Wrangle Report

In the udacity course “Data Analyst” project 5 was about data wrangling. In this report I will summarize the steps I took during this project. A Jupyter Notebook was used for the detailed documentation.

I structured the project in the following way:

1. Gathering
2. Assessing
3. Cleaning
4. Storing
5. Analysis
6. Conclusion

In the first step (Gathering) there were three sub steps. First a csv file was provided by the course which I transformed into a data frame. Here were information about the tweets from the Twitter account WeRateDogs included. Second I downloaded a tsv file using the requests library from a provided url and transformed it into a data frame as well. This data frame contained data from a neural net in order to identify the breed of the dogs. The third step used the Twitter API. After creating a developer account on Twitter I used the tweepy library to query the Twitter API. I used a list of tweet IDs from the provided csv file in order to get all the status information from the API and stored these in a json file. Afterwards I read the needed information out of json file (twitter ID, likes, retweets) into a list and created a new data frame out of this list. This concluded the gathering process and I started with the next step assessing the data.

With various functions I inspected the data and made notes about subjects I wanted to clean in the next step. This resulted in eleven quality issues and three tidiness issues. The quality issues contained things like replies and retweets were not useful for the analysis of the data, NaN values were stored as strings “None” and some problems with the used rating system (decimal numbers weren’t sliced correctly out of the text). The tidiness issues were mainly that the data was spread over three data sets and should be combined into one and different values which described the stage of the dog were in different columns.

In step three I cleaned the identified issues from step two. I always documented the issue, a definition how to clean it, the code and finally a test to check if the cleaning was successful. Most quality issues were solved by dropping rows and/or columns. In regards to the rating system a normalized value was calculated and wrong values (due to decimals) were replaced. The tidiness issues were resolved by merging the data sets and the columns with the dog stages.

In the fourth step the final data set was stored into a csv file called twitter_archive_master.csv.

During the analysis I looked at outliers in the rating, correlation between likes and rating as well as retweets and rating and the distribution of dogs according to their breed.

Finally I summed up the findings from the analysis in the conclusion. For detailed formation please refer to the Jupyter Notebook itself.