Exercise Sheet 2 - Working With Data

Timo Meiendresch

WS 20/21

1 Data Frames

1.1 Exercise: `flights` data frame

- Install (if necessary) and load the library `nycflights13`
- Have a look at the dataset `flights`, which is included in the library
  - Check the data using `?flights`. According to the description, what is the "Format"?
  - Simply print `flights` to the console. Compare this to the call `head(flights)`.
  - Check the class of the `flights` data.
  - Note that `tibbles` can be used similar to how we worked with `data.frames()`
- How many variables and observations does the dataset have?
  - The dataset has [...] observations and [...] variables.
  - Each observation (row) represents [...]
- What is the mean (median) `distance` of all the flights in the dataset?
  - What is the standard deviation of the `distance` variable of the flights?
  - How many flights are farther than the mean distance? Use vector arithmetic!
- How many flights had at least 20 minutes delay at departure?
  - What is the mean distance conditional on flights that had at least 20 minutes delay at departure?
  - Hint: The sum and mean function can strip `NA` values.
- Dealing with missings (`NA`s): Check the description of the `is.na()` function.
  - Use the `is.na()`-function to count the number of missing values (`NA`) of the variable `dep_time`

Good job! We'll revisit this dataset in later sections of the course.

1.2 Exercise: Data Frames, Loops, `lapply`

- Create a data.frame, `df` according the following specifications:
  - `df` contains four variables, `a`, `b`, `c`, `d`
  - Each variable contains 10 random values of the standard normal distribution (`rnorm()`)
  - Use a for loop to loop over `df` and print the median
  - Hint: Using `[i]` preserves the `data.frame` structure, while `[[i]]` returns a vector
  - Modify the loop to save your results as a vector.
  - Hint: Initialize the vector before using it in the loop. For this, you may use the function `vector()`, `numeric()`, `rep()`, etc.
- Instead of a for loop, use `lapply()` on your previously defined Data Frame `df` to achieve the same result as before, i.e. save your results as vector with the name `output2`.
  - Hint: `lapply` returns a list, so you have to simplify it to a vector here.
  - Compare `output` and `output2`

1.3 Exercise: Functional Programming

R is a functional programming language. Therefore, for-loops are not as important as in other programming languages and can easily be avoided. To show this, we refer back to previous exercise sheet and use `lapply`

instead of a loop.

Given is the following list:

```
cities <- list("Barbados", "Sankt Augustin", "Aachen", "Cologne")
```

- Use `lapply` on the `cities` list to return the number of characters as a vector

The result should look like this:

```
[1]  8 14  6  7
```

1.4 Exercise `web-browsers.csv`

- Read in the `web-browsers.csv` file

This main variable `spend` captures the time people spend online (in hours per year) and a handful of socio-demographic variables.

- Check the structure of the dataset
  - How many variables and observations do we have?
  - Of what class are these variables?
  - How many people of `hispanic` descent are in the dataset?
  - What is the overall percentage of people having broadband?
  - Calculate two conditional means of time spent online. Conditional on i) `anychildren == 1` vs. ii) `anychildren == 0`?
- The next questions relate to the `spend` variable:
  - What is the standard deviation of the `spend` variable?
  - What is the max and min value?
  - Compare mean and median value
  - Are mean and median equal? If not, try to explain your findings.
  - What is a quantile and its relation to the median?
  - Find a function to calculate quantiles in R and apply it to the `spend` variable

Histograms are used to visualize the distribution of single variables.

- Check the distribution of the variable `spend` using a simple (base R) histogram (`hist`-function)
  - Apply a log transformation

Boxplots are used to compare quantiles of continous variables versus group membership of another variable, i.e. continuous vs. categorical variables.

- Plot logarithmic spending (`spend`) versus the `anychildren`-variable to compare time spent online in the group of people with vs. without children.
  - Hint: Remember the `factor()` function?
- Based on the results so far: Do you think that time spent online depends on whether people have children or not?
- Use a linear regression model (`lm`) and try to explain quantitatively time spent online (logarithmic) with the following variables:
  - `anychildren`, `broadband`, `hispanic`
  - Hint: The argument `formula` within the `lm()`-call starts with `log(spend) ~`

1.5 Exercise: `murders` data frame

- Install and load the library: `dslabs`
- Check the description of the `murders` dataset

  - Hint: `?murders`

- How many observations and variables do we have?
  - Of what data type is the variable `state`?
  - How many levels does the variable `region` have?
- Take a closer look at the `population` variable:
  - What is the mean population size?
  - How many states have a population of at least 3 times the overall population mean?
  - What are the names of these states?
- Have a closer look at the `state` variable
  - print out the first five states
  - Which states had a higher murder total than the average murder total?
- Why is it not a "fair" comparison to compare murder totals of states to the mean murder?
  - What is a better comparison?
  - Implement your idea!
- Show number of states per region using the `table` function
- Sort total murders using `sort()`-function
  - What are the highest (lowest) numbers of total murders?
  - Can you identify the respective states?
  - Can you sort the whole `murders` data frame according to the variable `total`?
  - Hint: Have a look at the `order()` function.
- The functions `which.min()`, `which.max()` determine the location, i.e. index of the (first) minimum or maximum of a vector.
  - Use these functions to return the name of the state with the lowest (highest) murder total
- Define a murder rate variable (`rate`)
  - Add this variable to the `murders` data using the `$` operator.

Murder rate per 100.000:

$$murder\ rate = \frac{murders}{population} * 100,000$$

- Identify states below a murder rate of 0.42
  - How many of these relatively safe states can you find?

2 Find a data visualization

Find a data visualization in a medium of your choice.

- Answer the following questions:
  - What do you expect the underlying dataset "looks" like? Variables and number of observations?
  - Mapping of the variables to which element of the visualization (i.e. x-axis, y-axis, color, size)
  - Type of plot? Line plot, Scatterplot (points), columns, bars,...
  - What story does the data visualization try to deliver?

Please email me **only** the data visualization (without any explanation), included in the mail corpus by wednesday 23.59h to timo.meiendresch@fit.fraunhofer.de

In addition, please indicate how much time you spend on this exercise sheet 2, so I can adjust the workload accordingly.

Thank you!