Wrangle Report - WeRateDogs

This report outlines the data gathering, assessment, and cleaning process for the WeRateDogs dataset.

1. Data Gathering

Three datasets were used in this project:

- Twitter Archive (twitter-archive-enhance.csv): Provided by Udacity, downloaded and loaded into a dataframe named archive.
- Image Predictions (imagepredictions.tsv): Retrieved via a URL provided by Udacity and loaded into imgPredict.
- Tweet Data (JSON Format): Due to Twitter API limitations, data was extracted from a pre-provided JSON file, downloaded and stored as jsonlist.

2. Data Assessment

- Each dataset was assessed both visually and programmatically to identify inconsistencies and messy data.
- Key Issues Identified:
    - Inconsistent data types (e.g., timestamp formats, numerical columns).
    - Invalid values (e.g., incorrect dog names, abnormal ratings).
    - Irrelevant or redundant columns.

3. Data Cleaning

The cleaning process included the following transformations and modifications:

Data Type Corrections

- Converted timestamp columns to datetime.
- Changed rating numerators to float.
- Updated tweet_id to string format across all datasets.

Data Standardization & Fixes

- Adjusted rating denominators higher than 10 to 10, modifying numerators accordingly.
- Replaced invalid dog names with "No Name".
- Removed tweets and replies that had non-null retweet or reply values.

Column Removals

- Dropped irrelevant columns:
    - in_reply_to_status_id
    - in_reply_to_user_id
    - retweeted_status_id
    - retweeted_status_user_id
    - retweeted_status_timestamp
- Removed HTML tags from the source column.
- Removed expanded URLs due to missing values.

Structural Adjustments

- Merged doggo, floofer, pupper, and puppo columns into a single column.
- Created copies of the cleaned data and merged all three datasets into a final dataframe named df_main.
- Saved the final cleaned dataset as twitter_archive_master.csv.

Final Output

After cleaning, the processed dataset is stored in twitter_archive_master.csv, ready for analysis.