Wrangle report

By: (T.J.Adel) Al-Dajani

Date: January the 2nd 2022

The project I had was challenging and it helped to sharpen my data wrangling skills so much.

For this project data was gathered from three different places. WeRateDogs gave Udacity exclusive access to their twitter archive for this project in the form of a csv file. This file contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. The second source of data is tweet images that were run through a convolutional neural network to analyze these images and identifies the breed of the dog, and then it was programmatically downloaded as a tsv file.

The third and final source of data is in a file called tweet-json.txt that can be acquired from the twitter API.

After the data was gathered I started phase two of the data wrangling process which is assessing the data.

In this phase visual and programmatic assessments were made to identify some quality and tidiness issues as follows.

Quality issues

1. We have 181 retweeted_status_id entries in twitter_archive_enhanced that needs to be excluded from our data

2. change (tweet_id) column in image_predictions and twitter_archive_enhanced to object data type

3. find the source name from the source column

4. Some records with all-None stages (doggo,floofer,pupper,puppo) can be extracted from the (text) column to find the stage

5. some of the rating_denominator values are not 10 in twitter_archive_enhanced

6. Having missed names in the (name) column

7. remove columns that will not be used for the analysis

8. capitalize the first letter of each word in the p1, p2, p3 columns in the impage_predictions table

Tidiness issues

1. Change the rating_numerator and the rating_denominator to floats

2. make one column for all dog stages and drop the 4 exicting column

I handled these problem one by one in the same order as I listed them.

Cleaning this data was challenging some times I had to use stack overflow and geeks for geeks to help me with some functions this made me more resourceful and knowledgeable in multiple python libraries specially pandas.