PIMPRI CHINCHWAD EDUCATION TRUST'S

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING



**Department of Information Technology**

**Report of IE-2 activity**

**Subject: Statistical Data Analysis using R**

**Academic Year 2022-23**

**Semester I**

**Submitted by**

| Name of the student | Roll number | Branch |
| --- | --- | --- |
| Viraj Balasaheb Patil | SYITB96 | Information Technology |

**Submitted to : Mrs. A.Vaze**

**Date:**

**Sign:**

**Project Title : Comparative Analysis of 3 Supermarkets Sale in The Year 2019**

**Data used:** Dataset used - supermarket sales

**Details of the data :** Data contains 1000 observation and 17 variables as follows

1. Invoice ID:Computer generated sales slip invoice identification number

2. Branch:Branch of Supercenter (A,B,C)

3.City: Location of supercenters (Yangon , Mandalay , Naypyidaw)

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

5.Gender: Gender type of customer

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

7.Unit price: Price of each product in $

8. Quantity: Number of products purchased by customer

9. Tax: 5% tax fee for customer buying

10. Total: Total price including tax

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

12.Day:Day of purchase

13.Month:Month of purchase

14.Time: Purchase time (10am to 9pm)

15.Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)

16.COGS: Cost of goods sold

17.Gross margin percentage: Gross margin percentage

18.Gross income: Gross income

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

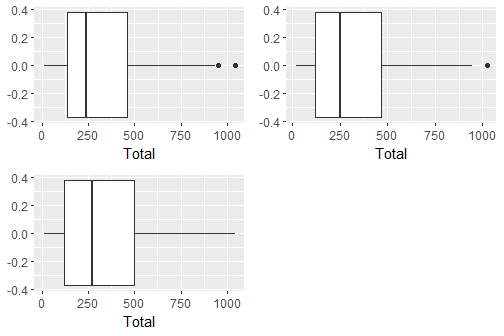
Acknowledgements

**Problem Statement 1:What is the overall total of sales trend for each of the 3 supermarkets ?**

The data is grouped by Branches and average cost of goods sold is calculated, results are as follows,

|  | **Branch** | **average (Total)** | **Average(Quantity)** |
| --- | --- | --- | --- |
| 1 | A | 312.3540 | 340 |
| 2 | B | 319.8725 | 332 |
| 3 | C | 337.0997 | 328 |

It shows that, average of total is less for supercenter A and average total is high for supercenter C . It can be visualise by boxplot as follows :



grid info - (1,1) - A

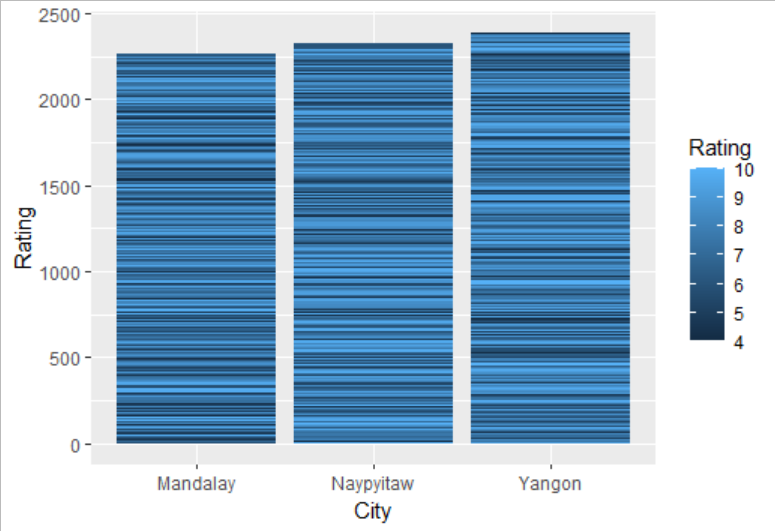
(1,2) - B

(2.1) - C

It clearly shows that supercenter A has some outside values of total and its average total is less than other supercenters . Supercenter C has slightly higher total , therefore supercenter C is performing better in the sense of total . Supercenter A has highest sales in the sense of quantity and Supercenter C has lowest .

**Problem Statement 2: Find the trend of rating between cities and whether the customers are satisfied or not ?**

we can understand trend of rating between cities from following graph :

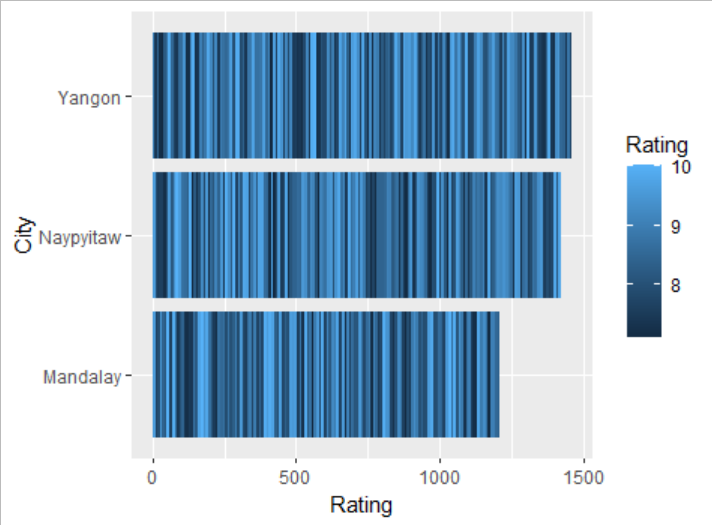


Yangon city has highest rating than other cities , Naypyitaw has intermediate and Mandalay has lowest rating but it is in the sense of total rating

p=ggplot(salesF,aes(x=City,y=Rating,fill=Rating))+geom\_bar(stat="identity")

**print**(p)

for finding customer satisfaction we can refer following graph :



From the above graph we can see that Yangen and Naypyidaw have higher customer satisfaction than Mandalay . This means that supercenter A has a higher customer satisfaction rating.

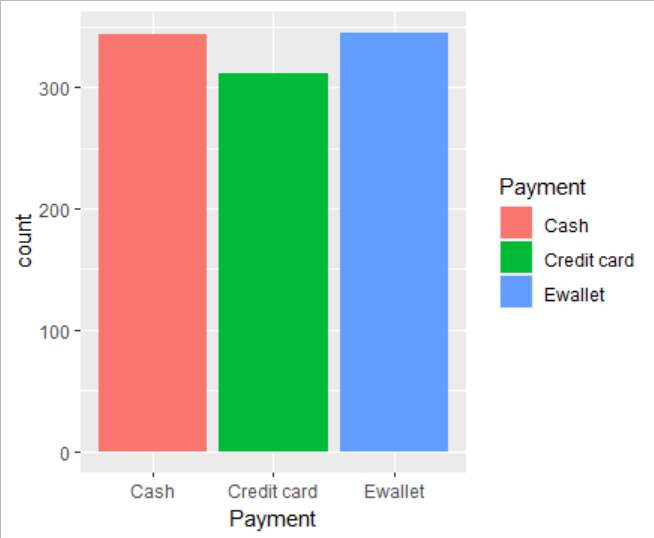
AB=**filter**(salesF,Rating>7)

**View**(AB)

p1=ggplot(AB,aes(x=Rating,y=City,fill=Rating))+geom\_bar(stat="identity")

**print**(p1)

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

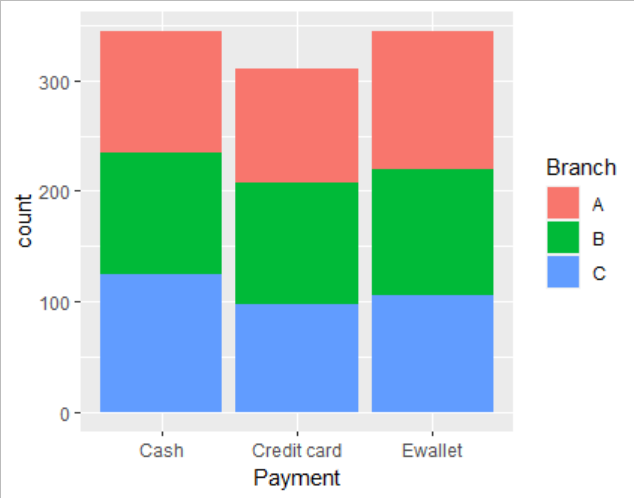
****

From the above graph we can see that cash and E Wallet payment methods are popular methods used by customers. The most popular payment method is the E wallet . Credit cards are used by less customers.

It states that in these cities cash and E wallets are mostly used. From this we can assume that citizens do not have credit cards or they don't want to use them because of credit card bills.

BC=ggplot(salesF,aes(Payment,fill=Payment))+geom\_bar()

**print**(BC)

****

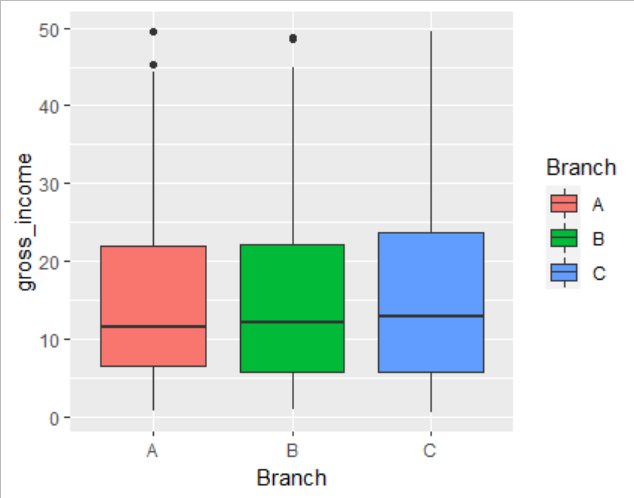
We can see that in Supercentre C people used cash payments more than other payments .In Supercentre A , mostly the E wallet payments method is used.Credit card payments are on the lower side .

**BC=ggplot(salesF,aes(Payment,fill=Branch))+geom\_bar()**

**print(BC)**

**Problem statement 4: Which branch is the most profitable?**

we can find most profitable branch (supercenter) by using boxplot :

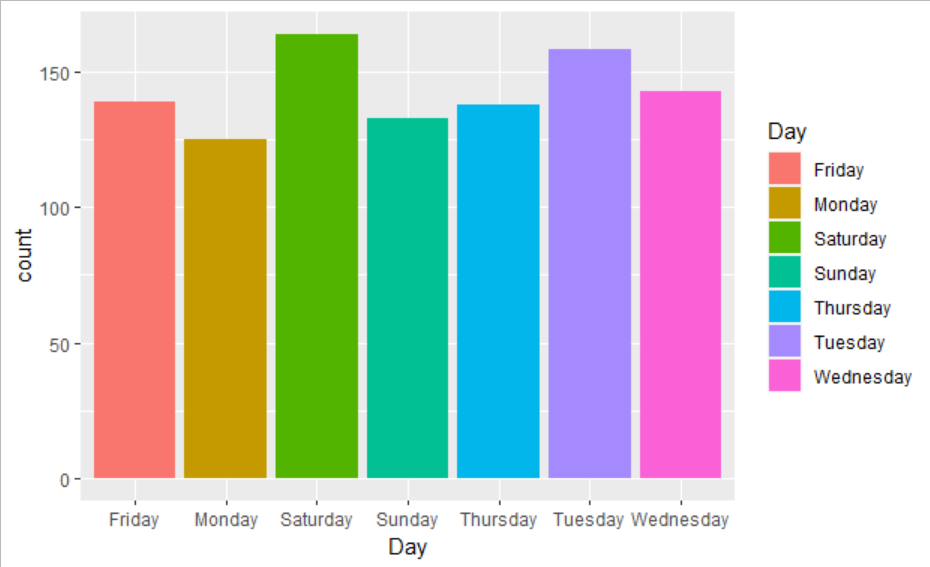
****

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. Naypyidaw is the most profitable branch in terms of gross income.

**BA=ggplot(salesF,aes(Branch,gross\_income,fill=Branch))+geom\_boxplot()**

**print(BA)**

**Problem statement 5: Which day of the week has maximum sales?**

****

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 the start of the working week.

**BD=ggplot(salesF,aes(Day,fill=Branch))+geom\_bar()**

**print(BD)**

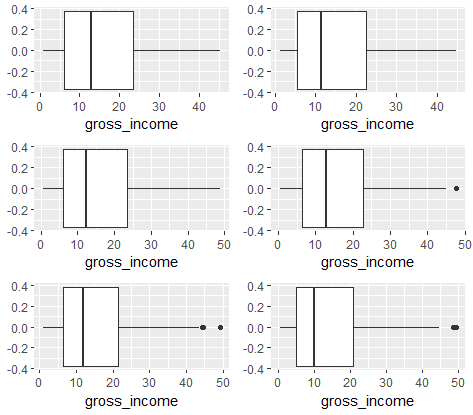
**Problem Statement 6:Which product line gives the highest gross income ?**

The data is grouped by Product lines and mean of gross income and standard deviation of gross income is calculated, results are as follows,

|  | **Product line** | **Mean (Gross income)** | **Standard Deviation (Gross income)** |
| --- | --- | --- | --- |
| 1 | Electronic accessories | 15.22060 | 11.71170 |
| 2 | Fashion accessories | 14.52806 | 11.59829 |
| 3 | Food and beverages | 15.36531 | 11.76942 |
| 4 | Health and beauty | 15.41157 | 11.31137 |
| 5 | Home and lifestyle | 16.03033 | 12.12384 |
| 6 | Sports and travel | 15.81263 | 11.82740 |

Above table shows that Home and lifestyle product line have the highest mean Gross income . but it has high standard deviation also. Health and beauty has lowest standard deviation , gross income by Health and beauty product line is less deviated.

It can be visualise by Boxplot as follows :



Grid info - (1,1) - Electronic accessories

(1,2) - Fashion accessories

(2,1) - Food and beverages

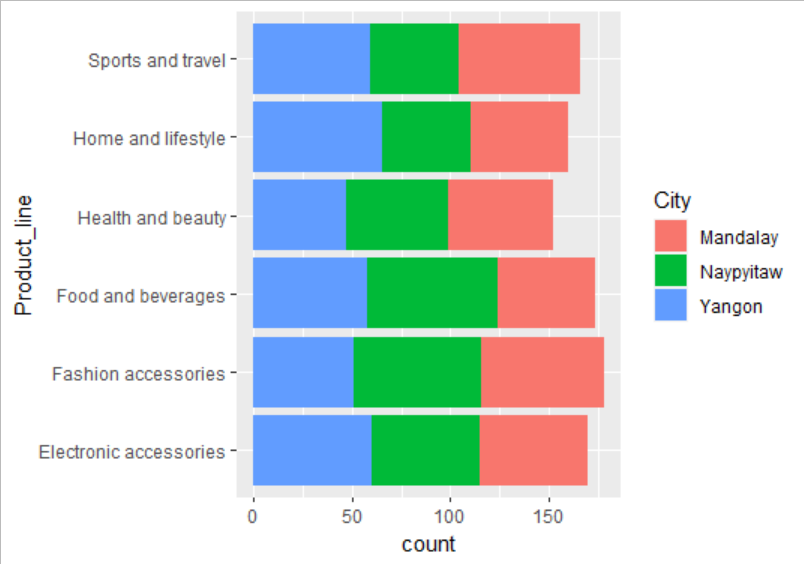
(2,2) - Health and beauty

(3,1) - Home and lifestyle

(3,2) - Sports and travel

Health and beauty, Home and lifestyle and Sports and travel have some outliers . It seems that gross income from Electronic accessories, Food and beverages and Home and lifestyle is slightly better than than other product lines.

**Problem Statement 7 : Find the product line and branch trend.**

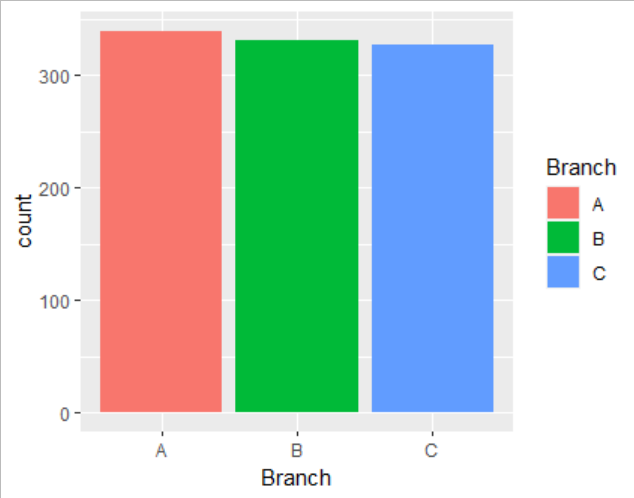
****

* Supercenter B leads Sports and Travels and also Health and Beauty.
* Supercenter C leads Foods and Beverage and also Fashion and accessories.
* Supercenter A leads Home and Lifestyle and also Electronic accessories.

**ABC=ggplot(salesF, aes(y=Product\_line))+geom\_bar(aes(fill = Branch))**

**print(ABC)**

**Problem Statement 8 : which branch is most busiest ?**

****

from above graph supercenter A is the busiest supercenter .

**PlotAB = ggplot(salesF, aes(x = Branch))+geom\_bar(aes(fill = Branch))**

**print(PlotAB)**

**Overall Conclusion:**

**1.** From 3 cities,Branch C has a higher overall total than other branches , i.e.Naypyidaw has higher sales than other cities.

**2.** Customers rating is less uniform and its mean is around 7. Supercenter A has a slightly higher rating than other supercenters. So,we can conclude that in supercenter A customers have more satisfaction.

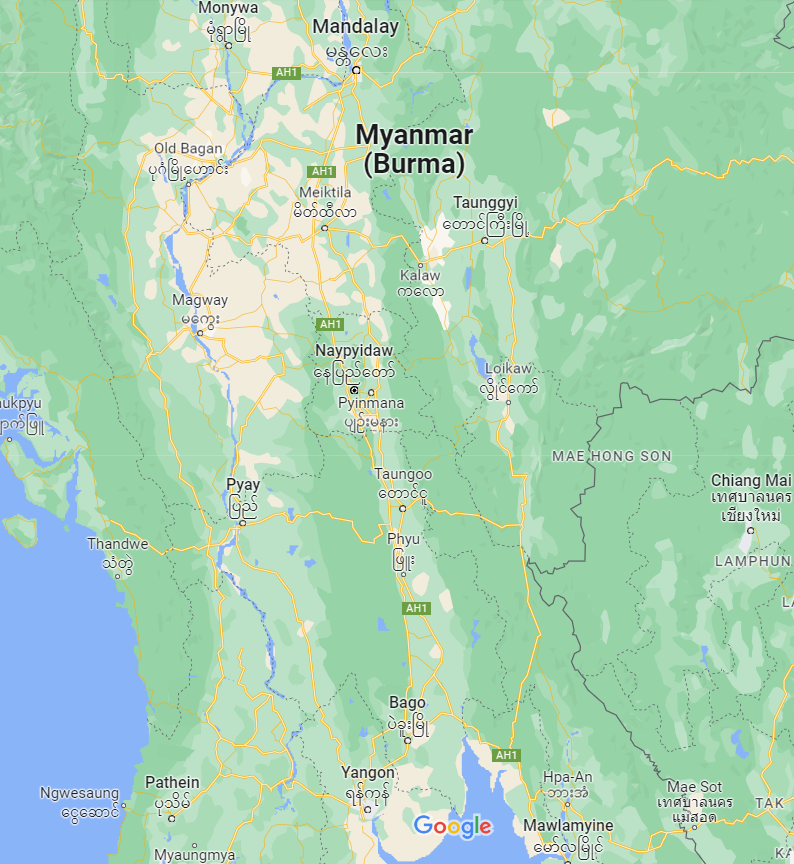
**3.**The E wallet and cash payments are very popular payment methods used by customers.

**4.** Naypyidaw is the most profitable branch in terms of gross income.

**5.**Sales are higher on Saturday and Tuesday , And on Monday sales are low because it is the start of the working week .Maybe on Tuesday there is a discount on products .

**6.** Food and beverages and Home and lifestyle have higher sales than the rest of the product lines.

**7.**We can conclude that yangon city people are technical , Mandalay people are sport oriented and Naypyidaw people are foodies who love Fashion.

****