Who’s the best Doggo!!

So, after cleaning the data as explained in our wrangle_report file, it is time to extract some insights, with some visualizations!!.

Namely, we will investigate how dog breeds “identified by the algorithm” fair against 3 metrics

1. Rating “rating numerator”
2. Favorite count
3. Retweet count

And whether the rating numerator correlates with the other two metrics.

First: Filter out the data that the algorithm misclassified as non dogs:

In some cases, the algorithm that provided the image prediction algorithm didn't correctly classify some pictures as dogs. So we will filter them out.

```
dogs_algorithm = tweet_df.query("prediction_1_dog == True")
```

Second: Investigate the existence of correlation between rating and either of retweets or favorites

We can do this using the correlation matrix as follows:

```
dog_df.corr().rating_numerator
```

Which outputs the following result:

```
favorite_count        0.021619
retweet_count         0.022603
rating_numerator      1.000000
img_num              -0.000328
prediction_1_conf    -0.009290
prediction_2_conf    -0.013584
prediction_3_conf    -0.004347
Name: rating_numerator, dtype: float64
```

So there is a very weak positive correlation between the rating and both the number of retweets and the number of favorites
So using the rating metric will not cause redundancy

Third: group by dog breed, then aggregate by the mean

Now that we have our filtered data, we can safely group by the dog breed, and aggregate by the mean. This will result in a dataframe with index as the dog breeds, and columns as any numeric column. We can specify the columns of interest: rating_numerator, favorite_count, retweet_count

```
dog_breed_means=dogs_algorithm.groupby("prediction_1").mean().loc[:,["favorite_count", "retweet_count", "rating_numerator"]]
```

The following image shows the first 5 columns of this dataframe:

prediction_1	favorite_count	retweet_count	rating_numerator
Afghan_hound	15500.666667	5102.000000	9.666667
Airedale	4739.833333	1197.333333	9.833333
American_Staffordshire_terrier	5587.833333	1615.666667	10.833333
Appenzeller	6585.500000	1236.000000	11.000000
Australian_terrier	10009.500000	2670.000000	11.500000

Fourth: Sort and Plot !

In order to give the bar plots a nicer look and feel, it is favorable to sort the values descendingly before plotting.
To do that, we used the sort_values function as follows:

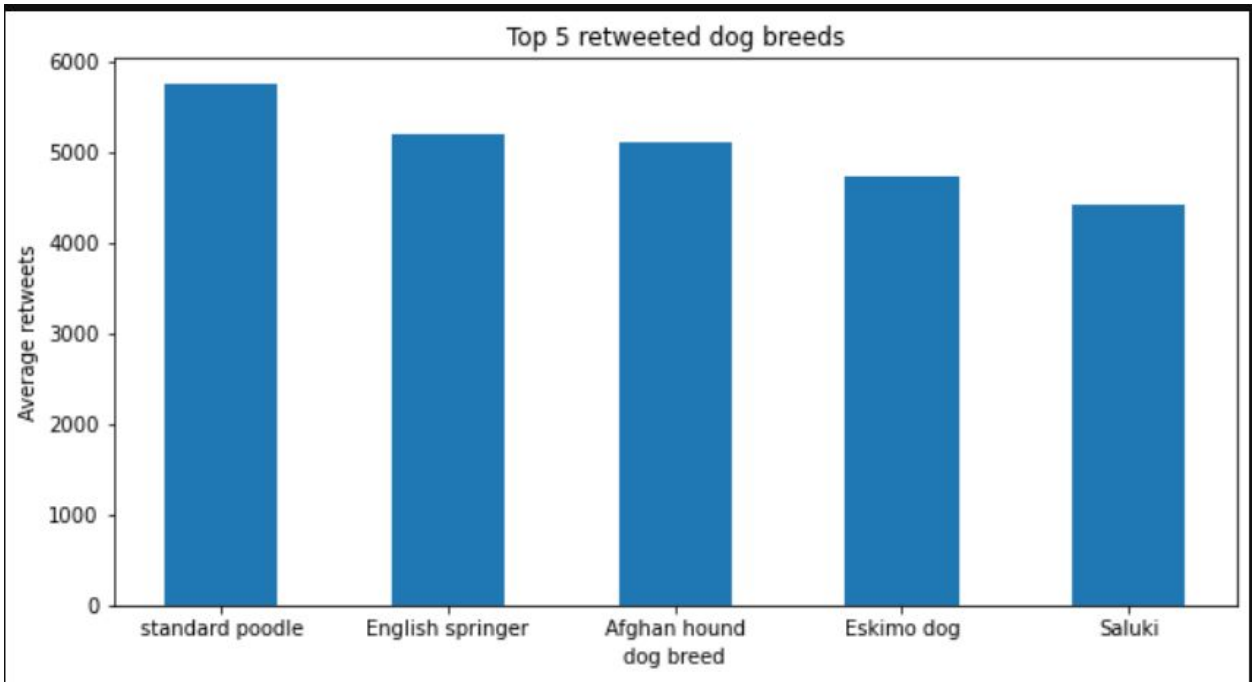
```
dataframe.sort_values(by="retweet_count", ascending=False)
```

In addition, some dog breed names have underscores, which is not pretty for displaying insights. We can replace them with spaces as follows:

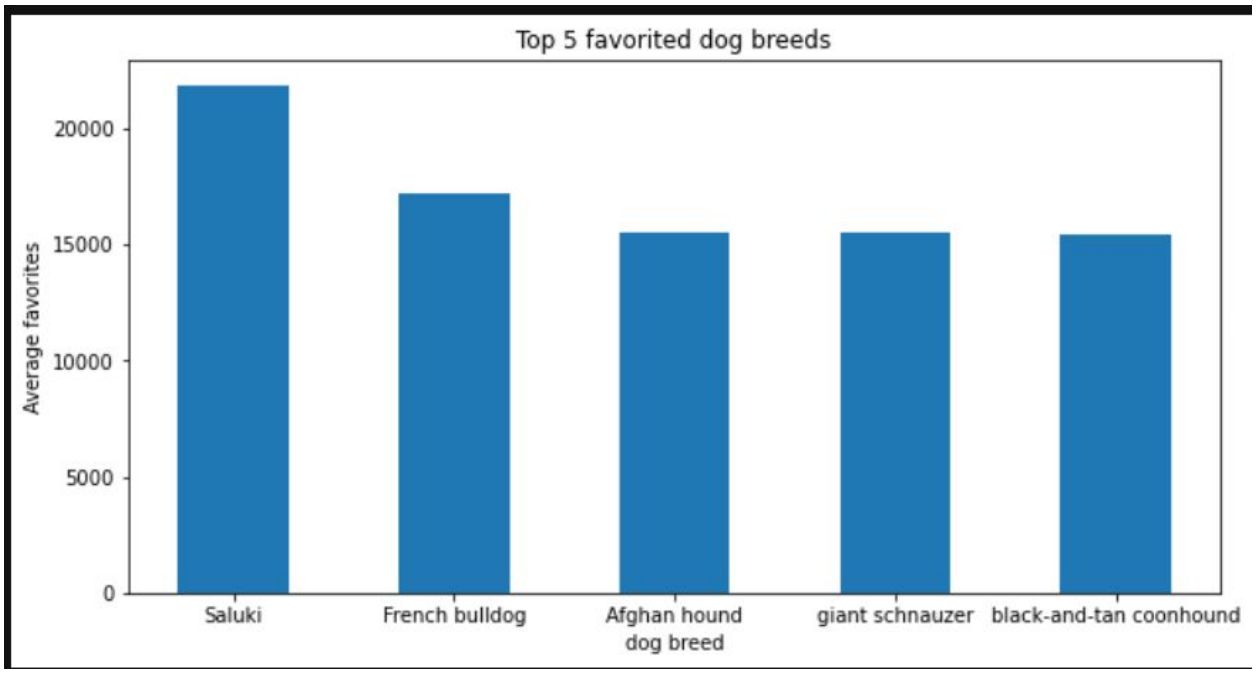
```
columnname.str.replace("_", " ")
```

Now we will find out the top 5 champions in each category!!!

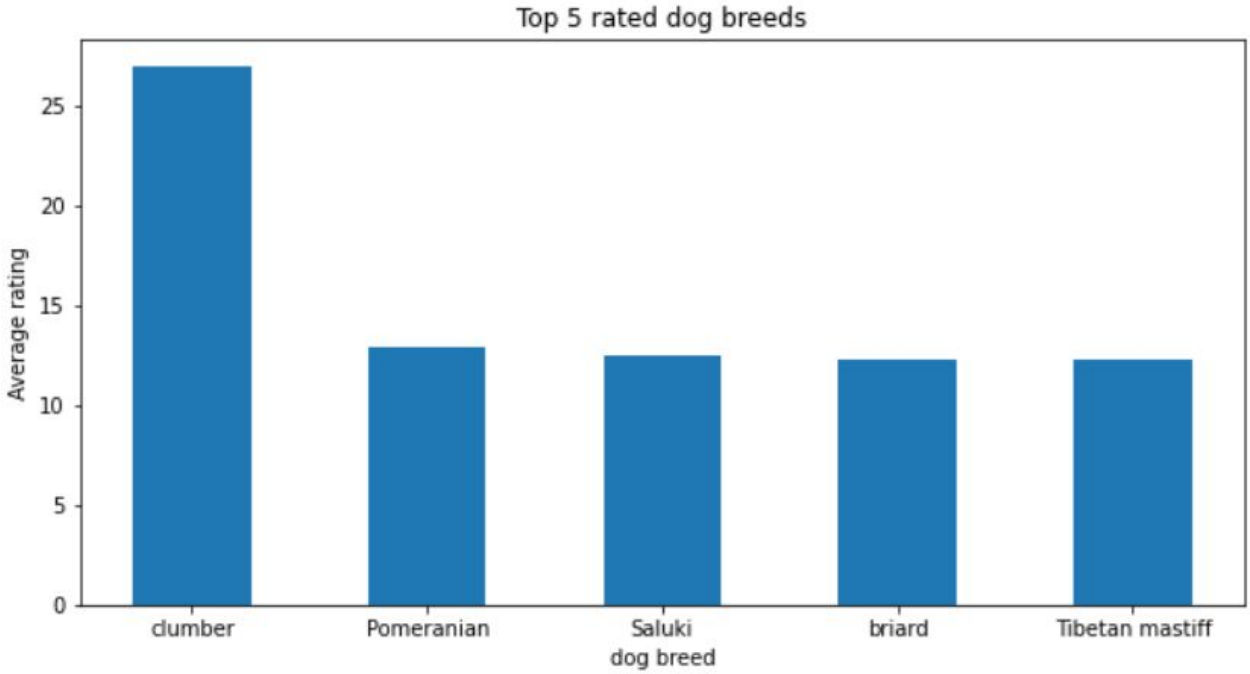
Retweet Champions:



Favorite Champions:



Rating champions:



Reflection:

From the previous sets of champions, one dog breed stands out, which is the **Saluki** breed. It made it to the top 5 in retweets, favorites and ratings!

Pretty fascinating indeed!