Querying CSVs and Plot Graphs with LLMs

SUBMITTED BY: AMITHESH Y

1. Overview

The CSV Analyzer and Question Answering Chatbot is a web-based application designed to perform statistical analysis on CSV files, generate dynamic insights through visualizations, and answer user questions about the data using natural language processing. The application is built using Python and leverages several key libraries and APIs to provide a comprehensive data analysis solution.

2. Key Components

2.1 Data Handling and Analysis:

- Pandas: Used for reading, manipulating, and analyzing CSV data.
- NumPy: Employed for numerical computations and statistical calculations.

2.2 Visualization:

- Matplotlib and Plotly Express: Utilized to create interactive and dynamic data visualizations.

2.3 Natural Language Processing:

- Groq API: Integrated to power the question-answering chatbot functionality using (Mixtral-8x7b-32768) LLM model.

2.4 User Interface:

- Streamlit: Serves as the framework for building the web-based user interface.

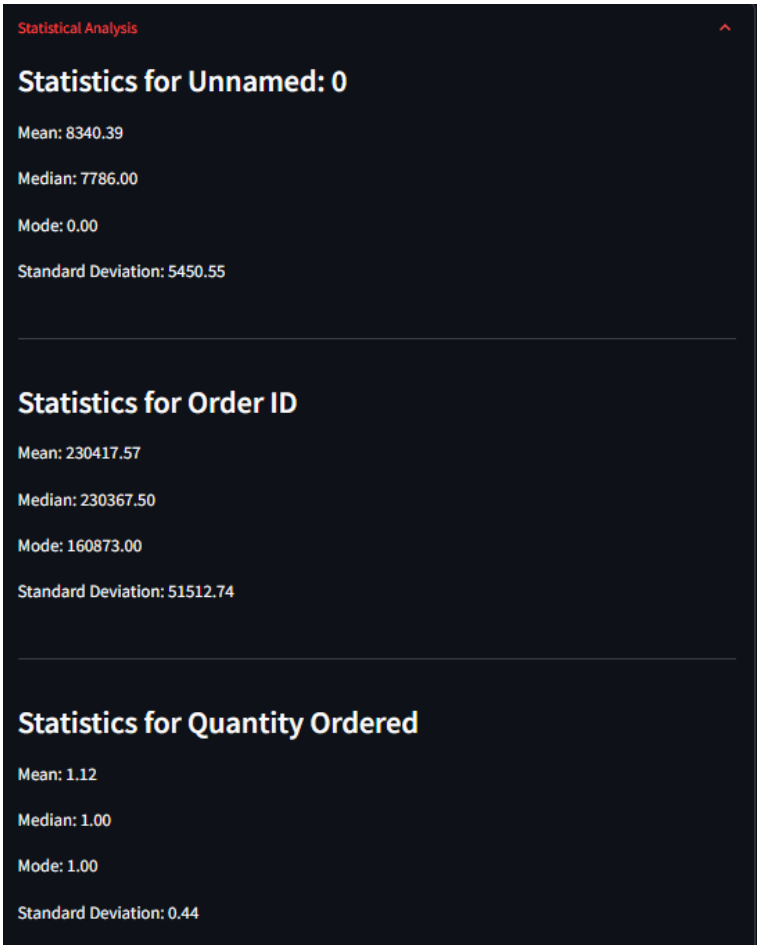
3. Core Functionalities

3.1 CSV File Upload and Parsing:

- Users can upload CSV files through the Streamlit interface.
- Pandas is used to read and parse the CSV data into a DataFrame.

3.2 Statistical Analysis:

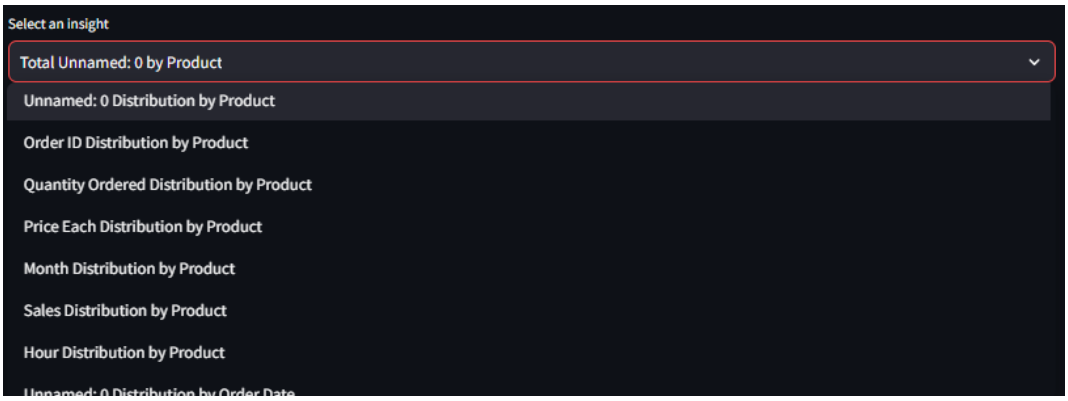
- Basic statistical measures (mean, median, mode, standard deviation) are calculated for numeric columns.
- Results are displayed in an expandable section of the UI.



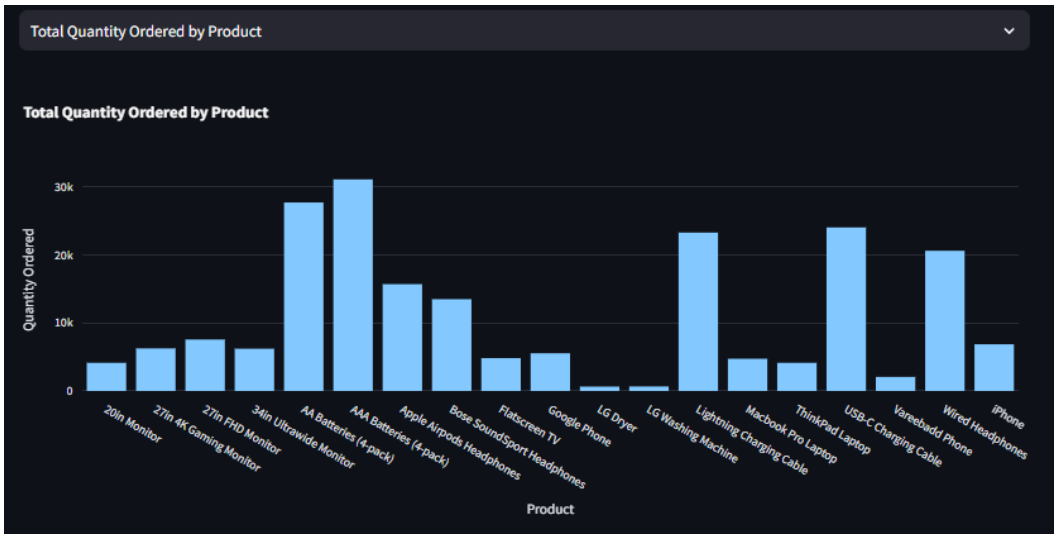
Snippet of statistical analysis in the UI

3.3 Dynamic Insights Generation:

- The application automatically generates various types of plots based on the data structure:
    i. Time series plots for numeric columns if a date column is present.
    ii. Bar charts showing totals of numeric columns grouped by categorical columns.
    iii. Box plots displaying the distribution of numeric columns across categories.
    iv. Histogram plots for every column.
    v. Correlation heatmap for numeric columns.
- Users can select and view different insights through a dropdown menu.



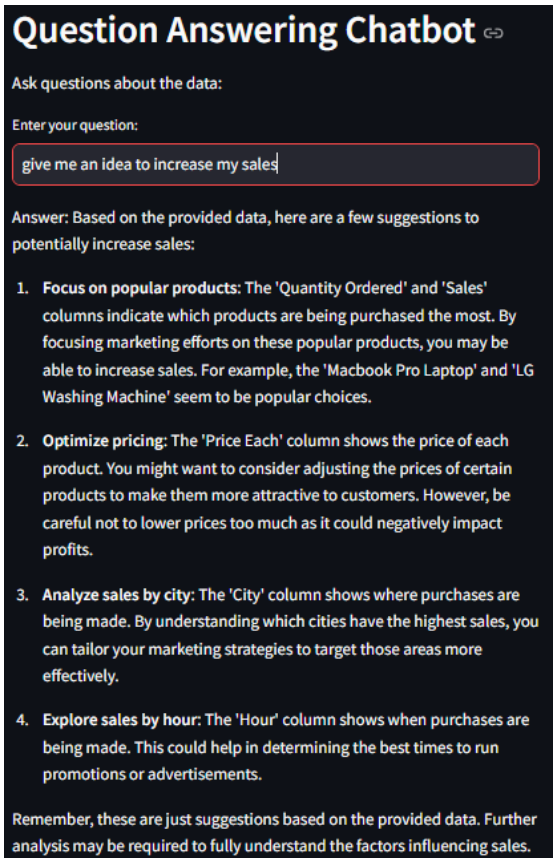
dropdown menu for different plots from the dataset



snippet of the selected graph from the dropdown menu

3.4 Question Answering Chatbot:

- Utilizes the Groq API with the Mixtral-8x7b-32768 model for natural language processing.
- Generates responses to user questions about the data using prompt template and context summary.
- Generates plots based on the user's query.
- Maintains a chat history for context and user reference.



snippet of the answer generated by the LLM

4. Implementation Approach

4.1 Data Processing Pipeline:

1. CSV file upload and parsing using Pandas.
2. Automatic detection of column types (numeric, categorical, datetime).
3. Calculation of basic statistics for numeric columns.
4. Generation of dynamic insights based on data structure.

4.2 Visualization Strategy:

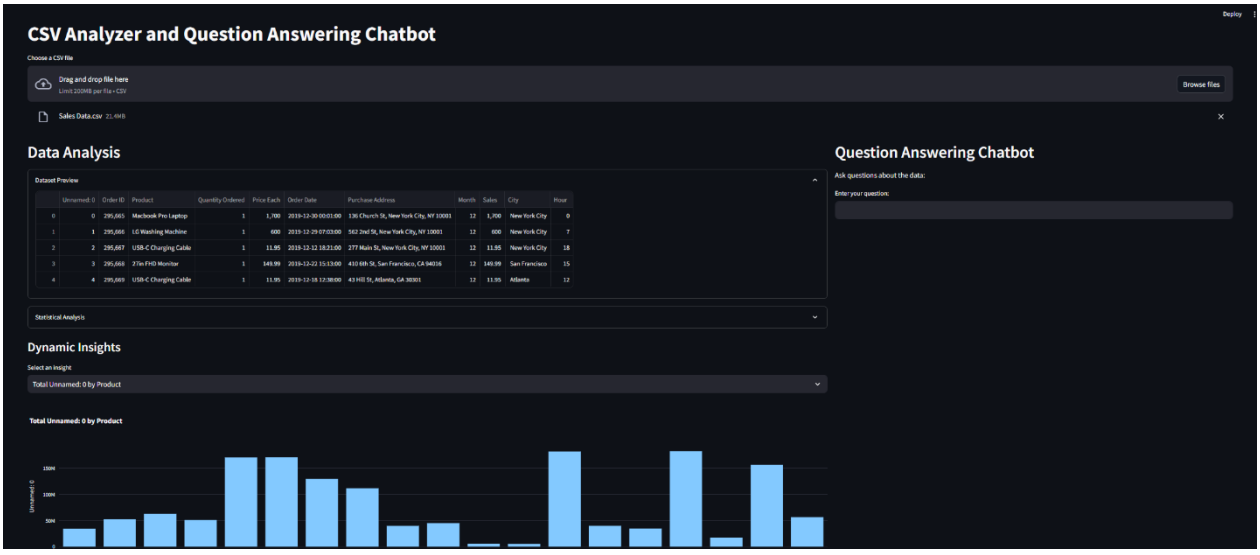
- Employ Plotly Express to create interactive, publication-quality graphs.
- Dynamically generate relevant plot types based on the data characteristics.
- Provide a selection mechanism for users to explore different insights.

4.3 Natural Language Processing Integration:

- Prepare a context-rich prompt including data summary and sample for the LLM.
- Send user questions along with the prepared context to the Groq API.
- Process and display the API's response in the chat interface.

4.4 User Interface Design:

- Implement a two-column layout for efficient space utilization:
  - Left column: Data analysis results and visualizations.
  - Right column: Question-answering chatbot interface.
- Use expandable sections to organize information and reduce clutter.



5. Error Handling and Edge Cases

- Implement try-except blocks to catch and display API errors gracefully.
- Handle various data types and structures in CSV files.
- Provide fallback options for visualizations when certain data types are not present.

6. Scalability and Performance Considerations

- Limit the number of rows displayed in the data preview to enhance performance.
- Generate visualizations on-demand rather than precomputing all possible insights.
- Utilize Streamlit's caching mechanism to improve responsiveness for repetitive operations.

7. Future Enhancements

- Implement more advanced statistical analyses and machine learning capabilities.
- Expand the range of visualization options and customization features.
- Enhance the NLP model's context understanding for more accurate and detailed responses.
- Add support for multiple file formats beyond CSV.

8. Conclusion

The CSV Analyzer and Question Answering Chatbot provides a comprehensive solution for data analysis and exploration. By combining powerful data processing capabilities, dynamic visualizations, and natural language interaction, the application offers users an intuitive and insightful way to understand and query their data. The modular design and use of popular libraries ensure that the solution is both robust and extensible for future improvements.