Data Visualization – Assignment 2

– Exploratory Data Analysis –

P. Rüdiger, J. Lukasczyk · H. Leitte – winter term 2018/19

The goal of the second assignment is to explore a dataset on your own, learn to ask the right questions, and develop good designs that tell a compelling story.

On the technical side, we will now transition to python using jupyter notebooks, the pandas library for data analysis, and bokeh for plotting. Combining these tools, we will combine analysis and documentation into a single document containing nicely formatted text, code, and interactive graphics.

The assignment is structured in four exercises: In the first, you can get used to the new environment and start thinking about your dataset. In the second, you have to formulate questions with respect to the data. In the fourth, you will use bokeh to explore the data with respect to your questions. And in the last, you will summarize your findings.

Due date: 22. November 2018, 22:00 (10pm)

Submission instructions:

- Work in **teams of 2**.
- Please submit your solutions as a **single jupyter notebook** plus an **HTML-export of the notebook** per team ("Submissions → Assignment 2").
- We provide you with an **initial jupyter notebook** for your submission.
  - Please keep on working on this notebook in this assignment.
  - Delete sample code you do not need.
- If you are new to jupyter notebooks, here is the **quick start guide**: `https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/index.html`

Exercise 1 – Choose a dataset

1 points

Before we get into the analysis, take a look at both datasets. You can find the two datasets in the material folder. You can open the files with an editor or Excel. In the provided jupyter notebook we give you code to open the data in it and print some statistics:

- **Baseball**:
  - Brief description: This dataset contains information about baseball players, their physical characteristics, and their performance.
  - Variables: name, handedness (right or left handed), height (in inches), weight (in pounds), batting average, and number of home runs
- **Titanic**:
  - Brief description: This dataset lists all passengers on the the ocean liner Titanic along with their economic status (class), gender, age and survival.
  - Data description `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html`
  - Variables `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/Ctitanic3.html`

Rename the provided jupyter notebook (`Assignment2_....ipynb`) to match your names. Open a terminal and go to the folder that contains the notebook. Enter the following command to start jupyter notebooks:

```
$ jupyter notebook
```

Click on the notebook you want to open to start it. In the open notebook, update the header, take a look at the data, and enter the selected data set name in exercise 1.

Exercise 2 – Data queries

5 points

Before starting with a detailed exploration of the data, think about the questions a person may have with respect to your selected data set. List six analytics queries or questions that can be answered using the given variables or by values derived from them.

Exercise 3 – Data analysis

12 points

In the following you shall design visualizations using Bokeh that help answer your questions. Share the questions between the two team members - each takes three of them and do the following for each question:

- Write down the question.
- Create one or multiple charts.
- Try to answer the posed question using your designed chart.
- Take notes on difficulties, ideas for improvements, ideas for further investigations, etc.

Requirements:

- Use Bokeh for the visualizations. See the provided jupyter notebook for an example.
- Design the charts on your own using Bokeh's class figure. Do **not** use high-level charts!

Exercise 4 – Write a summary

6 points

Write a one paragraph summary about what you found out about your dataset, your difficulties and how you would continue the exploration.

Contact: ruediger@rhrk.uni-kl.de, leitte@cs.uni-kl.de (Please only use in urgent cases. Questions are answered after the lecture and during the exercises.)