## **Objective Question-Answers**

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

**To check duplicates,** we use the **primary key**.

For example:

Select customer\_id, count(\*)

from customer

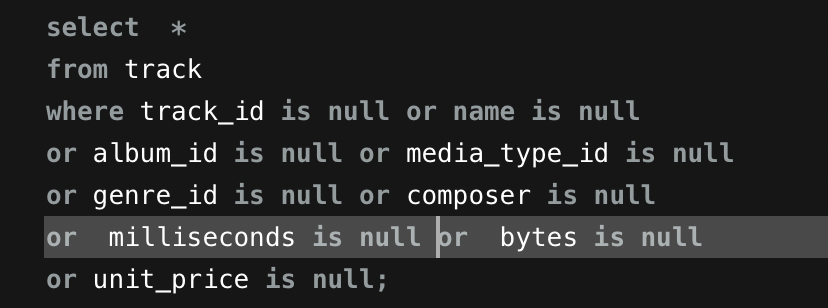
group by customer\_id

having count(\*) > 1;

This approach can be repeated for other tables by using the primary keys of those tables.

The dataset does not show any duplication issues.

To check for missing values, for example in the ‘track’ table:



Missing values exist in the **customer, employee, and track** tables.

They were handled by assigning 0 to numeric fields and "Unknown"/"None" to categorical ones.

## 

## 

## 

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

**SQL Query**

select il.track\_id, t.name as track, sum(il.unit\_price\*quantity) as total\_rev,

sum(il.quantity) as number\_of\_tracks,

a.name as artist, g.name as genrename

from invoice\_line il

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

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

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

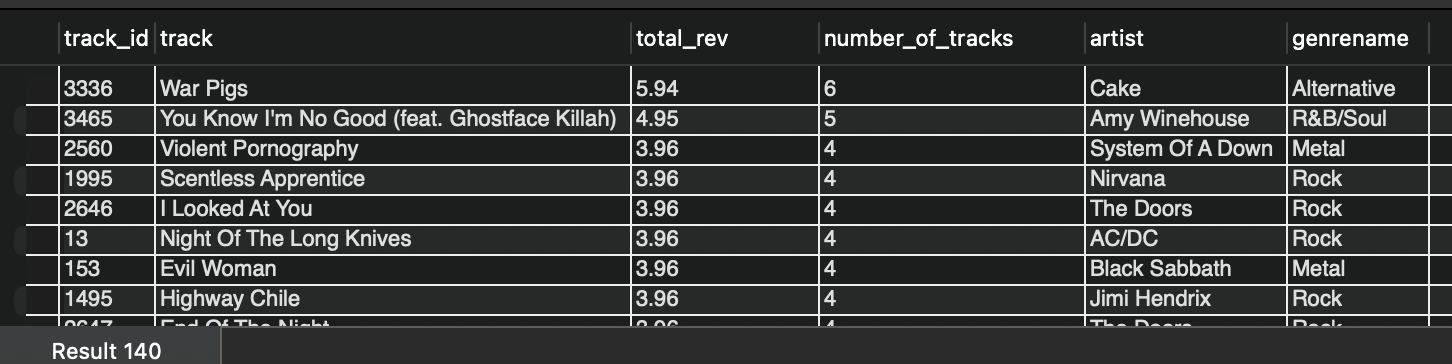
join artist a on al.artist\_id=a.artist\_id

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

where i.billing\_country='USA'

group by il.track\_id, t.name, a.name, g.name

order by total\_rev desc, number\_of\_tracks desc;



This Query:

* Filters sales data to **USA** transactions.
* Aggregates revenue and units sold at the **track level**.
* Joins ensure each track is correctly associated with its **artist** and **genre**.
* Results are ranked by **total revenue**, with quantity used as a tie-breaker.

### **Key Findings**

* *War Pigs* by **Cake** (Alternative) is the **top-selling track** in the USA.
* Other top tracks are largely from **Rock, Metal, and Alternative/R&B** genres.
* Artists such as **Nirvana, System Of A Down, AC/DC, and Amy Winehouse** appear frequently.

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

**The dataset does NOT contain age/gender data**

**SQL query:**

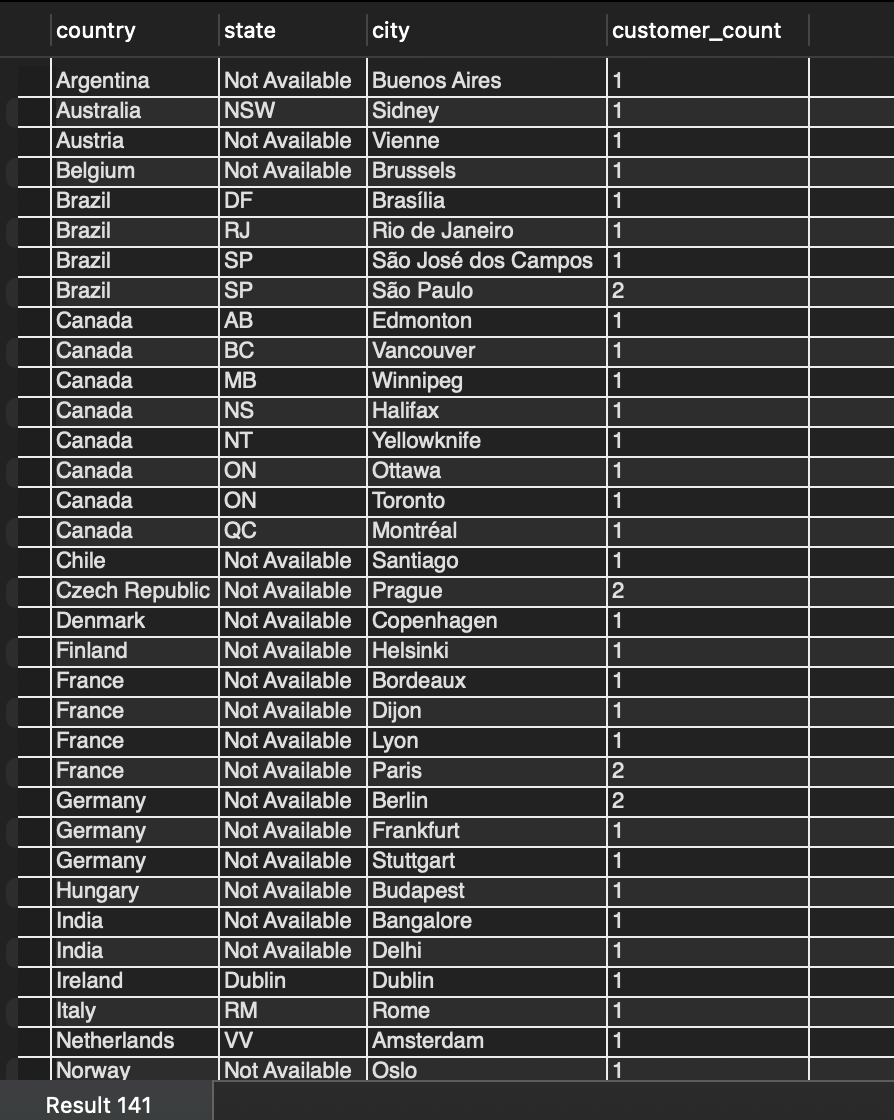
select country, coalesce(state,'Not Available') as state, city,

count(distinct customer\_id) as customer\_count

from customer

group by country, state, city

order by country;



* The query aggregates customer data by **country, state, and city** to understand geographic distribution.
* COALESCE(state, 'Not Available') standardizes missing state values for cleaner reporting.
* COUNT(DISTINCT customer\_id) ensures each customer is counted once per location.

**Key Findings**

* Chinook’s customer base is **globally distributed**.
* **USA** show the highest concentration of customers, across multiple cities.
* Several countries have customers concentrated in a **single major city**, indicating localized markets.
* Missing state-level data in some regions suggests **varying data completeness** across countries.

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

**SQL query:**

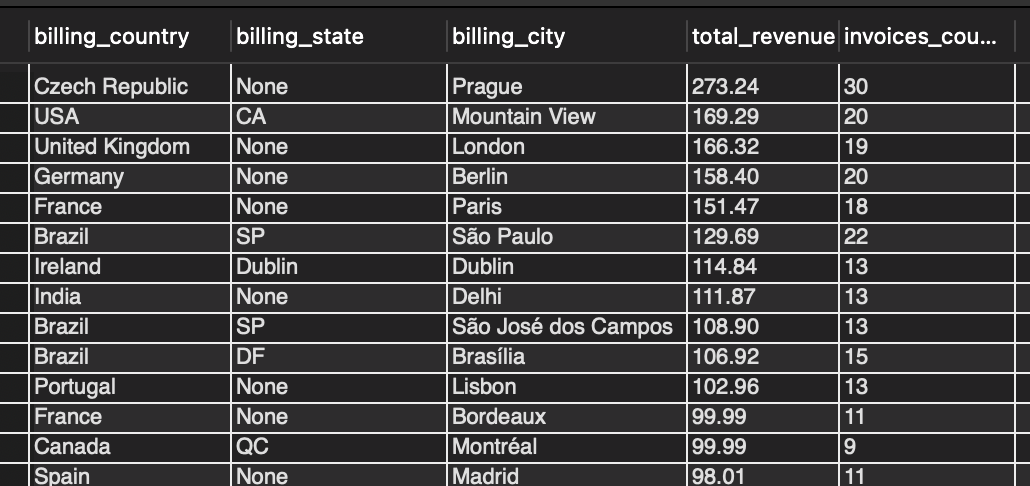
select billing\_country,billing\_state, billing\_city, sum(total) as total\_revenue,

count(distinct invoice\_id) as invoices\_count

from invoice

group by billing\_country,billing\_state, billing\_city

order by total\_revenue desc;



* The query aggregates invoice data at the **country–state–city** level.
* sum(total) calculates total revenue generated from each location.
* count(distinct invoice\_id) measures the number of transactions per location.
* Results are ordered by **total revenue** to identify top-performing regions.

### **Key Findings**

* **Prague (Czech Republic)** generates the highest total revenue
* Locations with higher invoice counts generally correspond to higher total revenue, showing a strong link between transaction volume and revenue.

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

**SQL query:**

with total as(

select c.country, i.customer\_id, concat(c.first\_name,' ',c.last\_name) as full\_name,

sum(i.total) as total\_rev

from invoice i

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

group by c.country, i.customer\_id, concat(c.first\_name,' ',c.last\_name)

),

ranked as(

select \*,

rank() over(partition by country order by total\_rev desc) as rnk

from total

)

select \*

from ranked

where rnk<=5;



### 

* Customer-level revenue is first aggregated by **country and customer**.
* A **window function (RANK())** is applied to rank customers **within each country** based on total revenue.
* PARTITION BY country ensures rankings are calculated independently per country.
* The final filter (rnk <= 5) selects the **top 5 revenue-generating customers** for each country.

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

**SQL Query:**

with total as(

select c.customer\_id, concat(c.first\_name,' ',c.last\_name) as full\_name, il.track\_id, t.name,

sum(il.quantity) as total\_units

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 c.customer\_id, il.track\_id, t.name

),

ranked as(

select \*,

row\_number() over(partition by customer\_id order by total\_units desc, track\_id) as rnk

from total

)

select \*

from ranked

where rnk=1

order by total\_units desc;



* Track-level purchase quantities are aggregated **per customer**.
* ROW\_NUMBER() is used with PARTITION BY customer\_id to rank tracks independently for each customer.
* Ordering by total\_units DESC identifies the most purchased track.
* track\_id is used as a secondary sort key to ensure deterministic tie-breaking.
* Filtering on rnk = 1 guarantees **exactly one top-selling track per customer**.

**Key Findings**

* Each customer is associated with a **single top-selling track**, based on total units purchased.
* Most customers’ top tracks have **low unit counts (1–2 units)**, reflecting the single-track purchase behavior typical of the dataset.
* The result provides a clear view of **individual customer preferences**, useful for personalization and recommendation use cases.

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

with customer\_spending as(

select customer\_id,

sum(total) as total\_spent,

round(sum(total)/count(distinct invoice\_id),2) as avg\_order\_value,

max(invoice\_date) as last\_purchase\_date,

count(distinct invoice\_id) as n\_transactions

from invoice

group by customer\_id

),

basket\_size as(

select i.customer\_id, i.invoice\_id, count(distinct il.invoice\_line\_id) as basket\_s

from invoice i

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

group by i.customer\_id, i.invoice\_id

),

avg\_basket\_size as(

select customer\_id, round(avg(basket\_s),2) as average\_basket\_size

from basket\_size

group by customer\_id

)

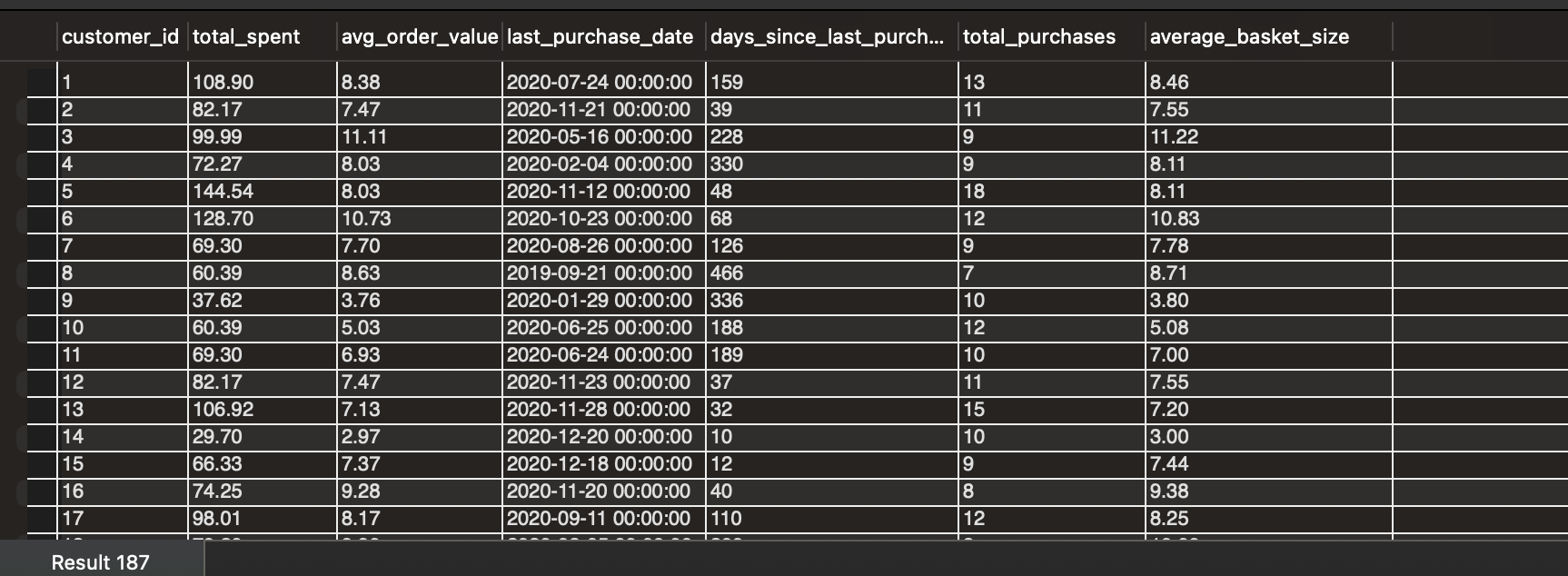
select cs.customer\_id, cs.total\_spent, cs.avg\_order\_value,

cs.last\_purchase\_date, datediff('2020-12-30',cs.last\_purchase\_date) as days\_since\_last\_purchase,

cs.n\_transactions as total\_purchases, bs.average\_basket\_size

from customer\_spending cs

join avg\_basket\_size bs on cs.customer\_id=bs.customer\_id



The query aggregates transaction data per customer to calculate:

* Total purchases (frequency)
* Average order value
* Average basket size (intensity of purchase)
* Total spending
* Days since last purchase (engagement trend)

**Key Insights**

* **Purchase Frequency Drives Total Spend**Customers with higher total\_purchases consistently show higher total\_spent.
* **Average Order Value** Is Relatively Stable
* **Basket Size Is Consistent Across Customers:** This indicates a **stable purchasing depth**, where customers tend to buy a similar number of tracks per transaction.
* **Recency Highlights Churn Risk**Customers with very high days\_since\_last\_purchase (200+ days) often have: Lower recent engagement, Fewer total purchases and Moderate or declining total spend

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

**SQL Query:**

with purchased\_earlier as(

select customer\_id, count(\*) as n\_purchases

from invoice

where invoice\_date < date\_sub('2020-12-30', interval 6 month)

group by customer\_id

),

purchased\_in\_last6\_months as(

select customer\_id, count(\*) as purchases

from invoice

where invoice\_date between date\_sub('2020-12-30', interval 6 month) and '2020-12-30'

group by customer\_id

),

retained\_cust as(

select l.customer\_id

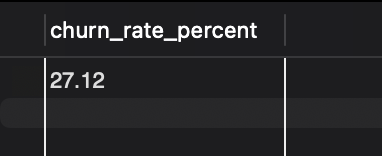
from purchased\_earlier e

join purchased\_in\_last6\_months l on e.customer\_id=l.customer\_id

)

select

round(((select count(\*) from purchased\_earlier) - (select count(\*) from retained\_cust))\*100.0/(select count(\*) from purchased\_earlier),2) as churn\_rate\_percent;



**Definition used**: A customer is considered **churned** if they made at least one purchase **before the last 6 months** but made **no purchases in the most recent 6-month period**.

Reference date used: **30 December 2020** (latest invoice date).

* Customers with purchases **before the 6-month cutoff** form the eligible churn base.
* Customers with purchases **in the last 6 months, who purchased in the earlier period as**  are treated as retained.
* For this, **inner join** is used.
* Churned customers are those present earlier but absent in the recent period.
* Churn rate is calculated as:  
    
   **(Churned Customers ÷ Customers Active Before Period) × 100**

### **Interpretation**

* Approximately **27% of previously active customers did not return** in the last 6 months.
* This indicates a **moderate churn level**, highlighting the need for targeted retention and re-engagement strategies.

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

**SQL Query:**

with genre\_sales as(

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

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

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

join artist ar on al.artist\_id=ar.artist\_id

where i.billing\_country='USA'

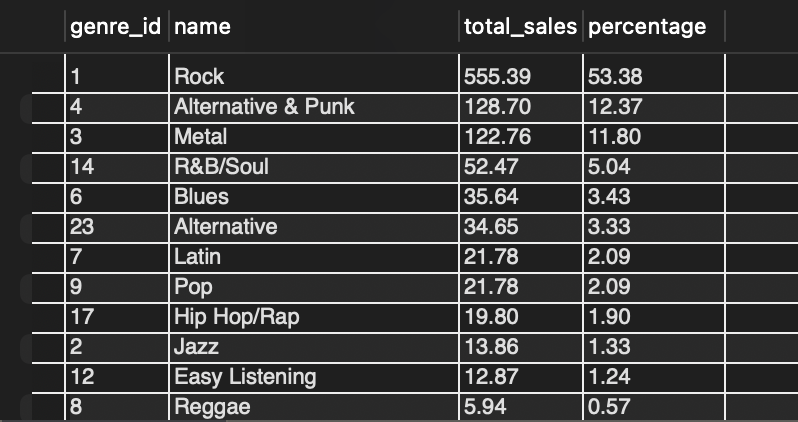
group by g.genre\_id

)

select \*, round((total\_sales\*100.0/(select sum(total) from invoice where billing\_country=’USA’)),2) as percentage

from genre\_sales

order by total\_sales desc;



#### **Key Findings**

* **Rock** is the dominant genre in the USA, contributing **53.38%** of total sales.
* **Alternative & Punk (12.37%)** and **Metal (11.80%)** are the next strongest contributors.
* All other genres individually contribute around <5% indicating a highly concentrated genre preference.

SQL Query for best selling genre and artist :

with artist\_sales 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 total\_rev,

row\_number() 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

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

join artist ar on al.artist\_id=ar.artist\_id

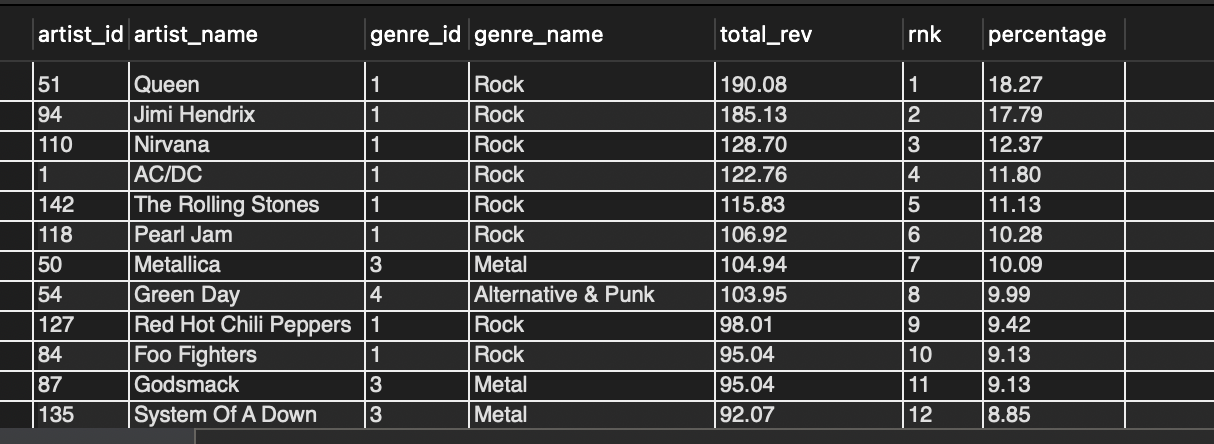
group by ar.artist\_id, ar.name, g.genre\_id, g.name

)

select \*, round((total\_rev\*100.0/(select sum(total) from invoice)),2) as percentage

from artist\_sales

order by total\_rev desc;



‘Group by’ is applied on artist\_id and genre\_id to get the desired result

#### **Key Findings**

* **Queen** is the top-selling artist, contributing **18.27%** of total revenue.
* Other leading artists include **Jimi Hendrix, Nirvana, AC/DC, The Rolling Stones, and Pearl Jam**.
* The majority of top artists belong to the **Rock genre**, reinforcing its dominance.

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

select c.customer\_id, concat(c.first\_name,' ', c.last\_name) as name, count(distinct t.genre\_id) as number\_of\_genres

from customer c

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

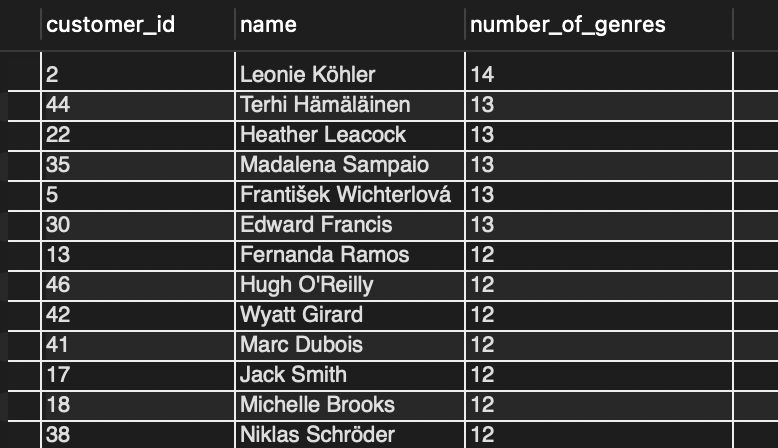
join invoice\_line il on i.invoice\_id=il.invoice\_id

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

group by c.customer\_id, concat(c.first\_name,' ', c.last\_name)

having count(distinct t.genre\_id)>=3

order by number\_of\_genres desc;



1. The results are grouped by customer (customer\_id and customer\_name) to aggregate purchases for each individual.
2. The COUNT(DISTINCT t.genre\_id) function calculates the total number of unique genres from which the customer has purchased tracks.

### **Key Findings**

* Several customers have purchased music across **10+ different genres**, indicating highly diverse listening behavior.
* Customers such as **Leonie Köhler, Terhi Hämäläinen, and Heather Leacock** show the broadest genre engagement.

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

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

dense\_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 t.genre\_id, [g.](http://g.name)name;



* DENSE\_RANK() assigns rankings based on total genre sales, allowing ties to share the same rank.

### **Key Findings**

* **Rock** is the top-performing genre in the USA, ranking **#1** with the highest total sales.
* **Alternative & Punk** and **Metal** rank **#2 and #3**, respectively, forming the second tier of strong performers.

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

**SQL Query:**

with req\_customers as(

select c.customer\_id, concat(c.first\_name,' ', c.last\_name) as Customer\_name

from customer c

left join invoice i on c.customer\_id=i.customer\_id and i.invoice\_date between date\_sub('2020-12-30', interval 3 month) and '2020-12-30'

where i.invoice\_id is null

)

select r.customer\_id, r.Customer\_name, max(i1.invoice\_date) as last\_purchase\_date

from req\_customers r

join invoice i1 on r.customer\_id=i1.customer\_id

where i1.invoice\_date< date\_sub('2020-12-30', interval 3 month)

group by r.customer\_id, r.Customer\_name

order by last\_purchase\_date desc;



* A **LEFT JOIN** is used to identify customers with **no purchases in the last 3 months**.
* Only those are kept where i.invoice\_id IS NULL, thus we get the inactive users

### **Key Findings**

* The output lists customers who have been **inactive for at least 3 months**.
* Also shown their last purchase date.

## 

## 

## 

## 

## **SUBJECTIVE QUESTION-ANSWERS**

## **Q1: 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.**

## **Objective**

To recommend three albums from a *new record label* that should be prioritized for advertising and promotion in the USA, using historical **genre-level sales performance** to guide decision-making.

## **Assumption & Interpretation**

The Chinook database does not contain an explicit *record label* or *release status* attribute. Therefore, for this analysis:

**“New albums” are defined as albums whose tracks have not recorded any historical sales.**

This assumption is necessary given the schema and ensures that recommendations focus on albums with no prior market exposure.

**SQL Query:**

with genre\_data as(

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

sum(il.quantity) as tracks\_sold,

(sum(il.unit\_price\*il.quantity))/sum(il.quantity) as price\_per\_track

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 t.genre\_id, g.name

),

new\_records as(

select t.track\_id, t.name, t.album\_id, t.media\_type\_id, t.genre\_id, t.unit\_price

from track t

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

where il.invoice\_line\_id is null

),

albums\_data as(

select album\_id, genre\_id, count(distinct track\_id) as count\_tracks, sum(unit\_price) as price

from new\_records

group by album\_id , genre\_id

),

indexing as(

select a.album\_id, sum(a.price\*g.genre\_sales) as new\_index

from albums\_data a

join genre\_data g on a.genre\_id=g.genre\_id

group by a.album\_id)

select i.album\_id , al.title as album\_title

from indexing i

join album al on i.album\_id=al.album\_id

order by new\_index desc

limit 3;

## **Approach**

1. **USA Genre Sales Analysis**Sales data was filtered to the USA market and aggregated at the genre level to identify which genres generate the highest historical revenue.
2. **Identification of New Albums**Tracks that have never appeared in the invoice\_line table were identified as unsold. Albums composed entirely of such tracks were treated as *new albums*.
3. **Album-Level Aggregation**New albums were aggregated by album and genre, and their total catalog price (sum of track prices) was calculated.
4. **Promotion Priority Index**Each new album was assigned a **promotion priority index**, computed by weighting the album’s catalog price by its genre’s historical USA sales.  
     
   This index serves as a **relative ranking metric** to align new releases with proven genre demand.
5. **Final Selection**Albums were ranked by this index, and the top three were selected for recommendation.

## **Key Findings**

### **1. Top Genres in the USA (from Genre Sales Analysis)**

Based on historical USA sales, the three highest-performing genres are:

****

These genres clearly dominate the USA market and represent the strongest demand segments.

### **2. New Albums Aligned with High-Performing Genres:** After restricting the analysis to *new albums* (no historical sales) and ranking them using the promotion priority index, the following albums emerged as top candidates:



These albums belong to genres with strong historical demand in the USA, making them suitable candidates for promotional investment despite having no prior sales history.

## **Recommendations:**

### **Prioritize the Following 3 Albums for Promotion in the USA**

1. **Greatest Hits:** This album is associated with one of the top-performing genres in the USA. Promoting it aligns new content with proven market demand.
2. **Compositores:** Despite being unsold historically, its genre-level performance in the USA suggests strong potential when supported by targeted promotion.
3. **King For A Day Fool For A Lifetime:** This album benefits from alignment with a high-revenue genre, making it a strong candidate for launch-focused marketing campaigns.

## **Actionable Strategy**

* **Genre-Led Promotion:** Concentrate marketing spend on these albums within the USA, leveraging the popularity of their respective genres.
* **Targeted Campaigns:** Use genre-specific advertising (e.g., Rock- or Alternative-focused playlists and promotions) to accelerate initial adoption.
* **Risk Mitigation:** By prioritizing new albums tied to historically successful genres, Chinook reduces the uncertainty typically associated with promoting new releases.

## **Conclusion**

In the absence of explicit record-label or release metadata, defining new albums as those without historical sales provides a clear and data-driven framework. By combining this definition with USA genre sales performance, the analysis identifies three new albums—*Greatest Hits*, *Compositores*, and *King For A Day Fool For A Lifetime*—that are best positioned for successful promotion in the USA market.

**Q2: Determine the top-selling genres in countries other than the USA and identify any commonalities or differences.**

## **Objective**

To identify the top-selling music genres in countries **outside the USA** and analyze **commonalities and differences** in genre preferences across international markets, using the USA as a benchmark.

**SQL Query:**

-- USA

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

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 t.genre\_id, g.name

order by genre\_sales desc;

-- other countries

with sales\_data as(

select i.billing\_country, g.name as genre\_name, sum(il.unit\_price\*il.quantity) as genre\_sales\_country

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

),

ranked\_data as(

select billing\_country, genre\_name, genre\_sales\_country,

dense\_rank() over(partition by billing\_country order by genre\_sales\_country desc) as rnk

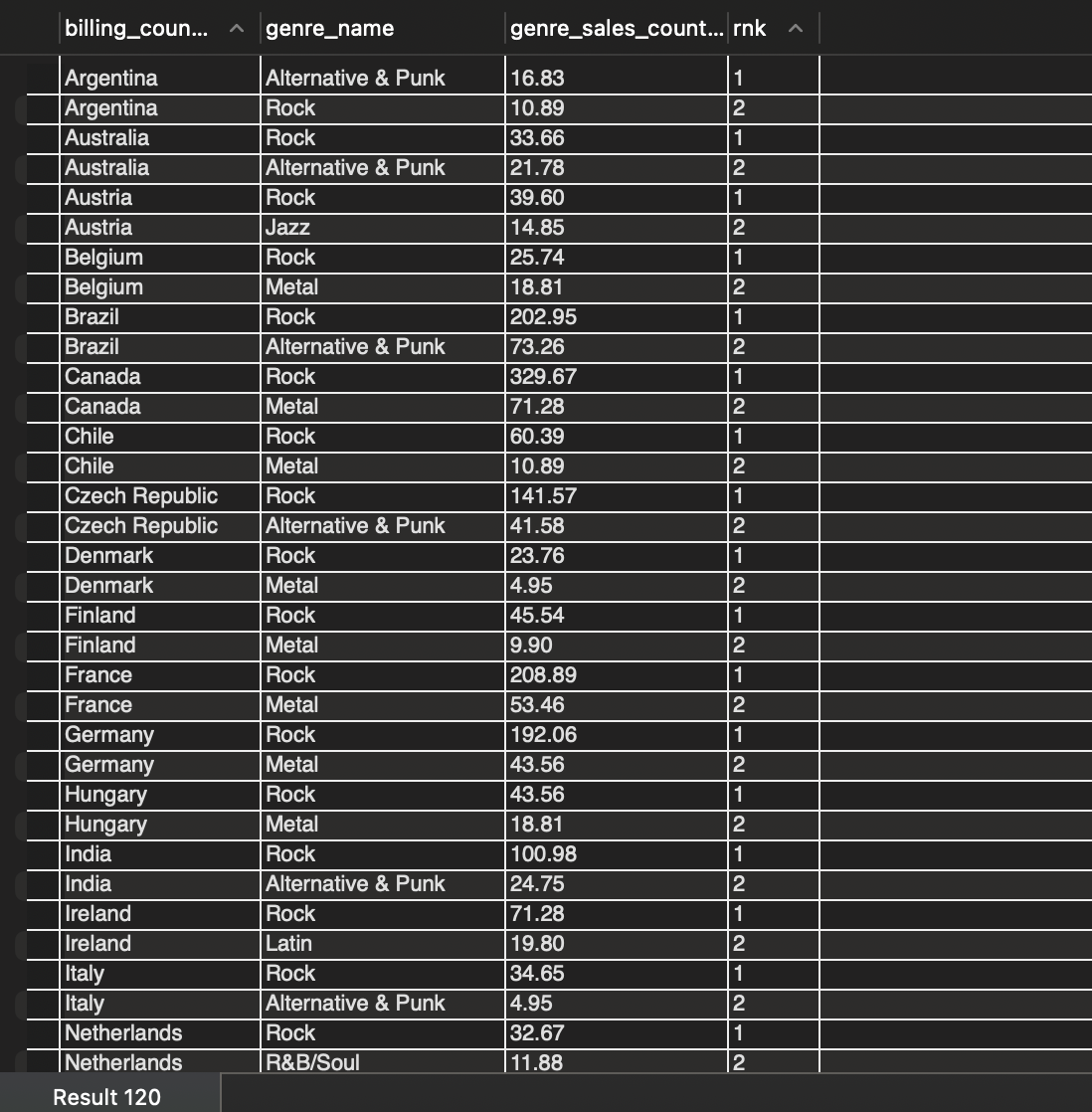
from sales\_data

)

select \*

from ranked\_data

where rnk<=2



## **Approach**

1. **USA Genre Benchmarking**Genre-level sales were first calculated for the USA to establish a reference distribution of genre performance.
2. **Non-USA Country Analysis**Sales data was filtered to exclude the USA and aggregated at the **country–genre** level.
3. **Top Genre Identification per Country**For each non-USA country, genres were ranked by total sales using DENSE\_RANK(), and the **top two genres per country** were selected to capture primary and secondary preferences.
4. **Comparative Evaluation**The resulting dataset was analyzed to identify:
   * Genres that consistently rank at the top across countries (commonalities)
   * Genres that appear only in specific countries or regions (differences)
5. **Visualization for Support**The final country-level top-genre output was exported to Excel and visualized to support interpretation and communication of cross-country patterns.

## **Key Findings**

### **1. USA Genre Landscape (Benchmark)**

The USA market shows a highly concentrated genre structure, led by:

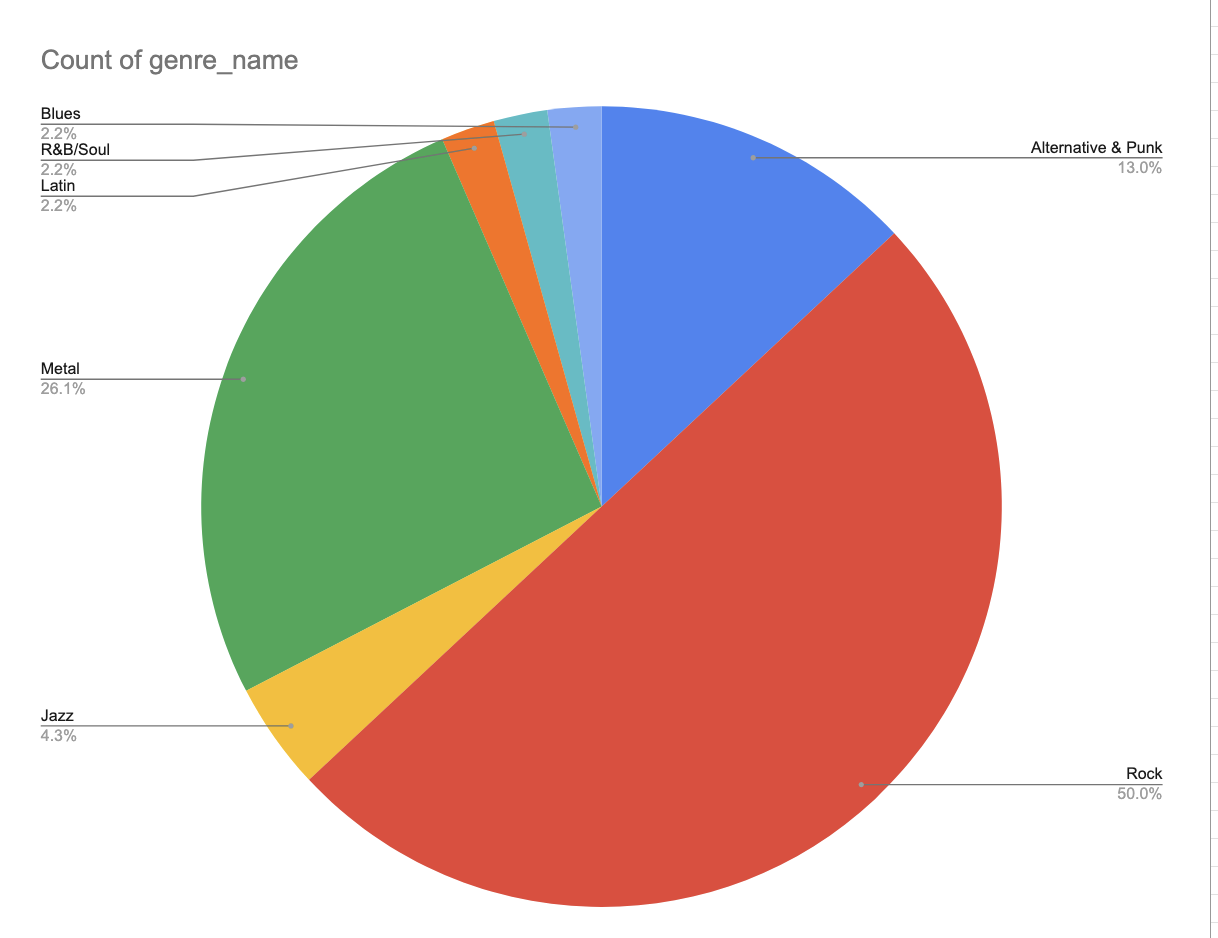
* **Rock** – $555.39
* **Alternative & Punk** – $128.70
* **Metal** – $122.76

These genres form the baseline for international comparison.

### **2. Strong Global Commonality: Rock Dominance**

* SQL analysis shows that **Rock ranks as the #1 genre in 22 out of 23 non-USA countries** analyzed.
* This makes Rock the **most universally dominant genre** across international markets, mirroring its leadership in the USA.

This dominance is further reinforced by the country-level genre distribution visualization, where Rock accounts for the largest share of top-ranked genre appearances



**Figure**: Count of countries where a genre appears in the top 2

### **3. Secondary Genre Preferences Vary Significantly**

While Rock is consistently dominant, **second-ranked genres differ across countries**, indicating localized taste patterns:

* **Metal** frequently appears as the second-ranked genre, particularly across European countries (e.g., Germany, France, Finland, Hungary).
* **Alternative & Punk** emerges as a strong secondary genre in countries such as Argentina, Australia, Brazil, India, and Italy.
* Other genres such as **Jazz, Latin, and R&B/Soul** appear as secondary preferences in a small number of countries, reflecting niche or region-specific appeal.

This variation is also visible in the visualization, where Metal and Alternative & Punk form substantial but clearly secondary shares compared to Rock.

### **4. Difference from the USA Market**

* Both the USA and non-USA markets share **Rock as the primary genre**, indicating strong global alignment at the top level.
* However, non-USA markets exhibit **greater diversity in secondary genre preferences**, whereas the USA shows stronger concentration in its top three genres.

## **Recommendations**

### **1. Global Content Strategy**

* **Prioritize Rock Globally**Since Rock ranks as the **#1 genre in 22 out of 23 non-USA countries**, all major international marketing initiatives—catalog expansion, featured releases, and promotional campaigns—should be heavily weighted toward Rock to **maximize global revenue impact**.
* **Secondary Global Focus**

Secondary genres such as Metal and Alternative & Punk should be promoted selectively based on country-specific preferences rather than through a uniform global campaign.Maintain strong international visibility for **Metal and Alternative & Punk**, as these genres consistently appear as **Rank 2** across most countries, making them reliable secondary revenue drivers.

### **2. Country-Specific Marketing Strategy (Differences)**

* **Argentina Specialization** Argentina stands out as a unique market where **Alternative & Punk ranks above Rock**. Marketing investment and featured album placements should therefore **prioritize Alternative & Punk over Rock** to align with local consumer preferences.
* **Targeted Niche Promotions** In countries where the second-ranked genre is more niche (e.g., **Jazz, Latin, R&B/Soul**), deploy **smaller, low-cost digital campaigns** targeted at those segments. This enables the company to capture localized demand without diluting global marketing resources.

## **Conclusion**

Across countries outside the USA, **Rock overwhelmingly dominates as the top-selling genre**, ranking first in 22 out of 23 markets analyzed. This highlights a strong global commonality in primary genre preference. However, meaningful differences emerge at the secondary genre level, where regional and cultural factors influence music consumption. A combined strategy—global consistency for Rock and localized targeting for secondary genres—best aligns with the observed international purchasing behavior.

## **Question 3: 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?**

## **Objective**

To compare the purchasing habits of **long-term customers** and **new customers** across key behavioral dimensions—**purchase frequency, basket size, and spending**—and derive insights that inform **customer loyalty and retention strategies**.

SQL Query:

with overall\_data as(

select

c.customer\_id, c.first\_name, c.last\_name, i.invoice\_id, i.invoice\_date,

il.invoice\_line\_id, il.track\_id, il.unit\_price, il.quantity,

(il.unit\_price \* il.quantity) as total

from customer c

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

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

),

customers\_data as(

select customer\_id,

min(invoice\_date) as first\_purchase\_date,

max(invoice\_date) as recent\_purchase\_date,

datediff(max(invoice\_date), min(invoice\_date)) as purchasing\_period,

count(distinct invoice\_id) as count\_of\_invoices,

sum(total) as total\_amount\_spent

from overall\_data

group by customer\_id

),

categorised\_final\_data as (

select \*,

case

when first\_purchase\_date < '2017-12-31' and purchasing\_period > 365 \* 2 then 'long term customer'

when first\_purchase\_date > '2017-12-31' then 'new customer'

else null

end as customer\_category

from customers\_data

),

basket\_sizes as (

select

c.customer\_id, o.invoice\_id,

count(distinct o.track\_id) as basket\_size,

sum(o.total) as basket\_price

from categorised\_final\_data c

join overall\_data o on c.customer\_id = o.customer\_id

group by c.customer\_id, o.invoice\_id

),

avg\_basket\_sizes as (

select

customer\_id,

avg(basket\_size) as avg\_basket\_size

from basket\_sizes

group by customer\_id

),

parameters\_compared as (

select

c.customer\_id, c.customer\_category, b.avg\_basket\_size,

round((count\_of\_invoices \* 365.0 / purchasing\_period), 2) as frequency\_yearly,

total\_amount\_spent,

(total\_amount\_spent / count\_of\_invoices) as avg\_order\_value

from categorised\_final\_data c

join avg\_basket\_sizes b

on c.customer\_id = b.customer\_id

)

select customer\_category,

count(\*) as count\_customers,

round(avg(avg\_basket\_size), 2) as avg\_cat\_basket\_size,

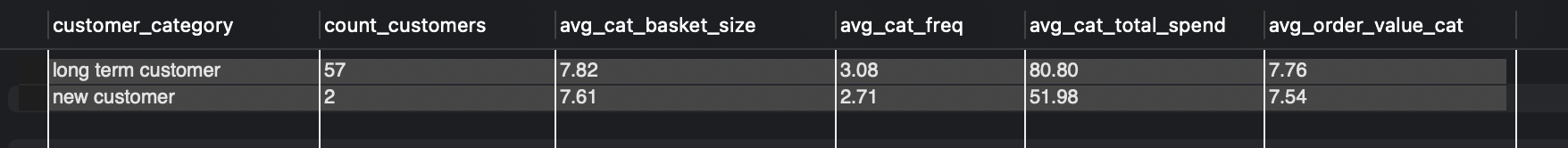
round(avg(frequency\_yearly), 2) as avg\_cat\_freq,

round(avg(total\_amount\_spent), 2) as avg\_cat\_total\_spend,

round(avg(avg\_order\_value), 2) as avg\_order\_value\_cat

from parameters\_compared

group by customer\_category;



## **Approach**

## **Customer Classification Criteria:**

**Long-term customers -** Customers whose:

* + **First purchase date** occurred before **31 December 2017**, and
  + **Purchasing period** (difference between first and most recent purchase) exceeded **two years (730 days)**.

**New customers-** Customers whose:

* + **First purchase date** occurred after **31 December 2017**.

This classification ensures that long-term customers represent **sustained engagement over time**, while new customers reflect **recently acquired users**.

1. **Metric Construction (Customer-Level First)** For each customer, the following metrics were calculated:
   * **Average basket size** (distinct tracks per invoice)
   * **Purchase frequency** (annualized, normalized by customer purchasing period)
   * **Total lifetime spend**
   * **Average order value**
2. **Segment-Level Comparison:** Customer-level metrics were then aggregated by customer category (long-term vs new) to ensure a fair, behavior-focused comparison.
3. **Visualization** Bar charts were created to visually compare these metrics.

## 

## **Key Findings**

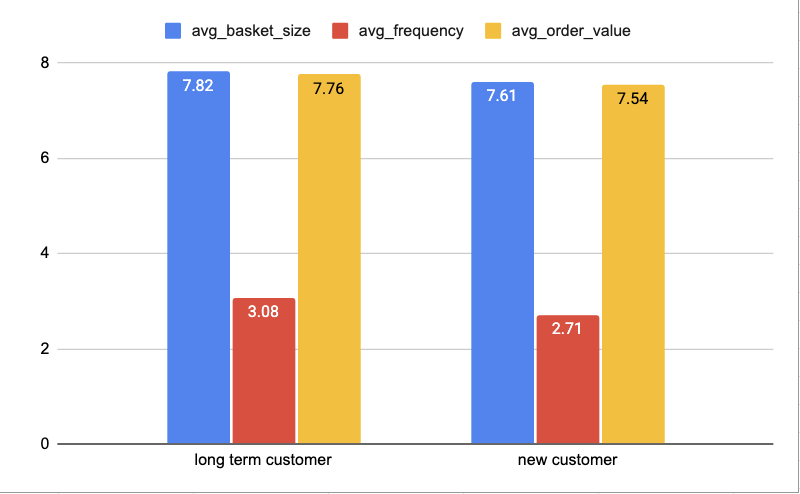
### **1. Customer Base Composition**

* The customer base is heavily skewed toward **long-term customers (57)**, with only **2 new customers**.
* This highlights that Chinook’s revenue is currently driven primarily by an **established customer base**, rather than continuous inflow of new users.

### **2. Basket Size & Order Value Are Similar**

* **Average basket size** is nearly identical:
  + Long-term customers: **7.82**
  + New customers: **7.61**
* **Average order value** is also very similar:
  + Long-term: **7.76**
  + New: **7.54**

**Insight:** Once customers decide to make a purchase, **order composition does not vary significantly by tenure**. New customers behave similarly to long-term customers at the transaction level.

**

**Figure**: KPIs(Avg Basket Size, Avg Frequency, Avg Order Value) comparison between the two customer segments(long term and new)

### **3. Purchase Frequency Is Higher for Long-Term Customers**

* **Annualized purchase frequency**:
  + Long-term customers: **3.08**
  + New customers: **2.71**

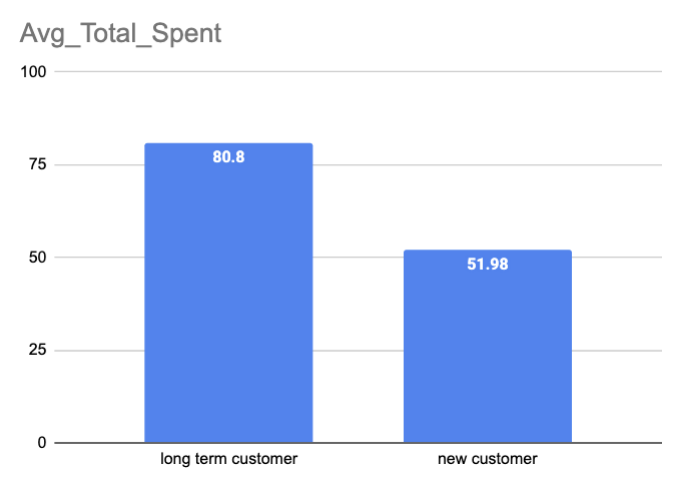
**Insight:** Customer loyalty manifests primarily through **repeat purchasing behavior**, not larger baskets or higher order values.

**4. Lifetime Spend Shows the Largest Gap**

* **Average total spend per customer**:
  + Long-term customers: **80.80**
  + New customers: **51.98**

**Insight:**

Higher lifetime value among long-term customers is driven by **repeat engagement over time**, rather than higher spend per transaction.

**

**Figure**: Comparison of Avg. Total Amount Spent by Customers(Long term Vs New)

## **What This Tells Us About Loyalty**

* Loyalty is **behavioral**, not transactional.
* Long-term customers are not spending more per order—they are **coming back more often**.
* New customers already show **healthy purchasing patterns**, suggesting strong early product-market fit.

## **Retention & Growth Recommendations**

### **1. Strengthen Retention Over Basket Expansion**

* Since basket size and order value are already stable across segments, efforts should focus on:
  + Increasing **purchase frequency**
  + Reducing time between purchases

Examples: Reminder emails, Personalized recommendations, Time-based incentives

### **2. Accelerate New-Customer Conversion to Long-Term**

* New customers show comparable basket behavior but lower frequency.
* Early-life engagement strategies (first 90 days) are critical.

Examples:

* Onboarding offers
* Loyalty point multipliers for early repeat purchases
* Curated playlists or bundles after first purchase

### **3. Protect Long-Term Customer Value**

* Given their significantly higher lifetime spend, long-term customers should be:
  + Prioritized for loyalty programs
  + Offered exclusive content or early access
  + Actively monitored for drops in purchase frequency (early churn signals)

## **Conclusion**

The key difference between long-term and new customers lies not in how much they spend per purchase, but in **how often they return**. Long-term customers generate higher lifetime value through consistent repeat behavior, while new customers already demonstrate strong transactional intent. Retention strategies should therefore prioritize **frequency-building mechanisms**, ensuring new customers are converted into long-term, loyal buyers as early as possible.

**Q4: 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?**

### **Objective**

To identify **music genres, artists, and albums that are frequently purchased together** and to understand how these co-purchase patterns can be leveraged for **product recommendations and cross-selling initiatives**.

SQL Query:

-- genre\_pairs

with combined\_data as(

select

il.invoice\_id, il.track\_id,

t.genre\_id, g.name as genre\_name

from invoice\_line il

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

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

),

genre\_pairs as(

select c1.invoice\_id,

c1.genre\_id as genre1\_id, c1.genre\_name as genre1\_name,

c2.genre\_id as genre2\_id, c2.genre\_name as genre2\_name

from combined\_data c1

join combined\_data c2 on c1.invoice\_id=c2.invoice\_id and c1.genre\_id<c2.genre\_id

)

select genre1\_id, genre1\_name,

genre2\_id, genre2\_name, count(distinct invoice\_id) as together\_occur\_in\_invoices,

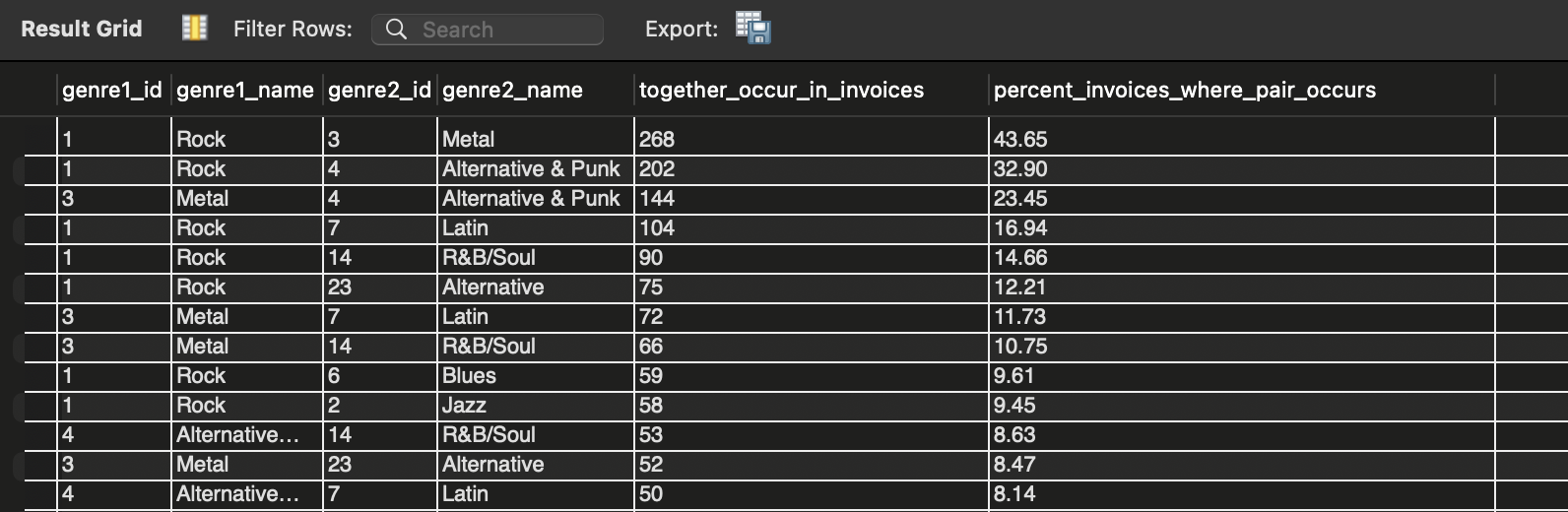
round(count(distinct invoice\_id)\*100.0/(select count(distinct invoice\_id) from invoice\_line),2) as percent\_invoices\_where\_pair\_occurs

from genre\_pairs

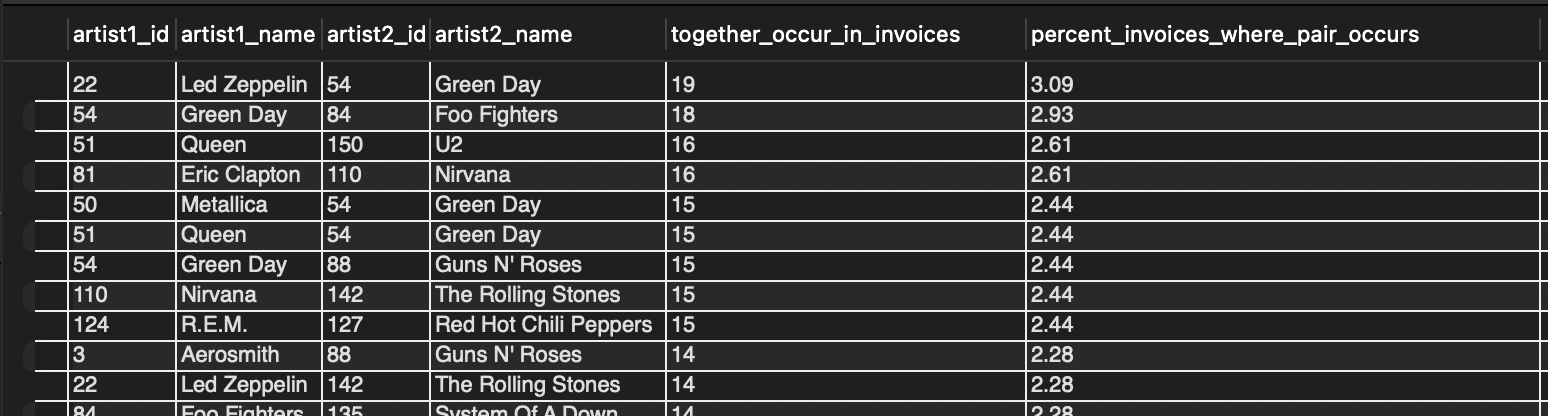
group by genre1\_id, genre2\_id

order by together\_occur\_in\_invoices desc;

Similarly for album\_pairs and artist\_pairs.







**Approach**

Invoices were treated as individual purchase baskets. Items appearing together within the same invoice were paired to identify co-occurrence patterns. Three levels of affinity analysis were performed:

1. **Genre-level affinity** to capture broad listening behavior.
2. **Album-level affinity** to identify specific product bundling opportunities.
3. **Artist-level affinity** to uncover cross-artist listening patterns.  
     
    For each pair, the number of invoices in which both items appeared together was calculated, along with the percentage of total invoices where the pair occurred.

## **Key Findings**

### **1. Genre Affinity Patterns**

Genre-level analysis reveals very strong and consistent co-purchase behavior centered around Rock:

* **Rock & Metal** is the most frequent genre pair, co-occurring in **268 invoices (43.65%)**.
* **Rock & Alternative & Punk** appears in **202 invoices (32.90%)**.
* **Metal & Alternative & Punk** follows with **144 invoices (23.45%)**.
* Several secondary combinations such as Rock with Latin, R&B/Soul, Blues, and Jazz also show meaningful overlap, though at lower frequencies.

**Insight:** Rock acts as a central “anchor genre,” frequently purchased alongside multiple other genres. Customers who buy Rock tracks are highly likely to explore adjacent genres, particularly Metal and Alternative & Punk.

### **2. Album Co-Purchase Patterns**

Album-level affinity highlights specific cross-selling opportunities:

* **‘Are You Experienced?’ & ‘Mezmerize’** is the most common album pair, appearing together in **14 invoices (2.28%)**.
* Multiple greatest-hits and classic albums frequently co-occur, such as:
  + *Dark Side of the Moon & The Singles*
  + *Mezmerize & The Police Greatest Hits*
  + *Big Ones* paired with several popular rock albums

**Insight:** Customers tend to bundle **iconic or greatest-hits albums** together, indicating a strong preference for curated, familiar collections rather than isolated album purchases.

### **3. Artist Affinity Patterns**

Artist-level analysis shows strong cross-artist listening behavior:

* **Led Zeppelin & Green Day** is the most frequent artist pair with **19 co-occurrences (3.09%)**.
* Other highly recurring combinations include:  
  + Green Day & Foo Fighters
  + Queen & U2
  + Eric Clapton & Nirvana
  + Green Day & Guns N’ Roses
  + Nirvana & The Rolling Stones

**Insight:** Customers frequently mix **classic rock and modern rock artists** within the same purchase, suggesting that listening behavior spans generations and styles rather than being artist-exclusive.

## **Recommendations:**

### **Product Recommendations**

* Recommend **Metal and Alternative & Punk** tracks immediately after Rock purchases.
* Suggest complementary artists (e.g., Green Day → Foo Fighters, Queen → U2) based on strong artist-level affinity.

### **Cross-Selling & Bundling**

* Create **Rock-centric bundles** that include Metal and Alternative & Punk tracks.
* Promote **greatest-hits album bundles**, as these albums frequently appear together in purchases.

### **Playlist & Merchandising Strategy**

* Curate playlists that combine classic and modern rock artists to mirror observed co-listening behavior.
* Feature album pairings prominently in “Customers Also Bought” sections to increase basket size.

## **Conclusion**

The affinity analysis shows that customer purchasing behavior is highly interconnected, with Rock serving as a central genre that bridges multiple styles, artists, and albums. Strong genre-level, artist-level, and album-level co-purchase patterns indicate clear opportunities for intelligent recommendations, curated bundles, and targeted cross-selling strategies. Leveraging these affinities can significantly enhance customer experience while increasing average order value and overall revenue.

**Q5: Do customer purchasing behaviors and churn rates vary across different geographic regions? How might these correlate with local demographic or economic factors?**

### **Objective**

To examine whether **customer purchasing behavior** and **churn rates** vary across geographic regions (billing countries) and to interpret how these differences may relate to **market maturity, customer concentration, and local economic or demographic factors**.

**SQL Query:**

-- purchasing behaviors as per location

select i.billing\_country,

count(distinct i.invoice\_id) as total\_transactions,

count(distinct i.customer\_id) as customer\_count,

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

round(count(il.invoice\_line\_id)/count(distinct il.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;



-- churn rates as per location

with earlier\_customers as(

select billing\_country, customer\_id, count(distinct invoice\_id) as number\_purchases

from invoice

where invoice\_date< date\_sub('2020-12-30', interval 6 month)

group by billing\_country, customer\_id

),

last6\_months\_customers as(

select billing\_country, customer\_id, count(distinct invoice\_id) as number\_purchases

from invoice

where invoice\_date between date\_sub('2020-12-30', interval 6 month) and '2020-12-30'

group by billing\_country,customer\_id

),

churned\_customers as(

select e.billing\_country,e.customer\_id

from earlier\_customers e

left join last6\_months\_customers l on e.customer\_id=l.customer\_id and e.billing\_country=l.billing\_country

where l.billing\_country is null

),

countrywise\_count\_churned as(

select billing\_country, count(\*) as n\_cust\_churned

from churned\_customers

group by billing\_country

),

countrywise\_count\_earlier as(

select billing\_country, count(\*) as n\_cust\_in\_earlier\_period

from earlier\_customers

group by billing\_country

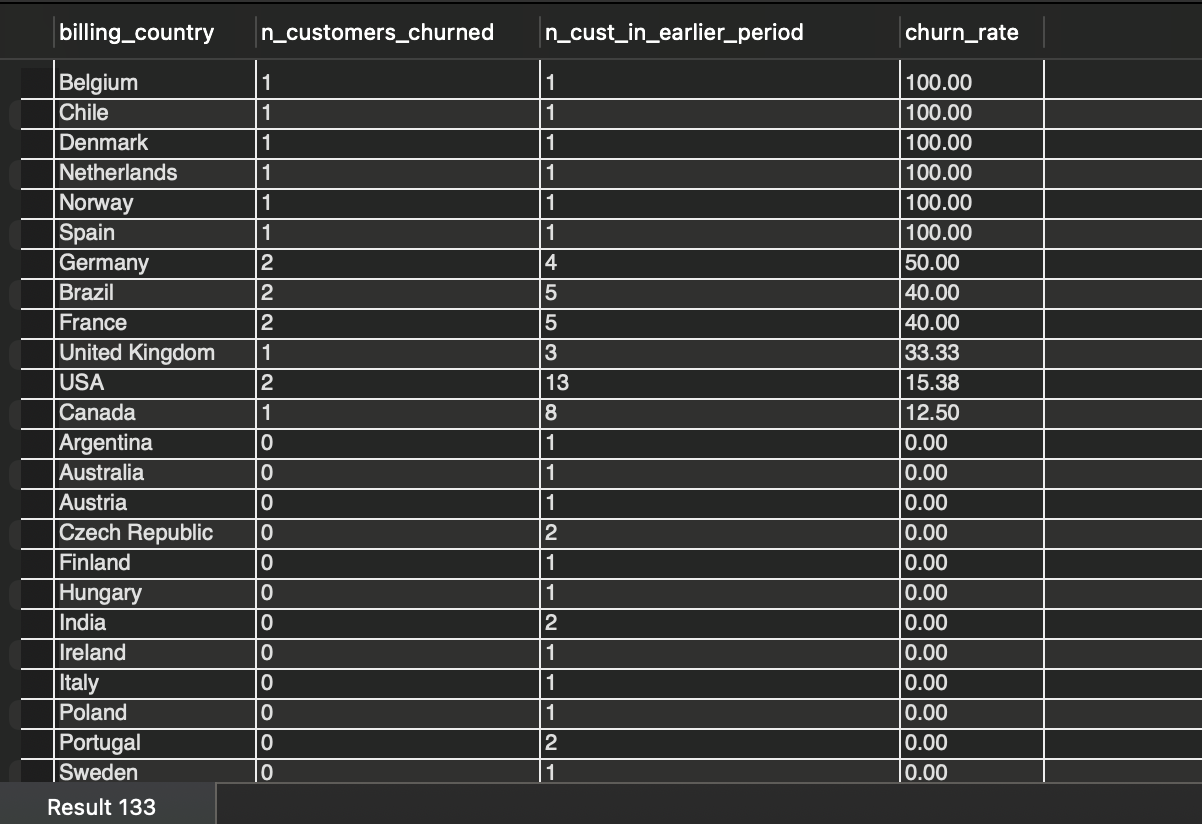
)

select ce.billing\_country, coalesce(n\_cust\_churned,0) as n\_customers\_churned, n\_cust\_in\_earlier\_period, coalesce(round((n\_cust\_churned\*100.0/n\_cust\_in\_earlier\_period),2),0) as churn\_rate

from countrywise\_count\_earlier ce

left join countrywise\_count\_churned cc on ce.billing\_country=cc.billing\_country

order by churn\_rate desc, billing\_country asc;



A customer is defined as **churned** if:

* They made at least one purchase **before 30 June 2025** (i.e., earlier than 6 months from the reference date), and
* They made **no purchases between 30 June 2025 and 30 December 2025**, where **30 December 2025** is the latest invoice date in the dataset.

## **Approach**

1. **Purchasing Behavior Analysis** For each country, the following metrics were calculated:
   * Total transactions
   * Number of unique customers
   * Average sales per customer
   * Average basket size (invoice lines per transaction)
2. **Churn Rate Analysis**Customers were segmented by country and evaluated for churn based on the above definition, producing country-level churn rates.
3. **Comparative Interpretation**Purchasing intensity and churn rates were compared across countries to identify patterns related to customer concentration and market stability.

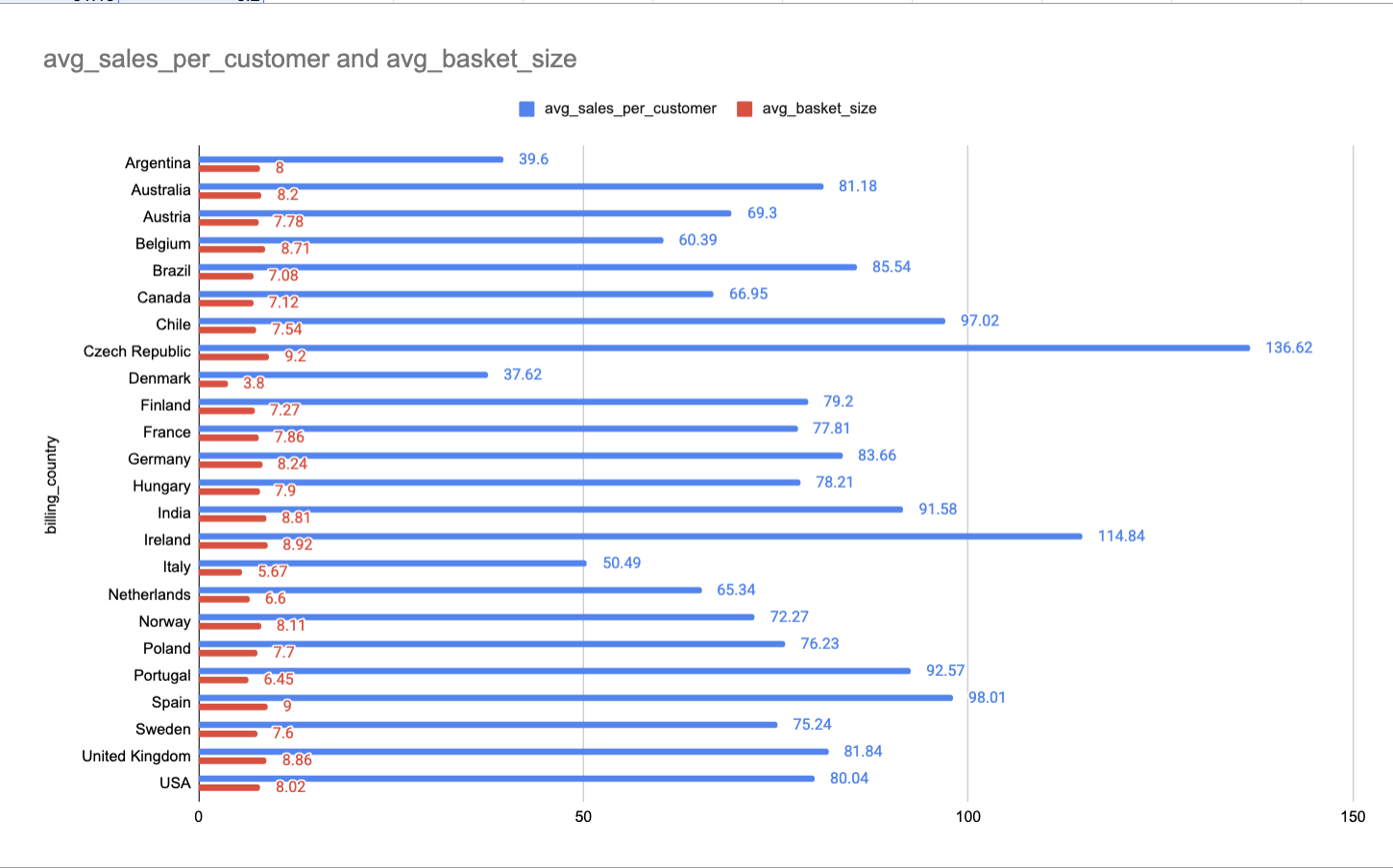
## **Key Findings**

### **1. Regional Differences in Purchasing Behavior**

* **High-volume markets** such as the **USA, Canada, Brazil, France, and Germany** show:  
  + Higher transaction counts
  + Larger customer bases
  + Stable average basket sizes (generally between 7–9 items)
* **Smaller markets** (e.g., Belgium, Denmark, Norway, Spain) typically have:  
  + One or very few customers
  + Moderate to high average basket sizes
  + Lower transaction volumes

**Key Insight:**

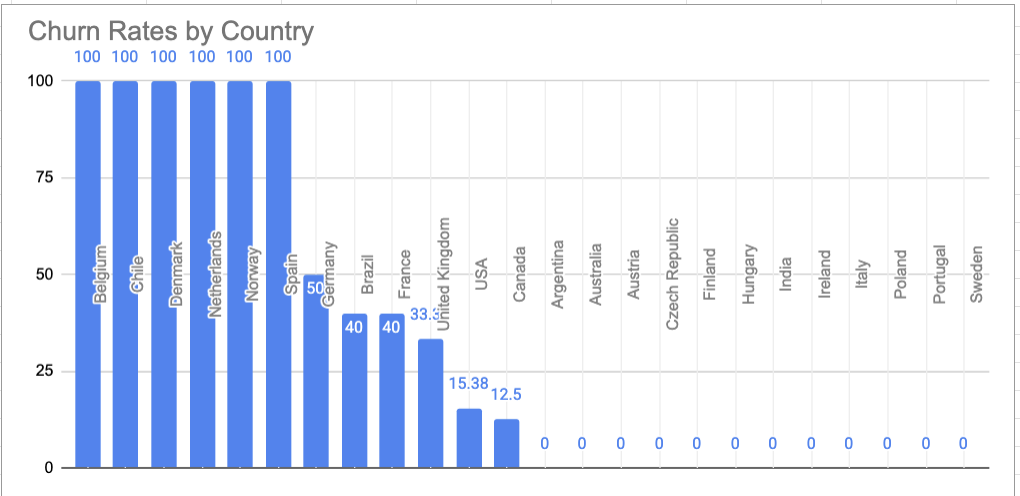
* Average basket size remains consistent across regions (typically 7–9 items).
* Significant variation exists in average sales per customer across countries.
* This indicates that **revenue differences are driven by customer volume and repeat purchases, not basket expansion**.

  
 **Figure:** Comparison of average sales per customer and average basket size across different billing countries.

While basket size remains relatively consistent across regions, **total spend and transaction volume scale with customer base size**, indicating higher engagement in mature markets.

### **2. Significant Regional Variation in Churn Rates**

* Several countries show **100% churn rates** (Belgium, Chile, Denmark, Netherlands, Norway, Spain).   
   These countries all have **only one active customer in the earlier period**, making churn highly sensitive to single-customer inactivity.



**Figure**: Comparison of Churn Rates by Country

* **Large markets** exhibit **lower churn rates**:
  + USA: **15.38%**
  + Canada: **12.50%**
  + United Kingdom: **33.33%**
* Many countries (e.g., India, Australia, Czech Republic, Finland, Portugal) show **0% churn**, reflecting continued engagement from their limited customer bases.

**Insight:** Churn rates are **inversely related to customer base size**. Smaller markets appear volatile due to low customer counts, while larger markets demonstrate greater stability.

### **3. Relationship Between Purchasing Behavior and Churn**

* Countries with **higher transaction density and more customers** tend to have **lower churn rates**, suggesting stronger customer retention.
* High churn in small markets does not necessarily indicate poor performance, but rather **data sparsity and customer concentration risk**.

**Correlation with Demographic and Economic Factors**

* **Mature markets** (e.g., USA, Canada, Western Europe) likely benefit from:
  + Higher digital adoption
  + Greater familiarity with music platforms  
    More consistent purchasing habits
* **Emerging or niche markets** may experience:
  + Irregular purchasing patterns
  + Higher churn volatility due to limited customer bases
  + Greater sensitivity to individual customer behavior

## **Recommendations**

* **Market-Specific Retention Strategy**
  + Focus retention efforts on **large, stable markets** where churn reduction yields meaningful revenue impact.
  + Treat churn metrics in low-customer countries cautiously; prioritize **acquisition over retention** in these regions.
* **Resource Allocation**
  + Invest marketing and loyalty programs more heavily in regions with demonstrated repeat purchasing behavior.
  + Use targeted, low-cost campaigns for smaller markets to reduce customer concentration risk.
* **Churn Monitoring**
  + Churn should be interpreted **in context of customer base size**, especially when comparing regions.

## **Conclusion**

Customer purchasing behavior and churn rates vary significantly across geographic regions. Larger, mature markets show higher transaction volumes and lower churn, indicating stable customer engagement. In contrast, smaller markets display high churn volatility driven primarily by low customer counts rather than disengagement. Regional strategies should therefore balance churn management with customer base size, market maturity, and acquisition potential.

**Q6 : 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?**

### **Objective**

To identify **customer segments that are more likely to churn or reduce spending**, based on customer purchase behavior and location, and to understand the **key factors contributing to customer risk**. This analysis helps prioritize retention efforts and design targeted intervention strategies.

**SQL query:**

with customer\_spending as(

select c.customer\_id, i.billing\_country,

sum(i.total) as total\_spent,

round(sum(i.total)/count(distinct i.invoice\_id),2) as avg\_spend\_per\_transaction,

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

count(distinct i.invoice\_id) as n\_transactions

from customer c

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

group by c.customer\_id, i.billing\_country

),

basket\_size as(

select i.customer\_id, i.invoice\_id, count(distinct il.invoice\_line\_id) as basket\_s

from invoice i

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

group by i.customer\_id, i.invoice\_id

),

avg\_basket\_size as(

select customer\_id, round(avg(basket\_s),2) as average\_basket\_size

from basket\_size

group by customer\_id

),

combined\_data as(

select cs.customer\_id, cs.billing\_country, cs.total\_spent, cs.avg\_spend\_per\_transaction,

cs.last\_purchase\_date, datediff('2020-12-30',cs.last\_purchase\_date) as days\_since\_last\_purchase,

cs.n\_transactions as total\_purchases, bs.average\_basket\_size,

ntile(3) over(order by cs.n\_transactions) as freq\_tile,

ntile(3) over(order by cs.total\_spent) as spend\_tile

from customer\_spending cs

join avg\_basket\_size bs on cs.customer\_id=bs.customer\_id

),

profiling as(

select customer\_id, billing\_country, total\_spent, avg\_spend\_per\_transaction, last\_purchase\_date,

days\_since\_last\_purchase, total\_purchases, average\_basket\_size,

case

when days\_since\_last\_purchase>180 and freq\_tile = 1 then 'Critical Risk'

when days\_since\_last\_purchase>180 then 'High Risk(churn risk)'

when freq\_tile = 1 then 'Medium Risk(low freq risk)'

when spend\_tile = 1 then 'Low Spender(value risk)'

else 'Low Risk'

end as Risk\_profiling

from combined\_data)

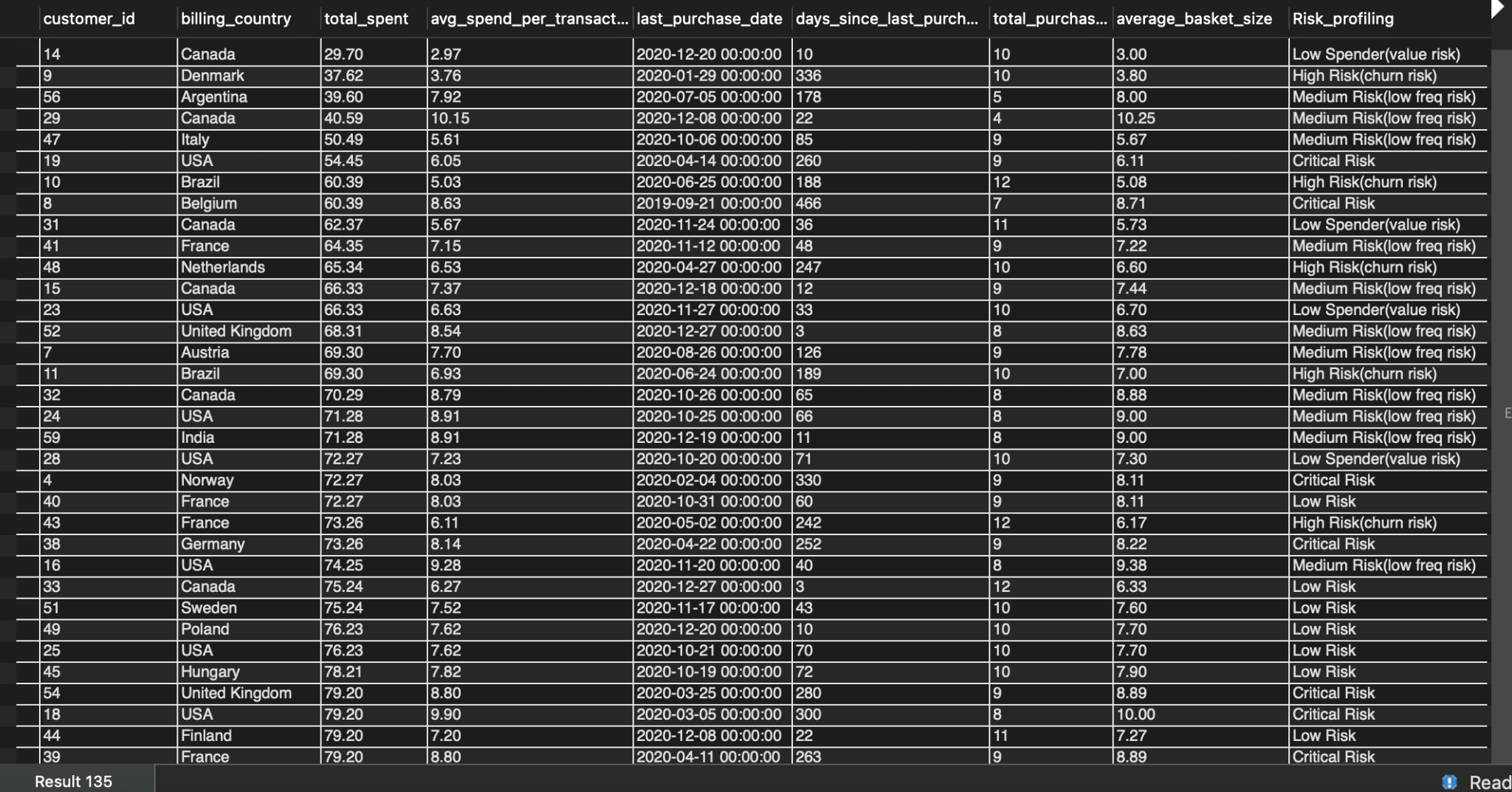
select Risk\_profiling, avg(total\_spent) as avg\_total\_spent\_of\_profile, avg(avg\_spend\_per\_transaction) as avg\_spent\_per\_transac\_of\_profile,

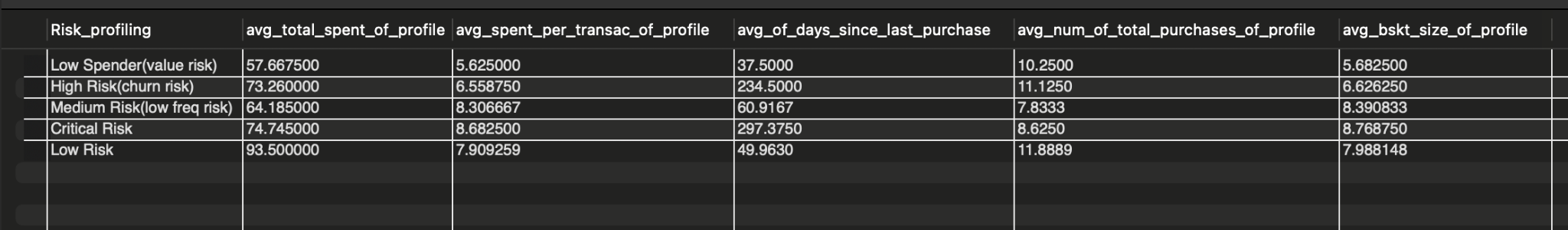
avg(days\_since\_last\_purchase) as avg\_of\_days\_since\_last\_purchase, avg(total\_purchases) as avg\_total\_purchases\_of\_profile,

avg(average\_basket\_size) as avg\_bskt\_size\_of\_profile

from profiling

group by Risk\_profiling;





## 

## **Approach**

1. **Behavioral Feature Construction {RFM Framework(Recency, Frequency, Monetary Value)}** Customer profiles were built using purchase history and engagement metrics:  
   * **Recency**: Days since last purchase (calculated using 30 December 2020 as the reference date).
   * **Frequency**: Total number of purchases per customer.
   * **Monetary Value**: Total lifetime spend per customer.  
     **Basket Size**: Average number of items purchased per transaction.
2. **Relative Segmentation Using Quantiles** To avoid arbitrary thresholds, customers were ranked using:  
   * NTILE(3) on **purchase frequency** (low → high engagement)
   * NTILE(3) on **total spend** (low → high value)
3. **Risk Classification Logic (RFM-based)** Customers were classified into risk segments using a **Recency → Frequency → Monetary** prioritization:  
   * **Critical Risk**: Inactive for over 180 days and in the lowest frequency tier.
   * **High Risk (Churn Risk)**: Inactive for over 180 days regardless of spend.
   * **Medium Risk (Low Frequency Risk)**: Active recently but in the lowest frequency tier.
   * **Low Spender (Value Risk)**: Active and frequent but in the lowest spending tier.
   * **Low Risk**: Active, frequent, and higher-spending customers.

## **Aggregated Risk Segment Characteristics (from Output)**

### **1. Low Risk (Best Segment):**

* **Highest avg total spend:** ~93.5
* **High purchase frequency:** ~11.9 purchases
* **Recent activity:** ~50 days since last purchase
* **Healthy basket size:** ~8 items

**Interpretation:** These customers are engaged, spend consistently, and purchase frequently. They represent the **most stable profile.**

### **2. Critical Risk**

* **Very high inactivity:** ~297 days since last purchase
* **Moderate historical spend:** ~74.7
* **Lower purchase frequency:** ~8.6

**Interpretation:** These customers were previously valuable but are now deeply disengaged. They pose the **highest churn risk with the highest potential loss**.

### **3. High Risk (Churn Risk)**

* **High inactivity:** ~234 days
* **Reasonable purchase count:** ~11.1
* **Moderate spend:** ~73.3

**Interpretation:** Customers in this group still have strong historical engagement but have recently stopped purchasing. This indicates **early-to-mid churn**, where intervention can still be effective.

### **4. Medium Risk (Low Frequency Risk)**

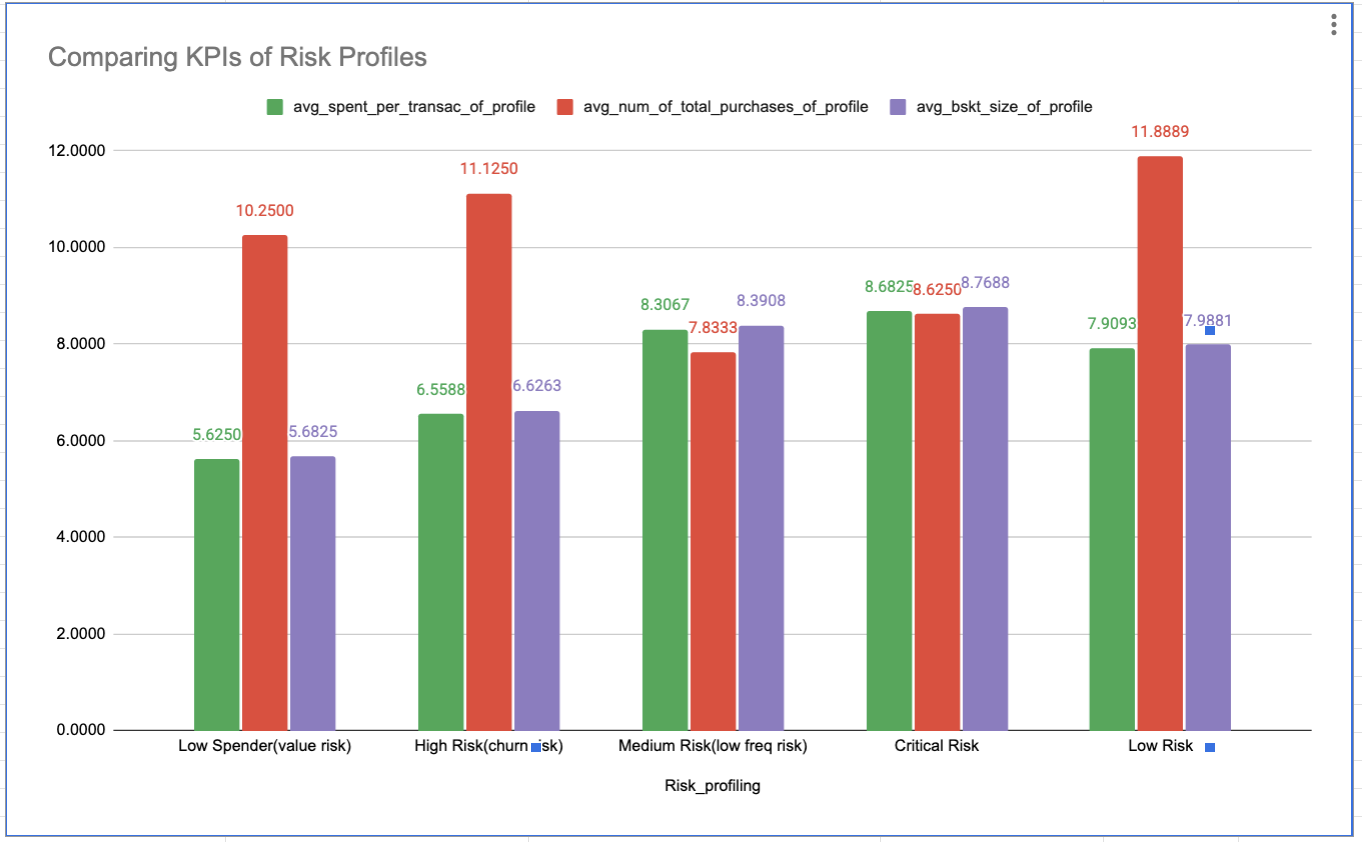
* **Moderate recency:** ~61 days
* **Lower purchase frequency:** ~7.8
* **Decent basket size:** ~8.4

**Interpretation:**These customers are active but purchase infrequently. Their risk is not churn yet, but **revenue stagnation** due to low repeat behavior.

### **5. Low Spender (Value Risk)**

* **Lowest avg total spend:** ~57.7
* **Recent activity:** ~37 days
* **High purchase count but small baskets:** ~5.7 items

**Interpretation:** These customers are active but generate limited revenue per transaction. The risk here is **low CLV**, not churn.



**Figure:** Comparison of purchase frequency, transaction value, and basket size across customer risk profiles.

* **Low Risk customers** show the highest purchase frequency and strong spending consistency, confirming them as the most stable segment.
* **Critical and High Risk customers** have similar average spend per transaction but sharply lower recency and frequency, indicating disengagement rather than reduced willingness to spend.
* **Medium Risk customers** remain active but exhibit noticeably lower purchase frequency, signaling early revenue stagnation.
* **Basket size remains relatively stable across all risk profiles**, reinforcing that churn risk is driven by engagement decline, not order depth.

**Key Factors Driving Customer Risk**

1. **Recency is the strongest churn signal**Risk increases sharply once inactivity crosses ~180 days, regardless of past spend.
2. **Frequency differentiates stable vs risky active customers**Active customers with low frequency drift into medium risk quickly.
3. **Basket size is not the primary risk driver**Basket sizes remain relatively stable across segments; engagement matters more than order depth.
4. **Historical value does not guarantee retention**High and Critical Risk groups still show decent historical spend, proving disengagement—not price sensitivity—is the issue.

## **Business Recommendations**

1. **Low Risk → Retention & Loyalty**
   * Loyalty programs, early access, exclusive releases
   * Goal: protect high-CLV customers
2. **Medium Risk → Frequency Uplift**
   * Personalized recommendations, reminder campaigns
   * Goal: increase repeat purchases
3. **High & Critical Risk → Win-Back Campaigns**
   * Time-bound offers, genre-based personalization
   * Goal: reactivate historically valuable customers before permanent churn
4. **Low Spender → Value Expansion**
   * Bundles, cross-sell strategies
   * Goal: improve basket value and CLV

## **Conclusion**

Customer churn and revenue risk are primarily driven by **behavioral engagement patterns**, with **recency and purchase frequency** being the strongest indicators of risk. While geography influences customer concentration, it is not the dominant driver of churn. An RFM-based risk framework enables Chinook to differentiate between customers who need **urgent retention**, those requiring **engagement nudges**, and those suitable for **revenue expansion**, allowing for more efficient and targeted customer lifecycle management.

**Q7: 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?**

**Objective**

To leverage customer tenure, purchase history, and engagement metrics to approximate **customer lifetime value (CLV**), segment customers into meaningful value groups, and identify common behavioral patterns among customers who have stopped purchasing. The insights are used to inform targeted marketing, retention, and loyalty strategies.

**SQL query:**

with customer\_spending as(

select c.customer\_id, i.billing\_country,

sum(i.total) as total\_spent,

round(sum(i.total)/count(distinct i.invoice\_id),2) as avg\_spend\_per\_transaction,

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

count(distinct i.invoice\_id) as n\_transactions

from customer c

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

group by c.customer\_id, i.billing\_country

),

basket\_size as(

select i.customer\_id, i.invoice\_id, count(distinct il.invoice\_line\_id) as basket\_s

from invoice i

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

group by i.customer\_id, i.invoice\_id

),

avg\_basket\_size as(

select customer\_id, round(avg(basket\_s),2) as average\_basket\_size

from basket\_size

group by customer\_id

),

combined\_data as(

select cs.customer\_id, cs.billing\_country, cs.total\_spent, cs.avg\_spend\_per\_transaction,

cs.last\_purchase\_date, datediff('2020-12-30',cs.last\_purchase\_date) as days\_since\_last\_purchase,

cs.n\_transactions as total\_purchases, bs.average\_basket\_size,

ntile(3) over (order by cs.total\_spent) as total\_spent\_tile,

ntile(3) over (order by cs.avg\_spend\_per\_transaction) as avg\_spend\_tile,

ntile(3) over (order by cs.n\_transactions) as purchase\_freq\_tile,

ntile(3) over (order by bs.average\_basket\_size) as basket\_size\_tile

from customer\_spending cs

join avg\_basket\_size bs on cs.customer\_id=bs.customer\_id

)

select customer\_id, billing\_country, total\_spent, avg\_spend\_per\_transaction, last\_purchase\_date,

days\_since\_last\_purchase, total\_purchases, average\_basket\_size,

case

when total\_spent\_tile = 1 and avg\_spend\_tile = 1 then 'Low Value'

when total\_spent\_tile = 3 and avg\_spend\_tile = 3 then 'High Value'

else 'Medium Value'

end as Cust\_segment,

case

when days\_since\_last\_purchase <= 180 then 'Active'

else 'Inactive'

end as activity\_status,

case

when basket\_size\_tile = 1 then 'Small Basket Purchaser'

when basket\_size\_tile = 2 then 'Medium Basket Purchaser'

else 'Large Basket Purchaser'

end as Basket\_size\_classified,

case

when purchase\_freq\_tile = 1 then 'Less Frequent purchaser'

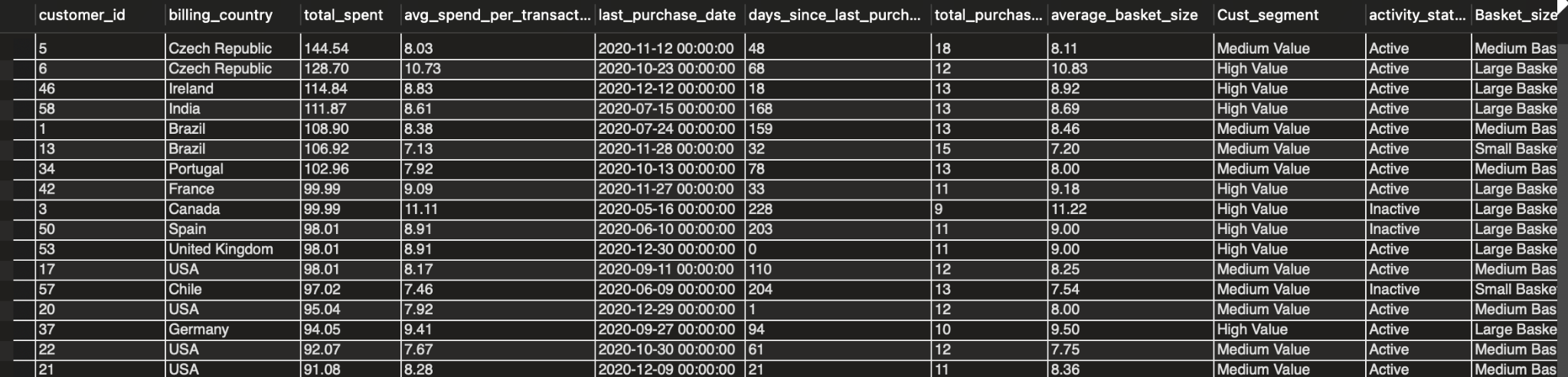
when purchase\_freq\_tile = 2 then 'Moderately Frequent Purchaser'

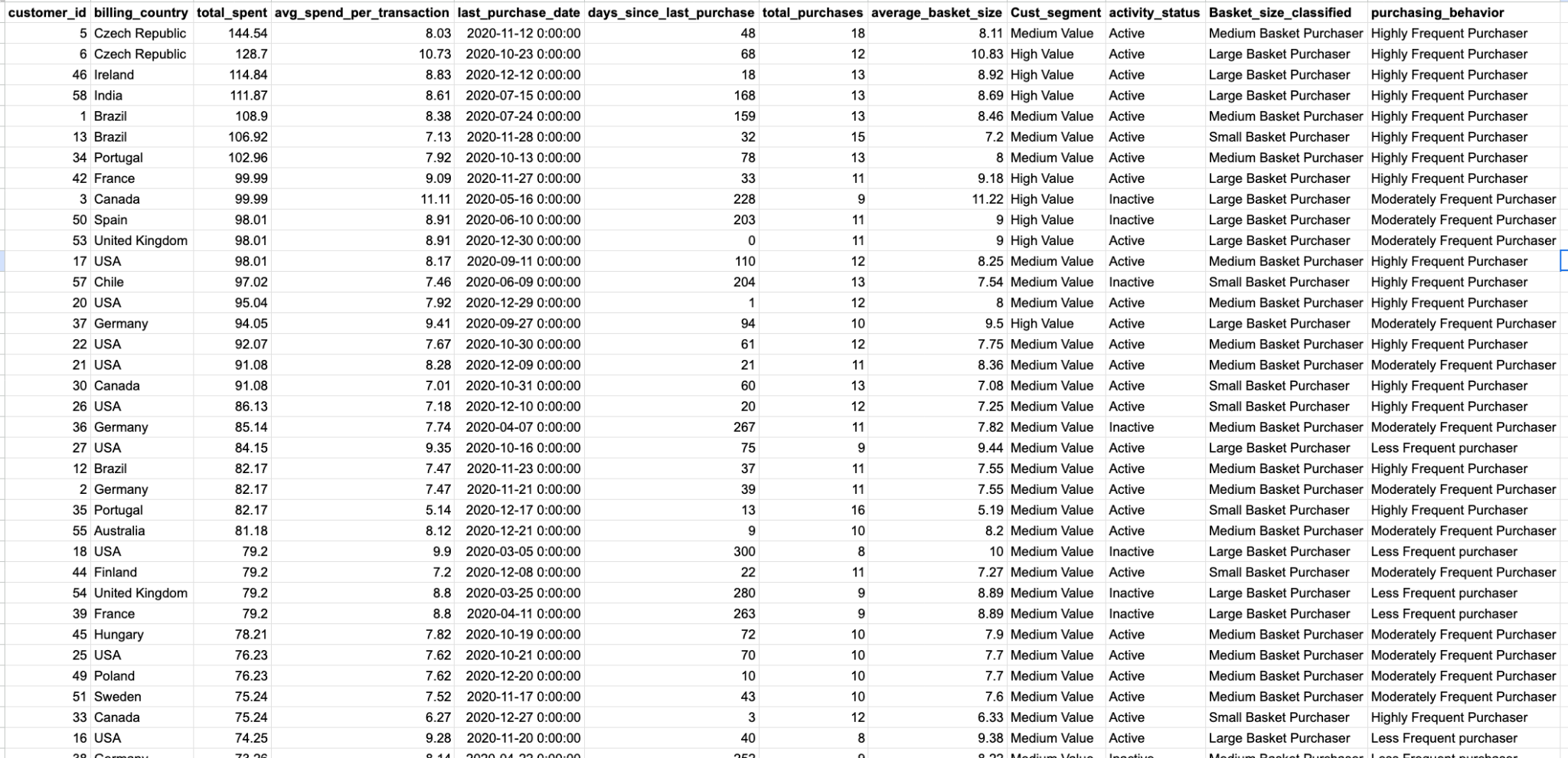
else 'Highly Frequent Purchaser'

end as purchasing\_behavior

from combined\_data

order by total\_spent desc;





## **Approach**

### **1. Deriving Key Customer Metrics**

Customer-level metrics were derived using historical transaction data:

* Monetary Value
  + total\_spent: total revenue generated by the customer
  + avg\_spend\_per\_transaction: average transaction value
* Frequency
  + total\_purchases: number of distinct invoices per customer
* Engagement / Depth
  + average\_basket\_size: average number of items purchased per transaction
* Recency (Tenure Proxy)
  + days\_since\_last\_purchase: days elapsed since the most recent purchase

These metrics collectively approximate the key drivers of CLV: recency, frequency, and monetary value, along with engagement intensity.

### **2. Value Segmentation (CLV Proxy)**

Customers were segmented into Low, Medium, and High Value groups using a percentile-based (NTILE) approach on:

* total\_spent
* avg\_spend\_per\_transaction

This ensures data-driven segmentation rather than arbitrary thresholds:

* Low Value: bottom tier on both spending dimensions
* High Value: top tier on both spending dimensions
* Medium Value: remaining customers

This segmentation serves as a practical CLV proxy suitable for SQL-based analysis.

### **3. Behavioral Tagging**

Customers were independently tagged across three behavioral dimensions:

* Activity Status (Churn Indicator)  
  + *Active*: last purchase within 180 days
  + *Inactive*: no purchase in the last 180 days
* Purchase Frequency  
  + Less Frequent
  + Moderately Frequent
  + Highly Frequent

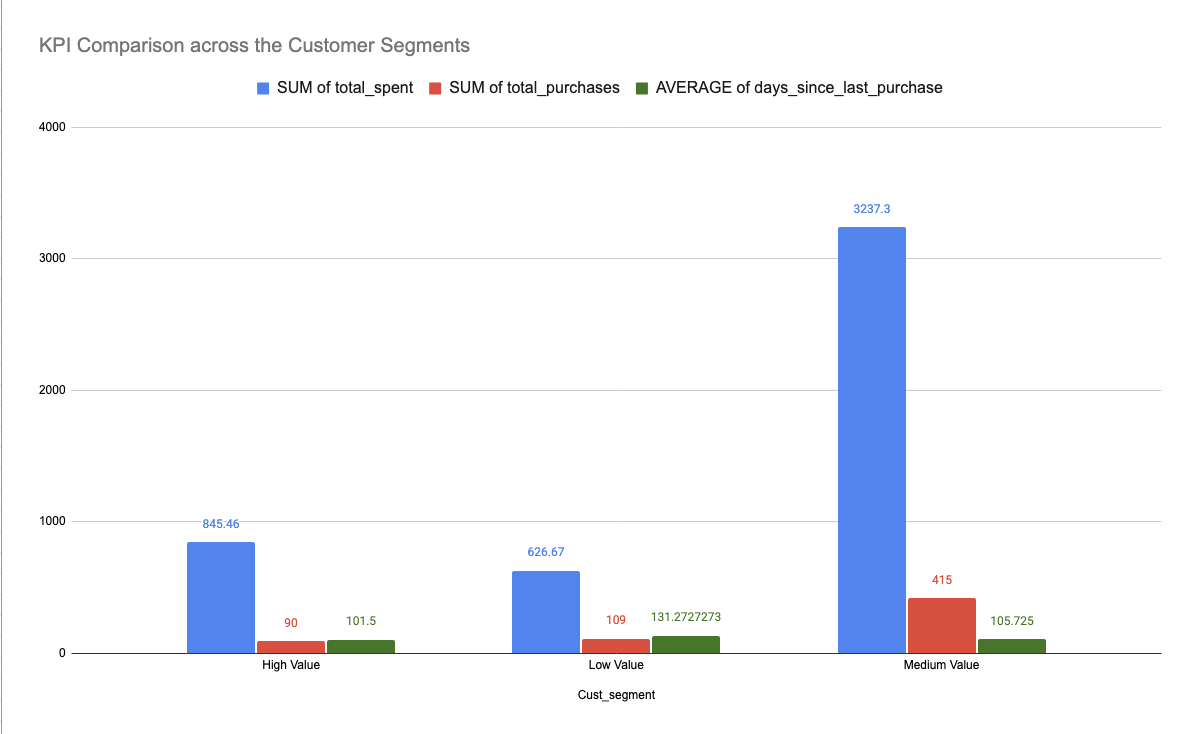
(derived using percentile ranking on total purchases)

* **Basket Size Tagging**
  + Small Basket Purchaser
  + Medium Basket Purchaser
  + Large Basket Purchaser

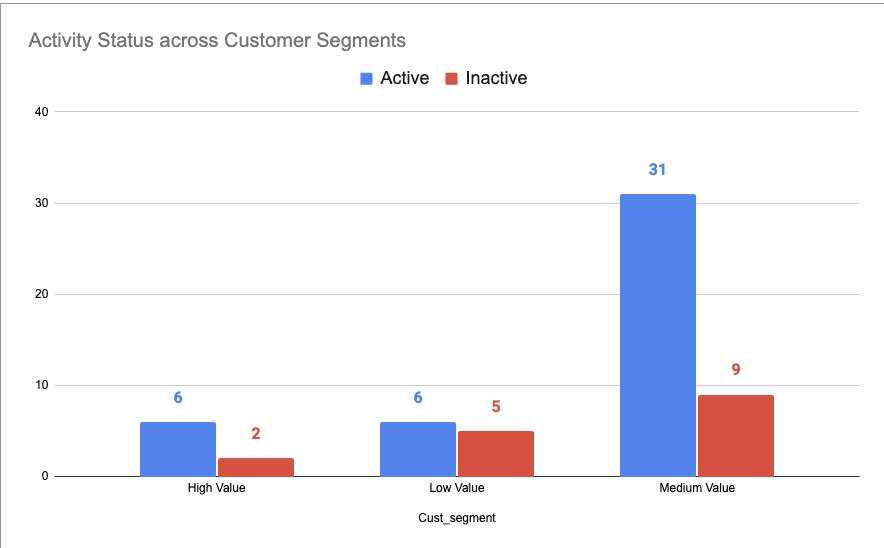
(derived using percentile ranking on average basket size)

This multi-dimensional tagging enables granular behavioral analysis beyond spend alone.

## **Key Insights**



**Figure**: Comparison of revenue contribution, purchase frequency, and recency across customer value segments



**Figure**: Activity status distribution across customer segments shows churn risk is concentrated in Medium and Low Value customers, despite their revenue contribution

### **1. Medium Value customers drive the majority of revenue—but with higher risk**

* Medium Value segment contributes the **highest total spend and purchase volume**.
* However, they also show **higher inactivity compared to High Value customers**.
* This confirms that Medium Value customers are **commercially important but behaviorally unstable**.

**Implication:** Retaining Medium Value customers has the **highest short-term revenue impact**.

### **2. High Value customers show stronger retention stability**

* High Value customers generate strong revenue with **lower inactivity rates**.
* Their recency and frequency metrics indicate **healthy engagement**.
* This validates them as the **most reliable CLV segment**.

**Implication:** Focus should be on **retention and exclusivity**, not discounts.

### **3. Low Value customers show limited upside despite activity**

* Low Value segment has:
  + Lower total spend
  + Higher average recency (longer gaps)
* Even when active, their **monetary contribution remains low**.

**Implication:** Heavy retention investment here yields **low ROI**.

### **4. Inactivity is not proportional to historical spend**

* Inactive customers exist across **all segments**, including High and Medium Value.
* This reinforces that **recency decline**, not spend, is the strongest churn signal.

**Implication:** CLV prediction must prioritize **recency and frequency over spend alone**.

**Business Recommendations**

#### **1. High Value Segment – Retention & Loyalty**

Action: Launch a premium loyalty or “VIP” program e.g. Exclusive releases/Early access

**Goal:** Preserve high-CLV customers and prevent early disengagement.

#### **2. Medium Value Segment – Stabilization & Upsell (Highest Priority)**

Action: Targeted engagement campaigns

* Personalized recommendations
* Bundled offers to increase purchase frequency

**Goal**: Reduce churn risk and convert Medium Value customers into High Value customers.

#### **3. Inactive High-Value Customers – Win-Back Strategy**

Action: Personalized re-engagement campaigns

* Time-limited discounts
* Genre- or artist-based personalized offers

**Goal**: Reactivate customers with proven CLV potential.

## **Conclusion**

By leveraging customer purchase history, engagement depth, and recency metrics, CLV can be effectively approximated using rule-based, data-driven segmentation. The analysis reveals that purchase frequency and recency are the strongest indicators of long-term customer value, while inactive customers—especially those with strong historical spend—represent a high-impact opportunity for targeted retention and reactivation strategies.

**Q8: If data on promotional campaigns (discounts, events, email marketing) is available, how could you measure their impact on customer acquisition, retention, and overall sales?**

## **Measuring the Impact of Promotional Campaigns**

If data on promotional campaigns such as discounts, events, or email marketing is available, their impact can be evaluated across **customer acquisition, retention, revenue, and behavioral change** by comparing customer activity **before, during, and after** each campaign period.

### **1. Customer Acquisition Impact**

To measure acquisition effectiveness:

* Identify **new customers whose first-ever purchase occurred during a promotion**.
* Compare the **new customer acquisition rate** during promotional periods against historical non-promotion baselines.
* Evaluate acquisition quality by tracking whether newly acquired customers make repeat purchases after the campaign ends.

**Insight:** This reveals whether promotions attract **high-quality long-term customers** or merely short-term deal seekers.

### **2. Customer Retention & Loyalty Impact**

Retention impact can be measured by analyzing existing customers’ behavior:

* Segment customers into **active, inactive, and high-value cohorts** prior to the campaign.
* Compare **repeat purchase rates, purchase frequency, and average basket size** before vs. after the promotion.
* Measure **re-activation rates** for previously inactive customers.
* Track post-campaign engagement to see if increased activity is sustained or temporary.

**Insight:** Effective promotions should improve **long-term engagement**, not just short-term spikes.

### **3. Sales & Revenue Impact**

Promotional effectiveness on revenue can be evaluated using:

* **Total revenue** during the campaign period
* **Incremental revenue**, calculated as sales above expected baseline levels
* **Average Order Value (AOV)** and **units per transaction**
* Discount and coupon **redemption rates**
* Campaign-specific conversion metrics (email open → click → purchase)

Where possible, compare:

* Promoted vs. non-promoted products
* Discounted vs. full-price transactions

**Insight:** This distinguishes revenue growth driven by **true demand uplift** versus margin erosion due to discounts.

### **4. Behavioral & Purchase Pattern Analysis**

Promotions may alter customer behavior beyond immediate sales:

* Analyze whether promotions drive **one-time purchases or repeat buying**.
* Track changes in **purchase cadence** and **category exploration**.
* Evaluate whether promotions lead to **larger baskets**, cross-selling, or product bundling.
* Observe post-campaign decay to assess habit formation vs. short-term incentive effects.

**Insight:** Promotions that change behavior are more valuable than those that simply pull demand forward.

### **5. Comparative Campaign Performance**

To identify the most effective promotional strategies:

* Compare **discounts, events, and email campaigns** on acquisition, retention, and revenue metrics.
* Normalize performance using metrics like **revenue per customer**, **repeat rate**, or **incremental lift**.
* Assess ROI by factoring in campaign cost, discount depth, and operational expenses.
* Identify which channels perform best for different customer segments (e.g., high-value vs. low-value customers).

**Insight:** This enables optimization of the promotional mix toward campaigns that balance **growth, profitability, and loyalty**.

### **Conclusion**

By combining cohort-based analysis, before-and-after comparisons, and incremental lift measurement, promotional campaigns can be evaluated not only on immediate sales impact but also on their ability to drive sustainable customer acquisition and retention. This holistic approach ensures promotions support **long-term CLV growth**, not just short-term revenue spikes.

**Q9: How would you approach this problem, if the objective and subjective questions weren't given?**

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

Begin by clarifying business goals such as revenue growth, market expansion, customer retention, or product optimization. Review the available datasets to understand customer demographics, transactions, products (tracks, albums, genres), geography, and time coverage.

**Step 2: Data Preparation & Initial Exploration**

Clean and validate the data by handling missing values, duplicates, and inconsistencies. Perform exploratory analysis to understand customer purchase behavior, sales distribution across regions, and product popularity. Use descriptive statistics to identify early trends and anomalies.

**Step 3: Define Core Metrics**

Establish key performance metrics aligned with business goals:

* Revenue metrics: total revenue, average order value, revenue by region
* Customer metrics: acquisition, retention, churn, and lifetime value
* Product metrics: top-performing tracks, albums, genres, and artists

**Step 4: Formulate Key Business Questions**

Based on exploration, develop hypotheses and guiding questions such as:

* Which regions and customer segments drive the most revenue?
* What products and genres perform best across geographies?
* How does customer spending and engagement change over time?
* Are there seasonal or lifecycle-based sales patterns?

**Step 5: Deep-Dive Analysis**

Use SQL to segment customers by geography, spending level, and preferences. Compare high- and low-performing regions, products, and customer segments. Identify behavioral patterns such as repeat purchasing, genre affinity, and shifts in engagement.

**Step 6: Generate Actionable Insights**

Translate analytical findings into business insights:

* Identify high-value and at-risk customers
* Detect opportunities for cross-selling, bundling, or targeted promotions
* Highlight growth opportunities in under-penetrated regions or genres

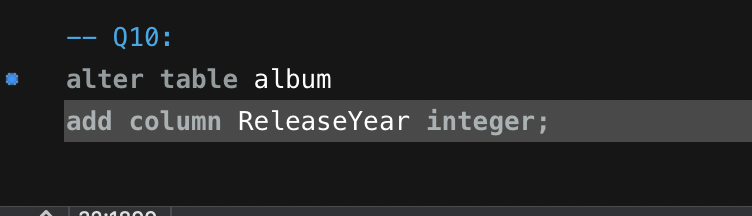
**Step 7: Retention & Growth Strategy**

Focus on retention by identifying churn-risk customers and designing re-engagement strategies. Recommend loyalty programs, personalized offers, or win-back campaigns for high-value customers.

**Step 8: Visualization & Communication**

Summarize findings using dashboards and visualizations showing revenue trends, customer retention, and product performance. Present clear, actionable recommendations aligned with business objectives.

**Q10: 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?**



**Approach:**

Use the ALTER TABLE command to modify the existing table structure by adding a new column named ReleaseYear to store the release year of each album.

**Summary:**

The ReleaseYear column enables analysis of albums by time periods, such as identifying albums released in a specific year or decade.

**Q11: 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.**

## **Objective**

To understand how customer purchasing behavior varies across countries by analyzing average customer spending, customer base size, and average number of tracks purchased per customer.

SQL Query:

with customers\_data as(

select i.customer\_id, i.billing\_country,

sum(il.unit\_price\*il.quantity) as total\_spent,

count(il.track\_id) as tracks\_purchased

from invoice i

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

group by i.customer\_id, i.billing\_country

)

select billing\_country, round(avg(total\_spent),2) as avg\_total\_spent\_per\_customer,

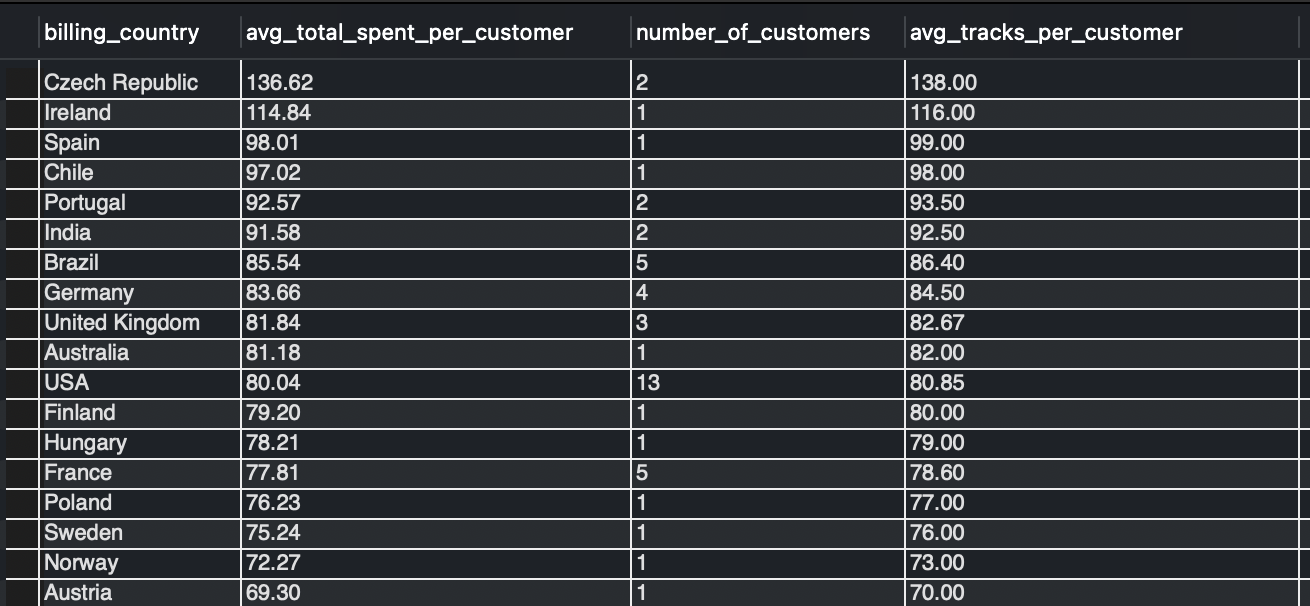
count(distinct customer\_id) as number\_of\_customers,

round(avg(tracks\_purchased),2) as avg\_tracks\_per\_customer

from customers\_data

group by billing\_country

order by avg\_total\_spent\_per\_customer desc;



## **Approach**

* Customer-level metrics were first calculated, including **total spend per customer** and **total tracks purchased per customer**, grouped by billing country.
* These customer-level values were then aggregated at the **country level** to compute:  
  + Average total amount spent per customer
  + Number of customers per country
  + Average number of tracks purchased per customer
* Results were ranked by **average total spend per customer** to highlight high-value geographies.

## **Key Findings**

### **1. High-Value but Small Customer Markets**

* Countries such as **Czech Republic, Ireland, Spain, and Chile** show the **highest average spend per customer** and **high average tracks purchased**, despite having very few customers.
* This indicates **strong per-customer monetization** but limited market penetration.

### **2. Large Customer Base with Moderate Spend**

* The **USA** has the **largest customer base**, but its average spend and average tracks purchased per customer are **moderate compared to smaller markets**.
* This suggests a **volume-driven market** rather than high per-customer spending.

### **3. Strong Balance Markets**

* Countries like **Brazil, Germany, United Kingdom, and Canada** show a healthy balance of:  
  + Reasonable customer counts
  + Consistent average spending
  + Stable track purchasing behavior
* These regions represent **reliable revenue contributors** with growth potential.

### **4. Correlation Between Spend and Tracks Purchased**

* Across nearly all countries, **average total spend closely tracks the average number of tracks purchased**.
* This indicates that **higher spending is primarily driven by purchasing more tracks**, rather than higher per-track pricing.

## **Business Implications & Recommendations**

* **Market Expansion Opportunities**:  
  High-spend, low-customer countries (e.g., Czech Republic, Ireland) are strong candidates for **targeted customer acquisition campaigns**.
* **Revenue Optimization in Large Markets**:  
  In high-volume regions like the USA and Canada, strategies should focus on **increasing average basket size or purchase frequency** rather than pure acquisition.
* **Product & Pricing Strategy**:  
  Since revenue growth is driven by track volume, **bundles, playlists, and album-based promotions** could be effective across markets.

## **Conclusion**

Customer purchasing behavior varies significantly by geography. Smaller markets tend to generate higher revenue per customer, while larger markets rely on customer volume. Understanding these differences enables more effective **country-specific marketing, pricing, and growth strategies**.