Data Wrangling Project

Introduction

The main purpose of this project is to use real world data to wrangle (gather, assess, clean) and then apply analysis with visualizations. The data used was from the Twitter account 'WeRateDogs'

Our Data

- Enhanced Twitter Archive
- Additional Data via the Twitter API
- Image Predictions File

1st we have imported Libraries then moved to 2nd part of gathered need data

➢ Data Gathering

1- Importing the Enhanced Twitter archive file

tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id	re
892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	Twitter for iPhone	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	NaN	
892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	Twitter for iPhone	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10 https://t.co/0Xxu71qeIV	NaN	
891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	Twitter for iPhone	This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co/wUnZnhtVJB	NaN	
891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	Twitter for iPhone	This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD36da7qLQ	NaN	
891327558926688256	NaN	NaN	2017-07-29 16:00:24	Twitter for iPhone	This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and should be respected as such.	NaN	

2- Importing the image predictions file:

```
images_predictions=pd.read_csv('image_predictions_download.tsv',sep='\t')
print(images_predictions.shape)
images_predictions.columns
```

Out[111]:

jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
/pbs.twimg.com/media/CUx2F6IVEAAvFev.jpg	1	web_site	0.901552	False	borzoi	0.026660	True	Chihuahua	0.012438	True
bs.twimg.com/media/C3MVTeHWcAAGNfx.jpg	2	Eskimo_dog	0.524454	True	Siberian_husky	0.467678	True	malamute	0.004976	True
s.twimg.com/media/CYASi6FWQAEQMW2.jpg	1	minibus	0.401942	False	llama	0.229145	False	seat_belt	0.209393	False
bs.twimg.com/media/CdeUKpcW0AAJAWJ.jpg	1	borzoi	0.490783	True	wire-haired_fox_terrier	0.083513	True	English_setter	0.083184	True
//pbs.twimg.com/media/CY816snW8AYltrQ.jpg	1	Cardigan	0.614231	True	skunk	0.139392	False	toilet_tissue	0.031158	False
pbs.twimg.com/media/C33P8PrUcAMiQQs.jpg	3	patio	0.272972	False	window_screen	0.131295	False	boathouse	0.046393	False
://pbs.twimg.com/media/CVZjOktVAAAtigw.jpg	1	Pembroke	0.582560	True	Cardigan	0.258869	True	nipple	0.033835	False
s.twimg.com/media/CnnKCKNWgAAcOB8.jpg	2	golden_retriever	0.872385	True	Labrador_retriever	0.099963	True	cocker_spaniel	0.006051	True
//pbs.twimg.com/media/CUi5M7TXIAAY0gj.jpg	1	Arabian_camel	0.999614	False	bison	0.000228	False	llama	0.000067	False
bs.twimg.com/media/CZhn-QAWwAASQan.jpg	1	Lakeland_terrier	0.530104	True	Irish_terrier	0.197314	True	Airedale	0.082515	True
pbs.twimg.com/media/CV_cnjHWUAADc-c.jpg	1	dough	0.806757	False	bakery	0.027907	False	French_loaf	0.018189	False
bs.twimg.com/media/CU3be0SWEAEqb7I.jpg	1	window_shade	0.583427	False	giant_schnauzer	0.062215	True	window_screen	0.039941	False
bs.twimg.com/media/CUdloW8WEAAxB_Y.jpg	1	bustard	0.380772	False	pelican	0.100554	False	crane	0.084713	False
'pbs.twimg.com/media/CVHIhi2WsAEgdKk.jpg	1	park_bench	0.194211	False	water_bottle	0.071870	False	beacon	0.053433	False
//pbs.twimg.com/media/C8vgfTsXgAA561h.jpg	3	Shetland_sheepdog	0.759907	True	collie	0.107405	True	Pembroke	0.052335	True

3- Importing the Json file:

	tweet_id	retweet_count	favorite_count
0	892420643555336193	8853	39467
1	892177421306343426	6514	33819
2	891815181378084864	4328	25461
3	891689557279858688	8964	42908
4	891327558926688256	9774	41048

- **Assessing data**

We have loaded and checked data from our files using jupyter notebook as our IDE then we started to i nvestigate and to assess our date including tweet id's – none values – duplication too in addition to numerators rating values and denominators, we have also checked a random dog image then moved to Quality issues and tidiness

➢ Quality

1-Source Data cannot be read easily.

2- we have retweets and replies, we should keep original tweets only.

3-the invalid tweet_id data type (intger instead of string)

4- we have false names in the archived_twitter due to false extraction from the tweet text (ex: a, an , ...)

5- The name column in the archived_twitter has names(None) , and must be be replaced with nan for dataset to be more consist .

6- Pertaining the numerator values in the archived twitter they are not correct also some of them are a possible outliers (1766, 420) and should be invistigated and removed in case they're outliers.

7- The numerator values in the archived_twitter have some wrong extracted values and should be corrected

8- Some tweets don't have images included and should be deleted

9-Pertaining the denominator in the archived_twitter values they're a large value which represented many dogs rating.

10- In the dog's type columns (doggo, floffer, popper,) whiche have a None values instead of nan so we need to create a single column "dog_type " to merge all of values in our four columns

11- extract the source of the tweet (iphone, ...) from the source column from the twiitter_data

12-Adjust the timestamp to datetime format

13-440 Rating numerators are lesser than 10

➢ Tidiness issues

1- Dog type data are 4 “doggo, floofer, pupper, puppo” columns and should be merged into single column

2- Merge the 3 dataframes into single dataframe

➢ cleaning Data

1- Copying the dataframes before the the cleaning process of our gahtered data

2- Merge the 3 dateframe (archived_twitter,images_predictions, twitter_data)into single dataframe

3-Removing tweets that doesn't have images included

4- Adjust the timestamp to datetime format

5- consolidate the dog types in single column (dog_classification) instead the 4 dog types (doggo,floofer, pupper, puppo)

6- Remove the retweets and replies and consequently remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp

7- Creating a new dog_classification column using the image prediction data to merge the dog types in a single column then remove p1 , p1_conf , p2_conf , p2 , p3 , p3_conf columns

8-adjusting the column Name

9-adjusting numerator and denominator column

At last all data have been stored in CSV file 'twitter_archive_master.csv'