DATA MINING & VISUALIZATION

ANALYSIS OF SUPERMARKET SALES IN R

By:

Devin Augustin - 2440094352

COMPUTER SCIENCE & STATISTICS

BINUS UNIVERSITY

A. Introduction

Dalam mengerjakan UAS Data Mining and Visualization, kita perlu menggunakan bahasa pemrograman R. Apa itu R? R merupakan bahasa pemrograman yang dapat kita gunakan untuk melakukan komputasi statistik dan presentasi grafis untuk melakukan analisis dan visualisasi terhadap data. R menyediakan banyak banyak teknik statistic dan memiliki banyak package untuk memecahkan masalah yang berbeda. R bersifat open-source dan juga gratis sehingga bisa dipakai oleh siapa saja.

Dataset yang akan saya gunakan merupakan data penjualan supermarket yang berasa dari 3 cabang supermarket selama 3 bulan dan kita membuat model untuk memprediksi rating supermarket menggunakan Decision Tree.

B. Data Description

Dataset supermarket_sales.csv terdiri dari 17 kolom dan 100 baris.

Deskripsi dari dataset Supermaket Sales:

- Invoice ID (char) : Nomor identifikasi slip penjualan yang dihasilkan oleh komputer.
- Branch (char) : Cabang dari supermarket (3 cabang dibagi menjadi A, B, dan C).
- City (char) : Lokasi dari supermarket.
- Customer Type (char) : Tipe customer, yang dibagi menjadi Member bagi yang menggunakan kartu member dan Normal bagi yang tidak menggunakan kartu member.
- Gender (char) : Jenis kelamin customer.
- Product Line (char) : Grup kategorisasi item umum.
- Unit Price (num) : Harga setiap produk.
- Quantity (num) : Jumlah produk yang dibeli oleh customer.
- Tax (num) : 5% pajak untuk customer yang belanja.
- Total (num) : Harga total termasuk pajak.
- Date (char) : Tanggal pembelian.
- Time (‘hms’ num) : Waktu pembelian.
- Payment (char) : Metode pembayaran oleh customer (Cash, Credit Card, dan Ewallet)
- COGS (num) : Cost of goods sold.
- Gross margin percentage (num) : Gross margin percentage.
- Gross income (num) : Gross income.
- Rating (num) : Rating customer dari keseluruhan pengalaman belanja.

C. Data Exploration and Visualization

1. Reading the Dataset

```
> head(sales)
# A tibble: 6 x 17
  `Invoice ID` Branch City    `Customer type` Gender `Product line` `Unit price` Quantity `Tax 5%` Total Date  Time  Payment
  <chr>        <chr>  <chr>   <chr>           <chr>  <chr>                 <dbl>    <dbl>    <dbl> <dbl> <chr> <time> <chr>
1 750-67-8428  A      Yangon  Member          Female Health and be~         74.7        7    26.1   549. 1/5/~ 13:08 Ewallet
2 226-31-3081  C      Naypy~  Normal          Female Electronic ac~         15.3        5     3.82   80.2 3/8/~ 10:29 Cash
3 631-41-3108  A      Yangon  Normal          Male   Home and life~         46.3        7    16.2    341.  3/3/~ 13:23 Credit~
4 123-19-1176  A      Yangon  Member          Male   Health and be~         58.2        8    23.3    489.  1/27~ 20:33 Ewallet
5 373-73-7910  A      Yangon  Normal          Male   Sports and tr~         86.3        7    30.2    634.  2/8/~ 10:37 Ewallet
6 699-14-3026  C      Naypy~  Normal          Male   Electronic ac~         85.4        7    29.9    628.  3/25~ 18:30 Ewallet
# ... with 4 more variables: cogs <dbl>, gross margin percentage <dbl>, gross income <dbl>, Rating <dbl>
> str(sales)
spec_tbl_df [1,000 x 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Invoice ID            : chr [1:1000] "750-67-8428" "226-31-3081" "631-41-3108" "123-19-1176" ...
 $ Branch                : chr [1:1000] "A" "C" "A" "A" ...
 $ City                  : chr [1:1000] "Yangon" "Naypyitaw" "Yangon" "Yangon" ...
 $ Customer type         : chr [1:1000] "Member" "Normal" "Normal" "Member" ...
 $ Gender                : chr [1:1000] "Female" "Female" "Male" "Male" ...
 $ Product line          : chr [1:1000] "Health and beauty" "Electronic accessories" "Home and lifestyle" "Health and beauty" ...
 $ Unit price            : num [1:1000] 74.7 15.3 46.3 58.2 86.3 ...
 $ Quantity              : num [1:1000] 7 5 7 8 7 7 6 10 2 3 ...
 $ Tax 5%                : num [1:1000] 26.14 3.82 16.22 23.29 30.21 ...
 $ Total                 : num [1:1000] 549 80.2 340.5 489 634.4 ...
 $ Date                  : chr [1:1000] "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" ...
 $ Time                  : 'hms' num [1:1000] 13:08:00 10:29:00 13:23:00 20:33:00 ...
  ..- attr(*, "units")= chr "secs"
 $ Payment               : chr [1:1000] "Ewallet" "Cash" "Credit card" "Ewallet" ...
 $ cogs                  : num [1:1000] 522.8 76.4 324.3 465.8 604.2 ...
 $ gross margin percentage: num [1:1000] 4.76 4.76 4.76 4.76 4.76 ...
 $ gross income          : num [1:1000] 26.14 3.82 16.22 23.29 30.21 ...
 $ Rating                : num [1:1000] 9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
```

Tabel di atas menunjukkan data yang berada di dalam dataset. Beberapa variabel pada data diatas berkaitan secara langsung, seperti Branch dan City.

2. Missing Data

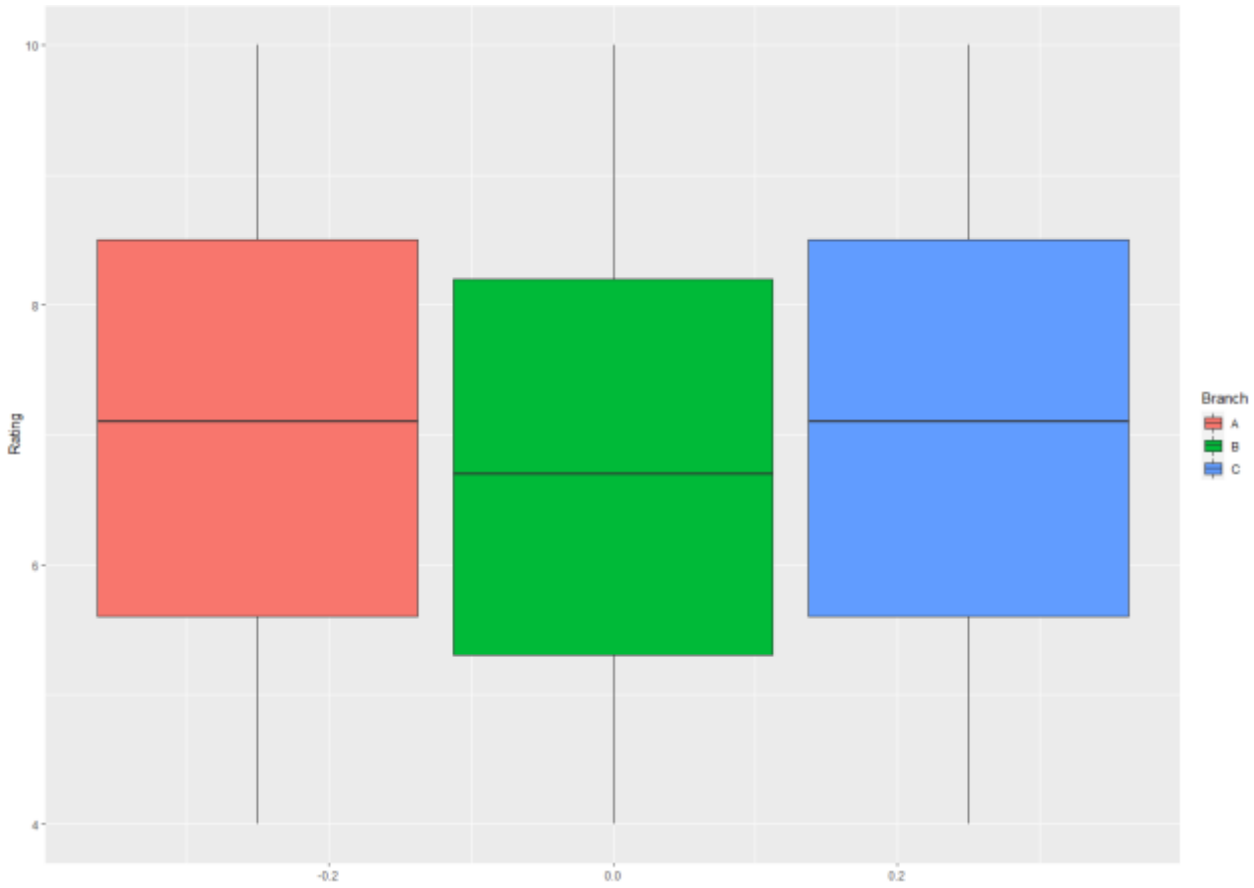
```
> data.frame("Variable"=c(colnames(sales)),
+            "Missing values count"=sapply(sales, function(x)sum(is.na(x))),
+            row.names = NULL)
                 Variable Missing.values.count
1              Invoice ID                     0
2                  Branch                     0
3                    City                     0
4           Customer type                     0
5                  Gender                     0
6            Product line                     0
7              Unit price                     0
8                Quantity                     0
9                   Tax 5%                    0
10                   Total                     0
11                    Date                     0
12                    Time                     0
13                 Payment                     0
14                    cogs                     0
15 gross margin percentage                     0
16            gross income                     0
17                  Rating                     0
```

Dari laporan data di atas, tidak ada missing value pada dataset ini.

3. Outliers

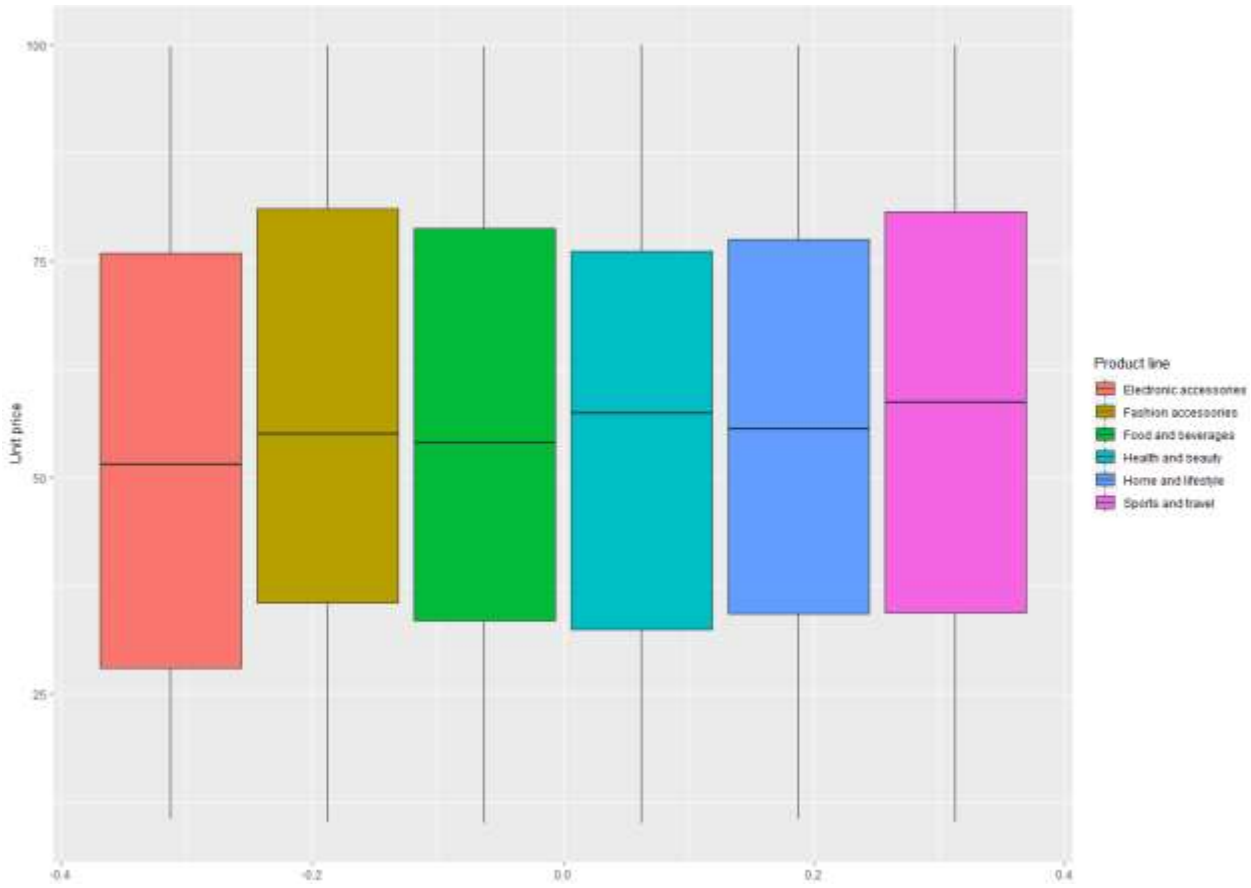
- Rating From Every Branch

```
> ggplot(data=sales)+
+   geom_boxplot(mapping = aes(x=Rating, fill=Branch))+
+   coord_flip()
```



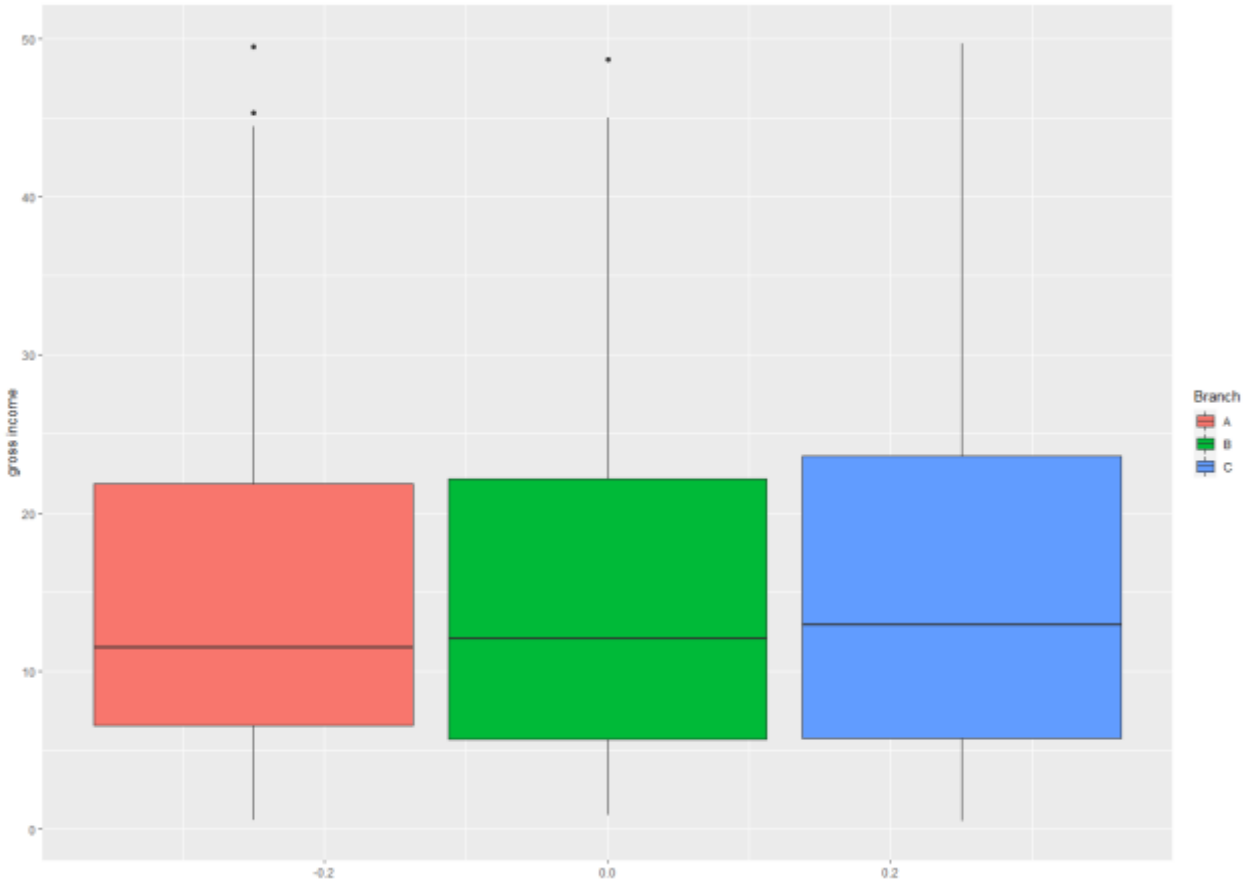
- Unit Price per Product Line

```
> ggplot(data=sales)+
+   geom_boxplot(mapping = aes(x=`Unit price`, fill=`Product line`))+
+   coord_flip()
```



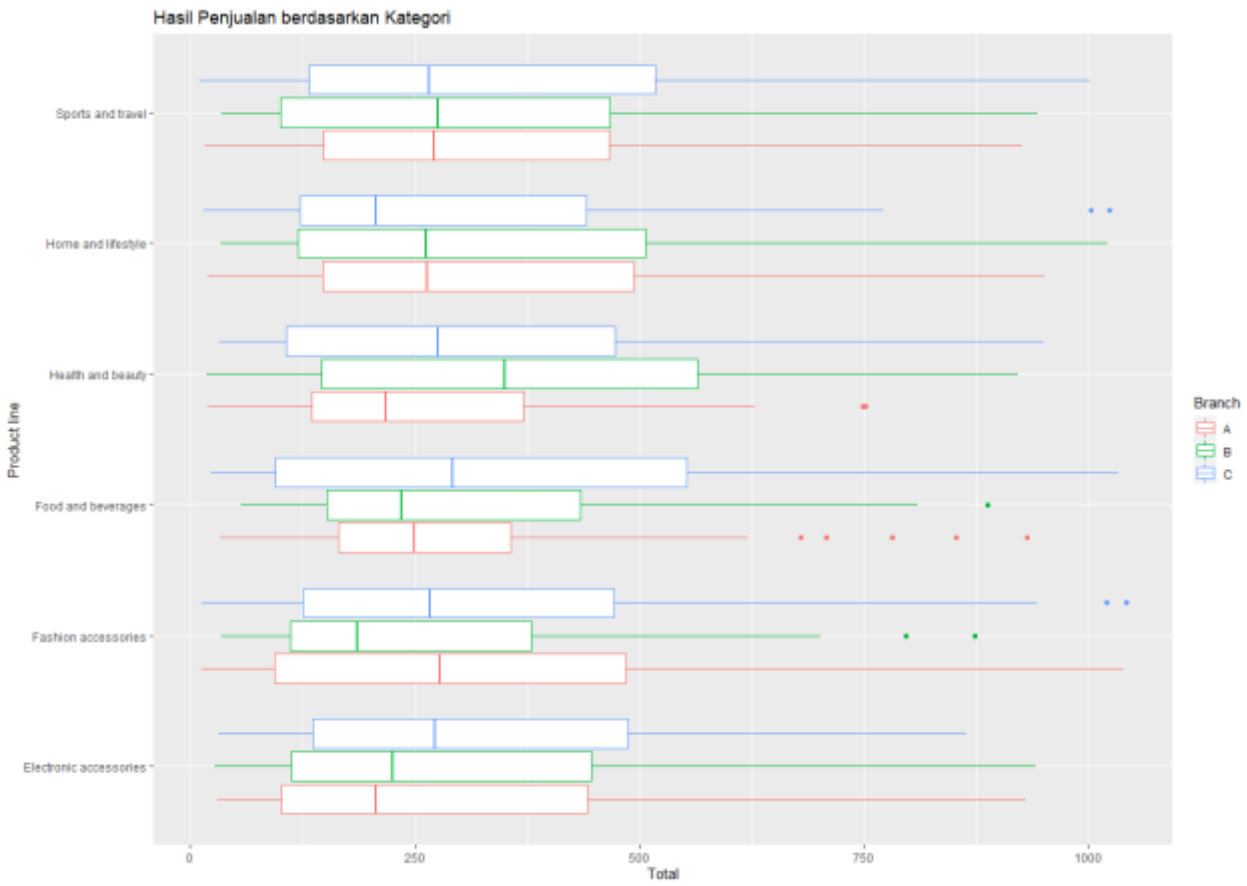
- Gross income From Every Branch

```
> ggplot(data=sales)+
+   geom_boxplot(mapping = aes(x=`gross income`, fill=Branch))+
+   coord_flip()
```



- Hasil Penjualan berdasarkan Kategori

```
> ggplot(data=sales)+
+   geom_boxplot(mapping = aes(x=`Product line`, y=Total, color= Branch))+
+   labs(title = "Hasil Penjualan berdasarkan Kategori")+
+   coord_flip()
```

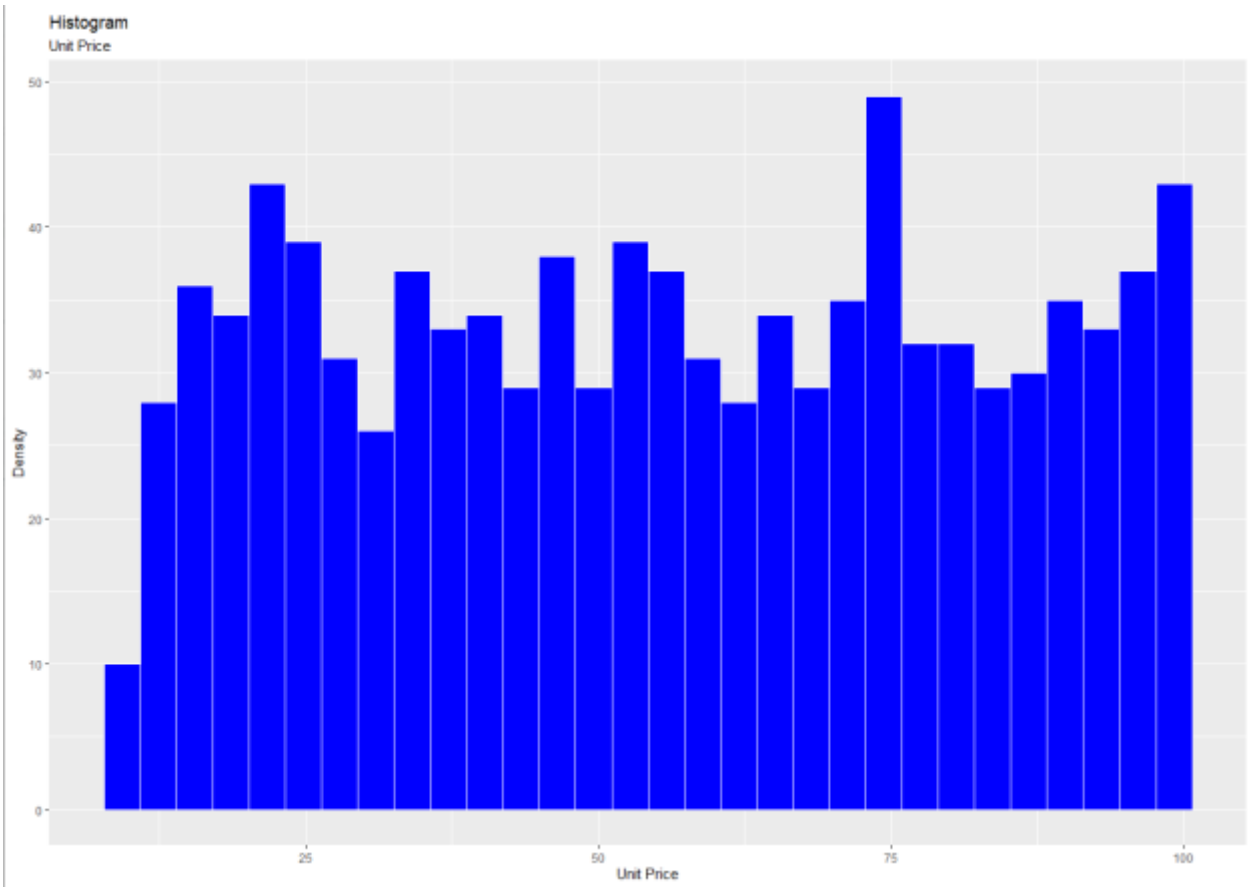


Dari hasil visualisasi data di atas, tidak ada outliers di dalam variabel rating terhadap branch dan juga variabel unit price terhadap product line. Tetapi, terdapat outliers di dalam variabel gross income terhadap branch. Di sini kita bisa melihat bahwa outliers hanya terdapa pada cabang A dan B saja terkait gross income. Untuk hasil penjualan, terdapat beberapa outliers dari masing-masing cabang supermarket khususnya pada produk kategori Food and Beverages.

4. Shape of Distribution

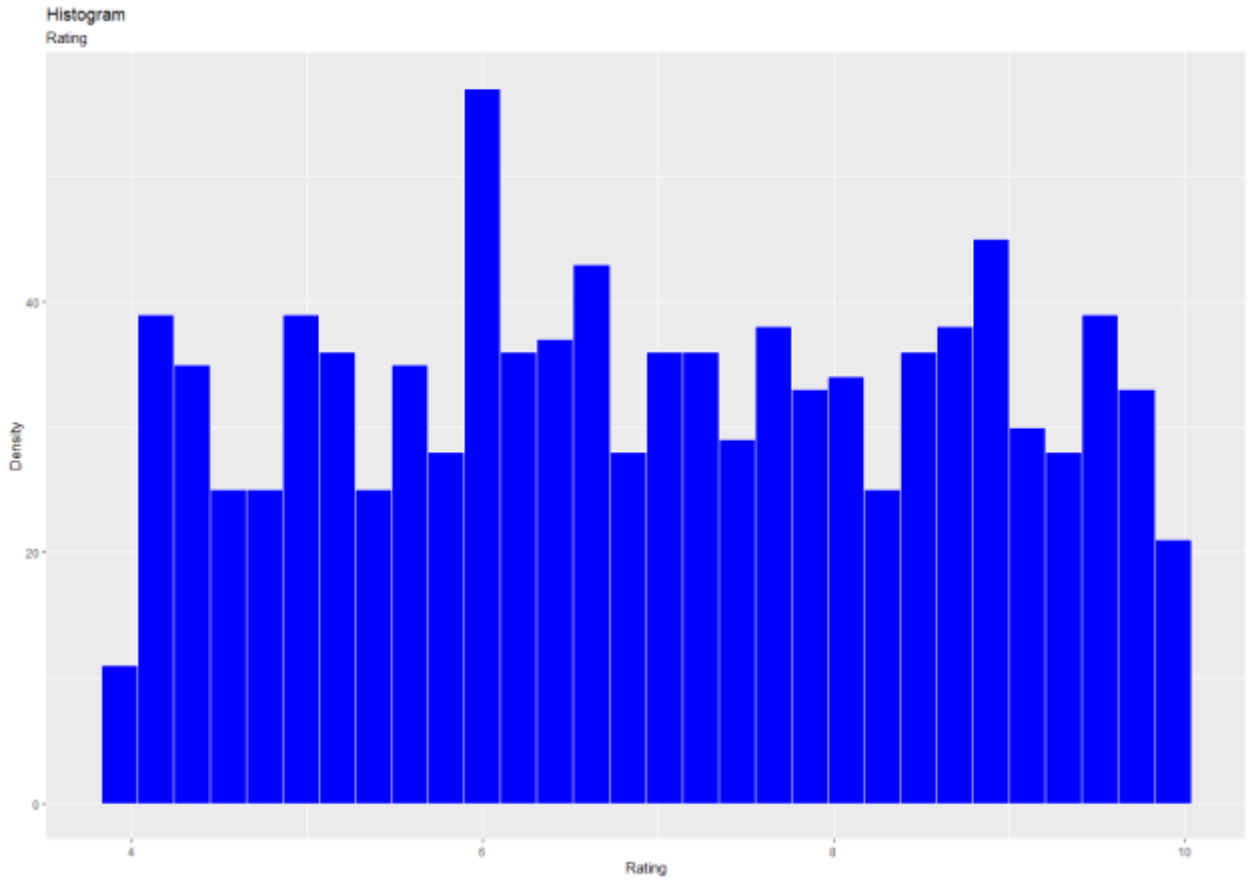
- Unit Price

```
> ggplot(data = sales)+
+   geom_histogram(mapping = aes(x=`Unit price`), col="white",fill="Blue")+
+   labs(x = "Unit Price",
+        y = "Density",
+        title = "Histogram",
+        subtitle = "Unit Price")
```



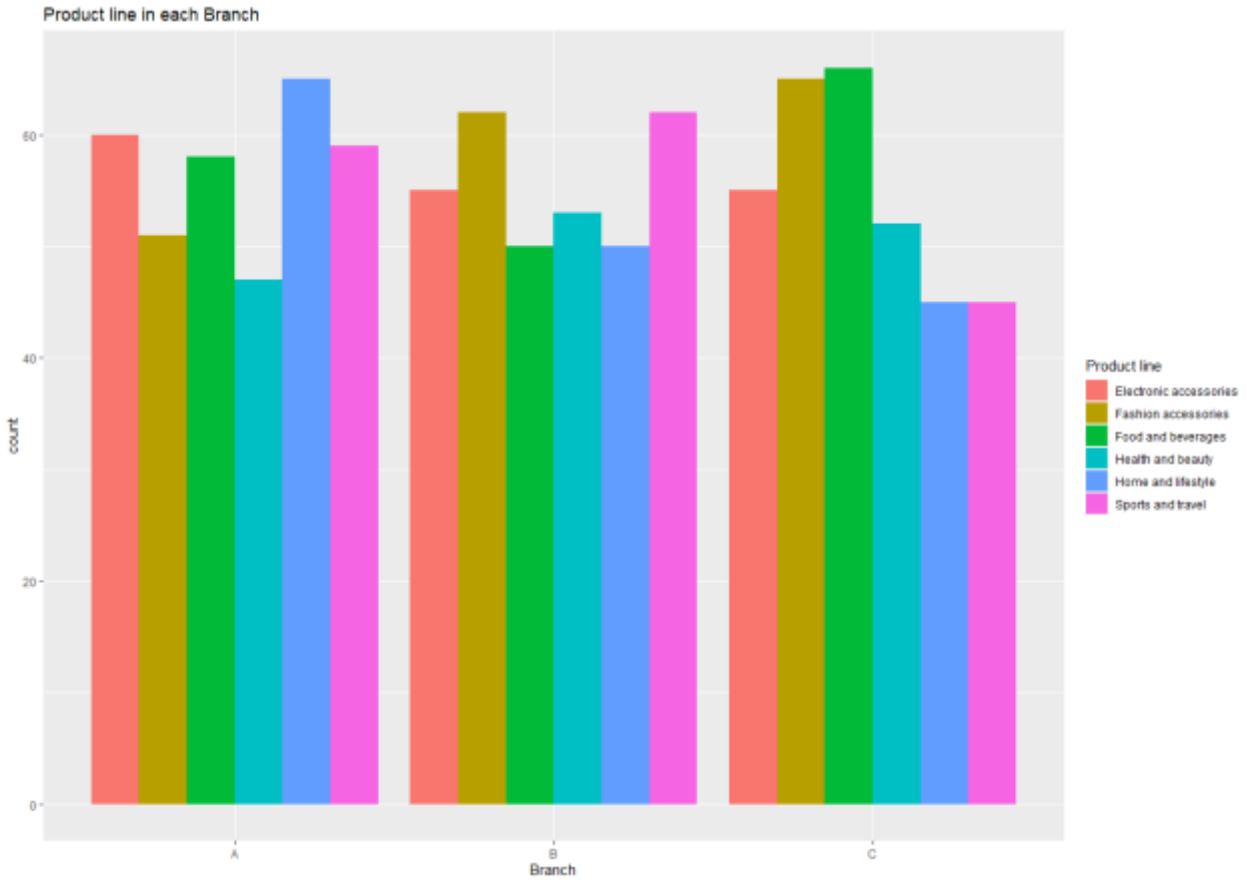
- Rating

```
> ggplot(data = sales)+
+   geom_histogram(mapping = aes(x=Rating), col="white",fill="Blue")+
+   labs(x = "Rating",
+        y = "Density",
+        title = "Histogram",
+        subtitle = "Rating")
```



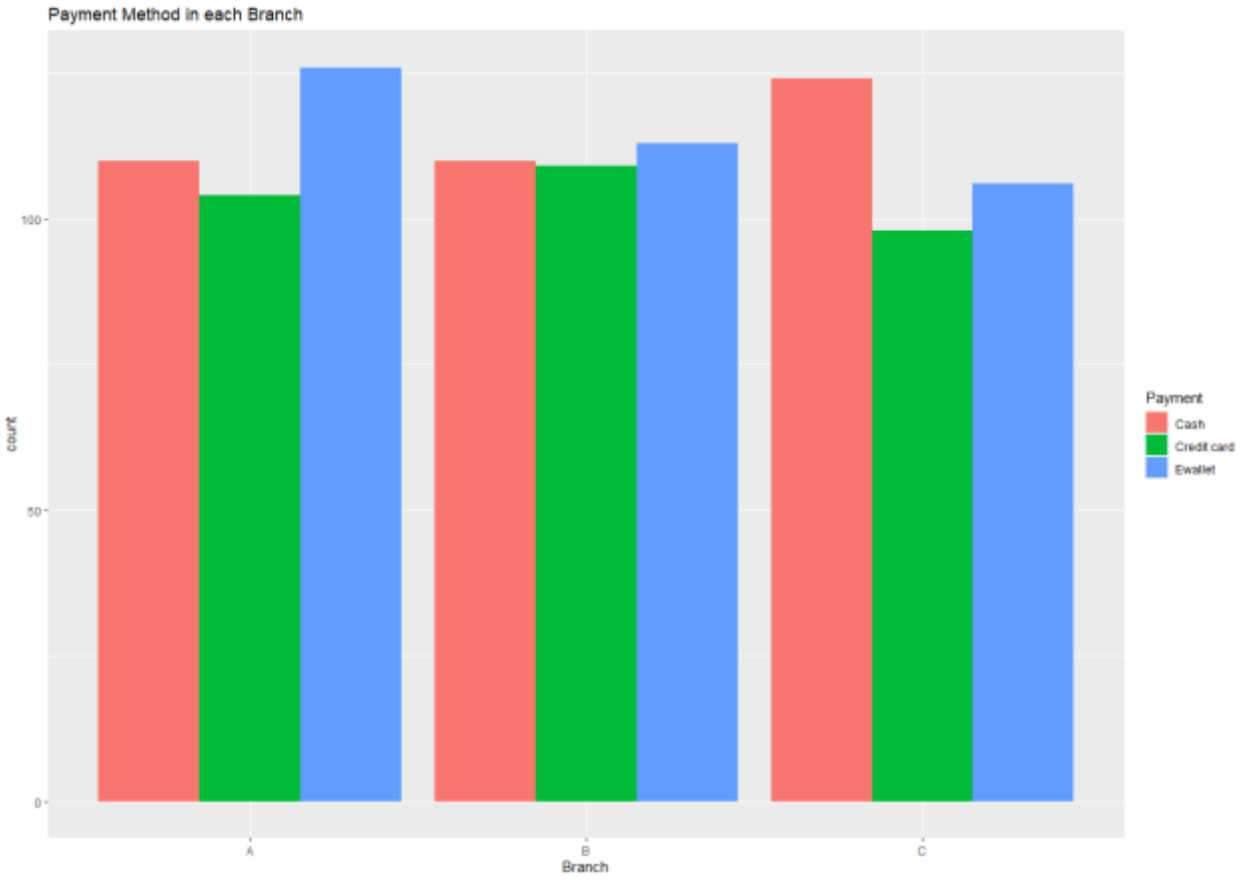
- Product Line in each Branch

```
> ggplot(data=sales)+
+   geom_bar(mapping=aes(x=Branch, fill=`Product line`), position = 'dodge')+
+   labs(title="Product line in each Branch")
```



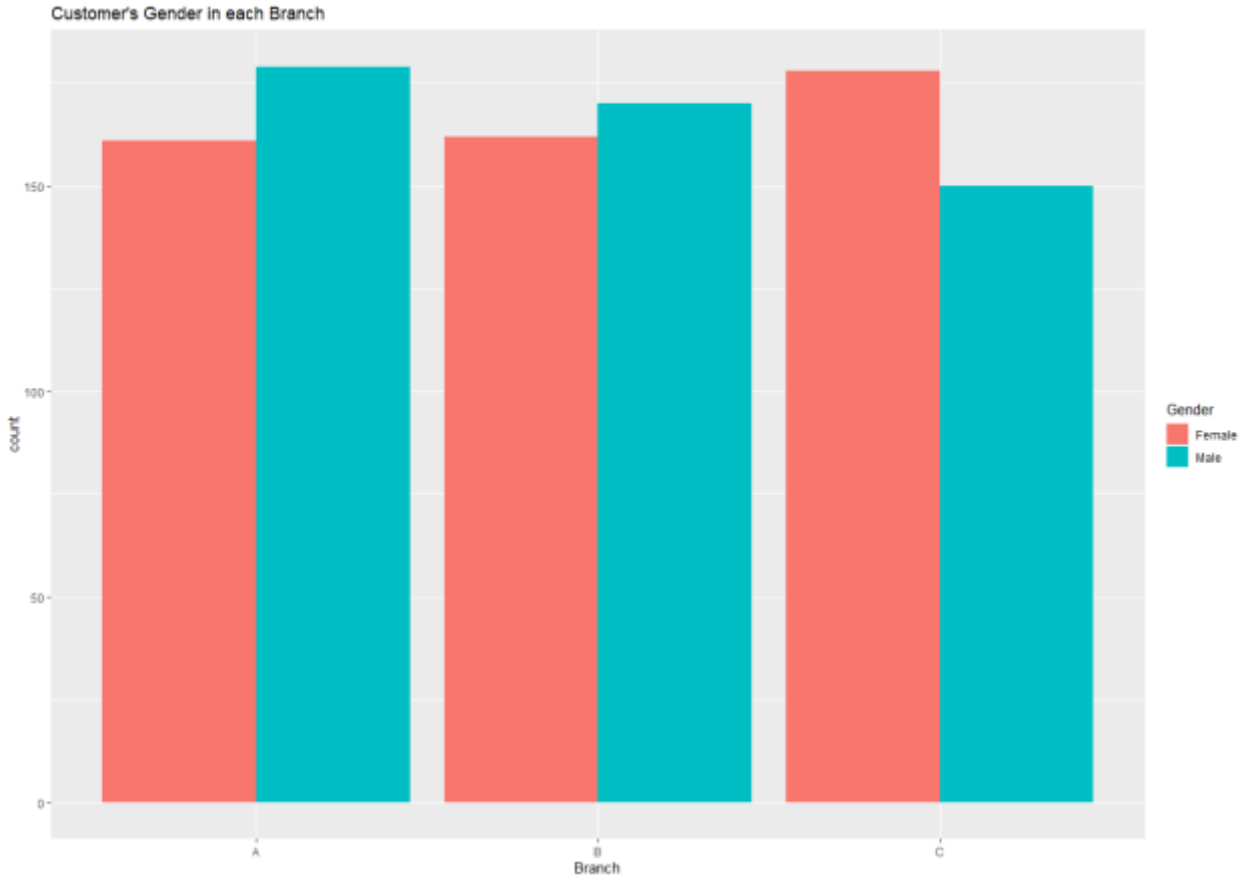
- Payment Method

```
> ggplot(data=sales)+
+   geom_bar(mapping=aes(x=Branch, fill=Payment), position = 'dodge')+
+   labs(title="Payment Method in each Branch")
```



- Customer's Gender

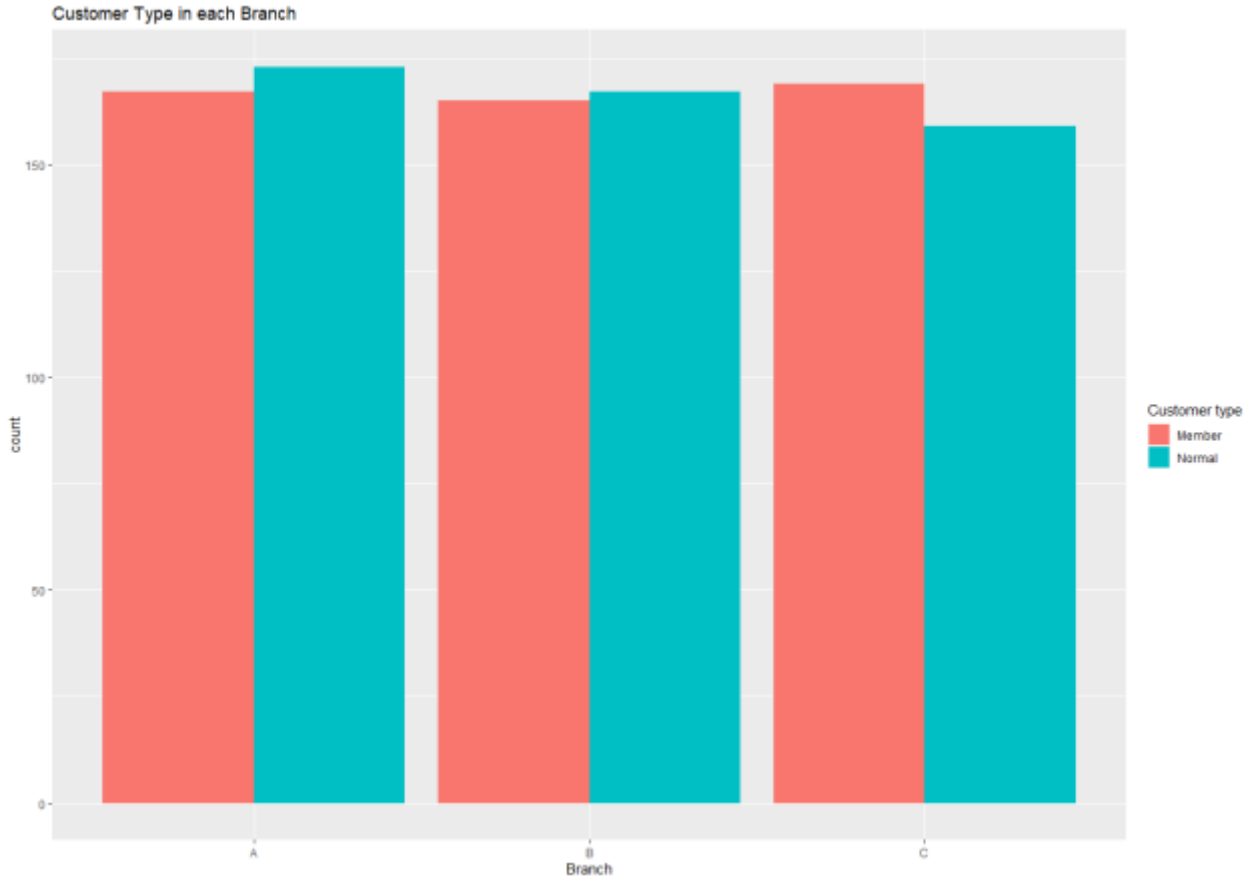
```
> ggplot(data=sales)+
+   geom_bar(mapping=aes(x=Branch, fill=Gender), position = 'dodge')+
+   labs(title="Customer's Gender in each Branch")
```



Di cabang A dan B, customer pria cenderung lebih banyak dibandingkan wanita

- Customer Type

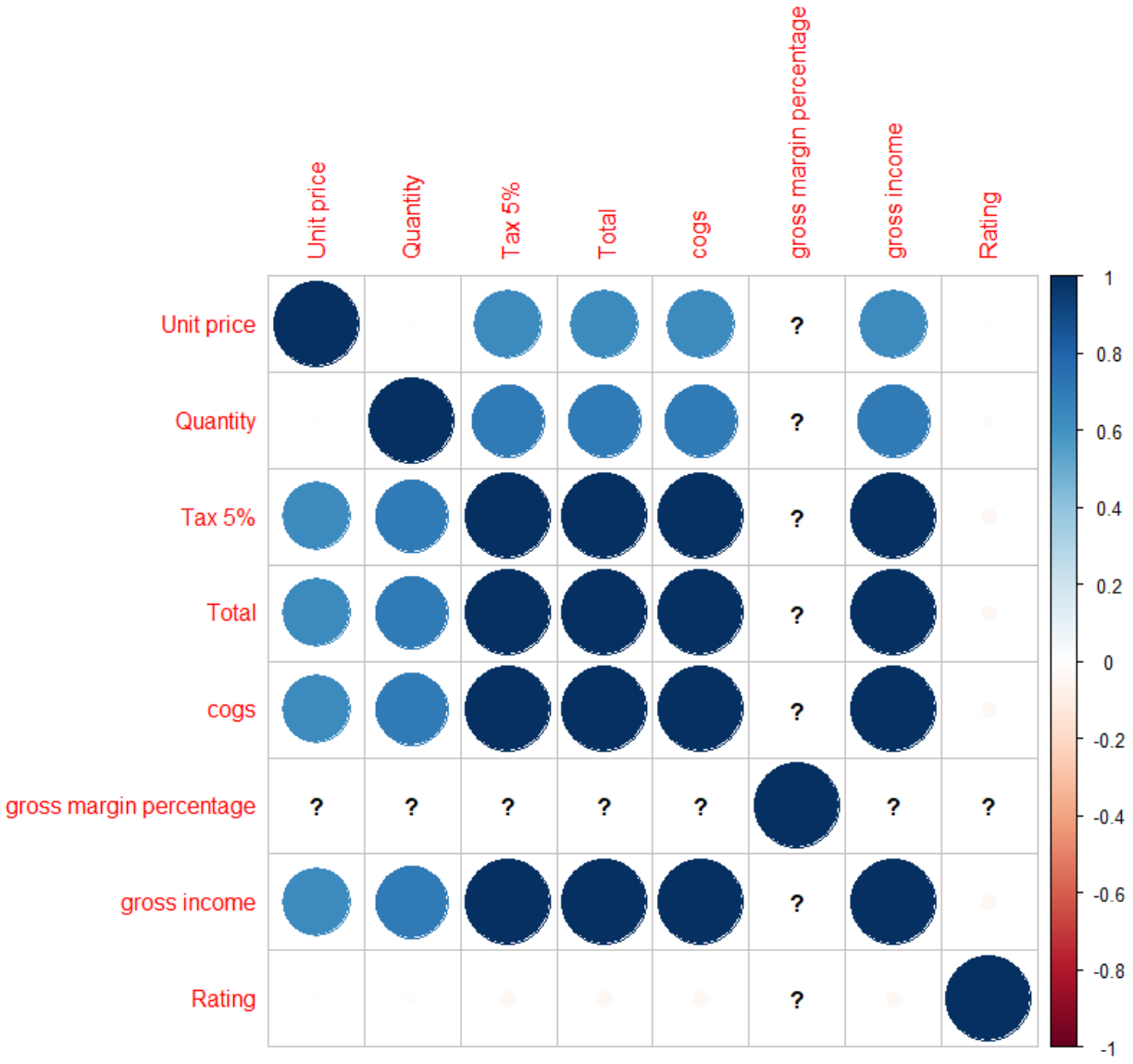
```
> ggplot(data=sales)+
+   geom_bar(mapping=aes(x=Branch, fill=`Customer type`), position = 'dodge')+
+   labs(title="Customer Type in each Branch")
```



Berdasarkan bagan di atas, customer yang berbelanja dengan kartu member lebih sedikit dibandingkan dengan yang berbelanja tanpa kartu member di cabang A dan B kecuali di cabang C.

5. Relationship to Other Numeric Variables

```
> sales.numeric<-sales[sapply(sales, is.numeric)]
> corrplot::corrplot(cor(sales.numeric))
```



6. Prediction Model Using Decision Tree

Kita akan menggunakan C5.0 Decision Tree untuk menentukan variabel yang siginifikan yang bisa memprediksi customer rating. Untuk Decision Tree, kita butuh mengubah kategori 'Rating' menjadi 'Yes' atau 'No' di mana rating di atas 7.5 merupakangood rating dan di bawahnya bad.

```
retail2<-mutate(sales,
                goodrating= ifelse(Rating>7.5, "Yes", "No"))

retail2$goodrating <- as.factor(retail2$goodrating)

retail3<-select(retail2, Branch, 'Customer type', Gender, 'Product line', 'Unit price', Quantity, Total, Month, goodrating)
```

Kemudian bagi data menjadi dua untuk training dan testing data

```
set.seed(321)
train_sample <- sample(1000,800)

retail_train <- retail3[train_sample, ]
retail_test <- retail3[-train_sample,]
```

```
> retail_model <- C5.0(retail_train[-9], retail_train$goodrating)
> retail_model

Call:
C5.0.default(x = retail_train[-9], y = retail_train$goodrating)

Classification Tree
Number of samples: 800
Number of predictors: 8

Tree size: 118

Non-standard options: attempt to group attributes
```

```
> retail_pred <- predict(retail_model, retail_test)
> CrossTable(retail_test$goodrating, retail_pred,
+            prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
+            dnn = c('actual default', 'predicted default'))


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  200


               | predicted default
actual default |        No |       Yes | Row Total |
---------------|-----------|-----------|-----------|
            No |        83 |        42 |       125 |
               |     0.415 |     0.210 |           |
---------------|-----------|-----------|-----------|
           Yes |        53 |        22 |        75 |
               |     0.265 |     0.110 |           |
---------------|-----------|-----------|-----------|
  Column Total |       136 |        64 |       200 |
---------------|-----------|-----------|-----------|
```

```
Evaluation on training data (800 cases):

           Decision Tree
         ----------------
         Size      Errors

          118  141(17.6%)   <<


           (a)   (b)    <-classified as
         ----  ----
          428    47    (a): class No
           94   231    (b): class Yes


        Attribute usage:

        100.00% Branch
        100.00% Product line
         53.38% Gender
         39.63% Unit price
         39.50% Month
         38.38% Customer type
         29.50% Quantity
         14.38% Total


Time: 0.0 secs
```

D. Discussion / Analysis

Dari hasil analisa dataset di atas, kita dapat melihat supermarket A dan C memiliki rating yang serupa, sedangkan supermarket B memiliki rating terendah. Barang yang dijual pun memiliki harga yang serupa jika dilihat dari rata-rata harga tiap kategori produk yang tersedia. Pada variable 'Gross Income', supermarket cabang C memiliki pendapatan yang paling banyak dibandingkan dengan supermarket cabang A dan B.

Untuk total penjual berdasarkan kategori, kita dapat melihat produk *health and beauty* dan *food and beverages* memiliki angka yang relatif rendah pada cabang A. Di cabang B, produk yang memiliki penjualan rendah adalah produk *Fashion Accessories* sedangkan pada cabang C produk yang memiliki total penjualan rendah adalah *Home and Lifestyle.* Hal ini bisa saja terjadi karena faktor lokasi dari masing-masing cabang supermarket.

Untuk prediction modeling, di sini kita menggunakan C5.0 Decision Tree untuk menentukan apakah ada variabel yang cukup siginifikan yang bisa memprediksi customer rating. Dari 17 variabel di dataset, kita hanya akan menggunakan 7 variabel, yaitu 'Branch', 'Product Line', 'Gender', 'Unit Price', 'Month', 'Customer Type', 'Quantity', dan "Total'. Dari hasil modeling 800 total sampel, kita mendapatkan error rate sebesar 17.6%. Walaupun demikian, kita dapat melihat bahwa variabel 'Branch' dan 'Product Line' memiliki peran yang sangat penting bagi 'Rating' dari customer.