SUPER MARKET SALES

Name : Thileti Sai Sruthi

Student Id : 11592257

Instructor : Mr Gahangir Hossain

Introduction:

A supermarket is a self-service business that provides a wide range of product lines, such as those for health and beauty, electronics, sports, meals, drinks, and home furnishings, all of which are divided into different sections. Supermarkets are expanding nowadays, and there is fierce market rivalry in most major urban areas. Supermarket Sales Analytics is used to track the sales.

Dataset:

The dataset in which i was working is collected from the kaggle website The dataset is about sales and done in three distinct branches, comprising of various metrics and dimensions. The characteristics in this dataset, which has 1000 rows, stand for total sales, quantity, customer type, payment method, and total revenue across three different branches.

Link:

Supermarket sales | Kaggle

Attributes:

There are of total 17 attributes.

1. **Invoice id:** Identification number for the sales

2. **Branch:** We have three branches in the dataset: A, B, and C. It shows the branch names of the super center.

3. **City:** provides details on the location of supermarkets.

4. **Customer type:** Customers are classified as either Members or Normal depending on whether they are using a normal card

5. **Gender:** describes the customer's gender (male or female).

6. **Product line:** There are six distinct product categories, including technology accessories, apparel accessories, food and drink, health and beauty, home and leisure, and travel and sports.

7. **Unit price:** Price of each item in US dollars.

8. **Quantity:** It gives information about how many things did the consumer buy?

9. **Tax 5%:** tax of 5% for customers purchasing.

10. **Total:** It gives details about overall product sales.

11. **Gross income:** Sum of all the incomes.

Tools used:

Python

It allows the development of programs using an object-oriented approach. It offers many high-level data structures and is straightforward and simple to learn.

Data analysis and visualization regularly make use of the well-known programming language Python. It includes several libraries with a focus on data analysis, such as NumPy, Pandas, and Matplotlib. These modules work together to forge a strong environment for Python data visualization and analysis.

Tableau

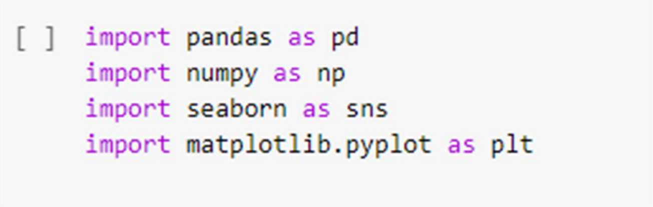
A business intelligence application called Tableau enables us to evaluate data visually using things like graphs and reports. You may use tableau as a very effective tool to analyze this data in the form of various graphs and reports. From several data sources, users may use Tableau to create dashboards, reports, and interactive and dynamic representations. Users may simply connect to data sources, create charts, maps, and tables, and modify the visualizations to meet their needs thanks to its drag-and-drop interface.

Data Cleaning:

The dataset contains no missing values, therefore I intend to continue on with the data visualizations.

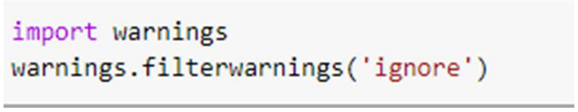
EDA:

First we Import the required libraries.



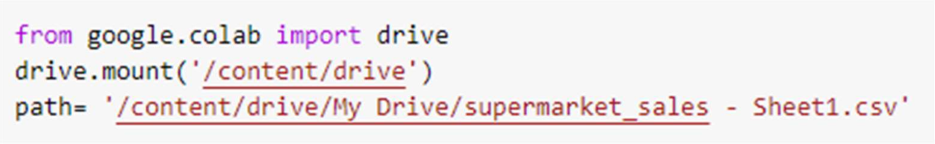
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Second, we import warnings so that all warnings can be ignored.

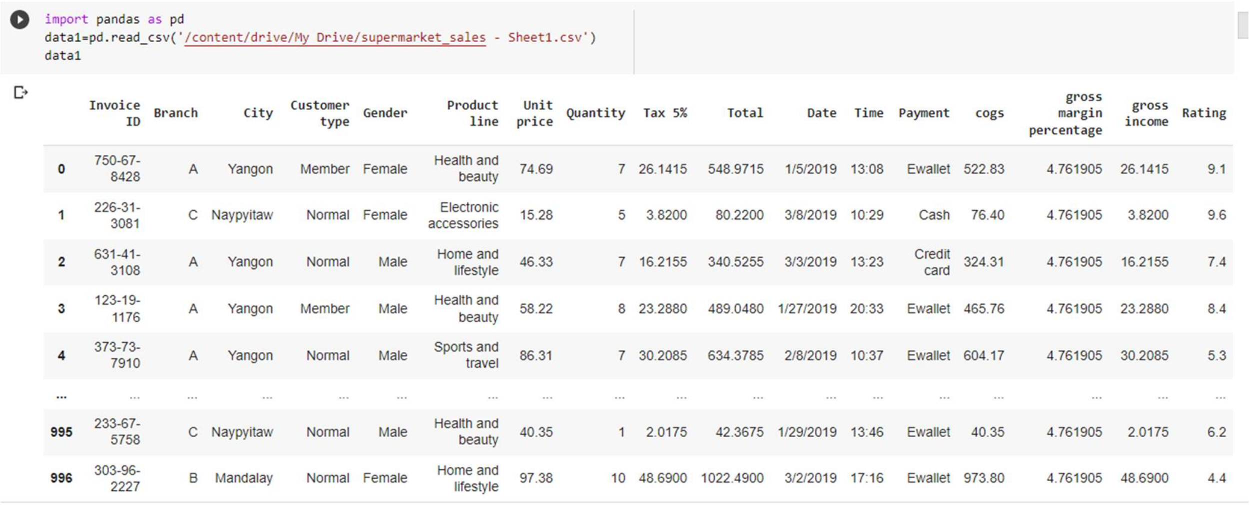


```
import warnings
warnings.filterwarnings('ignore')
```

Then loading the data into csv file



```
from google.colab import drive
drive.mount('/content/drive')
path= '/content/drive/My Drive/supermarket_sales - Sheet1.csv'
```

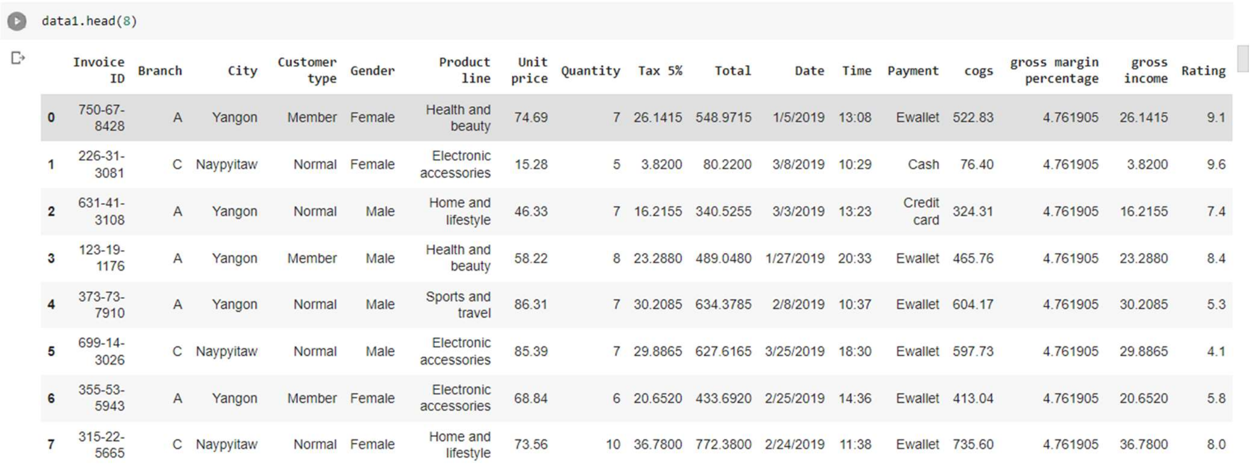


```
import pandas as pd
data1=pd.read_csv('/content/drive/My Drive/supermarket_sales - Sheet1.csv')
data1
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	233-67-5758	C	Naypyitaw	Normal	Male	Health and beauty	40.35	1	2.0175	42.3675	1/29/2019	13:46	Ewallet	40.35	4.761905	2.0175	6.2
996	303-96-2227	B	Mandalay	Normal	Female	Home and lifestyle	97.38	10	48.6900	1022.4900	3/2/2019	17:16	Ewallet	973.80	4.761905	48.6900	4.4

Data1.head(8)

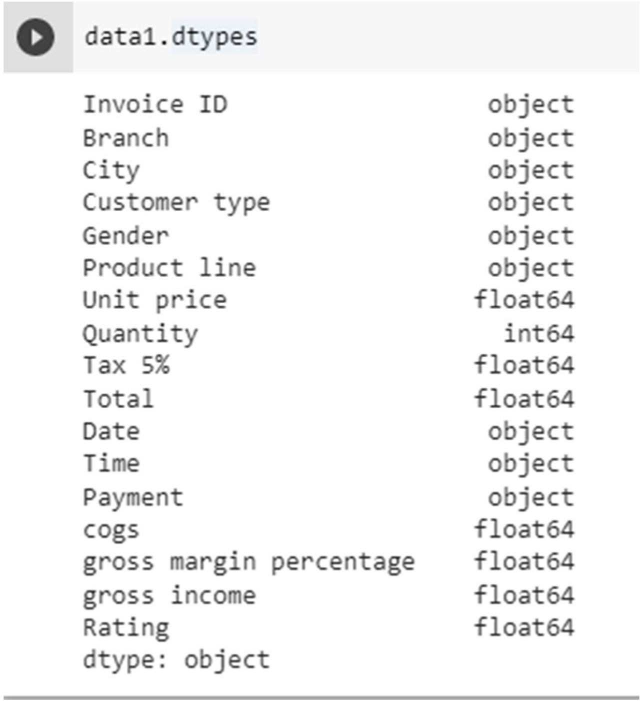
Gives information of first eight rows of the data.



```
data1.head(8)
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
5	699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	3/25/2019	18:30	Ewallet	597.73	4.761905	29.8865	4.1
6	355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.6520	433.6920	2/25/2019	14:36	Ewallet	413.04	4.761905	20.6520	5.8
7	315-22-5665	C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.7800	772.3800	2/24/2019	11:38	Ewallet	735.60	4.761905	36.7800	8.0

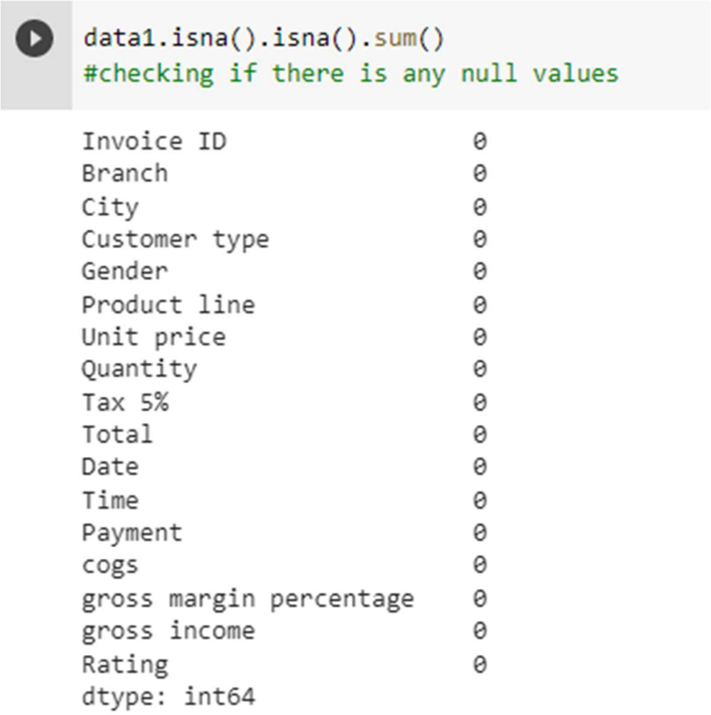
Information about each attribute data type.



```
data1.dtypes
```

```
Invoice ID                  object
Branch                      object
City                        object
Customer type               object
Gender                      object
Product line                object
Unit price                 float64
Quantity                     int64
Tax 5%                     float64
Total                      float64
Date                        object
Time                        object
Payment                     object
cogs                       float64
gross margin percentage    float64
gross income               float64
Rating                     float64
dtype: object
```

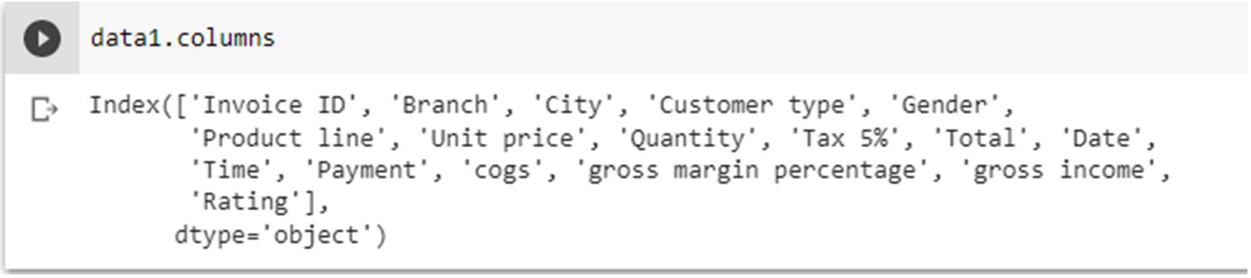
Checking if the data has any null values and from below picture, we can say that the dataset has no missing values.



```
data1.isna().isna().sum()
#checking if there is any null values
```

```
Invoice ID                 0
Branch                     0
City                       0
Customer type              0
Gender                     0
Product line               0
Unit price                 0
Quantity                   0
Tax 5%                     0
Total                      0
Date                       0
Time                       0
Payment                    0
cogs                       0
gross margin percentage    0
gross income               0
Rating                     0
dtype: int64
```

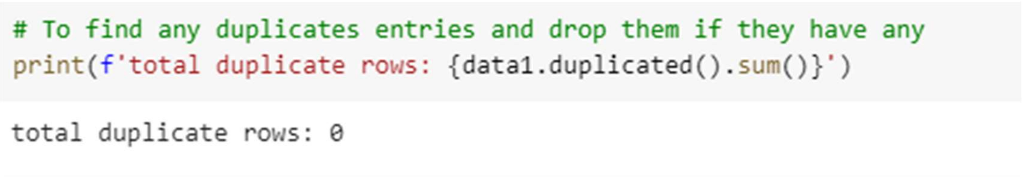
To get information about columns



```
data1.columns
```

```
Index(['Invoice ID', 'Branch', 'City', 'Customer type', 'Gender',
       'Product line', 'Unit price', 'Quantity', 'Tax 5%', 'Total', 'Date',
       'Time', 'Payment', 'cogs', 'gross margin percentage', 'gross income',
       'Rating'],
      dtype='object')
```

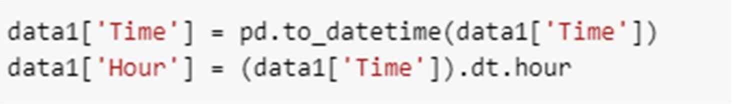
To find if there is any duplicate values.



```
# To find any duplicates entries and drop them if they have any
print(f'total duplicate rows: {data1.duplicated().sum()}')
```

```
total duplicate rows: 0
```

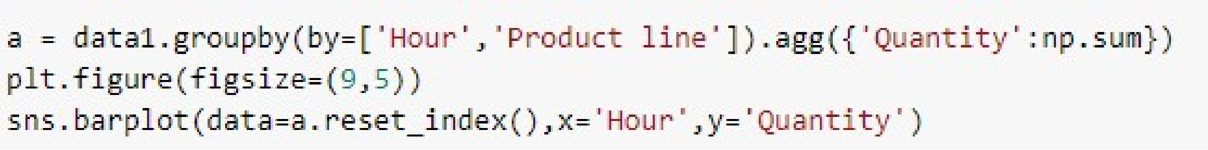
Examining it revealed that the time column type is an object, thus I'm switching it to Datetime datatype.



```
data1['Time'] = pd.to_datetime(data1['Time'])
data1['Hour'] = (data1['Time']).dt.hour
```

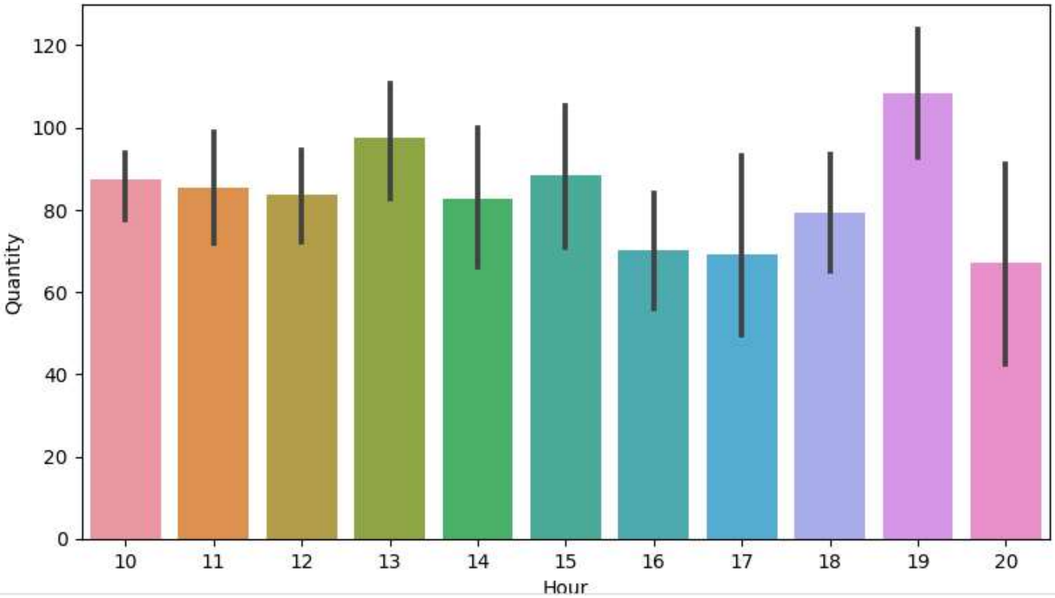
Visualizations.

At which hour the product sales are more?

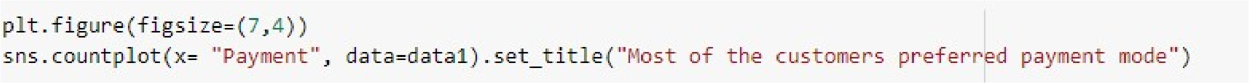


```
a = data1.groupby(by=['Hour','Product line']).agg({'Quantity':np.sum})
plt.figure(figsize=(9,5))
sns.barplot(data=a.reset_index(),x='Hour',y='Quantity')
```

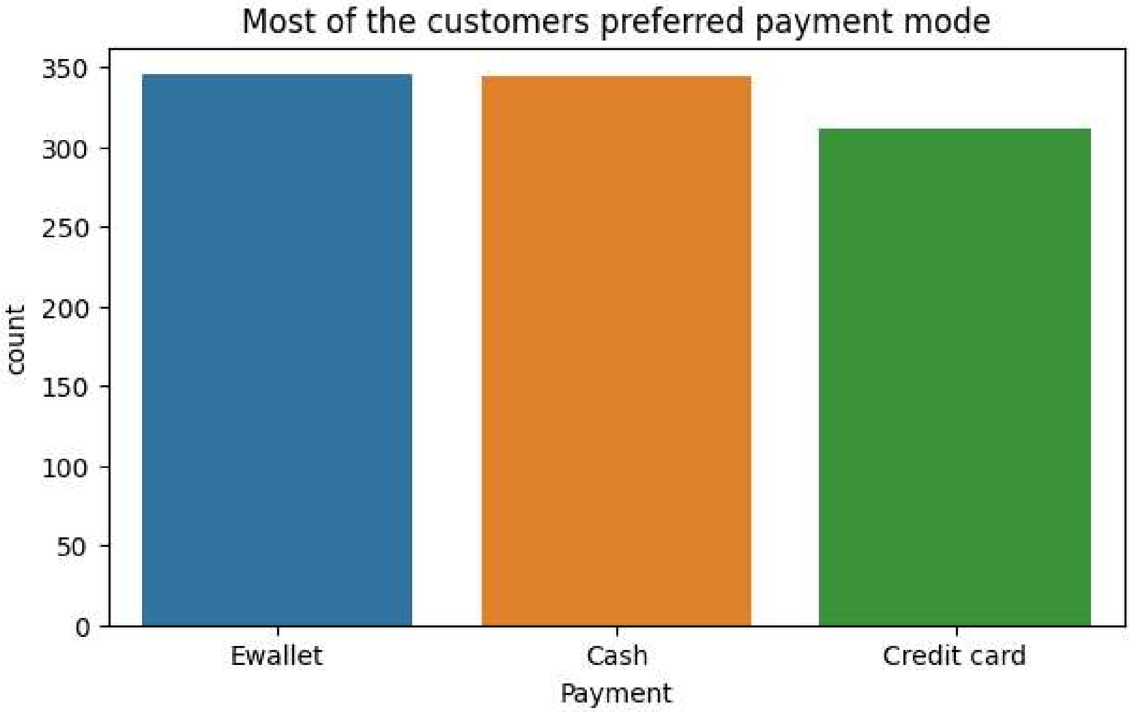
We can say that more things are being sold around 7:00/19:00 pm based on the results of the visualization that is displayed below.



More Payments are done in which Payment Mode?



```
plt.figure(figsize=(7,4))
sns.countplot(x= "Payment", data=data1).set_title("Most of the customers preferred payment mode")
```

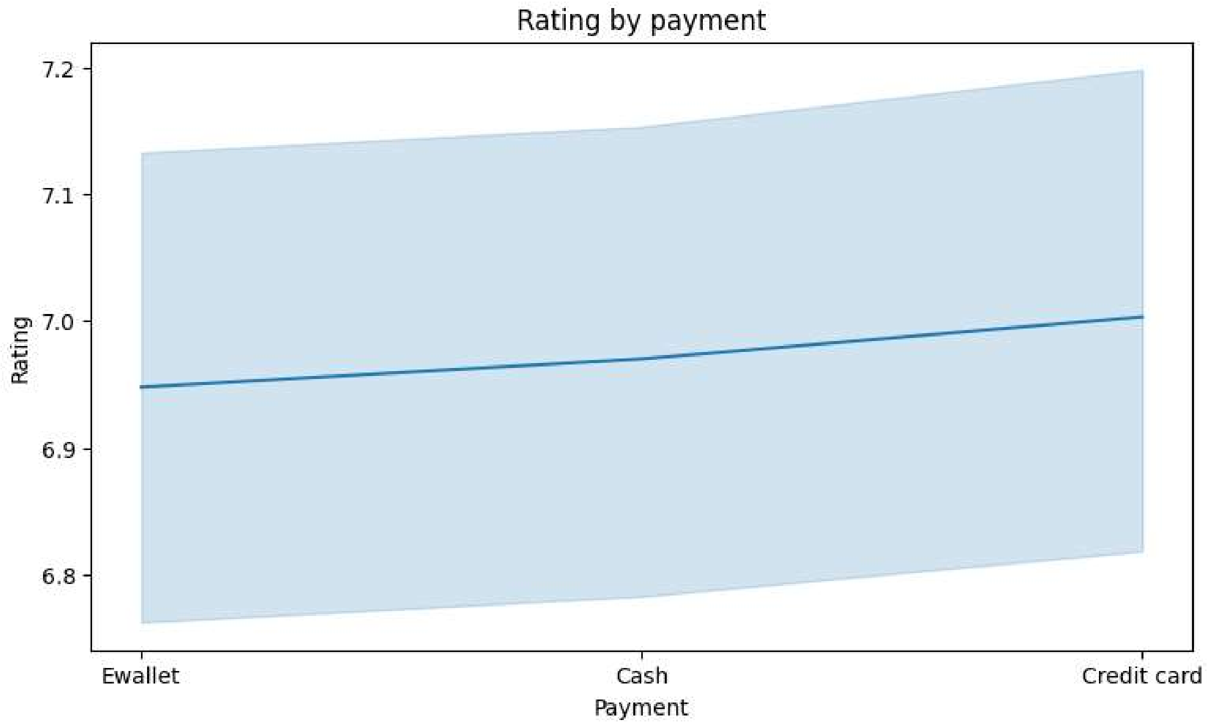


Result:

It is evident from the graphic representation above that cash is the preferred method of payment, followed by eWallets and credit cards.

Which payment mode as More Rating? and how much is the Highest Average rating?

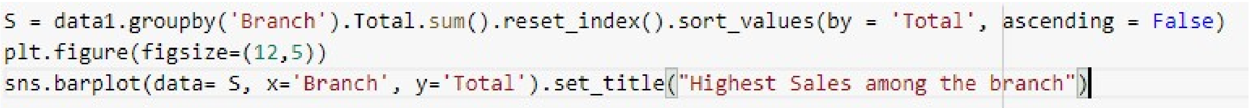
```
plt.figure(figsize=(9,5))
sns.lineplot(x="Payment",y="Rating" , data=data1).set_title("Rating by payment")
```



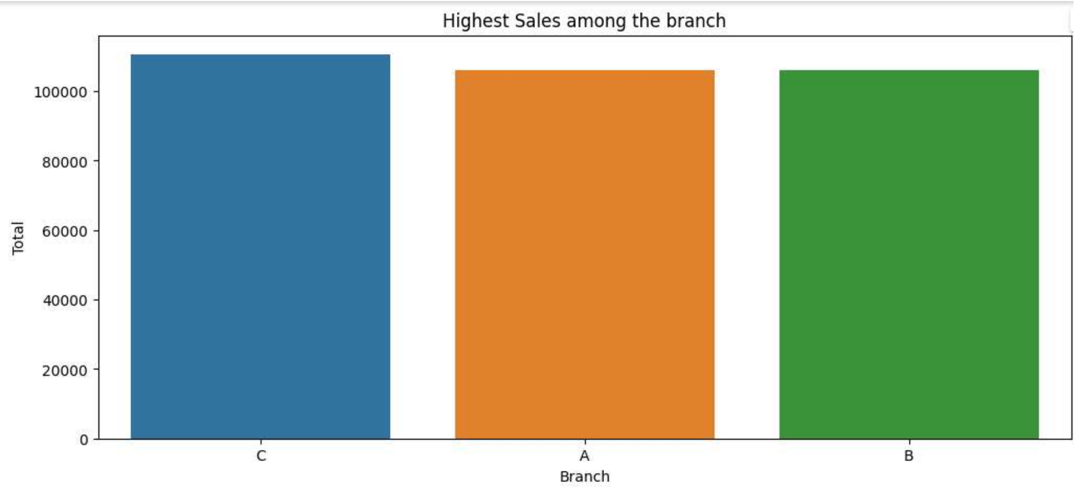
Result:

From the Above representation, we can see that credit card payment has more rating.

Which Branch has the highest and lowest Total Sales?



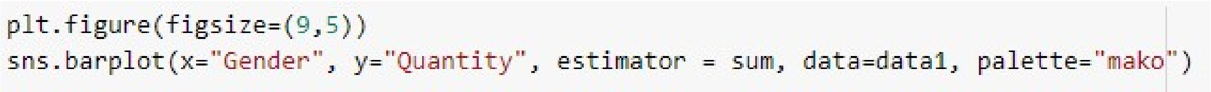
```
S = data1.groupby('Branch').Total.sum().reset_index().sort_values(by = 'Total', ascending = False)
plt.figure(figsize=(12,5))
sns.barplot(data= S, x='Branch', y='Total').set_title("Highest Sales among the branch")
```



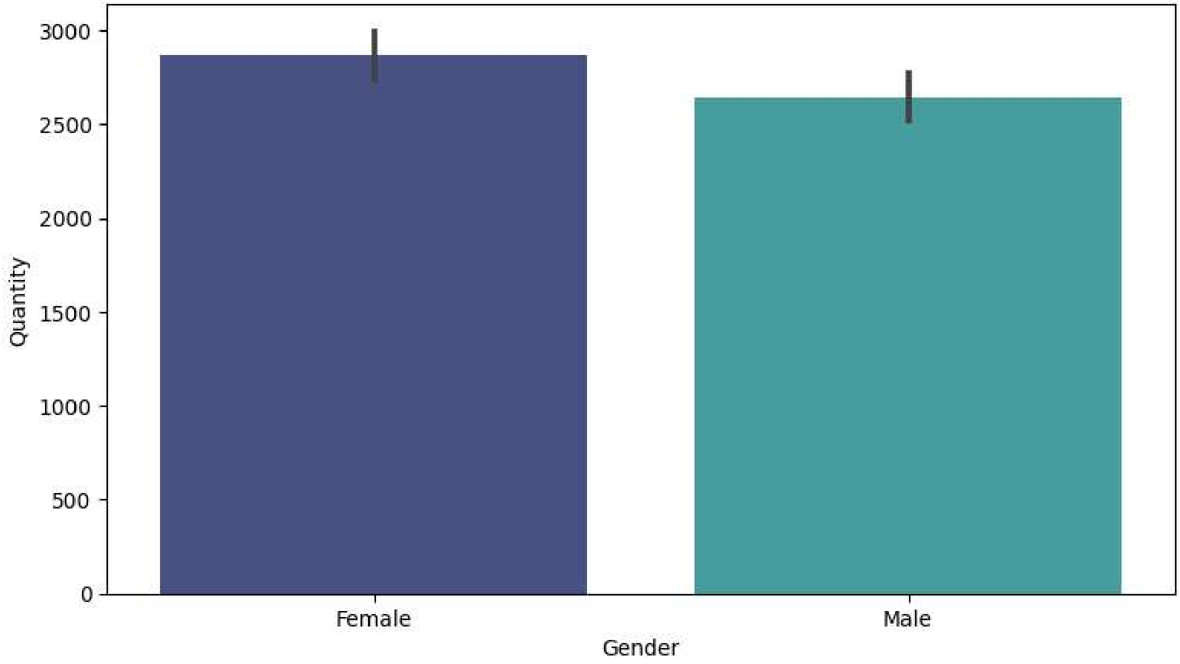
Result:

From the Above visualization, which is sorted in the descending order we can tell that branch c has the highest sales followed by A and B.

Which Type of gender purchased more goods?



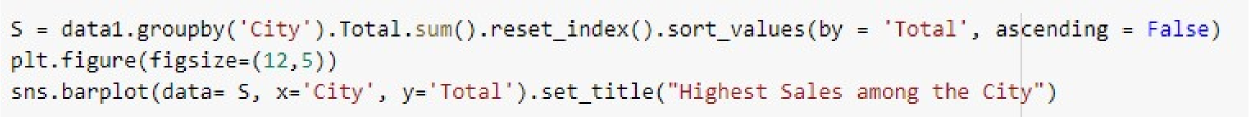
```
plt.figure(figsize=(9,5))
sns.barplot(x="Gender", y="Quantity", estimator = sum, data=data1, palette="mako")
```



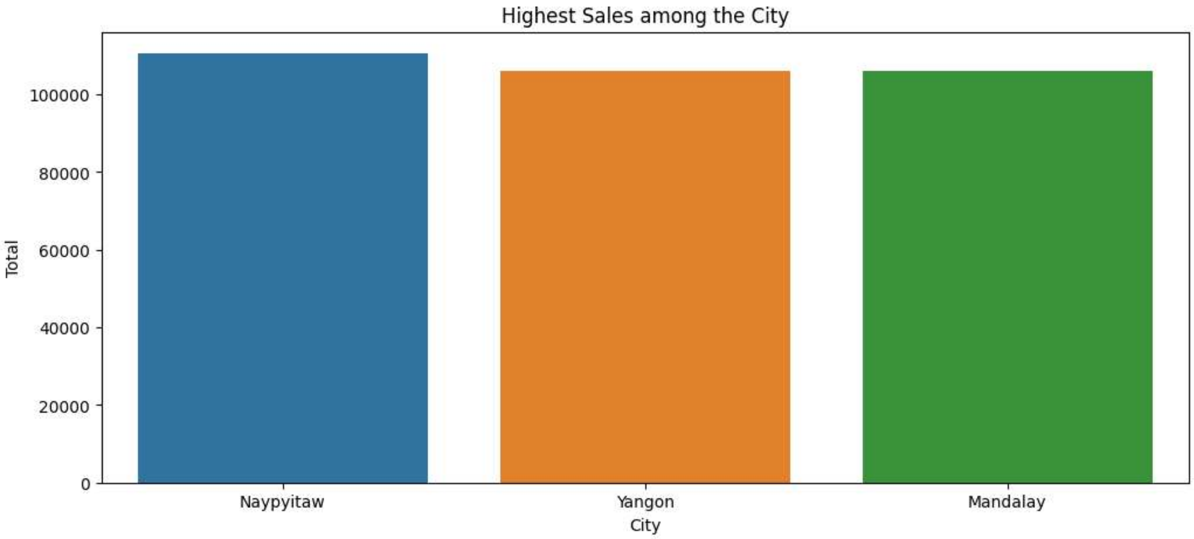
Result:

From the above visualization we can tell those customers with gender type female bought most of the products in comparison with customers with gender type male.

Which City which spends more?



```
S = data1.groupby('City').Total.sum().reset_index().sort_values(by = 'Total', ascending = False)
plt.figure(figsize=(12,5))
sns.barplot(data= S, x='City', y='Total').set_title("Highest Sales among the City")
```



Result:

From the above visualization Naypyitaw city spends more.

HYPOTHESIS

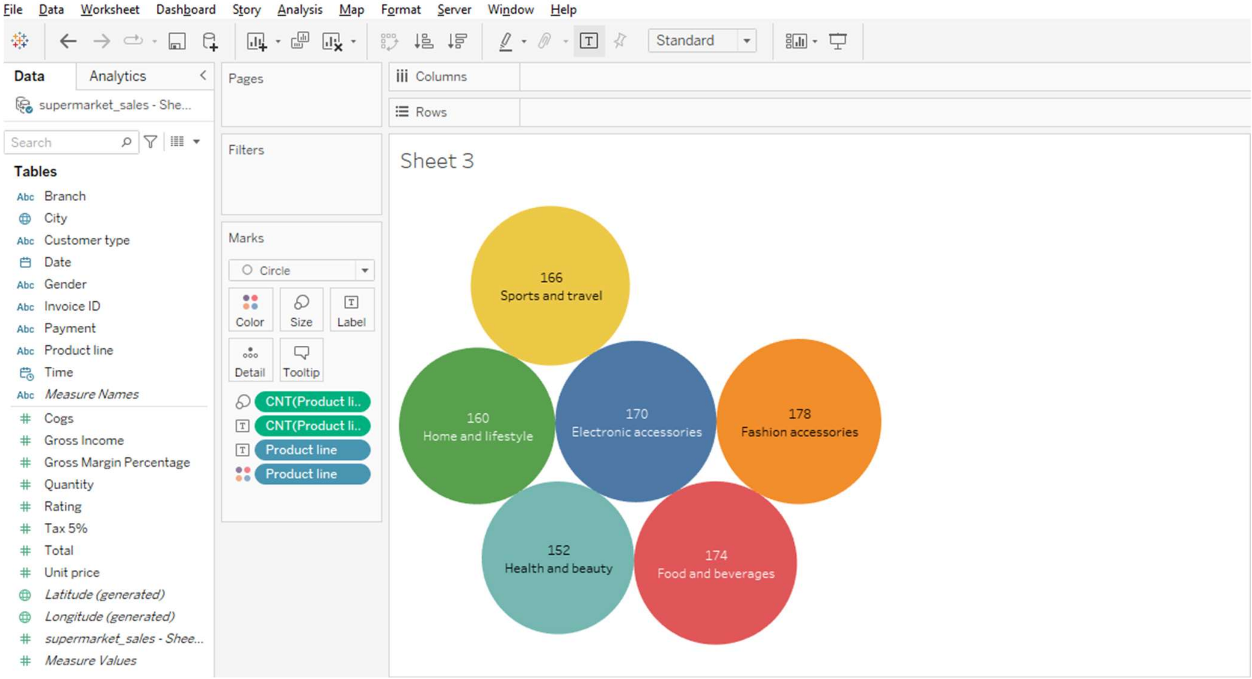
1. Which product line most and which is least bought by customers?
2. Identify the branch where most electronics and accessory products are sold, along with the time of day.
3. Which city has a higher proportion of female shoppers? Additionally, what product categories do men and women tend to buy more of? What their spending was?
4. Analyze whether the price and ratings of the product line affecting the sales?
5. Which product line most and which is least bought by customers? Which product line produces more tax and which is least.

1. Which product line most and which is least bought by customers?

For the following question I am using packed bubbles

Packed bubbles

To present data as a collection of circles, use packed bubble charts. Measures specify the size and color of the individual circles, whereas dimensions specify the individual bubbles.



The above Visualization is done by dragging product line in columns and quantity in rows. And keeping quantity in size . By clicking on quantity and selecting count from the measure option and dragging measure option to labels we can get the above bubble chart.

Results

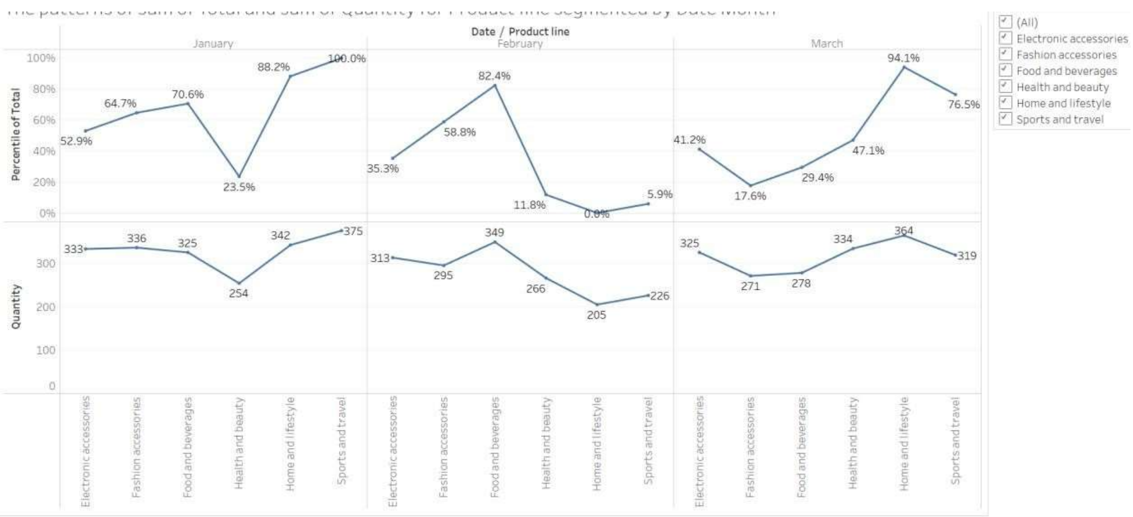
We can say that sports and travel has more quantity which has value 166 and the least is health and beauty (152) when compared with home and lifestyle ,electronic accessories, fashion accessories, food and beverages.

2. Analyze which payment mode is most used in Branch C? Furthermore, Identify, in which branch, at what hour, most electronics and accessories products are being sold?

I have used line and area graph to analyze the data.

Line graph:

The line graph below shows patterns of total sales and quantities of the products lines over a three-month period.



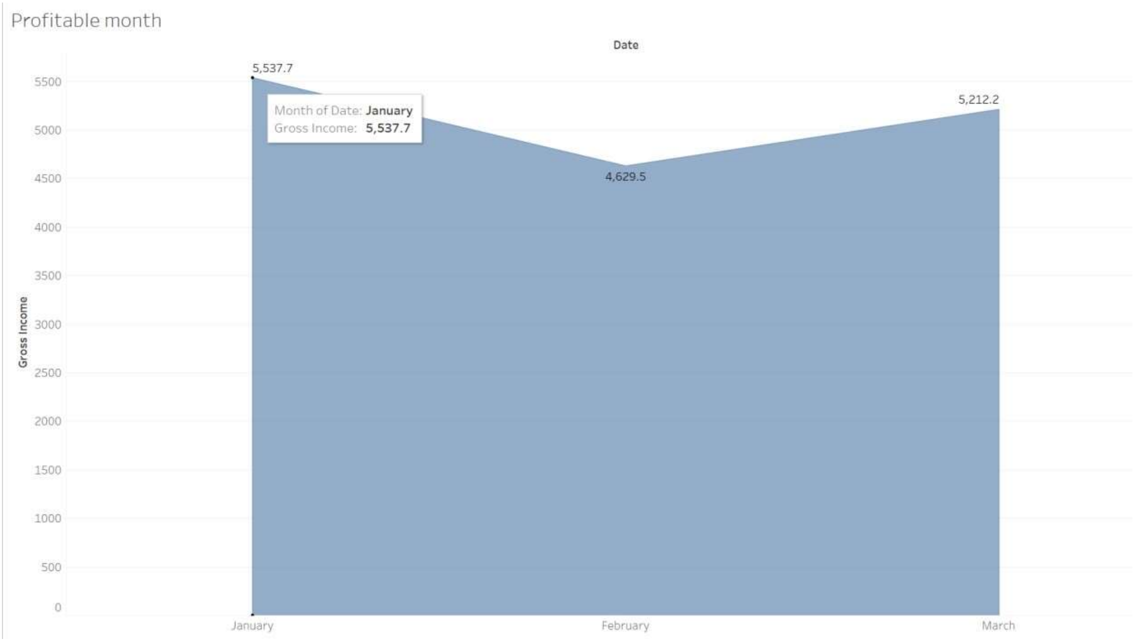
To create the visualization seen above, first change the date characteristics' object-based datatype to datetime. I then used the date attribute to placed in the month, quarter, and year-specific column. I added the Product line to the column, and I added total and quantity to the rows. By clicking on the quick table calculation and choosing percentages, I changed total to percentages. In order to allow users to check according to their preferences, I also included a product line to the filter. Finally, I added the total in the label to the marks.

Result:

The above visualization shows the overall sales of several product lines over a three-month period in percentages and volumes. According to the visualization, in January, February, and March, respectively, food and beverages (82.4%), sports and travel (100%), and home and lifestyles (92.1%) had the greatest overall sales percentages. While the lowest overall sales percentages in January, February, and March were in the health and beauty, home and lifestyle, and fashion accessories categories.

Area Graph:

Profitable month:



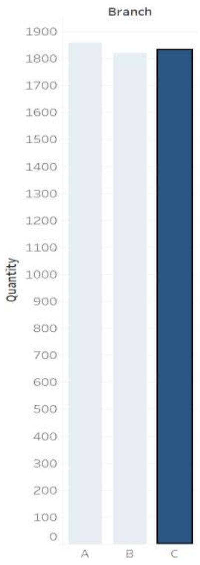
By adding month to the columns and gross income to the rows, this image is produced. Additionally, I changed the word "Automatic" to "Area" to alter the manner in which marks are represented. included gross income to the labels, as well.

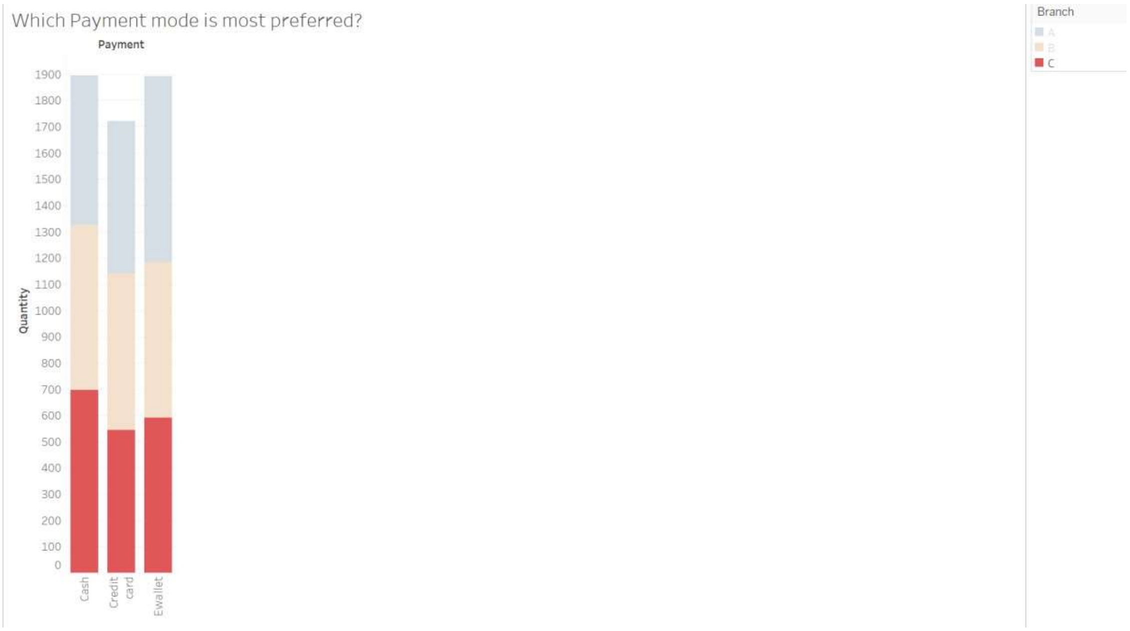
Result:

Which month is more profitable is seen in the visualization above.
It is obvious from the graphic portrayal that January was a more profitable month.

3 .Analyze which city has more female shopping? Also, inspect on what product line male and female shops more? How much they spent?

To Analyze the data, I am using interactive feature and scatter plot. Interactive feature:



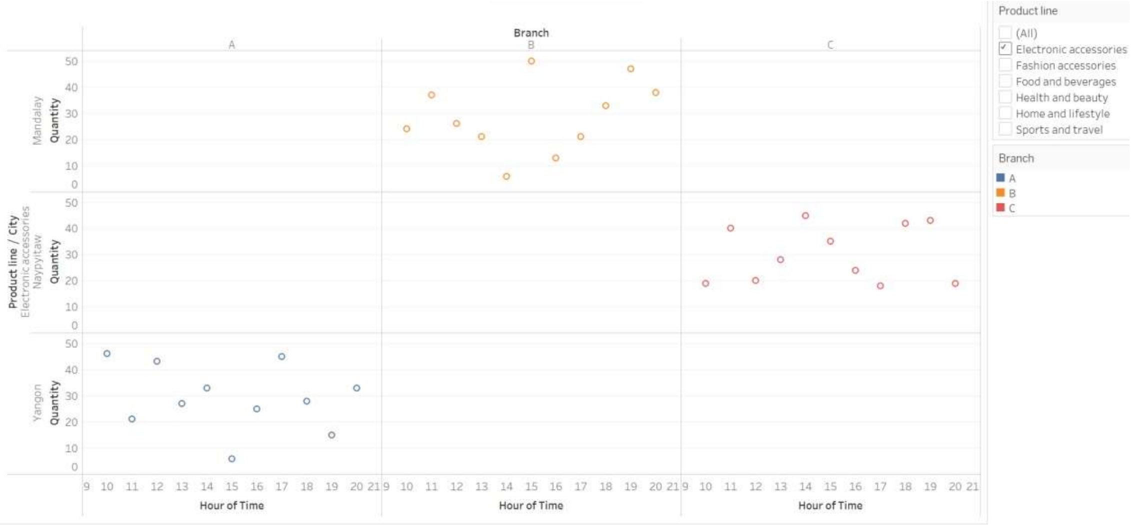


By using two sheets, the visualization shown above is made possible. The branch and quantity attributes are organized in rows and columns, respectively, in sheet 1. By allocating gross revenue to the colors in the Mark, the plot can be generalized to demonstrate which branch is more profitableLet's move on to Sheet 2, where Branch is given a color to aid distinguish between the many branches, and Payment and Quantity are divided into columns and rows, respectively. The graph on Sheet 2 will now be more dynamic. Pick "Worksheets" from the drop-down menu, then choose "Actions," "Add Action," "Set the Source and Target," and "OK."

Result:

The easiest technique to provide a rapid comparison between two visualizations is interactively. Here, branches vs. quantity are depicted, and selecting branch C takes us to a visualization of payments vs. quantity. We can see that Branch C uses cash for more transactions.

Scatterplot:



The plot is done for branch vs product lines is shown above. Product line is placed on the filter. Branch is positioned in the colors in the marks.

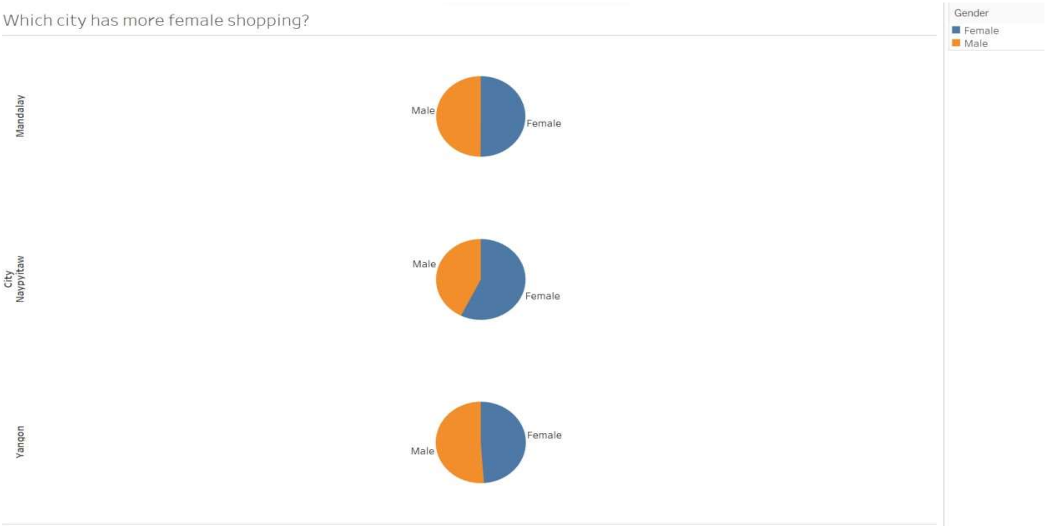
Result:

The majority of the electronics and accessory products in Branch B were sold at 14:00 p.m.

4. Analyze whether the price and ratings of the product line affecting the sales?

For better visualization I have used pie chart

Pie chart:



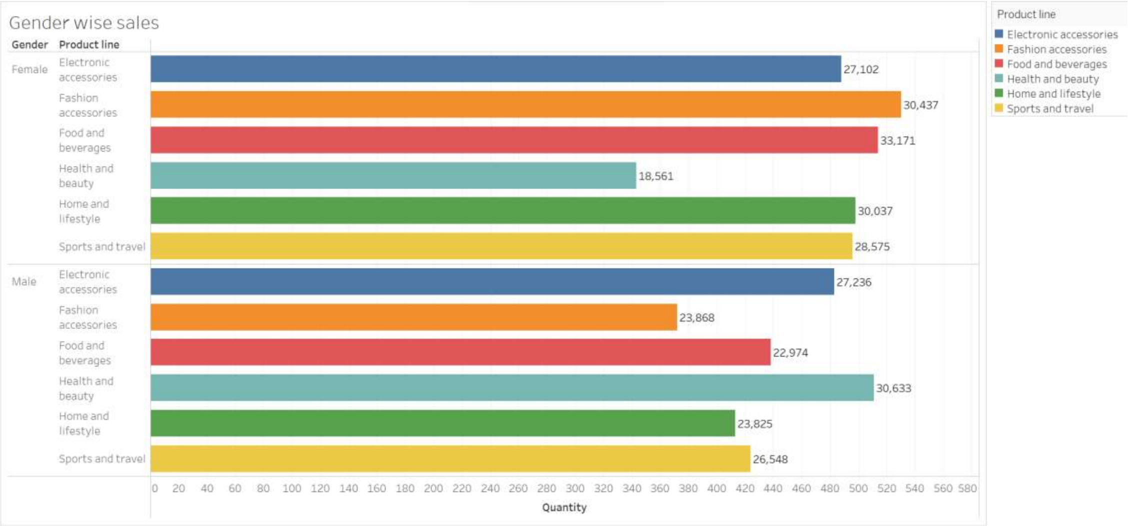
In order to create the above visualization, we must first select the gender, city, and quantity by holding down the control button. Next, we can choose from a variety of graphic representations in the Show Me option. I chose a pie chart in this case.

Result:

The above visualization shows the gender shopping trends in each city. We can determine that Naypyitaw has more female shoppers based on the visualization.

Horizontal bars:

By arranging number in rows and product line, gender in rows, the graphic shown below is created. To distinguish between product lines, product lines are marked with distinct colors. Labels are often printed with totals to show how men and women differ in their buying habits.



Result:

The product line sales broken down by gender are shown in the visualization above. The most popular product category among females is fashion and accessories, where they spent about $30.4k over a three-month period. On the other hand, men prefer to purchase in the health and beauty category. Over a three-month period, they had spent about $30.6k.

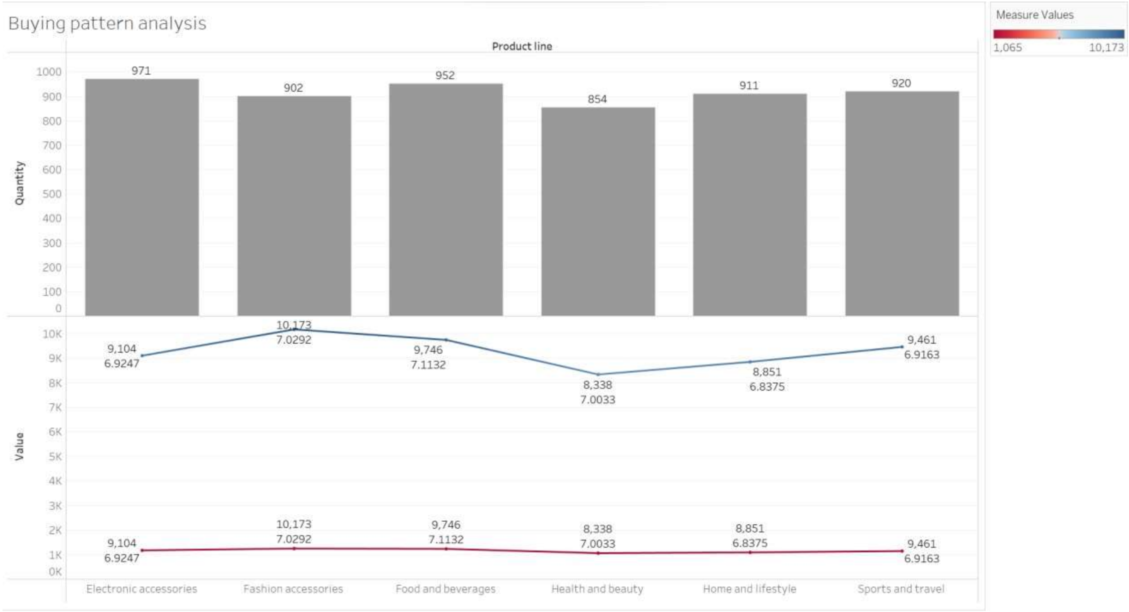
5. Which product line most and which is least bought by customers? Which product line produces more tax and which is least.

Comparison chart:

The sum of the quantities sold is depicted in the bar chart below, while the unit price and rating are shown in the line graphs. To determine if unit pricing and rating have an impact on sales, the bar chart and line graph are compared.

Result:

It is evident from the graphic below that the ratings and unit pricing of the product lines have no bearing on Sales.



Discussions:

1) The ratings and their unit pricing have little impact on sales.
2) In general, women shop more than men. While men are more interested in health and beauty, women are more interested in fashion and accessories.
3) Cash transactions are the preferred method of payment.
4) The busiest time is around 7:00 pm, and January is the most successful month.
5) The most well-liked product category is food and drink.
6) The Branch C average rating is the highest.

Conclusion:

Numerous factors, including customer type, payment method, busiest times, and others, were examined in this project. To provide a thorough understanding of how supermarket sales were in the first three months of 2019, the analysis includes a variety of factors, including sales, volumes, gross income, and more.

References:

1) 25 Appetizing U.S. Food Retail Industry Statistics [2022]: Facts About The Stores Where We Shop – Zippia

2) www.statista.com/topics/1563/supermarkets-in-the-us/