Project #5: Wrangle and Analyze Data (Act_Report)

Date: Oct26-2019
Name: Sami Adham
Program: Data Analysis Nanodegree

Introduction:

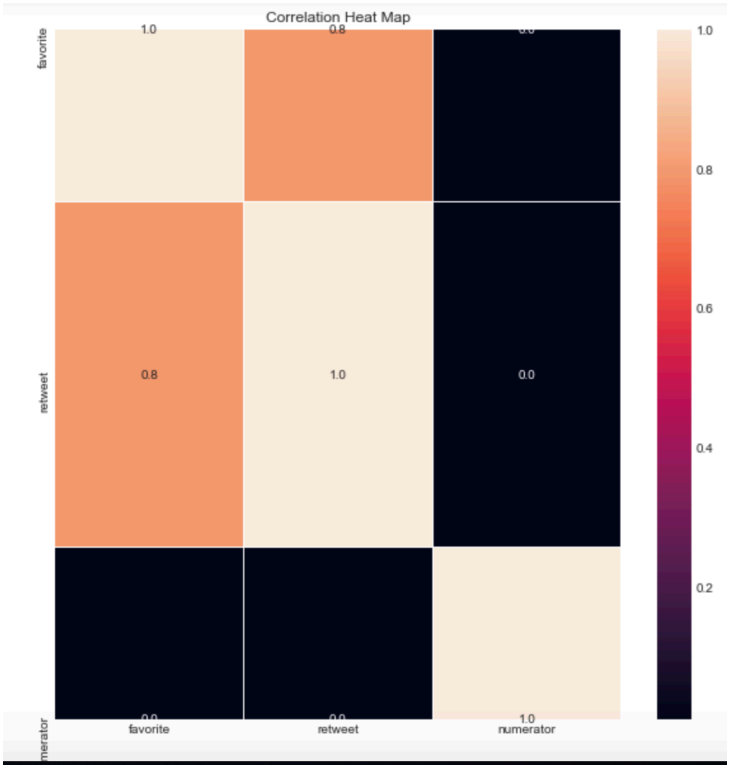
Wrangle and Analyze Data project required us to create report for wrangled data and act data, first report to show the process and implementation of how we gathering data from 3 sources, assessing data and clean it. Then we need to store cleaned dataset in CSV file before analyzing and visualizing dataset. After analyzing the data we come up with the following steps:

Project Steps:

Step1: Find the relationship between variables using correlation function

```
In [439]: #By using seasborn to find the strongest correlation between columns as below:
          f,ax = plt.subplots(figsize=(10, 10))
          sns.heatmap(tw_archive_clean[['source', 'favorite','retweet',
                        'numerator']].corr(), annot=True, linewidths=.5, fmt= '.1f')

          plt.title('Correlation Heat Map');
```

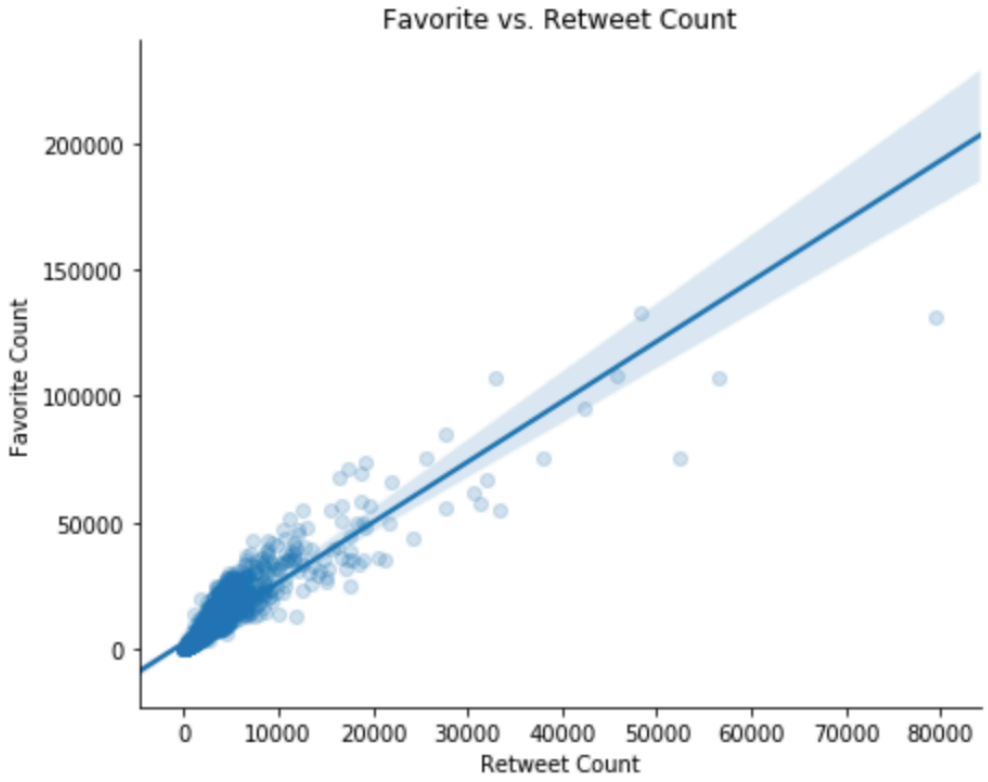


Insights:

- The strongest relationship in the dataset features are Favorites vs Retweet.
- The relationship is positive relationship as long favorites increase as long as retweet is increase.
- Numerator do not affect any other variables.

- Step2: Favorite vs Retweet

```
# Plot scatterplot of retweet vs favorite count
sns.lmplot(x="retweet",
           y="favorite",
           data=tw_archive_clean,
           size = 5,
           aspect=1.3,
           scatter_kws={'alpha':1/5})
plt.title('Favorite vs. Retweet Count')
plt.xlabel('Retweet Count')
plt.ylabel('Favorite Count');
```



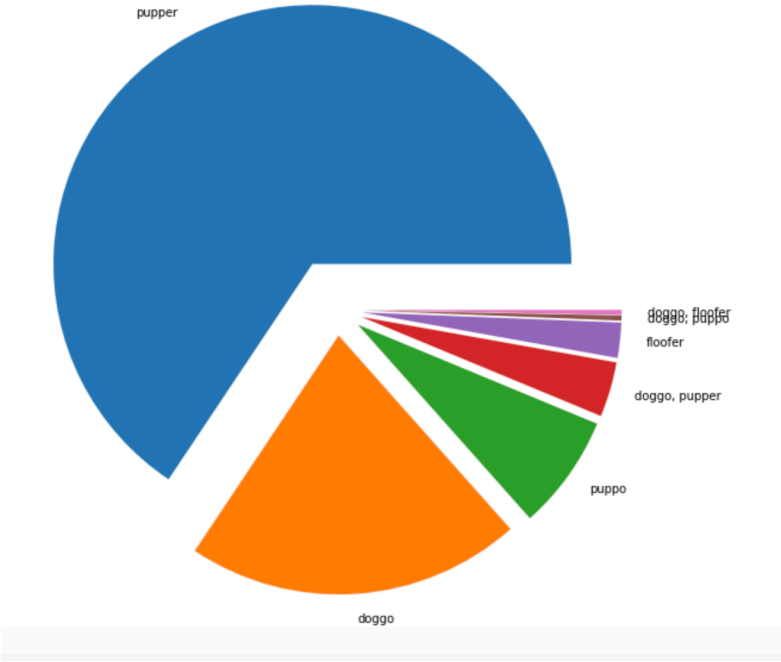
This graph shows how the relation is strong which is normal as we expected.

Step 3: Category of Dog Stage if not None

In [289]:

```
# Plot pie chart
dog_stage_count = list(tw_archive_clean[tw_archive_clean['dog_stage'] != '']['dog_stage'].value_counts())[0:7]
dog_stages = tw_archive_clean[tw_archive_clean['dog_stage'] != '']['dog_stage'].value_counts().index.tolist()[0:7]
explode = (0.2, 0.1, 0.1, 0.1,0.1,0.1,0.1)

fig1, ax1 = plt.subplots(figsize=(10,10))
ax1.pie(dog_stage_count, explode = explode, labels = dog_stages);
```



This show the most popular dog stage in our dataset. Pupper is the most variable in our data unlike foofer which is the weakest.

Tools:

1. python through Jupyter notebook (ANACONDA) to take a chance to practice in this first project for future projects.
2. Visualization libraries:
    a. Matplotlib
    b. Seaborn
3. Upload work in github." https://github.com/SamiAdham/TMDb-movie-data/blob/master/investigate-a-dataset-%5BSami%20Adham%5D.ipynb"
4. Microsoft Word to prepared 'PDF'report

References:

GitHub:
https://github.com/SamiAdham/TMDb-movie-data/blob/master/investigate-a-dataset-%5BSami%20Adham%5D.ipynb"