Overall description

For the following two tasks, please write a reproducible PDF report (a report that generates at compilation a PDF document with output based on provided data). You can choose for this task one or more of scripting languages, such as R, Python, or Julia (including all libraries that are available for that language e.g., C libraries in R) and upload all raw files (without the data files!) into the Github repository https://github.com/dkalisch/nordlb_YOURNAME. The raw files need to be able to compile on any given computer that has the languages and dependencies installed!

You may **NOT** ask anyone else for help during this exercise. You may **NOT** include any content that is not your own, and you may **NOT** include any content that you generated prior to the exam period for class, work, in preparation for this exam, or for any other purpose. You are **NOT** allowed to share any part of this exercise, neither data nor description, with another person.

You will have 72 hours to complete the exercise, but you may return the exercise at any point during that period. This exercise should not take the entire time – we give you ample time to help you fit this exercise into your schedule.

Please keep in mind that this exercise is not designed to be overly burdensome in terms of the time commitment. If you are devoting over 15 hours of work on this exercise, you are probably spending too much time and going beyond the scope of this exercise.

If you have questions during the exercise, please email dominik.kalisch@nordlb.de

Flight delays

Objective

For the first part of the exercise, we ask you to download (with a script if you can) the *complete* data from https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time for the entire year 2016 and write a script that imports the data into a PostgreSQL database with adjustable specifications. Your objectives are to create a script that performs a *full* descriptive statistics for the flight delay in San Francisco. Try to offload as much of the calculations as possible to the database level and investigate what the most common factors for a departure flight delay at the San Francisco airport are. Analyze if there is a significant difference of delay between the carriers at the San Francisco airport in 2016. If so which carriers are better and which worse? Conduct a *complete* analysis for this part. Finally, create an interactice dashboard that dynamicly showcast your work.

Report Writing

After completing the analysis, you are asked to write a report in either German or English explaining how you solved the task, describing the outcome of the descriptive analysis, the factor and group comparison, as well as the model construction. Please discuss any data, technical or statistical challenges you encountered while working on this task, and how you resolved these issues. This model should convey technical/statistical concepts while still being accessible to an educated but non-technical audience. Here are some topics that should be addressed in the report:

- *Language Selection*: There are multiple scripting/programming languages to solve this task. Please describe which language you used and why.
- *Variable Selection*: The downloaded files contain a lot of data and not all variables are necessary to for this task. Please describe the main variables that you used and included in your analysis. You should also discuss some of the variables you considered for inclusion but ultimately excluded.
- *Technical Aspects*: This task is a more technical one where you need to weigh different options. Please describe your thought process and why you choose a certain procedure over a different one.
- *Assumptions*: Analyses like this are based on certain assumptions. Please describe the assumptions that you made and how you tested them.
- *Visualization*: Your final report should include at least one data visualization for each part that could be used as a diagnostic tool to evaluate the assumptions and to describe the data.