Step-by-Step Guide: AI-Powered Data Analysis Tool using Streamlit

1. Project Setup

1.1 Install Required Dependencies

- Python – 3.11 version – make sure you installed python 3.11 in your system and add it in your path
- Install pycharm or vs code
- Inside the pycharm or vs code create one requirements.txt file with the given libraries

  ```
  streamlit
  pandas==1.5.3
  openai
  pandasai
  matplotlib
  seaborn
  speechrecognition
  python-dotenv
  openpyxl
  numpy==1.25.2
  pyaudio
  ```

1.2 Create a .env File

The .env file will securely store your OpenAI API key. Create this file in your project root directory and add:

```
OPENAI_API_KEY=”your_openai_api_key”
```

1.3 Create virtual environment and activate it

```
python -m venv myenv
```

```
venv\Scripts\activate
```

1.4 Installation of required library

We have mentioned all required libraries in requirements.txt now we have to install it with the given command

```
pip install -r requirements.txt
```

1.5 Create app.py

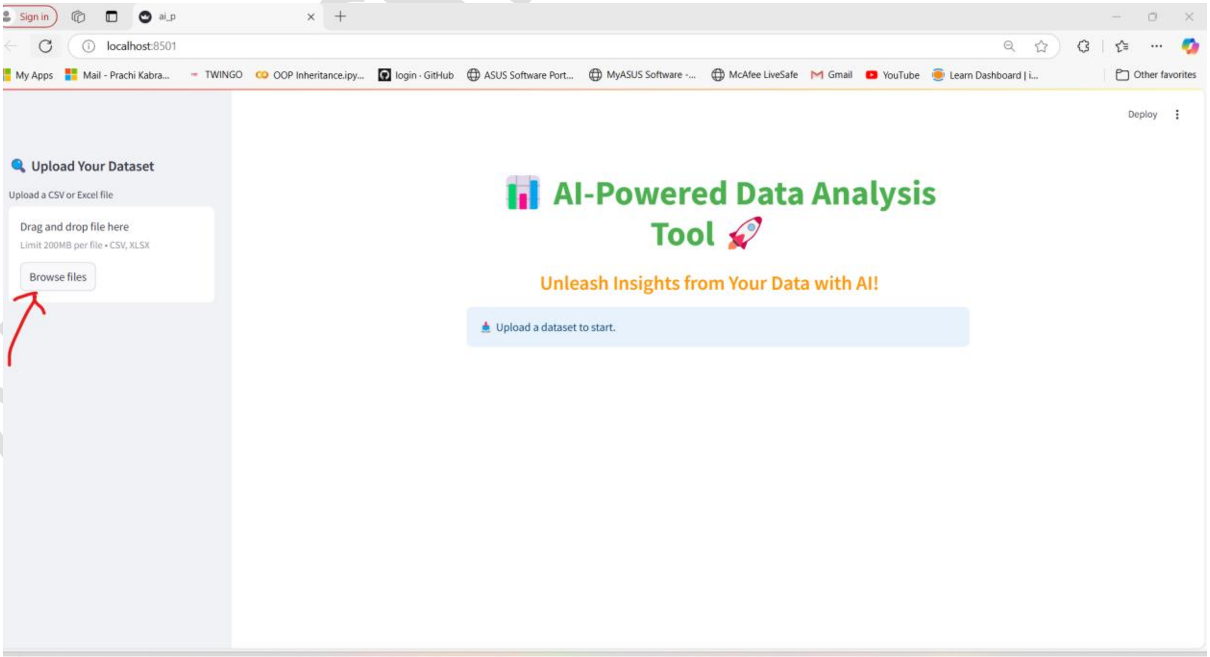
Inside app.py write your all logic to build the project

2. Technologies Used and Why

1. **Streamlit** - Used to create an interactive web-based UI for data analysis, making the tool accessible through a browser.
2. **Pandas** - Provides powerful data manipulation and analysis capabilities, allowing operations like filtering and aggregation.
3. **PandasAI** - An AI-powered wrapper around Pandas that integrates OpenAI to provide intelligent insights and suggestions.
4. **OpenAI API** - Used for AI-driven summarization, querying, and recommendations for handling data.
5. **Matplotlib & Seaborn** - Enable data visualization with various charting capabilities.
6. **SpeechRecognition** - Allows voice-based input for querying data, making the tool more interactive.
7. **Python-Dotenv** - Manages environment variables securely, ensuring API keys are not exposed in the code.
8. **Openpyxl** - Supports reading and writing Excel files for broader dataset compatibility.

3. How It Works

1. **File Upload**: Users can upload CSV or Excel datasets, which are then processed into a Pandas DataFrame (data size should be upto 200 mb only).



🔍 Upload Your Dataset

Upload a CSV or Excel file

Drag and drop file here

Limit 200MB per file • CSV, XLSX

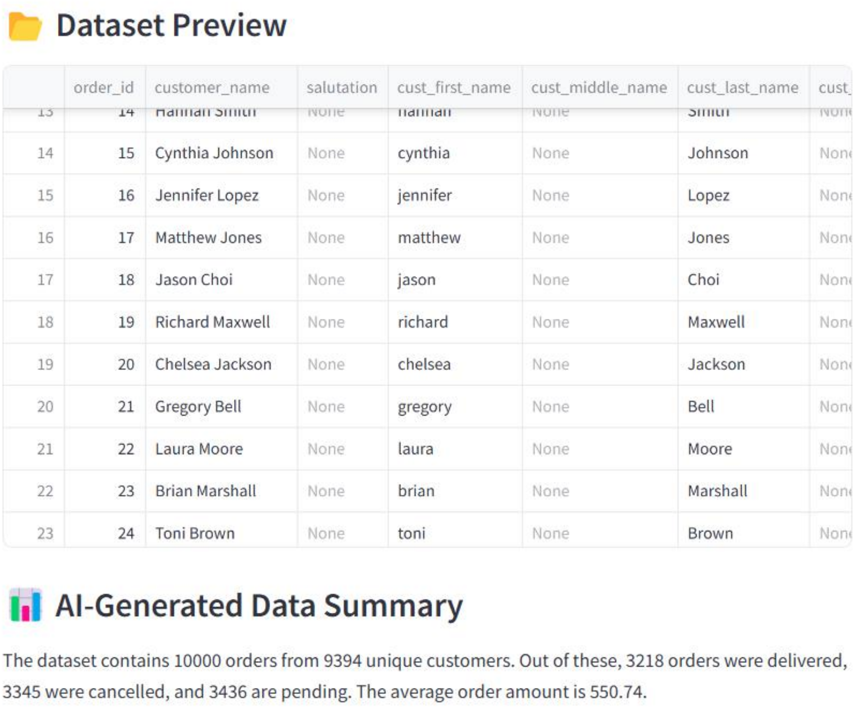
Browse files

📊 AI-Powered Data Analysis Tool 🚀

Unleash Insights from Your Data with AI!

Upload a dataset to start.

2. **SmartDataframe Conversion**: The dataset is wrapped using PandasAI to enable AI-powered queries and insights.



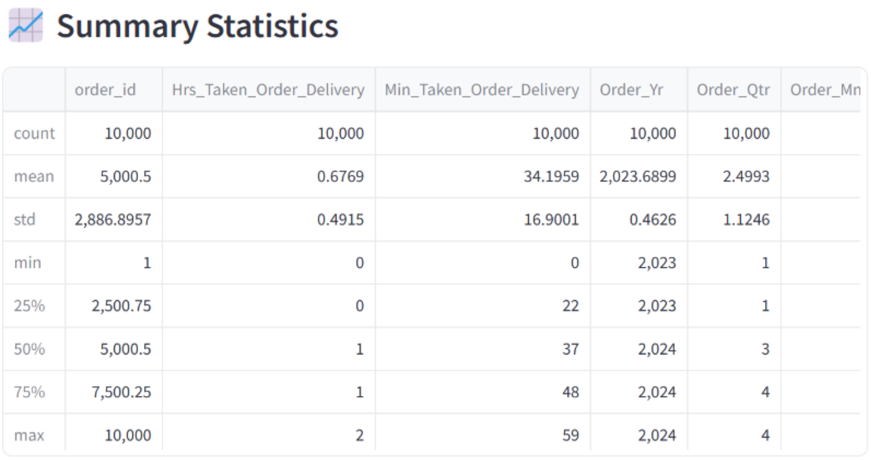
📁 Dataset Preview

	order_id	customer_name	salutation	cust_first_name	cust_middle_name	cust_last_name	cust_
13	14	Hannah Smith	None	hannah	None	Smith	Non
14	15	Cynthia Johnson	None	cynthia	None	Johnson	Non
15	16	Jennifer Lopez	None	jennifer	None	Lopez	Non
16	17	Matthew Jones	None	matthew	None	Jones	Non
17	18	Jason Choi	None	jason	None	Choi	Non
18	19	Richard Maxwell	None	richard	None	Maxwell	Non
19	20	Chelsea Jackson	None	chelsea	None	Jackson	Non
20	21	Gregory Bell	None	gregory	None	Bell	Non
21	22	Laura Moore	None	laura	None	Moore	Non
22	23	Brian Marshall	None	brian	None	Marshall	Non
23	24	Toni Brown	None	toni	None	Brown	Non

📊 AI-Generated Data Summary

The dataset contains 10000 orders from 9394 unique customers. Out of these, 3218 orders were delivered, 3345 were cancelled, and 3436 are pending. The average order amount is 550.74.

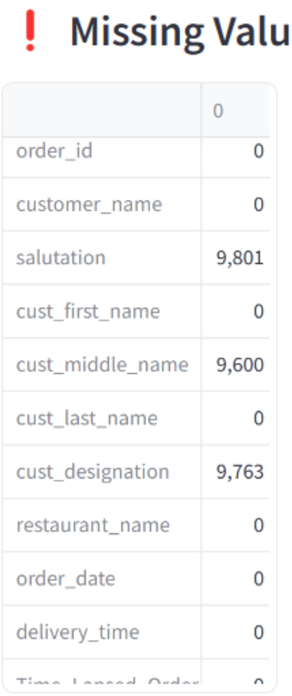
3. **AI-Powered Summarization**: OpenAI provides a concise summary of the dataset’s key aspects.



📈 Summary Statistics

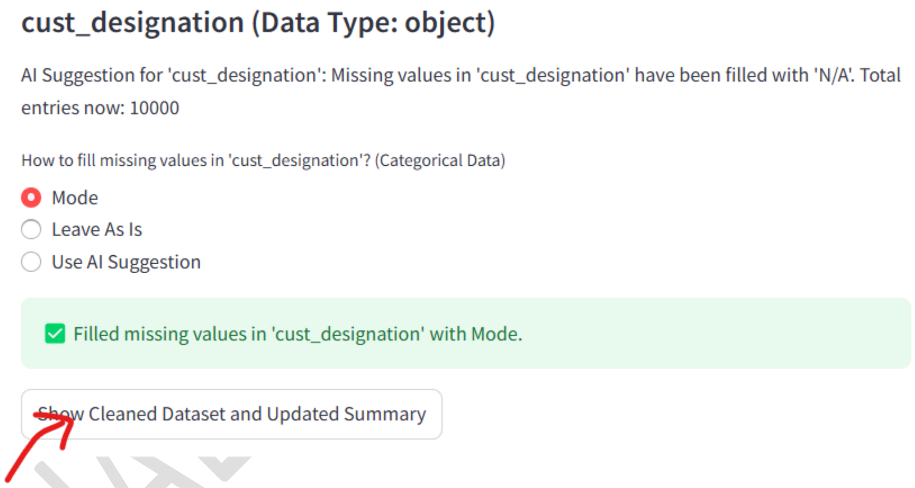
	order_id	Hrs_Taken_Order_Delivery	Min_Taken_Order_Delivery	Order_Yr	Order_Qtr	Order_Mn
count	10,000	10,000	10,000	10,000	10,000	
mean	5,000.5	0.6769	34.1959	2,023.6899	2.4993	
std	2,886.8957	0.4915	16.9001	0.4626	1.1246	
min	1	0	0	2,023	1	
25%	2,500.75	0	22	2,023	1	
50%	5,000.5	1	37	2,024	3	
75%	7,500.25	1	48	2,024	4	
max	10,000	2	59	2,024	4	

4. **Handling Missing Values**: AI suggests the best approach for handling missing data based on the dataset’s characteristics.



❗ Missing Values

	0
order_id	0
customer_name	0
salutation	9,801
cust_first_name	0
cust_middle_name	9,600
cust_last_name	0
cust_designation	9,763
restaurant_name	0
order_date	0
delivery_time	0
Time_Lapsed_Order	0



cust_designation (Data Type: object)

AI Suggestion for 'cust_designation': Missing values in 'cust_designation' have been filled with 'N/A'. Total entries now: 10000

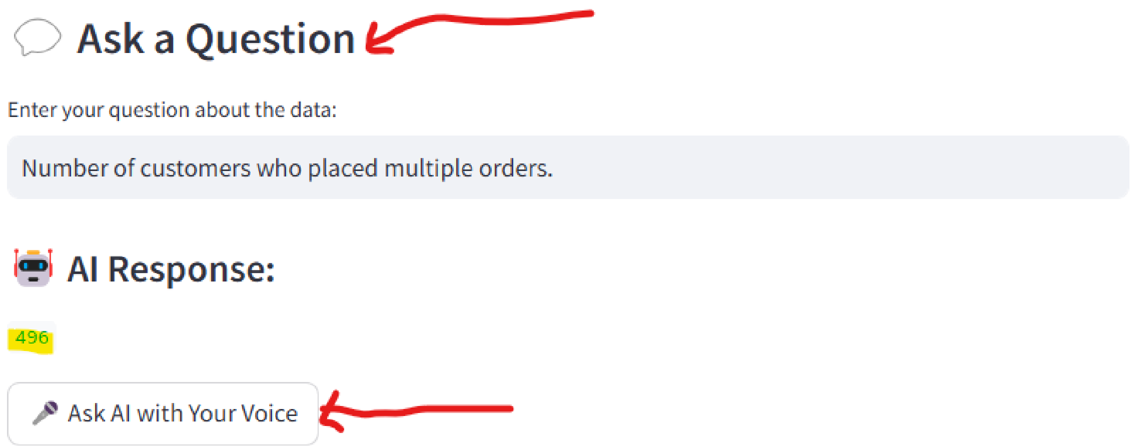
How to fill missing values in 'cust_designation'? (Categorical Data)

● Mode
○ Leave As Is
○ Use AI Suggestion

☑ Filled missing values in 'cust_designation' with Mode.

Once missing values filled you can show cleaned dataset and download as well

5. **AI Query System**: Users can either type queries or use speech recognition to ask data-related questions.



💬 Ask a Question

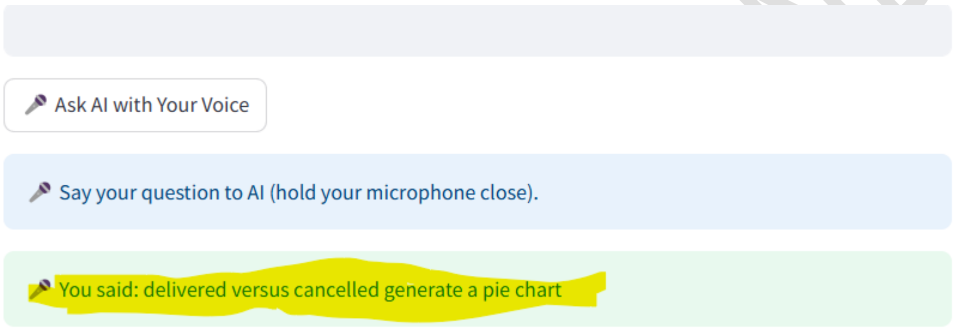
Enter your question about the data:

Number of customers who placed multiple orders.

🤖 AI Response:

496

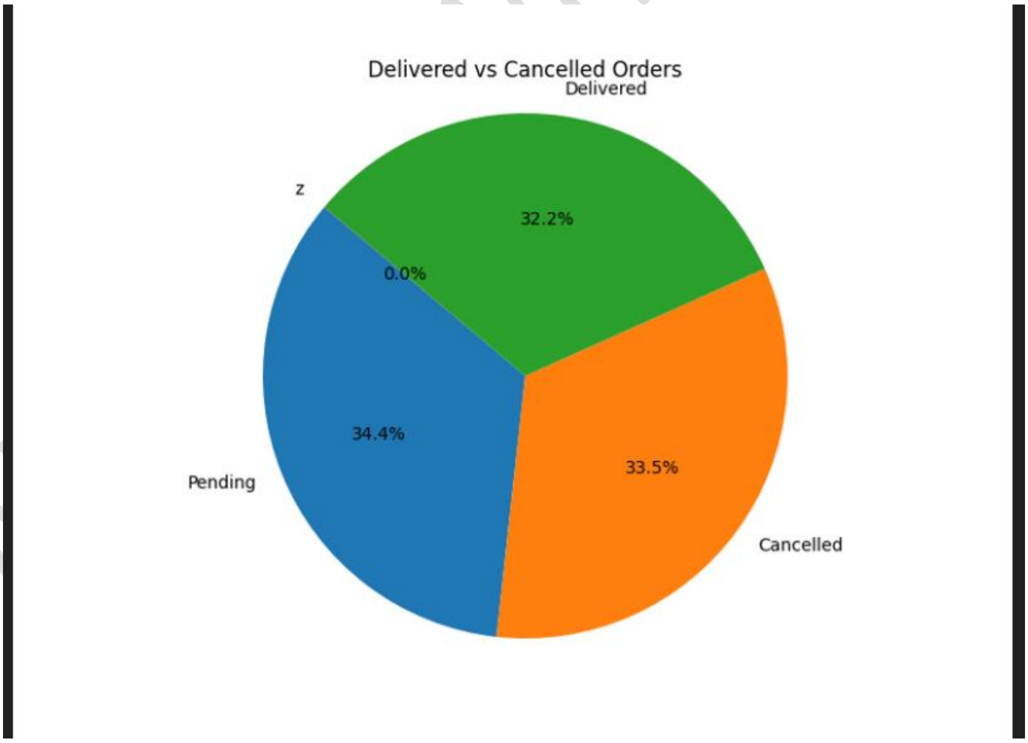
🎤 Ask AI with Your Voice



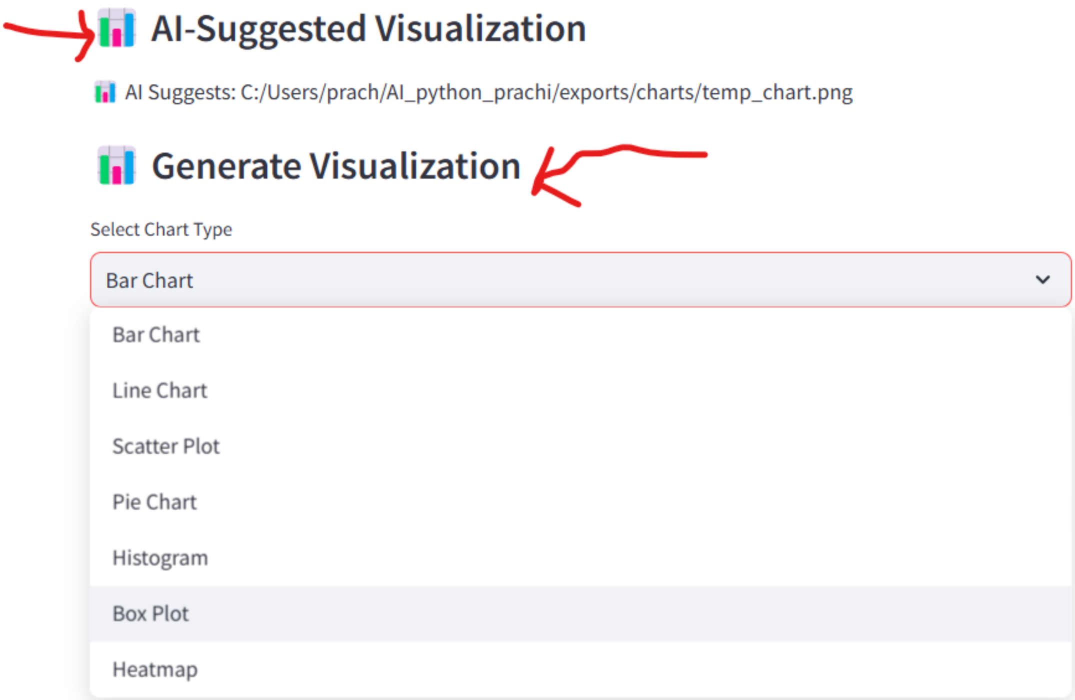
🎤 Ask AI with Your Voice

🎤 Say your question to AI (hold your microphone close).

🎤 You said: delivered versus cancelled generate a pie chart



6. **Data Visualization**: AI suggests the most relevant chart types, and users can generate and download visualizations.

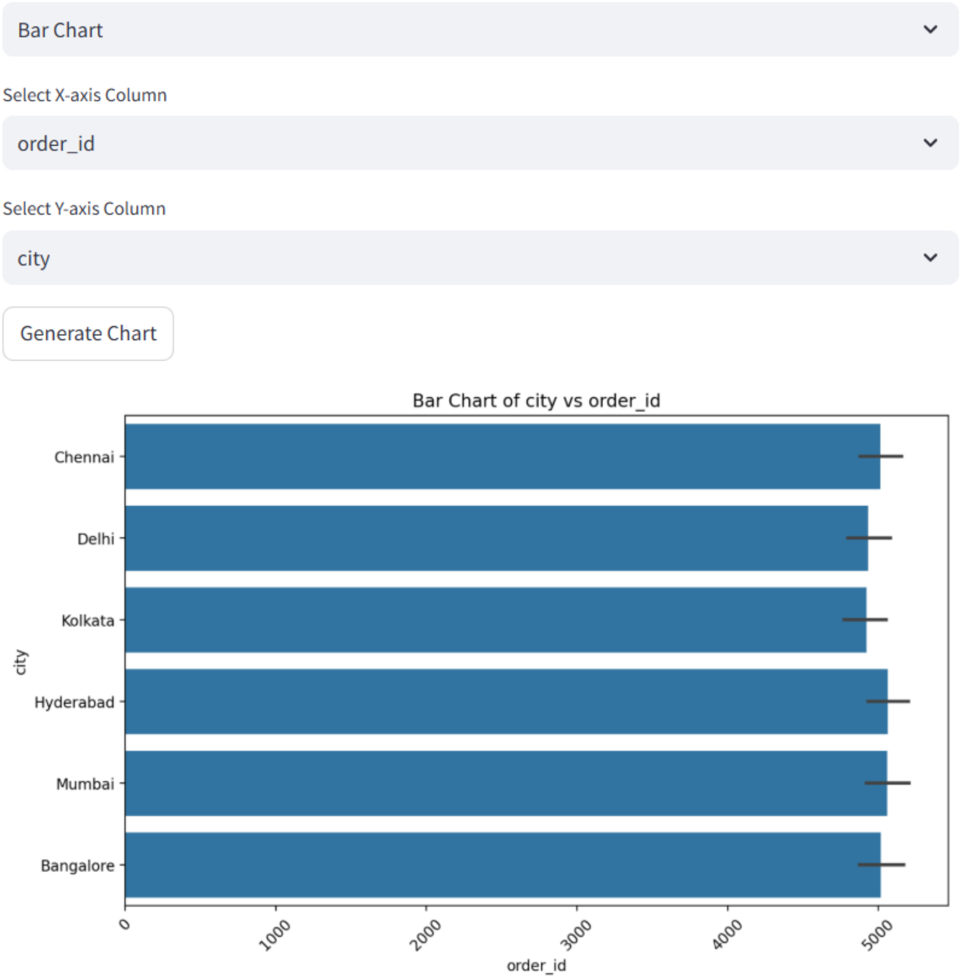


📊 AI-Suggested Visualization

📊 AI Suggests: C:/Users/prach/AI_python_prachi/exports/charts/temp_chart.png

📊 Generate Visualization

Select Chart Type: Bar Chart



Bar Chart

Select X-axis Column: order_id

Select Y-axis Column: city