

R Project: Insights from Customer Shopping Trends

California State University, Los Angeles

CIS 5250 Visual Analytics

Dr. Shilpa Balan

Hsiu-Ping Lin & Povrotanak Cheatha

December 10, 2023

Contents

A.	Introduction	3
B.	Dataset Source/URL.....	5
C.	Data Description.....	5
D.	Data Cleaning.....	6
	1. Categorical Value Mapping	6
	2. Missing Value Imputation	8
	3. Data Splitting.....	9
E.	Analysis & Visualizations	10
	1. What is the proportion of Total Purchase Amount based on Gender?	10
	2. What are the key seasonal patterns in cumulative previous purchases across different product categories?	12
	3. Which colors are the most popular among customers, as indicated by the top 10 colors with the highest purchase amounts?	14
F.	Statistical Summary	16
	1. Analysis Variable: Review Rating.....	16
	2. Analysis Variable: Previous Purchase	17
G.	User-defined Function.....	18
	Function – highlight:.....	18
H.	Reference:	19

R Project: Insights from Customer Shopping Trends

A. Introduction

The project revolves around the exploration and analysis of the Customer Shopping Preferences Dataset, a rich resource encapsulating diverse features such as age, gender, purchase history, payment preferences, and more. In a retail landscape constantly shaped by evolving consumer behaviors, this dataset of 3900 records serves as a crucial foundation for data-driven decision-making. The motivation behind this project is rooted in the need for businesses to understand and adapt to customer preferences, optimizing products and refining marketing strategies. The dataset not only empowers businesses to enhance customer experience and satisfaction but also facilitates the identification of trends, enabling companies to stay flexible and responsive to the dynamic demands of the market.

Moving through the project stages, our data cleaning process involves essential steps such as Categorical Value Mapping, where we transform the "Category" column for better comprehension, Missing Value Imputation to address numerical column inconsistencies, and Data Splitting to separate "Age" and "Gender" for clearer analysis.

The subsequent phase of Analysis & Visualizations delves into key aspects of customer behavior. We examine the proportion of Total Purchase Amount based on Gender, uncover seasonal patterns in cumulative previous purchases across different product categories, and identify the most popular colors among customers through insightful visualizations. The study of seasonal patterns based on previous purchases is inspired by an article titled "Best Times To Go Shopping," which explains the difference between fashion seasons and calendar seasons. The new fashion season always starts two months earlier than the official change of the season. This mainly means that sellers mostly supply clothes earlier to customers ahead of time, as each season passes

in the blink of an eye, and they also prepare for customers to stock up. The article further mentions that to get the best deal on clothing, one should look for the best times to shop. Usually, the best time to shop is at the end of the season when there are many leftover clothes that sellers may need to throw away before the arrival of the new season. This presents a great opportunity to stock up on clothes for the following year (Fowler, 2022). The analysis of the most popular colors among customers is inspired by an article, namely, "The role of color in fashion." The article mentions the way color impacts marketing and customer behavior, as it is not only attractive in the head but also to the heart of customers, as colors should be used in different contexts. For example, black, red, and dark blue are mostly used in a business context. Yellow refers to sunshine, happiness, optimism, and vitality, which is cheerful and positive, and red refers to love and passion, but the meaning of colors changes according to countries. Thus, color is able to create different emotions, behaviors, as well as convey messages, in which customers would prefer different colors for their dressing. By understanding the most popular color is very crucial for marketing and sales of an entity for their strategic planning (Blaazer, 2022). Another analysis is further conducted, Statistical Summary of key variables, such as "Review Rating" and "Previous Purchase," provides a comprehensive understanding of the dataset's central tendencies, variability, and customer transaction patterns. The "Review Rating" analysis reveals a well-balanced distribution, while the "Previous Purchase" analysis emphasizes the diversity in customer transaction histories. To streamline and enhance our visualization process, we introduce a user-defined function called "highlight." This function proves particularly useful in creating bar charts for categorical columns with numerous labels, allowing us to isolate and emphasize the top categories efficiently.

In essence, our project aims to empower businesses with actionable insights derived from the Customer Shopping Preferences Dataset, guiding them towards more informed decision-making in the competitive and dynamic landscape of the retail industry (Banerjee, 2023).

B. Dataset Source/URL

<https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset>

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Customer	Age	Gender	Item Purchased	Category	Purchase Amount	Location	Size	Color	Season	Review Rating	Subscription	Shipping Method	Discount Received	Promotional Code	Previous Purchases	Payment Method	Frequency of Purchases
2	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
3	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
4	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
5	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day	Yes	Yes	49	PayPal	Weekly

C. Data Description

The dataset we used is called "Customer Shopping Trends Dataset," sourced from Kaggle, and was designed to offer a comprehensive view of customer behavior and preferences. It contains various crucial features for businesses aiming to enhance their insights into their customer base. This dataset contains 19 columns and 3900 rows, which include information such as customer age, gender, purchase amount, payment methods, frequency of purchases, and review ratings, presented in both numerical and categorical formats. Besides, the dataset also provides valuable insights into the types of items purchased, shopping frequency, shopping seasons, and interactions with promotional offers. As companies adapt to changing customer preferences, this dataset is crucial for creating effective marketing plans and satisfying customers (Banerjee, 2023).

No.	Field Name	Description
1	Customer ID	Unique identifier for each customer
2	Age	Age of the customer
3	Gender	Gender of the customer (Male/Female)

4	Item Purchased	The item purchased by the customer
5	Category	Category of the item purchased
6	Purchase Amount (USD)	The amount of the purchase in USD
7	Location	Location where the purchase was made
8	Size	Size of the purchased item
9	Color	Color of the purchased item
10	Season	Season during which the purchase was made
11	Review Rating	Rating given by the customer for the purchased item
12	Subscription Status	Indicates if the customer has a subscription (Yes/No)
13	Shipping Type	Type of shipping chosen by the customer
14	Discount Applied	Indicates if a discount was applied to the purchase (Yes/No)
15	Promo Code Used	Indicates if a promo code was used for the purchase (Yes/No)
16	Previous Purchases	The total count of transactions concluded by the customer at the store, excluding the ongoing transaction
17	Payment Method	Customer's most preferred payment method
18	Frequency of Purchases	Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly)

D. Data Cleaning

1. Categorical Value Mapping

R code

```
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated_precleaning.csv")
> data$Category <- ifelse(data$Category == "A", "Accessories",
+                         ifelse(data$Category == "C", "Clothing",
+                         ifelse(data$Category == "F", "Footwear",
+                         ifelse(data$Category == "O", "Outerwear", data$Category))))
```

Pre-cleaning

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
1	1	55-Male	Blouse	C	53	Kentucky	L
2	2	19-Male	Sweater	C	64	Maine	L
3	3	50-Male	Jeans	C	73	Massachusetts	S
4	4	21-Male	Sandals	F	90	Rhode Island	M
5	5	45-Male	Blouse	C	49	Oregon	M
6	6	46-Male	Sneakers	F	20	Wyoming	M
7	7	63-Male	Shirt	C	85	Montana	M
8	8	27-Male	Shorts	C	34	Louisiana	L
9	9	26-Male	Coat	O	97	West Virginia	L
10	10	57-Male	Handbag	A	31	Missouri	M

Post-cleaning

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
1	1	55-Male	Blouse	Clothing	53	Kentucky	L
2	2	19-Male	Sweater	Clothing	64	Maine	L
3	3	50-Male	Jeans	Clothing	73	Massachusetts	S
4	4	21-Male	Sandals	Footwear	90	Rhode Island	M
5	5	45-Male	Blouse	Clothing	49	Oregon	M
6	6	46-Male	Sneakers	Footwear	20	Wyoming	M
7	7	63-Male	Shirt	Clothing	85	Montana	M
8	8	27-Male	Shorts	Clothing	34	Louisiana	L
9	9	26-Male	Coat	Outerwear	97	West Virginia	L
10	10	57-Male	Handbag	Accessories	31	Missouri	M

Explanation

Our first step in data cleaning is Categorical Value Mapping. Upon reviewing the pre-cleaning screenshot, it becomes evident that the "Category" column exclusively employs capital English alphabet characters to denote respective categories. This representation may pose challenges for comprehension. Consequently, we employ the `ifelse` function in R Studio to transform this column. Specifically, we map "A" to "Accessories," "C" to "Clothing," "F" to "Footwear," and "O" to "Outerwear." This mapping enhances the clarity and interpretability of the data.

2. Missing Value Imputation

R code

```
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated_precleaning.csv")
> ave_amount <- round(mean(data$Purchase_Amount_USD, na.rm = TRUE))
> data$Purchase_Amount_USD[is.na(data$Purchase_Amount_USD)] <- ave_amount
```

Pre-cleaning

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
3834	3834	44-Female	Jacket	Outerwear	93	Arizona	L
3835	3835	43-Female	Hoodie	Clothing	NA	Tennessee	M
3836	3836	58-Female	Sandals	Footwear	58	Pennsylvania	L
3837	3837	62-Female	Skirt	Clothing	84	Alaska	M

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
3848	3848	60-Female	Sweater	Clothing	34	Delaware	S
3849	3849	22-Female	Jewelry	Accessories	NA	New Hampshire	M
3850	3850	46-Female	Hoodie	Clothing	68	Florida	S
3851	3851	27-Female	Jewelry	Accessories	74	Mississippi	L

Post-cleaning

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
3834	3834	44-Female	Jacket	Outerwear	93	Arizona	L
3835	3835	43-Female	Hoodie	Clothing	60	Tennessee	M
3836	3836	58-Female	Sandals	Footwear	58	Pennsylvania	L
3837	3837	62-Female	Skirt	Clothing	84	Alaska	M

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
3848	3848	60-Female	Sweater	Clothing	34	Delaware	S
3849	3849	22-Female	Jewelry	Accessories	60	New Hampshire	M
3850	3850	46-Female	Hoodie	Clothing	68	Florida	S
3851	3851	27-Female	Jewelry	Accessories	74	Mississippi	L

Explanation

Addressing missing values is a common challenge in data cleaning that requires resolution before conducting data analysis. Our dataset includes the numerical column 'Purchase Amount USD,' which exhibits instances of missing values, as evident in the pre-cleaning screenshot displaying 'NA' values. To manage this, we employed a common approach, filling the missing values with the column's average. In R Studio, we used the

mean function to calculate the average value, storing it in the variable 'ave_amount.' To conform to the dataset format, we utilized the round function to convert the average to an integer. Subsequently, the post-cleaning screenshot illustrates the successful replacement of all initial missing values with the rounded-up average value of 60.

3. Data Splitting

R code

```
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated_precleaning.csv")
> library(tidyr)
> data <- separate(data, Age_Gender, c("Age", "Gender"), sep="-")
```

Pre-cleaning

	Customer_ID	Age_Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
1	1	55-Male	Blouse	Clothing	53	Kentucky	L
2	2	19-Male	Sweater	Clothing	64	Maine	L
3	3	50-Male	Jeans	Clothing	73	Massachusetts	S

Post-cleaning

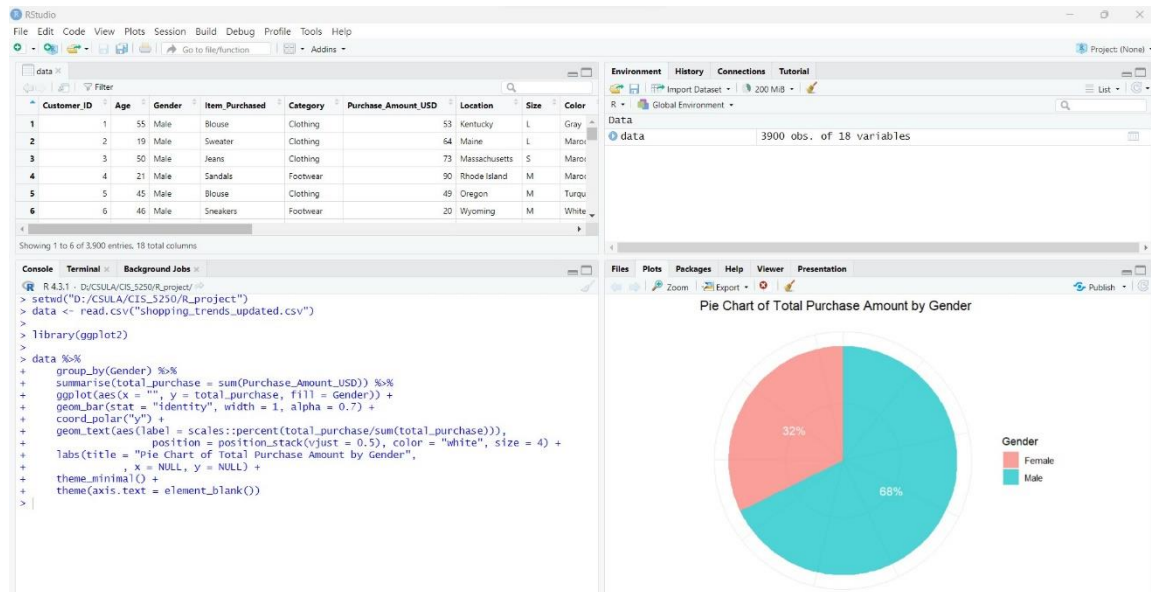
	Customer_ID	Age	Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size
1	1	55	Male	Blouse	Clothing	53	Kentucky	L
2	2	19	Male	Sweater	Clothing	64	Maine	L
3	3	50	Male	Jeans	Clothing	73	Massachusetts	S

Explanation

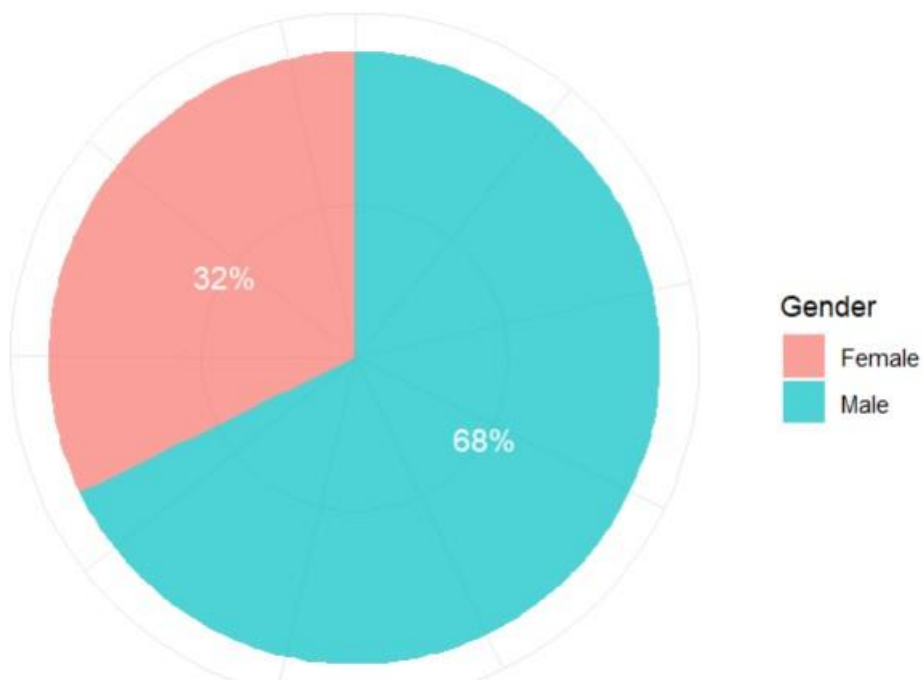
Our third data cleaning approach involves data splitting. Within our dataset, the "Age" and "Gender" feature is combined in a single column using a hyphen, as observed in the pre-cleaning screenshot. To address this issue, we implemented the "tidyr" package in R Studio. Following a successful installation, we employed the library function to activate the package and utilized the separate function to split the column based on the hyphen, creating distinct "Age" and "Gender" columns. The post-cleaning screenshot now reflects the dataset with separate and distinct "Age" and "Gender" columns.

E. Analysis & Visualizations

1. What is the proportion of Total Purchase Amount based on Gender?



Pie Chart of Total Purchase Amount by Gender



R code

```
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated.csv")

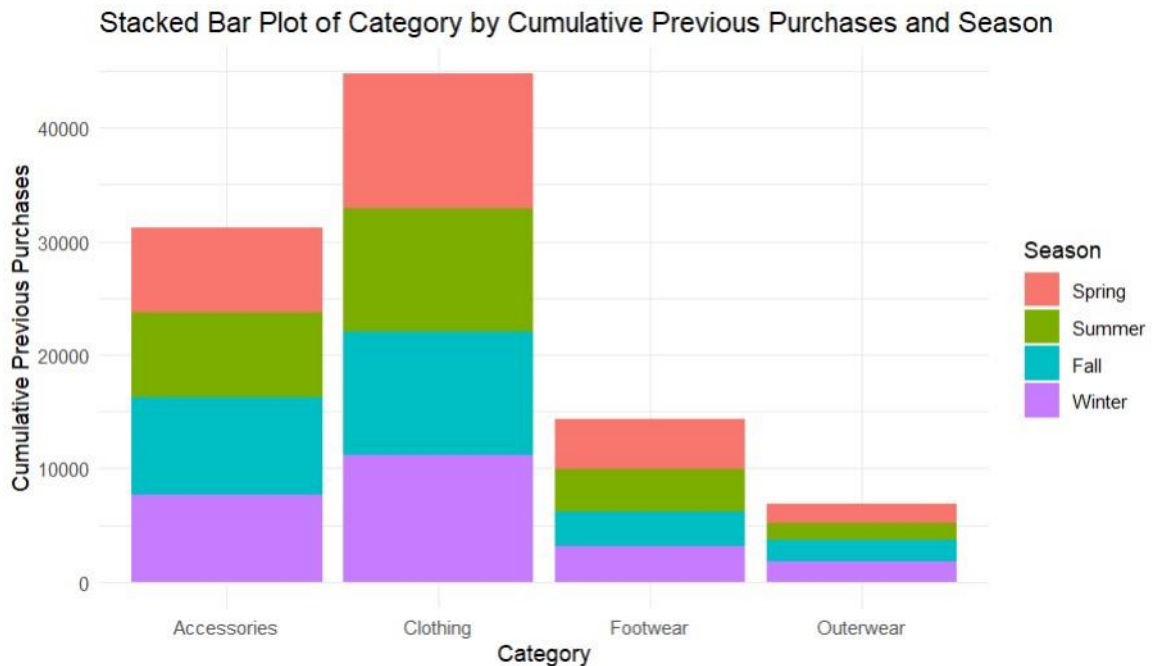
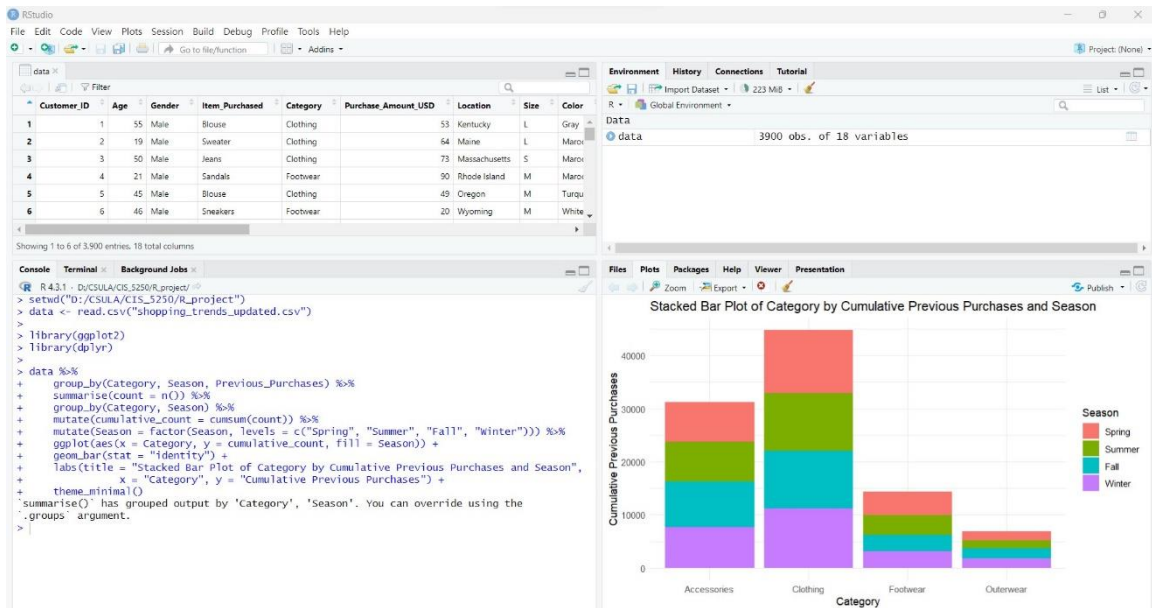
> library(ggplot2)
> library(dplyr)

> data %>%
+   group_by(Gender) %>%
+   summarise(total_purchase = sum(Purchase_Amount_USD)) %>%
+   ggplot(aes(x = "", y = total_purchase, fill = Gender)) +
+   geom_bar(stat = "identity", width = 1, alpha = 0.7) +
+   coord_polar("y") +
+   geom_text(aes(label = scales::percent(total_purchase/sum(total_purchase))),
+             position = position_stack(vjust = 0.5), color = "white", size = 4) +
+   labs(title = "Pie Chart of Total Purchase Amount by Gender",
+        , x = NULL, y = NULL) +
+   theme_minimal() +
+   theme(axis.text = element_blank())
```

Explanation

The pie chart depicting total purchase amounts by gender shows a noticeable difference, with males constituting 68% of the total and females representing 32%. Analyzing the underlying factors contributing to this difference, such as product preferences or marketing effectiveness, could provide valuable insights for businesses aiming to optimize their strategies. Additionally, adjusting marketing approaches to connect with the identified majority (Male) might enhance overall sales performance. This analysis underscores the significance of understanding the customer base to refine marketing strategies, ultimately fostering a more inclusive and practical approach in the competitive market.

2. What are the key seasonal patterns in cumulative previous purchases across different product categories?



R code

```
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated.csv")

> library(ggplot2)
> library(dplyr)

> data %>%
+   group_by(Category, Season, Previous_Purchases) %>%
+   summarise(count = n()) %>%
+   group_by(Category, Season) %>%
+   mutate(cumulative_count = cumsum(count)) %>%
+   mutate(Season = factor(Season, levels = c("Spring", "Summer", "Fall",
"Winter"))) %>%
+   ggplot(aes(x = Category, y = cumulative_count, fill = Season)) +
+   geom_bar(stat = "identity") +
+   labs(title = "Stacked Bar Plot of Category by Cumulative Previous Purchases and
Season",
+        x = "Category", y = "Cumulative Previous Purchases") +
+   theme_minimal()
```

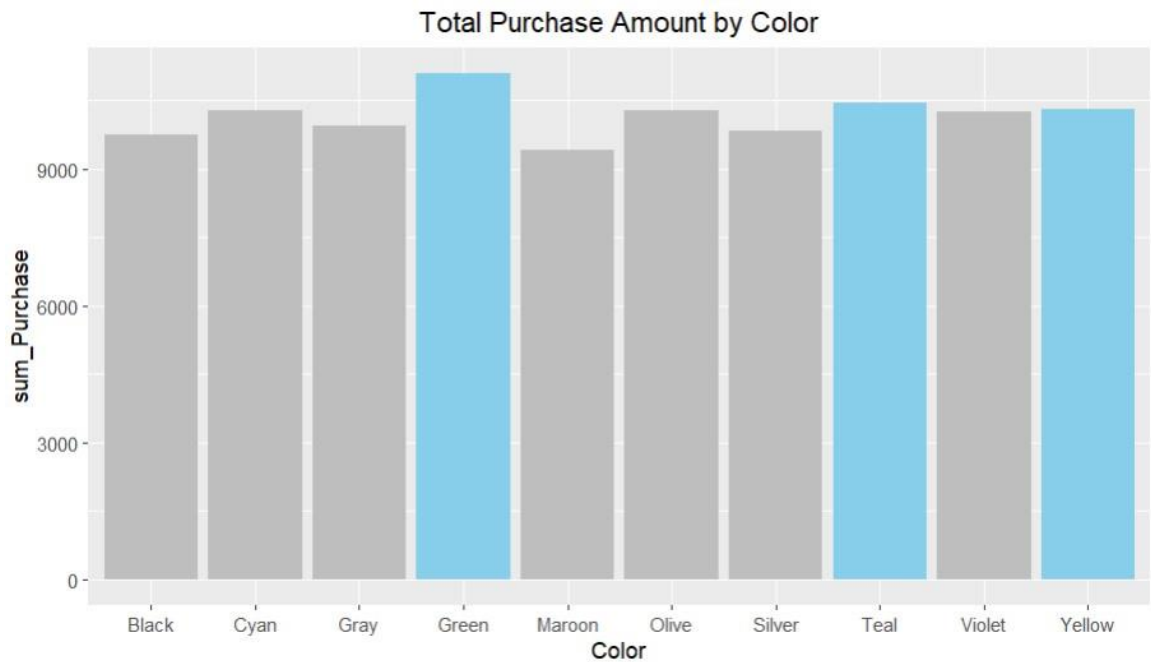
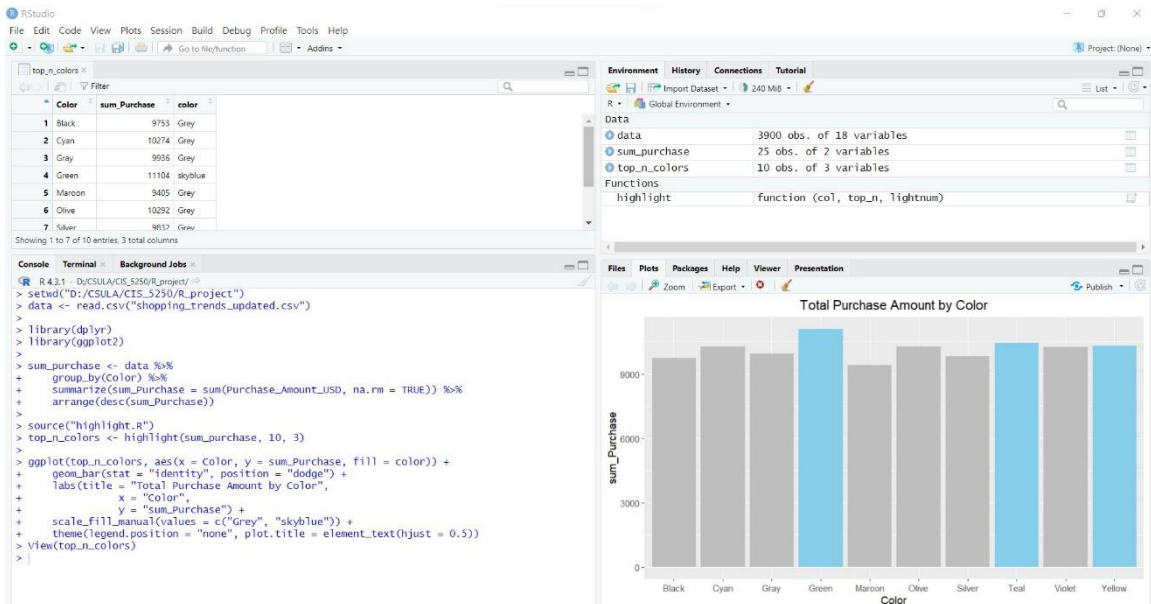
Explanation

The stack bar plot describes the relationship between "Category" and "Cumulative Previous Purchase" across the four seasons and reveals interesting patterns within the dataset. "Clothing" emerges as the predominant category with the highest cumulative previous purchase, surpassing 45,000 transactions, while "Accessories" follows closely as the second most popular category. In contrast, "Outerwear" has the lowest cumulative count, less than 10,000 transactions.

Furthermore, individual seasons within each category reveal a consistent trend. Regardless of the product type, each category maintains a relatively similar count of transactions across all four seasons. The robust stability of transaction counts throughout the seasons suggests that customer preferences and purchasing behaviors for

"Accessories," "Clothing," "Footwear," and "Outerwear" remain relatively constant, presenting an opportunity for targeted marketing strategies and inventory management.

3. Which colors are the most popular among customers, as indicated by the top 10 colors with the highest purchase amounts?



R code

```
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated.csv")

> library(dplyr)
> library(ggplot2)

> sum_purchase <- data %>%
+   group_by(Color) %>%
+   summarize(sum_Purchase = sum(Purchase_Amount_USD, na.rm = TRUE)) %>%
+   arrange(desc(sum_Purchase))

> source("highlight.R")
> top_n_colors <- highlight(sum_purchase, 10, 3)

> ggplot(top_n_colors, aes(x = Color, y = sum_Purchase, fill = color)) +
+   geom_bar(stat = "identity", position = "dodge") +
+   labs(title = "Total Purchase Amount by Color",
+        x = "Color",
+        y = "sum_Purchase") +
+   scale_fill_manual(values = c("Grey", "skyblue")) +
+   theme(legend.position = "none", plot.title = element_text(hjust = 0.5))
```

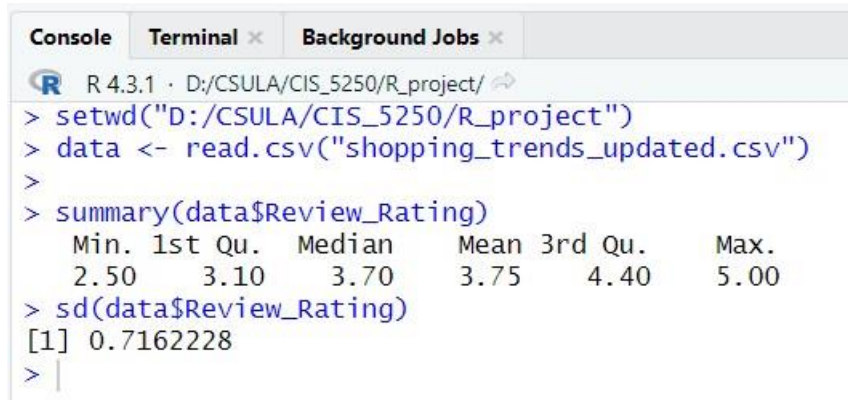
Explanation

This analysis reveals interesting insights into the top 10 purchase amounts across various colors. The selected colors, including Black, Cyan, Gray, Green, Maroon, Olive, Silver, Teal, Violet, and Yellow, provide a focused view on the most popular choices among customers. To enhance clarity, a strategic visualization approach was employed, highlighting the top 3 colors—Green, Teal, and Yellow—in a distinctive light blue, while the remaining colors were converted to grey. Despite subtle variations in the purchase amounts of the top 10 colors, the overall observation suggests a relatively balanced distribution of consumer preferences. The slight differences in the bar chart imply that customers exhibit comparable interest across the range of colors, with no single color

dominating significantly over others. This harmonious distribution could indicate a diverse customer base with varied preferences, contributing to a well-rounded product selection.

F. Statistical Summary

1. Analysis Variable: Review Rating

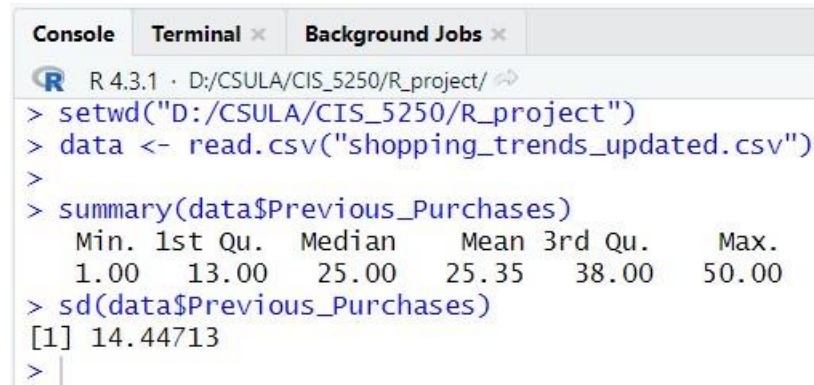


```
R 4.3.1 · D:/CSULA/CIS_5250/R_project/
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated.csv")
>
> summary(data$Review_Rating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.50   3.10   3.70   3.75   4.40   5.00
> sd(data$Review_Rating)
[1] 0.7162228
> |
```

Explanation

The "Review Rating" column in the dataset displays a range of values, ranging from a minimum of 2.5 to a maximum of 5.0. Summary statistics highlight a central tendency, with a mean rating of 3.75, close to the median of 3.7. This implies a well-balanced rating distribution, indicating that half of the reviews fall above and below this central point. Reviewing the interquartile range, from the first quartile at 3.10 to the third at 4.40, provides valuable insights into the middle 50% of the data, showcasing a dispersion within this interval. Since the standard deviation is 0.7162, we know there is a moderate level of variability in "Review Ratings" around the mean. Furthermore, the dataset mainly reflects above-average sentiments, with most ratings clustering between 3.10 and 4.40.

2. Analysis Variable: Previous Purchase



```
R 4.3.1 · D:/CSULA/CIS_5250/R_project/
> setwd("D:/CSULA/CIS_5250/R_project")
> data <- read.csv("shopping_trends_updated.csv")
>
> summary(data$Previous_Purchases)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.00  13.00   25.00   25.35   38.00   50.00
> sd(data$Previous_Purchases)
[1] 14.44713
> |
```

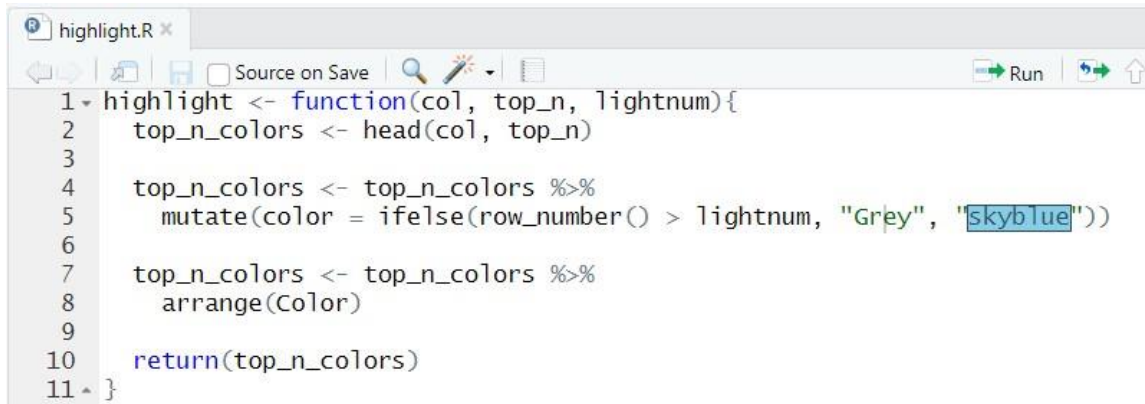
Explanation

The "Previous Purchases" column ranges from a minimum of 1.00 to a maximum of 50.00. Within this range, some customers have only made a single previous purchase, while others have engaged in up to 50 transactions. The median, positioned at 25.00, serves as the pivotal point, indicating that half of the customers have made fewer than 25 previous purchases. The mean, calculated at 25.35, suggests a slightly right-skewed distribution, pointing to a concentration of customers with fewer transactions. The interquartile range, from 13.00 to 38.00, expresses the central 50% of the data, depicting moderate variation. Furthermore, with a standard deviation of 14.45, there is a notable emphasis on dispersion, highlighting the diverse transaction histories within the dataset.

The analysis highlights a diverse customer base characterized by varying transaction patterns. Although a substantial portion falls within the mid-range, the broad spread of the data emphasizes the need for tailored strategies to accommodate different customer segments.

G. User-defined Function

Function – highlight:

A screenshot of an R script editor window titled 'highlight.R'. The editor shows a function definition for 'highlight'. The function takes three arguments: 'col', 'top_n', and 'lightnum'. It first selects the top 'top_n' rows from 'col'. Then, it uses 'mutate' to assign a 'color' column based on the 'row_number()' relative to 'lightnum'. Rows with 'row_number()' greater than 'lightnum' are colored 'Grey', and others are colored 'skyblue'. Finally, it uses 'arrange' to sort the selected rows by the 'Color' column and returns the result.

```
1 highlight <- function(col, top_n, lightnum){  
2   top_n_colors <- head(col, top_n)  
3  
4   top_n_colors <- top_n_colors %>%  
5     mutate(color = ifelse(row_number() > lightnum, "Grey", "skyblue"))  
6  
7   top_n_colors <- top_n_colors %>%  
8     arrange(Color)  
9  
10  return(top_n_colors)  
11 }
```

Explanation

Our user-defined function, "highlight," is beneficial before creating bar charts for categorical columns, especially those with many labels. It requires three inputs: "col" for the data to visualize in the bar chart, "top_n" as an integer input enabling the selection of the highest labels to display, and "lightnum," another integer input determining how many bars to highlight among the selected top_n. Additionally, it reorders the top n labels alphabetically. To illustrate its application, consider its use in the third visualization of part E. The bar chart showcases ten bars, with three highlighted in a distinctive light blue. The original data comprises 25 labels in "Color" column, and by employing the highlight function, we effortlessly isolate the top 10 labels and accentuate the top 3 among them.

H. Reference:

Banerjee, S. (2023, September). *Customer Shopping Trends Dataset*. [Www.kaggle.com](https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data).

<https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data>

Blaazer, E. (2022, December 19). *The role of color in fashion*. Fashion United.

<https://fashionunited.com/news/background/the-role-of-color-in-fashion/2022121951314>

Fowler, J. (2022, April 25). *Best Times To Go Shopping*. Investopedia.

<https://www.investopedia.com/financial-edge/0212/best-times-to-go-shopping.aspx>