# **Problem 1 (Meta Hard Level)**

A table named “famous” has two columns called user id and follower id. It represents each user ID has a particular follower ID. These follower IDs are also users of #Facebook / Meta. Then, find the famous percentage of each user.

Famous Percentage = number of followers a user has / total number of users on the platform.

**Explanation:**

1. distinct\_users CTE: Combines user\_id and follower\_id using UNION to get all unique users on the platform. This helps us determine the total number of users.
2. follower\_count CTE: Counts the number of followers for each user\_id by grouping the rows in the famous table. This gives a list of users with their follower counts.
3. Final SELECT Statement: Uses the data from follower\_count and distinct\_users to calculate the famous percentage for each user.

**MySQL Solution:**



**MSSQL Solution:**



# **Problem 2 (Amazon Hard Level)**

Given a table 'sf\_transactions' of purchases by date, calculate the month-over-month percentage change in revenue. The output should include the year-month date (YYYY-MM) and percentage change, rounded to the 2nd decimal point, and sorted from the beginning of the year to the end of the year. The percentage change column will be populated from the 2nd month forward and calculated as

.

**Explanation:**

1. MonthlyRevenue CTE: Aggregates the total revenue for each month using FORMAT to convert the created\_at date to the format YYYY-MM.
2. RevenueChange CTE: Adds a column previous\_revenue using the LAG function, which fetches the total revenue of the previous month for each row.
3. Final SELECT: Calculates the percentage change as ((total\_revenue - previous\_revenue) / previous\_revenue) \* 100. The ROUND function ensures the percentage is rounded to two decimal places. The output is ordered by year\_month to display the data chronologically.

**MSSQL Server Solution:**

****

**MySQL Solution:**

****

# **Problem 3 (Google Medium Level)**

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.

**Explanation:**

1. The CTEs (karl\_friends and hans\_friends) efficiently find all friends for Karl and Hans, respectively.
2. The main query joins these CTEs with the users table to find the users who are present in both Karl's and Hans's friend lists (mutual friends).



# **Problem 4 (Uber Hard Level)**

Some forecasting methods are extremely simple and surprisingly effective. Naïve forecast is one of them. To create a naïve forecast for "distance per dollar" (defined as distance\_to\_travel / monetary\_cost), first sum the "distance to travel" and "monetary cost" values monthly. This gives the actual value for the current month. For the forecasted value, use the previous month's value. After obtaining both actual and forecasted values, calculate the root mean squared error (RMSE) using the formula

Report the RMSE rounded to two decimal places.



**MySQL Solution:**



**MSSQL Solution:**



# **Problem 5 (Microsoft Medium Level)**

Given a list of projects and employees mapped to each project, calculate by the amount of project budget allocated to each employee. The output should include the project title and the project budget rounded to the closest integer. Order your list by projects with the highest budget per employee first.

Explanation:

1. Joining Tables: The initial step involves joining the ms\_projects and ms\_emp\_projects tables on the project ID to combine project details (including titles and budgets) with employee assignments.
2. Grouping and Aggregating: The data is then grouped by project title and budget, allowing for the calculation of budget per employee by dividing the total budget of each project by the count of employees assigned to that project.
3. Rounding and Ordering: Finally, the computed budget per employee is rounded to the nearest integer, and the results are ordered in descending order to prioritize projects with the highest budget allocation per employee.



# **Problem 6 (Airbnb Medium Level)**

Find the total number of available beds per hosts' nationality. Output the nationality along with the corresponding total number of available beds. Sort records by the total available beds in descending order.

**Explanation**

1. Joining Tables: The first step involves joining the airbnb\_apartments and airbnb\_hosts tables on the host\_id. This allows us to combine the apartment details (such as the number of beds) with the host's nationality information.
2. Grouping and Aggregating: Next, the data is grouped by the host's nationality, so that the total number of beds available for each nationality can be calculated. The SUM() function is used to add up the beds (n\_beds) for all apartments hosted by individuals of the same nationality.
3. Sorting the Results: Finally, the results are ordered in descending order based on the total number of available beds.



# **Problem 7 (IBM Hard Level)**

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.

𝐓𝐨 𝐬𝐨𝐥𝐯𝐞 𝐭𝐡𝐢𝐬 𝐩𝐫𝐨𝐛𝐥𝐞𝐦, 𝐰𝐞 𝐧𝐞𝐞𝐝 𝐭𝐨:  
1. Identify the Fridays in the first quarter (Q1) of the year.  
2. Calculate the week number for each of these Fridays.  
3. Group the purchases by week number and calculate the average amount spent per order.

**MSSQL Solution**



**MySQL Solution:**



# **Problem 8 (Tesla Medium Level)**

You are given a table of product launches by company by year. Write a query to count the net difference between the number of products companies launched in 2020 with the number of products companies launched in the previous year. Output the name of the companies and a net difference of net products released for 2020 compared to the previous year.

**Explanation:**

1. Counting Products per Year: Using SUM with CASE statements, we count the number of products launched in 2020 and 2019 separately for each company.
2. Calculating Net Difference: We calculate the difference between 2020 and 2019 product counts to get the net change.
3. Ordering: The results are ordered by net\_difference in descending order to show companies with the highest increase first.



# **Problem 9 (Netflix Hard Problem)**

Find the genre of the person with the most number of oscar winnings.  
If there are more than one person with the same number of oscar wins, return the first one in alphabetic order based on their name. Use the names as keys when joining the tables.

**Explanation:**

1. WinnerCount CTE: Calculates the total Oscar wins for each nominee by counting rows where winner = 1.
2. Final Selection: The TOP 1 clause fetches all rows with the highest total\_wins, sorted alphabetically by name to handle ties. We join the WinnerCount CTE with nominee\_information on the nominee’s name to retrieve the top\_genre for the top nominee(s) in terms of Oscar wins.

**MSSQL Solution**



**MySQL Solution**



# **Problem 10 (Amazon Medium Level)**

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.

**MSSQL Solution**

****

**MySQL Solution**

****

# **Problem 11 (Nvidia, Microsoft Medium Level)**

Find the number of transactions that occurred for each product. Output the product name along with the corresponding number of transactions and order records by the product id in ascending order. You can ignore products without transactions.

**Explanation:**

1. Joining Tables: The INNER JOIN between excel\_sql\_inventory\_data (aliased as inv) and excel\_sql\_transaction\_data (aliased as trans) matches records by product\_id. This way, only products with transactions are included.
2. Counting Transactions: Using COUNT(trans.transaction\_id) counts the number of transactions for each product.
3. Grouping and Ordering: GROUP BY inv.product\_id, inv.product\_name groups by product\_id and product\_name to get the transaction count per product. ORDER BY inv.product\_id ASC sorts the output by product\_id in ascending order.



# **Problem 12 (LinkedIn, Dropbox Basic Level)**

Write a query that calculates the difference between the highest salaries found in the marketing and engineering departments. Output just the absolute difference in salaries.

**Explanation**

1. CASE is used to selectively get the salary for the "marketing" and "engineering" departments.
2. MAX is applied to retrieve the highest salary in each department.
3. ABS calculates the absolute difference between the two maximum values.



# **Problem 13 (Expedia, Airbnb Basic Level)**

Find the number of rows for each review score earned by 'Hotel Arena'. Output the hotel name (which should be 'Hotel Arena'), review score along with the corresponding number of rows with that score for the specified hotel.



# **Problem 14 (Amazon, Salesforce Basic Level)**

What is the total sales revenue of Samantha and Lisa?



# **Problem 15 (Google Medium Level)**

Find all records from days when the number of distinct users receiving emails was greater than the number of distinct users sending emails.  
**Explanation:**

1. The distinct\_counts CTE calculates the number of distinct to\_user and from\_user for each day.
2. The main query joins the original google\_gmail\_emails table with distinct\_counts on the day field, selecting only records where distinct\_receivers is greater than distinct\_senders.



# **Problem 16 (JP Morgan, Chase, Bloomberg Medium Level)**

Bank of Ireland has requested that you detect invalid transactions in December 2022. An invalid transaction is one that occurs outside of the bank's normal business hours. The following are the hours of operation for all branches:  
Monday - Friday 09:00 - 16:00  
Saturday & Sunday Closed  
Irish Public Holidays 25th and 26th December  
Determine the transaction ids of all invalid transactions.

**Explanation:**

1. MONTH(time\_stamp) = 12 AND YEAR(time\_stamp) = 2022: This filters transactions to include only those in December 2022.
2. DATEPART(WEEKDAY, time\_stamp) IN (1, 7): This checks if the transaction occurred on a Saturday (7) or Sunday (1).
3. CAST(time\_stamp AS TIME) < '09:00:00': This checks if the transaction time is before the opening hours.
4. CAST(time\_stamp AS TIME) > '16:00:00': This checks if the transaction time is after the closing hours.
5. (DATEPART(DAY, time\_stamp) IN (25, 26) AND MONTH(time\_stamp) = 12): This checks if the transaction occurred on the public holidays of December 25th or 26th.

**MSSQL Solution:**

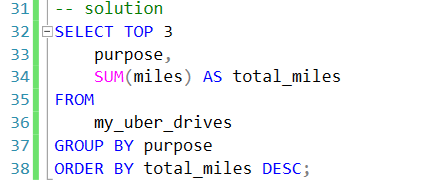


**MySQL Solution**

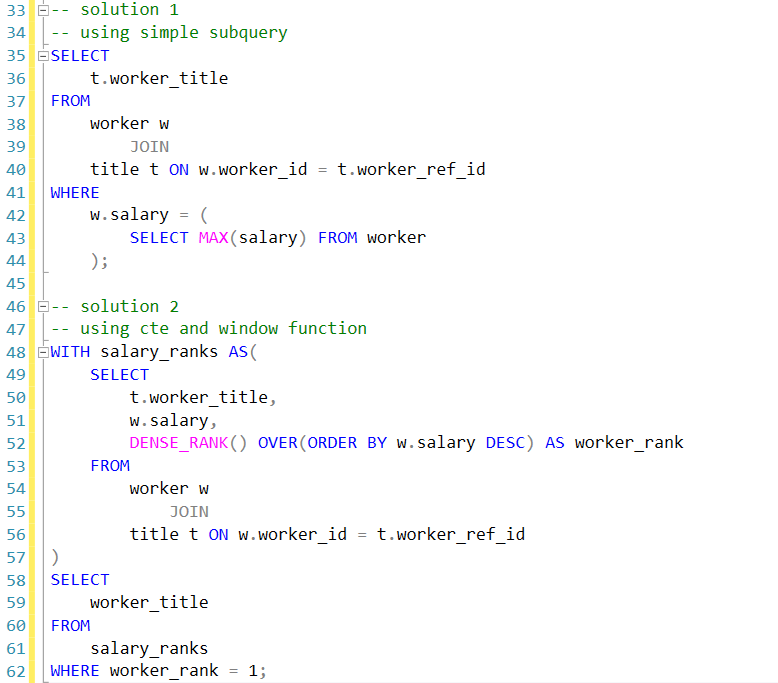


# **Problem 17 (Uber Medium Level)**

You’re given a table of Uber rides that contains the mileage and the purpose for the business expense. You’re asked to find business purposes that generate the most miles driven for passengers that use Uber for their business transportation. Find the top 3 business purpose categories by total mileage.



**Problem 18 (Amazon, Doordash Medium Level)**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.

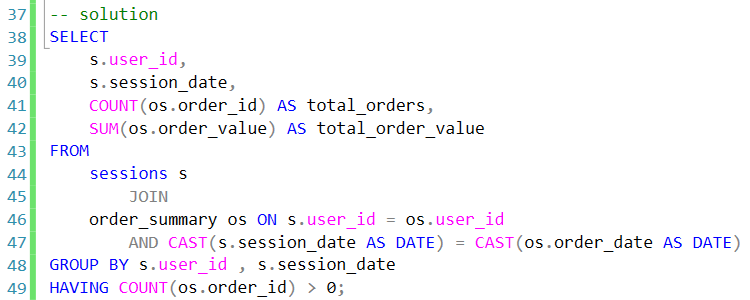


# **Problem 19 (Walmart Hard Level)**

Identify users who started a session and placed an order on the same day.

For these users, calculate the total number of orders and the total order value for that day.

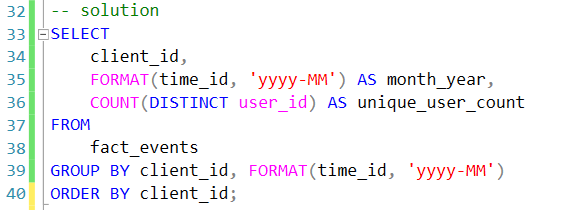
Your output should include the user, the session date, the total number of orders, and the total order value for that day.



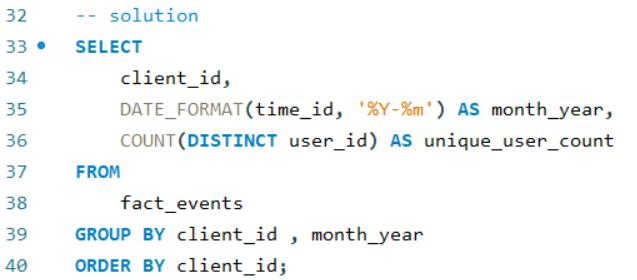
# **Problem 20 (Apple, Microsoft, Dell Easy Level)**

Write a query that returns the number of unique users per client per month

**MSSQL Solution**



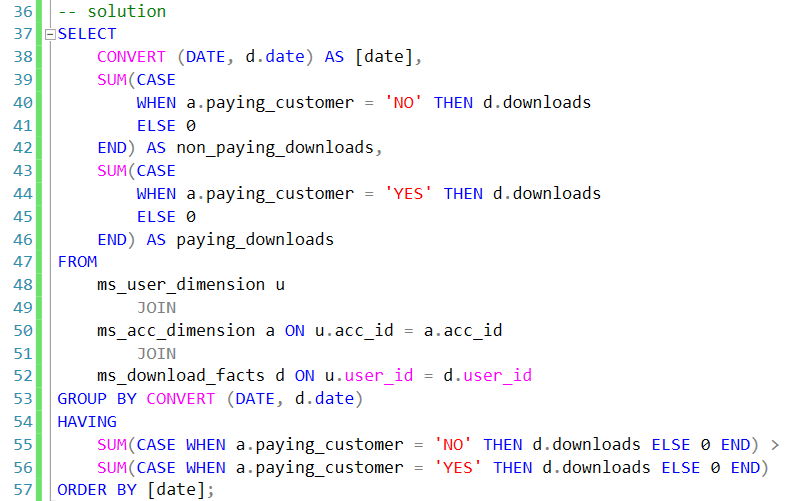
**MySQL Solution**



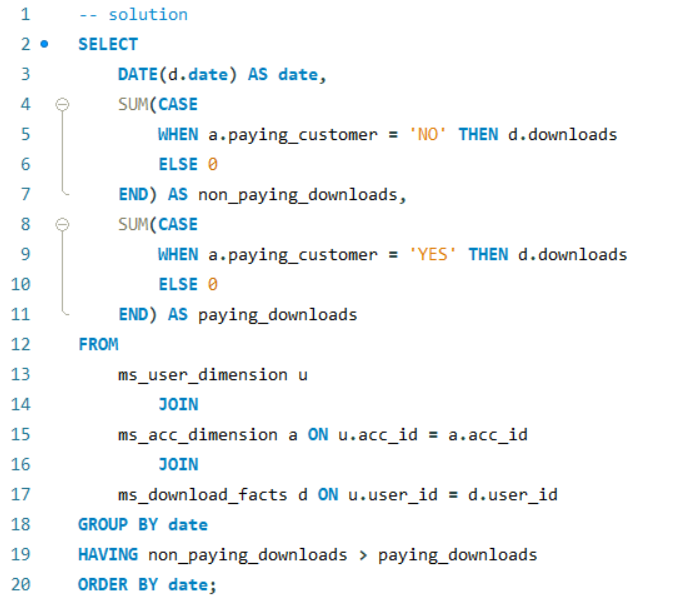
# **Problem 21 (Microsoft Hard Level)**

Find the total number of downloads for paying and non-paying users by date. Include only records where non-paying customers have more downloads than paying customers. The output should be sorted by earliest date first and contain 3 columns date, non-paying downloads, paying downloads.

**MSSQL Solution:**

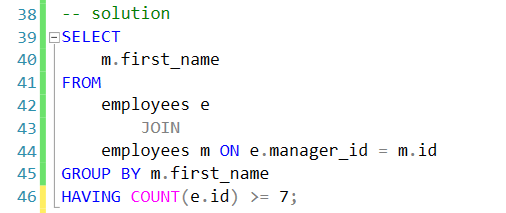


**MySQL Solution:**

****

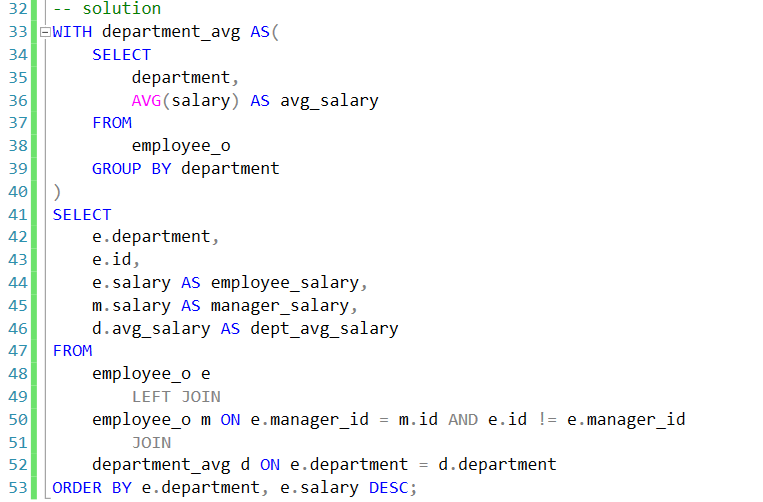
# **Problem 22 (Walmart, Paypal Medium Level)**

Find managers with at least 7 direct reporting employees. In situations where user is reporting to himself/herself, count that also.  
Output first names of managers.



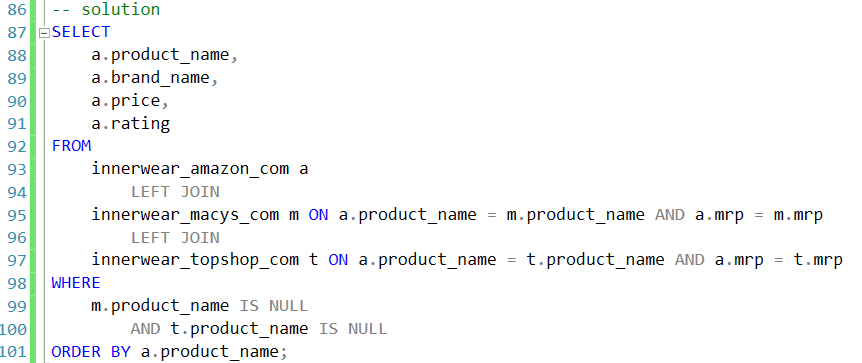
# **Problem 23 (Oracle Hard Level)**

Write a query that compares each employee's salary to their manager's and the average department salary (excluding the manager's salary). Display the department, employee ID, employee's salary, manager's salary, and department average salary. Order by department, then by employee salary (highest to lowest).



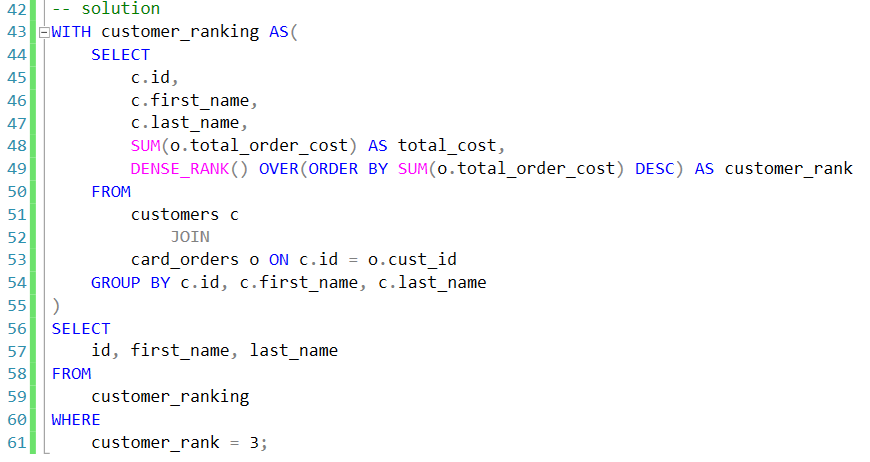
# **Problem 24 (Amazon Hard Level)**

Find products which are exclusive to only Amazon and therefore not sold at Top Shop and Macy's. Your output should include the product name, brand name, price, and rating.  
Two products are considered equal if they have the same product name and same maximum retail price (mrp column).



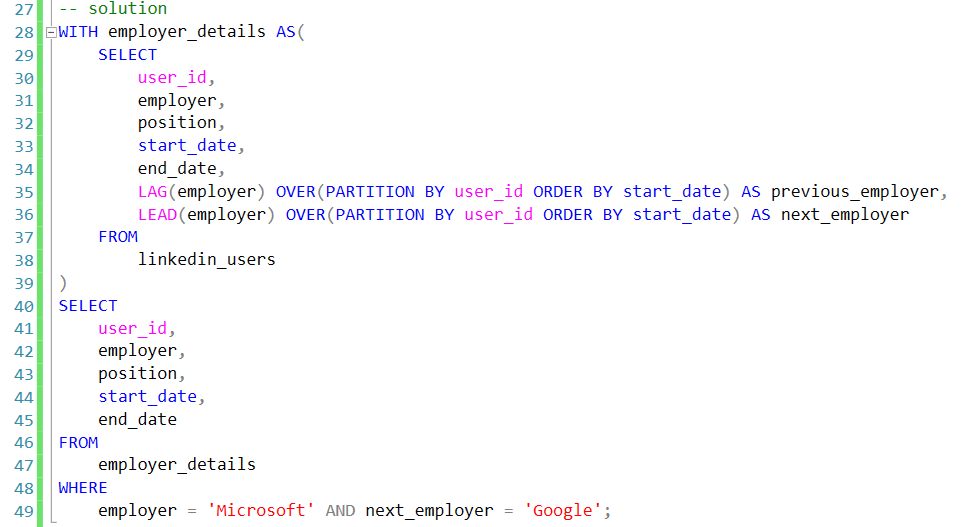
# **Problem 25 (American Express Medium Level)**

American Express is reviewing their customers' transactions, and you have been tasked with locating the customer who has the third highest total transaction amount. The output should include the customer's id, as well as their first name and last name. For ranking the customers, use type of ranking with no gaps between subsequent ranks.



# **Problem 26 (LinkedIn Hard Level)**

Consider all LinkedIn users who, at some point, worked at Microsoft. For how many of them was Google their next employer right after Microsoft (no employers in between)?

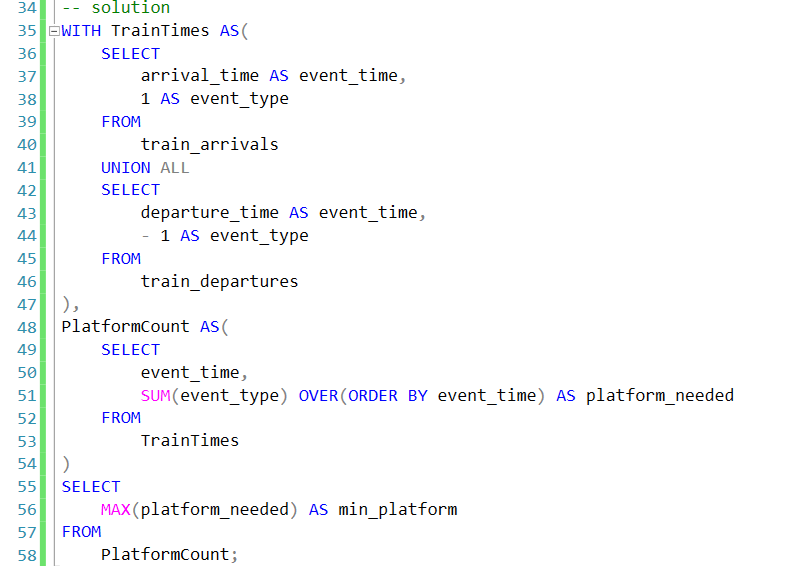


# **Problem 27 (Goldman Sachs, Deloitte Hard Level)**

You are given a day worth of scheduled departure and arrival times of trains at one train station. One platform can only accommodate one train from the beginning of the minute it's scheduled to arrive until the end of the minute it's scheduled to depart. Find the minimum number of platforms necessary to accommodate the entire scheduled traffic.

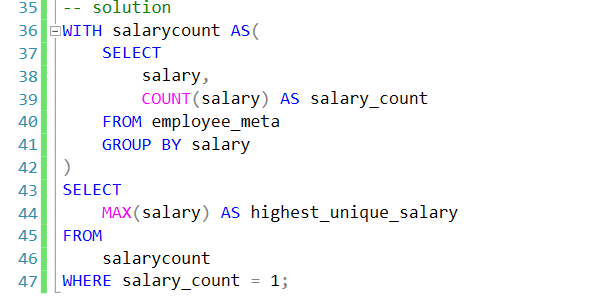
**Explanation:**

1. TrainTimes CTE: We combine both the arrival and departure times into one unified dataset. For arrivals, we use 1 as the event\_type to indicate the need for a platform. For departures, we use -1 as the event\_type to indicate the freeing of a platform.
2. PlatformCount Subquery: We use a SUM with a window function (OVER clause) to maintain a running count of the platforms needed. The event\_type for arrival adds one platform and the event\_type for departure subtracts one.
3. Max(platforms\_needed): Finally, we get the maximum value from the platforms\_needed, which represents the maximum number of platforms required at any point in time.



# **Problem 28 (Meta, Salesforce Hard Level)**

Find the highest salary among salaries that appears only once.



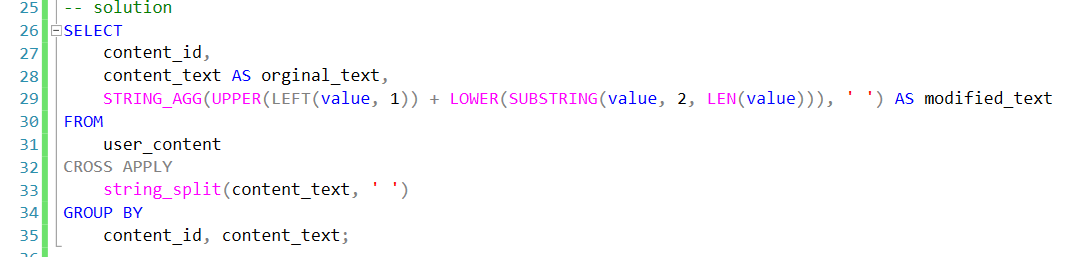
# **Problem 29 (Cisco Hard Level)**

Convert the first letter of each word found in content\_text to uppercase, while keeping the rest of the letters lowercase. Your output should include the original text in one column and the modified text in another column.

**Explanation:**

1. STRING\_SPLIT(content\_text, ' ') splits content\_text by spaces, breaking it down into words.
2. CROSS APPLY allows the function to be applied to each row, splitting content\_text into individual words as rows.
3. UPPER(LEFT(value, 1)) converts the first letter of each word to uppercase.
4. LOWER(SUBSTRING(value, 2, LEN(value))) converts the rest of each word to lowercase.
5. STRING\_AGG(..., ' ') aggregates the words back into a single string, with each word separated by a space.
6. The result is grouped by content\_id and content\_text, displaying both the original and modified text.

**MSSQL Solution:**

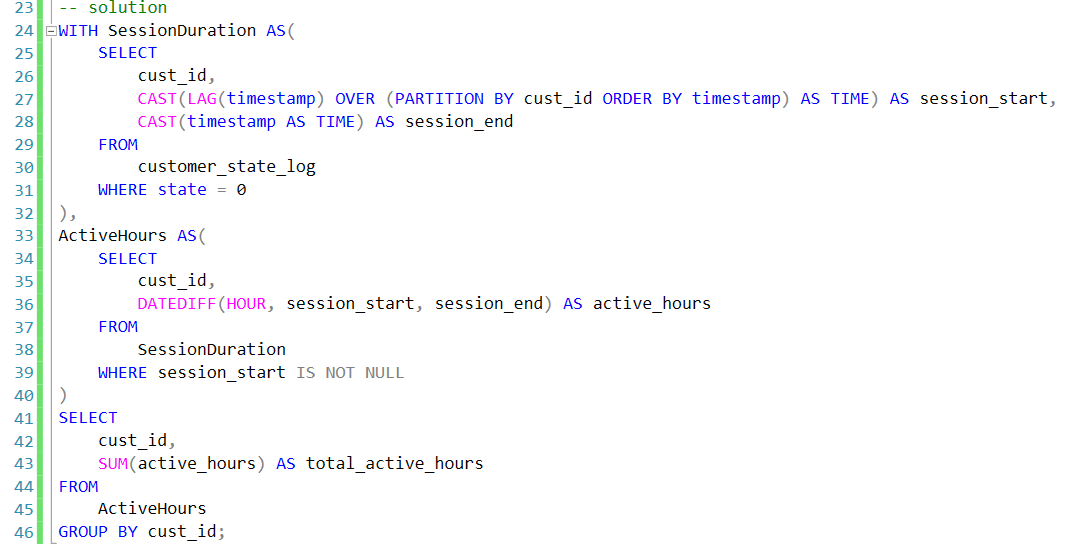


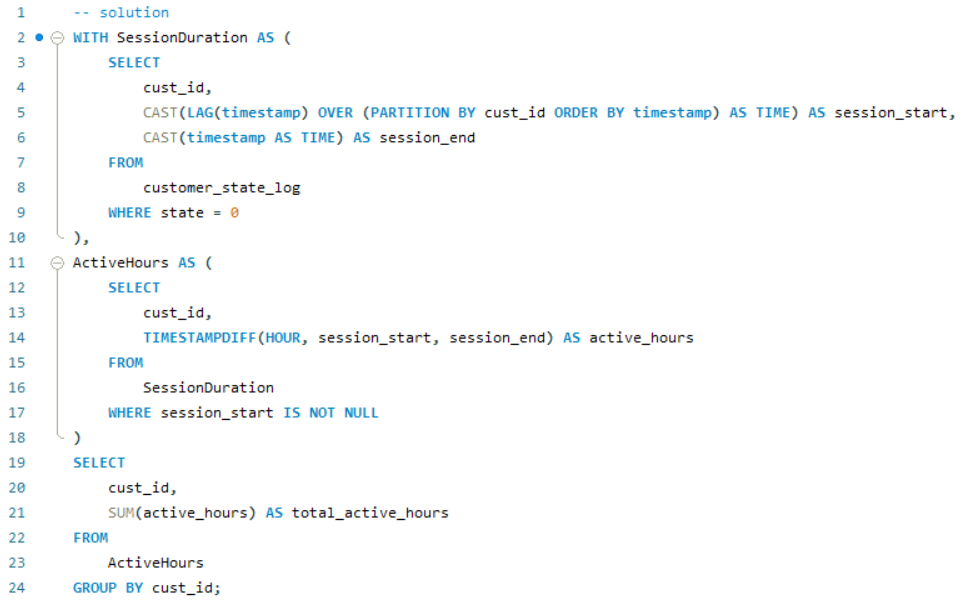
# **Problem 30 (Amazon Hard Level)**

Given the users' sessions logs on a particular day, calculate how many hours each user was active that day. Note: The session starts when state=1 and ends when state=0.

**Explanation:**

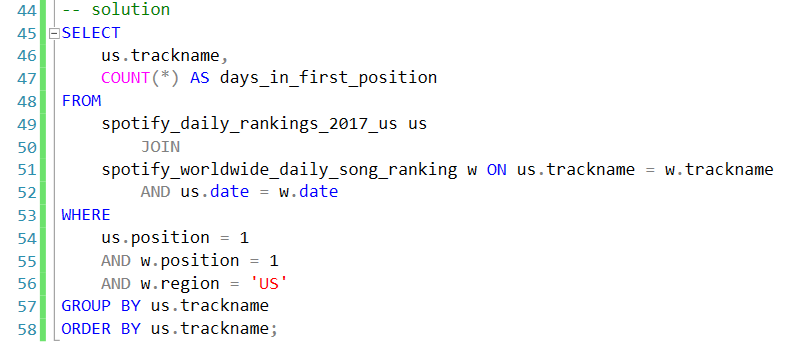
1. SessionDurations CTE: Use the LAG() function to get the timestamp of the previous state for each cust\_id. Filter rows where state=0 because this indicates the end of a session. The session\_start is derived from the LAG() value, and session\_end is the current timestamp.
2. ActiveHours CTE: Calculate the active duration for each session in minutes using DATEDIFF(MINUTE, session\_start, session\_end).
3. Final SELECT: Sum the active minutes for each cust\_id. Divide the total minutes by 60 to convert them to hours.





# **Problem 31 (Spotify Hard Level)**

Find the number of days a US track has stayed in the 1st position for both the US and worldwide rankings on the same day. Output the track name and the number of days in the 1st position. Order your output alphabetically by track name. If the region 'US' appears in dataset, it should be included in the worldwide ranking



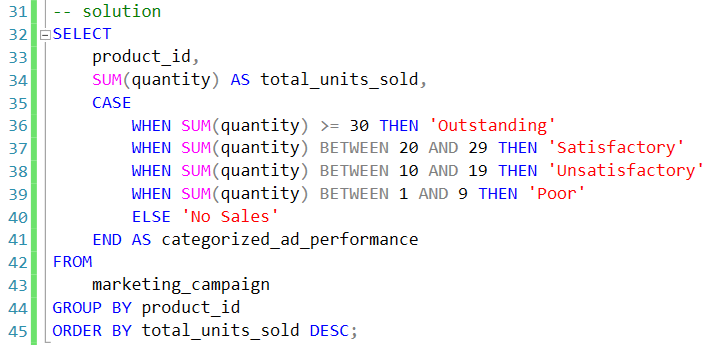
# **Problem 32 (Accenture Medium Level)**

Following a recent advertising campaign, the marketing department wishes to classify its efforts based on the total number of units sold for each product.

You have been tasked with calculating the total number of units sold for each product and categorizing ad performance based on the following criteria for items sold:

* Outstanding: 30+
* Satisfactory: 20 - 29
* Unsatisfactory: 10 - 19
* Poor: 1 - 9

Your output should contain the product ID, total units sold in descending order, and its categorized ad performance.



# **Problem 33 (Google Hard Level)**

Calculate the average session distance traveled by Google Fit users using GPS data for two scenarios:

Considering Earth's curvature (Haversine formula).

Assuming a flat surface.

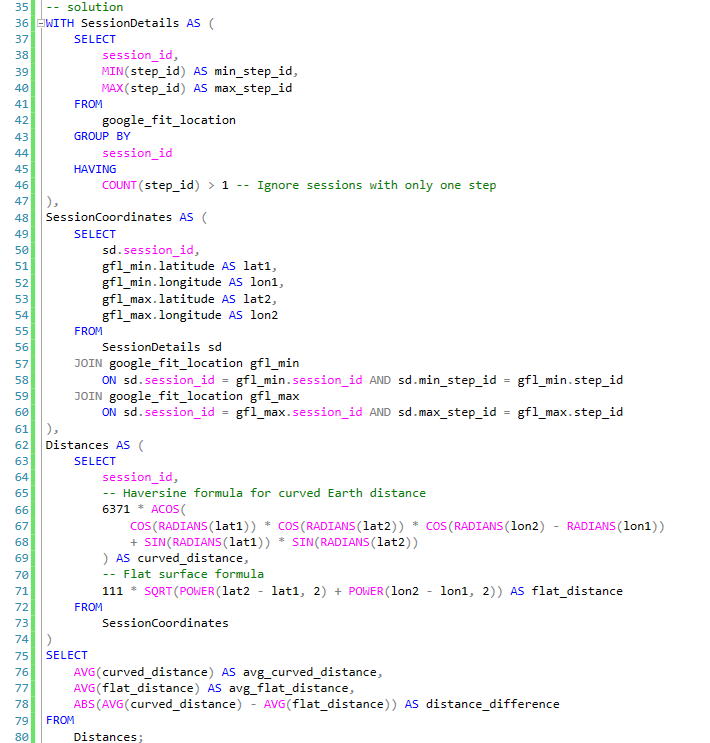
For each session, use the distance between the highest and lowest step IDs, and ignore sessions with only one step.

Calculate and output the average distance for both scenarios and the difference between them.

Formulas:

1.

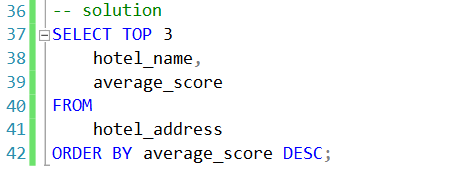
2.

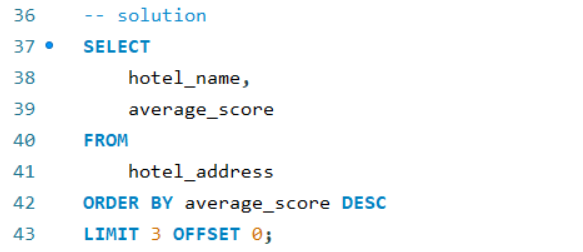


# **Problem 34 (Google, Airbnb, Expedia Medium Level)**

Find the three ten hotels with the highest ratings. Output the hotel name along with the corresponding average score.

Sort records based on the average score in descending order.

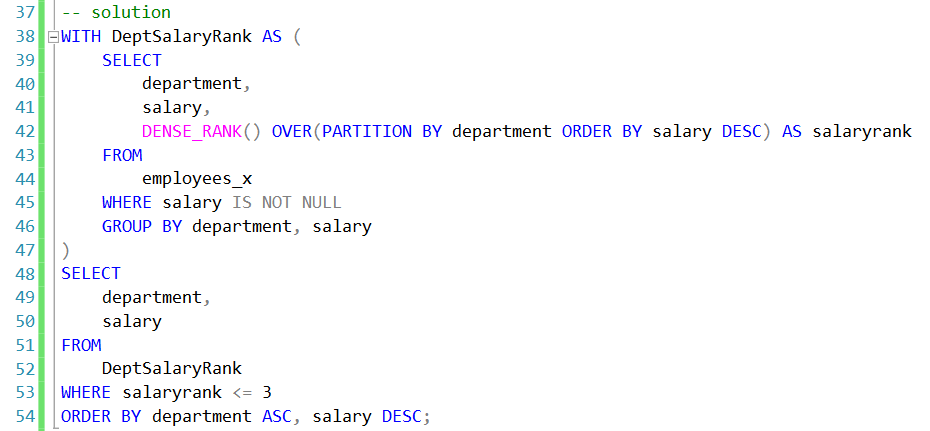




# **Problem 35 (Twitter Hard Level)**

Find the top three distinct salaries for each department. Output the department name and the top 3 distinct salaries by each department.

Order your results alphabetically by department and then by highest salary to lowest.

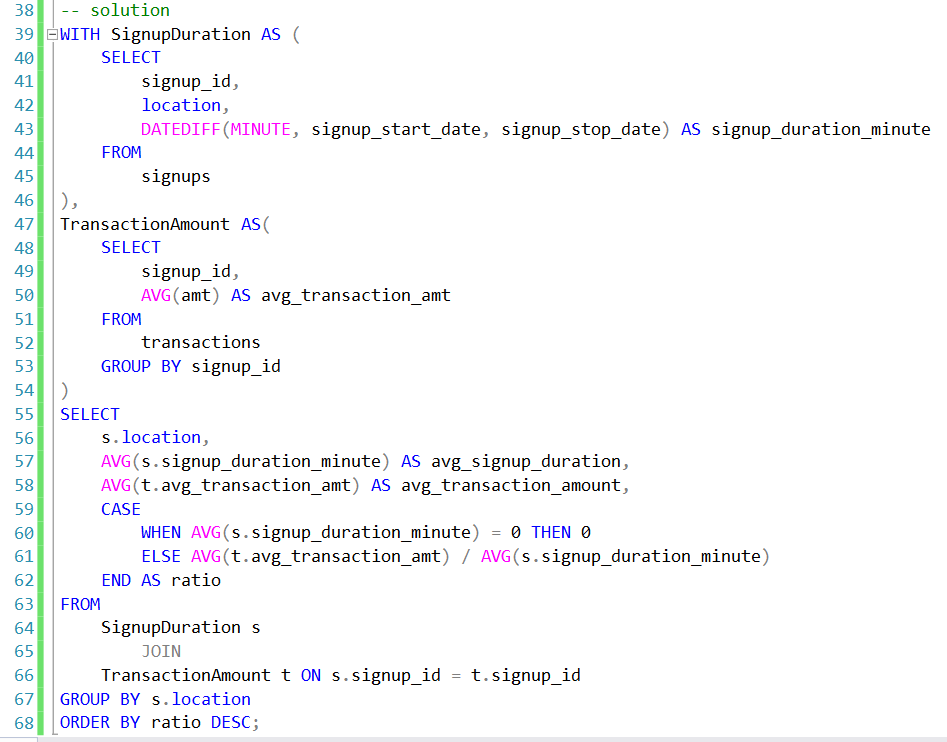


# **Problem 36 (Uber Hard Level)**

Find the most profitable location.

Write a query that calculates the average signup duration and average transaction amount for each location,

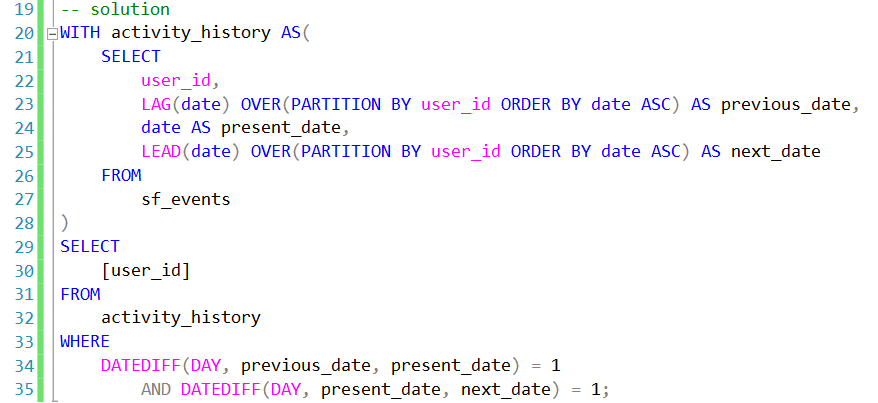
and then compare these two measures together by taking the ratio of the average transaction amount and average duration for each location.



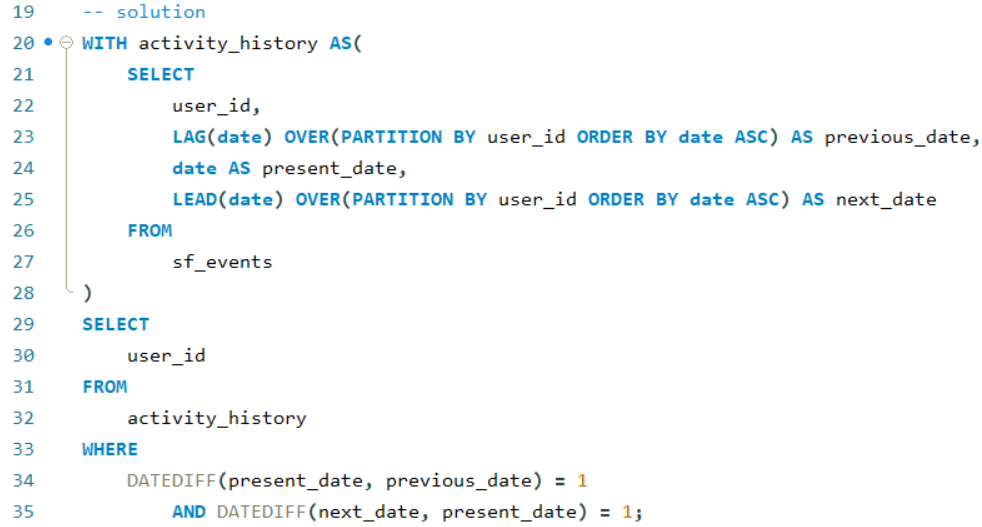
# **Problem 37 (Netflix Hard Level)**

Find all the users who were active for 3 consecutive days or more.

**MSSQL Solution:**



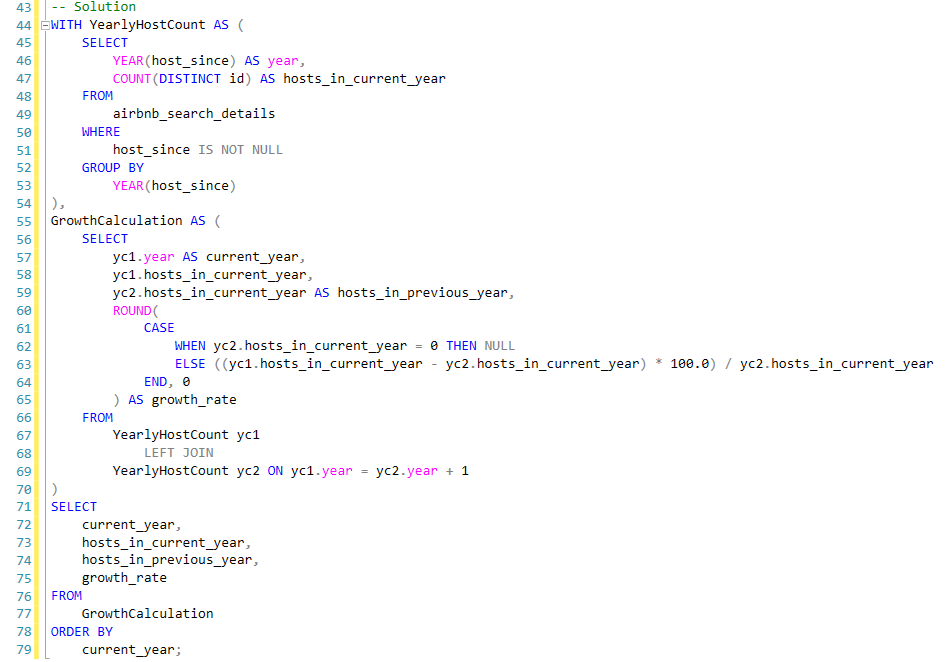
**MySQL Solution:**



# **Problem 38 (Airbnb Hard Level)**

Estimate the growth of Airbnb each year using the number of hosts registered as the growth metric. The rate of growth is calculated by taking

Output the year, number of hosts in the current year, number of hosts in the previous year, and the rate of growth. Round the rate of growth to the nearest percent and order the result in the ascending order based on the year.



**Problem 39 (Walmart (Hard Level)**   
Identify users who started a session and placed an order on the same day. For these users, calculate the total number of orders and the total order value for that day. Your output should include the user, the session date, the total number of orders, and the total order value for that day.)

A screenshot of a computer code

Description automatically generated

**Problem 40 (Walmart (Hard Level)**   
Identify users who started a session and placed an order on the same day. For these users, calculate the total number of orders and the total order value for that day. Your output should include the user, the session date, the total number of orders, and the total order value for that day.)

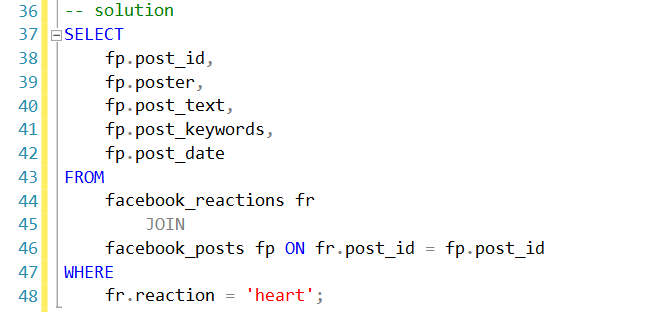
A screenshot of a computer code

Description automatically generated

# **Problem 41 (Meta Easy Level)**

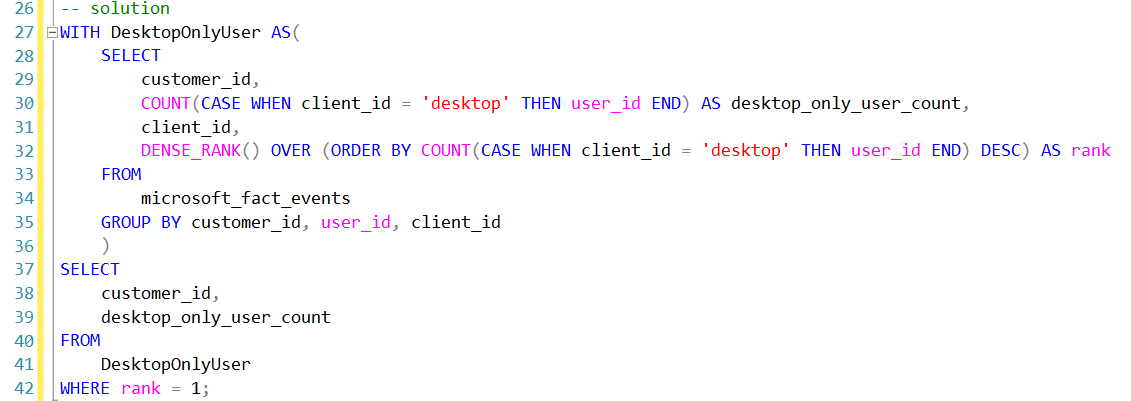
Meta/Facebook (Easy Level)

Find all posts which were reacted to with a heart. For such posts output all columns from facebook\_posts table.



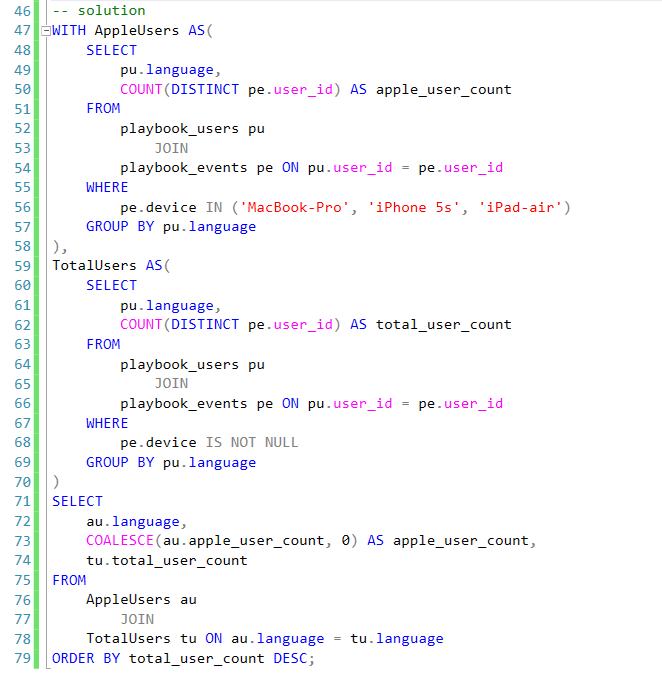
# **Problem 42 (Microsoft Medium Level)**

Write a query that returns the company (customer id column) with highest number of users that use desktop only.



# **Problem 43 (Apple Hard Level)**

Find the number of Apple product users and the number of total users with a device and group the counts by language. Assume Apple products are only MacBook-Pro, iPhone 5s, and iPad-air. Output the language along with the total number of Apple users and users with any device. Order your results based on the number of total users in descending order.



# **Problem 44 (Amazon Hard Level)**

You are given the table with titles of recipes from a cookbook and their page numbers. You are asked to represent how the recipes will be distributed in the book. Produce a table consisting of three columns: left\_page\_number, left\_title and right\_title.

The k-th row (counting from 0), should contain the number and the title of the page with the number 2×k in the first and second columns respectively, and the title of the page with the number 2×k+1 in the third column.

A screenshot of a computer program

Description automatically generated

# **Problem 45 (Visa Medium Level)**

Identify the top 3 areas with the highest customer density.

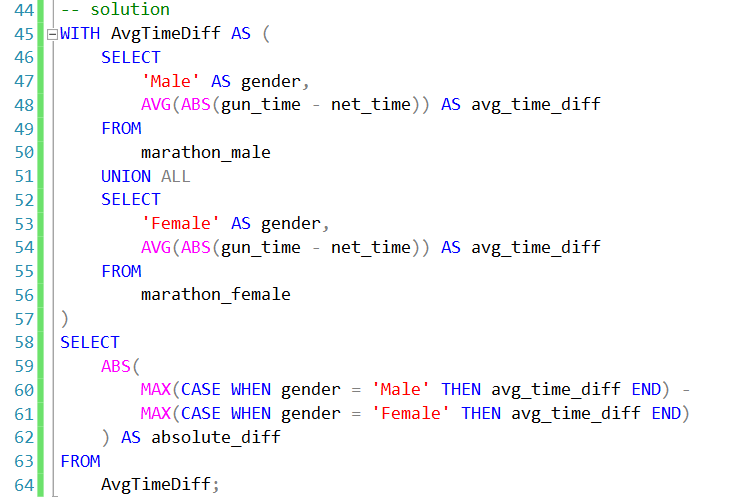
Your output should include the area name and its calculated customer density.

A computer code with black text

Description automatically generated with medium confidence

# **Problem 46 (EY, TCS, Deloitte Medium Level)**

In a marathon, gun time is counted from the moment of the formal start of the race while net time is counted from the moment a runner crosses a starting line. Both variables are in seconds.  
You are asked to check if the interval between the two times is different for male and female runners. First, calculate the average absolute difference between the gun time and net time. Group the results by available genders (male and female). Output the absolute difference between those two values.



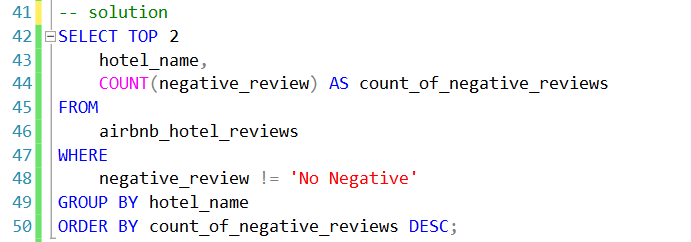
# **Problem 47 (Expedia, Airbnb, Tripadvisor Medium Level)**

Find the top two hotels with the most negative reviews.

Output the hotel name along with the corresponding number of negative reviews.

Negative reviews are all the reviews with text under negative review different than "No Negative".

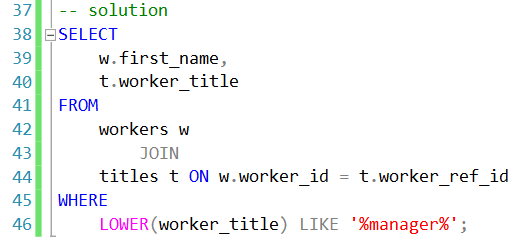
Sort records based on the number of negative reviews in descending order.



# **Problem 48 (Amazon, Doordash, Bosch Medium Level)**

Find all employees who have or had a job title that includes manager.

Output the first name along with the corresponding title.

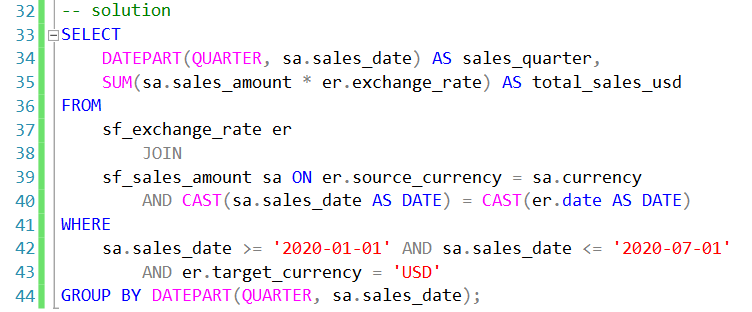


# **Problem 49 (Goldman Sachs Medium Level)**

You work for a multinational company that wants to calculate total sales across all their countries they do business in.

You have 2 tables, one is a record of sales for all countries and currencies the company deals with, and the other holds currency exchange rate information.

Calculate the total sales, per quarter, for the first 2 quarters in 2020, and report the sales in USD currency.



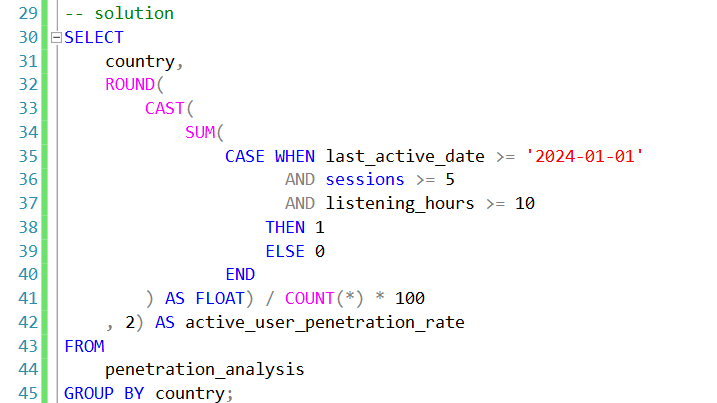
# **Problem 50 (Meta Hard Level)**

Market penetration is an important metric for Spotify's growth in different regions. As part of the analytics team,

calculate the active user penetration rate in specific countries. Active Users must meet these criteria:

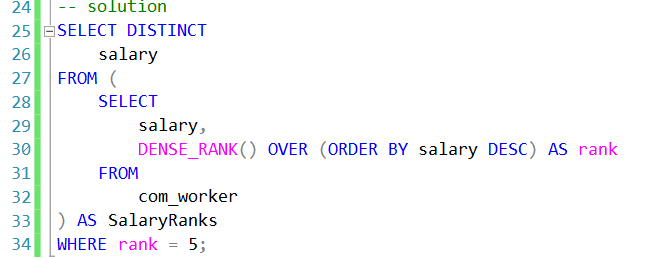
Interacted with Spotify within the last 30 days (last\_active\_date >= 2024-01-01). At least 5 sessions. At least 10 listening hours.

Output: country, active\_user\_penetration\_rate (rounded to 2 decimals).



# **Problem 51 (Amazon Hard Level)**

You have been asked to find the fifth highest salary without using TOP or LIMIT. Note: Duplicate salaries should not be removed.



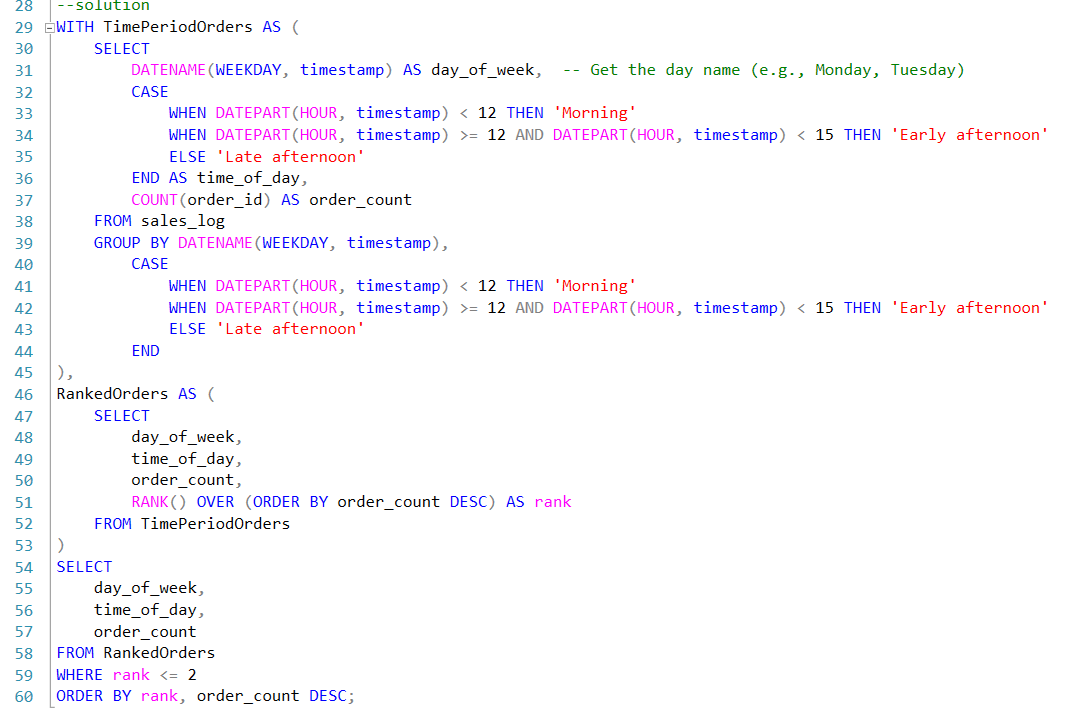
# **Problem 52 (Tesla Hard Level)**

The company you are working for wants to anticipate their staffing needs by identifying their top two busiest times of the week. To find this, each day should be segmented into differents parts using following criteria:

* Morning: Before 12 p.m. (not inclusive)
* Early afternoon: 12 -15 p.m.
* Late afternoon: after 15 p.m. (not inclusive)

Your output should include the day and time of day combination for the two busiest times, i.e. the combinations with the most orders, along with the number of orders (e.g. top two results could be Friday Late afternoon with 12 orders and Sunday Morning with 10 orders).

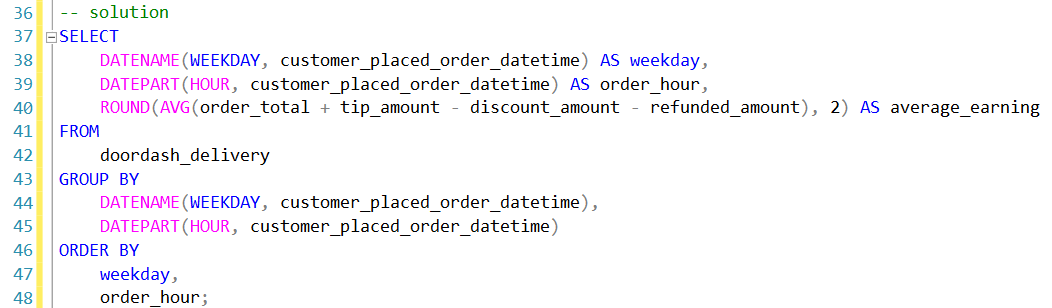
The company has also requested that the day be displayed in text format (i.e. Monday).



# **Problem 53 (DoorDash Medium Level)**

Calculate the average net earnings per order grouped by weekday (in text format, e.g., Monday) and hour from customer\_placed\_order\_datetime.

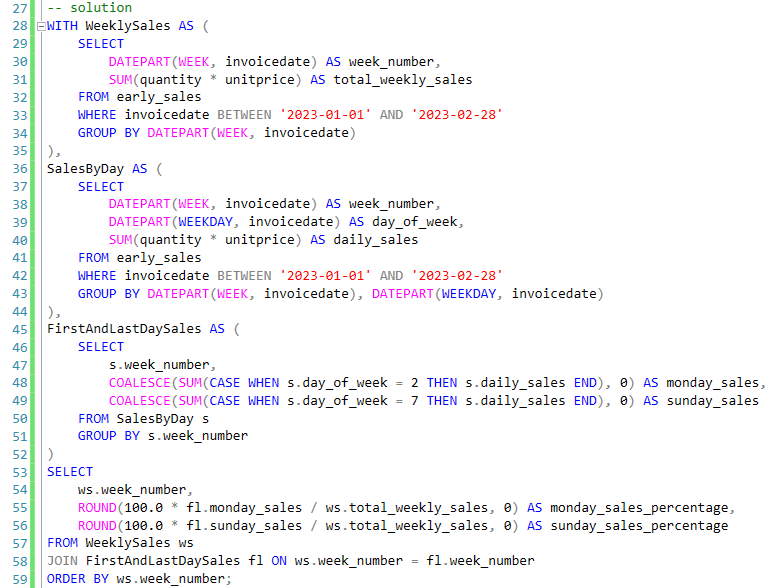
The net earnings are computed as: order\_total + tip\_amount - discount\_amount - refunded\_amount. Round the result to 2 decimals.



# **Problem 54 (Meta Hard level)**

The sales department has given you the sales figures for the first two months of 2023. You've been tasked with determining the percentage of weekly sales on the first and last day of every week. Consider Sunday as last day of week and Monday as first day of week.

In your output, include the week number, percentage sales for the first day of the week, and percentage sales for the last day of the week. Both proportions should be rounded to the nearest whole number.

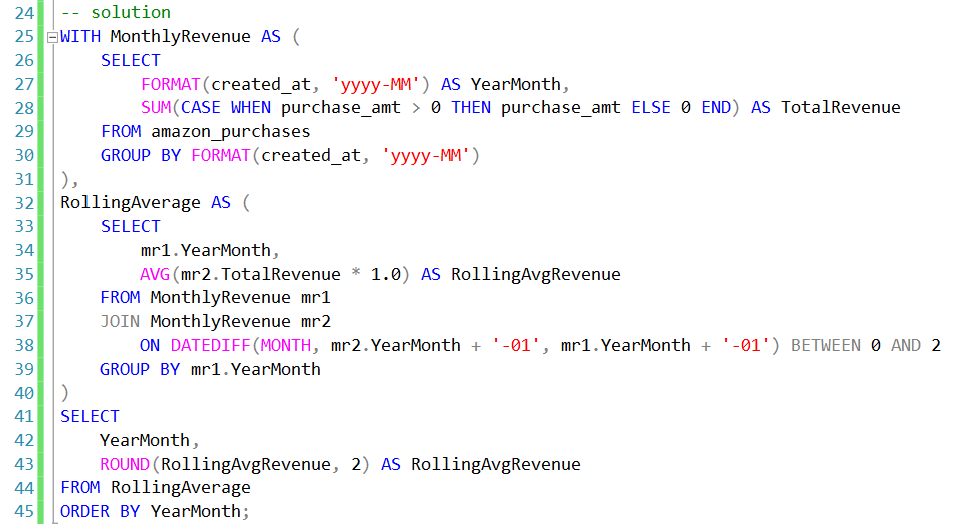


# **Problem 55 (Amazon Hard Level)**

Find the 3-month rolling average of total revenue from purchases given a table with users, their purchase amount, and date purchased. Do not include returns which are represented by negative purchase values. Output the year-month (YYYY-MM) and 3-month rolling average of revenue, sorted from earliest month to latest month.

A 3-month rolling average is defined by calculating the average total revenue from all user purchases for the current month and previous two months.

The first two months will not be a true 3-month rolling average since we are not given data from last year. Assume each month has at least one purchase.



# **Problem 56 (ESPN Hard Level)**

Find the quarterback who threw the longest throw in 2016. Output the quarterback name along with their corresponding longest throw.

The 'lg' column contains the longest completion by the quarterback.

