WeRateDogs Data Wrangling Report

Audi Chandra

This wrangling report describe the three main process of the data wrangling of WeRateDogs archive twitter:

1. Gathering the data
2. Assessing the data
3. Cleaning the data

Gathering

Gathering the data involved with the process of obtaining the data which is needed to be assesed. There are three data involved with this project. The first one is the twitter_archive_enchanced.csv, the csv file contained the archive of WeRateDogs from 2015-2017 which I downloaded it manually from Udacity. The second one is the image_predictions.tsv contained tweet image prediction according to a neural network which I get from Udacity's server by using Requests library programmatical download. The third and final one is tweet_json.txt containded retweet and favorite counts data which I downloaded directly from Twitter by using Twitter API (Tweepy).

Assessing

Assessing the data means that I need to check the data visually and programmatically for any quality and tidiness problem. After using .info(), .describe(), .head() and .value_counts(), I manage to summarise some of the problems:

Quality:

1. Wrong datatypes (integer for id, string for timestamp, etc.)
2. Some of the tweets are retweets of the tweet, not the original post
3. Some of the values are the wrong values and needs to be re-checked
4. Extracting the necessary values only (source column needs to have only source text, not need for url)
5. Some of the values format needs to be based on one standard (dog breed)
6. The name values have impossible names

Tidiness:

1. The structures for dog stages just need to have one columns instead of the dummy format of four columns
2. The three dataframes can be combined into one dataframe

3. The dog breed prediction of three columns can be combined into one column for the analysis purpose

Cleaning

Cleaning the data means applying the solution to each of the problems stated in the Assessing steps so that the data can be properly analyzed. The steps that are taken in this process are:

1. Applying the right datatype
2. Combining three dataframe into one master dataframe
3. Merging the 4 dog stages columns into one main column
4. Dropping the non-original post (Retweet)
5. Replacing the wrong rating value
6. Replacing some of the dog name column value
7. Extracting the source column from the url of source
8. Combine the three breed columns into one column and fixing the values into one standard

In the end, after managed to clean the master data, I saved the master data into twitter_archive_master.csv.