EASTERN INTERNATIONAL UNIVERSITY

BECAMEX BUSINESS SCHOOL

DATA MART CASE STUDY #5

Course : **MIS 443 - BUSINESS DATA MANAGEMENT**

Lecturers : **Mr. Dang Thai Doan**

Prepared by : **Group 2**

Quarter : **4, 2024-2025**

Name	IRN
Nguyễn Quang Trường	2132309001
Nguyễn Trần Ngọc Thy	2232300307
Huỳnh Thuỵ Bảo Trâm	2132300228
Huỳnh Trung Hậu	2132309003

TABLE OF CONTENTS

I. Introduction

1. Overview of the Case Study

Data Mart is a multi-platform supermarket chain that operates through both traditional physical stores and an online platform (Shopify). This case study focuses on analyzing consumer purchasing behavior and identifying how a shift to sustainable packaging (implemented from June 15, 2020) impacted sales performance. Through structured SQL-based exploration, we aim to uncover trends, seasonality, and behavioral shifts in customer data.

2. Database Overview

Original Columns (7):

- "week_date" – the date representing the start of the week (always Monday)
- "region" – geographic area (e.g., Asia, Europe)
- "platform" – sales channel (Retail or Shopify)
- "segment" – customer segmentation code (e.g., C1, F3)
- "customer_type" – type of customer (e.g., New or Existing)
- "sales" – total weekly sales revenue
- "transactions" – number of transactions in the week

Derived Columns (6):

- "week_number" – ISO-compliant week number (1–53)
- "month_number" – calendar month (1–12)
- "calendar_year" – extracted year (2018, 2019, 2020)
- "age_band" – customer age group (e.g., Young Adults, Retirees) derived from segment
- "demographic" – customer type (e.g., Couples, Families) derived from segment
- "avg_transaction" – average transaction value (sales / transactions, rounded to 2 decimals)

3. Project Objective

- Clean and transform the original weekly sales data into a structured format

- Explore customer behavior across time, region, platform, and demographics
- Assess the before-and-after impact of sustainable packaging policy (using 2020-06-15 as the baseline)
- Provide actionable insights and recommendations for Data Mart's management team

II. Case Study Questions & Analysis Plan

A. Data Cleansing Steps

The process of **data cleansing** is essential to ensuring that the dataset is accurate, consistent, and structured in a way that facilitates effective analysis. In this case study, we focus on cleaning the **weekly_sales** dataset to prepare it for analysis. The steps outlined below describe the transformation and cleaning processes carried out on the dataset to achieve high-quality, usable data.

1. Conversion of week_date to Date Format

The first step in data cleansing was to convert the week_date column, which originally contained text values in the format DD/MM/YY, into a proper date format. This conversion ensures that the data can be manipulated more easily and that time-based analysis can be performed without errors. The new date format adopted for the week_date column is YYYY/MM/DD, which aligns with standard conventions and provides a clear, consistent reference for temporal analysis.

SQL Used:

```
TO_DATE(week_date, 'DD/MM/YY') AS week_date
```

This transformation was necessary to facilitate accurate filtering, grouping, and comparison of date-related information, especially for weekly, monthly, and yearly analyses.

2. Creation of Temporal Variables: week_number, month_number, and calendar_year

Once the week_date column was cleaned, three new columns were derived from it to enhance the dataset's temporal analysis capabilities:

- **week_number**: This column represents the week number within a given year. It was extracted using SQL's EXTRACT function to ensure that each row represents a unique week.

- **month_number**: This column provides the month number (1 through 12) extracted from the week_date, making it possible to perform month-over-month comparisons.

- **calendar_year**: This column was created to capture the year in which the data was recorded, extracted directly from the week_date field.

These temporal fields are crucial for conducting time-series analyses, tracking trends, and comparing performance across different time periods (such as monthly or yearly comparisons).

SQL Used:

```
EXTRACT(WEEK FROM TO_DATE(week_date, 'DD/MM/YY')) AS week_number,
EXTRACT(MONTH FROM TO_DATE(week_date, 'DD/MM/YY')) AS month_number,
EXTRACT(YEAR FROM TO_DATE(week_date, 'DD/MM/YY')) AS calendar_year
```

3. Classification of Customer Data: age_band and demographic

The **segment** column in the dataset contains encoded information about the customers. This data was difficult to interpret directly, so we broke it down into two more readable columns: **age_band** and **demographic**. These columns were derived using the last and first characters of the **segment** code, respectively.

- **age_band**: The value for this column was determined by examining the last character of the **segment** code. Depending on the character, customers were categorized as follows:
    - 1 = Young Adults
    - 2 = Middle Aged
    - 3, 4 = Retirees
    - Any other value was classified as unknown.
- **demographic**: The first character of the **segment** code was used to determine whether the customer was part of a **Couples** or **Families** group:

- C = Couples
- F = Families
- Any other value was classified as unknown.

By creating these two new columns, we were able to better understand customer demographics, which is key for analyzing trends and making targeted business decisions.

SQL Used:

```
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
```

4. Handling NULL Values: Replacing with unknown

Another crucial step in data cleansing was to handle the **NULL** values present in the **segment**, **age_band**, and **demographic** columns. In many cases, missing or NULL values could introduce errors into analyses, so it was necessary to standardize these values. We replaced all NULL or missing values in these fields with the string "unknown". This practice helps maintain consistency and ensures that missing data does not skew results.

SQL Used:

```
CASE

  WHEN segment = 'null' THEN 'unknown'
```

ELSE segment

END AS segment

5. Calculation of avg_transaction

Finally, the **avg_transaction** column was created to calculate the average value of a transaction for each week. This was done by dividing the **sales** value by the **transactions** count, providing a measure of how much revenue was generated per transaction. The result was rounded to two decimal places to maintain precision and ensure that the value is easily interpretable.

This metric is important for understanding customer spending behavior and can help identify trends in customer purchasing habits.

SQL Used:

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

Output:

	week_date date	week_number numeric	month_number numeric	calendar_year numeric	region character varying (13)	platform character varying (7)	segment character varying	age_band text	demographic text	customer_type character varying (8)	transactions integer	sales integer	avg_transaction numeric
1	2020-08-31	36	8	2020	ASIA	Retail	C3	Retirees	Couples	New	120631	3656163	30.31
2	2020-08-31	36	8	2020	ASIA	Retail	F1	Young Adults	Families	New	31574	996575	31.56
3	2020-08-31	36	8	2020	USA	Retail	unknown	unknown	unknown	Guest	529151	16509610	31.20
4	2020-08-31	36	8	2020	EUROPE	Retail	C1	Young Adults	Couples	New	4517	141942	31.42
5	2020-08-31	36	8	2020	AFRICA	Retail	C2	Middle Aged	Couples	New	58046	1758388	30.29
6	2020-08-31	36	8	2020	CANADA	Shopify	F2	Middle Aged	Families	Existing	1336	243878	182.54
7	2020-08-31	36	8	2020	AFRICA	Shopify	F3	Retirees	Families	Existing	2514	519502	206.64
8	2020-08-31	36	8	2020	ASIA	Shopify	F1	Young Adults	Families	Existing	2158	371417	172.11
9	2020-08-31	36	8	2020	AFRICA	Shopify	F2	Middle Aged	Families	New	318	49557	155.84
10	2020-08-31	36	8	2020	AFRICA	Retail	C3	Retirees	Couples	New	111032	3888162	35.02
11	2020-08-31	36	8	2020	USA	Shopify	F1	Young Adults	Families	Existing	1398	260773	186.53
12	2020-08-31	36	8	2020	OCEANIA	Shopify	C2	Middle Aged	Couples	Existing	4661	882690	189.38
13	2020-08-31	36	8	2020	SOUTH AMERICA	Retail	C2	Middle Aged	Couples	Existing	1029	38762	37.67
14	2020-08-31	36	8	2020	SOUTH AMERICA	Shopify	C4	Retirees	Couples	New	6	917	152.83
15	2020-08-31	36	8	2020	EUROPE	Shopify	F3	Retirees	Families	Existing	115	35215	306.22
16	2020-08-31	36	8	2020	OCEANIA	Retail	F3	Retirees	Families	Existing	551905	30371770	55.03
17	2020-08-31	36	8	2020	ASIA	Shopify	C3	Retirees	Couples	Existing	1969	374327	190.11
18	2020-08-31	36	8	2020	AFRICA	Retail	F1	Young Adults	Families	Existing	97604	5185233	53.13
19	2020-08-31	36	8	2020	OCEANIA	Retail	C2	Middle Aged	Couples	New	111219	2980673	26.80
20	2020-08-31	36	8	2020	USA	Retail	F1	Young Adults	Families	New	11820	463738	39.23
21	2020-08-31	36	8	2020	SOUTH AMERICA	Retail	F3	Retirees	Families	Existing	1363	65730	48.22
22	2020-08-31	36	8	2020	AFRICA	Retail	C3	Retirees	Couples	Existing	284971	14430196	50.64
23	2020-08-31	36	8	2020	ASIA	Retail	F2	Middle Aged	Families	New	70496	2176980	30.88

6. Summary of Data Cleansing Steps

Through the above steps, the data was transformed into a clean, structured format that is ready for further analysis. Each of these transformations was performed to ensure:

- **Temporal consistency** through the conversion of dates and creation of time-based variables.
- **Improved customer segmentation** through the extraction of demographic and age information from the segment code.
- **Handling missing or NULL data** with consistent replacements to ensure accurate analysis.
- **Derived metrics**, such as avg_transaction, to facilitate deeper insights into customer behavior.

With these steps, the dataset is now structured, consistent, and ready to be used for meaningful exploration and analysis in the following sections of the case study.

B. Data Exploration

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

All week_date values in the dataset fall on a Monday. This consistency indicates that the data is recorded on a weekly basis, using Mondays as the reference date for each week. This aligns with the ISO 8601 standard for week-based analysis, where weeks start on Mondays.

SQL Used:

```
SELECT DISTINCT week_date,
    TO_CHAR(week_date, 'Day') AS day_of_week
FROM data_mart.clean_weekly_sales
ORDER BY week_date;
```

Output:

	week_date date	day_of_week text
62	2020-06-22	Monday
63	2020-06-29	Monday
64	2020-07-06	Monday
65	2020-07-13	Monday
66	2020-07-20	Monday
67	2020-07-27	Monday
68	2020-08-03	Monday
69	2020-08-10	Monday
70	2020-08-17	Monday
71	2020-08-24	Monday
72	2020-08-31	Monday

Insight:

All week_date values fall on Monday, which means:

- The dataset is standardized for weekly reporting.
- Each row represents data for a week starting from Monday.
- This ensures temporal consistency for time-based analysis (e.g. week-on-week trends).

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

We generated a list of all ISO week numbers from 1 to 53 and compared them against the week_number values present in the dataset. The result shows that the dataset is missing 29 week numbers, primarily outside the range 13 to 36, which is the actual range of weeks present in the dataset.

SQL Used:

```
WITH all_weeks AS (
  SELECT generate_series(1, 53) AS week_number
),
```

```
actual_weeks AS (
  SELECT DISTINCT week_number
  FROM data_mart.clean_weekly_sales
)
SELECT aw.week_number
FROM all_weeks aw
LEFT JOIN actual_weeks acw
  ON aw.week_number = acw.week_number
WHERE acw.week_number IS NULL;
```

Output:

	week_number integer
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	37
14	38
15	39
16	40
17	41
18	42
19	43
20	44
21	45
22	46
23	47
24	48
25	49
26	50
27	51
28	52
29	53

Insight:

The dataset should contain 53 week numbers (1–53) for the year, but only 24 are present.

A total of 29 week numbers are missing.

This reveals incomplete weekly data, which can affect trend analysis, seasonality, or forecasting.

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

We summed all transactions and grouped the result by calendar_year. This gives an overview of platform engagement across years.

SQL Used:

```
SELECT calendar_year,
    SUM(transactions) AS total_transactions
FROM data_mart.clean_weekly_sales
GROUP BY calendar_year
ORDER BY calendar_year;
```

Output:

	calendar_year numeric	total_transactions bigint
1	2018	346406460
2	2019	365639285
3	2020	375813651

Insight:

There is a steady year-over-year growth in the number of transactions.

From 2018 to 2020, total transactions increased by approximately 8.5%.

This trend may reflect:

- Business expansion
- Improved customer engagement
- Increase in digital sales channels (e.g., Shopify)

Business Implication:

The consistent growth in total transactions highlights positive performance. It may also suggest:

- A strong customer base
- Effective marketing or promotions
- Potential scalability of operations

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

We calculated total sales by grouping data by region and month_number. This helps assess geographic sales trends and seasonal performance.

SQL Used:

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

Output:

	region character varying (13)	month_number numeric	total_sales bigint
1	AFRICA	3	567767480
2	AFRICA	4	1911783504
3	AFRICA	5	1647244738
4	AFRICA	6	1767559760
5	AFRICA	7	1960219710
6	AFRICA	8	1809596890
7	AFRICA	9	276320987
8	ASIA	3	529770793
9	ASIA	4	1804628707
10	ASIA	5	1526285399
11	ASIA	6	1619482889

Visualization:

The provided line chart visualizes total sales across regions for each month (3 to 9)

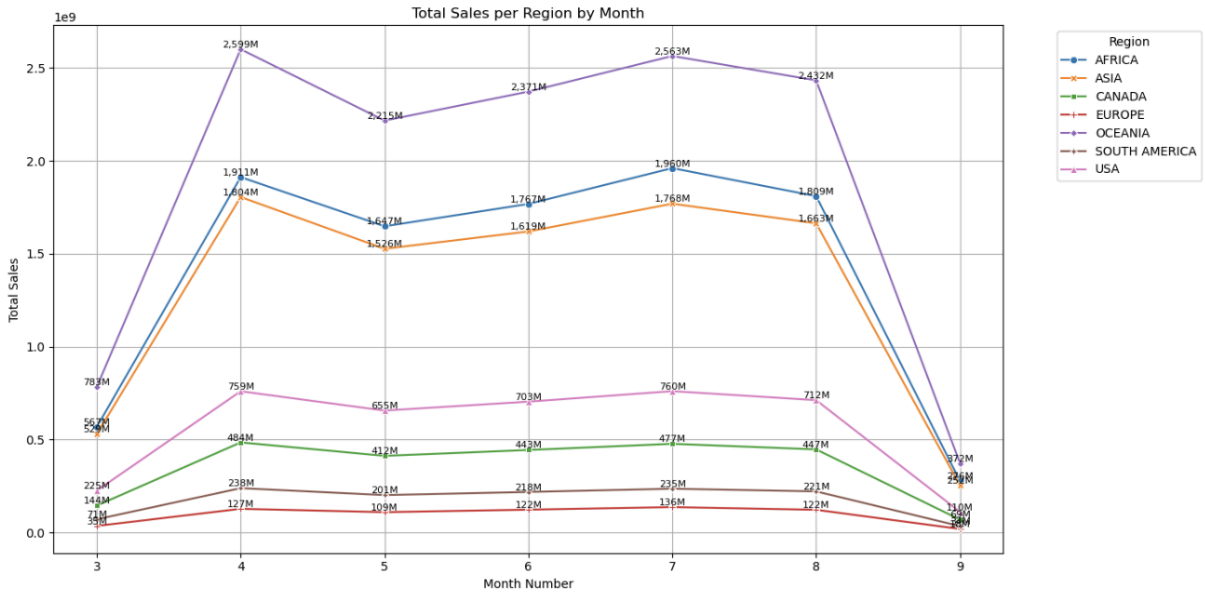


Figure 1: Total sales per region per month

Key Insight:

From your query output, we observe:

- All regions show monthly variations in total sales.
- Some regions have strong seasonal trends or standout months with peak performance.

For example:

Insight	Explanation
1. Oceania dominates	Oceania consistently shows the highest sales among all regions in every month except September (month 9), where all regions dropped drastically.
2. Peak in Month 4 (April)	Most regions (especially Oceania, Africa, and Asia) saw peak sales in month 4, suggesting a seasonal or promotional spike.
3. Sudden drop in Month 9 (September)	All regions experienced a sharp sales drop in month 9, possibly due to incomplete data, end of fiscal cycle, or external disruptions.
4. Consistency in trends	Africa and Asia follow very similar sales patterns throughout, suggesting possibly similar market behaviors or dependencies.
5. Lower performing regions	Europe and Canada consistently have the lowest total sales, indicating either smaller markets or fewer operations there.

Table 1: The insight of total sales per region per month

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

We aggregated the total number of transactions for each platform (Retail and Shopify).

SQL Used:

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

Output:

	platform character varying (7)	total_transactions bigint
1	Shopify	5925169
2	Retail	1081934227

Insight:

1. **Retail dominates** the transaction volume with nearly **95% of all transactions**, indicating it is the primary sales channel.
2. **Shopify only accounts for ~5%**, suggesting it's either a newer platform or plays a supplementary role.

Strategic Implications:

- If the company wants to **increase online presence**, investment in **Shopify optimization or promotion** could be beneficial.

- Retail's dominance suggests it's critical to maintain **strong in-store operations**, logistics, and customer service.

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

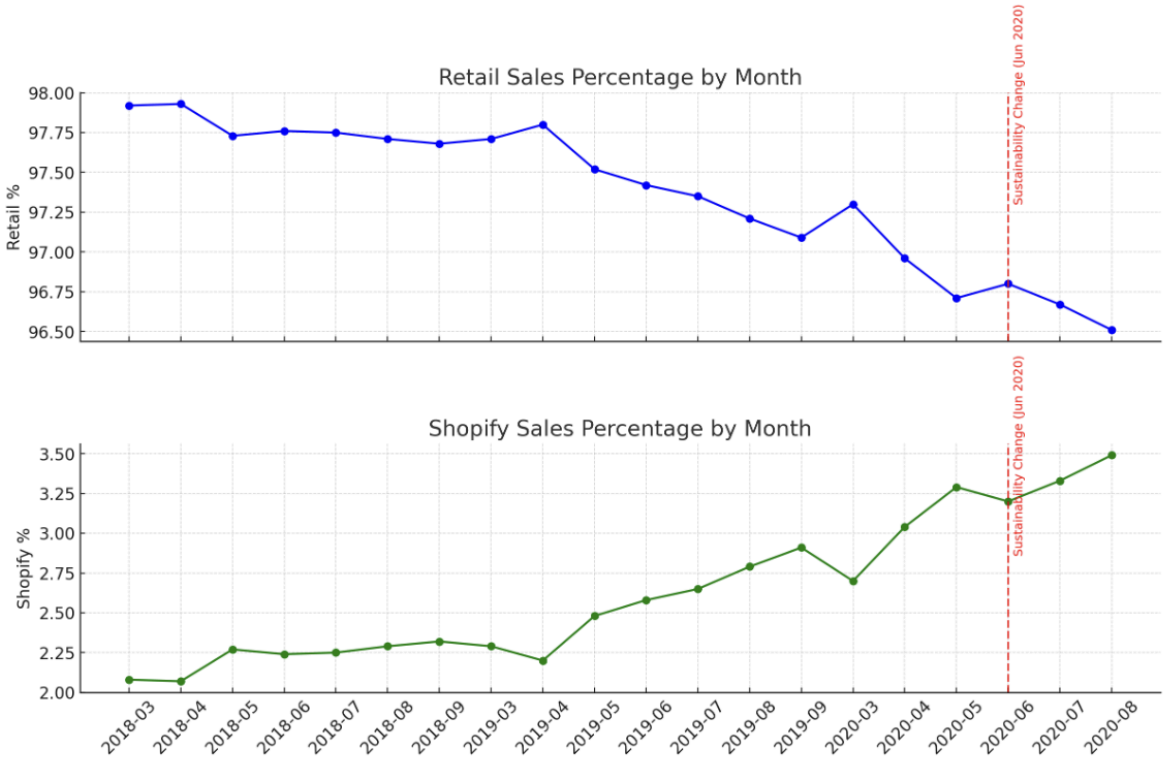
6.1 Retail Dominance

- Retail consistently holds the **majority share** — between **96.5% and 97.9%** — across all months shown.

- Shopify remains a small contributor, ranging from **~2.08% to 3.49%**.

6.2 Trend Over Time

- From early 2018 to mid-2020, Retail's share shows a **slow but steady decline**, while Shopify's share rises gradually.



-

Figure 2: Retail vs Shopify sales percentage trend chart.

- After June 2020, Shopify's share increases slightly faster — from **3.20% in June** to **3.49% in August**.

SQL:

```
SELECT
   calendar_year,
   month_number,
   ROUND( SUM(CASE WHEN platform = 'Retail' THEN sales ELSE 0 END) *100.0/
            SUM(sales), 2) AS retail_percentage,
   ROUND( SUM(CASE WHEN platform = 'Shopify' THEN sales ELSE 0 END) *100.0/
            SUM(sales), 2) AS shopify_percentage
FROM data_mart.clean_weekly_sales
GROUP BY calendar_year, month_number
ORDER BY calendar_year, month_number;
```

Output

	calendar_year numeric	month_number numeric	retail_percentage numeric	shopify_percentage numeric
1	2018	3	97.92	2.08
2	2018	4	97.93	2.07
3	2018	5	97.73	2.27
4	2018	6	97.76	2.24
5	2018	7	97.75	2.25
6	2018	8	97.71	2.29
7	2018	9	97.68	2.32
8	2019	3	97.71	2.29
9	2019	4	97.80	2.20
10	2019	5	97.52	2.48
11	2019	6	97.42	2.58
12	2019	7	97.35	2.65
13	2019	8	97.21	2.79
14	2019	9	97.09	2.91
15	2020	3	97.30	2.70
16	2020	4	96.96	3.04
17	2020	5	96.71	3.29
18	2020	6	96.80	3.20
19	2020	7	96.67	3.33
20	2020	8	96.51	3.49

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

Year	Couples sales (%)	Families sales (%)	Unknown sales (%)
2018	**26.38**	**31.99**	**41.63**
2019	**27.28**	**32.73**	**40.25**
2020	**28.72**	**32.73**	**38.55**

Table 2: The percentage of sales by demographic for each year

From 2018 to 2020, there is a gradual shift in the sales distribution across demographics.

- Couples increased from 26.38% in 2018 to 28.72% in 2020, suggesting this group became a more significant customer segment over time.

- Families remained relatively stable, rising slightly from 31.99% in 2018 to 32.73% in 2020, indicating consistent engagement from this demographic.

- Unknown customers declined from 41.63% in 2018 to 38.55% in 2020, which could mean better demographic data collection or a smaller share of sales from customers without recorded demographic information.

Overall, the trend shows a modest but steady growth in sales from Couples, stable sales from Families, and a gradual decrease in the share of Unknowns, potentially reflecting improved customer profiling and targeted marketing.

SQL:

```
SELECT
   calendar_year,
   ROUND(SUM(CASE WHEN demographic = 'Couples' THEN sales ELSE 0 END) * 100.0 /
SUM(sales), 2) AS couples_sales_percent,
```

```
    ROUND(SUM(CASE WHEN demographic = 'Families' THEN sales ELSE 0 END) * 100.0 /
SUM(sales), 2) AS families_sales_percent,
    ROUND(SUM(CASE WHEN demographic = 'unknown' THEN sales ELSE 0 END) * 100.0 /
SUM(sales), 2) AS unknown_sales_percent
FROM data_mart.clean_weekly_sales
GROUP BY calendar_year
ORDER BY calendar_year;
```

Output:

	calendar_year numeric	couples_sales_percent numeric	families_sales_percent numeric	unknown_sales_percent numeric
1	2018	26.38	31.99	41.63
2	2019	27.28	32.47	40.25
3	2020	28.72	32.73	38.55

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

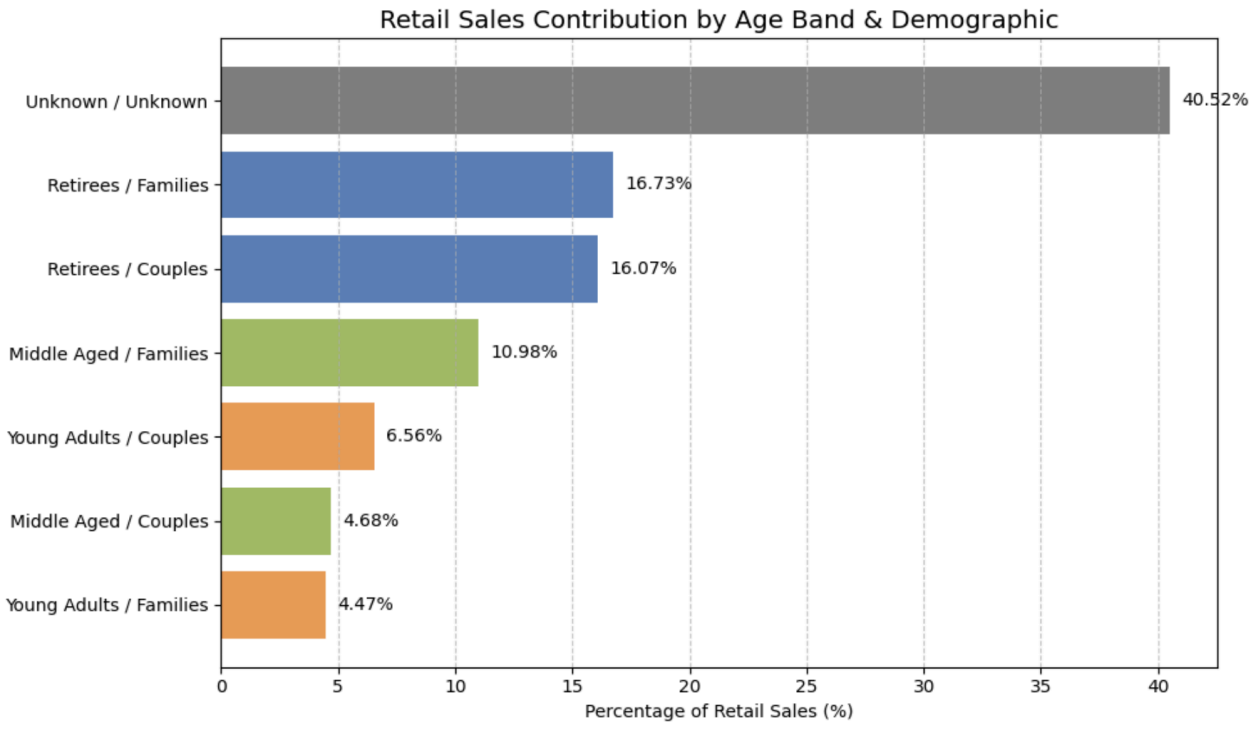


Figure 3: Rental sales contribution by Age Band and Demographic

Retail sales are heavily concentrated in the "Unknown" age_band and demographic group, which accounts for 40.52% of total sales—likely due to incomplete or missing customer profile data. Among identified groups, Retirees contribute significantly, with Families at 16.73% and Couples at 16.07% of sales, together making up over one-third of known customer segments. Middle Aged Families (10.98%) and Young Adult Couples (6.56%) are also notable contributors, while Young Adult Families represent a smaller share at 4.47%. This distribution suggests that targeted marketing efforts toward Retirees and Middle Aged Families could yield substantial returns, while improving customer data collection could unlock deeper demographic insights.

SQL:

```
SELECT
   age_band,
   demographic,
   SUM(sales) AS total_sales,
   ROUND(
      (SUM(sales) * 100.0) /
      SUM(SUM(sales)) OVER (),
      2
   ) AS percentage_sales
FROM data_mart.clean_weekly_sales
WHERE platform = 'Retail'
GROUP BY age_band, demographic
ORDER BY total_sales DESC;
```

Output:

	age_band text	demographic text	total_sales bigint	percentage_sales numeric
1	unknown	unknown	16067285533	40.52
2	Retirees	Families	6634686916	16.73
3	Retirees	Couples	6370580014	16.07
4	Middle Aged	Families	4354091554	10.98
5	Young Adults	Couples	2602922797	6.56
6	Middle Aged	Couples	1854160330	4.68
7	Young Adults	Families	1770889293	4.47

C. Before & After Analysis (Event Impact)

1. Identify the problem

This part presents a detailed quantitative analysis of the sales impact resulting from the implementation of sustainable packaging at Data Mart on June 15, 2020. The initiative, while environmentally motivated, may have affected consumer behavior and thus influenced revenue outcomes. This analysis is aimed at identifying and quantifying any significant change in sales patterns before and after the transition, over both short-term (4 weeks) and medium-term (12 weeks) windows. Additionally, the performance of these periods in 2020 is benchmarked against equivalent periods in the years 2018 and 2019 to account for typical seasonal variation and historical trends.

2. Methodology

The following steps were taken to ensure a consistent and reliable evaluation:

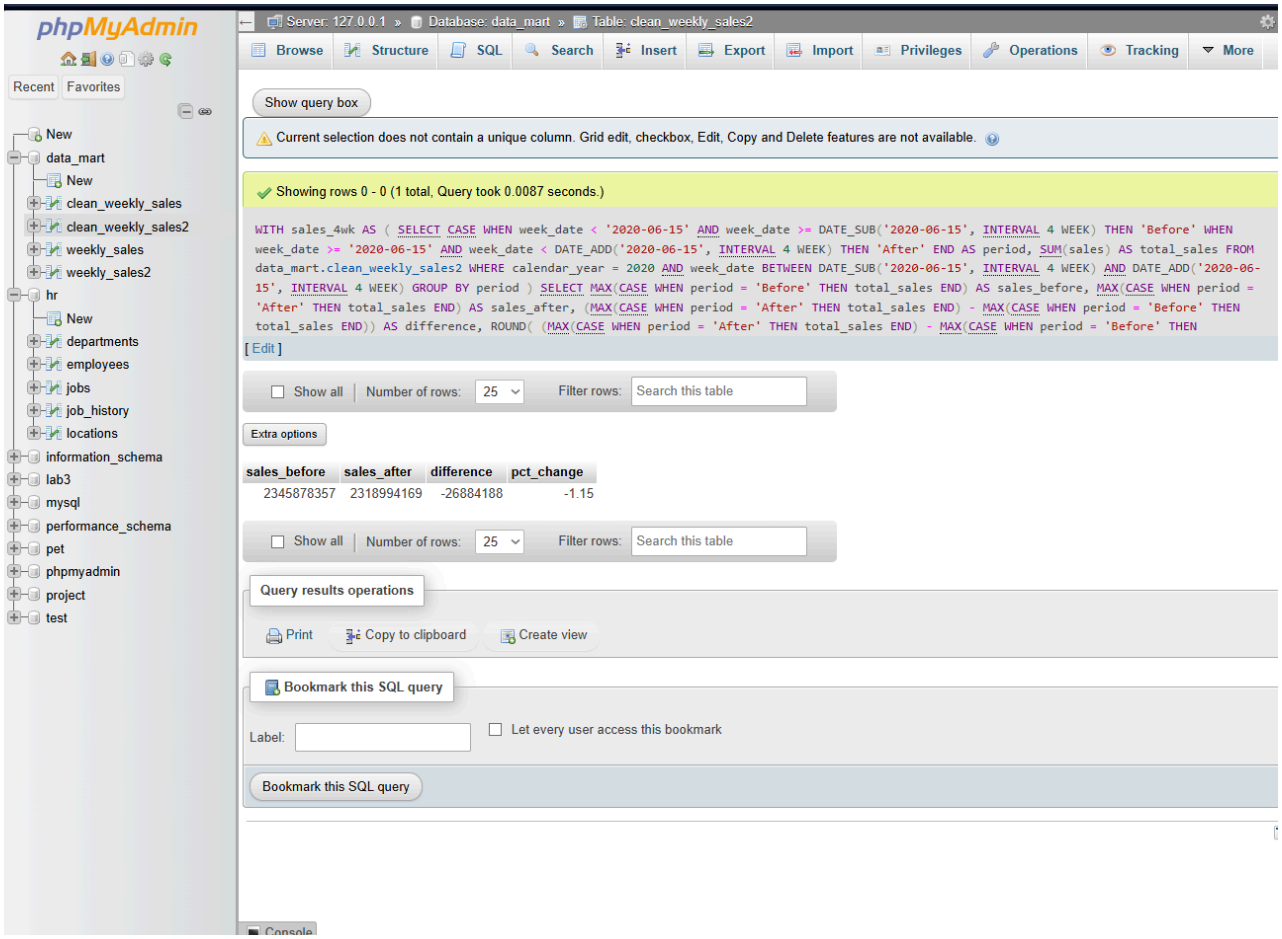
- The baseline date for the packaging change was established as **2020-06-15**.
- Two timeframes were defined:
    - **Short-term impact**: 4 weeks before and 4 weeks after June 15
    - **Medium-term impact**: 12 weeks before and 12 weeks after June 15

- Data was sourced from the clean_weekly_sales2 table within the data_mart schema, containing weekly aggregated sales, transaction, and customer segmentation data.
- SQL queries were executed to:
    - Filter week_date for the desired periods around June 15 for each year
    - Calculate total sales for the "Before" and "After" periods separately
    - Derive the **absolute difference** (sales_after - sales_before) and the **percentage change** ((difference / sales_before) * 100)
- Comparisons were made across three calendar years: **2018**, **2019**, and **2020**.

This method provides both intra-year (before vs after) and inter-year (2020 vs 2018–2019) comparisons.

3. Results for 2020: Immediate Sales Impact of Sustainable Packaging

3.1 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?



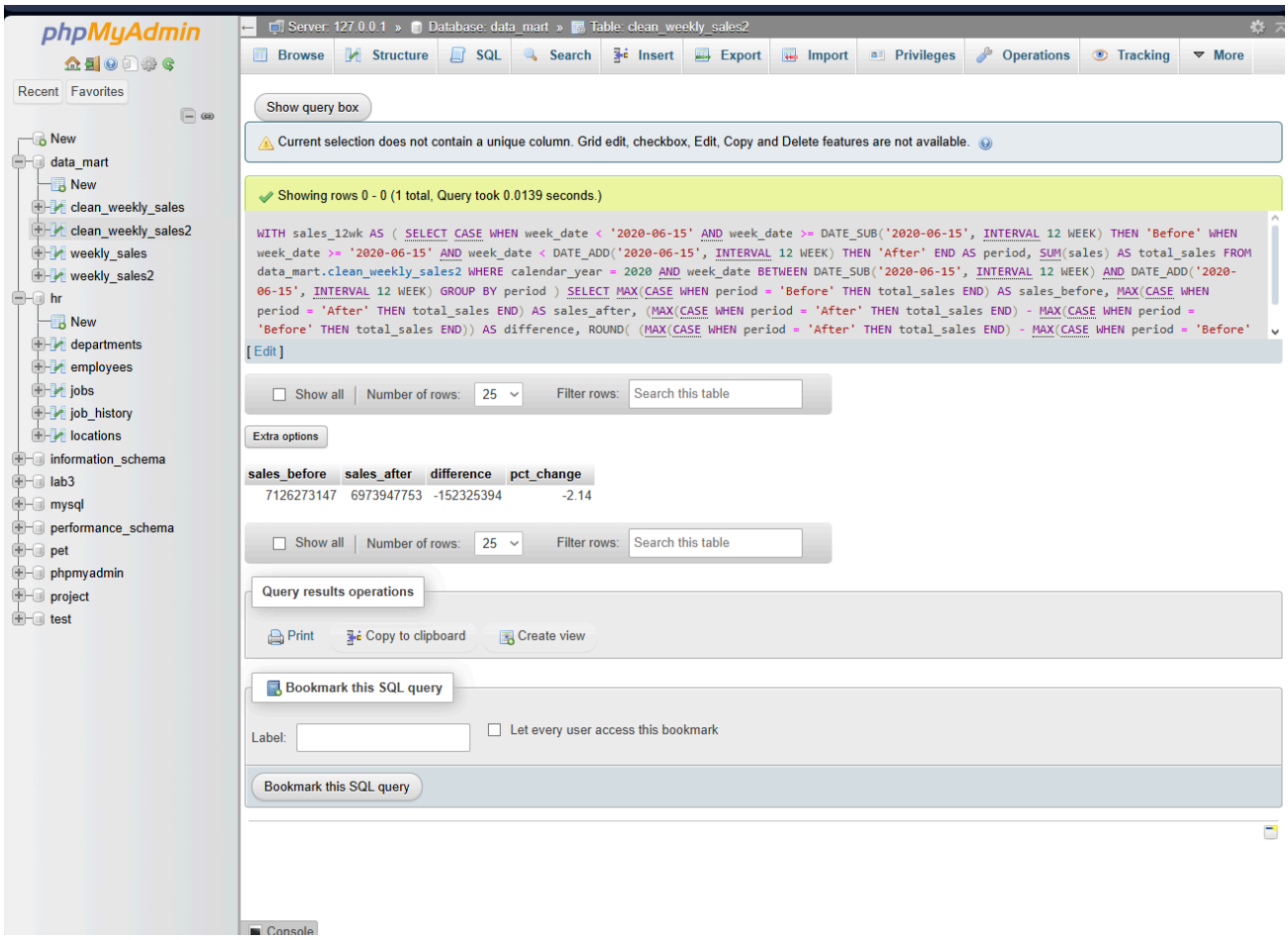
Insight: There was a 1.15% decrease in sales in the 4-week period immediately following the packaging change. This indicates a potentially negative customer reaction or market disruption resulting from the transition.

SQL Used:

```
WITH sales_4wk AS (
  SELECT
    CASE
      WHEN week_date::date < DATE '2020-06-15'
```

```
        AND week_date::date >= (DATE '2020-06-15' - INTERVAL '4 weeks') THEN 'Before'
    WHEN week_date::date >= DATE '2020-06-15'
        AND week_date::date < (DATE '2020-06-15' + INTERVAL '4 weeks') THEN 'After'
  END AS period,
  SUM(sales) AS total_sales
 FROM data_mart.clean_weekly_sales
 WHERE calendar_year = 2020
  AND week_date::date BETWEEN (DATE '2020-06-15' - INTERVAL '4 weeks')
                  AND (DATE '2020-06-15' + INTERVAL '4 weeks')
 GROUP BY period
)
SELECT
 MAX(CASE WHEN period = 'Before' THEN total_sales END) AS sales_before,
 MAX(CASE WHEN period = 'After' THEN total_sales END) AS sales_after,
 (MAX(CASE WHEN period = 'After' THEN total_sales END) -
  MAX(CASE WHEN period = 'Before' THEN total_sales END)) AS difference,
 ROUND(
  ((MAX(CASE WHEN period = 'After' THEN total_sales END) -
    MAX(CASE WHEN period = 'Before' THEN total_sales END))::numeric /
    NULLIF(MAX(CASE WHEN period = 'Before' THEN total_sales END), 0) * 100),
  2
 ) AS pct_change
FROM sales_4wk;
```

3.2. 12-Week Period (Medium-Term Impact)



Insight: While the drop in the 12-week comparison is smaller in relative terms, it still represents a sustained decrease in sales that did not recover post-transition within the medium term.

SQL USED:

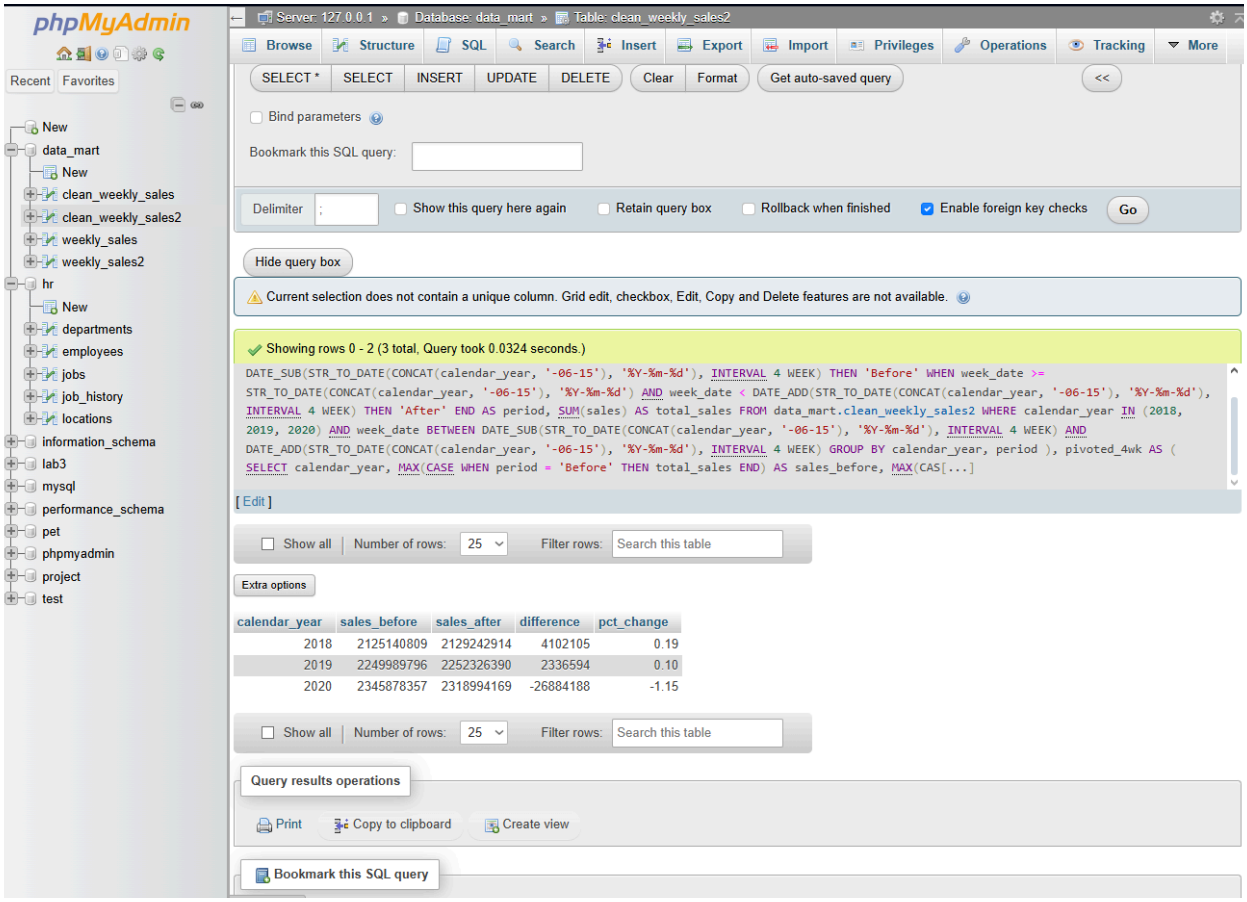
```
WITH sales_12wk AS (
  SELECT
    CASE
      WHEN week_date < DATE '2020-06-15'
          AND week_date >= (DATE '2020-06-15' - INTERVAL '12 weeks') THEN 'Before'
      WHEN week_date >= DATE '2020-06-15'
          AND week_date < (DATE '2020-06-15' + INTERVAL '12 weeks') THEN 'After'
```

```
    END AS period,
    SUM(sales) AS total_sales
  FROM data_mart.clean_weekly_sales
  WHERE calendar_year = 2020
    AND week_date BETWEEN (DATE '2020-06-15' - INTERVAL '12 weeks')
                      AND (DATE '2020-06-15' + INTERVAL '12 weeks')
  GROUP BY period
)
SELECT
  MAX(CASE WHEN period = 'Before' THEN total_sales END) AS sales_before,
  MAX(CASE WHEN period = 'After' THEN total_sales END) AS sales_after,
  (MAX(CASE WHEN period = 'After' THEN total_sales END) -
   MAX(CASE WHEN period = 'Before' THEN total_sales END)) AS difference,
  ROUND(
    ((MAX(CASE WHEN period = 'After' THEN total_sales END) -
      MAX(CASE WHEN period = 'Before' THEN total_sales END))::numeric /
      NULLIF(MAX(CASE WHEN period = 'Before' THEN total_sales END), 0) * 100),
    2
  ) AS pct_change
FROM sales_12wk;
```

3.3 Comparative Analysis: 2018 & 2019 vs 2020

To isolate the effect of the packaging change, the same analysis was performed on historical sales data for 2018 and 2019 during equivalent 4-week and 12-week periods centered around June 15. This helps determine if the decline in 2020 is part of a broader seasonal pattern or unique to the year of implementation.

3.3.1. 4-Week Comparison



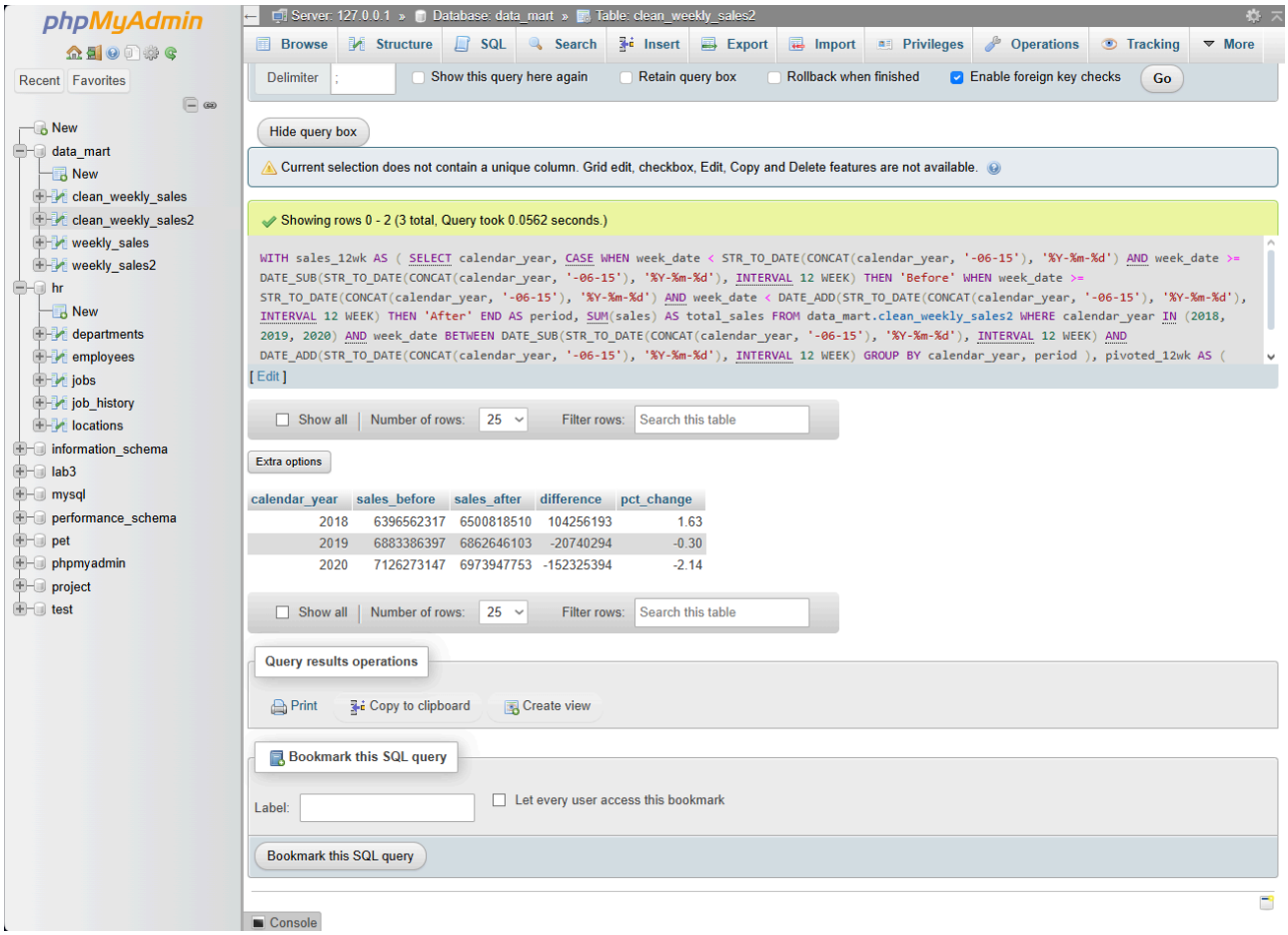
Insight: In both 2018 and 2019, there were slight increases in sales following the same 4-week period of June 15. However, in 2020, when the sustainable packaging was introduced, there was a notable sales drop of 26.88 million VND, representing a 1.15% decline. This deviation may be linked to the initial reaction to packaging changes, potential disruption in customer expectations, or short-term supply chain effects.

SQL USED:

```
WITH sales_4wk AS (
  SELECT
    calendar_year,
    CASE
      WHEN week_date < STR_TO_DATE(CONCAT(calendar_year, '-06-15'), '%Y-%m-%d')
        AND week_date >= DATE_SUB(STR_TO_DATE(CONCAT(calendar_year, '-06-15'),
'%Y-%m-%d'), INTERVAL 4 WEEK) THEN 'Before'
```

```
    WHEN week_date >= STR_TO_DATE(CONCAT(calendar_year, '-06-15'), '%Y-%m-%d')
      AND week_date < DATE_ADD(STR_TO_DATE(CONCAT(calendar_year, '-06-15'),
'%Y-%m-%d'), INTERVAL 4 WEEK) THEN 'After'
  END AS period,
  SUM(sales) AS total_sales
 FROM data_mart.clean_weekly_sales
 WHERE calendar_year IN (2018, 2019, 2020)
  AND week_date BETWEEN DATE_SUB(STR_TO_DATE(CONCAT(calendar_year,
'-06-15'), '%Y-%m-%d'), INTERVAL 4 WEEK)
                AND DATE_ADD(STR_TO_DATE(CONCAT(calendar_year, '-06-15'),
'%Y-%m-%d'), INTERVAL 4 WEEK)
 GROUP BY calendar_year, period
),
pivoted_4wk AS (
 SELECT
  calendar_year,
  MAX(CASE WHEN period = 'Before' THEN total_sales END) AS sales_before,
  MAX(CASE WHEN period = 'After' THEN total_sales END) AS sales_after
 FROM sales_4wk
 GROUP BY calendar_year
)
SELECT
 calendar_year,
 sales_before,
 sales_after,
 (sales_after - sales_before) AS difference,
 ROUND((sales_after - sales_before) / NULLIF(sales_before, 0) * 100, 2) AS pct_change
FROM pivoted_4wk
ORDER BY calendar_year;
```

3.3.2. 12-Week Comparison



Insight:

- In **2018**, sales **increased significantly** after mid-June (up 1.63%), showing a healthy growth trend.
- **2019** showed a minor dip (-0.30%) but still fairly stable.
- **2020**, however, saw a sharp **decline of over 152 million VND**, a **2.14% drop**, following the packaging change.

This reinforces the pattern from the 4-week view and highlights a **larger, longer-term negative impact** in 2020 compared to previous years.

SQL USED:

```
WITH sales_12wk AS (
  SELECT
    calendar_year,
    CASE
      WHEN week_date < TO_DATE(calendar_year || '-06-15', 'YYYY-MM-DD')
          AND week_date >= (TO_DATE(calendar_year || '-06-15', 'YYYY-MM-DD') -
INTERVAL '12 weeks') THEN 'Before'
      WHEN week_date >= TO_DATE(calendar_year || '-06-15', 'YYYY-MM-DD')
          AND week_date < (TO_DATE(calendar_year || '-06-15', 'YYYY-MM-DD') + INTERVAL
'12 weeks') THEN 'After'
    END AS period,
    SUM(sales) AS total_sales
  FROM data_mart.clean_weekly_sales
  WHERE calendar_year IN (2018, 2019, 2020)
    AND week_date BETWEEN (TO_DATE(calendar_year || '-06-15', 'YYYY-MM-DD') -
INTERVAL '12 weeks')
                  AND (TO_DATE(calendar_year || '-06-15', 'YYYY-MM-DD') + INTERVAL '12
weeks')
  GROUP BY calendar_year, period
),
pivoted_12wk AS (
  SELECT
    calendar_year,
    MAX(CASE WHEN period = 'Before' THEN total_sales END) AS sales_before,
    MAX(CASE WHEN period = 'After' THEN total_sales END) AS sales_after
  FROM sales_12wk
  GROUP BY calendar_year
)
SELECT
  calendar_year,
  sales_before,
```

```
  sales_after,
  (sales_after - sales_before) AS difference,
  ROUND(((sales_after - sales_before)::numeric / NULLIF(sales_before, 0) * 100), 2) AS
pct_change
FROM pivoted_12wk
ORDER BY calendar_year;
```

III. Conclusion

The analysis of Data Mart's sales from 2018 to 2020 reveals several key patterns in platform performance, customer demographics, and the impact of the sustainable packaging change introduced on June 15, 2020. Retail remains the dominant channel, consistently accounting for over 96% of sales, though Shopify shows a slow but steady growth trend. Demographic insights indicate that Retirees, particularly in the Families and Couples categories, are the largest known contributors to Retail revenue, while a significant 40% of sales come from customers with unknown profiles — representing both a data gap and a growth opportunity.

The before-and-after impact assessment shows that the packaging change coincided with a 1.15% sales drop in the immediate 4 weeks and a 2.14% decline over 12 weeks, diverging from positive or stable trends in prior years. This suggests a short- to medium-term negative reaction, potentially from customer perception changes or operational disruptions during the transition.

To sustain growth while advancing sustainability goals, Data Mart should strengthen customer profiling, target marketing toward high-value segments like Retirees and Middle Aged Families, and proactively manage communication around operational changes to reduce sales volatility. Expanding Shopify's role could also diversify revenue streams and mitigate Retail channel risks in future transitions.