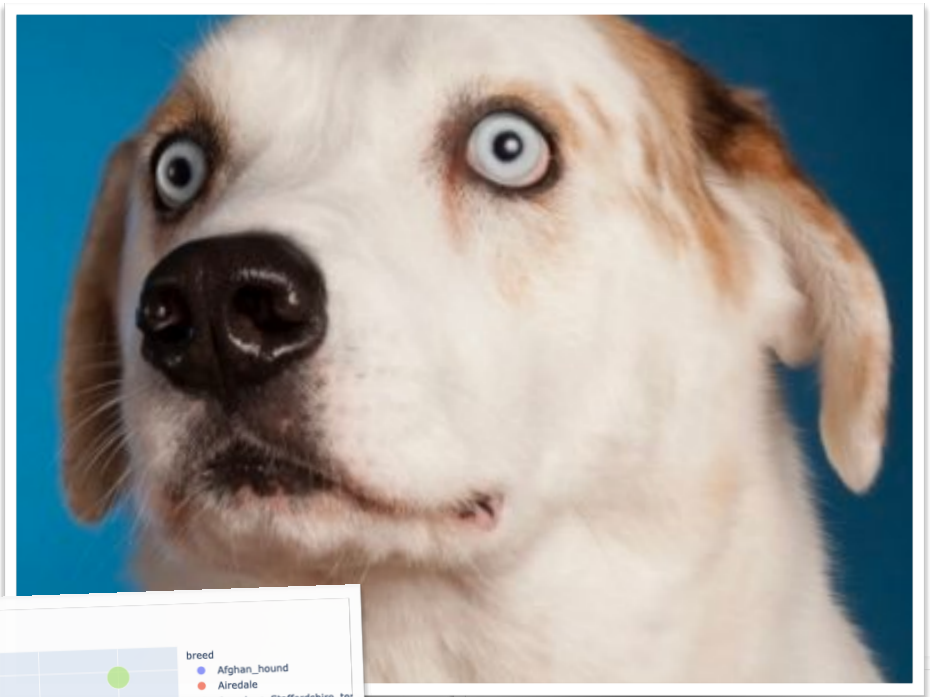
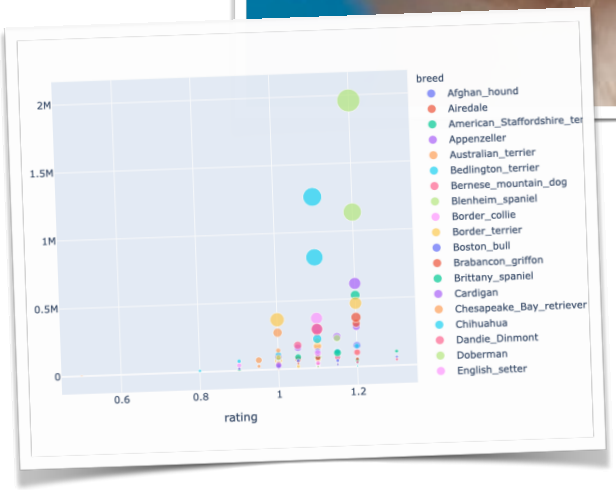
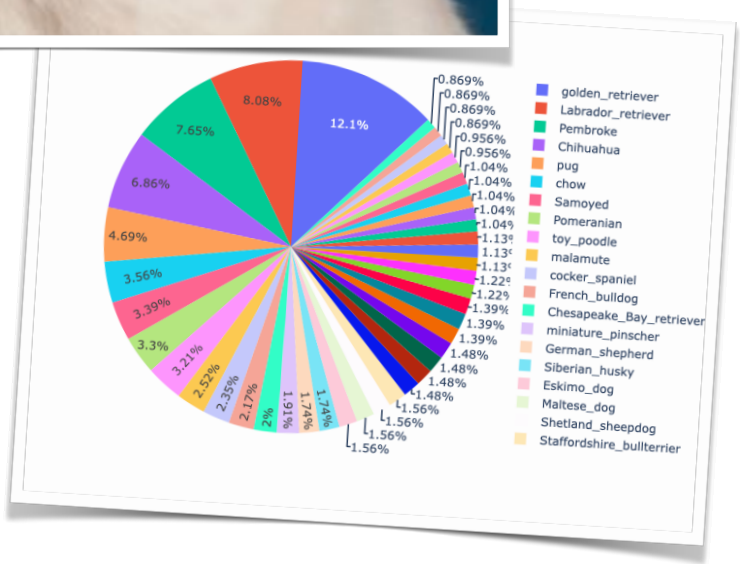
Eric Melillanca Torres
10 March 2021

Analising @WeRateDogs







Wrangle and Analise Data

Eric Melillanca Torres
10 March 2021

Analising @WeRateDogs

Wrangle and Analise Data

Introduction

This paper presents the work of data wrangling and exploratory data analysis of different datasets from WeRateDogs' @dog_rates Twitter account. This account rates images of dogs as well as adding funny pavement comments about each dog.

The work is developed from a primary file, sourced from Twitter and provided by Udacity; however this file does not contain all the information from the tweets. The second file corresponds to the product delivered by a learning machine based on an image recognition neural network, with the breed of each of the dogs. Finally, a third file, obtained directly from the Twitter API, is downloaded from the ID of each tweet in the primary file.

Objectives

This analysis, seek to answer the following questions.

1. What are the most common dog breeds on WeRateDogs?
2. What are the most common names?
3. Which breeds achieve the highest ratings?

4. Which breeds get the most reactions (Retweets and Likes)?

5. Is there a relationship between number of reactions and rating?

Data Wrangling

Gather

In the gather stage, all the data needed to work are obtained. This involves interfacing with different sources; for this study we read data from TSC and CSV text files, as well as reading directly from the Twitter API.

To fulfil this part of the Data Wrangling process, the files provided by Udacity were read through the read_csv method of the Pandas library; checking the necessary option in the case of TSV files.

It was also necessary to consult the tweets directly from the Twitter API, in order to complete missing information in the sources already available.

The file obtained from the Twitter API is a JSON file, which contains a large amount of information about each tweet. It was therefore necessary to select the precise columns to meet the objectives, which also reduced the complexity of the dataset and its processing.

Finally, these sources were merged into a single dataset, which was saved as a text file in CSV format, in order to avoid repeating this process and to start from the saved file in the following steps.

Assess

Once the data has been collected in a single dataframe, the process continues with its evaluation.

In the assess stage, the quality of the data and its tidiness is evaluated.

- From the quality point of view:
- Missing data
- Invalid data, e.g.
  - Weight expressed in negative numbers
  - Use of length measures where a mass measure should be
  - Use of string data types where a numeric one should be, among others
- Inaccurate data, e.g.
  - When the number reported is not correct
- Inconsistent data, e.g.
  - When pounds and kilograms are used in the same column

In terms of tidiness, the dataset is expected to be neat and tidy, complying with:

- Each variable forms a column

- Each observation forms a row
- Each type of observational unit forms a table

The dataset was assessed visually and programmatically; both its numerical and categorical variables. The following anomalies were detected:

- There are three columns with non-descriptve names.
- There are tweets, which correspond to retweets (values of 'retweeted_status_id' different to 'NaN'), although this is not an inaccuracy or a wrong data in itself, given the requirement for this project not to consider retweets, it is also something that needs to be improved.
- There are few missing data. Strictly speaking, there are NaN values and others not reported, but given the context it is accepted that this is the case. Some examples are found in the columns 'in_reply_to...' and 'retweeted_status_...'; in these cases the existence of NaN values indicates that it does not correspond to a reply to another tweet or that it does not correspond to a retweet, respectively.
- There are unreported values for 'expanded_urls'
- For the sizes ('doggo','floofer','pupper','puppo'), some rows from the file with the dog breed predictions are not informed, which is due to the fact that not all tweets obtained from the API have a correspondence in the file with the breed predictions.

- Dog sizes are in separate columns, not complying with the tidiness principle of "Each variable forms a column", it should be a single variable, these columns are untidy.

- Some strange names for dog breeds are observed, such as 'orange', 'paper_towel', 'basset', among others. However, the breed predictions are marked with the result of the test to the prediction, through a boolean value.

- The column 'expanded_urls' in some cases consists of several URLs, so the condition "Each variable forms a column" is not fulfilled.

- There are evident outliers for the numerator and denominator of the rating.

- The 'source' format is not very readable, as it comes inside html tags.

Clean

This stage of the data wrangling process involves improving the quality of the data, based on the observations detected in the assess stage.

In this process:

- Non-descriptive names of some columns were corrected.

- Prior to the cleaning process, it was detected that the column 'retweeted_status_id' contained null values when this row did not correspond to a retweet. This facilitated the detection of retweets, which for this study should not be considered. Thus, the non-null values for this column were removed.

- Failed predictions of the machine learning in terms of image recognition were corrected. Some of the images of puppies included not only puppies, but also objects, which surely caused confusion for the neural network. For this reason, in the dog breed column, objects such as oranges, towels, and others were reported. Fortunately, a column with the result of the recognition test was included and the erroneous values of the breed column are shown as False in the testing column.

- The readability of the source column was improved.

- The ouliers for rating values were corrected, values beyond two standard deviations above or below the median were considered outlier. They were replaced by the median.

- The dataset was sorted. Firstly, a single column was left for reporting the size of the dog; previously each possible size was reported in a separate column. Finally, the resulting column was removed because it contained too many null values.

- Continuing with the order of the dataset, the 'expanded_urls' field was separated. In many rows it contained several URLs in the same record. It was separated into different columns.

- Repeated columns with the ID of each tweet, product of successive joins, were eliminated.

- Finally, this cleaned dataset was saved in a CSV file with the name 'twitter_archive_master.csv'.

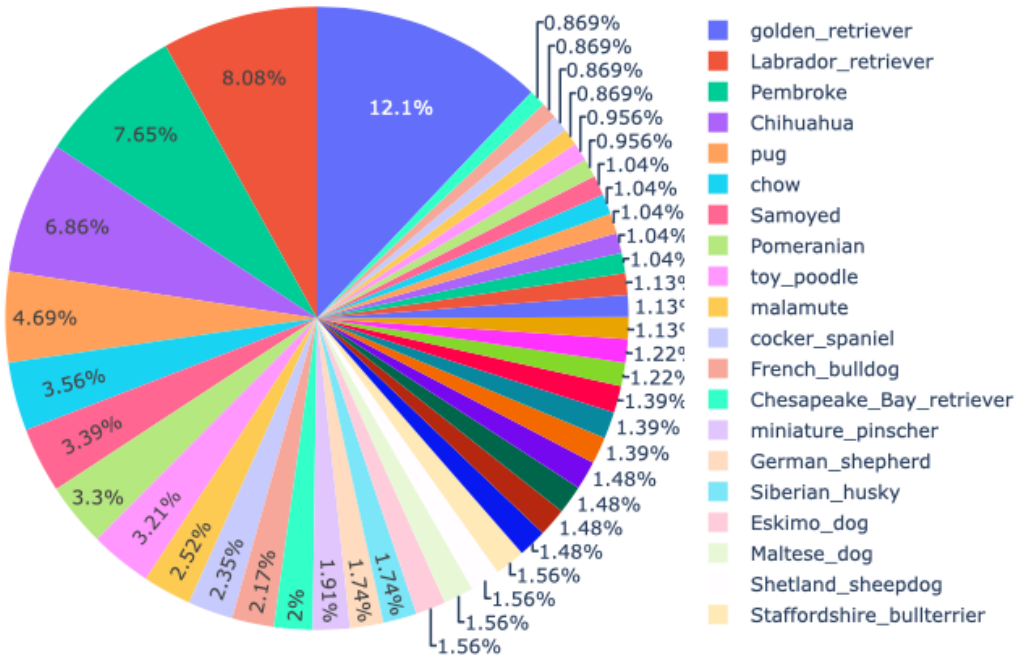
Exploratory Data Analysis

This is the final stage, where with a tidy dataset, it is possible to study, analyse and answer the questions initially posed.

What are the most common dog breeds on WeRateDogs?

A new dataset was generated from counting each breed by grouping.

In order to achieve a readable graph, the breeds were limited to those that were in at least ten records.

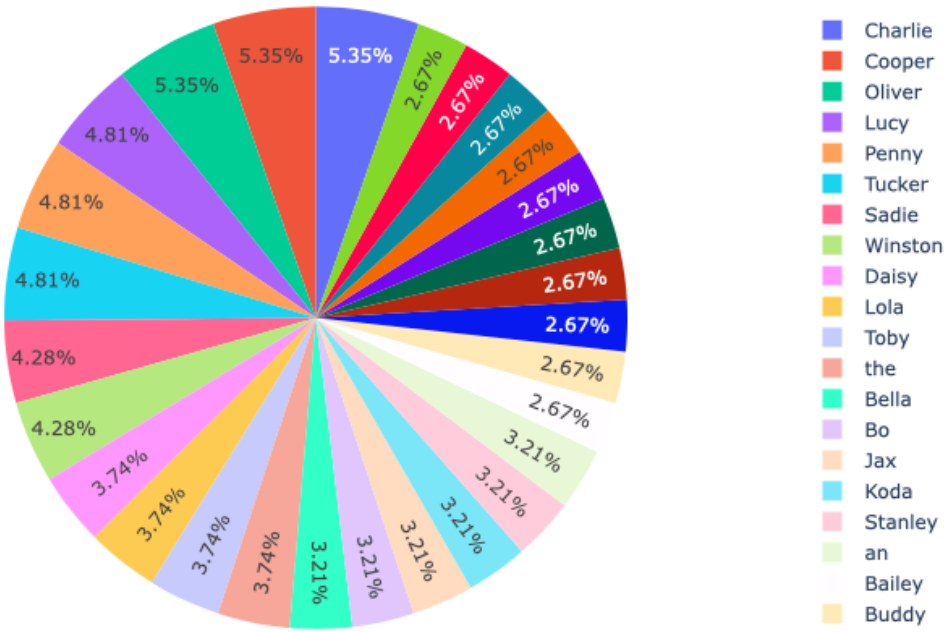


Most common breeds

Most common breed, are the Retrievers. Golden Retriever with 12.1% and Labrador Retriever with 8.08% are most preferred; together they add up to 20% aprox.

What are the most common names?

Similar to the breeds, a dataset was generated with the names and the count of each of them, those that were repeated at least 5 times were kept in the analysis.

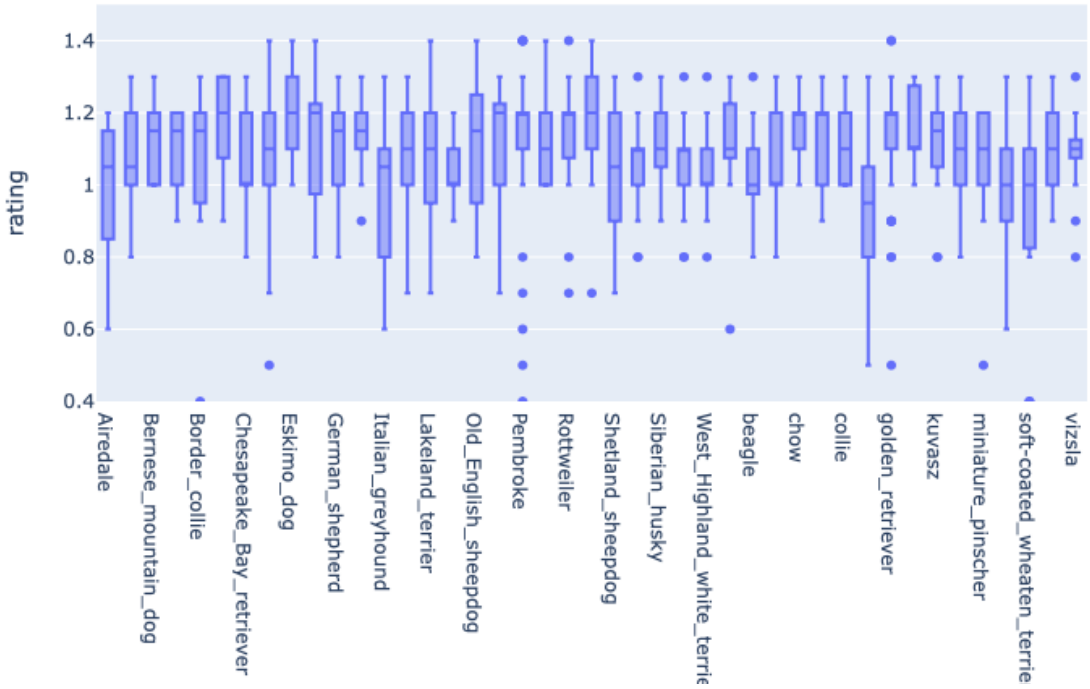


Most common names

Charlie, Cooper and Oliver are the most common names of the dogs in WeRateDogs.

Which breeds achieve the highest ratings?

To determine the rating, the numerator was divided by the denominator already reported in the dataset, generating a new column with this calculated field.

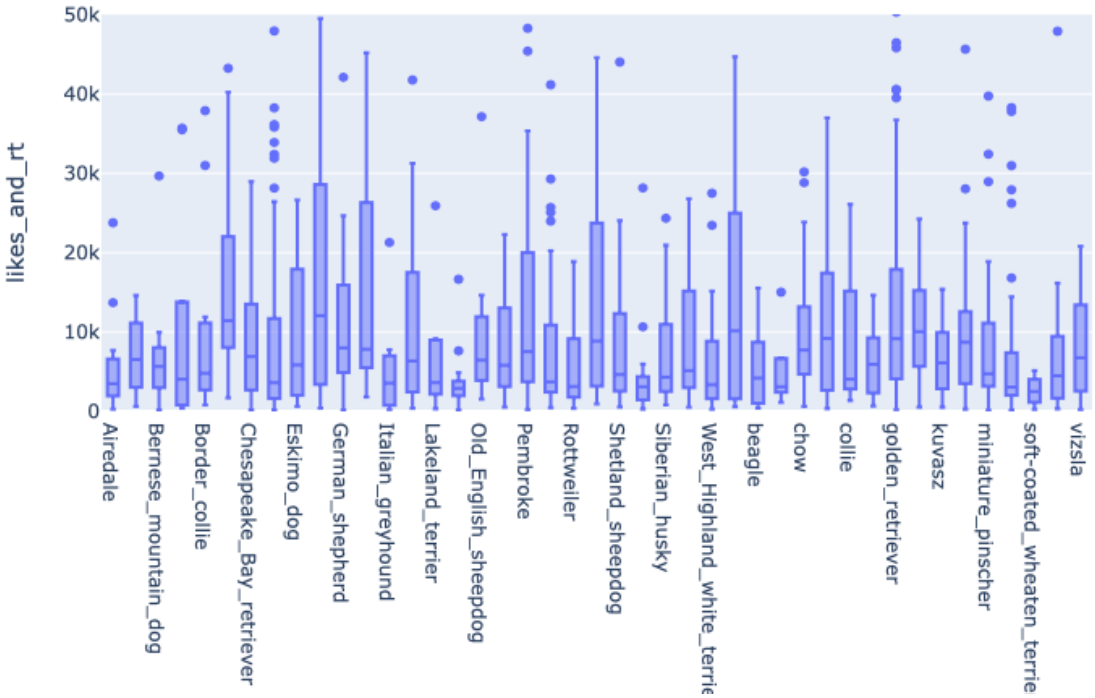


Ratings by breed

Eskimo_dog, Samoyed and Cardigan are the top breeds, rated by the Twitter account @dog_rates.

Which breeds get the most reactions (Retweets and Likes)?

Similar to the previous case, to determine user reactions, a field calculated by adding the retweets and favourites of each record was generated. There is probably some information that gives greater value to one or the other, but for this work we use the assumption that both have equal value.

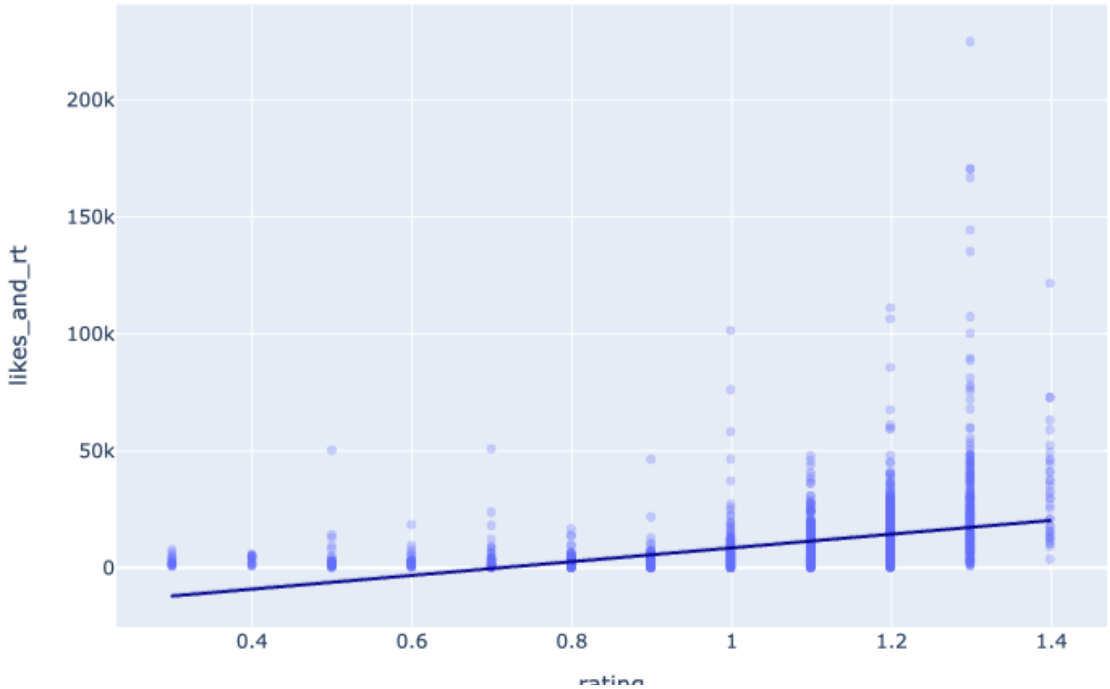


Reactions by breed

Breeds more reacted by users in Twitter are: French Bulldog, Cardigan, Basset.

Is there a relationship between number of reactions and rating?

Both variables were plotted and a linear regression line was drawn.

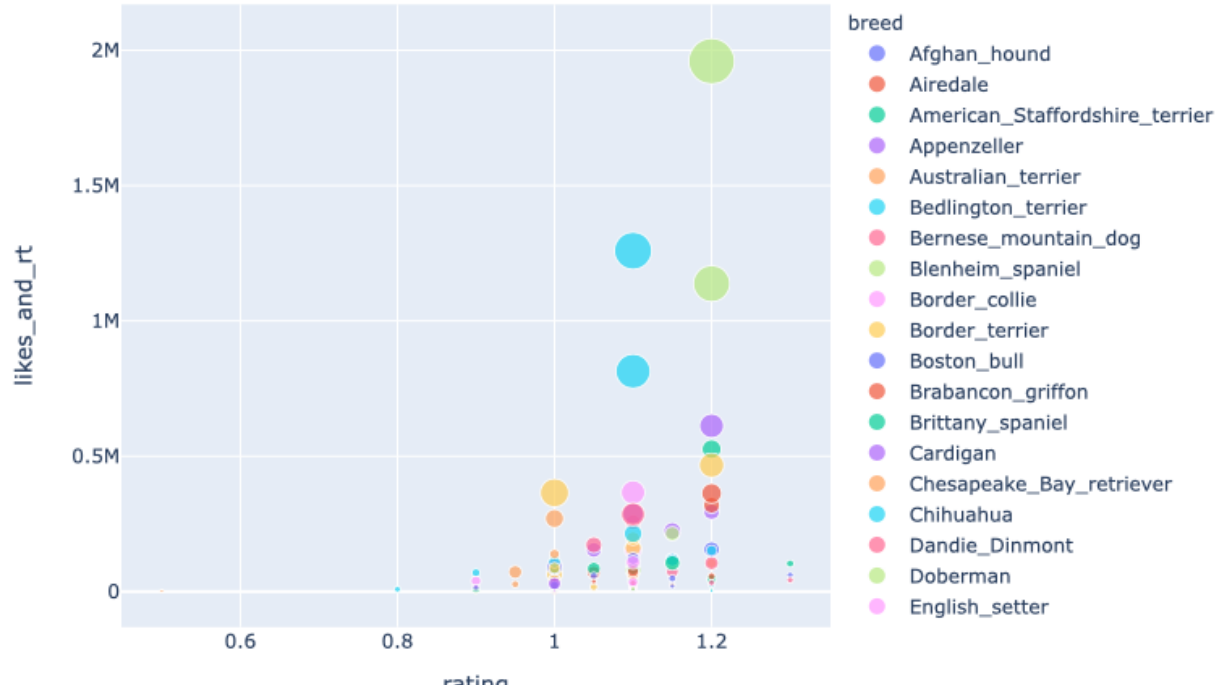


Rating vs Reactions

The graph shows no major relationship between the preferences of WeRateDogs and the preferences of the users. Although a positive relationship is generated, its slope does not show a steepness, which indicates that it does not reflect much of a relationship between the two, and its r^2 value is low, making the line unrepresentative.

The big picture

A bubble chart is presented, where four variables are shown on the same graph.



Big picture

Each bubble represents one breed. Size indicates the number of records (rows) with its breed. X-Axis shows the rating given by WeRateDogs and Y-Axis the reactions by Twitter users. With this information is possible affirm that Golden Retriever is the most common, valued by users and by WeRateDogs.

Conclusions

This work consisted mainly of the application of Data Wrangling techniques. Data Wrangling involves three main stages: Gather, Assess and Clean; all of them were applied in this study.

In addition to Data Wrangling, Exploratory Data Analysis work is carried out.

This study consisted of the analysis of the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

A total of three data sources were considered. One file in CSV format, which in this paper is referenced under the name 'primary', is provided by Udacity; it contains incomplete information on the Tweets to be analysed and the file described in the previous paragraph. The second file is in TSV format, it contains the identification of the breed of each dog published in the CSV file, this identification was performed by machine learning; it is also provided by Udacity. The third source of data corresponds to one generated by this same student, from the data obtained from the Twitter API, to achieve this the ID of each tweet is sent and the detail is received in JSON format, this allows the missing information in the primary file to be completed.

The following data quality problems were identified and corrected:

1. Non-descriptive name of column 'p1'.

2. Non-descriptive name of column 'p1_conf'.

3. Non-descriptive name of column 'p1_dog'.

4. Filter tweets that correspond to retweets, as it is requested in the statement of this project, not to consider these tweets.

5. Correct unreported values for expanded_urls.

6. Correct null values for the columns corresponding to dog breed ('doggo','floofer','pupper','puppo').

7. Correct values for dog breeds that do not correspond to dog breeds, but to other English words. Apparently, objects that the learning machine identified instead of the breed of the dog.

8. Corrected outliers for the column 'rating_numerator'.

9. Corrected outliers for the column 'rating_denominator'.

10. Improved readability of the 'source' column.

Identified and corrected tidiness issues.

1. Dog sizes correspond to the same variable, which is the size of the dog; however, it is reported in different columns.

2. The column 'expanded_urls', which contains several URLs in the same cell, is separated to leave one in each column.

As for the questions initially raised, they can now be answered:

1. What are the most common dog breeds on WeRateDogs?

R: Most common breed, are the Retrievers. Golden Retriever with 12.1% and Labrador Retriever with 8.08% are most preferred; together they add up to 20% aprox.

2. What are the most common names?

R: Charlie, Cooper and Oliver are the most common names of the dogs in WeRateDogs.

3. Which breeds achieve the highest ratings?

R: Eskimo_dog, Samoyed and Cardigan are the top breeds, rated by the Twitter account @dog_rates.

4. Which breeds get the most reactions (Retweets and Likes)?

R: Breeds more reacted by users in Twitter are: French Bulldog, Cardigan, Basset.

5. Is there a relationship between number of reactions and rating?

R: Based on the information available, it can be stated that WeRateDogs' preferences are not necessarily the preferences of its audience.