Amanda Zhou
August 28th, 2020

Project: Wrangle and Analyze Data

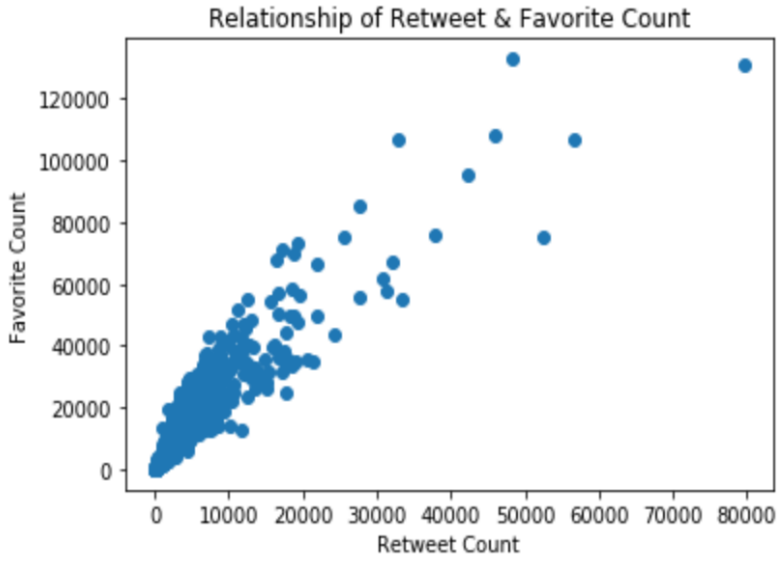
(External Document)

Introduction

WeRateDogs' data have been wrangled. This report will communicate the insights and display the visualizations produced from the data.

Insights and Visualizations

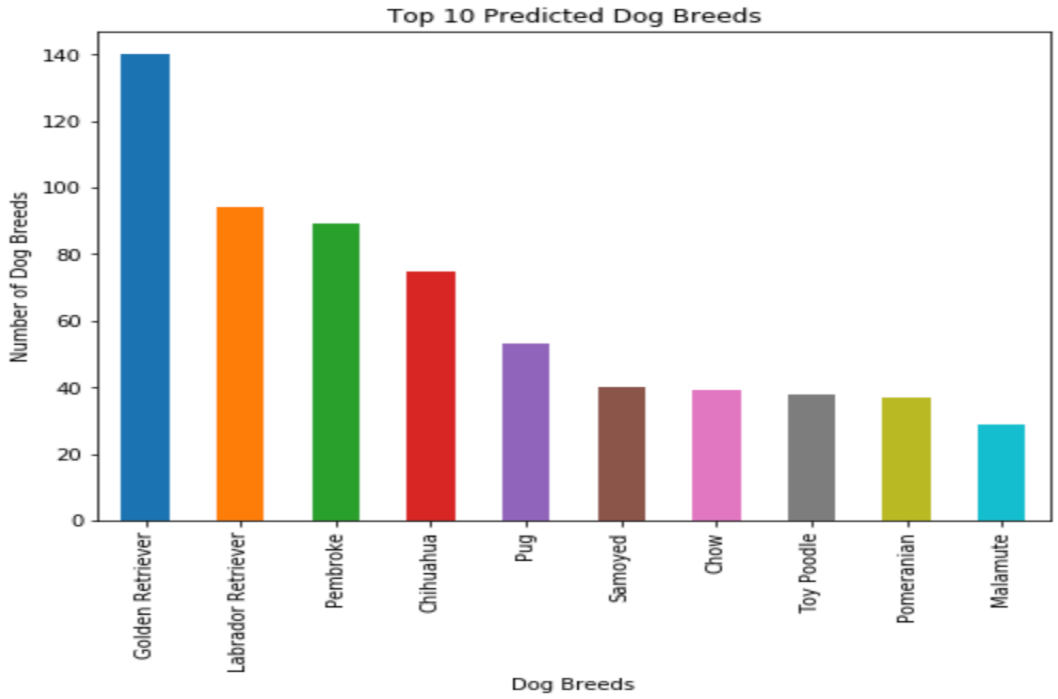
Finding 1: Relationship Between Retweet Count and Favorite Count



Summary: From the plot, it shows a highly linear postive relationship between Retweet Count and Favorite Count.

Amanda Zhou
August 28th, 2020

Finding 2: Top 10 Predicted Dog Breeds



Summary: From the chart, Golden Retriever has the highest number, following by Labrador Retriever and Pembroke.

Finding 3: Dog Breed and Retweet and Favorite Counts

```
df_count.sort_values('retweet_count', ascending = False).head(3)
```

	breed_pred	retweet_count	favorite_count
775	Labrador Retriever	79515	131075
400	Chihuahua	56625	107015
810	Eskimo Dog	52360	75163

```
df_count.sort_values('favorite_count', ascending = False).head(3)
```

	breed_pred	retweet_count	favorite_count
309	NaN	48265	132810
775	Labrador Retriever	79515	131075
58	English Springer	45849	107956

- **By checking each tweet, the one has the most retweet counts and favorite counts is Labrador Retriever**

```
#check which Top 3 breed have the most retweet_count
df_count_sum.sort_values('retweet_count', ascending = False).head(3)
```

breed_pred	retweet_count	favorite_count
Golden Retriever	484238	1677881
Labrador Retriever	366440	1123354
Pembroke	281669	1001934

```
#check which Top 3 breed have the most favorite_count
df_count_sum.sort_values('favorite_count', ascending = False).head(3)
```

breed_pred	retweet_count	favorite_count
Golden Retriever	484238	1677881
Labrador Retriever	366440	1123354
Pembroke	281669	1001934

- **By checking the sum of each dog breed, the one has the most retweet counts and favorite counts is Golden Retriever**

```
#check which Top 3 breed have the most retweet_count
df_count_mean.sort_values('retweet_count', ascending = False).head(3)
```

breed_pred	retweet_count	favorite_count
Afghan Hound	8017.500000	22451.000000
Standard Poodle	6631.857143	15786.000000
English Springer	5973.555556	15657.222222

```
#check which Top 3 breed have the most favorite_count
df_count_mean.sort_values('favorite_count', ascending = False).head(3)
```

breed_pred	retweet_count	favorite_count
Saluki	5133.750000	24060.00000
Afghan Hound	8017.500000	22451.00000
French Bulldog	4576.185185	17468.37037

- **By checking the mean of each dog breed, the one has the most retweet counts is Afghan Hound, and has the most favorite counts is Saluki**

Amanda Zhou
August 28th, 2020

Summary: When checking each simple tweet, Labrador Retriever is the one has the most retweet counts and favorite counts. When looking at the sum of retweet counts and favorite counts for each dog breed, Golden Retriever has the most in both, then it is Labrador Retriever. However, when checking by the mean of retweet counts and favorite counts for each dog breed, the result has big difference due to sample sizes of each bread are different

Finding 4: Dog Breed and Rating

	tweet_id	retweet_count	favorite_count	rating_numerator	rating_denominator	img_num
count	1.994000e+03	1994.000000	1994.000000	1994.000000	1994.0	1994.000000
mean	7.358508e+17	2766.753260	8895.725677	11.692076	10.0	1.203109
std	6.747816e+16	4674.698447	12213.193181	40.670663	0.0	0.560777
min	6.660209e+17	16.000000	81.000000	0.000000	10.0	1.000000
25%	6.758475e+17	624.750000	1982.000000	10.000000	10.0	1.000000
50%	7.084748e+17	1359.500000	4136.000000	11.000000	10.0	1.000000
75%	7.877873e+17	3220.000000	11308.000000	12.000000	10.0	1.000000
max	8.924206e+17	79515.000000	132810.000000	1776.000000	10.0	4.000000

```
df_rate_mean = df[['breed_pred','rating_numerator']].groupby('breed_pred').mean()
df_rate_mean.sort_values('rating_numerator', ascending = False).head(10)
```

breed_pred	rating_numerator
Clumber	27.000000
Afghan Hound	13.000000
Pomeranian	12.891892
Saluki	12.500000
Briard	12.333333
Tibetan Mastiff	12.250000
Kuvasz	12.062500
Standard Schnauzer	12.000000
Toy Terrier	12.000000
Scottish Deerhound	12.000000

Summary: As we know that all the denominators are 10, if the numerator is equal to or greater than 10, it means the dog is given a rating of above 100% overall. From the analysis above, the mean of rating_numerator is 11.692076, and majority of dog breads have a high rating (i.e. rating_numerator is at least 10 or overall above 100%). Even though there are few outliers in the dataset, we can neglect it at this moment. In general, Golden Retriever and Labrador Retriever are given a good rating.