**Wrangle\_report**

**GATHERING**

The three data sets in the wrangling notebook have been gathered differently:

a). The 'WeRateDogs' twitter archive Directly downloaded from the link provided in Udacity classroom and read into my notebook using pandas.

b). The Tweet Image Predictions

Downloaded programmatically using the Requests library and the URL provided in Udacity classroom. The URL is:<https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv>

c). Additional data from the Twitter API

Downloaded tweet\_json.txt the resulting data from twitter\_api.py, then read this tweet\_json.txt file line by line into a pandas Data Frame with (at minimum) tweet ID, retweet count, and favorite count.

**ASSESING**

**General issue**

* The "tweet\_id' column should be of the 'string' datatype and not 'INT'
* Datasets are not of the same length
* The data frames ‘twitter\_archive\_enhanced’, ‘image\_predictions’ and ‘twitter\_api’ should be one table.

**Twitter Archive Data**

**Quality**

* The datatype of timestamp to “datetime”
* The columns “in\_reply\_to\_status\_id”, “in\_reply\_to\_user\_id”, “retweeted\_status\_id”, “retweeted\_status\_user\_id”, and “retweeted\_status\_timestamp” have lots of NA values.
* The column 'source' contains html residues.
* The column 'expanded\_urls' have NaN values and also some rows contain double 'expanded\_urls'
* A total of 440 entries have a Numerator value less than 10.
* A total of 23 ratings do not have the denominator value as 10.
* The "name" column has a total of 745 'None' strings and multiple stop word present.

**Tidiness**

* The column 'text' in has two variables, the “text” and the “tweet\_link”
* The observations doggo, floofer, pupper and puppo are represented as four variables.

**Image Predictions Data**

**Quality**

* The prediction is an uppercase and lowercase mix, also there are "\_" in the breed name.
* There are 324 records without the prediction of the dog breed

**Tidiness**

* The column heads 'p1', 'p1\_conf', 'p1\_dog', 'p2', 'p2\_conf', 'p2\_dog', 'p3', 'p3\_conf', 'p3\_dog' are not informative.

**CLEANING**

* Change the datatype of ‘tweet\_id’ to string in the 3 data frames (‘twitter\_archive\_enhanced’, ‘image\_predictions’, ‘twitter\_api’) using a for loop and ‘astype’.
* Change the datatype of timestamp to ‘datetime’.
* Drop ‘in\_reply\_to\_status\_id’, ‘in\_reply\_to\_user\_id’, ‘retweeted\_status\_id’, ‘retweeted\_status\_user\_id’ & ‘retweeted\_status\_timestamp’ after filtering them to only tweets.
* Used regular expression to extract the source.
* To clean NaN values in 'expanded \_urls' I used ‘notna()’ to retain the rows with the value present. I however did not manage to clean for the double 'expanded\_urls'.
* Use a for loop and regular expression to detect all the text with decimal numbers as the rating numerator, Print the text, numerator and the text index. I then manually replace the rating numerator with the decimal value.
* Find the texts with two denominator ratings using a for loop and a list of all denominators not equal to zero then manually clean by choosing a rating with a denominator of 10.
* For the 745 'None' names and stop words in the column 'name', I sampled a few text to see the pattern or the placing of the dog name then used regex to extract the correct dog name.
* Create a new column with the tweet\_link that is extracted from the column 'text' using regex.
* Create a new column dog\_stage with all the four dog stages ('doggo', 'floofer', 'pupper', 'puppo') then drop the 4 columns ('doggo', 'floofer', 'pupper', 'puppo'). Use a for loop to replace all the "none" with a space
* Change the values under the columns 'p1', 'p2' and 'p3' to be all lowercase using .str.lower()

For the quality issue 324 records missing I was undecided whether to drop then or impute with the modal class so I left it like that

* Rename 'p1', 'p1\_conf', 'p1\_dog', 'p2', 'p2\_conf', 'p2\_dog', 'p3', 'p3\_conf', 'p3\_dog'

**Merging and storing the data**

* merge the data frames using an inner join
* stored the cleaned and merged data as "twitter\_archive\_master.csv".

**References**

1.<https://classroom.udacity.com/nanodegrees/nd002-alg-t2/parts/cd0015/modules/356d26d4-2237-46a0-ab78-e0ea93d9501a/lessons/ls2232/concepts/c32c5e23-8f28-45e1-8143-fb5ab20953e7>

2.<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.apply.html>

3.<https://www.w3schools.com/python/python_regex.asp>

4.<https://rachelchen0104.medium.com/weratedogs-project-part-i-data-assessment-7bf5f59b0b5e>

5.<https://www.guru99.com/python-regular-expressions-complete-tutorial.html>

6.<https://www.kaggle.com/code/rihamrafat/we-rate-dogs-data-wrangling>

7.<https://github.com/meet3012/We-Rate-Dogs-Twitter-Archive-Data-Wrangling>

​