WeRateDogs Twitter

Data Wrangle Report :-

1- Data Gathering :-

A) `Twitter_archive_enhanced.csv` gathered from Udacity's classroom contains the archived tweets data.

B) `Tweet_json.txt` gathered from Udacity's classroom because twitter didn't accept the API request because of the recent security preach in their API, contains the archived twitter information like retweets counts, favorites counts , media url , etc.

C) `image_predictions.tsv` gathered from Udacity's classroom , contains the tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network

2-Data Assessment:-

2.1)Quality Issues

A) The Twitter Archive Dataframe arch_df (2356 row)

- Missing Values in in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp , probably will be dropped

```
6   retweeted_status_id         181 non-null    float64
7   retweeted_status_user_id    181 non-null    float64
8   retweeted_status_timestamp  181 non-null    object
```

- source column is written in the html format needs to be string

source
<a href="http://twitter.com/download/iphone" r...

- timestamp column needs to be transformed into a timestamp data type instead of string

```
3   timestamp                   2356 non-null   object
```

- expanded_urls

  1. 59 Missing values
  2. Duplicated links in a single cell
  3. 137 duplicated rows with the same expanded url
  4. Non twitter Links
  5. Some links aren't from twitter

```
1-5 Random samples
====================
https://twitter.com/dog_rates/status/780931614150983680/photo/1
====================
https://twitter.com/dog_rates/status/718460005985447936/photo/1
====================
https://twitter.com/dog_rates/status/875144289856114688/video/1
====================
https://twitter.com/dog_rates/status/817777686764523521/video/1
====================
https://twitter.com/dog_rates/status/710283270106132480/photo/1,https://twitter.com/dog_rates/status/710283270106132480/photo/1
====================
=====================================
2.Missing Values = 59 out of 2356
=====================================
3-Duplicated Values = 137 out of 2356
```

- remove retweets spoted in the api_df.retweet_status isn't null by their id
- some values in rates_denomenator and rating_numerator doesn't match the rates in text
- text : Contains some ads not ratings in the column "WeRateDogs stickers are here and they're 12/10! Use code "puppers" at checkout 🐶🐾 Shop now: https://t.co/k5xsufRKYm https://t.co/ShXk46V13r" id = 709901256215666688
- name : Null (missing) values are expressed as None (745) Value , Some Names are only one letter "a" (55 name)

B) The Twitter API Dataframe api_df

- The Data Frame missing the following tweets with ids 771004394259247104, 888202515573088257 compared to the archived Data Frame arch_df
- Not all tweets has images
- 79 tweets are retweets which we don't want

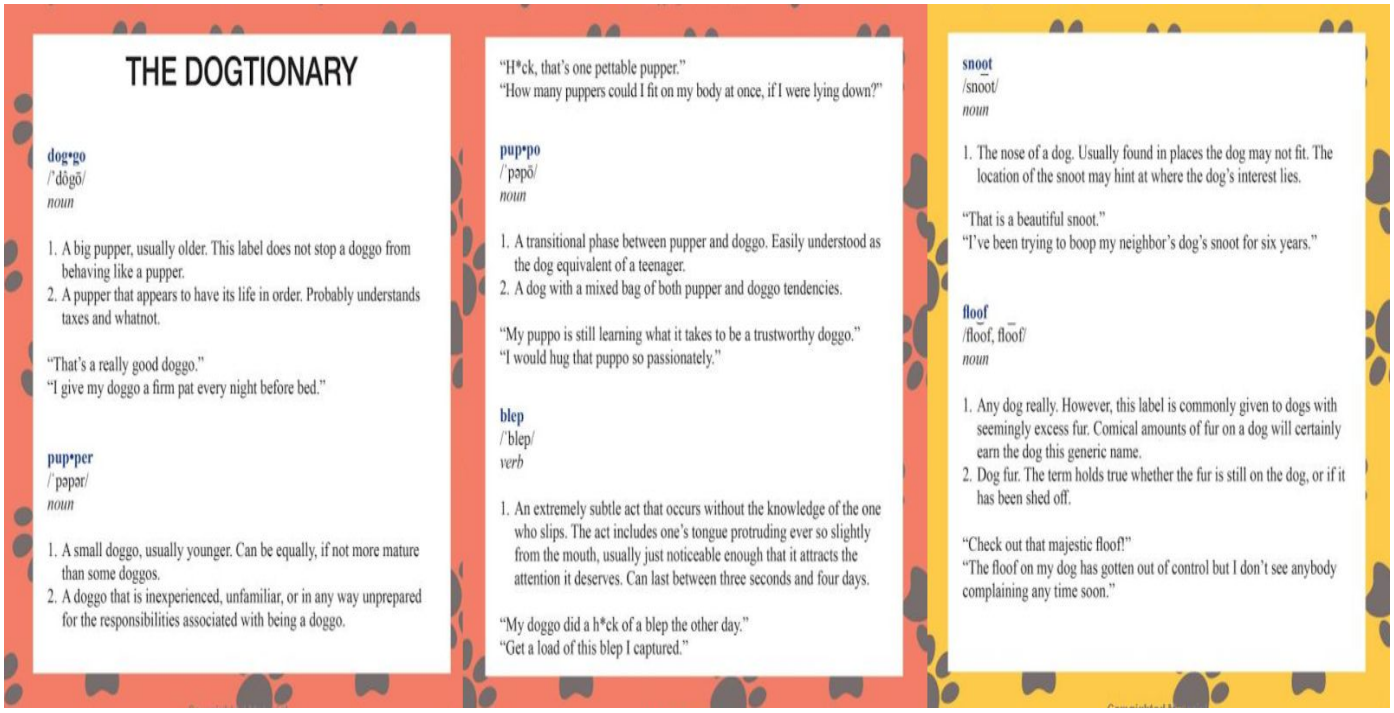
C) The Image Predictions Dataframe predictions_df

- 281 Missing Predictions compared to the archived tweets data set
- 66 Duplicated jpg_url .. since there's no duplicated tweet_id , so the duplicated image url is either a an error of Original Data gathering for the data set feeding the neural network , or they're retweets, this can be checked from the api_df dataframe
- Some Dog breed names in p1 isn't Capitalized and the delimiter is “_”.

2.2)Tidiness Issues

1. The Twitter Archive Dataframe arch_df

- text column : </br> 1. Hashtag need to be in a single column (entities column in api_df will help)
- expanded_urls column : </br> 1. Some links are Fundraising need to be in a single column
- doggo,floof(er),pupper,puppo columns needs to be melted into a single column "Dog_Stage" and extract another values snoot , blebif available according to the dictionary



THE DOGTIONARY

dog•go
/ˈdôgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

“That’s a really good doggo.”
“I give my doggo a firm pat every night before bed.”

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

“H*ck, that’s one pettable pupper.”
“How many puppers could I fit on my body at once, if I were lying down?”

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

“My puppo is still learning what it takes to be a trustworthy doggo.”
“I would hug that puppo so passionately.”

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one’s tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

“My doggo did a h*ck of a blep the other day.”
“Get a load of this blep I captured.”

snoot
/snoot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog’s interest lies.

“That is a beautiful snoot.”
“I’ve been trying to boop my neighbor’s dog’s snoot for six years.”

floof
/floof, floof/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

“Check out that majestic floof!”
“The floof on my dog has gotten out of control but I don’t see anybody complaining any time soon.”

2. The Twitter API Dataframe `api_df`

- The `entities` column will help us to attaining the missing values of expanded urls in the `arch_df` data frame , also the hashtags

3. All of the dataframes need to be merged into one dataframe since we only have one observational unit ,the "Tweet".

3-Data Cleaning :-

All the quality and tidiness problems addressed in the assessment section have been cleaned successfully and the final product of the cleaning section is the `twitter_archive_master.csv(1992 entry)` , that contains the cleaned and combined version of the 3 files addressed above.