Order (Case Study Questions):

The following case study questions require some data cleaning steps before we start to unpack key business questions in more depth. In a single query, perform the following operations and generate a new table in the data\_mart schema named clean\_weekly\_sales:

1. Convert the week date to a DATE format.

UPDATE test

SET date = STR\_TO\_DATE(date, '%d/%m/%y');

1. Add a week\_number as the second column for each week\_date value, for example any value from the 1st of January to 7th of January will be 1, 8th to 14th will be 2 etc.

-- Add the week\_number column

ALTER TABLE weekly\_sales

ADD week\_number INT;

-- Update the week\_number column

UPDATE weekly\_sales

SET week\_number = ceiling((DAY(week\_date)) / 7);

-- Update column position into 2nd column position

ALTER TABLE weekly\_sales

MODIFY COLUMN week\_number INT AFTER week\_date;

1. Add a month\_number with the calendar month for each week\_date value as the 3rd column.

-- Add the month\_number column

ALTER TABLE weekly\_sales

ADD month\_number INT;

-- Update column position into 3rd column position

ALTER TABLE weekly\_sales

MODIFY COLUMN month\_number INT AFTER week\_number;

-- Update the month\_number column

UPDATE weekly\_sales

SET month\_number = MONTH(week\_date);

1. Add a calendar\_year column as the 4th column containing either 2018, 2019 or 2020 values.

-- Add the calendar\_year column

ALTER TABLE weekly\_sales

ADD calendar\_year INT;

-- Update the calendar\_year column

UPDATE weekly\_sales

SET calendar\_year = YEAR(week\_date);

-- Update column position into 4th column position

ALTER TABLE weekly\_sales

MODIFY COLUMN calendar\_year INT AFTER month\_number;

1. Add a new column called age\_band after the original segment column using the following mapping on the number inside the segment value:

A black background with white text

Description automatically generated

ALTER TABLE weekly\_sales

ADD COLUMN age\_band VARCHAR(255) AFTER segment;

UPDATE weekly\_sales

SET age\_band =

CASE

WHEN segment LIKE '%1%' THEN 'Young Adults'

WHEN segment LIKE '%2%' THEN 'Middle Aged'

WHEN segment LIKE '%3%' OR segment LIKE '%4%' THEN 'Retirees'

END

WHERE segment IS NOT NULL;

1. Add a new demographic column using the following mapping for the first letter in the segment values:

A black screen with white text

Description automatically generated

ALTER TABLE weekly\_sales

ADD COLUMN demographoc VARCHAR(255) AFTER segment;

update weekly\_sales

SET demographic =

CASE

WHEN segment LIKE '%C%' THEN 'Couples'

WHEN segment LIKE '%F%' THEN 'Families'

END

WHERE segment IS NOT NULL;

1. Ensure all null string values with an "unknown" string value in the original segment column as well as the new age\_band and demographic columns

UPDATE weekly\_sales

SET

segment = COALESCE(segment, 'unknown'),

age\_band = COALESCE(age\_band, 'unknown'),

demographic = COALESCE(demographic, 'unknown');

1. Generate a new avg\_transaction column as the sales value divided by transactions rounded to 2 decimal places for each record

ALTER TABLE weekly\_sales

ADD avg\_transaction DECIMAL(10, 2) after sales;

UPDATE weekly\_sales

SET avg\_transaction = ROUND(sales / transactions, 2);

Segment 1 Data Exploration:

1. What day of the week is used for each week\_date value?

select count(week\_date) as jumlah, dayname(week\_date) as hari from weekly\_sales group by hari;

result : only monday

1. What range of week numbers are missing from the dataset?

select count(week\_number) as jumlah, week\_number from weekly\_sales group by week\_number;

None

1. How many total transactions were there for each year in the dataset?

select sum(transactions) as total\_transaksi , year(week\_date) as tahun from weekly\_sales group by tahun;

'375813651', '2020'

'365639285', '2019'

'346406460', '2018'

1. What is the total sales for each region for each month?

select distinct region, monthname(week\_date) as nama\_bulan, count(sales) as jumlah from weekly\_sales group by nama\_bulan,region;

region month total\_sales

|  |  |  |
| --- | --- | --- |
| ASIA | August | 442 |
| USA | August | 442 |
| EUROPE | August | 439 |
| AFRICA | August | 442 |
| CANADA | August | 442 |
| OCEANIA | August | 442 |
| SOUTH AMERICA | August | 440 |
| AFRICA | July | 476 |
| CANADA | July | 476 |
| USA | July | 476 |
| EUROPE | July | 475 |
| OCEANIA | July | 476 |
| SOUTH AMERICA | July | 475 |
| ASIA | July | 476 |
| OCEANIA | June | 442 |
| USA | June | 442 |
| SOUTH AMERICA | June | 442 |
| EUROPE | June | 441 |
| ASIA | June | 442 |
| CANADA | June | 442 |
| AFRICA | June | 442 |
| EUROPE | May | 406 |
| USA | May | 408 |
| SOUTH AMERICA | May | 407 |
| OCEANIA | May | 408 |
| ASIA | May | 408 |
| AFRICA | May | 408 |
| CANADA | May | 408 |
| CANADA | April | 476 |
| SOUTH AMERICA | April | 473 |
| EUROPE | April | 474 |
| ASIA | April | 476 |
| OCEANIA | April | 476 |
| AFRICA | April | 476 |
| USA | April | 476 |
| AFRICA | March | 136 |
| OCEANIA | March | 136 |
| EUROPE | March | 135 |
| SOUTH AMERICA | March | 136 |
| USA | March | 136 |
| CANADA | March | 136 |
| ASIA | March | 136 |
| OCEANIA | September | 68 |
| EUROPE | September | 66 |
| ASIA | September | 68 |
| SOUTH AMERICA | September | 68 |
| USA | September | 68 |
| AFRICA | September | 68 |
| CANADA | September | 68 |

1. What is the total count of transactions for each platform

select sum(transactions) as jumlah\_transaksi, platform from weekly\_sales group by platform;

Transactions Platform

|  |  |
| --- | --- |
| 1081934227 | Retail |
| 5925169 | Shopify |

1. Which age\_band and demographic values contribute the most to Retail sales?

select demographic, age\_band, sum(sales) from weekly\_sales where platform = 'Retail' group by demographic,age\_band order by sum(sales) asc limit 1;

demographic age\_band sales

|  |  |  |
| --- | --- | --- |
| Families | Young Adults | 1770889293 |

1. Can we use the avg\_transaction column to find the average transaction size for each year for Retail vs Shopify? If not - how would you calculate it instead?

select avg(avg\_transaction) as rata\_rata\_transaksi, platform from weekly\_sales group by platform;

avg transaction platfom

|  |  |
| --- | --- |
| 41.838224 | Retail |
| 180.226179 | Shopify |

Segment 2 Before & After Analysis:

This technique is usually used when we inspect an important event and want to inspect the impact before and after a certain point in time.

Taking the week\_date value of 2020-06-15 as the baseline week where the Data Mart sustainable packaging changes came into effect. We would include all week\_date values for 2020-06-15 as the start of the period **after** the change and the previous week\_date values would be **before.**

Using this analysis approach - answer the following questions:

1. What is the total sales for the 4 weeks before and after 2020-06-15? What is the growth or reduction rate in actual values and percentage of sales?

SELECT DISTINCT week\_number

FROM clean\_weekly\_sales

WHERE week\_date = '2020-06-15'

AND calendar\_year = '2020';

Result:

WITH packaging\_sales AS (

SELECT

week\_date,

week\_number,

SUM(sales) AS total\_sales

FROM clean\_weekly\_sales

WHERE (week\_number BETWEEN 21 AND 28)

AND (calendar\_year = 2020)

GROUP BY week\_date, week\_number

)

, before\_after\_changes AS (

SELECT

SUM(CASE

WHEN week\_number BETWEEN 21 AND 24 THEN total\_sales END) AS before\_packaging\_sales,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 28 THEN total\_sales END) AS after\_packaging\_sales

FROM packaging\_sales

)

SELECT

after\_packaging\_sales - before\_packaging\_sales AS sales\_variance,

ROUND(100 \*

(after\_packaging\_sales - before\_packaging\_sales)

/ before\_packaging\_sales,2) AS variance\_percentage

FROM before\_after\_changes;

1. What about the entire 12 weeks before and after?

WITH packaging\_sales AS (

SELECT

week\_date,

week\_number,

SUM(sales) AS total\_sales

FROM clean\_weekly\_sales

WHERE (week\_number BETWEEN 13 AND 37)

AND (calendar\_year = 2020)

GROUP BY week\_date, week\_number

)

, before\_after\_changes AS (

SELECT

SUM(CASE

WHEN week\_number BETWEEN 13 AND 24 THEN total\_sales END) AS before\_packaging\_sales,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 37 THEN total\_sales END) AS after\_packaging\_sales

FROM packaging\_sales

)

SELECT

after\_packaging\_sales - before\_packaging\_sales AS sales\_variance,

ROUND(100 \*

(after\_packaging\_sales - before\_packaging\_sales) / before\_packaging\_sales,2) AS variance\_percentage

FROM before\_after\_changes;

1. How do the sale metrics for these 2 periods before and after compare with the previous years in 2018 and 2019?

WITH changes AS (

SELECT

calendar\_year,

week\_number,

SUM(sales) AS total\_sales

FROM clean\_weekly\_sales

WHERE week\_number BETWEEN 21 AND 28

GROUP BY calendar\_year, week\_number

)

, before\_after\_changes AS (

SELECT

calendar\_year,

SUM(CASE

WHEN week\_number BETWEEN 13 AND 24 THEN total\_sales END) AS before\_packaging\_sales,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 28 THEN total\_sales END) AS after\_packaging\_sales

FROM changes

GROUP BY calendar\_year

)

SELECT

calendar\_year,

after\_packaging\_sales - before\_packaging\_sales AS sales\_variance,

ROUND(100 \*

(after\_packaging\_sales - before\_packaging\_sales)

/ before\_packaging\_sales,2) AS variance\_percentage

FROM before\_after\_changes;