Q1. **Spotify Penetration Analysis**

Market penetration is an important metric for understanding Spotify's performance and growth potential in different regions. You are part of the analytics team at Spotify and are tasked with calculating the active user penetration rate in specific markets.

For this task, 'active\_users' are defined based on the following criterias:

• last\_active\_date: The user must have interacted with Spotify within the last 30 days.

• monthly\_active\_sessions: The user must have engaged with Spotify for at least 5 sessions in the past month.

• listening\_hours: The user must have spent at least 10 hours listening on Spotify in the past month.

Based on the condition above, calculate the active 'user\_penetration\_rate' by using the following formula.

• Active User Penetration Rate = (Number of Active Spotify Users in the Market / Total Population of the Market)

Total Population of the market is based on both active and passive users. ​ The output should contain 'country' and 'active\_user\_penetration\_rate'. Make sure that all countries that appear in the dataset are also present in the output of your solution. Ensure there are 10 decimal places in your solution.

Let us assume the current\_day is 2024-01-31.

Q2. **Population Density**

You are working on a data analysis project at Deloitte where you need to analyze a dataset containing information about various cities. Your task is to calculate the population density of these cities, rounded to the nearest integer, and identify the cities with the minimum and maximum densities. The population density should be calculated as (Population / Area).

The output should contain 'city', 'country', 'density'.

Q3. **Aggregate Listening Data**

You are tasked with analyzing a Spotify-like dataset that captures user listening habits. For each user, calculate the total listening time and the count of unique songs they have listened to. In the database duration values are displayed in seconds. Round the total listening duration to the nearest whole minute.

The output should contain three columns: 'user\_id', 'total\_listen\_duration', and 'unique\_song\_count'.

Q4. **Customer Feedback Analysis**

Capital One's marketing team is working on a project to analyze customer feedback from their feedback surveys.

The team sorted the words from the feedback into three different categories;

• short\_comments • mid\_length\_comments • long\_comments

The team wants to find comments that are not short and come from social media. The output should include 'feedback\_id,' 'feedback\_text,' 'source\_channel,' and a calculated category.

Q5. **Common Friends Script**

You are analyzing a social network dataset at Google. Your task is to find mutual friends between two users, Karl and Hans. There is only one user named Karl and one named Hans in the dataset.

The output should contain 'user\_id' and 'user\_name' columns.

Q6. **Friday's Likes Count**

You have access to Facebook's database which includes several tables relevant to user interactions. For this task, you are particularly interested in tables that store data about user posts, friendships, and likes. Calculate the total number of likes made on friend posts on Friday.

The output should contain two different columns 'likes' and 'date'.

Q7. **Weekly Orders Report**

For each week, find the total number of orders. Include only the orders that are from the first quarter of 2023.

The output should contain 'week' and 'quantity'.

Q8. **Top Monthly Sellers**

You are provided with a transactional dataset from Amazon that contains detailed information about sales across different products and marketplaces. Your task is to list the top 3 sellers in each product category for January.

The output should contain 'seller\_id' , 'total\_sales' ,'product\_category' , 'market\_place', and 'month'.

Q9. **Peak Online Time**

You are given a dataset from Amazon that tracks and aggregates user activity on their platform in certain time periods. For each device type, find the time period with the highest number of active users.

The output should contain 'user\_count', 'time\_period', and 'device\_type', where 'time\_period' is a concatenation of 'start\_timestamp' and 'end\_timestamp', like ; "start\_timestamp to end\_timestamp".

Q10. **Movie Duration Match**

As a data scientist at Amazon Prime Video, you are tasked with enhancing the in-flight entertainment experience for Amazon’s airline partners. Your challenge is to develop a feature that suggests individual movies from Amazon's content database that fit within a given flight's duration. For flight 101, find movies whose runtime is less than or equal to the flight's duration.

The output should list suggested movies for the flight, including 'flight\_id', 'movie\_id', and 'movie\_duration'."

Q11. **Friday Purchases**

IBM is working on a new feature to analyze user purchasing behavior for all Fridays in the first quarter of the year. For each Friday separately, calculate the average amount users have spent per order. The output should contain the week number of that Friday and average amount spent.

Q12. **Finding Doctors**

Find doctors with the last name of 'Johnson' in the employee list. The output should contain both their first and last names.

Q13. **Employees With Same Birth Month**

Identify the number of employees within each department that share the same birth month. Your output should list the department, birth month, and the number of employees from that department who were born in that month. If a month has no employees born in it within a specific department, report this month as having 0 employees. The "profession" column stores the department names of each employee.

Q14. **Most Profitable Companies**

Find the 3 most profitable companies in the entire world. Output the result along with the corresponding company name. Sort the result based on profits in descending order.

Q15. **Workers With The Highest Salaries**

You have been asked to find the job titles of the highest-paid employees.

Your output should include the highest-paid title or multiple titles with the same salary.

Q16. **Users By Average Session Time**

Calculate each user's average session time. A session is defined as the time difference between a page\_load and page\_exit. For simplicity, assume a user has only 1 session per day and if there are multiple of the same events on that day, consider only the latest page\_load and earliest page\_exit, with an obvious restriction that load time event should happen before exit time event . Output the user\_id and their average session time.

Q17. **Activity Rank**

Find the email activity rank for each user. Email activity rank is defined by the total number of emails sent. The user with the highest number of emails sent will have a rank of 1, and so on. Output the user, total emails, and their activity rank. Order records by the total emails in descending order. Sort users with the same number of emails in alphabetical order. In your rankings, return a unique value (i.e., a unique rank) even if multiple users have the same number of emails. For tie breaker use alphabetical order of the user usernames.

Q18. **Algorithm Performance**

Meta/Facebook is developing a search algorithm that will allow users to search through their post history. You have been assigned to evaluate the performance of this algorithm.

We have a table with the user's search term, search result positions, and whether or not the user clicked on the search result.

Write a query that assigns ratings to the searches in the following way: • If the search was not clicked for any term, assign the search with rating=1 • If the search was clicked but the top position of clicked terms was outside the top 3 positions, assign the search a rating=2 • If the search was clicked and the top position of a clicked term was in the top 3 positions, assign the search a rating=3

As a search ID can contain more than one search term, select the highest rating for that search ID. Output the search ID and its highest rating.

*Example:* The search\_id 1 was clicked (clicked = 1) and its position is outside of the top 3 positions (search\_results\_position = 5), therefore its rating is 2.

Q19. **Distances Traveled**

Find the top 10 users that have traveled the greatest distance. Output their id, name and a total distance traveled.

Q20. **Finding User Purchases**

Write a query that'll identify returning active users. A returning active user is a user that has made a second purchase within 7 days of any other of their purchases. Output a list of user\_ids of these returning active users.