Documentation for data wrangling steps

Data Gathering Steps:

1) Downloaded the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv) and read it into pandas data frame df
2) Used the Requests library to download the tweet image prediction (image_predictions.tsv) and read it into data frame predictions
3) Reading provided tweet_json.txt file and appending it to an empty list. Then converting the list to a pandas data frame and selecting the id, favourite count and retweet count columns

Data Assessing Steps:

TWITTER ARCHIVE ENHANCED DATA

1) The .info method used on the Twitter archive dataset showed that The Twitter archive enhanced data has 2356 entries. The timestamp column has an object data type that needs to be fixed. We can also see that some columns have missing values, but for our analysis, the columns with missing values such as in_reply_to_status_id, in_reply_to_user_id retweeted_status_user_id, retweeted_status_timestamp, expanded URLs are of little interest.
2) The .describe method used on the Twitter Archive Dataset gives us a sense of the distribution of the quantitative variables in our data. We can see that the rating numerator and denominator columns have maximum values of 1776 and 170 respectively. This does not conform to the expected values. Also, the minimum values of the rating numerator and denominator are 0. This does not also conform to the expected values. Weratedogs use a specific rating system where the denominator is always 10 and the numerator is always greater than 10 but can never be as high as 170 or 1776. Since we are not giving a maximum value to our grading system, we will explore more the see the distribution of the ratings in both columns.
3) We can see that the count of ratings not equal to 10 is 23 from further assessment. We will drop entries with these values since they are small compared to the size of our dataset.
4) We see the distribution of the numerator ratings above has a lot of outliers that can greatly skew our analysis if not controlled. Since Weratedogs always rate dogs at or above 10. We select values from 10 to 15 for the numerator since these seem more plausible. Rating 10 has a count of 261 and rating 15 has a count of 2. The count of values between 10 and 15 is 1890. Out of 2356 entries, we have 1890 that we are going to work with which seem sensible.
5) Looking at the name column in the twitter enhanced dataset, we see that a lot of names which are not dog names are present such as infuriating etc and also None values are present. We have to find a way to handle this during the cleaning operation

6) Looking through the dog stages columns(floofer, pupper, puppo, doggo, puppo), we see that None values are present in these columns. We will be looking to merge these columns into one column for a tidier data
7) We can see that the expanded URL column has some URLs that occur twice. Upon further investigation, we see that these entries are retweets. Without digging too much, we will keep an eye out for this and upon deleting retweets confirm that all URLs occur once in the dataset

PREDICTIONS DATASET

8) Upon initial observation, The predictions dataset looks pretty neat. No null values noticed and the data types are appropriate
9) Range values seem plausible and are consistent with our data description
10) Looking at our jpg URL column, we see that some jpg URLs have a frequency of 2 and upon further investigation, they appear to be duplicate entries as all other parameters match except tweet ids. There are 66 instances
11) Looking at our p1,p2 and p3 values, we see that some names are not dog names such as web_site, seat_belt, electric_fan and so on. We will need to find a way to deal with this during our cleaning operation

TWEET JSON DATA

12) No null values present. Data types are on point.
13) We will need to rename the id column to tweet_id to merge all 3 datasets
14) No duplicates entries present

Data Cleaning Steps:
TWITTER ENHANCED ARCHIVE DATA

1) Made a copy of the original data for the 3 different datasets
2) Changed data type of Retweeted status id column to int type
3) Dropped rows with retweeted status id not equal to nan
4) Combined dog stages into one column and string values
5) Converted Time stamp column to date time format
6) Split timestamp column to year, month and day columns

7) Dropped columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp, doggo, puppo, pupper, floofer columns, expanded urls and source

PREDICTIONS DATA

8) Dropped some rows that have the jpg_urls repeated for different tweet ids(duplicate entries)
9) Dropped Img_num column in Predictions table

TWEET JSON DATA

10) Rename the id column to tweet_id

Merge all three Datasets