Market

Gözde Nur Özdemir

2023-09-12

MARKET DATA SET

This data set is taken from Kaggle

Power BI

This graph is made by POWER BI.

Preperation

```
# Please first download required library
library(tidyverse)
library(ggplot2)
library(dplyr)
library(rvest)
library(stringr)
library(corrplot)
```

General Information About Data Set

```
data<-read.csv("C:/Users/gozde/Desktop/market/supermarket_sales.csv")
# Assuming 'data' is your data frame
cols_to_factor <- c("Gender", "Customer.type", "City", "Branch", "Product.line", "Payment")

# Use lapply to factorize selected columns
data[cols_to_factor] <- lapply(data[cols_to_factor], factor)
```

```r
# View the summary
summary(data)
```

```
##   Invoice.ID        Branch         City     Customer.type    Gender
##  Length:1000        A:340   Mandalay :332   Member:501    Female:501
##  Class :character   B:332   Naypyitaw:328   Normal:499    Male  :499
##  Mode  :character   C:328   Yangon   :340
##
##
##
##                  Product.line   Unit.price       Quantity         Tax.5.
##  Electronic accessories:170   Min.   :10.08   Min.   : 1.00   Min.   : 0.5085
##  Fashion accessories   :178   1st Qu.:32.88   1st Qu.: 3.00   1st Qu.: 5.9249
##  Food and beverages    :174   Median :55.23   Median : 5.00   Median :12.0880
##  Health and beauty     :152   Mean   :55.67   Mean   : 5.51   Mean   :15.3794
##  Home and lifestyle    :160   3rd Qu.:77.94   3rd Qu.: 8.00   3rd Qu.:22.4453
##  Sports and travel     :166   Max.   :99.96   Max.   :10.00   Max.   :49.6500
##      Total             Date              Time                  Payment
##  Min.   :  10.68   Length:1000        Length:1000        Cash       :344
##  1st Qu.: 124.42   Class :character   Class :character   Credit card:311
##  Median : 253.85   Mode  :character   Mode  :character   Ewallet    :345
##  Mean   : 322.97
##  3rd Qu.: 471.35
##  Max.   :1042.65
##       cogs        gross.margin.percentage  gross.income         Rating
##  Min.   : 10.17   Min.   :4.762           Min.   : 0.5085   Min.   : 4.000
##  1st Qu.:118.50   1st Qu.:4.762           1st Qu.: 5.9249   1st Qu.: 5.500
##  Median :241.76   Median :4.762           Median :12.0880   Median : 7.000
##  Mean   :307.59   Mean   :4.762           Mean   :15.3794   Mean   : 6.973
##  3rd Qu.:448.90   3rd Qu.:4.762           3rd Qu.:22.4453   3rd Qu.: 8.500
##  Max.   :993.00   Max.   :4.762           Max.   :49.6500   Max.   :10.000
```

Features:

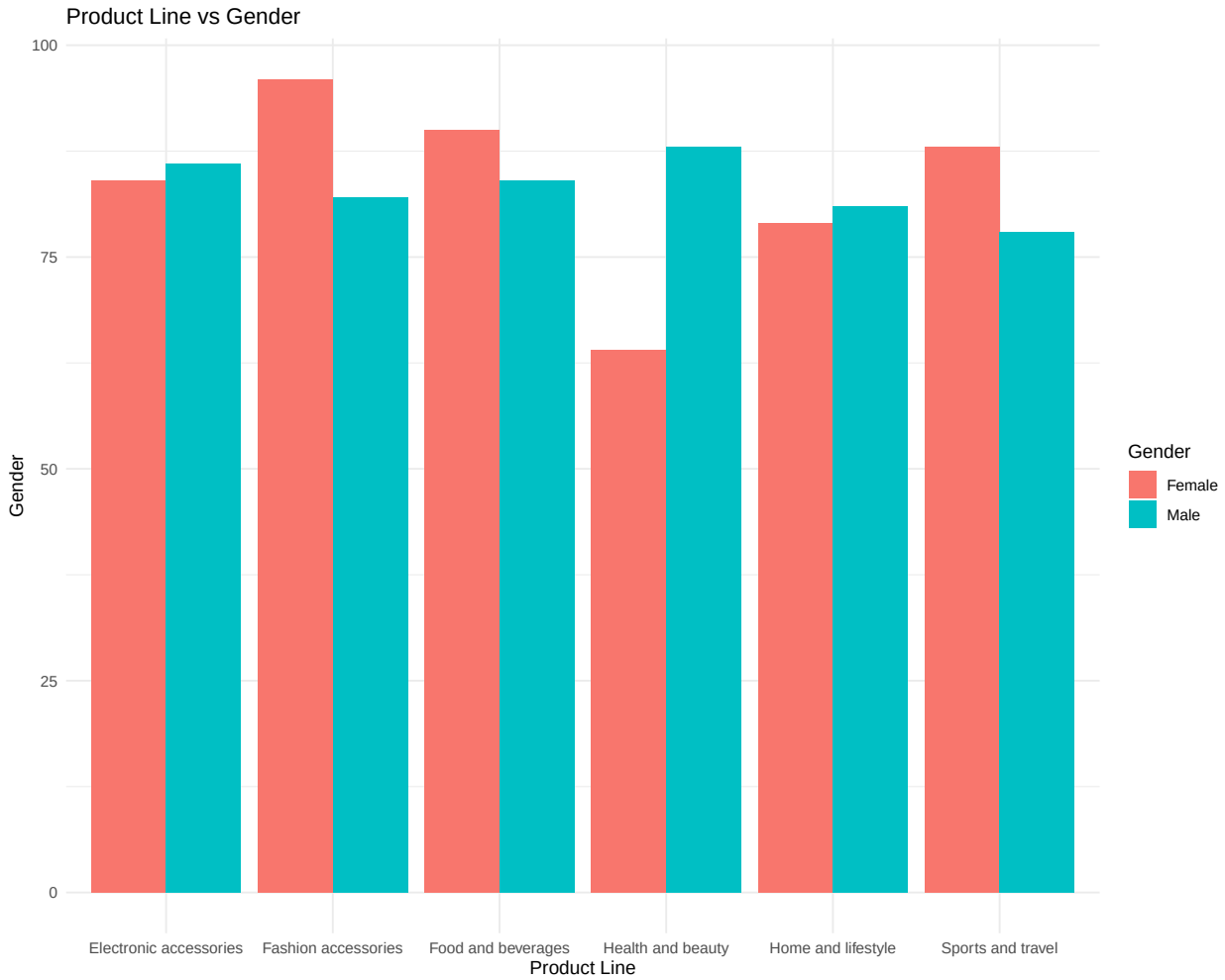
1. Invoice ID: Unique identifier for each invoice.
2. Branch: The branch where the purchase was made.
3. City: The city where the purchase was made.
4. Customer Type: Type of customer, e.g., “Member” or “Normal.”
5. Gender: Gender of the customer.
6. Product Line: The category or type of product.
7. Unit Price: Price per unit of the product.
8. Quantity: Number of units purchased.
9. Tax 5%: Tax amount as a percentage of the total.
10. Total: Total amount including tax.

Plot of Market Sales Data

Gender vs Product Line

```
ggplot(data = data, aes(x=Product.line, fill = Gender)) +
  geom_histogram(stat="count",position="dodge")+theme_minimal()+
  xlab("Product Line")+ylab("Gender")+labs(title="Product Line vs Gender")
```

```
## Warning in geom_histogram(stat = "count", position = "dodge"): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```



As we can see from the plot, the Majority of Females tend to buy Fashion accessories. But Males tend to buy health and beauty stuff.

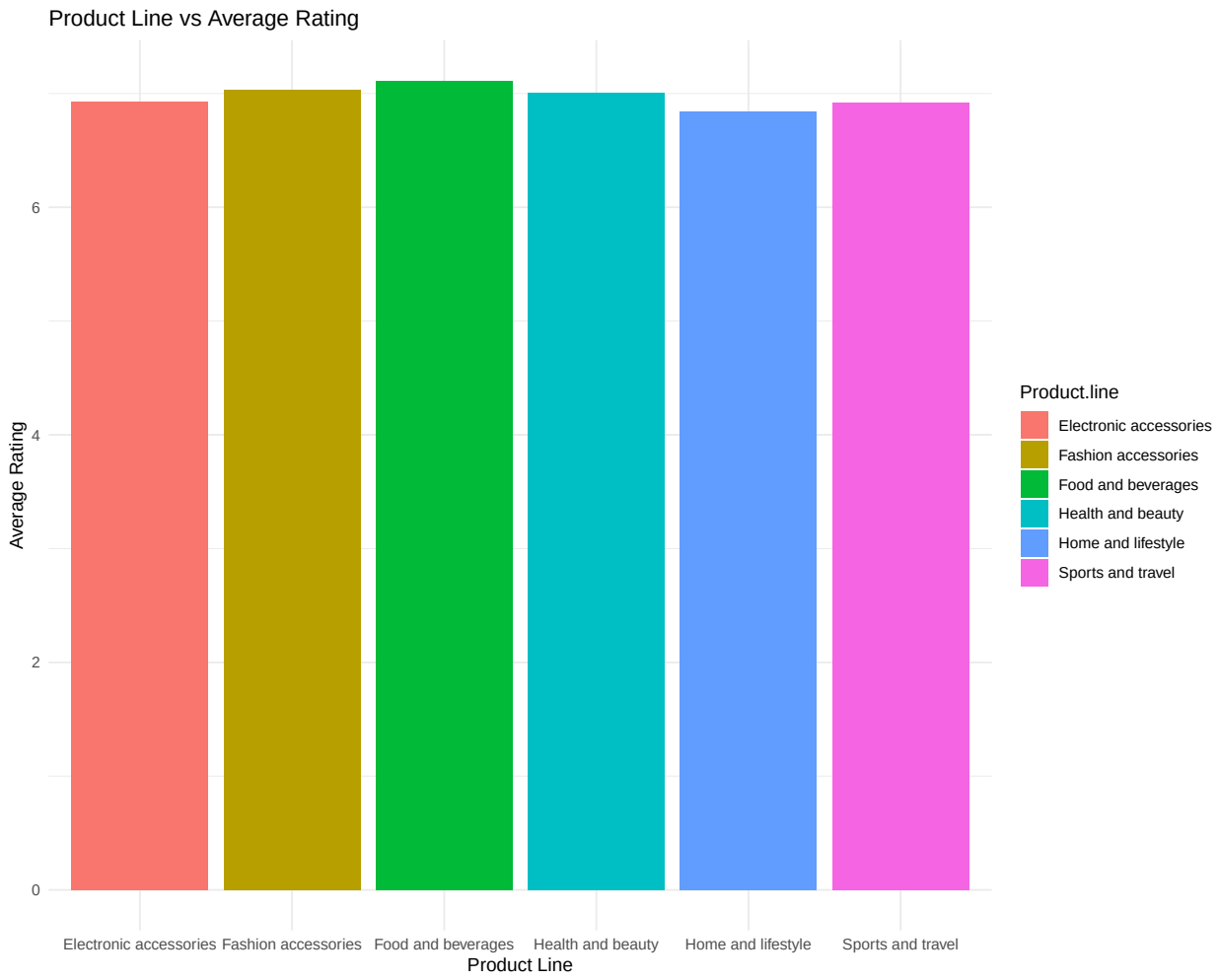
Average Rating vs Product Line

```
average_ratings <- data %>%
  group_by(Product.line) %>%
```

```
  summarize(Average_Rating = mean(Rating, na.rm = TRUE))



# Create a bar plot using ggplot2 with average rating on the y-axis
ggplot(data = average_ratings, aes(x = Product.line, y = Average_Rating, fill = Product.line)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  xlab("Product Line") +
  ylab("Average Rating") +
  labs(title = "Product Line vs Average Rating")
```

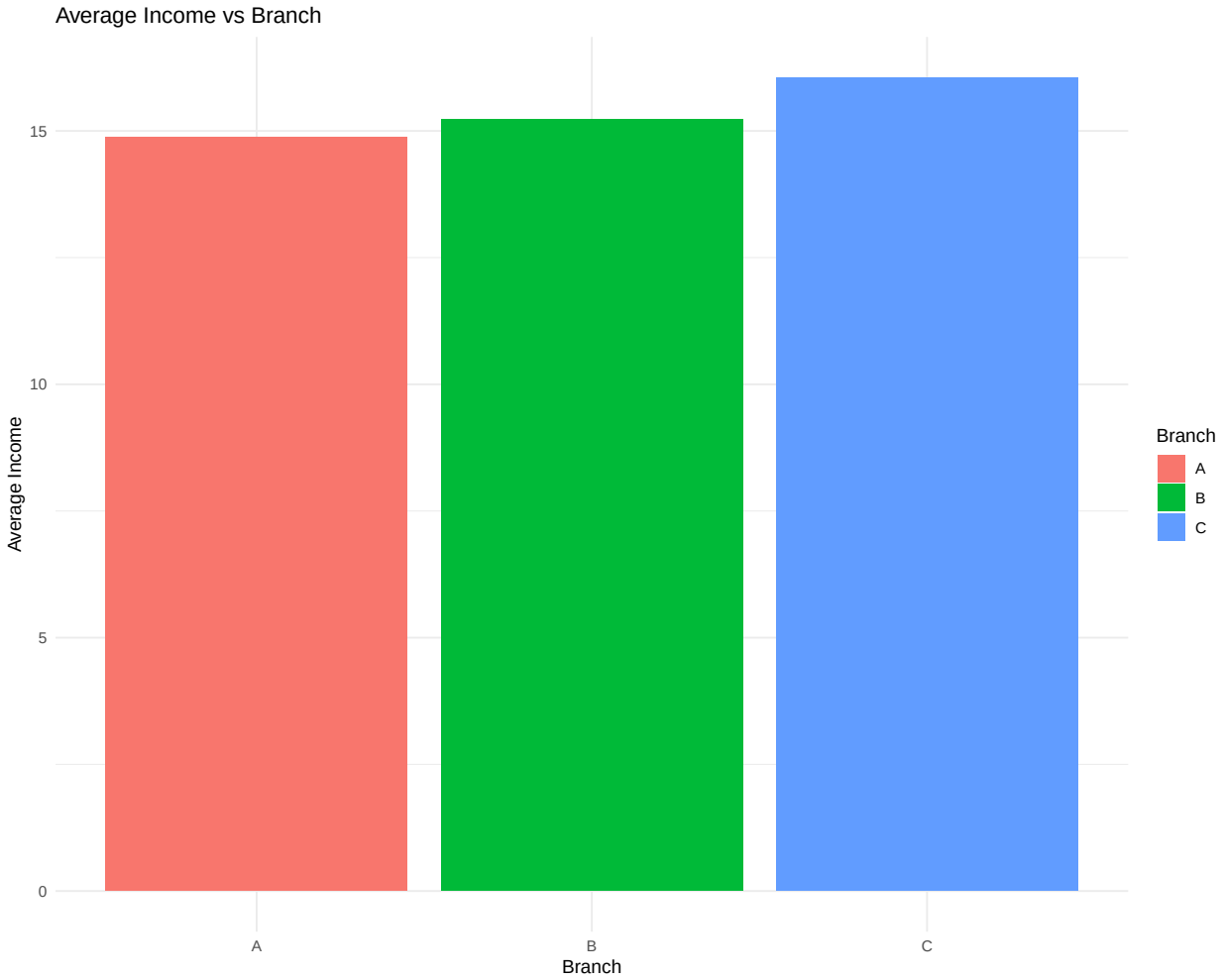


Food and beverages are mostly likeable.

Average Income vs Branch

```
average_income=data %>% group_by(Branch) %>%
  summarize(average=mean(gross.income))
```

```
ggplot(data = average_income, aes(y = average ,x=Branch, fill = Branch)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  xlab("Branch") +
  ylab("Average Income") +
  labs(title = "Average Income vs Branch")
```

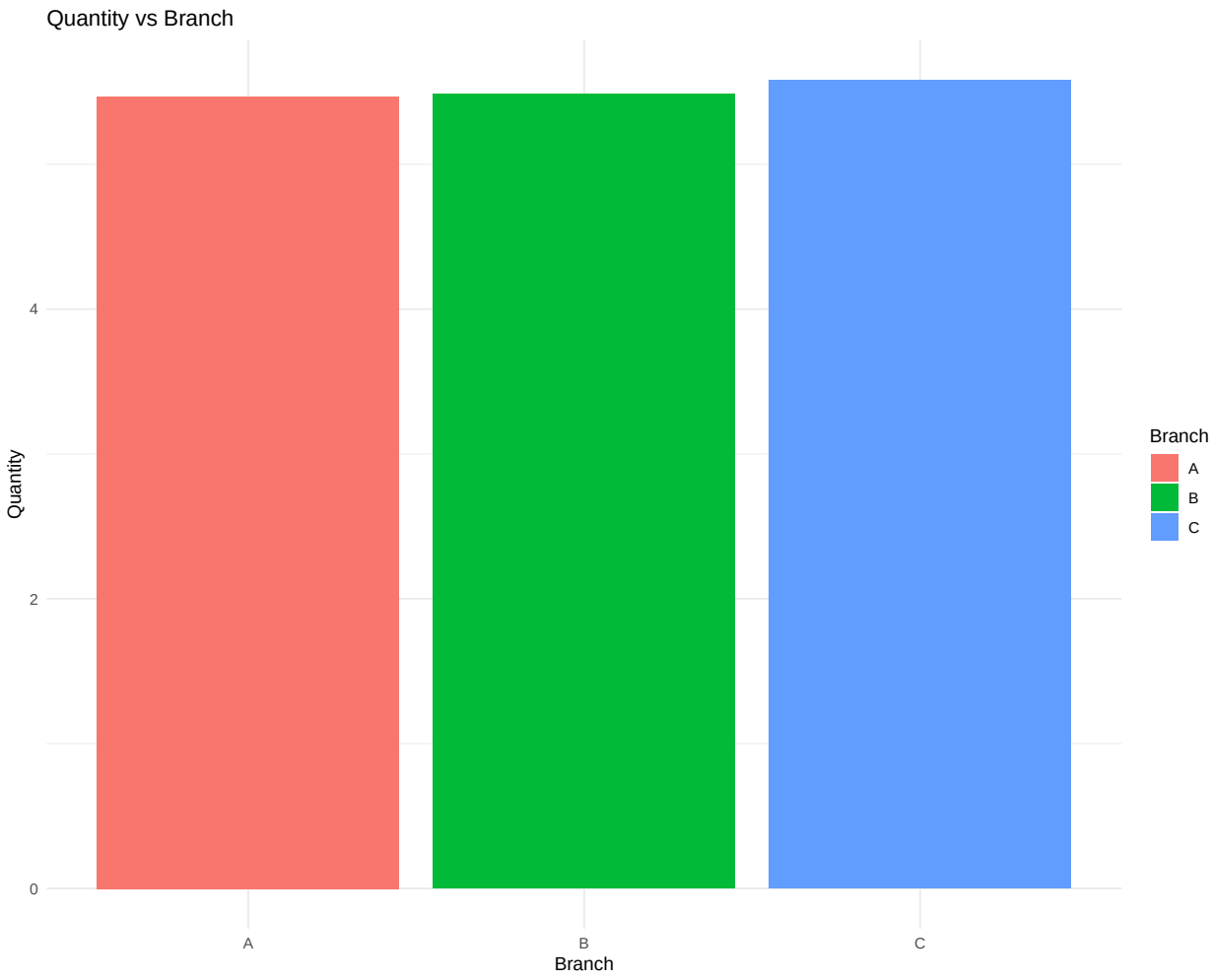


The branch C is makes the most income other than the Branch A and B.

Quantity vs Branch

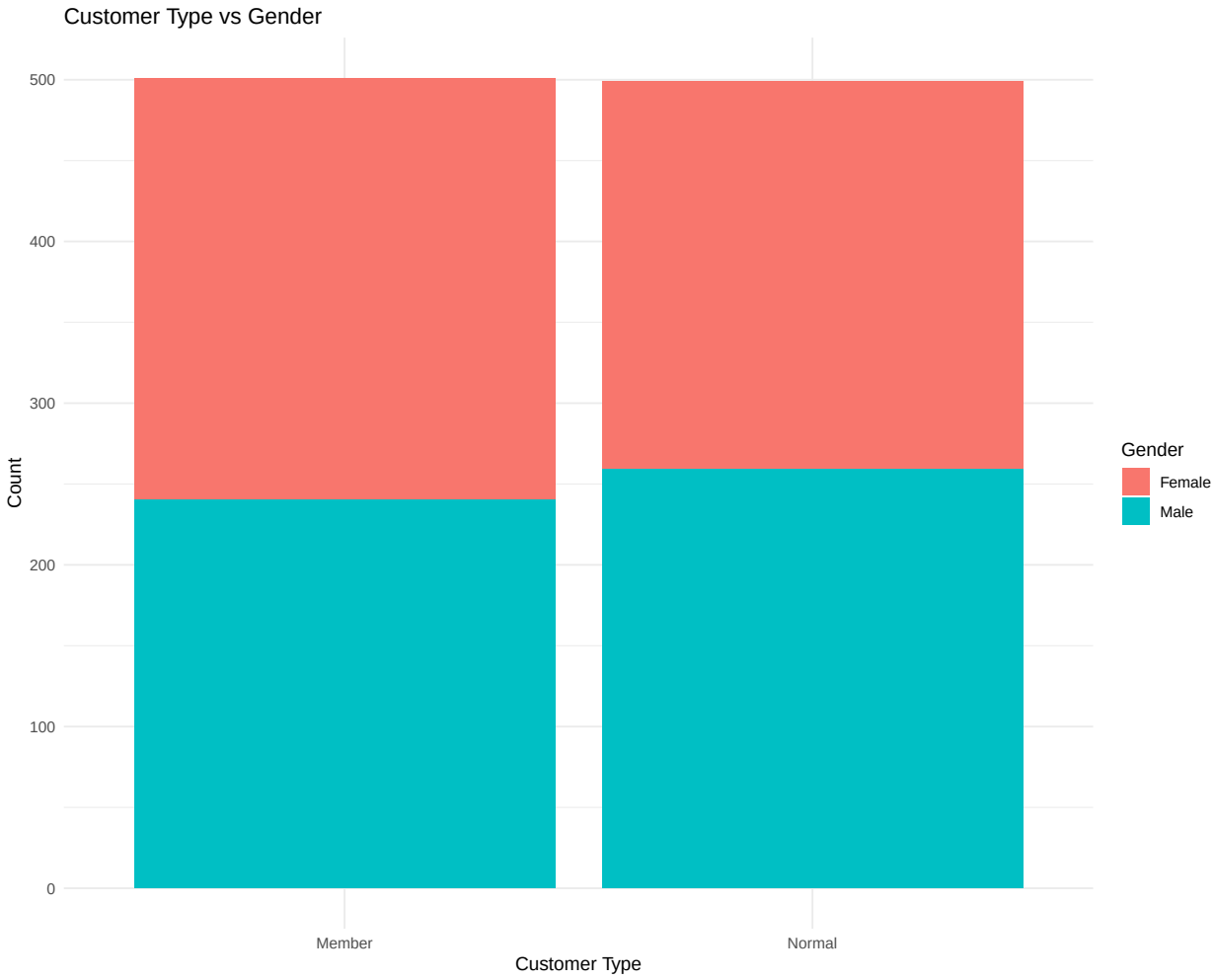
```
average_quantity=data %>% group_by(Branch) %>%
  summarize(average=mean(Quantity))

ggplot(data = average_quantity, aes(y = average ,x=Branch, fill = Branch)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  xlab("Branch") +
  ylab("Quantity") +
  labs(title = "Quantity vs Branch")
```



Customer Type vs Gender

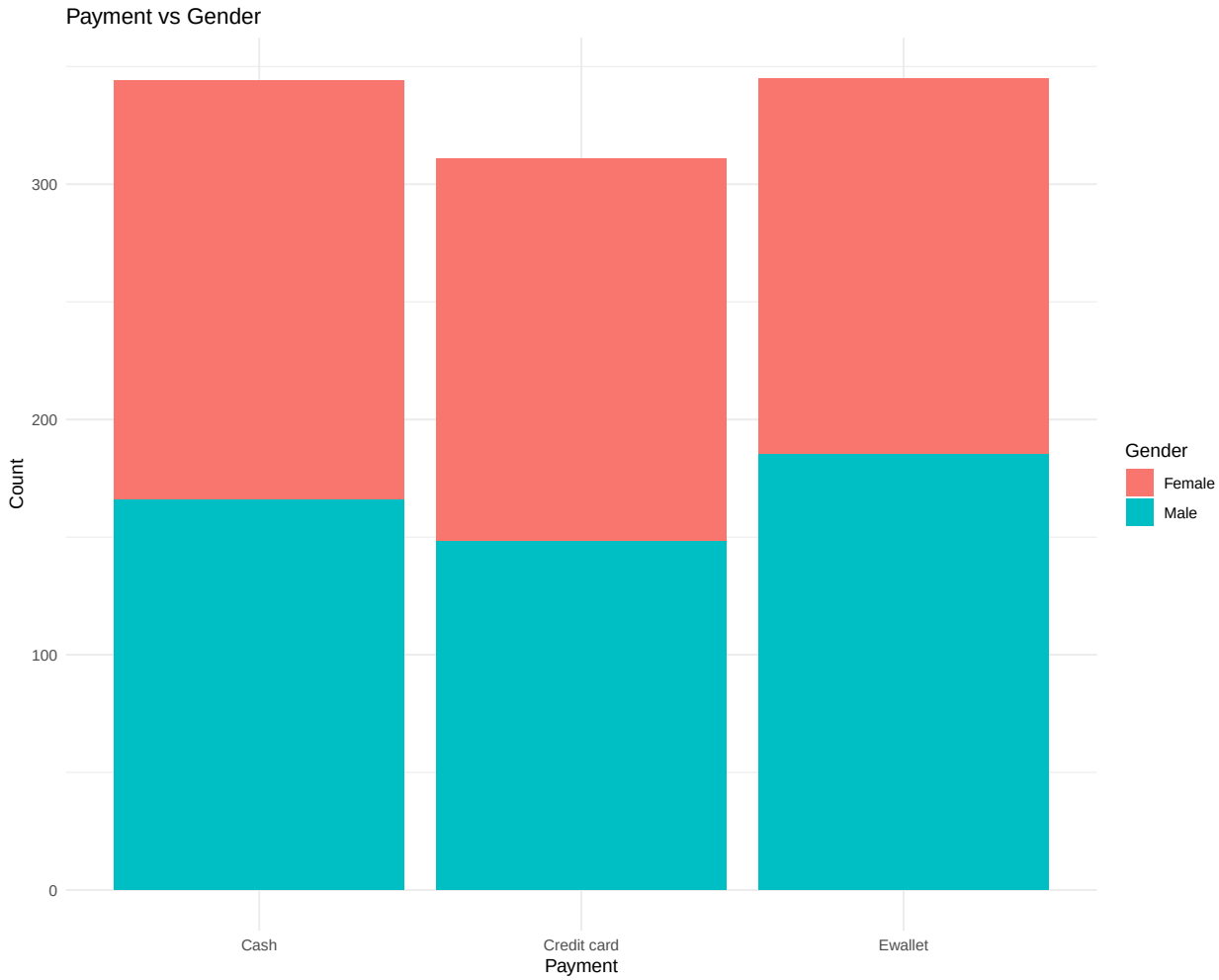
```
ggplot(data = data, aes(x = Customer.type, fill = Gender)) +
  geom_bar() +
  theme_minimal() +
  xlab("Customer Type") +
  ylab("Count") +
  labs(title = "Customer Type vs Gender")
```



Females are more tend to be members. But, Males are more tend to be Normal customers.

Gender vs Payment

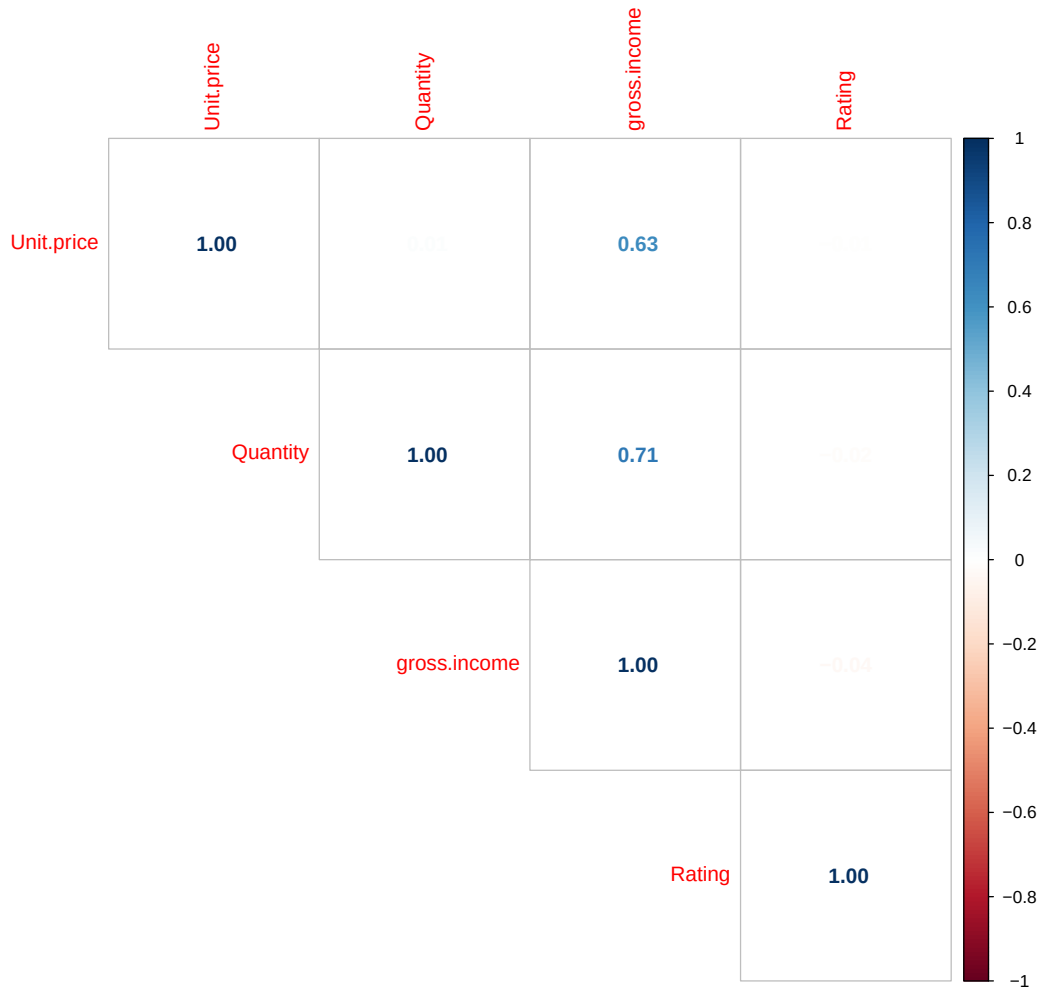
```
ggplot(data = data, aes(x = Payment, fill = Gender)) +
  geom_bar() +
  theme_minimal() +
  xlab("Payment") +
  ylab("Count") +
  labs(title = "Payment vs Gender")
```



E-wallet is more used by males. Credit card is more used by females.

Correlation Matrix

```
# Assuming 'data' is your data frame
cor_data <- data[c(7, 8, 16, 17)]
# Compute the correlation matrix, specifying 'use' to handle missing values
cor_data <- cor(cor_data, use = "pairwise.complete.obs")
# Round the correlation matrix
cor_data <- round(cor_data, 2)
# Create the correlation plot with a different font
corrplot(cor_data, method = "number",type="upper") # You can try different fonts
```



There is strong positive relation between Quantity and Gross income.

Logistic Regression

Product line by Gender

```
log1 <- glm(Product.line ~ Gender, data = data, family = "binomial")
print(log1)
```

```
##
## Call:  glm(formula = Product.line ~ Gender, family = "binomial", data = data)
##
## Coefficients:
## (Intercept)   GenderMale
##     1.60227     -0.03317
##
## Degrees of Freedom: 999 Total (i.e. Null);  998 Residual
## Null Deviance:       911.8
## Residual Deviance: 911.7     AIC: 915.7
```

```
#Product.line=1.60227-0.03317*(GenderMale)
```

Customer Type and Product Line

```
log2 <- glm(Product.line ~ Customer.type, data = data, family = "binomial")
print(log2)
```

```
##
## Call:  glm(formula = Product.line ~ Customer.type, family = "binomial",
##     data = data)
##
## Coefficients:
##         (Intercept)  Customer.typeNormal
##              1.6907              -0.2036
##
## Degrees of Freedom: 999 Total (i.e. Null);  998 Residual
## Null Deviance:       911.8
## Residual Deviance: 910.3     AIC: 914.3
```

```
#product.line=1.6907-0.2036*NormalCustomer
```

Branch and Gender

```
log3 <- glm(Branch~ Gender, data = data, family = "binomial")
print(log1)
```

```
##
## Call:  glm(formula = Product.line ~ Gender, family = "binomial", data = data)
##
## Coefficients:
## (Intercept)   GenderMale
##     1.60227     -0.03317
##
## Degrees of Freedom: 999 Total (i.e. Null);  998 Residual
## Null Deviance:       911.8
## Residual Deviance: 911.7     AIC: 915.7
```

```
# Branch=0.7475-0.1666*GenderMale
```

Payment and Gender

```
log4 <- glm(Gender~ Payment, data = data, family = "binomial")
summary(log4)
```

```
##
## Call:
## glm(formula = Gender ~ Payment, family = "binomial", data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.240  -1.148  -1.137   1.207   1.219
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.06980    0.10790  -0.647    0.518
## PaymentCredit card -0.02674    0.15663  -0.171    0.864
## PaymentEwallet      0.21498    0.15264   1.408    0.159
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.3  on 999  degrees of freedom
## Residual deviance: 1383.3  on 997  degrees of freedom
## AIC: 1389.3
##
## Number of Fisher Scoring iterations: 3
```

```
# GenderMale=-0.0698-0.0267*CreditCard+0.21498*Ewallet
```

Branch and Quantity

```
data$Quantity<-factor(data$Quantity)
data$Quantity<-relevel(data$Quantity,ref=1)
log5 <- glm(Quantity~ Branch, data = data, family = "binomial")
print(log5)
```

```
##
## Call:  glm(formula = Quantity ~ Branch, family = "binomial", data = data)
##
## Coefficients:
## (Intercept)      BranchB      BranchC
##      2.1335       0.1045      -0.2687
##
## Degrees of Freedom: 999 Total (i.e. Null);  997 Residual
## Null Deviance:       701.4
## Residual Deviance: 698.8     AIC: 704.8
```

```
# Quantity=2.13+0.1045*BranchB-0.2687*BrancC
```

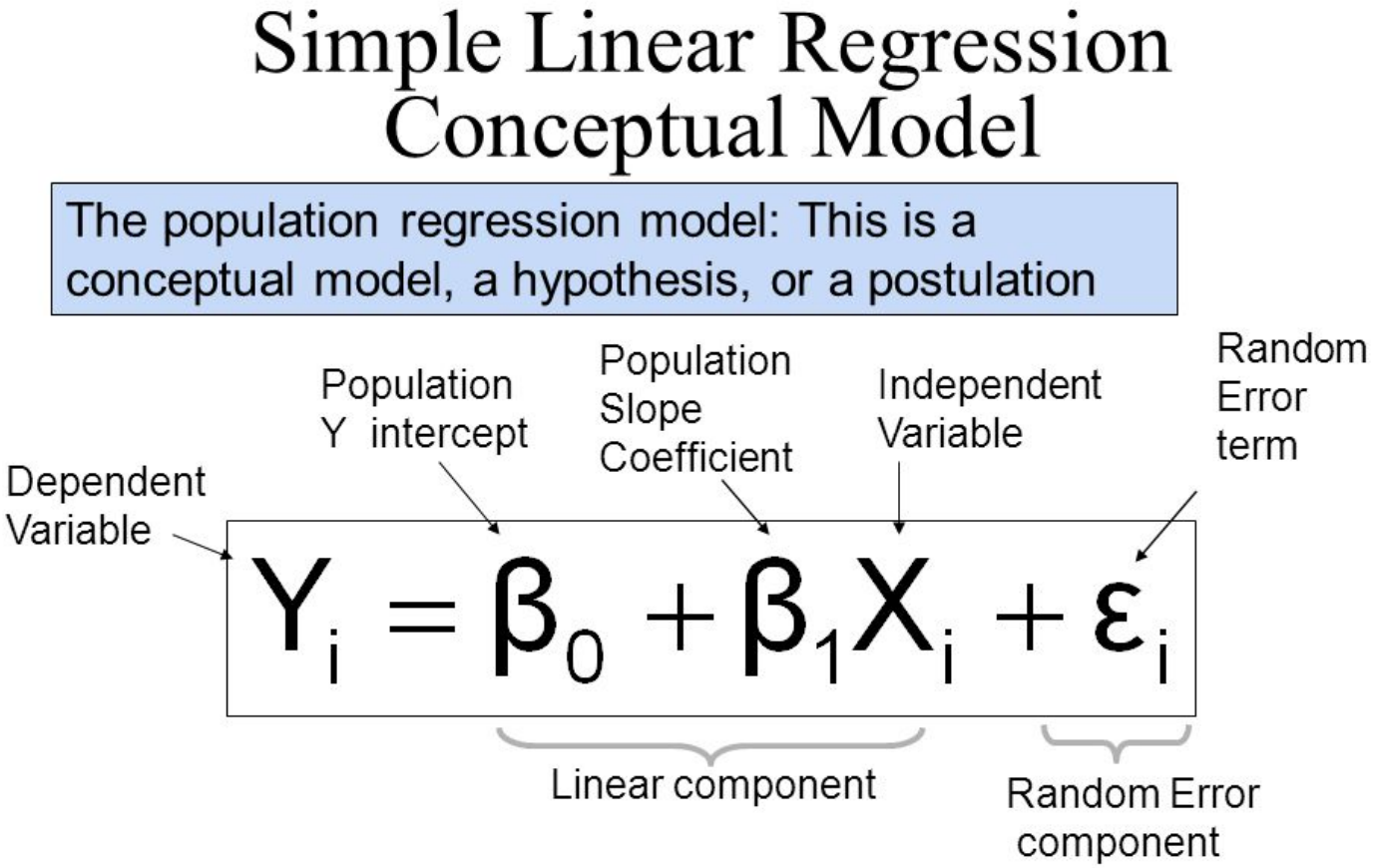
Customer Type and Branch

```
data$Customer.type<-relevel(data$Customer.type,ref="Member")
log6 <- glm(Customer.type~ Branch, data = data, family = "binomial")
print(log6)
```

```
##
## Call:  glm(formula = Customer.type ~ Branch, family = "binomial", data = data)
##
## Coefficients:
## (Intercept)      BranchB      BranchC
##     0.03530     -0.02325     -0.09629
##
## Degrees of Freedom: 999 Total (i.e. Null);  997 Residual
## Null Deviance:       1386
## Residual Deviance: 1386  AIC: 1392
```

```
#Member=0.035-0.0235BranchB-0.09629BranchC
```

Linear Regression



Quantity- Unit Price

```
data$Quantity<-as.numeric(data$Quantity)
l1<-lm(Quantity~Unit.price,data=data)
print(l1)
```

```
##
## Call:
## lm(formula = Quantity ~ Unit.price, data = data)
##
## Coefficients:
## (Intercept)   Unit.price
##    5.443795     0.001189
```

```
# Quantity=5.4437+0.001189*UnitPrice
```

Quantity-Rating

```
l2<-lm(Quantity~Rating,data=data)
print(l2)
```

```
##
## Call:
## lm(formula = Quantity ~ Rating, data = data)
##
## Coefficients:
## (Intercept)       Rating
##      5.6976      -0.0269
```

```
#Quantity=5.6976-0.0269*Rating
```

This is an interesting result for me. Since, when the rating is increasing 1 unit the corresponding quantity is decreasing as 0.0269

Gross Margin Percentage-Gross Income

```
l3<-lm(gross.margin.percentage~gross.income,data=data)
print(l3)
```

```
##
## Call:
## lm(formula = gross.margin.percentage ~ gross.income, data = data)
##
## Coefficients:
##  (Intercept)  gross.income
##    4.762e+00    -1.236e-16
```

```
# gross.margin.percentage=4.762e+00-1.236e-16 * gross.income
```

This is very low, so we can say that there is no linear relation between two variables.

Total-Unit Price

```
l4<-lm(Total~Unit.price,data=data)
print(l4)
```

```
##
## Call:
## lm(formula = Total ~ Unit.price, data = data)
##
## Coefficients:
## (Intercept)   Unit.price
##      -4.582        5.884
```

```
#Total=-4.582+5.884*UnitPrice
```

This isn't reliable since intercept is minus, but total can't be negative.

Total-Rating

```
l5<-lm(Total~Rating,data=data)
print(l5)
```

```
##
## Call:
## lm(formula = Total ~ Rating, data = data)
##
## Coefficients:
## (Intercept)       Rating
##     359.322       -5.214
```

```
#Total=359.322-5.214*Rating
```

When rating is increasing 1 unit, total is decreasing as 5.214 unit.