Act Report

Introduction and Background

Real-world data rarely comes clean. Using Python and its libraries, I gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then I cleaned it. This is called data wrangling. Here is the documentation of my act effort which was done in a Jupyter Notebook with markdown and then downloaded as *pdf* file to showcase the analyses and visualizations using Python (and its libraries).

The dataset that was wrangled, analyzed and visualized 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 downloaded their Twitter archive and sent it to Udacity via email exclusively for you to 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. Here's an example:

This project works through the data wrangling process, which are sub divided into three tasks called Gather, Assess and Clean. Then follows interesting and trustworthy analyses and visualizations

Necessary Libraries

The following packages (libraries) were used:

- pandas
- NumPy
- requests
- tweepy
- json

Data Wrangling

Gather

In this project, I gathered data from the following sources:

- The WeRateDogs Twitter archive. The *twitter_archive_enhanced.csv* file was provided to Udacity students. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to 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.
- The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was was downloaded programatically using requests. get() function from requests which was then saved in a tsv file named *image_predictions.tsv*.
- Twitter API by creating a Twitter developer account to generate my consumer key, consumer secret, access token and access secret. Also, Python's Tweepy library was used to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting which was extracted into the *tweet_json.txt* file.

Assess

Assessing data requires data analysts to evaluate a data set on quality and tidiness issues. The four (4) main data quality dimensions are:

- Completeness
- Validity
- Accuracy
- Consistency

And there are three (3) requirements for tidiness:

- Each variable forms a column

- Each observation forms a row
- Each type of observational unit forms a table

At least eight (8) quality issues and two (2) tidiness issues were detected and documented appropriately

Clean

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, or duplicated. Having bad quality data can be disastrous to your processes and analysis. The cleaning process involves three steps:

- **Define:** Determine exactly what needs to be cleaned, and how
- **Code:** Programmatically clean with code
- **Test:** Evaluate the code to ensure the dataset was properly cleaned

This method was used to clean all issues that were spotted in the assessment part

Data Analyses and Visualizations

Data analysis is a process for obtaining raw data, and subsequently converting it into information useful for decision-making by users.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization helps to see and understand trends, outliers, and patterns or insights in data. In this project, the clean dataset is stored in a master file named *twitter_archive_master.csv* file before the analysis. At least three (3) insights and one (1) visualization are provided

Drawing conclusions and creating visuals to communicate results. The following questions were addressed:

Q1: What are the features that influence retweet count and favorite count?

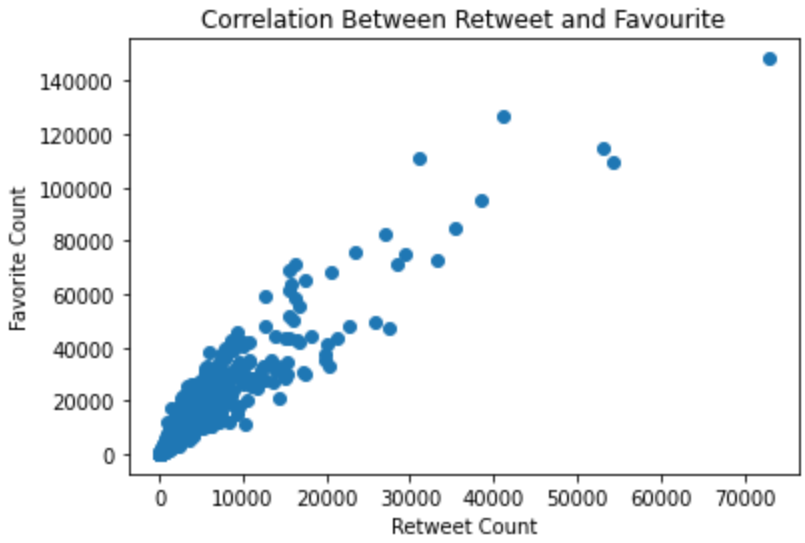
Q2: Is rating influenced by dog stage? What are the dog stages with the highest rating?

Q3: What is the most popular dog name?

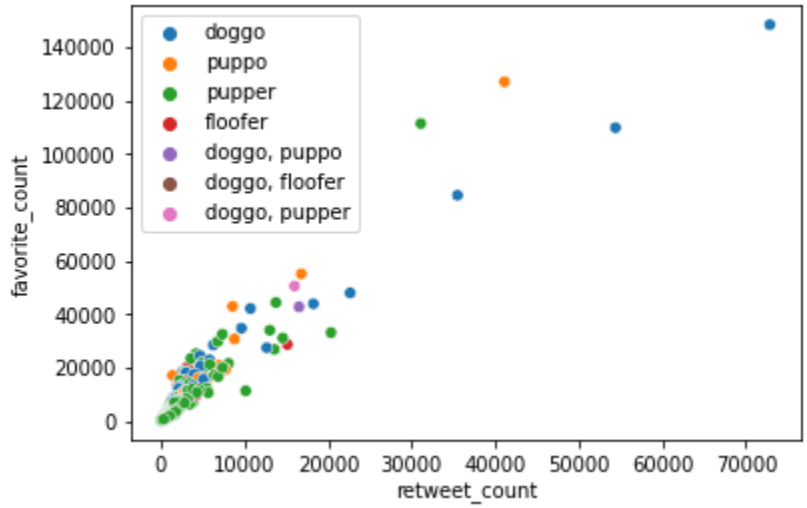
Insights and Visualizations

Q1: What are the features that influence retweet count and favorite count?

There is a positive correlation between retweet count and favorite count. This correlation helps the owner of WeRateDogs Twitter account to understand posts that generate user traffic on their page in order to improve and recommend future contents.

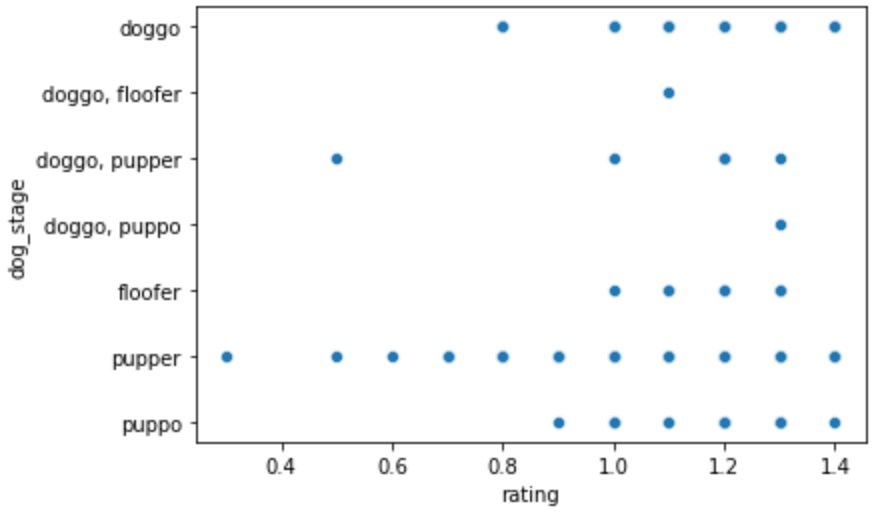


Based on the visual below, we can see that there is a positive correlation between retweet count and favorite count across the dog stages, with **doggo** having the strongest positive correlation. This is also important as it gives clue to the stage of dog that generate more user traffic on their page and could be used to tailor future trends



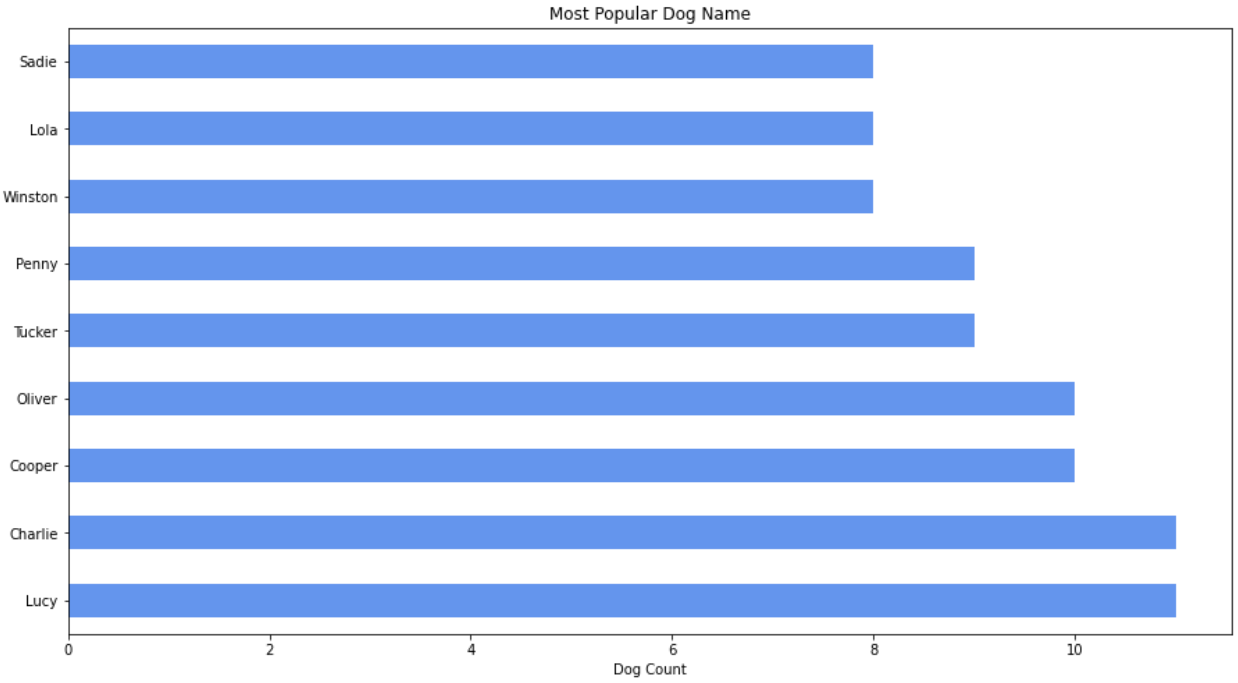
Therefore, variables like dog stage influence retweet count and favorite count.

Q2: Is rating influenced by dog stage? What are the dog stages with the highest rating?

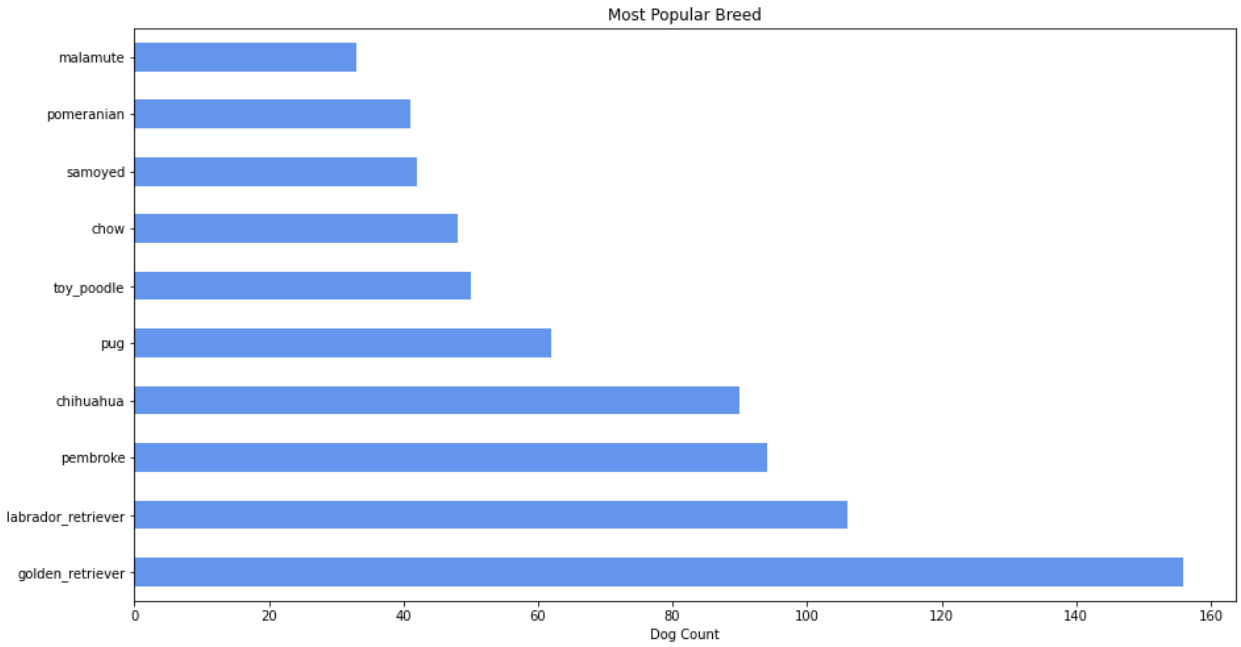


Doggo, pupper and puppo are the three dog stages with the highest rating. The account owner could use this information to create a target campaign for the stage with the lower rating to attract more users that may own dogs in any of these stages and also could decide to focus more on creating contents that will generate more traffic based on this insight.

Q3: What are the most popular dog names and dog breed?



From this visual, the most popular dog names are Lucy and Charlie, followed by Cooper and Oliver.



The Most Popular breeds are Golden Retriever, Labrador Retriever and Pembroke in that order. The owner of WeRateDogs Twitter account could use this information to create target campaign for breeds that have low popularity to increase their popularity, but also utilize the breeds that are proven to be popular to drive user traffic to the page.

Conclusion

This dataset provides interesting insights and visualizations which can be further explored with some additional data like information of dog owners, their locations, their culture etc., to recommend more value-added services that WeRateDogs can render to dog owners and pet stores around the world.