Wrangle Report

By: Brandon Chung

April 29, 2018

Introduction

For the WeRateDogs Twitter project we are using different wrangling techniques to proactively gather, assess and clean our data. We are gathering the necessary data files from different sources, assessing the data quality issues and tidiness of the structure and lastly cleaning the data to make sure the dataframe is error free for analyzing.

Gathering

In this section we gathered data three datasets to use from different sources.

1. WeRateDogs provided us with a csv file linked to twitter-archive-enhanced.csv. The archive dataframe included tweet ids, ratings, dog names, sources, retweets, URLs etc
2. The tweet image prediction file (image_predictions.tsv) was downloaded programmatically using the python request library. Each tweet image analyzes the images of the dog through a neural network and correctly identifying each type of breed
3. The last source was querying from the Twitter API where each tweet was stored in a JSON text file (tweet_json.txt) using the Tweepy Library. Because we used the tweet ids from the archive file we were able to query each tweet's JSON data and store it in a dictionary list and add it to the new lines of the dataframe. This extracted data includes the tweet ids, favorites and retweets.

Assessing

After successfully gathering all the data it is time to assess the quality and tidiness of the data where we will look to fix all the issues and documenting each of them so we can keep track of which ones are resolved.

Quality

Outlining some of the quality issues observed we can see there were dog names in the column that were missing names or they were not real dog names like "a", "very", and "quite" which all appeared to be in lowercase letter than uppercase. Looking at the tweet ids that have the same jpg_url images there were 132 duplicates spotted. Rating numerators and denominators had to be converted to a float to account decimal based ratings. Several empty replay and retweet columns including in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp had to be removed because we only want the original ratings. Some expanded_urls images were missing in the dataset so those rows had to be removed as well. On the image_df dataframe we have spotted in columns p1, p2 and p3 underscores which needed to be corrected without it.

Tidiness

As for the structure of dataframe the columns with 'doggo', 'floofer', 'pupper', and 'puppo' were integrated into one variable to showcase the stages of the dog for each tweet id. Same goes for creating a ratings column by dividing the numerator by the denominator to get a final rating score. When merging the three data sources we needed to reference the tweet id in the tweet_df and image_df dataframe to match the archive_df to create a new master. By doing so we had to convert the tweet ids to a 'str' object type from an integer. Lastly we want to merge the different dog breed columns to simplify our analysis for later.

Cleaning

The last step in the data wrangling process is assessing our data which we will need to clean our dataset. Our process is to Define, Code and Test each fix and we will want to create a copy of our dataframe to use for these changes. Walking through this process helps us observe each of the data quality errors that have been cleaned and how the final dataframe looks like before and after the cleansing process. Predominately most of the changes have come from the twitter archive dataframe that contain missing data, mislabeled names, unwanted columns, and conversion of data types. Finally merging the data sources into the final master dataframe allows us to use it for analyzing and visualizations.

Conclusion

The biggest challenges in terms of the data wrangling process was trying to figure out the code to query the Twitter API and store each retweet in a JSON data list as this is technically challenging even for avid python programmers and especially for a person like me who does not use twitter at all makes it much harder.