Reporting: wrangle_report

For the Twitter WeRateDogs Data Wrangling Project, much work has gone into gathering, assessing and cleaning 3 datasets to with accompanying exploratory data analysis, visualizations and conclusions drawn from the datasets to gain insights into the project's nature. I will describe in some detail the steps taken to accomplish this analysis in this report in the steps below.

Data Gathering:

Three datasets were involved in the data gathering phase. The first was an archive contributed to Udacity by the WeRateDogs moderators and gathered via direct download. The retweets dataset was collected using the pandas Tweepy library. Lastly, the image predictions dataset was collected using the pandas Request library.

Data Assessing:

Eight quality and two tidiness issues were identified that affected the datasets' overall usability for analysis. These were determined with a combination of visual and programmatic assessments. The main quality issues were dataframe column datatype errors, readability errors and erroneously entered information. The tidiness issues consisted of redundant dog type columns and determining that merging the three datasets into one would improve analysis later.

Data Cleaning:

The datatype quality issues were largely corrected via programmatic datatype conversion. Errors in the confidence interval columns were rounded to improve readability. The denominator column was reset to 10 to follow the WeRateDogs rating convention. The three datasets were joined and the dog types joined into a single column in order to address the tidiness issues. Some columns with very few rows were dropped as they contributed nothing to the analysis. Lastly, the confidence intervals were rounded to have five decimal places for ease of reading.

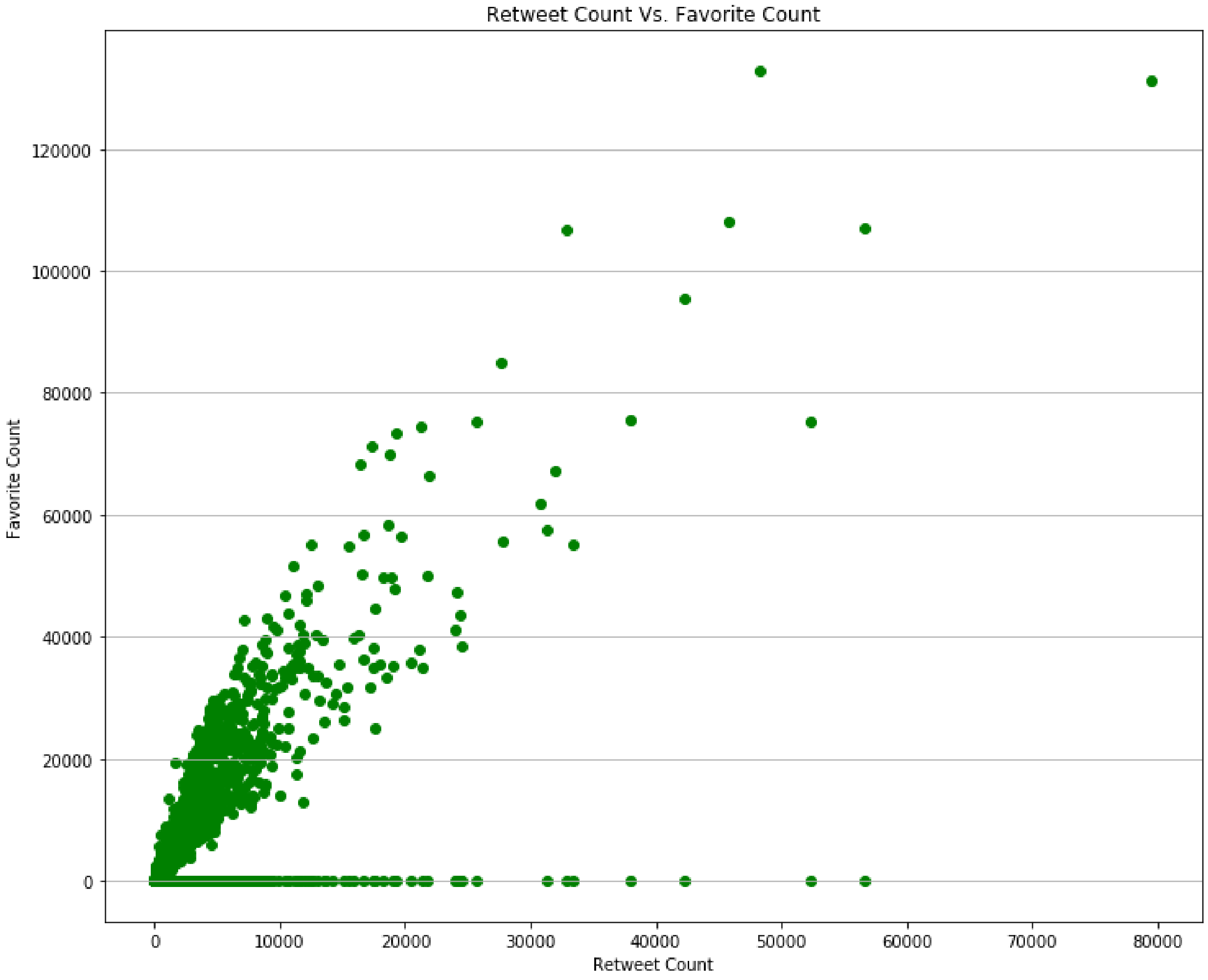
Exploratory Data Analysis (EDA):

A few EDA techniques such as the pandas sample, nlargest and describe methods and using value_counts with normalization helped me determine the most common ratings to focus on for columns like the ratings. A scatter plot visualization was useful for determining the correlation between favorite tweets and retweets. Insights gathered are based on the outputs of these EDA tools.

Insights and Visualization:

1. Scores with ranges of 10-13 are the most common, so clearly most voters are reserved in how they vote. These account for 79% of all votes. The mean score according to the describe method seems to be 15/10 across all data points. This, however, is skewed by a handful of large (and possibly erroneous) observation values, so observing the normalized frequency of the 10-13 values may be of greater insight.
2. The most frequent prediction choices for pet tweets are some combination of Golden and Labrador Retrievers and Chihuahuas.
3. According to the scatter plot below, there is a positive correlation between retweet_count and favorite_count, which probably isn't surprising since tweets that re favorited intuitively have higher retweets (shares).

Reporting: wrangle_report


Retweet Count Vs. Favorite Count
Favorite Count
Retweet Count