act_report

September 16, 2022

0.1 Report: act_report

```
In [2]: # Importing relevent libraries for visualization.
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
```

The shift from traditional way of doing things from shopping, communication and even travelling, to the new Digital world made loads of data to be easily collected. But the sad part is that most if not all this collected data is not clean, and it is our task to clean it in order to take advantage and use it to find useful insights.

The original data came from different sources and file formats but,after performing all the teps of the data wrangling of the dataset from a Twitter page that rate dogs and then share mostly harilious comments, the master dataframe was achieved with more tidy and clean attributes and structure. After analysis some insights were found and below are the visualization to some of the insights.

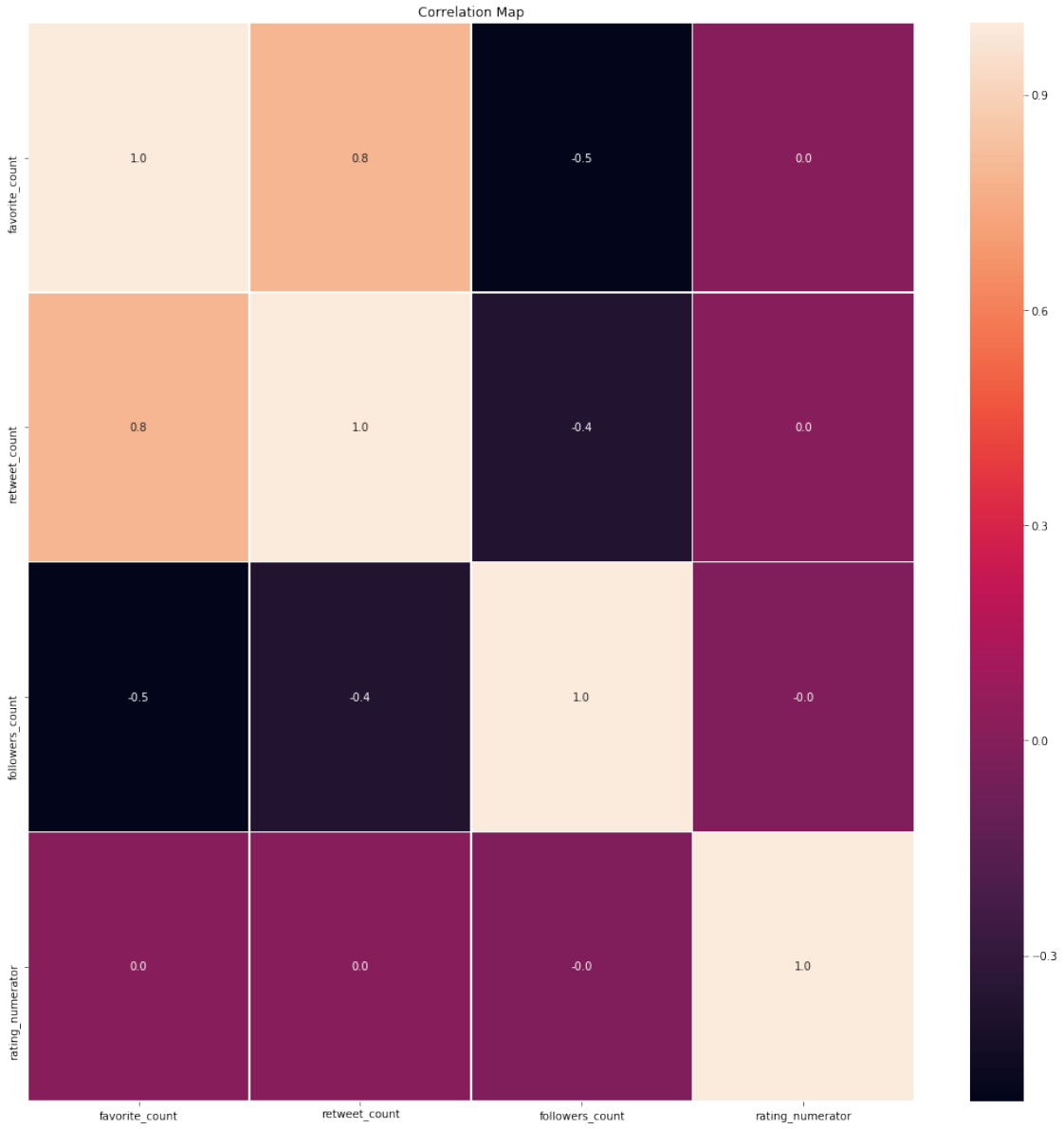
```
In [15]: ## Importing the already cleaned dataset

         final_cleaned=pd.read_csv('twitter_archive_master.csv')
```

0.1.1 Correlation between variables

```
In [14]: ##hitmap between few selectected numerical variables
         f,ax = plt.subplots(figsize=(18, 18))
         sns.heatmap(final_cleaned[['favorite_count',
                                    'retweet_count', 'followers_count',
                                    'rating_numerator']].corr(), annot=True, linewidths=.5, fmt= '.1
         plt.title('Correlation Map')

Out[14]: Text(0.5,1,'Correlation Map')
```

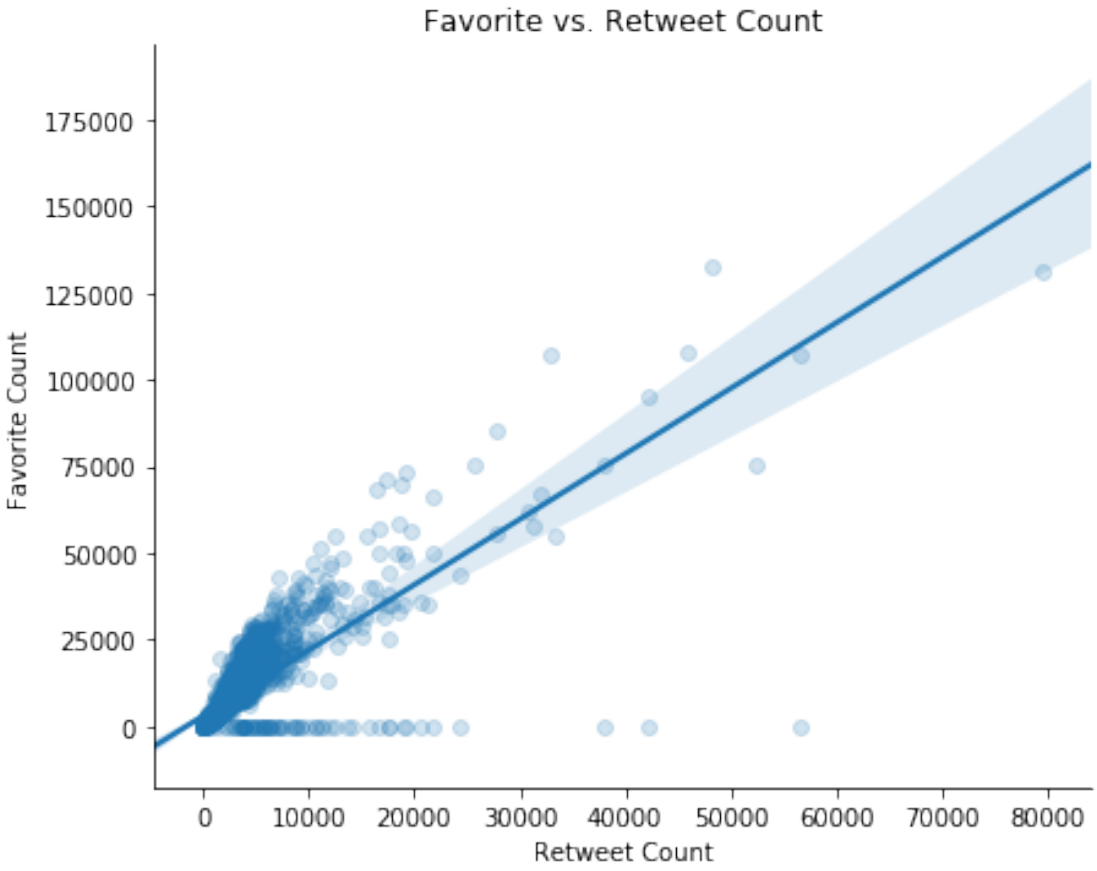


There is a strong correlation between favorite count and retweeted count, and that means 'liked' tweets were also retweeted.

0.1.2 Retweet vs Favorite count

```
In [4]: # Plot scatterplot of retweet vs favorite count
        sns.lmplot(x="retweet_count",
                   y="favorite_count",
                   data=final_cleaned,
                   size = 5,
                   aspect=1.3,
                   scatter_kws={'alpha':1/5})
```

```
plt.title('Favorite vs. Retweet Count')
plt.xlabel('Retweet Count')
plt.ylabel('Favorite Count');
```



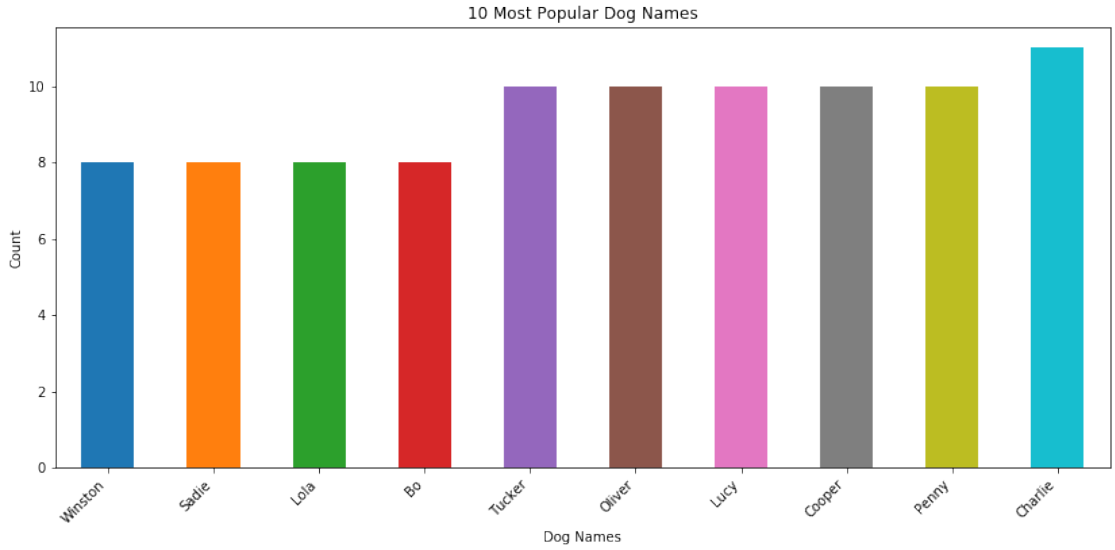
There is a clear and visible corelation between Favorite count and retweeted count, and that backs the correlation heatmap.

0.1.3 Ten most popular Dog names

```
In [8]: # filter dataset for entries with dog names only
        dogs_with_name = final_cleaned[final_cleaned.name.notna()]

        # Drawing the plot for 10 most frequent names
        most_frequent = dogs_with_name.name.value_counts(ascending=True)[-10:]

        plt.figure(figsize=(14, 6))
        most_frequent.plot(kind='bar')
        plt.title("10 Most Popular Dog Names")
        plt.xlabel("Dog Names")
        plt.ylabel("Count")
        plt.xticks(rotation=45, horizontalalignment='right');
```



Charlie is the most popular dog name

0.1.4 Conclusion

Lot of insights and questions can be addressed from the cleaned dataset, and above is the few from many beautiful visualiations that can be used for conclusions. And without the cleaned dataset it is impossible to achieve these quality visualizations and insights and hence that tells us the importance of the data wrangling stage in a project. By doing more of data wrangling these can be improved as there is still lot of issues that can be addressed on the data, especially variables that were not relevent for this specific type of project.