Title: Data Cleaning Assignment

Dataset: Use the attached dataset

Tasks:

Data Exploration: [2 marks]

a. Load the dataset into your preferred data analysis tool.

b. Perform an initial exploration of the dataset, including checking the dimensions, previewing a few rows, and identifying the data types of each column.

**Answer:**

Loaded the dataset and checked the no of rows and columns and checked the datatypes and printed the first and last 15 rows and also checked the Indexes, datatypes and memory information’s.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Handling Missing Values: [3 marks]

a. Identify columns with missing values.

b. Evaluate the extent of missing values in each column and decide on an appropriate strategy for handling them (e.g., imputation or removal).

c. Implement the chosen strategy and explain your reasoning behind it.

d. Validate the changes made and ensure that no missing values remain.

**Answer:**

Identified the missing values based on the percentage more than 20% and dropped only those column and also dropped the rows where it is null and now the data is clean with no null values.

A screenshot of a computer

Description automatically generated

Removing Duplicates: [3 marks]

a. Check for duplicate rows in the dataset.

b. If duplicates exist, remove them, and justify your approach.

c. Verify that duplicate rows have been successfully removed.

**Answer:**

Identified that there were 11 duplicate rows and therefore, removed those duplicate rows and validated that there are no more duplicate rows.



Dealing with Inconsistent Data: [3 marks]

a. Identify columns that contain inconsistent or erroneous data.

b. Implement necessary corrections to resolve inconsistencies (e.g., standardizing formats, correcting typos).

c. Document the changes made and explain how they improve the data quality.

**Answer:**

Identified few special characters in Genre field and replaced the special characters with empty string and So that when the data is summarized and used for visualization, it will be clean for the business tool to understand better and generate the corresponding reports.

A screenshot of a computer

Description automatically generated

Handling Outliers: [2 marks]

a. Identify columns that may contain outliers.

b. Analyze the outliers and decide on an appropriate strategy for handling them (e.g., removal or transformation).

c. Implement the chosen strategy and validate the changes made.

**Answer:**

Identified if there are Outliers in Rating field by checking if there are any values greater than 10 and there are no Outliers in Rating field therefore no action is required and I am proceeding further with my analysis.



Data Validation: [2 marks]

a. Perform any additional checks or validations that you think are necessary to ensure the overall quality of the dataset.

b. Document the validation steps taken and their outcomes.

**Answer:**

Performed one more additional checks to ensure the overall quality of the dataset

A screenshot of a computer

Description automatically generated

Observations: Share any interesting findings or patterns you discovered during the data cleaning process.

* One Pattern which I could discover was though the ratings were high above 8 rating when it is compared to few fields of vote it is less. So I feel rating field is negatively correlated with vote field in the dataset.

Conclusion: Summarize the importance of data cleaning and the impact it has on data analysis.

* Data Cleaning is one of the mandatory and first step towards data analysis and with the raw data having null values, duplicate values and Outliers doesn’t allow us to give any concrete findings or conclusions.

Submission Guidelines:

Submit your code and any relevant visualizations in the report or provide a separate file with the code.

Make sure to provide clear explanations and justifications for your data cleaning decisions.

Note: Use Python to complete this assignment.

Submitted by : Vidhya Venugopal (8902970)