Analyzing Customer Purchase Behavior

Paulette Rodriguez

March 2025

Table of Contents

Business Problem

For this project, we will be working with the retail data set. This data set represents a series of customer transactions for a retail business across different product categories such as Beauty, Clothing, and Electronics.

As data analysts, we are tasked with identifying key trends in customer purchasing behavior that can help the company improve marketing strategies, optimize pricing, and enhance customer targeting.

Our objective for this project is to be able to explore patterns in customer purchases to understand and answer the following questions:

- Which products or product categories generate the most revenue?
- What are the purchasing patterns by gender and age group?
- Are there any seasonal trends in purchasing behavior?
- What is the average purchase value for different customer segments?
- Can we predict potential future trends based on the available data?

Objective

Our objective is to perform a basic data analysis on the retail data set to answer the questions listed above. With the use of R, we will import the data, clean the data and visualize the data to help us uncover insights that can help the business make informed decisions.

The Data Analysis

Importing & Cleaning Data

First, we will begin by importing and cleaning the data set. This will help make sure the data set is running properly and any mistakes are fixed before we begin the data analysis. After importing the data, we take care of cleaning the data and making sure that the columns we will need are in the correct format. For example, after importing we want to remove any of the dollar signs attached to the revenue and price columns and convert them to a numeric form. This will help us when running our analysis because the formulas needed to get results require the values to be in numeric form.

After taking care of the dollar signs in the revenue and price columns, we want to make sure all of our spelling in the data set is correct and consistent. Within the data, one of the main columns we will be using is the Product.Category column with categories such as: Beauty, Clothing and Electronics, but we run into some problems as this column has spelling errors such as beautie, beautyy, clothi, electronixc, and electronicss to define the categories for the customers.

The following script will allow us to import and clean the correct data we will be using moving forward.

```
#importing data
retail <- read.csv("/Users/pauletterodriguez/Downloads/Final Project
(Google)/R Project/retail_sales.csv")

#removing dollar signs and converting to numeric for revenue column
retail$Total.Amount <- gsub("\\$","",retail$Total.Amount)
retail$Total.Amount <- as.numeric(retail$Total.Amount)

#removing dollar signs and converting to numeric for price column
retail$Price.per.Unit <- gsub("\\$", "", retail$Price.per.Unit)
retail$Price.per.Unit <- as.numeric(retail$Price.per.Unit)

#fixing spelling errors
retail <- mutate(retail, Product.Category = recode(.x=Product.Category,
"Beautie"="Beauty"))
retail <- mutate(retail, Product.Category = recode(.x=Product.Category,
"Beautyy"="Beauty"))
retail <- mutate(retail, Product.Category = recode(.x=Product.Category,
"Clothi"="Clothing"))
retail <- mutate(retail, Product.Category = recode(.x=Product.Category,
"Electronicss"="Electronics"))
retail <- mutate(retail, Product.Category = recode(.x=Product.Category,
"Electronixc"="Electronics"))
```

Question 1: What is the total revenue generated by each product category?

By calculating the total revenue generated by each product category we want to try understanding the overall revenue contribution by product category and how that can help identify which categories are the most profitable. First, we will calculate the total venue of the sales, in this case, the customer transactions. Then, we will calculate the average revenue per transaction in order to calculate the total revenue per product category type. In the data set, there are a total of three product categories as mentioned above. The product categories in the data set are: Beauty, Clothing and Electronics.

The following script will help us answer which product category out of the three generated more revenue for the company.

```
#calculating total revenue of sales
sum(retail$Total.Amount)

## [1] 456000

#calculating avg revenue per transaction
mean(retail$Total.Amount)

## [1] 456

#calculating total revenue per product category type
total_revenue <- retail %>%
  group_by(Product.Category)%>%
  summarize(totalrevenue = sum(Total.Amount))
print(total_revenue)

## # A tibble: 3 × 2
##   Product.Category totalrevenue
##   <fct>                   <dbl>
## 1 Beauty                143515
## 2 Clothing              155580
## 3 Electronics           156905

#Bar Plot for total quantity sold per product category
ggplot(total_revenue, aes(x = Product.Category, y = totalrevenue)) +
  geom_bar(stat = "identity", fill = "#76D7C4") +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Total Revenue by Product Category",
       x = "Product Category", y = "Revenue",
       subtitle = "Electronics stands as the top-selling category, while
Beauty ranks as the least sold") +
  theme(axis.text.x = element_text(size = 8, color = "#39858C", face =
"bold")) +
  theme(plot.title = element_text(color = "#3498DB", face = "bold")) +
  theme(plot.subtitle = element_text(size = 8, color = "#5F75CE", face =
"italic"))
```

Given our analysis, we can conclude that the overall total revenue of sales is 456,000 dollars with the average amount sold per transaction being 456 dollars. Given our business problem, we wanted to find what amount of the overall revenue belongs to each product category. We were able to conclude a complete breakdown of the revenues by each product type and per the analysis results here is what we found. Beauty made a total of 143,515 dollars in revenue, Clothing made a total of 155,580 dollars in revenue, and Electronics made a total of 156,905 dollars in revenue with the highest total of revenue overall.

We were able to visualize this result with a bar plot, where it is evident that electronics is higher in total revenue than the other products, with clothing coming in second.

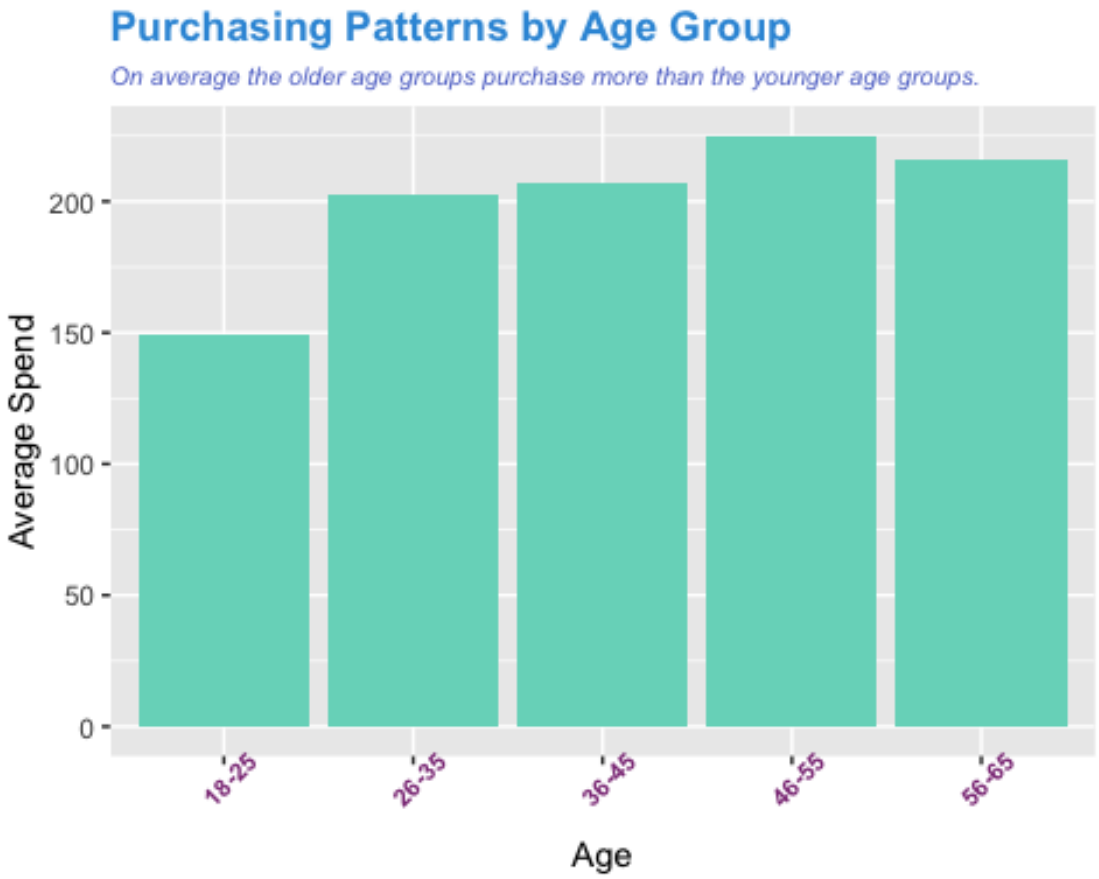
Question 2: What are the spending patterns by gender and age group?

Age Group Analysis:

By analyzing purchase amounts across different genders and age groups, we will reveal target customer segments that the business might want to focus on. We first begin by categorizing all of the ages into groups. For example, we will summarize all of the ages in the data set into the following age groups: 18-25, 26-35, 36-45, 56-65, and 65+. This will help us visualize the age groups better and in a neater way beneficial for analysis in both a table and a bar plot. We begin our age group analysis, which we have named

age_group_analysis for analysis purposes. This will help us group the ages, the average revenue and total purchases to analyze purchasing behavior.

```
#Categorizing Age into groups
retail$Age <- cut(retail$Age, breaks = c(18, 25, 35, 45, 55, 65, Inf),
                  labels = c("18-25", "26-35", "36-45", "46-55", "56-65",
"65+"), right = FALSE)

#purchasing patterns by age group
age_group_analysis <- retail %>%
  group_by(Age) %>%
  summarise(AverageSpend = mean(Total.Amount),
            TotalPurchases = n())

#Bar Plot of Age Group Analysis
ggplot(age_group_analysis, aes(x = Age, y = TotalPurchases)) +
  geom_bar(stat = "identity", fill = "#76D7C4") +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Purchasing Patterns by Age Group",
       x = "Age", y = "Average Spend",
       subtitle = "On average the older age groups purchase more than the
younger age groups.") +
  theme(axis.text.x = element_text(size = 8, color = "#8c3985", face =
"bold")) +
  theme(plot.title = element_text(color = "#3498DB", face = "bold")) +
  theme(plot.subtitle = element_text(size = 8, color = "#5F75CE", face =
"italic"))
```




```
print(age_group_analysis)
```

```
## # A tibble: 5 × 3
##   Age   AverageSpend TotalPurchases
##   <fct>        <dbl>          <int>
## 1 18-25         501.            149
## 2 26-35         478.            203
## 3 36-45         468.            207
## 4 46-55         432.            225
## 5 56-65         418.            216
```

We conclude, that the older age groups purchase more than younger age groups. The age group with the highest total amount of purchases are the customers within the ages of 46-55 and 56-65. On the other hand, on average the customers within the age groups of 18-25 and 36-45 spend more as compared to other age groups.

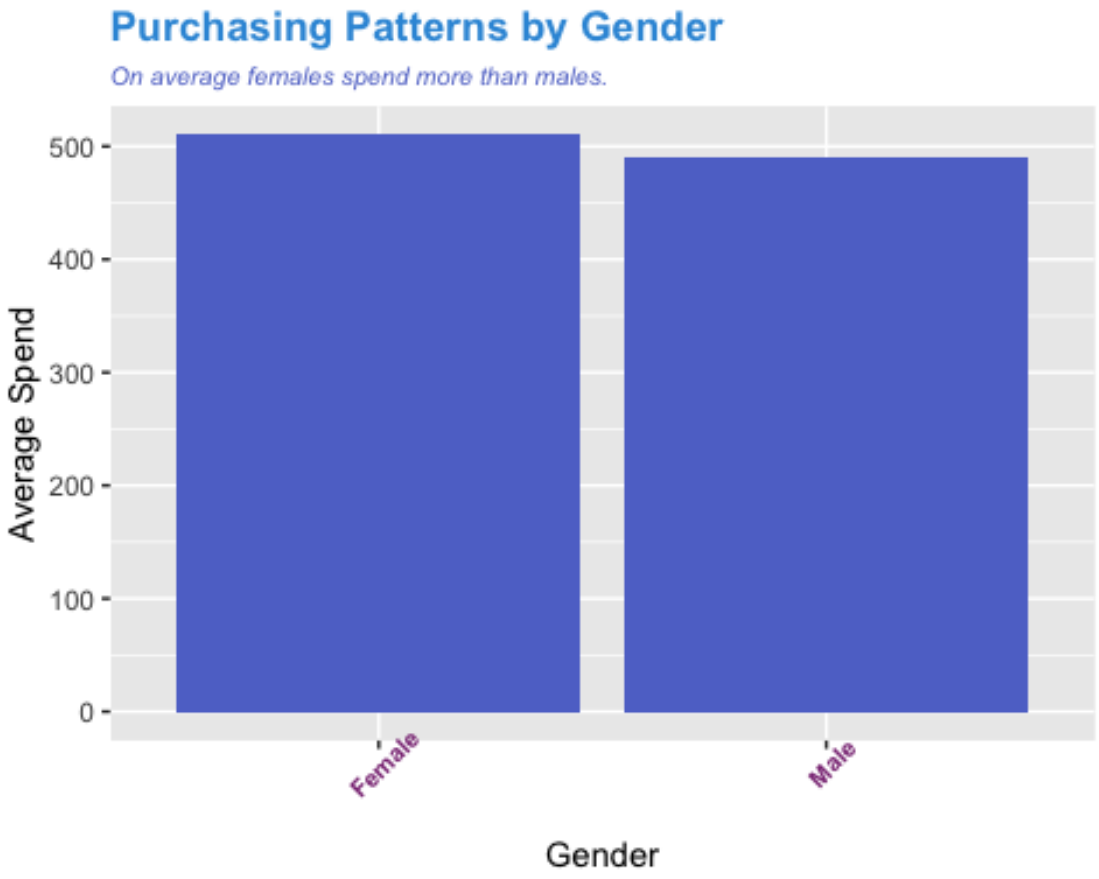
Gender Analysis:

We will then continue by performing an analysis based on the genders, named gender_analysis. We begin our gender analysis by grouping the genders by the average revenue and total purchases to analyze purchasing behavior. This will help us visualize the genders in a neater way beneficial for analysis in both a table and a bar plot.

```r
#purchasing behavior by gender
gender_analysis <- retail %>%
  group_by(Gender) %>%
  summarise(AvgSpend = mean(Total.Amount),
            TotalPurchase = n())

print(gender_analysis)
```

```
## # A tibble: 2 × 3
##   Gender AvgSpend TotalPurchase
##   <fct>     <dbl>         <int>
## 1 Female     457.           510
## 2 Male       455.           490
```

```r
#Bar Plot of Gender Analysis
ggplot(gender_analysis, aes(x = Gender, y = TotalPurchase)) +
  geom_bar(stat = "identity", fill = "#5F75CE") +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Purchasing Patterns by Gender",
       x = "Gender", y = "Average Spend",
       subtitle = "On average females spend more than males.") +
  theme(axis.text.x = element_text(size = 8, color = "#8c3985", face =
"bold")) +
  theme(plot.title = element_text(color = "#3498DB", face = "bold")) +
  theme(plot.subtitle = element_text(size = 8, color = "#5F75CE", face =
"italic"))
```



We conclude, that on average females purchase and spend more as compared to males. As per the analysis, females on average spent 457 dollars with an overall of 510 purchases as compared to the males who on spent on average of 455 dolars with an overall of 490 purchases.

Question 3: What is the average transaction amount by customer?

By analyzing the average transaction amount per customer we will give insight into customer spending habits and help to identify high-value customers. We begin by calculating the total transactions, and total revenue to perform and calculate the average transaction value (ATV) for the customers.

```
#calculate total transactions
trans <- length(retail$Transaction.ID)

#calculate total revenue
rev <- sum(retail$Total.Amount)

#calculate average transaction value
avg_trans_val <- rev/trans
print(avg_trans_val)

## [1] 456
```

Question 4: Is there any seasonality or trend in the transactions based on the date?

Date Analysis:

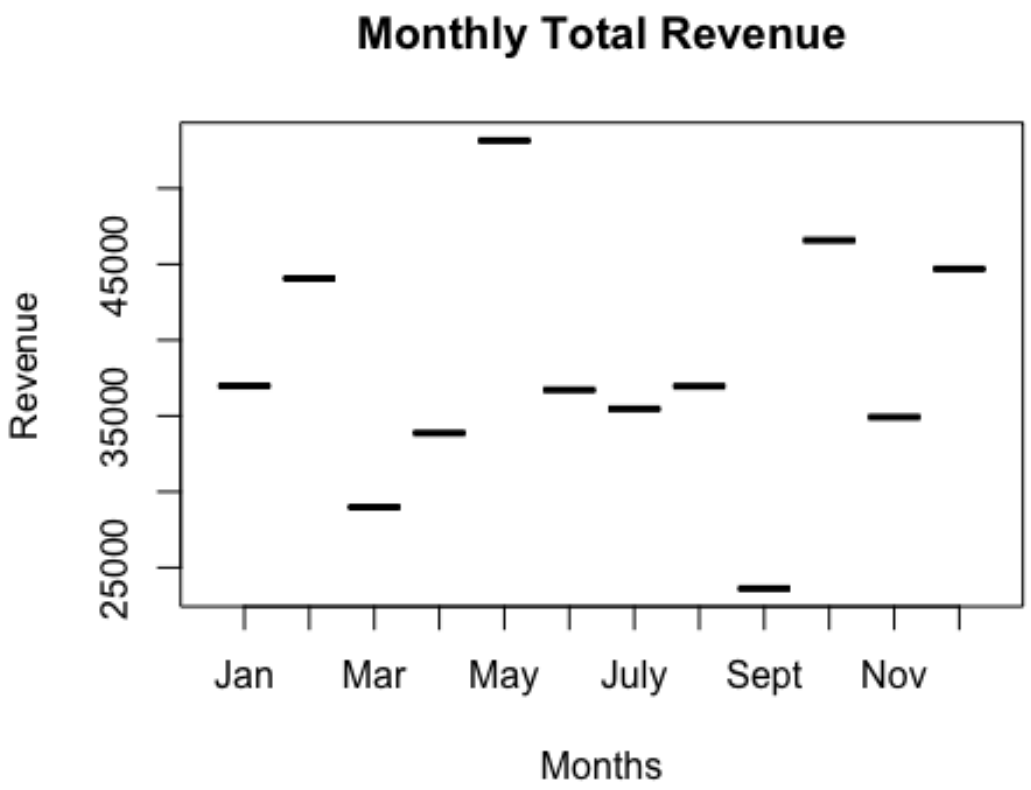
By examining purchase dates we can reveal any seasonal trends in sales (e.g., higher sales in certain months or around specific events).

We begin by categorizing the dates by month, and year. This will help keep the data we're working in a lot smaller as we will group all of the months together to list out all twelve months of the year instead of focusing on each individual data point in the retail data set. We start by defining a new data frame with the fixed dates. We are creating a new column for purchase date to change it into the first of the month and filtering any unfinished month transaction.

```
new_data <- retail %>% mutate(month = format(as.Date(Date), "%m")) %>%
  mutate(year_number = format(as.Date(Date), "%Y"))
new_data <- new_data[order(new_data$Date), ]

#create new column for purchase date to change it into first of the month
new_data$Purchase_Date <- paste(new_data$year_number, new_data$month, '01',
sep = "-")
new_data$Purchase_Date <- as.Date(new_data$Purchase_Date)

#filter unfinished month transaction
new_data <- new_data %>% filter(Purchase_Date < as.Date("2023-09-01"))
new_data$month = as.numeric(new_data$month)
new_data$month <- cut(new_data$month, breaks = c(01, 02, 03, 04, 05, 06, 07,
08, 09, 10, 11, 12, Inf),
                      labels = c("Jan", "Feb", "Mar", "Apr", "May", "June",
"July", "Aug", "Sept", "Oct", "Nov", "Dec"), right = FALSE)

summary(new_data$month)

##  Jan  Feb  Mar  Apr  May June July  Aug Sept  Oct  Nov  Dec
##   78   85   73   86  105   77   72   94   65   96   78   91
```

We then run a date analysis, similar to that of age group and gender analysis we did earlier. We visualize this in both a table and a scatter plot to finish the analysis for the months.

```
date_analysis <- new_data %>%
  group_by(month) %>%
  summarise(#TotalQuantity = sum(Quantity))
            TRev = sum(Total.Amount))

plot(date_analysis, col='blue', pch=19, cex=2, main="Monthly Total Revenue",
     xlab="Months", ylab="Revenue")
```




```
head(n=12,date_analysis)
```

```
## # A tibble: 12 × 2
##    month  TRev
##    <fct> <dbl>
##  1 Jan   36980
##  2 Feb   44060
##  3 Mar   28990
##  4 Apr   33870
##  5 May   53150
##  6 June  36715
##  7 July  35465
##  8 Aug   36960
##  9 Sept  23620
## 10 Oct   46580
## 11 Nov   34920
## 12 Dec   44690
```

From both our output and scatter plot, we conclude that the month of May has the highest values in both transactions, 105, and total revenue overall of 53,150 dollars.

In general, the top five months with highest revenues are May, October, December, February, and January. For the company, this means that the products are best selling in

the spring and towards the winter seasons. This may be due to the holidays, gift buying seasons, nice weather, and the new year.

Question 5: Which customers made the most purchases, and what is the frequency of their transactions?

Identifying the most frequent buyers and their behavior can help the business with customer loyalty programs or targeted marketing efforts.

For our final output, we can conclude which customers made the most purchases overall out of the entire date set.

```
retail$Customer.ID <- as.character(retail$Customer.ID)
retail[which.max(retail$Quantity),]$Customer.ID

## [1] "CUST008"

retail[which.min(retail$Quantity),]$Customer.ID

## [1] "CUST003"
```

We conclude that Customer 008 has the maximum amount of purchases, with Customer 003 coming in last.

Conclusion:

For this project, we were able to identify key trends in customer purchasing behavior to help the company improve on marketing strategies, optimizing pricing, and enhancing customer targeting.

Given the data analysis performed we were able to explore patterns in customer purchases to understand and answer the following questions:

- Which products or product categories generate the most revenue?

Per the analysis performed, we were able to conclude that the product Electronics generated the most revenue for the company as compared to rest of the products Beauty and Clothing.

- What are the purchasing patterns by gender and age group?

By analyzing purchasing patterns by gender and age group, we were able to conclude that on average females purchase and spend more than males do. This may mean that the company may want to market more to females, or on the other hand, the company may want to work on bettering their interactions and marketing techniques to cater to the male population. By the analysis we were also able to conclude that on average the older customers purchase more than the younger customers even though the younger age groups on average spend more. This may be due to the fact that the older customers may have better income, work full-time, or have financial independence and are able to manage and afford a bigger quantity of certain products as compared to the younger customers. From this we may conclude that the company should cater their pricing a little better to

meet with the younger groups' expectations and ability to buy certain products, and keep up their marketing technique for the older groups who tend to buy more in quantity than others.

- Are there any seasonal trends in purchasing behavior?

Given the performed analysis, we were able to conclude that the months of May, October, December, February, and January have the highest revenues. We can see a seasonal trend in purchasing behavior here, especially during the winter months, probably due to the holidays and gift buying that happens during these months.

- What is the average purchase value for different customer segments?

Given the analysis, the average purchase value for different customer segments turned out to be 456.

- Can we predict potential future trends based on the available data?

Yes, based on the available data we can conclude and see potential future trends in purchasing habits and total sales by age groups, gender, and time of year.

Given the performed data analysis, the company has great potential to keep on growing and skyrocketing in this field. By continuing to work with marketing, sales, finance and the business aspect of the company, there is potential in meeting certain expectations and achieving their goals.