Wrangleing Report

This report summarizes data wrangling efforts done to enable the analysis of WeRateDogs Twitter data. It involved data collection from multiple sources, assessment of quality and tidiness, and cleaning for analysis.

Data Gathering

Three datasets were retrieved:

Twitter archive: This was provided as a CSV file.

Image Predictions: Downloaded from URL using the requests library

Extra tweet data (favorites count,retweet count): Gathered using the Twitter API.

Certainly, I'll rewrite this in a more comprehensive way with additional details in Markdown format:

Data Quality and Tidiness Assessment

Quality Issues

1. Accuracy

Twitter Archive Table

- **Name column**: Contains incorrect or placeholder names (e.g., "a").
- **Rating columns**:
  - `rating_numerator` and `rating_denominator` contain unrealistic values, including zeros.

2. Completeness

Twitter Archive Table

- Nine columns have more than 2,000 null values each.
  - Specific columns should be identified.
  - The total number of rows in the dataset should be provided for context.
  - The percentage of missing data for each affected column should be calculated.

Extra Archive Table

- 13 columns have over 90% null values.
- Three columns are entirely composed of NaN values (100% missing).
  - These columns should be listed explicitly.

3. Consistency

Twitter Archive Table

- The `expanded_urls` column contains duplicate values.
  - The frequency and nature of these duplicates should be analyzed.

Image Predictions Table

- `jpg_url` column inconsistencies:
  - Two images are PNG files instead of JPG.
  - Some URLs are duplicated, potentially causing confusion.

4. Validity

Twitter Archive Table

- `timestamp` column is not in the correct datetime format.

Extra Archive Table

- `id_str` is stored as an integer instead of a string.

Tidiness Issues

1. Twitter Archive Table

- Dog "stage" columns (`doggo`, `floofer`, `pupper`, `puppo`) should be consolidated into a single column.

2. Extra Archive Table

- Column names need to be renamed to better indicate the nature of their values.

Certainly! I'll rewrite the cleaning steps as if you're speaking, outlining your plan to address the issues:

Data Cleaning

1. Twitter Archive Table

Accuracy Issues

1. I'm going to start by fixing that `name` column. I'll scan through it and replace any obviously incorrect names like "a" with NaN. If there's a pattern to these errors, I might be able to automate this.

2. For the `rating_numerator` and `rating_denominator` columns, I'll:

   - Remove any ratings with zeros
   - Cap the ratings at a reasonable maximum (probably 15/10, since we're dealing with good dogs here)

    - If there are any negative ratings, I'll investigate those individually

Completeness Issues

1. I've got 9 columns with over 2000 null values. I need to:
    - List out these columns
    - Calculate the percentage of nulls for each
    - Decide whether to drop columns that are mostly empty or if I can fill in some of the missing data

Consistency Issues

1. For the `expanded_urls` column:
    - I'll check for exact duplicates and remove them
    - If there are near-duplicates (like the same URL with different parameters), I'll standardize them

Validity Issues

1. I'm going to convert that `timestamp` column to a proper datetime format. I'll use pandas for this - something like `pd.to_datetime()` should do the trick.

Tidiness Issues

1. Those dog "stage" columns (`doggo`, `floofer`, `pupper`, `puppo`) need to be combined:
    - I'll create a new `dog_stage` column
    - Then I'll fill it with the value from whichever of the four columns isn't null
    - If multiple stages are present, I'll concatenate them with a separator
    - Finally, I'll drop the original four columns

2. Extra Archive Table

Completeness Issues

1. I've got 13 columns that are more than 90% nulls, and 3 that are completely empty:
    - I'll list out all these problematic columns
    - For the completely empty ones, I'm just going to drop them
    - For the others, I'll decide case-by-case if they're worth keeping

Validity Issues

1. The `id_str` column needs to be a string, not an int:
    - I'll convert it using `df['id_str'] = df['id_str'].astype(str)`
    - Then I'll check to make sure no data was lost in the conversion

Tidiness Issues

1. Time to rename some columns:
    - I'll make a list of all the ambiguous column names
    - Then I'll come up with more descriptive names for each
    - Finally, I'll use `df.rename(columns={...})` to apply the new names

3. Image Predictions Table

Consistency Issues

1. For the `jpg_url` column:
    - I'll identify those two PNG files and decide whether to keep them or not
    - I'll find and list out all the duplicate URLs
    - Depending on how many dupes there are, I might keep them and add a flag, or I might remove them

4. Final Steps

Storing

The 3 cleaned data sets were combined into one master dataset which was stored as 'twitter_archive_master.csv'.