WeRateDogs Project
Udacity Nanodegree Data Analyst

Wrangle Report

1. Introduction:
2. Gathering data
3. Assessing data
4. Cleaning data

1. Introduction:

Within the project there the data of tweet archive of Twitter user @dog_rates, also known as WeRateDogs was wrangled. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The Wrangling Process is separated in the Gathering, Assessing and Cleaning Part of the data, which are are described in the following sections with regard to the WeRateDogs Project.

2. Gathering Data:

In the project three different kind of data sources has been used, needed therefore to be gathered and was imported to the jupyter notebook.

A. **Twitter_archive_enhaced.csv** - The WeRateDogs Twitter archive, which was manually downloaded to the **data** subdir in of the git repository.

B. **image_predictions.tsv** - The tweet image predictions provided by Udacity. The result of predictions of a neural network for objects of an image. The file was programmatically downloaded to the **data** subdir in of the git repository. The URL used was the following: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

C. **tweet_json.txt** - The tweets retweet count, likes and additional data, which had to be provided by Udacity. For personal reasons I am not allowed to reactivate my existing twitter account. The file was manually downloaded to the **data** subdir in of the git repository.

3. Assessing Data:

Subsequently to the gathering process the data was assessed visually and programatically for quality and titiness issues. This was done separately for each of the three datasets.

A. Following Issues were found for the **twitter archive dataframe (df_ta)**:

Quality Issues (minimum 8):

- wrong datatype for
  - *tweet_id* (should be a string)
  - *timestamp* (should be datetime)
  - *retweeted_status_id* (should be string)
  - *retweeted_status_timestamp* (should be datetime)
- timestamp encoding with reduntant +0000
- some dog names in name column consists only of one letter
- rating numerator scale exeeds 10 and should be adjusted
- string columns have strings filled with "None" (instead of NaNs)
- non string columns have strings filled with "NaN" (instead of NaN)

Tidiness Issues (minimum 2):

- text includes many kind of information and should be split up
- unused columns:
  - *in_reply_to_status_id*
  - *in_reply_to_user_id*
- merge dog "stages" columns in one column
- amnt of rows is different for the 3 datasets
- dognames somtimes start with lower and sometimes with upper case letter

B. Following Issues were found for the **image prediction dataframe (df_ip)**:

Quality Issues (minimum 8):

- some information refer not to dogs

Tidiness Issues (minimum 2):

- dognames somtimes starte with lower or upper case letter

C. Following Issues were found for the **tweet json dataframe (df_tj)**:

Quality Issues (minimum 8):

- N/A

Tidiness Issues (minimum 2):

- N/A

4. Cleaning Data:

After all a sufficient amount of quality and titiness issues were identified, they could be cleaned within the the cleaning data process. The process itself was structured into a **Define**, a **Code** and a **Test** part.

Since the tweet_id could be used a common foreign key for all the data-frames each cleaning step could be performed after all the data-frames were merged successfully.

Each cleaning operation was done already with rough idea in mind what will be important for the Exploratory Data Analysis in advance.

It has to be mentioned that the identified quality and titiness issues are not complete. Cleaning each dataset completely would require a much higher effort and was not nesscary for the following Exploratory Data Analysis done in **Act_report.pdf**.