Wrangle Report

By: Ahmed Elsayed
December 2020

I am writing this report to summarize the wrangling efforts involved in completing the project of (We Rate Dogs) as a part of Udacity's Nanodegree in Data Analysis.

This project wrangling process consists of four main parts as the following:

1. Data Gathering
2. Data Assessing
3. Data cleaning
4. Data Storing, Analyzing, and Visualizing.

1. Data Gathering

While gathering, I used three different sources of data. Some of them were used manually and others were used programmatically. These three sources are:

Twitter Archive

Directly downloaded CSV file Using pd.read_csv import into pandas data frame.

Image Predictions

Downloaded programmatically from Udacity's server.

The (Image Predictions) file represents the images used in each tweet according to a neural network. This file located on Udacity's server and downloaded programmatically from this URL into the workspace using the requests library.

Tweets Json File

A query from Twitter API

Using the tweet IDs that are in the archive of the WeRateDogs Twitter account, a query from Twitter API is used for each tweet using Python's Tweepy library and stored the entire data in a file named **tweet-json.txt**.

2. Data Assessing

The three saved data frames were first assessed programmatically in Jupyter Notebook with pandas, then visually in Excel/Google Sheets.

The following issues found during the assessment:

Quality Issues

Twitter Archive

1. Dog stage has 'None' instead of np.nan.
2. Only keep the original ratings.
3. Don't need the following coloumns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'img_num', 'expanded_urls' and 'jpg_url'.
4. Timestamp have extra "+0000".
5. The datatype of timestamp should be converted to "datatime".
6. All the denominators of ratings should be "10".
7. Some numerators of ratings are huge values.
8. Since all the denominator is 10 after the last step, we can get rid of rating_denominator column and change rating_numerators to 'rating'.
9. Many dog names are messed up, such as "such" "a" "quite".
10. Source data columns are not clear.

Image Predictions

1. Columns (p1, p2, and p3) contain underscores instead of spaces in their names.
2. Image predictions should be summarized to one column 'dog_breed'.

Tweet Json File

The column id changed into tweet_id for easier merging.

Tidiness Issues

Twitter Archive

1. Columns (doggo, floofer, pupper,and puppo) refer to the same unit, dog stage.

Image Predictions

No tidiness issues found.

Tweet Json File

No Tidiness issues found.

Data Cleaning

All the issues found during the Data Assessing process are cleaned during this process.

Data Storing, Analyzing, and Visualizing

The clean data are then stored in a CSV format file named (twitter_archive_master) using .to_csv. And then this data was analyzed and visualized using python’s seaborn library.