Final Project Wrangling and Analyzing Twitter Data

Table of Contents:

1- Data Gathering
2- Data Assessing
3- Data Cleaning
4- Analyzing and Visualizing Data

Project Overview:

This project includes several datasets from Twitter which include data about a Twitter account called WeRateDogs, in this project I am demanded to assess, clean, and visualize the data with the things I have studied along the course.

Gathering Data:

The first step in any data-wrangling process is to gather the data which have many ways to do this either by reading a CSV file, gathering data by web scrapping, gathering data from other tables, online by downloading data, a given data by your company, or by an API if you have the keys. In this project, I used 3 types of data gathering:

- Reading a CSV file ("twitter-archive-enhanced")
- Download online ("image-predictions")
- Twitter API ("tweet_json")

The first step is done now as I gathered the data I want let's move to the next step.

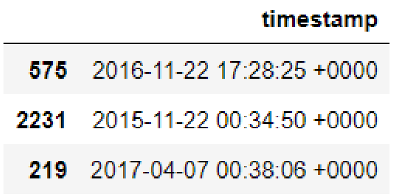
By: Omar Mohammed

Assessing Data:

This is one of the most important steps as with it I will search for the issues and misleading data. There are 2 types of assessing data: Visual Assessment, and Programmatic assessment, in the following there are the issues I discovered and their classification.

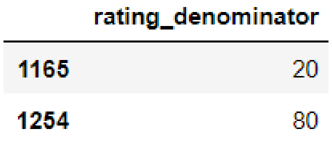
Visual Assessment:

1- timestamp column contains +0000 at their ends.



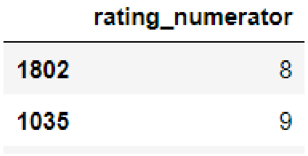
	timestamp
575	2016-11-22 17:28:25 +0000
2231	2015-11-22 00:34:50 +0000
219	2017-04-07 00:38:06 +0000

2- The rating denominator is not always 10 as it should be.



	rating_denominator
1165	20
1254	80

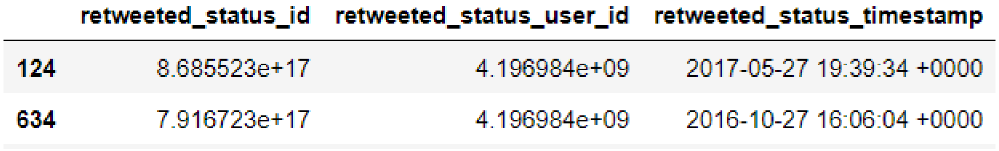
3- The rating numerator has outliers as the range must be from 10 to 20.



	rating_numerator
1802	8
1035	9

4- The source, in_reply_to_status_id, and in_reply_to_user_id columns don’t contain any important information.

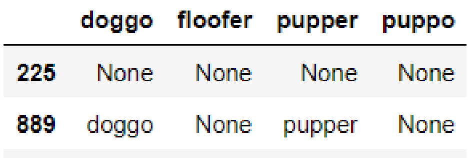
5- retweets are included in the database, and they must be removed.



	retweeted_status_id	retweeted_status_user_id	retweeted_status_timestamp
124	8.685523e+17	4.196984e+09	2017-05-27 19:39:34 +0000
634	7.916723e+17	4.196984e+09	2016-10-27 16:06:04 +0000

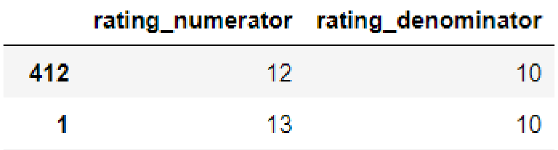
6- There is no use for the column of retweets after we delete retweets.

7- The columns named "doggo", "floofer", "pupper", and "puppo" can all be joined into 1 column.



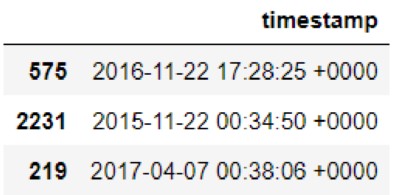
	doggo	floofer	pupper	puppo
225	None	None	None	None
889	doggo	None	pupper	None

8- The rating_denominator and rating_numerator column can both be joined into 1 column.



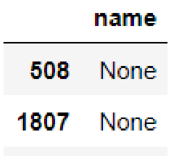
	rating_numerator	rating_denominator
412	12	10
1	13	10

9- the column time_stamp can be split into 2 columns to avoid confusion



	timestamp
575	2016-11-22 17:28:25 +0000
2231	2015-11-22 00:34:50 +0000
219	2017-04-07 00:38:06 +0000

10- In the dog name's column missing data is labeled as "None" which is wrong as the computer reads it as data and not null.



	name
508	None
1807	None

11- Tweets beyond August 1st, 2017 are not needed as demanded from the project because the algorithm won't work on them.

12- some columns I want from tweet_json file into the tweet archive file so I can analyze only 1 file.

Programmatic Assessment:

1- The column tweet_id dtype is an integer which is wrong it must be object.

2- Missing data.

By: Omar Mohammed

Quality Issues	Tidiness Issues
1. timestamp column contains +0000 at their ends 2. The rating denominator is not always 10 as it should be 3. the rating numerator have outliers as the range must be from 10 to 20 but it appears there were numbers above 20 and others lower than 10 4. some dog names contains only letters such as "a" or "O" which can't be a dog name 5. The source, in_reply_to_status_id, and in_reply_to_user_id columns doesn't contain any important information 6. tweet_id dtype is integer which is wrong it must be object 7. retweets are included in the database and they must be removed 8. we can remove all three columns related to the retweets as we won't use them after we distinguish the retweeted rows and delete them. 9. There is some missing data in some columns	1. The columns named "doggo", "floofer", "pupper", and "puppo" can all be joined into 1 column 2. The rating_denominator and rating_numerator column can both be joined into 1 column 3. the column time_stamp can be split into 2 columns to avoid confusion 4. In dog name's column missing data is labled by "None" which is wrong as the computer reads it as a data and not nulls 5. Tweets beyond August 1st, 2017 are not needed as demanded from the project because the algorithm won't work on them 6. some columns i want from tweet_json file into the tweet archive file so i can analyaze on only 1 file

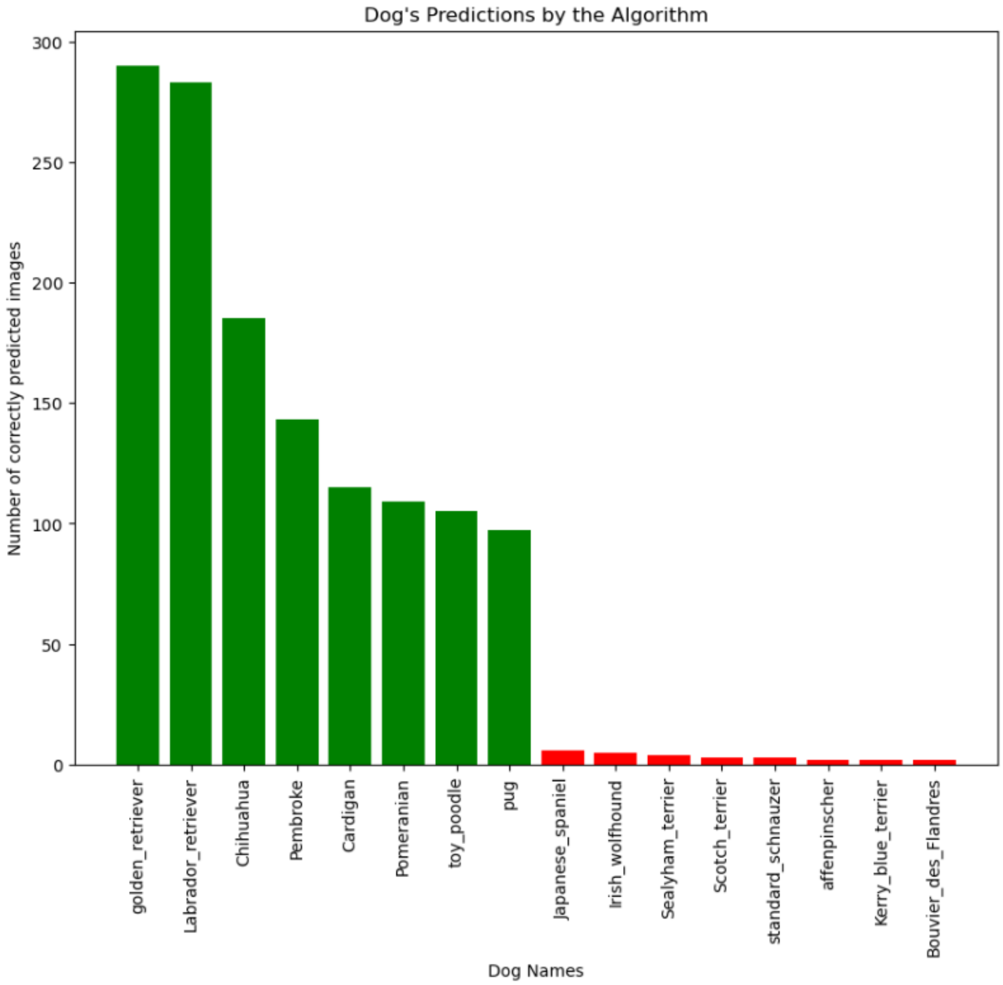
Cleaning Data:

Before cleaning data the most important thing is to take a copy of the original data so that if something goes wrong your original data will not be messed. I cleaned most of the issues I said above except 2 issues which are the dog names and the rating numerators outliers, at first sight, I thought that these numbers were wrong numbers it appeared that these ratings are normal and were written in the Tweet itself I tried to solve this problem for so much time but in the end, I discovered it's not a problem from the beginning, and for the dog names it seems that i can't gather the names of these dogs as they are not either available in the tweet or any other table so i will only turn them to nulls

Analyzing and Visualizing Data:

Insight 1 Dog's Predictions by the Algorithm:

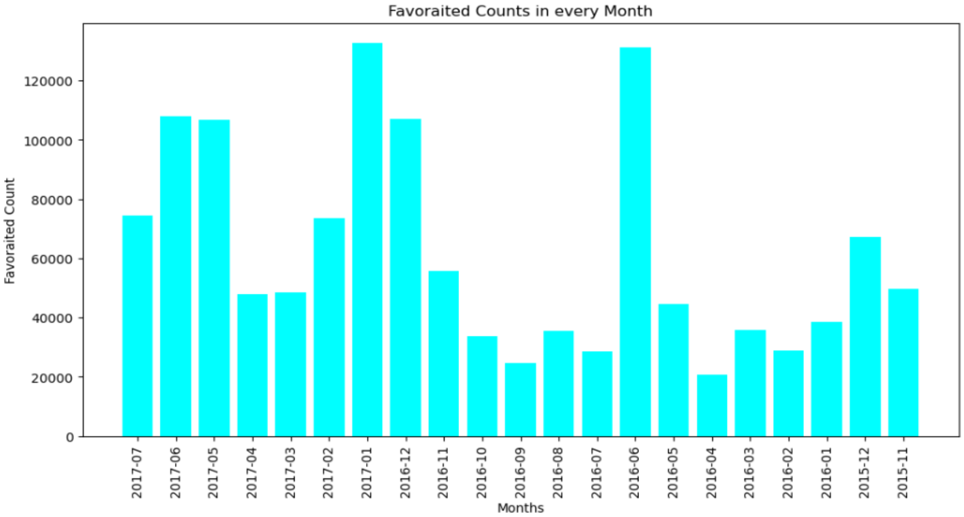
We can see from the graph above contains the highest correct times the algorithm predicted the dog type correctly and the lowest times the algorithm predicted correctly, in the above graph the dog types that the algorithm predicts correctly it seems that the golden retriever is the easiest for the algorithm to predict and the hardest to predict is the bouvier des flandres, so we can provide more photos of the bouvier des flandres dog breed to the algorithm until it gets them right and same goes for the dog breeds which have a low success rate.



By: Omar Mohammed

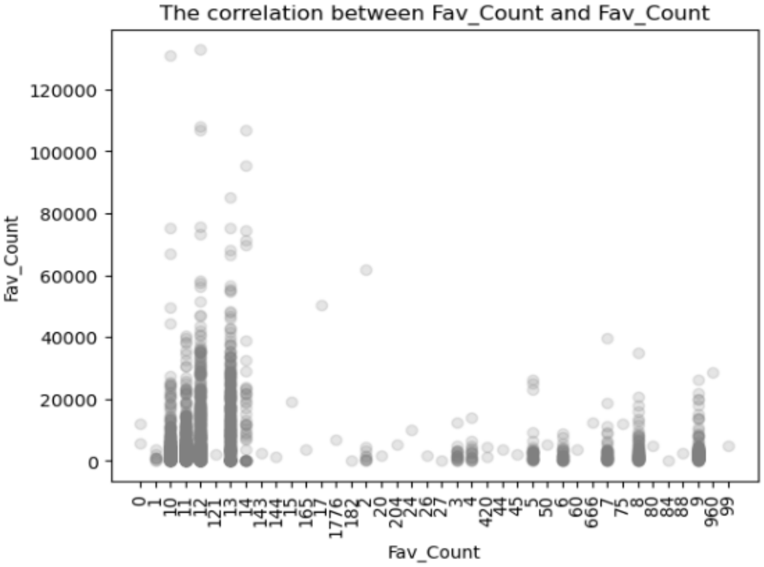
Insight 2 Favorited Counts in every Month:

the graph above represents the number of favorites on tweets for every month in the 3 years 2015, 2016, and 2017 although not all months is appearing because of the data, I can detect from this graph that every year there are 2 months when the favorite counts increase so much which is in both December and June, I think it's related to vacations because at the end of the year, there is a vacation and when working people sit home they usually use their phones so they will interact more with social media, that may be a weak reason but that's the only thought i can gather about this.



Insight 3 is there a correlation between the rating of the dogs and the number of people who fav the tweet?

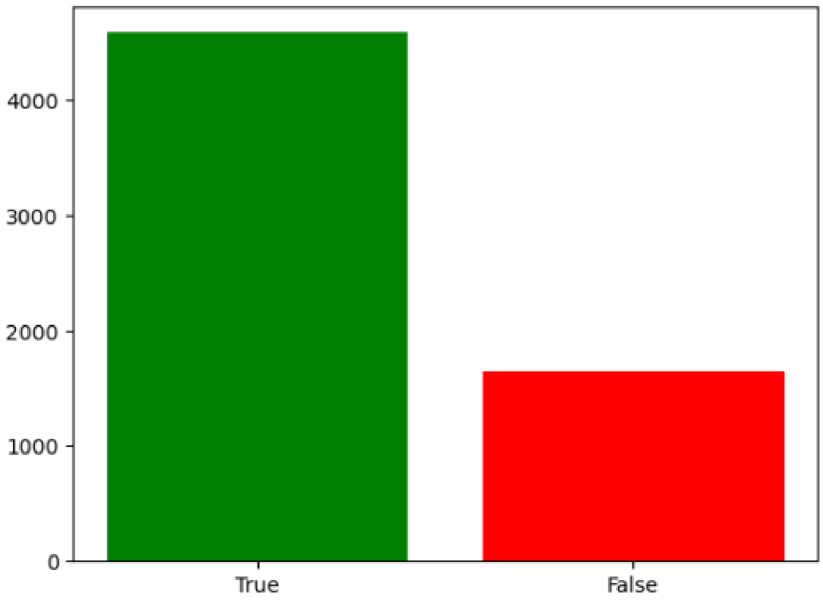
it doesn't seem there is any correlation between the rating given and the number of favorites on the tweet.



By: Omar Mohammed

Visualization: Model's efficiency

it seems that the model works pretty well he got more than half the answers right.



In Conclusion that was it all about my project I hope you liked it.

maybe at the assessing part otherwise nothing particular

By: Omar Mohammed