# Objective Questions:

**Q1. Does any table have missing values or duplicates? If yes how would you handle it ?**

No, we don’t have any duplicates but we do have some null values in different tables and replaced by according to their datatype and appropriate values.

**Example:**

Employee table - Reports\_to column null values are replaced by 0 as the datatype is integer.

Customer table - State column null values are replaced by “None” so as to match it by billing\_state column in invoice

Some null values are replaced by “NA”.

**Query:**

**update employee set reports\_to = '0' where reports\_to is null and employee\_id is not null;**

**update customer set company = 'NA' where company is null and customer\_id is not null;**

**update customer set fax = 'NA' where fax is null and customer\_id is not null;**

**update customer set phone = 'NA' where phone is null and customer\_id is not null;**

**update customer set state = 'None' where state is null and customer\_id is not null;**

**update track set composer = 'None' where composer is null and track\_id is not null;**

**Q2. Find the top-selling tracks and top artist in the USA and identify their most famous genres.**

**Query:**

**top selling tracks in USA – top 10**

**select l.track\_id, t.name, sum(i.total) as total\_revenue**

**from invoice\_line l join invoice i on i.invoice\_id = l.invoice\_id join track t on l.track\_id=t.track\_id**

**where billing\_country = "USA"**

**group by l.track\_id, t.name**

**order by total\_revenue desc, track\_id limit 10;**

**OUTPUT:**

**Top-10 artist in USA with their most famous genre**

**with top\_artist as (select a.artist\_id, a.name, sum(l.quantity) as total\_tracks\_sold**

**from artist a join album al on a.artist\_id = al.artist\_id**

**join track t on al.album\_id = t.album\_id**

**join invoice\_line l on t.track\_id = l.track\_id**

**join invoice i on l.invoice\_id= i.invoice\_id**

**where i.billing\_country = "USA"**

**group by a.artist\_id, a.name**

**order by total\_tracks\_sold desc)**

**select a.artist\_id, a.name, g.genre\_id, g.name, sum(il.quantity) as total\_genre\_quantity from genre g**

**join track t on g.genre\_id = t.genre\_id**

**join album al on al.album\_id = t.album\_id**

**join artist a on a.artist\_id=al.artist\_id**

**join invoice\_line il on il.track\_id=t.track\_id**

**join invoice i on i.invoice\_id = il.invoice\_id**

**where a.artist\_id in (select artist\_id from top\_artist)**

**and i.billing\_country = "USA"**

**group by a.artist\_id, a.name, g.genre\_id, g.name**

**order by total\_genre\_quantity desc limit 10;**

**OUTPUT**

**Q3. What is the customer demographic breakdown (age, gender, location) of Chinook's customer base?**

**Query**

**customer base demographic breakdown on age, gender and location**

**-based on location (customer data for age and gender not available)**

**with location\_breakdown as(**

**select customer\_id, concat(first\_name, " ", last\_name) as full\_name, city, coalesce(state, "NA") as state, country**

**from customer**

**group by country, state, city, customer\_id, full\_name**

**order by country, state, city**

**)**

**Select country, count(\*) as total\_customers from location\_breakdown group by country order by total\_customers desc;**

**OUTPUT**

**Q4. Calculate the total revenue and number of invoices for each country, state, and city**

**total revenue for each customer**

**-- country-wise**

**select i.billing\_country, sum(il.unit\_price\*il.quantity) as revenue, count(distinct(i.invoice\_id)) as total\_orders**

**from invoice i join invoice\_line il on i.invoice\_id = il.invoice\_id**

**group by billing\_country**

**order by billing\_country;**

**OUTPUT**

**-- state-wise**

**select i.billing\_state, sum(il.unit\_price\*il.quantity) as revenue, count(distinct(i.invoice\_id)) as total\_orders**

**from invoice i join invoice\_line il on i.invoice\_id = il.invoice\_id**

**group by billing\_state**

**order by billing\_state;**

**OUTPUT**

**-- city-wise**

**select i.billing\_city, sum(il.unit\_price\*il.quantity) as revenue, count(distinct(i.invoice\_id)) as total\_orders**

**from invoice i join invoice\_line il on i.invoice\_id = il.invoice\_id**

**group by billing\_city**

**order by billing\_city;**

**OUTPUT**

**Q5. Find the top 5 customers by total revenue in each country**

**top 5 customers by total revenue in each country**

**with total\_customers as (**

**select i.billing\_country, concat(c.first\_name, " ", c.last\_name) as customer\_name, sum(i.total) as revenue,**

**rank() over(partition by billing\_country order by sum(i.total) desc) as rnk**

**from invoice i join invoice\_line il on i.invoice\_id = il.invoice\_id**

**join customer c on c.customer\_id=i.customer\_id**

**group by i.billing\_country, c.customer\_id, c.first\_name, c.last\_name**

**)**

**select billing\_country, customer\_name, revenue from total\_customers**

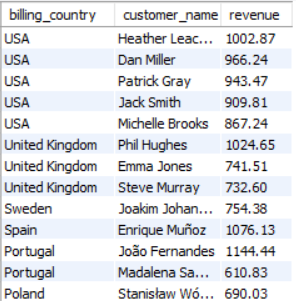
**where rnk<=5**

**order by billing\_country desc, rnk;**

**OUTPUT**

**Top 5 customers by total\_revenue in each country:**

**output is too large so a small amount of data is visualised**



**Q6. Identify the top-selling track for each customer**

**top-selling track for each customer**

**with quantity\_sold as (**

**select c.customer\_id, concat(c.first\_name, c.last\_name) as Full\_Name, t.track\_id, t.name, sum(il.total) as revenue, sum(il.quantity) as total\_quantity,**

**rank() over(partition by c.customer\_id order by sum(il.total) desc) as rnk**

**from customer c join invoice i on c.customer\_id = i.customer\_id**

**join invoice\_line il on il.invoice\_id = i.invoice\_id**

**join track t on t.track\_id = il.track\_id**

**group by c.customer\_id, c.first\_name, c.last\_name, t.track\_id, t.name**

**)**

**select Full\_Name, name as track\_name, quantity**

**from quantity\_sold**

**where rnk = 1**

**order by Full\_Name;**

**OUTPUT:**

**output is too large so a small amount of data is selected for better visualisation**

****

**query for highest amount spent on a track:**

**with quantity\_sold as (**

**select c.customer\_id, concat(c.first\_name, " ", c.last\_name) as Full\_Name, t.track\_id, t.name, sum(i.total) as revenue, sum(il.quantity) as total\_quantity,**

**rank() over(partition by c.customer\_id order by sum(i.total) desc) as rnk**

**from customer c join invoice i on c.customer\_id = i.customer\_id**

**join invoice\_line il on il.invoice\_id = i.invoice\_id**

**join track t on t.track\_id = il.track\_id**

**group by c.customer\_id, c.first\_name, c.last\_name, t.track\_id, t.name**

**)**

**select distinct Full\_Name, revenue, total\_quantity**

**from quantity\_sold**

**where rnk = 1**

**order by revenue desc;**

**OUTPUT:**

**Q7. Are there any patterns or trends in customer purchasing behavior (e.g., frequency of purchases, preferred payment methods, average order value)?**

**Query:**

**WITH PurchaseFrequency AS (**

**SELECT**

**c.customer\_id,**

**CONCAT(c.first\_name,' ',c.last\_name) AS customer\_name,**

**COUNT(i.invoice\_id) AS total\_purchases,**

**MIN(DATE(i.invoice\_date)) AS first\_purchase\_date,**

**MAX(DATE(i.invoice\_date)) AS latest\_purchase\_date,**

**ROUND(**

**DATEDIFF(MAX(DATE(i.invoice\_date)),MIN(DATE(i.invoice\_date))) /**

**COALESCE(COUNT(i.invoice\_id)-1, 0),0) AS avg\_days\_bet\_purchases**

**FROM customer c**

**JOIN invoice i ON c.customer\_id = i.customer\_id**

**GROUP BY 1,2**

**)**

**SELECT \* FROM PurchaseFrequency**

**ORDER BY avg\_days\_bet\_purchases, total\_purchases DESC;**

**OUTPUT:**

**output is too large so a small amount of data is selected for better visualisation**

**-- 2. Average Order Value**

**WITH CustomerPurchases AS (**

**SELECT**

**c.customer\_id, concat(c.first\_name, " ", c.last\_name) as full\_name,**

**SUM(i.total) AS total\_order\_value,**

**COUNT(i.invoice\_id) AS total\_purchases,**

**ROUND(AVG(i.total),2) AS avg\_order\_value**

**FROM customer c**

**LEFT JOIN invoice i ON c.customer\_id = i.customer\_id**

**GROUP BY c.customer\_id, c.first\_name, c.last\_name**

**)**

**SELECT \* FROM CustomerPurchases**

**ORDER BY avg\_order\_value DESC;**

**Q8. What is the customer churn rate?**

**with yearly\_customers as (select extract(year from invoice\_date) as purchase\_year, count(distinct customer\_id) as total\_customers,**

**lag(count(distinct customer\_id)) over(order by extract(year from invoice\_date)) as past\_year\_customers from invoice**

**group by purchase\_year**

**order by purchase\_year)**

**select purchase\_year, total\_customers, past\_year\_customers, (past\_year\_customers-total\_customers) as churned\_customers,**

**case**

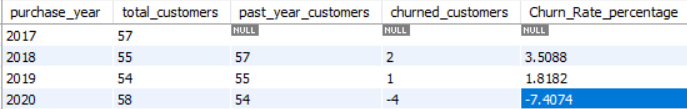
**when past\_year\_customers is Null then null**

**else ((past\_year\_customers-total\_customers)/past\_year\_customers)\*100**

**END as Churn\_Rate\_percentage**

**from yearly\_customers;**

**OUTPUT:**



**Q9. Calculate the percentage of total sales contributed by each genre in the USA and identify the best-selling genres and artists.**

**percentage of total sales contributed by each genre in the USA and identify best selling genre and artists**

**with total\_revenue as (select g.genre\_id, g.name, sum(i.total) as total\_sales\_per\_genre**

**from genre g join track t on g.genre\_id=t.genre\_id**

**join invoice\_line il on t.track\_id=il.track\_id**

**join invoice i on i.invoice\_id=il.invoice\_id**

**where i.billing\_country = "USA"**

**group by g.genre\_id, g.name**

**order by total\_sales\_per\_genre desc)**

**select genre\_id, name, total\_sales\_per\_genre,**

**(total\_sales\_per\_genre/(select sum(i.total) from invoice i**

**join invoice\_line il on i.invoice\_id=il.invoice\_id**

**where i.billing\_country = "USA"))\*100 as percent\_sales**

**from total\_revenue;**

**OUTPUT**

-- **for best selling genre and their artists**

**with total\_revenue as (select g.genre\_id, g.name, sum(il.unit\_price\*il.quantity) as total\_sales\_per\_genre**

**from genre g join track t on g.genre\_id=t.genre\_id**

**join invoice\_line il on t.track\_id=il.track\_id**

**join invoice i on i.invoice\_id=il.invoice\_id**

**where i.billing\_country = "USA"**

**group by g.genre\_id, g.name**

**order by total\_sales\_per\_genre desc),**

**percent\_total\_sales as (select genre\_id, name, total\_sales\_per\_genre,**

**(total\_sales\_per\_genre/(select sum(il.unit\_price\*il.quantity) from invoice I join invoice\_line il on i.invoice\_id=il.invoice\_id**

**where i.billing\_country = "USA"))\*100 as percent\_sales**

**from total\_revenue limit 1)**

**select distinct p.genre\_id, p.name, a.name from percent\_total\_sales p join genre g on p.genre\_id=g.genre\_id**

**join track t on g.genre\_id=t.genre\_id join album al on al.album\_id=t.album\_id**

**join artist a on a.artist\_id = al.artist\_id;**

**OUTPUT:**

**output is too large so a small amount of data is selected for better visualisation**

|  |  |  |
| --- | --- | --- |
| Genre\_ID | Name | Artist\_Name |
| 1 | Rock | AC/DC |
| 1 | Rock | Accept |
| 1 | Rock | Aerosmith |
| 1 | Rock | Alanis Morissette |
| 1 | Rock | Alice In Chains |
| 1 | Rock | Audioslave |
| 1 | Rock | Led Zeppelin |
| 1 | Rock | Frank Zappa & Captain Beefheart |
| 1 | Rock | Queen |
| 1 | Rock | Kiss |
| 1 | Rock | David Coverdale |
| 1 | Rock | Deep Purple |

**Q10. Find customers who have purchased tracks from at least 3 different+ genres**

**customers who have purchased tracks from at least 3 different+ genres**

**SELECT**

**c.customer\_id,**

**CONCAT(first\_name, ' ', last\_name) AS full\_name,**

**count(Distinct t.track\_id) as total\_tracks,**

**COUNT(distinct t.genre\_id) AS total\_genre**

**FROM**

**customer c**

**JOIN**

**invoice i ON c.customer\_id = i.customer\_id**

**JOIN**

**invoice\_line il ON i.invoice\_id = il.invoice\_id**

**JOIN**

**track t ON il.track\_id = t.track\_id**

**JOIN**

**genre g ON t.genre\_id = g.genre\_id**

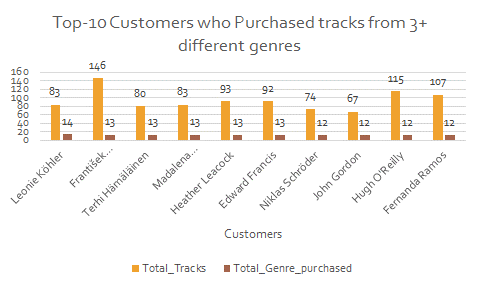
**GROUP BY c.customer\_id , c.first\_name , c.last\_name**

**HAVING COUNT(DISTINCT t.genre\_id) >= 3**

**ORDER BY total\_genre DESC;**

**Output:**

**output is too large so a small amount of data is selected for better visualisation**



**Q11. Rank genres based on their sales performance in the USA**

**Rank genres based on their sales performance in the USA**

**select g.genre\_id, g.name, sum(il.unit\_price\*il.quantity) as total\_sales\_per\_genre,**

**rank() over(order by sum(il.unit\_price\*il.quantity) desc) as rnk**

**from genre g join track t on g.genre\_id=t.genre\_id**

**join invoice\_line il on t.track\_id=il.track\_id**

**join invoice i on i.invoice\_id=il.invoice\_id**

**where i.billing\_country = "USA"**

**group by g.genre\_id, g.name**

**order by genre\_id;**

**Q12. Identify customers who have not made a purchase in the last 3 months**

**customers who have not made any purchase in last 3 months**

**select distinct c.customer\_id, concat(c.first\_name, " " , c.last\_name) as full\_name, i.invoice\_date**

**from customer c join invoice i on c.customer\_id=i.customer\_id**

**where i.invoice\_date > date\_sub((select max(invoice\_date) from invoice), interval 3 month);**

**Output :**

**output is too large so a small amount of data is selected for better visualisation**

|  |
| --- |
| Madalena Sampaio |
| Lucas Mancini |
| Emma Jones |
| Joakim Johansson |
| João Fernandes |
| Patrick Gray |
| Stanisław Wójcik |
| Jennifer Peterson |
| Ladislav Kovács |
| Julia Barnett |
| Victor Stevens |
| Helena Holý |
| Frank Ralston |
| Aaron Mitchell |
| František Wichterlová |

# 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.**

**SELECT g.name, a.album\_id, a.title, SUM(i.total) AS total\_sales**

**FROM invoice\_line il**

**JOIN track t ON il.track\_id = t.track\_id**

**JOIN album a ON t.album\_id = a.album\_id**

**JOIN genre g ON t.genre\_id = g.genre\_id**

**JOIN invoice i ON il.invoice\_id = i.invoice\_id**

**WHERE i.billing\_country = 'USA'**

**GROUP BY g.genre\_id, a.album\_id, a.title**

**ORDER BY total\_sales DESC;**

**OUTPUT :**

**INSIGHTS:**

Based on the Genre Sales Analysis, the following 3 albums can be prioritised for advertising and promotion in USA:

* Seek And Shall Find: More Of The Best (1963-1981) – R&B Soul
* From The Muddy Banks Of The Wishkah [live] - Rock
* House of Pain – Hip Hop Rap

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

**with top\_selling\_genre as (SELECT i.billing\_country, g.name AS genre\_name,**

**SUM(i.total) AS genre\_sales,**

**dense\_rank() OVER (PARTITION BY i.billing\_country ORDER BY SUM(i.total) DESC) AS genre\_rank**

**FROM invoice i**

**JOIN invoice\_line il ON i.invoice\_id = il.invoice\_id**

**JOIN track t ON il.track\_id = t.track\_id**

**JOIN genre g ON t.genre\_id = g.genre\_id**

**WHERE i.billing\_country <> 'USA'**

**GROUP BY i.billing\_country, g.name**

**ORDER BY i.billing\_country, genre\_rank),**

**country\_wise\_genre\_sales as**

**(select billing\_country, genre\_name, genre\_sales from top\_selling\_genre order by billing\_country, genre\_sales desc**

**)**

**select genre\_name, count(billing\_country) as total\_countries, sum(genre\_sales) as total\_sales from country\_wise\_genre\_sales**

**group by genre\_name order by total\_sales desc;**

**OUTPUT:**

**INSIGHTS**

**Commonalities Observed:**

 **Dominance of Rock**:

* "Rock" is overwhelmingly the most popular genre globally (excluding the USA).
* Indicates a strong global preference for rock music, suggesting it has broad international appeal.

 **High Revenue in Some Countries**:

* Countries like **Canada, France,** and **Brazil c**ontribute significantly to Rock's global sales.
* Suggests larger customer bases or higher per-user purchasing in those countries.

**Differences Observed:**

* **There exists a huge difference in the sales of Rock and other genres**.
  + This might reflect regional musical preferences or trends.
  + Revenue is notably lower, suggesting a smaller market or niche demand for Electronics/ Dance, Alternative, Easy Listening, etc..

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?**

**WITH in\_invoice AS (**

**SELECT invoice\_id, COUNT(invoice\_line\_id) AS items\_per\_invoice**

**FROM invoice\_line**

**GROUP BY invoice\_id**

**),**

**invoice\_sum AS (**

**SELECT invoice\_id, SUM(quantity \* unit\_price) AS invoice\_total**

**FROM invoice\_line**

**GROUP BY invoice\_id**

**),**

**customer\_metrics as (SELECT**

**i.customer\_id,**

**MIN(i.invoice\_date) AS first\_purchase,**

**MAX(i.invoice\_date) AS last\_purchase,**

**COUNT(DISTINCT i.invoice\_id) AS total\_orders,**

**AVG(ii.items\_per\_invoice) AS avg\_basket\_size,**

**AVG(im.invoice\_total) AS avg\_spending\_per\_purchase,**

**SUM(im.invoice\_total) AS total\_amount\_spent**

**FROM invoice i**

**JOIN in\_invoice ii ON i.invoice\_id = ii.invoice\_id**

**JOIN invoice\_sum im ON i.invoice\_id = im.invoice\_id**

**GROUP BY i.customer\_id**

**),**

**segmented\_customers as (**

**select \*,**

**case when datediff(date(last\_purchase), date(first\_purchase)) > 1100 then "Long-Term Customers"**

**Else "Short-Term\_Customers"**

**End as Customer\_Classification**

**from customer\_metrics**

**)**

**select customer\_classification, count(customer\_id),**

**avg(total\_orders) as avg\_orders,**

**avg(avg\_basket\_size) as avg\_basket\_size,**

**avg(avg\_spending\_per\_purchase) as avg\_spending\_per\_purchase,**

**avg(total\_amount\_spent) as total\_amount\_spent**

**from segme nted\_customers**

**group by customer\_classification;**

**OUTPUT:**

**INSIGHTS:**

1. **Long-Term Customers Are More Valuable**

* They **order more often, buy more per order,** and **spend more**.
* Their **total lifetime value (LTV) is ~34% higher** than that of short-term customers.

**Strategy**: Focus on **retention programs** to increase long-term customer base — e.g., loyalty points, exclusive offers, or personalized communication.

2. **Short-Term Customers Show Moderate Engagement**

* Despite being short-term, they still place ~8.67 orders and spend a fair amount.
* Indicates initial interest or potential if nurtured correctly.

**Strategy**: Implement **onboarding campaigns**, targeted re-engagement emails, and **special discounts** after first few purchases to encourage repeat behavior.

3. **Basket Size and Spending Patterns Are Predictable**

Long-term customers show **higher average basket size and spending**, which is expected due to established trust or brand affinity.

**Strategy**: For new customers, provide **product recommendations** or **bundled offers** to increase basket size.

4. **Segment-Based Campaigning Is Effective**

The SQL segmentation (based on the time between first and last purchase) helps differentiate customer types meaningfully.

**Strategy**: Use this segmentation logic in your CRM or marketing tools to tailor experiences.

**Conclusion:**

This analysis highlights the **strong impact of customer longevity** on business performance. Long-term customers bring in **more frequent, higher-value transactions**. Investing in **retention strategies** and optimizing the early experience for new customers can help **turn short-term buyers into long-term loyalists,** directly boosting revenue.

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?**

**-- genre purchased together**

**WITH track\_combinations AS (**

**SELECT il1.track\_id AS track\_id\_1, il2.track\_id AS track\_id\_2, COUNT(\*) AS times\_purchased\_together**

**FROM invoice\_line il1**

**JOIN invoice\_line il2 ON il1.invoice\_id = il2.invoice\_id AND il1.track\_id < il2.track\_id**

**GROUP BY il1.track\_id, il2.track\_id**

**),**

**genre\_combinations AS (**

**SELECT t1.genre\_id AS genre\_id\_1, t2.genre\_id AS genre\_id\_2, COUNT(\*) AS times\_purchased\_together**

**FROM track\_combinations tc**

**JOIN track t1 ON tc.track\_id\_1 = t1.track\_id**

**JOIN track t2 ON tc.track\_id\_2 = t2.track\_id**

**WHERE t1.genre\_id <> t2.genre\_id**

**GROUP BY t1.genre\_id, t2.genre\_id**

**)**

**SELECT g1.name AS genre\_1, g2.name AS genre\_2, gc.times\_purchased\_together**

**FROM genre\_combinations gc**

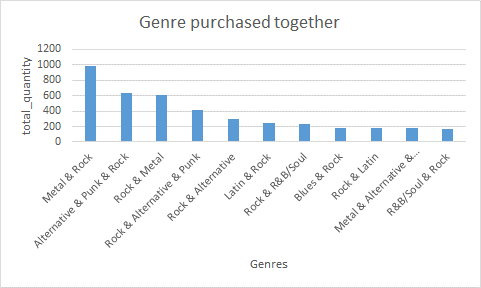
**JOIN genre g1 ON gc.genre\_id\_1 = g1.genre\_id**

**JOIN genre g2 ON gc.genre\_id\_2 = g2.genre\_id**

**ORDER BY gc.times\_purchased\_together DESC;**

**OUTPUT:**

**output is too large so a small amount of data is selected for better visualisation**





**-- artist purchased together**

**WITH track\_combinations AS (**

**SELECT il1.track\_id AS track\_id\_1, il2.track\_id AS track\_id\_2, COUNT(\*) AS times\_purchased\_together**

**FROM invoice\_line il1**

**JOIN invoice\_line il2 ON il1.invoice\_id = il2.invoice\_id AND il1.track\_id < il2.track\_id**

**GROUP BY il1.track\_id, il2.track\_id**

**),**

**artist\_combinations AS (**

**SELECT a1.artist\_id AS artist\_id\_1, a2.artist\_id AS artist\_id\_2, COUNT(\*) AS times\_purchased\_together**

**FROM track\_combinations tc**

**JOIN track t1 ON tc.track\_id\_1 = t1.track\_id**

**JOIN album al1 ON t1.album\_id = al1.album\_id**

**JOIN artist a1 ON al1.artist\_id = a1.artist\_id**

**JOIN track t2 ON tc.track\_id\_2 = t2.track\_id**

**JOIN album al2 ON t2.album\_id = al2.album\_id**

**JOIN artist a2 ON al2.artist\_id = a2.artist\_id**

**WHERE a1.artist\_id <> a2.artist\_id**

**GROUP BY a1.artist\_id, a2.artist\_id**

**)**

**SELECT a1.name AS artist\_1, a2.name AS artist\_2, ac.times\_purchased\_together**

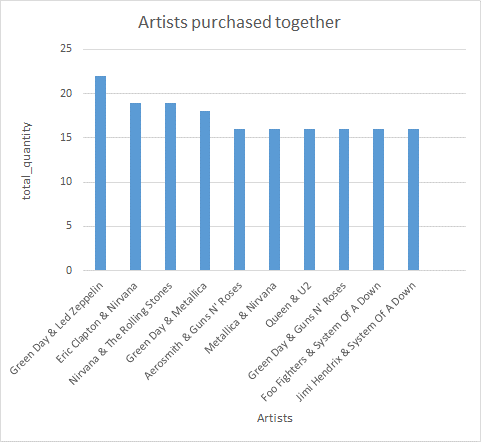
**FROM artist\_combinations ac**

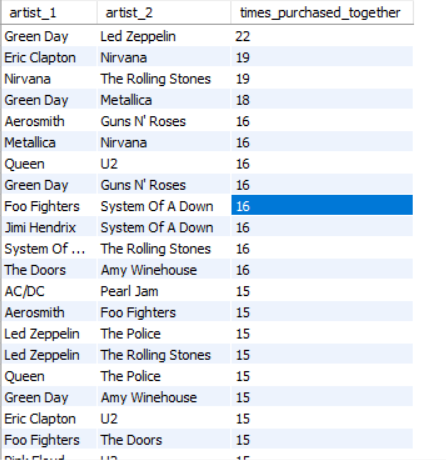
**JOIN artist a1 ON ac.artist\_id\_1 = a1.artist\_id**

**JOIN artist a2 ON ac.artist\_id\_2 = a2.artist\_id**

**ORDER BY ac.times\_purchased\_together DESC;**

**OUTPUT: output is too large so a small amount of data is visualised**



****

**-- albums purchased together**

**WITH track\_combinations AS (**

**SELECT il1.track\_id AS track\_id\_1, il2.track\_id AS track\_id\_2, COUNT(\*) AS times\_purchased\_together**

**FROM invoice\_line il1**

**JOIN invoice\_line il2 ON il1.invoice\_id = il2.invoice\_id AND il1.track\_id < il2.track\_id**

**GROUP BY il1.track\_id, il2.track\_id**

**),**

**album\_combinations AS (**

**SELECT al1.album\_id AS album\_id\_1, al2.album\_id AS album\_id\_2, COUNT(\*) AS times\_purchased\_together**

**FROM track\_combinations tc**

**JOIN track t1 ON tc.track\_id\_1 = t1.track\_id**

**JOIN album al1 ON t1.album\_id = al1.album\_id**

**JOIN track t2 ON tc.track\_id\_2 = t2.track\_id**

**JOIN album al2 ON t2.album\_id = al2.album\_id**

**WHERE al1.album\_id <> al2.album\_id**

**GROUP BY al1.album\_id, al2.album\_id**

**)**

**SELECT al1.title AS album\_1, al2.title AS album\_2, ac.times\_purchased\_together**

**FROM album\_combinations ac**

**JOIN album al1 ON ac.album\_id\_1 = al1.album\_id**

**JOIN album al2 ON ac.album\_id\_2 = al2.album\_id**

**ORDER BY ac.times\_purchased\_together DESC;**

**OUTPUT: output is too large so a small amount of data is visualised**

****

**INSIGHTS:**

**A. Genre Affinity Insights**

From genre data:

**Most frequent pairs**:

* Metal & Rock: 986 purchases
* Alternative & Punk & Rock: 629 purchases
* Rock & Metal: 613 purchases

Other popular combinations: Rock with Latin, R&B/Soul, Classical, Pop, Jazz

* **Rock** acts as a central genre — most genres are paired with it.
* Cross-genre interest exists between adjacent genres (e.g., Metal & Alternative & Punk, Jazz & Rock).

**Strategic Implications:**

* Build **cross-genre bundles** (e.g., Metal + Rock playlists or album combos).
* Implement genre-aware **recommendation engines**: if a user buys Rock, suggest Metal or Alternative & Punk.
* Promote genre diversity with curated "Transition" playlists (e.g., from Classical to Rock, or Blues to Rock).

**B. Artist Affinity Insights**

From your artist data:

Top pairings:

* **Green Day & Led Zeppelin** (22 times)
* **Eric Clapton & Nirvana** (19)
* **Nirvana & The Rolling Stones** (19)

Recurring artists: Green Day, Nirvana, Foo Fighters, Led Zeppelin, System Of A Down

* Users have **diverse artist tastes**, blending classic rock with alternative/grunge.
* Green Day, Foo Fighters, and Nirvana appear frequently — strong artist anchors.

**Strategic Implications:**

* Use **“People who bought X also bought Y”** recommendations (artist-to-artist).
* Create **collaborative artist bundles** (e.g., Nirvana + Foo Fighters).
* Cross-promote artists during anniversaries or events — e.g., System of A Down + Foo Fighters promo packs.

**C. Album Affinity Insights**

From your album data:

Top combos:

* **Facelift** & **Mezmerize**, **The Doors** & **Back to Black** (10 times)
* **Big Ones** appears frequently (paired with 5+ other albums)
* **Dark Side of The Moon** co-purchased with several albums (The Police, My Generation, etc.)

:

* Some albums (like Big Ones, Dark Side of The Moon) serve as **purchase anchors**.
* Pairings often span **eras or genres**, suggesting cross-generational interest.

**Strategic Implications:**

Offer **"album duos"** or curated bundles — e.g., Back to Black + The Doors as “Timeless Voices.”

Feature "Top 10 Frequently Bought Together Albums" as a dynamic storefront module.

Use these combos for **email promotions or discounts** to drive AOV (Average Order Value).

**Overall Recommendation System Strategy**

| **Data Type** | **Use in Recommendation** | **Cross-Selling Strategy** |
| --- | --- | --- |
| **Genres** | Recommend based on genre proximity | Curated genre bundles, playlist starters |
| **Artists** | Artist-to-artist pairing | Suggest “You may also like…” |
| **Albums** | Specific album duos | Package deals, discount for buying both |

**Final Recommendations**

**Dynamic Affinity Engine**  
Use your purchase combination logic to power personalized real-time suggestions (genre/artist/album).

**Cross-Sell During Checkout**

"You might also like..." based on purchase behavior

Bundle recommendations (e.g., Buy Big Ones, get 20% off Jagged Little Pill)

**Targeted Email Campaigns**

Send tailored promotions based on genre/artist affinity

E.g., “Fans of Led Zeppelin also loved Eric Clapton — explore now!”

**UI Integration**

Show affinity-based suggestions in:

Product detail pages

Cart/checkout pages

Music player interfaces

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?**

**yearly revenue per country per year and churn analysis**

**with revenue\_per\_country\_per\_year as (select c.country, extract(year from i.invoice\_date) as invoice\_year,**

**count(distinct c.customer\_id) as total\_customers,**

**sum(quantity\*unit\_price) as total\_revenue,**

**round(sum(quantity\*unit\_price)/count(distinct c.customer\_id), 2) as avg\_revenue\_per\_customer,**

**lag(count(distinct c.customer\_id)) over(partition by country order by extract(year from i.invoice\_date)) as past\_year\_customers**

**from invoice i join invoice\_line il on i.invoice\_id=il.invoice\_id**

**join customer c on c.customer\_id=i.customer\_id**

**group by country, invoice\_year**

**order by country desc, invoice\_year, total\_revenue desc)**

**Select \*,**

**case**

**when past\_year\_customers is Null then 0**

**when total\_customers > past\_year\_customers then 0**

**Else (past\_year\_customers-total\_customers)**

**END as churned\_customers**

**from revenue\_per\_country\_per\_year;’**

**OUTPUT:**

**Churn-Analysis Summary:**

| **Country** | **Churned Customers** | **Notable Trend** |
| --- | --- | --- |
| **USA** | Low (1 in 2019) | High retention, high revenue |
| **Canada** | **High in 2018** | Revenue dropped in 2018, churn recovery in 2019 |
| **France** | Medium (1 in 2019) | Churn affects revenue quickly |
| **India** | 1 in 2019 | Small base, churn impact magnified |
| Most Other Countries | 0 | Stable, low-volume bases |

**Geographic Comparison – Key Takeaways**

* **Developed markets** (USA, Canada, France, Germany):
  + Higher revenue/customer.
  + Some churn but better retention.
  + Worth optimizing via personalization, premium plans, and bundles.
* **Emerging or small markets** (India, Czech Republic, Poland, etc.):
  + Low churn due to small base but vulnerable.
  + Focus should be on **acquisition** and **brand building**.

**Business Strategy Recommendations**

| **Opportunity Type** | **Target Regions** | **Strategy** |
| --- | --- | --- |
| **Retention & Loyalty** | USA, Canada, France | Loyalty rewards, personalized emails, “Thank You” campaigns |
| **Re-engagement** | Canada (2018 churn), India (2019) | Win-back offers, reminders, reactivation discounts |
| **Acquisition & Growth** | Czech Republic, Poland, India | Awareness campaigns, referral programs, localized content |
| **High-Value Upsell** | USA, Germany, France | Cross-sell albums/artists frequently bought in those regions |

**yearly revenue per city per year and churn analysis**

**with revenue\_per\_city\_per\_year as (select billing\_city, extract(year from i.invoice\_date) as invoice\_year,**

**count(distinct customer\_id) as total\_customers,**

**sum(quantity\*unit\_price) as total\_revenue,**

**round(sum(quantity\*unit\_price)/count(distinct customer\_id), 2) as avg\_revenue\_per\_customer,**

**lag(count(distinct customer\_id)) over(partition by billing\_city order by extract(year from i.invoice\_date)) as past\_year\_customers**

**from invoice i join invoice\_line il on i.invoice\_id=il.invoice\_id**

**group by billing\_city, invoice\_year**

**order by billing\_city desc, invoice\_year, total\_revenue desc)**

**Select \*,**

**case**

**when past\_year\_customers is Null then 0**

**Else (past\_year\_customers-total\_customers)**

**END as churned\_customers**

**from revenue\_per\_city\_per\_year;**

**OUTPUT:**

**Geographic Summary:**

| **Region** | **Strong Cities** | **Volatile/At-Risk Cities** | **Strategic Focus** |
| --- | --- | --- | --- |
| **North America** | Mountain View, Yellowknife | Vancouver, Toronto, Ottawa | Improve consistency, retarget churn |
| **Europe** | Prague, Paris, Dublin | Copenhagen, Amsterdam, Rome | Upsell and re-engagement |
| **South America** | São Paulo, Brasília | Buenos Aires, Rio de Janeiro | Grow base, build loyalty |
| **Asia** | Delhi, Bangalore | Limited volume | Acquisition + engagement |

**Business Strategy Recommendation:**

| **Segment** | **Criteria** | **Action Plan** |
| --- | --- | --- |
| **Growth Cities** | Rising revenue & low churn | Invest in content, premium tiers, local partnerships |
| **Churn Risk Cities** | High churn or declining returners | Launch email win-back campaigns; re-engagement promos |
| **Volatile Cities** | Revenue swings without churn | Test loyalty rewards, offer subscription savings for monthly buyers |
| **Stable Low Yield** | Flat or declining spend over 4 years | Try A/B testing new product formats or UI nudges for higher cart value |

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 ?**

**SELECT**

**c.country,**

**COUNT(DISTINCT c.customer\_id) AS total\_customers,**

**COUNT(i.invoice\_id) AS total\_orders,**

**ROUND(SUM(i.total), 2) AS total\_revenue,**

**ROUND(SUM(i.total)/COUNT(DISTINCT c.customer\_id), 2) AS avg\_spend\_per\_customer,**

**ROUND(COUNT(i.invoice\_id)/COUNT(DISTINCT c.customer\_id), 2) AS avg\_orders\_per\_customer**

**FROM customer c**

**JOIN invoice i ON c.customer\_id = i.customer\_id**

**GROUP BY c.country**

**ORDER BY avg\_orders\_per\_customer ASC;**

**OUTPUT:**

**Key Metrics Explained:**

| **Metric** | **Meaning** |
| --- | --- |
| Avg\_orders\_per\_customer | Indicates **engagement level** — how often a customer buys |
| avg\_spend\_per\_customer | Indicates **monetary value** of a customer |
| total\_customers | Shows market size — useful for prioritizing actions |
| total\_revenue | Aggregated monetary contribution by country |

## **High-Risk Customer Segments**

**These are low order frequency and/or low spend per customer segments, especially in small markets.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Avg Orders** | **Avg Spend** | **Risk Factors** |
| Argentina | 5 | 39.6 | Very low order frequency, very small base |
| Belgium | 7 | 60.39 | Low order rate, low revenue |
| Italy | 9 | 50.49 | Weak engagement + lower spend |
| Austria | 9 | 69.3 | Decent spend, but few orders — at risk if not re-engaged |
| Denmark | 10 | 37.62 | Lowest spend per customer in the entire dataset |
| Netherlands | 10 | 65.34 | Low order count & moderate spend |

**Insight**:  
These customers are not deeply engaged and spend less. They are most likely to **churn or drop off** unless reactivated with targeted offers.

**Growth-Oriented Segments (Low Churn Risk)**

These are high **order frequency and high spend** customers **— your most valuable.**

| **Country** | **Avg Orders** | **Avg Spend** | **Notes** |
| --- | --- | --- | --- |
| **Czech Republic** | 15.00 | 136.62 | Highest loyalty + highest value |
| **Portugal** | 14.50 | 92.57 | High spend, high repeat rate |
| **Chile & Ireland** | 13.00 | 97.02–114.84 | Extremely valuable despite small size |
| **Brazil** | 12.20 | 85.54 | Large and loyal base |

**Strategy**:  
Retain aggressively. Offer **exclusive benefits,** early access, or **referral rewards.**

**At-Risk Mid-Tier Segments:**

Moderate order frequency (9–10), decent customer base, but potentially at risk without ongoing engagement

| **Country** | **Avg Orders** | **Avg Spend** | **Risk Notes** |
| --- | --- | --- | --- |
| **Germany** | 10.25 | 83.66 | Strong, but slipping could be costly |
| **USA** | 10.08 | 80.04 | Large base, but not very frequent |
| **Canada** | 9.50 | 66.95 | Decent value but could decline |
| **UK** | 9.33 | 81.84 | Low frequency, potential drop |

**Strategy**:  
Push **personalized reminders**, **rewards for reactivation**, and **subscription nudges** to avoid revenue drop-off.

**What Contributes to Churn Risk?**

| Based on trends across metrics  **Risk Factor**: | **Description & Indicators** |
| --- | --- |
| **Low Avg Orders per Customer** | Signals weak engagement → more likely to churn |
| **Low Avg Spend per Customer** | Customers don’t see enough value or aren’t upsold |
| **Small Market Size** | Higher volatility risk — one churn = major % drop |
| **Volatility in Spending** | Erratic behavior may mean external or seasonal factors |
| **Lack of Local Content/Support** | Possibly less localization → higher churn |

**Recommended Segmentation Actions**

| **Segment** | **Action** |
| --- | --- |
| **High Risk** | Incentivized reactivation, usage nudges, loyalty boosts |
| **Growth Tier** | Premium loyalty programs, referrals, VIP treatment |
| **At-Risk Mid Tier** | Targeted discounts, bundles, long-term commitment discounts |
| **Low Spend / Low Orders** | Consider exit surveys or understand product-market fit locally |

1. **Customer Lifetime Value Modelling: 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?**

**with customer\_lifestyle\_analysis as (**

**select**

**c.customer\_id,**

**concat(c.first\_name, ' ', c.last\_name) as customer\_name,**

**c.country, coalesce(c.state,'Not Available') as state, c.city,**

**min(date(i.invoice\_date)) as first\_purchase\_date,**

**max(date(i.invoice\_date)) as last\_purchase\_date,**

**datediff(max(i.invoice\_date),min(i.invoice\_date)) as customer\_tenure\_days,**

**sum(i.total) as total\_purchase, sum(i.total)/count(i.invoice\_id) as avg\_order\_value,**

**case**

**when max(i.invoice\_date) < date\_sub(curdate(), interval 1 year) then 'Churn' else 'Active'**

**end as status,**

**case**

**when datediff(max(i.invoice\_date),min(i.invoice\_date)) >= 365 then 'Long term' else 'Short term'**

**end as customer\_segment,**

**sum(i.total)/greatest(datediff(max(i.invoice\_date),min(i.invoice\_date)),1)\* 365 as predicted\_annual\_value,**

**sum(i.total) as lifetime\_value**

**from invoice i**

**join customer c on c.customer\_id = i.customer\_id**

**group by c.customer\_id**

**),**

**segment\_analysis as (**

**select**

**customer\_segment,**

**customer\_status,**

**count(customer\_id) as num\_customer,**

**avg(customer\_tenure\_days) as avg\_tenure\_days,**

**avg(total\_spending) as avg\_lifetime\_value,**

**avg(predicted\_annual\_value) as avg\_predicted\_annual\_value**

**from customer\_lifestyle\_analysis**

**group by customer\_segment, customer\_status**

**),**

**churn\_analysis as (**

**select**

**country,**

**state, city,**

**customer\_segment,**

**count(customer\_id) churned\_customer,**

**avg(total\_spending) avg\_lifetime\_value**

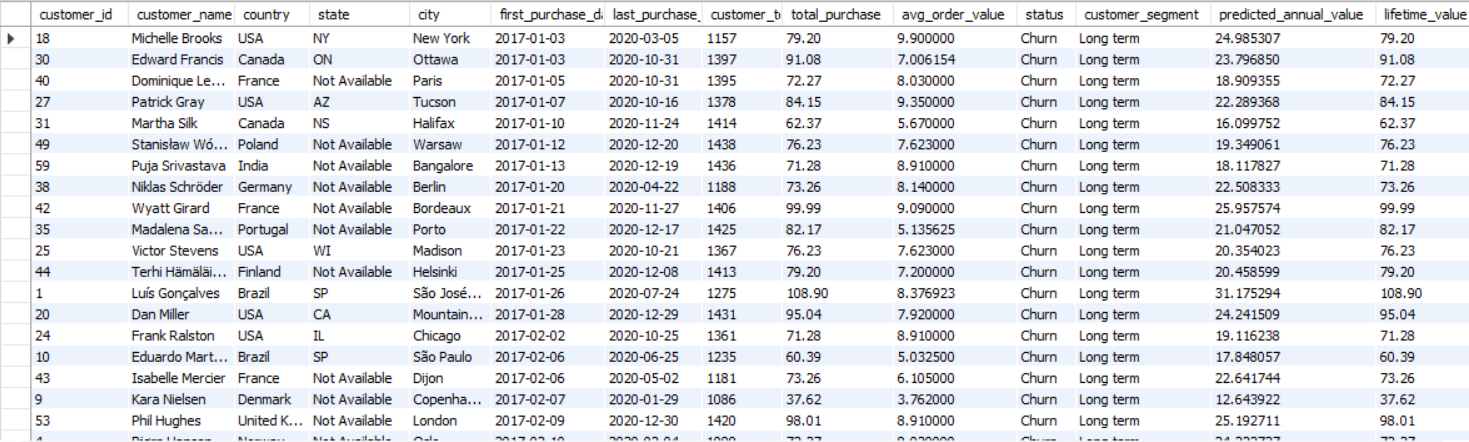
**from customer\_lifestyle\_analysis**

**)**

**select \* from customer\_lifestyle\_analysis;**

**select \* from segment\_analysis;**

**select \* from churn\_analysis;**



**APPROACH:**

* **Customer Lifecycle Analysis**: Calculate tenure, total spending, and classify customers by status (Active or Churn) and segment (Long-term or Short-term).
* **Segment and Churn Analysis**: Group data by customer segment, status, and region, computing lifetime value and predicted annual value.
  + Created a profile for each customer with purchase dates, tenure, number of purchases, total spend, and order average.
  + **Labeled** customers as churned or active based on their last purchase.
  + Grouped customers into long-term and short-term segments using tenure duration.
  + Estimated annual value and total lifetime value for each customer.
  + Summarized customer segments and status to understand trends in spending and retention.
  + **Analyzed** churned customers across locations and segments to detect patterns.

**INSIGHTS:**

* Many churned customers were long-term and high-value, indicating significant lost potential revenue across diverse regions.

**RESULT:**

This output provides:

* Customer count, average tenure, lifetime value, and predicted annual value by segment and status.
* Counts and average lifetime value of churned customers by region and segment.
* These insights aid in identifying high-value customer profiles for targeted marketing and understanding characteristics linked to churn for retention strategies.

**RECOMMENDATIONS:**

1. **Proactive Retention:** Monitor active long-term customers for signs of inactivity (e.g., 9–10 months without purchase) and intervene with personalized engagement.
2. **High-Value Win-Back:** Prioritize targeted campaigns with strong incentives to reactivate churned customers who previously had high total spending.
3. **Tiered Loyalty:** Design loyalty programs that reward customers based on their predicted Lifetime Value to encourage continued high engagement and spending.
4. **If data on promotional campaigns (discounts, events, email marketing) is available, how could you measure their impact on customer acquisition, retention, and overall sales?**

**How to Measure Campaign Impact**

**1. Define Metrics for Success**

**Acquisition:** Number of *new customers* acquired during and after campaign periods.

**Retention:** Rate at which existing customers make repeat purchases after the campaign.

**Sales:** Total sales revenue or order volume uplift during and after campaigns.

**Engagement:** Opens, clicks, and conversions (for email or digital campaigns).

**2. Data Collection & Integration**

Link campaign identifiers to:

* Customers (who received/used the promotion)
* Invoices/orders (made during/after campaign)
* Track timestamps (campaign start/end, purchase dates).

**3. Analyze Acquisition**

* Compare new customers acquired during campaign vs. baseline periods.
* Attribution models to assign new customers to campaigns (e.g., first purchase within X days after exposure).
* Example query metric:

Count of customer\_ids with first invoice during campaign window.

**4. Analyze Retention**

* Track repeat purchases from customers exposed to campaigns vs. non-exposed.
* Calculate retention rate or purchase frequency pre- and post-campaign.
* Example metric:

Percentage of customers with purchases in the 3 months following campaign.

**5. Analyze Sales Impact**

* Calculate incremental sales attributed to campaign periods.
* Compare average order value (AOV) and total revenue for customers using promo codes vs. those who don’t.
* Analyze sales lift relative to baseline (pre-campaign or similar periods without campaigns).

**6. Segment Analysis**

* Break down impact by:
* Customer segments (new vs. existing, location, purchase history)
* Campaign types (discount, event, email)
* Channels (online, in-store, email)

**7. Advanced Approaches**

* **A/B Testing:** Randomly assign customers to campaign vs. control groups to directly measure lift.
* **Time Series Analysis:** Use models (e.g., ARIMA) to isolate campaign effects over time.
* **Propensity Scoring & Matching:** Compare similar customers who received/didn't receive the campaign.
* **Cohort Analysis:** Track groups of customers acquired via campaigns over time.

**Example Queries and Analysis**

-- New customers acquired during campaign period

SELECT COUNT(DISTINCT customer\_id)

FROM invoice

WHERE invoice\_date BETWEEN 'campaign\_start' AND 'campaign\_end'

AND customer\_id NOT IN (

SELECT customer\_id FROM invoice WHERE invoice\_date < 'campaign\_start'

);

-- Repeat purchases post-campaign for customers exposed to campaign

SELECT customer\_id, COUNT(\*) as repeat\_orders

FROM invoice

WHERE invoice\_date > 'campaign\_end'

AND customer\_id IN (

SELECT DISTINCT customer\_id FROM campaign\_exposure WHERE campaign\_id = X

)

GROUP BY customer\_id;

**Summary**

* Use customer-level linking between campaigns and orders.
* Measure before, during, and after campaign behavior.
* Use control groups or historical baselines for accurate attribution.
* Combine quantitative metrics with segmentation to tailor marketing strategies.

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

1. **Understand the Business Context**

* Chinook is a **digital music store** with customers, invoices, tracks, artists, albums, and employees.
* Goals likely include increasing **revenue,** improving **customer retention**, and optimizing **product offerings.**

2. **Explore and Audit the Data**

Start by checking:

* **Missing or duplicate values**
* **Table sizes and row counts**
* **Key relationships** (foreign keys, one-to-many, many-to-many)

**SQL Examples:**

-- **Check for missing customer emails**

SELECT COUNT(\*) FROM customer WHERE email IS NULL;

-- **Check for duplicate track names**

SELECT name, COUNT(\*) FROM track GROUP BY name HAVING COUNT(\*) > 1;

3. **Profile Customers and Their Behavior**

Using customer, invoice, and invoice\_line, analyze:

* **Revenue per customer**
* **Average order value**
* **Number of purchases per customer**
* **Time since last purchase**
* **Country-wise customer distribution**

**SQL Example:**

Average revenue per customer

SELECT customer\_id, SUM(total) AS total\_spent

FROM invoice

GROUP BY customer\_id;

4. **Perform Geographic Analysis**

From the customer and invoice tables:

* Revenue by **country, state,** and **city**
* Regional engagement patterns

**SQL Example:**

SELECT country, COUNT(DISTINCT customer\_id) AS customers,

SUM(total) AS revenue

FROM customer

JOIN invoice USING(customer\_id)

GROUP BY country;

5. **Analyze Product Trends**

Use track, invoice\_line, genre, artist to evaluate:

* Top-selling tracks, albums, artists, and genres
* Average selling price per track/genre
* Genre performance by region

**SQL Example:**

**Top-selling genres**

SELECT g.name AS genre, SUM(il.unit\_price \* il.quantity) AS revenue

FROM invoice\_line il

JOIN track t ON il.track\_id = t.track\_id

JOIN genre g ON t.genre\_id = g.genre\_id

GROUP BY g.name

ORDER BY revenue DESC;

6. **Identify Retention & Churn Patterns**

* Define churn (e.g. no purchase in 90 days)
* Calculate retention rate
* Identify customers at risk (e.g. few purchases, long gap since last)

**SQL Example:**

**Customers who haven’t purchased in last 3 months**

SELECT c.customer\_id, MAX(i.invoice\_date) AS last\_purchase

FROM customer c

JOIN invoice i ON c.customer\_id = i.customer\_id

GROUP BY c.customer\_id

HAVING MAX(i.invoice\_date) < DATE('now', '-90 day');

7. **Cluster & Segment Customers**

Segment customers based on:

* Spend level (high, medium, low)
* Location
* Frequency of purchases
* Product preferences (genre/artist)

This can guide targeted marketing and churn prevention.

8. **Design and Analyze KPIs**

Create key metrics:

* Average revenue per user (ARPU)
* Customer lifetime value (LTV)
* Churn rate
* Top products by contribution

9. **Explore Predictive Opportunities**

If external or temporal data is added (e.g., campaign, feedback, browsing), you can:

* Build churn prediction models
* Forecast genre trends
* Evaluate promo campaign impact

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

The query for adding column will be:

# ALTER TABLE album

ADD COLUMN Release\_Year INTEGER;

1. **Chinook is interested in understanding the purchasing behaviour 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 a SQL query to provide this information.**

**with temp as (**

**select c.country, c.customer\_id, sum(il.unit\_price\*il.quantity) as total\_amount, count(distinct il.track\_id) as total\_tracks**

**from customer c join invoice i on c.customer\_id=i.customer\_id**

**join invoice\_line il on i.invoice\_id=il.invoice\_id**

**group by country, customer\_id**

**)**

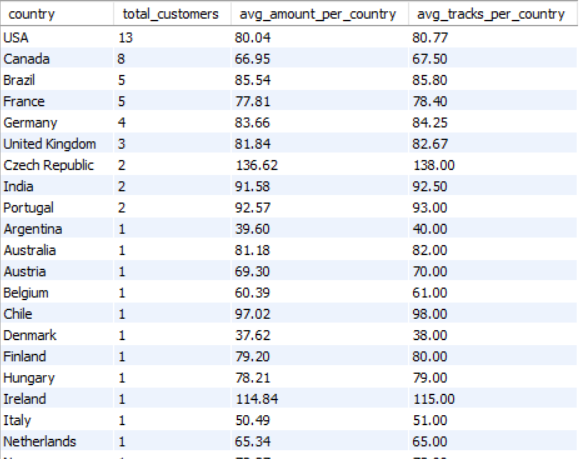
**select country, count(customer\_id) as total\_customers, round(avg(total\_amount),2) as avg\_amount\_per\_country, round(avg(total\_tracks),2) as avg\_tracks\_per\_country**

**from temp**

**group by country**

**order by total\_customers desc;**

**OUTPUT:**

****

**INSIGHTS:**

* **High-Value Pockets:** Countries like **Czech Republic** and **Ireland** show highest average spending per customer, despite few customers.
* **High Track Volume:** **India** leads in average tracks purchased per customer.
* **Volume vs. Value:** The **USA** has the most customers but lower average spending/tracks per customer.

**RESULT:**

The resulting dataset offers valuable insights into country-specific customer purchasing behavior. It specifically highlights:

* Countries with the highest average spending per customer.
* Countries with the highest average track purchases per customer.

**RECOMMENDATIONS:**

* **Target High-Value Markets:** Focus customer acquisition efforts in high-spending countries (e.g., Czech Republic, Ireland).
* **Tailor Promotions:** Offer track-focused deals in countries like India, and upsell/cross-sell in high-volume markets like the USA to boost per-customer value.
* **Localize:** Adapt strategies to capitalize on each region's unique buying habits.