This documentation is to capture the analysis and visualization steps carried in the We Rate Dogs project

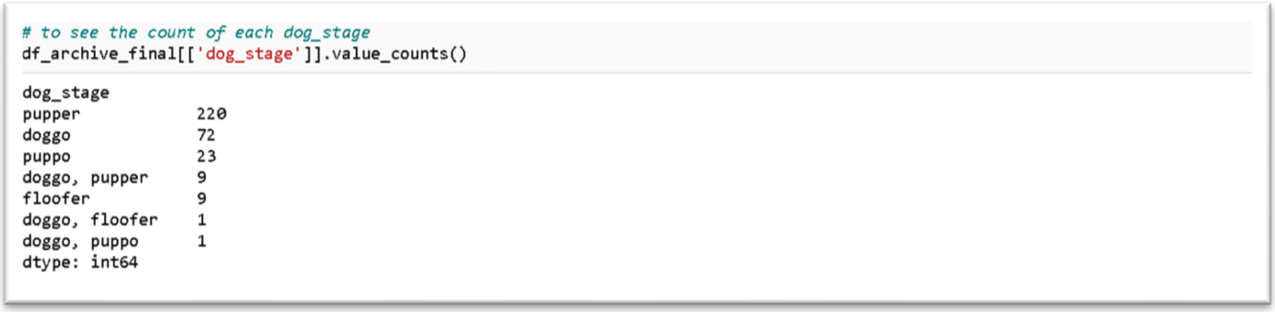
The final data ‘ **twitter_archive_master.csv’** consists of 1954 records with 29 columns

Analysis and Visualization for insights

1. Viewing the number of different dog_stages:

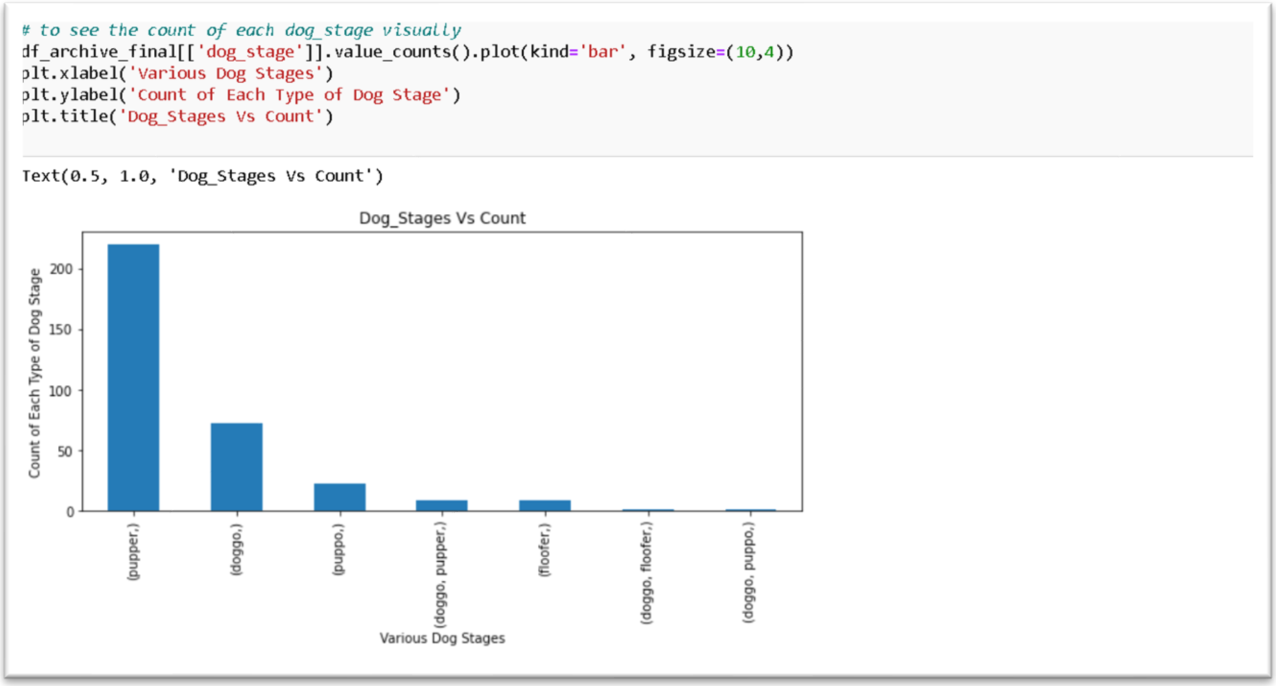
There are four different dog_stages such as **pupper**, **doggo**, **puppo** and **floofer**. I have used value counts () method to see the numbers as well bar plot to see the count.

It is confirmed that the pupper is the highest dog_stages among other



```
# to see the count of each dog_stage
df_archive_final[['dog_stage']].value_counts()
```

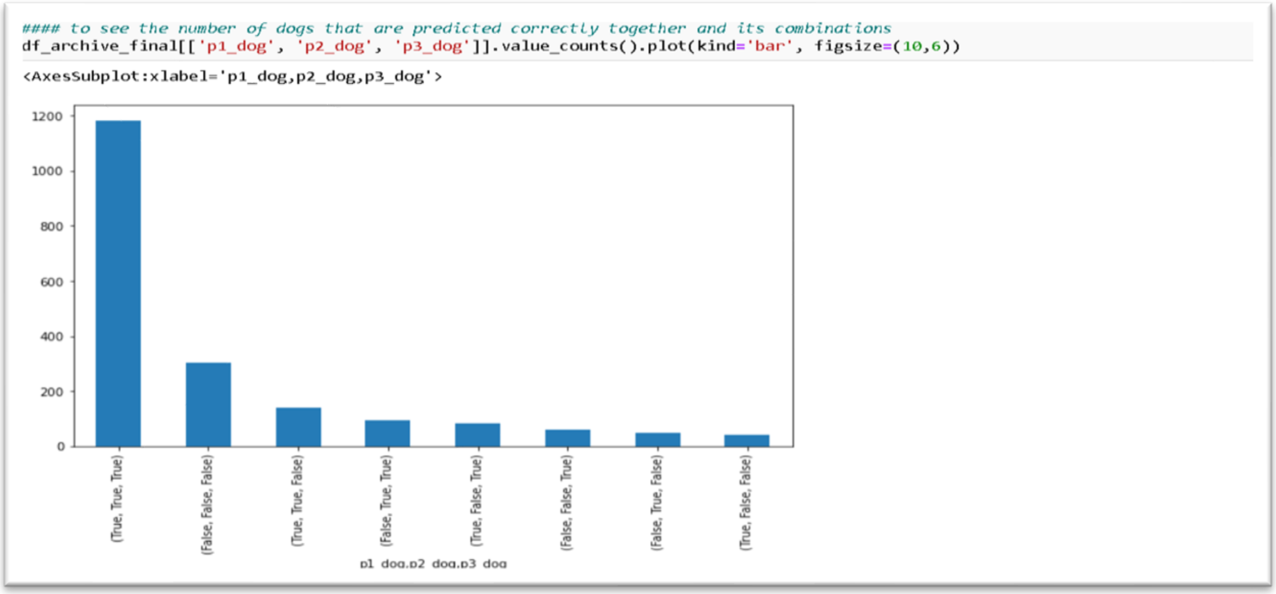
```
dog_stage
pupper            220
doggo              72
puppo              23
doggo, pupper       9
floofer             9
doggo, floofer      1
doggo, puppo        1
dtype: int64
```



```
# to see the count of each dog_stage visually
df_archive_final[['dog_stage']].value_counts().plot(kind='bar', figsize=(10,4))
plt.xlabel('Various Dog Stages')
plt.ylabel('Count of Each Type of Dog Stage')
plt.title('Dog_Stages Vs Count')
```

```
Text(0.5, 1.0, 'Dog_Stages Vs Count')
```

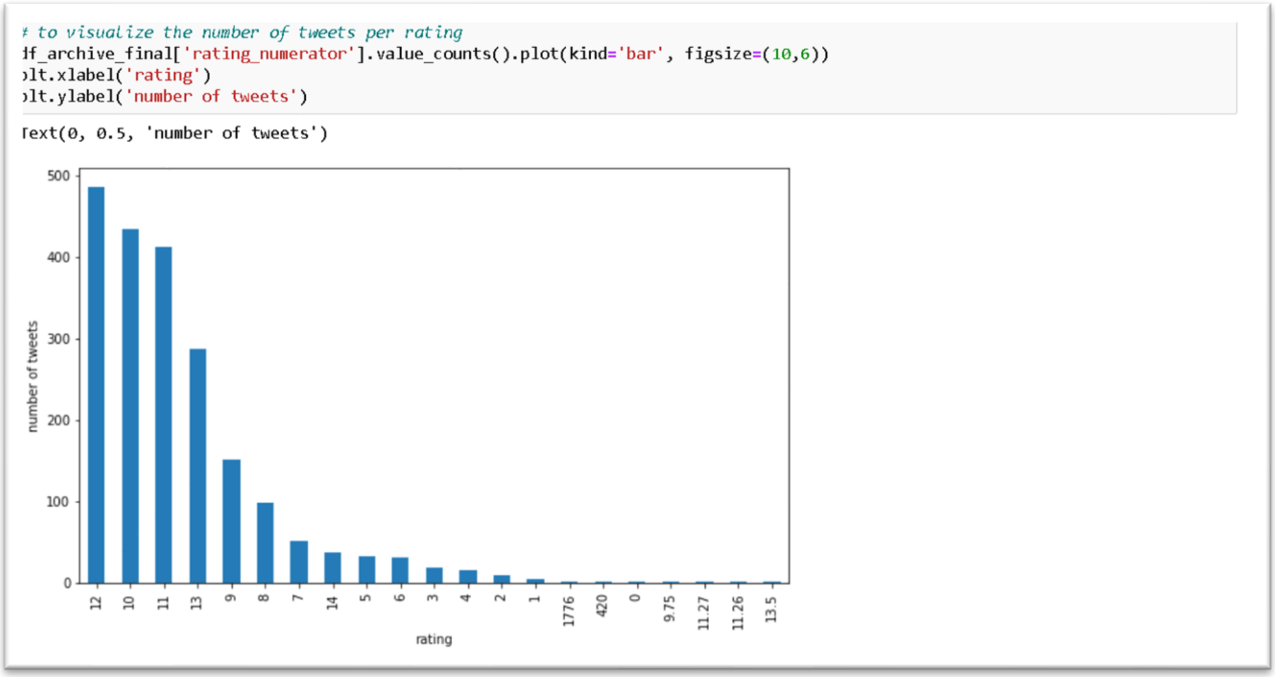
2. To see the combinations of prediction of dog breed by three algorithms
   Here the number of correct predictions (True, True, Ture) by all three algorithms is much higher than others.



```
#### to see the number of dogs that are predicted correctly together and its combinations
df_archive_final[['p1_dog', 'p2_dog', 'p3_dog']].value_counts().plot(kind='bar', figsize=(10,6))
```

```
<AxesSubplot:xlabel='p1_dog,p2_dog,p3_dog'>
```

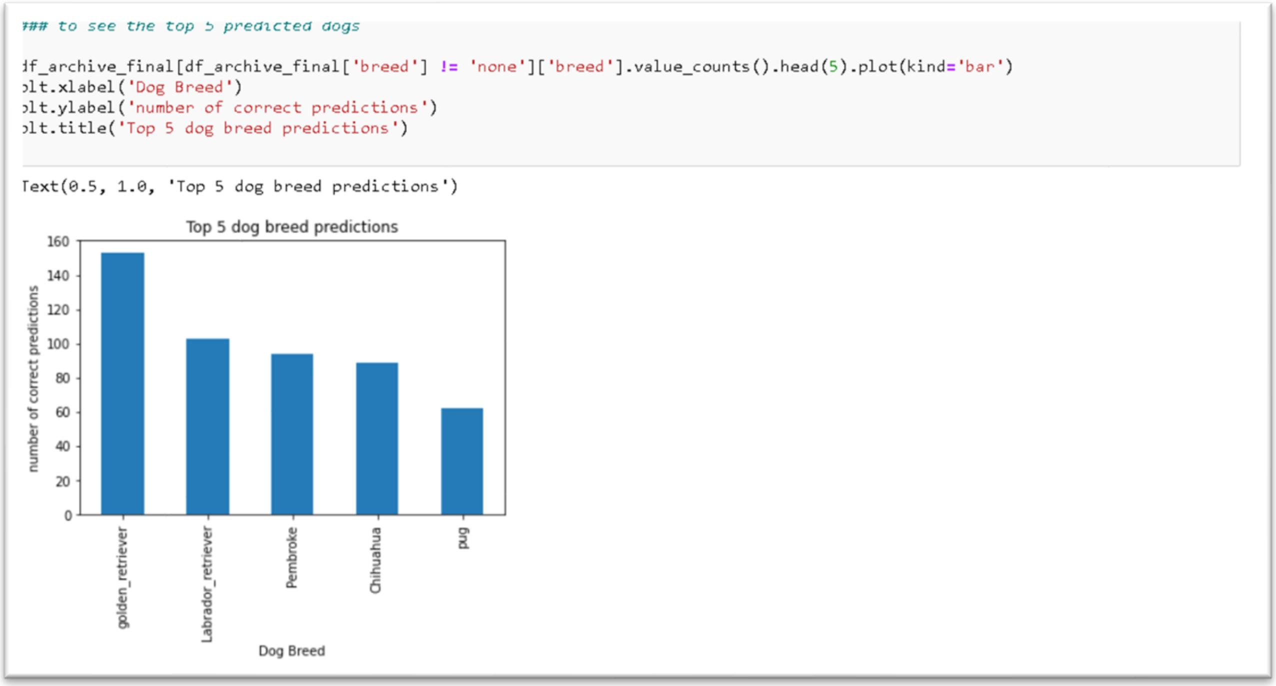
3. To see how the tweets are distributed against rating
   Its clearly evident that the ratings between 12 to 9 have more tweets.



```
# to visualize the number of tweets per rating
df_archive_final['rating_numerator'].value_counts().plot(kind='bar', figsize=(10,6))
plt.xlabel('rating')
plt.ylabel('number of tweets')
```

```
Text(0, 0.5, 'number of tweets')
```

4. Top 5 breeds that are predicted correct
   Its clear that golden retriever is the top breed that is predicted right by the algorithms



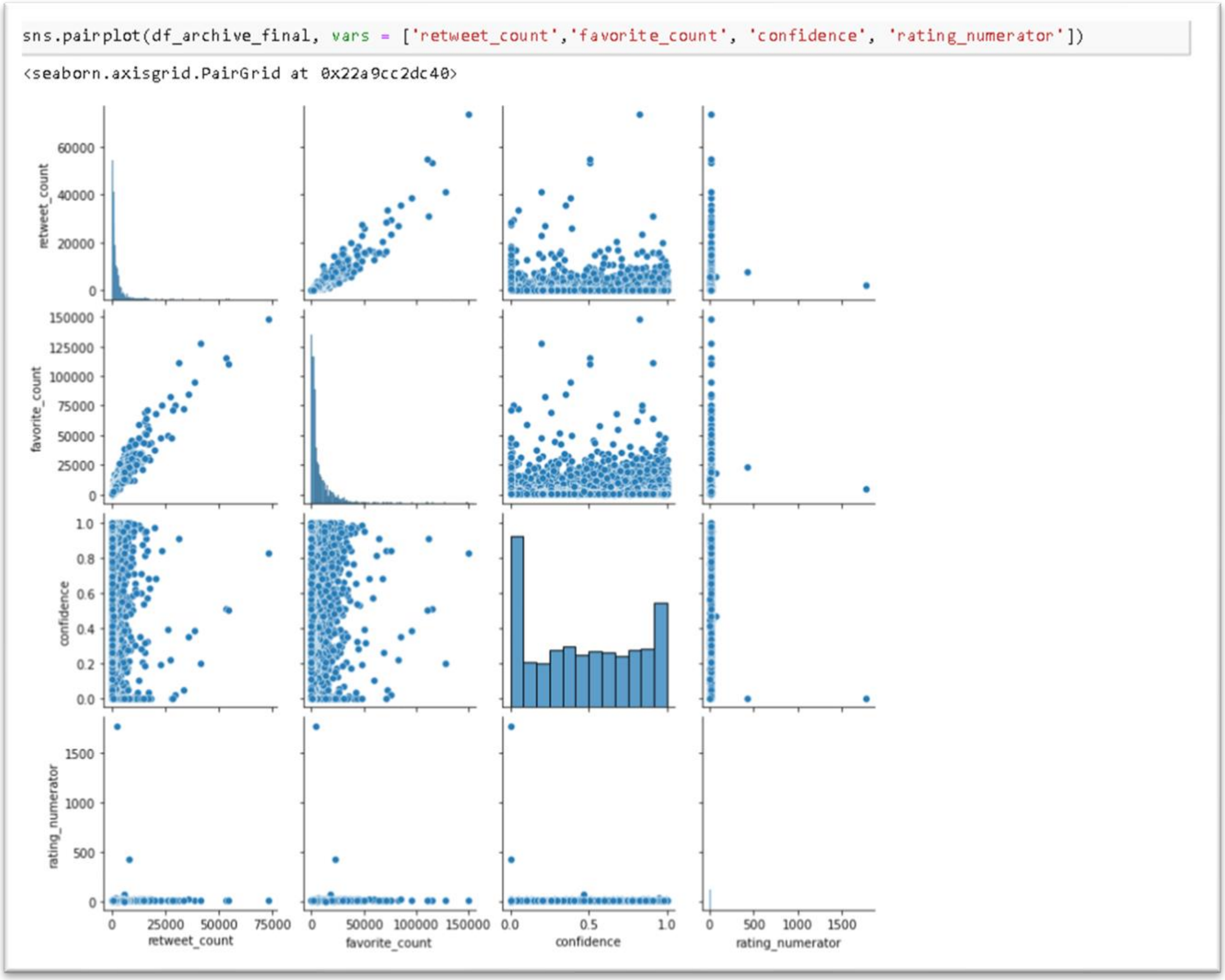
```
### to see the top 5 predicted dogs

df_archive_final[df_archive_final['breed'] != 'none']['breed'].value_counts().head(5).plot(kind='bar')
plt.xlabel('Dog Breed')
plt.ylabel('number of correct predictions')
plt.title('Top 5 dog breed predictions')
```

```
Text(0.5, 1.0, 'Top 5 dog breed predictions')
```

5. Correlation between various variables like retweet_count, favorite_count, confidence and rating_numerator

From the pair plot below, its understood that the retweet_count and favorite_count have positive correlation



```
sns.pairplot(df_archive_final, vars = ['retweet_count','favorite_count', 'confidence', 'rating_numerator'])
```

```
<seaborn.axisgrid.PairGrid at 0x22a9cc2dc40>
```