Homework 1

General instructions

- The following homework is to be done in pairs (groups of two students). The outcome to be submitted is a single zip-file containing exactly one .Rmd file, one PDF file, one .rds file and one .RProj file.
- Create an R-project to solve the tasks below. Zip the directory containing the .Rmd, .pdf, .rds and .Rpoj files and upload this single zip file as your solution. The .pdf-file is the output of your .Rmd file and should not exceed 11 pages.
- Only libraries introduced in class or mentioned in this homework are allowed. Refer explicitly to the slides you used for each task.
- In the last page describe your collaboration as well as the workload (individual and overall) for this homework.

Tasks

1. Using a single pipeline and `nycflights13::weather`, find out the number of days on which the temperature was below 10°C at Newark-Airport, broken down by month. Do these months correspond to what you expected?
2. Using a single pipeline and `nycflights13::flights`, find out January flights missing `dep_time` and get all other variables the flights missing `dep_time` are also missing. What might these rows represent?
3. Select all variables from `nycflights13::weather` except those between `year` and `hour`. Use solely this selection to solve the tasks a), b) and c) below. For the diagrams in a) and b), label the months with their full name. For the diagram in c), use number to label the months.
    a. Display the distribution of the temperature faceted on months
    b. Use `ggridges::geom_density_ridges()` to get a plot similar to the one in a).
    c. Aggregate your data into averaged monthly temperature, including their standard deviation. Use the aggregated data and `geom_line()` to display monthly average temperature surrounded by the $\pm 3 \cdot \frac{s}{\sqrt{n}}$ upper and lower 99% confidence lines, where s is a monthly standard deviation and n the number of observation per month.
4. `Data_HW2.xlsx` has in the sheet `A very small sample` some data collected on BFH-W 3rd semester students a couple of years ago. Most of the variables are self explaining.

- `foot` gives foot size measured in cm without shoes.
- `hair` is the length of respondent's longest hair measured in cm.
- `transport` is the main method of transport used by the respondent to get to the BFH.
- `cash (CHF)` gives how much money cash in CHF the student had with her/him at the time she/he was filling the survey.

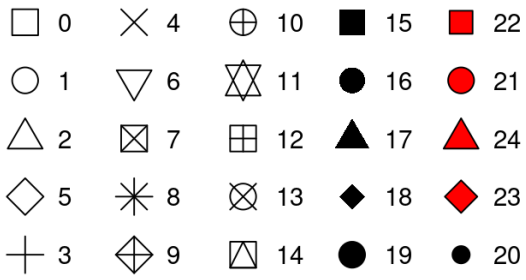
Import this data into `R` and inspect it. Describe any problem you notice and propose a solution for it. Save the cleaned data as `.rds` in your R-project directory.

5. `anscombe` illustrates how graphs are essential to good statistical analysis. These data are from four samples of eleven observations each. The index of `x` and `y` represents the corresponding sample.

For example the first observation in sample 1 has the x-value 10 and the y-value 8.04, and the first observation in sample 2 has the x-value 10 and the y-value 9.14.

a) Explain briefly why this dataset is not tidy. In a single pipeline, transform it to get a tidy-dataset with exactly three variables: `sample`, `x` and `y`. Do not output the tidy-dataset, but save it for future use.

b) In a single pipeline, use the tidy-data from a) to get a scatterplot that includes a line of best fit for `y` on `x`. The sample should be distinguished by color and shape. Use only the shapes 12, 10, 9 and 7 below.



c) In a single pipeline, summarize the tidy-data from a) to get the means, standard deviations and correlation coefficients. I expect from you the table below. You can use `kableExtra::kable()`.

sample	mean_x	mean_y	sd_x	sd_y	corr_xy
1	9	7.501	3.317	2.032	0.816
2	9	7.501	3.317	2.032	0.816
3	9	7.500	3.317	2.030	0.816
4	9	7.501	3.317	2.031	0.817

d) In a single pipeline, use `facet_grid()` to display the scatterplots per sample (including lines of best fit), and summarize the main message on your plot. You can get the paper by F.J.Anscombe here.