ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΠΛΗΡΟΦΟΡΙΚΗ

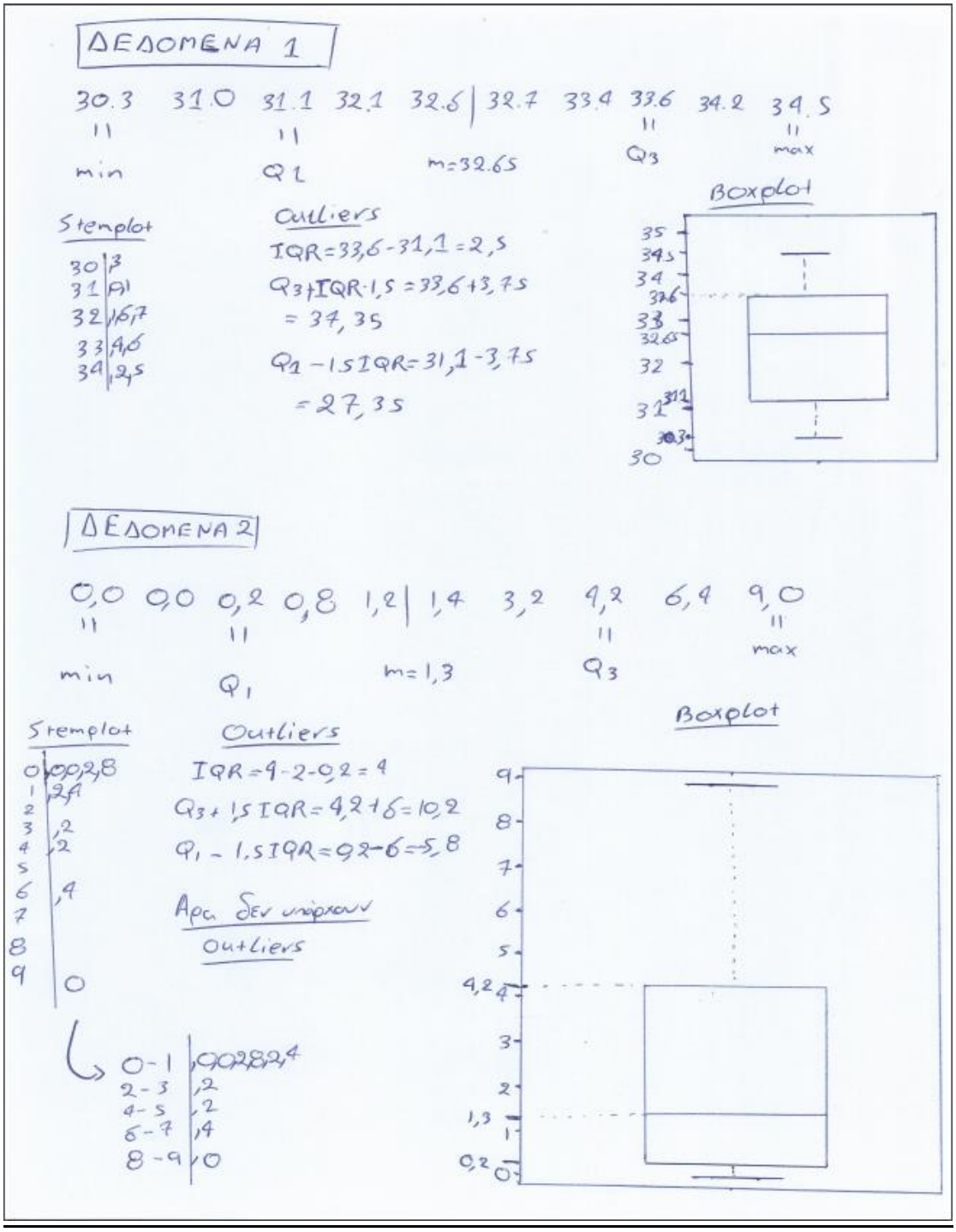
ΕΡΓΑΣΙΑ 5

ΜΕΛΗ ΟΜΑΔΑΣ:

ΑΘΑΝΑΣΙΟΣ ΚΛΕΤΤΑΣ: 3180079

ΙΑΣΩΝ ΖΙΩΓΑΣ: 3180057

ΑΣΚΗΣΗ 1:



ΔΕΔΟΜΕΝΑ 3

0 1 6 8 10 13 15 16 17 17 18 18 20 20
21 25 26 30 35 39 40 41 43 44 46 48 52 54 58
59 59 60 66 81 86 87 88 89 94 96

min = 0 $Q_1 = 17,5$ $m = 39,5$ $Q_3 = 59$ max = 96

Stemplot

0	0168
1	03567788
2	00156
3	059
4	013468
5	24899
6	06
7	
8	16789
9	46

Outliers

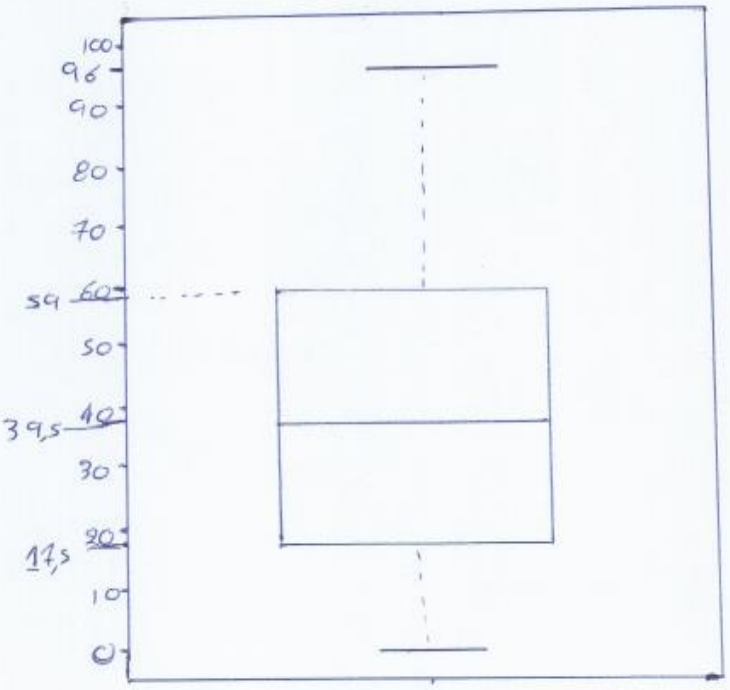
$IQR = Q_3 - Q_1 = 59 - 17,5 = 41,5$

$Q_3 + 1,5\,IQR = 59 + 62,25 = 121,25$

$Q_1 - 1,5\,IQR = 17,5 - 62,25 = -44,75$

Άρα δεν υπάρχουν outliers

BOXPLOT



b)

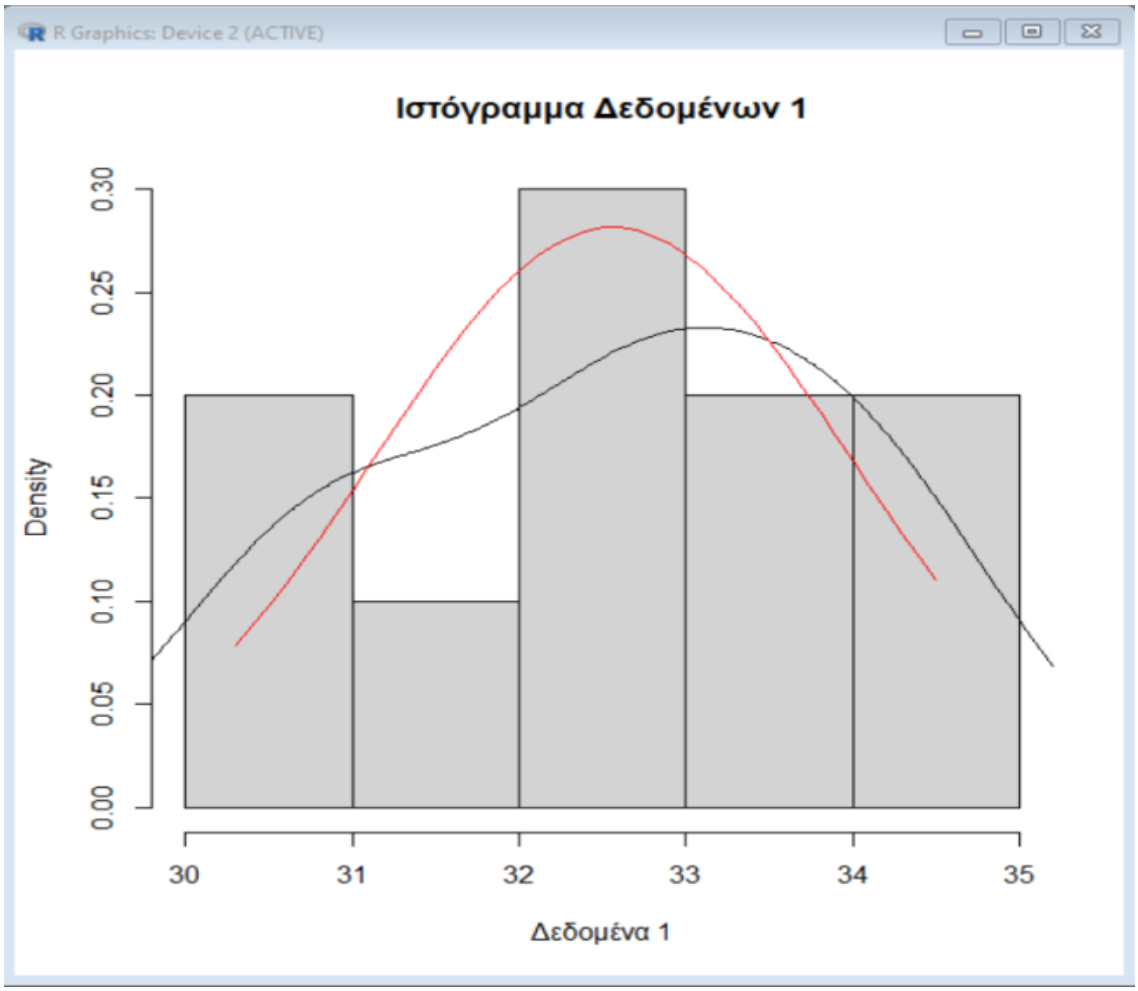
Δεδομένα 1: Παρατηρούμε στα δεδομένα ότι είναι συμμετρικά κατανεμημένα και δεν περιέχονται ισχυρά outliers οπότε συνοψίζει καλύτερα την κατανομή η μέση τιμή και τυπική απόκλιση

Δεδομένα 2: Παρατηρούμε ότι τα δεδομένα δεν είναι συμμετρικά κατανεμημένα πράγμα που σημαίνει ότι επηρεάζεται η μέση τιμή άρα η σύνοψη των 5 αριθμών συνοψίζει καλύτερα την κατανομή

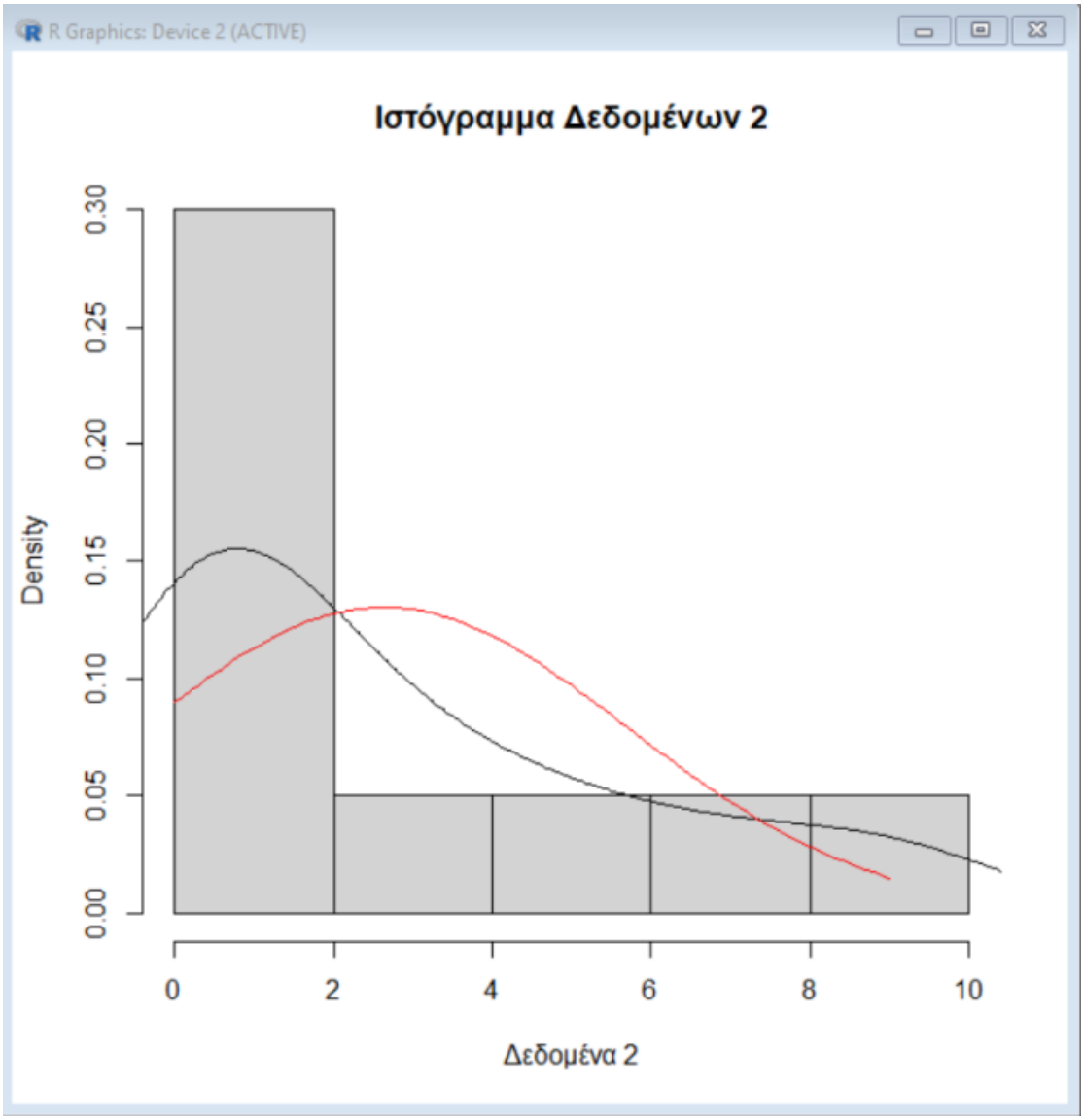
Δεδομένα 3: Παρατηρούμε στα δεδομένα ότι είναι συμμετρικά κατανεμημένα και δεν περιέχονται ισχυρά outliers οπότε συνοψίζει καλύτερα την κατανομή η μέση τιμή και τυπική απόκλιση

c)

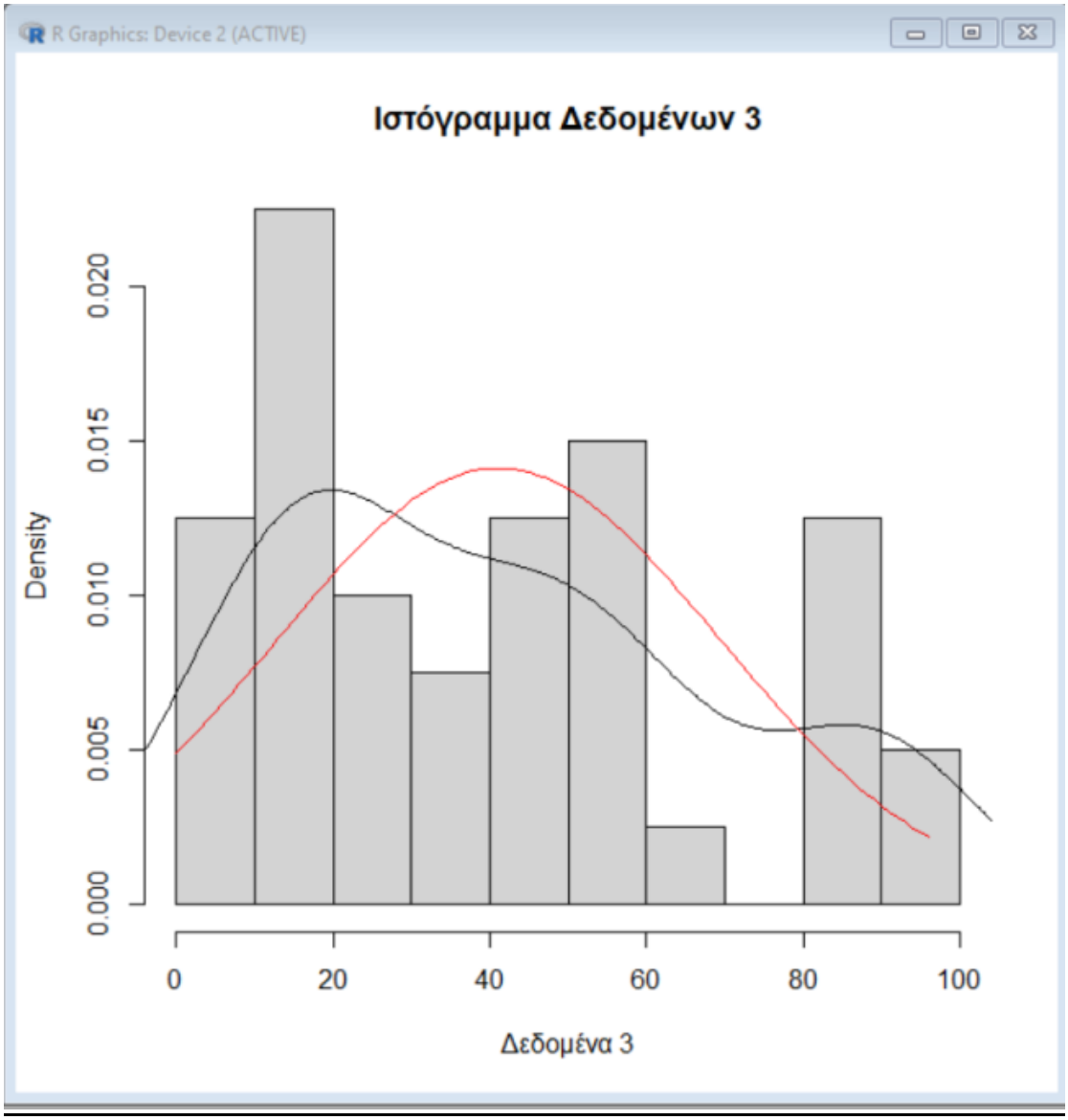
Καμπύλη πυκνότητας των δεδομένων (μαύρη) και καμπύλη πυκνότητας κανονικής κατανομής(κόκκινη) :



Υπολογίζοντας για τα δεδομένα 1 το ποσοστό των τιμών μέσα στο διάστημα (μ-σ,μ+σ) βρίσκουμε πως είναι το 50%, πράγμα που σημαίνει πως δεν προσεγγίζεται από μια καμπύλη πυκνότητας κανονικής κατανομής καθώς θα έπρεπε να είναι ‘κοντά’ στο 68%.



Υπολογίζοντας για τα δεδομένα 2 το ποσοστό των τιμών μέσα στο διάστημα (μ-σ,μ+σ) βρίσκουμε πως είναι το 80%, πράγμα που σημαίνει πως δεν προσεγγίζεται από μια καμπύλη πυκνότητας κανονικής κατανομής καθώς θα έπρεπε να είναι ‘κοντά’ στο 68%.



Υπολογίζοντας για τα δεδομένα 3 το ποσοστό των τιμών μέσα στο διάστημα (μ-σ,μ+σ) βρίσκουμε πως είναι το 70%, ποσοστό ‘κοντά’ στο 68%. Με περεταίρω υπολογισμούς βλέπουμε πως το ποσοστό των τιμών που βρίσκονται στο διάστημα (μ-2*σ,μ+2*σ) και στο (μ-3*σ,μ+3*σ) είναι το 100% το οποίο είναι κοντά στο 95% και 99.7% αντίστοιχα πράγμα που σημαίνει πως η καμπύλη πυκνότητας των δεδομένων προσεγγίζεται από μια καμπύλη πυκνότητας κανονικής κατανομής.

ΑΣΚΗΣΗ 2:

a)

Τα στοιχεία μας προέρχονται από τον Μετεωρολογικό Σταθμό Αχαρνών και από την Ελληνική Στατιστική Υπηρεσία. Στα δεδομένα μας περιέχονται 8 περιπτώσεις.

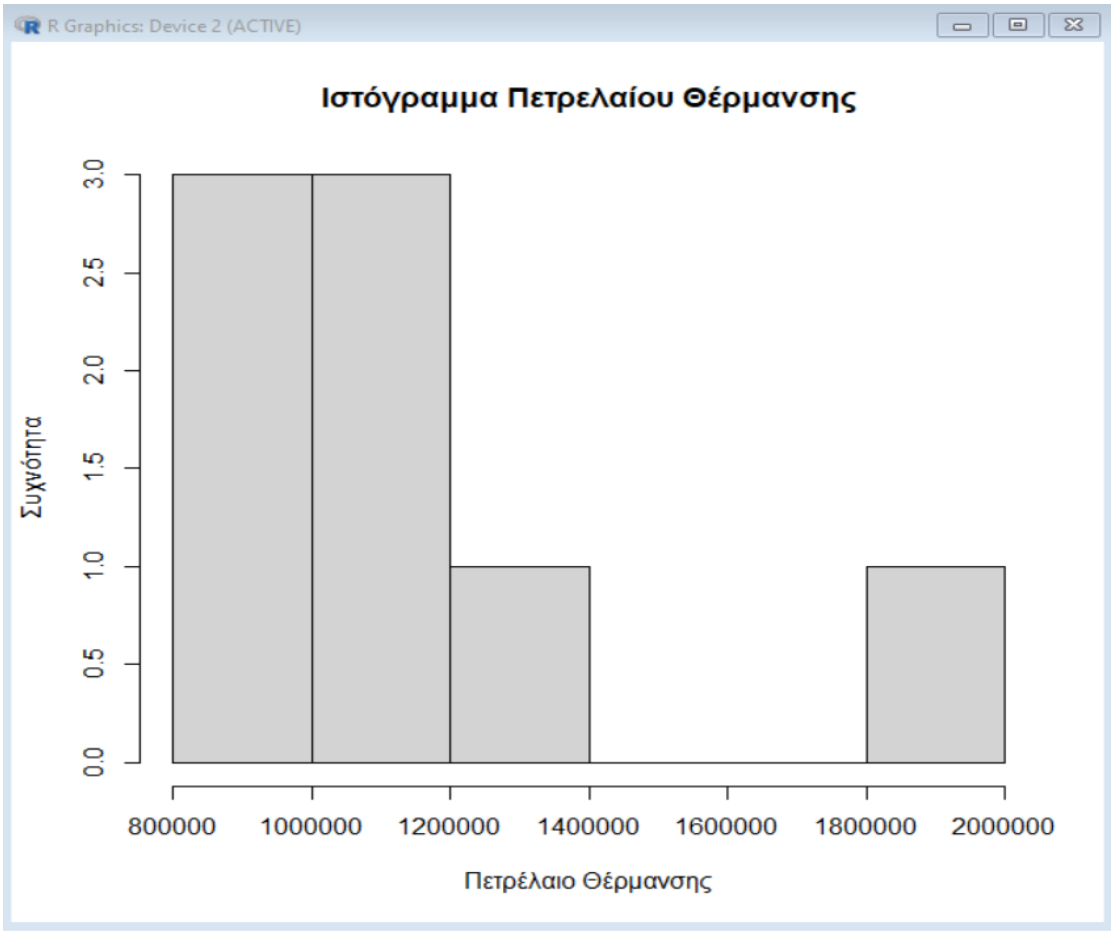
b)

Η κατηγορική μεταβλητή είναι τα χρόνια από 2012 εώς 2019. Οι ποσοτικές μεταβλητές είναι η συνολική κατανάλωση πετρελαίου θέρμανσης (τόνοι) στην Ελλάδα και η μέση τιμή θερμοκρασίας των τριών μηνών του Χειμώνα ανά έτος. Στην ουσία μια περίπτωση των δεδομένων μας περιλαμβάνει την χρονιά, την κατανάλωση πετρελαίου θέρμανσης και την θερμοκρασία του Χειμώνα η οποία προέκυψε όπως αναφέρθηκε παραπάνω.

c)

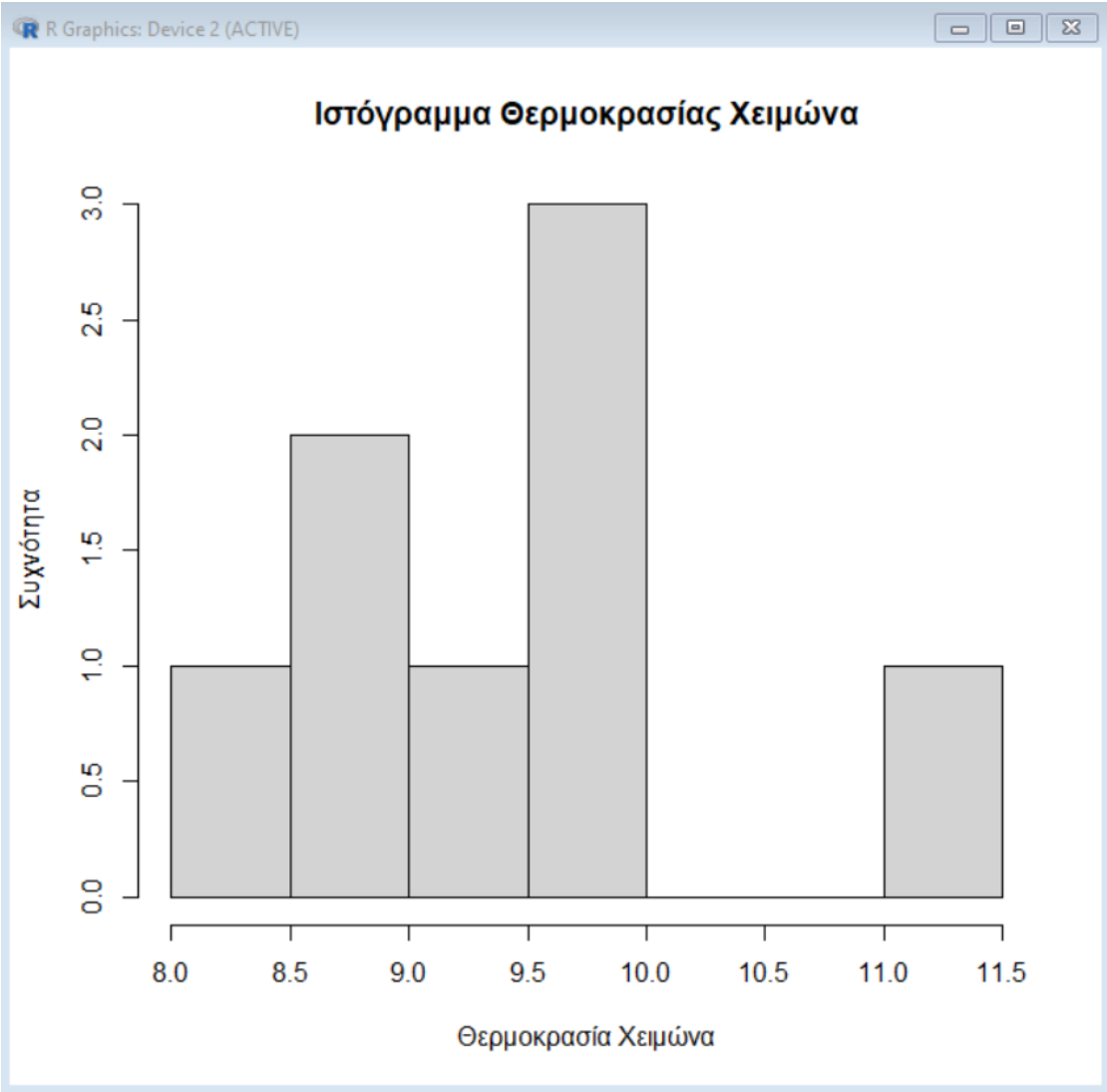
Κατανομές μεταβλητών σε γραφική μορφή :

- Κατανάλωση Πετρελαίου Θέρμανσης :



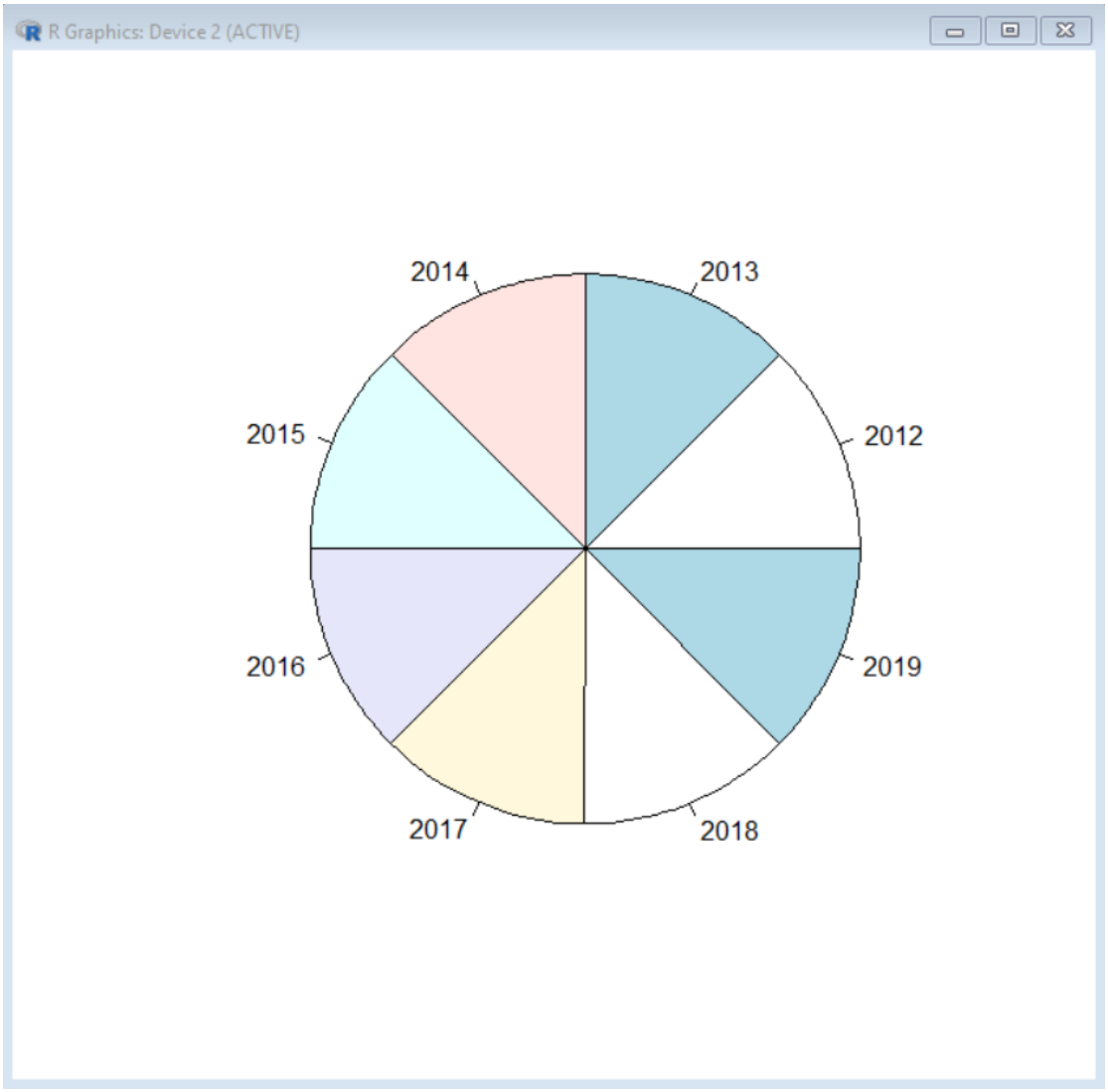
Παρατηρούμε ότι η κατανάλωση πετρελαίου θέρμανσης δεν είναι σταθερή. Επίσης μπορούμε να διακρίνουμε ένα ατυπικό σημείο. Αυτό μπορεί να οφείλεται στην τιμή του πετρελαίου η οποία αυξομειώνεται ανά τα χρόνια , στις καιρικές συνθήκες οι οποίες επικρατούν ή μπορεί και ακόμα στην θερμοκρασία του Χειμώνα (θα το μελετήσουμε παρακάτω). Οι περισσότερες τιμές είναι συγκεντρωμένες στο διάστημα [800000 , 1400000].

- Θερμοκρασία του Χειμώνα :



Παρατηρούμε ότι η Θερμοκρασία Χειμώνα βρίσκεται μεταξύ 8 και 11.5 βαθμούς Κελσίου παρουσιάζει μικρές διακυμάνσεις. Ανάμεσα στο (8.5 , 10] είναι συγκεντρωμένες οι περισσότερες τιμές. Η μορφή του οφείλεται στο ότι η θερμοκρασία δεν αλλάζει αισθητά ανά τα χρόνια είναι περίπου η ίδια. Ατυπικά σημεία δεν υπάρχουν.

Χρονιά :



Παρατηρούμε ότι τα χρόνια στο Τομεόγραμμα είναι ομοιόμορφα κατανεμημένα και αυτό είναι λογικό καθώς κάθε τιμή είναι μοναδική.

d)

α)

Μέση Τιμή Ποσοτικών Μεταβλητών

Κατανάλωση Πετρελαίου Θέρμανσης : m = 1212956

Θερμοκρασία Χειμώνα : m = 9.525

Τυπική Απόκλιση Ποσοτικών Μεταβλητών

Κατανάλωση Πετρελαίου Θέρμανσης : s = 318105.7

Θερμοκρασία Χειμώνα : s = 0.8892212

β)

Σύνοψη των Πέντε Αριθμών :

Κατανάλωση Πετρελαίου Θέρμανσης :

min : 959233

max : 1965436

Q1 : 968234.5

Q2 : 1127147

Q3 : 1294107.5

Θερμοκρασία Χειμώνα :

min : 8.4

max : 11.2

Q1 : 8.85

Q2 : 9.5

Q3 : 9.95

Σχολιασμός α),β)

Κατανάλωση Πετρελαίου Θέρμανσης :

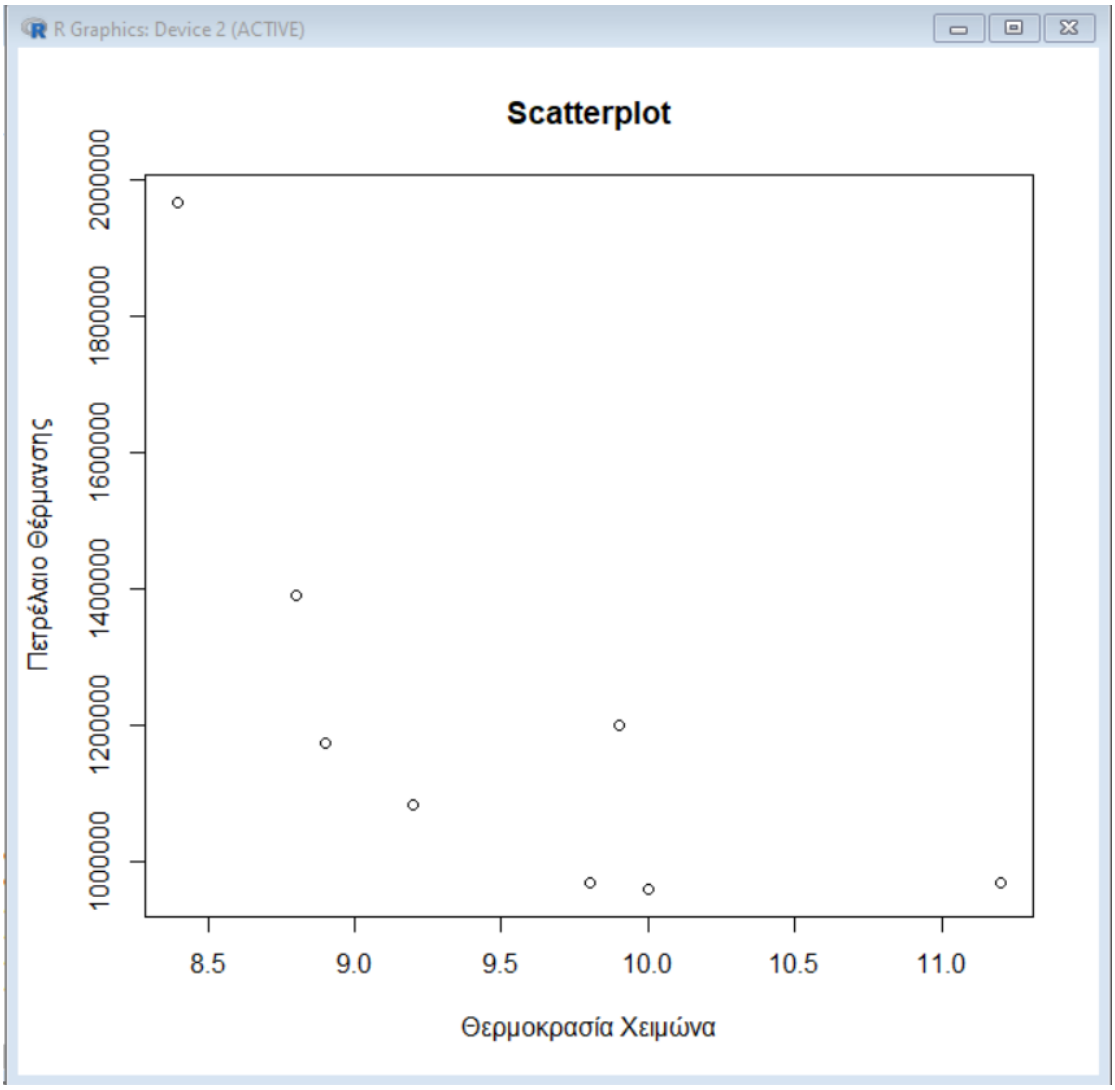
Η μέση τιμή και η τυπική απόκλιση δεν είναι κατάλληλες για τα δεδομένα της μεταβλητής κατανάλωση πετρελαίου θέρμανσης καθώς υπάρχει ένα outlier (1965436) το οποίο θα επηρεάσει την μέση τιμή και κατ'επέκταση την τυπική απόκλιση άρα δεδομένου του outlier η σύνοψη 5 αριθμών θα είναι πιο κατάλληλη.

Θερμοκρασία Χειμώνα :

Παρατηρώντας τα δεδομένα της Θερμοκρασίας Χειμώνα αντιλαμβανόμαστε ότι όχι μόνο δεν υπάρχουν ισχυρά outliers αλλά είναι και συμμετρικά κατανεμημένα. Αυτό σημαίνει ότι η μέση τιμή και η τυπική απόκλιση είναι κατάλληλα ενώ η σύνοψη των 5 αριθμών δεν είναι.

e)

Γραφική μορφή σχέσης μεταβλητών : Κατανάλωση Πετρελαίου Θέρμανσης, Θερμοκρασία Χειμώνα :

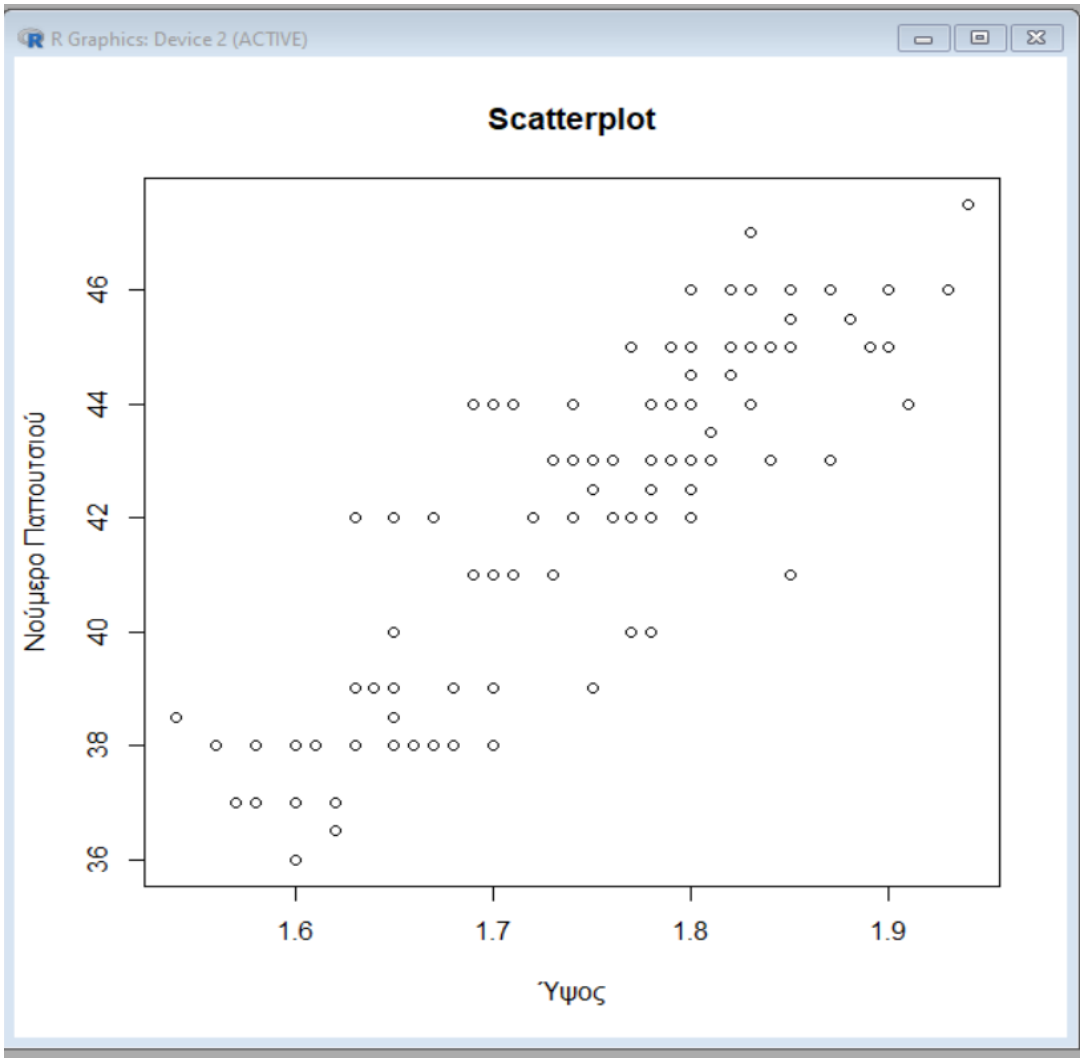


Παρατηρούμε ότι όσο οι τιμές της θερμοκρασίας του χειμώνα αυξάνονται οι τιμές κατανάλωσης πετρελαίου μειώνονται. Αυτό σημαίνει ότι η σχέση μεταξύ των 2 μεταβλητών είναι αιτιατή διότι η τιμή της μεταβλητής θερμοκρασία του χειμώνα επηρεάζει την τιμή της μεταβλητής κατανάλωση πετρελαίου

ΑΣΚΗΣΗ 3:

a)

Οι δύο ποσοτικές μεταβλητές που επιλέξαμε να διερευνήσουμε την σχέση τους από το ερωτηματολόγιο του 2020 είναι το ύψος και το νούμερο παπουτσιού (height,shoe). Το Scatterplot που προκύπτει είναι το εξής :



Παρατηρούμε ότι η μορφή του είναι γραμμική, δεν υπάρχουν ατυπικά σημεία. Η κατεύθυνση της σχέσης είναι αύξουσα και η δύναμη της ισχυρή.

b)

Ο συντελεστή συσχέτισης r είναι 0.8665411 (κοντά στο 1) και αυτό παρατηρούμε ότι είναι λογικό καθώς η δύναμη της σχέσης είναι ισχυρή.

Με την εκτέλεση της γραμμικής παλινδρόμησης ελαχίστων τετραγώνων προκύπτει η εξίσωση: $y = (0.0273)x + 0.6018$.