أكاديمية مسك
MISK ACADEMY

UDACITY

Act Report

Data Analysis Nano Degree – DAND

Prepared by:

Abdulmajeed Alharbi

alharbimabd@gmail.com

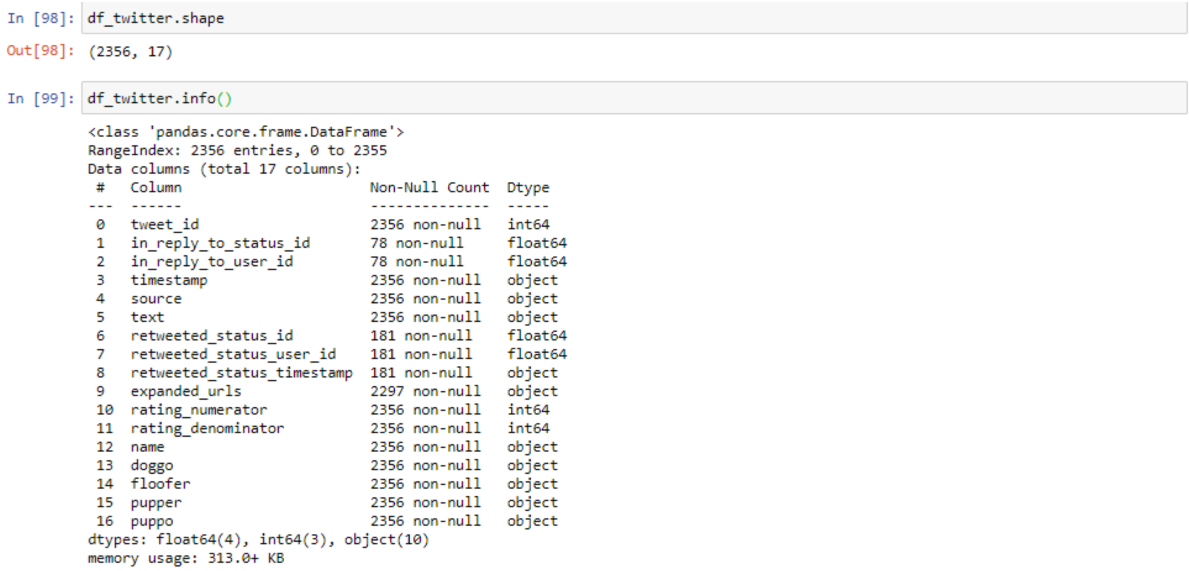
Data Analysis Nano Degree Program – 23 July of 2020

Act Report

This dataset that we are wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

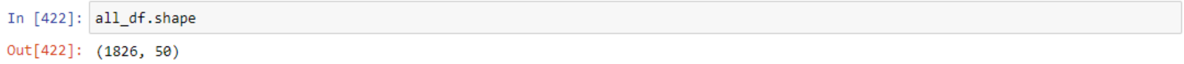
In this report I will communicates the insights and displays the visualizations in my analysis, in this part you will see how much the data wrangling process is very important and how the data wrangling impact your result.

Our dataset include 2356 rows but after cleaning and wrangling the data we have 1928 rows left, see the attachment below:



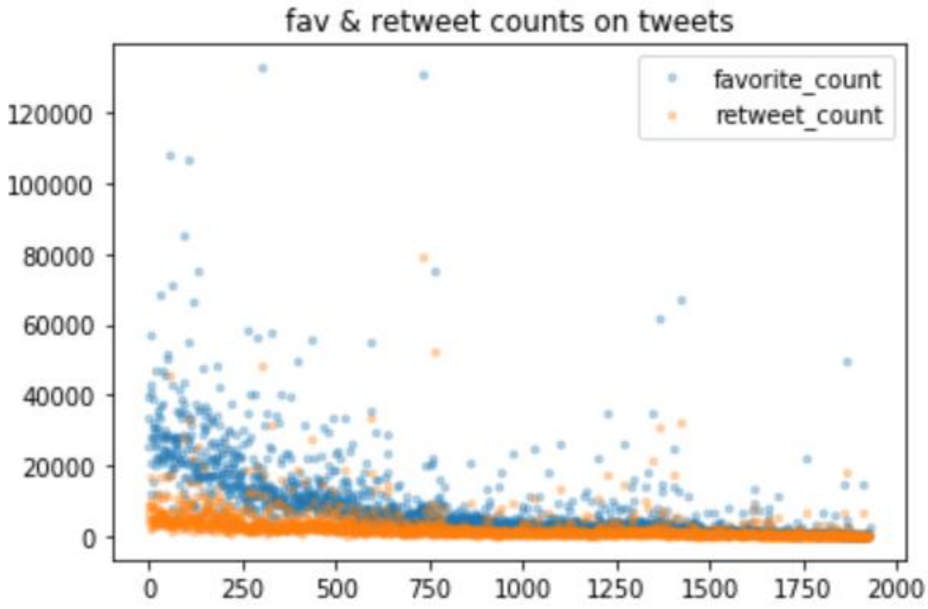
```
In [98]: df_twitter.shape
Out[98]: (2356, 17)

In [99]: df_twitter.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

And after cleaning and merging all dataset into one dataset:

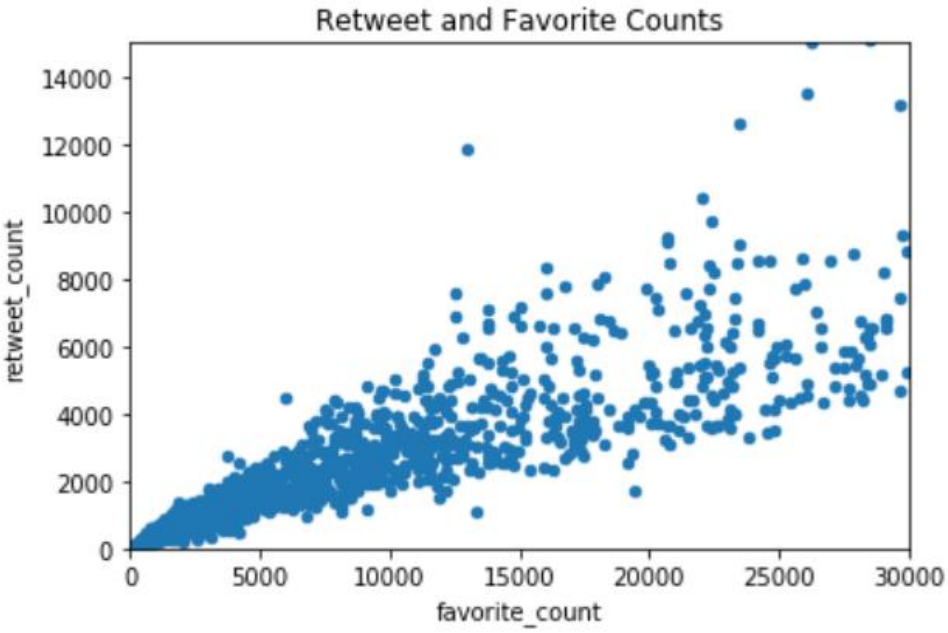


```
In [422]: all_df.shape
Out[422]: (1826, 50)
```

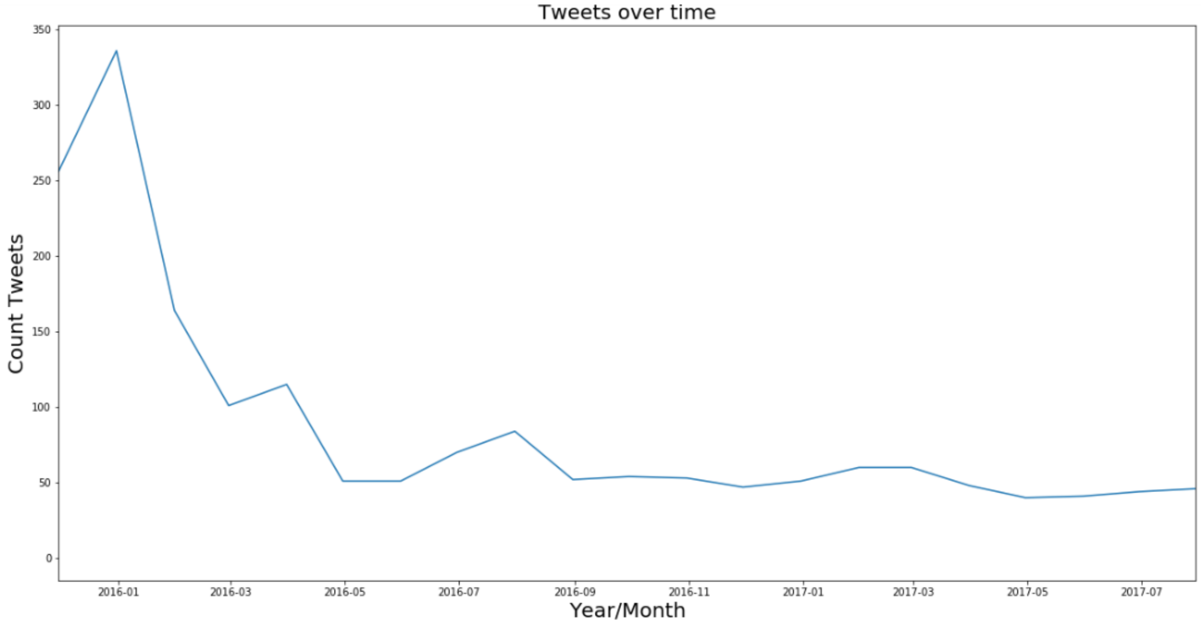
You can notice that retweet and favorite have good relationship, so in the below we will see the counts of retweet and favorite on tweets and we can see the counts of favorite become more than retweets because a lot of people make a like on this rating to keep it with his accounts and to come back again on tweet if he want.



- from above, we can see that there is a relationship between favoriie and retweets counts, so once the number of retweet increase the number of favorite also will increase, you can also check this insigths from below chart:

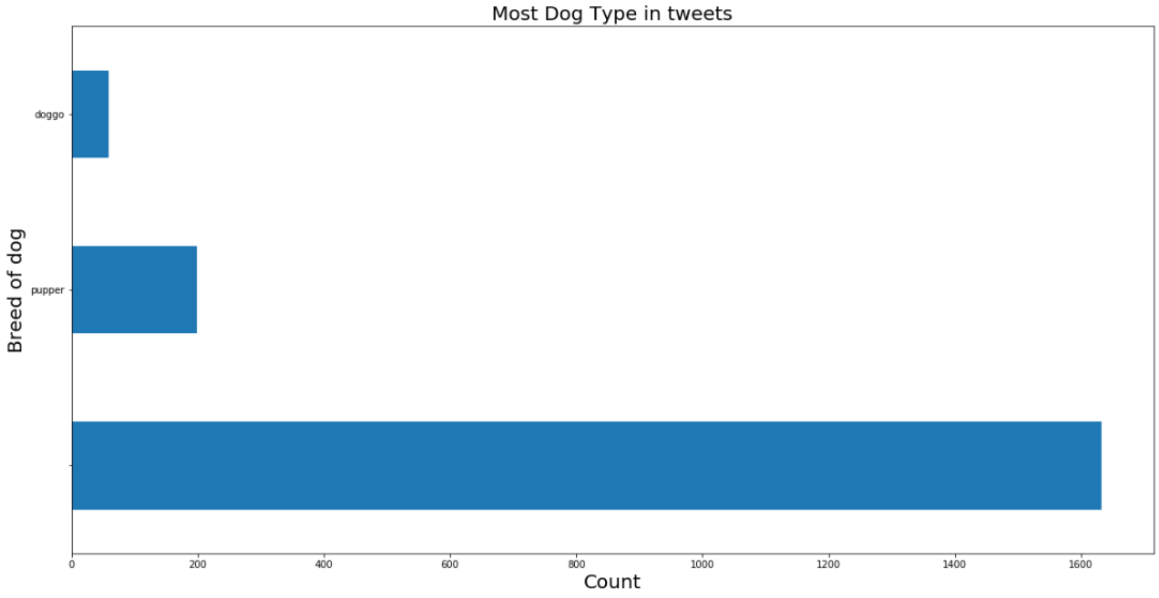


In the next visualization you will see the counts of tweets over the time

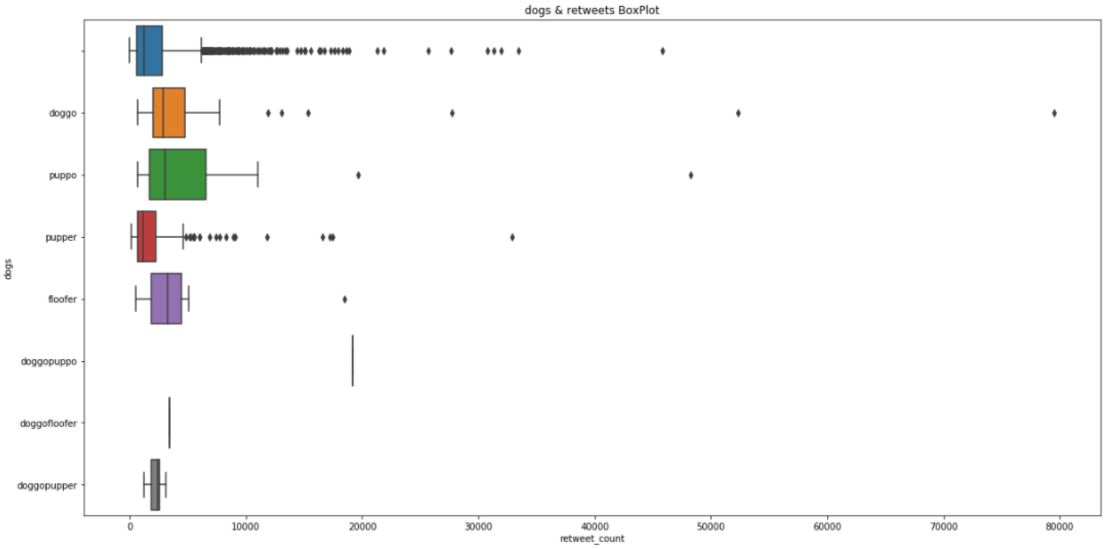


From 01 month to 05 the count of tweet are drop and still have a same range until 07 month.

In our data to make it more easy and clear we are merge all type of dogs into one column, that's will make our data easy to analysis it and easy to see it and visualize it, this is actually the value of data wrangling, below is the visualization of the top type of dogs like:



In this visualization we will see the box plot of most dogs and count of retweets



from above box plot, the most dogs are in Puppo but the highest retweetes in doggo.