INTRODUCTION

In this wrangling project, I wrangled a live dataset from the Twitter account of @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

The 3 datasets used in the project are:

1. The Twitter archive from the WeRateDogs Twitter account.
2. Additional data from Twitter API.
3. An image prediction file.

In wrangling the data, I made extensive use of the following Python libraries:

1. Pandas
2. NumPy
3. Matplotlib
4. Seaborn
5. Requests
6. JSON

PROCESS

I started the wrangling process by first getting and importing the data from various sources into my workspace. I then used various pandas exploratory methods and functions such as .info(), .head(), .describe(), .sample(), .shape(), and .duplicated() to get an overall feel of the datasets.

The supplied datasets have data quality and tidiness issues and, I was able to identify 8 data quality issues and 2 tidiness issues. In addressing these issues, I first made a copy of the datasets and promptly proceeded to clean the data through the following broad steps:

1. Converted all dataset feature into the correct datatype using pandas .astype() function for string-based features and .to_datetime() for date like features. I observed that the presence of null values makes some operations difficult; especially when performing comparison-based operations.
2. Removed all features that are not needed in the project, such as features containing information about tweets that are retweets.

3. I also dropped records that contain null values in specific feature columns and these null values cannot be safely imputed through other means. In addition, some encodings for null values were incorrect, I had to resolve such encodings such as wherever the text 'None' was used to represent the object 'np.nan'.
4. In addition, to ensure that the project adheres as closely as possible to the rating system used, I removed most values that did not obey the unique rating system where the numerator is greater than the denominator or cases where the denominator was not equal to 10.
5. I also split the data in the expanded URL feature column to ensure that there is only one URL for each tweet. This is to ensure that our dataset is tidy.
6. I also merged the 3 datasets into a single dataset with the joint key being the tweet id feature column. This allowed me to have a single dataset that contains full tweet information including tweet retweet count, favourite count, and dog breed name.
7. The final cleaned dataset was then saved for future use. This future use included my use of the cleaned dataset to answer some questions such as:
    a. What is the highest rating for any tweet?
    b. What is the most retweeted tweet?
    c. What is the most favourited tweet?
    d. Is there a relationship between retweet count and favourite count? (It turns out that there is a strong positive relationship).
    e. What are the most tweeted dog breeds?

INSIGHTS

Using the cleaned data, the following insights were observed in the data:

1. The highest rating for any dog tweet as communicated by my cleaned data is *14.0/10 and 24 dog tweets have that rating.*
2. The most retweeted tweet (https://twitter.com/dog_rates/status/744234799360020481) of a ***Labrador Retriever***, with a favorite_count of ***131,075*** was retweeted ***79,515*** times and has a rating of 13.0/10. While the least retweeted tweet (https://twitter.com/dog_rates/status/666102155909144576) of an

English Setter, with a favorite_count of ***81*** was retweeted ***16*** times and has a rating of 11.0/10 .

3. The most favourited tweet (https://twitter.com/dog_rates/status/822872901745569793) of a ***Lakeland Terrier***, with a retweet_count of ***48,265*** was favourited ***132,810*** times and has a rating of 13.0/10. While ***64*** tweets were favourited ***0*** times.
4. The ***5*** most tweeted dog_breed are *Golden Retriever (168 tweets), Labrador Retriever (108 tweets), Pembroke (94 tweets), Chihuahua (92 tweets), and Pug (64 tweets)*.

VISUALIZATIONS

I also did a visualization to show the 5 Most Tweeted Dog Breeds in the dataset using Matplotlib. This is shown in figure 1 below.

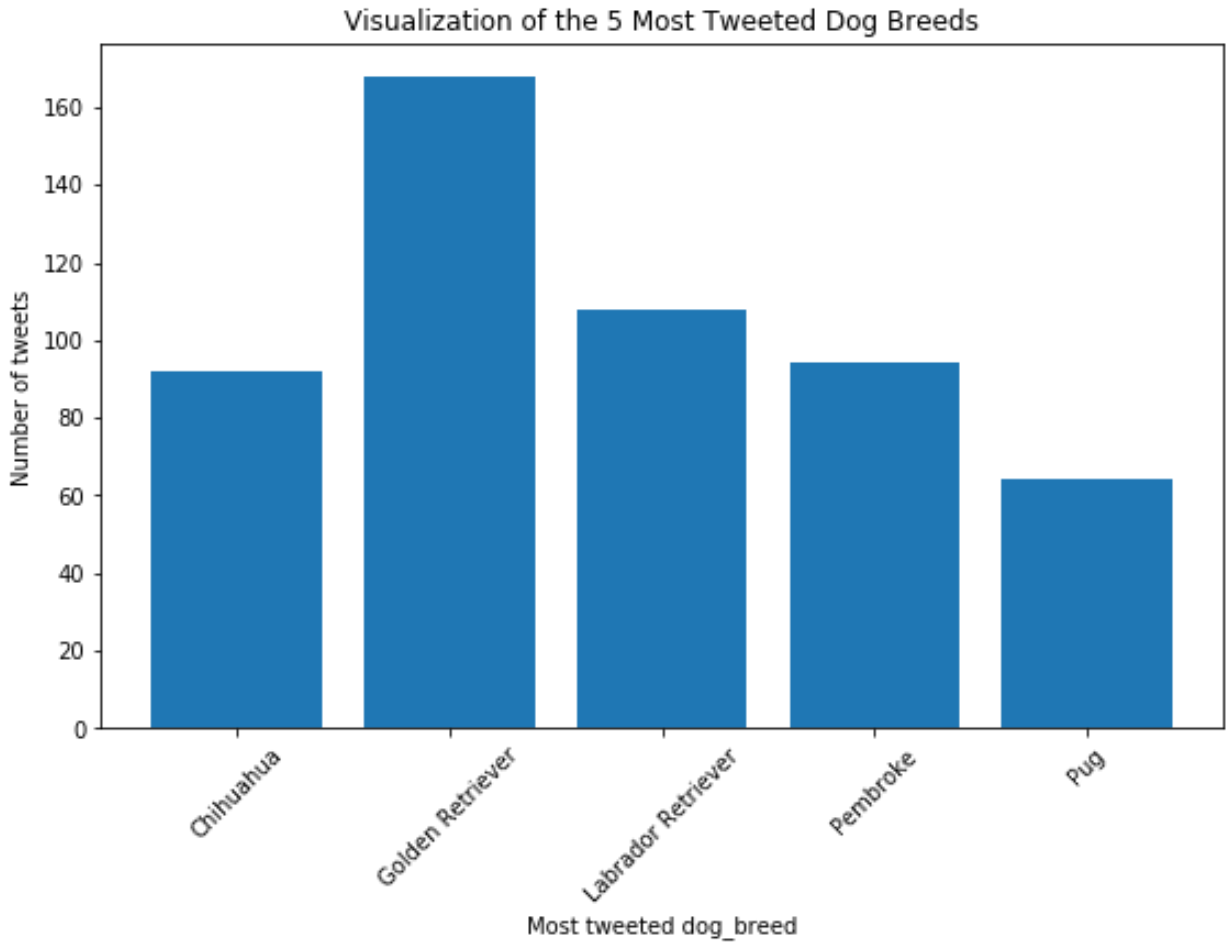

Figure 1: Visualization of the 5 Most Tweeted Dog Breeds

To view the Correlation Heatmap for all numeric variables, I plotted a Seaborn heatmap of the dataset as shown in figure 2 below.

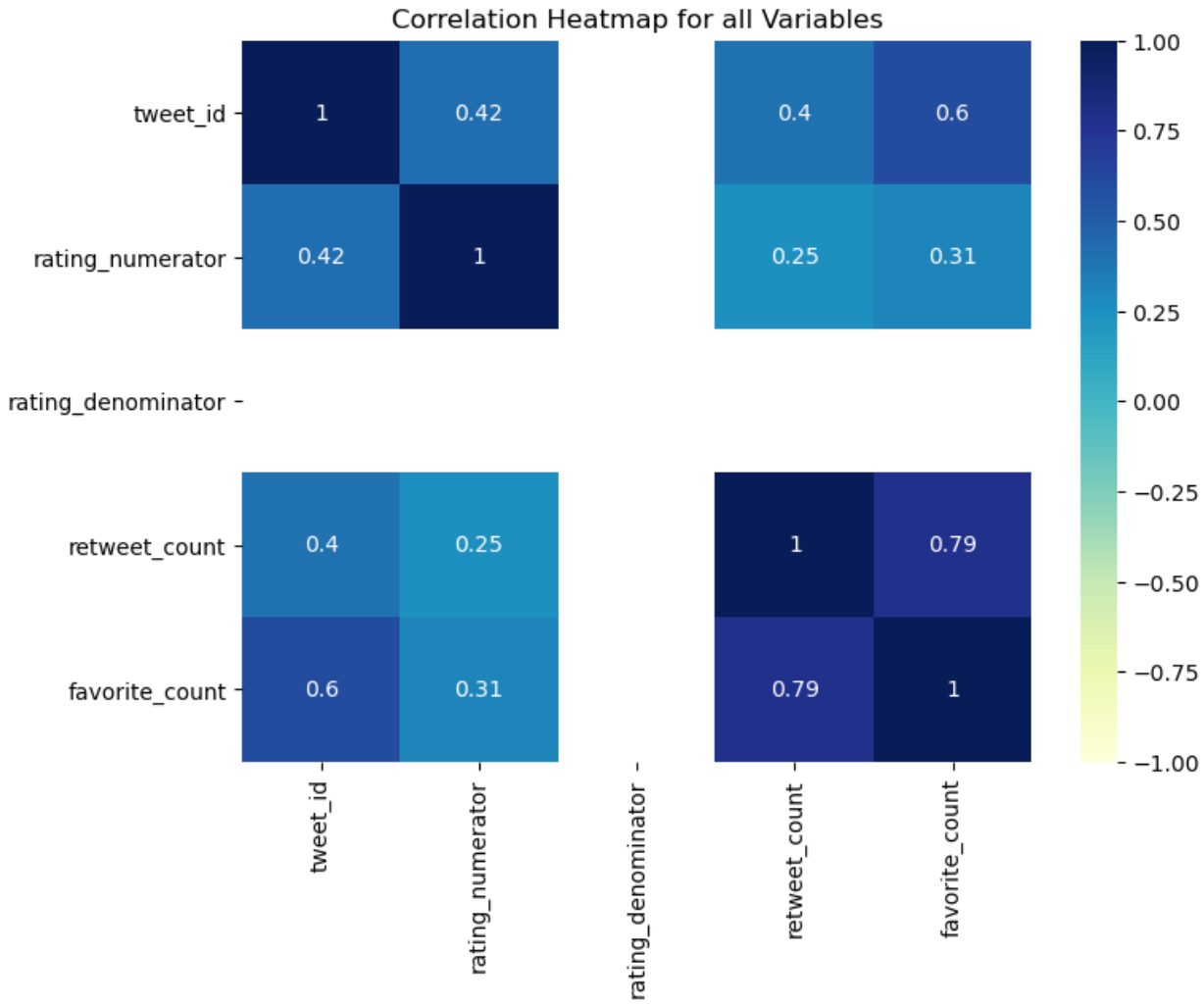

Figure 2: Correlation Heatmap for all Variables

Since I observed a correlation between retweet count and favourite count, I decided to plot a Seaborn scatterplot of the two features. The result is shown in figure 3 below.

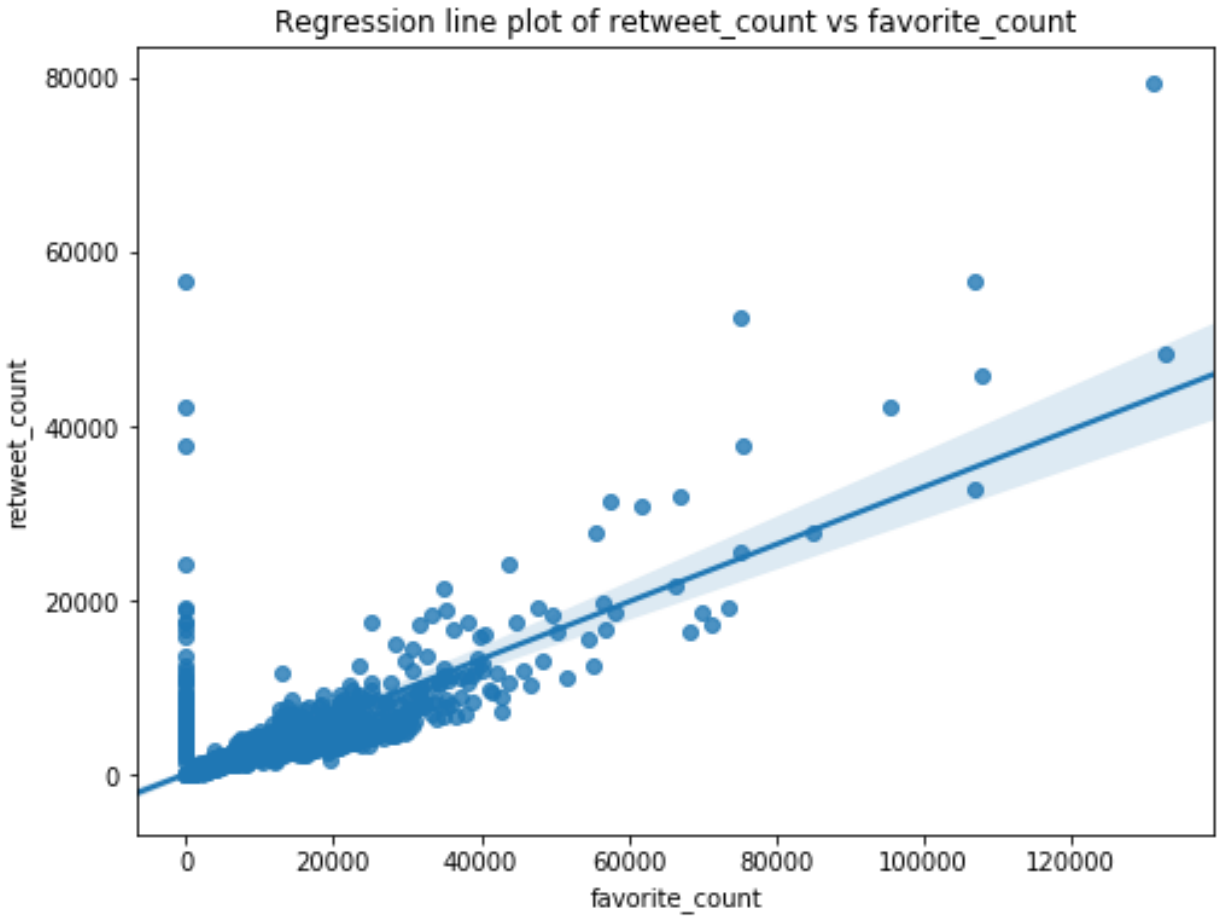

Figure 3: Scatter plot of retweet_count vs favorite_count

CONCLUSIONS

The opportunity to wrangle the data in this project is well appreciated. My most significant takeaway from this is that I should pay particular attention to null values whenever I am wrangling a dataset.