1. The dataframe `VIT2005` in the `PASWR2` package contains descriptive information and the appraised `totalprice` (in euros) for apartments in Vitoria, Spain.

   a) Characterize variable `totalprice` (find mean, standard deviation, verify symmetry), identify outliers, if any. Report the histogram and boxplot in a single display.

   b) Find the average age of apartments with no garage.

   c) Find the number apartments with no garage and a single storage unit.

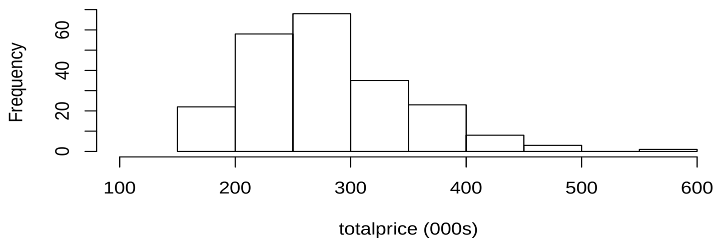
   d) Make a scatterplot of `totalprice` and `area`. Report the row number of outliers.

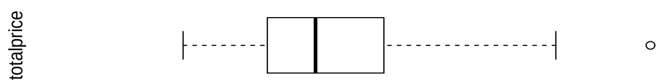
   e) Use variable `area` to predict the price of an apartment with 100 square meters.

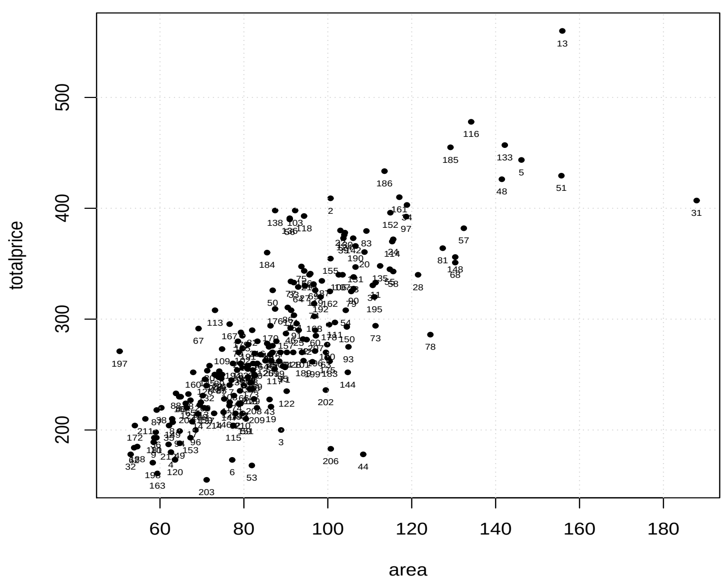
   f) Which is a better predictor of `totalprice`, area or age? why?

```
library(PASWR2)  # VIT2005
library(car)     # Boxplot()
d0=VIT2005
setwd("C:/Users/Cesar/Favorites/Downloads") # folder for saving

# a)
totalprice = d0$totalprice/1000
sd(totalprice)   # [1] 69.29846
summary(totalprice)
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  155.0   228.5   269.8   280.7   328.6   560.0

# Mean, Median different. totalprice is skewed

par(mfrow=c(2,1))
hist(totalprice,xlim=c(100,600),main="",xlab="totalprice (000s)")
Boxplot(totalprice,horizontal=T,axes=F,ylim=c(100,600))
# [1] 13
par(mfrow=c(1,1))
# right skewed population

# Price of apartment in row 13 is an outlier

#b) average age with 0,1,2 garage slots
tapply(d0$age,d0$garage,mean)
#       0        1        2
#21.59880 14.55102  9.00000

# average age is 21.6 years
```

```
#c) apartments by garage and storage units
table(d0$garage,d0$storage)
#      0   1   2
#  0  39 128   0
#  1   3  45   1
#  2   1   1   0

# there are 128 apartments

#d) scatterplot
area = d0$area
plot(area,totalprice,pch=19,cex=0.6)
grid()
text(area,totalprice,labels=rownames(d0),pos=1,cex=0.5)

identify(d0$area,d0$totalprice,rownames(d0),cex=0.6)

#e) price of apartment with 100 m2
m1=lm(totalprice~area,d0)
a = data.frame(area=100)
predict(m1,a)
#        1
# 311297.5

#f) best predictor, area or age?
summary(m1)
# Residual standard error: 40810 on 216 degrees of freedom
# Multiple R-squared:  0.6548,    Adjusted R-squared:  0.6532
# F-statistic: 409.8 on 1 and 216 DF,  p-value: < 2.2e-16

m2=lm(totalprice~age,d0)
summary(m2)
# Residual standard error: 66830 on 216 degrees of freedom
# Multiple R-squared:  0.07423,   Adjusted R-squared:  0.06994
# F-statistic: 17.32 on 1 and 216 DF,  p-value: 4.563e-05

# based on R-squared
# area is better predictor than age
```


Frequency
0
20
40
60
100
200
300
400
500
600
totalprice (000s)


totalprice

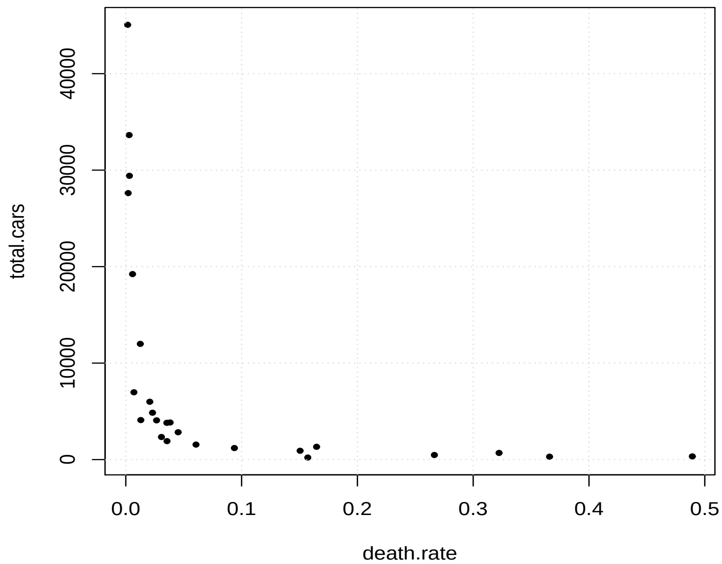

totalprice
200
300
400
500
60
80
100
120
140
160
180
area

2. (40 pts.) Use the CARS2004 data frame from the PASWR2 package, which contains the numbers of cars per 1000 inhabitants (cars), the total number of known mortal accidents (deaths), and the country population/1000 (population) for the 25 member countries of the European Union for the year 2004.
   Compute the total number of cars per 1000 inhabitants in each country, and store the result in a new column named total.cars. Determine the total number of known automobile fatalities in 2004 divided by the total number of cars for each country and store the result in a new column named death.rate.

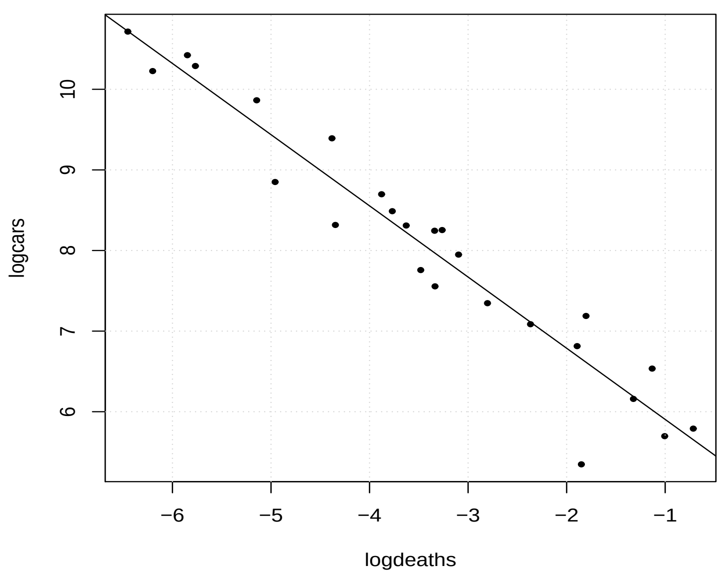
   a) Create a scatterplot of total.cars versus death.rate. How would you characterize the relationship between the two variables?
   b) Plot the natural logarithm of total.cars versus the natural logarithm of death.rate. How would you characterize the relationship?
   c) What are the least squares estimates for the regression of log(total.cars) on log(death.rate). Superimpose the least squares line on the previous scatterplot.
   d) What total number of cars (not log of cars) does this model predict for a country with a death.rate equal to 0.02305206?

```
library(PASWR2)
d0=CARS2004
dim(d0)        # [1] 25  4
total.cars = d0$cars * d0$population/1000
death.rate = d0$deaths/total.cars
d1 = data.frame(d0,totalcars=total.cars,deathrate=death.rate)
# head(d1)
#          country cars deaths population totalcars  deathrate
# 1        Belgium  467    112      10396  4854.932 0.02306932
# 2 Czech Republic  373    135      10212  3809.076 0.03544167
# 3        Denmark  354     68       5398  1910.892 0.03558548
# 4        Germany  546     71      82532 45062.472 0.00157559
# 5        Estonia  350    126       1351   472.850 0.26646928
# 6         Greece  348    147      11041  3842.268 0.03825865

# a)
plot(death.rate,total.cars,pch=19,cex=0.6)
grid()
# deaths decrease nonlinearly with more cars

#b)
logdeaths = log(death.rate)
logcars = log(total.cars)
plot(logdeaths,logcars,pch=19,cex=0.6)
grid()

# logdeaths decreases linearly when logcars increase
```

```
#c)
m1=lm(logcars~logdeaths)
# ls estimates
coefficients(m1)
# (Intercept)   logdeaths
#   5.0206666  -0.8833401

plot(logcars~logdeaths,pch=19,cex=0.6)
grid()
abline(m1)

#d) predict
a = data.frame(logdeaths=log(0.02305206))
b = predict(m1,a)      # 8.350859
exp(b)
# 4233.815 cars
```



total.cars
0
10000
20000
30000
40000
0.0
0.1
0.2
0.3
0.4
0.5
death.rate



logcars
6
7
8
9
10
−6
−5
−4
−3
−2
−1
logdeaths