



SQL PROJECT

Restaurant & Consumer Data Project!

Hello students! For this project, we'll be working with a set of data that describes consumers, restaurants, their preferences, and the ratings consumers have given to various restaurants. Imagine this is the kind of data a company might use to build a restaurant recommendation system, understand customer behavior, or analyze restaurant performance.

Our goal is to take this raw data (currently in CSV files), organize it into a structured SQL database, and then learn how to ask interesting questions and get insights from it using SQL queries.

Understanding Your Data Files:

You have been provided with several CSV (Comma Separated Values) files. Each file represents a different aspect of our restaurant and consumer world:

<https://drive.google.com/drive/folders/1gDUe1mUA1cW8YgrsJIXYImwR0pORnzjc?usp=sharing>

If the data is too large, follow the steps in the document [Click Here](#)

1. **consumers.csv:**

- **What it's about:** This file contains information about individual consumers.
- **Key Identifier:** Consumer_ID (a unique ID for each consumer).
- **What you'll find inside:** Details like where each consumer lives (City, State, Country, Latitude, Longitude), their lifestyle (Smoker, Drink_Level), how they get around (Transportation_Method), personal details (Marital_Status, Children, Age), their job (Occupation), and their spending habits (Budget).
- **Think of it as:** A customer profile table.

2. consumer_preferences.csv:

- **What it's about:** This file links consumers to the types of cuisines they like.
- **Key Identifiers:** Consumer_ID (to link back to the consumers.csv file) and Preferred_Cuisine.
- **What you'll find inside:** Rows that show which consumer likes which type of food. A single consumer might have many preferred cuisines, so they might appear multiple times in this file, each time with a different cuisine.
- **Think of it as:** A table showing what kinds of food each customer enjoys.

3. restaurants.csv:

- **What it's about:** This file lists various restaurants and their details.
- **Key Identifier:** Restaurant_ID (a unique ID for each restaurant).
- **What you'll find inside:** Information like the restaurant's name, location (City, State, Country, Zip_Code, Latitude, Longitude), services offered (Alcohol_Service, Smoking_Allowed), price range (Price), whether it's a franchise (Franchise), the type of area it's in (Area), and parking availability (Parking).
- **Think of it as:** A restaurant directory.

4. restaurant_cuisines.csv:

- **What it's about:** This file links restaurants to the types of cuisines they serve.
- **Key Identifiers:** Restaurant_ID (to link back to the restaurants.csv file) and Cuisine.
- **What you'll find inside:** Rows indicating which restaurant serves which type of cuisine. A single restaurant might offer multiple cuisines, so it could appear multiple times here.
- **Think of it as:** The menu specialities of each restaurant.

5. ratings.csv:

- **What it's about:** This is a very important file! It records the ratings that consumers have given to specific restaurants.

- **Key Identifiers:** Consumer_ID (linking to consumers.csv), Restaurant_ID (linking to restaurants.csv).
- **What you'll find inside:** For each rating, it shows which consumer rated which restaurant, and the scores they gave for Overall experience, Food, and Service.
- **Think of it as:** The review scores table, connecting customers and restaurants.

6. **data_dictionary.csv:**

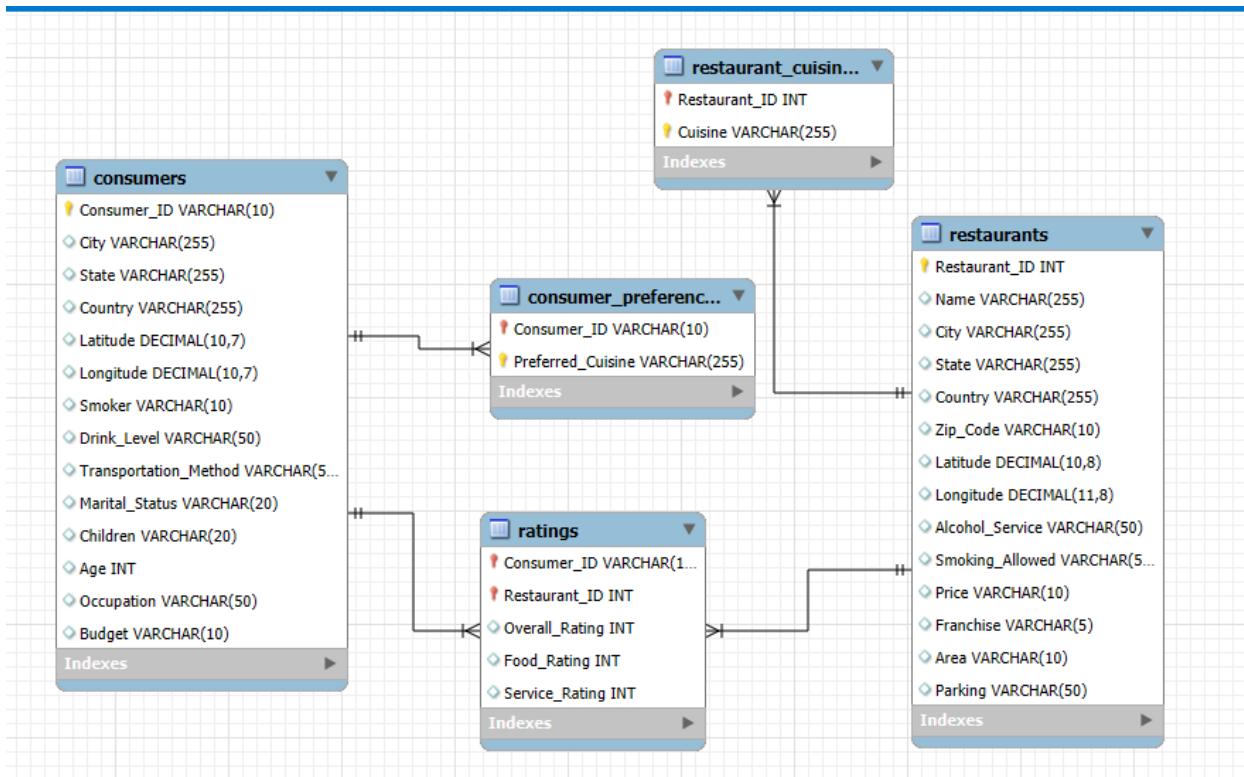
- **What it's about:** This is your "map" or "legend" for all the other files.
- **What you'll find inside:** It describes each column (Field) in every table, explaining what the data in that column means.
- **Think of it as:** A guidebook that helps you understand the meaning of every piece of data. Always refer to this if you're unsure about a column!

The Big Picture:

These files are related to each other. For example:

- A Consumer_ID in the consumers.csv file will also appear in consumer_preferences.csv to tell us what that specific consumer likes, and in ratings.csv to show us which restaurants that consumer has rated.
- A Restaurant_ID in restaurants.csv will also appear in restaurant_cuisines.csv to list its food types, and in ratings.csv to show its ratings from different consumers.

Just understand the ER and create the query



Objective:

Using the WHERE clause to filter data based on specific criteria.

1. List all details of consumers who live in the city of 'Cuernavaca'.
2. Find the Consumer_ID, Age, and Occupation of all consumers who are 'Students' AND are 'Smokers'.
3. List the Name, City, Alcohol_Service, and Price of all restaurants that serve 'Wine & Beer' and have a 'Medium' price level.
4. Find the names and cities of all restaurants that are part of a 'Franchise'.
5. Show the Consumer_ID, Restaurant_ID, and Overall_Rating for all ratings where the Overall_Rating was 'Highly Satisfactory' (which corresponds to a value of 2, according to the data dictionary).

Questions JOINs with Subqueries

1. List the names and cities of all restaurants that have an Overall_Rating of 2 (Highly Satisfactory) from at least one consumer.
2. Find the Consumer_ID and Age of consumers who have rated restaurants located in 'San Luis Potosi'.
3. List the names of restaurants that serve 'Mexican' cuisine and have been rated by consumer 'U1001'.
4. Find all details of consumers who prefer 'American' cuisine AND have a 'Medium' budget.

5. List restaurants (Name, City) that have received a Food_Rating lower than the average Food_Rating across all rated restaurants.
6. Find consumers (Consumer_ID, Age, Occupation) who have rated at least one restaurant but have NOT rated any restaurant that serves 'Italian' cuisine.
7. List restaurants (Name) that have received ratings from consumers older than 30.
8. Find the Consumer_ID and Occupation of consumers whose preferred cuisine is 'Mexican' and who have given an Overall_Rating of 0 to at least one restaurant (any restaurant).
9. List the names and cities of restaurants that serve 'Pizzeria' cuisine and are located in a city where at least one 'Student' consumer lives.
10. Find consumers (Consumer_ID, Age) who are 'Social Drinkers' and have rated a restaurant that has 'No' parking.

Questions Emphasizing WHERE Clause and Order of Execution

1. List Consumer_IDs and the count of restaurants they've rated, but only for consumers who are 'Students'. Show only students who have rated more than 2 restaurants.
2. We want to categorize consumers by an 'Engagement_Score' which is their Age divided by 10 (integer division). List the Consumer_ID, Age, and this calculated Engagement_Score, but *only* for consumers whose Engagement_Score would be exactly 2 and who use 'Public'

transportation.

3. For each restaurant, calculate its average Overall_Rating. Then, list the restaurant Name, City, and its calculated average Overall_Rating, but only for restaurants located in 'Cuernavaca' AND whose calculated average Overall_Rating is greater than 1.0.
4. Find consumers (Consumer_ID, Age) who are 'Married' and whose Food_Rating for any restaurant is equal to their Service_Rating for that same restaurant, but only consider ratings where the Overall_Rating was 2.
5. List Consumer_ID, Age, and the Name of any restaurant they rated, but only for consumers who are 'Employed' and have given a Food_Rating of 0 to at least one restaurant located in 'Ciudad Victoria'.

Advanced SQL Concepts: Derived Tables, CTEs, Window Functions, Views, Stored Procedures

1. Using a CTE, find all consumers who live in 'San Luis Potosi'. Then, list their Consumer_ID, Age, and the Name of any Mexican restaurant they have rated with an Overall_Rating of 2.
2. For each Occupation, find the average age of consumers. Only consider consumers who have made at least one rating. (Use a derived table to get consumers who have rated).
3. Using a CTE to get all ratings for restaurants in 'Cuernavaca', rank these ratings within each restaurant based on Overall_Rating (highest first). Display Restaurant_ID, Consumer_ID, Overall_Rating, and the RatingRank.
4. For each rating, show the Consumer_ID, Restaurant_ID, Overall_Rating, and also display the average Overall_Rating given by that specific consumer across all their ratings.
5. Using a CTE, identify students who have a 'Low' budget. Then, for each of these students, list their top 3 most preferred cuisines based on the order they appear in the Consumer_Preferences table (assuming no explicit preference order, use Consumer_ID, Preferred_Cuisine to define order for ROW_NUMBER).
6. Consider all ratings made by 'Consumer_ID' = 'U1008'. For each rating, show the Restaurant_ID, Overall_Rating, and the Overall_Rating of the *next* restaurant they rated (if any), ordered by Restaurant_ID (as a proxy for time if rating time isn't available). Use a derived table to filter for the consumer's ratings first.
7. Create a VIEW named HighlyRatedMexicanRestaurants that shows the Restaurant_ID, Name, and City of all Mexican restaurants that have an average Overall_Rating greater than 1.5.

8. First, ensure the HighlyRatedMexicanRestaurants view from Q7 exists. Then, using a CTE to find consumers who prefer 'Mexican' cuisine, list those consumers (Consumer_ID) who have *not* rated any restaurant listed in the HighlyRatedMexicanRestaurants view.
9. Create a stored procedure GetRestaurantRatingsAboveThreshold that accepts a Restaurant_ID and a minimum Overall_Rating as input. It should return the Consumer_ID, Overall_Rating, Food_Rating, and Service_Rating for that restaurant where the Overall_Rating meets or exceeds the threshold.
10. Identify the top 2 highest-rated (by Overall_Rating) restaurants for each cuisine type. If there are ties in rating, include all tied restaurants. Display Cuisine, Restaurant_Name, City, and Overall_Rating.
11. First, create a VIEW named ConsumerAverageRatings that lists Consumer_ID and their average Overall_Rating. Then, using this view and a CTE, find the top 5 consumers by their average overall rating. For these top 5 consumers, list their Consumer_ID, their average rating, and the number of 'Mexican' restaurants they have rated.
12. Create a stored procedure named GetConsumerSegmentAndRestaurantPerformance that accepts a Consumer_ID as input.

The procedure should:

1. Determine the consumer's "Spending Segment" based on their Budget:
 - o 'Low' -> 'Budget Conscious'
 - o 'Medium' -> 'Moderate Spender'
 - o 'High' -> 'Premium Spender'
 - o NULL or other -> 'Unknown Budget'
2. For all restaurants rated by this consumer:
 - o List the Restaurant_Name.
 - o The Overall_Rating given by this consumer.
 - o The average Overall_Rating this restaurant has received from *all* consumers (not just the input consumer).
 - o A "Performance_Flag" indicating if the input consumer's rating for that restaurant is 'Above Average', 'At Average', or 'Below Average' compared to the restaurant's overall average rating.
 - o Rank these restaurants for the input consumer based on the Overall_Rating they gave (highest rating = rank 1).

Challenges You Might Face

- Understanding table relationships and applying correct joins.
- Ensuring data consistency with foreign key constraints.
- Handling aggregation across joined tables.
- Extracting time-based trends from date data (especially if in VARCHAR format).

Project Presentation Template

As part of this project, you are required to create and present the analysis findings. Use the following PowerPoint template to structure your presentation:

 [**Click here to find the PPT Template for the Project Presentation**](#)

Submission

After completion of the project Zip the **.sql query file** and **PPT** upload the zip file with your name and batch number. In LMS.