Data Wrangling Report: WeRateDogs Twitter Account

Data Wrangling Process was divided into 3 Parts:

a) Data Gathering
b) Data Assessment
c) Data Cleaning

The below python libraries have been imported into the project to assist with Data Wrangling:

a) Pandas
b) Requests
c) Numpy

Data Gathering

As part of the Data Gathering process, we have imported the below files into the project that contain our data from different sources:

a) twitter-archive-enhanced.csv
b) responses.tsv
c) tweet_json.txt

Each of the files contains fields as per below, that each hold data.

a) twitter-archive-enhanced.csv:

tweet_id,in_reply_to_status_id,in_reply_to_user_id,timestamp,source,text,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,expanded_urls,rating_numerator,rating_denominator,name,doggo,floofer,pupper & puppo

b) Responses.tsv

tweet_id,jpg_url,img_num,p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf & p3_dog

c) Tweet_json.txt

id,retweet_count,favorite_count,created_at & full_text

Data Assessment

It was observed that the gathered data had the below Quality and Tidiness Issues which were noted using visual and programmatic assessment.

a)Quality Issues

- The timestamp columns in the twitter-archive-enhanced.csv file was noted to have trailing zeros and '+' symbol in each of the date values.
- b)Some of the records in p1,p2 & p3 columns of the responses.tsv file had values with Title case, while others were using Lower case. For uniformity, all the values were changed to Title case

- Some of the records in the name column of the twitter-archive-enhanced.csv file had values with Title case, while others were using Lower case. For uniformity, all the values were changed to Title case
- The Datatype of the timestamp column in the twitter-archive-enhanced.csv file was an object. The Datatype was changed to datetime for date and time values
- The Datatype of the retweeted_status_timestamp column in the twitter-archive-enhanced.csv file was an object. The Datatype was changed to datetime for date and time values
- The None values in the doggo, floofer,pupper,puppo,name columns in the twitter-archive-enhanced.csv file were changed to Null values
- The Datatype of the tweet_id column in the twitter-archive-enhanced.csv file was changed to the correct Datatype
- The Datatype on the id column in the tweet_json.txt file needed to be changed to the correct Datatype
- The Datatype on the tweet_id column in the response.tsv need to be changed to the correct Datatype
- Duplicate values were also checked across the imported files.

b) Tidiness Issues

- Twitter-archive-enhanced.csv,tweet_json.txt and responses.tsv files needed to be merged into one master file.
- Redundant columns in the master file needed to be dropped.

Data Cleaning

Results:

- All the observed issues in the dataset were resolved.
- No duplicate values were noted in the dataset.
- All the 3 gathered files were merged into one master dataset for analysis.