Wrangle report

The wrangle project was done by gathering, assessing, cleaning and analyzing the data from WeRateDogs Twitter. Therefore each step was implemented one after another in order to build a high quality and tidy data base to make interesting and trustworthy analyses.

Data Gathering

Three different data sources were provided and loaded into a dataframe in a Jupyter Notebook. The twitter archive was provided as csv file and could be gathered easily into the Notebook. Second, a image prediction file needed to be gathered by downloading it via the given url and safing it into a dataframe. This information enriched the basic tweet information from the archive file. Additionally, the tweets list could be accesed via the Twitter API and safed as json format into a txt file. From there, a json file was loaded into the Notebook and safed as a dataframe to access the information.

Assessing the data

The data was inspected visually and programmatically. Various quality and tidiness issues could be detected and the observations were defined for the cleaning process. Dirty Data is low quality data and needs to be removed from the data set. For example for this special purpose only original tweets were allowed for the analysis so retweets and replies needed to be deleted. Also for a smooth analyzing process the correct data type of the variables, e.g. datetime, and intuitive and correct column names needed to be ensured as well as the uniform spelling. Also for the purpose of this project, not necessary information like the url was dropped so the dataframe was good to handle. Some tidiness issues were adressed. Above all the dog class was put into one column as each variable needs to form only one column. Therefore some information was seperated into two columns, e.g. date of the tweet and time of the tweet. Dataframes were merged together as each type of observational unit should be listet in one table.

Cleaning the data

Data cleaning was achieved through programmatic code. The observations from the data assessment step were defined and coded as well as tested. In this process

some issues occured, e.g. in some cases a merge based on tweet ids was not successful as there was different ids and the data needed to be saved in different tables.

Analysis and Visualization

Finally the data could be analysed. The analysis is in a seperate report.