BRACT's

Vishwakarma Institute of Information Technology, Pune - 411048

Department of CSE (Artificial Intelligence)

Context AI Data Analyst

Name	PRN	Roll Number
Arya Yemul	22310290	381015
Akash Patel	22310745	381029
Prachi Shedge	22311647	381050
Srinidhi Soundarrajan	22311837	381064

Guided By - Anuradha Yenkikar Ma'am

Pranjal Pandit Ma'am

Problem Statement & Proposed Solution

Data analysis is time-consuming and requires technical skills. Non-technical users struggle to explore data, generate insights, and create reports without help from data analysts.

Our ContextAI powered by an open-source LLM that:

- Understands the data - schema and columns along with business context.
- Analyzes and visualizes data
- Generates insights and professional reports automatically

Implementation Details

Automatic File Type Detection: Supports CSV, Excel, JSON, PDF

LLM-Powered Cleaning: Context-aware data cleaning strategies

Outlier Detection: Intelligent identification of anomalies

Data Profiling: Comprehensive statistical summaries Contextual Question Answering Natural Language Queries: Ask questions in plain English

Numbered Questions: Quick access to pre-generated questions

Context Indicators: Visual feedback when using historical insights

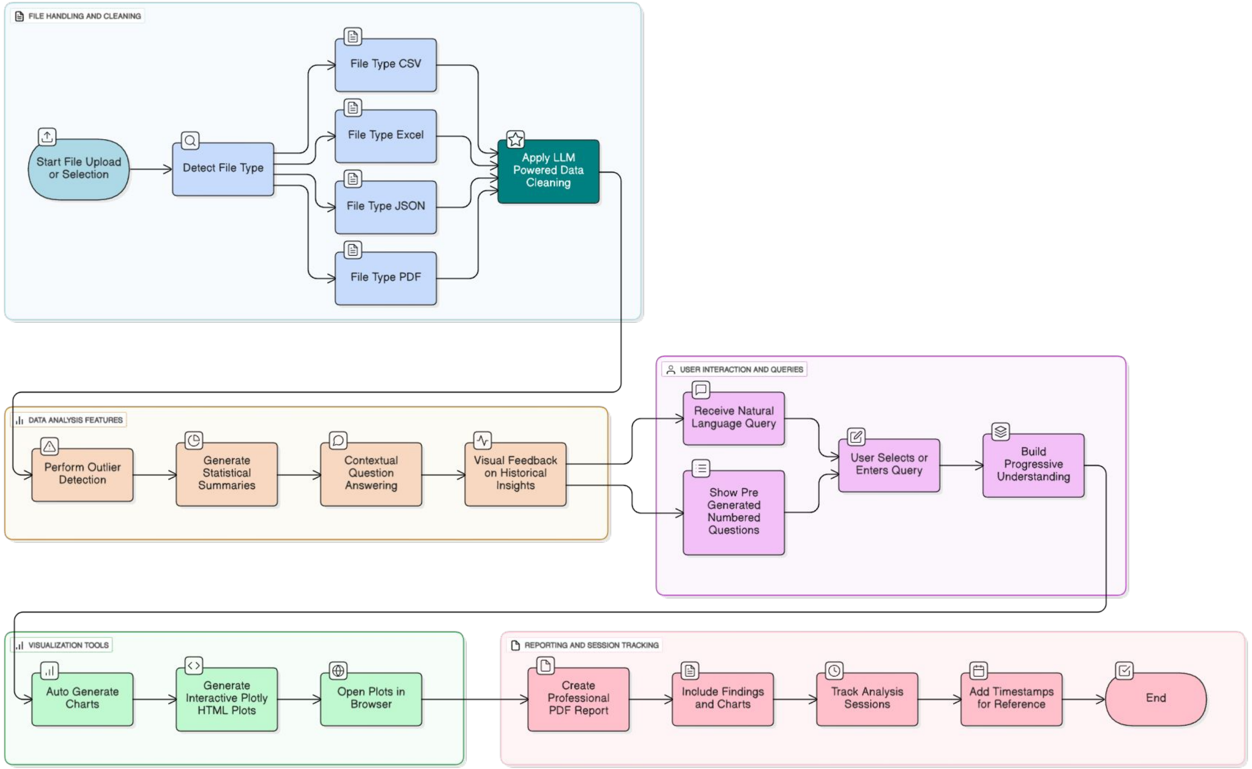
Progressive Analysis: Builds sophisticated understanding over time

Auto-Generated Visualizations: Charts created automatically with insights

Interactive HTML Plots: Plotly-powered visualizations that open in browser

Comprehensive PDF Reports: Professional reports with findings and charts Timestamp Tracking: All analysis sessions are dated and tracked summarize this

Flow of the System



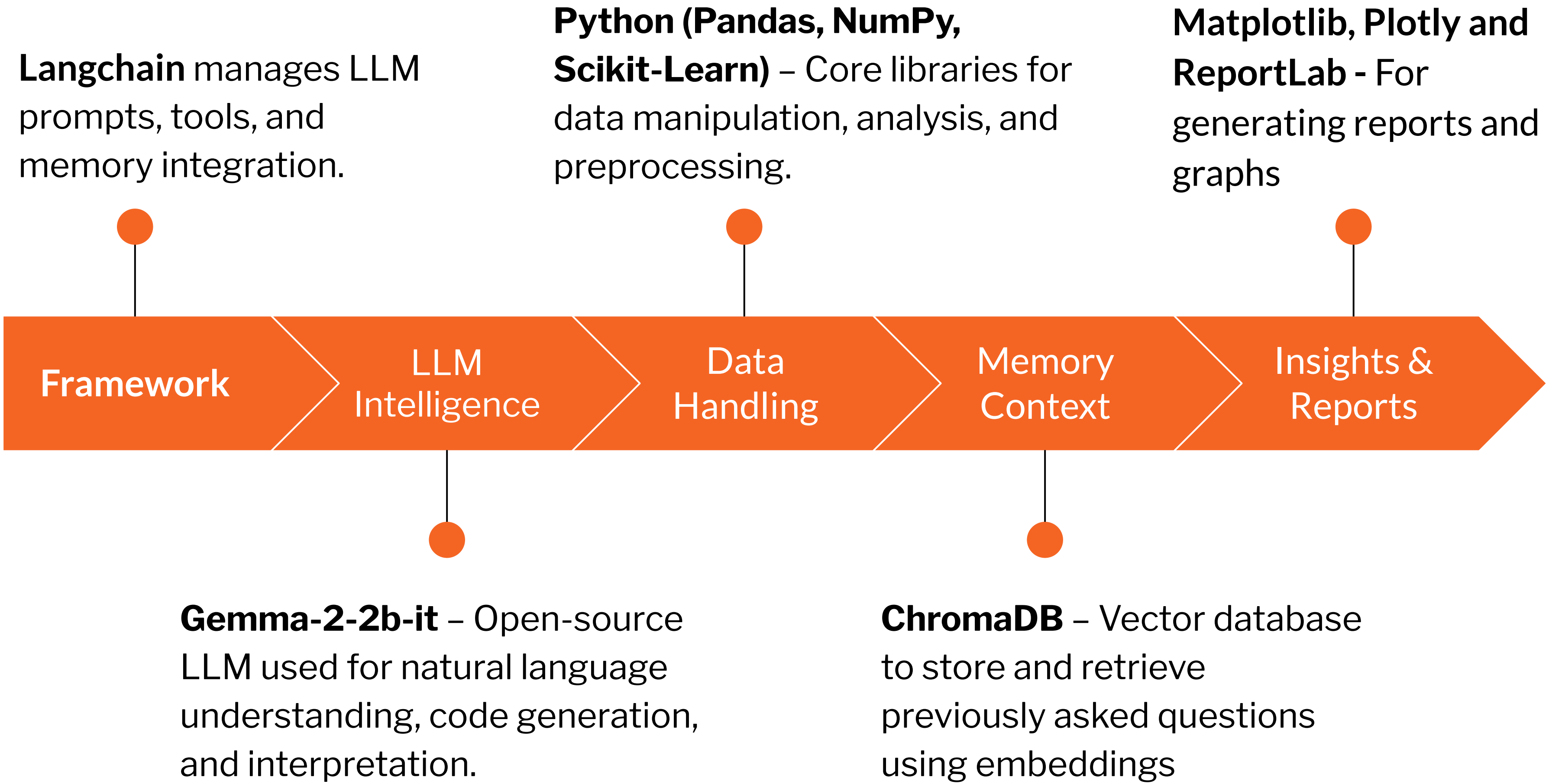
FILE HANDLING AND CLEANING
Start File Upload or Selection
Detect File Type
File Type CSV
File Type Excel
File Type JSON
File Type PDF
Apply LLM Powered Data Cleaning
DATA ANALYSIS FEATURES
Perform Outlier Detection
Generate Statistical Summaries
Contextual Question Answering
Visual Feedback on Historical Insights
USER INTERACTION AND QUERIES
Receive Natural Language Query
Show Pre Generated Numbered Questions
User Selects or Enters Query
Build Progressive Understanding
VISUALIZATION TOOLS
Auto Generate Charts
Generate Interactive Plotly HTML Plots
Open Plots in Browser
REPORTING AND SESSION TRACKING
Create Professional PDF Report
Include Findings and Charts
Track Analysis Sessions
Add Timestamps for Reference
End

Literature Survey

Research Title	Author	Summary
Data Formulator AI-powered Concept-driven Visualization	Chenglong Wang , John Thompson ,andBongshin Le Link - https://arxiv.org/pdf/2309.10094	Data Formulator uses AI and LLMs to transform data for visualization through natural language and Programming-by-Example. It provides multiple transformation options with visual and code-based feedback for user refinement.
Autonomous-AI-Agents-for-Real-Time-Data-Transformation-and-ETL-Automation	Raghavender Maddali Link - https://www.researchgate.net/publication/390363475_Autonomous_AI_Agents_for_Real-Time_Data_Transformation_and_ETL_Automation	The project employs a multi-layered agentic architecture to automate data analysis and decision-making across industries. By integrating ETL, real-time analytics, and adaptive feedback loops, it enhances speed, accuracy, and operational efficiency.

LLMs for Science: Usage for Code Generation and Data Analysis	Mohamed Nejjar,Luca Zacharias, Fabian Stiehle,Ingo Weber Link - https://arxiv.org/pdf/2311.16733	The paper assesses LLMs like ChatGPT and Copilot for scientific coding across 20 cases, highlighting both their potential and limitations. It emphasizes responsible use, improved evaluation, and further research for effective integration into scientific workflows.
Generative AI in Data Science: Applications in Automated Data Cleaning and Preprocessing for Machine Learning Models	Jeshwanth Reddy Machireddy, Prabu Ravichandran, Sareen Kumar Rachakatla Link - https://biotechjournal.org/index.php/jbai/article/view/71	The paper explores the use of Generative AI for data cleaning and preprocessing, emphasizing tasks like error detection, imputation, and synthetic data generation. Through case studies, it showcases the potential of models like GANs and VAEs to improve data quality, especially in healthcare and finance.
A Multi-Agentic AI Framework for Autonomous and Collaborative Data Science Workflows	Chirag,Prof. Sarab Nihal Singh Nagra Link - https://gcared.ganitara.com/proceedings/gcared25/papers/P31.pdf	The paper introduces a modular multi-agent AI framework that automates end-to-end data science tasks using specialized agents coordinated via LangGraph and FastAPI. Built with Python and LangChain, it enhances scalability, accuracy, and transparency while reducing development time by 60%.

Tech Stack Used



Framework

Langchain manages LLM prompts, tools, and memory integration.

LLM Intelligence

Gemma-2-2b-it – Open-source LLM used for natural language understanding, code generation, and interpretation.

Data Handling

Python (Pandas, NumPy, Scikit-Learn) – Core libraries for data manipulation, analysis, and preprocessing.

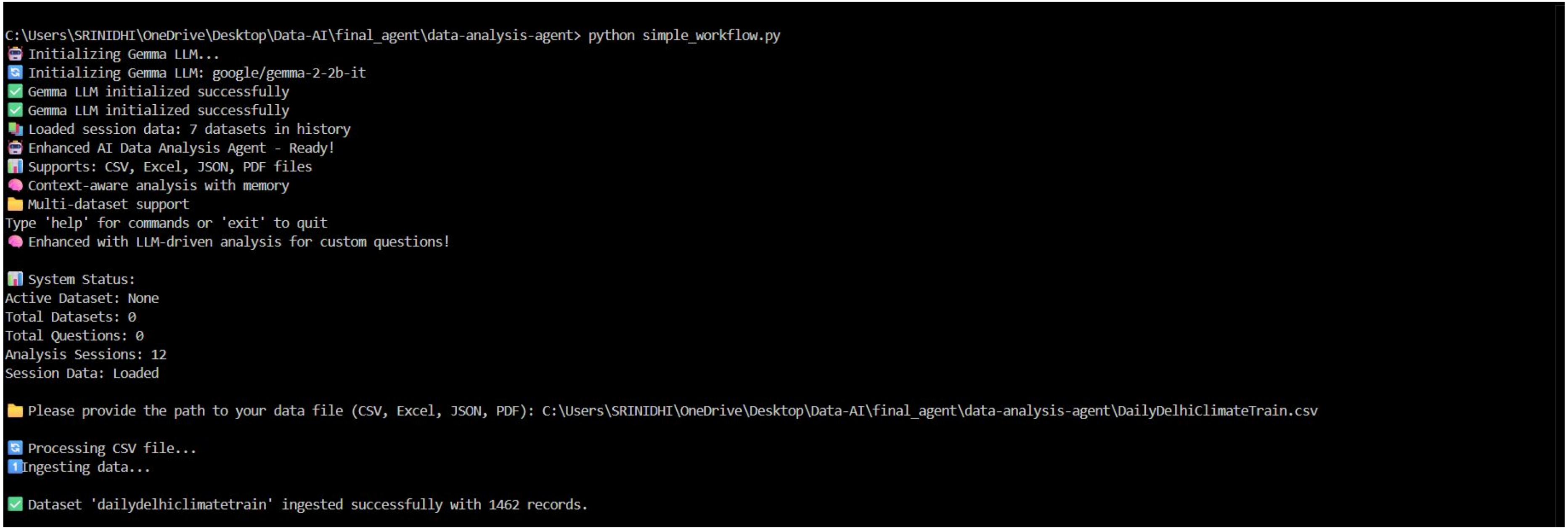
Memory Context

ChromaDB – Vector database to store and retrieve previously asked questions using embeddings

Insights & Reports

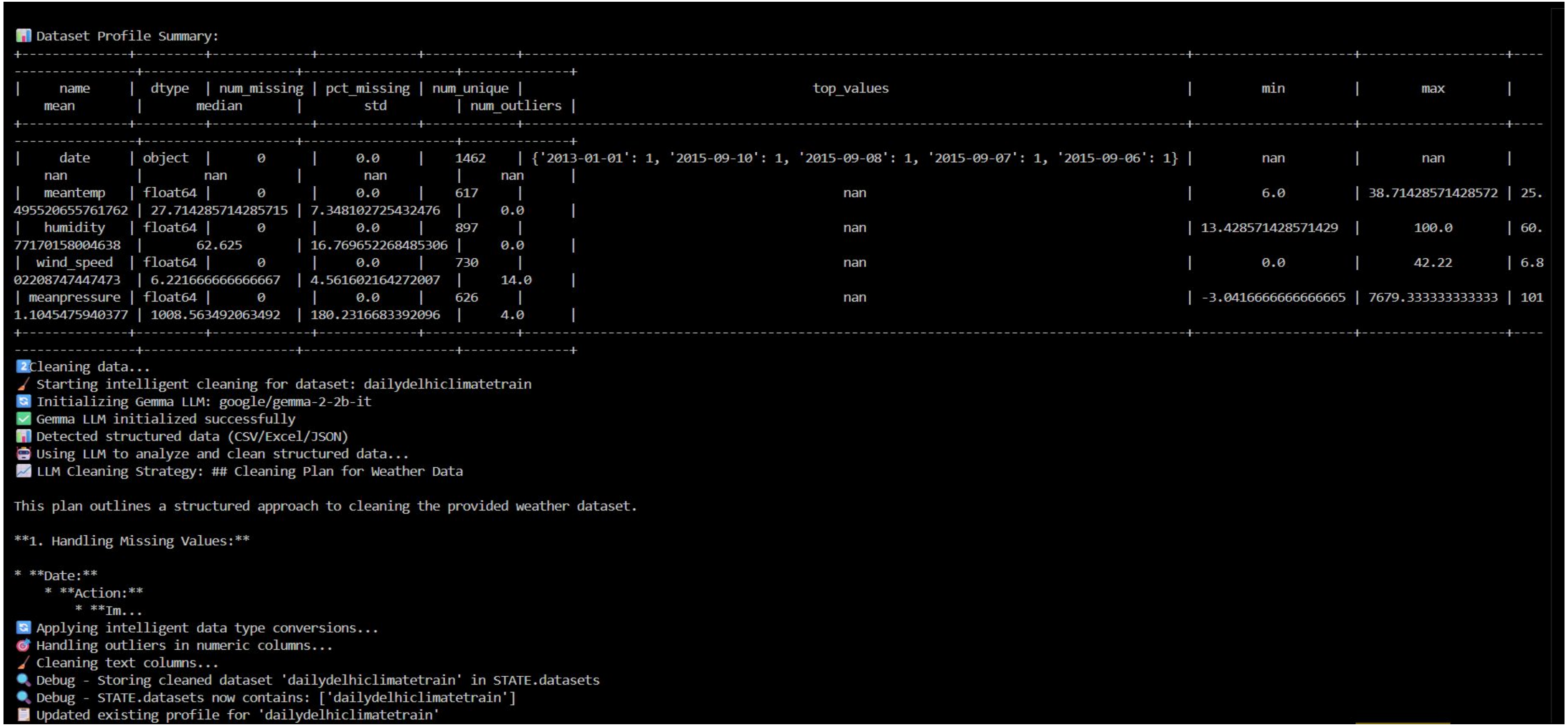
Matplotlib, Plotly and ReportLab - For generating reports and graphs

Results



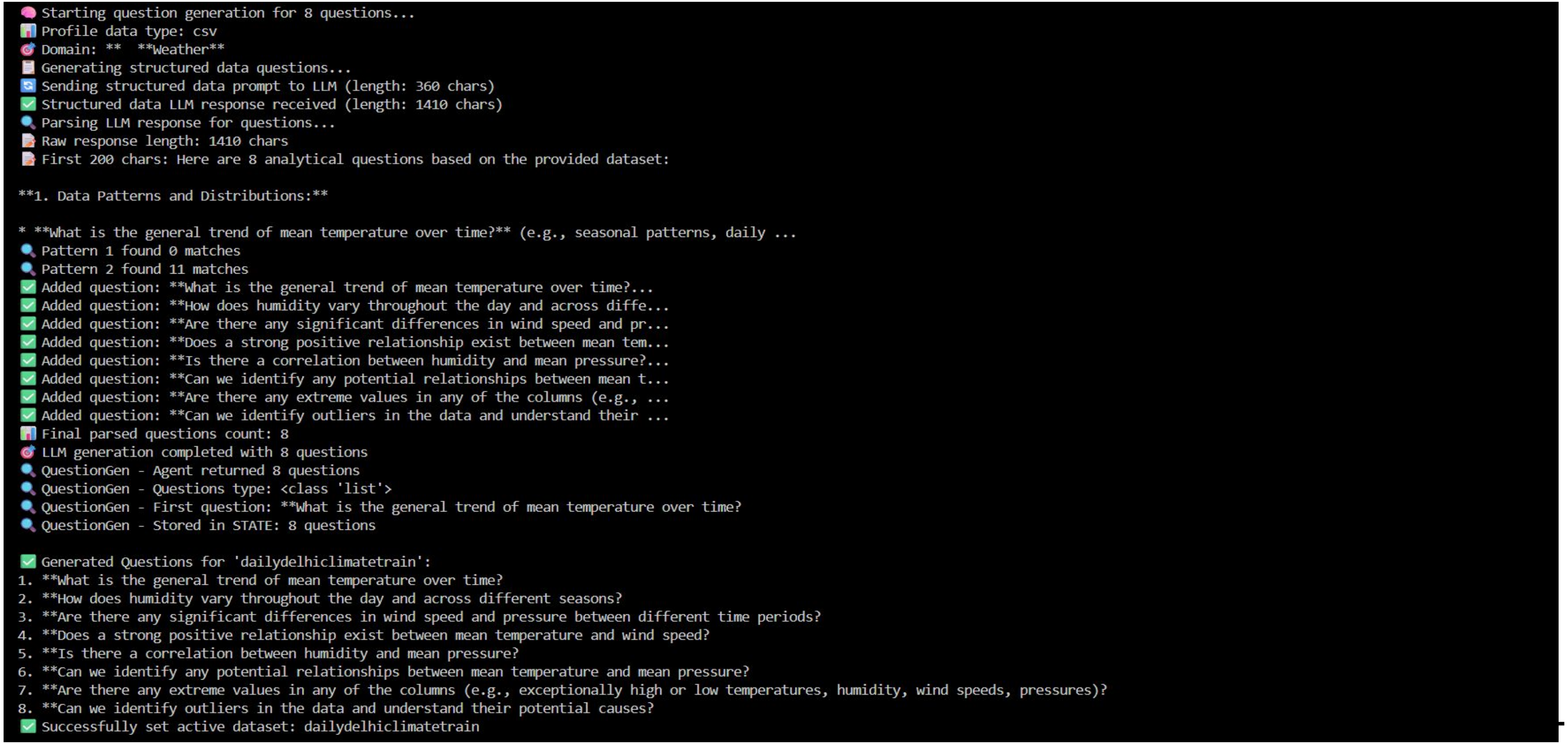
```
C:\Users\SRINIDHI\OneDrive\Desktop\Data-AI\final_agent\data-analysis-agent> python simple_workflow.py
🤖 Initializing Gemma LLM...
🔄 Initializing Gemma LLM: google/gemma-2-2b-it
✅ Gemma LLM initialized successfully
✅ Gemma LLM initialized successfully
📚 Loaded session data: 7 datasets in history
🤖 Enhanced AI Data Analysis Agent - Ready!
📊 Supports: CSV, Excel, JSON, PDF files
🧠 Context-aware analysis with memory
📁 Multi-dataset support
Type 'help' for commands or 'exit' to quit
🧠 Enhanced with LLM-driven analysis for custom questions!

📊 System Status:
Active Dataset: None
Total Datasets: 0
Total Questions: 0
Analysis Sessions: 12
Session Data: Loaded

📁 Please provide the path to your data file (CSV, Excel, JSON, PDF): C:\Users\SRINIDHI\OneDrive\Desktop\Data-AI\final_agent\data-analysis-agent\DailyDelhiClimateTrain.csv

🔄 Processing CSV file...
1️⃣Ingesting data...

✅ Dataset 'dailydelhiclimatetrain' ingested successfully with 1462 records.
```

Results



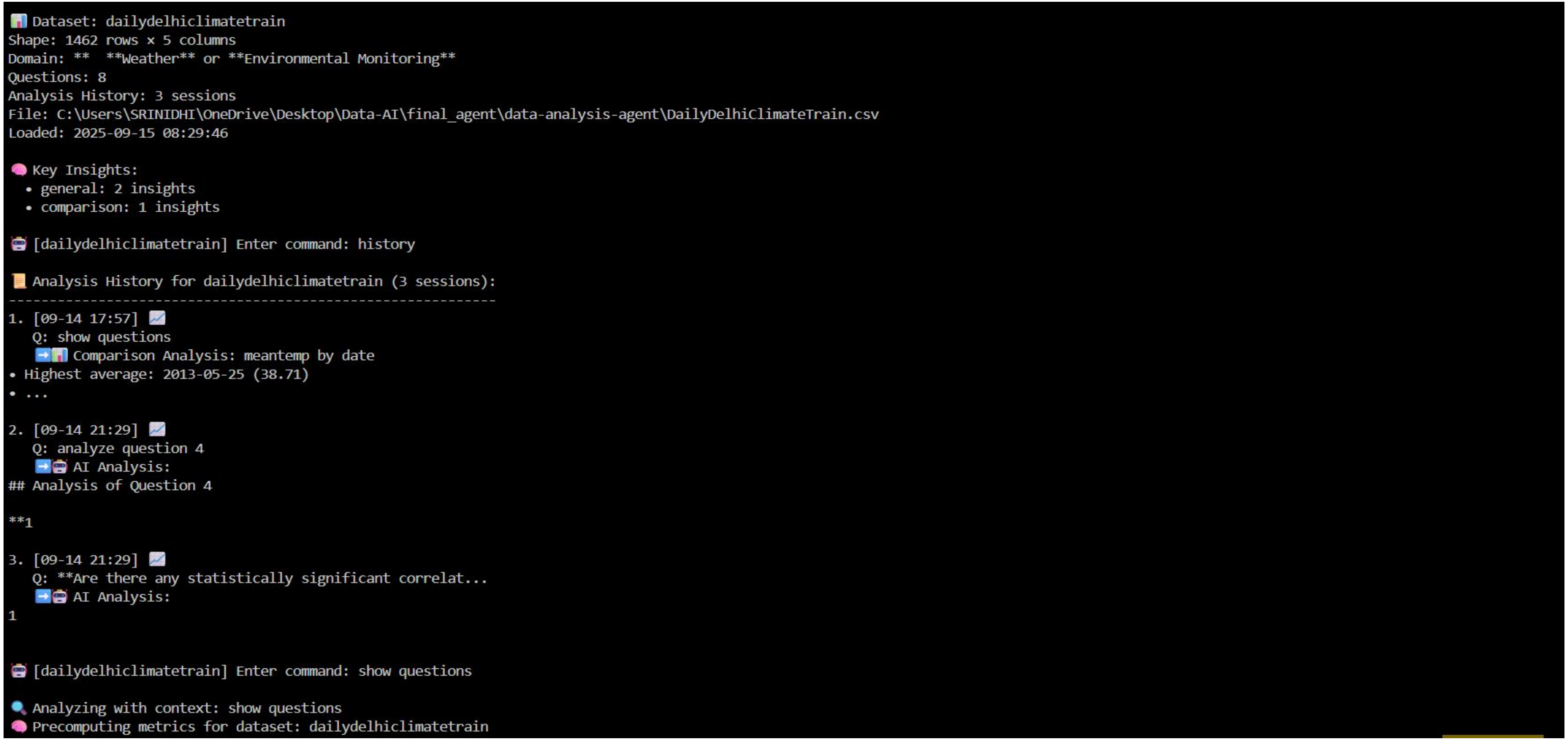
```
📊 Dataset Profile Summary:
+--------------+---------+-------------+-------------+------------+------------------------------------------------------------------------------------------+-------------------+-------------------+----
---------------+--------------------+--------------------+-------------+
|     name     |  dtype  | num_missing | pct_missing | num_unique |                                        top_values                                        |        min        |        max        |
    mean       |       median       |        std         | num_outliers |
+--------------+---------+-------------+-------------+------------+------------------------------------------------------------------------------------------+-------------------+-------------------+----
---------------+--------------------+--------------------+-------------+
|     date     | object  |      0      |     0.0     |    1462    | {'2013-01-01': 1, '2015-09-10': 1, '2015-09-08': 1, '2015-09-07': 1, '2015-09-06': 1} |        nan        |        nan        |
   nan         |        nan         |        nan         |     nan      |
|   meantemp   | float64 |      0      |     0.0     |    617     |                                           nan                                            |        6.0        | 38.71428571428572 | 25.
495520655761762 | 27.714285714285715 | 7.348102725432476  |     0.0      |
|   humidity   | float64 |      0      |     0.0     |    897     |                                           nan                                            | 13.428571428571429 |       100.0       | 60.
77170158004638  |       62.625       | 16.769652268485306 |     0.0      |
|  wind_speed  | float64 |      0      |     0.0     |    730     |                                           nan                                            |        0.0        |       42.22       | 6.8
02208747447473  | 6.221666666666667  | 4.561602164272007  |     14.0     |
| meanpressure | float64 |      0      |     0.0     |    626     |                                           nan                                            | -3.0416666666666665 | 7679.333333333333 | 101
1.1045475940377 | 1008.563492063492  | 180.2316683392096  |     4.0      |
+--------------+---------+-------------+-------------+------------+------------------------------------------------------------------------------------------+-------------------+-------------------+----
---------------+--------------------+--------------------+-------------+
2️⃣Cleaning data...
🧹 Starting intelligent cleaning for dataset: dailydelhiclimatetrain
🔄 Initializing Gemma LLM: google/gemma-2-2b-it
✅ Gemma LLM initialized successfully
📊 Detected structured data (CSV/Excel/JSON)
🤖 Using LLM to analyze and clean structured data...
📈 LLM Cleaning Strategy: ## Cleaning Plan for Weather Data

This plan outlines a structured approach to cleaning the provided weather dataset.

**1. Handling Missing Values:**

* **Date:**
    * **Action:**
        * **Im...
🔄 Applying intelligent data type conversions...
🎯 Handling outliers in numeric columns...
🧹 Cleaning text columns...
🔍 Debug - Storing cleaned dataset 'dailydelhiclimatetrain' in STATE.datasets
🔍 Debug - STATE.datasets now contains: ['dailydelhiclimatetrain']
📋 Updated existing profile for 'dailydelhiclimatetrain'
```

Results



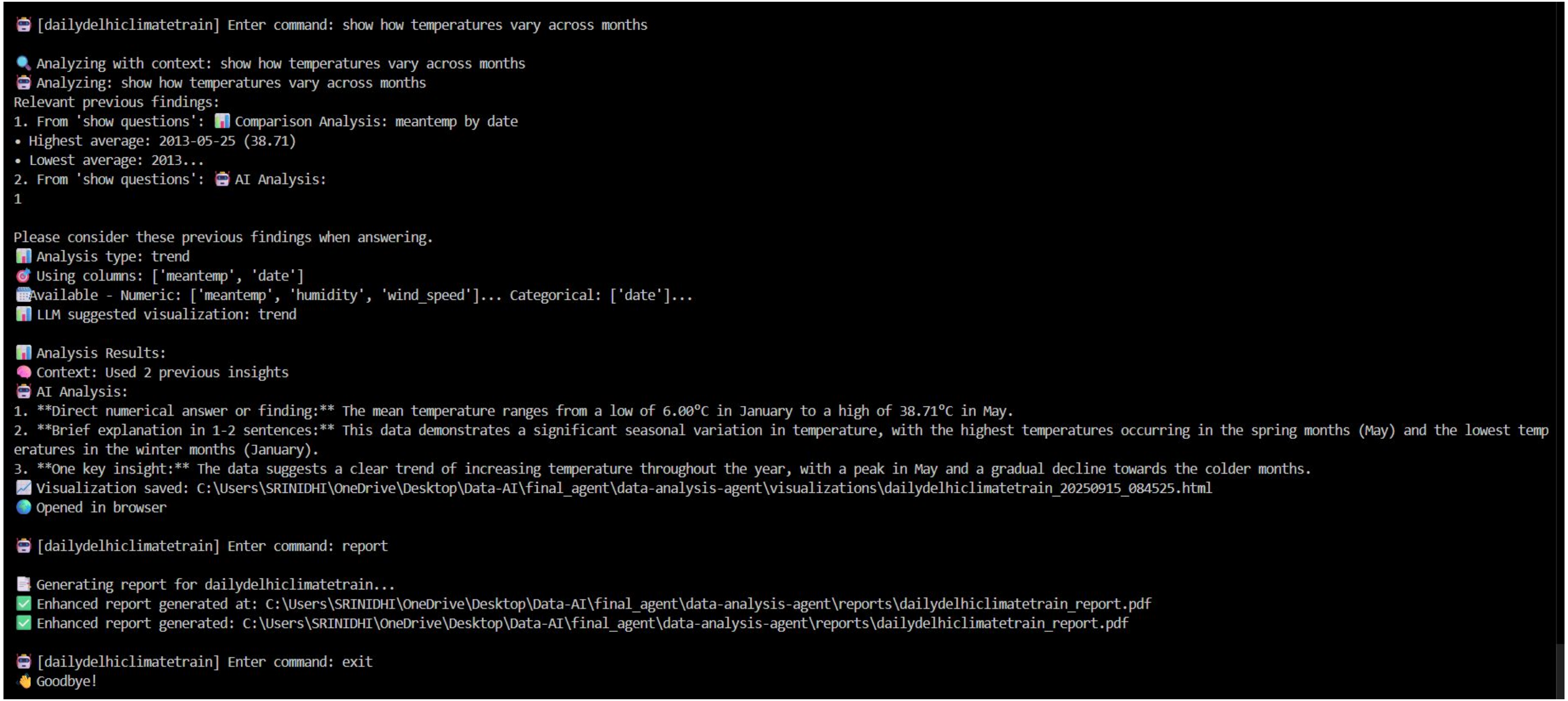
```
🧠 Starting question generation for 8 questions...
📊 Profile data type: csv
🎯 Domain: **  **Weather**
📋 Generating structured data questions...
🔄 Sending structured data prompt to LLM (length: 360 chars)
✅ Structured data LLM response received (length: 1410 chars)
🔍 Parsing LLM response for questions...
📝 Raw response length: 1410 chars
📝 First 200 chars: Here are 8 analytical questions based on the provided dataset:

**1. Data Patterns and Distributions:**

* **What is the general trend of mean temperature over time?** (e.g., seasonal patterns, daily ...
🔍 Pattern 1 found 0 matches
🔍 Pattern 2 found 11 matches
✅ Added question: **What is the general trend of mean temperature over time?...
✅ Added question: **How does humidity vary throughout the day and across diffe...
✅ Added question: **Are there any significant differences in wind speed and pr...
✅ Added question: **Does a strong positive relationship exist between mean tem...
✅ Added question: **Is there a correlation between humidity and mean pressure?...
✅ Added question: **Can we identify any potential relationships between mean t...
✅ Added question: **Are there any extreme values in any of the columns (e.g., ...
✅ Added question: **Can we identify outliers in the data and understand their ...
📊 Final parsed questions count: 8
🎯 LLM generation completed with 8 questions
🔍 QuestionGen - Agent returned 8 questions
🔍 QuestionGen - Questions type: <class 'list'>
🔍 QuestionGen - First question: **What is the general trend of mean temperature over time?
🔍 QuestionGen - Stored in STATE: 8 questions

✅ Generated Questions for 'dailydelhiclimatetrain':
1. **What is the general trend of mean temperature over time?
2. **How does humidity vary throughout the day and across different seasons?
3. **Are there any significant differences in wind speed and pressure between different time periods?
4. **Does a strong positive relationship exist between mean temperature and wind speed?
5. **Is there a correlation between humidity and mean pressure?
6. **Can we identify any potential relationships between mean temperature and mean pressure?
7. **Are there any extreme values in any of the columns (e.g., exceptionally high or low temperatures, humidity, wind speeds, pressures)?
8. **Can we identify outliers in the data and understand their potential causes?
✅ Successfully set active dataset: dailydelhiclimatetrain
```

Results



```
📊 Dataset: dailydelhiclimatetrain
Shape: 1462 rows × 5 columns
Domain: **  **Weather** or **Environmental Monitoring**
Questions: 8
Analysis History: 3 sessions
File: C:\Users\SRINIDHI\OneDrive\Desktop\Data-AI\final_agent\data-analysis-agent\DailyDelhiClimateTrain.csv
Loaded: 2025-09-15 08:29:46

🧠 Key Insights:
  • general: 2 insights
  • comparison: 1 insights

🤖 [dailydelhiclimatetrain] Enter command: history

📜 Analysis History for dailydelhiclimatetrain (3 sessions):
------------------------------------------------------------
1. [09-14 17:57] 📈
   Q: show questions
   ➡️📊 Comparison Analysis: meantemp by date
• Highest average: 2013-05-25 (38.71)
• ...

2. [09-14 21:29] 📈
   Q: analyze question 4
   ➡️🤖 AI Analysis:
## Analysis of Question 4

**1

3. [09-14 21:29] 📈
   Q: **Are there any statistically significant correlat...
   ➡️🤖 AI Analysis:
1


🤖 [dailydelhiclimatetrain] Enter command: show questions

🔍 Analyzing with context: show questions
🧠 Precomputing metrics for dataset: dailydelhiclimatetrain
```

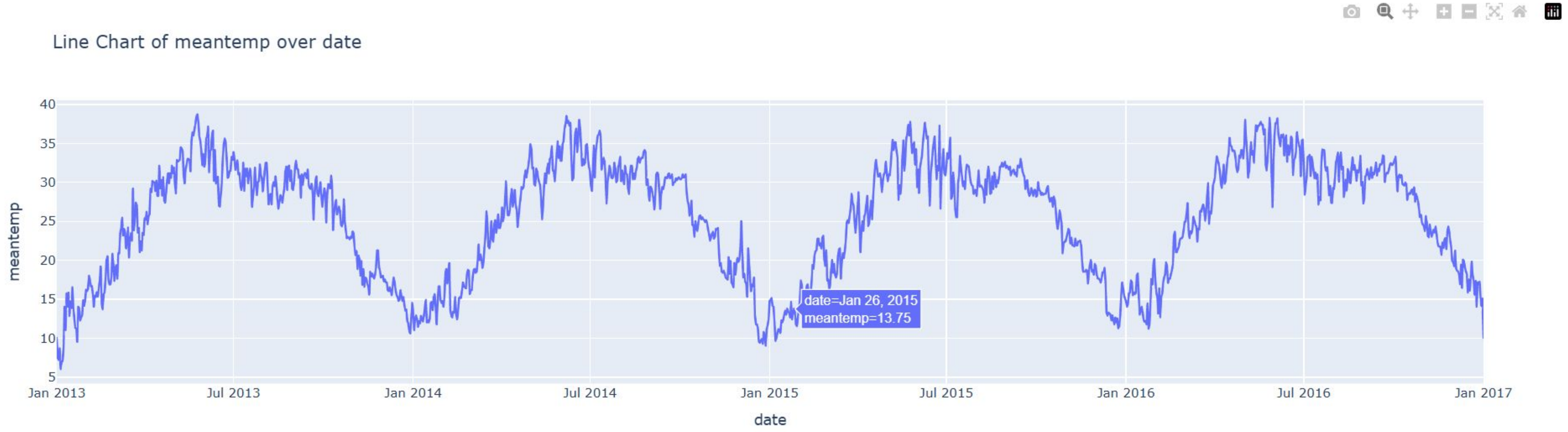
Results



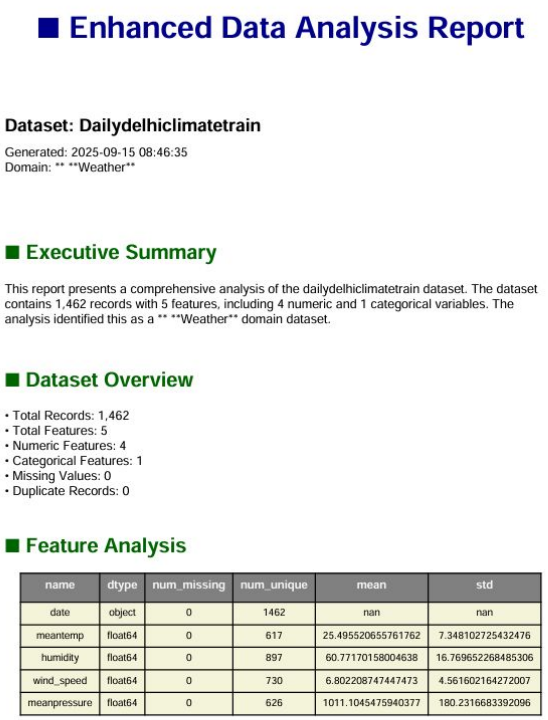
```
🤖 [dailydelhiclimatetrain] Enter command: show how temperatures vary across months

🔍 Analyzing with context: show how temperatures vary across months
🤖 Analyzing: show how temperatures vary across months
Relevant previous findings:
1. From 'show questions': 📊 Comparison Analysis: meantemp by date
• Highest average: 2013-05-25 (38.71)
• Lowest average: 2013...
2. From 'show questions': 🤖 AI Analysis:
1

Please consider these previous findings when answering.
📊 Analysis type: trend
🎯 Using columns: ['meantemp', 'date']
📅Available - Numeric: ['meantemp', 'humidity', 'wind_speed']... Categorical: ['date']...
📊 LLM suggested visualization: trend

📊 Analysis Results:
🧠 Context: Used 2 previous insights
🤖 AI Analysis:
1. **Direct numerical answer or finding:** The mean temperature ranges from a low of 6.00°C in January to a high of 38.71°C in May.
2. **Brief explanation in 1-2 sentences:** This data demonstrates a significant seasonal variation in temperature, with the highest temperatures occurring in the spring months (May) and the lowest temp
eratures in the winter months (January).
3. **One key insight:** The data suggests a clear trend of increasing temperature throughout the year, with a peak in May and a gradual decline towards the colder months.
📈 Visualization saved: C:\Users\SRINIDHI\OneDrive\Desktop\Data-AI\final_agent\data-analysis-agent\visualizations\dailydelhiclimatetrain_20250915_084525.html
🌐 Opened in browser

🤖 [dailydelhiclimatetrain] Enter command: report

📑 Generating report for dailydelhiclimatetrain...
✅ Enhanced report generated at: C:\Users\SRINIDHI\OneDrive\Desktop\Data-AI\final_agent\data-analysis-agent\reports\dailydelhiclimatetrain_report.pdf
✅ Enhanced report generated: C:\Users\SRINIDHI\OneDrive\Desktop\Data-AI\final_agent\data-analysis-agent\reports\dailydelhiclimatetrain_report.pdf

🤖 [dailydelhiclimatetrain] Enter command: exit
👋 Goodbye!
```

Results

Question: show how temperatures vary across months



Results



■ Enhanced Data Analysis Report

Dataset: Dailydelhiclimatetrain

Generated: 2025-09-15 08:46:35
Domain: ** **Weather**

■ Executive Summary

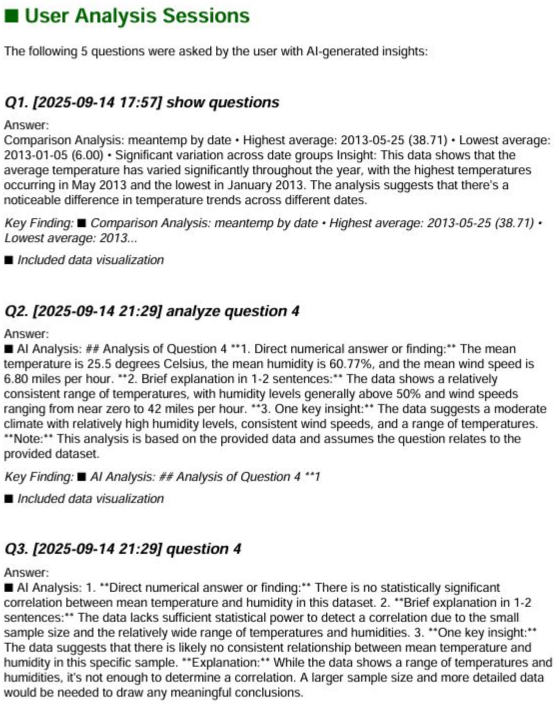
This report presents a comprehensive analysis of the dailydelhiclimatetrain dataset. The dataset contains 1,462 records with 5 features, including 4 numeric and 1 categorical variables. The analysis identified this as a ** **Weather** domain dataset.

■ Dataset Overview

• Total Records: 1,462
• Total Features: 5
• Numeric Features: 4
• Categorical Features: 1
• Missing Values: 0
• Duplicate Records: 0

■ Feature Analysis

name	dtype	num_missing	num_unique	mean	std
date	object	0	1462	nan	nan
meantemp	float64	0	617	25.495520655761762	7.348102725432476
humidity	float64	0	897	60.77170158004638	16.769652268485306
wind_speed	float64	0	730	6.802208747447473	4.561602164272007
meanpressure	float64	0	626	1011.1045475940377	180.2316683392096



■ User Analysis Sessions

The following 5 questions were asked by the user with AI-generated insights:

Q1. [2025-09-14 17:57] show questions

Answer:
Comparison Analysis: meantemp by date • Highest average: 2013-05-25 (38.71) • Lowest average: 2013-01-05 (6.00) • Significant variation across date groups Insight: This data shows that the average temperature has varied significantly throughout the year, with the highest temperatures occurring in May 2013 and the lowest in January 2013. The analysis suggests that there's a noticeable difference in temperature trends across different dates.

Key Finding: ■ Comparison Analysis: meantemp by date • Highest average: 2013-05-25 (38.71) • Lowest average: 2013...

■ *Included data visualization*

Q2. [2025-09-14 21:29] analyze question 4

Answer:
■ AI Analysis: ## Analysis of Question 4 **1. Direct numerical answer or finding:** The mean temperature is 25.5 degrees Celsius, the mean humidity is 60.77%, and the mean wind speed is 6.80 miles per hour. **2. Brief explanation in 1-2 sentences:** The data shows a relatively consistent range of temperatures, with humidity levels generally above 50% and wind speeds ranging from near zero to 42 miles per hour. **3. One key insight:** The data suggests a moderate climate with relatively high humidity levels, consistent wind speeds, and a range of temperatures. **Note:** This analysis is based on the provided data and assumes the question relates to the provided dataset.

*Key Finding: ■ AI Analysis: ## Analysis of Question 4 **1*

■ *Included data visualization*

Q3. [2025-09-14 21:29] question 4

Answer:
■ AI Analysis: 1. **Direct numerical answer or finding:** There is no statistically significant correlation between mean temperature and humidity in this dataset. 2. **Brief explanation in 1-2 sentences:** The data lacks sufficient statistical power to detect a correlation due to the small sample size and the relatively wide range of temperatures and humidities. 3. **One key insight:** The data suggests that there is likely no consistent relationship between mean temperature and humidity in this specific sample. **Explanation:** While the data shows a range of temperatures and humidities, it's not enough to determine a correlation. A larger sample size and more detailed data would be needed to draw any meaningful conclusions.

Conclusion

We have developed an **AI-powered data analyst** that can automatically explore datasets, understand user questions, and generate insights through text or visuals. Using LangChain, and open-source LLMs, it will simplify the entire analysis process through a conversational interface — making data analysis easier and faster.

Thank You!