**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\_number,

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 'Retirees'

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.calendar\_year AND cws.month\_number = tsem.month\_number

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.calendar\_year

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 percentage

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\_before \* 100,2) as variance\_percentage

FROM

(SELECT

SUM(CASE

WHEN week\_number BETWEEN 21 AND 24 THEN sales END) AS total\_sales\_before,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 28 THEN sales END) AS total\_sales\_after

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\_before \* 100,2) as variance\_percentage

FROM

(SELECT

SUM(CASE

WHEN week\_number BETWEEN 13 AND 24 THEN sales END) AS total\_sales\_before,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 36 THEN sales END) AS total\_sales\_after

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\_before \* 100,2) as variance\_percentage

FROM

(SELECT

calendar\_year,

SUM(CASE

WHEN week\_number BETWEEN 21 AND 24 THEN sales END) AS total\_sales\_before,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 28 THEN sales END) AS total\_sales\_after

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\_before \* 100,2) as variance\_percentage

FROM

(SELECT

calendar\_year,

SUM(CASE

WHEN week\_number BETWEEN 13 AND 24 THEN sales END) AS total\_sales\_before,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 36 THEN sales END) AS total\_sales\_after

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\_before \* 100,2) as variance\_percentage

FROM

(SELECT

region,

platform,

age\_band,

demographic,

customer\_type,

SUM(CASE

WHEN week\_number BETWEEN 13 AND 24 THEN sales END) AS total\_sales\_before,

SUM(CASE

WHEN week\_number BETWEEN 25 AND 36 THEN sales END) AS total\_sales\_after

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