Individual Report PAD Project

Analysis of Beijing Air Quality Data

Group 09: Annemarie Witschas, 59051

20 April 2020

In our group (Group 9) we have searched for a data set together and while Aron and Rebekka started on the Linear Regression part, I was working on the PCA task. Later, when the second part of the project was added, Aron resumed to work on the Clustering task. This way, we have mainly split up each of the three parts onto one person, but we were also discussing questions of the different parts frequently with each other.

I was also preprocessing the data (filtering relevant columns and rows, removing missing values, normalizing, calculating the groups ect). I have to say that I learned a lot already on that part, because I have not worked much with the pandas library before and through the routine of this project I now feel profoundly more confident handling large dataframes.

During the PCA task I have applied Singular Value Decomposition. Although I have used the inbuilt numpy function, I still had to think a lot about the background of PCA and SVD: when PCA/SVD shows a good projection, how standardizing the data can affect it, ect.

While visualizing the results I have also become a lot more accustomed with Matplotlib and seaborn than I have been before. After reading many pages of API and stackoverflow questions, I have acquired a much better expertise with designing and adapting my own graphs and visualizations. These skills seem particularly useful to me, as they can be of great use in any domain.

Moreover, what I also appreciate to having learnt throughout the project is the whole workflow surrounding data analysis. In previous courses in my home university due to time reasons we have often only done single parts ourselves, such as implementing parts of algorithms. To get a feeling for the larger process involved, starting with finding a suitable data set (which is not as trivial as it sounds, as I know now) over getting to know approaches on how to investigate that data all the way to visualizing and discussing the results is a great aid in getting a better understanding of the field of Data Analysis.