EASTERN INTERNATIONAL UNIVERSITY

AUGUST, 2025

CASESTUDY #5

DATA MART

MIS 443 - BUSINESS DATA MANAGEMENT

LECTURER: MR. DANG THAI DOAN

Quarter 4, 2024-2025

TEAM MEMBERS

HUYNH THUY BAO TRAM	2132300228
NGUYEN TRAN NGOC THY	2232300307
HUYNH TRUNG HAU	2132309003
NGUYEN QUANG TRUONG	2132309001

TABLE
of contents

1. Introduction

- "Data Mart" is a supermarket chain with both physical stores and an online platform (Shopify)
- "This case study" explores consumer behavior and the impact of switching to sustainable packaging
- "Packaging change" took effect on June 15, 2020
- "SQL analysis" was used to uncover trends, seasonality, and behavioral shifts

Database Overview

~17,117 transactions

Original Columns (7):

- "week_date" weekly sales records across platforms – 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 (2020 is a leap year)
- "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)

2. Objectives

1. Clean and transform the original weekly sales data into a structured format
2. Explore customer behavior across time, region, platform, and demographics
3. Assess the before-and-after impact of sustainable packaging policy (using 2020-06-15 as the baseline)
4. Provide actionable insights and recommendations for Data Mart’s management team

3. DATA CLEANING

- Purpose: Prepare the data for analysis by ensuring accuracy, consistency, and structure.
- Focus: Weekly sales data.
- Steps Covered:
    a. Date format conversion
    b. Creation of temporal variables
    c. Customer data classification
    d. Handling NULL values
    e. Calculation of average transactions

Conversion of week_date to Date Format

- Original Format: DD/MM/YY (text)
- New Format: YYYY/MM/DD (date)
- Reason: Ensures time-based analysis can be performed accurately.
- SQL Used:

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

Creation of Temporal Variables

- New Variables Created:
  - week_number: Extracted from week_date (ISO week number)
  - month_number: Extracted from week_date (1-12)
  - calendar_year: Extracted year from week_date
- Reason: Enables time-series analysis for weekly, monthly, and 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
```

Classification of Customer Data

Derived Columns:

- age_band: Categorizes customers based on the last character of the segment code (e.g., Young Adults, Retirees).
- demographic: Categorizes customers based on the first character of the segment code (e.g., Couples, Families).

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
```

Handling NULL Values & Calculation of avg_transaction

- **Issue**: Missing data in segment, age_band, and demographic columns.
- **Solution**: Replaced all NULL values with "unknown" to ensure consistency.
- **SQL Used**:

```
CASE WHEN segment = 'null' THEN 'unknown' ELSE segment END AS segment
```

- **Metric Created:** Average transaction value for each week.
- **Formula**: avg_transaction = sales / transactions
- **Reason**: Provides insights into customer spending behavior.
- **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

4. DATA EXPLORATION

1. What day of the week is used for each week_date value?
2. What range of week numbers are missing from the dataset?
3. How many total transactions were there for each year in the dataset?
4. What is the total sales for each region for each month?
5. What is the total count of transactions for each platform
6. What is the percentage of sales for Retail vs Shopify for each month?
7. What is the percentage of sales by demographic for each year in the dataset?
8. Which age_band and demographic values contribute the most to Retail sales?
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?

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

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).

```sql
-- PART B
-- Q1: What day of the week is used for each week_date?
SELECT
    DISTINCT week_date,
    TO_CHAR(week_date, 'Day') AS day_of_week
FROM data_mart.clean_weekly_sales
ORDER BY week_date;

-- Monday is used for the week_date value.
```

	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

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

The dataset should contain 53 week numbers (1–53) for the year, but only 24 are present (13-36).

- A total of 29 week numbers are missing.
- Examples of missing week numbers: 1-12; 37-53

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

```sql
-- Q2: What range of week numbers are missing from the dataset?

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;

--- The dataset is missing a total of 29 week_number records.
```

Data Output | Messages | Notifications

Showing rows: 1 to 29

	week_number integer
11	11
12	12
13	37
14	38
15	39
16	40
17	41
18	42

SQL CODE

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

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

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

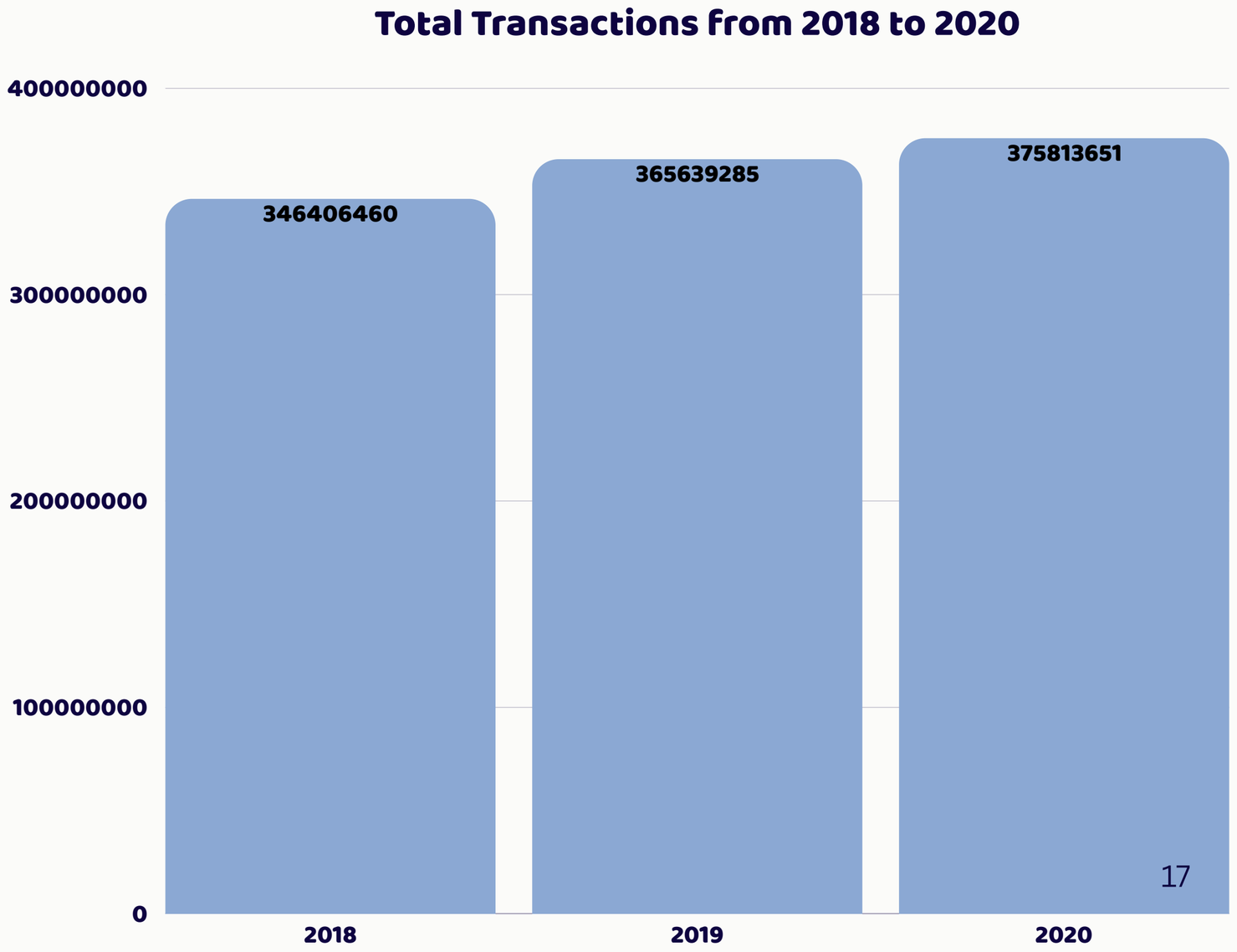
```
SELECT
  region,
  month_number,
  SUM(sales) AS total_sales
FROM data_mart.clean_weekly_sales
GROUP BY region, month_number
ORDER BY region, month_number;
```

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

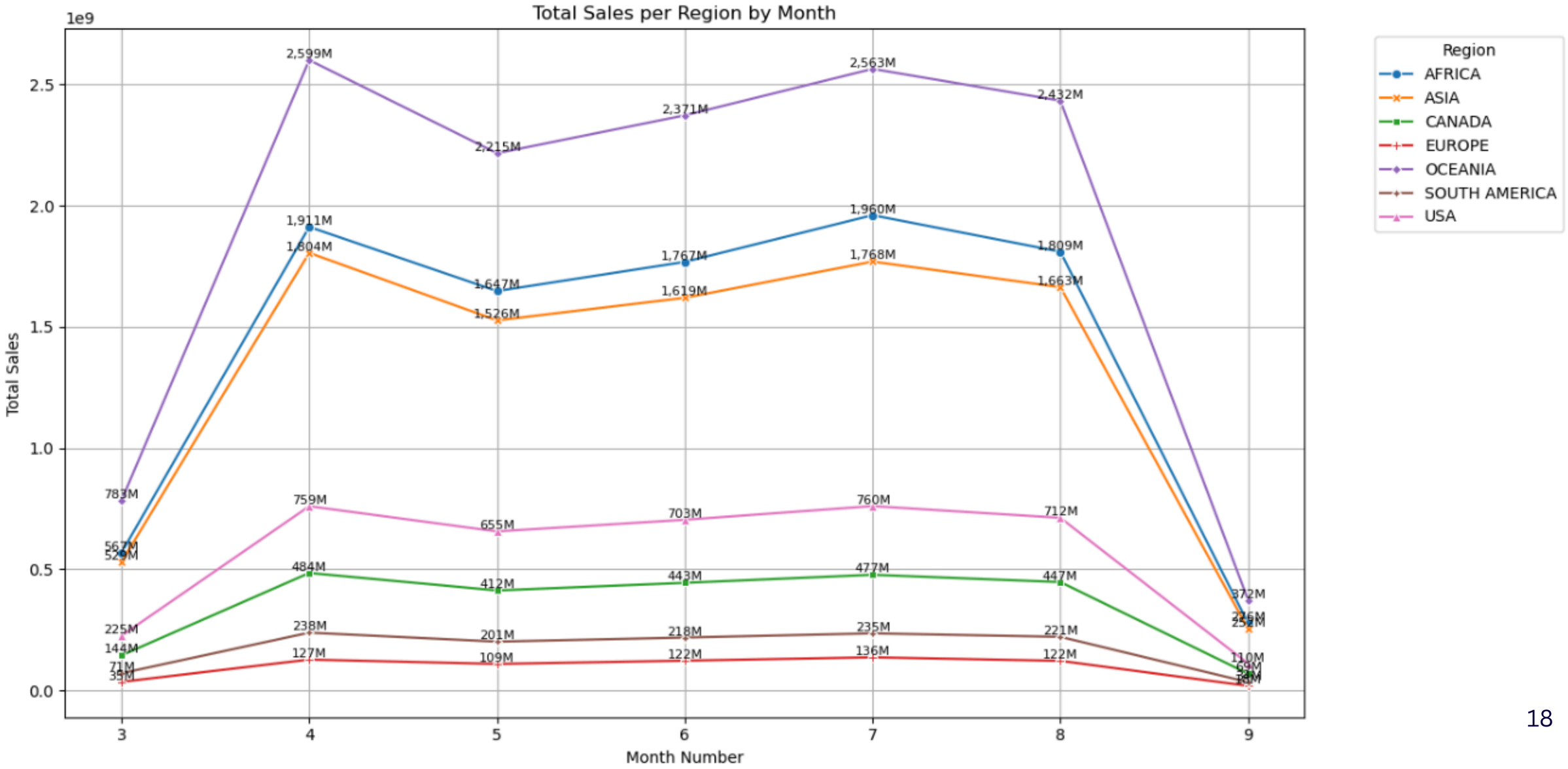
From 2018 to 2020, total transactions increased by approximately 8.5%.

This trend may reflect:

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



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



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

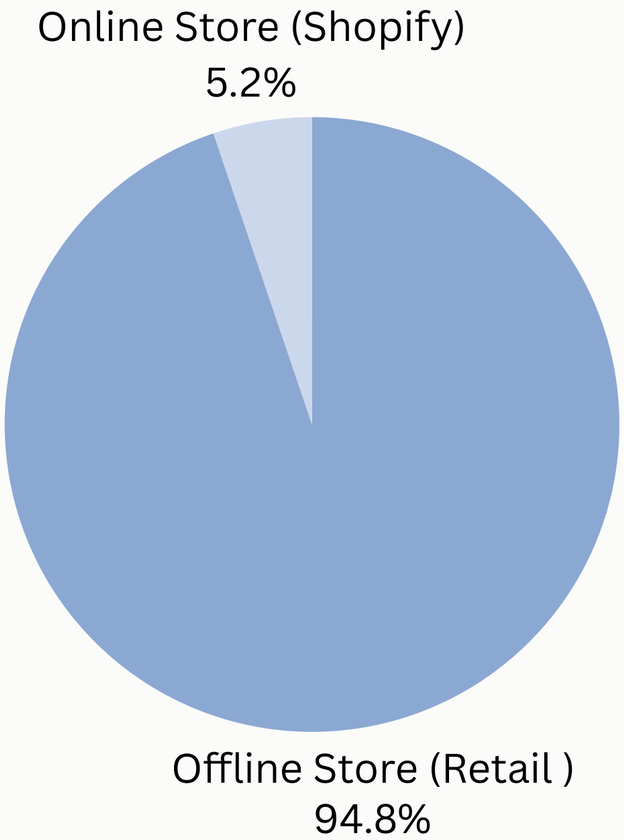
Insight	Explanation
1. Oceania dominates	Oceania consistently shows the highest sales among all regions in every month except September (month 9), where all regions dropped drastically.
2. Peak in Month 4 (April)	Most regions (especially Oceania, Africa, and Asia) saw peak sales in month 4, suggesting a seasonal (e.g., Easter in Europe/USA, Golden Week in parts of Asia, end of fiscal year sales in some countries) or promotional spike - many companies run Q2 kickoff sales or clearance sales after Q1, boosting April numbers.
3. Sudden drop in Month 9 (September)	All regions experienced a sharp sales drop in month 9, possibly due to incomplete data. (stop at 31/8/2020)
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.

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

```sql
--- Q5. What is the total count of transactions
SELECT
    platform,
    SUM(transactions) AS total_transactions
FROM data_mart.clean_weekly_sales
GROUP BY platform;
```

Data Output | Messages | Notifications

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



1. Retail dominates the transaction volume with nearly 95% of all transactions → the primary sales channel.
2. Shopify only accounts for ~5% → likely a newer channel or supplemental to in-store sales.

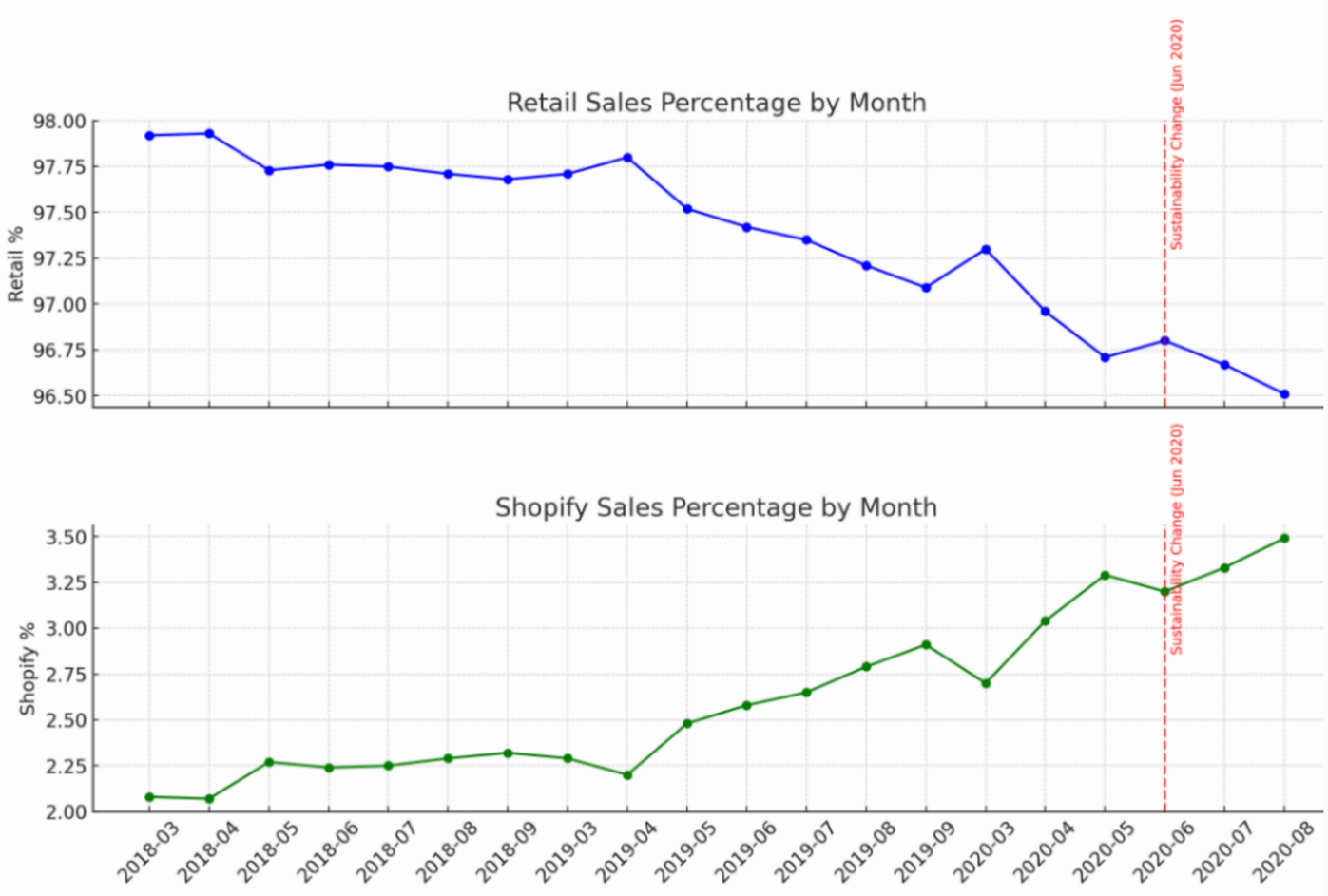
- If the company wants to increase online presence → investment in Shopify optimization and marketing
- Maintain strong in-store operations → logistics, and customer service to protect the Retail base.

SQL CODE

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

```
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;
```

6. Retail vs Shopify Sales Share for each month from 2018 to 2020



- Retail Dominance:
  - Retail holds 96.5%–97.9% share throughout the period.
  - Shopify remains small (~2.1%-3.5%)
- Trend Over Time:
  - Gradual decline in Retail share, increase in Shopify share.
  - Post June 2020 sustainability change:
    - Shopify growth accelerates from 3.20% → 3.49% (Jun–Aug 2020).

SQL CODE

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

```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;
```

7. Sales Distribution by Demographic (2018–2020)

Overall: Growth in Couples, stability in Families, and a reduction in Unknowns indicate better customer targeting and profiling.

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

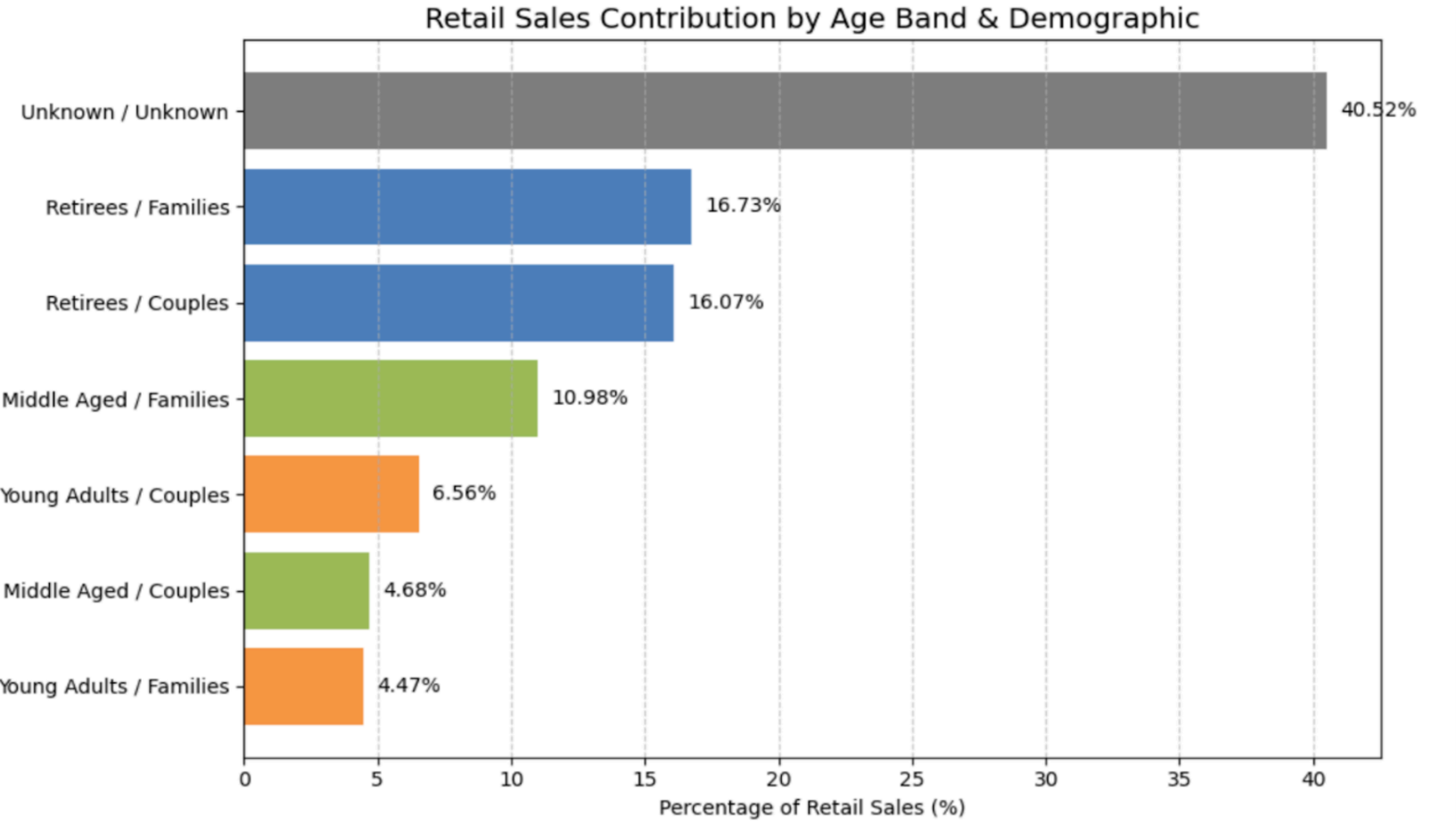
- Couples: Increased by 2.4%, becoming a stronger customer segment.
- Families: Stable around 32%, showing consistent engagement.
- Unknown: Declined about 3%, suggesting improved data collection or profiling.

SQL CODE

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

```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;
```

8. Top Contributors to Retail Sales by Age & Demographic



- Unknown / Unknown is the largest share at **40.52%**, likely due to missing profile data.

- Top identified groups:
  Retirees / Families – **16.73%**
  Retirees / Couples – **16.07%**
- Together, Retirees make up ~**33%** of all customers.

- Middle Aged Families – **10.98%**.
- Young Adult Couples – **6.56%**.

- Young Adult Families – **4.47%**, smallest contributor.

Implication:

- **Targeted marketing to** Retirees & Middle Aged Families could drive strong returns.
- I**mproved customer data collection** could unlock deeper insights into the large “Unknown” segment.

5. 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?
2. What about the entire 12 weeks before and after?
3. How do the sale metrics for these 2 periods before and after compare with the previous years in 2018 and 2019?

FOCUS POINTS

- Query Structure and Logic:

The use of CTEs (Common Table Expressions) like sales_4wk and sales_12wk organizes the data into "Before" and "After" periods, making the analysis clear and modular.

Talking Point: CTEs improve readability and allow for step-by-step data processing, which is useful for debugging or modifying the queries.

- Dynamic Date Handling:

Queries 3 and 4 use STR_TO_DATE(CONCAT(calendar_year, '-06-15'), '%Y-%m-%d') to dynamically set the reference date (June 15) based on the calendar_year, enabling analysis across multiple years.

Talking Point: This approach, allowing the same query to be reused for different years without hardcoding dates.

- Percentage Change Calculation:

The ROUND((sales_after - sales_before) / NULLIF(sales_before, 0) * 100, 2) formula handles division by zero and provides a precise percentage change.

Talking Point: the importance of this calculation for business decision-making and how NULLIF prevents errors when sales_before is zero.

4-Week Analysis for 2020

Extra options

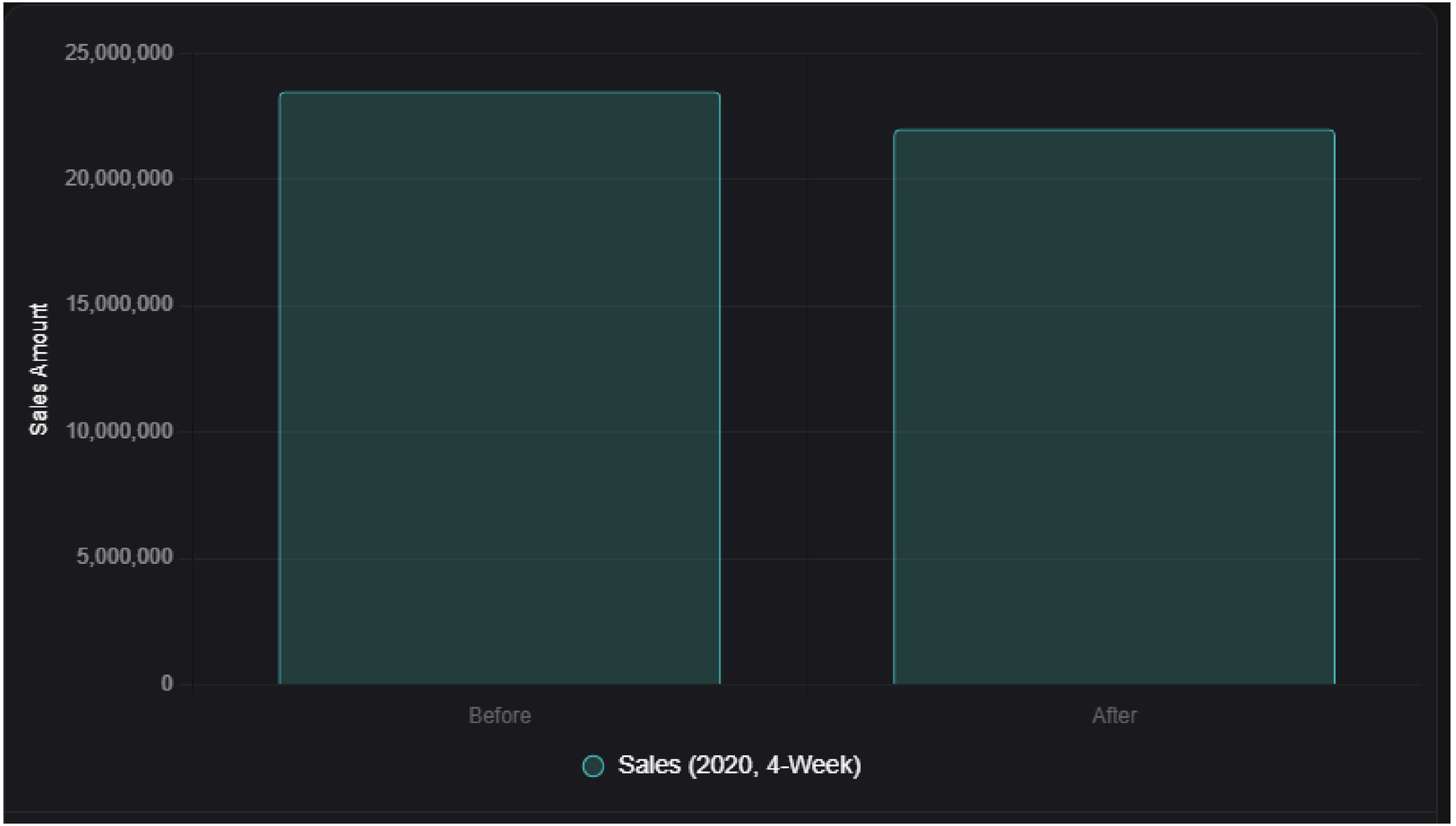
sales_before	sales_after	difference	pct_change
2345878357	2318994169	-26884188	-1.15

Show all | Number of rows: 25 Filter rows: Search this table

Our data shows total sales before June 15, 2020, were 2,345,878,357, while sales after dropped to 2,318,994,169. This results in a difference of -26,884,188 with a percentage change of -1.15%.

The significant -1.15% decline in sales post-June 15, 2020, stands out. This could be linked to global events like the COVID-19 pandemic, which began impacting markets around mid-2020.

VISUALIZATION



12-Week Analysis for 2020

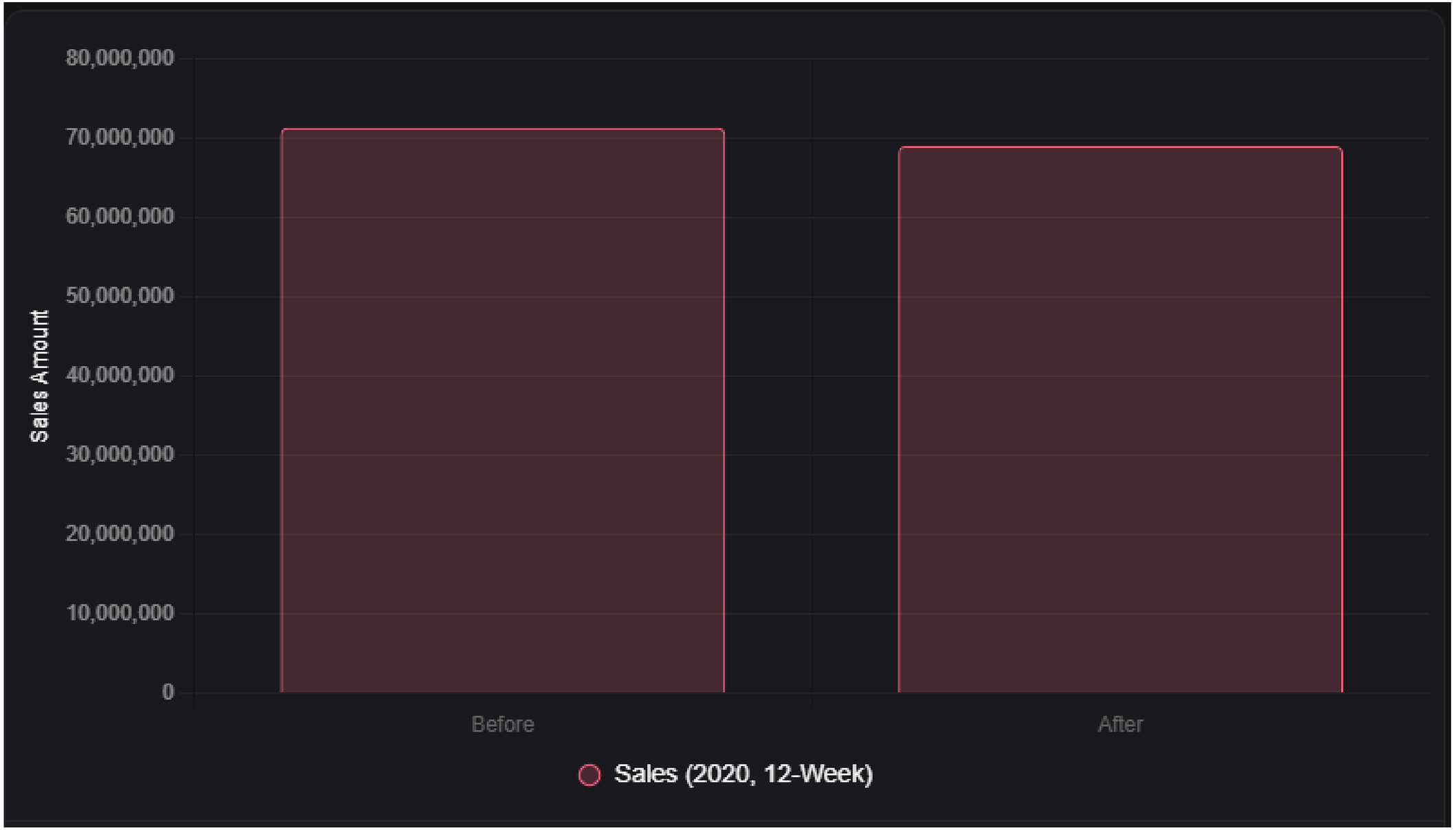
sales_before	sales_after	difference	pct_change
7126273147	6973947753	-152325394	-2.14

Show all | Number of rows: 25 Filter rows: Search this table

The data indicates sales before June 15, 2020, were 7,126,273,147, dropping to 6,973,947,753 afterward. This gives a difference of -152,325,394 and a percentage change of -2.14%.

Insight: The -2.14% decline over 12 weeks is more severe than the 4-week drop. This suggests a not possible stabilization or recovery after an initial sharp fall, possibly due to the company can not adapt to the market changes.

VISUALIZATION



4-Week Analysis for 2020 compared to 2018,2019

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

calendar_year	sales_before	sales_after	difference	pct_change
2018	2125140809	2129242914	4102105	0.19
2019	2249989796	2252326390	2336594	0.10
2020	2345878357	2318994169	-26884188	-1.15

In 2018, sales went from 2,125,140,809 to 2,129,242,914 (+4,102,105, 0.19%). In 2019, they rose from 2,249,989,796 to 2,252,326,390 (+2,336,594 +0.10%). In 2020, they fell from 2,345,878,357 to 2,318,994,169 (-26,884,188 -1.15%).

Insights:

"The 2020 drop of 1.15% is a stark contrast to the near-stable 2019 and slight growth in 2018. This anomaly likely reflects the economic disruption of 2020."

12-Week Analysis for 2020 compared to 2018,2019

Show all | Number of rows: 25 | Filter rows: Search this table

Extra options

calendar_year	sales_before	sales_after	difference	pct_change
2018	6396562317	6500818510	104256193	1.63
2019	6883386397	6862646103	-20740294	-0.30
2020	7126273147	6973947753	-152325394	-2.14

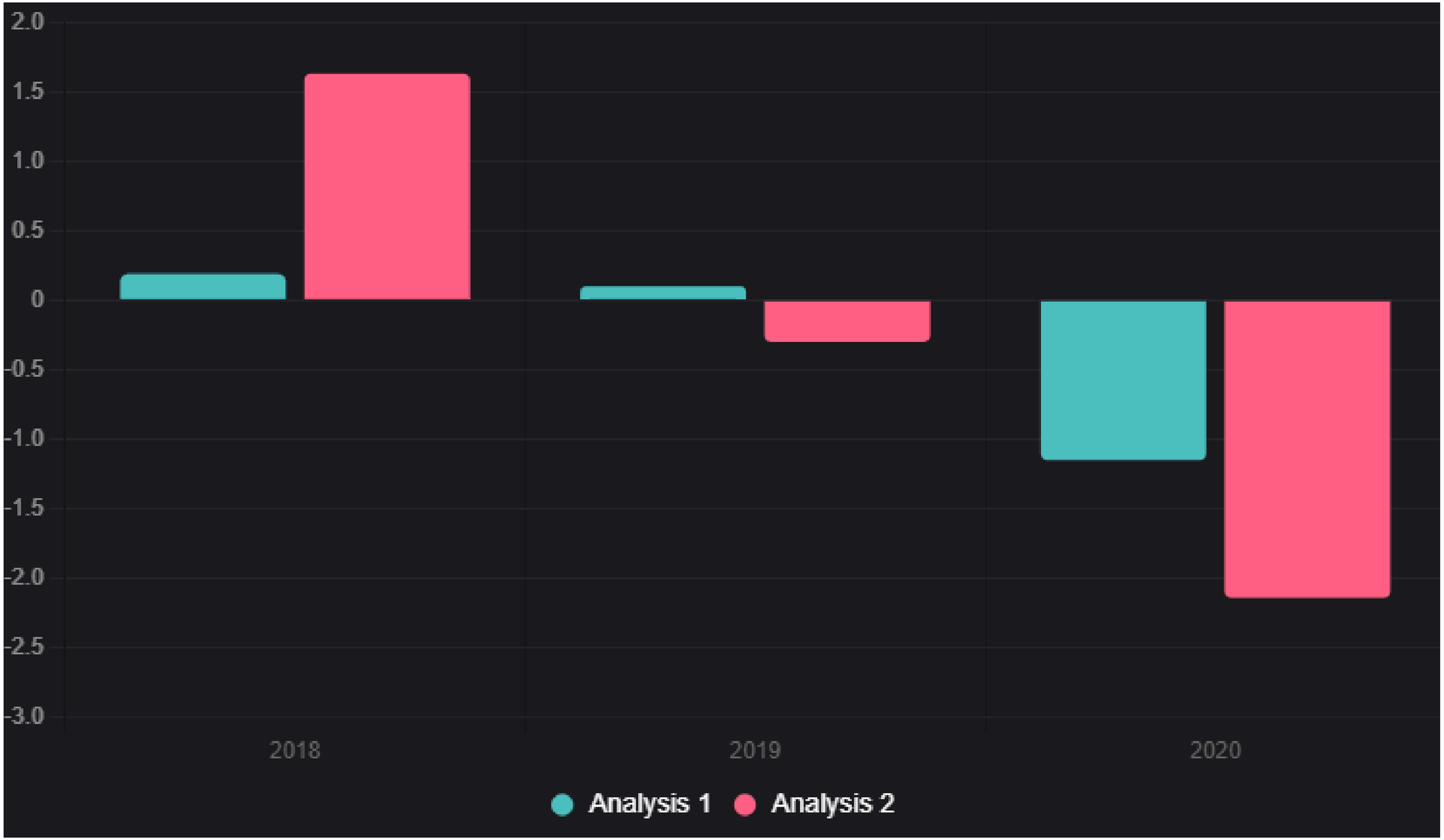
Show all | Number of rows: 25 | Filter rows: Search this table

In 2018, sales increased from 6,396,562,317 to 6,500,818,510 (+104,256,193 +1.63%). In 2019, they dipped from 6,883,386,397 to 6,862,646,103 (-207,402,294 -0.30%). In 2020, they dropped from 7,126,273,147 to 6,973,947,753 (-152,325,394 ,-2.14%).

Insights

"The 2020 decline of -2.14% is significant and even more drastic than the 4-week drop. Meanwhile, 2018 shows strong growth, and 2019 a slight dip, indicating varied yearly dynamics."

VISUALIZATION



6. Conclusions

Overall Findings

- Retail dominance: Retail drives over 96% of sales; Shopify is small but steadily growing.
- Customer segments: Retirees (Families & Couples) are the top known contributors; 40% of sales come from customers with unknown profiles, showing a major data opportunity.
- Packaging impact: Sustainable packaging change (June 15, 2020) led to a short-term sales drop of 1.15% and a medium-term decline of 2.14%, unlike previous years' stable or positive trends.

Recommendations

1. Improve customer profiling to reduce "unknown" share.

2. Focus marketing on Retirees and Middle Aged Families.

3. Strengthen change communication to minimize sales disruption.

4. Expand Shopify to diversify revenue and reduce channel risk.

THANK YOU!