WeRateDogs Analysis

WeRateDogs is a Twitter account that rates people's dogs with a humorous content about the dog. This is a Udacity Data Analyst Nanodegree project that enriches one with the knowledge of data wrangling and analysis.

In the real world, data rarely comes clean. Hence, the need to undergo the data analysis processes. These processes can be iterative depending on how "messy" the data is. For WeRateDogs, the wrangling and analysis was done in four stages: Gathering data, Assessing data, Cleaning data, and Analyzing and Visualizing data.

Gathering Data

The data was gathered from disparate sources in different formats. The sources are the WeRateDogs Twitter Archive file that was given to Udacity by WeRateDogs which was stored as a csv (comma-separated values) file. This file was then read using the pandas library. The data contained the names of the dogs, timestamp of tweets, dog ratings etc. Another source was the image prediction file hosted on Udacity's servers in a tsv (tab-separated values) file that was downloaded using the requests library. This file was read using pandas. The file contained the name of the predicted images and their confidence levels. The third source was Twitter's API to query additional data. This was queried using the tweepy library to get each tweet's json data on the tweet's id, favorite count and retweet count.

Assessing Data

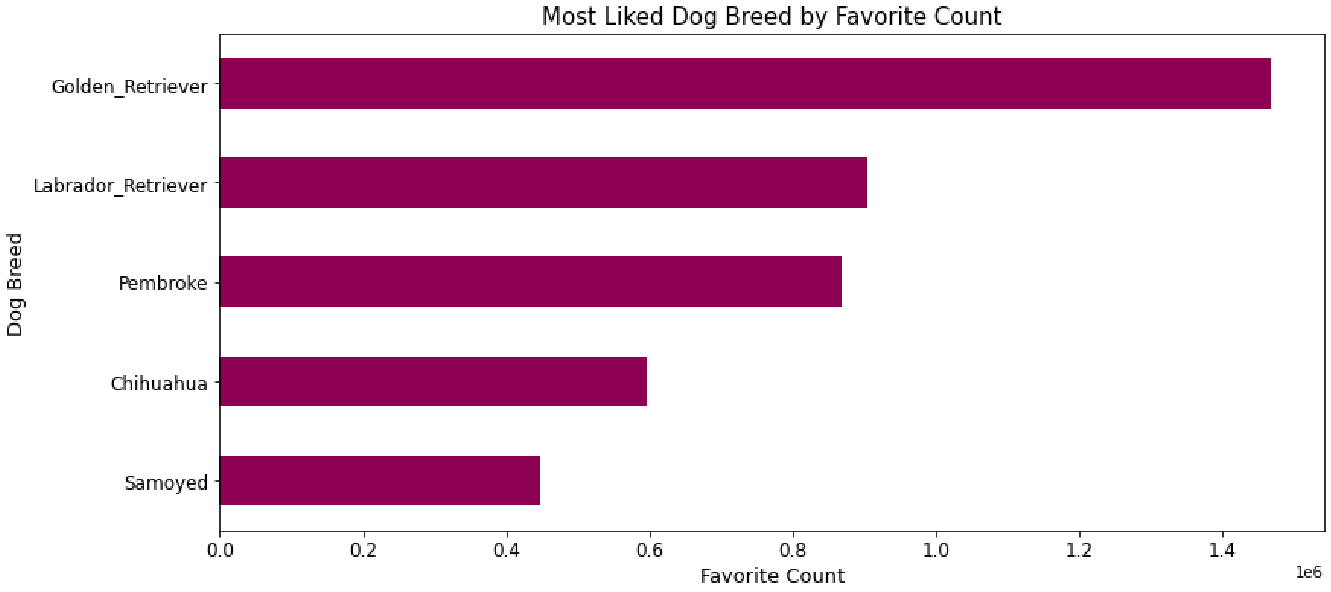
The data was assessed both visually and programmatically. A spreadsheet tool was used to visually assess the data. The programmatic assessment was done by writing a few lines of code to explore the data. This led to the documentation of eight (8) data quality issues and two (2) tidiness issues.

Cleaning Data

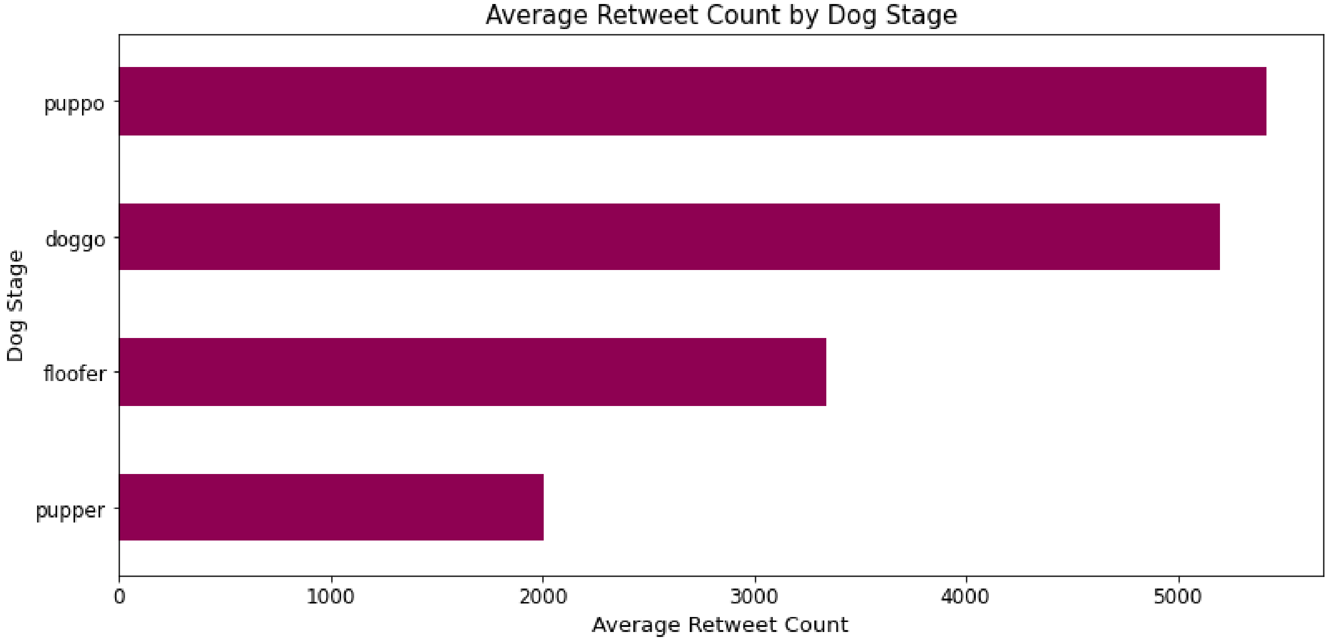
Each documented quality and tidiness issue was cleaned using the **define**, **code**, and **test** framework. The identified issues were defined into action statements of the cleaning process, which was translated into codes to resolve the issues and tested to ensure the codes worked and the cleaning operation was achieved. Once the entire dataset was cleaned, and the three files were merged as a single document, it was then stored as a single fle to be used for analysis and visualization.

Analyzing and Visualizing Data

After cleaning the data and analyzing the data. A few insights were generated. The most popular ratings given to dogs is **12**. The top four(4) most common names given to dogs are **Lucy**, **Charlie**, **Oliver**, and **Cooper**. The **Golden-Retriever** breed is the most liked dog breed by the sum of favorite count.



From the data, it can be observed that the month of **June** is the month with the highest average number of retweets. The **Puppo** and **Doggo** stage are the most popular dog stage based on the average retweet count, having an average retweet count of over 5000 retweets.



In conclusion, data wrangling and analysis is an important process of data analysis. This process enables data analysts to gather their data from different sources and in different formats for analysis.