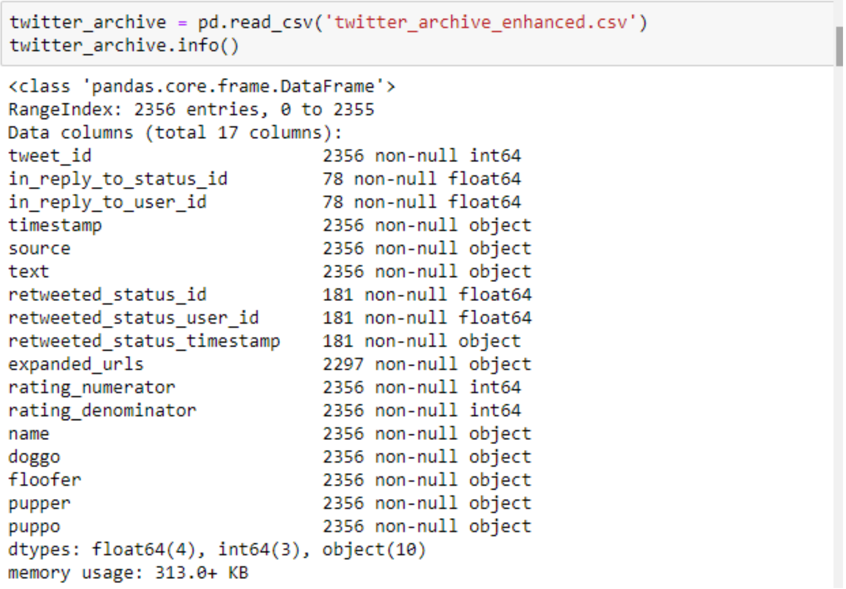
Act Report

WeRateDogs Data

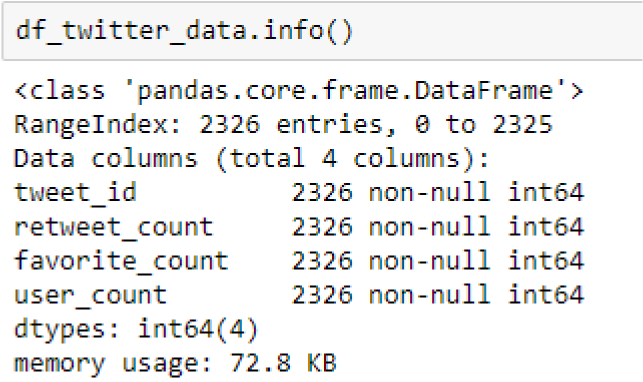
- The **WeRateDogs** Enhanced Twitter archive contains data extracted from 2356 of the 5000+ tweets from *@dog_rates* twitter account, posted between November 15, 2015, and August 1, 2017. This data consists of dog ratings that were taken from the text of the tweet along with the dog's name and dog stage if present.



```
twitter_archive = pd.read_csv('twitter_archive_enhanced.csv')
twitter_archive.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
rating_denominator            2356 non-null int64
name                          2356 non-null object
doggo                         2356 non-null object
floofer                       2356 non-null object
pupper                        2356 non-null object
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

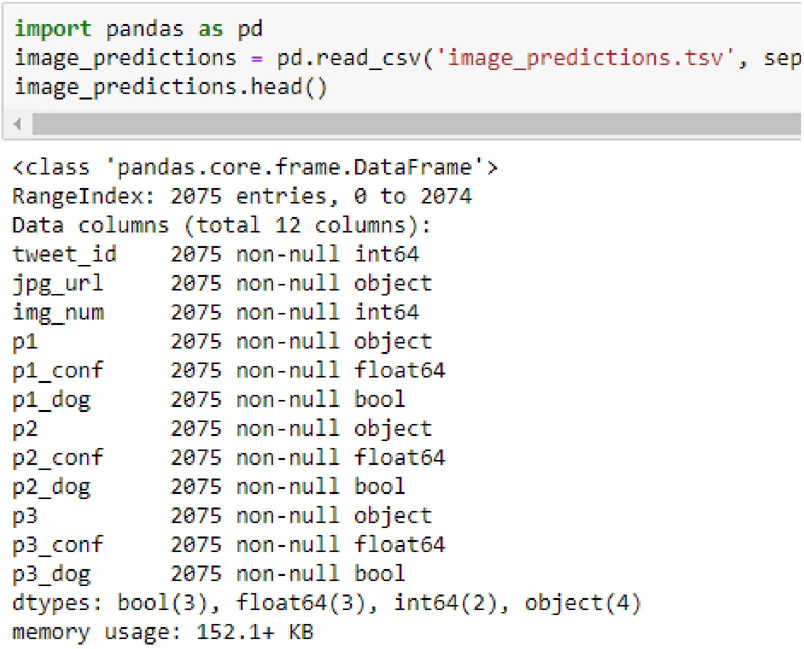
The retweet count and favourite count for each tweet were not included in the enhanced archive, and so I downloaded this additional information from the twitter account using the tweet ID from the archive.



```
df_twitter_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2326 entries, 0 to 2325
Data columns (total 4 columns):
tweet_id          2326 non-null int64
retweet_count     2326 non-null int64
favorite_count    2326 non-null int64
user_count        2326 non-null int64
dtypes: int64(4)
memory usage: 72.8 KB
```

Along with the Twitter data, I also downloaded an image predictions file from Udacity servers containing the top 3 predictions for dog breeds based on the image's tweets.



```
import pandas as pd
image_predictions = pd.read_csv('image_predictions.tsv', sep
image_predictions.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

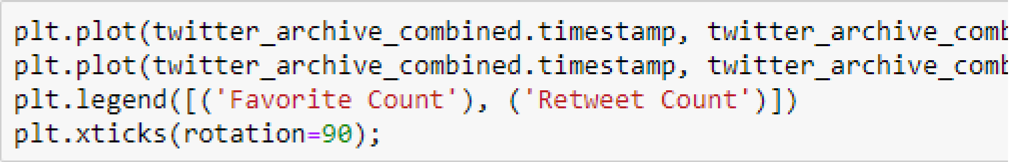
Wrangling Data

Before I could begin the analysis, the data had to be wrangled into shape to make it easier. I assessed the data both visually and programmatically for quality and tidiness; the quality of data is determined mainly by looking at several aspects or dimensions to ensure that it is complete, valid, accurate and consistent.

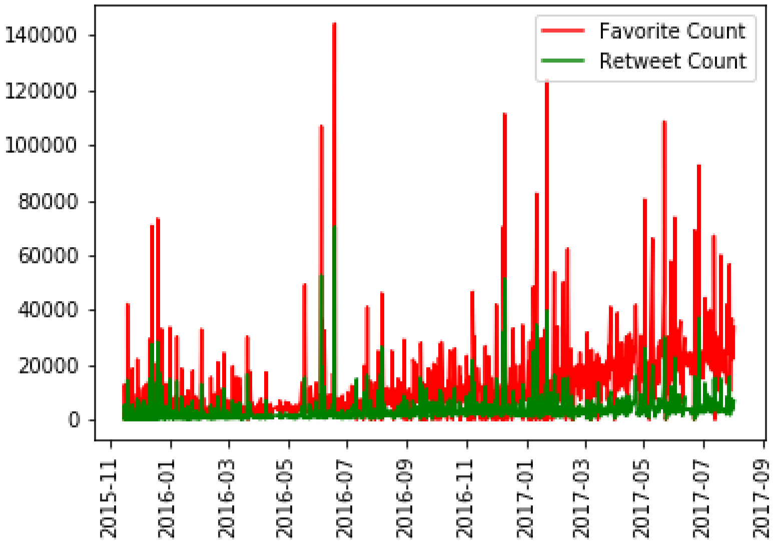
After cleaning the identified issues during my assessment, I came up with the below Insights

Insights

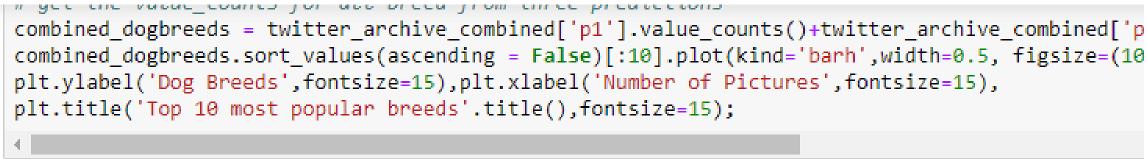
1. From the below graph you can see that the favorite and retweet counts were increasing over time.



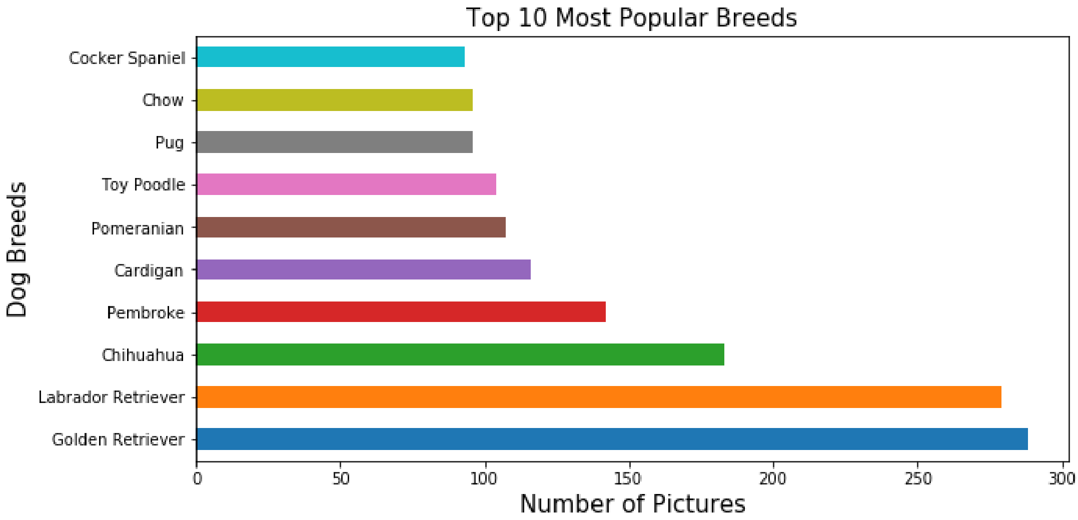
```
plt.plot(twitter_archive_combined.timestamp, twitter_archive_comb
plt.plot(twitter_archive_combined.timestamp, twitter_archive_comb
plt.legend([('Favorite Count'), ('Retweet Count')])
plt.xticks(rotation=90);
```



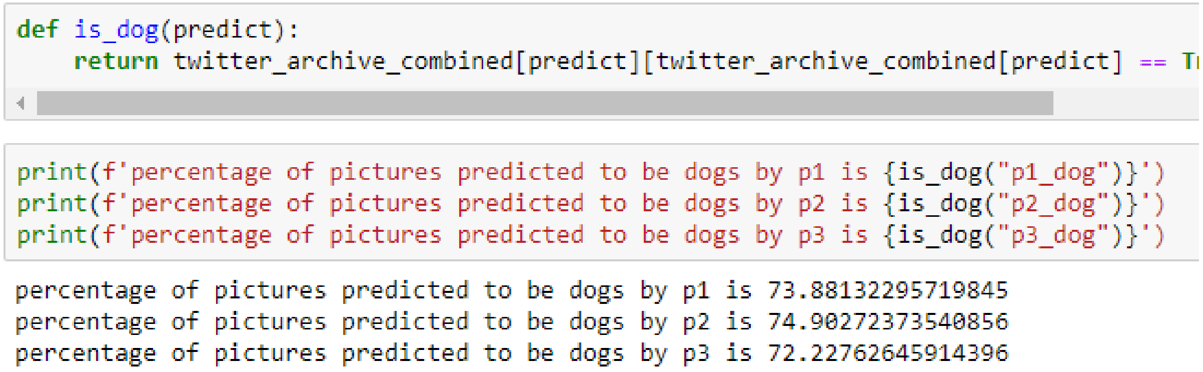
2. From the graph you can see that the Golden Retriever is one of the most popular dog breeds



```
# get the value_counts for all breed from three predictions
combined_dogbreeds = twitter_archive_combined['p1'].value_counts()+twitter_archive_combined['p
combined_dogbreeds.sort_values(ascending = False)[:10].plot(kind='barh',width=0.5, figsize=(10
plt.ylabel('Dog Breeds',fontsize=15),plt.xlabel('Number of Pictures',fontsize=15),
plt.title('Top 10 most popular breeds'.title(),fontsize=15);
```



3. On average it seems that the second prediction was more accurate in predicting whether a picture was a dog or not



```
def is_dog(predict):
    return twitter_archive_combined[predict][twitter_archive_combined[predict] == Tr
```

```
print(f'percentage of pictures predicted to be dogs by p1 is {is_dog("p1_dog")}')
print(f'percentage of pictures predicted to be dogs by p2 is {is_dog("p2_dog")}')
print(f'percentage of pictures predicted to be dogs by p3 is {is_dog("p3_dog")}')
```

```
percentage of pictures predicted to be dogs by p1 is 73.88132295719845
percentage of pictures predicted to be dogs by p2 is 74.90272373540856
percentage of pictures predicted to be dogs by p3 is 72.22762645914396
```