Wrangling Act Report

By: Sarah Mitchell

Introduction

This project consists of data from Twitter called We Rate Dogs. It is a combination of three different data frames. Each data frame brought additional information to the project. Because all three data frames are uncleaned, this project aims to gather, clean, and assess. Then I can show the insights I find with a few visualizations.

Gathering

The data was gathered from three different data frames provided by Udacity, they consist of the:

1."twitter-archive-enhanced.csv"

2."image_predictions.tsv"

3.'tweet_json.txt'

After importing pandas, seaborn, matplotlib.pyplot and a few others. Then I was able to pull the data frames. After I gathered all three, then decided to merge them into one data frame to begin assessing.

Assessing

Once I had one data frame, the next step would be to see how the data was inputted. Looking for quality and tidiness issues. To do that pulled the ‘head’, ‘tail’, ‘info’, and ‘shape’ to get a better understanding of the data. The issues I found were:

- There are names that are not actual names. Need to replace them with 'None' to show no actual name is there.
- The timestamp' and 'retweeted_timestamp' need to be changed to DateTime.
- Need to change out the '_' to a space between words in columns p1, p2, p3.
- Some of the column names can be changed for better understanding.
- Clean up the tweet column. Removing the 'https:' portion.
- Need to get rid of the 181 retweeted_status_id. Then need to drop the columns.
- Need to delete duplicate urls from the data frame.
- Need to set the rating_numerator to 10.
- Create a breed column and a confident_level column. Using the p1, p1_conf, p2, p2_conf, p3, p3_conf columns and combining them. Then dropping the old columns that are no longer needed.
- There are fewer lower cases in columns p1, p2, and p3, which need to be changed to have a Captial letter for the first letter.
- Columns: "doggo", "floofer", "pupper", "puppo" all have to same me

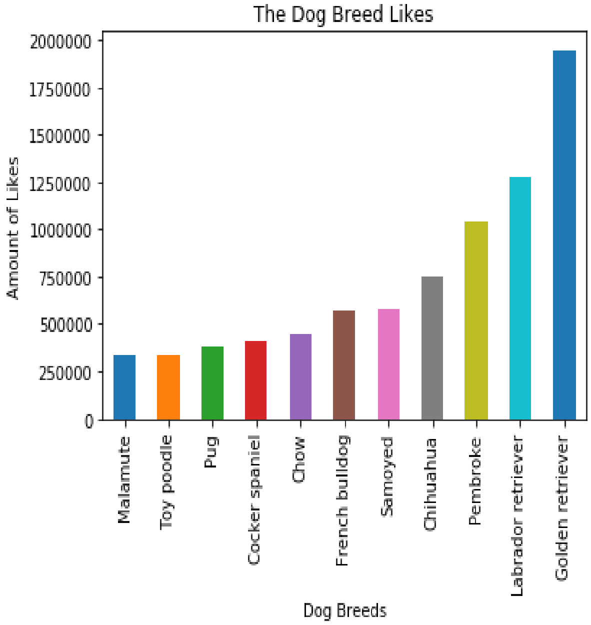
There were many more issues with the data I could see but with this project it was not necessary to fix everything.

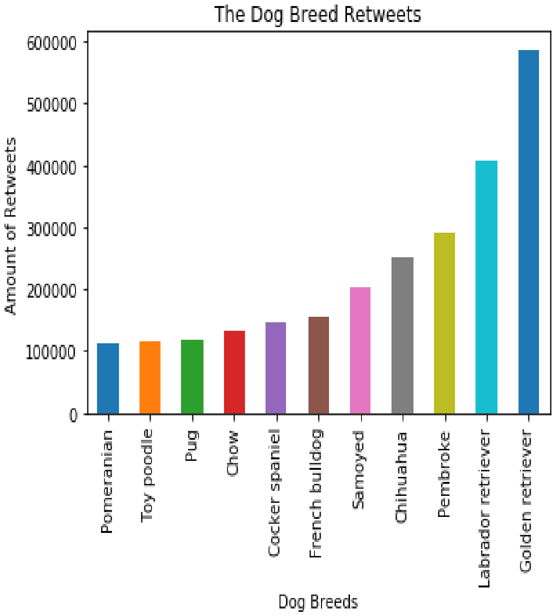
Cleaning

The cleaning process can be overwhelming. It is tedious and time-consuming. The best form for the process is to define, code and test. Define what the problem is and how you want it in the correct format to be. Code is finding the right code and executing it. This can be challenging because there are more ways than one to fix some issues, so you have to know how you want the data to look. Testing is looking at the changed data to make sure the code did what it was supposed to do. This is the practice I used for this project. It helps to stay organized and on task.

Analyze and Visualize

After I had a clean data frame it was time to analyze. There is an abundance of information in this data frame. I chose to analyze the information that I was curious about. I used matplot to visualize the information for the best understanding of it.





In this visual, I have compared the top 10 dog breeds' likes and retweets. The conclusion is easy to see that the Golden Retriever is the most popular dog. On the other side, the least popular dog is the Malamute. The interesting insight I noticed in this bar chart is the 2nd and 3rd popular dog breeds. The Pembroke and Labrador Retriever are equal in likes and retweets. As before the likes and retweets prove the more an item is liked the more likely it will be retweeted, as both charts show the same order of popularity in dog breeds. The other dogs do change also.

Conclusion

This data frame has so much more information one could find insights into. The steps of gathering, assessing, and cleaning is a process. Although it is a necessary process in order to get accurate information. Staying organized and doing the cleaning steps of defining, code, test is the key for any good analyst.