Ζέρβα Αθηνά-Μαρία : 3170207
Παπαγεωργίου Νικολέτα : 3170234

1η Άσκηση

1a)

Stemplot:

Για Δεδομένα Ι:

30 |3

31 |0 1

32 |1 6 7

33 | 4 6

34 |2 5

Για Δεδομένα Ι Ι:

0 |0 0 1 4 6 7

1 |6

2 |1

3 |2

4 | 5

Για Δεδομένα I I:

0 |

1 |0 3 5 6 7 8

2 |0 1 5 6

3 |0 5 9

4 |0 1 3 4 6 8

5 |2 4 8 9 9

6 |0 6

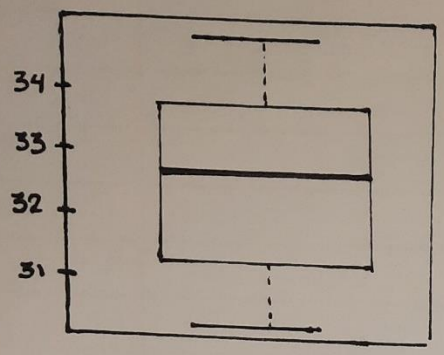
7 |

8 |1 6 7 8 9

9 |4 6

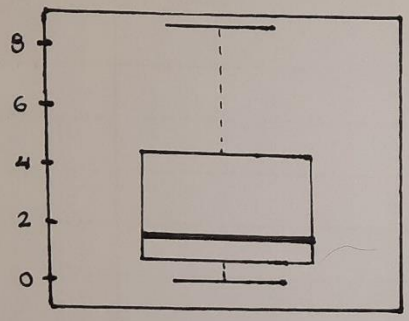
Boxplot

(I):



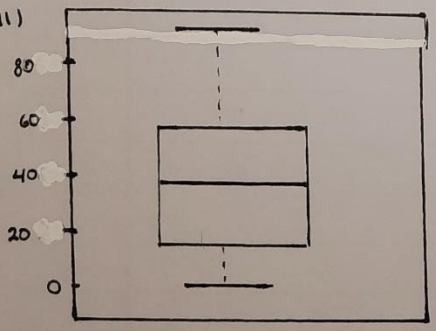
max = 34.5
Q_2 = 33.6
m = 32.65
Q_1 = 31.1
min = 30.3

(II)



max = 9.0
Q_2 = 4.2
m = 1.3
Q_1 = 0.2
min = 0.0

(III)



max = 96.0
Q_2 = 59.0
m = 39.5
Q_1 = 17.5
min = 0.0

1b)

Η ομάδα τιμών που περιγράφει καλύτερα τα δεδομένα είναι η σύνοψη των 5 αριθμών. Αυτό συμβαίνει επειδή λαμβάνοντας υπόψη αποκλειστικά το ζεύγος μέση τιμή - τυπική απόκλιση δε μπορούμε να αποφανθούμε με σιγουριά για το εύρος των τιμών του πειράματός μας και αυτό επειδή η κατανομή τους μπορεί να είναι εντελώς ανομοιόμορφη και διασκορπισμένη "άσχημα" μακριά από τις δύο τιμές που αναφέραμε παραπάνω. Αντίθετα, χρησιμοποιώντας την σύνοψη των 5 αριθμών, έχουμε πολύ περισσότερες πληροφορίες (min, Q1, median, Q3, max) και μπορούμε να έχουμε μια γενικότερη εικόνα της κατανομής που μας απασχολεί. Ξέρουμε την μεγαλύτερη και μικρότερη τιμή, καθώς και τη διάμεσο, και έχουμε μια επιπλέον πληροφορία για τις διαμέσους των επιμέρους τμημάτων (min,median), (median,max), κοινώς τα Q1, Q2. Επομένως έτσι έχουμε σφαιρικότερη άποψη για το εύρος των δυνατών τιμών μας καθώς επίσης και για τη διακύμανση αυτών.

1c)

Από τη θεωρία μας, και πιο συγκεκριμένα τον κανόνα 68-95-99.7, γνωρίζουμε ότι: στην κανονική κατανομή ισχύουν τα εξής:

- το 68% των παρατηρήσεων βρίσκεται στο διάστημα (μ – σ, μ + σ)
- το 95% των παρατηρήσεων βρίσκεται στο διάστημα (μ – 2σ, μ + 2σ)
- το 99,5% των παρατηρήσεων βρίσκεται στο διάστημα (μ – 3σ, μ + 3σ)

,όπου μ = μέση τιμή, σ = τυπική απόκλιση.

Για το i:

Επομένως θα εργαστούμε κι εμείς ανάλογα στο σημείο αυτό υπολογίζοντας την μέση τιμή και την τυπική απόκλιση των παρατηρήσεών μας και στη συνέχεια κάνοντας τις κατάλληλες συγκρίσεις. Εύκολα βρίσκουμε ότι μ = 32.55, σ = 1.419898 (πληκτρολογήσαμε στην R τις εντολές mean(x) και sd(x) αντίστοιχα).

Παρατηρούμε ότι στο διάστημα (31.1301, 33.9699), δηλαδή το αντίστοιχο (μ – σ, μ + σ) για τα δικά μας δεδομένα, βρίσκονται οι 6 από τις 10 παρατηρήσεις, ποσοστό της τάξης του 60% και όχι του αναμενόμενου 68% ώστε να έχουμε κανονική κατανομή ή κάποιας κοντινής σε αυτή.Έτσι, μπορούμε να αποφανθούμε ότι η προσέγγιση της κατανομής των δεδομένων μας από την καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές αποκλίσεις. Κάνοντας δοκιμές για τα αντίστοιχα διαστήματα (μ – 2σ, μ + 2σ) αλλά και (μ – 3σ, μ +3σ) είδαμε ότι στο πρώτο διάστημα (29.7102, 35.3898) βρίσκεται το 100% των παρατηρήσεών μας έναντι του 95% της Κανονικής Κατανομής, και στο (28.29031, 36.80969) το 100% αυτών έναντι του 99,5%.

Για το ii:

Επομένως θα εργαστούμε κι εμείς ανάλογα στο σημείο αυτό υπολογίζοντας την μέση τιμή και την τυπική απόκλιση των παρατηρήσεών μας και στη συνέχεια κάνοντας τις κατάλληλες συγκρίσεις. Εύκολα βρίσκουμε ότι μ = 2.64, σ = 3.059121 (πληκτρολογήσαμε στην R τις εντολές mean(x) και sd(x) αντίστοιχα).

Παρατηρούμε ότι στο διάστημα (-0.4191212, 5.699121), δηλαδή το αντίστοιχο (μ – σ, μ + σ) για τα δικά μας δεδομένα, βρίσκονται οι 8 απο τις 10 παρατηρήσεις, ποσοστό της τάξης του 80% και όχι του αναμενόμενου 68% ώστε να έχουμε κανονική κατανομή ή κάποιας κοντινής σε αυτή. Έτσι, μπορούμε να αποφανθούμε ότι η προσέγγιση της κατανομής των δεδομένων μας από την καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές αποκλίσεις. Κάνοντας δοκιμές για τα αντίστοιχα διαστήματα (μ – 2σ, μ + 2σ) αλλά και (μ – 3σ, μ +3σ) είδαμε ότι στο πρώτο διάστημα (-3.478242, 8.758242) βρίσκεται το 100% των παρατηρήσεών μας έναντι του 100% της Κανονικής Κατανομής, και στο (-6.537363, 11.81736) το 100% αυτών έναντι του 99,5%.

Για το iii:

Επομένως θα εργαστούμε κι εμείς ανάλογα στο σημείο αυτό υπολογίζοντας την μέση τιμή και την τυπική απόκλιση των παρατηρήσεών μας και στη συνέχεια κάνοντας τις κατάλληλες συγκρίσεις. Εύκολα βρίσκουμε ότι μ = 41.15, σ = 28.26754 (πληκτρολογήσαμε στην R τις εντολές mean(x) και sd(x) αντίστοιχα).

Παρατηρούμε ότι στο διάστημα (12.88246, 69.41754), δηλαδή το αντίστοιχο (μ – σ, μ + σ) για τα δικά μας δεδομένα, βρίσκονται οι 28 από τις 40 παρατηρήσεις, ποσοστό της τάξης του 70% και όχι του αναμενόμενου 68% ώστε να έχουμε κανονική κατανομή ,αλλά έχουμε κάποιας κοντινής σε αυτή. Έτσι, μπορούμε να αποφανθούμε ότι η προσέγγιση της κατανομής των δεδομένων μας από την καμπύλη πυκνότητας της Κανονικής Κατανομής δεν παρουσιάζει σημαντικές αποκλίσεις. Κάνοντας δοκιμές για τα αντίστοιχα διαστήματα (μ – 2σ, μ + 2σ) αλλά και (μ – 3σ, μ +3σ) είδαμε ότι στο πρώτο διάστημα (-15.38508, 97.68508) βρίσκεται το 100% των παρατηρήσεών μας έναντι του 95% της Κανονικής Κατανομής, και στο (-43.65262, 125.9526) το 100% αυτών έναντι του 99,5%.

2a)

Δώστε μια σύντομη περιγραφή από που προέρχονται τα δεδομένα και πόσες περιπτώσεις περιέχονται.

Σε ένα μάθημα προηγούμενο εξάμηνου μας είχε ζητηθεί να μελετήσουμε μια καινοτόμα εφαρμογή και στα πλαίσια της ερευνάς είχαμε φτιάξει ένα ερωτηματολόγιο με θέμα «Συμπληρώνοντας αυτό το ερωτηματολόγιο μας βοηθάτε στην αξιολόγηση της έξυπνης κλειδαριάς,AppLocker.

Πρόκειται για μια εφαρμογή για που σας παρέχει την ασφάλιση της μοτοσυκλέτας σας μέσω της smartphone συσκευής σας.»

Οι ερωτήσεις οι οποίες είχαν απαντηθεί από συμφοιτητές μας ήταν οι εξής:

1. Είστε κάτοχος μοτοσυκλέτας; (motorcycle)
2. Τι φύλο είστε; (sex)
3. Τι ηλικία έχετε;(age)
4. Χρησιμοποιείτε κλειδαριά;(locker)
5. Αν ναι,πόσο ικανοποιημένος είστε με την κλειδαριά που χρησιμοποιείτε ήδη; (satisfaction)
6. Η κλειδαριά σας είναι smart;(smart)
7. Αν όχι θα σας ενδιέφερε να αποκτήσετε;(interesting)
8. Τι θα θέλατε να σας παρέχει μια τέτοια εφαρμογή;(benefits)
9. Τι χρηματικό ποσό θα ήσασταν πρόθυμοι να διαθέσετε για μια τέτοια κλειδαριά;(amount of money)
10. Πόσο χρήσιμη θα σας φαινόταν μια τέτοια εφαρμογή;(useful)

2b)

Ποιες είναι κατηγορικές και ποιες ποσοτικές μεταβλητές; Δώστε μια σύντομη περιγραφή κάθε μιας από αυτές (ή ορισμένων εάν είναι πάρα πολλές).

Οι κατηγορικές μεταβλητές είναι τα motorcycle, sex, locker, smart ,interesting, benefits, amount of money ενώ ποσοτικές τα age, satisfaction, useful.

Motorcycle: απαντάει στο αν έχουν μηχανάκι η όχι.

Locker: απαντάει στο αν έχουν κλειδαριά

Smart: απαντάει στο αν είναι smart ή όχι

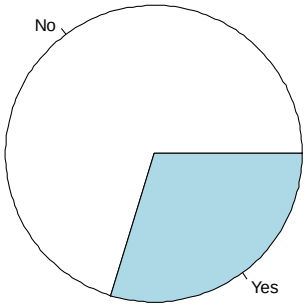
Interesting : αν τους ενδιαφέρει να αποκτήσουν

Benefits : Τι θα ήθελαν να παρέχει μια τέτοια εφαρμογή

2c)

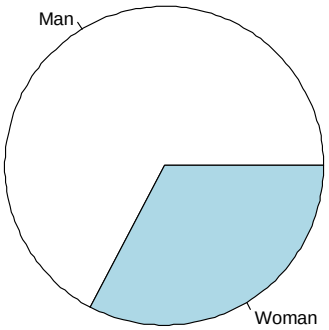
Οι κατανομές των κατηγορικών μεταβλητών :motorcycle,sex,locker,smart,interesting, φαίνονται από στα εξής τομεόγραμματα:

motorcycle

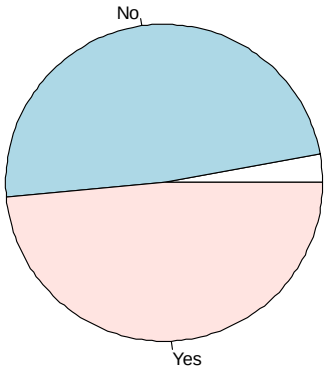


Sex

Φαίνεται πως περίπου το 60% των ατόμων που απάντησαν το ερωτηματολόγιο μας κι κατέχουν μηχανές είναι άντρες.

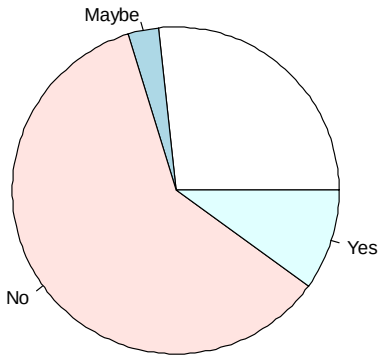


Locker

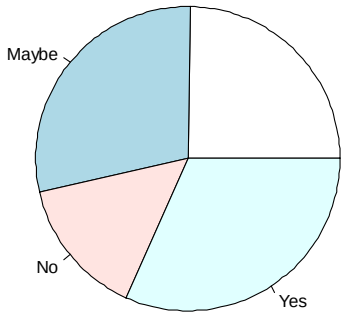


Smart

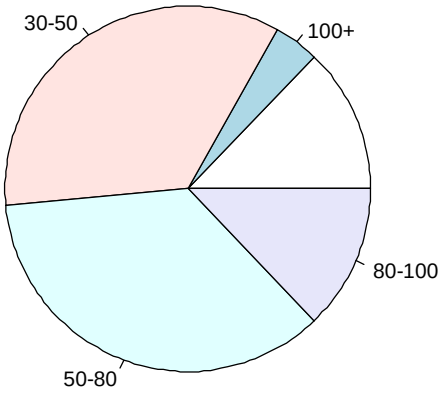
Περίπου το 60% δεν χρησιμοποιούν smart κλειδαριά ,κάτι που συμφωνεί με τη λόγο ορισμού της applocker ως καινοτομίας.



Interesting

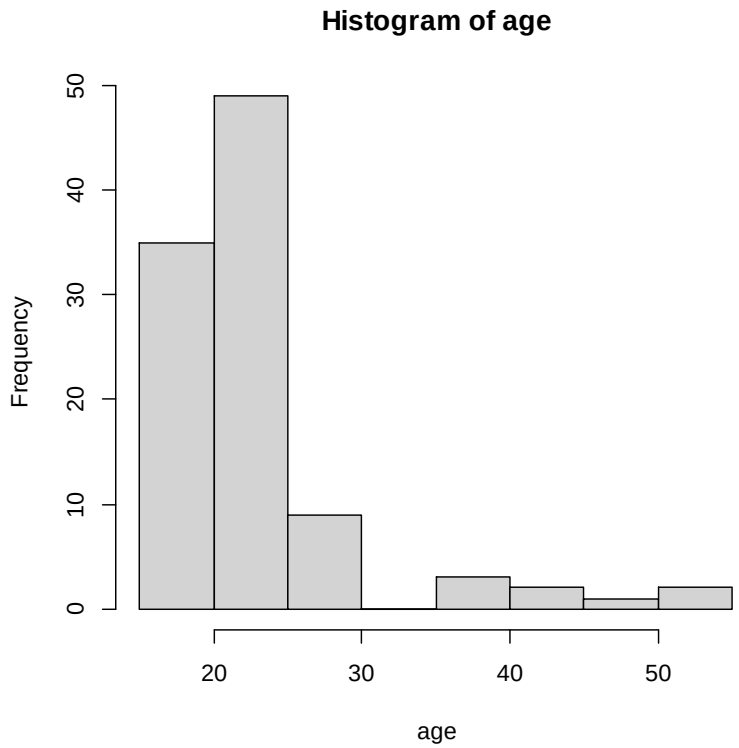


Amount.of.money



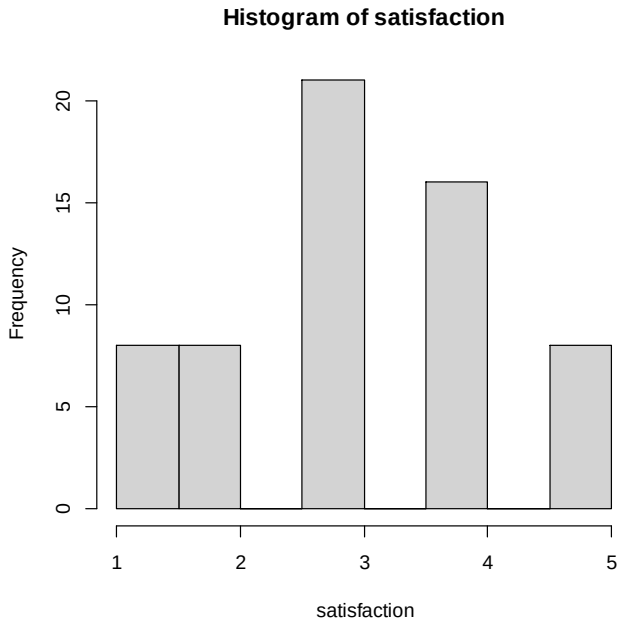
Ενώ των ποσοτικών μεταβλητών age,satisfaction,amount of money,useful από τα ιστογράμματα:

Age

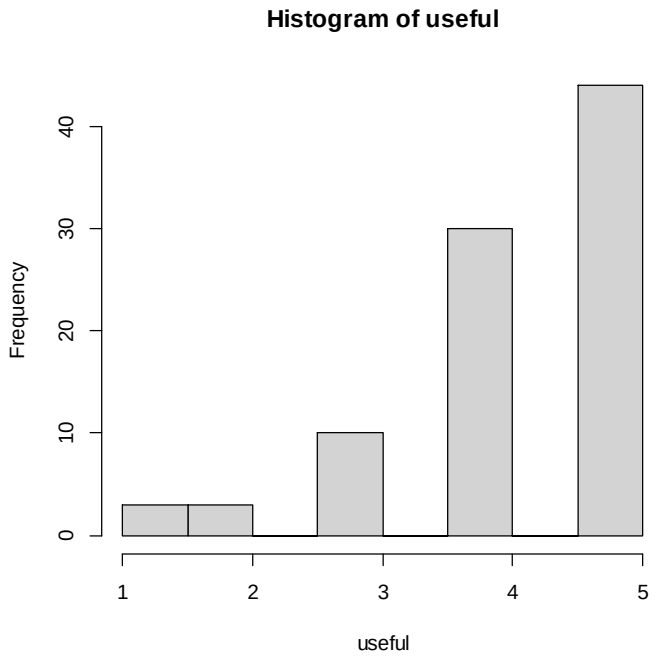


Οι ηλικίες κυμαίνονται κοντά στα 20 όπου έχουμε κάποιες αποκλίσεις από το 30 και έπειτα.

Satisfaction



Useful



2d)

Για κάθε ποσοτική μεταβλητή, υπολογίστε

α) τη μέση τιμή και τυπική απόκλιση, και

- **Age:** Η μέση τιμή είναι 23.0198 και η τυπική απόκλιση : 7.263581
- **Satisfaction:** Η μέση τιμή ισούται με 3.131148 κι η τυπική απόκλιση : 1.203819
- **Useful:** Η μέση τιμή είναι 4.211111κι η τυπική απόκλιση : 0.9999376

β) τη σύνοψη των πέντε αριθμών.

- fivenum(satisfaction)

[1] 1 2 3 4 5

- fivenum(useful)

[1] 1 4 4 5 5

- fivenum(age)

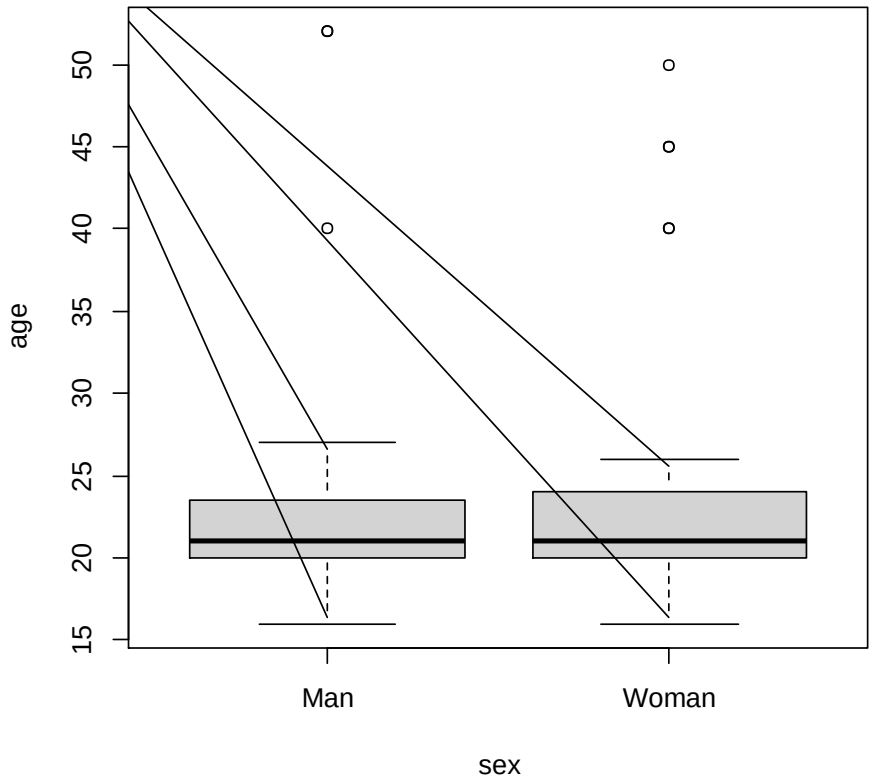
[1] 16 20 21 24 52

Σχολιάστε την καταλληλότητα των α), β) για κάθε μεταβλητή.

Κι οι τρεις μεταβλητές είναι κατάλληλες αφού μπορούν να υπολογιστούν ποσοστημόρια.

2e)

Θα διερευνήσουμε τη σχέση μεταξύ της κατηγορηματικής sex και ποσοτικής age.Ως επεξηγηματική μεταβλητή θα θεωρήσουμε το sex.Γραφικά αναπαριστούμε της σχέσης τους με πλάι πλάι boxplots :



Επόμενο είναι να παρουσιάζεται ίδια min τιμή για τη ηλικία και στα δυο φύλλα

Όμως φαίνεται πως οι άντρες του ερωτηματολογίου που κατέχουν μοτοσυκλέτα max age μεγαλύτερη των γυναικών.

Η σχέση φύλου και ηλικίας δεν είναι αιτιατή, καθώς δεν επηρεάζει το ένα το άλλο.

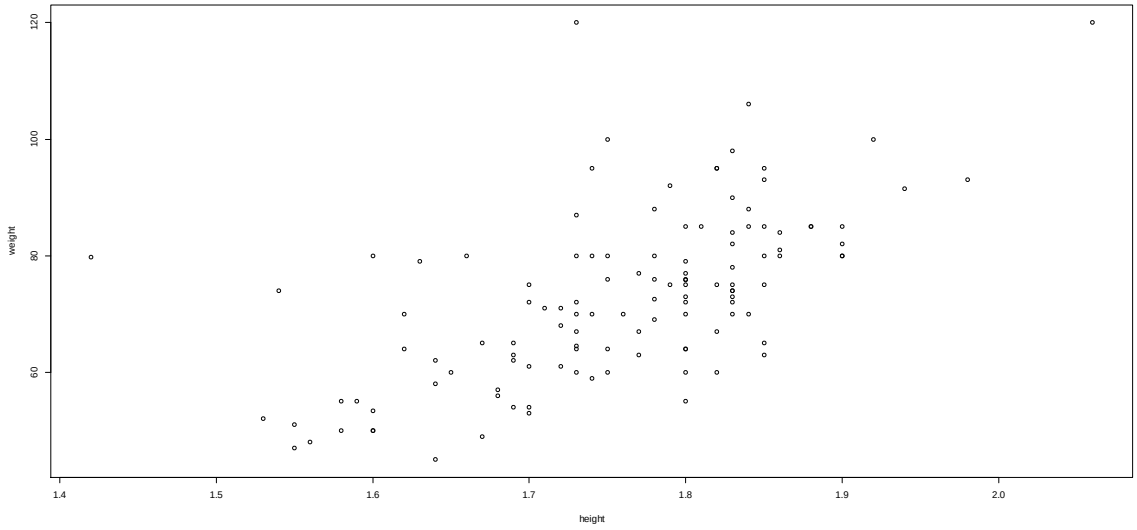
3)

Εδώ θα διερευνήσετε τη σχέση μεταξύ οποιονδήποτε δύο ποσοτικών μεταβλητών στα δεδομένα των απαντήσεων ερωτηματολογίου 2021 που βρίσκονται στο eclass.

a)

Δώστε το scatterplot και σχολιάστε τη μορφή, κατεύθυνση και δύναμη της σχέσης των δύο μεταβλητών.

plot(height, weight)



1. Μορφή σχέσης αύξουσα γραμμική, και **αρκετά ισχυρή**. Παρατηρούμε ότι ίσως να υπάρχουν 2 ομάδες, δεξιά και αριστερά. Αυτό συμβαίνει γιατί έχουμε πάρει ως ποσοτικές μεταβλητές τα βάρος και ύψος. Άρα είναι πιθανόν να χωρίζονται σε δυο κατηγορικές ομάδες αγόρια – κορίτσια. Πιο ψηλά και πιο βαριά τα αγόρια. Και πιο κοντά και πιο λεπτά τα κορίτσια.
   Έχουμε και μερικά ατυπικά σημεία, δηλαδή σημεία που απέχουν απτό σύνολο. Κάποια μεταβλητή έχει μεγάλο ύψος και κάποια άλλη μεγάλο βάρος , σε σχέση με τους υπόλοιπους.

b)

Υπολογίστε τον συντελεστή συσχέτισης και εκτελέστε γραμμική παλινδρόμηση ελαχίστων τετραγώνων.

plot(weight~height, xlim=c(1.5,2), ylim=c(40,110))

> cor(height, weight,use="complete.obs")

0.6090113

> model <- lm(weight~height)

> model

Call:

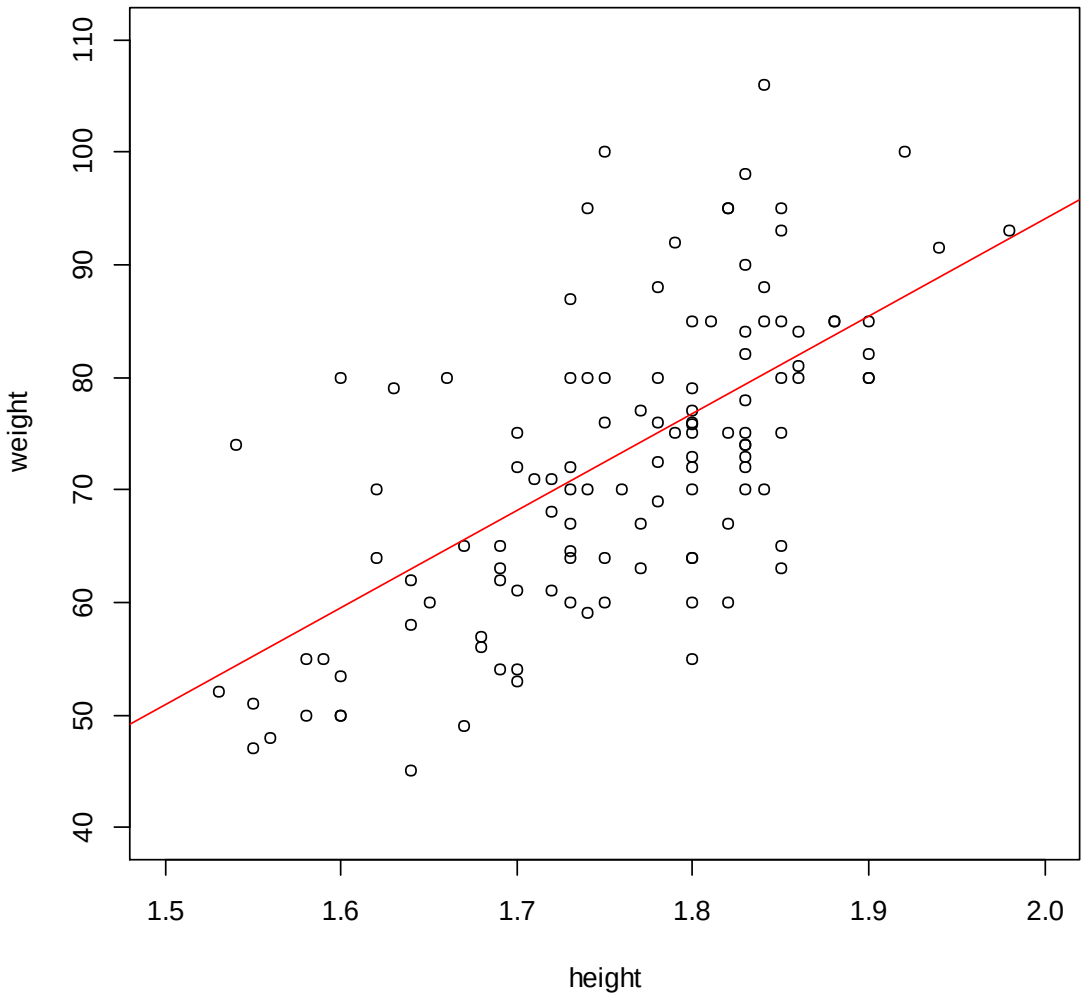
lm(formula = weight ~ height)

Coefficients:

(Intercept) height

-78.67 86.35

abline(model,col="red")


weight
height
40
50
60
70
80
90
100
110
1.5
1.6
1.7
1.8
1.9
2.0