Analysis and Visualization of customer shopping dataset in the United States

1. DATASET

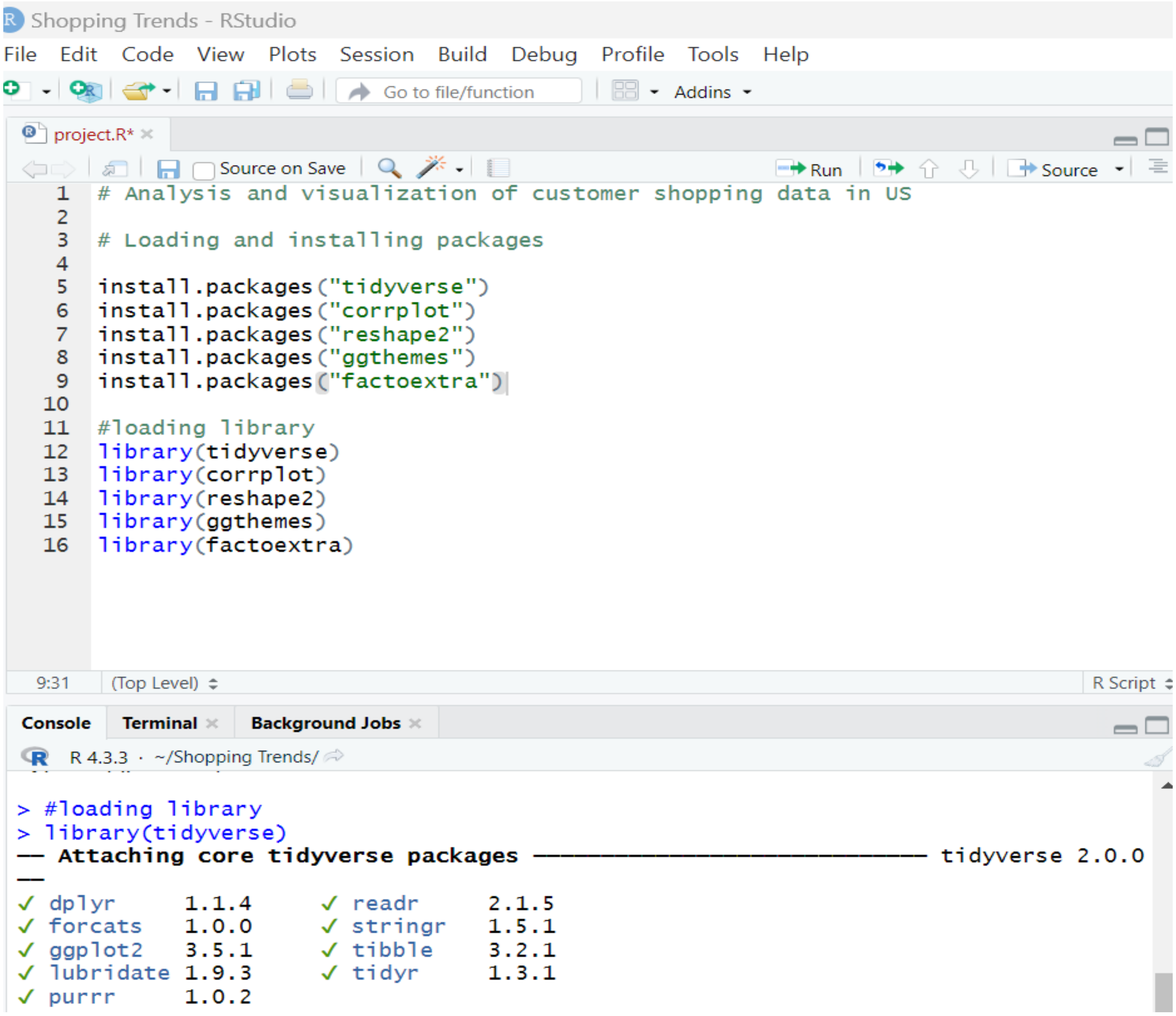
1.1 DESCRIPTION OF DATASET

The dataset shows the documentation on how people in the United States shop. It has information on 3,900 shopping trips (3,900 rows and 19 columns), with details like age, gender,

what was bought, how much it cost, and customer reviews. We can use this data to understand how people shop, what they like, and how their shopping habits change over time.
This data was gotten from Kaggle. It's called the "Customer Shopping Latest Trends Dataset" and can be viewed through this link:
https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset

1.2 INSTALLING PACKAGES AND LIBRARIES

Before I start my project, I get ready by installing packages that will help me understand and show the data. I installed packages like tidyverse and some other special tools. You can see some of the tools I installed in the picture below.



```r
# Analysis and visualization of customer shopping data in US

# Loading and installing packages

install.packages("tidyverse")
install.packages("corrplot")
install.packages("reshape2")
install.packages("ggthemes")
install.packages("factoextra")

#loading library
library(tidyverse)
library(corrplot)
library(reshape2)
library(ggthemes)
library(factoextra)
```

```
> #loading library
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────────── tidyverse 2.0.0
──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ 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
```

Figure 1

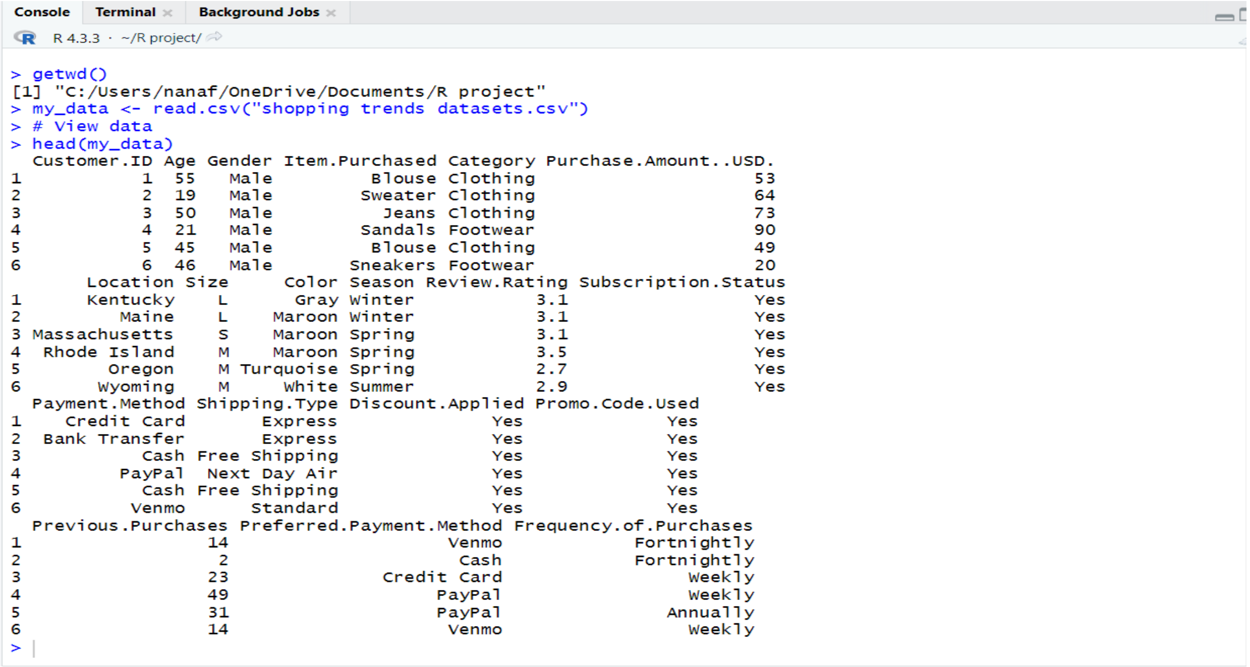
1.3 READING THE .CSV FILE

While I was adding my data to R, I used a function called getwd () to see where my files were saved on my computer.

```
> getwd()
[1] "C:/Users/nanaf/OneDrive/Documents/R project"
```

I used the head () function to view the first few rows (typically 6) of the data (see Figure 2). This helped me confirm that all the fields were imported successfully and appeared as expected in R.

I saved the data file (a .csv file) in my Documents folder. This way, it's easy to find when I'm working in R. Then, I used RStudio to open the data file. Here's how I did it:

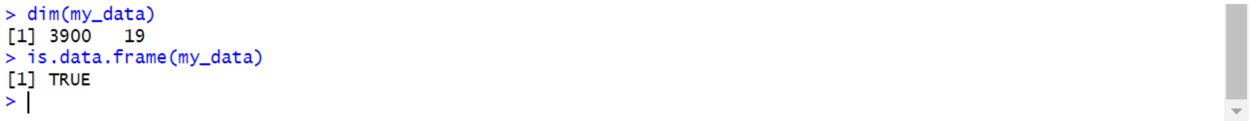


```
> getwd()
[1] "C:/Users/nanaf/OneDrive/Documents/R project"
> my_data <- read.csv("shopping trends datasets.csv")
> # View data
> head(my_data)
  Customer.ID Age Gender Item.Purchased Category Purchase.Amount..USD.
1           1  55   Male         Blouse Clothing                    53
2           2  19   Male        Sweater Clothing                    64
3           3  50   Male          Jeans Clothing                    73
4           4  21   Male        Sandals Footwear                    90
5           5  45   Male         Blouse Clothing                    49
6           6  46   Male       Sneakers Footwear                    20
       Location Size     Color Season Review.Rating Subscription.Status
1      Kentucky    L      Gray Winter           3.1                 Yes
2         Maine    L    Maroon Winter           3.1                 Yes
3 Massachusetts    S    Maroon Spring           3.1                 Yes
4  Rhode Island    M    Maroon Spring           3.5                 Yes
5        Oregon    M Turquoise Spring           2.7                 Yes
6       Wyoming    M     White Summer           2.9                 Yes
  Payment.Method Shipping.Type Discount.Applied Promo.Code.Used
1    Credit Card       Express              Yes             Yes
2  Bank Transfer       Express              Yes             Yes
3           Cash Free Shipping              Yes             Yes
4         PayPal  Next Day Air              Yes             Yes
5           Cash Free Shipping              Yes             Yes
6          Venmo      Standard              Yes             Yes
  Previous.Purchases Preferred.Payment.Method Frequency.of.Purchases
1                 14                    Venmo            Fortnightly
2                  2                     Cash            Fortnightly
3                 23              Credit Card                 Weekly
4                 49                   PayPal                 Weekly
5                 31                   PayPal               Annually
6                 14                    Venmo                 Weekly
> 
```

Figure 2

- I used the is.data.frame () function to verify that the data loaded as a data frame, the preferred format for working with data in R.

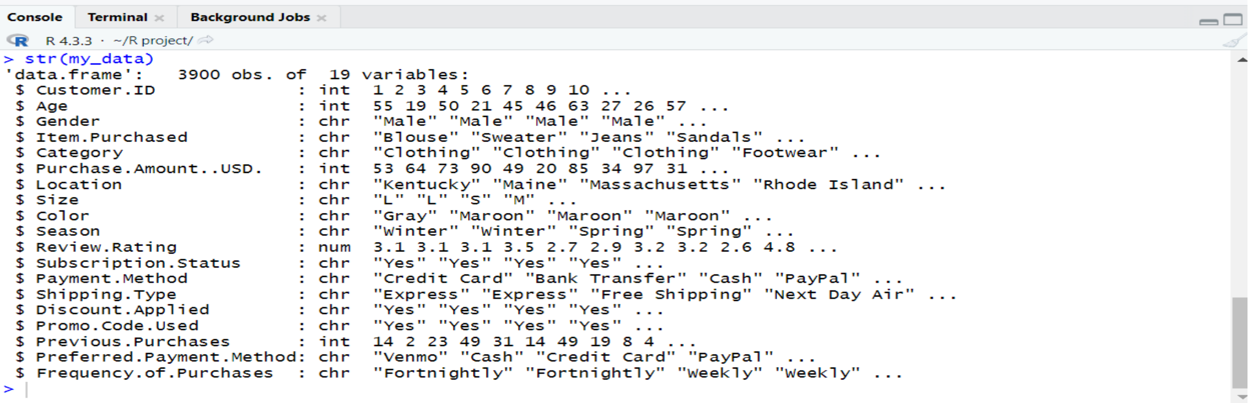
- Next, I used dim(my_data) to check the dimensions of the data, which tells me the number of rows (observations) and columns (variables).



```
> dim(my_data)
[1] 3900   19
> is.data.frame(my_data)
[1] TRUE
> 
```

Figure 3

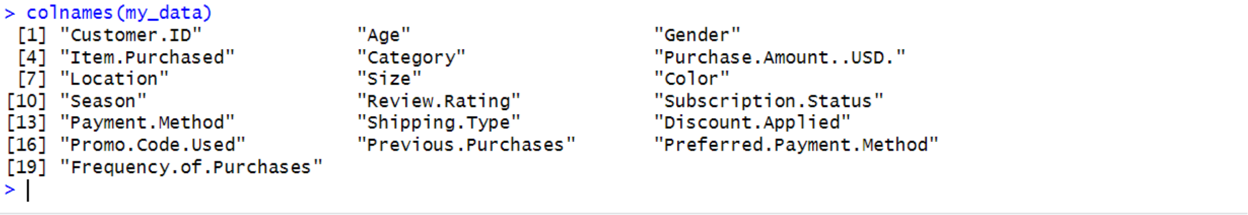
The str () function was used to examine the structure of the data, including the data types of each variable. This helped identify any variables that might need data type conversions or other adjustments.



```
> str(my_data)
'data.frame':	3900 obs. of  19 variables:
 $ Customer.ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Age                     : int  55 19 50 21 45 46 63 27 26 57 ...
 $ Gender                  : chr  "Male" "Male" "Male" "Male" ...
 $ Item.Purchased          : chr  "Blouse" "Sweater" "Jeans" "Sandals" ...
 $ Category                : chr  "Clothing" "Clothing" "Clothing" "Footwear" ...
 $ Purchase.Amount..USD.   : int  53 64 73 90 49 20 85 34 97 31 ...
 $ Location                : chr  "Kentucky" "Maine" "Massachusetts" "Rhode Island" ...
 $ Size                    : chr  "L" "L" "S" "M" ...
 $ Color                   : chr  "Gray" "Maroon" "Maroon" "Maroon" ...
 $ Season                  : chr  "Winter" "Winter" "Spring" "Spring" ...
 $ Review.Rating           : num  3.1 3.1 3.1 3.5 2.7 2.9 3.2 3.2 2.6 4.8 ...
 $ Subscription.Status     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Payment.Method          : chr  "Credit Card" "Bank Transfer" "Cash" "PayPal" ...
 $ Shipping.Type           : chr  "Express" "Express" "Free Shipping" "Next Day Air" ...
 $ Discount.Applied        : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Promo.Code.Used         : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Previous.Purchases      : int  14 2 23 49 31 14 49 19 8 4 ...
 $ Preferred.Payment.Method: chr  "Venmo" "Cash" "Credit Card" "PayPal" ...
 $ Frequency.of.Purchases  : chr  "Fortnightly" "Fortnightly" "Weekly" "Weekly" ...
> 
```

Figure 4

The colnames() function was used to list the names of all 19 variables in the dataset.



```
> colnames(my_data)
 [1] "Customer.ID"              "Age"                      "Gender"                  
 [4] "Item.Purchased"           "Category"                 "Purchase.Amount..USD."   
 [7] "Location"                 "Size"                     "Color"                   
[10] "Season"                   "Review.Rating"            "Subscription.Status"     
[13] "Payment.Method"           "Shipping.Type"            "Discount.Applied"        
[16] "Promo.Code.Used"          "Previous.Purchases"       "Preferred.Payment.Method"
[19] "Frequency.of.Purchases"  
> |
```

Figure 5

1.4 TABLE OF VARIABLES AND DESCRIPTION

Variable Name	Mode	Description
Customer ID	int	Unique identifier for each customer.
Age	int	The age of the customer in years.
Gender	chr	The gender of the customer (e.g., Male, Female).
Item Purchased	chr	Name of the item purchased by the customer.
Category	chr	The category of the purchased item (e.g., Clothing, Footwear).
Purchase Amount (USD)	int	The total amount spent by the customer in USD.
Location	chr	The geographic location of the customer.
Size	chr	The size of the purchased item (e.g., S, M, L).
Color	chr	The color of the purchased item.
Season	chr	The season in which the purchase was made (e.g., Spring, Winter).
Review Rating	num	The customer's rating for the product, ranging from 1 to 5.
Subscription Status	chr	Indicates whether the customer is subscribed (e.g., Yes, No).
Payment Method	chr	The payment method used (e.g., Credit Card, PayPal).
Shipping Type	chr	The shipping method selected (e.g., Free Shipping, Express).

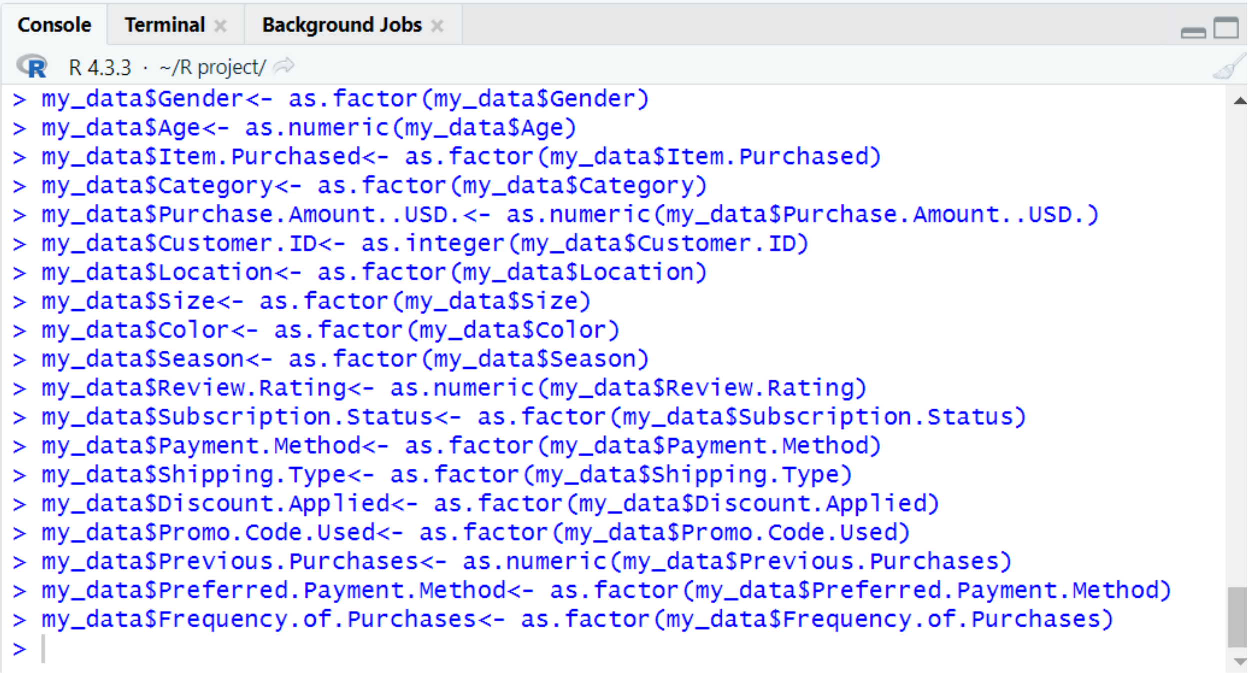
Discount Applied	chr	Indicates if a discount was applied (e.g., Yes, No).
Promo Code Used	chr	Indicates if a promotional code was used (e.g., Yes, No).
Previous Purchases	int	The number of prior purchases made by the customer.
Preferred Payment Method	chr	The customer's preferred payment method.
Frequency of Purchases	chr	How often the customer makes purchases (e.g., Weekly, Monthly, Annually).

1.5 EXPECTATION

I expect this analysis to reveal patterns in how customers shop, like the most popular product categories, seasonal trends, and preferences based on demographics. I also aim to find connections between factors such as review ratings, how much customers spend, and how often they make purchases.

1.6 DATA CLEANING

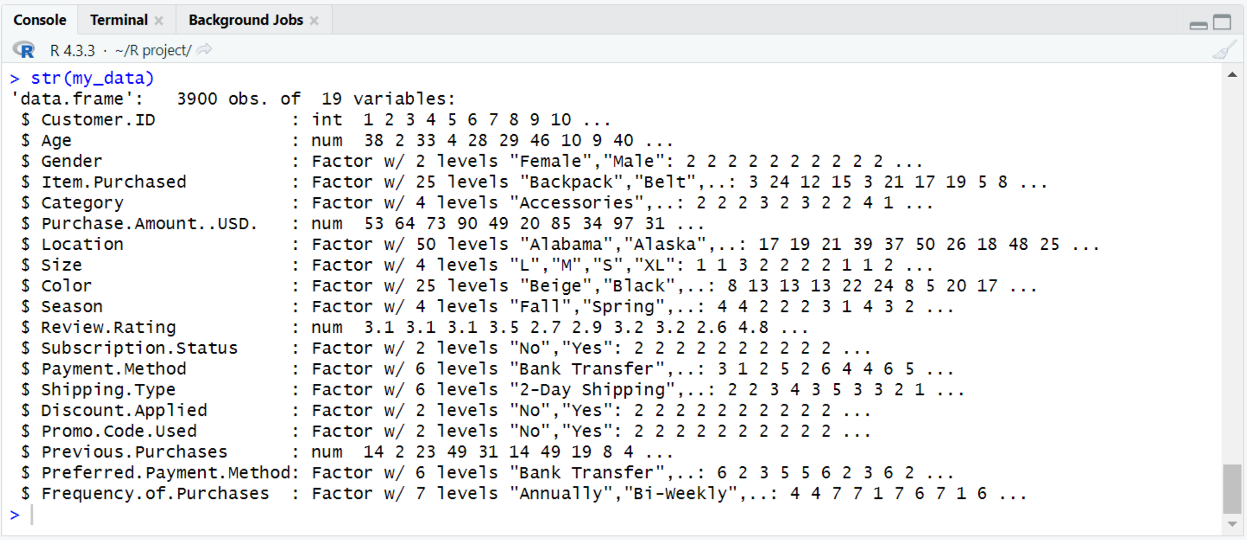
I checked the data to see if anything was missing. Everything appeared to be present, and there were no duplicates. However, the variables were not properly defined. To address this, I cleaned the data to make it clearer and easier to understand and analyze.
I changed the way some of the data were categorized. For example, I changed from “chr” to “Factor”; from “int” to “Factor”; from “int” to “num”, etc.)



```
> my_data$Gender<- as.factor(my_data$Gender)
> my_data$Age<- as.numeric(my_data$Age)
> my_data$Item.Purchased<- as.factor(my_data$Item.Purchased)
> my_data$Category<- as.factor(my_data$Category)
> my_data$Purchase.Amount..USD.<- as.numeric(my_data$Purchase.Amount..USD.)
> my_data$Customer.ID<- as.integer(my_data$Customer.ID)
> my_data$Location<- as.factor(my_data$Location)
> my_data$Size<- as.factor(my_data$Size)
> my_data$Color<- as.factor(my_data$Color)
> my_data$Season<- as.factor(my_data$Season)
> my_data$Review.Rating<- as.numeric(my_data$Review.Rating)
> my_data$Subscription.Status<- as.factor(my_data$Subscription.Status)
> my_data$Payment.Method<- as.factor(my_data$Payment.Method)
> my_data$Shipping.Type<- as.factor(my_data$Shipping.Type)
> my_data$Discount.Applied<- as.factor(my_data$Discount.Applied)
> my_data$Promo.Code.Used<- as.factor(my_data$Promo.Code.Used)
> my_data$Previous.Purchases<- as.numeric(my_data$Previous.Purchases)
> my_data$Preferred.Payment.Method<- as.factor(my_data$Preferred.Payment.Method)
> my_data$Frequency.of.Purchases<- as.factor(my_data$Frequency.of.Purchases)
> 
```

Figure 6

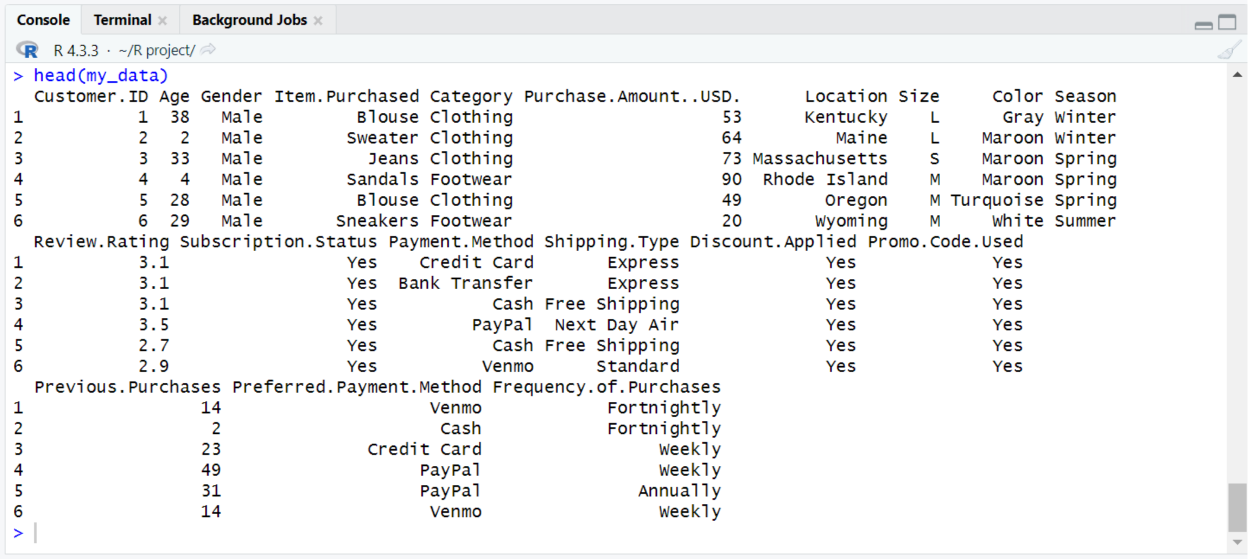
The cleaned data now includes 4 numerical variables and 15 categorical variables and the “str function” was used to show the new data, as detailed below:



```
> str(my_data)
'data.frame':	3900 obs. of  19 variables:
 $ Customer.ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Age                     : num  38 2 33 4 28 29 46 10 9 40 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Item.Purchased          : Factor w/ 25 levels "Backpack","Belt",..: 3 24 12 15 3 21 17 19 5 8 ...
 $ Category                : Factor w/ 4 levels "Accessories",..: 2 2 2 3 2 3 2 2 4 1 ...
 $ Purchase.Amount..USD.   : num  53 64 73 90 49 20 85 34 97 31 ...
 $ Location                : Factor w/ 50 levels "Alabama","Alaska",..: 17 19 21 39 37 50 26 18 48 25 ...
 $ Size                    : Factor w/ 4 levels "L","M","S","XL": 1 1 3 2 2 2 2 1 1 2 ...
 $ Color                   : Factor w/ 25 levels "Beige","Black",..: 8 13 13 13 22 24 8 5 20 17 ...
 $ Season                  : Factor w/ 4 levels "Fall","Spring",..: 4 4 2 2 2 3 1 4 3 2 ...
 $ Review.Rating           : num  3.1 3.1 3.1 3.5 2.7 2.9 3.2 3.2 2.6 4.8 ...
 $ Subscription.Status     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Payment.Method          : Factor w/ 6 levels "Bank Transfer",..: 3 1 2 5 2 6 4 4 6 5 ...
 $ Shipping.Type           : Factor w/ 6 levels "2-Day Shipping",..: 2 2 3 4 3 5 3 3 2 1 ...
 $ Discount.Applied        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Promo.Code.Used         : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Previous.Purchases      : num  14 2 23 49 31 14 49 19 8 4 ...
 $ Preferred.Payment.Method: Factor w/ 6 levels "Bank Transfer",..: 6 2 3 5 5 6 2 3 6 2 ...
 $ Frequency.of.Purchases  : Factor w/ 7 levels "Annually","Bi-Weekly",..: 4 4 7 7 1 7 6 7 1 6 ...
> 
```

Figure 7

Here's the output from the head () function and it did not change



```
> head(my_data)
  Customer.ID Age Gender Item.Purchased Category Purchase.Amount..USD.      Location Size     Color Season
1           1  38   Male         Blouse Clothing                    53      Kentucky    L      Gray Winter
2           2   2   Male        Sweater Clothing                    64         Maine    L    Maroon Winter
3           3  33   Male          Jeans Clothing                    73 Massachusetts    S    Maroon Spring
4           4   4   Male        Sandals Footwear                    90  Rhode Island    M    Maroon Spring
5           5  28   Male         Blouse Clothing                    49        Oregon    M Turquoise Spring
6           6  29   Male       Sneakers Footwear                    20       Wyoming    M     White Summer
  Review.Rating Subscription.Status Payment.Method Shipping.Type Discount.Applied Promo.Code.Used
1           3.1                 Yes    Credit Card       Express              Yes             Yes
2           3.1                 Yes  Bank Transfer       Express              Yes             Yes
3           3.1                 Yes           Cash Free Shipping              Yes             Yes
4           3.5                 Yes         PayPal  Next Day Air              Yes             Yes
5           2.7                 Yes           Cash Free Shipping              Yes             Yes
6           2.9                 Yes          Venmo      Standard              Yes             Yes
  Previous.Purchases Preferred.Payment.Method Frequency.of.Purchases
1                 14                    Venmo            Fortnightly
2                  2                     Cash            Fortnightly
3                 23              Credit Card                 Weekly
4                 49                   PayPal                 Weekly
5                 31                   PayPal               Annually
6                 14                    Venmo                 Weekly
> 
```

Figure 8

Table of Variables and Descriptions After cleaning

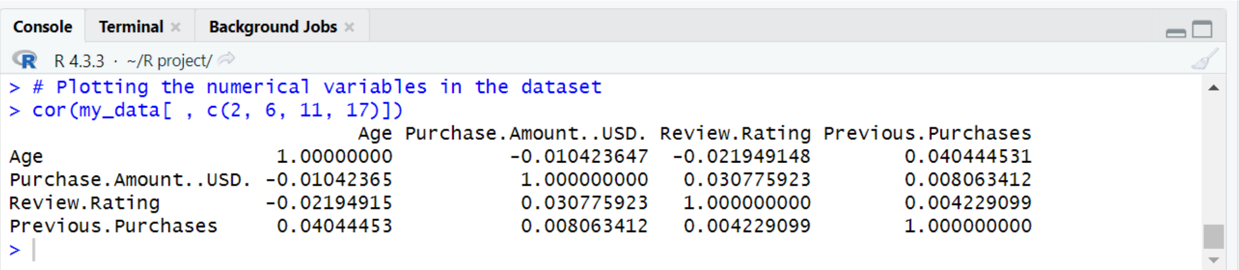
Variable Name	**Mode**	**Description**
Customer ID	int	Unique identifier for each customer.
Age	num	The age of the customer in years.
Gender	fac	The gender of the customer (e.g., Male, Female).
Item Purchased	fac	Name of the item purchased by the customer.
Category	fac	The category of the purchased item (e.g., Clothing, Footwear).
Purchase Amount (USD)	num	The total amount spent by the customer in USD.
Location	fac	The geographic location of the customer.

Size	fac	The size of the purchased item (e.g., S, M, L).
Color	fac	The color of the purchased item.
Season	fac	The season in which the purchase was made (e.g., Spring, Winter).
Review Rating	num	The customer's rating for the product, ranging from 1 to 5.
Subscription Status	fac	Indicates whether the customer is subscribed (e.g., Yes, No).
Payment Method	fac	The payment method used (e.g., Credit Card, PayPal).
Shipping Type	fac	The shipping method selected (e.g., Free Shipping, Express).
Discount Applied	fac	Indicates if a discount was applied (e.g., Yes, No).
Promo Code Used	fac	Indicates if a promotional code was used (e.g., Yes, No).
Previous Purchases	num	The number of prior purchases made by the customer.
Preferred Payment Method	fac	The customer's preferred payment method.
Frequency of Purchases	fac	How often the customer makes purchases (e.g., Weekly, Monthly, Annually).

2. DATA ANALYSIS

2.1 SCATTERPLOTS AND CORRELATION MATRICES OF NUMERIC VARIABLES

The dataset includes different types of variables, such as numeric and categorical variables. I focused on the numerical variables to find patterns and connections between them. To do this, I used the cor () function, which helps check how much the variables are related to one another.



```
> # Plotting the numerical variables in the dataset
> cor(my_data[ , c(2, 6, 11, 17)])
                            Age Purchase.Amount..USD. Review.Rating Previous.Purchases
Age                  1.00000000          -0.010423647  -0.021949148        0.040444531
Purchase.Amount..USD. -0.01042365           1.000000000   0.030775923        0.008063412
Review.Rating        -0.02194915           0.030775923   1.000000000        0.004229099
Previous.Purchases    0.04044453           0.008063412   0.004229099        1.000000000
> 
```

Figure 9

The correlation table reveals that the variables in the dataset have very weak relationships with each other. Age, Purchase Amount, Review Rating, and Previous Purchases are largely independent, showing almost no linear connection. This suggests that these variables don't significantly influence one another within the dataset.

```
> my_data_numericals<-data.frame(my_data[ , c(2,6,11,17)])
> plot(my_data_numericals)
```

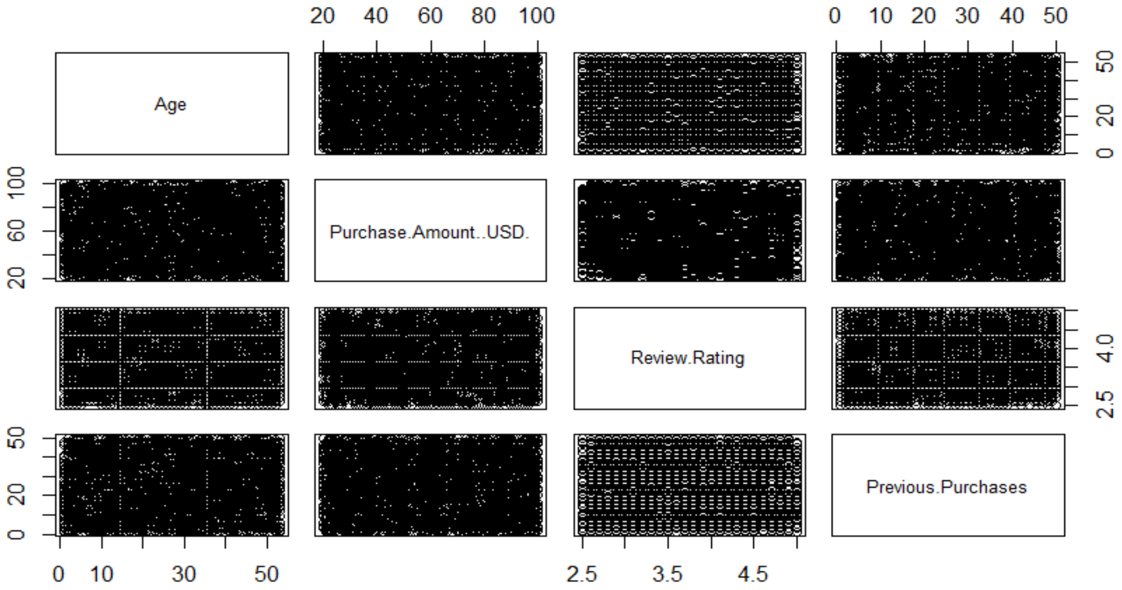


Figure 10. **Scatterplot Matrix**

Corplot was also used to visualize the correlation as shown below

```
corrplot(cor(my_data_numericals), type= "full", tl.col="navy", bg="white", co
l=NULL )
```

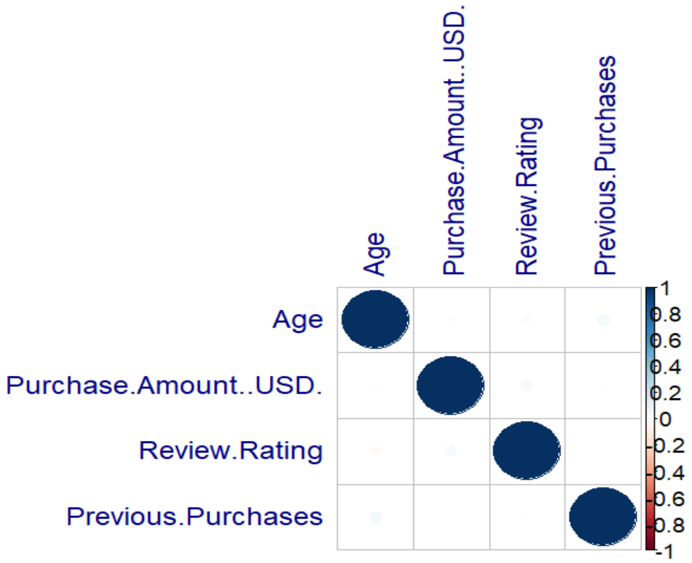


Figure 11. **Corplot**

2.1.1 CORRELATION BETWEEN PURCHASE AMOUNT (USD) AND REVIEW RATING

I created a scatter plot to visualize the relationship between Purchase Amount and Review Rating. While there's a slight positive trend, it's not a strong one. This suggests a weak positive correlation between the two variables.

```
> plot(my_data$Purchase.Amount..USD., my_data$Review.Rating, main ="Purchase
Amount and Review Rating", xlab="Purchase amount", ylab="Review Rating")
```

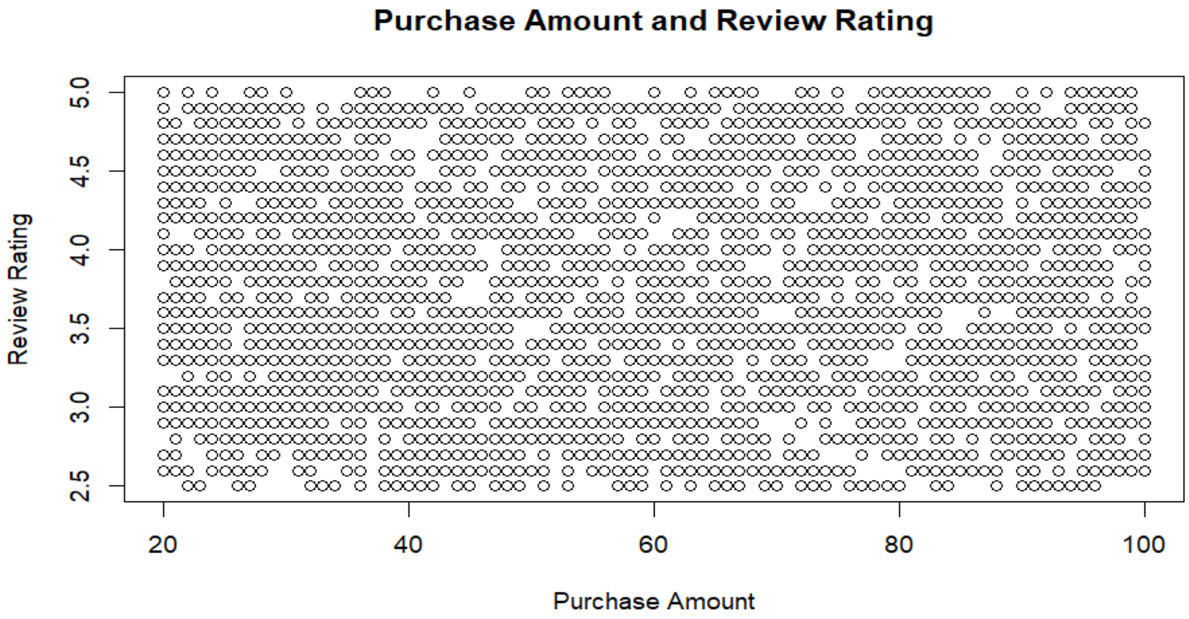


Figure 12

2.1.2 CORRELATION BETWEEN REVIEW RATING AND PREVIOUS PURCHASES

The scatter plot shows a weak positive link between Previous Purchases and Review Rating. As previous purchases increase, review ratings go up slightly, but the connection is not strong, and the points are very spread out.

```
> plot(my_data$Review.Rating, my_data$Previous.Purchases, main ="Review Ratin
g and Previous Purchase", xlab="Review Rating", ylab="Previous Purchase")
```

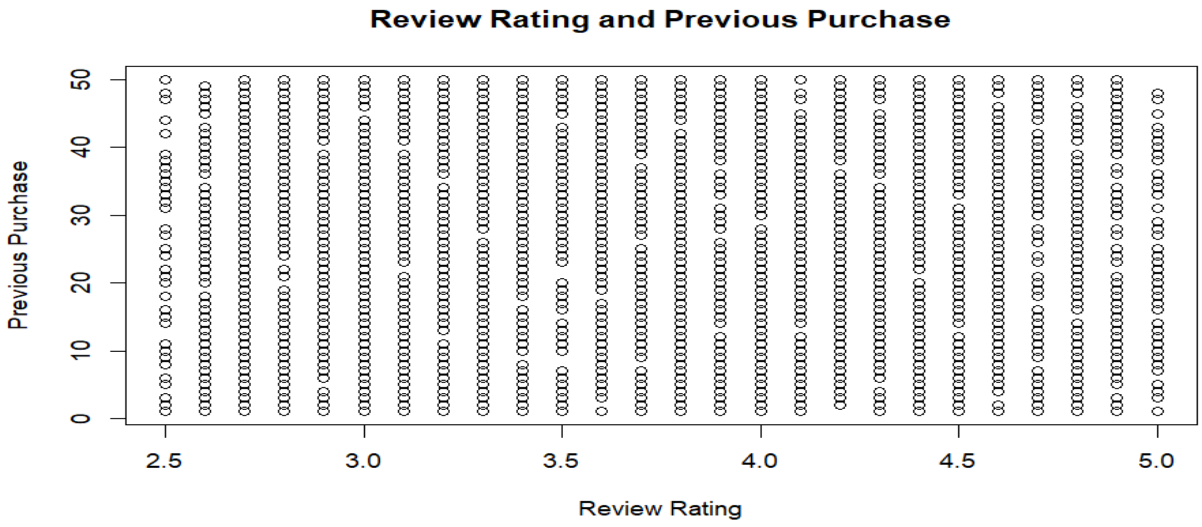


Figure 13

2.1.3 CORRELATION BETWEEN PREVIOUS PURCHASES AND PURCHASE AMOUNT (USD)

The scatter plot shows a weak positive relationship between Previous Purchases and Purchase Amount. As previous purchases increase, the purchase amount tends to rise slightly. However, the connection is not strong, and the data points are widely scattered.

```
> plot(my_data$Previous.Purchases, my_data$Purchase.Amount..USD., main ="Prev
ious Purchase and Purchase Amount", xlab="Previous Purchase", ylab="Purchase
Amount")
```

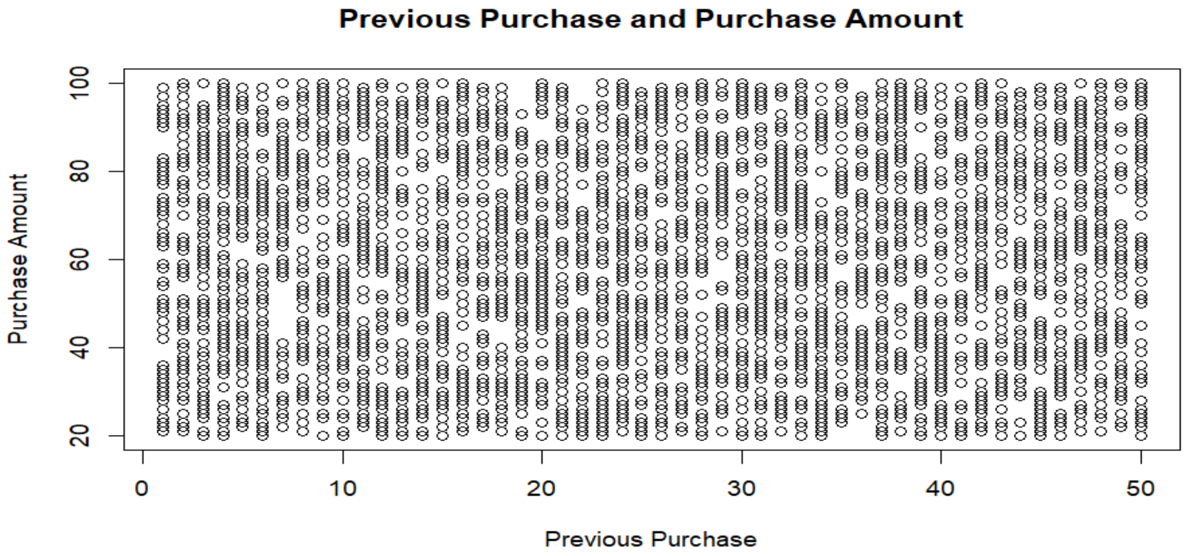
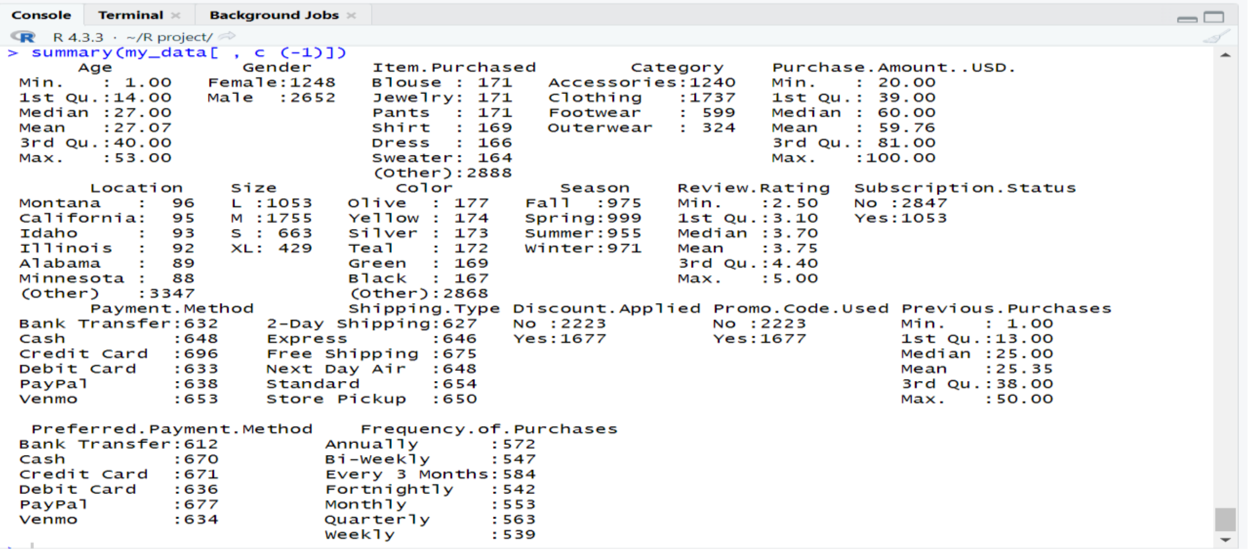


Figure 14

2.2 SUMMARY STATISTICS

The `summary () ` function in R was very useful for quickly analyzing the dataset. I excluded only the Customer.ID column and found some interesting patterns. It provided insights into customer categories, purchase frequency, sizes, gender, and more. These summaries would be helpful for visualizing trends and understanding customer segments better.

```
> summary(my_data[ ,c(-1)])
```

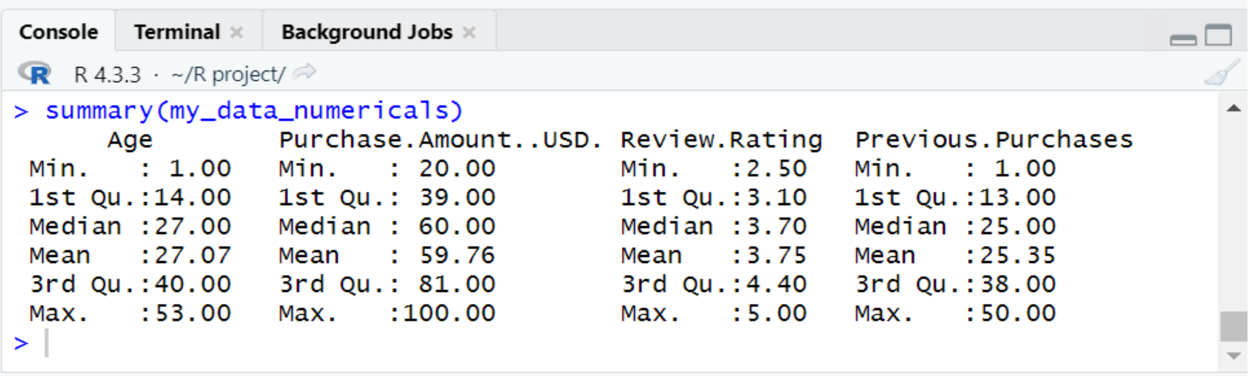


```
> summary(my_data[ , c (-1)])
      Age             Gender      Item.Purchased          Category      Purchase.Amount..USD.
 Min.   : 1.00     Female:1248    Blouse : 171     Accessories:1240     Min.   : 20.00
 1st Qu.:14.00     Male  :2652    Jewelry: 171     Clothing   :1737     1st Qu.: 39.00
 Median :27.00                    Pants  : 171     Footwear   : 599     Median : 60.00
 Mean   :27.07                    Shirt  : 169     Outerwear  : 324     Mean   : 59.76
 3rd Qu.:40.00                    Dress  : 166                          3rd Qu.: 81.00
 Max.   :53.00                    Sweater: 164                          Max.   :100.00
                                  (Other):2888
        Location     Size           Color         Season     Review.Rating    Subscription.Status
 Montana   :  96     L :1053    Olive  : 177     Fall  :975     Min.   :2.50     No :2847
 California:  95     M :1755    Yellow : 174     Spring:999     1st Qu.:3.10     Yes:1053
 Idaho     :  93     S : 663    Silver : 173     Summer:955     Median :3.70
 Illinois  :  92     XL: 429    Teal   : 172     Winter:971     Mean   :3.75
 Alabama   :  89                Green  : 169                    3rd Qu.:4.40
 Minnesota :  88                Black  : 167                    Max.   :5.00
 (Other)   :3347                (Other):2868
        Payment.Method          Shipping.Type  Discount.Applied   Promo.Code.Used    Previous.Purchases
 Bank Transfer:632      2-Day Shipping:627     No :2223           No :2223           Min.   : 1.00
 Cash         :648      Express       :646     Yes:1677           Yes:1677           1st Qu.:13.00
 Credit Card  :696      Free Shipping :675                                           Median :25.00
 Debit Card   :633      Next Day Air  :648                                           Mean   :25.35
 PayPal       :638      Standard      :654                                           3rd Qu.:38.00
 Venmo        :653      Store Pickup  :650                                           Max.   :50.00

  Preferred.Payment.Method       Frequency.of.Purchases
 Bank Transfer:612            Annually      :572
 Cash         :670            Bi-Weekly     :547
 Credit Card  :671            Every 3 Months:584
 Debit Card   :636            Fortnightly   :542
 PayPal       :677            Monthly       :553
 Venmo        :634            Quarterly     :563
                              Weekly        :539
```

Figure 15. Summary Statistic Data

Next, I calculated summary statistics for the numerical variables. These statistics provide key insights into the distribution and central tendency of numerical data.
Summary Statistics of numerical variables.

```
> summary(my_data_numericals)
```



```
> summary(my_data_numericals)
      Age        Purchase.Amount..USD.  Review.Rating    Previous.Purchases
 Min.   : 1.00   Min.   : 20.00         Min.   :2.50     Min.   : 1.00
 1st Qu.:14.00   1st Qu.: 39.00         1st Qu.:3.10     1st Qu.:13.00
 Median :27.00   Median : 60.00         Median :3.70     Median :25.00
 Mean   :27.07   Mean   : 59.76         Mean   :3.75     Mean   :25.35
 3rd Qu.:40.00   3rd Qu.: 81.00         3rd Qu.:4.40     3rd Qu.:38.00
 Max.   :53.00   Max.   :100.00         Max.   :5.00     Max.   :50.00
> |
```

Figure 16.

The summary displays the mean and median values, along with other statistical measures, for the numerical variables.

2.3 BAR PLOTS

In Figure 15 above, visualizing gender in a bar plot would be helpful as it can show which gender shops more.

I utilized the `table ()` function to extract the $Gender column, which I then used to generate the bar plot.

```
> table(my_data$Gender)

Female   Male
  1248   2652

> #Percentage Distribution:
> 1248/3900 #Female
[1] 0.32
> 2652/3900 #Male
[1] 0.68
>
> barplot(table(my_data$Gender), main = "Customer Shopping by Gender", xlab =
"Gender", ylab = "No. of Purchases")
```

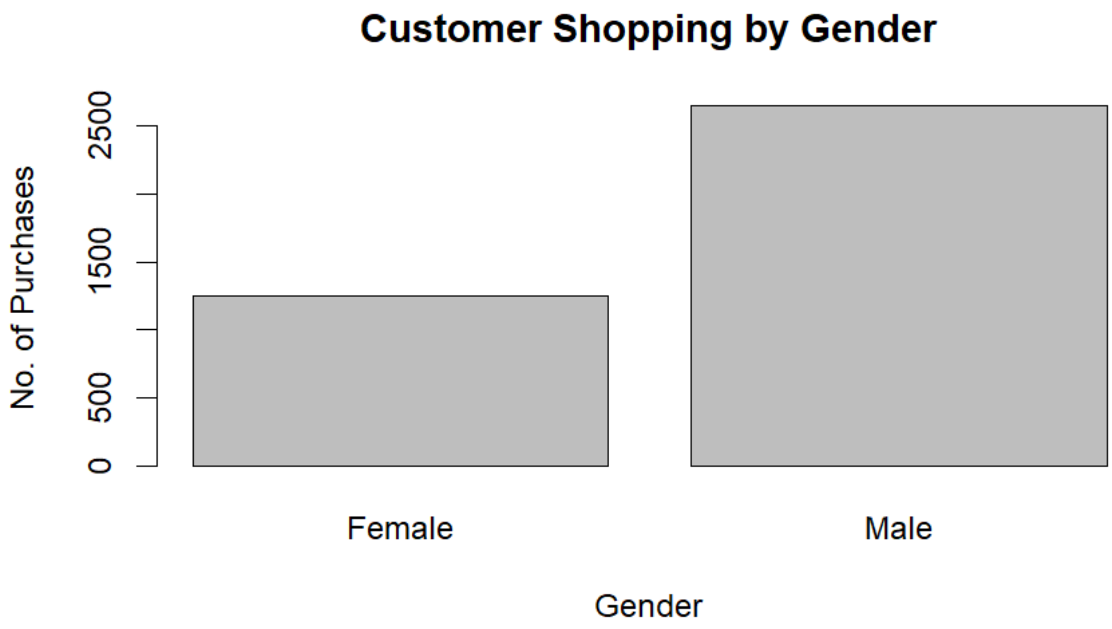


Figure 17. Barplot of Customers Shopping Base on Gender

(The bar chart shows the distribution of purchases by gender. Males seem to make significantly more purchases than females, suggesting a potential gender disparity in shopping behavior.)

I also created a bar plot for the product categories to provide more details about the types of products purchased.

```
> table(my_data$Category)

Accessories    Clothing    Footwear   Outerwear 
       1240        1737         599         324 
> 
> #Percentage Distribution:
> 1240/3900 #Accessories
[1] 0.3179487
> 1737/3900 #Clothing
[1] 0.4453846
> 599/3900 #Footwear
[1] 0.1535897
> 324/3900 #Outerwear
[1] 0.08307692
> 
> barplot(table(my_data$Category), main = "Products Shopped by Customers", xl
ab = "Category", ylab = "No. of Purchases")
```

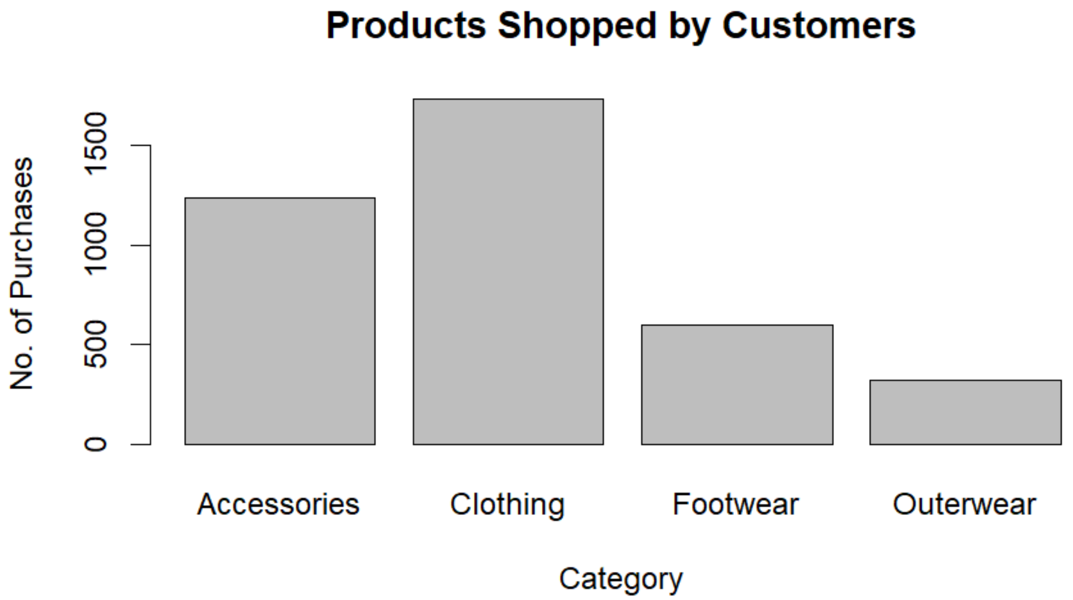


Figure 18. Barplot of What Customers Shopped For

(The plot above shows the distribution of purchases across various product categories. Clothing stands out as the most popular category, followed by accessories, footwear, and outerwear. This indicates that customers have a strong preference for clothing over the other categories.)

Another interesting visualization is the frequency of purchases, as customers are consistently making purchases.

```
> table(my_data$Frequency.of.Purchases)

      Annually      Bi-Weekly Every 3 Months    Fortnightly 
           572            547            584            542 
       Monthly      Quarterly         Weekly 
           553            563            539 
> 
```

```
> #Percentage Distribution:
> 572/3900 #Annually
[1] 0.1466667
> 553/3900 #Monthly
[1] 0.1417949
> 547/3900 #Bi-Weekly
[1] 0.1402564
> 563/3900 #Quarterly
[1] 0.144359
> 584/3900 #Every 3 Months
[1] 0.1497436
> 539/3900 #Weekly
[1] 0.1382051
> 542/3900 #Fortnightly
[1] 0.1389744
>
> barplot(table(my_data$Frequency.of.Purchases), main = "Shopping Base on Ret
urning Customers", xlab = "Frequency of Purchases", ylab = "No. of Purchases"
)
```

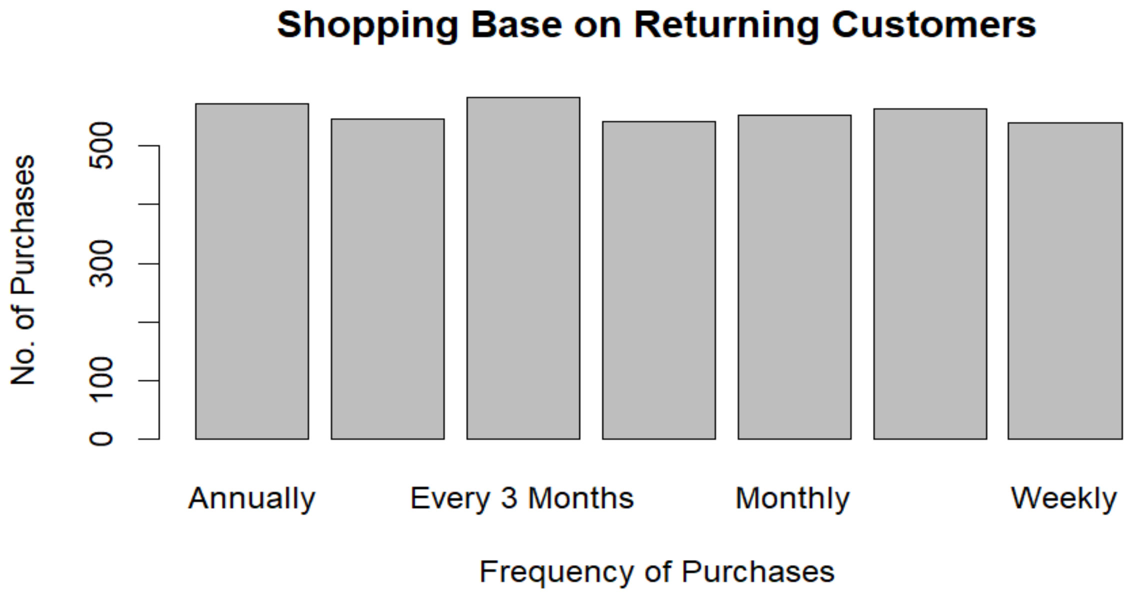


Figure 19. Barplot Showing Shopping Trends Base on Returning Customers

(The bar chart shows the distribution of purchases based on customer return frequency. Customers who shop every three months have the highest frequency, followed by annual and weekly shoppers. Monthly shoppers have the lowest frequency, suggesting that less frequent but consistent shoppers contribute more to overall sales.)

2.4 PIE CHART

Product sizes and the seasons when customers shop the most are important data points to explore and visualize. I used the `pie ()` function in R to create this visualization. It helps business owners understand shopping patterns and develop strategies accordingly.

Pie chart on Size

```
> table(my_data$Size)

   L    M    S   XL 
1053 1755  663  429 
> 
> #Percentage Distribution:
> 1053/3900 #L
[1] 0.27
> 1755/3900 #M
[1] 0.45
> 663/3900 #S
[1] 0.17
> 429/3900 #XL
[1] 0.11
> 
> pie(table(my_data$Size), main = "Customer Size")
```

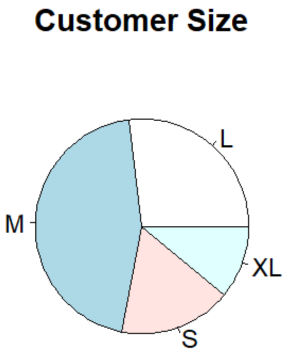


Figure 20. Pie Chart Showing Sizes Purchased the Most

(The pie chart shows that most customers prefer medium-sized clothing, followed by large and small sizes. Extra-large sizes are the least popular.)

Pie chart on Season

```
> table(my_data$Season)

  Fall Spring Summer Winter 
   975    999    955    971 
> 
> #Percentage Distribution:
> 975/3900 #Fall
[1] 0.25
> 999/3900 #Spring
[1] 0.2561538
> 955/3900 #Summer
[1] 0.2448718
> 971/3900 #Winter
```

```
[1] 0.2489744
>
> pie(table(my_data$Season), main = "Shopping Base on Season")
```

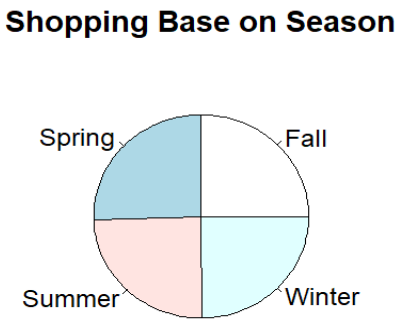


Figure 21. Pie Chart Showing Season Customers Shops the Most

(The pie chart shows that Fall and Winter are the most popular shopping seasons, followed by Spring and Summer, suggesting customers shop more during colder seasons.)

2.5 BOXPLOT

To identify outliers and inconsistencies between numerical and categorical variables, I used boxplots to examine discrepancies and explore the causes of the outliers.

The boxplots below display purchase amounts by customer gender, revealing that spending varies based on the shopper's gender.

```
my_data%>%
+   ggplot(aes(Gender, Purchase.Amount..USD.)) +
+   geom_boxplot()+
+   theme_bw()+
+   labs(x= "Gender", y="Purchase Amount")
```

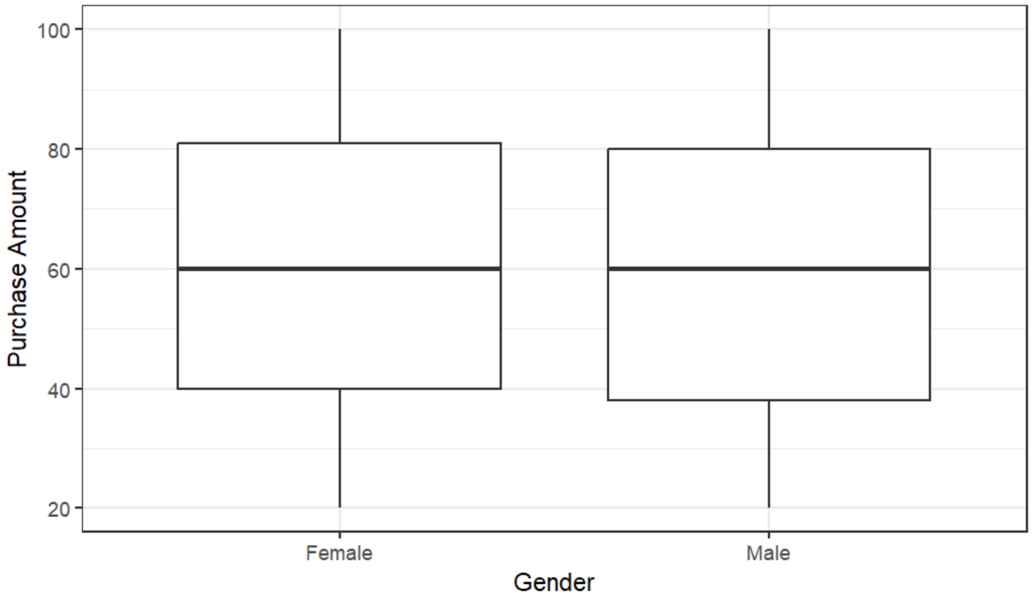


Figure 22

(The boxplot shows purchase amounts for females and males. Both have a similar median of about $60, but males have a wider range with some very high and very low purchases. Females have a more consistent range with fewer extreme values.)

Another variable that showed variation when plotted against customer gender is previous purchases.

```
> my_data%>%
+ ggplot(aes(Gender, Previous.Purchases)) +
+ geom_boxplot()+
+ theme_bw()+
+ labs(x= "Gender", y="Previous Purchases")
```

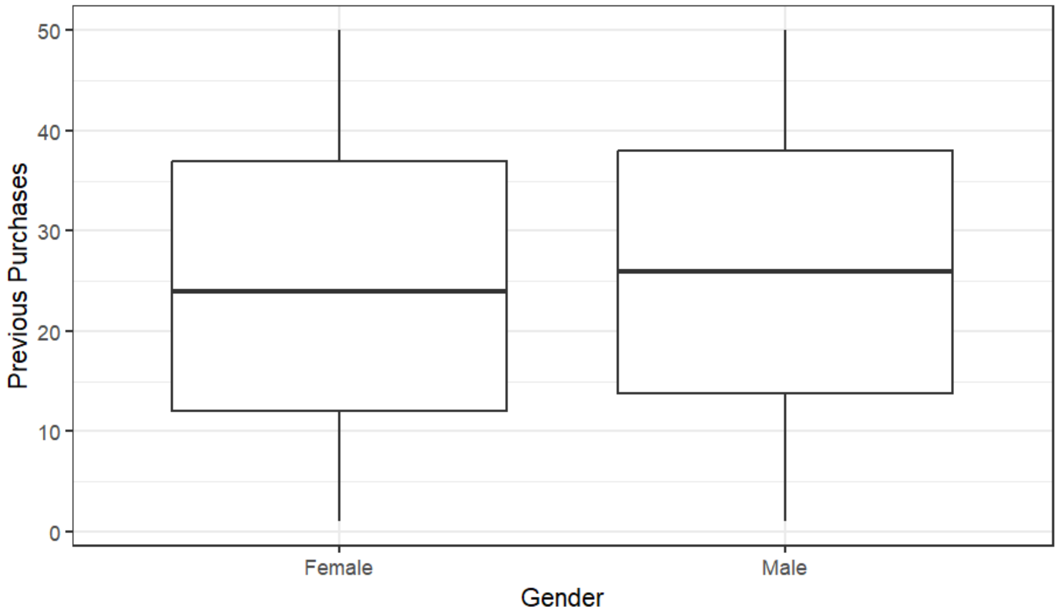


Figure 23

(The boxplot shows that both genders have a similar number of previous purchases on average. However, males have a wider range of purchases, with some extreme values on both ends. In contrast, females have a more consistent range with fewer outliers. This indicates that while both genders have a similar average, males show more variation in their purchasing behavior than females.)

2.6 CLOSER LOOK AT PURCHASE AMOUNT

I noticed a slight positive trend between Purchase Amount and Review Rating, but the relationship isn't strong. The number of previous purchases doesn't seem to significantly impact either variable. This suggests that while higher-priced items may slightly correlate with higher ratings, the number of previous purchases doesn't strongly influence this connection.

The bubble chart shows the relationship between Purchase Amount, Review Rating, and the number of Previous Purchases, with the bubble size representing the number of previous purchases.

```
> my_data %>%
+   ggplot(aes(Review.Rating, Purchase.Amount..USD.)) +
+   geom_point(aes(size= Previous.Purchases))+
+   coord_flip()+
+   theme_classic()
```

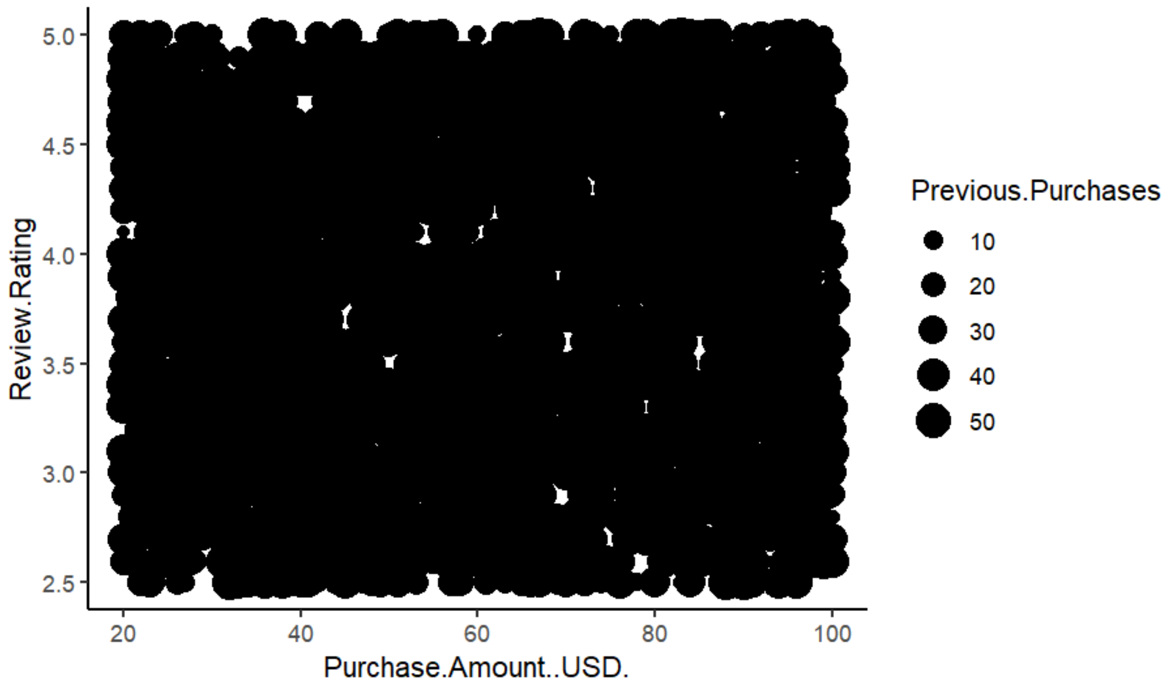


Figure 24. The Influence of Review Rating on Customer Behavior

From another notable analysis I did, there seems to be no notable difference in spending between states. The dot plot shows the distribution of purchase amounts across different states. The dots are evenly spread across the purchase amount range in each state, indicating that customers from various states generally spend similar amounts on average.

```
> my_data %>%
+   ggplot(aes(Location, Purchase.Amount..USD.)) +
+   geom_point()+
+   coord_flip()+
+   theme_gray()
```

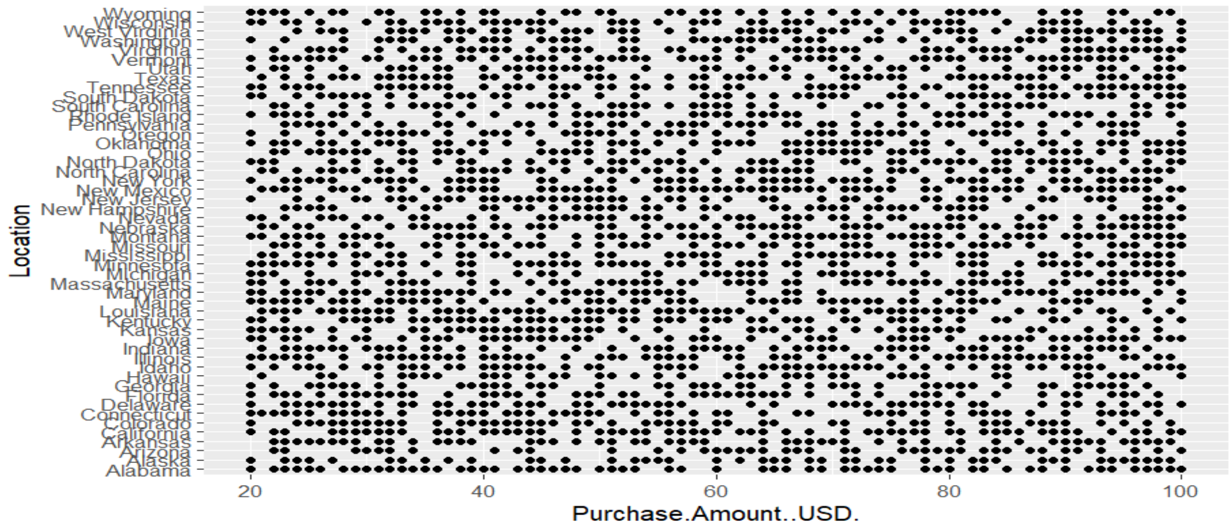


Figure 25 Customer Demographics on Purchase Patterns

Next, let's examine the product categories and identify which one was purchased the most. I created a density chart to compare the four categories.

```
> my_data %>%
+ ggplot(aes(Purchase.Amount..USD.)) +
+ geom_density()+
+ facet_wrap(~Category)
> theme_bw()
```

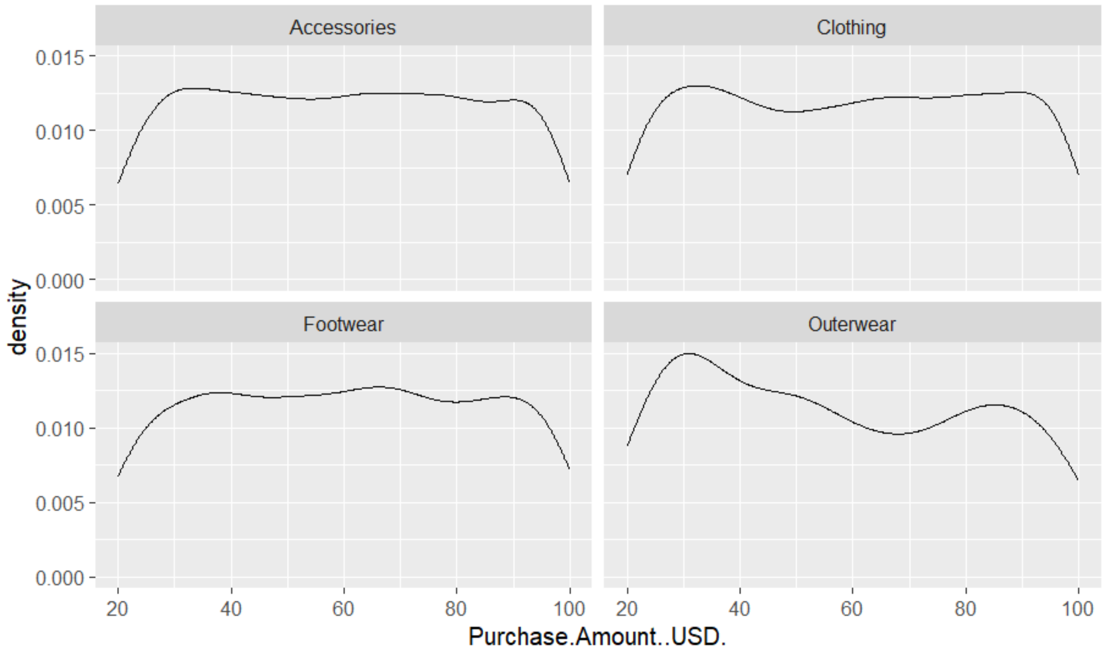


Figure 26. Purchase Per Category

(Accessories and Footwear have a wide range with no clear peak, while Clothing and Outerwear show lower-priced items, indicating a right-skewed distribution.)

2.7 HISTOGRAM AND DENSITY PLOT

I also analyzed the purchase behavior of all orders by customers, as this helps vendors track fast-moving products and develop effective strategies for better understanding their clients.

I plotted a histogram to show the number of orders placed at each purchase amount. The distribution appears relatively uniform, with no significant peaks or valleys. This suggests that orders are spread across a wide range of purchase amounts, with no specific amount being notably more or less popular.

```
my_data %>%
+     ggplot(aes(Purchase.Amount..USD.))+
+     geom_histogram(stat = "count")+
+     labs(x= "Purchase Amount",
+          y= "No. of Orders")+
+     theme_bw()
```

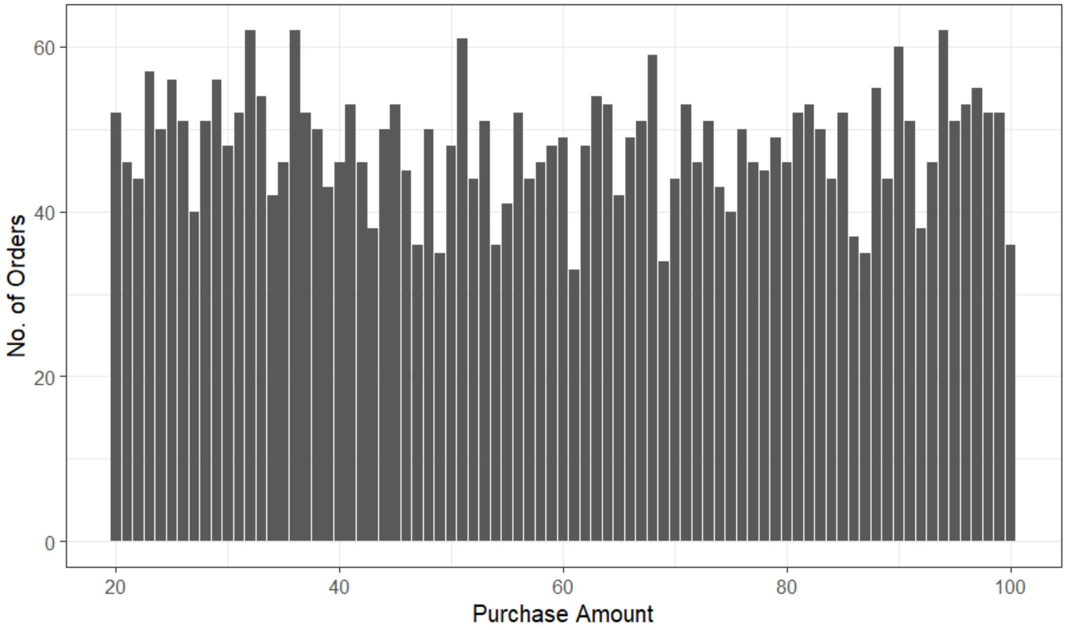


Figure 27. Purchase of Orders

This can also be represented in a density plot across various product categories. The results below highlight the items that customers most commonly purchase.

```
>   my_data %>%
+     ggplot(aes(Purchase.Amount..USD.))+
+     geom_density(colour = "navyblue", fill = "seagreen")+
+     facet_wrap(~Item.Purchased)+
+     theme_classic()
```

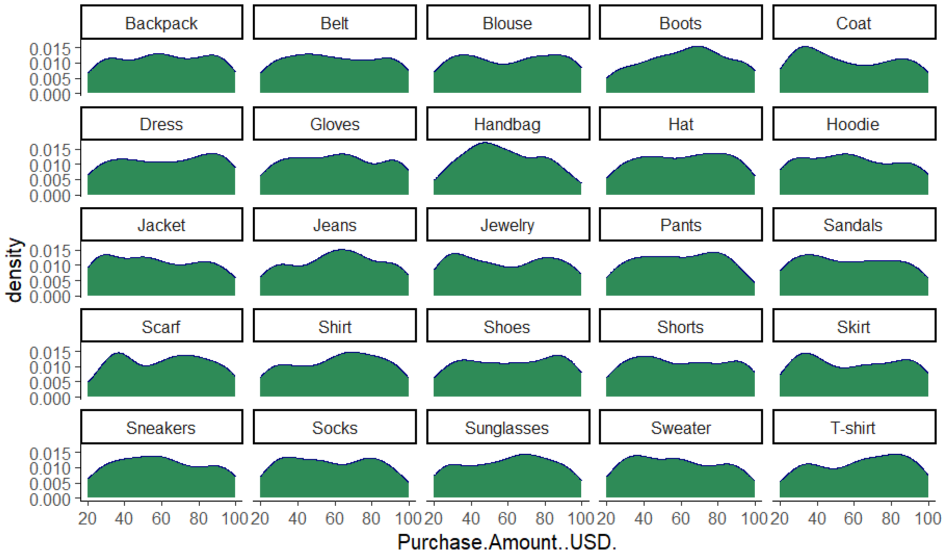


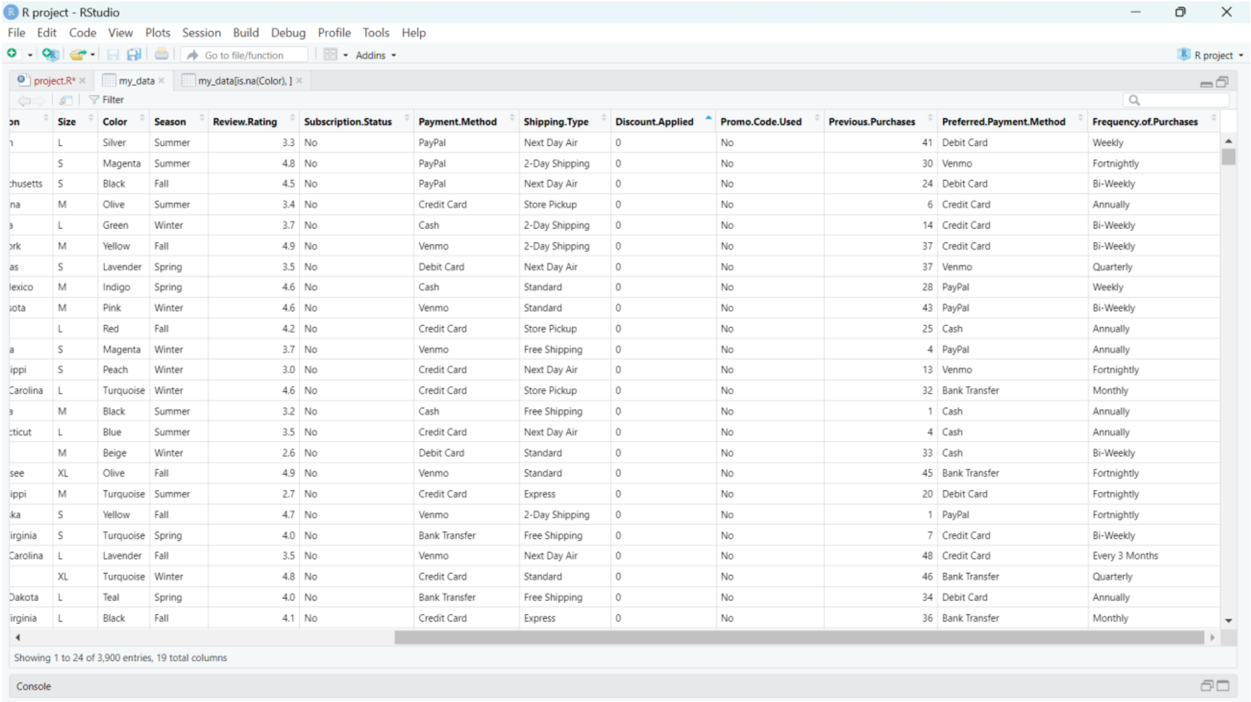
Figure 28. Density Plot Of Purchases Across Different Items

2.8 DATA TRANSFORMATION

I came across an interesting variable that I wanted to visualize, but it was categorical. To proceed, I converted it into a numerical variable. I aimed to visualize the distribution of purchase amounts across different discount levels and explore the relationship between purchase amounts and the discounts applied.

To convert it into a numerical variable, here's how I made the changes.

```
> my_data <- my_data %>%
+   mutate(Discount.Applied = ifelse(Discount.Applied == "Yes", 1, 0))
```



on	Size	Color	Season	Review.Rating	Subscription.Status	Payment.Method	Shipping.Type	Discount.Applied	Promo.Code.Used	Previous.Purchases	Preferred.Payment.Method	Frequency.of.Purchases
n	L	Silver	Summer	3.3	No	PayPal	Next Day Air	0	No	41	Debit Card	Weekly
	S	Magenta	Summer	4.8	No	PayPal	2-Day Shipping	0	No	30	Venmo	Fortnightly
chusetts	S	Black	Fall	4.5	No	PayPal	Next Day Air	0	No	24	Debit Card	Bi-Weekly
na	M	Olive	Summer	3.4	No	Credit Card	Store Pickup	0	No	6	Credit Card	Annually
a	L	Green	Winter	3.7	No	Cash	2-Day Shipping	0	No	14	Credit Card	Bi-Weekly
ork	M	Yellow	Fall	4.9	No	Venmo	2-Day Shipping	0	No	37	Credit Card	Bi-Weekly
as	S	Lavender	Spring	3.5	No	Debit Card	Next Day Air	0	No	37	Venmo	Quarterly
lexico	M	Indigo	Spring	4.6	No	Cash	Standard	0	No	28	PayPal	Weekly
sota	M	Pink	Winter	4.6	No	Venmo	Standard	0	No	43	PayPal	Bi-Weekly
	L	Red	Fall	4.2	No	Credit Card	Store Pickup	0	No	25	Cash	Annually
a	S	Magenta	Winter	3.7	No	Venmo	Free Shipping	0	No	4	PayPal	Annually
ippi	S	Peach	Winter	3.0	No	Credit Card	Next Day Air	0	No	13	Venmo	Fortnightly
Carolina	L	Turquoise	Winter	4.6	No	Credit Card	Store Pickup	0	No	32	Bank Transfer	Monthly
a	M	Black	Summer	3.2	No	Cash	Free Shipping	0	No	1	Cash	Annually
cticut	L	Blue	Summer	3.5	No	Credit Card	Next Day Air	0	No	4	Cash	Annually
	M	Beige	Winter	2.6	No	Debit Card	Standard	0	No	33	Cash	Bi-Weekly
see	XL	Olive	Fall	4.9	No	Venmo	Standard	0	No	45	Bank Transfer	Fortnightly
ippi	M	Turquoise	Summer	2.7	No	Credit Card	Express	0	No	20	Debit Card	Fortnightly
ka	S	Yellow	Fall	4.7	No	Venmo	2-Day Shipping	0	No	1	PayPal	Fortnightly
irginia	S	Turquoise	Spring	4.0	No	Bank Transfer	Free Shipping	0	No	7	Credit Card	Bi-Weekly
Carolina	L	Lavender	Fall	3.5	No	Venmo	Next Day Air	0	No	48	Credit Card	Every 3 Months
	XL	Turquoise	Winter	4.8	No	Credit Card	Standard	0	No	46	Bank Transfer	Quarterly
Dakota	L	Teal	Spring	4.0	No	Bank Transfer	Free Shipping	0	No	34	Debit Card	Annually
irginia	L	Black	Fall	4.1	No	Credit Card	Express	0	No	36	Bank Transfer	Monthly

Showing 1 to 24 of 3,900 entries, 19 total columns

Figure 29

2.8.1 HEAT MAP

Having successfully transformed the Discount Applied data into a numeric format, I will now analyze it alongside Purchase Amount to examine the density of the variables using the stat_density function.

```
>   my_data %>%
+     ggplot(aes(Discount.Applied, Purchase.Amount..USD.))+
+     stat_density_2d(geom = "tile", contour = FALSE, aes(fill=..density..))+
+     scale_fill_gradientn(colours = rainbow(5))+
+     labs(x="Discount Applied", y="Purchase Amount")+
+     theme_classic()
```

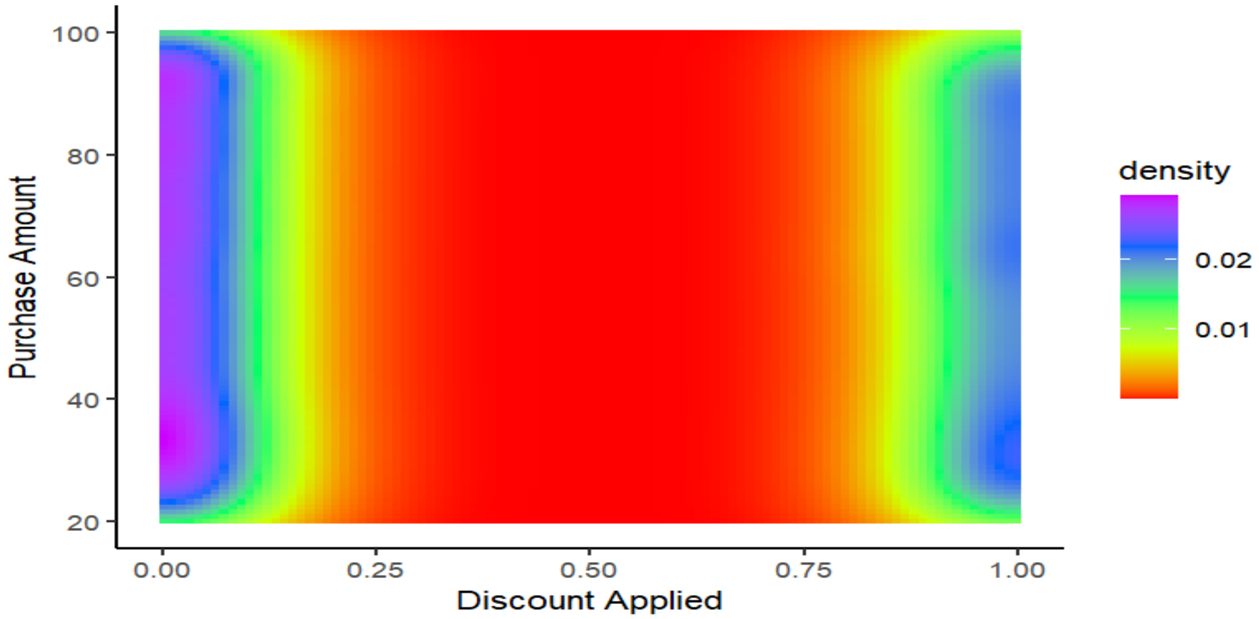


Figure 30. Purchases Over Discount

(The heatmap shows the distribution of purchase amounts across discount levels, with color intensity indicating density. Purchases are most frequent at lower amounts with higher discounts, while higher amounts see fewer purchases. This suggests that larger discounts encourage more purchases, particularly for lower-priced items.)

2.9 CLUSTER ANALYSIS

Analyzing the four numerical variables on their own didn’t provide enough insights, so I decided to combine them with other factors, such as customer location or product categories, for a deeper understanding. Variables like purchase amount or age should be more meaningful when analyzed alongside these categorical factors. However, since categorical variables cannot be directly used for cluster analysis, I first converted them into numerical values.

I then extracted the relevant categorical variables for clustering using the following command.

```
> Important_Categoricals <- data.frame(my_data[ , c(5, 7, 10)])
```

I already installed the package, “fastDummies” as I will need to convert the variables into numerics.

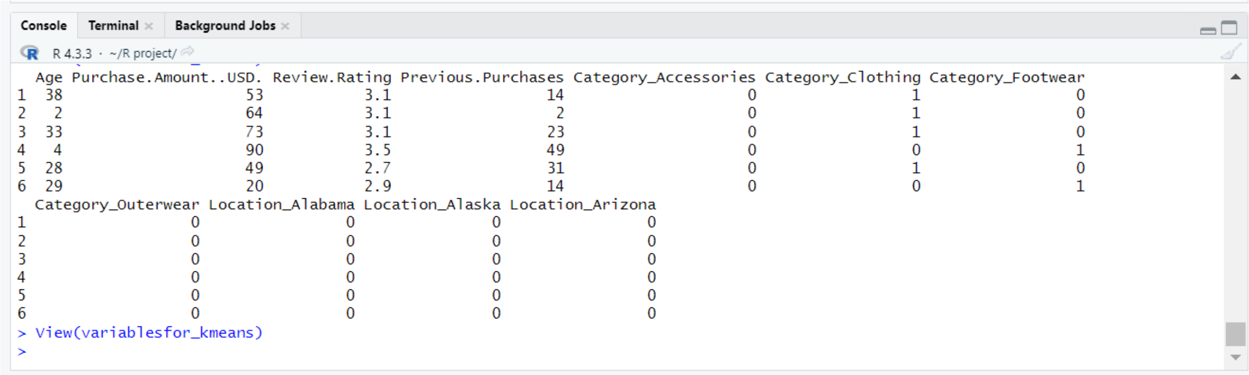
```
> library(fastDummies)
```

I converted the important_Categoricals data frame into numerical variables using the following command.

```
> Important_Categoricals <- dummy_cols(Important_Categoricals, remove_most_fr
equent_dummy = FALSE)
> View(Important_Categoricals)
>
```

I completed the dataset by merging the numerical variables with the transformed important categorical variables into one dataset for cluster analysis.

```
> variablesfor_kmeans <- cbind(my_data_numericals, Important_Categoricals[ ,
4:10])
> head(variablesfor_kmeans)
```



```
  Age Purchase.Amount..USD. Review.Rating Previous.Purchases Category_Accessories Category_Clothing Category_Footwear
1  38                    53           3.1                 14                    0                 1                 0
2   2                    64           3.1                  2                    0                 1                 0
3  33                    73           3.1                 23                    0                 1                 0
4   4                    90           3.5                 49                    0                 0                 1
5  28                    49           2.7                 31                    0                 1                 0
6  29                    20           2.9                 14                    0                 0                 1
  Category_Outerwear Location_Alabama Location_Alaska Location_Arizona
1                  0                0               0                0
2                  0                0               0                0
3                  0                0               0                0
4                  0                0               0                0
5                  0                0               0                0
6                  0                0               0                0
> View(variablesfor_kmeans)
>
```

Figure 31

I also verified that the clustered data maintained its data.frame class.

```
> class(variablesfor_kmeans)
[1] "data. frame"
```

The final step before performing cluster analysis is determining the optimal number of clusters. I used the elbow method for this purpose. To proceed, I installed the “factoextra” package and utilized it to identify the appropriate number of clusters.

```
> library(factoextra)
> fviz_nbclust(variablesfor_kmeans, kmeans, method = "wss" )+
+   labs(subtitle = "Elbow Method")
```

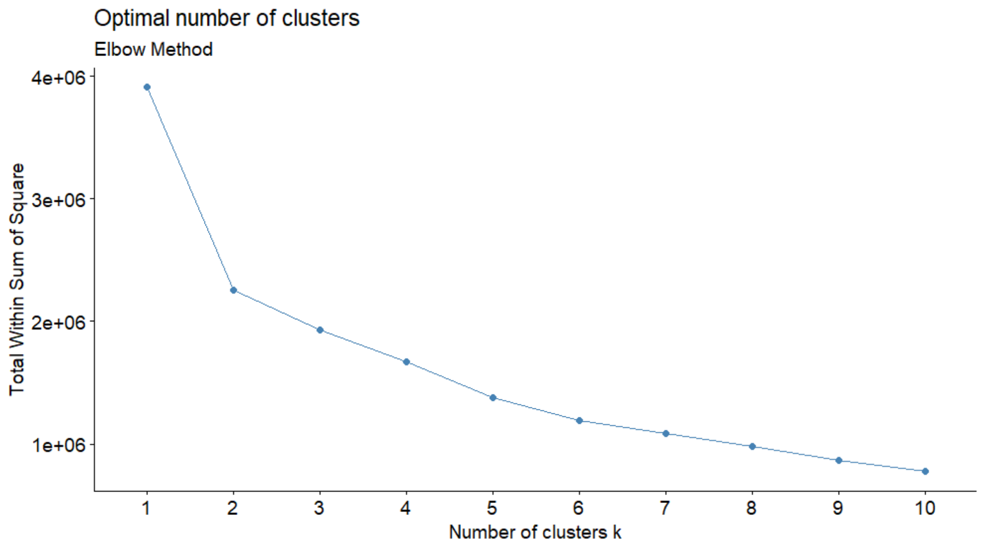
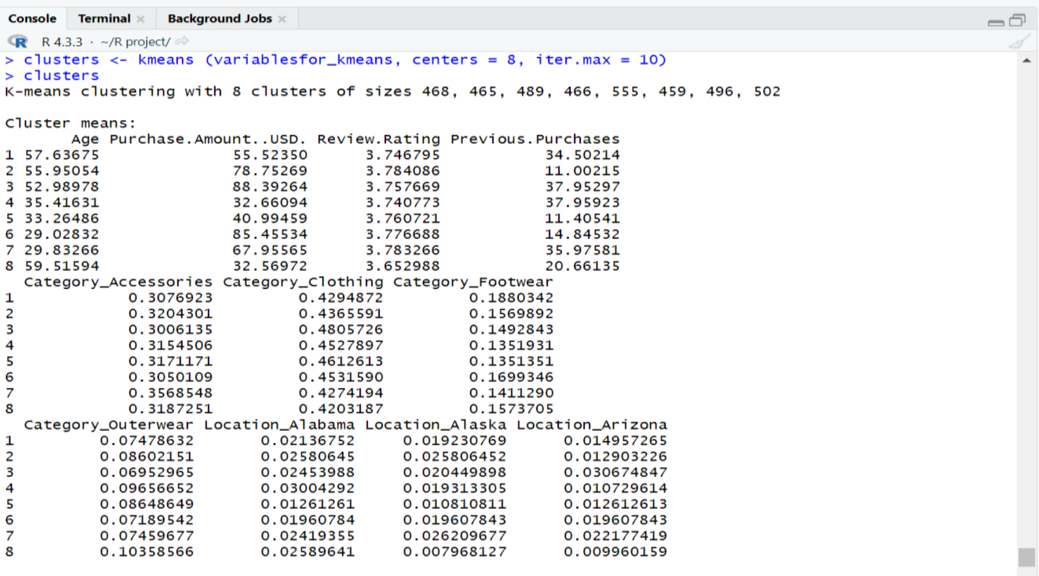


Figure 32

(The elbow plot shows how the total within-cluster sum of squares (WSS) decreases as the number of clusters increases. The decline slows after a certain point, creating an "elbow" shape. In this case, the elbow is around 3 clusters, suggesting this as the optimal number. Adding more clusters beyond this point provides minimal improvement in reducing WSS.)

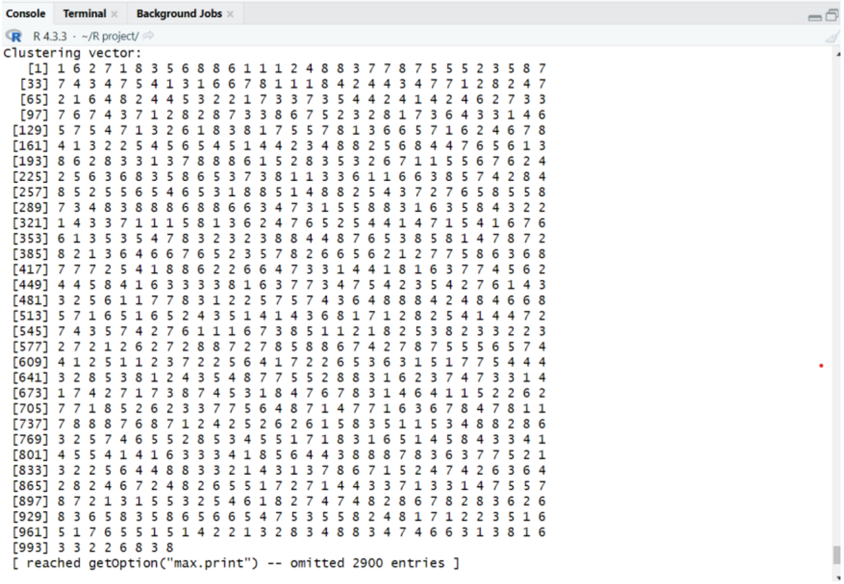
I executed the following command to perform the cluster analysis.

```
> clusters <- kmeans (variablesfor_kmeans, centers = 8, iter.max = 10)
> clusters
```



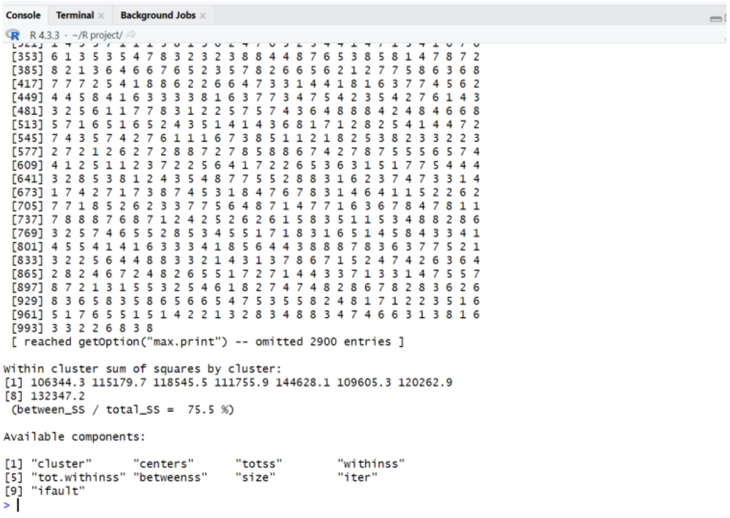
```
> clusters <- kmeans (variablesfor_kmeans, centers = 8, iter.max = 10)
> clusters
K-means clustering with 8 clusters of sizes 468, 465, 489, 466, 555, 459, 496, 502

Cluster means:
       Age Purchase.Amount..USD. Review.Rating Previous.Purchases
1 57.63675              55.52350      3.746795           34.50214
2 55.95054              78.75269      3.784086           11.00215
3 52.98978              88.39264      3.757669           37.95297
4 35.41631              32.66094      3.740773           37.95923
5 33.26486              40.99459      3.760721           11.40541
6 29.02832              85.45534      3.776688           14.84532
7 29.83266              67.95565      3.783266           35.97581
8 59.51594              32.56972      3.652988           20.66135
  Category_Accessories Category_Clothing Category_Footwear
1            0.3076923         0.4294872         0.1880342
2            0.3204301         0.4365591         0.1569892
3            0.3006135         0.4805726         0.1492843
4            0.3154506         0.4527897         0.1351931
5            0.3171171         0.4612613         0.1351351
6            0.3050109         0.4531590         0.1699346
7            0.3568548         0.4274194         0.1411290
8            0.3187251         0.4203187         0.1573705
  Category_Outerwear Location_Alabama Location_Alaska Location_Arizona
1         0.07478632       0.02136752     0.019230769      0.014957265
2         0.08602151       0.02580645     0.025806452      0.012903226
3         0.06952965       0.02453988     0.020449898      0.030674847
4         0.09656652       0.03004292     0.019313305      0.010729614
5         0.08648649       0.01261261     0.010810811      0.012612613
6         0.07189542       0.01960784     0.019607843      0.019607843
7         0.07459677       0.02419355     0.026209677      0.022177419
8         0.10358566       0.02589641     0.007968127      0.009960159
```

Figure 33



```
Clustering vector:
   [1] 1 6 2 7 1 8 3 5 6 8 8 6 1 1 1 2 4 8 8 3 7 7 8 7 5 5 5 2 3 5 8 7
  [33] 7 4 3 4 7 5 4 1 3 1 6 6 7 8 1 1 1 8 4 2 4 4 3 4 7 7 1 2 8 2 4 7
  [65] 2 1 6 4 8 2 4 4 5 3 2 2 1 7 3 3 7 3 5 4 4 2 4 1 4 2 4 6 2 7 3 3
  [97] 7 6 7 4 3 7 1 2 8 2 8 7 3 3 8 6 7 5 2 3 2 8 1 7 3 6 4 3 3 1 4 6
 [129] 5 7 5 4 7 1 3 2 6 1 8 3 8 1 7 5 5 7 8 1 3 6 6 5 7 1 6 2 4 6 7 8
 [161] 4 1 3 2 2 5 4 5 6 5 4 5 1 4 4 2 3 4 8 8 2 5 6 8 4 4 7 6 5 6 1 3
 [193] 8 6 2 8 3 3 1 3 7 8 8 8 6 1 5 2 8 3 5 3 2 6 7 1 1 5 5 6 7 6 2 4
 [225] 2 5 6 3 6 8 3 5 8 6 5 3 7 3 8 1 1 3 3 6 1 1 6 6 3 8 5 7 4 2 8 4
 [257] 8 5 2 5 5 6 5 4 6 5 3 1 8 8 5 1 4 8 8 2 5 4 3 7 2 7 6 5 8 5 5 8
 [289] 7 3 4 8 3 8 8 8 6 8 8 6 6 3 4 7 3 1 5 5 8 8 3 1 6 3 5 8 4 3 2 2
 [321] 1 4 3 3 7 1 1 1 5 8 1 3 6 2 4 7 6 5 2 5 4 4 1 4 7 1 5 4 1 6 7 6
 [353] 6 1 3 5 3 5 4 7 8 3 2 3 2 3 8 8 4 4 8 7 6 5 3 8 5 8 1 4 7 8 7 2
 [385] 8 2 1 3 6 4 6 6 7 6 5 2 3 5 7 8 2 6 6 5 6 2 1 2 7 7 5 8 6 3 6 8
 [417] 7 7 7 2 5 4 1 8 8 6 2 2 6 6 4 7 3 3 1 4 4 1 8 1 6 3 7 7 4 5 6 2
 [449] 4 4 5 8 4 1 6 3 3 3 3 8 1 6 3 7 7 3 4 7 5 4 2 3 5 4 2 7 6 1 4 3
 [481] 3 2 5 6 1 1 7 7 8 3 1 2 2 5 7 5 7 4 3 6 4 8 8 8 4 2 4 8 4 6 6 8
 [513] 5 7 1 6 5 1 6 5 2 4 3 5 1 4 1 4 3 6 8 1 7 1 2 8 2 5 4 1 4 4 7 2
 [545] 7 4 3 5 7 4 2 7 6 1 1 1 6 7 3 8 5 1 1 2 1 8 2 5 3 8 2 3 3 2 2 3
 [577] 2 7 2 1 2 6 2 7 2 8 8 7 2 7 8 5 8 8 6 7 4 2 7 8 7 5 5 5 6 5 7 4
 [609] 4 1 2 5 1 1 2 3 7 2 2 5 6 4 1 7 2 2 6 5 3 6 3 1 5 1 7 7 5 4 4 4
 [641] 3 2 8 5 3 8 1 2 4 3 5 4 8 7 7 5 5 2 8 8 3 1 6 2 3 7 4 7 3 3 1 4
 [673] 1 7 4 2 7 1 7 3 8 7 4 5 3 1 8 4 7 6 7 8 3 1 4 6 4 1 1 5 2 2 6 2
 [705] 7 7 1 8 5 2 6 2 3 3 7 7 5 6 4 8 7 1 4 7 7 1 6 3 6 7 8 4 7 8 1 1
 [737] 7 8 8 8 7 6 8 7 1 2 4 2 5 2 6 2 6 1 5 8 3 5 1 1 5 3 4 8 8 2 8 6
 [769] 3 2 5 7 4 6 5 5 2 8 5 3 4 5 5 1 7 1 8 3 1 6 5 1 4 5 8 4 3 3 4 1
 [801] 4 5 5 4 1 4 1 6 3 3 3 4 1 8 5 6 4 4 3 8 8 8 7 8 3 6 3 7 7 5 2 1
 [833] 3 2 2 5 6 4 4 8 8 3 3 2 1 4 3 1 3 7 8 6 7 1 5 2 4 7 4 2 6 3 6 4
 [865] 2 8 2 4 6 7 2 4 8 2 6 5 5 1 7 2 7 1 4 4 3 3 7 1 3 3 1 4 7 5 5 7
 [897] 8 7 2 1 3 1 5 5 3 2 5 4 6 1 8 2 7 4 7 4 8 2 8 6 7 8 2 8 3 6 2 6
 [929] 8 3 6 5 8 3 5 8 6 5 6 6 5 4 7 5 3 5 5 8 2 4 8 1 7 1 2 2 3 5 1 6
 [961] 5 1 7 6 5 5 1 5 1 4 2 2 1 3 2 8 3 4 8 8 3 4 7 4 6 6 3 1 3 8 1 6
 [993] 3 3 2 2 6 8 3 8
 [ reached getOption("max.print") -- omitted 2900 entries ]
```

Figure 34



```
 [353] 6 1 3 5 3 5 4 7 8 3 2 3 2 3 8 8 4 4 8 7 6 5 3 8 5 8 1 4 7 8 7 2
 [385] 8 2 1 3 6 4 6 6 7 6 5 2 3 5 7 8 2 6 6 5 6 2 1 2 7 7 5 8 6 3 6 8
 [417] 7 7 7 2 5 4 1 8 8 6 2 2 6 6 4 7 3 3 1 4 4 1 8 1 6 3 7 7 4 5 6 2
 [449] 4 4 5 8 4 1 6 3 3 3 3 8 1 6 3 7 7 3 4 7 5 4 2 3 5 4 2 7 6 1 4 3
 [481] 3 2 5 6 1 1 7 7 8 3 1 2 2 5 7 5 7 4 3 6 4 8 8 8 4 2 4 8 4 6 6 8
 [513] 5 7 1 6 5 1 6 5 2 4 3 5 1 4 1 4 3 6 8 1 7 1 2 8 2 5 4 1 4 4 7 2
 [545] 7 4 3 5 7 4 2 7 6 1 1 1 6 7 3 8 5 1 1 2 1 8 2 5 3 8 2 3 3 2 2 3
 [577] 2 7 2 1 2 6 2 7 2 8 8 7 2 7 8 5 8 8 6 7 4 2 7 8 7 5 5 5 6 5 7 4
 [609] 4 1 2 5 1 1 2 3 7 2 2 5 6 4 1 7 2 2 6 5 3 6 3 1 5 1 7 7 5 4 4 4
 [641] 3 2 8 5 3 8 1 2 4 3 5 4 8 7 7 5 5 2 8 8 3 1 6 2 3 7 4 7 3 3 1 4
 [673] 1 7 4 2 7 1 7 3 8 7 4 5 3 1 8 4 7 6 7 8 3 1 4 6 4 1 1 5 2 2 6 2
 [705] 7 7 1 8 5 2 6 2 3 3 7 7 5 6 4 8 7 1 4 7 7 1 6 3 6 7 8 4 7 8 1 1
 [737] 7 8 8 8 7 6 8 7 1 2 4 2 5 2 6 2 6 1 5 8 3 5 1 1 5 3 4 8 8 2 8 6
 [769] 3 2 5 7 4 6 5 5 2 8 5 3 4 5 5 1 7 1 8 3 1 6 5 1 4 5 8 4 3 3 4 1
 [801] 4 5 5 4 1 4 1 6 3 3 3 4 1 8 5 6 4 4 3 8 8 8 7 8 3 6 3 7 7 5 2 1
 [833] 3 2 2 5 6 4 4 8 8 3 3 2 1 4 3 1 3 7 8 6 7 1 5 2 4 7 4 2 6 3 6 4
 [865] 2 8 2 4 6 7 2 4 8 2 6 5 5 1 7 2 7 1 4 4 3 3 7 1 3 3 1 4 7 5 5 7
 [897] 8 7 2 1 3 1 5 5 3 2 5 4 6 1 8 2 7 4 7 4 8 2 8 6 7 8 2 8 3 6 2 6
 [929] 8 3 6 5 8 3 5 8 6 5 6 6 5 4 7 5 3 5 5 8 2 4 8 1 7 1 2 2 3 5 1 6
 [961] 5 1 7 6 5 5 1 5 1 4 2 2 1 3 2 8 3 4 8 8 3 4 7 4 6 6 3 1 3 8 1 6
 [993] 3 3 2 2 6 8 3 8
 [ reached getOption("max.print") -- omitted 2900 entries ]

Within cluster sum of squares by cluster:
[1] 106344.3 115179.7 118545.5 111755.9 144628.1 109605.3 120262.9
[8] 132347.2
 (between_SS / total_SS =  75.5 %)

Available components:

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

Figure 35