MS-C1620 Statistical Inference

Exercise 2

Homework exercise

To be solved at home before the exercise session.

---

1. Visit the website https://datavizproject.com/ and pick one data visualization/plot that interests you. Find out how it is drawn and what aspects of the data the different components represent. Be prepared to explain how your visualization of choice works in the class.
2. Type the command `data()` in **R** to show all data sets currently available in your installed packages. Go through the data sets and pick one that interests you. Check the help file of the data set using the command `?packagename` for more detailed information. Be prepared to describe your answers to the following questions in the class:
    - What is the purpose of the data? What kind of phenomenon does it describe?
    - What kind of study is behind the data (observational, controlled, simulation, survey or something else)?
    - How is the data represented in R (univariate, multivariate, time series…)?
    - What kind of plots would you use to best summarize the data?
    - What kind of numerical statistics would you use to best summarize the data?

Class exercise

To be solved at the exercise session.

*Note: all the needed data sets are available in base **R**.*

---

1. The data set `rivers` contains the lengths of 141 major rivers in North America.
    a. Find a suitable way to visualize the data and plot it.
    b. How are the lengths distributed based on your plot?
    c. Discretize the lengths into six classes: [min, 250], (250, 500], (500, 750], (750, 1000], (1000, 1250], (1250, max]. The function `cut` may prove helpful.
    d. Find a suitable way to visualize the discretized data and plot it.
    e. Which of the two visualizations is more informative?

---

2. The data set `islands` contains the areas of all landmasses in the world which exceed a certain threshold.
    a. Find a suitable way to visualize the data and plot it.
    b. How are the landmasses distributed based on your plot?
    c. Compute both robust and non-robust measures of location and scatter for the data.
    d. Remove some of the outliers (and think of a possible reason for justifying this!) from the data and compute the same measures as in part c.
    e. Compare the results of part c and part d.

---

3. The data set `Nile` contains yearly measurements of the flow of the river Nile.
    a. Find a suitable way to visualize the data and plot