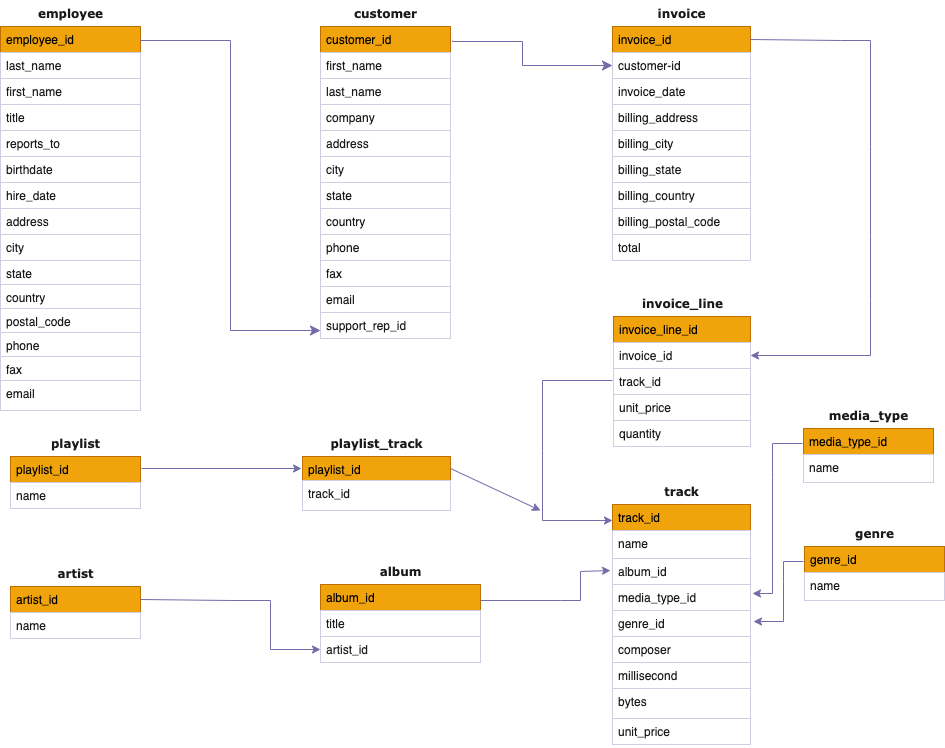


|  |
| --- |
| **PROJECT ON MySQL**  **THE ANALYSIS OF CHINOOK DATABASE** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Developed by**  **Members:**   |  |  |  | | --- | --- | --- | | **No.** | **Student Name** | **Student ID** | | 1 | Nguyen Truong Vu | Student1430272 | | 2 | Nguyen Quoc Bao | Student1439125 | | 3 |  |  | | 4 |  |  |  * **Class No.: DT2210L** * **Start Date:** * **End Date:** * **Name of the Coordinator: Mr. Ho Nhut Minh** * **Date of Submission:** |

|  |
| --- |
| **I. Introduction and Schema Diagram** |

1. **Entry Relationship Diagram of Chinook Database**



*Figure 1. ER Diagram of Chinook Database*

1. **Tables of Chinook Database**

The chinook database is designed for a digital media store, including over 15000 rows of data and 11 tables: Artist, Album, Genre, Media\_type, Playlist, Playlist\_track, Track, Invoice, Invoice\_line, Customer, and Employee.

|  |
| --- |
| **II. EXPLORATORY DATA ANALYSIS (EDA)** |

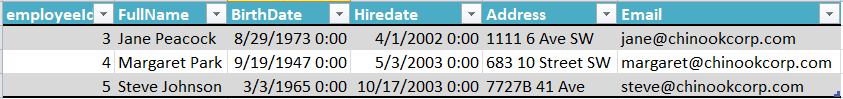
With the aim of mastering the information available in the Chinook database, authors mainly extract the data based on information of customers, employees, tracks, and artists to as have the most consolidated view as possible on the database. The below tables show the results after exploring data, and also used as references for further analysis in this project.

**2.1. Information of tracks sold**

****

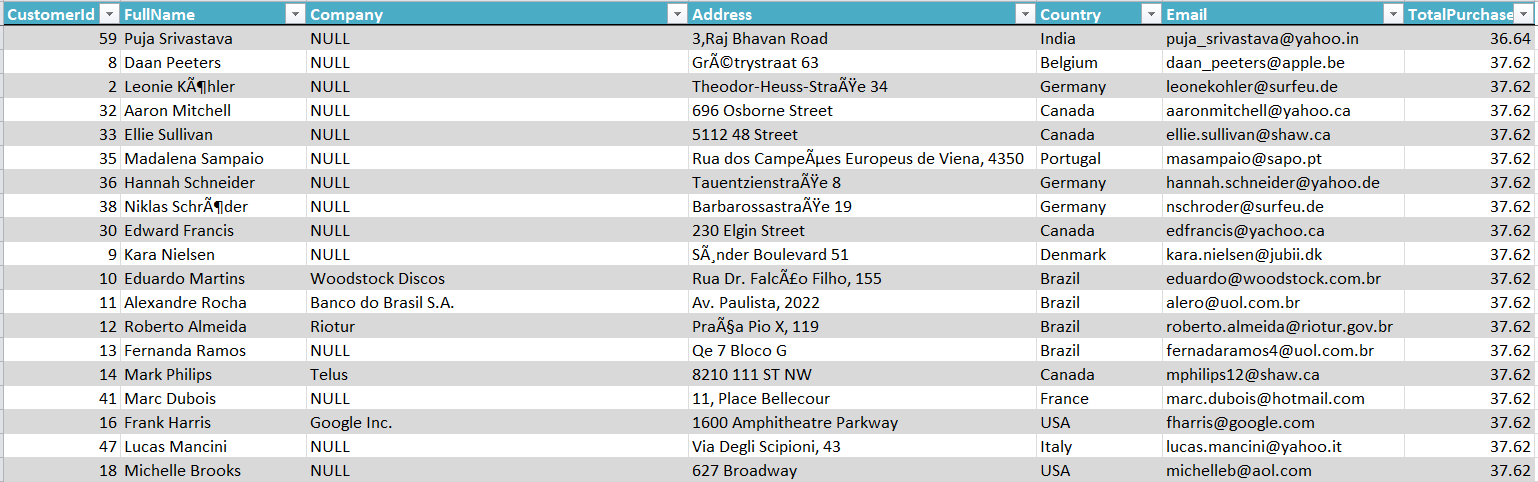
The above table shows tracks that are purchased of the store. Additional information related to genre, album, artist, and the number of purchases are also included. This piece of information is necessary for further analysis throughout the project.

**2.2. Information of employees who directly generate the sales for the store**



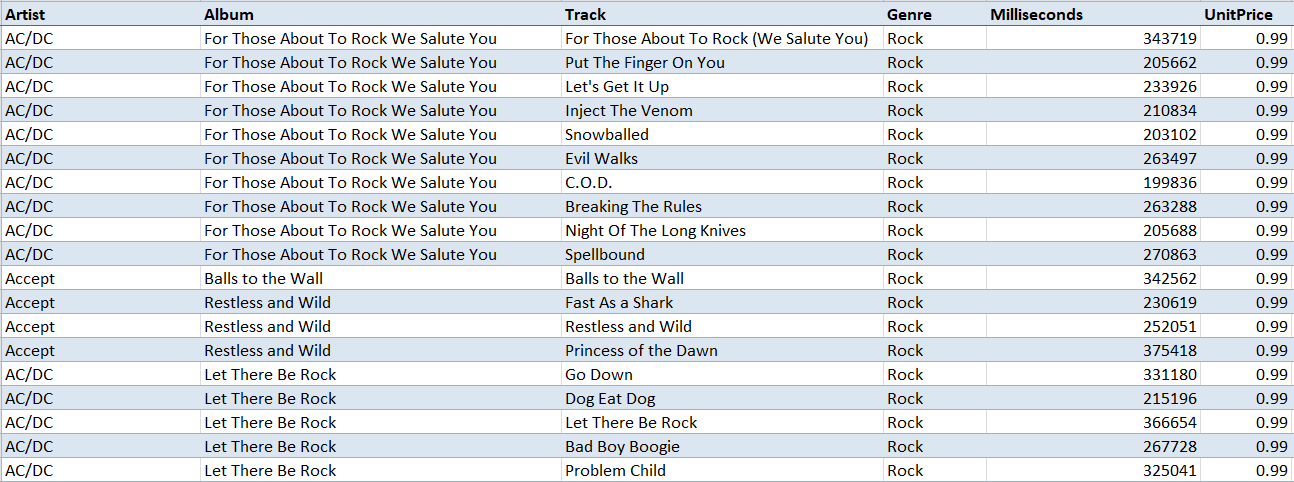
The mentioned results list three employees who directly take responsibility for selling product at the store. Despite 8 employees listed in the database, only three of them generate the direct sales for the store. This information will be used for analyzing the employee performance further in the section 4.

**2.3. Information of customers making purchases at the store**



The given table shows the information of customers making purchases, associated with some relevant basic information. Moreover, the total purchase column in the table illustrates the total amount of money customers have contributes when buying product for the store.

**2.4. Information of all tracks available at the store**



The given table basically list all tracks purchased by the store for selling. This table will be used to identify the question in the section 6.

|  |
| --- |
| **III. Selecting Albums to Purchase** |

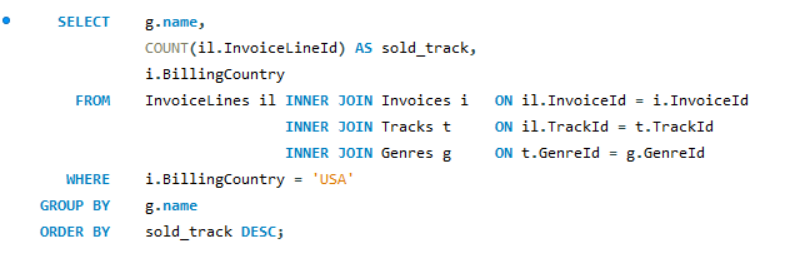
**3.1. The most popular Genres were sold the most in the USA**

The identification of the most popular genres sold in the USA is the main topic for this section. Authors aim to find out which tracks have the highest sales in the invoiceline table and then categorize those tracks by genres.

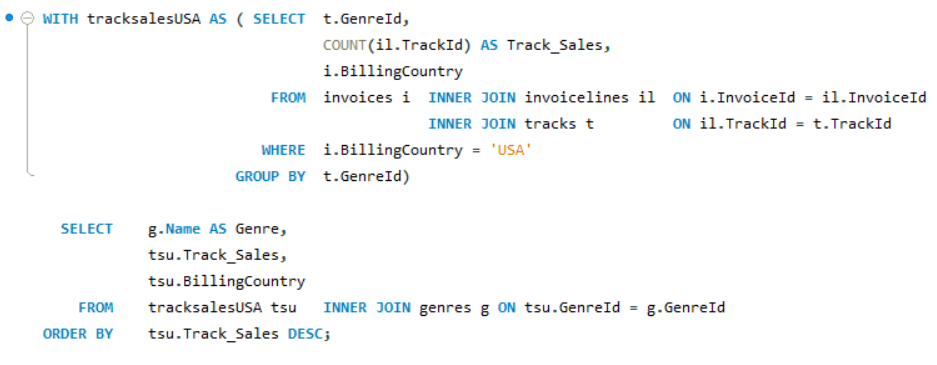
* **Syntax**

In terms of syntax for this part, authors have developed two ways of querying shown below. The first figure shows how to retrieve data by using JOIN statement and the second one is about using WITH clause to create a temporary table for further retrieving data.

The similarity of two mentioned way is that authors use INNER JOIN to collect data from three tables including inovice, track, and genre. Next, the COUNT function would be used on “il.invoicelineId” column and referred as “sold\_tracks”. Authors, then, use WHERE clause to get records of USA only. The result returns are grouped by genres.



There is also another way to retrieve data to find out the answer without using WITH clause. With the WITH clause, authors have created a new temp table for counting tracks sold in the USA and then grouped them into genre groups. At the outer query, authors use JOIN statement to connect this temp table to genre one in order to get the title of each genre.



* **Result**



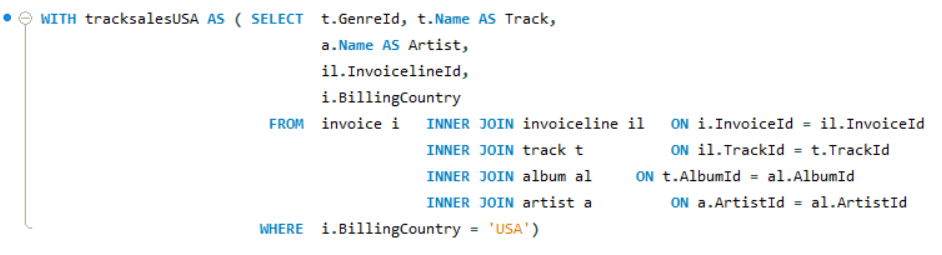
The result table shows that Rock, Latin, and Metal are three genres that have the highest sales, with 157, 91, and 64 tracks respectively.

**3.2. The recommendation for three artists whose albums should be purchased**

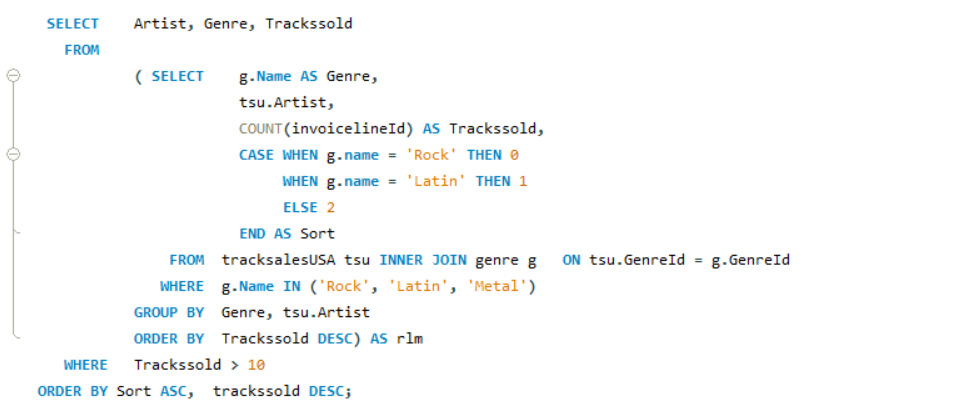
As an additional part of Question 3, after figuring out the most popular genres in the USA, authors aim to suggest the most favorable artists for the store to make purchase decisions of buying their albums. In other words, the findings of sold track of each artist is the final goal in this part.

* **Syntax**

To do so, authors intent to create a temporary table, namely tracksaleUSA in which information about genreId, track name, aritst name, invoicelineId, and billingcountry would be involved.

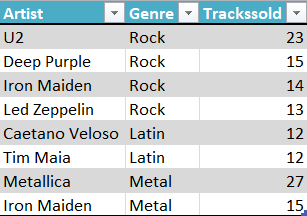


In this temp table, authors used JOIN function to gether data from five different tables together, including invoice, invoiceline, track, album, and artist. All Data in this table will be sorted by The USA.



Then, authors have used subquery to create another temp table, and the most outer query will be used to extract data from this one. In the subquery, the mentioned tracksalesUSA table will be joined with genre by genreId column. The WHERE clause has been used in this step in order to put the focus of the table on three popular genres including Rock, Latin, and Metal. The table, after executed, illustrates genre name, artist name, and Trackssold generated by counting records on invoicelineId column. In regard to sorting artist by genre, authors has created a another column, namely Sort.

* **Result**

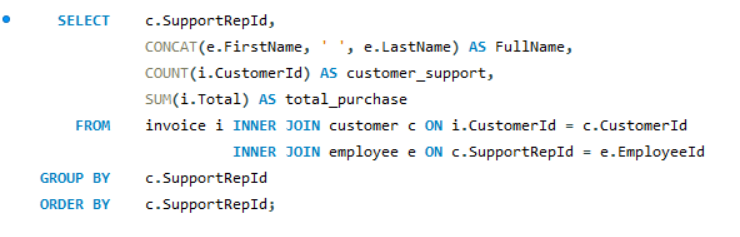
****

In the above figure, there are 4 artists of Rock, 3 artists of Latin, and 2 artists of Mental that are recommended for the store to purchase their albums.

|  |
| --- |
| **IV. Analyzing employee sales performance** |

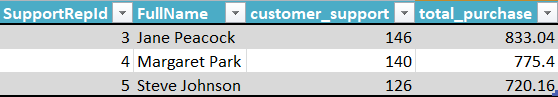
In this section, authors have to find out sales performance of each employees based on the total purchase of customers who they handle.

* **Syntax**



The syntax for this section is mainly using JOIN functions to gather data from invoice, customer, and employee table. Next, authors have performed SUM calculation to figure out the total purchase of customers which reflects the sales performance of employees and COUNT calculation to see whether or not the number of sales of each employee has impacts on their performance. In order to make the result table readable, authors has used CONCAT function to form a FullName column in which First Name and Last Name of employees put together.

* **Result**



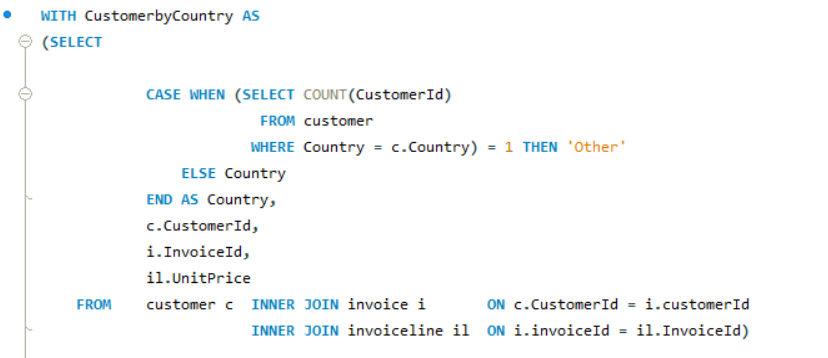
The result obtained shows that Jane peacock is an employee that have the highest sales performance with $833.04. Margaret Park follows jane with $775.4 and the last one is Steve with $720.16. Inspite of 8 employees given in the employee table, only three of them are assigned to support customers in sales. In comparison with customer\_support, its data seems to not reflect their impact on employee performance as the more customers employees can handle, the higher total sales employee can have. It is likely that hidden factors may exist in other data.

|  |
| --- |
| **V. Analyzing sales by Country** |

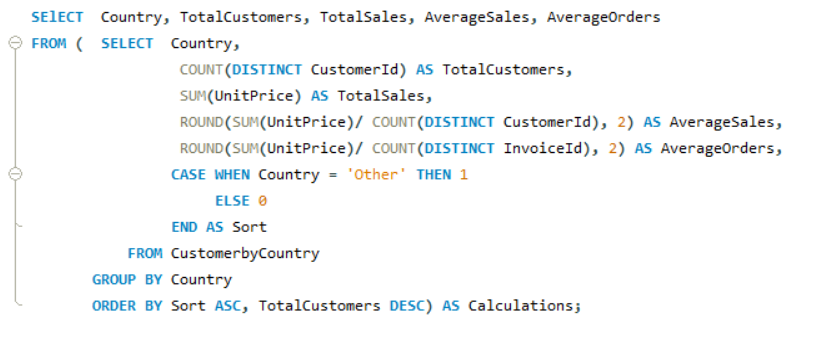
In this part, authors are going to calculate total number of customers, total value of sales, average value of sales per customer, and average order value in each country. The technique used for this part is joinning tables including Customers, Inovices, and Invoicelines.

* **Syntax**

The author will create a temporary table by using WITH clause and name the table as CustomerbyCountry. In this table, the author will group the number of customers into subgroup by countries and a new subgroup, namely “Other”, includes many countries that have only 1 customers.

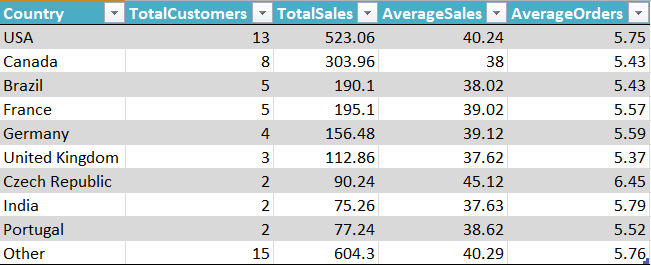


Next, authors will retrieve the requested data to answer the question of part 5 based on Data gathered in the above CustomerbyCountry table.



In this phase, author will use subquery to address the problem. First, three aggregated functions calculations will be performed, including Count, Sum, and Average and then, the obtained results are grouped by Country. Second, the author, then create a new column, namely “Sort” by using CASE Statement. This condition will assign “1” for Country having only 1 customer and “0” for the remainings, allowing authors arrange table in ascending order.

* **Result**



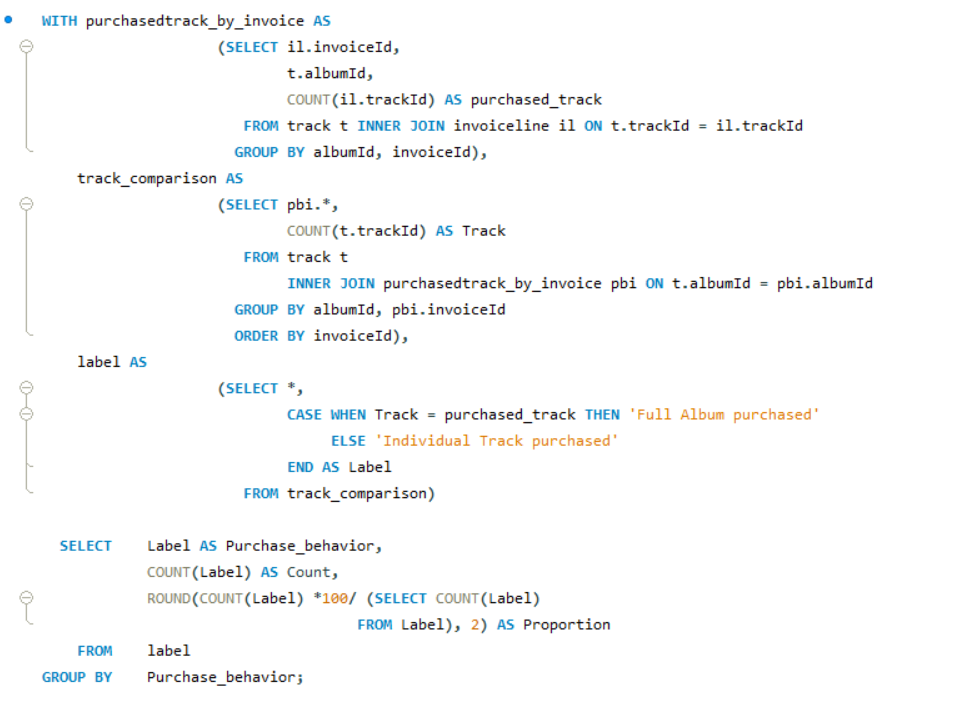
*Table of Sales Analysis by Country*

As shown in the table, The USA was the only country that had customer and Sale figures which is far exceeding compared to the remaining countries, with 13 customers and 523.06 dollars respectively. Follwing the USA were Canada ($303.96), Brazil ($190.1), and France ($195.1). The numbers of customers in those three countries were low, ranging from 8 to 2 customers. Finally, it is also known that 15 countries have only 1 customer

|  |
| --- |
| **VI. Albums and Individual tracks** |

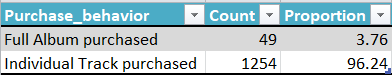
In this part, authors focus on answering the question whether customers prefer to buy tracks in full albums or just picking individual tracks. The syntax for this part is around creating temp tables, and using JOIN function.

* **Syntax**

****

Authors has created three tables for further analysis, including purchased tracks by invoice, available tracks by available album, and clssification for marking which invoice is purchased in full album. After creating temp tables, authors have performed the outer queries, retrieving data by counting and calculating its proportions.

* **Result**



It is noted that nearly 4% of albums are purchased in full version. Meanwhile, the figure for Individual tracks is completely far exceeding. From there, it can be concluded that not having full album available seems to not cause that much dissatsifaction of customers and sales of the store is not affected.

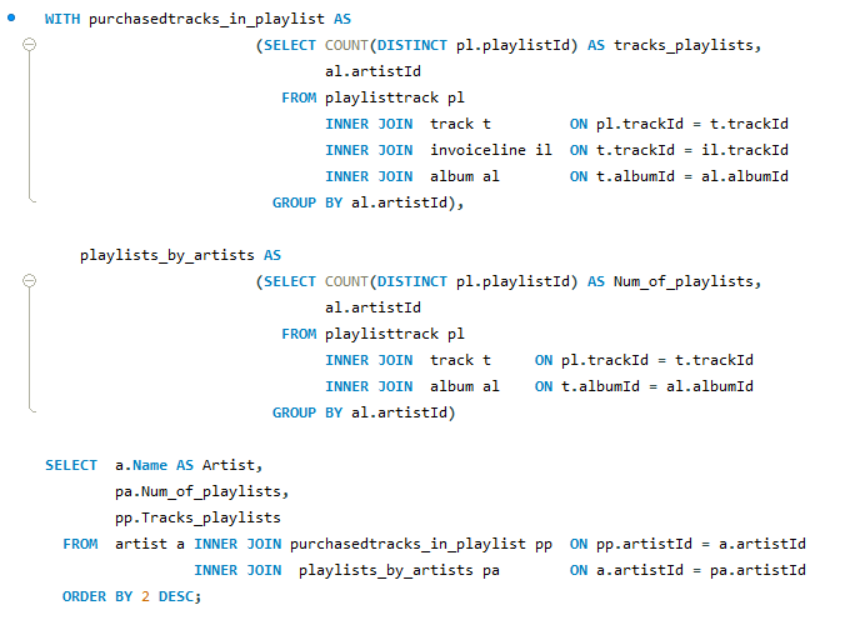
|  |
| --- |
| **VII. ADDITIONAL EXPLORATIONS** |

**7.1. The finding of artists whose sold tracks are used the most in playlists**

In this part, the aim of authors is figuring out artists who have the highest number of sold tracks in playlists.

* **Syntax**

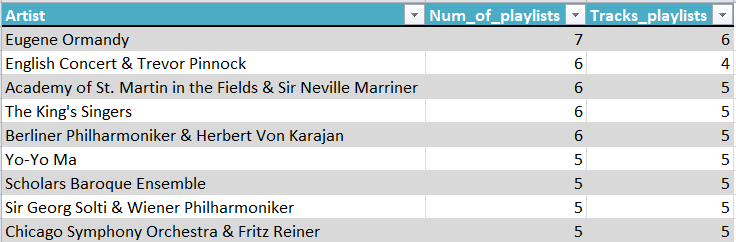
The technique used is creating a temp table, and joinning tables together, and then the authors will perform some of other COUNT calculations to gain the final result

****

The technique used is creating 2 temp tables, namely purchasedtracks\_by\_playlist, and playlists\_by\_artists. In those tables, the authors have joined invoiceline, track, playlisttrack, and album table together by using INNER JOIN. Those tables will include information related to the number of playlists containing sold tracks and non-sold tracks, which are then organized by artistId.

Next, authors will join above temp tables with artist table in order to list the name of artists and the time their tracks presenting in playlists.

* **Result**

****

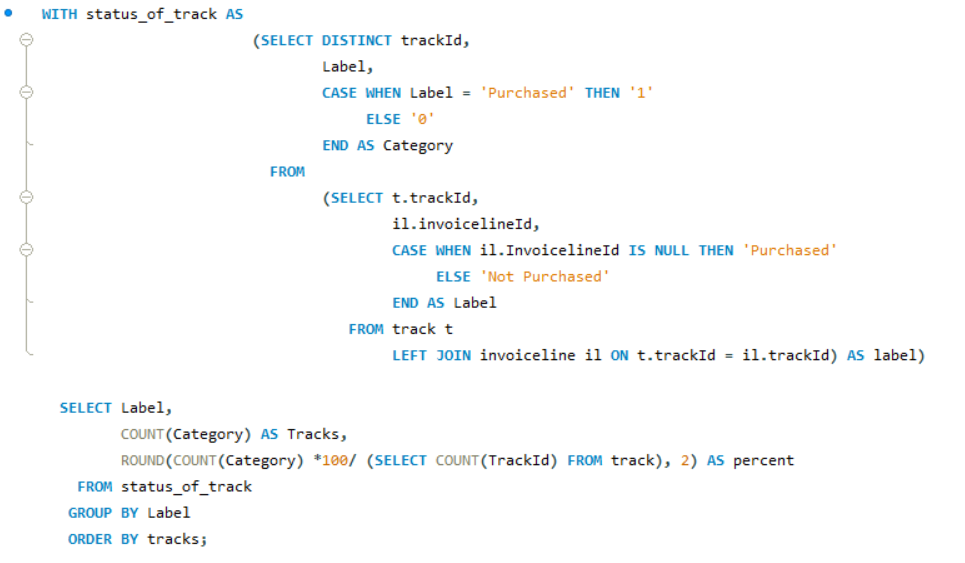
The given table provides the information related to artists whose tracks are available in most playlists. When it comes to sales, the ranks of those artist shows little change. Speaking of both columns, Eugene Ormandy is one of the artist’s name featured the most.

**7.2. The finding of purchased and non-purchased tracks**

The purpose of this part is finding out the number of purchased and non-purchased tracks for the store.

* **Syntax**

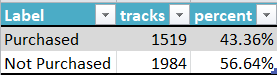
WITH function and COUNT calculations are performed to gain the final result. The systax is also presented by using subqueries.

****

First, authors will create a status\_of\_track table to present the data ralted to unique trackId associated with “purchased” and “not purchased” status. This step is conducted to sort which tracks are purchased.

Second, authors will retrieve the data from above table to count number of tracks asre purchased and that of non-purchased ones. Finally, result is also presented under the form of proportion.

* **Result**

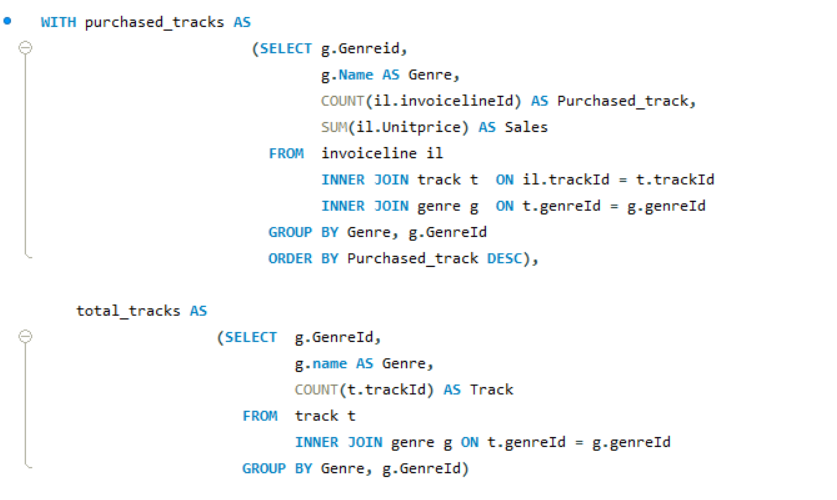
****

The result shows that nearly two thousands tracks (56.64%) are not purchased by customers. In this sense, it is true to say that the store still have not make an optimal use of their expenditure on choosing which tracks should be purchased. An adjustment on expenditure use is highly recommended.

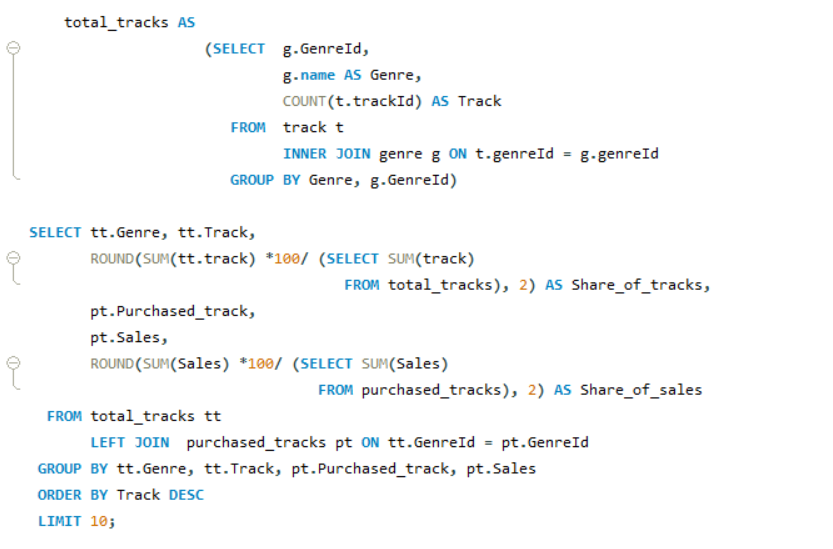
**7.3. The finding of relationships between track ranges and their sales popularity**

Before going deeper to analyze the data for this part, authors are going to verified track range is simply the share of each track in the total. In another word, the key technique to anser this question is finding out the percentage of tracks and sales for comparison.

* **Syntax**

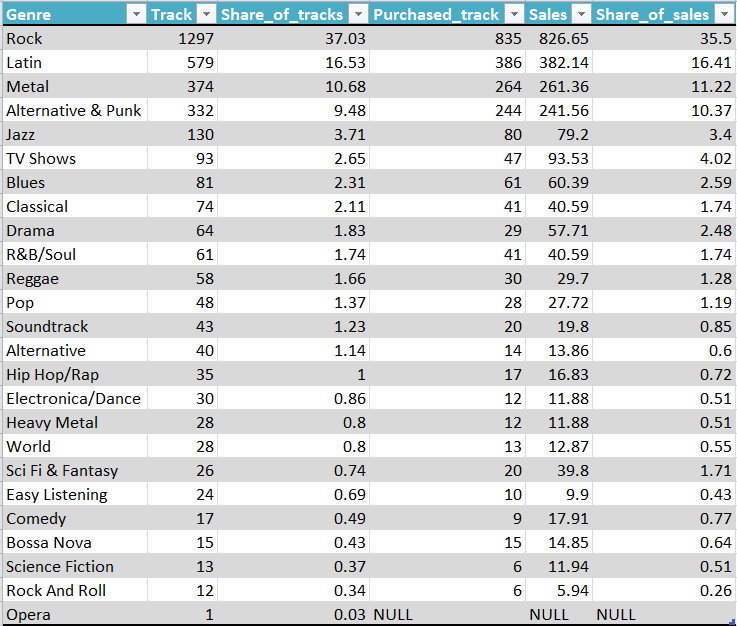
****

To do so, there are two temp tables created. The first one is about the number of purchased tracks grouped by Genres, while the other is of showing the total available tracks in the stores.

****

In the outer queries, authors are going to perform aggregated function to return the data under the form of percentage, called the share of tracks, and share of sales. Along, LEFT JOIN and subquery techniques are used to serve the hold code.

* **Result**

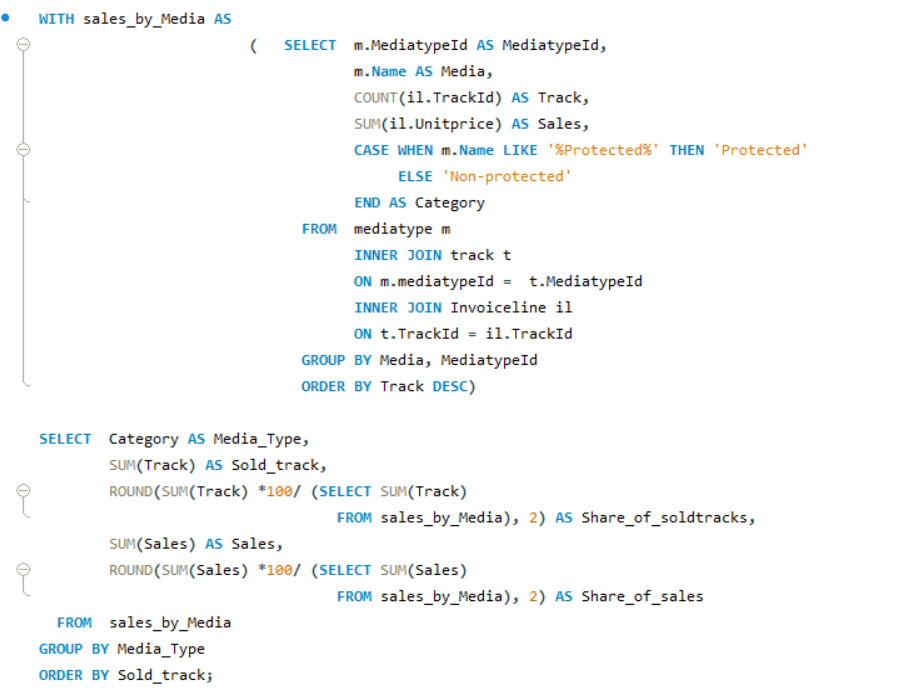
****

For the 10 first records included in the given table, it can be claimed that Rock, Latin, Mental are three genres having the most share of tracks, and share of sales over the total. It is also noticeable that almost figures for share of tracks are always lower than those of share of sales, but they share a little of sales. Only Opera genre is recorded to not witness any purchase from customers.

**7.4. The finding of relationships between Media types and track popolarity**

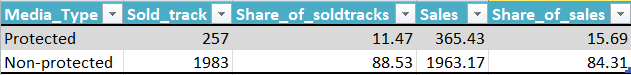
The aim of this part is finding out whether or not, the tracks of protected media types outnumbered that of non-protected ones. In general, protected media types are prefered as platforms that consumers have no capability of sharing any tracks with other people.

* **Syntax**

****

Following the same pattern of coding with previous part, a temporary table is created to calculate the count of trackId, and sum of unitprice. After that, the result obtained will be classified as Protected and non-protected media type. Authors continuou to present data under the form of percentage so see wich types of media preferred the most by consumers.

* **Result**

****

The result illustrate that the relationship between the sales of tracks and its media types they come from. In another word, customers seem to prefer using tracks that are non-protected (1983 tracks, or 88.53%).

|  |
| --- |
| **CONCLUSION** |

After going through analysises of Sale data of this music storem, there are some reccomendations for the store to consider.

When it comes to the sales in the USA which country has the highest sales, stores should focus on selling tracks of Rock, Latin, and Mental genre as those are the most popular genres. With each genre, there are some artists whose tracks are also preferred the most by the American. Thus, it is reasonable to store those tracks from those artists.

It is still necessary to obtain more data from employees and some context of sale in order to have a deeper analysis of employee performance.

There are some reveal about the preference of customers for buying tracks. First, they are likely to prefer buying tracks by picking individual tracks rather than buying them in a whole album. From there, it is still recommended the store to stock individual tracks which have the highest sales, but not the whole album. Second, the percentage of non-purchased tracks accounts for over 50%, which means that those tracks do not have any contribution for the sales of the stores. Thus, cutting down purchases for those tracks seems to be reasonable in order to make an optimal use of the store's expenditure. Finally, almost all customers prefer buying tracks of non-protected media rather than the protected one and it has potential impacts on the sales of the store.