Analytics SQL Questions

Here are some common data challenges that you as an Analyst in Digital Alchemy will encounter every day.

Any SQL Syntax can be used for the test. The objective is not to validate accuracy of syntax, but rather logical thought process in writing SQL to query relational tables.

**Tables**

**Member**

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute name** | **Type** |
| 1 | MEMBER\_ID | Number |
| 2 | MEMBER\_NAME | String |
| 3 | JOIN\_DATE | DATE |
| 4 | REGISTERED\_STORE\_ID | Number |

**Store**

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute name** | **Type** |
| 1 | STORE\_ID | Number |
| 2 | STORE\_NAME | String |
| 3 | STATE | String |
| 4 | STORE\_ONBOARD\_DATE | Date |

**Transation\_basket**

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute name** | **Type** |
| 1 | TXN\_ID | Number |
| 2 | MEMBER\_ID | Number |
| 3 | TXN\_DATE | String |
| 4 | STORE\_ID | Store |
| 5 | TXN\_AMOUNT | Number |

**Question 1:**

Write a SQL to select a list of member ids who joined in October 2020 and member\_name starts with ‘M’

SELECT MEMBER\_ID FROM Member WHERE JOIN\_DATE >= '2020-10-01' AND JOIN\_DATE <= '2020-10-31' AND MEMBER\_NAME LIKE 'M%';

**Question 2:**

Write a SQL to summarize the following stats by members’ registered store

1. how many new members who joined in October 2020

SELECT COUNT(\*) AS NewMembersCount FROM Member WHERE JOIN\_DATE >= '2020-10-01' AND JOIN\_DATE <= '2020-10-31';

1. how many members who made at least 1 transactions within 30 days after joining

SELECT COUNT(DISTINCT m.MEMBER\_ID) AS alteast\_1\_txn FROM Member m JOIN Transaction\_basket t ON m.MEMBER\_ID = t.MEMBER\_ID WHERE t.TXN\_DATE <= DATE\_ADD(m.JOIN\_DATE, INTERVAL 30 DAY);

3) average spending per transaction by store

SELECT t.STORE\_ID, s.STORE\_NAME, AVG(t.TXN\_AMOUNT) AS avg\_spending FROM Transaction\_basket t JOIN Store s ON t.STORE\_ID = s.STORE\_ID GROUP BY t.STORE\_ID, s.STORE\_NAME;

**Question 3:**

3.1 Write SQL to find the list of BKK store names that got more than 10,000 transactions in October and order by the number of transactions from highest to lowest

SELECT s.STORE\_NAME, COUNT(\*) AS txn\_count FROM

Store s JOIN Transaction\_basket t ON s.STORE\_ID = t.STORE\_ID

WHERE s.STATE = 'BKK' AND t.TXN\_DATE >= '2020-10-01' AND t.TXN\_DATE <= '2020-10-31'

GROUP BY s.STORE\_NAME HAVING COUNT(\*) > 10000 ORDER BY txn\_count DESC;

3.2 Find the distribution of monthly frequency of customers shopping in BKK in 2020. What does the distribution look like? What would be potential issues for this output?

Example output of 3.2:

|  |  |
| --- | --- |
| Avg number of transactions (monthly) | Number of customer |
| 1 | 50 |
| 2 | 30 |
| 3 | 20 |
| .  .  . | .  .  . |

SELECT Mon\_txn\_cnt AS "Avg number of transactions (monthly)",

COUNT(CustomerID) AS "Number of customers"

FROM (

SELECT MONTH(TXN\_DATE) AS Txn\_month, MEMBER\_ID AS CustomerID,

COUNT(\*) AS Mon\_txn\_cnt FROM Transaction\_basket t

JOIN Store s ON t.STORE\_ID = s.STORE\_ID WHERE s.STATE = 'BKK'

AND YEAR(TXN\_DATE) = 2020

GROUP BY Txn\_month, CustomerID) AS Mon\_txn\_cnts

GROUP BY Mon\_txn\_cnt

ORDER BY Mon\_txn\_cnt;

Distribution looks like No. of customers are decreasing and it may lead to less revenue.

**Question 4:**

We would like to send a campaign to the top 5 customers in each store, ranking by their total spending in October 2020.

WITH CustomerSpending AS (

SELECT t.MEMBER\_ID, t.STORE\_ID, SUM(t.TXN\_AMOUNT) AS TotalSpending

FROM Transaction\_basket t WHERE t.TXN\_DATE >= '2020-10-01' AND t.TXN\_DATE <= '2020-10-31'

GROUP BY t.MEMBER\_ID, t.STORE\_ID),

RankedCustomers AS (

SELECT \*, ROW\_NUMBER() OVER (PARTITION BY STORE\_ID ORDER BY TotalSpending DESC) AS Rank FROM CustomerSpending

)

SELECT rc.STORE\_ID, s.STORE\_NAME, rc.MEMBER\_ID, m.MEMBER\_NAME, rc.TotalSpending

FROM RankedCustomers rc JOIN Member m ON rc.MEMBER\_ID = m.MEMBER\_ID

JOIN Store s ON rc.STORE\_ID = s.STORE\_ID WHERE rc.Rank <= 5;

**Question 5:**

Write SQL to find the top 5 stores that total transaction value in 2020 increase the most compared to total transaction value in 2019

WITH TransactionTotals AS ( SELECT

STORE\_ID, YEAR(TXN\_DATE) AS TransactionYear, SUM(TXN\_AMOUNT) AS TotalTransactionAmount FROM Transaction\_basket WHERE YEAR(TXN\_DATE) IN (2019, 2020) GROUP BY STORE\_ID, YEAR(TXN\_DATE) ),

YearlyComparison AS ( SELECT STORE\_ID, SUM(CASE WHEN TransactionYear = 2019 THEN TotalTransactionAmount ELSE 0 END) AS TotalTransactionAmount\_2019, SUM(CASE WHEN TransactionYear = 2020 THEN TotalTransactionAmount ELSE 0 END) AS TotalTransactionAmount\_2020 FROM TransactionTotals GROUP BY STORE\_ID )

SELECT yt.STORE\_ID, s.STORE\_NAME, yt.TotalTransactionAmount\_2020 - yt.TotalTransactionAmount\_2019 AS TransactionIncrease FROM YearlyComparison yt JOIN Store s ON yt.STORE\_ID = s.STORE\_ID ORDER BY TransactionIncrease DESC LIMIT 5;

**Question 6:**

6.1 Write SQL to find the average value of the biggest transaction happening in 2019 in each store by state

WITH MaxTransactionPerStore AS (

SELECT t.STORE\_ID, s.STATE, MAX(t.TXN\_AMOUNT) AS MaxTransactionAmount

FROM Transaction\_basket t

JOIN Store s ON t.STORE\_ID = s.STORE\_ID WHERE YEAR(t.TXN\_DATE) = 2019

GROUP BY t.STORE\_ID, s.STATE),

AverageMaxTransactionPerState AS (

SELECT STATE, AVG(MaxTransactionAmount) AS AverageMaxTransaction

FROM MaxTransactionPerStore GROUP BY STATE )

SELECT STATE, AverageMaxTransaction FROM AverageMaxTransactionPerState;

6.2 Write SQL to find the average value of the top 3 biggest transactions happening in 2019 in each store by state

WITH Top3TransactionsPerStore AS (

SELECT t.STORE\_ID, s.STATE, t.TXN\_AMOUNT,

ROW\_NUMBER() OVER(PARTITION BY t.STORE\_ID, s.STATE ORDER BY t.TXN\_AMOUNT DESC) AS TransactionRank FROM Transaction\_basket t JOIN Store s ON t.STORE\_ID = s.STORE\_ID

WHERE YEAR(t.TXN\_DATE) = 2019 ),

Top3TransactionsPerState AS (

SELECT STORE\_ID, STATE, TXN\_AMOUNT FROM Top3TransactionsPerStore

WHERE TransactionRank <= 3)

SELECT STATE, AVG(TXN\_AMOUNT) AS AverageTop3TransactionValue

FROM Top3TransactionsPerState GROUP BY STATE;

**Question 7:**

Use the following 2 tables for this question:

PRODUCT

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute name** | **Type** |
| 1 | PRODUCT\_ID | Number |
| 2 | PRODUCT\_NAME | String |
| 3 | PRODUCT\_CATEGORY | String |

UPDATED\_CATEGORY

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute name** | **Type** |
| 1 | PRODUCT\_ID | Number |
| 2 | UPDATED\_CATEGORY | String |

There seem to be miscategorisations in the product category table. Create a new product table that contains PRODUCT\_ID, PRODUCT\_NAME and the correct categories using Updated Categories table and the logic below:

1) if the product category contains the word ‘DRINKS' or 'MILK' (that is not 'FLAVOURED MILK'), then categorise it as 'BEVERAGE'

2) if the product category is ‘FLAVOURED MILK', then categorise it as 'FRESH\_FOOD'

3) if the product category is 'DOG FOOD', then categorise it as 'HOUSEHOLD\_PERSONALCARE'

4) Otherwise, use the updated\_category field from the Updated Categories table.

CREATE TABLE NewProduct AS

SELECT

p.PRODUCT\_ID, p.PRODUCT\_NAME,

CASE

WHEN UPPER(p.PRODUCT\_CATEGORY) LIKE '%DRINKS%' OR UPPER(p.PRODUCT\_CATEGORY) LIKE '%MILK%'

THEN 'BEVERAGE'

WHEN UPPER(p.PRODUCT\_CATEGORY) = 'FLAVOURED MILK'

THEN 'FRESH\_FOOD'

WHEN UPPER(p.PRODUCT\_CATEGORY) = 'DOG FOOD'

THEN 'HOUSEHOLD\_PERSONALCARE'

ELSE uc.UPDATED\_CATEGORY

END AS CORRECT\_CATEGORY

FROM PRODUCT p

JOIN UPDATED\_CATEGORY uc ON p.PRODUCT\_ID = uc.PRODUCT\_ID;

Analytics Case Questions

**Question 1:**

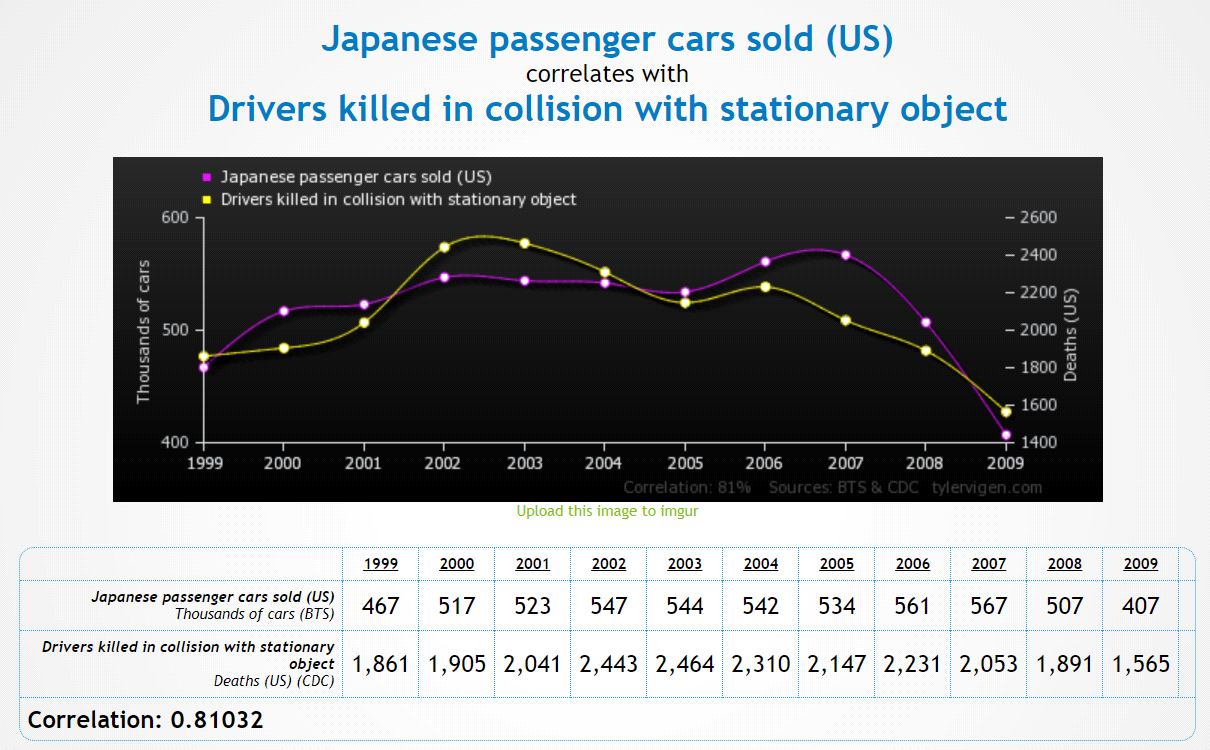
Our client is an online payment platform that facilitates payments between individuals and businesses. Your project is to drive the total payment volumes through the platform. What would be your strategy and how would you present it to the management? Which method/data would you use to support your argument? Please prepare slides to present to the management for approval of your project. (You can use data from any source or make an assumption on the number needed to support your model/calculation)

(You can put your own assumptions where needed, but kindly state clearly and **please create a slide of presentation**)

**Question 2:**

How should you conclude from the chart below? Which data would you present to proof/disproof the statement?

“Japanese car company intentionally adjust the car, which leads to people’s death”



Credit: <http://tylervigen.com/>

**Question 3:**

Our client is a clothing company in Thailand that has 3 product lines. Each product has a different margin. At the end of 2020, the head of marketing found that the average margin per unit is 11% lower than the forecasted number. How would you help the marketing team to explain the scenario to their management team?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **2020 Forecasted** | | **2020 Actual** | |
| **Product** | **Margin per unit** | **# unit sold** | **Total margin** | **# unit sold** | **Total margin** |
| T-shirt | 5 | 30 | 150 | 48 | 240 |
| Dress | 30 | 50 | 1,500 | 51 | 1,530 |
| Shirt | 12 | 20 | 240 | 22 | 264 |
| Average margin per unit |  | 18.9 | | 16.8 | |

**Question 4:**

Please make a slide to present the data below

**Share of revenue by product**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **2018** | **2019** |
| Total revenue (Millions USD) | | 1,000 | 1,200 |
| Food | Processed food | 15% | 16% |
| Organic food | 3% | 10% |
| Canned food | 20% | 14% |
| Dry food | 39% | 37% |
| Drink | Milk | 4% | 6% |
| Soda | 4% | 1% |
| Alcoholic drink | 6% | 3% |
| Juice | 8% | 9% |
| Fresh juice | 1% | 4% |