WeRateDogs Act Report

The three datasets that I worked with were twitter_archive, image_predictions, and tweet_json. In order for me to analyze and visualize my wrangled data, I had to first go through the following stages;

Gathering: the three datasets I worked with were all gathered in different ways. Twitter_archive was provided by Udacity and downloaded manually, image_predictions was downloaded programmatically using the requests library and a provided URL, while tweet_json was queried from the Twitter API after creating a developer account and obtaining credentials. After gathering, they were all loaded into pandas data frames.

Assessing: the data frames were then assessed for quality and tidiness issues.

Cleaning: the identified quality and tidiness issues were then cleaned following the define, code, and test framework of data cleaning. During this stage, I also merged the three different data frames into one on the basis of the column they had in common.

Storing: After cleaning, I stored the merged data frame into a CSV file ready for analysis and visualization.

After going through the above stages, I performed an analysis and visualization of the merged data frame archive_master.

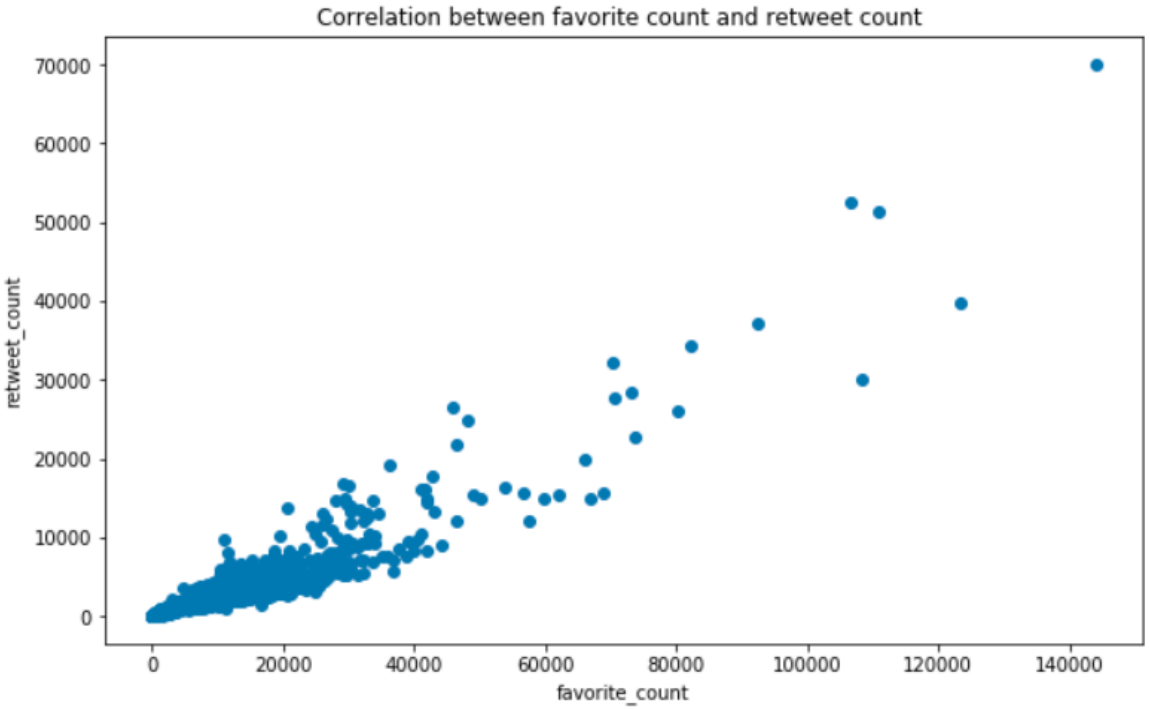
In order for me to get insights from the merged data frame, I used pandas methods such as .describe, and .shape.

Insights obtained include;

- The merged dataset archive_master had 1986 rows and 22 columns.
- Among all the columns of archive_master, dog_stage was the only one that had some missing values.
- The highest favorite count was 143944 while the lowest was 66.
- The highest retweet count was 70134 while the lowest was 11.

Visualization

1. There was a positive correlation between favorite count and retweet count where the tweets which were liked most were also retweeted most as observed from the graph below.



2. The most popular dog_stage is pupper, followed by doggo, and puppo.