

Report: act_report

This current notebook is organised into the following sections:

- [Introduction](#)
- [Accessing Data](#)
- [Insights](#)
 1. [insight 1](#)
 2. [insight 2](#)
 3. [insight 3](#)
- [Conclusion](#)

Introduction

- This document briefly communicates some insights and visuals from the wrangled data processed from the `wrangle_act` notebook.

Accessing the wrangled data

In order to generate insights and create visual, it is important that we read in some dependencies and our master 'CSV' file. Tools used for this process are:

1. Pandas
2. Numpy
3. Matplotlib

```
In [12]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

df = pd.read_csv('twitter_archive_master.csv')
```

Insights

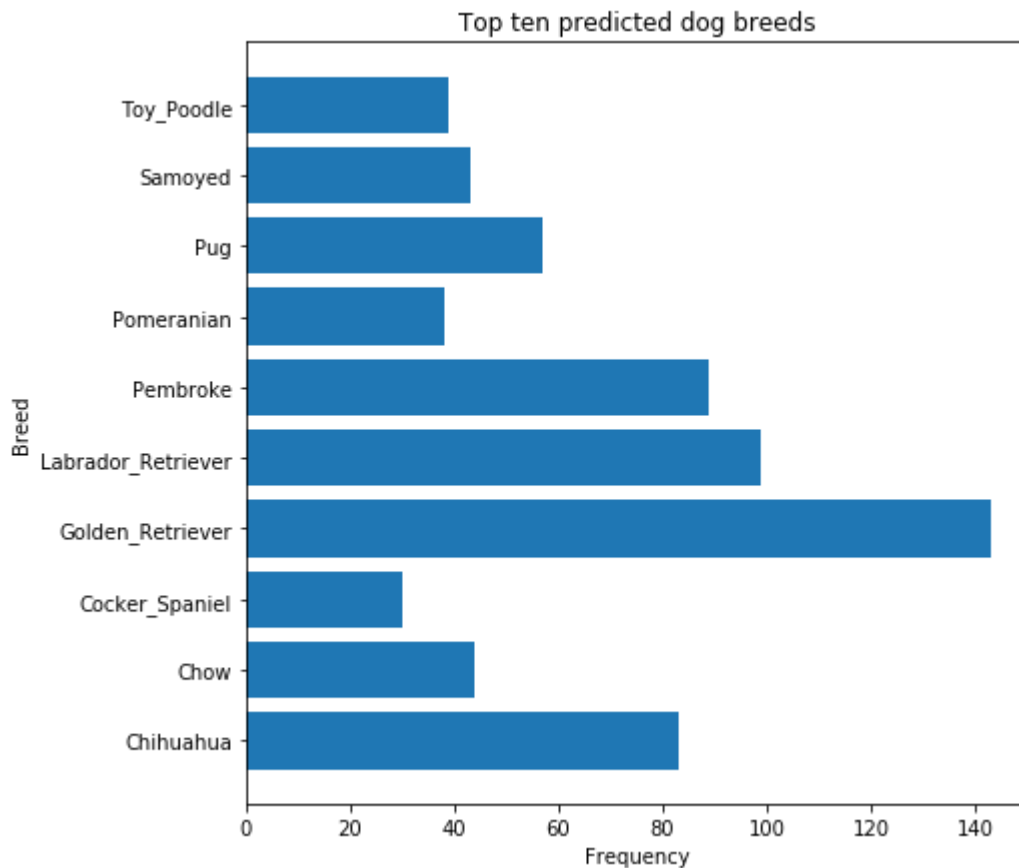
- Most frequent dog breeds predicted by the neural network.
- Most common dog stage present in the dataset.
- Exploring a possible relationship between retweets and likes in the dataset.

Insight 1

Most frequent dog breeds predicted by the neural network.

```
In [3]: top_dog_breed = df.p1.value_counts().head(10)
```

```
In [4]: plt.figure(figsize=(7,7))
y = top_dog_breed.index
x = top_dog_breed
plt.barh(y,x)
plt.title("Top ten predicted dog breeds")
plt.ylabel('Breed')
plt.xlabel('Frequency');
```



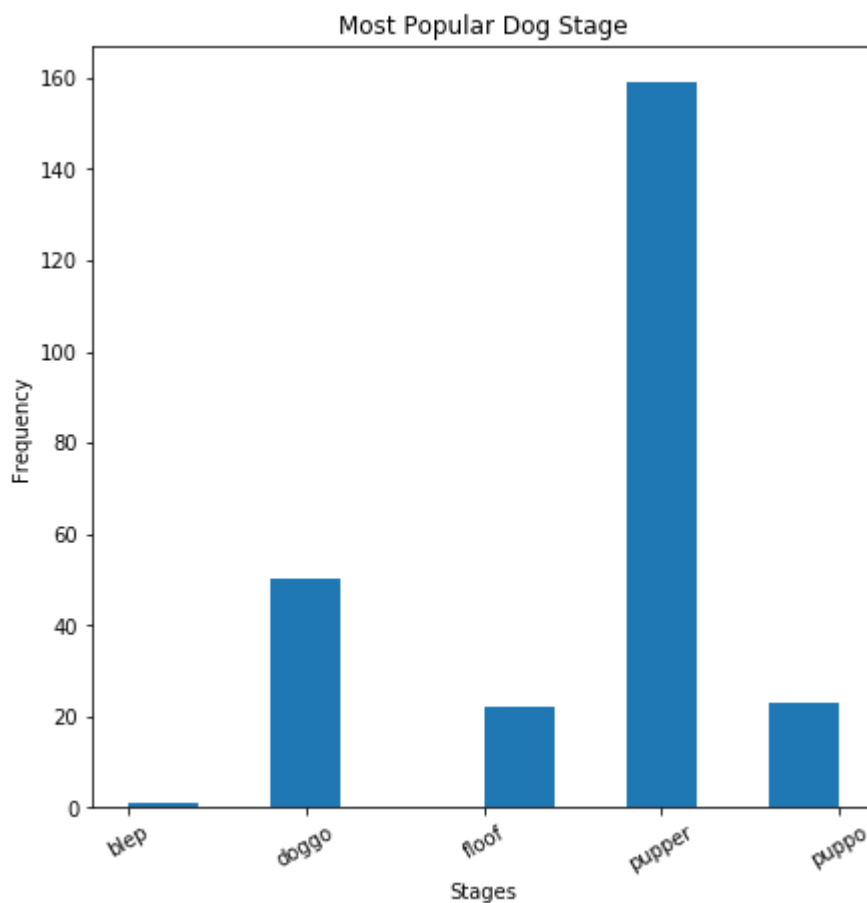
Of the top Ten most predicted dog breeds in our dataset, Golden Retriever tops the list

Insight 2

Most common dog stage present in the dataset.

```
In [5]: stages = list(df[~(df.stage.isna())]['stage'])
```

```
plt.figure(figsize=(7,7))
plt.hist(stages)
plt.title("Most Popular Dog Stage")
plt.ylabel('Frequency')
plt.xlabel('Stages')
plt.xticks(rotation = 30);
```

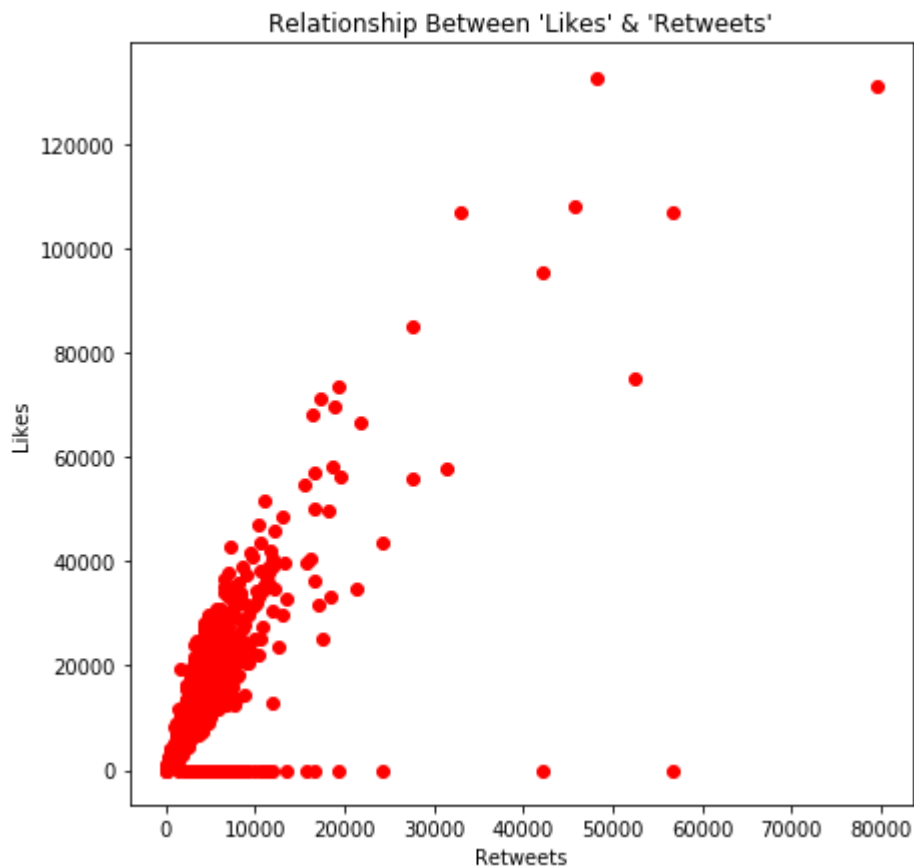


The chart above displays the frequency of dog stages present in our data set. Pupper appears to have the most occurrence. In other words, more dogs are in the pupper stage compared to other dog stages.

Insight 3

Exploring a possible relationship between retweets and likes in the dataset.

```
In [14]: x = df.retweets
y = df.number_of_likes
plt.figure(figsize=(7,7))
plt.scatter(x,y, c = 'red')
plt.title("Relationship Between 'Likes' & 'Retweets'")
plt.ylabel('Likes')
plt.xlabel('Retweets');
```



As seen in the scatter plot above, there appears to be a strong positive relationship between the `number_of_likes` and `retweets` columns in our dataset. However, there are notable presence of several outliers.

Conclusion

The `twitter_archive_master.csv` file was used to generate and communicate insights for this project. Some of the notable insights included in this document include: relationships between the `number_of_likes` and `retweets` and most frequent dog breeds predicted by the neural network. There is room for further analysis to understand other variables such as the ranking column in the dataset.