Data Wrangling Project – Udacity

Mansour Al Rajhi

Introduction

The project of data wrangling from Udacity consist of gathering data from the Twitter account WeRateDogs, assessing the data, cleaning the data, analyze the content, visualize the insights, and storing the results.

Gathering

The gathering stage involved downloading two csv files: twitter archive and tweet json. Also, downloading image prediction file from the internet using the requests function.

Assessment

In the assessment stage, we checked all the files to assess their shape, data type, and number of entries. A change of column display feature to be a max of 50 columns was necessary at this stage to inspect the data fully. In addition, we use Tweepy function to query Twitter API using the developer access keys and tokens provided by Twitter after approving the developer access account initiated for this project. Using Tweepy, the query failed many times and would cause the kernel to get stuck in Jupyter Notebook which required us to complete the project without benefiting from the function.

Cleaning

The project requires 2 tidiness issues and 8 quality issues. We decided to clean the following issues:

Tidiness:

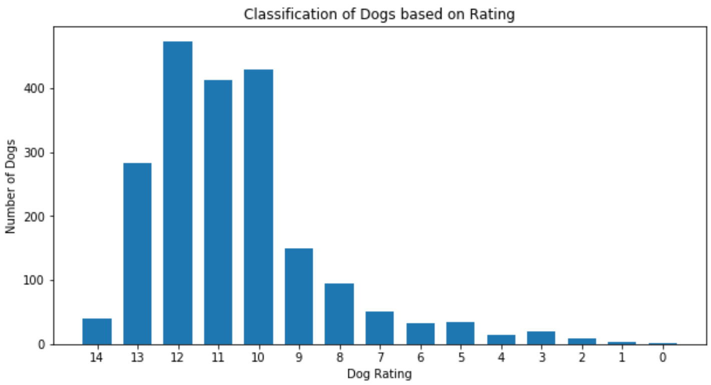
- join all three dataframes by tweet id
- create new column to classify dog type instead of having 4 columns that indicate dog type

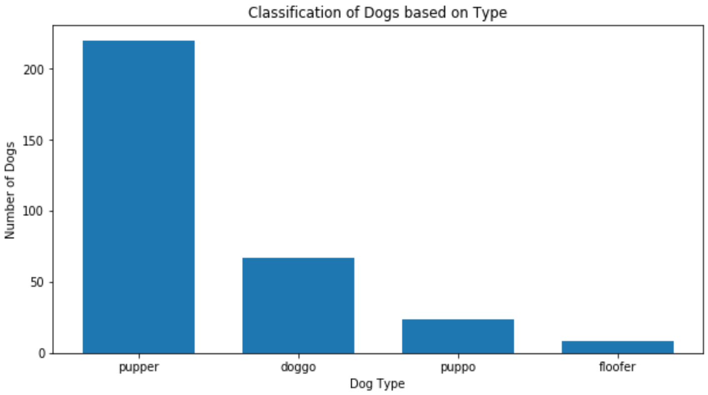
Quality:

- drop all columns that are missing too many values such as (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, etc...) and are not needed for analysis
- drop all rows that are inconsistant such as denominator with values not equal to 10
- drop all rows that are inconsistant such as numerator of high rating which seems to be above 14.
- change the names of some columns to more descriptive names.
- rename the column id to be tweet_id to facilitate merging
- drop the columns (doggo, floofer, pupper and puppo) since we combined them in a new column called dog type
- replace 'None' with np.nan to indicate the missing values

- drop all columns that are duplicate and give the same values such as timestamp and created_at

Analysis





We analyze the number of dogs in the data based on the dog type and dog rating and created two bar chart to visualize the results.

Also, we analyze the statistics of the account tweets favorite and retweet counts and showed the results of mean, standard deviation, minimum and maximum as well as the quartile breakdown.

Insights:

We came up with the following list of insights:

- most popular dog type is pupper
- the account WeRateDogs tend to rate above 10 for most of their rating
- at average, each tweet get around 8500 favorite
- at average, each tweet get around 2900 retweet

Storing:

We finally stored the final results in a csv file named twitter archive master.