Final Project

Shadin Chatila

2024-05-04

Introduction

For my final project, I have chosen to explore and analyze the key factors that influence consumer spending behavior across various products and demographics. I believe this topic is crucial, as understanding these factors can enable the development of more effective marketing strategies and enhance sales efficiency. By identifying what drives consumer decisions, businesses can tailor their approaches to meet the specific wants and needs of different consumer groups, thus enhancing their marketing efforts and optimizing overall profitability. This study aims to unravel the nuances of consumer behavior, providing actionable insights that can transform standard business practices.

Research Questions

1. Which demographic factors (such as location, gender, and age) have the most influence on consumer spending on different product categories?
2. How do seasonal changes and economic conditions affect consumer spending patterns?
3. How do marketing promotions influence consumer spending decisions across different demographics?
4. Is it possible to predict future spending behaviors based on the consumer loyalty as well as how frequently they make purchases?
5. How do different payment methods affect the spending habits/preferences of consumers?

Approach

In order to investigate how demographic factors, seasonal changes, marketing promotions, payment methods, and consumer loyalty impact consumer spending behaviors, I'll conduct a detailed exploratory data analysis using R. I will be following this overall templace when it comes to analyzing my datasets:

Cleaning Data:

1. Handling Missing Values: Assess the impact and frequency of the missing data. Remove/impute any missing data points based on their impact to the dataset and their frequency.
2. Data Type Conversion: Ensure that all data types are appropriately formatted for analysis. This could involve converting date strings or categorical variables.
3. Outliers: Identify and manage outliers using methods like IQR (Interquartile Range) or Z-scores. Decide whether to remove or adjust the outliers based on their effects to the data and the needs of the analysis.

Descriptive Statistics:

1. Central Tendencies: Evaluate the mean, median, and mode in order to understand the central values of data distributions. This will help determine the central location within the dataset.
2. Spread Metrics: Calculate standard deviations and variance in order to gauge the spread and dispersion of the data.

Visualization:

1. Distribution Visualizations: Use histograms, bar charts, and box plots to illustrate the data's distribution and identify any outliers.
2. Exploring Relationships: Use scatter plots and line graphs to investigate any trends and relationships between the variables.

Statistical Tests

1. t-tests: Use t-tests to compare the means of two different groups.
2. Chi-squared Tests: Conduct chi-squared tests on categorical data to analyze the associations between categories, like comparing types of payment methods with participation in loyalty programs.
3. Regression Analysis: Apply linear regression techniques, to explore how factors like age or the frequency of visits, are related to consumer spending.

Predictive Modeling:

1. Linear Regression Models: Develop models to predict spending based on linear relationships between variables.
2. Logistic Regression: If appropriate, use logistic regression to predict categorical outcomes (high vs low spending).

How This Approach Addresses the Problem

This method allows us to grasp a deep understanding of what influences consumer spending. By using both the descriptive and inferential statistics, we will be able to measure how difference factors impact spending and predict any future trends. This allows us to tackle a problem head on by providing insights which can help businesses with their strategies and marketing.

Data

Consumer Behavior and Shopping Habits Dataset

- https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset

1. Description: This dataset provides inisghts into the shopping habits of consumers and their behavior patterns. It also includes purchase history and other demographical data of the consumers.
2. Original Purpose: To analyze the consumer behavior in order to understand their shopping patterns and to predict their future buying patterns.
3. Variables: Product categories, amount spent, customer demographics, purchasing frequency

4. Data Peculiarities: Missing Values (no specification on how to care for missing values). Data imputation (No info provided on data imputation therefore signaling that the data could be complete)

Analyzing Customer Spending Habits

- https://www.kaggle.com/datasets/thedevastator/analyzing-customer-spending-habits-to-improve-sa

1. Description: This dataset zones into the shopping habits of consumers, more specifically aimed at trying to understand how the different factors like the seasons of the year, promotions and customer demographics influence consumer spending.
2. Original Purpose: This dataset was designed in order to help businesses improve their sales strategies based on consumer spending data.
3. Variables: Expenditure data by the customer, transaction details, promotions, customer demographics
4. Data Peculiarities: Data imputation (No info provided on data imputation therefore this may require implementing appropriate strategies during processing)

Customer Spend Dataset

- https://www.kaggle.com/datasets/manjeetsingh/retaildataset

1. Description: This dataset contains historical sales data from 45 stores, aiming to forecast future sales and understand the sales patterns related to holidays, store type, department details, and promotional activities. It includes weekly sales, holiday flags, and temperature data, providing a comprehensive view of the retail environment.
2. Original Purpose: The dataset is designed for tasks like sales forecasting and market analysis. It supports efforts to analyze the effectiveness of promotional strategies and to study the impacts of external factors such as holidays and economic fluctuations on sales.
3. Variables: Store, Dept, Date, Weekly_Sales, IsHoliday, Type, Size, Temperature, Fuel_Price, CPI, Unemployment
4. Data Peculiarities: Missing values and anomalies in weekly sales data could require imputation or careful outlier management to maintain the integrity of the analysis.

Packages

1. dplyr
2. ggplot2
3. tidyverse
4. lubridate
5. DataExplorer
6. caret

Plots and Tables

1. Histograms: to explore data distribution and outliers
2. Bar charts: compare spending across demographics and over time
3. Line graphs: compare spending across demographics and over time
4. Scatter plots: visualize correlations
5. Regression plots: visualize relationships and model fits

Skills and Knowledge to Develop

Advanced statistical analysis techniques in R, particularly in the context of predictive analytics as well as data modeling and learning more about machine learning techniques for predictive modeling in R.

Step 2

How did you import and clean your data?

```r
# Import the datasets
shopping_behavior <- read.csv("/Users/shadinchatila/Downloads/archive (1)/shopping_behavior_updated.csv
customer_spending <- read.csv("/Users/shadinchatila/Downloads/archive (8)/sales data-set.csv")
spending_habits <- read.csv("/Users/shadinchatila/Downloads/spending_habits.csv")
```

```r
# Checking the structure of each dataset
#str(shopping_behavior)
#str(customer_spending)
#str(spending_habits)
```

```r
# Viewing the first few rows to understand what the data looks like
#head(shopping_behavior)
#head(customer_spending)
#head(spending_habits)
```

```r
#summary(spending_habits)
```

```r
# Assuming missing values should be removed for simplicity
shopping_behavior <- na.omit(shopping_behavior)
customer_spending <- na.omit(customer_spending)
spending_habits <- na.omit(spending_habits)
```

```r
# Remove duplicates based on all columns
shopping_behavior <- unique(shopping_behavior)
customer_spending <- unique(customer_spending)
spending_habits <- unique(spending_habits)
```

```r
# Convert date from character to Date type
spending_habits$Date <- as.Date(spending_habits$Date, format="%m/%d/%y")

customer_spending$Date <- as.Date(customer_spending$Date, format="%d/%m/%Y")
```

```r
# Standardize text data to lower case
shopping_behavior$Gender <- tolower(shopping_behavior$Gender)
shopping_behavior$Item.Purchased <- tolower(shopping_behavior$Item.Purchased)
shopping_behavior$Category <- tolower(shopping_behavior$Category)
```

```r
# List of columns to convert to factors
columns_to_factor <- c("Gender", "Location", "Size", "Color", "Season",
                       "Subscription.Status", "Shipping.Type", "Discount.Applied",
                       "Promo.Code.Used", "Payment.Method", "Frequency.of.Purchases")

columns_to_factor2 <- c("Month", "Customer.Gender", "Country", "State", "Product.Category", "Sub.Catego

# Convert the columns to factors using lapply() function (forloop)
shopping_behavior[columns_to_factor] <- lapply(shopping_behavior[columns_to_factor], as.factor)
spending_habits[columns_to_factor2] <- lapply(spending_habits[columns_to_factor2], as.factor)
```

What does the final data set look like?

```r
# Final structure and summary check
str(shopping_behavior)
```

```
## 'data.frame':    3900 obs. of  18 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                : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ 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              : 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 ...
##  $ 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    : int  14 2 23 49 31 14 49 19 8 4 ...
##  $ 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 ...
```

```r
summary(shopping_behavior)
```

```
##   Customer.ID          Age            Gender     Item.Purchased
##  Min.   :   1.0   Min.   :18.00   female:1248   Length:3900
##  1st Qu.: 975.8   1st Qu.:31.00   male  :2652   Class :character
##  Median :1950.5   Median :44.00                 Mode  :character
##  Mean   :1950.5   Mean   :44.07
##  3rd Qu.:2925.2   3rd Qu.:57.00
##  Max.   :3900.0   Max.   :70.00
##
##    Category          Purchase.Amount..USD.       Location    Size
##  Length:3900        Min.   : 20.00        Montana   :  96   L :1053
##  Class :character   1st Qu.: 39.00        California:  95   M :1755
##  Mode  :character   Median : 60.00        Idaho     :  93   S : 663
```

```
##                      Mean   : 59.76        Illinois  :  92   XL: 429
##                      3rd Qu.: 81.00        Alabama   :  89
##                      Max.   :100.00        Minnesota :  88
##                                            (Other)   :3347
##      Color           Season     Review.Rating  Subscription.Status
##  Olive  : 177   Fall  :975   Min.   :2.50   No :2847
##  Yellow : 174   Spring:999   1st Qu.:3.10   Yes:1053
##  Silver : 173   Summer:955   Median :3.70
##  Teal   : 172   Winter:971   Mean   :3.75
##  Green  : 169                3rd Qu.:4.40
##  Black  : 167                Max.   :5.00
##  (Other):2868
##         Shipping.Type Discount.Applied Promo.Code.Used Previous.Purchases
##  2-Day Shipping:627   No :2223         No :2223        Min.   : 1.00
##  Express       :646   Yes:1677         Yes:1677        1st Qu.:13.00
##  Free Shipping :675                                    Median :25.00
##  Next Day Air  :648                                    Mean   :25.35
##  Standard      :654                                    3rd Qu.:38.00
##  Store Pickup  :650                                    Max.   :50.00
##
##        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
```

```r
str(customer_spending)
```

```
## 'data.frame':    421570 obs. of  5 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Dept        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : Date, format: "2010-02-05" "2010-02-12" ...
##  $ Weekly_Sales: num  24924 46039 41596 19404 21828 ...
##  $ IsHoliday   : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```

```r
summary(customer_spending)
```

```
##      Store          Dept            Date             Weekly_Sales
##  Min.   : 1.0   Min.   : 1.00   Min.   :2010-02-05   Min.   : -4989
##  1st Qu.:11.0   1st Qu.:18.00   1st Qu.:2010-10-08   1st Qu.:  2080
##  Median :22.0   Median :37.00   Median :2011-06-17   Median :  7612
##  Mean   :22.2   Mean   :44.26   Mean   :2011-06-18   Mean   : 15981
##  3rd Qu.:33.0   3rd Qu.:74.00   3rd Qu.:2012-02-24   3rd Qu.: 20206
##  Max.   :45.0   Max.   :99.00   Max.   :2012-10-26   Max.   :693099
##  IsHoliday
##  Mode :logical
##  FALSE:391909
##  TRUE :29661
##
##
##
```

```r
str(spending_habits)
```

```
## 'data.frame':    2574 obs. of  16 variables:
##  $ index           : int  312 313 314 315 316 317 318 319 320 321 ...
##  $ Date            : Date, format: "2016-01-11" "2016-01-11" ...
##  $ Year            : num  2016 2016 2016 2016 2016 ...
##  $ Month           : Factor w/ 12 levels "April","August",..: 5 5 5 5 5 5 5 5 4 8 ...
##  $ Customer.Age    : Factor w/ 52 levels "17","18","19",..: 24 24 24 24 24 24 24 24 24 24 ...
##  $ Customer.Gender : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Country         : Factor w/ 4 levels "France","Germany",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ State           : Factor w/ 29 levels "Alabama","Bayern",..: 29 29 29 29 29 29 29 29 29 29 ...
##  $ Product.Category: Factor w/ 3 levels "Accessories",..: 2 1 2 1 1 1 2 1 2 2 ...
##  $ Sub.Category    : Factor w/ 16 levels "Bike Racks","Bike Stands",..: 12 8 11 3 3 8 11 8 15 12 ...
##  $ Quantity        : num  3 2 2 2 1 2 1 3 1 2 ...
##  $ Unit.Cost       : num  567 192 1160 115 140 ...
##  $ Unit.Price      : num  790 199 1512 147 167 ...
##  $ Cost            : num  1701 385 2320 230 140 ...
##  $ Revenue         : num  2370 398 3023 294 167 ...
##  $ Column1         : num  2370 398 3023 294 167 ...
##  - attr(*, "na.action")= 'omit' Named int [1:32293] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:32293] "1" "2" "3" "4" ...
```

```r
summary(spending_habits)
```

```
##      index             Date                 Year          Month     
##  Min.   : 312.0   Min.   :2015-01-01   Min.   :2015   December: 270  
##  1st Qu.: 955.2   1st Qu.:2015-10-10   1st Qu.:2015   June    : 264  
##  Median :1598.5   Median :2016-01-04   Median :2016   January : 250  
##  Mean   :1598.5   Mean   :2016-01-05   Mean   :2016   August  : 222  
##  3rd Qu.:2241.8   3rd Qu.:2016-04-14   3rd Qu.:2016   May     : 221  
##  Max.   :2935.0   Max.   :2016-07-31   Max.   :2016   July    : 207  
##                                                       (Other) :1140  
##   Customer.Age  Customer.Gender           Country                  State    
##  39     : 167   F:1250          France        : 430   California      :860  
##  38     : 154   M:1324          Germany       : 251   Washington      :513  
##  34     : 151                   United Kingdom: 344   England         :344  
##  32     : 149                   United States :1549   Oregon          :164  
##  40     : 128                                         Hessen          : 90  
##  28     : 124                                         Seine Saint Denis: 76  
##  (Other):1701                                         (Other)         :527  
##     Product.Category            Sub.Category    Quantity       Unit.Cost      
##  Accessories:1653    Tires and Tubes   :895   Min.   :1.000   Min.   :   0.67  
##  Bikes      : 528    Helmets           :314   1st Qu.:1.000   1st Qu.:  46.00  
##  Clothing   : 393    Mountain Bikes    :305   Median :2.000   Median : 175.00  
##                      Bottles and Cages:241   Mean   :1.989   Mean   : 388.83  
##                      Jerseys           :217   3rd Qu.:3.000   3rd Qu.: 528.00  
##                      Road Bikes        :126   Max.   :3.000   Max.   :3120.00  
##                      (Other)           :476                                  
##    Unit.Price            Cost             Revenue           Column1      
##  Min.   :   0.667   Min.   :   2.0   Min.   :   2.0   Min.   :   2.0  
##  1st Qu.:  55.083   1st Qu.:  88.0   1st Qu.: 101.0   1st Qu.: 104.2  
##  Median : 194.250   Median : 300.0   Median : 354.5   Median : 390.5  
```

```
## Mean   : 426.595   Mean   : 642.1   Mean   : 703.7   Mean   : 688.1
## 3rd Qu.: 588.500   3rd Qu.: 850.0   3rd Qu.: 989.0   3rd Qu.: 975.8
## Max.   :3887.000   Max.   :3600.0   Max.   :4923.0   Max.   :3681.0
##
```

What information is not self-evident?

Things like interactions between variables, non-linear relationships, subgroup variations, influence of promotions and seasonal trends are all not self evident. In order to uncover the information that is not self-evident. The following techniques below can be used to uncover this information:

1. Advanced Analytical Techniques
2. Exploratiry Data Analysis (EDA)
3. Linear Regression Analysis
4. Logistic Regression Analysis
5. Predictive Modeling
6. Multivariate Regression
7. Machine Learning

What are different ways you could look at this data?

1. Which demographic factors (such as location, gender, and age) have the most influence on consumer spending on different product categories?

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```r
# Ensure Customer.Age is numeric
spending_habits$Customer.Age <- as.numeric(as.character(spending_habits$Customer.Age))

# Summary statistics by demographic factors
summary_by_demo <- spending_habits %>%
  group_by(Country, State, Customer.Gender, Age = cut(`Customer.Age`, breaks = c(18, 25, 35, 45, 55, 65
  summarise(Average_Spending = mean(Revenue, na.rm = TRUE),
            Count = n(),
            .groups = 'drop')

# Display the summary
```

```r
##print(summary_by_demo)

# ANOVA to check the effect of demographics on Revenue
anova_result <- aov(Revenue ~ Country + State + Customer.Gender + Customer.Age, data = spending_habits)
summary(anova_result)
```

```
##                   Df    Sum Sq  Mean Sq F value   Pr(>F)    
## Country            3 7.671e+07 25569143  42.265  < 2e-16 ***
## State             25 5.303e+07  2121183   3.506 1.03e-08 ***
## Customer.Gender    1 1.779e+04    17789   0.029    0.864    
## Customer.Age       1 1.294e+06  1294159   2.139    0.144    
## Residuals       2543 1.538e+09   604972                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. How do seasonal changes and economic conditions affect consumer spending patterns?

```r
# Group by 'Season' and calculate mean, median, and sum
Season_purchase_info <- aggregate(Purchase.Amount..USD. ~ Season, data = shopping_behavior,
                                  FUN = function(x) c(mean = mean(x), median = median(x), sum = sum(x)))

# Format output
Season_purchase_info <- do.call(data.frame, Season_purchase_info)
names(Season_purchase_info)[2:4] <- c("Mean", "Median", "Sum")

# Load the dplyr package
library(dplyr)

# Group by 'Season' and calculate mean, median, and sum
Season_purchase_info <- shopping_behavior %>%
  group_by(Season) %>%
  summarise(
    Mean = mean(Purchase.Amount..USD., na.rm = TRUE),
    Median = median(Purchase.Amount..USD., na.rm = TRUE),
    Sum = sum(Purchase.Amount..USD., na.rm = TRUE)
  )

# Print the result
print(Season_purchase_info)
```

```
## # A tibble: 4 x 4
##   Season  Mean Median   Sum
##   <fct>  <dbl>  <int> <int>
## 1 Fall    61.6     62 60018
## 2 Spring  58.7     58 58679
## 3 Summer  58.4     58 55777
## 4 Winter  60.4     62 58607
```

```r
# Analyze regional trends
# Display some entries for each location
location_groups <- shopping_behavior %>%
  group_by(Location) %>%
  slice_head(n = 300)  # This is similar to .head(300) for each group in pandas

# Analyze average price by region
avg_price <- shopping_behavior %>%
  group_by(Location) %>%
  summarise(Average_Price = mean(Purchase.Amount..USD., na.rm = TRUE)) %>%
  arrange(desc(Average_Price))

# Print the result
print(avg_price)
```

```
## # A tibble: 50 x 2
##    Location      Average_Price
##    <fct>                 <dbl>
##  1 Alaska                 67.6
##  2 Pennsylvania           66.6
##  3 Arizona                66.6
##  4 West Virginia          63.9
##  5 Nevada                 63.4
##  6 Washington             63.3
##  7 North Dakota           62.9
##  8 Virginia               62.9
##  9 Utah                   62.6
## 10 Michigan               62.1
## # i 40 more rows
```

```r
# Analyze category counts by region
category_counts <- shopping_behavior %>%
  count(Location, Category) %>%
  group_by(Location) %>%
  summarise(Max_Count = max(n), .groups = 'drop')  # Find the maximum count of categories in each locat

# Print the result
print(category_counts)
```

```
## # A tibble: 50 x 2
##    Location    Max_Count
##    <fct>           <int>
##  1 Alabama            41
##  2 Alaska             33
##  3 Arizona            32
##  4 Arkansas           37
##  5 California         47
##  6 Colorado           32
##  7 Connecticut        32
##  8 Delaware           41
##  9 Florida            30
## 10 Georgia            41
## # i 40 more rows
```

```r
# Extract month and year from Date column
customer_spending$Month <- format(customer_spending$Date, "%m")
customer_spending$Year <- format(customer_spending$Date, "%Y")

# Define the seasons based on month
customer_spending$Season <- cut(as.integer(customer_spending$Month),
                         breaks=c(0, 3, 6, 9, 12),
                         labels=c("Winter", "Spring", "Summer", "Autumn"),
                         include.lowest=TRUE)

# Aggregate data by season
seasonal_sales <- aggregate(Weekly_Sales ~ Season, data=customer_spending, FUN=sum)
seasonal_sales
```

```
##   Season Weekly_Sales
## 1 Winter   1494112230
## 2 Spring   1826615244
## 3 Summer   1841852365
## 4 Autumn   1574639148
```

```r
# Compare Holiday vs. Non-Holiday Sales
holiday_effect <- aggregate(Weekly_Sales ~ Season + IsHoliday, data = customer_spending, FUN = mean)
colnames(holiday_effect)[3] <- "Average_Sales"

holiday_effect
```

```
##   Season IsHoliday Average_Sales
## 1 Winter     FALSE      15214.66
## 2 Spring     FALSE      15913.64
## 3 Summer     FALSE      15660.24
## 4 Autumn     FALSE      16974.03
## 5 Winter      TRUE      16378.00
## 6 Summer      TRUE      15881.69
## 7 Autumn      TRUE      18386.36
```

3. How do marketing promotions influence consumer spending decisions across different demographics?

```r
shopping_behavior$Age_Group <- cut(shopping_behavior$Age, breaks=c(18, 25, 35, 45, 55, 65, 75), labels=

# Group by 'Promo.Code.Used', 'Gender', and 'Age_Group', then calculate mean, median, and sum
promo_influence <- shopping_behavior %>%
  group_by(Promo.Code.Used, Gender, Age_Group) %>%
  summarise(
    Mean = mean(Purchase.Amount..USD., na.rm = TRUE),
    Median = median(Purchase.Amount..USD., na.rm = TRUE),
    Total_Spending = sum(Purchase.Amount..USD., na.rm = TRUE),
    .groups = 'drop'
  )
```

```r
# Print the results
print(promo_influence)
```

```
## # A tibble: 21 x 6
##    Promo.Code.Used Gender Age_Group  Mean Median Total_Spending
##    <fct>           <fct>  <fct>     <dbl>  <dbl>          <int>
##  1 No              female 18-25      61.1   61             9342
##  2 No              female 26-35      62.1   64.5          15019
##  3 No              female 36-45      59.2   58            14394
##  4 No              female 46-55      58.9   57.5          14480
##  5 No              female 56-65      61.0   63            14648
##  6 No              female 66-75      58.8   59             6114
##  7 No              female <NA>       59.7   59.5           1194
##  8 No              male   18-25      61.6   63             8065
##  9 No              male   26-35      58.9   56            10431
## 10 No              male   36-45      59.5   60            10947
## # i 11 more rows
```

4. Is it possible to predict future spending behaviors based on the consumer loyalty as well as how frequently they make purchases?

```r
# Checking correlation matrix for age
cor_data <- shopping_behavior[, c("Purchase.Amount..USD.", "Age")]
cor_matrix <- cor(cor_data, use = "complete.obs") # Ensuring missing values are handled properly
cor_matrix
```

```
##                       Purchase.Amount..USD.         Age
## Purchase.Amount..USD.            1.00000000 -0.01042365
## Age                             -0.01042365  1.00000000
```

```r
# Average spending by Subscription Status and Purchase Frequency
shopping_behavior %>%
  group_by(Subscription.Status, Frequency.of.Purchases) %>%
  summarise(Average_Spending = mean(Purchase.Amount..USD., na.rm = TRUE),
            Count = n()) %>%
  arrange(desc(Average_Spending))
```

```
## `summarise()` has grouped output by 'Subscription.Status'. You can override
## using the `.groups` argument.
```

```
## # A tibble: 14 x 4
## # Groups:   Subscription.Status [2]
##    Subscription.Status Frequency.of.Purchases Average_Spending Count
##    <fct>               <fct>                             <dbl> <int>
##  1 Yes                 Quarterly                          61.0   140
##  2 No                  Bi-Weekly                          60.9   407
##  3 Yes                 Every 3 Months                     60.8   154
##  4 No                  Annually                           60.7   412
##  5 Yes                 Bi-Weekly                          60.1   140
##  6 No                  Every 3 Months                     59.8   430
```

```
##   7 No                    Quarterly                                59.7   423
##   8 No                    Fortnightly                              59.6   389
##   9 No                    Monthly                                  59.4   404
##  10 Yes                   Monthly                                  59.1   149
##  11 Yes                   Weekly                                   59.1   157
##  12 No                    Weekly                                   58.9   382
##  13 Yes                   Annually                                 58.8   160
##  14 Yes                   Fortnightly                              57.8   153
```

```r
# ANOVA for Subscription Status
anova_subscription <- aov(Purchase.Amount..USD. ~ Subscription.Status, data = shopping_behavior)
summary(anova_subscription)
```

```
##                       Df  Sum Sq Mean Sq F value Pr(>F)
## Subscription.Status    1     107   107.1   0.191  0.662
## Residuals           3898 2187223   561.1
```

```r
# ANOVA for Frequency of Purchases
anova_frequency <- aov(Purchase.Amount..USD. ~ Frequency.of.Purchases, data = shopping_behavior)
summary(anova_frequency)
```

```
##                          Df  Sum Sq Mean Sq F value Pr(>F)
## Frequency.of.Purchases    6    1371   228.5   0.407  0.875
## Residuals              3893 2185959   561.5
```

```r
# Linear regression model
model <- lm(Purchase.Amount..USD. ~ Subscription.Status + Frequency.of.Purchases, data = shopping_behav
summary(model)
```

```
##
## Call:
## lm(formula = Purchase.Amount..USD. ~ Subscription.Status + Frequency.of.Purchases,
##     data = shopping_behavior)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.783 -21.074  -0.072  20.827  41.272
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            60.2694     1.0194  59.125   <2e-16 ***
## Subscription.StatusYes                 -0.3444     0.8552  -0.403    0.687
## Frequency.of.PurchasesBi-Weekly         0.5134     1.4174   0.362    0.717
## Frequency.of.PurchasesEvery 3 Months   -0.0964     1.3942  -0.069    0.945
## Frequency.of.PurchasesFortnightly      -1.1187     1.4206  -0.787    0.431
## Frequency.of.PurchasesMonthly          -0.8457     1.4134  -0.598    0.550
## Frequency.of.PurchasesQuarterly        -0.1998     1.4072  -0.142    0.887
## Frequency.of.PurchasesWeekly           -1.1969     1.4227  -0.841    0.400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.7 on 3892 degrees of freedom
## Multiple R-squared:  0.0006685,  Adjusted R-squared:  -0.001129
## F-statistic: 0.3719 on 7 and 3892 DF,  p-value: 0.919
```

5

```r
# Summarize Average Purchase Amount by the Payment Method
average_spending_by_payment <- shopping_behavior %>%
  group_by(Payment.Method) %>%
  summarise(
    Average_Spending = mean(Purchase.Amount..USD., na.rm = TRUE),
    Count = n()
  ) %>%
  arrange(desc(Average_Spending))

# Print the result
print(average_spending_by_payment)
```

```
## # A tibble: 6 x 3
##   Payment.Method Average_Spending Count
##   <fct>                     <dbl> <int>
## 1 Debit Card                 60.9   636
## 2 Credit Card                60.1   671
## 3 Bank Transfer              59.7   612
## 4 Cash                       59.7   670
## 5 PayPal                     59.2   677
## 6 Venmo                      58.9   634
```

```r
# ANOVA test
anova_result <- aov(Purchase.Amount..USD. ~ Payment.Method, data = shopping_behavior)
summary(anova_result)
```

```
##                  Df  Sum Sq Mean Sq F value Pr(>F)
## Payment.Method    5    1514   302.8    0.54  0.746
## Residuals      3894 2185816   561.3
```

How do you plan to slice and dice the data?

Yes, slicing and dicing the data is very useful when it comes to grouping and making subsets of the data. More specifically, we have grouped data based on various categorical variables such as 'Season', 'IsHoliday', and 'Payment Method' in order to understand how these factors could affect consumer spending. This apporach of slicing and dicing helps isolate the effects of any specific categories on the spending behaviors. There are some instances where we specifically looked at subsets of data, like transactions during holidays or non-holidays in order to see if there were any notable differences in the spending habits which could be crucial when it comes to understanding the seasonal effects.

How could you summarize your data to answer key questions?

Q1: Influence of Demographic Factors on Spending

We grouped the data by country, state, gender, and age groups, calculating the average spending for each of the groups. This allowed us to observe how spending patterns varied accross the demographic segments, which provided a granular view of the consumer spending habits. The ANOVA test used showed that both country and state showed significant effect on spending. The p value was below the 0.05 threshold which indicates strong significance. Gender however did not show any significance. Age showed a very marginal effect (pvalue = 0.144), this indicates that there was a potential trend where the age could possibly effect the spending but not strong enough to be statistically significant.

Q2: Seasonal Analysis of Spending

We conducted an analysis using two of our data sets, shopping_behavior and customer_spending. To start, we aggregated the purchase amounts by season in order to calculate the mean, median and total spending for each season. Based on these results, we saw that Autumn had the highest average and median spending which pointed to the fact that the seasonal peak in consumer spending had to be during this period. We also analyzed the spending by location and noticed that certain states like Alaska and Pennsylvania had higher spending which could indicate economic strength. Next, we analyzed the customer_spending data. We segmented the sales data in order to compare the holiday vs non holiday sales within each season. The data showed that the sales during the holidays were consistently higher than the non holiday periods accross all of the seasons. By analyzing these two datasets, we confirmed that seasonal changes and specific economic conditions like holidays could influence spending by the average consumer.

Q3: Influence of Promotions on Consumer Spending

To prepare, we categorized the data (specifically the age category) into different groups, 18-25, 26-35, 36-45, 46-55, 56-65, and 66-75. The data was grouped by the use of the promo codes, gender, and the age groups. For each group, mean, median and total purchase amounts were calculated in order to asses the spending behavior. It was noticed that the use of promotions/sales influences the spending patterns significantly. Males aged 18-25 without a promo spent on average of $61.56, while those with a promo spent slightly less on average but ended up contributing more to the total spending because they had higher transaction volume or more frequent purchases. The analysis showed that marketing promotions have a varying impact on consumer spending across different demographics. It appears that younger, male consumers usually are more responsive to the promotions which results in their higher spending habits. Whereas the females and older consumers seem to have a steadier spending pattern.

Q4: Future Spending Based on Loyalty

Conducted a correlation analysis between the age and purhcase amount in order to determine if there was a direct relationship that might also imply predictability in spending behaviors. The correlation between purchase amount and age was very low (-0.0104) which suggests that there is no significant relationship. We then grouped the data by subscription status and the frequency of purchases to see if there are any spending patterns. Consumers that had a subscription status of “Yes” and purchasing “Weekly” showed the highest average spending at ~ 59.10. Non-subscribers purchasing weekly were at a spending average of ~ 58.92. This means there is a very small difference in spending habits based on the subscription status.

Q5: Influence of Payment Method on Spending

To address this question, we grouped the data by payment method, and then calculated the average purchase amount for each method. This is used to identify if any certain methods were associated with higher spending. We then used the ANOVA test to see any differences in their statistical significance. Based on the results, Debit cards and credit cards yielded the most average spending but the rest of the categories fell shortly behind. The ANOVA test showed a p-value greater than 0.05 (0.746), which means the differences in average spending across the payment methods are not statistically significant. The findings show that the consumer spending habits are relatively consistent across all payment methods which suggests that there are factors other than payment method that could be effecting the data.

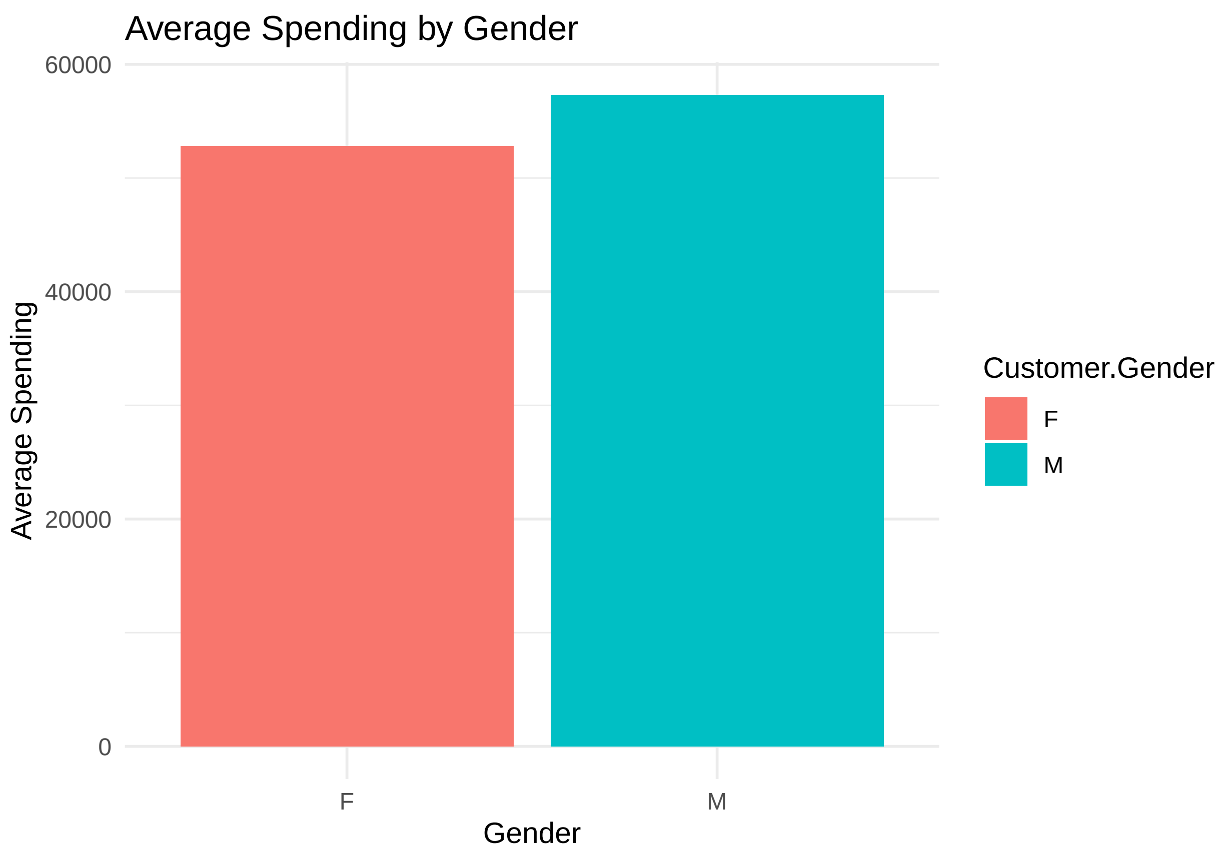
What types of plots and tables will help you to illustrate the findings to your questions?

Q1: Influence of Demographic Factors on Spending

```r
library(tidyverse)
```

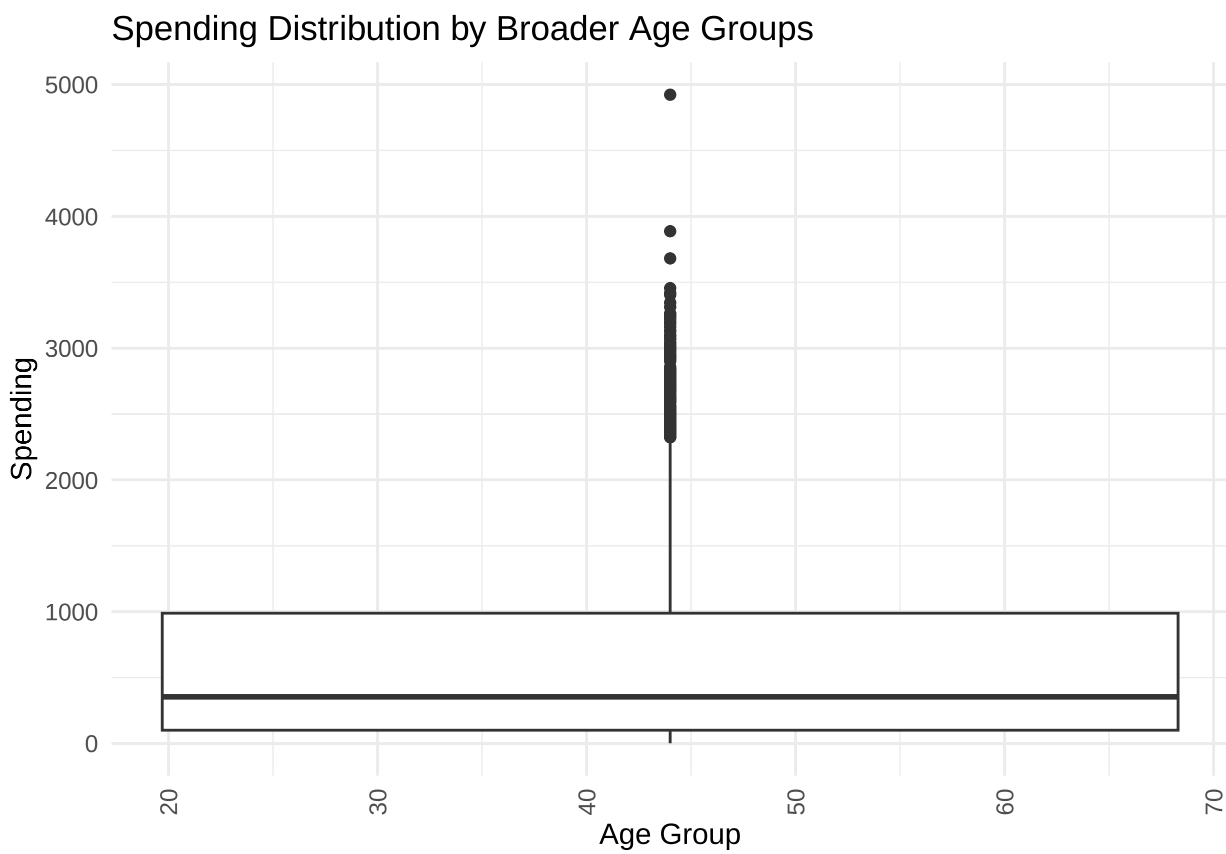
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Bar Chart for Average Spending by Gender
ggplot(data = summary_by_demo, aes(x = Customer.Gender, y = Average_Spending, fill = Customer.Gender)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Spending by Gender", x = "Gender", y = "Average Spending") +
  theme_minimal()
```

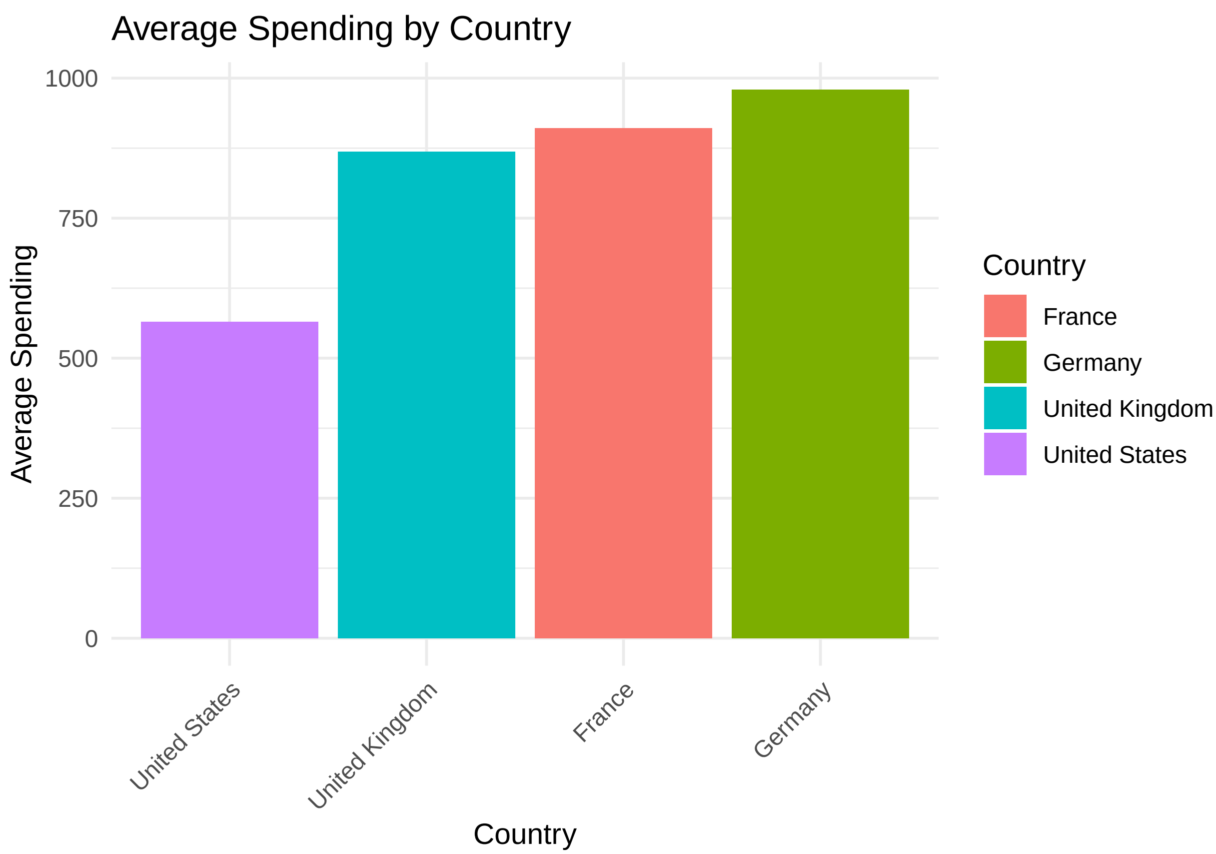



```
# Box Plot for Spending by Age Group
ggplot(data = spending_habits, aes(x = Customer.Age, y = Revenue, fill = Customer.Age)) +
  geom_boxplot() +
  labs(title = "Spending Distribution by Broader Age Groups", x = "Age Group", y = "Spending") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, vjust = 0.5),
    legend.position = "right",
    legend.margin = margin(t = 15, unit = "pt")
  )
```

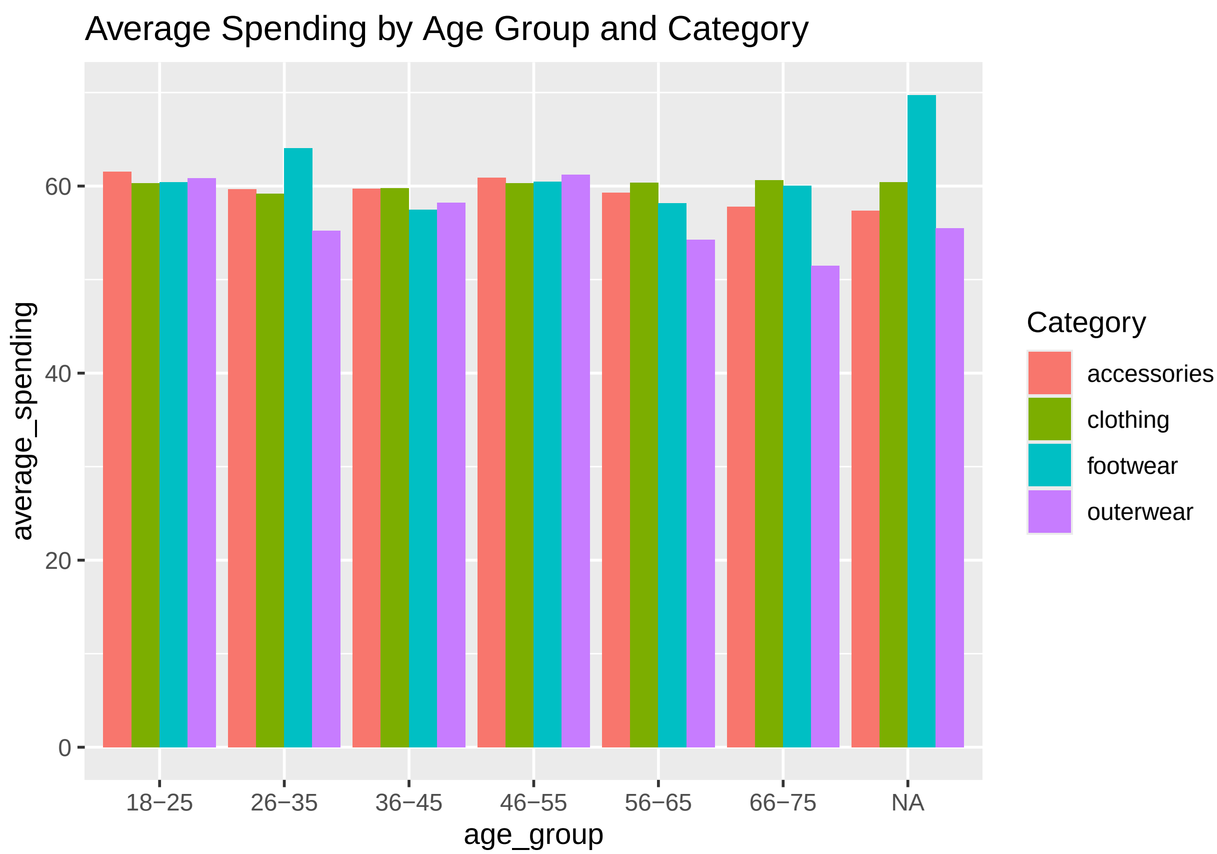
```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?

## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```r
# Creating summary data frame for average spending by country
country_spending <- spending_habits %>%
  group_by(Country) %>%
  summarise(Average_Spending = mean(Revenue, na.rm = TRUE)) %>%
  arrange(desc(Average_Spending))

# Plotting
ggplot(country_spending, aes(x = reorder(Country, Average_Spending), y = Average_Spending, fill = Count
  geom_col() +
  labs(title = "Average Spending by Country", x = "Country", y = "Average Spending") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

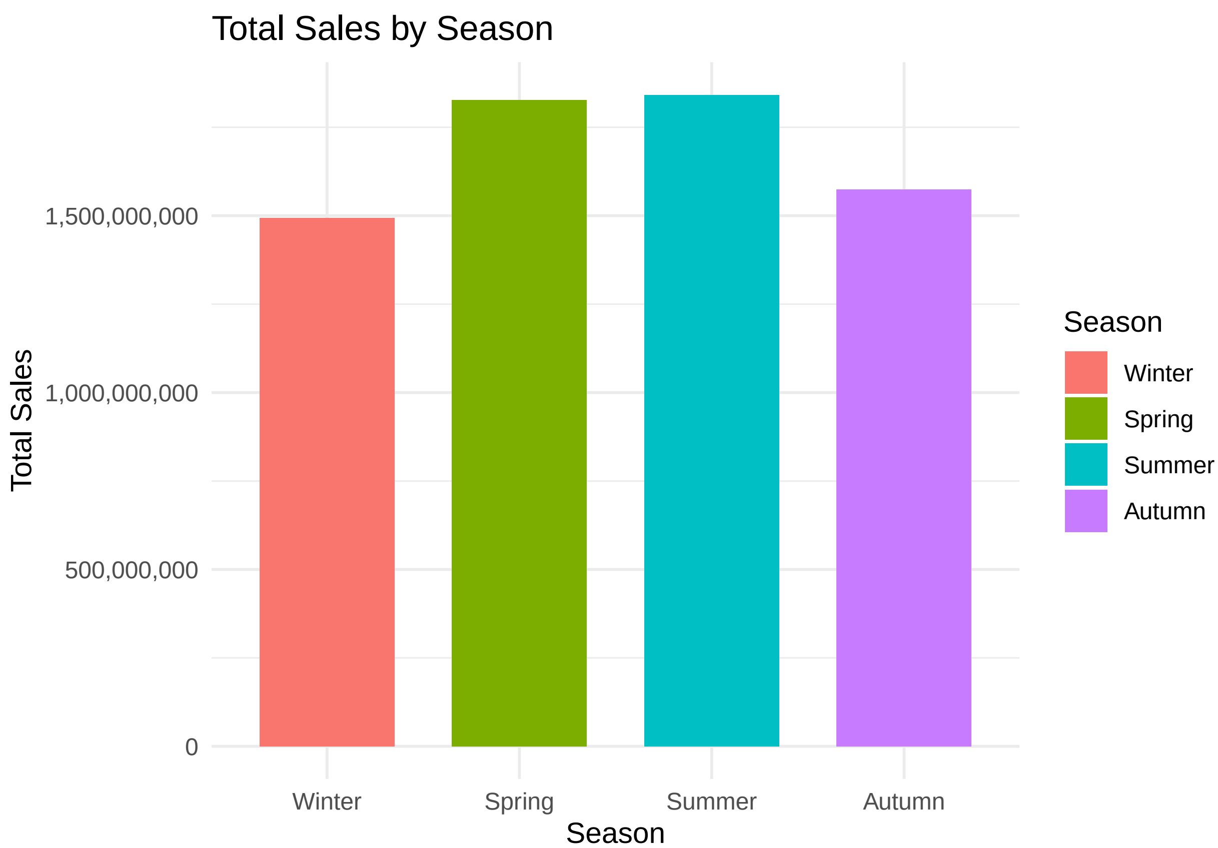



```
# Creating age groups
shopping_behavior <- shopping_behavior %>%
  mutate(age_group = cut(Age, breaks = c(18, 25, 35, 45, 55, 65, 75), labels = c("18-25", "26-35", "36-

# Analyzing spending by age group
age_group_analysis <- shopping_behavior %>%
  group_by(age_group, Category) %>%
  summarise(average_spending = mean(Purchase.Amount..USD.), .groups = 'drop')

# Plotting
ggplot(age_group_analysis, aes(x = age_group, y = average_spending, fill = Category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average Spending by Age Group and Category")
```

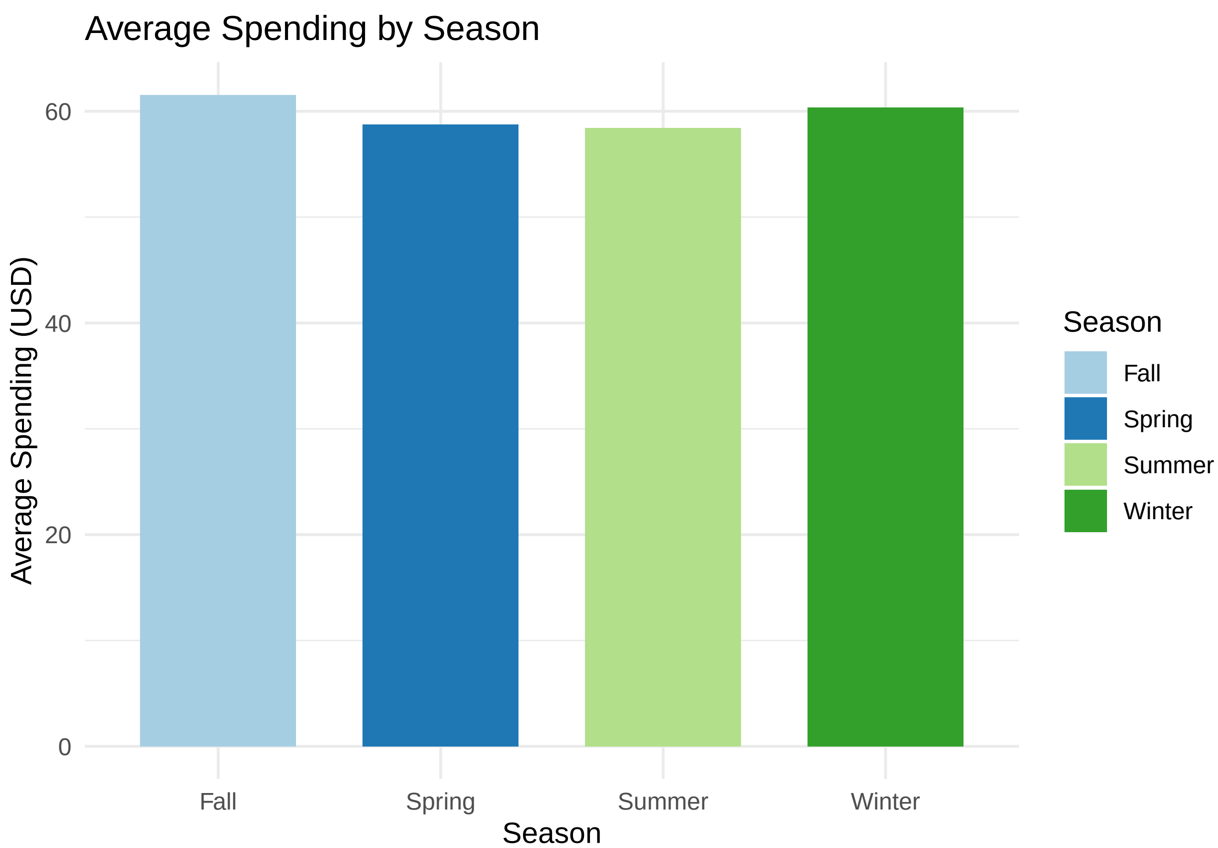


Q2: Seasonal Analysis of Spending

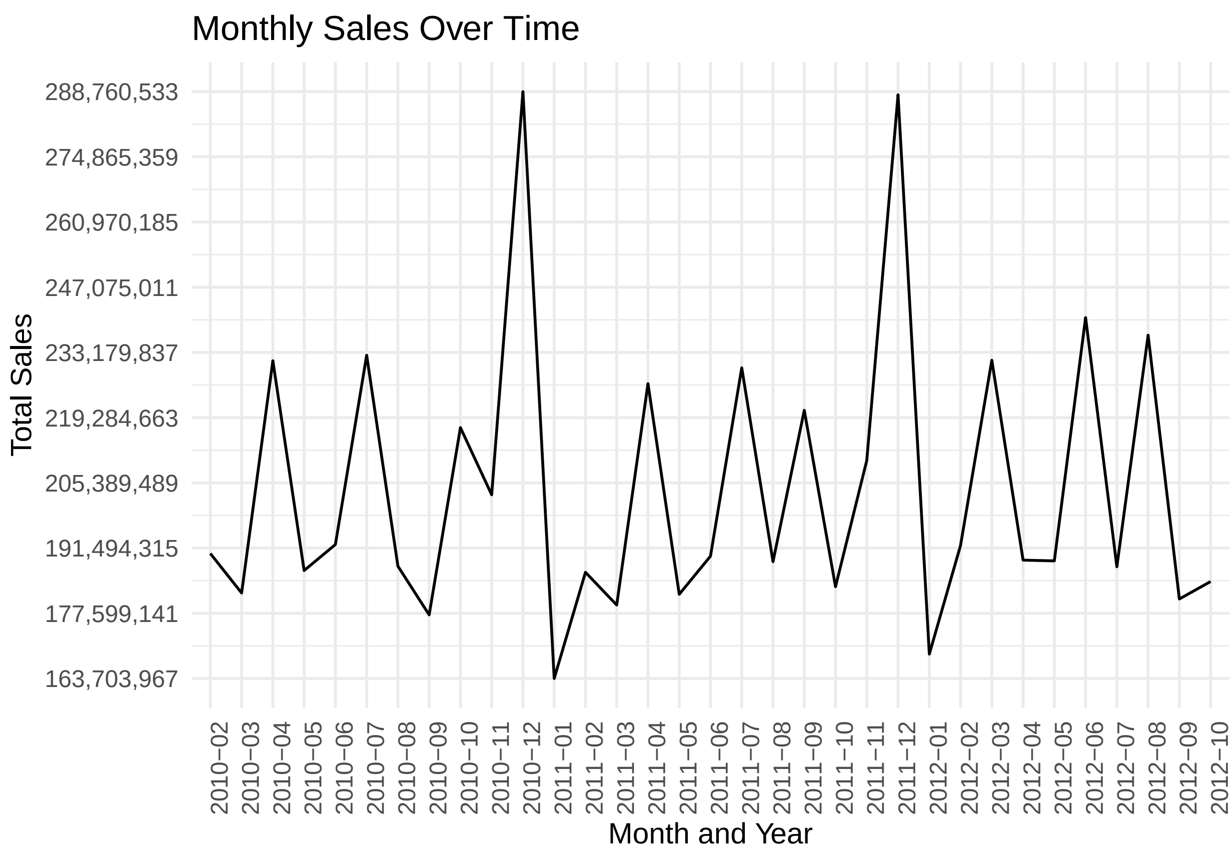
```r
# Calculate average spending by season
average_spending_by_season <- aggregate(Purchase.Amount..USD. ~ Season, data = shopping_behavior, mean)

ggplot(data = seasonal_sales, aes(x = Season, y = Weekly_Sales, fill = Season)) +
  geom_bar(stat = "identity", width = 0.7) +
  labs(title = "Total Sales by Season", x = "Season", y = "Total Sales") +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma)
```



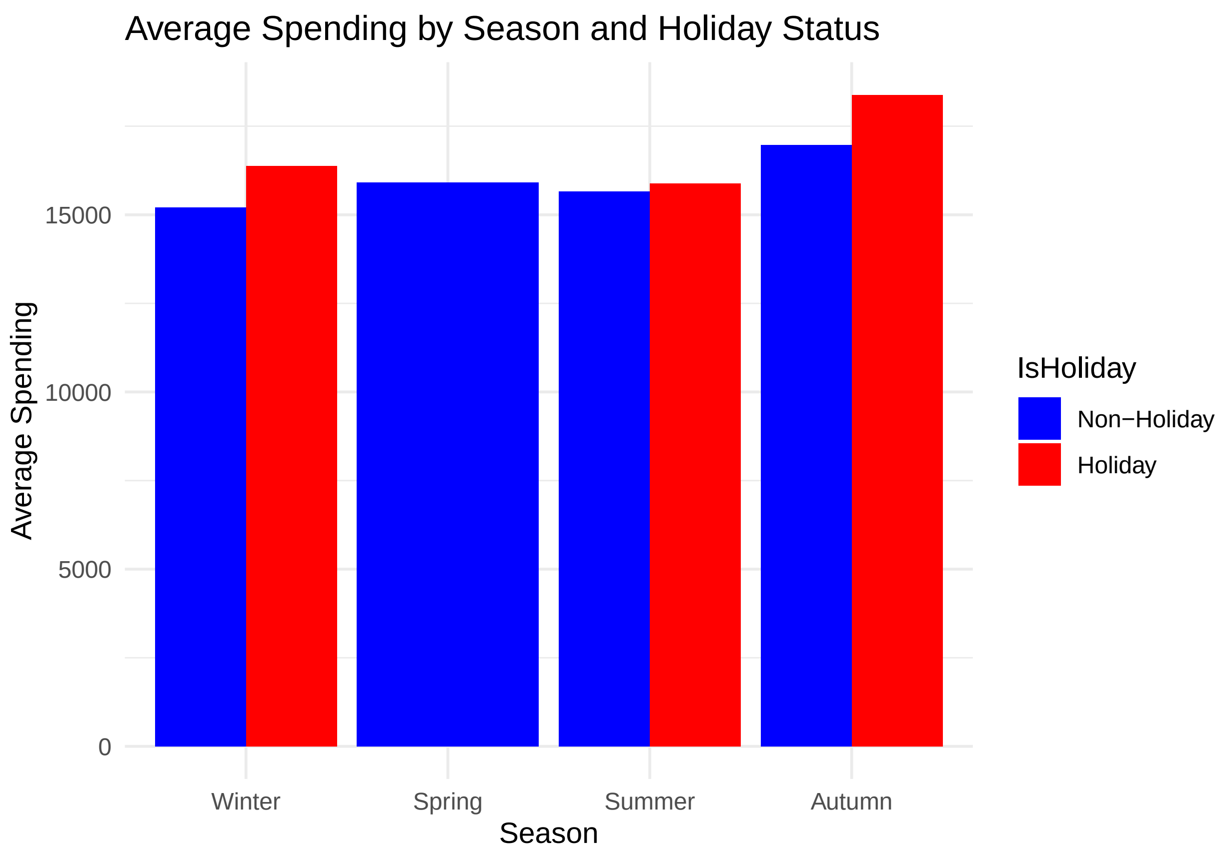

```r
# Create a bar plot to visualize average spending by season
ggplot(average_spending_by_season, aes(x = Season, y = Purchase.Amount..USD., fill = Season)) +
  geom_bar(stat = "identity", width = 0.7) +
  labs(title = "Average Spending by Season", x = "Season", y = "Average Spending (USD)") +
  theme_minimal() +
  scale_fill_brewer(palette = "Paired")
```




```
customer_spending$Date <- as.Date(customer_spending$Date, "%m/%d/%Y")
monthly_sales <- customer_spending %>%
  mutate(Month_Year = format(Date, "%Y-%m")) %>%
  group_by(Month_Year) %>%
  summarise(Total_Sales = sum(Weekly_Sales, na.rm = TRUE))


# Plotting monthly sales over time
ggplot(data = monthly_sales, aes(x = Month_Year, y = Total_Sales, group = 1)) +
  geom_line() +
  labs(title = "Monthly Sales Over Time", x = "Month and Year", y = "Total Sales") +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma, breaks = seq(min(monthly_sales$Total_Sales), max(monthly_sa
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

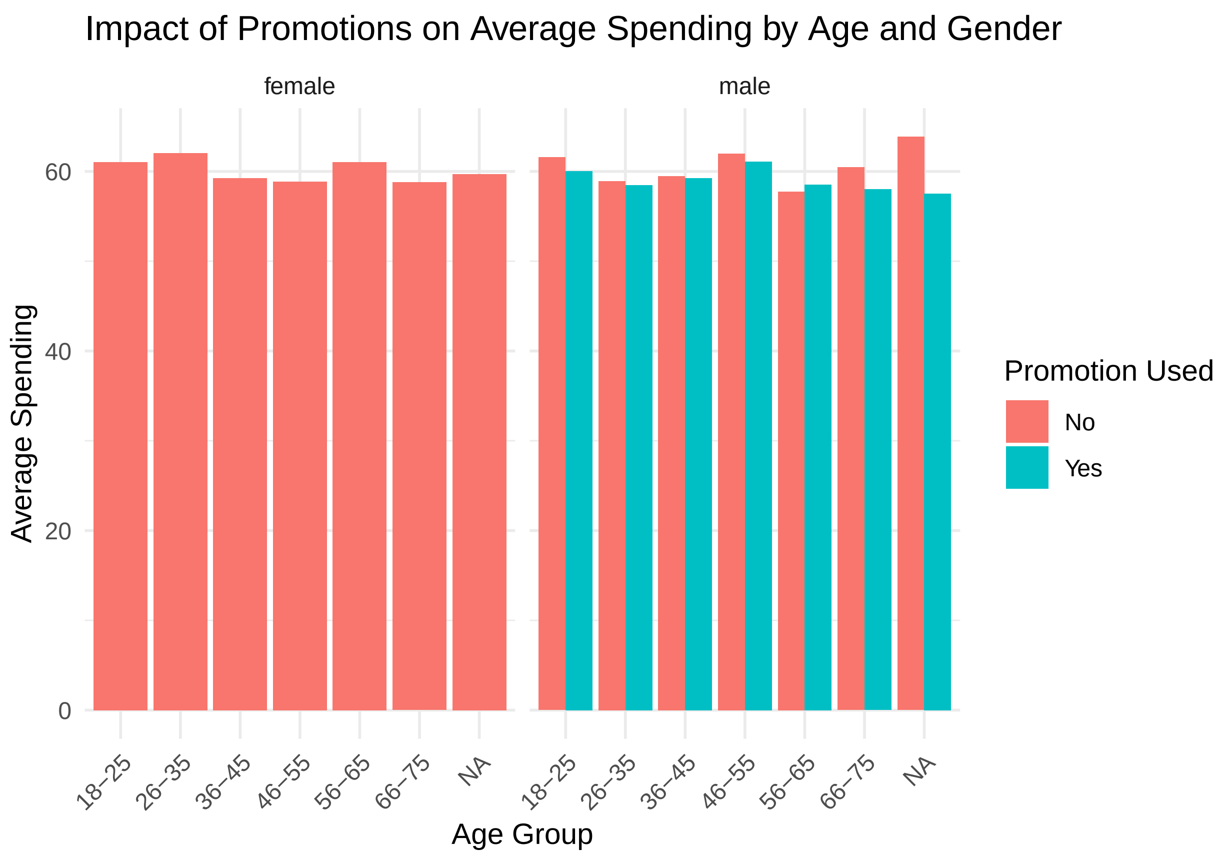



```r
# Plotting average sales by season with a distinction between holiday and non-holiday periods
ggplot(data = holiday_effect, aes(x = Season, y = Average_Sales, fill = IsHoliday)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Average Spending by Season and Holiday Status", x = "Season", y = "Average Spending") +
  scale_fill_manual(values = c("blue", "red"), labels = c("Non-Holiday", "Holiday")) +
  theme_minimal()
```

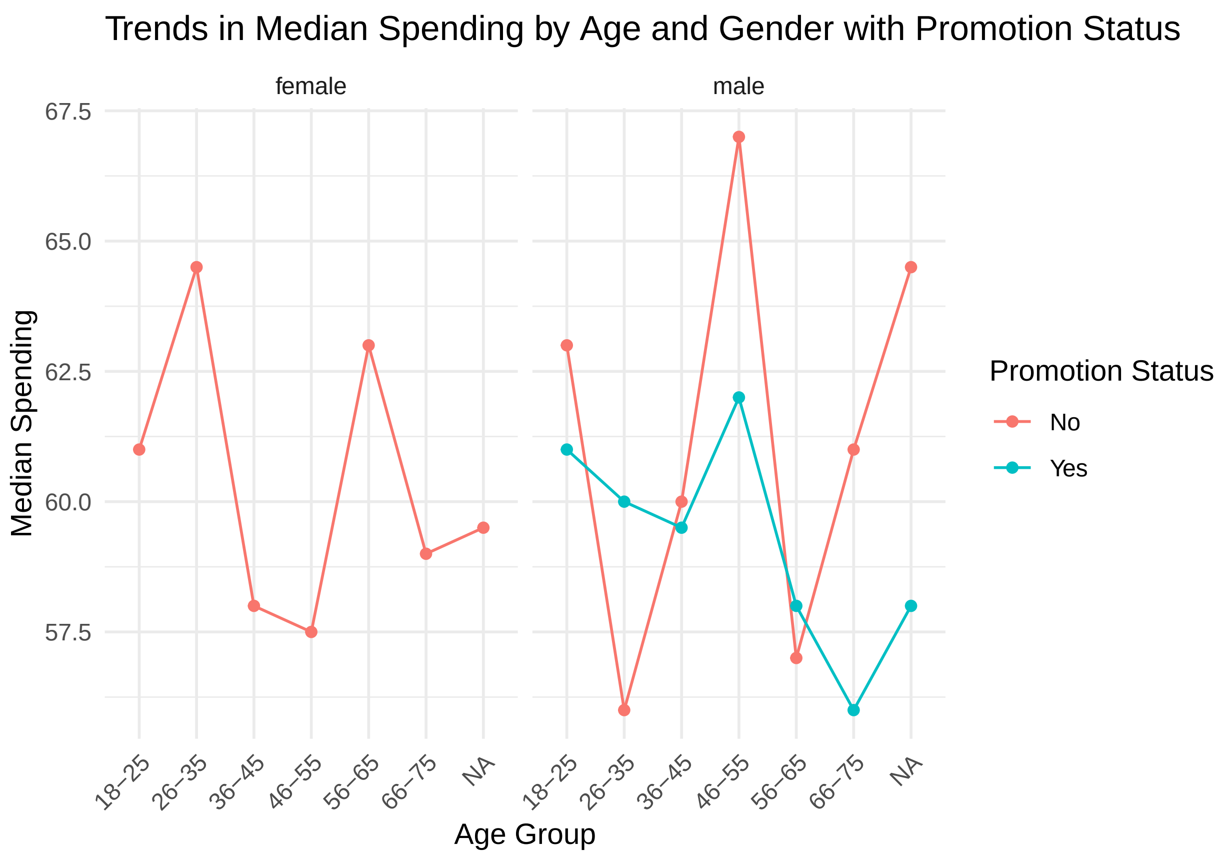


Q3: Influence of Promotions on Consumer Spending

```
library(ggplot2)

# Bar plot to compare the mean spending with and without promotions across demographics
ggplot(data = promo_influence, aes(x = Age_Group, y = Mean, fill = Promo.Code.Used)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  facet_wrap(~Gender, scales = "free_x") +
  labs(title = "Impact of Promotions on Average Spending by Age and Gender",
       x = "Age Group",
       y = "Average Spending",
       fill = "Promotion Used") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Line plot to show trends in median spending with promotions across age groups
ggplot(data = promo_influence, aes(x = Age_Group, y = Median, color = Promo.Code.Used, group = Promo.Cod
  geom_line() +
  geom_point() +
  facet_wrap(~Gender) +
  labs(title = "Trends in Median Spending by Age and Gender with Promotion Status",
       x = "Age Group",
       y = "Median Spending",
       color = "Promotion Status") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

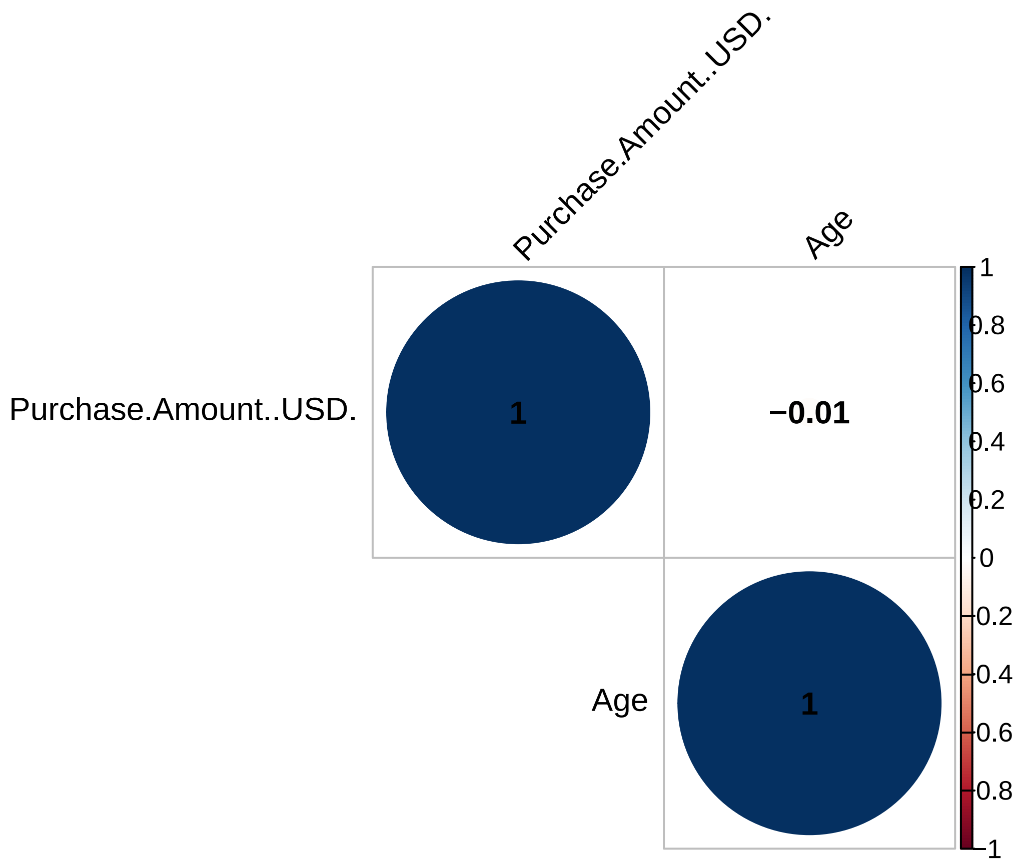


Q4: Future Spending Based on Loyalty

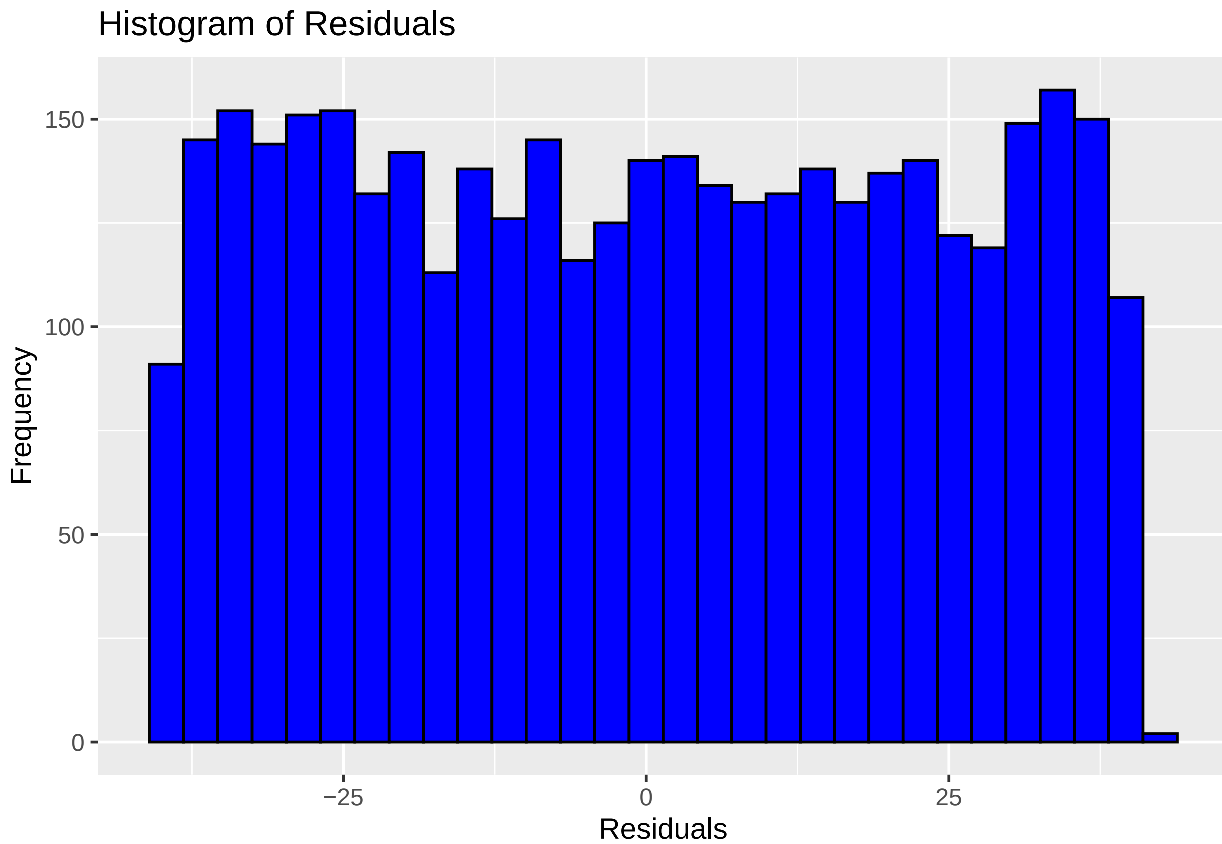
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

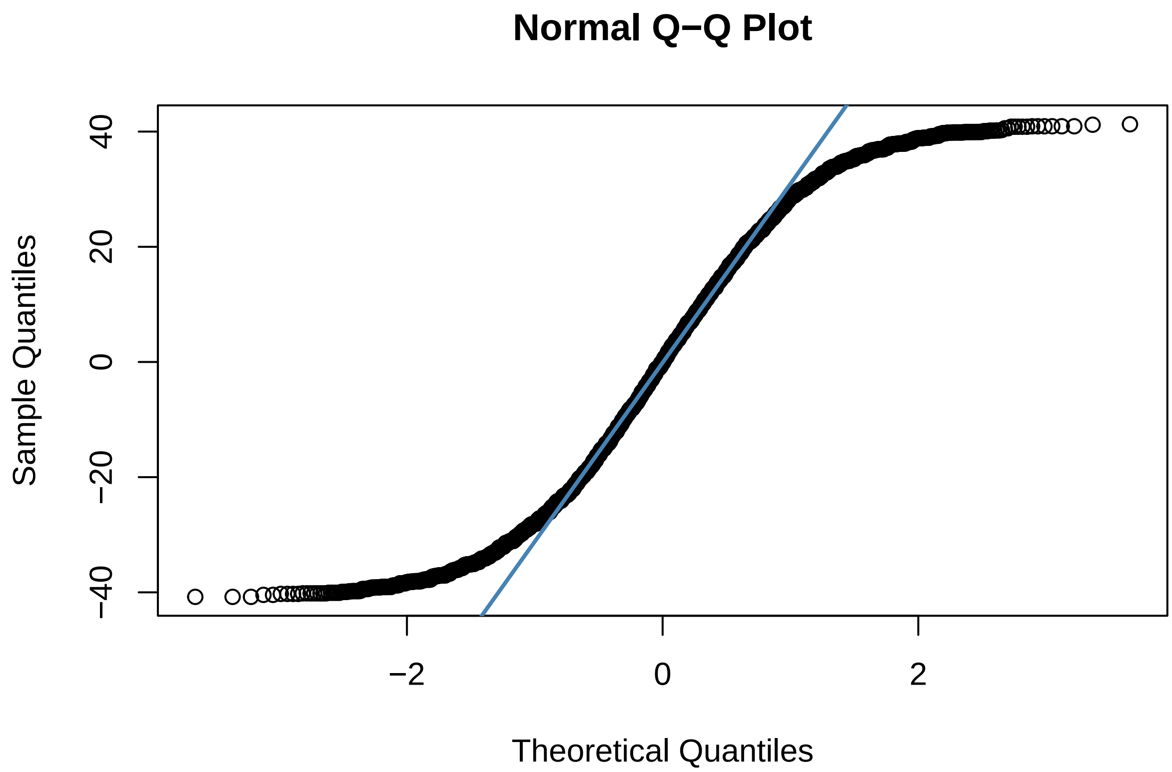
```
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```




```r
residuals_df <- data.frame(Residuals = residuals(model))

ggplot(residuals_df, aes(x = Residuals)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency")
```

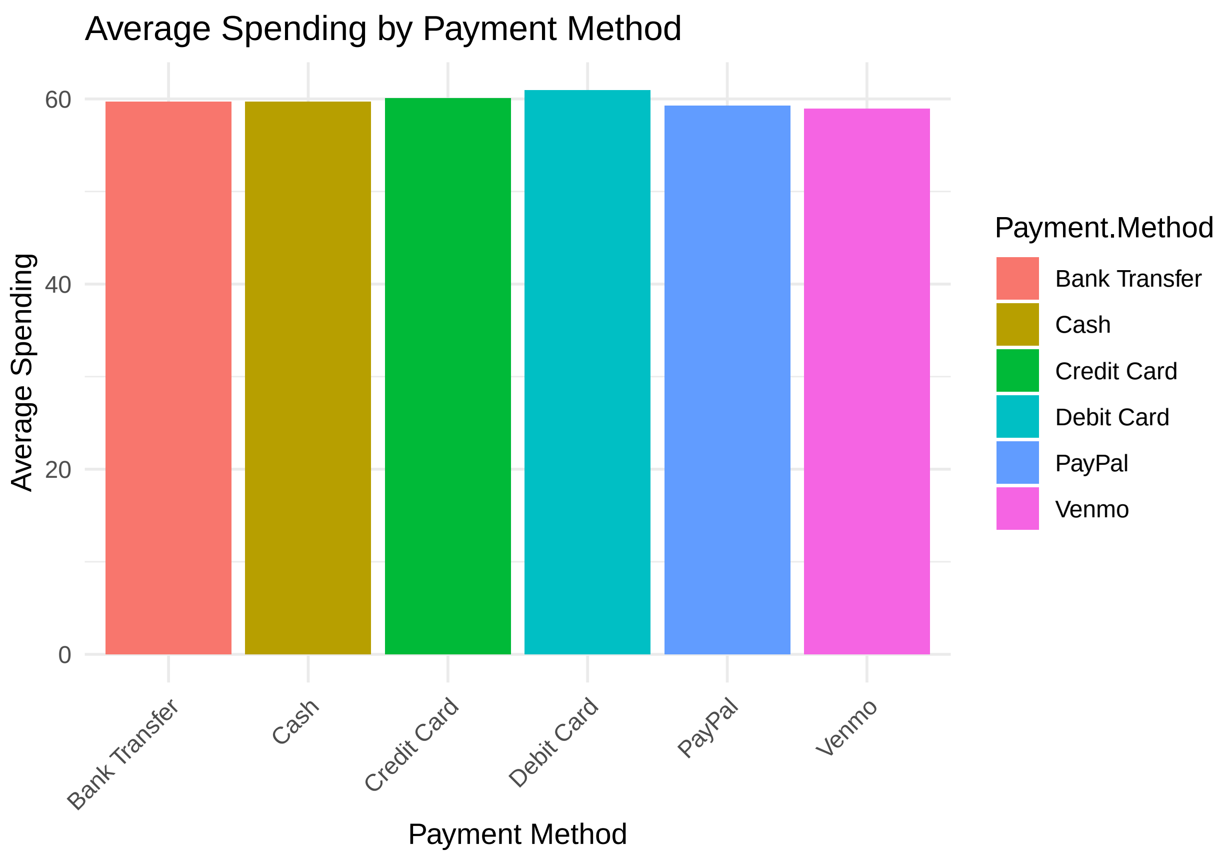



```r
# QQ plot of residuals
qqnorm(residuals(model))
qqline(residuals(model), col = "steelblue", lwd = 2)
```

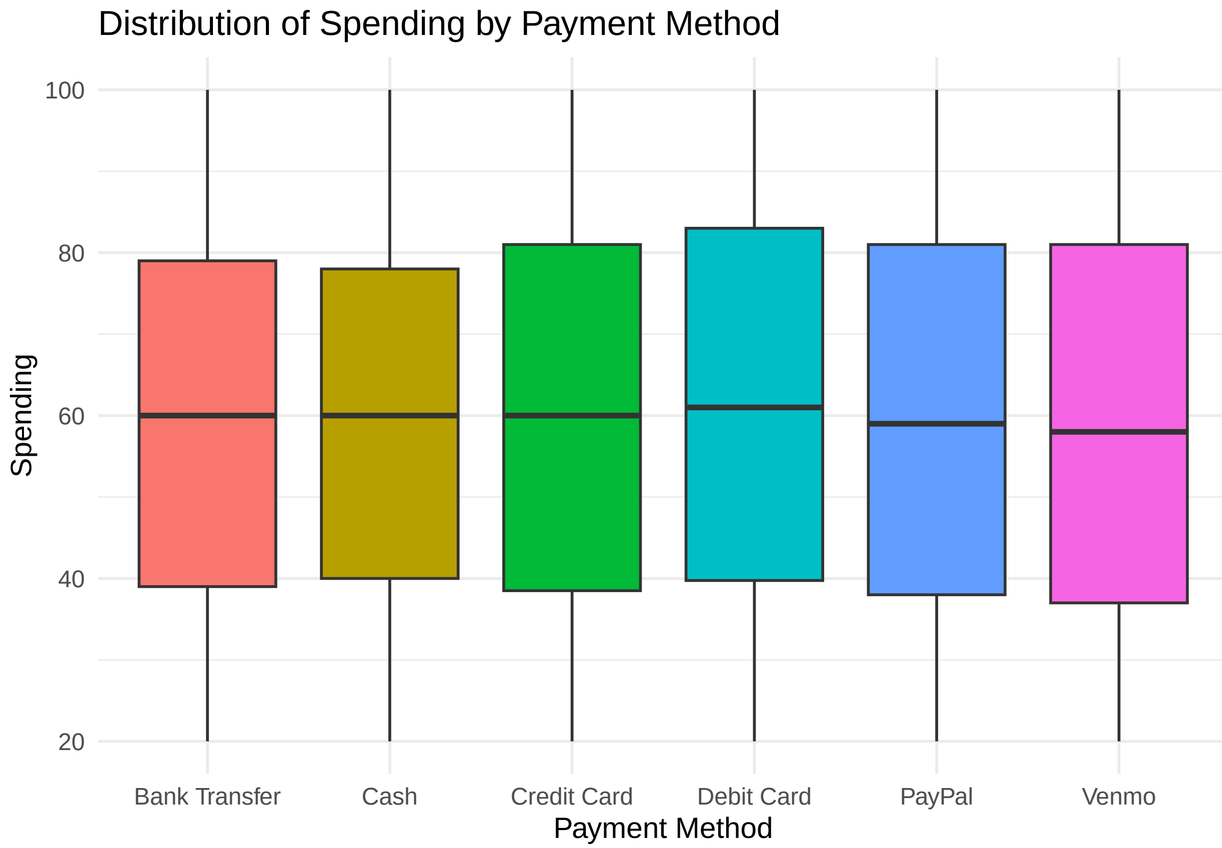


Q5: Influence of Payment Method on Spending

```
ggplot(average_spending_by_payment, aes(x = Payment.Method, y = Average_Spending, fill = Payment.Method
  geom_bar(stat = "identity") +
  labs(title = "Average Spending by Payment Method", x = "Payment Method", y = "Average Spending") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
ggplot(shopping_behavior, aes(x = Payment.Method, y = Purchase.Amount..USD., fill = Payment.Method)) +
  geom_boxplot() +
  labs(title = "Distribution of Spending by Payment Method", x = "Payment Method", y = "Spending") +
  theme_minimal() +
  theme(legend.position = "none")
```



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain

Incorporating machine learning could offer some huge inisghts especially when it comes to trying to predict future consumer behaviors and being able to enhance the precision of our analyses. We have tried to use some predictive modeling with regression analysis, we used the linear regression model in order to predict spending based on linear relationships like the impact of demographic factors or the pruchase frequency on spending. However, I do believe other techniques can be explored in order to better tackle these questions. Clustering/K-Clustering can be used to group consumers based on their similar criteria without the need of any predifned labels to understand patterns. Additionaly, a time series analysis could be utilized like ARIMA. This model could help to forecast future spending patterns especially in relation to the seasonal changes.

What questions do you have now, that will lead to further analysis or additional steps?

1. Segmentation Depth: Can deeper consumer segments/ more nuanced segments be identified that go beyond basic demographics, perhaps using advanced clustering techniques to reveal any patters or traits?
2. Marketing Response: Which consumer segments are most responsive to particular marketing promotions, and is there a model that can be developed to enhance the effectiveness of marketing strategies?

3. External Data: What external data can be brought in/joined in order to better make sense of the current data sets?
4. Product Preference: Which product categories are favored by different demographic groups, and how can these insights improve sales strategies?
5. Influence of Payment Methods: How do different payment methods impact loyalty? Would it be possible to predict these effects based on consumer behavior?
6. Enhancing Model Accuracy: What refinements can be made or implemented in our predictive models in order to improve their precision nd reliability in forecasting consumer spending patterns?

Milestone 3 - Final Analysis and Recommendations

Introduction

Building on our data analysis conducted in the steps above, this final section will give an overview of the insights that were derived from the data. We will summarize the projects scope, the methods used and the insights that were gained through our analysis.

Problem Statement Summary

Our initial goal from this project was to try and uncover the main factors which influenced consumer spending behaviors across different demographics, seasons, and promotional activities. By analyzing these variables, different businesses can use the insight to optimize their marketing strategies as well as improve sales efficiency.

Methodology Summary

1. Cleaning and reprocessing the data for accuracy and easier usability
2. Conducting statistical analysis to understand data distributions and their relationships
3. Applying the regression analyses and using machine learning techniques to help predict spending behaviors and analyze the impact of the different variables among one another.

Analysis Insights

1. The demographic factors like the location and the age impacted the spending habits significantly.
2. Changes in the season affect consumer spending, with some notable peaks during certain holidays.
3. Marketing promotions were effective in influencing the younger demographics.

Implications for Consumers

The analysis conducted provides the consumers with a better comprehension of how the external factors like the economic conditions or the marketing promotions could affect their spending behaviors. It also gives the businesses some data-driven insights that can be used to tailor their marketing approaches to consumer needs.

Limitations and Future Research

While the analysis done had some useful insights, it also had some limitations.

1. The predictive power of the models was very limited due to the unavailability of previous data, or lacking historical data.
2. The external factors, like economic downturns or unprecedented events (covid) were not fully taken into consideration.

In the future, it would be worth integrating additional data in order to refine the predictions and expand a little more on the impacts of external factors. We can also utilize some more advanced models or machine learning tools to enhance the accuracy of the predictive analytics.

Conclusion

This project has shown us the true power of data analysis in uncovering the hidden patters of consumer behavior. By proceeding to refine the data and the models that were used, businesses could really enhance their understanding of consumer needs which could lead to some more effective marketing strategies and improve their profitability as a whole.