Exploratory Data Analysis and Visualisation 11374/11517

Week 2 Lab Exercises

1. Work through the “Introduction to R” script. This will help you become more familiar with the capabilities of R and the coding syntax. You will also be introduced to some default code for creating plots.

2. Consider the “mammals” data frame from the MASS package in R, which gives body weights and brain weights for 62 mammals.
    a. Install and load the MASS package in R.
    b. Read the help description about the “mammals” data using “?mammals” in R.
    c. View summaries of the “mammals” dataset using the dim(), names(), summary() and head() R functions.
    d. Create a scatterplot of brain weight against body weight. Identify the outliers. Which mammal has the largest brain weight? (Hint: use the “identify” function – use “?identify” in R to read the help file).
    e. Which mammal has the largest brain weight relative to their body weight? Plot the brain to body weight ratio against brain weight.

3. Download the “airquality.csv” file from Canvas. It contains daily air quality measurements in New York, May to September 1973 on the following variables:
    - Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
    - Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
    - Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
    - Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport

    a. Import the “airquality.csv” file.
    b. Install and load the “dplyr” and “naniar” packages.
    c. There are missing values in this dataset. How many missing values are there for each variable? Use the “gg_miss_var” function to plot the number of missing values for each variable.
    d. Create two subsets of the original dataset, one called “airquality_missing” which only contains observations with missing values and the other called “airquality_notmissing” which does not contain the observations with missing values.
    e. Using the “airquality_notmissing” dataset, find the mean for each variable by month. Compare and rank.
    f. From the “airquality_missing” dataset, create a subset called “airquality_missing_Ozone” which only contains observations with missing Ozone values. What are some concerns you have regarding the missing Ozone values?
    g. Create a dataset called “airquality_Ozoneimputed” and impute the missing values for Ozone using median values by month.
    h. Create the final dataset called “airquality_final” which excludes the “Solar.R”, “Month” and “Day” variables.