Data Analysis and Visualization in R

Case Study I

Felix Brechtmann, Julien Gagneur

Case Study I: 2015 Flight Delays and Cancellations

from Kaggle:

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

Dataset

The data set contains 3 files:

- one with the airline names and codes
- one describing the airlines
- one listing the flights, their scheduled and actual times, and further information.

A more detailed explanation of the features can be found here: Kaggle (https://www.kaggle.com/usdot/flight-delays/)

Dataset 2

- We assembled a smaller data set containing all flights to and from Los Angeles (on moodle).
- You have to use at least this subset
- and can use the entire data set if you want to.
- You are allowed to use any additinal data which is publicly availiable.

Case study

Goal: Give a presentation where you show patterns or trends in the data you found interesting.

The presentation should take 10 min (7 min presentation + 3 min questions).

- You work in groups of 4 people.
- Sign up on the group sign-up sheet here: https://docs.google.com/spreadsheets/d/1YG6wilur52v99vO6M2BW4VvDH1eVCdMbC_Tf6qDjXz4/ec (https://docs.google.com/spreadsheets/d/1YG6wilur52v99vO6M2BW4VvDH1eVCdMbC_Tf6qDjXz4/e
- You should spend around 24h each on the case study.

Case study

- There will be no lecture on the 10th Dec 2019 instead the professor and some tutors will give advice on demand in the usual lecture hall.
- The deadline for submission is noon 11th Dec 2019, via moodle only.
- Presentations will be given in the tutorial rooms in the week from 9th to 13th Dec as listed in the group sign-up sheet.
- Submit one .zip file named as gXXX-Lastname1.zip where XXX is your group number (Example: g001-Smith.zip)
- This zip file should contain an R-markdown file which allows to regenerate all figures from raw data input files, the pdf output of this R-markdown and optionally the presentation as pdf slides and all external data sets used (if any).

Instructions for the presentations

The presentation should cover:

- 1-2 slides on data preparation
- result slides, where the title states the finding (what do we observe?) and not the methods (how do you do it?). For example a good title sates "The further away an airport, the longer the flight" and not "Scatter-plotting air time and distance".
- All labels should be legible for the audience (big fonts). Axis and colors and shapes should be labelled.
- Show only relevant code
- Ideally build up a story where you dig step by step into details.
- Finish with a conclusion slide recapping the main findings

An example slide is given next...

The further away an airport, the longer the flight

```
flightsLAX <- fread('./extdata/casestudy1/flightsLAX.csv')
ggplot(flightsLAX, aes(DISTANCE, AIR_TIME)) +
    geom_point(alpha=0.2) + geom_smooth(method='lm')
```