Data Wrangling Report

The primary goals of the project were to:

- Conduct data wrangling, including the collection, evaluation, and cleaning of the provided data sources.
- Store, analyze, and visualize the cleaned data.
- Produce reports on:
    1. The data wrangling process.
    2. The analysis and visualizations of the data.

Step 1: Data Collection

During this phase, the following datasets were collected and represented as pandas dataframes:

- **WeRateDogs Twitter Archive**: This file, named twitterarchiveenhanced.csv, was manually downloaded.
- **Tweet Image Predictions**: The image-predictions.tsv file was programmatically downloaded using the Requests library from a specified URL.

- **Tweet JSON Data**: Each tweet's JSON data, containing at least the tweet ID, retweet count, and favorite count, was stored in a file named tweet_json.txt. This data was retrieved using the Twitter API and Python's Tweepy library, with each tweet's JSON data written on a separate line.

Steps 2 and 3: Data Assessment and Cleaning

During the data assessment phase, several observations were made. The table below outlines these observations along with the corresponding actions taken during the cleaning process.

Quality

For the df_arch dataset:

- **Timestamp Format**: The timestamp was initially a string and was converted to the datetime data type using the pandas to_datetime function.
- **Retweets**: Rows containing retweets were removed, as only original tweets were of interest.
- **Replies**: Rows containing replies to original tweets were removed, focusing solely on original tweets.
- **Rating Numerator**: The rating_numerator column was corrected and converted to float. Rating scores were accurately extracted.
- **Rating Denominator**: Rows with a denominator greater than 10 were removed, as these likely indicated ratings for multiple dogs.

- **Expanded URLs**: Rows with missing URLs in the expanded_urls column were removed as they were invalid data.
- **Names**: None and invalid names in the name column were replaced with np.nan.
- **Doggo Columns**: The columns doggo, floofer, pupper, and puppo had None for missing values, which were replaced with np.nan.
- **Text Column**: The text column included tweet links and ratings at the end. These were removed using regular expressions (RegEx) and the pandas extract method.

.

Tidiness

In the df_arch dataset, the columns doggo, floofer, pupper, and puppo all represented variations of dog personalities. To streamline the data, these were merged into a single column named dog_stage, and the original four columns were removed.

In the df_pred dataset, the img_num column was unnecessary and removed. Only the id, retweet_count, and favorite_count columns were retained, with all other columns removed.

Result

A combined data set with all needed information was stored in a sqlite data base.