ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ

ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

‘‘Ανάπτυξη μεθοδολογίας πρόβλεψης κατανάλωσης νερού από οικιακούς χρήστες, αξιοποιώντας δημογραφικά στοιχεία’’

Δημήτριος Ροδόπουλος

Α.Μ. 2673

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

Βασίλειος Ασημακόπουλος

Καθηγητής ΕΜΠ

Αθήνα, Οκτώβριος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ

ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

‘‘Ανάπτυξη μεθοδολογίας πρόβλεψης κατανάλωσης νερού από οικιακούς χρήστες, αξιοποιώντας δημογραφικά στοιχεία’’

Δημήτριος Ροδόπουλος

Α.Μ. 2673

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

Βασίλειος Ασημακόπουλος

Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Β. Ασημακόπουλος	Ι. Ψαρράς	Δ. Ασκούνης
Καθηγητής ΕΜ	Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

Δημήτριος Ροδόπουλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Τεχνολογίας Υπολογιστών

Πολυτεχνικής Σχολής Πανεπιστημίου Πατρών

Αθήνα, Οκτώβριος 2017

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του ΔΠΜΣ "Τεχνο-Οικονομικά Συστήματα" στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, στο Εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης.

Το αντικείμενο της εργασίας αφορά στην εύρεση μεθοδολογίας πρόβλεψης κατανάλωσης νερού από οικιακούς χρήστες, έπειτα από επιλογή και κατάλληλη επεξεργασία διαθέσιμων καταγεγραμμένων δεδομένων καταναλώσεων του παρελθόντος. Στόχος είναι έπειτα από δοκιμές και ελέγχους βελτιστοποίησης να προταθεί μία μέθοδος η οποία θα προβλέπει μελλοντικές μηνιαίες καταναλώσεις νερού, δεχόμενη ως είσοδο συγκεκριμένα δημογραφικά στοιχεία των νοικοκυριών για τα οποία πρόκειται να πραγματοποιήσει την πρόβλεψη.

Υπεύθυνος καθηγητής στην εκπόνηση της εργασίας ήταν ο κύριος Βασίλειος Ασημακόπουλος, στον οποίο οφείλω ιδιαίτερες ευχαριστίες για την αποδοχή του αιτήματος μου και την ανάθεση της συγκεκριμένης εργασίας, δίνοντας μου έτσι την ευκαιρία να ασχοληθώ με το πιο ενδιαφέρον κατά τη γνώμη μου αντικείμενο από όσα περιλαμβάνει ο οδηγός σπουδών του ΔΠΜΣ.

Θα ήθελα επίσης να ευχαριστήσω θερμά τον επιβλέποντα της εργασίας Ευάγγελο Σπηλιώτη, για τον χρόνο που αφιέρωσε και για την πολύτιμη συμβολή στην προσπάθεια να εκπονηθεί η παρούσα διπλωματική εργασία, η οποία είναι αποτέλεσμα της άψογης συνεργασίας που είχαμε όλο αυτό το διάστημα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για τη συμπαράσταση της σε όλη τη διάρκεια της φοίτησης μου στο ΔΠΜΣ, ώστε να φτάσω με επιτυχία στο τέλος αυτής της δύσκολης αλλά όμορφης διαδρομής, μέσα από ένα απαιτητικό καθημερινό πρόγραμμα, που συνδύαζε σπουδές και εργασία.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική αφορά στην ανάλυση δεδομένων με σκοπό την εξαγωγή χρήσιμων αποτελεσμάτων και την πραγματοποίηση προβλέψεων. Με την εφαρμογή μαθηματικών μοντέλων και αλγορίθμων, θα χρησιμοποιήσουμε τα δεδομένα που έχουμε στη διάθεση μας ώστε να δούμε εάν αυτά σχετίζονται μεταξύ τους και με βάση τα δεδομένα αυτά θα κάνουμε προβλέψεις, τις οποίες και θα αξιολογήσουμε.

Στόχος της εργασίας είναι να καταλήξουμε σε μια μεθοδολογία, η οποία να προβλέπει σε ικανοποιητικό βαθμό την κατανάλωση νερού οικιακών καταναλωτών βάσει συγκεκριμένων δημογραφικών στοιχείων. Η μεθοδολογία αυτή πρόβλεψης της ζήτησης νερού, σε συνδυασμό με κάποια μέθοδο εκτίμησης των διαθέσιμων αποθεμάτων σε ορισμένα χρονικά διαστήματα, μπορεί να δώσει πολύ χρήσιμα συμπεράσματα για τη μελλοντική επάρκεια του νερού για οικιακή χρήση σε κάποιον οικισμό, δήμο ή πόλη.

Αρχικά γίνεται μία εισαγωγή του προβλήματος με το οποίο θέλουμε να ασχοληθούμε. Το πρόβλημα της λειψυδρίας είναι ένα μείζον θέμα της εποχής μας το οποίο δεν αφορά μόνο τις αναπτυσσόμενες χώρες, αλλά όλον τον Πλανήτη. Παρουσιάζονται οι τρόποι με τους οποίους μπορεί να αντιμετωπιστεί η λειψυδρία, σε τεχνικό επίπεδο αλλά και σε επίπεδο σωστής διαχείρισης των αποθεμάτων νερού. Η σωστή διαχείριση αφορά στην εκτίμηση των αποθεμάτων και στην πρόβλεψη της μελλοντικής ζήτησης. Η πρόβλεψη είναι μία διαδικασία την οποία και θα μελετήσουμε στην παρούσα εργασία.

Έπειτα περιγράφονται θεωρητικά οι 2 μέθοδοι που θα χρησιμοποιηθούν στην ανάλυση μας. Η πρώτη είναι η μέθοδος της γραμμικής παλινδρόμησης. Η μέθοδος αυτή μας βοηθά όταν έχουμε στη διάθεση μας γνωστές τιμές ανεξάρτητων μεταβλητών και των αντίστοιχων εξαρτημένης μεταβλητής, να κατασκευάσουμε μία μαθηματική σχέση που συνδέει την εξαρτημένη μεταβλητή με τις ανεξάρτητες. Η δεύτερη μέθοδος είναι η διαδικασία της ανάλυσης συστάδων. Παρουσιάζονται οι διάφορες τεχνικές αυτής της μεθόδου με τα βήματα που ακολουθούνται, καθώς και τα πλεονεκτήματα και μειονεκτήματα της κάθε τεχνικής.

Στη συνέχεια μπαίνουμε στο πρακτικό κομμάτι της εργασίας. Σε πρώτη φάση παρουσιάζονται τα δεδομένα που έχουμε στη διάθεση μας, επιλέγουμε ποια από αυτά θα χρησιμοποιήσουμε για την ανάλυση μας και τα επεξεργαζόμαστε αναλόγως. Κατόπιν εφαρμόζουμε στα δεδομένα μας τη μέθοδο της γραμμικής παλινδρόμησης ώστε να κατασκευάσουμε ένα βέλτιστο γραμμικό μοντέλο που να εκφράζει τη σχέση ανάμεσα στην εξαρτημένη και τις ανεξάρτητες μεταβλητές. Η διαδικασία αυτή κάνοντας χρήση της R. Έπειτα πραγματοποιούμε ανάλυση συστάδων για τα δεδομένα μας επίσης με τη βοήθεια της R, η οποία είναι μία εναλλακτική προσέγγιση για να βγάλουμε χρήσιμα συμπεράσματα από τα διαθέσιμα δεδομένα και με βάση αυτά να πραγματοποιήσουμε προβλέψεις.

Τέλος, αξιολογούμε τις 2 μεθόδους που εφαρμόσαμε στα δεδομένα μας, υπολογίζουμε το σφάλμα πρόβλεψης για την καθεμία, εντοπίζουμε εάν υπάρχουν περιπτώσεις καταναλωτών για τις οποίες υπάρχει μεγάλο σφάλμα και προσπαθούμε να εξηγήσουμε τον λόγο.

ABSTRACT

The present project has to do with data analysis, with a final cause to extract useful results and make forecasts. Applying mathematical models and algorithms, we will use our available data in order to observe if there is any relation between that data and make forecasts that we are going to evaluate.

This project's purpose is to conclude in a method, which can forecast household water consumptions in a satisfying way, based on specific demographic characteristics. A water consumption forecasting method, combined with a method of estimating water supplies stock during some specific time frames, could give some very useful conclusions about future water sufficiency for household usage in a settlement, municipality or a whole city.

At first, we make an introduction to the problem that we want to deal with. Water scarcity is nowadays a major issue that concerns the whole planet. We present some methods to come up against water scarcity not only in a technical level, but also by using water supplies in a correct way. Correct water supplies management includes the estimation of water supply and forecasting future demand. Forecasting future water demand is the procedure that we study in the present project.

Afterwards, we implement a theoretical description of the 2 methods that that we will apply to our analysis. The first method is linear regression. This method helps us to form a mathematical relationship between the dependent variable and independent variables, as long as we have available a sample of known values for these variables. The second method is the procedure of cluster analysis. Various techniques for this kind of analysis are presented, step by step. The presentation includes the advantages and disadvantages of each technique.

After the theoretical analysis, we get into the practical part of the project. Firstly, we present our available data, we choose which of them will participate in our analysis and we make any necessary data processing. Thereafter we apply the multiple linear regression method to our available data, trying to form an optimum linear model that indicates the relationship between the dependent variable and independent variables. We implement this application using R. . After that we apply a cluster analysis to our available data, also using R. This kind of analysis is an alternative approach in order to extract useful conclusions from our data and make forecasts.

Finally we evaluate the 2 methods that we applied to our data, we determine the forecasting error, we detect if there are houses with increased value of that error and we try to explain the reason.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

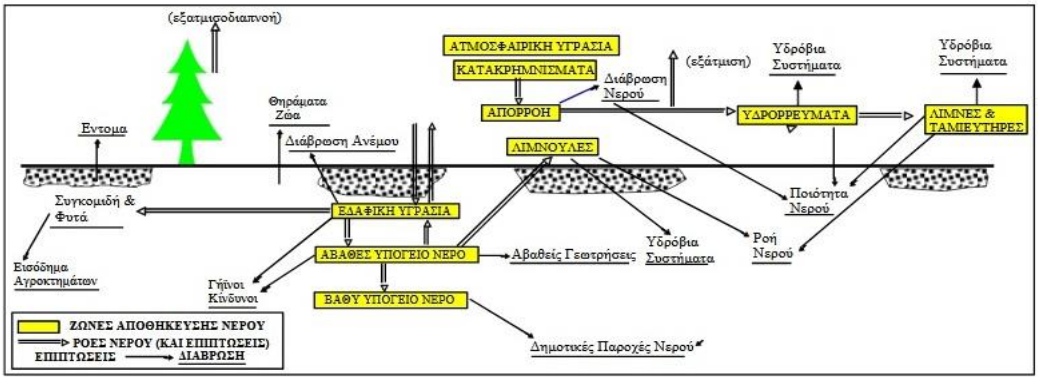
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ - ΠΑΡΟΥΣΙΑΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ -ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ

1.1 Εισαγωγή

Η **λειψυδρία** είναι ένα από τα βασικότερα προβλήματα που αντιμετωπίζει η ανθρωπότητα δημιουργώντας ένα δύσκολο παρόν και ένα αβέβαιο μέλλον. Ως λειψυδρία ορίζεται η έλλειψη ή η ανεπάρκεια του νερού, το οποίο προορίζεται για πόσιμο, για την καλλιέργεια των φυτών και γενικά για την χρήση του από τον άνθρωπο. Έλλειψη ή ανεπάρκεια νερού δημιουργείται είτε λόγω ξηρασίας, είτε από τη μη ορθολογική χρήση του στην περίπτωση που το έχουμε σε επάρκεια.

Η ξηρασία είναι ένα από τα πιο διαδεδομένα και ανησυχητικά προβλήματα στον πλανήτη, το οποίο πρέπει να αντιμετωπίσουν οι διαχειριστές νερού. Είναι ένα φαινόμενο το οποίο επιστημονικά ορίζεται ως εξής: ‘‘*Ένα χρονικό διάστημα, γενικά της τάξης των μηνών ή ετών σε διάρκεια, κατά το οποίο η πραγματική παροχή υγρασίας σε μια δεδομένη περιοχή υστερεί μάλλον σταθερά έναντι της διατιθέμενης παροχής υγρασίας από τον επικρατούντα κλιματικό τύπο.*’’ Θεωρείται το πιο πολύπλοκο υδρολογικό φαινόμενο και περιλαμβάνει ζητήματα που σχετίζονται με το κλίμα, τις χρήσεις γης, τους κανόνες χρήσεις νερού, καθώς και διαχειριστικά ζητήματα όπως η ετοιμότητα. Παρά το γεγονός ότι οι σοβαρές ξηρασίες είναι σπάνιες, στις ΗΠΑ παρατηρούνται κάθε χρόνο τουλάχιστον σε κάποιο βαθμό και σε άλλες χώρες οι θεομηνίες από ξηρασίες προκαλούν λιμούς και δυστυχία στις περισσότερες περιπτώσεις.

Η ξηρασία οφείλεται κατά βάση στη μειωμένη παροχή υγρασίας από την ατμόσφαιρα προς το έδαφος και τα ύδατα (ποτάμια, λίμνες, θάλασσες). Η μειωμένη αυτή παροχή υγρασίας με τη μορφή κατακρημνισμάτων (βροχή, χιόνι κτλ) επηρεάζει το υδρολογικό ισοζύγιο και κατά συνέπεια τον κύκλο του νερού. Στο σχήμα που ακολουθεί, φαίνεται ότι το έλλειμμα στα κατακρημνίσματα θα εμφανιστεί αργότερα στον υδρολογικό κύκλο.



Εικόνα 1*: Υδρολογικές συνθήκες που επηρεάζονται από την ξηρασία*

Ωστόσο, η ξηρασία δεν αποτελεί τη μοναδική αιτία για το πρόβλημα της λειψυδρίας. Εξαιτίας της ραγδαίας αύξησης του πληθυσμού της Γης, της μαζικής κατανάλωσης, της κατάχρησης των φυσικών πόρων, της ρύπανσης και μόλυνσης του νερού η διαθεσιμότητα του πόσιμου νερού δεν επαρκεί για να καλύψει τις ανάγκες της σύγχρονης εποχής και διαρκώς μειώνεται. Αυτό σημαίνει πως ακόμα και σε περιόδους μη ξηρασίας, παρότι η ποσότητα του νερού παραμένει σταθερή, άλλοι παράμετροι όπως η αύξηση του παγκόσμιου πληθυσμού, η βελτίωση του επιπέδου ζωής, η αλλαγή των καταναλωτικών συνηθειών, η επέκταση της αρδευόμενης γεωργίας και οι απαιτήσεις της βιομηχανίας **οδηγούν** στη διαρκή αύξηση της ζήτησης νερού διεθνώς και δυστυχώς αναμένεται να επιδεινώσουν το πρόβλημα μελλοντικά, με ότι αυτό συνεπάγεται. Για αυτό το λόγο, το νερό αποτελεί στρατηγικής σημασίας αγαθό σε όλη την υφήλιο.

Η ανεπάρκεια στην παροχή νερού μπορεί να έχει σοβαρές επιπτώσεις σε πόλεις, βιομηχανίες, και σε άλλους χρήστες νερού όπως άρδευση, παραγωγή ηλεκτρικής ενέργειας και άγρια ζωή. Παρά το γεγονός ότι η αποτελεσματική διαχείριση των υδατικών πόρων δεν μπορεί να αντιμετωπίσει ολοκληρωτικά τις ελλείψεις νερού, αυτή μπορεί να ελαχιστοποιήσει τα προβλήματα. Ελάχιστα εγχειρίδια και εκθέσεις παρέχουν οδηγίες για την αντιμετώπιση της ξηρασίας και οι ερευνητικές αναφορές ασχολούνται κυρίως με τη ξηρασία μετά την εμφάνισή της. Η ξηρασία είναι ένα κλιματολογικό χαρακτηριστικό. Το ότι αυτή αποτελεί ένα σοβαρό και συνεχιζόμενο παγκόσμιο ζήτημα, οφείλεται κατά κύριο λόγο στην έλλειψη αποτελεσματικού σχεδιασμού και διαχείρισης παρά στο κλίμα. Καμία περιοχή δεν είναι απαλλαγμένη από τα προβλήματα της λειψυδρίας.

1.2 Επιπτώσεις της λειψυδρίας στην ανθρώπινη ζωή

Σήμερα ένα ποσοστό περίπου 40% από τους ανθρώπους που ζουν στη γη δεν έχει επαρκές νερό ακόμα και για υποτυπώδη υγιεινή . Περισσότεροι από 2,2 εκατομμύρια άνθρωποι πέθαναν το 2000 από ασθένειες που σχετίζονται με την κατανάλωση μολυσμένου νερού ή με ξηρασία. Η εξάντληση των υδάτινων πόρων μετατρέπει το νερό σε πολύτιμο αγαθό για εμάς. Ήδη κάποια έθνη ήρθαν σε αντιπαραθέσεις για τον λόγο αυτό και προβλέπεται ότι στο μέλλον θα ξεσπάσουν διαμάχες ανάμεσα στα έθνη για την διεκδίκηση ποταμών και άλλων πραγμάτων που έχουν σχέση με το νερό.

Περίπου 500 εκατομμύρια άνθρωποι στον πλανήτη ζουν σε περιοχές, όπου η κατανάλωση νερού είναι διπλάσια από ότι η αναπλήρωση του γλυκού νερού μέσω των βροχών. Επιστήμονες προβλέπουν ότι το 2025 1,8 δισεκατομμύρια άνθρωποι θα ζουν σε χώρες ή περιοχές με πλήρη λειψυδρία, ενώ τα δυο τρίτα του παγκοσμίου πληθυσμού θα ζουν κάτω από δύσκολες συνθήκες. Η έλλειψη νερού φυσικά δεν αφορά μόνο τις χώρες της Αφρικής. Σημαντικός περιορισμός σε υδάτινα αποθέματα , υπάρχει ακόμα και στις πιο ανεπτυγμένες χώρες.

1.3 Χρήση νερού στη βιομηχανική και αγροτική παραγωγή

Η ποσότητα νερού που χρησιμοποιείται για βιομηχανική και αγροτική παραγωγή είναι τεράστια. Για την παραγωγή μόλις ενός κιλού κρέατος χρειάζονται 15.000 λίτρα νερού (σχεδόν όλο αυτό το νερό διοχετεύεται στην άρδευση των καλλιεργειών που παράγουν τις ζωοτροφές), ενώ για την παραγωγή ίδιας ποσότητας κρέατος από πουλερικά χρειάζονται 6.000 λίτρα νερού. Επίσης για την παραγωγή ενός κιλού εσπεριδοειδών, χρειάζονται περίπου 1000 λίτρα νερού.

Η κατανάλωση νερού στην βιομηχανική παραγωγή είναι επίσης σημαντική: Για έναν τόνο αλουμινίου χρειάζονται 1.500 λίτρα ενώ για έναν τόνο ατσάλι περίπου 6.000. Παρά τις όποιες προσπάθειες για μείωση της κατανάλωση διαμέσου αύξησης της αποδοτικότητας, η βιομηχανική παραγωγή είναι υπεύθυνη για την κατανάλωση του 22% της συνολικής ποσότητας νερού.

1.4 Το πρόβλημα της λειψυδρίας στον Πλανήτη

Σοβαρό πρόβλημα **λειψυδρίας** αντιμετωπίζουν σχεδόν τα **δύο τρίτα του πληθυσμού** (περίπου 4 δισεκατομμύρια άνθρωποι) για τουλάχιστον ένα μήνα κάθε χρόνο. Από τα συνολικά 4 δισεκατομμύρια αυτών των ανθρώπων, τα 2 δισεκατομμύρια ζουν στην **Κίνα** και στην **Ινδία.** Η ευρύτερη περιοχή της Μέσης Ανατολής καλύπτει τις ανάγκες της σε μεγάλο βαθμό αντλώντας από τα υπόγεια ύδατα και με αφαλατωμένο θαλασσινό νερό. Είναι προφανές πως λιγότερο νερό για τον διαρκώς αυξανόμενο πληθυσμό στη Μέση Ανατολή σημαίνει και μεγαλύτερη αστάθεια στην περιοχή. Ορισμένες χώρες μπορεί να οδηγηθούν στη μείωση της παραγωγής τροφίμων για να μειωθεί και η κατανάλωση νερού, μεγάλο μέρος του οποίου δαπανάται σε καλλιέργειες. Μάλιστα, ήδη η Σαουδική Αραβία ανακοίνωσε ότι από το 2016 ο πληθυσμός της θα εξαρτάται εξολοκλήρου από τις εισαγωγές σιτηρών. **Σοβαρό πρόβλημα αντιμετωπίζει και το μεγαλύτερο μέρος του πληθυσμού της Αφρικής. Πολλοί εξ' αυτών, προκειμένου** να βρουν νερό πηγαίνουν σε πηγάδια, ποταμούς και λίμνες που είναι αρκετά χιλιόμετρα μακριά από το σπίτι τους. Ωστόσο το νερό στην Αφρική είναι αρκετά βρώμικο και πολλοί άνθρωποι κάθε χρόνο χάνουν τη ζωή τους καταναλώνοντας το.

Η έλλειψη νερού βέβαια δεν αφορά μόνο τις χώρες της Ασίας και της Αφρικής. Σημαντικός περιορισμός σε υδάτινα αποθέματα, υπάρχει ακόμα και στις πιο ανεπτυγμένες χώρες. Έλλειψη νερού σε ανησυχητικά επίπεδα σημειώνεται και στις **ΗΠΑ**, ιδιαίτερα στις πολιτείες Καλιφόρνια, Τέξας και Φλόριντα, καθώς και στην **Αυστραλία**. Ο μέσος όρος των βροχοπτώσεων τα τελευταία χρόνια έχει μειωθεί κατά 30% στην δυτική ακτή της Αυστραλίας και κατά 15% στην ανατολική ακτή. Οι κάτοικοι της πολιτείας Κουήνσλαντ, στο βορειοανατολικό άκρο της ηπείρου, σύντομα θα καταναλώνουν ανακυκλωμένο νερό.

Από το πρόβλημα της λειψυδρίας φυσικά δεν είναι απαλλαγμένες και οι χώρες της Ευρώπης, συμπεριλαμβανομένης και της Ελλάδας. **Σε πολλές περιοχές η ισορροπία μεταξύ ζήτησης και διαθεσιμότητας νερού, έχει φτάσει σε κρίσιμο σημείο. Η γεωργία παραμένει ο σημαντικότερος χρήστης νερού (με 64%), ακολουθούμενη από την ενέργεια (20%), τη δημόσια παροχή νερού (12%) και τη βιομηχανία (4%). Μέχρι σήμερα έχουν προσδιοριστεί** 33 λεκάνες ποταμών που επηρεάζονται από την έλλειψη νερού. Αντιπροσωπεύουν μία συνολική έκταση 460.000 τετραγωνικών χιλιομέτρων (περίπου το 10% της συνολικής έκτασης της ΕΕ) και φιλοξενούν έναν συνολικό πληθυσμό 82 εκατομμυρίων ανθρώπων (περίπου το 16,5% του συνολικού πληθυσμού της Ε.Ε.). Από το 2000 ως το 2006 κατά μέσο όρο το 15% της συνολικής έκτασης της ΕΕ και κατά μέσο όρο το 17% του συνολικού πληθυσμού της Ευρωπαϊκής Ένωσης επηρεάστηκαν από τις ξηρασίες. Επίσης, ορισμένα κράτη μέλη αντιμετωπίζουν ήδη προβλήματα μόνιμης λειψυδρίας σε ολόκληρη την επικράτεια τους. Η Τσεχική Δημοκρατία ανέφερε την ύπαρξη περιοχών με συχνή λειψυδρία, ενώ η Γαλλία και το Βέλγιο επισήμαναν την υπερεκμετάλλευση υδροφόρων οριζόντων. Μελέτες κρούουν τον κώδωνα του κινδύνου ακόμα και για χώρες όπως η Αγγλία, αναφέροντας συγκεκριμένα ότι στο Λονδίνο η κατανάλωση νερού δεν είναι μακροπρόθεσμα βιώσιμη στα τρέχοντα επίπεδα.

1.5 Το πρόβλημα της λειψυδρίας στην Ελλάδα

Πολύ σοβαρό πρόβλημα λειψυδρίας αντιμετωπίζει και η χώρα μας, η οποία μάλιστα βρίσκεται ‘‘στο κόκκινο’’ σχετικά με τα αποθέματα των επιφανειακών της υδάτων, μαζί με άλλες 32 χώρες παγκοσμίως βάσει μελέτης που εκπονήθηκε από το Ινστιτούτο Παγκόσμιων Πόρων το 2015. Την ακρότητα του φαινομένου αντικατοπτρίζει η Αθήνα. Η πρωτεύουσα έχει συγκεντρώσει περίπου το 50% του πληθυσμού της χώρας. Όλοι οι υδάτινοι πόροι γύρω από την Αθήνα, αλλά και από πολύ μακρινές αποστάσεις, χρησιμοποιούνται για να καλύψουν τις ανάγκες αυτών των κατοίκων. Τα νερά της Υλίκης στην Βοιωτία και του Μόρνου, που έρχονται από πολύ μεγάλη απόσταση, επαρκούν οριακά και δεν είναι απίθανο στο άμεσο μέλλον να αναζητηθούν και νέα αποθέματα.

Ωστόσο εκτός από την Αθήνα και την ευρύτερη περιοχή της Αττικής, υπάρχουν και περιοχές όπως οι νομοί Κιλκίς και Αργολίδας που απειλούνται έντονα από την έλλειψη νερού. Η λειψυδρία αποτελεί τον μόνιμο πρόβλημα των κατοίκων και πολλών νησιών του Αιγαίου, που πασχίζουν με διάφορους τρόπους να αξιοποιούν τα νερά που τους προσφέρουν οι ελάχιστες βροχοπτώσεις του χειμώνα. Το καλοκαίρι πάλι, μεταφέρεται νερό από άλλες περιοχές με την βοήθεια ειδικών πλοίων (υδροφόρων). Σε πολλά νησιά, και κυρίως στην Κρήτη, από αιώνες χρησιμοποιούσαν

τη δύναμη του ανέμου για να αντλούν τα υπόγεια αποθέματα νερού, ώστε να αντιμετωπίσουν αυτή την έλλειψη. Σήμερα όμως, το πρόβλημα της λειψυδρίας έχει πάρει πολύ μεγάλες διαστάσεις.

Η αλλαγή των συνθηκών διαβίωσης, η αύξηση του πληθυσμού και του τουρισμού καθώς και η υπερκατανάλωση έχουν οδηγήσει και τα νησιά του Νότιου Αιγαίου σε έλλειψη υδάτινων πόρων. Έπειτα , οι γεωτρήσεις ξεπέρασαν τα 500 μέτρα βάθους κάτι που μέχρι τότε δεν είχε ξαναγίνει και τα υπόγεια νερά εξαντλήθηκαν ή υποβαθμίστηκαν . Έτσι , ξεκίνησαν σιγά-σιγά να γίνονται κάποια έργα και να δημιουργούνται καινούριοι κλάδοι που ασχολούνταν με το νερό και την εύρεση και την αποθήκευσή του . Πιο συγκεκριμένα , από το 1980 ξεκίνησαν σιγά-σιγά να κατασκευάζονται φράγματα και υδραγωγεία ώστε το νερό να συγκεντρώνεται σε μεγάλη ποσότητα και να αξιοποιείται αναλόγως στις καλλιέργειες και στα σπιτικά των ανθρώπων. Ακόμα, φορτηγά-πλοία άρχισαν τη μεταφορά και διανομή υπερμεγεθών όγκων νερού , από άλλες περιοχές και χώρες. Έπειτα, οι δήμοι των περιοχών αυτών του Αιγαίου εγκατέστησαν σύγχρονες μονάδες αφαλάτωσης , ώστε να μπορούν να κάνουν πόσιμο και χρήσιμο το νερό εκείνο, που πριν ήταν πλήρες σε άλατα , λάσπη και άλλα ανθυγιεινά στοιχεία.

Όμως η οικονομική κρίση των τελευταίων 8 ετών έχει χτυπήσει δυνατά την πόρτα της Ελλάδας . Οι δήμοι δε διαθέτουν τα απαιτούμενα χρήματα και εφόδια για να τα διαθέσουν ώστε να συνεχίσει η ανάπτυξη και η δημιουργία . Για αυτό έχει σαν αποτέλεσμα , τα έργα που έχουν κατασκευαστεί εδώ και χρόνια να μην μπορούν να συντηρηθούν και να συνεχίσουν τη σωστή λειτουργία τους . Δεν είναι λίγες οι φορές που οι γεωτρήσεις στερεύουν και οι παραγωγοί καταστρέφονται οικονομικά. Οι καλλιέργειες που εφαρμόζονται στις περισσότερες περιοχές, όπως για παράδειγμα το βαμβάκι, απαιτούν μεγάλες ποσότητες νερού. Οι παραδοσιακές καλλιέργειες που δεν είχαν αρδευτικές απαιτήσεις, όπως τα δημητριακά και τα όσπρια, δεν προσφέρουν στους αγρότες ικανοποιητικά ποσοστά κέρδους. Για το λόγο αυτό, όπου υπάρχει η δυνατότητα ποτίσματος, έχουν σχεδόν εγκαταλειφθεί, δίνοντας τη θέση τους σ' αυτές που με τη βοήθεια του νερού αποδίδουν υψηλά οφέλη.

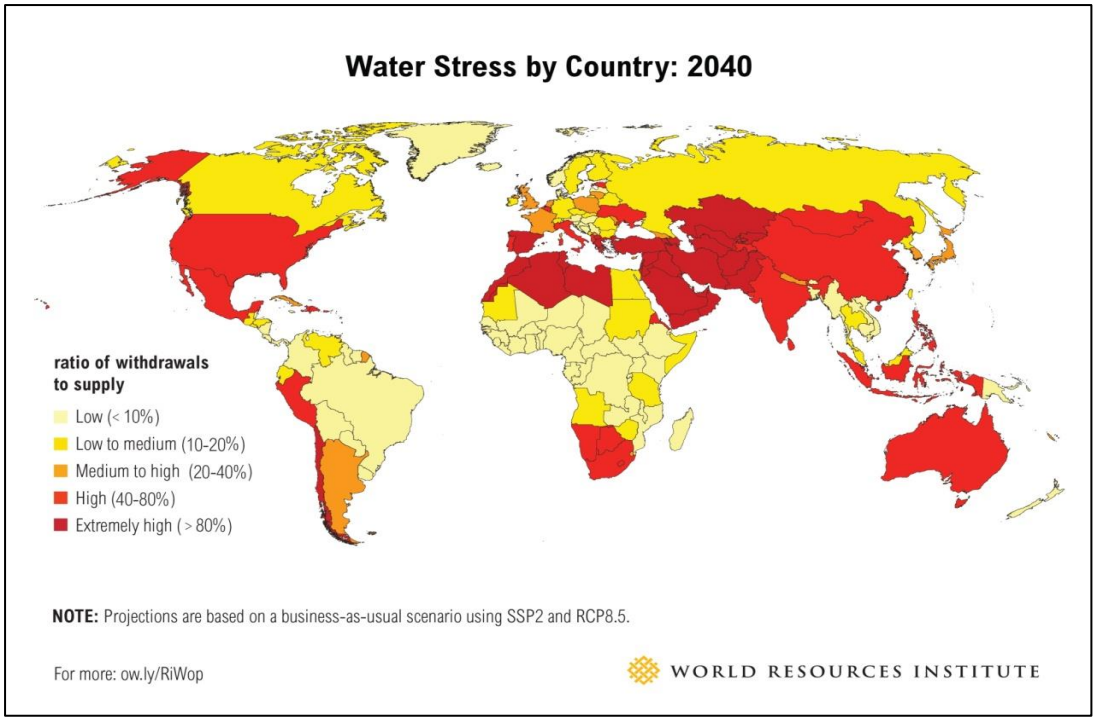
Πέραν όλων αυτών των παραγόντων που συντελούν στην έλλειψη νερού, τα τελευταία χρόνια παρατηρείται και μία διαταραχή των κλιματολογικών συνθηκών στη χώρα μας. Αυτό οφείλεται σε πολλούς παράγοντες, όπως η τρύπα του όζοντος και το φαινόμενο του θερμοκηπίου. Κυριότεροι όμως παράγοντες, που επηρεάζουν το κλίμα της χώρας μας, είναι δυστυχώς οι ανθρώπινες επεμβάσεις. Το κάψιμο των δασών, οι αποξηράνσεις των λιμνών, η συσσώρευση βιομηχανικών μονάδων σε μια περιοχή επιδρούν άμεσα. Η ατμόσφαιρα μολύνεται με την συσσώρευση του διοξειδίου του άνθρακα, που επιτείνει το φαινόμενο του θερμοκηπίου, τα ανύπαρκτα δάση δεν «τραβούν» τις βροχές και δεν μπορούν να συγκρατήσουν τον όγκο των βροχοπτώσεων, το στέγνωμα των λιμνών συνεπάγεται μείωση της υγρασίας στην περιοχή και σταδιακή υποχώρηση του υδροφόρου ορίζοντα. Όλα αυτά προκαλούν μεγάλες μεταβολές στο κλίμα της χώρας. Το αποτέλεσμα είναι σε άλλες περιοχές να έχουμε πλημμύρες, και αλλού να βρέχει σπάνια. Το τελικό ισοζύγιο είναι μάλλον

αρνητικό, αφού έχουν λιγοστέψει οι χιονοπτώσεις που συμβάλλουν στην ανανέωση των υπόγειων δεξαμενών. Αλλά και η έλλειψη βροχοπτώσεων τον καιρό που χρειάζονται (φθινόπωρο, άνοιξη), δημιουργούν μεγάλα προβλήματα τόσο στην καλλιέργεια της γης (όργωμα, σπορά), όσο και στην απόδοση της καλλιέργειας (παραγωγή) .

Επιπλέον, η μορφολογία της Ελλάδας δεν επιτρέπει τη δημιουργία μεγάλων ποταμών, γιατί τα νερά των βροχοπτώσεων διασχίζουν μικρές αποστάσεις και καταλήγουν τελικά στη θάλασσα. Τα λίγα μεγάλα ποτάμια που ευτυχώς διασχίζουν τη χώρα μας (Αχελώος, Αξιός κλπ), πηγάζουν και διασχίζουν σε μεγάλο μήκος τις χώρες των Βαλκανίων. Αυτό έχει ως αποτέλεσμα τη σχετική επάρκεια του νερού στις περιοχές της βόρειας και της δυτικής Ελλάδας. Στη Θεσσαλία όμως και την Ανατολική Στερεά, όπου βρίσκονται σημαντικές καλλιεργήσιμες εκτάσεις, αλλά και σε άλλες περιοχές, παρατηρείται έλλειψη των επιφανειακών υδάτινων πόρων.

Όλα τα παραπάνω μας δίνουν σε γενικές γραμμές τις διαστάσεις της λειψυδρίας στην Ελλάδα του σήμερα. Είναι ένα μεγάλο πρόβλημα που θα πρέπει να βρει μια άμεση και σωστή λύση. Οι ρυθμοί με τους οποίους ελαττώνονται αυτοί οι φυσικοί πόροι είναι ανησυχητικοί. Κάθε χρόνος που περνά, προσθέτει δυσαναπλήρωτα κενά και κάνει το πρόβλημα εντονότερο. Και φυσικά δεν μπορεί ατομικά κάθε ενδιαφερόμενος να λύσει το πρόβλημα , είτε είναι αγρότης, είτε είναι πόλη ή χωριό. Είναι ανάγκη να βρεθούν οι λύσεις, να γίνει ένας σχεδιασμός σε πανεθνικό επίπεδο και να εφαρμοστεί με υπευθυνότητα και αμεροληψία. Δεν πρέπει σε καμία περίπτωση να αδιαφορούμε μπροστά στα προβλήματα του τόπου μας, όταν νοιώθουμε ότι δεν μας αγγίζουν άμεσα. Αν σκεφτούμε ότι οι δαπάνες π.χ. των αγροτών για το πότισμα των καλλιεργειών έχουν σαν αποτέλεσμα την επιβάρυνση της τιμής των προϊόντων, τότε καταλαβαίνουμε ότι η λειψυδρία είναι πρόβλημα του κάθε πολίτη.

Σύμφωνα με μελέτη που εκπονήθηκε το 2014 από ερευνητές από το Ινστιτούτο Παγκόσμιων Πόρων, χρησιμοποιώντας ένα σύνολο κλιματικών μοντέλων και κοινωνικοοικονομικών σεναρίων και κατατάσσει 167 χώρες ανάλογα με τα αποθέματα των επιφανειακών τους υδάτων, η Ελλάδα βρίσκεται μαζί με άλλες 32 χώρες (στο γκρουπ των χωρών που εμφανίζουν τη μεγαλύτερη ανησυχία και αναμένεται να αντιμετωπίσουν σοβαρό πρόβλημα λειψυδρίας μέχρι το 2040. Εννέα από αυτές τις χώρες, που εμφανίζονται να αντιμετωπίζουν σοβαρό πρόβλημα έλλειψης νερού τα επόμενα 25 χρόνια βρίσκονται στη Μέση Ανατολή (Μπαχρέιν, Κουβέιτ, Κατάρ, Ηνωμένα Αραβικά Εμιράτα, Σαουδική Αραβία, Ομάν, Λίβανος, Ισραήλ και τα Παλαιστινιακά Εδάφη) Πέραν αυτού, σε παγκόσμιο επίπεδο εντείνεται ο προβληματισμός για την έλλειψη σταθερότητας που ήδη παρατηρείται σε ορισμένες περιοχές εξαιτίας χρόνιων συγκρούσεων, για τις οποίες μία σοβαρή αιτία είναι η κυριαρχία στην εκμετάλλευση των υδάτινων αποθεμάτων. Τα αποτελέσματα της μελέτης αποτυπώνονται στον επόμενο παγκόσμιο χάρτη, όπου οι χώρες έχουν επισημανθεί με αντίστοιχο χρώμα, ανάλογα με τον κίνδυνο που εμφανίζουν στην έλλειψη υδάτινων αποθεμάτων.



***Εικόνα 1**: Παγκόσμιος χάρτης κατηγοριοποίησης των χωρών, σε σχέση με τον κίνδυνο αντιμετώπισης λειψυδρίας μέχρι το 2040*

1.6 Αντιμετώπιση του προβλήματος

Διάφορες προσεγγίσεις έχουν προταθεί από τους επιστήμονες σχετικά με την εκμετάλλευση των υδάτινων πόρων και την αντιμετώπιση της λειψυδρίας. Η πιο απλή πρόταση είναι στις περιοχές εκείνες όπου το επιτρέπει μορφολογία του εδάφους, να κατασκευαστούν μικρά φράγματα για την αποθήκευση του νερού των χειμάρρων. Έτσι, σε κάθε περιοχή θα χρησιμοποιούνται κοντινοί πόροι, που θα ισορροπήσουν αρκετά το έλλειμμα που παρουσιάζεται, με αποτέλεσμα να ελαττωθεί σε ικανοποιητικό βαθμό η χρήση των γεωτρήσεων. Ταυτόχρονα, η λεκάνη συγκέντρωσης του νερού θα εμπλουτίσει τον υδροφόρο ορίζοντα της περιοχής και θα προσφέρει νέες προοπτικές για ανάπτυξη της χλωρίδας και της πανίδας γύρω από αυτή.

Μία άλλη πρόταση που έχει διατυπωθεί είναι η αξιοποίηση των όμβριων υδάτων που συγκεντρώνονται στα ειδικά αποχετευτικά δίκτυα των πόλεων. Μέσα από τον κατάλληλο σχεδιασμό είναι δυνατό να συγκεντρώνονται τα νερά αυτά για να χρησιμοποιούνται στην βιομηχανία ή την άρδευση. Αν σκεφτεί κανείς την έκταση που καταλαμβάνει μια μέτριου μεγέθους πόλη, μπορεί να καταλάβει πόσες χιλιάδες κυβικά μέτρα νερού χάνονται κάθε φορά που βρέχει, αφού το τσιμέντο των πόλεων δεν απορροφά σχεδόν τίποτα. Επίσης, μία από τις επικρατέστερες προτάσεις είναι και

η αφαλάτωση του θαλασσινού νερού, ωστόσο αυτή η λύση δεν είναι πανάκεια και σίγουρα δεν είναι αθώα. Για τη δημιουργία ενός κυβικού γλυκού νερού απαιτείται άντληση τριών περίπου κυβικών θαλασσινού ή υφάλμυρου νερού. Το υπολειπόμενο από την επεξεργασία νερό απορρίπτεται πάλι στη θάλασσα με τη μορφή του αλμόλοιπου. Αυτό είναι αυξημένης αλατότητας και επιβαρυμένο με σημαντικές ποσότητες χλωρίνης και χημικών. Ίσως δεν είναι ευρέως γνωστό, αλλά το αντλούμενο νερό προχλωριώνεται για την προστασία των μεμβρανών που χρησιμοποιούνται ως φίλτρα. Τα αποπλύματα των μεμβρανών των συστημάτων αφαλάτωσης θα απορρίπτονται και αυτά στη θάλασσα μαζί με το αλμόλοιπο. Όταν αυτό γίνεται επαναλαμβανόμενα σε βάθος χρόνου, είναι δύσκολο κανείς να προσδιορίσει τις μακροπρόθεσμες συνέπειες μια τέτοιας πρακτικής.

Πέρα όμως από την εξεύρεση υδάτινων πόρων, σημαντικό θέμα αποτελεί η δυνατότητα περιορισμού των απαιτήσεων σε νερό. Δε νοείται να καταφεύγουμε σε τεχνικές λύσεις όπως αυτές που αναφέρθηκαν στην προηγούμενη παράγραφο, όταν δεν έχουμε προηγουμένως προσπαθήσει να καθιερώσουμε ένα πιο οικολογικό προφίλ και να προσαρμόσουμε τις ανάγκες μας έτσι ώστε να αποφεύγουμε τη σπατάλη νερού. Η μείωση των απαιτήσεων σε νερό, μπορεί να επιτευχθεί με διάφορους τρόπους. Σε ό,τι αφορά τους οικιακούς καταναλωτές, ισχυρό κίνητρο για τον περιορισμό της σπατάλης θα μπορούσε να αποτελέσει για παράδειγμα η μεγάλη αύξηση της τιμής του νερού για πάνω από ορισμένη ποσότητα, ανάλογα βέβαια και με τον αριθμό των προσώπων που κατοικούν σε κάθε σπίτι. Στον γεωργικό τομέα, θα πρέπει η πολιτεία να δει τρόπους ενίσχυσης των καλλιεργειών που δεν έχουν καθόλου ή έχουν ελάχιστες απαιτήσεις σε νερό.

1.7 Σκοπός της παρούσας εργασίας

Προκειμένου να αναπτυχθεί από το καταναλωτικό κοινό μία κουλτούρα αποφυγής της σπατάλης και να καθιερωθεί ένας τρόπος ζωής στον οποίο η ζήτηση για το νερό δε θα ξεπερνά τη διαθέσιμη παροχή νερού, απαιτείται ένας ενιαίος σχεδιασμός και ορθότητα επιλογών μέσα από περιβαλλοντικά και κοινωνικά κριτήρια.

Οι διαχειριστές νερού, θα πρέπει να κάνουν ανάλυση της ζήτησης από τους καταναλωτές που εξυπηρετούν. Θα πρέπει να διαπιστώσουν εάν υπάρχουν περιπτώσεις όπου δεν είναι επαρκής η διαθέσιμη ποσότητα νερού για να ικανοποιηθεί η ζήτηση, επειδή οι παροχές νερού πέφτουν κάτω από τα προσδοκώμενα επίπεδα. Τα προσδοκώμενα επίπεδα είναι κοινωνικοοικονομική έννοια, επειδή οι προσδοκίες μπορούν να διευθετηθούν.

Η έλλειψη νερού, που οφείλεται στην ξηρασία καθώς και σε αυξημένες απαιτήσεις, είναι στο μεγαλύτερο μέρος, ένα πολύπλοκο διαχειριστικό πρόβλημα με ένα μεγάλο αριθμό δραστηριοποιημένων ατόμων. Η προετοιμασία για την αντιμετώπιση της λειψυδρίας απαιτεί ατομική και συλλογική δράση για την διασφάλιση των αναγκαίων παροχών νερού και για να γίνουν εγκαίρως τα σχέδια παροχής και συντήρησης όταν οι παροχές είναι ανεπαρκείς. Οποιαδήποτε διαχείριση λειψυδρίας ή μεθοδολογία σχεδιασμού της παροχής νερού απαιτούν να είναι γνωστή η σχέση ανάμεσα στην προσδοκώμενη παροχή και στη ζήτηση για να καθηλώσουμε σε σταθερά επίπεδα τον κίνδυνο της ανεπάρκειας στην παροχή νερού. Ένα σχέδιο ανάλυσης επάρκειας παροχής νερού θα πρέπει να προβλέπει και την παροχή και τη ζήτηση. Εάν όλες οι εγκαταστάσεις αποθήκευσης δεν αδειάζουν ποτέ αρκετά κατά την περίοδο της ανάλυσης, σύμφωνα με την πρόβλεψη, τότε η παροχή θεωρείται επαρκής.

Στην παρούσα εργασία θα επιχειρήσουμε να καλύψουμε το κομμάτι της πρόβλεψης της ζήτησης μέσα από κοινωνικά κριτήρια και δημογραφικά στοιχεία. Μέσα από ένα πλήθος νοικοκυριών για τα οποία γνωρίζουμε την κατανάλωση τους καθώς και αρκετά δημογραφικά χαρακτηριστικά, θα διερευνήσουμε εάν υπάρχει σχέση ανάμεσα στη ζήτηση νερού και στα δημογραφικά χαρακτηριστικά . Έπειτα θα μοντελοποιήσουμε την κατανάλωση νερού με παράμετρο αυτά τα χαρακτηριστικά. Η μοντελοποίηση αυτή θα μας βοηθήσει στο να ταξινομήσουμε κατάλληλα τα νοικοκυριά και να εντοπίσουμε εάν κάποια από αυτά ξεφεύγουν από τις απαιτήσεις νερού σε σχέση με τα υπόλοιπα νοικοκυριά της ανάλυσης. Στην περίπτωση αυτή μπορεί ο εκάστοτε διαχειριστής νερού να λάβει τα κατάλληλα μέτρα. Πέραν τούτου όμως, τα μοντέλα που θα αναπτύξουμε θα μας βοηθήσουν να κάνουμε και πρόβλεψη για μελλοντική ζήτηση σε νερό, με βάση πάντα τα δημογραφικά στοιχεία ενός νοικοκυριού.

Η πρόβλεψη αυτή, εφόσον κριθεί ικανοποιητική, θα μπορούσε σε συνδυασμό με κάποια μελέτη πρόβλεψης παροχής νερού να αποτελέσουν ένα σημαντικό εργαλείο για την πραγματοποίηση ανάλυσης επάρκειας παροχής νερού. Έτσι ο διαχειριστής θα μπορεί να διαπιστώσει εάν η παροχή νερού που διαθέτει είναι επαρκής καθ' όλη τη διάρκεια του χρόνου και αν θα χρειαστεί να αντλήσει αποθέματα από άλλες αποθήκες. Έτσι, θα έχει τη δυνατότητα να προγραμματίσει τις απαιτήσεις και τη συντήρηση του δικτύου του, αλλά και να χαράξει αντίστοιχη πολιτική ανάλογα με την καταναλωτική συμπεριφορά του κοινού που αυτός εξυπηρετεί.

ΚΕΦΑΛΑΙΟ 2: ΑΝΑΛΥΣΗ ΜΕΘΟΔΟΥ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα αναλυθεί η πρώτη από τις δύο διαφορετικές προσεγγίσεις με τις οποίες στη συνέχεια θα επεξεργαστούμε τα δεδομένα μας και βάσει αυτής της επεξεργασίας θα πραγματοποιήσουμε προβλέψεις.

Η γραμμική παλινδρόμηση αποτελεί ένα χρήσιμο εργαλείο στο να κατασκευάζουμε μοντέλα που θα πραγματοποιούν ποσοτικές προβλέψεις. Είναι πολύ χρήσιμη στην επιστήμη της στατιστικής και έχει αναλυθεί εκτενώς σε πάρα πολλά επιστημονικά συγγράμματα. Παρότι η επιστήμη της στατιστικής έχει εξάγει πολλές νέες μεθόδους πρόβλεψης και ανάλυσης, η γραμμική παλινδρόμηση είναι μία πολύ χρήσιμη και ευρέως χρησιμοποιούμενη μέθοδος. Επιπλέον, μπορούμε να πούμε ότι αποτελεί ένα σημείο εκκίνησης για την εξαγωγή νέων μεθόδων και προσεγγίσεων, καθώς πολλές μοντέρνες μέθοδοι στατιστικής ανάλυσης και πρόβλεψης μπορούν να χαρακτηριστούν ως γενικεύσεις ή επεκτάσεις της γραμμικής παλινδρόμησης. Συνεπώς θα ήταν χρήσιμο να εξηγήσουμε και να αναλύσουμε τη μέθοδο αυτή πριν προχωρήσουμε στην παρουσίαση και την αντιμετώπιση του προβλήματος με το οποίο έχουμε επιλέξει να ασχοληθούμε στην παρούσα εργασία.

2.2 Απλή Γραμμική Παλινδρόμηση

Σε γενικές γραμμές, η χρησιμότητα της μεθόδου γραμμικής παλινδρόμησης είναι στο ότι μας βοηθάει να συμπεράνουμε αν υπάρχει συσχέτιση ή όχι ανάμεσα σε μία ανεξάρτητη μεταβλητή (ή περισσότερες) και σε μία εξαρτημένη μεταβλητή/αποτέλεσμα. Για να γίνει πιο κατανοητή η μέθοδος αυτή καθώς και οι διαδικασίες που τη διέπουν, θα εξετάσουμε το παράδειγμα μίας εταιρίας η οποία θέλει να διερευνήσει το κατά πόσον οι πωλήσεις ενός προϊόντος της (αποτέλεσμα) σχετίζονται με τη διαφημιστική δαπάνη στην τηλεόραση. Με τη μέθοδο της γραμμικής παλινδρόμησης ένας αναλυτής μπορεί να συμπεράνει εάν υπάρχει συσχέτιση ανάμεσα στην ανεξάρτητη μεταβλητή που είναι το τι ποσά δαπανώνται σε τηλεοπτική διαφήμιση του προϊόντος και στο τι πωλήσεις παρουσιάζει το προϊόν (π.χ. εάν αύξηση της διαφημιστικής δαπάνης οδηγεί σε αύξηση των πωλήσεων) και πόσο ισχυρή είναι η συσχέτιση αυτή, καθώς και να εξετάσει αν η ανεξάρτητη και η εξαρτημένη μεταβλητή σχετίζονται γραμμικά.

Η απλή γραμμική παλινδρόμηση λοιπόν κάνει αυτό ακριβώς που λέει το όνομα της. Είναι μία ευθεία προσέγγιση που προβλέπει μία ποσοτική απόκριση Y (εξαρτημένη μεταβλητή) με βάση μία μόνο ανεξάρτητη μεταβλητή X. Υποθέτει δηλαδή ότι υπάρχει γραμμική σχέση ανάμεσα σε X και Y και κάνει προβλέψεις που βασίζονται σε αυτήν τη σχέση. Μαθηματικά, η γραμμική αυτή σχέση εκφράζεται ως εξής:

$$\boxed{\mathbf{Y = \alpha + \beta \cdot x}} \qquad \textbf{(2.1)}$$

Στην παραπάνω σχέση τα α και β αποτελούν τους συντελεστές ή παραμέτρους της γραμμικής σχέσης. Συγκεκριμένα :

- Το α είναι μία σταθερά, η οποία αντιστοιχεί στην τιμή του Y για x= 0 και
- το β είναι ο όρος που καθορίζει την κλίση της υποτιθέμενης ευθείας, είναι δηλαδή μία σταθερά που δείχνει τη μεταβολή της τιμής Y όταν το x μεταβάλλεται κατά μία μονάδα

Στο παράδειγμα της εταιρίας που πουλάει το προϊόν, η γραμμική σχέση ανάμεσα σε διαφημιστική δαπάνη και πωλήσεις εκφράζεται ως:

Πωλήσεις = α + β · Διαφημιστική Δαπάνη

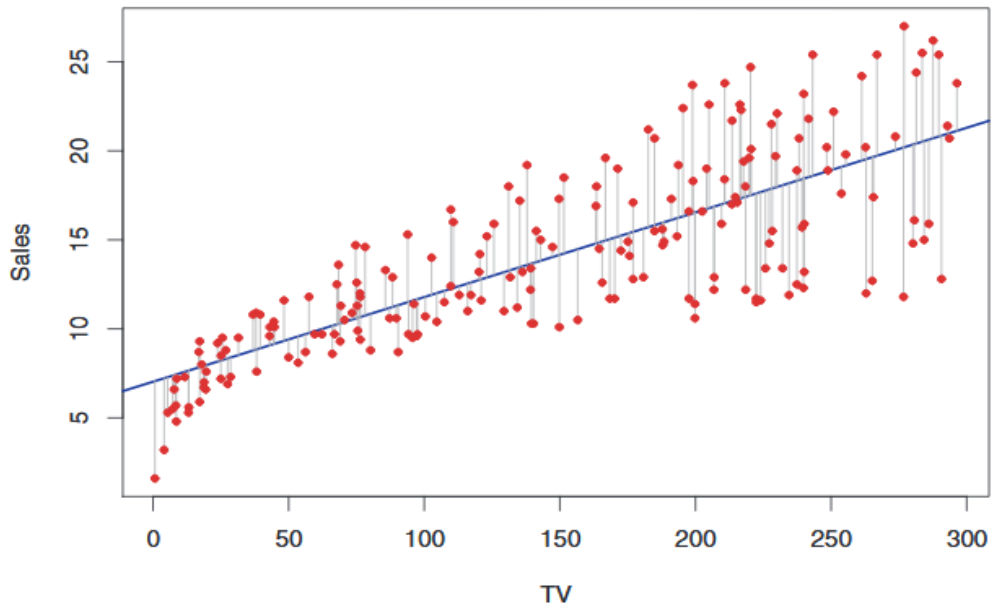
Εφόσον έχουμε στη διάθεση μας ένα πλήθος τιμών x και τις αντίστοιχες τιμές για το Y, χρησιμοποιούμε αυτά τα ζεύγη τιμών X-Y ώστε να εξάγουμε ένα γραμμικό μοντέλο ή για την ακρίβεια να εξάγουμε μία τιμή για τη σταθερά α και μία τιμή για τη σταθερά β οι οποίες θα ορίζουν τη γραμμική σχέση μεταξύ X και Y. Αφού υπολογίσουμε το α και το β, μπορούμε να κάνουμε πρόβλεψη της τιμής Y για οποιαδήποτε δεδομένη τιμή του x. Δηλαδή εάν μας δώσουν μία τιμή για το X, τότε γνωρίζουμε όλους τους όρους του 2ου μέλους της Σχέσης (2.1), επομένως μπορούμε να υπολογίσουμε και το 1ο μέλος, δηλαδή το Y.

Έτσι στο παράδειγμα που έχουμε αναφέρει, εφόσον η εταιρία έχει διατηρήσει ιστορικό μιας σειράς τιμών τηλεοπτικής διαφημιστικής δαπάνης για ένα προϊόν και των αντίστοιχων πωλήσεων για το προϊόν αυτό, ο αναλυτής έχει τη δυνατότητα να εξάγει ένα γραμμικό μοντέλο, δηλαδή τιμή για α και β και στη συνέχεια για διάφορες μελλοντικές τιμές της διαφημιστικής δαπάνης να κάνει πρόβλεψη για τα επίπεδα πωλήσεων του εν λόγω προϊόντος.

2.2.1 Υπολογισμός των συντελεστών α και β

Όπως εξηγήσαμε προηγουμένως, στα προβλήματα που συναντάμε στην πράξη, όπως στο παράδειγμα της εταιρίας που αναφέραμε, τα α και β είναι άγνωστα και σκοπός μας είναι να τα προσδιορίσουμε με βάση το ιστορικό των δεδομένων που έχουμε στην κατοχή μας. Έστω ότι έχουμε στη διάθεση μας ένα πλήθος από n ζεύγη τιμών για δαπάνη για τηλεοπτική διαφήμιση και επίπεδο πωλήσεων, δηλαδή ένα πλήθος ζευγών (x_1,Y_1) , (x_2,Y_2) , (x_3,Y_3) , ….. , (x_n,Y_n).

Κάθε ζευγάρι από αυτά αποτελεί μία απεικόνιση στον άξονα X/Y και μπορεί να αντιπροσωπεύει για παράδειγμα ένα ποσό για διαφημιστική δαπάνη και αντίστοιχο επίπεδο πωλήσεων στην ίδια αγορά για μία σειρά ετών ή ένα ποσό για διαφημιστική δαπάνη και αντίστοιχο επίπεδο πωλήσεων στο ίδιο έτος σε διάφορες χώρες. Αυτό που έχει σημασία είναι ότι η απεικόνιση των δεδομένων που διαθέτουμε είναι δισδιάστατη, εφόσον μιλάμε για απλή γραμμική παλινδρόμηση και φαίνεται στο επόμενο διάγραμμα:



Εικόνα 2 *:Απεικόνιση απλής γραμμικής παλινδρόμησης (Πωλήσεις – Δαπάνες για τηλεοπτική Διαφήμιση)*

Όπως αναφέραμε, ο σκοπός μας είναι να υπολογίσουμε μία βέλτιστη τιμή για το α και β από το ιστορικό των ζευγών (x_n,Y_n) που διαθέτουμε. Στην Εικόνα 2.1 απεικονίζεται με κόκκινο χρώμα κάθε ζεύγος (x_n,Y_n) που έχουμε στη διάθεση μας, ενώ η μωβ ευθεία είναι η ευθεία που ορίζεται από την τιμή των α και β που θα υπολογίσουμε βάσει όλων αυτών των ζευγών. Η ευθεία αυτή (ή αλλιώς το α και το β) θα εξαχθούν με τη βοήθεια της μεθόδου των ελαχίστων τετραγώνων, η οποία αναλύεται στη συνέχεια.

Όπως βλέπουμε στην Εικόνα 2.1, κάθε κόκκινο σημείο έχει μία κάθετη προβολή στη μωβ ευθεία. Ένα οποιοδήποτε κόκκινο σημείο είναι η απεικόνιση ενός πραγματικού ζεύγους (x_i, Y_i) ενώ η προβολή του στη μωβ ευθεία αντιπροσωπεύει το ζεύγος του ίδιου x_i, αλλά με το αντίστοιχο Y_i το οποίο θα υπολογίζεται από το γραμμικό μας μοντέλο και θα πρέπει να ανήκει στην ευθεία. Κάθε γκρι ευθύγραμμο τμήμα που ενώνει τα κόκκινα σημεία με τη μωβ ευθεία, αντιπροσωπεύει τη διαφορά (σφάλμα) ανάμεσα στην πραγματική τιμή του Y_i και την αντίστοιχη εκτιμώμενη τιμή του (προβολή στη μωβ ευθεία), για μία συγκεκριμένη τιμή του x_i. Έστω ότι κάθε τέτοιο σφάλμα έχει μία τιμή e_i. Έστω επίσης ότι ορίζουμε ένα μέγεθος το οποίο ισούται με το άθροισμα των τετραγώνων όλων αυτών των σφαλμάτων και το ονομάζουμε RSS (Residual Sum of Squares), δηλαδή:

$$\mathbf{RSS = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2} \qquad \mathbf{(2.2)}$$

Η βέλτιστη τιμή των α και β που θα πρέπει να υπολογίσουμε, θα πρέπει να είναι τέτοια ώστε αυτά να ορίζουν μία μωβ ευθεία η οποία θα ελαχιστοποιεί το RSS. Αυτή είναι και η μέθοδος των ελαχίστων τετραγώνων. Θα πρέπει ουσιαστικά να υπολογίσουμε μία μωβ ευθεία η οποία θα διέρχεται όσο το δυνατόν πιο κοντά από όλα τα κόκκινα σημεία ζευγών x_i και αντίστοιχων πραγματικών τιμών Y_i.

Κάθε σφάλμα e_i όπως το ορίσαμε παραπάνω, μπορεί να εκφραστεί μαθηματικά με τη βοήθεια των α και β ως εξής:

$$e_i = Y_i - (\alpha + \beta \cdot x_i) \rightarrow \boxed{\mathbf{e_i = Y_i - \alpha - \beta \cdot x_i}} \qquad \mathbf{(2.3)}$$

Με τη βοήθεια της Σχέσης (2.3) η Σχέση (2.2) γίνεται ως εξής:

$$\boxed{\mathbf{RSS = (Y_1 - \alpha - \beta \cdot x_1)^2 + (Y_1 - \alpha - \beta \cdot x_1)^2 + (Y_2 - \alpha - \beta \cdot x_2)^2 + \ldots + (Y_n - \alpha - \beta \cdot x_n)^2}} \quad \mathbf{(2.4)}$$

Η μέθοδος των ελαχίστων τετραγώνων όπως εξηγήσαμε υπολογίζει μία τιμή για το α και μία τιμή για το β τέτοιες ώστε το RSS να ελαχιστοποιείται. Έπειτα από υπολογισμούς, προκύπτει ότι οι βέλτιστες τιμές για τα α και ώστε το RSS να ελαχιστοποιηθεί δίνονται από τους παρακάτω τύπους:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \mathbf{(2.5)}$$

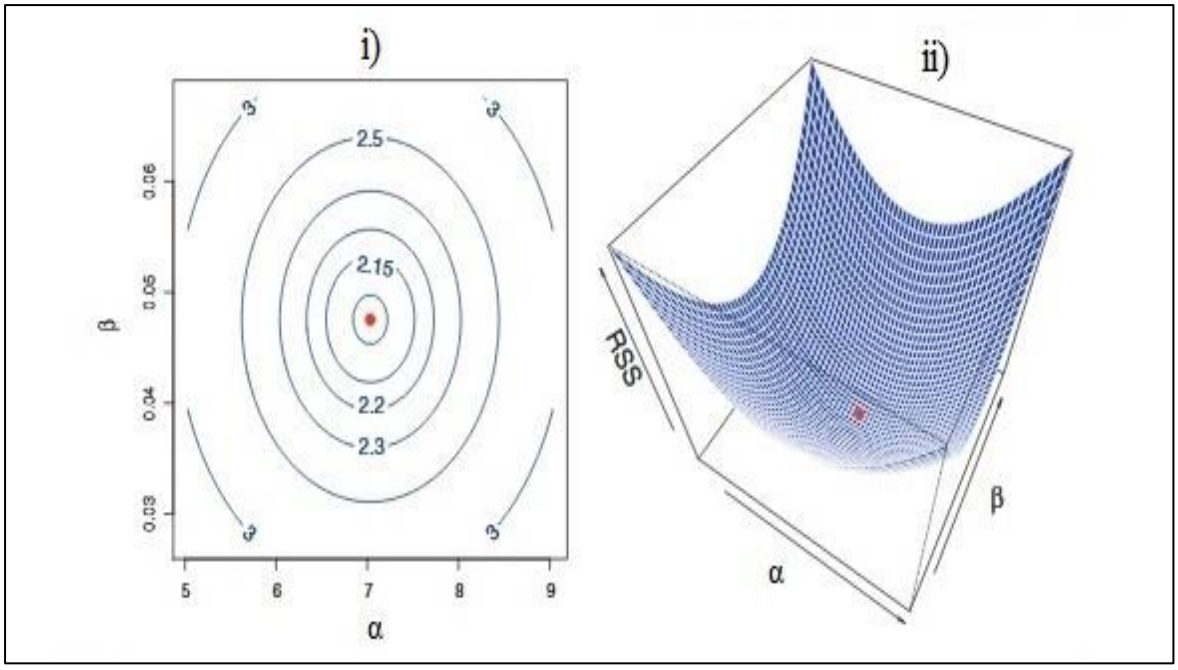
και

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

όπου $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ **και** $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$

Επομένως η Σχέση (2.5) ορίζει τους συντελεστές ελαχίστων τετραγώνων α και β μέσω των οποίων κατασκευάζεται η ευθεία των ελαχίστων τετραγώνων.

Στο επόμενο Σχήμα φαίνεται μία δισδιάστατη (i) και μία τρισδιάστατη (ii) απεικόνιση της μεταβολής του RSS, ανάλογα με το πώς μεταβάλλονται ταυτόχρονα τα α και β. Το κόκκινο σημείο στα 2 διαγράμματα αντιστοιχεί στο ζεύγος α,β και το οποίο επιτυγχάνεται η ελάχιστη τιμή του RSS. Φαίνεται πως εάν η τιμή για το α ή/και το β αυξηθούν ή μειωθούν αρκετά, πέραν του κόκκινου σημείου, τότε η τιμή του RSS αυξάνεται.



***Εικόνα 3**: **i)** Δισδιάστατη και **ii)** τρισδιάστατη απεικόνιση της μεταβολής του RSS ανάλογα με τη μεταβολή των α και β*

Από τη δισδιάστατη φαίνεται πως η βέλτιστη τιμή του RSS προκύπτει από μία τιμή του α κοντά στο 7 και μία τιμή και μία τιμή του β κοντά στο 0,0475. Με βάση το πώς ορίσαμε παραπάνω το τι αντιπροσωπεύουν ποιοτικά τα α και β, συμπεραίνουμε πως για μία αύξηση 1000 € στη διαφημιστική δαπάνη στην τηλεόραση, η εταιρία αναμένει αύξηση των πωλήσεων του προϊόντος κατά περίπου 47,5 τεμάχια. Αντίστοιχα, ακόμα και αν δε δαπανούσε τίποτα για τηλεοπτική διαφήμιση (x=0) θα ανέμενε την πώληση 7 τεμαχίων.

2.2.2 Αξιολόγηση των παραμέτρων α και β

Βρισκόμαστε στο σημείο όπου έχουμε υπολογίσει μία τιμή για τον συντελεστή α και μία τιμή για τον συντελεστή β, τέτοιες ώστε η τιμή του RSS να ελαχιστοποιείται. Επομένως με βάση την ανάλυση της προηγούμενης ενότητας έχουμε υποθέσει ότι υπάρχει μία ευθεία που ορίζεται από το α και το β που υπολογίσαμε, η οποία διέρχεται όσο το δυνατόν πλησιέστερα από όλα τα ζεύγη πραγματικών τιμών x_i, Y_i που έχουμε στη διάθεση μας. Έχουμε δηλαδή υποθέσει εξαρχής ότι οι πραγματικές τιμές του Y που διαθέτουμε, σχετίζονται γραμμικά με τις αντίστοιχες τιμές του x, μέσω της Σχέσης (2.1), δηλαδή: $Y = \alpha + \beta \cdot x$
Ωστόσο, καθώς η ευθεία ποτέ δεν μπορεί να διέρχεται ακριβώς από όλα τα σημεία των πραγματικών ζευγών x_i, Y_i, θεωρούμε πως στο γραμμικό μοντέλο που εξάγαμε υπάρχει πάντα ένα σφάλμα. Το σφάλμα αυτό θα το εισάγουμε στη γραμμική μας σχέση με τον όρο ε. Επομένως η σχέση (2.1) μεταβάλλεται ως εξής:

$$\boxed{\mathbf{Y = \alpha + \beta \cdot x + \varepsilon}} \qquad \textbf{(2.6)}$$

Καθώς το γραμμικό μοντέλο είναι στην ουσία πολύ απλουστευμένο, στην πράξη οι τιμές των x και Y πιθανότατα να μη συνδέονται εντελώς γραμμικά. Μπορεί δηλαδή αύξηση ή μείωση του X να συνεπάγεται αντίστοιχα αύξηση ή μείωση του Y, χωρίς όμως οι μεταβολές να γίνονται με γραμμικό τρόπο. Επομένως είναι βέβαιο ότι θα υπάρχει διαφορά ανάμεσα στην ευθεία που εξάγαμε και στις πραγματικές τιμές των ζευγών x_i, Y_i, γι αυτό και εισάγουμε στη γραμμική σχέση και τον όρο ε, ο οποίος θεωρούμε ότι είναι ανεξάρτητος του X.

Στο σημείο αυτό κρίνεται σκόπιμο να ορίσουμε κάποια υπολογιστικά σφάλματα για την τιμή του α και του β. Για κάθε μία από αυτές τις 2 τιμές που εκτιμήσαμε, με τη βοήθεια του γραμμικού μοντέλου το οποίο εξήχθη από τα δεδομένα μας, ορίζεται ένα σφάλμα Standard Error (SE) με βάση τους παρακάτω τύπους:

$$\mathrm{SE}(\alpha)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\beta)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \textbf{(2.7)}$$

Από τη σχέση (2.7) η μοναδική τιμή που δε γνωρίζουμε είναι το σ^2, το οποίο το ορίζουμε ως **Residual Standard Error** και μπορούμε να το υπολογίσουμε από τον τύπο:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad \textbf{(2.8)}$$

Το RSS έχει προσδιοριστεί στη Σχέση (2.2)

Με τη βοήθεια των Standard Errors μπορούμε να ορίσουμε τα αντίστοιχα διαστήματα εμπιστοσύνης για τις τιμές των α και β και συγκεκριμένα τα 95% διαστήματα εμπιστοσύνης για την κάθε σταθερά. Ένα διάστημα εμπιστοσύνης 95% ορίζεται ως ένα διάστημα τιμών με άνω και κάτω όριο, μέσα στο οποίο υπάρχει 95% πιθανότητα να βρίσκεται πραγματική τιμή της σταθεράς του γραμμικού μας μοντέλου.

Για τη σταθερά α, έστω ότι με βάση τα δεδομένα που έχουμε στη διάθεση μας υπολογίσαμε από τη Σχέση (2.5) μία τιμή α'. Το 95% διάστημα εμπιστοσύνης δίνεται από τον τύπο $\alpha' \pm 2 \cdot SE(\alpha')$, δηλαδή ορίζεται ως:

$$[\ \alpha' - 2 \cdot SE(\alpha')\ ,\ \alpha' + 2 \cdot SE(\alpha')\ ] \qquad \textbf{(2.9)}$$

Κατ' αντιστοιχία, το 95% διάστημα εμπιστοσύνης για τη σταθερά β ορίζεται ως:

$$[\ \beta' - 2 \cdot SE(\beta')\ ,\ \beta' + 2 \cdot SE(\beta')\ ] \qquad \textbf{(2.10)}$$

Έτσι είμαστε βέβαιοι κατά 95% ότι οι πραγματικές τιμές των α και β, θα βρίσκονται στα παραπάνω διαστήματα τα οποία ορίζονται από τις εκτιμώμενες τιμές α' και β' αντίστοιχα.

Στο παράδειγμα που χρησιμοποιήσαμε παραπάνω, για την εταιρία που θέλει να διερευνήσει τη σχέση ανάμεσα στη δαπάνη για τηλεοπτική διαφήμιση ενός προϊόντος και στη μεταβολή των πωλήσεων του προϊόντος αυτού, οι βέλτιστες εκτιμώμενες τιμές που προέκυψαν για το α και το β από τους τύπους της (2.5) ήταν περίπου 7 και 0,0475 αντίστοιχα. Από τη Βιβλιογραφία από την οποία αντλήσαμε το παράδειγμα βλέπουμε πως το 95% διάστημα εμπιστοσύνης για την πραγματική τιμή του α είναι [6,130 ,7,935] ενώ το αντίστοιχο διάστημα εμπιστοσύνης για την πραγματική τιμή του β είναι [0,042 , 0,053]. Έτσι η εταιρία μπορεί να αναμένει με 95% βεβαιότητα πως μία αύξηση της τάξεως των 1000 € στη δαπάνη για τηλεοπτική διαφήμιση θα έχει ως αποτέλεσμα την αύξηση των πωλήσεων του προϊόντος σε ένα εύρος τιμών που θα κυμαίνεται από 42 τεμάχια (στην χειρότερη περίπτωση) έως 53 τεμάχια (στην καλύτερη περίπτωση).

Ο υπολογισμός των Standard Errors μας επιτρέπει εκτός από την εκτίμηση των 95% διαστημάτων εμπιστοσύνης για τις πραγματικές τιμές των α και β, να εξετάσουμε και το κατά πόσο οι τιμές του x σχετίζονται γραμμικά με τις τιμές του Y μέσω του ελέγχου μηδενικής υπόθεσης. Ο έλεγχος μηδενικής υπόθεσης είναι μία συνήθης μέθοδος που χρησιμοποιείται για έλεγχο γραμμικές σχέσης βασίζεται σε απλές υποθέσεις:

- $\mathbf{H_0}$: Δεν υπάρχει καμία σχέση μεταξύ x και Y

ή εναλλακτικά

- $\mathbf{H_\alpha}$: Υπάρχει κάποια σχέση μεταξύ x και Y

Εάν υποτεθεί ότι ισχύει η συνθήκη H_0, τότε θεωρούμε ότι β=0 και επομένως η Σχέση (2.6) αλλάζει ως εξής: Y = α + ε και κατά συνέπεια η τιμή του Y είναι εντελώς ανεξάρτητη από την τιμή του x.

Εάν από την άλλη υποτεθεί ότι ισχύει η συνθήκη H_α, ότι δηλαδή υπάρχει κάποια γραμμική σχέση μεταξύ x και Y, τότε η τιμή του β θα είναι διάφορη του μηδενός, επομένως η Σχέση (2.6) θα ισχύει ως έχει. Για να μπορέσουμε να απαντήσουμε εάν αυτή η υπόθεση είναι σωστή, θα πρέπει να δούμε εάν η τιμή του β', δηλαδή η εκτιμώμενη βέλτιστη τιμή για το β που υπολογίσαμε, είναι αρκετά μακριά από το 0, ώστε να είμαστε σίγουροι ότι το β δεν έχει μηδενική τιμή. Το πόσο μακριά πρέπει να βρίσκεται η τιμή του β' από το μηδέν είναι κάτι σχετικό και εξαρτάται από την τιμή του SE(β'). Εάν η τιμή του SE(β') είναι μικρή, τότε ακόμα και μία σχετικά μικρή τιμή για το β' θα αρκούσε στο να φανεί ότι β ≠0 και επομένως υπάρχει κάποια γραμμική σχέση μεταξύ x και Y. Εάν η τιμή του SE(β') είναι μεγάλη, αυτό σημαίνει πως χρειαζόμαστε μεγάλη τιμή για το β', ώστε να φαίνεται ότι αυτό απέχει πολύ από το μηδέν και επομένως είναι διάφορο του μηδενός και έτσι θα ισχύει η υπόθεση ότι υπάρχει γραμμικής σχέση μεταξύ x και Y. Σε κάθε περίπτωση μία μεγάλη τιμή του για το β' είναι επιθυμητή προκειμένου να αποδειχθεί η ύπαρξη γραμμικότητας.

Για την ευκολότερη απόδειξη των παραπάνω, χρησιμοποιούμε τον όρο **t-statistic** ο οποίος εκφράζεται με τον τύπο:

$$t = \frac{\beta' - 0}{SE(\beta')} \qquad \textbf{(2.11)}$$

Ο όρος αυτός εκφράζει το κατά πόσο η τιμή β' απέχει από το 0. Ξεκινώντας με τον έλεγχο μηδενικής υπόθεσης και θεωρώντας ότι β' = 0, είναι απλό να υπολογίσουμε την πιθανότητα του να εντοπίσουμε κάποια τιμή ίση ή μεγαλύτερη του |t|. Η πιθανότητα αυτή ορίζεται ως **p-value** και ερμηνεύεται ως εξής: μία μικρή τιμή της p-value υποδεικνύει πως είναι απίθανο να παρατηρήσουμε κάποια ουσιώδη σχέση μεταξύ κάποιων μεμονωμένων ανεξάρτητων (x) και εξαρτημένων μεταβλητών (Y) λόγω τύχης και όχι στο σύνολό τους. Είναι δηλαδή απίθανο τα x και Y να μη σχετίζονται γραμμικά μεταξύ τους στο σύνολό τους, αλλά να βρούμε κάποια ελάχιστα ζεύγη που σχετίζονται κατά τύχη. Κατά συνέπεια, μία μικρή τιμή της πιθανότητας p-value μας βοηθάει να συμπεράνουμε πως υπάρχει γραμμική σχέση μεταξύ των x και των αντίστοιχων Y.

Επομένως σε περίπτωση που έχουμε μικρή τιμή για την p-value (κοντά στο 0), μπορούμε να απορρίψουμε τη μηδενική υπόθεση H_0, η οποία ξεκινάει με την παραδοχή ότι δεν υπάρχει καμία σχέση ανάμεσα στην ανεξάρτητες και τις εξαρτημένες τιμές. Ένα τυπικό άνω όριο για την τιμή της p-value ώστε να απορρίψουμε με ασφάλεια τη μηδενική υπόθεση, είναι το 1%, δηλαδή εάν η τιμή της

p-value είναι μικρότερη από 0,001, τότε μπορούμε να οδηγηθούμε στο συμπέρασμα πώς δεν ισχύει η μηδενική υπόθεση.

Επανερχόμαστε πάλι στο παράδειγμα με την εταιρία και τα έξοδα διαφήμισης, το οποίο βοηθάει στο να γίνουν κατανοητοί οι ορισμοί και οι συντελεστές που παρουσιάζουμε. Στον παρακάτω πίνακα φαίνονται βέλτιστες οι τιμές για το α′ και β′ που έχουν ήδη υπολογιστεί, καθώς και οι αντίστοιχες τιμές για Standard Error, t-statistic και p-value που έχουν αντληθεί έτοιμα από τη Βιβλιογραφία, καθώς στην παρούσα φάση μας ενδιαφέρει να γίνονται κατανοητοί οι συντελεστές και όχι να υπολογίζονται.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Πίνακας 1*: Πίνακας με τις τιμές για το α (Intercept) και β (TV) καθώς και τους αντίστοιχους όρους που είναι απαραίτητοι για την αξιολόγηση των τιμών αυτών*

Από τον παραπάνω Πίνακα διαπιστώνεται πως ενώ η τιμή του Standard Error β′ είναι αρκετά μικρή (0,0027), ο όρος t-statistic είναι αρκετά μακριά από το 0 σε σχέση με τη μικρή τιμή του STE(β′), ενώ θα μας αρκούσε ακόμα και μια πιο μικρή τιμή για το t-statistic ώστε να απορρίψουμε τη μηδενική υπόθεση. Επομένως με βάση την ανάλυση που έγινε παραπάνω και σε συνδυασμό με την πολύ μικρή τιμή του p-value (<0,0001) οδηγούμαστε με βεβαιότητα στο συμπέρασμα ότι η ανεξάρτητη τιμή x (TV), δηλαδή η δαπάνη για τηλεοπτική διαφήμιση συνδέεται με μια ισχυρή γραμμική σχέση με την εξαρτημένη μεταβλητή Y, δηλαδή την πώληση του διαφημιζόμενου προϊόντος.

2.2.3 Αξιολόγηση της ακρίβειας του μοντέλου μας

Σε μία οποιαδήποτε μελέτη ελέγχου γραμμικότητας, όπως και στο παράδειγμα της εταιρίας που πουλάει ένα προϊόν και θέλει να εξετάζει κατά πόσο η πώληση του σχετίζεται με τη διαφημιστική δαπάνη στην τηλεόραση, το οποίο και έχουμε αναλύσει, δεν αρκεί μόνο να αποδειχθεί ότι απορρίπτεται η μηδενική υπόθεση. Στο παράδειγμα μας, βρισκόμαστε στο σημείο όπου έχουμε καταλήξει στο συμπέρασμα πως τα 2 παραπάνω μεγέθη x-Y σχετίζονται γραμμικά μεταξύ τους και μάλιστα έχουμε κατασκευάσει ένα θεωρητικό γραμμικό μοντέλο (δηλαδή τις τιμές α′ και β′) που προσπαθεί να εκφράζει όσο καλύτερα μπορεί τη γραμμική αυτή σχέση. Το επόμενο στάδιο της ανάλυσης μας είναι να αξιολογήσουμε το μοντέλο μας, δηλαδή να εξετάσουμε το κατά πόσο τα σημεία της θεωρητικής αυτής ευθείας προσεγγίζουν τις αντίστοιχες πραγματικές τιμές των δεδομένων μας.

Η ακρίβεια ενός γραμμικού μοντέλου που έχει προκύψει με τη μέθοδο της γραμμικής παλινδρόμησης, αξιολογείται πρακτικά με τη βοήθεια 2 όρων οι οποίοι σχετίζονται μεταξύ τους:

- **Residual Standard Error (RSE)**

  Παρουσιάστηκε στην ενότητα 2.2.2 και υπολογίζεται από τα σφάλματα ανάμεσα στις πραγματικές τιμές των δεδομένων μας και τις αντίστοιχες τιμές του γραμμικού μοντέλου και εκφράζεται μαθηματικά από τη Σχέση (2.8). Ο όρος αυτός υποδηλώνει πως παρόλο που έχουμε υπολογίσει μία βέλτιστη τιμή για το α και μία αντίστοιχα για το β, πρακτικά οι τιμές της εξαρτημένης μεταβλητής του γραμμικού μοντέλου δε συμπίπτουν με τις αντίστοιχες πραγματικές τιμές, για δεδομένες τιμές της ανεξάρτητης μεταβλητής. Αυτό κατ' επέκταση σημαίνει πως με τη βοήθεια του γραμμικού μοντέλου μπορούμε να προβλέπουμε την τιμή του Y για μια δεδομένη τιμή του x, χωρίς ωστόσο η πρόβλεψη μας να είναι απόλυτα ακριβής. Ο όρος RSE στην ουσία υπολογίζει τον μέσο όρο της απόκλισης των τιμών που βρίσκονται πάνω στην ευθεία του μοντέλου μας, από τις αντίστοιχες πραγματικές τιμές του Y.

  Με άλλα λόγια το RSE υφίσταται λόγω του ότι στην πράξη, όσο ακριβές και να είναι το γραμμικό μοντέλο, οι εκτιμώμενες τιμές της εξαρτημένης μεταβλητής ποτέ δεν συμπίπτουν ακριβώς με τις αντίστοιχες πραγματικές τιμές και κατά συνέπεια η πρόβλεψη για μελλοντικές τιμές της εξαρτημένης μεταβλητής όσο καλή κι αν είναι, στην πράξη δε συμπίπτει ακριβώς με την πραγματική τιμή που θα έρθει στο μέλλον. Προφανώς, ελλείψει μίας τέλειας πρόβλεψης, σκοπός μας είναι το γραμμικό μοντέλο να έχει RSE όσο το δυνατόν πλησιέστερα στο 0, δηλαδή οι τιμές που προβλέπει το μοντέλο να είναι όσο το δυνατόν πιο κοντά με τις αντίστοιχες πραγματικές τιμές.

- **R^2 statistic**

  Ο όρος RSE όπως εξηγήθηκε, είναι μια απόλυτη τιμή που ουσιαστικά εκφράζει την απουσία μιας τέλειας προσαρμογής των τιμών που προβλέπονται από το γραμμικό μοντέλο με τις αντίστοιχες πραγματικές τιμές. Ωστόσο, καθώς είναι όρος που εκφράζεται στη μονάδα μέτρησης του Y, δεν είναι ξεκάθαρο ότι πάντα η τιμή του θα μας βοηθά να κρίνουμε αν το μοντέλο κάνει καλές προβλέψεις ή όχι. Είπαμε ότι σε γενικές γραμμές θέλουμε να έχει όσο το δυνατόν μικρότερη τιμή, όμως αυτό είναι κάποιες φορές σχετικό. Εξαρτάται καθαρά από τη φύση του προβλήματος για το αν μία μεγάλη τιμή του RSE μπορεί να είναι αποδεκτή ή αντίστροφα μία σχετικά μικρή τιμή του να βγαίνει εκτός αποδεκτών ορίων. Για το λόγο είναι προτιμότερο να μελετάμε τον όρο R^2 ο οποίος εκφράζεται σε ποσοστό, δηλαδή λαμβάνει τιμές

από 0 έως 1 και επομένως μας βοηθά να αξιολογήσουμε πιο σωστά το γραμμικό μας μοντέλο, εφόσον δεν εξαρτάται από τη μονάδα μέτρησης της μεταβλητής Y.

Η τιμή του R^2 υπολογίζεται από τον τύπο:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \textbf{(2.12)}$$

όπου

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

Ο όρος R^2 θα μπορούσαμε να πούμε ότι εκφράζει την ποιότητα της διακύμανσης των εκτιμώμενων τιμών του Y, βάσει των τιμών της ανεξάρτητης μεταβλητής x. Μία τιμή του R^2 κοντά στο 1, υποδηλώνει πως ένα μεγάλο ποσοστό της συνολικής διακύμανσης στην εξαρτημένη μεταβλητή, δικαιολογείται από τις δεδομένες τιμές του x. Αντιθέτως, μία τιμή του R^2 κοντά στο 0 ότι η παλινδρόμηση δεν μπορεί να δικαιολογήσει τη διακύμανση ανάμεσα στις διάφορες τιμές του Y, βάσει των τιμών του x, είτε διότι το γραμμικό μοντέλο που έχουμε εξάγει δεν είναι σωστό είτε γιατί το RSE και κατ' επέκταση το RSS έχουν μεγάλη τιμή.

<u>Σημείωση:</u> Γενικά, όταν θέλουμε να διερευνήσουμε την ύπαρξη γραμμικής σχέσης ανάμεσα σε μία εξαρτημένη μεταβλητή Y και μία ανεξάρτητη μεταβλητή X, χρησιμοποιούμε την έννοια της συσχέτισης (correlation) η οποία δίνεται από τον παρακάτω τύπο:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \textbf{(2.13)}$$

Όμως αποδεικνύεται ότι <u>στην περίπτωση της γραμμικής παλινδρόμησης και μόνο</u>, ο όρος $\text{Cor}^2(X,Y)$ ισούται με τον όρο R^2. Επομένως όποιον από τους δύο όρους και να υπολογίσουμε για την αξιολόγηση του μοντέλου που προέκυψε από απλή παλινδρόμηση, έχουμε το ίδιο αποτέλεσμα. Ωστόσο στην περίπτωση της πολλαπλής παλινδρόμησης που θα αναλύσουμε στην επόμενη ενότητα, όπου εξετάζεται η σχέση που συνδέει μία εξαρτημένη μεταβλητή με περισσότερες από μία ανεξάρτητες μεταβλητές, προφανώς δεν μπορεί να γίνει χρήση της Σχέσης (2.13) η οποία περιλαμβάνει μόνο μία ανεξάρτητη μεταβλητή. Για το λόγο αυτό θα πρέπει να υπολογίσουμε αναγκαστικά μόνο την τιμή του R^2.

2.3 Πολλαπλή Γραμμική Παλινδρόμηση

Όπως αναλύθηκε στην προηγούμενη ενότητα, η απλή Γραμμική Παλινδρόμηση αποτελεί μία πολύ χρήσιμη μέθοδο πρόβλεψης για την τιμή μίας εξαρτημένης μεταβλητής, η οποία όμως σχετίζεται με μία μόνο ανεξάρτητη μεταβλητή. Στην πράξη όμως, όταν θέλουμε να διερευνήσουμε τη μεταβολή μίας εξαρτημένης μεταβλητής, καλούμαστε να λάβουμε υπόψη μας περισσότερες από μία ανεξάρτητες μεταβλητές. Στο παράδειγμα με τη μεταβολή της δαπάνης για τηλεοπτική διαφήμιση ενός προϊόντος και τον αντίκτυπο που θα έχει η μεταβολή αυτή στις πωλήσεις του προϊόντος, μπορεί να έχουμε στη διάθεση μας και άλλα δεδομένα, όπως π.χ. η δαπάνη για διαφήμιση στο ραδιόφωνο και η δαπάνη για διαφήμιση στον έντυπο τύπο, επομένως θα κληθούμε στη διερεύνηση που θα κάνουμε να λάβουμε υπόψη και αυτά τα δεδομένα και να δούμε αν οι εν λόγω δαπάνες σχετίζονται και αυτές γραμμικά με τις πωλήσεις του διαφημιζόμενου προϊόντος.

Έτσι λοιπόν είναι κατανοητό πως στην πράξη μία διερεύνηση για ύπαρξη γραμμικότητας δε θα περιέχει μία μόνο ανεξάρτητη μεταβλητή, αλλά ένα διάνυσμα ανεξάρτητων μεταβλητών. Μία πρώτη προσέγγιση θα ήταν στο να διεξάγουμε τρεις ξεχωριστές απλές γραμμικές παλινδρομήσεις για καθεμία ανεξάρτητη μεταβλητή χωριστά, δηλαδή για κάθε διαφημιστική δαπάνη σε κάθε μέσο ενημέρωσης. Ωστόσο μία τέτοια προσέγγιση δεν κρίνεται ικανοποιητική και ενδέχεται να μας οδηγήσει και σε λανθασμένες εκτιμήσεις. Πρώτα απ' όλα εάν διεξάγουμε τρεις διαφορετικές απλές γραμμικές παλινδρομήσεις, δεν είναι καθόλου ξεκάθαρο το πώς θα μπορέσουμε να κάνουμε μία συνολική πρόβλεψη για τις πωλήσεις του προϊόντος, που αυτό είναι και το ζητούμενο, καθώς η δαπάνη σε κάθε μέσο ενημέρωσης θα σχετίζεται με τις πωλήσεις με μία διαφορετική σχέση παλινδρόμησης. Επίσης, κάθε μοντέλο που θα προκύπτει από την ανάλυση γραμμικής παλινδρόμησης, εφόσον έχει λάβει υπόψη του τη διαφημιστική δαπάνη για ένα μόνο μέσο ενημέρωσης, θα αγνοεί τη συμμετοχή και τη βαρύτητα που έχουν οι δαπάνες στα άλλα δύο μέσα στον υπολογισμό των συντελεστών της ευθείας.

Για τους παραπάνω λόγους είναι ξεκάθαρο πως δε συστήνεται η ανάλυση τριών ξεχωριστών απλών γραμμικών παλινδρομήσεων, αλλά η διεξαγωγή μίας πολλαπλής γραμμικής παλινδρόμησης, δηλαδή μίας μεθόδου που θεωρείται επέκταση της απλής γραμμικής παλινδρόμησης και μας δίνει τη δυνατότητα να διερευνήσουμε την ύπαρξη γραμμικότητας λαμβάνοντας υπόψη παράλληλα πολλές ανεξάρτητες μεταβλητές. Η σχέση που εκφράζει μαθηματικά τη μεταβολή της εξαρτημένης μεταβλητής ανάλογα με τη μεταβολή των εξαρτημένων μεταβλητών παράλληλα, προκύπτει ως επέκταση της Σχέσης (2.6) και είναι η εξής:

$$\boxed{Y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n + \varepsilon} \qquad \textbf{(2.14)}$$

Όπως εξηγήσαμε και στην απλή γραμμική παλινδρόμηση, το α είναι μία σταθερά που αντιστοιχεί στην τιμή του Y όταν $x_1 = x_2 = \ldots = x_n = 0$ και το β_i είναι η σταθερά που εκφράζει ποσοτικά τη σχέση ανάμεσα στην εξαρτημένη μεταβλητή και στην αντίστοιχη εξαρτημένη μεταβλητή x_i (π.χ. τη σχέση ανάμεσα στις πωλήσεις και τη διαφημιστική δαπάνη στο ραδιόφωνο).

Εάν θέλουμε στη Σχέση (2.14) να αντικαταστήσουμε τις παραμέτρους του παραδείγματος που εξετάζουμε, τότε θα έχουμε:

Πωλήσεις = α + β_1·Δαπάνη TV + β_2·Δαπάνη Ραδιοφώνου +…+ β_n· Δαπάνη Εφημερίδες + ε

2.3.1 Υπολογισμός των συντελεστών α, β_1,β_2,…β_n

Όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης, σκοπός μας είναι να υπολογίσουμε τις βέλτιστες τιμές για τους συντελεστές της γραμμικής εξίσωσης. Το γραμμικό μοντέλο μας θα περιγράφεται από τη σχέση $\mathbf{Y = \alpha' + \beta_1'\cdot x_1 + \beta_2'\cdot x_2 + \ldots + \beta_n'\cdot x_n + \varepsilon}$, όπου τα α΄, β_1', β_2',... , β_n' θα είναι οι βέλτιστες τιμές των συντελεστών που θα υπολογιστούν με τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης. Με τη βοήθεια του γραμμικού μας μοντέλου θα κάνουμε προβλέψεις στις τιμές του Y για δεδομένες τιμές των $x_1,x_2,\ldots,x_n$.

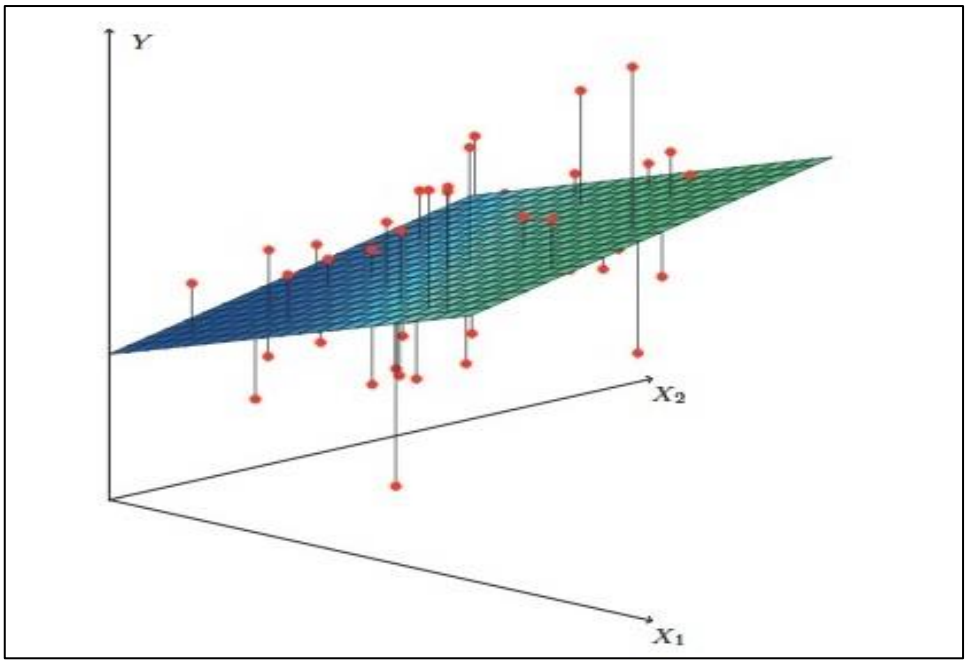
Όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης, οι βέλτιστες τιμές για τη γραμμική εξίσωση θα υπολογιστούν με τη βοήθεια των ελαχίστων τετραγώνων των σφαλμάτων μοντελοποίησης, όμως σε όχι σε μορφή ευθείας, εφόσον τώρα έχουμε περισσότερες από μία εξαρτημένες μεταβλητές, αλλά σε πιο σύνθετη μορφή. Έτσι η μέθοδος των ελαχίστων τετραγώνων στην πολλαπλή παλινδρόμηση θα εξάγει τις τιμές για του συντελεστές α, β_1, β_2,... , β_n για τις οποίες θα ελαχιστοποιείται το άθροισμα τετραγώνων των σφαλμάτων **RSS** (Residual Sum of Squares) που περιγράφεται από τον τύπο:

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

(2.15)

Έτσι, οι τιμές α', β_1', β_2',... , β_n' για τις οποίες ελαχιστοποιείται αυτό το άθροισμα, θα αποτελούν τους συντελεστές της πολλαπλής παλινδρόμησης ελαχίστων τετραγώνων, ή αλλιώς τους συντελεστές της Σχέσης (2.14). Οι συντελεστές αυτοί (σε αντίθεση με τους συντελεστές της απλής γραμμικής παλινδρόμησης που υπολογίζονται με μία διαίρεση αθροισμάτων) είναι πιο πολύπλοκο να υπολογιστούν μαθηματικά και απαιτούν τύπους γραμμικής άλγεβρας στους οποίους δε θα σταθούμε, καθώς δεν είναι αυτός ο σκοπός μας. Σκοπός μας είναι να μπορούμε να υπολογίζουμε γρήγορα και να αξιολογούμε τέτοιους συντελεστές πολλαπλής γραμμικής παλινδρόμησης με τη βοήθεια της R, μία διαδικασία που θα παρουσιαστεί αναλυτικά στο Κεφάλαιο 5, στο οποίο και θα εξάγουμε ένα γραμμικό μοντέλο πρόβλεψης από πραγματικά δεδομένα.

Έχουμε αναφέρει ότι η πολλαπλή γραμμική παλινδρόμηση είναι στην ουσία μια μέθοδος που θεωρείται επέκταση της απλής γραμμικής παλινδρόμησης. Επομένως εξετάζοντας την ελαχιστοποίηση του RSS, δεν προκύπτει μία ευθεία που θα διέρχεται με τον βέλτιστο δυνατό τρόπο από τα πραγματικά σημεία, καθώς πλέον μιλάμε για 3 ή περισσότερες διαστάσεις. Η μοναδική περίπτωση πολλαπλής παλινδρόμησης την οποία μπορούμε να αναπαραστήσουμε γραφικά είναι η πιο απλή μορφή της, δηλαδή η περίπτωση που έχουμε την εξαρτημένη μεταβλητή και 2 ανεξάρτητες μεταβλητές. Μία τέτοια αναπαράσταση φαίνεται στην ακόλουθη εικόνα, όπου δεν υπάρχει ευθεία, αλλά επίπεδο, το οποίο διέρχεται με τον βέλτιστο δυνατό τρόπο από τα πραγματικά σημεία:



***Εικόνα 4**: Απεικόνιση πολλαπλής γραμμικής παλινδρόμηση με 3 διαστάσεις (μία διάσταση για την εξαρτημένη μεταβλητή και 2 για τις ανεξάρτητες)*

2.3.2 Σημαντικά Ερωτήματα για την Ανάλυση

Συνήθως όταν εφαρμόζουμε τη μέθοδο πολλαπλής γραμμικής παλινδρόμησης πρέπει να απαντήσουμε με τη σειρά στα εξής σημαντικά ερωτήματα:

i) Έχει τουλάχιστον μία από τις ανεξάρτητες μεταβλητές $x_1,x_2,\ldots x_n$ σημαντικό ρόλο στην πρόβλεψη της εξαρτημένης μεταβλητής Y;

ii) Είναι σημαντικές όλες οι ανεξάρτητες μεταβλητές $x_1,x_2,\ldots x_n$ στη διαμόρφωση της τιμής της εξαρτημένης μεταβλητής Y ή μόνο κάποιες από αυτές;

iii) Πόσο καλά το γραμμικό μας μοντέλο προσεγγίζει τα πραγματικά δεδομένα;

Οι απαντήσεις στα παραπάνω ερωτήματα δίνονται με τη σειρά:

i) Έχει τουλάχιστον μία από τις ανεξάρτητες μεταβλητές $x_1,x_2,\ldots x_m$ σημαντικό ρόλο στην πρόβλεψη της εξαρτημένης μεταβλητής Y;

Έστω ότι έχουμε m ανεξάρτητες μεταβλητές και n διανύσματα πραγματικών δεδομένων ($x_1,x_2,..,x_m$,Y) τα οποία χρησιμοποιήσαμε για να εξάγουμε το γραμμικό μας μοντέλο**.** Για να απαντήσουμε στο συγκεκριμένο ερώτημα, θα πρέπει να ελέγξουμε εάν κάποιοι από τους συντελεστές $\beta_1',\beta_2', ..\beta_m'$ που έχουμε υπολογίσει είναι ίσοι με το μηδέν. Όπως έχουμε εξηγήσει, εάν ένας συντελεστής β_i' ισούται με μηδέν, αυτό θα συνεπάγεται πως το γινόμενο του με την αντίστοιχη ανεξάρτητη μεταβλητή x_i θα ισούται με μηδέν οποιαδήποτε τιμή και να λάβει το x_i, επομένως το Y δε θα επηρεάζεται από τη συγκεκριμένη ανεξάρτητη μεταβλητή.

Για να απαντήσουμε λοιπόν στο συγκεκριμένο ερώτημα θα εξετάσουμε το <u>σενάριο της μηδενικής υπόθεσης **Ho**</u> η οποία θεωρεί ότι $\beta_1=\beta_2=\ldots=\beta_m=0$ και το σενάριο $\mathbf{H_a}$ το οποίο υποθέτει ότι τουλάχιστον μία από τις τιμές των $\beta_1,\beta_2,\ldots,\beta_n$ είναι διάφορη του μηδενός. Το ποιο από τα 2 σενάρια ισχύει, θα απαντηθεί υπολογίζοντας την τιμή του όρου **F-statistic** από τη σχέση:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/\mathrm{m}}{\mathrm{RSS}/(\mathrm{n}-\mathrm{m}-1)} \qquad \textbf{(2.16)}$$

με το RSS να υπολογίζεται από τη Σχέση (2.15) και το TSS, όπως και στην απλή γραμμική παλινδρόμηση από τη σχέση:

$$\mathrm{TSS} = \sum(y_i - \bar{y})^2$$

Όταν δεν υπάρχει συσχέτιση ανάμεσα στην εξαρτημένη μεταβλητή και τις ανεξάρτητες μεταβλητές, ό όρος F λαμβάνει τιμή κοντά στη μονάδα .Σε αντίθετη περίπτωση, ένα δηλαδή ισχύει η H_{α}, η τιμή του F αναμένεται να είναι αρκετά μεγαλύτερη της μονάδας.

Στο παράδειγμα με την εταιρία το οποίο και εξετάζουμε, αντλούμε την πληροφορία ότι από την πολλαπλή γραμμική παλινδρόμηση με ανεξάρτητες μεταβλητές τη διαφημιστική δαπάνη σε 3 διαφορετικά μέσα ενημέρωσης, η τιμή του όρου F-statistic που υπολογίστηκε είναι 570. Αυτή είναι μία τιμή πολύ μεγαλύτερη της μονάδας, η οποία αποτελεί μία ισχυρή απόδειξη <u>ότι η μηδενική υπόθεση απορρίπτεται, συνεπώς μπορούμε να πούμε πως με βεβαιότητα πως η διαφημιστική δαπάνη έστω και σε ένα μέσο ενημέρωσης επηρεάζει τις πωλήσεις</u>.

Με βάση τα παραπάνω, γεννιέται το ερώτημα στο ποια θα είναι η απόφαση μας στην περίπτωση που η τιμή του F-statistic δεν προκύψει τόσο μεγάλη, καθώς και πόση πρέπει να είναι ελάχιστη τομή του ώστε να μπορούμε να αποκλείσουμε με ασφάλεια τη μηδενική υπόθεση Ho. Η απάντηση είναι πώς όλα εξαρτώνται από την τιμή του n: όταν η τιμή του n είναι μεγάλη, αρκεί μία τιμή του F-statistic λίγο μεγαλύτερη της μονάδας ώστε να απορριφθεί η Ho. Αντίστροφα, όταν το n έχει μικρή τιμή, χρειαζόμαστε μια τιμή για το F-statistic αρκετά μεγαλύτερη της μονάδας για να απορρίψουμε την Ho. Εάν επιθυμούμε να κάνουμε πιο ολοκληρωμένο έλεγχος σε αυτή τη φάση, μπορούμε παράλληλα με την τιμή του F-statistic να τσεκάρουμε και την τιμή ενός άλλου ενδεικτικού όρου, της πιθανότητας **p-value** όπως και στην απλή παλινδρόμηση. Για δεδομένες τιμές n και m, προκύπτει μία τιμή p-value για το μοντέλο, που σχετίζεται με τον όρο F-statistic, η οποία εάν είναι κοντά στο 0 μας επιβεβαιώνει πως η Ho απορρίπτεται. Στο κεφάλαιο 4 στους ελέγχουμε που θα κάνουμε, θα βρούμε με τη βοήθεια της R τις τιμές και για το F-statistic και για την p-value και θα αξιολογήσουμε ανάλογα το μοντέλο μας.

<u>Σημείωση:</u> Για την παραπάνω ανάλυση έχουμε θεωρήσει ότι έχουμε έναν λογικό αριθμό ανεξάρτητων μεταβλητών στο πρόβλημα που εξετάζουμε (<100) και έναν ικανοποιητικό αριθμό πραγματικών δεδομένων επομένως θα ισχύει n>p. Σε περίπτωση που n<m σημαίνει ότι ο αριθμός m των συντελεστών που πρέπει να υπολογίσουμε είναι μεγαλύτερος από τον αριθμό n των διανυσμάτων των δεδομένων με βάση τα οποία θα γίνουν οι υπολογισμοί. Επομένως η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης δεν είναι εφαρμόσιμη και επιλέγονται άλλες προσεγγίσεις, οι οποίες δεν αποτελούν αντικείμενο της παρούσας εργασίας.

ii) Είναι σημαντικές όλες οι ανεξάρτητες μεταβλητές $x_1,x_2,\ldots x_n$ στη διαμόρφωση της τιμής της εξαρτημένης μεταβλητής Y ή μόνο κάποιες από αυτές;

Στην πράξη, σε πολλές περιπτώσεις μπορεί να έχουμε στη διάθεση μας πληθώρα δεδομένων για ανεξάρτητες μεταβλητές ώστε να διερευνήσουμε το πώς αυτά επηρεάζουν μια εξαρτημένη μεταβλητή και να εξάγουμε ένα γραμμικό μοντέλο. Όμως το γεγονός ότι διαθέτουμε την πληροφορία για τις ανεξάρτητες μεταβλητές, δε συνεπάγεται απαραίτητα ότι όλες οι μεταβλητές είναι χρήσιμες για το μοντέλο μας. Όσο μεγαλύτερη πληροφορία έχουμε, τόσο το καλύτερο από άποψη ανάλυσης και ακρίβειας, ωστόσο είναι πολύ σημαντικό να ξεσκαρτάρουμε αυτήν την πληροφορία μ και να κρατήσουμε μόνο όποια δεδομένα συμμετέχουν στη βέλτιστη εξαγωγή του γραμμικού μοντέλου.

Στο προηγούμενο βήμα περιγράψαμε με ποιον τρόπο μπορούμε να αποδείξουμε ότι τουλάχιστον μία ανεξάρτητη μεταβλητή σχετίζεται γραμμικά με το μοντέλο. Αυτό όμως δεν είναι αρκετό, καθώς αφενός δε μας εξασφαλίζει ότι όλες οι ανεξάρτητες μεταβλητές συνδέονται γραμμικά με το μοντέλο και αφετέρου εάν υπάρχουν ανεξάρτητες μεταβλητές που πρέπει να μείνουν εκτός, δε γνωρίζουμε ποιες είναι αυτές. Ο λόγος για τον οποίο μπορεί να χρειαστεί να μη λάβουμε υπόψη μας μία ή περισσότερες ανεξάρτητες μεταβλητές στην κατασκευή του γραμμικού μοντέλου, είναι ότι μπορεί αυτές να μη σχετίζονται με την εξαρτημένη μεταβλητή και έτσι να επηρεάζουν με λάθος τρόπο το μοντέλο, αλλοιώνοντας την ακρίβεια του. Έτσι, είναι προφανές πως αν αυτές οι μεταβλητές δεν συμπεριληφθούν στη γραμμική Σχέση (2.14) το μοντέλο που θα προκύψει θα προσεγγίζει τις αντίστοιχες πραγματικές τιμές με πιο βελτιωμένο τρόπο απ' ότι ένα μοντέλο του οποίου η γραμμική σχέση θα περιελάμβανε αυτές τις μεταβλητές. Σκοπός μας είναι να μπορέσουμε να δούμε εάν υπάρχουν τέτοιες ανεξάρτητες μεταβλητές αλλά και σε περίπτωση που υπάρχουν να εντοπίσουμε ποιες είναι αυτές.

Μία ιδανική μέθοδος θα ήταν να κάνουμε δοκιμές με διάφορα γραμμικά μοντέλα, καθένα από τα οποία θα περιλαμβάνει ένα διαφορετικό υποσύνολο ανεξάρτητων μεταβλητών από τα υπόλοιπα. Για παράδειγμα, εάν έχουμε ένα μοντέλο με 2 ανεξάρτητες μεταβλητές x_1 και x_2, εάν θέλουμε να κάνουμε δοκιμές με όλους τους δυνατούς διαφορετικούς συνδυασμούς, θα πρέπει να εξάγουμε 4 μοντέλα: ένα που θα περιλαμβάνει και τις 2 μεταβλητές, ένα που θα περιλαμβάνει μόνο τη x_1, ένα που θα περιλαμβάνει μόνο τη x_2 και ένα που δε θα έχει καμία ανεξάρτητη μεταβλητή (ευθεία παράλληλη στον οριζόντιο άξονα). Αφού έχουμε κατασκευάσει όλα τα δυνατά μοντέλα, μπορούμε να επιλέξουμε το βέλτιστο από αυτά με διάφορα κριτήρια, όπως είναι η τιμή του όρου R^2 . Μία τέτοια προσέγγιση όμως, όπως είναι προφανές δεν είναι εφαρμόσιμη στην περίπτωση μεγάλου αριθμού ανεξάρτητων μεταβλητών, καθώς εάν έχουμε m ανεξάρτητες μεταβλητές, ο αριθμός των διαφορετικών γραμμικών μοντέλων που μπορεί να κατασκευαστούν είναι 2^m. Έτσι για έναν αριθμό π.χ. 10 ανεξάρτητων μεταβλητών θα πρέπει να κατασκευάσουμε 1024 διαφορετικά γραμμικά μοντέλα! Είναι ξεκάθαρο ότι πρέπει να βρούμε έναν πιο πρακτικό τρόπο για να εντοπίσουμε το βέλτιστο γραμμικό μοντέλο από όλα τα πιθανά.

Υπάρχουν συνολικά τρεις διαφορετικές μέθοδοι διαδοχικών δοκιμών τις οποίες μπορούμε να εφαρμόσουμε ώστε να εντοπίσουμε το βέλτιστο γραμμικό μοντέλο:

- **Διαδικασία Forward**: Ξεκινάμε με το μοντέλο το οποίο περιέχει μόνο τη σταθερά α και καθόλου ανεξάρτητες μεταβλητές. Στη συνέχεια κατασκευάζουμε τα m διαφορετικά μοντέλα, καθένα από τα οποία περιλαμβάνει μόνο μία από τις m ανεξάρτητες μεταβλητές, θα έχει δηλαδή τη μορφή $Yi = \beta_i \cdot x_i$. Υπολογίζουμε τον όρο RSS για κάθε μοντέλο και την ανεξάρτητη μεταβλητή του μοντέλου που έχει το χαμηλότερο RSS την προσθέτουμε στο μοντέλο που περιείχε μόνο τη σταθερά α. Έτσι πλέον έχουμε ένα γραμμικό μοντέλο που αποτελείται από τη σταθερά α και μία ανεξάρτητη μεταβλητή. Από τις m-1 εναπομείνασες ανεξάρτητες μεταβλητές προσθέτουμε μόνο μία κάθε φορά στο μοντέλο, δημιουργώντας m-1 διαφορετικά μοντέλα που περιλαμβάνουν το α, μία σίγουρη ανεξάρτητη μεταβλητή και άλλη μία υποψήφια. Κρατάμε το μοντέλο που δίνει το μικρότερο RSS. Συνεχίζουμε ούτω καθεξής μέχρι να ικανοποιηθεί κάποια συνθήκη που θα μας δείχνει ότι πρέπει να σταματήσουμε να προσθέτουμε ανεξάρτητες μεταβλητές. Για παράδειγμα, για κάθε φορά που προσθέτουμε μία ανεξάρτητη μεταβλητή μπορούμε να ελέγχουμε την τιμή του όρου R^2 (είναι δείκτης που αξιολογεί πόσο καλά το γραμμικό μοντέλο προσεγγίζει τις πραγματικές τιμές. Θα αναλύσουμε αυτόν τον όρο στο Κεφάλαιο 4 και θα τον υπολογίζουμε με τη βοήθεια της R). Όσο η τιμή του αυξάνεται με την προσθήκη ανεξάρτητων μεταβλητών, συνεχίζουμε τη διαδικασία. Όταν φτάσουμε σε σημείο που θα προσθέσουμε μία επιπλέον ανεξάρτητη μεταβλητή και ο όρος R^2 για το μοντέλου που θα προκύψει, θα είναι μικρότερος σε σχέση με το αμέσως προηγούμενο μοντέλο, συμπεραίνουμε πως στην αμέσως προηγούμενη δοκιμή επιτύχαμε το βέλτιστο μοντέλο και πρέπει να σταματήσουμε να προσθέτουμε άλλες ανεξάρτητες μεταβλητές, καθώς από εκεί και πέρα επηρεάζεται αρνητικά η ακρίβεια του. Ομοίως εάν στο μοντέλο προσθέτουμε ανεξάρτητες μεταβλητές και ο όρος R^2 δε βελτιώνεται αλλά παραμένει αμετάβλητος σημαίνει πως πρέπει να σταματήσουμε, καθώς κάνουμε πιο πολύπλοκο το μοντέλο χωρίς ουσιαστικά να βελτιώνουμε την ακρίβεια του

- **Διαδικασία Backward**: Η διαδικασία αυτή ξεκινά αντίστροφα από τη διαδικασία Forward, δηλαδή ξεκινάμε με το γραμμικό μοντέλο το οποίο περιέχει όλες τις ανεξάρτητες μεταβλητές. Στο σημείο αυτό πρέπει να αναφέρουμε πως σε κάθε ανεξάρτητη μεταβλητή αντιστοιχεί μία τιμή Pr, η οποία εκφράζει πιθανότητα, άρα λαμβάνει τιμές από 0 έως 1 και μπορούμε να την υπολογίσουμε με διάφορους τρόπους, όπως με τη βοήθεια της R . Εάν σε μία ανεξάρτητη μεταβλητή αντιστοιχεί τιμή της Pr κοντά στο 1, σημαίνει πως υπάρχει μεγάλη πιθανότητα να ισχύει η μηδενική υπόθεση για την εν λόγω

μεταβλητή, η οποία είναι πιθανόν να μην είναι σημαντική για το γραμμικό μοντέλο. Άρα ανεξάρτητες μεταβλητές, υποψήφιες για μη συμμετοχή στο μοντέλο είναι αυτές στις οποίες αντιστοιχεί μεγάλη τιμή του Pr. Έτσι είναι ξεκάθαρο πως σε πρώτη φάση αφαιρούμε από το πλήρες γραμμικό μοντέλο με το οποίο ξεκινήσαμε, την ανεξάρτητη μεταβλητή στην οποία αντιστοιχεί η μεγαλύτερη τιμή για την p-value. Έτσι προκύπτει ένα μοντέλο με m-1 ανεξάρτητες μεταβλητές. Συνεχίζουμε την ίδια διαδικασία, δηλαδή αφαιρούμε την μεταβλητή εκείνη στην οποία αντιστοιχεί η μεγαλύτερη τιμή της p-value. Η διαδικασία Backward συνεχίζεται έως ότου φτάσουμε σε μία επιθυμητή συνθήκη, δηλαδή για παράδειγμα όταν τα p-value που αντιστοιχούν στις εναπομείνασες ανεξάρτητες μεταβλητές είναι όλα μικρότερα από ένα ανώτατο κατώφλι που έχουμε θέσει. Φυσικά πρέπει να είμαστε σίγουροι ότι δε θα αφαιρέσουμε περισσότερες ανεξάρτητες μεταβλητές από αυτές που πρέπει, γιατί θα επηρεαστεί η ακρίβεια του μοντέλου. Έτσι π.χ. μπορούμε να πούμε ότι αφαιρούμε διαδοχικά ανεξάρτητες μεταβλητές που χαρακτηρίζονται από σχετικά μεγάλο p-value, όσο η τιμή του R^2 παραμένει βέλτιστη. Στη δοκιμή στην οποία το R^2 μειώνεται σε σχέση με πριν, σημαίνει ότι πρέπει να σταματήσουμε να αφαιρούμε ανεξάρτητες μεταβλητές γιατί πλέον το μοντέλο χάνει ακρίβεια.

- **Μικτή Διαδικασία:** Η διαδικασία αυτή αποτελεί έναν συνδυασμό των διαδικασιών Forward και Backward. Ξεκινάμε με την πιο απλή μορφή του γραμμικού μοντέλου, που δεν περιέχει καμία ανεξάρτητη μεταβλητή, παρά μόνο τη σταθερά α και προσθέτουμε με τη λογική Forward που περιγράψαμε τις ανεξάρτητες μεταβλητές διαδοχικά, μία σε κάθε βήμα. Παράλληλα, σε κάθε βήμα στο οποίο προστίθεται μία ανεξάρτητη μεταβλητή στο μοντέλο, πρέπει να ελέγχουμε το Pr καθεμίας, γιατί υπάρχει η περίπτωση όταν στο γραμμικό μοντέλο προστεθούν νέες ανεξάρτητες μεταβλητές, το Pr κάποιας να αυξηθεί. Εάν η αύξηση του Pr κάποιας μεταβλητής ξεπεράσει ένα ανώτατο κατώφλι, τότε αυτή βγαίνει εκτός του μοντέλου. Συνεχίζουμε τη μικτή διαδικασία, προσθέτοντας μεταβλητές και ελέγχοντας τη μεταβολή των Pr που τους αντιστοιχούν έως ότου καταλήξουμε σε ένα γραμμικό μοντέλο το οποίο περιέχει ανεξάρτητες μεταβλητές με Pr κοντά στο 0, ενώ όλες οι ανεξάρτητες μεταβλητές που έχουν μείνει εκτός, θα έχουν μεγάλο Pr εάν προστεθούν στο μοντέλο.

iii) Πόσο καλά το γραμμικό μας μοντέλο προσεγγίζει τα πραγματικά δεδομένα;

Όταν καταλήξουμε στο στάδιο στο οποίο έχουμε επιλέξει τη βέλτιστη εκδοχή του γραμμικού μοντέλου, επομένως δε θα επιδέχεται περαιτέρω διόρθωση, θα πρέπει να αξιολογήσουμε την ακρίβεια του, δηλαδή το κατά πόσο οι τιμές της εξαρτημένης μεταβλητής Y που ανήκουν στην ευθεία. προσεγγίζουν τις αντίστοιχες πραγματικές τιμές. Έχουμε αναφέρει πως ένας τέτοιος δείκτης είναι ο όρος RSE, με την αξιολόγηση του ωστόσο να είναι κάποιες φορές σχετική, καθώς εκφράζεται στη μονάδα μέτρησης του Y και εξαρτάται από τη φύση του προβλήματος για το αν η τιμή του στα πλαίσια αποδεκτών ορίων. Για το λόγο αυτό είναι προτιμότερο να δίνουμε βαρύτητα στον όρο R^2 ο οποίος εκφράζει ποσοστό, επομένως μπορεί να αξιολογηθεί χωρίς περεταίρω ανάλυση. Μία τιμή του R^2 κοντά στο 0, όπως έχουμε εξηγήσει, υποδηλώνει πως η παρατηρούμενη διακύμανση στις τιμές του Y, δεν μπορεί να δικαιολογηθεί από τις τιμές των ανεξάρτητων μεταβλητών. Αντίθετα μία τιμή του R^2 κοντά στη μονάδα εξηγεί καλύτερα τη διακύμανση αυτή.

Στο πρόβλημα που θα αναλύσουμε στο Κεφάλαιο 5, θα δούμε αναλυτικά τη διαδικασία κατασκευής ενός μοντέλου με τη βοήθεια της R, καθώς και τη διαδικασία επιλογής της βέλτιστης μορφής του μοντέλου αυτού με τους αντίστοιχους ελέγχους.

ΚΕΦΑΛΑΙΟ 3: ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ (CLUSTER ANALYSIS)

3.1 Εισαγωγή

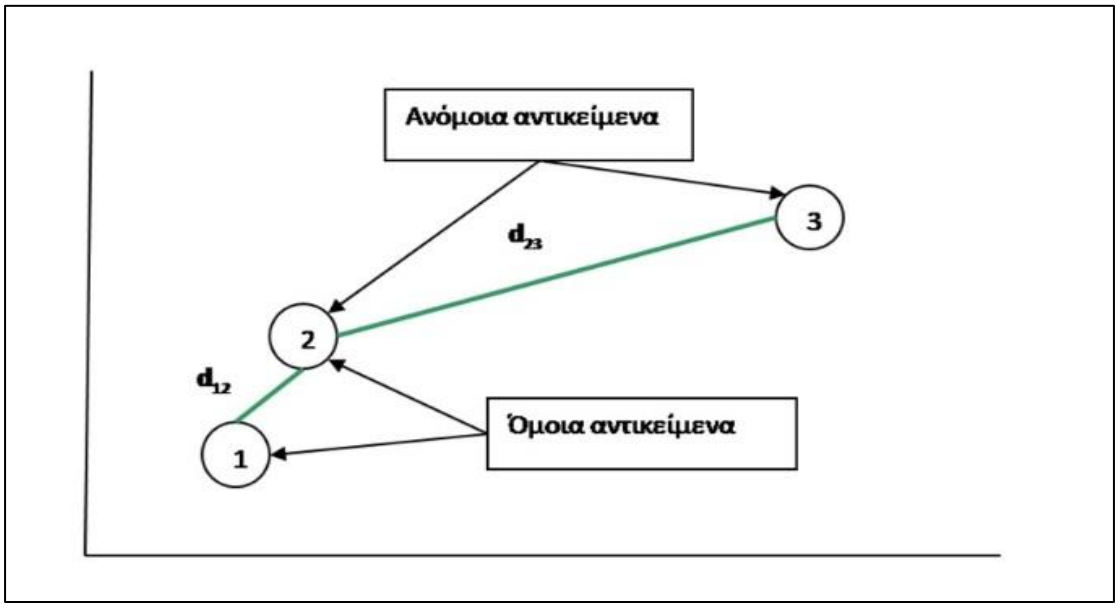
Η Ανάλυση Συστάδων (Cluster analysis) αποτελεί μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Είναι η διαδικασία κατά την οποία ένα σύνολο δεδομένων ομαδοποιείται με βάση κάποιο μέτρο ομοιότητας. Χρησιμοποιείται εκτεταμένα στην επιστημονική έρευνα, όπου υπάρχει η ανάγκη ταξινόμησης και κατάταξης των αντικειμένων μελέτης σε ομάδες. Η ομαδοποίηση αυτή εξαρτάται από το είδος των δεδομένων και τη φύση του προβλήματος στο οποίο εφαρμόζουμε αυτήν τη διαδικασία. Στα πλαίσια της Εξόρυξης Δεδομένων, η ανάλυση συστάδων έχει πολλαπλή χρησιμότητα. Ως αυτόνομη αναλυτική εργασία, επιτρέπει στον αναλυτή να επιμερίσει τα δεδομένα σε ομάδες ομοειδών παρατηρήσεων. Ακολούθως, ο αναλυτής μπορεί να επικεντρωθεί στην εκάστοτε ομάδα, να αναγνωρίσει τα κοινά χαρακτηριστικά της, και να εξάγει γνώση χρήσιμη για τη λήψη αποφάσεων. Για παράδειγμα, στα χρηματοοικονομικά οι μετοχές που διαπραγματεύονται σε ένα χρηματιστήριο μπορούν να ομαδοποιηθούν με βάση τη διακύμανση των τιμών τους και το βαθμό στον οποίο τείνουν να κινούνται προς την ίδια κατεύθυνση διαχρονικά. Στο μάρκετινγκ και την έρευνα αγοράς μια βασική ανάγκη είναι η τμηματοποίηση της αγορά-στόχου. Ο όρος τμηματοποίηση της αγοράς περιγράφει τον επιμερισμό των καταναλωτών σε ομάδες με όμοια καταναλωτική συμπεριφορά. Με την ομαδοποίηση αυτή, δίνεται η δυνατότητα στη συνέχεια να σχεδιαστούν πολιτικές μάρκετινγκ για κάθε τμήμα.

Η ανάλυση σε ομάδες έχει σκοπό να διαχωρίσει ένα σύνολο παρατηρήσεων σε φυσικές ομάδες, (υποσύνολα) έτσι ώστε τα μέλη κάθε ομάδας να είναι όσο το δυνατόν πιο όμοια μεταξύ τους, ενώ τα μέλη διαφορετικών ομάδων να είναι όσο το δυνατόν πιο ανόμοια. Δηλαδή μια επιτυχημένη ανάλυση συστάδων θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς, ενώ παράλληλα παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

Το πρώτο και σημαντικότερο βήμα κατά τη διαδικασία της ανάλυσης συστάδων είναι η περιγραφή των δεδομένων και η επιλογή των κατάλληλων χαρακτηριστικών. Στη συνέχεια, πρέπει να οριστεί το μέτρο ομοιότητας με το οποίο θα γίνονται οι συγκρίσεις μεταξύ των παρατηρήσεων. Βασικές έννοιες για την ανάλυση κατά συστάδες είναι οι έννοιες της απόστασης και της ομοιότητας. Είναι εύκολο να διαπιστωθεί πως αυτές οι δύο έννοιες είναι αντίθετες, παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα.

3.2 Η Έννοια της Απόστασης στην Ανάλυση Συστάδων

Η **απόσταση** στην ανάλυση συστάδων αποτελεί το μέτρο εγγύτητας που ποσοτικοποιεί την ομοιότητα ή την ανομοιότητα μεταξύ δύο, κάθε φορά, παρατηρήσεων στα δεδομένα. Κάθε παρατήρηση στον πολυδιάστατο χώρο των μεταβλητών έχει ένα σύνολο συντεταγμένων οι οποίες αποτελούνται από τις τιμές για όλες τις μεταβλητές, όπως είδαμε και το παράδειγμα με τη διαφημιστική δαπάνη στα διάφορα μέσα και τις πωλήσεις του προϊόντος. Με βάση τις συντεταγμένες αυτές υπολογίζονται αποστάσεις μεταξύ των παρατηρήσεων ή των ομάδων παρατηρήσεων. Με βάση την ανάλυση κατά συστάδες, όμοια αντικείμενα-παρατηρήσεις θεωρούνται 2 παρατηρήσεις που βρίσκονται σε κοντινές αποστάσεις μεταξύ τους, δηλαδή οι συντεταγμένες τους γειτνιάζουν, ενώ ανόμοια αντικείμενα είναι αυτά που έχουν μεγάλη απόσταση μεταξύ τους, σύμφωνα με το παρακάτω διάγραμμα:



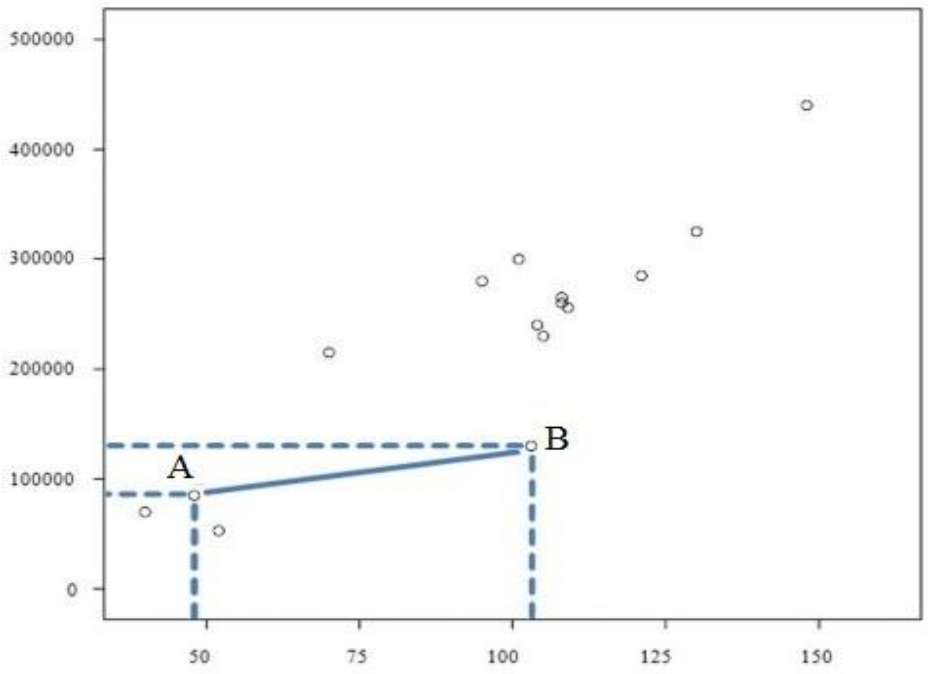
Εικόνα 5*: Παρατήρηση ομοιότητας αντικειμένων με κριτήριο την απόσταση τους*

Ο απλούστερος τρόπος μέτρησης της απόστασης είναι η επέκταση του Πυθαγορείου θεωρήματος σε πολλές διαστάσεις, ώστε για δύο παρατηρήσεις x και y προϊόντος η απόστασή τους d(x,y) (Ευκλείδια Απόσταση) δίνεται από τον τύπο:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2} = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

όπου m είναι ο αριθμός των μεταβλητών (m-διάστατος χώρος) και x_i,y_i οι τιμές των παρατηρήσεων.

Στο διάγραμμα 2 μεταβλητών (2-διάστατος χώρος) που ακολουθεί φαίνεται η απόσταση μεταξύ 2 παρατηρήσεων $A(x_1,x_2)$ και $B(y_1,y_2)$:



***Εικόνα 6**: Παρατήρηση ομοιότητας αντικειμένων με κριτήριο την απόσταση τους*

Γενικά καθώς το m μεταβάλλεται δίνει μεγαλύτερη βαρύτητα σε μεγαλύτερες ή μικρότερες διαφοροποιήσεις των παρατηρήσεων. Όπως γίνεται αντιληπτό, για περισσότερες από 2 παρατηρήσεις, θα πρέπει να υπολογιστεί ένας Πίνακας ο οποίος θα αποτελείται από τις τιμές όλων των Ευκλείδιων αποστάσεων κάθε παρατήρησης από τις υπόλοιπες. Με τη βοήθεια κατάλληλου λογισμικό, υπολογίζεται η απόσταση κάθε παρατήρησης από όλες τις υπόλοιπες παρατηρήσεις ώστε να εντοπιστεί ποιες παρατηρήσεις παρουσιάζουν μεγαλύτερη ομοιότητα. Εάν n είναι ο αριθμός των παρατηρήσεων, προκύπτει ένας Πίνακας nxn ο οποίος θα έχει μηδενικά στοιχεία στη διαγώνιο (εφόσον στη διαγώνιο θα είναι η απόσταση κάθε παρατήρησης από τον εαυτό της) και την απόσταση μεταξύ του i στοιχείου και του j στοιχείου (στην θέση (i, j). Οι μικρές αποστάσεις αντιστοιχούν σε παρόμοιες παρατηρήσεις, δηλαδή σε παρατηρήσεις που πιθανότατα θα ανήκουν στην ίδια ομάδα και οι μεγαλύτερες αποστάσεις σε παρατηρήσεις με μεγάλες διαφορές στις τιμές των μεταβλητών.

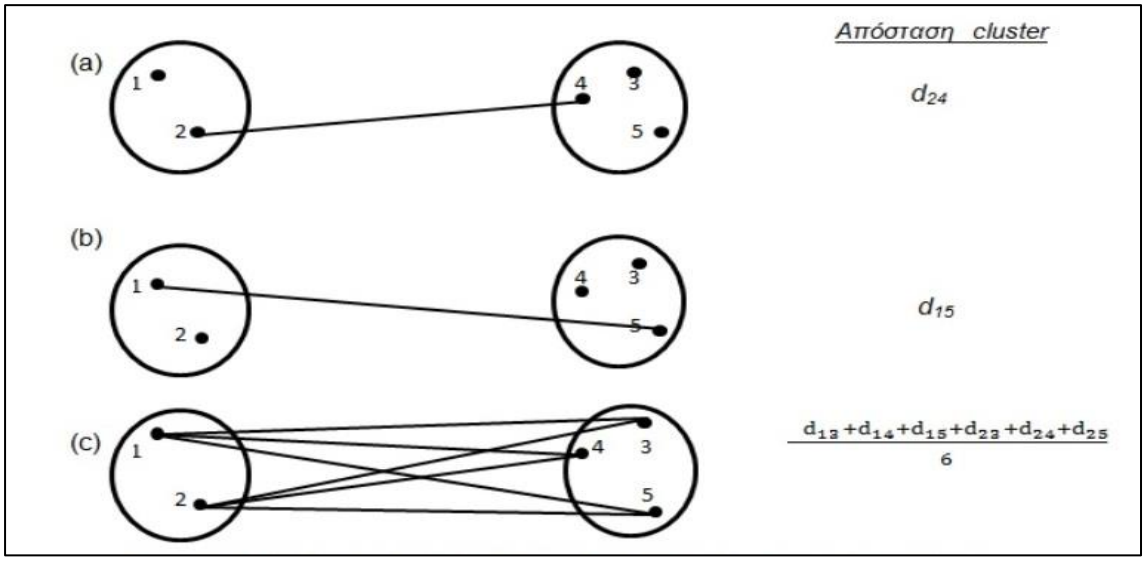
3.3. Μέθοδοι Υπολογισμού Απόστασης

Υπάρχουν διάφορες τεχνικές για τον υπολογισμό των αποστάσεων των μεταξύ παρατηρήσεων ή συστάδων. Οι πιο δημοφιλείς είναι οι εξής:

-Κριτήριο Κοντινότερου Γείτονα (Single Linkage): Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες U και V, ως τη μικρότερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Η μέθοδος έχει κάποιες χρήσιμες μαθηματικές ιδιότητες αλλά παράγει ομάδες που δεν είναι συμπαγείς και συνήθως δημιουργεί μερικές πολύ μεγάλες ομάδες και κάποιες πάλι πολύ μικρές.

-Κριτήριο πιο Απομακρυσμένου Γείτονα (Complete Linkage): Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες U και V, ως τη μεγαλύτερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση σε κ άλλη ομάδα. Οι ομάδες που δημιουργούνται είναι συνήθως συμπαγείς αλλά αποτυγχάνει να δημιουργήσει κάποιες μικρές αλλά πολύ συμπαγείς ομάδες.

-Average Linkage between Groups: Ως απόσταση μεταξύ δύο συστάδων U και V θεωρούμε την μέση απόσταση μεταξύ των δύο συστάδων (το άθροισμα όλων των αποστάσεων μεταξύ ενός στοιχείου της ομάδας U και ενός στοιχείου του V διά του γινομένου του πλήθους των στοιχείων της U επί του πλήθους των στοιχείων της V).



***Εικόνα 7**: Απεικόνιση απόστασης ανάμεσα σε clusters για τις μεθόδους (a) Single Linkage, (b) Complete Linkage, (c) Average Linkage between Group*

-Average within Groups: Ως απόσταση, λογίζεται ο μέσος όλων των αποστάσεων των στοιχείων, που προκύπτουν όταν ενώσουμε τις δύο ομάδες.

Μέθοδος του Ward: Η μέθοδος αυτή δεν υπολογίζει αποστάσεις ανάμεσα στις συστάδες. Σχηματίζει συστάδες στο εσωτερικό των υφιστάμενων συστάδων μεγιστοποιώντας την ομοιογένεια. Κριτήριο για τη δημιουργία συστάδων είναι η μεγιστοποίηση της ομοιογένειας στο εσωτερικό των συστάδων. Ως μέτρο ομοιογένειας χρησιμοποιείται το άθροισμα των τετραγώνων των αποστάσεων των στοιχείων μέσα σε μια δημιουργούμενη συστάδα, από το μέσο σημείο της συστάδας αυτής. Επομένως η μέθοδος του Ward προσπαθεί να ελαχιστοποιήσει το συνολικό άθροισμα τετραγώνων των αποστάσεων αυτών μέσα στο cluster.

Τα δημιουργούμενα clusters σε κάθε βήμα σχηματίζονται έτσι ώστε η λύση που προκύπτει να δίνει το μικρότερο δυνατό άθροισμα τετραγώνων μέσα στο αρχικό cluster. Τα αθροίσματα των τετραγώνων των αποστάσεων που ελαχιστοποιούνται είναι γνωστά και ως **αθροίσματα τετραγωνικών σφαλμάτων (Sum of Squared Errors -SSE).** Στη μέθοδο αυτή, ως απόσταση μεταξύ 2 συστάδων U και V λογίζεται η αύξηση που θα προκύψει στο SSE από την ένωση των δύο συστάδων. Ως απόσταση μεταξύ δύο συστάδων U και V θεωρούμε την απόσταση με τη μικρότερη τιμή από όλες τις πιθανές αποστάσεις μεταξύ ενός στοιχείου (ή συστάδας) του U και ενός στοιχείου (ή συστάδας) του V. Η μέθοδος, για να συνενώσει δύο συστάδες από συνολικό πλήθος k συστάδων, ελέγχει τα δυνατά $k\cdot(k-1)/2$ ζεύγη συστάδων τα οποία μπορούν να δημιουργηθούν και επιλέγει το ζεύγος, το οποίο όταν ενωθεί θα μας δώσει τη συστάδα με το ESS. Το ίδιο κριτήριο χρησιμοποιείται και από τον αλγόριθμο k-Means που θα αναλυθεί παρακάτω, οπότε η μέθοδος Ward μπορεί να θεωρηθεί το ιεραρχικό ανάλογο του k-Means.

Από τις παραπάνω μεθόδους η πιο απλή είναι η μέθοδος του κοντινότερου γείτονα η οποία όμως έχει το μειονέκτημα πως δίνει ομάδες με μεγάλες διαφορές ως προς το μέγεθος τους. Η μέθοδος του Ward έχει το πλεονέκτημα ότι μας δίνει περίπου ισοπληθείς ομάδες και για αυτό είναι προτιμότερη.

Οι ιεραρχικές μέθοδοι ομαδοποίησης έχουν τα εξής πλεονεκτήματα:

- ✓ Παρουσιάζουν καλή προσαρμοστικότητα. Μπορούν να εντοπίσουν καλά διαχωρισμένες, επιμήκεις και ομόκεντρες συστάδες.
- ✓ Δημιουργούν πολλαπλά επίπεδα φωλιασμένων συστάδων και επιτρέπουν στον χρήστη να επιλέξει το επίπεδο που αυτός επιθυμεί

Έχουν αναφερθεί όμως και αρκετά μειονεκτήματα της ιεραρχικής ομαδοποίησης. Ένα βασικό μειονέκτημα είναι ότι για μεγάλα σύνολα δεδομένων απαιτείται μεγάλος υπολογιστικός χρόνος. Επίσης, όπως ήδη αναφέραμε, οι ομάδες που δημιουργούνται στα αρχικά βήματα δεν μπορούν να μεταβληθούν στη συνέχεια. Κάθε ενέργεια, η οποία πραγματοποιείται σε ένα στάδιο, δεν είναι αντιστρέψιμη. Από τη στιγμή που

δύο αντικείμενα ενταχθούν στην ίδια ομάδα, θα παραμείνουν στην ίδια ομάδα, και δεν υπάρχει δυνατότητα να διαχωριστούν αργότερα και να ενταχθούν σε διαφορετικές ομάδες

Συχνά όμως, στη διαδικασία της ανάλυσης προκύπτει ανάγκη τροποποίησης των ομάδων, προκειμένου να δημιουργηθούν ομάδες οι οποίες να ανταποκρίνονται στα δεδομένα του ερευνητικού προβλήματος. Επίσης η ιεραρχική ταξινόμηση γενικά δημιουργεί μερικές ομάδες με πολλές παρατηρήσεις και αφήνει κάποιες παρατηρήσεις να αποτελούν μόνες τους μία ομάδα. Τέλος πρέπει να τονιστεί ότι η μέθοδος παράγει λύση για κάθε διαφορετικό αριθμό ομάδων, δηλαδή ο αριθμός των ομάδων δεν είναι γνωστός από πριν. Επομένως ο αναλυτής θα πρέπει να επιλέξει ποια ομαδοποίηση θα κρατήσει.

3.4 Μέθοδοι Ανάλυσης Κατά Συστάδες

Στην ανάλυση κατά συστάδες υπάρχουν 2 διαφορετικές βασικές προσεγγίσεις:

3.4.1 Ιεραρχικές Μέθοδοι Ομαδοποίησης

Οι ιεραρχικές μέθοδοι ομαδοποίησης προχωρούν είτε συσσωρευτικά, με μια σειρά διαδοχικών συγχωνεύσεων, είτε επιμεριστικά με μια σειρά διαδοχικών διαιρέσεων. Η μία κατηγορία μεθόδων δηλαδή ακολουθεί την αντίστροφη διαδικασία από την άλλη κατηγορία. Έτσι οι ιεραρχικές μέθοδοι ομαδοποίησης διαχωρίζονται στις συσσωρευτικές και στις διαιρετικές ή διαχωριστικές μεθόδους.

3.4.1.1 Συσσωρευτικές (Προσθετικές) μέθοδοι (Agglomerative Hierarchical Clustering)

Ξεκινούν με μεμονωμένες παρατηρήσεις, καθεμία από τις οποίες αποτελεί μία συστάδα. Κατά συνέπεια, αρχικά υπάρχουν τόσες συστάδες όσες και οι παρατηρήσεις. Οι πιο όμοιες παρατηρήσεις ομαδοποιούνται και οι αρχικές αυτές ομάδες συγχωνεύονται σύμφωνα με τις ομοιότητες τους στο επόμενο επίπεδο. Όσο προχωράμε σε επόμενα επίπεδα και η ομοιότητα μεταξύ των ομάδων μειώνεται, όλες οι ομάδες συγχωνεύονται σε μία ενιαία συστάδα. Τα βήματα που ακολουθούνται περιγράφονται περιληπτικά:

1) Αρχίζουμε με n συστάδες, με την κάθε μία να περιέχει μόνο ένα στοιχείο και έναν Πίνακα διαστάσεων n×n που τα στοιχεία του εκφράζουν αποστάσεις μεταξύ των συστάδων.

2) Βρίσκουμε στον πίνακα το ζεύγος U και V συστάδων με την μικρότερη απόσταση μεταξύ τους.

3) Ενώνουμε τις συστάδες U και V σε μια συστάδα, έστω UV. Ανανεώνουμε τον πίνακα αποστάσεων διαγράφοντας τις γραμμές και στήλες που αντιστοιχούν στις U και V και προσθέτοντας μια γραμμή και μια στήλη με τις αποστάσεις της UV από τις υπόλοιπες συστάδες.
4) Επαναλαμβάνουμε τα βήματα 2 και 3 (n−1) φορές μέχρι να υπάρχει μόνο μια συστάδα. Καταγράφουμε τις συστάδες που δημιουργήθηκαν κατά τη διάρκεια της διαδικασίας και το επίπεδο (απόσταση) στο οποίο δημιουργήθηκε η κάθε μία.

3.4.1.2 Διαιρετικές Μέθοδοι (Divisive Analysis Clustering)

Εδώ ακολουθείται η αντίστροφη διαδικασία, δηλαδή όλες οι παρατηρήσεις ανήκουν σε μία αρχική συστάδα, η οποία διαιρείται σε 2 υποομάδες. Η διάσπαση γίνεται με τέτοιον τρόπο, ώστε οι 2 υποομάδες που θα προκύψουν να έχουν μεταξύ τους τη μεγαλύτερη δυνατή ανομοιότητα. Αυτές οι 2 υποομάδες διασπώνται στη συνέχεια σε νέες όσο το δυνατόν ανόμοιες υποομάδες. Η διαδικασία συνεχίζεται έως ότου προκύψουν τόσες υποομάδες όσες και οι παρατηρήσεις μας.

Τα βήματα που ακολουθούνται τυπικά στις διαιρετικές μεθόδους ανάλυσης συστάδων περιγράφονται περιληπτικά:

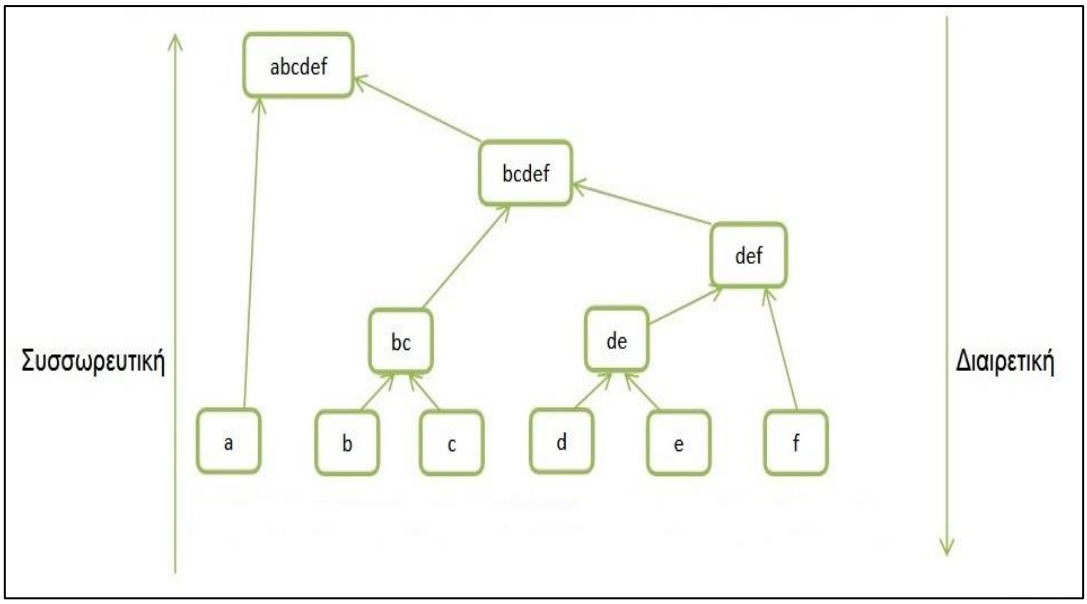
1) Αρχικά επιλέγουμε μια συστάδα.

2) Επιλέγουμε μετά το στοιχείο με τη μεγαλύτερη μέση απόσταση από τα υπόλοιπα στοιχεία της συστάδας, το οποίο γίνεται μια νέα συστάδα.

3) Κατανέμουμε τα στοιχεία της συστάδας είτε στην παλιά συστάδα είτε στην νέα, βάση της απόστασης του κάθε στοιχείου από τις συστάδες.

4) Επιλέγουμε τη συστάδα με τη μεγαλύτερη διάμετρο (μεγαλύτερη απόσταση μεταξύ δυο στοιχείων της συστάδας) και επιστρέφουμε στο βήμα 2 μέχρι να έχουμε τόσες συστάδες όσα τα στοιχεία μας.

Η διαδικασία των συσσωρευτικών και διαιρετικών μεθόδων συνοψίζεται στην ακόλουθη εικόνα:



***Εικόνα 8**: Αναπαράσταση Συσσωρευτικής και Διαιρετικής Ομαδοποίησης*

3.4.2 <u>Διαχωριστικές Μέθοδοι Ομαδοποίησης</u>

Στη μη ιεραρχική ομαδοποίηση, τα δεδομένα διαιρούνται σε k διαμερίσεις, καθεμία από τις οποίες αντιπροσωπεύει ένα cluster. Σε αντίθεση με την ιεραρχική ομαδοποίηση, ο αριθμός των k Cluster μπορεί είτε να διευκρινιστεί εκ των προτέρων, ή να καθοριστεί σαν μέρος της διαδικασίας ομαδοποίησης. Στις μεθόδους αυτής της κατηγορίας εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Ο πιο γνωστός αλγόριθμος αυτής της κατηγορίας είναι ο k-Means ο οποίος θα αναλυθεί στη συνέχεια.

Το πλεονέκτημα των μη ιεραρχικών μεθόδων έναντι των ιεραρχικών είναι πως μπορούν να εφαρμοστούν σε πολύ μεγαλύτερο όγκο δεδομένων. Ο λόγος είναι πως σε αντίθεση με τις ιεραρχικές μεθόδους, δε χρειάζεται να καθοριστεί ένας Πίνακας αποστάσεων μεταξύ των αντικειμένων που θέλουμε να οργανώσομε σε συστάδες και τα βασικά δεδομένα δε χρειάζεται να αποθηκευτούν στον υπολογιστή κατά το τρέξιμο του αλγορίθμου.

Οι τεχνικές διαχωριστικής ομαδοποίησης ακολουθούν κατά βάση τα εξής βήματα:

1) Επιλέγουμε k αρχικά κέντρα ή κομβικά σημεία (seed points) των clusters, όπου k είναι ο επιθυμητός αριθμός των clusters.
2) Αναθέτουμε κάθε παρατήρηση στο cluster, στο οποίο αυτή είναι η πλησιέστερη.
3) Αναθέτουμε εκ νέου ή ανακατανέμουμε κάθε παρατήρηση σε ένα cluster σύμφωνα με έναν προκαθορισμένο κανόνα τερματισμού.
4) Σταματάμε εάν δεν υπάρχει καμία ανακατανομή των παρατηρήσεων ή αν η κατανομή ικανοποιεί το σύνολο των κριτηρίων, από τον κανόνα τερματισμού. Αλλιώς επιστρέφουμε στο Βήμα 2.

Οι μη ιεραρχικές μέθοδοι αρχίζουν είτε:

α) από μια αρχική διαμέριση των αντικειμένων σε ομάδες είτε

β) από ένα αρχικό σύνολο κομβικών σημείων τα οποία θα διαμορφώσουν τον πυρήνα των συστάδων

Οι περισσότεροι από τους μη ιεραρχικούς αλγορίθμους διαφέρουν μεταξύ τους σε σχέση με τη μέθοδο που χρησιμοποιείται για την επιλογή των αρχικών σημείων και σε σχέση με τον κανόνα που χρησιμοποιείται για την ανακατανομή των παρατηρήσεων.

Οι βασικότερες μέθοδοι που χρησιμοποιούνται για τα αρχικά κομβικά σημεία, εάν έχουμε n παρατηρήσεις είναι οι εξής:

- Επιλέγονται οι k πρώτες παρατηρήσεις με μη ελλιπή δεδομένα σαν κομβικά σημεία για τα αρχικά cluster.
- Επιλέγεται η πρώτη παρατήρηση με μη ελλιπή δεδομένα σαν κομβικό σημείο για το πρώτο cluster. Το κομβικό σημείο για το δεύτερο cluster επιλέγεται έτσι ώστε η απόσταση του από το προηγούμενο κομβικό σημείο να είναι μεγαλύτερη από μία καθορισμένη απόσταση που έχει προκαθοριστεί. Το κομβικό σημείο του τρίτου cluster επιλέγεται έτσι ώστε η απόσταση του από τα άλλα 2 κομβικά σημεία που επιλέχθηκαν να είναι μεγαλύτερη από την καθορισμένη απόσταση και ούτω καθεξής.
- Βελτιώνουμε τα κομβικά σημεία που επιλέχθηκαν, χρησιμοποιώντας κάποιους κανόνες ώστε αυτά να είναι όσο το δυνατόν πιο απομακρυσμένα.
- Χρησιμοποιούμε κάποιον αλγόριθμο που προσδιορίζει τα κομβικά σημεία των cluster, ώστε αυτά να είναι όσο το δυνατόν απομακρυσμένα.
- Χρησιμοποιούμε κομβικά σημεία προμηθευμένα από τον ερευνητή

Μόλις προσδιοριστούν τα κομβικά σημεία, σχηματίζονται τα αρχικά cluster, αναθέτοντας τις υπόλοιπες n-k παρατηρήσεις στο κομβικό εκείνο σημείο στο οποίο η κάθε παρατήρηση είναι πλησιέστερη.

Οι μη ιεραρχικές μέθοδοι διαφέρουν επίσης σε ό,τι αφορά στη διαδικασία που χρησιμοποιούν για την ανακατανομή των παρατηρήσεων στα k cluster. Οι πιο συνήθεις κανόνες ανακατανομής είναι οι εξής:

- Υπολογίζουμε το κομβικό σημείο κάθε cluster και αναθέτουμε εκ νέου αντικείμενα, σε εκείνο το cluster που έχει το πλησιέστερο σε αυτά κομβικό σημείο. Τα κομβικά σημεία υπολογίζονται εκ νέου, αφού έχει γίνει η ανάθεση όλων των παρατηρήσεων. Εάν η μεταβολή τους είναι μεγαλύτερη από ένα κριτήριο σύγκλισης που έχουμε ορίσει, τότε αυτά επαναπροσδιορίζονται. Η διαδικασία ανακατανομής συνεχίζεται μέχρι η μεταβολή των κομβικών σημείων να είναι μικρότερη από το όριο του καθορισμένου κριτηρίου σύγκλισης.
- Υπολογίζουμε το κομβικό σημείο κάθε cluster και αναθέτουμε εκ νέου τα αντικείμενα στο cluster με το πλησιέστερο κομβικό σημείο. Για την ανάθεση κάθε αντικειμένου, υπολογίζουμε εκ νέου το κομβικό σημείο του cluster εκείνο στο οποίο καταχωρείται το αντικείμενο, καθώς και το cluster από το οποίο αποδίδεται το αντικείμενο. Η εκ νέου ανάθεση συνεχίζεται μέχρι η μεταβολή στα κομβικά σημεία να γίνει μικρότερη από το όριο του καθορισμένου κριτηρίου σύγκλισης.

3.4.2.1 Η μέθοδος k-means

Η μέθοδος k- Means προτάθηκε από τον MacQueen το 1967 και είναι η πιο γνωστή και διαδεδομένη διαιρετική μέθοδος ανάλυσης συστάδων. Στόχος της είναι να κατανείμει ένα σύνολο αντικειμένων σε έναν προκαθορισμένο αριθμό συστάδων, με τρόπο τέτοιο που να αυξάνει την ομοιότητα εντός των συστάδων. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας (centroid). Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο. Ο αλγόριθμος της μεθόδου αυτής έχει ως εξής:

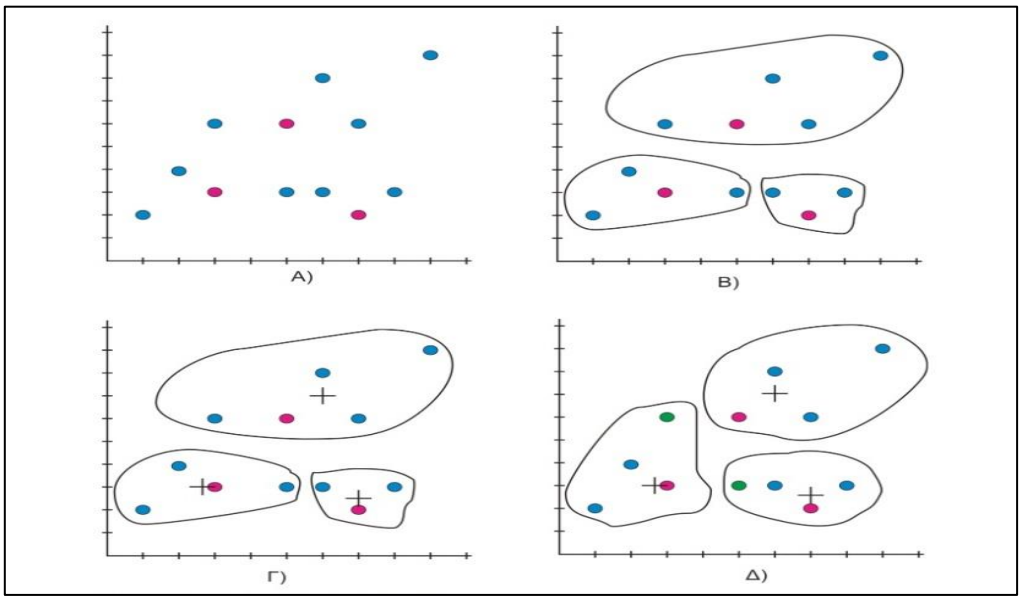
1) Αρχικά επιλέγονται τυχαία k αντικείμενα. Ο αριθμός k είναι το πλήθος των συστάδων που θα προκύψουν και προκαθορίζεται από τον χρήστη. Τα επιλεγμένα σημεία θεωρούνται τα κέντρα των συστάδων.
2) Κάθε αντικείμενο κατατάσσεται στη συστάδα, της οποίας το κέντρο είναι πλησιέστερα του. Για τον υπολογισμό της απόστασης συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση.
3) Τα κέντρα της κάθε συστάδας επαναϋπολογίζονται. Για κάθε διάσταση το κέντρο έχει τιμή ίση με τη μέση τιμή όλων των αντικειμένων, τα οποία ανήκουν στη συστάδα.
4) Τα προηγούμενα δύο βήματα επαναλαμβάνονται μέχρι να ικανοποιηθεί η συνθήκη εξόδου. Τυπικά, συνθήκη εξόδου είναι η ελαχιστοποίηση του

τετραγωνικού σφάλματος, όπως και στη μέθοδο του Ward, το οποίο ορίζεται από τη σχέση:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

Όπου C_i είναι οι συστάδες, x είναι οι παρατηρήσεις και m_i το κέντρο της συστάδας i .

Στο σχήμα που ακολουθεί, παρουσιάζεται ο σχηματισμός των συστάδων με τη μέθοδο k-Means. Στο τμήμα **Α)** παρουσιάζονται τα σημεία. Τα κόκκινα σημεία συμβολίζουν τα αρχικώς επιλεγμένα κέντρα. Στο τμήμα **Β)** σχηματίζονται οι συστάδες. Κάθε σημείο εντάσσεται στη συστάδα, στης οποίας το κέντρο βρίσκεται πλησιέστερα. Στο τμήμα **Γ)** υπολογίζονται τα νέα κέντρα των υφιστάμενων συστάδων. Τα νέα κέντρα συμβολίζονται με το σχήμα του σταυρού. Στο τμήμα **Δ)** επαναϋπολογίζεται η απόσταση των σημείων από τα νέα κέντρα, και τα σημεία επανεντάσσονται στις συστάδες. Τα δύο πράσινα σημεία αλλάζουν συστάδα.



***Εικόνα 9**: Διαδικασία δημιουργίας συστάδων με τη μέθοδο k-Means*

Ο αλγόριθμος k-Means διαθέτει τα παρακάτω πλεονεκτήματα:

- ✓ Είναι απλός και κατανοητός.
- ✓ Τα αντικείμενα μοιράζονται σε συστάδες με αυτόματο τρόπο.
- ✓ Είναι αρκετά γρήγορος, τουλάχιστον σε σχέση με τις ιεραρχικές μεθόδους. Ο χρόνος εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τα στοιχεία του προβλήματος, όπως το πλήθος των συστάδων k, το πλήθος των αντικειμένων n και το πλήθος των επαναλήψεων l.. Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι μικρότερη από άλλες μεθόδους και κυρίως τις ιεραρχικές.. Για τον λόγο αυτό, είναι πιο κατάλληλος από άλλες μεθόδους για την ομαδοποίηση μεγάλων συνόλων αντικειμένων.

Ωστόσο διακρίνεται και από τα παρακάτω μειονεκτήματα:

- Ο αριθμός των συστάδων πρέπει να προκαθοριστεί από τον χρήστη.
- Το τελικό αποτέλεσμα εξαρτάται σε σημαντικό βαθμό από την επιλογή των αρχικών κέντρων. Επιλογή διαφορετικών κέντρων μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες.
- Είναι πολύ ευαίσθητος στην ύπαρξη αντικειμένων με ακραίες τιμές. Λίγα αντικείμενα με πολύ μεγάλες τιμές μπορούν να επηρεάσουν σημαντικά τον υπολογισμό των νέων κέντρων και κατά συνέπεια τη διαμόρφωση των τελικών συστάδων.
- Έχει την τάση να δημιουργεί σφαιρικές και ίσου μεγέθους συστάδες. Για τον λόγο αυτό, δεν είναι κατάλληλος για συστάδες με περίπλοκα σχήματα ή με πολύ διαφορετικά μεγέθη.

3.4.2.2 Η μέθοδος k-Medoids

Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος k-Means είναι ευαίσθητος στην ύπαρξη εξαιρέσεων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η χρήση ως κέντρου, όχι ενός υπολογιζόμενου μέσου σημείου, αλλά ενός υπαρκτού σημείου δεδομένων. Ο αλγόριθμος k-Medoids ακολουθεί αυτήν την προσέγγιση. Μια από τις πρώτες εκδοχές του k-Medoids ήταν η μέθοδος Partitioning Around Medoids (Kaufman & Rousseeuw, 1990).Οι αλγόριθμοι k-Means και k-Medoids παρουσιάζουν αρκετές ομοιότητες:

- Αρχικά επιλέγονται αυθαίρετα τα κέντρα των συστάδων.
- Σε μια επαναληπτική διαδικασία τα κέντρα επαναπροσδιορίζονται.
- Σε κάθε επανάληψη μειώνεται το κριτήριο.
- Επιλογή διαφορετικών αρχικών κέντρων μπορεί να δώσει διαφορετικά αποτελέσματα.
- Δεν επιτυγχάνουν καθολικά βέλτιστα.

Αναλυτικότερα, στον αλγόριθμο k-Medoids επιλέγονται αρχικά k σημεία ως κέντρα (medoids). Τα υπόλοιπα σημεία κατατάσσονται στη συστάδα του πλησιέστερου κέντρου. Μια συνάρτηση κόστους μετρά το άθροισμα των αποστάσεων όλων των σημείων από το κέντρο της συστάδας τους. Σε μια επαναληπτική διαδικασία, σημεία τα οποία δεν είναι κέντρα δοκιμάζονται ως πιθανά κέντρα. Εάν για ένα σημείο το κόστος γίνεται μικρότερο, τότε το σημείο αυτό γίνεται το νέο κέντρο στη θέση του προηγούμενου.

Ο αλγόριθμος k-Medoids λειτουργεί πιο αποτελεσματικά από τον k-Means, όταν στα δεδομένα υπάρχουν αντικείμενα με ακραίες τιμές. Ωστόσο, το κόστος υπολογισμού των medoids είναι σημαντικά μεγαλύτερο από το κόστος υπολογισμού

των μέσων τιμών. Για τον λόγο αυτό, ο k-Medoids υπολείπεται του k-Means ως προς τον χρόνο επεξεργασίας μεγάλων συνόλων δεδομένων.

3.4.3 Άλλες μέθοδοι ομαδοποίησης

- **Μέθοδοι βασισμένες στην πυκνότητα**: Στις βασισμένες στην πυκνότητα μεθόδους (density based methods) ελέγχεται η πυκνότητα των αντικειμένων στον χώρο και δημιουργούνται συστάδες, οι οποίες καλύπτουν τις πυκνές περιοχές. Για κάθε παρατήρηση που ανήκει σε μια συστάδα, η γειτονιά της, η οποία είναι καθορισμένης διαμέτρου, πρέπει να περιλαμβάνει έναν ελάχιστο αριθμό παρατηρήσεων. Η συστάδα συνεχίζει να επεκτείνεται όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι αυτές μπορούν να δημιουργήσουν συστάδες με μη κυρτά και περίπλοκα σχήματα. Επίσης έχουν την ικανότητα να απομονώνουν τις εξαιρέσεις.

- **Μέθοδοι πλέγματος**: Οι μέθοδοι πλέγματος (grid based methods) επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα οποία συγκροτούν ένα πλέγμα. Τα αντικείμενα πλέον αντιπροσωπεύονται από τα κελιά στα οποία ανήκουν. Η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος και όχι στα αντικείμενα. Στις μεθόδους πλέγματος ο χρόνος επεξεργασίας εξαρτάται από το πλήθος των κελιών και όχι από το πλήθος των αντικειμένων. Επειδή κατά κανόνα ο αριθμός των κελιών είναι πολύ μικρότερος από τον αριθμό των αντικειμένων, οι μέθοδοι αυτές είναι σημαντικά ταχύτερες. Ένα σημαντικό ζήτημα είναι ο καθορισμός κελιών κατάλληλου μεγέθους.

- **Μέθοδοι βασισμένες σε μοντέλα**: Στις βασισμένες σε μοντέλα μεθόδους (model based methods), όπως υπονοεί το όνομα τους, γίνεται χρήση μοντέλων. Στόχος τους είναι να βελτιστοποιηθεί η προσαρμογή ανάμεσα στα δεδομένα και στα μοντέλα. Το μοντέλο εκπαιδεύεται με μη επιβλεπόμενη μάθηση σχετικά με τη συμμετοχή των παρατηρήσεων σε συστάδες. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι ένα ειδικός τύπος νευρωνικών δικτύων, που ονομάζονται Αυτοοργανούμενοι Χάρτες (Self Organizing Maps).

ΚΕΦΑΛΑΙΟ 4: ΕΠΙΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΜΑΣ

4.1 Περιγραφή Μελέτης

Για τη μελέτη που θέλουμε να υλοποιήσουμε, έχουμε στη διάθεση μας τις εβδομαδιαίες καταναλώσεις νερού από 67 νοικοκυριά της ισπανικής παραθαλάσσιας πόλης Alicante. Οι καταναλώσεις αυτές αφορούν το διάστημα Ιανουάριος 2015-Φεβρουάριος 2017. Επιπλέον, για όλα αυτά τα νοικοκυριά γνωρίζουμε ένα πλήθος δημογραφικών στοιχείων τα οποία θα χρησιμοποιήσουμε στην ανάλυση μας. Η μελέτη θα αποτελεί μία διερεύνηση της σχέσης των δημογραφικών χαρακτηριστικών των νοικοκυριών με την αντίστοιχη κατανάλωση που έχουν σε νερό και θα περιλαμβάνει όλα εκείνα τα βήματα που παρουσιάστηκαν στο Κεφάλαιο 2 τόσο για **τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης**, όσο και για τη **μέθοδο ανάλυσης συστάδων k-means**. Θα αναλύσουμε δηλαδή τα δεδομένα μας με 2 διαφορετικές εναλλακτικές μεθόδους και θα αξιολογήσουμε την ακρίβεια της καθεμίας.

Από το σύνολο των δημογραφικών που διαθέτουμε για τα νοικοκυριά που θα συμμετάσχουν στη μελέτη μας, θα λάβουμε υπόψη τα πιο αντικειμενικά, εκείνα δηλαδή που απαντώνται με ακρίβεια και δεν επιδέχονται προσωπικών εκτιμήσεων. Τα αυτά δημογραφικά χαρακτηριστικά με βάσει τα οποία θα επιχειρήσουμε να ομαδοποιήσουμε τα 67 νοικοκυριά είναι τα εξής:

- Μορφωτικό Επίπεδο Ιδιοκτήτη
- Πλήθος Ανηλίκων σε κάθε νοικοκυριό
- Μέγεθος Οικίας (σε m^2)
- Ηλικία Ιδιοκτήτη
- Αριθμός ατόμων νοικοκυριού
- Αριθμός θηλυκών μελών σε κάθε νοικοκυριό
- Ετήσιο εισόδημα νοικοκυριού (προ φόρων)
- Ιδιοκατοίκηση ή ενοικίαση οικίας

Παρότι έχουμε στη διάθεση μας και άλλες πληροφορίες σχετικά με την καθημερινότητα των νοικοκυριών, όπως πόσο συχνά χρησιμοποιούν το πλυντήριο ή πλένουν το αμάξι κτλ, αποφασίσαμε να μην τα λάβουμε υπόψη μας, γιατί οι απαντήσεις μπορεί να είναι κατά προσέγγιση και να επηρεάσουν τη μελέτη μας. Έτσι όπως είπαμε κρατήσαμε τα 8 πιο ‘‘αντικειμενικά χαρακτηριστικά’’ για κάθε νοικοκυριό. Παρόλα αυτά, όπως αναφέραμε και στο Κεφάλαιο 2, υπάρχει η περίπτωση βάσει των βημάτων που θα ακολουθήσουμε και που θα περιγραφούν

αναλυτικά στην πορεία, κάποιο ή κάποια από αυτά τα χαρακτηριστικά που κρατήσαμε να αποδειχθεί είτε πως έχει πολύ μικρή επιρροή στη μοντελοποίηση είτε πως επηρεάζει αρνητικά το μοντέλο και πρέπει να αγνοηθεί, επομένως να μη ληφθεί/ληφθούν υπόψη στη μελέτη μας.

4.2 Επεξεργασία Δεδομένων

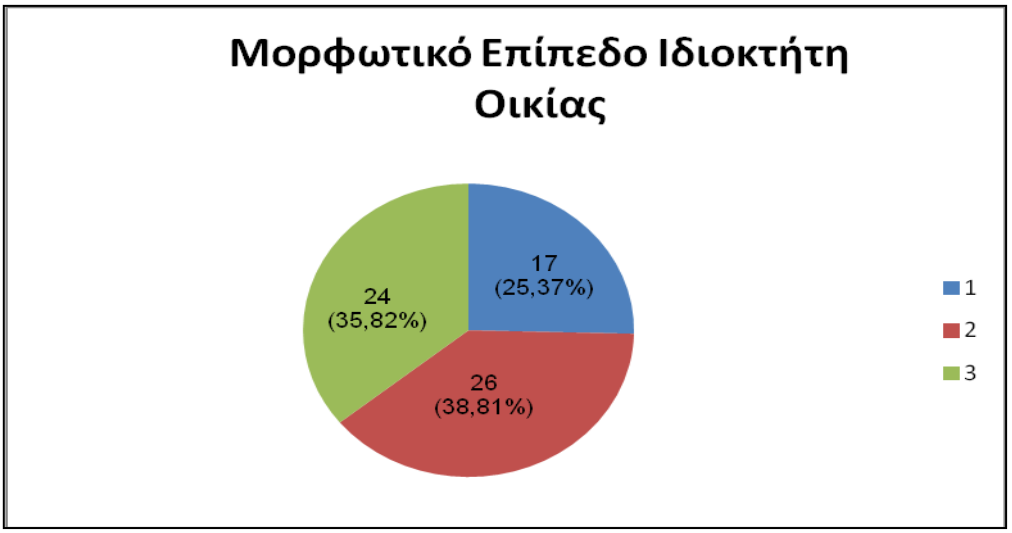
Στο σημείο αυτό, θα παρουσιάσουμε αναλυτικά την κατανομή των 67 νοικοκυριών με βάση τα δημογραφικά χαρακτηριστικά τους.

1) Μορφωτικό Επίπεδο Ιδιοκτήτη Οικίας

Τα νοικοκυριά χωρίστηκαν σε 3 Ομάδες με βάση το μορφωτικό επίπεδο του ιδιοκτήτη της οικίας, ο οποίος έδωσε και τις πληροφορίες για το νοικοκυριό:

- Ομάδα 1: Απόφοιτος Πρωτοβάθμιας ή Δευτεροβάθμιας Εκπαίδευσης (ουσιαστικά με μέγιστο επίπεδο μόρφωσης το απολυτήριο Λυκείου) Άτομα
- Ομάδα 2: Απόφοιτος Ανώτερης Εκπαίδευσης
- Ομάδα 3: Κάτοχος Μεταπτυχιακού τίτλου ή Διδακτορικού

Η κατανομή των ιδιοκτητών των οικιών με βάση το επίπεδο μόρφωσης φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα
:



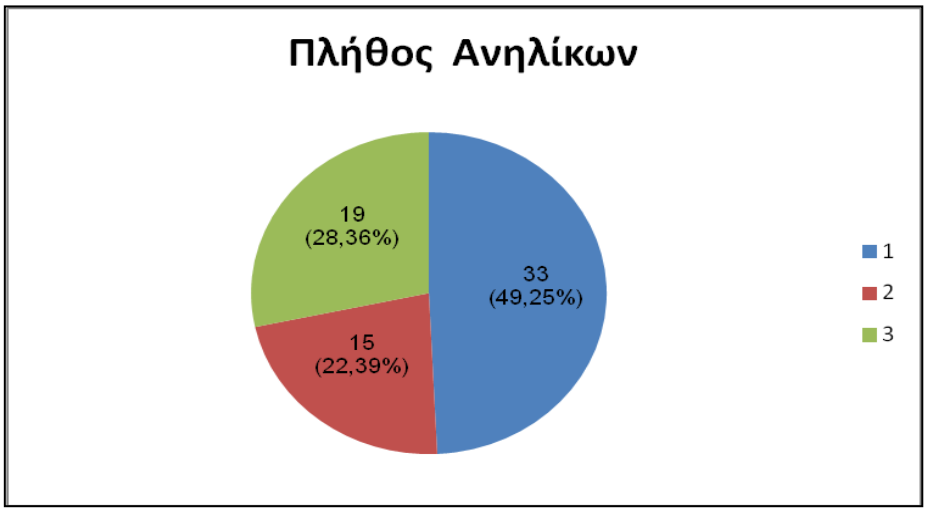
***Εικόνα 10**: Κατανομή νοικοκυριών βάσει επιπέδου μόρφωσης*

2) Πλήθος Ανηλίκων σε κάθε νοικοκυριό

Τα νοικοκυριά χωρίστηκαν σε 3 Ομάδες με βάση τον αριθμό των ανηλίκων:

- Ομάδα 1: Κανένας Ανήλικος
- Ομάδα 2: 1 Ανήλικος
- Ομάδα 3: 2 ή 3 Ανήλικοι

Η κατανομή των νοικοκυριών με βάση τον αριθμό των ενηλίκων σε αυτό φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



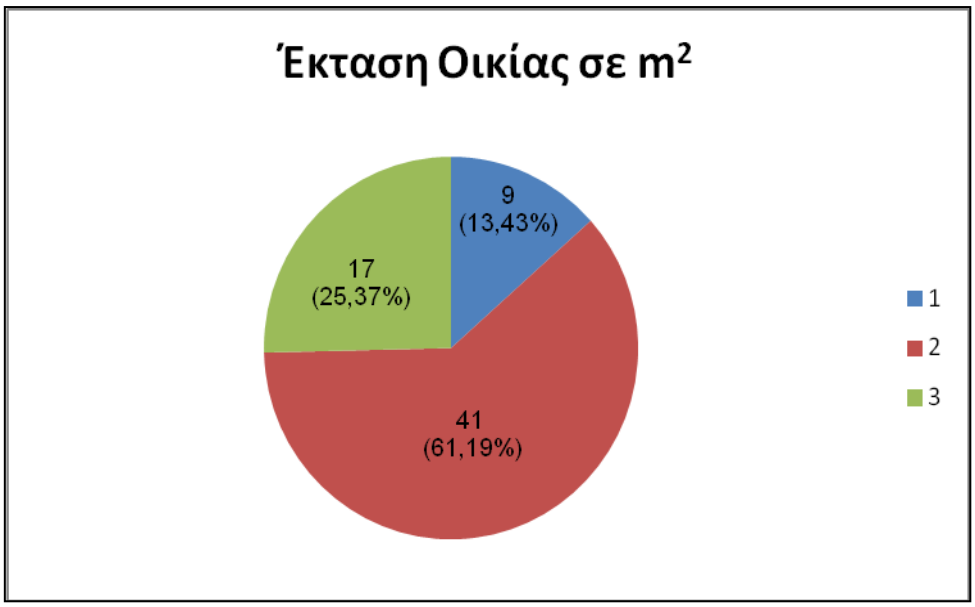
***Εικόνα 11**: Κατανομή νοικοκυριών βάσει του αριθμού των ενηλίκων μέσα σε αυτό*

3) Μέγεθος Οικίας σε Τετραγωνικά Μέτρα:

Τα νοικοκυριά χωρίστηκαν σε 3 Ομάδες με βάση το μέγεθος της οικίας:

- Ομάδα 1: 61-80 m^2
- Ομάδα 2: 81-110 m^2
- Ομάδα 3: Πάνω από 110 m^2

Η κατανομή των νοικοκυριών βάσει της έκτασης της οικίας, φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



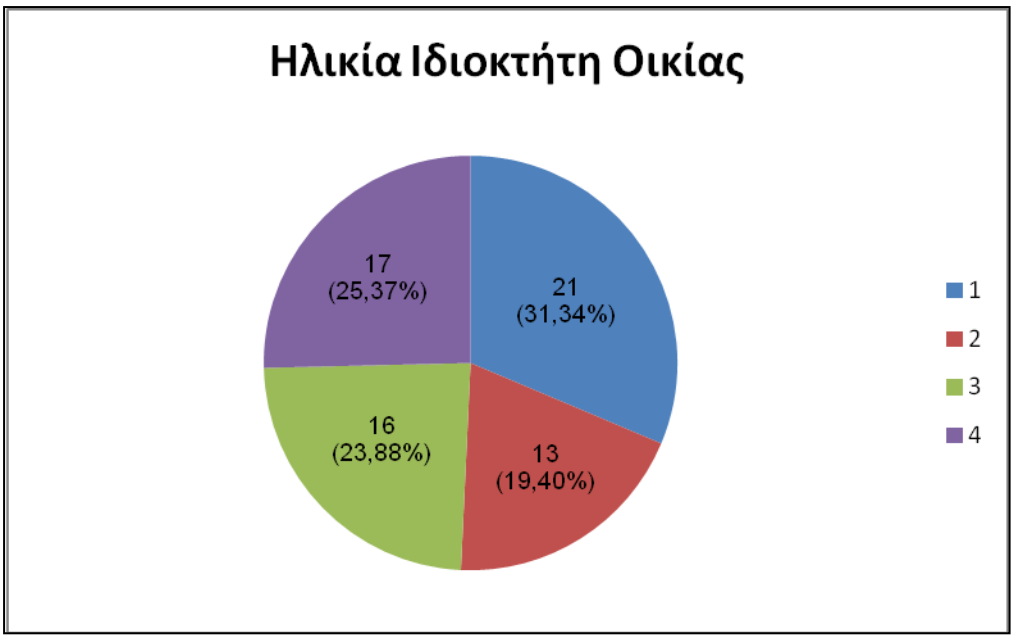
***Εικόνα 12**: Κατανομή νοικοκυριών βάσει έκτασης της οικίας*

4) Ηλικία Ιδιοκτήτη Οικίας:

Τα νοικοκυριά χωρίστηκαν σε 4 Ομάδες με βάση την ηλικία του ιδιοκτήτη της οικίας, ο οποίος έδωσε και τις πληροφορίες για το νοικοκυριό. Καθώς, όπως είναι λογικό, οι ηλικίες των ερωτηθέντων είχαν μεγάλη διασπορά, υπολογίσαμε τα όρια κάθε Ομάδας με τη βοήθεια της Εντολής ‘‘Quartile’’ στο Excel, η οποία ουσιαστικά χώρισε το δείγμα μας σε 4 ομάδες και μας έδειξε το όριο ηλικίας για κάθε ομάδα. Έτσι οι ομάδες που προέκυψαν είναι οι εξής:

- Ομάδα 1: Ηλικία ιδιοκτήτη έως 37 ετών
- Ομάδα 2: Ηλικία ιδιοκτήτη έως 43 ετών
- Ομάδα 3: Ηλικία ιδιοκτήτη έως 51 ετών
- Ομάδα 4: Ηλικία ιδιοκτήτη 51 -71 ετών

Η κατανομή των νοικοκυριών βάσει της ηλικίας του ιδιοκτήτη της οικίας, ο οποίος μας έδωσε και τις πληροφορίες, φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



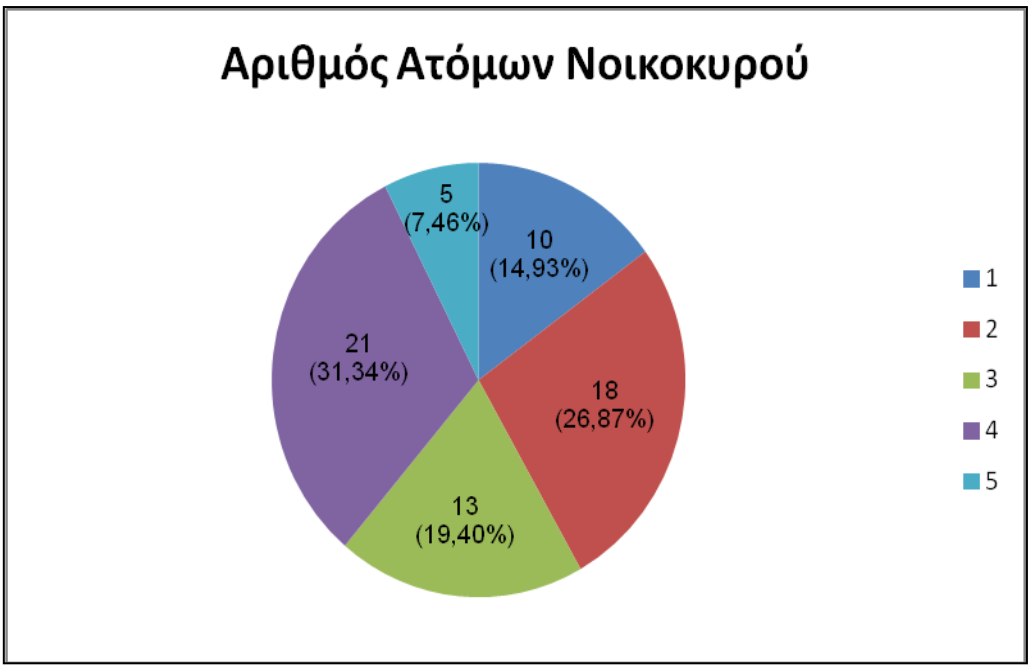
***Εικόνα 13**: Κατανομή νοικοκυριών βάσει ηλικίας του ιδιοκτήτη*

5) Συνολικός Αριθμός Ατόμων Νοικοκυριού:

Τα νοικοκυριά χωρίστηκαν σε 5 Ομάδες με βάση τον συνολικό αριθμό των ατόμων κάθε νοικοκυριού.

- Ομάδα 1: 1 Άτομο
- Ομάδα 2: 2 Άτομα
- Ομάδα 3: 3 Άτομα
- Ομάδα 4: 4 Άτομα
- Ομάδα 5: 5 Άτομα

Η κατανομή των νοικοκυριών με βάση τον αριθμό των ατόμων, φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



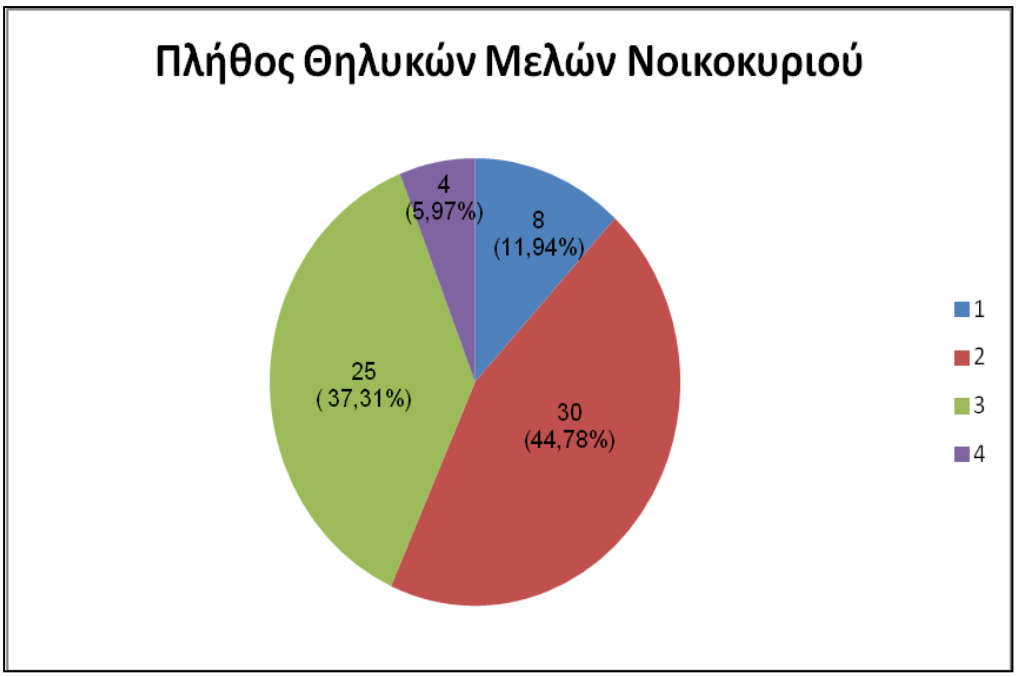
***Εικόνα 14**: Κατανομή νοικοκυριών βάσει του συνολικού ατόμων σε αυτό*

6) Αριθμός Θηλυκών Ατόμων Νοικοκυριού:

Τα νοικοκυριά χωρίστηκαν σε 4 Ομάδες με βάση τον αριθμό των θηλυκών μελών κάθε νοικοκυριού

- Ομάδα 1: Κανένα Θηλυκό Μέλος
- Ομάδα 2: 1 Θηλυκό Μέλος
- Ομάδα 3: 2 Θηλυκά Μέλη
- Ομάδα 4: 3 Θηλυκά Μέλη

Η κατανομή των νοικοκυριών βάσει τον αριθμό των θηλυκών μελών του, φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



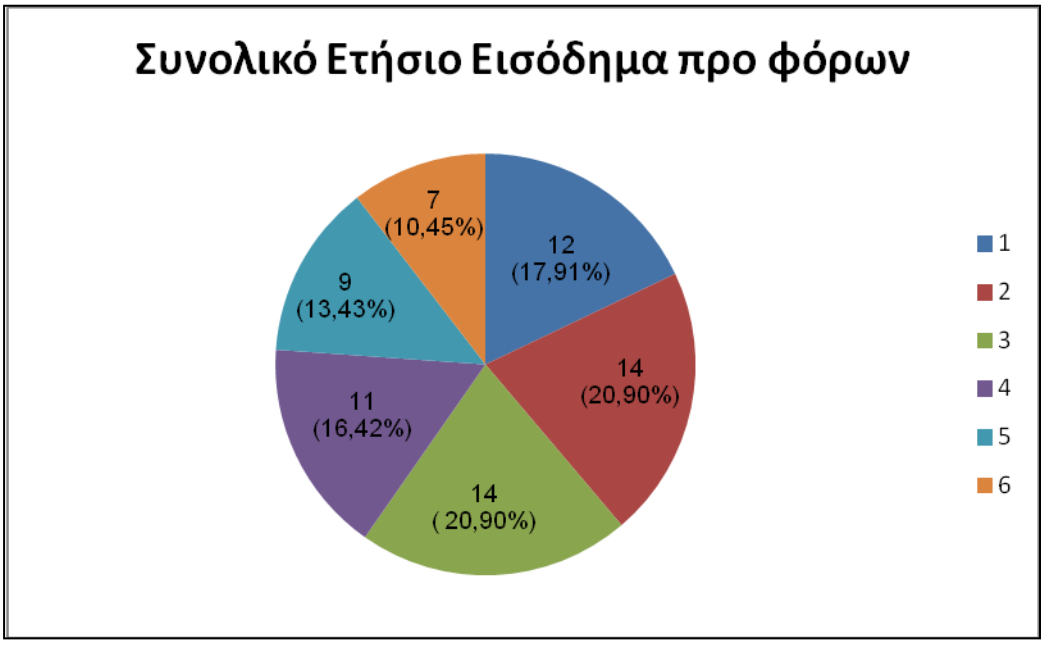
Εικόνα 15: *Κατανομή νοικοκυριών βάσει του αριθμού των θηλυκών ατόμων σε αυτό*

7) Συνολικό Ετήσιο Εισόδημα Νοικοκυριού προ φόρων:

Τα νοικοκυριά χωρίστηκαν σε 6 Ομάδες με βάση το συνολικό ετήσιο εισόδημα προ φόρων κάθε νοικοκυριού.

- Ομάδα 1: Έως 20.000 €
- Ομάδα 2: 20.000-25.000 €
- Ομάδα 3: 30.000-40.000 €
- Ομάδα 4: 40.000-50.000 €
- Ομάδα 5: 50.000-60.000 €
- Ομάδα 6: Πάνω από 60.000 €

Η κατανομή των νοικοκυριών βάσει το συνολικό ετήσιο εισόδημα προ φόρων κάθε νοικοκυριού, φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



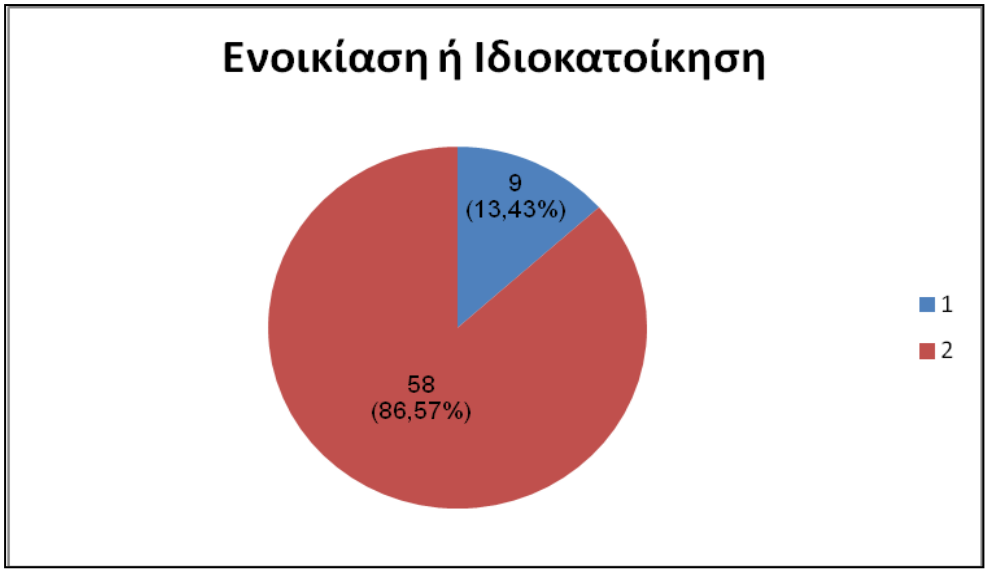
***Εικόνα 16**: Κατανομή νοικοκυριών βάσει ετήσιου εισοδήματος*

8) Ιδιοκατοίκηση ή ενοικίαση οικίας:

Τέλος, τα νοικοκυριά χωρίστηκαν σε 2 Ομάδες, ανάλογα με το αν τα άτομα ενοικιάζουν την οικία στην οποία ζουν ή την έχουν στην ιδιοκτησία τους.

- Ομάδα 1: Ενοικίαση
- Ομάδα 2: Ιδιοκατοίκηση

Η κατανομή των νοικοκυριών φαίνεται στο παρακάτω σχεδιάγραμμα, με κάθε Ομάδα να έχει αντίστοιχο χρώμα:



***Εικόνα 17**: Κατανομή νοικοκυριών βάσει ιδιοκατοίκησης ή ενοικίασης*

Όπως φαίνεται από την επιλογή και την επεξεργασία των δεδομένων που για τα 67 νοικοκυριά, όλα αφορούν ποιοτικά χαρακτηριστικά και όχι ποσοτικά, εφόσον κάποια από αυτά δεν είναι αριθμητικά χαρακτηριστικά, ενώ όσα από αυτά είναι αριθμητικά (π.χ. ετήσιο εισόδημα) είναι κατανεμημένα σε ομάδες και δε δίνεται η ακριβής τιμή τους. Αυτό σημαίνει πως εφόσον αυτά τα ποιοτικά χαρακτηριστικά δεν έχουν αριθμητικές τιμές θα πρέπει εμείς να δώσουμε κατάλληλες τέτοιες τιμές στα δεδομένα του δείγματος μας ώστε να προχωρήσουμε στην ανάλυση τόσο της πολλαπλής γραμμικής παλινδρόμησης όσο και στην ανάλυση συστάδων

ΚΕΦΑΛΑΙΟ 5: ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ ΜΕ ΤΗ ΜΕΘΟΔΟ ΠΟΛΛΑΠΛΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

5.1 Εισαγωγή

Η **γραμμική παλινδρόμηση** όπως έχουμε πει, είναι μια μαθηματική μέθοδος για τον προσδιορισμό της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής Y και μίας ή περισσότερων ανεξάρτητων μεταβλητών x ή $x_1...x_n$. Στην περίπτωση μιας ανεξάρτητης μεταβλητής, η μέθοδος ονομάζεται απλή γραμμική παλινδρόμηση. Για περισσότερες από μία ανεξάρτητες μεταβλητές (όπως στη δική μας περίπτωση), η διαδικασία ονομάζεται πολλαπλή γραμμική παλινδρόμηση.

Στην πολλαπλή γραμμική παλινδρόμηση, η οποία και θα μας απασχολήσει, επιχειρούμε να βρούμε μία γραμμική σχέση ανάμεσα στην εξαρτημένη μεταβλητή Y και τις ανεξάρτητες μεταβλητές $x_1 \ldots x_n$ που θα το απαρτίζουν. Το μοντέλο που θα υπακούει στη σχέση αυτή, θα έχει τη μορφή:

$$Yi = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n$$

Όπου

- α είναι μία σταθερά, η οποία αντιστοιχεί στην τιμή του Y για $x_1=x_2=\ldots=x_n=0$
- το β_1 είναι μία σταθερά που δείχνει τη μεταβολή της τιμής Y όταν το x_1 μεταβάλλεται κατά μία μονάδα, ενώ τα $x_2 \ldots x_n$ παραμένουν σταθερά,
- το β_2 είναι μία σταθερά που δείχνει τη μεταβολή της τιμής Y όταν το X_2 μεταβάλλεται κατά μία μονάδα, ενώ τα $x_1 \ldots x_n$ παραμένουν σταθερά,
- και κατ' επέκταση το β_n είναι σταθερά που δείχνει τη μεταβολή της τιμής Y όταν το x_n μεταβάλλεται κατά μία μονάδα, ενώ τα $x_1 \ldots x_{n-1}$ παραμένουν σταθερά.

Το μοντέλο πολλαπλής γραμμικής παλινδρόμησης μας βοηθάει στην πρόβλεψη τις τιμής που αντιστοιχεί στο Y, δηλαδή εφόσον κατασκευάσουμε το μοντέλο μας, μπορούμε να προβλέψουμε για μελλοντικές τιμές των $x_1 \ldots x_n$ τι τιμές θα πάρει το Y. Μπορούμε επίσης εφόσον έχουμε κατασκευάσει το μοντέλο μας για ένα δείγμα πληθυσμού (μίας πόλης, ενός σχολείου κτλ) να εφαρμόσουμε το μοντέλο μας σε ένα

άλλο αντίστοιχο δείγμα πληθυσμού, δηλαδή να προβλέψουμε τις τιμές του Y εφόσον θα γνωρίζουμε τις τιμές των $x_1 \ldots x_n$ για το άλλο δείγμα πληθυσμού.

5.2 Ύπαρξη Ποιοτικών Ανεξάρτητων Μεταβλητών

Στη θεωρητική ανάλυση της γραμμικής παλινδρόμησης, είτε απλής είτε της πολλαπλής εξετάσαμε μία περίπτωση στην οποία έχουμε ανεξάρτητες μεταβλητές που μπορούν να πάρουν οποιαδήποτε αριθμητική τιμή. Σε ορισμένες περιπτώσεις όμως μπορεί να κληθούμε να διερευνήσουμε ένα πρόβλημα στο οποίο κάποιες από τις ανεξάρτητες μεταβλητές δε θα είναι ποσοτικές αλλά ποιοτικές.

Τέτοιες περιπτώσεις μπορεί να συναντήσουμε όταν έχουμε στη διάθεση μας τα χαρακτηριστικά για ένα δείγμα πληθυσμού και θέλουμε να δούμε αν αυτά σχετίζονται γραμμικά με την τιμή μίας ανεξάρτητης μεταβλητής Y. Τα χαρακτηριστικά που διαθέτουμε θα πρέπει να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές, ωστόσο μπορεί να μην είναι ποσοτικά. Παραδείγματα τέτοιων ποιοτικών χαρακτηριστικών μπορεί να είναι τα εξής:

-Εθνικότητα

-Επίπεδο Εκπαίδευσης

-Θρησκεία

-Φύλο

-Κάποιο κατηγορία ή ομάδα

Για τα 2 πρώτα παραδείγματα ποιοτικών χαρακτηριστικών, είναι προφανές ότι αυτά δεν μπορούν να εκφραστούν με αριθμούς. Μπορούμε ωστόσο να φτιάξουμε ομάδες με αύξων αριθμό, κάθε μια από τις οποίες θα αντιπροσωπεύει το ποιοτικό χαρακτηριστικό για την αντίστοιχη ανεξάρτητη μεταβλητή. Για παράδειγμα σε ό,τι αφορά το επίπεδο μόρφωσης, μπορούμε να πούμε ότι η ανεξάρτητη μεταβλητή λαμβάνει την τιμή 1 εάν κάποιος από το δείγμα μας είναι απόφοιτος πρωτοβάθμιας εκπαίδευσης, την τιμή 2 εάν είναι απόφοιτος δευτεροβάθμιας εκπαίδευσης, την τιμή τρία εάν είναι απόφοιτος τριτοβάθμιας εκπαίδευσης και την τιμή 4 εάν έχει μεταπτυχιακές σπουδές ή είναι κάτοχος διδακτορικού διπλώματος. Για μία τέτοια μαθηματική έκφραση φυσικά δεν υπάρχει μπορεί μοναδικός τρόπος, αλλά είναι στην ευχέρεια του αναλυτή.

Ομοίως, μπορεί να έχουμε στη διάθεση μας στοιχεία π.χ. για την οικονομική κατάσταση των ατόμων του δείγματος μας χωρίς να γνωρίζουμε ακριβή ποσά ετήσιου εισοδήματος, αλλά τα δεδομένα να είναι ανά κατηγορίες, π.χ. $1^η$ κατηγορία έως 10.000 € /έτος, $2^η$ κατηγορία 10.000 -20.000 €/έτος κτλ). Σε αυτή την περίπτωση θα πρέπει να βρούμε πόσες διαφορετικές κατηγορίες έχουμε και να δώσουμε στην

καθεμία έναν αύξων αριθμό, ο οποίος θα είναι και η τιμή της αντίστοιχης ανεξάρτητης μεταβλητής.

Τέλος υπάρχουν και ποιοτικά δεδομένα όπως π.χ. το φύλο ή η οικογενειακή κατάσταση, τα οποία μπορούν να εκφραστούν μόνο με δυαδικό τρόπο. Η αντίστοιχη μεταβλητή μπορεί έτσι να πάρει μόνο δύο τιμές, όπως φαίνεται στην ακόλουθη σχέση:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Στο πρόβλημα που θα εξετάσουμε στο Κεφάλαιο αυτό, έχουμε διάφορα ποιοτικά χαρακτηριστικά για τα νοικοκυριά τα οποία θα κατατάξουμε σε κατηγορίες ώστε να δώσουμε τιμές στις αντίστοιχες ανεξάρτητες μεταβλητές και να προχωρήσουμε στη διεξαγωγή πολλαπλής γραμμικής παλινδρόμησης.

5.3 Προετοιμασία των δεδομένων μας με σκοπό την εισαγωγή τους στη Γραμμική Παλινδρόμηση

Στην ανάλυση, οι ανεξάρτητες μεταβλητές $x_1 \ldots x_8$ θα αντιπροσωπεύουν ποσοτικά χαρακτηριστικά για τις ομάδες στις οποίες έχουμε χωρίσει τα νοικοκυριά και οι οποίες παρουσιάστηκαν στο προηγούμενο Κεφάλαιο. Κάθε μεταβλητή θα παίρνει την τιμή της Ομάδας στην οποία ανήκει το αντίστοιχο νοικοκυριό.

Όπως είδαμε στο Κεφάλαιο 3, οι ανεξάρτητες μεταβλητές μας $x_1 \ldots x_8$ θα λαμβάνουν τις εξής τιμές, ανάλογα με τον αριθμό των ομάδων στον οποίο χωρίστηκαν τα νοικοκυριά για κάθε δημογραφικό χρατκηριστικό:

- **x_1** →Μορφωτικό Επίπεδο Ιδιοκτήτη (Λαμβάνει Τιμές από 1-3)
- **x_2** →Πλήθος Ανηλίκων σε κάθε νοικοκυριό (Λαμβάνει Τιμές από 1-3)
- **x_3**→Μέγεθος Οικίας (σε m^2) (Λαμβάνει Τιμές από 1-3)
- **x_4**→Ηλικία Ιδιοκτήτη (Λαμβάνει Τιμές από 1-4)
- **x_5** →Αριθμός ατόμων νοικοκυριού (Λαμβάνει Τιμές από 1-5)
- **x_6** →Αριθμός θηλυκών μελών σε κάθε νοικοκυριό (Λαμβάνει Τιμές από 1-4)
- **x_7** →Ετήσιο εισόδημα νοικοκυριού (προ φόρων) (Λαμβάνει Τιμές από 1-6)
- **x_8** →Ιδιοκατοίκηση ή ενοικίαση οικίας (Λαμβάνει τιμές 1-2)

Στον παρακάτω Πίνακα φαίνεται ένα νοικοκυριό του δείγματός μας και οι αντίστοιχες τιμές που λαμβάνουν οι ανεξάρτητες μεταβλητές $x_1 \ldots x_8$, ανάλογα με τις απαντήσεις που έχει δώσει ο ιδιοκτήτης:

SWM ID	What is the highest degree or level of school you have completed?	X1	How many minors live in your household (i.e. of age less than 18 years old)?	X2	What is the size of your appartment/house in square meters/feet?	X3	What is your age?	X4	How many members are there in your household?	X5	How many of your household members are females?	X6	What is approximately your yearly household income before tax?	X7	Do you own or lease your residence?	X8
C13UA198207D	University Degree	2	0	1	81 - 110 square meters	2	55	4	3	3	2	3	20.000€ - 25.000€	2	Own	2

Πίνακας 2*: Πίνακας με τα ποιοτικά χαρακτηριστικά ενός νοικοκυριού (ανεξάρτητες μεταβλητές) επεξεργασμένα ώστε να εισαχθούν στη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης. Με τον ίδιο τρόπο έχουν επεξεργαστεί όλες οι γραμμές του Πίνακα*

Η τιμή της μεταβλητής Y αντιπροσωπεύει τη μέση μηνιαία κατανάλωση νερού για το νοικοκυριό. Ωστόσο, για κάθε νοικοκυριό, έχουμε 12 τιμές για το Y σε ένα έτος. Οι ανεξάρτητες μεταβλητές όμως $x_1 \ldots x_8$ παραμένουν σταθερές, εφόσον αντιπροσωπεύουν δημογραφικά χαρακτηριστικά, τα οποία δε μεταβάλλονται στη διάρκεια του έτους για ένα νοικοκυριό.

Με βάση τα παραπάνω, εφόσον διαθέτουμε την πληροφορία μέσης μηνιαίας κατανάλωσης για ένα νοικοκυριό για διάστημα 26 μηνών (Ιανουάριος 2015-Φεβρουάριος 2017), κάθε νοικοκυριό στο συνολικό δείγμα μας αποτελείται από 26 παρατηρήσεις (γραμμές), στις οποίες οι ανεξάρτητες μεταβλητές $x_1 \ldots x_8$ παραμένουν σταθερές σε όλες τις γραμμές, ενώ η εξαρτημένη μεταβλητή Y είναι προφανώς διαφορετική σε κάθε γραμμή.

Για το νοικοκυριό με κωδικό μετρητή κατανάλωσης C13UA198207D που δείξαμε στον προηγούμενο Πίνακα, το πλήρες δείγμα για το διάστημα 1/1/2015 έως 28/2/2017 θα έχει ως εξής:

USER ID	X1	X2	X3	X4	X5	X6	X7	X8	Μήνας	Y
C13UA198207D	2	1	2	4	3	3	2	2	1	10200
C13UA198207D	2	1	2	4	3	3	2	2	2	9660
C13UA198207D	2	1	2	4	3	3	2	2	3	9397
C13UA198207D	2	1	2	4	3	3	2	2	4	8142
C13UA198207D	2	1	2	4	3	3	2	2	5	7844
C13UA198207D	2	1	2	4	3	3	2	2	6	5780
C13UA198207D	2	1	2	4	3	3	2	2	7	8513
C13UA198207D	2	1	2	4	3	3	2	2	8	7510
C13UA198207D	2	1	2	4	3	3	2	2	9	5258
C13UA198207D	2	1	2	4	3	3	2	2	10	7056
C13UA198207D	2	1	2	4	3	3	2	2	11	7128
C13UA198207D	2	1	2	4	3	3	2	2	12	8767
C13UA198207D	2	1	2	4	3	3	2	2	1	9546
C13UA198207D	2	1	2	4	3	3	2	2	2	7429
C13UA198207D	2	1	2	4	3	3	2	2	3	7968
C13UA198207D	2	1	2	4	3	3	2	2	4	7493
C13UA198207D	2	1	2	4	3	3	2	2	5	7746
C13UA198207D	2	1	2	4	3	3	2	2	6	6448
C13UA198207D	2	1	2	4	3	3	2	2	7	7012
C13UA198207D	2	1	2	4	3	3	2	2	8	5401
C13UA198207D	2	1	2	4	3	3	2	2	9	6115
C13UA198207D	2	1	2	4	3	3	2	2	10	8209
C13UA198207D	2	1	2	4	3	3	2	2	11	8328
C13UA198207D	2	1	2	4	3	3	2	2	12	7198
C13UA198207D	2	1	2	4	3	3	2	2	1	7870
C13UA198207D	2	1	2	4	3	3	2	2	2	6885

***Πίνακας 3**: Πίνακας με τα ποιοτικά χαρακτηριστικά ενός νοικοκυριού (ανεξάρτητες μεταβλητές) επεξεργασμένα ώστε να εισαχθούν στη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης. Με τον ίδιο τρόπο έχουν επεξεργαστεί όλες οι γραμμές του Πίνακα*

Κατά παρόμοιο τρόπο αναπτύσσονται και οι Πίνακες για τα υπόλοιπα νοικοκυριά.

Στο σημείο αυτό φαίνεται ότι εισάγεται στη μελέτη μας άλλη μία ανεξάρτητη μεταβλητή, η οποία είναι ο αύξων αριθμός του μήνα και η οποία θα αντιπροσωπευτεί με την ανεξάρτητη μεταβλητή X_9. Για παράδειγμα, η σχέση που δείχνει την εξάρτηση ανάμεσα σε Y και $X_1 \ldots X_8$ για το νοικοκυριό με κωδικό μετρητή κατανάλωσης C13UA198207D για τον Ιούνιο του 2015 είναι η εξής:

$5780 = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \beta_6 \cdot x_6 + \beta_7 \cdot x_7 + \beta_8 \cdot x_8 + \beta_9 \cdot x_9 \rightarrow$

$5780 = \alpha + \beta_1 \cdot 2 + \beta_2 \cdot 1 + \beta_3 \cdot 2 + \beta_4 \cdot 4 + \beta_5 3 + \beta_6 \cdot 3 + \beta_7 \cdot 2 + \beta_8 \cdot 2 + \beta_9 \cdot 6$

Προφανώς για κάθε μήνα, οι τιμές που θα αλλάζουν θα είναι η μέση μηνιαία κατανάλωση Y και ο Α/Α του μήνα (x_9).

Εάν ασχολούμασταν με το συγκεκριμένο νοικοκυριό, θα μπορούσαμε να εξάγουμε ένα γραμμικό μοντέλο με δεδομένα τις 26 γραμμές που παρουσιάστηκαν παραπάνω. Κατ' επέκταση, προκειμένου να διερευνήσουμε τη σχέση ανάμεσα στην εξαρτημένη μεταβλητή Y και τις ανεξάρτητες μεταβλητές $x_1 \ldots x_9$ για όλο το δείγμα μας, ο στόχος μας είναι να υπολογίσουμε με τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης ένα γραμμικό μοντέλο, δηλαδή τους συντελεστές $\alpha, \beta_1, .., \beta_9$, οι οποίοι θα εξαχθούν έχοντας λάβει υπόψη όλα νοικοκυριά.

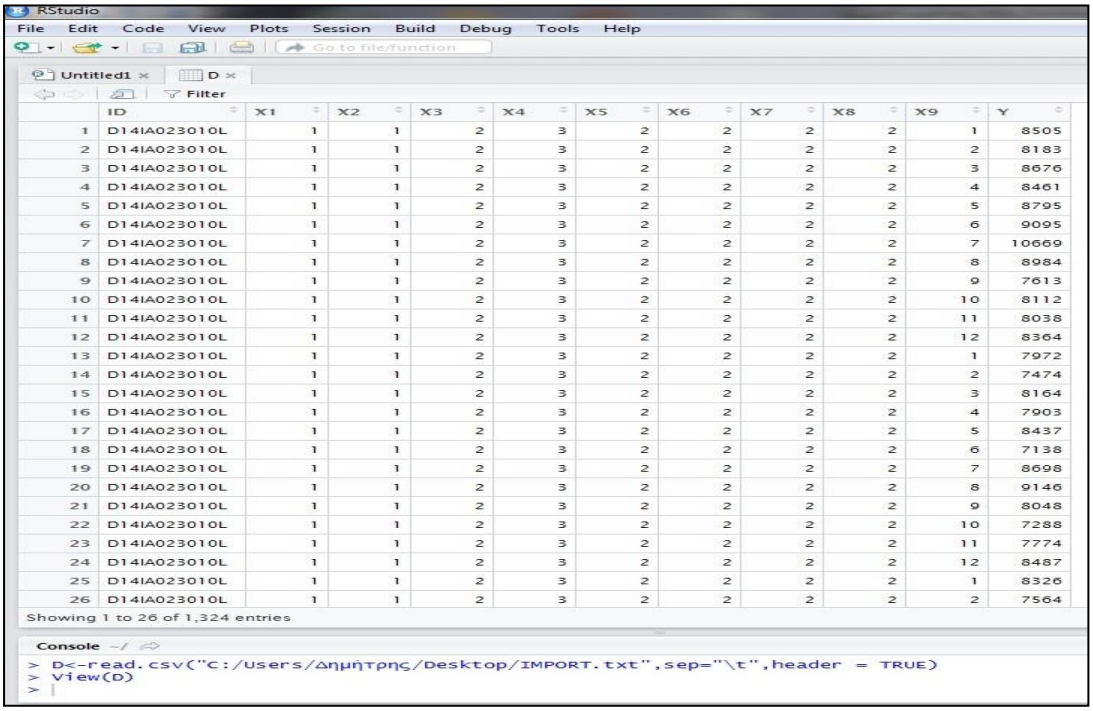
5.4 Παρουσίαση Μεθόδου Πολλαπλής Γραμμικής Παλινδρόμησης με τη Βοήθεια της R

Σε πρώτη φάση πρέπει να εισάγουμε το αρχείο των μέσω μηνιαίων καταναλώσεων για όλα τα νοικοκυριά σε μορφή CSV ή txt.

Σημείωση: Προκειμένου να διερευνήσουμε την ακρίβεια του μοντέλου που θα διερευνήσουμε, θα πρέπει να κρατήσουμε κάποια νοικοκυριά εκτός, τα οποία στη συνέχεια θα τα εισάγουμε στο μοντέλο μας ώστε να δούμε κατά πόσο οι τιμές των καταναλώσεων που προβλέψαμε έχουν σχέση με τις πραγματικές τιμές των καταναλώσεων. Επομένως θα κρατήσουμε εκτός ένα 15% του συνολικού αριθμού των νοικοκυριών (10 νοικοκυριά) τα οποία δε θα συμμετέχουν στην εξαγωγή του γραμμικού μοντέλου μας, αλλά θα χρησιμεύσουν στην πορεία στον έλεγχο της ακρίβειας του μοντέλου. Το γραμμικό μοντέλο επομένως θα εξαχθεί με τη βοήθεια της R με εισαγωγή των καταναλώσεων 57 νοικοκυριών.

5.4.1 Εισαγωγή Αρχείου

Εισάγουμε το αρχείο txt με τους κωδικούς μετρητών των καταναλωτών και τις αντίστοιχες μέσες μηνιαίες καταναλώσεις. Οι αντίστοιχες εντολές, καθώς και τα αποτελέσματα στην R φαίνονται στην παρακάτω εικόνα:



***Εικόνα 18**: Αποτέλεσμα στην R, μετά την εισαγωγή του αρχείου με τα επεξεργασμένα δεδομένα*

Όπως φαίνεται, τα δεδομένα έχουν εισαχθεί στη μορφή που αναλύσαμε παραπάνω, δηλαδή για έναν νοικοκυριό υπάρχουν 26 τιμές στο αρχείο. Οι ανεξάρτητες μεταβλητές $x_1 \ldots x_8$ παραμένουν σταθερές σε όλες τις γραμμές, εφόσον αποτελούν τα ποιοτικά χαρακτηριστικά του νοικοκυριού. Η ανεξάρτητη μεταβλητή x_9 εκφράζει τον Α/Α του κάθε μήνα και η εξαρτημένη μεταβλητή Y δείχνει την αντίστοιχη μέση μηνιαία κατανάλωση για το νοικοκυριό.

5.4.2 Εξαγωγή Γραμμικού Μοντέλου

Επόμενο βήμα είναι η εξαγωγή του γραμμικού μοντέλου με τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης. Οι εντολές και τα αντίστοιχα αποτελέσματα παρουσιάζονται στην παρακάτω εικόνα:

```
> D<-read.csv("C:/Users/Δημήτρης/Desktop/IMPORT.txt",sep="\t",header = TRUE)
> View(D)
>
> Model<-lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9,data=D)
>
> Model

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
    data = D)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5           X6
     641.02      -229.71     -1593.58       957.18       351.88      1991.74       509.78
         X7           X8           X9
     582.54      -707.83        20.29

>
```

***Εικόνα 19**: Εξαγωγή γραμμικού μοντέλου, με σταθερά και 9 ανεξάρτητες μεταβλητές*

Δημιουργήσαμε ένα Γραμμικό μοντέλο με την ονομασία ‘‘Model’’ το οποίο εξήχθη λαμβάνοντας υπόψη τις ανεξάρτητες μεταβλητές $x_1 \ldots x_9$ και την εξαρτημένη τιμή Y των 57 νοικοκυριών συνολικά. Στην εικόνα φαίνονται ουσιαστικά οι συντελεστές α και $\beta_1 \ldots \beta_9$ για το μοντέλο μας.

Σημείωση: Στα αποτελέσματα που μας εμφανίζει η R, οι σταθεροί συντελεστές β1…β9 των ανεξάρτητων μεταβλητών εμφανίζονται ως X1…X9. Για τον λόγο στο εξής και εμείς θα αναφερόμαστε σε αυτές με αυτόν τον συμβολισμό.

5.4.3 Έλεγχος και Διάγνωση Μοντέλου

Στο σημείο αυτό, εφόσον έχουμε εξάγει το γραμμικό μοντέλο, δηλαδή τους συντελεστές-σταθερές α και $\beta_1 \ldots \beta_9$, θα πρέπει να ελέγξουμε κατά πόσο οι συντελεστές αυτοί επηρεάζουν σημαντικά το μοντέλο μας και αν το επηρεάζουν με σωστό τρόπο. Όπως αναφέραμε στην εισαγωγή του Κεφαλαίου, υπάρχει περίπτωση κάποιος συντελεστής (ή και περισσότεροι) και κατ' επέκταση κάποιο δημογραφικό χαρακτηριστικό, να αποδειχθεί είτε ότι έχει πολύ μικρή επιρροή είτε ότι επηρεάζει με αρνητικό τρόπο το μοντέλο μας. Στην περίπτωση αυτή θα πρέπει να εξάγουμε ξανά το μοντέλο μας, χωρίς να λάβουμε υπόψη αυτόν τον συντελεστή (ή τους συντελεστές).

Με την εντολή ‘‘**Summary (Model)** ’’ στην R, όπου ‘‘Model’’ είναι η γραμμική συνάρτηση που έχουμε εξάγει, θα έχουμε την εξής εικόνα για τους σταθερούς συντελεστές των ανεξάρτητων μεταβλητών καθώς και για τους όρους που αναλύθηκαν στο Κεφάλαιο 2, οι οποίοι μας βοηθούν στην αξιολόγηση του μοντέλου:

```
> summary (Model)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
    data = D)

Residuals:
   Min     1Q Median     3Q    Max
 -8950  -1793    -43   1585  25994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   641.02     595.07   1.077  0.28158
X1           -229.71     128.71  -1.785  0.07454 .
X2          -1593.58     171.84  -9.274  < 2e-16 ***
X3            957.18     165.42   5.786 8.98e-09 ***
X4            351.88      81.60   4.312 1.74e-05 ***
X5           1991.74     163.73  12.165  < 2e-16 ***
X6            509.78     180.68   2.821  0.00485 **
X7            582.54      67.01   8.693  < 2e-16 ***
X8           -707.83     241.60  -2.930  0.00345 **
X9             20.29      23.20   0.874  0.38204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3045 on 1314 degrees of freedom
Multiple R-squared:  0.4042,    Adjusted R-squared:  0.4001
F-statistic: 99.05 on 9 and 1314 DF,  p-value: < 2.2e-16
```

***Εικόνα 20**: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου με σταθερά και 9 ανεξάρτητες μεταβλητές καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Οι όροι που φαίνονται στην παραπάνω Εικόνα έχουν αναλυθεί στο Κεφάλαιο 2. Στη συνέχεια θα κάνουμε μία σύντομη αναφορά σους πιο σημαντικούς από αυτούς για την αξιολόγηση της προσαρμοστικότητας του μοντέλου που εξήχθη.

5.4.3.1 Επεξήγηση Συντελεστών

- **Residual Standard Error**

Όπως έχει αναφερθεί και στο Κεφάλαιο 2, το RSA είναι ένας όρος που δείχνει την ‘‘ποιότητα’’ της γραμμικής παλινδρόμησης που επιτεύχθηκε. Στην πραγματικότητα, κάθε γραμμικό μοντέλο υποτίθεται ότι περιέχει έναν συντελεστή σφάλματος Ε. Λόγω της παρουσίας αυτού του σφάλματος, δεν μπορούμε να προβλέψουμε τέλεια την εξαρτημένη μεταβλητή Y με τη βοήθεια των ανεξάρτητων μεταβλητών. Στην ουσία το Residual Standard Error αντιπροσωπεύει τη μέση τιμή των αποκλίσεων των πραγματικών τιμών του Y, από το γραμμικό μοντέλο που έχουμε κατασκευάσει (ή αλλιώς τη μέση τιμή των αποκλίσεων των πραγματικών τιμών του Y από τις τιμές του Y που έχουμε προβλέψει, οι οποίες ανήκουν στο γραμμικό μοντέλο που έχουμε εξάγει). Στην περίπτωση του μοντέλου που εξάγαμε αρχικά, το οποίο συμπεριλαμβάνει όλους τους συντελεστές, φαίνεται ότι κατά μέσο όρο οι πραγματικές τιμές της μηνιαίας κατανάλωσης νερού των 57 νοικοκυριών που μελετάμε, διαφέρουν κατά 3045 λίτρα από τις αντίστοιχες τιμές τις οποίες έχουμε προβλέψει. Στα αποτελέσματά μας φαίνεται πως το Residual Standard Error υπολογίστηκε με 1314 ‘‘βαθμούς ελευθερίας’’. Οι βαθμοί ελευθερίας είναι ο αριθμός των σημείων (data points) που λήφθηκαν υπόψη στον υπολογισμό των σταθερών $X_1 \ldots X_9$ και Intercept, αφού αφαιρέσουμε το πλήθος των μεταβλητών αυτών. Δηλαδή το αρχείο με τις καταναλώσεις που εισάγαμε προς επεξεργασία είχε 1324 γραμμές-σημεία, (Εικόνα 4.3) μείον 10 μεταβλητές (Intercept και $X_1 \ldots X_9$), άρα 1314 βαθμοί ελευθερίας.

- **Multiple R-squared και Adjusted R-squared**

Ο συντελεστής R-squared (R^2) παρέχει μία εκτίμηση για το κατά πόσο το μοντέλο μας προσεγγίζει τις πραγματικές τιμές. Εκφράζει ποσοστό διακύμανσης, επομένως λαμβάνει τιμές μεταξύ 0 και 1. Ο συντελεστής αυτός αποτελεί ένα μέτρο για το κατά πόσο υπάρχει γραμμική σχέση ανάμεσα στις ανεξάρτητες και την εξαρτημένη μεταβλητή Y, αυτή δηλαδή που δείχνει το αποτέλεσμα. Μία τιμή του R^2 κοντά στο 0 υποδηλώνει πως οι τιμές των ανεξάρτητων μεταβλητών στο γραμμικό μας μοντέλο, δεν μπορούν να δικαιολογήσουν τις διακυμάνσεις στις αντίστοιχες τιμές τη μεταβλητής Y. Αντίθετα, με μία τιμή του R^2 κοντά στη δικαιολογείται καλύτερα η διακύμανση αυτή. Στην περίπτωση μας έχουμε π.χ. ένα R^2 (είτε δούμε το Multiple R^2 είτε το Adjusted R^2) με τιμή περίπου 0,4. Αυτό πρακτικά σημαίνει πως μόνο το 40% από τη συνολική διακύμανση που βρέθηκε στο σύνολο των εξαρτημένων μεταβλητών Y μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές του μοντέλου μας.

Σημείωση: Στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, όπως και στη δική μας περίπτωση, ο συντελεστής Multiple R^2 αυξάνει όσο μεγαλύτερος είναι ο αριθμός των ανεξάρτητων μεταβλητών του γραμμικού μοντέλου. Για το λόγο αυτό προτιμούμε να δίνουμε σημασία στον συντελεστή Adjusted R^2 (προσαρμοσμένο R^2)

ο οποίος προσαρμόζεται έτσι ώστε να μην επηρεάζεται από τον αριθμό των ανεξάρτητων μεταβλητών.

Προφανώς, μία τιμή του R^2 κοντά στο 40% δεν είναι ικανοποιητική στην αξιολόγηση του μοντέλου μας, γι αυτό θα πρέπει να διερευνήσουμε εάν και πόσες από τις ανεξάρτητες μεταβλητές θα πρέπει να παραληφθούν από το μοντέλο μας.

- **F-Statistic:**

Αποτελεί έναν δείκτη για το κατά πόσο υπάρχει συσχέτιση ανάμεσα στην εξαρτημένη και τις ανεξάρτητες μεταβλητές μας. Όσο μεγαλύτερος της μονάδας είναι αυτός ο συντελεστής, τόσο το καλύτερο για το μοντέλο μας. Ωστόσο το πόσο μεγάλο πρέπει να είναι το F-statistic, εξαρτάται τόσο από τα data points που χρησιμοποιούμε, όσο και από τον αριθμό των ανεξάρτητων μεταβλητών του μοντέλου μας. Γενικά, όταν ο αριθμός των data points είναι μεγάλος (όπως και στην περίπτωσή μας), μία τιμή για τον συντελεστή F-statistic λίγο μεγαλύτερη της μονάδας θεωρείται ικανοποιητική ώστε να αποκλείσουμε τη μηδενική υπόθεση (δηλαδή την υπόθεση που θεωρεί πως δεν υπάρχει καμία συσχέτιση ανάμεσα στις ανεξάρτητες και την εξαρτημένη μεταβλητή). Σε αντίθετη περίπτωση, όπου δηλαδή ο αριθμός των data points είναι μικρός, απαιτείται μία μεγάλη τιμή του F-statistic προκειμένου να εξετάσουμε εάν υπάρχει σχέση μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών. Στην περίπτωση που εξετάζουμε, έχουμε αφενός μεγάλο αριθμό data points (1324) και αφετέρου μεγάλη τιμή για τον συντελεστή F-statistic (ενώ είπαμε ότι αρκεί μία τιμή λίγο μεγαλύτερη της μονάδας), κάτι που μας οδηγεί στο συμπέρασμα ότι μπορούμε να υποθέσουμε ότι οι ανεξάρτητες μεταβλητές συσχετίζονται με την εξαρτημένη. Αυτό όμως δε σημαίνει ότι όλες οι ανεξάρτητες μεταβλητές $x_1 \ldots x_9$ συσχετίζονται με την εξαρτημένη μεταβλητή Y. Αυτό είναι κάτι που θα πρέπει να διερευνήσουμε, όπως θα εξηγηθεί παρακάτω.

5.4.3.2 Επεξήγηση Αποτελεσμάτων

Σύμφωνα με την R, οι συντελεστές που έχουν αστερίσκους στο τέλος είναι πολύ σημαντικοί για το γραμμικό μοντέλο μας, δηλαδή επηρεάζουν σημαντικά την εξαρτημένη μεταβλητή Y ή με άλλα λόγια υπάρχει μεγάλη συσχέτιση μεταξύ των συντελεστών αυτών και του Y. Οι συντελεστές χωρίς αστερίσκο δεν έχουν συσχέτιση με την έξοδο Y ή με άλλα λόγια δεν επηρεάζουν σωστά το μοντέλο μας ή το επηρεάζουν με λάθος τρόπο και επομένως κάποιος ή κάποιοι από αυτούς θα πρέπει να παραλειφθούν. Θα πρέπει δηλαδή να εξάγουμε ένα νέο μοντέλο το οποίο δε θα περιέχει τους συντελεστές αυτούς.

Ένας άλλος τρόπος να εξηγήσουμε μαθηματικά το κατά πόσο μία σταθερά από τις $X_1..X_9$ και Intercept είναι σημαντική για το μοντέλο, πέραν των αστερίσκων, είναι η τιμή (πιθανότητα) του **Pr (>ltl)**. Αυτή η τιμή απαντά στην υπόθεση που έχει γίνει στο κατά πόσον είναι σημαντική η αντίστοιχη σταθερά. Η υπόθεση που έχει γίνει (Έλεγχος Μηδενικής υπόθεσης) είναι πως ο η ανεξάρτητη μεταβλητή δεν έχει καθόλου συσχέτιση με την έξοδο Y, είναι δηλαδή εντελώς ασήμαντη για το μοντέλο μας, επομένως μεγάλη τιμή της Pr (κοντά στο 1) επιβεβαιώνει αυτή την υπόθεση. Σε περίπτωση που η τιμή της Pr είναι μικρή (κοντά στο 0) σημαίνει πως η υπόθεση απορρίπτεται, επομένως η σταθερά σημαντική για το μοντέλο. Διασταυρώνοντας άλλωστε και την ποιοτική με τη μαθηματική επεξήγηση στην περίπτωση μας, φαίνεται πως οι σταθερές με τους αστερίσκους έχουν πολύ μικρή, σχεδόν μηδενική τιμή της Pr σε αντίθεση με τους συντελεστές χωρίς αστερίσκο των οποίων η τιμή της Pr είναι μεγαλύτερη κατά τουλάχιστον 2 τάξεις μεγέθους.

5.4.4 Βελτίωση του Γραμμικού Μοντέλου

Με βάση την επεξήγηση τόσο των συντελεστών του γραμμικού μας μοντέλου, όσο και των αποτελεσμάτων, όπως αυτά αναπτύχθηκαν προηγουμένως, θα πρέπει να μελετήσουμε το εάν και πώς μπορούμε να βελτιώσουμε το μοντέλο μας, δηλαδή οι τιμές της εξαρτημένης μεταβλητής (μηνιαία κατανάλωση νερού) που περιέχει το μοντέλο να προσεγγίζουν όσο το δυνατόν τις αντίστοιχες πραγματικές τιμές. Η μελέτη αυτή θα γίνει με τη βοήθεια κάποιων από τους συντελεστές που εξηγήθηκαν παραπάνω. Θα δοκιμάζουμε να παραλείψουμε από το μοντέλο μας κάποιες ανεξάρτητες μεταβλητές (μία κάθε φορά) και θα ελέγχουμε πώς το μοντέλο μεταβάλλεται. Οι συντελεστές αυτοί, καθώς και ο τρόπος με τον οποίο θέλουμε να μεταβληθούν παρουσιάζονται στην επόμενη Ενότητα:

5.4.4.1 Κριτήρια Αξιολόγησης και Βελτίωσης του Γραμμικού Μοντέλου

Όταν κάνουμε δοκιμές ανάμεσα στις διαφορετικές εκδοχές του γραμμικού μοντέλου, οι όροι οι οποίοι πρέπει να εξετάζουμε προκειμένου να βγάλουμε συμπέρασμα για το αν μία δοκιμή βελτιώνει ή υποβαθμίζει την ακρίβεια του μοντέλου είναι οι παρακάτω:

1) Adjusted R-squared

Όπως εξηγήσαμε και πριν, επιλέγουμε τον συγκεκριμένο συντελεστή αντί του Multiple R-squared, καθώς δεν επηρεάζεται ανάλογα με τον αριθμό των ανεξάρτητων μεταβλητών ενός γραμμικού μοντέλου. Ό,τι τροποποίηση πραγματοποιήσουμε στο μοντέλο μας με σκοπό τη βελτιστοποίηση του, όπως αναλύθηκε παραπάνω, εάν ο συντελεστής αυτός αυξάνει και πλησιάζει στη μονάδα, σημαίνει ότι είμαστε σε καλό δρόμο. Σε αντίθετη περίπτωση, θα σημαίνει πως οι τροποποιήσεις που πραγματοποιούμε στο μοντέλο δεν το επηρεάζουν με σωστό τρόπο.

2) Pr (>ItI) σε συνδυασμό με αριθμό αστερίσκων

Όπως εξηγήσαμε εάν μία ή περισσότερες από τις σταθερές Intercept και $X_1 \ldots X_9$ δεν έχει αστερίσκο στην παρουσίασης ανάλυση των αποτελεσμάτων στην R, σημαίνει πως δεν είναι σημαντική για το γραμμικό μοντέλο ή το επηρεάζει με λάθος τρόπο. Αυτό συνδυάζεται και με την τιμή του Pr (>ItI). Αυτό μας δίνει τη δυνατότητα να έχουμε μια αφετηρία για τη βελτιστοποίηση του μοντέλου, δοκιμάζοντας να παραλείψουμε στην αρχή κάποια σταθερά που δεν έχει αστερίσκους βλέποντας πώς αλλάζει το μοντέλο μας χωρίς αυτή τη σταθερά. Μπορούμε να κάνουμε δοκιμές παραλείποντας κάποια άλλη σταθερά στη συνέχεια (μία κάθε φορά) και να δούμε πάλι πώς αλλάζει το μοντέλο μας.

Στόχος μας είναι να φτάσουμε σε ένα σημείο στο οποίο όλες οι σταθερές θα έχουν αστερίσκο στην ανάλυση του μοντέλου, άρα οι αντίστοιχες ανεξάρτητες μεταβλητές θα είναι σημαντικές για το μοντέλο μας και να έχουμε επιτύχει τη βέλτιστη τιμή του συντελεστή Adjusted R-squared

5.4.4.2 Δοκιμές για τη Βελτίωση του Μοντέλου

Δοκιμή #1: Παραλείπουμε τη σταθερά X_1

Μία πρώτη σκέψη που θα έκανε κάποιος, θα ήταν πως το μορφωτικό επίπεδο του ιδιοκτήτη μιας κατοικίας δεν έχει τόσο άμεση σχέση με την κατανάλωση νερού, όσο έχουν άλλες παράμετροι του μοντέλου μας, όπως ο αριθμός των ατόμων ή το μέγεθος της οικίας. Αυτό άλλωστε αποτυπώνεται και στα αποτελέσματα της αρχικής ανάλυσης που έγινε με τη βοήθεια της R, (Εικόνα 5.5) όπου πράγματι φαίνεται πως η σταθερά X_1 δεν έχει αστερίσκο, άρα δεν είναι σημαντική για το μοντέλο μας ή δεν το

επηρεάζει με σωστό τρόπο. Έτσι θα δοκιμάσουμε να εξάγουμε ένα γραμμικό μοντέλο το οποίο δε θα περιέχει τη σταθερά X_1. Οι αντίστοιχες εντολές, καθώς και τα αποτελέσματα της R φαίνονται στην Εικόνα που ακολουθεί:

```
> ModelNew1 <- lm (Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9, data =D)
>
> ModelNew1

Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9, data = D)

Coefficients:
(Intercept)           X2           X3           X4           X5           X6           X7
     298.37     -1570.48       954.43       385.87      1953.74       548.99       531.92
         X8           X9
    -746.08        19.71

>
> summary (ModelNew1)

Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-9065.6 -1846.4   -21.3  1559.3 26108.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   298.37     563.72   0.529  0.59669
X2          -1570.48     171.49  -9.158  < 2e-16 ***
X3            954.43     165.55   5.765 1.01e-08 ***
X4            385.87      79.41   4.859 1.32e-06 ***
X5           1953.74     162.47  12.025  < 2e-16 ***
X6            548.99     179.49   3.059  0.00227 **
X7            531.92      60.77   8.754  < 2e-16 ***
X8           -746.08     240.85  -3.098  0.00199 **
X9             19.71      23.22   0.849  0.39607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3048 on 1315 degrees of freedom
Multiple R-squared:  0.4028,    Adjusted R-squared:  0.3991
F-statistic: 110.9 on 8 and 1315 DF,  p-value: < 2.2e-16
```

***Εικόνα 21**: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου με σταθερά και 8 ανεξάρτητες μεταβλητές, καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Από την ανάλυση του γραμμικού μοντέλου που προέκυψε χωρίς τη σταθερά X_1, βλέπουμε ότι αφενός υπάρχουν πάλι σταθερές χωρίς αστερίσκο (Intercept και X_9) και αφετέρου ο συντελεστής Adjusted R-squared μειώθηκε ελάχιστα σε σχέση με πριν. Οι βαθμοί ελευθερίας προφανώς αυξήθηκαν κατά μία μονάδα, αφού οι συντελεστές των ανεξάρτητων μεταβλητών μειώθηκαν κατά μία μονάδα. Οι υπόλοιποι συντελεστές δεν άλλαξαν σημαντικά σε σχέση με πριν. Επομένως, λόγω της μείωσης του Adjusted R-squared, αντί της επιθυμητής αύξησής του, θεωρούμε ότι το γραμμικό μοντέλο δε βελτιώθηκε και προχωράμε σε επόμενη δοκιμή.

<u>Δοκιμή #2:</u> Παραλείπουμε τη σταθερά X_9

Επόμενη δοκιμή είναι να παραλείψουμε την επόμενη σταθερά που δεν έχει αστερίσκο, η οποία είναι η X_9. Αυτό σημαίνει ότι εφόσον θα βγει η σταθερά που αντιπροσωπεύει τους μήνες, στο γραμμικό μοντέλο μας θα φαίνεται ότι κάθε νοικοκυριό έχει σταθερή κατανάλωση σε ένα έτος. Αυτό όμως δεν παίζει σημαντικό

ρόλο για το μοντέλο μας, γιατί αυτό που μας νοιάζει είναι το κατά πόσο το μοντέλο μας προσεγγίζει τις πραγματικές τιμές. Οι αντίστοιχες εντολές, καθώς και τα αποτελέσματα της R φαίνονται στην Εικόνα που ακολουθεί:

```
>  ModelNew3 <- lm (Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 , data =D)
>
> ModelNew3

Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8, data = D)

Coefficients:
(Intercept)           X2           X3           X4           X5           X6           X7
      417.2      -1569.1        957.0        385.4       1954.7        546.2        531.4
         X8
     -744.7

> summary(ModelNew3)

Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-9096.2 -1842.8    -5.2  1554.6 26124.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   417.15     546.02   0.764  0.44501
X2          -1569.10     171.47  -9.151  < 2e-16 ***
X3            957.04     165.50   5.783 9.18e-09 ***
X4            385.44      79.40   4.854 1.35e-06 ***
X5           1954.69     162.45  12.033  < 2e-16 ***
X6            546.21     179.44   3.044  0.00238 **
X7            531.37      60.76   8.746  < 2e-16 ***
X8           -744.72     240.82  -3.093  0.00203 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3048 on 1316 degrees of freedom
Multiple R-squared:  0.4024,    Adjusted R-squared:  0.3993
F-statistic: 126.6 on 7 and 1316 DF,  p-value: < 2.2e-16
```

Εικόνα 22*: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου με σταθερά και 8 ανεξάρτητες μεταβλητές, καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Όπως φαίνεται από την ανάλυση, ο συντελεστής Adjusted R-squared δε μεταβλήθηκε σε σχέση με τον αντίστοιχο συντελεστή του αρχικού γραμμικού μοντέλου, οι υπόλοιπο συντελεστές παρέμειναν στα ίδια επίπεδα, ενώ οι σταθερές Intercept και X1 εξακολουθούν να μην έχουν αστερίσκο. Αυτό σημαίνει πως η παράλειψη του X_9 δεν επέφερε κάποια βελτίωση στο γραμμικό μας μοντέλο σε σχέση με την αρχική του μορφή.

Στο σημείο αυτό, διαπιστώνουμε πως εάν παραλείψουμε μία από τις δύο σταθερές του μοντέλου που δεν έχουν αστερίσκο, δεν επέρχεται καμία βελτίωση του γραμμικού μας μοντέλου σε σχέση με την αρχική του μορφή. Γι αυτό θα δοκιμάσουμε να παραλείψουμε και τις 2 σταθερές.

<u>Δοκιμή #3:</u> Παραλείπουμε τις σταθερά X_1 και X_9

Οι αντίστοιχες εντολές, καθώς και τα αποτελέσματα της R φαίνονται στην Εικόνα που ακολουθεί:

```
>  ModelNew3 <- lm (Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 , data =D)
>
> ModelNew3

Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8, data = D)

Coefficients:
(Intercept)           X2           X3           X4           X5           X6           X7
      417.2      -1569.1        957.0        385.4       1954.7        546.2        531.4
         X8
     -744.7

> summary(ModelNew3)

Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-9096.2 -1842.8    -5.2  1554.6 26124.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   417.15     546.02   0.764  0.44501
X2          -1569.10     171.47  -9.151  < 2e-16 ***
X3            957.04     165.50   5.783 9.18e-09 ***
X4            385.44      79.40   4.854 1.35e-06 ***
X5           1954.69     162.45  12.033  < 2e-16 ***
X6            546.21     179.44   3.044  0.00238 **
X7            531.37      60.76   8.746  < 2e-16 ***
X8           -744.72     240.82  -3.093  0.00203 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3048 on 1316 degrees of freedom
Multiple R-squared:  0.4024,    Adjusted R-squared:  0.3993
F-statistic: 126.6 on 7 and 1316 DF,  p-value: < 2.2e-16
```

Εικόνα 23*: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου με σταθερά και 7 ανεξάρτητες μεταβλητές, καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Παρατηρώντας την ανάλυση του γραμμικού μοντέλου, φαίνεται ότι ούτε παραλείποντας τις σταθερές X_1 και X_9 από το γραμμικό μοντέλο, επέρχεται κάποια βελτίωση σε σχέση με την αρχική μορφή. Γι αυτό θα ελέγξουμε την αλλαγή του μοντέλου με την παράλειψη κάποιας άλλης σταθεράς.

<u>Δοκιμή #4:</u> Παραλείπουμε τη σταθερά Intercept

Η σταθερά Intercept όπως έχουμε εξηγήσει, είναι μία σταθερά στο γραμμικό μας μοντέλο, στην οποία δεν αντιστοιχεί κάποια ανεξάρτητη μεταβλητή, και ισούται με την τιμή του Y, στην περίπτωση που όλες οι ανεξάρτητες μεταβλητές είναι μηδέν. Είναι η Τρίτη σταθερά του γραμμικού μοντέλου η οποία δε φαίνεται να επηρεάζει σημαντικά το μοντέλο ή να το επηρεάζει με λάθος τρόπο. Η παρουσία του στο μοντέλο μας, σημαίνει πρακτικά πως ακόμα και αν είχαμε ένα φανταστικό

νοικοκυριό με 0 άτομα, μηδενική έκταση οικίας χωρίς κανένα εισόδημα κτλ, η μηνιαία κατανάλωση νερού δε θα ήταν μηδενική, αλλά θα είχε πάλι κάποια σταθερή τιμή. Επομένως θα δοκιμάσουμε να εξάγουμε ένα γραμμικό μοντέλο το οποίο δε θα περιέχει καθόλου αυτή τη σταθερά. Οι αντίστοιχες εντολές, καθώς και τα αποτελέσματα της R φαίνονται στην Εικόνα που ακολουθεί:

```
> ModelNew4 <- lm (Y ~ -1 + X1+ X2 + X3 + X4 + X5 + X6 + X7 + X8 +X9 , data =D)
>
> ModelNew4

Call:
lm(formula = Y ~ -1 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
    X9, data = D)

Coefficients:
      X1         X2         X3         X4         X5         X6         X7         X8         X9
 -184.97  -1561.23    1022.91     371.18    1940.28     578.07     574.85    -561.24      26.05

> summary(ModelNew4)

Call:
lm(formula = Y ~ -1 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
    X9, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-8965.4 -1745.9   -18.9  1614.8 25918.1

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X1  -184.97     121.83  -1.518 0.129195
X2 -1561.23     169.21  -9.227  < 2e-16 ***
X3  1022.91     153.76   6.653 4.22e-11 ***
X4   371.18      79.61   4.662 3.44e-06 ***
X5  1940.28     156.61  12.389  < 2e-16 ***
X6   578.07     169.21   3.416 0.000654 ***
X7   574.85      66.64   8.627  < 2e-16 ***
X8  -561.24     199.64  -2.811 0.005008 **
X9    26.05      22.58   1.154 0.248868
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3046 on 1315 degrees of freedom
Multiple R-squared:  0.8796,    Adjusted R-squared:  0.8788
F-statistic:  1068 on 9 and 1315 DF,  p-value: < 2.2e-16
```

Εικόνα 24*: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου χωρίς σταθερά, με 9 ανεξάρτητες μεταβλητές, καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Παραλείποντας τη σταθερά Intercept από το γραμμικό μας μοντέλο, βλέπουμε μία εντυπωσιακή αύξηση του συντελεστή Adjusted R-squared, ο οποίος φαίνεται ότι διπλασιάστηκε. Επίσης ο συντελεστής F-statistic αυξήθηκε πάρα πολύ σε σχέση με τα μοντέλα των προηγούμενων δοκιμών, σχεδόν διπλασιάστηκε. Ωστόσο όπως έχουμε πει, εμείς θα ασχοληθούμε μόνο με τη μεταβολή του Adjusted R-squared, γιατί η τιμή του F-statistic δε μας βοηθά πάντα να αξιολογήσουμε το μοντέλο μας και άλλωστε και στις προηγούμενες δοκιμές η τιμή του ήταν αρκετά μεγαλύτερη από το 0, χωρίς αυτό να σημαίνει ότι οι τιμές του προσεγγίζουν με ικανοποιητικό τρόπο τις αντίστοιχες πραγματικές τιμές.

Συμπέρασμα: Έπειτα από τις διαγνωστικές δοκιμές που πραγματοποιήσαμε για την αξιολόγηση και την πιθανή βελτίωση του μοντέλου μας, διαπιστώθηκε πώς με την παράλειψη της σταθεράς Intercept βελτιώνεται εξαιρετικά ο συντελεστής Adjusted R-squared πλησιάζοντας πολύ περισσότερο τη μονάδα σε σχέση με τις προηγούμενες δοκιμές. Ωστόσο οι σταθερές X_1 και X_9 εξακολουθούν να μην έχουν κανέναν αστερίσκο, κάτι που σημαίνει πως παρόλο που το μοντέλο βελτιώθηκε σημαντικά, οι 2 αυτές σταθερές φαίνεται πως συνεχίσουν να μην είναι σημαντικές για το μοντέλο ή να μην το επηρεάζουν με σωστό τρόπο. Τίθεται λοιπόν το ερώτημα, εάν εκτός από τη βέβαιη παράλειψη του Intercept, θα προχωρήσουμε και την παράλειψη των 2 σταθερών X1 και X9 ή κάποιας εκ των 2. Για να απαντήσουμε σε αυτό το ερώτημα, θα προχωρήσουμε σε ένα δεύτερο γύρο δοκιμών του γραμμικού μοντέλου, όπου σίγουρα δε θα υπάρχει η σταθερά Intercept.

Επομένως, με βάση τα παραπάνω, θα κάνουμε διάφορες δοκιμές με κοινό σημείο αναφοράς την απουσία του Intercept και θα αξιολογήσουμε τα αποτελέσματα. Οι δοκιμές που ακολούθησαν είναι οι παρακάτω:

Δοκιμή Α: Παραλείπουμε Intercept και X_1

```
> ModelNew5 <- lm (Y~ -1 +X2+ X3 +X4 + X5 + X6 + X7 +X8 +X9 ,data=D)
>
> ModelNew5

Call:
lm(formula = Y ~ -1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
    data = D)

Coefficients:
      X2         X3         X4         X5         X6         X7         X8         X9
-1556.01     988.86     392.46    1930.85     580.50     533.04    -666.05      22.76

> summary(ModelNew5)

Call:
lm(formula = Y ~ -1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
    data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-9062.1 -1799.2   -14.5  1594.4 26057.4

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X2 -1556.01     169.26  -9.193  < 2e-16 ***
X3   988.86     152.19   6.497 1.16e-10 ***
X4   392.46      78.41   5.005 6.33e-07 ***
X5  1930.85     156.56  12.333  < 2e-16 ***
X6   580.50     169.28   3.429 0.000624 ***
X7   533.04      60.71   8.780  < 2e-16 ***
X8  -666.05     187.42  -3.554 0.000393 ***
X9    22.76      22.49   1.012 0.311606
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3047 on 1316 degrees of freedom
Multiple R-squared:  0.8794,    Adjusted R-squared:  0.8787
F-statistic:  1200 on 8 and 1316 DF,  p-value: < 2.2e-16
```

***Εικόνα 25**: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου χωρίς σταθερά, με 8 ανεξάρτητες μεταβλητές, καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Παρατηρούμε μία πολύ ανεπαίσθητη μείωση του συντελεστή Adjusted R-squared. Η σταθερά X_9 εξακολουθεί να μην έχει αστερίσκο, επομένως θα δοκιμάσουμε να την παραλείψουμε.

<u>Δοκιμή Β:</u> Παραλείπουμε Intercept και X_9

```
> ModelNew6 <- lm (Y~ -1 +X1 +X2+ X3 +X4 + X5 + X6 + X7 +X8 ,data=D)
>
> ModelNew6

Call:
lm(formula = Y ~ -1 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,
    data = D)

Coefficients:
     X1       X2       X3       X4       X5       X6       X7       X8
 -171.5  -1550.9   1043.2    375.8   1928.2    591.9    571.7   -522.5

> summary(ModelNew6)

Call:
lm(formula = Y ~ -1 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,
    data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-9013.1 -1764.5     0.7  1600.6 25922.1

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X1  -171.50     121.29  -1.414  0.15760
X2 -1550.88     168.99  -9.177  < 2e-16 ***
X3  1043.19     152.77   6.828 1.31e-11 ***
X4   375.79      79.52   4.726 2.54e-06 ***
X5  1928.19     156.28  12.338  < 2e-16 ***
X6   591.88     168.80   3.506  0.00047 ***
X7   571.67      66.59   8.585  < 2e-16 ***
X8  -522.53     196.82  -2.655  0.00803 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3046 on 1316 degrees of freedom
Multiple R-squared:  0.8795,    Adjusted R-squared:  0.8788
F-statistic:  1201 on 8 and 1316 DF,  p-value: < 2.2e-16
```

***Εικόνα 26**: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου χωρίς σταθερά, με 8 ανεξάρτητες μεταβλητές, καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

Ο συντελεστής Adjusted R-squared έχει ίδια τιμή με τη δοκιμή στην οποία παραλείψαμε μόνο τη σταθερά Intercept. Η σταθερά X_1 που παραλείψαμε στην προηγούμενη δοκιμή εξακολουθεί να μην έχει αστερίσκο.

Βλέπουμε πως παραλείποντας τη σταθερά Intercept και μία εκ των X_1 ή X_9, το μοντέλο ουσιαστικά δεν αλλάζει, ενώ η σταθερά που δεν παραλείφθηκε εξακολουθεί να μην επηρεάζει σημαντικά ή να επηρεάζει με λάθος τρόπο το μοντέλο μας. Επομένως θα πρέπει να εξετάσουμε και το μοντέλο που θα προκύψει εάν παραλείψουμε και τις τρεις σταθερές.

<u>Δοκιμή Γ:</u> Παραλείπουμε Intercept, X_1 και X_9

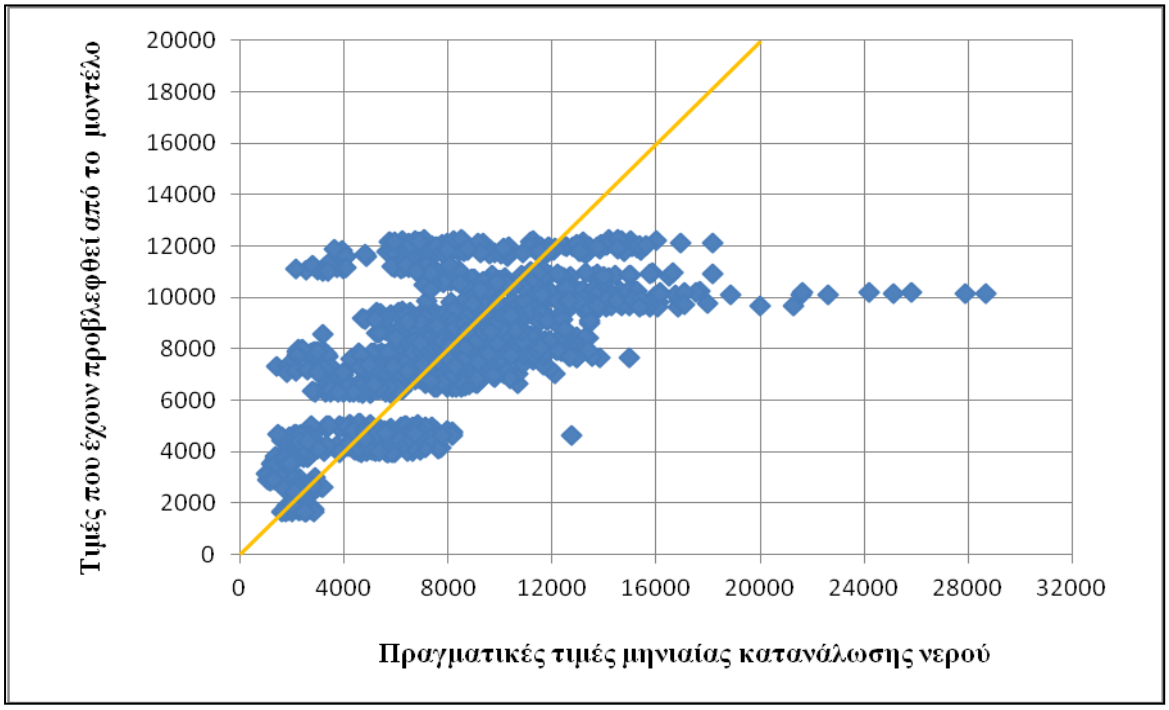
```
> ModelNew7 <- lm( Y~ -1  + X2 + X3 + X4 + X5 +X6 + X7 + X8 ,data=D )
>
> ModelNew7

Call:
lm(formula = Y ~ -1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = D)

Coefficients:
     X2       X3       X4       X5       X6       X7       X8
-1547.2   1008.9    395.2   1920.8    592.5    532.9   -625.2

> summary(ModelNew7)

Call:
lm(formula = Y ~ -1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-9098.0 -1806.8    -6.6  1571.0 26052.0

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X2 -1547.22     169.03  -9.153  < 2e-16 ***
X3  1008.93     150.90   6.686 3.37e-11 ***
X4   395.16      78.36   5.043 5.23e-07 ***
X5  1920.79     156.25  12.293  < 2e-16 ***
X6   592.52     168.87   3.509 0.000465 ***
X7   532.92      60.71   8.778  < 2e-16 ***
X8  -625.18     183.02  -3.416 0.000655 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3047 on 1317 degrees of freedom
Multiple R-squared:  0.8793,    Adjusted R-squared:  0.8787
F-statistic:  1371 on 7 and 1317 DF,  p-value: < 2.2e-16
```

Εικόνα 27*: Αναλυτική παρουσίαση των συντελεστών του γραμμικού μοντέλου χωρίς σταθερά, με 7 ανεξάρτητες μεταβλητές καθώς και των όρων που βοηθούν στην αξιολόγηση του μοντέλου*

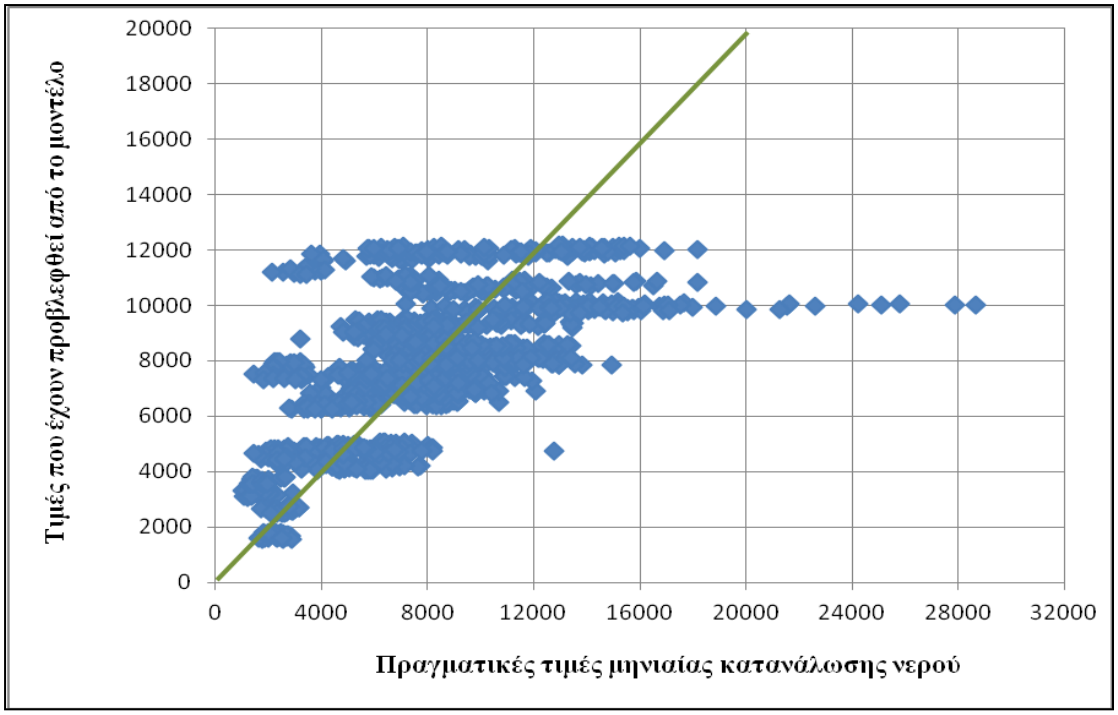
Εδώ παρατηρούμε πώς ο συντελεστής Adjusted R-squared έχει σχεδόν την ίδια τιμή που είχε και στις προηγούμενες δοκιμές. Η δεύτερη σημαντική παρατήρηση είναι πως όπως ήταν αναμενόμενο, εφόσον παραλείφθηκαν οι άλλες 2 σταθερές που δεν είχαν αστερίσκο, πλέον το μοντέλο μας αποτελείται από 7 μεταβλητές οι οποίες όλες είναι σημαντικές για το μοντέλο και τα επηρεάζουν με σωστό τρόπο. Επομένως με παράλειψη Intercept, X1 και X9, το μοντέλο φαίνεται πως δεν αλλοιώθηκε και μάλλον οδηγήθηκε στη βέλτιστη μορφή του.

Στο σημείο αυτό θα πρέπει να αποφασίσουμε τελικά πόσες σταθερές θα παραλείψουμε από το μοντέλο, εκτός από την Intercept. Ουσιαστικά, μόνο με την παράλειψη και των X_1 και X_9 το μοντέλο παύει να έχει συντελεστές χωρίς αστερίσκο, ωστόσο αυτό δε σημαίνει ότι από ένα αρχικό γραμμικό μοντέλο μπορούμε να διαγράφουμε όσες σταθερές θέλουμε, γιατί υπάρχει το ενδεχόμενο το μοντέλο που προκύπτει αλλάξει σημαντικά σε σχέση με την αρχική του μορφή.

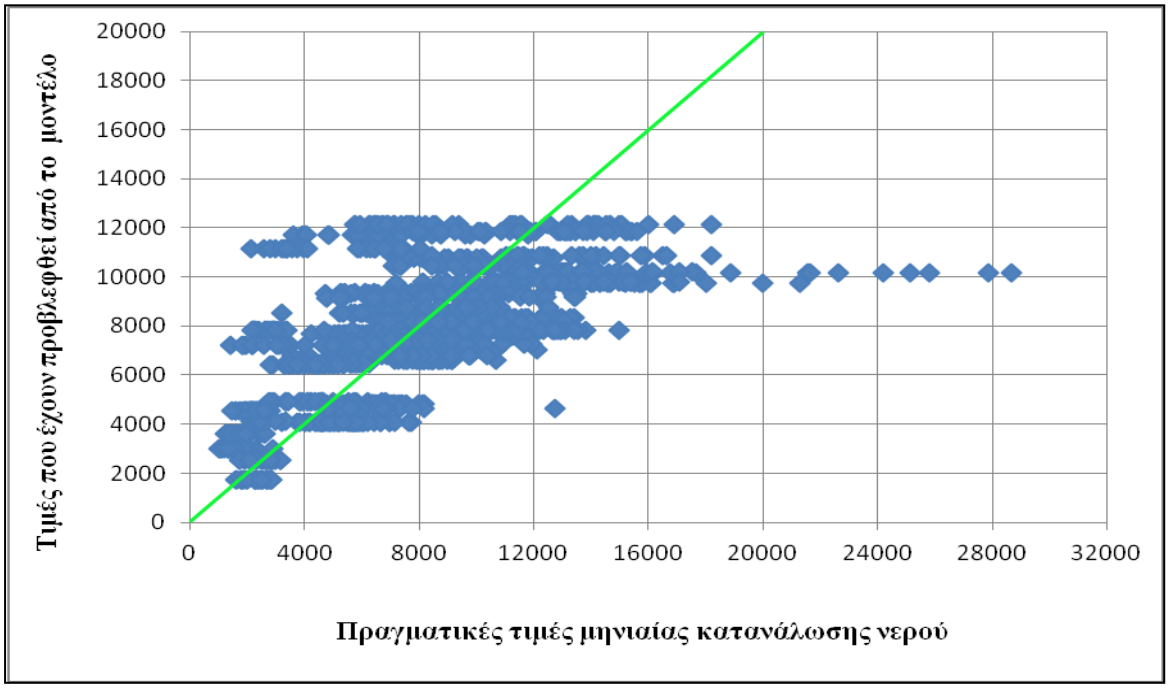
Μπορούμε να ρίξουμε μια ματιά στις κάτωθι γραφικές παραστάσεις, στις οποίες φαίνεται το πώς οι τιμές του κάθε μοντέλου προσεγγίζουν τις αντίστοιχες πραγματικές τιμές. Σε καθεμία από τις γραφικές παραστάσεις που ακολουθούν, στον οριζόντιο άξονα είναι καταχωρημένες οι πραγματικές τιμές κατανάλωσης νερού, ενώ στον κάθετο άξονα είναι οι τιμές που έχουν προβλεφθεί από τα αντίστοιχα μοντέλα που έχουν προκύψει από τις δοκιμές μας. Έτσι έχει προκύψει ένα διάγραμμα διασποράς (διάγραμμα X/Y), δηλαδή θα μπορούσαμε να πούμε ένα διάγραμμα μία συνάρτησης, στην οποία για κάθε πραγματική τιμή μηνιαίας κατανάλωσης νερού ενός νοικοκυριού (X) έχει προβλεφθεί μία αντίστοιχη τιμή από κάθε μοντέλο. Στο ίδιο σύστημα αξόνων για κάθε διάγραμμα διασποράς, έχει σχεδιαστεί και η ευθεία Y=X ως αναφορά. Όσο πιο κοντά στην ευθεία αυτή βρίσκεται κάθε ζεύγος X/Y (πραγματική τιμή/προβλεφθείσα τιμή) μηνιαίας κατανάλωσης, σημαίνει πως τόσο η τιμή που έχει προβλεφθεί από το μοντέλο προσεγγίζει την αντίστοιχη πραγματική τιμή.



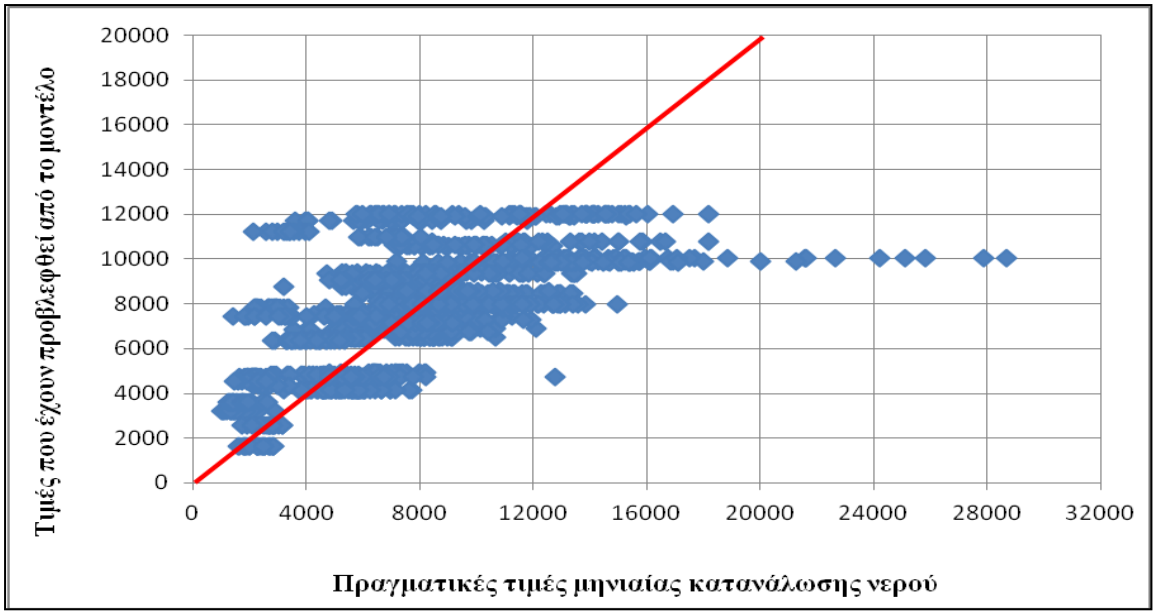
***Εικόνα 28**: Διάγραμμα διασποράς για γραμμικό μοντέλο χωρίς Intercept*



***Εικόνα 29**: Διάγραμμα διασποράς για γραμμικό μοντέλο χωρίς Intercept και X1*



***Εικόνα 30**: Διάγραμμα διασποράς για γραμμικό μοντέλο χωρίς Intercept και X9*



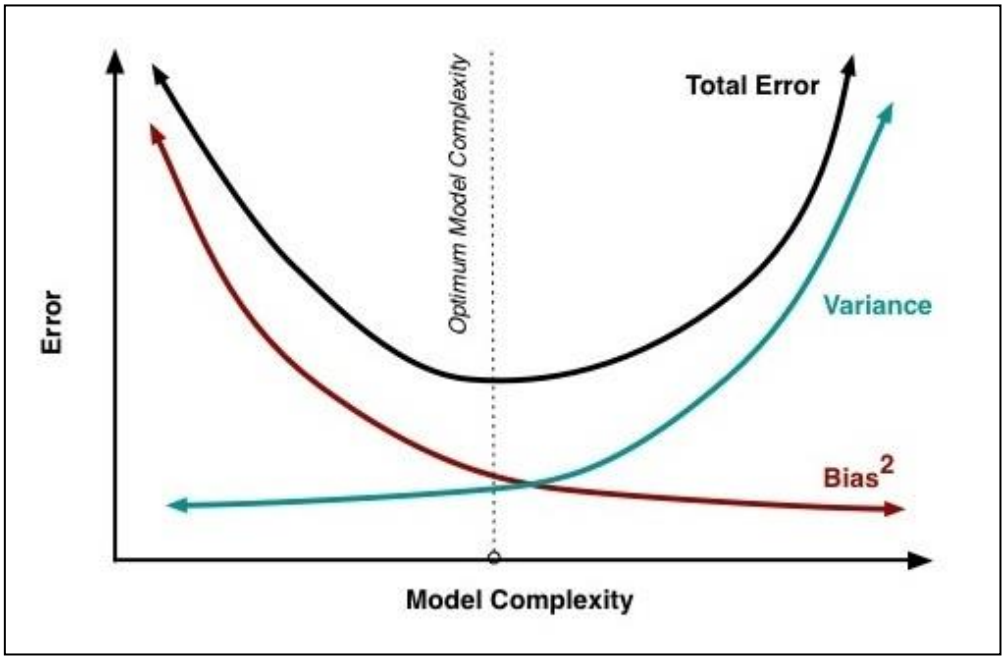
***Εικόνα 31**: Διάγραμμα διασποράς για γραμμικό μοντέλο χωρίς Intercept, X1 και X9*

Από τα διαγράμματα διασποράς δεν είναι ξεκάθαρο ποια μορφή του γραμμικού μοντέλου θα ήταν καλύτερο να επιλέξουμε, καθώς τα διαγράμματα μοιάζουν αρκετά μεταξύ τους. Θα σκεφτόταν κανείς πως η καλύτερη λύση θα ήταν να παραλειφθούν και οι 3 σταθερές, εφόσον η τιμή του συντελεστή Adjusted R-squared δεν αλλάζει όσο προχωράμε στις διάφορες δοκιμές. Δηλαδή αν ξεκινήσουμε με τη δοκιμή όπου παραλείπεται μόνο το Intercept και προχωράμε παραλείποντας επιπλέον και X_1 και X_9, βλέπουμε ότι το Adjusted R-squared παραμένει στα ίδια επίπεδα. Επομένως εφόσον με την παράλειψη του Intercept πετύχαμε την υψηλότερη τιμή για το Adjusted R-squared και από κει και πέρα παραμένει σταθερός με την παράλειψη των X_1 και X_9 που δεν έχουν αστερίσκο, θα ήταν λογικό να επιλέξουμε να παραλειφθούν και αυτές οι 2 σταθερές.

5.5 Θεωρία Πολυπλοκότητας ενός Μοντέλου (Over- and Under-Fitting) και τελική επιλογή βέλτιστου μοντέλου

Για να επιβεβαιώσουμε ότι η απόφαση να επιλέξουμε το γραμμικό μοντέλο όπου έχουν παραληφθεί εκτός από τη σταθερά Intercept και οι σταθερές X_1 και X_9 , θα εξετάσουμε τη θεωρία της πολυπλοκότητας ενός μοντέλου. Κάθε μαθηματικό μοντέλο πρόβλεψης χαρακτηρίζεται από την προκατάληψη και τη διακύμανση που αυτό εμφανίζει. Όσο η πολυπλοκότητα ενός μοντέλου αυξάνεται, δηλαδή ο αριθμός των παραμέτρων που σχηματίζουν το μοντέλο, τόσο η προκατάληψη μειώνεται όπως είναι λογικό, γιατί στην πρόβλεψη συμμετέχουν πολλές μεταβλητές. Από την άλλη η

διακύμανση αυξάνεται όσο αυξάνεται και η πολυπλοκότητα. Αυτός είναι και ο βασικός προβληματισμός που αντιμετωπίζουμε όταν κατασκευάζουμε ένα μοντέλο, ότι δηλαδή από τη μία θέλουμε ένα μοντέλο να είναι ακριβές, δηλαδή να αποτελείται από έναν ικανοποιητικό αριθμό παραμέτρων και άρα να μη χαρακτηρίζεται από προκατάληψη, ωστόσο από ένα σημείο και έπειτα όσο το πλήθος των παραμέτρων μεγαλώνει, τόσο αυξάνεται και η διακύμανση (over-fitting). Η θεωρία αυτή απεικονίζεται στο παρακάτω διάγραμμα, όπου φαίνονται και τα μεγέθη που αναφέραμε:



Εικόνα 32*: Διακύμανση και Προκατάληψη ενός Μοντέλου πρόβλεψης και η συνεισφορά τους στο συνολικό σφάλμα, ανάλογα με την πολυπλοκότητα του μοντέλου*

Από τη γραφική παράσταση φαίνεται πως η προκατάληψη ενός μοντέλου έχει αρνητική παράγωγο α' τάξης σε συνάρτηση με την πολυπλοκότητα του, ενώ αντίθετα η διακύμανση έχει θετική παράγωγο α' τάξης. Όπως γίνεται κατανοητό λοιπόν, θα πρέπει να βρεθεί ένα ''σημείο ισορροπίας'', δηλαδή ένα επίπεδο πολυπλοκότητας για το μοντέλο , ώστε η αύξηση της προκατάληψης να είναι ισοδύναμη με τη μείωση της διακύμανσης. Αυτό μαθηματικά εκφράζεται ως εξής:

$$\frac{dBias}{dComplexity} = -\frac{dVariance}{dComplexity}$$

Το σημείο ισορροπίας το οποίο φαίνεται στην παραπάνω γραφική παράσταση ορίζεται ως το **''βέλτιστο επίπεδο πολυπλοκότητας''** του μοντέλου, στο οποίο αντιστοιχεί η ελάχιστη τιμή του συνολικού σφάλματος του μοντέλου. Εάν η πολυπλοκότητα του μοντέλου υπερβεί το βέλτιστο επίπεδο, τότε μιλάμε για over-fitting μοντελοποίηση, ενώ αντίστοιχα εάν η πολυπλοκότητα του μοντέλου δε φτάνει το επιθυμητό επίπεδο, μιλάμε για under-fitting μοντελοποίηση. Στην πράξη, δεν υπάρχει κάποιος αναλυτικός τρόπος ώστε να εντοπίσουμε το βέλτιστο επίπεδο πολυπλοκότητας ενός μοντέλου. Για να εκτιμήσουμε το βέλτιστο επίπεδο, κάνουμε διαδοχικές δοκιμές τροποποιώντας την πολυπλοκότητα του μοντέλου που μελετάμε και επιλέγουμε τον αριθμό των παραμέτρων που θα αποτελούν το μοντέλο, όταν φτάσουμε στο σημείο στο οποίο ο δείκτης ο οποίος αξιολογεί κάθε μοντέλο αποκτήσει τη βέλτιστη τιμή του (μέγιστη ή ελάχιστη, ανάλογα με τον δείκτη).

Στην περίπτωση μας, όπως έχουμε ήδη αναφέρει, ο δείκτης που αξιολογεί το μοντέλο που προκύπτει από κάθε δοκιμή είναι ο συντελεστής Adjusted R-squared. Όπως διαπιστώσαμε, είτε παραλείψουμε μόνο τη σταθερά Intercept, είτε τη σταθερά Intercept και μία εκ των σταθερών X_1 και X_9, είτε και τις 3 σταθερές, η τιμή του δείκτη Adjusted R-squared παραμένει σταθερή και ίση με 0,8787. Αυτό πρακτικά σημαίνει πως πέρα από τη σταθερά Intercept που είναι βέβαιο πως πρέπει να παραλειφθεί, όσο συνεχίζουμε και διαγράφουμε μεταβλητές (δηλαδή μειώνουμε την πολυπλοκότητα του μοντέλου) η αξιολόγηση του δεν αλλάζει. Αυτό μας επιτρέπει να κάνουμε με ασφάλεια την υπόθεση πως όσο το μοντέλο περιελάμβανε τις σταθερές X_1 και X_9 χαρακτηριζόταν από πολυπλοκότητα που εκτεινόταν πέρα από το βέλτιστο επίπεδο, ενώ παραλείποντας τις 2 αυτές σταθερές, μειώσαμε την πολυπλοκότητα, φτάνοντας την στο επιθυμητό-βέλτιστο επίπεδο. Με άλλα λόγια, τα μοντέλο που περιέχουν τις σταθερές X_1 και X_9 ή μία εκ των δύο σταθερών, δεν προσφέρουν τίποτα παραπάνω σε σχέση με το μοντέλο που δεν περιέχει καμία από τις 2 σταθερές. Γι αυτό επιλέγουμε να αγνοήσουμε και τις 2 σταθερές, ώστε να μειώσουμε τη διακύμανση και να θεωρήσουμε πως προσεγγίζουμε το βέλτιστο επίπεδο πολυπλοκότητας.

Με βάση την παραπάνω ανάλυση επιλέγουμε το γραμμικό μοντέλο που δεν περιλαμβάνει τις σταθερές Intercept, X_1 και X_9 και έχει τον εξής μαθηματικό τύπο βάσει των αποτελεσμάτων της R:

$$\mathbf{Y = -1547{,}22 \cdot X_2 + 1008{,}93 \cdot X_3 + 395{,}16 \cdot X_4 + 1920{,}79 \cdot X_5 + 592{,}52 \cdot X_6 + 532{,}92 \cdot X_7 - 625{,}18 \cdot X_8}$$

ΚΕΦΑΛΑΙΟ 6: ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ ΜΕ ΤΗ ΜΕΘΟΔΟ K-MEANS

6.1 Εισαγωγή

Έχουμε αναφέρει ότι για τη μέθοδο της k-means χρειάζεται να έχουμε προκαθορίσει εμείς τον αριθμό των cluster με βάση τα οποία θα ομαδοποιηθούν τα δεδομένα μας. Αυτό βέβαια δε θα το κάνουμε αυθαίρετα, αλλά με τη βοήθεια της R, που θα επεξεργαστεί τα δεδομένα μας και θα μας δείξει ποιος είναι ο προτιμότερος αριθμός cluster. Η διαδικασία που θα ακολουθήσουμε για την ομαδοποίηση των δεδομένων έχει ως εξής:

1ο Βήμα: Βρίσκουμε βέλτιστο αριθμό συστάδων

2ο Βήμα: Ομαδοποιούμε και απεικονίζουμε τα δεδομένα μας ανάλογα με τη συστάδα στην οποία ανήκουν, με τις κατάλληλες εντολές στην R.

3ο Βήμα: Για τα 10 νοικοκυριά τα οποία έχουμε κρατήσει εκτός, ώστε να κάνουμε πρόβλεψη των καταναλώσεων τους και σύγκριση με τις πραγματικές τους τιμές, βρίσκουμε την απόσταση των ανεξάρτητων μεταβλητών τους από το κέντρο κάθε συστάδας. Κάθε νοικοκυριό θα θεωρήσουμε ότι ανήκει στη συστάδα εκείνη από την οποία του απέχει τη μικρότερη απόσταση.

4ο Βήμα: Εφόσον ομαδοποιήσουμε τα 10 νοικοκυριά σε cluster, θα κάνουμε πρόβλεψη για τη μηνιαία κατανάλωση τους, θεωρώντας ότι κάθε νοικοκυριό έχει ακριβώς την ίδια κατανάλωση με όλα τα νοικοκυριά του ίδιου cluster. Αυτή θα είναι και η πρόβλεψη μας. Η κατανάλωση του κάθε cluster θα ισούται με τη μέση κατανάλωση όλων των νοικοκυριών τα οποία ανήκουν σε αυτό, εφόσον θεωρούμε ότι τα νοικοκυριά κάθε cluster έχουν την ίδια συμπεριφορά.

5ο Βήμα: Συγκρίνουμε τις τιμές τις οποίες προβλέψαμε, με τις αντίστοιχες πραγματικές τιμές που έχουμε στη διάθεση μας.

Στο σημείο αυτό θα πρέπει αν διευκρινίσουμε πως η ανάλυσης κατά συστάδες θα γίνει με κριτήριο μόνο τα δημογραφικά χαρακτηριστικά κάθε νοικοκυριού, χωρίς να ληφθεί υπόψη η κατανάλωση τους. Τα νοικοκυριά θα ομαδοποιηθούν βάσει των ανεξάρτητων μεταβλητών που αντιπροσωπεύουν τα αντίστοιχα δημογραφικά χαρακτηριστικά και όπως είπαμε θα θεωρηθεί ότι έχουν όλα πανομοιότυπη καταναλωτική συμπεριφορά. Επίσης στην ανάλυση δε θα συμμετέχει προφανώς ούτε η ανεξάρτητη μεταβλητή X_9, η οποία αντιπροσωπεύει τους μήνες, καθώς τα δημογραφικά χαρακτηριστικά δεν αλλάζουν κατά τη διάρκεια του έτους.

6.2 Εφαρμογή της μεθόδου k-means με τη βοήθεια της R

6.2.1 Εισαγωγή Αρχείου

Με βάση τα όσα αναφέρθηκαν στην παραπάνω εισαγωγή, είναι προφανές ότι στο αρχείο που θα κάνουμε Import, σε κάθε νοικοκυριό θα αντιστοιχεί μία μόνο γραμμή. Αρχικά θέλουμε και τις 67 γραμμές, ώστε να κάνουμε scale τα δεδομένα όλων των νοικοκυριών με την ίδια κλίμακα, επομένως το αρχείο μας θα έχει τη μορφή που φαίνεται στην παρακάτω εικόνα:

```
>  D2<-read.csv("C:/Users/Δημήτρης/Desktop/IMPORTFORSCALE.txt",sep="\t",header = TRUE)
>  D2<-read.csv("C:/Users/Δημήτρης/Desktop/IMPORTFORSCALE.txt",sep="\t",header = TRUE)
> D2
> D2
             ID X1 X2 X3 X4 X5 X6 X7 X8
1   D14IA023010L  1  1  2  3  2  2  2  2
2   D12NA111684K  1  1  1  1  1  1  1  2
3   D12NA055486K  1  1  2  4  2  2  3  2
4   C12FA154674R  2  1  3  4  2  2  6  2
5   D12NA073420U  3  2  2  4  3  3  6  2
6   D14IA053695N  2  3  3  2  4  3  6  2
7   I14FA021789A  3  1  2  1  1  1  2  2
8   D12NA055665L  3  3  3  3  4  2  5  2
9   I13FA056306V  1  1  3  1  2  2  4  2
10  C14FA107096K  1  1  2  4  1  2  1  2
11  C12UA215866D  2  1  3  4  4  2  2  2
12  C13UA198418M  2  1  3  3  2  2  1  2
13  I13FA056611B  1  2  1  3  3  3  4  2
14  D12NA080070L  1  1  3  1  1  1  2  1
15  I14FA079412B  3  2  3  4  3  3  5  2
16  C12FA152684L  1  1  2  4  2  2  2  2
17  C12UA223429J  3  2  2  2  3  3  4  1
18  D13IA605405Z  2  1  2  4  5  3  2  2
19  I14FA043963O  3  1  2  1  2  2  6  1
20  C12UA115300Q  2  3  3  2  5  4  4  2
21  D13IA577080J  2  1  1  1  2  2  1  2
22  I13FA049641O  2  1  1  1  1  1  1  2
23  C12UA224149I  1  1  1  3  1  1  2  1
24  D12TA581630Z  2  1  1  1  2  2  1  2
25  C12FA152699S  3  3  2  2  4  3  4  2
26  C12FA154665Q  2  3  2  1  4  3  4  2
27  D12NA031884U  2  1  2  1  2  2  3  1
28  C14FA107301W  3  3  2  1  4  3  3  2
29  C13UA198207D  2  1  2  4  3  3  2  2
30  C13UA198977M  1  2  2  3  3  3  2  2
31  C12UA218724C  2  2  3  3  4  3  5  2
32  D13IA514382E  2  2  2  3  3  2  4  2
33  D13IA591157Z  3  2  2  3  4  3  3  2
34  C12UA115931P  1  1  2  4  4  3  1  2
35  D12TA539481L  3  1  2  3  2  2  5  2
36  I13FA049239I  2  3  2  3  3  3  3  2
```

***Εικόνα 33**: Εισαγωγή αρχείου για επεξεργασία (συνολικά 67 γραμμές)*

Έπειτα όπως είπαμε θα κάνουμε scale όλες τις ανεξάρτητες μεταβλητές των νοικοκυριών, ώστε ο αλγόριθμος της K-means να τις επεξεργαστεί σωστά. Στην εικόνα που ακολουθεί φαίνονται οι αντίστοιχες εντολές στην R, μαζί με τα απαραίτητα σχόλια.

6.2.2 Επεξεργασία δεδομένων

Επόμενο βήμα όπως είπαμε είναι να κάνουμε scale όλες τις ανεξάρτητες μεταβλητές των νοικοκυριών, ώστε ο αλγόριθμος της K-means να τις επεξεργαστεί σωστά. Στην εικόνα που ακολουθεί φαίνονται οι αντίστοιχες εντολές στην R, μαζί με τα απαραίτητα σχόλια.

```
> mydata <- D2
>
> #Βγάζουμε εκτός την πρώτη στήλη με τα IDs γιατί δεν είναι NUMERIC VALUES
>
> ID <- mydata$ID ; mydata$ID <- NULL
>
> mydata = as.data.frame(unclass(mydata))
>
> summary(mydata)
       X1              X2              X3              X4              X5              X6
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:1.500   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
 Median :2.000   Median :2.000   Median :2.000   Median :2.000   Median :3.000   Median :2.000
 Mean   :2.104   Mean   :1.791   Mean   :2.119   Mean   :2.433   Mean   :2.896   Mean   :2.373
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.500   3rd Qu.:3.500   3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :4.000   Max.   :5.000   Max.   :4.000
       X7              X8
 Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000
 Median :3.000   Median :2.000
 Mean   :3.179   Mean   :1.866
 3rd Qu.:4.000   3rd Qu.:2.000
 Max.   :6.000   Max.   :2.000
>
> dim(mydata)
[1] 67  8
> #Άρα έχουμε 67 παρατηρήσεις, καθεμία από τις οποίες χαρακτηρίζεται από 8 ανεξάρτητες μεταβλητές
>
> #Έπειτα πρέπει να αφαιρέσουμε από τα δεδομένα μας τυχόν κενές (NA) τιμές. Στη συγκεκριμένη περίπτωση δεν
 έχουμε κενές τιμές, απλά αναφέρουμε το βήμα
>
> mydataClean = na.omit(mydata)
>
> dim(mydataClean)
[1] 67  8
>
> #Έπειτα κάνουμε Scale όλες τις παρατηρήσεις, ώστε να μπορούν να διαβαστούν σωστά από τον αλγόριθμο της
k-means
>
> scaled_data = as.matrix (scale(mydataClean))
```

***Εικόνα 34**: Κατάλληλη επεξεργασία των δεδομένων μας πριν εφαρμόσουμε τη μέθοδο k-means*

Στην επόμενη εικόνα φαίνονται τα δεδομένα μας, όπως έχουν προκύψει έπειτα από το scaling που υπέστησαν, (λείπει η στήλη με τα ID των μετρητών κατανάλωσης κάθε νοικοκυριού, την οποία βγάλαμε εκτός, καθώς δεν περιέχει αριθμητικές τιμές).

Untitled1 × D × scaled_data ×

Filter

	X1	X2	X3	X4	X5	X6	X7	X8
1	-1.4139867	-0.9173700	-0.1938334	0.4792174	-0.73390679	-0.4812792	-0.7350054	0.3909686
2	-1.4139867	-0.9173700	-1.8171880	-1.2106544	-1.55343604	-1.7711074	-1.3583644	0.3909686
3	-1.4139867	-0.9173700	-0.1938334	1.3241533	-0.73390679	-0.4812792	-0.1116464	0.3909686
4	-0.1337555	-0.9173700	1.4295212	1.3241533	-0.73390679	-0.4812792	1.7584306	0.3909686
5	1.1464757	0.2423241	-0.1938334	1.3241533	0.08562246	0.8085490	1.7584306	0.3909686
6	-0.1337555	1.4020183	1.4295212	-0.3657185	0.90515171	0.8085490	1.7584306	0.3909686
7	1.1464757	-0.9173700	-0.1938334	-1.2106544	-1.55343604	-1.7711074	-0.7350054	0.3909686
8	1.1464757	1.4020183	1.4295212	0.4792174	0.90515171	-0.4812792	1.1350716	0.3909686
9	-1.4139867	-0.9173700	1.4295212	-1.2106544	-0.73390679	-0.4812792	0.5117126	0.3909686
10	-1.4139867	-0.9173700	-0.1938334	1.3241533	-1.55343604	-0.4812792	-1.3583644	0.3909686
11	-0.1337555	-0.9173700	1.4295212	1.3241533	0.90515171	-0.4812792	-0.7350054	0.3909686
12	-0.1337555	-0.9173700	1.4295212	0.4792174	-0.73390679	-0.4812792	-1.3583644	0.3909686
13	-1.4139867	0.2423241	-1.8171880	0.4792174	0.08562246	0.8085490	0.5117126	0.3909686
14	-1.4139867	-0.9173700	1.4295212	-1.2106544	-1.55343604	-1.7711074	-0.7350054	-2.5195751
15	1.1464757	0.2423241	1.4295212	1.3241533	0.08562246	0.8085490	1.1350716	0.3909686
16	-1.4139867	-0.9173700	-0.1938334	1.3241533	-0.73390679	-0.4812792	-0.7350054	0.3909686
17	1.1464757	0.2423241	-0.1938334	-0.3657185	0.08562246	0.8085490	0.5117126	-2.5195751
18	-0.1337555	-0.9173700	-0.1938334	1.3241533	1.72468096	0.8085490	-0.7350054	0.3909686
19	1.1464757	-0.9173700	-0.1938334	-1.2106544	-0.73390679	-0.4812792	1.7584306	-2.5195751
20	-0.1337555	1.4020183	1.4295212	-0.3657185	1.72468096	2.0983773	0.5117126	0.3909686
21	-0.1337555	-0.9173700	-1.8171880	-1.2106544	-0.73390679	-0.4812792	-1.3583644	0.3909686
22	-0.1337555	-0.9173700	-1.8171880	-1.2106544	-1.55343604	-1.7711074	-1.3583644	0.3909686
23	-1.4139867	-0.9173700	-1.8171880	0.4792174	-1.55343604	-1.7711074	-0.7350054	-2.5195751
24	-0.1337555	-0.9173700	-1.8171880	-1.2106544	-0.73390679	-0.4812792	-1.3583644	0.3909686
25	1.1464757	1.4020183	-0.1938334	-0.3657185	0.90515171	0.8085490	0.5117126	0.3909686
26	-0.1337555	1.4020183	-0.1938334	-1.2106544	0.90515171	0.8085490	0.5117126	0.3909686
27	-0.1337555	-0.9173700	-0.1938334	-1.2106544	-0.73390679	-0.4812792	-0.1116464	-2.5195751

Showing 1 to 27 of 67 entries

Εικόνα 35*: Τα δεδομένα μας έπειτα από scaling (συνολικά 67 γραμμές)*

6.2.3 Εύρεση βέλτιστου αριθμού συστάδων

Στο σημείο αυτό είμαστε έτοιμοι να βρούμε τον προτεινόμενο από την R αριθμό των clusters στα οποία θα ομαδοποιήσουμε τα δεδομένα μας. Αυτό θα γίνει με διάφορους τρόπους, οι οποίοι θα παρουσιαστούν ώστε να διαπιστώσουμε εάν συμφωνούν και συγκλίνουν στον ίδιο προτεινόμενο clusters. Τα δεδομένα τα οποία θα επεξεργαστούμε θα είναι αυτά της μεταβλητής ‘‘Scaled_data’’ που φαίνονται στην Εικόνα 6.3, αφού αφαιρέσουμε τα 10 νοικοκυριά τα οποία θα κρατήσουμε εκτός για να αξιολογήσουμε την πρόβλεψη μας όταν ολοκληρωθεί η ανάλυση. Οι 57 από τις 67 πλέον γραμμές, θα καταχωρηθούν στη μεταβλητή ‘‘data’’.

1ος Τρόπος : Elbow Method

Με τη μέθοδο Elbow θα αναπαραστήσουμε γραφικά την καμπύλη του αθροίσματος των τετραγώνων των σφαλμάτων των αποστάσεων των αντικειμένων σε ένα cluster, από το μέσο του cluster (**ESS**), σε συνάρτηση με τον αριθμό των clusters. Στην ουσία, γίνεται εφαρμογή της μεθόδου μέτρησης απόστασης του Ward, που αναφέρθηκε στο 3ο Κεφάλαιο. Από τη γραφική παράσταση, ανάλογα με το σημείο καμπής της (Elbow) θα προσπαθήσουμε να βρούμε ποιοτικά έναν αριθμό clusters, ο οποίος θα ικανοποιεί ικανοποιητικά τα δεδομένα μας.

Ο αλγόριθμος της k-means όπως είπαμε, προσπαθεί να ελαχιστοποιήσει το συνολικό άθροισμα τετραγώνων των μέσα στο cluster. Για να συνενωθούν δύο συστάδες από συνολικό πλήθος k συστάδων, ελέγχονται τα δυνατά K·(K-1)/2 ζεύγη συστάδων τα οποία μπορούν να δημιουργηθούν και επιλέγεται το ζεύγος, το οποίο όταν ενωθεί θα μας δώσει τη συστάδα με το ελάχιστο ESS. Τα σχηματιζόμενα clusters σε κάθε βήμα σχηματίζονται έτσι ώστε η λύση που προκύπτει να δίνει το μικρότερο δυνατό ESS μέσα στο αρχικό cluster.

Οι εντολές που χρησιμοποιήσαμε στην R για την κατασκευή της καμπύλης του ESS, συναρτήσει του αριθμού k των συστάδων, καθώς και τα αντίστοιχα σχόλια φαίνονται στο επόμενο Σχήμα, όπου όπως φαίνεται, ψάχνουμε τον προτεινόμενο αριθμό συστάδων μέσα σε ένα εύλογο εύρος, δηλαδή μεταξύ 2 και 20.

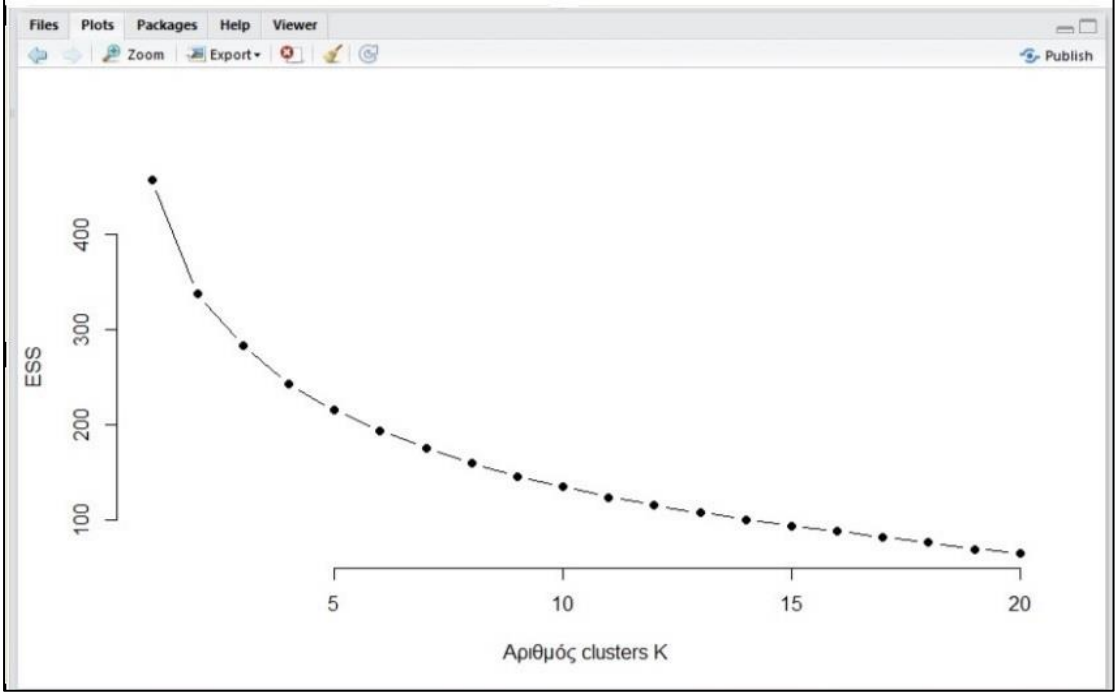
```
> #ΜΕΘΟΔΟΣ ELBOW
>
> data <- D3
>
> set.seed(123)
>
> # Υπολογισμός του ESS συναρτήσει του αριθμού K clusters, για K από 2 έως 20
>
> k_max <- 20
>
> ESS <- sapply(1:k.max,  function(k){kmeans(data, k, nstart=50,iter.max = 15 )$tot.withinss})
>
> ESS
 [1] 456.26573 336.79975 282.87360 242.61355 214.88677 194.01719 175.26352 159.46904 146.03690 134.69828 123.89366 115.40484
[13] 107.81183 100.75718  94.15848  88.79244  82.13737  76.32987  69.50973  65.71861
```

Εικόνα 36*: Κατασκευή της συνάρτησης του ESS με μεταβλητή των αριθμό των συστάδων και εξαγωγή αποτελεσμάτων για K= 2 έως 20 συστάδες*

Είναι προφανές πως όσο αυξάνεται το K, το ESS μειώνεται, όπως άλλωστε περιμέναμε.

Η μείωση του ESS ως αποτέλεσμα της αύξησης των συστάδων φαίνεται και στην γραφική παράσταση που ακολουθεί, μαζί με την αντίστοιχη εντολή στην R για την εξαγωγή της.

```
plot(1:k.max, ESS,
                type="b", pch = 19, frame = FALSE,
                xlab="Αριθμός clusters K",
                ylab="ESS")
```



Εικόνα 37*: Γραφική παράσταση του ESS για 2-20 clusters*

Στο διάγραμμα φαίνεται πως για τουλάχιστον K=2 clusters, το ESS παρουσιάζει αρκετή μείωση (από περίπου 450 → σε περίπου 330). Από εκεί και πέρα, η αύξηση του K συνεπάγεται περεταίρω μείωση του ESS, όχι όμως με τον ίδιο ρυθμό. Επομένως θα πρέπει να χρησιμοποιήσουμε τουλάχιστον 2 ή 3 συστάδες για να ομαδοποιήσουμε τα δεδομένα μας, όχι όμως πολύ περισσότερες, διότι έτσι θα δημιουργηθούν περισσότερα clusters από αυτά που είναι απαραίτητα, δηλαδή θα προκύψουν υποομάδες μέσα στις βασικές συστάδες, αυξάνοντας την πολυπλοκότητα του αλγορίθμου και τις υπολογιστικές απαιτήσεις. Επομένως δεν υπάρχει λόγος να επεκταθούμε πολύ περισσότερο από τις 2-3 συστάδες, καθώς μετά η ανάλυση γίνεται πιο πολύπλοκη, χωρίς ιδιαίτερο λόγο, αφού η ακρίβεια της ανάλυσης δεν αλλάζει τόσο σημαντικά. Με άλλα λόγια, για περεταίρω αύξηση του K, η μικρή βελτίωση στην ακρίβειας της μεθόδου, δεν αντισταθμίζει την πολυπλοκότητα της ανάλυσης και τις αυξημένες υπολογιστικές απαιτήσεις. Για να το επιβεβαιώσουμε αυτό, θα εξετάσουμε και άλλους τρόπους εκτίμησης του βέλτιστου K.

2η Μέθοδος: Εύρεση του βέλτιστου αριθμού Συστάδων με τη βοήθεια της μεθόδου ‘‘Mclust’’

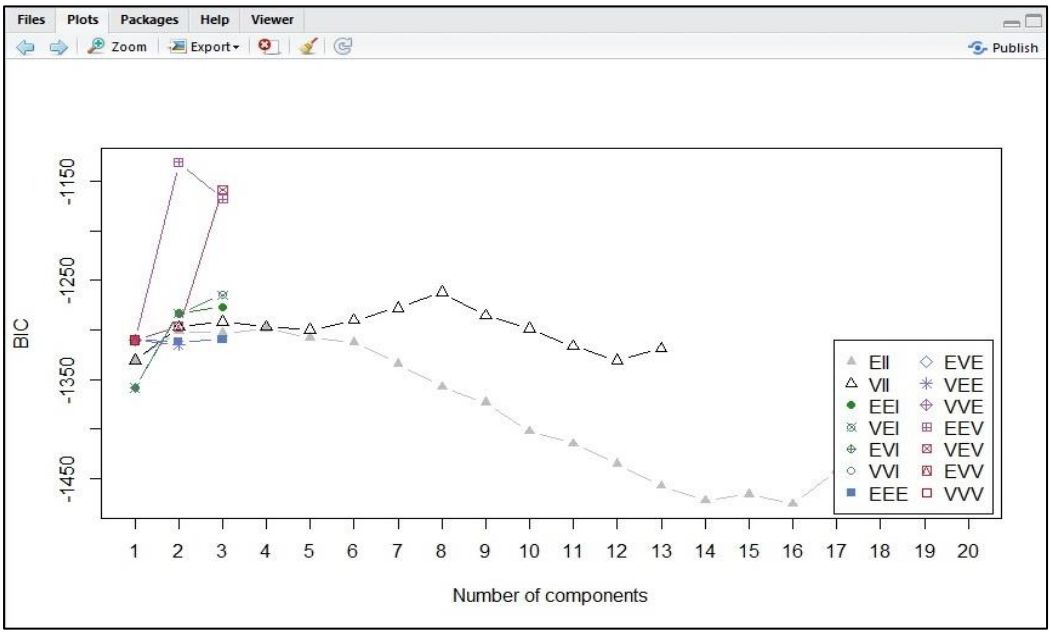
Θα πρέπει να εγκαταστήσουμε το πακέτο ‘‘mclust’’ και να χρησιμοποιήσουμε τη μέθοδο Mclust που αυτό περιέχει. Η μέθοδος αυτή θα μας δώσει το βέλτιστο K με βάση το κριτήριο μοντελοποίησης BIC (Bayesian Inference Criterion) το οποίο έχει εξαχθεί από την ιεραρχική ανάλυση συστάδων σε σύνθετα παραμετροποιημένα μοντέλα Gauss. Οι εντολές που θα χρησιμοποιήσουμε στην R παρουσιάζονται στην επόμενη εικόνα:

```
> library (mclust)
    __  _____________  __  ___________
   /  |/  / ____/ / /  / / / / ___/ _ |/ /
  / /|_/ / /   / /  / / / / /\__ \ / /
 / /  / / /___/ /___/ /_/ / /___/ // /
/_/  /_/\_____/_____/\____/_____//_/    version 5.2.3
Type 'citation("mclust")' for citing this R package in publications.
Warning message:
package 'mclust' was built under R version 3.2.5
>
> #Ελεγχος για ευερση βέλτιστου αριθμού clusters, από 2 έως 20
>
> #Ελεγχος για εύρεση βέλτιστου αριθμού clusters, από 2 έως 20
>
> K_clusters <- Mclust(as.matrix(data) , G=1:20 )
>
> K_best <- dim( K_clusters$z)
>
> cat("Βέλτιστος αριθμός cluster:", K_best, )
Error in cat("Βέλτιστος αριθμός cluster:", K_best, ) :
  argument is missing, with no default
> cat("Βέλτιστος αριθμός cluster:", K_best )
Βέλτιστος αριθμός cluster: 57 2
```

Εικόνα 38*: Εύρεση βέλτιστου αριθμού clusters με τη βοήθεια της μεθόδου Mclust*

<u>Επομένως η μέθοδος Mclust μας προτείνει ως βέλτιστο αριθμό cluster, K=2</u>.

Κάνοντας plot τα αποτελέσματα της μεθόδου Mclust με βάση το κριτήριο BIC, παίρνουμε το εξής γράφημα:



Εικόνα 39*: Γραφικά αποτελέσματα της μεθόδου Mclust με κριτήριο το BIC*

Στην ουσία κάθε μοντέλο που περιέχεται μέσα στη μέθοδο Mclust προτείνει έναν δικό του βέλτιστο αριθμό K. Στο διάγραμμα φαίνεται ότι για όλα τα μοντέλα, εκτός του VII, το κριτήριο της BICμεγιστοποιείται για 2 ή 3 cluster. Επομένως επιβεβαιώνεται και το αποτέλεσμα της μεθόδου ELBOW που εξετάσαμε πριν.

3η Μέθοδος: Γραφική Απεικόνιση των Συστάδων

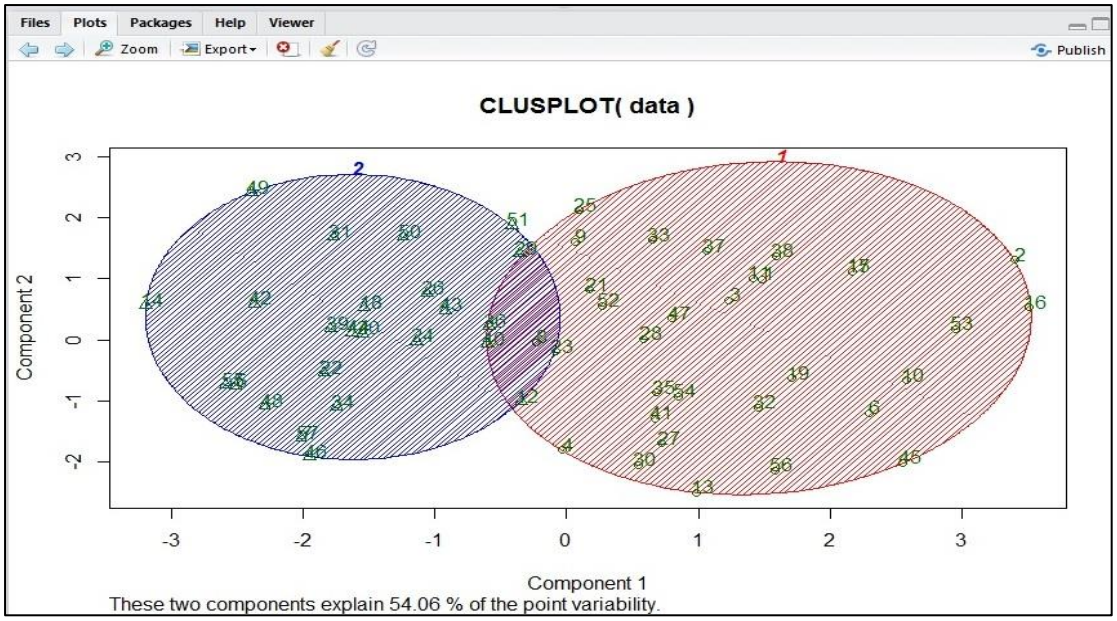
Στη σημείο αυτό θα εφαρμόσουμε τη μέθοδο k-means στα δεδομένα μας, τα οποία έχουμε καταχωρήσει στη μεταβλητή ‘‘ data’’. Η μέθοδος θα εφαρμοστεί μία φορά για την ομαδοποίηση των δεδομένων σε 2 συστάδες και μία για την ομαδοποίηση τους σε 3 συστάδες και θα απεικονίσουμε γραφικά αυτές τις 2 ομαδοποιήσεις, ώστε να δούμε το κατά πόσο γίνεται πιο ακριβής η ανάλυση με την αύξηση του K. Ακολουθούν εικόνες με την ομαδοποίηση των δεδομένων μας σε 2 ή 3 clusters σε δισδιάστατη απεικόνιση και οι αντίστοιχες εντολές.

```
> Model_fit2 <- kmeans(data,2)
>
> Model_fit2
K-means clustering with 2 clusters of sizes 32, 25

Cluster means:
          X1         X2         X3          X4         X5         X6         X7          X8
1 -0.2137699 -0.6999274 -0.3460229 -0.02246329 -0.6570759 -0.6022006 -0.2285262 -0.15475834
2  0.2759185  1.0309162  0.3905743 -0.12913645  0.9379329  0.7569559  0.3122377  0.04170336

Clustering vector:
 [1] 1 1 1 1 2 1 2 1 1 1 1 2 1 2 1 1 1 2 1 2 1 2 1 2 1 2 1 1 2 1 2 1 1 2 1 2 1 1 2 2 1 2 1 2 2 2 2 1 1 1 2 1 2

Within cluster sum of squares by cluster:
[1] 210.6397 126.1600
 (between_SS / total_SS =  26.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
> clusplot(data, Model_fit2$cluster, color=TRUE, shade=TRUE,
+          labels=2, lines=0)
```

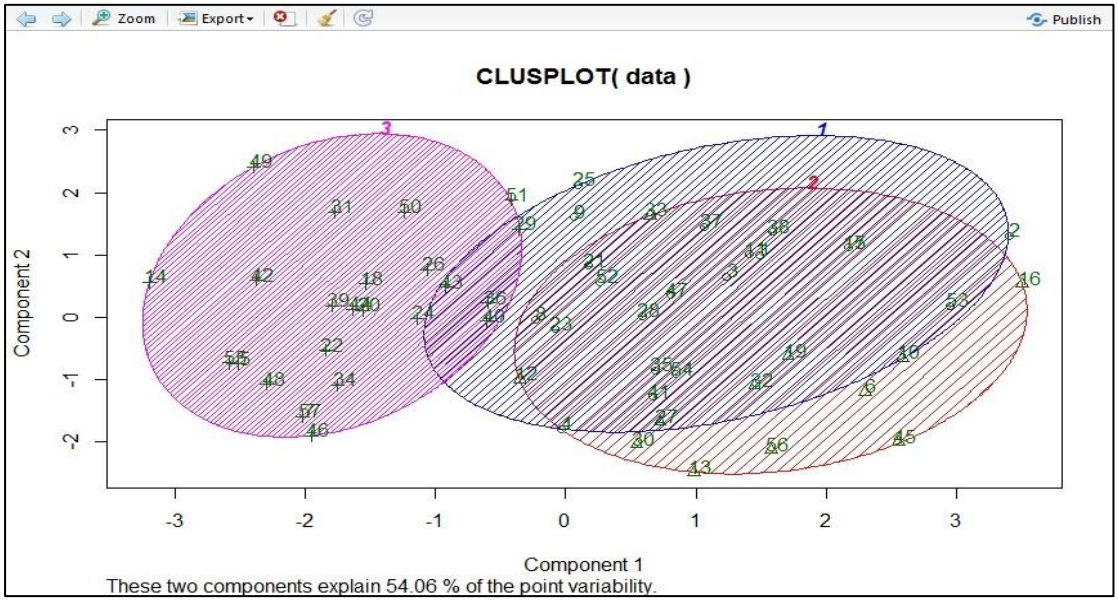
Εικόνα 40*: Ομαδοποίηση των δεδομένων σε 2 συστάδες και παρουσίαση των ανεξάρτητων μεταβλητών X1…X8 που χαρακτηρίζουν την κάθε ομάδα*

```
> Model_fit3 <- kmeans(data,3)
>
> Model_fit3
K-means clustering with 3 clusters of sizes 21, 12, 24

Cluster means:
          X1         X2          X3         X4         X5         X6          X7         X8
1 -0.4995358 -0.6412524 -0.50304380  0.4792174 -0.4217052 -0.4198588 -0.37880026  0.3909686
2  0.3996742 -0.7240876 -0.05855385 -0.9290091 -1.0070832 -0.8037363  0.09613993 -1.3068486
3  0.2396453  1.0637742  0.41492457 -0.1192789  0.9734458  0.7548062  0.30392627  0.1484233

Clustering vector:
 [1] 1 1 1 1 3 2 3 1 1 2 1 2 2 3 1 2 1 3 2 3 1 3 1 3 1 3 2 1 3 2 3 2 2 3 1 3 1 1 3 3 1 3 3 3 2 3 1 3 3 3 3 1 1 1 3 2 3

Within cluster sum of squares by cluster:
[1]  92.89806  78.59610 116.67433
 (between_SS / total_SS =  36.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
> clusplot(data, Model_fit3$cluster, color=TRUE, shade=TRUE,
+                    labels=2, lines=0)
```

Εικόνα 41*: Ομαδοποίηση των δεδομένων σε 3 συστάδες και παρουσίαση των ανεξάρτητων μεταβλητών X1…X8 που χαρακτηρίζουν την κάθε ομάδα*



***Εικόνα 42**: Δισδιάστατη απεικόνιση ομαδοποίησης των δεδομένων μας σε 2 συστάδες*



***Εικόνα 43**: Δισδιάστατη απεικόνιση ομαδοποίησης των δεδομένων μας σε 3 συστάδες*

Παρατηρώντας τις 2 απεικονίσεις, βλέπουμε ότι τα δεδομένα μας λόγω και του μικρού σχετικά πλήθους τους, ομαδοποιούνται ικανοποιητικά σε 2 συστάδες. Εάν πάμε σε περαιτέρω αύξηση των συστάδων δημιουργούνται υποομάδες πέραν των απαραίτητων ομάδων, με αποτέλεσμα να αυξάνεται η πολυπλοκότητα της ανάλυσης, χωρίς ωστόσο να αυξάνεται τόσο σημαντικά η ακρίβεια της.

Λαμβάνοντας λοιπόν υπόψη τα αποτελέσματα των 3 μεθόδων που παρουσιάστηκαν, καταλήγουμε πως η ανάλυση μας θα βασιστεί σε 2 συστάδες.

Στην επόμενη Εικόνα φαίνεται η κατανομή των 57 νοικοκυριών του δείγματος μας στα 2 cluster. Όπως φαίνεται και από το clustering vector στην Εικόνα 6.8, 32 από αυτά ανήκουν στο cluster 1 και 25 στο cluster 2.

ID	Cluster	Y mean
D14IA023010L	1	8304
D12NA111684K	1	2299
D12NA055486K	1	8388
C12FA154674R	1	18669
I14FA021789A	1	1710
C12UA215866D	1	6956
I13FA056611B	1	5921
D12NA080070L	1	4462
C12FA152684L	1	7833
I14FA043963O	1	11545
D13IA577080J	1	4690
C12UA224149I	1	1714
D12TA581630Z	1	6032
D12NA031884U	1	6182
C13UA198207D	1	7650
D13IA514382E	1	8914
C12UA115931P	1	12763
D14IA046257C	1	8992
D14IA023473M	1	9003
I14FA010583F	1	2780
I14FA023382P	1	2193
C14FA491863M	1	8354
D12TA574423B	1	7444
D13IA514808J	1	6793
C13FA252179O	1	4041
I15FA004337N	1	10422
C12UA215140Y	1	2626
I15FA037976N	1	2168
D15IA029133C	1	7453
C13FA251475P	1	2593
I14FA055790Y	1	3821
C12FA154530C	1	6489
	Μέση Μηνιαία Κατανάλωση cluster	6538

ID	Cluster	Y mean
D14IA053695N	2	9364
D12NA055665L	2	10693
C12UA223429J	2	6831
C12UA115300Q	2	7112
C12FA154665Q	2	8485
C14FA107301W	2	7106
C12UA218724C	2	13657
D13IA591157Z	2	14565
I13FA049239I	2	5224
C14FA107308D	2	7971
D12TA539595U	2	7342
D13IA514470D	2	11163
I13FA033530F	2	8105
I14FA088018W	2	6793
I14FA069949Y	2	9575
D15IA030006R	2	11103
D15IA039508S	2	8837
D15TA079300I	2	11550
D13IA590120I	2	13757
C15FA223026F	2	3265
D15IA029394V	2	13928
D15IA046023V	2	8765
D15IA001849P	2	9551
C15FA219301Y	2	6923
I14FA096902M	2	13598
	Μέση Μηνιαία Κατανάλωση	9410

Πίνακας 4*: Κατανομή των 57 νοικοκυριών στα 2 cluster και υπολογισμός μέσης μηνιαίας κατανάλωσης κάθε cluster*

6.3 Ομαδοποίηση στα cluster των νοικοκυριών προς έλεγχο

Το πώς κατανέμονται τα 57 νοικοκυριά στις 2 συστάδες φαίνεται στο Clustering Vector που μας έδωσε η R. Τώρα πρέπει να αποφασίσουμε πώς θα κατανεμηθούν τα άλλα 10 νοικοκυριά που αφήσαμε εκτός για να κάνουμε την πρόβλεψη της κατανάλωσης τους. Όπως έχουμε αναφέρει, το κριτήριο μας θα είναι η απόσταση των ανεξάρτητων μεταβλητών κάθε νοικοκυριού, από το αρχείο με τα scaled data, από τις ανεξάρτητες μεταβλητές κάθε συστάδας. Όπως έχει αναφερθεί στο Κεφάλαιο 3, ως απόσταση θα ορίσουμε τη ρίζα του συνολικού αθροίσματος των τετραγώνων κάθε ανεξάρτητης μεταβλητής ενός νοικοκυριού, μείον την αντίστοιχη ανεξάρτητη μεταβλητής της συστάδας.

Η απόσταση επομένως ενός νοικοκυριού με scaled ανεξάρτητες μεταβλητές $X_1,X_2,\ldots X_8$ από μία συστάδα με scaled ανεξάρτητες μεταβλητές $Xc_1,Xc_2,\ldots Xc_8$ θα ισούται με:

$$d = \sqrt{(X_1 - Xc_1)^2 + (X_2 - Xc_2)^2 + \ldots + (X_8 - Xc_8)^2}$$

Στην επόμενη Εικόνα φαίνονται οι scaled ανεξάρτητες μεταβλητές των 10 νοικοκυριών, καθώς και η απόστασή τους από τα 2 cluster (d1, d2).

	X1	X2	X3	X4	X5	X6	X7	X8	X1	X2	X3	X4	X5	X6	X7	X8		
Νοικοκυριό 1	1,1464757	0,2423241	-0,1938334	1,3241533	0,08562246	0,8085490	1,7584306	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	d1	3,3708
	1,1464757	0,2423241	-0,1938334	1,3241533	0,08562246	0,8085490	1,7584306	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	**d2**	2,6031
Νοικοκυριό 2	-1,4139867	-0,9173700	1,4295212	-1,2106544	-0,7339068	-0,4812792	0,5117126	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	d1	2,6303
	-1,4139867	-0,9173700	1,4295212	-1,2106544	-0,7339068	-0,4812792	0,5117126	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	d2	3,6593
Νοικοκυριό 3	-1,4139867	-0,9173700	-0,1938334	1,3241533	-1,5534360	-0,4812792	-1,3583644	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	**d1**	2,391
	-1,4139867	-0,9173700	-0,1938334	1,3241533	-1,5534360	-0,4812792	-1,3583644	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	d2	4,445
Νοικοκυριό 4	-0,1337555	-0,9173700	1,4295212	0,4792174	-0,7339068	-0,4812792	-1,3583644	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	**d1**	2,2478
	-0,1337555	-0,9173700	1,4295212	0,4792174	-0,7339068	-0,4812792	-1,3583644	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	d2	3,5573
Νοικοκυριό 5	1,1464757	0,2423241	1,4295212	1,3241533	0,08562246	0,8085490	1,1350716	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	d1	3,5218
	1,1464757	0,2423241	1,4295212	1,3241533	0,08562246	0,8085490	1,1350716	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	**d2**	2,4697
Νοικοκυριό 6	-0,1337555	-0,9173700	-0,1938334	1,3241533	1,7246810	0,8085490	-0,7350054	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	d1	3,1792
	-0,1337555	-0,9173700	-0,1938334	1,3241533	1,7246810	0,8085490	-0,7350054	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	**d2**	2,8736
Νοικοκυριό 7	-0,1337555	-0,9173700	-1,8171880	-1,2106544	-1,5534360	-1,7711074	-1,3583644	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	**d1**	2,7155
	-0,1337555	-0,9173700	-1,8171880	-1,2106544	-1,5534360	-1,7711074	-1,3583644	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	d2	5,0516
Νοικοκυριό 8	1,1464757	1,4020183	-0,1938334	-0,3657185	0,90515171	0,8085490	0,5117126	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	d1	3,4185
	1,1464757	1,4020183	-0,1938334	-0,3657185	0,90515171	0,8085490	0,5117126	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	**d2**	1,2077
Νοικοκυριό 9	-1,4139867	0,2423241	-0,1938334	0,4792174	0,08562246	0,8085490	-0,7350054	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	**d1**	2,3873
	-1,4139867	0,2423241	-0,1938334	0,4792174	0,08562246	0,8085490	-0,7350054	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	d2	2,4773
Νοικοκυριό 10	1,1464757	-0,9173700	-0,1938334	0,4792174	-0,7339068	-0,4812792	1,1350716	0,3909686	-0,2137699	-0,69992740	-0,3460229	-0,02246329	-0,6570759	-0,6022006	-0,2285262	-0,15475834	**d1**	2,0857
	1,1464757	-0,9173700	-0,1938334	0,4792174	-0,7339068	-0,4812792	1,1350716	0,3909686	0,2759185	1,0309162	0,3905743	-0,12913645	0,9379329	0,7569559	0,3122377	0,04170336	d2	3,2238

Cluster 1 Components
Cluster 2 Components
Shorter Distance

Πίνακας 5*: Αποστάσεις κάθε νοικοκυριού από τα 2 cluster και επιλογή της μικρότερης απόστασης*

Από την επιλογή των ελάχιστων αποστάσεων, βλέπουμε πως στο cluster 1 ανήκουν 6 νοικοκυριά και στο cluster 2 ανήκουν 4 νοικοκυριά. Επομένως και τα 6 νοικοκυριά του πρώτου cluster, θεωρούμε πως θα έχουν την ίδια κατανάλωση, δηλαδή τη μέση κατανάλωση του cluster,η οποία θα είναι 6538 λίτρα. Ομοίως και τα άλλα 4 νοικοκυριά του δεύτερου cluster, θεωρούμε πως θα έχουν την ίδια κατανάλωση, δηλαδή τη μέση κατανάλωση του cluster,η οποία θα είναι 9410 λίτρα. Επομένως η πρόβλεψη μας θα υπαγορεύει πως τα νοικοκυριά που ανήκουν στο Cluster 1 θα έχουν όλο το χρόνο μηνιαία κατανάλωση ίση με 6538 λίτρα και τα νοικοκυριά που ανήκουν στο cluster 2 θα έχουν μηνιαία κατανάλωση ίση με 9410.

ΚΕΦΑΛΑΙΟ 7: ΑΞΙΟΛΟΓΗΣΗ ΑΚΡΙΒΕΙΑΣ ΠΡΟΒΛΕΨΕΩΝ ΓΙΑ ΤΙΣ 2 ΜΕΘΟΔΟΥΣ

7.1 Εισαγωγή

Εφόσον έχουμε κατασκευάσει τα 2 μοντέλα πρόβλεψης και έχουμε κάνει τις προβλέψεις για τη μηνιαία κατανάλωση των 10 νοικοκυριών, είμαστε σε θέση να υπολογίσουμε το σφάλμα πρόβλεψης για κάθε μέθοδο, δηλαδή τη διαφορά ανάμεσα στις τιμές που προέβλεψε κάθε μοντέλο και στις πραγματικές τιμές των καταναλώσεων.

Για καθεμία από τις 2 μεθόδους, έχουμε συνολικά 2 στήλες με 260 γραμμές η καθεμία. Η μία στήλη **F_i** περιέχει τις προβλέψεις κατανάλωσης που έγιναν για καθένα από τα 10 νοικοκυριά για σύνολο 26 μηνών (Ιανουάριος 2015-Φεβρουάριος 2017) ενώ η στήλη **Y_i** περιέχει τις αντίστοιχες πραγματικές τιμές. Με βάση τις 2 αυτές στήλες, από το σφάλμα $Y_i - F_i$ για καθεμία από τις 260 γραμμές θα υπολογίσουμε για κάθε μέθοδο το **Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error –MAPE).**

$$\mathrm{MAPE} = \frac{100}{N} \times \sum_{i=1}^{N} \left| \frac{Y_i - F_i}{Y_i} \right|$$

Όπως έχουμε αναφέρει, όταν θέλουμε να υπολογίσουμε ένα σφάλμα, είναι προτιμότερο αυτό να εκφράζεται σε μορφή ποσοστού ώστε να αξιολογείται καλύτερα, παρά να εκφράζεται στη μονάδα μέτρησης της εξαρτημένης μεταβλητής Y, καθώς τότε εξαρτάται από τη φύση του προβλήματος εάν το η τιμή του σφάλματος βρίσκεται σε αποδεκτά όρια

7.2 Υπολογισμός Σφάλματος Πρόβλεψης με την Πολλαπλή Γραμμική Παλινδρόμηση

Στο Κεφάλαιο 5, έπειτα από διάφορες δοκιμές καταλήξαμε στο βέλτιστο γραμμικό μοντέλο, το οποίο δίνεται από τη σχέση:

$$\mathbf{Y = -1547{,}22 \cdot X_2 + 1008{,}93 \cdot X_3 + 395{,}16 \cdot X_4 + 1920{,}79 \cdot X_5 + 592{,}52 \cdot X_6 + 532{,}92 \cdot X_7 - 625{,}18 \cdot X_8}$$

Στις ανεξάρτητες μεταβλητές αντικαθιστούμε την τιμή του αντίστοιχου δημογραφικού χαρακτηριστικού του νοικοκυριού.

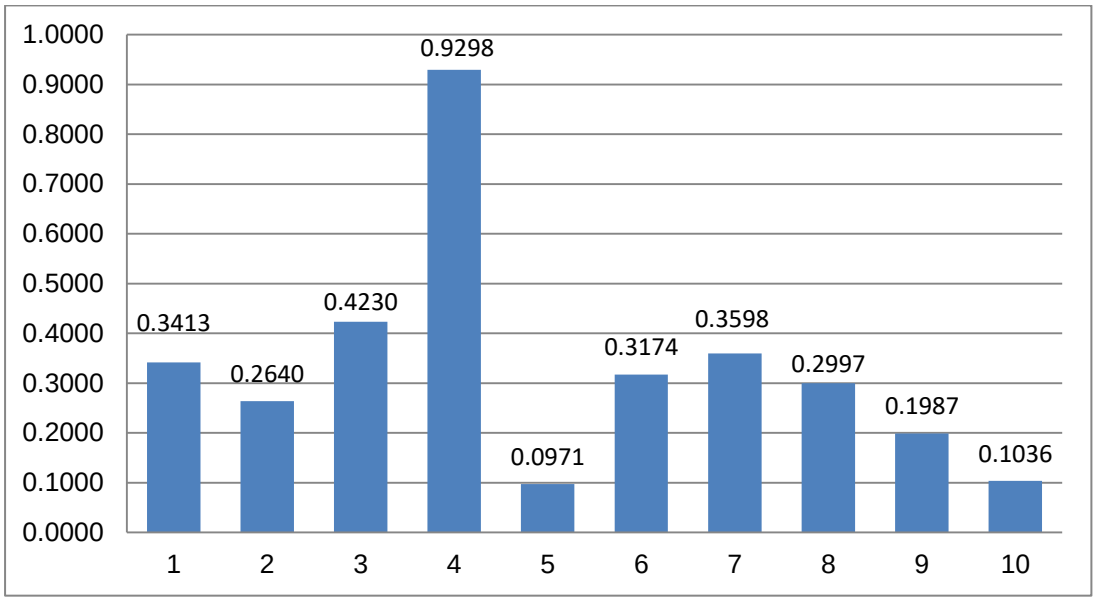
Όπως έχουμε πει και φαίνεται και από την παραπάνω σχέση, το βέλτιστο γραμμικό μοντέλο δεν περιέχει την ανεξάρτητη μεταβλητή που εκφράζει τον μήνα, επομένως το μοντέλο μας προβλέπει ότι κάθε νοικοκυριό θα έχει την ίδια κατανάλωση για όλους τους μήνες του έτους. Αυτό βέβαια δεν μπορεί να ισχύει στην πραγματικότητα, ωστόσο αυτό που μας ενδιαφέρει είναι η τιμή του MAPE συνολικά η οποία θέλουμε να είναι όσο το δυνατόν μικρότερη. Στην επόμενη Εικόνα φαίνεται ο Πίνακας με τα δημογραφικά χαρακτηριστικά και τις στήλες Y_i και F_i για ένα νοικοκυριό για το διάστημα των 26 μηνών που μελετάμε. Ομοίως έχουν συμπληρωθεί οι αντίστοιχες στήλες για τα υπόλοιπα 9 νοικοκυριά.

ID	X1	X2	X3	X4	X5	X6	X7	X8	Yi		Fi
D12NA073420U	3	2	2	4	3	3	6	2	8145		9991
D12NA073420U	3	2	2	4	3	3	6	2	8804		9991
D12NA073420U	3	2	2	4	3	3	6	2	10247		9991
D12NA073420U	3	2	2	4	3	3	6	2	10483		9991
D12NA073420U	3	2	2	4	3	3	6	2	8587		9991
D12NA073420U	3	2	2	4	3	3	6	2	9223		9991
D12NA073420U	3	2	2	4	3	3	6	2	10539		9991
D12NA073420U	3	2	2	4	3	3	6	2	10076		9991
D12NA073420U	3	2	2	4	3	3	6	2	8869		9991
D12NA073420U	3	2	2	4	3	3	6	2	7546		9991
D12NA073420U	3	2	2	4	3	3	6	2	8201		9991
D12NA073420U	3	2	2	4	3	3	6	2	8568		9991
D12NA073420U	3	2	2	4	3	3	6	2	7968		9991
D12NA073420U	3	2	2	4	3	3	6	2	7440		9991
D12NA073420U	3	2	2	4	3	3	6	2	8312		9991
D12NA073420U	3	2	2	4	3	3	6	2	7975		9991
D12NA073420U	3	2	2	4	3	3	6	2	8286		9991
D12NA073420U	3	2	2	4	3	3	6	2	7822		9991
D12NA073420U	3	2	2	4	3	3	6	2	10634		9991
D12NA073420U	3	2	2	4	3	3	6	2	7553		9991
D12NA073420U	3	2	2	4	3	3	6	2	6461		9991
D12NA073420U	3	2	2	4	3	3	6	2	4765		9991
D12NA073420U	3	2	2	4	3	3	6	2	4330		9991
D12NA073420U	3	2	2	4	3	3	6	2	6242		9991
D12NA073420U	3	2	2	4	3	3	6	2	5434		9991
D12NA073420U	3	2	2	4	3	3	6	2	5018		9991

***Πίνακας 6**: Πίνακας με τα δημογραφικά χαρακτηριστικά και τις τιμές Yi και Fi για ένα από τα 10 νοικοκυριά, με βάση τα οποία θα αξιολογήσουμε την ακρίβεια πρόβλεψης του γραμμικού μοντέλου*

<u>Το συνολικό σφάλμα πρόβλεψης MAPE προκύπτει ότι είναι ίσο με 33,34%.</u>

Για να δούμε τη συνεισφορά καθενός από τα 10 νοικοκυριά στην τιμή του MAPE, υπολογίζουμε το MAPE καθενός νοικοκυριού ξεχωριστά.



***Εικόνα 44**: Γράφημα που απεικονίζει τη συνεισφορά του κάθε νοικοκυριού στη διαμόρφωση της τιμής του συνολικού MAPE*

Βλέπουμε πως τη μεγαλύτερη συνεισφορά στο συνολικό MAPE την έχει το 4ο νοικοκυριό, ενώ οι τιμές του MAPE για τα υπόλοιπα νοικοκυριά ξεχωριστά κυμαίνονται από περίπου 10% έως 42%.

Στην επόμενη Εικόνα φαίνονται οι πραγματικές τιμές και οι τιμές που προβλέφθηκαν για την κατανάλωση του 4ου νοικοκυριού

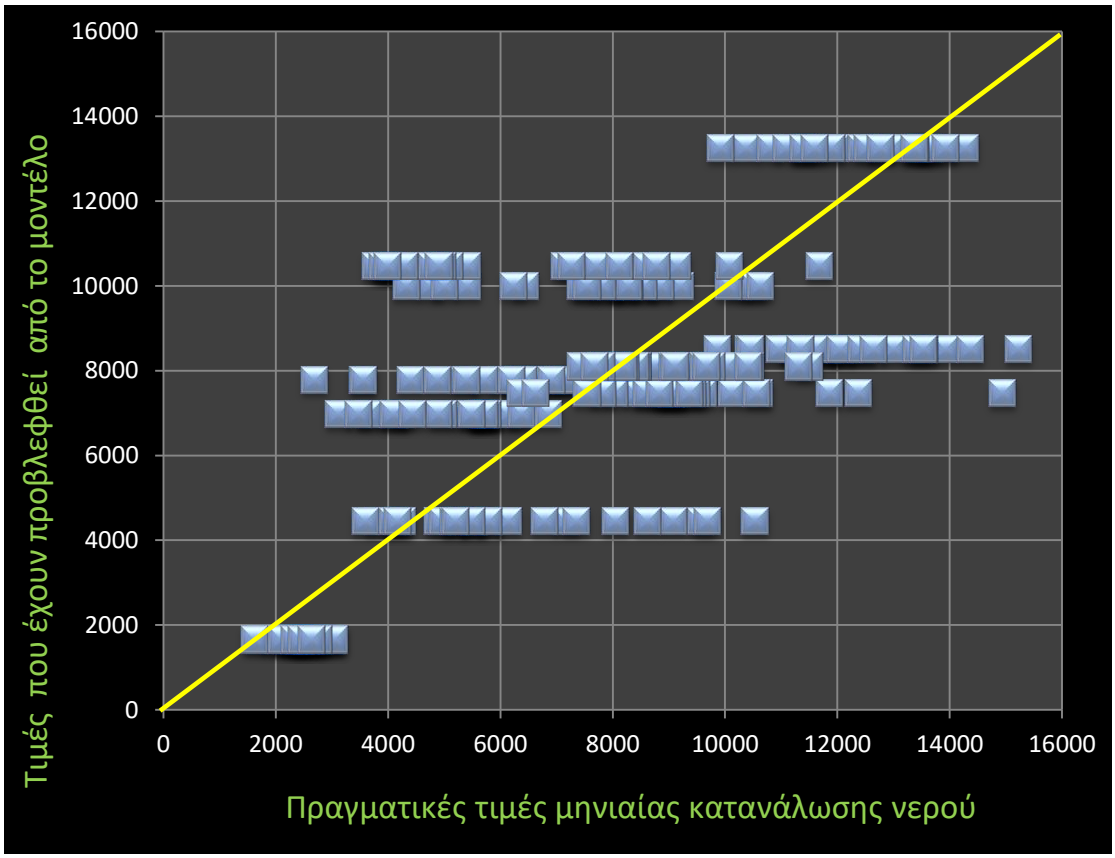
I14FA079412B	3	2	3	4	3	3	5	2	1	5425		10467
I14FA079412B	3	2	3	4	3	3	5	2	2	8612		10467
I14FA079412B	3	2	3	4	3	3	5	2	3	8893		10467
I14FA079412B	3	2	3	4	3	3	5	2	4	10081		10467
I14FA079412B	3	2	3	4	3	3	5	2	5	11690		10467
I14FA079412B	3	2	3	4	3	3	5	2	6	9172		10467
I14FA079412B	3	2	3	4	3	3	5	2	7	7724		10467
I14FA079412B	3	2	3	4	3	3	5	2	8	8797		10467
I14FA079412B	3	2	3	4	3	3	5	2	9	7160		10467
I14FA079412B	3	2	3	4	3	3	5	2	10	7276		10467
I14FA079412B	3	2	3	4	3	3	5	2	11	5102		10467
I14FA079412B	3	2	3	4	3	3	5	2	12	4011		10467
I14FA079412B	3	2	3	4	3	3	5	2	1	3958		10467
I14FA079412B	3	2	3	4	3	3	5	2	2	3789		10467
I14FA079412B	3	2	3	4	3	3	5	2	3	4641		10467
I14FA079412B	3	2	3	4	3	3	5	2	4	3910		10467
I14FA079412B	3	2	3	4	3	3	5	2	5	4919		10467
I14FA079412B	3	2	3	4	3	3	5	2	6	4882		10467
I14FA079412B	3	2	3	4	3	3	5	2	7	4789		10467
I14FA079412B	3	2	3	4	3	3	5	2	8	8135		10467
I14FA079412B	3	2	3	4	3	3	5	2	9	4971		10467
I14FA079412B	3	2	3	4	3	3	5	2	10	4126		10467
I14FA079412B	3	2	3	4	3	3	5	2	11	4094		10467
I14FA079412B	3	2	3	4	3	3	5	2	12	4920		10467
I14FA079412B	3	2	3	4	3	3	5	2	1	4305		10467
I14FA079412B	3	2	3	4	3	3	5	2	2	3998		10467

***Πίνακας 7**: Πίνακας με τις πραγματικές και εκτιμώμενες τιμές για το νοικοκυριό με τη μεγαλύτερη τιμή MAPE*

Από τον παραπάνω Πίνακα, φαίνεται πως ατομικά το συγκεκριμένο νοικοκυριό έχει τόσο μεγάλη τιμή MAPE καθώς δεν παρουσιάζει σταθερά επίπεδα κατανάλωσης, καθώς μπορούμε να δούμε πως για δεδομένα του, εμφανίζει είτε ιδιαίτερα υψηλή κατανάλωση την άνοιξη του 2015 είτε ιδιαίτερα χαμηλή κατανάλωση από τον Ιανουάριο του 2016 και μετά.

Συμπέρασμα: Ενδεικτικά, εάν στην πρόβλεψη μας δε συμπεριλάβουμε το νοικοκυριό αυτό, η τιμή του συνολικού σφάλματος πρόβλεψης MAPE πέφτει στο 24% . Επομένως, βάσει της φύσης του προβλήματος και δεδομένου ότι πάντα θα υπάρχουν περιπτώσεις νοικοκυριών με ασταθή καταναλωτική συμπεριφορά, θεωρούμε ότι το γραμμικό μοντέλο μας πραγματοποίησε μία ικανοποιητική πρόβλεψη με τιμή του MAPE που κυμαίνεται σε αποδεκτά όρια.

Στην επόμενη εικόνα φαίνεται στο διάγραμμα διασποράς για πραγματικές και εκτιμώμενες τιμές με τη της πολλαπλής γραμμικής παλινδρόμησης:



***Εικόνα 45**: Διάγραμμα διασποράς για πραγματικές και εκτιμώμενες τιμές με τη μέθοδο k-means*

7.3 Υπολογισμός Σφάλματος Πρόβλεψης με τη μέθοδο k-means

Όπως είδαμε στο προηγούμενο Κεφάλαιο, οι 57 παρατηρήσεις μας ομαδοποιήθηκαν σε 2 συστάδες, 35 παρατηρήσεις στην πρώτη και 22 στη δεύτερη. Επίσης, βάση της ελάχιστης απόστασης, δηλαδή της πιο πανομοιότυπης συμπεριφοράς, από τα 10 νοικοκυριά που κρατήσαμε για να αξιολογήσουμε την πρόβλεψη της μεθόδου, εκτιμήσαμε ότι τα 6 θα ανήκουν στην πρώτη συστάδα και τα 4 στη δεύτερη συστάδα. Η πρόβλεψη μας υποθέτει πως η μηνιαία κατανάλωση κάθε νοικοκυριού θα είναι μόνιμα ίση με τη μέση μηνιαία κατανάλωση της συστάδας στην οποία ανήκει. Επομένως και τα 6 νοικοκυριά του πρώτου cluster, θεωρούμε πως θα έχουν την ίδια κατανάλωση, δηλαδή τη μέση κατανάλωση του cluster,η οποία θα είναι 6538 λίτρα. Ομοίως και τα άλλα 4 νοικοκυριά του δεύτερου cluster, θεωρούμε πως θα έχουν την ίδια κατανάλωση, δηλαδή τη μέση κατανάλωση του cluster,η οποία θα είναι 9410 λίτρα.

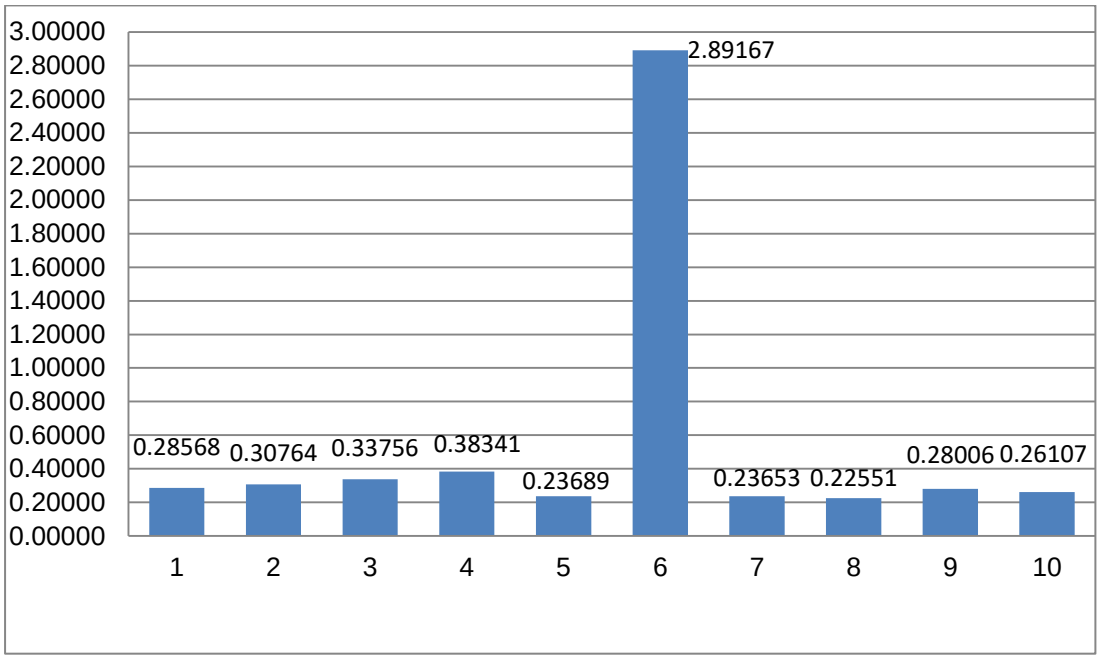
Στον επόμενο Πίνακα φαίνεται ενδεικτικά το ίδιο νοικοκυριό που παρουσιάστηκε και στην περίπτωση της γραμμικής παλινδρόμησης, με τα δημογραφικά χαρακτηριστικά και τις στήλες Y_i και F_i για το διάστημα των 26 μηνών που μελετάμε. Ομοίως έχουν συμπληρωθεί οι αντίστοιχες στήλες για τα υπόλοιπα 9 νοικοκυριά.

ID	X1	X2	X3	X4	X5	X6	X7	X8		Yi		Fi
D12NA073420U	3	2	2	4	3	3	6	2		8145		9410
D12NA073420U	3	2	2	4	3	3	6	2		8804		9410
D12NA073420U	3	2	2	4	3	3	6	2		10247		9410
D12NA073420U	3	2	2	4	3	3	6	2		10483		9410
D12NA073420U	3	2	2	4	3	3	6	2		8587		9410
D12NA073420U	3	2	2	4	3	3	6	2		9223		9410
D12NA073420U	3	2	2	4	3	3	6	2		10539		9410
D12NA073420U	3	2	2	4	3	3	6	2		10076		9410
D12NA073420U	3	2	2	4	3	3	6	2		8869		9410
D12NA073420U	3	2	2	4	3	3	6	2		7546		9410
D12NA073420U	3	2	2	4	3	3	6	2		8201		9410
D12NA073420U	3	2	2	4	3	3	6	2		8568		9410
D12NA073420U	3	2	2	4	3	3	6	2		7968		9410
D12NA073420U	3	2	2	4	3	3	6	2		7440		9410
D12NA073420U	3	2	2	4	3	3	6	2		8312		9410
D12NA073420U	3	2	2	4	3	3	6	2		7975		9410
D12NA073420U	3	2	2	4	3	3	6	2		8286		9410
D12NA073420U	3	2	2	4	3	3	6	2		7822		9410
D12NA073420U	3	2	2	4	3	3	6	2		10634		9410
D12NA073420U	3	2	2	4	3	3	6	2		7553		9410
D12NA073420U	3	2	2	4	3	3	6	2		6461		9410
D12NA073420U	3	2	2	4	3	3	6	2		4765		9410
D12NA073420U	3	2	2	4	3	3	6	2		4330		9410
D12NA073420U	3	2	2	4	3	3	6	2		6242		9410
D12NA073420U	3	2	2	4	3	3	6	2		5434		9410
D12NA073420U	3	2	2	4	3	3	6	2		5018		9410

***Πίνακας 8**: Πίνακας με τα δημογραφικά χαρακτηριστικά και τις τιμές Yi και Fi για ένα από τα 10 νοικοκυριά, με βάση τα οποία θα αξιολογήσουμε την ακρίβεια πρόβλεψης της μεθόδου k-means*

Το συνολικό σφάλμα πρόβλεψης MAPE προκύπτει ότι είναι ίσο με 54,46%.

Για να δούμε τη συνεισφορά καθενός από τα 10 νοικοκυριά στην τιμή του MAPE, υπολογίζουμε το MAPE καθενός νοικοκυριού ξεχωριστά.



***Εικόνα 46**: Γράφημα που απεικονίζει τη συνεισφορά του κάθε νοικοκυριού στη διαμόρφωση της τιμής του συνολικού MAPE*

Βλέπουμε πως τη μεγαλύτερη συνεισφορά στο συνολικό MAPE την έχει το 6ο νοικοκυριό, ενώ οι τιμές του MAPE για τα υπόλοιπα νοικοκυριά ξεχωριστά κυμαίνονται από περίπου 23% έως 38%.

Στην επόμενη Εικόνα φαίνονται οι πραγματικές τιμές και οι τιμές που προβλέφθηκαν για την κατανάλωση του 6ου νοικοκυριού:

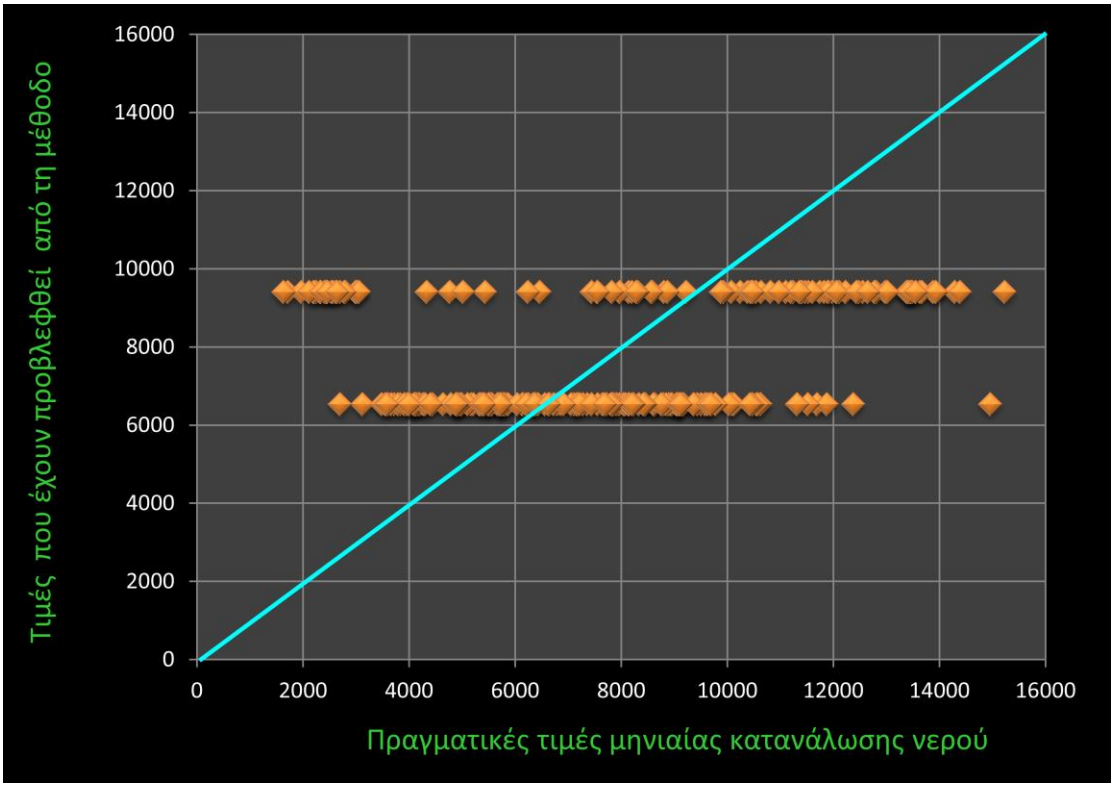
I13FA049641O	2	1	1	1	1	1	1	2		2988	9410
I13FA049641O	2	1	1	1	1	1	1	2		2699	9410
I13FA049641O	2	1	1	1	1	1	1	2		2533	9410
I13FA049641O	2	1	1	1	1	1	1	2		2568	9410
I13FA049641O	2	1	1	1	1	1	1	2		1716	9410
I13FA049641O	2	1	1	1	1	1	1	2		1969	9410
I13FA049641O	2	1	1	1	1	1	1	2		2801	9410
I13FA049641O	2	1	1	1	1	1	1	2		3026	9410
I13FA049641O	2	1	1	1	1	1	1	2		2620	9410
I13FA049641O	2	1	1	1	1	1	1	2		2198	9410
I13FA049641O	2	1	1	1	1	1	1	2		2253	9410
I13FA049641O	2	1	1	1	1	1	1	2		2617	9410
I13FA049641O	2	1	1	1	1	1	1	2		2323	9410
I13FA049641O	2	1	1	1	1	1	1	2		1633	9410
I13FA049641O	2	1	1	1	1	1	1	2		3008	9410
I13FA049641O	2	1	1	1	1	1	1	2		2214	9410
I13FA049641O	2	1	1	1	1	1	1	2		2118	9410
I13FA049641O	2	1	1	1	1	1	1	2		3057	9410
I13FA049641O	2	1	1	1	1	1	1	2		2777	9410
I13FA049641O	2	1	1	1	1	1	1	2		2429	9410
I13FA049641O	2	1	1	1	1	1	1	2		2449	9410
I13FA049641O	2	1	1	1	1	1	1	2		2409	9410
I13FA049641O	2	1	1	1	1	1	1	2		2341	9410
I13FA049641O	2	1	1	1	1	1	1	2		2564	9410
I13FA049641O	2	1	1	1	1	1	1	2		2466	9410
I13FA049641O	2	1	1	1	1	1	1	2		2642	9410

***Πίνακας 9**: Πίνακας με τις πραγματικές και εκτιμώμενες τιμές για το νοικοκυριό με τη μεγαλύτερη τιμή MAPE*

Από τον παραπάνω Πίνακα, φαίνεται πως ατομικά το συγκεκριμένο νοικοκυριό έχει τόσο μεγάλη τιμή MAPE καθώς σε γενικές γραμμές οι μηνιαίες καταναλώσεις του είναι αρκετά πιο χαμηλές από τη μέση τιμή μηνιαίας κατανάλωσης του 2^{ου} cluster στο οποίο ανήκει. Έτσι φαίνεται πως παρότι βάσει δημογραφικών χαρακτηριστικών έχει ομαδοποιηθεί σε ένα cluster, εντούτοις η καταναλωτική του συμπεριφορά δε συμβαδίζει με τα συμπεριφορά των υπόλοιπων νοικοκυριών του ίδιου cluster.

Συμπέρασμα: Ενδεικτικά, εάν στην πρόβλεψη μας δε συμπεριλάβουμε το νοικοκυριό αυτό, η τιμή του συνολικού σφάλματος πρόβλεψης MAPE πέφτει στο 25,54%, που είναι πολύ κοντά με το MAPE που υπολογίστηκε στη μέθοδο της γραμμικής παλινδρόμησης. Επομένως, βάσει της φύσης του προβλήματος και δεδομένου ότι πάντα θα υπάρχουν περιπτώσεις νοικοκυριών με ασταθή καταναλωτική συμπεριφορά, θεωρούμε ότι το γραμμικό μοντέλο μας πραγματοποίησε μία ικανοποιητική πρόβλεψη με τιμή του MAPE που κυμαίνεται σε αποδεκτά όρια.

Στην επόμενη εικόνα φαίνεται στο διάγραμμα διασποράς για πραγματικές και εκτιμώμενες τιμές με τη μέθοδο k-means:



***Εικόνα 47**: Διάγραμμα διασποράς για πραγματικές και εκτιμώμενες τιμές με τη μέθοδο k-means*

ΚΕΦΑΛΑΙΟ 8: ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΕΚΤΑΣΕΙΣ

Στο προηγούμενο Κεφάλαιο, όπου αξιολογήσαμε τις προβλέψεις των δύο προσεγγίσεων, είδαμε πως υπήρξαν μεμονωμένα νοικοκυριά τα οποία αύξησαν την τιμή του συνολικού MAPE, το οποίο σε περίπτωση τα νοικοκυριά αυτά δε συμμετείχαν στην πρόβλεψη, θα κυμαινόταν στο 25% περίπου. Το ποσοστό αυτό κινείται σε αποδεκτά πλαίσια για τη φύση του προβλήματος με το οποίο ασχοληθήκαμε. Ωστόσο, όπως και στη μελέτη μας, πάντα υπάρχει η πιθανότητα ύπαρξης νοικοκυριών με καταναλωτική συμπεριφορά που είναι δύσκολο να προβλεφθεί, η παρουσία των οποίων αυξάνει την τιμή MAPE. Αυτό όμως δεν είναι πρόβλημα των τεχνικών πρόβλεψης που παρουσιάσαμε, οι οποίες σε γενικές γραμμές προσομοίωσαν ικανοποιητικά τα δεδομένα μας και έκαναν προβλέψεις με αποδεκτά όρια σφαλμάτων.

Όπως εξηγήσαμε, σε καθένα από τα 2 μοντέλα υπήρξε μία περίπτωση νοικοκυριού, διαφορετική σε κάθε περίπτωση, η πραγματική κατανάλωση του οποίου απείχε αρκετά από την τιμή που είχε προβλεφθεί. Η εμφάνιση μεγάλου σφάλματος συνέβη σε κάθε περίπτωση για διαφορετικό λόγο και οφείλεται στον τρόπο με τον οποίο λειτουργεί κάθε μοντέλο. Στην περίπτωση της γραμμικής παλινδρόμησης το νοικοκυριό παρουσίασε πολύ ασταθή κατανάλωση στο διάστημα των 26 μηνών, εμφανίζοντας καταναλώσεις και πολύ χαμηλές αλλά και αρκετά υψηλές σε σχέση με τη μέση κατανάλωση του. Αυτό είχε σαν αποτέλεσμα να υπάρχουν τιμές της κατανάλωσης οι οποίες απείχαν αρκετά από τη βέλτιστη ευθεία των ελαχίστων τετραγώνων, η οποία δεν ήταν δυνατό να προσεγγίσει μια τόσο ασταθή συμπεριφορά. Στη περίπτωση του αλγορίθμου k-means, είχαμε το αντίθετο φαινόμενο, δηλαδή το νοικοκυριό το οποίο εκτίναξε την τιμή του MAPE είχε μεν σταθερά χαμηλή κατανάλωση, όμως απείχε αρκετά από τη μέση μηνιαία κατανάλωση του cluster στο οποίο ομαδοποιήθηκε. Ενώ δηλαδή βάσει δημογραφικών στοιχείων εμφάνιζε ομοιότητα με άλλα νοικοκυριά, παρόλα αυτά η κατανάλωση του ήταν πολύ μικρότερη από τη μέση κατανάλωση που είχαν τα υπόλοιπα νοικοκυριά της ομάδας.

Σε όλα τα παραπάνω ζητήματα έρχεται να προστεθεί και το γεγονός ότι δεν είχαμε πολύ μεγάλο αριθμό νοικοκυριών για την ανάλυση μας. Πιθανότατα η ύπαρξη περισσότερων νοικοκυριών με σχετικά ομαλή καταναλωτική συμπεριφορά να συντελούσε σε μία καλύτερη γραμμική εξίσωση, από την οποία θα προέκυπτε η ευθεία ελαχίστων τετραγώνων. Σε ό,τι αφορά την ανάλυση συστάδες, η ύπαρξη περισσότερων παρατηρήσεων ενδεχομένως να οδηγούσε σε περισσότερες και πιο καλά διαφοροποιημένες ομάδες, τα μέλη των οποίων θα είχαν μια πιο κοντινή τιμή κατανάλωσης με τη μέση τιμή κάθε ομάδας. Από την άλλη, θα μπορούσε κάποιος να ισχυριστεί πως όσα περισσότερα νοικοκυριά λαμβάνουμε υπόψη στην ανάλυση μας, τόσο αυξάνονται και οι πιθανότητες να εμφανιστούν νοικοκυριά με ασταθή συμπεριφορά τα οποία θα αυξάνουν την τιμή του συνολικού σφάλματος πρόβλεψης.

Με βάση όλα τα παραπάνω, η εντατική παρακολούθηση και καταγραφή των καταναλώσεων των κατοίκων διάφορων περιοχών κρίνεται ιδιαίτερα ενδιαφέρουσα και είναι βέβαιο πως όσο πιο πολλά δεδομένα καταγράφονται και μελετώνται, τόσο πιο χρήσιμα συμπεράσματα θα εξάγονται. Μια μελέτη θα μπορούσε πέρα από τα δημογραφικά στοιχεία να επεκταθεί και σε άλλους παράγοντες ανάμεσα σε διαφορετικές περιοχές, όπως το κλίμα, οι καιρικές συνθήκες κτλ. Η δική μας ανάλυση επικεντρώθηκε στα δημογραφικά χαρακτηριστικά, καθώς αφορούσε νοικοκυριά της ίδιας πόλης, επομένως θεωρήσαμε πως εάν η κατανάλωση επηρεάστηκε από την αλλαγή καιρικών συνθηκών, αυτό θα εμφανιζόταν ούτως ή άλλως στα δεδομένα μας με την αλλαγή των μηνών. Αξιολογώντας την ανάλυση που έγινε στην παρούσα εργασία και βλέποντας ότι έγινε μία σχετικά ικανοποιητική πρόβλεψη και με τις 2 μεθόδους χωρίς να έχουμε μεγάλος πλήθος διαθέσιμων δεδομένων, καταλαβαίνει εύκολα κάποιος πόσο ενδιαφέρον μπορεί να κρύβει η τακτική καταγραφή και μελέτη δεδομένων και ο εντοπισμός ιδιαίτερων περιπτώσεων όπως τα 2 νοικοκυριά που εντοπίσαμε, καθώς και πόσο πιο βελτιωμένες προβλέψεις μπορούν να πραγματοποιηθούν όσο το πλήθος των διαθέσιμων δεδομένων αυξάνεται.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- **An Introduction to Statistical Learning with applications in R** - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

- **The Use of Population Forecasting and Cluster Analysis in Water Utility Master Planning** -Lillian Pierson, April 2013

- **Cluster analysis of urban water supply and demand: Toward large-**scale comparative sustainability planning-Karen Noiva, John E. Fernández, James L. Wescoat Jr., 2016

- **Identifying Typical Urban Water Demand Patterns for a Reliable Short-Term Forecasting** – The Icewater Project Approach - A. Candelieri, F. Archetti, 2016

- **Exploring Patterns in Water Consumption by Clustering** – Chrysi Laspidou, Elpiniki Papageorgiou, Konstantinos Kokkinos, Sambit Sahu, Arpit Gupta, LeandrosTassiulas

- **Assessment of water use efficiency in the household using cluster analysis** - Catarina Jorge, Paula Vieira, Margarida Rebelo, Dídia Covas

- **Multivariate Statistical Analysis for Water Demand Modeling** - C. M. Fontanazza, V. Notaro, V. Puleo,G. Fren

- **Water demand pattern classification from smart meter data** - S.A. McKenna, F. Fusco, B.J. Eck