Act Report

WeRateDogs Project Analysis

1.1. Introduction

According to Wikipedia, *"Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions and supporting decision-making".* This definition captures the intriguing fact about what data analysis is all about.

Recently, the importance of analyzing our data has grown tremendously and many firms and individuals have taken hold of this opportunity in order to make significant decision on their daily activities ranging from being business specifics down to just finding interesting facts.

1.2. The Data Analysis process

The process that are involved in data analysis can be ideally organized into five (5) steps, which are:

- Asking relevant questions about a given data or merely just aroused to discover to a particular fact and gathering relevant data to provide some insights.

- The next step is wrangling a data and this process tends to be the most tedious of all, as it is argued to be the stage where much data professionals spend most of their time. It involves gathering, assessing and cleaning the dataset that will be used for analysis.

- EDA (Exploratory Data Analysis) is next after a successful data wrangling and here, our cleaned data are been explored and augmented to maximize the potential of our analysis, visualizations and models.

- Drawing conclusion is the next step and this could practically be approached with a machine learning or inferential statistics process. We can as well tend to proceed with carrying out descriptive statistics here to build up tangible insights.

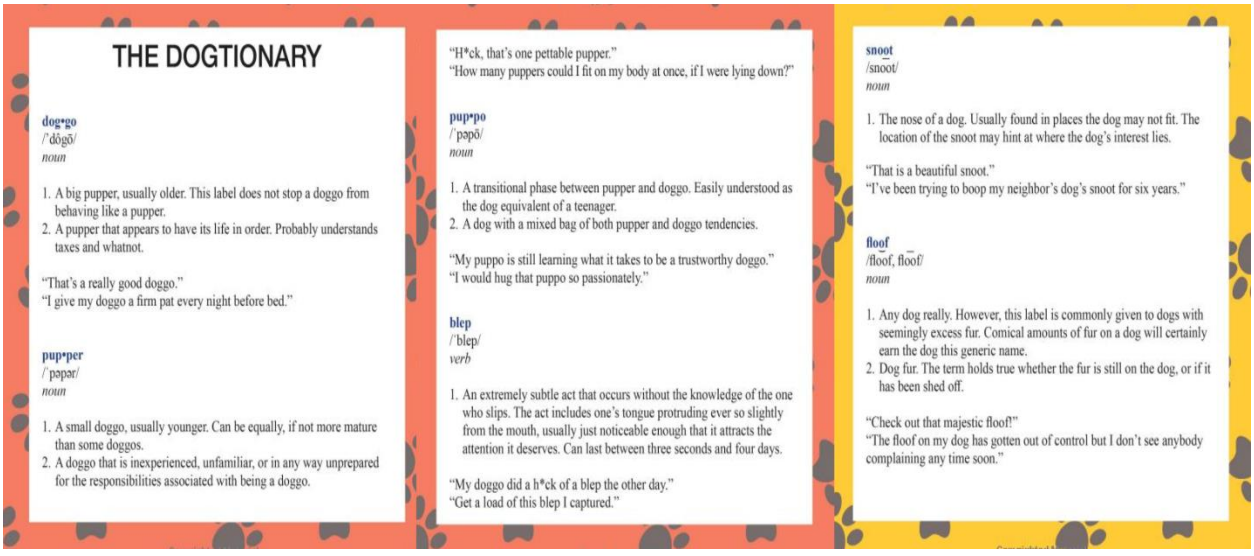
- Result communication is the final step to every data analysis process and it could be carried out using reports, slide decks, block posts, emails etc.

1.3. The WeRateDogs Project Analysis Report.

This project is majorly based on the data wrangling process and every dataset used was thoroughly cleaned to enable us to carry out analysis that led to drawing useful insights on the project.

in this project, the tweet archive of Twitter user @dog_rates, also known as WeRateDogs is being wrangled. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The ratings almost have a denominator of 10. The numerators caries values that are almost greater than 10 (11/10,12/10,13/10) etc. This is because "They're good dogs Brent."

The project introduced to us the different stages of dogs (i.e., doggo, floofer, pupper and puppo) as well as different types of dog breeds.



- **Doggo:** A doggo is a full-size pupper.
- **Floofer:** A very fluffy dog such as a Samoyed or Pomeranian.
- **Puppo:** A transitional phase between pupper and doggo.
- **Pupper:** A small doggo, usually younger, can be equally, if not more mature than some doggos.

Twitter APIs was queried to extract JSON data using specified tweet IDs that is needed for the project. From these tweets, different variables such as text, name, image_url, ratings, dog_breeds were extracted as well.

From this WeRateDogs tweet, we can see the name of the dog as well as the ratings given to it.

In our project analysis, we were able to get insights about dog_stage with the highest ratings, the dog_breeds that has the highest average tweet's-likes and so on.

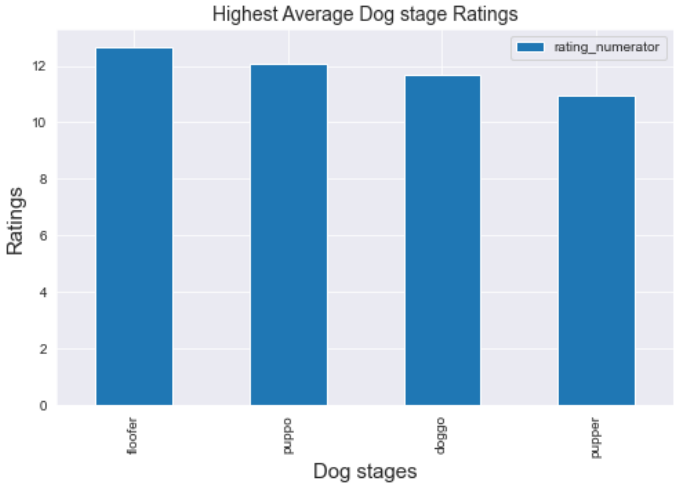
1.4. Analysis and Insights

The first insight we explored in our analysis process was to discover the dog_stage that has the highest average ratings.

Q1. What Dog stage has the highest average ratings?

```
In [106]: Dog_stage_rating = twitter_archive_clean.groupby(['dog_stages'])['rating_numerator'].mean().sort_values(ascending=False)
          Dog_stage_rating

Out[106]: dog_stages
          floofer    12.666667
          puppo      12.076923
          doggo      11.689655
          pupper     10.930533
          Name: rating_numerator, dtype: float64
```



From this analysis and our chart, it is evident that Floofer is the dog_stage with tweets that has the highest average ratings, having an approximate value of 12.67.

We went further to estimate the top-5 dog_breeds that has the tweet with the highest average likes (favourite_count)

Q2. What Dog breed has the highest average likes (favorite_count)? Top 5

```
In [109]: twitter_archive_clean.groupby(['dog_breed'])['favorite_count'].mean().astype(int).nlargest(n=5)

Out[109]: dog_breed
          Bedlington_terrier         21454
          Saluki                     20722
          French_bulldog             17464
          Afghan_hound               14652
          black-and-tan_coonhound    14602
          Name: favorite_count, dtype: int32
```

Tweets on **Bedlington terrier** Dog Breed has the highest average likes

Out of these fives dog breeds, Bedlington Terrier came out on top as the dog_breeds with the highest average tweet likes

"*An Interesting fact about a Bedlington Terrier is, today Bedlington(s) excel as companions in the show ring*".

The popular names given to dogs were also analyzed and we were able figure out the top ten among these names.

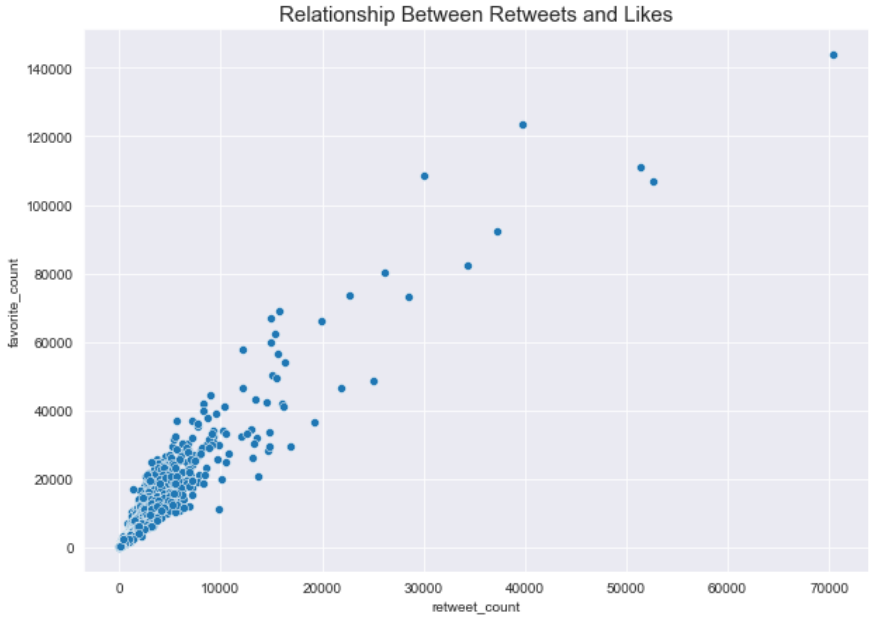
Q3. What is the top 10 most popular Dog names?

```
In [110]: twitter_archive_clean['name'].value_counts().nlargest(n=10)

Out[110]: Cooper     10
          Oliver      9
          Charlie     9
          Tucker      9
          Penny       8
          Lucy        8
          Daisy       7
          Winston     7
          Sadie       7
          Toby        6
          Name: name, dtype: int64
```

According to our analysis, Cooper is the name given to most dogs, follow by Oliver and so on.

One more insight that our analysis revealed, was the relationship between every tweet likes and retweets. In finding out about this, we performed a correlation analysis between the favorite_count and retweet_count variables.



There is a strong positive correlation between the likes and retweets of the overall tweets with a *correlation value* of **(0.9290594621813724).**

We also created a WordCloud visualization of every dog breed that appeared in the tweets used for the project.

1.5. Conclusion

Data analysis is a very essential technical skill and it has been optimally utilized in our project, revealing interesting insights from our data and allowing us to make significant interpretations.