Act Report

After generating the Master dataframe, I was able to generate a number of questions to test through Pandas code and plots and to generate some insights on the data we have at hand.

The questions were:

- Is there a relationship between the retweet and favourite counts?
- Taking the retweet count as a measure, is there a relationship between the number of retweets and the rating?
- What are the types of dogs that are more frequently featured in the account based on the prediction algorithm (p1)?
- What types of dogs generally have higher rating (p1)?

Q1/ Is there a relationship between the retweet and favourite counts?

A scatterplot was developed to plot both the retweets and favorites count.

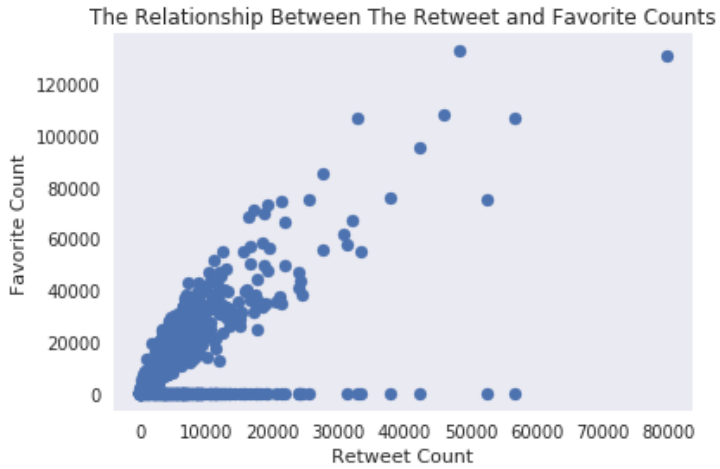
```
plt.scatter(master.retweet_count, master.favorite_count);

plt.title("The Relationship Between The Retweet and Favorite Counts");

plt.xlabel("Retweet Count");

plt.ylabel("Favorite Count");
```

The resulting chart showed a strong positive relationships between the count of retweets and favourites. Although there are outliers, where the favorite count is zero. Apparently this data extraction error since this is not normal on Twitter, normally the favorite count exceeds the retweet counts such as in our case.

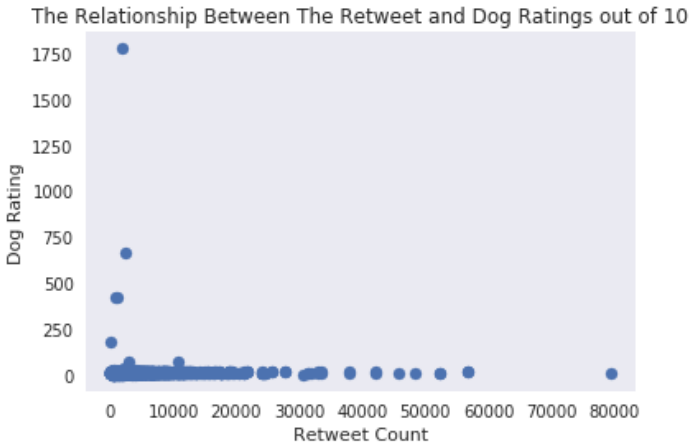


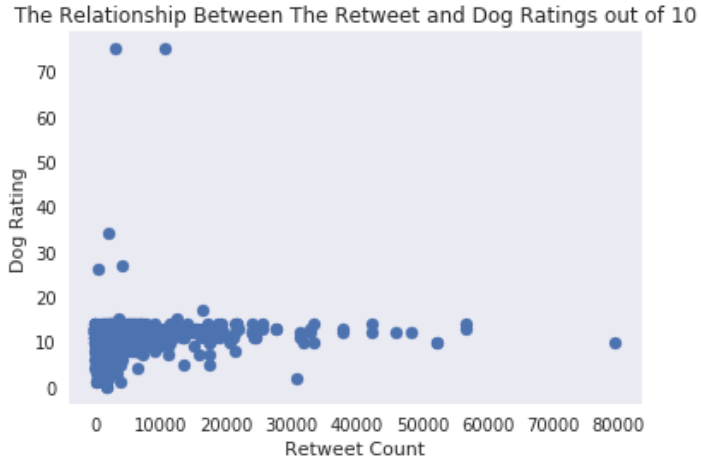
Q2/ Taking the retweet count as a measure, is there a relationship between the number of retweets and the rating?

A scatter plot was developed to accommodate the question purpose, with a small code. After seeing that there are outliers distorting the code

```
master_2=master.query('rating_numerator<100')
plt.scatter(master_2.retweet_count, master_2.rating_numerator);
plt.title("The Relationship Between The Retweet and Dog Ratings out of 10");
plt.xlabel("Retweet Count");
plt.ylabel("Dog Rating");
```

Apparently there is no strong relationship between the dog rating and retweet count. In fact the outliers with big rating appear to gather fewer retweets.

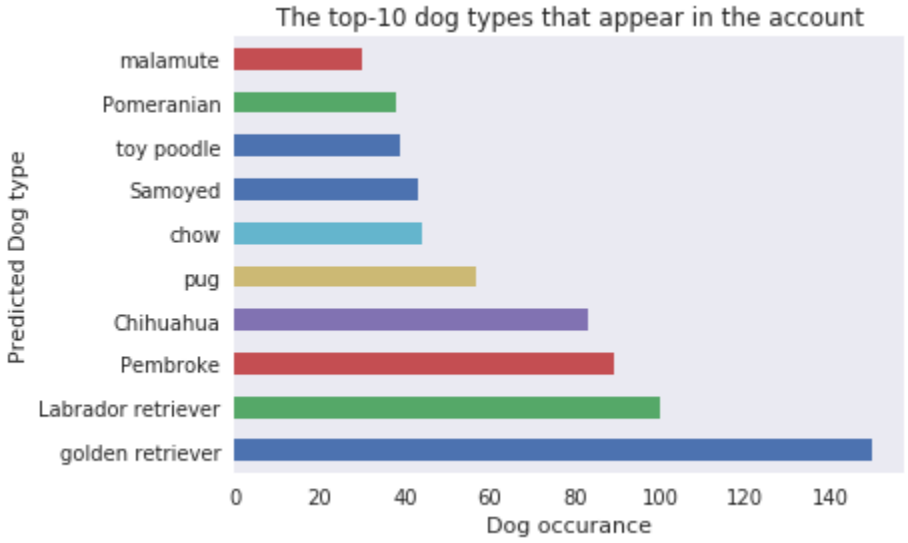




Q3/ What are the types of dogs that are more frequently featured in the account based on the prediction algorithm (p1)?

A horizontal bar chart was chosen for this analysis.

```
master.p1.value_counts().head(10).plot(kind='barh')
plt.title("The top-10 dog types that appear in the account");
plt.ylabel("Predicted Dog type");
plt.xlabel("Dog occurance");
```



From the graph, it appears that the retriever dogs are the one most ocurring in the account (mid-size dog, could be argued that they are most likly to meet in daily life), followed by small-size dogs (Pembroke, Chihuahua, Pug)

Q4/ What types of dogs generally have higher rating (p1)?.

A horizontal bar chart was chosen for this question.

```
master.groupby(['p1']).rating_numerator.mean().sort_values(ascending=False).
head(10).plot(kind='barh');

plt.title("The top-10 dog types in average rating as predicted");

plt.ylabel("Predicted Dog type");

plt.xlabel("Average rating score");
```

There are a number of predicted values here that do not appear to be a type of dog (Orange, Refrigerator, Crane). Since there is no simple programatic way of cleaning these, we will leave them as-is. Other than that the Blenhein Spaniel, Barbancon Griffon, and Airedale dogs are amongsth the highset rated dogs on average (although they are outliers since most ratins do not exceed 25)