Data Wrangling Project [Data-Analyst-Nanodegree]: WeRateDogs®

By: Lina AlKhodair

Introduction:

This report is intended to document the data wrangling steps performed on the dataset of Twitter account @WeRateDogs tweets. The Twitter account rates people's dogs with a humorous comment about the dog. Throughout the wrangling process the typical data wrangling steps were performed rigorously, starting off with gathering, assessing and finally cleaning.

1. Gathering Data:

In order to analyze the Twitter account, multiple resources have been used.

- ***The WeRateDogs Twitter archive:***

This file was manually downloaded and has been provided by Udacity. The archive has basic tweets data, timestamp, source, and retweeted status etc.

- ***Extracted data from server:***

The second dataset was extracted from Udacity's servers using the 'Requests' library.

- ***Twitter API:***

The final dataset was extracted using Twitter's API and Tweepy library that stores it as a JSON file.

2. Assessing Data:

After gathering each of the above pieces of data, we moved on to assess them visually and programmatically for quality and tidiness issues. A summary of the findings is as follows:

Quality Issues:

1. Drop unneeded columns in the data frames like source, img_num, expanded_urls, source, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, and in_reply_to_user_id.
2. Change data type of timestamp column.
3. Column tweet_id is int, change to type object as no calculation is needed.
4. The name column has invalid data for a name, such as "a".
5. For the column "rating_denominator" there is a zero value that should be removed.
6. Remove retweets.
7. p1,p2,p3 columns dog breeds are inconsistent in lower/upper case.
8. Remove tweets that are not rating (tweets without images).

Tidiness Issues:

1. The columns (doggo, floofer, pupper and puppo) do not need to be separated. Each dog will be classified as one of these classifications, so should be collapsed into one column.
2. All of dataframes should be merged into one, since they hold information about the same context.
3. Rename the column id to be tweet_id to facilitate merging in tweets dataframe.

3. Cleaning Data:

In this step all of the quality and tidiness issues have been addressed and the cleaning process was conducted by following the (define, code and test) method. Before cleaning out the data I have created a copy of each dataframe. The following is some of the cleaning steps that have been performed:

- Dropping unneeded columns such as expanded_urls, source and in_reply_to_status_id using .drop function
- Filtering data of the tweets dataset to only id, retweet_count and favorite_count
- Changing datatype of timestamp to type datetime, and id to type str instead of int.
- Cleaning the name column since it contained some invalid values, such as ‘a’. So, I have created a function that splits the the text into words, then iterating those words and looking for phrases that are commonly used before the dog name.
- Consistent format of lower cases.
- Removing retweets since they are considered as duplicates.

There is more to the cleaning process, however, this is just to give an idea.
After all cleaning is complete, the new merged complete dataset is saved to a .csv file for later analysis and visualizations.