Wrangle Report

By Mariam Alabkari

The dataset wrangling project aimed to about using Twitter API to collect dataset from Twitter. The Twitter API will help to collect the dogs images form 'We Rate Dogs' Twitter account. This account provides many dogs images with people rates.

The goal of this project is:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

This step required to use the Twitter API to collect the tweets and create a dataset. Since using Twitter API required to have a permit to be able to access Twitter data. I Used twitter_archive_enhanced-1.csv file which contains all the information that related to 'We Rate Dogs' account. This file contains information for more than 2356 tweets. Including, source, tweet id, timestamp, ratings, name, etc.

The second file named as tweet image predictions, which contains the predictions of the tweets. This is done by using neural network, which a technique that used to make a prediction. It used here to make a prediction for the breed of dog or the other objects in the images.

The third file tweets.json stores all the tweets that I found along with their information. Such as, id, full_text, retweeted, favorited, possibly_sensitive, retweeted_status, quoted_status, etc.

Assessing data

After collected the datasets, I started to assess the data and file quality and tidiness issues.

Quality issues:

- columns name is different in the three datasets.
- Timestamp data type is not correct.
- So many columns contain too many null values.
- There are some values that separated into different columns.
- jpg_url column has some duplicated rows.
- Source column contains non-readable texts.
- Some columns contain non-convenient names
- There are some unnecessary columns.
- Rating numerators column contains some decimal values.
- The dataset contains some non-dogs images.

Tidiness issues:

- The need to combine the three datasets together.
- The need to combine dog stages to one column.

Cleaning data

After I discover the datasets, I started to clean the issues one by one. Also, after each process, I tested the effectiveness of each step.

- Rename column id to be able to combine the datasets
- Convert the timestamp column form object to datetime format
- Drop duplicated rows for jpg_url column
- Remove the columns that contains too many null values
- Combine source x and source y in one column
- Replace the values in source column with readable texts
- Change the names of the columns to convenient names
- Delete unnecessary columns
- Clean decimal values in rating numerators
- Combine the three datasets together
- Combine dog stages to one column