EDA ON SUPERMARKET SALES DATASET IN R

Created by Reneeka, Kshitij, Rahil

Table of Content

This is the material point that will be delivered in the presentation.

Problem Statement

About Data

Implementation

Analysis

Bibliography

Conclusion

Problem Statement

Our aim is to undertake a comprehensive analysis of sales data for a supermarket chain's three branches. This mystical quest will involve traversing the ethereal realm of sales data, where we'll delve into the tapestry woven by these branches from January to March 2019. Our goal is to observe the enigmatic dance of the bivariate entities, ultimately gaining valuable insights.

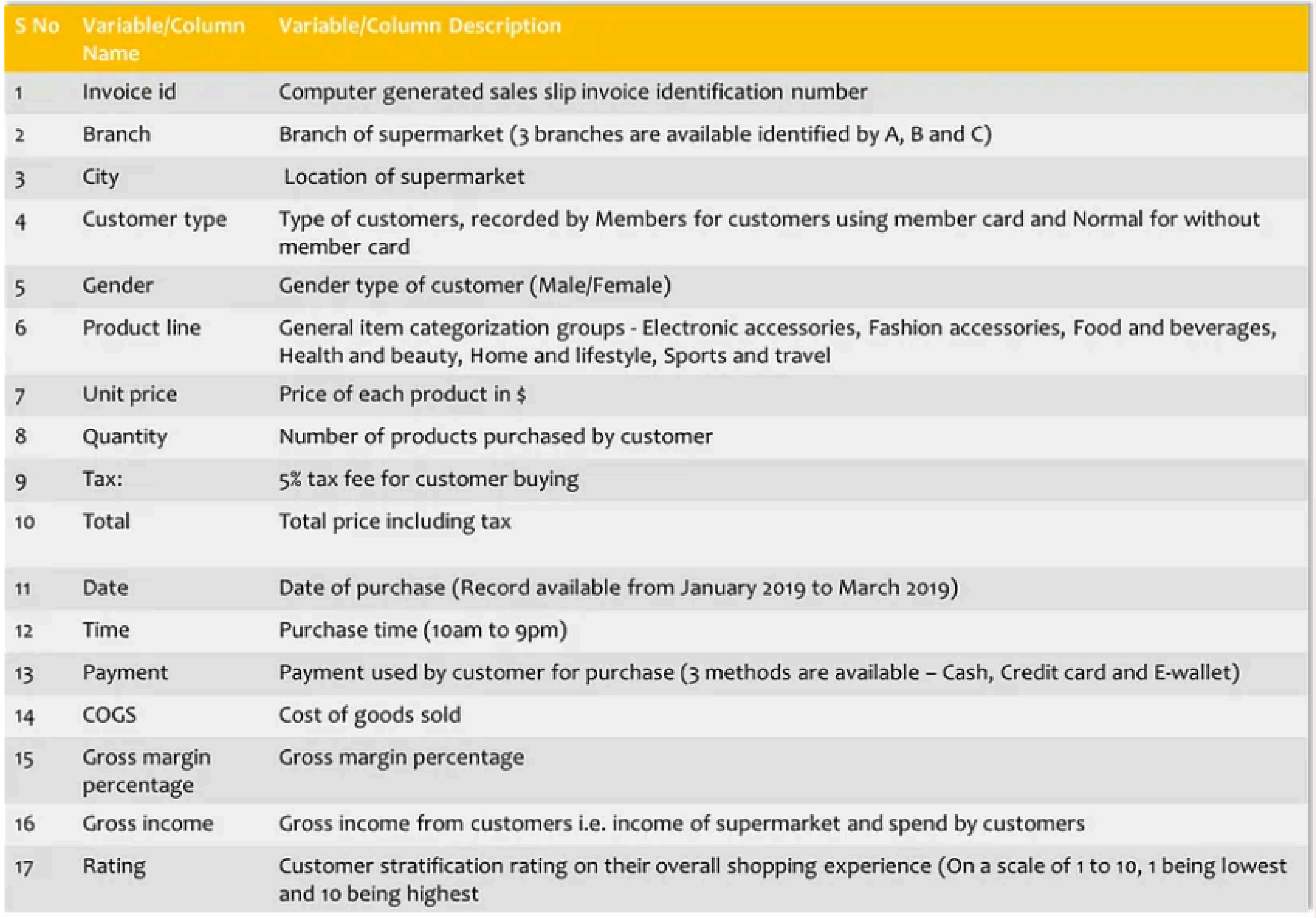
Dataset Selection

The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data. Predictive data analytics methods are easy to apply with this dataset.

kaggle

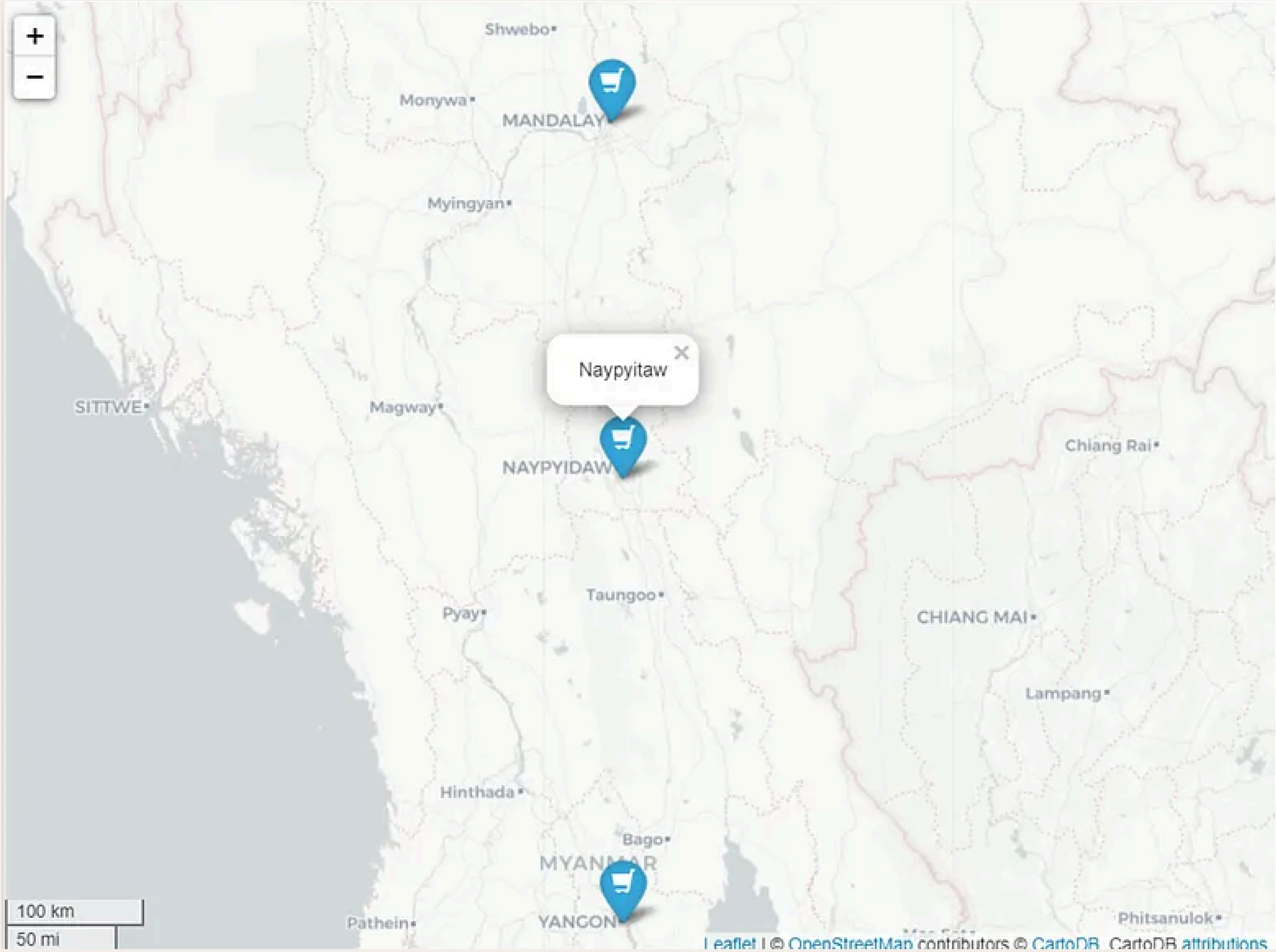
Origin of the Dataset.

Familiarizing variables

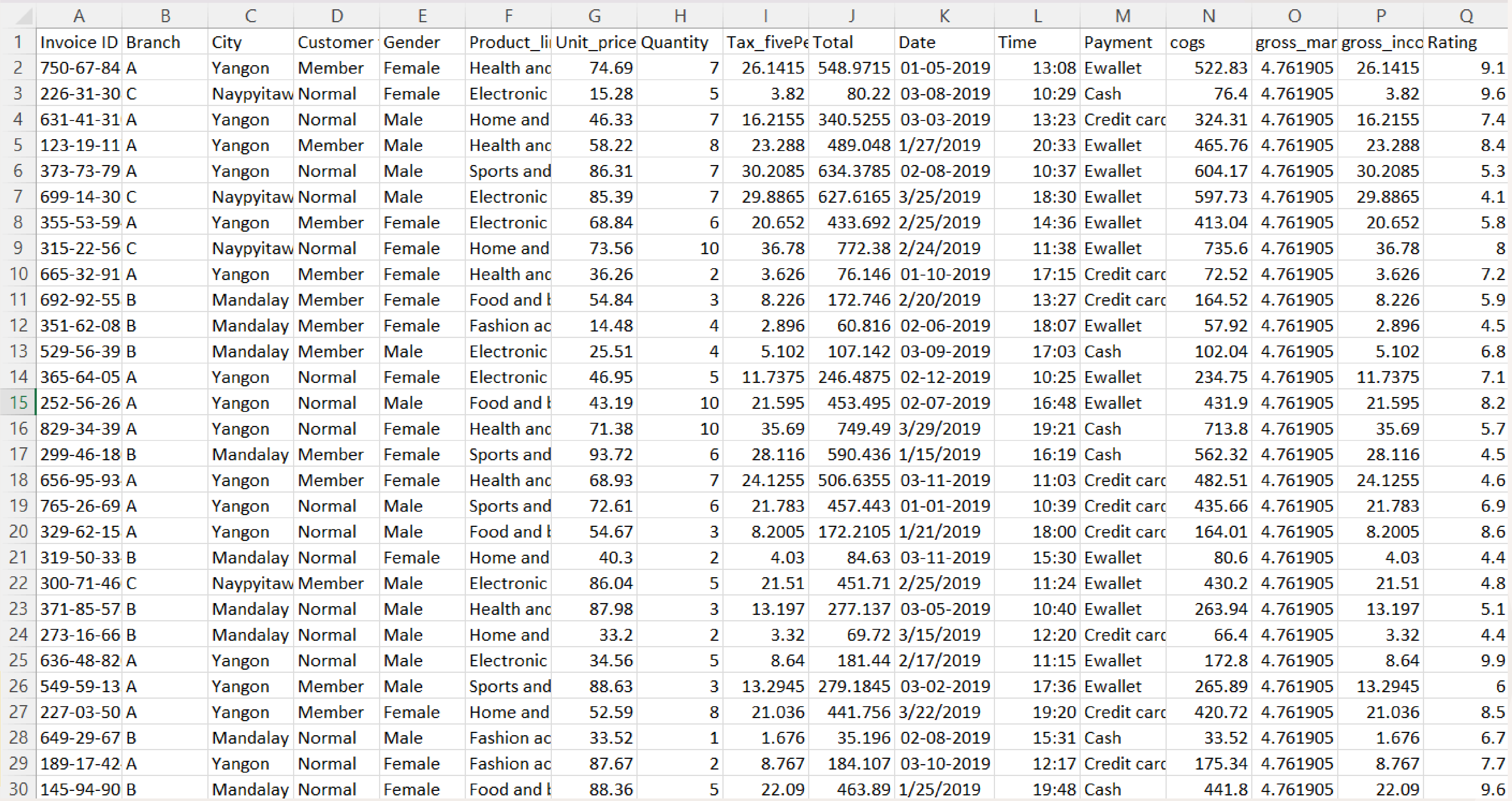


S No	Variable/Column Name	Variable/Column Description
1	Invoice id	Computer generated sales slip invoice identification number
2	Branch	Branch of supermarket (3 branches are available identified by A, B and C)
3	City	Location of supermarket
4	Customer type	Type of customers, recorded by Members for customers using member card and Normal for without member card
5	Gender	Gender type of customer (Male/Female)
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	Time	Purchase time (10am to 9pm)
13	Payment	Payment used by customer for purchase (3 methods are available – Cash, Credit card and E-wallet)
14	COGS	Cost of goods sold
15	Gross margin percentage	Gross margin percentage
16	Gross income	Gross income from customers i.e. income of supermarket and spend by customers
17	Rating	Customer stratification rating on their overall shopping experience (On a scale of 1 to 10, 1 being lowest and 10 being highest

Familiarizing variables



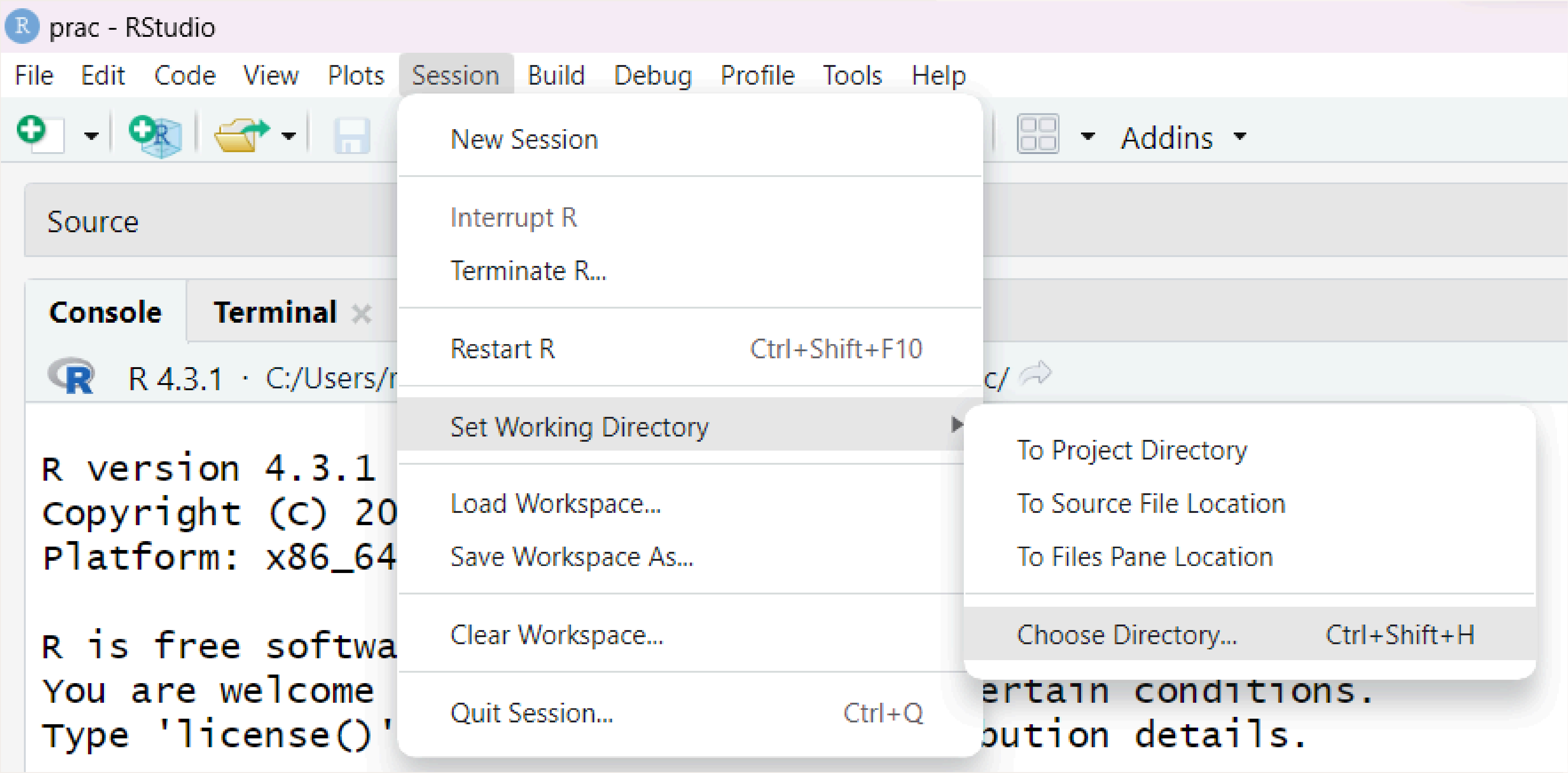
Preview of the Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Invoice ID	Branch	City	Customer	Gender	Product_li	Unit_price	Quantity	Tax_fivePe	Total	Date	Time	Payment	cogs	gross_mar	gross_inco	Rating
2	750-67-84	A	Yangon	Member	Female	Health anc	74.69	7	26.1415	548.9715	01-05-2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
3	226-31-30	C	Naypyitaw	Normal	Female	Electronic	15.28	5	3.82	80.22	03-08-2019	10:29	Cash	76.4	4.761905	3.82	9.6
4	631-41-31	A	Yangon	Normal	Male	Home and	46.33	7	16.2155	340.5255	03-03-2019	13:23	Credit carc	324.31	4.761905	16.2155	7.4
5	123-19-11	A	Yangon	Member	Male	Health anc	58.22	8	23.288	489.048	1/27/2019	20:33	Ewallet	465.76	4.761905	23.288	8.4
6	373-73-79	A	Yangon	Normal	Male	Sports and	86.31	7	30.2085	634.3785	02-08-2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
7	699-14-30	C	Naypyitaw	Normal	Male	Electronic	85.39	7	29.8865	627.6165	3/25/2019	18:30	Ewallet	597.73	4.761905	29.8865	4.1
8	355-53-59	A	Yangon	Member	Female	Electronic	68.84	6	20.652	433.692	2/25/2019	14:36	Ewallet	413.04	4.761905	20.652	5.8
9	315-22-56	C	Naypyitaw	Normal	Female	Home and	73.56	10	36.78	772.38	2/24/2019	11:38	Ewallet	735.6	4.761905	36.78	8
10	665-32-91	A	Yangon	Member	Female	Health anc	36.26	2	3.626	76.146	01-10-2019	17:15	Credit carc	72.52	4.761905	3.626	7.2
11	692-92-55	B	Mandalay	Member	Female	Food and b	54.84	3	8.226	172.746	2/20/2019	13:27	Credit carc	164.52	4.761905	8.226	5.9
12	351-62-08	B	Mandalay	Member	Female	Fashion ac	14.48	4	2.896	60.816	02-06-2019	18:07	Ewallet	57.92	4.761905	2.896	4.5
13	529-56-39	B	Mandalay	Member	Male	Electronic	25.51	4	5.102	107.142	03-09-2019	17:03	Cash	102.04	4.761905	5.102	6.8
14	365-64-05	A	Yangon	Normal	Female	Electronic	46.95	5	11.7375	246.4875	02-12-2019	10:25	Ewallet	234.75	4.761905	11.7375	7.1
15	252-56-26	A	Yangon	Normal	Male	Food and b	43.19	10	21.595	453.495	02-07-2019	16:48	Ewallet	431.9	4.761905	21.595	8.2
16	829-34-39	A	Yangon	Normal	Female	Health anc	71.38	10	35.69	749.49	3/29/2019	19:21	Cash	713.8	4.761905	35.69	5.7
17	299-46-18	B	Mandalay	Member	Female	Sports and	93.72	6	28.116	590.436	1/15/2019	16:19	Cash	562.32	4.761905	28.116	4.5
18	656-95-93	A	Yangon	Member	Female	Health anc	68.93	7	24.1255	506.6355	03-11-2019	11:03	Credit carc	482.51	4.761905	24.1255	4.6
19	765-26-69	A	Yangon	Normal	Male	Sports and	72.61	6	21.783	457.443	01-01-2019	10:39	Credit carc	435.66	4.761905	21.783	6.9
20	329-62-15	A	Yangon	Normal	Male	Food and b	54.67	3	8.2005	172.2105	1/21/2019	18:00	Credit carc	164.01	4.761905	8.2005	8.6
21	319-50-33	B	Mandalay	Normal	Female	Home and	40.3	2	4.03	84.63	03-11-2019	15:30	Ewallet	80.6	4.761905	4.03	4.4
22	300-71-46	C	Naypyitaw	Member	Male	Electronic	86.04	5	21.51	451.71	2/25/2019	11:24	Ewallet	430.2	4.761905	21.51	4.8
23	371-85-57	B	Mandalay	Normal	Male	Health anc	87.98	3	13.197	277.137	03-05-2019	10:40	Ewallet	263.94	4.761905	13.197	5.1
24	273-16-66	B	Mandalay	Normal	Male	Home and	33.2	2	3.32	69.72	3/15/2019	12:20	Credit carc	66.4	4.761905	3.32	4.4
25	636-48-82	A	Yangon	Normal	Male	Electronic	34.56	5	8.64	181.44	2/17/2019	11:15	Ewallet	172.8	4.761905	8.64	9.9
26	549-59-13	A	Yangon	Member	Male	Sports and	88.63	3	13.2945	279.1845	03-02-2019	17:36	Ewallet	265.89	4.761905	13.2945	6
27	227-03-50	A	Yangon	Member	Female	Home and	52.59	8	21.036	441.756	3/22/2019	19:20	Credit carc	420.72	4.761905	21.036	8.5
28	649-29-67	B	Mandalay	Normal	Male	Fashion ac	33.52	1	1.676	35.196	02-08-2019	15:31	Cash	33.52	4.761905	1.676	6.7
29	189-17-42	A	Yangon	Normal	Female	Fashion ac	87.67	2	8.767	184.107	03-10-2019	12:17	Credit carc	175.34	4.761905	8.767	7.7
30	145-94-90	B	Mandalay	Normal	Female	Food and b	88.36	5	22.09	463.89	1/25/2019	19:48	Cash	441.8	4.761905	22.09	9.6

Implementation

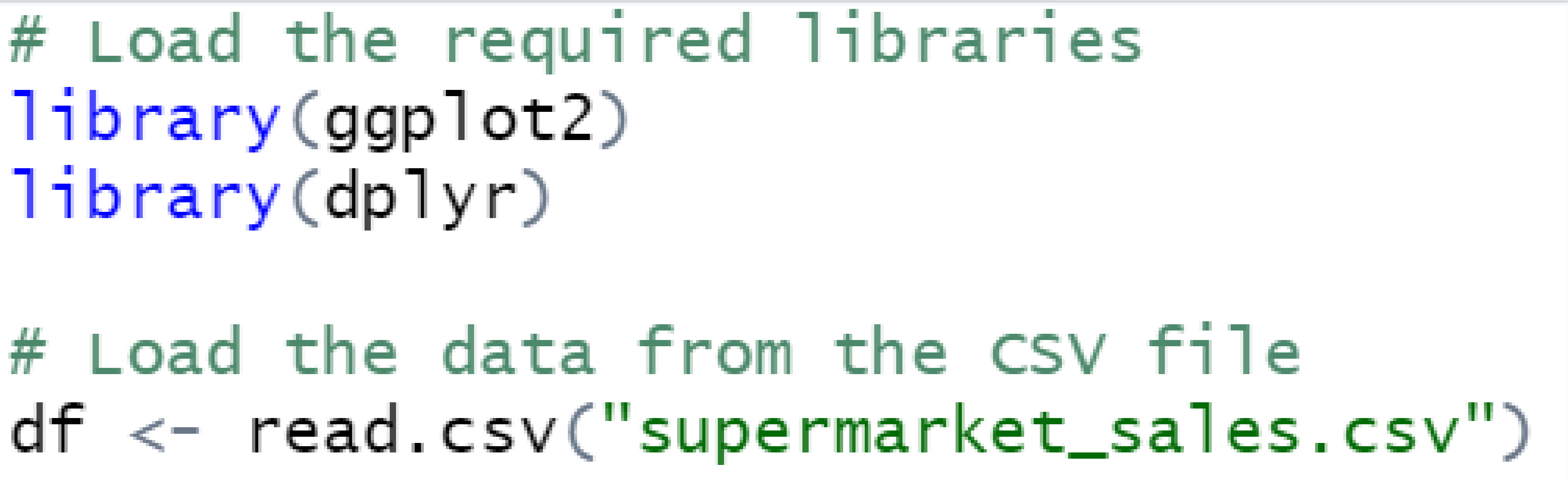
Establishing the working directory of the project.



```
R version 4.3.1
Copyright (C) 20
Platform: x86_64

R is free softwa
You are welcome                                  ertain conditions.
Type 'license()'                                 bution details.
```

Implementation

Loading the relevant libraries and datasets.

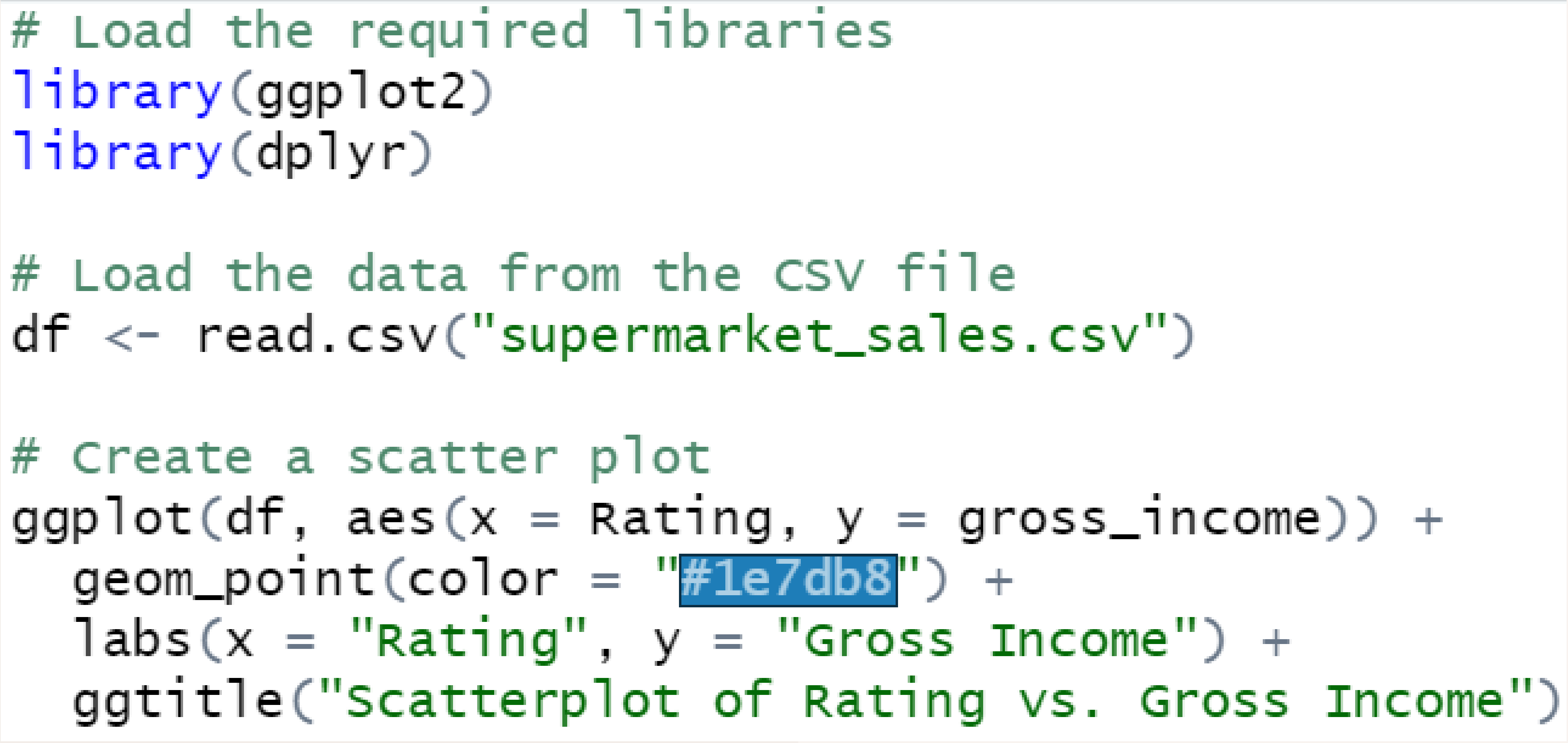


```
# Load the required libraries
library(ggplot2)
library(dplyr)

# Load the data from the CSV file
df <- read.csv("supermarket_sales.csv")
```

Bi-Variate Analysis

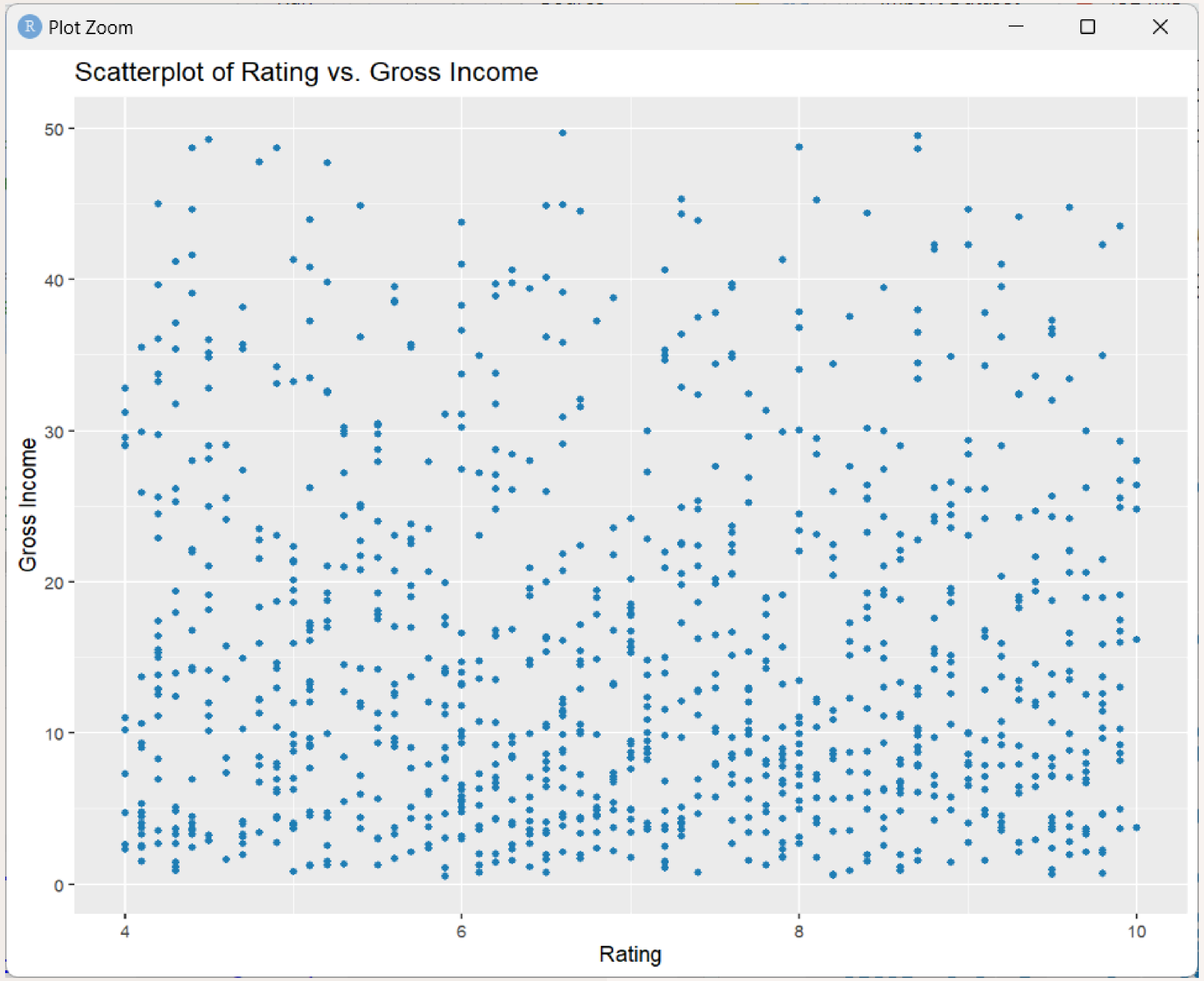
Bivariate analysis entails a statistical approach that examines the correlation or connection between two variables, with the aim of comprehending their interdependence or mutual influence.

Does gross income affect the ratings that the customers provide?

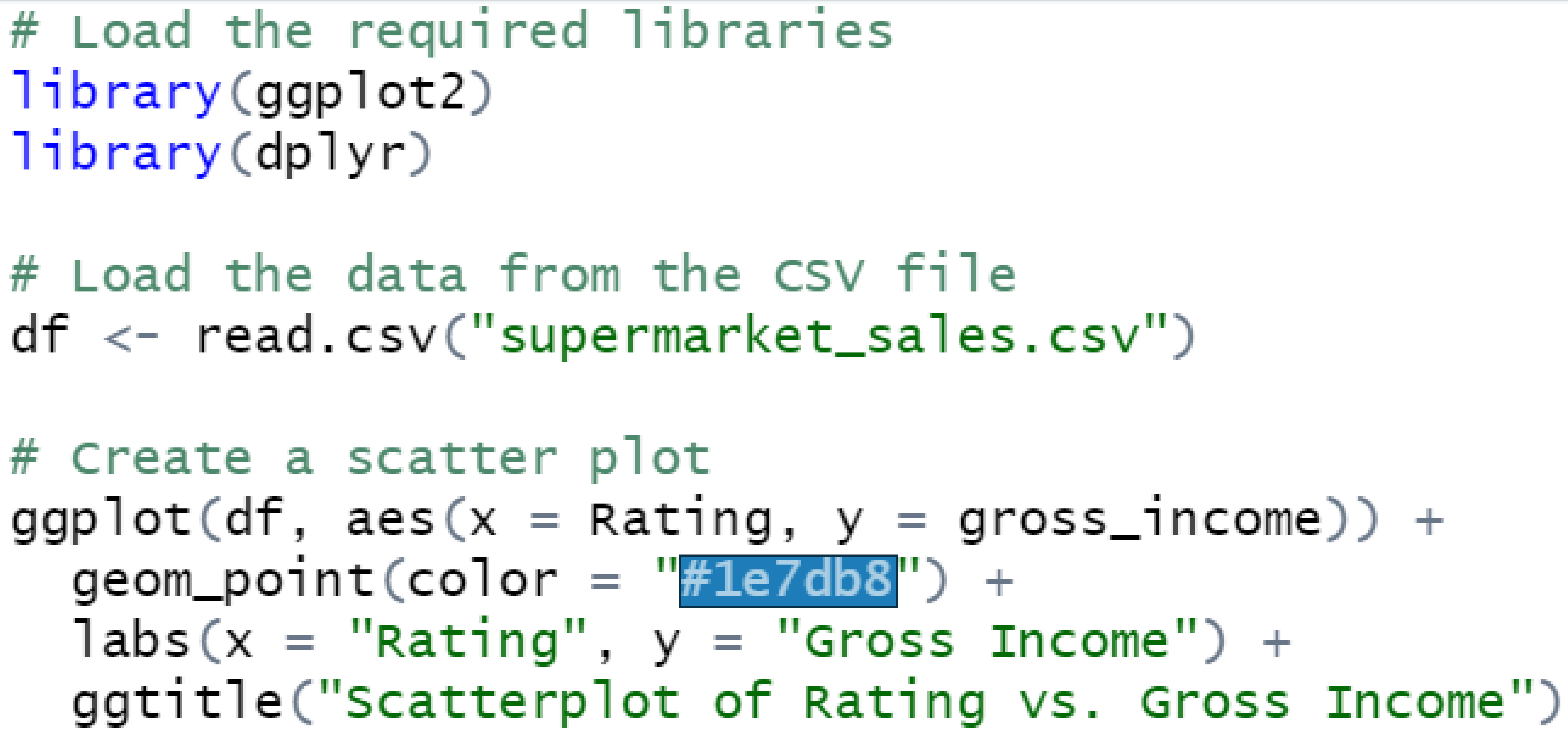


```r
# Load the required libraries
library(ggplot2)
library(dplyr)

# Load the data from the CSV file
df <- read.csv("supermarket_sales.csv")

# Create a scatter plot
ggplot(df, aes(x = Rating, y = gross_income)) +
  geom_point(color = "#1e7db8") +
  labs(x = "Rating", y = "Gross Income") +
  ggtitle("Scatterplot of Rating vs. Gross Income")
```

Does gross income affect the ratings that the customers provide?

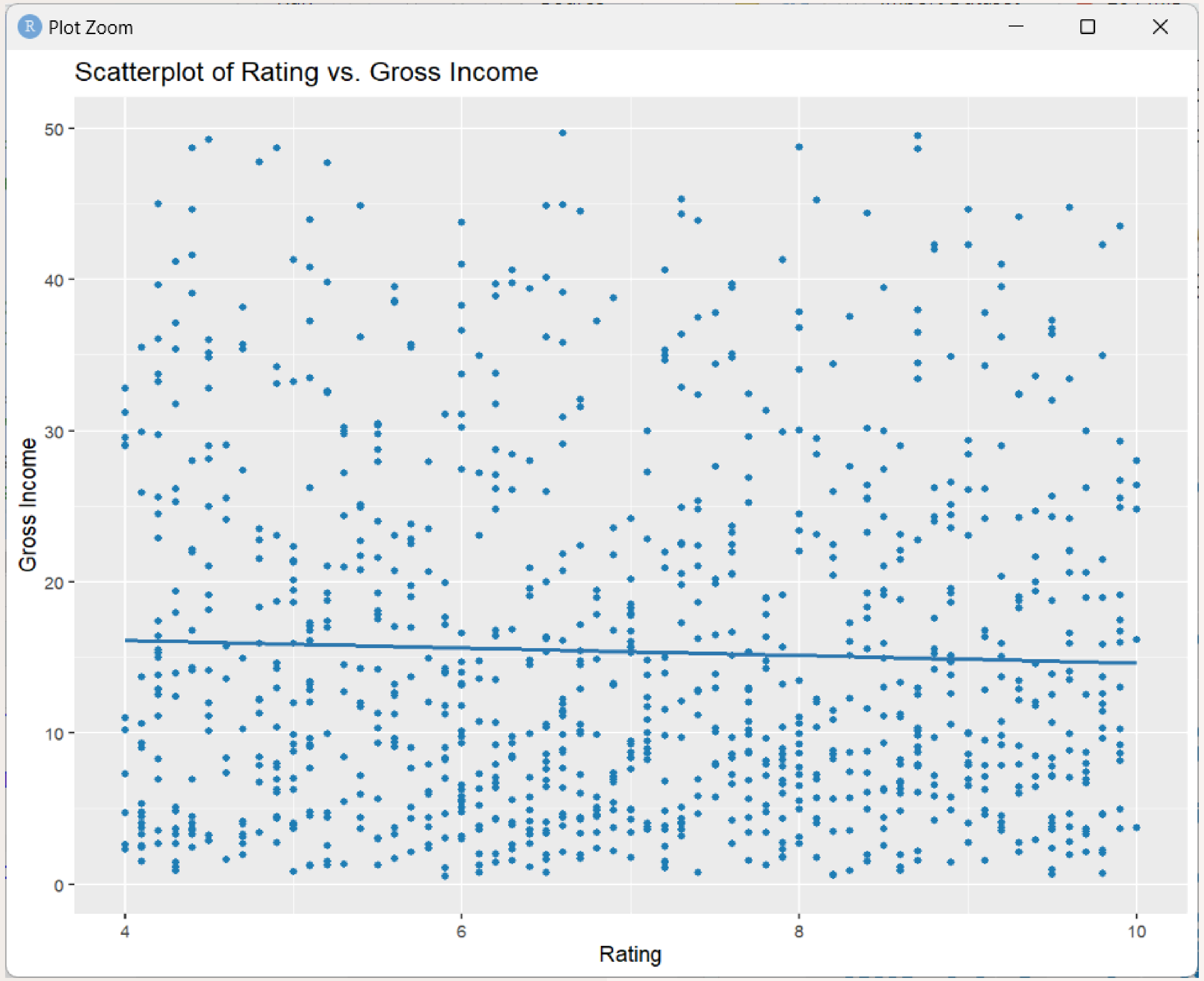


Does gross income affect the ratings that the customers provide?

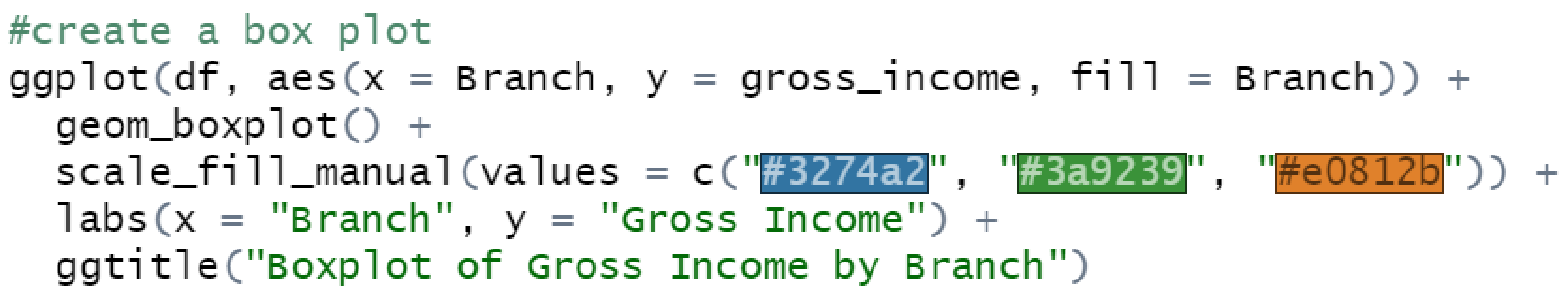


```
# Load the required libraries
library(ggplot2)
library(dplyr)

# Load the data from the CSV file
df <- read.csv("supermarket_sales.csv")

# Create a scatter plot
ggplot(df, aes(x = Rating, y = gross_income)) +
  geom_point(color = "#1e7db8") +
  labs(x = "Rating", y = "Gross Income") +
  ggtitle("Scatterplot of Rating vs. Gross Income")
```

Does gross income affect the ratings that the customers provide?

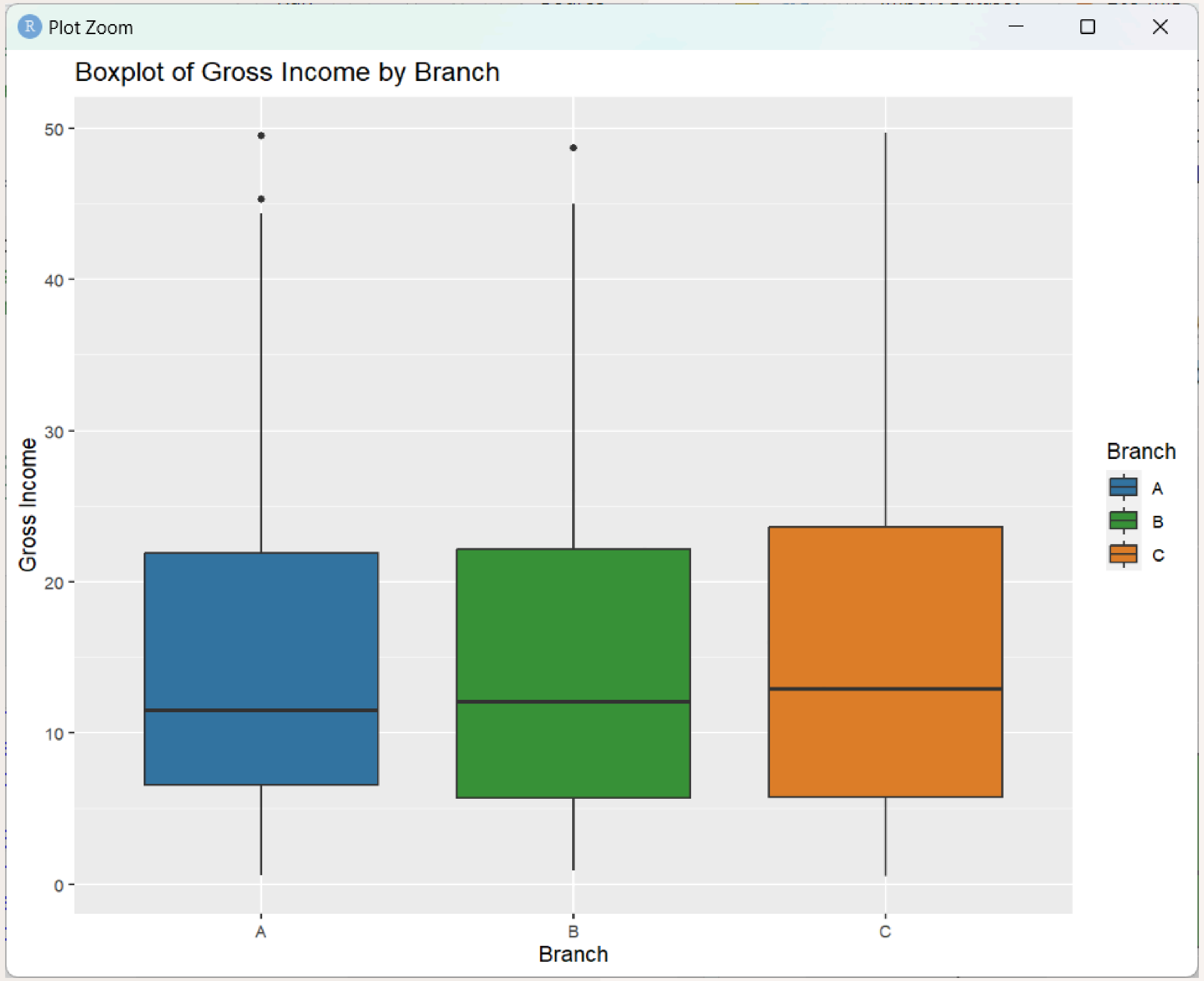


Which branch is the most profitable?

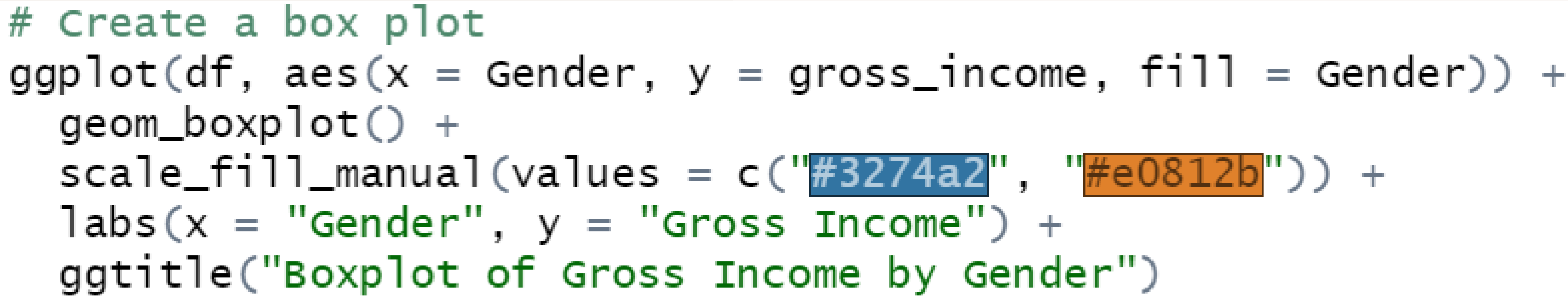


```r
#create a box plot
ggplot(df, aes(x = Branch, y = gross_income, fill = Branch)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#3274a2", "#3a9239", "#e0812b")) +
  labs(x = "Branch", y = "Gross Income") +
  ggtitle("Boxplot of Gross Income by Branch")
```

Which branch is the most profitable?

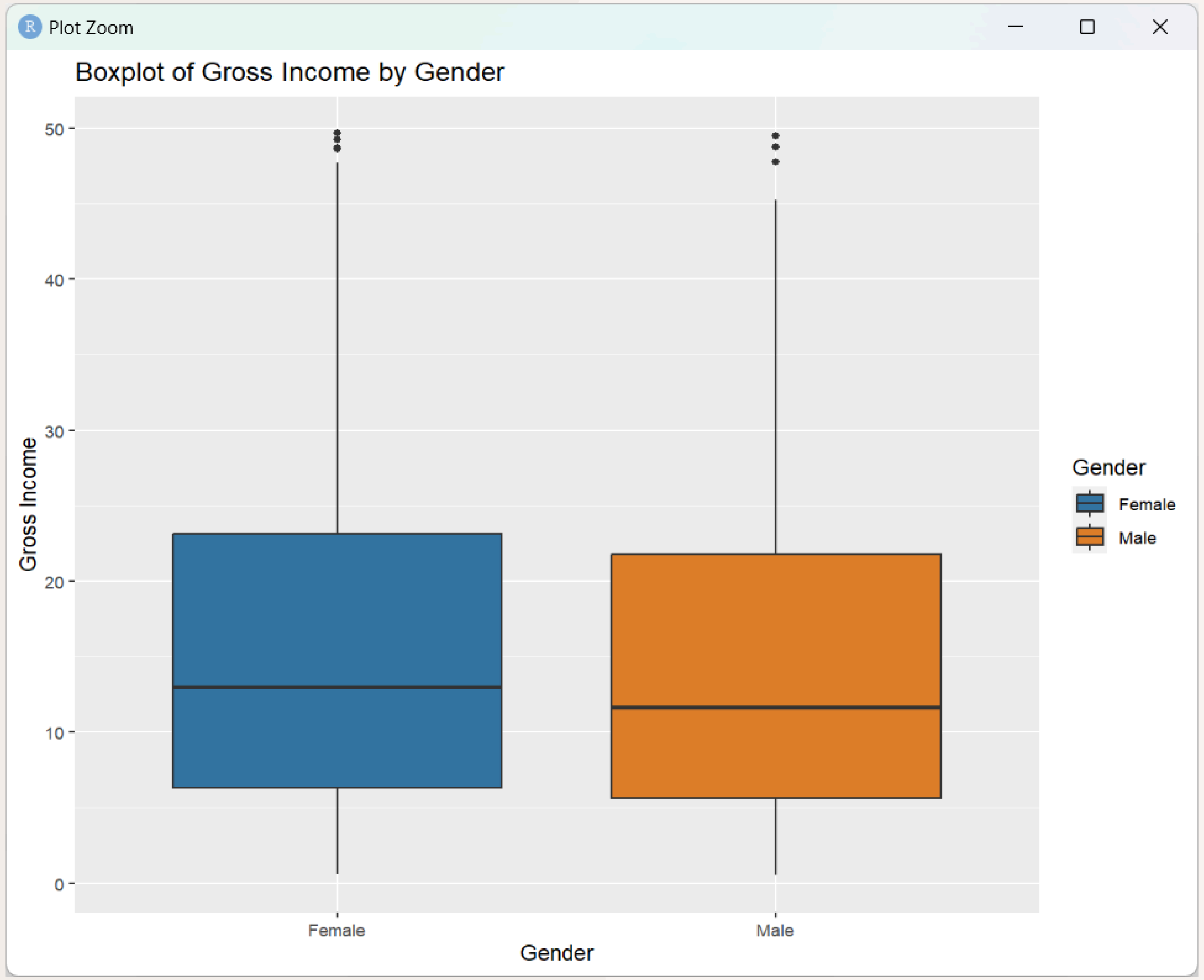


Is there any relationship between Gender and Gross income?

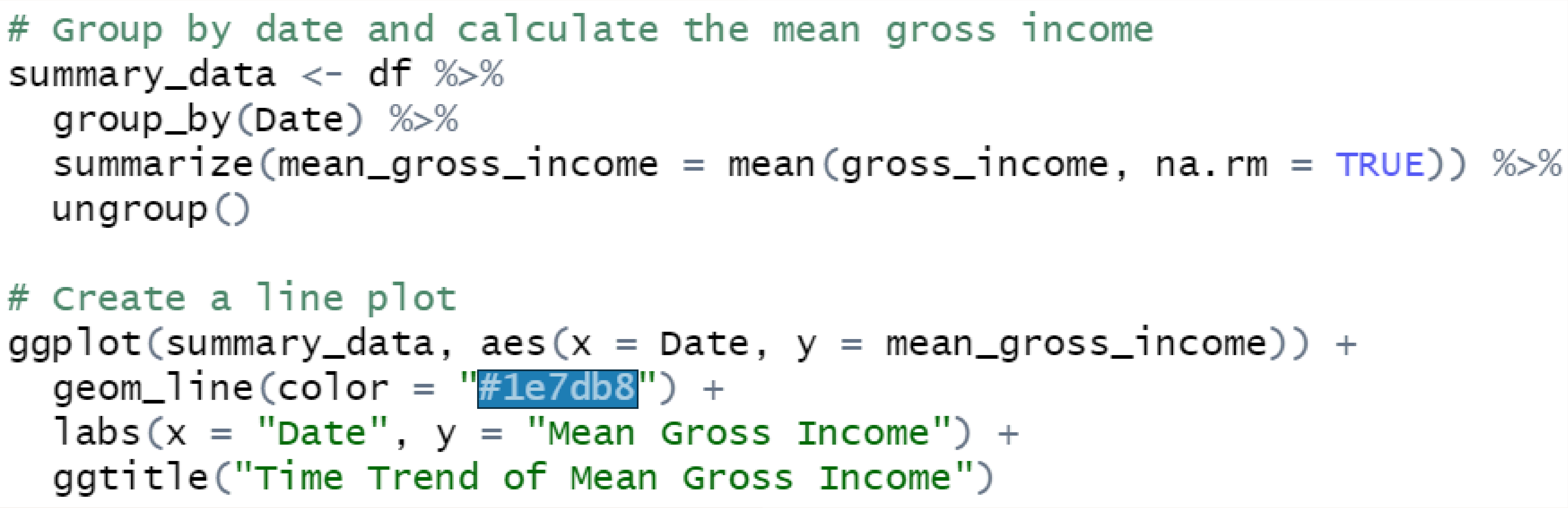


```r
# Create a box plot
ggplot(df, aes(x = Gender, y = gross_income, fill = Gender)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#3274a2", "#e0812b")) +
  labs(x = "Gender", y = "Gross Income") +
  ggtitle("Boxplot of Gross Income by Gender")
```

Is there any relationship between Gender and Gross income?

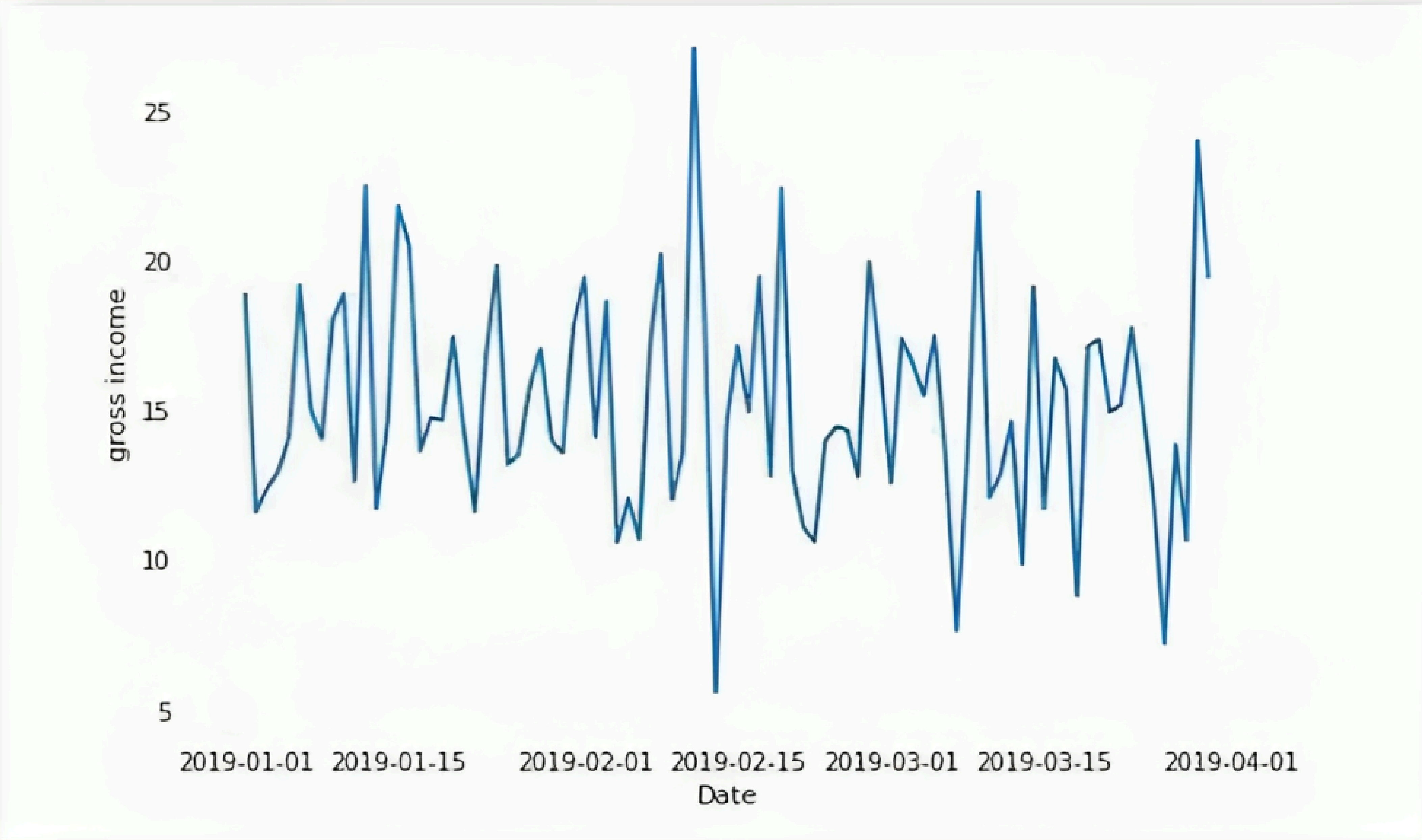


Is there any time trend in gross income?

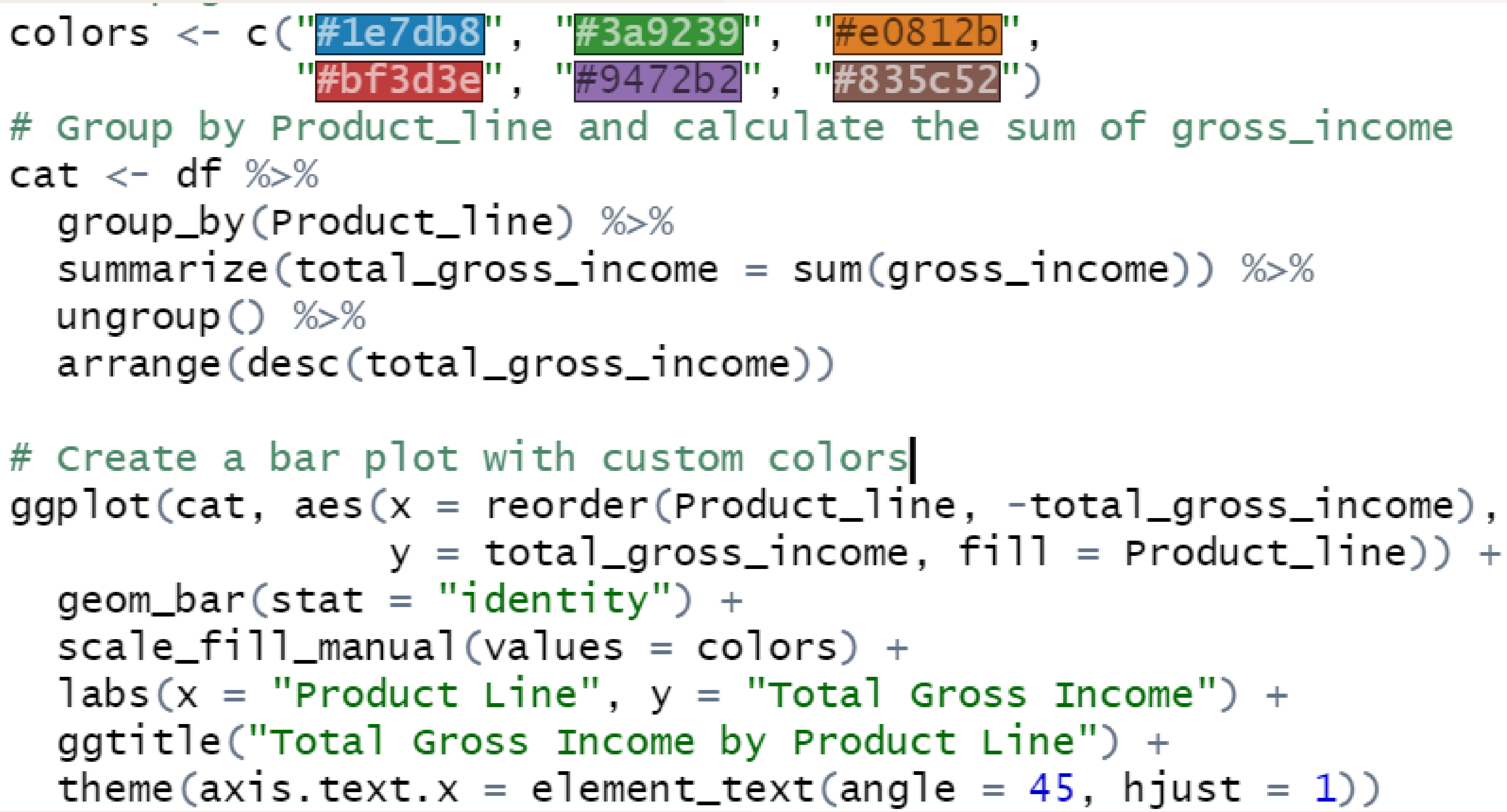


```
# Group by date and calculate the mean gross income
summary_data <- df %>%
  group_by(Date) %>%
  summarize(mean_gross_income = mean(gross_income, na.rm = TRUE)) %>%
  ungroup()

# Create a line plot
ggplot(summary_data, aes(x = Date, y = mean_gross_income)) +
  geom_line(color = "#1e7db8") +
  labs(x = "Date", y = "Mean Gross Income") +
  ggtitle("Time Trend of Mean Gross Income")
```

Is there any time trend in gross income?

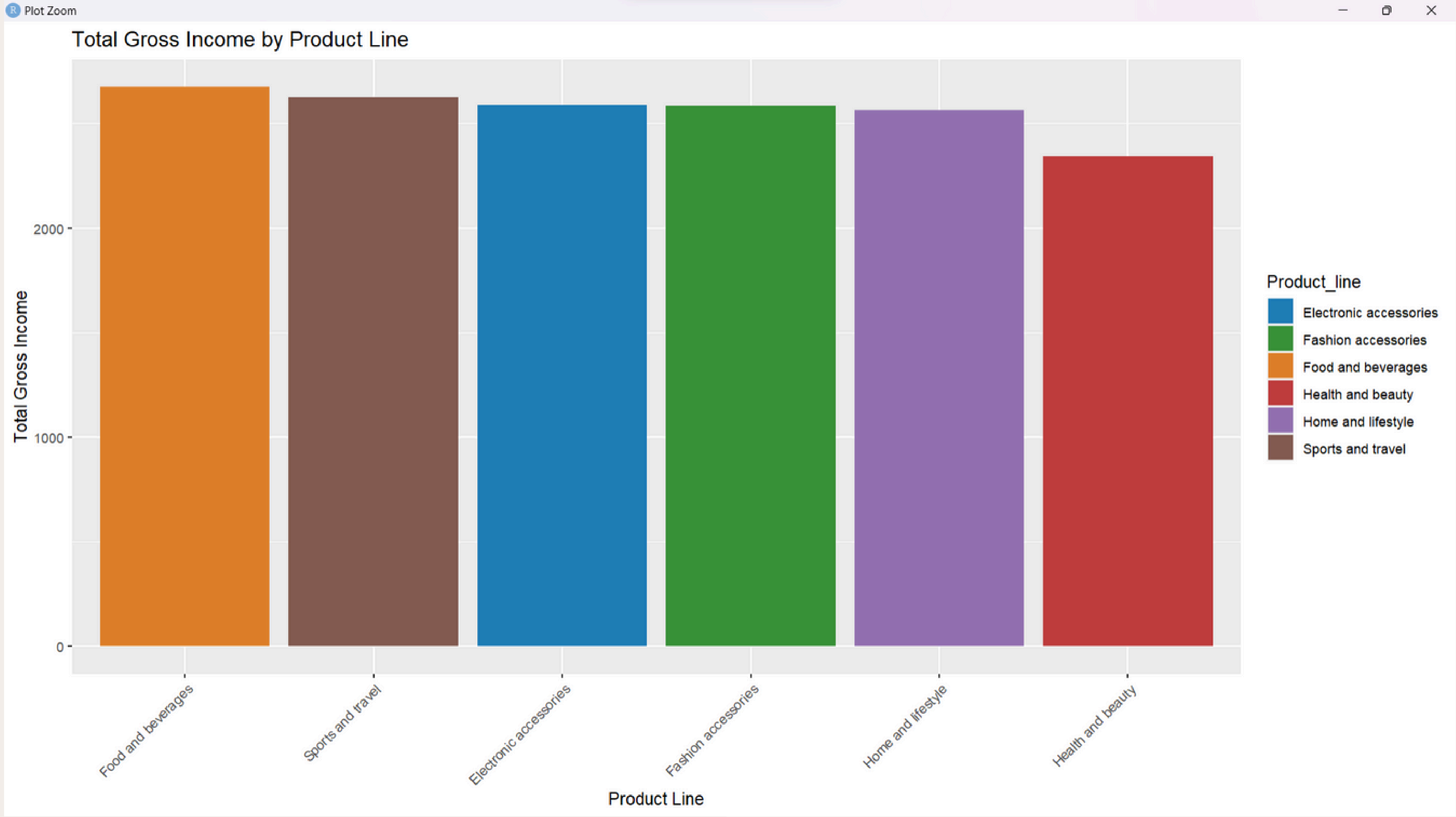


Which product line generates most income?



```
colors <- c("#1e7db8", "#3a9239", "#e0812b",
            "#bf3d3e", "#9472b2", "#835c52")
# Group by Product_line and calculate the sum of gross_income
cat <- df %>%
  group_by(Product_line) %>%
  summarize(total_gross_income = sum(gross_income)) %>%
  ungroup() %>%
  arrange(desc(total_gross_income))

# Create a bar plot with custom colors
ggplot(cat, aes(x = reorder(Product_line, -total_gross_income),
                y = total_gross_income, fill = Product_line)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = colors) +
  labs(x = "Product Line", y = "Total Gross Income") +
  ggtitle("Total Gross Income by Product Line") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Which product line generates most income?



Bibliography

01 https://towardsdatascience.com/exploratory-data-analysis-using-spermarket-sales-data-in-python-e99d329a07fc

02 https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

03 Kaggle Image Source

Thank You

We appreciate your patience.