WeRateDogs Twitter Archive - Act Report

By Elliot M Sithole

06/09/22

Synopsis

The objective of this project was to offer a real case on how to wrangle data; that is gathering, assessing, cleaning for improved quality and tidiness and then analysing data, iteratively. The data used was sourced from three pieces on Twitter database; WeRateDogs. First, an enhanced twitter archive which contains extracted data from 5000+ tweets @thedog_rate twitter account (November 15, 2015, to August 1, 2017). The second, was gathering additional data through utilizing the tweet IDs in the WeRateDogs Twitter archive, to query the Twitter API for each tweet's JSON data using Python's *Tweepy library,* by downloading to incorporate re-tweet count and favourite count which was excluded from the original archive file. The third was programmatically downloading an image predictions file from Udacity servers containing the top three image predictions for dog breeds based on the images from tweets.

After data gathering, I had to use both visual and programmatic assessments to assess the data and detect and document eight quality and five tidiness issues. Through an iterative process of assessing and cleaning, I managed to clean my datasets, merged and stored them into one file named *twitter_archive_master.csv.*

In the data analysis step I managed to pose a few questions as a guide to discover important insights:

- What are the top five common dog names?
- Which dog breeds are the most rated from the data set?
- Which dog breed has the most average favourite count?
- Is there any relationship between dog re-tweet counts and favourite counts?

1. Introduction

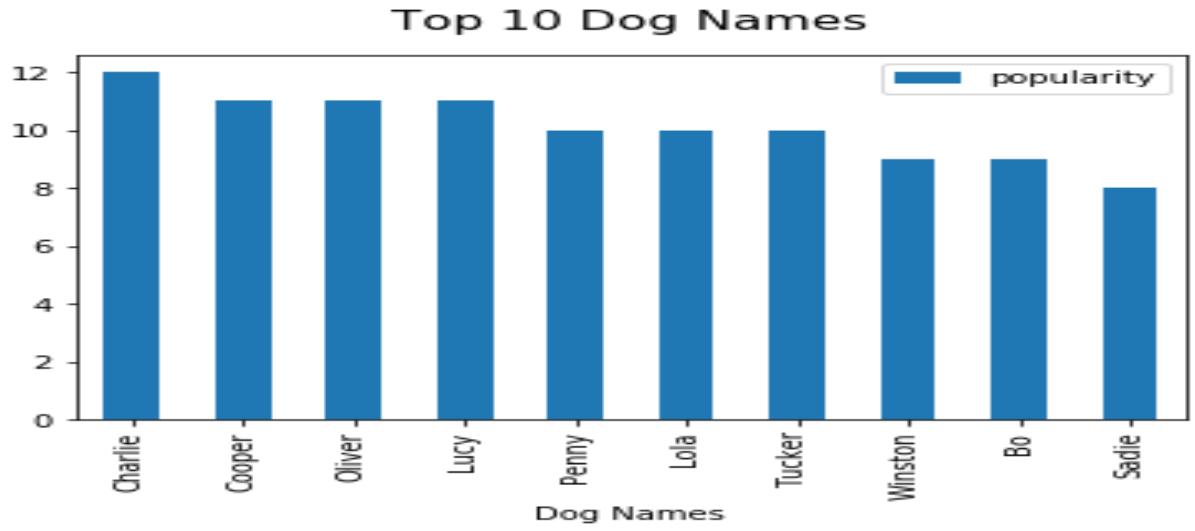
This report is an additional complement to the insights and trends observed on wrangle_act.ipynb notebook file for data wrangling and analysis on the Udacity Data Science Nanodegree project. The most exciting element of the WeRateDogs tweets is the humorous and non-standard rating system, with the numerator almost always exceeding the denominator at 10.

2. Exploratory Data Analysis

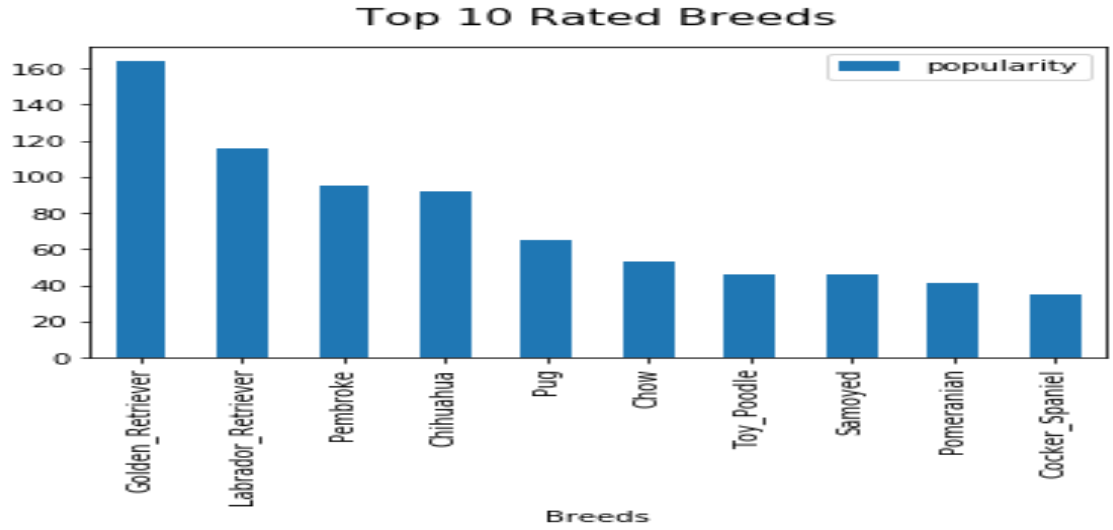
After a thorough visual and programmatic assessment, several issues were documented, about quality and tidiness shortfalls. The final cleaned file was saved as master dataset to a CSV file named "twitter_archive_master.csv" ready for further analysis and visualizations.

2.1 Insights

1. The top 5 most common dog names as observed in Figure1 were; Charlie, Oliver, Cooper, Lucy and Lola. However, I should emphasize, that most dogs' names were not mentioned, hence this analysis may not be representative of the facts at hand.



2. Golden_Retriever and Labrador_Retriever are the top two highly rated breeds, check *Figure 2* for visual presentation among the top 10 breeds.
   Figure 2: Top ten rated breeds.



3. The Saluki breed has the most average favourite count, although one, the Lakeland Terrier breed had a maximum favourite count from the data set. See Figures 3 and 4 respectively.

   Figure: 3

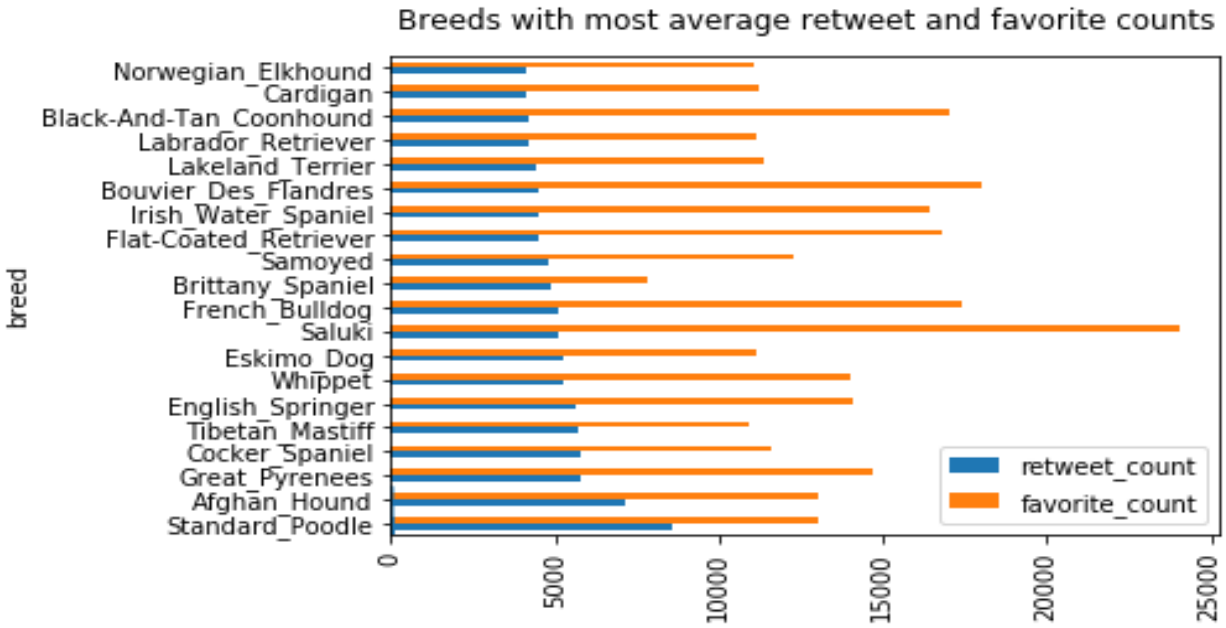
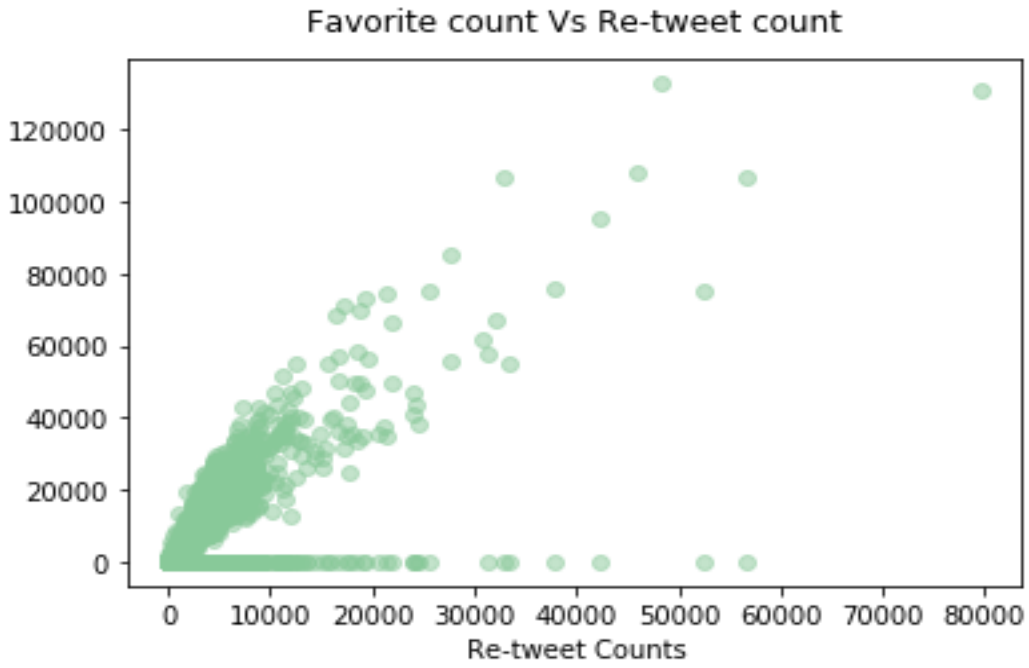


Figure 4: The most liked dog, a Lakeland Terrier breed.

4. The Standard Poodle breed has on average more re-tweet counts, with one the Labrador Retriever having the maximum re-tweet counts. Figure 5

Figure 5: The most re-tweeted dog; a Labrador Retriever breed

5. There is a general positive correlation between dog re-tweet counts and favourite counts.
   Figure 6: Presents a straightforward scatterplot graphic of favourite-count vs re-tweet count to visualize any pattern.



Conclusions

This project aimed to perform data wrangling and analysis for the twitter WeRateDodog account. The steps taken included a rigorous iterative process from data gathering, assessing, cleaning and finally exploratory data analysis. All data manipulations were handled on Jupyter Notebook, utilizing python and its libraries, pandas, numpy, matplotlib, requests, tweepy and json features. Based on the data gathered I assessed a few evident issues, which included addressing the data shortfalls in quality and tidiness, that required a thorough manipulation to obtain an enhanced cleaned data frame (twitter_archive_master.csv). After fixing 8 quality issues and 5 tidiness issues I managed to do an exploratory data analysis that reviewed important insights including; highlighting the top 5 most common dog names as Charlie, Oliver, Cooper, Lucy, Lola, the breeds Golden_Retriever and Labrador_Retriever as the top two highly rated, also the Saluki breed had the most average favourite count, hence most likes on twitter albert the Lakeland_Terrier had a maximum favourite count. The Standard Poodle breed had on average more re-tweet counts, with the Labrador_Retriever with the maximum re-tweet counts. And finally, I discovered a general positive correlation between dog re-tweet counts and favourite counts.