Evaluating Large Language Models for SQL-Based Business Analysis

Kenny Okeke

MKT 188 SQL For Business and Marketing

McCombs School of Business
The University of Texas at Austin

February 22, 2026

Executive Summary

This report evaluates the effectiveness of three leading large language models (LLMs) for SQL-based business analytics: GPT-5.2 Plus (OpenAI), Gemini 3 Pro (Google), and Claude Sonnet 4.6 (Anthropic). The analysis was designed from the perspective of a private organization seeking to understand whether modern AI tools can support common data analysis tasks and whether premium models provide meaningful advantages over free alternatives.

An experiment was conducted using a structured sales dataset representative of data commonly analyzed within business environments. Each model was given identical prompts and asked to generate SQL queries and interpret the resulting outputs across several analytical scenarios. These tasks included identifying top-performing product lines, determining the highest-value customers, and evaluating revenue by geographic region. Model responses were evaluated using a standardized scoring rubric that assessed SQL correctness, logical accuracy, query efficiency, schema understanding, and usability.

The results show that all three models were able to successfully complete the tasks and generate valid SQL queries that produced accurate analytical outputs. Differences between the models were relatively small. GPT-5.2 Plus consistently generated reliable queries and concise business explanations. Gemini 3 Pro demonstrated strong SQL reasoning but occasionally required minor adjustments to align with the dataset structure. Claude Sonnet 4.6 performed comparably to the premium models and often provided detailed interpretations of the results.

A key finding from the experiment is that the free-tier model performed competitively with the paid tools. This suggests that organizations may not always need the most advanced or expensive models to support routine SQL-based analysis. However, premium platforms may still provide advantages in reliability, governance, security features, and enterprise integration.

Beyond performance, the report also evaluates privacy considerations, operational limitations, and cost factors associated with adopting LLMs for analytics workflows. Because these tools typically operate through cloud-based systems, companies must consider how data is shared, stored, and protected. Enterprise versions of these platforms often provide stronger data governance controls and may be better suited for organizations handling sensitive information.

Overall, the findings indicate that modern LLMs can be effective assistants for SQL-based business analysis. While no single model clearly outperformed the others, each demonstrated the ability to support common analytics tasks. As a result, organizations should base adoption decisions not only on technical capability but also on factors such as privacy protections, operational cost, and compatibility with existing analytics processes.

Table of Contents

List of Tables

List of Figures

Large Language Models (LLMs)

Organizations increasingly rely on data analysis to support strategic and operational decision making. Many business analysts use Structured Query Language (SQL) to retrieve, summarize, and interpret data stored in organizational databases. However, writing SQL queries can be time consuming and requires technical expertise, particularly when analysts are exploring unfamiliar datasets or attempting to answer complex business questions. As large language models (LLMs) have advanced, many companies have begun experimenting with using these systems to generate SQL queries and assist with data analysis tasks. This raises an important question for businesses: which LLMs are most effective for supporting SQL-based analytics workflows?

Three leading LLMs were selected for this analysis: GPT-5.2 Plus from OpenAI, Gemini 3 Pro from Google, and Claude Sonnet 4.6 from Anthropic. Two of the models (GPT-5.2 Plus and Gemini 3 Pro) represent premium, paid versions of their respective platforms, while Claude Sonnet 4.6 represents a widely available free-tier model. Including both paid and free models allows the experiment to evaluate whether higher-cost premium models provide a meaningful advantage when performing SQL-based data analysis tasks. From a private business perspective, this distinction is important because organizations must weigh the benefits of more advanced paid tools against the cost savings of using free or lower-cost alternatives.

Experiment

To evaluate the effectiveness of large language models for SQL-based data analysis, an experiment was conducted using a real-world sales dataset commonly used for business analytics. The dataset contains transactional sales records including product information, customer data, geographic attributes, and revenue fields, enabling realistic analytical queries similar to those performed in private organizations.

Three large language models were tested. Each model received the same dataset and the same prompt describing a business analytics task in order to evaluate how effectively the models could analyze the data and answer business questions using SQL.

Multiple analytical scenarios were included to better reflect real-world business analysis. These tasks included identifying top-performing product lines by revenue, determining the highest-value customers, and analyzing sales performance across geographic regions. Using several tasks provides a broader assessment of how well each model supports common business intelligence workflows.

For each task, the dataset was uploaded to the model and the prompt was used without modification. The models were asked to generate the SQL required to answer the question and produce the resulting output from the dataset. The generated queries were then reviewed and tested where possible to confirm that they would run successfully and return the intended results.

Model outputs were evaluated using criteria relevant to real business usage: SQL correctness, logical accuracy, query efficiency, schema understanding, and usability. SQL correctness evaluates whether the query is syntactically valid and executable. Logical accuracy measures whether the query correctly answers the business question. Query efficiency considers whether the SQL is structured clearly and avoids unnecessary complexity. Schema understanding evaluates whether the model correctly uses existing dataset fields without inventing new ones. Usability assesses whether the query and results are organized in a way that a business analyst could realistically use.

To ensure consistent comparison, outputs were scored using a standardized rubric. Each criterion was rated from 1 to 5, with higher scores indicating stronger performance. Scoring was conducted manually using the same rubric across all models to maintain consistency and reduce subjectivity.

Finally, the experiment was designed to be reproducible. By using the same dataset, prompts, and scoring framework, the evaluation can be replicated with additional models in future analyses.

Analysis

The results of the experiment show that all three large language models were capable of generating logically correct SQL queries and producing accurate outputs from the dataset. Across the tasks, each model correctly interpreted the business questions and constructed appropriate aggregation queries using grouping, ordering, and limiting. The overall performance comparison is summarized in Table 1, while detailed scoring for each model and task can be found in Table 2 through Table 10. The corresponding model outputs are shown in Figure 1 through Figure 9. These results indicate that modern LLMs are generally reliable for basic analytical SQL tasks such as ranking, summarizing, and identifying top performers within a dataset.

ChatGPT consistently produced queries that executed correctly without modification and clearly aligned with the structure of the dataset (see Figure 1 through Figure 3 and Table 2 through Table 4). In addition to generating valid SQL, it also provided concise explanations of the results and highlighted the key business implications. This made the outputs particularly useful from a business perspective because the model did not only perform the analysis but also translated the findings into insights that a decision maker could understand quickly.

Gemini also demonstrated strong SQL reasoning and produced logically correct queries for each task (see Figure 4 through Figure 6 and Table 5 through Table 7). However, it repeatedly referenced a slightly different table name than the one used in the experiment. While this was a minor issue that could easily be corrected, it does indicate a small reliability gap when compared to the other models. In practical settings, this type of mismatch could require an analyst to make adjustments before executing the query.

Gemini's explanations were also somewhat less detailed in some cases, which slightly reduced the usability of the output.

Claude Sonnet performed at a similarly high level to ChatGPT (see Figure 7 through Figure 9 and Table 8 through Table 10). The model generated correct SQL queries that worked with the dataset and provided detailed interpretations of the results. In several cases, Claude expanded on the implications of the analysis, such as identifying revenue concentration among top customers or geographic dependence on specific markets. These types of observations are valuable in a business environment because they move beyond raw numbers and help identify potential strategic considerations.

An interesting outcome of the experiment is that the free-tier model (Claude Sonnet 4.6) performed as well as the premium models in terms of SQL accuracy and analytical usefulness (see Table 1 for comparative totals). This suggests that organizations may not necessarily need the most advanced paid models to perform routine SQL-based analysis. While premium models may still offer advantages in more complex analytical or technical scenarios, the results here indicate that a well-performing free model can still deliver strong analytical outputs.

Overall, the experiment demonstrates that large language models can be effective tools for supporting SQL-based data analysis. All three systems successfully completed the assigned tasks, though small differences emerged in reliability and clarity of interpretation. These findings suggest that LLMs can serve as useful assistants for analysts, particularly for generating queries and summarizing results, while still benefiting from human review to ensure accuracy and execution compatibility.

Privacy and Data Use Policies

When evaluating large language models for business analytics tasks, organizations must consider how each platform handles privacy, data usage, and security. While GPT-5.2 Plus, Gemini 3 Pro, and Claude Sonnet 4.6 are capable of performing SQL-based analysis, their data handling policies play an important role in determining whether they are appropriate for real business environments.

Because these tools operate through cloud-based systems, datasets and prompts submitted for analysis may leave a company's internal infrastructure. For organizations working with proprietary or customer information, this raises concerns about data retention, access, and whether submitted information could be used to improve future versions of the models. Companies therefore need to carefully review provider policies and determine whether the level of protection aligns with their internal data governance requirements.

One practical solution is to use business or enterprise versions of these platforms rather than consumer tools. Enterprise offerings from providers such as OpenAI, Google, and Anthropic generally provide stronger controls around data retention, administrative access, and model training. In many cases, commercial plans specify that customer data

is not used to train models by default and may include additional options such as limited or zero retention, encryption controls, and administrative governance features. For smaller organizations, these plans can provide meaningful privacy protections without requiring the company to build its own internal AI models.

Beyond vendor agreements, companies can further reduce risk through operational practices. Analysts can limit the amount of information shared with a model, remove personally identifiable information, or use sampled or synthetic datasets during exploration. Another common approach is to have the model generate SQL queries while the actual database execution remains inside the company's secure environment, preventing the AI system from accessing the full underlying dataset.

In this experiment, privacy concerns were limited because the dataset used was publicly available and contained no sensitive information. However, the workflow closely resembles how analysts might interact with these tools in a real business context. As organizations increasingly integrate AI into analytics processes, careful vendor evaluation, internal governance policies, and responsible data handling practices will be necessary to ensure that sensitive information remains protected.

Overall, while GPT-5.2 Plus, Gemini 3 Pro, and Claude Sonnet 4.6 demonstrate strong analytical capabilities, privacy and security considerations remain an important factor when deciding how these tools should be used within a business environment.

Caveats and Concerns

While the experiment provides useful insights into how large language models perform SQL-based analytics tasks, several limitations should be considered when interpreting the results.

First, the dataset used in the experiment was relatively clean and structured. Real business datasets are often significantly more complex, containing missing values, inconsistent formatting, multiple related tables, and evolving schemas. Because the models were tested using a single structured dataset, the results may represent a best-case scenario compared to real organizational data environments.

Second, the evaluation relied on a limited number of analytical tasks. Although the tasks were designed to reflect common business questions such as identifying top product lines, customers, and geographic performance, they do not capture the full range of SQL problems analysts face. More complex tasks involving joins across multiple tables, nested queries, or data cleaning may produce different performance differences between the models.

Another caveat is that the scoring process involved manual evaluation. While a standardized rubric was used to maintain consistency, some level of subjective judgment is unavoidable when assessing factors such as usability and query structure. Different evaluators may assign slightly different scores even when reviewing the same outputs.

There are also limitations related to how large language models operate. LLMs generate responses probabilistically, meaning the same prompt can occasionally produce slightly different outputs across runs. In a production environment, this variability could affect reliability if organizations rely heavily on automated query generation without verification.

Additionally, the models did not directly connect to a live database environment. Instead, they were asked to analyze an uploaded dataset and generate SQL-style solutions. In real business workflows, models may interact with database tools, dashboards, or analytics platforms, which could influence both the accuracy and usefulness of the generated queries.

Finally, the comparison included two premium models and one free-tier model. While this allowed the study to explore potential differences between paid and free tools, pricing tiers do not perfectly represent capability differences across all use cases. Model performance may vary depending on updates, configuration, or the specific interfaces through which organizations access the systems.

Despite these caveats, the experiment still provides a practical comparison of how modern large language models can support SQL-based business analysis tasks and highlights areas where human validation and governance remain important.

Cost Analysis

Cost is another important factor for organizations considering whether to use large language models for business analytics. While the experiment compared model performance, companies must also evaluate how much these tools cost per user and whether premium or enterprise offerings provide sufficient value.

For individual users, the pricing across the major platforms is relatively similar. OpenAI's ChatGPT Plus subscription is approximately $20 per user per month, or about $240 per year. Google's Gemini Advanced offering is priced at roughly $19.99 per month, which results in a similar annual cost of around $240 per user. Anthropic's Claude Pro plan is also typically priced at about $20 per month. At the individual level, this means that the three major providers fall into a similar pricing range, making cost differences between premium subscriptions relatively small for a single analyst.

However, businesses typically do not rely on individual subscriptions when adopting AI tools at scale. Instead, companies often use team or enterprise versions that include administrative controls, security features, and collaboration capabilities. These plans generally cost slightly more per user but provide stronger governance and privacy protections. ChatGPT Team plans are commonly reported in the range of about $25 to $30 per user per month. Claude Team offerings are typically in a similar range. Google's enterprise-oriented Gemini tools vary more widely depending on the specific product and environment, but reported pricing often ranges from roughly $20 to $55 per user per month.

When scaled across an organization, these costs can increase quickly. For example, a team of ten analysts using a $20 per month plan would cost roughly $2,400 per year. If the organization instead adopted enterprise-tier tools at $30 to $50 per user per month, the same team could cost between approximately $3,600 and $6,000 annually. While these costs are not extremely high compared to many enterprise software platforms, they are still significant enough that companies must consider whether the productivity benefits justify the expense.

It is also important to note that subscription pricing does not always represent the full cost of using large language models. Many companies integrate these systems through APIs, where pricing is based on usage rather than fixed monthly subscriptions. In those cases, costs depend on the volume of prompts, size of datasets analyzed, and how frequently automated workflows call the models. Organizations performing large-scale analytics tasks may therefore experience higher operational costs than what simple subscription pricing suggests.

Overall, the pricing across OpenAI, Google, and Anthropic related offerings is relatively comparable at the individual level, while enterprise deployments introduce additional variation depending on governance, security, and usage requirements. As a result, businesses evaluating these tools must balance performance, privacy protections, and operational cost when deciding which platform to adopt.

Recommendation

The results of this experiment suggest that modern large language models are broadly capable of supporting SQL-based business analysis. Across all three tasks, GPT-5.2 Plus, Gemini 3 Pro, and Claude Sonnet 4.6 were able to correctly interpret the dataset, generate functional SQL queries, and produce meaningful analytical results. While minor differences appeared in formatting, explanation depth, and structure, the overall analytical accuracy was consistently strong across the models. This indicates that basic business intelligence tasks such as aggregations, rankings, and exploratory data analysis can be effectively supported by a range of current LLM platforms.

One of the most notable findings from the experiment is that the free-tier model performed competitively with the premium offerings. Claude Sonnet 4.6, despite being used in a free configuration, produced results that were comparable to those generated by GPT Plus and Gemini 3 Pro. In several cases it also provided detailed interpretations of the results that resembled the type of reasoning a business analyst might include in a report. This suggests that organizations may not necessarily need the most expensive tools to benefit from AI-assisted analytics, particularly for relatively straightforward SQL tasks.

At the same time, premium models still offer advantages that may be important in a business setting. Paid versions of these platforms typically provide higher usage limits, faster responses, more consistent performance, and access to enterprise features such as administrative controls, security configurations, and integration with other tools. For

companies planning to integrate AI into regular analytics workflows, these operational benefits may justify the additional cost even if raw SQL performance is similar across models.

From a private business perspective, the decision of which model to adopt should therefore not rely solely on query accuracy. Instead, organizations should evaluate the broader ecosystem surrounding each platform, including privacy protections, governance capabilities, reliability, and cost. The experiment demonstrates that multiple models are capable of producing accurate analytical outputs, meaning that factors such as data security policies, enterprise integration, and overall workflow compatibility may ultimately be more important than small differences in SQL generation.

Overall, the findings indicate that GPT-5.2 Plus, Gemini 3 Pro, and Claude Sonnet 4.6 are all viable tools for SQL-based analysis. Businesses can successfully leverage any of these platforms depending on their budget, security requirements, and operational needs. Rather than there being a single “best” model, the results suggest that organizations should focus on selecting the platform that best fits their internal policies, data governance standards, and long-term analytics strategy.

Appendix A – Prompt Template for the Experiment

You are a data analyst working for a private company.

I have uploaded a dataset called sales_data_sample.csv. The dataset contains transactional sales data including customers, products, locations, and revenue.

Using the dataset, write and execute SQL-style analysis to answer the business question below.

Requirements:

- Base your analysis only on the data provided.
- Use the existing columns in the dataset.
- Clearly show the SQL query used.
- Provide the final result from the data.

Business Question:

[INSERT TASK HERE]

Task 1 Prompt

Business Question:

Identify the top 5 product lines by total sales revenue in the United States. Show the product line, total revenue, and number of orders. Sort from highest to lowest revenue.

Task 2 Prompt

Business Question:

Find the top 5 customers by total sales revenue. Return the customer name, total revenue, and number of orders.

Task 3 Prompt

Business Question:

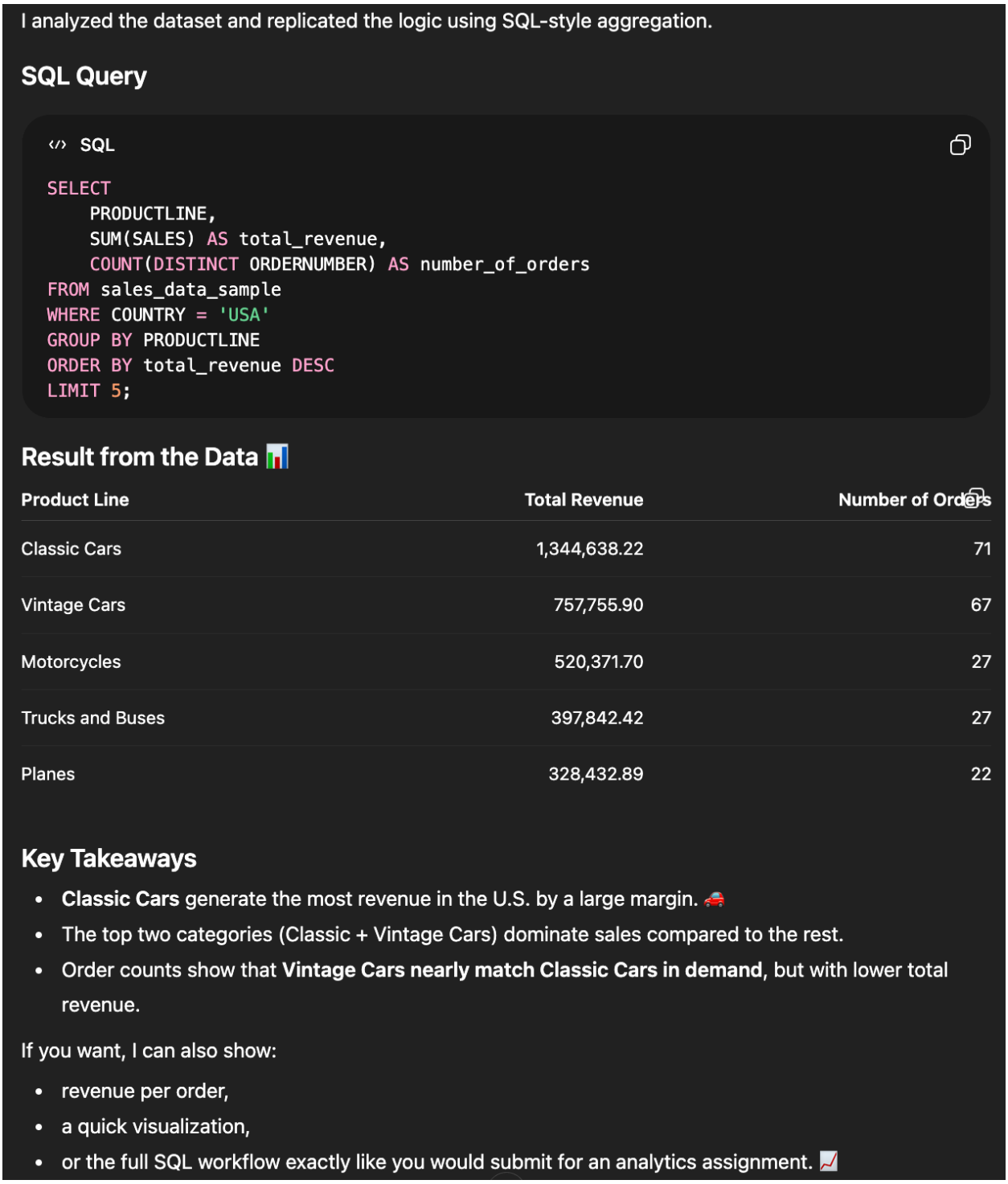
Calculate the total sales by country and return the top five countries by total revenue, sorted from highest to lowest.

Appendix B – Model Scoring

Table 1: Comparative Performance of LLMs on SQL Analysis Tasks

Model	Task 1	Task 2	Task 3	Total	Score (% of Maximum)
GPT-5.2 Plus	25	25	25	75	100.00%
Gemini 3 Pro	24	23	23	70	93.33%
Claude Sonnet 4.6	25	25	25	75	100.00%

Appendix C – GPT-5.2 Plus Task Outputs



I analyzed the dataset and replicated the logic using SQL-style aggregation.

SQL Query

SQL

```sql
SELECT
    PRODUCTLINE,
    SUM(SALES) AS total_revenue,
    COUNT(DISTINCT ORDERNUMBER) AS number_of_orders
FROM sales_data_sample
WHERE COUNTRY = 'USA'
GROUP BY PRODUCTLINE
ORDER BY total_revenue DESC
LIMIT 5;
```

Result from the Data 📊

Product Line	Total Revenue	Number of Orders
Classic Cars	1,344,638.22	71
Vintage Cars	757,755.90	67
Motorcycles	520,371.70	27
Trucks and Buses	397,842.42	27
Planes	328,432.89	22

Key Takeaways

- **Classic Cars** generate the most revenue in the U.S. by a large margin. 🚗
- The top two categories (Classic + Vintage Cars) dominate sales compared to the rest.
- Order counts show that **Vintage Cars nearly match Classic Cars in demand**, but with lower total revenue.

If you want, I can also show:

- revenue per order,
- a quick visualization,
- or the full SQL workflow exactly like you would submit for an analytics assignment. 📈

Figure 1: GPT-5.2 Plus Output for Task 1

Table 2: SQL Task Evaluation Results for GPT-5.2 Plus (Task 1)

Evaluation Factor	**Score**
SQL Correctness	5
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	25

SQL Query

```sql
SELECT
    CUSTOMERNAME,
    SUM(SALES) AS total_revenue,
    COUNT(DISTINCT ORDERNUMBER) AS number_of_orders
FROM sales_data_sample
GROUP BY CUSTOMERNAME
ORDER BY total_revenue DESC
LIMIT 5;
```

Result from the Data 📊

Customer Name	Total Revenue	Number of Orders
Euro Shopping Channel	912,294.11	26
Mini Gifts Distributors Ltd.	654,858.06	17
Australian Collectors, Co.	200,995.41	5
Muscle Machine Inc	197,736.94	4
La Rochelle Gifts	180,124.90	4

Summary

- **Euro Shopping Channel** is the largest customer by revenue, generating over **$912K** across **26 orders**.
- The top two customers contribute significantly more revenue than the rest.
- After the top two, revenue drops sharply, indicating a **high concentration of sales among a few customers**.

Figure 2: GPT-5.2 Plus Output for Task 2

Table 3: SQL Task Evaluation Results for GPT-5.2 Plus (Task 2)

Evaluation Factor	**Score**
SQL Correctness	5
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	25

SQL Query

SQL

```
SELECT
    COUNTRY,
    SUM(SALES) AS total_revenue
FROM sales_data_sample
GROUP BY COUNTRY
ORDER BY total_revenue DESC
LIMIT 5;
```

Result from the Data 📊

Country	Total Revenue
USA	3,627,982.83
Spain	1,215,686.92
France	1,110,916.52
Australia	630,623.10
UK	478,880.46

Summary

- The **United States** generates by far the most revenue, accounting for over **$3.6M** in sales.
- **Spain and France** are the next strongest markets but are significantly behind the U.S.
- The top five countries represent the majority of total sales, showing that revenue is concentrated in a few geographic markets. 📈

Figure 3: GPT-5.2 Plus Output for Task 3

Table 4: SQL Task Evaluation Results for GPT-5.2 Plus (Task 3)

Evaluation Factor	**Score**
SQL Correctness	5
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	25

Appendix D – Gemini 3 Pro Task Outputs

Based on the dataset provided, here is the SQL-style analysis to identify the top 5 product lines by total sales revenue in the United States.

SQL Query Used

SQL

```sql
SELECT
    PRODUCTLINE,
    ROUND(SUM(SALES), 2) AS Total_Revenue,
    COUNT(DISTINCT ORDERNUMBER) AS Number_Of_Orders
FROM sales_data
WHERE COUNTRY = 'USA'
GROUP BY PRODUCTLINE
ORDER BY Total_Revenue DESC
LIMIT 5;
```

(Note: `COUNT(DISTINCT ORDERNUMBER)` is used to count the unique number of orders rather than individual line items.)

Final Result

Rank	Product Line	Total Revenue (USD)	Number of Orders
1	Classic Cars	$1,344,638.22	71
2	Vintage Cars	$757,755.90	67
3	Motorcycles	$520,371.70	27
4	Trucks and Buses	$397,842.42	27
5	Planes	$328,432.89	22

Export to Sheets

Insights:

Classic Cars are by far the highest-revenue product line in the United States, generating over $1.34 million across 71 distinct orders. They are followed by **Vintage Cars** ($757k) and **Motorcycles** ($520k).

Figure 4: Gemini 3 Pro Output for Task 1

Table 5: SQL Task Evaluation Results for Gemini 3 Pro (Task 1)

Evaluation Factor	**Score**
SQL Correctness	4
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	24

Based on the dataset provided, here is the SQL-style analysis to identify the top 5 customers by total sales revenue.

SQL Query Used

SQL

```sql
SELECT
    CUSTOMERNAME,
    ROUND(SUM(SALES), 2) AS Total_Revenue,
    COUNT(DISTINCT ORDERNUMBER) AS Number_Of_Orders
FROM sales_data
GROUP BY CUSTOMERNAME
ORDER BY Total_Revenue DESC
LIMIT 5;
```

Final Result

Rank	Customer Name	Total Revenue (USD)	Number of Orders
1	Euro Shopping Channel	$912,294.11	26
2	Mini Gifts Distributors Ltd.	$654,858.06	17
3	Australian Collectors, Co.	$200,995.41	5
4	Muscle Machine Inc	$197,736.94	4
5	La Rochelle Gifts	$180,124.90	4

Export to Sheets

Figure 5: Gemini 3 Pro Output for Task 2

Table 6: SQL Task Evaluation Results for Gemini 3 Pro (Task 2)

Evaluation Factor	**Score**
SQL Correctness	4
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	4
Total	23

Based on the dataset provided, here is the SQL-style analysis to identify the top 5 countries by total sales revenue.

SQL Query Used

SQL

```sql
SELECT
    COUNTRY,
    ROUND(SUM(SALES), 2) AS Total_Revenue
FROM sales_data
GROUP BY COUNTRY
ORDER BY Total_Revenue DESC
LIMIT 5;
```

Final Result

Rank	Country	Total Revenue (USD)
1	USA	$3,627,982.83
2	Spain	$1,215,686.92
3	France	$1,110,916.52
4	Australia	$630,623.10
5	UK	$478,880.46

Export to Sheets

Figure 6: Gemini 3 Pro Output for Task 3

Table 7: SQL Task Evaluation Results for Gemini 3 Pro (Task 3)

Evaluation Factor	**Score**
SQL Correctness	4
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	4
Total	23

Appendix E – Claude Sonnet 4.6 Task Outputs

SQL Query (equivalent logic applied in Python/pandas):

```sql
SELECT
    PRODUCTLINE,
    ROUND(SUM(SALES), 2)          AS TOTAL_REVENUE,
    COUNT(DISTINCT ORDERNUMBER)   AS NUMBER_OF_ORDERS
FROM sales_data_sample
WHERE COUNTRY = 'USA'
GROUP BY PRODUCTLINE
ORDER BY TOTAL_REVENUE DESC
LIMIT 5;
```

Result — Top 5 Product Lines by Revenue (USA)

Rank	Product Line	Total Revenue	Number of Orders
1	Classic Cars	$1,344,638.22	71
2	Vintage Cars	$757,755.90	67
3	Motorcycles	$520,371.70	27
4	Trucks and Buses	$397,842.42	27
5	Planes	$328,432.89	22

Key Takeaways:

- **Classic Cars** is the dominant product line in the US, generating over **$1.3M** — nearly double the second-place category.
- **Vintage Cars** comes in a strong second at ~$758K, also with a high order volume (67 orders), suggesting broad customer demand.
- **Motorcycles**, **Trucks and Buses**, and **Planes** trail significantly but each still represent meaningful revenue segments.
- The top 2 product lines alone account for roughly **65%** of total US revenue, indicating a heavily concentrated product mix.

Figure 7: Claude Sonnet 4.6 Output for Task 1

Table 8: SQL Task Evaluation Results for Claude Sonnet 4.6 (Task 1)

Evaluation Factor	**Score**
SQL Correctness	5
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	25

SQL Query:

```sql
SELECT
    CUSTOMERNAME,
    ROUND(SUM(SALES), 2)        AS TOTAL_REVENUE,
    COUNT(DISTINCT ORDERNUMBER) AS NUMBER_OF_ORDERS
FROM sales_data_sample
GROUP BY CUSTOMERNAME
ORDER BY TOTAL_REVENUE DESC
LIMIT 5;
```

Result — Top 5 Customers by Total Revenue

Rank	Customer Name	Total Revenue	Number of Orders
1	Euro Shopping Channel	$912,294.11	26
2	Mini Gifts Distributors Ltd.	$654,858.06	17
3	Australian Collectors, Co.	$200,995.41	5
4	Muscle Machine Inc	$197,736.94	4
5	La Rochelle Gifts	$180,124.90	4

Key Takeaways:

- **Euro Shopping Channel** is by far the highest-value customer, generating over **$912K** across 26 orders — making them both the top spender and the most frequent buyer in the dataset.
- **Mini Gifts Distributors Ltd.** is a strong second at ~$655K with 17 orders, also demonstrating consistent, high-volume purchasing behavior.
- There is a significant **revenue cliff** between the top 2 and the remaining 3 customers — ranks 3–5 each sit around **$180K–$201K**, roughly a quarter of the #1 customer's revenue.
- Customers in ranks 3–5 each placed only **4–5 orders**, suggesting higher average order values compared to their total spend level.
- The top 2 customers alone represent a substantial concentration of revenue, which flags a **customer dependency risk** worth monitoring strategically.

Figure 8: Claude Sonnet 4.6 Output for Task 2

Table 9: SQL Task Evaluation Results for Claude Sonnet 4.6 (Task 2)

Evaluation Factor	**Score**
SQL Correctness	5
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	25

SQL Query:

```sql
SELECT
    COUNTRY,
    ROUND(SUM(SALES), 2)        AS TOTAL_REVENUE,
    COUNT(DISTINCT ORDERNUMBER) AS NUMBER_OF_ORDERS
FROM sales_data_sample
GROUP BY COUNTRY
ORDER BY TOTAL_REVENUE DESC
LIMIT 5;
```

Result — Top 5 Countries by Total Revenue

Rank	Country	Total Revenue	Number of Orders
1	USA	$3,627,982.83	112
2	Spain	$1,215,686.92	36
3	France	$1,110,916.52	37
4	Australia	$630,623.10	19
5	UK	$478,880.46	13

Key Takeaways:

- The **USA dominates** global sales by a wide margin, generating over **$3.6M** — nearly **3x more** than the second-ranked country, Spain, and representing the lion's share of total company revenue.
- **Spain and France** are closely competitive at ~$1.2M and ~$1.1M respectively, and interestingly France has slightly more orders (37 vs. 36), yet lower revenue — pointing to a marginally lower average order value per transaction.
- **Australia** ranks 4th at ~$631K with 19 orders, showing solid mid-tier performance, while the **UK** rounds out the top 5 at ~$479K across just 13 orders — the highest average revenue per order among the top 5.
- The heavy concentration of revenue in the USA signals a **geographic dependency risk** — a strategic diversification into other markets could help balance the revenue portfolio.

Figure 9: Claude Sonnet 4.6 Output for Task 3

Table 10: SQL Task Evaluation Results for Claude Sonnet 4.6 (Task 3)

Evaluation Factor	**Score**
SQL Correctness	5
Logical Accuracy	5
Query Efficiency	5
Schema Understanding	5
Usability	5
Total	25