WE RATE DOGS

Data Gathering

The data for this project was collected from three different sources, i.e. twitter_archive_enhanced.csv, image-predictions.tsv and twitter API. Finally all were merged into a single file called twitter_archive_master.vsv.

Tweepy library was used to download tweet information from Twitter.

Data Assessing

I assessed the data both visually and programmatically and found out some issues as follows:

Data quality

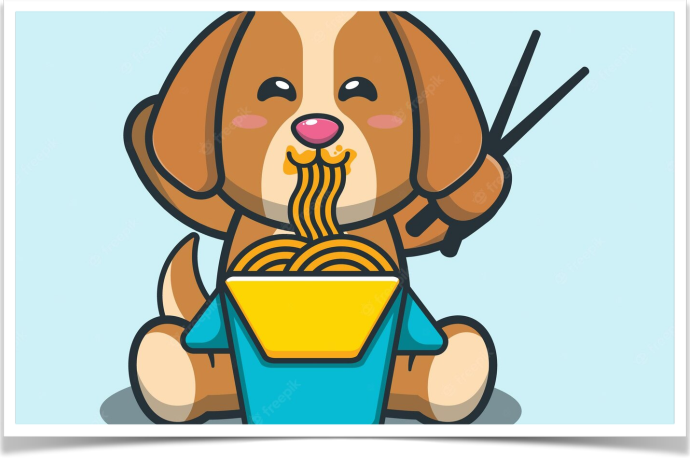
Invalid values: Too high values or non-standard values (e. g. dog's names such a, an, very, etc.);
Wrong data type: Date Time variables as a string.

Data tidiness

Converting several columns into one column, and;
Merging tables.

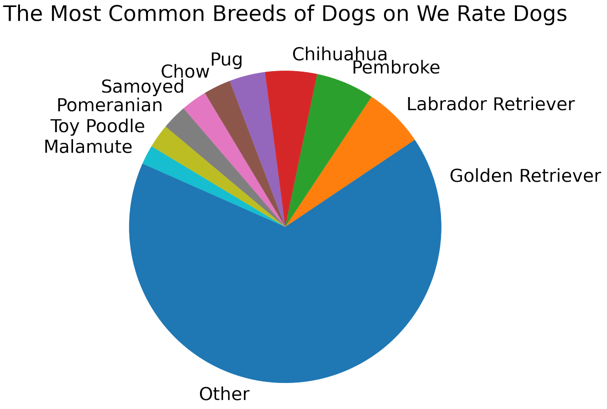
Data Cleaning

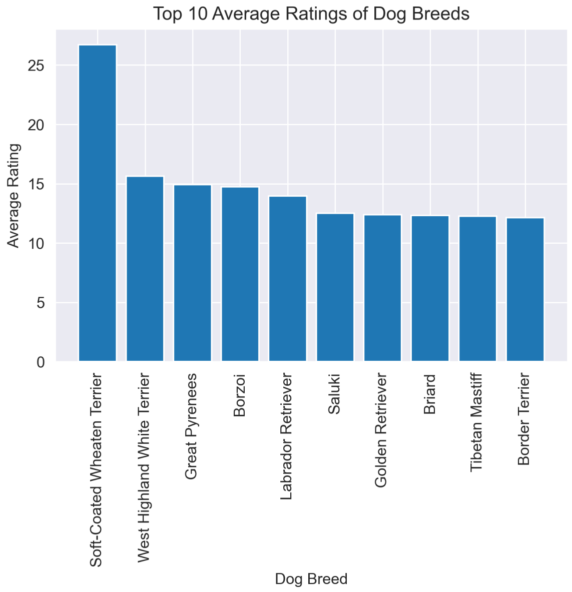
- fixed problems in rating_numerator and rating _denominator values resulted by a non well calibrated regular expression to extract the rating from the text column.



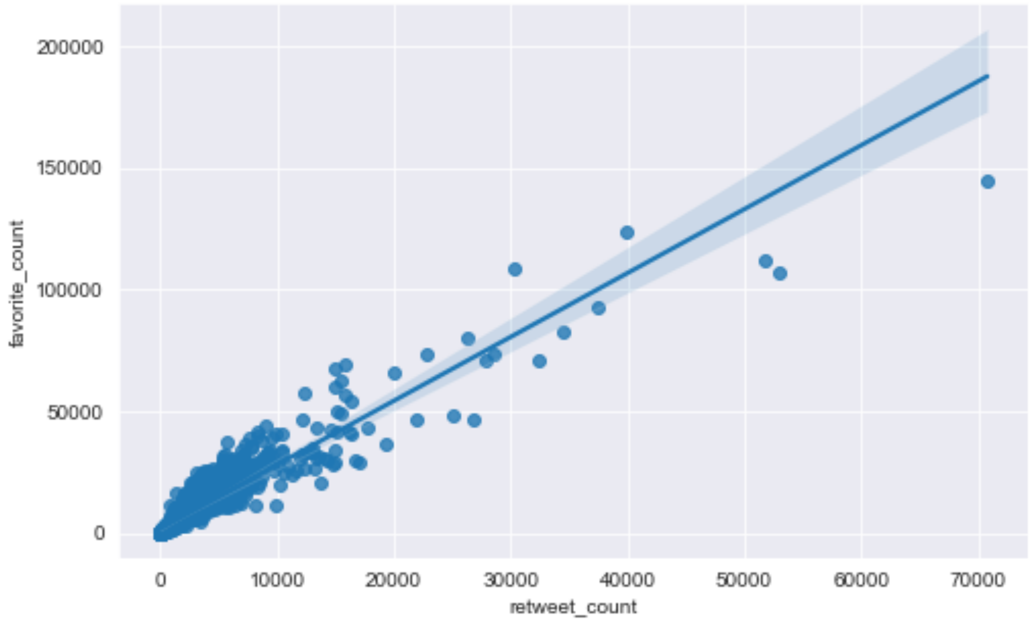
Data Assessing

Following charts were generated based on the data gathered in Data Gathering stage.





A Correlation of Retweets and Likes (Favorites)



Conclusion

- This project aims to perform the Data Wrangling and the Exploratory Data Analysis in the WeRateDogs Twitter account.

- The Data Gathering process englobed three different tasks, the first one download file from URL and later loading to the my working environment, i.e. Dataspell, which requires a manual step, the second downloading a file programmatically, and the third gathering data from the Twitter API.

- Based on the data gathered, I have assessed the most evident issues (11 issues in total) and documented it to create a record of modifications. Later, in Data Cleaning process I have fixed all identified issues to complete, and I have also merged separated data frame into one and added some missing values. The final data frame was stored as twitter_archive_master.csv.

- In the Data Analysis and Visualization, which I have interpreted as Exploratory Analysis, I have posed few questions to guide my analysis. I have found strong evidence of:
  1. Soft-coated Wheaten Terrier has the highest average rating.
  2. A positive correlation between the number of retweets and the number of favourites
  3. The most common breed of dogs is Labrador Retriever.