Data Wrangling Report

Project Objectives:

The objectives was to gather, assess and clean the WeRateDogs twitter data. Then store, analyze, visualize, and give insights on the visualization.

Gathering Data steps:

1. ‘twitter_archive_enhanced.csv’ was manually downloaded.

2. ‘image_predictions.tsv’ file was programmatically downloaded using the Requests library.

3. For gathering tweets from WeRateDogs profile, I used Twitter API and python’s tweepy library. All those data were store in ‘tweet_json.txt’ with only tweet_id, retweet_count, and favorite_count column.

Assessing and Cleaning steps:

During the assessment stage I have found several quality and tidiness issue of data, which I will describe below.

Quality issues

1. tweet_id has dtype int64 and should be object.

   Solution: By changing the variable type to object.

2. Remove all retweets (status and reply) rows--> retweeted_status_id, in_reply_to_status_id.

   Solution: We have seen on the assessment that there were some non-null values on these columns. So, by only keeping the null value rows we achieved it.

3. Need to drop unwanted columns --> retweeted_status_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp.

   Solution: achieved by dropping those columns.

4. Timestamp is in str format not datetime format.

   Solution: By changing the format to datetime.

5. tweets with no image (out of 2356 rows, 2297 have image).

   Solution: by dropping the NaN value rows from expanded_urls and jpg_url columns.

6. Many dog names are not correct or missing. Several has names like 'a', 'quite’, 'an’, 'the' etc.

   Solution: From the assessment we have seen that all these non-names were lower case. So, I created a variable stop_word where I stored all those names. Then finding the index of those rows and replace those names with ‘None’ and the finally replace with ‘NaN’ value.

7. p1, p2, p3 columns should be all lower case and '_', '-' needs to be removed.

   Solution: On these columns dog_breed names were inconsistent. Some name started with upper case, and some were lower case. Also, some names had ‘_’ and ‘-‘ in them. So, I removed those underscore and dash by using regular expression and replace() method and capitalized all name by using capitalized() method.

8. For some dog style (doggo, floofer, pupper, puppo) few has more than one style.

   Solution: For this issue, there were some rows/dog_breed who were given 2 dog_class (doggo, floofer, pupper, puppo) value. So, I first get the index of the rows that had more than 2 dog_class (doggo, floofer, pupper, puppo) values. Then replace the wrong value with empty string (I figured the correct value from the text column).

Tidiness issues

1. Group the 4 different class of dog (doggo, floofer, pupper, puppo) into one column

   Solution: First I replaced all ‘None’ value with empty string. Then I concatenated all four (doggo, floofer, pupper, puppo) columns into one column name dog_class. Then I dropped the original (doggo, floofer, pupper, puppo) columns. Then I replaced empty string with ‘Nan’ value.

2. Merge twitter_archive_enhanced.csv, image_predictions.tsv and tweet_json.txt into one table.

   Solution: All three data that were gathered, they many unnecessary columns and a tweet_id column that they all had in common. So, per tidiness rule I merged all three table into one dataframe.