Wrangle Report

By: Mahmoud Esmail

Introduction:

WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, One column the archive does contain though: each tweet's text, which we 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, data file have been filtered for tweets with ratings only (there are 2356)

Data Wrangling Details:

- Gathering data
- Assessing data
- Cleaning data

1. Gathering Data:

Data were gathered from different resources:

1) The WeRateDogs Twitter archive. The filename **twitter_archive_enhanced.csv**
   Were already given so we just imported to the workspace and used
   By the help of pandas dataframe
   This file contains various columns that will help us in our future work

Ex: tweet_id, timestamp, source, text, rating_numerator, rating _denominator, etc.

2) The tweet image predictions file **image_predictions.tsv** , we downloaded it programmatically by the help of requests library.
   The file contains columns of hyper links of dog photos
   And their tweet_id and also the confidence of prediction of the dog photo being correct

3) The JSON File which is named tweet_json.txt
   Contains information about the tweets itself
   Ex: tweet_id, favorite_count, retweet_count
   I used the read_json method from pandas library
   To read the values in the text and saved it into a dataframe

Assessing Data

- I assessed this dataset after merging all files and removed unnecessary columns from the dataset
- The tool that helped me in visual assessment was printing out the complete dataset and look up for any unusual occurrence or faulty error
- Some tools that helped me in programmatic assessment were .info(), value_counts(), .sum(), .duplicates() , etc..
  Most of these tools are to figure the outliers in the data, typos and unneeded occurrencies

Cleaning Data:

Here's the data issues that I figured out and then cleaned most of them

archive_df :

- Remove Data contains retweets and leave the original content
- change column timestamp data type to appropriate data type
- change the form of source column to contain only the needed text
  that describes the source of navigating.
  in other words, Remove the HTML tags that are not needed
- change data type of tweet_id into object
- some inaccurate names like: 'a', 'an', 'the', 'very', 'by', 'al'
- Replace name values which equal to None
- unnecessary columns can be removed
- change the rating_denominator and numerator datatype into float

image_predictions_df :

- Remove jpg_url duplicated items
- Remove unnecessary columns
- tweet_id is int

tweet_counts :

- Remove retweets and keep original tweets

Tidiness Issues:

- The data in 3 tables could be merged into 1 dataset, because they describe one thing
- We have dog stages doggo, floofer, pupper and puppo in 4 different columns instead of one column in **archive_df**

After All, I finished cleaning data and stored a copy of dataframe into a csv file called
'twitter_archive_master.csv'

Conclusion:

Data wrangling is an essential skill for any data analyst, propably most of data analysts use this skill in every project they go through since the data is ubiquitous and most of the data in the world is untidy, so it requires some modern solutions to be able to wrangle data before making exploration and visualization to your data to discover the patterns in it