Data Visualization – Assignment 3

– Visualizing multivariate data –

P. Rüdiger, J. Lukasczyk · H. Leitte – winter term 2017/18

The goal of the third assignment is to give you practical experience in analyzing high-dimensional data. You will implement statistical charts which will serve as input for a given generalized pair plot. Using these techniques, you will also analyze tabular data.

The assignment is structured in three exercises: In the first one, you will implement statistical charts to extend the bokeh plotting facilities. You will have to combine skills in data analysis (pandas side) and plotting (bokeh). We also want you to think of shortcomings, test cases, and possible extensions to your implementation. The second exercise targets the analysis of individual plot types and aggregation of information across multiple plots. In the last exercise, you will work on a new dataset and summarize the findings you can make using your SPLOM.

Due date: 06. December 2017, 22:00

Exercise 1 – Statistical charts

12 points

Your first task is to build the chart components for the generalized pair plot. You are given the code for the generalized SPLOM (splom.py, see next exercise), but it currently misses the code for the individual charts especially those handling categorical data. Write routines for the following statistical charts:

a) **Scatterplot** (2P)

b) **Boxplot** (5P)

c) **Histogram** (5P)

More details and specifications of each method are given in the jupyter notebook. The test cases ensure that it is compatible with the current splom-implementation.

Exercise 2 – Working with SPLOMs

6 points

Next we will look at how to combine findings of data with multiple variables. In the jupyter notebook, you find code that loads the SPLOM, the respective data, and renders a SPLOM of it (if you implemented all subplots correctly.) In the assignment folder you will also find the respective images to check your implementation and work on the exercise if your code does not work (correctly).

Exercise 2a) (3P): Use the SPLOM of the Baseball dataset and characterize each histogram. For each histogram, do the following

- Describe the shape.
- Speculate what could be the reason for this shape.
- Look if other plots give evidence for your assumption.

Exercise 2b) (3P): Use the SPLOM of the Titanic dataset and characterize the passengers in each class and compare them:

- What can you tell about passengers in each class?
- Where are differences between the classes visible?

Exercise 3 – Analyzing data with SPLOMs

6 points

For the third exercise you are given two datasets:

- **Auto MPG** The cars dataset we partly looked at in the lecture. The goal is to analyze fuel consumption of approx. 400 cars based on 8 variables. Information: `https://archive.ics.uci.edu/ml/datasets/Auto+MPG`
- **Iris** The iris flower dataset is a second classic in data analysis. The goal is to characterize and distinguish the three types of iris flowers. Information: `https://archive.ics.uci.edu/ml/datasets/Iris`

Select one of the datasets and analyze it. Your analysis shall include the following:

- A **dataset description** (1P) (what is the data about, which variables)
- An **analysis protocol** (2P). Document your analysis process: Which plot are you looking at? What did you find? Which conclusions did you draw? What is unusual? Give your findings as a bullet list.
- A **summary** (3P). Based on the information you found and documented above, compile a summary analysis that guides the reader through the discoveries you made concerning the data.

This exercise shall take you roughly one hour. Do not make a full analysis. See how far you can go and document as described above.