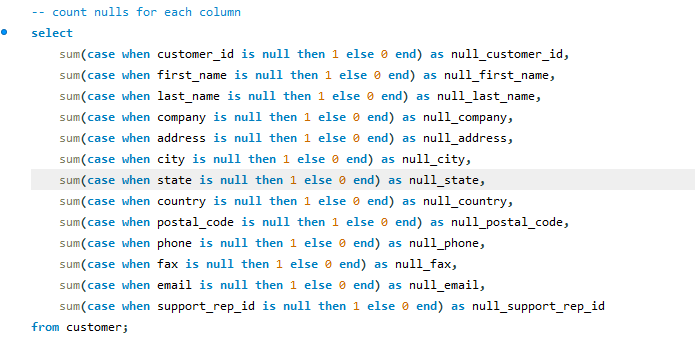
**Objective Questions**

1. Does any table have missing values or duplicates? If yes, how would you handle it?

Although, there are not any duplicate values, however we can see that there are numerous ‘NULL’ values in the data. The columns and tables with such information, along with their queries, are as follows:

*Customer Table*

Query:

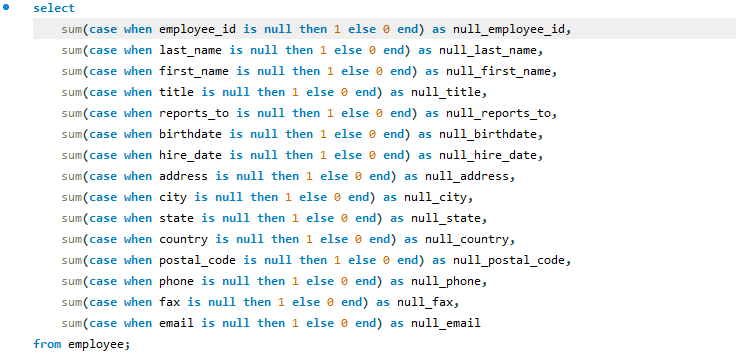


Output:



*Employee Table*

Query:

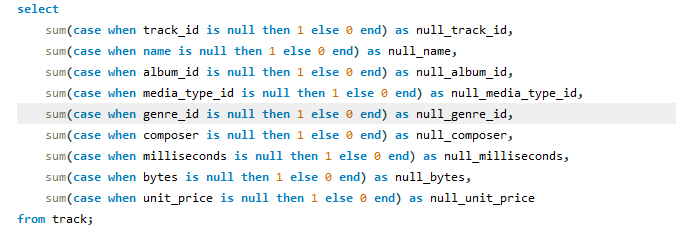
**

Output:

**

*Track Table*

Query:

**

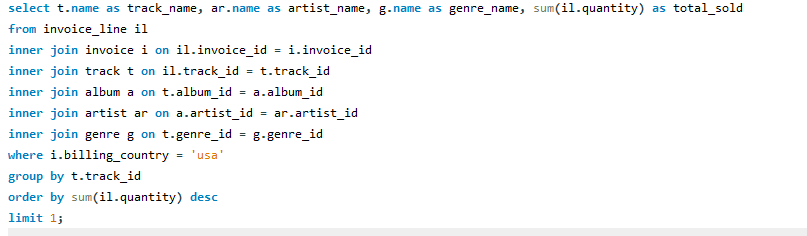
Output:

**

Handling missing data would involve updating or deleting rows with NULLs (if appropriate) and resolving duplicates by either updating or removing redundant entries. In this case, there are no duplicates and the NULL values can be handled by removing them as they constitute very little percentage of the total data which would not impact the result as a whole.

1. Find the top-selling tracks and top artist in the USA and identify their most famous genres.

Applying the following query



We get to know that (Output:)

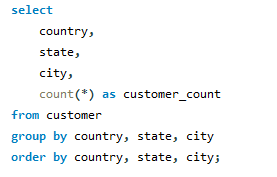


is the track name and artist name of the most sold track along with its genre name. In this query, I'm identified the top-selling track in the USA by summing the quantities sold. Firstly, by joining multiple tables to gather information about tracks, albums, artists, and genres, I linked them to invoice data. Then, I filtered the results to include only sales from the USA. After grouping the data by track, I calculated the total number sold for each track. Finally, I sorted these totals in descending order for the most sold part.

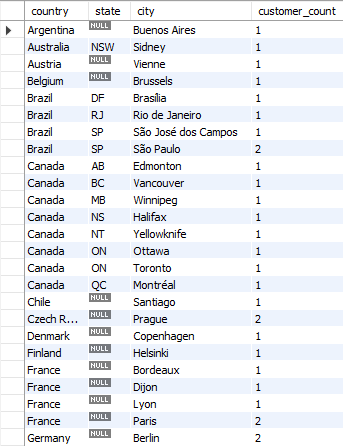
1. What is the customer demographic breakdown (age, gender, location) of Chinook's customer base?

As no birthdate or year is mentioned in the data about the customers, the age factor calculation was not plausible. However, below is the location wise count of customers:

Query:



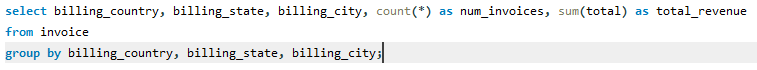
Output:

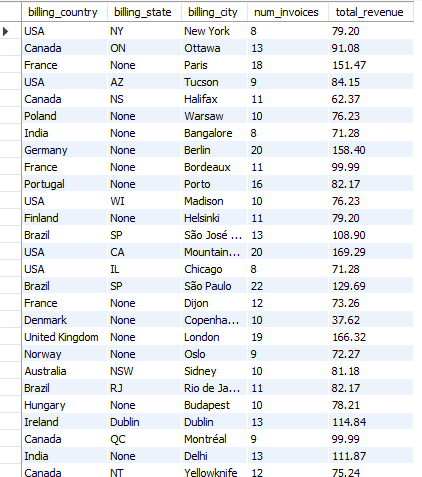
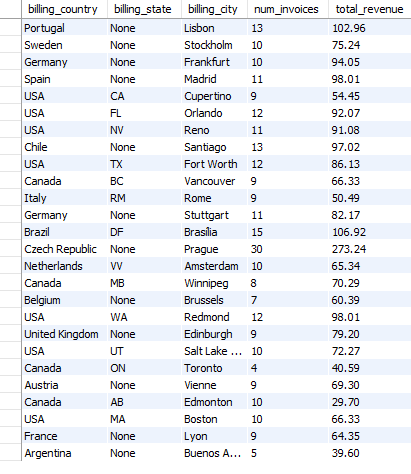
1. Calculate the total revenue and number of invoices for each country, state, and city:

I analysed total revenue and invoice count by grouping data as per country, state, and city, to identify top-performing locations.

Query:

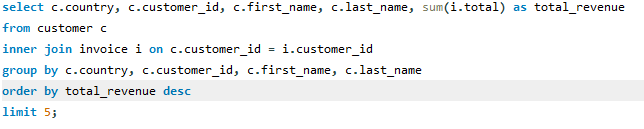


Output:

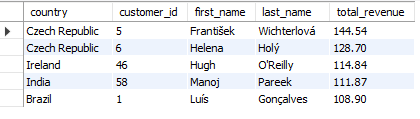
 

1. Find the top 5 customers by total revenue in each country

Query:



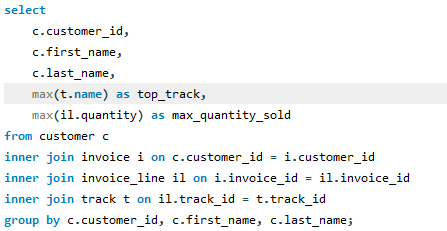
Output:



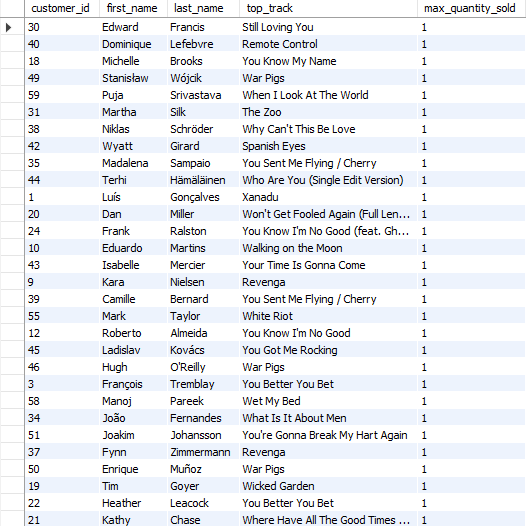
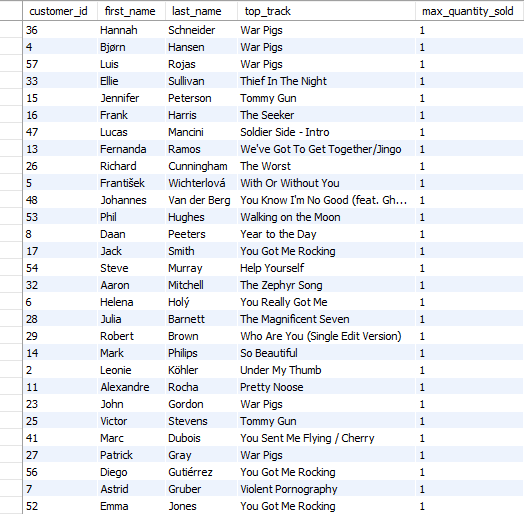
In this query, to identify the top five customers by total revenue, I, firstly, joined the customer and invoice tables using the customer ID, grouping them by country and customer details in order to sum their total purchases. Then, I ordered the results by total revenue in descending order and limited the output to highlight the most valuable customers as the question asks for top 5.

1. Identify the top-selling track for each customer

Query:



Output:

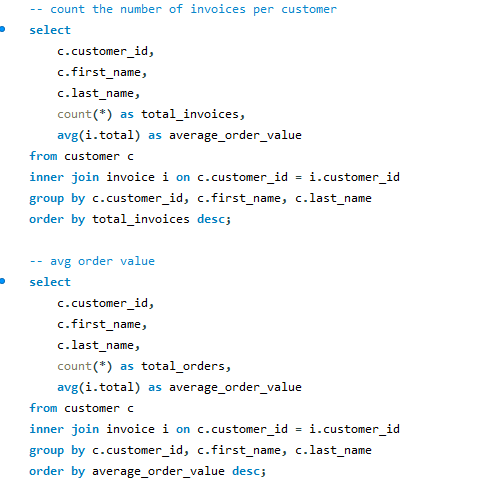
 

In this query, I found each customer's top track based on the maximum quantity sold. I joined the customer, invoice, invoice line, and track tables to gather all necessary data. Then, I grouped the results by customer ID and name, using `MAX` to determine the most sold track and the maximum quantity per customer.

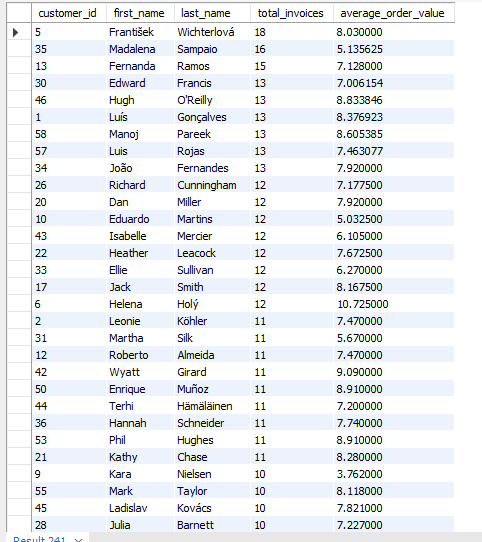
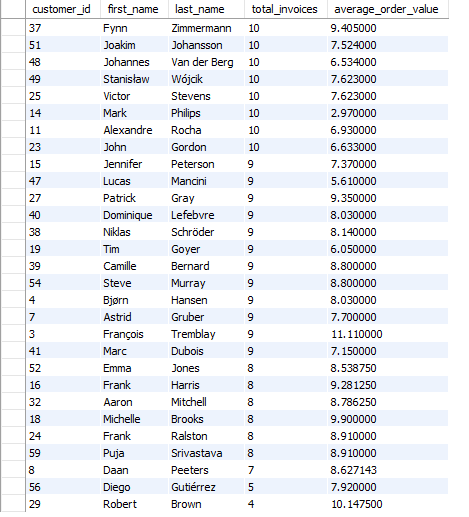
1. Are there any patterns or trends in customer purchasing behavior (e.g., frequency of purchases, preferred payment methods, average order value)?

For this calculation, a two part approach was followed.

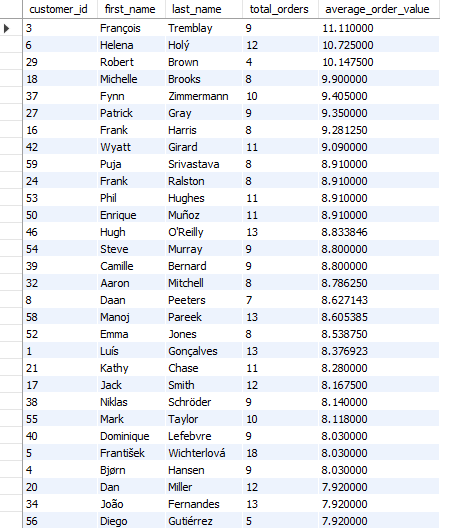
Queries:

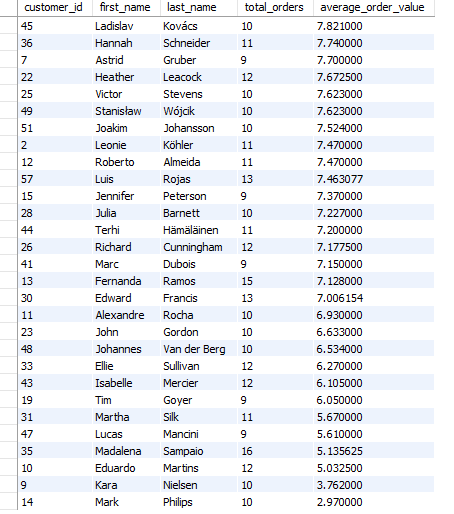


Output 1:

Output 2:

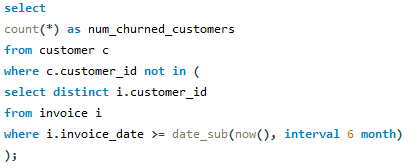




In the first query, the total number of invoices and the average order value for each customer were calculated by joining the customer and invoice tables on customer ID. Results were then grouped by customer ID and name then ordered by total invoices in descending order to highlight frequent buyers. In the second query, average order value was calculated, by grouping results as per customer ID and name, and ordering them by average order value in descending order to identify customers with higher average spending habits.

1. What is the customer churn rate?

Query:



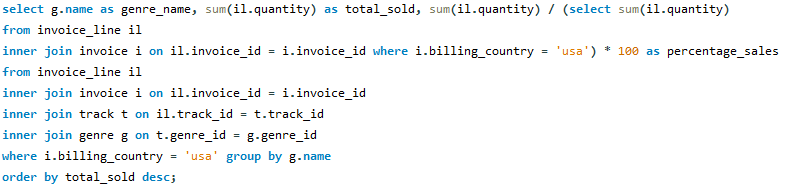
Output:



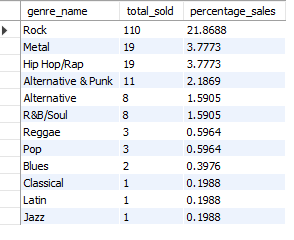
For this query, calculations are based on an assumption of 6 months that the customer stopped purchasing. I used a subquery to return distinct customer IDs from invoices dated within the last six months. Then, I excluded these active customers using a `NOT IN` clause in the customer table, and counted those who were not included.

1. Calculate the percentage of total sales contributed by each genre in the USA and identify the best-selling genres and artists.

Query:



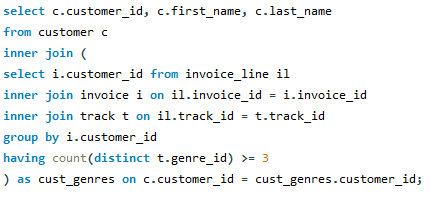
Output:



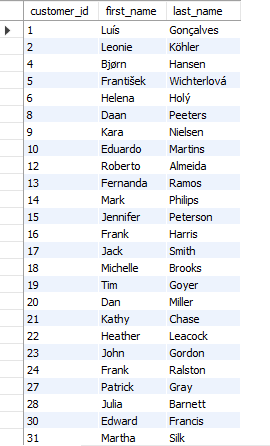
Here, by joining invoice line, invoice, track, and genre tables, I aggregated the quantity of tracks sold for each genre. The percentage sales calculation within each genre was calculated by dividing the total quantity sold by the overall quantity sold in the USA market, then multiplying by 100 to get a percentage.

1. Find customers who have purchased tracks from at least 3 different genres

Query:



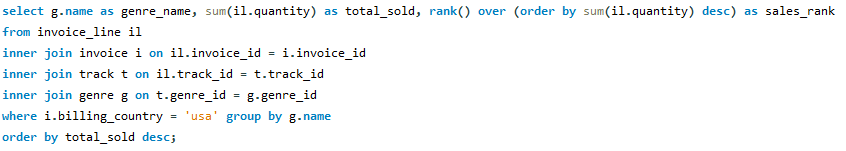
Output:

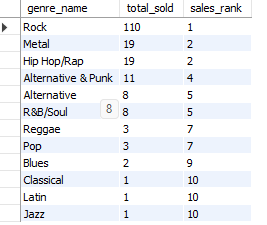
Firstly, I identified the customers who purchased tracks from at least three different genres. For this, I used a subquery to find customer IDs from invoices where tracks were joined with their genres, grouping by customer ID and filtering those who purchased from three or more distinct genres (`HAVING count(distinct t.genre\_id) >= 3`). Then, I joined this result back to the customer table based on customer ID to fetch the customer details (`customer\_id, first\_name, last\_name`).

1. Rank genres based on their sales performance in the USA

Query:



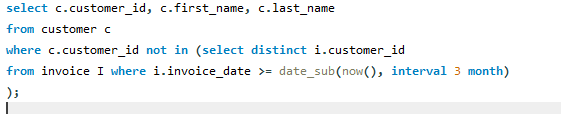
Output:



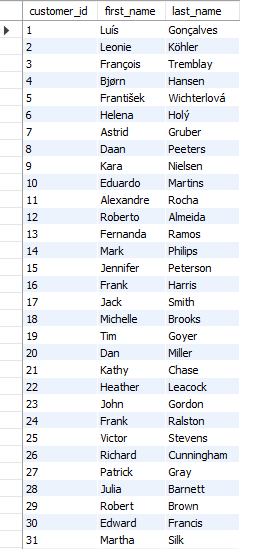
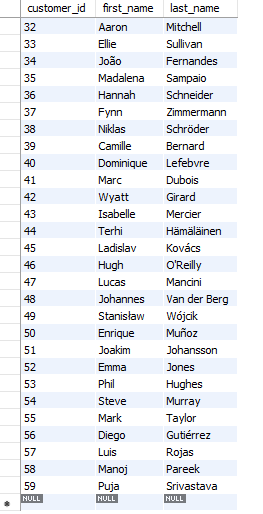
In this query, I joined invoice line, invoice, track, and genre tables, and aggregated the quantity of tracks sold for each genre (`sum(il.quantity)`). Using the `RANK()` window function, I assigned a ranking to each genre based on descending total sales (`order by sum(il.quantity) desc`).

1. Identify customers who have not made a purchase in the last 3 months.

Query:



Output:

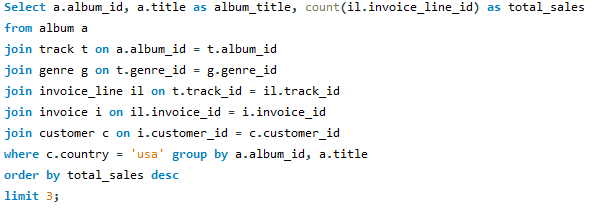
 

Here, I used a subquery to retrieve distinct customer IDs from invoices dated within the past three months. Then, I excluded these active customers using a `NOT IN` clause against the customer table, selecting those who haven't been included in the subquery result.

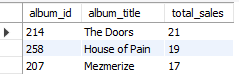
**Subjective Questions**

1. Recommend the three albums from the new record label that should be prioritised for advertising and promotion in the USA based on genre sales analysis.

Query:



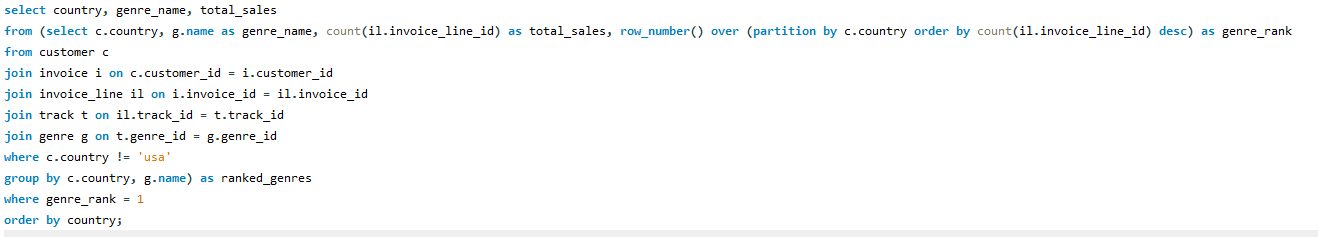
Output:



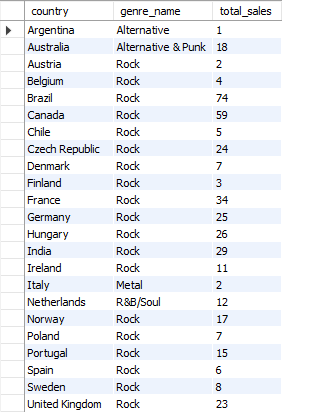
In this query, to identify the top-selling albums in the USA market I joined multiple tables to trace the sales of tracks back to their respective albums. Starting with the album table, I joined it with the track and genre tables to relate each track with its album and genre. Then, by joining with invoice\_line, invoice, and customer tables, I linked each track sale to the customer who purchased it and filtered for customers from the USA. Using `COUNT` and `GROUP BY`, I aggregated the total sales for each album and ordered them in descending order to identify the top three albums by sales volume.

1. Determine the top-selling genres in countries other than the USA and identify any commonalities or differences.

Query:

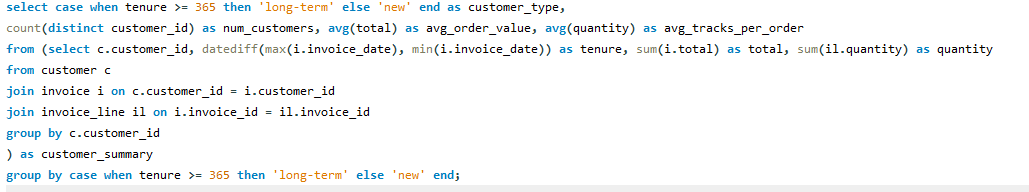


Output:



In this query, I identified the top-selling genre in each country outside of the USA. I achieved this by joining the `customer`, `invoice`, `invoice\_line`, `track`, and `genre` tables to link customer purchases with track genres. The `ROW\_NUMBER()` window function separated the results by country (`PARTITION BY c.country`) and ranked genres based on the count of invoice line items (`COUNT(il.invoice\_line\_id)`) in descending order (`ORDER BY count(il.invoice\_line\_id) DESC`). This ranking (`genre\_rank`) allowed me to filter for the top-ranked genre (`genre\_rank = 1`) in each country. The final result set contained columns for `country`, `genre\_name` (the top-selling genre), and `total\_sales` (the count of sales for that genre), ordered by country.

1. Customer Purchasing Behavior Analysis: How do the purchasing habits (frequency, basket size, spending amount) of long-term customers differ from those of new customers? What insights can these patterns provide about customer loyalty and retention strategies?

Query: 

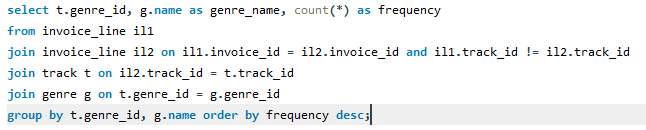
Output:



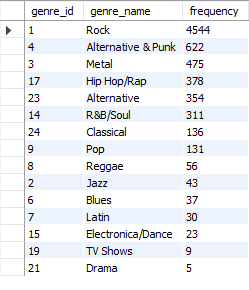
Here, I categorized customers based on their tenure with the business to analyse their purchasing behaviour differences. First, I calculated the difference between the latest and earliest invoice dates for each customer and called it tenure, along with their total expenses (`total`) and total tracks purchased (`quantity`). Then, I joined the `customer`, `invoice`, and `invoice\_line` tables and aggregated the data to the customer level. Using a subquery (`customer\_summary`), I summarized them, then, I used a `CASE` statement to organise customers as either "long-term" (tenure >= 365 days) or "new" (tenure < 365 days). Finally, I grouped the results by `customer\_type` to count the number of customers (`num\_customers`) and calculated the average order value (`avg\_order\_value`) and average tracks per order (`avg\_tracks\_per\_order`) for each customer type.

From this analysis, we get to know about customer loyalty and retention strategies. For example, understanding that long-term customers tend to have higher average order values and purchase more tracks per order suggests they are more engaged and likely to respond positively to loyalty programs or personalized offers. In contrast, new customers might benefit from targeted onboarding campaigns or introductory discounts to encourage repeat purchases and retention. Adjusting marketing strategies based on these insights can help improve customer lifetime value and improve overall customer retention rates.

1. Product Affinity Analysis: Which music genres, artists, or albums are frequently purchased together by customers? How can this information guide product recommendations and cross-selling initiatives?

Query: 

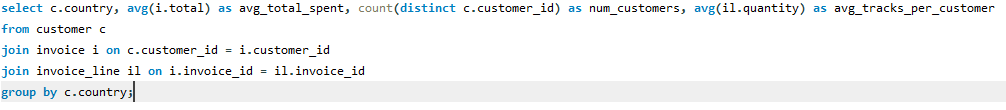
Output:



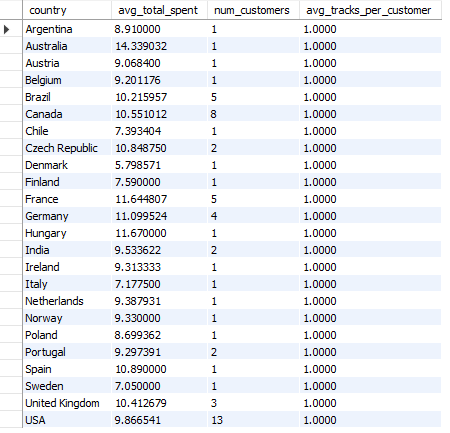
In this query, I identified the pairs of tracks (`il1` and `il2`) from the same invoice where each track belongs to a different genre (`genre\_id`). By joining the `invoice\_line` table (`il1` and `il2` on the same `invoice\_id` but different `track\_id`), I counted how often such pairs occur for each genre (`t.genre\_id`) and retrieved the corresponding genre names (`g.name`). This helped in determining which genres tended to be purchased together less frequently, highlighting likely opportunities for cross-selling between less connected genres.

This information can guide product recommendations and cross-selling initiatives by suggesting which genres might benefit from being featured together more prominently or bundled in promotions. For instance, if certain genres frequently appear together in purchases, recommending tracks from these genres when a customer purchases a track from one of them could increase the likelihood of additional sales. It could also inform strategic marketing campaigns that target customers interested in diverse music genres, potentially increasing overall sales and customer satisfaction.

1. Regional Market Analysis: Do customer purchasing behaviors and churn rates vary across different geographic regions or store locations? How might these correlate with local demographic or economic factors?

Query: 

Output:



First, I joined the `customer`, `invoice`, and `invoice\_line` tables based on their relationships. This allowed me to calculate aggregate data per country: average total spent (`avg\_total\_spent`), number of unique customers (`num\_customers`), and average tracks per customer (`avg\_tracks\_per\_customer`).

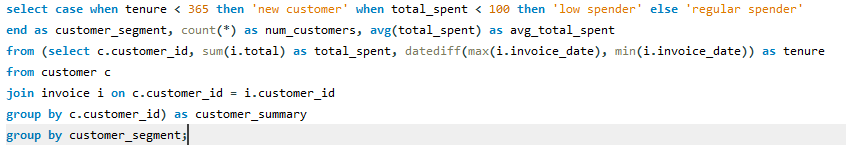
Based on the output which shows average total spent per customer, number of customers, and average tracks per customer across different countries, it can be seen that:

1. The average total spent per customer varies across different countries, ranging from a low of 5.05 in Denmark to a high of 14.34 in Australia. This variation suggests that customer spending habits differ significantly based on geographic location.

2. The number of customers also varies, with larger markets like the USA having 13 customers compared to smaller markets like Denmark or Hungary, which have only one customer each.

1. Customer Risk Profiling: Based on customer profiles (age, gender, location, purchase history), which customer segments are more likely to churn or pose a higher risk of reduced spending? What factors contribute to this risk?

Query:



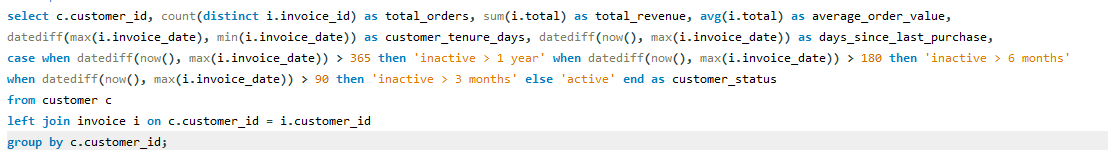
Output:



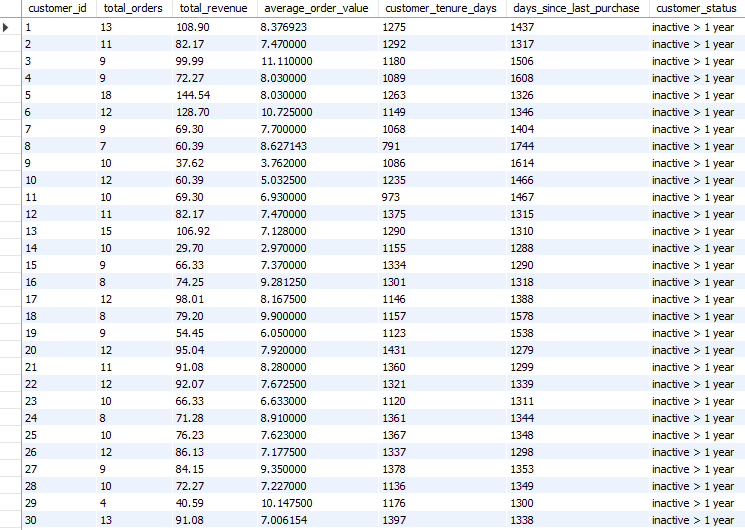
For this query, I calculated the total spent (`total\_spent`) and tenure (difference between the latest and earliest invoice dates) for each customer by joining the `customer` and `invoice` tables. Then, within the outer query, I used a `CASE` statement to assign customers to segments: 'new customer' for those with a tenure less than 365 days, 'low spender' for those whose total spent is less than 100, and 'regular spender' for all others. Grouping by `customer\_segment`, I counted the number of customers (`num\_customers`) and calculated the average total spent (`avg\_total\_spent`) for each part.

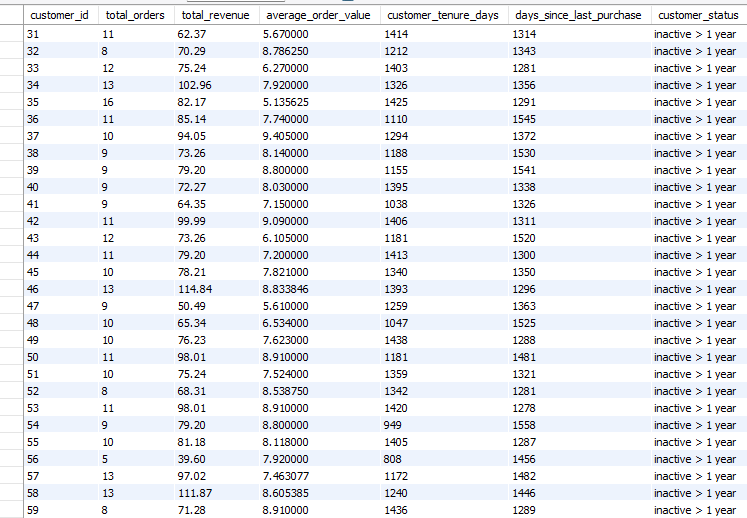
Based on this separation, low spenders are the largest segment in terms of customer count, indicating a potential risk of reduced spending. Factors contributing to this risk could include demographics (like age and income level), purchasing history (frequency and average order value), and possibly external economic factors affecting disposable income.

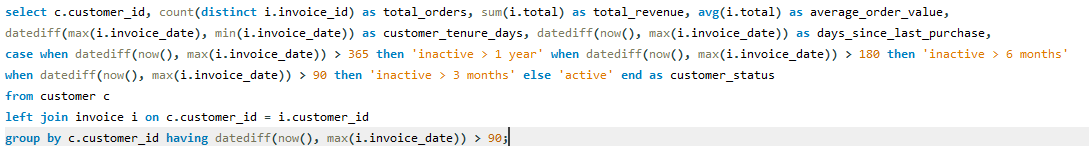
1. Customer Lifetime Value Modeling: How can you leverage customer data (tenure, purchase history, engagement) to predict the lifetime value of different customer segments? This could inform targeted marketing and loyalty program strategies. Can you observe any common characteristics or purchase patterns among customers who have stopped purchasing?

Query1: 

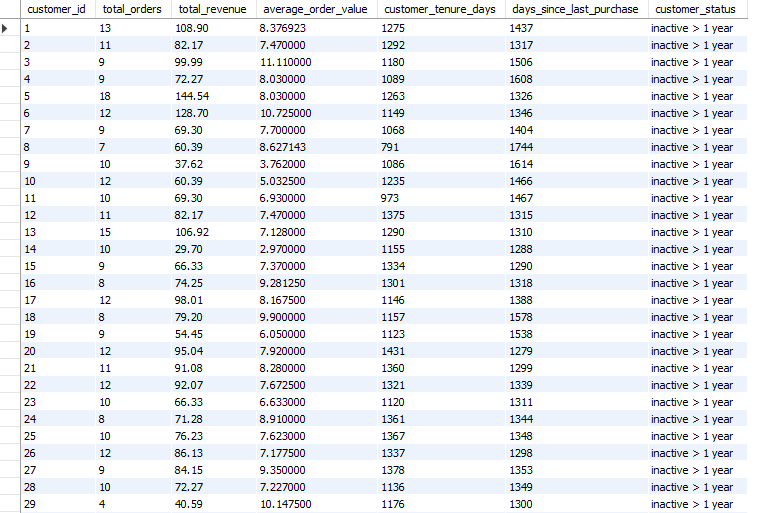
Output1:

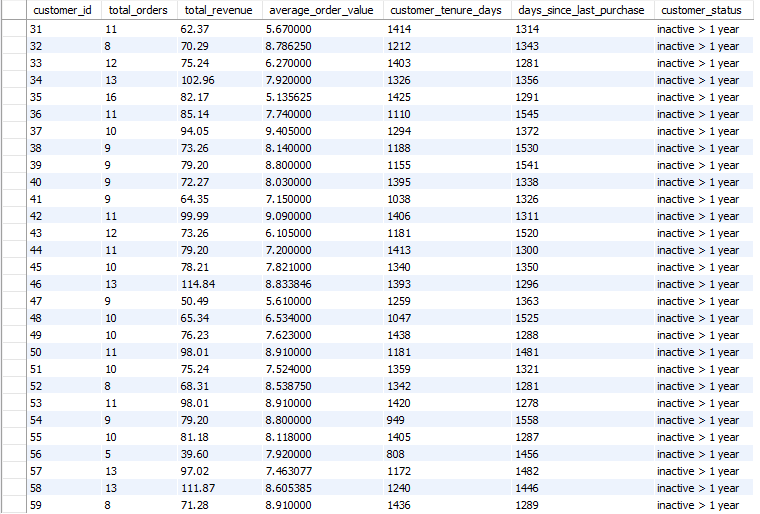




Query2: 

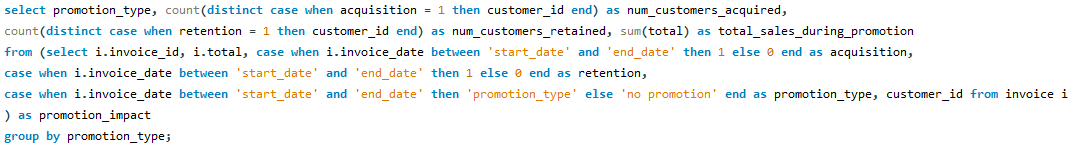
Output2:





First, analysed customers' orders, revenue, and tenure. Through this, I categorized customers based on their activity status—whether they were active or inactive over specific timeframes like 3 months, 6 months, or 1 year. Additionally, I identified at-risk customers by filtering out those who hadn't made a purchase in the last 90 days, which provided valuable insights for churn analysis. The second query was similar to Step 1 but with a HAVING clause to filter out customers who are considered at-risk of churn.

1. If data on promotional campaigns (discounts, events, email marketing) is available, how could you measure their impact on customer acquisition, retention, and overall sales?

Query: 

Output:



Here, first, I identified invoices within specified promotion periods (`start\_date` to `end\_date`) and categorized them based on whether they contributed to customer acquisition (`acquisition = 1`) or retention (`retention = 1`). Each invoice was labeled with a `promotion\_type` to differentiate between different promotional efforts. In the outer query, I grouped the results by `promotion\_type` to summarize the impact: the number of customers acquired (`num\_customers\_acquired`), the number retained (`num\_customers\_retained`), and the total sales (`total\_sales\_during\_promotion`) generated during each promotion type.

The output shows the impact for invoices during non-promotional periods (`no promotion`), where no specific promotion was identified within the specified date range.

1. How would you approach this problem, if the objective and subjective questions weren't given?

In approaching the analysis without specific objectives or questions provided, I would have begun by understanding the business context and the structure of the available data schema. This involves closely examining key tables such as customers, invoices, tracks, and genres to grasp their relationships and relevance to business operations. Then, I would have researched the data to uncover trends, anomalies, and potential insights. From this, theories naturally emerge regarding customer behaviour, sales patterns, and market preferences.

Using SQL queries and statistical methods, I would have authenticated these theories, testing assumptions about customer segments, purchasing habits, and genre preferences. This approach directs strategic decisions, allowing the development of targeted marketing strategies, customer retention steps, and improvements of operations based on information gathered by all the provided data.

1. How can you alter the "Albums" table to add a new column named "ReleaseYear" of type INTEGER to store the release year of each album?

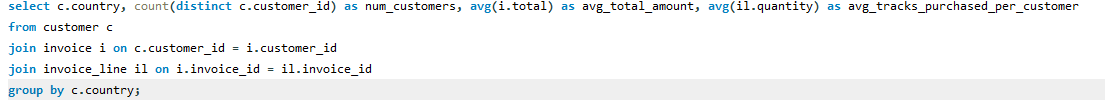
Query:



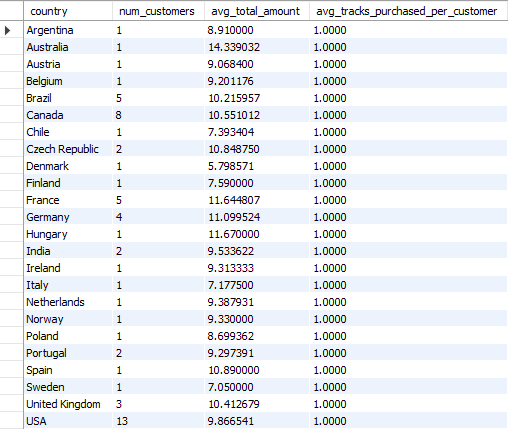
Output: (cannot be shown)

To alter the "Albums" table and add a new column named "ReleaseYear" of type INTEGER to store the release year of each album, I typed the ALTER TABLE statement for the "Albums" table, specifying the new column name "ReleaseYear" with the INTEGER data type.

1. Chinook is interested in understanding the purchasing behavior of customers based on their geographical location. They want to know the average total amount spent by customers from each country, along with the number of customers and the average number of tracks purchased per customer. Write an SQL query to provide this information.

Query: 

Output:



To provide Chinook with understanding in customer purchasing behaviour across different countries, my SQL query joins the necessary tables—Customer, Invoice, and Invoice Line— to combine data based on geographic location. The query calculated the average total amount spent per customer, the number of distinct customers, and the average number of tracks purchased per customer for each country.