Udacity Data Analyst Nanodegree

04 Wrangle And Analyze Data Part 2

DAVID LASSIG

02-28-2019

Contents

1 Project 4: Wrangle and Analyze Data Part 2: Analysis

After cleaning and storing the data into one CSV based master dataset I will try now to draw several conclusions from the datasets. I’m using Pandas DataFrames to store the CSV dataset after reading it.

```
tam_df.head(1)
```

	tweet_id	timestamp	text	name	dogtype	rating	retweet_count_x	favorite_count_x	retweet_count_y	favorite_count_y	retweet_count	favorite_count	jpg_url	p1	p1_conf	p1_dog
0	892420643555336193	2017-08-01 16:23:56	This is Phineas. He's a mystical boy. Only eve...	Phineas	none	1.3	8264.0	37871.0	8264.0	37871.0	8264.0	37871.0	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	orange	0.097049	False

1.1 First Insight: Correlation of dog rating and Retweet Count/Favourite Count

1.1.1 Assumption

My first insight will examine the correlation between the **Rating** and the **retweet_count** and **favourite_count**. My obvious assumption is that dogs with higher ratings will have higher amounts of retweets and favorite markings. Therefore I’m assuming that I can interprete the rating as a single value by taking the division of numerator and denominator. The resulting correlation matrix between every column in my dataset is as following:

	tweet_id	rating	retweet_count_x	favorite_count_x	retweet_count_y	favorite_count_y	retweet_count	favorite_count	p1_conf
tweet_id	1.000000	0.409900	0.400022	0.553450	0.400022	0.553450	0.400022	0.553450	0.149871
rating	0.409900	1.000000	0.229428	0.259820	0.229428	0.259820	0.229428	0.259820	0.097778
retweet_count_x	0.400022	0.229428	1.000000	0.821638	1.000000	0.821638	1.000000	0.821638	0.062440
favorite_count_x	0.553450	0.259820	0.821638	1.000000	0.821638	1.000000	0.821638	1.000000	0.084060
retweet_count_y	0.400022	0.229428	1.000000	0.821638	1.000000	0.821638	1.000000	0.821638	0.062440
favorite_count_y	0.553450	0.259820	0.821638	1.000000	0.821638	1.000000	0.821638	1.000000	0.084060
retweet_count	0.400022	0.229428	1.000000	0.821638	1.000000	0.821638	1.000000	0.821638	0.062440
favorite_count	0.553450	0.259820	0.821638	1.000000	0.821638	1.000000	0.821638	1.000000	0.084060
p1_conf	0.149871	0.097778	0.062440	0.084060	0.062440	0.084060	0.062440	0.084060	1.000000

1.1.2 Evaluation

The only strong correlation with a practical relevance can be observated between the **retweet_count** and the **favorite_count**. That means a dog tweet that will be often retweeted will also be often favorised. I guess this correlation is reasonable. Unfortunately there is a almost non-existent correlation between **retweet_count** and the rating. Although there is a small correlation between the rating and the favorite and retweet counts it's not enough relevance for pointing this out.

1.2 Second Insight: High Confidence in Image Classification for "Dogtionary" labeled dogs

1.2.1 Assumption

Now I will try to verify my assumption, that dogs that got a label from "WeRateDogs" like **"doggo"** should have a very high level of mean confidence compared to the entries without a label. I have this assumption, because I believe that a dog that can be labeled according to the "WeRateDogs Dogtionary", should have a very typical dog appearance and thus can be classified by a estimator with a high confidence.

Mean over all confidence levels of estimated dog pictures: 0.60205079452519372

Mean over confidence level with a label from "WeRateDogs": 0.60181255880000006

1.2.2 Evaluation

There is a slightly decrease of the mean confidence for the dogs that got a label according to the "WeRateDogs Dogtionary". For me this isn't any evidence for a better estimation confidence for dogs with a classification label.

1.3 Third Insight and Visualisation: Increase of Social Media Activity over time

1.3.1 Assumption

As many social media trends become more succesful respectively get more user interaction simply by their level of awareness over the user base it's an interesting assumption for me, that the interaction with "WeRateDog" postings will increase at all independently from the objective quality of a posting.

1.3.2 Visualisation

I did reorder the rows in the dataset for following the progress in time for having no issues to plot this relation:

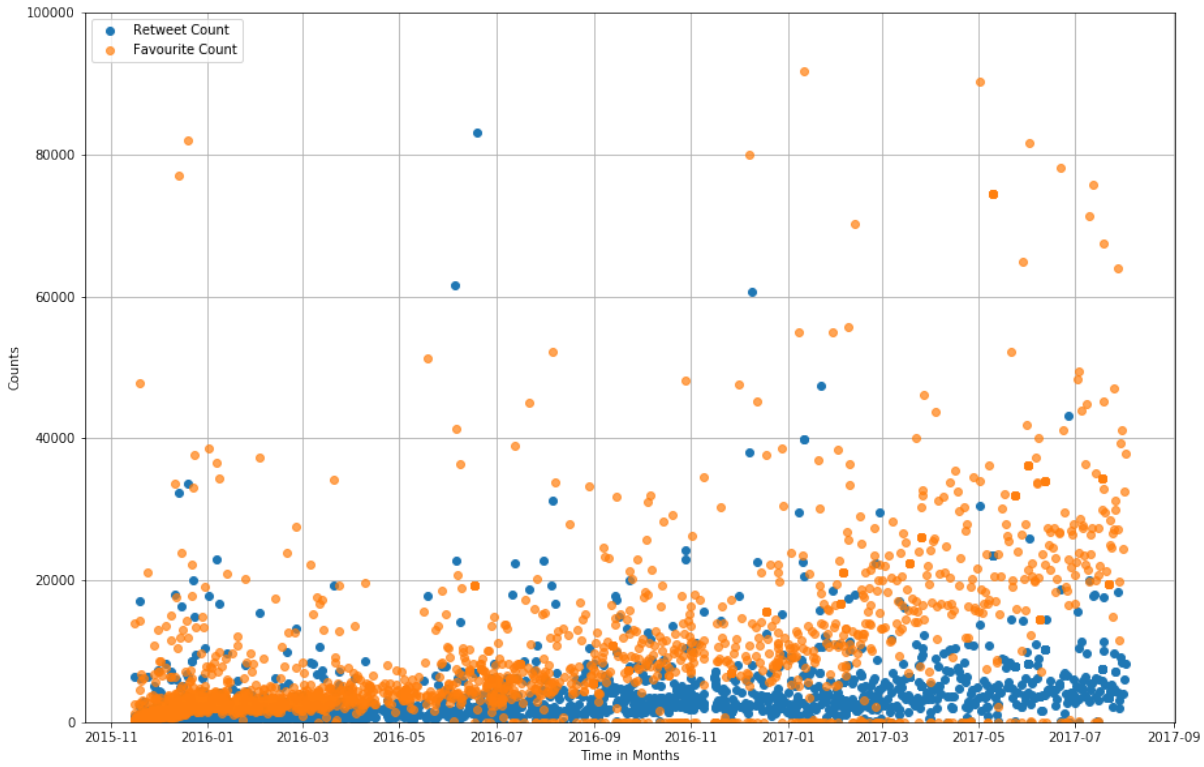

Figure 1: Matplotlib Visualisation

It seems, that the Favourite Count will increase at all with progress of time. To proof this observation I will calculate the **Linear Regression** for both, the Favourite Count and the Retweet Count over the time to verify it.

1.3.3 Visualisation with Linear Regression

For this step it's crucial to remove any null or infinite values in the used columns of the dataset. Otherwise the calculation of the Linear Regression it will only produce nonsense data.

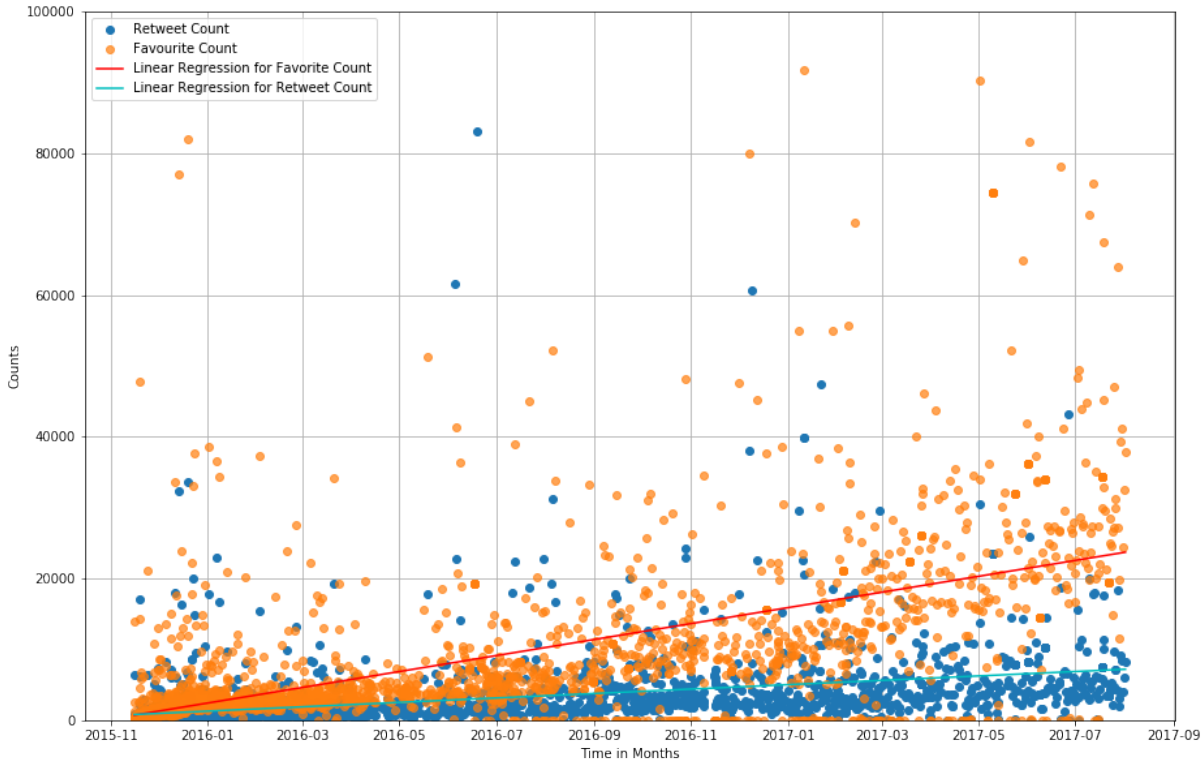

Figure 2: Matplotlib Visualisation

Actually, there is an obvious increase for the Favourite account over time. It seems, that a quadratic polynomial regression would match the course of the dots in a slightly more accurate way but I think it's goood enough to verify the first observation.

1.3.4 Evaluation

To explain the difference in increase we can ask for our own behaviour on social media like Twitter. The effort of will to like or favourise a post is much smaller than retweeting it and reveal a stronger commitment, i.e. a stronger opinion to the content of the post.

Important: Overall we have to take into consideration a specific factor to correct the rating of "WeRateDogs" posts based on Retweet Counts and Favourite Counts if observing the whole timeline.

1.4 Conclusion

I was able to point out a relationship of social media activity to the progress of time. This is a very interesting insight, as it qualifies the evaluation of social media metrics in different points of time.