wrangle_report

February 10, 2019

1 Wrangling Report

2 Introduction

project for wrangle and Analyze Data using Python for The dataset which to be from account @dog_rates also known as WeRateDogs , i got the dataset from Udacity. in this procject we have Data wrangling, which consists of:

1- Gathering data.
2- Assessing data.
3- Cleaning data.

3 1- Gathering data:

we have three les:

1- twitter-archive-enhanced.csv.
2- image_predictions.tsv.
3- tweet-json.txt. I gathered each of the three les.

4 2- Assessing Data:

Visual assess: I assessed data visually and programmatically to nd quality issues and 2 tidiness issues:

5 A) Quality Issues

There is a lot of missing values.

tweet_id should be string , not object.

Drop (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) , i think no need.

Data type wrong for timestamp.

Missing value expanded_urls 2297 only. Missing data in doggo, oofer, pupper and puppo.

doggo, oofer, pupper and puppo: Dog Phase, should be one column.

Split timestamp column to dates and time.

6 B) Tidiness issues:

One dataset between df and image_predictions because the link is tweet_id.
One dataset between df and df_titter because the link is tweet_id and id.

7 3- Clean Data:

I have 8 qualit issues and 2 tidiness issues, i clean all by using some functiono like:
Replace..
Astype.
isnull.
drop.
to_datetime.
dropna.
merge.
There are a lot of missing data, i converted to 0, and we have some variables wrong data type like tweet_id and timestamp , i converted. i deleted in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp because its not necessery. wehave4coloumndoggo,oofer,pupperandpuppowhichconvertedtoonecolumnitname dog_stageandideletedfor4variables. Iseperatedtimestamptotwocolumns(dates,time)andi deleted timestamp column.

8 Saving Data:

I saved all main dataframe to le:
df_master = pd.read_csv('df_master.csv') image_predictions_master = pd.read_csv('image_predictions_master.csv') df_twitter_master = pd.read_csv('df_twitter_master.csv')

9 Visulization

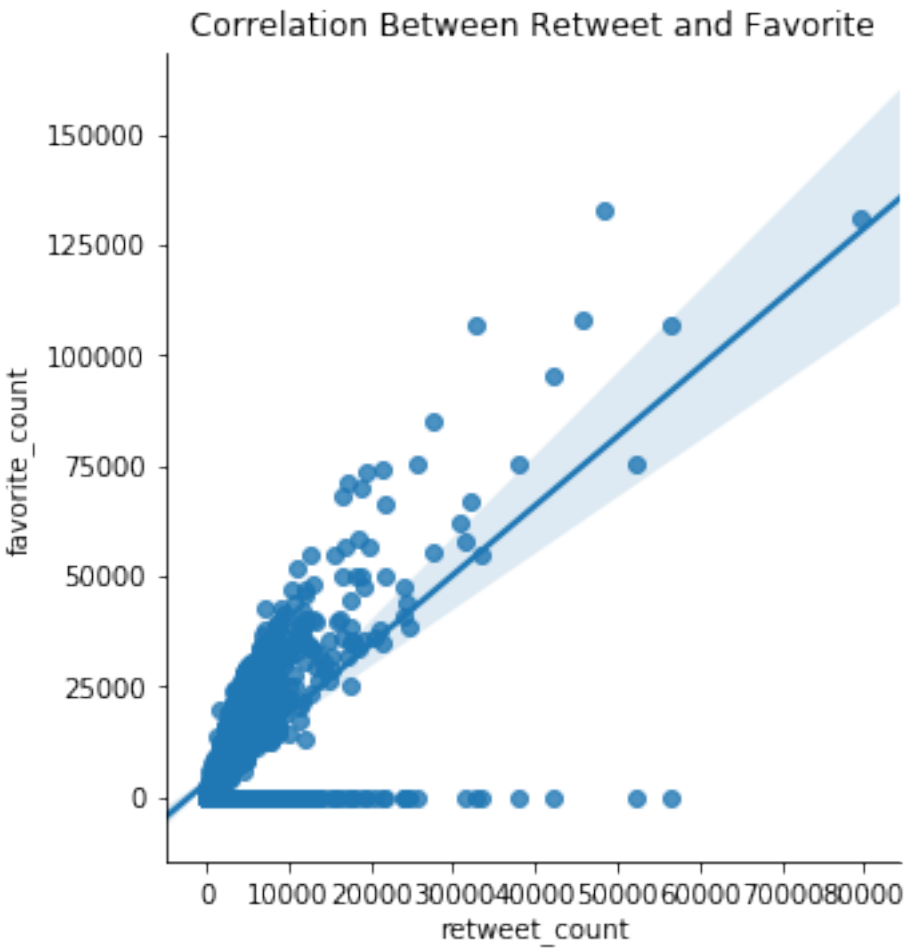
we have visualization, Correlation Between Retweet and Favorite.
and we have three insight that mean we want who the most retweet and favorite and what the most dog it appear?

10 Final

I have enjoyed analyzing the data for these dataser and i used several methods that explained in this report.

```
In [44]: # Relayionship between retweet_count and favorite_count
         sns.lmplot(x="retweet_count", y="favorite_count", data=df_twitter_master);
         plt.title('Correlation Between Retweet and Favorite')
```

```
Out[44]: Text(0.5,1,'Correlation Between Retweet and Favorite')
```



11 Resources:

https://seaborn.pydata.org/tutorial/regression.html
positive relationship between retweet_count and favorite_count.