STORE, ANALYZE & VISUALIZE DATA REPORT

Author: IRAKLIS KALAMAS

Introduction

Quoting from project overview page:

”Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. The dataset that you will be wrangling (and analyzing and visualizing) 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.”

This project works through the data wrangling process, i.e. gathering, assessing and cleaning of data. Then analysis with corresponding visualization and observation results are documented as well. This report focuses on the latter part of the whole process.

- For a detailed documentation of how the datasets, used for the analysis, has been produced, please refer to “wrangle_report.pdf”.

Storing

Having the three datasets (“twitter-archive-enhanced.csv”, “image_predictions.csv” and “tweet_json.txt”) ready, I merged them to produce that final dataset for analysis: “twitter_archive_master.csv”.

Analysis & Visualization

In short my analysis includes the following:

1. Visualizing the most popular dog breed
2. Visualizing the most popular dog names
3. Visualizing the total number of tweets over time to see whether that number increases, or decreases, over time
4. Visualizing the retweet counts, and favorite counts comparison over time

1. Visualizing the most popular dog breed

The most popular dog breed is golden retriever (image 1), with a Labrador retriever coming in as the second most popular breed. The results could be used for marketing purposes or to drive user traffic to the pages with the least rated dog breeds in order to enhance their popularity.

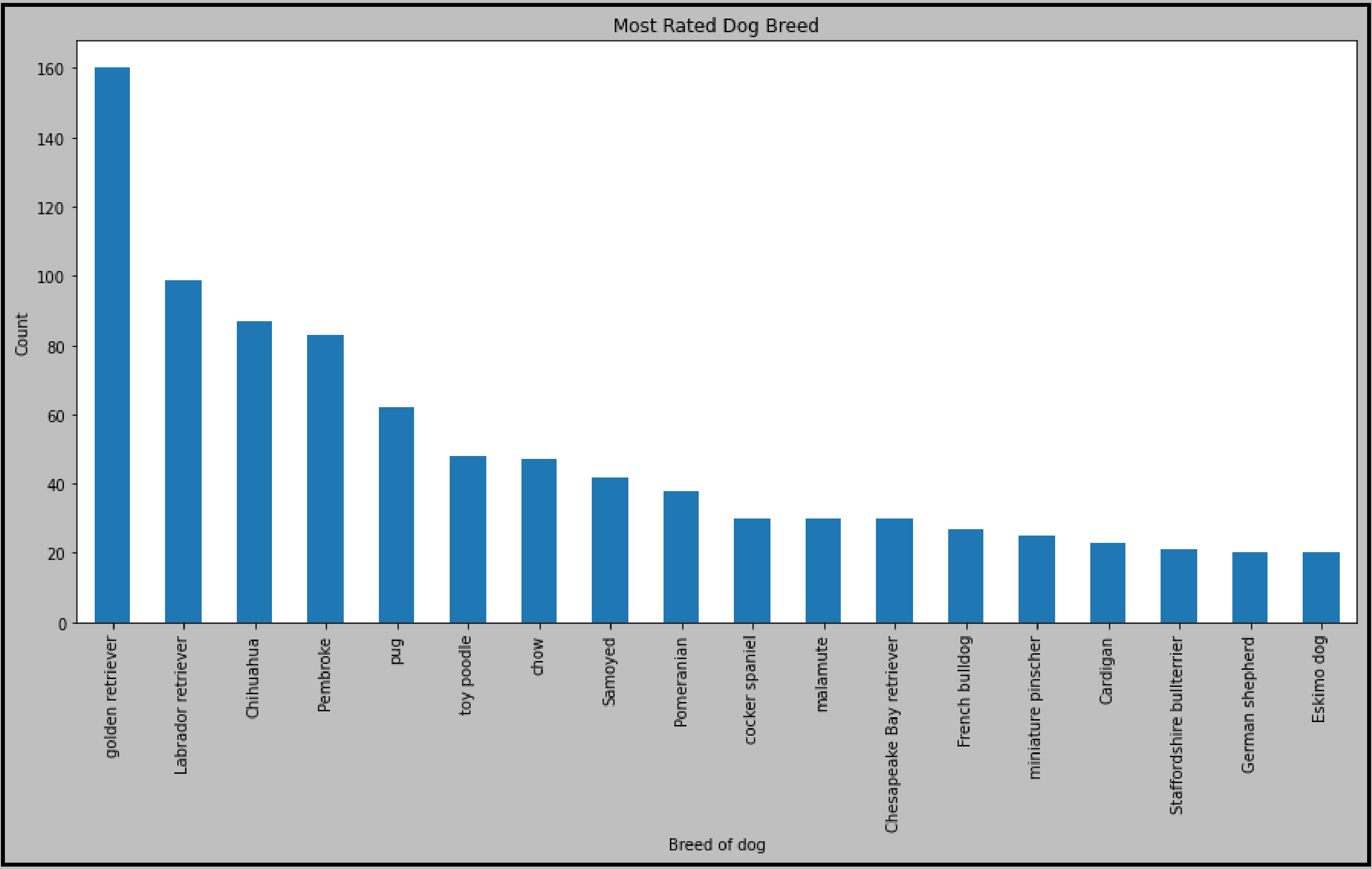

image 1

An image of the “winner”

2. Visualizing the most popular dog names

The most common dog names are Lucy & Charlie with Oliver & Cooper as runner-ups as the following image show. This information can be used only for marketing and promotional cases.

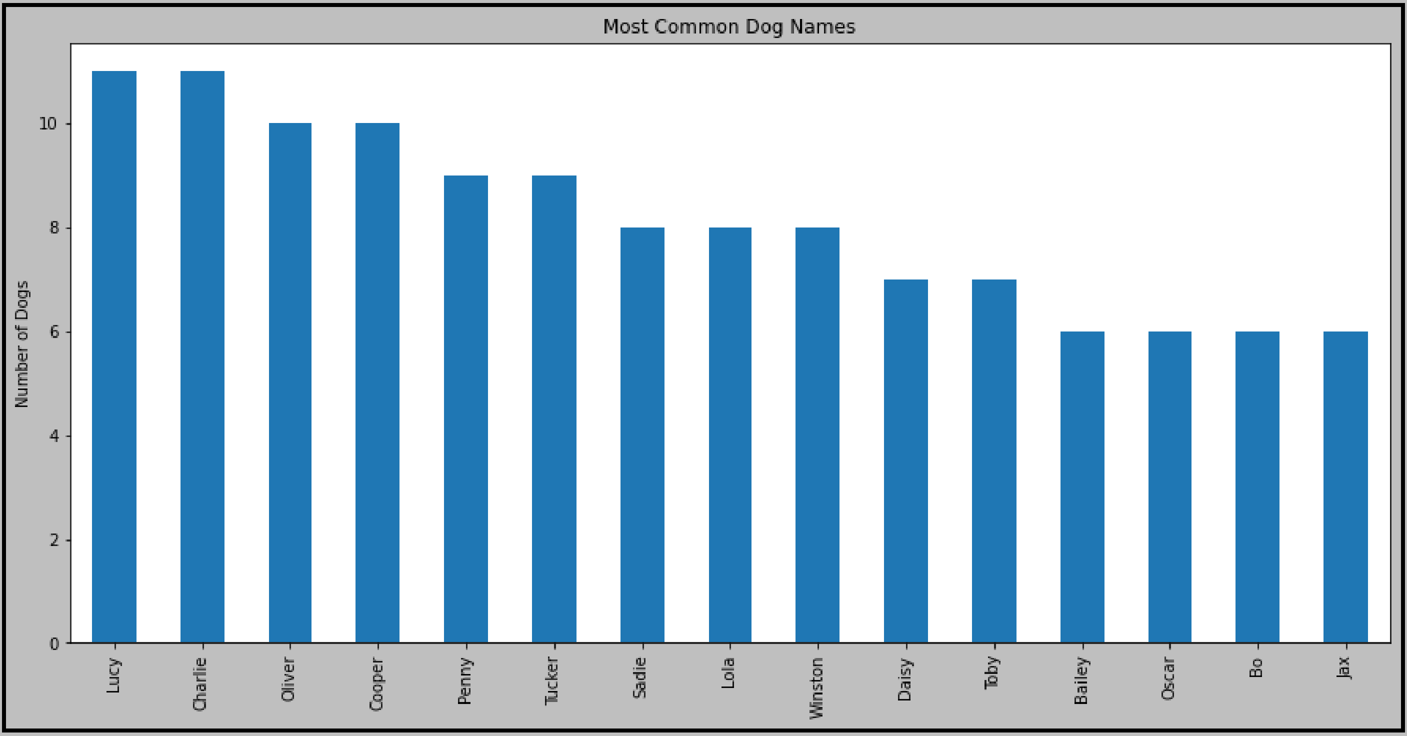

image 2

3. Visualizing the total number of tweets over time to see whether that number increases, or decreases, over time

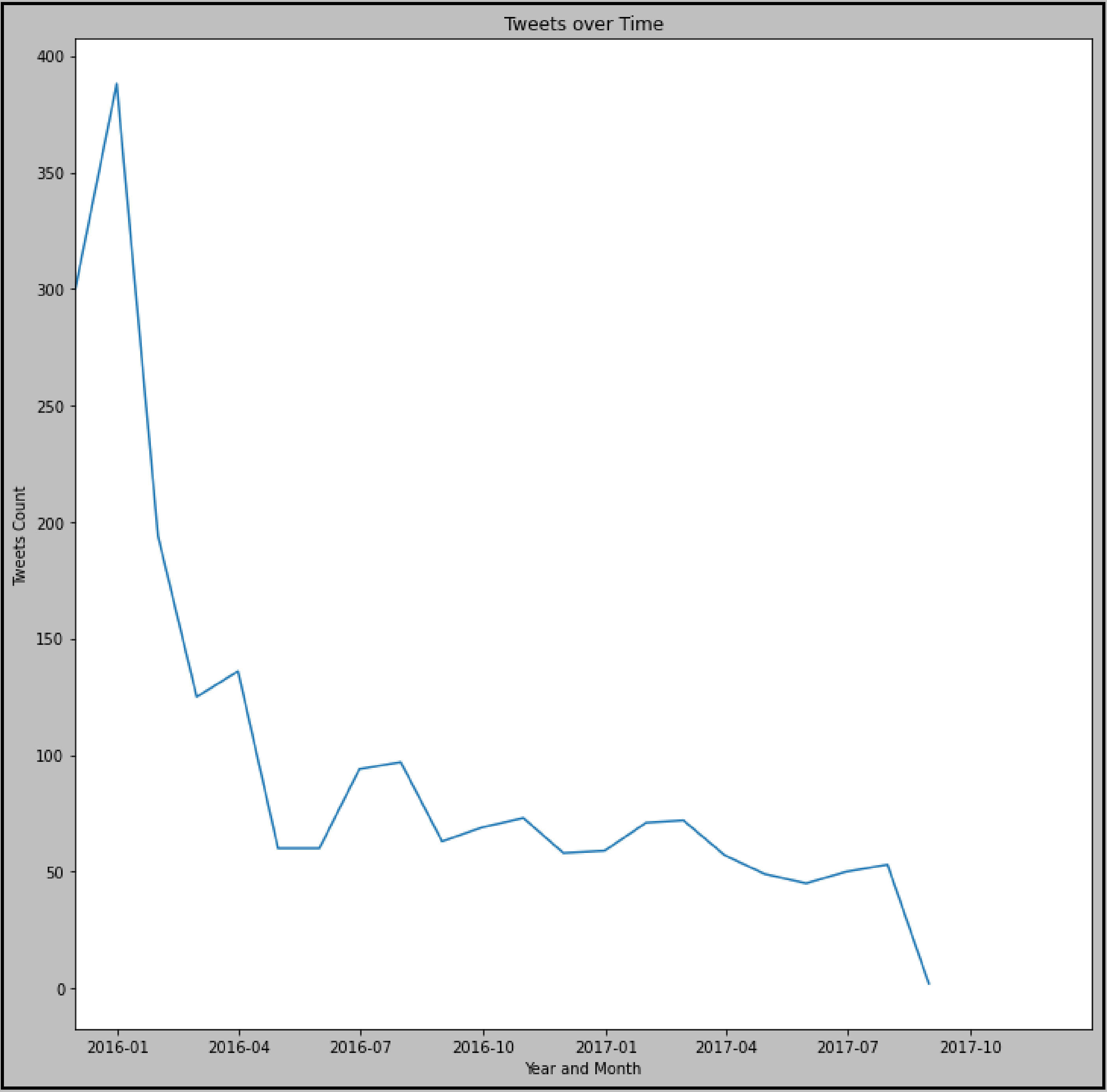

image 3

From the above image we can see that tweets are drastically decreasing over time. Specifically the tweets from 2016 onwards are dropping sharply although there are some spikes in activity. This result shows that something must be done in order to stop the decreasing trend in the tweets.

4. Visualizing the ‘retweet counts’, and ‘favorite counts’ comparison over time

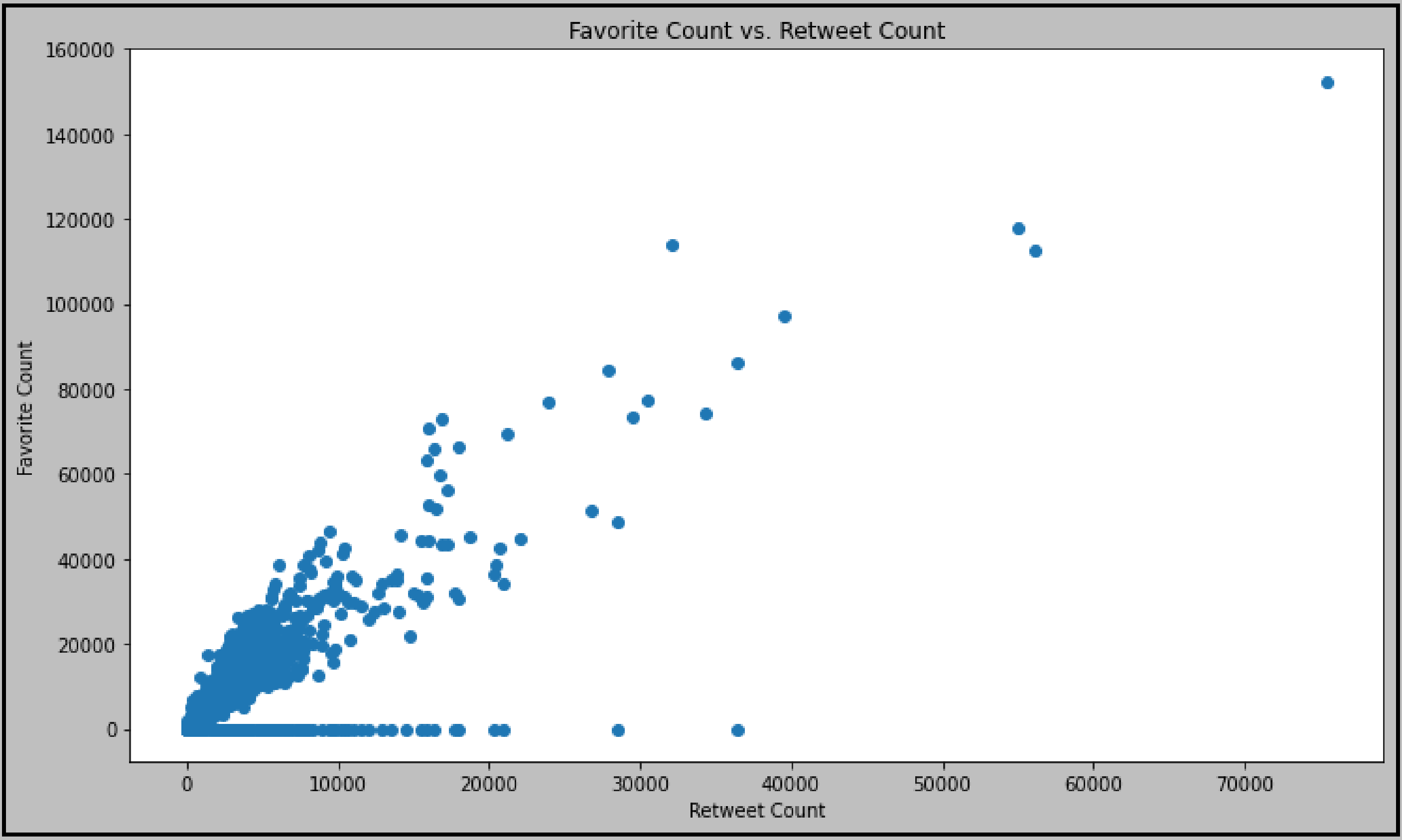

image 4

The final image shows that there is a positive correlation between Retweet Counts and Favorite Counts. This correlation is important in order to understand which method is determining to increase users’ traffic on the page.

Final word

This analysis has been done for the purposes of the specific project and its results may be altered or validated negative if other wrangling and cleaning process takes place.