wrangle-report

July 13, 2024

1 Wrangle Report

2 Title: Wrangling Twitter Data

3 Introduction

This document outlines the steps and processes involved in the wrangling of three datasets related to tweets from the Twitter platform. The goal of the wrangling effort was to clean, merge, and prepare the data for subsequent analysis and visualization.

4 Datasets

Twitter Archive Enhanced: Contains metadata for tweets, including tweet ID, timestamp, text, and ratings. Image Predictions: Contains predictions about the image content in tweets, including confidence scores and whether the image contains a dog. Tweet JSON: Contains detailed information about each tweet, including retweet count and favorite count. Accessing Data

Loading Data: The datasets were loaded using pandas. The Twitter Archive Enhanced data was loaded from a CSV file, the Image Predictions data from a TSV file, and the Tweet JSON data from a JSON file. Initial Inspection: Basic inspection was conducted using head(), info(), describe(), and isna().sum() methods to understand the structure, data types, and missing values. Cleaning Data

5 Quality Issues:

Columns with excessive missing values were identified and removed. Data types were standardized (e.g., converting IDs to strings and timestamps to datetime objects). Boolean columns were converted to proper boolean data types. Tidiness Issues:

The datasets were merged into a single DataFrame to facilitate analysis. The merge was performed on the tweet_id column. Unnecessary columns were dropped to reduce data redundancy and improve clarity. Storing Data The cleaned and merged dataset was stored as a new CSV file named twitter_archive_master.csv.

6 Conclusion

The wrangling process involved cleaning the data for quality issues, merging multiple datasets into a single cohesive DataFrame, and storing the cleaned data for further analysis. The next step is to analyze and visualize the cleaned data to extract meaningful insights