Wrangle Report

By Collins Kimotho

Introduction

The following wrangle report is part of the Udacity's Data Analyst project on wrangling and analyzing data. The dataset utilized for the wrangling efforts was the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. The WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The rates are usually out of 10 and should be a whole number such as 11, 12,13 etc. The following wrangle report documents the three steps utilized for the data wrangling process of the WeRateDogs twitter account: gathering data, assessing data and cleaning the data.

Gathering Data

The data required for the project was gathered from several sources in a number of different formats.

1. The WeRateDogs Twitter archive data. This data was sent to Udacity via email and as part of the class, I was to download this download the data manually from a link provided by the class. After downloading the data, I uploaded the files to the Jupyter Notebook workspace.
2. The tweet image predictions. This data was present in each tweet according to a neural network which is hosted on Udacity's servers. The data was meant to be downloaded programmatically using the Requests library and a link provided by the class.
3. Additional data from the Twitter API. I gathered the tweet's retweet count and favorite(like) count utilizing the Twitter API. I used the tweet IDs in the WeRateDogs

Twitter archive, then query the Twitter API for each tweet's JSON data using Pythons' Tweepy library then store each tweet's entire set of JSON data in a txt file.

Assessing Data

This was the next step after gathering all the three pieces of data. Assessing the data was done visually and programmatically and I was able to identify the following quality and Tidiness issues;

Tidiness Issues

1. ***Twitter Archive Data***
    1. Delete retweets
    2. Null values in the in_reply_status_id, in_reply_user_id and retweeted_status_id.
    3. Source column can be improved to show relevant information.
    4. 109 irrelevant dog names in the name column
    5. The timestamp column is in string format instead of timedate format.
    6. There are 23 tweets whose rating denominator is NOT 10. The ones with zero as a denominator should be corrected or deleted.
    7. There are 28 tweets whose rating numerator is greater than 15. These tweets have been identified as outliers from the analysis done above. The ones with zero as a numerator should be corrected or deleted.

2. ***Image Prediction Data***
    8. The data from the image prediction table reveals that there are 2075 images that have been retrieved. This means that we are 281 columns less than that of the twitter archive table. We will classify these are missing values.
    9. Drop the unnecessary columns that will not be used for our analysis.

3. ***Twitter API Data***

    10. Just like the image_prediction dataset, the tweet_data is also has missing data of 50 rows
    11. The tweet_id datatype should be a string and not an integer. This step should be done after all datasets have been combined.

Tidiness Issues

1. There are 4 columns for dog stages i.e., doggo, floofer, pupper, and puppo. These 4 dog stages should be in one column.
2. The three datasets should be combined into one since they are all related.

Cleaning Data

It is important to note that there were other issues with the dataset that were not mentioned. The reason why other issues were not covered is because I placed more efforts in the data issues that proved to be vital for the analysis. There are three cleaning steps for each data issue mentioned above, define step where I defined what the issue was, the code step where I programmatically solved the issue and the test step where I ensured that the code was successful in achieving the define step results. A lot of reassessment and iteration was done to ensure that the data was clean. After the cleaning process, the clean datasets were stored in a csv file.