Data Wrangling Report

Introduction

This report outlines the comprehensive data wrangling efforts undertaken to prepare a dataset for analysis. The process involved gathering data from multiple sources, assessing it for various quality and tidiness issues, and meticulously cleaning the data to ensure it was suitable for insightful analysis. The final clean dataset served as the foundation for deriving meaningful insights and creating visualizations.

Gathering

Data was collected from three primary sources:

- **The Twitter Archive dataset** provided tweet-level data, including dog ratings, tweet timestamps, and text content. This dataset formed the core of the analysis, offering insights into user interactions and the popularity of different dog stages.
- **The Image Predictions dataset** was sourced from a neural network that predicted dog breeds from images. This dataset added a layer of analysis, allowing for the exploration of breed-specific trends and their correlation with engagement metrics.
- Additionally, **a JSON file** provided extra data, including engagement metrics like retweet and favorite counts, which were essential for analyzing the popularity of different tweets over time.

Assessing

The assessment phase involved a thorough evaluation of the datasets for both quality and tidiness issues. This included:

- **Quality Assessment**: Identifying missing data, inconsistencies, and inaccuracies. For instance, it was found that some dog names were incorrect, and some tweets had missing or erroneous data in the dog stage column.

- **Tidiness Assessment**: Checking the structure of the datasets to ensure that each variable was in its own column, each observation in its own row, and that the datasets could be easily merged or joined for analysis. This step revealed that the dog stage information was spread across multiple columns, which required consolidation into a single column for clarity and ease of analysis.

Cleaning

The cleaning process was methodical and addressed all identified issues to prepare the data for analysis. Key actions included:

- **Correcting Data Types**: Conversion of columns to appropriate data types, such as ensuring that timestamp data was properly formatted as datetime objects.

- **Handling Missing and Duplicate Data**: Missing data was either filled using appropriate imputation methods or removed if it was deemed irrelevant. Duplicates were identified and removed to prevent skewed analysis results.

- **Merging Datasets**: The cleaned datasets were merged to create a master dataset, which facilitated comprehensive analysis across different variables, including tweet content, dog stages, and image predictions.

Conclusion

The data wrangling process was essential in transforming raw and unstructured data into a clean, analyzable format. This process enabled the extraction of valuable insights, such as the identification of patterns in user engagement across different dog stages and the determination of the most frequently mentioned dog breeds. The resulting clean dataset provided a solid foundation for generating accurate and meaningful visualizations, ultimately supporting data-driven decision-making.