Data Analysis practice

Data Wrangling practice

We rate dogs dataset analysis

Alejandro Sierra
Email: asierra21@gmail.com

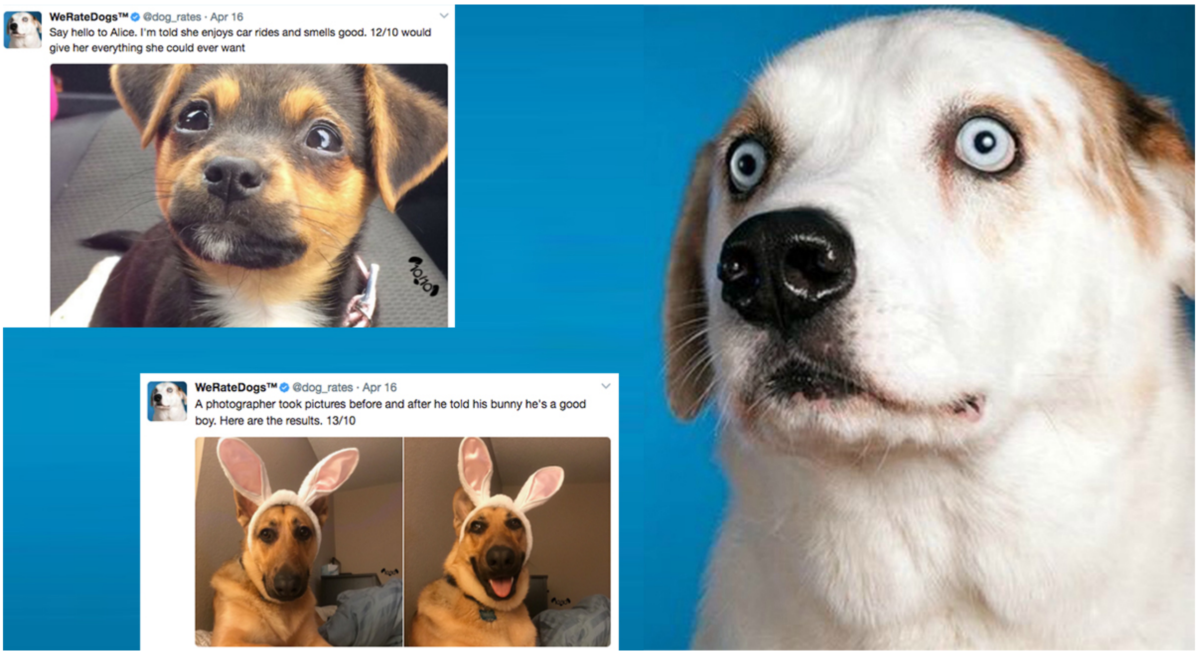
27th December 2020

Contents

1 Introduction

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user dog_rates, also known as WeRateDogs.
WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent."

WeRateDogs has over 4 million followers and has received international media coverage.



WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for our use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

The main goal for this analysis is to gather data from different sources which include:

- .csv files provided by Udacity
- .tsv files to be downloaded pragmatically from provided url.
- .txt file to be generated by extracting information from twitter API and storaged in JSON format.

Then this gathered data needs to be assessed and cleanend in order to deep into the information and present insights with their respective visualizations.

2 Gather data

2.1 twitter-archive-enhanced.csv

This file was provided by Udacity so we just need to read it into the python script via pandas library.

2.2 image-predictions.tsv

In this case, the file of interest is located in the following url.
'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'
We need to download it programatically to our working directory folder and then load it into the python script via pandas library considering tabular separation.

2.3 tweet_json.txt

This last file will be created from data extracted by the python script through the Twitter API and Tweepy library.
We are interested in reteweet count and like count for each of the tweet ids listed in the file provided by We rate dogs itself.
Some of the tweets have been deleted, so exceptions have to be handled.
Initially we will create a .txt file containing the JSON format data, save it to the working directory and load it back into the python script. By this practice, we avoid downloading the data everytime we run the script, since it verifies the existance of the tweet_json.txt before executing the API extraction.

3 Data assessment

The goal for this chapter is to assess the data visually and programatically in order to document the necessary wrangling for the dataset provided.

As a reminder, we will define the following terms:

- **Dirty data or Low quality data:** Content issues such as inaccurate data, corrupted data or duplicate data
- **Messy data or Untidy data:** Structural or organizational issues. Violates the tidiness definition where each variable forms a column, each observation forms a row and each type of observational unit forms a table.

Both of this data issues are going to be checked both visually and programatically.

3.1 Feature description

Once the dataset is conformed, we need to understand each of the variables:

twitter_archive:
The loaded file contains 2356 observations with 17 different features listed below.

- **tweet_id**
- **in_reply_to_status_id** tweet id to which the tweed is a reply
- **in_reply_to_user_id** user id to which the tweed is a reply
- **timestamp** time of the tweet
- **source** additional information that provides context about the Tweet and its author
- **text** text of the Tweet
- **retweeted_status_id** Id of the parent tweet if handling a retweet.
- **retweeted_status_user_id** User Id of the parent tweet if handling a retweet.
- **retweeted_status_timestamp** Timestamp of the parent tweet if handling a retweet.
- **expanded_urls**
- **rating_numerator** We rate dogs numerator
- **rating_denominator** We rate dogs denominator. Set to 10.
- **name** Dog´s name
- **doggo** A doggo is a full-size pupper.
- **floofer** A very fluffy dog
- **pupper** Puppy
- **puppo** Slang term for Puppy

image_predictions:
This file contains predictions for up to three different dog photos for each tweet.

- **tweet_id**
- **jpg_url** url of the image
- **img_num** number of image
- **p1** the algorithm's 1 prediction for the image in the tweet
- **p1_conf** how confident the algorithm is in its 1 prediction
- **p1_dog** whether or not the 1 prediction is a breed of dog
- **p2** the algorithm's 1 prediction for the image in the tweet
- **p2_conf** how confident the algorithm is in its 1 prediction
- **p2_dog** whether or not the 1 prediction is a breed of dog
- **p3** the algorithm's 1 prediction for the image in the tweet
- **p3_conf** how confident the algorithm is in its 1 prediction
- **p3_dog** whether or not the 1 prediction is a breed of dog

tweet_json:
This dataframe created from data extracted from Twitter API contains the following information.

- **tweet_id**
- **retweet_count** the ammount of retweets the tweet got.
- **favorite_count** the ammount of likes the tweet got.

3.2 Untidy data

Below, every finding is going to be documented.

- Files from different sources result in three different tables, when in fact, only one is needed.
- twitter_archive table : dog stage is listed in four different columns when it could be just one "dog_stage" column where the observation input is taken form the list {'doggo', 'floofer', 'pupper', 'puppo'}.

3.3 Low quality data

Also, it is very much recommended to perform a programmatic assessment using Pandas library.

- inconsistent number of observations among tables. Minimum one is tweet_json with 1813 entrys.
- twitter_archive table: Null observations for columns, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user, retweeted_status_timestam
expanded_urls.
- twitter_archive table: timestamp format adds +0000 in every tweet.

- twitter_archive table: missing dog name in many observations. Others have name "a", "this", or "name"
- twitter_archive table: expanded url with more than one url for some observations. Information is duplicated in the cell.
- twitter_archive table: In source column extract information dropping link.
- twitter_archive table: standard deviation for ratings is 45.9.
- twitter_archive table: Min rating is 0 which seems strange for this system.
- twitter_archive table: Max rating is 1776, which needs to be checked.
- twitter_archive table: Min denominator is 0, which is incorrect.
- twitter_archive table: Max denominator is 170 which is incorrect.
- image_predictions: Several predictions are not dog breeds for columns p1, p2 and p3.
- tweet_json: Seems strange that tweet with ID 841833993020538882 has 14.427 retweets but 0 favorites, this needs to be checked for more cases in the dataset.

4 Clean data

4.1 Combine tables

First cleaning action will be to combine the three tables provided into one called "master_clean" joined using tweet_id as foreign key. We will choose an inner join, addressing also the inconsistency among the three tables.
This table results in 1606 observations given the inner join restriction to have tweet_id present in all three tables.

This new dataframe contains only 47 observations for the retweet_status information, which will not be adressed until confirmation of the usage of these three columns.

4.2 Timestamp format

Timestamp column will be converted to date type format instead of object. The resulting column is dtype datetime64.

4.3 Zero rating

We will check the minimum of 0 for column ratings.
We can see that there are two observations with 0 rating. Checking the rest of the features, everything seems correct, conatains a denominator of 10, contains at least one image and the tweet contains text.
Considering this facts, we will consider the zero rating as a valid number.

4.4 High ratings

In this case it is hard to determine what a high rating is.
In the statistics for this column, we see that the third quartile is determined in a rating of 12 points.
Being conservative and taking as a "high rating" a value of 15, we see that there are 12 observations that go beyond that number.
Among them, we find that most of this cases, 9 out of 12, have a denominator that does not correspond to the value 10.
This issue is to be discussed in the following section.

4.5 Non standard denominator

As we could see before, there are observations where the denominator is different than the standard 10. These observations could be standardized and take the denominator to 10, or this could be considered to be decimal ponits in the numerator that were considered as a separated field by the parser when loading the data. So standardization will be performed when denominator is not 10, and decimal convertion and round to integer will be applied when denominator is already 10.

4.6 High ratings revisited

As an interative process, we return to analize the high ratings after denominator standarization.

Now we have only three values to take care of, with indexes {426, 612, 1328}.This observations are highly inconsistent with the rest of the dataset, therefore I consider the best option is to remove them.

4.7 Ratings standard deviaton

After handling outliers and standarization, we see that the standard deviaton for ratings is now 2.27 out of the 10 scale.

4.8 Source information

Source information is presented within a link, which makes it hard to find the information presented. We need to extract the source from the link provided.

We can see that we have three different sources, "Twitter for iphone", "Twitter web client" and "TweetDeck" being by far, the first category the predominant one.

4.9 Dog stage column

Four different columns describe Dog stage. We are going to tidy it into just one column that shows the result out of the four.

After mergin the four columns into one, we check the value counts and see that there are 1300 of "None" fields.

No action will be taken regarding this issue.

4.10 Expanded urls

There are some fields where url information is duplicated.

After running a for loop checking whether the information matches and can be considered duplicated we see that for 27 observations this is not the case.

These 27 cases, anyway, have starting information that gets splitted by the "," but that

do not correspond to a url.
Checking the following information we have a complete dataset match for duplicated information when we have more than one entry for the url.

4.11 Dog names

Dog names are not always accurate. In this dataset we have 428 "None" and 48 "a" names.
In this case, this values represent a high percentage of the whole dataset, therefore, we will keep the column as it is.
Anyway, while analyzing the text in look for the dog name, we found that in the text column, we can find many tweets that do not correspond to a dog, and therefore, the text mentions explicitly "We only rate dogs". Example "This is a taco", or "This is a carrot" We will look for this phrase in the text column for the whole dataset.

There are 42 observations that contain the text "We only rate dogs." so we could consider these rows as misleading information.
By checking at the information for this rows, we can see that most of them have a rating over 10, a considerable ammount of favorites and retweets, Anyway, I consider the best option to drop this rows in order to keep the dataset dog related.

4.12 Zero retweets and likes

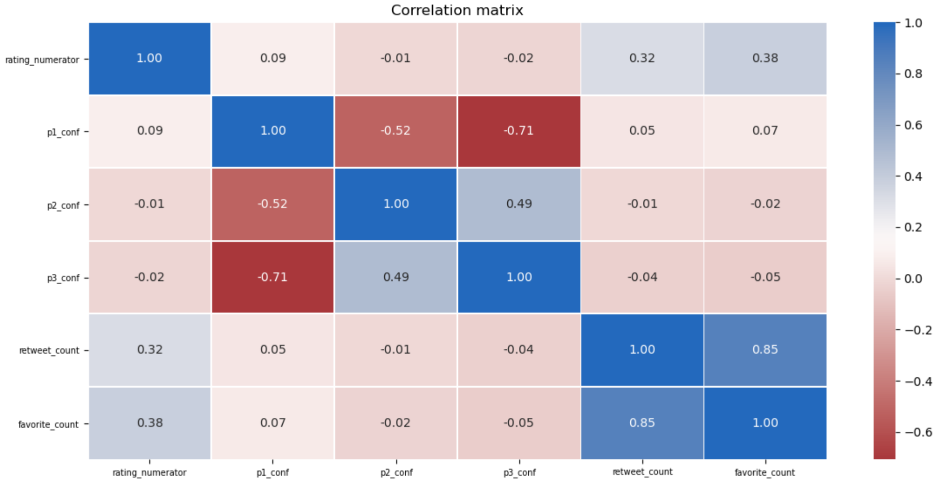
There are 44 observations where likes count is 0. No observations have 0 retweets.
After analyzing text, retweets, timestamp and ratings, there is no evidence for this tweets to be missleading so we will consider them as valid information.

5 Data exploration

In this chapter we are going to study the data and perform an analysis on the variables rating, favorite_count and retweet_count.
We would like to determine whether rating system and favorites are correlated.
Besides, we are going to explore what breed of dog gets the higher ratings and favorites or retweets.

5.1 Feature correlation

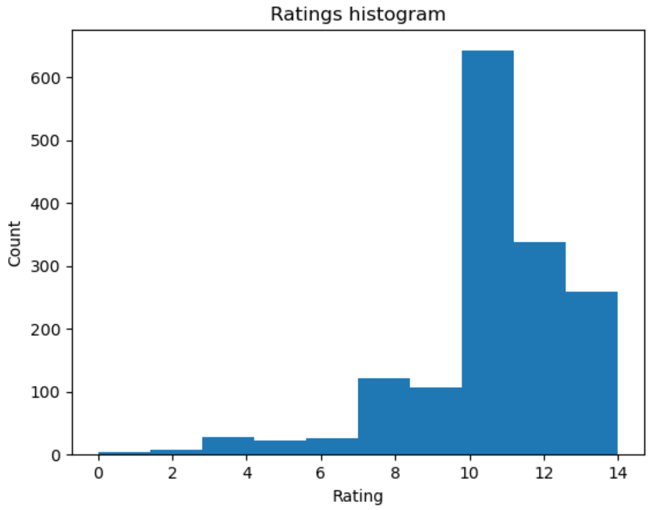
First thing we are going to explore is the correlation matrix.



As we can see in the correlation matrix, the only high interesting relation is, as expected, between favorite counts and retweet counts.
There is a low correlation also between rating and both favorite count and retweet count.

5.2 Ratings

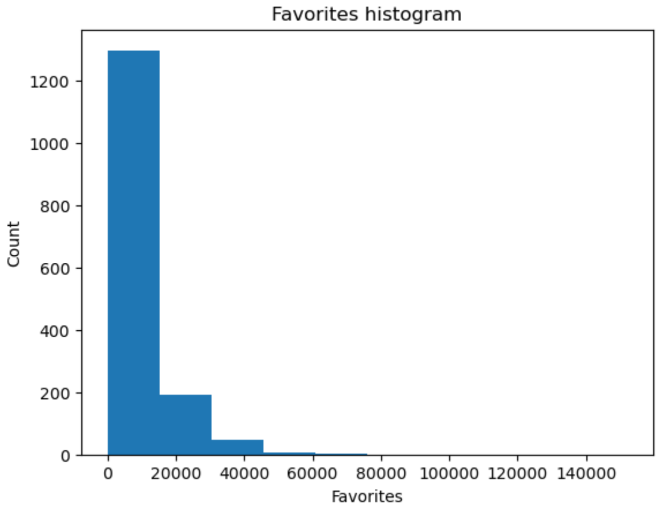
The univariate exploration for numerical variables is best explained by histograms.



As we can see, ratings higher than 10 are really common in this popular site, being the rating 10/10 the most used.

5.3 Favorites

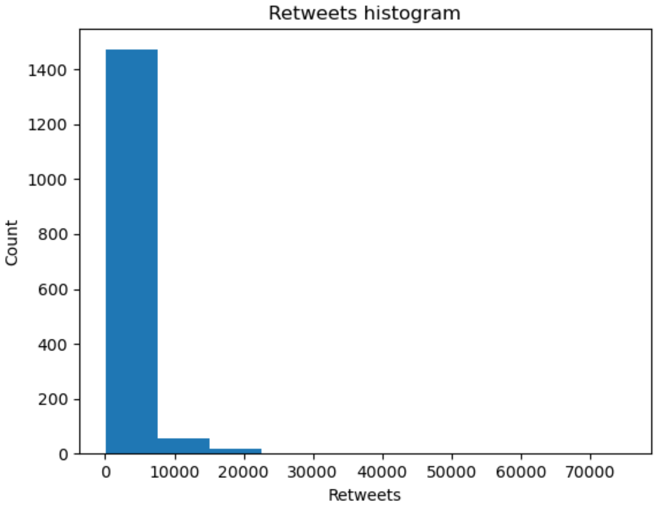
The following plot will show the likes distribution a tweet gets.



We can see in this plot that the ammount of favorites tweets get is commonly in the range 0-1000.
It is not usual to get over 4000 likes for a tweet in this account.

5.4 Retweets

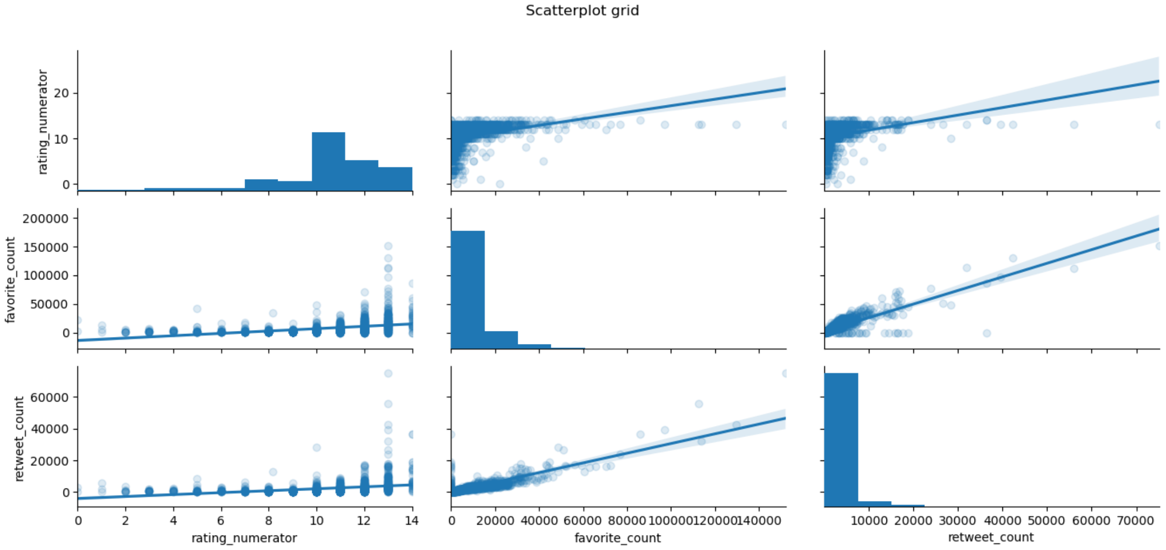
We repeat the analysis for retweets to check their distribution.



As we saw with favorites, retweets behave in the same way. More than half of the dataset gets around 500 retweets or less.

5.5 Bivariate exploration

Ratings, retweets and favorites:
Now we are going to analyze how these variables relate to each other with a scatterplot grid.



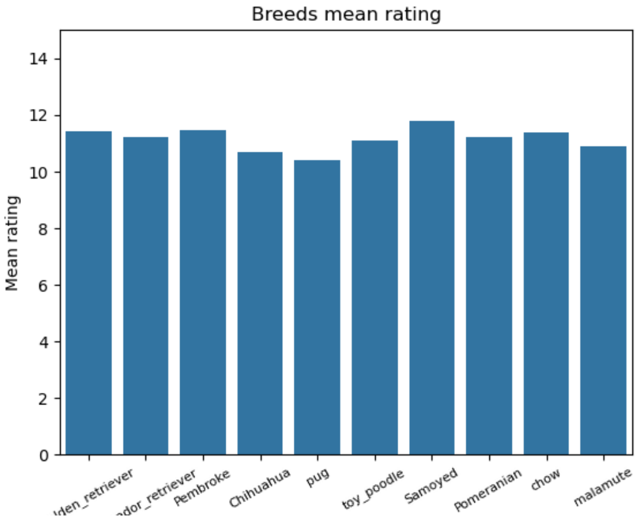
We can see that for ratings higher than 10, both favorites and retweets start to increase. Below this threshold, they show a stable behavior with fewer interactions.
Favorites and retweets seems positively correlated with a linear relation with the key aspect of the slope not going through the origin.

Ratings by dog breed:

We are going to group by the first prediction made for the tweets picture and see what is the top ten dog breeds that show up.

Breed	Count
Golden_retriever	114
Labrador_retriever	66
Pembroke	64
Chihuahua	61
Pug	47
Toy_poodle	34
Samoyed	33
Pomeranian	29
Chow	28
Malamute	20

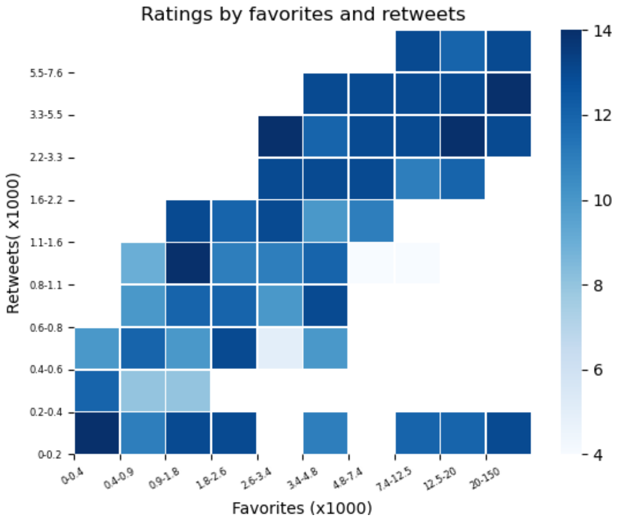
And now we are going to plot the mean rating each of those breeds recived.



This plot that there is not a considerable difference between dog breeds that could determine a rating increase or decrease.

5.6 Multivariate exploration

Ratings by favorites and retweets:
We are going to see how ratings relate to both variables, favorites and retweets.



The heatmap shows, as expected, that with the increase of favorites and retweets, we have darker blues corresponding to higher ratings in the top right corner.
Relation seems linear across the diagonal for the heatmap.

6 Conclusions

The main conclusion for this dataset is regarding the ratings system, dogs usually get a rating over 10, higher than the scale maximum.

Another main finding is that there is not a prevalent breed in terms of rating mean, while there is one in terms of tweet counts, being the Golden Retriever the most popular one.

There is a positive correlation between likes and retweets, and that correlation is accompanied by the ratings, which increase with the other two variables.

Most tweets get between 0-1000 likes and between 0-500 retweets.