

**PROJECT REPORT**

CSE 457 - ELECTIVE III : Data Analytics - Use of Select Tools and Techniques

Name : Gokul P

Unique ID : E0119056

Year : 2022

Quarter : IV

Department : CSE (AI & ML)

Faculty Name : Prof. Shiyamala

# Academic year : 2022-2023

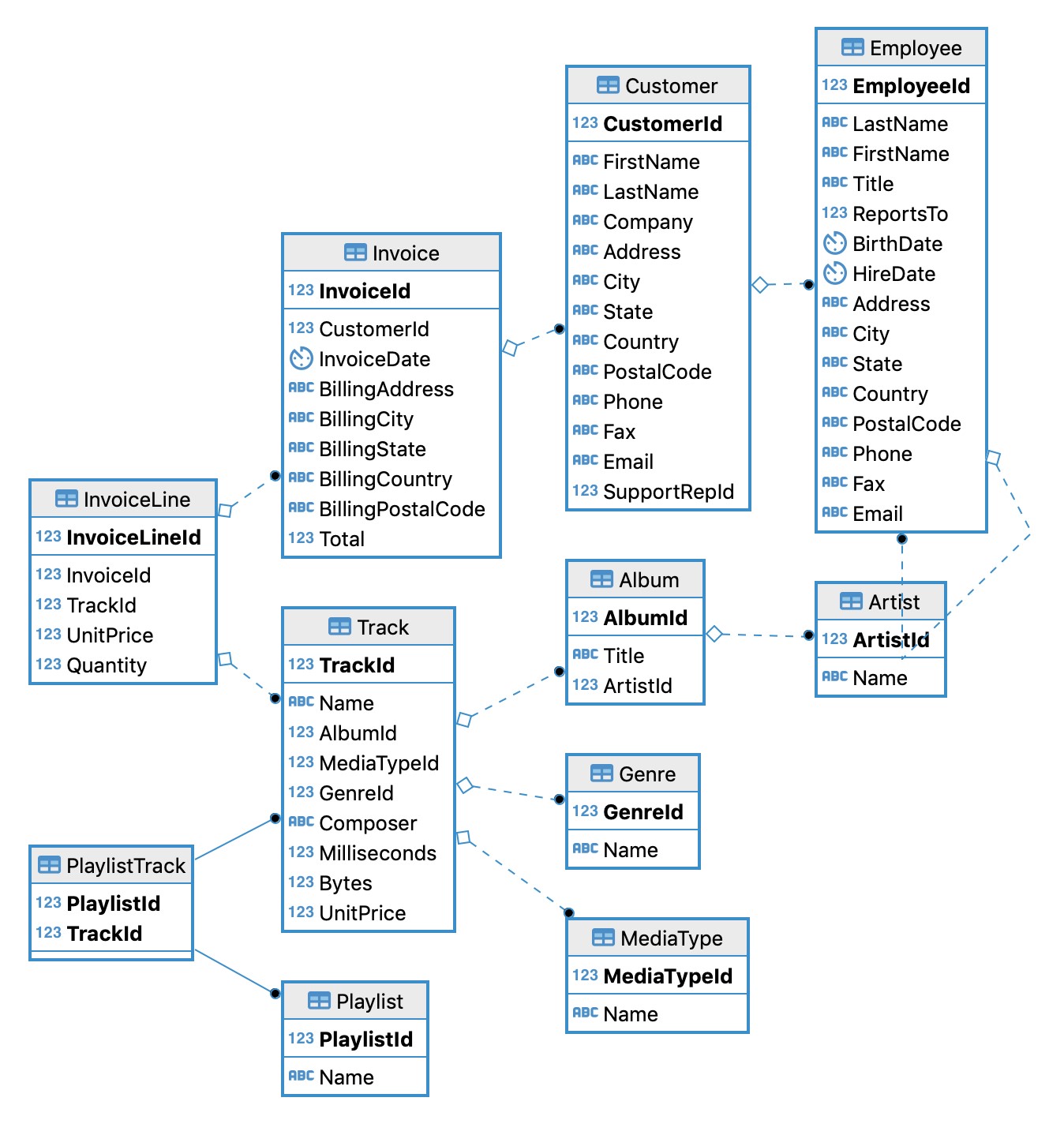
**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **CHAPTER** | **PAGE NO.** |
| **1** | **Abstract** | 3 |
| **2** | **Introduction** | 4 |
| **3** | **Literature Review** | 5 |
| **4** | **Results and Discussions** | 6 |
| **5** | **Conclusion** | 28 |

## ABSTRACT

**Chinook Database:**

The Chinook data model represents a digital music media store, including tables for artists, albums, media tracks, invoices and customers. This is a medium sized database consisting of 11 tables of about 15,000 rows approx in each.



**CARS INDIA DATASET:**

The dataset consists of 156 rows and 25 columns pertaining information on each car’s model type, makers, no.of cylinders, height, width etc.

## INTRODUCTION

**POSTGRES:**

PostgreSQL is an advanced, enterprise-class, and open-source relational database system. PostgreSQL supports both SQL (relational) and JSON (non-relational) querying. PostgreSQL is a highly stable database backed by more than 20 years of development by the open-source community. PostgreSQL is used as a primary database for many web applications as well as mobile and analytics applications.

PostgreSQL has many advanced features that other enterprise-class database management systems offer, such as: User-defined types, Table inheritance, Sophisticated locking mechanism, Foreign key referential integrity, Views, rules, subquery, Nested transactions (savepoints), Multi-version concurrency control (MVCC), Asynchronous replication.

PostgreSQL is designed to be extensible. PostgreSQL allows you to define your own data types, index types, functional languages, etc. Users can develop a custom plugin to enhance it to meet your requirements e.g., adding a new optimizer.

**MONGODB:**

MongoDB is an open-source document database and leading NoSQL database. MongoDB is written in C++. MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose. A document is a set of key-value pairs.

Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data.

Mongodb is widely used as it supports Document Oriented Storage − Data is stored in the form of JSON style documents, Index on any attribute, Replication and high availability, Auto-Sharding, Rich queries, fast in-place updates, professional support by MongoDB.

**APACHE SPARK:**

Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters. It provides essential features such as Batch/streaming data which Unifies the processing of your data in batches and real-time streaming, using our preferred

language: Python, SQL, Scala, Java or R. SQL analytics with

Execute fast, distributed ANSI SQL queries for dashboarding and ad-hoc reporting. Runs faster than most data warehouses.

Data science at scale performs Exploratory Data Analysis (EDA) on petabyte-scale data without having to resort to downsampling.

**SPLUNK:**

Splunk is a software used to search and analyze machine data. This machine data can come from web applications, sensors, devices or any data created by the user. It serves the needs of IT infrastructure by analyzing the logs generated in various processes but it can also analyze any structured or semi-structured data with proper data modeling. It has built-in features to recognize the data types, field separators and optimize the search processes. It also provides data visualization on the search results.

Splunk provides the essential features such as Data Ingestion which can ingest a variety of data formats like JSON, XML and unstructured machine data like web and application logs. The unstructured data can be modeled into a data structure as needed by the user. Secondly, Data Indexing which performs ingested data is indexed by Splunk for faster searching and querying on different conditions. Data Searching in Splunk involves using the indexed data for the purpose of creating metrics, predicting future trends and identifying patterns in the data.

Splunk alerts can be used to trigger emails or RSS feeds when some specific criteria are found in the data being analyzed. Splunk Dashboards can show the search results in the form of charts, reports and pivots, etc. The indexed data can be modeled into one or more data sets that are based on specialized domain knowledge.

**DATAMEER:**

An all-in-one solution for exploring, preparing, visualizing and cataloging Snowflake insights. It enables data engineers and analysts to transform data directly in Snowflake via a simple SQL code or no-code interface.

Datameer offers an innovative data modeling and transformation toolset that brings non-coders into the analytic engineering process and fosters collaboration between citizen data users and engineers to accelerate the creation and delivery of datasets for analytics, machine learning, and reporting.

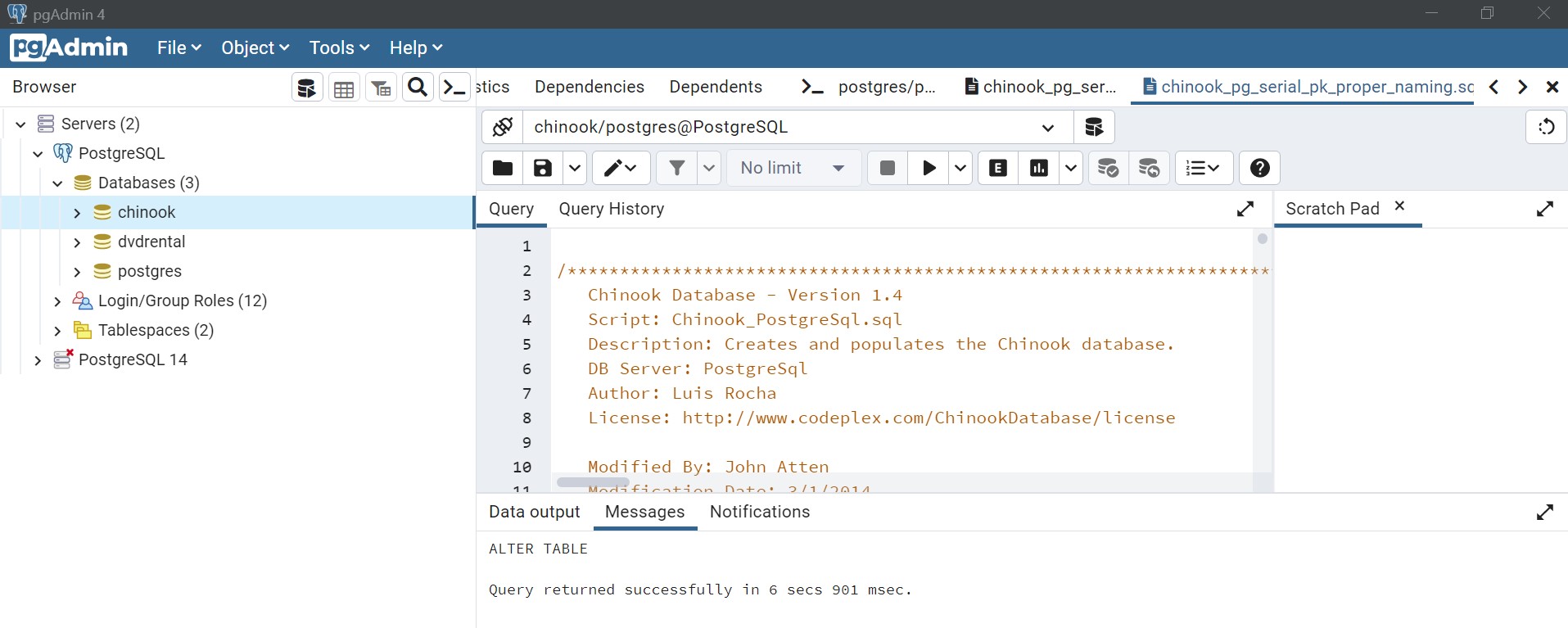
With Datameer, business teams with no prior SQL coding experience can self-serve and create their own datasets inside of Snowflake via a no-code visual interface. Data engineers can review, contribute or optimize these datasets if needed using a more traditional SQL editor. Datameer gives data engineers the control they require and analysts the ability to support a collaborative data modeling and transformation process.

## RESULTS AND DISCUSSIONS

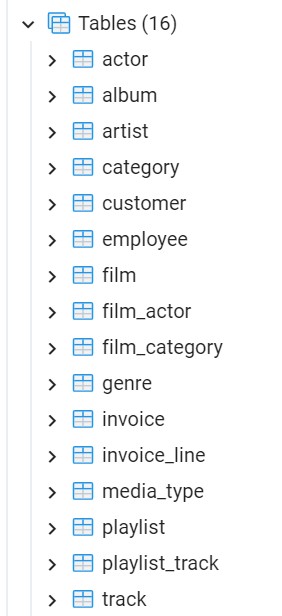
**LOADING THE CHINOOK DATASET:**

On creating an empty database named chinook in postgres admin, we can import the db directly from sql query tab. The chinook database has a .sql script that was imported into pgAdmin4 postgres and this sql code was run to create all the tables with values. Using this link, [https://raw.githubusercontent.com/xivSolutions/ChinookDb\_Pg\_Modified/master/chinook\_pg\_se rial\_pk\_proper\_naming.sql](https://raw.githubusercontent.com/xivSolutions/ChinookDb_Pg_Modified/master/chinook_pg_serial_pk_proper_naming.sql)

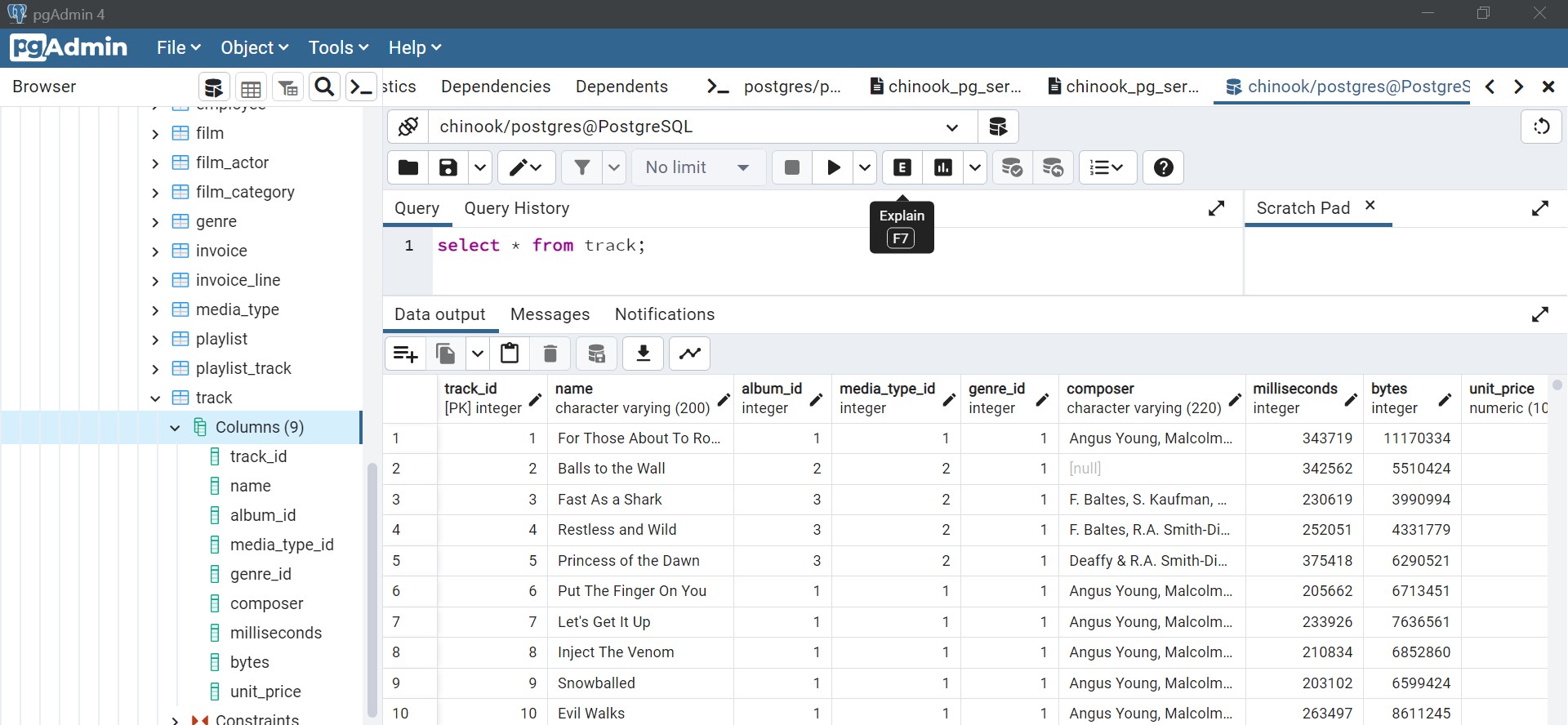
SQL script imported into query tool



On running this sql script, the following table were created with populated values.

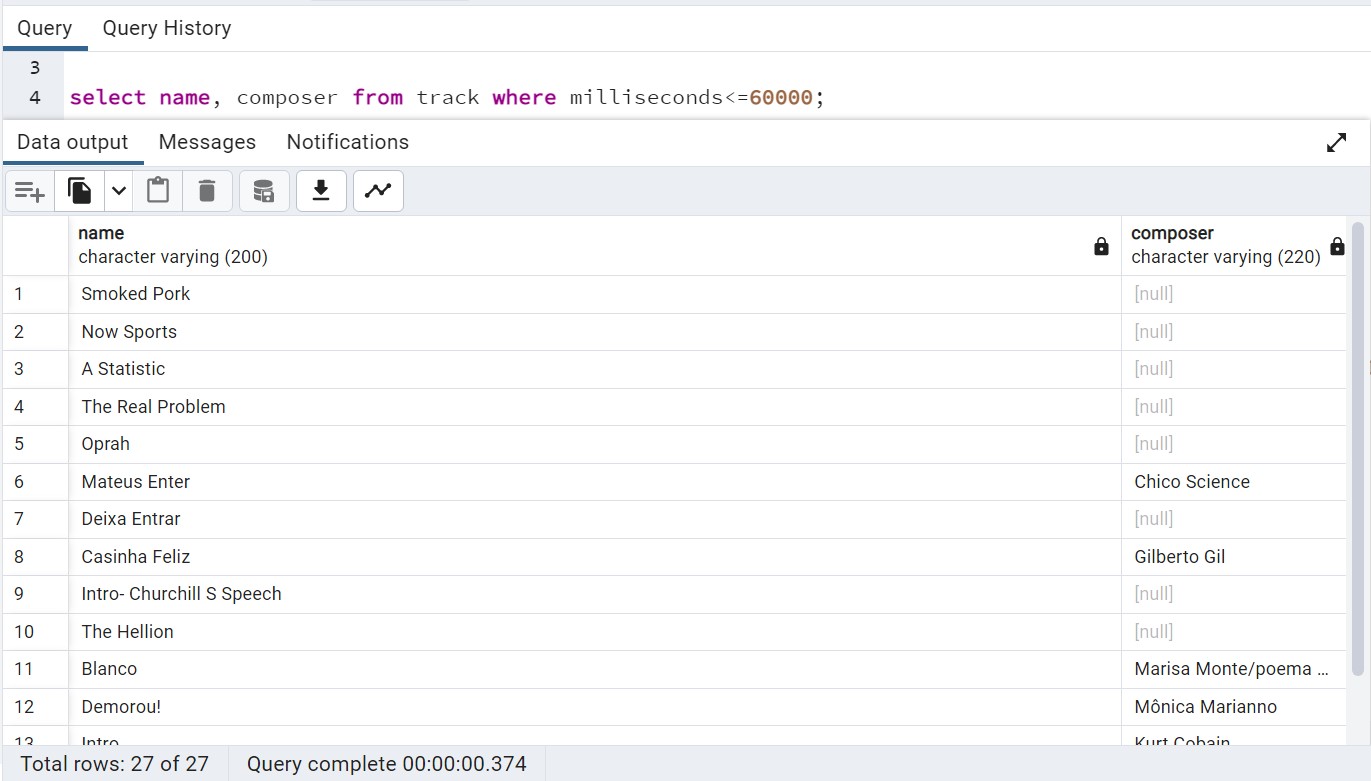


Displaying the data in “track” table using the select statement.



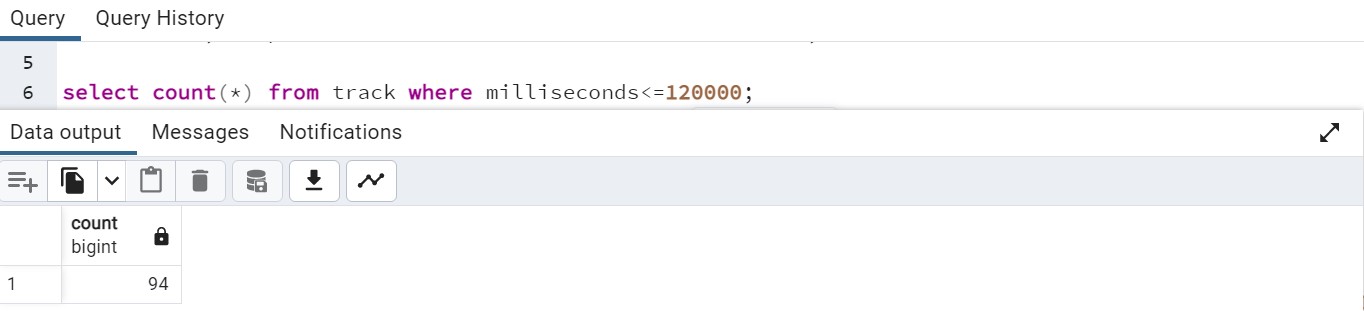
**Queries implemented:**

1. List of all songs whose duration is less than 1 min or 60000ms



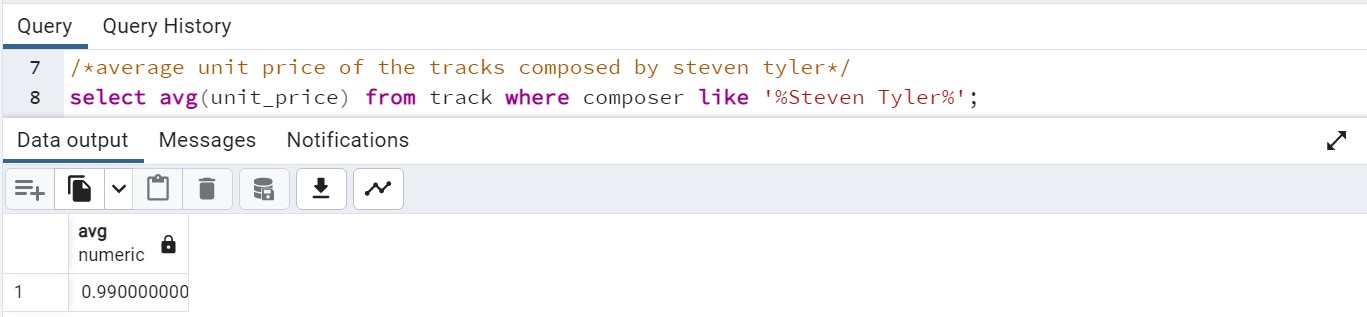
* + - Using where to filter values based on milliseconds column values less than or equal to 1min or 60000ms
    - Select statement to define which column’s data to display

1. No.of songs whose duration is less than or equal to 2 mins or 120000ms



* + count(\*) to count the number of records whose value in milliseconds column is less than or equal to 2 mins or 120000 ms

1. Average price of the music tracks composed by Steven Tyler



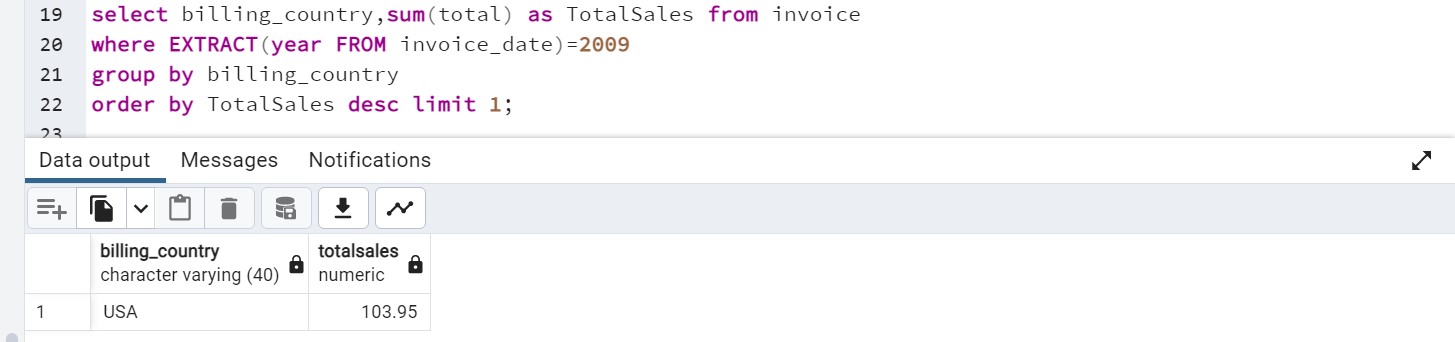
* + avg( ) function was used to compute mean price

1. Country wise total sales in the year 2011



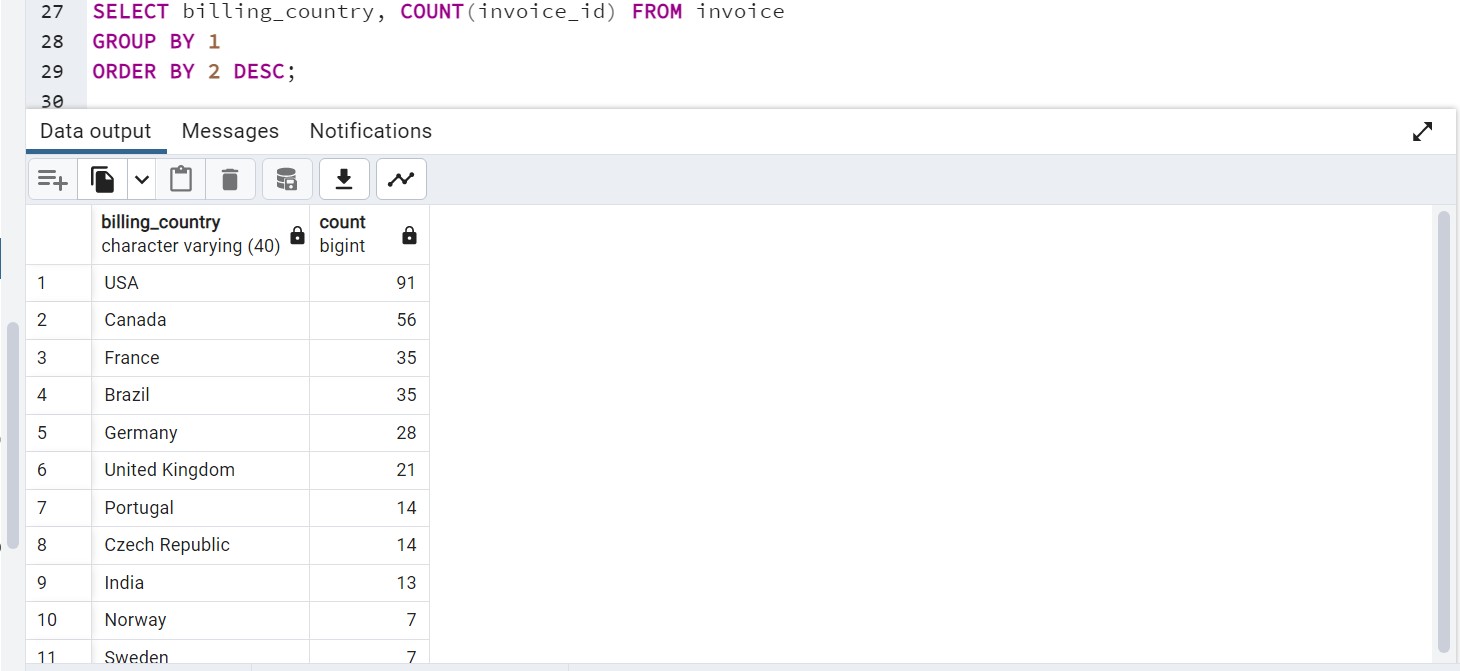
* + extract( ) function separates the datetime stamp into year,month and date, using this 2011 year is filtered out
  + All the records are grouped by billing\_country to get country eise total sales

1. Country with highest sales in the year 2009



* + Filtering the records for the year 2009 using extract( ) function, grouping the records wrt billing\_country to get country wise sales
  + Sorting the total sales in descending order and limit the output record by 1 to get the country with maximum most sales.

1. Country with most invoices

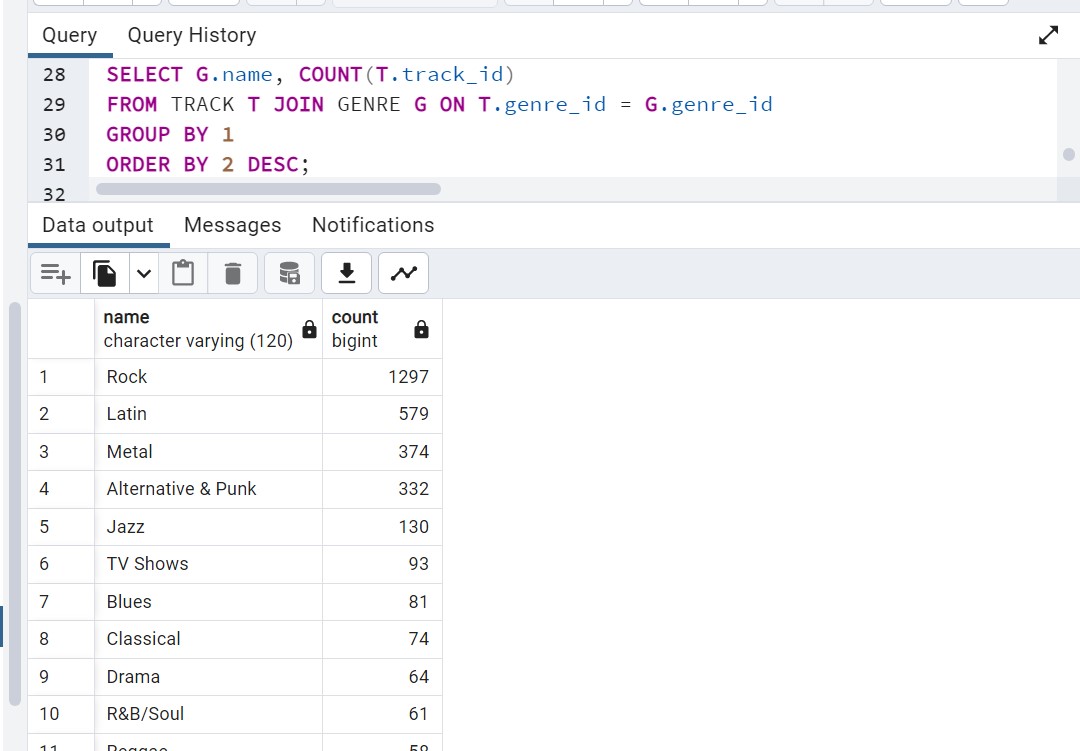


* + Grouping the records country wise using group by keyword and sorting wrt no.of invoices in descending order.

1. Highest no.of music tracks purchased by a single customer

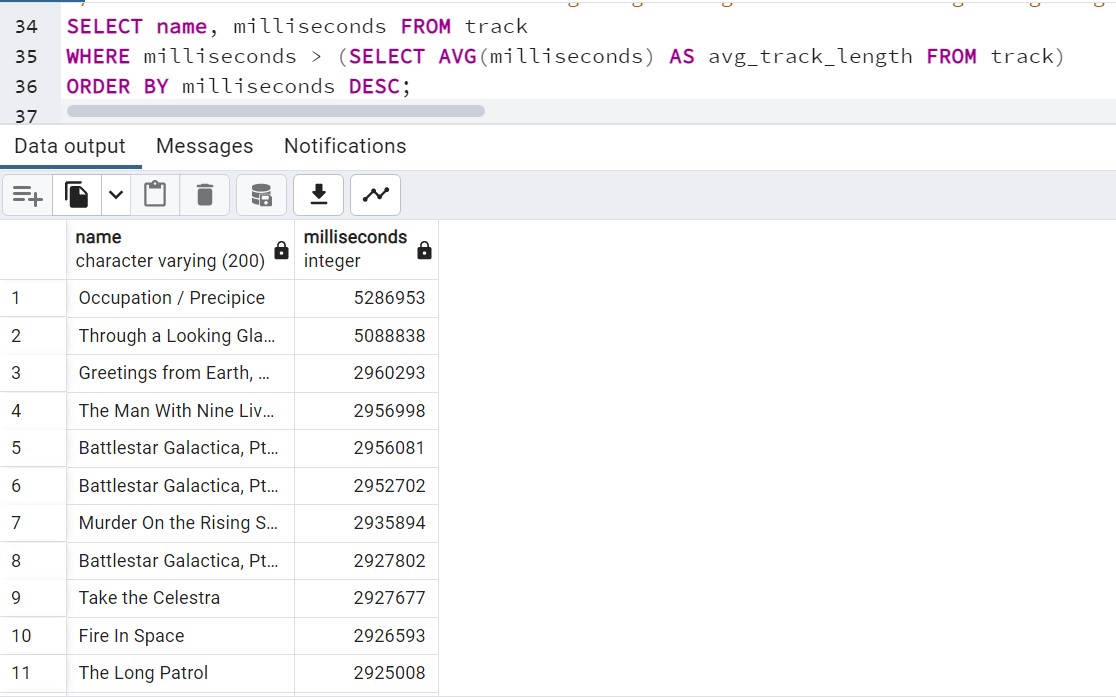


* + Joining the customer and invoice tables wrt customer\_id to get invoice details for each customer
  + Grouping the records by each customer id and sorting by no.of purchases by each customer in descending order and limiting the result to 1 record 8. No.of songs per Genre



* + Joining tables track and genre based on genre\_id to get genre wise track details

9. List of all music tracks that has a song length greater than avg song length



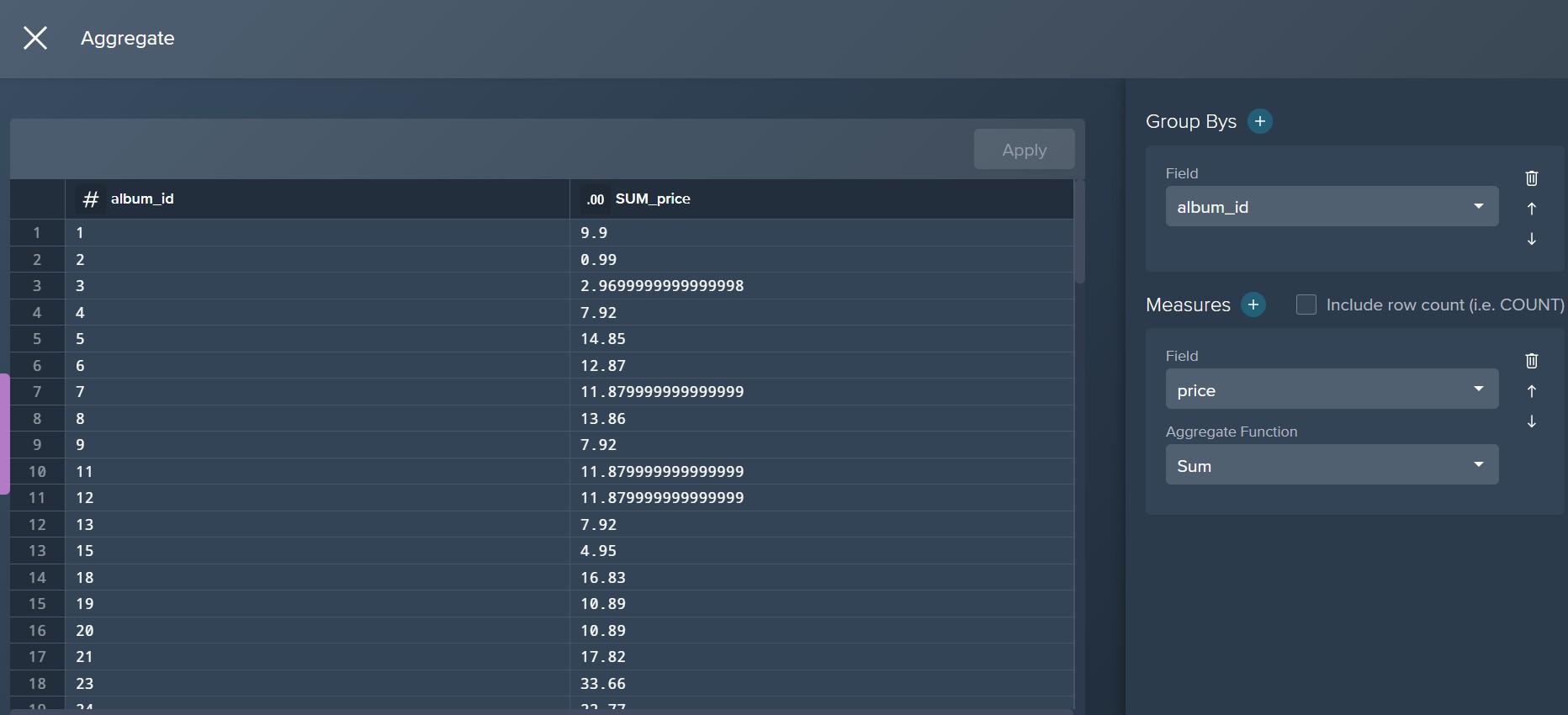
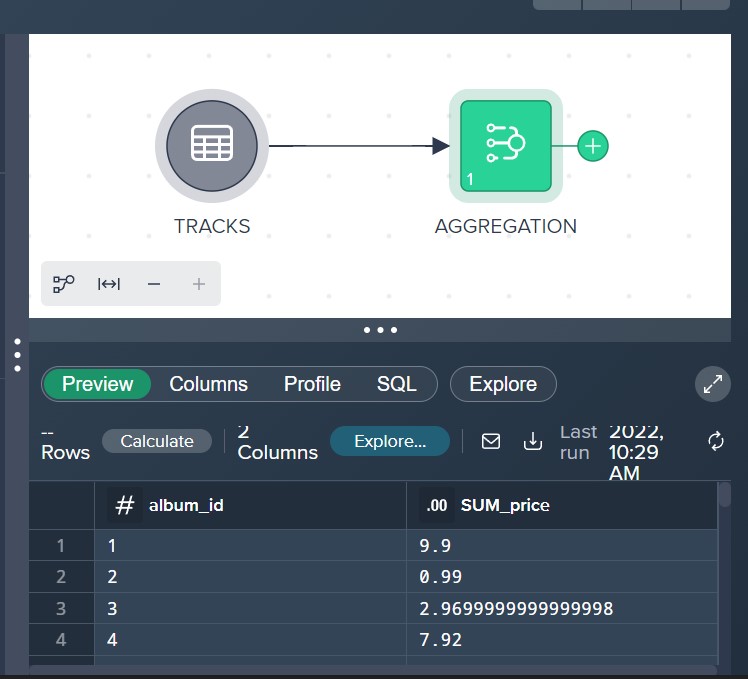
● Using a sub query to fetch the average duration of songs from track table, using this value then records can be filtered for all songs with a higher duration that this value in the outer query

**DATAMEER**

Implementation:

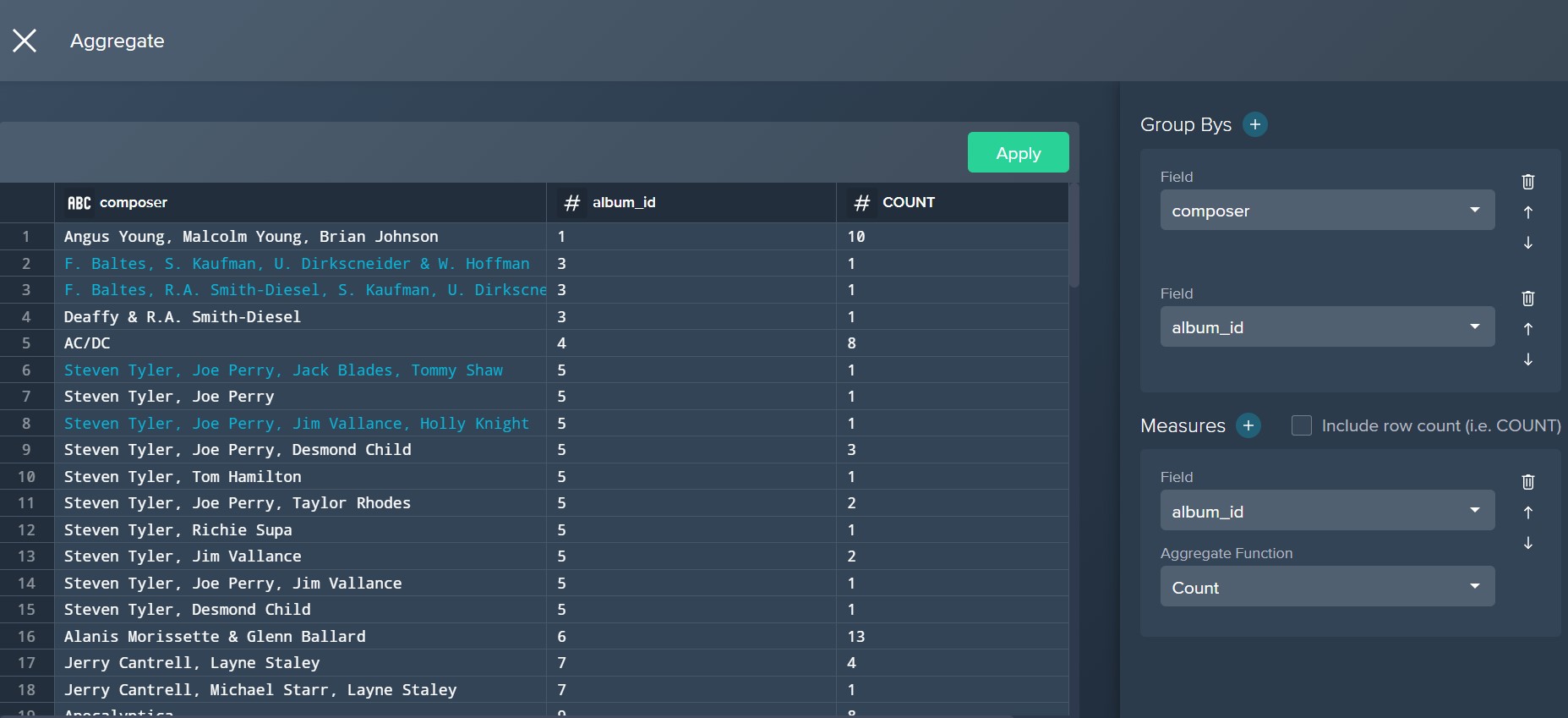
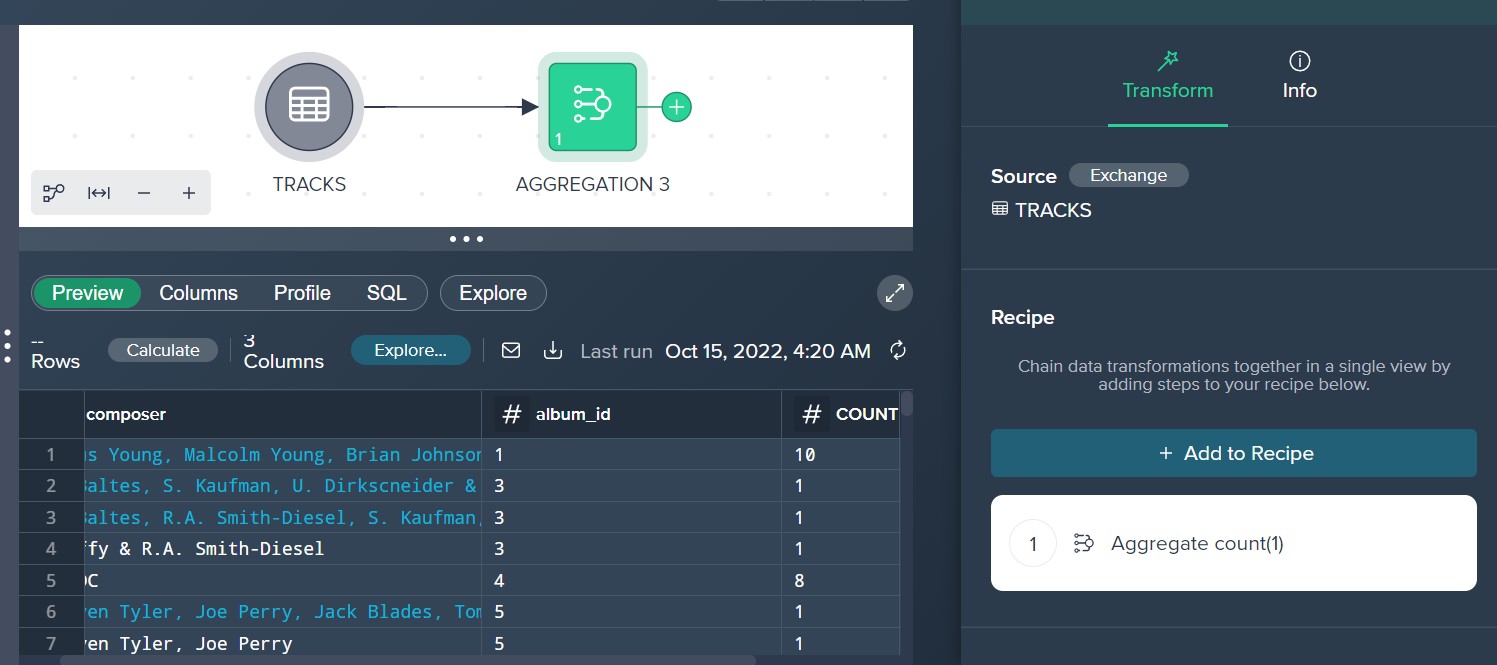
1. Album wise total price of music tracks

Using aggregation option on tracks table to fetch total price of al tracks album wise by performing group by on album id and and sum measure on price



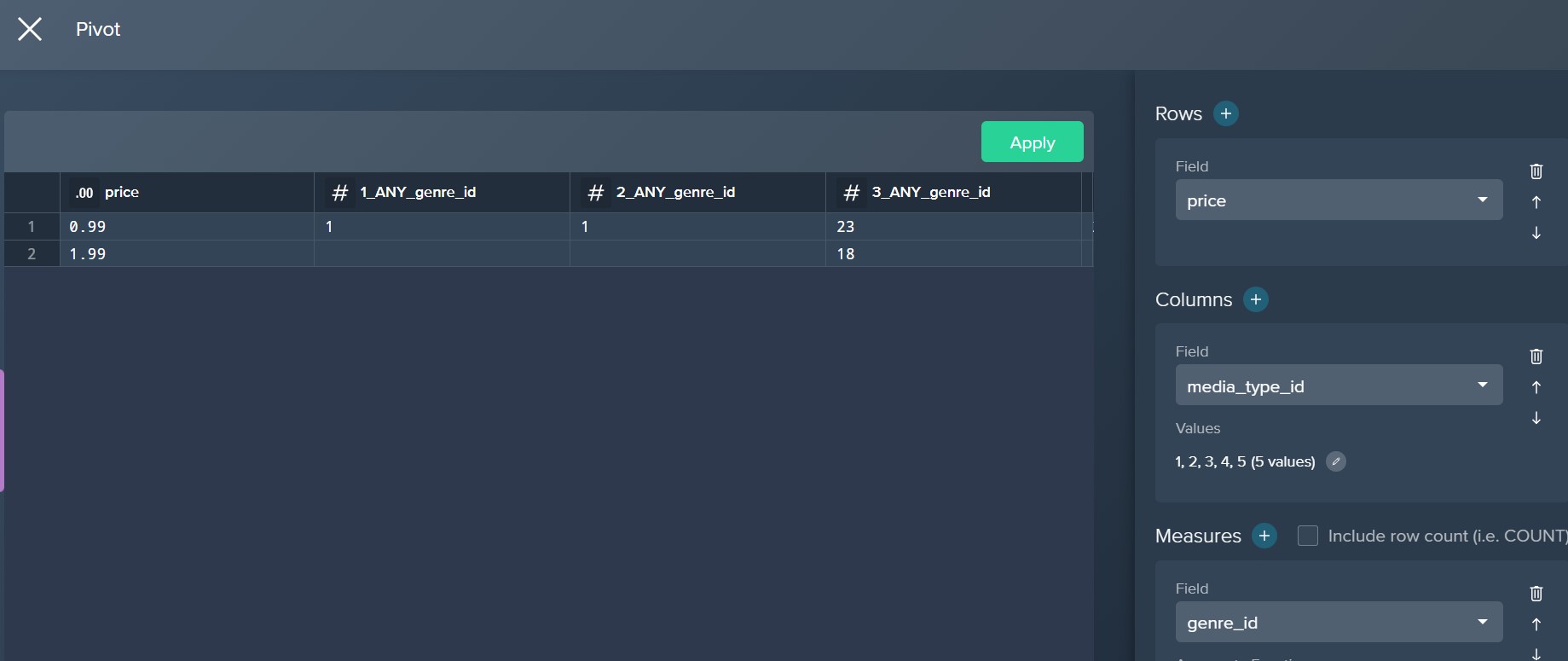
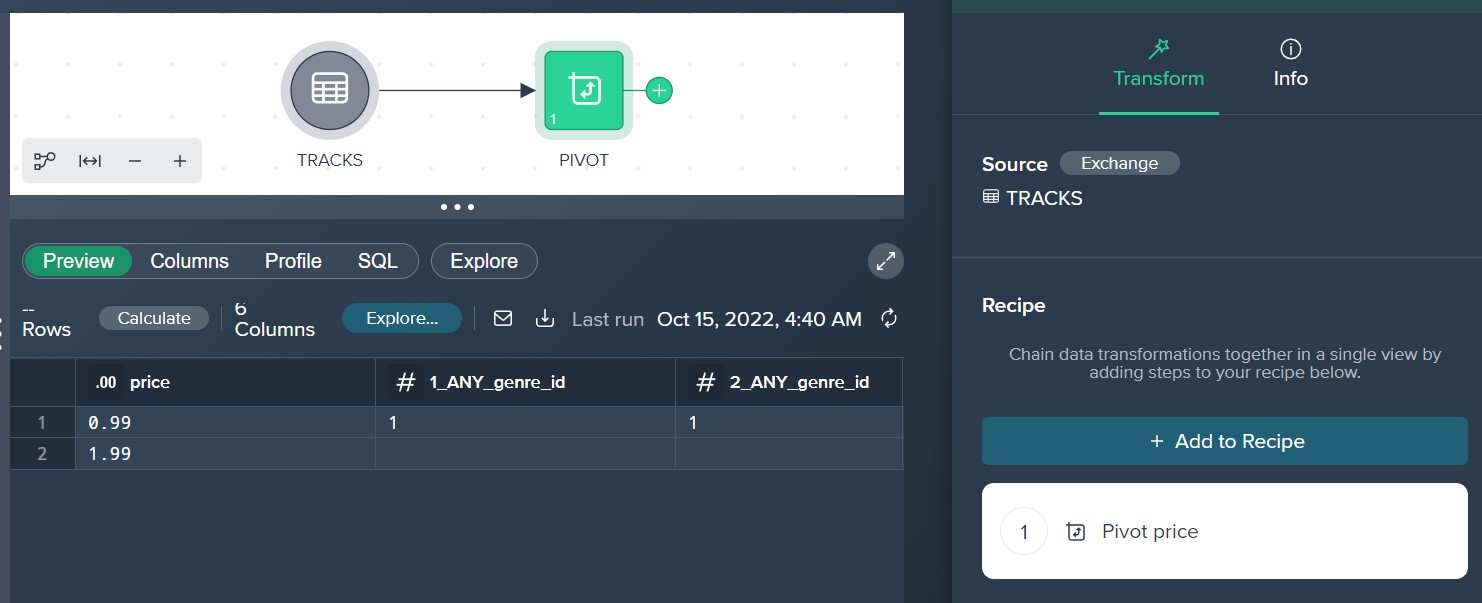
1. No.of songs composed by each composer album wise

Using aggregation option to group by compser\_id and album\_id, count measure on album\_id to get no.of songs for each composer album wise



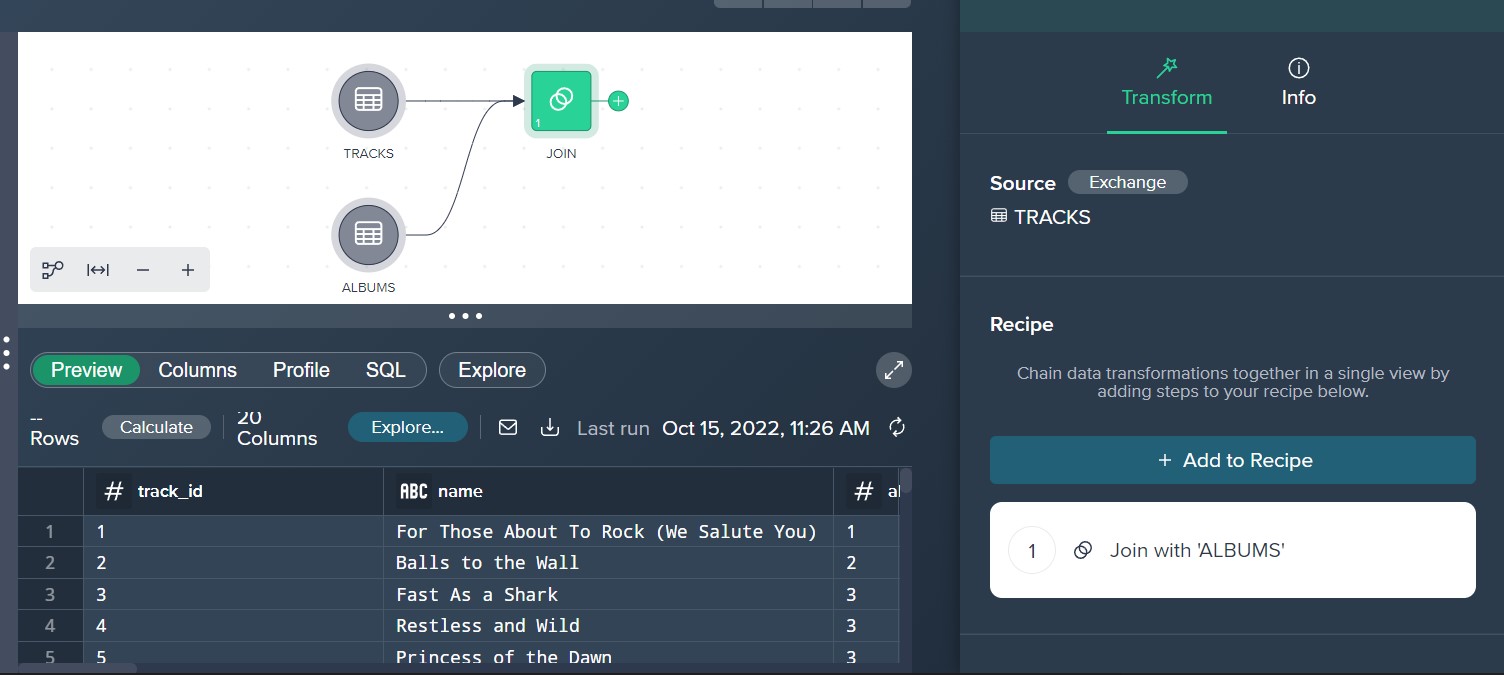
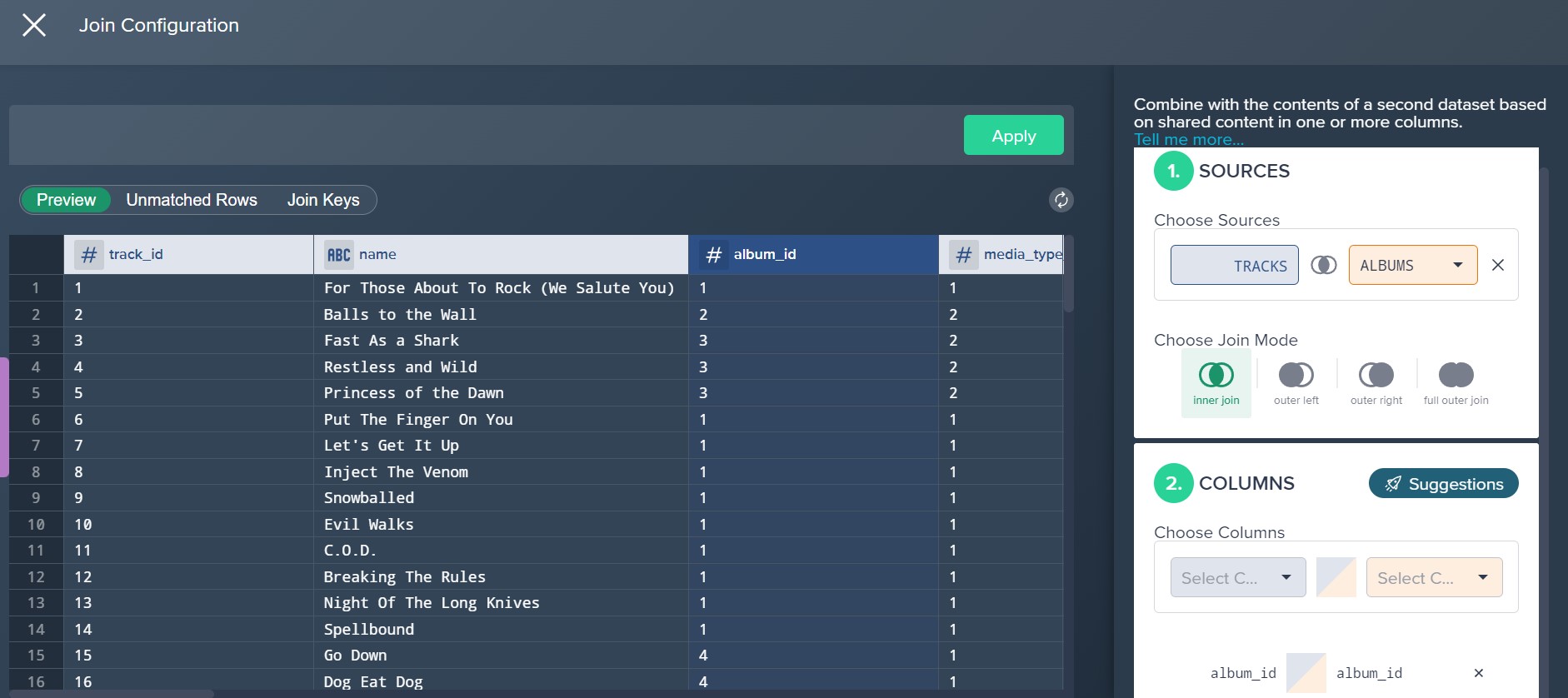
1. No.of songs genre and price wise - using pivot table

Using pivot option to create a pivot table to group price, genre\_id with no.of genres in each



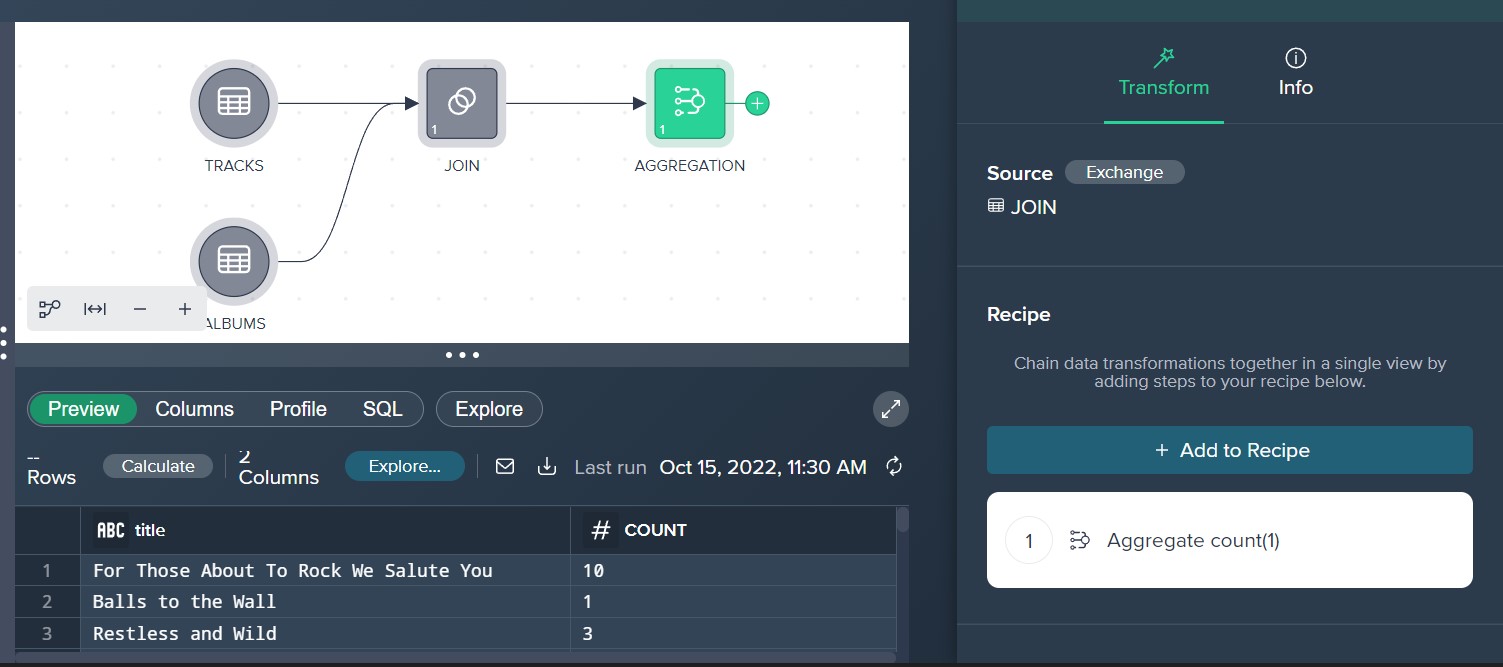
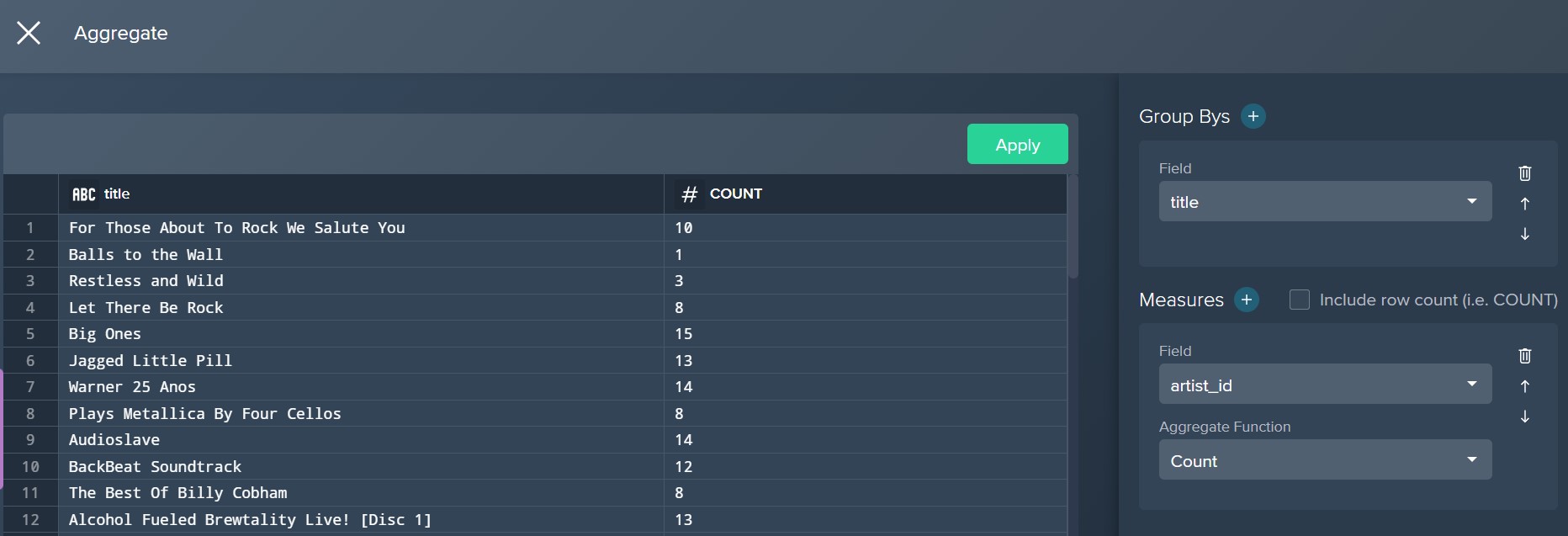
1. Joining tables track and album

Using join key word to join tracks and albums wrt album\_id column



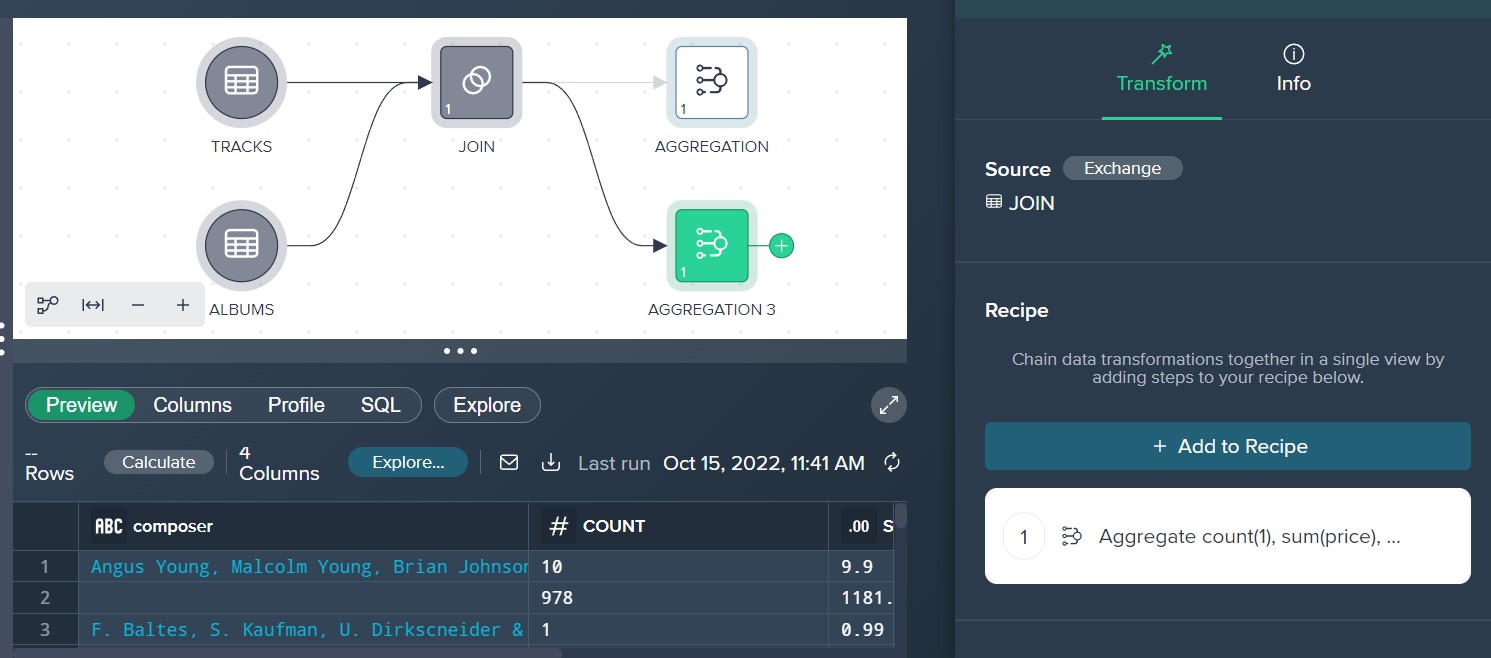
1. No.of artists who sang in each album - using both track and album tables

Using the above joined table, in grouping the records by track title with no.of artists in each song

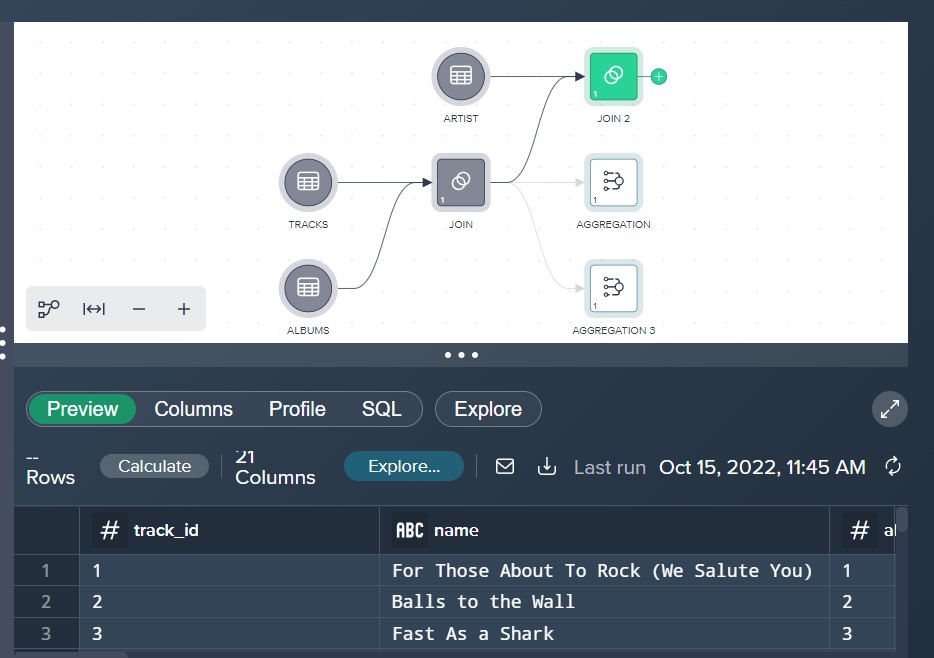
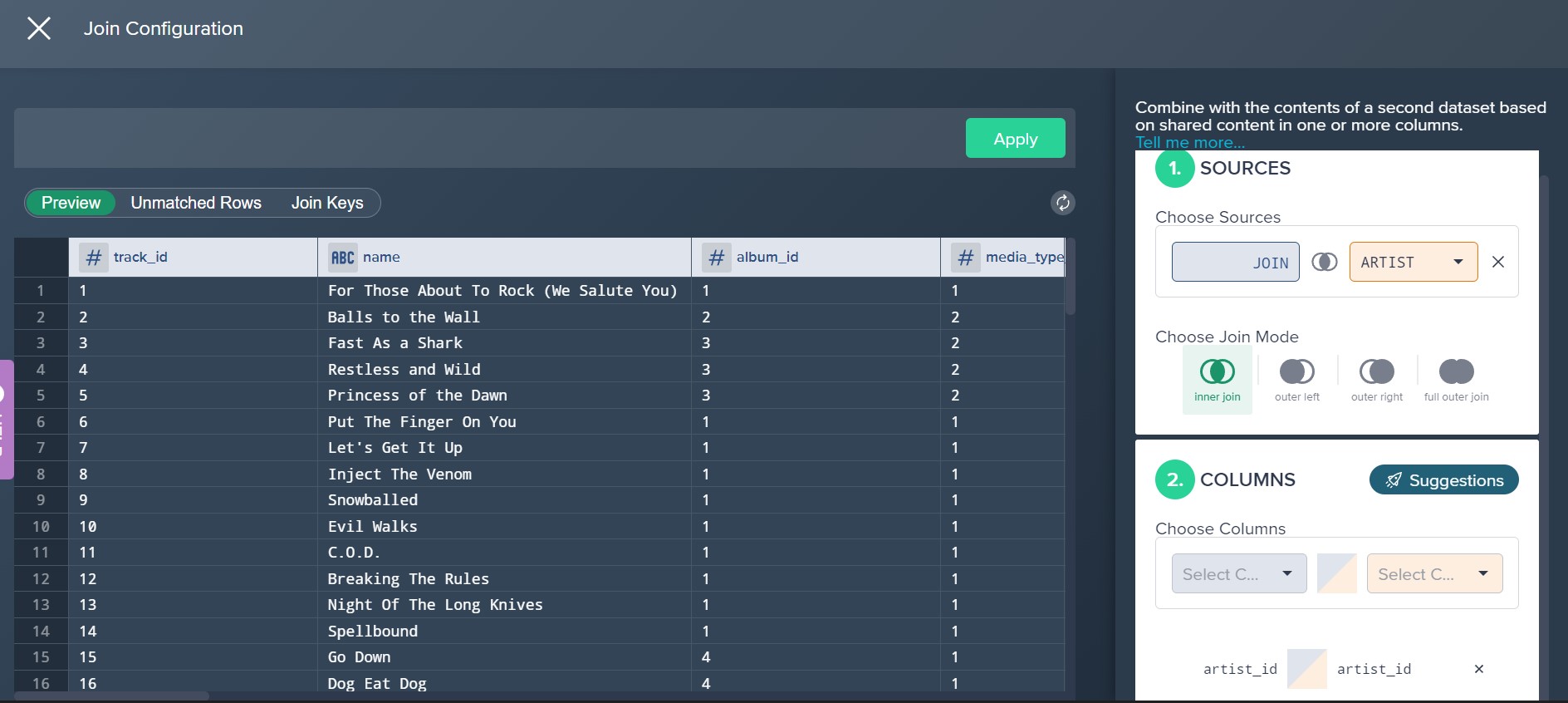


1. No.of albums composed by each composer along with total price of all the songs and total no.of genres composed

Using aggregation option to group by composer and count by album\_id and genre\_id, sum measure on price column



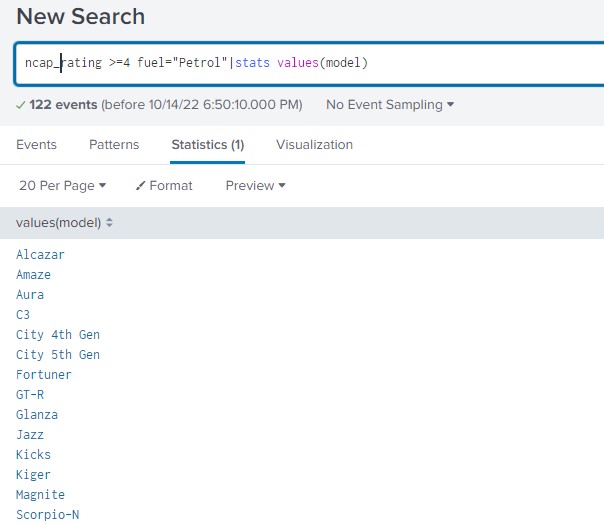
1. Joining 3 tables - track, album and artist



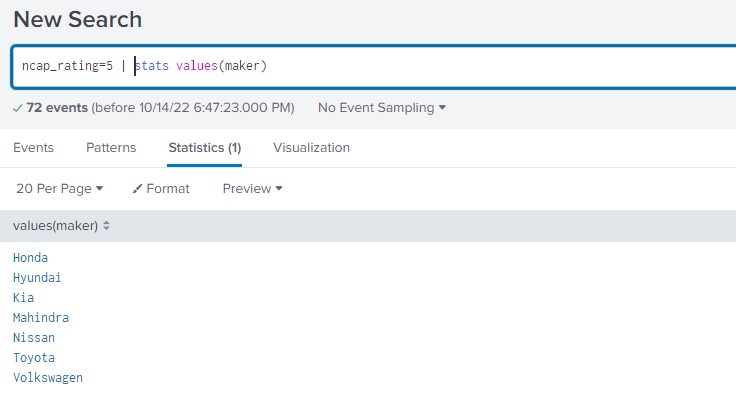
### SPLUNK

1. List of all petrol car models with high safety measures i.e NCAP Rating greater than or equal to 4

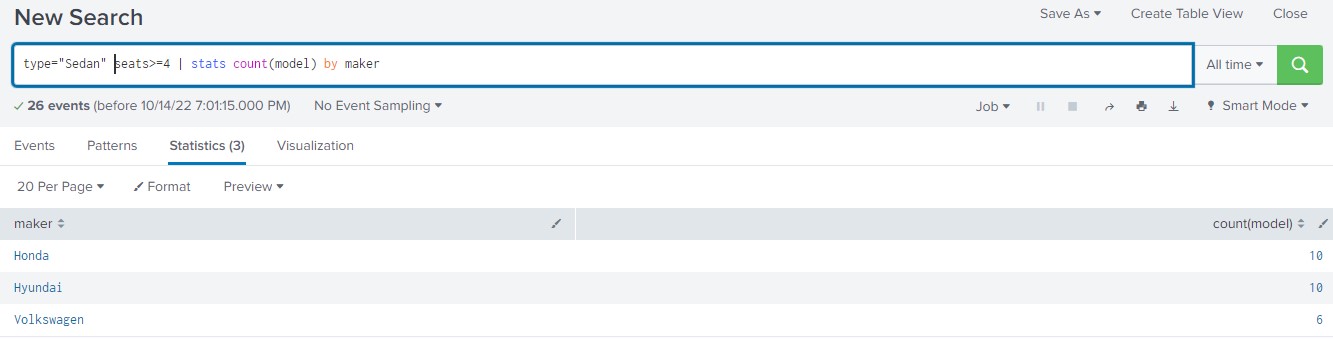
Filtering ncap\_rating greater than or equal to 4, fuel as petrol and retrieving all the distinct models



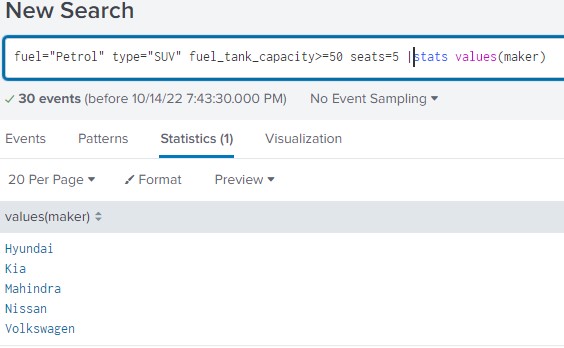
1. List of all car makers who produce cars with high safety measure with an NCAP Rating of 5



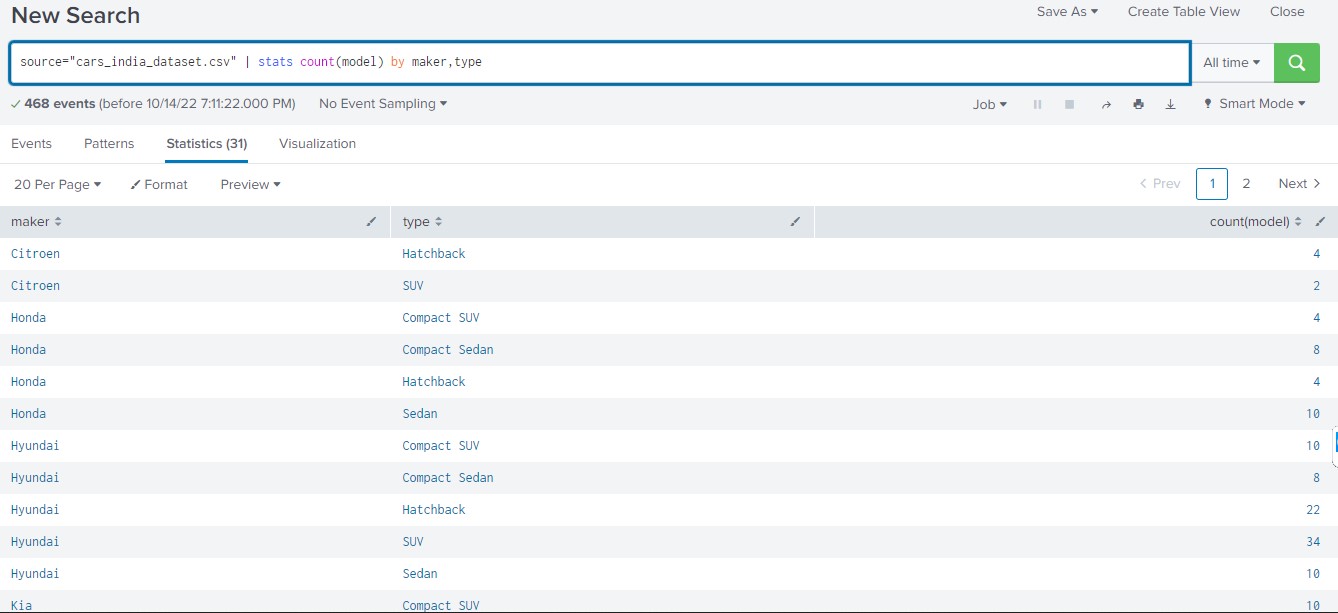
1. No.of different Sedan type car models produced by each maker with more than 4 seats



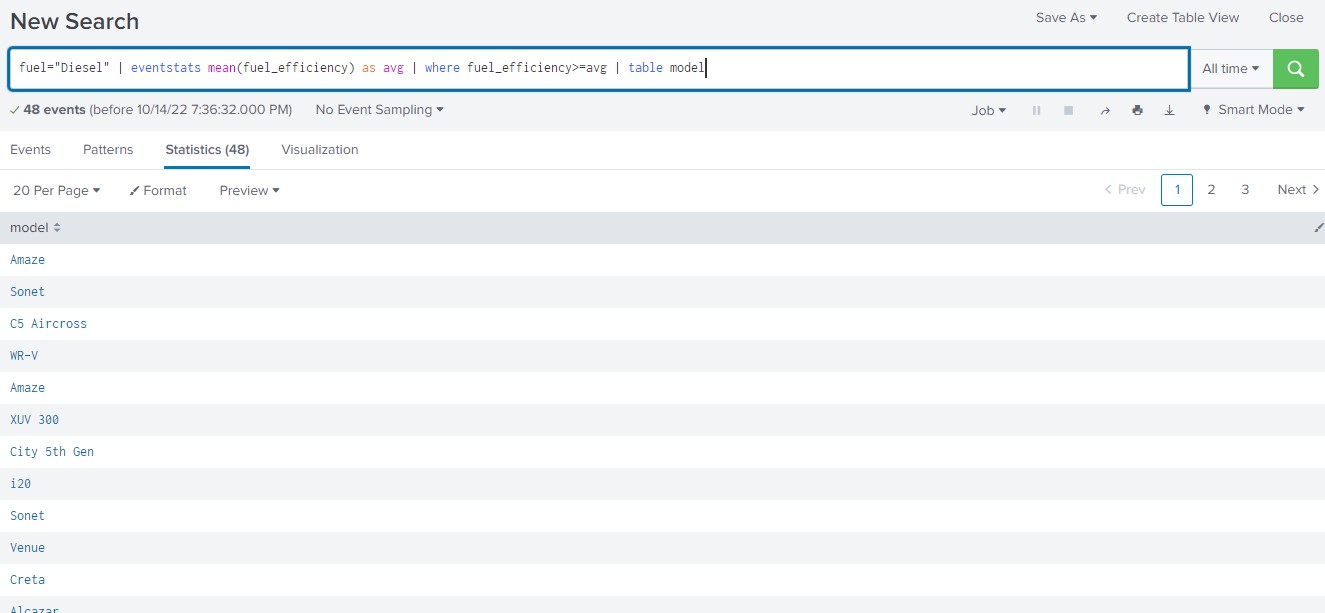
1. Find list of types of Petrol based SUV cars that have fuel tank capacity >= 50 and more than 5 seats



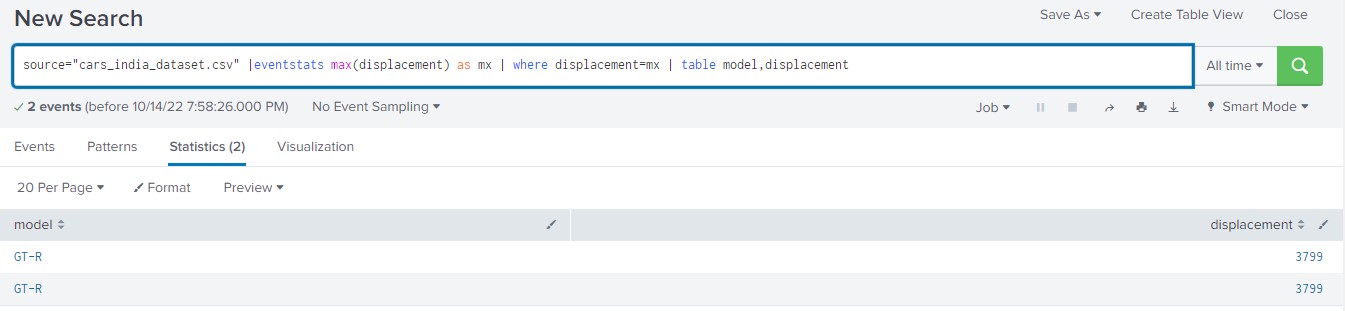
1. Type wise no.of cars produced by each maker



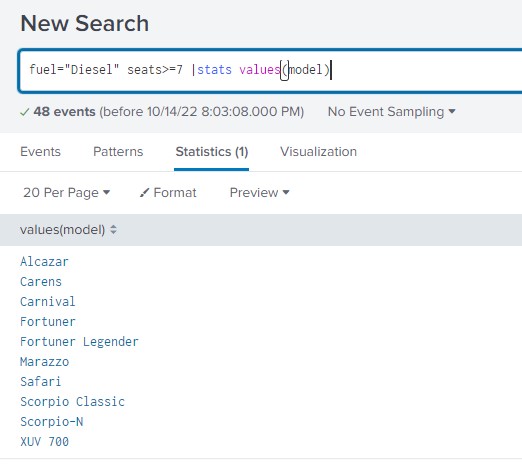
1. List of all Diesel cars that have a fuel efficiency greater than the average fuel efficiency



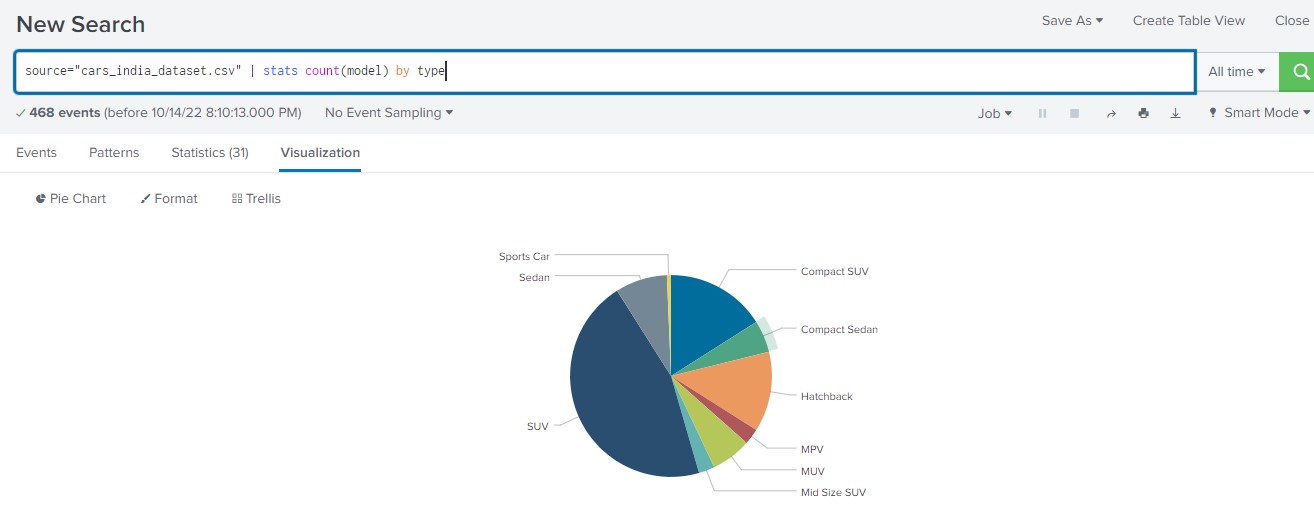
1. Model of the car with highest displacement



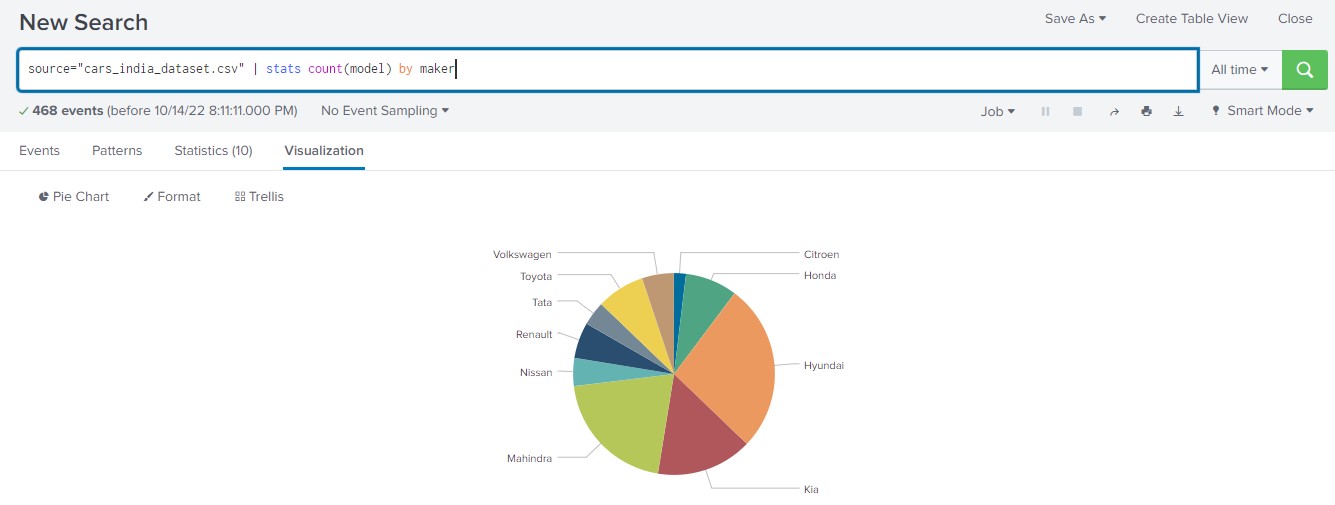
1. Distinct car models that has 7 seats and uses diesel consumption



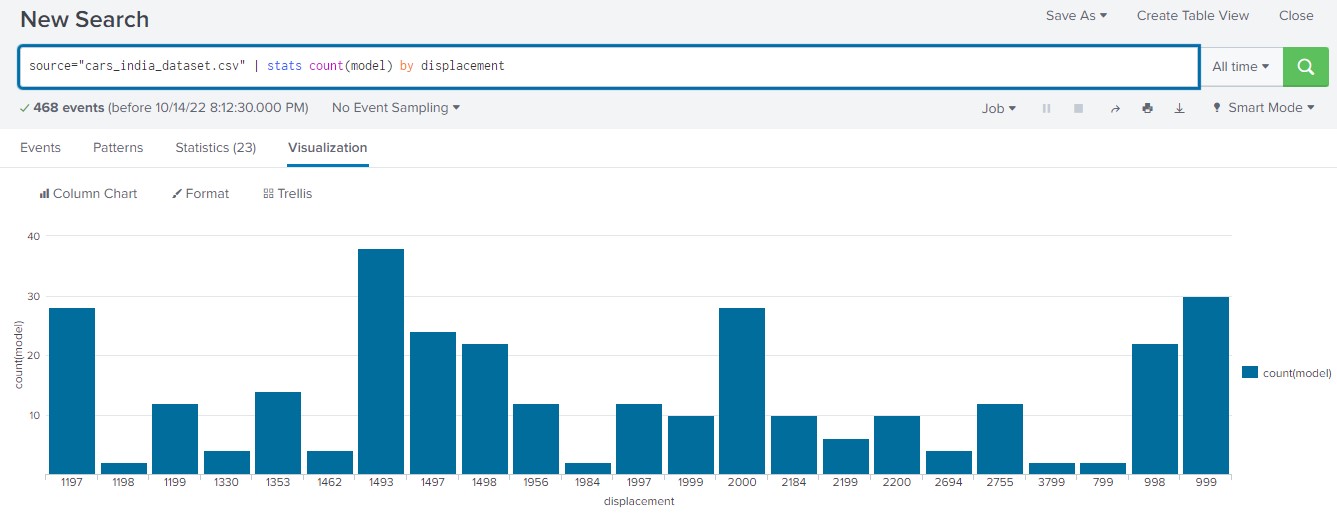
1. Visualizing types of car models via a pie chart



1. Visualizing different car makers via a pie chart



1. Visualizing no.of cars under each displacement via a bar chart

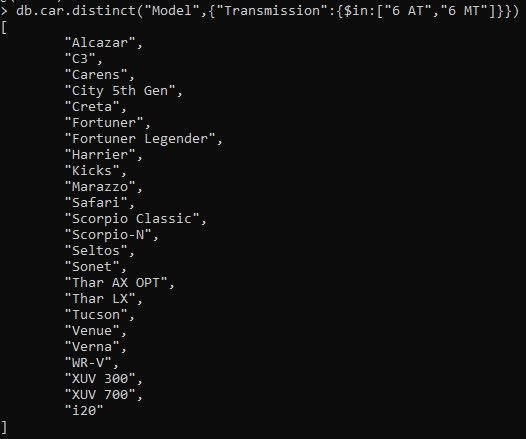


### MONGODB

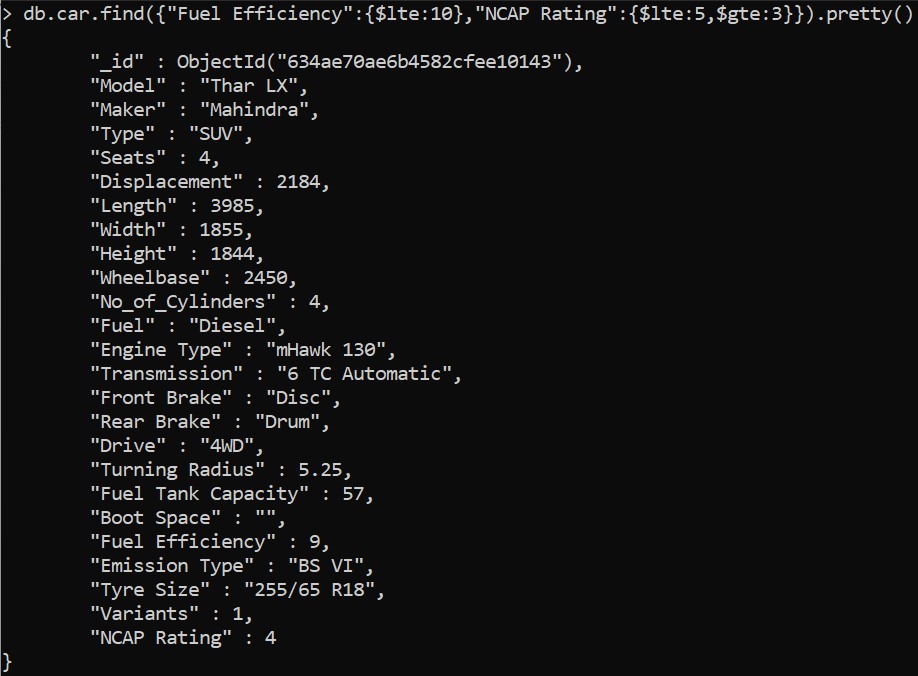
1. Updating the categorical values such as “Not Rated”, “Not Tested” as 0 in NCAP Rating column



1. Display all car models whose transmission is 6AT or 6MT



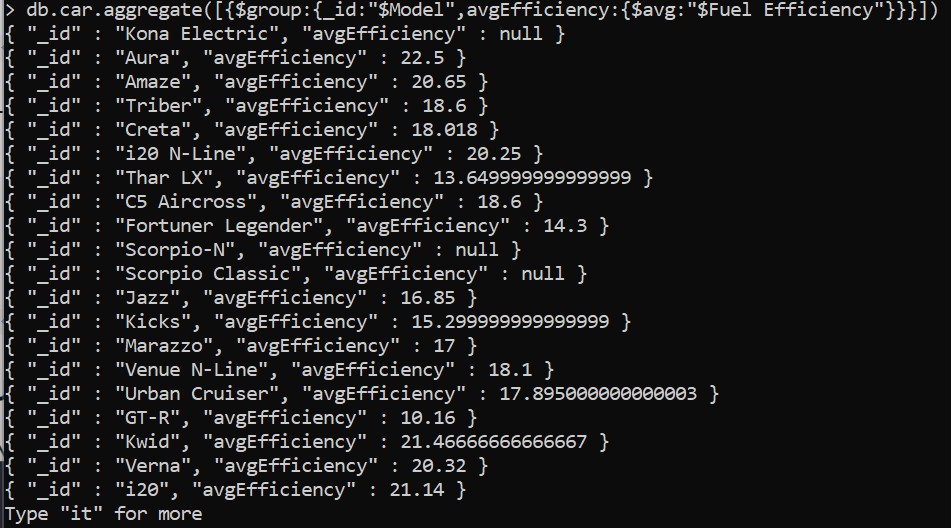
1. Find all car makers who produce cars with NCAP Rating between 3 and 4. And fuel efficiency is less than 10



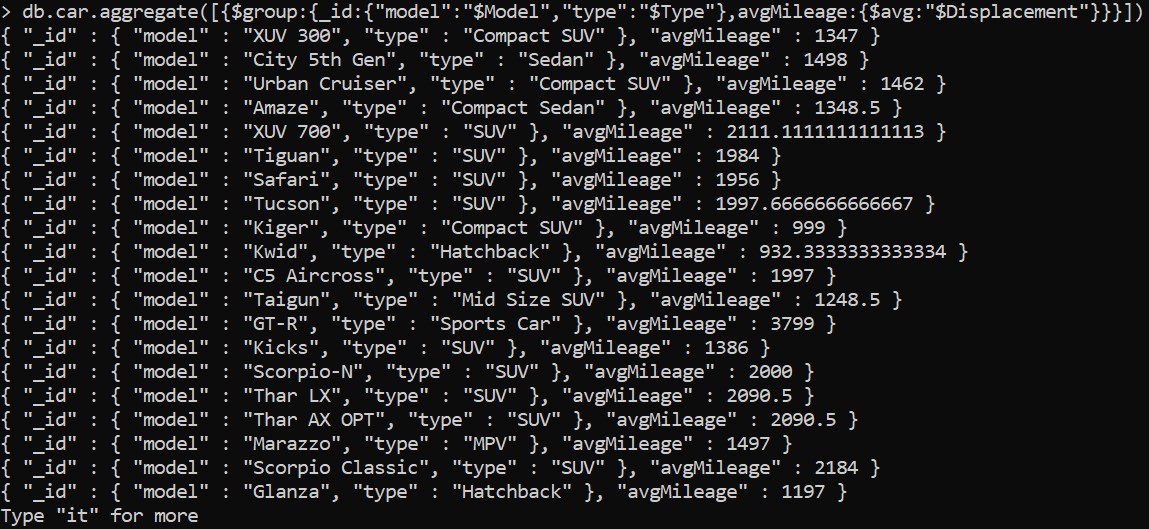
1. No.of different types of Petrol cars manufactured by Mahindra car maker



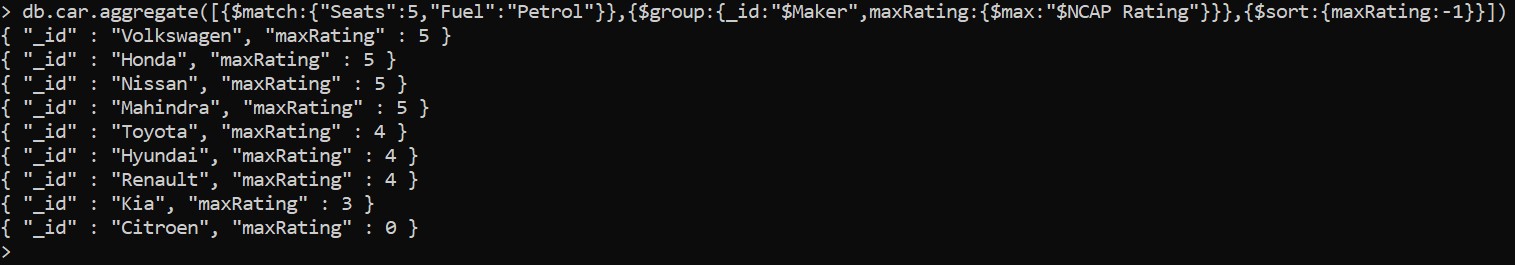
1. Average fuel efficiency Model wise



1. Model and type of car wise avg mileage - group by model and type of car



1. Max NCAP Rating according to car maker wise, among all those 5 seater Petrol cars



### CONCLUSION

On exploration of these data analysis tools a clear understanding towards various data retrieval mechanisms with modern tools and techniques was achieved. Different ways to apply the methodologies for data integration and manipulation using complex queries were implemented. And how to apply scalable data analytics framework in industry applications in a broader perspective was acquired and overall different ways to query and manage data integration, data manipulation using complex queries and data retrieval methods.