12/20/2022

M-Smart Store Analysis with R

Final Project Report

BY
JEENESH R JAIN
NANDAN S
HITESH S
NEELESH

Dataset:

The Dataset is of a retail store, this data set contains 17 columns and 1000 rows of sales information which was recorded for a certain period of time. The dataset has the mix of columns. Glimpse of the dataset is the below screenshot for the reference.

```
> summary(Sales)
       Invoice.ID  Branch         City      Customer.type    Gender                    Product.line   Unit.price       Quantity
 101-17-6199:  1   A:340   Mandalay :332   Member:501    Female:501   Electronic accessories:170   Min.   :10.08   Min.   : 1.00
 101-81-4070:  1   B:332   Naypyitaw:328   Normal:499    Male  :499   Fashion accessories   :178   1st Qu.:32.88   1st Qu.: 3.00
 102-06-2002:  1   C:328   Yangon   :340                               Food and beverages    :174   Median :55.23   Median : 5.00
 102-77-2261:  1                                                      Health and beauty     :152   Mean   :55.67   Mean   : 5.51
 105-10-6182:  1                                                      Home and lifestyle    :160   3rd Qu.:77.94   3rd Qu.: 8.00
 105-31-1824:  1                                                      Sports and travel     :166   Max.   :99.96   Max.   :10.00
 (Other)    :994
     Tax.5.            Total                Date          Time            Payment         cogs        gross.margin.percentage
 Min.   : 0.5085   Min.   :  10.68   2/7/2019 : 20   14:42  :  7   Cash       :344   Min.   : 10.17   Min.   :4.762
 1st Qu.: 5.9249   1st Qu.: 124.42   2/15/2019: 19   19:48  :  7   Credit card:311   1st Qu.:118.50   1st Qu.:4.762
 Median :12.0880   Median : 253.85   1/8/2019 : 18   17:38  :  6   Ewallet    :345   Median :241.76   Median :4.762
 Mean   :15.3794   Mean   : 322.97   3/14/2019: 18   10:11  :  5                     Mean   :307.59   Mean   :4.762
 3rd Qu.:22.4453   3rd Qu.: 471.35   3/2/2019 : 18   11:40  :  5                     3rd Qu.:448.90   3rd Qu.:4.762
 Max.   :49.6500   Max.   :1042.65   1/23/2019: 17   11:51  :  5                     Max.   :993.00   Max.   :4.762
                                     (Other)  :890   (Other):965
  gross.income         Rating
 Min.   : 0.5085   Min.   : 4.000
 1st Qu.: 5.9249   1st Qu.: 5.500
 Median :12.0880   Median : 7.000
 Mean   :15.3794   Mean   : 6.973
 3rd Qu.:22.4453   3rd Qu.: 8.500
 Max.   :49.6500   Max.   :10.000
```

Libraries:

For calling any functions, Libraries and Packages play a very important role in R. Hence all the required libraries are imported in this Analysis. Tidyverse, Cowplot, dplyr and corrplot..etc are some important libraries amongst all. These are just quoted for the reference.

Pre-Processing and Data Transformation –

Once the Libraries are imported, it is now the Data Import and Data Cleaning operation that needs to be done. Below is the Source of the file that is being called from.

```
Sales = read.csv("C:/Users/Admin/Desktop/SMart_sales.csv")
```

For us to initially look at the data in R, Below function is used as this defines and explains clearly about each attributes and Classes.

```
# To list out the type of each attribute used in this dataset
sapply(Sales, class)
```

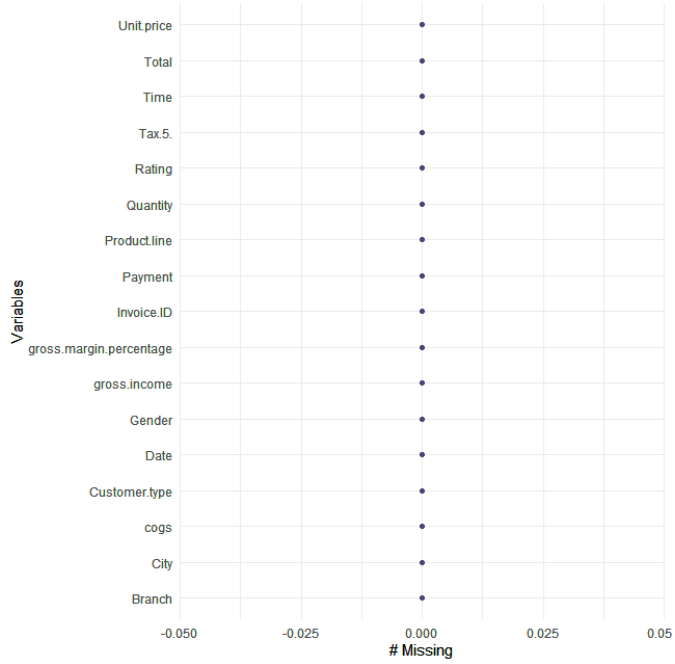
Data Cleaning: In this Module, Missing values are replaced and some data changes in the column has been performed.

```
# Replacing the Values
Sales <- Sales %>% mutate(Customer.type = replace(Customer.type,Customer.type == "Subscribed User","Member"))
Sales <- Sales %>% mutate(Customer.type = replace(Customer.type,Customer.type == "Un Subscribed User","Normal"))
```

Cleaning Data –

Data Cleaning is one of the crucial steps in the analysis. Here the primary objective of Data cleaning is to make the model to get the most out of the dataset. So All the assumptions are listed in this section before filling any values in the dataset. Below is the dataset values after cleaning the dataset. It is clearly visible from the below image that there are no missing/Blank values.

Result After Cleaning Missing Values:



Data Analysis:

Here we perform some of the operations to see the current visualizations and understand the metrics. As R is known for Statistical Analysis, Some analysis were performed on this existing data to check and draw some insights from this business.

1. Payments Frequency:

```
# Now consider the number of instances (rows) that belong to each class.
# Class distribution Summarization

percentage <- prop.table(table(Sales$Payment)) * 100
cbind(freq=table(Sales$Payment), percentage=percentage)
```

Result:

```
> cbind(freq=table(Sales$Payment), percentage=percentage)
            freq percentage
Cash         344       34.4
Credit card  311       31.1
Ewallet      345       34.5
>
```

2. Summarizing Data:

When the Business tries to Analyze the city rating, this could be the best approach. As the below insight gives a clear image on the ratings given. Even though this is not that great visual to capture all metrics in one view, Rounding off the values would help business to understand these numbers better.

```
table(Sales$City, Sales$Rating)
```

Result:

```
> table(Sales$City, Sales$Rating)

            4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9  5 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9  6 6.1 6.2 6.3 6.4 6.5 6.6 6.7
  Mandalay  5   6   7   7   5   7   4   5   5   8  7  10   5   4   4   5   6   4   4   4  8   6   7   7   8  10   5   9
  Naypyitaw 4   6   4   5   6   4   1   2   4   6  6   6   6   5   6   7   5   5   2   7 10   5   6   6   5   6   8   5
  Yangon    2   5  11   6   6   6   3   5   4   4  8   5   4   2   4   7   5   5   8   6  8   3   8   2   3   5  11   5

           6.8 6.9  7 7.1 7.2 7.3 7.4 7.5 7.6 7.7 7.8 7.9  8 8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9  9 9.1 9.2 9.3 9.4 9.5
  Mandalay   4   5  9   4   4   5   3   3   6   7   5   7  9   4   3   3   5   4   6   9   1  10  6   1   2   3   3   7
  Naypyitaw  4   5  4   8   6   5   4   5   6   4   3   7  6   6   6   4   9   7   8   5   5   3  5   5   6   7   6  10
  Yangon     5   5  7   4   8   8   9   5   8   7   7   4  6   3   5   4   4   7   4   6   8   3  4   8   8   6   3   5

           9.6 9.7 9.8 9.9 10
  Mandalay   6   3   5   5  3
  Naypyitaw  5   5   6   4  1
  Yangon     6   6   8   7  1
```

Unit Price Comparison with Gross Income:

Here we analyze the unit price and its contribution in the gross income for the organization. Below are the screen shots for your reference.

```
#Unit Price comparision with gross income
x <- Sales$Unit.price
y <- Sales$gross.income
plot(x, y, col = "Dark green", main="Unit Price Vs Gross Income",
     xlab="Unit Price", ylab="Gross Income")
```

Result: Unit Price vs Gross Income:

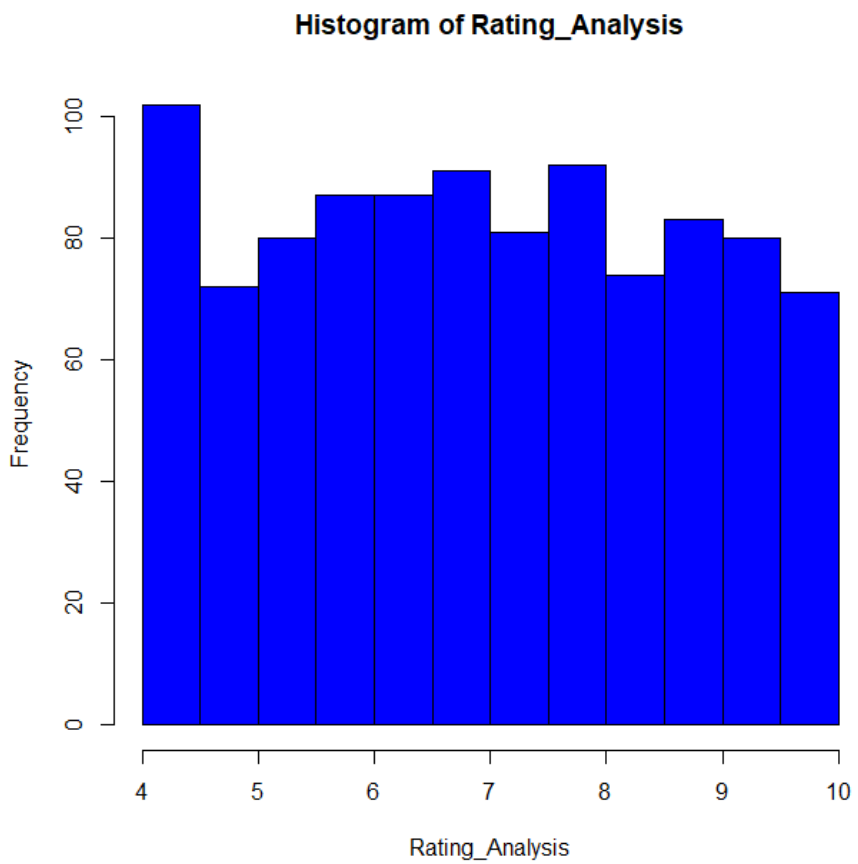
Visualization:

It is time to explore some visuals and see the results to understand the metrics better. Below are some analysis, which shows Histograms, Graphs..Etc.

Rating Analysis:

```
#Rating Analysis

Rating_Analysis <- Sales$Rating
hist(Rating_Analysis, col = 'Blue')
```

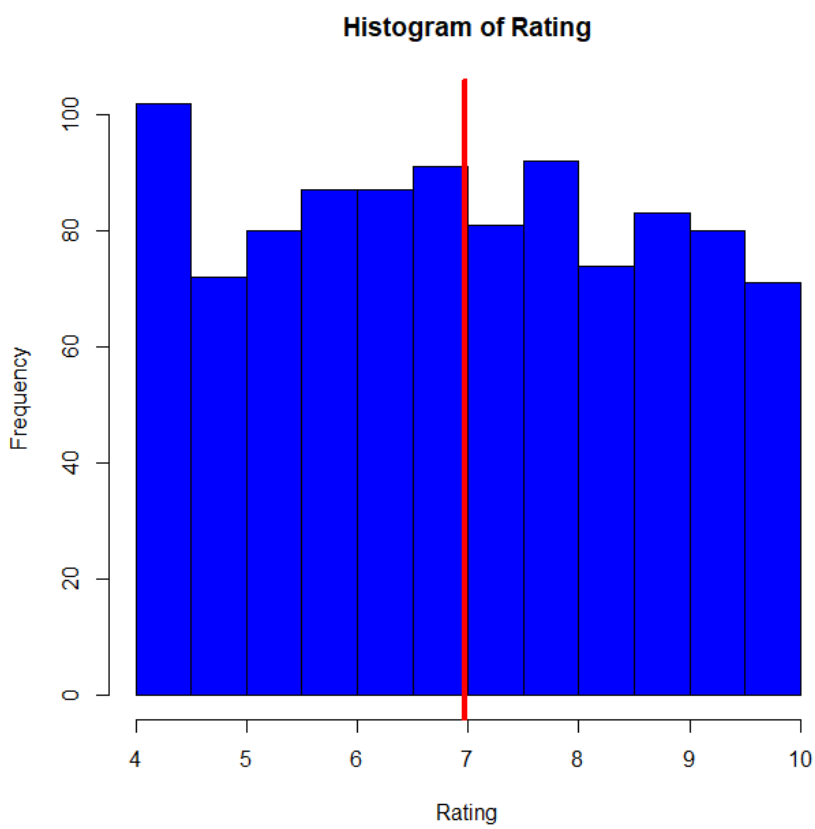
Result:



Mean of Average: Overall if we need an average rating for the data that we have analyzed so far, then this is definitely one of the methods. Overall Average rating has come up to 6.9 and close to 7.

```
#Taking Average of the Rating - Overall Average

Rating_Analysis <- Sales$Rating
average <- mean(Rating)
hist(Rating, col = 'Blue')
abline(v= average, col='Red', lwd =4)
```

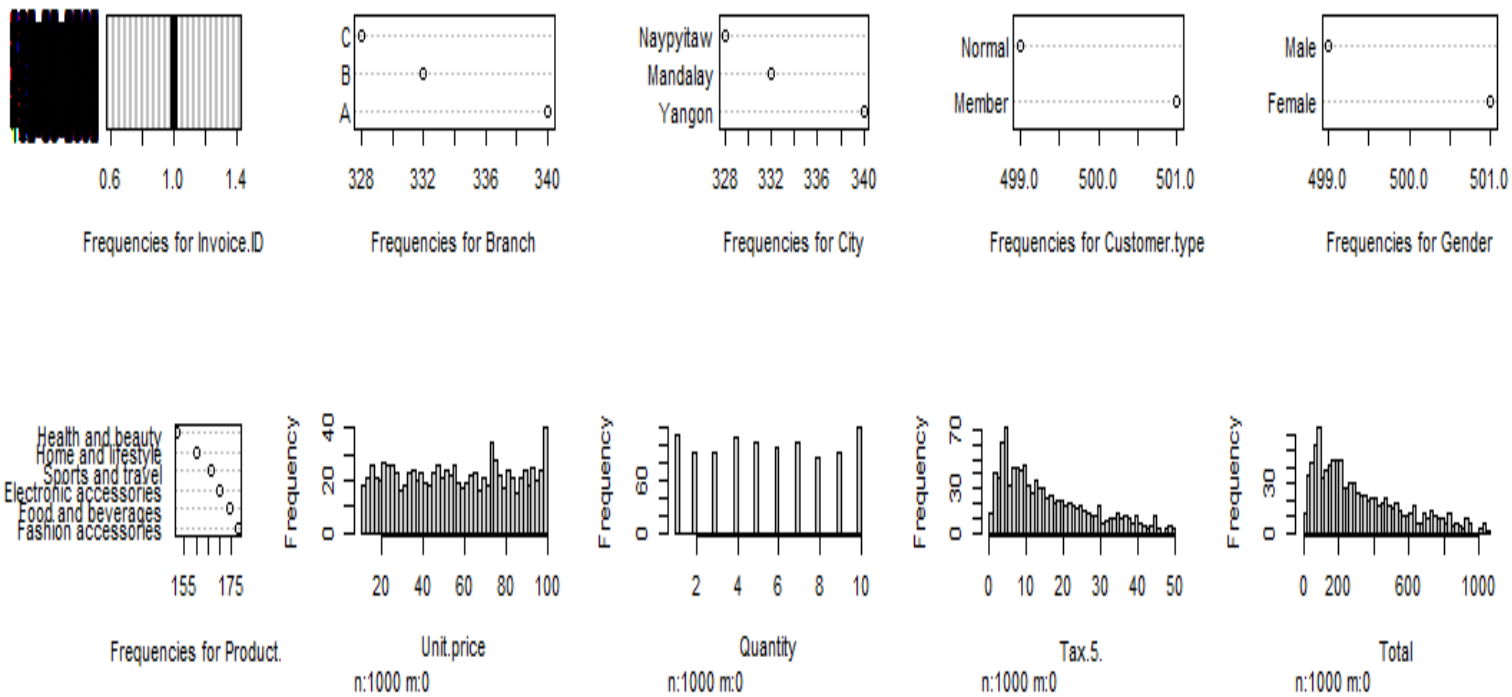
Result:



Summarized ViOverall Dataset:

```
# To plot the Histogram for the whole dataset
hist.data.frame(sales)
```

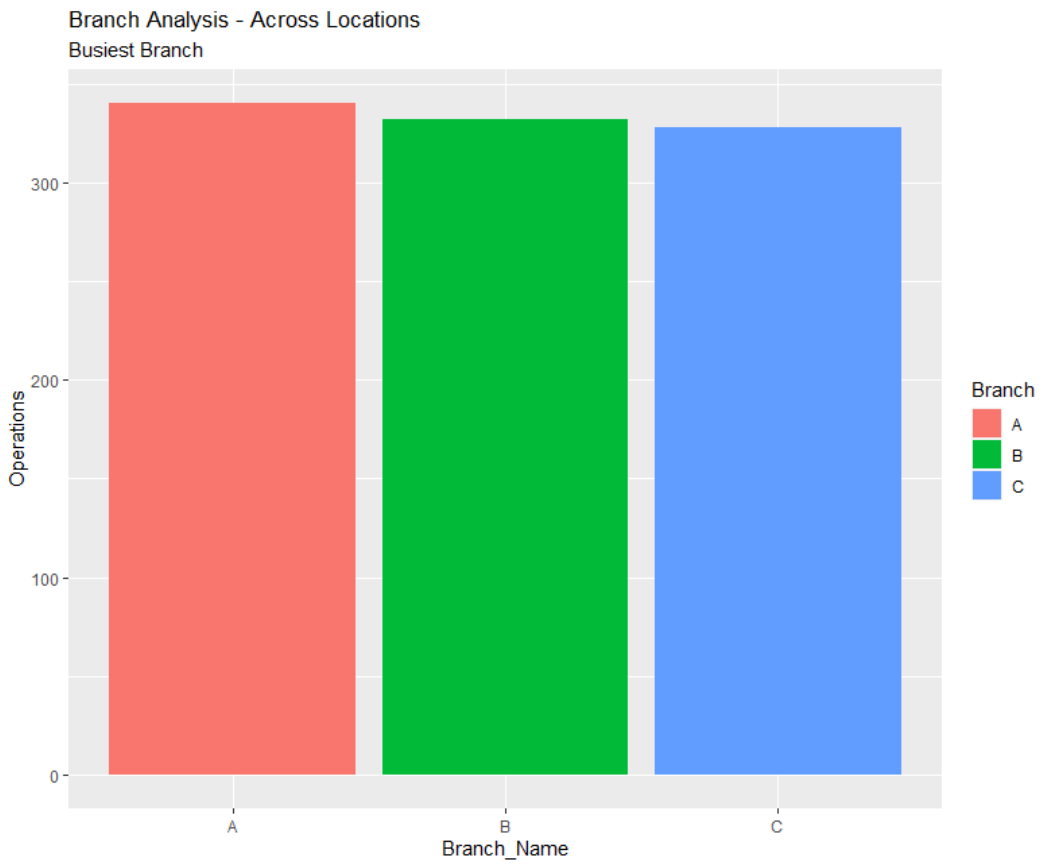
Result:



Branch Analysis :

There is competition everywhere, for business to review the performance, it is very important to have a insight of all the individual performances, which shall later be also considered for the growth calculations and revenue contribution for organization. To find out the busiest branch according to the operations.

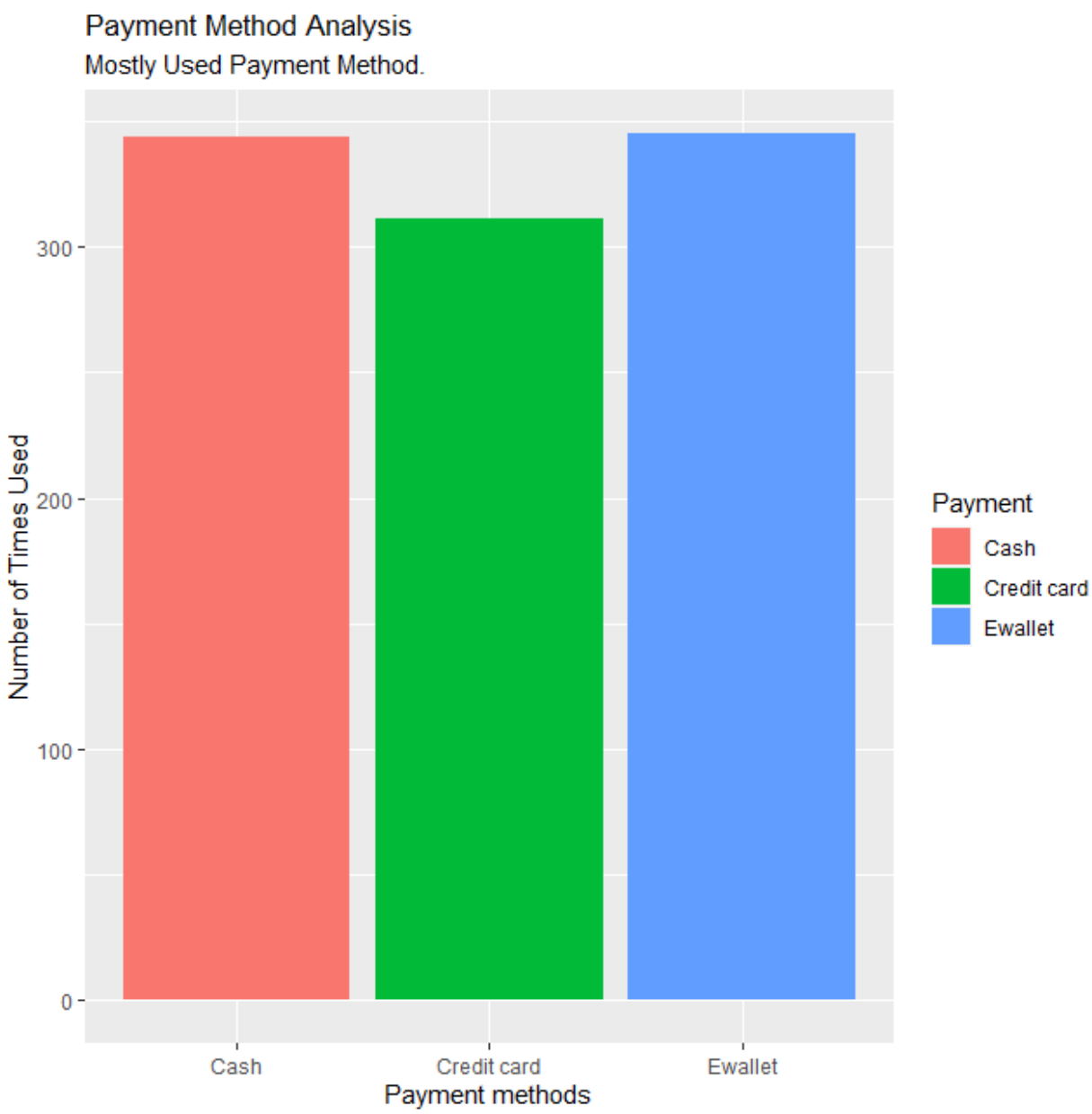
```
# Branch Analysis - Operations Persepective to Analyse the Busiest Branch

ggplot(data = Sales, aes(x = Branch)) +
  geom_bar(aes(fill = Branch)) +
  labs(title = "Branch Analysis - Across Locations",
       subtitle = "Busiest Branch",
       x = "Branch_Name",
       y = "Operations")
```



Payment Method Analysis: Due to Digitalization, it is observed that E wallet payments has increased.

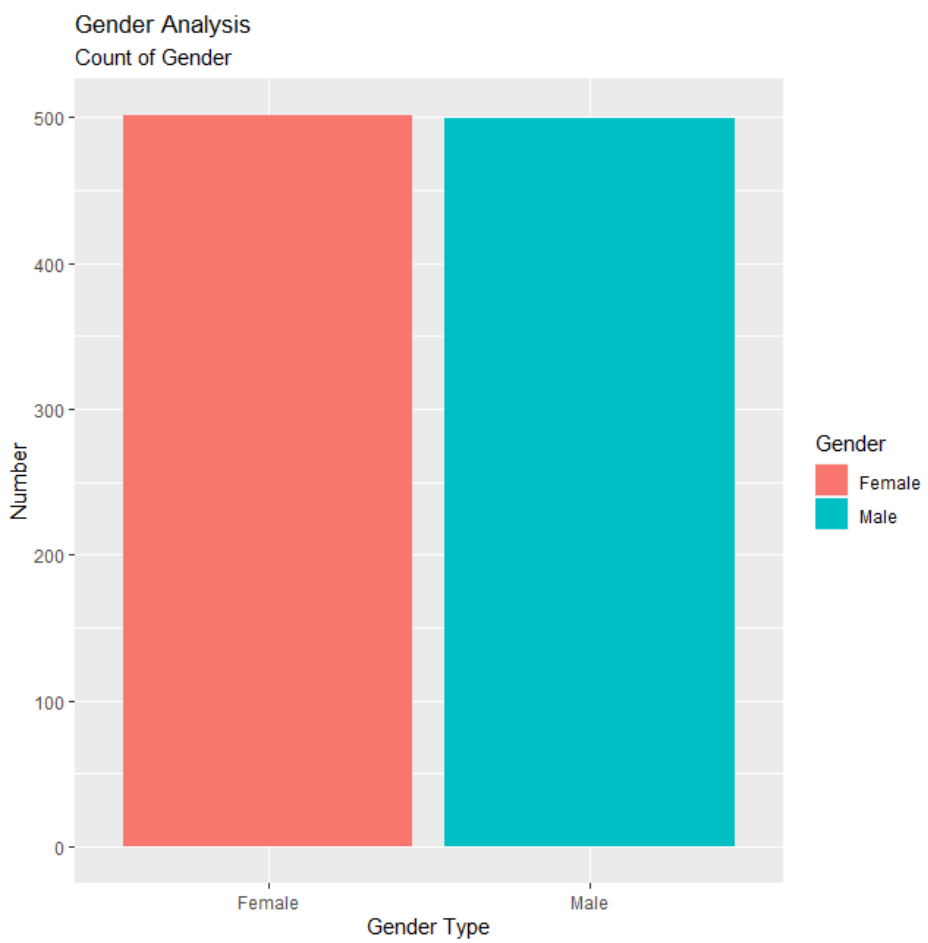
```
# payment Method Analysis - To Validate and the understand the most used Payment method

ggplot(data = Sales, aes(x = Payment)) +
  geom_bar(aes(fill = Payment)) +
  labs(title = "Payment Method Analysis",
       subtitle = "Mostly Used Payment Method.",
       x = "Payment methods",
       y = "Number of Times Used")
```

Result:



Gender Analysis:

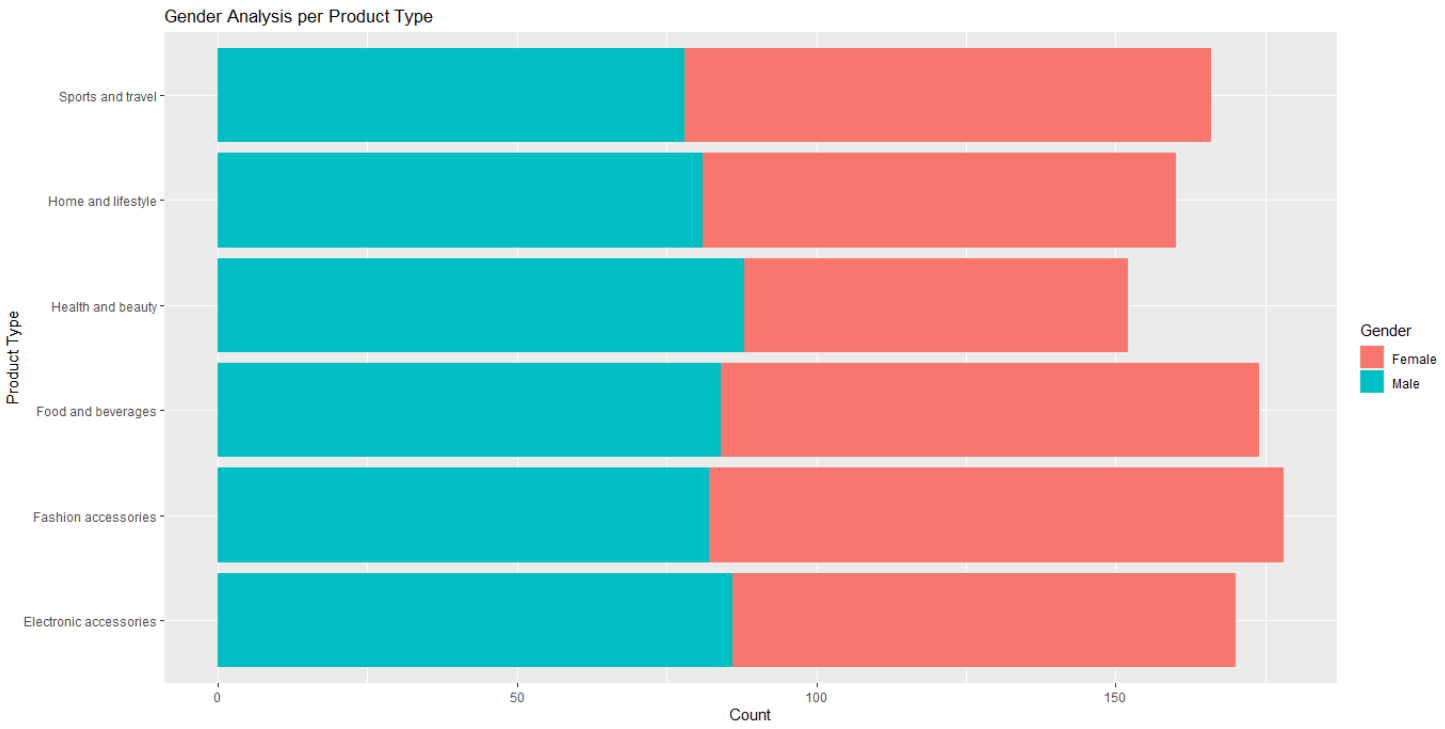
```
# Gender Analysis - To check on which gender is giving Most sales

ggplot(data = Sales, aes(x = Gender)) +
  geom_bar(aes(fill = Gender)) +
  labs(title = "Gender Analysis",
       subtitle = "Count of Gender",
       x = "Gender Type",
       y = "Number")
```



Product By Gender:

```
#Gender Analysis by Product

ggplot(Sales, aes( y= 'Product.line)) +
  geom_bar(aes(fill = Gender)) +
  labs(title = "Gender Analysis per Product Type",
       x = "Count",
       y = "Product Type")
```

Result:



Correlation Matrix:

A correlation matrix is essentially a table that displays the correlation coefficients for various variables. The matrix displays the relationship between all possible pairings of values in a table. It is an effective tool for summarizing a huge dataset as well as identifying and visualizing trends in the data.

Correlation Formulae:

```
#Correlation Analysis
cor = cor(Sales$Quantity,Sales$gross.income)
cor
```

Result of Correlation Matrix between these two parameters:

```
> #Correlation Analysis
> cor = cor(Sales$Quantity,Sales$gross.income)
> cor
[1] 0.7055102
```

Regression Models:

1. Simple Regression Model:

Regression Models are applied on the continuous solutions for better solutions. This analysis benefits such analysis and gives good results. Please find the algorithm applied below

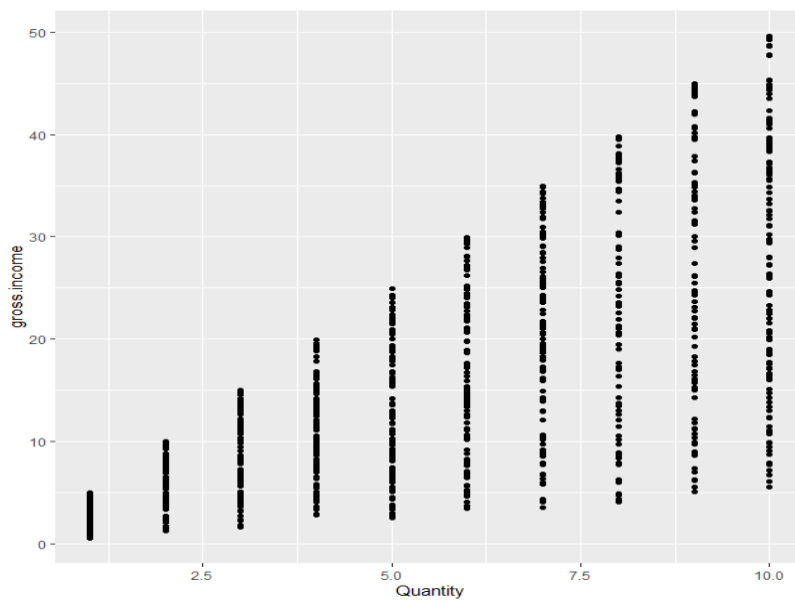
```
#Model 1: Linear Regression - To validate Quantity Vs Total

Sales_Quantity.lm <- lm(Quantity ~ Total, data = Sales)

summary(Sales_Quantity.lm)
```

Results:

```
> summary(Sales_Quantity.lm)

Call:
lm(formula = Quantity ~ Total, data = Sales)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6789 -1.6822 -0.5127  1.2203  6.2338 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.8009236  0.1082462   25.88   <2e-16 ***
Total       0.0083881  0.0002667   31.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.073 on 998 degrees of freedom
Multiple R-squared:  0.4977,	Adjusted R-squared:  0.4972 
F-statistic:   989 on 1 and 998 DF,  p-value: < 2.2e-16
```

The considerable thing to be observed here is the p value is <2.2e-16) which says whether the model fits the data well or not

The value of y-intercept is 2.8009236 and the residual standard error is 2.073. T value is 31.45



2. Multiple Regression Model:

```
#Model 2 : # Multiple Regression : To Validate Gross Income vs Total

Sales_Gross_Income.lm <- lm(gross.income ~ Quantity + Total, data = Sales)

summary(Sales_Gross_Income.lm)

plotmulti.data <- expand.grid(Total = seq(min(Sales$gross.income), max(Sales$Quantity), length.out=30),
                              Year=c(min(Sales$Total), mean(Sales$Total), max(Sales$Total)))
```

Result:

```
> Sales_Gross_Income.lm <- lm(gross.income ~ Quantity + Total, data = Sales)
>
> summary(Sales_Gross_Income.lm)

Call:
lm(formula = gross.income ~ Quantity + Total, data = Sales)

Residuals:
       Min         1Q     Median         3Q        Max
-1.388e-12 -1.300e-16  9.100e-16  2.410e-15  4.268e-14

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept) -4.294e-14  2.976e-15 -1.443e+01   <2e-16 ***
Quantity     6.152e-16  6.732e-16  9.140e-01    0.361
Total        4.762e-02  8.004e-18  5.950e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.408e-14 on 997 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.524e+31 on 2 and 997 DF,  p-value: < 2.2e-16
```

The considerable thing to be observed here is the p value is <2.2e-16) which says whether the model fits the data well.

The value of y-intercept is and the residual standard error is 8.004e-18. T value is 5.950

Conclusion and Future Work:

This Regression Models gave better results and would definitely need some more data for the model to predict the data. Retail Analysis is needed for all the businesses to understand the insights and take decisions. Technology is helping businesses to make the decisions by proving such a beautiful analysis.

This project can be extended by continuing this analysis on training the datasets. We have started working on Arima and Naïve Forecasting Models.