|  |  |  |
| --- | --- | --- |
| IBM-Logo - Chicago Innovation | DGT introduces high end diploma courses - digitalLEARNING Magazine | Edunet Foundation-Delhi- CSR Organization profile |

**SUPERMARKET SALES ANALYSIS (Project Report)**

Submitted in partial fulfillment of completion of the course

Advanced Diploma in IT, Networking and Cloud

Submitted by:

**Team Member- Anubhav Srivastava**

**Team Member- Asheesh Singh**

**Team Member- Haleema Khalid**

**Team Member- Pravitha CP**

**Team Member – Himanshu Srivastava**

Under Guidance of:

**IBM Mentor – (Arpita Banerjee)**

**Edunet Mentor – (MD Sarwar Babu)**

Year 2022

**Abstract**

In this project we will Analyse the sells dataset of a super market and try to analyse how it can develop in future and also what are the necessary steps needed to be taken for it's betterment and customer satisfaction. This will be followed in a step-by-step analysis report from the historical data and its implementation in future.

**Acknowledgement**

It is a great pleasure for me to undertake this project. I feel highly doing the project entitled - **“ Supermarket sales analysis”.**

I am grateful to my project guide **Miss. Arpita Banerjee** teammember of **IBM** and **Mr. MD Sarwar Babu,** Edu net Mentor.

The project would not have completed without their enormous help and worthy experience. Whenever I was in need, they were there behind me.

Although, this report has been prepared with utmost care and deep routed interest. Even then I accept respondent and imperfection.

I also wish to express my sincere thanks to the National Skill Training Institute, Hydrabad for accepting me into the diploma program. In addition, I am deeply indebted to the Ministry of Skill Development & Entrepreneurship and IBM for granting me the diploma course. This technical and financial support has enabled me to complete my diploma course studies successfully. Also, I am grateful the Mr. Rakesh faculty of NSTI Vidyanagar for supporting me for course completion in the specific subject.

**TEAM – 67**

**Vidya Nagar,Hydrabad**

**INDEX**

|  |  |
| --- | --- |
| **Sr. No.** | **Table of Contents** |
| 1 | Chapter 1: Introduction |
| 2 | Chapter 2: Services and Tools Required |
| 3 | Chapter 3: Project Architecture |
| 4 | Chapter 4: Model Building |
| 5 | Chapter 5:Proposed Methodology |
| 6 | Chapter 7: Conclusion |

**CHAPTER 1. INTRODUCTION**

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 for 3 months data. Predictive data analytics methods are easy to apply with this dataset.

Attribute information

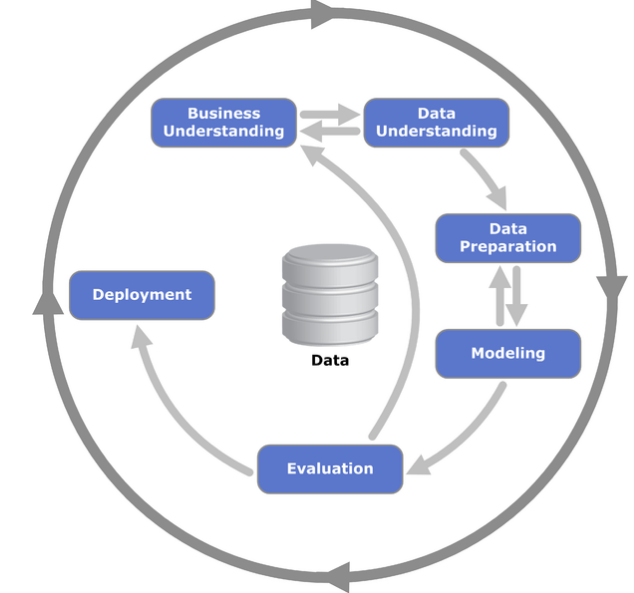
* Invoice id: Computer generated sales slip invoice identification number.
* Branch: Branch of supercenter (3 branches are available identified by A, B and C).
* City: Location of supercenters.
* Customer type: Type of customers, recorded by Members for customers using member card and Normal for without 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 9pm).
* Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet).
* COGS: Cost of goods sold.
* Gross margin percentage: Gross margin percentage.
* Gross income: Gross income.

**CHAPTER 2 SERVICES AND TOOLS REQUIRED**

**The Tools which are using are given below:-**

1. EXCEL
2. MATPLOTLIB
3. SEABORN
4. NUMPY
5. PYPLOT
6. TABLEAU

**CHAPTER 3 PROJECT ARCHITECTURE**



**Chapter 4: Model Building**

This is most important phase which includes model building for prediction of Supermarket Sale. In this we have implemented

various machine learning algorithms which are discussed above for Supermarket prediction.

**Procedure of Proposed Methodology-**

**Step1:** Import required libraries, Import Supermarket dataset.

**Step2:** Pre-process data to remove missing data. Without a strong research problem or problem statement, your team could end up spending resources unnecessarily, or coming up with results that aren’t actionable - or worse, harmful to your business - because the field of study is too broad.

**Step3:** Businesses today have so much data that it can be difficult to know which problems to address first. Researchers also have business stakeholders who come to them with problems they would like to have explored. A researcher’s job is to sift through these inputs and discover exactly what higher-level trends and key concepts are worth investing in.

**Step4:** Select the machine learning algorithm i.e. K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

**Step5:** Build the classifier model for the mentioned ma-chine learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned machine learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

**Step8:** After analyzing based on various measures conclude the best performing algorithm.

**Chapter 5: Proposed Methodology**

**Data gathering and Importing libraries.**

All the standard libraries like NumPy, pandas, matplotlib and seaborn are imported in this step.

We use NumPy for linear algebra operations, pandas for using data frames, matplotlib and seaborn for plotting graphs.

The dataset is imported using the pandas command *read\_csv()*.

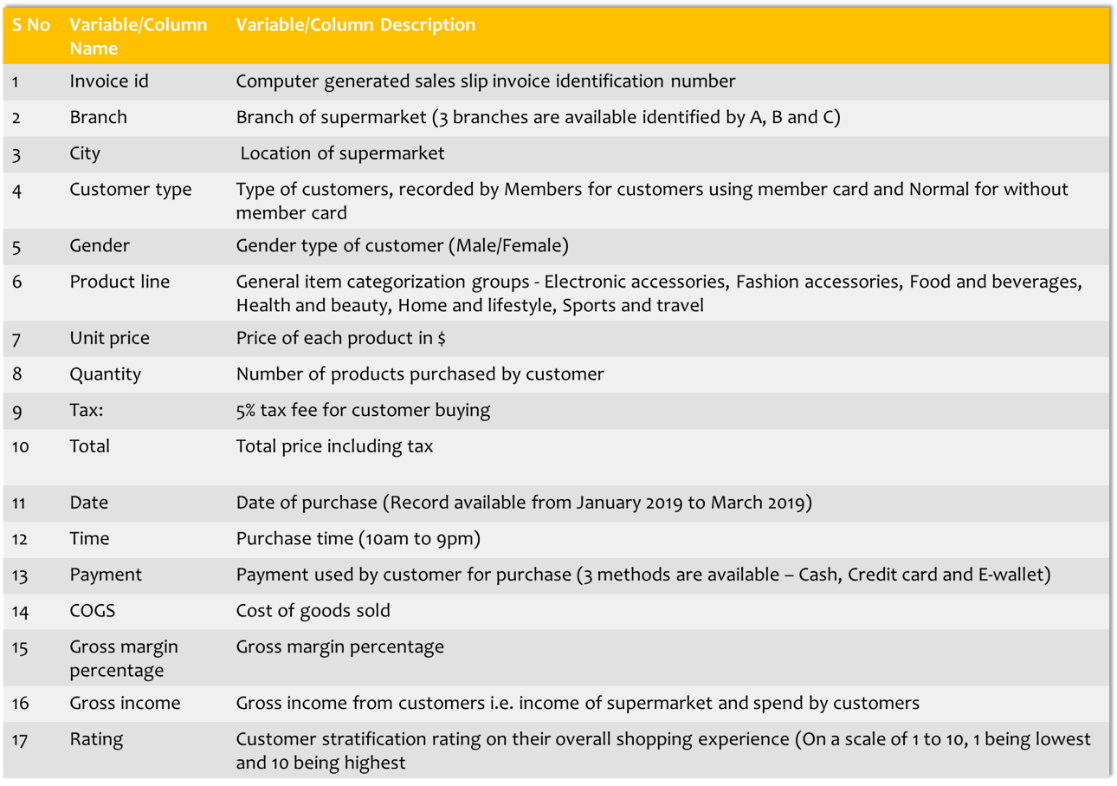
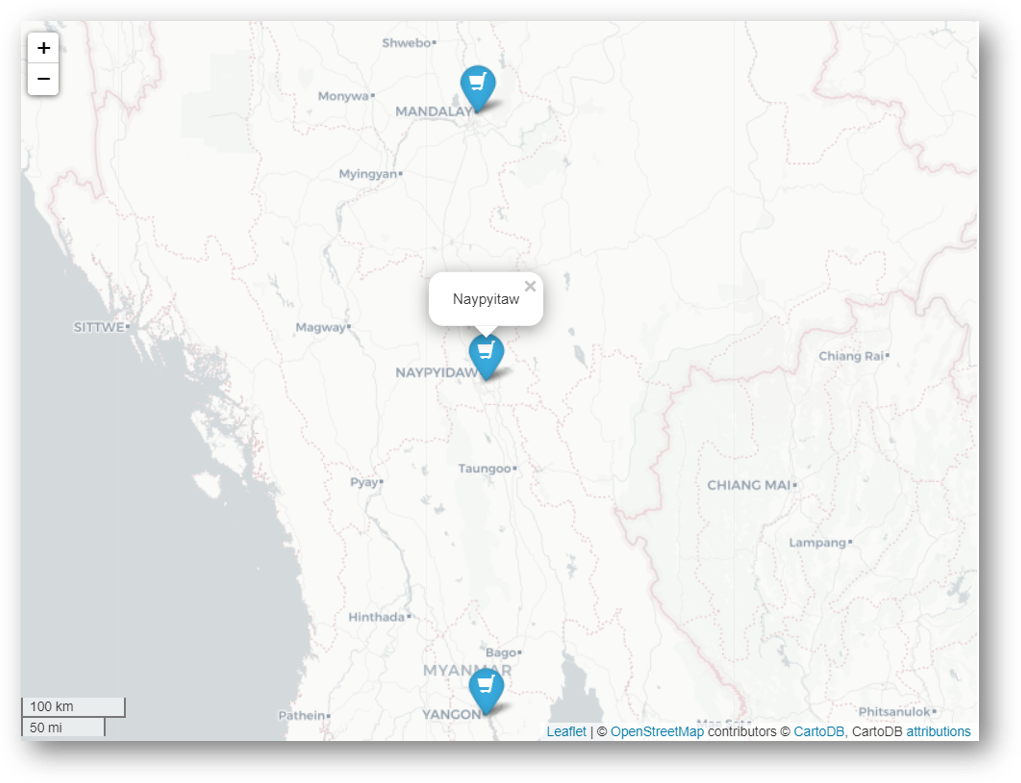


Table 1: Data set with variables



Location of 3 branches/cities

The dataset consists of data from 3 cities or 3 branches in Myanmar as given below-

a) Branch A (Yangoon)

b) Branch B (Mandalay)

c) Branch C (Naypyitaw)

# **Uni-variate analysis**

Uni-variate analysis is the analysis involving a single variable (‘uni’) without considering relationships with other variables. This is the stand-alone analysis of a variable/column without considering any casual relationships with other variables. We will see how a simple uni-variate analysis can help to get more insights into the data.

**The Question behind this Analysis are as :-**

**Question 1**: **What does the customer rating look like and is it skewed?**

The lines of code below can be used to answer this:

sns.distplot(df['Rating'])  
plt.axvline(x=np.mean(df['Rating']), c='red', ls='--', label='mean')  
plt.axvline(x=np.percentile(df['Rating'],25),c='green', ls='--', label = '25th percentile:Q1')  
plt.axvline(x=np.percentile(df['Rating'],75),c='orange', ls='--',label = '75th percentile:Q3' )  
plt.legend()

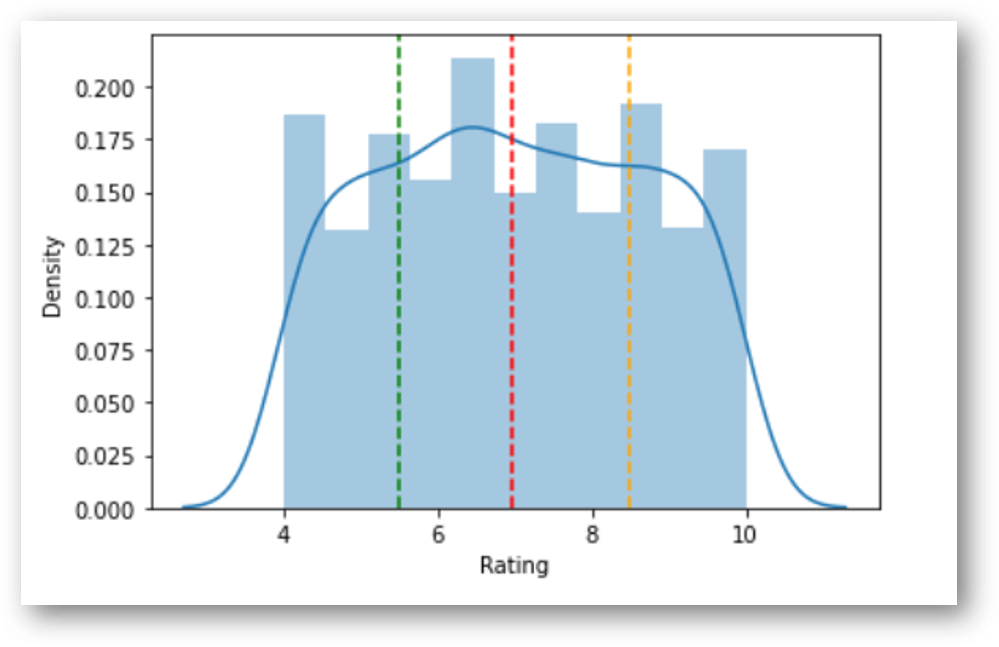


Fig 1: Distribution of Rating

***The rating distribution looks uniform and there seems to be no skewness on the left or right side of the distribution. We can plot the mean and the percentiles to show this as well. The red dotted lime is the mean and the green and orange lines indicate the 25th and 75th percentiles respectively. The mean rating is ~7 across products.***

**Question 2**: **Is there any difference in aggregate sales across branches?**

Next we would like to know whether there is any difference in aggregate sales across the branches. This can be achieved by a count plot as given below:

sns.countplot(df['Branch'])  
df['Branch'].value\_counts()

A 342  
B 333  
C 328  
Name: Branch, dtype: int64

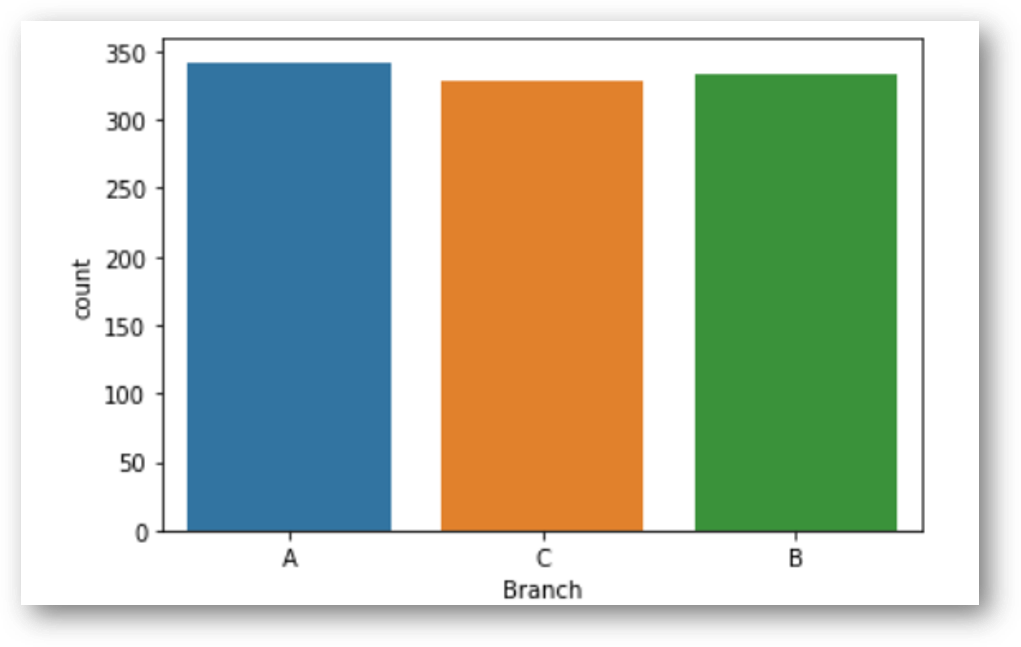


Fig 2: Sales by Branch

***There is not much difference in sales across the 3 branches of A, B and C. The sales in branch A is a bit higher than the rest of the branches.***

**Question 3**: **Which is the most popular payment method used by customers?**

We can again do this using count plot as given below:

sns.countplot(df['Payment'])

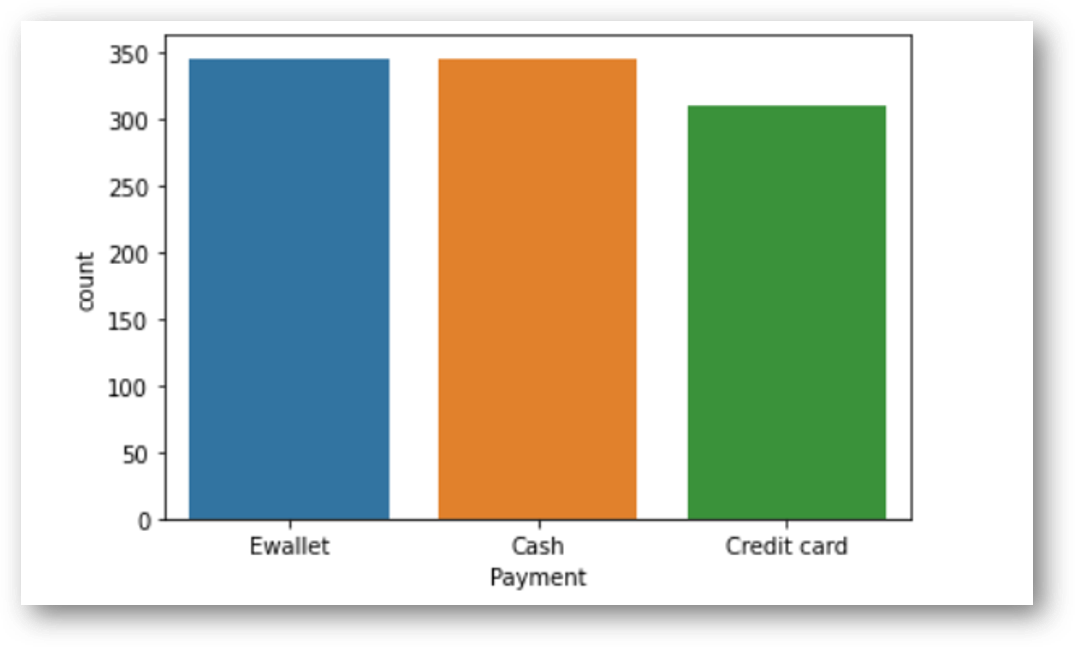


Fig 3: Most popular payment method (E wallet)  
 ***The most popular payment method is in-fact E-wallet and not credit cards.Cash payment is also popular.***

# **3. Bi-variate analysis**

Let us now consider two (bi) variables together and look at the interaction/relationship between them. This is bi-variate analysis that can help to draw important conclusions about the data.

Let us first consider gross income and try to answer the following questions:

**Question 4**: **Does gross income affect the ratings that the customers provide?**

We can use scatter plot and regression plot to answer this question. We can use the following code:

sns.scatterplot(df['Rating'], df['gross income'])

The output will look like this:

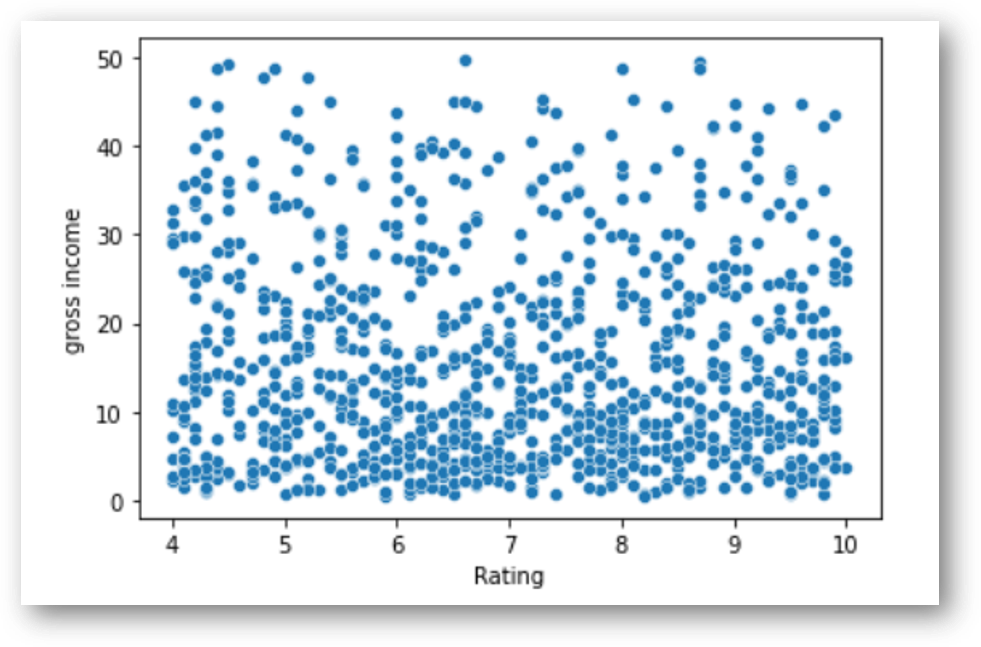


Fig 4: Scatter plot: Gross income vs Rating

We can also fit a trend-line to this plot using the regression plot as given below.

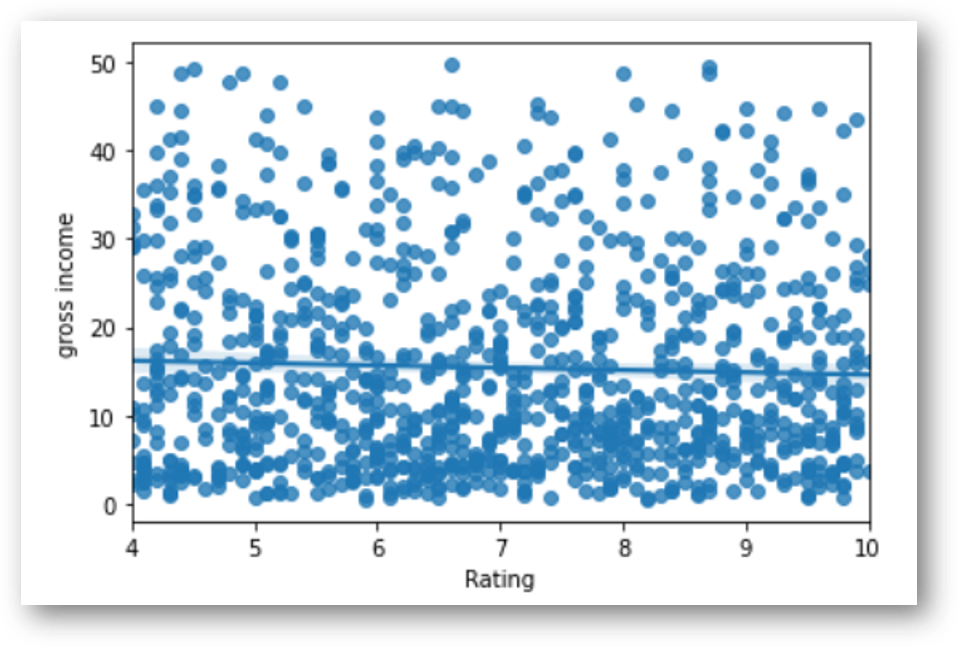


Fig 5: Regression plot

***As you can see from the scatter plot and the trend line which is pretty flat that there is no relationship between gross income of a customer and his rating.***

**Question 5**: **Which branch is the most profitable?**

We can use use the box plot given below for this.

sns.boxplot(x=df['Branch'], y=df['gross income'])

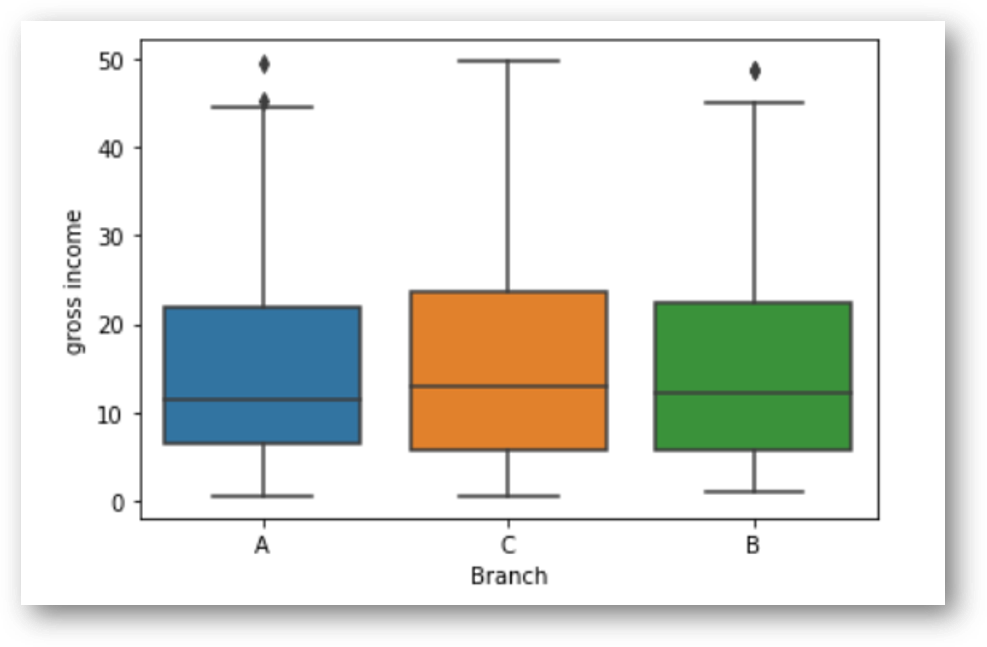


Fig 6: Gross income by branches

***There is not much difference in gross income by branches at an average level. Branch C has a slightly higher income than A or B.As observed earlier,though branch A has slightly higher sales than the rest,C i.e. Naypyitaw is the most profitable branch in terms of gross income.***

**Question 6: Is there any relationship between Gender and Gross income?**

Similar to the branch analysis earlier,we can use a box plot to answer this question.

sns.boxplot(x=df['Gender'], y=df['gross income'])

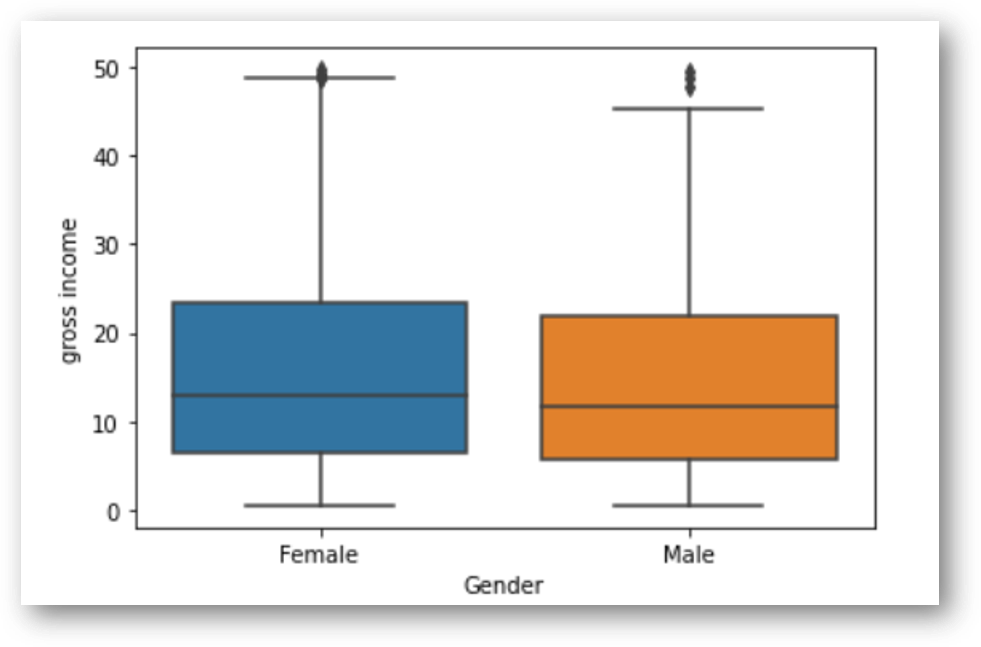


Fig 7: Gross income by Gender

***Gross income is similar for both male and female, though female customers spend a bit higher at the 75th percentile.***

**Question 7: Is there any time trend in gross income?**

There are multiple dates per customers, so we have to summarize the data to find the time trend of gross income. We will use ‘**date**’ as the index for this and for summarization we will use mean value of the variables.

sns.lineplot(x= df.groupby(df.index).mean().index,   
 y = df.groupby(df.index).mean()['gross income'])

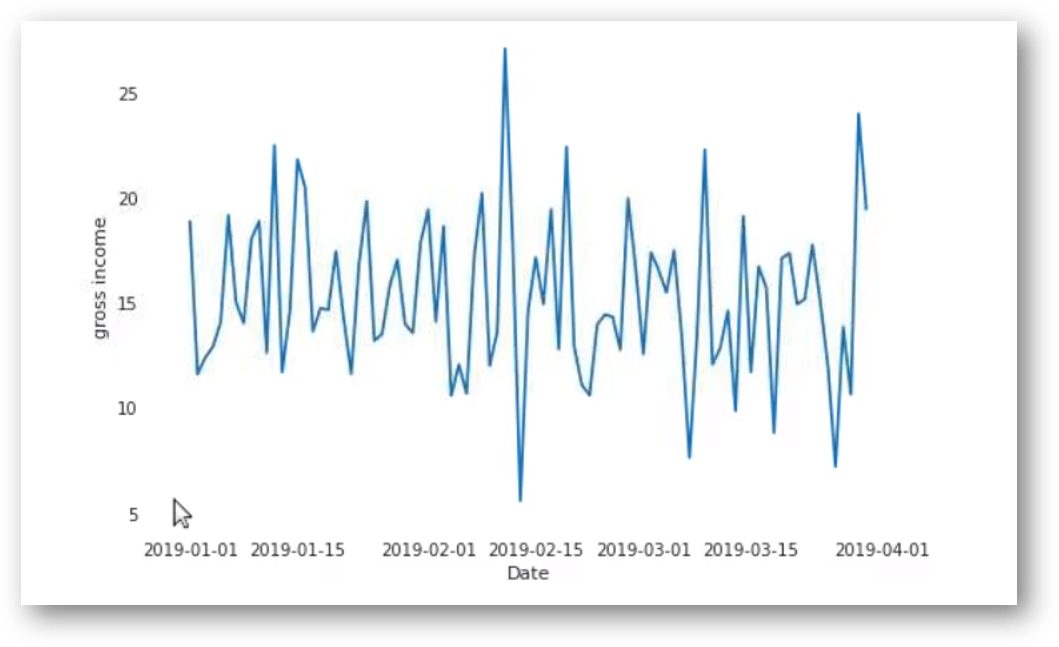


Fig 8: Time trend of gross income

***No particular time trend is observed except for some days when the gross income is pretty high or pretty low. Overall it remains at a certain average level.***

**Question 8: Which product line generates most income?**

cat=df[["Product line", "gross income"]].groupby(['Product line'], as\_index=False).sum().sort\_values(by='gross income', ascending=False)  
plt.figure(figsize=(20,8))sns.barplot(x='Product line', y='gross income', data=cat)

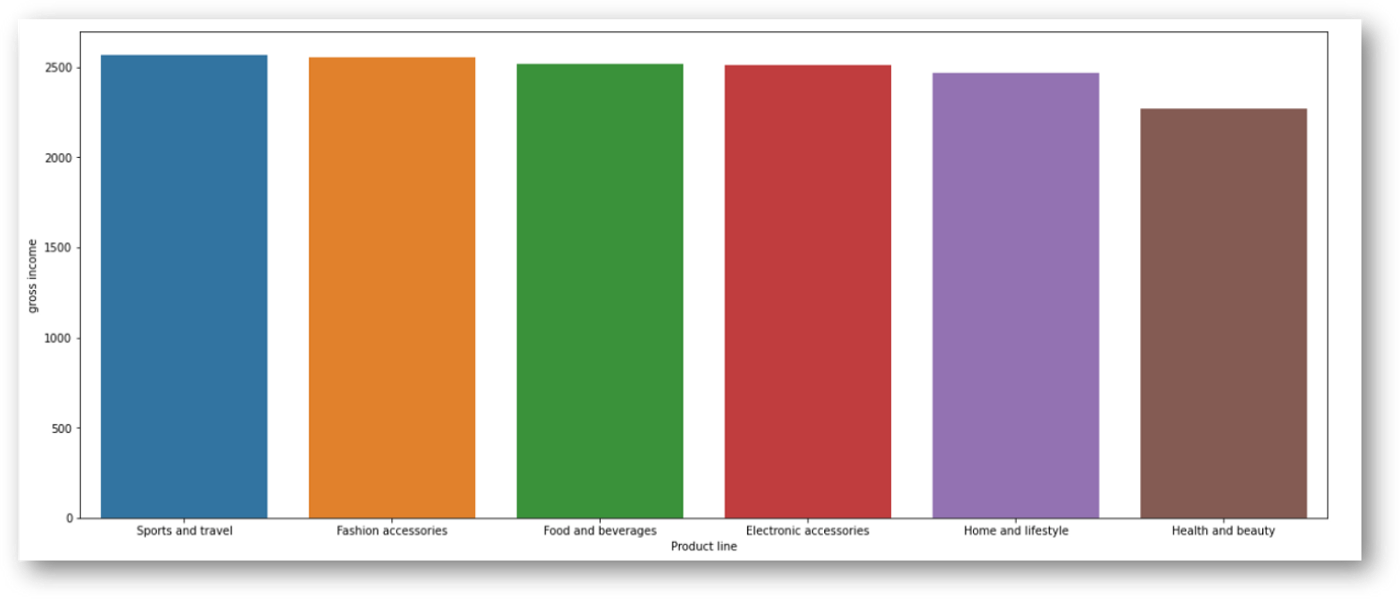
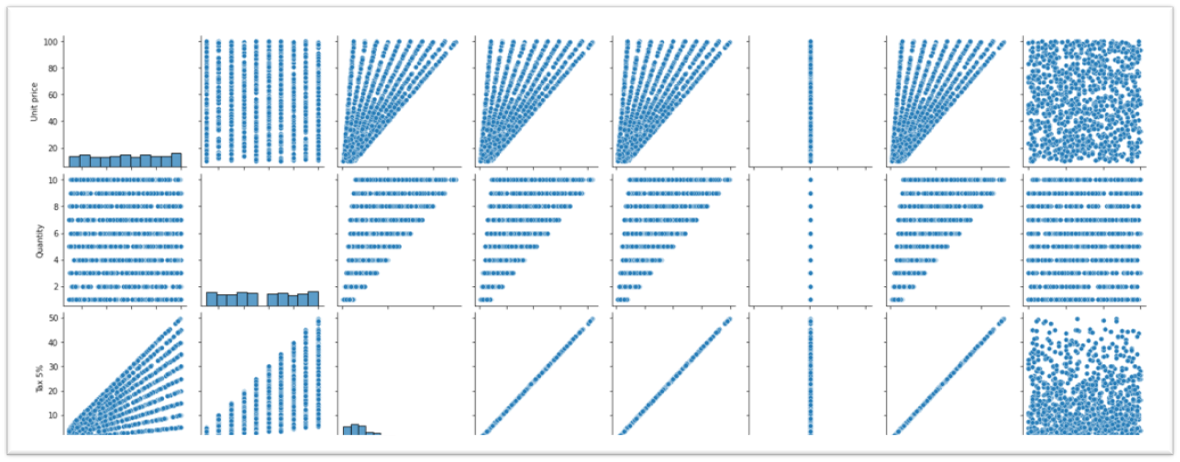


Fig 9: Gross income by products

***Gross income is highest in sports and travel.***

## **Pairwise plot: We can look at all the bi-variate relationships in the data using pairwise plot in one go. But it is recommended not to use this method in large datasets because it provides all possible variable combinations. A snippet of our data today is given below:**



# **3. Correlation analysis**

Correlation is the degree to which a pair of variables is linearly related.

Correlation analysis indicates the positive or negative or zero correlation between the variables. A positive correlation means that the values of the 2 variables increase together and negative correlation means the values of one variable decrease with the other. Zero correlation implies no correlation.

The seaborn heat-map can be used to visualize the correlations among variables.

sns.heatmap(np.round(df.corr(),2), annot=True)

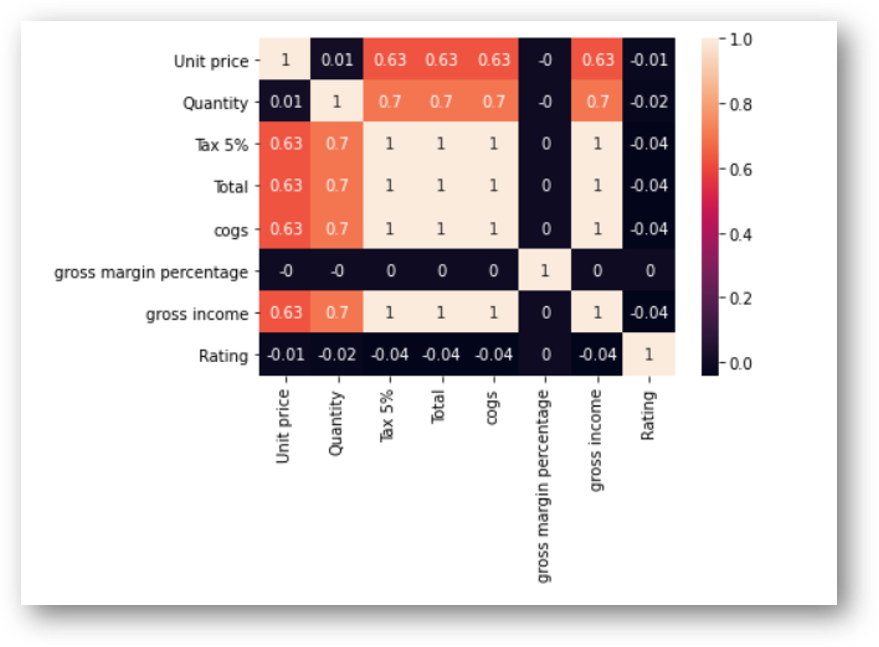


Fig 10: Correlation analysis

As you can see, unit price is positively correlated to cogs with 63% correlation. Another interesting observation is that ‘Ratings’ hardly has any correlation with any other variables.

As is obvious, Quantity and gross income has very high correlation of 70%.

# **4. Other analysis**

**Question 9: What is the spending pattern of females and males and in which category do they spend a lot?**

For the analysis on spending patterns of males and females, we can create create a dummy variable for Gender and concatenate it with the original data.

We can also utilize the ‘hue’ option to get this information in one chart.

plt.figure(figsize=(12, 6))  
plt.title('Total Monthly transaction by Gender')  
sns.countplot(df['Product line'], hue = df.Gender)

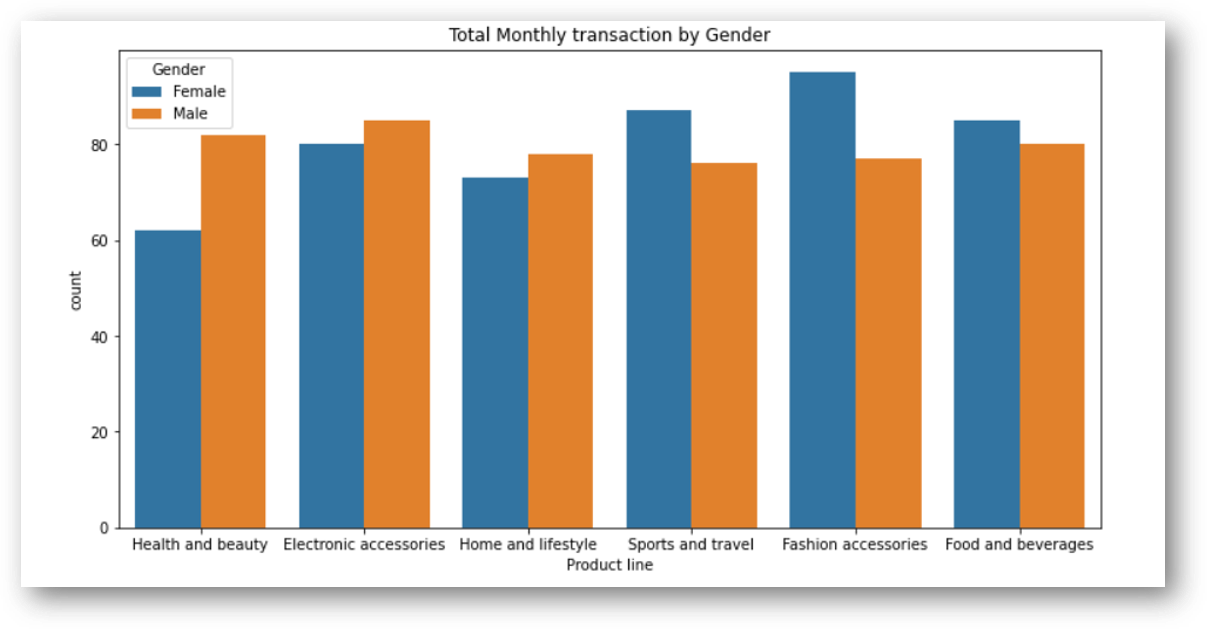


Fig 11: Spending pattern: Males vs Female

***Females spend on 'fashion accessories' the most and for males it is 'Health and beauty'. Females also spend more on 'Sports and travel'.***

**Question 10: How many products are bought by customers?**

Using distribution plot and heat map we can explore the number of products/quantities that most customers buy.

xdata = [1,2,3,4,5,6,7,8,9,10]  
plt.figure(figsize = (12,6))  
sns.distplot(df['Quantity'])  
plt.xticks(xdata)

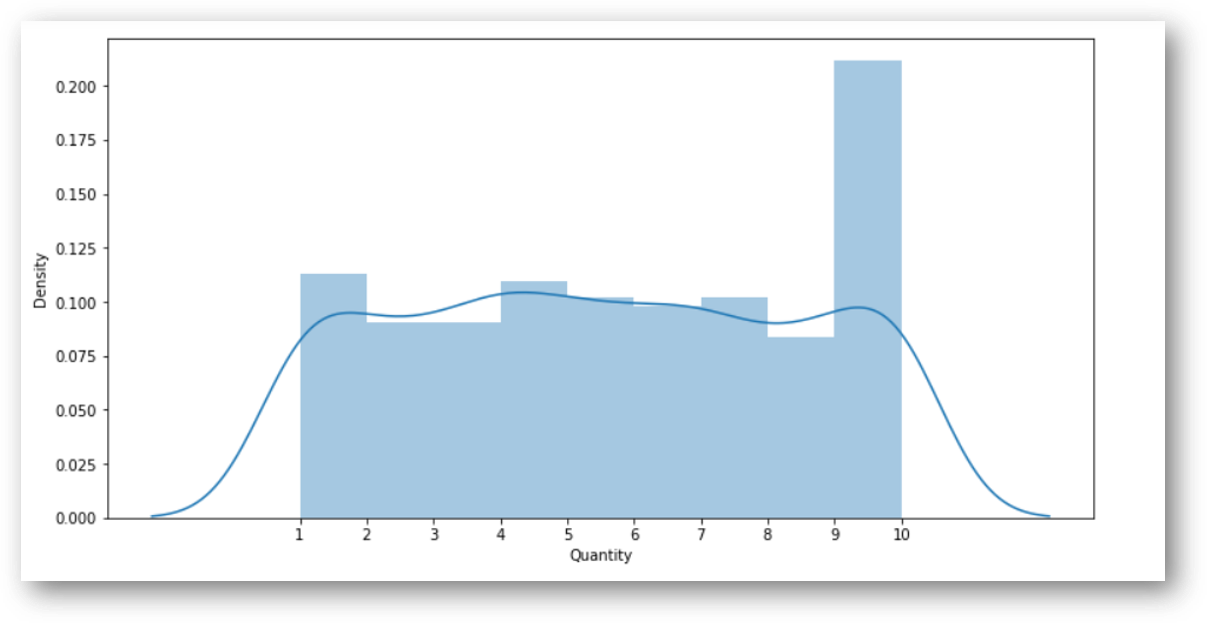


Fig 12: Quantity distribution Plot

***Most of the customers buy 10 quantities.***

**Question 11: Which day of the week has maximum sales?**

Let us now explore if there is any particular day of the week when the sales is higher.

plt.figure(figsize=(8, 6))  
plt.title('Daily Sales by Day of the Week')  
sns.countplot(df['weekday'])

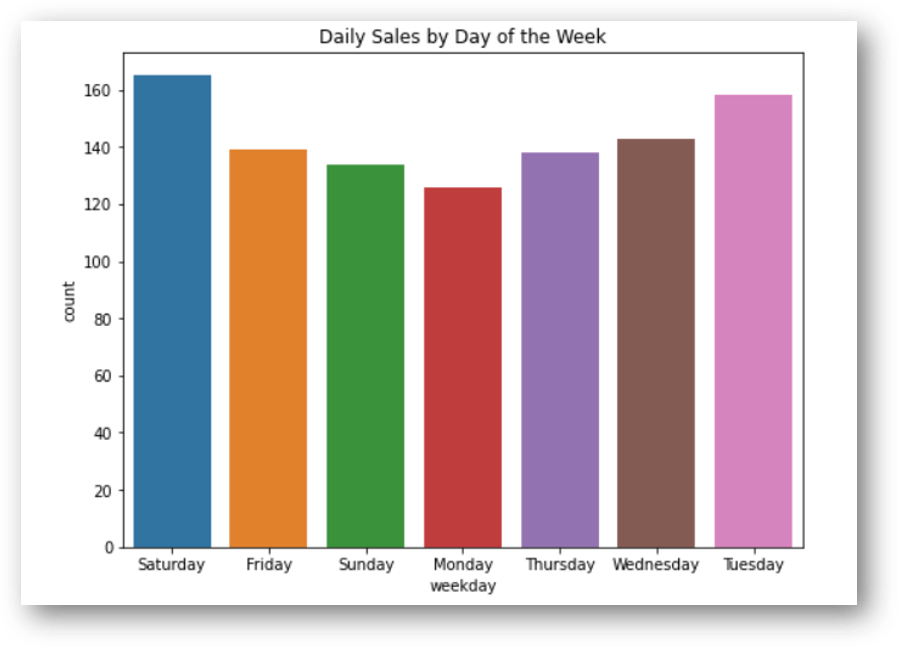


Fig 13: Daily sales by day of the week

***Sales is highest on Saturdays probably because it is the weekend. Interestingly, Tuesdays is a close second. Mondays is the lowest in sales, probably because it is start of the working week.***

**Question 12: Which hour of the day is the busiest?**

We need to extract the hour from the ‘Date’ variable to do this analysis. This can be done in the following way.

df['Time'] = pd.to\_datetime(df['Time'])  
df['Hour'] = (df['Time']).dt.hour  
df['Hour'].unique()

Plotting the data we get the below output:

sns.lineplot(x="Hour", y = 'Quantity',data =df).set\_title("Product Sales per Hour")

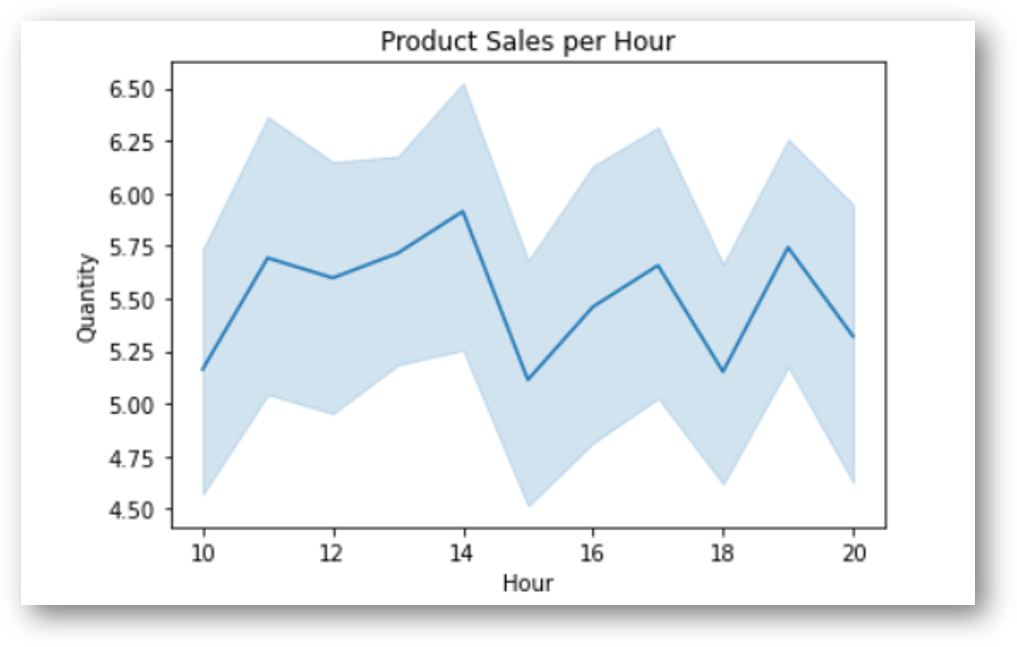


Fig 14: Hourly sales of products

***Peak is observed in the 14th hour i.e 2 pm of the day. Hence, sales is typically higher in the afternoons.***

**Question 13: Which product line should the supermarket focus on?**

To answer this question, let us look at 2 graphs below:

a) **Rating of products**: Rating is similar across products and at around 7.

xdata = [0,1,2,3,4,5,6,7,8,9,10]  
plt.figure(figsize = (12,6))  
sns.barplot(y = df['Product line'], x = df['Rating'])  
plt.xticks(xdata)

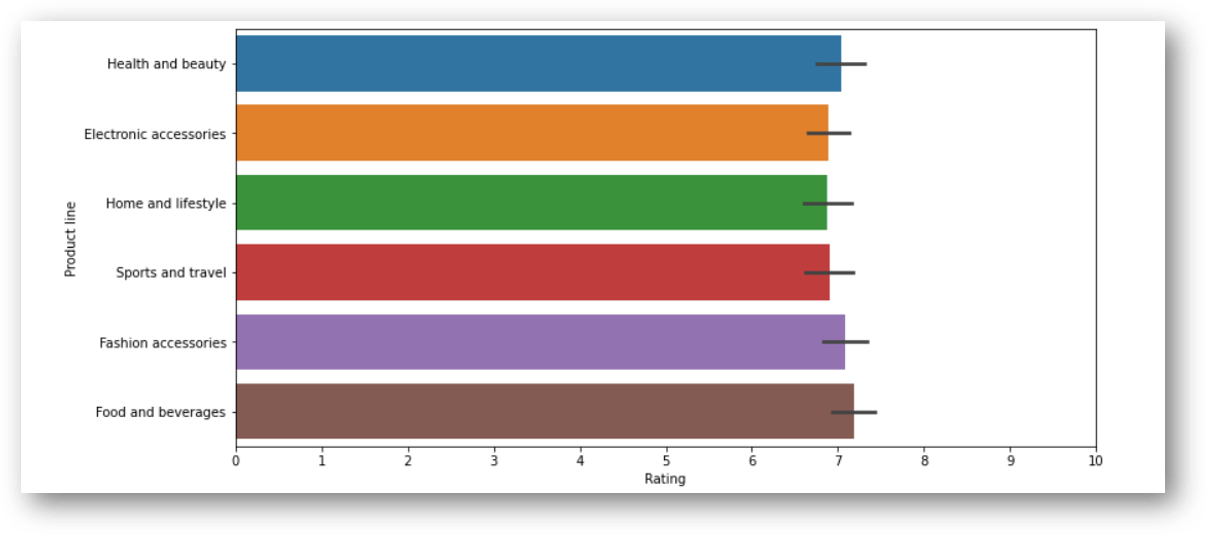


Fig 15: Ratings across product lines

b) **Quantity purchased by product**: The mean quantity is lower for ‘fashion accessories’ and ‘food and beverages’

sns.boxenplot(y = 'Product line', x = 'Quantity', data=df )

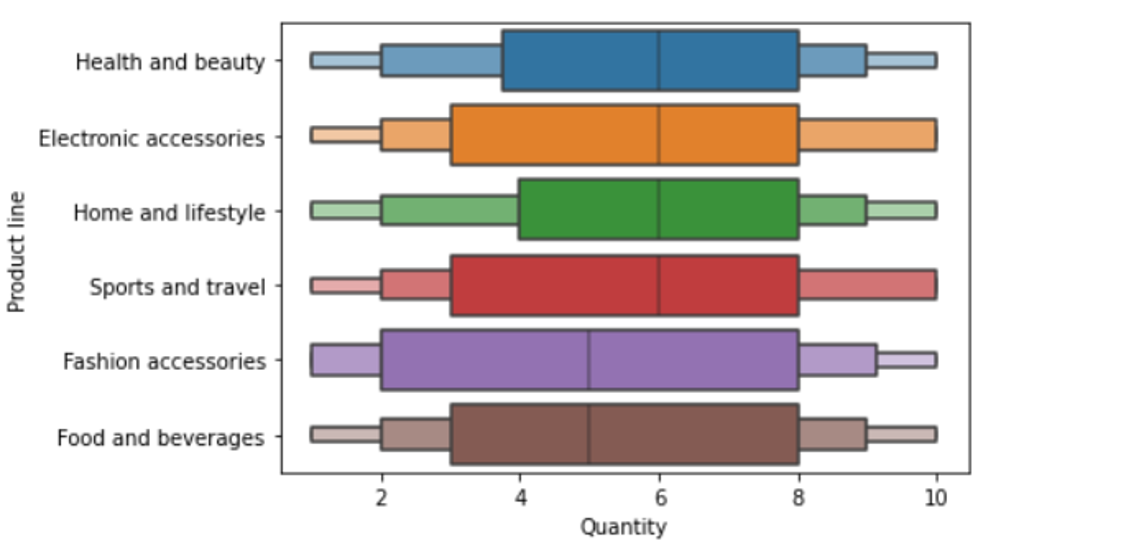


Fig 16: Quantity of products

***Though the rating for ‘fashion accessories’ and ‘food and beverages’ is high but quantity purchased is low. Hence, supply for these products need to be increased.***

**Question 14: Which city should be chosen for expansion and which products should it focus on?**

Let us first answer which city should be chosen for expansion:

plt.figure(figsize=(20,7))  
sns.barplot(df['City'],df['gross income'],palette='Set1')  
plt.xlabel('City name',fontsize='16')  
plt.xticks(fontsize='16')  
plt.ylabel('Gross income',fontsize='16')  
plt.yticks(fontsize='16')

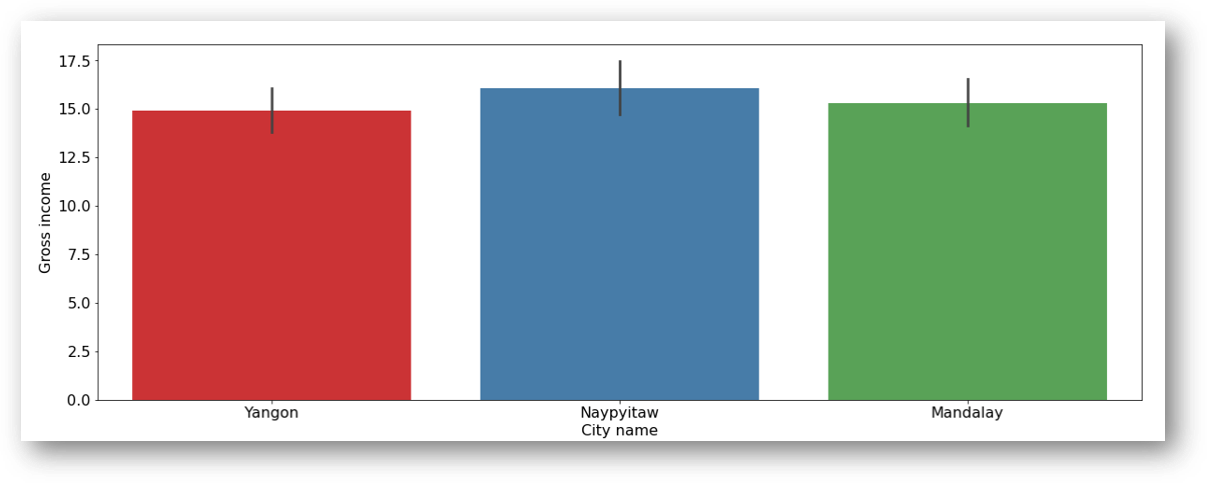


Fig 17: Gross income across branches/cities:

***It is obvious that Naypyitaw is the most profitable city, hence the expansion plan should be based on this city.***

Let us look at the products sold at ‘***Naypyitaw’***

plt.figure(dpi=125)  
sns.countplot(y ='Product line', hue = "City", data = df)   
plt.xlabel('Count')  
plt.show()

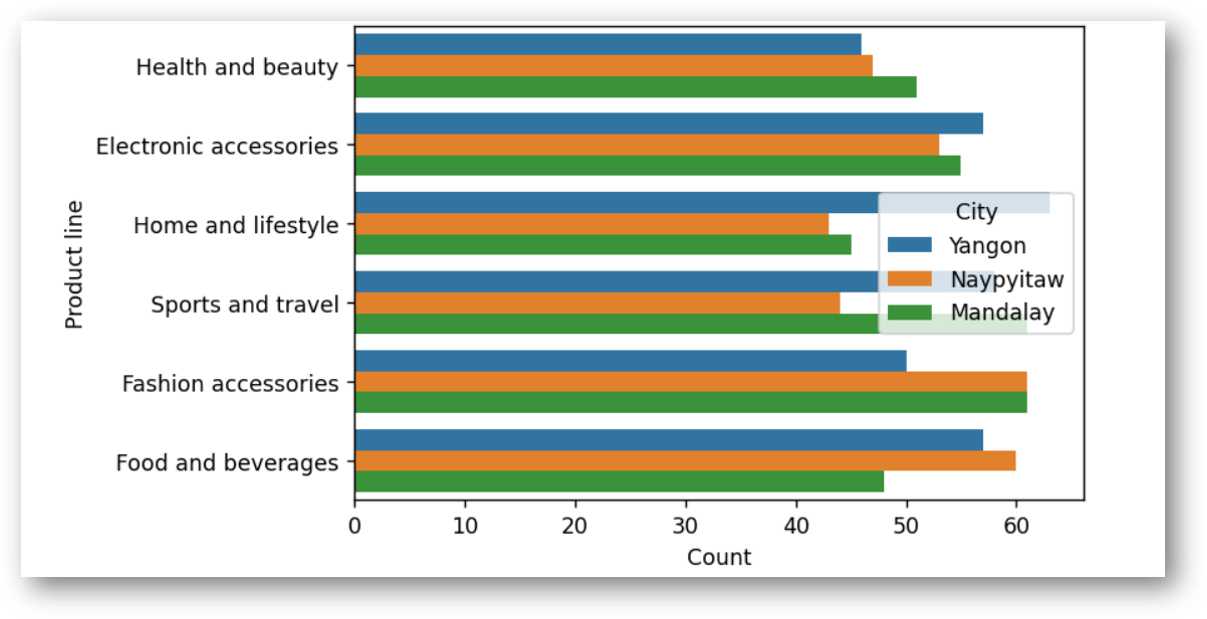


Fig 18: Sales of products across cities/branches

***Fashion accessories and food and beverages are the most sold product in Naypyitaw and these products should be focused on for expansion along with electronic accessories.***

# **Conclusion:**

We used uni-variate, bi-variate and correlation analysis to perform basic EDA on the supermarket sales data.

To summarize below are some of the findings/observations from the data:

1. The customer rating is more or less uniform with the mean rating being around 7 and there is no relationship between gross income and customer ratings.
2. The data consists of 3 cities/branches. Though branch A has slightly higher sales than the rest, C i.e. ***Naypyitaw*** is the most profitable branch in terms of gross income.
3. Fashion accessories and food and beverages are the most sold product in Naypyitaw and these products should be focused on along with electronic accessories.
4. The most popular payment method is E-wallet and cash payment is also on the higher side.
5. There is no particular time trend that can be observed in gross income.
6. At an overall level, ‘Sports and Travel’ generates highest gross income.
7. Gross income is similar for both male and female, though female customers spend a bit higher at the 75th percentile. Females spend on ‘fashion accessories’ the most and for males surprisingly it is ‘Health and beauty’. Females also spend more on ‘Sports and travel’ which generates highest income overall.
8. Using the correlation analysis, one interesting observation has emerged that customer ratings is not related to any variable.
9. Most of the customers buy 10 quantities and busiest time of the day is afternoon i.e. around 2 pm which records highest sales. Sales is higher on Tuesdays and Saturdays compared to the rest of the week.
10. Though the rating for ‘fashion accessories’ and ‘food and beverages’ is high but the quantity purchased is low. Hence, supply for these products need to be increased.