Wrangle Report

Introduction

The purpose of this document is to summarize the data wrangling efforts for the Twitter archive dataset. This dataset required gathering data from various sources, assessing its quality and tidiness, and then cleaning the identified issues to create a high-quality, tidy master dataset.

Data Gathering

Three main data sources were gathered for this project:

1. `twitter-archive-enhanced.csv`: Contains basic tweet data.
2. `image-predictions.tsv`: Provides predictions about the breed of dogs based on images.
3. `tweet-json.txt`: Additional tweet data obtained via Twitter API.

Data Assessment

Quality Issues Identified:

1. **Missing and Sparse Values**: Identified columns with significant null values, handled by dropping irrelevant columns and filling missing data where appropriate.
2. **Redundant Columns**: Removed duplicate or unnecessary columns to streamline the dataset.
3. **Columns with List Values**: Split list values into separate columns to improve readability and usability.
4. **Column Name Consistency**: Renamed columns to ensure consistency and clarity throughout the dataset.
5. **Timestamp Format Issues**: Standardized timestamp formats to facilitate time-based analysis.
6. **Text Formatting Issues in `source` Column**: Extracted relevant information from HTML entities to improve data clarity.
7. **Duplicate Entries**: Removed duplicate rows to maintain data integrity.
8. **Data Type Inconsistencies**: Converted data types to their appropriate formats for better analysis and storage efficiency.

Tidiness Issues Identified:

1. **Consolidate Dog Stage Columns**: Combined multiple columns related to dog stage into a single column to reduce redundancy and improve dataset structure.
2. **Merge Dog Breed Prediction Columns**: Integrated separate columns related to dog breed predictions into a more cohesive and understandable format.

Data Cleaning Techniques

Removing Retweets

Before the initial cleaning process, it was necessary to remove retweets. Retweets were identified by non-null values in the `retweeted_status_id`, `retweeted_status_user_id`, and `retweeted_status_timestamp` columns. Removing retweets is important because we only want to analyze original tweets to ensure the integrity and accuracy of our analysis. Retweets might duplicate the same tweet multiple times, skewing the results and insights derived from the data.

Quality Issue Handling Techniques:

1. **Handling Missing Values**: Addressed missing values by either dropping irrelevant columns or filling them with appropriate default values.
2. **Redundant Columns**: Removed redundant columns that did not contribute to the analysis or were duplicates.
3. **Handling Columns with List Values**: Split list values into separate columns to provide a more granular view of the data.
4. **Standardizing Column Names**: Renamed columns to ensure consistency and clarity in data interpretation.
5. **Timestamp Format Standardization**: Converted timestamps into a consistent format for ease of time-based analysis.
6. **Text Formatting in `source` Column**: Extracted pertinent information from HTML entities in the `source` column to improve data readability.
7. **Duplicate Entries**: Removed duplicate rows to maintain data integrity and accuracy.
8. **Data Type Standardization**: Ensured data types were correctly assigned to each column for efficient storage and analysis.

Tidiness Issue Handling Techniques:

1. **Consolidating Dog Stage Columns**: Merged multiple columns related to dog stages into a single column to simplify dataset structure and improve analysis efficiency.
2. **Merging Dog Breed Prediction Columns**: Combined separate columns related to dog breed predictions into a cohesive format to enhance data coherence and interpretation.

Conclusion

The cleaning process addressed the identified quality and tidiness issues in the Twitter archive dataset, resulting in a refined dataset (`twitter_archive_master.csv`) ready for further analysis and visualization. By systematically addressing each issue using appropriate techniques, the cleaned dataset now provides a reliable foundation for insightful data exploration and modeling.