Wrangle Report

1. Data Gathering

The data was gathered from 3 different sources with different formats

1.1 Enhanced Twitter Archive

- The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).
- The data is downloaded manually by this link : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
- Output : archive_df

1.2 Image Predictions File

- 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.
- The data is downloaded programmatically by the url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Output : image_prediction_df

1.3 Additional Data via the Twitter API

- The twitter api data integrate with the enhanced twitter archive that it provides favorite counts and retweet counts for each tweet and we can get more if wanted.
- I got the data without twitter account and I have used:
  1. twitter_api.py: This is the Twitter API code to gather some of the required data for the project.
  2. tweet_json.txt: This is the resulting data from twitter_api.py.
- output : api_df

2. Data Assessment

- I have used both of the assessment methods (visual and programmatic)
- In the visual method I have used the excel application.

- In the programmatic method I have used jupyter notebook cells and pandas library functions ex: head(), describe(), duplicated(), ..etc

Quality issues:

1.archive_df

- tweet_id is int not string
- timesamp column is object type not datetime type
- none value instead of nan in name, doggo, floofer, pupper and puppo columns
- missing values in name column
- missing values in expanded_urls column
- validity issue in rating_numerator that is less than 10
- validity issue in rating_denominator that is more than 10
- 78 rows of replies and 181 rows of retweets

2.image_prediction_df

- missing rows (2075 instead of 2356)
- int data type for tweet_id column instead of object
- lowercase and uppercase in p1, p2 and p3 columns

3.api_df

- missing 2 rows (2354 instead of 2356)
- tweet_id is int not string

Tidiness issues:

1. dog stages presented in 4 columns instead of one column
2. p1,p2 and p3 values represented as columns
3. all the data can be represented in 1 dataframe instead of 3 dataframes

3. Clean Data

- In this process we use clean workflow (define, code and test)
- I started solving the tidiness issues at first and the I solved the quality issues

Tideness issues:

Issue	Solution
1.dog stages values presented in 4 columns instead of one column	- replace the 4 columns value with one column named dog_stage
2. p1,p2 and p3 values represented as columns	- convert p1, p2 and p3 columns value to one prediction column
3. all the data can be represented in 1 dataframe instead of 3 dataframes	• merge the archive_df and api_df on tweet_id and present as all_tweet_data • merge all_tweet_data and imag_predictions_df and present as master_data

Quality issues :

Issue	Solution
1. 78 rows of replies and 181 rows of retweets	- Drop the the retweets and replies and keep original tweets only
2. validity issue in rating_numerator that is less than 10 and weird values	- delet all the rows with rating_numerator values more than 15 and less than 6

3. validity issue in rating_denominator that is more than 10	- set the denominator value to 10 for all the rows
4. missing values in name column	- replace all the none values, lowercase strings and even the strings with length less than 3 with nan if name is the tweet text
5. tweet_id is int not string	- change the tweet_id data type to string
6. timesamp column is object type	- change timestamp data type to datetime
7. prediction level is int	- change the prediction level column data type to string
8. mix of lowercase and uppercase names in the prediction column	- lowecase the prediction column values