Chinook Music Store Project

Objective Questions:

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

Answer

- **Query**

```
-- NULLs in CUSTOMER
DESC customer;

SELECT 'customer.customer_id' AS column_name, COUNT(*) FROM customer WHERE customer_id IS NULL
UNION ALL
SELECT 'customer.first_name', COUNT(*) FROM customer WHERE first_name IS NULL
UNION ALL
SELECT 'customer.last_name', COUNT(*) FROM customer WHERE last_name IS NULL
UNION ALL
SELECT 'customer.company', COUNT(*) FROM customer WHERE company IS NULL
UNION ALL
SELECT 'customer.address', COUNT(*) FROM customer WHERE address IS NULL
UNION ALL
SELECT 'customer.city', COUNT(*) FROM customer WHERE city IS NULL
UNION ALL
SELECT 'customer.state', COUNT(*) FROM customer WHERE state IS NULL
UNION ALL
SELECT 'customer.country', COUNT(*) FROM customer WHERE country IS NULL
UNION ALL
SELECT 'customer.postal_code', COUNT(*) FROM customer WHERE postal_code IS NULL
UNION ALL
SELECT 'customer.phone', COUNT(*) FROM customer WHERE phone IS NULL
UNION ALL
SELECT 'customer.fax', COUNT(*) FROM customer WHERE fax IS NULL
UNION ALL
SELECT 'customer.email', COUNT(*) FROM customer WHERE email IS NULL
UNION ALL
SELECT 'customer.support_rep_id', COUNT(*) FROM customer WHERE support_rep_id IS NULL;

-- NULLs in EMPLOYEE
DESC employee;

SELECT 'employee.employee_id', COUNT(*) FROM employee WHERE employee_id IS NULL
UNION ALL
SELECT 'employee.last_name', COUNT(*) FROM employee WHERE last_name IS NULL
UNION ALL
SELECT 'employee.first_name', COUNT(*) FROM employee WHERE first_name IS NULL
UNION ALL
SELECT 'employee.title', COUNT(*) FROM employee WHERE title IS NULL
UNION ALL
SELECT 'employee.reports_to', COUNT(*) FROM employee WHERE reports_to IS NULL
UNION ALL
SELECT 'employee.birth_date', COUNT(*) FROM employee WHERE birthdate IS NULL
UNION ALL
SELECT 'employee.hire_date', COUNT(*) FROM employee WHERE hire_date IS NULL
UNION ALL
SELECT 'employee.address', COUNT(*) FROM employee WHERE address IS NULL
UNION ALL
SELECT 'employee.city', COUNT(*) FROM employee WHERE city IS NULL
UNION ALL
```

```
SELECT 'employee.state', COUNT(*) FROM employee WHERE state IS NULL
UNION ALL
SELECT 'employee.country', COUNT(*) FROM employee WHERE country IS NULL
UNION ALL
SELECT 'employee.postal_code', COUNT(*) FROM employee WHERE postal_code IS NULL
UNION ALL
SELECT 'employee.phone', COUNT(*) FROM employee WHERE phone IS NULL
UNION ALL
SELECT 'employee.fax', COUNT(*) FROM employee WHERE fax IS NULL
UNION ALL
SELECT 'employee.email', COUNT(*) FROM employee WHERE email IS NULL;

-- NULLs in ARTIST
DESC artist;

SELECT 'artist.artist_id', COUNT(*) FROM artist WHERE artist_id IS NULL
UNION ALL
SELECT 'artist.name', COUNT(*) FROM artist WHERE name IS NULL;

-- NULLs in ALBUM
DESC album;

SELECT 'album.album_id', COUNT(*) FROM album WHERE album_id IS NULL
UNION ALL
SELECT 'album.title', COUNT(*) FROM album WHERE title IS NULL
UNION ALL
SELECT 'album.artist_id', COUNT(*) FROM album WHERE artist_id IS NULL;

-- NULLs in TRACK
DESC track;

SELECT 'track.track_id', COUNT(*) FROM track WHERE track_id IS NULL
UNION ALL
SELECT 'track.name', COUNT(*) FROM track WHERE name IS NULL
UNION ALL
SELECT 'track.album_id', COUNT(*) FROM track WHERE album_id IS NULL
UNION ALL
SELECT 'track.media_type_id', COUNT(*) FROM track WHERE media_type_id IS NULL
UNION ALL
SELECT 'track.genre_id', COUNT(*) FROM track WHERE genre_id IS NULL
UNION ALL
SELECT 'track.composer', COUNT(*) FROM track WHERE composer IS NULL
UNION ALL
SELECT 'track.milliseconds', COUNT(*) FROM track WHERE milliseconds IS NULL
UNION ALL
SELECT 'track.bytes', COUNT(*) FROM track WHERE bytes IS NULL
UNION ALL
SELECT 'track.unit_price', COUNT(*) FROM track WHERE unit_price IS NULL;

-- NULLs in INVOICE
DESC invoice;

SELECT 'invoice.invoice_id', COUNT(*) FROM invoice WHERE invoice_id IS NULL
UNION ALL
SELECT 'invoice.customer_id', COUNT(*) FROM invoice WHERE customer_id IS NULL
UNION ALL
SELECT 'invoice.invoice_date', COUNT(*) FROM invoice WHERE invoice_date IS NULL
UNION ALL
SELECT 'invoice.billing_address', COUNT(*) FROM invoice WHERE billing_address IS NULL
UNION ALL
SELECT 'invoice.billing_city', COUNT(*) FROM invoice WHERE billing_city IS NULL
UNION ALL
SELECT 'invoice.billing_state', COUNT(*) FROM invoice WHERE billing_state IS NULL
UNION ALL
SELECT 'invoice.billing_country', COUNT(*) FROM invoice WHERE billing_country IS NULL
UNION ALL
SELECT 'invoice.billing_postal_code', COUNT(*) FROM invoice WHERE billing_postal_code IS NULL
```

```
UNION ALL
SELECT 'invoice.total', COUNT(*) FROM invoice WHERE total IS NULL;

-- NULLs in INVOICE_LINE
DESC invoice_line;

SELECT 'invoice_line.invoice_line_id', COUNT(*) FROM invoice_line WHERE invoice_line_id IS NULL
UNION ALL
SELECT 'invoice_line.invoice_id', COUNT(*) FROM invoice_line WHERE invoice_id IS NULL
UNION ALL
SELECT 'invoice_line.track_id', COUNT(*) FROM invoice_line WHERE track_id IS NULL
UNION ALL
SELECT 'invoice_line.unit_price', COUNT(*) FROM invoice_line WHERE unit_price IS NULL
UNION ALL
SELECT 'invoice_line.quantity', COUNT(*) FROM invoice_line WHERE quantity IS NULL;

-- Duplicates in CUSTOMER (excluding customer_id)
SELECT first_name, last_name, company, address, city, state, country, postal_code, phone, fax, email, support_rep_id, COUNT(*) AS
occurrences
FROM customer
GROUP BY first_name, last_name, company, address, city, state, country, postal_code, phone, fax, email, support_rep_id
HAVING occurrences > 1;

-- Duplicates in EMPLOYEE (excluding employee_id)
SELECT last_name, first_name, title, reports_to, birthdate, hire_date, address, city, state, country, postal_code, phone, fax, email,
COUNT(*) AS occurrences
FROM employee
GROUP BY last_name, first_name, title, reports_to, birthdate, hire_date, address, city, state, country, postal_code, phone, fax, email
HAVING occurrences > 1;

-- Duplicates in ARTIST (excluding artist_id)
SELECT name, COUNT(*) AS occurrences
FROM artist
GROUP BY name
HAVING occurrences > 1;

-- Duplicates in ALBUM (excluding album_id)
SELECT title, artist_id, COUNT(*) AS occurrences
FROM album
GROUP BY title, artist_id
HAVING occurrences > 1;

-- Duplicates in TRACK (excluding track_id)
SELECT name, album_id, media_type_id, genre_id, composer, milliseconds, bytes, unit_price, COUNT(*) AS occurrences
FROM track
GROUP BY name, album_id, media_type_id, genre_id, composer, milliseconds, bytes, unit_price
HAVING occurrences > 1;

-- Duplicates in INVOICE (excluding invoice_id)
SELECT customer_id, invoice_date, billing_address, billing_city, billing_state, billing_country, billing_postal_code, total, COUNT(*)
AS occurrences
FROM invoice
GROUP BY customer_id, invoice_date, billing_address, billing_city, billing_state, billing_country, billing_postal_code, total
HAVING occurrences > 1;

-- Duplicates in INVOICE_LINE (excluding invoice_line_id)
SELECT invoice_id, track_id, unit_price, quantity, COUNT(*) AS occurrences
FROM invoice_line
GROUP BY invoice_id, track_id, unit_price, quantity
HAVING occurrences > 1;

-- Duplicates in PLAYLIST (excluding playlist_id)
SELECT name, COUNT(*) AS occurrences
FROM playlist
GROUP BY name
HAVING occurrences > 1;
```

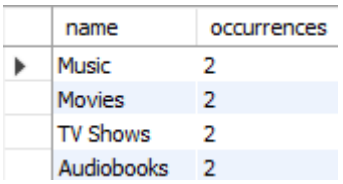
```
-- Duplicates in MEDIA_TYPE (excluding media_type_id)
SELECT name, COUNT(*) AS occurrences
FROM media_type
GROUP BY name
HAVING occurrences > 1;

-- Duplicates in GENRE (excluding genre_id)
SELECT name, COUNT(*) AS occurrences
FROM genre
GROUP BY name
HAVING occurrences > 1;
```

- There are 11 tables in the Chinook dataset provided.
    - **Missing Values / NULLs:**
        - Most tables in Chinook are designed with optional columns, and hence there are some NULLs, but they are expected and not an error. These occur in fields like company, fax, state etc. which are not always applicable to every record.
        - These are fine unless the business rule explicitly requires them to be filled.
        - Currently, I am leaving them as is and will revisit them if a specific use case or reporting requirement necessitates filling or replacing these NULLs.
    - **Duplicates:**
        - The primary keys are not considered in the duplicate check, as it is already enforced by the database and guarantees uniqueness.
        - In the invoice_line table, there are several rows where all columns except the primary key appear identical. These duplicates are likely not data entry errors but rather the result of how purchases are logged, the same track might be added multiple times to an invoice without aggregation.

invoice_id	track_id	unit_price	quantity	occurrences
20	3297	0.99	1	2
49	146	0.99	1	2
224	2600	0.99	1	2
261	1122	0.99	1	2
261	2402	0.99	1	2
353	1470	0.99	1	2
401	3282	0.99	1	2
421	60	0.99	1	2
426	2558	0.99	1	2
456	3495	0.99	1	2
517	2573	0.99	1	2
593	1042	0.99	1	2

    - The analysis of the playlist table revealed that several playlist names are duplicated, even though their id remains unique.

Specifically, the names **Music**, **Movies**, **TV Shows**, and **Audiobooks** each appear **twice**, indicating that the same name has been assigned to different playlists. While this does not violate database integrity (since the primary key remains unique), it may lead to confusion in reporting or user-facing applications. These duplicates should be reviewed to determine if they are intentional (e.g., representing different versions or contexts of the same playlist) or if they should be consolidated for consistency.



name	occurrences
Music	2
Movies	2
TV Shows	2
Audiobooks	2

- No other table in the database shows duplicates on business-relevant columns, so this seems acceptable given the sample nature of the data.

2. Find the top-selling tracks and top artists in the USA and identify their most famous genres.

Answer

- **Query**

```
-- Creating a View for sales corresponding to USA
CREATE VIEW usa_sales AS
SELECT il.invoice_line_id,
                i.invoice_id,
                i.billing_country,
                il.track_id,
                il.unit_price,
                il.quantity,
                t.name AS track_name,
                t.genre_id,
                al.album_id,
                al.artist_id,
                ar.name AS artist_name,
                g.name AS genre_name
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 ar ON al.artist_id = ar.artist_id
JOIN genre g ON t.genre_id = g.genre_id
WHERE i.billing_country = 'USA'

-- Top selling tracks in USA
WITH
top_tracks AS (
SELECT track_id, track_name,
                SUM(quantity) AS total_quantity,
                SUM(unit_price * quantity) AS total_sales,
    DENSE_RANK() OVER(ORDER BY SUM(unit_price * quantity) DESC) AS track_rank
FROM usa_sales
```

```
GROUP BY track_id, track_name
)
SELECT * FROM top_tracks
WHERE track_rank <= 3

-- Top artists in USA
WITH
top_artists AS (
        SELECT artist_id, artist_name,
                SUM(quantity) AS total_quantity,
                SUM(unit_price * quantity) AS total_sales,
    DENSE_RANK() OVER(ORDER BY SUM(unit_price * quantity) DESC) AS artist_rank
        FROM usa_sales
        GROUP BY artist_id, artist_name
)
SELECT * FROM top_artists
WHERE artist_rank <=5

-- Most famous genres of top artist
WITH
top_artists AS (
        SELECT artist_id,
    DENSE_RANK() OVER(ORDER BY SUM(unit_price * quantity) DESC) AS artist_rank
        FROM usa_sales
        GROUP BY artist_id
),
top_artist_selection AS (
        SELECT artist_id FROM top_artists
        WHERE artist_rank <=5
),
top_artist_genre AS (
        SELECT artist_id, artist_name, genre_name,
                SUM(quantity) AS total_quantity,
                SUM(unit_price * quantity) AS total_sales,
                RANK() OVER(PARTITION BY artist_id ORDER BY SUM(unit_price * quantity) DESC) AS genre_rank
        FROM usa_sales
        WHERE artist_id IN (SELECT * FROM top_artist_selection)
        GROUP BY artist_id, artist_name, genre_name
)
SELECT artist_id, artist_name, genre_name, total_quantity, total_sales
FROM top_artist_genre
WHERE genre_rank = 1
ORDER BY total_sales DES;
```

- **Output**

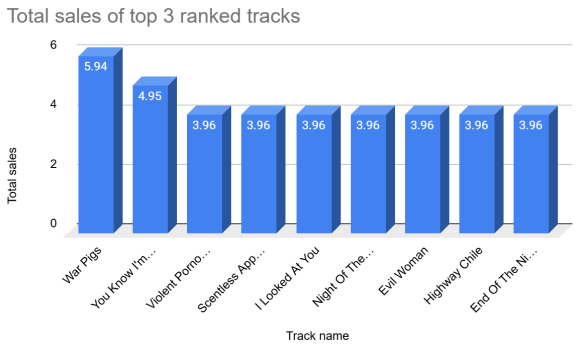
track_id	track_name	total_quantity	total_sales	track_rank
3336	War Pigs	6	5.94	1
3465	You Know I'm No Good (feat. Ghostface Killah)	5	4.95	2
2560	Violent Pornography	4	3.96	3
1995	Scentless Apprentice	4	3.96	3
2646	I Looked At You	4	3.96	3
13	Night Of The Long Knives	4	3.96	3
153	Evil Woman	4	3.96	3
1495	Highway Chile	4	3.96	3
2647	End Of The Night	4	3.96	3

artist_id	artist_name	total_quantity	total_sales	artist_rank
152	Van Halen	43	42.57	1
124	R.E.M.	38	37.62	2
142	The Rolling Stones	37	36.63	3
110	Nirvana	35	34.65	4
84	Foo Fighters	34	33.66	5
81	Eric Clapton	34	33.66	5

artist_id	artist_name	genre_name	total_quantity	total_sales
152	Van Halen	Rock	43	42.57
142	The Rolling Stones	Rock	37	36.63
110	Nirvana	Rock	35	34.65
124	R.E.M.	Alternative & Punk	32	31.68
81	Eric Clapton	Blues	31	30.69
84	Foo Fighters	Rock	21	20.79

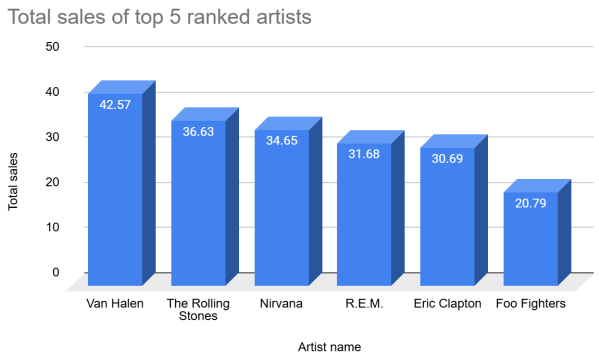
- **Top 3 Ranked Tracks in USA:**

  The highest-selling track in the USA is *War Pigs* (track_id: 3336), which sold 6 units and generated $5.94 in sales. It is followed by *You Know I'm No Good (feat. Ghostface Killah)* (track_id: 3465) with 5 units and $4.95 in sales. Several other tracks, including *Violent Pornography*, *Scentless Apprentice*, *I Looked At You*, *Night Of The Long Knives*, *Evil Woman*, *Highway Chile*, and *End Of The Night*, are tied at rank 3, each selling 4 units and earning $3.96.



- **Top 5 Ranked Artists in USA:**

  The most successful artist in the USA is *Van Halen* (artist_id: 152), with 43 units sold and $42.57 in revenue. *R.E.M.* (artist_id: 124) ranks second with 38 units and $37.62, followed by *The Rolling Stones* (artist_id: 142) with 37 units and $36.63. *Nirvana* (artist_id: 110) comes in fourth, selling 35 units for $34.65, while both *Foo Fighters* (artist_id: 84) and *Eric Clapton* (artist_id: 81) are tied at fifth place, each achieving 34 units and $33.66 in sales.



- **Most Famous Genres of Top Artists in USA:**

  For the top-ranked artists, the most prominent genre is overwhelmingly *Rock*, represented by *Van Halen*, *The Rolling Stones*, *Nirvana*, and *Foo Fighters*, who collectively dominate the sales in this category. *R.E.M.* stands out with *Alternative & Punk* as their most popular genre, contributing significantly to their sales. Meanwhile, *Eric Clapton* shines in the *Blues* genre, which accounts for the majority of his success in the USA market.

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

   Answer

   - **Query**

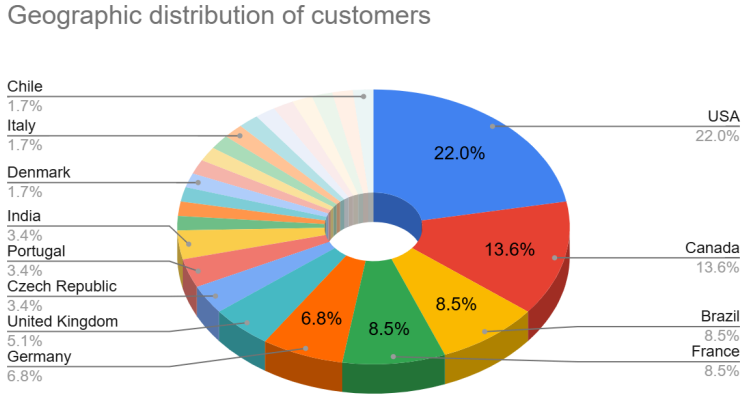
```
-- Total customers
SELECT COUNT(customer_id) AS customer_count
FROM customer

-- Customers by country
SELECT country, COUNT(*) AS customers_count
FROM customer
GROUP BY country
ORDER BY customers_count DESC

-- Customers by country & state
SELECT country, state, COUNT(*) AS customers_count
FROM customer
GROUP BY country, state
ORDER BY customers_count DESC

-- Customers by country, state, and city
SELECT country, state, city, COUNT(*) AS customers_count
FROM customer
GROUP BY country, state, city
ORDER BY customers_count DESC
```

   - **Output**

country	customers_count
USA	13
Canada	8
Brazil	5
France	5
Germany	4
United Kingdom	3
Czech Republic	2
Portugal	2
India	2
Norway	1
Austria	1
Belgium	1

country	state	city	customers_count
Czech Republic	NULL	Prague	2
Brazil	SP	São Paulo	2
USA	CA	Mountain View	2
Germany	NULL	Berlin	2
France	NULL	Paris	2
United Kingdom	NULL	London	2
Canada	QC	Montréal	1
Norway	NULL	Oslo	1
Austria	NULL	Vienne	1
Belgium	NULL	Brussels	1
Denmark	NULL	Copenhagen	1
Brazil	RJ	Rio de Janeiro	1

- The customer demographic breakdown of Chinook's customer base is primarily defined by their geographical location, as the database does not include age or gender information.
- Out of the 59 customers, the majority are concentrated in the **USA**, followed by significant numbers in **Canada**, **Brazil and France**, with smaller clusters in several other countries.
- Within these countries, customers are distributed across various states and cities, reflecting a diverse and international customer base.
- This location-based segmentation provides an opportunity to tailor marketing strategies and promotions according to regional preferences and concentrations.



4. Calculate the total revenue and number of invoices for each country, state and

city.

Answer

- **Query**

```
– By Country
SELECT billing_country AS country,
    COUNT(*) AS invoice_count,
    SUM(total) AS total_revenue
FROM invoice
GROUP BY billing_country
ORDER BY total_revenue DESC;

-- By Country and state
SELECT billing_country AS country,
    billing_state AS state,
    COUNT(*) AS invoice_count,
    SUM(total) AS total_revenue
FROM invoice
GROUP BY billing_country, billing_state
ORDER BY total_revenue DESC;

-- By Country, state and city
SELECT billing_country AS country,
    billing_state AS state,
    billing_city AS city,
    COUNT(*) AS invoice_count,
    SUM(total) AS total_revenue
FROM invoice
GROUP BY billing_country, billing_state, billing_city
ORDER BY total_revenue DESC;
```

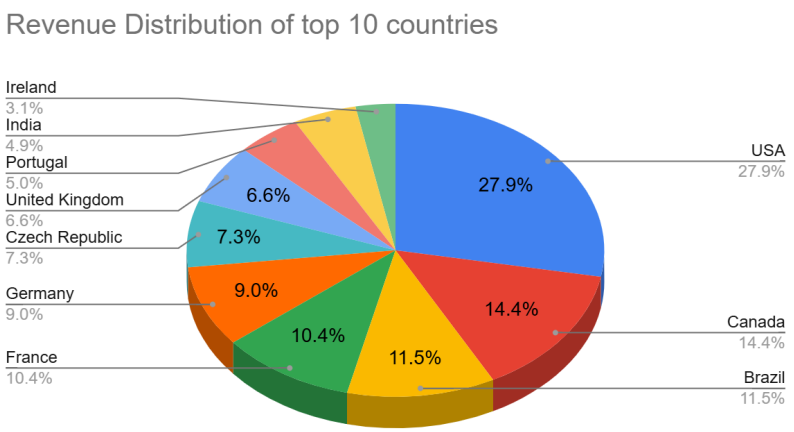
- **Output**

country	invoice_count	total_revenue
USA	131	1040.49
Canada	76	535.59
Brazil	61	427.68
France	50	389.07
Germany	41	334.62
Czech Republic	30	273.24
United Kingdom	28	245.52
Portugal	29	185.13
India	21	183.15
Ireland	13	114.84

country	state	city	invoice_count	total_revenue
Czech Republic	None	Prague	30	273.24
USA	CA	Mountain View	20	169.29
United Kingdom	None	London	19	166.32
Germany	None	Berlin	20	158.40
France	None	Paris	18	151.47
Brazil	SP	São Paulo	22	129.69
Ireland	Dublin	Dublin	13	114.84
India	None	Delhi	13	111.87
Brazil	SP	São José dos Campos	13	108.90
Brazil	DF	Brasília	15	106.92
Portugal	None	Lisbon	13	102.96

- The USA leads Chinook's customer base with the highest revenue of $1,040.49 from 131 invoices, followed by Canada and Brazil with notable contributions of $535.59 and $427.68, respectively.

- At the country & state level, the USA's California and Washington stand out among specific states, while São Paulo in Brazil and Ontario in Canada also show strong sales figures.
- When further drilled down to the city level, Prague (Czech Republic) emerges as the single most lucrative city with $273.24 from 30 invoices, followed by key cities like Mountain View (California, USA), London (UK), Berlin (Germany), Paris (France), and São Paulo (Brazil).
- These insights highlight that while the USA dominates overall, certain cities and regions outside the USA also contribute significantly, indicating diverse customer engagement across global locations.



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

Answer

- **Query**

```
WITH customer_revenue AS (
  SELECT c.customer_id,
      CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
      c.country,
      SUM(i.total) AS total_revenue,
      DENSE_RANK() OVER (PARTITION BY c.country ORDER BY SUM(i.total) DESC) AS revenue_rank
  FROM customer c
  JOIN invoice i ON c.customer_id = i.customer_id
  GROUP BY c.customer_id, customer_name, c.country
)
SELECT country, customer_id, customer_name, total_revenue, revenue_rank
FROM customer_revenue
WHERE revenue_rank <= 5
ORDER BY country, revenue_rank;
```

- **Output**

country	customer_id	customer_name	total_revenue	revenue_rank
Argentina	56	Diego Gutiérrez	39.60	1
Australia	55	Mark Taylor	81.18	1
Austria	7	Astrid Gruber	69.30	1
Belgium	8	Daan Peeters	60.39	1
Brazil	1	Luís Gonçalves	108.90	1
Brazil	13	Fernanda Ramos	106.92	2
Brazil	12	Roberto Almeida	82.17	3
Brazil	11	Alexandre Rocha	69.30	4
Brazil	10	Eduardo Martins	60.39	5
Canada	3	François Tremblay	99.99	1
Canada	30	Edward Francis	91.08	2
Canada	33	Ellie Sullivan	75.24	3
Canada	32	Aaron Mitchell	70.29	4
Canada	15	Jennifer Peterson	66.33	5

- The analysis of the top 5 customers by total revenue in each country provides valuable insights into Chinook's most important clientele across different regions.
- In each country, a small group of customers accounts for a significant share of sales, underscoring the importance of identifying and retaining these high-value individuals.
- The rankings show which customers consistently generate the highest revenue within their country, enabling Chinook to tailor loyalty programs, exclusive offers, and personalized communication to these segments.
- Moreover, understanding the distribution of top customers geographically allows the business to identify patterns, such as countries where revenue is more concentrated among a few customers versus countries with a more distributed customer base.
- These insights can guide strategic decisions on where to focus customer relationship management efforts, develop premium services, and allocate marketing resources to maximize returns from these priority customer segments.

6. Identify the top-selling track for each customer.

Answer

- **Query**

```
– Top selling track
WITH customer_tracks AS (
  SELECT c.customer_id,
      CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
      t.track_id,
      t.name AS track_name,
      SUM(il.quantity) AS total_quantity,
      SUM(il.unit_price * il.quantity) AS total_revenue,
      RANK() OVER (PARTITION BY c.customer_id ORDER BY SUM(il.unit_price * il.quantity) DESC) AS track_rank
  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, customer_name, t.track_id, t.name
```

```
)
SELECT customer_id, customer_name, track_id, track_name, total_quantity, total_revenue
FROM customer_tracks
WHERE track_rank = 1
ORDER BY customer_id;

-- Top selling genre
WITH customer_genres AS (
   SELECT c.customer_id,
        CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
        g.genre_id,
        g.name AS genre_name,
        SUM(il.quantity) AS total_quantity,
        SUM(il.unit_price * il.quantity) AS total_revenue,
        RANK() OVER (PARTITION BY c.customer_id ORDER BY SUM(il.quantity) DESC) AS genre_rank
   FROM customer c
   JOIN invoice i ON c.customer_id = i.customer_id
   JOIN invoice_line il ON i.invoice_id = il.invoice_id
   JOIN track t ON il.track_id = t.track_id
   JOIN genre g ON t.genre_id = g.genre_id
   GROUP BY c.customer_id, customer_name, g.genre_id, g.name
)
SELECT customer_id, customer_name, genre_name, total_quantity, total_revenue
FROM customer_genres
WHERE genre_rank = 1
ORDER BY customer_id;

-- Top selling artist
WITH customer_artists AS (
   SELECT c.customer_id,
        CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
        a.artist_id,
        a.name AS artist_name,
        SUM(il.quantity) AS total_quantity,
        SUM(il.unit_price * il.quantity) AS total_revenue,
        RANK() OVER (PARTITION BY c.customer_id ORDER BY SUM(il.quantity) DESC) AS artist_rank
   FROM customer c
   JOIN invoice i ON c.customer_id = i.customer_id
   JOIN invoice_line il ON i.invoice_id = il.invoice_id
   JOIN track t ON il.track_id = t.track_id
   JOIN album al ON t.album_id = al.album_id
   JOIN artist a ON al.artist_id = a.artist_id
   GROUP BY c.customer_id, customer_name, a.artist_id, a.name
)
SELECT customer_id, customer_name, artist_name, total_quantity, total_revenue
FROM customer_artists
WHERE artist_rank = 1
ORDER BY customer_id;
```

- # Output

customer_id	customer_name	track_id	track_name	total_quantity	total_revenue
1	Luís Gonçalves	572	Put Your Lights On	1	0.99
1	Luís Gonçalves	2229	Speak To Me/Breathe	1	0.99
1	Luís Gonçalves	2365	The Righteous & The Wicked	1	0.99
1	Luís Gonçalves	2113	Shining In The Light	1	0.99
1	Luís Gonçalves	3003	All I Want Is You	1	0.99
1	Luís Gonçalves	3087	Atomic Punk	1	0.99
1	Luís Gonçalves	2558	Radio/Video	1	0.99
1	Luís Gonçalves	38	All I Really Want	1	0.99
1	Luís Gonçalves	2733	My Generation	1	0.99
1	Luís Gonçalves	2600	Train In Vain	1	0.99
1	Luís Gonçalves	1757	Chuva No Brejo	1	0.99
1	Luís Gonçalves	419	A Kind Of Magic	1	0.99
1	Luís Gonçalves	2307	So Fast, So Numb	1	0.99

customer_id	customer_name	genre_name	total_quantity	total_revenue
1	Luís Gonçalves	Rock	72	71.28
2	Leonie Köhler	Rock	45	44.55
3	François Tremblay	Rock	75	74.25
4	Bjørn Hansen	Rock	40	39.60
5	František Wichterlová	Rock	67	66.33
6	Helena Holý	Rock	76	75.24
7	Astrid Gruber	Rock	40	39.60
8	Daan Peeters	Rock	26	25.74
9	Kara Nielsen	Rock	24	23.76
10	Eduardo Martins	Metal	25	24.75
11	Alexandre Rocha	Alternative ...	30	29.70
12	Roberto Almeida	Rock	41	40.59

customer_id	customer_name	artist_name	total_quantity	total_revenue
1	Luís Gonçalves	The Cult	18	17.82
2	Leonie Köhler	Audioslave	14	13.86
3	François Tremblay	The Who	20	19.80
4	Bjørn Hansen	Guns N' Roses	12	11.88
5	František Wichterlová	Kiss	20	19.80
6	Helena Holý	Red Hot Chili Peppers	20	19.80
7	Astrid Gruber	Miles Davis	15	14.85
8	Daan Peeters	Godsmack	12	11.88
9	Kara Nielsen	Jamiroquai	4	3.96
10	Eduardo Martins	Metallica	15	14.85

- In the **Chinook dataset**, each customer seems to have purchased each track only once (i.e., no customer has bought the same track multiple times).
- This means that when you group by customer_id and track_id, the SUM(quantity) is always 1, and the total_revenue simply equals the track price.
- Since no customer shows a preference for a track by buying it repeatedly, identifying the “top” track per customer is effectively just picking one of the tracks they bought, which doesn’t indicate any real preference or behavior.
- In reality, such analysis is useful if customers have bought multiple tracks and some tracks multiple times, signaling popularity, loyalty, or preference.
- Instead of identifying a customer’s top-selling track (since all are unique), it would be more insightful to explore their preferred **genres** and **artists**.
- Each customer tends to favor a specific genre, with ***Rock*** emerging as the most common favorite across the customer base.
- Similarly, each of the customers show loyalty to certain artists.
- These insights into customer-level preferences for genres and artists provide an opportunity to personalize marketing, curate targeted playlists, and design promotions that resonate with each customer’s unique tastes, thereby enhancing engagement and retention.

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

Answer

- **Query**

```
-- Total purchases, AOV, Total spent
SELECT c.customer_id,
    CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
    COUNT(i.invoice_id) AS total_purchases,
    AVG(i.total) AS avg_order_value,
    SUM(i.total) AS total_spent
FROM customer c
JOIN invoice i ON c.customer_id = i.customer_id
GROUP BY c.customer_id, customer_name
ORDER BY total_purchases DESC;

-- Purchase Frequency: Avg gap, Min gap, Max gap
WITH customer_invoices AS (
  SELECT
    customer_id,
    invoice_date,
    LAG(invoice_date) OVER (PARTITION BY customer_id ORDER BY invoice_date) AS prev_invoice_date
  FROM invoice
),
gaps AS (
  SELECT
    customer_id,
    DATEDIFF(invoice_date, prev_invoice_date) AS gap_days
  FROM customer_invoices
  WHERE prev_invoice_date IS NOT NULL
)
SELECT
  c.customer_id,
  CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
  AVG(g.gap_days) AS avg_gap_days,
  MIN(g.gap_days) AS min_gap_days,
  MAX(g.gap_days) AS max_gap_days
FROM gaps g
JOIN customer c ON g.customer_id = c.customer_id
GROUP BY c.customer_id, customer_name
ORDER BY c.customer_id;
```

- **Output**

customer_id	customer_name	total_purchases	avg_order_value	total_spent
5	František Wichterlová	18	8.030000	144.54
35	Madalena Sampaio	16	5.135625	82.17
13	Fernanda Ramos	15	7.128000	106.92
30	Edward Francis	13	7.006154	91.08
46	Hugh O'Reilly	13	8.833846	114.84
1	Luís Gonçalves	13	8.376923	108.90
58	Manoj Pareek	13	8.605385	111.87
57	Luis Rojas	13	7.463077	97.02
34	João Fernandes	13	7.920000	102.96
26	Richard Cunningham	12	7.177500	86.13

customer_id	customer_name	avg_gap_days	min_gap_days	max_gap_days
1	Luís Gonçalves	106.2500	2	330
2	Leonie Köhler	129.2000	2	373
3	François Tremblay	147.5000	2	287
4	Bjørn Hansen	136.1250	2	440
5	František Wichterlová	74.2941	8	307
6	Helena Holý	104.4545	8	194
7	Astrid Gruber	133.5000	33	257
8	Daan Peeters	131.8333	11	307
9	Kara Nielsen	120.6667	2	201
10	Eduardo Martins	112.2727	35	279

- Total Purchases, AOV and Total Spent
    - The purchasing patterns of Chinook's customers show a strong concentration of moderate-frequency buyers with steady spending habits.
    - The majority of customers make between 8–13 purchases, with average order values (AOV) consistently between $7–$9 per invoice, reflecting uniform pricing and buying behavior.
    - Notably, some customers, such as *Robert Brown*, stand out for having high AOVs despite fewer purchases, indicating premium spending tendencies.
    - Overall, these insights suggest a stable and predictable customer base, with opportunities to further engage high-AOV customers and encourage more frequent purchases from lower-frequency segments.
- Purchase Frequency
    - The analysis of purchase frequency, measured as the average, minimum, and maximum gap (in days) between consecutive purchases, reveals considerable variation across customers.
    - Most customers exhibit an average gap of around 100–150 days, indicating fairly infrequent, periodic buying behavior.
    - The minimum gap for many customers is as low as 1–8 days, showing occasional bursts of activity, whereas maximum gaps can stretch to over 900 days in extreme cases, highlighting periods of inactivity.
    - These insights point to a predominantly irregular purchasing pattern, with opportunities to engage customers more consistently and reduce the long dormant periods between purchases.

8. What is the customer churn rate?

Answer

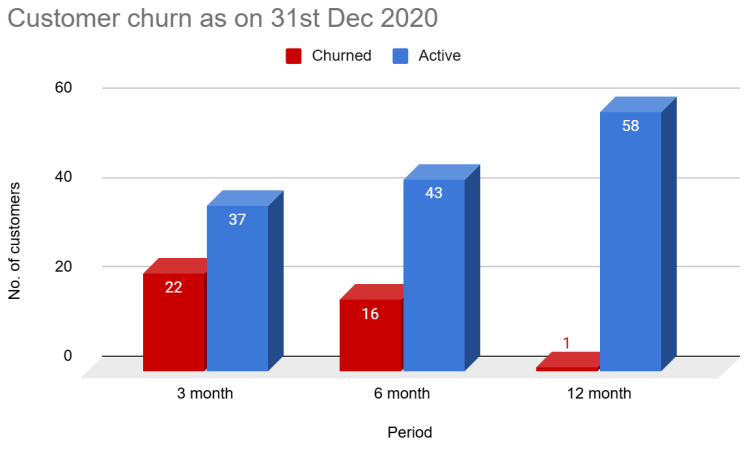
- **Query**

```
WITH last_purchase AS (
    SELECT customer_id, MAX(invoice_date) AS last_invoice_date
    FROM invoice
    GROUP BY customer_id
)

SELECT
    COUNT(*) AS total_customers,

    SUM(CASE WHEN last_invoice_date < '2020-10-01' THEN 1 ELSE 0 END) AS churned_3m,
    CONCAT(ROUND(SUM(CASE WHEN last_invoice_date < '2020-10-01' THEN 1 ELSE 0 END) / COUNT(*) * 100, 2), '%') AS
churn_rate_3m,

    SUM(CASE WHEN last_invoice_date < '2020-07-01' THEN 1 ELSE 0 END) AS churned_6m,
    CONCAT(ROUND(SUM(CASE WHEN last_invoice_date < '2020-07-01' THEN 1 ELSE 0 END) / COUNT(*) * 100, 2), '%') AS
churn_rate_6m,

    SUM(CASE WHEN last_invoice_date < '2020-01-01' THEN 1 ELSE 0 END) AS churned_12m,
    CONCAT(ROUND(SUM(CASE WHEN last_invoice_date < '2020-01-01' THEN 1 ELSE 0 END) / COUNT(*) * 100, 2), '%') AS
churn_rate_12m

FROM last_purchase;
```

- **Output**

	total_customers	churned_3m	churn_rate_3m	churned_6m	churn_rate_6m	churned_12m	churn_rate_12m
▶	59	22	37.29%	16	27.12%	1	1.69%

- Since most of the last invoices for customers were dated during the last quarter of 2020, we assume that this analysis was conducted as of **31-Dec-2020**, making it a reasonable cutoff point for evaluating customer activity and churn.
- Over the preceding 3 months, **37.29% (22 out of 59)** of customers had not made a purchase, indicating a significant short-term drop-off.
- At the 6-month mark, churn improved slightly, with **27.12% (16 customers)** inactive, and over the 12-month period, churn was very low at just **1.69% (1 customer)**.
- This suggests that while some customers lapse temporarily, many remain engaged over the long term, and targeted strategies to re-engage short-term inactive customers could help further improve retention.



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

Answer

- **Query**

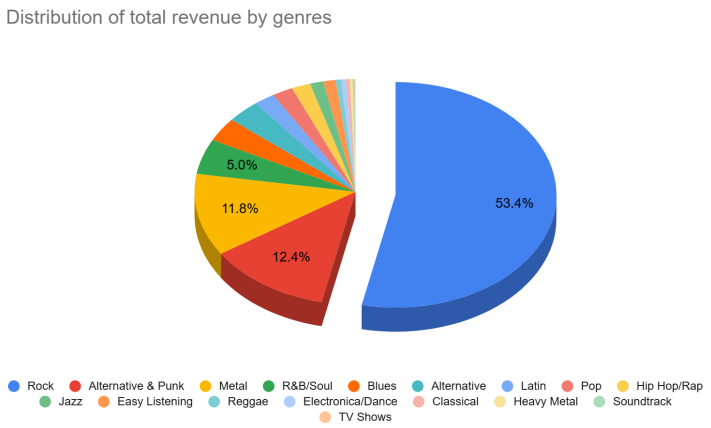
```
-- Genre level sales in USA
WITH usa_sales AS (
    SELECT il.invoice_line_id, il.unit_price, il.quantity,
        i.billing_country, t.genre_id, g.name AS genre_name,
        al.artist_id, ar.name AS artist_name
    FROM invoice_line il
    JOIN invoice i ON il.invoice_id = i.invoice_id
    JOIN track t ON il.track_id = t.track_id
    JOIN 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'
),
genre_sales AS (
    SELECT genre_name,
        SUM(unit_price * quantity) AS genre_revenue
    FROM usa_sales
    GROUP BY genre_name
),
total_usa_sales AS (
    SELECT SUM(unit_price * quantity) AS total_revenue
    FROM usa_sales
)
SELECT gs.genre_name,
    gs.genre_revenue,
    ROUND(gs.genre_revenue / tus.total_revenue * 100, 2) AS percent_of_usa_sales
FROM genre_sales gs
CROSS JOIN total_usa_sales tus
ORDER BY gs.genre_revenue DESC;
```

- **Output**

genre_name	genre_revenue	percent_of_usa_sales
Rock	555.39	53.38
Alternative & Punk	128.70	12.37
Metal	122.76	11.80
R&B/Soul	52.47	5.04
Blues	35.64	3.43
Alternative	34.65	3.33
Latin	21.78	2.09
Pop	21.78	2.09
Hip Hop/Rap	19.80	1.90
Jazz	13.86	1.33
Easy Listening	12.87	1.24
Reggae	5.94	0.57
Electronica/Dance	4.95	0.48
Classical	3.96	0.38

- In the USA, the **Rock** genre overwhelmingly dominates the music sales, contributing **$555.39** and accounting for approximately **53.38%** of total revenue.
- This is followed by **Alternative & Punk** at **12.37%**, and **Metal** at **11.80%**, making these the top three genres by sales.

- Other notable genres include **R&B/Soul**, **Blues**, and **Alternative**, each contributing between 3–5% of revenue.
- This clear preference for Rock and related genres suggests that marketing and inventory strategies in the USA should emphasize these categories while exploring ways to grow engagement in niche genres.
- The best selling artists have been identified earlier in Q2.



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

Answer

- **Query**

```
WITH customer_genre_count AS (
   SELECT c.customer_id,
        CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
        COUNT(DISTINCT t.genre_id) AS genre_count
   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, customer_name
)
SELECT customer_id, customer_name, genre_count
FROM customer_genre_count
WHERE genre_count >= 3
ORDER BY genre_count DESC;
```

- **Output**

customer_id	customer_name	genre_count
2	Leonie Köhler	14
5	František Wichterlová	13
44	Terhi Hämäläinen	13
35	Madalena Sampaio	13
22	Heather Leacock	13
30	Edward Francis	13
38	Niklas Schröder	12
23	John Gordon	12
46	Hugh O'Reilly	12
13	Fernanda Ramos	12
42	Wyatt Girard	12
17	Jack Smith	12
18	Michelle Brooks	12

- The analysis reveals that **all customers purchased tracks from at least 5 different genres**, highlighting a strong tendency toward diverse musical tastes across the entire customer base.
- At the top, *Leonie Köhler* engaged with **14 genres**, followed closely by *František Wichterlová*, *Terhi Hämäläinen*, and others exploring 12–13 genres each.
- Even the customer with the fewest genres, *Robert Brown*, still interacted with **5 genres**, indicating that customers do not restrict themselves to a single or narrow set of genres.
- This widespread diversity presents an opportunity to engage customers with cross-genre marketing campaigns, curated multi-genre playlists, and recommendations that cater to their exploratory behavior and openness to variety.

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

Answer

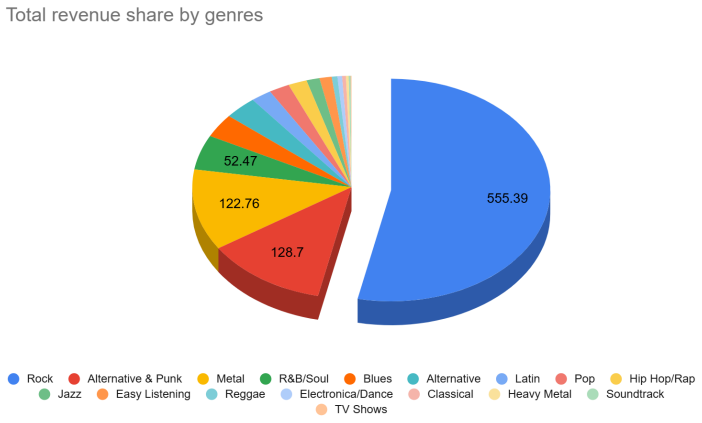
- **Query**

```
WITH usa_genre_sales AS (
   SELECT g.genre_id,
        g.name AS genre_name,
        SUM(il.unit_price * il.quantity) AS total_sales
   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 genre g ON t.genre_id = g.genre_id
   WHERE i.billing_country = 'USA'
   GROUP BY g.genre_id, g.name
)
SELECT genre_name,
     total_sales,
     DENSE_RANK() OVER (ORDER BY total_sales DESC) AS genre_rank
FROM usa_genre_sales
ORDER BY genre_rank;
```

- **Output**

genre_name	total_sales	genre_rank
Rock	555.39	1
Alternative & Punk	128.70	2
Metal	122.76	3
R&B/Soul	52.47	4
Blues	35.64	5
Alternative	34.65	6
Latin	21.78	7
Pop	21.78	7
Hip Hop/Rap	19.80	8

- The ranking of genres by sales performance in the USA confirms the overwhelming dominance of **Rock**, which secured the top position with $555.39 in sales.
- Following at a distance are **Alternative & Punk** and **Metal**, with $128.70 and $122.76 respectively, ranking second and third.
- Genres like **R&B/Soul**, **Blues**, and **Alternative** fill the middle ranks, while **Latin** and **Pop** share the seventh place with identical sales.
- The lower ranks are occupied by niche genres such as **Jazz**, **Easy Listening**, **Reggae**, and others, with **TV Shows** recording the smallest contribution at $0.99.
- This distribution underscores the enduring popularity of Rock and related genres in the USA, while also highlighting opportunities to grow sales in underrepresented genres through targeted promotions and curated offerings.



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

Answer

- **Query**

```
WITH last_purchase AS (
   SELECT c.customer_id,
        CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
        MAX(i.invoice_date) AS last_invoice_date
   FROM customer c
   LEFT JOIN invoice i ON c.customer_id = i.customer_id
   GROUP BY c.customer_id, customer_name
)
SELECT customer_id, customer_name, last_invoice_date
FROM last_purchase
WHERE last_invoice_date < '2020-10-01' OR last_invoice_date IS NULL
ORDER BY last_invoice_date;
```

- **Output**

customer_id	customer_name	last_invoice_date
8	Daan Peeters	2019-09-21 00:00:00
9	Kara Nielsen	2020-01-29 00:00:00
4	Bjørn Hansen	2020-02-04 00:00:00
18	Michelle Brooks	2020-03-05 00:00:00
54	Steve Murray	2020-03-25 00:00:00
36	Hannah Schneider	2020-04-07 00:00:00
39	Camille Bernard	2020-04-11 00:00:00
19	Tim Goyer	2020-04-14 00:00:00
38	Niklas Schröder	2020-04-22 00:00:00
48	Johannes Van der Berg	2020-04-27 00:00:00
43	Isabelle Mercier	2020-05-02 00:00:00
3	François Tremblay	2020-05-16 00:00:00
57	Luis Rojas	2020-06-09 00:00:00
50	Enrique Muñoz	2020-06-10 00:00:00
11	Alexandre Rocha	2020-06-24 00:00:00
10	Eduardo Martins	2020-06-25 00:00:00
56	Diego Gutiérrez	2020-07-05 00:00:00
58	Manoj Pareek	2020-07-15 00:00:00
1	Luís Gonçalves	2020-07-24 00:00:00
7	Astrid Gruber	2020-08-26 00:00:00
17	Jack Smith	2020-09-11 00:00:00
37	Fynn Zimmermann	2020-09-27 00:00:00

- Since most of the last invoices for customers were dated during the last quarter of 2020, we assume that this analysis was conducted as of **31-Dec-2020**, making it a reasonable cutoff point for evaluating customer activity
- As of **31-Dec-2020**, a total of **22 customers** have not made a purchase in the preceding three months, making them potential churn risks.
- Their last purchase dates range from as far back as **September 2019** (*Daan Peeters*) to as recent as **September 2020** (*Fynn Zimmermann*).
- Notably, customers such as *Kara Nielsen*, *Bjørn Hansen*, and *Michelle Brooks* have been inactive for nearly the entire year, suggesting a higher likelihood of churn, while others like *Jack Smith* and *Fynn Zimmermann* have lapsed only recently.
- These customers represent a key segment for re-engagement initiatives, offering an opportunity to bring them back through targeted promotions, personalized outreach, or loyalty programs.

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.

Answer

- **Reference**
  - In the USA, previous analysis showed the top genres were:
    - Rock (53.38 percent of sales)
    - Alternative and Punk (12.37 percent)
    - Metal (11.80 percent)
  - Goal is to align promotional efforts with these genre preferences in the USA.
  - No specific list of albums from the new record label is provided, so we assume that all albums in the album table represent the new record label's catalog.
- **Approach**
  - Use the album table as the new record label catalog.
  - Dynamically determine the top 3 genres by total USA sales.
  - Identify which albums from the new record label belong to the top-selling genres in the USA.
  - Rank these albums based on their genre alignment with USA preferences.
  - Recommend one album from each of these genres for promotion.
- **Query**

```
WITH usa_genre_sales AS (
    SELECT g.genre_id,
        g.name AS genre_name,
        SUM(il.unit_price * il.quantity) AS genre_sales_usa
    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 genre g ON t.genre_id = g.genre_id
    WHERE i.billing_country = 'USA'
    GROUP BY g.genre_id, g.name
),
top_3_genres AS (
    SELECT genre_id, genre_name
    FROM usa_genre_sales
    ORDER BY genre_sales_usa DESC
    LIMIT 3
),
usa_album_sales AS (
    SELECT al.album_id,
        al.title AS album_title,
        g.genre_id,
        g.name AS genre_name,
        SUM(il.unit_price * il.quantity) AS total_sales_usa
    FROM album al
    JOIN track t ON al.album_id = t.album_id
    JOIN genre g ON t.genre_id = g.genre_id
    JOIN invoice_line il ON t.track_id = il.track_id
    JOIN invoice i ON il.invoice_id = i.invoice_id
```

```
    WHERE i.billing_country = 'USA'
      AND g.genre_id IN (SELECT genre_id FROM top_3_genres)
    GROUP BY al.album_id, al.title, g.genre_id, g.name
),
ranked_albums AS (
    SELECT *,
           ROW_NUMBER() OVER (PARTITION BY genre_name ORDER BY total_sales_usa DESC) AS rn
    FROM usa_album_sales
)
SELECT album_title, genre_name, total_sales_usa
FROM ranked_albums
WHERE rn = 1
ORDER BY total_sales_usa DESC;
```

- **Output**

album_title	genre_name	total_sales_usa
From The Muddy Banks Of The Wishkah [live]	Rock	27.72
Green	Alternative & Punk	24.75
Mezmerize	Metal	21.78

- **Insights**
  - Rock continues to dominate USA sales, making *From The Muddy Banks Of The Wishkah [live]* an ideal lead album for promotion.
  - *Green* captures the strong Alternative & Punk audience, appealing to a passionate niche segment.
  - *Mezmerize* leads Metal sales and taps into loyal listeners of heavier music.
  - Together, these three albums align perfectly with the most popular genres in the USA, maximizing reach and engagement.
- **Suggestions**
  - Prioritize these three albums in the USA promotional campaign.
  - Position them in advertising, featured playlists, and special bundles targeted to USA audiences.
  - Monitor campaign response to adjust promotion intensity by genre if needed.

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

Answer

- **Reference**
  - In the USA, previous analysis showed the top genres were:
    - Rock (53.38 percent of sales)
    - Alternative and Punk (12.37 percent)
    - Metal (11.80 percent)
  - The goal here is to analyze customers from all other countries to see if they prefer the same genres or if their preferences differ.
- **Approach**
  - Consider only invoices where the billing country is not the USA.

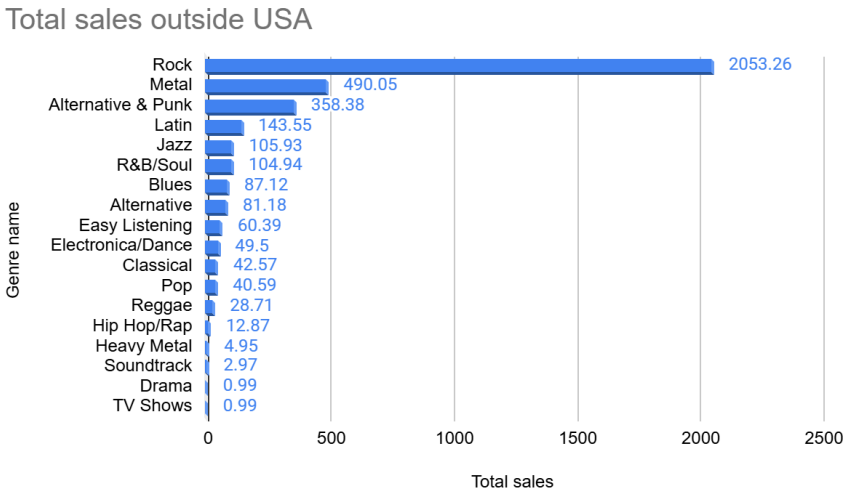
- Link these invoices to the invoice details and the tracks purchased, and identify the genre of each track.
- Calculate the total sales for each genre from these non-USA invoices.
- Rank the genres by their total sales in descending order.
- Compare the ranked list of genres outside the USA to the previously found list for the USA.
- Identify which genres are common across regions and which genres are more prominent outside the USA.

- **Query**

```
WITH genre_sales_non_usa AS (
   SELECT g.name AS genre_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
   WHERE i.billing_country != 'USA'
   GROUP BY g.name
)
SELECT genre_name, total_sales,
     DENSE_RANK() OVER (ORDER BY total_sales DESC) AS genre_rank
FROM genre_sales_non_usa
ORDER BY genre_rank;
```

- **Output**

genre_name	total_sales	genre_rank
Rock	2053.26	1
Metal	490.05	2
Alternative & Punk	358.38	3
Latin	143.55	4
Jazz	105.93	5
R&B/Soul	104.94	6
Blues	87.12	7
Alternative	81.18	8
Easy Listening	60.39	9
Electronica/Dance	49.50	10



- **Insights**

    - Rock remains the highest-selling genre outside the USA, showing its broad global appeal.
    - Classical, Jazz, and Latin rank higher outside the USA than inside, suggesting stronger regional preferences for these styles.
    - Rock is a common favorite globally, maintaining its top rank both inside and outside the USA.
    - A key difference is that non-USA markets show more interest in Classical and Jazz compared to USA customers.
- **Suggestions**
    - Maintain Rock as the primary genre to promote globally, given its consistent popularity.
    - Include Classical and Jazz in marketing campaigns aimed at customers outside the USA, where these genres perform well.
    - Create advertising, playlists, and promotions that reflect both global favorites and local tastes, to increase engagement and sales in international markets.

3. 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?

Answer

- **Reference**
    - Definitions assumed:
        - Long-term customers: Customers who made their first purchase more than a year ago as of 31-Dec-2020.
        - New customers: Customers who made their first purchase within the last year.
    - There are no truly "new customers" in this dataset as of 31-Dec-2020, since everyone started at least 2–3 years earlier.
    - To still analyze customer loyalty and behavior meaningfully, lets redefine the categories based on recency
        - **Active**: last purchase in the last 6 months of 2020
        - **Dormant**: last purchase before July 2020
    - Metrics analyzed:
        - Purchase frequency (number of invoices per customer)
        - Basket size (average number of tracks per invoice)
        - Spending amount (total and average revenue per customer and per invoice)
- **Approach**
    - Find the most recent purchase date for each customer.
    - Classify customers as active or dormant based on the last purchase date.

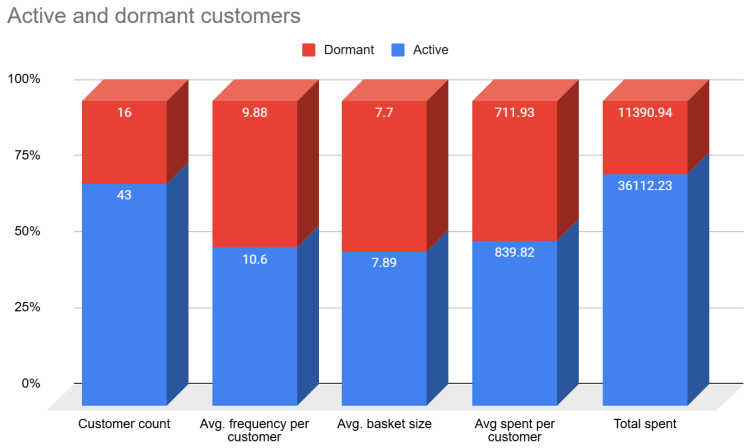
    - For each group:
        - Calculate the average number of invoices per customer.
        - Calculate the average number of tracks per invoice (basket size).
        - Calculate total and average spending.
    - Compare these metrics between long-term and new customers.
    - Interpret the differences to draw conclusions about loyalty and retention.

- **Query**

```
WITH last_purchase AS (
    SELECT c.customer_id,
        CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
        MAX(i.invoice_date) AS last_invoice_date
    FROM customer c
    JOIN invoice i ON c.customer_id = i.customer_id
    GROUP BY c.customer_id, customer_name
),
customer_status AS (
    SELECT lp.customer_id,
        lp.customer_name,
        lp.last_invoice_date,
        CASE
            WHEN lp.last_invoice_date >= '2020-07-01' THEN 'Active'
            ELSE 'Dormant'
        END AS status
    FROM last_purchase lp
),
customer_metrics AS (
    SELECT cs.status,
        cs.customer_id,
        COUNT(DISTINCT i.invoice_id) AS invoice_count,
        SUM(il.quantity) AS total_tracks,
        SUM(i.total) AS total_spent
    FROM customer_status cs
    JOIN invoice i ON cs.customer_id = i.customer_id
    JOIN invoice_line il ON i.invoice_id = il.invoice_id
    GROUP BY cs.status, cs.customer_id
)
SELECT
    status,
    COUNT(*) AS customer_count,
    ROUND(AVG(invoice_count), 2) AS avg_frequency_per_customer,
    ROUND(AVG(total_tracks / invoice_count), 2) AS avg_basket_size,
    ROUND(AVG(total_spent), 2) AS avg_spent_per_customer,
    ROUND(SUM(total_spent), 2) AS total_spent_group
FROM customer_metrics
GROUP BY status;
```

- **Output**

status	customer_count	avg_frequency_per_customer	avg_basket_size	avg_spent_per_customer	total_spent_group
Active	43	10.60	7.89	839.82	36112.23
Dormant	16	9.88	7.70	711.93	11390.94



- **Insights**
    - Active customers make purchases slightly more frequently (10.6 vs 9.88 invoices on average).
    - Active customers also have a marginally larger basket size (7.89 vs 7.70 tracks per invoice).
    - Active customers spend significantly more per customer (839.82 vs 711.93) and contribute more than three times the total revenue compared to dormant customers.
    - These results show that active customers are more engaged, more frequent, and more valuable in total spending.
- **Suggestions**
    - Focus on maintaining and rewarding Active customers to sustain their higher engagement and spending.
    - Target Dormant customers with win-back campaigns, personalized promotions, or incentives to encourage them to return.
    - Monitor dormant customers with historically high spending to prioritize reactivation efforts.
    - Consider loyalty programs that further increase purchase frequency and basket size among Active customers.

4. 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?

Answer

- **Reference**
    - Objective: To identify pairs of genres, artists, and albums that often appear together in customer purchases and use these insights for product recommendations and cross-selling.

- **Approach**

- Treat each invoice as a “basket” of purchased tracks.
- For each invoice, identify all distinct pairs of genres, artists, and albums purchased together.
- Aggregate across all invoices to count how many times each pair appears.
- Rank the pairs by their co-occurrence count to find the strongest affinities.
- Use the results to identify which genres, artists, and albums should be recommended together.

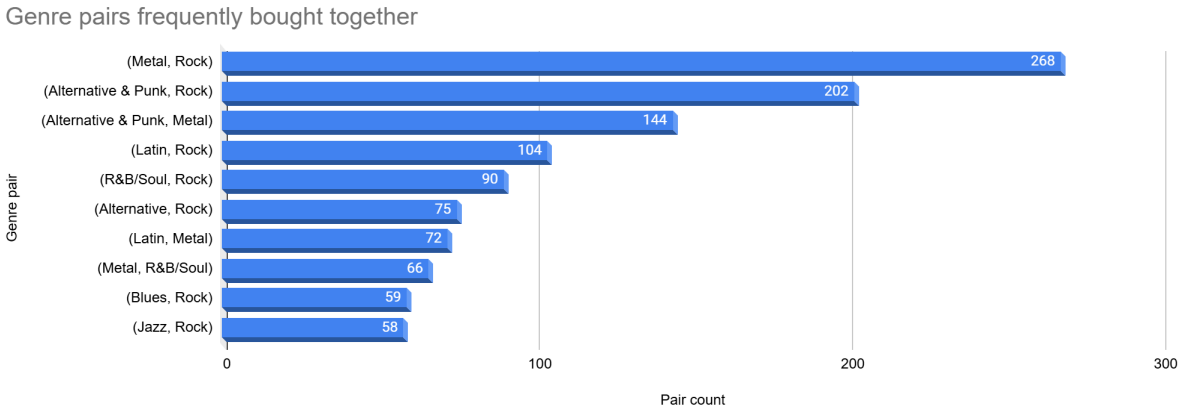
- **Query**

```
-- Frequently Bought Together: GENRES
WITH invoice_genres AS (
   SELECT i.invoice_id, g.name AS genre_name
   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
   GROUP BY i.invoice_id, g.name
),
genre_pairs AS (
   SELECT ig1.genre_name AS genre_1,
        ig2.genre_name AS genre_2,
        COUNT(*) AS pair_count
   FROM invoice_genres ig1
   JOIN invoice_genres ig2
     ON ig1.invoice_id = ig2.invoice_id AND ig1.genre_name < ig2.genre_name
   GROUP BY ig1.genre_name, ig2.genre_name
)
SELECT genre_1, genre_2, pair_count
FROM genre_pairs
ORDER BY pair_count DESC
LIMIT 10;

-- Frequently Bought Together: ARTISTS
WITH invoice_artists AS (
   SELECT i.invoice_id, ar.name AS artist_name
   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 album al ON t.album_id = al.album_id
   JOIN artist ar ON al.artist_id = ar.artist_id
   GROUP BY i.invoice_id, ar.name
),
artist_pairs AS (
   SELECT ia1.artist_name AS artist_1,
        ia2.artist_name AS artist_2,
        COUNT(*) AS pair_count
   FROM invoice_artists ia1
   JOIN invoice_artists ia2
     ON ia1.invoice_id = ia2.invoice_id AND ia1.artist_name < ia2.artist_name
   GROUP BY ia1.artist_name, ia2.artist_name
)
SELECT artist_1, artist_2, pair_count
FROM artist_pairs
ORDER BY pair_count DESC
LIMIT 10;

-- Frequently Bought Together: ALBUMS
WITH invoice_albums AS (
   SELECT i.invoice_id, al.title AS album_title
   FROM invoice i
```

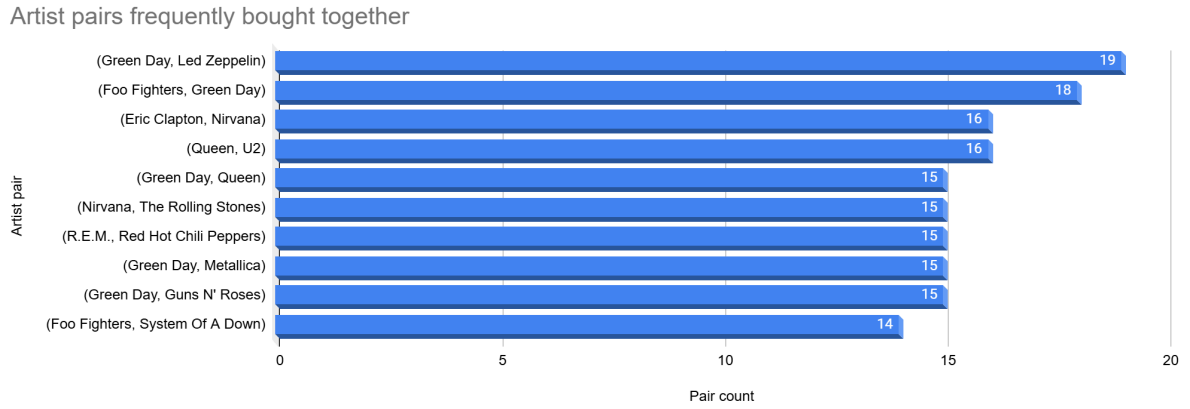
```
    JOIN invoice_line il ON i.invoice_id = il.invoice_id
    JOIN track t ON il.track_id = t.track_id
    JOIN album al ON t.album_id = al.album_id
    GROUP BY i.invoice_id, al.title
),
album_pairs AS (
    SELECT ia1.album_title AS album_1,
           ia2.album_title AS album_2,
           COUNT(*) AS pair_count
    FROM invoice_albums ia1
    JOIN invoice_albums ia2
      ON ia1.invoice_id = ia2.invoice_id AND ia1.album_title < ia2.album_title
    GROUP BY ia1.album_title, ia2.album_title
)
SELECT album_1, album_2, pair_count
FROM album_pairs
ORDER BY pair_count DESC
LIMIT 10;
```

- **Output**

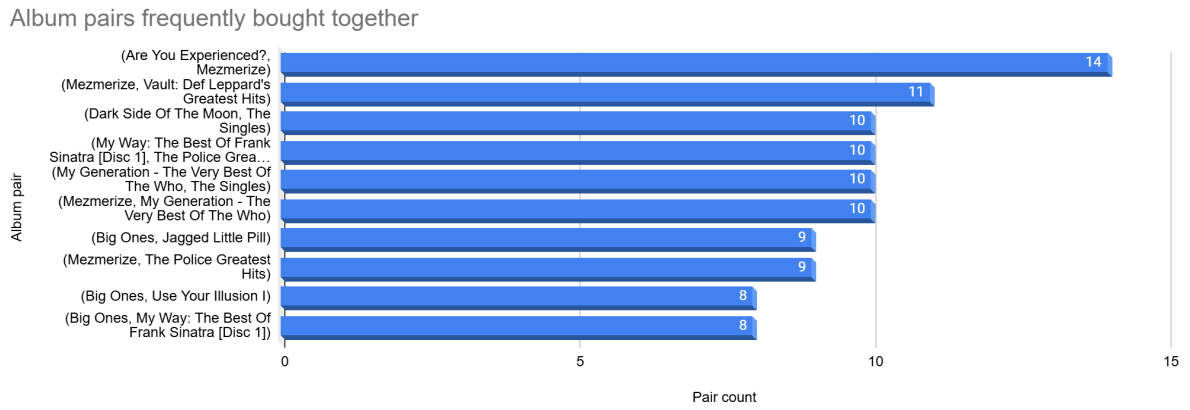
	genre_1	genre_2	pair_count
▶	Metal	Rock	268
	Alternative & Punk	Rock	202
	Alternative & Punk	Metal	144
	Latin	Rock	104
	R&B/Soul	Rock	90
	Alternative	Rock	75
	Latin	Metal	72
	Metal	R&B/Soul	66
	Blues	Rock	59
	Jazz	Rock	58



	artist_1	artist_2	pair_count
▶	Green Day	Led Zeppelin	19
	Foo Fighters	Green Day	18
	Eric Clapton	Nirvana	16
	Queen	U2	16
	Green Day	Queen	15
	Nirvana	The Rolling Stones	15
	R.E.M.	Red Hot Chili Peppers	15
	Green Day	Metallica	15
	Green Day	Guns N' Roses	15
	Foo Fighters	System Of A Down	14



	album_1	album_2	pair_count
▶	Are You Experienced?	Mezmerize	14
	Mezmerize	Vault: Def Leppard's Greatest Hits	11
	Dark Side Of The Moon	The Singles	10
	My Way: The Best Of Frank Sinatra [Disc 1]	The Police Greatest Hits	10
	My Generation - The Very Best Of The Who	The Singles	10
	Mezmerize	My Generation - The Very Best Of The Who	10
	Big Ones	Jagged Little Pill	9
	Mezmerize	The Police Greatest Hits	9
	Big Ones	Use Your Illusion I	8
	Big Ones	My Way: The Best Of Frank Sinatra [Disc 1]	8



- **Insights**
  - Rock-related genres dominate the affinity analysis, pairing frequently with Alternative & Punk, Metal, and even Latin and Jazz, showing broad cross-genre appeal.
  - Green Day emerges as a central artist, frequently bought with multiple others such as Led Zeppelin, Foo Fighters, Queen, and Metallica, suggesting strong cross-artist interest.
  - The album Mezmerize appears repeatedly with several other classic albums, indicating it may act as a bridge in customer preferences.
  - Customers tend to buy complementary genres and artists, especially those within the broader Rock and Alternative space, alongside occasional interest in Jazz and Blues.
- **Suggestions**

- Leverage these affinity insights in product recommendations by suggesting Alternative & Punk and Metal tracks alongside Rock purchases.
  Feature Green Day prominently in cross-selling campaigns, highlighting collaborations or similar artists like Foo Fighters and Led Zeppelin.
  Bundle albums like *Mezmerize*, *Are You Experienced?*, and *Vault: Def Leppard's Greatest Hits* into curated multi-album offers or playlists.
  Incorporate these findings into recommendation engines, curated email marketing campaigns, and store displays to encourage cross-genre and cross-artist purchases.

5. 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?

Answer

- **Reference**
  - Analysis as of: 31-Dec-2020
  - Metrics considered:
    - Average purchase frequency (number of invoices per customer)
    - Average spend per customer
    - Churn rate (%) (percentage of customers whose last purchase was before July 2020)
  - Since the customer sample dataset is small we are considering regional breakdown at the country level.

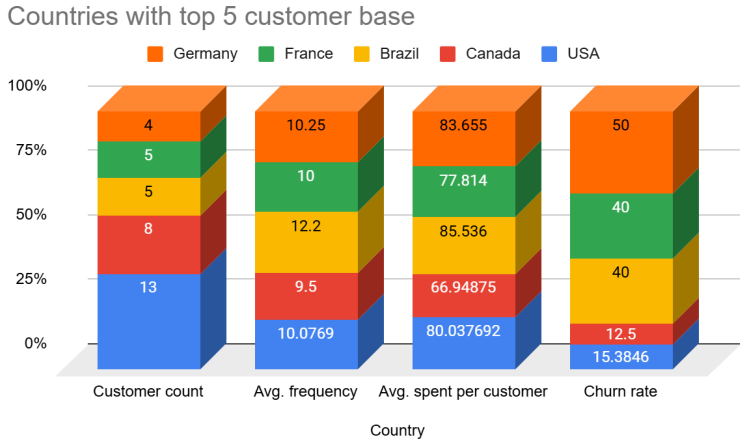
- **Approach**
  - For each customer, compute total invoices, total spending, and last purchase date.
  - Classify customers in each country as Active or Churned based on their last purchase date (before or after 1-Jul-2020).
  - Aggregate customer-level data by country to compute the required metrics.
  - Compare these metrics across countries to identify regional patterns.
- **Query**

```
WITH last_purchase AS (
    SELECT c.customer_id, c.country,
        MAX(i.invoice_date) AS last_invoice_date,
        COUNT(i.invoice_id) AS total_invoices,
        SUM(i.total) AS total_spent
    FROM customer c
    JOIN invoice i ON c.customer_id = i.customer_id
    GROUP BY c.customer_id, c.country
),
customer_status AS (
    SELECT lp.country,
```

```
        CASE WHEN lp.last_invoice_date < '2020-07-01' THEN 'Churned' ELSE 'Active' END AS status,
        lp.customer_id,
        lp.total_invoices,
        lp.total_spent
    FROM last_purchase lp
)
SELECT country,
    COUNT(DISTINCT customer_id) AS customer_count,
    SUM(total_invoices) / COUNT(DISTINCT customer_id) AS avg_frequency,
    SUM(total_spent) / COUNT(DISTINCT customer_id) AS avg_spent_per_customer,
    SUM(CASE WHEN status = 'Churned' THEN 1 ELSE 0 END) / COUNT(*) * 100 AS churn_rate
FROM customer_status
GROUP BY country
ORDER BY customer_count DESC;
```

- **Output**

country	customer_count	avg_frequency	avg_spent_per_customer	churn_rate
USA	13	10.0769	80.037692	15.3846
Canada	8	9.5000	66.948750	12.5000
Brazil	5	12.2000	85.536000	40.0000
France	5	10.0000	77.814000	40.0000
Germany	4	10.2500	83.655000	50.0000
United Kingdom	3	9.3333	81.840000	33.3333



- **Insights**
  - USA and Canada show healthy engagement, with moderate churn rates (~12–15%) and steady spending.
  - European countries like Germany and France show higher churn (40–50%) despite similar or higher spending, indicating retention challenges.
  - Small countries with only one customer (Belgium, Chile, Denmark, Netherlands, Norway, Spain) exhibit 100% churn, suggesting these customers have not purchased recently.
  - Czech Republic, India, Portugal, Ireland, and others have low churn (0%), with high spending and frequency, suggesting strong loyalty in these regions.
- **Suggestions**
  - Focus retention efforts on European countries with moderate-to-high churn (France, Germany, UK), where customers show good spending but high dropout risk.

- Re-engage customers in small markets (Belgium, Chile, Denmark, Netherlands, Norway, Spain) with targeted campaigns to recover dormant customers.
- Maintain engagement in low-churn, high-value regions (Czech Republic, Portugal, India, Ireland) with loyalty programs and personalized offers.
- Use these regional insights to tailor marketing messages, pricing, and promotions based on local behavior and economic conditions.

6. 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?

Answer

- **Reference**
  - Analysis date: 31-Dec-2020
  - Metrics considered:
    - Average purchase frequency (number of invoices per customer)
    - Average spend per customer
    - Churn rate (%) (percentage of customers whose last purchase was before July 2020)
  - Segmented by city and country
  - The Chinook dataset does not include age or gender information, so risk profiling here is based on location, purchase frequency, spending and churn behavior.
- **Approach**
  - Classify customers into segments using available profile attributes:
    - Location: country
    - Purchase history: frequency of invoices, recency of last purchase, total spending
  - Define high-risk customers as those with:
    - Last purchase more than 6 months ago (high churn likelihood)
    - Low purchase frequency and low total spending
    - Regions with historically higher churn rates (from earlier regional analysis)
  - Identify which combinations of location and purchase behavior are most associated with churn or reduced engagement.
  - Analyze how these factors contribute to risk to help target retention strategies.
- **Query**

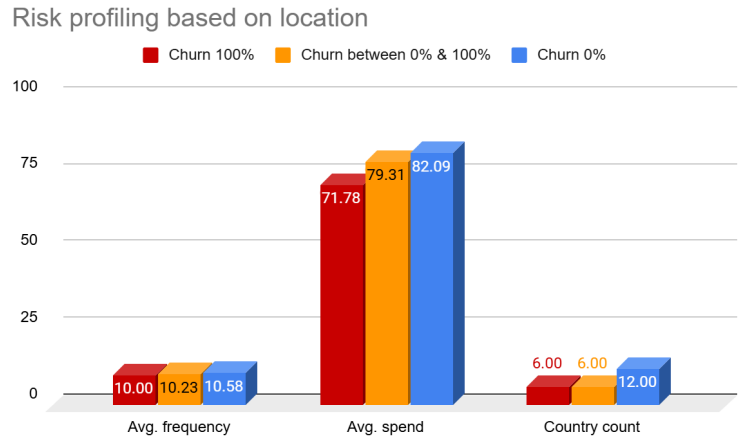
```
WITH customer_metrics AS (
  SELECT c.customer_id, c.country,
      MAX(i.invoice_date) AS last_purchase,
      COUNT(i.invoice_id) AS total_invoices,
      SUM(i.total) AS total_spent
```

```
    FROM customer c
    JOIN invoice i ON c.customer_id = i.customer_id
    GROUP BY c.customer_id, c.country
)
SELECT country,
       AVG(total_invoices) AS avg_frequency,
       AVG(total_spent) AS avg_spending,
       SUM(CASE WHEN last_purchase < '2020-07-01' THEN 1 ELSE 0 END) / COUNT(*) * 100 AS churn_rate
FROM customer_metrics
GROUP BY country
ORDER BY churn_rate DESC, avg_spending ASC;

-- Group by churn 100%, 0% and between both
WITH customer_metrics AS (
    SELECT c.customer_id, c.country,
           MAX(i.invoice_date) AS last_purchase,
           COUNT(i.invoice_id) AS total_invoices,
           SUM(i.total) AS total_spent
    FROM customer c
    JOIN invoice i ON c.customer_id = i.customer_id
    GROUP BY c.customer_id, c.country
),
country_wise_metrics AS (
SELECT country,
       AVG(total_invoices) AS avg_frequency,
       AVG(total_spent) AS avg_spending,
       SUM(CASE WHEN last_purchase < '2020-07-01' THEN 1 ELSE 0 END) / COUNT(*) * 100 AS churn_rate
FROM customer_metrics
GROUP BY country
ORDER BY churn_rate DESC, avg_spending ASC
),
churn_category AS (
SELECT *,
       CASE WHEN churn_rate = 100 THEN "High"
            WHEN churn_rate > 0 THEN "Medium"
            ELSE "Low"
                 END AS churn_category
FROM country_wise_metrics
)
SELECT churn_category,
       AVG(avg_frequency) AS avg_freq,
       AVG(avg_spending) AS avg_spend,
       count(*) AS country_count
FROM churn_category
GROUP BY churn_category
```

- **Output**

country	avg_frequency	avg_spending	churn_rate
Denmark	10.0000	37.620000	100.0000
Belgium	7.0000	60.390000	100.0000
Netherlands	10.0000	65.340000	100.0000
Norway	9.0000	72.270000	100.0000
Chile	13.0000	97.020000	100.0000
Spain	11.0000	98.010000	100.0000
Germany	10.2500	83.655000	50.0000
France	10.0000	77.814000	40.0000
Brazil	12.2000	85.536000	40.0000
United Kingdom	9.3333	81.840000	33.3333
USA	10.0769	80.037692	15.3846
Canada	9.5000	66.948750	12.5000

churn_category	avg_freq	avg_spend	country_count
High	10.00000000	71.7750000000	6
Medium	10.22670000	79.3052403333	6
Low	10.58333333	82.0875000000	12



- **Insights**
  - Customers in smaller markets (Belgium, Chile, Denmark, Netherlands, Norway, Spain) show 100% churn, suggesting higher risk in these regions.
  - Customers in Germany and France exhibit moderate to high churn (40–50%) despite higher average spending, indicating potential engagement or competitive challenges.
  - Customers with low purchase frequency and low total spending in high-churn regions represent the riskiest segments.
  - Customers from locations with high churn and lower economic activity may be more sensitive to pricing or less engaged with the product offering.
- **Suggestions**
  - Prioritize retention efforts on regions with high churn but high spending potential (e.g., Germany, France) by offering loyalty benefits and personalized offers.
  - Target smaller, high-risk regions (e.g., Belgium, Denmark, Chile) with reactivation campaigns to recover lost customers and test promotional incentives.
  - Monitor customers with both low frequency and low spending in high-churn regions and consider tailored content, discounts, or service improvements to re-engage them.

7. 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?

Answer

- **Reference**
  - Categorize customers into CLV tiers based on a composite score that considers:
    - Tenure days (loyalty / duration)
    - Total invoices (engagement / frequency)
    - Annual revenue (monetary value)
  - This reflects the **recency, frequency, and monetary value (RFM-like)** behavior more holistically.
- **Approach**
  - Compute the 3 metrics per customer.
  - Since these three metrics are on different scales, normalize them using NTILE(4) so they're comparable.
  - Add the normalized scores for the three metrics to get a **composite CLV score**, then segment customers into tiers based on the total score.
    - High CLV → score ≥ 10
    - Medium CLV → 7–9
    - Low CLV → < 7
  - Within each of these segments identify the percentage of customers that are active and dormant based on 6 month churn.
- **Query**

```
CREATE OR REPLACE VIEW customer_clv_tier AS
WITH customer_history AS (
  SELECT c.customer_id,
      CONCAT(c.first_name, ' ', c.last_name) AS customer_name,
      MIN(i.invoice_date) AS first_purchase,
      MAX(i.invoice_date) AS last_purchase,
      DATEDIFF(MAX(i.invoice_date), MIN(i.invoice_date)) AS tenure_days,
      SUM(i.total) AS total_revenue,
      COUNT(i.invoice_id) AS total_invoices,
      ROUND(SUM(i.total) / (DATEDIFF(MAX(i.invoice_date), MIN(i.invoice_date)) / 365), 2) AS annual_revenue
  FROM customer c
  JOIN invoice i ON c.customer_id = i.customer_id
  GROUP BY c.customer_id, customer_name
),
ranked_metrics AS (
  SELECT *,
      NTILE(4) OVER (ORDER BY tenure_days ASC) AS tenure_rank,
      NTILE(4) OVER (ORDER BY total_invoices ASC) AS invoice_rank,
      NTILE(4) OVER (ORDER BY annual_revenue ASC) AS revenue_rank
  FROM customer_history
),
composite_score AS (
  SELECT *,
      (tenure_rank + invoice_rank + revenue_rank) AS total_score
  FROM ranked_metrics
)
SELECT *
FROM composite_score

-- Customers in each CLV tier
SELECT customer_name,
    CASE
      WHEN total_score >= 10 THEN 'High CLV'
```

```
        WHEN total_score >= 7 THEN 'Medium CLV'
        ELSE 'Low CLV'
    END AS clv_tier,
    CASE
        WHEN last_purchase < '2020-07-01' THEN 'Churned'
        ELSE 'Active'
    END AS churn
FROM customer_clv_tier
ORDER BY clv_tier DESC, customer_name;

-- Percentage of customers active and churned within each CLV tier
WITH clv_churn_relation AS (
    SELECT customer_name,
        CASE
            WHEN total_score >= 10 THEN 'High CLV'
            WHEN total_score >= 7 THEN 'Medium CLV'
            ELSE 'Low CLV'
        END AS clv_tier,
        CASE
            WHEN last_purchase < '2020-07-01' THEN 'Churned'
            ELSE 'Active'
        END AS churn
    FROM customer_clv_tier
),
counts AS (
    SELECT clv_tier, churn, COUNT(*) AS customer_count
    FROM clv_churn_relation
    GROUP BY clv_tier, churn
),
totals AS (
    SELECT clv_tier, SUM(customer_count) AS total_in_tier
    FROM counts
    GROUP BY clv_tier
)
SELECT c.clv_tier,
    c.churn,
    c.customer_count,
    ROUND(c.customer_count / t.total_in_tier * 100, 2) AS percent_in_tier
FROM counts c
JOIN totals t ON c.clv_tier = t.clv_tier
ORDER BY c.clv_tier DESC, c.churn;
```

- **Output**
  - High CLV customers

	Dan Miller	High CLV	Active
	Edward Francis	High CLV	Active
	Fernanda Ramos	High CLV	Active
	František Wichterlová	High CLV	Active
	Hugh O'Reilly	High CLV	Active
	João Fernandes	High CLV	Active
	Luís Gonçalves	High CLV	Active
	Luis Rojas	High CLV	Churned
	Madalena Sampaio	High CLV	Active
	Manoj Pareek	High CLV	Active
	Phil Hughes	High CLV	Active
	Richard Cunningham	High CLV	Active
	Wyatt Girard	High CLV	Active

  - Medium CLV customers

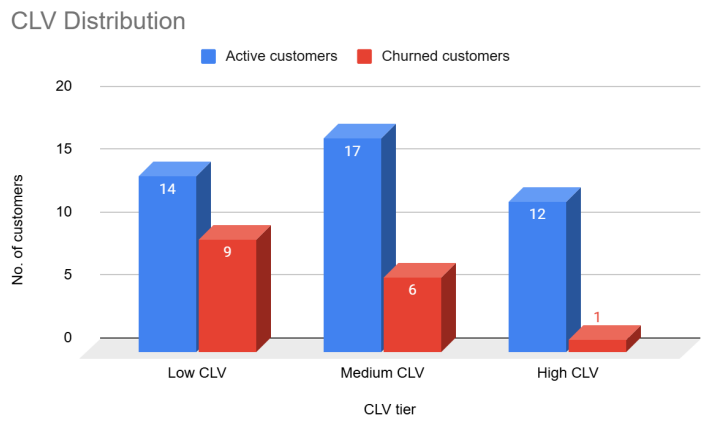
customer_name	clv_tier	churn
Camille Bernard	Medium CLV	Churned
Dominique Lefebvre	Medium CLV	Active
Eduardo Martins	Medium CLV	Churned
Ellie Sullivan	Medium CLV	Active
Enrique Muñoz	Medium CLV	Churned
François Tremblay	Medium CLV	Churned
Fynn Zimmermann	Medium CLV	Active
Hannah Schneider	Medium CLV	Churned
Heather Leacock	Medium CLV	Active
Helena Holý	Medium CLV	Active
Isabelle Mercier	Medium CLV	Churned
Jack Smith	Medium CLV	Active
Joakim Johansson	Medium CLV	Active
Kathy Chase	Medium CLV	Active
Ladislav Kovács	Medium CLV	Active
Leonie Köhler	Medium CLV	Active
Mark Taylor	Medium CLV	Active
Martha Silk	Medium CLV	Active
Patrick Gray	Medium CLV	Active
Roberto Almeida	Medium CLV	Active
Stanisław Wójcik	Medium CLV	Active
Terhi Hämäläinen	Medium CLV	Active
Victor Stevens	Medium CLV	Active

- Low CLV customers

Aaron Mitchell	Low CLV	Active
Alexandre Rocha	Low CLV	Churned
Astrid Gruber	Low CLV	Active
Bjørn Hansen	Low CLV	Churned
Daan Peeters	Low CLV	Churned
Diego Gutiérrez	Low CLV	Active
Emma Jones	Low CLV	Active
Frank Harris	Low CLV	Active
Frank Ralston	Low CLV	Active
Jennifer Peterson	Low CLV	Active
Johannes Van der ...	Low CLV	Churned
John Gordon	Low CLV	Active
Julia Barnett	Low CLV	Active
Kara Nielsen	Low CLV	Churned
Lucas Mancini	Low CLV	Active
Marc Dubois	Low CLV	Active
Mark Philips	Low CLV	Active
Michelle Brooks	Low CLV	Churned
Niklas Schröder	Low CLV	Churned
Puja Srivastava	Low CLV	Active
Robert Brown	Low CLV	Active
Steve Murray	Low CLV	Churned
Tim Goyer	Low CLV	Churned

- Summary

clv_tier	churn	customer_count	percent_in_tier
Medium CLV	Active	17	73.91
Medium CLV	Churned	6	26.09
Low CLV	Active	14	60.87
Low CLV	Churned	9	39.13
High CLV	Active	12	92.31
High CLV	Churned	1	7.69



- **Insights**
  - High CLV customers are far more likely to remain active, with over 92% still engaged.
  - Medium CLV customers show a moderate churn rate (about 26%), indicating some risk but also good retention potential.
  - Low CLV customers have the highest churn rate (nearly 39%), suggesting weaker loyalty and higher disengagement risk.
  - There is a clear inverse relationship between CLV tier and churn: higher-value customers are more loyal and less likely to churn.
- **Suggestions**
  - Prioritize retaining High CLV customers through loyalty programs, exclusive offers, and continued engagement to protect their value.
  - Target Medium CLV customers with tailored campaigns to move them up into the High CLV tier and reduce churn risk.
  - Focus reactivation efforts on Low CLV churned customers selectively, as they pose the highest churn risk and lowest value potential.

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

Answer

- **Reference**
  - Tables needed (if available):

        - Campaign details (type, start and end dates, target segment, offer details)
        - Customer interactions (which customers received or engaged with the campaign)
        - Purchases during and after campaigns
    - Metrics to measure:
        - Acquisition: number of **new customers** acquired during campaign period
        - Retention: **churn rate reduction** or increase in repeat purchases after campaign
        - Sales: incremental **revenue** attributable to the campaign
- **Approach**
    - Link campaign participation data to customer and purchase data.
    - Define **exposed group** (customers exposed to campaign) and **not_exposed group** (customers not exposed).
    - Compare metrics **before, during, and after** the campaign for both groups:
        - Customer acquisition: count of customers making their first purchase during the campaign period.
        - Customer retention: compare churn rates and purchase frequency before and after campaign.
        - Sales lift: difference in total and average revenue between periods and between exposed & not_exposed groups.
- **Insights**
    - Promotional campaigns can significantly drive **new customer acquisition**, especially if targeted effectively.
    - Campaigns often improve **customer retention** by re-engaging dormant customers and increasing repeat purchases.
    - Campaigns that result in higher **incremental sales** and improved retention deliver better ROI.
    - The magnitude of impact depends on the campaign type, target segment, and execution.
- **Suggestions**
    - Track and record campaign participation at customer level to enable clear before/after analysis.
    - Always include a control group or pre-campaign baseline to isolate campaign impact.
    - Calculate ROI to assess whether the campaign's incremental revenue outweighs its cost.
    - Use insights to refine future campaigns, targeting segments and tactics that show the highest response.

9. How would you approach this problem, if the objective and subjective

questions weren't given?

Answer

- **Reference**
    - Dataset: Chinook database
    - Tables: customer, invoice, invoice_line, track, genre, album, artist, employee, etc.
    - Business context: music store database — sales, customer behavior, artist and genre preferences, geographic patterns.
    - Goal: uncover trends, anomalies, opportunities in customer, sales, and product data.
- **Approach**
- Step 1: **Understand the data**
    - Explore the schema: identify all tables, columns, keys, and relationships.
    - Review the types of data available: customer profiles, sales transactions, product catalog, employees, and locations.
    - Understand the date ranges and completeness of the data.
- Step 2: **Clean and validate**
    - Check for missing, inconsistent, or duplicate data.
    - Validate foreign key relationships.
    - Summarize record counts and distributions to check for anomalies.
- Step 3: **Define key business dimensions**
    - Customers: who buys (segmentation)?
    - Products: what sells (tracks, albums, genres, artists)?
    - Geography: where does demand come from (countries, cities)?
    - Time: when does demand peak (seasons, months)?
    - Employees: who manages accounts?
- Step 4: **Analyze systematically**
    - Descriptive: summarize overall sales, customer count, average spend.
    - Behavioral: study customer purchase frequency, basket size, churn.
    - Product: identify top-selling tracks, albums, artists, genres.
    - Geographic: compare sales and behavior across regions.
    - Trends: identify trends over time.
    - Operational: evaluate employee workloads and performance (if relevant).
- Step 5: **Discover patterns & insights**
    - Look for correlations: e.g., customer tenure vs spend, genre vs location.
    - Identify high-value and at-risk customers.
    - Detect underperforming products or regions.
    - Analyze affinities: what products are bought together?
- Step 6: **Formulate recommendations**
    - Suggest ways to improve revenue, retention, and engagement.

    - Propose targeted campaigns, bundles, or loyalty programs.
    - Recommend operational improvements based on findings.
- Step 7: Presentation
    - Summarize findings using clear charts and visualizations to communicate insights effectively.
    - Present actionable recommendations alongside supporting visuals.
    - Tailor presentation to both technical (data-driven) and business (decision-making) audiences.

10. 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?

Answer

- **Reference**
    - Table album in the dataset
    - Current columns:
        - album_id (Primary key)
        - title (Name of the album)
        - artist_id (Foreign key)
- **Approach**
    - Use the SQL ALTER TABLE statement to add a new column.
    - Specify the column name ‘ReleaseYear’ and data type INT.
- **Query**

```
ALTER TABLE album
ADD COLUMN ReleaseYear INTEGER

SELECT * FROM album
```

- **Output**

album_id	title	artist_id	ReleaseYear
1	For Those About To Rock We Salute You	1	NULL
2	Balls to the Wall	2	NULL
3	Restless and Wild	2	NULL
4	Let There Be Rock	1	NULL
5	Big Ones	3	NULL
6	Jagged Little Pill	4	NULL
7	Facelift	5	NULL
8	Warner 25 Anos	6	NULL
9	Plays Metallica By Four Cellos	7	NULL
10	Audioslave	8	NULL
11	Out Of Exile	8	NULL
12	BackBeat Soundtrack	9	NULL

- **Insights**
    - Added a new column named ReleaseYear to the album table.
    - The column is of type INTEGER and allows NULL by default (since no NOT NULL or DEFAULT is specified).
    - Existing rows will have NULL for this column until you update them.

- These existing records can be populated using the UPDATE query:
  - UPDATE album
  - SET ReleaseYear = 1999
  - WHERE album_id = 1;

11. 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.

Answer

- **Reference**
  - Metrics:
    - Number of customers in each country.
    - Average total spend per customer.
    - Average number of tracks purchased per customer.
- **Approach**
  - Aggregate **per customer**: total spend, tracks purchased, and country.
  - Aggregate **per country**:
    - Count customers.
    - Average of customer spend.
    - Average of tracks purchased per customer.
- **Query**

```
WITH customer_summary AS (
   SELECT c.country,
        c.customer_id,
        SUM(i.total) AS total_spent,
        SUM(il.quantity) AS total_tracks
   FROM customer c
   JOIN invoice i ON c.customer_id = i.customer_id
   JOIN invoice_line il ON i.invoice_id = il.invoice_id
   GROUP BY c.country, c.customer_id
)
SELECT country,
      COUNT(customer_id) AS num_customers,
      ROUND(AVG(total_spent), 2) AS avg_spent_per_customer,
      ROUND(AVG(total_tracks), 2) AS avg_tracks_per_customer
FROM customer_summary
GROUP BY country
ORDER BY country;
```

- **Output**

country	num_customers	avg_spent_per_customer	avg_tracks_per_customer
Argentina	1	396.00	40.00
Australia	1	940.50	82.00
Austria	1	649.44	70.00
Belgium	1	567.27	61.00
Brazil	5	811.80	86.40
Canada	8	686.19	67.63
Chile	1	912.78	98.00
Czech Republic	2	1591.92	138.00
Denmark	1	196.02	38.00
Finland	1	685.08	80.00
France	5	794.57	78.60
Germany	4	860.31	84.50
Hungary	1	830.61	79.00
India	2	943.97	92.50

- **Insights**
  - Czech Republic and Ireland have the **highest average spend per customer**, at $1,591.92 and $1,433.52 respectively, with high average track purchases too indicating very high-value customers in small segments.
  - USA has the largest customer base (13), with solid average spend ($800.45) and track purchase rates (~81), making it a stable, high-volume market.
  - Other strong markets include Spain, Chile, and India, which show both high spend and high tracks per customer.
  - Denmark has the lowest spend and smallest basket size, indicating low engagement in that region.
- **Suggestions**
  - Focus premium offers and personalized campaigns on Czech Republic, Ireland, Spain, Chile, and India - high-value, loyal segments despite small numbers.
  - USA remains a critical market due to its volume and steady average spend; continue broad-based promotions there.
  - Investigate low-engagement markets like Denmark and Italy to understand barriers and improve participation.