Wrangle Report

The project is about gathering, cleaning, analyzing and visualizing data from a Twitter account: @WeRateDogs, that post tweets containing funny pictures of dogs together with captions describing the pictures.

I have gathered data from three different sources: Twitter API, Twitter Archive and image prediction data from Udacity.

I first started off gathering the data, then I loaded it into a Jupyter notebook. From there, I assessed the data visually and programmatically to find quality and tidiness issues. Moving from that stage, I chose eight quality issues and three tidiness issues to work with. The issues I chose were:

Quality issues, *twitter archive* table

- "timestamp" and "retweeted_status_timestamp data type is string object and not date type
- There are dog names with first letter in lowercase that are not names
- Some rows in "id"-columns are retweets and can be dropped
- The ratings in the tweets sometimes do not correspond to the ratings in "rating_denominator" and "rating_numerator". For example, the rating for row 46 is 13.5/10, but it says 5 in the column "rating_denominator".
- The datatype for the “id”-columns are either int or float. Can be changed to string
- "source"-column contain url-addresses besides the source

Quality issues, *image predictions* table

- "tweet_id" is int and can be changed to string object
- The column-names are not so easy to understand. Would be better to rename them

Quality issues, *tweets* table

- "tweet_id" is int and can be changed to string object

Tidiness issues, *twitter archive* table

- The columns: "doggo", "floofer", "pupper", "puppo" can be merged into one column
- The three datasets can be merged into one master dataset

I made copies of the data and then I wrote code to clean the data and finally I tested the results.

When I prepared the data to be analyzed and visualized, I came up with some interesting questions to answer after the wrangling process. The questions were:

- What tweets have the highest likes and retweets?
- What are the most common dog names?
- What dog breeds are most favorited?
- What dog breeds have the highest level of confidence from the image prediction table?
- What dog breeds were predicted most often?