Data Gathering

Three different data sources were used in the project.

The first one is *WeRateDogs* Twitter archive. Since this file is available in CSV format, it was read directly with the *pandas* library. This file contains information such as rates of the dogs, dog stage, the text of the tweet etc.

The second source was hosted on Udacity's servers. Its URL address was given. It had to be downloaded programmatically using the *Requests* library. This file was in text file format. It was then converted to CSV format and read. In this dataset, the types of dogs were estimated using the neural network.

The last source is data taken from the internet using the Twitter API. To be able to download this data, authorization from Tweeter was required. Since my application was rejected, I could not download the data with the code. Instead, I had to use the ready data provided by Udacity. The last one was a text file in JSON format. The file was read line by line and required data was taken.

Data Assessing

Data sets were not ready to use because they were dirty and messy. All data sets were inspected visually and programmatically, respectively. Quality and tidiness problems were defined.

In the first dataset, several columns had missing values. Datatypes of some columns were erroneous. Values were messy and unnecessarily long in some columns. Some dog names were erroneous.

In the second dataset, datatype errors seemed in some columns. Names of some columns were not descriptive enough.

In the third dataset, again there are datatype problems.

Data Cleaning

Firstly, missing data were dealt with in the first dataset. Non-empty rows of retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp were removed because retweets and actual tweets are duplicated. Most columns with the missing data had very few rows with values and filling these values were not possible. Therefore, these rows were removed. expanded_urls column had missing values at only few rows so these rows were dropped also.

After missing data, tidiness problems were solved. More than one variable in the same column were seperated into different columns. In the first dataframe, four different columns existed for only dog stage. All dog stages were collected under one column. All datasets were combined based on tweet_id column. In this way, only one dataset survived.

Finally, quality issues were handled. Gibberish values in source column were trimmed. The columns related with breeds and breed probabilities were renamed with descriptive ones. Since dog breeds were determined by a neural network, some dog breeds could not be identified here. They were defined “not identified”. Of the dogs whose breed could be identified, the most likely breed was assigned to the dog. Dog breeds do not have a unique format so they were standardised. Erroneous datatypes are corrected with astype method of pandas. Some dog names which are“a”, “an”, “the” etc were removed.