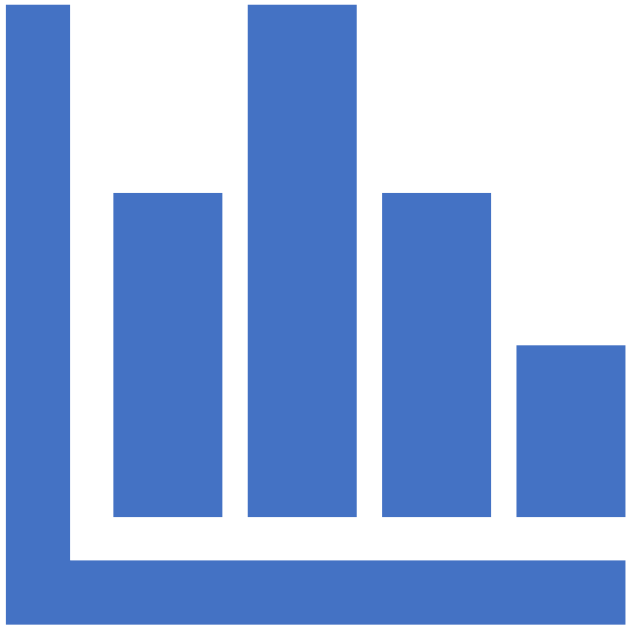
Alexandria University
Faculty of Computer and Data Science
Department: Data Science
Course Title: Data Science 2023-2024



Introduction to Data Science Course

Code: 02-24-00104

Super Market Sales Report

Members Names and Role

Name	ID	Role
1. Youssef Khaled Abdel Aziz	22010303	Research, data visualization and UI
2. Salsabil Mohamed Askar	22010110	Supervised and unsupervised methods
3. Mazen Ayman Jalal Fouad	22010198	Data visualization and UI
4. Malak Ali Ahmed	22010265	Data cleaning
5. Youssef Abdel Wahab Mohammed	22010308	Data visualization
6. Abdul Rahman Hisham	22010136	Reporting

Introduction:

•Objective:

- The objective of this data science project is to analyze and gain insights from a dataset containing information about **supermarket sales**. By employing data science techniques, we aim to uncover patterns, trends, and key factors that contribute to the sales performance of the supermarket. This analysis can aid in strategic decision-making, marketing optimization, and overall business improvement.

Input:

1-Invoice ID : Computer-generated invoice ID number

2-Branch : Branch of the Supermarket chain where the transaction took place (A, B or C)

3-City : City where the Supermarket store is located

4-Customer Type : The type of customer who made the purchase (Member or Normal)

5-Gender : Gender type of customer (Female or Male)

6-Product Line : The Product line to which the product purchased belongs to, according to the Supermarket chain categorization SOP

7-Unit Cost : Product cost, or what the Supermarket chain paid for it

8-5pct_markup : 5% markup on the purchase total cost

9-Revenue : Revenue from the purchase (includes markup)

10-Date : The date when the purchase took place

11-Time : The time when the purchase took place

12-Quantity : Quantity purchased

13-Payment Method : The payment method used to conduct the purchase

14-COGS (Cost of Goods Sold) : Cost of Goods Sold, calculated as unit cost times quantity

15-gm_pct : Gross Margin percentage. Calculated as 1 - cogs/revenue

16- Gross income : Gross Income, calculated as revenue minus COGS

Introduction

17_rating : Rating assigned to the purchase

Output:

This project aims to provide valuable insights to supermarket stakeholders, helping them optimize sales strategies, enhance customer experience, and improve overall business performance.

Methodologies used

- **Project Title:** Supermarket Sales Analysis for Business Optimization

Project Steps:

- **Data Collection**:

  Gathered a comprehensive dataset containing information on supermarket sales, including transaction details, customer attributes, and product-related variables from Kaggle.

- **Data Cleaning and Pre-processing:**

  Checked for missing values, outliers, and inconsistencies in the dataset.
  Handled missing data using appropriate imputation techniques.
  Standardized and normalized numerical features to ensure uniformity.

- **Data Visualization:**

  Created visualizations to communicate key insights and complex patterns and insights in data to both technical and non-technical audiences.

- **Decision Tree Modeling:**

  Built a decision tree model to predict key factors influencing sales, such as Gross Income, product preferences, and branch.

- **Cluster Analysis:**

  Applied k-means clustering to group customers based on similar purchasing behavior.

Methodologies and Techniques:

- Data cleaning
- K-means
- Decision tree
- Data visualization and ui

Challenges in the dataset

- **Data Quality Issues:**
  - The dataset may have contained missing values, outliers, or inaccuracies that required careful handling during the data cleaning process.
- **Complexity of Customer Behavior:**
  - Understanding and accurately capturing the diverse and complex patterns of customer behavior can be challenging, especially in a supermarket setting where customers may exhibit varied preferences.
- **Choosing the Optimal Number of Clusters (k):**
  - Selecting the right number of clusters for the k-means algorithm can be subjective and may impact the quality of the cluster analysis.
- **Interpreting and Communicating Complex Models:**
  - Decision tree models, while interpretable, can become complex and challenging to communicate effectively.
- **Ensuring Business Relevance:**
  - Aligning analysis and findings with the specific business goals and needs of the supermarket stakeholders.
- **Resource Constraints:**
  - Limited computational resources may impact the scale and complexity of the analysis,

Data Cleaning

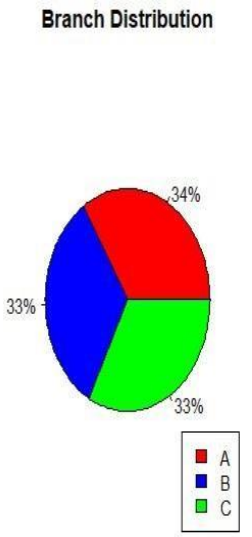
```
13  #check data
14  dim(supermarket_sales)
15  str(supermarket_sales)
16  ----------------------------------------------
```

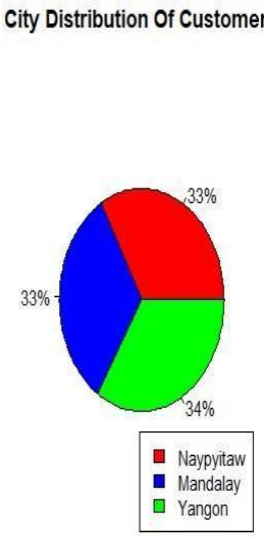
```
> dim(supermarket_sales)
[1] 1000   17
> str(supermarket_sales)
'data.frame':   1000 obs. of  17 variables:
 $ invoice_id     : chr  "750-67-8428" "226-31-3081" "631-41-3108" "123-19-1176" ...
 $ branch         : chr  "A" "C" "A" "A" ...
 $ city           : chr  "Yangon" "Naypyitaw" "Yangon" "Yangon" ...
 $ customer_type  : chr  "Member" "Normal" "Normal" "Member" ...
 $ gender_customer: chr  "Female" "Female" "Male" "Male" ...
 $ product_line   : chr  "Health and beauty" "Electronic accessories" "Home and lifestyle" "Health and beauty" ...
 $ unit_cost      : num  74.7 15.3 46.3 58.2 86.3 ...
 $ quantity       : int  7 5 7 8 7 7 6 10 2 3 ...
 $ X5pct_markup   : num  26.14 3.82 16.22 23.29 30.21 ...
 $ revenue        : num  549 80.2 340.5 489 634.4 ...
 $ date           : chr  "01/05/19" "03/08/19" "03/03/19" "1/27/2019" ...
 $ time           : chr  "13:08" "10:29" "13:23" "20:33" ...
 $ payment_method : chr  "Ewallet" "Cash" "Credit card" "Ewallet" ...
 $ cogs           : num  522.8 76.4 324.3 465.8 604.2 ...
 $ gm_pct         : num  4.76 4.76 4.76 4.76 4.76 ...
 $ gross_income   : num  26.14 3.82 16.22 23.29 30.21 ...
 $ rating         : num  9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
>
```

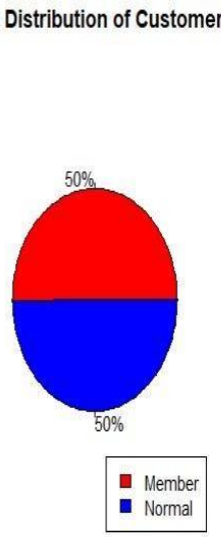
```
17  #Data cleaning
18  sum(duplicated(supermarket_sales))
19  sum(is.na(supermarket_sales))
20
```

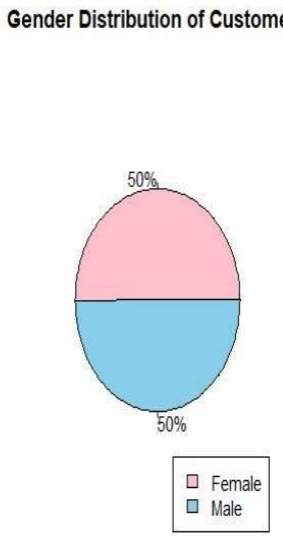
```
> sum(duplicated(supermarket_sales))
[1] 0
> sum(is.na(supermarket_sales))
[1] 0
>
```

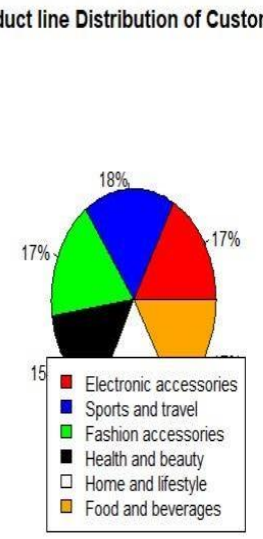
Data Visualisation

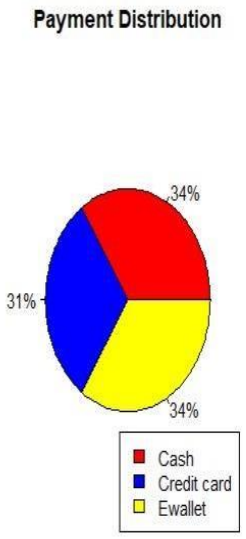






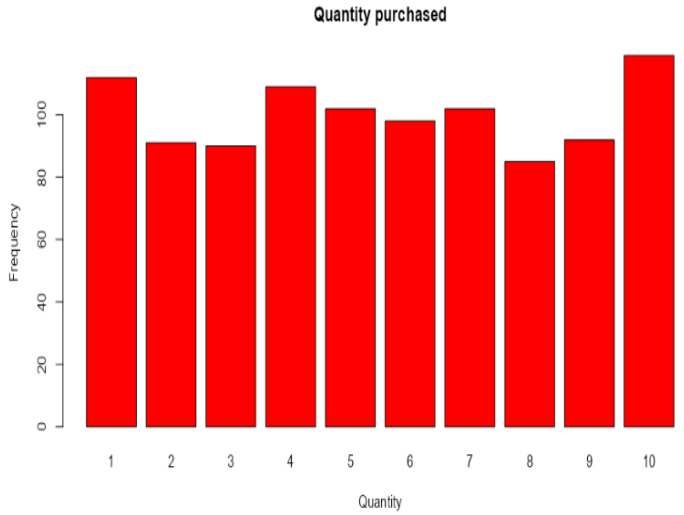


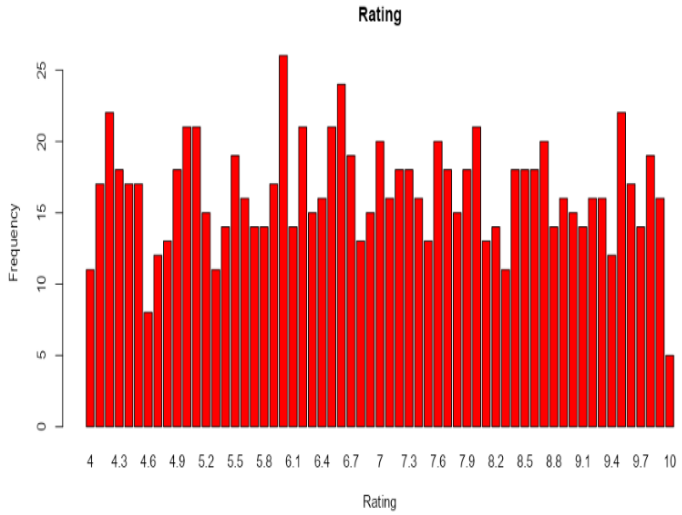


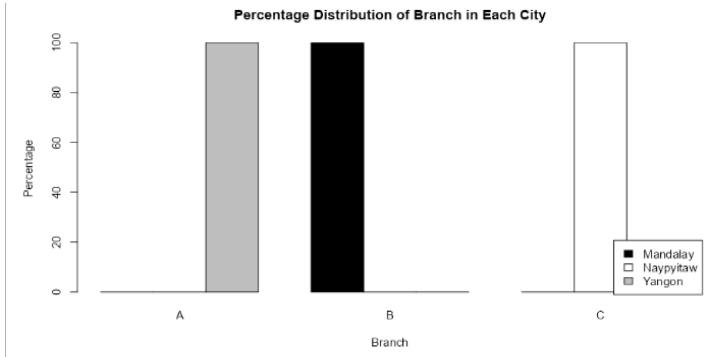


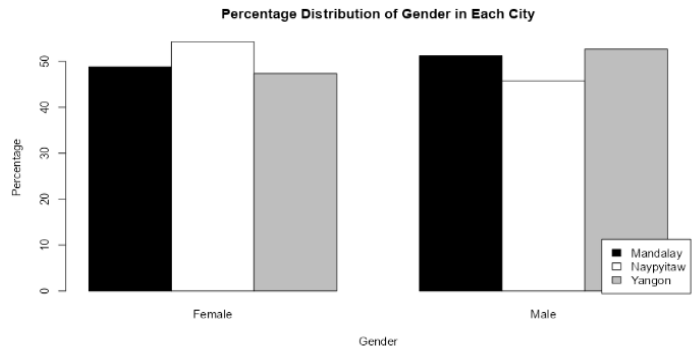
1 : The highest percentage of customers from branch A
2 : The highest percentage of customers from branch Yangoo
3 : The proportion of customer type is equal
4 : The proportion of customer gender is equal
5 : The highest percentage of payment method is Cash and Ewallet

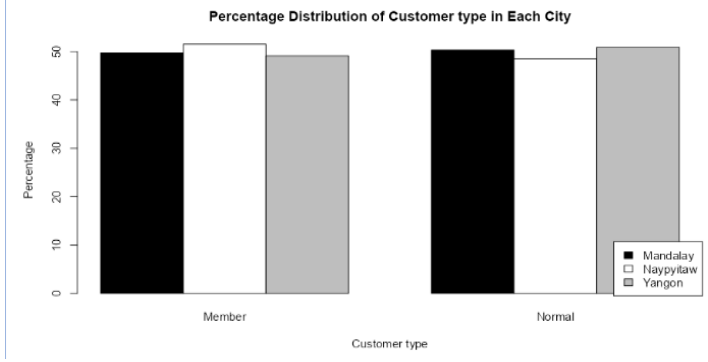
.Data Visualisation

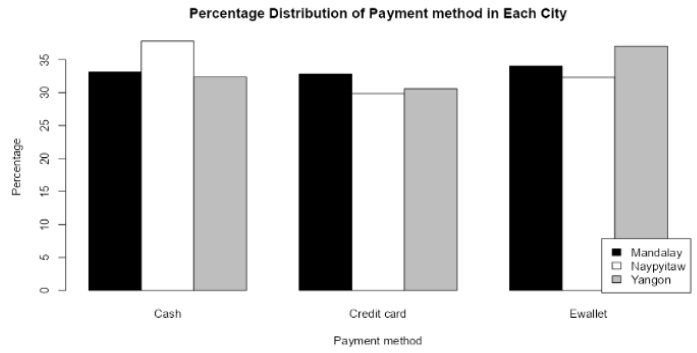












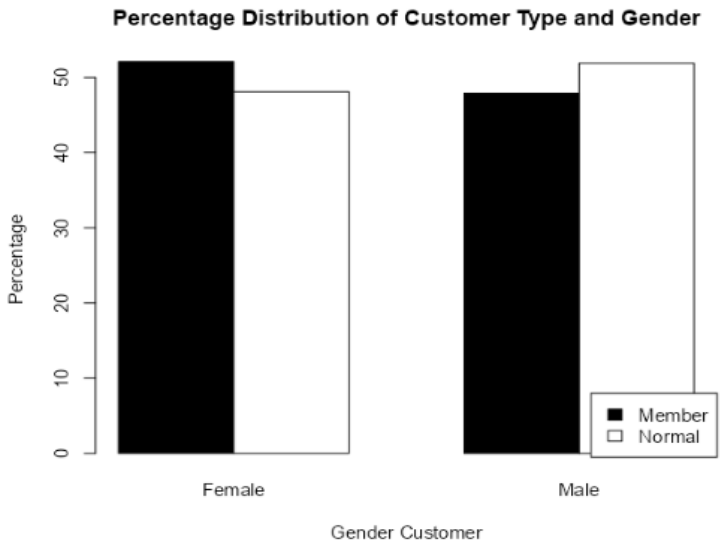
This page shows the frequency of quantity and rating.

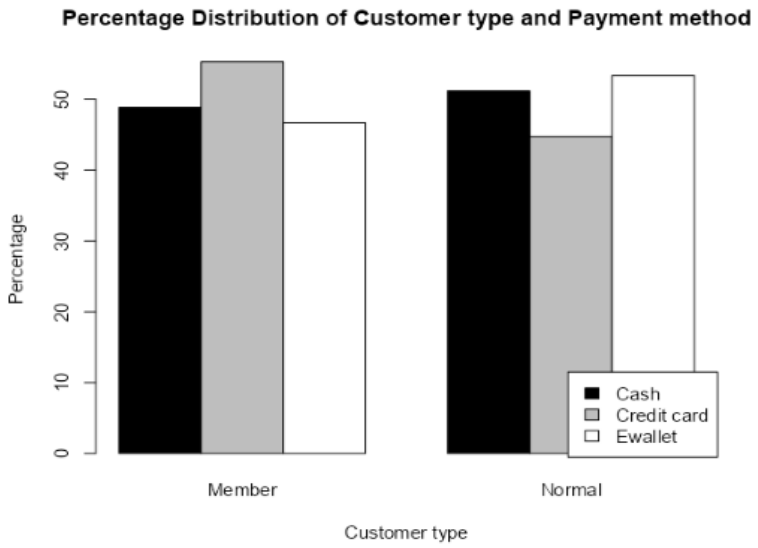
1. The average of ratings between 6 to 7.
2. The rating below 4 is non existence.
3. Most quantity purchased 10.

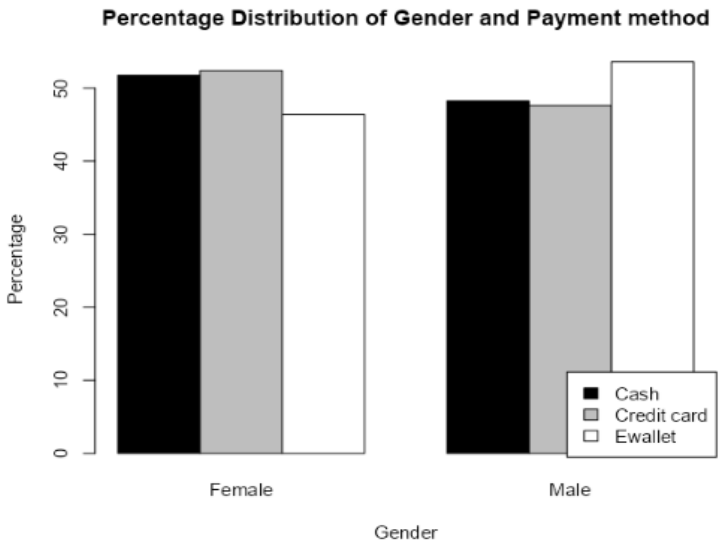
This page shows the Relationships of multiple elements with cities.

1.The first visualization shows that the branch A is in Yangon, the branch B is in Mandalay and the branch C is in Naypyitaw.
2.The largest percentage of females is in Naypyitaw.
3.The largest percentage of cash method is in Naypyitaw although the percentage of cash and ewallet are equal.
4.The highest percentage of member customers is in Naypyitaw
5.The percentage of normal customers in Yangoo and Mandalay is equal

Data Visualisation





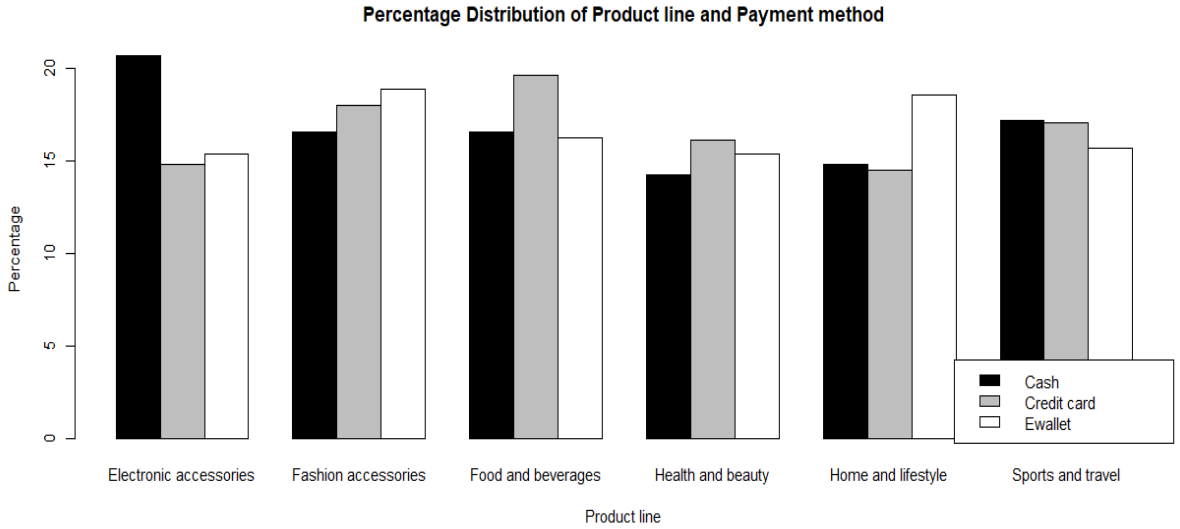


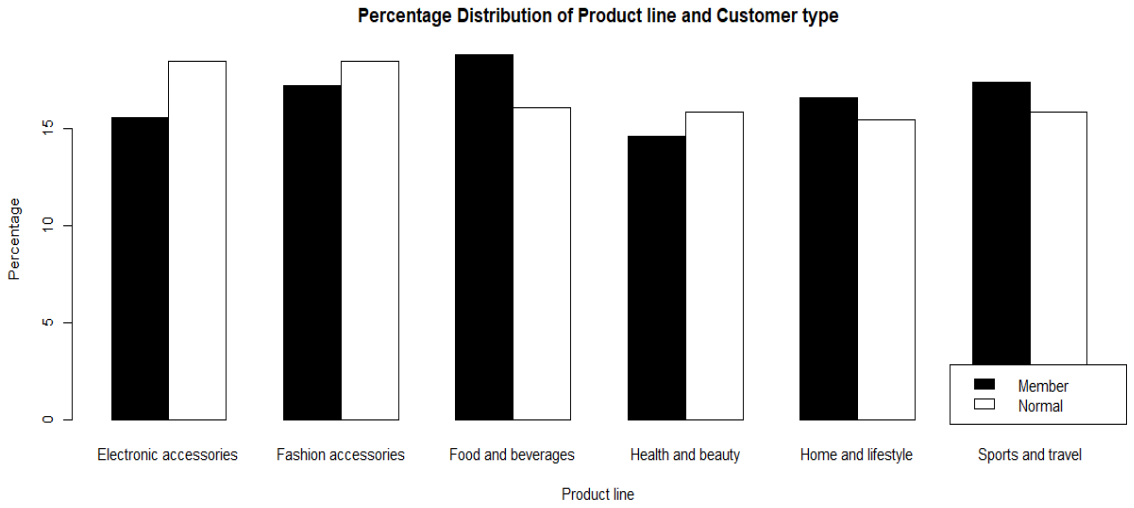
- This page shows the Relationships of multiple elements with each other.

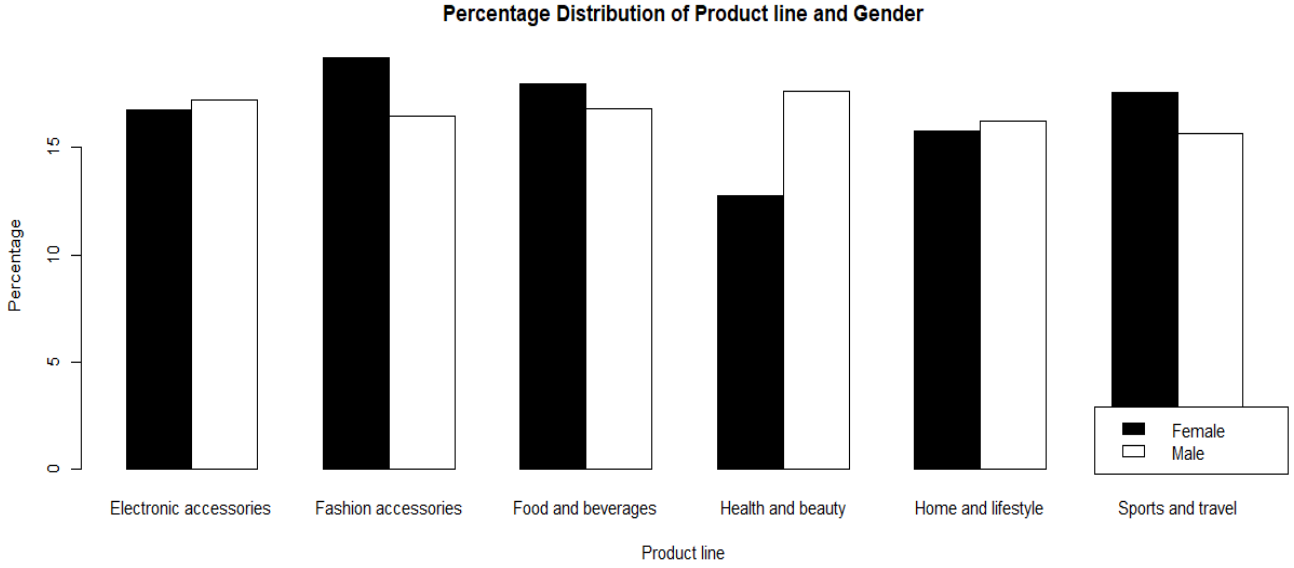
1. The members prefers using Credit card but the normal people prefer Ewallet and Cash.
2. The males prefer using Ewallet but the females prefer Credit card and Cash.

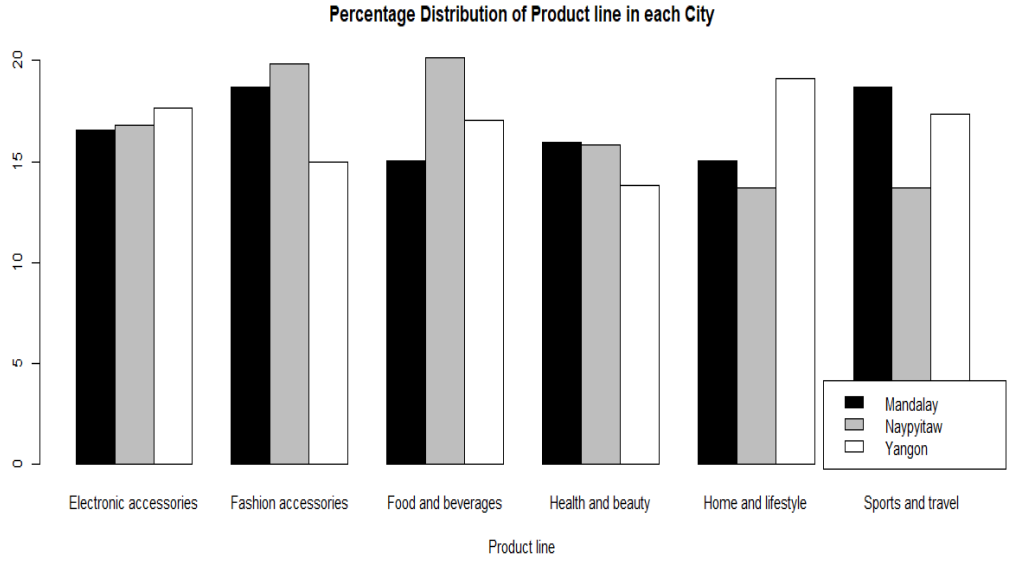
- ***As we see in this page the proportions are overlapping and that helps to bring the proportions are close***
- ***Female = 50% Male = 50% (From Percentage page).***
- ***Member = 50% Normal = 50% (From Percentage page).***
- ***Cash = 34% Ewallet = 34% Credit card =31% (From Percentage page).***

Data Visualisation



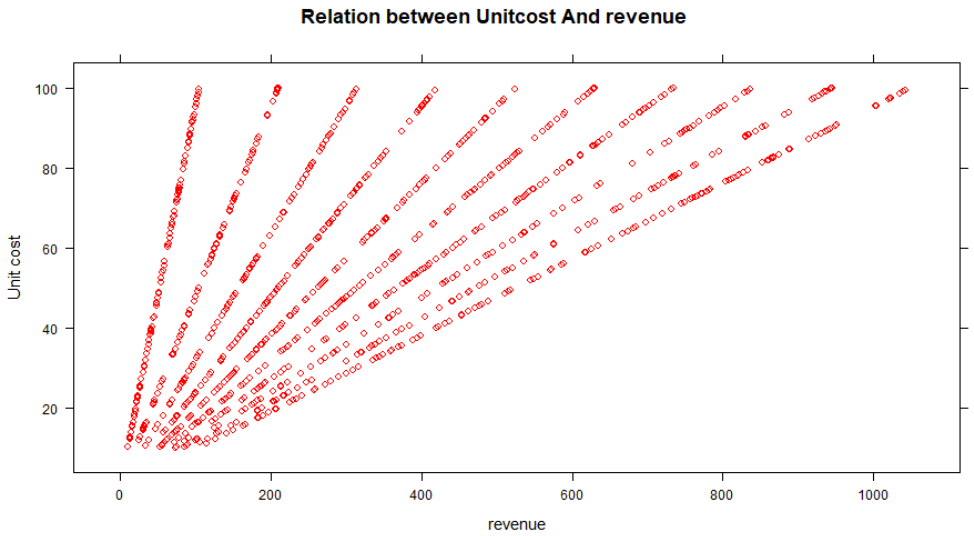


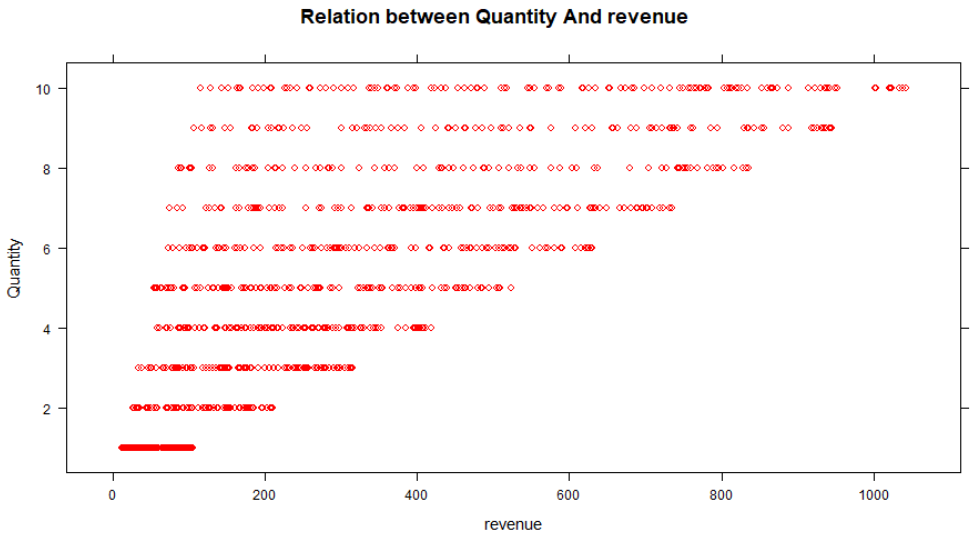


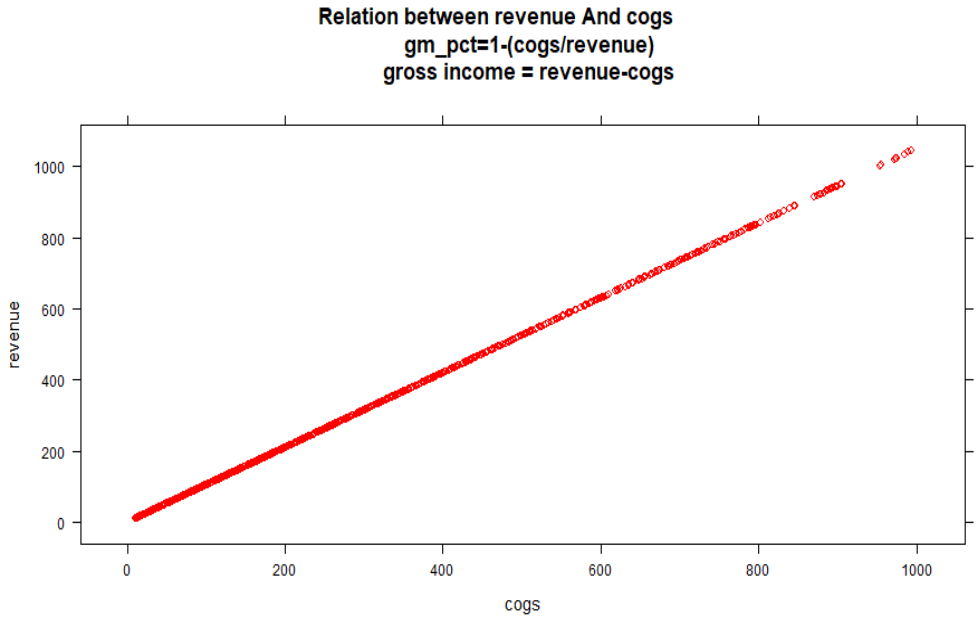


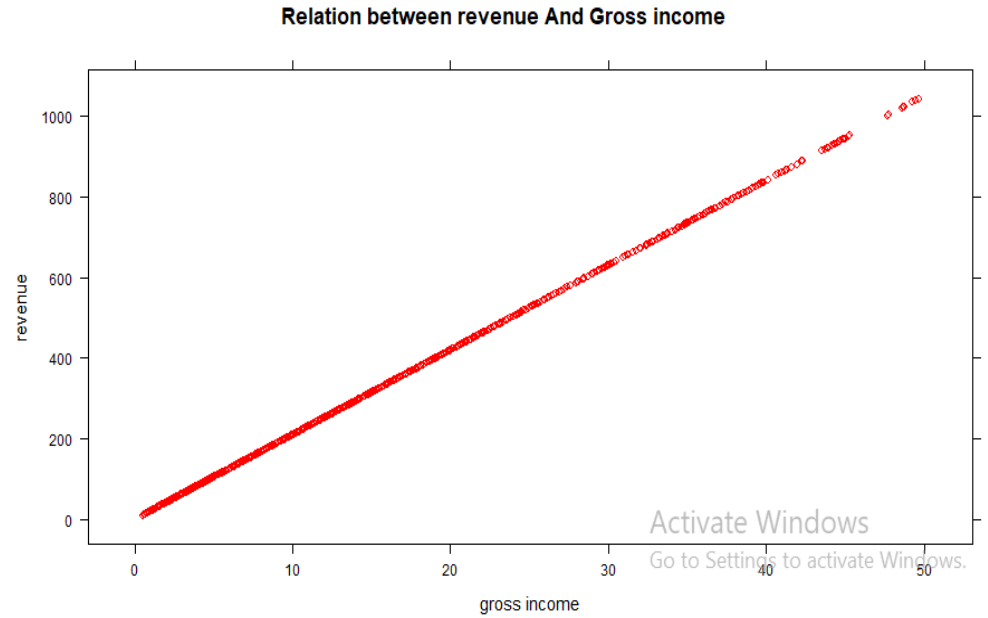
This page shows the Relationships of multiple elements with Product line

.Data Visualisation



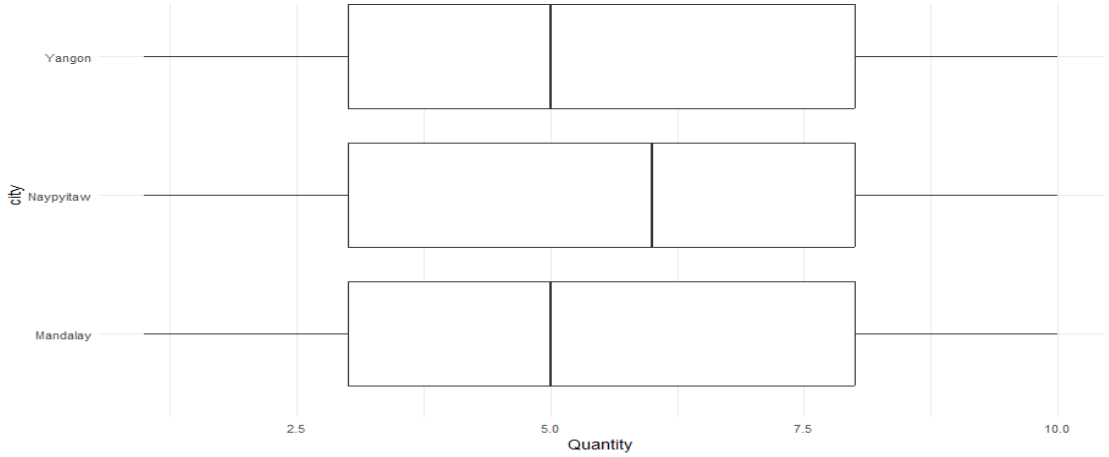


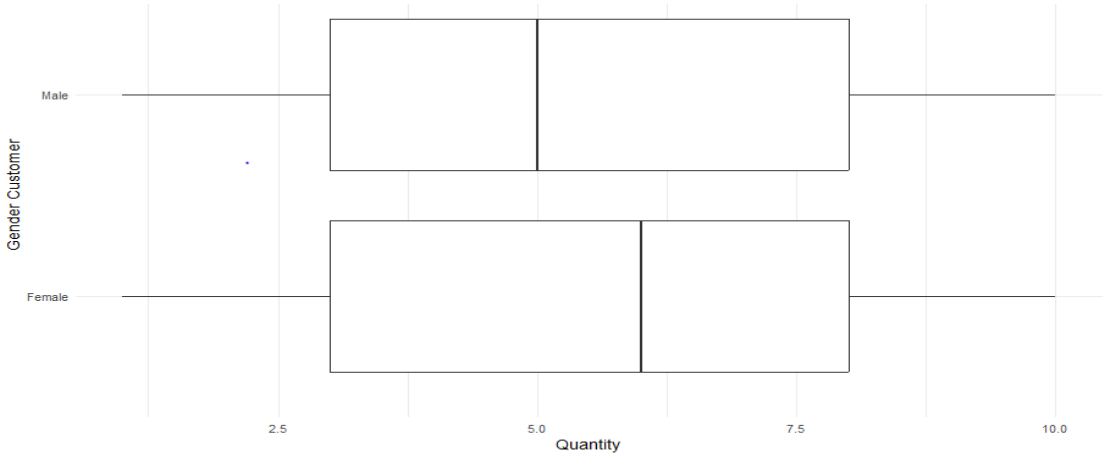


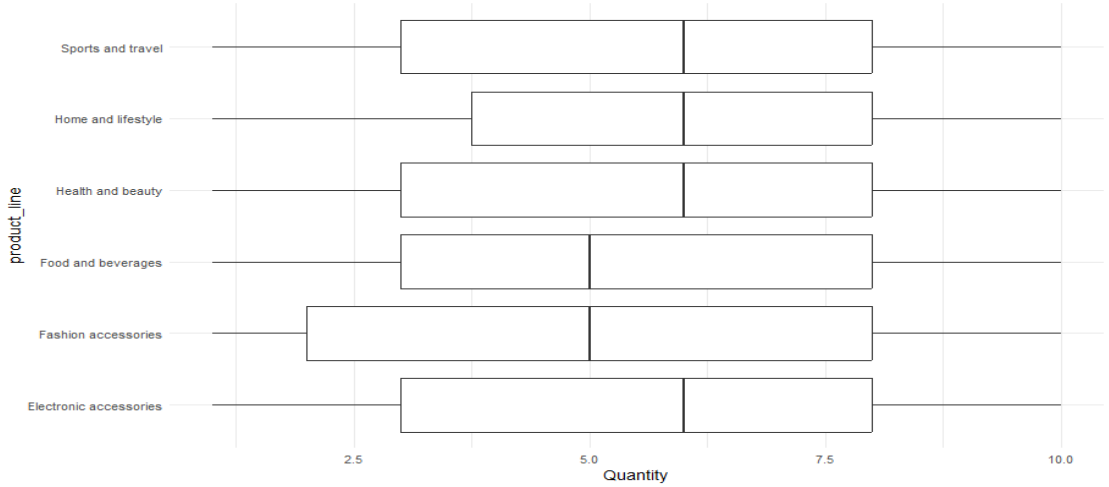


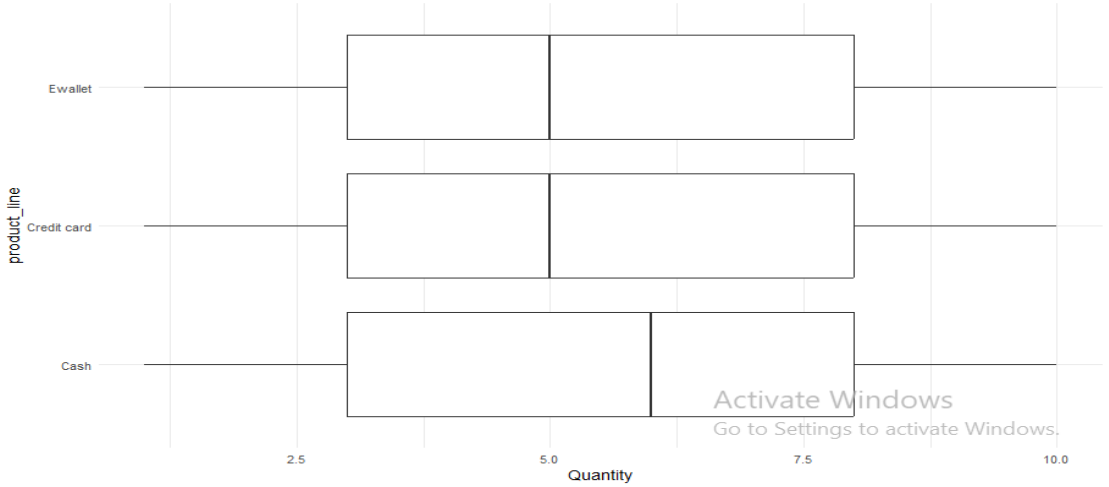
This page shows the Relationships of multiple elements with each other.

Data Visualisation



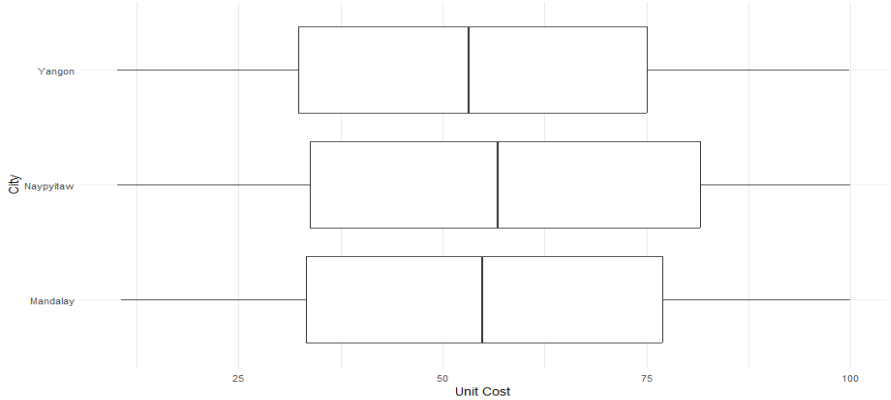


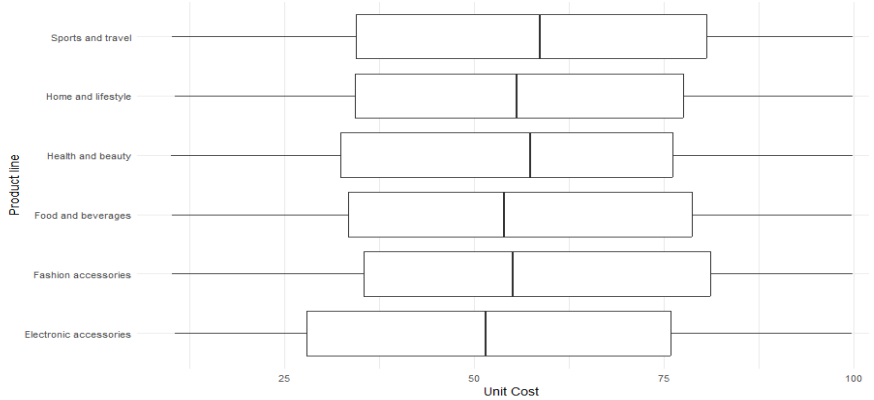




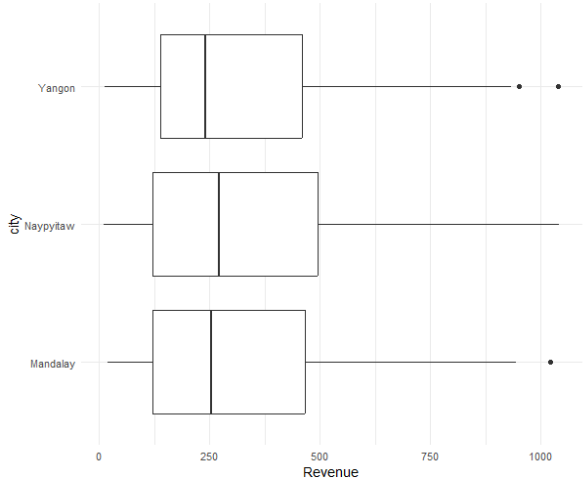
This page shows the Relationships of multiple elements with Quantity.

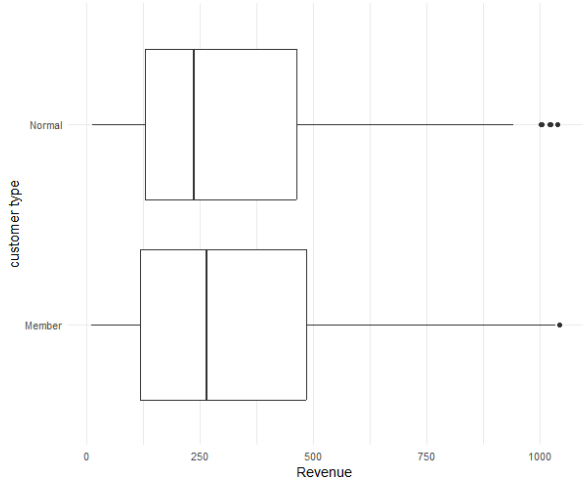
Data Visualisation

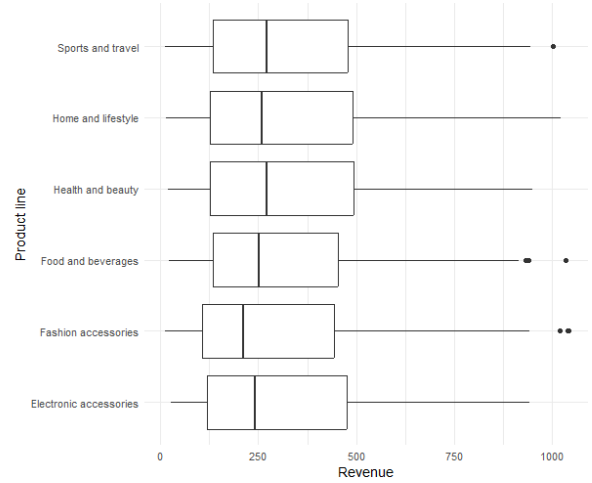




This page shows the Relationships of multiple elements with unit cost

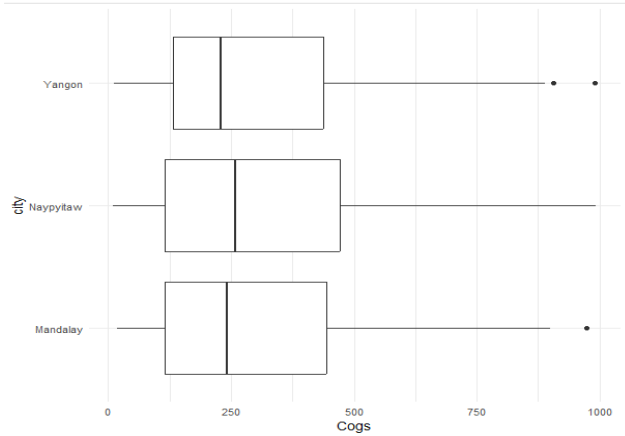


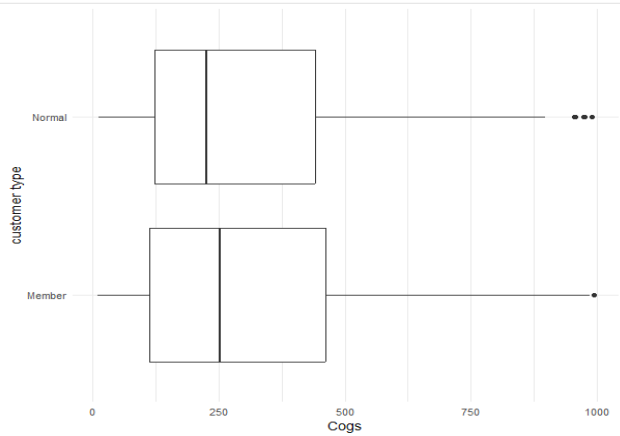


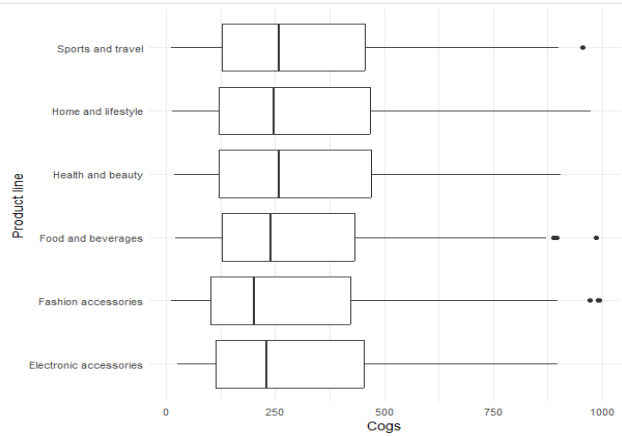


This page shows the Relationships of multiple elements with revenue

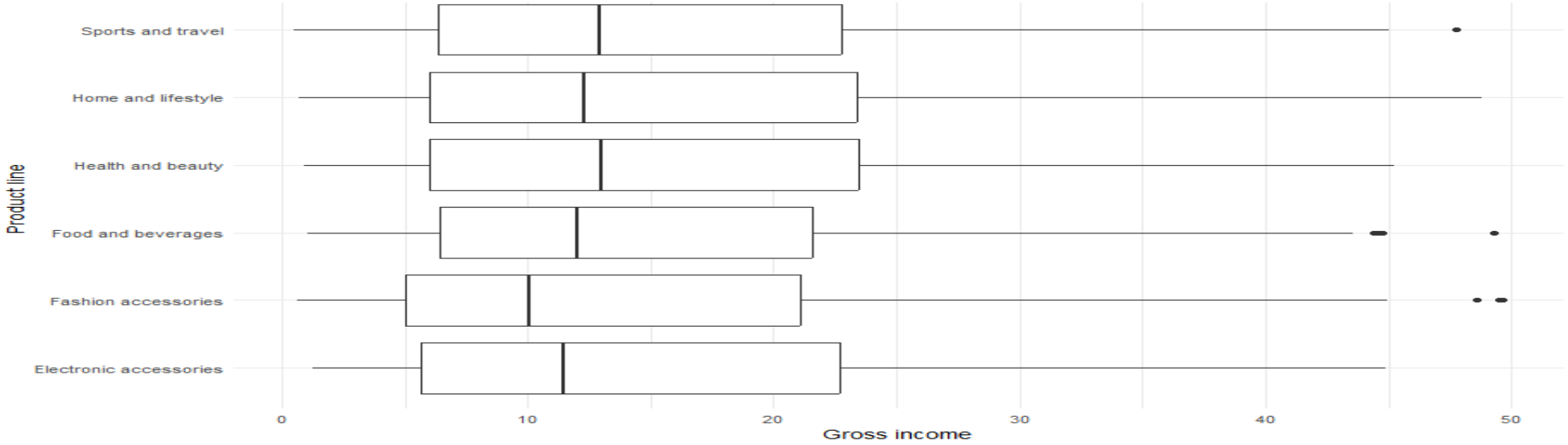
Data Visualisation







This page shows the Relationships of multiple elements with cogs



This page shows the Relationships between productline and gross income

Supervised methods

- **Decision tree:**
- **For gross_income**
  - gross_income is depended for two filed in super market sales (quantity & unit_cost) based on them we make decisions
  - EX : if quantity smaller than 6 and greater than 3 and unit cost lease than 61
    - Then income gross ranged between (1577.3860 , 7.040667)

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

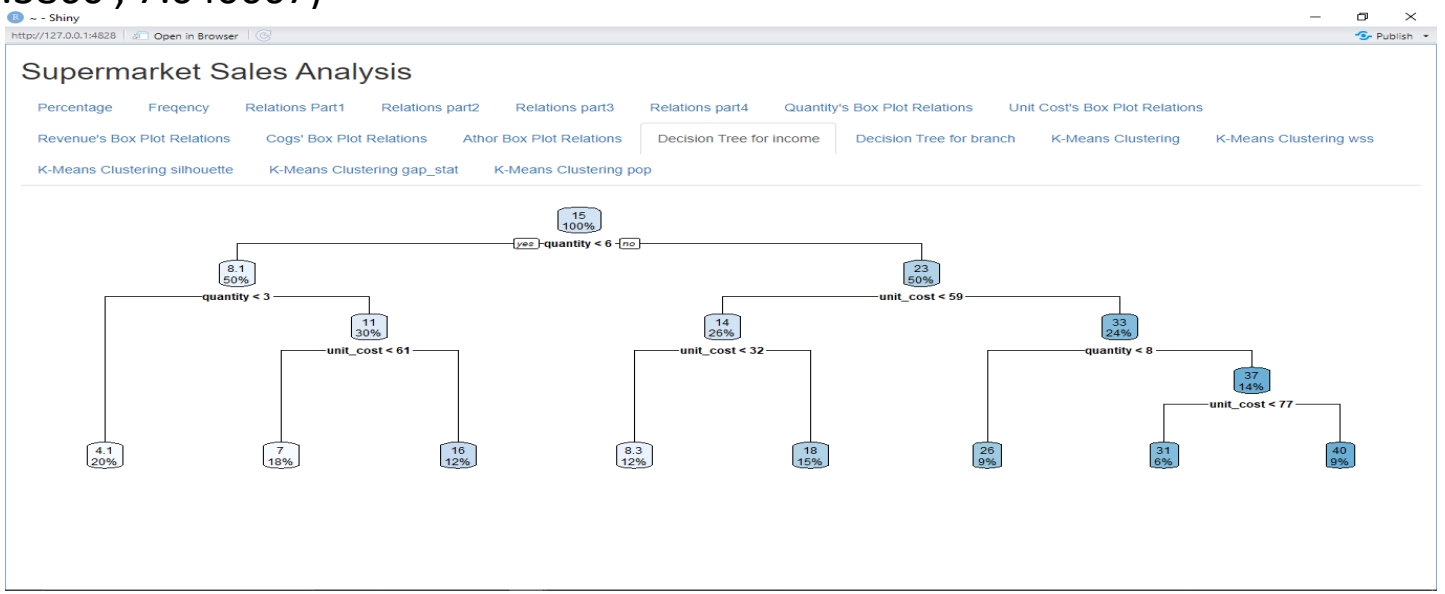
Go to file/function | Addins

Source

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
>   #decision tree for gross_income
>   library("rpart.plot")
> tree_to_income<- rpart( gross_income  ~ unit_cost  + quantity  , data = su
permarket_sales , minsplit = 5)
> tree_to_income
n= 1000

node), split, n, deviance, yval
      * denotes terminal node

 1) root 1000 136959.5000 15.379370
   2) quantity< 5.5 504  16309.8300  8.108294
     4) quantity< 2.5 203   1317.8150  4.090347 *
     5) quantity>=2.5 301   9504.5920 10.818070
      10) unit_cost< 60.88 177   1577.3860  7.040667 *
      11) unit_cost>=60.88 124   1796.5760 16.210010 *
   3) quantity>=5.5 496  66928.4100 22.767720
     6) unit_cost< 58.755 261   9719.3260 13.807750
      12) unit_cost< 31.7 115    836.4740  8.291070 *
      13) unit_cost>=31.7 146   2626.2190 18.153080 *
     7) unit_cost>=58.755 235  12984.1000 32.719010
      14) quantity< 7.5 93   1637.9840 26.463700 *
      15) quantity>=7.5 142   5323.8520 36.815800
        30) unit_cost< 76.5 56    781.0682 31.214740 *
        31) unit_cost>=76.5 86   1641.9850 40.462990 *
```



Through this analysis of the given data we were able to conclude the best plan for income growth

```
> rpart.plot(tree_to_income)
>
> data_to_income<- data.frame( unit_cost = 74.69 , quantity = 5 )
> predict(tree_to_income , newdata = data_to_income)
       1
16.21001
>
```

For branch

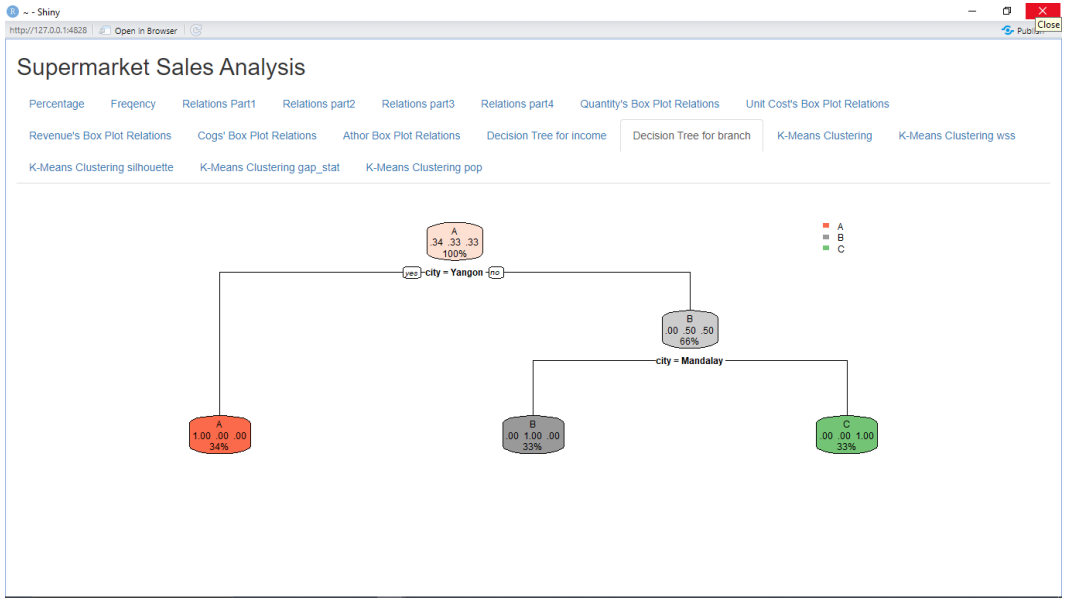
Branch depend on (city) based on them we make decisions

EX: if city is not Yangon and city is not Mandalay

Then branch is C

```
R 4.2.2 · ~/
> #decision tree for branch
> tree_to_branch<- rpart( branch  ~ city  , data = supermarket_sales , minsplit = 5)
> tree_to_branch
n= 1000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 1000 660 A (0.3400000 0.3320000 0.3280000)
  2) city=Yangon 340   0 A (1.0000000 0.0000000 0.0000000) *
  3) city=Mandalay,Naypyitaw 660 328 B (0.0000000 0.5030303 0.4969697)
    6) city=Mandalay 332   0 B (0.0000000 1.0000000 0.0000000) *
    7) city=Naypyitaw 328   0 C (0.0000000 0.0000000 1.0000000) *
> rpart.plot(tree_to_branch)
>
```

Tree in ui



Through this analysis of the given data we were able to conclude the best plan for branch

```
>
> data_to_branch<- data.frame(city = "Yangon" )
> predict(tree_to_branch , newdata = data_to_branch)
  A B C
1 1 0 0
>
```

unsupervised methods

- **K_means clustering**
  - Data for K_means from super market sales data is (unit_cost quantity X5pct_markup revenue cogs gm_pct gross_income rating)
  - Dimantion of K_means data is rows = 1000 and columns = 8

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

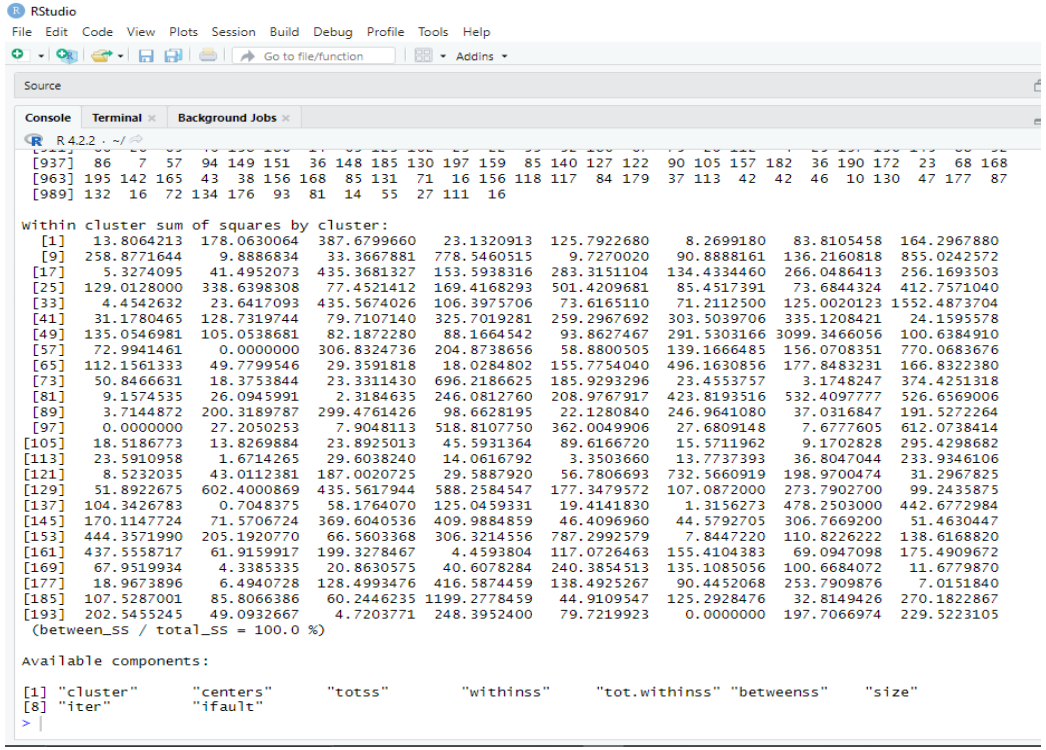
Source

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
> #k_means clustering
> k_data <- supermarket_sales[ , -c(1,2,3,4,5,6,11,12,13)]
> str(k_data)
'data.frame':   1000 obs. of  8 variables:
 $ unit_cost   : num  74.7 15.3 46.3 58.2 86.3 ...
 $ quantity    : int  7 5 7 8 7 7 6 10 2 3 ...
 $ X5pct_markup: num  26.14 3.82 16.22 23.29 30.21 ...
 $ revenue     : num  549 80.2 340.5 489 634.4 ...
 $ cogs        : num  522.8 76.4 324.3 465.8 604.2 ...
 $ gm_pct      : num  4.76 4.76 4.76 4.76 4.76 ...
 $ gross_income: num  26.14 3.82 16.22 23.29 30.21 ...
 $ rating      : num  9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
> dim(k_data)
[1] 1000    8
>
> k_clustering <- kmeans(k_data , centers = 200)
> k_clustering
K-means clustering with 200 clusters of sizes 3, 7, 5, 3, 4, 4, 3, 5, 7, 2, 5, 8, 3, 10, 6, 8, 5, 3, 6, 5, 5, 7, 6,
4, 6, 7, 8, 3, 9, 4, 5, 8, 3, 3, 8, 4, 8, 3, 4, 13, 4, 5, 5, 5, 8, 5, 6, 3, 6, 5, 4, 3, 4, 6, 9, 6, 4, 1, 5, 5, 3,
4, 4, 8, 3, 4, 3, 6, 8, 6, 6, 4, 5, 2, 6, 5, 5, 4, 3, 7, 4, 5, 2, 8, 6, 5, 8, 8, 3, 6, 5, 4, 5, 6, 3, 10, 1, 6, 3,
5, 6, 3, 3, 5, 3, 5, 3, 5, 4, 3, 3, 6, 4, 3, 4, 3, 2, 3, 5, 4, 2, 4, 4, 4, 3, 8, 5, 5, 6, 7, 5, 8, 4, 4, 5, 4, 4,
2, 4, 5, 4, 3, 6, 5, 5, 5, 5, 8, 5, 3, 6, 7, 6, 5, 4, 8, 7, 3, 10, 3, 8, 7, 4, 2, 7, 6, 4, 8, 2, 2, 5, 5, 9, 5, 3,
4, 5, 3, 6, 8, 4, 3, 7, 2, 7, 4, 2, 14, 3, 6, 4, 8, 6, 3, 4, 5, 7, 1, 5, 6

Cluster means:
    unit_cost  quantity X5pct_markup    revenue      cogs   gm_pct gross_income   rating
1    49.83667  4.000000     9.967333  209.31400 199.34667 4.761905     9.967333 7.333333
2    29.24571  4.428571     6.394357  134.28150 127.88714 4.761905     6.394357 7.671429
3    40.57800  9.400000    18.909000  397.08900 378.18000 4.761905    18.909000 8.120000
4    26.36667  2.000000     2.636667   55.37000  52.73333 4.761905     2.636667 5.233333
5    69.45250  4.000000    13.890500  291.70050 277.81000 4.761905    13.890500 5.900000
6    43.64000  2.000000     4.364000   91.64400  87.28000 4.761905     4.364000 5.650000
7    88.58333  5.000000    22.145833  465.06250 442.91667 4.761905    22.145833 7.133333
8    36.14000  9.200000    16.514500  346.80450 330.29000 4.761905    16.514500 5.600000
9    24.39143  4.428571     5.181857  108.81900 103.63714 4.761905     5.181857 6.657143
10   99.75500  7.000000    34.914250  733.19925 698.28500 4.761905    34.914250 6.850000
11   12.31200  1.000000     0.615600   12.92760  12.31200 4.761905     0.615600 8.320000
12   87.00000  5.375000    23.208250  487.37325 464.16500 4.761905    23.208250 6.787500
```

K_means clustering with 200 clusters and cluster means = 100% explained in photo



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

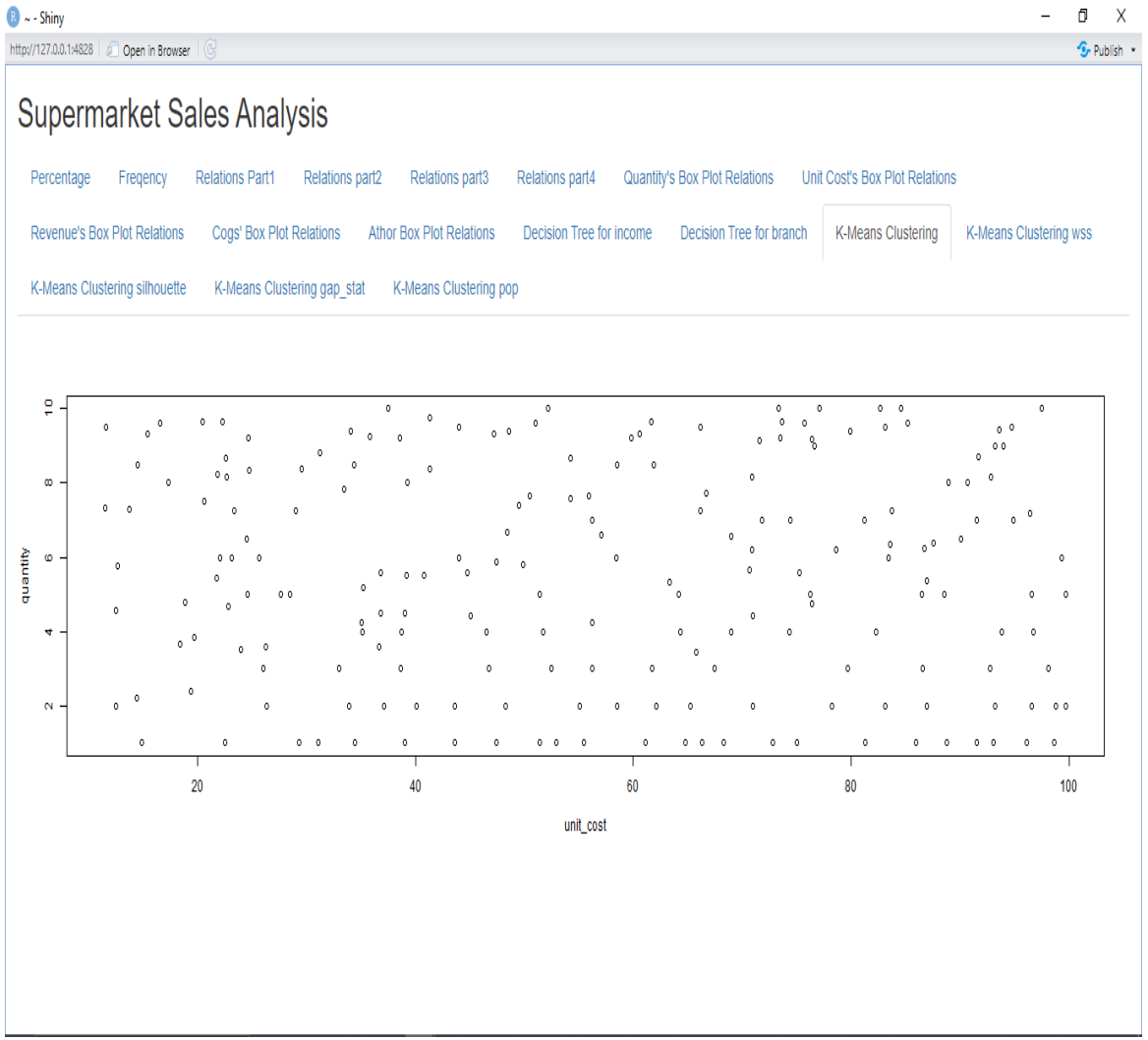
Source

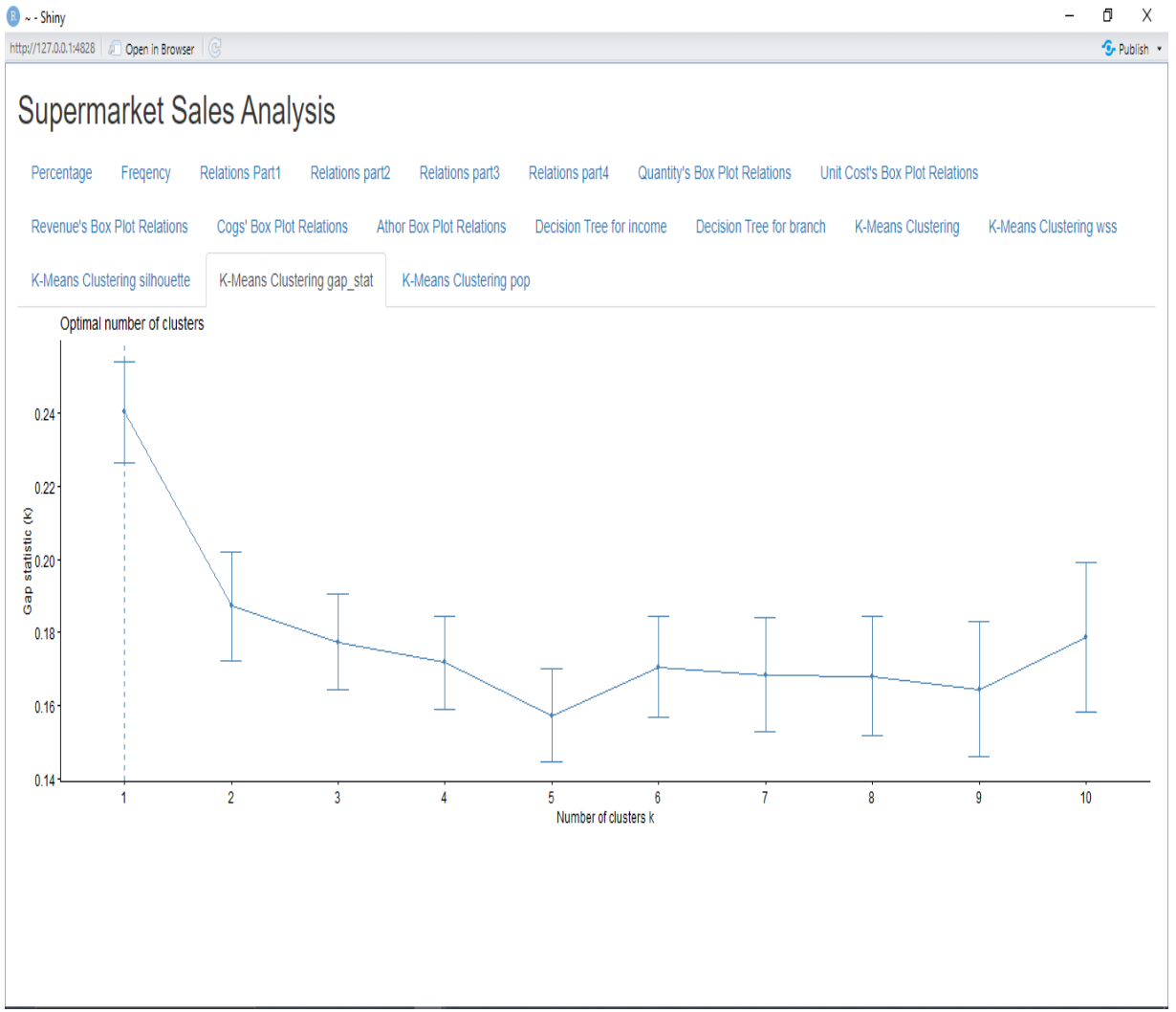
Console | Terminal | Background Jobs

R 4.2.2 · ~/

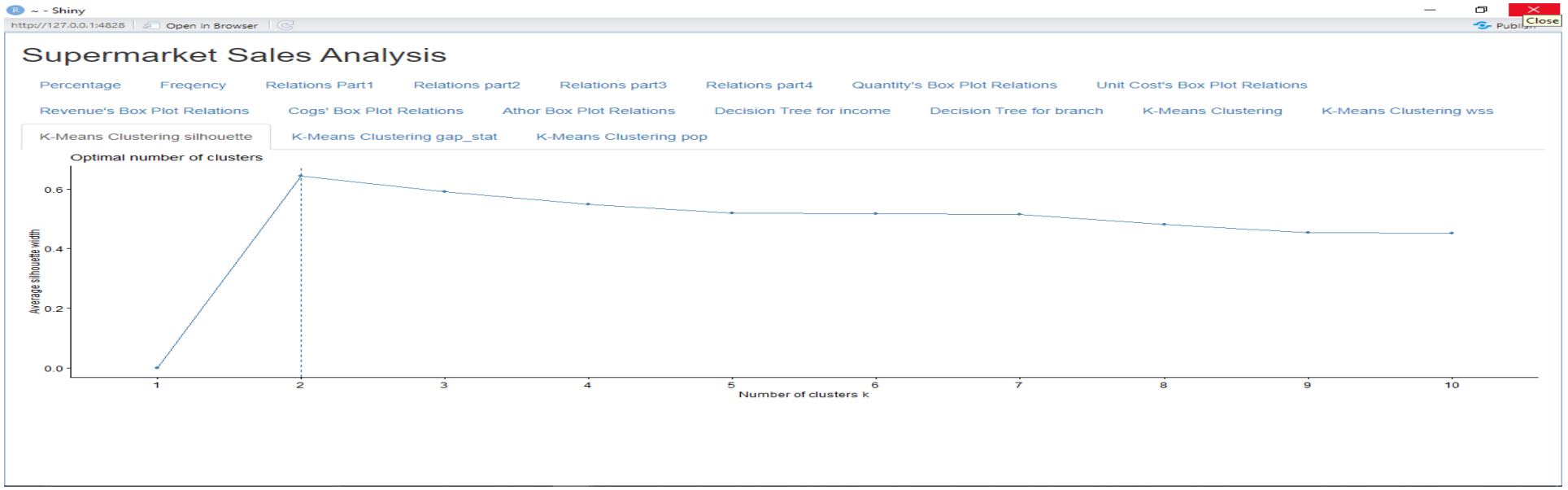
```
 [937]  86   7  57  94 149 151  36 148 185 130 197 159  85 140 127 122  90 105 157 182  36 190 172  23  68 168
 [963] 195 142 165  43  38 156 168  85 131  71  16 156 118 117  84 179  37 113  42  42  46  10 130  47 177  87
 [989] 132  16  72 134 176  93  81  14  55  27 111  16

Within cluster sum of squares by cluster:
  [1]   13.8064213  178.0630064  387.6799660   23.1320913  125.7922680    8.2699180   83.8105458  164.2967880
  [9]  258.8771644    9.8886834   33.3667881  778.5460515    9.7270020   90.8888161  136.2160818  855.0242572
 [17]    5.3274095   41.4952073  435.3681327  153.5938316  283.3151104  134.4334460  266.0486413  256.1693503
 [25]  129.0128000  338.6398308   77.4521412  169.4168293  501.4209681   85.4517391   73.6844324  412.7571040
 [33]    4.4542632   23.6417093  435.5674026  106.3975706   73.6165110   71.2112500  125.0020123 1552.4873704
 [41]   31.1780465  128.7319744   79.7107140  325.7019281  259.2967692  303.5039706  335.1208421   24.1595578
 [49]  135.0546981  105.0538681   82.1872280   88.1664542   93.8627467  291.5303166 3099.3466056  100.6384910
 [57]   72.9941461    0.0000000  306.8324736  204.8738656   58.8800505  139.1666485  156.0708351  770.0683676
 [65]  112.1561333   49.7799546   29.3591818   18.0284802  155.7754040  496.1630856  177.8483231  166.8322380
 [73]   50.8466631   18.3753844   23.3311430  696.2186625  185.9293296   23.4553757    3.1748247  374.4251318
 [81]    9.1574535   26.0945991    2.3184635  246.0812760  208.9767917  423.8193516  532.4097777  526.6569006
 [89]    3.7144872  200.3189787  299.4761426   98.6628195   22.1280840  246.9641080   37.0316847  191.5272264
 [97]    0.0000000   27.2050253    7.9048113  518.8107750  362.0049906   27.6809148    7.6777605  612.0738414
[105]   18.5186773   13.8269884   23.8925013   45.5931364   89.6166720   15.5711962    9.1702828  295.4298682
[113]   23.5910958    1.6714265   29.6038240   14.0616792    3.3503660   13.7737393   36.8047044  233.9346106
[121]    8.5232035   43.0112381  187.0020725   29.5887920   56.7806693  732.5660919  198.9700474   31.2967825
[129]   51.8922675  602.4000869  435.5617944  588.2584547  177.3479572  107.0872000  273.7902700   99.2435875
[137]  104.3426783    0.7048375   58.1764070  125.0459331   19.4141830    1.3156273  478.2503000  442.6772984
[145]  170.1147724   71.5706724  369.6040536  409.9884859   46.4096960   44.5792705  306.7669200   51.4630447
[153]  444.3571990  205.1920770   66.5603368  306.3214556  787.2992579    7.8447220  110.8226222  138.6168820
[161]  437.5558717   61.9159917  199.3278467    4.4593804  117.0726463  155.4104383   69.0947098  175.4909672
[169]   67.9519934    4.3385335   20.8630575   40.6078284  240.3854513  135.1085056  100.6684072   11.6779870
[177]   18.9673896    6.4940728  128.4993476  416.5874459  138.4925267   90.4452068  253.7909876    7.0151840
[185]  107.5287001   85.8066386   60.2446235 1199.2778459   44.9109547  125.2928476   32.8149426  270.1822867
[193]  202.5455245   49.0932667    4.7203771  248.3952400   79.7219923    0.0000000  197.7066974  229.5223105
 (between_SS / total_SS = 100.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
>
```

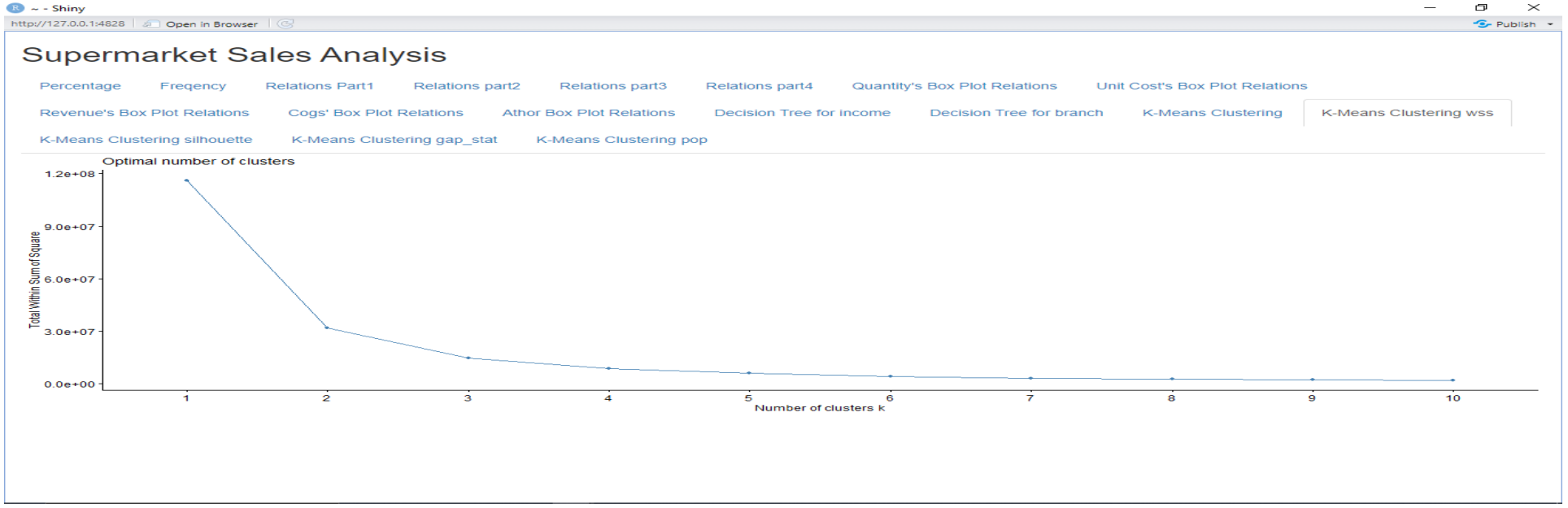
K:mean


~ - Shiny
http://127.0.0.1:4828
Open in Browser
Publish
Supermarket Sales Analysis
Percentage
Freqency
Relations Part1
Relations part2
Relations part3
Relations part4
Quantity's Box Plot Relations
Unit Cost's Box Plot Relations
Revenue's Box Plot Relations
Cogs' Box Plot Relations
Athor Box Plot Relations
Decision Tree for income
Decision Tree for branch
K-Means Clustering
K-Means Clustering wss
K-Means Clustering silhouette
K-Means Clustering gap_stat
K-Means Clustering pop
quantity
unit_cost


~ - Shiny
http://127.0.0.1:4828
Open in Browser
Publish
Supermarket Sales Analysis
Percentage
Freqency
Relations Part1
Relations part2
Relations part3
Relations part4
Quantity's Box Plot Relations
Unit Cost's Box Plot Relations
Revenue's Box Plot Relations
Cogs' Box Plot Relations
Athor Box Plot Relations
Decision Tree for income
Decision Tree for branch
K-Means Clustering
K-Means Clustering wss
K-Means Clustering silhouette
K-Means Clustering gap_stat
K-Means Clustering pop
Optimal number of clusters
Gap statistic (k)
Number of clusters k

K_means


~ - Shiny
http://127.0.0.1:4828
Open in Browser
Close
Supermarket Sales Analysis
Percentage
Freqency
Relations Part1
Relations part2
Relations part3
Relations part4
Quantity's Box Plot Relations
Unit Cost's Box Plot Relations
Revenue's Box Plot Relations
Cogs' Box Plot Relations
Athor Box Plot Relations
Decision Tree for income
Decision Tree for branch
K-Means Clustering
K-Means Clustering wss
K-Means Clustering silhouette
K-Means Clustering gap_stat
K-Means Clustering pop
Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
9
10
Number of clusters k


~ - Shiny
http://127.0.0.1:4828
Open in Browser
Publish
Supermarket Sales Analysis
Percentage
Freqency
Relations Part1
Relations part2
Relations part3
Relations part4
Quantity's Box Plot Relations
Unit Cost's Box Plot Relations
Revenue's Box Plot Relations
Cogs' Box Plot Relations
Athor Box Plot Relations
Decision Tree for income
Decision Tree for branch
K-Means Clustering
K-Means Clustering wss
K-Means Clustering silhouette
K-Means Clustering gap_stat
K-Means Clustering pop
Optimal number of clusters
Total Within Sum of Square
0.0e+00
3.0e+07
6.0e+07
9.0e+07
1.2e+08
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Conclusion:

- The supermarket has a balanced customer distribution across branches, suggesting a widespread customer base.
- Naypyitaw stands out as the dominant city in terms of customer representation, indicating potential opportunities for targeted marketing or branch-specific strategies.
- Gender parity among customers suggests that the supermarket appeals to a diverse audience.
- Fashion accessories are the top-selling product line, emphasizing the importance of this category in the supermarket's offerings.
- A significant portion of customers prefer cash transactions, highlighting the importance of maintaining efficient cash-handling processes.
- These findings provide valuable insights for strategic decision-making, enabling the supermarket to tailor its marketing efforts, optimize inventory for popular product lines, and enhance the overall customer experience based on payment preferences