Day 1 Closing Assignment

Pieter Schoonees, Dennis Fok & Andreas Alfons
Erasmus University Rotterdam

Create a new R script (that is, a `.R` file) in RStudio which will contain the code to answer the questions below. For all questions, include the code that you use to answer the question in your R script file.

Remember to include comments (lines that start with a `#`) in this text file so that you can remember which code relates to which exercise. Save your work under an appropriate file name, and remember to resave regularly.

1 House Prices

Exercise 1.1 (Data loading and exploration)**.** Download the house price data in the file `houseprice.RData` from the Canvas page for the *Programming & Visualization for Business Analytics* module by saving it directly to disk. Store it in an appropriate folder.

(a) Load the `houseprice.RData` file into your R session. Include the code to do so in your R script.

(b) Open the resulting data frame object `houseprice` using RStudio's viewer with the `View()` function (as always, include the code). Make sure you have reasonable answers to the following questions:

- What does a single observation (row) in this data set represent?
- Which variables do the data set contain?
- What units are these measured in?
- What types of variables are these?

(c) How many observations (rows) are there in this data set? How many variables (columns) are there? Include code to obtain these values.

(d) Use the `plot()` and `summary()` functions on the data frame object to see what output this gives you. Note down some of the key points that you have learned about the data from the output of these two functions.

Exercise 1.2 (Univariate summary statistics)**.** Now that you are familiar with the basics of the data, let's answer some simple but important questions. Remember that you can extract a single column vector from a data frame using the `$` operator, as in `dataname$variablename`.

(a) What is the mean price of a house in this data set? Include code to calculate this using the `mean()` function.

(b) How does that compare to the median price? Can you say something about the expected skewness of the data based on the mean and median (advanced)? We will look into this more closely when doing plots in the next exercise.

(c) What is the price of the cheapest house? And of the most expensive house? Remember the functions `min()`, `max()` and `range()`. What results do these respective functions report?

(d) Calculate the range of the house prices, as well as the standard deviation.

(e) What values do the variables `airco`, `driveway`, `fullbase`, `gashw`, `prefarea` and `recroom` take? Make frequency tables of counts for each of these variables separately using the `table()` function. What types of variables are these?

(f) How many houses have six bedrooms? How many have five bedrooms? Include code that shows that this is the case.

Exercise 1.3 (Univariate plots)**.** Let's investigate the marginal (univariate) distributions of some of the variables using graphs.

(a) Graphically show the distribution of the house prices in this data set using a histogram. Play around with the number of bins and / or the width of these bins.

(b) Produce a bar plot of the `stories` variable. How many houses have four stories?

(c) Is the `lotsize` variable normally distributed? Investigate using a quantile-quantile plot and a density plot.

Exercise 1.4 (Bivariate relationships)**.** Finally, let's look at binary relationships between variables using graphs of two variables and cross-tabulations.

(a) Investigate the relation between `price` and `lotsize` using a scatterplot with `price` on the vertical axis. Can you describe the general relationship in words?

(b) Use the `table()` function with two arguments (as in `table(x, y)`) to investigate the relation between the number of bedrooms and bathrooms in this data set. How many houses are there with four bedrooms and two bathrooms?

(c) Investigate whether houses with many bedrooms tend to be in preferred areas or not. *Bonus*: Use the `prop.table()` function to turn the counts in your contingency table into proportions.

(d) One may expect that houses with many places in the garage also have a large lot size. Create a graph to investigate this idea, using conditional boxplots.

2 Additional exercise: Volkswagen prices

Exercise 2.1. The `euR` package contains a data set `vwgolf` containing the prices of used VW Golf cars. Perform a similar analysis as above on this data set. For example, answer the following questions in your commented R script.

(a) Load the data from the package, and read the help file. Include the commands used.

(b) Produce summaries of the univariate distributions of the variables. Are there missing values? Are there any outliers?

(c) Create a scatterplot of `Mileage` against `AskingPrice`. Why do you get a warning message? Can you ignore it?

(d) Create a scatterplot of `Mileage` against `PriceNew` minus `AskingPrice`.

(e) Create a histogram and density plot of `Mileage`.

(f) Create boxplots of `Mileage` conditional on `Fuel`.

(g) What is the minimum, median and maximum number of owners a car has had?

(h) What are the quantiles of `PriceNew - AskingPrice`?

(i) How many diesel cars have automatic transmission?

(j) What is the minimum and maximum of `Mileage`? Ignore the missing values on this variable using the additional argument `na.rm = TRUE`, as in `mean(x, na.rm = TRUE)`.

Exercise 2.2. Now, get creative and come up with your own questions to investigate.