REPORT: WRANGLE AND ANALYSE DATA PROJECT

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Image via Boston Magazine

This report comprises of the summary of the data wrangling of WeRateDogs Twitter archive Data.
I worked with three datasets for this project, they include:

A file on hand provided by Udacity, which I downloaded manually, twitter_archive_enhanced.csv. It contains basic tweet data of 2356 tweets from November 15th 2015 to August 1st 2017.

A file hosted on the Udacity server that I downloaded programmatically, image_predictions.tsv. It contains a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction.

A file named tweet_json.txt containing data including but not limited to tweet IDs, retweet count and favorite count. This was gotten using Python's Tweepy library to access Twitter's API.

While assessing this datasets, I came across eleven quality issues and two tidiness issues, which I cleaned up and then I merged the dataset into one master dataframe.

Quality issues

1.archive missing values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp

2.archive timestamp is object datatype instead of datetime

3.archive there are 181 retweeted tweets

4.json_data missing values across multiple columns

5.json_data user column is a duplicate of id and id_str

6.json_data id and id_str columns have the same values

7.json_data created_at and archive table timestamp have the same values but different titles

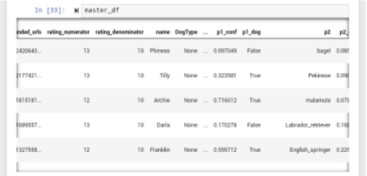
8.json_data has three columns with no values

9.json_data id should be tweet_id

Tidiness issues

1.archive table doggo, floofer, pupper and puppo columns should be one column

2.predictions table tweet_id column is arranged serially



```
In [33]: master_df
```

nded_urls	rating_numerator	rating_denominator	name	DogType	...	p1_conf	p1_dog	p2	p2_
2420643...	13	10	Phineas	None	...	0.097049	False	bagel	0.085
2177421...	13	10	Tilly	None	...	0.323581	True	Pekinese	0.090
1815181...	12	10	Archie	None	...	0.716012	True	malamute	0.078
1689557...	13	10	Darla	None	...	0.170278	False	Labrador_retriever	0.168
1327558...	12	10	Franklin	None	...	0.555712	True	English_springer	0.225

Master dataframe

Insights

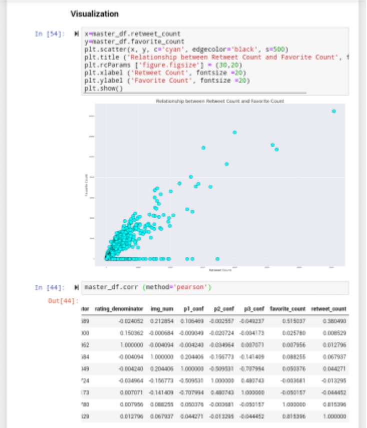
1.Favorite count has a minimum of 0, a maximum of 144914 and a mean of 6704

2.Retweet count has a minimun of 1, a maximum of 70811 and a mean of 2325

3.The master dataset has a total colunm number of 22 which includes tweet_id, timestamp, source, text, expanded_urls, rating_numerator, rating_denominator, name, dogtype, jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog, favorite_count, retweet_count.

4.P2 has the highest True count next to P1 and then P3 5.The most occurring rating numerator is 12 6.Using df.corr() to clarify the result of the scatterplot, it shows that retweet count and favorite count having a correlation of 0.815396 which is greater than zero confirms it has a positive correlation.

Visualization



Visualization

In [54]:

```
x=master_df.retweet_count
y=master_df.favorite_count
plt.scatter(x, y, c='cyan', edgecolor='black', s=500)
plt.title ('Relationship between Retweet Count and Favorite Count', f
plt.rcParams ['figure.figsize'] = (30,20)
plt.xlabel ('Retweet Count', fontsize =20)
plt.ylabel ('Favorite Count', fontsize =20)
plt.show()
```

In [44]:

```
master_df.corr (method='pearson')
```

Out[44]:

tor	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	favorite_count	retweet_count
389	-0.024052	0.212854	0.106469	-0.002557	-0.049237	0.515037	0.380490
000	0.150362	-0.000684	-0.009049	-0.020724	-0.004173	0.025780	0.008529
362	1.000000	-0.004094	-0.004240	-0.034964	0.007071	0.007956	0.012796
384	-0.004094	1.000000	0.204406	-0.156773	-0.141409	0.088255	0.067937
349	-0.004240	0.204406	1.000000	-0.509531	-0.707994	0.050376	0.044271
724	-0.034964	-0.156773	-0.509531	1.000000	0.480743	-0.003681	-0.013295
173	0.007071	-0.141409	-0.707994	0.480743	1.000000	-0.050157	-0.044452
780	0.007956	0.088255	0.050376	-0.003681	-0.050157	1.000000	0.815396
329	0.012796	0.067937	0.044271	-0.013295	-0.044452	0.815396	1.000000

Linear correlation between Retweet count and Favorite count

This project was all together engaging and fun.