Title: Data Science Assignment 2
Student Name: Minahil Irfan

Overview

Data cleaning is the process of identifying and correcting inaccurate records from a dataset. It ensures the quality, consistency, and accuracy of data used for analysis and model training.

Key Concepts:

- **Removing Duplicates:**
  Duplicate rows can distort results and bias model outcomes. Using Pandas (df.drop_duplicates()), we can remove repeated entries and keep only unique records.
- **Handling Missing Values:**
  Missing data can occur due to collection errors or incomplete entries.
  - Remove rows/columns with too many missing values (df.dropna()).
  - Fill missing values with the mean, median, or mode (df.fillna()).
- **Treating Outliers:**
  Outliers are extreme values that differ significantly from other observations. They can be detected using boxplots or statistical methods like the IQR (Interquartile Range) rule, and treated by capping, transforming, or removing them.

Importance:
Clean data ensures reliable and valid analysis. Without cleaning, even the most powerful models will produce misleading results.

Task Overview

The goal of this assignment was to perform data cleaning on the selected project dataset. The tasks included removing duplicates, handling missing values, treating outliers, and generating a “before vs after cleaning” comparison report.

Activity Log

Step 1: Import and Extract Data

- Imported required Python libraries: *pandas*, *zipfile*, and *os*.
- Extracted the dataset from a ZIP file and loaded it into a DataFrame using Pandas.
- Added column names as per the Kaggle dataset description.

Step 2: Review Before Cleaning

- Checked dataset shape, missing values, and duplicates.
- Observed initial data issues, including duplicate records and possible outliers in the text column.

Step 3: Data Cleaning Process

- **Removed duplicates** using drop_duplicates().
- **Handled missing values** by removing rows with null data using dropna().
- **Treated outliers** by filtering tweets with text lengths between 3 and 280 characters.

Step 4: Compare Before vs After

- Created a Pandas DataFrame report summarizing differences in dataset shape, missing values, and duplicates before and after cleaning.
- Noted significant reduction in total rows after cleaning, confirming successful removal of duplicates and invalid data.

Step 5: Save and Upload

- Saved the cleaned dataset as twitter_sentiment_cleaned.csv.
- Uploaded the cleaned dataset to the GitHub project repository.

Summary

This activity focused on improving data quality by cleaning the dataset. By removing duplicates, handling missing values, and treating outliers, the dataset became more accurate and consistent for analysis. The before and after report clearly showed the improvement achieved through cleaning.