WRANGLING ACT REPORT

Introduction

The project aims at wrangling 3 sources of data. Data cleaning process is done meticulously with aim of attaining a tidy master data frame for analysis.

Gathering data

The project wrangles 3 sources of data; Twitter archive enhanced.csv that has been provided by Udacity. It was sourced from we rate dogs, a page on twitter. Image predictions.tsv that is accessed through web scraping from the udacity servers and the third data frame, tweet df, is directly twitter accessed data. Authorization for elevated access of twitter data was processed before querying twitter favorite count, retweet count and full text based on the tweet id provided in the twitter archive data frame.

WRANGLING ACT REPORT

Assessing data

The 3 data frames were assessed during the cleaning process. However, twitter archive had more issues arising as compared to the other data frames. This were the issues that were flagged during the assessment

1)**Quality**

a) Twitter archive table

- Anomalous numerator ratings that are above 30
- Anomalous denominator ratings that are not equal to 10
- Finding tweet id with original rating (no retweet)
- Nulls represented as None in columns puppo, floofer, pupper and doggo
- Nulls represented as 'none' in the column name. **Given the available data, nothing much can be done about this**
- Erroneous datatypes for columns puppo, floofer, pupper and doggo
- Dogs named as 'a'
- Retaining only relevant columns for analysis
- Erroneous datatypes for timestamp column

b) Image predictions table

- Some predictions are not for a dog
- Inconsistencies of p1_dog, p2_dog and p3_dog values

c)tweet df

- Rename 'id' column to tweet id for uniformity across all the data frames

2)Tidiness

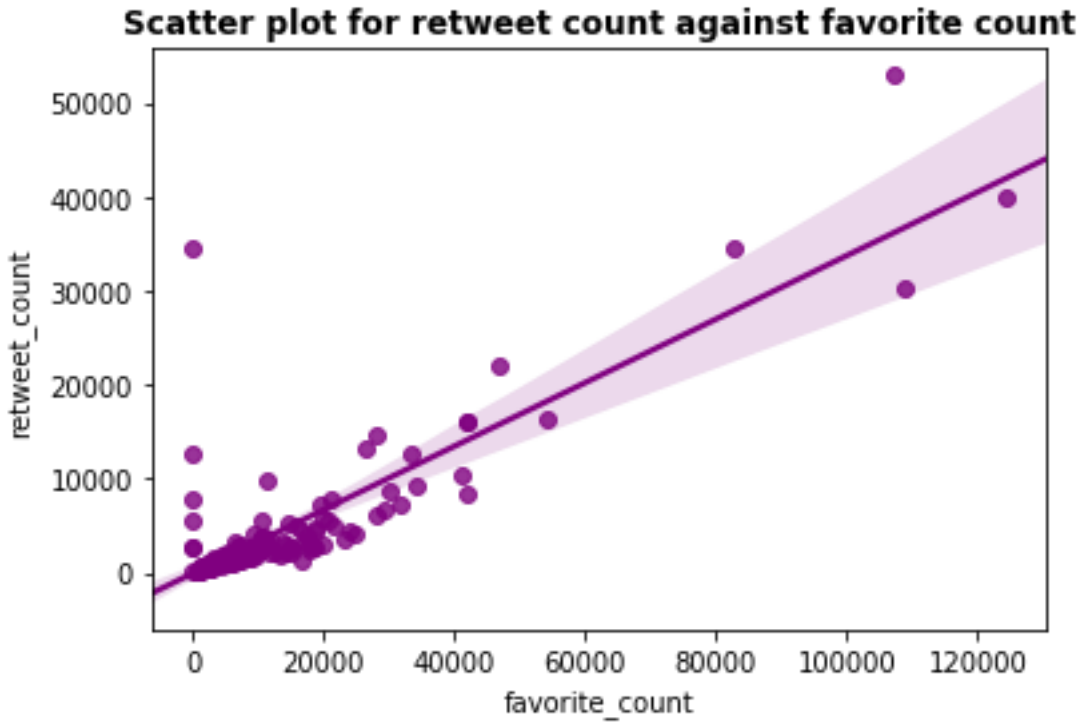
- Eliminate duplicated tweet id column in the 3 data frames by having a single data frame
- Columns puppo, doggo, floofer and pupper should be one column called dog stage

The final tidy master data frame was twitter archive master. This was the piece that was used to transform the data to valuable insights through analysis and visualization.

WRANGLING ACT REPORT

Insights.

- pupper has the highest favorite count sum of 889,118 followed by pupper while floofer has the lowest favorite sum. The margin between doggo and pupper for this variable is relatively large with a figure of over 100,000.¶
- pupper has the highest retweet count sum of 289,546 and floofer has the lowest retweet count sum. The margin for doggo and pupper for this variable is less than 2% which is relatively small¶
- There's no association between numerator rating and favorite count. Most ratings populated around 11 and 14
- Favourite count sum and retweet count sum have a positive correlation as shown in the diagram below.



WRANGLING ACT REPORT

Conclusion:

Pupper which has the highest numerator rating is the most favorite dog on twitter while floofer is the least preferred.

Weakness of the data

It’s important to note that a large majority of the dogs missed information about the dog stages. That was a major setback for our analysis as this reduced our sample size tremendously.