# ACT REPORT

By Thu Phung Ngoc Minh

May 2020

From Wrangling Process from Udacity project WeRateDogs, I take chance to apply the lessons not only into Wrangling Process implementation, but also deep dive in analysis and take results from what I observed. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

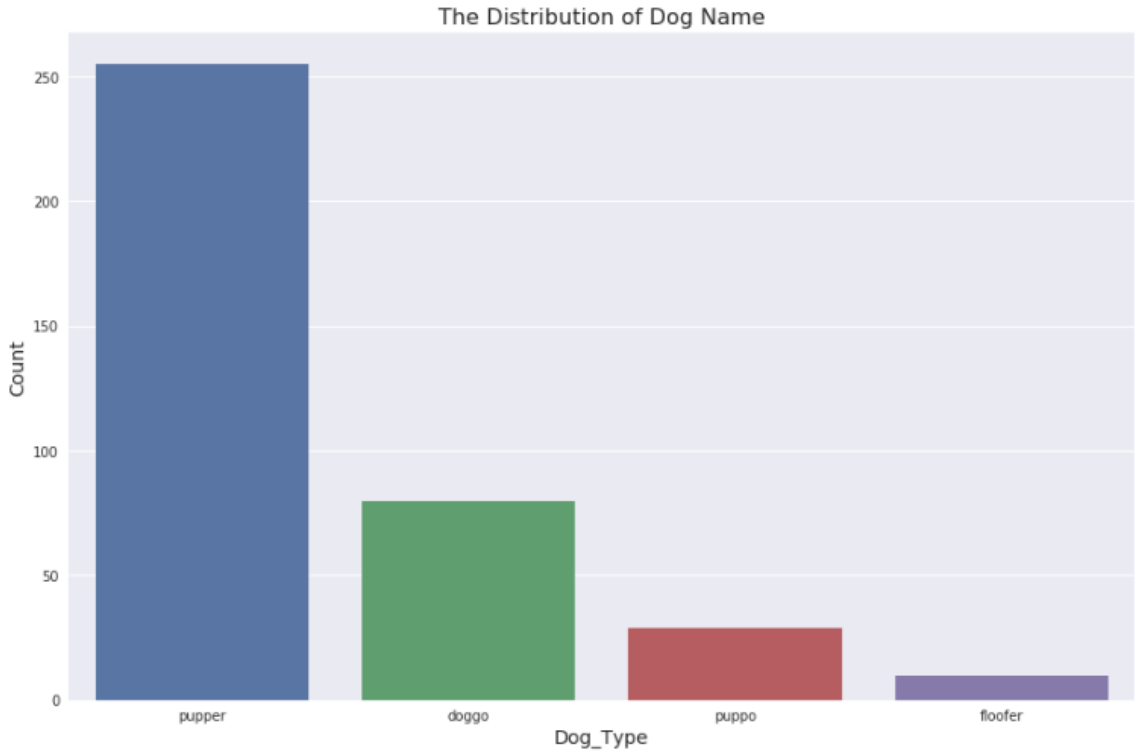
WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson (Winkipedia).

Question 1: What are the most dog names?

```
None       780
Charlie     12
Lucy        11
Oliver      11
Cooper      11
Penny       10
Lola        10
Tucker      10
Bo           9
Winston      9
Name: name, dtype: int64
```

After cleaning the weird names such as ‘a’,’an’,’all’,….the beside result showed that accepted the undefined names of dog, which accounted the large proportion. ‘Charlie’ is the most name which is used by many twitter accounts, followed by Lucy and Oliver.

Question 2: How is the distribution of Dog Type?



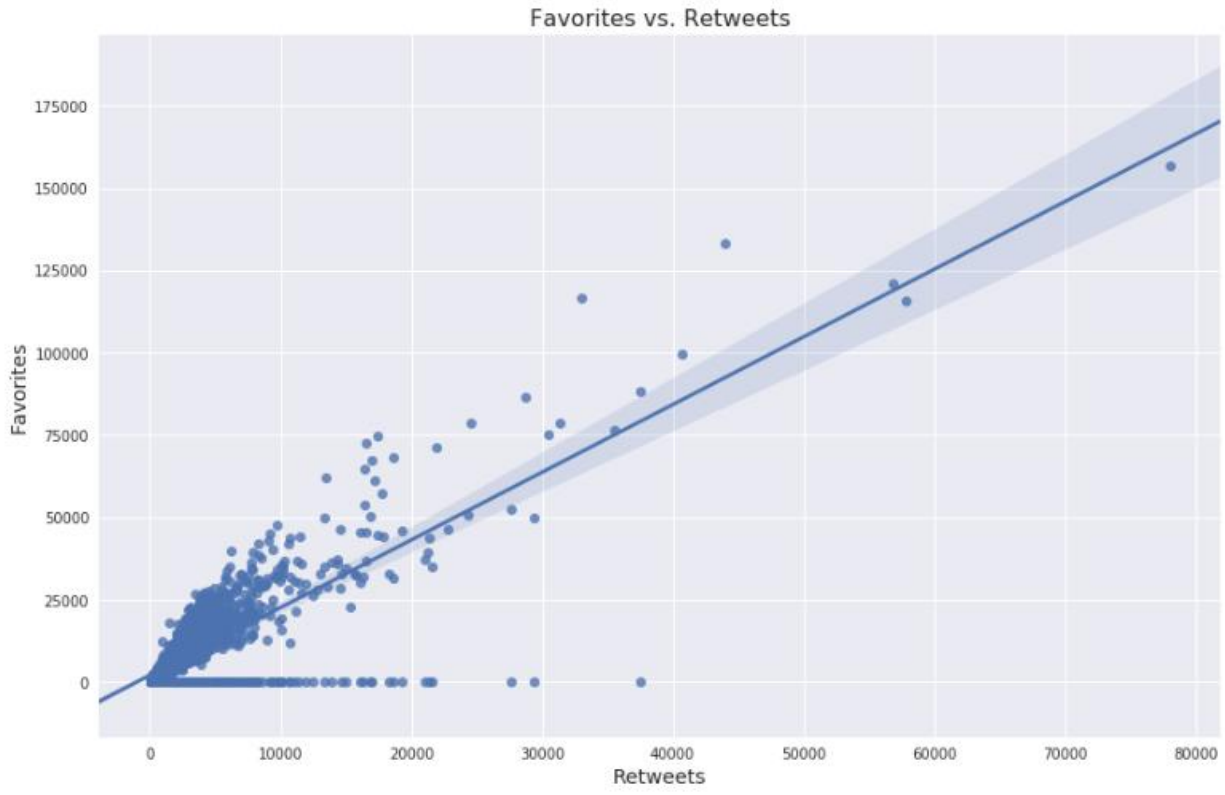

Mostly the type of dog is pupper: a small doggo, but usually younger. And follow by Doggo and Puppo. May be the pupper dogs are smaller and cutter than other dogs. However there are still big amount of None values, means the type of dogs cannot be defined. So the above result may be affected and may not reflect the right distribution.

Question 3: Correlation between variables

Firstly, by using the function describe(), we can see all variables are correlated or not. In all variables in below table, retweet_counts and favorite_counts have strong relationship (r^2 = 0.8).

| | tweet_id | rating_numerator | rating_denominator | favorite_count | retweet_count | img_num | p1_conf | p2_conf | p3_conf |
|---|---|---|---|---|---|---|---|---|---|
| tweet_id | 1.000000 | 0.023986 | -0.022078 | 0.519600 | 0.389042 | 0.206521 | 0.101821 | 0.002012 | -0.043424 |
| rating_numerator | 0.023986 | 1.000000 | 0.188685 | 0.016965 | 0.017749 | 0.000363 | -0.009744 | -0.020481 | -0.006127 |
| rating_denominator | -0.022078 | 0.188685 | 1.000000 | -0.021235 | -0.020809 | -0.000878 | -0.006174 | -0.037538 | -0.007066 |
| favorite_count | 0.519600 | 0.016965 | -0.021235 | 1.000000 | 0.800720 | 0.125228 | 0.066669 | -0.017833 | -0.046613 |
| retweet_count | 0.389042 | 0.017749 | -0.020809 | 0.800720 | 1.000000 | 0.105860 | 0.043437 | -0.004793 | -0.031613 |
| img_num | 0.206521 | 0.000363 | -0.000878 | 0.125228 | 0.105860 | 1.000000 | 0.203571 | -0.159956 | -0.139622 |
| p1_conf | 0.101821 | -0.009744 | -0.006174 | 0.066669 | 0.043437 | 0.203571 | 1.000000 | -0.511298 | -0.709449 |
| p2_conf | 0.002012 | -0.020481 | -0.037538 | -0.017833 | -0.004793 | -0.159956 | -0.511298 | 1.000000 | 0.479027 |
| p3_conf | -0.043424 | -0.006127 | -0.007066 | -0.046613 | -0.031613 | -0.139622 | -0.709449 | 0.479027 | 1.000000 |

Therefore, to make the correlation is clearer between these 2 variables. I visualize by using regplot from seaborn library.

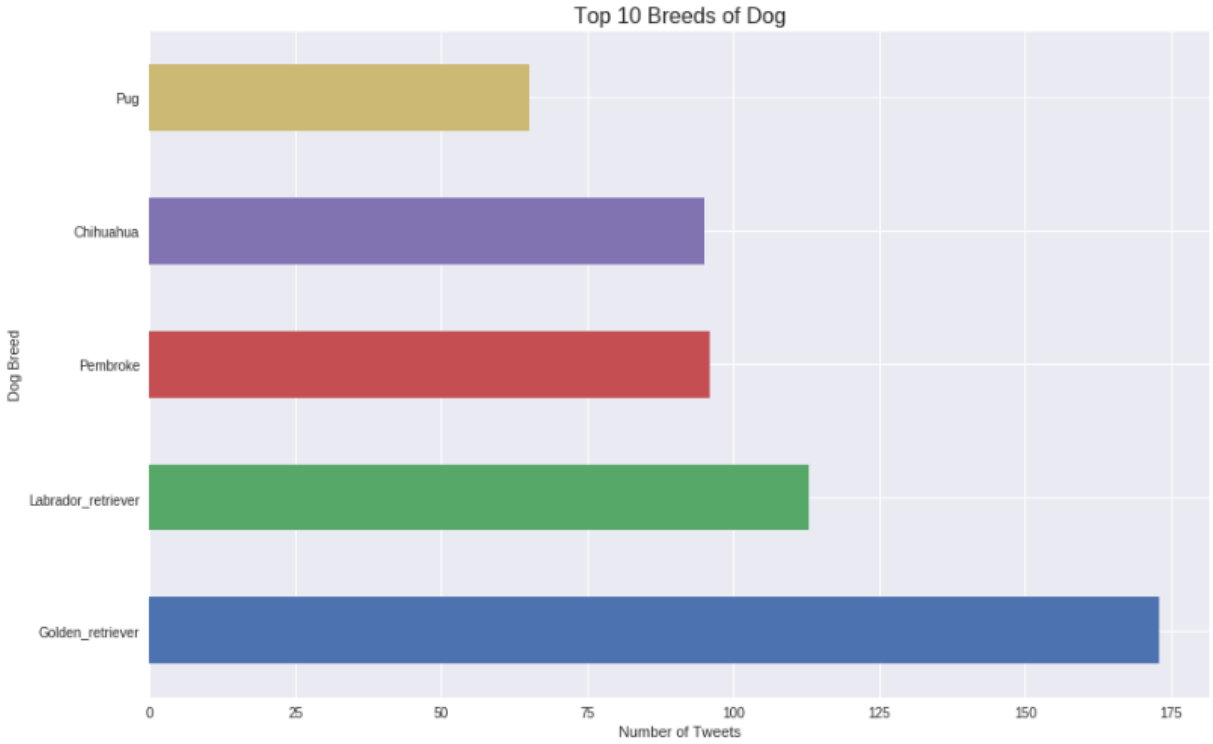



The strong correlation above may reflect the right reality that most of the interesting tweets always get high retweets and favorites. Retweets are used to share the references to everybody who has same interests, and favorites are used to feedback the likes/interests/ or even thanks to the poster.

Question 4: Distribution of dog breed?

Firstly I need to use loop to define the new columns which merge 3 columns P1 P2 P3, to define which kinds of dog breed are most popular.

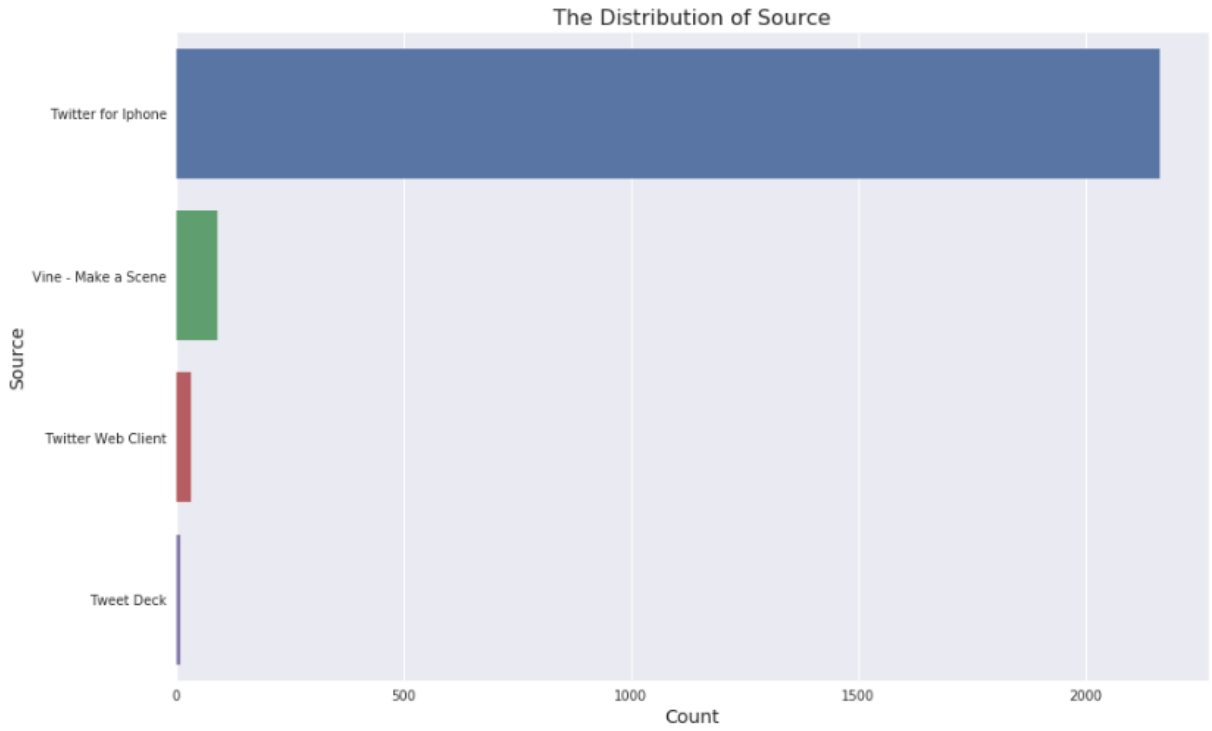
Then I use function value_coutns() and nlargest(10) to see top 10 of dog breeds are most popular.




I can define that the major prediction is from golden_retriever, which is most popular choice for dog breed. Followed by Labrador_Retriever and Pembroke – these 2 breeds of dog are mostly small dogs.



Question 5: How is the distribution of source?




I can see from chart that the most popular interaction from the people who use Twitter are by using Iphone, and by Twitter application. That means most of information are posted or interacted such as posting, retweets, favorites by Twitter app more than other functions like Web Client or Tweet desk.