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`

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.

Example Rows

10 random rows are shown in the table output below from `data_mart.weekly_sales`:

week_date	region	platform	segment	customer_type	transactions	sales
9/9/20	OCEANIA	Shopify	C3	New	610	1100033.89
29/7/20	AFRICA	Retail	C1	New	110692	3053771.19
22/7/20	EUROPE	Shopify	C4	Existing	24	8101.54
13/5/20	AFRICA	Shopify	null	Guest	5287	1003301.37
24/7/19	ASIA	Retail	C1	New	127342	3151780.41
10/7/19	CANADA	Shopify	F3	New	51	8844.93
26/6/19	OCEANIA	Retail	C3	New	152921	5551385.36

29/5/19	SOUTH AMERICA	Shopify	null	New	53	10056.2
22/8/18	AFRICA	Retail	null	Existing	31721	1718863.58
25/7/18	SOUTH AMERICA	Retail	null	New	2136	81757.91

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

- 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

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?