WeRateDogs Twitter archive data wrangling.

Data wrangling, also known as data cleaning, data remediation, or data munging, refers to a set of methods used to convert raw data into more usable formats. The specific strategies vary based on the data you're utilizing and the aim you're attempting to achieve.

The project aims to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The wrangling process consists of data gathering, assessing data, and cleaning data.

For the data gathering process, 3 datasets have been used, the first one is the twitter_archive_enhanced.csv dataset has been manually downloaded and loaded in Jupiter Notebook using pandas library, it contains 2356 entries and 17 columns which include Twitter id, time, ratings, name of the dog, dog category, etc. The second one is image_predictions.tsv which has to be downloaded programmatically on: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using request library, and it contains 2075 entries and 12 columns such as Twitter id, image URL, etc. The third data set is a JSON file where Each tweet's JSON data should be written to its own line and a data frame of tweet ID, retweet count, and favorite count have to be extracted from it.

For the data assessment process, the visual assessments have been done in Microsoft excel and python, while programmatic assessments have been done in Jupiter Notebook by using different methods such as dataframe.info(), dataframe.duplicated(), dataframe.head(),dataframe.isnull().sum(), etc.

Different issues have been found:

a) Quality issues

1. There are missing values in The WeRateDogs Twitter archive in IDs columns.
2. In Excel , the text column , some texts are hidden in one cell while others occupy several cells.
3. Character mixing in p1 ,p1_conf,p2 p2_conf columns in image dataset where some names start with capital letters others do not, underscore characters, etc.
4. In the dog category columns(doggo ,floofer,pupper ,puppo) have 'None' values instead of NaN in the Twitter archive dataset.
5. Timestamp is not in DateTime but it is the object.
6. IDs in the twitter_archive dataset are numbers that can be operated but they must be strings.
7. Tweet in image_data dataset is an integer that can be operated but must be a string
8. There are Duplicates in the image URL column in image datasets.
9. There are Invalid names in Name columns such as a, quiet, the, an, not ,etc.

10. Some dogs have no names.
11. Some dog ratings do not follow general rules.
12. Mixed data format in dog ratings.

b) Tidiness issues

1. There are 3 different datasets referring to one general project. All datasets should be merged since they are all referring to about one thing for easy computation.
2. The dog stage is one variable and hence should form single column. But this variable is spread across 4 columns - doggo, floofer, pupper, puppo

For data cleaning processes, all the 3 datasets have been copied and the cleaning processes were done on copies, The mentioned issues have been cleaned for each dataset separately, and after getting cleaned datasets, they have been merged into one clean dataset named twitter_archive_master which was stored back in CSV file format.

The limitation of the data wrangling process is that we have a vague goal for this project which make it hard to know which columns to focus on, which ones to drop, etc.
Another limitation is that there aren't enough metadatas to explain each column and their role in the datasets.