1. Δημιουργήστε το διάνυσμα (ακεραίων) και υπολογίστε τον μέσο όρο και την τυπική απόκλιση

my.vector1<-c(4,6,8,3,5,1,6)

mean(my.vector1)

sd(my.vector1)

2. Δημιουργήστε το διάνυσμα (δεκαδικών) και υπολογίστε τον μέσο όρο και την τυπική απόκλιση ()

my.vector2<-c(3.1,1,ΝΑ,1.3,ΝΑ,2.1,4.5)

mean(my.vector2) #δεν θα δουλέψει,

sd(my.vector2) #δείτε την βοήθεια των συναρτήσεων mean και sd, tip: na.rm=T

3. Δημιουργήστε ένα διάνυσμα χαρακτήρων και υπολογίστε την συχνότητα των λέξεων

my.vector3<-c("a","b","c","c","c","b","a")

table(my.vector2)

4. Δημιουργήστε ένα data.frame από τα τρία προηγούμενα διανύσματα

my.df = data.frame(name=my.vector3, vec1=my.vector1, vec2=my.vector2)

my.df

5. Δείτε στα γρήγορα τα χαρακτηριστικά του data.frame που δημιουργήσατε

summary(my.df)

str(my.df)

6. Ανακτήστε το διάνυσμα της πρώτης στήλη του data.frame

my.df$name

my.df[[1]]

7. Ανακτήστε τις τρείς πρώτες γραμμές του data.frame

my.df[1:3,]

8. Ανακτήστε τις γραμμές του data.frame που η δεύτερη στήλη είναι μεγαλύτερη από 4

my.df[my.df$vec1>4,]

9. Ανακτήστε τις δύο πρώτες στήλες του data.frame σε μία καινούργια μεταβλητή. Δείτε τί είδους τύπος είναι και τί είδους αντικείμενο

my.df2=my.df[,1:2]

class(my.df2)

typeof(my.df2)

10. Δείτε ποια στοιχεία της τρίτης στήλης του data.frame είναι NA. Υπολογίστε τον μέσο όρο των υπολοίπων

vec2.na=is.na(my.df$vec2)

mean(my.df[!vec2,3])

mean(my.df[!is.na(my.df$vec2),3])

11. Υπολογίστε την συσχέτιση των στηλών vec1 και vec2, μόνο για τις παρατηρήσεις που δεν έχουν στοιχεία NA

full.rows=(rowSums(is.na(my.df))==0)

full.rows

cor(my.df[full.rows,"vec1"],my.df[full.rows,"vec2"])

cor(my.df$vec1,my.dfs$vec2,use="")

12. Βρείτε τον μέσο όρο των δύο στηλών για κάθε level της 1ης στήλης

aggregate(my.df[,2:3],by=list(my.df$name),mean,na.rm=T)

13. Εισάγεται τα δεδομένα από το αρχείο «2004\_aua\_meteo.csv». Δείτε τι στήλες έχουν και τι είδους δεδομένα έχουν

real.df=read.table("data/2004\_aua\_meteo.csv",sep=",",header=T, as.is=T)

str(real.df)

summary(real.df)

14. Δείτε τον μέσο όρο της θερμοκρασίας ανά μήνα στις 13.00 το μεσημέρι και φτιάξτε ένα γράφημα. Δείτε τα τεταρτημόρια με βάση τον μήνα και φτιάξτε ένα boxplot

temp.agr.mean=aggregate(real.df[real.df$Hour==13,"AirT.1.5"],by=list(real.df[real.df$Hour==13,"Month"]),mean,na.rm=T)

plot(temp.agr.mean)

temp.agr.quant=aggregate(real.df[real.df$Hour==13,"AirT.1.5"],by=list(month=real.df[real.df$Hour==13,"Month"]),quantile,na.rm=T)

boxplot(AirT.1.5~Month,real.df[real.df$Hour==13,])

15. Βρείτε την αθροιστική βροχόπτωση ανά μήνα

rain.aggr=aggregate(list(rain=real.df$Rain),by=list(month=real.df$Month),sum,na.rm=T)

rain.aggr

plot(rain.aggr)

16. Συγκρίνετε τις θερμοκρασίες στις 13.00 και στις 20.00 για κάθε μήνα. Φτιάξτε ένα boxplot,

boxplot(AirT.1.5~Month\*Hour,real.df[real.df$Hour%in%c(13,20),],col=c("green","red"),main="Temp 13.00 and 20.00", xlab="Month")

17. Εγκαταστήστε και φορτώστε τις βιβλιοθήκες reshape2 και ggplot2

install.packages("reshape2"); install.packages("ggplot2");

library(reshape2); library(ggplot2)

18. Εγκαταστήστε και φορτώστε το πακέτο car. Δείτε αν έχει datasets διαθέσιμα. Θα δουλέψουμε με το dataset Prestige. Δείτε την περιγραφή του πακέτου στην βοήθεια.

install.packages("car"); library(car)

data(package="car")

?Prestige

19. Δείτε γραφικά την συσχέτιση όλων των μεταβλητών μεταξύ τους.

plot(Prestige)

20. Δείτε αν το prestige μίας δουλειάς εξαρτάται από την εκπαίδευση, το εισόδημα και το φύλο, τρέχοντας μία απλή γραμμική παλινδρόμηση

res.lm=lm(prestige~education+income+women,data=Prestige)

summary(res.lm)

plot(res.lm)

21. Εισάγεται τα δεδομένα από το αρχείο «cluster\_data\_lessons\_owned.csv». Δείτε τι στήλες έχουν και τι είδους δεδομένα έχουν

lessons.df=read.table("data/cluster\_data\_lessons\_owned.csv",sep="\t",header=T,as.is=T)

str(lessons.df)

22. Υπολογίστε το ποσοστό των μαθημάτων που έχουν περαστεί (σε σχέση με αυτά που θα έπρεπε να έχουν περαστεί), δηλαδή το ποσοστό ολοκλήρωσης των σπουδών

lessons.df$perc1= lessons.df$passed1/ lessons.df$ideal1

lessons.df$perc2= lessons.df$passed2/ lessons.df$ideal2

lessons.df$perc3= lessons.df$passed3/ lessons.df$ideal3

lessons.df$perc4= lessons.df$passed4/ lessons.df$ideal4

lessons.df$perc5= lessons.df$passed5/ lessons.df$ideal5

23. Επιλέξτε μόνο τις στήλες με τα ποσοστά σε ένα νέο data.frame

lessons.df.perc= lessons.df[,c("id",perc1","perc2","perc3","perc4","perc5")]

24. Υπολογίστε τις ομάδες με βάση την εξέλιξη του ποσοστού ολοκλήρωσης μαθημάτων. Επιλέξτε 4 ομάδες. Βρείτε τα χαρακτηριστικά των ομάδων σε σχέση με το ποσοστό ολοκλήρωσης (see https://www.statmethods.net/advstats/cluster.html)

fit <- kmeans(lessons.df.perc[,2:6],4)

fit$centers

25. Σε ένα boxplot διάγραμμα, δείξτε την εξέλιξη ολοκλήρωσης σπουδών για τις διαφορετικές ομάδες. Τρέξτε το προηγούμενο και αυτό το ερώτημα για 5 ομάδες.

lessons.df.perc$cluster=fit$cluster

lessons.df.perc.melted=melt(lessons.df.perc,id.vars = c("id","cluster"))

ggplot(lessons.df.perc.melted,aes(x=variable,y=value,fill=factor(cluster)))+geom\_boxplot()