act_report

January 2, 2018

1 WeRateDogs

For this project 2 sources of data were combined together:

1. WeRateDogs Twitter archive which was enhanced by extracting each tweet's retweet count and favorite as extracted from the twitter API.
2. Tweet image predictions based on a neural network that can classify breeds of dogs.

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline

        twitter_data = pd.read_csv('twitter_archive_master.csv')
        dog_info = pd.read_csv('dog_info.csv')

        full_dataset = dog_info.merge(twitter_data, how='inner', on='tweet_id')
        grouped_full = full_dataset.groupby(by='breed')['rating_numerator',
                       'favorite_count', 'retweet_count'].aggregate(['mean',
                                                                     'count'])
        grouped_full = grouped_full[grouped_full['rating_numerator']['count']>2]
```

This data data can be used to extract some information on the most popular dog breeds out there. The best rated dogs in the twitter archive are:

```
In [2]: print '\n'.join(grouped_full.sort_values([('rating_numerator' ,'mean')],
                                                  ascending=False)[:3].index)
```

```
Pomeranian
Saluki
briard
```

The worst rated dogs are:

```
In [3]: print '\n'.join(grouped_full.sort_values([('rating_numerator' ,'mean')],
                                                  ascending=True)[:3].index)
```

```
soft-coated_wheaten_terrier
Walker_hound
Ibizan_hound
```

Alternatively we could look at the favorite and or retweet count to get an idea on the popularity of the dogs. The top 3 most reteweed dogs are:

```
In [4]: print '\n'.join(grouped_full.sort_values([('retweet_count' ,'mean')],
                                                  ascending=False)[:3].index)
```

```
Bedlington_terrier
Afghan_hound
standard_poodle
```

The top 3 most favorited dogs are:

```
In [5]: print '\n'.join(grouped_full.sort_values([('favorite_count' ,'mean')],
                                                  ascending=False)[:3].index)
```

```
Saluki
Bedlington_terrier
French_bulldog
```

Finally it is interesting to take a look at the most added dogs on the twitter account. The top 3 consists of:

```
In [6]: print '\n'.join(grouped_full.sort_values([('favorite_count' ,'count')],
                                                  ascending=False)[:3].index)
```

```
golden_retriever
Labrador_retriever
Pembroke
```

The analysis above indicates that ratings, retweets and favorites do not always seem to match. This is investiged further in the following visual assessement. Additionnally the correlation coefficients are determined.

Both the visual assessement as well as the correlation analysis show a significant correlation between the ratings and favorite and retweet counts. This correlation is more obvious with the favorite count compared to the retweet count.

```
In [7]: plt.figure(figsize=(16, 8))
        plt.scatter(x=grouped_full['rating_numerator']['mean'],
                    y=grouped_full['favorite_count']['mean'],
                    s=grouped_full['favorite_count']['count'])
        plt.scatter(x=grouped_full['rating_numerator']['mean'],
                    y=grouped_full['retweet_count']['mean'],
```

```
                s=grouped_full['retweet_count']['count'])
    plt.legend(['favorite_count', 'retweet_count'])
    plt.grid()
    plt.xlabel('rating_numerator [-]')
    plt.ylabel('Count [-]')
```

```
Out[7]: <matplotlib.text.Text at 0x9aebf60>
```