

## 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.

## 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!

data_mart.weekly_sales	
week_date	VARCHAR(7)
region	VARCHAR(13)
platform	VARCHAR(7)
segment	VARCHAR(4)
customer_type	VARCHAR(8)
transactions	INTEGER
sales	INTEGER



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

#### 2. 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?

