LOVELY PROFESSIONAL UNIVERSITY

Course: DATA EXPLORATION AND PREPARATION

Course Code: CAP482

CA 4

Dated: - 29/Apr/2024

Submitted by

Name: Hammad Raza Khan
Roll No: 49
Reg: 12222969
Section: DE419, Group: 1

Submitted to

Ms. Ranjit Kaur Walia
UID: 28632
Assistant Professor
SCA, LPU

Lovely Faculty of Technology & Sciences

School of Computer Applications

Lovely Professional University

Punjab

Customer_Data

Hammad Raza Khan

2024-04-29

roll_NO: '49'

reg_NO: '12222969'

https://rpubs.com/hammadrazakhann/1179536

#Adding the Dataset

```
library(readr)
customer_data <- read_csv("C:/Users/hamma/Downloads/customer_data.csv")

## New names:
## Rows: 202 Columns: 16
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (7): gender, education, region, loyalty_status, purchase_frequency,
prod... dbl
## (6): id, age, income, purchase_amount, promotion_usage, satisfaction_score
lgl
## (3): ...13, ...14, ...15
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...13`
## • `` -> `...14`
## • `` -> `...15`
## • `` -> `...16`

View(customer_data)
```

Topic / Dataset = "CUSTOMER_DATA"

1. PRE-PROCESSING OF DATA

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
```

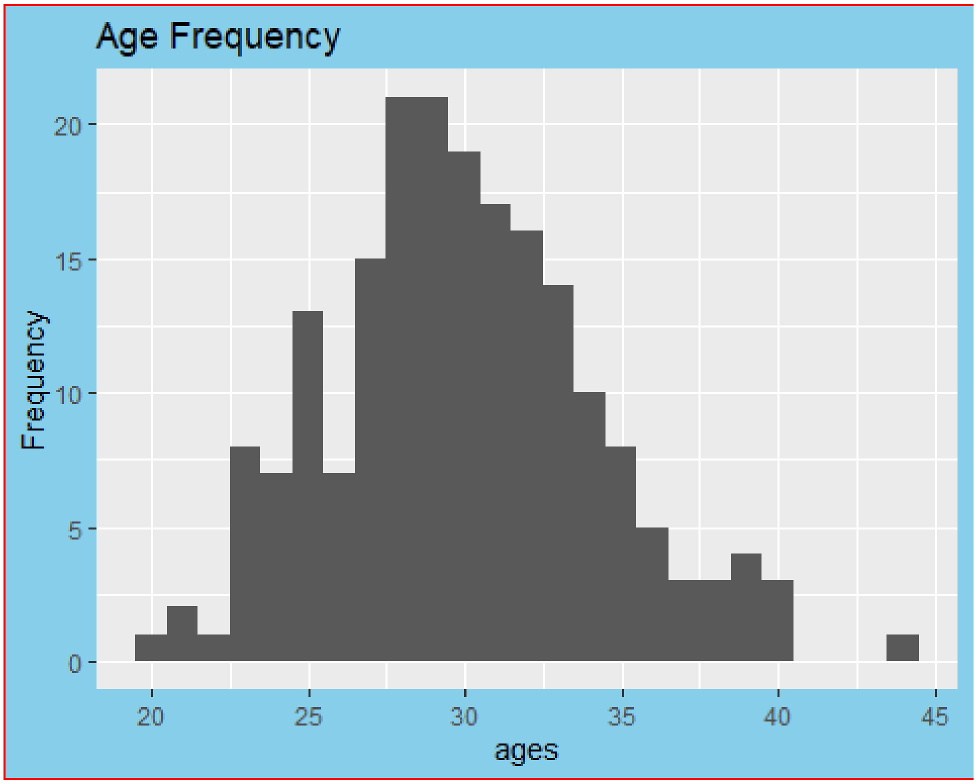
```
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
library(tidyr)
library(ggplot2)
library(knitr)

# DATA PRE-PROCESSING
# deleting null values rows

customer_data = customer_data[-200, ]


view(customer_data)
```

ANALYSIS OF DATA

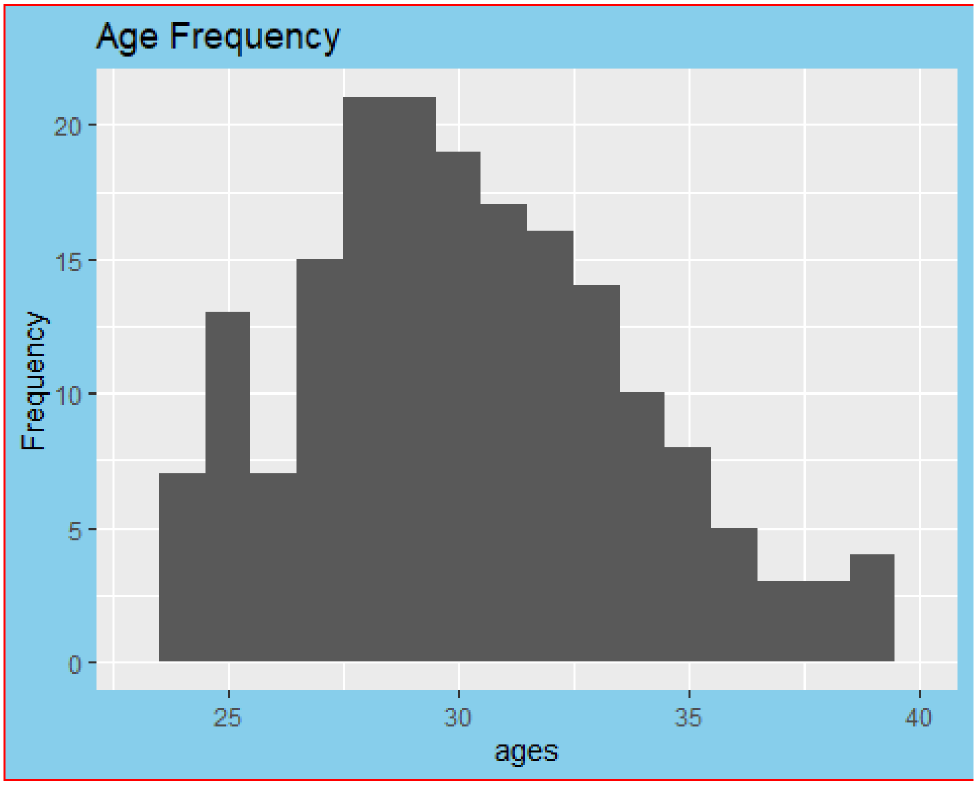
ANALYSIS - 1

```r
#ANALYSIS 1

# Age Distribution: What is the distribution of ages among customers?

ggplot(data = customer_data, aes(x = age)) +
  geom_histogram(binwidth = 1) +
  theme(plot.background = element_rect(fill = "skyblue", color = "red")) +
  labs(title = "Age Frequency",
       x = "ages",
       y = "Frequency")
```



##Since there are outliners, they should be taken care of -

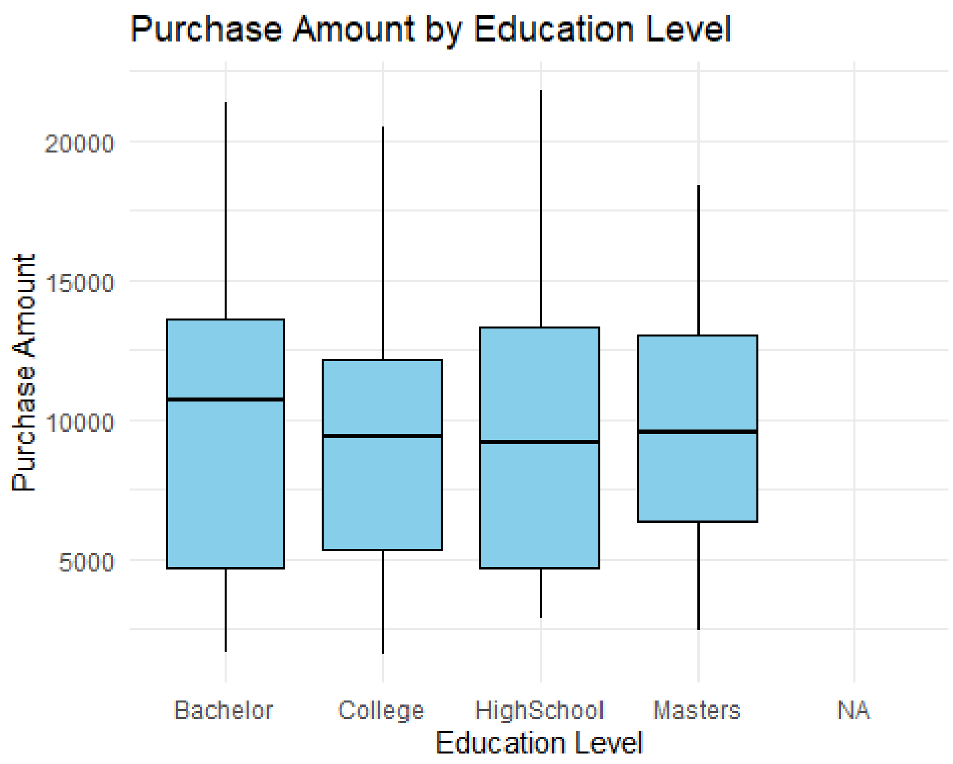
```
customer_data %>%
  group_by(age) %>%
  summarise(n())

## # A tibble: 23 × 2
##      age `n()`
##    <dbl> <int>
##  1    20     1
##  2    21     2
##  3    22     1
##  4    23     8
##  5    24     7
##  6    25    13
##  7    26     7
##  8    27    15
##  9    28    21
## 10    29    21
## # ℹ 13 more rows

#After excluding Out-liners
ggplot(data = customer_data, aes(x = age)) +
  geom_histogram(binwidth = 1) +
  xlim(23, 40) +
  theme(plot.background = element_rect(fill = "skyblue", color = "red")) +
  labs(title = "Age Frequency",
       x = "ages",
       y = "Frequency")
```



##OUTCOME: After using the group_by() function, we could easily define the OUTLINERS of the data and remove them from the histogram graph. The ages between 25 and 30 tend to have the highest frequency in the dataset.

#ANALYSIS 2

```
# Purchase Amount by Education Level: How does the education level of
customers affect
# their purchase amount?

# Visualization Tool Used : BOX PLOT

ggplot(customer_data, aes(x = education, y = purchase_amount)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Purchase Amount by Education Level",
      x = "Education Level", y = "Purchase Amount") +
  theme_minimal()
```



OUTCOME: Since, the box-plot is divided into 4 Quarters, we can analyze the Quarters for each of the category as follows

	Bachelors	College	High School	Masters
1st Quarter:	2500-5000	2500-5100	2600-4900	2500-6000
2nd Quarter:	5000-10,200	5100-9000	4900-8000	6000-9000
3rd Quarter:	10,200-11,500	9000-12000	8000-12600	9000-13000
4th Quarter:	11,500-22,000	12000-21000	12600-23000	13000-17000
Median :	10,500	9000	8000	9000

Note: (These are approx values)

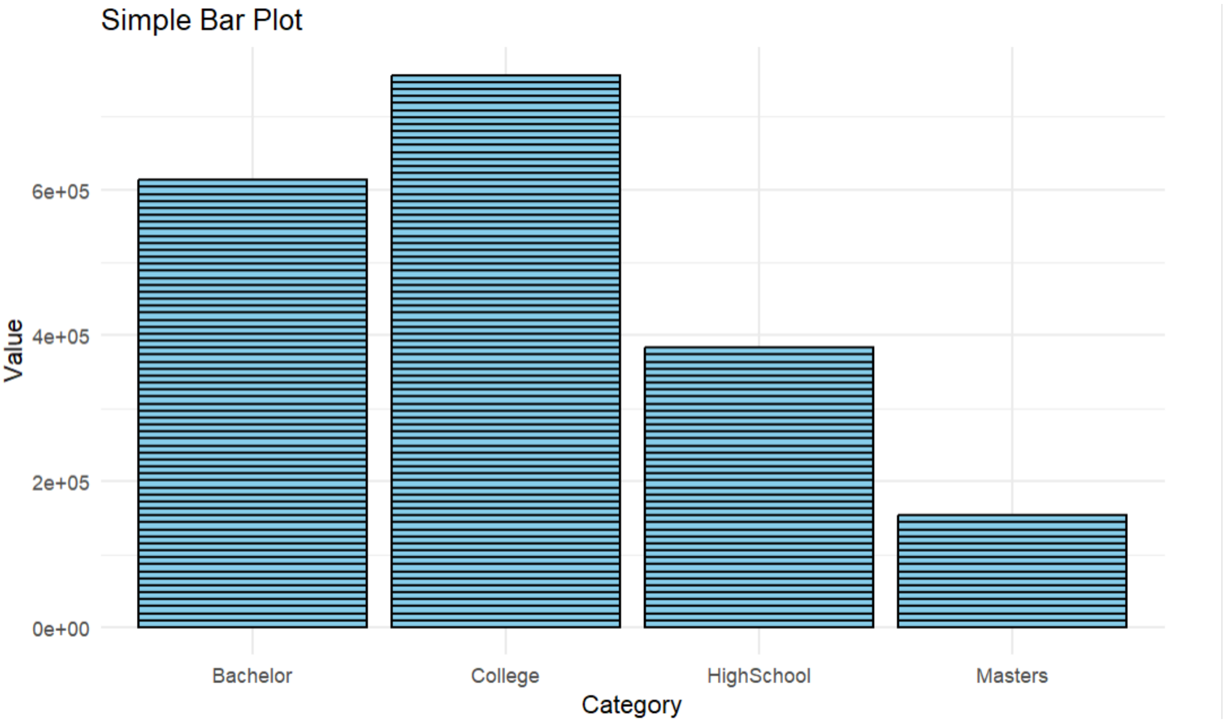
##Analysis-3

##Now, Using Box-Plot for the same info to get a better idea of the data.

```
s = mean(customer_data$purchase_amount)

customer_data %>%
  group_by(education) %>%
  summarise(s)
```

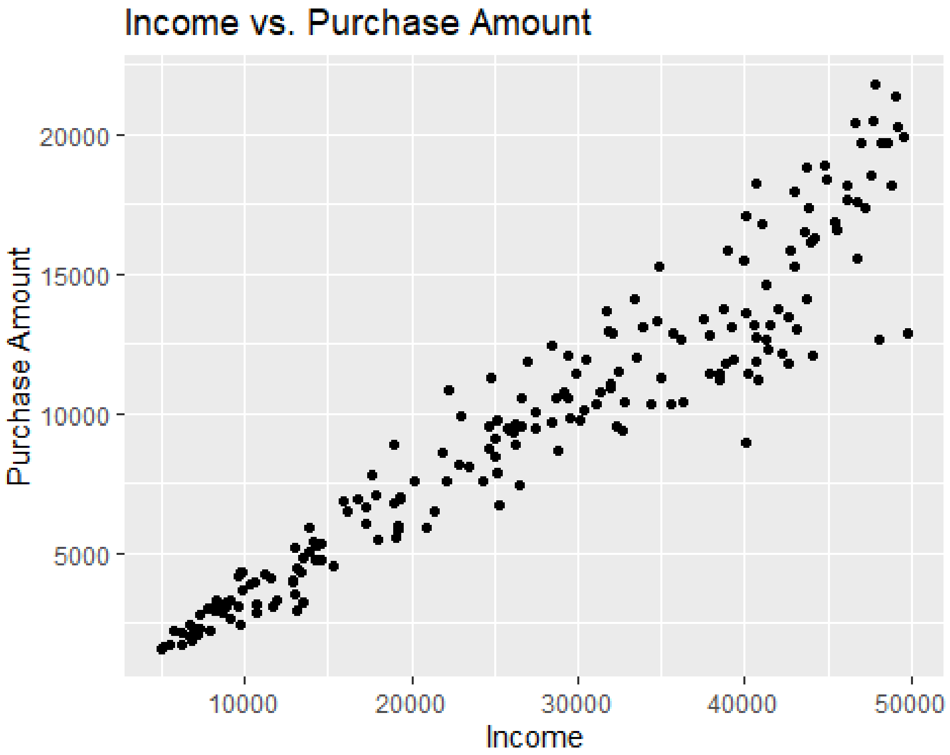
```
## # A tibble: 5 × 2
##   education      s
##   <chr>      <dbl>
## 1 Bachelor      NA
## 2 College       NA
## 3 HighSchool    NA
## 4 Masters       NA
## 5 <NA>          NA
```

```
ggplot(customer_data, aes(x = education, y = s)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Simple Bar Plot", x = "Category", y = "Value") +
  theme_minimal()
```



#ANALYSIS - 4 ##Income vs. Purchase Amount: Is there a relationship between income and purchase amount?

```
# Visualization Tool Used:  scatter plot
ggplot(data = customer_data, aes(x = income, y= purchase_amount)) +
  geom_point() +
  labs(title = "Income vs. Purchase Amount",
       x = "Income",
       y = "Purchase Amount")
```



OUTCOME: Since the dots can be seen in a linear direction; If the variables correlate they will fall along a line or curve. The stronger the correlation the tighter the data points will follow the line or curve. Therefore, these columns are correlatable…

dependent variables (y-axis): purchase_amount

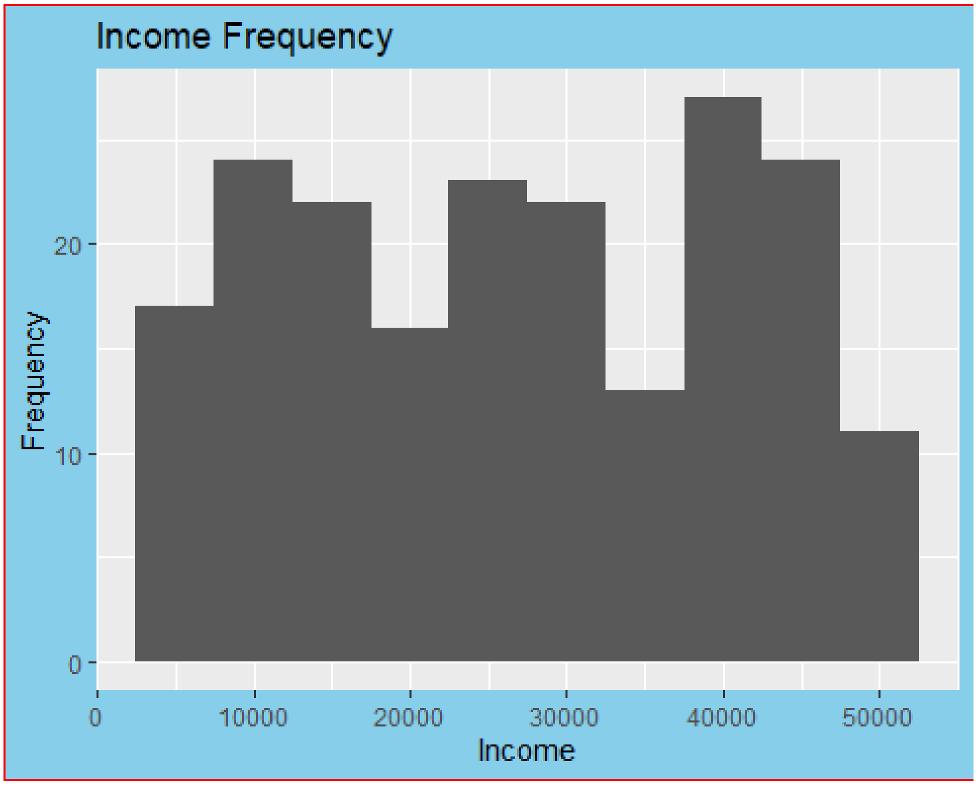
independent variables(x-axis): income

CORRELATION: POSITIVE

Analysis-5

# Income Distribution: What is the distribution of income among the customers?

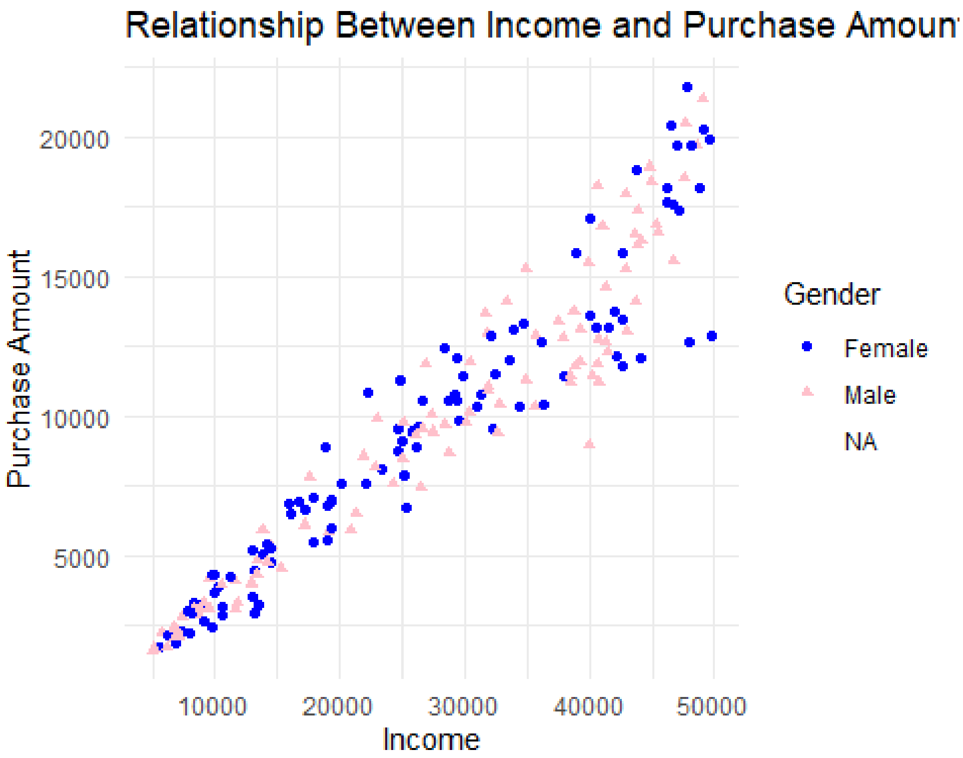
```r
ggplot(data = customer_data, aes(x = income)) +
  geom_histogram(binwidth = 5000) +
  theme(plot.background = element_rect(fill = "skyblue", color = "red")) +
  labs(title = "Income Frequency",
       x = "Income",
       y = "Frequency")
```



OUTCOME: The horizontal axis, labeled “Income,” shows different income ranges, while the vertical axis, labeled “Frequency,” represents how many customers have incomes within each range. Each bar on the chart corresponds to a particular income range, and its height indicates the number of customers falling within that range. This visualization helps you understand the distribution of income among the customers in the dataset.

#Analaysis-6

```r
# Income vs. Purchase Amount by Gender: How does the relationship between
# income and purchase amount vary between different genders?

#Visual Tool Used: Scatter Plot
ggplot(customer_data, aes(x = income, y = purchase_amount, color = gender,
shape = gender)) +
  geom_point() +
  labs(title = "Relationship Between Income and Purchase Amount by Gender",
       x = "Income", y = "Purchase Amount",
       color = "Gender", shape = "Gender") +
  scale_color_manual(values = c("blue", "pink")) +
  scale_shape_manual(values = c(16, 17)) +
  theme_minimal()
```



OUTCOME: Since the dots can be seen in a linear direction; If the variables corelate they will fall along a line or curve. The stronger the correlation the tighter the data points will follow the line or curve. Therefore, these columns are correlatable…

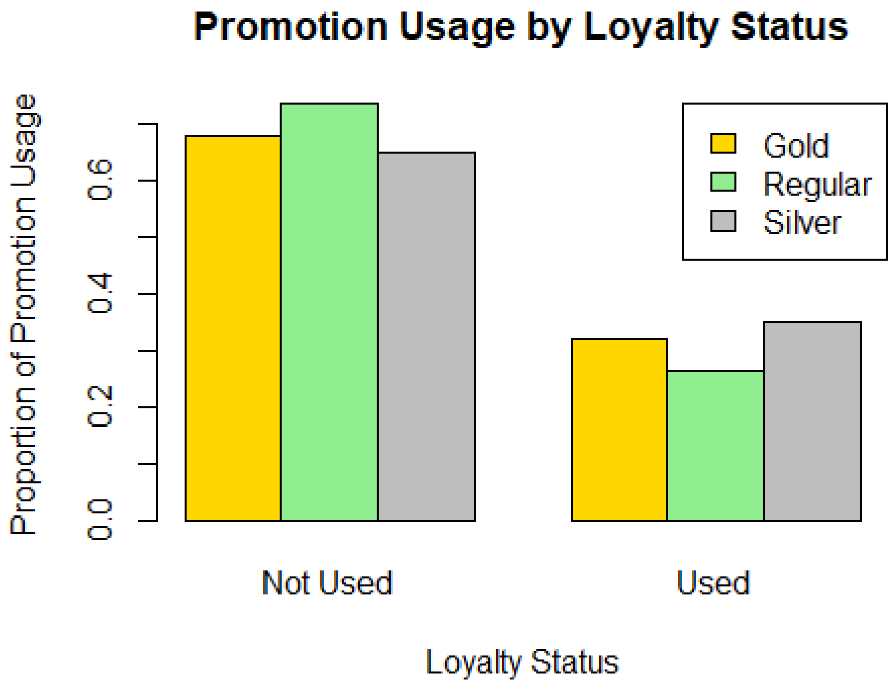
dependent variables (y-axis): purchase_amount

independent variables(x-axis): income

CORRELATION: POSITIVE

Analysis-7

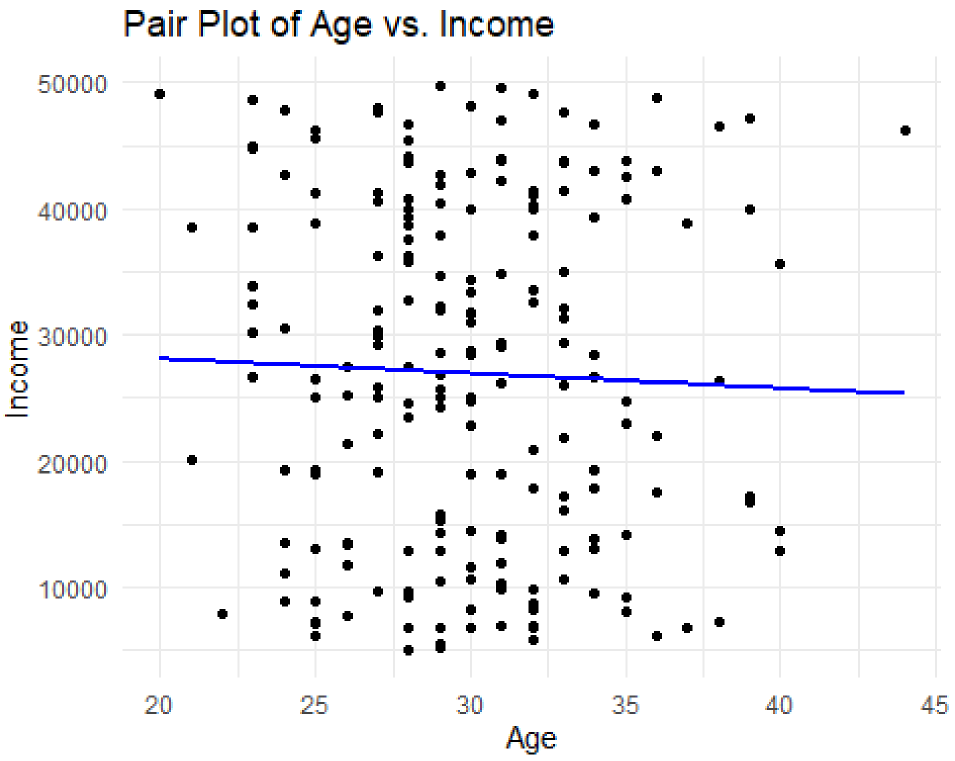
```
# Calculate proportion of promotion usage for each loyalty status
# Create a table of counts for each loyalty status and promotion usage
table_data = table(customer_data$loyalty_status,
                    ifelse(customer_data$promotion_usage >= 0.5, "Used", "Not 
Used"))

# Convert counts to proportions
prop_data = prop.table(table_data, margin = 1)

# Create a stacked bar plot
barplot(prop_data,
        beside = TRUE,
        main = "Promotion Usage by Loyalty Status",
        xlab = "Loyalty Status",
        ylab = "Proportion of Promotion Usage",
        col = c("gold", "lightgreen","grey"),
        legend = rownames(prop_data),
        args.legend = list(x = "topright")
)
```



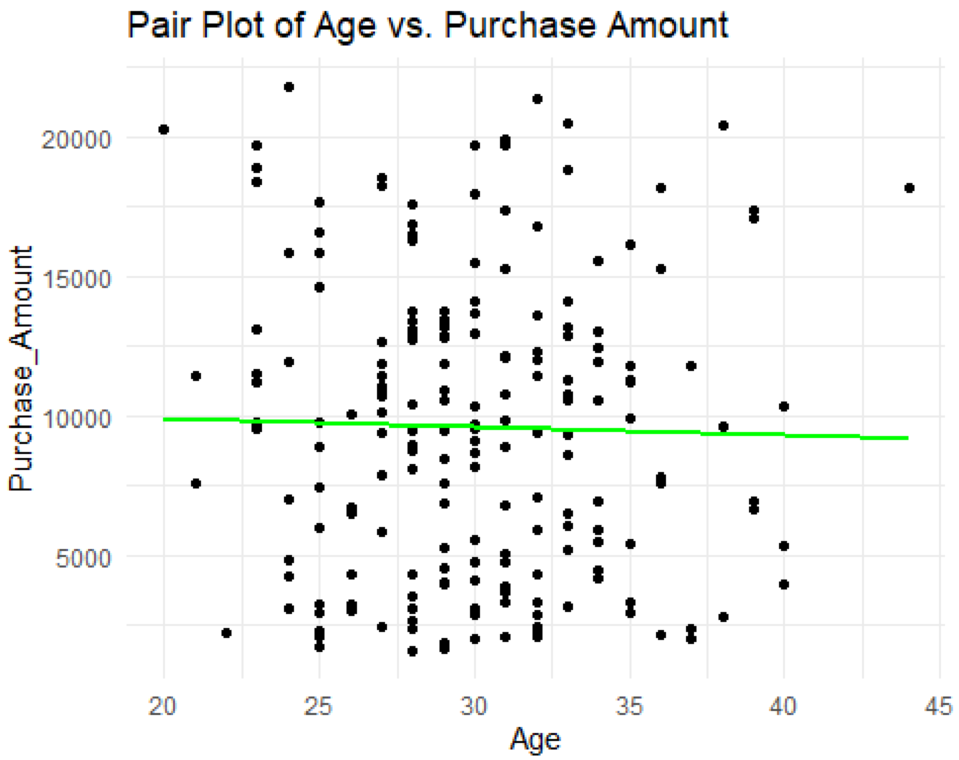
OUTCOME: This code will generate a grouped bar chart where each bar represents a different loyalty status level, and within each bar, there are stacked bars representing the proportion of promotion usage. Adjust the data and plot settings according to your actual data and preferences.

#ANALYSIS-8

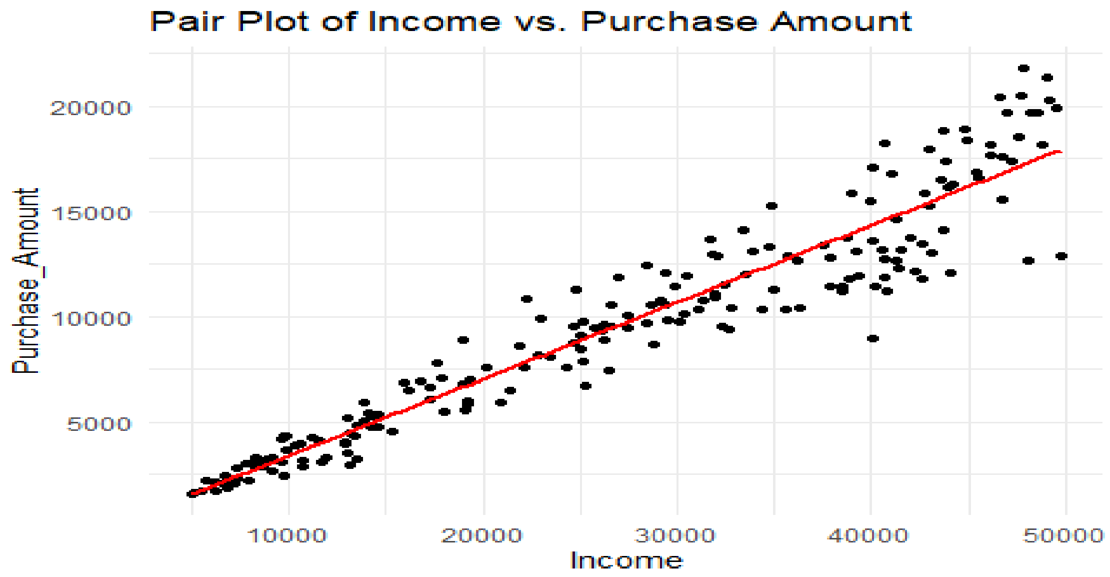
```
# Relationship Between Age, Income, and Purchase Amount:
# How do age, income, and purchase amount relate to each other in the
dataset?


data <- data.frame(Age = customer_data$age, Income = customer_data$income,
Purchase_Amount = customer_data$purchase_amount)

# Create a pair plot
ggplot(data, aes(x = Age, y = Income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Pair Plot of Age vs. Income") +
  theme_minimal()
```



```
ggplot(data, aes(x = Age, y = Purchase_Amount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  labs(title = "Pair Plot of Age vs. Purchase Amount") +
  theme_minimal()
```



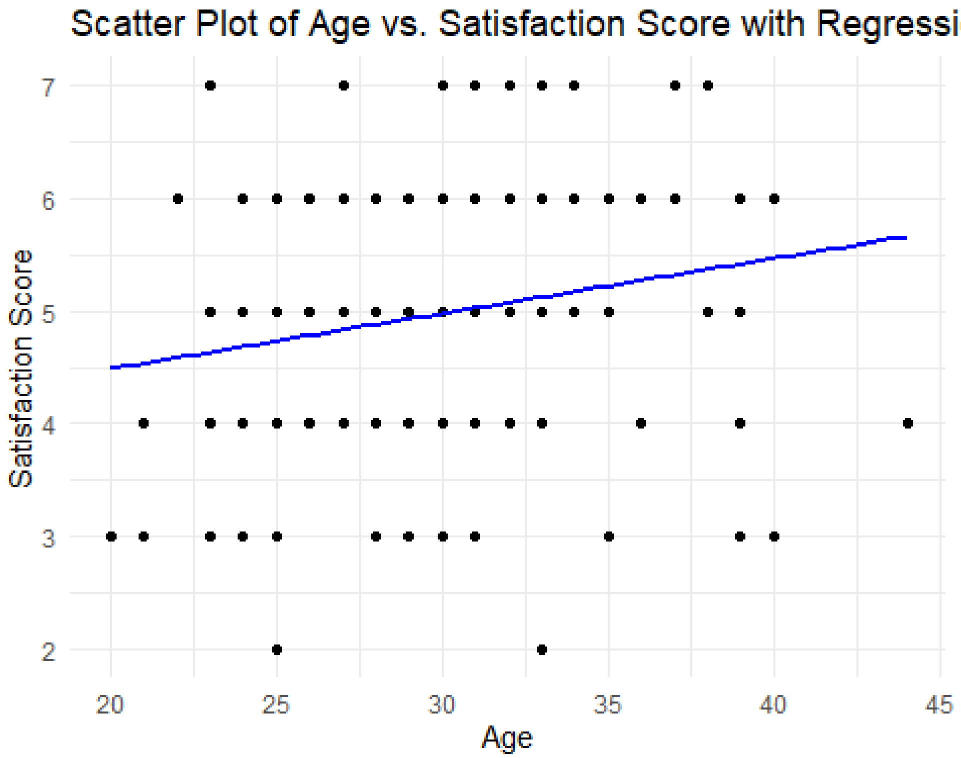
```r
ggplot(data, aes(x = Income, y = Purchase_Amount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Pair Plot of Income vs. Purchase Amount") +
  theme_minimal()
```



Correlation : Linear correlation found between Income and Purchase Amount

#ANALYSIS-9

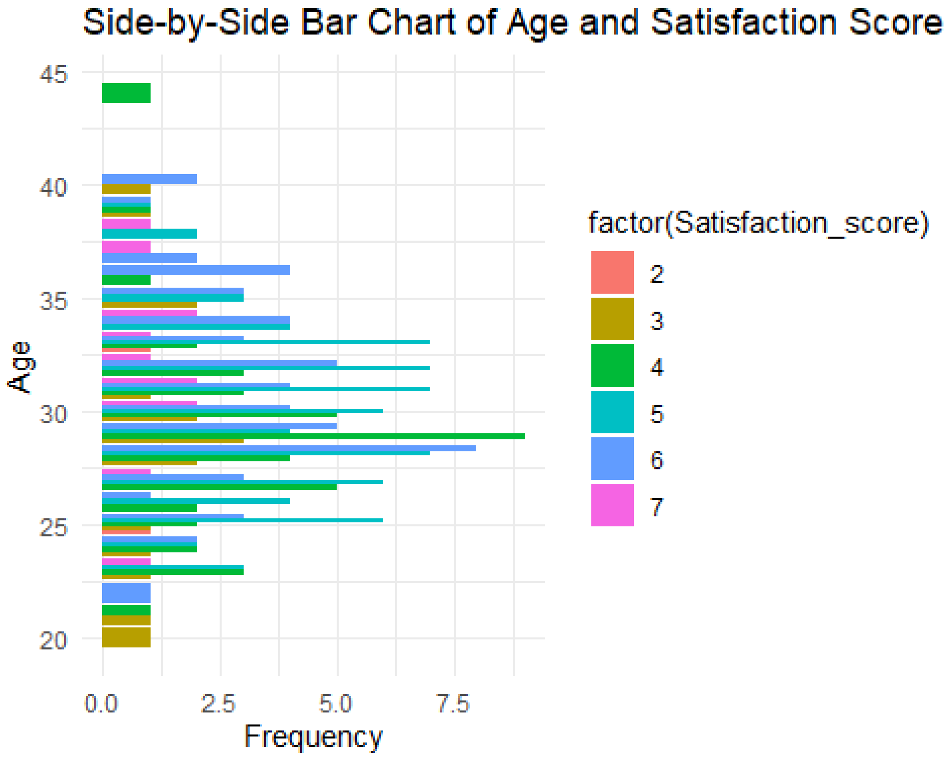
```
# Predicting Satisfaction Income on Purhcase_Amount: Can we predict the
satisfaction score of customers based on their age?
# VU: A single regression analysis with age as the independent variable and
# satisfaction score as the dependent variable. Interpret the regression
coefficients and assess the predictive power of the model.

data <- data.frame(Age = customer_data$age, Satisfaction_Score =
customer_data$satisfaction_score)

# linear regression
model <- lm(Satisfaction_Score ~ Age, data = data)

summary(model)


ggplot(data, aes(x = Age, y = Satisfaction_Score)) +
  geom_point() +  # Scatter plot
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Regression line
  labs(title = "Scatter Plot of Age vs. Satisfaction Score with Regression
Line",
       x = "Age", y = "Satisfaction Score") +  # Axes labels
  theme_minimal()
```



#Analysis-10

It was This is a poor representation; so I decided to represent it in side-by-side bar graph instead.

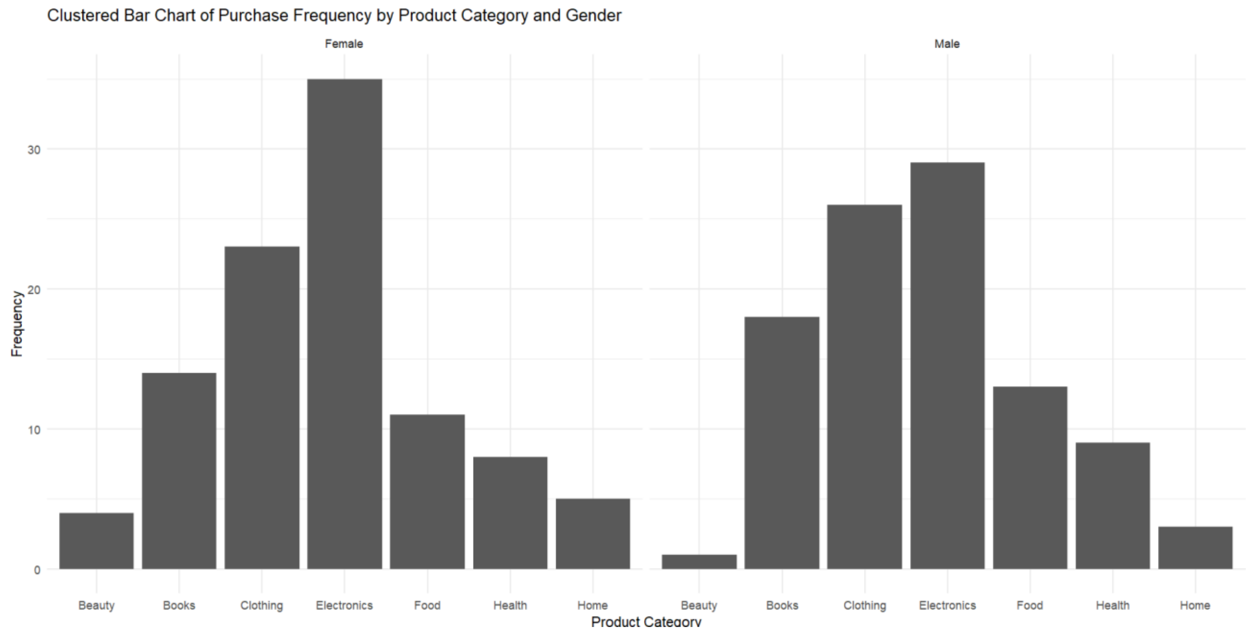
```r
data <- data.frame(
  Age = customer_data$age,
  Satisfaction_score = customer_data$satisfaction_score)

model <- lm(Satisfaction_score ~ Age, data = data)

ggplot(data, aes(x = Age)) +
  geom_bar(aes(fill = factor(Satisfaction_score)), position = "dodge") +
  labs(title = "Side-by-Side Bar Chart of Age and Satisfaction Score",
       x = "Age", y = "Frequency") +
  theme_minimal()+
  coord_flip()
```



Prediction: Customers between the range of 35-40 tend to have the highest satisfaction of their product.

#ANALYSIS-11

```
#Correlation Between Purchase Frequency, Product Category, and Income:
# Is there a visual correlation between purchase frequency, product category,
and income?
# VU: Clustered bar chart to visualize the relationship between these
variables.

ggplot(customer_data, aes(x = customer_data$product_category, fill =
customer_data$income)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ customer_data$gender) +
  labs(title = "Clustered Bar Chart of Purchase Frequency by Product Category
and Gender",
       x = "Product Category", y = "Frequency",
       fill = "Income") +
  theme_minimal()
```

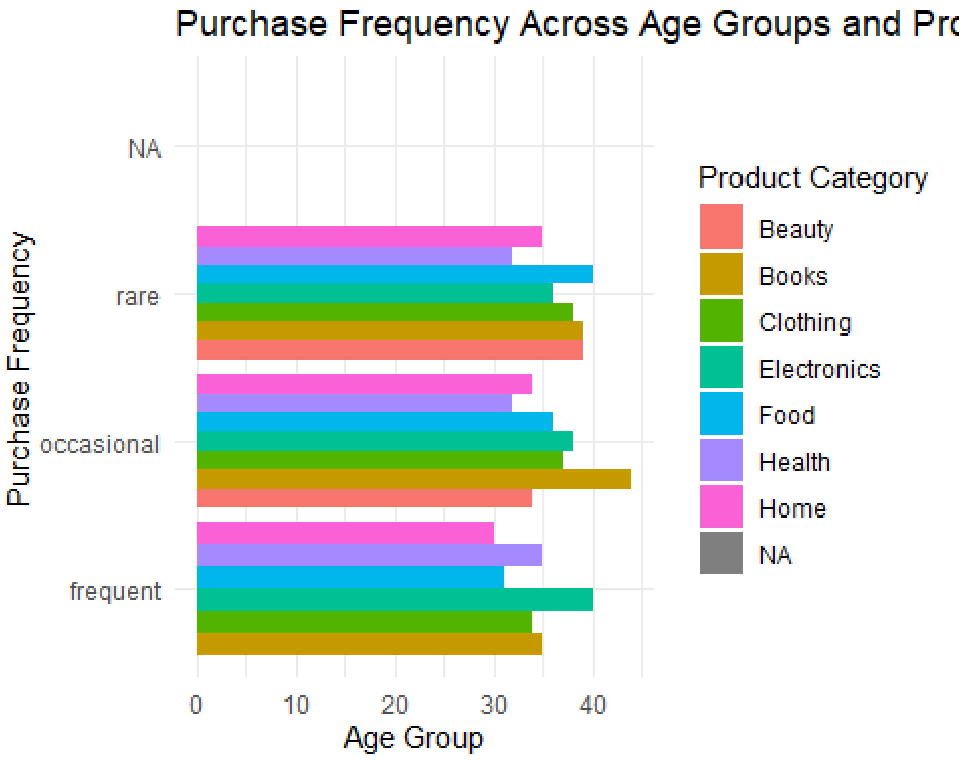


Expected Outcome: This code creates a clustered bar chart using ggplot2 to visualize the relationship between purchase frequency, product category, income, and gender in the customer_data dataset.y. Additionally, the plot is faceted by gender, allowing for separate visualizations of purchase frequency for each gender. This visualization helps to explore how purchase frequency varies across different product categories, income levels, and genders in the dataset.

Outcome Highlights: Female tend to have higher interest in Electronics then Male; while Male tend to have higher interest in beauty!

#ANALYSIS-12

```
#Comparison of Purchase Frequency Across Different Age Groups and Product
Categories:
#How does purchase frequency vary across different age groups and product
categories?
# VU: A grouped bar chart

purchase_data <- data.frame(
  Age_group = customer_data$age,
  Product_category = customer_data$product_category,
  Purchase_frequency = customer_data$purchase_frequency)

# Created a grouped bar chart
ggplot(purchase_data, aes(x = Age_group, y = Purchase_frequency, fill =
Product_category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Purchase Frequency Across Age Groups and Product Categories",
       x = "Age Group", y = "Purchase Frequency",
       fill = "Product Category") +
  theme_minimal()
```

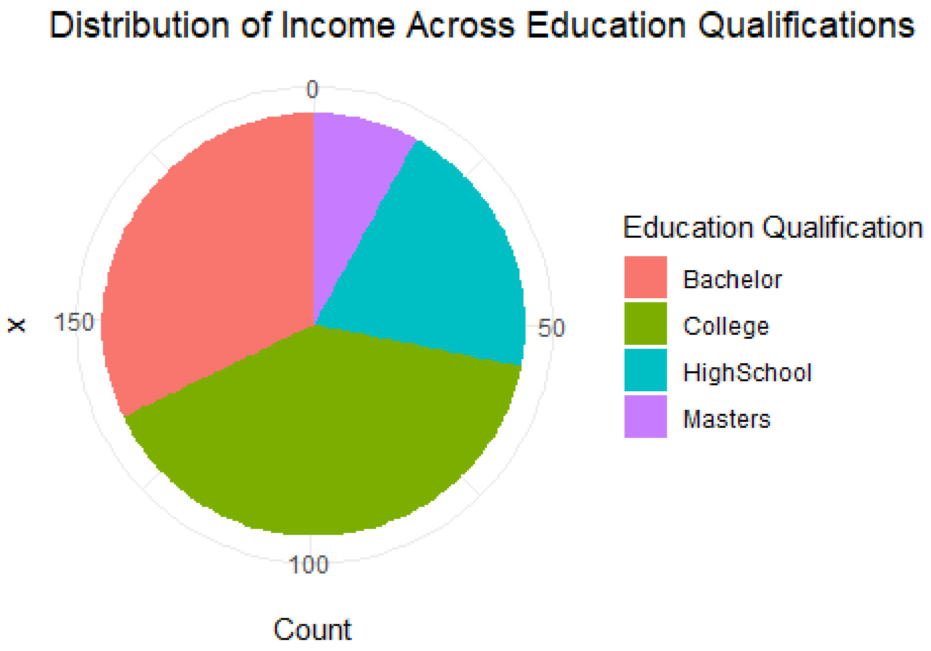


Outcome: a grouped bar chart showing the purchase frequency for different product categories across different age groups. Each group of bars represents a specific age group, and within each group, bars of different colors represent different product categories. This visualization helps to explore how purchase frequency varies across age groups and product categories in the dataset.

Outcome: Highlights: The age group of age higher than 40 is more interested in Books. Electronics are the most often bought products.

#ANALYSIS-13

```r
# Multivariate Analysis of Customer Behavior: How do multiple variables
collectively influence purchasing behavior?
#  VU: Multidimensional scatter plot or ternary plot to visualize the
relationships between these variables and their impact on purchasing
behavior.

library(ggplot2)

cu <- data.frame(
  Income = customer_data$income,
  Education_Qualification = customer_data$education)

# Count the number of observations for each combination of income and
education qualification
income_education_count <- table(customer_data$income,
customer_data$education)

# Convert the table to a data frame
income_education_count <- as.data.frame(income_education_count)
colnames(income_education_count) <- c("Income", "Education_Qualification",
"Count")

# Create a pie chart
ggplot(income_education_count, aes(x = "", y = Count, fill =
Education_Qualification)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of Income Across Education Qualifications",
       fill = "Education Qualification") +
  theme_minimal() +
  theme(legend.position = "right")
```



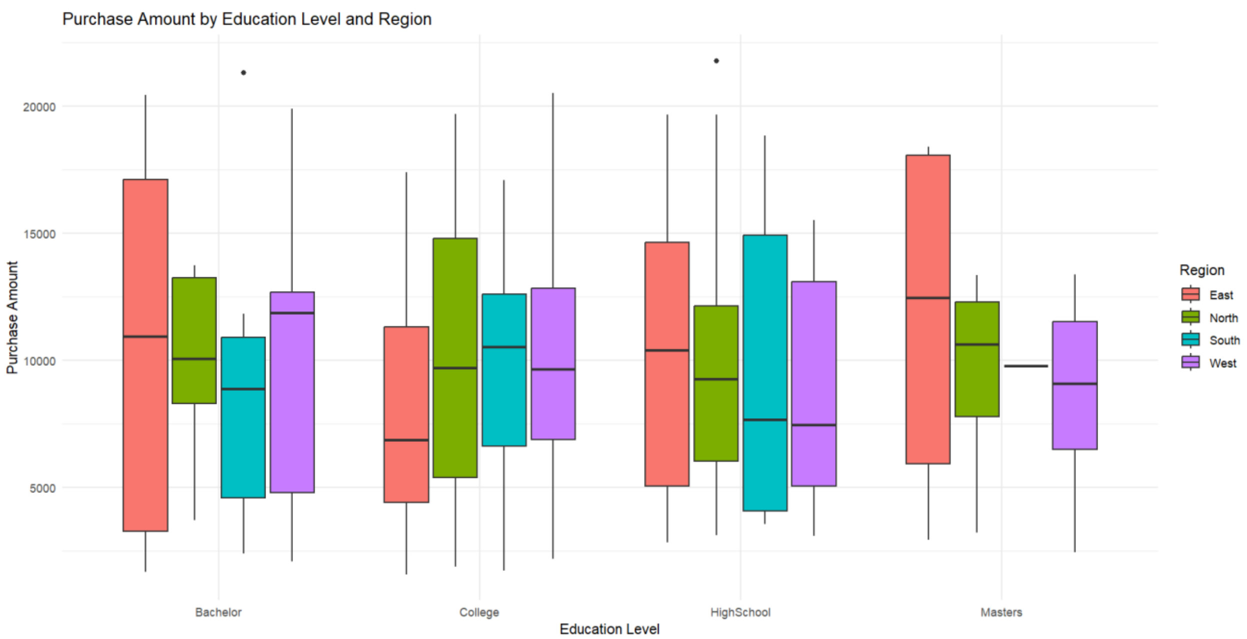
Expected Outcome: . Each segment of the pie chart represents a specific education qualification, and the size of each segment corresponds to the count of observations for that qualification. This visualization helps to understand how income is distributed among individuals with different education qualifications.

Outcome Highlights: College Degree Education tend to have the Income in the customer data; while the Masters degree customers are having the lowest ones.

#ANALYSIS-14

```
# Perform ANOVA
anova_result <- aov(customer_data$purchase_amount ~ customer_data$education *
customer_data$region, data = customer_data)

# Summary of ANOVA
summary(anova_result)

##                                                  Df    Sum Sq  Mean Sq F
value
## customer_data$education                           3 3.890e+07 12966775
0.449
## customer_data$region                              3 2.293e+07  7643301
0.265
```

```
## customer_data$education:customer_data$region   9 1.259e+08 13984331
0.485
## Residuals                                    183 5.281e+09 28857971
##                                              Pr(>F)
## customer_data$education                       0.718
## customer_data$region                          0.851


# Visualize the results with boxplots
ggplot(customer_data, aes(x = education, y = purchase_amount, fill = region))
+
  geom_boxplot() +
  labs(title = "Purchase Amount by Education Level and Region",
       x = "Education Level", y = "Purchase Amount",
       fill = "Region") +
  theme_minimal()
```

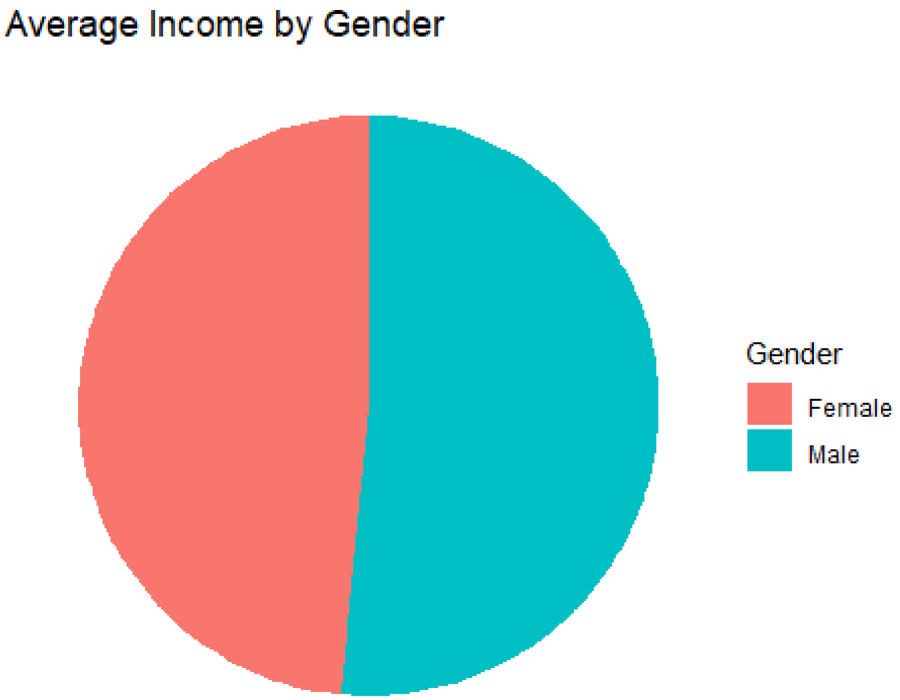


Expected Outcome: A visualization of purchase amount by education level and region using boxplots. This allows you to explore the variation in purchase amount across different education levels and regions, and assess whether there are significant differences in purchase amount between groups.

Pie chart showing the distribution of income across different education qualifications in the dataset. Each slice of the pie represents a specific education qualification, and the size of the slice corresponds to the proportion of individuals with that education qualification in each income category.

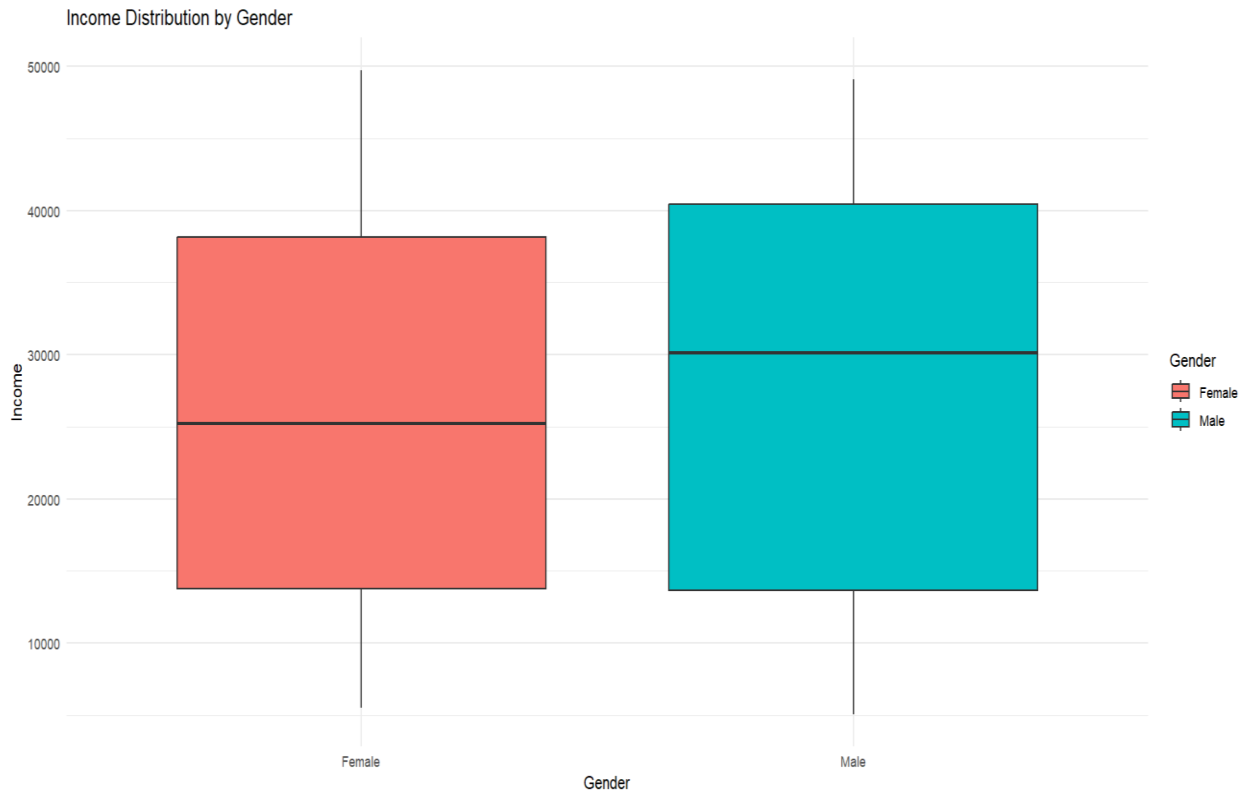
#ANALYSIS-15

```r
library(ggplot2)


average_income = aggregate(income ~ gender, data = customer_data, FUN = mean)

# Created pie chart
ggplot(average_income, aes(x = "", y = income, fill = gender)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Average Income by Gender",
       x = NULL, y = NULL,
       fill = "Gender") +
  theme_void() +
  theme(legend.position = "right")
```



#ANALYSIS-16

```
library(ggplot2)


ggplot(customer_data, aes(x = gender, y = income, fill = gender)) +
  geom_boxplot() +
  labs(title = "Income Distribution by Gender",
       x = "Gender", y = "Income",
       fill = "Gender") +
  theme_minimal()
```


Income Distribution by Gender
Income
Female
Male
Gender
Gender
🟥 Female
🟦 Male