Teesside University

# BIG DATA AND BUINESS INTELLIGENCE
## (CIS 4008 – N)


**NAME: EUCHARIA MADU**

**STUDENT ID: C2082260**


**PROJECT TOPIC: GLOBAL SUPERSTORE ANALYSIS**

Teesside University

# Table of Contents

# 1. EXECUTIVE SUMMARY

This report is based on a Global super market sales analysis from 2011 to 2015. The end goal is to see the future trend and market sale based on the analysis. Now this analysis is helpful from the seller point of view as it shows the sales and profit growth and trend with respect to the products, customers and their various locations. With the help of this analysis, it is easy to see what customer exactly likes and dislikes and which product is more purchased by the customers. So that in future, demand can be fulfilled. The analysis shows that the store made the most profit $31,045.99 from Cisco Machine and. The analysis was also able to deduce that sales and revenue is highest in October and lowest in march with October having a revenue of $2,531,678,422.10 and sale of $9,494,097.34. March has a Revenue of $40,147,956.93 and a sale of $200,389.89. further analysis also shows that the Consumer segment made a purchase of $6,507,949.42 which is the highest purchase and covers 51.48% of the total sales made by the store. It was also seen that the store made the highest sale from Europe with a sale of $2,822,302.52.

The analysis shows that the products that yield more profit for the store are the technology products so there is a need to explore options to increase sales in this category. Sales also increases as the number of customers increase, so it is imperative for the store to increase their marketing skills and explore different ways to get more buying customers from countries with low purchase.

The store analysis will enlighten the store to know that while giving discounts increases sales it also reduces profit for most of the categories so the store will need to focus on giving discounts to technology category because the analysis done shows that that was the only category that recorded an increase in profit with a discount.

## 1.1. Introduction

Several studies and analysis have been done on the Global super store data. From those studies different results and insights are taken. But in this project, analysis is primarily done in two parts one is sales with regards to segment, product and category while the other is Profit with regards to segment product and category. For the analysis, different countries were taken into consideration and different parameters were used for get meaningful information. This analysis is mainly providing country wise sale and revenue, profit across the countries, highest customers in particular country and many more others.

## 1.2. Data Source

The Global super store data was sourced from Kaggle and it contains information related to a superstore sales. It contains almost 5 years of data taken into consideration for the analysis. The data set contains 24 columns and 51,291 rows. The link to the data set is seen below.

https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset?select=Global_Superstore2.csv

### 1.2.1. Description

| S/No | Column Name | Description |
|---|---|---|
| 1 | Row ID | A Unique ID identifier for each row |
| 2 | Order ID | Unique ID for each order |
| 3 | Order Date | Date the order was made |
| 4 | Ship Date | Date the order was shipped |
| 5 | Ship Mode | The mode for the the shipping |
| 6 | Customer ID | Unique ID for the customer |
| 7 | Customer Name | Name of the Customer |
| 8 | Segment | Split of customers based on their spend |
| 9 | City | A more exact location |
| 0 | State | The Stae where the purchase was made |
| 11 | Country | The country where the order was made from |
| 12 | Postal Code | The address for delivery |
| 13 | Market | A code that describes the region |
| 14 | Region | The area from which the order was made |
| 15 | Product ID | Unique ID for the products |
| 16 | Category | Division of products for easy identification |
| 17 | Sub Category | A smaller division of product |
| 18 | Product Name | Name for each product |
| 19 | Sales | The value of money received for a product |
| 20 | Quantity | The number of products bought |
| 21 | Discount | A deduction from the origingal cost |
| 22 | Profit | The amount gained after decucting the cost |
| 23 | Shipping Cost | Cost for shipping |
| 24 | Order Priority | The order with which the request is processed |

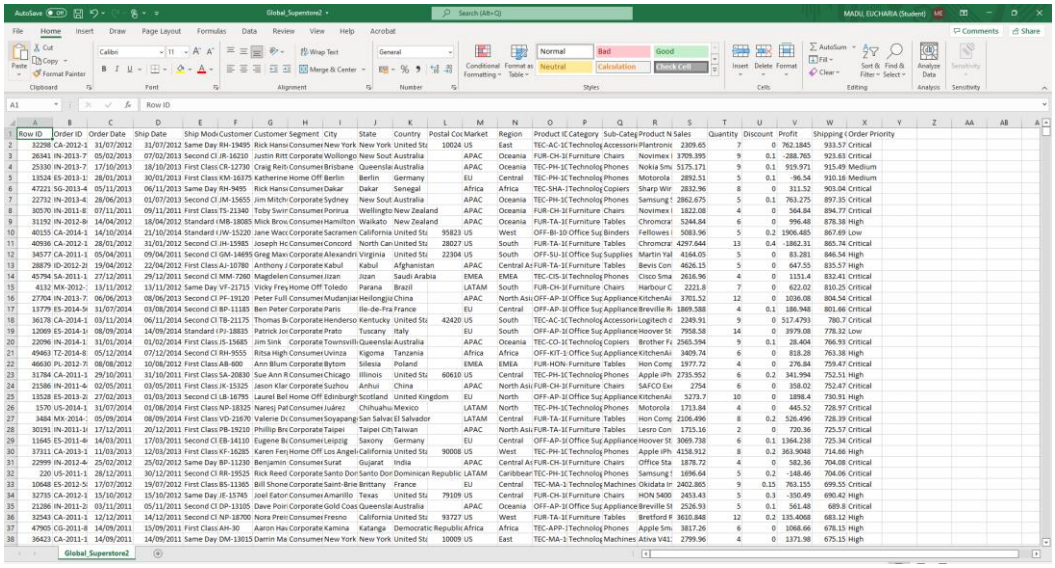

*Figure 1 Showing the Columns in the Data set*

## 1.3. Business Intelligence requirements and questions

The objective of this project is to provide the store with insights on how to deal with customers as per their demands and also show the segments of customers and product type that yield the highest revenue and sales. This project seeks to answer the below questions.

1. What are the Top profit Making products types?
2. What is the sales and Revenue Pattern Month on month?
3. What region gives the store the most sales on a yearly basis?
4. What segment makes the most purchase?
5. What Country gives the store the highest sales on a yearly basis?
6. What Category of products do the store record the most sales?
7. What Countries do the store make the most profit from?
8. What customer segment yield the most profit for the company?
9. Determine the profit by category and discount by year?
10. When is Revenue likely to be low?

This analysis is basically focused on the sales, profit and revenue for the store within a 5 years period. It is intended for the store to know what country, region, product and segment they make more revenue and profit from and to know how the discounts they offer affects their profit as the major aim or goal of every business is to make profit, in the bid to do this customers need to also get fulfilment during purchase and this analysis gives the company insight on the region, country, city and product to focus their marketing skills.

## 1.4. Findings based on Analysis and Evaluation

After proper analysis, different charts were plotted as shown below which explains the outcome of the analysis donewhich answers the business questions that will help the store to know the market trend.

Figure 2 was plotted using a tree map displaying the top 5 products based on the profit yielded by the store from them it was used to answer the first business question and it shows that the store made the most profit $31,045.99 from Cisco Machine followed by Canon Image with a profit of $25,199.93. The third product with the highest profit is Nokia smart phone which gave the company a profit of $25,043.41. Saunder Class and Hewlett wireless gave the store a profit of $23,371.95 and $21.372.97 respectively.

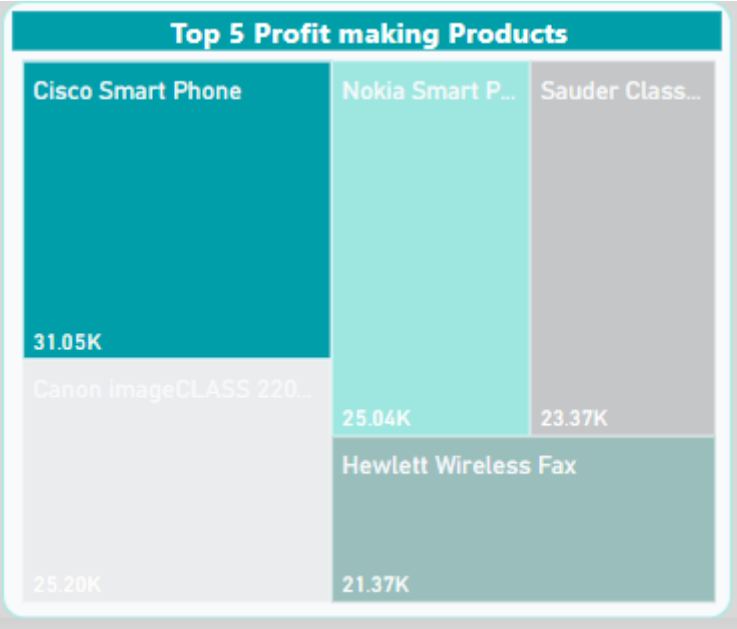


*Figure 2: Tree Map of Top 5 profit making Products*

Figure 3 below was plotted using a line chart to compare the pattern between sales and revenue. This was used to answer the second business question. From the analysis it shows that sales and revenue is highest in October and lowest in march with October having a revenue of $2,531,678,422.10 and sale of $9,494,097.34. March has a Revenue of $40,147,956.93 and a sale of $200,389.89.

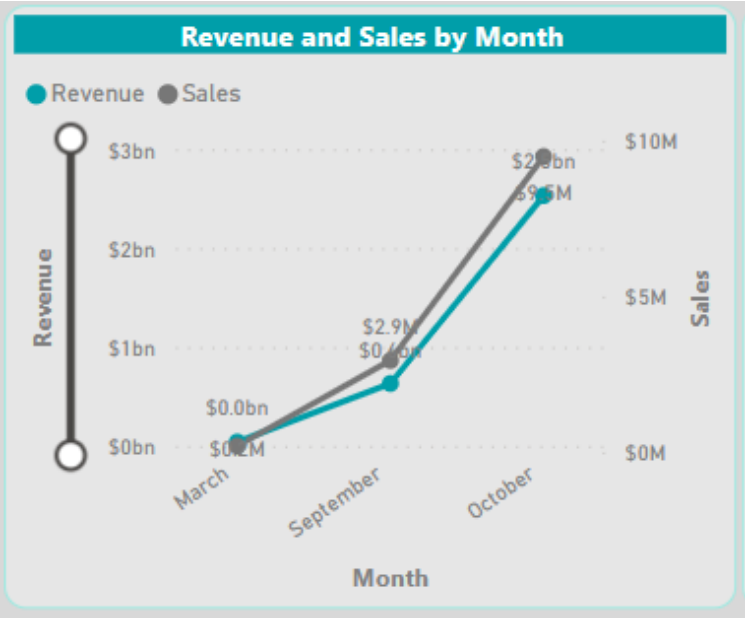


*Figure 3: Line chart showingnRevenue and sales by Month*

Figure 4 is a map used to visualise spatial data and it represented the amount of sales made per region. This chart was used to answer the third business question.

It shows that store made the most sales from Europe with a sale of $2,822,302.52 and made the least sale from Caribbean with a sale of $324,280.86


*Figure 4: Map Showing Sales by Region*

Figure 5 is a pie chart that was used to represent the sales made per segment, this chart wasused to answer the 4th business question. From the analysis, it shows that Consumer segment made a purchase of $6,507,949.42 which is the highest purchase and covers 51.48% of the total sales made by the store.

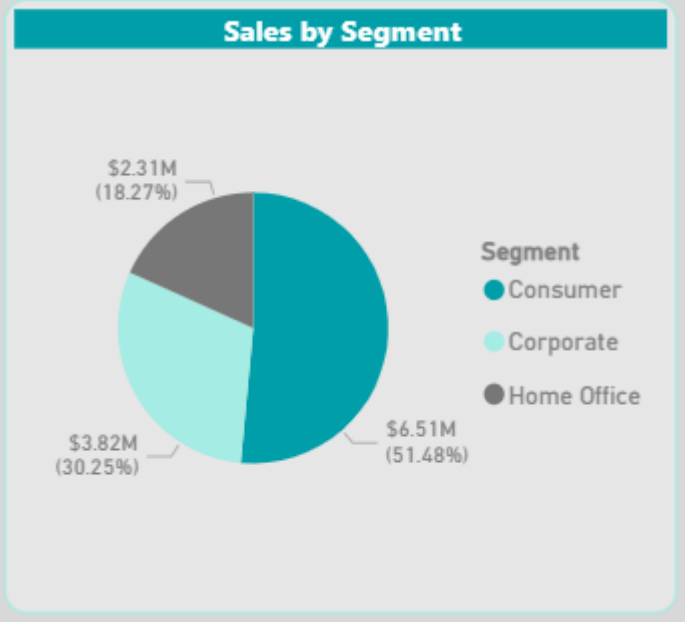


*Figure 5: Pie Chart showing Sales by Segment*

Figure 6 is a clustered bar chart that was used to represent the amount of sales made per country. The analysis from this visualization was used to answer the 5th business question and from this analysis it shows that the store made the most sales from USA with a sale of $2,297,200.86 in 2011. The analysis also shows that USA made the most purchase across the other 4 years.

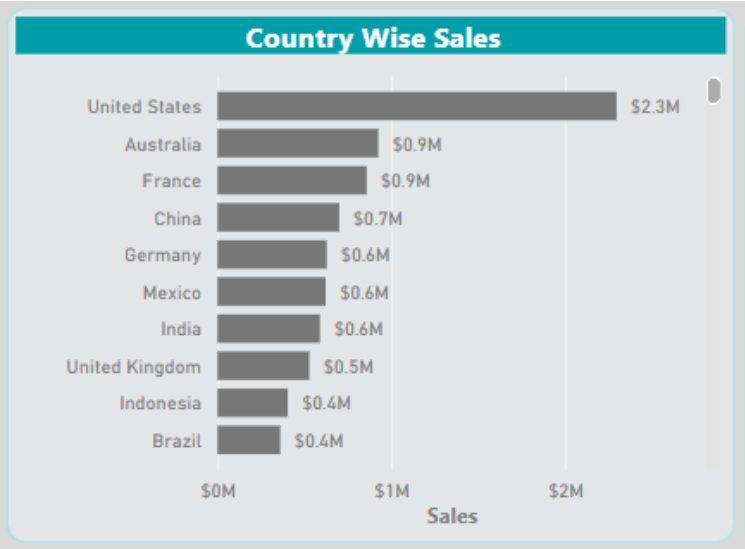


*Figure 6: Clustered Bar chart showing Country wise Sales*

Figure 7 below is a donut chart and it was used to visualize the amount of sales that the store made by category. After the analysis this was used to answer the 6th business question. From the analysis it shows that the store made the most sales from Technology with a sales amount of $4,744,557.50 which is 35.53% of the total sales made per for the whole category.

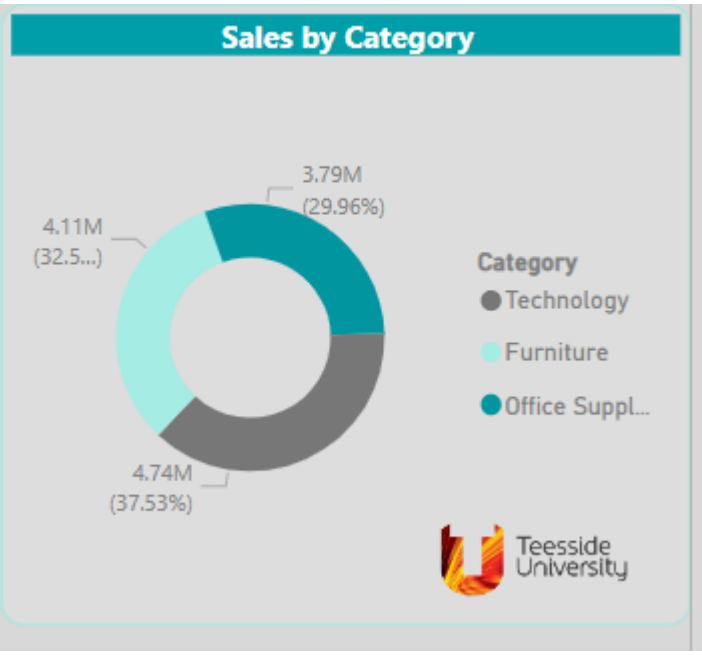


*Figure 7: Donut Chart showing Sales by Category*

Figure 8 is a stacked column chart that visualises profit by country with more focus on the top 5 countries, Analysis done was used to answer the 7th business question. From the chart it can be seen that the store made more profit from the sales made by customers from USA with a total profit of $286,397.02.

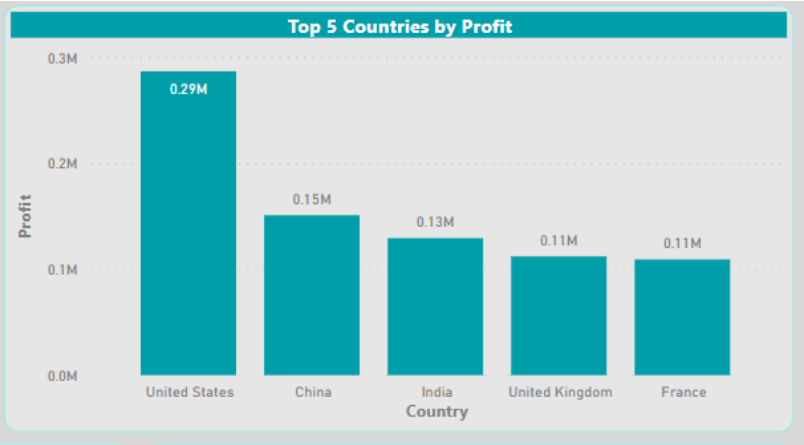


*Figure 8: Stacked Column Chart showing Top 5 Countries by Profit*

Figure 9 is a clustered bar chart that visualizes top 3 segments by profit. The analysis done on this chart is used to answer the 8th business question and it shows that the store made the most profit from the consumer segment with a profit of $749,239.78.

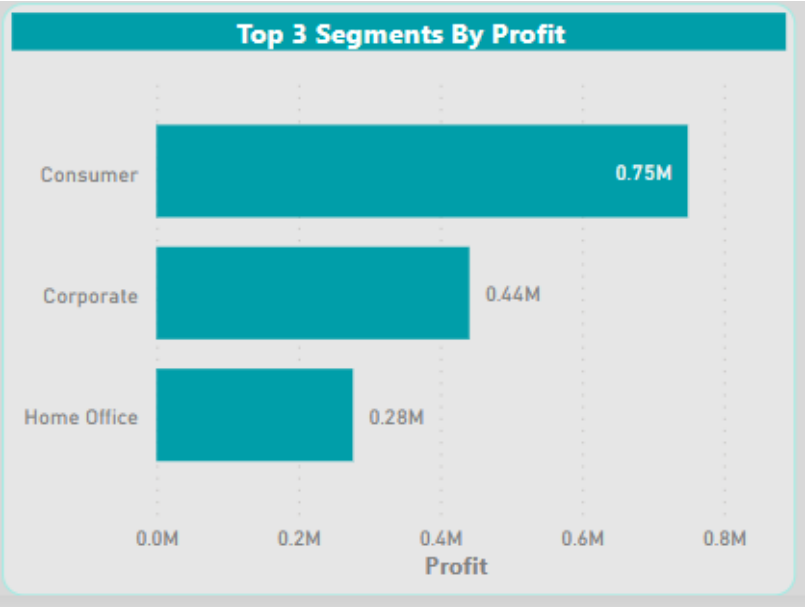


*Figure 9:Clustered Bar Chart showing Top 3 Segments by Profit*

Figure 10 is a waterfall chart that was plotted to visualize profit by category and discount. Analysis done on this chart is used to answer the 9th business question. It shows that there was a 26.96% increase in the profit made from the sale of technology products with a 0.1 discount.


*Figure 10: Waterfall Chart showing profit by category and Discount*

Figure 11 is a decomposition tree that visualizes the sales hierarchy by category, product name, country, city and customer name. it basically shows a high level of the information that is gotten using the data set.


*Figure 11 Decomposition Chart showing Sales Hierachy*

## 1.5. Recommendation

- The analysis shows that the products that yield more profit for the store are the technology products so there is a need to explore options to increase sales in this category
- Sales increases as the number of customers increase, so it is imperative for the store to increase their marketing skills and explore different ways to get more buying customers from countries with low purchase.
- From the analysis it shows that while giving discounts increases sales it also reduces profit for most of the categories so the store will need to focus on giving discounts to technology category because the analysis done shows that that was the only category that recorded an increase in profit with a discount.

## 1.6. Conclusion

In conclusion my knowledge and skill for power BI has improved massively and I have this project to thank for that. My data analytics and business intelligent skills have also improved greatly. I encountered a few challenges during the cause of this work from deciding on which data set to choose to the pre processing of the data set and the analysis. Thankfully the workbook from the lab sessions helped to overcome some of these challenges and research done also helped greatly.

It will be my pleasure to apply the knowledge and skills that I have gained in this module in the real world as they will definitely be used to address various problems .

# Section 2  BUSINESS INTELLIGENCE DESIGN

## 2. Data Pre-Processing and Cleaning

Before analysis cleaning & transformation is one of the crucial or necessary step which helps in removing redundancy and gives the idea of useful parameters across the dataset. There are several ways by which data can be imported into power BI desktop like CSV mode, SQL Server, JSON, excel, oracle and many more.

### 2.1 Loading the Data

The data to be used for this report is in Excel format and it was loaded by clicking the get data drop down and clicking on excel workbook.

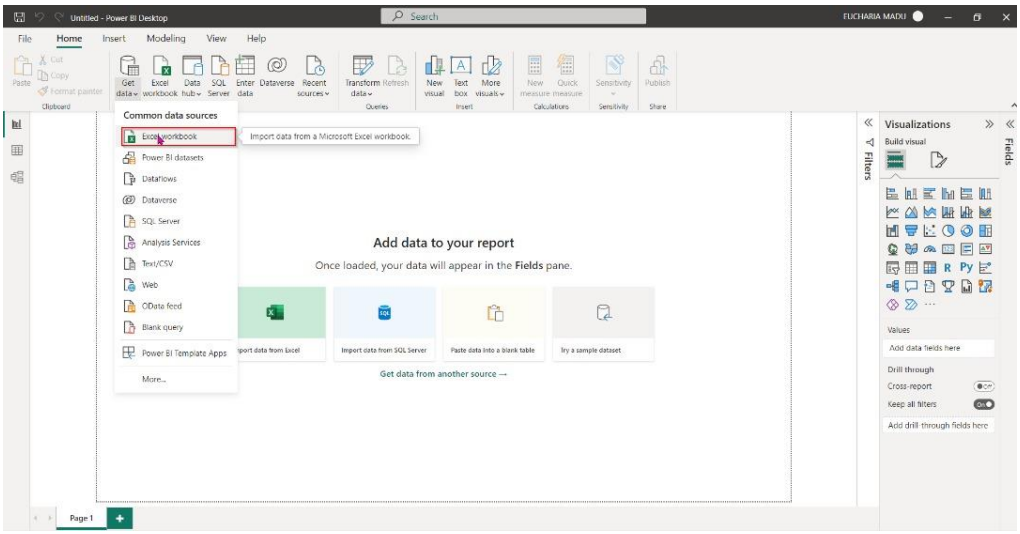


*Figure 12: Showing how the data was Loaded*

The Global superstore data was selected. This opened a dialog box showing the top 21 rows and 24 columns of the data set and a dialog box showing the option to click on Load, transform data or cancel came up.

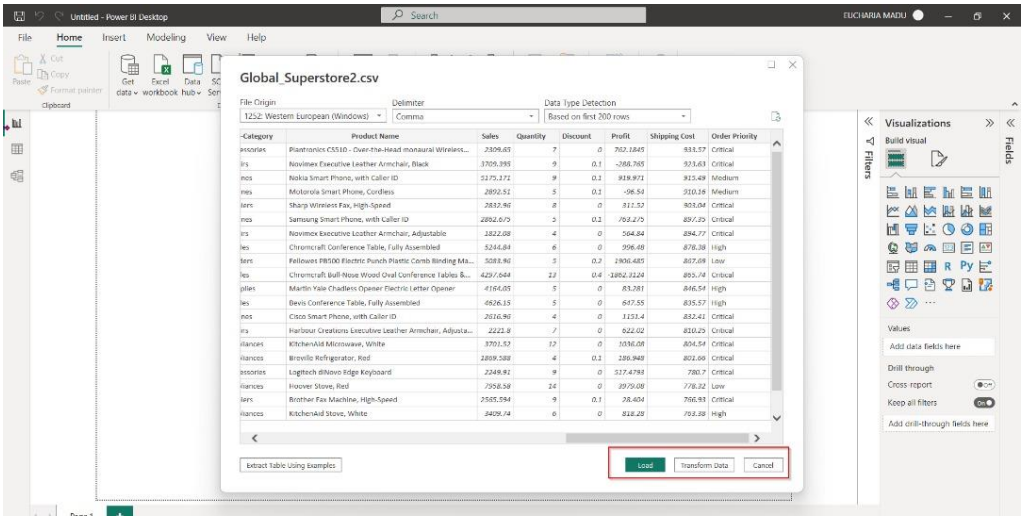


*Figure 13: showing how the data was loaded*

After importing the data into the Power BI Desktop, a new window pops up which shows initial rows of the dataset. There are two ways to deal with the data like first load then Transform or first transform then load the data to perform the visualization. But the best practice is to first transform the data then load it for further analysis. The power query editor mode is where most of the transformation happens before the analysis.

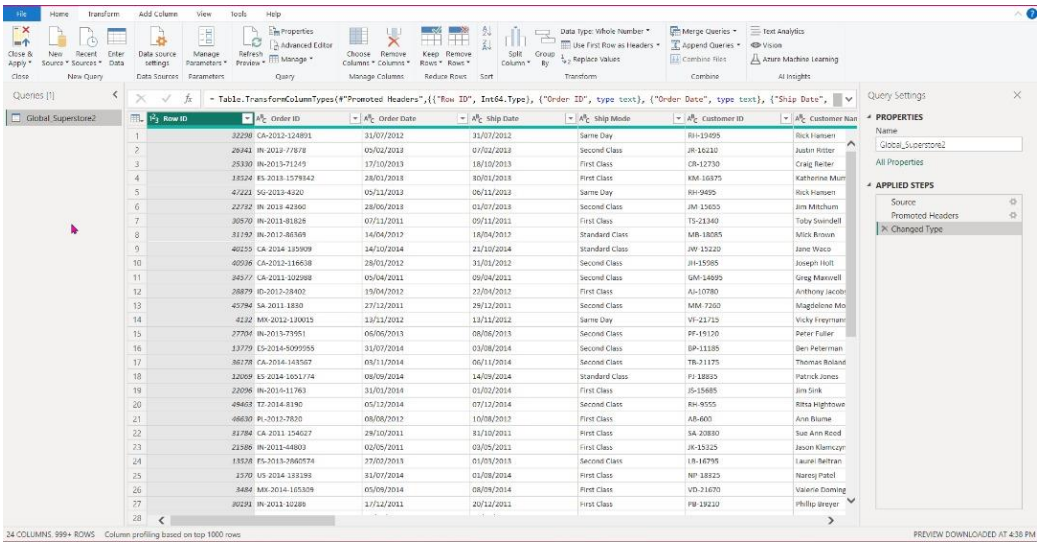

*Figure 14: Showing the power query editor*

## 2.2 Data Cleaning

To get the basic idea about the data, column quality option has been enabled under the view Ribbon bar. After enabling this option, it is easy to know null and blank values as well as the relevant data within a column.

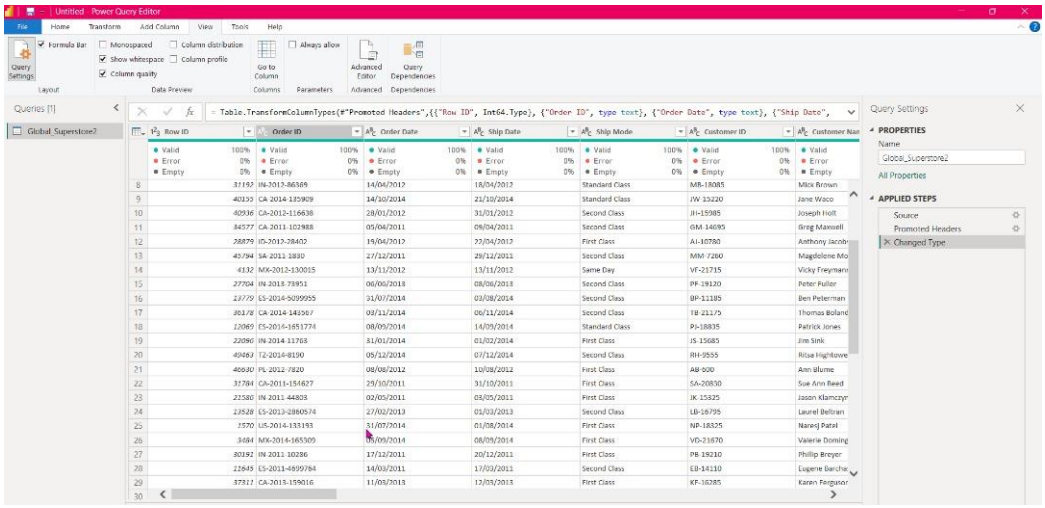

*Figure 15: showing the column quality*

From the figure it can be seen in all columns empty percentage is zero except postal code which contains so many null values. It happens in very rare cases where data is already cleaned and free of blank and null values.

### 2.2.1 Removing Errors/Columns

Postal code column contains so many null values and we can deal with this to remove null values. But in this project postal code is not used anywhere for the analysis and due to this postal code is removed from the dataset.

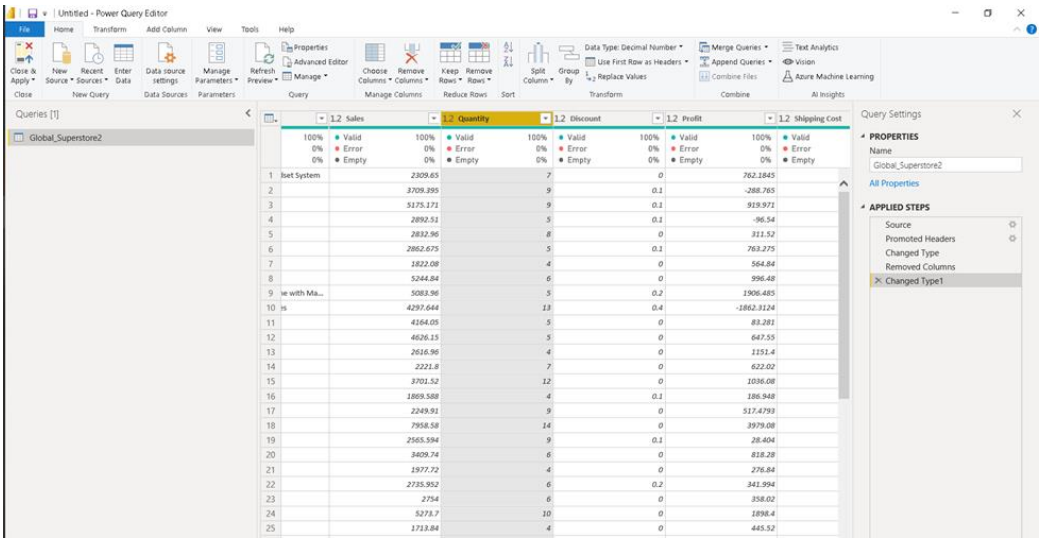


*Figure 16: Showing the removal of Postal code column*

To delete any column from the dataset select particular column then right click on to it, a new box will appear on screen which contains so many options like remove, data type change, replace value and many more. From figure 2.4 it can be seen that this column has been removed from the dataset by clicking on the remove button.

### 2.2.2 Making a column Data Relevant

In the next step, data type among all the columns is checked where it is found that almost all columns contain relevant data type with respect to the data. But the data type of the Quantity column is in decimal number and quantity cannot be decimal number to change it into whole number. There is an option under Home Ribbon bar named data type which shows current data type of column from there it changed into whole number see figure 2.5 for the reference.

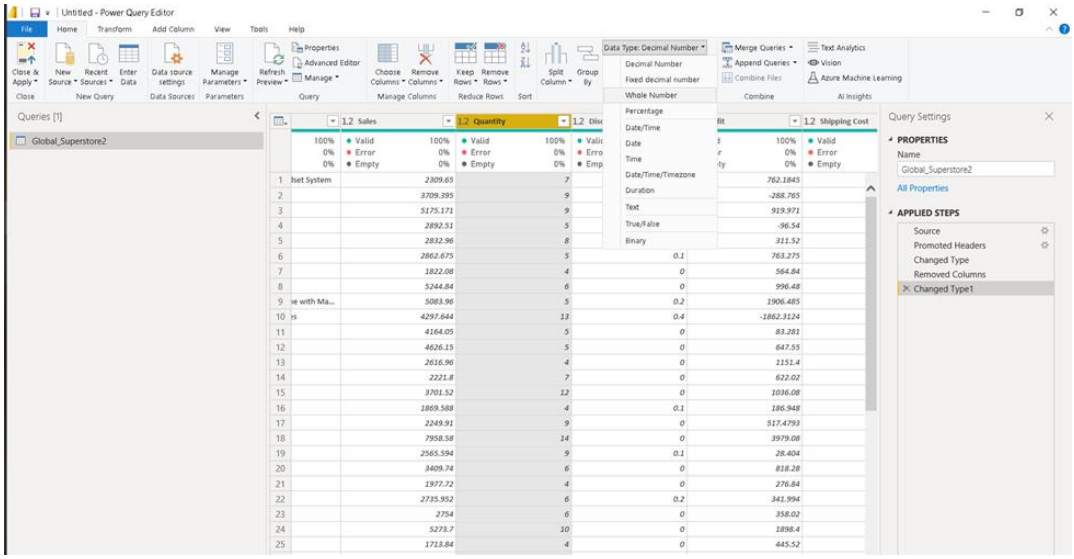


*Figure 17: showing the change of data type to whole number*

### 2.2.3 Formating Columns

The Sales, shipping cost and revenue columns in the sales table were formatted by clicking the column tools and selecting currency under format to change the format to currency. This is illustrated in the figure below.

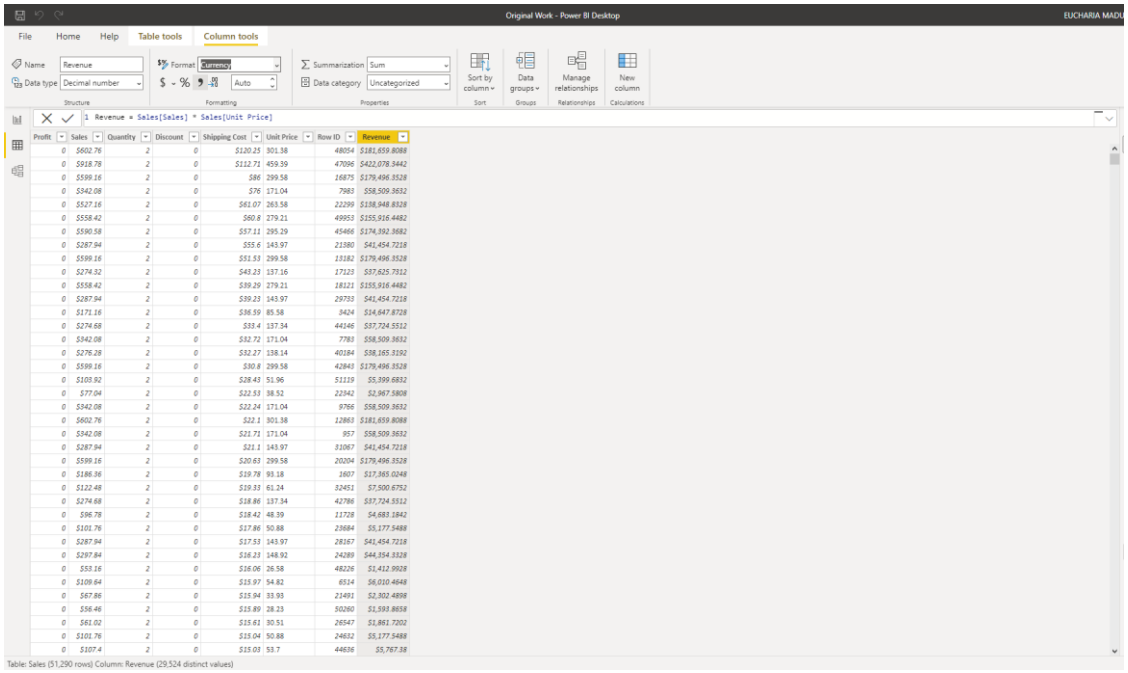


*Figure 18: Illustrating Column Formatting*

## 2.3 Data Modelling and Business intelligence

For this project, the modelling type used is Star schema and this schema is considered as one of the best schemas in modelling. The Fact table is Global sales table which is connected to the 5 different dimension tables.

One of the crucial things is cardinality which is also many to one in all the tables. This type of cardinality is considered as one of the best cardinality among the 4 different cardinality types.

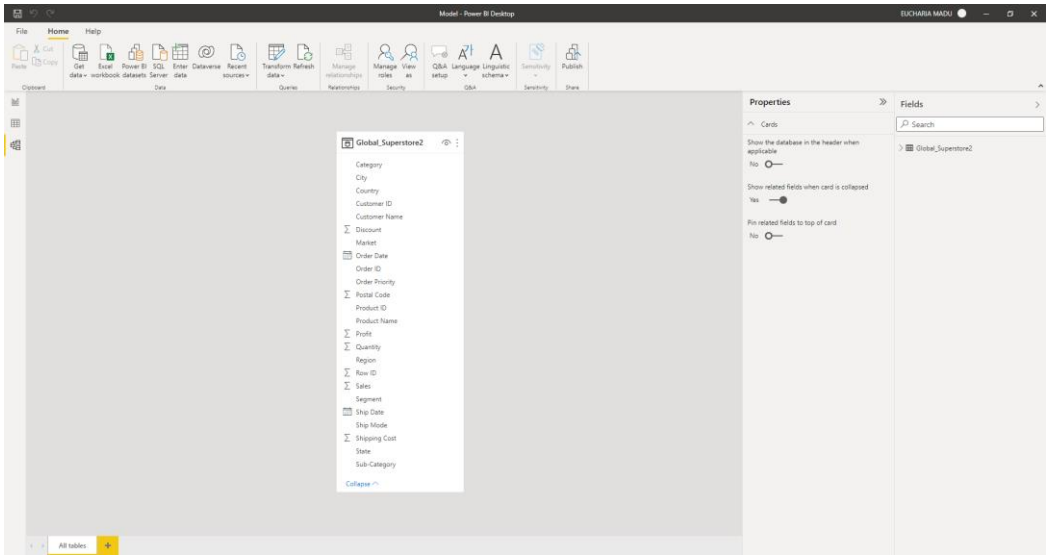


*Figure 19: Showing Global superstore as the Fact table*

5 tables were created from the global sales table which are:

1. Customer Information Table
2. Product Information Table
3. Sales Table
4. Order Information Table
5. Date Table

To create the dimensions table, the Global superstore table was duplicated and renamed. The required columns were selected, the columns not required were deleted by right clicking on the selected columns and clicking on remove other columns

The step above was repeated to get the remaining 4 tables.

I attempted to remove duplicates by highlighting the columns in the table, right clicking on the highlighted columns and clicking remove duplicates, the data had no duplicates.

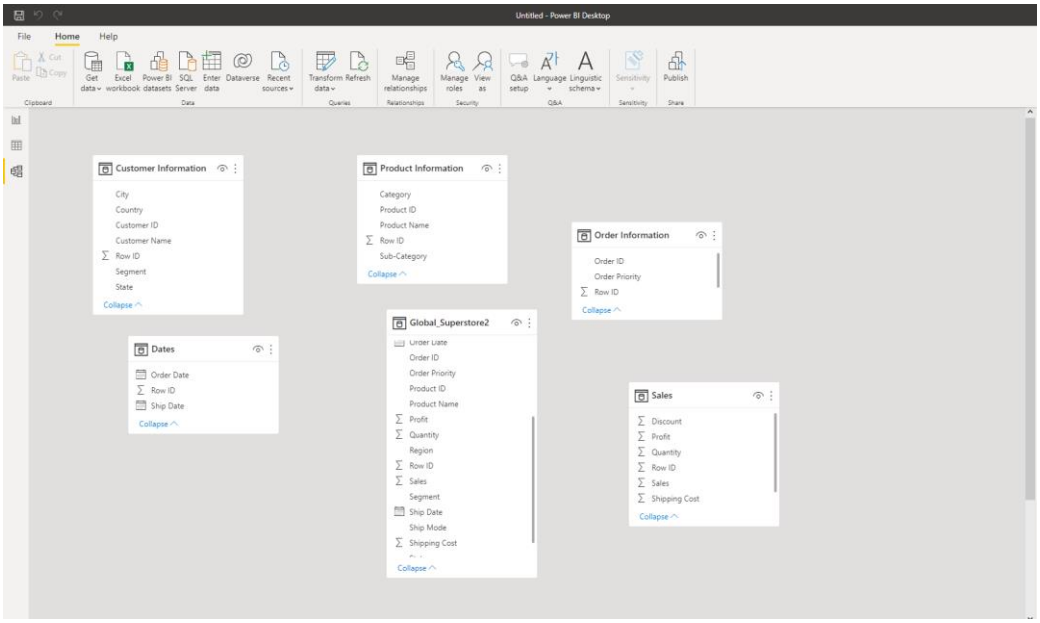


*Figure 20: showing the Dimensions table created*

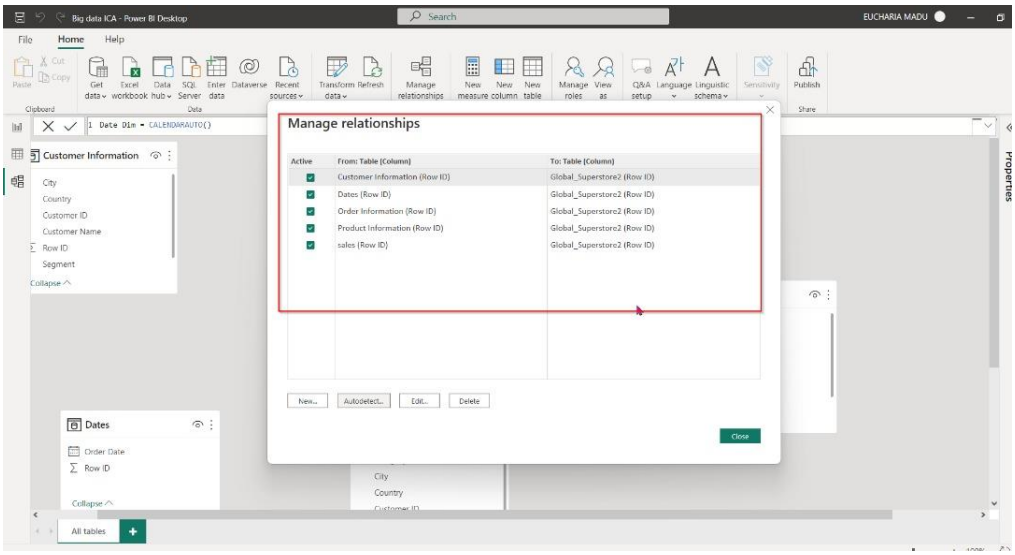


*Figure 21: Managing Relationships*

Figure 21 above shows that the relationship between the fact table and the dimensions table was created using row ID which is a unique identifier. Figure 22 shows the relationship created.

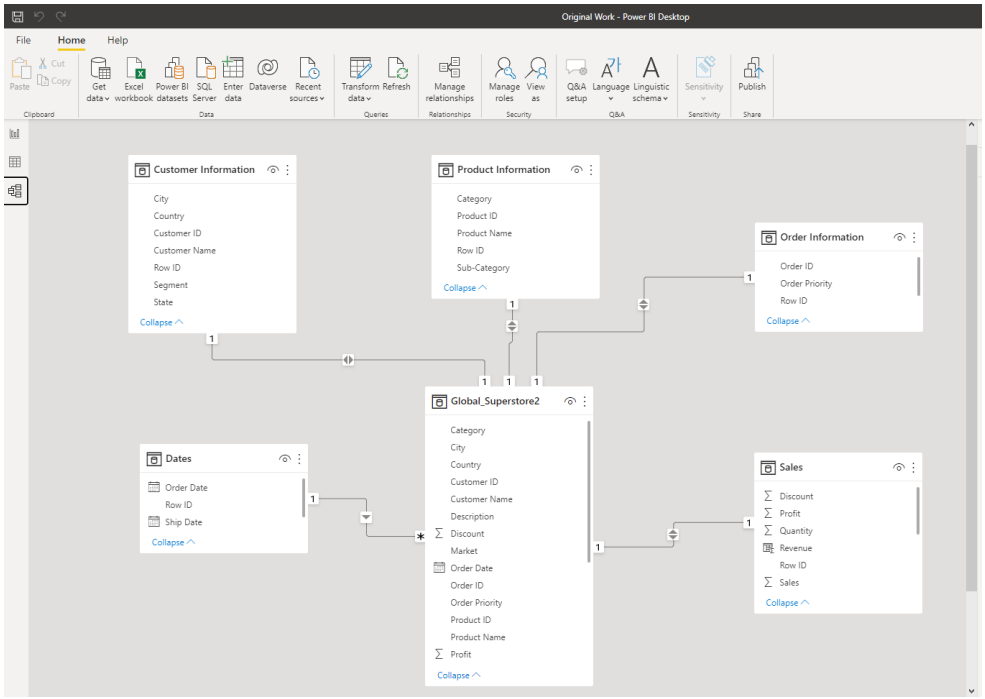


*Figure 22: Showing the Star schema data Modelling*

## 2.4 DAX and M Language

Fig 23 shows the M Language as the dataset which is used in this project is already cleaned and free of duplicate values. Due to these, few steps are applied to transform and clean the data. Few transformation steps are manually written in the advance editor window like changing of the datatype. The product name column is split into two columns product name and description based on the delimiter. This transformation is done using M code in advance editor.

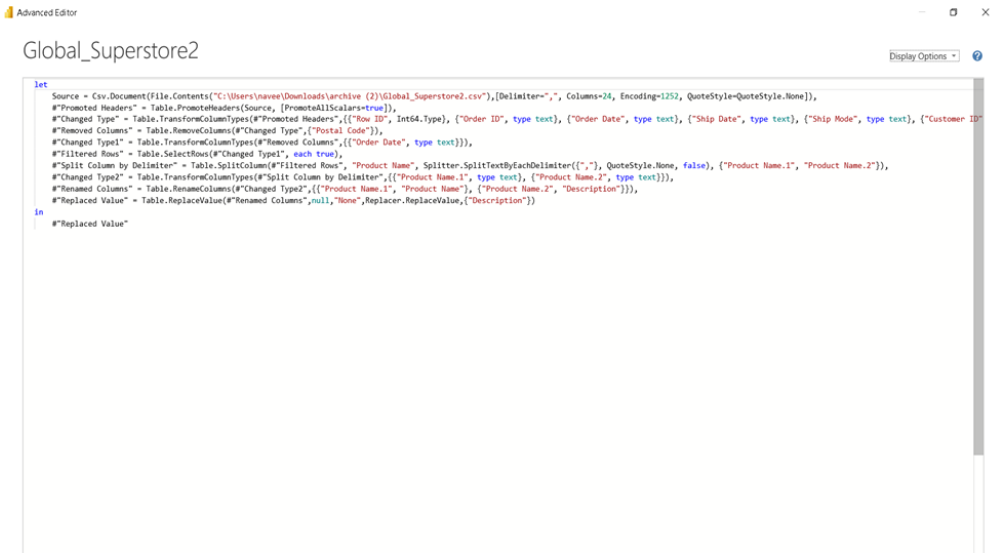


*Figure 23: showing the M Language*

| Product Name | Description |
|---|---|
| Valid 100% | Valid 100% |
| Error 0% | Error 0% |
| Empty 0% | Empty 0% |
| Plantronics CS510 - Over-the-Head monaural Wireless Headset System | None |
| Novimex Executive Leather Armchair | Black |
| Nokia Smart Phone | with Caller ID |
| Motorola Smart Phone | Cordless |
| Sharp Wireless Fax | High-Speed |
| Samsung Smart Phone | with Caller ID |
| Novimex Executive Leather Armchair | Adjustable |
| Chromcraft Conference Table | Fully Assembled |
| Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Ma... | None |
| Chromcraft Bull-Nose Wood Oval Conference Tables & Bases | None |
| Martin Yale Chadless Opener Electric Letter Opener | None |
| Bevis Conference Table | Fully Assembled |
| Cisco Smart Phone | with Caller ID |
| Harbour Creations Executive Leather Armchair | Adjustable |
| KitchenAid Microwave | White |
| Breville Refrigerator | Red |
| Logitech diNovo Edge Keyboard | None |
| Hoover Stove | Red |
| Brother Fax Machine | High-Speed |
| KitchenAid Stove | White |
| Hon Computer Table | with Bottom Storage |
| Apple iPhone 5S | None |
| SAFCO Executive Leather Armchair | Black |
| KitchenAid Refrigerator | Black |
| Motorola Smart Phone | Full Size |
| Hon Computer Table | Fully Assembled |
| Lesro Conference Table | with Bottom Storage |
| Hoover Stove | Red |

*Figure 24: Showing the Split column and the new column called Description*

When the product name column is split based on delimiter which is comma (,) then some of the values in description are found null. This null value is replaced by the None because their description is already present in the product name column.

### 2.4.1 Data Analytics Expression (DAX)

A new custom column called Unit price was created using the Dax formula below

Unit Price = [Sales]/[Quantity]

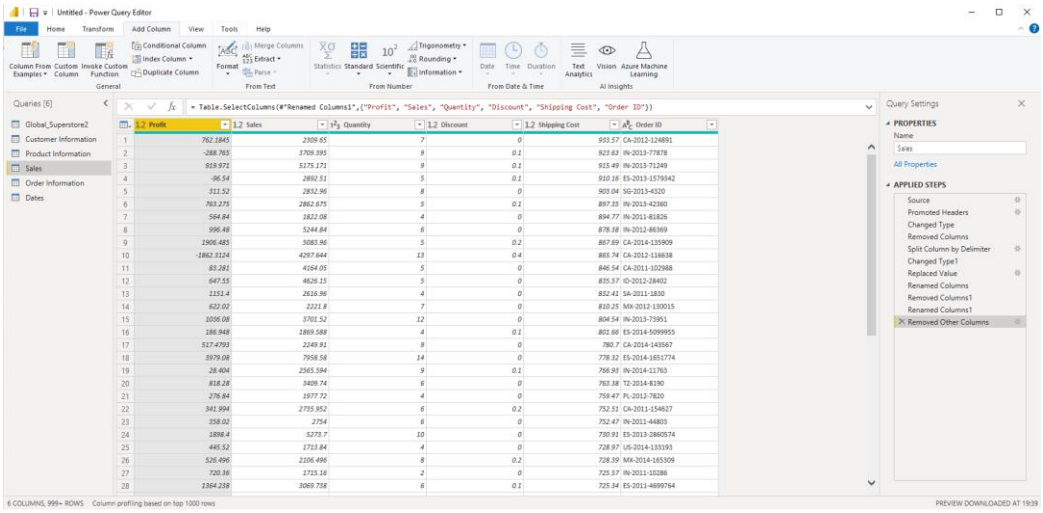

*Figure 25: Showing the Sales table before the addition of unit price column*

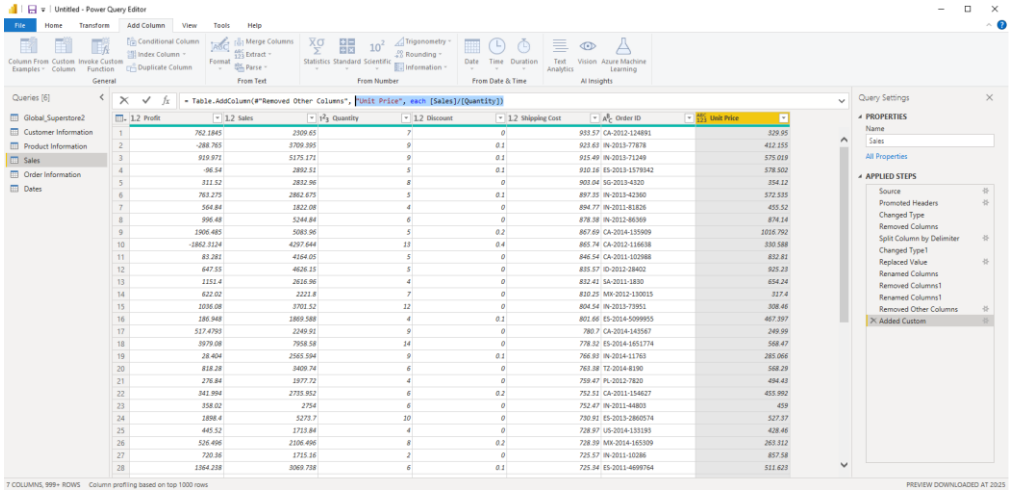

*Figure 26: Showing the sales table after the Unit price column was created*

Four New Measures were created using Dax formular on the sales table as shown below.

1. Total Sales = SUM(Global_Superstore2[Sales])
2. Total Revenue = SUM(Sales[Quantity])
3. Revenue = Sales[Sales] * Sales[Unit Price]
4. Profit Sales = CALCULATE(SUM(Global_Superstore2[Profit]))

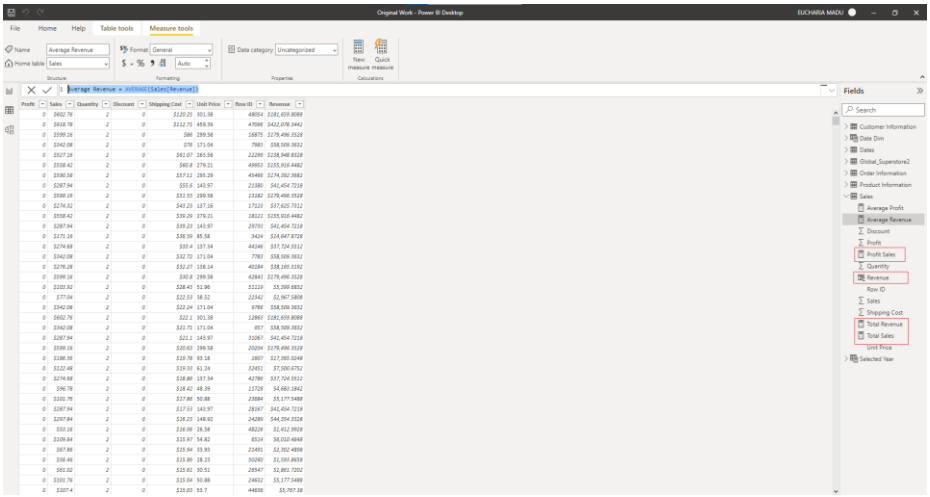

*Figure 27: Showing the 4 Measures created*

## 2.5. Dashboard

The Dashboard was designed to analyse the dataset and answer vital business questions.

### 2.5.1. Home Page

- The first page of the dashboard is the home page. The home page displays the details below.
- The Title of the project work
- The Course Tiltle and Course code
- My information (Name and Student ID)
- It contains buttons that work as page navigatiors for other pages of the dashboard
- A button that takes you to the next page
- Finally the school logo is also displayed at the bottom of the home page.

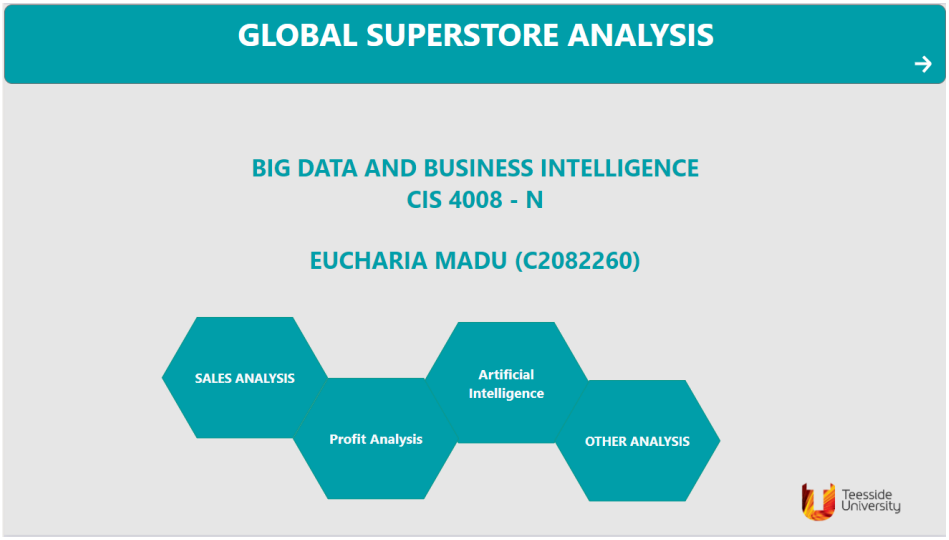


*Figure 28: Showing the Home page of the Dashboard*

### 2.5.2. Sales Analysis Page

The sales analysis page is the second page on the dashboard and it contains various charts and analysis as illustrated below.

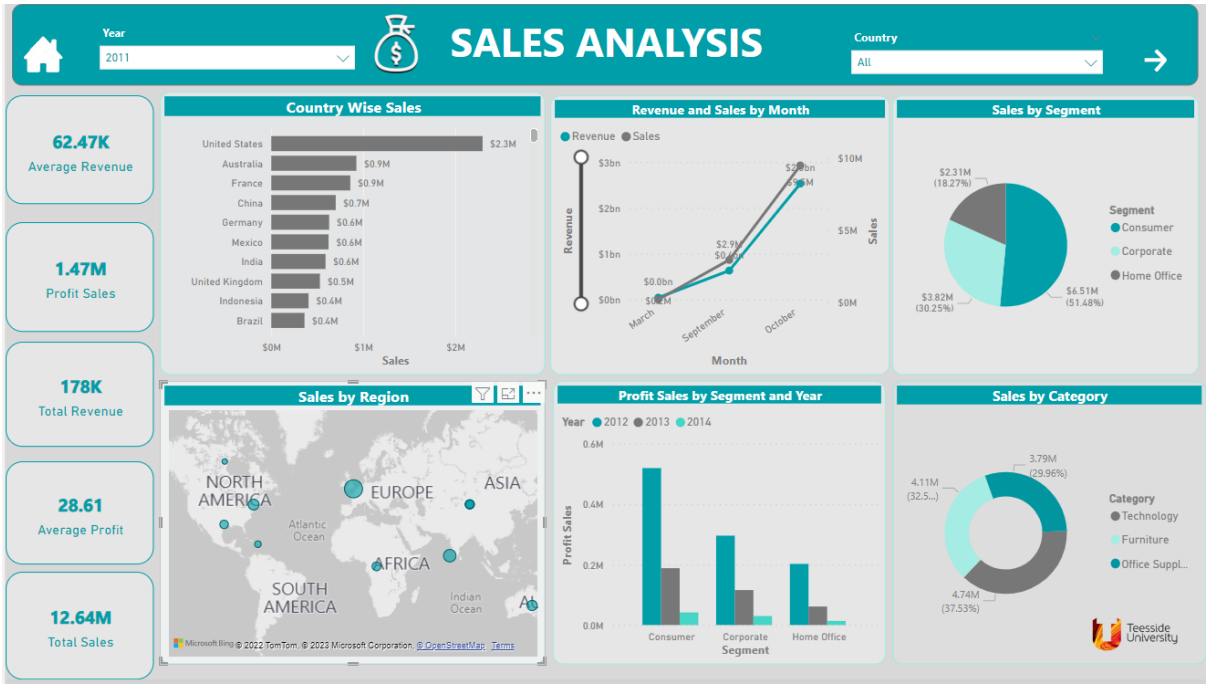


*Figure 29: Showing the sales Analysis page of the Dashboard*

The first visualisation on the sales analysis page is a clustered bar chart visualising country wise sales. The chart was plotted using Country as the x-axis and sales as the y-axis and it displays the amount of sales that happens per country on the data set. This sale is not the complete sale throughout the years, this is year wise sales, A slicer was used to filter by year. The figure below shows the sales for the year 2011 where United States is seen to have the highest sales across all the other countries. US is not only top in 2011, but it is also top in all the 4 years i.e., 2011, 2012, 2013 & 2014. There are several other countries as well but the graph only shows the top 10 countries based on their sales. Due to this it can be concluded that US is generating good revenue compared to the other sales.

The sales Analysis page contains the below visualisations

- A line chart was also used to do a comparison between Revenue and sales by month
- A Clustered bar chart was used to represent country wise sales
- A Pie chart was used to represent sales by segment
- A Map showing sales by region
- A Clustered Column chart used to represent profit sales by segment and year
- A doughnut Chart used to visualise sales by category
- 5 Cards showing different measures that assist to analyse the data
- 2 slicers that was used to filter the year and Countries.
- A home button that navigates to the home page
- A next button that navigates to the next page

### 2.5.3. Profit Analysis Table

The Profit Analysis page displays various visualization plotted with regards to profit, the different chart used are listed below.

- An Area Chart showing count of Order by Category
- Stacked column chart showing the Top 5 countries by profit
- A Tree map that visualises the top 5 profit making products
- A clustered bar chart showing the top 3 segments by profit
- A Clustered column chart displaying the top 5 products by profit
- A home button that navigates to the home page
- A next and a back button that navigates to the next and previous page respectively.

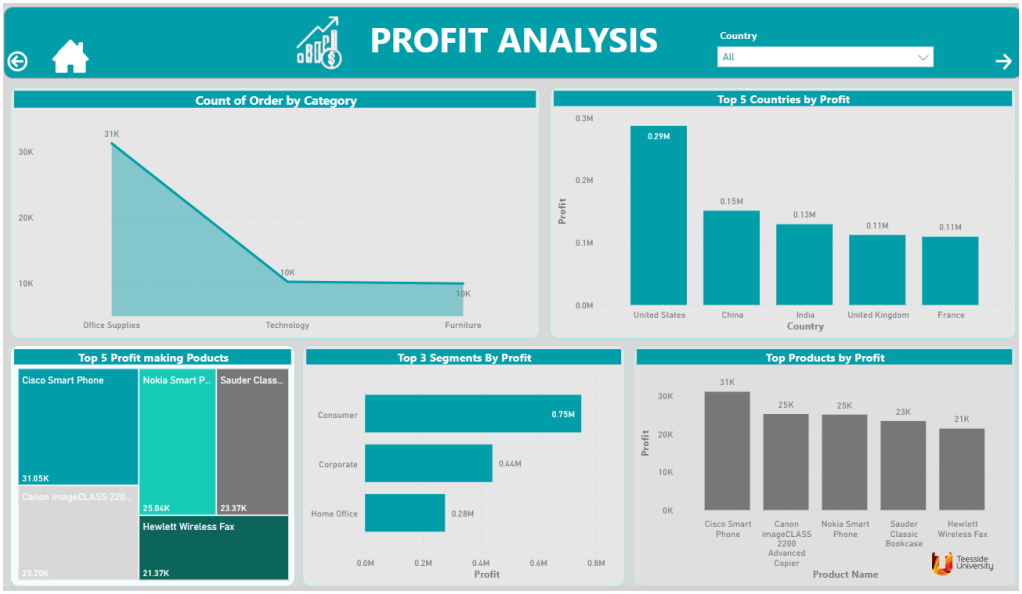


*Figure 30: Showing the Profit Analysis Page of the Dashboard*

### 2.5.4. Artificial Intelligence

A decomposition tree was used to analyse sales explained by category, product name, country, city and customer name.

Key influencers was also used to visualise the key influencers with regards to revenue and also to know the key influencers based on segment.

A home button that navigates to the home page, a next button that navigates to the next page and a back button that navigates to the previous page was also added to this page.

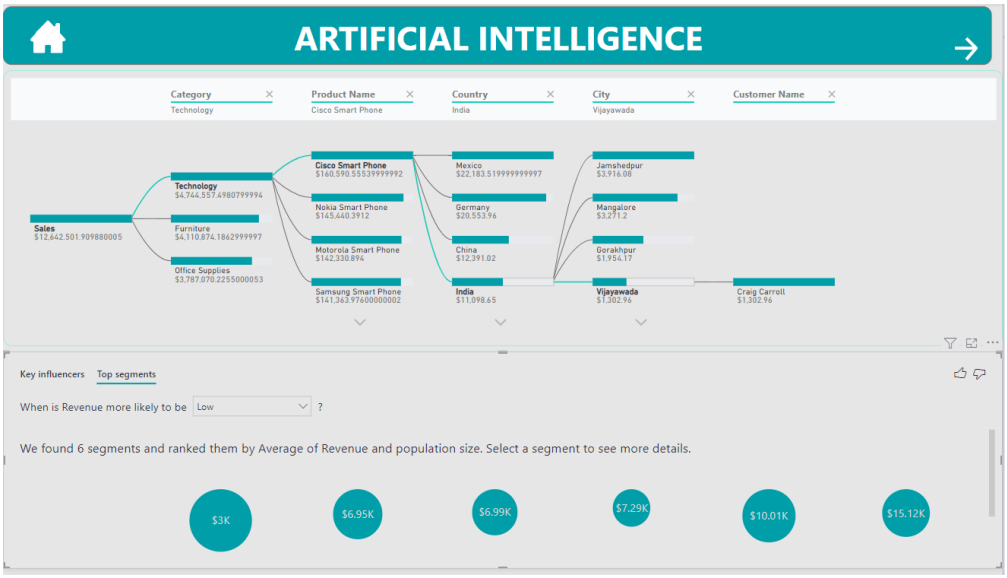


*Figure 31: Showing the Artificial Intelligence page of the Dashboard*

### 2.5.5. Further Analysis

On this page, the below visualizations were done.

- An Animated Bar chart Race that represents total sales by product and year
- A Drill down radial barchart that was used to represent the product quantity purchased per segment
- A waterfall chart was used to represent profit by category and discount
- A back button that navigates to the previous page
- A home button that navigates to the home page

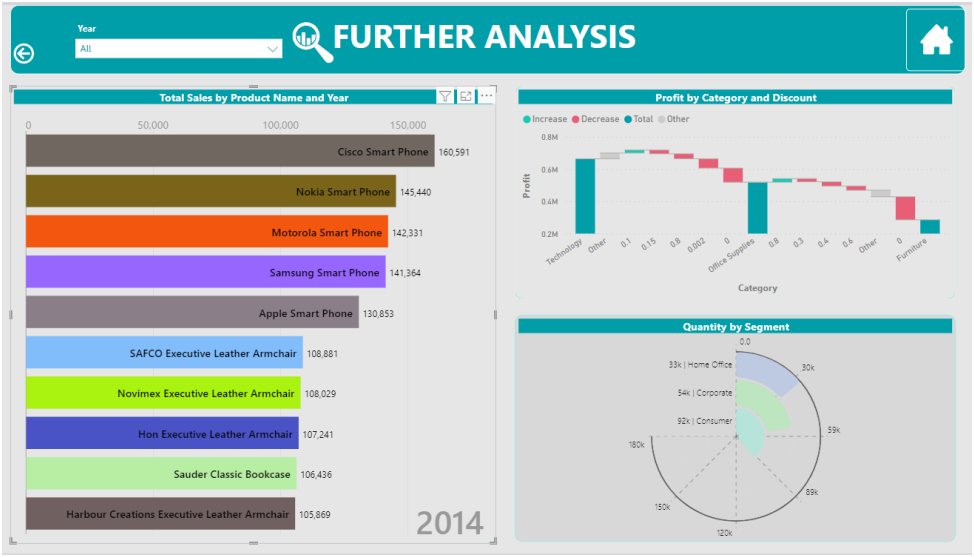


*Figure 32: Showing the Further Analysis page of the Dashboard*

| Report Section | Description | Grade your work from 0 to 100 |
|---|---|---|
| Report Structure | The report is well-written, and it contains all the relevant sections | 95 |
| Data Pre-processing and Data Modelling | Many pre-processing steps have been applied. The data model is well-structured | 85 |
| Dax and M language | Both DAX and M Language have been extensively used in the report | 80 |
| Dashboard Design | The dashboard contains a variety of charts, including advanced ones. | 95 |
| Average | | Add below the average of the four cells above:<br><br>88.75 |