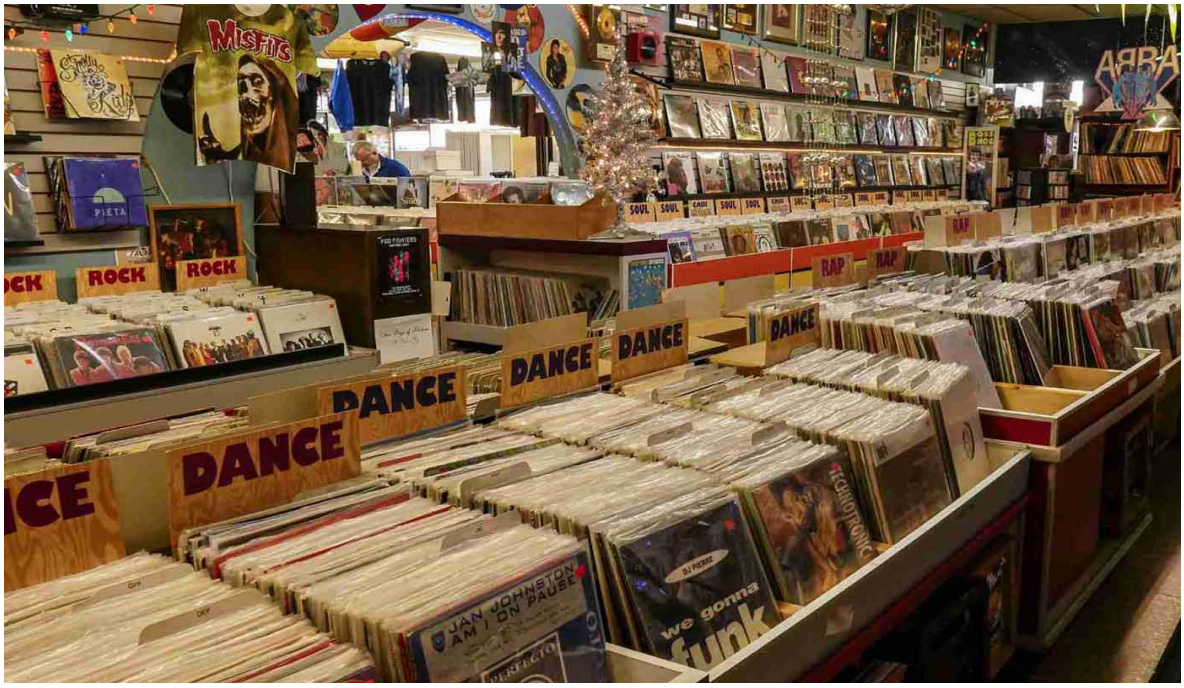
Music Store Data Analysis – SQL & Tableau project

This SQL & Tableau project is aimed at analysing Music Store data to glen meaningful information that the company can use to further its business (such as promotional activities, stock up on popular artist/genre etc).



The analysis will be carried with Excel, SQL and finally Tableau. Excel will serve as first repository for our data, SQL will give meaning to our data and Tableau will give a clear face to our data.

The following music store data analysis will follow the 6 steps of Data Analysis which are: Ask, Prepare, Process, Analyse, Share and Act.

Step 1: Ask

In this step, we will define the business objectives and deliverables expected out of this analysis. The music store, having global presence, would be keen on using customer spending data to further improve its sales and possibly look at regions not performing and drill down on the reason for the same.

Business objectives:

- How can we optimize our profits?
- What are the emerging trends that we can we identify?
- How can we take these insights to build recommendations?

We can address the business objectives by answering the following questions using SQL:

1. Which is the most profitable region & what is the top 5 most popular genre there?
2. Which are the top 5 non-performing region?
3. Where are the top 10 customers located?
4. Which genre has the most customers & which artist has sold the most records?
5. What is the top spending customer spending in each country?

Deliverables:

- A clear summary of the business objectives.
- A full documentation of all the data cleaning, manipulation and analysis.
- A dashboard with visualizations and main outcomes.
- Recommendations based on insights and analysis.

Step 2: Prepare

In this phase, we will identify and assess the features of our Music Store Dataset:

- The data is freely available to download from https://shorturl.at/DJMUX
- The data set has 12 CSV files, which contain the album details, invoice details, customer details, genre details, playlist details as well as artist and track details.
- The limitation with this data set is that we would have to individually clean each CSV file. Another limitation is that the customer data is limited to 59 customers, which would result in conclusions base only on a small data set.

Step 3: Process

We will process and clean our data with the help of Excel as the file is already a CSV file so a look through of our data with Excel can be ideal to Observe our data, check for missing data with the help of conditional formatting, Remove duplicate rows & Correctly format columns for easy SQL analysis.

While exploring our dataset, we can perform and notice the following;

- Our data looks correct and consistent.
- We utilized Excel's conditional formatting feature to highlight cells with empty values (missing data) within our dataset. Conditional formatting is a versatile tool that allows you to visually emphasize specific data based on predefined conditions. In this case, our objective was to pinpoint and accentuate cells lacking data (blanks) by using orange colour. There were no blank cells in our data
- Then we checked for duplicate data where applicable i.e., in genre, playlist & track Dataset.
- Finally, made sure all the data was in format that it represents i.e., album ID, customer ID, track ID etc had only numbers and address section had country details etc.

Now that our dataset has been appropriately configured, it is ready for analysis. This analysis will allow us to unearth connections, recognize patterns, and detect trends that can give us a competitive edge and effectively address our business objectives.

Step 4: Analyse

During the analysis stage, our focus will be on isolating and emphasizing the critical elements of our data to provide answers to our business objectives.
Let’s load our data into SQL and check the first 5 rows to make sure it imported well.

We will be using Postgre SQL 15 to run queries and check the Album dataset.

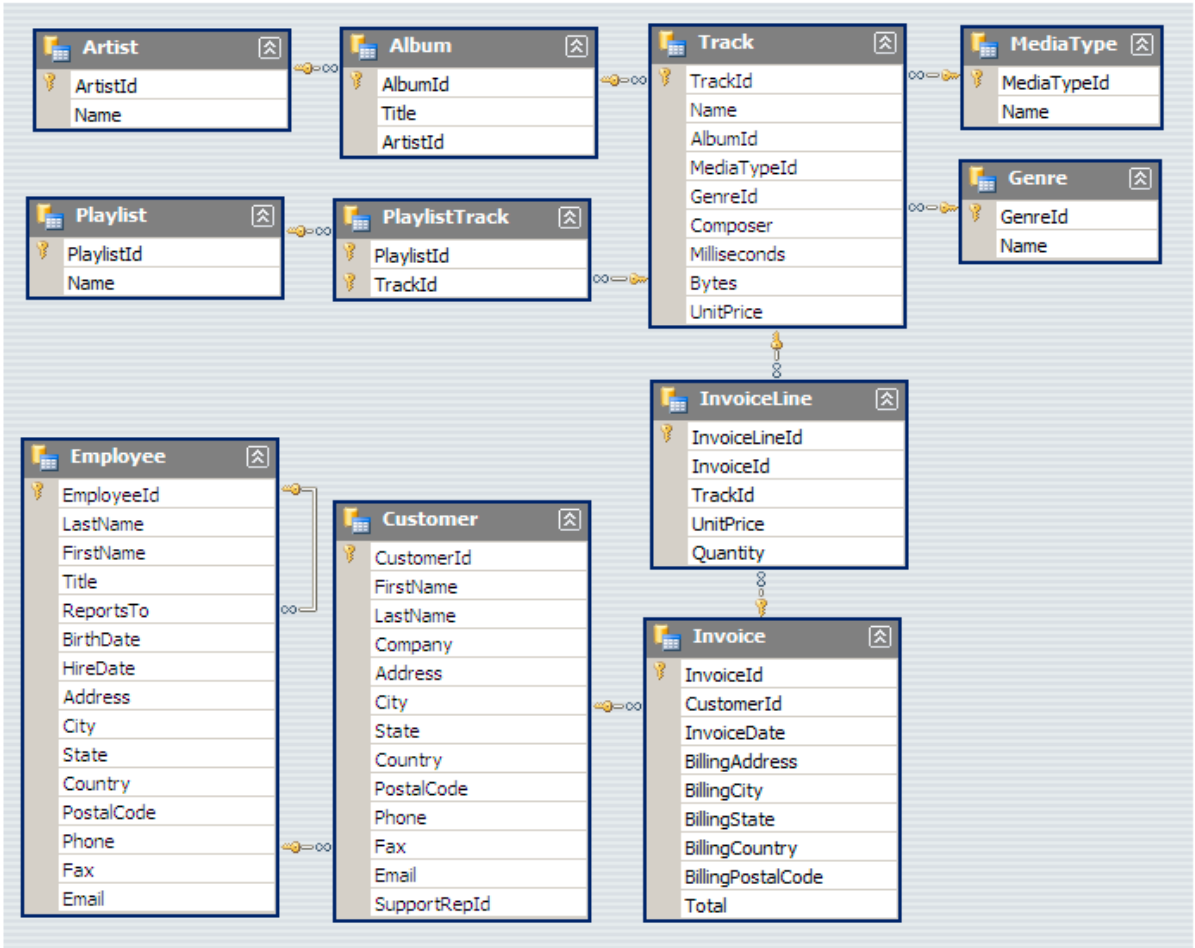
Query | Query History

```
1  Select *
2  From album
3  Limit 5;
```

Data Output | Messages | Notifications

	album_id [PK] character varying (50)	title character varying (120)	artist_id character varying (30)
1	1	For Those About To Rock We Salute You	1
2	2	Balls to the Wall	2
3	3	Restless and Wild	2
4	4	Let There Be Rock	1
5	5	Big Ones	3

Before starting the analysis, we will chart out Schema, which will help us make connections with other data bases and guide us on which tables to join in order to make our analysis.



At the conclusion of our analysis, we will transition to a dashboard that presents the key components essential for addressing the business problem entrusted to us.

1. ***Which is the most profitable region & what is the top 5 most popular genre there?***
   We will first look at the most profitable region. For this we will have to select the invoice table and make a summation of the 'Total' column and group the data by billing country.

```
SELECT billing_country, Round(Sum(total)) AS Total_Amount
FROM invoice
GROUP BY billing_country
ORDER BY Total_Amount DESC;
```

The query without the 'Limit' command gives total for all the 24 countries where the company has sold the albums. We will limit the query to 1 so as to get most top most profitable region.

Query Query History

```
SELECT billing_country, Round(Sum(total)) AS Total_Amount
FROM invoice
GROUP BY billing_country
ORDER BY Total_Amount DESC
Limit 1;
```

Data Output Messages Notifications

	billing_country character varying (30)	total_amount double precision
1	USA	1040

After running the query, we get that USA is the most profitable country, with $ 1040 in total sales. Next, we will look at which cities are most profitable within USA. This will give us a deeper insight into the cities that contributed the most to the total sales in the USA.

```
SELECT billing_country, billing_city, Round(Sum(total)) AS Total_Amount
FROM invoice
where billing_country = 'USA'
GROUP BY billing_city, billing_country
ORDER BY Total_Amount DESC
Limit 5;
```

Data Output Messages Notifications

	billing_country character varying (30)	billing_city character varying (30)	total_amount double precision
1	USA	Mountain View	169
2	USA	Redmond	98
3	USA	Orlando	92
4	USA	Reno	91
5	USA	Fort Worth	86

Mountain View, Redmond, Orlando, Reno & Fort Worth are the top 5 city that generated the most revenue within USA.

Next step is to find out the 5 popular genres in USA. For this we will have to make multiple joins. Looking at the schema, to get the Genre, we need to join customer table with invoice table, invoice table with invoice line table, track table with invoice line table and finally track table with genre table.

```
WITH popular_genre AS
( SELECT customer.country, genre.name As Genre_name, COUNT(invoice_line.quantity) AS Total_purchases
  FROM invoice_line
  JOIN invoice ON invoice.invoice_id = invoice_line.invoice_id
  JOIN customer ON customer.customer_id = invoice.customer_id
  JOIN track ON track.track_id = invoice_line.track_id
  JOIN genre ON genre.genre_id = track.genre_id
  GROUP BY 1, 2
  ORDER BY 1 ASC, 3 DESC
)
SELECT * FROM popular_genre WHERE country = 'USA';
```

Data Output | Messages | Notifications

	country character varying (50)	genre_name character varying (120)	total_purchases bigint
1	USA	Rock	561
2	USA	Alternative & Punk	130
3	USA	Metal	124
4	USA	R&B/Soul	53
5	USA	Blues	36

What we have done in the above query is used CTE method to pull out the 5 most popular genre in USA. We define most popular genre by the one which has sold the greatest number of quantities. Group by 1, 2 means that we have asked the output to be grouped by column 1 which is Country and column 2 which is the Genre name. In the same way we have Sorted the output by total purchases (which is column no. 3) in Descending order to get the information that we want.

Using the above query, we get the 5 most popular genre in USA to be Roack, Alternative & Punk, Metal, R&B and Blues.

2. Which are the top 5 non-performing region?

Non-performing region can be defined as the one which has generated the least amount of revenue. We will use the Invoice table to get the data from.

```
SELECT billing_country, Round(sum(total)) As Total_Amount
FROM Invoice
GROUP BY billing_country
ORDER BY Total_Amount ASC
LIMIT 5
```

Data Output | Messages | Notifications

	billing_country character varying (30)	total_amount double precision
1	Denmark	38
2	Argentina	40
3	Italy	50
4	Belgium	60
5	Netherlands	65

We have used the same code as we did to answer part 1 of the 1st question, but modified it to show countries with the least amount of invoice amount. The output shows Denmark, Argentina, Italy, Belgium & Netherlands as the countries with the lowest revenue generated.

3. Where are the top 10 customers located?

```
SELECT first_name, last_name, ROUND(SUM(total)) AS total_spending, country
FROM customer
JOIN invoice ON customer.customer_id = invoice.customer_id
GROUP BY 1,2,4
ORDER BY total_spending DESC LIMIT 10;
```

Data Output | Messages | Notifications

	first_name character	last_name character	total_spending double precision	country character varying (50)
1	R	Madhav	145	Czech Republic
2	Helena	Holý	129	Czech Republic
3	Hugh	O'Reilly	115	Ireland
4	Manoj	Pareek	112	India
5	Luís	Gonçalves	109	Brazil
6	Fernanda	Ramos	107	Brazil
7	João	Fernandes	103	Portugal
8	Wyatt	Girard	100	France
9	François	Tremblay	100	Canada
10	Jack	Smith	98	USA

Running this query gives us a surprising result. The top spending customers are from other regions than USA (which was the best performing region in terms of sales revenue). Therefore, USA must have a greater number of clients than other regions to generate the highest number of revenues.

We can check this by totalling the number of customers and listing them region wise.

```
SELECT COUNT (DISTINCT(customer.customer_id)) AS Total_Customers, country, ROUND(AVG(Total)) AS Average_Spending
FROM customer
JOIN invoice ON customer.customer_id = invoice.customer_id
GROUP BY 2
ORDER BY 1 DESC, 2 DESC LIMIT 10;
```

Data Output | Messages | Notifications

	total_customers bigint	country character varying (50)	average_spending double precision
1	13	USA	8
2	8	Canada	7
3	5	France	8
4	5	Brazil	7
5	4	Germany	8
6	3	United Kingdom	9
7	2	Portugal	6
8	2	India	9
9	2	Czech Republic	9
10	1	Sweden	8

The above query gives us list of unique customers in all the regions and their average spending. USA has a greater number of customers than other regions.

If we remove the Distinct function from the query above, we get the following result

```
SELECT COUNT (customer.customer_id) AS Total_Customers, country, ROUND(AVG(Total)) AS Average_Spending
FROM customer
JOIN invoice ON customer.customer_id = invoice.customer_id
GROUP BY 2
ORDER BY 1 DESC, 2 DESC LIMIT 10;
```

Data Output | Messages | Notifications

	total_customers bigint	country character varying (50)	average_spending double precision
1	131	USA	8
2	76	Canada	7
3	61	Brazil	7
4	50	France	8
5	41	Germany	8
6	30	Czech Republic	9
7	29	Portugal	6
8	28	United Kingdom	9
9	21	India	9
10	13	Ireland	9

4. ***Which genre has sold the most & which artist has sold the most records?***

We can solve this by CTE method as in the question 1. With this query we get Rock as the most popular genre i.e., the genre which has sold the greatest number of quantities.

Query | Query History

```
WITH popular_genre AS
(   SELECT genre.name AS Genre_Name, COUNT(invoice_line.quantity) AS Total_Purchases
    FROM invoice_line
    JOIN invoice ON invoice.invoice_id = invoice_line.invoice_id
    JOIN customer ON customer.customer_id = invoice.customer_id
    JOIN track ON track.track_id = invoice_line.track_id
    JOIN genre ON genre.genre_id = track.genre_id
    GROUP BY 1
    ORDER BY 2 DESC
)
SELECT * FROM popular_genre LIMIT 1
```

Data Output | Messages | Notifications

	genre_name character varying (120)	total_purchases bigint
1	Rock	2635

We can change the LIMIT to 5 and see the top 5 genres too.

Query | Query History

```
WITH popular_genre AS
(    SELECT genre.name AS Genre_Name, COUNT(invoice_line.quantity) AS Total_Purchases
     FROM invoice_line
     JOIN invoice ON invoice.invoice_id = invoice_line.invoice_id
     JOIN customer ON customer.customer_id = invoice.customer_id
     JOIN track ON track.track_id = invoice_line.track_id
     JOIN genre ON genre.genre_id = track.genre_id
     GROUP BY 1
     ORDER BY 2 DESC
)
SELECT * FROM popular_genre LIMIT 5
```

Data Output | Messages | Notifications

	genre_name character varying (120)	total_purchases bigint
1	Rock	2635
2	Metal	619
3	Alternative & Punk	492
4	Latin	167
5	R&B/Soul	159

To check the Artist with the greatest number of sold records, we can do a simple query and know that 'Queen' artist has sold the maximum number of records and unsurprisingly it is from the 'Rock' genre. We can also see that the top 5 artist are from 'Rock' genre as it is the most popular segment that people prefer.

Query | Query History

```
SELECT artist.name AS artist_name, COUNT(invoice_line.quantity) AS total_sales, genre.name as Genre_Name
FROM invoice_line
JOIN track ON track.track_id = invoice_line.track_id
JOIN album ON album.album_id = track.album_id
JOIN artist ON artist.artist_id = album.artist_id
JOIN genre ON genre.genre_id = track.genre_id
GROUP BY 1, 3
ORDER BY 2 DESC
LIMIT 5
```

Data Output | Messages | Graph Visualiser | Notifications

	artist_name character varying (120)	total_sales bigint	genre_name character varying (120)
1	Queen	192	Rock
2	Jimi Hendrix	187	Rock
3	Nirvana	130	Rock
4	AC/DC	124	Rock
5	The Rolling Stones	117	Rock

5. ***What is the top spending customer spending in each country?***

   We can use the CTE method again to find out the what the top spending customer in each country are spending. We use the ROW method to limit the data to only the top spending in each country. If we do not put in the ROW clause, the query will give multiple results for a specific country. Since the focus is to just get top spending customer from each country, we use the ROW method.

Query | Query History

```sql
WITH top_customer_spendings AS
(SELECT first_name, last_name, ROUND(SUM(total)) AS total_spending_by,country,
ROW_NUMBER() OVER(PARTITION BY customer.country ORDER BY COUNT(total) DESC) AS RowNo
FROM customer
JOIN invoice ON customer.customer_id = invoice.customer_id
GROUP BY 1,2,4
ORDER BY 3 DESC )
SELECT * FROM top_customer_spendings WHERE RowNo <= 1
```

Using the above query, we get the following result

Data Output | Messages | Graph Visualiser | Notifications

	first_name character	last_name character	total_spending_by_top_customer double precision	country character varying (50)	rowno bigint
1	R	Madhav	145	Czech Republic	
2	Hugh	O'Reilly	115	Ireland	
3	Manoj	Pareek	112	India	
4	Fernanda	Ramos	107	Brazil	
5	Phil	Hughes	98	United Kingdom	
6	Enrique	Muñoz	98	Spain	
7	Luis	Rojas	97	Chile	
8	Heather	Leacock	92	USA	
9	Edward	Francis	91	Canada	
10	Madalena	Sampaio	82	Portugal	
11	Leonie	Köhler	82	Germany	
12	Mark	Taylor	81	Australia	
13	Terhi	Hämäläinen	79	Finland	
14	Ladislav	Kovács	78	Hungary	
15	Stanisław	Wójcik	76	Poland	

Total rows: 24 of 24 Query complete 00:00:00.184

Step 5: Share

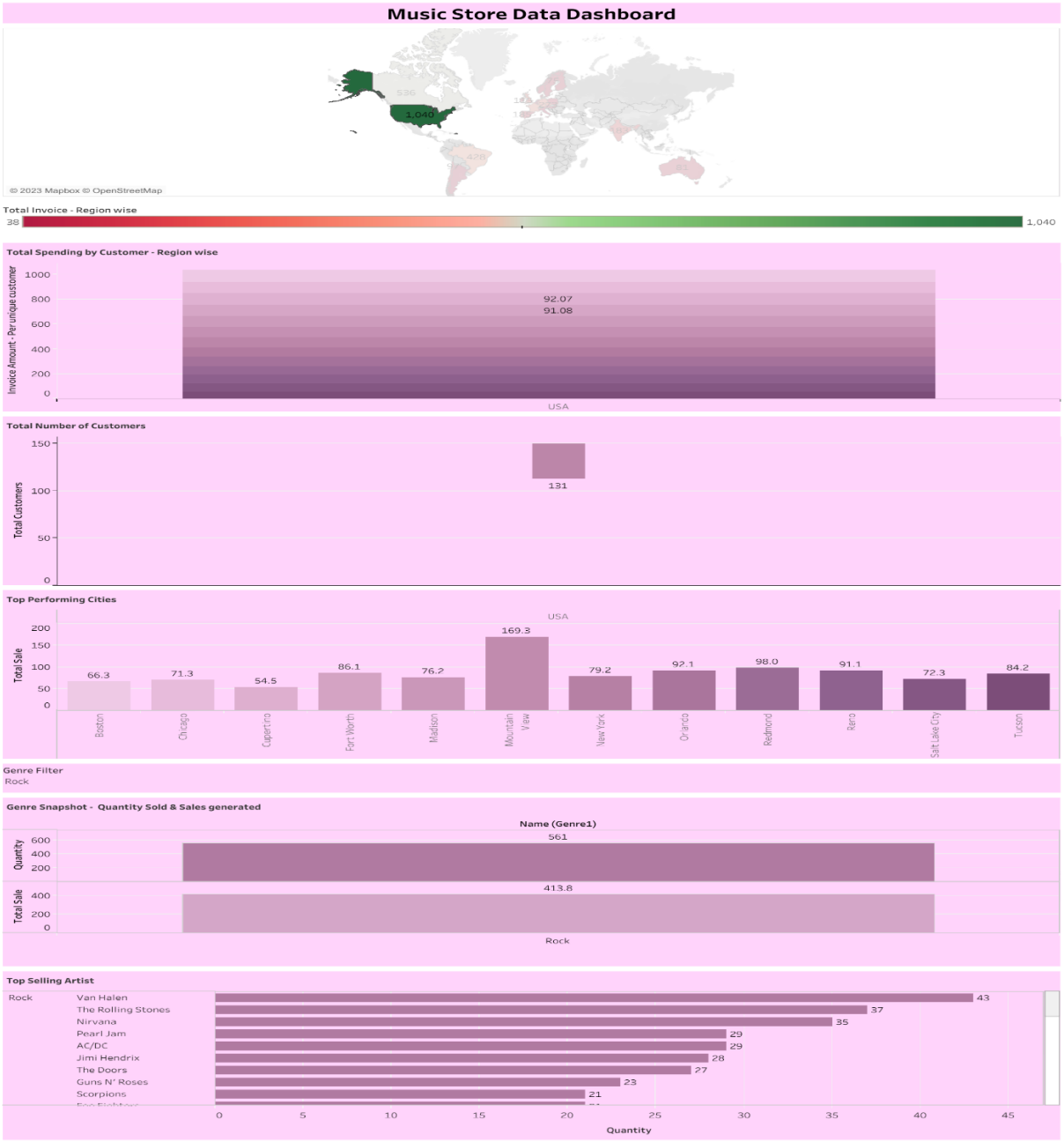
Onto sharing our observations, below you will find a picture of the interactive dashboard that represents the main KPIs and information on the collected Music store data which was realized and demonstrated with Tableau.

The link to the interactive dashboard will be found under the screenshot below.

Here are our findings:

- USA is the most profitable region with Mountain View, generating the maximum sales within USA.
- Denmark, Denmark, Argentia, Italy, Belgium & Netherlands were the least profitable regions, with Denmark only generating 38$ in total sales for the year.
- Suprisingly, the top spending customers are from Czech Republic, Ireland, India, Brazil & Portugal.
- USA became the top sales generator on back of high number of customers as compared to other regions. The number of unique customers in USA was 13, as compared to 2 each in Czech Republic, Portugal & India.

- Overall, the number of customer count in USA was 131, followed by 76 in Canada, 61 in Brazil, 50 in France. The rest of the regions were below the 50 counts.
- We then look at the top spending customers in each region, and while the top customer in Czech Republic, Ireland, India & Brazil spent 145$, 115$, 112$ & 107$ respectively, the top customers in USA and Canada spent 92$ and 91$ respectively.
- We then looked at the most popular genre and the artist who sold the most albums in that genre. The most popular genre was Roack followed by Metal and the most bought artist in the Rock genre was 'Queens'.
- Our tableau dashboard not only gives the insight we sought, but is dynamic to show details for other queries that may arise, for e.g., who is the most sold artist from Metal genre. The dashboard gives selection to filter the query and get the answer require.



https://public.tableau.com/app/profile/shrinidhi.krishnan8091/viz/MusicStoreDashboard_16972368656870/Dashboard1

Step 6: Act

With everything that was covered, here are our conclusions and future recommendations for the success of the Music Store

- The store can consider shutting the non-performing regions as it has a very low customer base. Instead of spreading itself thin, it can concentrate on pushing sales in USA and Canada which already has a good rapport with the customers as it gets lot of repeat sales from a single customer.
- The store should also look at developing Brazil Market as it has got customers who are ready to spend in the store as well as give repeat business.
- The store, with the help of our Dynamic dashboard, can see the popular genre and artist in the various regions and can look to run promotions or offers, like inviting the artist who sold the most albums or run a promotion to win tickets for concert etc to attract more customers.
- Pulling out from non-performing markets may give the store resource to do more advertisement in the popular region. This will in long run, build customer base and thereby push up profits.
- Making a focused effort on building a base will be more beneficial to the store rather than just expanding it presence.

The recommendations come with the limitation of analysis data only for 1 year. A year-on-year data would have given a better insight. Given the data set, the store is recommended to take a long hard look at regions with low customer base and sale and focusing on developing the regions where it has got a good customer base.

Thank You!!