

SOLUTION APPROACH

Objective:

- Develop a system that reads CSV data, performs statistical analysis, visualizes the data, and uses GPT or LLaMA models to generate text-based insights.

Key Components:

- **Data Handling:** Efficiently reading and processing CSV files with multiple encoding formats.
- **Statistical Analysis:** Extracting meaningful statistics from the dataset.
- **Visualization:** Creating plots to represent data trends and correlations.
- **Text Generation:** Using GPT or LLaMA models to answer questions based on data context.

Challenges with CSV Files:

- CSV files often come with different encodings, leading to potential reading errors.
- Encodings like UTF-8, Latin1, and CP1252 are common but can cause UnicodeDecodeError if mismatched.

Solution:

- The read_csv function iterates over a list of potential encodings to read the file correctly.
- **Example:**
 - The function tries UTF-8, and if it fails, it moves to Latin1, and so on until the file is successfully read or all encodings fail.

Purpose of Statistical Analysis:

- Provides a quick overview of the dataset, highlighting key metrics like mean, median, mode, and standard deviation.
- Helps in identifying trends, outliers, and relationships between variables.

Function Breakdown:

- **Mean:** Average value, useful for understanding central tendency.
- **Median:** The middle value, which is less sensitive to outliers than the mean.
- **Mode:** The most frequently occurring value, important in categorical data analysis.
- **Standard Deviation:** Measures data spread, indicating variability.
- **Correlation:** Shows the relationship between variables, important for predictive modeling.