Exercises 2

1. Variables in literature

Identify the variables used in the given parts of the following two articles and determine their types. (NOTE: You may have to browse through the whole article in order to fully understand the study.)

- Article 1 (whole study): L.E. Juarez-Orozco *et al.*, Machine learning in the integration of simple variables for identifying patients with myocardial ischemia, DOI: https://doi.org/10.1007/s12350-018-1304-x
- Article 2 (Figures 1 & 2): W. He *et al.*, High-salt diet inhibits tumour growth in mice via regulating myeloid-derived suppressor cell differentiation, DOI: https://doi.org/10.1038/s41467-020-15524-1

2. Cyclists

The files `cyclists-helsinki.csv` and `cyclists-espoo.csv` contain daily numbers of cyclists spotted on selected streets in Helsinki and Espoo.

- Load the files and merge the data into a single data frame.
- For how many days were observations made in total?
- How many observation days were there for each street?
- On how many days were all streets observed simultaneously?
- Which street was the busiest in terms of the total number of cyclists?
- Filter out the dates which have one or more missing values. Does this affect your conclusion about the busiest street? Why or why not?

3. Human heights

- Create a histogram and a density plot of the following two sets of data points, which contain human heights measured in centimeters:
  - 170, 192, 184, 168, 176, 181, 163
  - 170, 170, 170, 170, 192, 192, 192, 192, 184, 184, 184, 184, 168, 168, 168, 168, 176, 176, 176, 176, 181, 181, 181, 181, 163, 163, 163, 163
- Based on the plots, would you consider the distributions to be normal? How confident are you about your conclusion?
- The data sets contain similar values, but your conclusions may differ. Can you explain such a difference in your results?
- (BONUS) Use a statistical test to assess the normality of the distributions. (NOTE: The test needed here will be introduced in the topic 4.)

4. World temperature

- Download the data set from https://climate.nasa.gov/vital-signs/global-temperature/ and consider the `No_smoothing` variable.
- Calculate the mean and the median of the data.
- Create a histogram and a density plot for the pre-2000 measurements. Does the variable seem to be normally distributed?
- Create a histogram and a density plot for the measurements from year 2000 onwards. Does the variable seem to be normally distributed?
- (BONUS) Use a statistical test to assess the normality of the distributions. (NOTE: The test needed here will be introduced in the topic 4.)

5. Electric bikes (continues)

Continue to analyse the data you handled in the earlier exercise. The descriptions of the variables are given below.

- What types are the variables? (Consider as many categorisations as possible.)
- Check that the data types and values in the data you have loaded match the variable types. Fix if needed.

Variable	Description
ticket	ticket type
cost	paid fee in euros
month	calendar month during which the trip was made
location_from	start location of the trip

Variable	Description
location_to	end location of the trip
duration	travel time in seconds
distance	travel distance in meters
assistance	status of electric assistance (0 = disabled, 1 = enabled)
energy_used	energy consumed by the bike in watt-hours
energy_collected	energy collected by the bike in watt-hours