Problem Set: Base R

Cristian Huse

(Based on James et al's ISRL, 2021)

This problem set uses the well-known Boston data set.

Questions

1. Load the Boston data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

Hint:

```
library(MASS) #alternatively, library(ISLR2) attach(Boston)
?Boston
Boston
```

2. Calculate the correlation matrix of the data. How does the outcome variable medv (what is it?) relate to the average number of rooms per dwelling and to pollution? Display these two relationships graphically.
3. How many of the suburbs in this data set are bound the Charles river?
4. What is the median pupil-teacher ratio among the towns in this data set?
5. How many of the areas average more than seven rooms per dwelling? More than eight rooms per dwelling? Report descriptive statistics for the suburbs that average more than eight rooms per dwelling.
6. Create a dummy variable indicating that an observation is in a high pollution area. To do so, define "hp" as having NOx levels in the upper quartile of the sample distribution. After defining is as a categorical variable, graphically compare median property values in high and low pollution areas.

Hint:

```
. . .
hp<-ifelse(. . . ,“Yes”,“No”)
dv_hp<-as.factor(hp)
plot(dv_hp,medv)
```

7. Now define a categorical version of the NOx variable according to the quartile of the observation. Define "noxq" and "noxq.f" as the numerical and categorical variables mapping nox onto quartiles. Graphically compare median property values across quartiles of local air pollution using "noxq" and "noxq.f". Think about the differences of those two variables.

Hint:

```
. . .
noxq<-ifelse(. . . possibly nested?. . . )
noxq.f<-as.factor(noxq)
```