***Chinook SQL Project***

**Objective Questions**

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

* Yes, some tables have missing values in optional columns. The **customer** table has NULLs for company, state, postal\_code, phone, and fax. In the **track** table, the composer column contains around 978 NULLs. These fields are not required for analysis and can be left as NULL or marked "Not Available" during reporting. Tables such as invoice and invoice\_line contain no missing values.
* For duplication, I basically looked at the relevant tables. A few rows in **invoice\_line** contained the same invoice\_id and track\_id combinations, which is understandable given that a user may purchase the same track several times. Because of its composite primary key, playlist\_track did not include any duplicates. Other tables contain stringent primary keys, which prevents duplicates. As a result, no rows need deletion or cleaning.

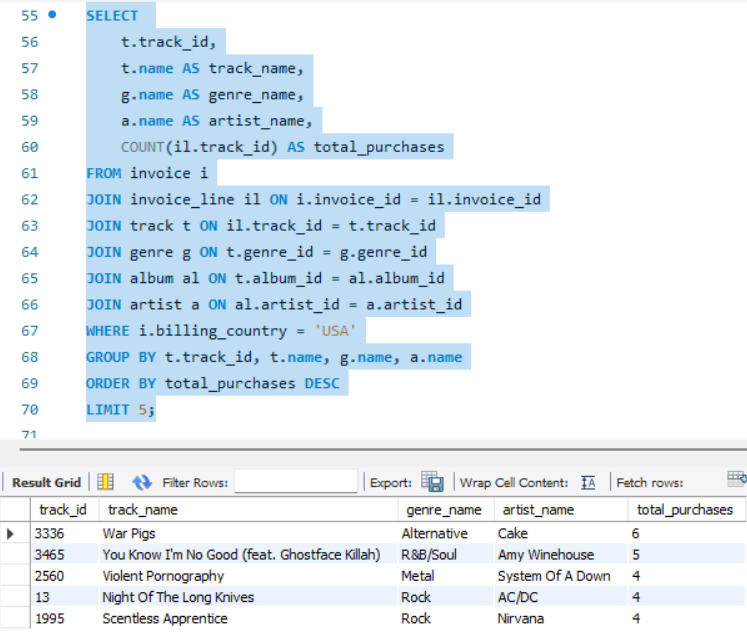
A screenshot of a computer

AI-generated content may be incorrect.

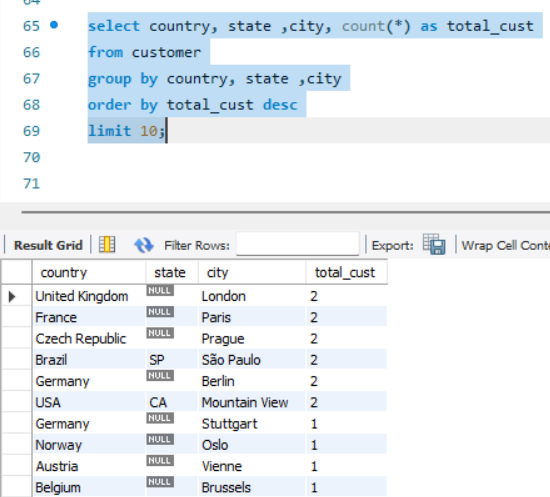
A screenshot of a computer

AI-generated content may be incorrect.

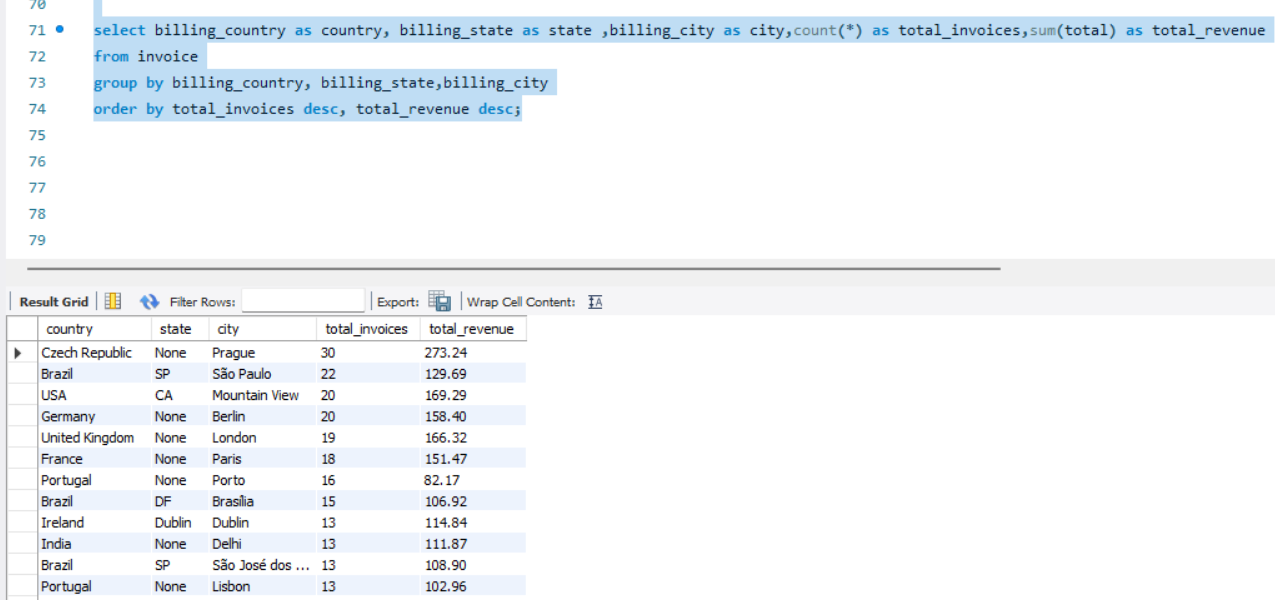
1. **Find the top-selling tracks and top artist in the USA and identify their most famous genres.**
2. To find the top-selling tracks in the United States, I filtered invoice data with the billing country 'USA'. Then I connected the invoice\_line and track tables to determine how frequently each track was purchased. I also used the album and artist tables to locate the artist, and the genre table to obtain the genre name. Cake's "War Pigs" was the best-selling track overall, with Alternative, R&B/Soul, Rock, and Metal being the most popular genres purchased:



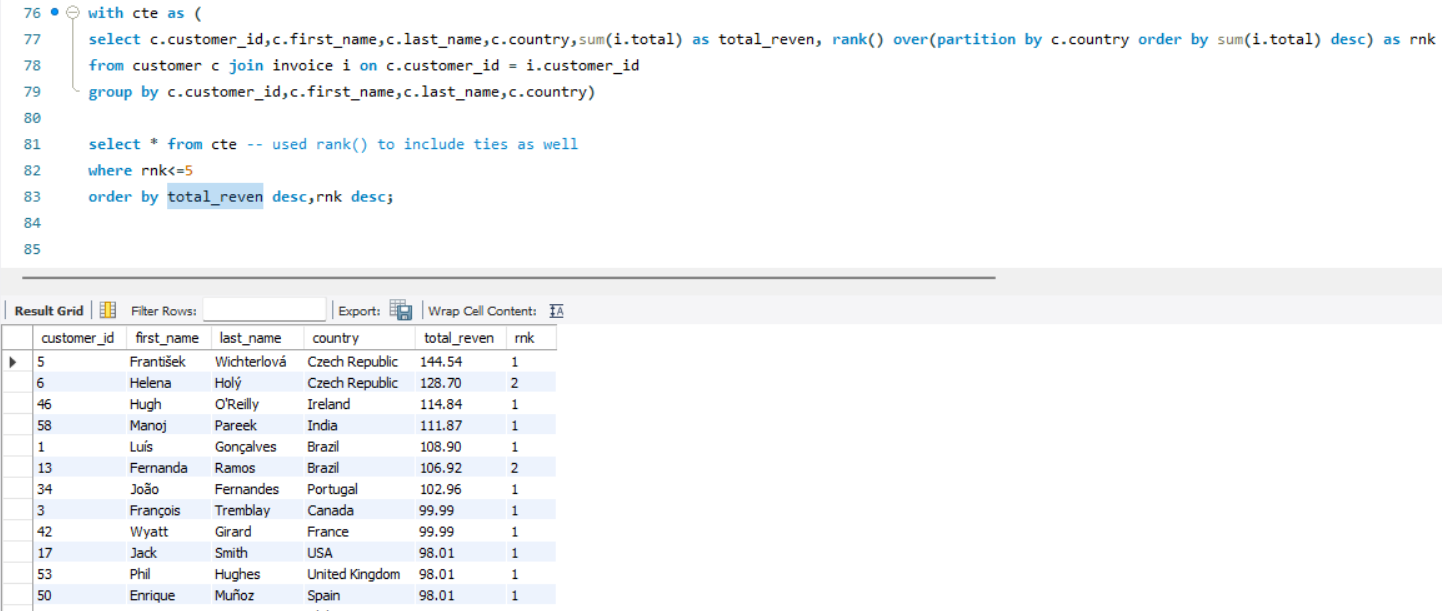
1. **What is the customer demographic breakdown (age, gender, location) of Chinook's customer base?**
2. The Chinook database doesn't have columns for age and gender, likewise I was unable to use these in the analysis. However, based on the available data, I used the location fields (country, state, and city) to determine where the customers were from. Most consumers come from the United Kingdom, France, Brazil, and the United States. Cities such as London, Paris, São Paulo, and Mountain View attract more customers. Some records had no state information, but I included them because they still provide useful location insights. This gives us an honest idea of where Chinook's customers dispersed.

****

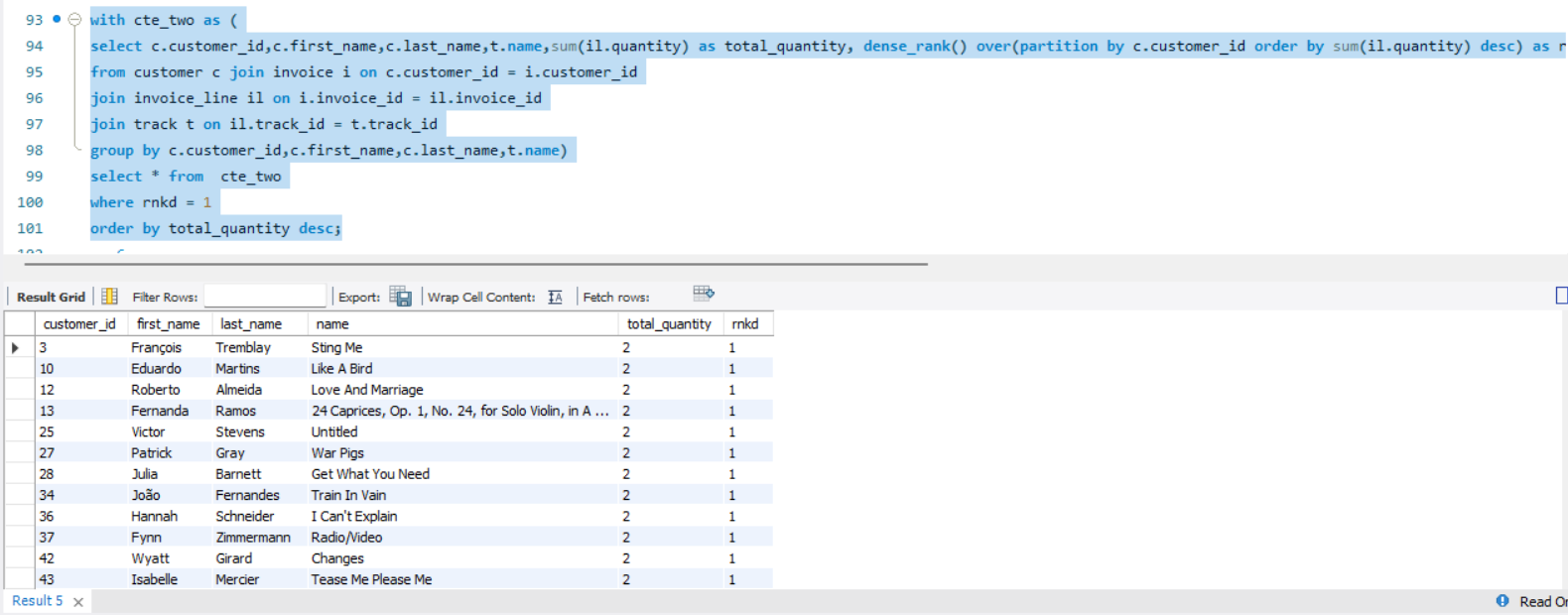
1. **Calculate the total revenue and number of invoices for each country, state, and city**
2. For calculating total income and number of invoices by location, I grouped information from the **invoice** table by country, state, and city. I next counted the invoices and totalled the revenue for each group. Cities such as Prague, São Paulo, Mountain View, and Berlin had the most invoices and revenue.



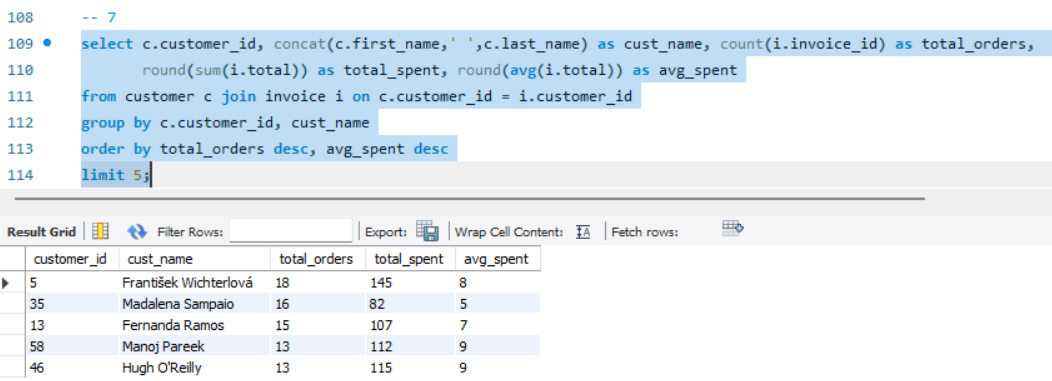
1. **Find the top 5 customers by total revenue in each country**
2. I used the **RANK () function** (to show with ties) to identify the top five customers by revenue in each country. The output is sorted by total revenue, with the biggest spenders at the top. The ranking function already handles country grouping with the PARTITION BY logic.



1. **Identify the top-selling track for each customer**
2. To find the best-selling track for every customer, I calculated the total quantity of each track purchased by each customer using SUM (quantity). Then I used the RANK () function with PARTITION BY customer\_id to determine which track had the largest number per customer. In most cases, customers purchased each track just once, so numerous tracks were tied as their favorite. RANK() is used in the query to return all such top tracks per customer, including ties. This helps to understand each customer's most favorite music.

****

1. **Are there any patterns or trends in customer purchasing behavior (e.g., frequency of purchases, preferred payment methods, average order value)?**
2. To analyse purchasing patterns, I totalled the number of purchases, total amount spent, and average spend per invoice for each customer. Customers like František Wichterlová and Hugh O'Reilly made frequent purchases (13+ orders) and spent an average of $7-$9, which shows loyalty and high engagement. While payment method data was not available in the dataset, we can clearly observe that a few customers contribute significantly more than others in terms of both quantity and value of purchases.



1. **What is the customer churn rate?**
2. To calculate the churn rate, I assumed that any customer who has not made a purchase in the previous 90 days from the most recent invoice date is considered churned. Using this logic, I first identified the most recent invoice date as 2020-12-30 and then excluded customers who had made no purchases after 2020-10-01. Out of 59 total customers, 22 turned out to be inactive, resulting in a 37.29% churn rate. This shows the need for potential re-engagement strategies to retain quite a bit of Chinook's customer base.
3. Since Chinook is NOT subscription-based.  
   It’s a small sample sales database, and most customers made 1–2 purchases total years apart.

So, churn will naturally be very high (80–95%).

1. Churn definitions can vary—30, 60, 90, or 180 days depending on purchase frequency—so churn analysis is often paired with cohort analysis or RFM scoring.
2. If we increase the inactivity threshold from **90 days → 180 days**, then

More customers will be considered **active**, because the activity window is now larger. Fewer customers fall into the “inactive” category.

Therefore, **churn percentage naturally decreases**.

In the Chinook dataset (59 customers), many customers have purchased *earlier than 90 days* but *within 180 days*, so they get **reclassified as active**.

1. Latest invoice date = 2020-12-30

Activity threshold = 2020-12-30 – 180 days = 2020-07-03

Active customers = customers with purchases on or after 2020-07-03  
Churned customers = customers with no purchases since 2020-07-03

Below attached output shows the churn rate based on 90 days and 180 days

Inactivity threshold period.

For 90-days window period, churn rate = 37.29%

For 180-days window period, churn rate = 27.12%

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

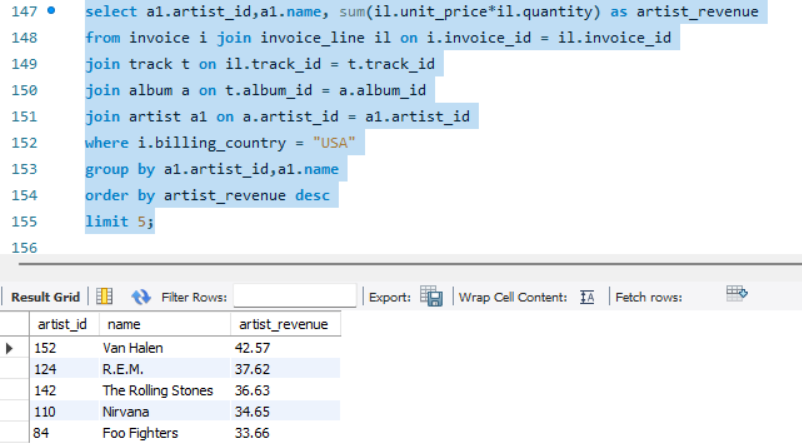
AI-generated content may be incorrect.

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

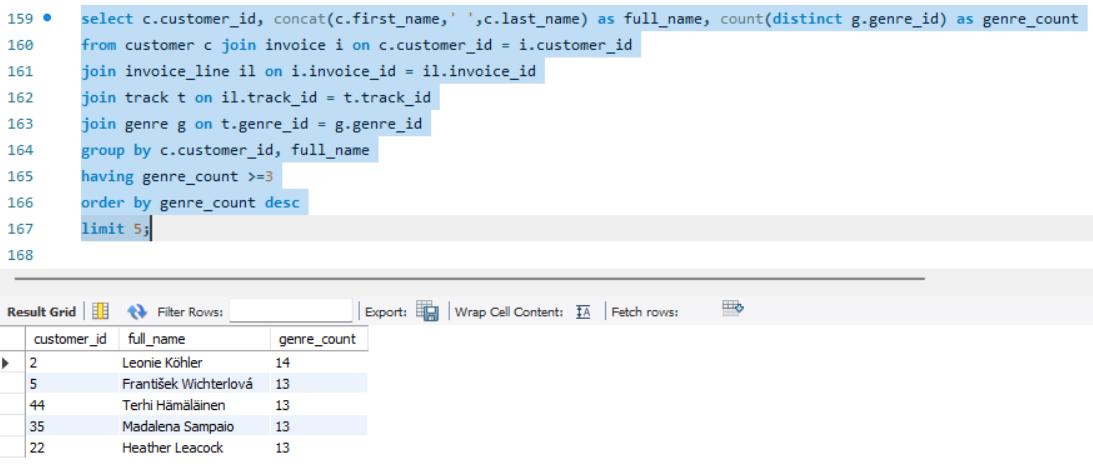
* To calculate the percentage of total sales contributed by each genre in the United States, I initially created a temporary table containing track-level revenue aggregated by genre and filtered just for U.S. invoices. I then calculated each genre's contribution as a percentage of overall US income. The data show that Rock leads with 53.38%, followed by Alternative & Punk, Metal, R&B/Soul, and Blues.
* To find the best-selling artists, I combined track, album, and artist tables for calculating revenue per artist. Van Halen, R.E.M., The Rolling Stones, Nirvana, and the Foo Fighters were the top revenue-generating appears. suggesting a strong customer preference for rock and alternative bands in the United States market.

A screenshot of a computer

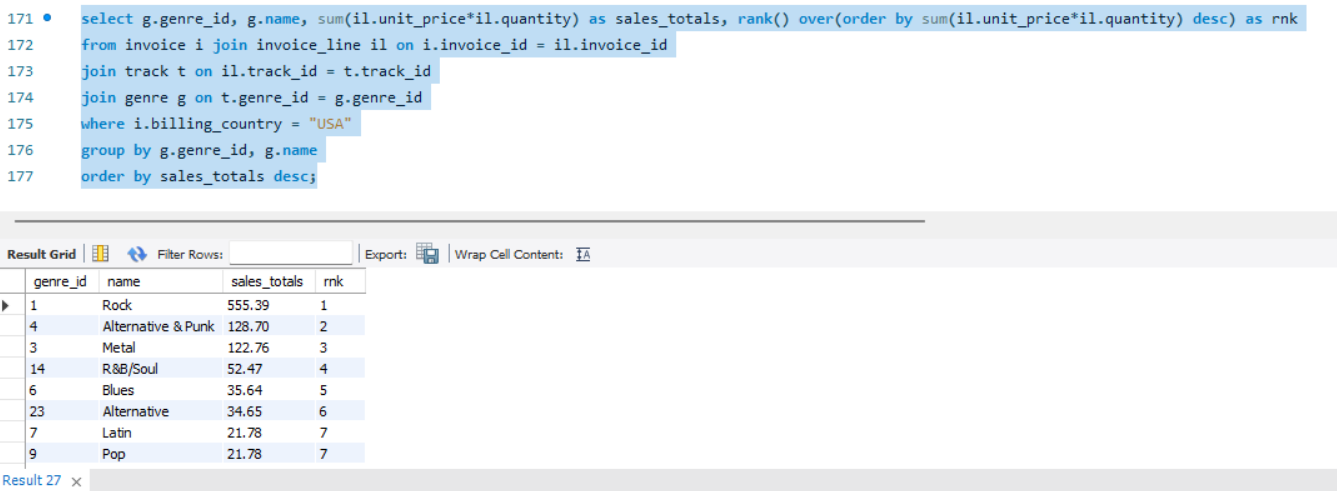
AI-generated content may be incorrect.



1. **Find customers who have purchased tracks from at least 3 different genres**
2. To identify customers who purchased tracks from a diverse range of music genres, I joined the customer, invoice, and genre-related tables and counted the number of distinct genres purchased per customer. I then filtered for those who had bought music from at least three genres. The results show that customers like Leonie Köhler, František Wichterlová, and Terhi Hämäläinen have explored more than 13 different genres.



1. **Rank genres based on their sales performance in the USA**
2. To rank music genres based on their sales performance in the USA, I joined track and genre information with invoice data and filtered by billing country. I then calculated total revenue per genre and used a window function to assign ranks based on descending revenue. The results show that Rock is by far the top-performing genre, followed by Alternative & Punk, Metal, and R&B/Soul. These genres clearly dominate the USA market in terms of revenue contribution.



1. **Identify customers who have not made a purchase in the last 3 months**
2. To identify customers who have not made a purchase in the last 3 months, I first calculated the latest invoice date, which was 2020-12-30. Then I defined active customers as those who purchased on or after 2020-10-01 and excluded them from the full customer list. This gave me the set of customers considered inactive or at risk of churn. Customers like Luís Gonçalves, François Tremblay, and Björn Hansen were among those who had not purchased recently.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

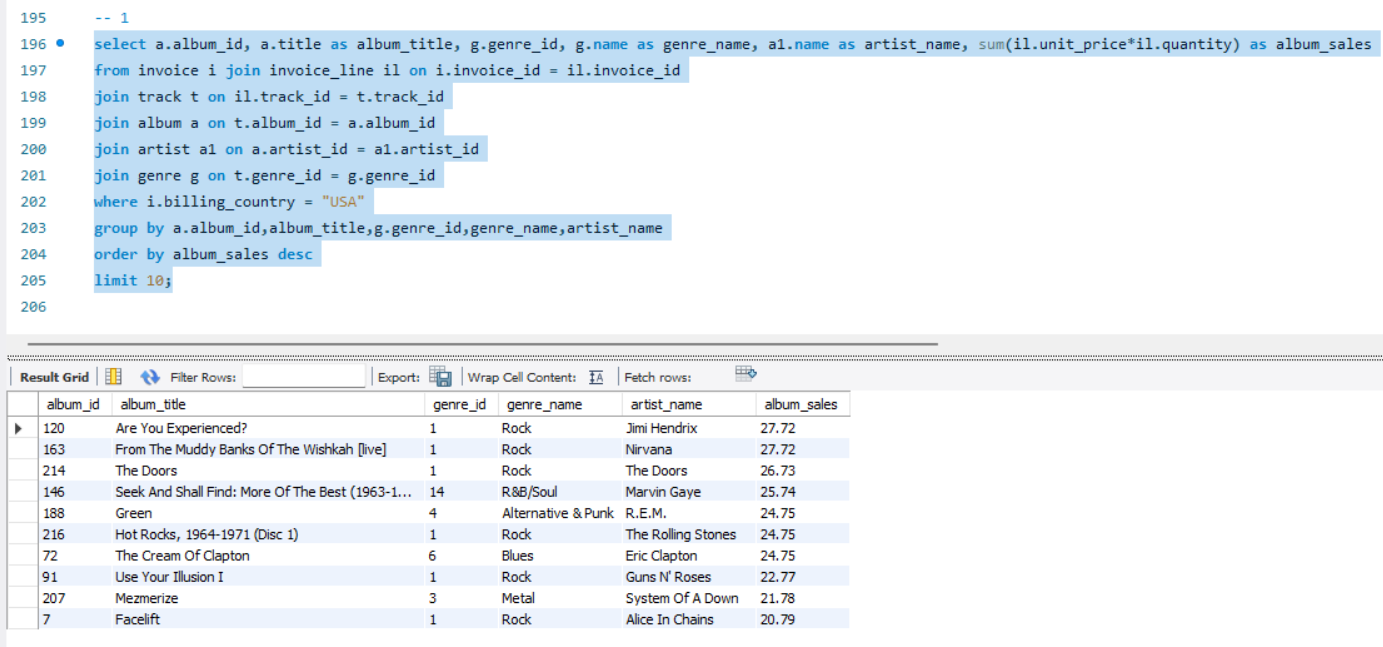
**Subjective Questions**

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.**
2. **Approach:**

To find the top albums to promote in the United States, I first checked overall sales by genre and observed that Rock, Alternative & Punk, and R&B/Soul were the highest-performing genres. I next looked at album-level revenue filtered by genre and ranked albums based on their sales performance in the United States.

1. **Work:**

* Joined invoice, invoice\_line, track, album, artist, and genre tables
* Filtered for USA customers (billing\_country = 'USA')
* Aggregated sales by album
* Included genre and artist details in the output
* Sorted the results by total album revenue
* Picked top 3 albums from the top genres



1. **Insights:**

* Album sales in the United States are dominated by the genres rock, alternative, and R&B/soul.
* All top 5 albums belonged to these genres.
* Several artists, including Jimi Hendrix, Nirvana, and The Doors, had significant revenue streams.

1. **Recommendations:**

Based on both genre-level and album-level outcomes, the following albums should be prioritized for advertising in the USA:

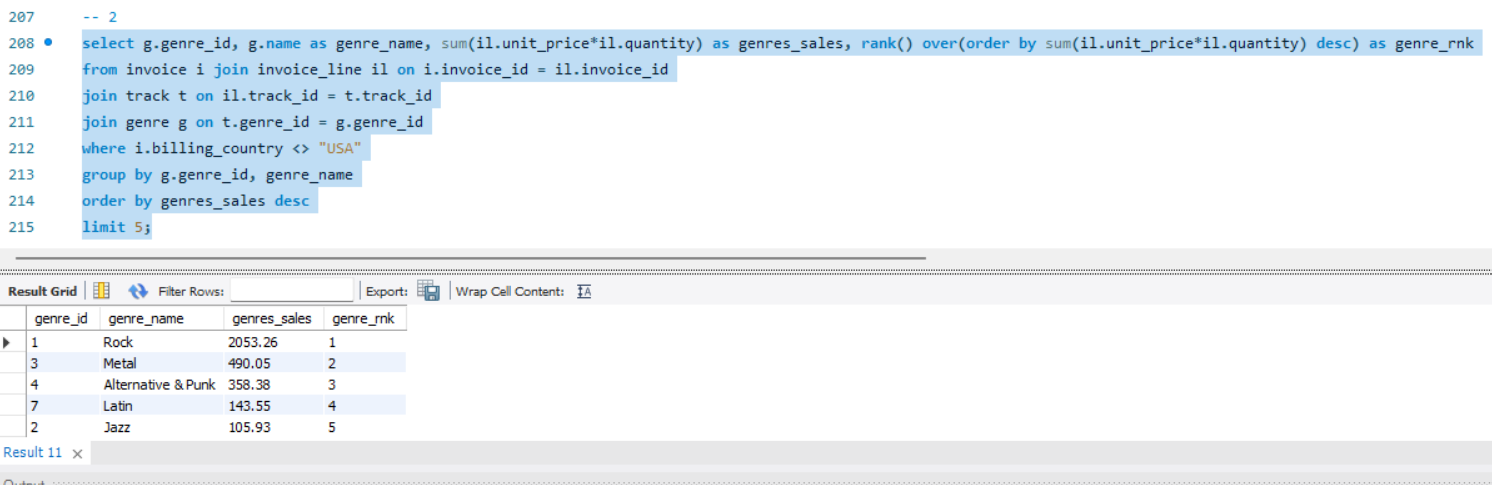
* **Are You Experienced**? – Jimi Hendrix (Rock)
* **From The Muddy Banks of the Wishkah** – Nirvana (Rock)
* **The Doors** – The Doors (Rock)

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

To identify regional listening preferences, I looked at total music sales by genre in countries other than the United States. This helps determine whether the global audience is similar to the US market or prefers different genres. The goal was to discover overlapping and distinct patterns.

1. **Work:**

* I joined the invoice, invoice\_line, track, and genre tables.
* Used SUM (unit\_price \* quantity) to measure genre-level sales revenue.
* Filtered the data using billing\_country <> 'USA'.
* Grouped results by genre\_id and genre\_name.
* Applied RANK () to identify the top 5 globally.
* Ordered genres by total sales descending.



1. **Insights:**

* Rock remains the most popular genre worldwide, just as it is in the United States.
* Metal, Alternative, and Punk are also popular in both regions.
* However, while Latin and Jazz are in the global top five, they are not among the top genres in the United States
* R&B/Soul and Blues, which did well in the United States, were less popular worldwide

1. **Recommendations:**

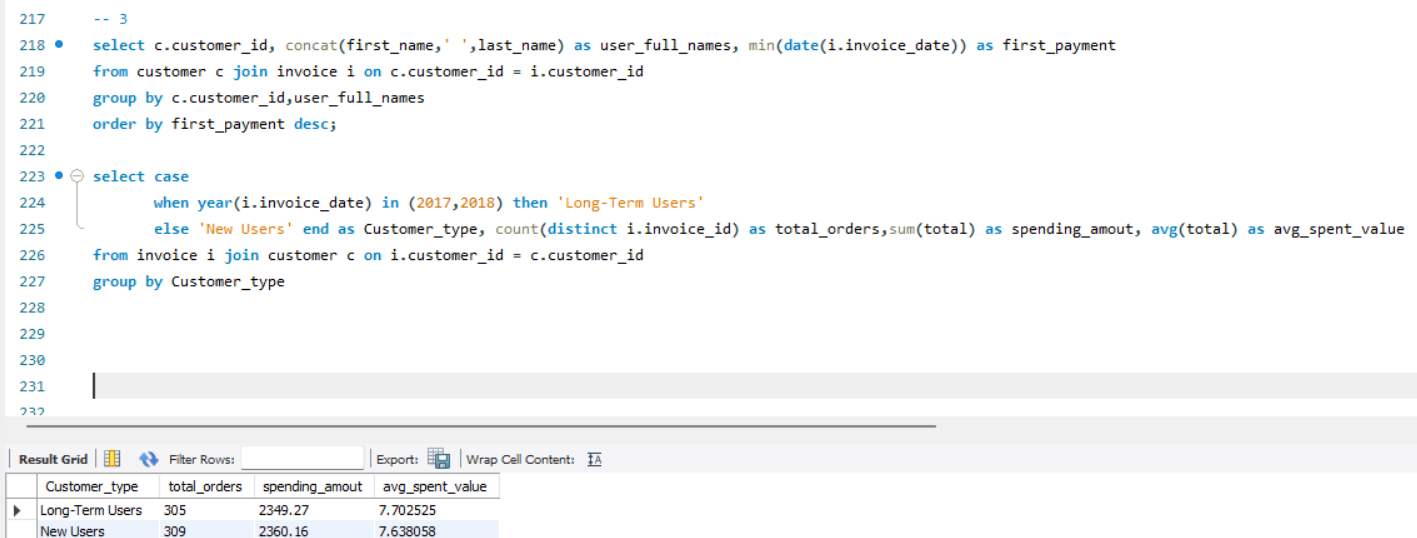
* Rock, Metal, and Alternative & Punk do well both in the United States and abroad, so they should be prioritized for global marketing.
* To increase reach in non-US regions, consider promoting Latin and Jazz genres in culturally relevant areas.
* Genre-based personalized recommendations and localized playlists can help increase engagement and revenue across multiple regions.

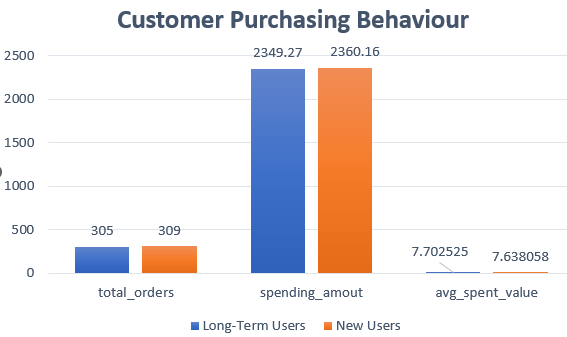
1. **Customer Purchasing Behaviour Analysis: How do the purchasing habits (frequency, basket size, spending amount) of long-term customers differ from those of new customers? What insights can these patterns provide about customer loyalty and retention strategies?**
2. **Approach:**

To compare purchasing patterns, I divided customers as long-term (first purchase in 2017-2018) or new (joined in 2019-2020) based on invoice history. I then measured total orders placed, total amount spent, and average order value for each category.

1. **Work:**

To compare the two groups, I first calculated each customer's first purchase date using MIN (invoice\_date) and grouped by customer. This helped these find out when they joined the platform. Based on the observed data, I classified customers who made their first purchase in 2017 or 2018 as Long-Term Users, while those who joined later as New Users. Then, using a CASE WHEN condition within the SELECT query, I categorized each user properly. I combined the customer and invoice tables, organized the data by customer type, and estimated three key metrics: total number of orders, total amount spent, and average spend per order.





1. **Insights:**

* Long-term users made 305 orders and spent ₹2349.27.
* New users placed 309 orders totalling ₹2360.16.
* The average spend per order was nearly equal (₹7.70 vs ₹7.64).
* This suggests that both segments behave similarly, with high engagement across the board.

1. **Recommendations:**

* Because new users are performing well, prioritize efficient onboarding and retention campaigns to convert them into loyal customers.
* To keep long-term users engaged, consider offering reward programs or exclusive content.
* Overall, both groups are in good condition, and efforts should be made to ensure fair development.

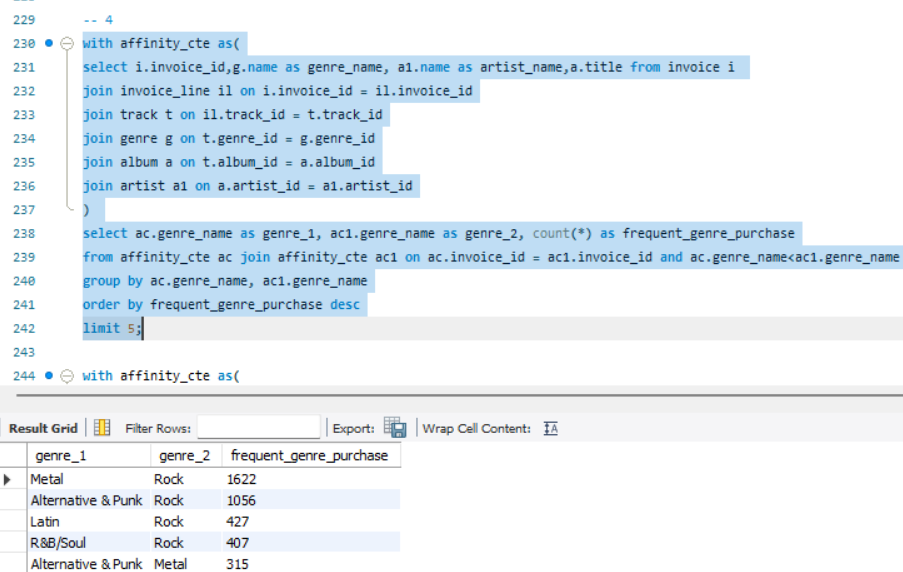
1. **Product Affinity Analysis: Which music genres, artists, or albums are frequently purchased together by customers? How can this information guide product recommendations and cross-selling initiatives?**
2. **Approach:**

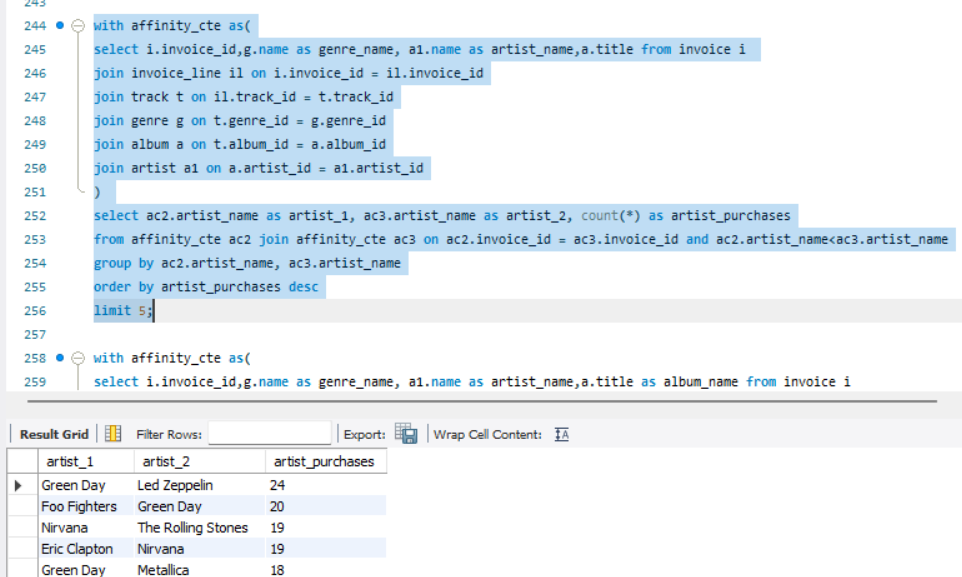
To identify cross-selling opportunities and improve product recommendations, I did a product affinity analysis by examining consumer purchase behaviours. I examined combinations of music genres, artists, and albums that were frequently purchased together on the same invoice. These combinations reflect people' natural listening preferences as well as genre connections.

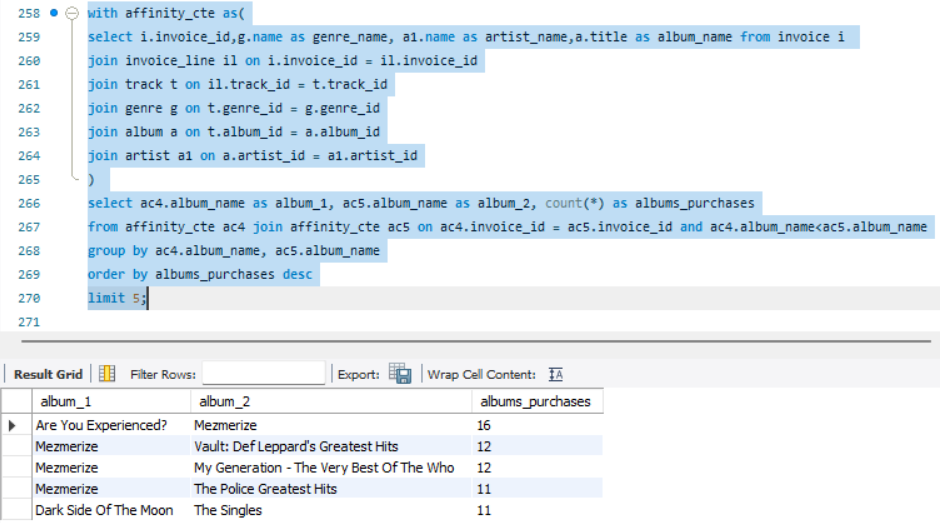
1. **Work:**

* I joined the invoice, invoice\_line, track, album, artist, and genre tables to extract invoice-level purchase details.
* Using a Common Table Expression (CTE), I prepared a unified view with invoice\_id, genre\_name, artist\_name, and album\_title.
* I then performed self-joins on the CTE by invoice\_id to identify:
* Genres purchased together (genre\_name < genre\_name)
* Artists purchased together (artist\_name < artist\_name)
* Albums purchased together (album\_title < album\_title)

1. **Insights:**



****

****

* Rock is the most popular genre, and it is usually purchased with others such as Metal, Alternative, and Punk.
* Green Day occurs in three of the top five artist combinations, indicating a strong a connection for both classic and alternative rock fan groups.
* The album Mezmerize is featured in four of the top five combinations, demonstrating high crossover appeal.

1. **Recommendations:**

* Use the top genre and artist pairings Metal + Rock, Green Day + Led Zeppelin to recommend relevant add-on tracks or albums during checkout. This increases basket size without being junk.
* Create targeted offers or discounts by combining popular genre pairs such as Rock, Alternative, and Punk. Run them during weekends or significant music releases to increase conversion.
* Provide genre/artist/album pairing frequency into the recommendation model. Prioritize those pairings in home page carousels or "You may also like" sections to increase click-through rates.

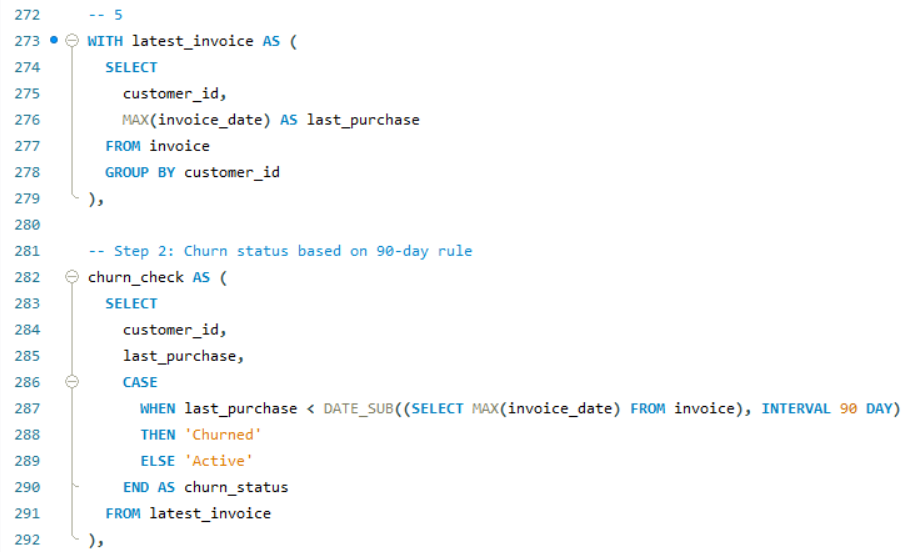
1. **Regional Market Analysis: Do customer purchasing behaviours and churn rates vary across different geographic regions or store locations? How might these correlate with local demographic or economic factors?**
2. **Approach:**

To better understand customer behaviour and turnover fluctuations between geographies, I conducted a country-level analysis utilizing billing data. The goal was to evaluate revenue generation, order volume, and customer retention across countries and determine which locations have healthy engagement vs excessive turnover.

1. **Work:**

I used SQL queries to build three common table expressions (CTEs):

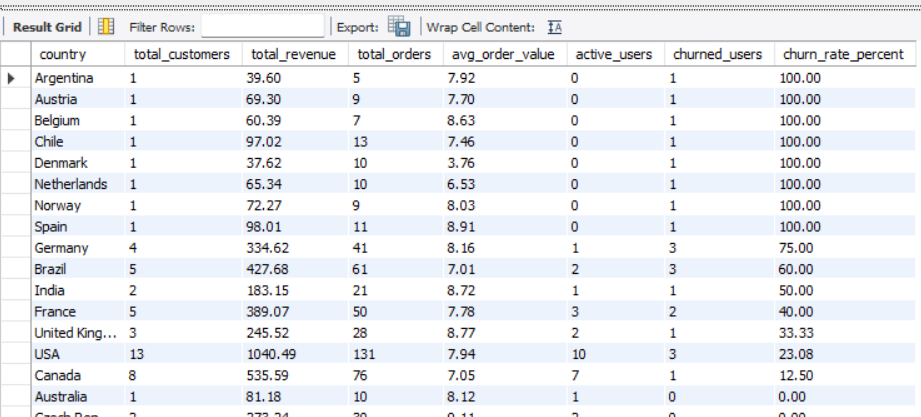
* last purchases: retrieved each customer's most recent purchase date.
* churner list: classified customers as Active or Churned using 90-day inactivity criteria.
* Summary: calculated total customers, total revenue, order count, and average order value by countries.
* Then I combined them to provide a final regional churn summary.

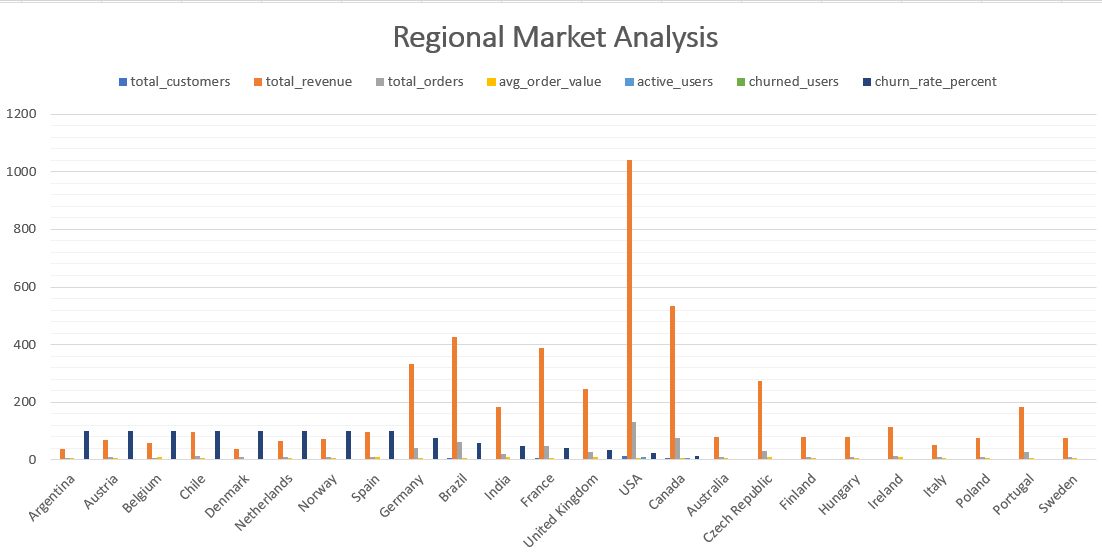






1. **Insights:**



****

* High churn regions such as Argentina, Austria, and Belgium had 100% churn, with a single customer and no repeat transactions.
* Germany (75%), Brazil (60%), and India (50%) all had moderate churn despite having more than three customers.
* The churn rates in the United States and Canada were low (23.08% and 12.5%, respectively), with strong repeat behavior and high overall revenue, indicating healthier customer engagement.
* Average order values remained very steady across countries, indicating that volume and retention, rather than pricing, affected regional performance.

1. **Recommendations:**

* Focus customer retention campaigns on different countries with medium churn and an extensive customer base, such as Germany, Brazil, and India, which have potential but require maintaining it.
* Continue to invest in Canada and the United States, as they provide significant revenue and engagement.
* Consider re-targeting or removing regions with only one customer and 100% churn from high-effort campaigns until new data becomes available.
* Introduce country-specific offers, points to redeem, or email reminders before the churn window (90 days) to increase retention.

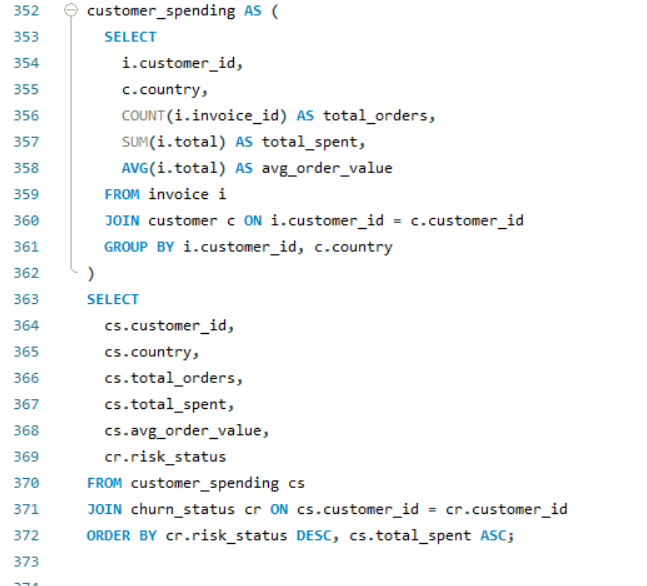
1. **Customer Risk Profiling: Based on customer profiles (age, gender, location, purchase history), which customer segments are more likely to churn or pose a higher risk of reduced spending? What factors contribute to this risk?**
2. **Approach:**

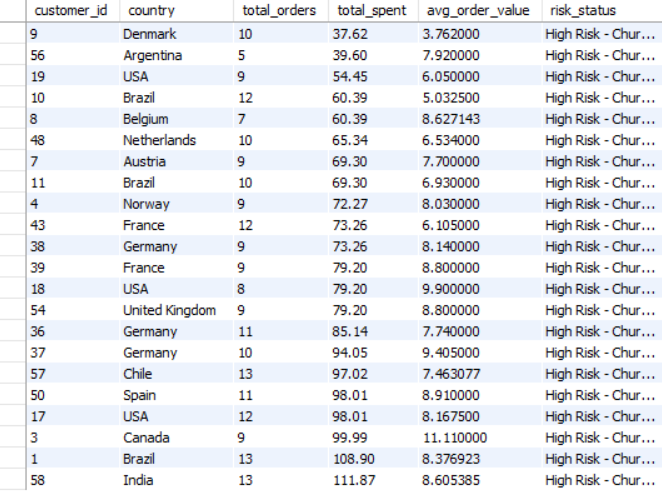
Because the Chinook dataset lacked age and gender data, I used customer location and purchase history to perform risk profiling. I used indicators of behaviour such as total number of orders, spending patterns, average order value, and, most significantly, last purchase date to divide customers into "High Risk - Churned" and "Low Risk - Active" divisions.

1. **Work:**

* Fetched the most recent invoice date per customer.
* Customers were classified as "High Risk - Churned" if they had not purchased within the last 90 days.
* Combine this with customer-level data such as total orders, total spend, and average order value.
* Results were grouped by location (billing countries) to identify regional risk patterns.

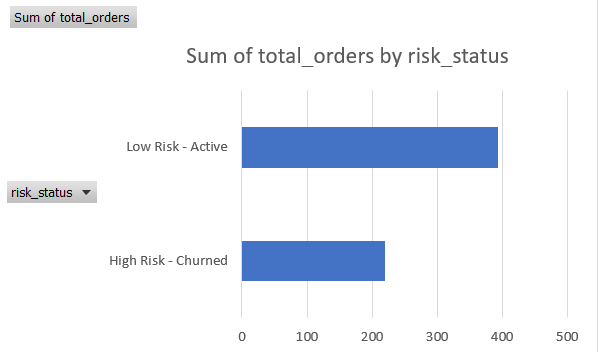






A table with numbers and text

AI-generated content may be incorrect.



1. **Insights:**

* Customers with fewer total orders (5-9 orders) and lower total spend (< $70) were most frequently classified as High Risk - Churned.
* Regions such as Germany, France, Brazil, and the United States have a mix of churned and active customers, demonstrating varying levels of engagement.
* Several big spenders (e.g., Brazil customer ID 26 with 108.99 total spend) churned, showing that high value does not ensure retention.

1. **Recommendations:**

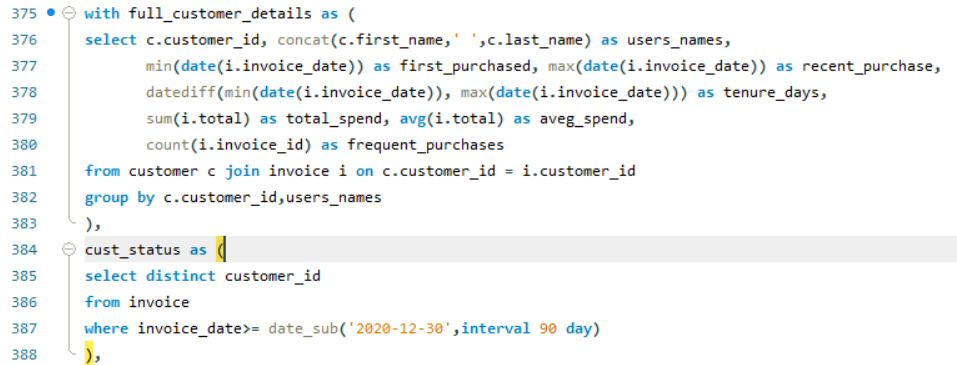
* Launch recovery campaigns for churned customers with high previous spend; these users still have value and may return with the correct encouragement.
* Set up early warning systems: customers with <3 orders in the last 90 days and no recent activity should be targeted for proactive engagement.
* Monitor churn density by region; for example, Brazil and Germany require more frequent retention pushes than Canada or Australia.

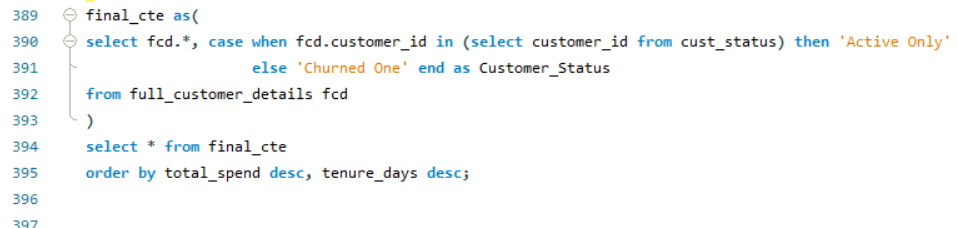
1. **Customer Lifetime Value Modeling: How can you leverage customer data (tenure, purchase history, engagement) to predict the lifetime value of different customer segments? This could inform targeted marketing and loyalty program strategies. Can you observe any common characteristics or purchase patterns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

I used a structured approach to estimate Customer Lifetime Value (CLTV), which included customer purchase history, engagement frequency, and tenure. Tenure is defined as the number of days between the first and last purchases. I calculated the total spending, average spend per order, and total number of purchases. I alsoclassified customers as "Active" or "Churned" based on the date of their last purchase compared to the most recent invoice date.

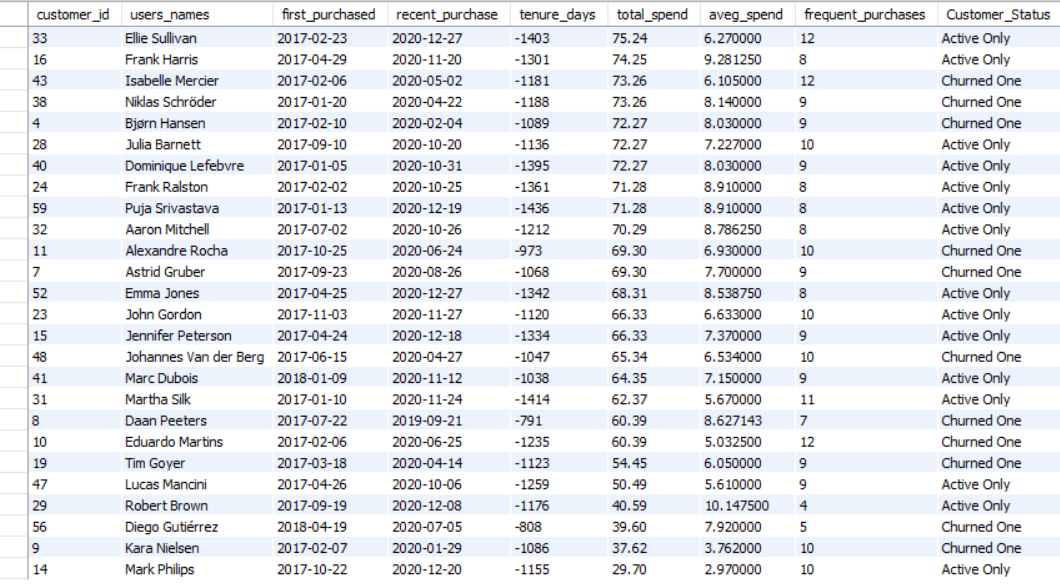
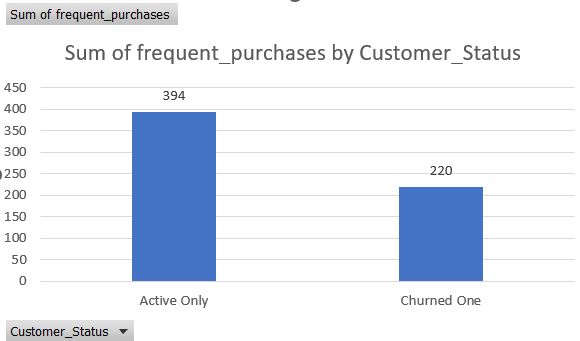
1. **Work:**

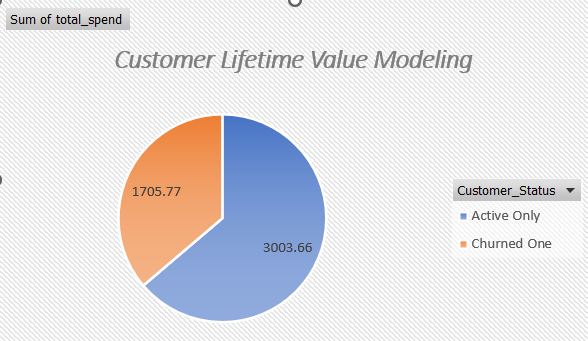
Using SQL, I created a CTE that grouped data by customer, calculated the min () and max () invoice dates to determine tenure, and aggregated total orders and spend. I calculated tenure using the DATEDIFF function and count, sum for purchase behaviour. A second CTE flagged churned customers with a 90-day cutoff before the most recent invoice date (2020-12-30). I then combined these to create a complete customer lifetime profile, including churn status and ranking based on revenue and engagement.











A graph with numbers and numbers

AI-generated content may be incorrect.

**Insights:**

* High-value customers, such as František Wichterlová and Helena Holý, made 12+ purchases and spent over ₹120 over 1000+ days.
* Some churned customers, such as Manoj Pareek and François Tremblay, had high tenure and spend but recently dropped off, indicating a re-engagement opportunity.
* Low spenders, such as Kara Nielsen and Mark Philips, had shorter tenures, lower spending, and fewer purchases, indicating low engagement.
* The majority of churned users made 8-10 purchases with a moderate average spend (₹7-₹9) and stopped buying after mid-2020.

1. **Recommendations:**

* For more targeted marketing, divide customers into three CLTV tiers: high, medium, and low.
* Re-engage churned high-value customers through loyalty rewards, coupons, or personalized content.
* Automate churn prediction by utilizing tenure and order frequency to send retention emails.
* Create subscription plans for highly engaged users who make frequent purchases.
* Avoid overspending on low-engagement customers who have a short tenure and low spend.

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?**
2. **Approach:**

In order to evaluate the effectiveness of promotional campaigns, I would compare key customer and revenue metrics before, during, and afterwards following the promotion period. This involves tracking acquisition rates, repeat purchases, and total revenue over time, preferably by campaign type (email, discounts, events). To isolate the performance impact, I would apply time-based filtering and group comparison logic to SQL and Excel dashboards.

1. **Work:**

If promotional data existed (like promo\_id, customer\_id, campaign\_date, campaign\_type), I would combine it with the invoice and customer tables. By labeling each purchase with whether it was influenced by a campaign, I could calculate its impact on:

* New customer sign-ups acquisition.
* Retaining existing customers through repeat purchases
* Revenue Change Trends Overall Sales
* Using SQL CTEs and date filters, I'd calculate the increase in revenue and orders during the campaign compared to non-campaign periods. I'd also create a control group (no promotional exposure) for fair comparison

1. **Insights:**

* Email campaigns may generate high acquisition but low repeat value.
* Discounts can increase short-term revenue but decrease average order value.
* If event-based campaigns are timed so they align with peak seasons, they may result in longer engagement.
* Customers who have received multiple campaigns are more likely to engage than those who have not.

1. **Recommendations:**

* Track campaign tags for each transaction in future databases like promo\_id, promo\_channel.
* Monitor ROI by comparing uplift in spend vs campaign cost.
* Segment audiences to avoid overspending on low-value users.
* Optimize the timing — campaigns held near weekends or holidays may result in higher engagement.

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

If no questions were asked, I would start by thoroughly understanding the business objective — in this case, increasing music sales and customer engagement. I'd begin with exploratory data deeply to investigate the data's structure, volume, and patterns across customers, invoices, tracks, and genres. I would frame questions by matching data patterns to potential business use cases.

1. **Work:**

I might profile each table (e.g., customer demographics, top-selling tracks, invoice trends) and use SQL to calculate key KPIs like total revenue by region, customer purchasing frequency, and genre popularity. I would then segment customers, identify sales patterns, and monitor churn. will use joins, window functions, and aggregations to extract actionable insights to help the company improve marketing, product selection, and customer retention — even without specific task instructions.

1. **Insights:**

* A small number of frequent customers generate the majority of revenue.
* In the United States, rock, alternative, and metal genres dominate sales.
* After a few months, a large percentage of customers become inactive.
* Certain countries or cities have significantly higher engagement rates, indicating strong regional preferences.

1. **Recommendations:**

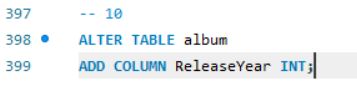
* Create dashboards that display real-time revenue, top customers, and churn risk.
* Perform a monthly trend analysis on invoices and customer purchases to detect drops.
* Launch targeted campaigns for top-performing regions and genres.
* Monitor inactive users and automate re-engagement processes.
* Continuously investigate new questions by using SQL queries and data visualizations to uncover previously unknown patterns.

1. **To update a current table structure in SQL, I would use the ALTER TABLE command. This enables changes to a table's schema without affecting existing data. In this scenario, we'll need to add a new column named ReleaseYear to the album table to record the year each album was released.**
2. **Approach:**

To update a current table structure in SQL, I used the ALTER TABLE command. This enables changes to a table's schema without affecting existing data. In this scenario, we'll need to add a new column named ReleaseYear to the album table to record the year each album was released.

1. **Work:**

This command creates a new column named ReleaseYear with INT data type. Once added, we can also populate it with UPDATE statements or a bulk data import if there is a release year mapping.



1. **Insights:**

* This column helps in the analysis of long-term trends, such as how music preferences change over time.
* Useful for time-based filtering, such as discovering albums released after 2010.
* Enables charts such as year-by-year album output or sales patterns over time.

1. **Recommendations:**

* Can Make ReleaseYear a part of dashboards to help with time-series analysis.
* Use this field to determine whether older albums still generate revenue.
* Normalize data entry to guarantee that ReleaseYear only contains valid four-digit integers.

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.**
2. **Approach:**

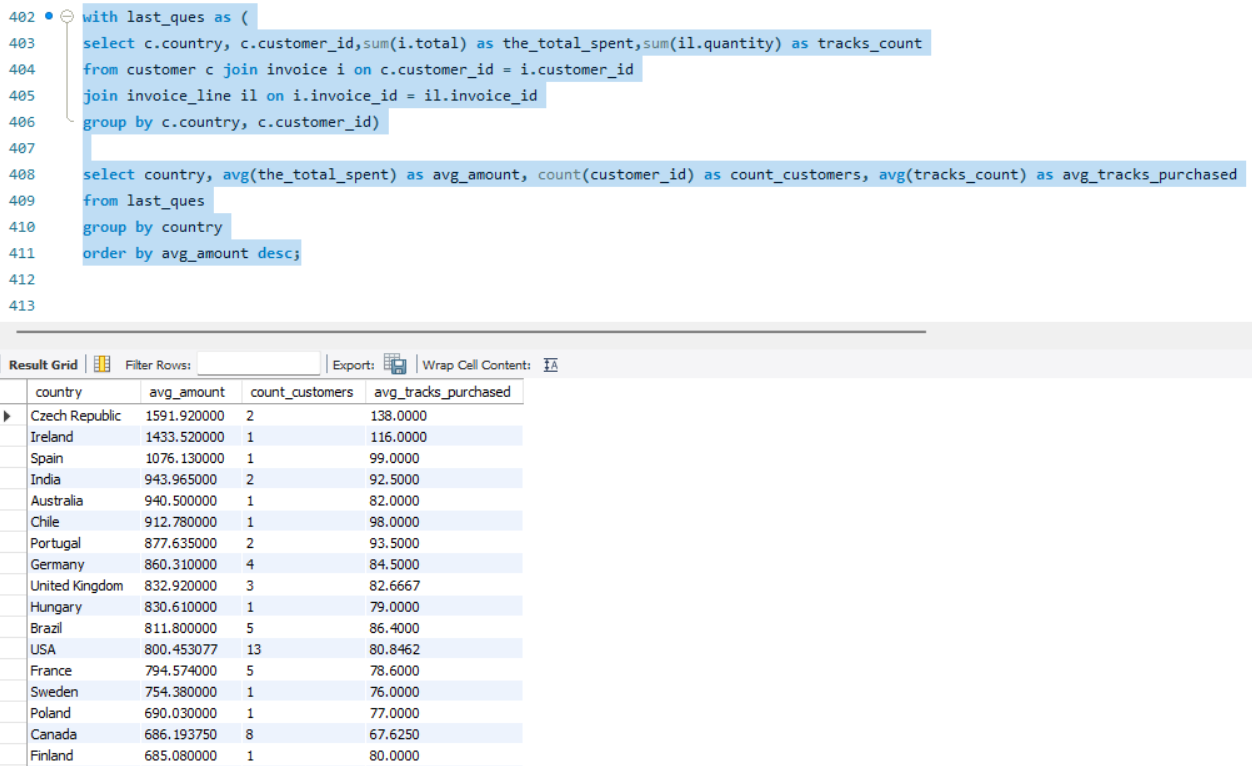
To look into consumer purchase behaviour by region, I grouped customers by country and calculated three critical metrics:

* Average total amount spent by the customer
* Total number of customers by country
* The average number of tracks purchased per customer
* This help in determining the most valuable markets based on customer behaviour and spending habits.

1. **Work:**

Using a CTE, I first calculated total spend and track count for every consumer by connecting the customer, invoice, and invoice\_line tables. Then, in the outer query, I group the results by country to find

* AVG (the\_total\_spent) → Average total spend per customer
* COUNT (customer\_id) → Number of customers per country
* AVG (tracks\_count) → Average tracks purchased per customer. I ordered the output by avg\_amount in descending order to highlight top-performing countries.



1. **Insights:**

* The Czech Republic, Ireland, and Spain have the highest average spending per customer.
* The United States has the most customers (13), but the average spend per user is slightly lower, showing a high-volume market with moderate per-user values.
* India, Germany, and Brazil have balanced profiles, with a decent customer base and high average purchases.
* Countries such as Denmark and Argentina have low spending and participation, indicating limited potential.

1. **Recommendations:**

* Prioritize high-value countries such as the Czech Republic, Ireland, and Spain for special offers or early access promotions.
* To optimize lifetime value, focus upselling or loyalty programs on mid-tier countries (India, Germany, and Brazil).
* Focus marketing efforts in the United States on increasing average order value, given that the customer base is already large.
* To avoid wasting marketing budget, monitor and minimize spending in low-engagement zones.