Analyzing and Visualizing WeRateDogs

Introduction

The main purpose of this paper is to describe the brief analysis and visualizations done in this project, and I'll be briefly describing the wrangling procedure put in practice in the data wrangling section of Udacity Data Analyst Nanodegree. The dataset used in this project is a tweet archive from user @dog_rates or WeRateDogs. This account rates people's dogs with humorous content.

Gathering

This project consists of three different datasets that are complementary to each other. A brief explanation on how to describe them is:

1. **Twitter archive file:** This dataset has been provided by Udacity may be found in this github directory as twitter_archive_enhanced.csv.
2. **Twitter image prediction:** It helps to categorize what breed is present in each tweet according to a neural network algorithm. This file is hosted by Udacity and can be downloaded programmatically through requests.
3. **Twitter API and JSON:** This is a compilation of JSON files. Through the tweet id acquired from twitter archive file, it is queried each tweet from each tweet id through Python's Tweepy library and then stored in a txt file. It is then read and transformed into a dataframe with selected columns.

Assessing

After gathering the three datasets, the datasets were assessed as follow:

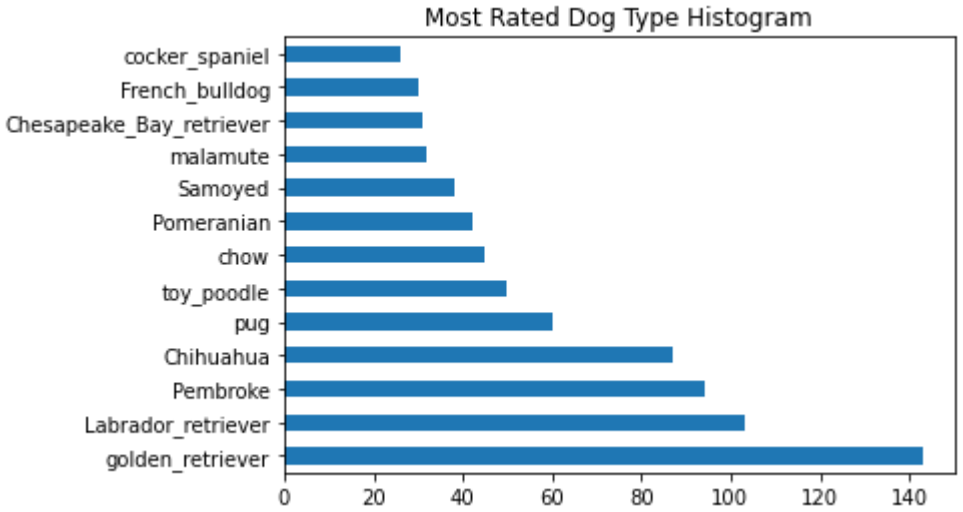
- Visually, each dataset was printed separately in the Jupyter Notebook
- Programmatically, each dataset was explored through different methods such as the functions .info(), value_counts().

Through the methods used above it was possible to identify possible problems in the datasets that could be classified as **quality issues or tidiness issues**.

Visualizing and Analyses

Most common dog breeds

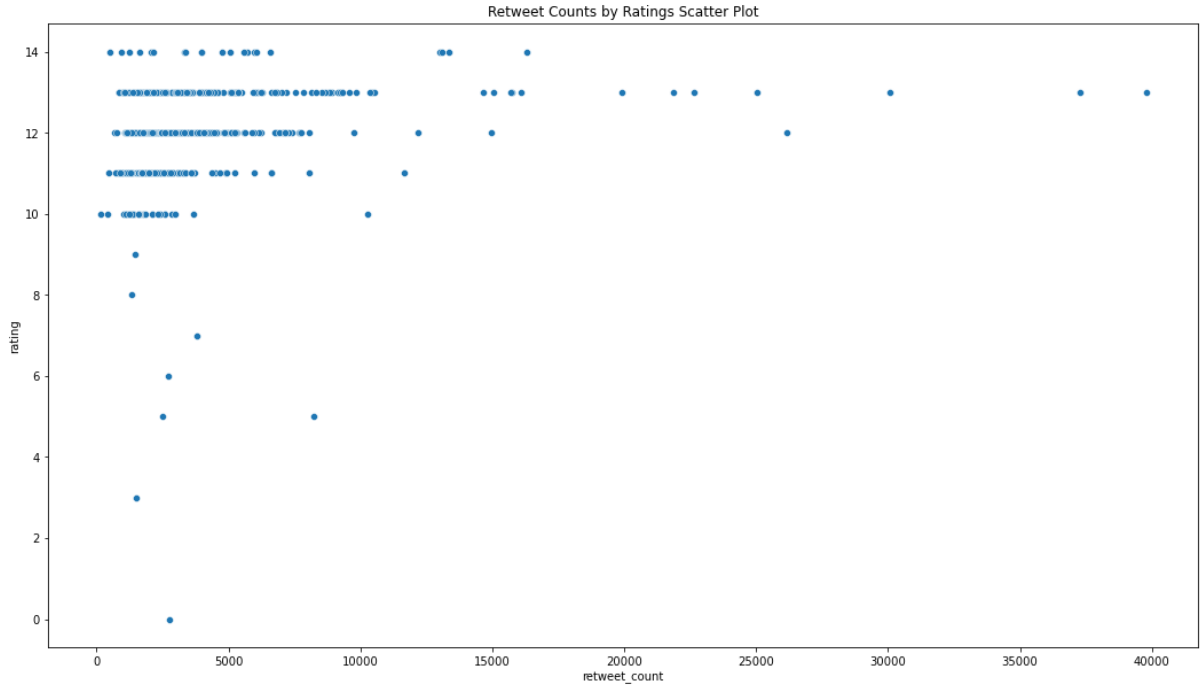
There is over 6000 tweets within WeRateDogs, and I managed to collect around 750 of them. We can see below a graph with the most prominent races and it may reflect all the tweets.



As the graph can show us we may conclude that golden_retriever is by far the dog breed with most rating among all.

Retweet Counts per Rating - Scatterplot

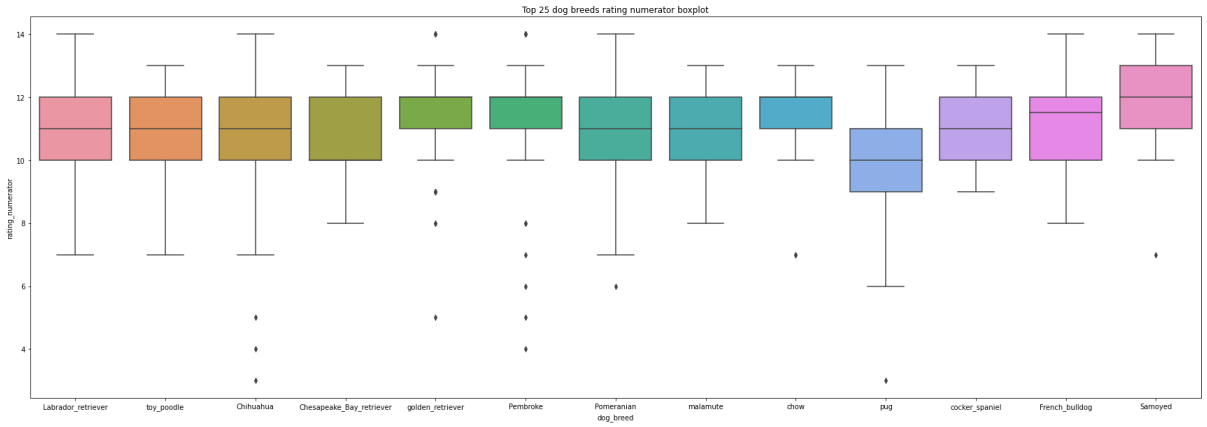
I wanted to know how the dispersion of our retweets per each rating in our dataset.



It seems that the bigger the rating, the bigger the amount of retweets. It seems logical, but we could check afterwards if this behavior will appear again in a micro analysis where we look for this dispersion within each dog breed.

Numerator dispersion per dog breed - Boxplot

This graph will show us what is the numerator dispersion between the dog breeds, and this way we may look at the differences between them.



This is a big graphic. This is a boxplot and what we're looking for visually is to the extremes, so that we can guess which dog breeds have the biggest interquartile range. The dog breeds with the most spread out ratings are pug, pomeranian and chihuaha. The dog breeds with the most concentrated reviews are golden_retriever, pembroke and chow.