Wrangling data(We rate dogs project)

first of all, I gathered three different source of data, the first one:

from the given CSV file (twitter-archive-enhanced.csv), second, from the link('https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'), I download tsv file by using request library to read image_predictions.tsc file.

Lastely: after getting twitter developer account, I used Tweetpy to gather JSON data and access to API and save it at json.txt. Then I take some attributes such as tweet_id , favorites and retweets rates.

After gathering data, I assessed data by using pandas functions such as, info(), describe, sample(),...

Then I started cleaning part, so that I defined which data have issues, which are:

From Twitter-archive table:

Quality¶

`Twitter-archive` table:¶

- None values in name of dogs and incorrect dogs name like: "a", "an", "such","the", "very"...
- tweet_ id should convert to str.
- Useless columns such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id ,retweeted_status_timestamp should be removed.
- timestamp should convert to timestamp datatype.
- there is no particular limit or specific numeric values of rating denominator such as 0, 2,7,...
- rating_numerator and rating_denominator should convert to float.

`image_predictions` table:¶

- tweet_id should convert to str

`tweets_info` table:¶

- favorite and retweets columns should convert to int datatype.

Tidiness :¶

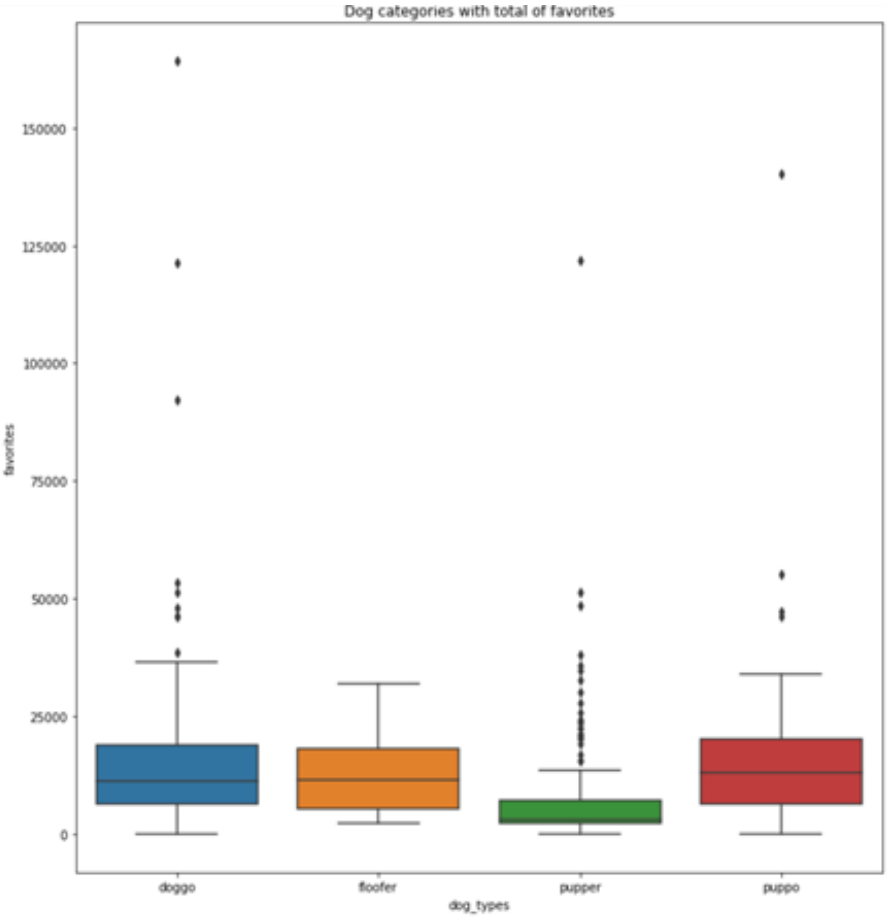
- We have 3 datasets, `twitter_archive` , `image_predictions`, and `tweets_json` dataset , we should merge it in one dataframe.
- We have three seperate columns of dog categories.

After finishing wrangling part which is encompass the given data, here we want to answer some questions by using visualizing.

We used some attributes after we gathered ,assessed, and cleaned our data, such as, retweets rate , favorites rates, dog categories and dog names.

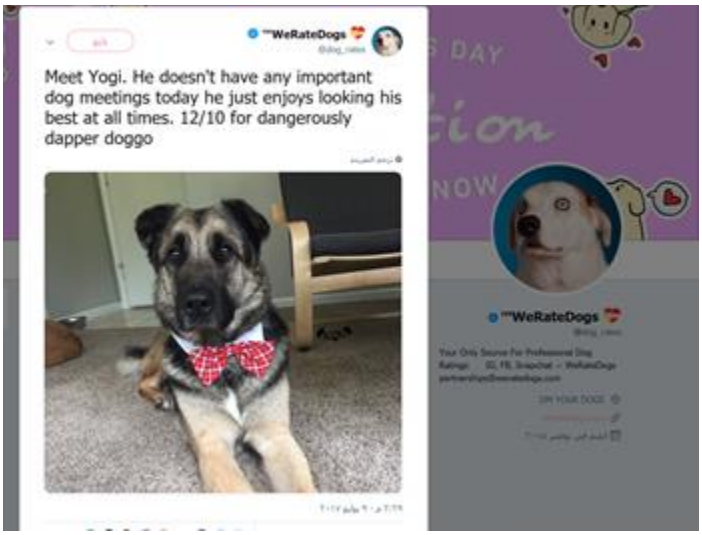
Find out the dog category with the highest favorites rate:

In this first part ,we want to find out the category which is recorded the highest rate of favorites, after we removing None values, and visualizing the the chart, we can say that , puppo dog recorded the highest rate, then doggo, floofer, and pupper.



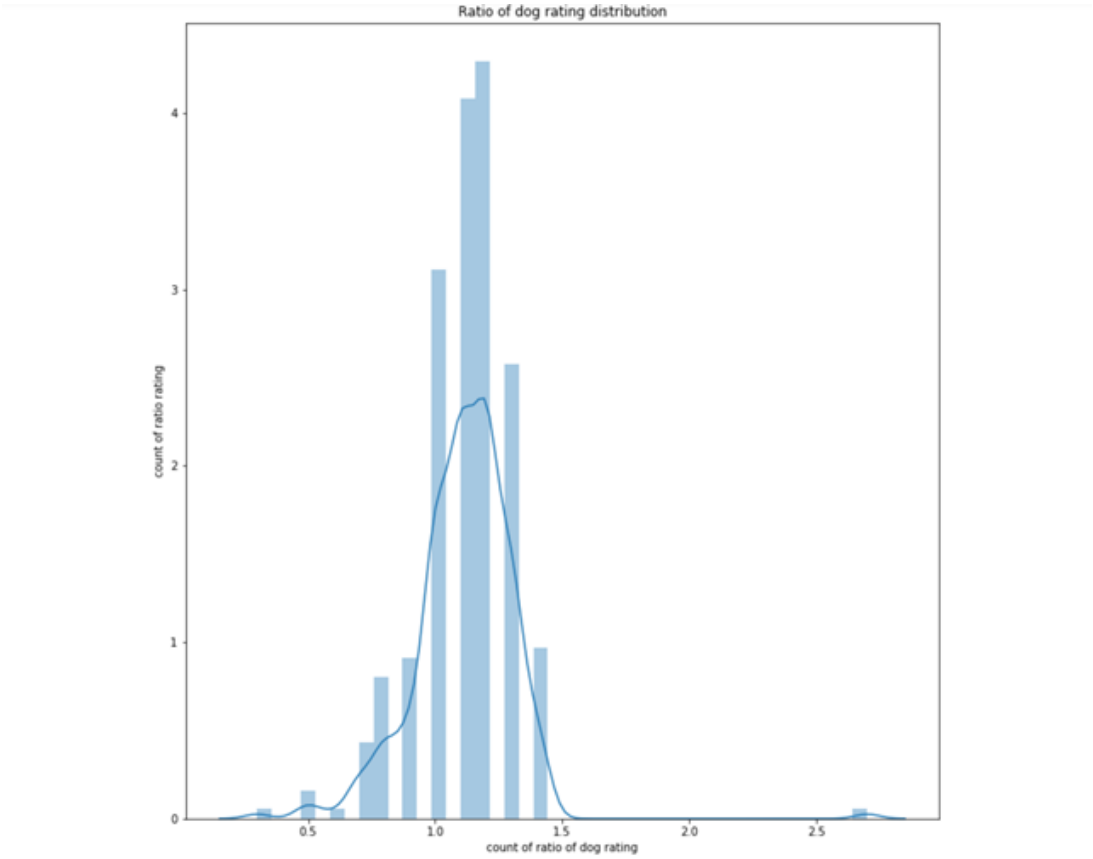
Here we want to see dog categories which is take the highest rate of favorites:

As we can see at the image below:



Discover the ratio of dog rating distribution:

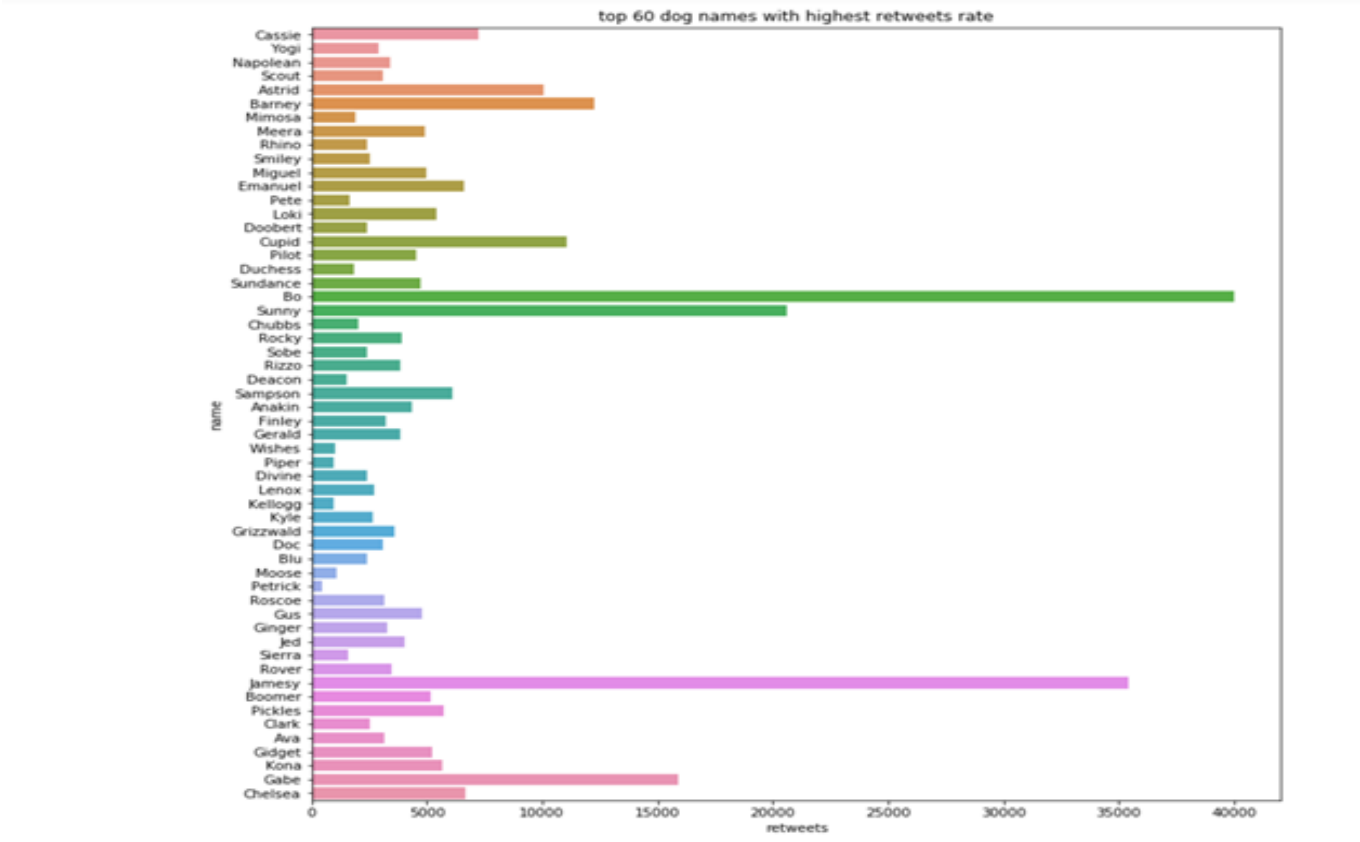
Here after we calculated the ratio of dog rating from(rating numerator /rating denominator)rate, we can see that the chart takes normally distribution of dogs rating.



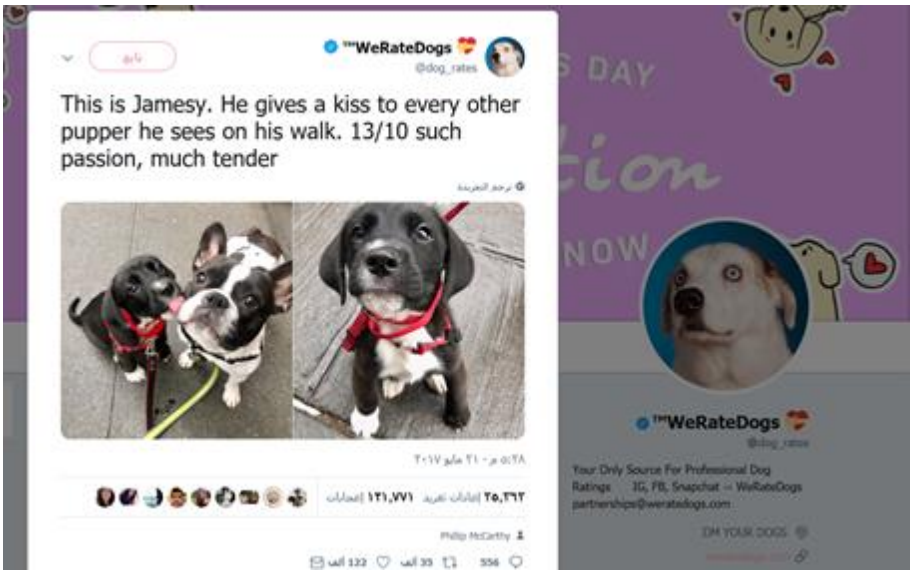
Find out the top 60 of dog names with retweets rate:

In the last part, after removed None values, and removed inappropriate dog names such as ("a", "an", "the", "very", "quite",... we, want to find out the top 60 of dog names which are recorded the highest retweets rating.

As we can see from the chart the name of dog ("Bo") recorded the highest rate of retweets, then the name of dog "Jamesy".



Here the images of B and Jamesy dogs:



Limitation that I found:

As a one of a community that does not use dogs except in guard, I have encountered many obstacles in understanding these data so that I can analyze them. For example: I was not familiar with dog categories and types.in other hand, using twitter's API, that make a sense of the given data, such as finding out the retweets and favorites rate and accessing of tweet Id, to see the image of each dog , in which helped me to understand how can I start analyzing then finding out the result that I want to found.

Resources:

https://www.udemy.com/data-wrangling-in-pandas-for-machine-learning-engineers/learn/v4/t/lecture/8758244?start=1

https://classroom.udacity.com/nanodegrees/nd002-mena-connect/parts/71de6fde-0474-4933-85c8-312aa416cbfe/modules/564dc4fd-c702-4871-980c-4ca5605f91c6/lessons/29f2ae4c-ed5b-4fc2-981d-76dc308a1b4b/concepts/9d3c80f7-3b67-42bc-8513-7fe0c71ef83e