UDACITY – ACT REPORT

TO: UDACITY REVIEWER

FROM: JHONATAN NAGASAKO

SUBJECT: WRANGLE REPORT (>250 WORDS)

DATE: 27-FEB-2021

PURPOSE

The purpose of this project was to analyze real-world data. This data analyzed took the form of:

1. Meta twitter data (@dog_rates)
2. WeRateDogs twitter
3. Udacity Neural network algorithm (to predict dog breed)

This paper is an extension of the first report written called, "wrangle_report", which explains further motivation of the project and concise review of the data analyzed.

DATA EXPLORATION

POPULAR DOGS

One of the questions asked while review the data was understanding the distribution of different classification of dogs. From the cleaned twitter dataset, the following dog classes was evaluated (per the dogtionary):

1. Floof: Referred to as a miscellaneous or "any dog really" classification
2. Pupper: A small doggo
3. Doggo: A bigg pupper
4. Puppo: a transitional phase between pupper and doggo

Review and the counts for each of these dog classes are as shown in Table 1. The corresponding graph is as shown in Figure 1.

Table 1: Dog Class Count [output 105]

```
Out[105]:  Dog_Class
           floof      1748
           pupper      209
           doggo        67
           puppo        23
           Name: tweet_id, dtype: int64
```

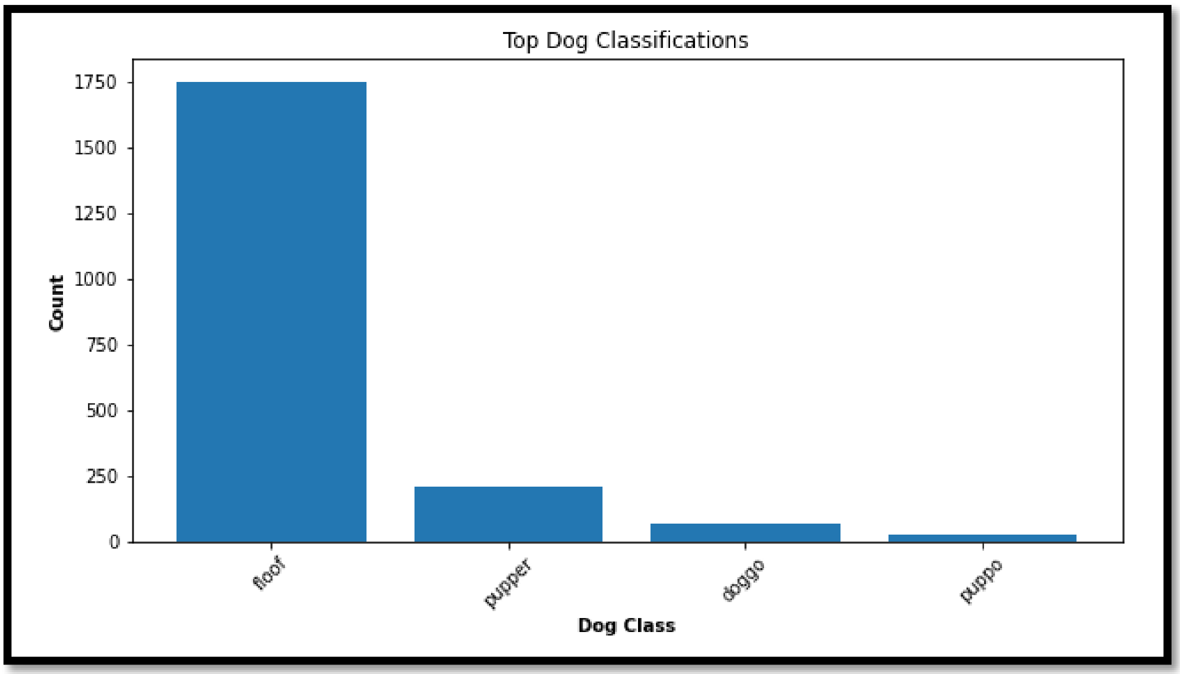

Figure 1: Dog Class Count bar graph [output 105]

POPULAR DOG NAMES

In addition to the different classification of dogs, the review of popular dog names was also evaluated. The results are as shown in Table 2 and Figure 2. Additionally, the least favorite—or very least common—names were also reviewed, as shown in Table 3 and Figure 3.

Table 2: Top dog names (shown here is top 5) [output 113]

```
Out[113]: name
          A          55
          CHARLIE    11
          PENNY      10
          COOPER     10
          LUCY       10
          Name: tweet_id, dtype: int64
```

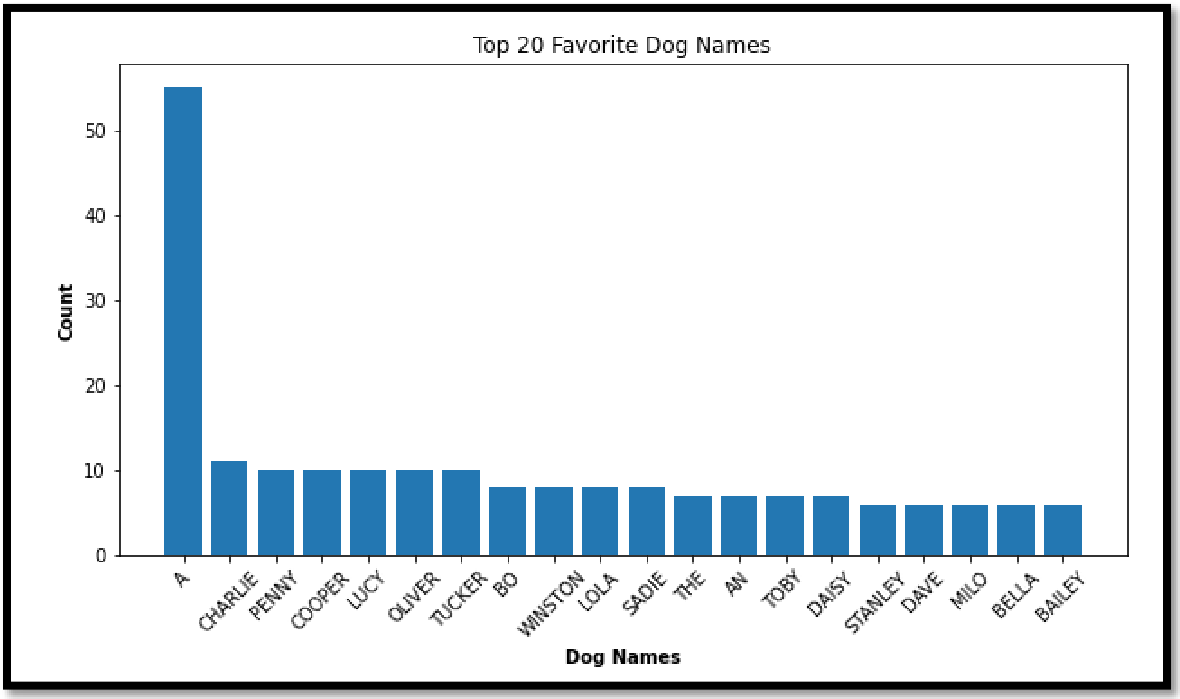

Figure 2: Top dog names (shown here is top 20) [output 113]

Table 3: Top LEAST favorite/common dog names (shown here is top 5) [output 115]

```
Out[115]: name
          KAYLA      1
          MARTY      1
          MARVIN     1
          MARY       1
          MASON      1
          Name: tweet_id, dtype: int64
```

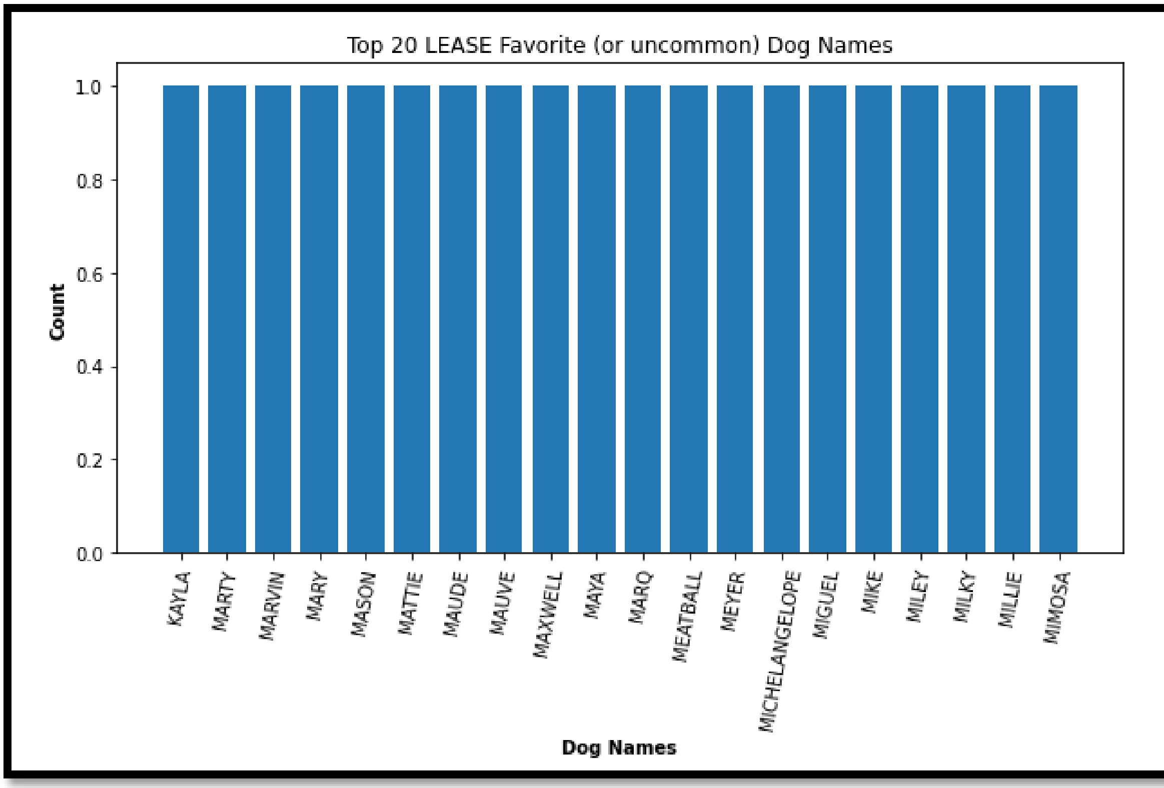

Figure 3: Top LEAST favorite/common dog names (shown here is top 20) [output 115]

NEURAL NETWORK ALGORITHM EFFICIENCY

Lastly, the Udacity neural network algorithm was evaluated to determine algorithm efficiency to discern various pictures of dogs. The reference “p1”, “p2”, and “p3” are different “prediction” success rates—the iteration of numbers is the algorithms 1st, 2nd, and 3rd attempt to classify the dog in the picture. The following results each prediction success is as shown in Figure 4, Figure 5, and Figure 6.

```
In [120]:  df_AI.groupby('p1_dog')['tweet_id'].count()

Out[120]:  p1_dog
           False     539
           True     1508
           Name: tweet_id, dtype: int64

In [121]:  print('First try accuracy = ', 1515/2054)

           First try accuracy =  0.737585199610516
```

Figure 4: Prediction 1 (p1) success rate

```
In [122]:  df_AI.groupby('p2_dog')['tweet_id'].count()

Out[122]:  p2_dog
           False     516
           True     1531
           Name: tweet_id, dtype: int64

In [123]:  print('Second try accuracy = ', 1538/2054)

           Second try accuracy =  0.7487828627069133
```

Figure 5: Prediction 2 (p2) success rate

```
In [124]:  df_AI.groupby('p3_dog')['tweet_id'].count()

Out[124]:  p3_dog
           False     568
           True     1479
           Name: tweet_id, dtype: int64

In [125]:  print('Third try accuracy = ', 1485/2054)

           Third try accuracy =  0.7229795520934762
```

Figure 6: Prediction 3 (p3) success rate

DATA EXPLANATORY

The review of the dog classification data reveals that majority of the dog classification falls into the floof category, followed by pupper. This floof category high rating may be due to the fact its classification is a “catch-all” for situation which the dog is not captured in the other three categories. The second higher pupper class indicates that most twitter users are either rating, tweeting, or taking pictures of “smaller dogs” or perhaps “puppies”. More analysis of the actual dog pictures is required to make that determination.

The review of popular dog names indicates that “A” is by far the most popular name. The data scientist lends itself to think this maybe a typo, but respects dog owner’s decision to name their dogs whatever-way they please. It is interesting to note that many of these dog names—good or bad/least-common—sound like “real people” names (e.g., good set: Charlie, Penny, etc. Then bad/least-common set: Kayla, Marty, etc.).

Lastly, the review of the neural network algorithm reveals average 73.63% efficiency rating to determine the correct bread of dog in various pictures. Further classification dataset and training can be explored to improve this efficiency—but is out of scope for this analysis and report.

CONCLUSION

The data scientist was able to explore the provided twitter datasets to create informative plots and descriptive statistics.