Wrangle Report

Introduction

This project focuses on wrangling WeRateDogs Twitter data to create compelling and trustworthy analyses and visualizations. The initial Twitter archive offers a solid foundation but lacks detailed information, necessitating further gathering, assessing, and cleaning.

The primary tasks of this project are:

- **Data Wrangling**, which encompasses:
  - Gathering data
  - Assessing data
  - Cleaning data
  - Storing, analyzing, and visualizing the cleaned data

Gathering Data

Data was collected from three sources:

1. **Twitter Archive**: A file named "twitter_archive_enhanced.csv" containing basic tweet information.
2. **Tweet Image Predictions**: Predictions of dog breeds or other objects present in tweets using a neural network, stored in "image_predictions.tsv" and downloaded programmatically.
3. **Tweet JSON Data**: Data downloaded from Udacity.

Each dataset was imported into separate pandas DataFrames for initial assessment.

Assessing Data

The data from the three sources were visually and programmatically assessed to identify quality and tidiness issues.

Quality Issues

1. **"in_reply_to_status_id" column**: Contains 2278 null/NaN values.
2. **"in_reply_to_user_id" column**: Contains 2278 null/NaN values.
3. **'timestamp' column**: Data type is object instead of datetime.
4. **Replies and Retweets**: Data includes unnecessary replies and retweets.
5. **Duplicate URLs**: jpg_url column contains 66 duplicates.
6. **Ambiguous Column Names**: Column names are unclear and provide limited information.
7. **Incorrect Ratings**: Some ratings are inaccurately extracted.
8. **Missing Data**: Some records have missing values in critical fields.

Tidiness Issues

1. **Unnecessary Columns**: Drop columns that do not contribute to the analysis.
2. **Dataset Merging**: Merge datasets for a cohesive master dataset.

Cleaning Data

The data cleaning process involved defining, coding, and testing the cleaning steps, which included:

- **Removing Retweets and Replies**: Filtering out unnecessary retweets and replies.

- **Dropping Unnecessary Columns**: Removing columns not needed for the analysis.
- **Correcting Data Types**: Converting the 'timestamp' column to datetime.
- **Addressing Duplicates**: Removing duplicate URLs in the jpg_url column.
- **Merging Datasets**: Combining the datasets into a single master DataFrame.
- **Handling Missing Values**: Addressing null values appropriately.
- **Renaming Columns**: Renaming columns to be more descriptive and informative.

Conclusion

The cleaned data was stored in a CSV file for further analysis and visualization. This report outlines the data wrangling process, highlighting the steps taken to ensure the data's quality and tidiness. The foundation set by this process enables reliable and insightful analyses and visualizations.

By addressing these issues, we ensured the dataset's integrity, making it suitable for generating meaningful insights and visualizations. The final cleaned dataset provides a robust basis for subsequent analysis, leading to compelling and accurate results.