

ICA

**BIG DATA AND BUSINESS
INTELLIGENCE**

NAME: AYOOLUWA DORCAS BABALOLA

STUDENT ID: B1202001

Table of Contents

SECTION 1	3
BUSINESS INTELLIGENCE DESIGN.....	3
1 BI Source Description and Questions.....	3
1.1 Data Source Description	3
1.2 Objectives and Approach	3
1.3 Scope.....	4
1.4 Rationale	4
1.5 Outcomes.....	4
1.6 Description of the Dataset	4
2 BI Data Pre-Processing and Data Cleansing	6
2.1 Loading the Data	6
2.2 Data Pre-processing and Cleaning	8
2.2.1 Removing Null Values	9
2.2.2 Changing Data Type	11
2.3 Generating Fact and Dimension Tables	15
2.3.1 Creating Location Dimension Table	17
2.3.2 Further Pre-processing.....	22
3 Data Modelling.....	26
SECTION 2	32
BUSINESS REPORT	32
1 EXECUTIVE SUMMARY	33
2 INTRODUCTION.....	34
3 NEW VISUALIZATION USED IN THIS PROJECT:	35
4 KEY FINDINGS.....	38
5 REPORT SUPPORT INFORMATION	50

SECTION 1

BUSINESS INTELLIGENCE DESIGN

1 BI Source Description and Questions

1.1 Data Source Description

The dataset used in this project is called "superstore_data" and is available for free and public use on Kaggle. The following is a link to the dataset:

[superstore_data | Kaggle](#)

The superstore data collection has about 52000 instances with 24 attributes enclosed in three tables. It is a customer-centric data set that includes information on all orders placed through various vendors and markets from 2011 to 2015 in different countries.

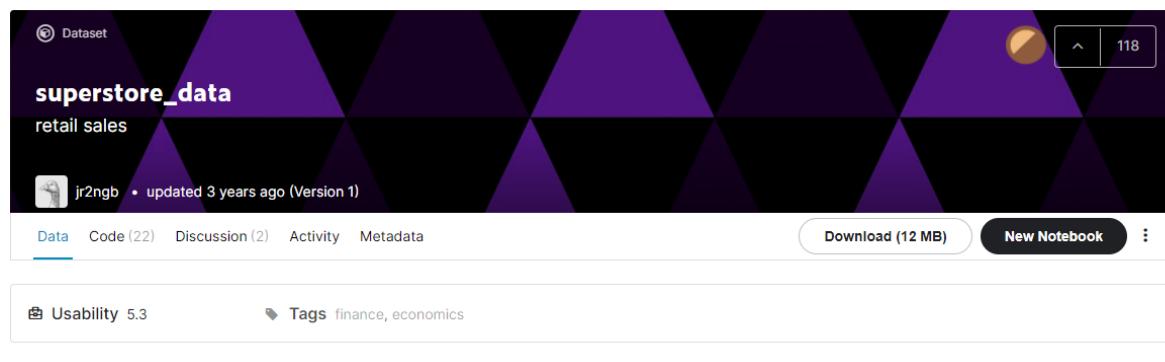


Fig: Pictorial representation of the data source

1.2 Objectives and Approach

A superstore is a major industry that relies on cutting-edge technology and data analytics to maintain its competitive advantage. It's critical that they enlist strategic tools to deliver meaningful insights into their sales in their numerous centres, broken down by product category and subcategory.

Power BI and its report were utilised in this BI project to deliver insights and the analysis focuses on the following aspects:

- predicting the superstore's sales and profit trends
- to assess the product performance
- to assess the store performance based on location

- to calculate the order cycle of orders i.e. product delivery time
- to understand customers' preferences as regards category of products
- to determine customer's most preferred ship mode

1.3 Scope

This project follows the described strategy, except in section 1 it focuses on data pre-processing and data modelling. The major goal of section 2 is to provide business insights and analytical reports utilising interactive dashboards to enhance performance evaluation based on the questions above.

1.4 Rationale

For two reasons, this dataset was chosen for this study. For starters, Power BI appears to be the best business intelligence solution for this dataset because it is both user-friendly and cost-effective. It's also the ideal tool for creating sophisticated visualisations for the superstore's sales performance analysis. The dashboards and report can be customised to meet the needs of the store and its future. Second, the dataset is commercial and contains genuine data that may be utilised to develop analytical applications in a real-world setting. This is an excellent opportunity for me to hone my business analytics skills by putting them to use while preparing for future positions.

1.5 Outcomes

These skills would have been fully shown by the end of this project:

- Data pre-processing and cleaning such as importing data, generating new columns, altering data types, and deleting null values.
- Creating and managing relationships as a method of data modelling
- Power query with M language
- Use DAX to carry out computations
- Use of various visualisations and Artificial Intelligence tools to do critical analysis
- Using the data from the report to create an interactive dashboard
- Report design and publication

1.6 Description of the Dataset

The dataset has three tables, and all of them were used. In tabular form, the tables and their attributes are summarised below.

Column	Definition
Row ID	Unique ID of each row
Order ID	Unique identifier of an order
Order Date	Date when the customer placed their order
Ship Date	Actual shipping date to the customer
Ship Mode	Mode of shipping of each order
Customer ID	Customer ID number
Customer Name	Names of individual customers
Segment	Segment to which each product belongs
Postal Code	Post code of each customer
City	The city where the customer lives
State	State where the customer lives
Country	Country name of each customer
Region	Region where each customer is based
Market	Market to which the customer belongs
Product ID	Unique identifier of a product
Sub-Category	Sub-category to which each product belongs
Product Name	Unique name of each product

Sales	Sales value of each transaction
Profit	Gain value of each transaction
Quantity	The quantities of each product per transaction
Discount	Deduction on each transaction
Order Priority	Priority of each unique order
Shipping cost	Cost of shipping each unique order

The dataset is contained in a CSV file and the figure below shows a screenshot of the dataset

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Segment	City	State	Country	Postal Code	Market	Region	Product ID	Category	Sub-Category	Product N	Sales	Quantity	Discount	Profit	Shipping	Order Pric
2	42433	AG-2011-2040	01/01/2011	06/01/2011	Standard	ITB-11280	Toby Brau Consumer	Constanti/Constantine		Algeria		Africa	OFF-TEN-10000025	Office Svc Storage	408.3	2	0	106.14	35.46	Medium			
3	22253	IN-2011-47883	01/01/2011	08/01/2011	Standard	IJH-15985	Joseph Hc Consumer	Wagga Wi New South Wales		Australia		APAC	OFF-SU-10000018	Office Svc Supplies	Acme Trin	120.366	3	0.1	36.036	9.72	Medium		
4	48853	IN-2011-364782	01/01/2011	03/01/2011	Second	CDX-11410	Eugene M Home Office	Stockholm/Stockholm		Sweden		EMEA	OFF-PR-10000085	Office Svc Storage	Tenex Box	66.12	4	0	29.64	8.17	High		
5	22255	IN-2011-364782	01/01/2011	03/01/2011	Second	CDX-11410	Eugene M Home Office	Stockholm/Stockholm		Sweden		EMEA	OFF-PR-10000085	Office Svc Paper	Office Paper	44.486	3	0.1	-26.05	4.87	High		
6	22255	IN-2011-47883	01/01/2011	08/01/2011	Standard	IJH-15985	Joseph Hc Consumer	Wagga Wi New South Wales		Australia		APAC	OFF-FU-10003447	Furniture	Furnishing Eldon Ligh	113.67	5	0.1	37.77	4.7	Medium		
7	22254	IN-2011-47883	01/01/2011	08/01/2011	Standard	IJH-15985	Joseph Hc Consumer	Wagga Wi New South Wales		Australia		APAC	OFF-FU-10001948	Office Svc Paper	Eaton Con	55.242	2	0.1	15.342	1.1	Medium		
8	21613	IN-2011-30733	01/02/2011	03/02/2011	Second	PD-18865	Patrick O'Consumer	Dhaka	Dhaka	Bangladesh		APAC	Central AsTec-CO-10002316	Technolo Copiers	Brother Pi	285.78	2	0	71.4	57.3	Critical		
9	34662	CA-2011-115168	01/02/2011	03/02/2011	First Class	LC-17050	Uz CarliShi Consumer	Mission V California		United States	92091	US	West	PUR-BO-10003916	Furniture	Bookcase/Sauder Fa	290.666	2	0.15	3.4196	54.64	High	
10	44508	AO-2011-1390	01/02/2011	04/02/2011	Second	DC-3115	David Ken Corporate	Luanda	Luanda	Angola		Africa	OFF-FEL-10001541	Office Svc Storage	Fellowes I	206.4	1	0	92.88	53.08	Critical		
11	23688	ID-2011-56493	01/02/2011	04/02/2011	Second	CD-20850	Stephanie Hc Consumer	Yingcheng/Hubei		China		APAC	North Asia-Off-ST-10002161	Office Svc Storage	Tenex Tra	162.72	3	0	68.31	44.36	Critical		
12	25293	IN-2011-36074	01/02/2011	05/02/2011	Second	CD-11350	David Ken Corporate	Chongqin/Chongqing		China		APAC	North Asia-Off-AP-10002534	Office Svc Supplies	KitchenAri	352.35	5	0	137.4	33.15	Medium		
13	24843	US-2011-65509	01/02/2011	05/02/2011	Second	PO-18860	Patrick O'Consumer	San Miguel/Panama		Panama		LATAM	Central-Off-AP-10002534	Office Svc Storage	GlobeWe	290.666	2	0.4	20.024	21.38	Medium		
14	24843	US-2011-65509	01/02/2011	05/02/2011	Second	PO-18860	Patrick O'Consumer	Marlowe/Marlowe/Khorasan		Iran		EMEA	OFF-ADY-10002040	Office Svc Storage	Hamilton	400.75	6	0	102.04	100.07	Medium		
15	16777	ES-2011-526849	01/02/2011	03/02/2011	Second	CO-11485	Gene Hall Corporate	La Rochelle/Poitou-Charentes		France		EMEA	OFF-AR-10001539	Office Svc Art	Bonney & C	139.65	5	0	15.3	15.29	High		
16	21615	IN-2011-30733	01/02/2011	03/02/2011	Second	PD-18865	Patrick O'Consumer	Dhaka	Dhaka	Bangladesh		APAC	Central-Off-SU-10000484	Office Svc Supplies	Kleeneut I	40.68	3	0	11.79	11.13	Critical		
17	8484	US-2011-118892	01/02/2011	06/02/2011	Standard	LOM-13075	Dave Hall Corporate	San Miguel/Panama		Panama		LATAM	Central-TEC-10001221	Technolo Accessori Memori	81.984	2	0.4	-19.136	6.21	Medium			
18	15796	ES-2011-5460465	01/02/2011	05/02/2011	Standard	FR-151315	Ralph Ritt Consumer	Parma	Emilia-Romagna	Italy		EMEA	OFF-AR-10000980	Office Svc Paper	Sandford Pi	78.3	3	0	20.34	6.03	Medium		
19	21614	IN-2011-30733	01/02/2011	03/02/2011	Second	PD-18865	Patrick O'Consumer	Dhaka	Dhaka	Bangladesh		APAC	Central-Off-BI-10003012	Office Svc Binders	Wilson Pi	22.65	5	0	9.6	5.29	Critical		
20	21616	IN-2011-30733	01/02/2011	03/02/2011	Second	PD-18865	Patrick O'Consumer	Dhaka	Dhaka	Bangladesh		APAC	Central-Off-FA-10001292	Office Svc Labels	Smedai Fl	20.34	3	0	9.9	3.76	Critical		
21	16726	ES-2011-5268483	01/02/2011	03/02/2011	Second	GH-14845	Gene Hall Corporate	La Rochelle/Poitou-Charentes		France		EMEA	Central-Off-EN-10004597	Office Svc Envelopes	GlobeWei	21.39	1	0	0	3.34	High		
22	14413	ES-2011-2205482	01/02/2011	07/02/2011	Standard	IM-13055	Iona/MG Consumer	Halle/North Rhine-Westphalia		Germany		EMEA	OFF-DEY-10002040	Office Svc Art	Aczo Hole	21.06	3	0	10.53	1.86	Medium		
23	24843	US-2011-118892	01/02/2011	06/02/2011	Standard	DIH-13079	Dave Hall Corporate	San Miguel/Panama		Panama		LATAM	Central-Off-PR-10000710	Office Svc Binders	Wing Po	5.576	6	0	-4.394	0.83	Medium		
24	24843	US-2011-118892	01/02/2011	06/02/2011	Standard	DIH-13079	Dave Hall Corporate	San Miguel/Panama		Panama		EMEA	OFF-FEL-10001710	Office Svc Storage	Fellowes I	55.16	4	0	71.64	164.36	High		
25	44228	CA-2011-1800	01/03/2011	04/03/2011	First Class	TP-111415	Tom Prey Consumer	Toronto	Ontario	Canada		EMEA	OFF-FEL-10001405	Office Svc Storage	Fellowes I	1244.16	6	0	211.5	60.76	Medium		
26	13130	ES-2011-1705541	01/03/2011	06/03/2011	Standard	TS-21370	Todd Sum Corporate	Farnborou	England	United Kingdom		EMEA	OFF-BO-10002259	Furniture	Bookcase/Safco Clas	1314.45	3	0	341.73	150.4	High		
27	48595	UP-2011-3730	01/03/2011	05/03/2011	Standard	RD-9900	Ruben Daio Consumer	Vinnitsya/Vinnitsya		Ukraine		EMEA	TEC-LOG-10003896	Technolo Accessori Logitech R	1470.78	6	0	264.6	146.53	Medium			
28	15218	ES-2011-3833440	01/03/2011	05/03/2011	Standard	TB-21400	Tony Boe/Consumer	Berlin	Berlin	Germany		EMEA	OFF-AP-10002546	Office Svc Appliance	Hamilton	364.416	8	0.2	45.454	80.87	High		
29	37844	CA-2011-113880	01/03/2011	05/03/2011	Standard	TP-111415	Ticky Prey Home Office	Elmhurst	Illinois	United States	60126	US	Central-FUR-CH-10000863	Furniture	Chairs Novitex I	634.116	6	0.3	-172.117	70.05	High		
30	44230	CA-2011-1800	01/03/2011	04/03/2011	First Class	TP-111415	Tony Prey Consumer	Toronto	Ontario	Canada		EMEA	FUR-HAR-10001792	Furniture	Chairs Harbour C	246.48	4	0	105.5	65.81	High		
31	14413	ES-2011-1705541	01/03/2011	06/03/2011	Standard	TS-21370	Todd Sum Corporate	Farnborou	England	United Kingdom		EMEA	FUR-CH-10002830	Furniture	Chairs Office Sta	704.55	5	0	288.75	64.4	High		
32	48595	UP-2011-3730	01/03/2011	06/03/2011	Standard	RD-9900	Ruben Daio Consumer	Vinnitsya/Vinnitsya		Ukraine		EMEA	OFF-CH-10002857	Office Svc Storage	Fellowes I	1244.16	6	0	211.5	60.76	Medium		
33	14545	ES-2011-5268483	01/03/2011	05/03/2011	First Class	TP-111405	Ruthie Dan Corporate	Audra/Australia		Australia		EMEA	OFF-MAX-10002522	Furniture	Tables Global	218.98	2	0.1	22.0	53.07	Medium		
34	14545	CA-2011-104269	01/03/2011	06/03/2011	Second	CD-18360	Dave Bro Consumer	Seattle	Washington	United States	98115	US	West-FUR-CH-10004063	Furniture	Chairs Global	457.568	2	0.2	51.4764	47.89	Medium		
35	48595	UP-2011-3730	01/03/2011	05/03/2011	Standard	RD-9900	Ruben Daio Consumer	Vinnitsya/Vinnitsya		Ukraine		EMEA	TEC-SHA-10004874	Technolo Copiers	Sharp Fax	587.7	2	0	123.36	42.88	Medium		
36	39607	CA-2011-168312	01/03/2011	07/03/2011	Standard	GW-14405	Giulietta I Consumer	Houston	Texas	United States	77036	US	Central-FUR-CH-10001866	Furniture	Tables Bevis Rou	376.509	3	0.3	-43.0296	32.7	Medium		
37	13132	FS-2011-1705541	01/03/2011	06/03/2011	Standard	TD-21370	Todd Sum Corporate	Farnborou	England	United Kingdom		EMEA	OFF-AR-10004511	Office Svc Art	RIC/Sketch	194.64	4	0	34.52	25.39	High		

Fig: Pictorial illustration of the data file

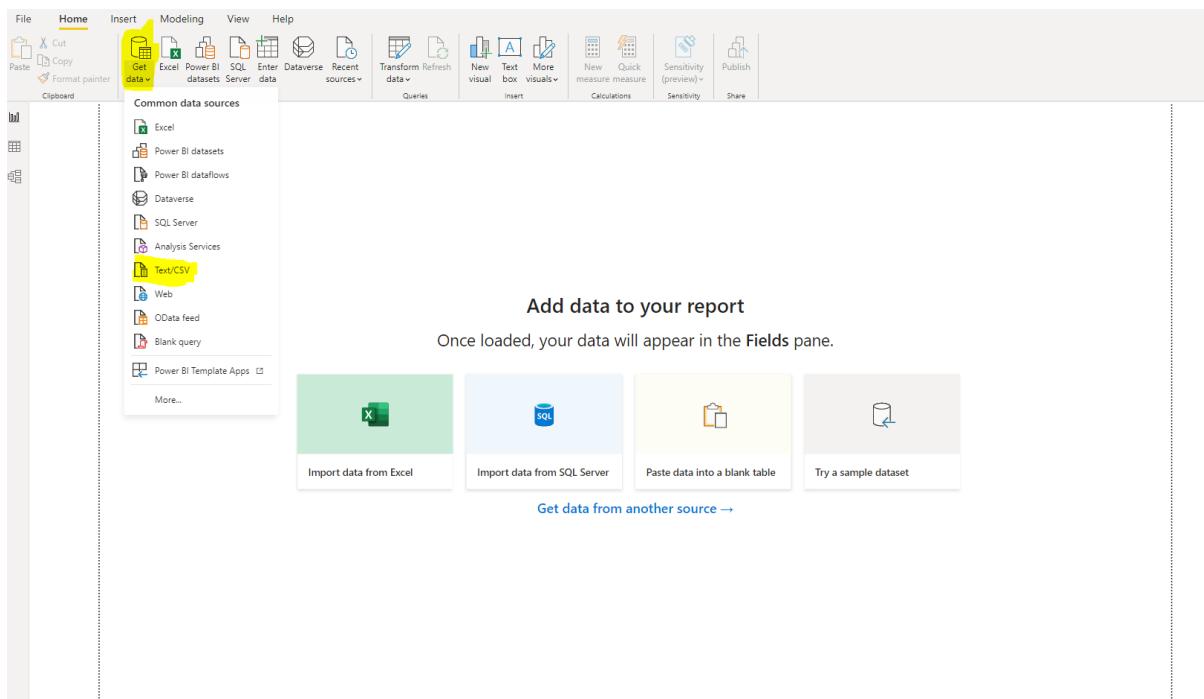
2 BI Data Pre-Processing and Data Cleansing

2.1 Loading the Data

The initial step in performing data analysis and developing business intelligence is to load the raw data. Excel, csv, space separated format, the online API,

databases, and a variety of other sources and files are supported by Power BI to import datasets.

When the application is launched, a dialogue box appears, allowing you to load an existing project or use the '**Get data**' button on the '**Home**' ribbon. The diagram below depicts the framework through which we can import data and how to load it from multiple sources. The data for this dataset came from a csv file.



The data is then loaded into Power BI after the dataset has been imported. The loading of data is depicted in the diagram below.

The screenshot shows the Microsoft Power BI Data Editor interface. The ribbon at the top has tabs for File, Home, Insert, Modeling, View, and Help. The 'Home' tab is currently selected. Below the ribbon, there are several data import options: Get data (with sub-options for Excel, Power BI datasets, SQL Server data, Enter data, Dataverse, and Recent sources), Transform Refresh, New visual, Text box, More visuals, New measure, Quick measure, Sensitivity (preview), Publish, and Share. The 'Data' tab is also selected. In the center, a modal window titled 'superstore_dataset2011-2015.csv' is open, showing a preview of the dataset. The preview includes a header row and 20 sample rows. The columns listed are Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, City, and State. The 'Data Type Detection' dropdown shows 'Comma' and 'Based on first 200 rows'. At the bottom of the modal, there are buttons for 'Extract Table Using Examples', 'Load', 'Transform Data', and 'Cancel'. On the left side of the main Power BI window, there is a green sidebar with the text 'Import data from' and an 'Excel' icon.

Then the data is fully loaded

File	Home	Help	Table tools														
Name	superstore_dataset...	Mark as date table	Manage relationships														
Structure	Calendars	New measure	Quick measure														
Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	Country	Postal Code	Market	Region	Product ID	Category	Sub-Category	Product Name
17897	ES-2021-1457147	03 November 2021	07 November 2021	Standard Class	CD-1950	Michael Osman	Consumer	Belfort	Franche-Comté	France	EU	Central	Office Supplies	Binders	Wilson Jones Binding D		
15040	ES-2021-5665333	03 August 2021	10 August 2021	Standard Class	KB-16315	Karen Braun	Consumer	Lis Willingam	Le Blanc-Mesnil	France	EU	Central	Office Supplies	Binders	Illico 3-Hole Punch, C		
18948	ES-2021-4472315	07 April 2021	13 April 2021	Standard Class	UW-1725	Lisa Williamson	Consumer	Le Blanc-Mesnil	France	EU	Central	Office Supplies	Binders	Cardinal Hole Binder, Clear			
11076	ES-2021-1724946	01 December 2021	07 December 2021	Standard Class	CD-1950	Carlos Daly	Consumer	Saint-Malo	Brittany	France	EU	Central	Office Supplies	Binders	Cardinal Binder, Recycl		
11076	ES-2021-1724946	01 December 2021	07 December 2021	Standard Class	CD-1950	Carlos Daly	Consumer	Saint-Malo	Brittany	France	EU	Central	Office Supplies	Binders	Acco Binder Covers, Cl		
12558	ES-2021-2555887	06 June 2022	12 June 2022	Standard Class	SS-20590	Sonia Sunley	Consumer	Palaisseu	Île-de-France	France	EU	Central	Office Supplies	Binders	Acco Index Tab, Clear		
18861	ES-2021-4472315	10 August 2022	14 August 2022	Standard Class	SD-18075	Sung Sharlai	Consumer	Songtou	Provence-Alpes-Côte d'Azur	France	EU	Central	Office Supplies	Binders	Cardinal Hole Reinforc		
14257	ES-2021-4229484	10 November 2022	14 November 2022	Standard Class	RB-18933	Randy Bradley	Consumer	Watrellos	Nord-Pas-de-Calais	France	EU	Central	Office Supplies	Binders	Wilson Jones Binder, D		
18021	ES-2021-2445045	10 November 2022	14 November 2022	Standard Class	RA-19904	Ryan Akim	Consumer	Montrion	Rhône-Alpes	France	EU	Central	Office Supplies	Binders	Illico Binder Covers, D,		
16801	ES-2021-2029398	08 June 2013	13 June 2013	Standard Class	SV-2085	Seth Vernon	Consumer	Beyonne	Aquitaine	France	EU	Central	Office Supplies	Binders	Cardinal Binder Covers		
13551	ES-2021-1724946	15 June 2013	20 June 2013	Standard Class	CD-1950	Gary Heung	Consumer	Boulogne-sur-Mer	Nord-Pas-de-Calais	France	EU	Central	Office Supplies	Binders	Wilson Jones Binding D		
16851	ES-2021-1724946	01 December 2021	07 December 2021	Standard Class	SD-18075	Stephen Thompson	Consumer	Montreuil	Île-de-France	France	EU	Central	Office Supplies	Binders	Wilson Jones Binding D		
17008	ES-2021-4482031	02 August 2024	07 August 2024	Standard Class	AC-18960	Anne Chung	Consumer	Grasse	Provence-Alpes-Côte d'Azur	France	EU	Central	Office Supplies	Binders	Wilson Jones Binder, E		
10021	ES-2021-4482037	04 August 2024	10 August 2024	Standard Class	CM-17845	Michael Chen	Consumer	Sorèze	Île-de-France	France	EU	Central	Office Supplies	Binders	Illico Hole Tief, Clear		
16350	ES-2021-2244655	06 April 2024	10 April 2024	Standard Class	TB-21335	Tilly Ritter	Consumer	Clichy	Île-de-France	France	EU	Central	Office Supplies	Binders	Acco Binding Machine		
18402	ES-2021-3989358	06 October 2024	12 October 2024	Standard Class	HD-14993	Henia Zeffil	Consumer	Reusvill	Picardy	France	EU	Central	Office Supplies	Binders	Acco 3-Hole Punch, C		
18402	ES-2021-3989358	06 October 2024	12 October 2024	Standard Class	HD-14993	Henia Zeffil	Consumer	Reusvill	Picardy	France	EU	Central	Office Supplies	Binders	Illico Binder Covers, E		
16812	ES-2021-117053	07 November 2024	12 November 2024	Standard Class	LT-17110	Li Thompson	Consumer	Nice	Provence-Alpes-Côte d'Azur	France	EU	Central	Office Supplies	Binders	Acco 3-Hole Punch, Ec		
18898	ES-2021-1418330	21 September 2024	27 September 2024	Standard Class	MN-17935	Michael Nguyen	Consumer	Dijon	Burgundy	France	EU	Central	Office Supplies	Binders	Illico Binder Covers, Ec		
18522	ES-2024-2742229	09 June 2024	15 June 2024	Standard Class	BD-11330	Bill Donated	Consumer	Landerneau	Brittany	France	EU	Central	Office Supplies	Binders	Illico Indre Tab, Clear		
15478	ES-2024-1020861	01 December 2024	18 December 2024	Standard Class	PH-14381	Fred Harton	Consumer	Champlieu-sur-Marme	Île-de-France	France	EU	Central	Office Supplies	Binders	Cardinal Indre Hole, Recycl		
15804	ES-2024-2018801	10 December 2024	14 December 2024	Standard Class	WB-21850	William Brown	Consumer	Oyonnax	Rhône-Alpes	France	EU	Central	Office Supplies	Binders	Wilson Jones 3-Hole Pi		
12537	ES-2024-1181938	12 February 2024	16 February 2024	Standard Class	AB-10015	Aaron Bergman	Consumer	Vincennes	Île-de-France	France	EU	Central	Office Supplies	Binders	Cardinal Indre Binder Covers		
15586	ES-2024-2024232	12 December 2024	17 December 2024	Standard Class	HD-18940	Roy Fransisco	Consumer	Luminoz	Picardy	France	EU	Central	Office Supplies	Binders	Avery Binder, Clear		
14847	ES-2024-1486782	13 August 2024	18 August 2024	Standard Class	TS-21340	Troy Sennfeld	Consumer	Millau	Mid-Pyrénées	France	EU	Central	Office Supplies	Binders	Wilson Jones Binder Cr		
12000	ES-2024-1486782	13 August 2024	18 August 2024	Standard Class	TS-21340	Troy Sennfeld	Consumer	Chambéry	Haute-Savoie	France	EU	Central	Office Supplies	Binders	Avery Binder, Clear		
14753	ES-2024-1516216	15 March 2025	20 March 2025	Standard Class	TS-21340	Troy Sennfeld	Consumer	Conceng	Nord-Pas-de-Calais	France	EU	Central	Office Supplies	Binders	Wilson Jones Indre Tab, Clear		
16924	ES-2024-3084941	15 May 2024	20 May 2024	Standard Class	CA-13775	Carina Amstein	Consumer	Meysse	Provence-Alpes-Côte d'Azur	France	EU	Central	Office Supplies	Binders	Avery Binder, Covers, Ec		
18977	ES-2024-4548007	15 November 2024	20 November 2024	Standard Class	LG-17140	Logan Currie	Consumer	Cholet	Pays de la Loire	France	EU	Central	Office Supplies	Binders	Avery 3-Hole Punch, Bi		
18977	ES-2024-4548007	15 November 2024	20 November 2024	Standard Class	LG-17140	Logan Currie	Consumer	Cholet	Pays de la Loire	France	EU	Central	Office Supplies	Binders	Avery Binding Machine		
19670	ES-2024-2718754	18 January 2025	23 January 2025	Standard Class	SF-20200	Sheri Foster	Consumer	Laon	Picardy	France	EU	Central	Office Supplies	Binders	Wilson Jones Indel Tab		
15591	ES-2024-3084941	15 January 2025	22 January 2025	Standard Class	DU-15630	Doug Jacob	Consumer	Paris	Île-de-France	France	EU	Central	Office Supplies	Binders	Avery Binder, Covers, R		
18881	ES-2024-2709384	15 September 2024	22 September 2024	Standard Class	KD-16270	Karen Daniels	Consumer	Argenteuil	Île-de-France	France	EU	Central	Office Supplies	Binders	Cardinal Indre Tab, Ecc		
18847	ES-2024-3084941	17 April 2024	24 April 2024	Standard Class	DC-12881	Dan Campbell	Consumer	Les Pavillons-sous-Bois	Île-de-France	France	EU	Central	Office Supplies	Binders	Avery 3-Hole Punch, Cl		
14421	ES-2024-3213169	21 September 2024	24 September 2024	Standard Class	EY-19815	Emily Phran	Consumer	Villejuif	Île-de-France	France	EU	Central	Office Supplies	Binders	Avery Index Tab, Durat		
11075	ES-2024-5064111	17 December 2022	22 December 2022	Standard Class	CA-12265	Christine Anderson	Consumer	Commercy-en-Parisis	Île-de-France	France	EU	Central	Office Supplies	Binders	Wilson Jones 3-Hole Pi		
18578	ES-2024-5111882	18 February 2023	23 February 2023	Standard Class	AB-10165	Alan Barnes	Consumer	Paris	Île-de-France	France	EU	Central	Office Supplies	Binders	Avery Hole Reinforc		

2.2 Data Pre-processing and Cleaning

Following the loading of the data, cleaning and pre-processing begins, which includes removing columns, removing null values, promoting column names,

altering data type, and more. We'll use the '**Transform Data**' button in the Home ribbon for this project, as seen in the figure below.

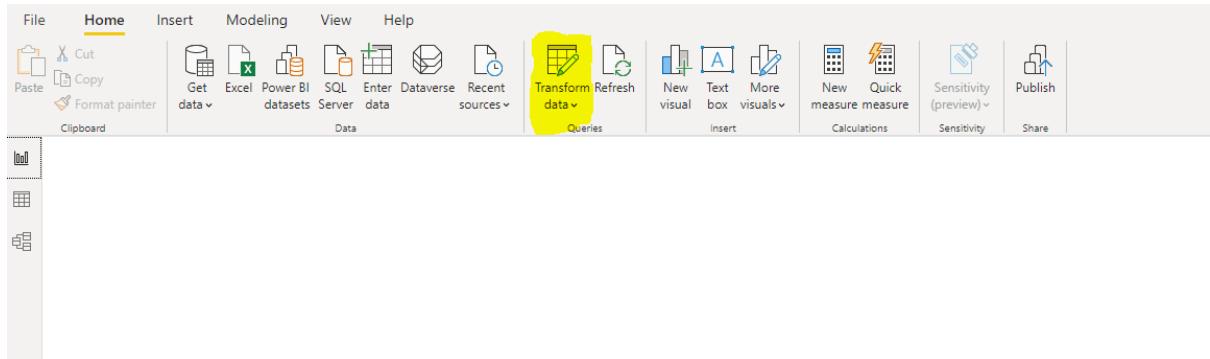


Fig: Selecting the Transform Data option

2.2.1 Removing Null Values

Firstly, we notice that the 'Postal code' column has many null values.

The screenshot shows the Microsoft Power Query Editor interface. A table is open with columns: State, Country, Postal Code, Market, Region, Product ID, Category, and Sub-Category. The 'Postal code' column contains many null values. The 'Applied Steps' pane on the right shows a step named 'Changed Type' applied to the 'Postal code' column. The 'Properties' pane shows the column name as 'superstore_dataset2011-2015[Postal code]'.

We then remove the null values by right clicking and using the '**fill up**' option as shown below

After which we close and apply to see the effect

The figure below shows the data after filling up the null values

2.2.2 Changing Data Type

Changing the data type is one of the pre-processing steps that some tables require. It is critical to update the data type of currency-related columns in this dataset from just whole number. The images below show how to change the data type.

Row ID	Sales	Quantity	Discount	Profit	Shipping Cost
39159	82	2	0	0	4.93
14372	56	2	0	0	1.6
31860	4	2	0	0	1.24
38055	82	2	0	0	9.4
41296	27	2	0	0	3.91
48226	53	2	0	0	16.06
43228	56	2	0	0	8.13
23644	228	2	0	0	8.4
51119	104	2	0	0	28.43
48100	24	2	0	0	1.84
45860	24	2	0	0	3.86
16900	100	2	0	0	5.46
50506	97	2	0	0	8.89
48054	603	2	0	0	120.25
5034	15	2	0	0	2.59
42091	111	2	0	0	5.8
50588	59	2	0	0	7.26
3820	9	2	0	0	1.64
48126	100	2	0	0	8.4
28748	26	2	0	0	1.53
12836	107	2	0	0	13.31
8663	59	2	0	0	8.12
49673	97	2	0	0	9.97
22849	28	2	0	0	1.86
19934	59	2	0	0	2.26
21380	288	2	0	0	55.6
49959	27	2	0	0	3.29
12863	603	2	0	0	22.1
5266	15	2	0	0	5.6
7783	342	2	0	0	32.72
43272	142	2	0	0	10.18
16826	22	2	0	0	1.06
42843	599	2	0	0	30.8
13439	44	2	0	0	1.32
12175	44	2	0	0	4.14
8503	20	2	0	0	1.16
24289	298	2	0	0	16.23
12153	18	2	0	0	1.32

Fig: Data type of sales before changing it

We select the drop-down arrow in the format and select currency since the sales is related to it and would be better for the analysis.

File Home Help Table tools Column tools

Name Sales Format Whole number \sum S
Data type Decimal number $\$ \vee \%$ General D
Structure Currency

Row ID Sales Quantity Discount Whole number Cost Lo

Row ID	Sales	Quantity	Discount	Whole number	Cost	Lo
39159	82	2		Percentage	4.93	
14372	56	2		Scientific	1.6	
31860	4	2			1.24	
38055	82	2	0	0	9.4	
41296	27	2	0	0	3.91	
48226	53	2	0	0	16.06	
43228	56	2	0	0	8.13	
23644	228	2	0	0	8.4	
51119	104	2	0	0	28.43	
48100	24	2	0	0	1.84	
45860	24	2	0	0	3.86	
16900	100	2	0	0	5.46	
50506	97	2	0	0	8.89	
48054	603	2	0	0	120.25	
5034	15	2	0	0	2.59	
42091	111	2	0	0	5.8	
50588	59	2	0	0	7.26	
3820	9	2	0	0	1.64	
48126	100	2	0	0	8.4	
28748	26	2	0	0	1.53	
12836	107	2	0	0	13.31	
8663	59	2	0	0	8.12	
49673	97	2	0	0	9.97	
22849	28	2	0	0	1.86	
19934	59	2	0	0	2.26	
21380	288	2	0	0	55.6	
49959	27	2	0	0	3.29	
12863	603	2	0	0	22.1	
5266	15	2	0	0	5.6	
7783	342	2	0	0	32.72	
43272	142	2	0	0	10.18	
16826	22	2	0	0	1.06	
42843	599	2	0	0	30.8	
13439	44	2	0	0	1.32	
12175	44	2	0	0	4.14	
8503	20	2	0	0	1.16	
24289	298	2	0	0	16.23	
12153	18	2	0	0	1.32	

Fig: Changing the type of sales column to currency

File Home Help Table tools Column tools

Name Sales Format Currency

Data type Decimal number \$ % .00 Auto

Structure Formatting

Row ID Sales Quantity Discount Profit Shipping Cost L

Row ID	Sales	Quantity	Discount	Profit	Shipping Cost	L
39159	\$81.96	2	0	0	4.93	
14372	\$56.45	2	0	0	1.6	
31860	\$3.96	2	0	0	1.24	
38055	\$81.96	2	0	0	9.4	
41296	\$27.18	2	0	0	3.91	
48226	\$53.16	2	0	0	16.06	
43228	\$56.45	2	0	0	8.13	
23644	\$227.82	2	0	0	8.4	
51119	\$103.92	2	0	0	28.43	
48100	\$24.42	2	0	0	1.84	
45860	\$24.42	2	0	0	3.86	
16900	\$99.72	2	0	0	5.46	
50506	\$96.9	2	0	0	8.89	
48054	\$602.76	2	0	0	120.25	
5034	\$14.8	2	0	0	2.59	
42091	\$111.12	2	0	0	5.8	
50588	\$59.1	2	0	0	7.26	
3820	\$9.32	2	0	0	1.64	
48126	\$99.72	2	0	0	8.4	
28748	\$25.56	2	0	0	1.53	
12836	\$107.4	2	0	0	13.31	
8663	\$58.6	2	0	0	8.12	
49673	\$96.9	2	0	0	9.97	
22849	\$28.44	2	0	0	1.86	
19934	\$59.1	2	0	0	2.26	
21380	\$287.94	2	0	0	55.6	
49959	\$27.18	2	0	0	3.29	
12863	\$602.76	2	0	0	22.1	
5266	\$15.32	2	0	0	5.6	
7783	\$342.08	2	0	0	32.72	
43272	\$142.08	2	0	0	10.18	
16826	\$21.54	2	0	0	1.06	
42843	\$599.16	2	0	0	30.8	
13439	\$44.34	2	0	0	1.32	
12175	\$44.34	2	0	0	4.14	
8503	\$20.4	2	0	0	1.16	
24289	\$297.84	2	0	0	16.23	
12153	\$18	2	0	0	1.32	

Table: SalesFact (51,290 rows) Column: Sales (22,995 distinct values)

Fig: screenshot of the output of the new data type of sales

This method was used to change the data type of profit, shipping cost and discount too.

Row ID	Sales	Quantity	Discount	Profit	Shipping Cost	
32964	\$74.352	3	\$0.20	\$23.24	\$22.49	
32965	\$14.04	3	\$0.20	\$1.58	\$3.69	
36997	\$40.752	3	\$0.20	\$15.28	\$2.49	
32962	\$4.344	3	\$0.20	\$0.87	\$1.42	
32356	\$4.416	3	\$0.20	\$1.60	\$0.31	
31850	\$10.824	3	\$0.20	\$2.57	\$0.88	
32555	\$540.048	3	\$0.20	(\$47.25)	\$65.74	
36445	\$842.376	3	\$0.20	\$105.30	\$70.22	
35917	\$90.48	3	\$0.20	\$33.93	\$15.95	
32803	\$17.88	3	\$0.20	\$5.59	\$1.07	
40152	\$21.312	3	\$0.20	\$7.99	\$1.07	
34612	\$14.376	3	\$0.20	\$4.85	\$0.98	
32827	\$74.352	3	\$0.20	\$23.24	\$5.16	
36401	\$16.776	3	\$0.20	\$5.03	\$2.4	
33173	\$470.376	3	\$0.20	\$47.04	\$20.65	
33911	\$230.28	3	\$0.20	\$23.03	\$34.42	
39127	\$311.976	3	\$0.20	\$39.00	\$21.99	
39125	\$63.936	3	\$0.20	\$6.39	\$4.7	
39128	\$50.352	3	\$0.20	\$17.62	\$3.33	
1701	\$271.488	3	\$0.20	\$84.83	\$14.2	
38185	\$73.344	3	\$0.20	\$27.50	\$4.08	
38036	\$102.624	3	\$0.20	\$7.70	\$7.26	
38038	\$13.392	3	\$0.20	\$3.18	\$1.02	
18547	\$315.432	3	\$0.20	(\$7.94)	\$52.44	
40424	\$50.136	3	\$0.20	(\$11.28)	\$1.65	
38446	\$971.88	3	\$0.20	\$109.34	\$95.63	
2337	\$852.48	3	\$0.20	\$213.12	\$149.86	
33291	\$11.304	3	\$0.20	(\$2.12)	\$0.56	
35135	\$674.352	3	\$0.20	(\$109.58)	\$45.54	
33640	\$123.552	3	\$0.20	(\$29.34)	\$9.62	
40110	\$266.352	3	\$0.20	\$13.32	\$25.36	
38731	\$9.144	3	\$0.20	\$3.09	\$0.5	
9179	\$223.2	3	\$0.20	(\$39.06)	\$28.28	
38080	\$37.752	3	\$0.20	\$4.25	\$2.8	
36765	\$585.552	3	\$0.20	\$73.19	\$179.73	
33839	\$143.952	3	\$0.20	\$14.40	\$7.77	
7146	\$57.36	3	\$0.20	(\$11.52)	\$5.66	
39023	\$331.536	3	\$0.20	(\$82.88)	\$33.17	

2.3 Generating Fact and Dimension Tables

Superstore_data is a single flat table with 24 columns and 51290 rows. Order ID, customer ID, customer name, nationality, product name, sales, discount, profit, category, and others are among the records in this table.

This table is renamed by right clicking in the transform data option to **superstore** before generating the fact and dimension tables.

The superstore table is now our flat table. It will be difficult to handle, maintain, and create business-related questions if all the data is contained in a single flat file or table. The table is also unusually huge because it contains duplicated records and information. We therefore split it into fact and dimension tables.

The screenshot shows the Microsoft Power Query Editor interface. The 'Queries' ribbon tab is selected. A context menu is open over the 'superstore' table, with the 'Duplicate' option highlighted. The 'Properties' pane on the right shows the table name as 'superstore'. The 'Applied Steps' pane shows a step named 'Filtered Rows'.

Fig: Creating duplicate table

Right-clicking on the cloned table and selecting the '**Rename**' option from the context menu renamed it into '**SalesFact**' which is the fact table for this model.

Then, we right click on superstore and delete while continuing the work with the fact table that would be formed.

The screenshot shows the Microsoft Power Query Editor interface. The 'Queries' list on the left contains two items: 'superstore' and 'SalesFact'. The 'SalesFact' item is highlighted with a red box. The main area displays the data from the 'SalesFact' table, which has 24 columns and over 999 rows. The columns include Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, and Segment. A preview of the data is shown at the bottom right.

Fig: Generating the fact table

This screenshot shows the same Power Query Editor interface as above, but with a different context menu open. The 'SalesFact' entry in the 'Queries' list has a red box around it. A context menu is open, and the 'Duplicate' option is highlighted with a red box. Other options in the menu include 'Copy', 'Paste', 'Delete', 'Enable load', 'Include in report refresh', 'Reference', 'Move To Group', 'Move Up', 'Move Down', 'Create Function...', 'Convert To Parameter...', 'Advanced Editor', and 'Properties...'. The main data grid and preview area are visible below the menu.

Fig: Remove the superstore table

In generating our dimension tables, we transform data and make duplicates of the fact table to select columns that would be essential for each dimension table. We go ahead to generate the first dimension table. We right-click the SalesFact table in the query and select ‘Duplicate’.

The screenshot shows the Microsoft Power Query Editor interface. The 'File' tab is selected in the ribbon. In the 'Queries' ribbon tab, a context menu is open over a row in the 'SalesFact' table. The 'Duplicate' option is highlighted with a red box. Other options visible in the menu include 'Copy', 'Delete', 'Rename', 'Enable load', 'Include in report refresh', 'Reference', 'Move To Group', 'Move Up', 'Move Down', 'Create Function...', 'Convert To Parameter', 'Advanced Editor', and 'Properties...'. The main table view shows a large dataset with columns like Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, and Segment. The 'APPLIED STEPS' pane on the right shows a step labeled 'Filtered Rows'.

Fig: Duplicate option in the query

2.3.1 Creating Location Dimension Table

A duplicate of the original flat table was generated by right clicking the '**Duplicate**' option and rename as shown below to construct the '**LocationDim**' dimension table. Columns relating to location such as county, region, state, city and market were left in this table while other columns were removed.

To minimise ambiguity, it is critical that the dimension tables have unique rows. We selected all the columns, then right clicked and selected '**Remove Duplicates**' which removed duplicate records from the table.

The screenshot shows the Microsoft Power Query Editor interface. A context menu is open over the 'LocationDim' table, specifically over the 'Region' column. The menu item 'Remove Duplicates' is highlighted with a red box. Other options in the menu include 'Remove Columns', 'Remove Other Columns', 'Add Column From Examples...', 'Replace Values...', 'Fill', 'Change Type', 'Merge Columns', 'Group By...', 'Unpivot Columns', 'Unpivot Only Selected Columns', and 'Move'. To the right of the menu, the 'APPLIED STEPS' pane shows a step named 'Removed Duplicates'. The main table view shows data from the 'SalesFact' and 'LocationDim' tables.

Fig: Removing duplicates from the location dimension table

There is a need for the dimension table to relate to the fact table, we there go to ‘**Add column**’ ribbon and select index from 1 which would correspond the columns in the location dimension and have a connection with the fact table.

The screenshot shows the Microsoft Power Query Editor interface. The 'Add Column' ribbon tab is selected, and the 'From 1' option is highlighted with a red box. Other options in the ribbon include 'Index Column', 'From Text', 'From Number', 'From Date & Time', and 'Custom...'. The main table view shows data from the 'SalesFact' and 'LocationDim' tables. The 'APPLIED STEPS' pane shows a step named 'Removed Duplicates'.

Fig: Adding a new column to the location dimension table

We go further by renaming the index column to ‘LocationInd’ standing for Location Index. This was regarded as the table’s primary key.

The screenshot shows the Power Query Editor interface with two queries: SalesFact and LocationDim. The SalesFact query contains columns: City, State, Country, Postal Code, Market, Region, and Index. A context menu is open over the 'Index' column, with the 'Rename...' option highlighted and a red box drawn around it. Other options like 'Remove', 'Remove Duplicates', and 'Replace Values...' are also visible.

Fig: Rename the index column to LocationInd

Furthermore, we go back to the SalesFact table and merge the new dimension table ‘LocationDim’ to it using the ‘merge queries’ option. We then select the common columns and select ‘OK’ to merge.

The screenshot shows the Power Query Editor with the 'Merge' dialog open. It displays two tables: SalesFact and LocationDim. The SalesFact table has columns: Row ID, Order ID, City, State, Country, Postal Code, Market, Region, Product ID, and Category. The LocationDim table has columns: LocationID, City, State, Country, Postal Code, Market, and Region. The 'Merge' dialog shows the selected columns for both tables. The 'Join Kind' dropdown is set to 'Left Outer (all from first, matching from second)'. The 'OK' button is highlighted with a red box. The 'APPLIED STEPS' pane on the right shows the step 'Filled Up'.

Fig: Matching columns between the sales fact table and the location dimension table.

Merging the location dimension table to the sales fact table reduces the model's complexity and creates a relationship between them. A table is formed by merging and we collapse the table by expanding with the '**LocationInd**' we have created. A table is created in another column and when expanded, only the primary key '**LocationInd**' was selected to represent the other attributes. The columns '**country, state, city, postal code, market and region**' were then deleted from the **Orders** table.

The screenshot shows the Power Query Editor interface with two queries loaded: SalesFact and LocationDim. The SalesFact query is the current focus, displaying a table with columns: Sales, Quantity, Discount, Profit, Shipping Cost, Order Priority, and LocationDim. The LocationDim column is highlighted with a yellow box. A context menu is open over the LocationDim column, with the 'Expand' option selected. The 'LocationInd' checkbox is checked, indicating that the primary key of the LocationDim table will be used to represent its attributes. The 'OK' button at the bottom of the dialog is highlighted with a red box.

We then remove the columns that has been merged from the fact table

The screenshot shows the Power Query Editor interface with the SalesFact query selected. A context menu is open over the LocationDim column, with the 'Remove Columns' option highlighted with a red box. Other options visible in the menu include Copy, Remove Other Columns, Add Column From Examples..., Replace Values..., Replace Type..., Merge Columns, Group By..., Unpivot Columns, Unpivot Other Columns, Unpivot Only Selected Columns, Move, and Furniture. The 'APPLIED STEPS' pane on the right shows the 'Expanded LocationDim' step.

Fig: Selecting the merge queries option to merge LocationDim to superstore

The screenshot shows the Microsoft Power Query Editor interface. The ribbon at the top has tabs like File, Home, Insert, Design, Layout, References, Mailings, Review, View, Help, RCM, and Acrobat. The 'File' tab is selected. In the center, there's a table with columns: Sales, Quantity, Discount, Profit, Shipping Cost, Order Priority, and LocationID. On the right, there's a 'Query Settings' pane titled 'APPLIED STEPS' which lists steps like 'Source', 'Promoted Headers', 'Changed Type', 'Filtered Rows', 'Merged Queries', and 'Expanded LocationDim'. At the bottom left, it says '25 COLUMNS, 999+ ROWS' and 'Column profiling based on top 1000 rows'. At the bottom right, it says 'PREVIEW DOWNLOADED AT 15:31'.

We then close and apply

Similar approaches were performed to generate the other dimension tables: **DateDim**, **CustomerDim** and **ProductDim**. As a result, a star schema was created by normalising the flat table into several dimension tables and one fact table.

For the **OrderDim**, make use of **M Language** to remove the columns that have been merged through indexing to the fact table, we go click on '**Advanced Editor**' in the home ribbon

The screenshot shows the Microsoft Power Query Editor interface with the 'File' tab selected. The ribbon has various buttons for file operations like Close, New Source, Refresh Preview, and Advanced Editor. The 'Advanced Editor' button is highlighted with a red box. The main area shows a table with columns: Sales, Quantity, Discount, Profit, Shipping Cost, Order Priority, and LocationID. The 'Query Settings' pane on the right is visible.

The columns Order ID, Order Date, Ship Date, Ship Mode and Order Priority were removed from the columns through the M language below

Advanced Editor

SalesFact

Display Options ?

```
let
    Source = Csv.Document(File.Contents("C:\Users\b1202001\OneDrive - Teesside University\Big Data and Business Intelligence\ICA Datasets\Global Sales [1].csv")),
    #Promoted Headers = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #Changed Type = Table.TransformColumnTypes(#"Promoted Headers",{{"Row ID", Int64.Type}, {"Order ID", type text}, {"Order Date", type date}, {"Postal Code", type text}, {"Country", type text}, {"State", type text}, {"City", type text}, {"Region", type text}, {"Market", type text}, {"Segment", type text}, {"Category", type text}, {"Sub-Category", type text}, {"Product Name", type text}, {"Product ID", type text}, {"Product Ind", type text}, {"Customer ID", type text}, {"Customer Name", type text}, {"Order Ind", type text}, {"Order Date", type date}, {"Ship Date", type date}, {"Ship Mode", type text}, {"Order Priority", type text}}),
    #Filled Up = Table.FillUp(#"Changed Type", {"Postal Code"}),
    #Merged Queries = Table.NestedJoin(#"Filled Up", {"City", "State", "Country", "Postal Code", "Market", "Region"}, LocationDim, {"City", "State", "Country", "Postal Code", "Market", "Region"}, {"LocationInd"}, {"LocationInd"}),
    #Expanded LocationDim = Table.ExpandTableColumn(#"Merged Queries", "LocationDim", {"LocationInd"}, {"LocationInd"}),
    #Removed Columns = Table.RemoveColumns(#"Expanded LocationDim", {"City", "State", "Country", "Postal Code", "Market", "Region"}),
    #Merged Queries1 = Table.NestedJoin(#"Removed Columns", {"Segment", "Product ID", "Category", "Sub-Category", "Product Name"}, ProductDim, {"Segment", "Product ID", "Category", "Sub-Category", "Product Name"}, {"ProductInd"}, {"ProductInd}),
    #Expanded ProductDim = Table.ExpandTableColumn(#"Merged Queries1", "ProductDim", {"Product Ind", "Product Name", "Category", "Sub-Category", "Product Name"}),
    #Removed Columns1 = Table.RemoveColumns(#"Expanded ProductDim", {"Segment", "Product ID", "Category", "Sub-Category", "Product Name"}),
    #Merged Queries2 = Table.NestedJoin(#"Removed Columns1", {"Customer ID", "Customer Name"}, CustomerDim, {"Customer ID", "Customer Name"}, {"CustomerInd"}, {"CustomerInd"}),
    #Expanded CustomerDim = Table.ExpandTableColumn(#"Merged Queries2", "CustomerDim", {"Customer Ind", "Customer Name"}, {"CustomerInd"}, {"CustomerInd"}),
    #Removed Columns2 = Table.RemoveColumns(#"Expanded CustomerDim", {"Customer ID", "Customer Name"}),
    #Merged Queries3 = Table.NestedJoin(#"Removed Columns2", {"Order ID", "Order Date", "Ship Date", "Ship Mode", "Order Priority"}, OrderDim, {"Order Ind", "Order Date", "Ship Date", "Ship Mode", "Order Priority"}),
    #Expanded OrderDim = Table.ExpandTableColumn(#"Merged Queries3", "OrderDim", {"Order Ind", "Order Date", "Ship Date", "Ship Mode", "Order Priority"}),
    #Removed Columns3 = Table.RemoveColumns(#"Expanded OrderDim", {"Order ID", "Order Date", "Ship Date", "Ship Mode", "Order Priority"})
in
    #Removed Columns3"
```

Fig: M language to remove columns

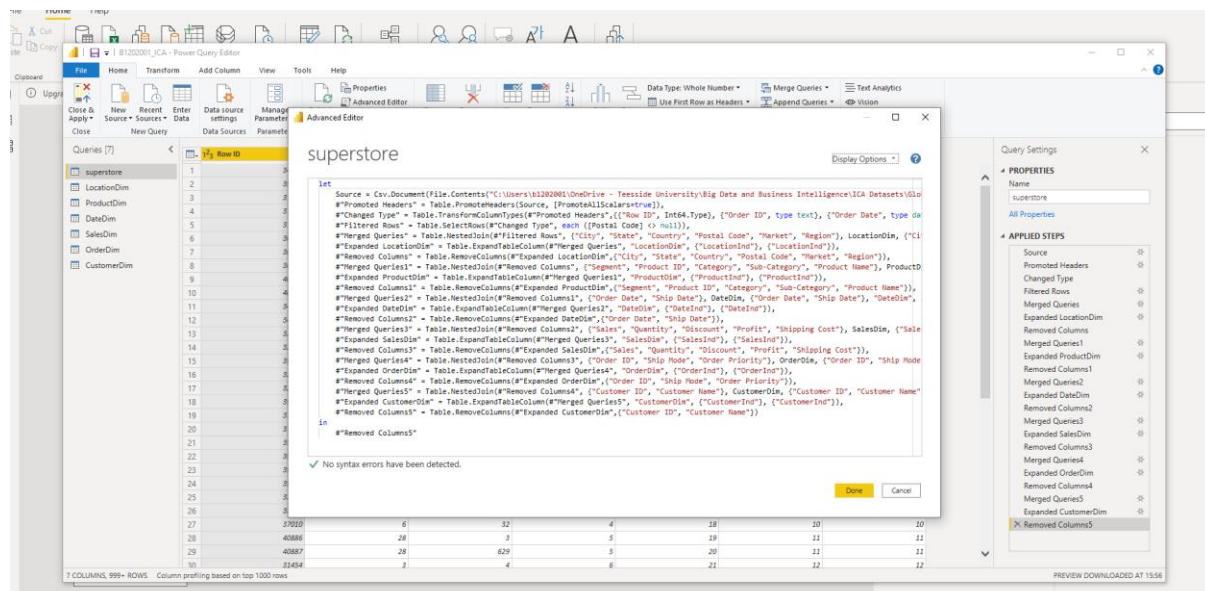


Fig: M language showing how all the dimension tables were generated

2.3.2 Further Pre-processing

In the DateDim table, we computed three new columns called ‘Day’, ‘Month’ and ‘Year’ by clicking on the new column button and writing a DAX formula

to show the corresponding days, months and years of the order date as depicted below.

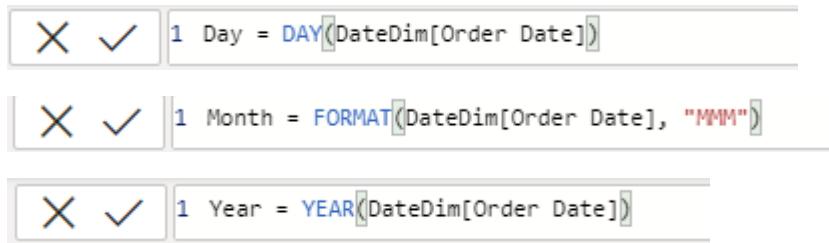


Fig: DAX expression to extract day, month and year of the order

Additionally, the order date and ship date in the date dimension table was used to generate the number of days the delivery took using the **custom column option**. This is illustrated in the figure below

The screenshot shows the Power Query Editor interface. On the left, the 'Queries' pane lists several tables: Orders Date, 01 February, 02 November, 01 June, 01 Month, 11 Mo, LocationDim, ProductDim, SalesDim, OrderDim, CustomerDim, and DateDim. The 'DateDim' table is currently selected. In the main area, a 'Custom Column' dialog box is open. It contains fields for 'New column name' (set to 'Delivery Days') and 'Formula' (set to '(Ship Date) - [Order Date]'). A preview table shows the calculated values for the first few rows. The 'Available columns' list includes DateId, Order Date, and Ship Date. The 'APPLIED STEPS' pane on the right shows the history of operations, including 'Reordered Columns'. The status bar at the bottom right indicates 'PREVIEW DOWNLOADED AT 16:09'.

Fig: Creating a new column using power query formulas

After generating the ‘Delivery Days’ column, we then change the type to whole numbers by clicking the little type box.

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table titled "DateDim" with columns: DateKey, Date, Year, Quarter, Month, WeekNumber, Day, WeekDay, IsWeekend, IsHoliday, DayName, MonthName, and QuarterName. The "Date" column contains dates from 2011-01-01 to 2014-09-01. The "Year" column shows years 2011 through 2014. The "Month" column shows months from January to September. The "Day" column shows days from 1 to 30. The "WeekDay" column shows days of the week. The "IsWeekend" and "IsHoliday" columns are binary values. The "DayName", "MonthName", and "QuarterName" columns provide textual representations of the date components.

The ribbon at the top has tabs for File, Home, Insert, Design, Layout, References, Mailings, Review, View, Help, RCM, and Acrobat. The "File" tab is selected. The "Home" tab has options like Cut, Copy, Paste, and Format. The "Transform" tab has options like Add Column, View, Tools, and Help. The "Format" tab includes sections for General, Conditional Column, Merge Columns, Extract, Statistics, Standard, Scientific, Rounding, Information, Date, Time, Duration, Text Analytics, Vision, Azure Machine Learning, and AI Insights.

The "Properties" pane on the right shows the query name as "DateDim". The "Applied Steps" pane lists various steps taken during the transformation process, such as Promoted Headers, Changed Type, Filtered Rows, Merged Queries, Expanded LocationDim, Removed Columns, Merged Query1, Expanded ProductDim, Removed Other Columns, Removed Duplicates, Added Index, Renamed Columns, Reordered Columns, and Added Custom.

The final **DateDim** table:

Table: DateDim (3,473 rows) Column: Year (4 distinct values)

Furthermore, three new measures were created, Total Sales, Total Profit, Profit Ratio and Sales by City using the **DAX expressions**.

'Total Sales' was generated using the function SUM on Sales table

```
Total Sales = SUM([SalesFact[Sales]])
```

'Total Profit' was generated using the function SUM on Profit table

```
Total Profit = SUM([SalesFact[Profit]])
```

'Profit Ratio' was generated using the DIVIDE function on Total Profit by Total Sales

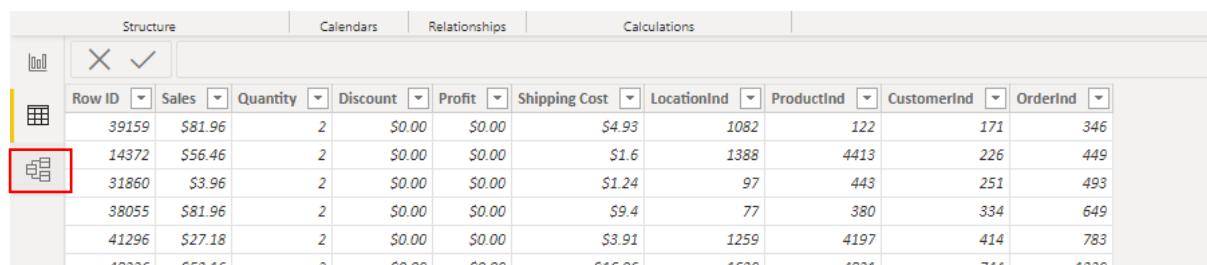
```
Profit Ratio = DIVIDE([Total Profit],[Total Sales],0)
```

'Sales by City' was generated with the top 5 cities using City in the LocationDim table and Total Sales.

```
1 Sales by City =
2 CALCULATE([Total Sales],
3 TOPN(5,LocationDim,LocationDim[City],ASC))
```

3 Data Modelling

Power BI is capable of detecting relationships between the models automatically as long as we have created the fact and dimension tables relationally. For the sake of this video, we would create the relationship between the fact and dimension tables manually.



Row ID	Sales	Quantity	Discount	Profit	Shipping Cost	LocationInd	ProductInd	CustomerInd	OrderInd
39159	\$81.96	2	\$0.00	\$0.00	\$4.93	1082	122	171	346
14372	\$56.46	2	\$0.00	\$0.00	\$1.6	1388	4413	226	449
31860	\$3.96	2	\$0.00	\$0.00	\$1.24	97	443	251	493
38055	\$81.96	2	\$0.00	\$0.00	\$9.4	77	380	334	649
41296	\$27.18	2	\$0.00	\$0.00	\$3.91	1259	4197	414	783
48226	\$53.16	2	\$0.00	\$0.00	\$15.06	1630	4821	744	1270

Fig: Screenshot of how to assess the model

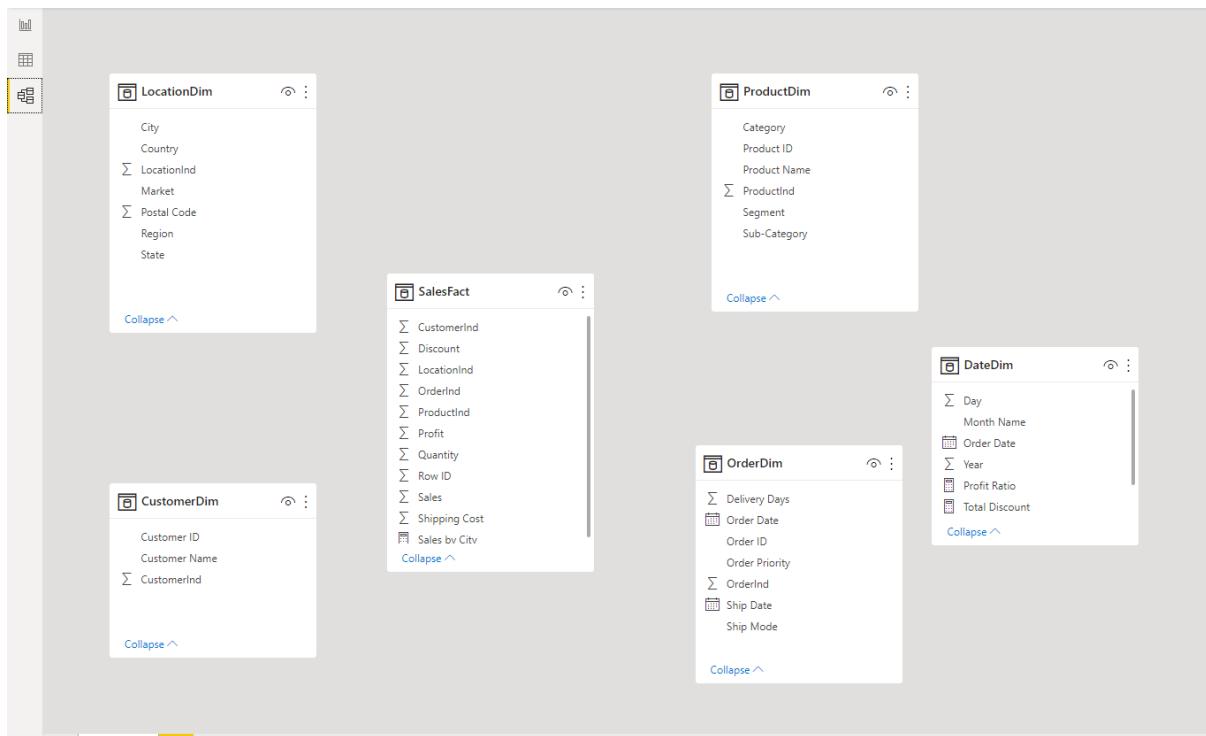
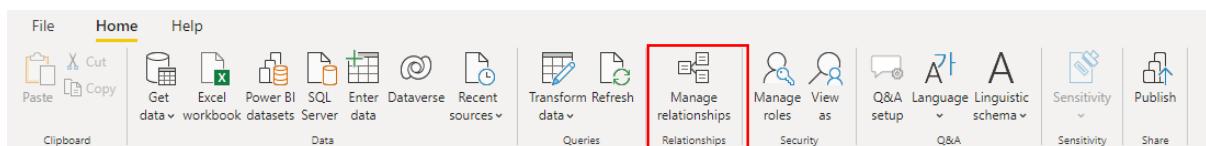
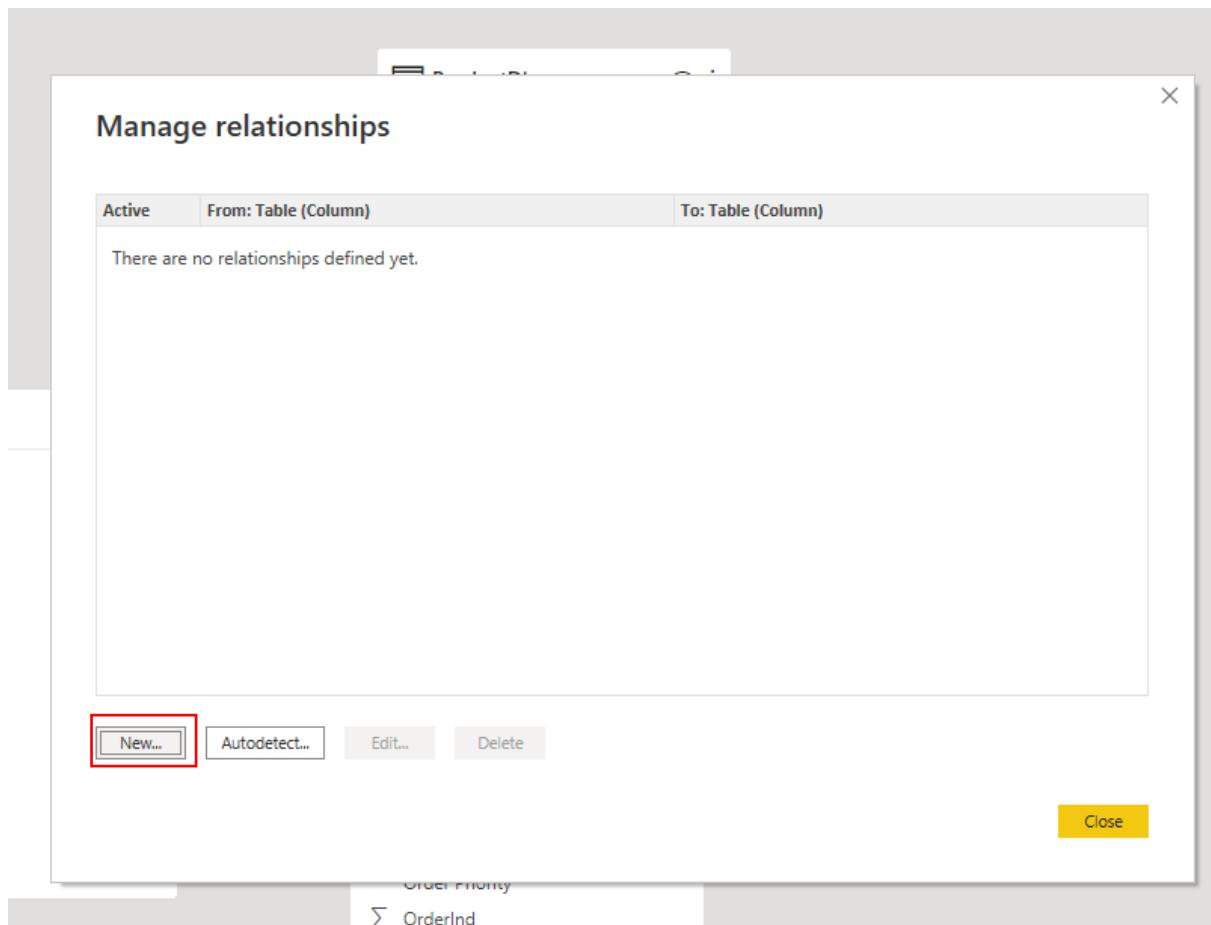


Fig: Screenshot of what the model looks like without relationships

There are three ways by which we can create relationships between these tables, we can either '**drag and drop**' the relational columns or '**manage relationship**' in the ribbon by auto-detecting or creating new. For this sake of this project, we would go with the latter method.





Create relationship

Select tables and columns that are related.

SalesFact

Row ID	Sales	Quantity	Discount	Profit	Shipping Cost	LocationInd	ProductInd	CustomerInd
39159	\$81.96	2	\$0.00	\$0.00	\$4.93	1082	122	171
14372	\$56.46	2	\$0.00	\$0.00	\$1.6	1388	4413	226
31860	\$3.96	2	\$0.00	\$0.00	\$1.24	97	443	251

LocationDim

City	State	Country	Postal Code	Market	Region	LocationInd
Paris	Ile-de-France	France	53711	EU	Central	198
Paris	Ile-de-France	France	43229	EU	Central	217
Paris	Ile-de-France	France	20735	EU	Central	385

Cardinality

Many to one (*:1)

Cross filter direction

Single

Make this relationship active

Apply security filter in both directions

Assume referential integrity

OK

Cancel

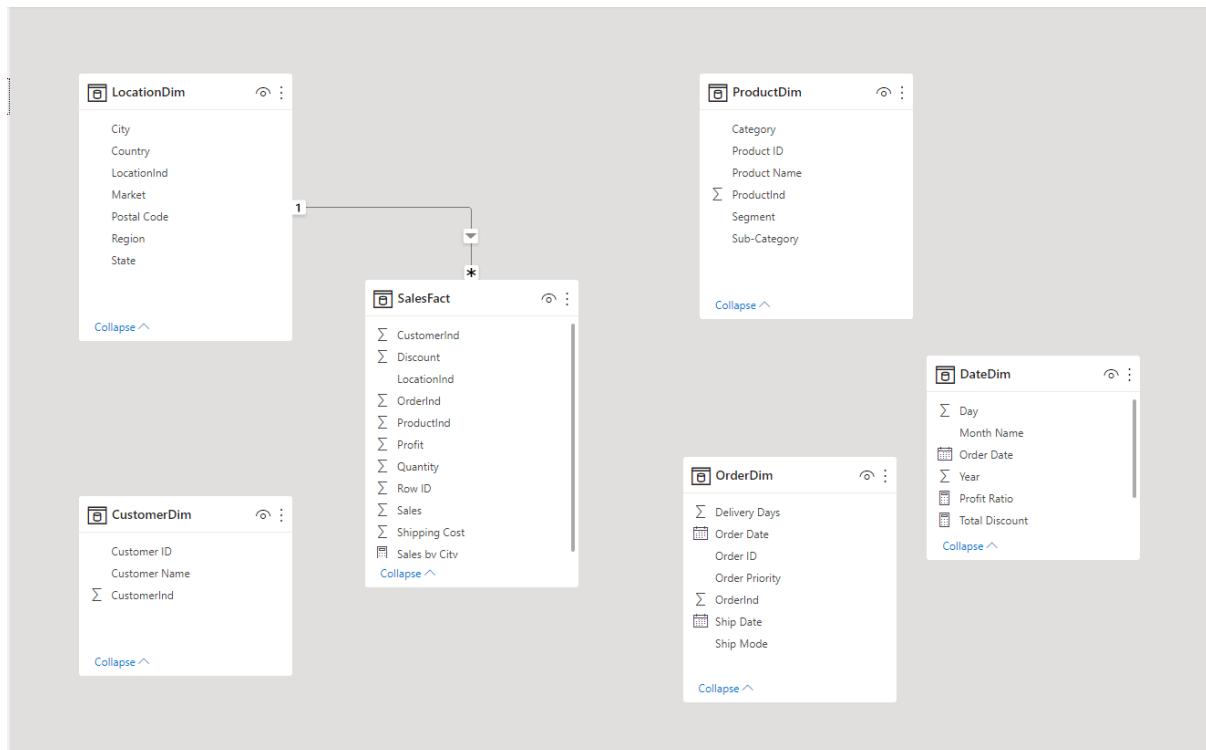
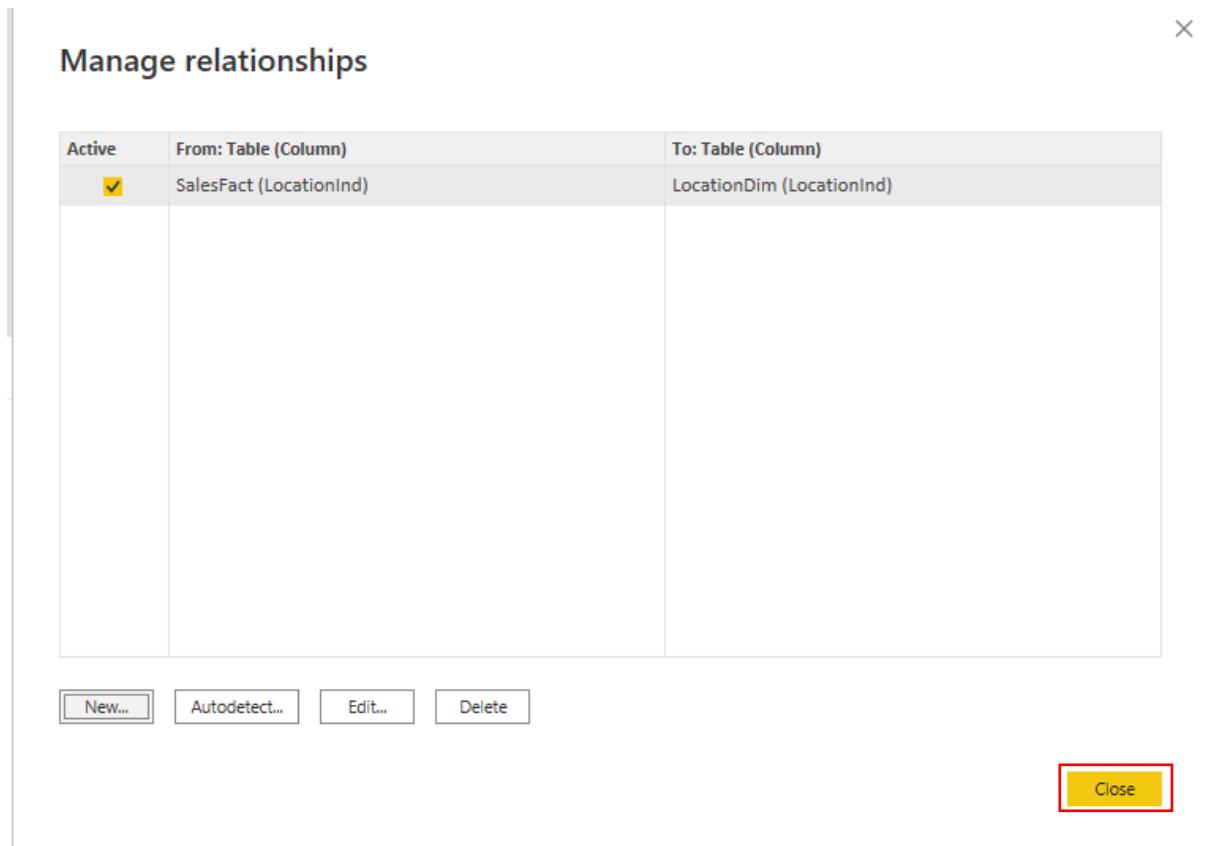


Fig: the first relationship is formed between **SalesFact** and **LocationDim**

We do through these methods to create a relationship between SalesFact and OrderDim, CustomerDim, OrderDim and ProductDim, then OrderDim and DateDim. Below is the data model after creating the relationships.

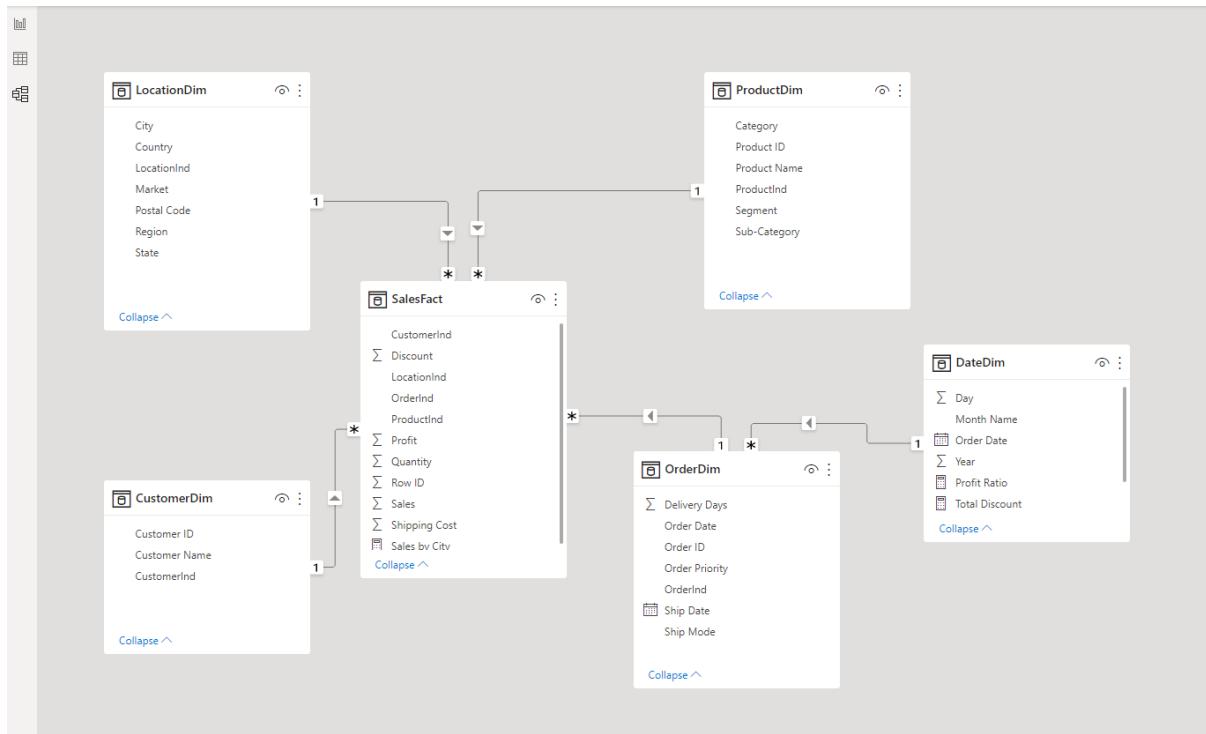


Fig: Data Modelling showing the fact and dimension tables

In the diagram above, the data modelling is a **Snowflake Schema**. The snowflake schema has one or more fact tables linked to numerous dimension tables; however, the dimension table has additional dimensions too. The SalesFact table is linked to other dimension tables in this model, including CustomerDim, OrderDim, ProductDim, and LocationDim. The OrderDim is then linked to the Date dimension, which refers to the Order Date. This model includes one fact table, four dimension tables, and a fifth dimension table linked to one of the preceding dimension tables.

SECTION 2

BUSINESS REPORT

SUPERSTORE SALES ANALYSIS

NAME: AYOOLUWA DORCAS BABALOLA

STUDENT ID: B1202001

1 EXECUTIVE SUMMARY

Superstores are a multi-faceted enterprise that provides a wide range of items in enormous quantities. A superstore is a collection of large stores that sell food, pharmaceuticals, cosmetics, health products, games and toys, furniture, appliances, and a variety of other items. Because of the nature of the industry, most of these products are available in big and bulk quantities. In addition, superstores have locations in other countries and places. The dataset used in this study includes information on client orders from 147 countries and 7 markets. To find insights in superstore sales, many analytical and visualisation approaches are applied.

This report gives insight on the following business questions:

- The trend of sales over the years and the forecast of sales in the next two years
- The market in which the most and least profit was made
- Reasons for increase and decrease in sales by country
- Category that generated the most revenue
- The time of the year that has the most sale
- Cities in which the most sales were made
- Top customers by sales and profit

Findings and Conclusions - A Power BI dashboard was created to critically analyse the report's questions.

- As the year progresses, sales and profit rise.
- The average number of days it takes to deliver the products is four.
- Technological products are the best-selling products over the years.
- There are usually more sales in the last quarter of the year.
- The more discounts each product group received, the more profit they made. In retrospect, the consumer segment products received the highest discount of \$3.8k and, as a result, the highest profit of \$750k.
- Customers who contributed the most to superstore sales are not the reason for superstore profit because they are the ones who order the most discounted items.

Recommendations for Improving Performance

- Reducing the number of delivery days to two days would increase customer satisfaction and possibly the sales too.
- To increase profitability, EMEA, African, and Canadian markets should be targeted.
- Customers should be permitted to rate products because this may assist the superstore choose which products to focus on more and which products are the best sellers.
- Customers would have better shopping experiences and sales if they are segmented based on their profile and references.
- Some of the clients who contribute the most to profit aren't regulars. To improve profits, several tactics such as introducing membership cards and promotions can be used to keep customers.

2 INTRODUCTION

A major industry that relies on cutting-edge technology and data analytics to preserve its competitive advantage is the superstore. This initiative intends to boost the superstore's sales and profitability across all products, categories, and countries. This report also examines sales performance from 2011 to 2014 in order to assess effectiveness and identify areas for improvement. The report was created using the dataset specified in Section 1, and the results of the data model are shown below.

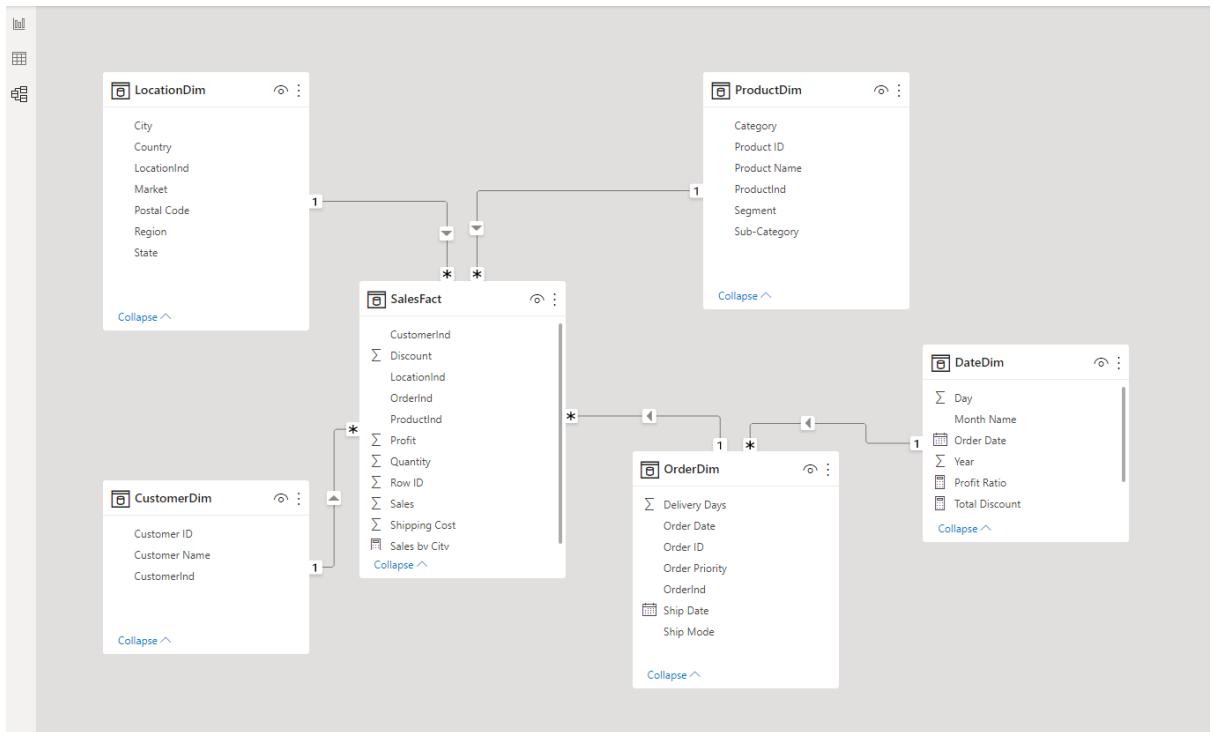


Fig: Data modelling

This model includes six tables named SalesFact, OrderDim, ProductDim, LocationDim, CustomerDim and Date. The data in these tables give detailed information about the orders, sales, profit, products, customers, order and ship dates of the superstore in four years.

3 NEW VISUALIZATION USED IN THIS PROJECT:

According to the project's specifications, a new visualisation tool was imported from the Power BI Visuals package. The '**Scroller**' visual tool aids in the visualisation of data as an animation scrolling text. Its pace, size, status, indicator, and colouring, among other things, can all be adjusted. It can also be used to scroll through personalised messages, notification changes, and receive updates. It was utilised in this report to illustrate the sales of each individual sub-category. Below is an illustration of how it was imported: First, we go to the visualization pane and click on the 3 dots '...' to see more options, we then select '**Get more visuals**'

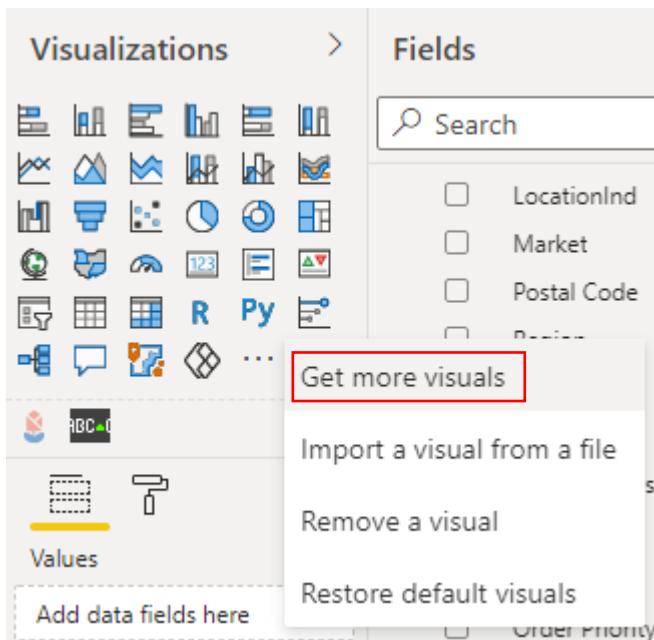


Fig: How to get more visuals

This then takes us to Power BI Visuals page which contains many different visualisation tools, for the sake of this project, we search for ‘Scroller’ and click ‘Add’ to add it to the visualization pane automatically.

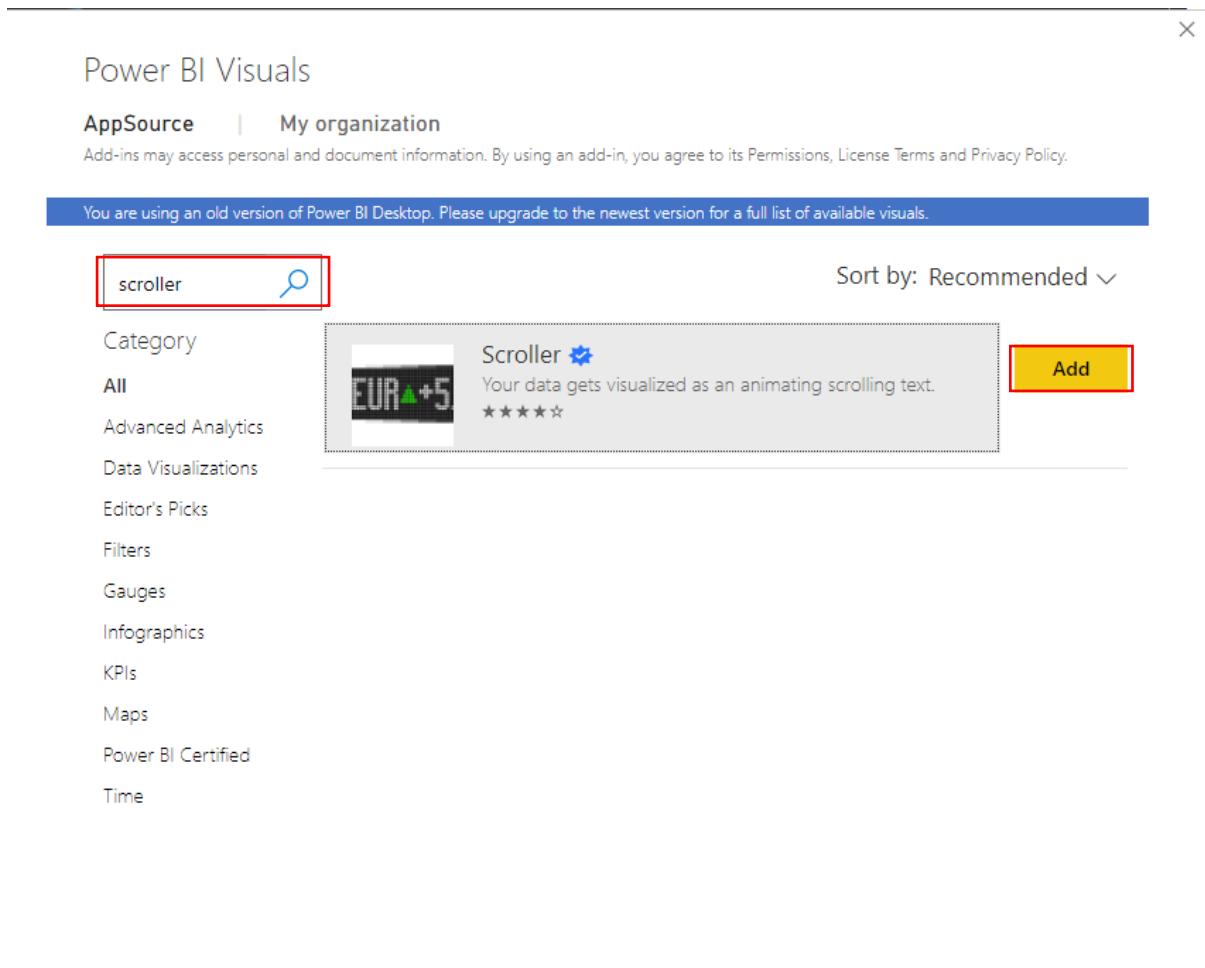
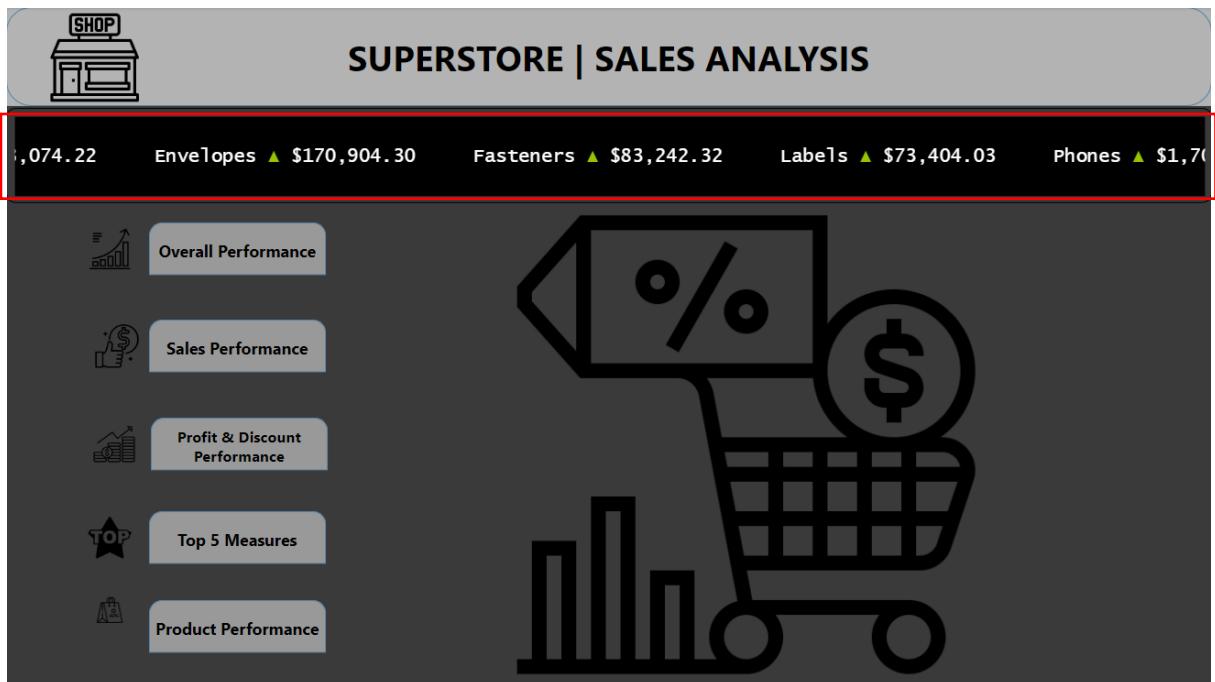


Fig: Importing Scroller as our new visualisation tool

In this project, we can find the scroller being used on the 'Home Page' dashboard.



4 KEY FINDINGS

- i. One important crucial finding we can observe on the dashboard over the years is the increase in sales as the year progresses. The overall sales increased by more than 50%. Plotting a clustered column chart and using the trend line in the analytics tab were used to examine the pattern of sales by year. **Total sales** (set to sum) were used as a measure, and **Year** was used as an axis. The superstore began with a total sale of \$2.3 million in 2011, and by the end of 2014, it had increased to \$4.3 million. Providing a broader selection of items, establishing new domestic markets, improving customer service, participating in marketing efforts, and many other factors will all contribute to increased sales over time. In the diagram below, the trend line provides extra information. A simple DAX formula was used to calculate the total sales of the given data.

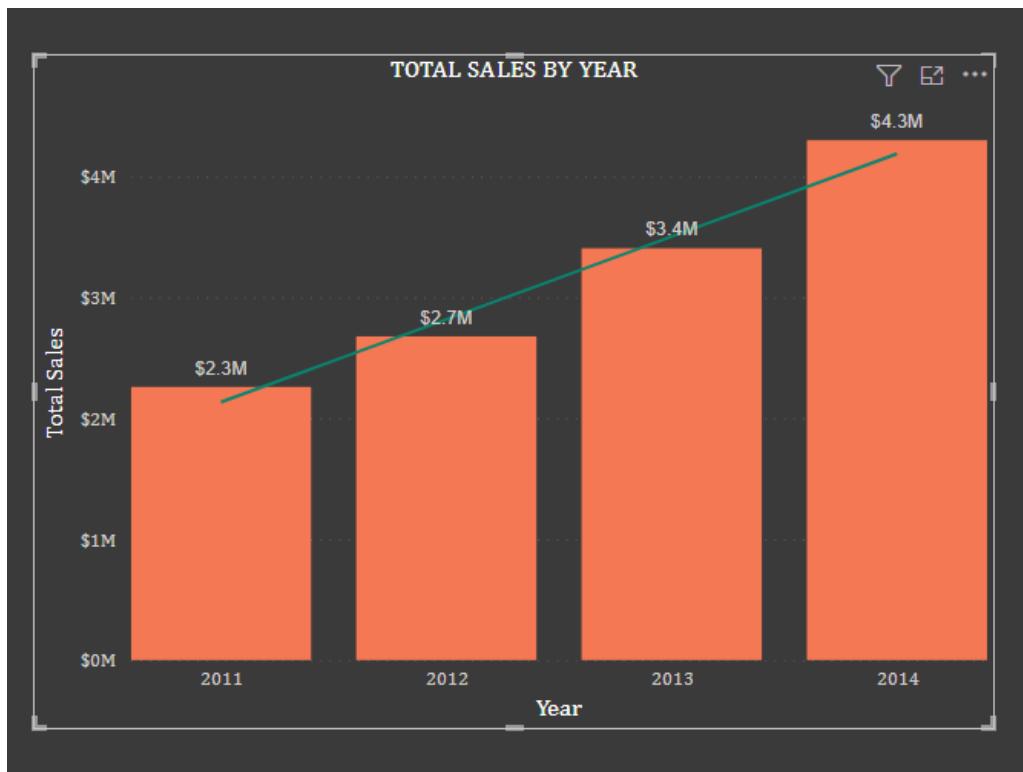


Fig: showing the trend line of sales over the years

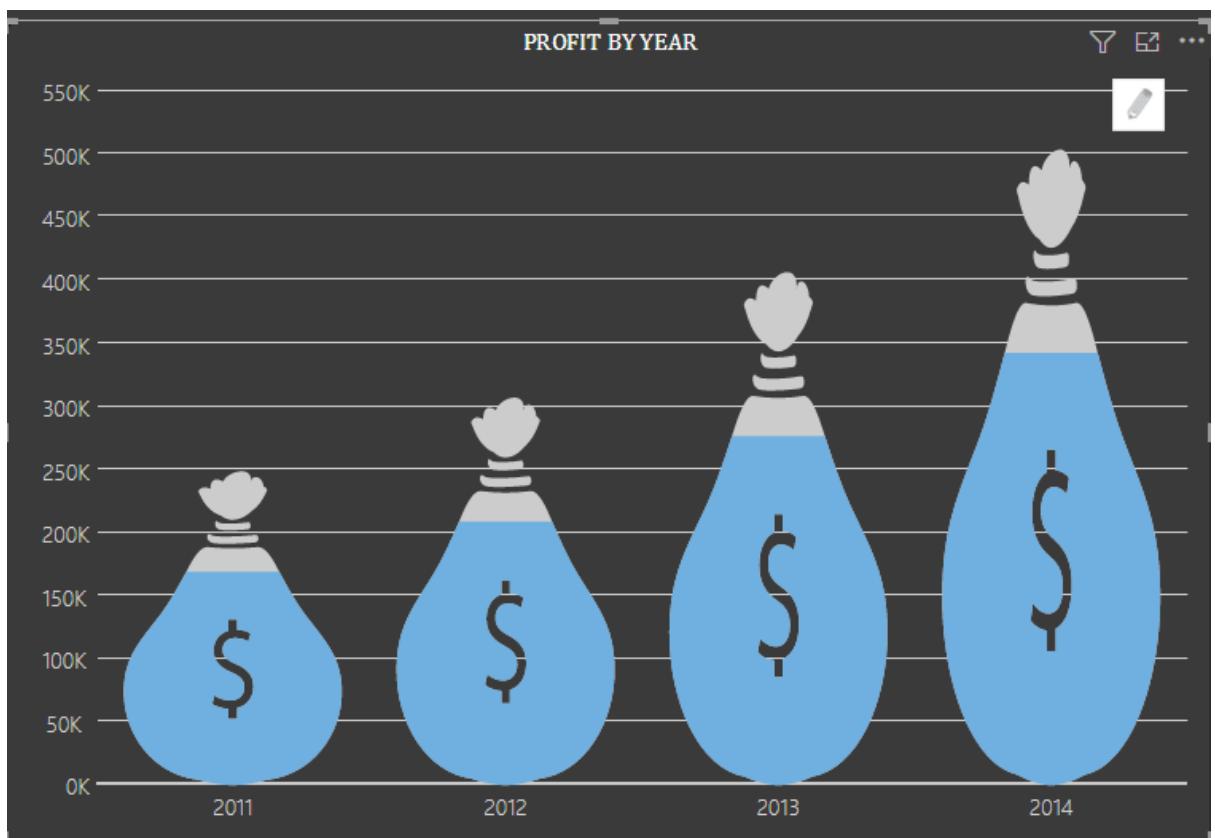


Fig: The profit also progresses with the year

- ii. Due to a strong trend in total sales over the years, a prognosis was created for the next two years, with total sales increasing to roughly \$10 million by the end of 2016, which is beneficial for the superstore. This is illustrated in the figure below where the **Sales** was plotted against the **Order Date**.

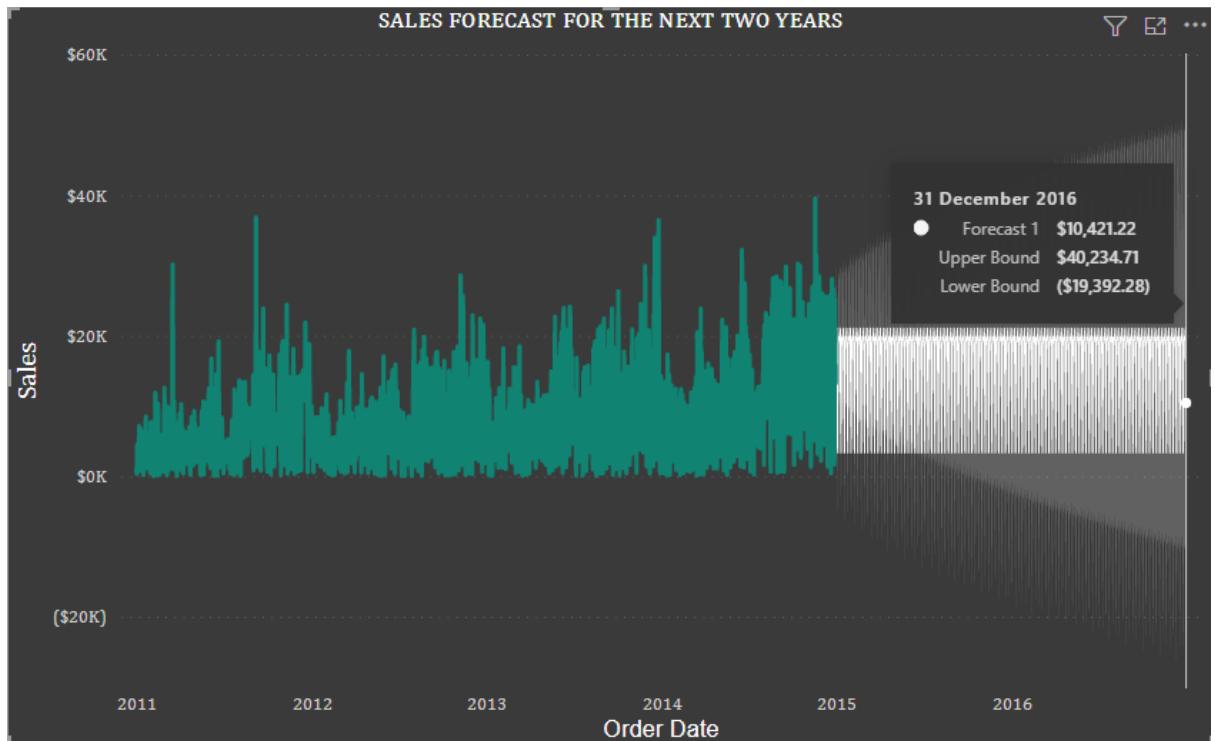


Fig: Showing a forecast of sales by 2016

- iii. More profit was made in the APAC market (Asian Pacific Market) while less profit was made in the Canadian market. While concentrating on the market that brings on the most profit, measures should be taken to expand the profit in Canada. As illustrated in the picture below, a waterfall chart was utilised to have a deeper look at the profit by market. **Profit** was used in the SalesFact table, and **Market** was used in the LocationDim table to construct the chart. The order of markets that made more profit too are the EU, US, LATAM (Latin America), Africa and EMEA (Europe, the Middle East and Africa) respectively.



Fig: Profit by Market

- iv. The total sales by month varied all round the year however the lowest sales were made between August and September and the highest sale was made between October and November. A line chart was used to analyse the sales through the eyes of the month and quarter, as seen in the image below. The month served as the axis, and the total sales served as the values. Further down, the significant drop and spike were analysed in relation to the countries.

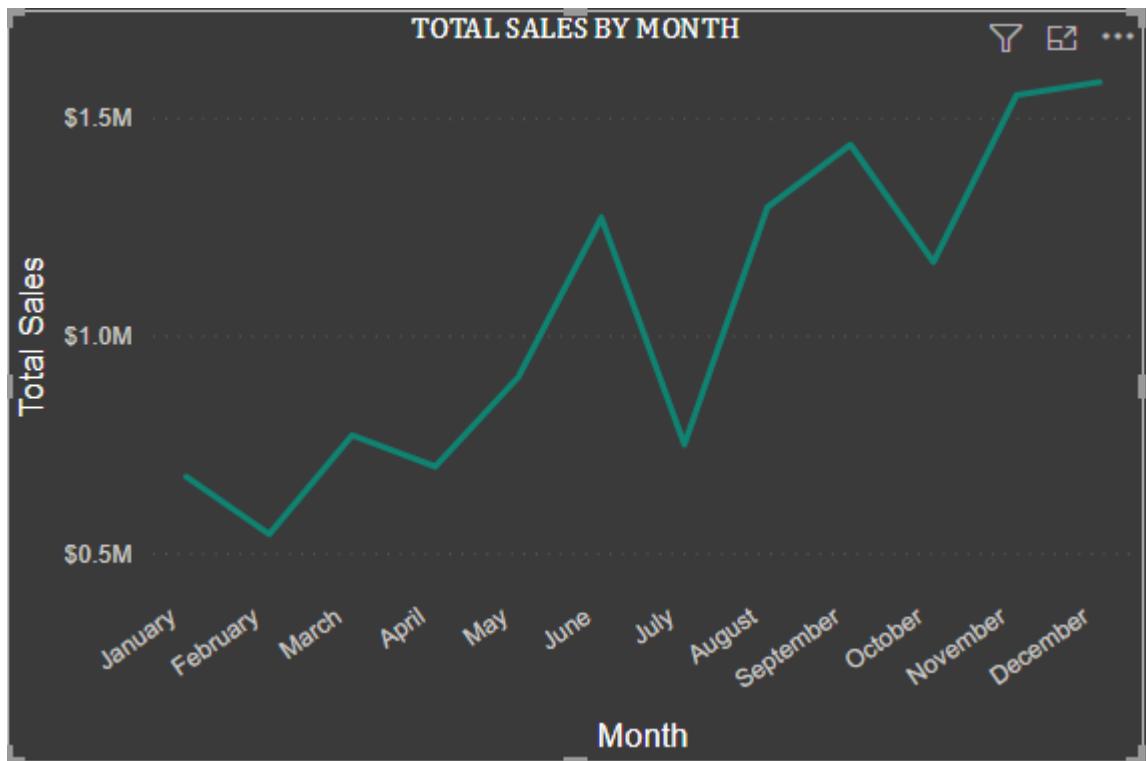


Fig: Total sales by month

Using waterfall charts depending on countries, the explanation for the growth and decline at these times was studied further. Between August and September, the United States, Mexico, and Nicaragua saw the biggest growth among the countries, offsetting China's decline. Furthermore, the United States', China's, and France's relative contributions changed the greatest. This is depicted in the graph below.

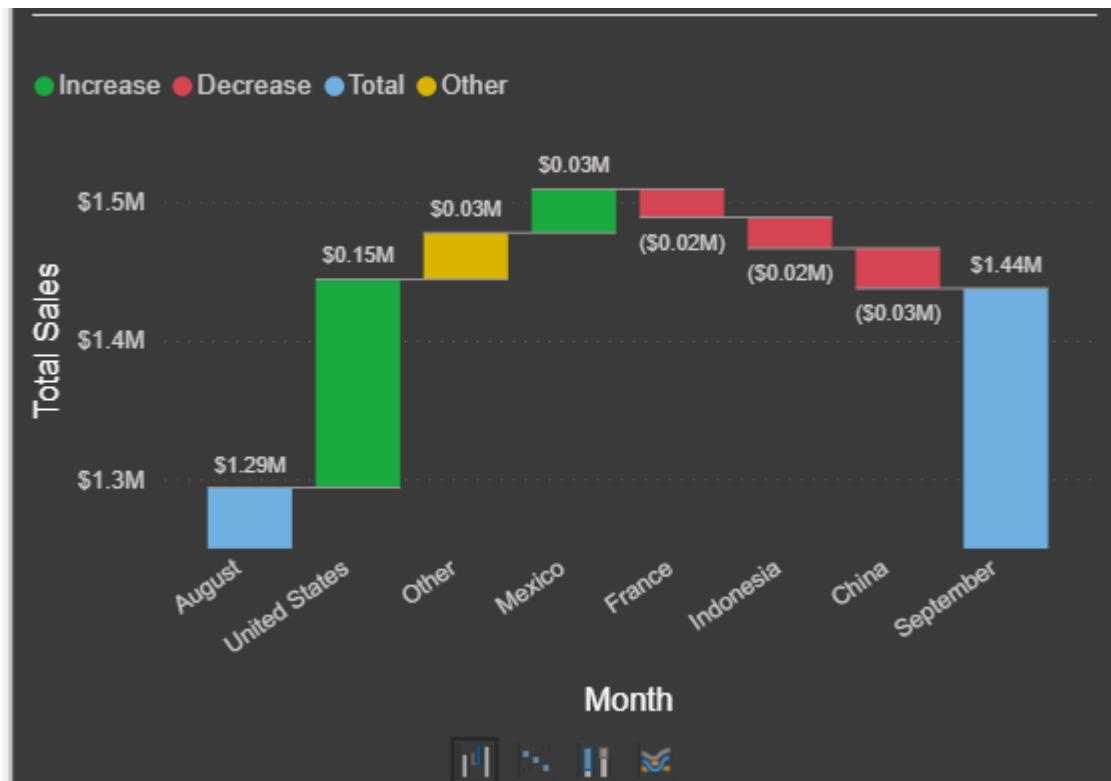


Fig: Increase in sales among countries

Also, Australia, France and United Kingdom had the largest decrease among the countries. The relative contribution made by the United States, Australia and United Kingdom changed the most. This is illustrated in the chart below.

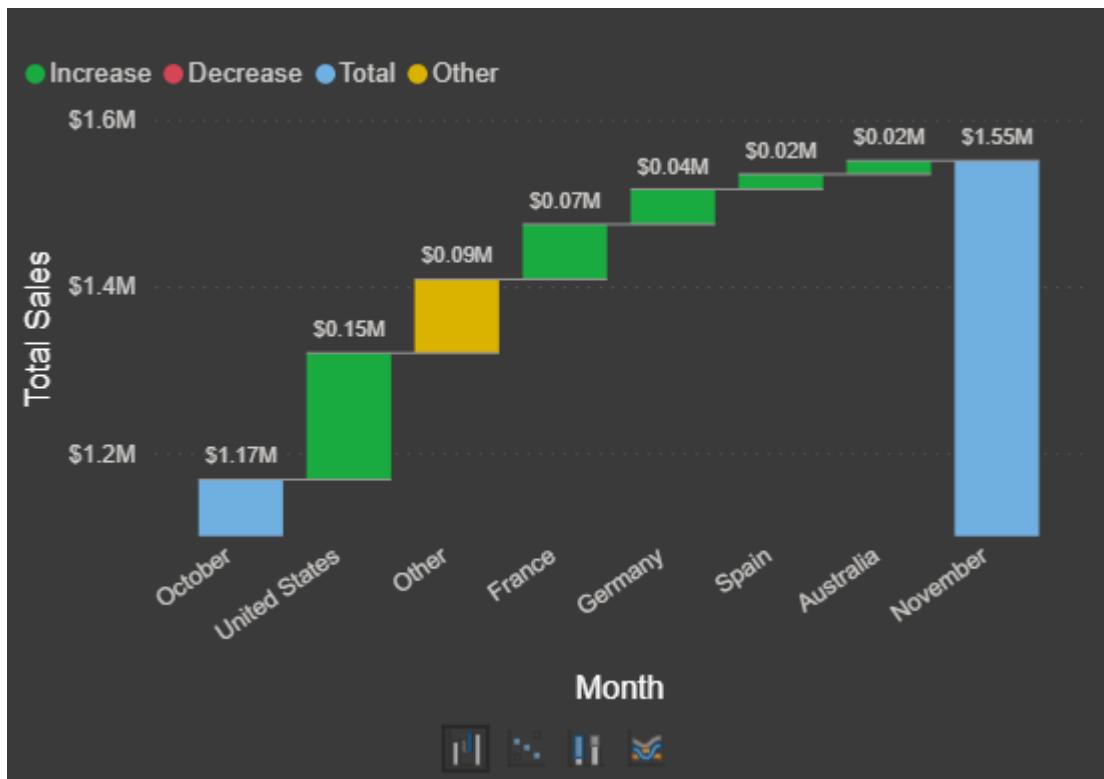


Fig: Decrease in total sales by country

- v. The technology category had the highest sales and, as a result, the most profit. Even though office supplies had the lowest sales, they made more money than furniture. As a result, in order to enhance revenues, the store should focus more on furnishings too. A clustered bar chart was used to analyse and visualize this chart with the category as axis and profit and sales as values.

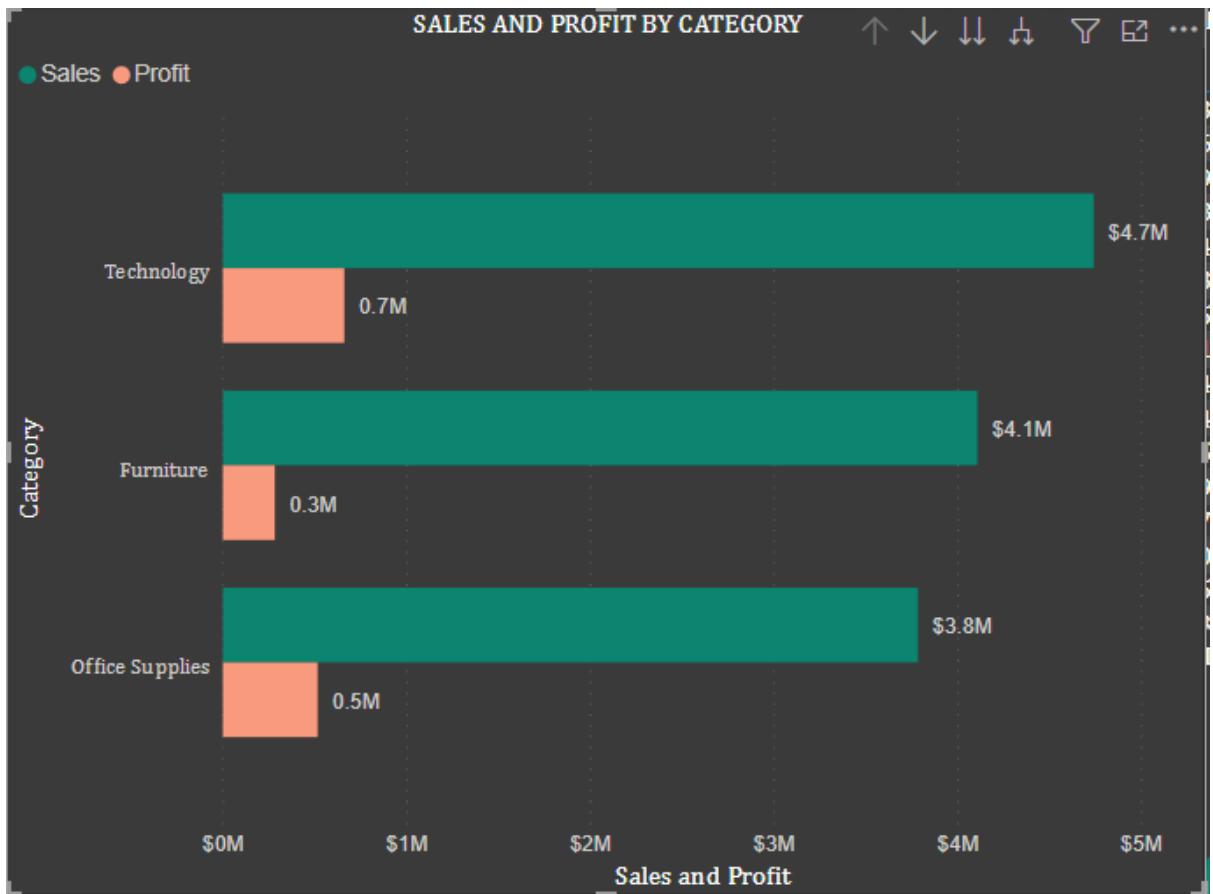
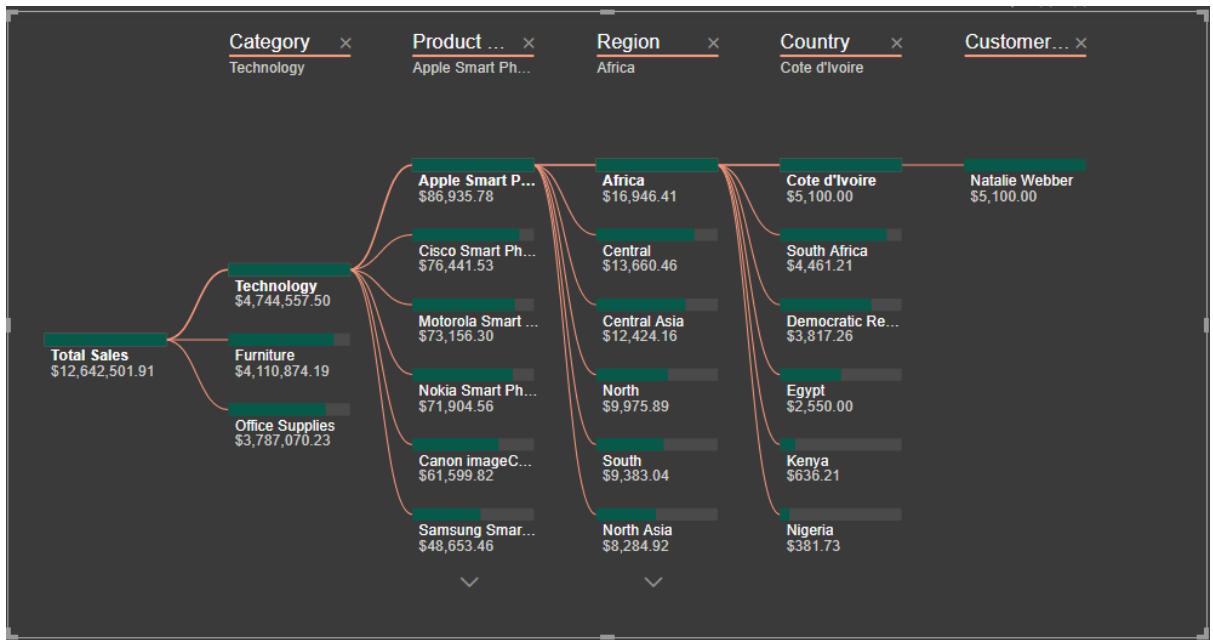


Fig: Sales and Profit by Category

Furthermore, a decomposition tree has been used to buttress the total sales and it was expanded by the category, product, region, country and customer with Technology being the category with the most sales, an Apple smart phone as the highest selling product in Cote d'Ivoire in Africa and Natalie Webber as the customer.



- vi. It can be seen from the figure below that profit is higher in the fourth quarter of the year. This could be due to a variety of factors, but further investigation reveals that accessories, appliances, and paper generate more earnings.

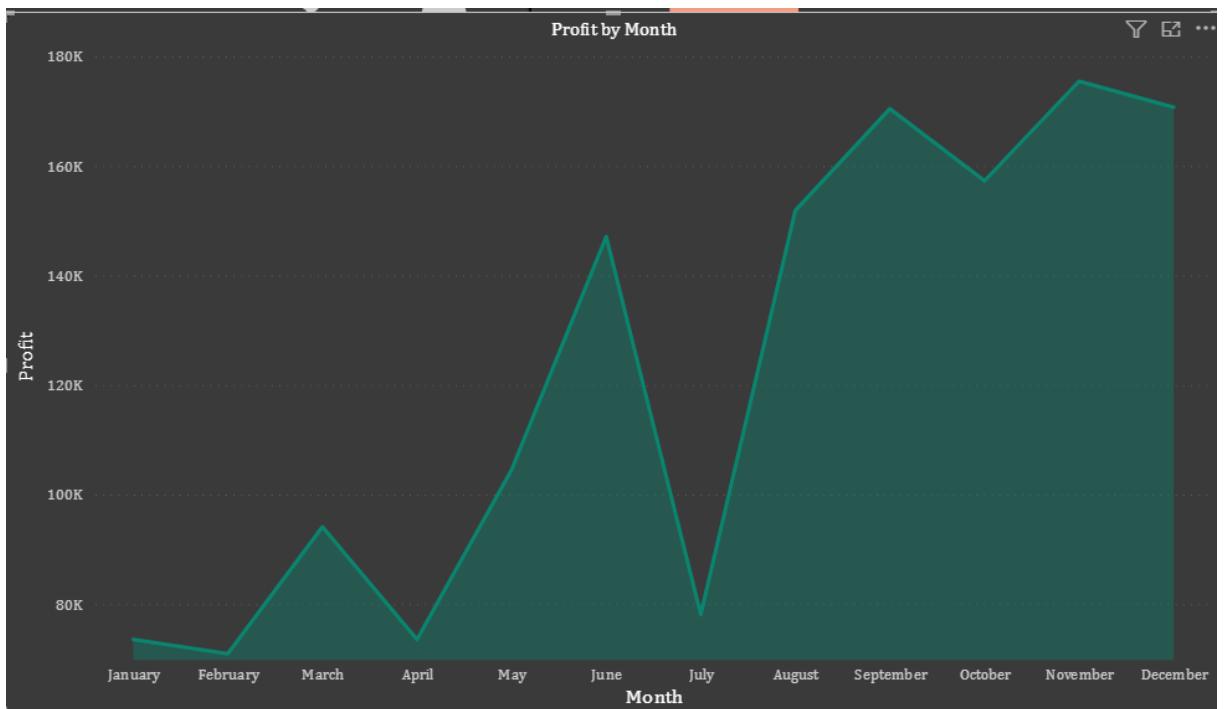


Fig: Trend of profit by month

- vii. While analysing the top sales by cities, it was discovered that the top 5 cities with the most sales are all in the United States, or more precisely,

four of the top five cities with the highest sales are all in the United States, with Manilla being the exception. To increase revenue, the superstore must ensure that the standard is maintained while also aiming to expand its locations in other cities.

We first create a measure using the DAX formula called Sales by city by generating the Top 5 cities in ascending order after calculating the Total Sales.

```
1 Sales by City =  
2 CALCULATE([Total Sales],  
3 TOPN(5, LocationDim, LocationDim[City], ASC))
```

Moving forward, a map was used in the visualization pane by using the city as the location and legend and the Sales by City measure in sizes. This is illustrated in the figure below.

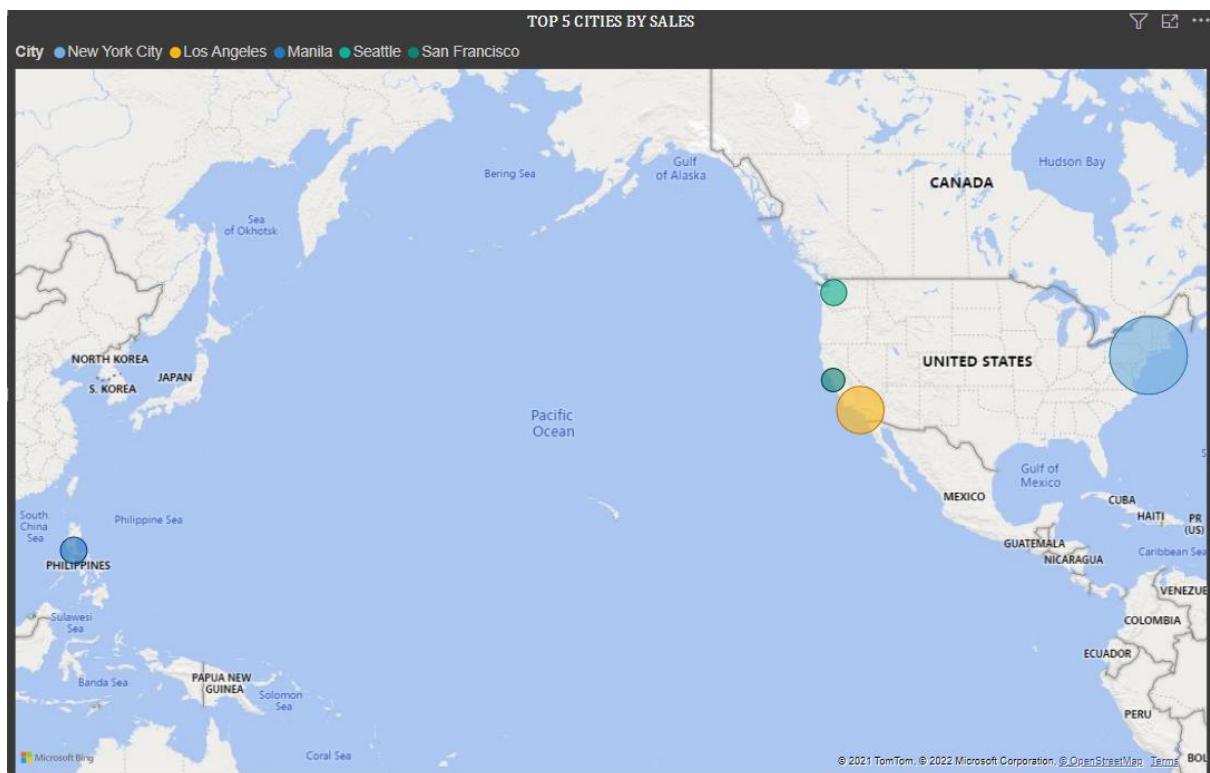


Fig: Top 5 cities with the most sales

- viii. Using a clustered column chart to visualize the number of delivery days by the order priority, we see that the low priority order took the most days to deliver, and in that light, critical orders were delivered faster which took

about 2 days. We also use a ‘**Trend line**’ in the analytics tab to find the average number of delivery days.

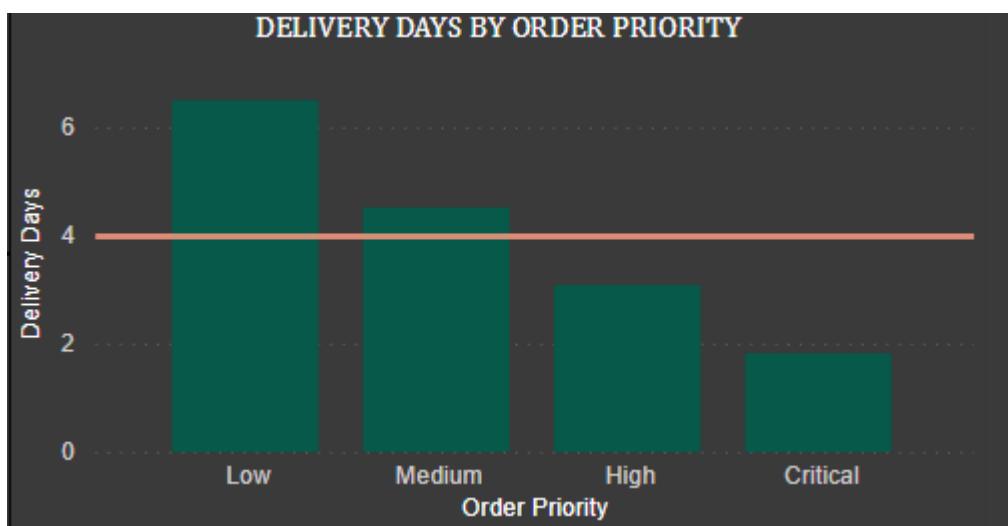
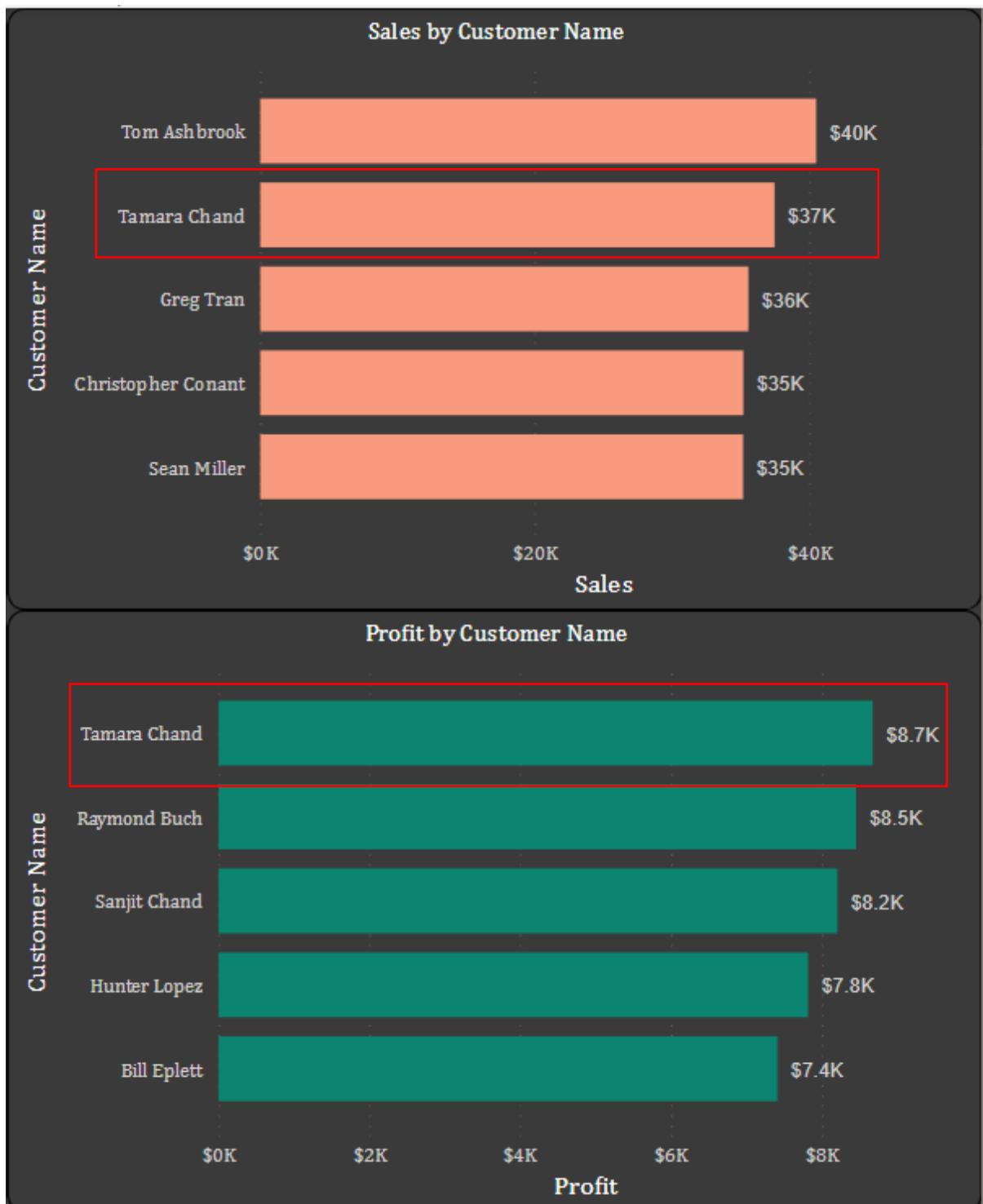
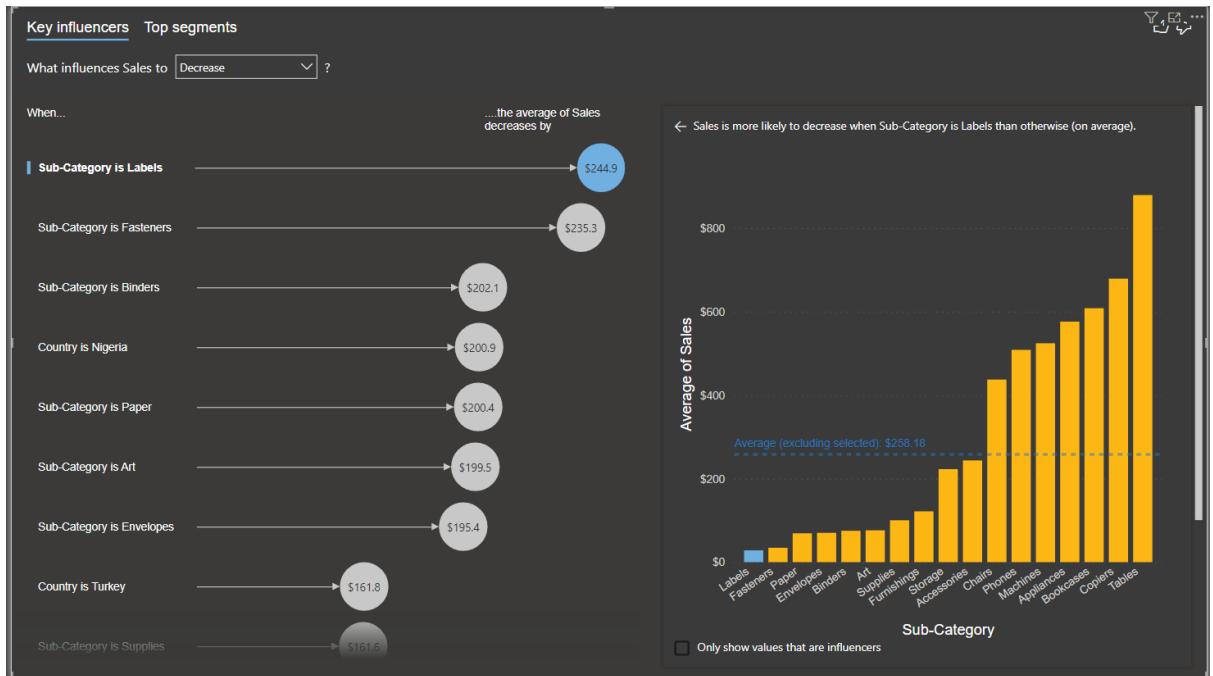


Fig: The average delivery days by order priority

To analyse sales by country, delivery days, sub-category, ship mode, and segment, a '**Key Influencer**' visualisation was used for its description in the image below. When we look at what causes sales to be low or drop, we notice that the sub-category **Labels** has the most impact, reducing sales by \$244.9 more than any other aspect. In the right panel, we will find a more detailed graphic about the sales situation. It goes into greater depth on how each item in the subcategory influences sales and causes them to fall. The second most important factor causing a decline in sales is the sub-category of **Fasteners**, which shows a \$235.3 decrease in sales. This indicates that the product is not in high demand among buyers, and as a result, fewer sales are generated. Apart from the sub-categories having a significant impact on sales, **Nigeria, Iraq, and Turkey** do not generate many sales. Understanding these important elements will aid in determining the best way to boost sales performance.

- ix. The top customers who bring in more sales are not necessarily the reason for the superstore profit. The other customers who order often and bring in more profit are not regular customers. From the charts below, we see that only **Tamara Chand** orders more product and is more profitable to the superstore. A stacked bar chart was used to visualize the top 5 customers by sales and by profits.



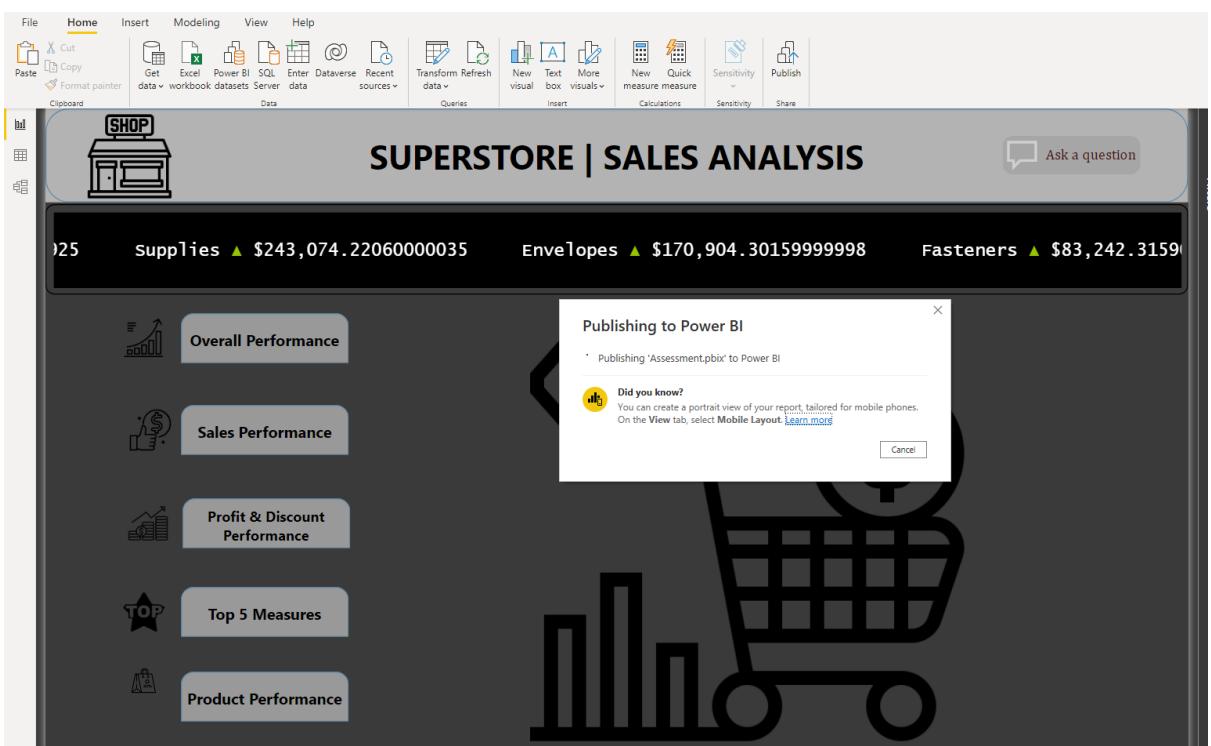
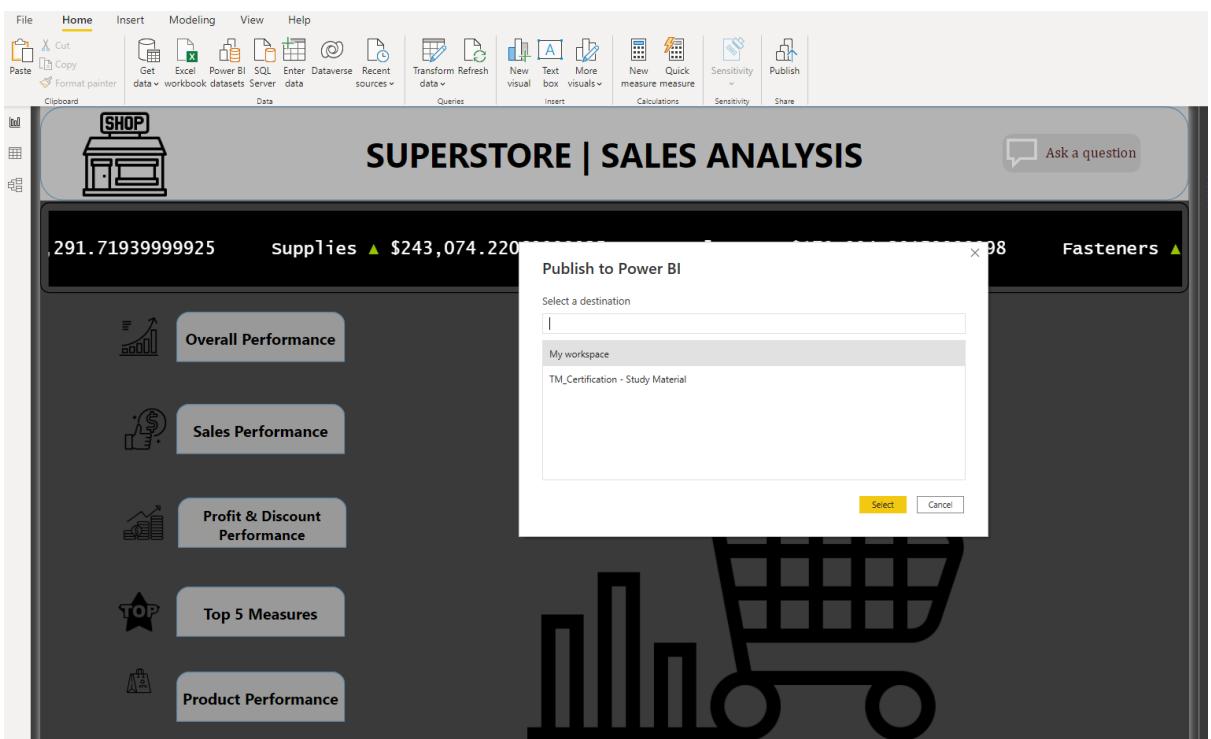


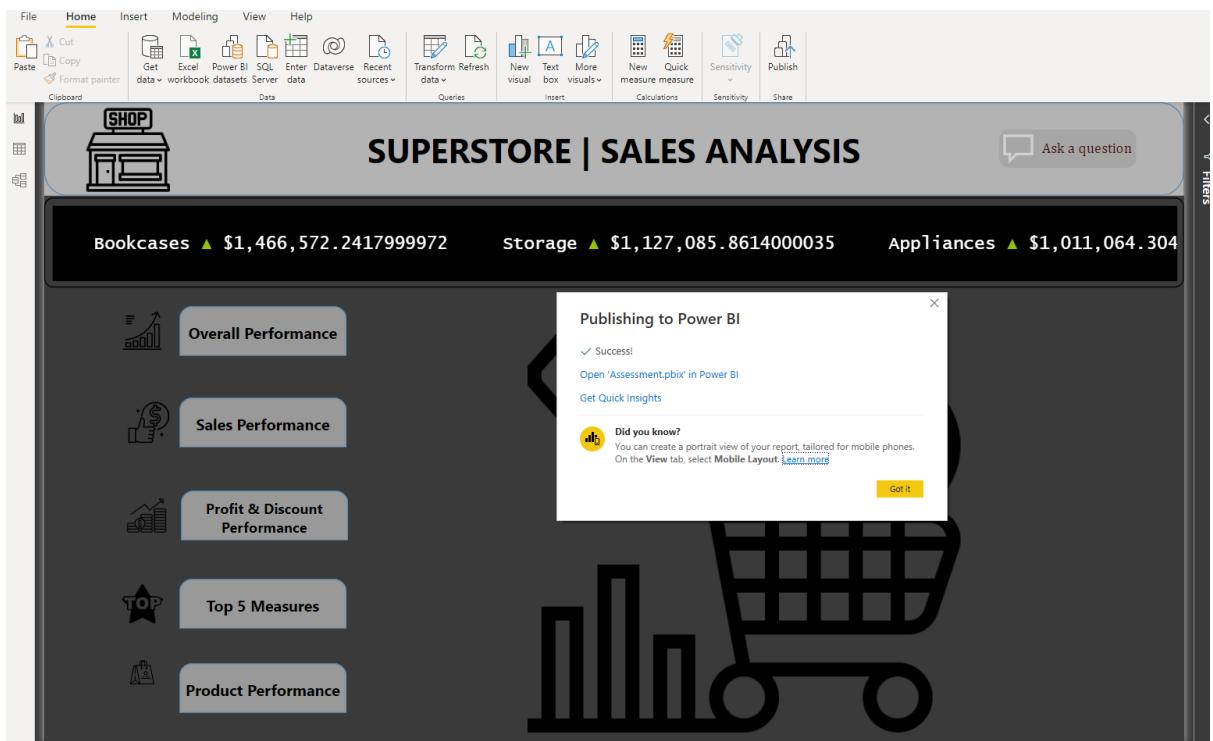
Figs: Key Influencers to decrease in sales

5 REPORT SUPPORT INFORMATION

This report includes a Power BI report (*.pbix) that contains the analysis. This report is divided into several pages, each of which focuses on a different area of sales performance, such as product, location, customer, profit, and discount. The report is published using the 'Publish' button in

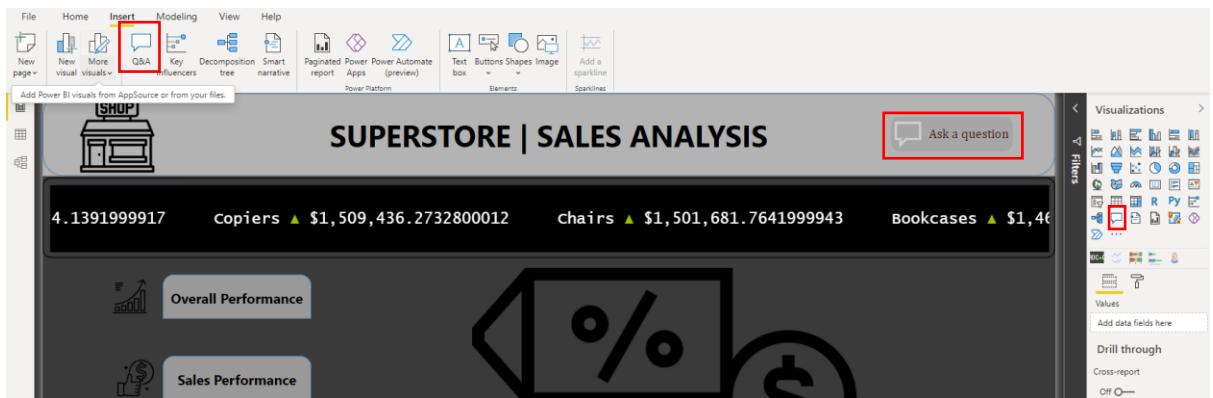
the **Home** ribbon, as shown in the figure below, which illustrates the report's processing and successful completion.





Figs: Publishing a report

The report and dashboard are also supported to provide a viewers to ‘ask a question about the data using the ‘QnA’ button and this gives a prompt answer to data-related questions.



Power BI also proposes some data-related questions. If we wish to know the top countries by sales by city, Power BI will quickly present the answer to the inquiry, as shown below, using the most appropriate visualisation tool:

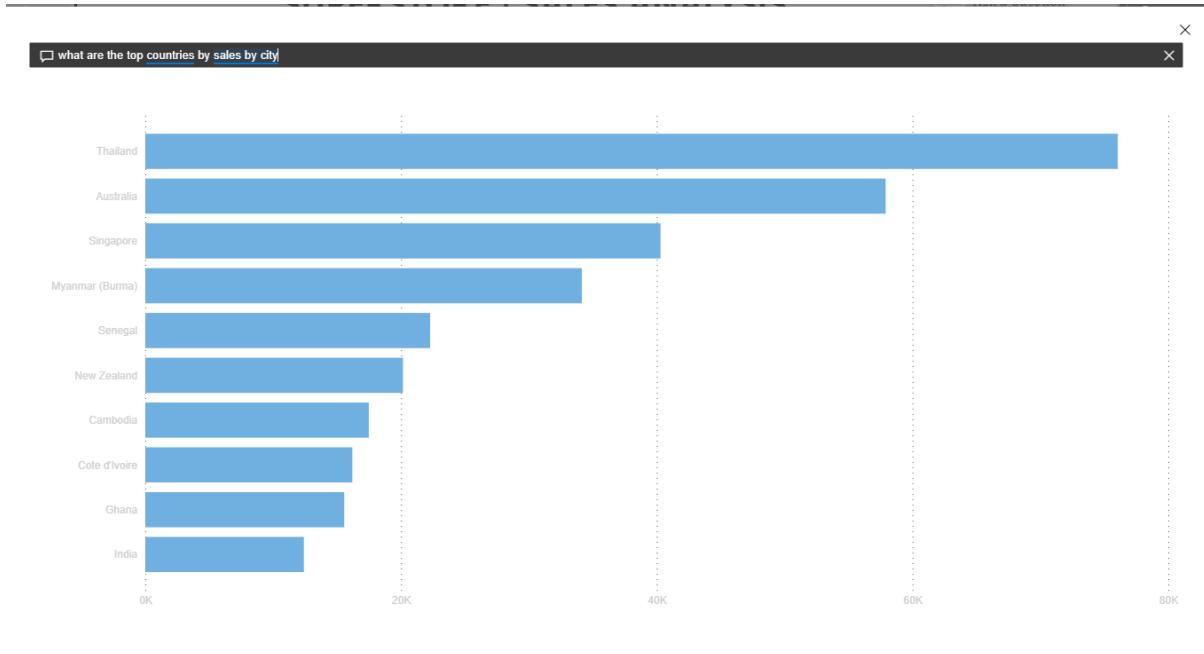


Fig: Power BI suggested question using QnA

Conclusion

- As the year progresses, sales and profit rise.
- The average number of days it takes to deliver the products is four.
- Technological products are the best-selling products over the years.
- There are usually more sales in the last quarter of the year.
- The more discounts each product group received, the more profit they n retrospect, the consumer segment products received the highest discount of \$3.8k and, as a result, the highest profit of \$750k.
- Customers who contributed the most to superstore sales are not the reason for superstore profit because they are the ones who order the most discounted items.

Recommendations

- Reducing the number of delivery days to two days would increase customer satisfaction and possibly the sales too.
- To increase profitability, EMEA, African, and Canadian markets should be targeted.

- Customers should be permitted to rate products because this may assist the superstore choose which products to focus on more and which products are the best sellers.
- Customers would have better shopping experiences and sales if they are segmented based on their profile and references.
- Some of the clients who contribute the most to profit aren't regulars. To improve profits, several tactics such as introducing membership cards and promotions can be used to keep customers.