Objective Questions

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

Missing values exist in the **customer, employee, and track** tables, handled by assigning 0 to numeric fields and "Unknown"/"None" to categorical ones. The dataset does not show any duplication issues.

To check duplicates, we use the **primary key**. For example:

select

customer\_id,

count(\*)

from customer

group by customer\_id

having count(\*) > 1;

This approach can be repeated for other tables by replacing the primary key.

To check for missing values, for example in the invoice table:

SELECT \*

FROM invoice

WHERE invoice\_id IS NULL

OR customer\_id IS NULL

OR invoice\_date IS NULL

OR billing\_address IS NULL

OR billing\_city IS NULL

OR billing\_state IS NULL

OR billing\_country IS NULL

OR billing\_postal\_code IS NULL

OR total IS NULL;

This method can be used for other tables also

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

SQL query

SELECT

t.name AS track,

at.name AS artist,

g.name AS genre,

SUM(quantity) AS total\_count,

SUM(quantity \* il.unit\_price) AS total\_revenue,

RANK() OVER(ORDER BY SUM(quantity \* il.unit\_price) DESC,SUM(quantity) DESC) AS RNK

FROM invoice\_line il

JOIN track t

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

JOIN genre g

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

JOIN invoice i

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

JOIN album a

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

JOIN artist at

ON at.artist\_id = a.artist\_id

WHERE billing\_country = 'USA'

GROUP BY t.name, g.name, at.name

ORDER BY total\_count DESC, total\_revenue DESC

LIMIT 5;



* This query lists the top 5 selling tracks in the USA, along with their artist, genre, sales count, revenue, and rank.
* The RANK() window function assigns a ranking based on total revenue first, and then total sales count.
* By joining across invoice\_line, track, album, artist, and genre, it ensures that every track is tied to its artist and genre.
* The output highlights:
  + Which tracks are most popular in the USA.
  + Who the top artists are based on sales.
  + Which genres dominate sales.

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

SQL query:

SELECT

country,

COALESCE(state, "N/A") AS state,

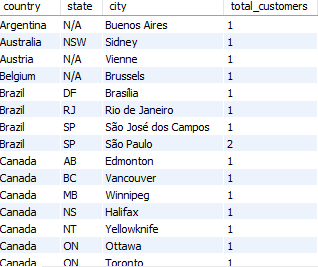
city,

COUNT(\*) AS total\_customers

FROM customer

GROUP BY country, state, city

ORDER BY country;



* This query groups customers by country, state, and city to show the geographic demographic breakdown of Chinook’s customer base.
* COALESCE(state, "N/A") ensures that if a customer record does not have a state, it is labeled as “N/A” instead of being left blank.
* The output highlights:
  + How many customers exist in each country.
  + Breakdown within states and cities.
* While the question mentions age and gender, Chinook’s default schema doesn’t include those fields. Hence, the analysis is focused on location demographics (country, state, city).

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

Sql query

SELECT

billing\_country AS country,

billing\_state AS state,

billing\_city AS city,

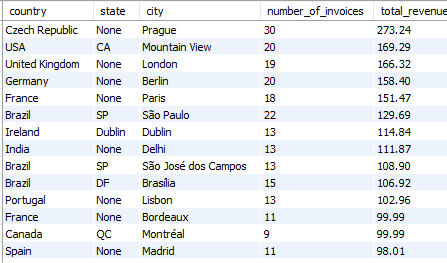
COUNT(invoice\_id) AS number\_of\_invoices,

SUM(total) AS total\_revenue

FROM invoice

GROUP BY billing\_country, billing\_state, billing\_city

ORDER BY total\_revenue DESC;



* The query calculates total revenue and number of invoices for each country, state, and city by grouping the invoice table.
* COUNT(invoice\_id) gives the total number of invoices, while SUM(total) gives the total revenue generated.
* Ordering by total\_revenue DESC helps identify the highest-revenue locations at the top.

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

WITH cust\_total\_rev AS (

SELECT

i.customer\_id,

CONCAT(first\_name, " ", last\_name) AS name,

billing\_country AS country,

SUM(total) AS total\_revenue,

DENSE\_RANK() OVER (PARTITION BY country ORDER BY SUM(total) DESC) AS rnk

FROM invoice i

JOIN customer c

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

GROUP BY customer\_id, billing\_country, CONCAT(first\_name, " ", last\_name)

)

SELECT

name,

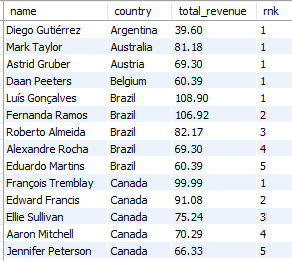
country,

total\_revenue,

rnk

FROM cust\_total\_rev

WHERE rnk <= 5;



* The query first calculates total revenue per customer in each country using a CTE (cust\_total\_rev).
* CONCAT(first\_name, " ", last\_name) creates the full customer name.
* SUM(total) computes the total revenue generated by each customer.
* The outer query uses DENSE\_RANK() partitioned by country to rank customers based on their total revenue within each country.
* The WHERE rnk <= 5 clause selects the top 5 customers per country.

1. Identify the top-selling track for each customer

SQL query

WITH cust\_qnty\_sum AS (

SELECT

CONCAT(first\_name, " ", last\_name) AS customer\_name,

t.name AS track\_name,

SUM(il.quantity) AS sum,

ROW\_NUMBER() OVER (PARTITION BY CONCAT(first\_name, " ", last\_name) ORDER BY SUM(il.quantity) DESC) AS rnk

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

GROUP BY CONCAT(first\_name, " ", last\_name), t.name

)

SELECT

customer\_name,

track\_name,

sum AS total\_quantity

FROM cust\_qnty\_sum

WHERE rnk = 1

ORDER BY total\_quantity DESC;



* The query first calculates the total quantity purchased for each track by each customer using a CTE (cust\_qnty\_sum).
* CONCAT(first\_name, " ", last\_name) creates the customer’s full name.
* SUM(il.quantity) totals how many copies of each track a customer has purchased.
* ROW\_NUMBER() with PARTITION BY customer\_name ranks the tracks for each customer in descending order of quantity purchased.
* The outer query selects only the top track per customer (WHERE rnk = 1).
* Ordering by total\_quantity DESC lists the most purchased tracks across all customers first.

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

Sql query:

WITH cust\_trend\_format AS (

SELECT

customer\_id,

DATE(invoice\_date) AS date,

DATE(LAG(invoice\_date) OVER (PARTITION BY customer\_id ORDER BY invoice\_date)) AS prev\_date,

TIMESTAMPDIFF(DAY, LAG(invoice\_date) OVER (PARTITION BY customer\_id ORDER BY invoice\_date), invoice\_date) AS date\_dif,

total,

billing\_country

FROM invoice i

)

SELECT

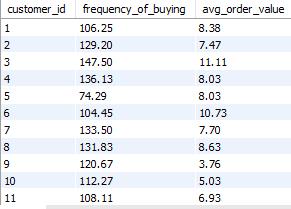
customer\_id,

ROUND(AVG(date\_dif), 2) AS frequency\_of\_buying,

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

FROM cust\_trend\_format

GROUP BY customer\_id;



* Most customers have a frequency of buying between ~100 to 200 days and an average order value around 6–10.
* There is no strong linear trend, meaning customers who buy more frequently do not necessarily spend more per order.
* Common Table Expression (CTE): It uses the LAG() window function partitioned by customer\_id and ordered by invoice\_date. This retrieves the date of the customer's previous purchase.
* Time Difference Calculation: The TIMESTAMPDIFF(DAY, ...) function calculates the time in days (date\_dif) between consecutive orders.
* Outer Query Aggregation: The main query groups the results by customer\_id and calculates two averages:
* frequency\_of\_buying: The average of date\_dif, which indicates the mean number of days between purchases (lower is better).
* avg\_order\_value: The average of total, which represents the customer's average spending per transaction.

1. What is the customer churn rate?

WITH churned\_cust AS

(

SELECT

customer\_id

FROM invoice i

GROUP BY customer\_id

HAVING MAX(invoice\_date) < DATE\_SUB((SELECT MAX(invoice\_date) FROM invoice), INTERVAL 6 MONTH)

)

SELECT ROUND(100\*(SELECT COUNT(\*) FROM churned\_cust)/(SELECT COUNT(DISTINCT customer\_id) FROM customer),2) AS churn\_rate;



* This query calculates the Passive Customer Churn Rate by defining churn as a period of inactivity (6 months) and then finding the percentage of the customer base that meets this criteria.

1 **Common Table Expression (CTE) - churned\_cust:**

* This step identifies the set of churned customers.
* It calculates the **last purchase date** (MAX(invoice\_date)) for every customer\_id.
* The HAVING clause applies the churn definition: A customer is considered churned if their last purchase date is older than 6 months prior to the absolute latest transaction date in the entire invoice table (SELECT MAX(invoice\_date) FROM invoice).

2 **Outer Query:**

* The final query calculates the churn rate percentage
* SELECT COUNT(\*) FROM churned\_cust provides the numerator (the number of customers identified as churned).
* SELECT COUNT(DISTINCT customer\_id) FROM customer provides the denominator (the total number of customers in the system).

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

-- View Creation

CREATE VIEW usa\_genre\_sales AS

SELECT

g.genre\_id,

i.billing\_country AS country,

g.name AS genre,

ROUND(100 \* SUM(il.unit\_price \* il.quantity) / (SELECT SUM(unit\_price \* quantity) FROM invoice\_line il\_sub JOIN invoice i\_sub ON il\_sub.invoice\_id = i\_sub.invoice\_id WHERE i\_sub.billing\_country='USA'), 2) AS total\_sales\_per,

SUM(il.unit\_price \* il.quantity) AS total\_sales,

RANK() OVER(ORDER BY SUM(il.unit\_price \* il.quantity) DESC) AS rnk

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, g.genre\_id

ORDER BY i.billing\_country, total\_sales DESC;

-- 1st part of question : total sales contributed by each genre in the USA

SELECT \* FROM usa\_genre\_sales;

-- 2nd part of question : identify the best-selling genres and artists.

WITH artist\_rnk\_by\_revenue AS

(

SELECT

ar.artist\_id,

ar.name AS artist\_name,

g.genre\_id,

g.name AS genre\_name,

SUM(il.unit\_price \* il.quantity) AS revenue,

ROW\_NUMBER() OVER(PARTITION BY g.genre\_id ORDER BY SUM(il.unit\_price \* il.quantity) DESC) AS rnk

FROM invoice\_line il

JOIN track t

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

JOIN album al

ON t.album\_id = al.album\_id

JOIN artist ar

ON al.artist\_id = ar.artist\_id

JOIN genre g

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

GROUP BY ar.artist\_id, ar.name, g.genre\_id, g.name

)

SELECT

artist\_name,

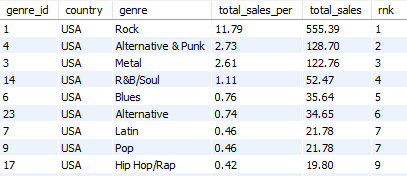
genre\_name,

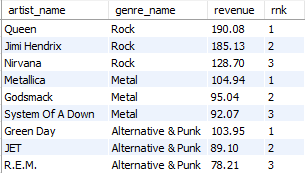
revenue,

rnk

FROM artist\_rnk\_by\_revenue

WHERE rnk <= 3 AND genre\_id IN (SELECT genre\_id FROM usa\_genre\_sales WHERE rnk <= 3);





* This two-part analysis first determines the genre sales performance in the USA and then identifies the top artists driving revenue within the most successful genres.

1. **Genre Sales View (usa\_genre\_sales):** This view filters all sales data to the USA (billing\_country = 'USA') and calculates the total revenue and the percentage of total US sales contributed by each genre. The RANK() function is applied to identify the top-selling genres.
2. **Top Artists CTE:** The Common Table Expression calculates the total revenue for every artist across all genres. It uses the **ROW\_NUMBER()** window function, partitioned by genre\_id, to rank artists *within* each genre based on their revenue.
3. **Final Selection:** The final query combines the rankings, filtering for artists who are both:
   * Ranked in the top 3 sellers within *their specific genre* (rnk <= 3).
   * Belong to one of the *overall top 3 selling genres in the USA* (by referencing the usa\_genre\_sales view).
4. Find customers who have purchased tracks from at least 3 different genres

SELECT

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

COUNT(DISTINCT t.genre\_id) AS num\_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

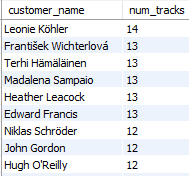
JOIN track t

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

GROUP BY i.customer\_id, CONCAT(c.first\_name , " ", c.last\_name)

HAVING COUNT(DISTINCT t.genre\_id) > 2

ORDER BY num\_tracks DESC;



* This query identifies customers with diverse purchasing habits by counting the unique genres they have bought from.

1. **Data Linking:** It joins the customer, invoice, invoice\_line, and track tables to link each customer's purchase with the corresponding track's genre.
2. **Grouping:** The results are grouped by customer (customer\_id and customer\_name) to aggregate purchases for each individual.
3. **Distinct Genre Count:** The COUNT(DISTINCT t.genre\_id) function calculates the total number of unique genres from which the customer has purchased tracks.
4. **Filtering (The Condition):** The HAVING clause filters these aggregated results, retaining only those customers where the count of distinct genres is **greater than 2** (> 2), thus meeting the "at least 3 different genres" requirement.
5. Rank genres based on their sales performance in the USA

SELECT

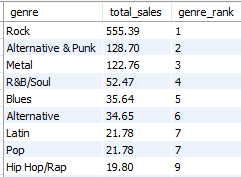
genre,

total\_sales,

rnk AS genre\_rank

FROM usa\_genre\_sales

ORDER BY rnk;



This query utilizes the pre-calculated genre sales and ranking data from the view, streamlining the output to focus only on the essential ranking metrics.

1. **View Selection:** It selects data from the usa\_genre\_sales view, which contains the US genre sales performance metrics.
2. **Column Restriction:** It explicitly selects only three columns: genre (the genre name), total\_sales (the revenue figure), and rnk.
3. **Aliasing:** The pre-calculated rank column, rnk, is aliased as genre\_rank for clarity in the final output.
4. **Ordering:** The results are ordered by the rank (rnk), ensuring the highest-selling genres are displayed at the top.
5. Identify customers who have not made a purchase in the last 3 months

SELECT

c.customer\_id,

CONCAT(first\_name," ",last\_name) AS name,

DATE(recent\_date) AS last\_purchase\_date

FROM

(SELECT

customer\_id,

MAX(invoice\_date) AS recent\_date

FROM invoice

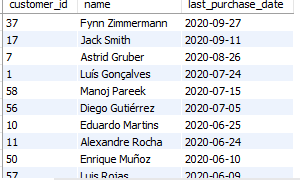
GROUP BY customer\_id) t

JOIN customer c

ON t.customer\_id = c.customer\_id

WHERE TIMESTAMPDIFF(MONTH, recent\_date, (SELECT MAX(invoice\_date) FROM invoice)) >= 3

ORDER BY last\_purchase\_date DESC;



This query identifies customers who have been inactive for a period of 3 months or more, relative to the latest recorded transaction date.

1. **Find Last Purchase Date (Subquery t):** An inner subquery calculates the MAX(invoice\_date) (aliased as recent\_date) for every customer, effectively finding each customer's last purchase date.
2. **Reference Date:** A scalar subquery (SELECT MAX(invoice\_date) FROM invoice) establishes the absolute latest transaction date in the entire dataset.
3. **Customer Linking:** The subquery results are joined with the customer table to retrieve the customer's full name.
4. **Inactivity Filter:** The WHERE clause uses **TIMESTAMPDIFF(MONTH, ...)** to calculate the difference in months between the customer's recent\_date and the overall latest date. It filters for customers where this difference is **3 months or greater** (>= 3), identifying them as inactive.
5. **Ordering:** The results are ordered by last\_purchase\_date to show the longest-inactive customers first.

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

genre\_name,

album\_name,

total\_sales,

rnk AS rank\_by\_sales

FROM

(SELECT

g.genre\_id,

g.name AS genre\_name,

a.title AS album\_name,

SUM(il.quantity \* il.unit\_price) AS total\_sales,

DENSE\_RANK() OVER(PARTITION BY genre\_id ORDER BY SUM(il.quantity \* il.unit\_price) DESC) AS rnk

FROM track t

JOIN genre g

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

JOIN album a

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

JOIN invoice\_line il

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

JOIN invoice i

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

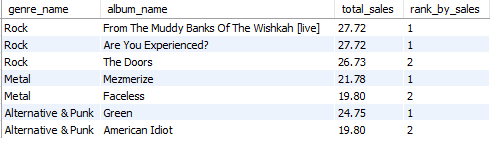
WHERE i.billing\_country = 'USA'

AND g.genre\_id IN (SELECT DISTINCT genre\_id FROM usa\_genre\_sales WHERE rnk <= 3)

GROUP BY g.genre\_id, g.name, a.title

ORDER BY genre\_id, total\_sales DESC) AS t

WHERE rnk IN (1, 2);



**Approach**:

1. **Genre and Location Filtering:** The inner subquery joins the track, genre, album, invoice\_line, and invoice tables. It strictly filters the sales data to transactions occurring only in the **USA** (i.billing\_country = 'USA').
2. **Top Genre Restriction:** It further restricts the data by ensuring only albums belonging to the **top 3 selling genres** in the USA are considered. This is achieved by using a subquery that selects the genre\_id values from the previously created usa\_genre\_sales view where the sales rank (rnk) is 3 or less.
3. **Album Revenue Calculation:** Data is grouped by genre\_id, genre\_name, and album\_name, and the total sales revenue (SUM(il.quantity \* il.unit\_price)) is calculated for each album within the selected top genres.
4. **Ranking Albums by Sales:** The **DENSE\_RANK()** window function is applied. It *partitions* the results by genre\_id (ranking albums independently within each top genre) and *orders* them by total\_sales descending. This assigns a rank to each album based on its revenue within its genre.
5. **Final Selection:** The outer query filters the results to only include the **top two albums** (where rnk is 1 or 2) from each of the top three genres. This provides a focused list of the highest-revenue-generating albums in the most important genre segments.

**Insights**

* **Genre Dominance:** The analysis confirms that the Rock, Alternative & Punk, and Metal genres are the top revenue drivers in the USA, making them the most critical segments for promotional efforts.
* **Highest Sales Concentration:** Rock is the most profitable genre, containing the two highest-selling albums in the entire set, both generating $27.72 and tying for the Rank 1 position within the Rock genre.
* **Key Albums Identified:** The query successfully narrowed down the focus to six top-performing albums within the three key US genres. The top revenue drivers across all identified albums are:
  1. *From The Muddy Banks Of The Wishkah [live]* ($27.72)
  2. *Are You Experienced?* ($27.72)
  3. *Green* (Alternative & Punk, $24.75)
* **High Potential in Alternative & Punk:** The album *Green* shows exceptionally strong sales for its genre ($24.75), placing it as the third highest-selling album overall in this filtered analysis. This suggests a responsive market for this specific album.
* **Consistency in Metal:** While Metal sales are lower than Rock and Alternative & Punk, *Mezmerize* ($21.78) is clearly the leader for that genre, indicating it should be the focus of any Metal-specific promotion.

**Recommendation**

Based on the goal of prioritizing the **three** albums with the highest proven sales performance in the US market, the recommendation focuses on the highest-ranking albums across all top genres:

**Prioritize these 3 Albums for Immediate Promotion in the USA:**

1. **Album:** *From The Muddy Banks Of The Wishkah [live]* (Rock)
   * **Reasoning:** As one of the two tied highest-selling albums across all top genres (Total Sales: $27.72), it is a clear revenue champion and must be placed at the forefront of Rock promotions.
2. **Album:** *Are You Experienced?* (Rock)
   * **Reasoning:** Tied with the first recommendation, this album confirms the extreme profitability of Rock in the USA. Promotional efforts should bundle or feature both Rank 1 Rock albums heavily.
3. **Album:** *Green* (Alternative & Punk)
   * **Reasoning:** This is the next highest-selling album overall (Total Sales: $24.75), demonstrating exceptional performance as the top seller in the Alternative & Punk category. This album offers the best potential for capturing sales outside the dominant Rock segment.

**Actionable Strategy:**

* **Bundling:** Create promotional bundles that combine the two top Rock albums to maximize average order value.
* **Targeted Campaigns:** Run separate, high-impact digital and in-store campaigns specifically targeting the Alternative & Punk demographic with *Green*.
* **Resource Allocation:** Allocate the majority of the US marketing budget toward the promotion of these three albums, as they represent the most reliable and highest-generating product units within the company's best market segments.

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

WITH top\_genre\_per\_country AS

(

SELECT

billing\_country AS country,

g.genre\_id,

g.name,

SUM(il.quantity \* il.unit\_price) AS total\_sales,

DENSE\_RANK() OVER(PARTITION BY billing\_country ORDER BY SUM(il.quantity \* il.unit\_price) DESC) AS rankk

FROM track t

JOIN genre g

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

JOIN invoice\_line il

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

JOIN invoice i

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

WHERE billing\_country != 'USA'

GROUP BY billing\_country, g.genre\_id , g.name

ORDER BY billing\_country

)

SELECT

country,

name AS genre\_name,

rankk AS rnk

FROM top\_genre\_per\_country

WHERE rankk <= 2;

**Output**

****

**Approach:**

The objective of this query is to analyze international sales trends by identifying the two top-selling genres in every country **excluding the USA**. This is achieved using window functions to rank genres based on revenue within each country.

1. **Data Joining and Filtering:** The query joins the track, genre, invoice\_line, and invoice tables to link genre names to sales revenue. A WHERE clause filters out all transactions where the billing\_country is 'USA', focusing exclusively on the international market.
2. **Sales Aggregation:** The data is grouped by billing\_country, genre\_id, and genre\_name to calculate the total\_sales revenue for each genre within each country.
3. **In-Country Ranking (CTE):** The **DENSE\_RANK()** window function is applied within the Common Table Expression (top\_genre\_per\_country).
   * **PARTITION BY billing\_country**: This is critical, as it resets the rank calculation for every new country.
   * **ORDER BY SUM(...) DESC**: This ranks the genres by their total sales revenue, with rank 1 being the highest-selling genre in that specific country.
4. **Final Selection:** The outer query selects the country, genre\_name, and the calculated rank (rnk), filtering the results to include only the **top two** genres per country (rankk <= 2).

**Insights :**

* **Global Preference for Rock:** Rock is the undisputed top-selling genre (Rank 1) in **21 out of the 23 countries** analyzed (91.3%). This indicates a massive global commonality in musical taste among the customer base (with USA).
* **Key Supporting Genres (Commonalities):** The Rank 2 positions reveal a strong, consistent pattern across the world, primarily split between three genres:
  + **Metal:** Appears as Rank 2 in 10 countries (e.g., Canada, Germany, UK, Nordic countries).
  + **Alternative & Punk:** Appears as Rank 2 in 7 countries (e.g., Argentina, Australia, Brazil, India).
* **Regional Differences (Diversity):** While Rock dominates, Rank 2 highlights slight regional nuances and differences in secondary preferences:
  + **Jazz & Blues:** Appear as Rank 2 in only two countries each (Jazz: Austria, Spain; Blues: Poland).
  + **Latin:** Only appears as Rank 2 in Ireland.
  + **R&B/Soul:** Only appears as Rank 2 in the Netherlands.
* **Differences :** Only Argentina deviates from the global trend, with **Alternative & Punk** as the Rank 1 genre and Rock as the Rank 2. This suggests a unique market preference in Argentina that requires a specialized approach.

**Recommendation :**

1 **Global Content Strategy (Commonality):**

* + **Prioritize Rock:** Since Rock is the Rank 1 genre in almost every country, all major international marketing campaigns, content releases, and catalog acquisitions should be heavily weighted towards Rock music to maximize global revenue.
  + **Secondary Focus:** Ensure strong international promotional coverage for Metal and Alternative & Punk, as these are the reliable Rank 2 genres across most territories.

2 **Country-Specific Marketing (Differences):**

* + **Argentina Specialization:** Argentina should be treated as a unique market. Marketing spend and featured album placement should prioritize **Alternative & Punk** over Rock, reflecting the local Rank 1 preference.
  + **Targeted Niche Promotions:** For countries where Rank 2 is a niche genre (e.g., Jazz in Spain, Blues in Poland), run smaller, low-cost digital campaigns specifically targeting those local segments. This allows the company to capture these specialized markets without diverting significant global resources.

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 data\_min\_date AS

(

SELECT

\*,

(SELECT MIN(invoice\_date) FROM invoice WHERE customer\_id = o.customer\_id) AS min\_date

FROM invoice o

ORDER BY customer\_id, invoice\_date

),

cust\_data\_with\_status AS

(

SELECT

customer\_id,

billing\_country,

total,

YEAR(min\_date) AS year,

CASE

WHEN YEAR(min\_date) <= 2017 THEN 'old customer'

ELSE 'new customer'

END AS status

FROM data\_min\_date

),

cust\_agg\_data AS

(

SELECT

status,

COUNT(DISTINCT customer\_id) AS total\_customer,

SUM(total) AS total\_spent,

ROUND(SUM(total) / COUNT(DISTINCT customer\_id), 2) AS avg\_spending

FROM cust\_data\_with\_status

GROUP BY status

),

cust\_order\_frequency AS

(

SELECT

status,

round(AVG(frequency),2) AS avg\_orders\_per\_customer

FROM

(

SELECT

customer\_id,

status,

COUNT(\*) AS frequency

FROM cust\_data\_with\_status

GROUP BY customer\_id, status

) AS t

GROUP BY status

),

basket\_size\_intr AS

(

SELECT

i.customer\_id,

i.invoice\_id,

status,

COUNT(\*) AS basket\_size

FROM invoice i

JOIN invoice\_line il

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

JOIN cust\_data\_with\_status cs

ON cs.customer\_id = i.customer\_id

GROUP BY i.customer\_id, i.invoice\_id, status

ORDER BY i.customer\_id, i.invoice\_id

),

basket\_size\_agg AS

(

SELECT

status,

ROUND(AVG(basket\_size), 2) AS avg\_basket\_size

FROM basket\_size\_intr

GROUP BY status

)

-- final query

SELECT

cd.status,

cd.total\_customer,

cd.total\_spent,

cd.avg\_spending,

cf.avg\_orders\_per\_customer,

ba.avg\_basket\_size

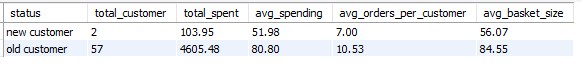
FROM cust\_agg\_data cd

JOIN cust\_order\_frequency cf

ON cd.status = cf.status

JOIN basket\_size\_agg ba

ON cd.status = ba.status;



**Approach:**

This comprehensive query utilizes five Common Table Expressions (CTEs) to segment the customer base into 'old' and 'new' based on their initial purchase year (≤2017 vs. >2017) and calculates distinct purchasing behavior metrics for each group.

1. **Customer Segmentation (data\_min\_date & cust\_data\_with\_status):** The first two CTEs determine the date of the first invoice (min\_date) for every customer and use the year of that first purchase to classify customers into **'old'** (purchased ≤2017) and **'new'** (purchased after 2017) segments.
2. **Spending and Segment Size (cust\_agg\_data):** This CTE calculates the total number of distinct customers, the total cumulative revenue (total\_spent), and the **average spending per customer** (avg\_spending) for each status group.
3. **Purchase Frequency (cust\_order\_frequency):** This CTE first counts the number of invoices (orders) per individual customer, and then calculates the **average number of orders per customer** (avg\_orders\_per\_customer) for the entire 'old' and 'new' groups.
4. **Average Basket Size (basket\_size\_intr & basket\_size\_agg):** These CTEs calculate the number of tracks (items) per invoice (basket\_size) and then aggregate these to find the **average basket size** (avg\_basket\_size) for each customer status group.
5. **Final Synthesis:** The final query joins the results from the three main aggregated CTEs (cust\_agg\_data, cust\_order\_frequency, and basket\_size\_agg) on the status column to present all comparative metrics in a single, comprehensive output row per customer segment.

**Insights :**

* **Financial Disparity (Loyalty Pays):** Long-term ('old') customers are the backbone, generating the vast majority of revenue ($4,605.48 vs. $103.95) and spending $28.82 more per customer on average ($80.80 vs. $51.98).
* **Superior Purchase Habits:** Old customers exhibit superior habits, purchasing significantly more often (avg\_orders\_per\_customer 10.53 vs. 7.00) and buying larger volumes per transaction (avg\_basket\_size 84.55 vs. 56.07).
* **Acquisition/Churn Warning:** The 'new customer' segment is critically small (only 2 customers). This indicates a severe issue with customer acquisition or rapid early-stage churn, making the growth engine highly vulnerable.
* **CLV Model:** The data confirms that **Customer Lifetime Value (CLV) is directly correlated with customer longevity**. Customers who are retained become high-frequency, high-volume purchasers.

**Recommendation :**

* **Prioritize Retention (Tier 1 Focus):** Immediately implement a high-touch **VIP/Loyalty program** for all 'old customers'. Use personalized offers, early access, and exclusive rewards to safeguard this high-CLV segment.
* **Address Early Churn (Tier 2 Focus):** Launch a **structured onboarding and engagement campaign** targeting customers in their first year:
  + **Incentivize Frequency:** Offer small, targeted rewards after the first purchase to encourage the second and third, bridging the gap toward the 10.53 average order frequency.
  + **Increase Basket Size:** Utilize bundling and volume discounts (e.g., "Buy 4 singles, get the 5th free") to encourage new customers to increase their transaction size toward the 84.55 average.
* **Investigate Acquisition Funnel:** Initiate an urgent investigation into the **customer acquisition strategy**.

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?

SELECT

g1.name AS genre\_1,

g2.name AS genre\_2,

COUNT(\*) AS paired\_purchase\_count

FROM invoice\_line i1

JOIN invoice\_line i2

ON i1.invoice\_id = i2.invoice\_id

AND i1.invoice\_line\_id < i2.invoice\_line\_id -- Ensures unique pairing

JOIN track t1

ON t1.track\_id = i1.track\_id

JOIN track t2

ON t2.track\_id = i2.track\_id

JOIN genre g1

ON t1.genre\_id = g1.genre\_id

JOIN genre g2

ON t2.genre\_id = g2.genre\_id

WHERE g1.genre\_id < g2.genre\_id -- Ensures a consistent pair order

GROUP BY g1.name, g2.name

ORDER BY paired\_purchase\_count DESC;

SELECT

ar1.name AS artist\_1,

ar2.name AS artist\_2,

COUNT(\*) AS paired\_purchase\_count

FROM invoice\_line i1

JOIN invoice\_line i2

ON i1.invoice\_id = i2.invoice\_id

AND i1.invoice\_line\_id < i2.invoice\_line\_id -- Ensures unique pairing

JOIN track t1

ON i1.track\_id = t1.track\_id

JOIN track t2

ON i2.track\_id = t2.track\_id

JOIN album al1

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

JOIN album al2

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

JOIN artist ar1

ON al1.artist\_id = ar1.artist\_id

JOIN artist ar2

ON al2.artist\_id = ar2.artist\_id

WHERE ar1.artist\_id < ar2.artist\_id -- Ensures a consistent pair order

GROUP BY ar1.name, ar2.name

ORDER BY paired\_purchase\_count DESC;

SELECT

al1.title AS album\_1,

al2.title AS album\_2,

COUNT(\*) AS paired\_purchase\_count

FROM invoice\_line i1

JOIN invoice\_line i2

ON i1.invoice\_id = i2.invoice\_id

AND i1.invoice\_line\_id < i2.invoice\_line\_id -- Ensures unique pairing

JOIN track t1

ON i1.track\_id = t1.track\_id

JOIN track t2

ON i2.track\_id = t2.track\_id

JOIN album al1

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

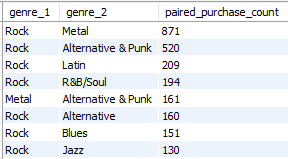
JOIN album al2

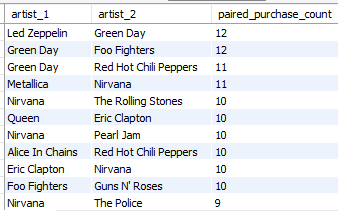
ON t2.album\_id = al2.album\_id

WHERE al1.album\_id < al2.album\_id -- Ensures a consistent pair order

GROUP BY al1.title, al2.title

ORDER BY paired\_purchase\_count DESC;







**Approach :**

The analysis uses **Market Basket Analysis** to find co-occurring purchases at three levels (Genre, Artist, Album).

1. **Item Pairing:** The core mechanism identifies every possible two-item combination within a single customer order (invoice).
2. **Unique Counting Logic:** A conditional check (ID < ID on the paired entities) is implemented to ensure that each unique combination is counted only once, thus providing an accurate paired\_purchase\_count.
3. **Ranking:** The resulting counts are ranked to show the **strongest product affinities** for driving cross-selling strategies using order by.

**Insights :**

* **Genre: The Heavy Music Cluster**
  + **Rock → Metal (871)** and **Rock → Alternative & Punk (520)** represent the highest-confidence cross-sell opportunities, confirming a dominant customer segment for hard/heavy music.
  + This segment has significant purchasing overlap, making genre-based bundling highly effective.
* **Artist: Classic vs. Modern Rock Loyalty**
  + Top affinities pair classic rock (Led Zeppelin, Queen, Eric Clapton) with modern/grunge rock (Green Day, Foo Fighters, Nirvana).
  + This suggests customers are seeking *style continuity* across eras, not just within a single artist's catalog.
  + The high pairings of Green Day with Foo Fighters and Red Hot Chili Peppers (all tied for top spots) confirms a strong "90s/00s Alternative" cluster.
* **Album: Specific Band Cross-Pollination**
  + Album affinities are specific and generally low-count (max 8), suggesting most customers buy full albums from a single band, but the top pairs indicate highly focused cross-selling:
    - **Mezmerize** (System of a Down) and **My Generation** (Limp Bizkit) shows a nu-metal affinity.
    - The frequent pairing involving albums by **The Police** indicates a strong, high-volume catalog affinity around that band.

**Recommendations**

**High-Confidence Cross-Sell (Auto-Recommendation):**

* + Implement immediate, automated pop-up recommendations: If a customer views/adds *any* Rock item, recommend the top two **Metal** albums/tracks, leveraging the 871 count affinity.
* **Themed Bundles:**
  + Create **"Style Bridge" bundles** pairing classic artist albums (e.g., Led Zeppelin's top album) with a modern affiliated artist (e.g., a top Green Day album) based on the Artist Affinity data. Offer a 10% discount on the second album in the pair.
* **Product Placement & Playlists:**
  + Use the **Mezmerize** and **The Police** album data to create highly specific, genre-mixing playlists ("Nu-Metal's Top Influences") to drive traffic to low-volume but highly affiliated products.
* **Landing Page Optimization:**
  + Design dedicated landing pages (e.g., "The Alternative Rock Hub") that feature the top five paired artists (Nirvana, Green Day, Foo Fighters, Red Hot Chili Peppers, Metallica) together to maximize discovery within known segments.

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?

SELECT

i.billing\_country AS country,

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

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

-- Calculates Average Sales per Customer (Total Revenue / Total Customers)

ROUND(SUM(il.quantity \* il.unit\_price) / COUNT(DISTINCT i.customer\_id), 2) AS avg\_sales\_per\_customer,

-- Calculates Average Basket Size (Total Items / Total Transactions)

ROUND(COUNT(il.invoice\_line\_id) / COUNT(DISTINCT i.invoice\_id), 2) AS avg\_basket\_size

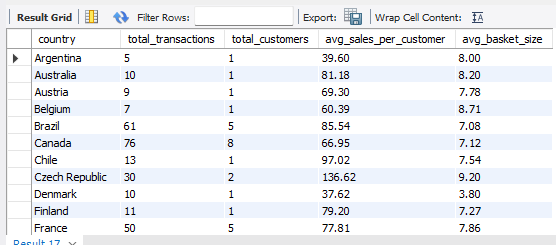
FROM invoice i

JOIN invoice\_line il

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

GROUP BY i.billing\_country

ORDER BY country;



WITH cust\_lastest\_purchase\_date AS

(

SELECT

customer\_id,

billing\_country,

MAX(invoice\_date) AS last\_purchase\_date

FROM invoice

GROUP BY customer\_id, billing\_country

),

churned\_cust\_count AS

(

SELECT

billing\_country AS region,

COUNT(DISTINCT customer\_id) AS total\_customer,

SUM(CASE WHEN TIMESTAMPDIFF(MONTH, last\_purchase\_date, (SELECT MAX(invoice\_date) FROM invoice)) >= 6 THEN 1 ELSE 0 END) AS churned\_cust

FROM cust\_lastest\_purchase\_date

GROUP BY billing\_country

)

SELECT

region,

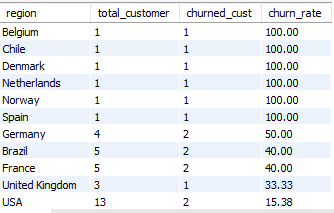
total\_customer,

churned\_cust,

ROUND(100.0 \* churned\_cust / total\_customer, 2) AS churn\_rate

FROM churned\_cust\_count

ORDER BY churn\_rate DESC;



**Churn Rate by Country :**

1. **Define Customer Inactivity:** The first CTE (cust\_lastest\_purchase\_date) establishes the most recent purchase date for *every* unique customer, grouped by their country.
2. **Determine Measurement Period:** The subquery (SELECT MAX(invoice\_date) FROM invoice) dynamically finds the last transaction date in the entire business history. This sets a consistent "Today's Date" for all churn calculations.
3. **Flag Churners:** In the second CTE (churned\_cust\_count), a CASE statement uses TIMESTAMPDIFF(MONTH, ...) to compare the customer's last purchase against the maximum date. If the difference is ≥6 months, the customer is counted as churned\_cust (flagged with 1).
4. **Final Calculation:** The final query aggregates these counts by country and applies the formula: ChurnRate=(ChurnedCustomers/TotalCustomers)×100.

**Insights**

* **Critical Retention Failure in Micro-Markets:** Six countries with a single customer experienced a **100% churn rate** (Belgium, Chile, Denmark, Netherlands, Norway, Spain). This indicates that, for the smallest markets, the company has completely failed to retain the initial buyer.
* **Highest Volume Churners:** While the churn rate is not the highest, **Germany** (50% churn on 4 customers) and the larger markets of **Brazil** and **France** (40% churn on 5 customers each) represent the largest absolute loss of existing customers outside the USA/Canada cluster.
* **Anchor Markets:** The two largest customer bases, the **USA (15.38%)** and **Canada (12.5%)**, exhibit the lowest churn rates among high-volume regions. These markets are relatively stable and form the retention backbone of the business.
* **Perfect Loyalty Illusion:** Twelve different single- or two-customer markets show a **0% churn rate**. Given the small sample size, this is likely a statistical artifact, not a genuine indicator of high loyalty.

**Recommendations**

* **Tiered Win-Back Protocol:**
  + **Level 1 (100% Churn):** Stop all new acquisition spend in these micro-markets (Belgium, Chile, Denmark, etc.) until a cost-effective win-back campaign can be executed to recover the single lost customer.
  + **Level 2 (High Volume Loss):** Launch aggressive, high-value win-back campaigns (e.g., 20% off plus a free track) targeting the churned customers in **Germany, Brazil, and France** to stabilize the business in these medium-sized regions.
* **Retention Focus (Anchor Markets):**
  + Invest in proactive retention strategies (loyalty points, exclusive offers) in the **USA and Canada** to keep churn below the current 15% threshold. These markets are critical for sustainable growth.
* **Investigate Drop-off:** Conduct a post-mortem analysis on why the 100% churn customers failed to return, specifically looking for regional issues such as **payment processing failure**, local taxation issues, or a sudden rise in local competitor popularity.

**Regional Purchasing Behavior :**

1. **Define Aggregation Level:** The query groups all transactional data by billing\_country to measure metrics at the regional level.
2. **Market Size:** COUNT(DISTINCT i.customer\_id) gives the total size of the customer base in that market.
3. **Sales Value (CLV Proxy):** ROUND(SUM(revenue) / COUNT(DISTINCT i.customer\_id), 2) measures the average revenue generated per customer. This is a proxy for Customer Lifetime Value (CLV) in each region.
4. **Transaction Quality:** ROUND(COUNT(items) / COUNT(DISTINCT i.invoice\_id), 2) calculates the average number of items per order (avg\_basket\_size), indicating customer engagement and upselling potential on a per-transaction basis.

**Insights**

* **High-Value Micro-Markets:** Markets like the **Czech Republic** (avg sales: $136.62; basket size: 9.20 items) and **Ireland** (avg sales: $114.84) have exceptionally high spending and basket sizes despite having very few customers. This indicates a highly engaged, high-value customer profile.
* **Largest Markets:** The **USA** (13 customers) and **Canada** (8 customers) are the largest markets by customer volume, but their average sales per customer ($80.04 and $66.95, respectively) are mid-range.
* **Low-Value Markets:** **Denmark** (avg sales: $37.62; basket size: 3.80 items) is a clear outlier, showing significantly lower engagement and spending.

**Recommendations**

* **Protect High-Value Markets:** Launch hyper-focused loyalty campaigns in high-spending markets like the **Czech Republic** and **Ireland** to nurture the existing high-value customers.
* **Market Optimization:** Focus marketing efforts in the **USA** and **Canada** on increasing the average order value (AOV) by promoting premium bundles, as these markets have the necessary customer base but lack top-tier spending.
* **Investigate Low-Value Regions:** Analyze the local competitive landscape in **Denmark** to understand why both spending and basket size are so low. The low basket size suggests poor product discoverability or high friction during checkout.

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?

-- Per customer purchasing behavior and risk segmentation

WITH cust\_spending AS (

SELECT

i.customer\_id,

i.billing\_country,

SUM(i.total) AS total\_spent,

ROUND(SUM(i.total) / COUNT(DISTINCT i.invoice\_id), 2) AS avg\_spending,

MAX(invoice\_date) AS last\_purchase\_date

FROM invoice i

GROUP BY i.customer\_id, i.billing\_country

),

cust\_frequency AS (

SELECT

customer\_id,

COUNT(DISTINCT invoice\_id) AS frequency

FROM invoice

GROUP BY customer\_id

),

basket\_size\_intr AS (

SELECT

i.customer\_id,

i.invoice\_id,

COUNT(il.invoice\_line\_id) AS basket\_size

FROM invoice i

JOIN invoice\_line il

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

GROUP BY i.customer\_id, i.invoice\_id

),

basket\_size\_agg AS (

SELECT

customer\_id,

ROUND(AVG(basket\_size)) AS avg\_basket\_size

FROM basket\_size\_intr

GROUP BY customer\_id

)

SELECT

cs.customer\_id,

cs.billing\_country,

cs.total\_spent,

cs.avg\_spending,

cf.frequency AS total\_purchases,

ba.avg\_basket\_size,

cs.last\_purchase\_date,

TIMESTAMPDIFF(DAY, cs.last\_purchase\_date, (SELECT MAX(invoice\_date) FROM invoice)) AS days\_since\_last\_purchase,

CASE

-- Priority 1: High Risk (Recency - Inactive for 6+ months)

WHEN TIMESTAMPDIFF(DAY, cs.last\_purchase\_date, (SELECT MAX(invoice\_date) FROM invoice)) > 180 THEN 'High Risk (Inactive)'

-- Priority 2: Medium Risk (Frequency - Low purchases)

WHEN cf.frequency <= 2 THEN 'Medium Risk (Low Frequency)'

-- Priority 3: Low Spender (Monetary - Below overall average)

WHEN cs.total\_spent < (SELECT AVG(total\_spent) FROM cust\_spending) THEN 'Low Spender'

ELSE 'Low Risk'

END AS risk\_segment

FROM cust\_spending cs

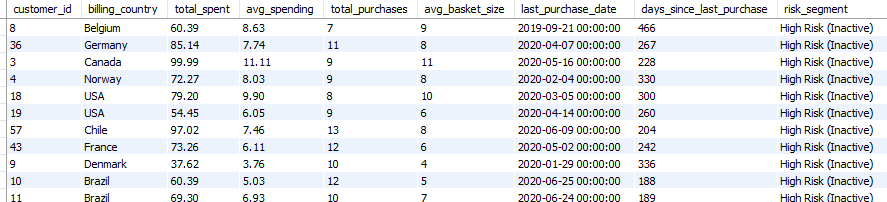
JOIN cust\_frequency cf

ON cs.customer\_id = cf.customer\_id

JOIN basket\_size\_agg ba

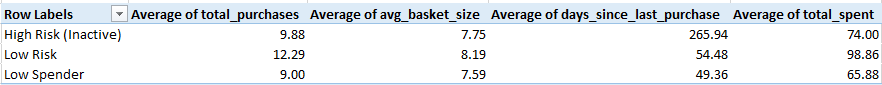
ON cs.customer\_id = ba.customer\_id

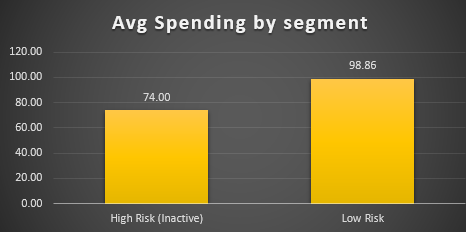
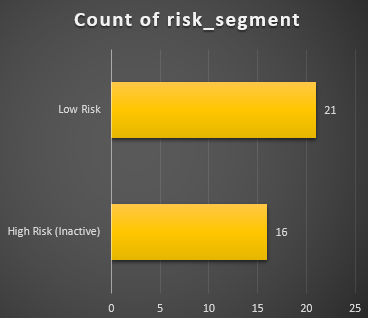
ORDER BY risk\_segment;



Approach:

* **RFM Calculation:** Calculates key purchasing metrics for every customer: days\_since\_last\_purchase (Recency), total\_purchases (Frequency), and total\_spent (Monetary).
* **Risk Segmentation:** Applies a hierarchical CASE statement to assign one of 3 risk labels, prioritizing the most critical risk factors: **Inactivity (Churn Risk)** → **Low Frequency** → **Low Spending**.
* **Factor Analysis:** The final analysis will examine the financial and behavioral characteristics of each resulting segment to identify the specific factors (e.g., country, spending habits) that are highly correlated with the assigned risk.



**Insights :**

* **Primary Risk Factor is Recency:** The **High Risk (Inactive)** segment is defined by its average of **265.94 days since the last purchase**, confirming that **inactivity** is the single most significant indicator of high churn risk.
* **Inactive Customers are Low-to-Average Value:** Despite being frequent shoppers (Avg. 9.88 purchases), the High Risk segment's average lifetime spend is only **$74.00**, which is **below the overall customer average** ($79.82). This indicates the company is losing its lower-value customers first.
* **Low Spenders are High-Value Potential:** The **Low Spender** segment has the best recency (49.36 days) and a decent frequency (9.00 purchases), but the lowest average spending at **$65.88**. This group is engaged but fails to convert to higher-value purchases, representing a huge upsell opportunity.
* **Low Risk Defined by Spending:** The **Low Risk** group is characterized primarily by its high spending, with an average of **$98.86 total spent** and the highest frequency (12.29 purchases). **High Monetary Value and High Frequency** are the key factors for a retained customer.

**Recommendations**

* **High Risk (Inactive) :**

**Strategy:** Offer a **deep, time-bound discount (e.g., 20% off)** or a **free, personalized full album** to re-engage these customers before the 6-month window is significantly exceeded.

* **Low Spender :**
  + **Strategy:** Since this group buys frequently but cheaply, use **Artist and Album Affinity bundles** (e.g., Green Day + Foo Fighters bundle) priced just above their average spending limit to encourage full album or catalog purchases
* **Low Risk** 
  + **Strategy:** Reward this segment's high spending and high frequency with early access to new releases, exclusive merchandise, or a premium loyalty tie

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?

WITH cust\_spending AS (

-- Calculate key monetary and recency metrics per customer

SELECT

i.customer\_id,

i.billing\_country,

SUM(i.total) AS total\_spent,

ROUND(SUM(i.total) / COUNT(DISTINCT i.invoice\_id), 2) AS avg\_spending,

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

FROM invoice i

GROUP BY i.customer\_id, i.billing\_country

),

cust\_frequency AS (

-- Calculate total number of purchases (Frequency)

SELECT

customer\_id,

COUNT(DISTINCT invoice\_id) AS total\_purchases

FROM invoice

GROUP BY customer\_id

),

basket\_size\_intr AS (

-- Calculate basket size for each individual invoice

SELECT

i.customer\_id,

i.invoice\_id,

COUNT(il.invoice\_line\_id) AS basket\_size

FROM invoice i

JOIN invoice\_line il

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

GROUP BY i.customer\_id, i.invoice\_id

),

basket\_size\_agg AS (

-- Calculate the average basket size per customer

SELECT

customer\_id,

ROUND(AVG(basket\_size), 2) AS avg\_basket\_size

FROM basket\_size\_intr

GROUP BY customer\_id

),

-- Final CLV table with calculated metrics and segmentation

customer\_lifetime\_value AS (

SELECT

cs.customer\_id,

cs.billing\_country,

cs.total\_spent,

cs.avg\_spending,

cf.total\_purchases,

ba.avg\_basket\_size,

cs.last\_purchase\_date,

-- Recency: Days since last purchase

TIMESTAMPDIFF(DAY, cs.last\_purchase\_date, (SELECT MAX(invoice\_date) FROM invoice)) AS days\_since\_last\_purchase,

-- 1. Simple Value Segmentation (CLV Proxy)

CASE

WHEN cs.avg\_spending < 3 AND cs.total\_spent < 60 THEN 'Low Value'

WHEN cs.avg\_spending < 6 AND cs.total\_spent < 80 THEN 'Medium Value'

ELSE 'High Value'

END AS customer\_segment,

-- 2. Activity Flag (Churn Indicator)

CASE

WHEN TIMESTAMPDIFF(DAY, cs.last\_purchase\_date, (SELECT MAX(invoice\_date) FROM invoice)) <= 180 THEN 'Active'

ELSE 'Inactive'

END AS activity\_status,

-- 3. Basket Behavior (Purchase Depth)

CASE

WHEN ba.avg\_basket\_size < 2 THEN 'Small Basket Shopper'

WHEN ba.avg\_basket\_size BETWEEN 2 AND 5 THEN 'Medium Basket Shopper'

ELSE 'Bulk Buyer'

END AS basket\_behavior,

-- 4. Frequency Tag (Commitment Level)

CASE

WHEN cf.total\_purchases < 8 THEN 'Not Frequent Buyer'

ELSE 'Frequent Buyer'

END AS purchase\_behavior

FROM cust\_spending cs

JOIN cust\_frequency cf

ON cs.customer\_id = cf.customer\_id

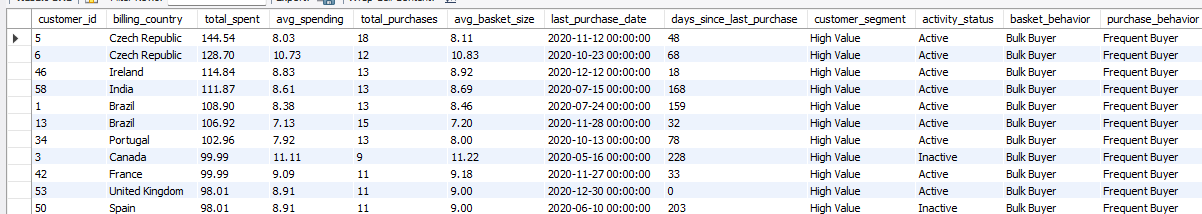
JOIN basket\_size\_agg ba

ON cs.customer\_id = ba.customer\_id

)

SELECT \* FROM customer\_lifetime\_value

ORDER BY total\_spent DESC;



Approach:

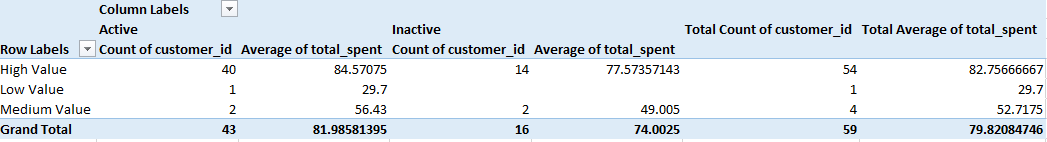
1. **Value Segmentation:**

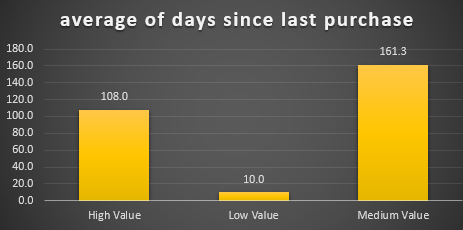
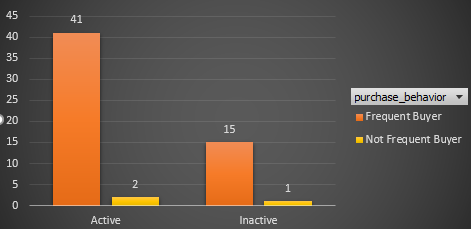
Customers are segmented into **Low, Medium, or High Value** based on a combination of their **Total Spent** and **Average Spending per Transaction**. This establishes the core monetary worth of each customer.

2. **Behavioral Tagging:**

Each customer is independently tagged along three behavioral dimensions, using fixed thresholds:

* **Activity:** An **Inactive** flag is set for customers with ≥180 days since their last purchase, acting as the primary **Recency/Churn indicator**.
* **Frequency:** Tags distinguish **Frequent** from **Not Frequent** buyers based on the total number of purchases (≥8).
* **Basket Size:** Tags define customers as **Small, Medium, or Bulk Buyers** based on their average transaction item count.



Insights

**1. CLV Predictors & Segment Health**

* **CLV is Driven by Retention:** The High Value segment is the healthiest, showing the lowest average Recency (107.96 days), confirming that keeping customers engaged is key to maximizing lifetime value.
* **The Medium Value Flaw:** The Medium Value segment is highly unstable, with a **50% churn rate** (2 active, 2 inactive). This tier represents the highest short-term risk to the current revenue base.
* **Universal Bulk Buying:** Across all segments, the **Average Basket Size is high** (7.59 to 8.19). This means customers are open to purchasing bundles/albums; the limiting factor is **frequency**, not transaction size.

**2. Characteristics of Churned (Inactive) Customers**

* **The Churn Crisis is High-Value:** **14 of 16** total churned customers were originally High Value. This is the primary retention failure point.
* **High Historical Value:** Inactive customers had a high average historical spend ($74.00) and frequency (9.88 purchases). They are not low-quality customers, but **lapsed committed buyers**.
* **Pre-Churn Warning Signal:** High Value customers reduce their spending before they leave; Inactive High Value customers averaged **$77.57** compared to **$84.57** for Active High Value customers.

Recommendation :

1. **High Value Segment → Maximized Retention (Loyalty):**

* **Action:** Immediately launch a **"VIP Retention Program"** for the 40 Active High Value customers. Offer exclusive, high-value perks (e.g., concert pre-sales, limited-edition bundles) to maintain their low Recency (107.96 days).
* **Goal:** Protect the core revenue engine and prevent them from showing the pre-churn spending drop.

2. **Medium Value Segment → Stabilization & Upsell:**

* **Action:** Focus resources on stabilizing this highly volatile segment (50% churn rate) with a personalized 3-month nurturing track. Use their existing willingness to bulk-buy (Avg. Basket Size 7.59) to push premium, high-margin bundles.
* **Goal:** Increase their total spend and reduce the churn rate immediately.

3. **Inactive Segment → High-Value Win-Back:**

* **Action:** Initiate a high-budget **Win-Back Campaign** specifically targeting the 14 lost High Value customers. Offer a significant, personalized incentive (e.g., 50% off their next *full album* purchase) to reactivate their high historical frequency (9.88 purchases).
* **Goal:** Recover the customers who have the highest proven CLV potential.

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?

* **Customer Acquisition:** Identify new customers who placed their first order during the promotion. Compare this with the usual acquisition rate to measure the campaign’s effectiveness in attracting fresh buyers.
* **Customer Retention:** Segment existing customers and analyze their behavior before, during, and after the promotion. Track repeat purchase rates, order frequency, and changes in average basket size to see if the promotion improved loyalty.
* **Sales Impact:** Measure key financial metrics such as total revenue, incremental revenue (sales above normal levels), and average order value. Track coupon/discount code redemption, event participation, or email campaign conversions to directly link results to specific promotions.
* **Behavioral Trends:** Analyze whether promotions shifted customer habits—for example, driving one-time purchases vs. encouraging long-term engagement.
* **Comparative Analysis:** Compare different promotions (discounts vs. events vs. email campaigns) to see which channel delivers the best balance of acquisition, retention, and revenue growth.

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

**Approach if No Specific Questions Are Given**

**Step 1: Understand Business Objectives & Data**

* Clarify the company’s goals: revenue growth, market expansion, retention, or product focus.
* Review datasets: customer demographics, invoices, tracks, genres, albums, regions, and time-based sales trends.

**Step 2: Data Preparation & Exploration**

* Clean the data: handle null values, duplicates, and inconsistent formats.
* Explore customer segments, purchase frequency, and product popularity.
* Perform descriptive statistics to identify early patterns in sales and customer behavior.

**Step 3: Define Core Metrics**

* Revenue-related: total revenue, average order value, revenue per country/region.
* Customer-related: acquisition, retention, churn rate, lifetime value.
* Product-related: best-selling tracks, genres, albums, and most profitable artists.

**Step 4: Develop Key Business Questions**

* Which countries/regions drive maximum sales?
* What are the most popular tracks, albums, and artists by region?
* How do sales vary by genre and location?
* What are the seasonal or time-based sales trends?
* How does customer spending evolve over time?

**Step 5: Perform Deeper Analysis**

* Use SQL queries to segment customers by geography, purchase volume, and genre preference.
* Conduct affinity analysis to detect patterns (e.g., customers who buy X also tend to buy Y).
* Compare high-performing vs. low-performing regions and products.

**Step 6: Generate Insights for Engagement**

* Identify frequent buyers vs. one-time buyers.
* Recommend loyalty programs, point systems, or targeted discounts to improve repeat purchases.
* Propose cross-selling and bundling strategies to maximize revenue per order.

**Step 7: Retention & Win-Back Strategies**

* Spot churn-risk customers (long gaps in purchase activity).
* Run re-engagement campaigns (emails, offers, or special bundles).
* Offer exclusive promotions for high-value or loyal customers.

**Step 8: Visualization & Recommendations**

* Use Excel/BI dashboards to plot:
  + Revenue by country and genre.
  + Customer churn vs. retention trends.
  + Sales performance over time.
* Convert findings into actionable strategies:
  + Boost investment in top genres and regions.
  + Design regional promotions based on preferences.
  + Focus on customer lifetime value through loyalty programs and personalized offers.

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?

ALTER TABLE album

ADD COLUMN ReleaseYear INT(4);

**Approach:**  
Use the ALTER TABLE command to modify the existing table structure. Here, we add a new column ReleaseYear with the data type INT(4) to the album table.

**Summary**

The ReleaseYear column stores the year an album was released, helping in organizing, filtering, and analyzing albums based on their release year (e.g., finding albums released in a specific year or decade).

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.

WITH CustomerMetrics AS (

-- Step 1: Calculate Total Tracks and Total Spending per individual Customer

SELECT

i.customer\_id,

i.billing\_country,

-- Total spent is calculated by summing the line item prices (robust)

SUM(il.unit\_price \* il.quantity) AS total\_spent\_by\_customer,

-- Total number of tracks purchased by the customer

COUNT(il.track\_id) AS total\_tracks\_purchased

FROM invoice i

JOIN invoice\_line il

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

GROUP BY i.customer\_id, i.billing\_country

)

-- Step 2: Aggregate the customer metrics by Country

SELECT

billing\_country,

COUNT(customer\_id) AS number\_of\_customers,

-- Average of the 'total\_spent\_by\_customer' for all customers in that country

AVG(total\_spent\_by\_customer) AS average\_total\_amount\_spent,

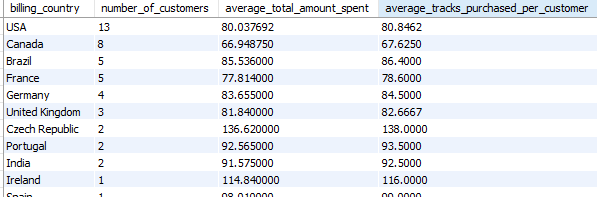
-- Average of the 'total\_tracks\_purchased' for all customers in that country

AVG(total\_tracks\_purchased) AS average\_tracks\_purchased\_per\_customer

FROM CustomerMetrics

GROUP BY billing\_country

ORDER BY number\_of\_customers DESC, average\_total\_amount\_spent DESC;



Insights

1 **High-value small markets** (high spend per customer, few customers):

* Czech Republic: avg spend 136.62, avg tracks 138
* Ireland: avg spend 114.84, avg tracks 116
* Portugal: avg spend 92.56, avg tracks 93.5
* India: avg spend 91.57, avg tracks 92.5

2 **Medium-value markets** (good customer base, decent spend):

* USA: 13 customers, avg spend 80.04, avg tracks 80.8
* Canada: 8 customers, avg spend 66.95, avg tracks 67.6
* Brazil: 5 customers, avg spend 85.54, avg tracks 86.4
* France: 5 customers, avg spend 77.81, avg tracks 78.6
* Germany: 4 customers, avg spend 83.65, avg tracks 84.5

3 **Low-value markets** (low spend per customer, few customers):

* Argentina: avg spend 39.6, avg tracks 40
* Denmark: avg spend 37.62, avg tracks 38
* Italy: avg spend 50.49, avg tracks 51

4 **Strategic insight:**

* Countries with **high customer count and decent spending** (USA, Canada, Brazil, Germany, France) are **priority markets for advertising and promotions**.
* Countries with **high spend but few customers** (Czech Republic, Ireland, Portugal, India) are ideal for **premium offerings and upselling strategies**.
* Countries with **low spend and few customers** (Argentina, Denmark, Italy) may require **targeted campaigns, discounts, or local partnerships** to improve engagement.

**Recommendations**

1. **Focus on growth in strong markets:**
   * The **USA and Canada** have the largest customer bases; targeted promotions and loyalty programs here could further increase revenue.
   * **Brazil, France, and Germany** also have strong potential with a good balance of customer numbers and high spend.
2. **Nurture high-spending small markets:**
   * Countries like **Czech Republic, Ireland, India, and Portugal** show **very high spending per customer**.
   * Chinook could explore **premium offerings, exclusive releases, or bundled packages** here to maximize value.
3. **Improve weaker markets:**
   * Countries such as **Argentina, Denmark, and Italy** show low spending.
   * Strategies could include **localized pricing, promotional discounts, or partnerships with local distributors** to boost engagement.
4. **Track long-term trends:**
   * If high-spending small markets (like Czech Republic) grow in customer base, they could become **priority markets for expansion**