# Project 5: Wrangle and Analyze Data (Act Report)

## Data Analyst Nanodegree Project

**Date 19th August 2020**

**By Wisatat Komwatcharapong**

# Act Report (by Wisatat Komwatcharapong)

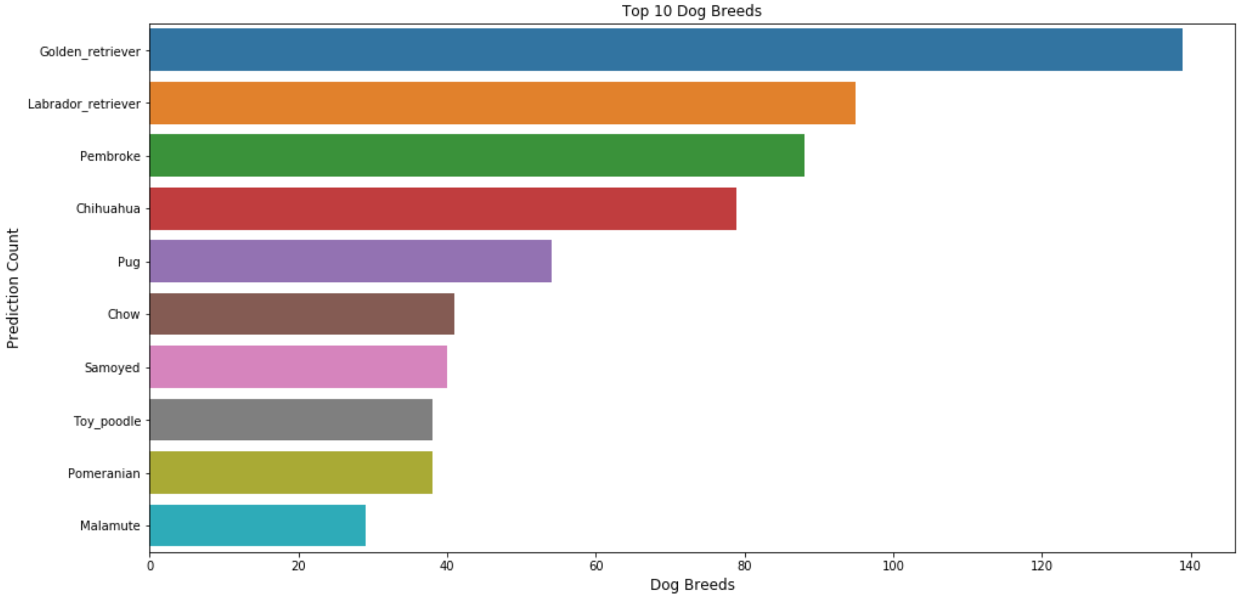
## Introduction

This report is part of the wrangle and analyze data project from the Udacity Data Analysis Nanodegree program. The dataset that used for this project is from twitter user @dog_rates, also known as WeRateDogs, indeed rates people's dog with diverting comment about the dog. These ratings almost always have a denominator of 10. Nonetheless, the numerators are almost always greater than 10! Generally, ratings should be 1 to 10. However, they admit almost dogs deserve a 10 and more than that!

Therefore, we will use this dataset to analyze various insights and display the results with pictures according to the following.

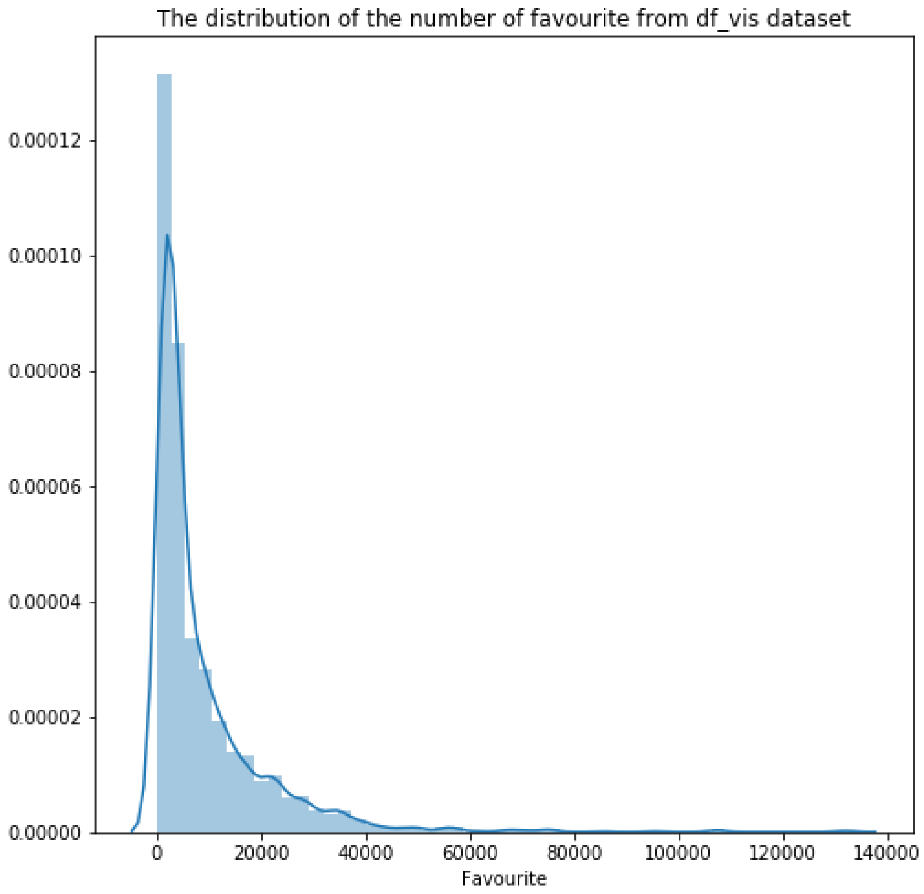
## Top 10 of the most popular dog breeds

I have analyzed this dataset and got results like this figure.



The Golden Retriever is the most popular dog breed with 139 favorites, Labrador Retriever and Pembroke are 2nd and 3rd popular respectively.

## The distribution of favorite from dataset.

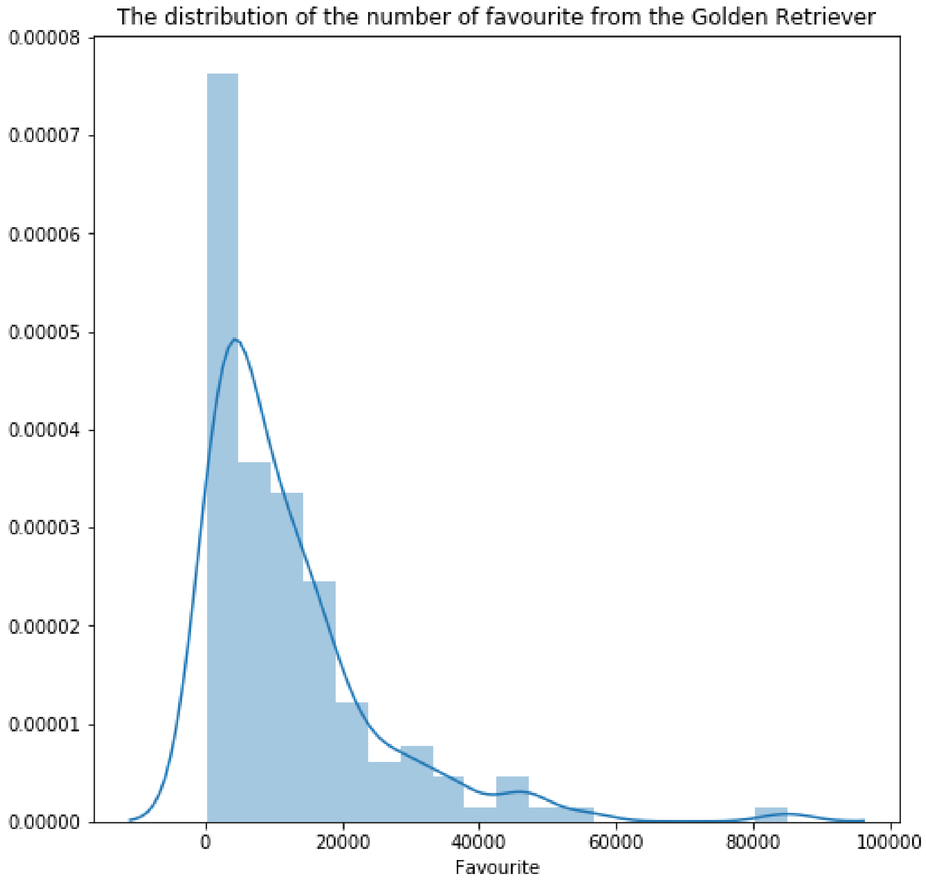


From the figure, it shows that there is a right skew distribution. We can find other detail values from the describe() command, which in the following results:

```
count    1995.000000
mean     8891.310777
std      12211.722542
min      81.000000
25%      1976.000000
50%      4134.000000
75%      11306.000000
max      132810.000000
Name: favorite_count, dtype: float64
```

## The distribution of favorite from Golden Retriever

We can analyze similarly to plot the distribution as in the previous article, but we must choose a specific dog breed for the Golden Retriever.

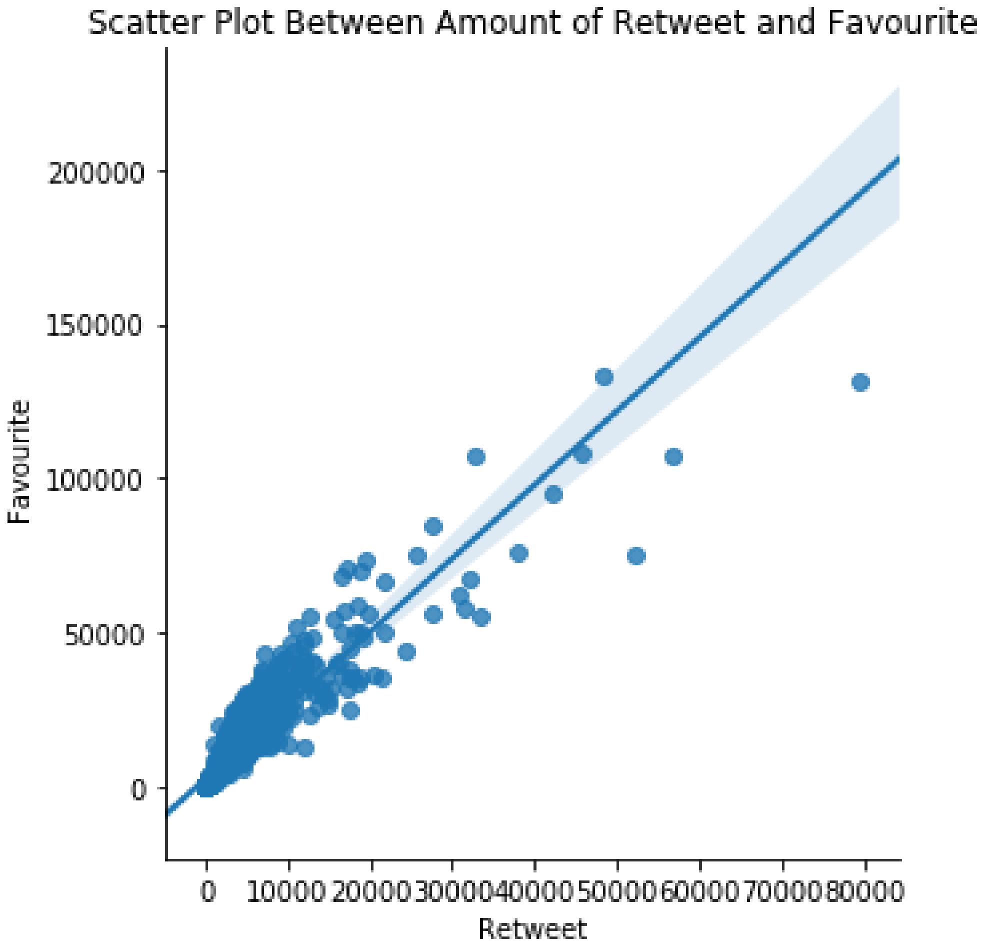


From the figure, it shows that there is a right skew distribution. We can find other detail values from the describe() command, which in the following results:

```
count     139.000000
mean      12205.949640
std       12986.086286
min       198.000000
25%       3554.500000
50%       8046.000000
75%       16247.000000
max       85011.000000
Name: favorite_count, dtype: float64
```

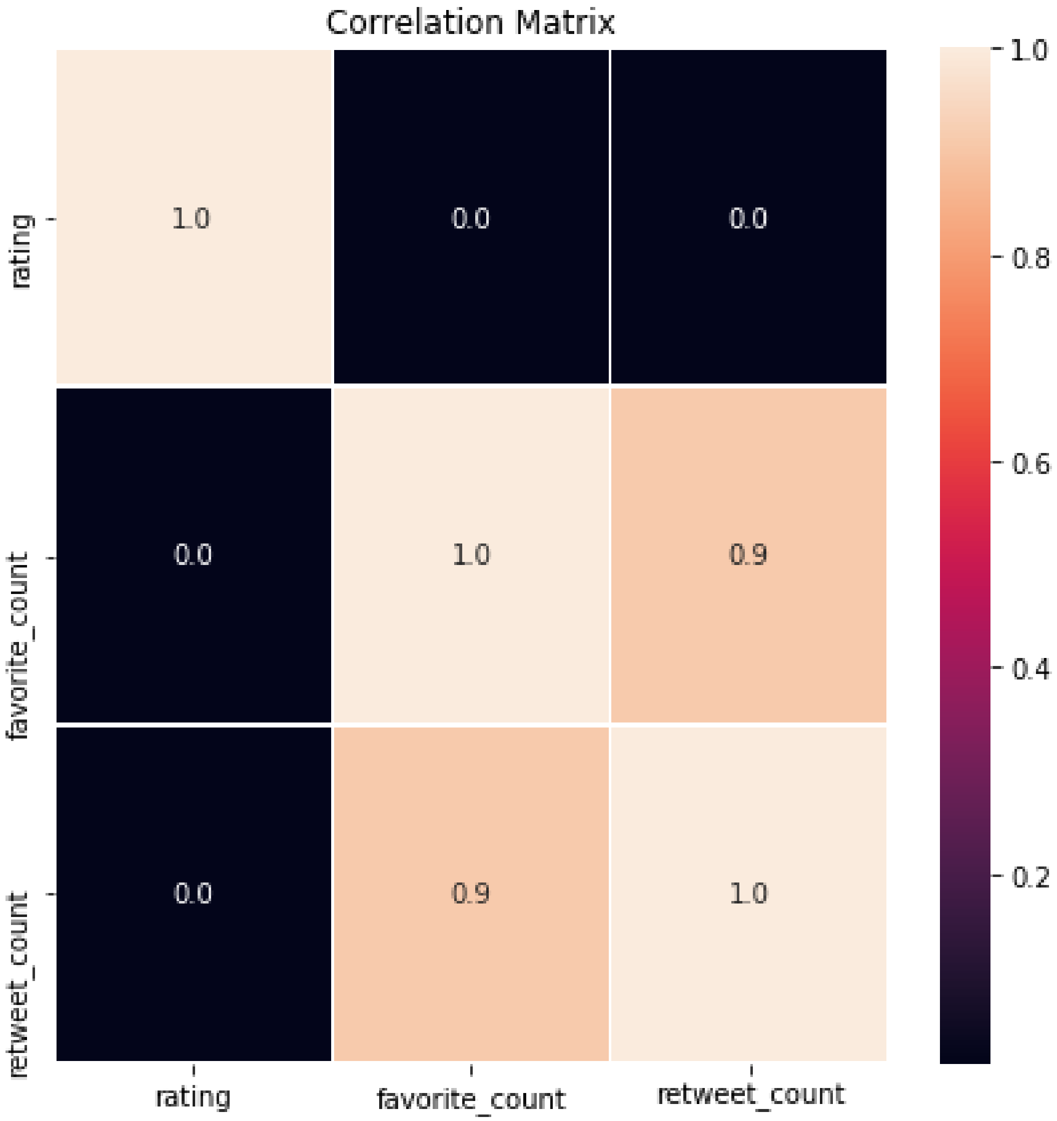
## Find the relationship between rating and retweet by scatter plot.

We analyze how the data is distributed and look at the trends to each other.



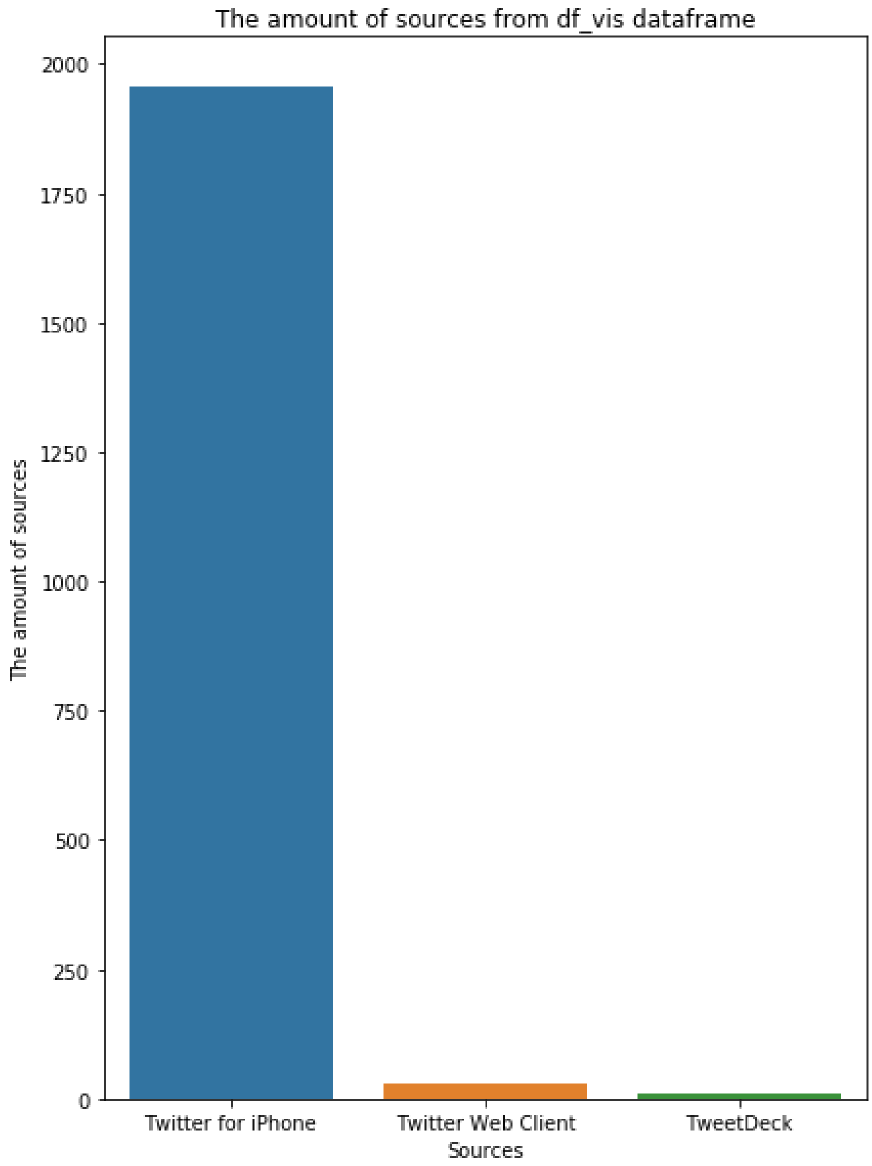
From the figure, it shows that favorite and retweet are positively correlated, and they are mostly around 40000 favorite and 15000 retweet.

**Find the relationship between retweet, favorite and rating by heatmap (Correlation Matrix).**



From the figure, it can be seen that retweet and rating, favorite and rating, rating and retweet are not correlated. On the other hand, favorite and retweet is huge correlated.

## How many sources came from by bar chart



From the figures, it can be seen that Twitter from iPhone is far more numerous than other sources. However, we can analyzed by value_counts() command to show, , which in the following results:

```
Twitter for iPhone    1956
Twitter Web Client    28
TweetDeck             11
Name: source, dtype: int64
```