Wrangle and Analyze Data

Sourojyoti Paul

Act Report

Synopsis

This project aims to perform the *Data Wrangling and the Exploratory Data Analysis* in the WeRateDogs™ Twitter account.
It focused on how to gather, assess, clean, and analyze the data, in other words it englobes the Data Wrangling and Exploratory Data Analysis. The database used as an example is about the WeRateDogs™ (https://twitter.com/dog_rates) Twitter user, this account has more than 7,572,000 followers, 9,500 tweets, and 141,000 likes.

The Data Gathering process bundled three different tasks, the first one downloading from URL and later loading to the Jupyter Notebook, which requires a manual step, the second downloading a le programmatically, and the third gathering data from the Twitter API. This step is also required to save this data in a local machine.
Based on the data gathered, assessed the most evident issues and documented it to create a record of modifications. Later, in the Data *Cleaning process* we have fixed all identified issues, and also merged (the two downloaded les from the Data Gathering process) into one and added some missing values (from the archive downloaded from the Twitter API). The nal data frame was stored as **twitter_archive_master.csv**.

In the Data Analysis and Visualization, we have posed few questions to guide my analysis, which lead to some strong evidence of:

- Seasonality in the number of tweets along the week and along the year
- A positive correlation between the number of retweets and the number of favorites
- No correlation between the algorithm output used to predict the dog breed.

1. Introduction

This report is a resume of the Wrangle and Analyze Data Project from the Udacity Data Science Nanodegree, and aims to shows the insights observed in the wrangle_act.ipynb.One of the funniest parts of the WeRateDogs tweets is about the humor content in almost all tweets and the very non-standard rating systems, which generally exceed the limits, in other words, from 0 to 10 the dogs usually receive a rating above 10.

2. Data Wrangling

For a better understanding, this chapter has been partitioned into three:

- Data Gathering
- Data Assessing
- Data Cleaning.

2.1. Data Gathering

The bedrock of this project is a combination of the twitter_archive_enhanced.csv and image_predictions.tsv les, which are here named as twitter_archive_master.csv. Later, added to the twitter_archive_master.csv two new features gathered from Twitter using the tweepy json file, one feature is the number of retweets and the other is the number of favorites.
The twitter_archive_master.csv has information about [WeRateDogs™][dog_rate], including the predicted dog's breeds by three dierents algorithms. Let's investigate this data frame to identify any kind of pattern or relationship between the recorded data.

2.2. Data Assessing

Following the good practices, we have documented each issue found before fixing it. Below are the found issues:

Quality issues

twitter_archive_enhanced.csv

1. Invalid names or non-standard names.
2. Invalid ratings numerator. Value varies from 1776 to 0. Data Structure must be converted from int to float.
3. Invalid denominator, I expected a fixed base. Data Structure must be converted from int to float.
4. timestamp is a String, It needs to be converted to date.
5. tweet_id must be a string.
6. retweeted_status_id : The same dog could be recorded twice or more in cases of retweets.
7. in_reply_to_status_id : The same dog could be recorded twice or more in cases of reply.
8. source column is having HTML tags, URL, and content in a single column.

image_predictions.tsv

9. Columns p1,p2 & p3 : Dog's breed has no standard. Capital letter or lowercase names.
10. tweet_id needs to be converted as string
11. Column jpg_url has duplicated images and consequently double entry.

Tidiness issues

twitter_archive_enhanced.csv

1. doggo, floofer, pupper, and puppo. These are categorical variable, can be combined into one column.
2. text : There is two information in a single column. need to split the text from the URL.

2.3. Data Cleaning

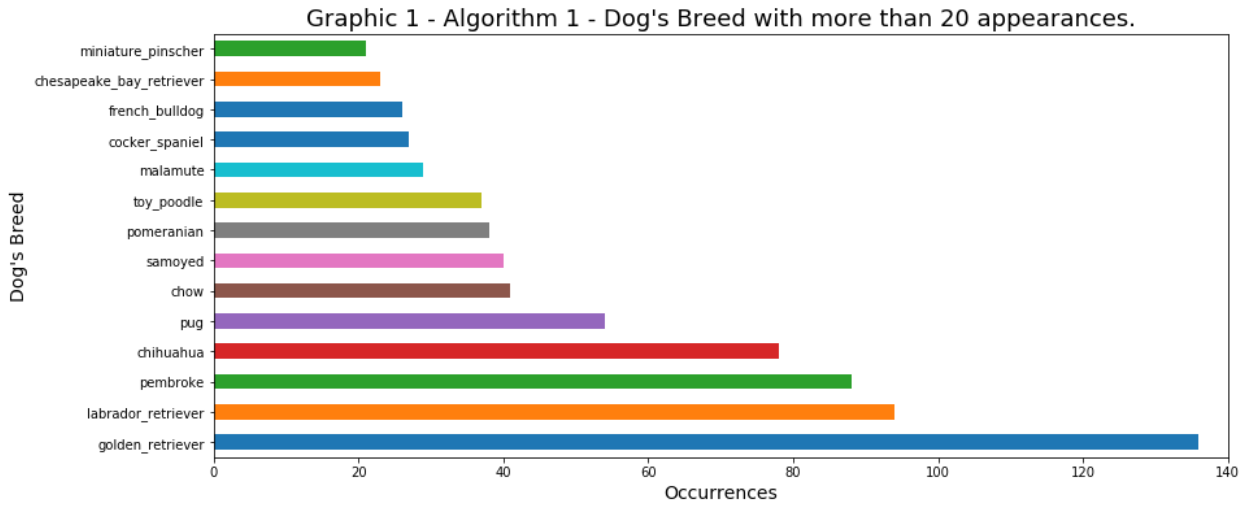
Along the process of Cleaning, there have been fixed problems in rating_numerator and rating_denominator values resulting from a non well calibrated regular expression to extract the rating from the text column. Similarly, the dog's name was fixed due to a problem gathering ordinary words instead of the dog's name.

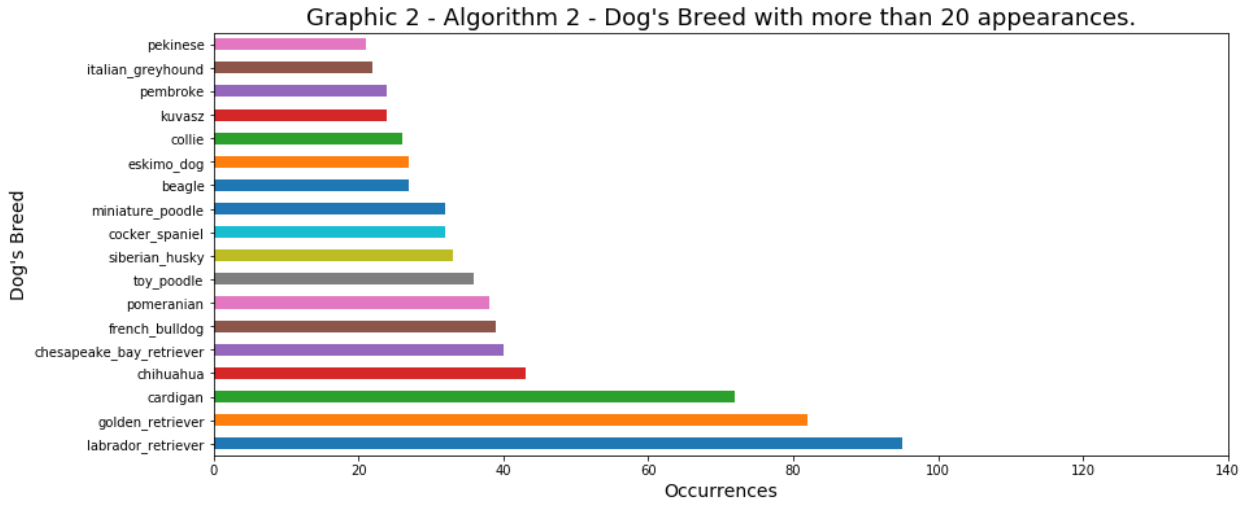
3. Exploratory Analysis

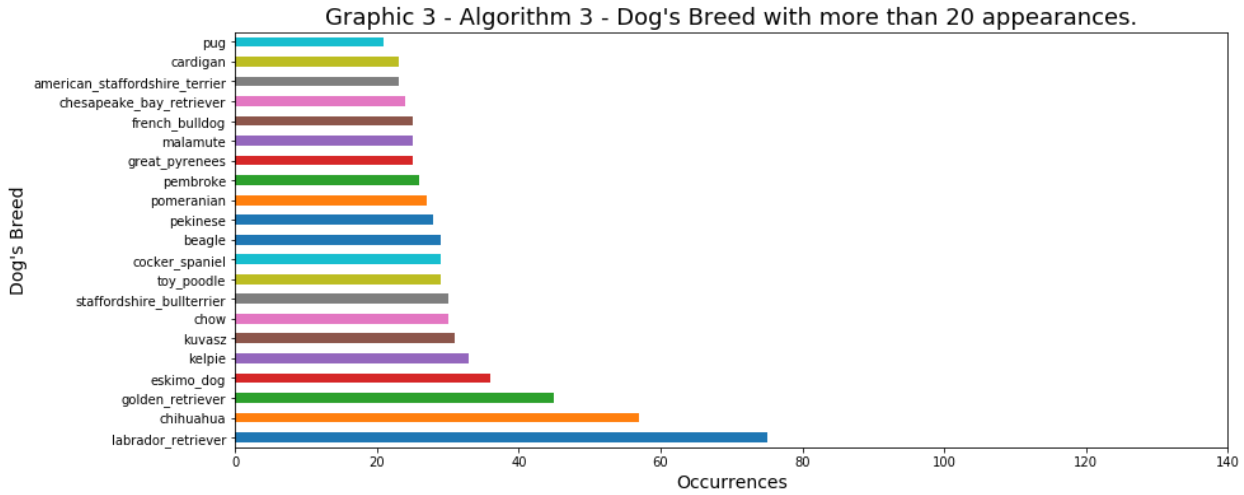
Based on a data frame of several tweets from WeRateDogs™ (provided by twitter_archive_master.csv le), investigated the output of each algorithm employed to predict the dog's breed.

3.1. Predict Dog Breed Algorithm

The Graphics 1, 2, and 3 show the rst 20 breeds with more appearance.





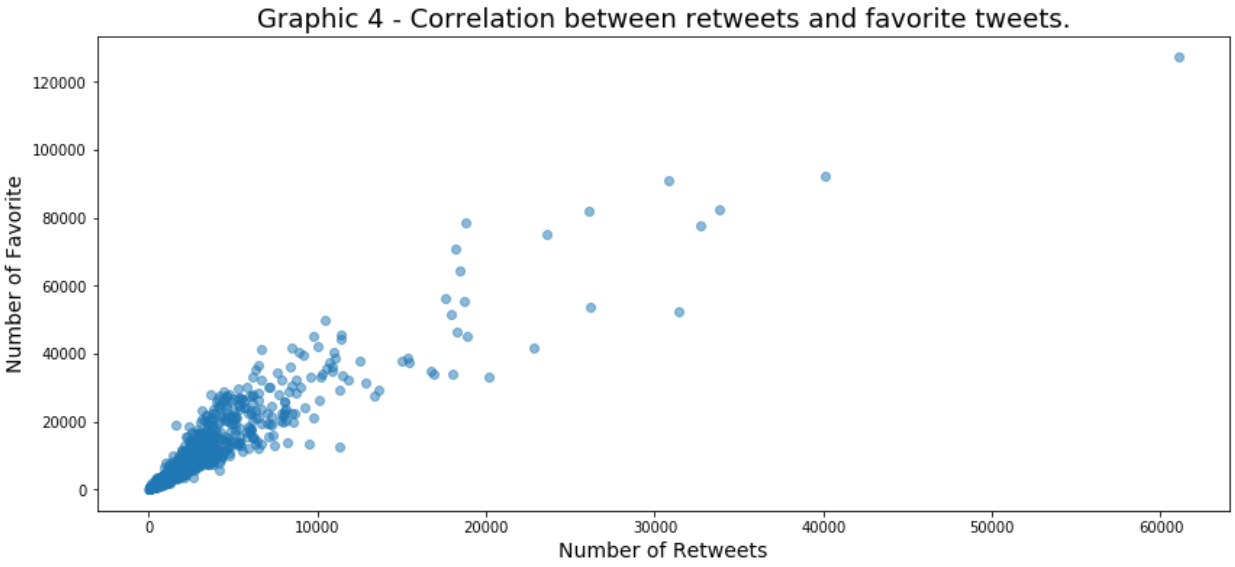


As you can see, the P1 algorithm has the lowest quantity of breed with more than 20 appearance, only 14 breeds, and has the breed with more appearance between the three algorithms. On the other hand, the algorithm P3 has 21 breeds with more than 20 appearances, which also results the lowest between the top scored breeds.

Conclusion: *The algorithm 1 tends to concentrate the breed classification in a few breeds, whereas algorithm P3 does the opposite, spreading the classification in more breeds. The algorithm P2 is a midterm between P1 and P3 algorithms.*

3.2. Retweets vs Favorite

The graphics 4 and 5, presents a straightforward scatterplot graphic of favorite_count vs retweet_count to visualize any pattern.

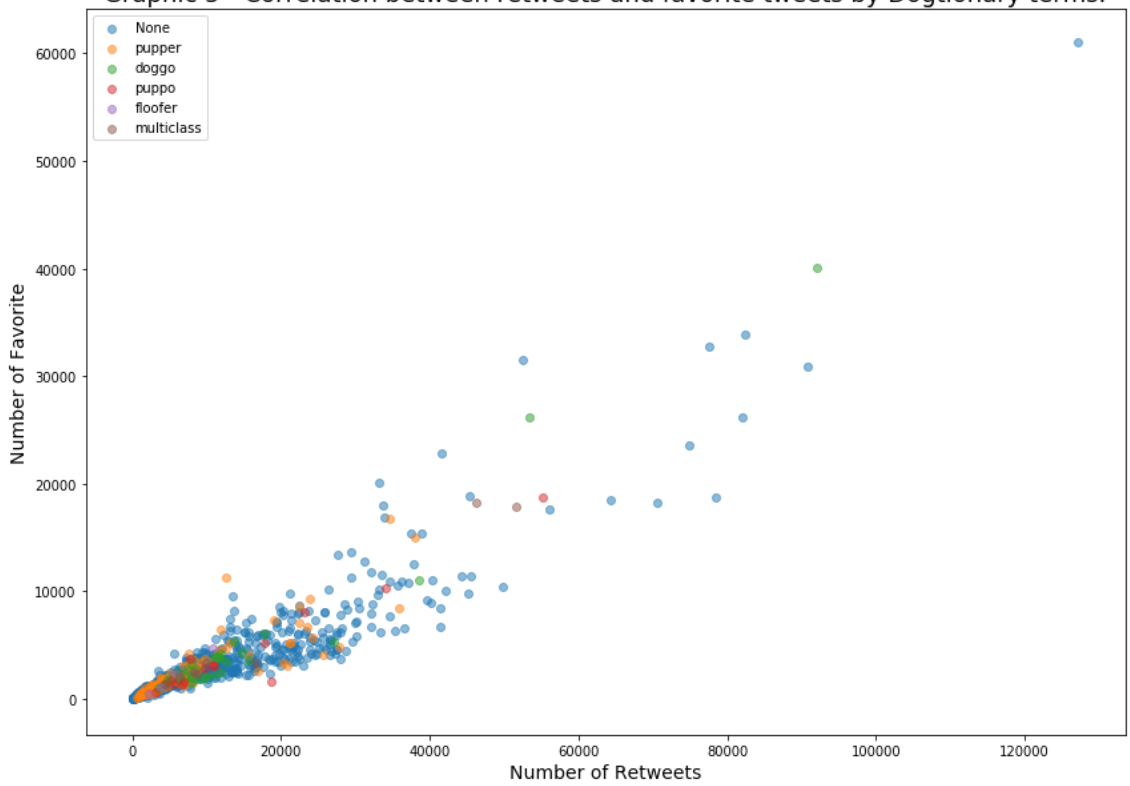


Conclusion: *There is a strong and positive correlation between both variables.*

Puppo, Pupper, Doggo, Multiclass, or Floofer

We wanted to visualize if there is any pattern using the Dictionary as a classier.

Graphic 5 - Correlation between retweets and favorite tweets by Dogtionary terms.



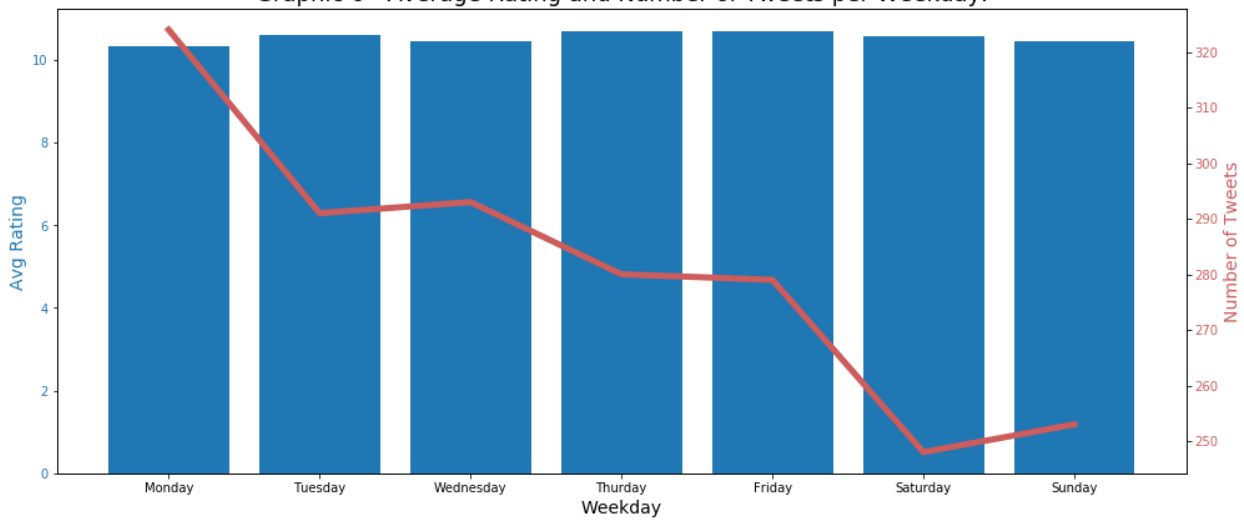
Conclusion: *It is not possible to identify any pattern using the dictionary classification.*

3.3. Tweet's Seasonality

To determine if there is any seasonality in the tweets behavior.

The Graphic 6 shows the Tweets' number during the week.

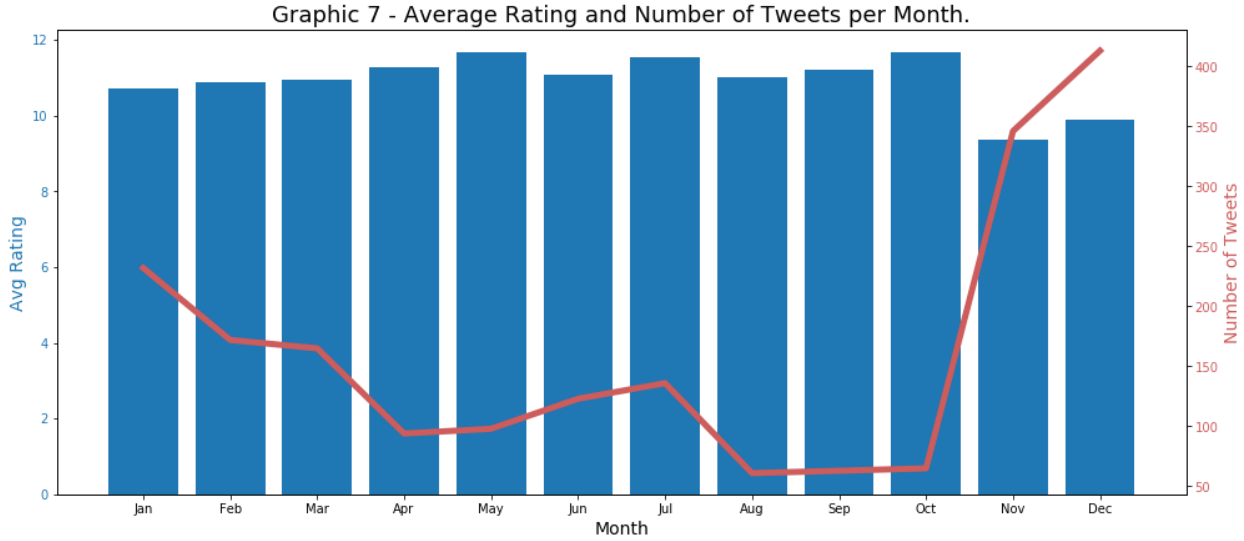
Graphic 6 - Average Rating and Number of Tweets per Weekday.



Conclusion: *There is a seasonality along the week because the WeRateDogs tend to tweet in average more on Mondays, Tuesday, and Wednesday. Although the number of tweets is dierent during the week, the rating stays steadily*

The Graphic 7 shows the Tweets' number along the year.

Graphic 7 - Average Rating and Number of Tweets per Month.



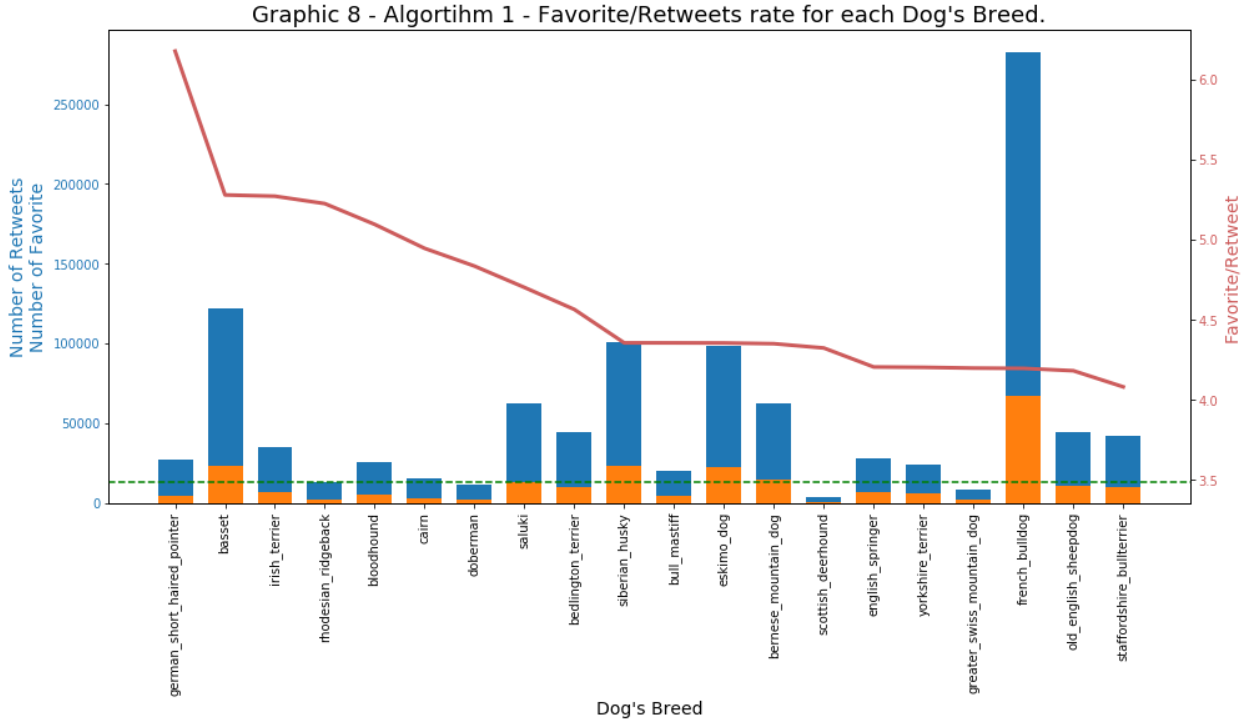
Conclusion: *The tweets also have seasonality during the year, there are much more tweets in December and November. A strange characteristic is the average rating in these two months, which is the lowest value over the year.*

3.4. Dog's Breed Appeal

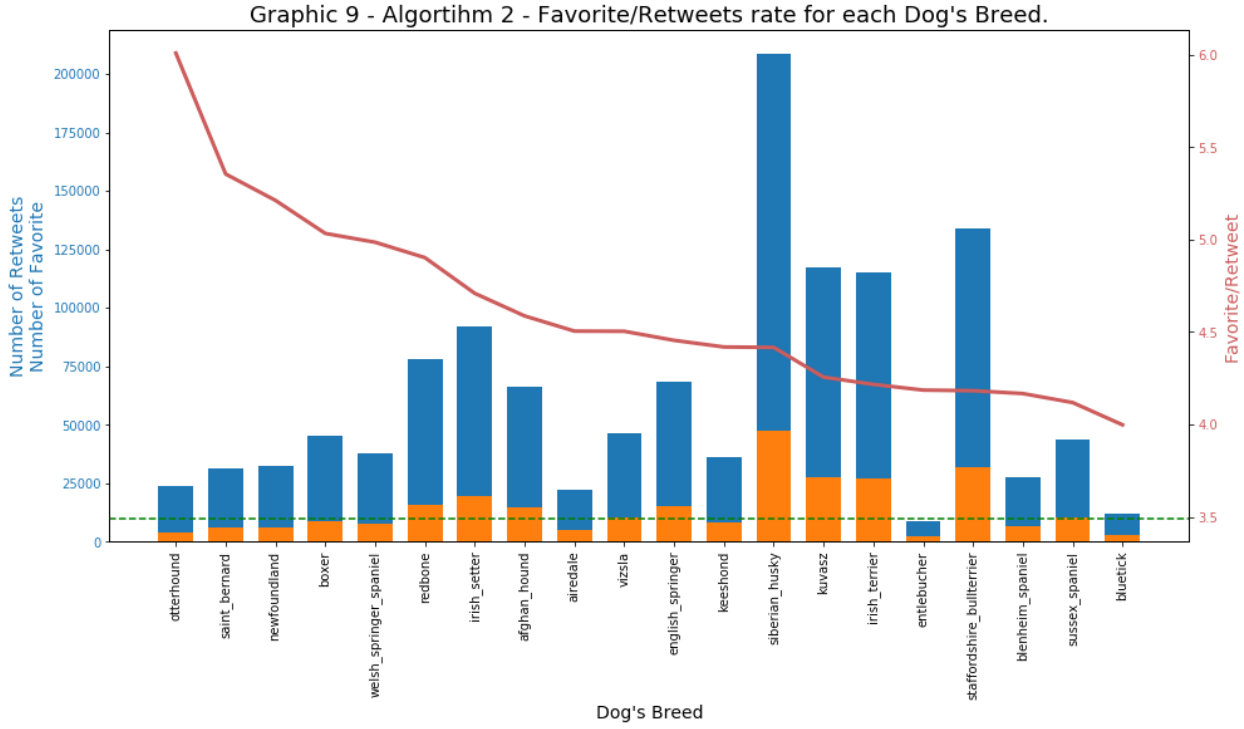
To determine the dog breed with more impact.

The graphics 8, 9, and 10 show the behavior of the relation between favorite and retweet. The red line represents the rate of favorite by each retweet, green line average rate, the blue bars (number of Favorites), and orange bars (number of retweets).

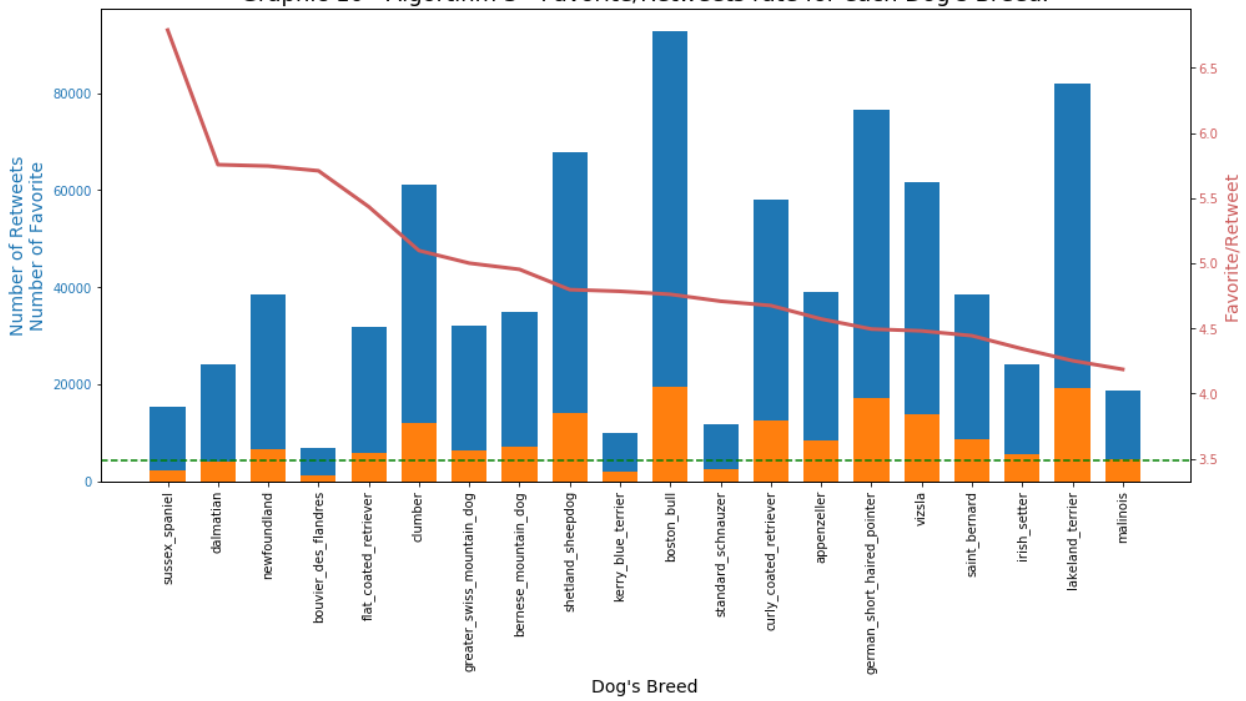
Graphic 8 - Algortihm 1 - Favorite/Retweets rate for each Dog's Breed.



Graphic 9 - Algortihm 2 - Favorite/Retweets rate for each Dog's Breed.



Graphic 10 - Algortihm 3 - Favorite/Retweets rate for each Dog's Breed.



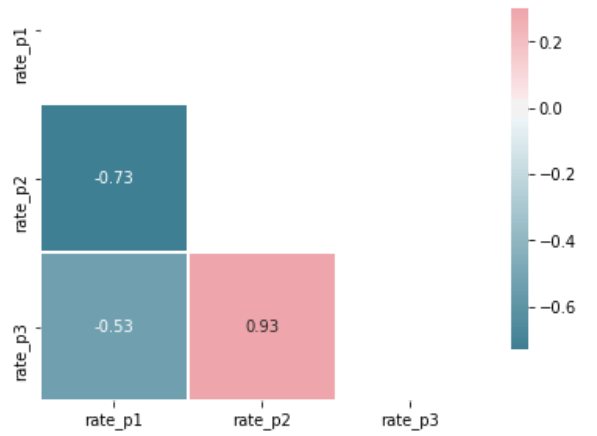
Conclusion: *Unfortunately, it is impossible to make any conclusion based on these three graphics, but these results open an opportunity to pose a new question. How different are the results of these three algorithms?*

3.5. Algorithm Correlation

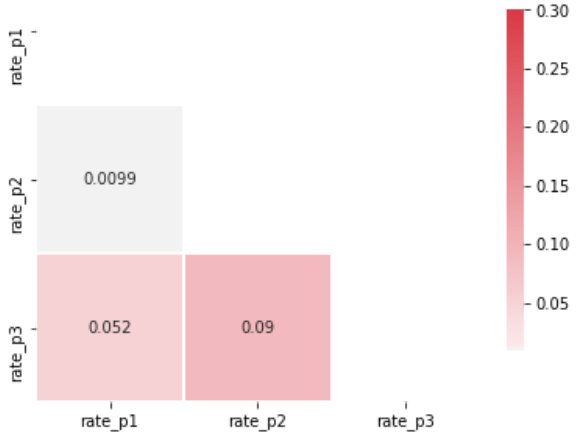
To determine the performance of the algorithms

Found on the 3.4. item, where it is not possible to identify what is the breed with better performance with respect to the rate (the relationship between favorite and retweet). There is no dog's breed which appears in the three algorithms in the top 20. For this reason, we increased the threshold to 40 (top 40) and later I have used all breeds. The graphic 11 shows a comparison between the rate of the same breed for all algorithms.

Graphic 11a - Correlation Map - Threshold 40



Graphic 11b - Correlation Map - All Breeds



Conclusion: *There is no correlation between the results of the three algorithms. An analysis oblique using only the 40 breeds with the highest rate could lead us to an erroneous conclusion. The Correlation Map using all breeds gave a good measure of (un)similarities between these algorithms.*

4. Conclusions

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

The Data Gathering process engulfed three different tasks, the first one downloading file from URL and later loading to the Jupyter Notebook, 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, We have assessed the most evident issues (17 issues in total) and documented it to create a record of modifications. Later, in the Data Cleaning process I have fixed all identified issues to complete, and I have also merged separated dataframe 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 we have interpreted as Exploratory Analysis, We have posed a few questions to guide my analysis. Also have found strong evidence of:

- Seasonality in the number of tweets along the week and along the year
- A positive correlation between the number of retweets and the number of favorites
- No correlation between the algorithms output used to predict the dog breed

---

Thank You