Dublin Business School

DESING AND DEVELOPMENT OF DATA STORAGE SOLUTIONS FOR ANALYSIS

BY:

Aditya Anilkumar

Table of Contents

1. Introduction 2

1.1 Reasons for selection the subject area and data 2

2. Schema 2

3. ETL 4

4. Visualizations and Reports 7

4.1 DashBoard 7

4.2 Total Sales per branch 8

4.3 Product Sold 8

4.4 Products Sales per week 10

4.4 Monthly Sales and taxes 12

5. Graph Databases 18

6. Conclusions 20

7. Bibliography 20

8. Appendix B – Neo4j code 20

1. Introduction

After going through an analysis process, it is necessary to carry out a logical and physical design for data storage. To carry out this design, it is recommended to use dimensional modeling so that these data can be found by users in an intuitive and fast way.

In the following steps, the use of dimensional modeling will be explained in such a way as to analyze the data of our origin (Supermarket Sales) that refer to facts, whether economic or of other types, from the perspective of their components or dimensions (using for this purpose a metric or business measure).

- **Reasons for selection the subject area and data**

Because we had many source data options, none of them matched in at least two dimensions. Finally, the sales area is selected, because it was the closest data source to apply the tasks required for this assignment. In addition, this data can be used to analyze the following:

- In which weeks of 2019 were there more sales?
- In which branch is the company experiencing the most success?
- What product should you continue to invest in to sell?

2. Schema

The datasets have been downloaded from Kaggle. This dataset was then stored in Google Drive. Links for these datasets are as follows:

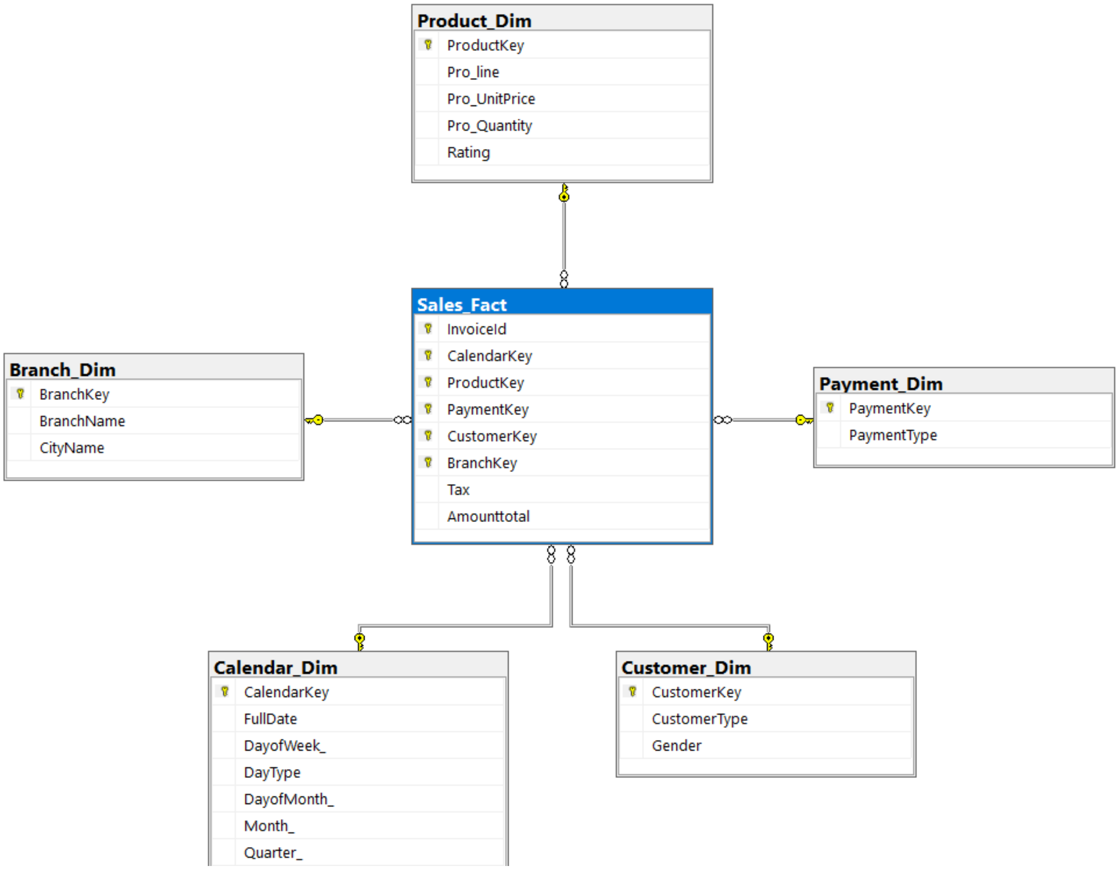
- www.kaggle.com (Samek, 2020)

- https://drive.google.com/drive/folders/1pWfz7ms90lk_rrdVnP2LVOc8LGIdJQK1?usp=sharing

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Qty	Tax 5%	Total	Date	Time
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	01/05/2019	13:08
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	03/08/2019	10:29
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	03/03/2019	13:23

Payment	cogs	gross income	Rating
Ewallet	522.83	26.1415	9.1
Cash	76.4	3.82	9.6
Credit card	324.31	16.2155	7.4

With the following data, we have dimensions such as Branch, Customer, Product, Calendar.



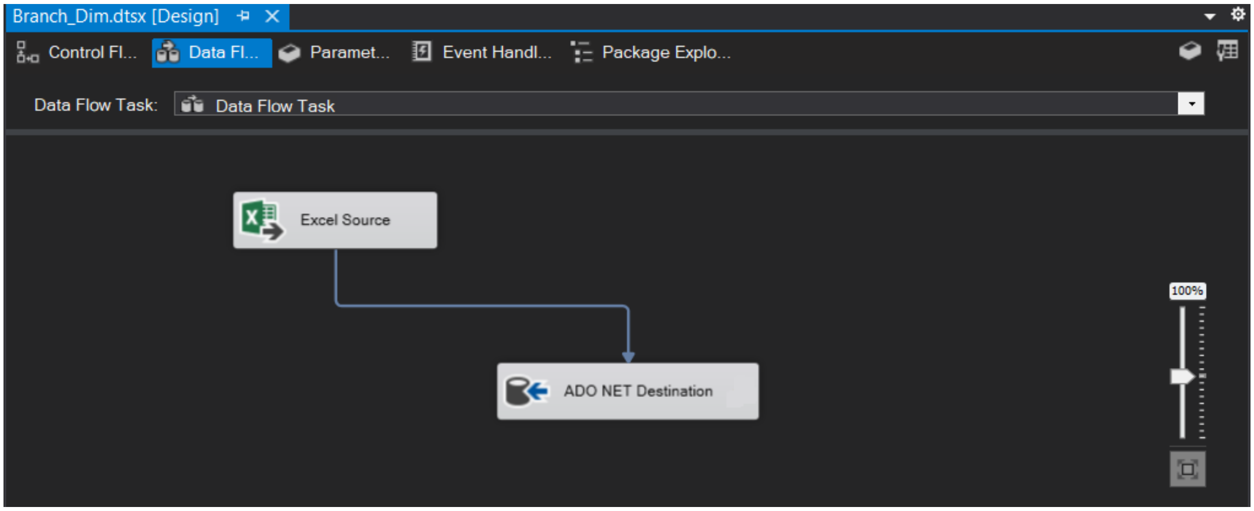
As it is a small data and little detail. Because we lack location information, it was decided that cities should be included in the branch. Like Gender and Customer type, they are included in the Customer dimension. Invoice can be a dimension; in this case it is included as a dimension and it is included as a primary key in the fact table, since it involves all the other data. Analyzing the data, we have Total and tax as measure.

The Scheme used here is a star design. It is composed of the act table Sales_Fact and the tables connected to it for Customer_Dim, Payment_Dim, Branch_Dim, Calendar_Dim, and Product_Dim.

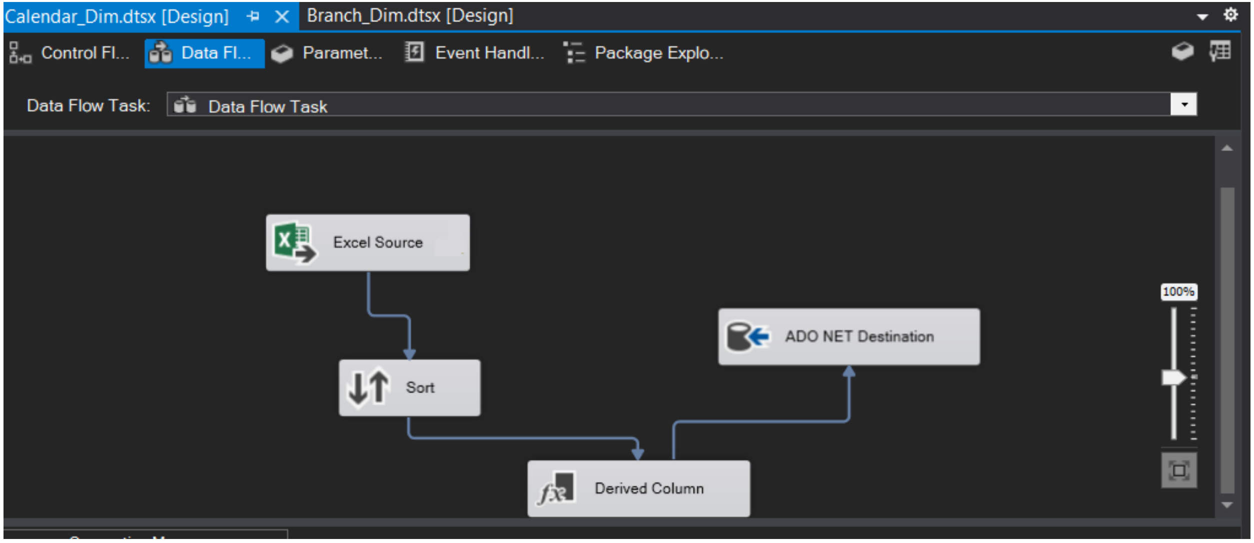
3. ETL

After we've defined the dimensional model schema, we'll ETL our data from its source.

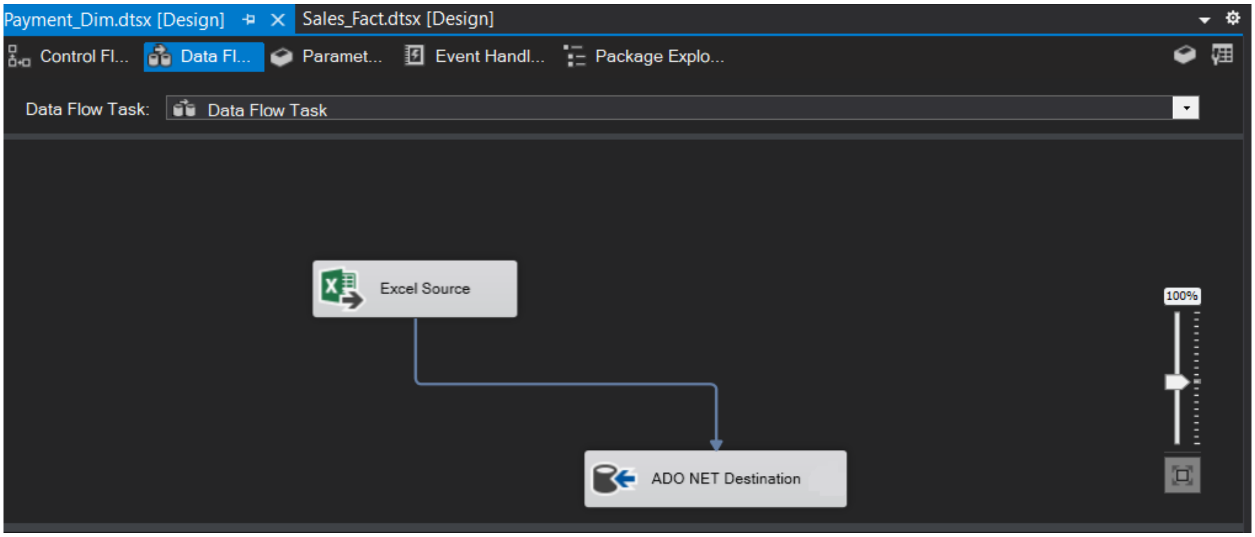
The source in this case is an Excel sheet, and the steps are as follows:



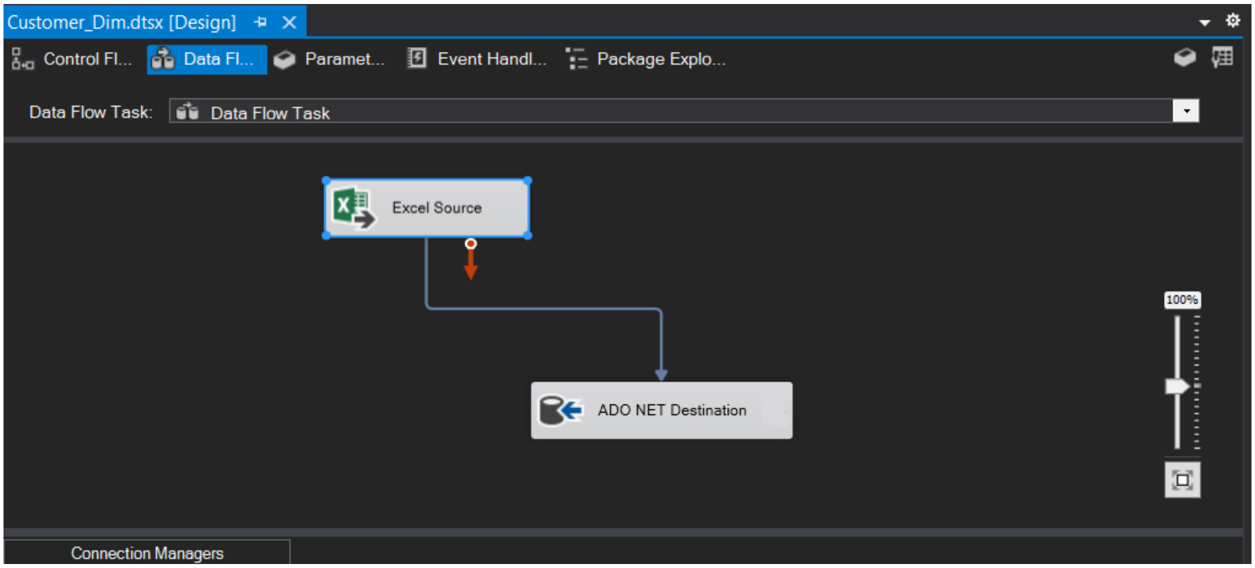
SELECT Branch, City
FROM ['supermarket_sales - Sheet1$']
GROUP BY Branch, City



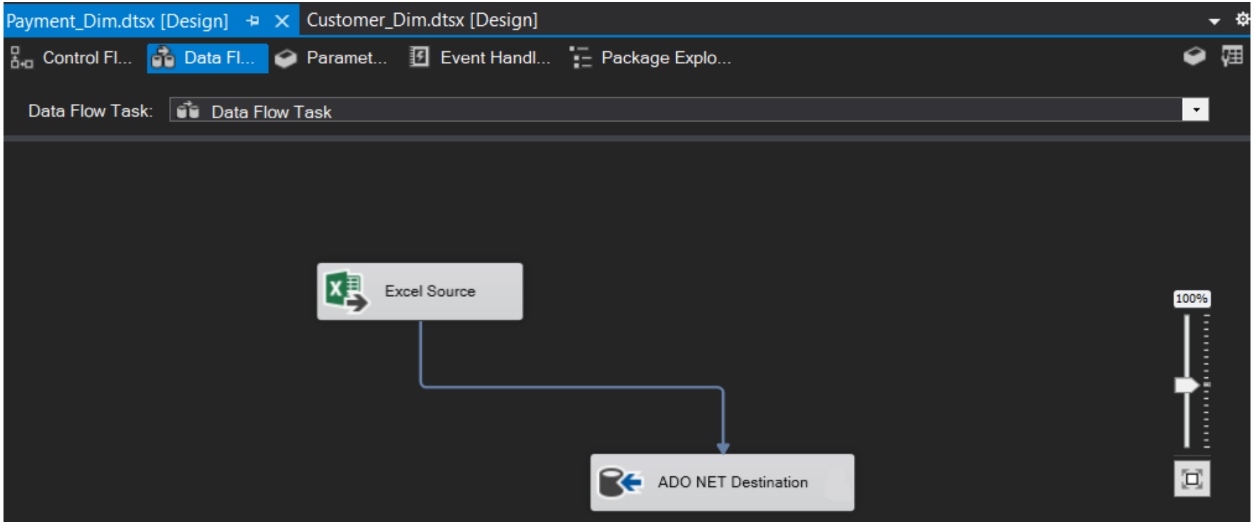
```
(DATEPART("dw",TDate)==1? "Sunday": DATEPART("dw",TDate)==2?
"Monday":DATEPART("dw",TDate)==3?
"Tuesday":DATEPART("dw",Date)==4?"Wednesday":DATEPART("dw",Date)==5?
"Thursday":DATEPART("dw",Date)==6? "Friday":DATEPART("dw",Date)==7?
"Saturday":"")

(DATEPART("dw",Date)==1? "Weekend": DATEPART("dw",Date)==2?
"Weekday":DATEPART("dw",Date)==3?
"Weekday":DATEPART("dw",Date)==4?"Weekday":DATEPART("dw",Date)==5?
"Weekday":DATEPART("dw",Date)==6? "Weekday":DATEPART("dw",Date)==7?
"Weekend":"")

(DATEPART("qq",Date)==1? "Q1": DATEPART("qq",Date)==2?
"Q2":DATEPART("qq",Date)==3? "Q3":DATEPART("qq",Date)==4?"Q4":"")
```



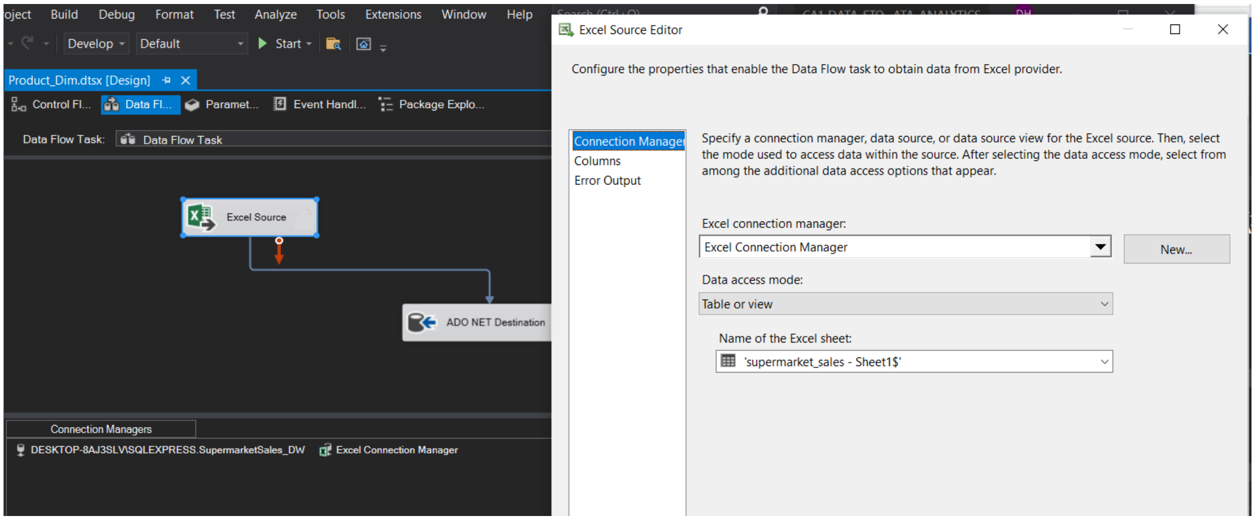
```
SELECT Payment
FROM     ['supermarket_sales - Sheet1$']
GROUP BY Payment
```



SELECT Customertype, Gender
FROM ['supermarket_sales - Sheet1$']
GROUP BY Customertype, Gender



SELECT Payment
FROM ['supermarket_sales - Sheet1$']
GROUP BY Payment



Excel Source Editor

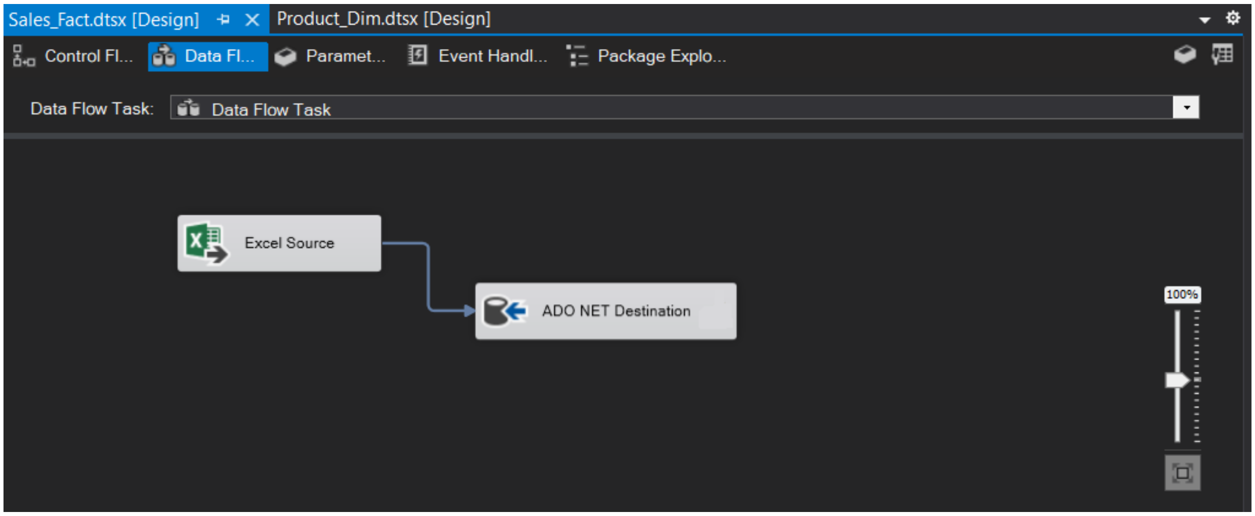
Configure the properties that enable the Data Flow task to obtain data from Excel provider.

Specify a connection manager, data source, or data source view for the Excel source. Then, select the mode used to access data within the source. After selecting the data access mode, select from among the additional data access options that appear.

Excel connection manager: Excel Connection Manager

Data access mode: Table or view

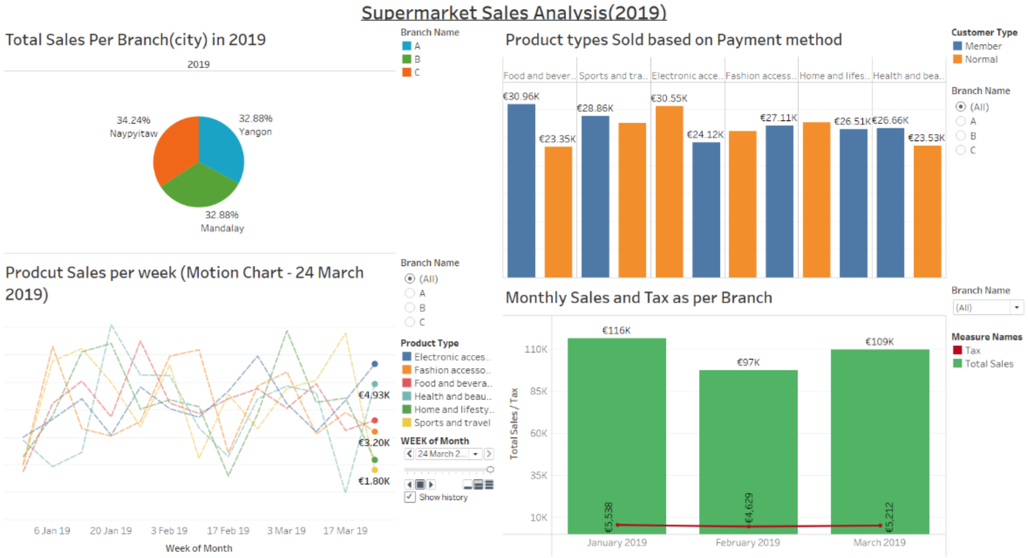
Name of the Excel sheet: 'supermarket_sales - Sheet1$'



4. Visualizations and Reports

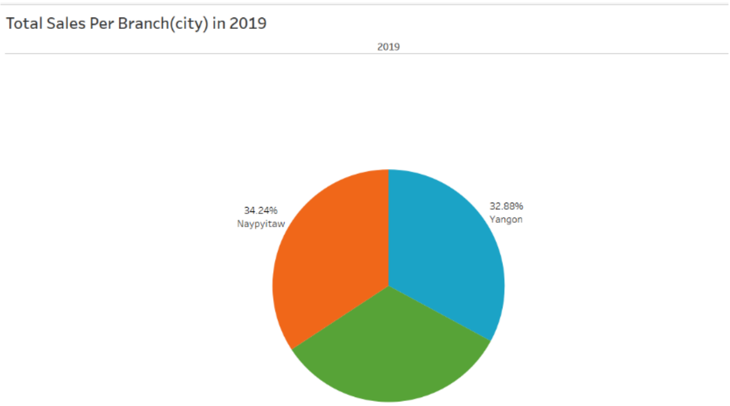
For the visualization step, Tableau and SSRS reports were used:

4.1 Dashboard Tableau



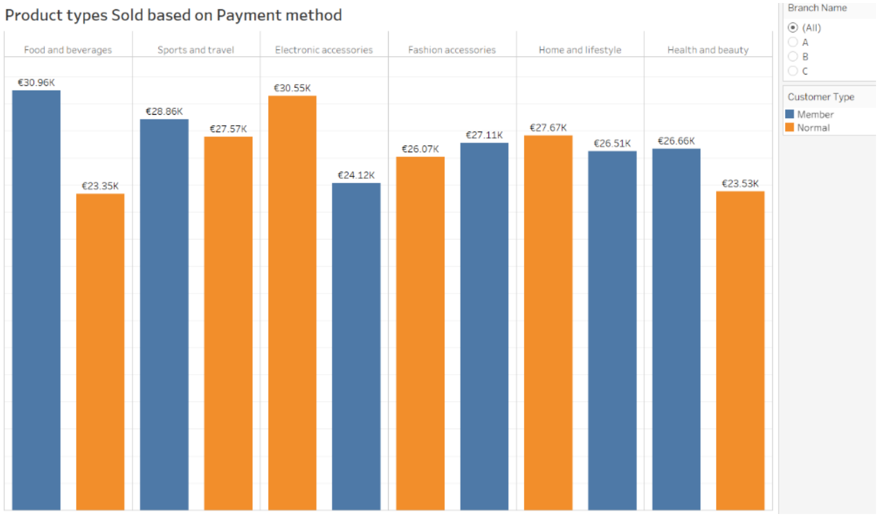
This visualization is about a supermarket and the sales of things in that supermarket based on the sorts of products sold in cities in Myanmar (Naypyitaw, Yangon, and Mandalay) in the year 2019 between the ranges shown (Jan-March).

4.2 Total Sales per branch

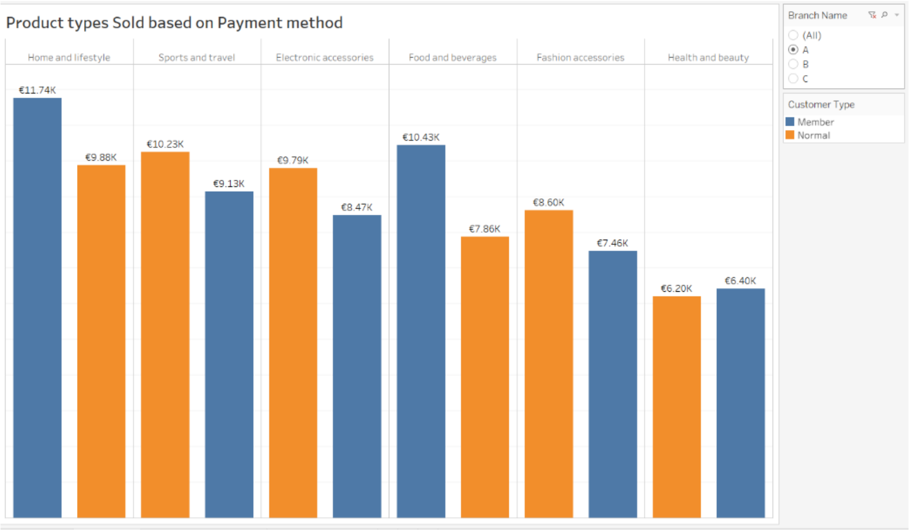


As a result of this, we can observe that it has a somewhat larger number of sales in Naypyitaw than the other two locations since it is the administrative capital of Myanmar. Now that we have everything under control, the team can concentrate on marketing the store in other cities.

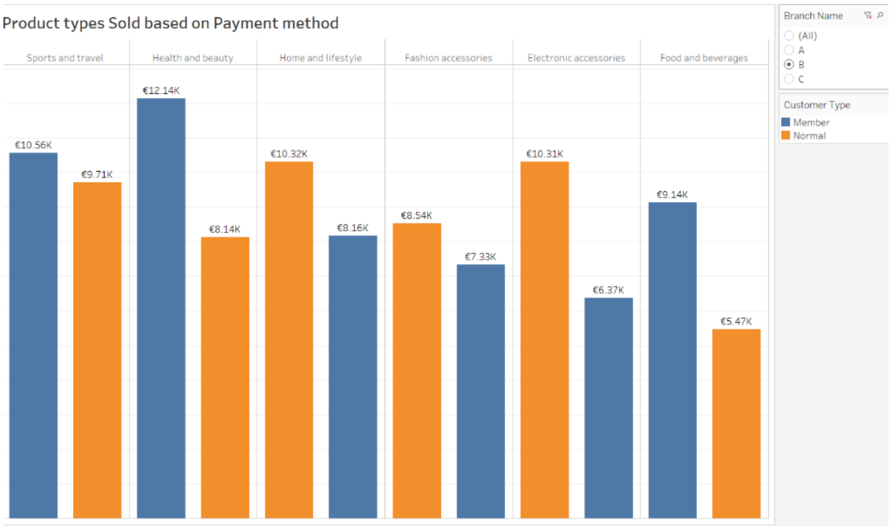
4.3 Product Sold



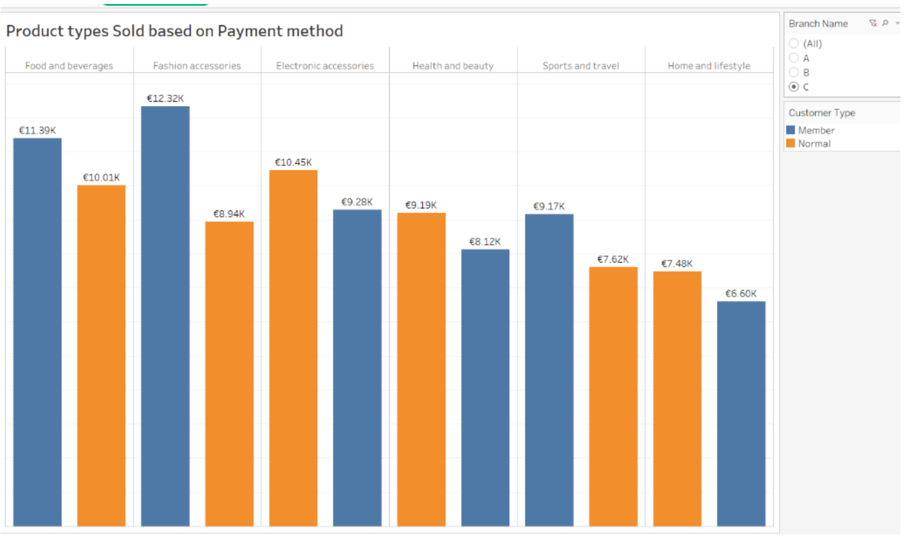
When all three locations are considered together, the most popular products to sell are those related to food and beverages, which generate a revenue of 30,96K for a member client. Normal consumers tend to purchase electronic accessories considering all three locations, this results in a total income of 30,550.



The category that sells the most at branch A is home and lifestyle; nevertheless, member customers have a tendency to spend more money here in comparison to regular clients.

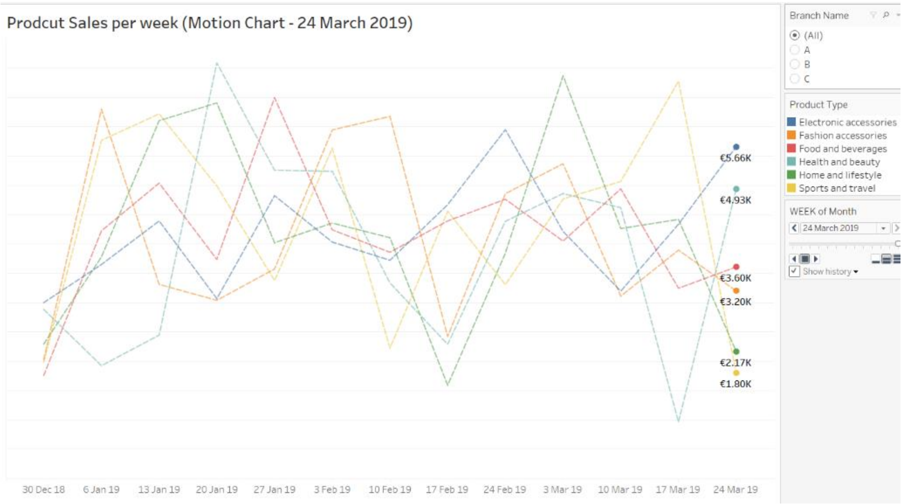


Within department B, sales of health and beauty items brought in the most money, totaling $12,14k for member clients.

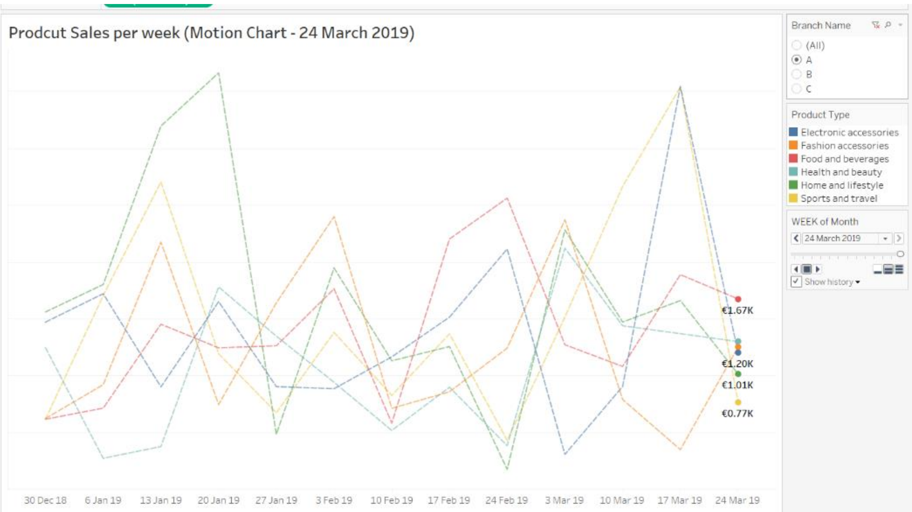


Food and drinks brought in 21.40 thousand dollars in sales for branch C, while home and lifestyle brought in over 15,000 dollars, making it the least profitable branch.

- **4.4 Products Sales per week**



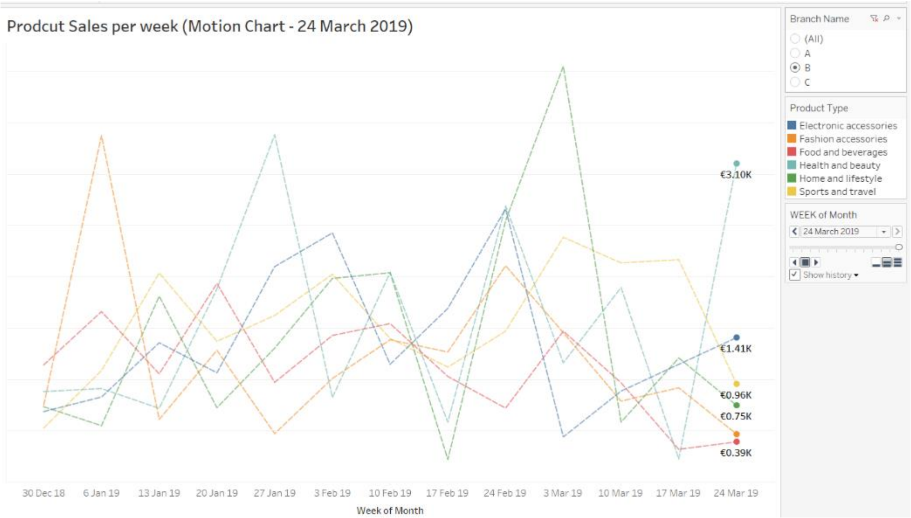
Based on the aggregated statistics from all three departments, we can conclude that the Electronics accessories department brought in the highest sales (5.66k), while the Sports and Travel department only had sales of 1.80k.



Branch Name
○ (All)
● A
○ B
○ C

WEEK of Month
24 March 2019
☑ Show history

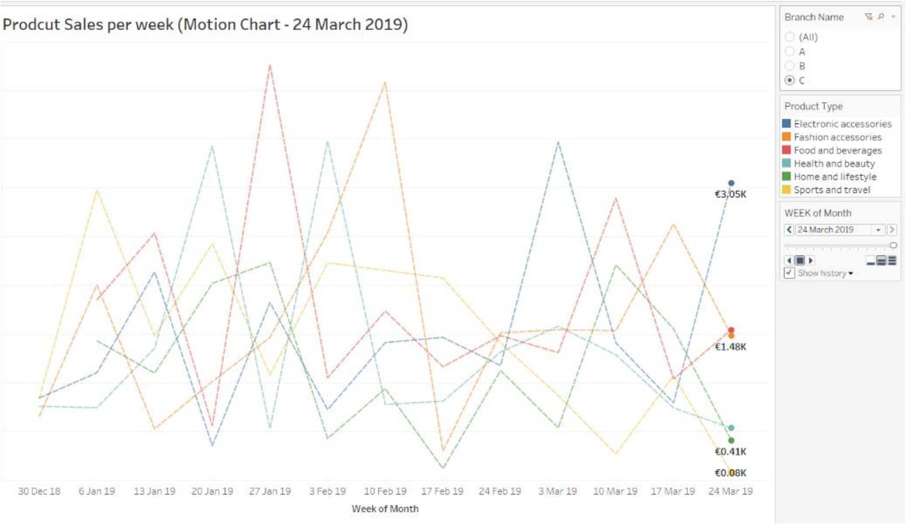
When we look at branch A, we can see that the trend of home and leisure items reaches its highest point during the sale season and then begins to decline as soon as the sale is over. The sales of all items are very typical, exhibiting both highs and lows.



Branch Name
○ (All)
○ A
● B
○ C

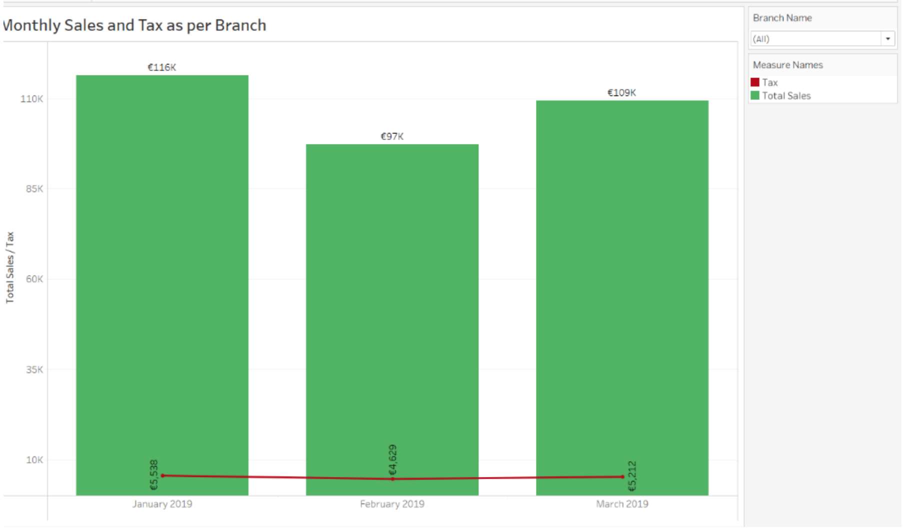
WEEK of Month
24 March 2019
☑ Show history

For Branch B, we can see that the sales of health and beauty items and fashion accessories both increased significantly during the month of January. This trend continued throughout the period.

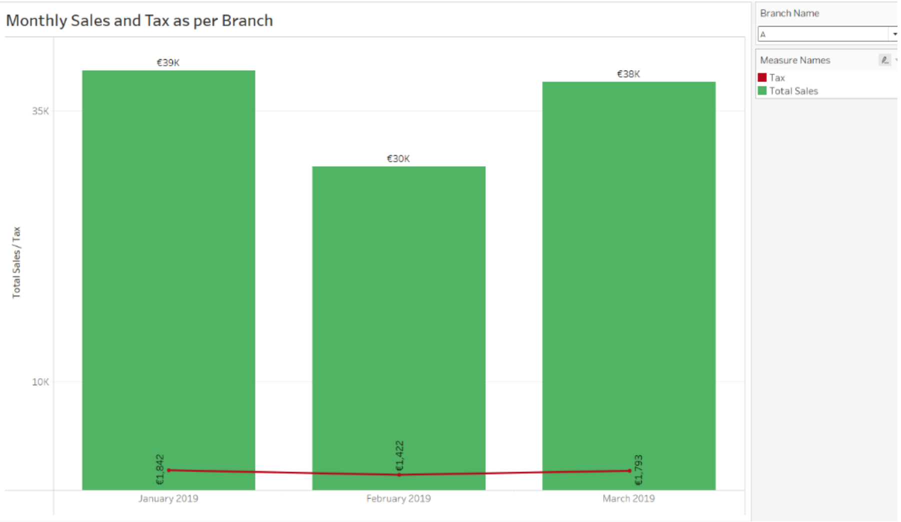


At the conclusion of the time, the branch C that deals in electronic accessories has the largest sales, although throughout the period that deals in food and drinks, the branch C that deals in fashion accessories has exhibited an increasing tendency.

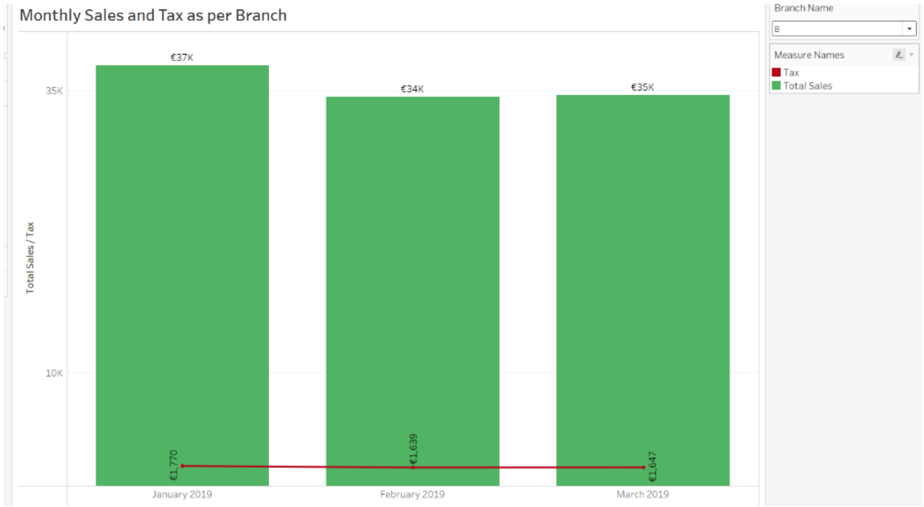
4.4 Monthly Sales and taxes



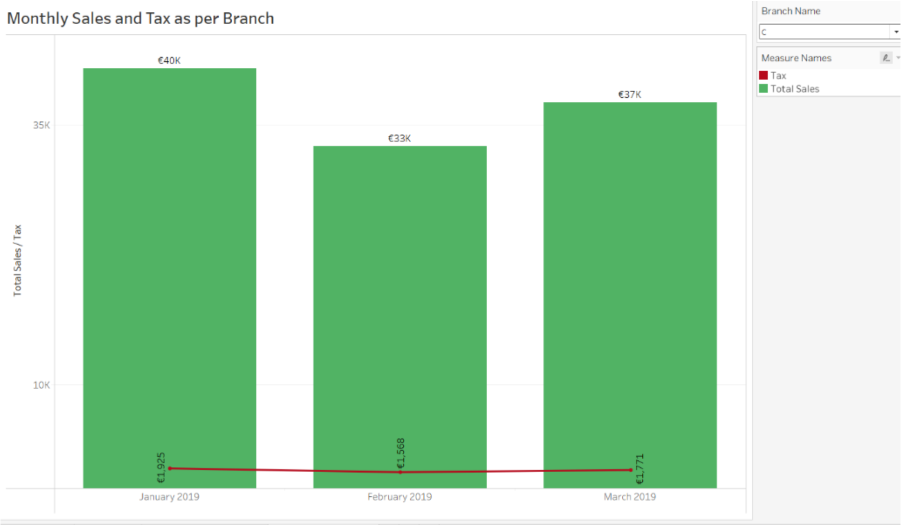
The combined taxes of the supermarket chain's three locations were highest in the month of January since it was the month that brought in the most revenue. The tax rate is just 5%.



When compared to the other branches, Branch A was required to pay a lower amount of tax and had the fewest sales during the month of February. However, Branch A paid the highest tax during the month of January.



For Branch B, the total amount of taxes paid for all three months is more than 1,500, and its sales total more than 33,000.

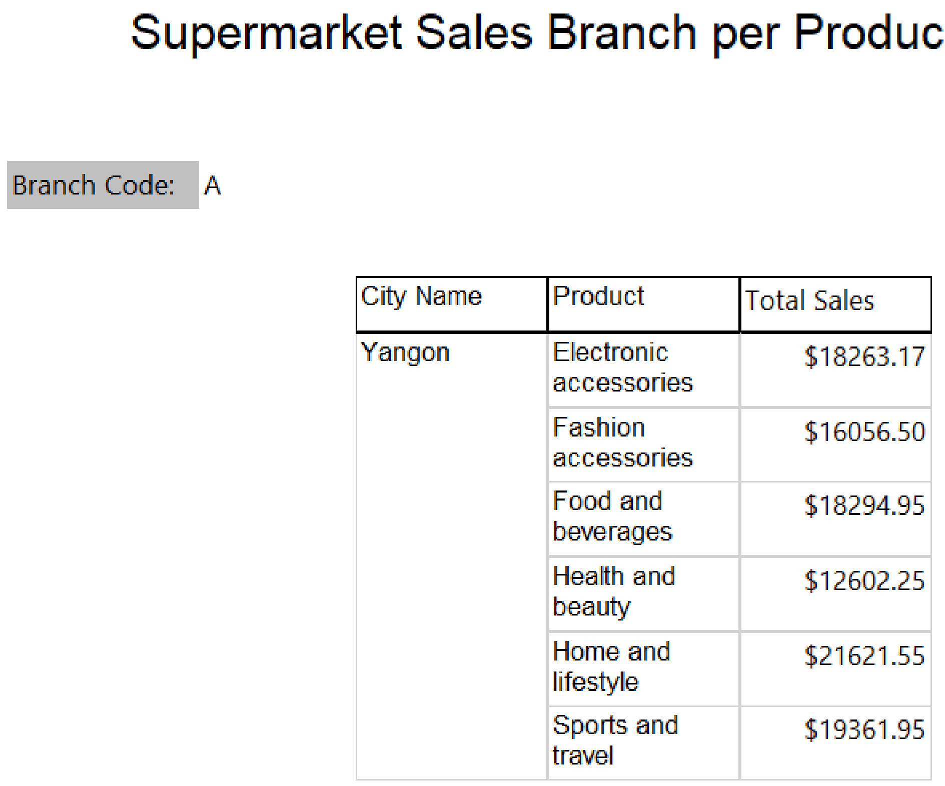


Because Jan had the largest sales in branch C (40,000), he was responsible for paying the most tax (2,000).

- **SSRS (SQL Sever Reporting Services)**

We have used the reporting services extension in Visual Studio 2019 for the preparation of reports.

In this report, the total of each product is obtained according to the branches of Naypyitaw, Mandalay, and Yangon.



Supermarket Sales Branch per Product Report

Branch Code: A

City Name	Product	Total Sales
Yangon	Electronic accessories	$18263.17
	Fashion accessories	$16056.50
	Food and beverages	$18294.95
	Health and beauty	$12602.25
	Home and lifestyle	$21621.55
	Sports and travel	$19361.95

Supermarket Sales Branch per Product Report

Branch Code: B

City Name	Product	Total Sales
Mandalay	Electronic accessories	$16678.38
	Fashion accessories	$15871.80
	Food and beverages	$14609.89
	Health and beauty	$20280.03
	Home and lifestyle	$18479.51
	Sports and travel	$20278.06

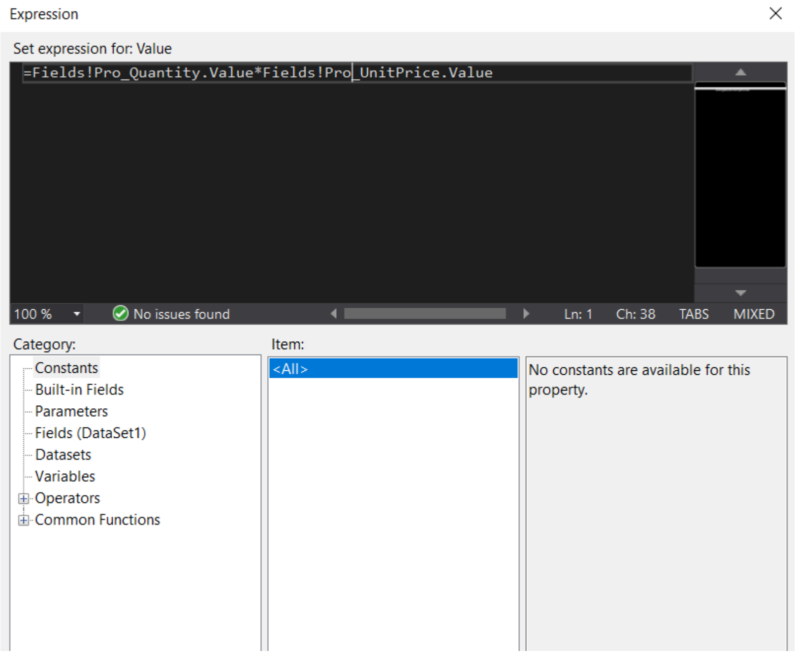
Supermarket Sales Branch per Product Report

Branch Code: C

City Name	Product	Total Sales
Naypyitaw	Electronic accessories	$19733.93
	Fashion accessories	$21258.83
	Food and beverages	$21404.53
	Health and beauty	$17309.28
	Home and lifestyle	$14076.60
	Sports and travel	$16785.52

The following report details the monthly sales and cost of goods sold (COGS), which is the sum of all direct costs associated with making a product.

To obtain the cost of goods sold, the following calculation expression was made.



Expression

Set expression for: Value

```
=Fields!Pro_Quantity.Value*Fields!Pro_UnitPrice.Value
```

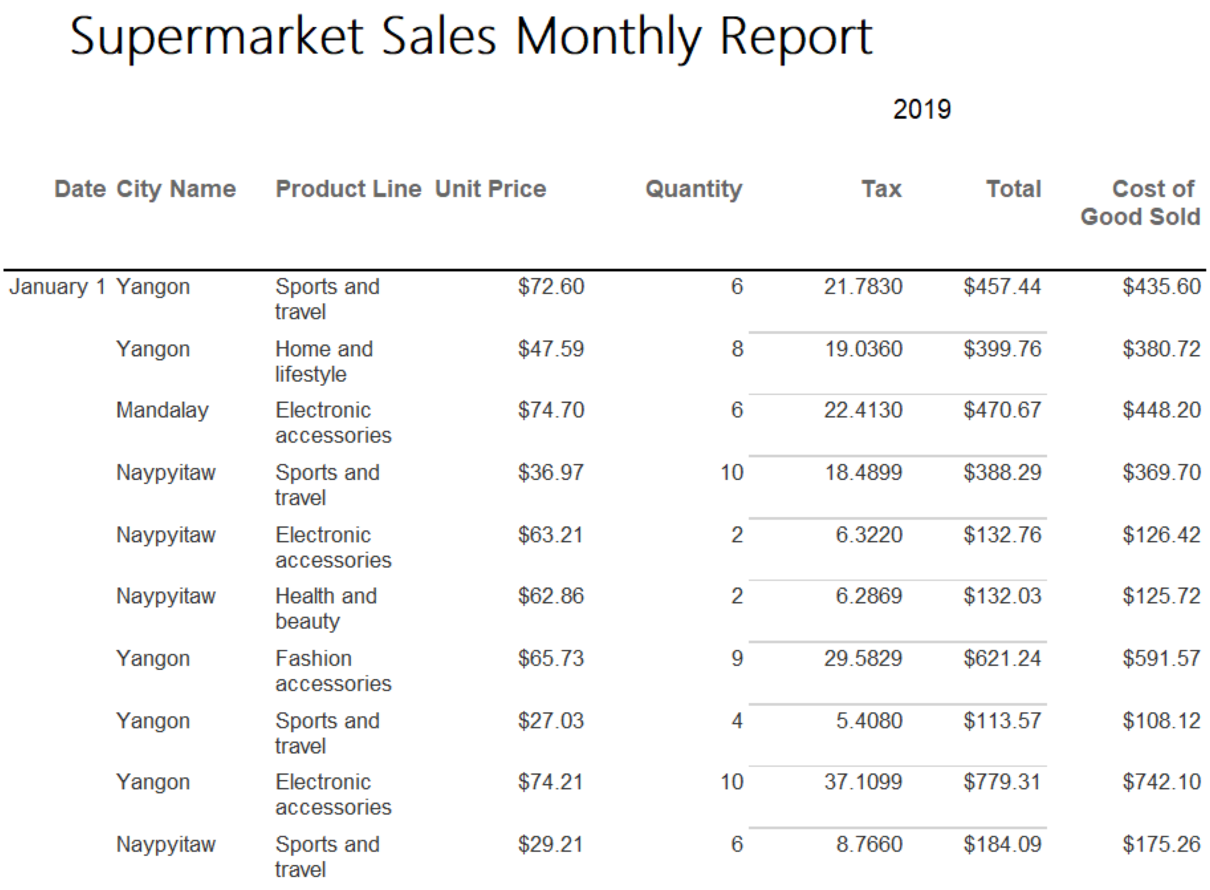
100 % · No issues found · Ln: 1 Ch: 38 TABS MIXED

Category:

- Constants
- Built-in Fields
- Parameters
- Fields (DataSet1)
- Datasets
- Variables
- Operators
- Common Functions

Item: <All>

No constants are available for this property.

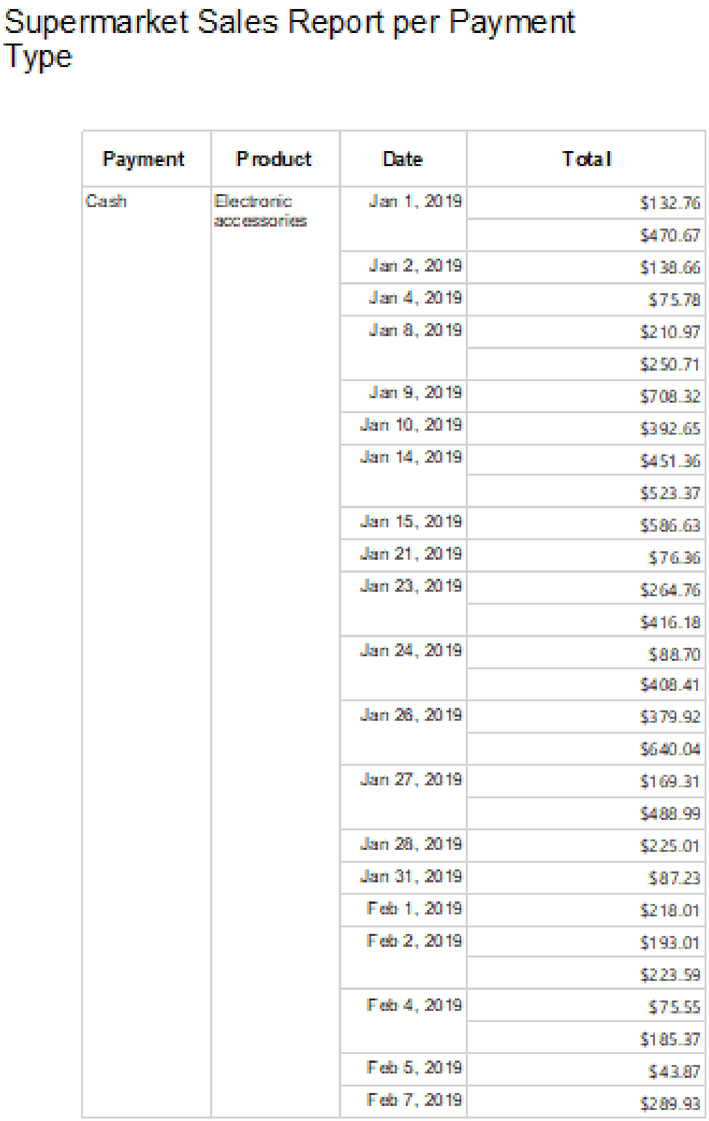


Supermarket Sales Monthly Report

2019

Date	City Name	Product Line	Unit Price	Quantity	Tax	Total	Cost of Good Sold
January 1	Yangon	Sports and travel	$72.60	6	21.7830	$457.44	$435.60
	Yangon	Home and lifestyle	$47.59	8	19.0360	$399.76	$380.72
	Mandalay	Electronic accessories	$74.70	6	22.4130	$470.67	$448.20
	Naypyitaw	Sports and travel	$36.97	10	18.4899	$388.29	$369.70
	Naypyitaw	Electronic accessories	$63.21	2	6.3220	$132.76	$126.42
	Naypyitaw	Health and beauty	$62.86	2	6.2869	$132.03	$125.72
	Yangon	Fashion accessories	$65.73	9	29.5829	$621.24	$591.57
	Yangon	Sports and travel	$27.03	4	5.4080	$113.57	$108.12
	Yangon	Electronic accessories	$74.21	10	37.1099	$779.31	$742.10
	Naypyitaw	Sports and travel	$29.21	6	8.7660	$184.09	$175.26

The following report details sales by type of payment, according to the date.



Supermarket Sales Report per Payment Type

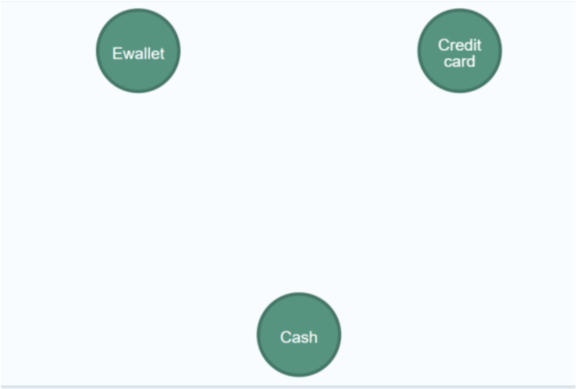
Payment	Product	Date	Total
Cash	Electronic accessories	Jan 1, 2019	$132.76
			$470.67
		Jan 2, 2019	$138.66
		Jan 4, 2019	$75.78
		Jan 8, 2019	$210.97
			$250.71
		Jan 9, 2019	$708.32
		Jan 10, 2019	$392.65
		Jan 14, 2019	$451.36
			$523.37
		Jan 15, 2019	$586.63
		Jan 21, 2019	$76.36
		Jan 23, 2019	$264.76
			$416.18
		Jan 24, 2019	$88.70
			$408.41
		Jan 26, 2019	$379.92
			$640.04
		Jan 27, 2019	$169.31
			$488.99
		Jan 28, 2019	$225.01
		Jan 31, 2019	$87.23
		Feb 1, 2019	$218.01
		Feb 2, 2019	$193.01
			$223.59
		Feb 4, 2019	$75.55
			$185.37
		Feb 5, 2019	$43.87
		Feb 7, 2019	$289.93

5. Graph Databases

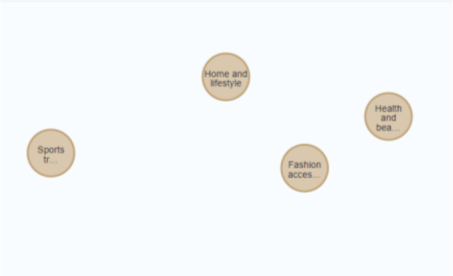
Neo4j is the world’s leading graph database. The architecture is designed for optimal management, storage, and traversal of nodes and relationships. The graph database takes a property graph approach, which is beneficial for both traversal performance and operation runtime. Neo4j offers dedicated memory management and memory-efficient operations. Neo4j is scalable and can be set up as a single server or on a group of machines in a production environment that can handle failures. Other features for production applications include hot backups and extensive monitoring. (Neo4j, 2022)

With this tool, we proceed to create all the graphs of our data:

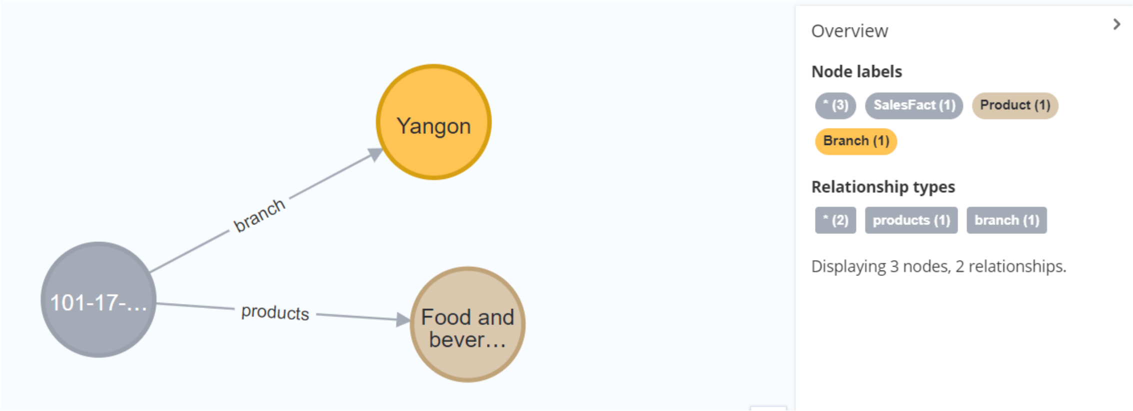
See APPENDIX B – Code Branch



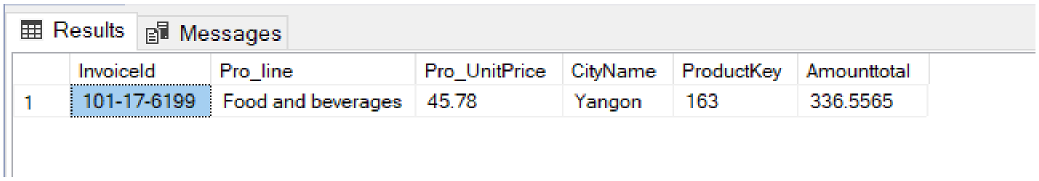
See APPENDIX B – Code Payment



See APPENDIX B – Code Product.



Displaying 3 nodes, 2 relationships.



	InvoiceId	Pro_line	Pro_UnitPrice	CityName	ProductKey	Amounttotal
1	101-17-6199	Food and beverages	45.78	Yangon	163	336.5565

See APPENDIX B – Code Relationship types (Products and branch)

6. Conclusions

In conclusion, what was done in this assignment shows that data warehouses become important in data analysis because they allow us to integrate data from different sources and provide us with tools to make use of this information and support decision-making. Using SSIS made it easy to link our databases, which let us make a number of visual reports that helped us get a quick and accurate idea of how much money supermarkets made.

In the different steps elaborated, they are based on the analysis and understanding of the information, and for this, different tools are used for data collection, ETL, visualization, reporting of the results, and other alternatives for data analysis such as Neo4j.

7. Bibliography

-

Neo4j. (2022, 11 8). *Neo4J*. Retrieved from https://neo4j.com/docs/operations-manual/current/introduction/

Samek, J. (2020). *Supermarket Sales Data Analysis*. Retrieved from https://www.kaggle.com/code/jordansamek/supermarket-sales-data-analysis/data

8. Appendix B – Neo4j code

Code Branch

```
LOAD CSV WITH HEADERS FROM "file:///Branch.csv" as row CREATE (b:Branch) SET b=
{branch_id:row.BranchKey, nameBranch:row.BranchName, nameCity: row.CityName} return b
```

Code Payment

```
LOAD CSV WITH HEADERS FROM "file:///Payment.csv" as row CREATE(pa:Payment) SET
 pa = {PaymentId:row.PaymentKey, PaymentType:row.PaymentType} return pa
```

Code Calendar

```
LOAD CSV WITH HEADERS FROM "file:///Calendar.csv" as row CREATE (c:Calendar) SET
c= {calendarid:row.CalendarKey, fullDate:row.fullDate, DayofWeek:row.DayofWeek_,
DayType:row.DayType, DayofMonth:row.DayofMonth_, Month:row.Month,
Quarter:row.Quarter_, Year:row.Year_} return c
```

Code SalesFact

```
LOAD CSV WITH HEADERS FROM "file:///SalesFact.csv" as row CREATE (sf:SalesFact) SE
T sf = {invoiceId:row.InvoiceId, CalendarId:row.CalendarKey, ProductId:row.ProductKey, Pay
mentId:row.PaymentKey,CustomerId:row.CustomerKey,BranchId:row.BranchKey,Tax:row.Tax,
 Amounttotal:row.Amounttotal} RETURN sf
```

Code Product

```
LOAD CSV WITH HEADERS FROM "file:///Product.csv" as row CREATE (p:Product) SET p
= {productid:row.ProductKey, Proline:row.Pro_line, ProUnitPrice:row.Pro_UnitPrice, ProQuant:
row.Pro_Quantity, Rating:row.Rating} return p
```

Code Customer

```
LOAD CSV WITH HEADERS FROM "file:///Customer.csv" as row CREATE(cu:Customer)
SET cu = {Customerid:row.CustomerKey, CustomerType:row.CustomerType,
Gender:row.Gender} return cu
```

Code Relationship types (Products and branch)

```
MATCH (sf:SalesFact), (b:Branch), (p:Product) WHERE  sf.BranchId = b.branchid and sf.ProductId = p.productid and sf.invoiceId='101-17-6199' CREATE (sf)-[r:branch{CityName:b.nameCity, ProductName:p.Proline, ProductUnitPrice:p.ProUnitPrice}] ->(b) return sf,p,b
```

```
SELECT sf.InvoiceId, p.Pro_line, p.Pro_UnitPrice, b.CityName, p.ProductKey,
sf.Amounttotal
  FROM Sales_Fact sf
INNER JOIN Product_Dim p
    ON sf.ProductKey = p.ProductKey
INNER JOIN Branch_Dim b ON sf.BranchKey = b.BranchKey
      WHERE sf.InvoiceId = '101-17-6199'
```