act_report

May 26, 2020

1 Visualize and Analyze

1.1 Shakhawat Hassan

```
In [70]: df_new.describe()
```

```
Out[70]:        rating_numerator  rating_denominator      p1_conf       p2_conf  \
         count       2073.000000         2073.000000  2073.000000  2.073000e+03
         mean          12.265798           10.511819     0.594532  1.346665e-01
         std           40.699924            7.180517     0.271234  1.006830e-01
         min            0.000000            2.000000     0.044333  1.011300e-08
         25%           10.000000           10.000000     0.364095  5.390140e-02
         50%           11.000000           10.000000     0.588230  1.186220e-01
         75%           12.000000           10.000000     0.843911  1.955730e-01
         max         1776.000000          170.000000     1.000000  4.880140e-01

                     p3_conf  retweet_count  favorite_count
         count  2.073000e+03    2073.000000     2073.000000
         mean   6.034005e-02    2976.089243     8556.718283
         std    5.092769e-02    5054.897526    12098.640994
         min    1.740170e-10      16.000000        0.000000
         25%    1.619920e-02     634.000000     1674.000000
         50%    4.947150e-02    1408.000000     3864.000000
         75%    9.193000e-02    3443.000000    10937.000000
         max    2.734190e-01   79515.000000   132810.000000
```

- At 75 percentile, most dogs get at the scale of 12 on rating numerator.
- At 75 percentile, most dogs get at the scale of 11 on rating denominator.
- There are more favorite counts than retweet counts.

```
In [ ]:
```

1.1.1 Most Popular Names

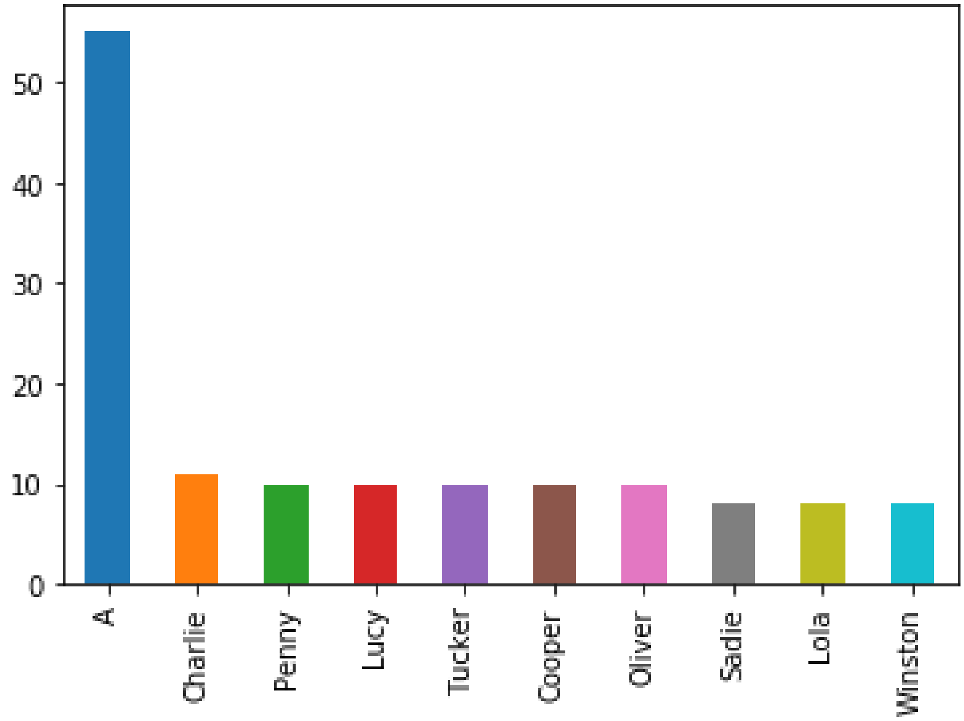
```
In [72]: common_names = df_new['name'].value_counts().nlargest(10)
         common_names
```

```
Out[72]: A          55
         Charlie    11
         Penny      10
         Lucy       10
         Tucker     10
         Cooper     10
         Oliver     10
         Sadie       8
         Lola        8
         Winston     8
         Name: name, dtype: int64
```

Top 10 dog names

```
In [73]: common_names.plot.bar()

Out[73]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe1fc392048>
```



Unrecorded names with 'A' dogs' have the highest number of names among all other names.

1.1.2 Dog Stages

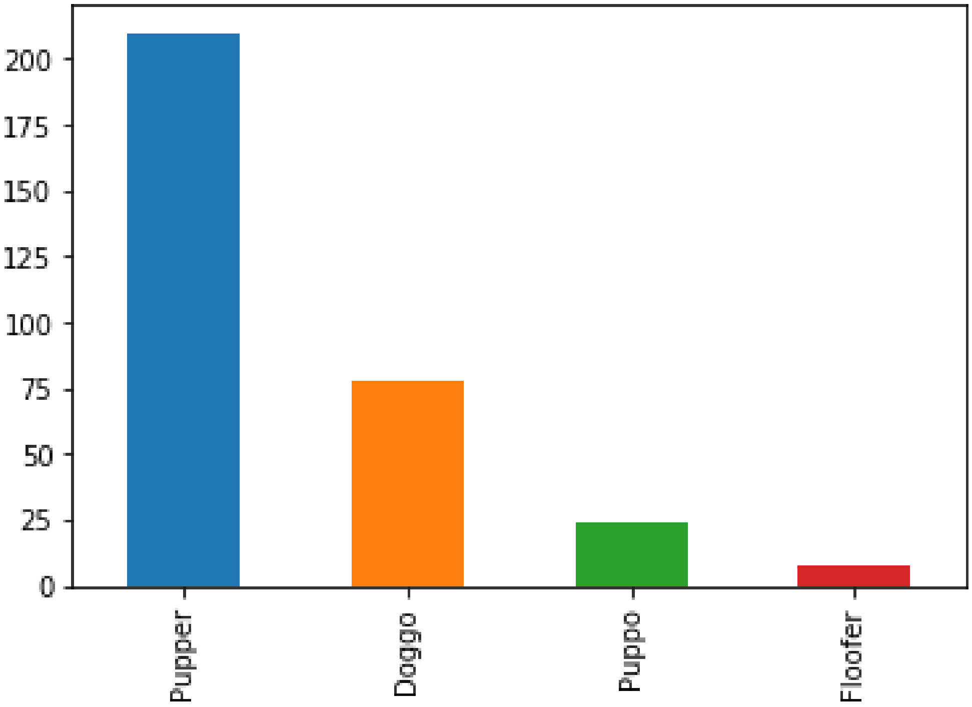
```
In [74]: dog_stages = df_new['stage'].value_counts()
         dog_stages
```

```
Out[74]: Pupper     210
         Doggo       78
         Puppo       24
         Floofer      8
         Name: stage, dtype: int64
```

Pupper stage has the highest number of dogs

```
In [75]: dog_stages.plot.bar()
```

```
Out[75]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe1fc5079b0>
```



Pupper stage has the highest number of dogs

```
In [76]: df_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2073 entries, 0 to 2072
Data columns (total 18 columns):
tweet_id                 2073 non-null object
timestamp                2073 non-null datetime64[ns]
text                     2073 non-null object
rating_numerator         2073 non-null int64
rating_denominator       2073 non-null int64
name                     1496 non-null object
```

```
stage                     320 non-null object
p1                        2073 non-null object
p1_conf                   2073 non-null float64
p1_dog                    2073 non-null bool
p2                        2073 non-null object
p2_conf                   2073 non-null float64
p2_dog                    2073 non-null bool
p3                        2073 non-null object
p3_conf                   2073 non-null float64
p3_dog                    2073 non-null bool
retweet_count             2073 non-null int64
favorite_count            2073 non-null int64
dtypes: bool(3), datetime64[ns](1), float64(3), int64(4), object(7)
memory usage: 265.2+ KB
```

1.1.3 Favorite Tweets vs Retweets

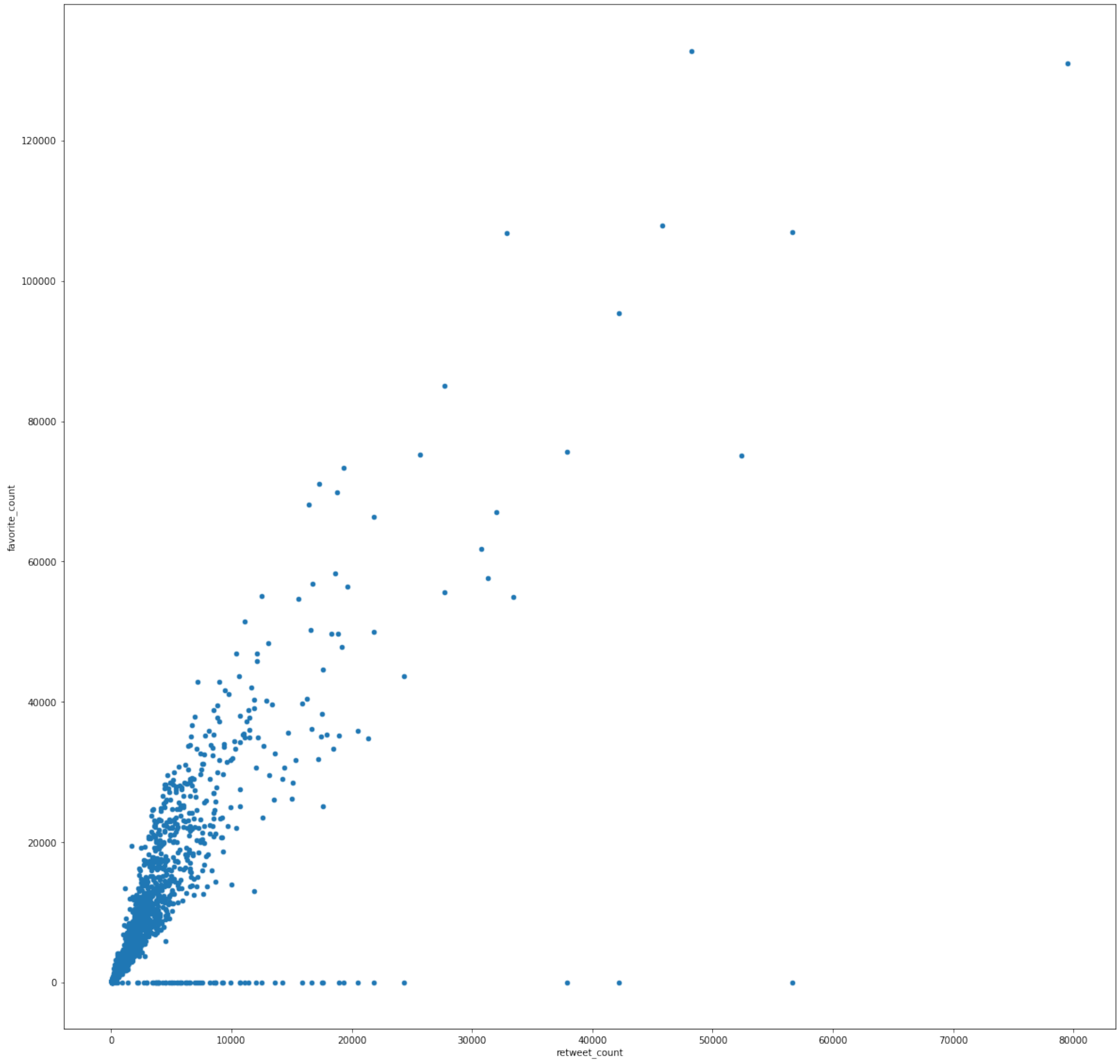
```
In [77]: df_new['retweet_count'].describe()
```

```
Out[77]: count     2073.000000
         mean      2976.089243
         std       5054.897526
         min         16.000000
         25%        634.000000
         50%       1408.000000
         75%       3443.000000
         max      79515.000000
         Name: retweet_count, dtype: float64
```

```
In [78]: df_new['favorite_count'].describe()
```

```
Out[78]: count      2073.000000
         mean       8556.718283
         std       12098.640994
         min           0.000000
         25%        1674.000000
         50%        3864.000000
         75%       10937.000000
         max      132810.000000
         Name: favorite_count, dtype: float64
```

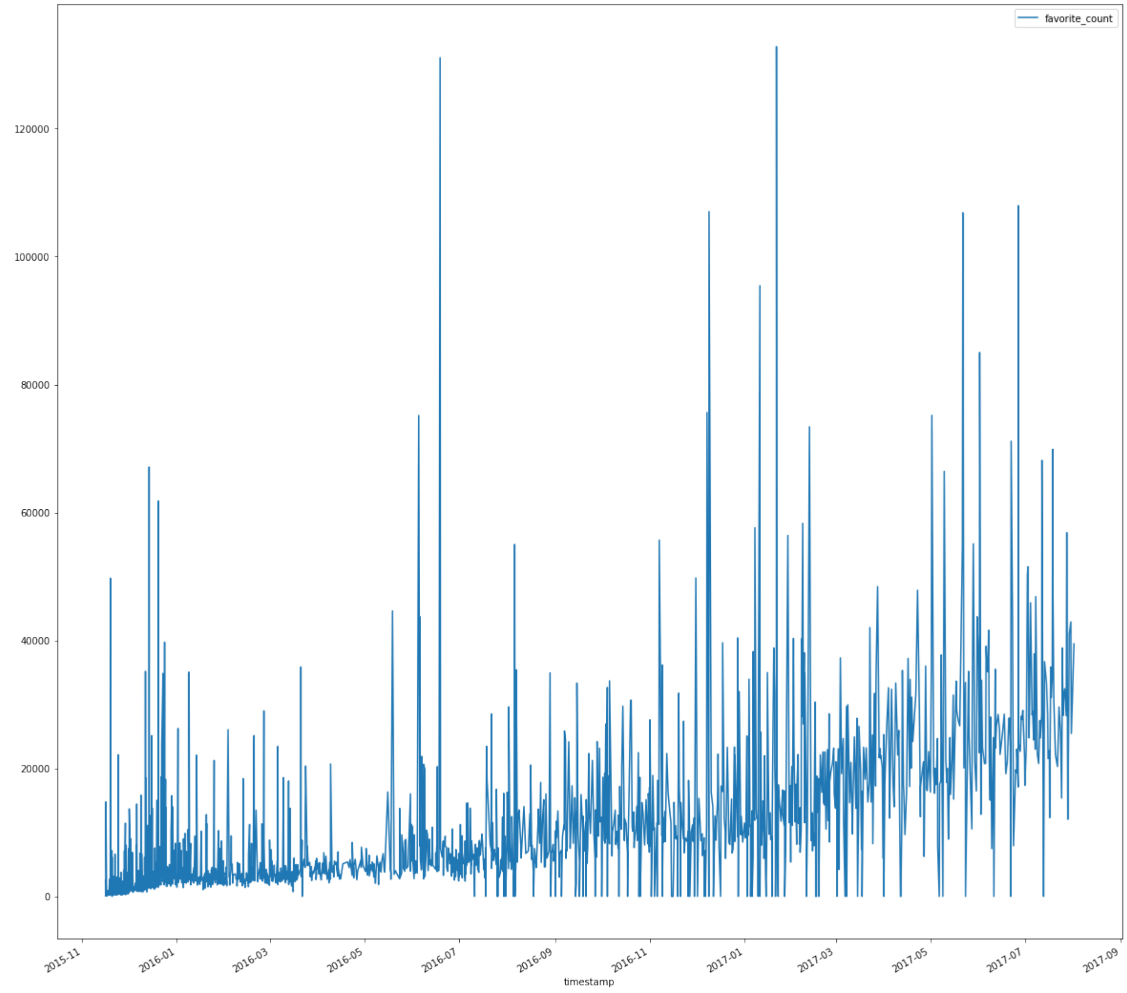
```
In [52]: df_new.plot(x = 'retweet_count', y= 'favorite_count' , kind = 'scatter', figsize= (20,
         plt.show()
```



There is a more positive correlation towards the favorite count side.

1.1.4 Favorite tweets vs Timestamp

```
In [53]: df_new.plot(x = 'timestamp', y= 'favorite_count' , kind = 'line', figsize = (20,20))
         plt.show()
```



As time goes on, more people are liking a tweet than retweeting that tweet.

1.2 Conclusion:

Data Wrangling: After cleaning all the datasets, merged all three datasets into one single dataset. I get 320 dogs', 1496 dogs' names, other than these two rows. I have all other rows with 2073 rows. #### Analysis: At 75 percentile, most dogs get at the scale of 12 on rating numerator. At 75 percentile, most dogs get at the scale of 11 on rating denominator. There are more favorite counts than retweet counts. Top 5 dog names are A (unrecorded name), Charlie, Penny, Charlie, Lucy. The pupper stage has the highest number of dogs among all other stages (210). As time goes on, more people are favoriting a tweet than retweeting a tweet.

1.3 Limitations:

There is about 55 dog names which are named with 'A' which does not tell us what's the real name of an 'A' dog's name. There are about 1496 dog names wheres rows are 2073. Dog stages are only about 320 while having 1496 dogs' names.

```
In [ ]:
```