**Exploratory Data Analysis:**

1. **Dataset analysis and Data Preparation**

* **Customers dataset**: Information about customers

|  |  |  |
| --- | --- | --- |
| ***Column name*** | ***Type of data*** | ***Data preparation performed*** |
| CustomerKey | Primary key to identify each customer |  |
| Gender | Category data |  |
| Name |  |  |
| City |  |  |
| State Code | Code to identify the state | Dropped |
| State | Categorical data |  |
| Zip Code |  |  |
| Country | Categorical data |  |
| Continent | Categorical data |  |
| Birthday |  | Changed to datatype datetime |

* **Products dataset**: Details of products

|  |  |  |
| --- | --- | --- |
| ***Column name*** | ***Type of data*** | ***Data preparation performed*** |
| ProductKey | Primary key to identify the different products |  |
| Product Name |  |  |
| Brand | Categorical data |  |
| Color | Categorical data |  |
| Unit Cost USD |  | Changed to datatype numeric after removing special characters |
| Unit Price USD |  | Changed to datatype numeric after removing special characters |
| SubcategoryKey |  |  |
| Subcategory | Categorical data |  |
| CategoryKey |  |  |
| Category | Categorical data |  |

* **Sales dataset**: Transaction data

|  |  |  |
| --- | --- | --- |
| ***Column name*** | ***Type of data*** | ***Data preparation performed*** |
| Order Number | Unique value to identify the order. |  |
| Line Item | In conjugation with order number identifies the rows in the dataframe/table |  |
| Order Date |  | Changed to datatype datetime |
| Delivery Date |  | Changed to datatype datetime  Interpolation is used to fill the missing values. |
| CustomerKey | Foreign key to identify the customers |  |
| StoreKey | Foreign key to identify the store |  |
| ProductKey | Foreign key to identify the product |  |
| Quantity |  |  |
| Currency Code | Code to identify the exchange rate along with the order\_date |  |

* **Stores dataset**: Information on store locations and attributes

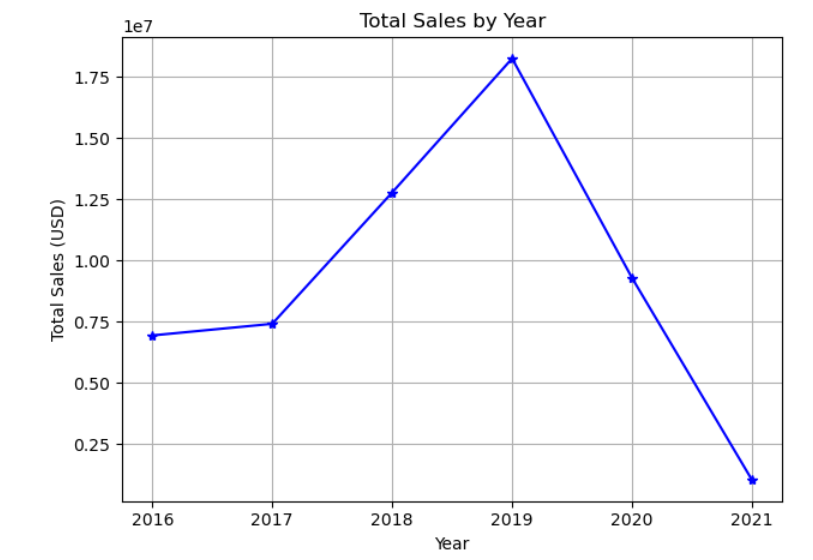
|  |  |  |
| --- | --- | --- |
| ***Column name*** | ***Type of data*** | ***Data preparation performed*** |
| StoreKey | Primary key to identify the different stores |  |
| Country | Categorical data |  |
| State | Analogous to the store name as only one entry is available for each store key |  |
| Square Meters | Numeric data | Assigned 0 for online  Changed to it datatype rather than float |
| Open Date |  | Changed to datatype datetime |

* **Exchange rates dataset**: Information about the exchange rates for the various currencies for the given time period

|  |  |  |
| --- | --- | --- |
| ***Column name*** | ***Type of data*** | ***Data preparation performed*** |
| Date |  | Changed to datatype datetime |
| Currency |  |  |
| Exchange |  |  |

The aforementioned datasets are cleaned and the preprocessed data are written into .csv files.

Further these datasets can be combined to perform analysis using matplotlib. Appropriately joining the datasets allows us to determine the total sales and plot them year-wise.

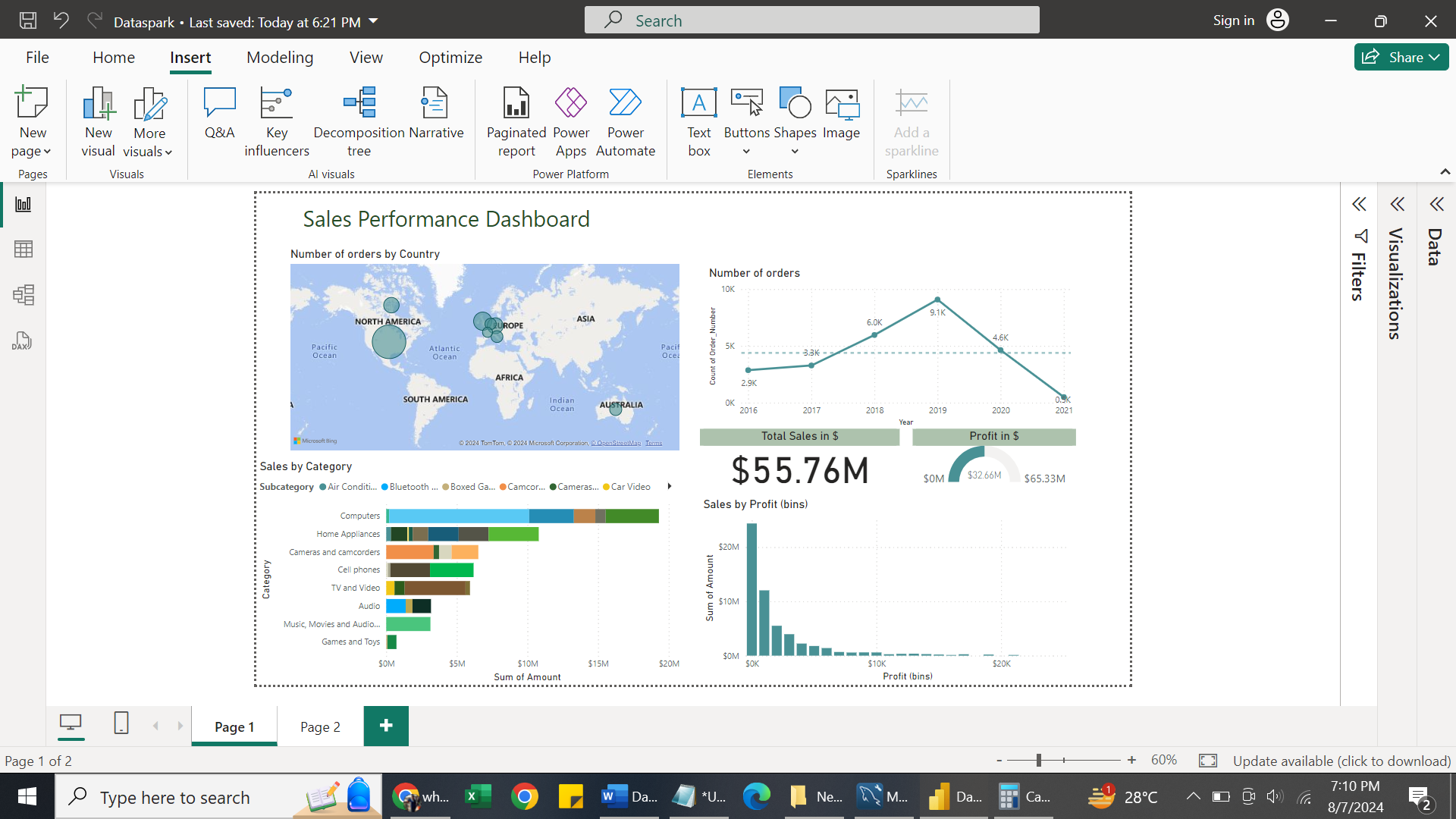


1. **Load data**

5 Tables are created to load the cleaned and preprocessed data. We can see that star schema is formed with the given tables with the Sales table being the fact table and the rest are dimension tables. Sales table can be joined with the exchange rates table considering the order date & currency code columns in the sales table and date & currency code column in the exchange rates table. So appropriate keys should be set accordingly.

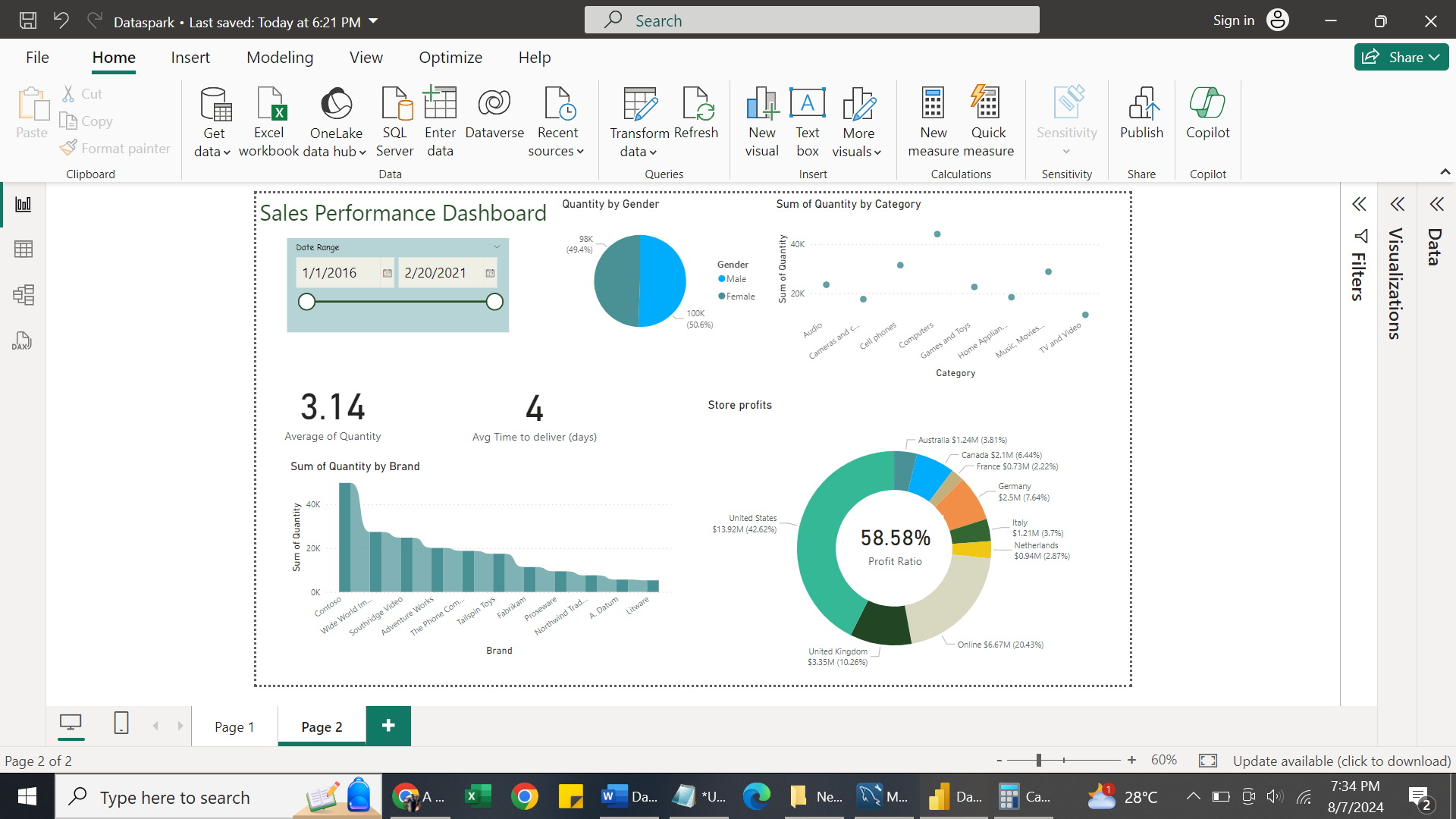
**Visualizations:**

The below visualizations are obtained after loading the dtaa stored in sql tables.



Following observations can be made from the above visualization

1. United States customers have ordered the most no. of products followed by United Kingdom which very much trails behind by over 10k. On the other hand, the least number of orders were from France and Netherlands.
2. The number of orders has dipped since the pandemic and the first quarter 2 months are not performing as expected compared to the previous years.
3. The most liked category of products are the “Computers” while the least is the Game and toys in terms of ordered quantity
4. From the profit bins it is evident that the sales have peaked with less profit and vice-verse on the other case
5. Overall sales is $55.76M generating a profit around 58% till date.



Following observations can be made from the above visualization over time

1. Regardless of the gender the quantity ordered remains same for each category of products.
2. Average days to deliver the product is around 4 days and average quantity ordered is 3.14 .
3. The most liked product brand are the “Contoso” (available across all category )while the least is “Litware”(Home appliances and TV&Video category)
4. Stores in the United States have contributed the most to the sales profit and the least by the stores in France.

**SQL*:***

1. ***Customer Analysis***

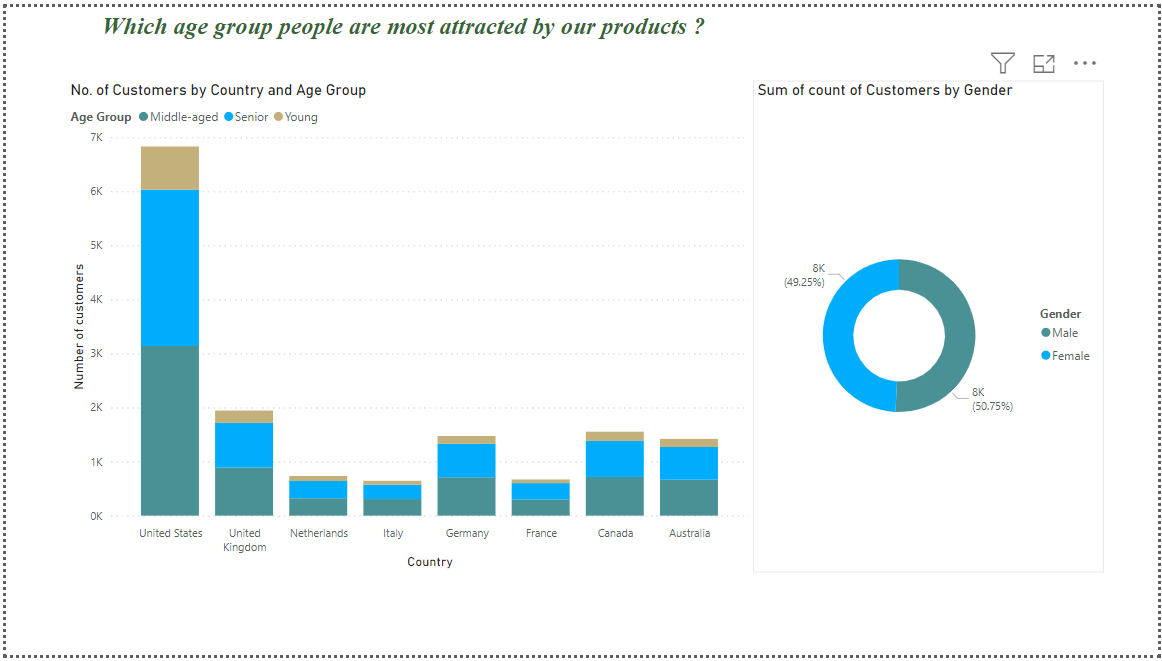
* SELECT gender, CASE

WHEN floor(datediff(curdate(), birthday) / 365) < 30 THEN 'young'

WHEN floor(datediff(curdate(), birthday) / 365) between 30 and 60 THEN 'middle-aged'

ELSE 'senior' END AS age\_group, country, count(\*)

FROM customers GROUP BY gender, age\_group, country;



Observation:

Regardless of the country the middle-aged (30-60) category have aligned to buy our products followed by the senior age group (> =60). Analysing in terms of gender category also seems to be around 50-50 attracted by our products.

* SELECT customers.Country, customers.State,

sum(Quantity\*(products.Unit\_Price\_USD)) AS Sales\_amount,

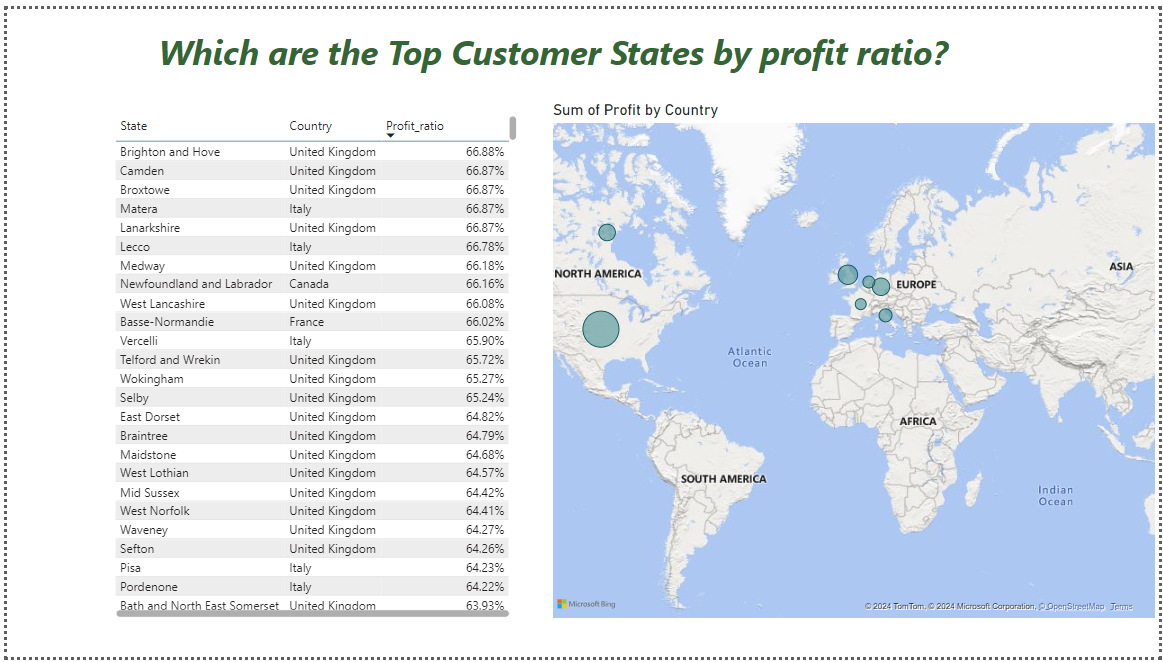
sum(Quantity\*(products.Unit\_Price\_USD-products.Unit\_Cost\_USD)) AS Profit,

sum(Quantity\*(products.Unit\_Price\_USD-products.Unit\_Cost\_USD))/sum(Quantity\*(products.Unit\_Price\_USD)) AS Profit\_ratio, count(\*) FROM sales

JOIN products ON sales.ProductKey = products.ProductKey

JOIN customers ON sales.CustomerKey = customers.CustomerKey

GROUP BY customers.State,customers.Country;



Observation:

The highest profit ratio was from Brighton and Hove (United Kingdom) customers followed by the Camdem(UK) , Broxtowe (UK).

***2.Sales Analysis***

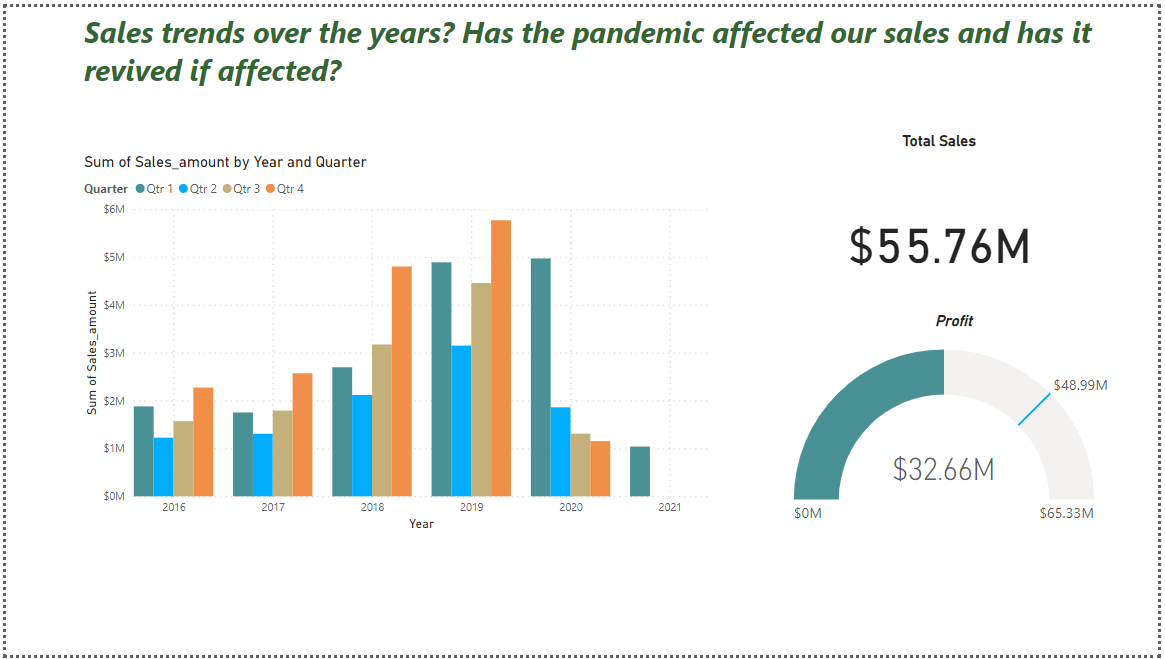
* SELECT Order\_Date, sum(Quantity\*(products.Unit\_Price\_USD)) AS Sales\_amount,

sum(Quantity\*(products.Unit\_Price\_USD-products.Unit\_Cost\_USD)) AS Profit

FROM sales

JOIN products ON sales.ProductKey = products.ProductKey

GROUP BY Order\_Date;

******

Observation:

We can see that the sales has increased the initial years but is not performing well post pandemic. Though 2/3 of first quarter data is available for 2021 , it does not suffice our expectations. It can also be observed that the pattern for sales amount is Q4>Q1>Q3>Q2 during the years 2016-2019.Post pandemic it has been reducing by never rising and it is expected this trend to continue.

* SELECT products.Product\_Name, SUM(Quantity) AS total\_quantity\_sold, SUM(Quantity \* products.Unit\_Price\_USD) AS total\_revenue,

SUM(Quantity \*(products.Unit\_Price\_USD-products.Unit\_Cost\_USD)) as profit

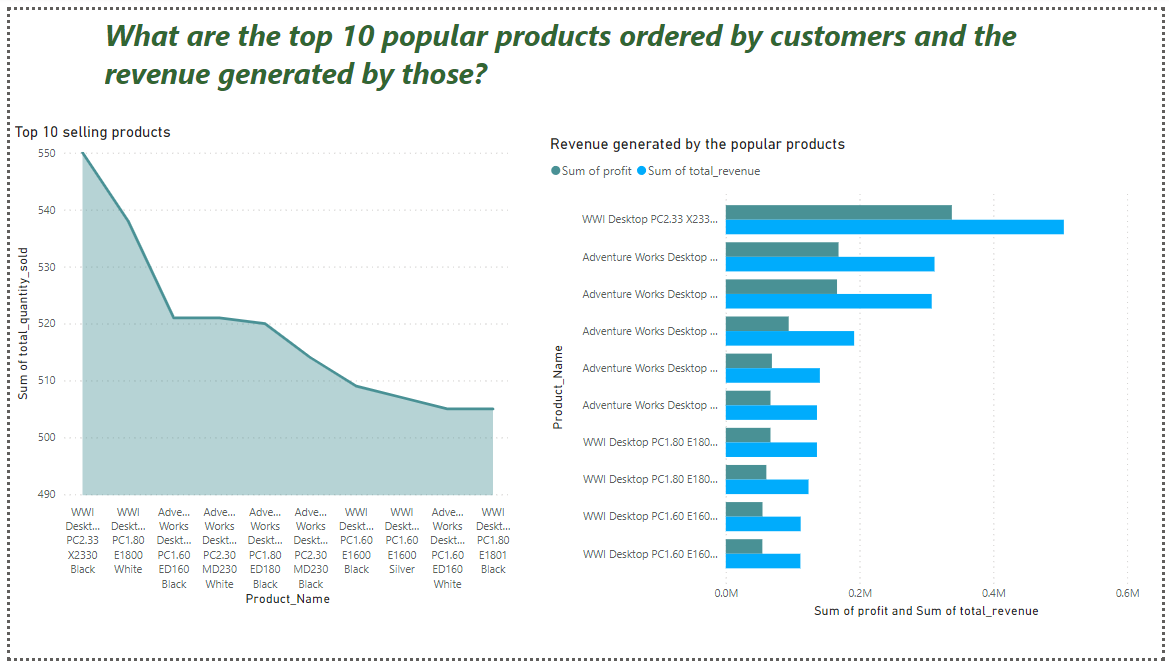
FROM sales

JOIN products ON sales.ProductKey = products.ProductKey

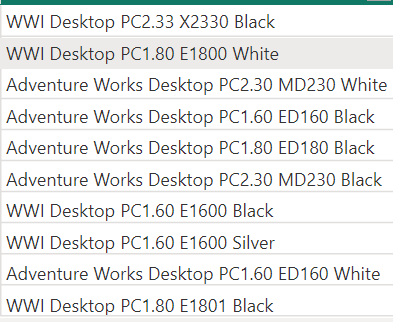
GROUP BY products.Product\_name

ORDER BY total\_quantity\_sold DESC

LIMIT 10;

******

Observation:



The top selling product was WWI Desktop PC2.33 X2330 Black and it has generated the most revenue amongst the top 10 popular products. The 2nd most ordered product was WWI Desktop PC1.80 E1800 White( which secured the 8th place in the revenue generated among popular products ). The 3rd most ordered product was Adventure Works Desktop PC2.30 MD230 White( which secured the 5th place in the revenue generated among popular products). The 2nd and the 3rd revenue generated products were the 4th and 6th most ordered products.

* SELECT stores.Country,stores.State, SUM(Quantity) AS total\_quantity\_sold,

SUM(Quantity \* products.Unit\_Price\_USD) AS total\_revenue

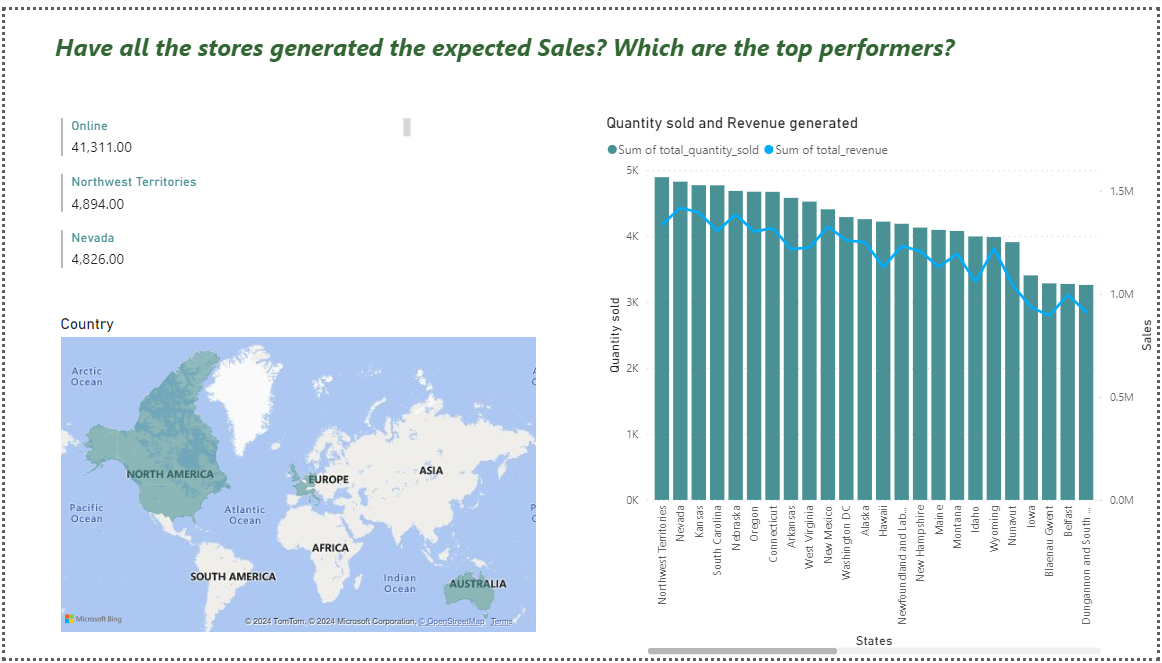
FROM sales

JOIN stores ON sales.StoreKey = stores.StoreKey

JOIN products ON sales.ProductKey = products.ProductKey

GROUP BY Country,State

ORDER BY total\_quantity\_sold DESC;

******

Observation:

Northwest Territories (Canada),Nevada(US), Kansas(US) have secured the 2nd-4th place in terms of quantity sold following the Online store. Northern territory (Australia) has sold the least amount of products. And one more noteworthy point to consider about this Northern territory store on that downside is all these products were sold only in the year 2016.

* SELECT exchange\_rates.Currency\_code,avg(exchange\_rates.Exchange),

SUM(Quantity) AS total\_quantity\_sold,

SUM(Quantity \* products.Unit\_Price\_USD) AS total\_revenue,

SUM(Quantity \* products.Unit\_Price\_USD\* exchange\_rates.Exchange) AS total\_revenue\_ex,

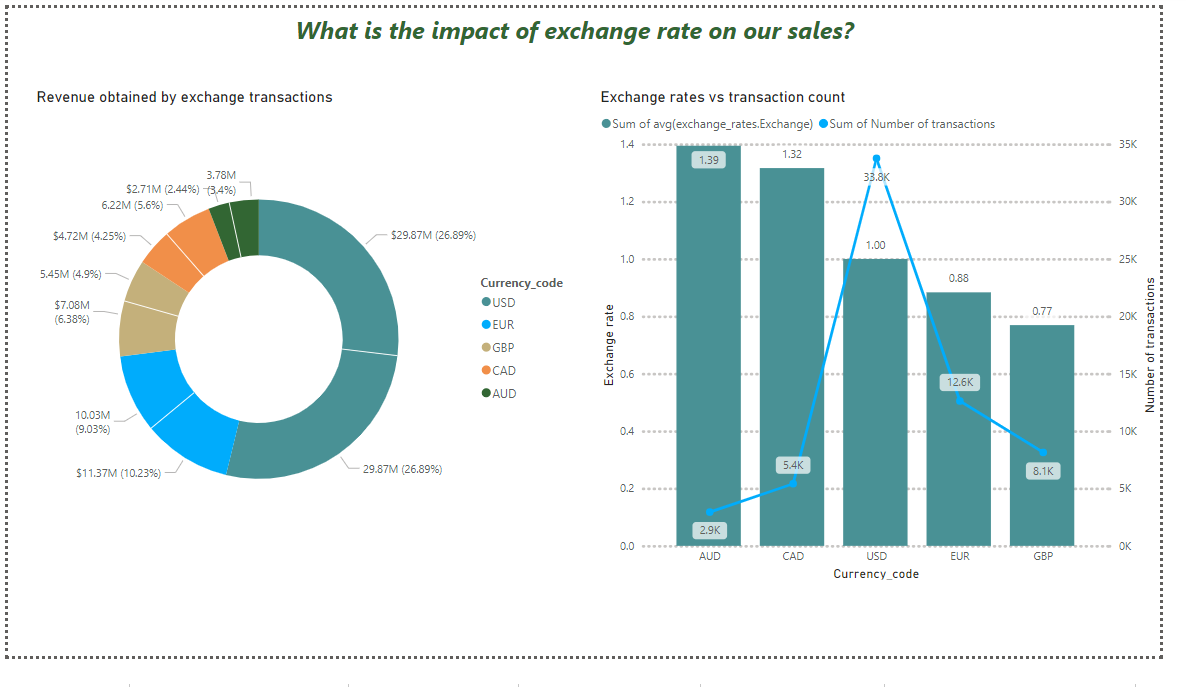
AVG(Quantity \* products.Unit\_Price\_USD\* exchange\_rates.Exchange) AS avg\_sales\_per\_transaction,count(\*)

FROM sales

JOIN exchange\_rates ON sales.Order\_Date = exchange\_rates.Date and sales.Currency\_code=exchange\_rates.Currency\_code

JOIN products ON sales.ProductKey = products.ProductKey

GROUP BY Currency\_code;

******

Observation:

Since all the revenue comparison are made with the revenue generated in US currency $, it is significant to consider the currency exchange rates. Considering the average conversion rates over the time and the no.of transactions made, it is evident that conversion rates affects the amount of transactions made.

Case 1: When the user shells out 1.4 AUD for 1 USD to buy the product, the number of transactions made is less. Same for CAD.

Case 2: When the user shells out 0.77 GDP for 1 USD to buy the product, the number of transactions made is relatively 2 times higher than case 1. Though UK has ordered the 2nd highest number of products overtimes, the number of transactions is less. Same for EUR.

***3.Product Analysis***

* SELECT products.Category, products.Subcategory,

SUM(Quantity) AS total\_quantity\_sold,

SUM(Quantity \* products.Unit\_Price\_USD) AS total\_revenue,

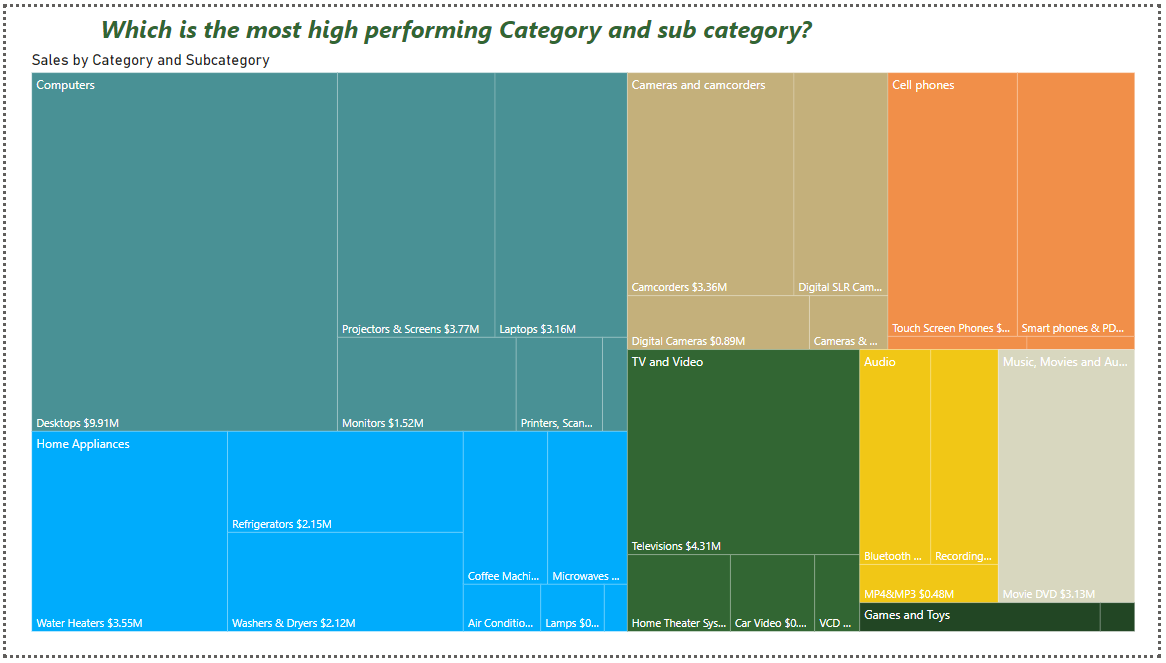
SUM(Quantity \* products.Unit\_Cost\_USD) AS total\_cost,

avg((products.Unit\_Price\_USD-products.Unit\_Cost\_USD)/products.Unit\_Price\_USD) AS Profit\_margin

FROM sales

JOIN products ON sales.ProductKey = products.ProductKey

GROUP BY Category,Subcategory;

******

Observation:

The highest performing category is “Computers ” followed by “Home Appliances” and “Cameras and camcorders” . The least performing category is Games and Toys. Amongst the subcategories, the top 3 performing subcategories are Desktops, Televisions and Projectors & Screens.

* SELECT products.Product\_Name,customers.Country,SUM(Quantity) AS total\_quantity\_sold

FROM sales

JOIN products ON sales.ProductKey = products.ProductKey

JOIN customers ON sales.CustomerKey = customers.CustomerKey

GROUP BY customers.Country,products.Product\_Name;

******

Observation:

The following table shows the top 3 products for each country.

|  |  |
| --- | --- |
| ***Country*** | ***Products*** |
| United States | WWI Desktop PC2.33 X2330 Black, Adventure Works Desktop PC2.30 MD230 White, WWI Desktop PC1.60 E1600 Red |
| United Kingdom | Adventure Works Desktop PC1.60 ED White, WWI Desktop PC3.0 M0300 Black, Contoso DVD External DVD Burner M200 Grey |
| Canada | Adventure Works Desktop PC2.30 MD230 White , SV DVD Player M120 White, Adventure Works Desktop PC3.0 MS300 Black |
| Germany | Adventure Works Desktop PC2.30 MD230 Black, SV DVD 14-Inch Player Portable L100 Silver, WWI Desktop PC1.80 E1800 White |
| Australia | WWI Desktop PC2.33 X2330 Black, Adventure Works Desktop PC2.30 MD230 White, WWI Desktop 1.60 ED1600 Red |
| Italy | Adventure Works Desktop PC1.80 E180 Silver, Adventure Works Desktop PC2.33 XD233 Silver, Adventure Works Desktop PC1.80 E180 White |
| France | Contoso DVD 12-Inch Player Portable M400 White, Contoso DVD Recorder L230 Grey, Adventure Works Desktop 1.60 ED160 Brown |
| Netherlands | WWI Desktop PC1.80 E1800 White, Adventure Works Desktop PC1.80 ED180 White, Contoso DVD 38DVD Storage Binder E25 Red |

***4.Store Analysis***

* SELECT products.Product\_Name,customers.Country,

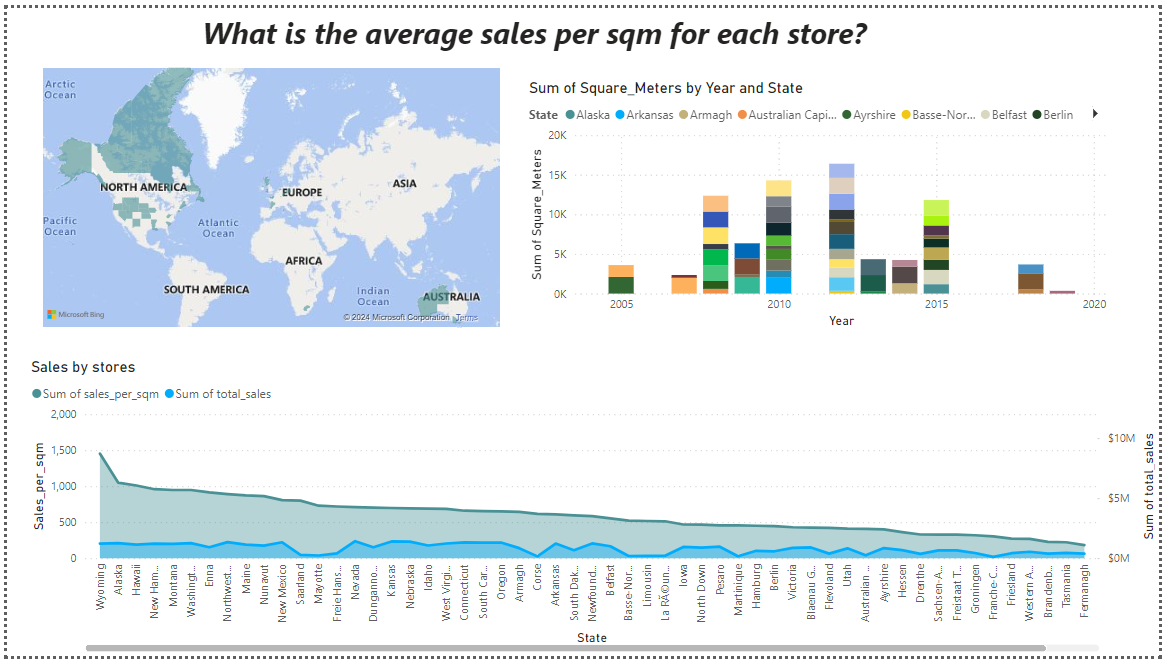
SUM(Quantity) AS total\_quantity\_sold

FROM sales

JOIN products ON sales.ProductKey = products.ProductKey

JOIN customers ON sales.CustomerKey = customers.CustomerKey

group by customers.Country,products.Product\_Name;

******

Observation:

The store in Wyoming(US) which was opened in 2014 has made highest Sales\_per\_sqm. Followed by Alaska and Hawaii stores(2015,US).

* SELECT s.State, year(Order\_Date) AS order\_year,

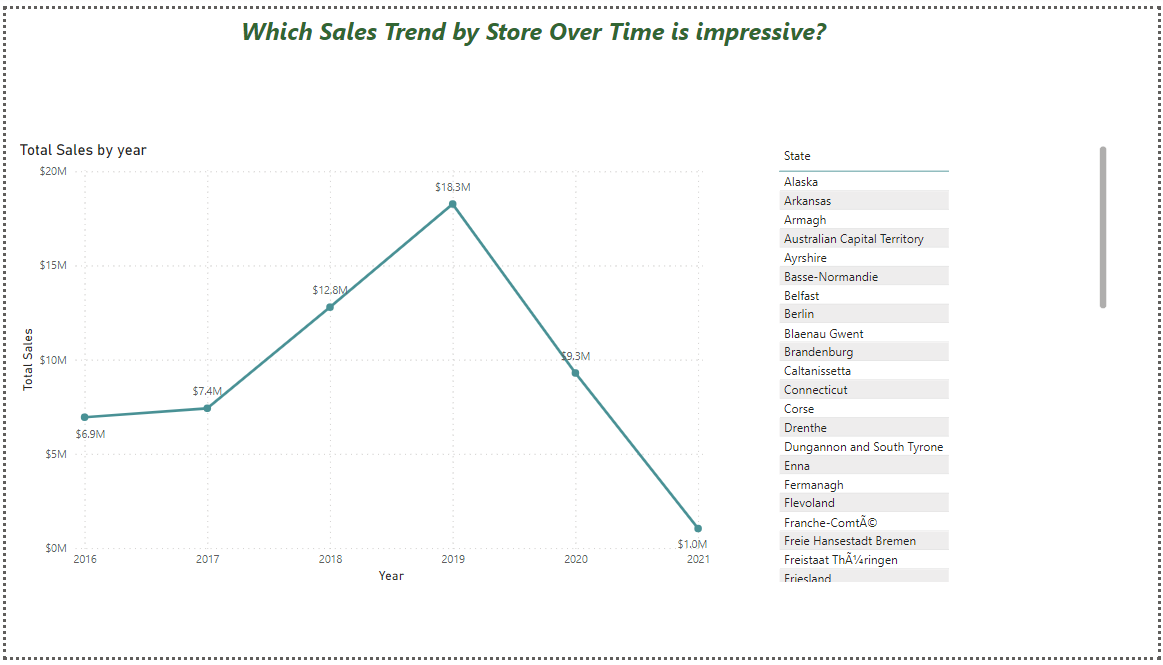
SUM(Quantity \* products.Unit\_Price\_USD) AS total\_sales

FROM sales

JOIN stores s ON sales.StoreKey = s.StoreKey

JOIN products ON sales.ProductKey = products.ProductKey

GROUP BY s.State, order\_year;

******

Observation:

Almost all the stores have performed well throughout the timeframe considered. Except the following stores- Caltanissetta, Fermanagh, Northern Territory

