Data Wrangling Project 'We Rate Dogs'

By Sarah El Shatby

Introduction

This project is the full analysis of more than 2000 tweets from the WeRateDogs account on Twitter which is dedicated to dogs from different breads with every Tweet having important information such as the Dog's Name, Life Stage, Image Prediction for each posted image... etc.

The project is divided into 3 Stages:

1- Wrangling

In which we gathered 3 Datasets:

- Twitter_archive_enhanced.csv
- Image_predictions.tsv
- Twitter_api

The first 2 Datasets were downloaded manually and programmatically and the Twitter_api file was downloaded through querying the Twitter API (which wasn't possible in my case due to a problem in authentication).

2- Assessing

In this step we assessed the gathered datasets in detail displaying features and MetaData for each and the following Insights were noticed:

- The Twitter Archive DataSet has a total of 2356 records(tweets).
- Only 181 records have (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) meaning those are retweets which are not required in our analysis.
- There're 3 Data Types in the Twitter Data Set: (Float, Integer, Object "mainly string") that will be modified in the cleaning step.
- Each Tweet has a rating numerator & denominator which indicates how popular this dog is.
- In the Image Predictions DataSet, each image has a confidence ratio which refers to what extent the dog's image was predicted correctly.
- In the Tweet Info file each Tweet has a favorite count & a retweet count (which is merged with the Twitter Archive Dataset to ease the cleaning stage).

3- Cleaning

In this stage we use what we documented in the Assessing step to clean and refine the Data so we can perform our Analysis & Visualization.
The following issues are noted in this stage:

Quality

Master DF & Image_predictions:

1- There're 59 missing expanded_urls which refer to tweets without photos so we have to remove those records.

2- There're 181 retweet records that should be removed.

3- retweet_counts & favorite_counts columns should be in Integer form not float and remove Null values in both columns.

4- timestamp should be in datetime format.

5- Correct wrong rating values & change 'rating_numerator' & 'rating_denominator' to float.

6- Make the 'source' more readable and then convert it to category.

7- Tweet_id in image_predictions and Twitter Archive should be 'str'

8- Change missing values in the 'name' column to NaN and remove incorrect names.

9- Correct Ratings that were extracted incorrectly from the text.

Tideness

1- Create a column (Dog Stage) instead of the 4 classification columns [doggo, floofer, pupper, puppo] and merge a ‘multiple’ dog stage to the table.

2- Join Twitter_archive, tweet_info and image predictions into one DataFrame (Final DF).