**Case Study #5 - Data Mart**

Danny Ma · June 20, 2021



**Introduction**

Data Mart is Danny’s latest venture and after running international operations for his online supermarket that specialises in fresh produce - Danny is asking for your support to analyse his sales performance.

In June 2020 - large scale supply changes were made at Data Mart. All Data Mart products now use sustainable packaging methods in every single step from the farm all the way to the customer.

Danny needs your help to quantify the impact of this change on the sales performance for Data Mart and it’s separate business areas.

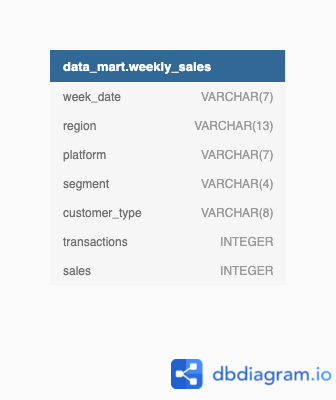
The key business question he wants you to help him answer are the following:

* What was the quantifiable impact of the changes introduced in June 2020?
* Which platform, region, segment and customer types were the most impacted by this change?
* What can we do about future introduction of similar sustainability updates to the business to minimise impact on sales?

**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 |

* Add a new demographic column using the following mapping for the first letter in the segment values:

|  |  |
| --- | --- |
| 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

**-- 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;**

**2. Data Exploration**

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;**



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

**WITH week\_number\_cte AS (**

**SELECT GENERATE\_SERIES(1,52) AS week\_number**

**)**

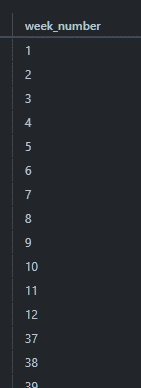
**SELECT DISTINCT week\_no.week\_number**

**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`.**



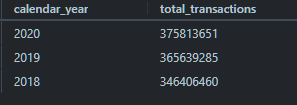
Note: make sure that you have retrieved 28 rows!

1. **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**

**ORDER BY total\_transactions DESC ;**



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

**SELECT**

**month\_number,**

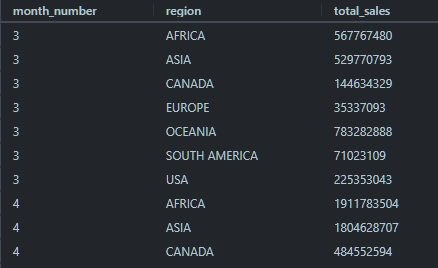
**region,**

**SUM(sales) AS total\_sales**

**FROM clean\_weekly\_sales**

**GROUP BY month\_number, region**

**ORDER BY month\_number, region;**



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

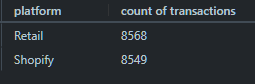
**SELECT**

**platform,**

**SUM(transactions) AS total\_transactions**

**FROM clean\_weekly\_sales**

**GROUP BY platform;**



1. **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**

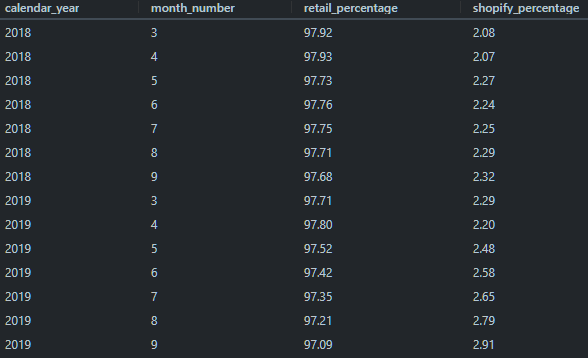
**WHEN platform = 'Shopify' THEN monthly\_sales ELSE NULL END)**

**/ SUM(monthly\_sales),2) AS shopify\_percentage**

**FROM monthly\_transactions**

**GROUP BY calendar\_year, month\_number**

**ORDER BY calendar\_year, month\_number;**



1. **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**

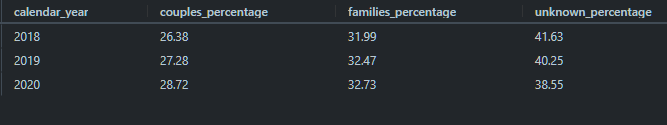
**(CASE**

**WHEN demographic = 'unknown' THEN yearly\_sales ELSE NULL END)**

**/ SUM(yearly\_sales),2) AS unknown\_percentage**

**FROM demographic\_sales**

**GROUP BY calendar\_year;**



1. **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 (),**

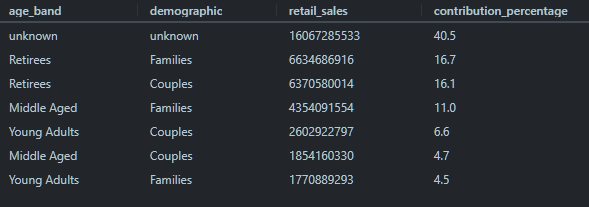
**1) AS contribution\_percentage**

**FROM clean\_weekly\_sales**

**WHERE platform = 'Retail'**

**GROUP BY 1,2**

**ORDER BY Retail\_Sales Desc**



**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;**

* 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:

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 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**

**)**

**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;**



**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.

1. **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 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**

**)**

**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;**



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.

1. **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 1

)

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;



* **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**

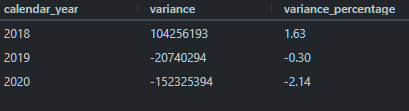
**GROUP BY calendar\_year**

**)**

**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;**



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 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?

* 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

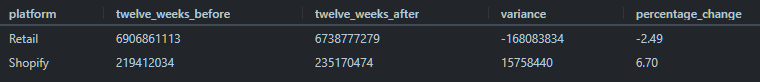
ORDER BY percentage\_change;

By Region:



* 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:



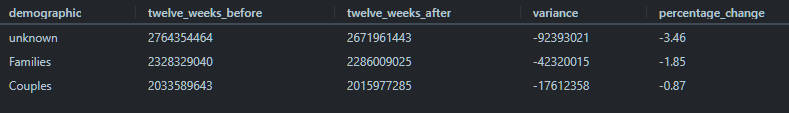
**Retail shopping had the highest negative impact between platforms**

By Age Bound:



**Unknown** age\_band was **the** most impacted negatively among the age bands

By Demographic:



Unknown demographic was the most negatively impacted demographic