DATA ANALYST NANODEGREE

WeRateDogs - Twitter Data

ANALYZING AND COMMUNICATING DATA

Introduction

Real-world data rarely comes clean. Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset that we wrangled (and analyzed and visualized) was the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Account details:

name: WeRateDogs™ (author)
screen_name: dog_rates
description: #1 Source for Professional Dog Ratings | STORE: @ShopWeRateDogs | IG, FB & SC: WeRateDogs MOBILE APP: @GoodDogsGame | Business: dogratingtwitter@gmail.com
verified: True
followers_count(Till 2017-08-01 16:23:56) : 3200889
friends_count(Till 2017-08-01 16:23:56): 104

Analysis and Visualization

- **Dataset was provided from date '2015-11-15 22:32:08' to '2017-08-01 16:23:56'.**
- **The number of followers decreased by 129 during this time period even though they posted several post during this period.**

Time period of Dataset

```
In [251]: (merge1.timestamp[len(merge1)-1],merge1.timestamp[0])
Out[251]: ('2015-11-15 22:32:08 +0000', '2017-08-01 16:23:56 +0000')
```

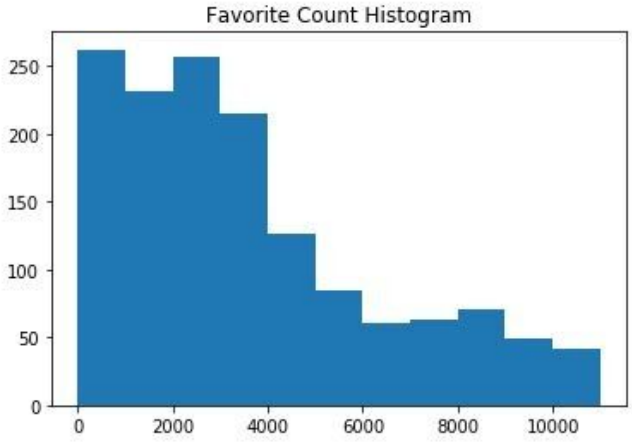
Change in No. of Followers during the time period '2015-11-15 22:32:08' - '2017-08-01 16:23:56'

```
In [252]: merge1.followers_counttweet[0]-merge1.followers_counttweet[len(merge1)-1]
Out[252]: -129
```

Note: The number of followers decreased by 129.

- **For most of the tweets favorites lies between 0-5000.**

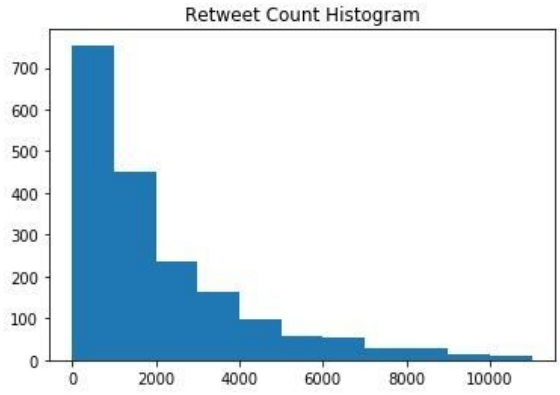
```
In [247]: plt.hist(merge1.favorite_count,bins=np.arange(0,12000,1000))
          plt.title("Favorite Count Histogram")
Out[247]: Text(0.5, 1.0, 'Favorite Count Histogram')
```



- **For most of the tweets, retweets lie between 0-3000.**

```
In [248]: plt.hist(merge1.retweet_count,bins=np.arange(0,12000,1000))
          plt.title("Retweet Count Histogram")
```

```
Out[248]: Text(0.5, 1.0, 'Retweet Count Histogram')
```



- **Top 10 tweets based on favorite count**

```
In [249]: top=merge1.sort_values(by=['favorite_count'],ascending=False).reset_index(drop=True)
          top10=top[:10]
          top10[["tweet_id","full_text","favorite_count","name"]]
```

Out[249]:

	tweet_id	full_text	favorite_count	name
0	822872901745569793	Here's a super supportive puppo participating ...	132810	Unknown
1	744234799360020481	Here's a doggo realizing you can stand in a po...	131075	Unknown
2	879415818425184262	This is Duddles. He did an attempt. 13/10 some...	107956	Duddles
3	807106840509214720	This is Stephan. He just wants to help. 13/10 ...	107015	Stephan
4	866450705531457537	This is Jamesy. He gives a kiss to every other...	106827	Jamesy
5	819004803107983360	This is Bo. He was a very good First Doggo. 14...	95450	Bo
6	870374049280663552	This is Zoey. She really likes the planet. Wou...	85011	Zoey
7	806629075125202948	"Good afternoon class today we're going to lea...	75639	Unknown
8	859196978902773760	We only rate dogs. This is quite clearly a smo...	75193	Unknown
9	739238157791694849	Here's a doggo blowing bubbles. It's downright...	75163	Unknown

- **Top 10 tweets based on retweet count**

```
In [250]: retop=merge1.sort_values(by=['retweet_count'],ascending=False).reset_index(drop=True)
          retop10=retop[:10]
          retop10[["tweet_id","full_text","retweet_count","name"]]
```

Out[250]:

	tweet_id	full_text	retweet_count	name
0	744234799360020481	Here's a doggo realizing you can stand in a po...	79515	Unknown
1	807106840509214720	This is Stephan. He just wants to help. 13/10 ...	56625	Stephan
2	739238157791694849	Here's a doggo blowing bubbles. It's downright...	52360	Unknown
3	822872901745569793	Here's a super supportive puppo participating ...	48265	Unknown
4	879415818425184262	This is Duddles. He did an attempt. 13/10 some...	45849	Duddles
5	819004803107983360	This is Bo. He was a very good First Doggo. 14...	42228	Bo
6	806629075125202948	"Good afternoon class today we're going to lea...	37911	Unknown
7	761672994376806400	Ohboyohboyohboyohboyohboyohboyohboyohboyohboyo...	33421	Unknown
8	866450705531457537	This is Jamesy. He gives a kiss to every other...	32883	Jamesy
9	676219687039057920	This is Kenneth. He's stuck in a bubble. 10/10...	31989	Kenneth

- **Number of tweets without dog names- 1971 tweets**

```
n [253]: named=merge1[merge1.name=="unknown"].tweet_id.count()
         unnamed=merge1.shape[0]-named
         unnamed
ut[253]: 1971
```

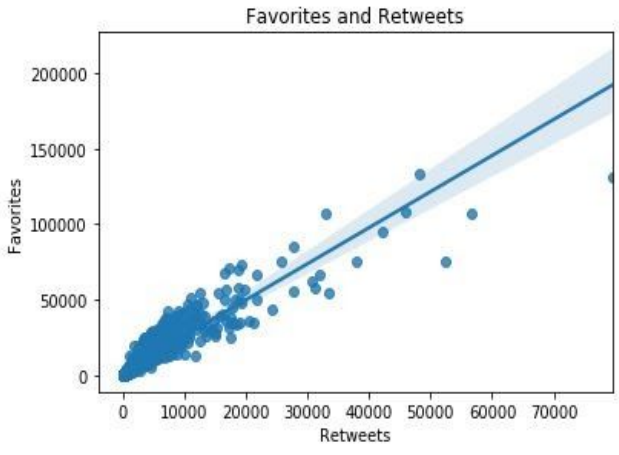
- **Correlation between "retweet_count","favorite_count", "rating_numerator" :**
  There is strong correlation between retweets and favorites. The correlation value is 0.91. Therefore, with more number of retweets, the favorite for the tweet increases.

Heatmap for correlation

```
In [255]: sns.heatmap(merge1[column].corr(),cmap="rocket_r",annot=True,vmin=0)
Out[255]: <matplotlib.axes._subplots.AxesSubplot at 0x2baf07cadc8>
```

- **Linear relationship between favorite count and retweet count.**

```
In [256]: graph = sns.regplot(x=merge1.retweet_count, y=merge1.favorite_count)
          plt.title("Favorites and Retweets")
          plt.xlabel('Retweets')
          plt.ylabel('Favorites')
Out[256]: Text(0, 0.5, 'Favorites')
```



Note: As the number of retweets increases, the number of favorites also increases.

The most favorited tweet:

Here's a super supportive puppo participating in the Toronto #WomensMarch today. 13/10 https://t.co/nTz3FtorBc

The most retweeted tweet:

Here's a doggo realizing you can stand in a pool. 13/10 enlightened af (vid by Tina Conrad) https://t.co/7wE9LTEXC4