AN EXPLORATORY ANALYSIS ON SUPERMARKET SALES IN MYANMAR

Project Report submitted to the

SDM COLLEGE(Autonomous)

in partial fulfillment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

by

Dhanashree D

Under the guidence of

Ms. Supriya S.P

Department Of Post Graduate Studies in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE - 574240

Karnataka, INDIA

NOVEMBER 2021

Contents

1 Chapter 1

INTRODUCTION

1.1 Introduction

Supermarket
A supermarket is a self-service shop offering a wide variety of food, beverages and household products, organized into sections. This kind of store 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. Consumers have wide variety of choice to select the product they want. Consumers are very sensitive on what they are buying. Thus it is important for the supermarkets to keep all the products in stock.
Supermarkets always monitor the consumer buying trend and always keep the certain products in stock all the time. It is also important that they should keep all the products in stock at all their stores. Now supermarkets even sells the ethnic foods like, Indian, African to attract ethnic customers.
They use all sorts of marketing strategy to attract the customers. Their adverts are more customer centric which they like attract. For e.g., Asda attracts customers with the slogan 'Always low price'. Tesco slogan is 'Every little helps' as they want to attract a large economical consumer group which cannot afford premium supermarkets. Price is the major factor which influence to consumers to switch to other supermarkets or retailer.
Supermarkets compete with each other to sell the products at the best rate to the consumers. Tesco and Asda is the biggest competitor in the economical consumer range. They are trying to sell the products to best price to retail the customer loyalty to supermarket.
During the economic downturn consumers tends to limit on their spending and looking for the cheap bargain products. Tesco has large consumer group. It is easy for them to buy large quantity to meet the demand in order to meet the consumer requirement at a lower price. So they can offer the products at a lower price to customers. Consumer can also compare the price of the products they are buying to make it easy for the them to select the products.
As modernisation and advancements occur rapidly in every field globally, the shade of markets have also changed a lot in the last decade.To meet the versatile needs of the consumers, supermarkets have become quite common in all parts of the world. These supermarkets have not only changed the shade of the market but affected the customers and shopping habits also.

As the setup of the supermarket is large, the process of its operation is complicated. To improve customer base, a supermarket always offers exciting offers and deals. Moreover, the supermarkets concentrate on the aspect of maintaining good hygiene, unlike the traditional open fresh markets. Furthermore, customers get a lot of opportunities to try different brands and products that are normally unavailable in local markets.

One of the primary concerns of a supermarket is to sell farmed products, much like a fresh market. This makes up a significant portion of the stock of a supermarket.

In the traditional market, majority payments are cash-based. The scenario is however different in case of supermarkets in metro and urban cities. Customers can pay online or through debit and credit cards in a secured manner. This saves their time as well.

Supermarkets are the most important aspect of food production and distribution because they are the interface between supply and demand. There are alternatives such as farmers markets, but not in reach. They 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. Overall it's all about economies of scale. The scale of operation makes grocery a profitable business.

1.1.1 Literature Review

In 2018, M. Guruprasad carried out a research to understand customer preference and perceptions on D Mart products and services. The study was conducted in the area of karjat and badlapur. The core objective is to offer customers good products at great value. The study shows that factors which influence customer to shop are discount, convivences and offerings/variety. Large majority of consumer are loyal to brand, also what contribute to the popularity of the demand is the service provided by D'mart coupled with the attractive pricing strategy followed by it. It has also been observed during the study, that the presence of the D'mart has made a impact on the retailer.

In 2000, Martin pesendorfer carried out a study on pricing behaviour in supermarkets. This paper examines supermarket price for ketchup products to explain temporary price reductions. The study reveals that the demand at low price levels depends on past price, suggesting intertemporal effects in demand. Based on descriptive data analysis, a model of intertemporal pricing is considered which assumes that demand at low prices accumulates in the time elapsed since last scale. The predicted probability of a sale increases in the time elapsed since last sale.

In 2018, Mansour Rasoly conducted a research on consumer satisfaction of supermarket, which gives a comprehensive information and data regarding consumers supermarket preference in the area of mysore city. He is trying to examine the consumer price satisfaction and also their preference about imported products or domestic products. Based on his study it is identified that the consumer are facing problems interms of parking facilities and bill counter of supermarket. Also he had suggested some suitable solution to overcome these problem. The study shows that the customers now prefer the organized retailers than unorganized retailers and it is because of products quality, products availability, international and recognized products.

In 2013, Gyeong-Cho Kim carried out a study on the effects of super-supermarket Service quality on satisfaction in Store Selection. The purpose of this study is to analyze the relationship between SSM's service factors and following customer satisfaction and how satisfaction difference influences customer's repurchase intention. This study examines the effects of service quality and product quality of SSMs on customer satisfaction levels, and analyzes whether these factors affect customer's revisit intention directly. The study reveals that 44.1% of them were satisfied with price. Regarding the product quality of SSM, 54.4% were neutral and 33.8% responded to feel reliable about the quality. Overall, 88.2% showed positive reaction toward product quality.

Objectives and Scope of Study

I am doing a project entitled "**An Exploratory Analysis of Supermarket Sales in Myanmar**"

1.1.2 Objectives

1. To compare sales trend of different branch.
2. To find out which city should be chosen for expansion and the products to be focused on.
3. To gain insights on customer purchasing behaviour based on gender, membership status (Member / Normal)
4. To know whether the payment method has any impact on customer rating.
5. To study about customer preference on different product line based on gender are independent or not.
6. To study whether unit price and tax has any influence on the quantity of product being purchased by customer.

1.1.3 Scope of Study

The result obtained from this project helps the supermarket to improve their sales and also intergrate their marketing strategies. This project gives an adequate information on customer purchasing pattern. Also it is useful for those are interested to know about the customer preference and their needs.

2 Chapter 2

2.1 About the Data

A secondary data has been collected from the website of "Kaggle".
The dataset consists of historical sales data of a supermarket company from three different branches in Myanmar as given below :

1. Branch A (Yangoon)
2. Branch B (Mandalay)
3. Branch C (Naypyitaw)

This data was collected during the month Jan-March 2019.The data consists of 1000 observations and 17 variables. The description about the variables considered in the analysis are given as follows:

- **Branch** : It indicates the branch of supermarket (A, B, C).
- **City** : It indicates the location of supermarket (Yangon, Naypyitaw, Mandalay).
- **Customer type** : It indicates the type of customers recorded by 'Members' for customer using member card and 'Normal' for customer without member card.
- **Gender** : It indicates the Gender of the customers (Male, Female).
- **Product line** : It indicates the general item categorization groups - (Food and beverages, Health and beauty, electronic accessories, Fashion accessories, Home and lifestyle, Sports and travel).
- **Unit price** : It indicates the price of products in US dollars.
- **Quantity** : It indicates the number of products purchased by each customer.
- **Tax** : It indicates the 5% tax fee on top of purchase amount.
- **Total** : It indicates the total price of the purchase including tax.
- **Date** : It indicates the date of purchase (January 2019 to March 2019).
- **Payment** : It indicates the payment methods used by customer for purchase. Three methods are available E-wallet, Cash and Credit card.
- **Rating** : It indicates the satisfaction rating given by each customer on their overall shopping experience.

2.2 Statistical Techniques Used

'Python' and 'R' language has been used to carry out the analysis and the interpretation of the data. The statistical methods used in order to carry out the analysis are given as follows :

2.2.1 Chi-Square Test of Independence

The chi-square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of each category for one nominal variable is compared across the categories of the second nominal variable. The data can be displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable. The hypothesis under consideration are given as follow :

H_0 : The two categorical variables are independent.
H_1 : The two categorical variable are dependent.

The test statistic for the chi-square test of independence is denoted χ^2, and is computed as:

$$\chi^2 = \frac{\sum (O - E)^2}{E}$$

where O represents the observed frequency, E is the expected frequency under the null hypothesis and is computed as follows :

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

The expected frequency count for each cell of the table must be atleast 5.
The test procedure is to reject the null hypothesis H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$, where $\chi^2_{\alpha,(r-1)(c-1)}$ is the upper α^{th} percentile value of the central chi-square distribution with $(r-1)(c-1)$ degrees of freedom, r is the number of rows and c is the number of columns.
Also using the $p - value$ we can draw conclusion about the test. If the $p - value$ is less than 0.05, then reject the null hypothesis H_0.

2.2.2 Correlation

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. However, in general, the presence of a correlation is not sufficient to infer the presence of a causal relationship.

The most familiar measure of dependence between two quantities is the "Pearson's correlation coefficient", commonly called simply "the correlation coefficient". Mathematically, it is defined as the quality of least squares fitting to the original data. It is obtained by taking the ratio of the covariance of the two variables in question of our numerical dataset, normalized to the square root of their variances.
The population correlation coefficient ρ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Correlation coefficient shows how strong a relationship is between data. It takes value between -1 and 1, where:

- A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases.
- A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.
- A zero correlation exists when there is no relationship between two variables.

If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the matrix data structure is used. Such matrix is called as **correlation matrix.** Graphical representation of correlation matrix representing correlation between different variables is known as **correlation heatmap**.

2.2.3 Poisson Regression Model

The poisson distribution models the probability of y events (i.e., failure, death, or existence) with the formula:

$$P(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables. he general mathematical form of Poisson Regression model is:

$$log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

where,

- y: Is the response variable
- α and β : are numeric coefficients, α being the intercept, sometimes α also is represented by β_0.
- x is the predictor/explanatory variable

The coefficients are calculated using methods such as Maximum Likelihood Estimation(MLE).
Consider an equation with one predictor variables and one response variable:

$$\log(y) = \alpha + \beta(x)$$

This is equivalent to:

$$y = e^{(\alpha+\beta(x))} = e^{\alpha} + e^{\beta x}$$

Assumption :

Much like linear least squares regression (LLSR), using Poisson regression to make inferences requires model assumptions.

- Poisson Response: The response variable is a count per unit of time or space, described by a Poisson distribution.
- Independence: The observations must be independent of one another.
- Mean=Variance: By definition, the mean of a Poisson random variable must be equal to its variance.
- Linearity: The log of the mean rate, $log(\lambda)$, must be a linear function of x.

One of the most important characteristics for Poisson distribution and Poisson Regression is equidispersion, which means that the mean and variance of the distribution are equal.
If the mean (λ) is denoted by $E(X)$

$$E(X) = \lambda$$

For Poisson Regression, mean and variance are related as:

$$var(X) = \sigma^2 E(X)$$

Where σ^2 is the dispersion parameter. Since $var(X) = E(X)$ must hold for the Poisson model to be completely fit, σ^2 must be equal to 1.
When variance is greater than mean, that is called over-dispersion and it is greater than 1. If it is less than 1 than it is known as under-dispersion.

Quasi-Poisson Regression

The quasi-poisson regression is a generalization of the Poisson regression and is used when modeling an overdispersed count variable. A quasi-poisson model, assumes that the variance is a linear function of the mean, is more appropriate. One way of dealing with over-dispersion is to use the mean regression function and the variance function from the 'Poisson GLM' but to leave the dispersion parameter ϕ unrestricted. Thus, ϕ is not assumed to be fixed at 1 but is estimated from the data. This strategy leads to the same coefficient estimates as the standard poisson model but inference is adjusted for over-dispersion. Consequently, both models (quasi-poisson and sandwich-adjusted poisson) adopt the estimating function view of the poisson model and do not correspond to models with fully specified likelihoods.

2.2.4 Time Series

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.It is a 2-dimensional plot in which one axis, the time-axis, shows graduations at an appropriate scale (seconds, minutes, weeks, quarters, years), while the other axis shows the numeric values.Time series graphs can be used to visualize trends in counts or numerical values over time. Because date and time information is continuous categorical data (expressed as a range of values), points are plotted along the x-axis and connected by a continuous line.

Components of Time series:

1. Trend
   The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency.

2. Seasonal Variations
   These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months.

3. Cyclic Variations
   The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year.

4. Random or Irregular Movements
   This component is unpredictable. Every time series has some unpredictable component that makes it a random variable.

2.2.5 Bar Graph

A bar graph is a graphical representation of the categories as bars. This chart is used to show comparision between categorical variables. Each bar's height is proportional to the quantity of the category that it represents. Bar charts can be displayed with vertical columns, horizontal bars, comparative bars (multiple bars to show a comparison between values), or stacked bars (bars containing multiple types of information).

2.2.6 Pie-Chart

A pie chart is a circular chart looking like a pie divided into slices (sectors). Slices show the percentage each value contributes to a total, the area of each slice being proportional to the quantity it represents, and the circle representing 100%.

2.2.7 Line Chart

A line chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time - a time series.

2.2.8 Histogram

A histogram is basically used to represent data provided in a form of some groups.It is accurate method for the graphical representation of numerical data distribution.It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.

3 Chapter 3

3.1 Visualisation

3.1.1 Checking the Normality

The normality assumption of the variables are checked by plotting Histogram.

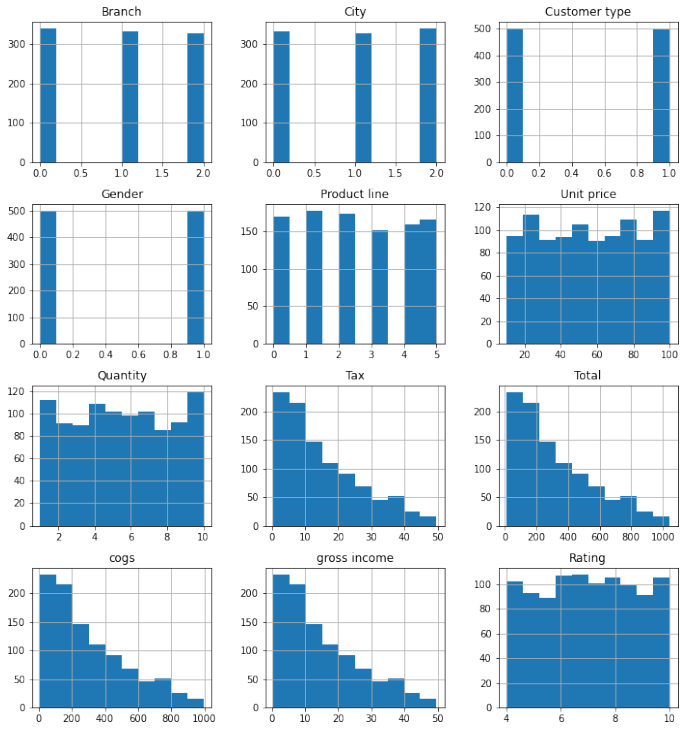

Figure 1:

From above figure we can observe that all variables are not normally distributed.

3.1.2 Correlation Analysis:

The heat-map is used to visualize the correlations among variables.

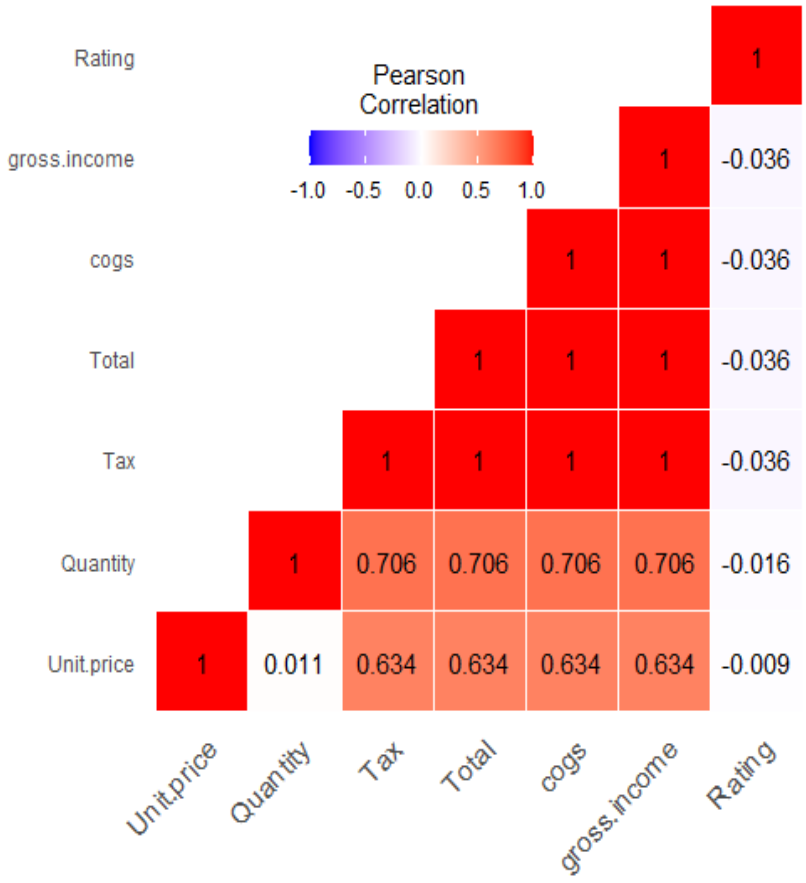

Figure 2: Correlation matrix

From heatmap we can observe that the unit price is positively correlated to cogs with 63% correlation and quantity and gross income has very high correlation of 70%. Also 'Ratings' hardly has any correlation with any other variables.

3.1.3 Branch Sales Comparison

1. **Comparing total sales of different branch.**

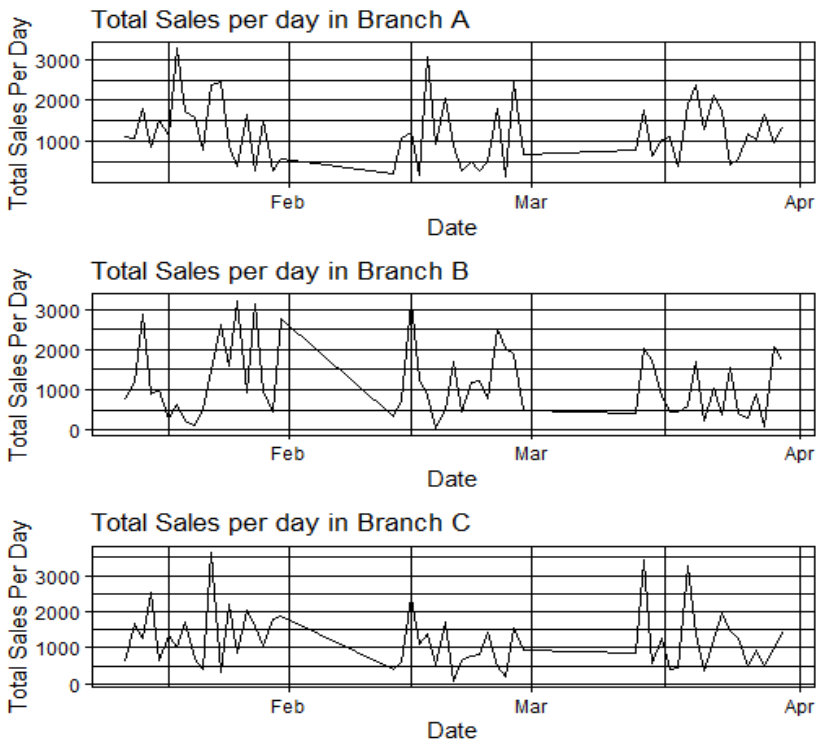

Figure 3:

- From above plot we can observe that at beginning of month (february and march) there is slight decline in sales of all three branch.
- Branch A and B attained high sales in the month of january and february, whereas branch C in the month of january and march.
- Also there is no particular trend in sales of three branch.

2. **The branch/city and the product line should be focused on for expansion.**

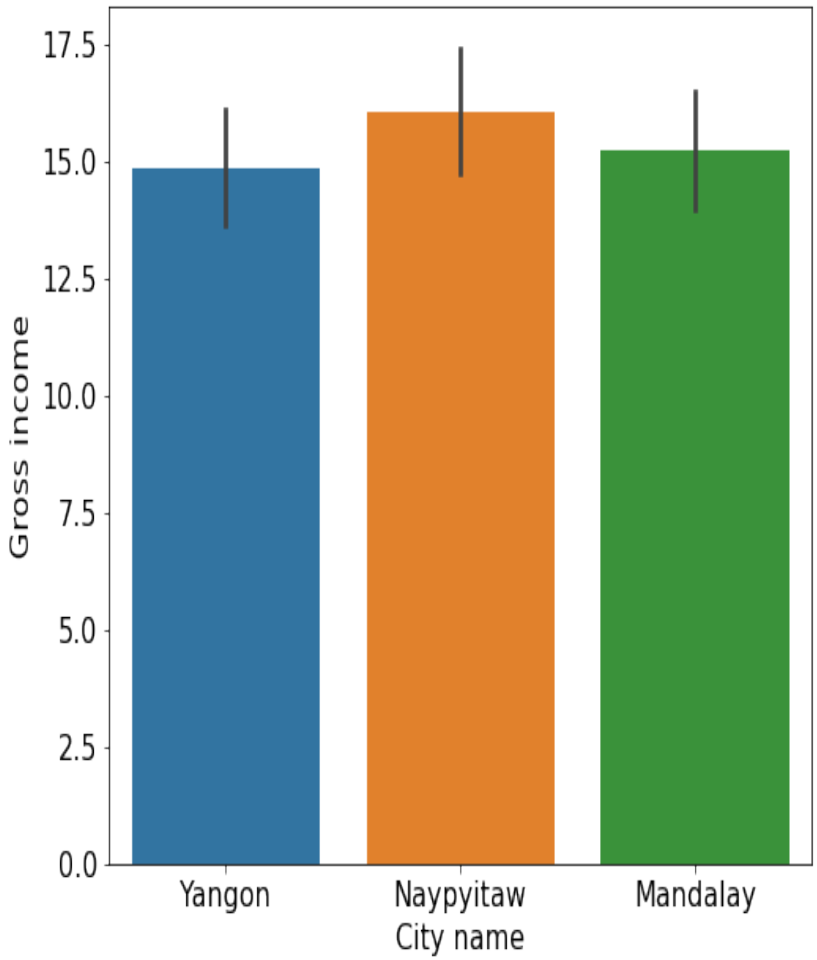

Figure 4: Gross income across branches/cities

From figure we can observe that Naypyitaw(Branch C) is the most profitable city, hence the expansion plan should be based on this city.

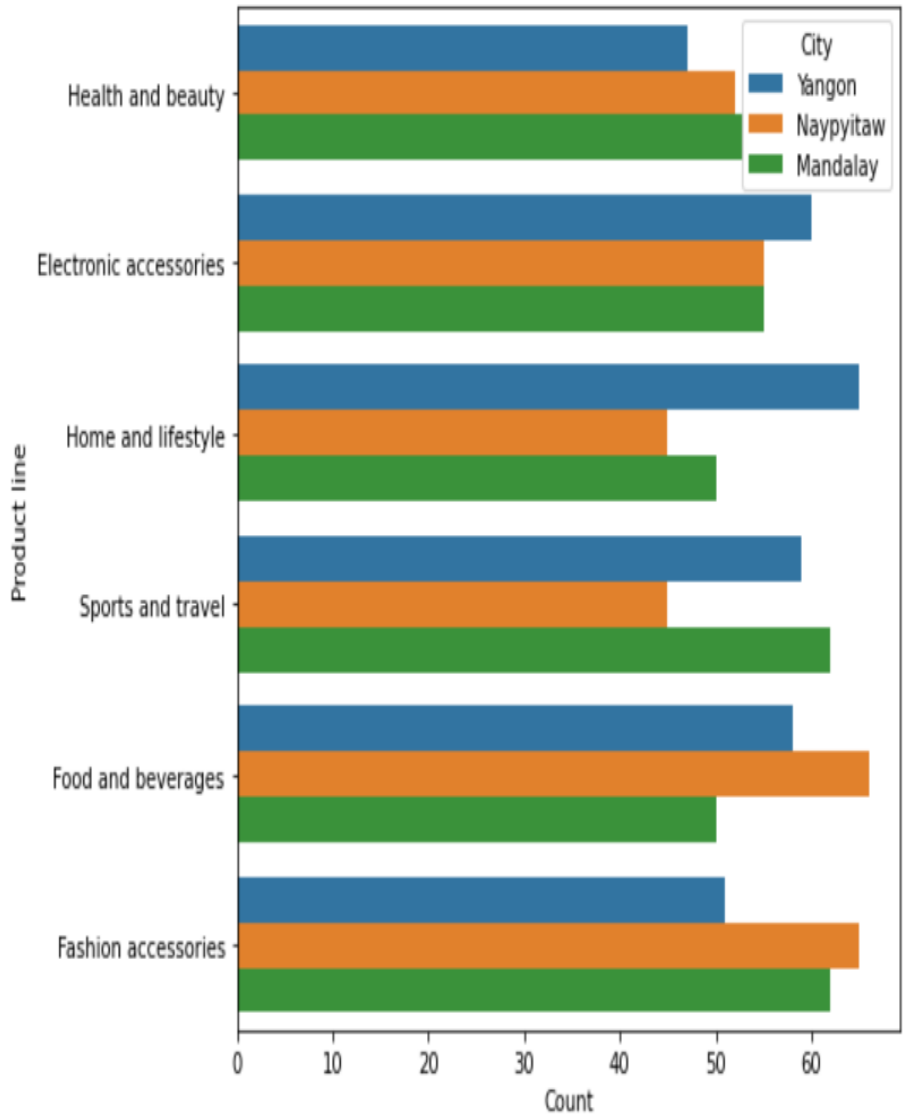

Figure 5: Sales of products across cities/branches

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

3.1.4 Customer Purchasing behaviour

Gain insights on how customer purchasing behaviour varies depending on gender, membership status (Member / Normal)

1. **Who are the leading buyers?**

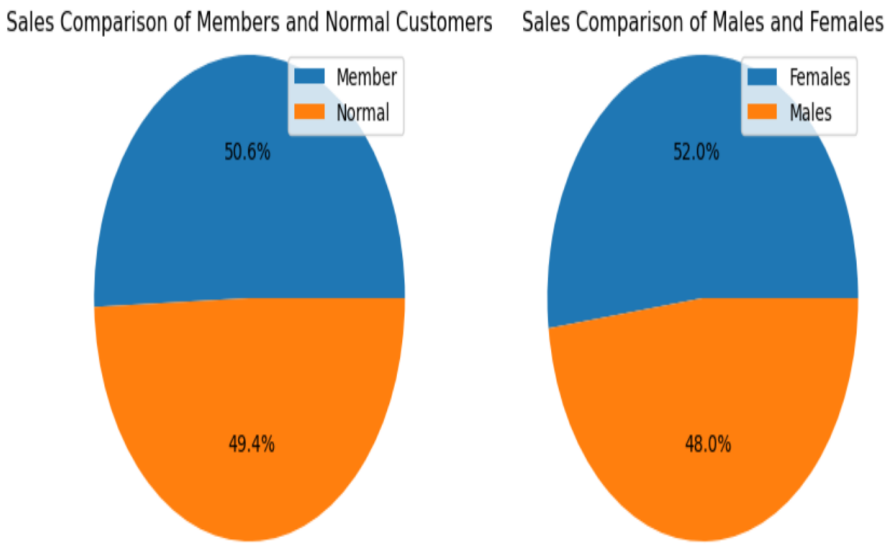

Figure 6: Sales of Respective Customer Segments

The sale from members and non-members is almost same. The sale from female customers is slightly more than the male customers

.

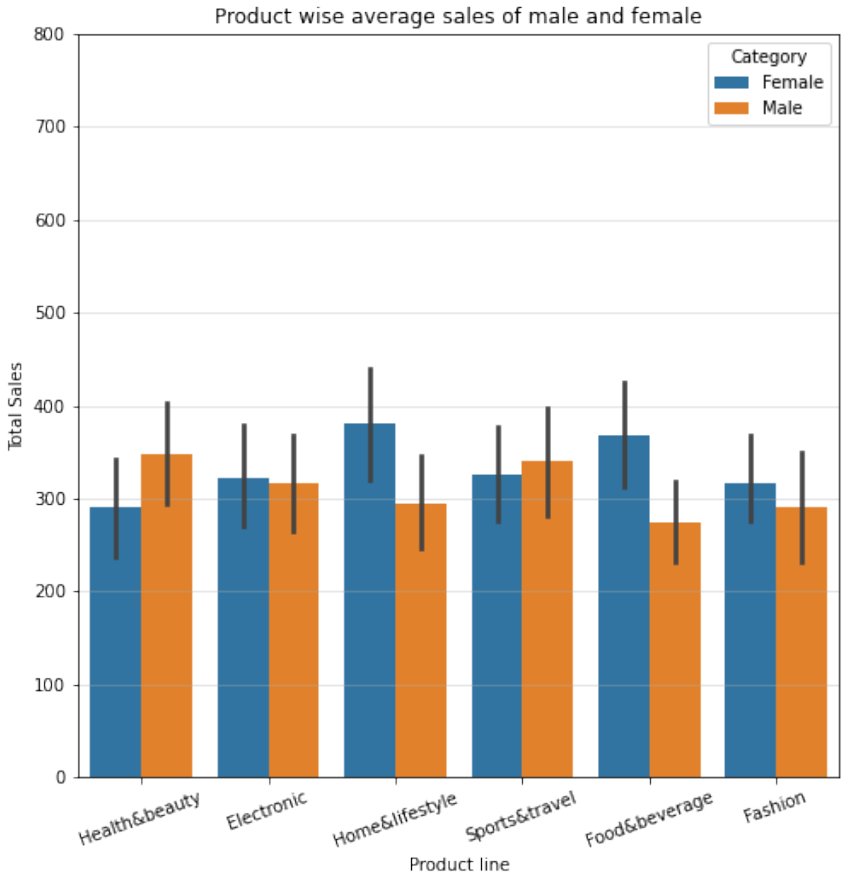

Figure 7: Sales of Respective Customer Segments

'Electronic Accessories' and 'Sports and Travel' items are almost equally purchased by males and females. Female customers contribute larger amount of sales from 'Fashion Accessories', 'Food and Beverages' and 'Home and Lifestyle'. Male customers contribute larger amount of sales from 'Health and Beauty' and 'Sprts and Travel'.

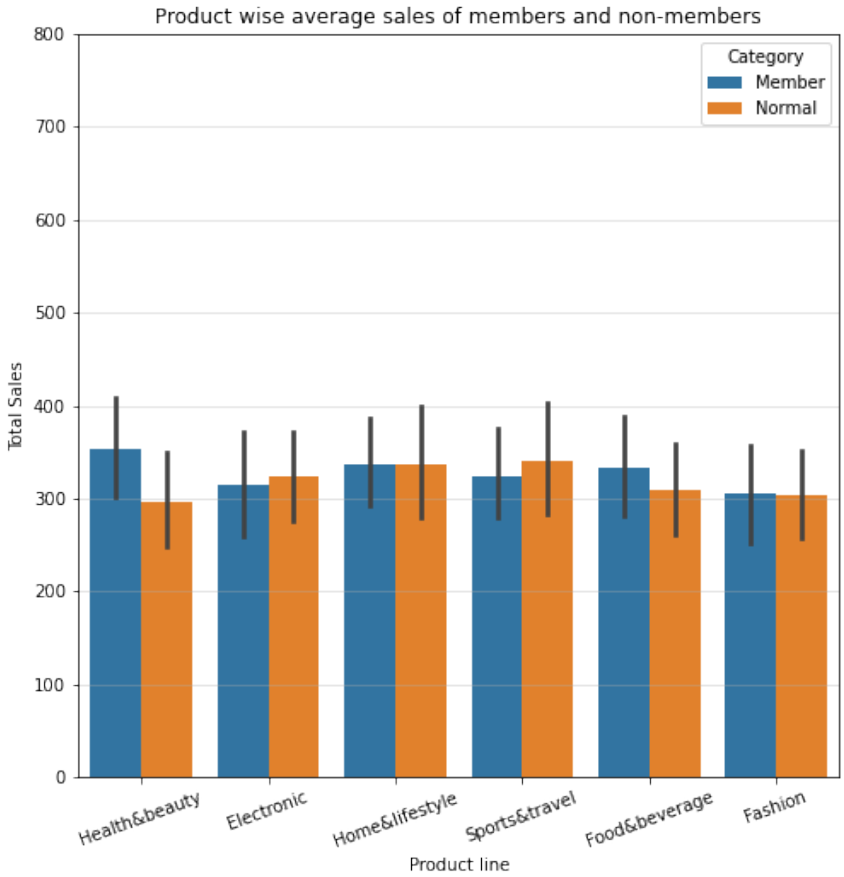

Figure 8: Sales of Respective Customer Segments

'Fashion Accessories', 'Electronic Accessories', 'Home and Lifestyle' and 'Sports and Travel' items are almost equally purchased by Members and Non-Members. Members contribute larger amount of sales from 'Health and Beauty' as well as from 'Food and Beverages'.

3.2 Summary of the Data

Table 1:

Description	Min	Max	Mean	Median	1st Quartile	3rd Quartile
City	0	2	1.008	1	0	2
Branch	0	2	0.988	1	0	2
Customer Type	0	1	0.499	0	0	1
Gender	0	1	0.499	0	0	1
Product Line	0	5	2.452	2	1	4
Unit Price	10.08	99.96	55.67	55.23	32.88	77.94
Tax	0.508	49.65	15.378	12.09	5.925	22.44
Quantity	1	10	5.51	5	3	8
Total Price	10.68	1042.65	322.97	253.85	124.42	471.35
Payment	0	2	1.001	1	0	2
COGS	10.17	993	307.59	241.76	118.5	448.9
Gross Income	0.5085	49.65	15.379	12.088	5.924	22.445
Rating	4.000	10.00	6.973	7.000	5.500	8.500

Here we can see the basic description of entire dataset.

3.3 Characteristics of Sample Customers

3.3.1 Branch Customer Count

The following table shows the customer distribution in different branch.

Table 2:

Gender	Frequency
Branch A	340
Branch B	328
Branch C	332

This dataset contains 1000 records with individual branch distribution being 340 Branch A records, 328 Branch B records, and 332 Branch C records.

3.3.2 Gender

The following table shows the distibution of gender of customer considered in this study.

Table 3:

Gender	Frequency
Male	499
Female	501

The number of male customer is 499 and the number of female customer is 501.

3.3.3 Customer Type

The following table shows the frequency of customer type considered in this study.

Table 4:

Customer type	Frequency
Member	501
Normal	499

The number of customer with member card is 501 and the number of customer without member card is 499.

3.3.4 Product line

The following table shows the distribution of products.

Table 5:

Product line	Branch			Total
	A	B	C	
Electronic accessories	60	55	55	170
Fashion accessories	51	62	65	178
Food and beverages	58	50	66	174
Health and beauty	47	53	52	152
Home and lifestyle	65	50	45	160
Sports and travel	59	62	45	166

The product line has a wide distribution with 170 Electronic accessories, 178 Fashion

accessories, 174 Food and beverages, 152 Health and beauty, 160 Home and lifestyle, and 166 Sports and travel records.

3.4 Chi-Square Test of Independence

1. **Affect of payment method on the ratings given by respondents using chi-square test of independence.**

   The following table shows payment type available and ratings given by respondents.

   Table 6:

	Rating		
Payment method	High	medium	low
Cash	116	119	109
Credit card	105	94	112
Ewallet	120	117	108

   The hypothesis are as follows:

   H_0 : The variable payment method and rating are independent.

   H_1 : The variables payment method and rating are dependent.

   The values obtained are as follows :

   - Chi-square test statistic: 2.486
   - Degrees of freedom : 4
   - p-value : 0.647

   We can observe that the obtained p-value is greater than 0.05. Hence we do not reject H_0. Thus, the variable 'payment method' and 'rating' are independent.

   Therefore, no relationship exists between the payment method available and the ratings given by respondents.

2. **Affect of customer gender on their product preference using chi-square test of independence.**

   The following table shows the gender and product category.

Table 7:

Gender	Product line					
	Electronic	Fashion	Food & Beverage	Health & Beauty	Home & Lifestyle	Sports & Travel
Female	84	96	90	64	79	88
Male	86	82	84	88	81	78

The hypothesis are as follows:

H_0 : The variable gender and product line are independent.

H_1 : The variables gender and product line are dependent.

The values obtained are as follows :

- Chi-square test statistic: 5.744
- Degrees of freedom : 5
- p-value : 0.3319

We can observe that the obtained p-value is greater than 0.05. Hence we do not reject H_0. Thus, the variable 'payment method' and 'rating' are independent.

Therefore, no relationship exists between the gender and their product preference.

3.5 Poisson Regression Model

To check whether the unit price and Tax has any affect on the quantity of product being purchased by customer.

Here we consider the study variable as 'Quantity' and predictor as 'Unit price' and 'Tax'.

The hypothesis are as follows:

H_0 : The variable 'Unit price' and 'Tax' has no affect on the quantity of product being purchased.

H_1 : The variables 'Unit price' and 'Tax' has affect on the quantity of product being purchased.

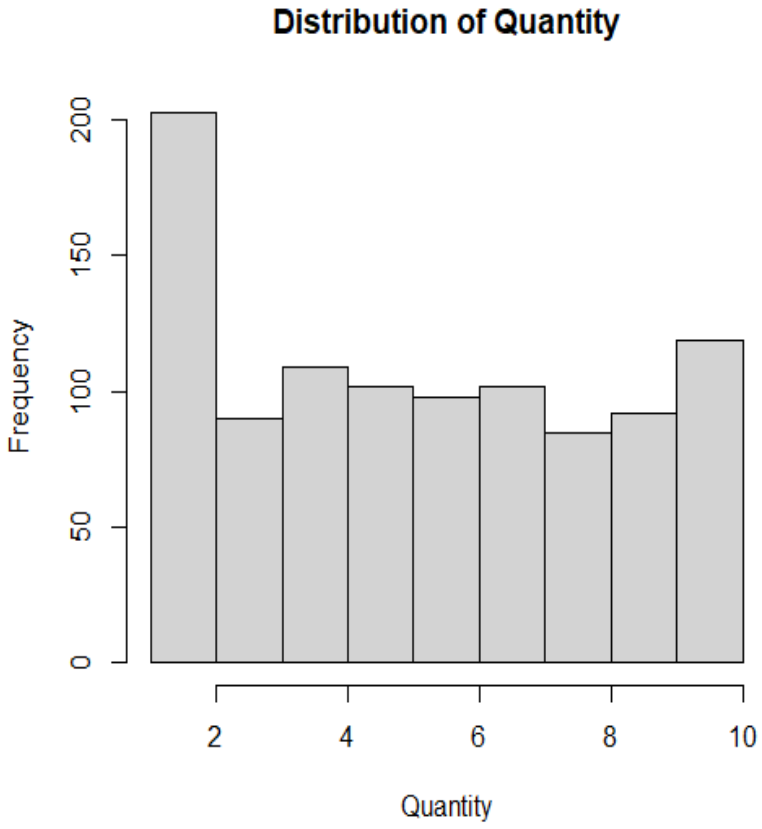

Figure 9:

The data is not in the form of a bell curve.
The following table shows the mean and variance of response variable(i.e Quantity).

Table 8:

Mean	Variance
5.51	8.546

The variance is greater than the mean, which suggests that we will have over-dispersion in the model.

Result obtained by fitting poisson model:

Table 9:

Deviance Residuals				
Min	1Q	Median	3Q	Max
-2.1973	-0.3734	0.1391	0.3823	1.2336

Table 10:

Coefficients:				
	Estimate	Std. Error	z value	Pr> $\|t\|$
Intercepts	1.7854	0.0315	56.67	<2e-16
Unit price	-0.0196	0.0008	-23.24	<2e-16
Tax	0.0574	0.0016	34.12	<2e-16

- Null deviance: 1722.36 on 999 degrees of freedom
- Residual deviance: 407.29 on 997 degrees of freedom

We can see that residual deviance is less than degrees of freedom. Thus we have under-disperion. Re-fit model using quasi poisson.

Result obtained by fitting quasipoisson model:

Table 11:

Coefficients:				
	Estimate	Std. Error	t value	Pr> $\|t\|$
Intercepts	1.7927	0.0217	82.513	<2e-16
Unit price	-0.0196	0.0005	-38.838	<2e-16
Tax	0.0573	0.001	56.983	<2e-16

- Null deviance: 1722.36 on 999 degrees of freedom

- Residual deviance: 407.29 on 997 degrees of freedom

From the table, we can observe that the p-value for 'Unit price' and 'Tax' variable is less than 0.05. Hence, we reject H_0.

Therefore 'Unit price' and 'Tax' effects quantity of product the customer purchase.

Table 12:

Predictor(x)	exp(x)
Unit.price	0.9805
Tax	1.058

Thus, each addition in unit price is associated with a decrease in the quantity of product being purchased by a factor of 2%.

3.6 Conclusion

1. From the heatmap, 'unit price' and 'COGS' is highly correlated with 63% correlation. Also quantity and gross income are very high positively correlated with correlation of 70%.

2. The customer ratings is not related to any of the variable, it hardly as any correlation with other variable.

3. On comparing total sales trend using line plot of three branches, it is observed that there is slight decline in sales at the beginning of month february and march.

4. Branch A and B attained high sales in the month of january and february, whereas branch C in the month of january and march.

5. There is no trend in total sales of three branch.

6. All three branch is doing well interms of sales. Based on gross income the most profitable branch was C i.e, 'Naypyitaw'. Also 'Fashion accessories' and 'Food and beverages' are the most sold product in Naypyitaw.

7. Females are the leading buyers compared to male.

8. Female customers contribute larger amount of sales from 'Fashion Accessories', 'Food and Beverages' and 'Home and Lifestyle'. Male customers contribute larger amount of sales from 'Health and Beauty'.

9. Non-Members contribute larger amount of sales from 'Health and Beauty' and Members contribute larger amount of sales from 'Food and Beverages'.

10. According to chi-square test of independence no relationship exists between the payment method and the ratings given by respondents.

11. According to chi-square test there is no relationship between the gender and their product preference.

12. According to poisson regression 'Unit price' and 'Tax' effects the 'quantity'. Each addition in unit price is associated with a decrease in the quantity of product being purchased by a factor of 2%.

4 Chapter 4

4.1 Conclusion

1. Gross income and quantity are related to each other. In business point of view, when more number of quantity is sold it leads to increase in gross income.

2. From business point of view customer ratings depends on satisfaction they get from branch service also the product price. But in our study, ratings hardly as any relation with unit price, tax, payment choice and so on.

3. Based on our study, customer product preference doesn't depends on their gender. Hence we cannot assume that women or men would buy certain products belonging to certain category.

4. From business point of view, customer considers price of the products before purchasing it in large quantity. That is true in our study.

5. In our study, when comparing purchasing behaviour based on gender, female are the leading buyers compared to male and they contribute larger amount of sales from 'Fashion Accessories', 'Food and Beverages' and 'Home and Lifestyle'. Thus, our target customers are females and supermarket must consider to provide discounts in these product line.

6. All three branch is doing well interms of sales. The most profitable branch was C i.e, 'Naypyitaw'.

4.2 Summary

A project entitled 'An Exploratory Analysis of Supermarket Sales in Myanmar' has been done. The dataset which is considered has been collected from website 'Kaggle'. The dataset consists of historical sales data of a supermarket company from 3 different branches over 3 months from Jan-March 2019.

'Python' and 'R' language has been used to carry out the analysis and the interpretation of the data. Also some statistical methods like correlation(heatmap), chi-square test of independence, poisson regression, time series, barplot and pie chart.

From the results obtained, we came to know that the customer ratings is not related to any of the variable. Also, there is no relationship between the gender and their product preference. The city chosen for expansion should be Naypyitaw . Opening the store in Naypyitaw with a full-backed inventory of goods in product categories - Food and Beverages, Fashion Assesories and Electronics leads in good income. The company should improve service in Electronics domain. The prime target for marketing should be women, persausive ads for offering memberships, more to women.

5 Bibliography

- https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/
- https://www.koreascience.or.kr/article/JAKO201317054958265.pdf
- https://www.newworldencyclopedia.org/entry/Supermarket
- http://www.ijrbsm.org/papers/v5-i5/4.pdf
- https://www.cmgconsulting.com/post/what-is-customer-analysis
- https://www.researchgate.net/publication/343569312_Study_on_the_Consumer_Preference_and_Perception_of_Supermarket_Chain_-Case_of_Dmart
- https://www.researchgate.net/publication/24103353_Retail_Sales_A_Study_of_Pricing_Behavior_in_Supermarkets
- https://www.kaggle.com/vaishnavi2209/data-visualizing-of-super-market-r

6 Appendix

Python Script :

1. **Library**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as sp
from scipy.stats import chi2
import statsmodels.api as sm
from statsmodels.iolib.summary2 import summary_col
```

2. **Importing dataset :**

```
df= pd.read_csv(r'D:\project\original.csv')

# Droping columns which are not required

df2=df.drop(["Invoice ID","gross margin percentage","City","cogs",
  "Customer type","Time"],axis=1)

# Renaming entries of Product line column

df["Product line"].replace({'Health and beauty':'Health&beauty',
                            'Electronic accessories':'Electronic',
                            'Home and lifestyle':'Home&lifestyle',
                            'Sports and travel':'Sports&travel',
                            'Food and beverages':'Food&beverage',
                            'Fashion accessories':'Fashion'}
                           , inplace=True)
```

Branch	City	Customer type	Gender	Product line	Unit price
A	Yangon	Member	Female	Health&beauty	74.69
C	Naypyitaw	Normal	Female	Electronic	15.28
A	Yangon	Normal	Male	Home&lifestyle	46.33

Quantity	Tax	Total	cogs	gross income	Rating
7	26.1415	548.9715	522.83	26.1415	9.1
5	3.8200	80.2200	76.40	3.8200	9.6
7	16.2155	340.5255	324.31	16.2155	7.4

3. **Checking distribution of each variable**
   `df.hist(figsize=(12,13))`

4. **Plotting barplot of 'Gross income' across 'Branches/cities'.**

```
plt.figure(figsize=(8,7))
sns.barplot(x='City',y='gross income',data=df2)
plt.xlabel('City name',fontsize='16')
plt.xticks(fontsize='16')
plt.ylabel('Gross income',fontsize='16')
plt.yticks(fontsize='16')
plt.savefig(r'D:\project\bar1.png')

# Sales of products across cities/branches

plt.figure(figsize=(7,7))
sns.countplot(y ='Product line', hue = "City", data = df2);
plt.xlabel('Count')
plt.savefig(r'D:\project\bar4.png')
```

5. **Customer Purchasing behaviour:**

   Grouping dataset based on customer type:

```
customer_type_df = df.groupby('Customer type').sum()
member_sales = customer_type_df.iloc[0]['Unit price']
normal_sales = customer_type_df.iloc[1]['Unit price']
member_normal_sales = [member_sales, normal_sales]
member_normal_legend = ['Member', 'Normal']
customer_type_df
```

Customer type	Unit price	Quantity	Tax	Total	cogs	gross income	Rating
Member	28159.70	2785	7820.164	164223.444	156403.28	7820.164	3477.1
Normal	27512.43	2725	7559.205	158743.305	151184.10	7559.205	3495.6

Grouping dataset based on customer gender:

```
customer_gender_df = df2.groupby('Gender').sum()
female_sales = customer_gender_df.iloc[0]['Total']
male_sales = customer_gender_df.iloc[1]['Total']
female_male_sales = [female_sales, male_sales]
female_male_legend = ['Females', 'Males']
customer_gender_df.head()
```

Gender	Unit price	Quantity	Tax	Total	cogs	Gross income	Rating
Female	27687.24	2869	7994.425	167882.925	159888.50	7994.425	3489.2
Male	27984.89	2641	7384.944	155083.824	147698.88	7384.944	3483.5

Pie chart

```
fig, axs = plt.subplots(1, 2, figsize=(9,9),dpi=100)
axs[0].pie(member_normal_sales, radius=1.2, autopct = "%0.1f%%")
axs[0].set_title('Sales Comparison of Members and Normal Customers')
axs[0].legend(member_normal_legend, loc = 'upper right')
axs[0].set_aspect('equal')

axs[1].pie(female_male_sales, radius=1.2, autopct = "%0.1f%%")
axs[1].set_title('Sales Comparison of Males and Females')
axs[1].legend(female_male_legend, loc = 'upper right')
axs[1].set_aspect('equal')

plt.subplots_adjust(left=0.1,
                    bottom=0.1,
                    right=0.9,
                    top=0.9,
                    wspace=0.4,
                    hspace=0.4)

plt.savefig(r'D:\project\pie1.png')
```

Bar plot on product wise distribution

```
# 1) Based on customer gender

plt.figure(figsize=(8,8))
plot=sns.barplot(x='Product line',y='Total',hue='Gender',data=df2)
plot.set_title('Product wise average sales of male and female')
plt.legend(loc="upper right",title="Category")
```

```
plt.ylabel("Total Sales")
plt.ylim(0,800);
plt.xticks(rotation=20)
plt.grid(axis='y',alpha=0.4)
plt.savefig(r'D:\project\bar7.png')

# 2) Based on customer type

plt.figure(figsize=(8,8))
plot=sns.barplot(x='Product line',y='Total',hue='Customer type',data=df2)
plot.set_title('Product wise average sales of members and non-members')
plt.legend(loc="upper right",title="Category")
plt.ylabel("Total Sales")
plt.ylim(0,800);
plt.xticks(rotation=20)
plt.grid(axis='y',alpha=0.4)
plt.savefig(r'D:\project\bar8.png')
```

Data Description

```
df2.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Branch	1000.0	0.988	0.818	0.000	0.00	1.000	2.000	2.00
City	1000.0	1.008	0.820	0.000	0.00	1.000	2.000	2.00
Customer type	1000.0	0.499	0.500	0.000	0.00	0.000	1.000	1.00
Gender	1000.0	0.499	0.500	0.000	0.00	0.000	1.000	1.00
Product line	1000.0	2.452	1.715	0.000	1.00	2.000	4.000	5.00
Unit price	1000.0	55.67	26.49	10.08	32.8	55.23	77.90	99.96
Quantity	1000.0	5.510	2.923	1.000	3.00	5.000	8.000	10.00
Tax	1000.0	15.37	11.70	0.508	5.92	12.08	22.40	49.65
Total	1000.0	322.9	245.8	10.67	124.4	253.8	471.3	1042.65
cogs	1000.0	307.5	234.1	10.17	118.4	241.7	448.9	993.00
gross income	1000.0	15.37	11.70	0.508	5.925	12.08	22.44	49.65
Rating	1000.0	6.973	1.719	4.000	5.500	7.000	8.500	10.00

Frequency table

```
# GENDER
dfm = df.groupby('Gender').count()
dfm[['Invoice ID']]

# BRANCH COUNT
dfb = df.groupby('Branch').count()
dfb[['Invoice ID']]

#PRODUCT COUNT
dfp =  df.groupby(['Branch','Product line'])['Product line'].count()
dfp[['Invoice ID']]

# CUSTOMER TYPE COUNT

dfc = df2.groupby(['Customer type']).count()
dfc[['Invoice ID']]
```

Chi-square test of independence

- Affect of payment method on ratings

```
Rating = pd.qcut(df2['Rating'], 3, labels = ['High', 'Medium', 'Low'])
data_t1 = pd.crosstab(df2['Payment'],Rating)

chi2, pval, dof, expected = sp.chi2_contingency(data_t1)
print("ChiSquare test statistic: ",chi2)
print("p-value                 : ",pval)
print("Degrees of freedom      : ",dof)

output:

ChiSquare test statistic:  2.486485891332456
p-value                 :  0.647057327588838
Degrees of freedom      :  4
```

- Affect of gender on their product preference

```
data_t2 = pd.crosstab(df2['Gender'],df2['Product line'])
chi2, pval, dof, expected = sp.chi2_contingency(data_t2)
print("ChiSquare test statistic: ",chi2)
print("p-value                 : ",pval)
print("Degrees of freedom      : ",dof)
```

```
output:
ChiSquare test statistic:  5.7444558595826445
p-value                 :  0.33188385805539106
Degrees of freedom      :  5
```

R script

1. **library**

```
library(lubridate)
library(ggplot2)
library(gridExtra)
```

2. **Read csv file**

```
> data1=read.csv("original.csv",header=TRUE)
> head(df)
  Branch City Customer.type Gender Product.line Unit.price Quantity
1      0    2             0      0            3      74.69        7
2      2    1             1      0            0      15.28        5
3      0    2             1      1            4      46.33        7
4      0    2             0      1            3      58.22        8
5      0    2             1      1            5      86.31        7
6      2    1             1      1            0      85.39        7
      Tax    Total   cogs gross.income Rating
1 26.1415 548.9715 522.83      26.1415    9.1
2  3.8200  80.2200  76.40       3.8200    9.6
3 16.2155 340.5255 324.31      16.2155    7.4
4 23.2880 489.0480 465.76      23.2880    8.4
5 30.2085 634.3785 604.17      30.2085    5.3
6 29.8865 627.6165 597.73      29.8865    4.1
```

3. **Plotting total sales of branch**

```
> data_removed <- data1[,-1] # removing first column

#changing date format

> data_removed$Date <- as.Date(data_removed$Date, "%m/%d/%y")
> year(data_removed$Date) <- 2019
```

```
> total_sales_per_day <- data.frame(xtabs(formula=Total~Date,
+ data=data_removed))
> total_sales_per_day$Date <- as.Date(total_sales_per_day$Date)

> A <- data_removed %>% filter(Branch == "A")
> total_A <- data.frame(xtabs(formula = Total~Date, data = A))
> total_A$Date <- as.Date(total_A$Date)

> B <- data_removed %>% filter(Branch == "B")
> total_B <- data.frame(xtabs(formula = Total~Date, data = B))
> total_B$Date <- as.Date(total_B$Date)

> C <- data_removed %>% filter(Branch == "C")
> total_C <- data.frame(xtabs(formula = Total~Date, data = C))
> total_C$Date <- as.Date(total_C$Date)

# creating plot

> plot1 <- ggplot(data = total_A, mapping = aes(x = Date, y = Freq))+
 geom_line()+
+ theme_linedraw()+ ggtitle("Total Sales per day in Branch A")
+ xlab("Date")+ ylab("Total Sales Per Day")

> plot2 <- ggplot(data = total_B, mapping = aes(x = Date, y = Freq))+
 geom_line()+ theme_linedraw()+
ggtitle("Total Sales per day in Branch B")+
xlab("Date")+ ylab("Total Sales Per Day")

> plot3 <- ggplot(data = total_C, mapping = aes(x = Date, y = Freq))+
 geom_line()+ theme_linedraw()+
 ggtitle("Total Sales per day in Branch C")+
 xlab("Date")+ ylab("Total Sales Per Day")

> grid.arrange(plot1, plot2, plot3)
$
```

4. **Poisson Regression model**

Affect of unit price and tax on quantity
Plotting histogram on variable 'Quantity'

```
> hist(df$Quantity)
```

Mean and variance of variable 'Quantity'

```
> mean(df$Quantity)
[1] 5.51

> var(df$Quantity)
[1] 8.546446
$
```

General poisson regression model

```
> poimodel=glm(Quantity~Unit.price+Tax,df,
family=poisson(link="log"))
> summary(poimodel)

Call:
glm(formula = Quantity ~ Unit.price + Tax,
 family = poisson(link = "log"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1973  -0.3734   0.1391   0.3823   1.2336

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.7854964  0.0315068   56.67   <2e-16 ***
Unit.price  -0.0196082  0.0008437  -23.24   <2e-16 ***
Tax          0.0574454  0.0016835   34.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1722.36  on 999  degrees of freedom
Residual deviance:  407.29  on 997  degrees of freedom
AIC: 3804.6

Number of Fisher Scoring iterations: 4
```

Quassi-poisson regression model

```
> poismodel2=glm(Quantity~Unit.price+Tax,data=df,
family=quasipoisson(link="log"))
```

```
> summary(poismodel2)

Call:
glm(formula = Quantity ~ Unit.price + Tax,
 family = quasipoisson(link = "log"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1973  -0.3734   0.1391   0.3823   1.2336

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.7854964  0.0188092   94.93   <2e-16 ***
Unit.price  -0.0196082  0.0005037  -38.93   <2e-16 ***
Tax          0.0574454  0.0010050   57.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.3563948)

    Null deviance: 1722.36  on 999  degrees of freedom
Residual deviance:  407.29  on 997  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```