Wrangle and analyze data project

Data wrangling report

- This project includes data wrangling processes through gathering, assessing, cleaning data, storing, and then the analysis and visualization of the results.

- WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Steps:

1- Gathering:

- We have here three data sources for this project
  - The WeRateDogs Twitter archive: The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. We have here as CSV file twitter_archive_enhanced.csv
  - The tweet image predictions: This file image_predictions.tsv, generated according to a neural network that predicts what breed of dog (or other object, animal, etc.), I downloaded the file programmatically using the Requests library and URL provided in project details
  - Additional Data via the Twitter API: I used here tweet_json.txt file as my source instead of using twitter API as I faced an issue creating an account. Using json library I created my data frame.

2- Assessing:

- Quality

  Here I checked the issues in data, like missing records (completeness), schema (Validity), inaccurate data (Accuracy) and Consistency

  Here is a list of what I resulted:

  In “twitter archive”

  - there are 181 retweets records
  - timestamp data type is object not datetime
  - name column contain some invalid real names as 55 values 'a'
  - some tweets have more than one dog stage.
  - in rating_numerator and rating_denominator columns invalid ratings appear
  - drop columns that holds data for retweets
  - split timestamp into two columns date and time
  - source column contain html URL

  In “image-predictions”

  - 100 tweets (no retweets) in archive not existed in image_predictions file
  - 66 image url duplicates

  In “tweet_json”

  - id column have to be renamed to tweet_id

- Tidiness
  I check here data structure

  - Dog Stage column added to merge doggo, floofer, pupper and puppo columns
  - Create one column that holds rating computed from rating_numerator and rating_denominator columns
  - Merge twitter_archive and image_predictions and twitter_api data by tweet_id value

3- Cleaning:

First I copied the three data sources I worked with in previous steps,
then performed the actions regarding the assessments I made to improve quality and tidiness.

After cleaning I made storing for cleaned and final data frame resulted from merging the three datasets.