Case Study #5 - Data Mart

A. Data Cleansing Steps

```
CREATE VIEW clean_weekly_sales AS
(SELECT
    STR_TO_DATE(week_date,'%d/%m/%y') AS week_date,
    WEEKOFYEAR(STR_TO_DATE(week_date,'%d/%m/%y')) as week_numbe
    MONTH(STR_TO_DATE(week_date,'%d/%m/%y')) as month_number,
    YEAR(STR_TO_DATE(week_date,'%d/%m/%y')) as calendar_year,
    region,
    platform,
    segment,
    (CASE
        WHEN segment LIKE '%1' THEN 'Young Adults'
        WHEN segment LIKE '%2' THEN 'Middle Aged'
        WHEN segment LIKE '%3' OR segment LIKE '%4' THEN 'Retire
        ELSE 'Unknown' END) as age_band,
    (CASE
        WHEN segment LIKE 'C%' THEN 'Couples'
        WHEN segment LIKE 'F%' THEN 'Families'
        ELSE 'Unknown' END) as demographic,
    transactions,
    ROUND(sales/transactions,2) AS avg_transaction,
    sales
FROM weekly_sales)
```

B. Data Exploration

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

```
SELECT
    DISTINCT DAYOFWEEK(week_date) AS week_day
```

```
FROM clean_weekly_sales
```

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

```
WITH RECURSIVE list_52_week AS (
    SELECT 1 AS week_number
    UNION ALL
    SELECT week_number + 1
    FROM list_53_week
    WHERE week_number < 52
)
SELECT COUNT(DISTINCT l.week_number)
FROM list_53_week AS l
LEFT JOIN clean_weekly_sales AS cws
  ON l.week_number = cws.week_number
WHERE cws.week_number IS NULL;
```

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

```
SELECT
    calendar_year,
    SUM(transactions) AS total_transaction
FROM clean_weekly_sales
GROUP BY calendar_year
```

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

```
SELECT
    region,
    month_number,
    SUM(sales) AS total_sales
FROM clean_weekly_sales
GROUP BY region, month_number
```

5. What is the total count of transactions for each platform?

```
SELECT
    platform,
    SUM(transactions) AS total_transactions
FROM clean_weekly_sales
GROUP BY platform
```

6. What is the percentage of sales for Retail vs Shopify for each month?

```
WITH total_sales_each_month AS (
    SELECT
        calendar_year,
        month_number,
        SUM(sales) AS total_sales
    FROM clean_weekly_sales
    GROUP BY calendar_year,month_number
    ORDER BY calendar_year,month_number
)
SELECT
    cws.calendar_year,
    cws.month_number,
    ROUND(SUM(
        IF(platform = 'Retail',sales,0)
    )/total_sales * 100,2) as percentage_retail,
    100 - ROUND(SUM(
        IF(platform = 'Retail',sales,0)
    )/total_sales * 100,2) as percentage_shopee
FROM clean_weekly_sales AS cws
JOIN total_sales_each_month AS tsem ON cws.calendar_year = tsem
GROUP BY cws.calendar_year, cws.month_number
ORDER BY calendar_year,month_number;
```

7. What is the percentage of sales by demographic for each year in the dataset?

```
WITH total_sales_each_year AS (
    SELECT
```

```
        calendar_year,
        SUM(sales) AS total_sales
    FROM clean_weekly_sales
    GROUP BY calendar_year
    ORDER BY calendar_year
)
SELECT
    cws.calendar_year,
    ROUND(SUM(
        IF(demographic = 'Couples',sales,0)
    )/total_sales * 100,2) as percentage_couples,
    ROUND(SUM(
        IF(demographic = 'Families',sales,0)
    )/total_sales * 100,2) as percentage_families,
    ROUND(SUM(
        IF(demographic = 'Unknown',sales,0)
    )/total_sales * 100,2) as percentage_unknow
FROM clean_weekly_sales AS cws
JOIN total_sales_each_year AS tsey ON cws.calendar_year = tsey.c
GROUP BY cws.calendar_year
ORDER BY calendar_year
```

8. Which age_band and demographic values contribute the most to Retail sales?

```
SELECT
    age_band,
    demographic,
    SUM(sales) AS total_sales,
    ROUND(SUM(sales)/SUM(SUM(sales)) OVER() * 100,1) AS percenta
FROM clean_weekly_sales
WHERE platform = 'Retail'
GROUP BY age_band,demographic
ORDER BY total_sales DESC
```

9. 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
    calendar_year,
    platform,
   ROUND(AVG(avg_transaction)) as avg_transaction

FROM clean_weekly_sales
GROUP BY calendar_year, platform
ORDER BY calendar_year
```

C. Before & After Analysis

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
    total_sales_after - total_sales_before as variance_sales,
    ROUND((total_sales_after - total_sales_before) / total_sales
FROM
    (SELECT
        SUM(CASE
            WHEN week_number BETWEEN 21 AND 24 THEN sales END) A
        SUM(CASE
            WHEN week_number BETWEEN 25 AND 28 THEN sales END) A
    FROM clean_weekly_sales
    WHERE calendar_year = 2020) as x
```

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

```
SELECT
    total_sales_after - total_sales_before as variance_sales,
    ROUND((total_sales_after - total_sales_before) / total_sales
FROM
    (SELECT
        SUM(CASE
```

```
            WHEN week_number BETWEEN 13 AND 24 THEN sales END) A
        SUM(CASE
            WHEN week_number BETWEEN 25 AND 36 THEN sales END) A
    FROM clean_weekly_sales
    WHERE calendar_year = 2020) as x
```

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

- **Part 1: How do the sale metrics for 4 weeks before and after compare with the previous years in 2018 and 2019?**

```
SELECT
    calendar_year,
    total_sales_after - total_sales_before as variance_sales,
    ROUND((total_sales_after - total_sales_before) / total_sales
FROM
    (SELECT
        calendar_year,
        SUM(CASE
            WHEN week_number BETWEEN 21 AND 24 THEN sales END) A
        SUM(CASE
            WHEN week_number BETWEEN 25 AND 28 THEN sales END) A
    FROM clean_weekly_sales
    WHERE calendar_year BETWEEN 2018 AND 2020
    GROUP BY calendar_year) as x
GROUP BY calendar_year
ORDER BY calendar_year
```

- **Part 2: How do the sale metrics for 12 weeks before and after compare with the previous years in 2018 and 2019?**

```
SELECT
    calendar_year,
    total_sales_after - total_sales_before as variance_sales,
    ROUND((total_sales_after - total_sales_before) / total_sales
```

```
FROM
    (SELECT
        calendar_year,
        SUM(CASE
            WHEN week_number BETWEEN 13 AND 24 THEN sales END) A
        SUM(CASE
            WHEN week_number BETWEEN 25 AND 36 THEN sales END) A
    FROM clean_weekly_sales
    WHERE calendar_year BETWEEN 2018 AND 2020
    GROUP BY calendar_year) as x
GROUP BY calendar_year
ORDER BY calendar_year
```

D. Bonus Question

- Which areas of the business have the highest negative impact in sales metrics performance in 2020 for the 12 week before and after period?

```
SELECT
    region,
    platform,
    age_band,
    demographic,
    customer_type,
    total_sales_after - total_sales_before as variance_sales,
    ROUND((total_sales_after - total_sales_before) / total_sales
FROM
    (SELECT
        region,
        platform,
        age_band,
        demographic,
        customer_type,
        SUM(CASE
            WHEN week_number BETWEEN 13 AND 24 THEN sales END) A
        SUM(CASE
```

```
            WHEN week_number BETWEEN 25 AND 36 THEN sales END) A
    FROM clean_weekly_sales
    WHERE calendar_year = 2020
    GROUP BY region,
        platform,
        age_band,
        demographic,
        customer_type) as x
GROUP BY region,
        platform,
        age_band,
        demographic,
        customer_type
ORDER BY variance_sales
LIMIT 1
```