Retail Insights Assistant Chatbot

Project Overview:

A multi-agent, AI-powered chatbot for querying and summarizing business sales data using LangChain + Google Gemini, Streamlit, and DuckDB.

System Architecture Overview

Efficient Data Loading

- Sales CSV files are imported into Pandas DataFrames, enabling efficient data handling for subsequent analysis.

In-Memory SQL Analytics

- DuckDB registers the data for fast, in-memory SQL operations, speeding up complex analytics tasks.

AI Coordination and Summarization

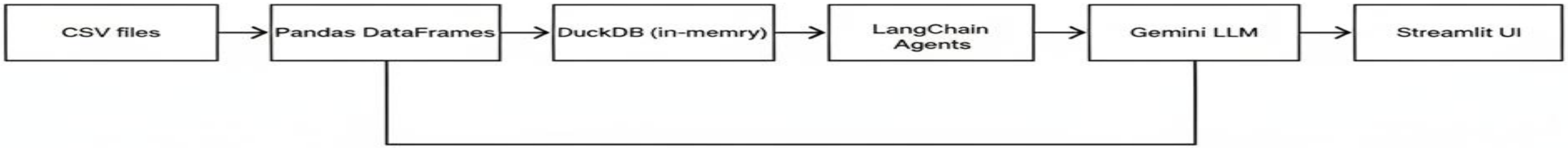
- LangChain Agents manage the workflow, using large language models to summarize analytics in natural language.

Interactive User Experience

- A Streamlit-powered user interface offers seamless and interactive chat for engaging with analytics and summaries.

System Architecture Overview

Flow diagram



CSV files

Pandas DataFrames

DuckDB (in-memry)

LangChain Agents

Gemini LLM

Streamlit UI

LLM Integration Strategy

- Application uses LangChain’s ChatGoogleGenerativeAI class to access Gemini API.
- LangChain agent coordinates tool execution (SQL queries, summaries) and manages dialogs.
- User queries are parsed and agent selects tools as needed.
- Insights and summaries generated by Gemini API delivered via Streamlit.

Data Storage, Indexing & Retrieval Design for 100GB Scale

- All CSV files loaded into DuckDB in memory for rapid analytics.
- For large datasets, consider partitioning, distributed DuckDB, or cloud warehouses.
- Frequently accessed columns should be indexed and results cached.
- LangChain tools map database queries for retrieval.

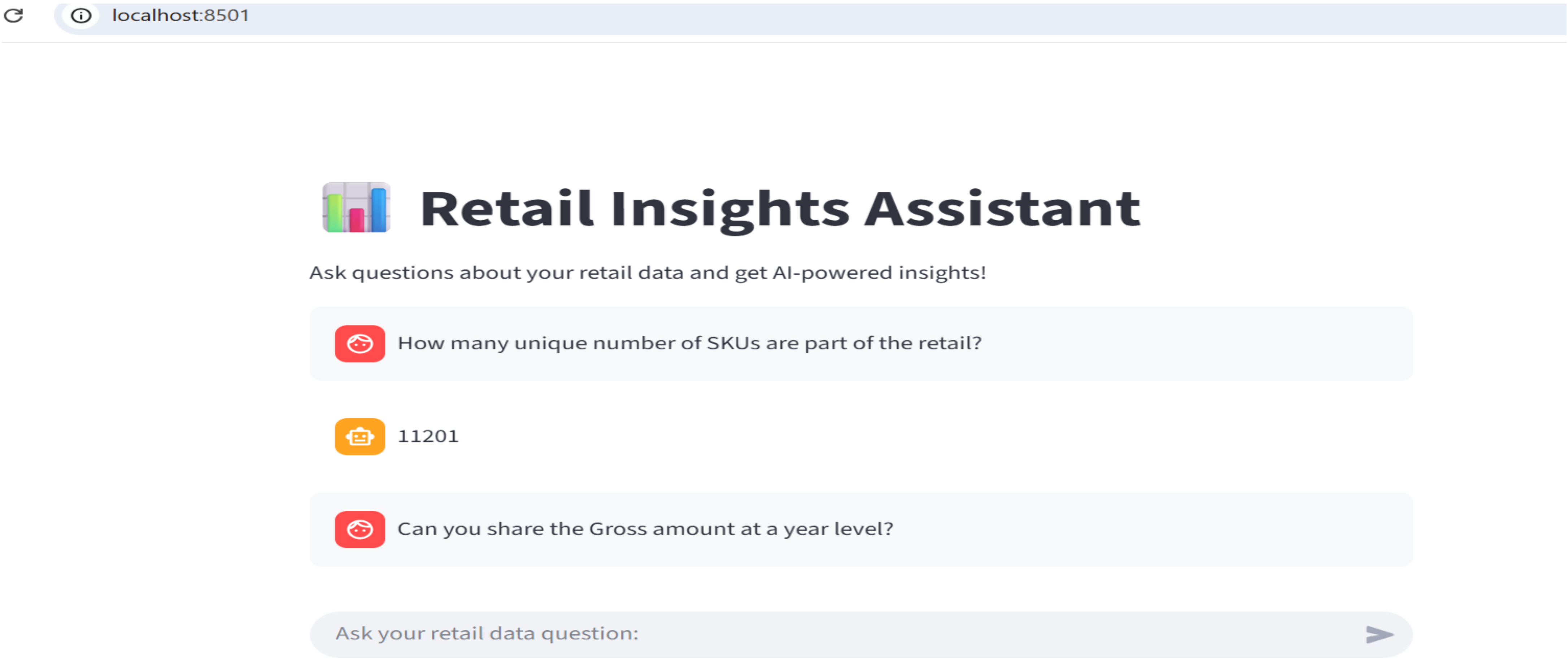
Example Query-Response Pipeline

- User enters sales-related question in Streamlit chat UI.
- Agent interprets intent with Gemini, forms SQL query, runs on DuckDB.
- Query results optionally summarized by LLM.
- Results presented in conversational Streamlit UI.

Cost and Performance Considerations

- Gemini API usage billed by tokens/requests—estimate per traffic.
- In-memory DuckDB fast for small/medium data but RAM limited.
- Distributed/cloud DBs handle massive data with higher cost/latency.
- Optimize queries, batch requests, cache for performance.

Demo & Screenshots



localhost:8501

Retail Insights Assistant

Ask questions about your retail data and get AI-powered insights!

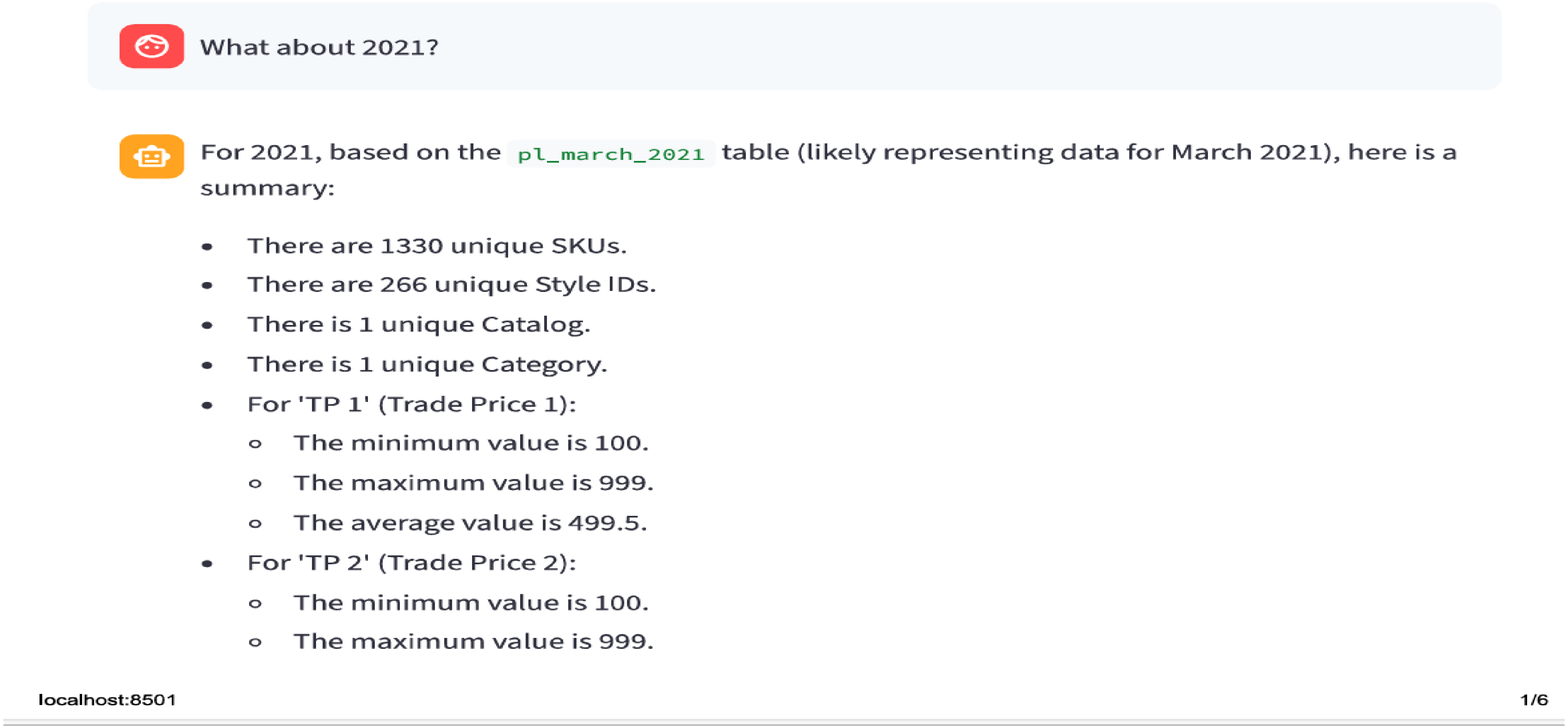
How many unique number of SKUs are part of the retail?

11201

Can you share the Gross amount at a year level?

Ask your retail data question:

Demo & Screenshots



What about 2021?

For 2021, based on the `pl_march_2021` table (likely representing data for March 2021), here is a summary:

- There are 1330 unique SKUs.
- There are 266 unique Style IDs.
- There is 1 unique Catalog.
- There is 1 unique Category.
- For 'TP 1' (Trade Price 1):
  - The minimum value is 100.
  - The maximum value is 999.
  - The average value is 499.5.
- For 'TP 2' (Trade Price 2):
  - The minimum value is 100.
  - The maximum value is 999.

localhost:8501 1/6

Summary & Next Steps

Retail Insights Assistant delivers AI-powered, multi-agent analysis over sales data. Scalable, modular, and cloud-ready (with suitable backend DB).

Next Steps:

- Add persistent conversation memory.
- Integrate data visualizations.
- Support for larger, distributed datasets.
- Implement memory or context management if multi-turn conversation is required.
- Add necessary error handling or fallback responses for missing/empty data.