Project: Wrangling and Analyzing Data

There is a saying that man's best friend is Dog. I often wonder why? We have many animals been domesticated in different homes. On some days, the Doctor Strange in me will wander into earth Z, located in a different universe in which there is a history where dogs came to rescue man from invading alien forces led by Thanos. Or perhaps, this is a function of numerous hollywood movies where I have seen dogs fought side by side with man. But I have also seen donkeys, birds, horses and cats fighting side by side with Man. What then is special about dogs and why are they called man's best friend?

Copied from Udacity

I have not yet figured out the answer to this question. However, what I have found is a twitter page in which dogs are rated. Before I wander again into another universe and bore you with some medieval contents, let me come back to my earth and talk about this twitter page, the thought process behind the analyses of the data extracted from this platform.

The weRateDogs twitter page is a platform that rates people's dog, adding humorous comments to it. I ventured to analyse extracted data from this page in continuation of the Udacity data analysis nanodegree program. I was involved in the data wrangling project. The project tested students' ability to go through the different data wrangling stages which include gathering, assessing and cleaning data.

Gathering Data.

To get insight from data, one needs to have access to this data. Data gathering is a step towards making the data available for analysis.

Firstly, I imported the pandas, requests, Json, wordcloud, numpy, seaborn and matplotlib libraries which will be used during the project. Thereafter, using the pandas library, I proceeded to reading the twitter archived and Json files provided. The third file was a tsv file, extracted from the image platform using the requests library.

Assessing Data

Assessing data requires been an investigator. Often times, data to be analysed are not always clean. Due to various technical, human and domain knowledge gaps, data gathered are not in a ready-made stage to analyse thereby requiring auditing skills to find out the messiness of the data.

Two key data assessment process include, the visual assessment and programmatic assessment. Each with different techniques. I employed both the visual and programmatic assessment techniques to investigate the tidy and qualitative issues of the data.

Visual Assessment: This technique involved scanning through the data using excel or jupyter notebook to find out obvious issues with the data. This method provided quick wins but is not scalable when the data is large. Mostly used on small data.

From this assessment, it can be seen that some twitter_archive_data columns have missing data. Some values in the name column does not look like real dog names and the presence of replied and retweeted tweets.

Programmatic Assessment: This was done using pandas tools such as the sample, head, info and describe methods. This brought up some quality issues such as timestamp data type, incorrect numerator and denominator values, some inconsistencies in the capitalization of values in the p1, p2, p3 columns.

Summary of issues with the data are;

Quality issues

- Twitter_archive_data has retweeted data
- Twitter_archive_data has replied data
- Twitter_archive_data has columns with missing values
- Twitter_archive_data timestamp has an object data type instead of datetime data type
- None values in twitter_archive_data
- Some numerator and denominator ratings are not correct
- Stop words seen in twitter_archive_data name colum
- Inconsistent capitalization of values in p1,p2 and p3 columns of image table
- Tweet_id columns in all tables is int instead of string

Tidiness issues

- Dog stage values in twitter_archive_data used as columns
- Image data columns for prediction should be renamed to be more descriptive
- tweet_id floating-point rounding issue.

Data Cleaning

After assessing the data, I proceeded to cleaning the issues identified in the assessment stage.

- First, the retweeted and replied columns and rows were dropped using pandas drop method as they were not needed in the data analysis.
- I converted the timestamp column to datetime datatype using pandas datetime function.
- None values in the dog stages columns were replaced with empty values.
- Using regex, I created numerator and denominator functions to extract the correct numerator and denominator numbers. Both functions were applied to the twitter_archive_date text columns. Extracted values were places in new numerator and denominator columns.
- The next cleaning step involved extracted wrong dog names and stop words. I created a for loop procedure for this and a replace method to replace these extracted wrong names with “No name” for easy identification.
- Using the string capitalize method, the first letter of the p1, p2 and p3 columns were capitalized.
- Tweet_id columns in the three tables were converted from int to string using the pandas astype method. This is will help in resolving one of the issues seen in the data assessment stage
- Next, I concatenated the dog stage columns into one column so as to clean up the tidiness issue
- Unclear column names in the image_data table was renamed to provide clarity.
- To resolve the tweet_id floating-point rounding issue, I created and applied a lambda function on the tweet_id column on all the tables to help in the merging

Storing Data

In the final step of my wrangling process, I stored the cleaned data in one table by merging all the table on the tweet_id column, thereafter, I proceeded to exporting this master data table to a csv file which will read using pandas to kick start the visualization stage.

Analysing and Visualizing Data

The next step in the second project of the data analyst nanodegree program involve analyzing and visualizing the details in the stored master data file extracted and cleaned as seen in the previous report.

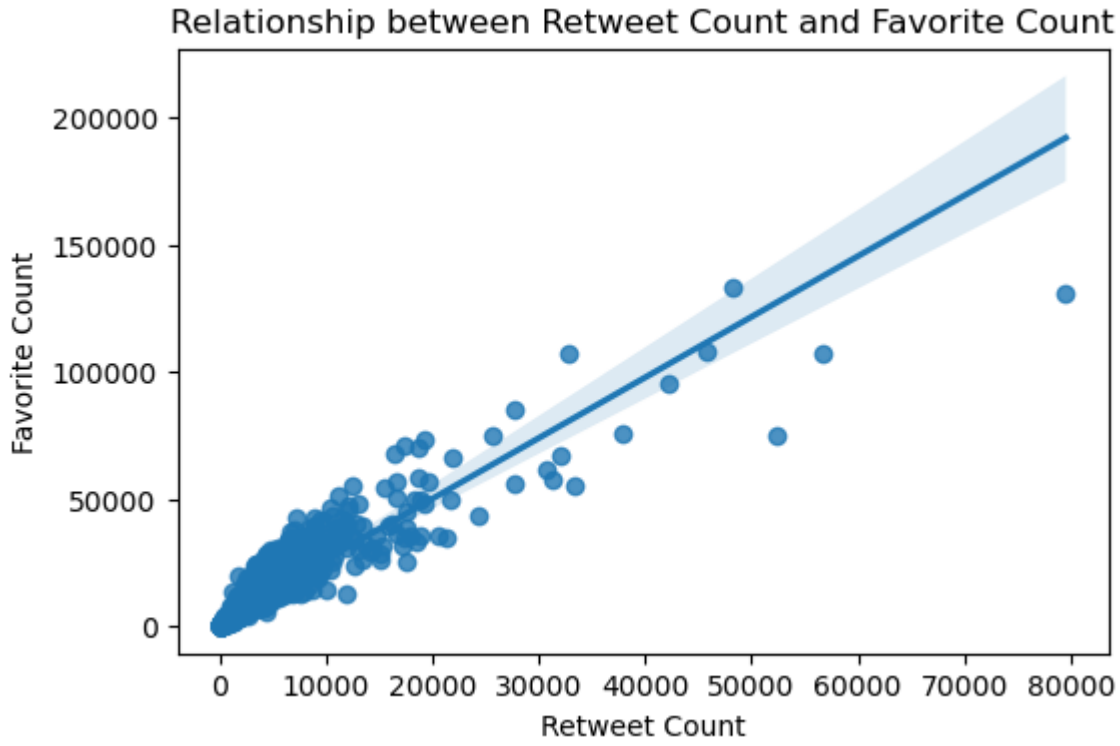
Importing/Reading the Data: In this step, I used the pandas read_csv to read the twitter master data. After which, columns not needed in the analysis phase were dropped.

As detailed in previous learning module, one of the first step in analysis process is to ask some relevant questions regarding the data. This will serve as a guide finding insights about the data. In view of this, the following questions were asked.

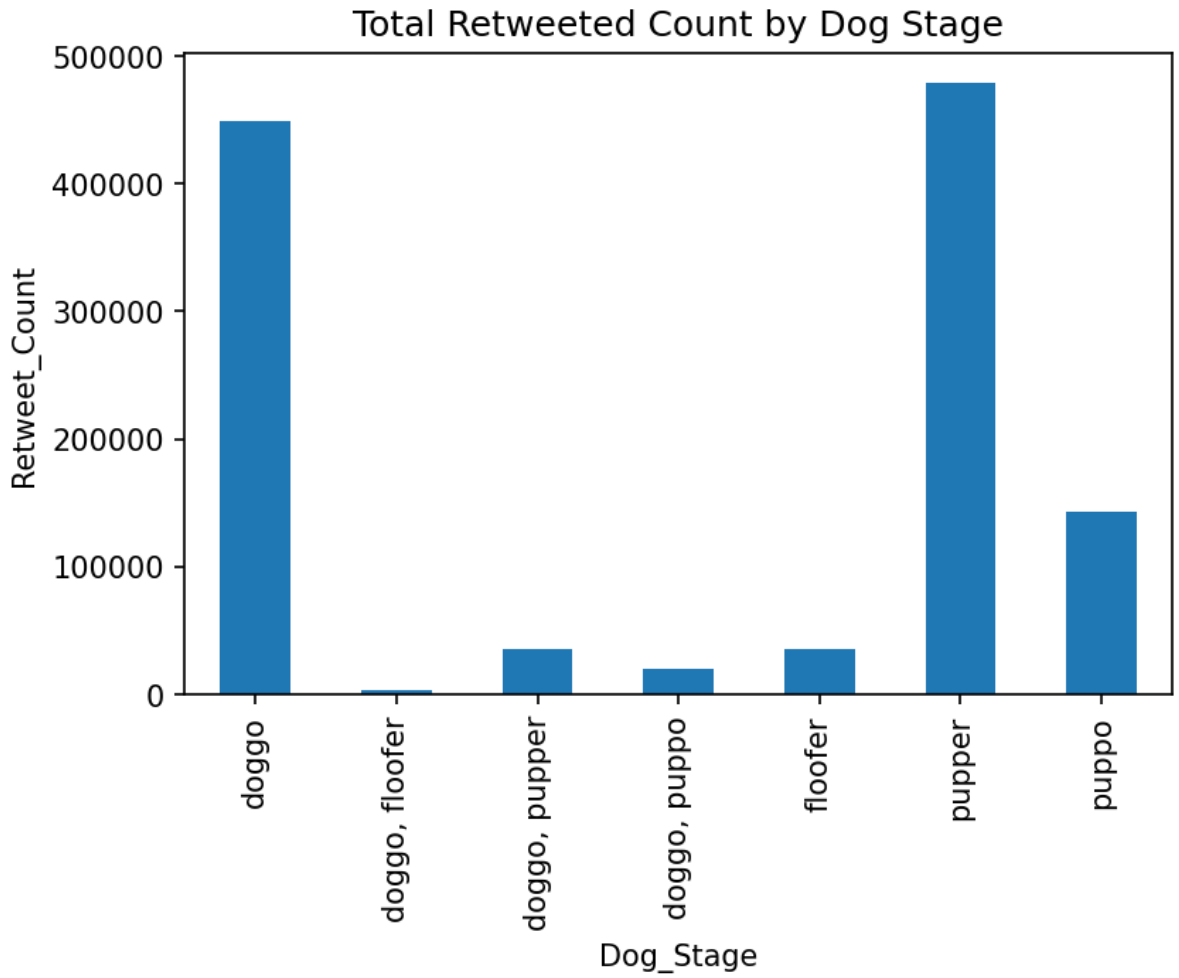
1: What is the relationship between retweet count and favorite count?
2: What is most retweeted dog stage?
3: Which dog stage is the most liked?
4: What is the most frequent dog stage?

5: Top ten is dog breed with the highest numerator rating on every prediction
6: Most used dog names?
7: Most predicted dog breeds across all the predictions
Before finding answers to these questions, I did an exploratory analysis to have a quick understanding of the data.

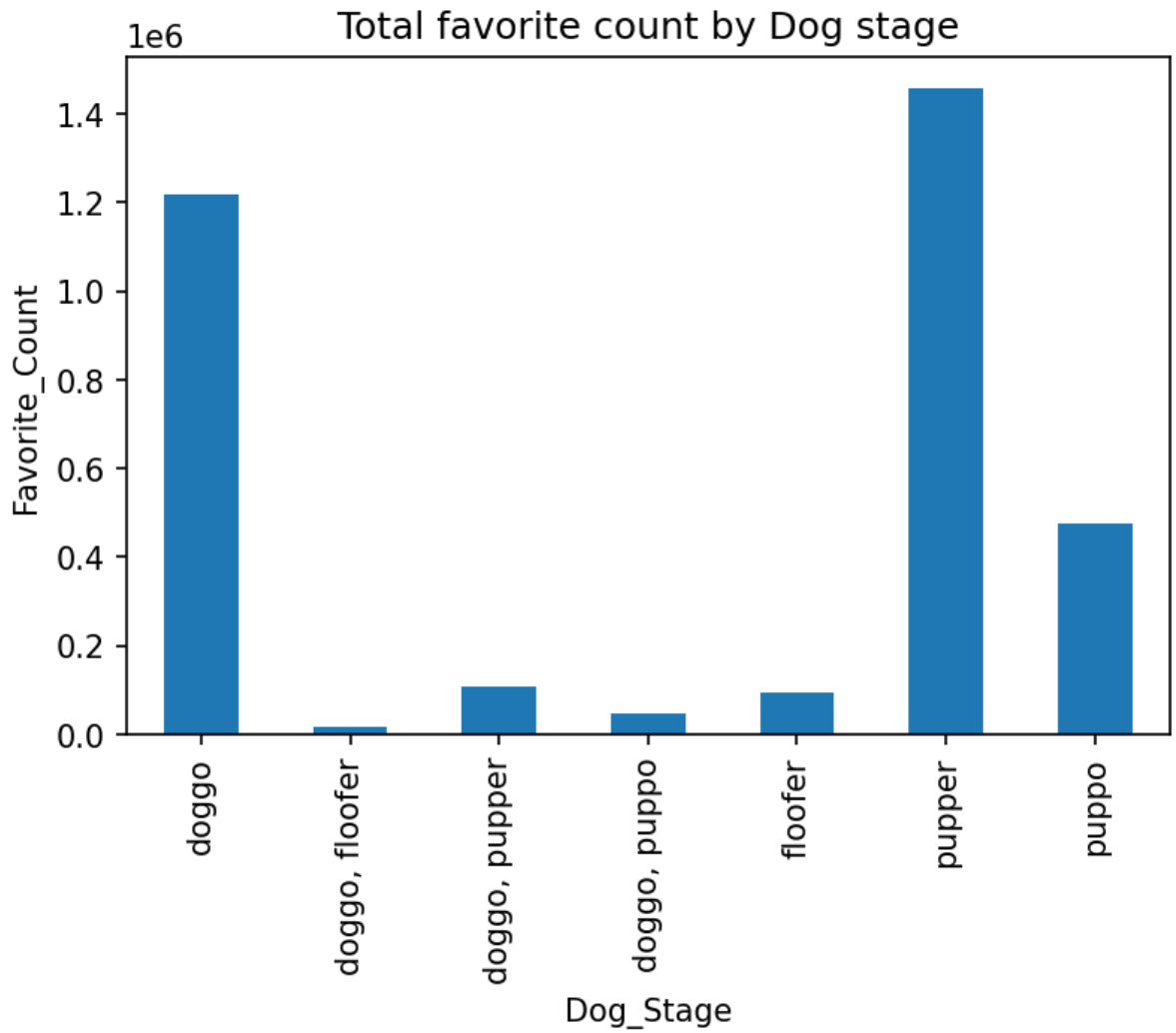
1: What is the relationship between retweet count and favorite count? Using seaborn library, I did a regplot visualization to see the correlation between these variables. From the below figure, there is a positive correlation between retweet count and favorite count. It can be inferred that the more a dog image is retweeted, the greater number of likes it get.



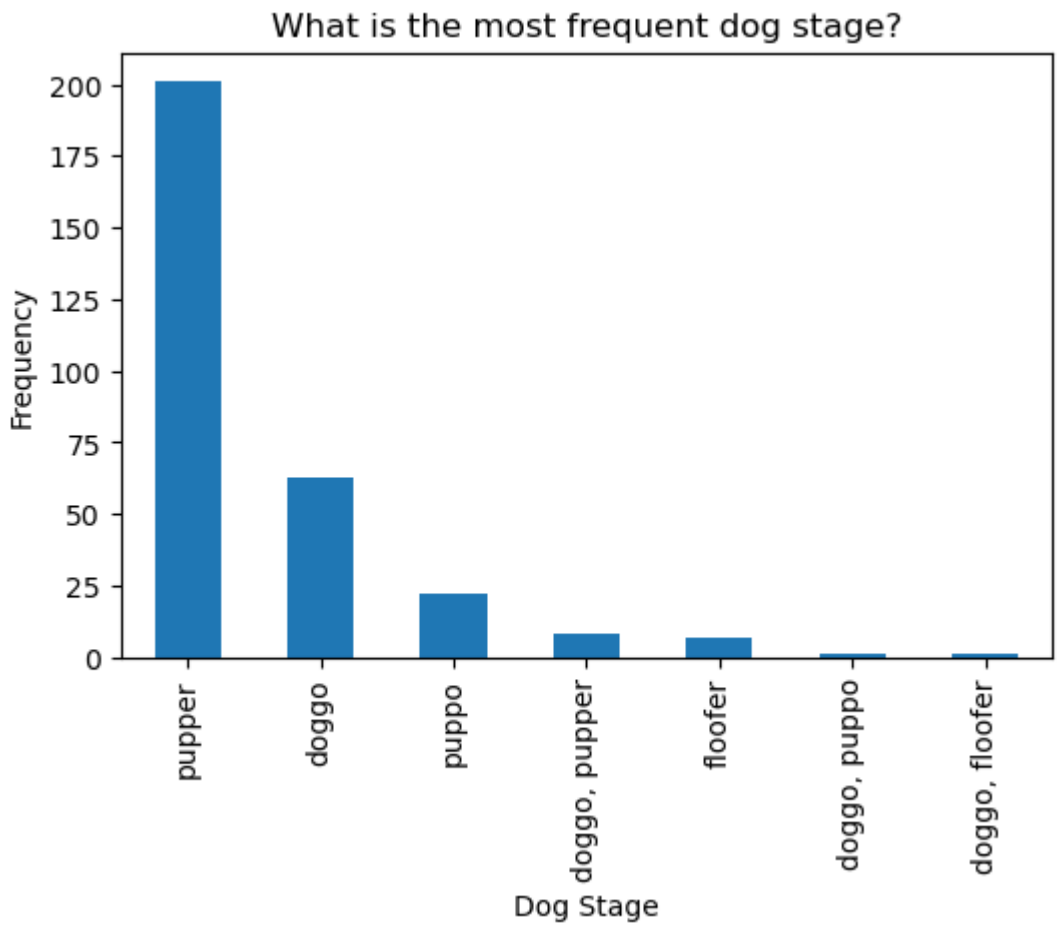
2: What is most retweeted dog stage? The master file was read and grouped by dog stage column, followed by the adding up the retweet_count variable on each dog stage. Given the figure below, pupper and doggo dog stages appear to be the most retweeted dog stage.



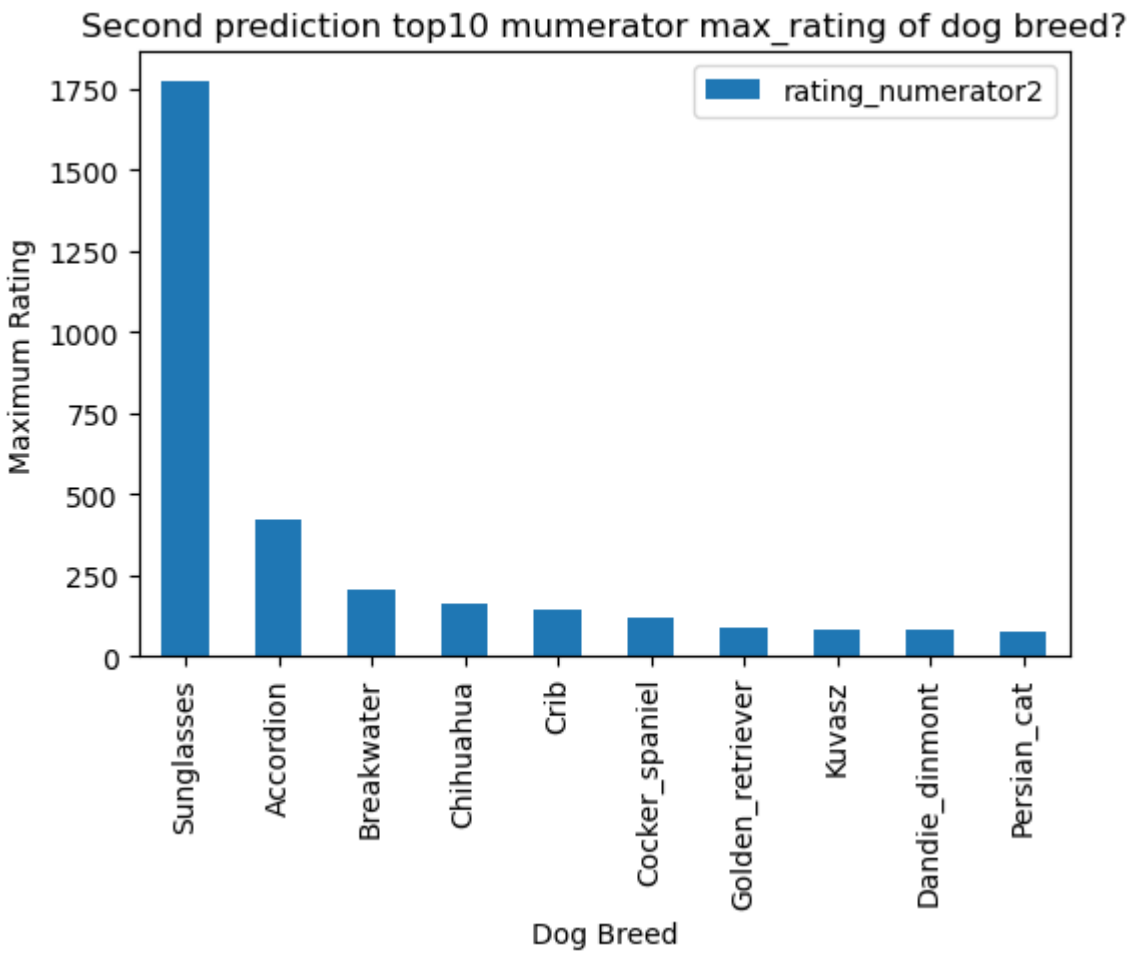
3: Which dog stage is the most liked? The master file was read and grouped by dog stage column, followed by the adding up the favorite_count variable on each dog stage. Given the figure below, pupper and doggo dog stages appear to be the most liked dog stage. This is in line with the insight gotten from the first visualization figure.

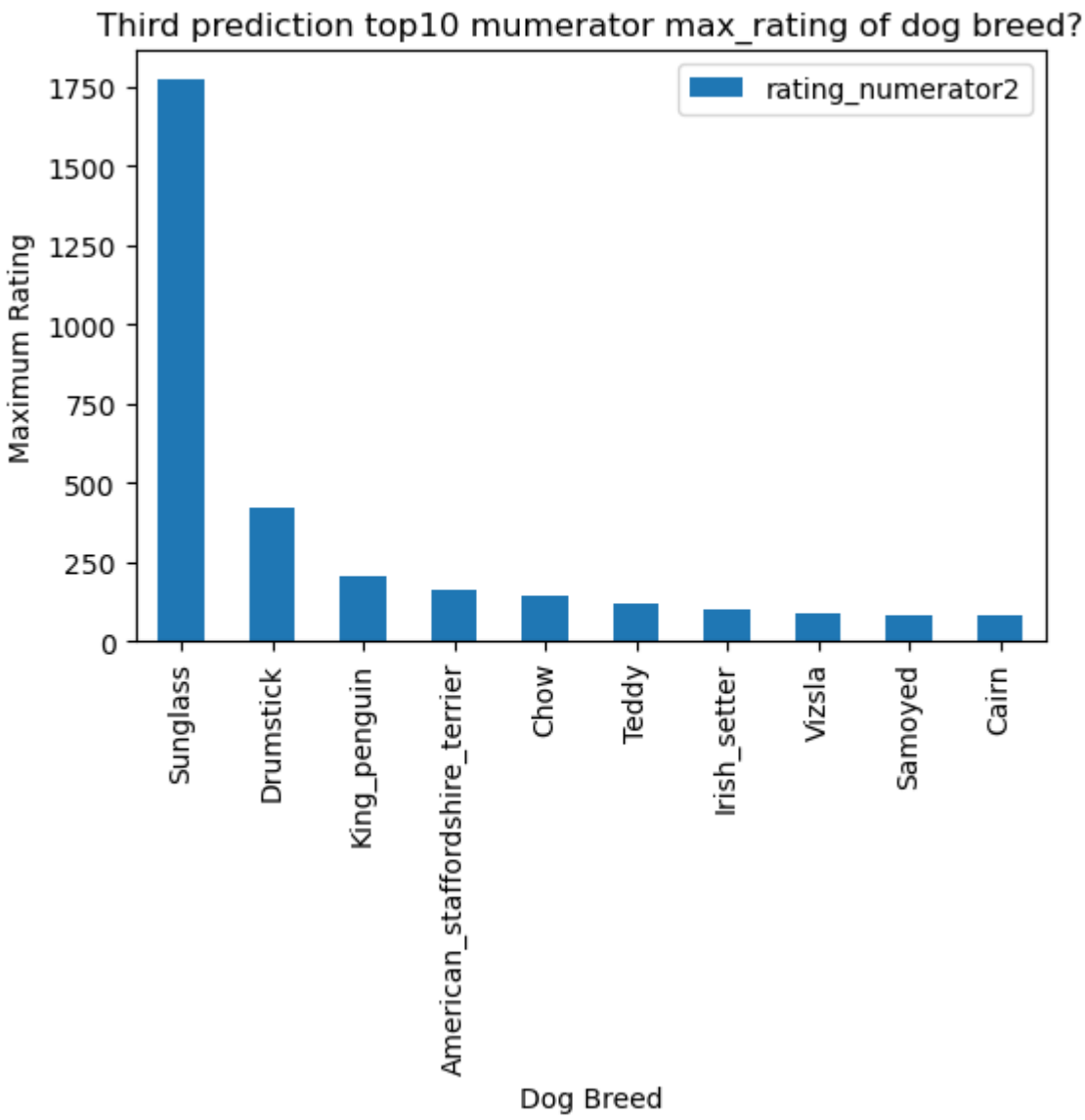


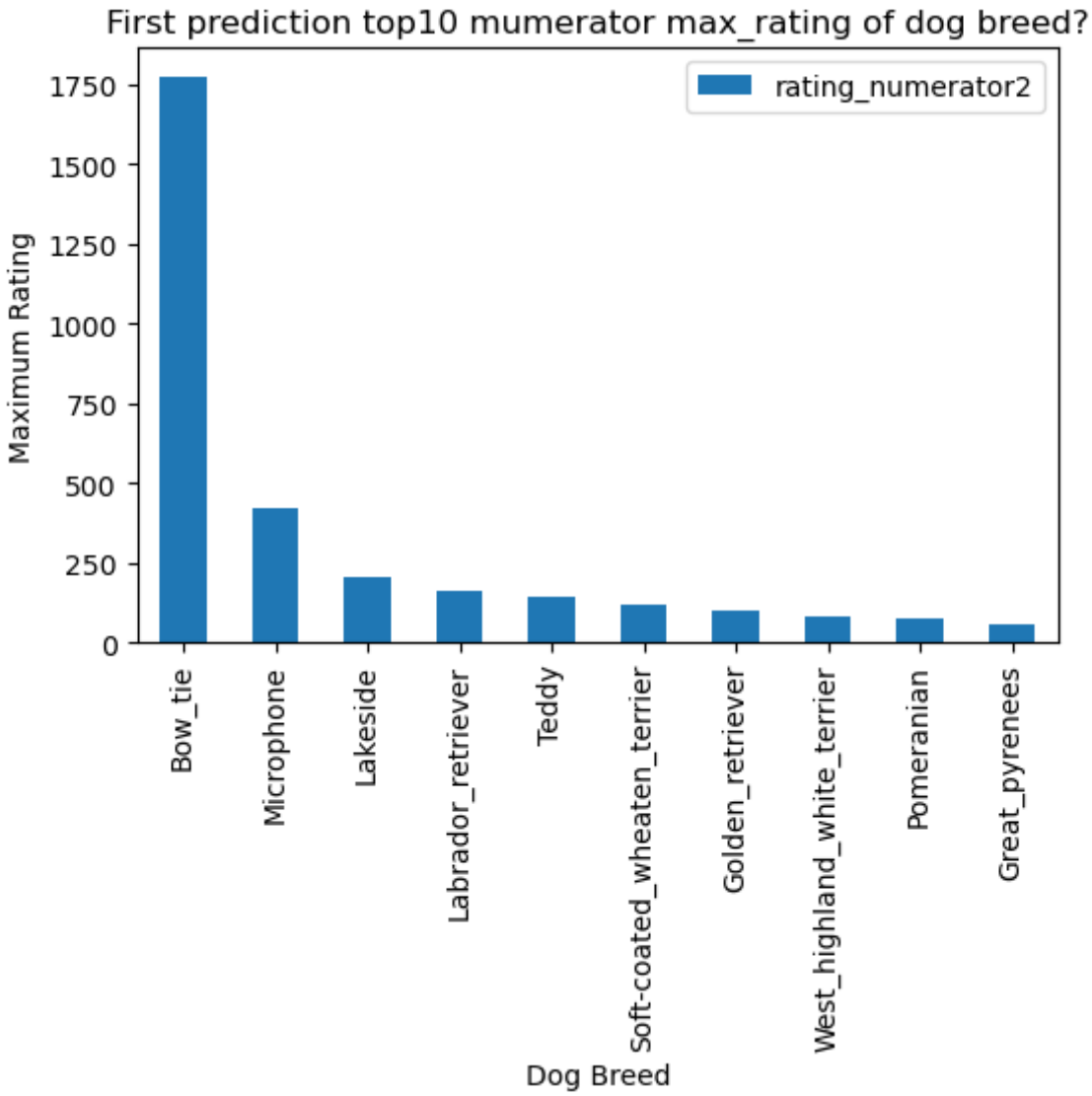
4: What is the most frequent dog stage? Among the dog stages, which of them appeared more frequently? The none values in the dog_stage column was excluded, after which, I plotted a bar chat of the value counts of each dog stage. Pupper appears to be the most frequent dog stage in the dataset.



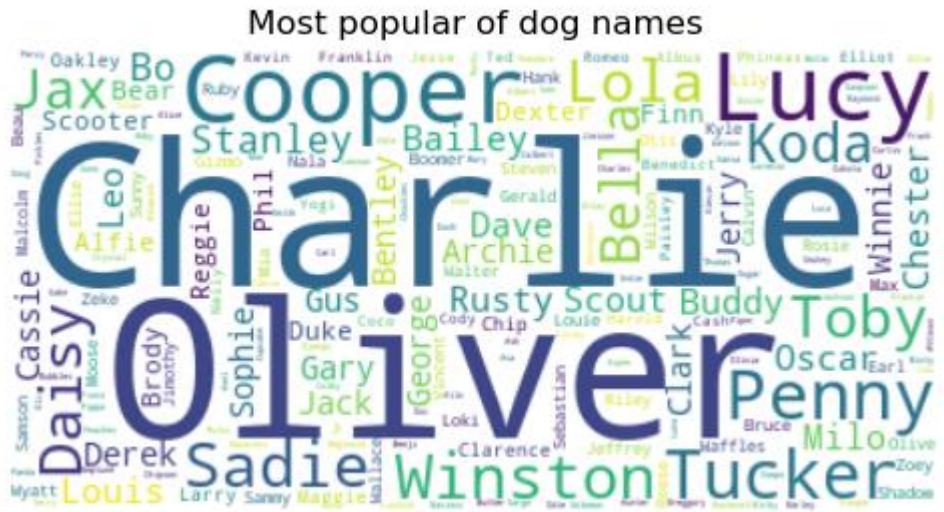
5: Top ten is dog breed with the highest numerator rating on every prediction: On every predicted dog breed column, I grouped the data on img_predicted column, got the maximun numerator rating and sorted it. This is was saved in a variable, after which I plotted a bar chat to view the top10 breed with the maximum numerator rating.


Second prediction top10 mumerator max_rating of dog breed?
rating_numerator2
Maximum Rating
1750
1500
1250
1000
750
500
250
0
Sunglasses
Accordion
Breakwater
Chihuahua
Crib
Cocker_spaniel
Golden_retriever
Kuvasz
Dandie_dinmont
Persian_cat
Dog Breed


Third prediction top10 mumerator max_rating of dog breed?
rating_numerator2
Maximum Rating
1750
1500
1250
1000
750
500
250
0
Sunglass
Drumstick
King_penguin
American_staffordshire_terrier
Chow
Teddy
Irish_setter
Vizsla
Samoyed
Cairn
Dog Breed



6: Most used dog names? All names in the name column was added to a name variable which was then passed to a word cloud generator to visualize the most used dog name in the dataset. From the below, it can be deduced Charlie and Oliver are the most used dog names in the dataset



7: Most predicted dog breeds across all the predictions: Each predicted dog breed was saved in a variable. The three variables were concatenated and then passed in the python word cloud library. Golden_retriever and Labrador_retriever appeared to be the most predicted dog breeds across all the predicted dog breeds in the dataset.