wrangle_report

October 1, 2023

0.1 Reporting: wragle_report

Title: Wrangling Report - Internal Document

Introduction: This report outlines the data wrangling efforts undertaken to clean and prepare the WeRateDogs Twitter archive data for analysis. The goal of this project was to assess, clean, and transform the data into a tidy and usable format, ensuring its quality and consistency. The data was collected through various sources, including Twitter's API, and comprises tweets and additional information related to dogs.

Data Gathering: The data collection process involved downloading the WeRateDogs Twitter archive data, including tweet IDs, timestamps, text content, and various dog-related columns such as dog stages and names. Additionally, we queried the Twitter API using the Tweepy library to gather retweet count, favorite count, and JSON data for each tweet. Furthermore, we downloaded the tweet image prediction data from a provided link, which contained predictions for dog breeds present in the tweeted images.

Data Assessment: Upon visual assessment, we identified several quality and tidiness issues. The 'archive' dataframe contained retweets, which we decided to exclude as we were interested only in original content. Additionally, the 'timestamp' column was in string format and needed conversion to datetime. The 'dog stages' columns were spread across separate columns, which should be combined into one categorical column. The 'name' column also had incorrect values like 'a' for names, suggesting data entry errors. In the 'predictions' dataframe, the prediction data was spread across multiple columns and needed to be organized.

Data Cleaning: Data cleaning was performed systematically to address the issues discovered during the assessment. We removed retweets by dropping rows with non-null values in the 'retweeted_status_id' column. We converted the 'timestamp' column to the datetime data type. For the 'dog stages', we combined the separate columns into one, using a categorical data type for better representation. Incorrect 'name' values were corrected, and 'None' values were replaced with NaNs for consistency.

For the 'predictions' dataframe, we melted the prediction columns ('p1', 'p2', 'p3') into a single column to specify the prediction number. This allowed us to create additional columns for the actual prediction, confidence, and whether it is a dog breed. We also standardized the capitalization of prediction labels.

Data Transformation: We merged the cleaned 'archive' dataframe with the gathered 'tweet data' and 'predictions' dataframes, based on their common tweet IDs. This process combined the information from various sources into a comprehensive dataset. We then filtered the data to retain only tweets with valid images ('expanded_urls' not null) to meet our analysis requirements.

Conclusion: The data wrangling efforts have resulted in a clean, tidy, and comprehensive dataset, ready for further analysis. The data is now devoid of retweets, and the timestamp and dog stage columns have appropriate data types. The prediction data is now organized in a structured format. The dataset can be effectively used for exploratory data analysis, insights extraction, and predictive modeling.

This internal document serves as a record of the data wrangling process undertaken, ensuring transparency, reproducibility, and a strong foundation for the subsequent analytical stages. The clean dataset will enable us to make data-driven decisions and uncover valuable insights about WeRateDogs tweets and their associated content.