Analysis

July 1, 2018

```
In [13]: import pandas as pd
         import matplotlib.pyplot as plt
         from IPython.display import Image
         from IPython.core.display import HTML
```

0.1 Analyze

```
In [14]: import matplotlib
         df = pd.read_csv('twitter_archive_master.csv')
```

```
In [15]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1993 entries, 0 to 1992
Data columns (total 24 columns):
tweet_id                 1993 non-null int64
in_reply_to_status_id    23 non-null float64
in_reply_to_user_id      23 non-null float64
timestamp                1993 non-null object
source                   1993 non-null object
text                     1993 non-null object
expanded_urls            1993 non-null object
rating_numerator         1993 non-null float64
rating_denominator       1993 non-null float64
name                     1371 non-null object
dog_stage                326 non-null object
retweet_count            1993 non-null int64
favorite_count           1993 non-null int64
jpg_url                  1993 non-null object
img_num                  1993 non-null int64
p1                       1993 non-null object
p1_conf                  1993 non-null float64
p1_dog                   1993 non-null bool
p2                       1993 non-null object
p2_conf                  1993 non-null float64
p2_dog                   1993 non-null bool
p3                       1993 non-null object
p3_conf                  1993 non-null float64
```

```
p3_dog                   1993 non-null bool
dtypes: bool(3), float64(7), int64(4), object(10)
memory usage: 332.9+ KB
```

```
In [16]: df.head(2)
```

```
Out[16]:             tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         0  892420643555336193                    NaN                  NaN
         1  892177421306343426                    NaN                  NaN

                      timestamp             source  \
         0  2017-08-01 16:23:56  Twitter for iPhone
         1  2017-08-01 00:17:27  Twitter for iPhone

                                                         text  \
         0  This is Phineas. He's a mystical boy. Only eve...
         1  This is Tilly. She's just checking pup on you...

                                                expanded_urls  rating_numerator  \
         0  https://twitter.com/dog_rates/status/892420643...              13.0
         1  https://twitter.com/dog_rates/status/892177421...              13.0

            rating_denominator     name dog_stage  retweet_count  favorite_count  \
         0                10.0  Phineas       NaN           8560           38693
         1                10.0    Tilly       NaN           6293           33168

                                                    jpg_url  img_num         p1  \
         0  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg        1     orange
         1  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg        1  Chihuahua

             p1_conf  p1_dog        p2   p2_conf  p2_dog        p3   p3_conf  p3_dog
         0  0.097049   False     bagel  0.085851   False    banana  0.076110   False
         1  0.323581    True  Pekinese  0.090647    True  papillon  0.068957    True
```

0.1.1 Define

- Who has the most favorited dog?
- What does their picture look like?

```
In [17]: pd.set_option('display.max_columns', None)
```

```
In [18]: df[df["favorite_count"]== 143024]
```

```
Out[18]:               tweet_id  in_reply_to_status_id  in_reply_to_user_id  \
         309  822872901745569793                    NaN                  NaN

                        timestamp             source  \
         309  2017-01-21 18:26:02  Twitter for iPhone
```

```
                                               text  \
309  Here's a super supportive puppo participating ...

                                       expanded_urls  rating_numerator  \
309  https://twitter.com/dog_rates/status/822872901...              13.0

     rating_denominator name dog_stage  retweet_count  favorite_count  \
309                10.0  NaN     puppo          48971          143024

                                         jpg_url  img_num  \
309  https://pbs.twimg.com/media/C2tugXLXgAArJO4.jpg        1

                   p1   p1_conf  p1_dog                  p2   p2_conf  p2_dog  \
309  Lakeland_terrier  0.196015    True  Labrador_retriever  0.160329    True

                p3   p3_conf  p3_dog
309  Irish_terrier  0.069126    True
```

In [19]:
```
#Let's pull his picture the dataset
img_url = str(df[df['tweet_id']==822872901745569793].jpg_url).split()[1]
print(img_url)
Image(img_url,width=300, height=300)
```

```
https://pbs.twimg.com/media/C2tugXLXgAArJO4.jpg
```

Out[19]:

0.1.2 Define

- What are the top 5 most popular dog names?

```
In [20]: from collections import Counter

         common_5_names = df[df.name.notnull()].name
         count = Counter(common_5_names)
         count.most_common(5)
```

```
Out[20]: [('Charlie', 11), ('Oliver', 10), ('Cooper', 10), ('Lucy', 10), ('Penny', 9)]
```

- Charlie, Oliver, Cooper, Lucy and Penny are the five most common name

0.2 Descriptive Statistical Analysis

```
In [21]: # Descriptive statistics
         stats= df.drop('tweet_id', axis=1)
         stats.describe()
```

```
Out[21]:        in_reply_to_status_id  in_reply_to_user_id  rating_numerator  \
         count           2.300000e+01         2.300000e+01       1993.000000
```

```
mean             6.978112e+17         4.196984e+09           12.206613
std              4.359384e+16         0.000000e+00           41.473096
min              6.671522e+17         4.196984e+09            0.000000
25%              6.732411e+17         4.196984e+09           10.000000
50%              6.757073e+17         4.196984e+09           11.000000
75%              7.031489e+17         4.196984e+09           12.000000
max              8.558181e+17         4.196984e+09         1776.000000

       rating_denominator  retweet_count  favorite_count      img_num  \
count         1993.000000    1993.000000     1993.000000  1993.000000
mean            10.511791    2708.934772     8827.983944     1.203211
std              7.262919    4677.697123    12537.586518     0.560899
min             10.000000      13.000000       80.000000     1.000000
25%             10.000000     606.000000     1913.000000     1.000000
50%             10.000000    1304.000000     4032.000000     1.000000
75%             10.000000    3119.000000    11113.000000     1.000000
max            170.000000   77143.000000   143024.000000     4.000000

           p1_conf       p2_conf       p3_conf
count  1993.000000  1.993000e+03  1.993000e+03
mean      0.593802  1.344685e-01  6.026575e-02
std       0.271951  1.006821e-01  5.089760e-02
min       0.044333  1.011300e-08  1.740170e-10
25%       0.362835  5.405530e-02  1.619070e-02
50%       0.587507  1.175080e-01  4.952370e-02
75%       0.845256  1.952180e-01  9.160200e-02
max       1.000000  4.880140e-01  2.734190e-01
```

Key points:

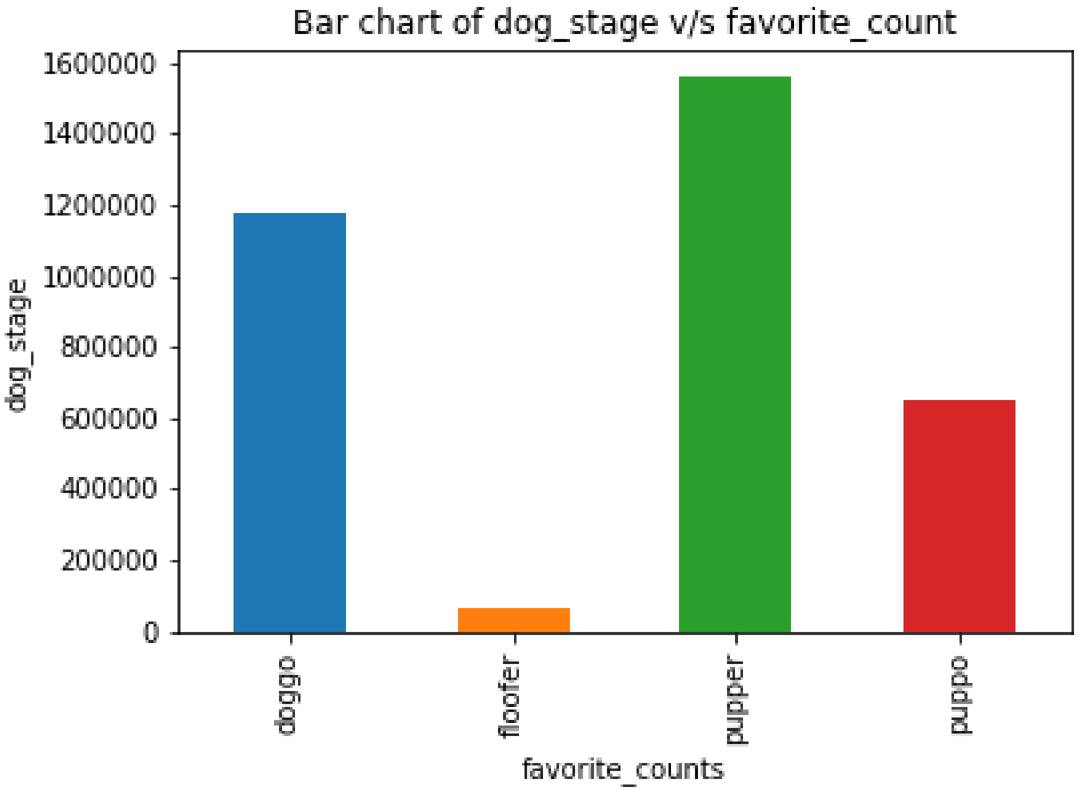
- The neural network performed the best on the 1st iteration with a mean prediciton of 0.59
- Mean rating for a dog is 12.207/10 with an outlier of 1776/10
- Mean retweet count for an original tweet was 2708 and a maximum value of 77143.
- Mean favorite count for an original tweet was 8827 and a maximum value of 143024.

0.2.1 dog_stage analysis

- Which dog_stage has got most favorite counts ?

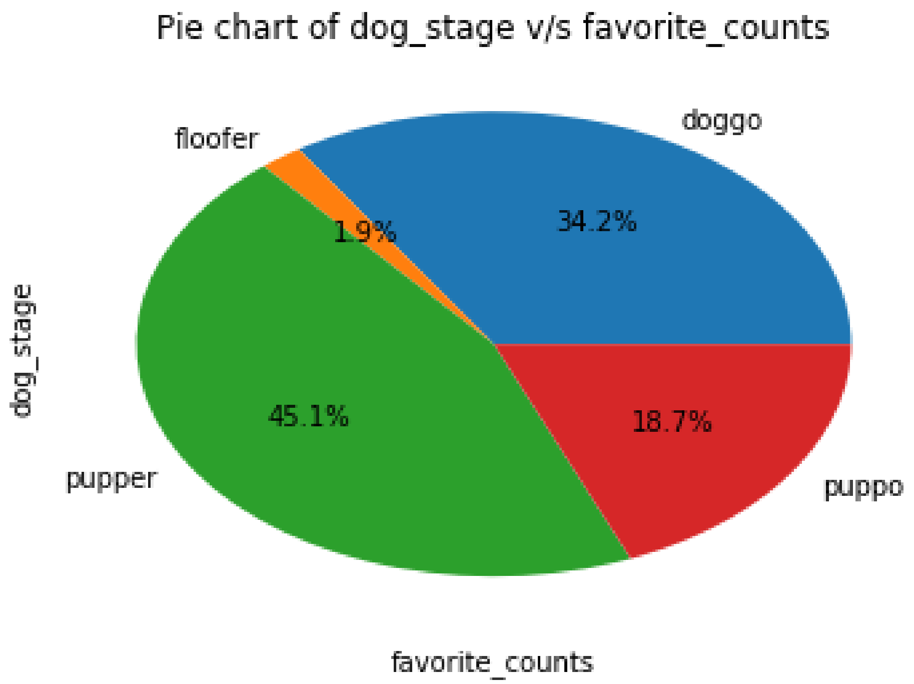
```
In [22]: top_dog_stage = df.groupby('dog_stage')['favorite_count'].sum()
         top_dog_stage.plot.bar()
         plt.title('Bar chart of dog_stage v/s favorite_count')
         plt.xlabel('favorite_counts')
         plt.ylabel('dog_stage')
```

```
Out[22]: Text(0,0.5,'dog_stage')
```



```
In [23]: top_dog_stage.plot(kind = 'pie', autopct='%1.1f%%')
         plt.title('Pie chart of dog_stage v/s favorite_counts')
         plt.xlabel('favorite_counts')
         plt.ylabel('dog_stage')

Out[23]: Text(0,0.5,'dog_stage')
```



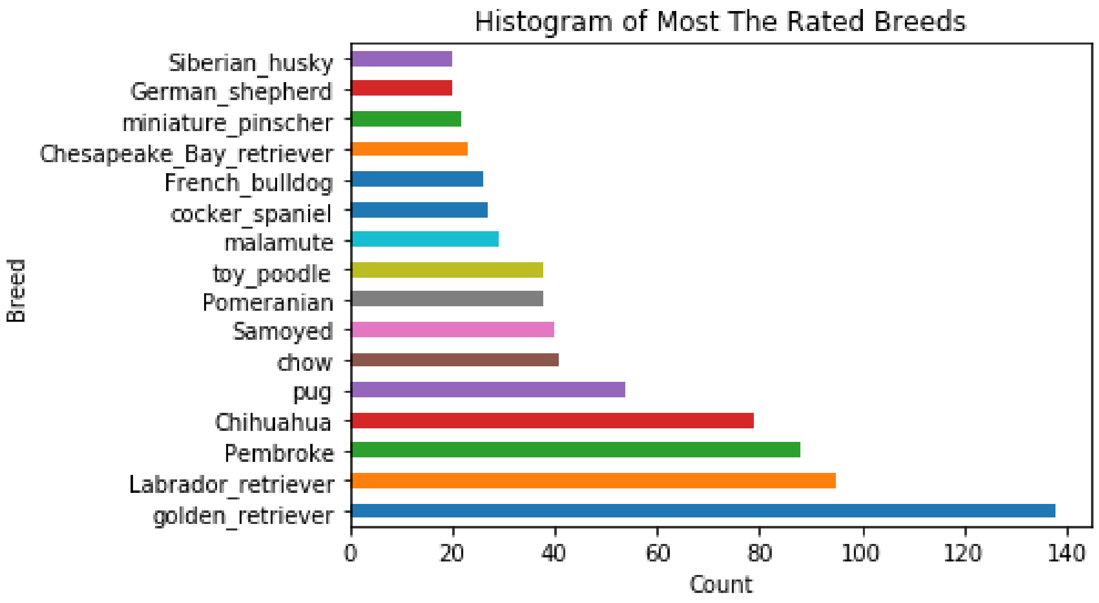
dog_stage pupper has the most favorite counts.

0.2.2 Most liked Breed

- What breed is having most favorite counts ?

```
In [24]: # Most liked breed
         top_breeds=df[df.p1_dog == True].groupby('p1').filter(lambda x: len(x) >= 20)
         top_breeds.p1.value_counts().plot(kind = 'barh')
         plt.title('Histogram of Most The Rated Breeds')
         plt.xlabel('Count')
         plt.ylabel('Breed')

Out[24]: Text(0,0.5,'Breed')
```



Golden_retriever is the most rated breed.

0.3 Conclusion

- The neural network performed the best on the 1st iteration with a mean prediciton of 0.59
- Mean rating for a dog is 12.207/10 with an outlier of 1776/10
- Mean retweet count for an original tweet was 2708 and a maximum value of 77143.
- Mean favorite count for an original tweet was 8827 and a maximum value of 143024.
- Most favorite dog tweet_id = 822872901745569793 with maximum value of favorite counts.
- Charlie, Oliver, Cooper, Lucy and Penny are the five most common name
- dog_stage pupper has the most favorite counts.
- Golden_retriever is the most rated breed.