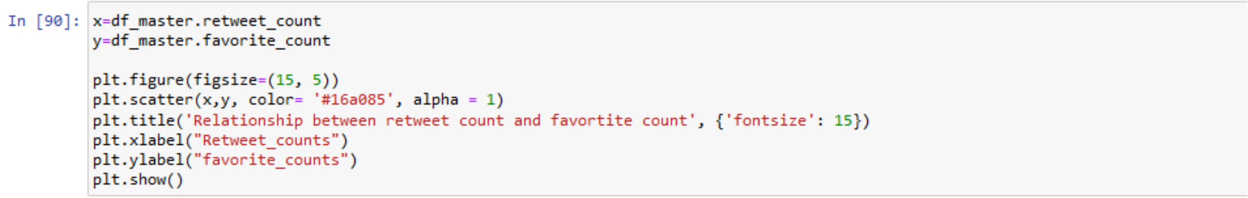
Act Report for WeRateDogs

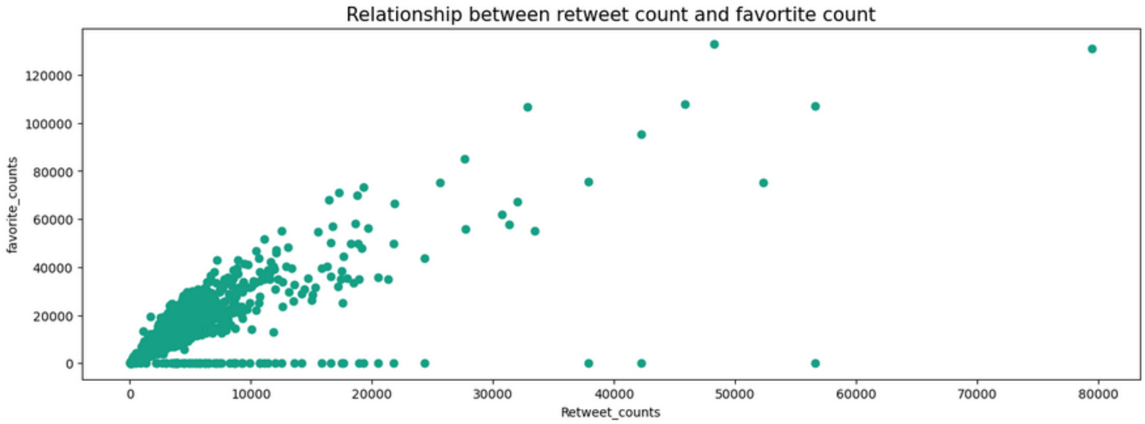
In this report, we analyzed and cleaned the data from the popular Twitter/x database of WeRateDogs. We cleaned data from multiple sources to gather the insights shown. These sources included image predictions as well as information from the Twitter API. We focused on this report's three main things: the retweet count related to the favorite count, the stage of dogs in the tweets, and finding what the most popular names for dogs in the database were.

Insight 1:

When going through the data, I realized that there could be a correlation between the retweet count and the favorite count. Once the data was all cleaned a merged into one master file, I checked this hunch by plotting a scatter chart. The results show that there was indeed a correlation.



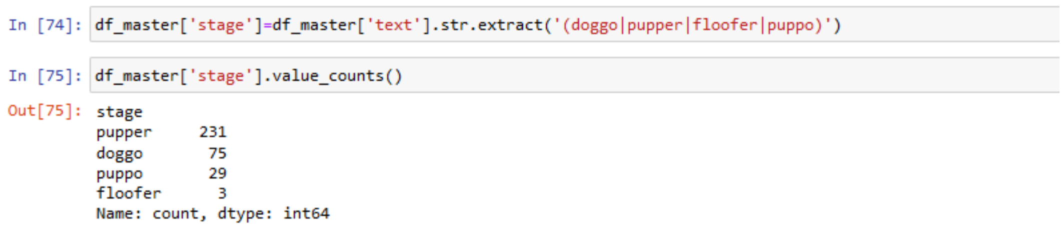
```
In [90]: x=df_master.retweet_count
         y=df_master.favorite_count

         plt.figure(figsize=(15, 5))
         plt.scatter(x,y, color= '#16a085', alpha = 1)
         plt.title('Relationship between retweet count and favortite count', {'fontsize': 15})
         plt.xlabel("Retweet_counts")
         plt.ylabel("favorite_counts")
         plt.show()
```



The scatter plot clearly shows that there is a positive correlation between retweet and favorite count. When the favorite count goes up, then the retweet counts will typically go up as well.

Insight 2:

In order to get the different stage counts, we first combined all the different stages into one field and then did a count of them in total.

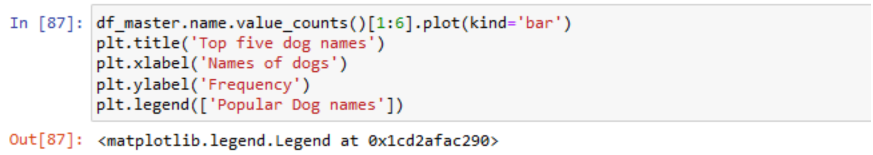


```
In [74]: df_master['stage']=df_master['text'].str.extract('(doggo|pupper|floofer|puppo)')

In [75]: df_master['stage'].value_counts()

Out[75]: stage
         pupper     231
         doggo       75
         puppo       29
         floofer      3
         Name: count, dtype: int64
```

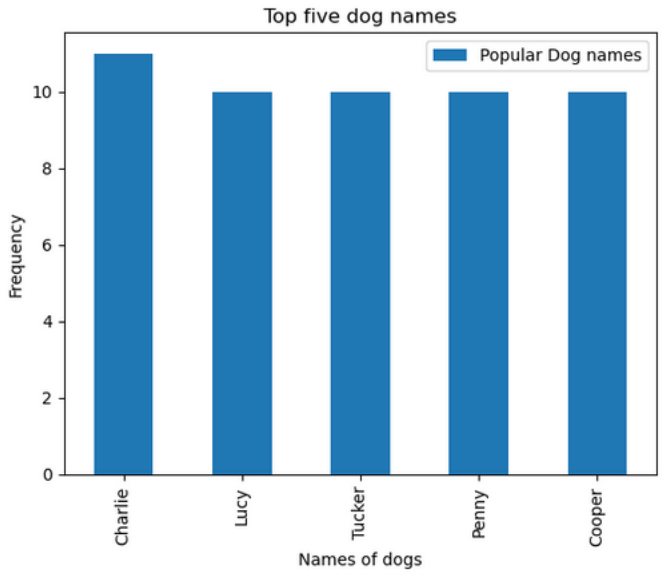
As we can see here, once we combined the different stages together, we could get an accurate count. At this point, it became easy to see that the pupper stage was far and away the most popular and the floofer stage is far and away the least popular.

Insight 3:

The most popular names for dogs in the database would be some good information to know. This can assist people to know if any specific name would be better than others for possibly gaining retweets. I wanted to know the top five names for this report.



```
In [87]: df_master.name.value_counts()[1:6].plot(kind='bar')
         plt.title('Top five dog names')
         plt.xlabel('Names of dogs')
         plt.ylabel('Frequency')
         plt.legend(['Popular Dog names'])

Out[87]: <matplotlib.legend.Legend at 0x1cd2afac290>
```



As we can see from the results, the most common name is Charlie, followed by Lucy, Tucker, Penny, and Cooper. All the top five names are close in frequency.

Conclusion:

While I know that there will be some errors in this data, I believe that this is a good representation of data gathering and reporting. I am open to any comments and recommendations to make this report better in any way.