Data wrangling report

The datasets gathered for this report are the

1. The Twitter-archive-enhanced.csv file.
2. The Tweet-json.txt file.
3. The Image-predictions.tsv file.

After the assessment of the first dataset, the following issues were discovered and resolved:

1. Quality issues
    a. The timestamp and retweeted_status_timestamp columns were of the string datatype instead of the DateTime datatype. It was converted to the correct datatype using the astype method.
    b. There were a lot of missing values in the in_reply_to_user_id, in_reply_to_status_id, retweeted_status_timestamp, retweeted_status_id, retweeted_status_user_id and expanded_urls columns, which resulted in them getting dropped.
    c. The rating_denominator column has a minimum value of 0, which is not a valid denominator. The row got deleted from the dataframe
    d. The puppo, floofer, pupper, name and doggo columns have the string 'None' representing the missing values, were replaced with NAN values and were eventually dropped.
    e. After a visual assessment of the unique elements in the name columns, the words 'a', 'not', 'one', 'an', 'very' and others were extracted being invalid names and were dropped.
2. Tidiness issues.
    a. Multiple variables in the text column were split into three columns.
    b. Column headers are values, not variable names. The doggo, floofer, pupper, and puppo columns were merged into a column, dog_type.
    c. Multiple variables in the source column were split into two distinctive columns.

When assessing the second, the following issues were discovered and fixed:

1. Quality issues
    a. The id and id_str columns contain duplicate values. One was dropped to avoid data redundancy.
    b. Incorrect data type: The created_at column has a string datatype (object) instead of the DateTime datatype. It was transformed to the proper datatype using the astype method.
    c. The truncated, favorited, retweeted, possibly_sensitive, and possibly_sensitive_appealable columns have string datatype instead of Boolean datatype. It was changed to the appropriate datatype using the astype method.
    d. There were lot of missing values in the coordinates, geo, contributors, place, quoted_status, quoted_status_id_str, quoted_status_id, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, retweeted_status, extended_entities, possibly_sensitive_appealable and possibly_sensitive columns, which resulted in them getting dropped.
    e. Incorrect datatype: The retweet_count and favorite_count columns have a string datatype rather than an integer datatype. It was changed to the correct datatype using the astype method.

    f. Every element in the truncated, retweeted columns has the value False. These Columns were dropped because they were too monotonous to offer any insight.

2. Tidiness issues.
    a. Multiple variables in the full_text column were separated into three columns.
    b. Multiple variables in the display_text_range column.
    c. Multiple variables in the entities, extended_entities, user, place, retweeted_status and quoted_status columns are contained in dictionaries and were used to form new tables and then dropped.
    d. Multiple variables in the source column were split into two distinct columns.

Upon the assessment of the third dataset, the following things were learned and fixed:

1. The values in the p1, p2, and p3 columns have inconsistent casing and were standardized.
2. The jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog columns were renamed for clarity.

After the cleaning, the datasets were merged and upon reassessment of the merged dataset the following issues were discovered and fixed:

1. The full_text, full_text_rating, full_text_link, tweet_id, created_at, source_text_x, source_link_x, and the text, text_rating, text_link, id, timestamp, source_text_y, source_link_y columns respectively are identical in terms of values, one of these identical columns in each case was dropped.
2. The rating_denominator and rating_numerator columns were extracted from the full_text_rating is the combination of the rating_denominator and rating_numerator columns. The rating_denominator and rating_numerator columns were dropped to avoid data redundancy.
3. The source_text_x and the source_link_x were renamed for clarity.