Introduction

The data wrangling for the WeRateDogs Twitter dataset aimed to prepare the data for analysis by cleaning, transforming, and merging various data sources. The process ensured data accuracy, completeness, and proper structure for insightful analysis.

Data Sources

1. Twitter Archive Data: Contains historical tweets, including tweet IDs, text, and metadata from the WeRateDogs account**.**

2. Image Predictions Data: Includes neural network predictions on tweet images, detailing predicted dog breeds and confidence levels.

3. Additional Data: Engagement metrics like retweet_count and favorite_count were missing initially and required additional processing.

Data Cleaning Steps

1. Handling Missing Values

- Dog Stage: Missing values in the dog_stage column were imputed or rows with significant gaps were removed.

- Retweet Count: The retweet_count column, initially absent, was added by reprocessing the original dataset.

2. Resolving Column Issues

- Missing Columns: Reprocessed the original data to include all necessary columns.

- Column Consistency: Renamed columns and ensured alignment with expected formats.

3. Data Type Conversions

- Numeric Conversion: Converted retweet_count and favorite_count to numeric types, handling non-numeric values.

- Date Conversion: Converted date columns to datetime format for temporal analysis.

4. Merging DataFrames

- Merging Data: Combined twitter_archive_clean with image_predictions on tweet_id, creating a unified dataset.

- Handling Merge Issues: Resolved discrepancies such as mismatched tweet IDs and missing values.

Challenges

- Missing Retweet Count: The absence of this column required reprocessing to include all necessary data.

- Data Consistency: Merging datasets with differing formats and incomplete information posed challenges.

Insights and Analysis

1. Tweet Engagement

- Most Popular Tweets: Identified top 10 tweets by retweet count to gauge engagement.

- Average Engagement by Dog Stage: Calculated average engagement metrics for different dog stages.

- Engagement Trends Over Time: Analyzed monthly engagement metrics to observe trends.

2. Rating Scores Insights

- Distribution of Rating Scores: Analyzed and visualized the frequency of rating scores**.**

- Average Rating Score by Dog Stage: Compared average ratings across dog stages.

Conclusion

The data wrangling process effectively prepared the dataset for analysis by addressing missing values, resolving column issues, and ensuring data consistency. The cleaned dataset supports meaningful analysis of tweet engagement and rating scores.