WRANGLE REPORT FOR THE WE_RATE_DOGS WRANGLING PROJECT

The wrangling process for this project was in three (3) main phases

Gathering

The datasets came in three different files, the first one is the archive of the WeRateDogs account on twitter. This first one is csv file which I opened using read.csv in pandas library. The second file is an image prediction tsv file which I programmatically downloaded using the requests library after which I read it into a dataframe using read.csv in pandas library. The last file is a JSON.txt file which I got from Udacity because I was unable to get my twitter API approval in time for the project. This I queried using the tweepy library.

Assessing

This phase is sub-divided into two (2) phases: visual and programmatic assessment. Kindly find listed below the issues I identified after carrying out both assessments in this phase:

Quality issues

1. Most of the rows in the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted status timestamp columns in df1 have null values.

2. The column names in df2 are not clear/easy to understand

3. In df2, there's insconsistency in the first letters in colums p1, p2 and p3

4. Timestamp is in the wrong format - string instead of datetime.

5. In the name column of df1 some names are 'such','none','a','an','this','his','quite','the' these appear to be missing names and should be represented by NaN

6. From column p1, it is clear that not all the urls are images of dogs

7. Some rows have denominator less or more than 10. Some are retweets and while others have irregular ratings different from the text.

8. The number of images per tweet img_num is not necessary for this analysis

Tidiness issues

1. The dog stage - doggo, floofer, puppa and poppo - can be one column since one dog is at one stage at a time.

2. The text column has both text and url mixed in the sentence

Cleaning

In this phase I cleaned all the issues I has identified in the Assessing phase after which I merged the files and stored as the master dataset “twitter_archive_master.csv”