Case Study #5

- Data Mart

Welcome to our comprehensive solution overview for **- Data Mart** case study.

Join us as we dive into the queries and unveil the valuable insights that can be derived from the provided tables: 'data_mart.weekly_sales'.

By Vijaykumar Chauhan

8WeekSQLChallenge.com

CASE STUDY #5

DATA MART

fresh is best

DataWithDanny.com

Introduction

Danny's latest venture, Data Mart, implemented significant changes in June 2020 by adopting sustainable packaging methods for all its products. This transition aimed to align the business with environmentally friendly practices throughout the supply chain. Now, Danny seeks assistance in analyzing the sales performance to understand the quantifiable impact of this sustainability initiative on Data Mart's various business aspects.

Problem Statement

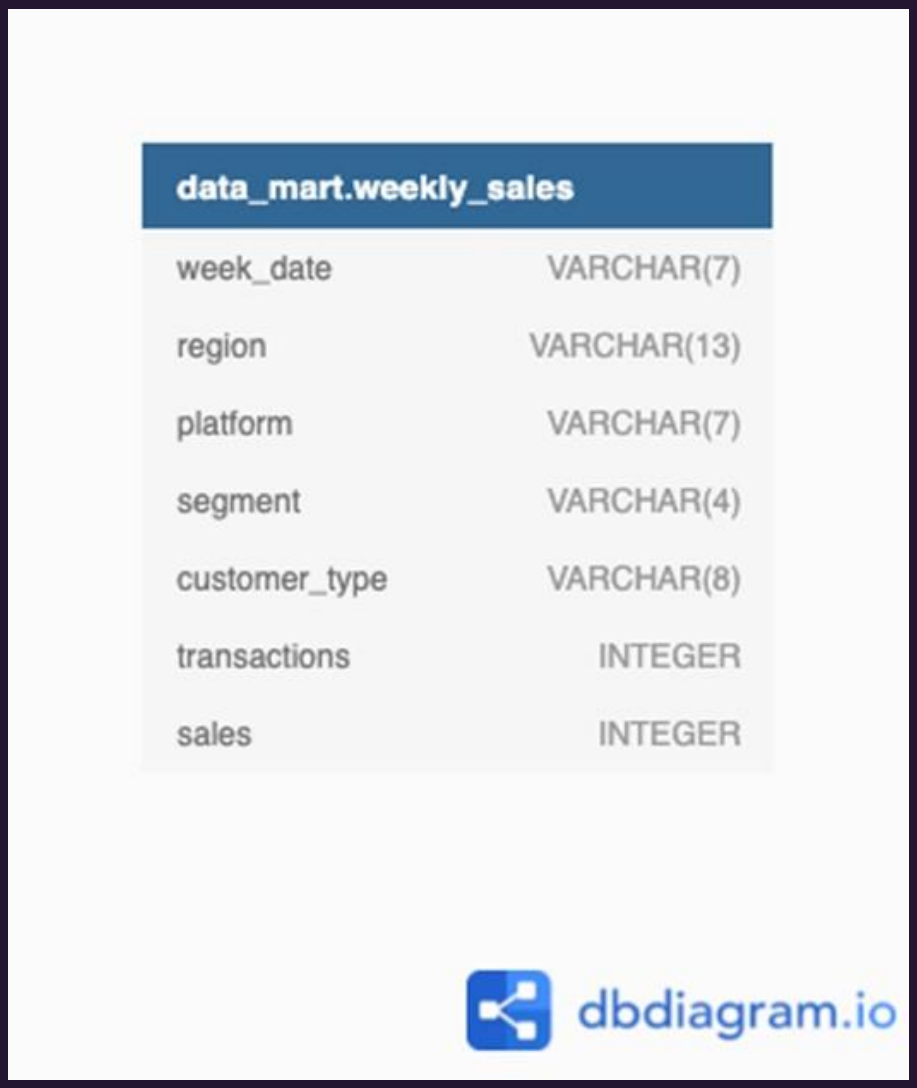
To effectively address Danny's concerns, we need to explore three key business questions:

1. **Quantifiable Impact Assessment:**
    - Analyze the sales data before and after the introduction of sustainable packaging in June 2020.
    - Quantify the impact of the sustainability changes on overall sales performance.
    - Examine changes in sales trends, revenue, and customer behavior to understand the direct effects of the initiative.
2. **Identification of Most Impacted Areas:**
    - Determine which platforms (online, offline), regions (domestic, international), product segments, and customer types experienced the most significant impact.
    - Use segmentation analysis to identify patterns and variations in sales performance across different business areas.
    - Provide insights into whether certain regions or customer segments responded more positively or negatively to the sustainability changes.
3. **Strategic Recommendations for Future Sustainability Updates:**
    - Propose strategies to minimize the potential negative impact on sales for future sustainability updates.
    - Consider tailoring communication strategies to different customer segments and regions.
    - Suggest ways to proactively address concerns or resistance from specific platforms, regions, or customer types.
    - Explore the possibility of conducting customer surveys or feedback sessions to gather insights and preferences regarding sustainability initiatives.

Available Data

For this case study there is only a single table: data_mart.weekly_sales

The Entity Relationship Diagram is shown below with the data types made clear, please note that there is only this one table - hence why it looks a little bit lonely!



Column Dictionary

The columns are pretty self-explanatory based on the column names but here are some further details about the dataset:

1. Data Mart has international operations using a multi-region strategy
2. Data Mart has both, a retail and online platform in the form of a Shopify store front to serve their customers
3. Customer segment and customer_type data relates to personal age and demographics information that is shared with Data Mart
4. transactions is the count of unique purchases made through Data Mart and sales is the actual dollar amount of purchases

Each record in the dataset is related to a specific aggregated slice of the underlying sales data rolled up into a week_date value which represents the start of the sales week.

Case Study Questions

The following case study questions require some data cleaning steps before we start to unpack Danny's key business questions in more depth.

1. Data Cleansing Steps

In a single query, perform the following operations and generate a new table in the data_mart schema named clean_weekly_sales:

- Convert the week_date to a DATE format
- 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 a month_number with the calendar month for each week_date value as the 3rd column
- Add a calendar_year column as the 4th column containing either 2018, 2019 or 2020 values
- Add a new column called age_band after the original segment column using the following mapping on the number inside the segment value segment age_band 1 Young Adults 2 Middle Aged 3 or 4 Retirees

segment	**age_band**
1	Young Adults
2	Middle Aged
3 or 4	Retirees

- Add a new demographic column using the following mapping for the first letter in the segment values: segment demographic C Couples F Families

segment	demographic
C	Couples
F	Families

- 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
- Generate a new avg_transaction column as the sales value divided by transactions rounded to 2 decimal places for each record

Solution Overview

In order to tackle this case study, we have devised various queries that will allow us to extract meaningful information from the available data. Let's take a closer look at each query and its purpose in our analysis.

Solution 1:

```
-- Drop the table if it exists

DROP TABLE IF EXISTS clean_weekly_sales;

-- Create the clean_weekly_sales table

CREATE TABLE clean_weekly_sales AS (

SELECT CAST(week_date AS DATE) AS week_date,

EXTRACT(WEEK FROM CAST(week_date AS DATE)) AS week_number,

EXTRACT(MONTH FROM CAST(week_date AS DATE)) AS month_number,

EXTRACT(YEAR FROM CAST(week_date AS DATE)) AS calendar_year,

region, platform, segment,

CASE WHEN RIGHT(segment,1) = '1' THEN 'Young Adults' WHEN RIGHT(segment,1) = '2' THEN 'Middle Aged'

WHEN RIGHT(segment,1) IN ('3','4') THEN 'Retirees' ELSE 'unknown' END AS age_band,

CASE WHEN LEFT(segment,1) = 'C' THEN 'Couples'

WHEN LEFT(segment,1) = 'F' THEN 'Families' ELSE 'unknown' END AS demographic,

transactions, ROUND((sales::NUMERIC/transactions), 2) AS avg_transaction,

sales FROM data_mart.weekly_sales );

-- Display the first few rows of the clean_weekly_sales table

SELECT * FROM clean_weekly_sales LIMIT 5;
```

Output

week_date	week_number	month_number	calendar_year	region	platform	segment
2020-08-31	36	8	2020	SOUTH AMERICA	Retail	C1
2020-08-31	36	8	2020	EUROPE	Retail	unknown
2020-08-31	36	8	2020	CANADA	Shopify	F3
2020-08-31	36	8	2020	USA	Shopify	F2
2020-08-31	36	8	2020	AFRICA	Retail	F2

age_band	demographic	transactions	avg_transaction	sales
unknown	unknown	97001	34.56	3352338
Retirees	Families	36	169.64	6107
Middle Aged	Families	2031	198.12	402380
Middle Aged	Families	277600	53.71	14910405
Young Adults	Couples	2466	21.26	52421

2. Data Exploration

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

```
SELECT DISTINCT (TO_CHAR(week_date, 'Day')) as Week_Day

FROM clean_weekly_sales;
```

week_day
Monday

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

```
WITH week_number_cte AS (

SELECT GENERATE_SERIES(1,52) AS week_number

)

SELECT DISTINCT week_no.week_number

FROM week_number_cte AS week_no

LEFT JOIN clean_weekly_sales AS sales

ON week_no.week_number = sales.week_number

WHERE sales.week_number IS NULL;

-- Filter to identify the missing week numbers where the values are Null.
```

week_number
1
2
3
4
5
6
7
8
9
10
11
12
37
38
39

Note: make sure that you have retrieved 28 rows!

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

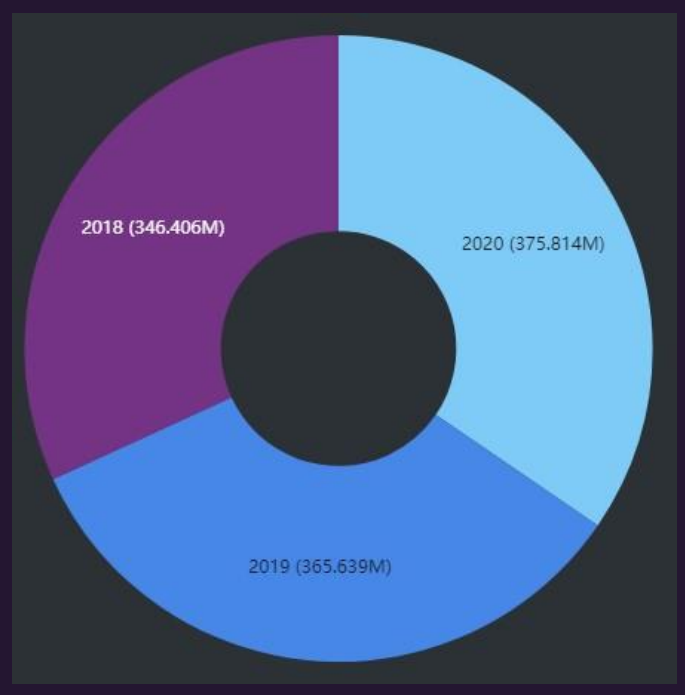
```
SELECT calendar_year,

SUM(transactions) AS total_transactions

FROM clean_weekly_sales

group by 1;
```

calendar_year	total_transactions
2020	375813651
2019	365639285
2018	346406460



Query 4: What is the total sales for each region for each month?

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

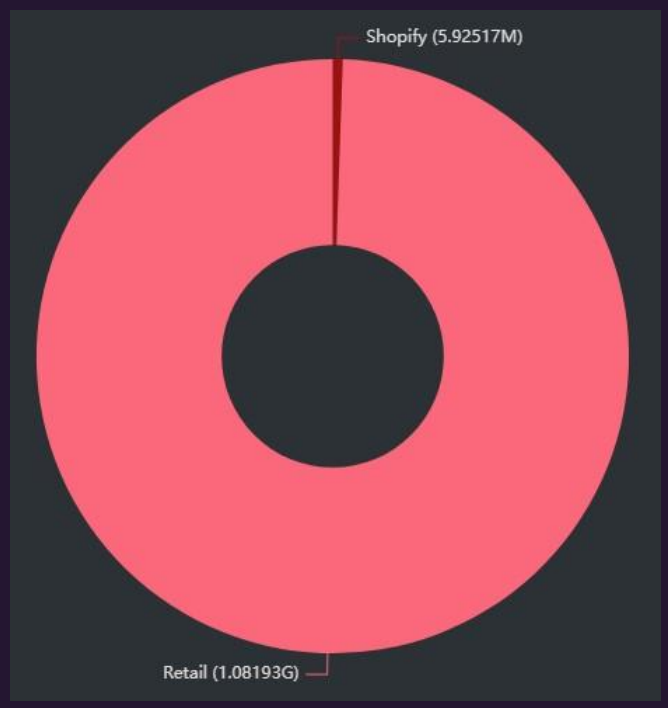
Note: make sure that you have retrieved 49 rows!

month_number	region	total_sales
3	AFRICA	567767480
3	ASIA	529770793
3	CANADA	144634329
3	EUROPE	35337093
3	OCEANIA	783282888
3	SOUTH AMERICA	71023109
3	USA	225353043
4	AFRICA	1911783504
4	ASIA	1804628707
4	CANADA	484552594

Query 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;
```

platform	total_transactions
Shopify	5925169
Retail	1081934227



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

```
WITH monthly_transactions AS (
SELECT  calendar_year,
month_number,  platform,
SUM(sales) AS monthly_sales
FROM clean_weekly_sales
GROUP BY calendar_year, month_number, platform)
SELECT calendar_year, month_number,
ROUND(100 * MAX (CASE WHEN platform = 'Retail' THEN monthly_sales ELSE NULL END)
/ SUM(monthly_sales),2) AS retail_percentage,
ROUND(100 * MAX
(CASE
WHEN platform = 'Shopify' THEN monthly_sales ELSE NULL END)
/ SUM(monthly_sales),2) AS shopify_percentage
FROM monthly_transactions
GROUP BY calendar_year, month_number
ORDER BY calendar_year, month_number;
```

calendar_year	month_number	retail_percentage	shopify_percentage
2018	3	97.92	2.08
2018	4	97.93	2.07
2018	5	97.73	2.27
2018	6	97.76	2.24
2018	7	97.75	2.25
2018	8	97.71	2.29
2018	9	97.68	2.32
2019	3	97.71	2.29
2019	4	97.80	2.20
2019	5	97.52	2.48
2019	6	97.42	2.58
2019	7	97.35	2.65
2019	8	97.21	2.79
2019	9	97.09	2.91

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

```
WITH demographic_sales AS (
SELECT calendar_year, demographic,
SUM(sales) AS yearly_sales
FROM clean_weekly_sales
GROUP BY calendar_year, demographic
)
SELECT calendar_year,
ROUND(100 * MAX (CASE
WHEN demographic = 'Couples' THEN yearly_sales ELSE NULL END)
/ SUM(yearly_sales),2) AS couples_percentage,
ROUND(100 * MAX
(CASE
WHEN demographic = 'Families' THEN yearly_sales ELSE NULL END)
/ SUM(yearly_sales),2) AS families_percentage,
ROUND(100 * MAX
(CASE
WHEN demographic = 'unknown' THEN yearly_sales ELSE NULL END)
/ SUM(yearly_sales),2) AS unknown_percentage
FROM demographic_sales
GROUP BY calendar_year;
```

calendar_year	couples_percentage	families_percentage	unknown_percentage
2018	26.38	31.99	41.63
2019	27.28	32.47	40.25
2020	28.72	32.73	38.55

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

```
SELECT age_band,demographic,sum(sales) AS Retail_Sales,

ROUND(100 * SUM(sales)::NUMERIC/ SUM(SUM(sales)) OVER ( ),1) AS contribution_percentage

FROM clean_weekly_sales

WHERE platform = 'Retail'

GROUP BY 1, 2

ORDER BY Retail_Sales Desc;
```

age_band	demographic	retail_sales	contribution_percentage
unknown	unknown	16067285533	40.5
Retirees	Families	6634686916	16.7
Retirees	Couples	6370580014	16.1
Middle Aged	Families	4354091554	11.0
Young Adults	Couples	2602922797	6.6
Middle Aged	Couples	1854160330	4.7
Young Adults	Families	1770889293	4.5

Query 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),0) AS avg_transaction_row,
SUM(sales) / sum(transactions) AS
avg_transaction_group
FROM clean_weekly_sales
GROUP BY calendar_year, platform
ORDER BY calendar_year, platform;
```

Output

calendar_year	platform	avg_transaction_row	avg_transaction_group
2018	Retail	43	36
2018	Shopify	188	192
2019	Retail	42	36
2019	Shopify	178	183
2020	Retail	41	36
2020	Shopify	175	179

The difference between avg_transaction_row and avg_transaction_group is as follows:

- avg_transaction_row calculates the average transaction size by dividing the sales of each row by the number of transactions in that row.
- On the other hand, avg_transaction_group calculates the average transaction size by dividing the total sales for the entire dataset by the total number of transactions.

For finding the average transaction size for each year by platform accurately, it is recommended to use avg_transaction_group.

3. 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:

Query 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?

NOTE: 2020-06-15 Week number is 25

```
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 1,  2

),  Before_After_sales as

( SELECT  SUM(CASE WHEN week_number BETWEEN 21 AND24 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) Variance,

ROUND(((After_packaging_sales-Before_packaging_sales)/Before_packaging_sales)*100,2) variance_percentage

FROM Before_After_sales;
```

variance	variance_percentage
-26884188	-1.15

Note: Since the implementation of the new sustainable packaging, there has been a decrease in sales amounting by **$26,884,188** reflecting a negative change at **1.15%.** Introducing a new packaging does not always guarantee positive results as customers may not readily recognise your product on the shelves due to the change in packaging.

Query 2: 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  1, 2

), Before_After_sales as

( SELECT    SUM(CASE WHEN week_number BETWEEN 13 AND24
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) Variance ,

ROUND(((After_packaging_sales-Before_packaging_sales)/Before_packaging_sales)*100,2)
variance_percentage

FROM Before_After_sales;
```

Output

variance	variance_percentage
-152325394	-2.14

Note: Looks like the sales have experienced a further decline, now at a negative **2.14%!** If I'm Danny's boss, I wouldn't be too happy with the results.

Query 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?

- Basically, the question is asking us to find the sales variance between 4 weeks before and after '2020-06-15' for years 2018, 2019 and 2020. Perhaps we can find a pattern here.
- We can apply the same solution as above and add calendar_year into the syntax.

```
WITH Packaging_sales as (

SELECT  week_date,  week_number,calendar_year,

SUM(sales) as total_sales

FROM clean_weekly_sales

WHERE week_number BETWEEN 21 AND 28

 AND calendar_year in (2020,2019,2018)

GROUP BY 1, 2,3

), Before_After_sales as (

SELECT calendar_year,

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

group by  calendar_year )

SELECT calendar_year ,

(After_packaging_sales - Before_packaging_sales) Variance,

ROUND(((After_packaging_sales-Before_packaging_sales)/Before_packaging_sales)*100,2)
variance_percentage

FROM Before_After_sales;
```

Output

calendar_year	variance	variance_percentage
2018	4102105	0.19
2019	2336594	0.10
2020	-26884188	-1.15

- In 2018, there was a sales variance of $4,102,105, indicating a positive change of 0.19% compared to the period before the packaging change.
- Similarly, in 2019, there was a sales variance of $2,336,594, corresponding to a positive change of 0.10% when comparing the period before and after the packaging change.
- However, in 2020, there was a substantial decrease in sales following the packaging change. The sales variance amounted to $26,884,188, indicating a significant negative change of -1.15%. This reduction represents a considerable drop compared to the previous years.

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

Use the same solution above and change to week 13 to 24 for before and week 25 to 37 for after.

```
WITH Packaging_sales as (

SELECT  week_date,     week_number,calendar_year,

SUM(sales) as total_sales

FROM  clean_weekly_sales

WHERE  week_number BETWEEN 13 AND 37

AND calendar_year in (2020,2019,2018)

GROUP BY   1,  2, 3

),Before_After_sales 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 37 THEN total_sales END) AS
After_packaging_sales

FROM Packaging_sales

GROUP BY calendar_year )

SELECT calendar_year,(After_packaging_sales - Before_packaging_sales)
Variance,

ROUND(((After_packaging_sales-
Before_packaging_sales)/Before_packaging_sales)*100,2)
variance_percentage

FROM Before_After_sales;
```

Output

calendar_year	variance	variance_percentage
2018	104256193	1.63
2019	-20740294	-0.30
2020	-152325394	-2.14

Note: There was a fair bit of percentage differences in all 3 years. However, now when you compare the worst year to their best year in 2018, the sales percentage difference is even more stark at a difference of **3.77% (1.63% + 2.14%).**

When comparing the sales performance across all three years, there were noticeable variations in the percentage differences. However, the most significant contrast emerges when comparing the worst-performing year in 2020 to the best-performing year in 2018. In this comparison, the sales percentage difference becomes even more apparent with a significant gap of **3.77% (1.63% + 2.14%).**

4.Bonus Questions

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

- region
- platform
- age_band
- demographic

```
WITH twelve_weeks_before AS (

SELECT DISTINCT week_date

FROM clean_weekly_sales

WHERE week_date BETWEEN (TO_DATE('2020-06-15', 'yy/mm/dd') - interval '12 weeks') AND (TO_DATE('2020-06-15', 'yy/mm/dd') - interval '1 week')

) , twelve_weeks_after AS(

SELECT DISTINCT week_date

FROM clean_weekly_sales

WHERE week_date BETWEEN TO_DATE('2020-06-15', 'yy/mm/dd') AND  (TO_DATE('2020-06-15', 'yy/mm/dd') + interval '11 weeks')

) ,  summations AS(

SELECT region, SUM(CASE WHEN week_date in (select * from twelve_weeks_before) THEN sales END) AS twelve_weeks_before,

SUM(CASE WHEN week_date in (select * from twelve_weeks_after) THEN sales END) AS twelve_weeks_after

FROM clean_weekly_sales

GROUP BY region  )

SELECT *,  twelve_weeks_after - twelve_weeks_before AS variance,

ROUND(100 * (twelve_weeks_after - twelve_weeks_before)::numeric/twelve_weeks_after, 2) AS percentage_change

FROM summations

ORDER BY percentage_change;
```

By Region:

region	twelve_weeks_before	twelve_weeks_after	variance	percentage_change
ASIA	1637244466	1583807621	-53436845	-3.37
OCEANIA	2354116790	2282795690	-71321100	-3.12
SOUTH AMERICA	213036207	208452033	-4584174	-2.20
CANADA	426438454	418264441	-8174013	-1.95
USA	677013558	666198715	-10814843	-1.62
AFRICA	1709537105	1700390294	-9146811	-0.54
EUROPE	108886567	114038959	5152392	4.52

- The only region that experienced positive growth is Europe, with a 4.52% increase.
- All other regions experienced negative growth, with Asia and Oceania experiencing the largest declines.
- The overall percentage change across all regions is -1.70%.

By Platform:

platform	twelve_weeks_before	twelve_weeks_after	variance	percentage_change
Retail	6906861113	6738777279	-168083834	-2.49
Shopify	219412034	235170474	15758440	6.70

Retail shopping had the highest negative impact between platforms

By Age Bound:

age_band	twelve_weeks_before	twelve_weeks_after	variance	percentage_change
unknown	2764354464	2671961443	-92393021	-3.46
Middle Aged	1164847640	1141853348	-22994292	-2.01
Retirees	2395264515	2365714994	-29549521	-1.25
Young Adults	801806528	794417968	-7388560	-0.93

Unknown age_band was **the** most impacted negatively among the age bands

By Demographic:

demographic	twelve_weeks_before	twelve_weeks_after	variance	percentage_change
unknown	2764354464	2671961443	-92393021	-3.46
Families	2328329040	2286009025	-42320015	-1.85
Couples	2033589643	2015977285	-17612358	-0.87

Unknown demographic was the most negatively impacted demographic

Thankyou