INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT

(Project Semester August-December 2020)

SUPERMARKET SALES ANALYSIS

Submitted by
Vaishnavi Jaiswal
11812089

B.Tech CSE (KM079)
Course Code INT217

Under the Guidance of
Ashu Mehta(23631)

Discipline of CSE/IT
Lovely School of Computer Science and Engineering
Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Vaishnavi Jaiswal bearing Registration no. 11812089 has completed Introduction to Data Management (INT 217) project titled, **“Supermarket Sales Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor Ashu Mehta
Assistant Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date:

DECLARATION

I, Vaishnavi Jaiswal, student of Bachelor in Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 4th December,2020

Signature

Registration No. 11812089 Vaishnavi Jaiswal

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher Mrs. Ashu Mehta who gave me the golden opportunity to do this wonderful project of analysis of the data of a superstore namely "SUPERMARKET SALES ANALYSIS" which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to her.

Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Table of Contents

INTRODUCTION

Data Analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

A supermarket is a self-service shop offering a wide variety of food, beverages and household products, organized into sections. It is larger and has a wider selection than earlier grocery stores, but is smaller and more limited in the range of merchandise than a hypermarket or big-box market. Supermarket are becoming a very important part of our lives. They provide the customer to pick the items from a huge collection and at a lesser price. This not only makes the customer happy and is also a hassle-free shopping technique.

Supermarkets typically are chain stores, supplied by the distribution centers of their parent companies, thus increasing opportunities for economies of scale. Supermarkets usually offer products at relatively low prices by using their buying power to buy goods from manufacturers at lower prices than smaller stores can. They also minimize financing costs by paying for goods at least 30 days after receipt and some extract credit terms of 90 days or more from vendors. Certain products (typically staple foods such as bread, milk and sugar) are very occasionally sold as loss leaders so as to attract shoppers to their store. Supermarkets make up for their low margins by a high volume of sales, and with of higher-margin items bought by the attracted shoppers. Self-service with shopping carts (trolleys) or baskets reduces labor costs, and many supermarket chains are attempting further reduction by shifting to self-service check-out.

Supermarkets and other grocery stores sales in the United States exceeded 650 billion U.S. dollars in 2019. Among the leading players in the food and grocery retail industry, Walmart alone generated sales worth over 270 billion U.S. dollars in that year. Hence, it is important to analyze the data generated by these because it helps in knowing the profits and income.

Along with that, it also helps in knowing the customers better and thus fulfilling their needs which helps in increasing profits and hence better revenue. The data is easy to analyze and gives a fruitful result which can be used in various ways. U.S has a large number of supermarkets which generate data and use them to know their consumer needs better.

Supermarket normally charge fewer prices than the traditional retailer and provide the large number of variety product and highly quality food. Price is the main object which attracts the customer of all classes. The customers always think about to save the money during shopping. According to reports the price of food in the supermarket is 15% lower and vegetable are 30% lower than the traditional retailer.

Thus, we will analyze the data generated by three cities namely Yangon, Naypyidaw, Mandalay. These 3 cities sell almost all items of everyday use and are categorized as: Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel. There are both male and female customers which shop and these stores provide payment facilities through cash, card and E-wallet. They also provide membership but the perks of the membership are not disclosed in the dataset.

SOURCE OF THE DATASET

Website:

https://www.kaggle.com/aungpyaeap/supermarket-sales

Publisher:

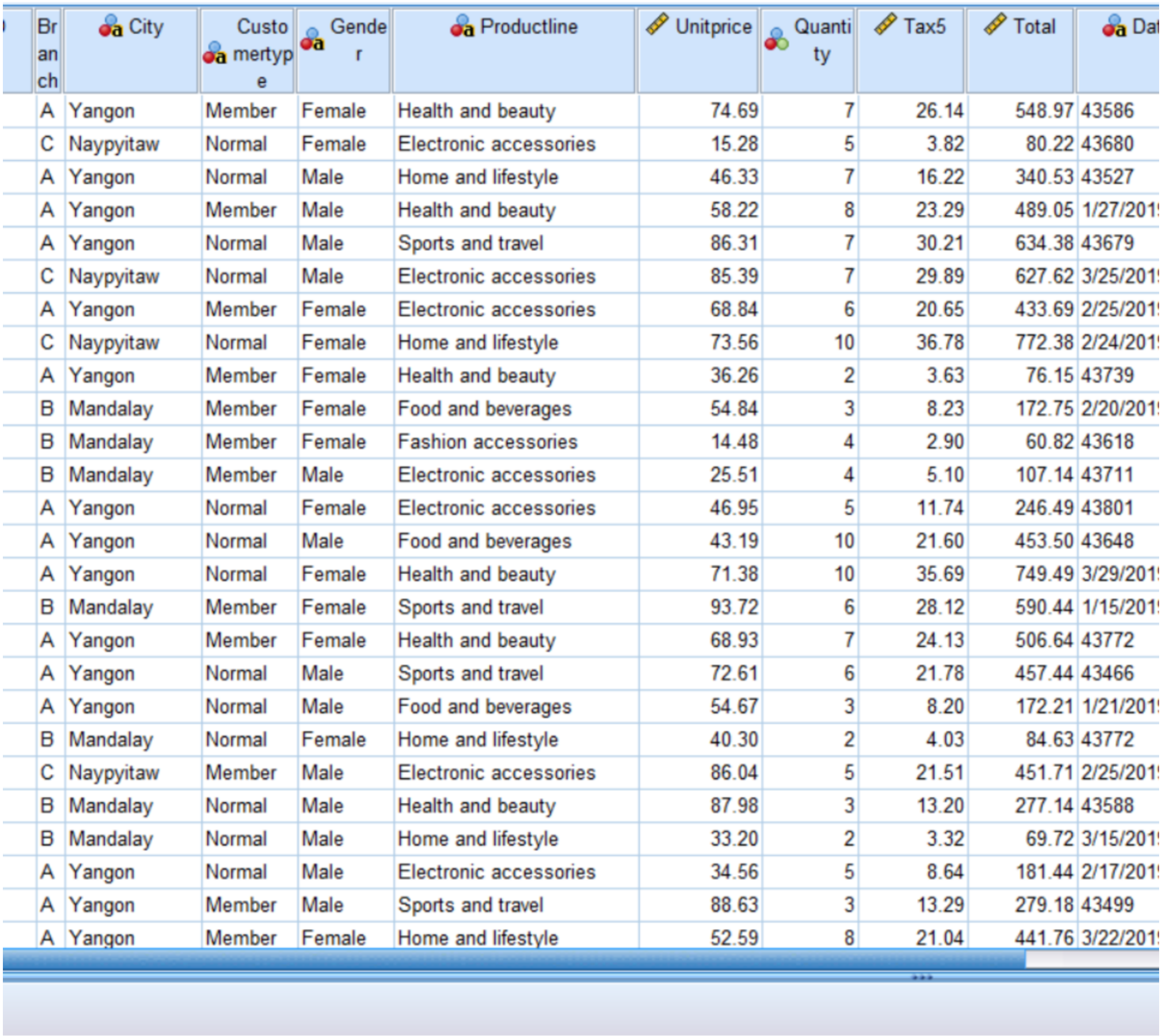
Aung Pyae

Studied M.Sc Data Science at Asia Pacific University (KL, Malaysia)

Mandalay, Mandalay Region, Myanmar (Burma)

Last Updated:

2 years ago.



Branch	City	Customertype	Gender	Productline	Unitprice	Quantity	Tax5	Total	Dat
A	Yangon	Member	Female	Health and beauty	74.69	7	26.14	548.97	43586
C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	43680
A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.22	340.53	43527
A	Yangon	Member	Male	Health and beauty	58.22	8	23.29	489.05	1/27/201
A	Yangon	Normal	Male	Sports and travel	86.31	7	30.21	634.38	43679
C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.89	627.62	3/25/201
A	Yangon	Member	Female	Electronic accessories	68.84	6	20.65	433.69	2/25/201
C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.78	772.38	2/24/201
A	Yangon	Member	Female	Health and beauty	36.26	2	3.63	76.15	43739
B	Mandalay	Member	Female	Food and beverages	54.84	3	8.23	172.75	2/20/201
B	Mandalay	Member	Female	Fashion accessories	14.48	4	2.90	60.82	43618
B	Mandalay	Member	Male	Electronic accessories	25.51	4	5.10	107.14	43711
A	Yangon	Normal	Female	Electronic accessories	46.95	5	11.74	246.49	43801
A	Yangon	Normal	Male	Food and beverages	43.19	10	21.60	453.50	43648
A	Yangon	Normal	Female	Health and beauty	71.38	10	35.69	749.49	3/29/201
B	Mandalay	Member	Female	Sports and travel	93.72	6	28.12	590.44	1/15/201
A	Yangon	Member	Female	Health and beauty	68.93	7	24.13	506.64	43772
A	Yangon	Normal	Male	Sports and travel	72.61	6	21.78	457.44	43466
A	Yangon	Normal	Male	Food and beverages	54.67	3	8.20	172.21	1/21/201
B	Mandalay	Normal	Female	Home and lifestyle	40.30	2	4.03	84.63	43772
C	Naypyitaw	Member	Male	Electronic accessories	86.04	5	21.51	451.71	2/25/201
B	Mandalay	Normal	Male	Health and beauty	87.98	3	13.20	277.14	43588
B	Mandalay	Normal	Male	Home and lifestyle	33.20	2	3.32	69.72	3/15/201
A	Yangon	Normal	Male	Electronic accessories	34.56	5	8.64	181.44	2/17/201
A	Yangon	Member	Male	Sports and travel	88.63	3	13.29	279.18	43499
A	Yangon	Member	Female	Home and lifestyle	52.59	8	21.04	441.76	3/22/201

OBJECTIVES/SCOPE OF THE ANALYSIS

Supermarket sales can be analyzed with the help of the data generated by them. In this particular dataset, the growth of supermarkets in most populated cities are increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches. There is only one sheet containing of 13 columns.

Attribute Information:

Invoice id: Computer generated sales slip invoice identification number

Branch: Branch of supercenter (3 branches are available identified by A, B and C).

City: Location of supercenters – Yangon, Naypyidaw, Mandalay.

Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card.

Gender: Gender type of customer

Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel

Unit price: Price of each product in $

Quantity: Number of products purchased by customer

Tax: 5% tax fee for customer buying

Total: Total price including tax

Date: Date of purchase (Record available from January 2019 to March 2019)

Time: Purchase time (10am to 8pm)

Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and E wallet)

COGS: Cost of goods sold

Gross margin percentage: Gross margin percentage

Gross income: Gross income

Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

There are 1000 rows and the data is cleaned.

The objectives of this analysis are:

- General Sales Analysis Branch Wise Based on Product Type
- Types of Payment used in Various Branches
- Gross Income Analysis Branch Wise Based on Product Type
- Analysis of Branch Wise Customer Rating
- Hourly Analysis Based on Number of Customers
- Overall Analysis of the Dataset

In this dashboard, I am using various slicers which will help us analyze the data in a well-mannered way. The slicers used are City, Customer Type, Gender, Product Line and Payment Type used.

NOTE: The cities and branches are namely

Yangon known as branch A

Mandalay known as branch B

Naypyidaw known as branch C

ETL APPLIED ON THE DATASET

ETL stands for extraction, transformation and loading. ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system.

Creating a Data warehouse is not simply extracting data from multiple sources and loading into database of a Data warehouse. This requires a complex ETL process.

The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

The ETL applied on the dataset are as follows:

The csv file from Kaggle website of Supermarket sales data is converted to xlxs file so that the analysis can be done easily. Since the data is now in a little bit proper form, it is easy to do changes and remove errors.

Now the file is converted to table so that we can easily apply filters on it. The table format also ensures that the new data fed inside the file is in the correct format. It also makes the sheet dynamic and it can be easily refreshed to update new values.

The sheet is checked for null values. There are 1000 rows with zero values which means that the data is not missing anywhere.

The cities and branches are matched using the filter option so that there is no mismatch of data in these.

The total and gross income is converted to currency format which makes charts and tables easy to evaluate.

ANALYSIS ON DATASET

1. *General Sales Analysis Product Wise*

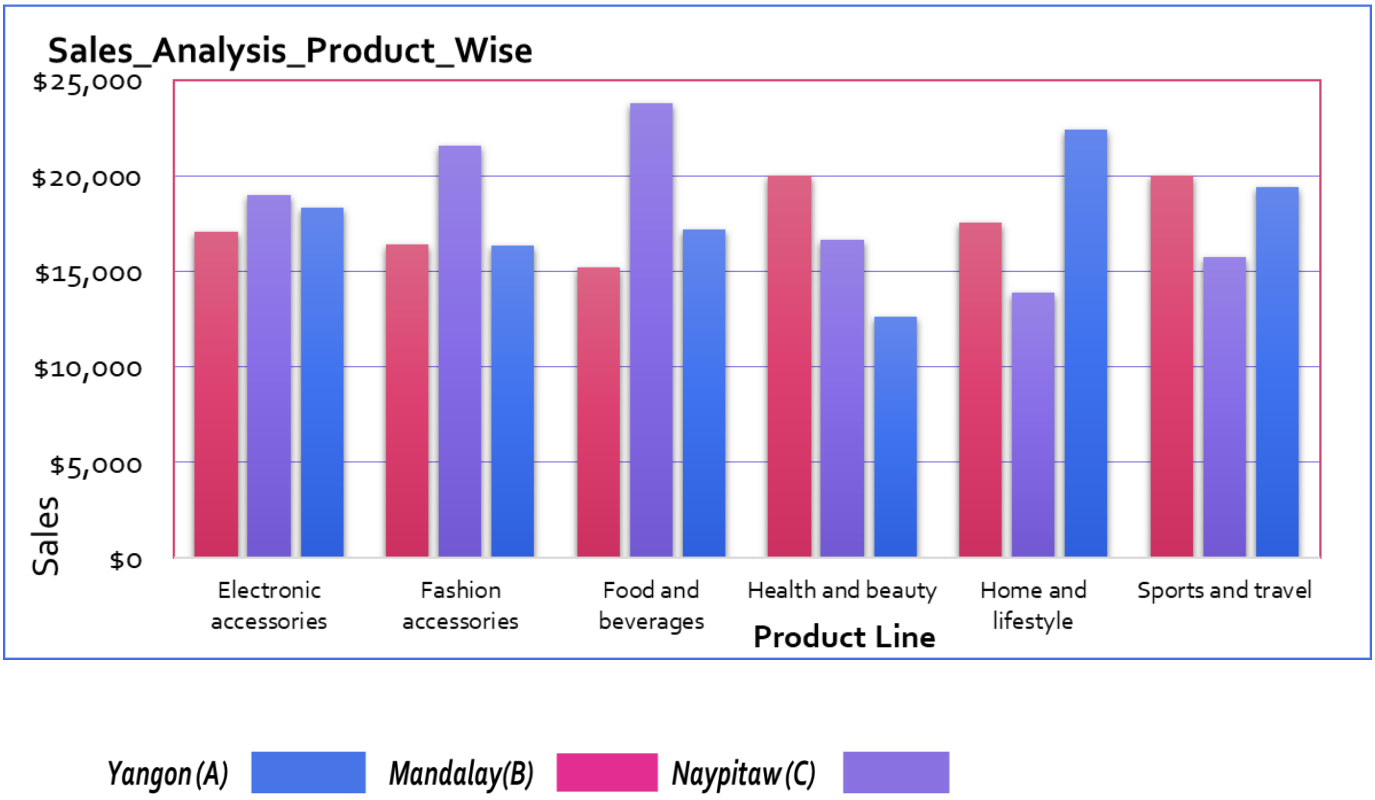
This analysis is used using the total sales generated by each branch on the basis of each product line.

Specific Requirements, functions and formulas:

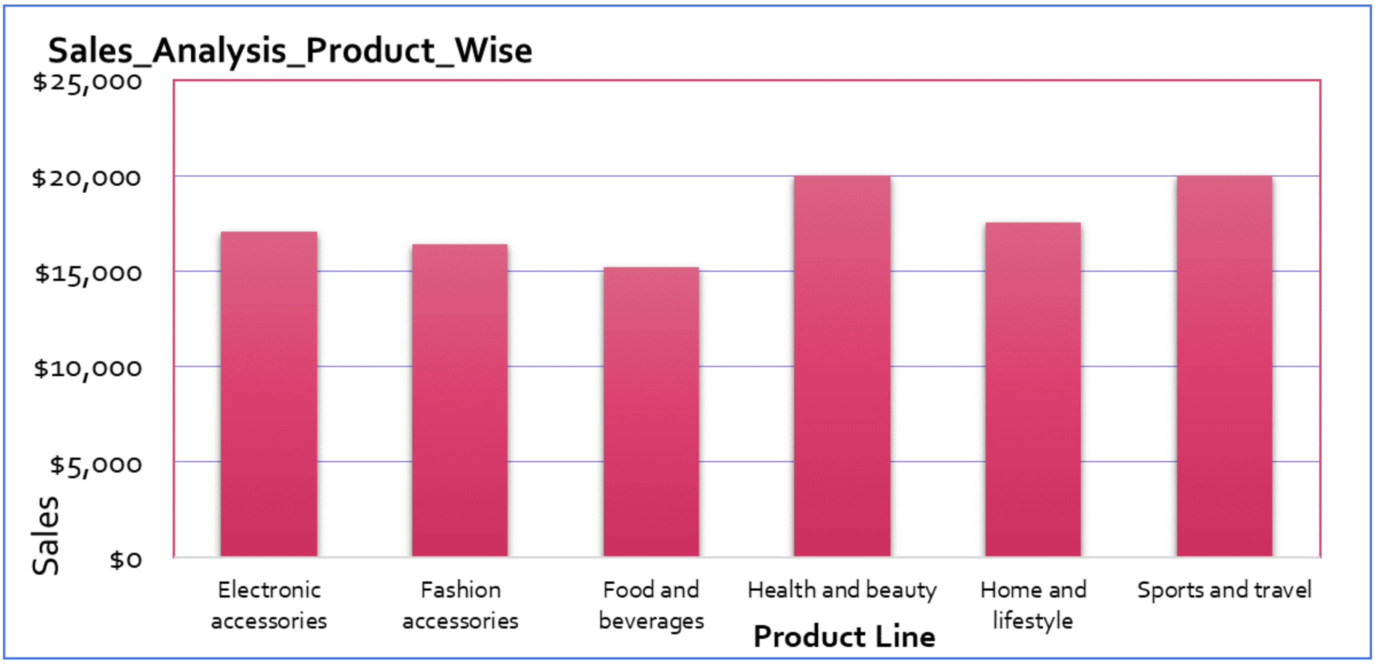
- ✓ Pivot Table with product line as rows, city as columns and total sales as values.
- ✓ Renaming the total sales values as 'sales analysis'.
- ✓ Changing the number format to the nearest decimal of zero.
- ✓ Converting the pivot table layout to tabular form.
- ✓ Adding a pivot chart of clustered column type.

Analysis Results and Visualizations:

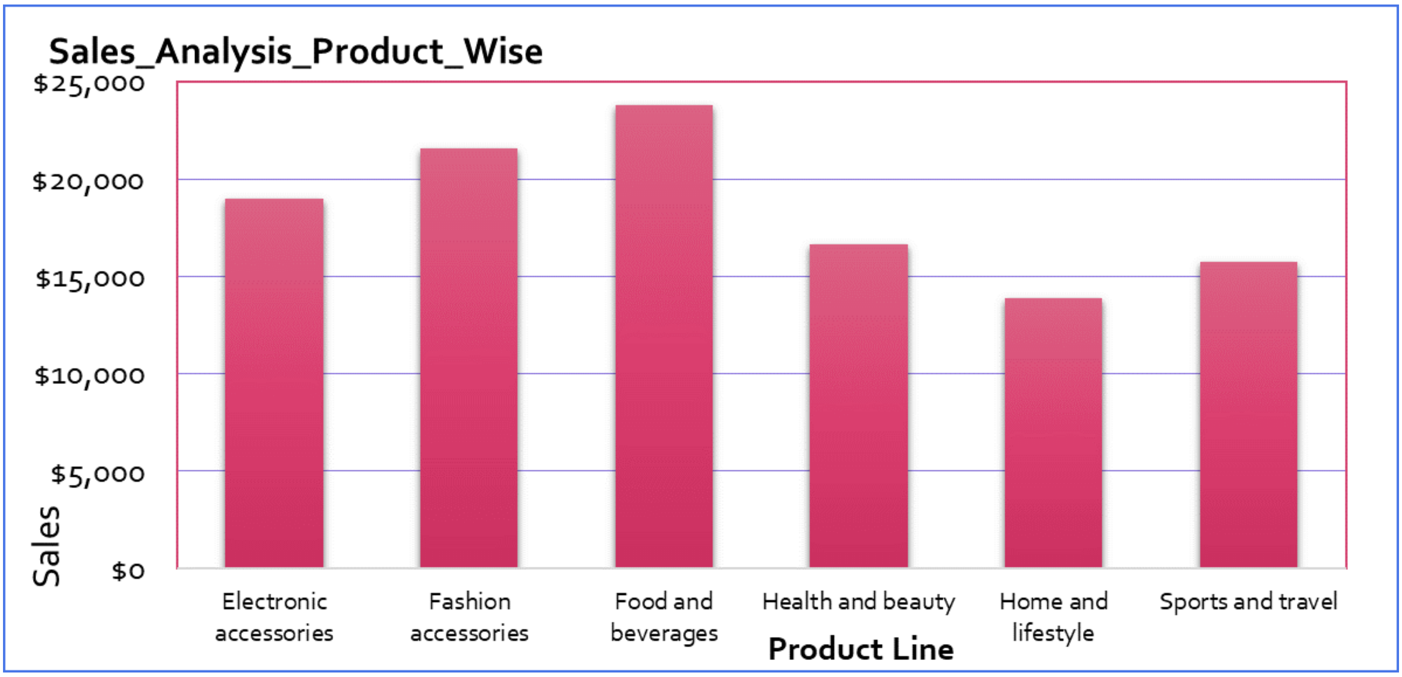
- ➢ The product line 'food and beverages' have the highest sales in all product lines in the branch 'Naypitaw'.



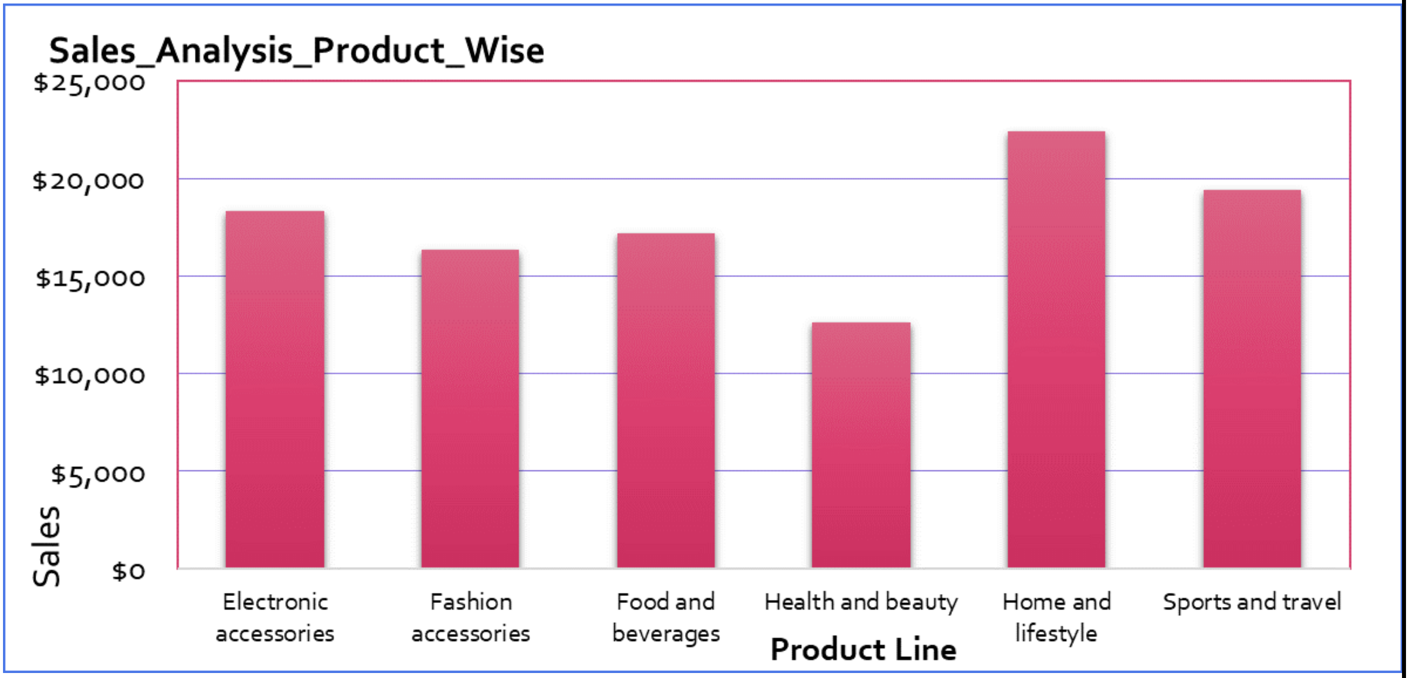
- Mandalay has the maximum sales in two product lines of ‘healthy and beauty’ & ‘sports and travel’.



- Naypitaw has the maximum sales in ‘food and beverages’ & next to it is ‘fashion accessories’.



- Yangon has the maximum sales in ‘home and lifestyle’ & just after it ‘sports and travel’



2. *Types of Payment used in various branches*

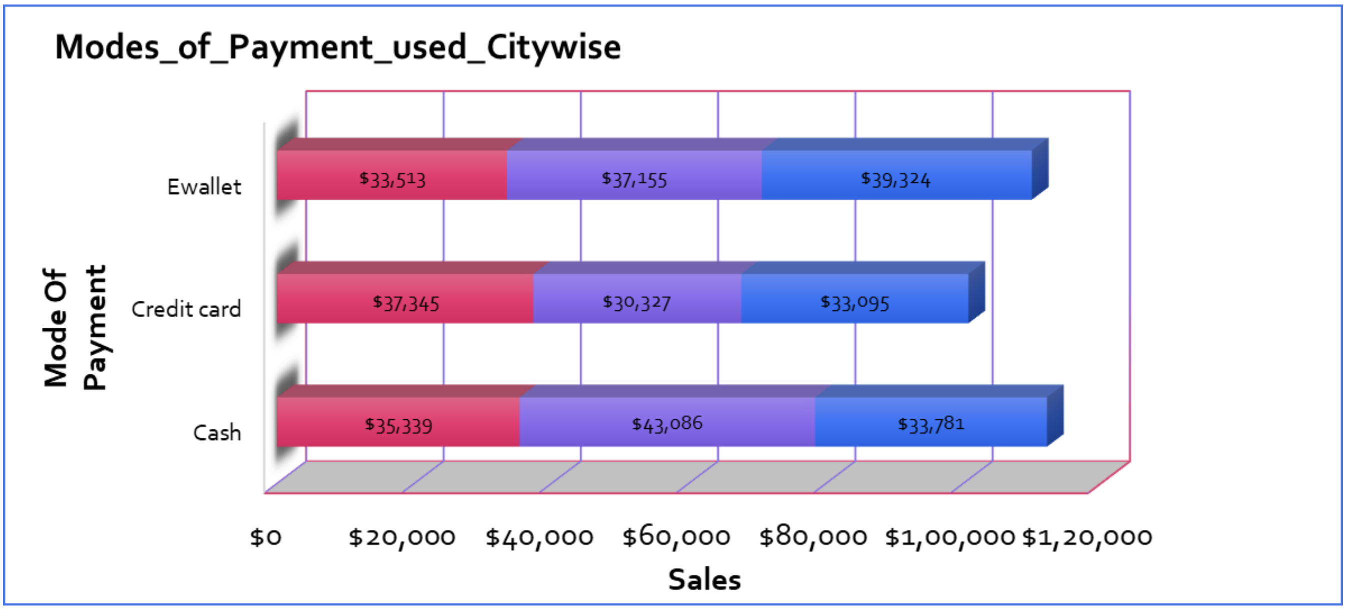
This analysis is used to find out which mode of payment is generally preferred by the customer on the basis of the branches. There are basically three modes of payment – E-wallet, cash and credit card.

Specific Requirements, functions and formulas:

- ✓ Pivot Table with Payment as rows, city as columns and total sales as values.
- ✓ Rename the total sales as ‘sales’
- ✓ Change the number format of sales up to zero decimal values.
- ✓ Converting the pivot table layout to tabular form.
- ✓ Use the pivot table to create 3-D stacked bar chart.

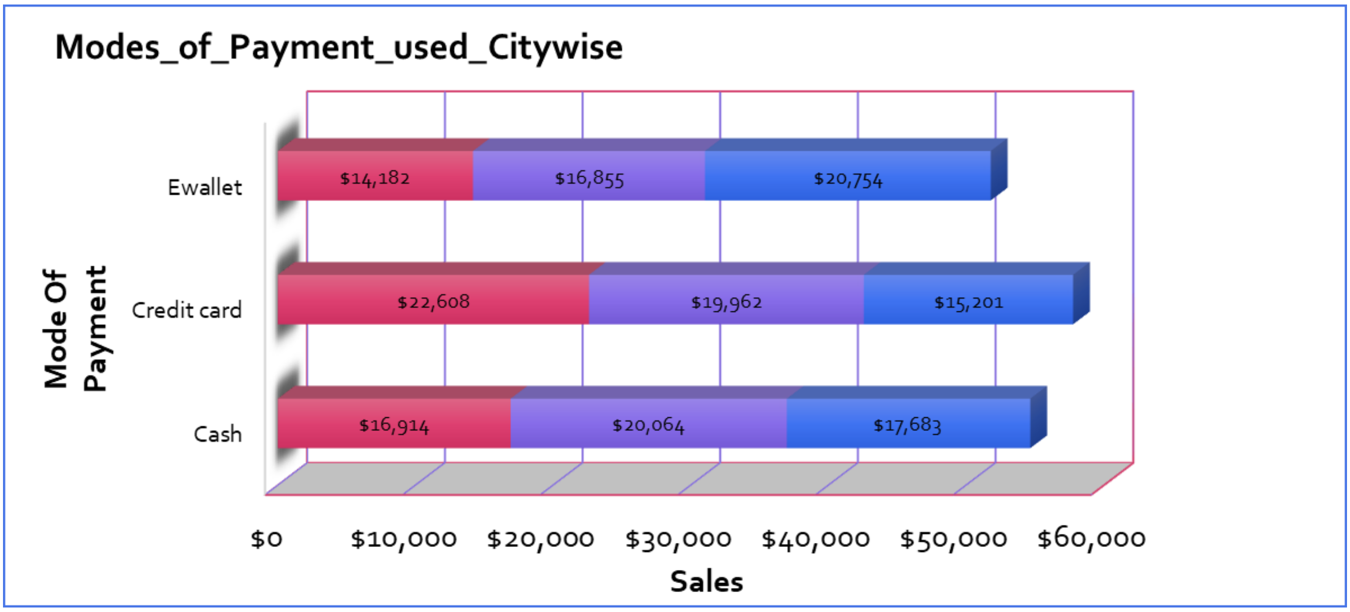
Analysis Results and Visualizations:

- ➢ The highest mode of payment used among all the three branches is Cash and following it is E-wallet.



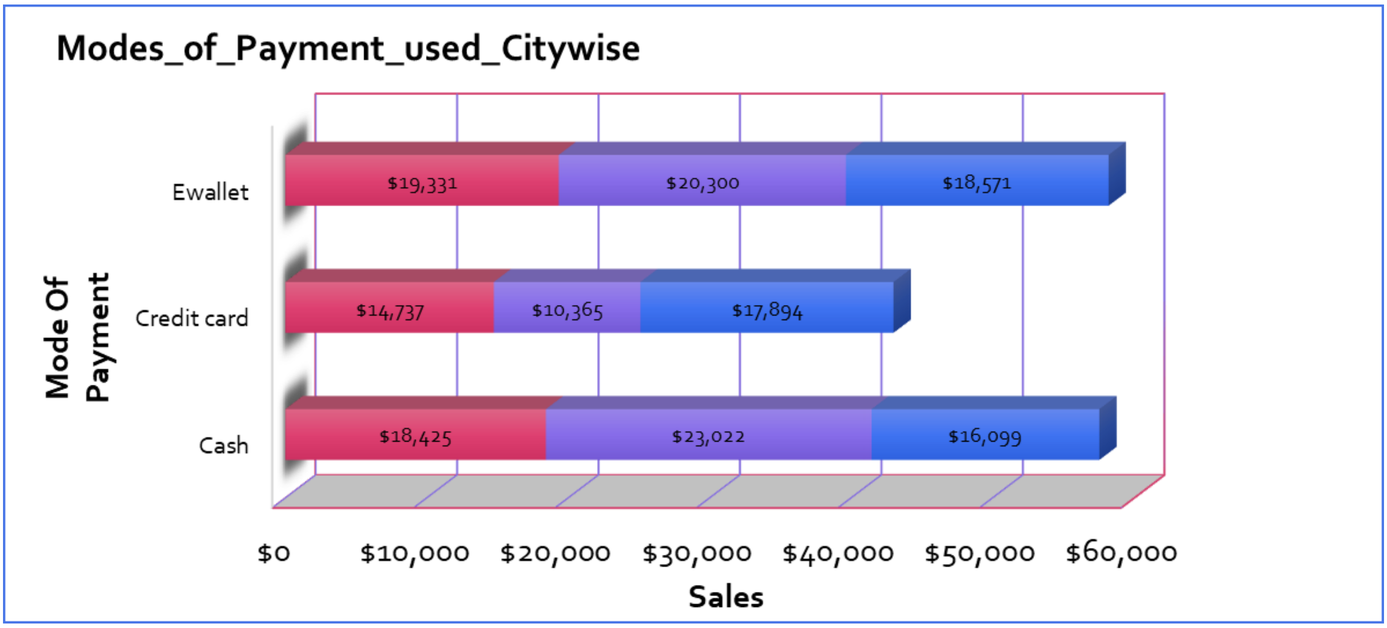
Yangon (A) 🟦 *Mandalay(B)* 🟥 *Naypitaw (C)* 🟪

- ➢ The members of the supermarket tend to use credit card rather than cash and E-wallet while the normal customers use credit card less to do payment.



Members use credit card rather than cash and E-wallet

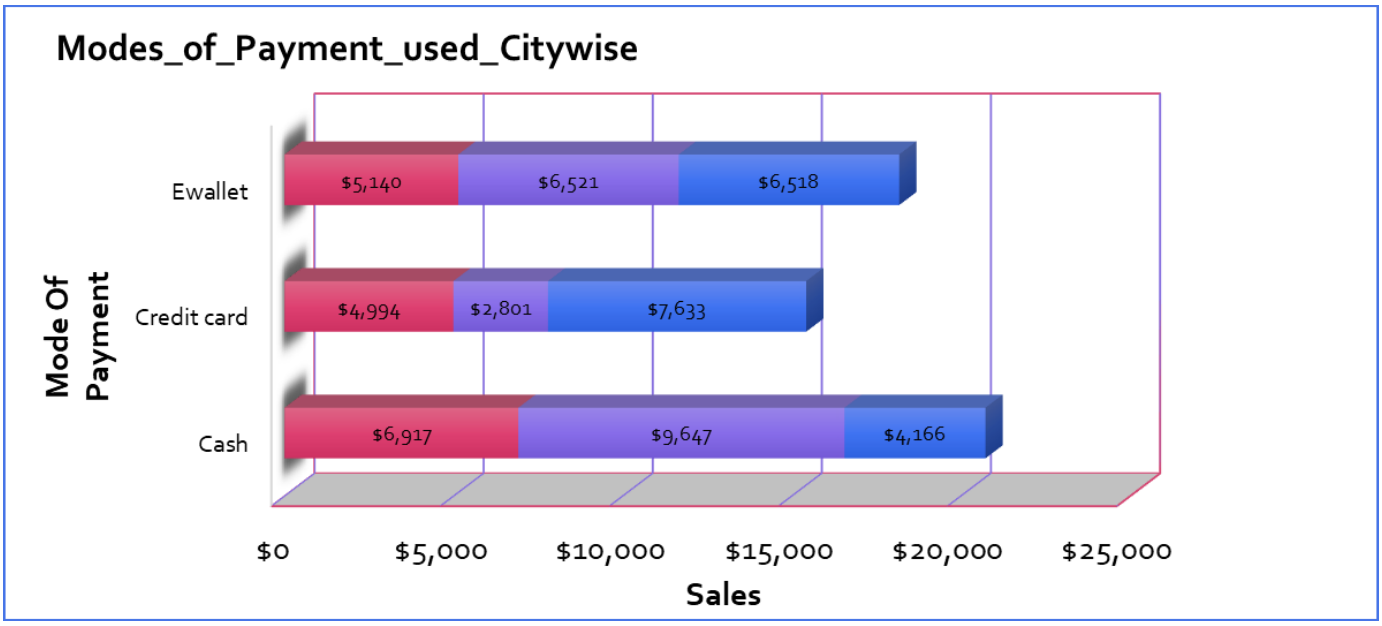
Yangon (A) 🟦 *Mandalay(B)* 🟥 *Naypitaw (C)* 🟪



Normal Customer use E-wallet and cash rather than card.

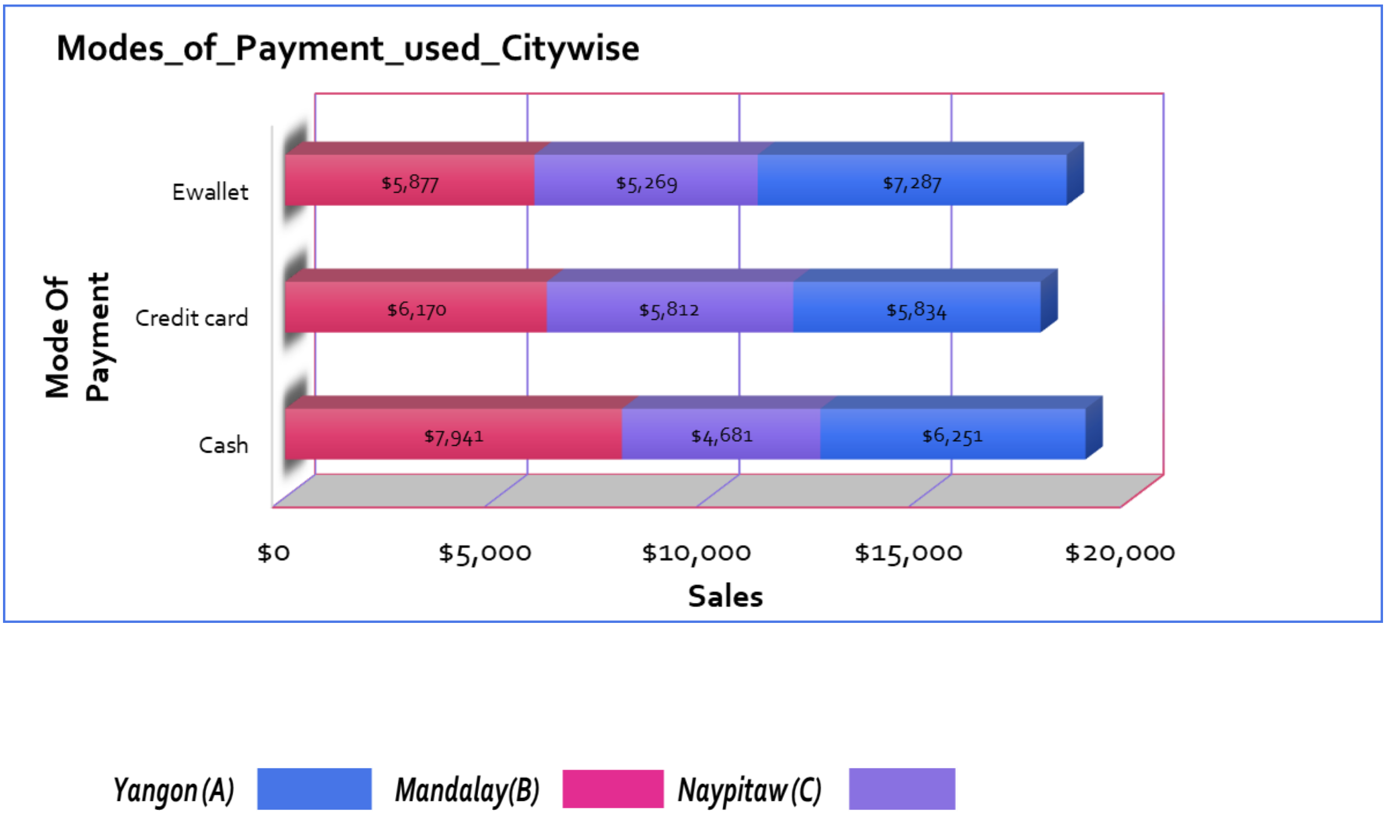
Yangon (A) 🟦 *Mandalay(B)* 🟥 *Naypitaw (C)* 🟪

- When customers are buying ‘electronic accessories’, they prefer paying cash.



Yangon (A) 🟦 *Mandalay(B)* 🟥 *Naypitaw (C)* 🟪

- There is very less margin in payment types used while buying ‘sports and travel’ products.



3. *Gross Income Analysis Branch Wise Based on Product Type*

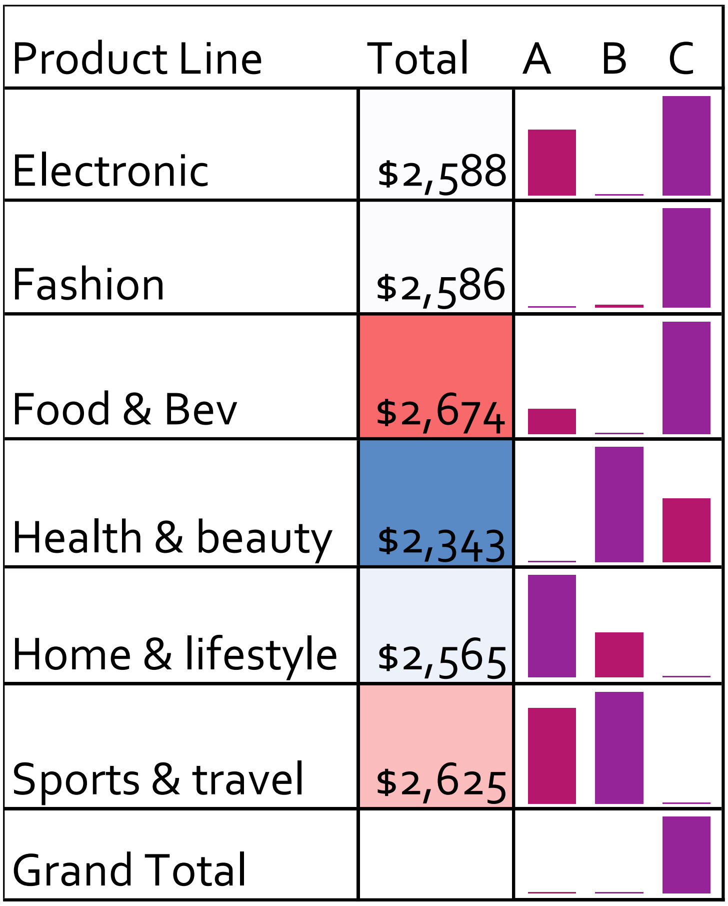
This analysis is used to know the gross income for each branch based on the product type. The gross income gives the overall income of these branches after deducting all the expenses and taxes.

Specific Requirements, functions and formulas:

- ✓ Pivot table using product line as row, city as columns and total gross income as values.
- ✓ Rename total gross income as gross income analysis.
- ✓ Change the number format up to zero decimals.
- ✓ Convert the layout of the table to Tabular.
- ✓ Use the formula “=IF(Gross_income!A5="","",Gross_income!A5)” in the dashboard so that it uses the pivot table for referencing the values.
- ✓ Use the value referencing the pivot table to create sparklines.

Analysis Results and Visualizations:

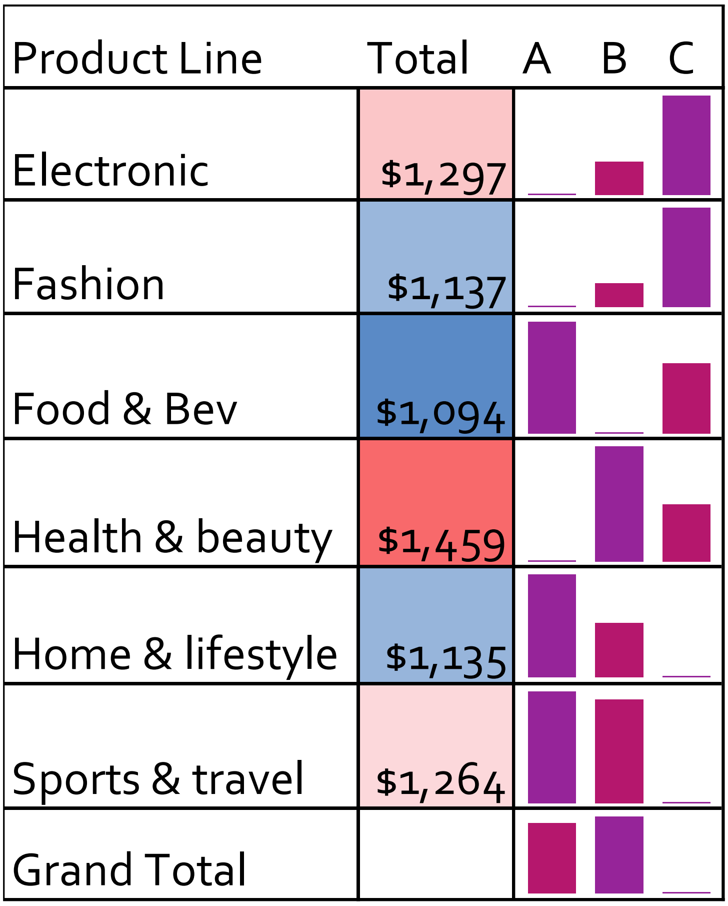
- Naypitaw has the highest gross income as compared to Mandalay and Yangon.



Product Line	Total	A	B	C
Electronic	$2,588			
Fashion	$2,586			
Food & Bev	$2,674			
Health & beauty	$2,343			
Home & lifestyle	$2,565			
Sports & travel	$2,625			
Grand Total				

Yangon (A) ***Mandalay(B)*** ***Naypitaw (C)***

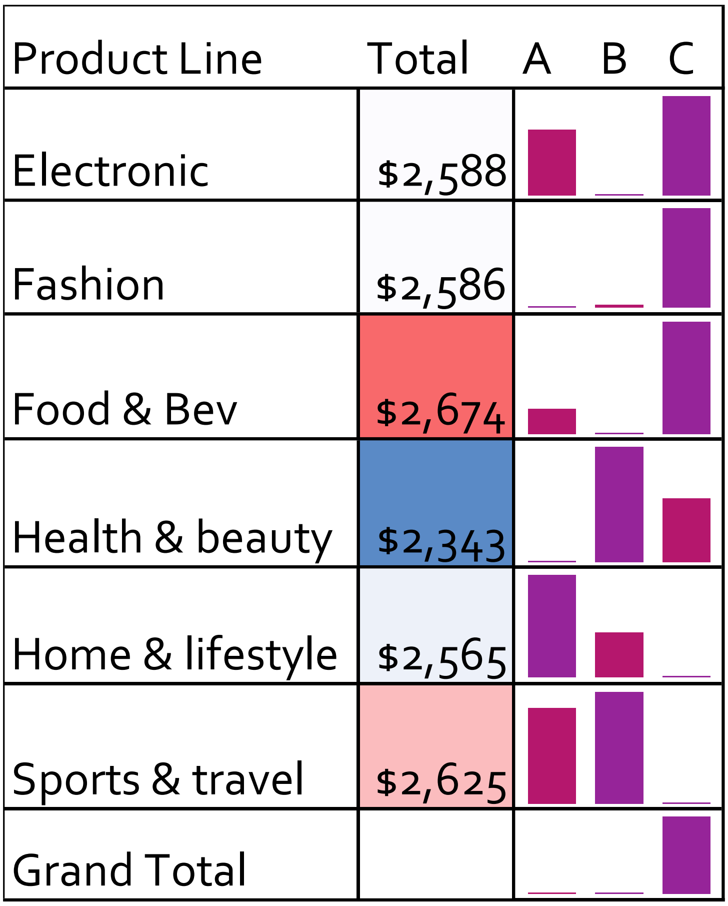
- The gross income from male customers is more in Yangon and Mandalay than Naypitaw.



Product Line	Total	A	B	C
Electronic	$1,297			
Fashion	$1,137			
Food & Bev	$1,094			
Health & beauty	$1,459			
Home & lifestyle	$1,135			
Sports & travel	$1,264			
Grand Total				

Yangon (A) ***Mandalay(B)*** ***Naypitaw (C)***

- Naypitaw has the highest gross income in ‘electronic accessories’, ‘fashion accessories’ and ‘food beverages’ while it has the lowest gross income in ‘home and lifetime’ and ‘sports and travel’.



Product Line	Total	A	B	C
Electronic	$2,588			
Fashion	$2,586			
Food & Bev	$2,674			
Health & beauty	$2,343			
Home & lifestyle	$2,565			
Sports & travel	$2,625			
Grand Total				

Yangon (A) ***Mandalay(B)*** ***Naypitaw (C)***

4. Analysis of Branch Wise Customer Rating

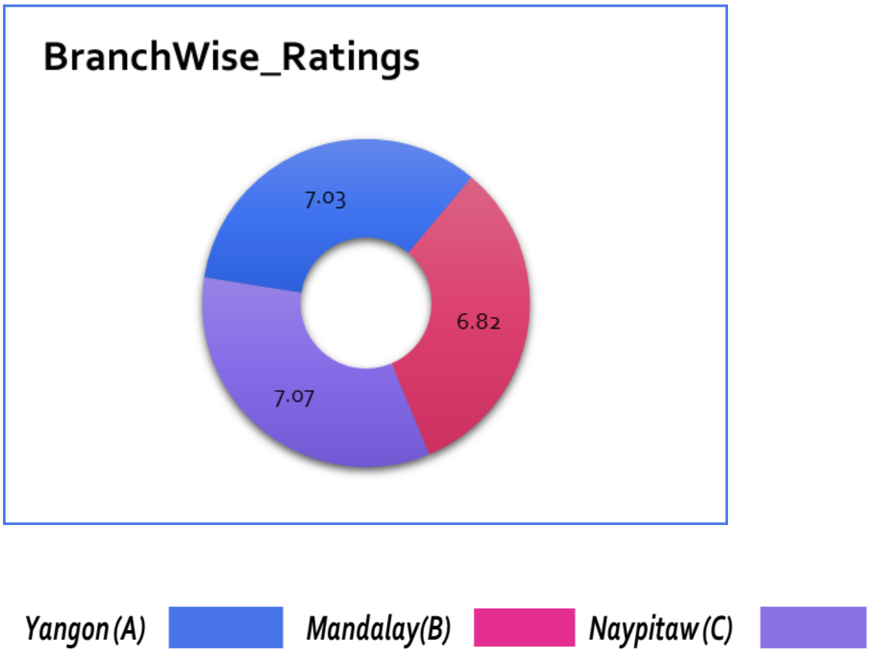
This analysis will help us know are the customers liking to visit the supermarket or not. It is done branch-wise so that we can know which branch gives a good customer experience.

<u>Specific Requirements, functions and formulas:</u>

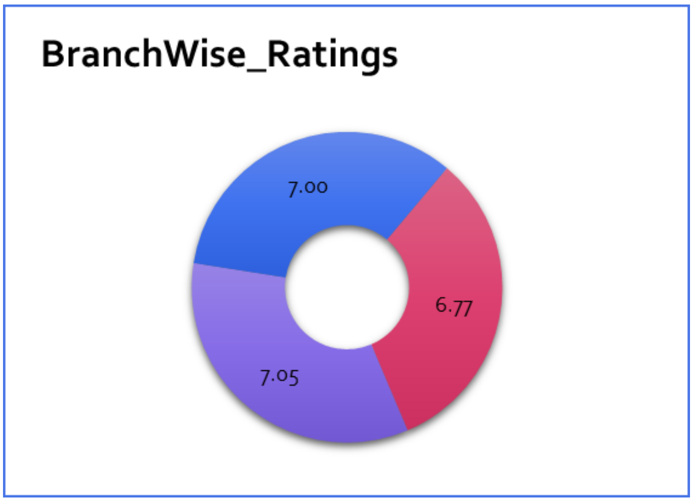
- ✓ Pivot table using city as rows and average of all ratings in values.
- ✓ The number format is set up to two decimal values.
- ✓ Doughnut Pie chart is used to visualize data.

<u>Analysis Results and Visualizations:</u>

- ➢ We saw that Naypitaw has the highest sales and gross income hence, the ratings are highest of the branch.

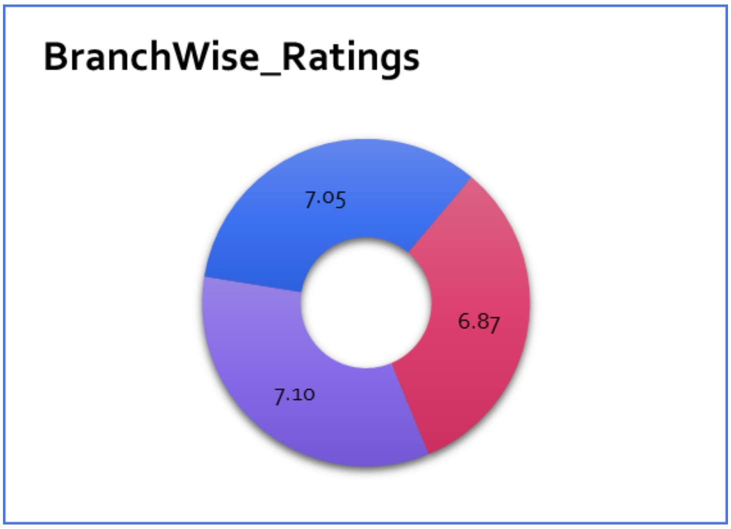


- The overall ratings in all branches given by members is less than given by the normal customers.



Ratings given By Member Customers.

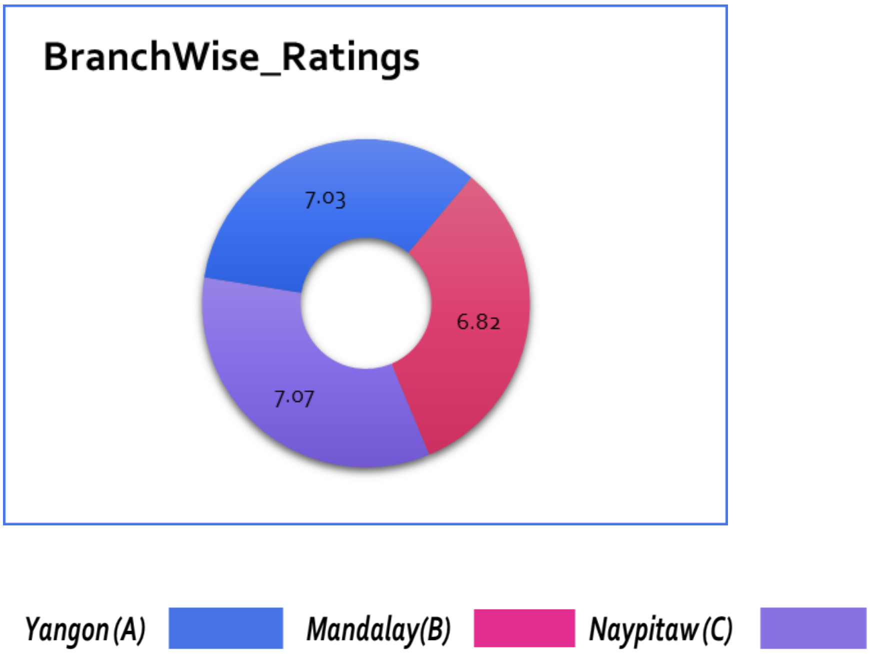
Yangon (A) 🟦 *Mandalay(B)* 🟥 *Naypitaw (C)* 🟪



Ratings given By Normal Customers.

Yangon (A) 🟦 *Mandalay(B)* 🟥 *Naypitaw (C)* 🟪

- There is not much difference between the ratings. They generally lie between 6.50 to 7.50. This clearly indicates that the customers generally give favorable ratings but they do not find the experience best.



5. *Hourly Analysis Based on Number of Customers*

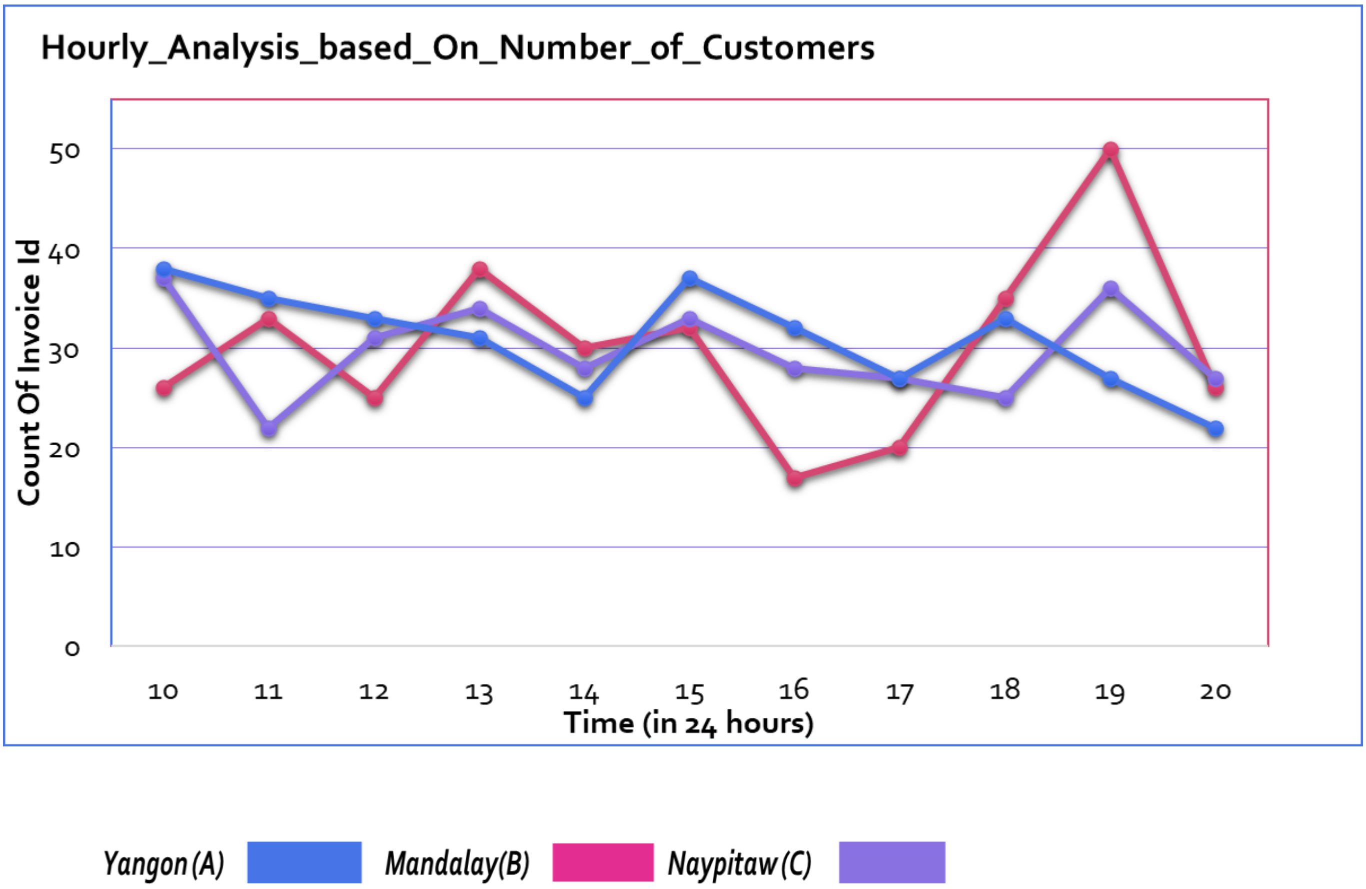
This analysis can be used to know when is the superstore crowded according to different branch. The time is in 24 hours format. The superstore is open from 10am to 8am.

Specific Requirements, functions and formulas:

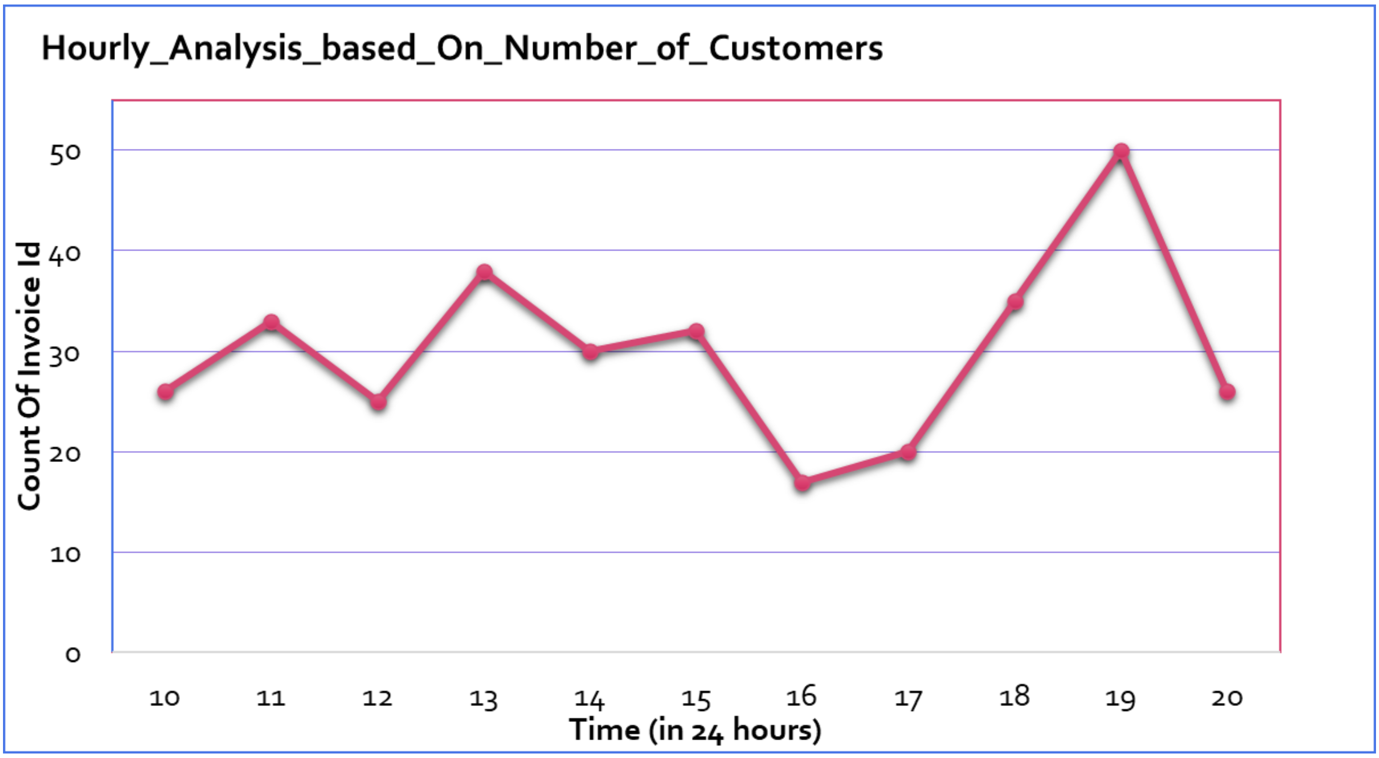
- ✓ Pivot Table using time as rows, count of Invoice ID as values and city as columns.
- ✓ Line chart for the pivot table.

Analysis Results and Visualizations:

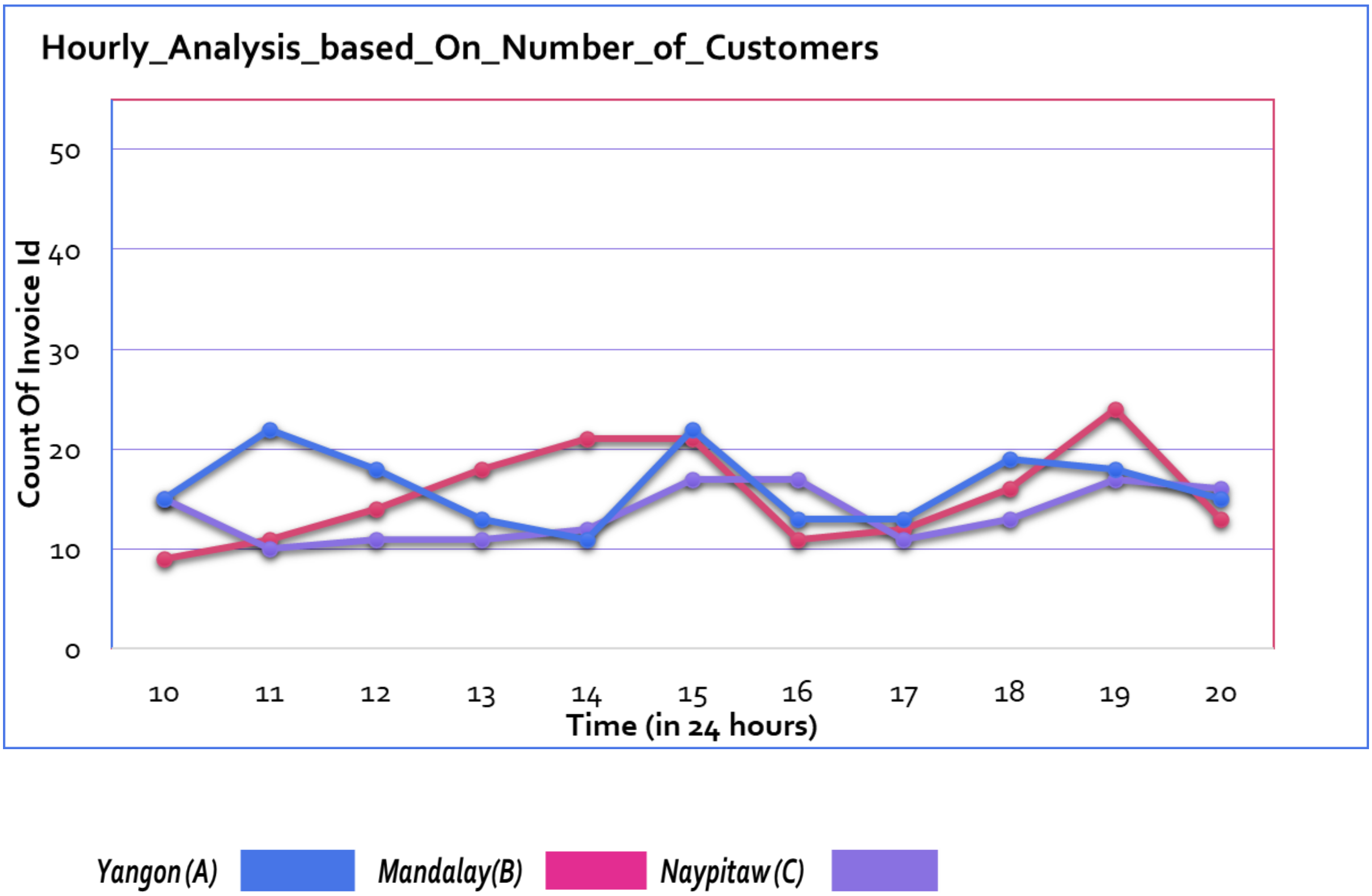
- The number of customers does not vary much between 12 pm to 3pm.



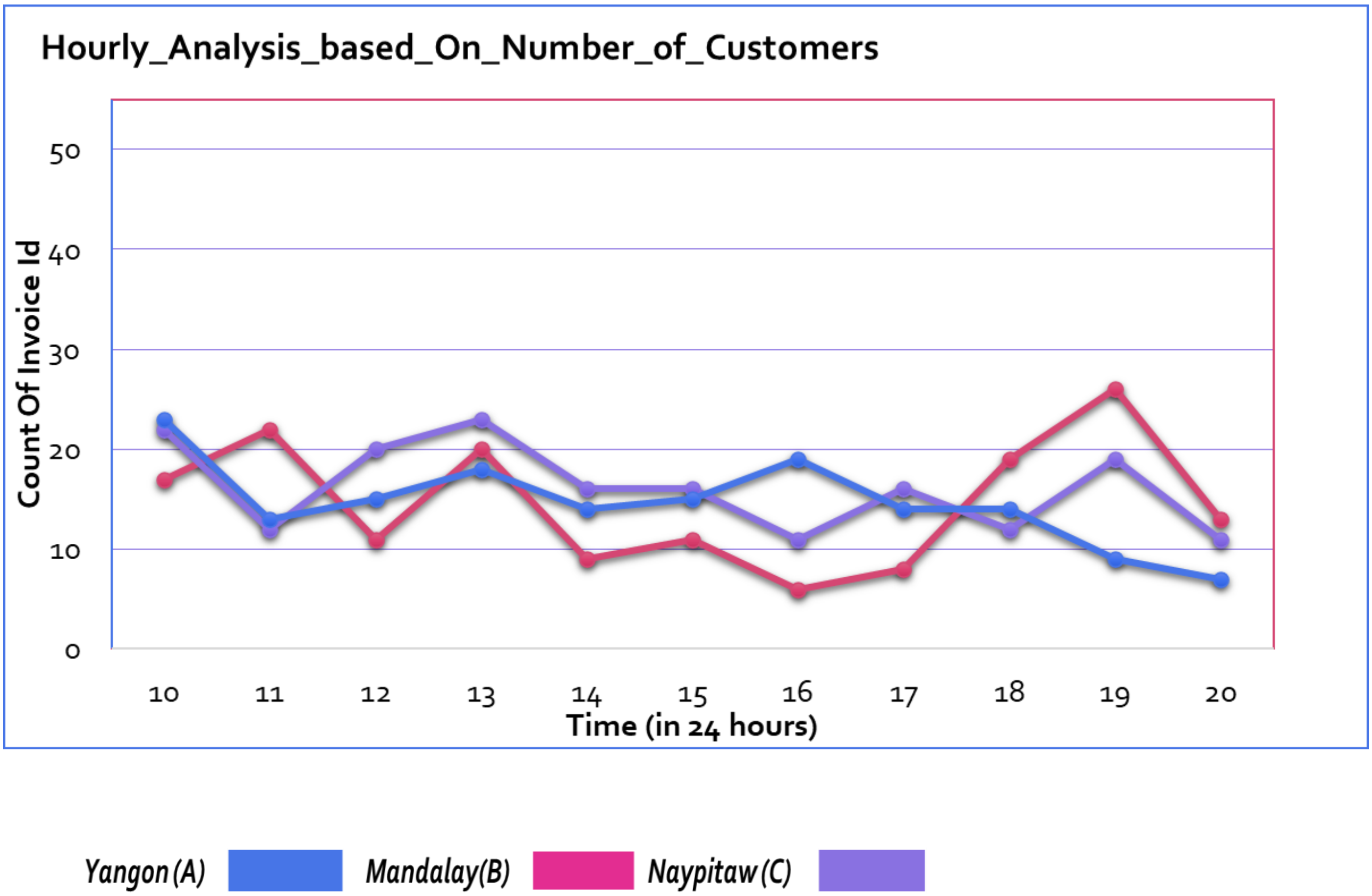
- The highest number of customers are at Mandalay i.e 50 at 7pm and below 20 at 4pm which is the lowest in all branches.



- The number of male customers remains almost constant at all times throughout the day between 10 to 20.

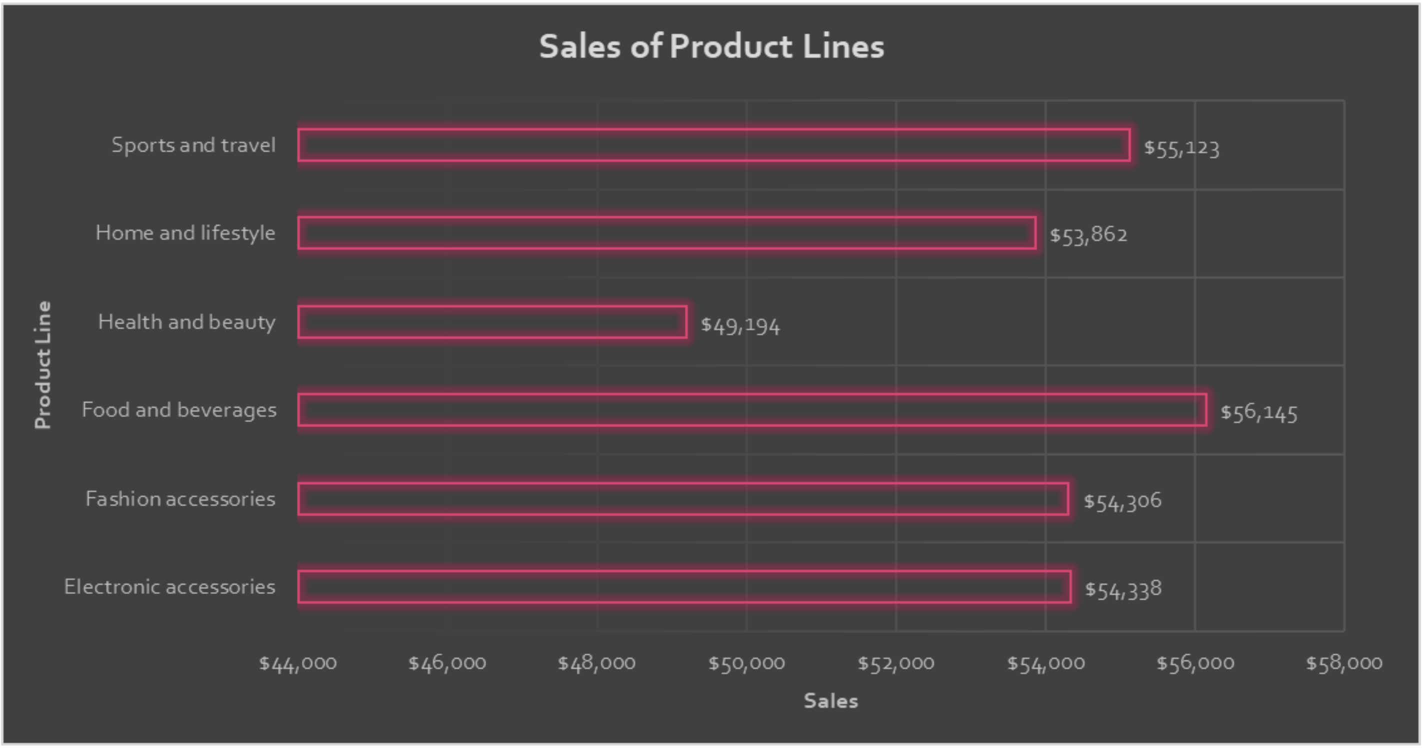


- The number of female customers vary greatly and there is no direct relation with time.

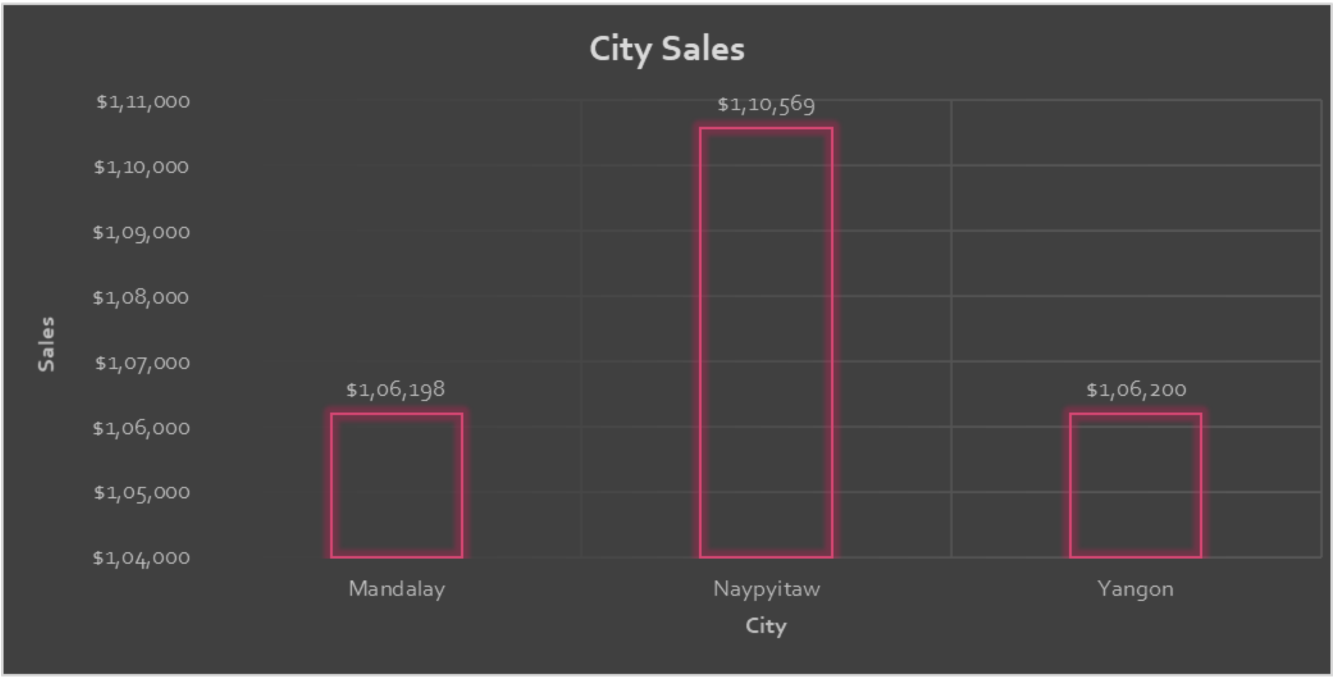


Overall Analysis:

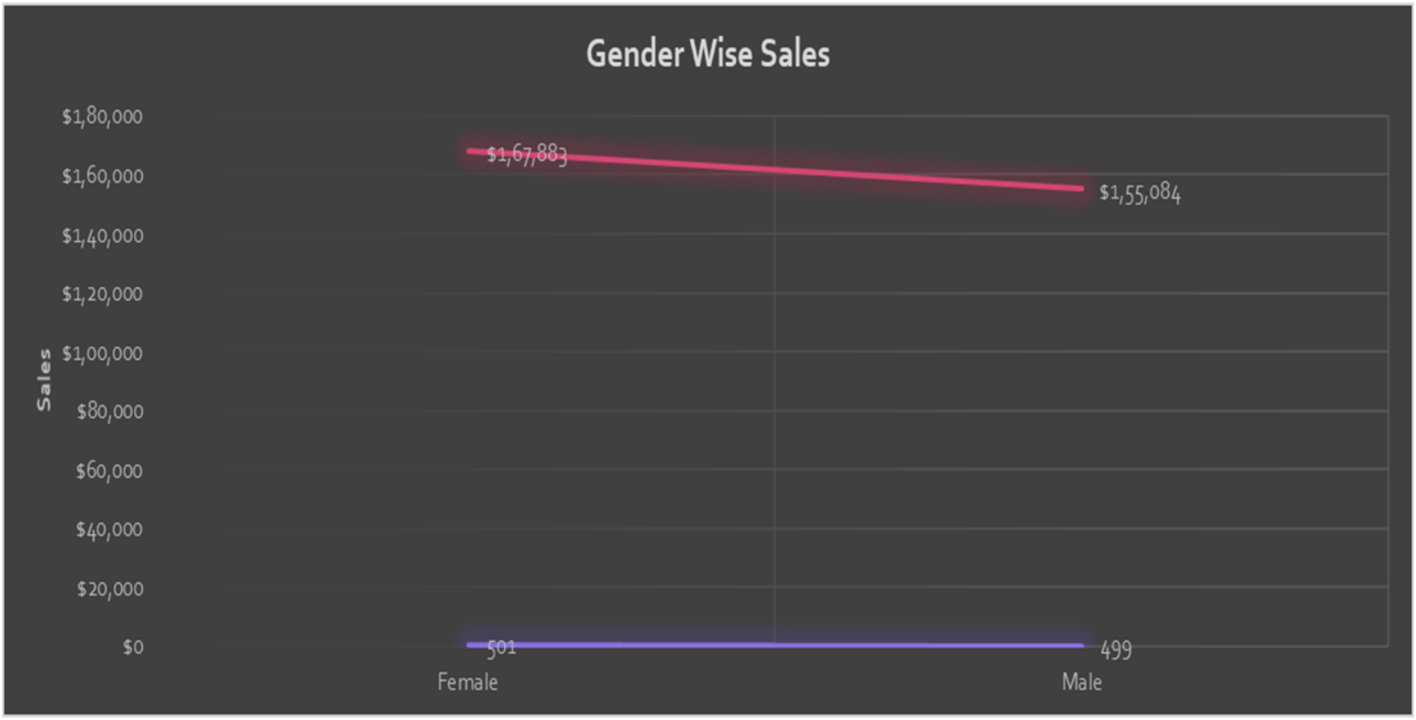
- The Product Line with highest sales is “Product & Beverages” while with the lowest sales is “Health & Beauty”



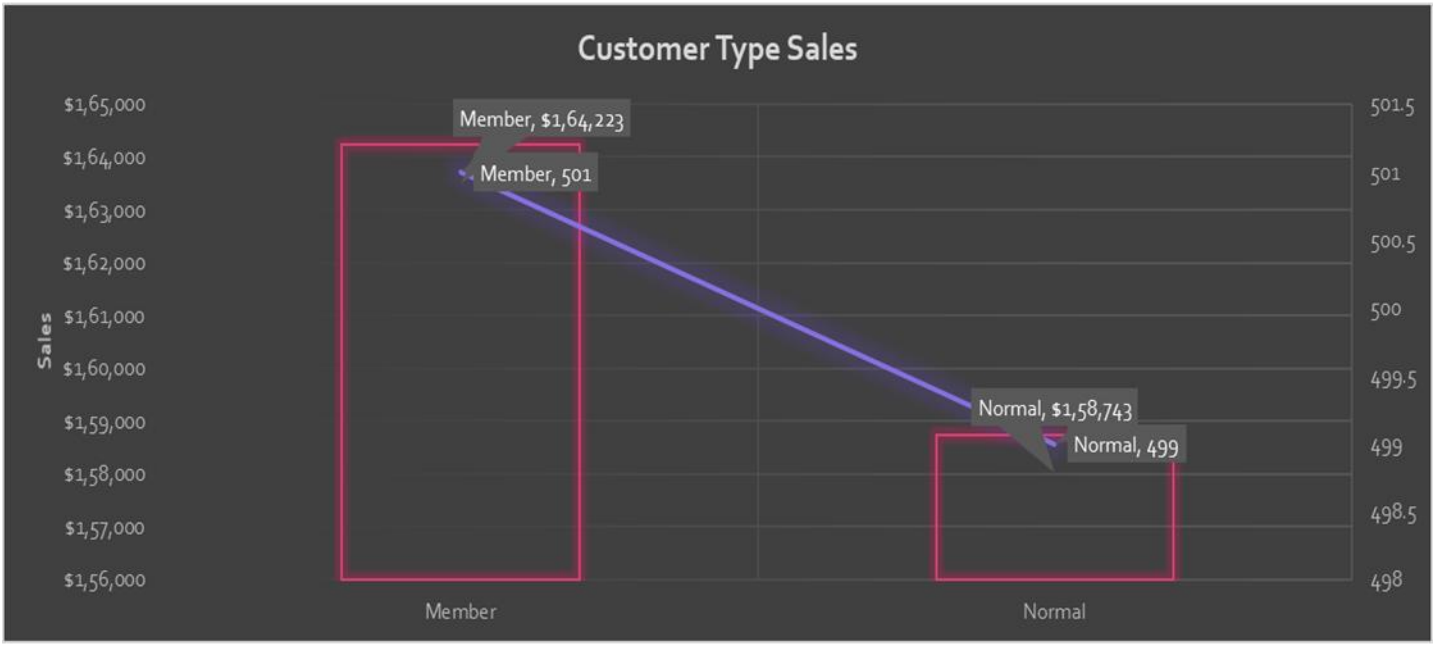
- The city with highest sales is ‘Naypitaw’ and the other two cities have approximately same amount of sales.



- The sales done by female customers is more as compared to male customers but there is only a slight difference between the number of customers.



- Member customer type does more shopping than Normal customers and there is a slight difference between their numbers.



LIST OF ANALYSIS WITH RESULTS

1. Product line ‘food and beverages’ have the highest sales in all product lines in the branch ‘Naypitaw’.
2. Mandalay has the maximum sales in two product lines of ‘healthy and beauty’ & ‘sports and travel’.
3. Naypitaw has the maximum sales in ‘food and beverages’ & next to it is ‘fashion accessories’.
4. Yangon has the maximum sales in ‘home and lifestyle’ & just after it ‘sports and travel’.
5. The highest mode of payment used among all the three branches is Cash and following it is E-wallet.
6. The members of the supermarket tend to use credit card rather than cash and E-wallet while the normal customers use credit card less to do payment.
7. When customers are buying electronic products, they prefer to pay in cash.
8. There is very less margin in payment types used while buying ‘sports and travel’ products.
9. Naypitaw has the highest gross income as compared to Mandalay and Yangon.
10. The gross income from male customers is more in Yangon and Mandalay than Naypitaw.
11. Naypitaw has the highest gross income in ‘electronic accessories’, ‘fashion accessories’ and ‘food beverages’ while it has the lowest gross income in ‘home and lifetime’ and ‘sports and travel’.
12. Naypitaw has the highest sales and gross income hence, the ratings are highest of the branch.
13. The overall ratings in all branches given by members is less than given by the normal customers.
14. There is not much difference between the ratings. They generally lie between 6.50 to 7.50. This clearly indicates that the customers generally give favorable ratings but they do not find the experience best.
15. The number of customers does not vary much between 12 pm to 3pm.
16. The highest number of customers are at Mandalay i.e 50 at 7pm and below 20 at 4pm which is the lowest in all branches.

17. The number of male customers remains almost constant at all times throughout the day between 10 to 20.
18. The number of female customers vary greatly and there is no direct relation with time.
19. The Product Line with highest sales is “Product & Beverages” while with the lowest sales is “Health & Beauty”.
20. The city with highest sales is ‘Naypitaw’ and the other two cities have approximately same amount of sales.
21. The sales done by female customers is more as compared to male customers but there is only a slight difference between the number of customers.
22. Member customer type does more shopping than Normal customers and there is a slight difference between their numbers.

REFERENCES

1. www.kaggle.com
2. www.youtube.com
3. www.google.com
4. www.stackoverflow.com
5. www.github.com

BIBLIOGRAPHY

- MICROSOFT EXCEL 2016 BIBLE: THE COMPREHENSIVE TUTORIAL RESOURCE by JOHN
- WALKENBACH, WILEY FUNDAMENTALS OF BUSINESS ANALYTICS by R.N. PRASAD, SEEMA ACHARYA, WILEY