Recent Customer Shopping Trends

Ava Cascario, Sydney Holt, and Kacie Rohn

2024-12-03

Research Topic: Latest Customer Shopping Trends

This research focuses on recent customer shopping trends, a topic that is both familiar and increasingly important to each member of our team. The rise of online shopping, particularly after the Covid-19 pandemic, combined with rapid technological advancements, has made this subject more relevant than ever. Our research will revolve around gaining a better understanding on what customers tend to purchase pertaining to gender, age, geographical location, season, item, price, and other features.Through a series of research questions and data visualizations, we will explore these relationships to uncover insights and draw connections between key attributes. Our goal is to contribute new knowledge to the reader and deepen our understanding of modern consumer behavior.

Research Questions

The first research question we will explore is, how do different demographics such as age, gender, location, and price affect the shopping trends of customers. We will create different visuals to present our findings and explain the correlation between each one of these features and customer shopping behavior. We are also curious to know, does gender have an affect on how much the customer spends, what item(s) they buy, and what reviews they left on the product. We predict that there will be large differences between the shopping trends of males versus the shopping trends of females, and intend to explore this further using multiple types of visualizations and tables. We must be aware of bias, as we are all females who experience the female shopping trends ourselves, and we cannot allow this to alter the conclusions we make. We are also curious on if the time of year (season) and geographical location of a customer changes what specific item they purchase. For example, does someone who is experiencing summer in Florida tend to buy something different from a customer experiencing winter in Maine? Overall, this is not an exhaustive list on what we intend to explore, as there are many different combinations of features that allow for different discoveries.

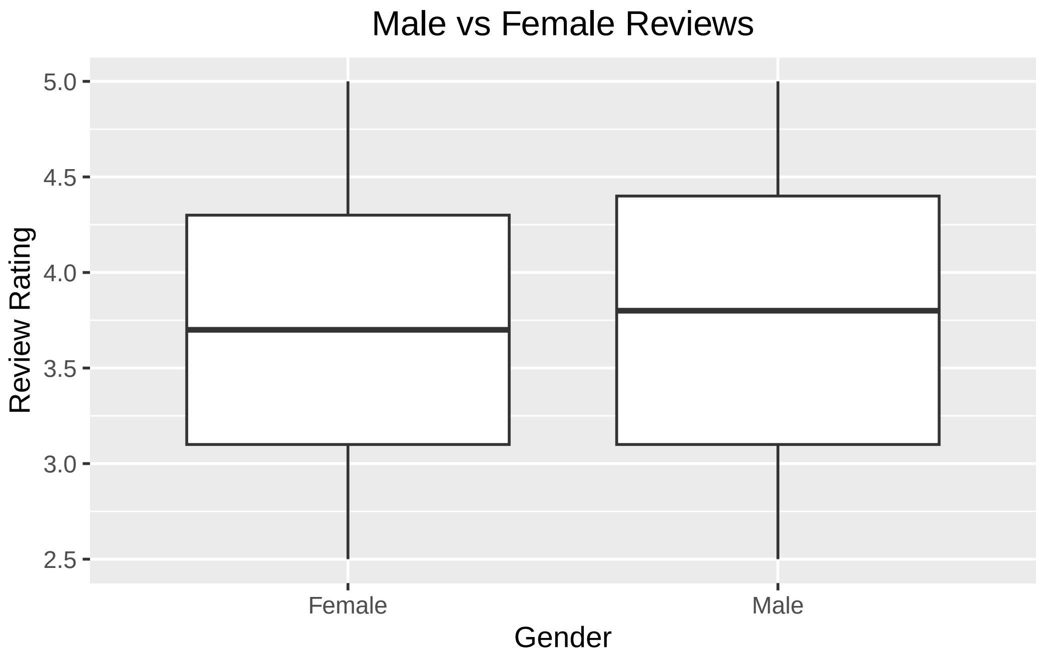
Provenance Of Our Data

We are utilizing a data set that we found on Kaggle. Kaggle is a website focused towards data scientists with a goal in helping others learn about data. The author of the data is Bhadra Mohit, and they describe it as offering a comprehensive view of consumer shopping trends, aiming to uncover patterns and behaviors in retail purchasing. It contains detailed transactional data across various product categories, customer demographics, and purchase channels. This data set was last updated 20 days ago, and is expected to be updated 4 times a year. This ensures that the data remains relevant and is as accurate as possible. In this data set, case is an individual transaction. This includes the attributes, customer ID, age, gender, item purchased, category, purchase amount USD, location, size, color, season, review rating, subscription status, payment method, shipping type, discount applies, promo code used, previous purchases, preferred payment method, and frequency of purchases. We intend to focus on age, gender, item purchased, location, season, review rating, and previous purchases. All of the attributes come from the data set, but we also created the attribute of region, which groups all fifty states into four regions: Northeast, South, Midwest, and West. This is based upon the national recognized regions in the United States.

FAIR Principles

The data we are utilizing meets the FAIR principles. The data is **findable**, and includes unique identifiers as well rich and substantial metadata. Each case is given an ID number, and there are numerous attributes that they are defined by. The data is **accessible** and can be found in our public repository. We also downloaded the data from Kaggle, which is public and we were able to easily locate it and access it. The data is **interoperable** and uses the R language which is widely known and accepted. This way our data can be exchanged between collaborators and allows for open communication. By citing our sources and explaining the provenance of our data, this ensure our data is also **reusable**. By meeting the FAIR principles we are ensuring that our data is universal, and can be easily understood.

Data Visualizations



`Figure One: Average Customer Reviews by Gender`

Figure One shows the average review ratings a customer leaves after making a purchase, by Gender. Each customer was able to leave a review rating between 0-5, and we took the average of both the female reviews, and the male reviews to graph. The results we found were different than our initial expectations. We can conclude from Figure One that males tend to leave slighter high reviews than women do. While we thought this to be true before graphing the data, we expected the results to be much more drastic. Females tend to leave, on average, a rating of approximately 3.7, while males leave approximately a 3.8. This is based on the grid lines and Y-axis scale. However, we found that there are also more men in our original data set which can lead to skewed and biased results. The unbalanced number of each gender could be what is making the males average appear higher than the females. Overall, figure one gives us an insight into the reviews typically left by each gender, and whether female or males tend to be "harsher" when reviewing a product. The use of a box plot appropriately displays the data and therefore the difference in average review ratings by gender can be easily compared.

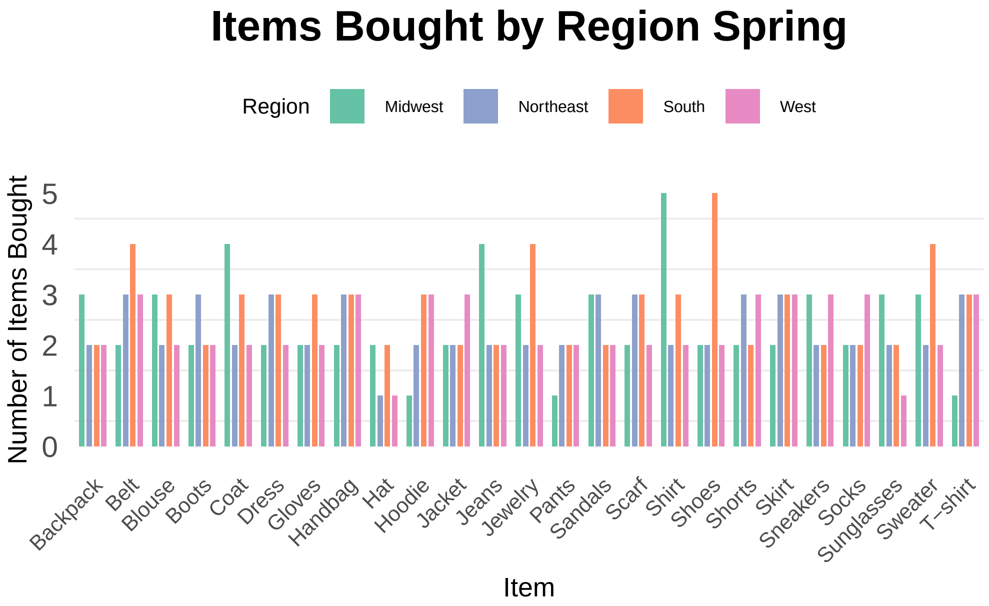
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59.8387097	0.6725026	88.9791512	0.0000000
categoryClothing	0.1866214	0.8804070	0.2119717	0.8321402
categoryFootwear	0.4167160	1.1783422	0.3536460	0.7236233
categoryOuterwear	-2.6658702	1.4775420	-1.8042602	0.0712677

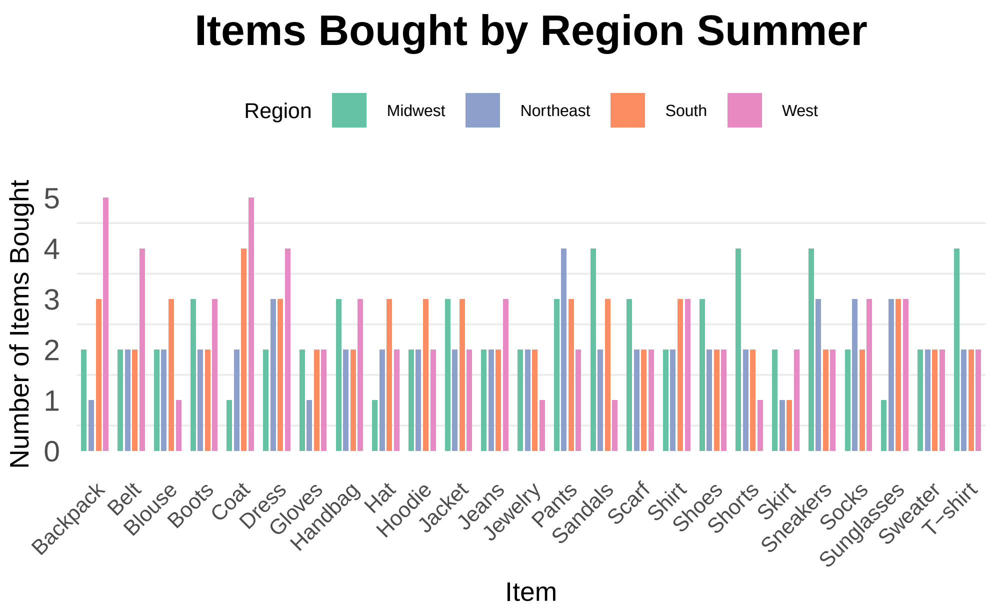
Another data point we wanted to research was the relationship between amount spent and what type of category of item it was. When looking at the regression summary we can see that Outerwear has the most affect on the regression line. This makes sense due to the fact that Outerwear tend to be more expensive items such as jackets and coats. Then Footwear which also tends to be expensive items. Finally Clothing affects the regression line least due to the fact that these items tend to be the least expensive.

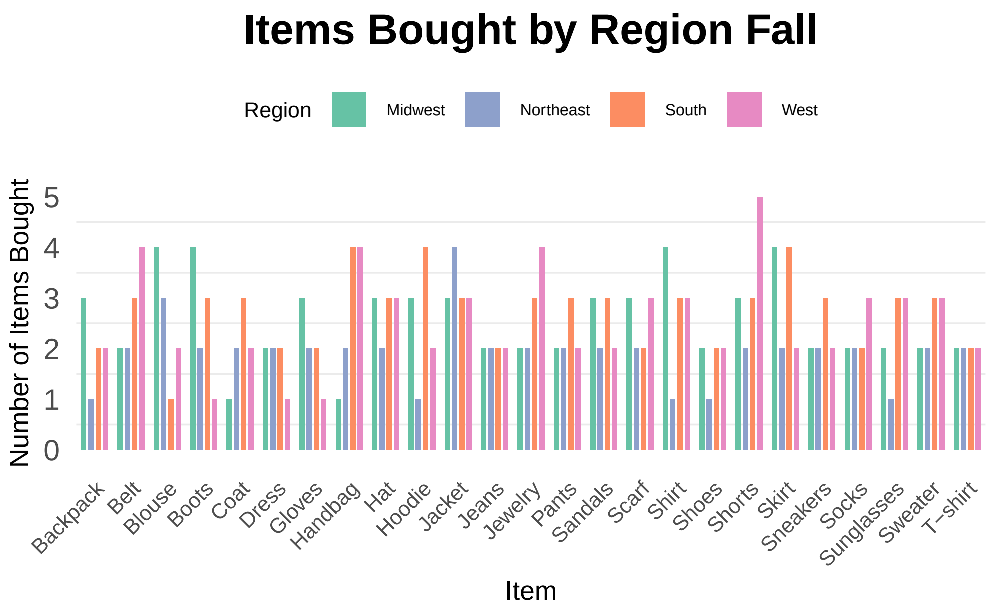
category	count	min	Q1	median	Q3	max	mad	mean	sd
Accessories	1240	20	39.75	60.0	80	100	29.6520	59.83871	23.30123
Clothing	1737	20	38.00	60.0	81	100	31.1346	60.02533	23.79246
Footwear	599	20	39.00	60.0	81	100	31.1346	60.25543	23.63844
Outerwear	324	20	34.00	54.5	80	100	33.3585	57.17284	24.59003

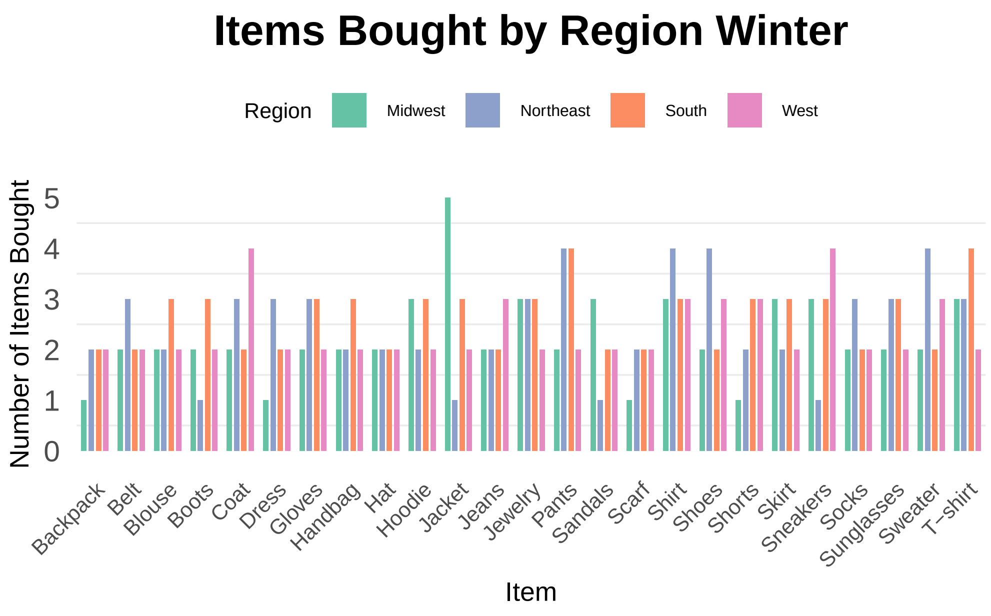
`Figure Two: Table Showing Purchase Amount by Clothing Category`

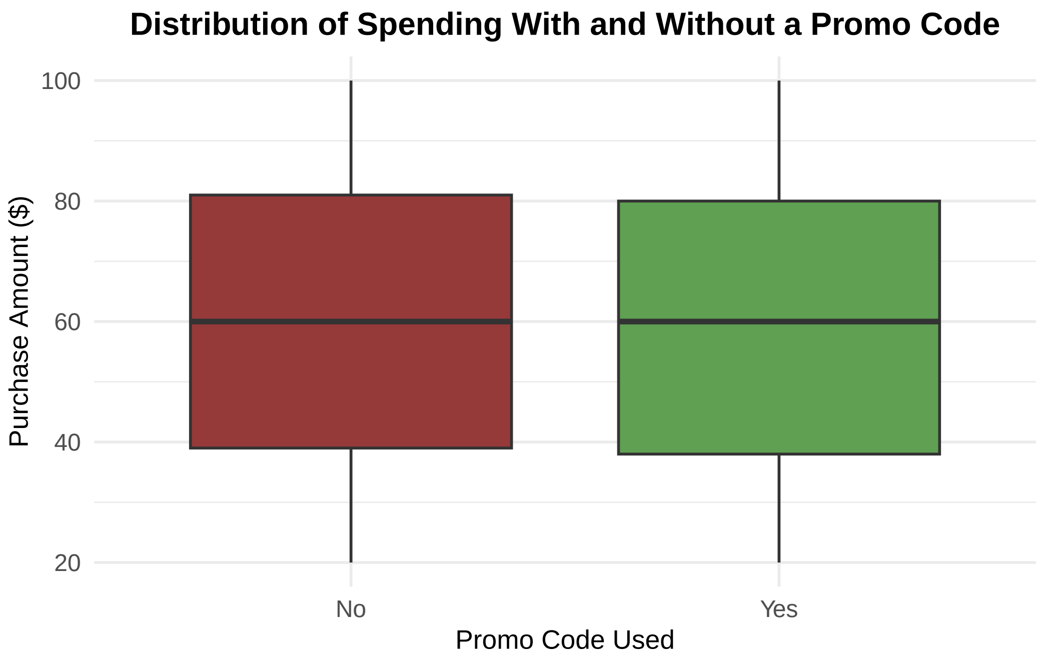
Figure Two shows a summary table including count, minimum, Q1, median, maximum, mean, and standard deviation of




Items Bought by Region Summer
Region
Midwest
Northeast
South
West
Number of Items Bought
0
1
2
3
4
5
Backpack
Belt
Blouse
Boots
Coat
Dress
Gloves
Handbag
Hat
Hoodie
Jacket
Jeans
Jewelry
Pants
Sandals
Scarf
Shirt
Shoes
Shorts
Skirt
Sneakers
Socks
Sunglasses
Sweater
T−shirt
Item


Items Bought by Region Fall
Region
Midwest
Northeast
South
West
Number of Items Bought
0
1
2
3
4
5
Backpack
Belt
Blouse
Boots
Coat
Dress
Gloves
Handbag
Hat
Hoodie
Jacket
Jeans
Jewelry
Pants
Sandals
Scarf
Shirt
Shoes
Shorts
Skirt
Sneakers
Socks
Sunglasses
Sweater
T−shirt
Item

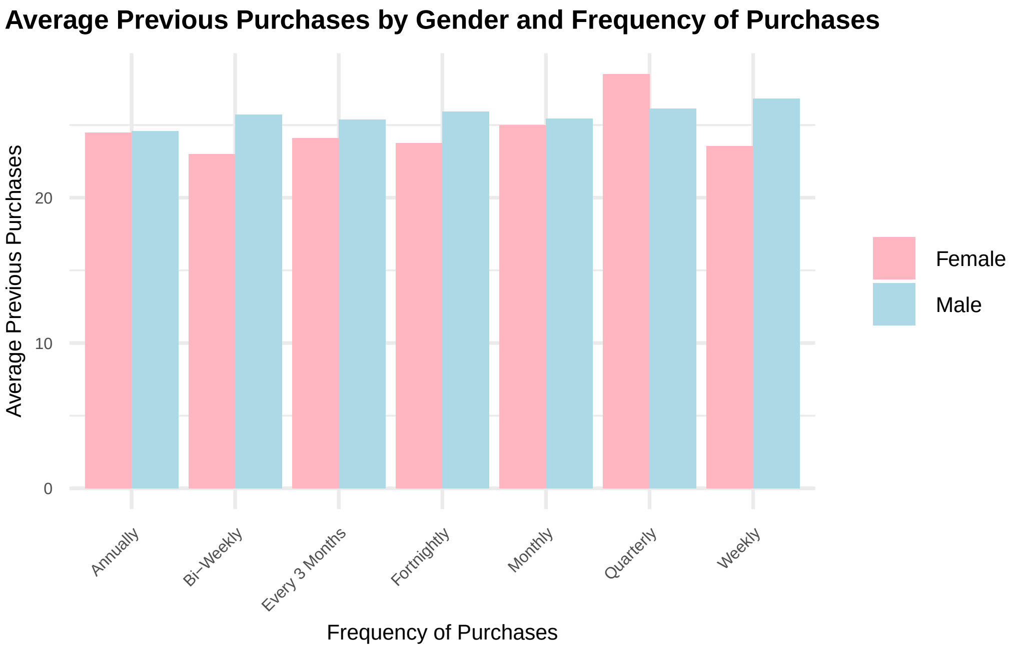




```r
# Load necessary packages ----
library(ggplot2)
library(dplyr)
library(knitr)
library(tinytex)
# Load shopping trends dataset ----
shopping_trends_raw <- read.csv(
  file = "shopping_trends.csv",
  header = TRUE,
  sep = ","
)
# Clean column names for consistency ----
shopping_trends_clean <- shopping_trends_raw %>%
  rename(
    customer_id = "Customer.ID",
    age = "Age",
    gender = "Gender",
    item_purchased = "Item.Purchased",
    category = "Category",
    purchase_amount_usd = "Purchase.Amount..USD.",
    location = "Location",
    size = "Size",
    color = "Color",
    season = "Season",
    review_rating = "Review.Rating",
    subscription_status = "Subscription.Status",
    payment_method = "Payment.Method",
    shipping_type = "Shipping.Type",
    discount_applied = "Discount.Applied",
    promo_code_used = "Promo.Code.Used",
    previous_purchases = "Previous.Purchases",
    preferred_payment_method = "Preferred.Payment.Method",
    frequency_of_purchases = "Frequency.of.Purchases"
  )

# Group data by gender and review rating ----
type_reviews <- shopping_trends_clean %>%
  select(
    gender,
    review_rating
  ) %>%
  group_by(
    gender,
    review_rating
  )
# Boxplot of review ratings by gender ----
ggplot(
  data = type_reviews,
  aes(
    x = gender,
    y = review_rating
  )
) +
  geom_boxplot()+
  labs(
    x = "Gender",
    y = "Review Rating",
    title = "Male vs Female Reviews"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
# Display regression summary for purchase amount by category ----
shopping_summary <- lm(formula = purchase_amount_usd ~ category, data = shopping_trends_clean)
shopping_summary_model <- summary(shopping_summary)

shopping_summary_model$coefficients %>%
  knitr::kable()
# Summary statistics for purchase amount by category ----
shopping_summary <- shopping_trends_clean %>%
  select(category, purchase_amount_usd) %>%
  group_by(category) %>%
  summarize(
    count = n(),
    min = min(purchase_amount_usd),
    Q1 = quantile(purchase_amount_usd, 0.25),
    median = median(purchase_amount_usd),
    Q3 = quantile(purchase_amount_usd, 0.75),
    max = max(purchase_amount_usd),
    mad = mad(purchase_amount_usd),
    mean = mean(purchase_amount_usd),
    sd = sd(purchase_amount_usd)
  )

shopping_summary %>%
  knitr::kable()
# Group data by location, season, and item ----
item_purchased_data <- shopping_trends_clean %>%
  group_by(
    location,
    season,
    item_purchased
  ) %>%
  summarize(
    item_count = n(),
    .groups = "drop"
  )
# Map U.S. states to regions ----
# Group States by Region
state_to_region <- c(
  "Maine" = "Northeast",
  "New Hampshire" = "Northeast",
  "Vermont" = "Northeast",
  "Massachusetts" = "Northeast",
  "Rhode Island" = "Northeast",
  "Connecticut" = "Northeast",
  "New York" = "Northeast",
  "New Jersey" = "Northeast",
  "Pennsylvania" = "Northeast",
  "Delaware" = "South",
  "Maryland" = "South",
  "Virginia" = "South",
  "North Carolina" = "South",
  "South Carolina" = "South",
  "Georgia" = "South",
  "Florida" = "South",
  "West Virginia" = "South",
  "Kentucky" = "South",
  "Tennessee" = "South",
  "Alabama" = "South",
  "Mississippi" = "South",
  "Arkansas" = "South",
  "Louisiana" = "South",
  "Oklahoma" = "South",
  "Texas" = "South",
  "Indiana" = "Midwest",
  "Illinois" = "Midwest",
  "Michigan" = "Midwest",
  "Ohio" = "Midwest",
  "Wisconsin" = "Midwest",
  "Missouri" = "Midwest",
  "Iowa" = "Midwest",
  "Minnesota" = "Midwest",
  "North Dakota" = "Midwest",
  "South Dakota" = "Midwest",
  "Nebraska" = "Midwest",
  "Kansas" = "Midwest",
  "Montana" = "West",
  "Wyoming" = "West",
  "Colorado" = "West",
  "Idaho" = "West",
  "Nevada" = "West",
  "Utah" = "West",
  "Arizona" = "West",
  "New Mexico" = "West",
  "Washington" = "West",
  "Oregon" = "West",
  "California" = "West",
  "Alaska" = "West",
  "Hawaii" = "West"
```

```
)
# Add region column based on location ----
item_purchased_data <- item_purchased_data %>%
  mutate(
    region = state_to_region[location]
  )
# Filter data by season ----
spring_data <- item_purchased_data %>%
  filter(season == "Spring")

summer_data <- item_purchased_data %>%
  filter(season == "Summer")

fall_data <- item_purchased_data %>%
  filter(season == "Fall")

winter_data <- item_purchased_data %>%
  filter(season == "Winter")
# Visualize data by region and season ----
plot_items_by_season <- function(
    item_purchased_data,
    season_name
    ) {
ggplot(
  item_purchased_data,
  aes(
    x = item_purchased,
    y = item_count,
    fill = region
  )
) +
  geom_bar(
    stat = "identity",
    position = position_dodge(width = 0.8),
    width = 0.6
  ) +
  scale_x_discrete(
    expand = expansion(add = c(0.5, 0.5))
  ) +
  labs(
    title = paste(
      "Items Bought by Region",
      season_name
    ),
    x = "Item",
    y = "Number of Items Bought",
    fill = "Region"
  ) +
  scale_fill_manual(
    values = c(
      "Northeast" = "#8DA0CB",
      "South" = "#FC8D62",
      "Midwest" = "#66C2A5",
      "West" = "#E78AC3"
    )
  ) +
  theme_minimal(base_size = 14) +
  theme(
    panel.grid.major = element_blank(),
    axis.text.x = element_text(
      angle = 45,
      hjust = 1,
      size = 8
    ),
    axis.title.x = element_text(size = 10),
    axis.title.y = element_text(size = 10),
    legend.position = "top",
    legend.text = element_text(size = 6),
    legend.title = element_text(size = 8),
    legend.key.size = unit(0.5, "cm"),
    plot.title = element_text(
      size = 16,
      hjust = 0.5,
      face = "bold"
    ),
    plot.margin = margin(15, 15, 15, 15)
  )
}
# Generate and display plots for each season ----
spring_plot <- plot_items_by_season(spring_data, "Spring")
summer_plot <- plot_items_by_season(summer_data, "Summer")
fall_plot <- plot_items_by_season(fall_data, "Fall")
winter_plot <- plot_items_by_season(winter_data, "Winter")

print(spring_plot)
print(summer_plot)
print(fall_plot)
print(winter_plot)

# Visualize purchase amount with and without discount codes ----
ggplot(
  shopping_trends_clean,
  aes(
    x = promo_code_used,
    y = purchase_amount_usd,
    fill = promo_code_used
  )
) +
  geom_boxplot(
   outlier.color = "black",
   outlier.size = 2
  ) +
  labs(
    title = "Distribution of Spending With and Without a Promo Code",
    x = "Promo Code Used",
    y = "Purchase Amount ($)"
  ) +
  scale_fill_manual(
    values = c(
      "Yes" = "#5fa052",
      "No" = "#963939"
    )
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    plot.title = element_text(
      hjust = 0.5,
      face = "bold",
      size = 12
    ),
    axis.title = element_text(size = 10)
  )

While we assumed there would be a greater difference between amount spent by customers when the

# Summarize data by gender and frequency of purchases ----
summary_by_gender_freq <- shopping_trends_clean %>%
  group_by(
    gender,
    frequency_of_purchases
  ) %>%
  summarize(
    average_previous_purchases = mean(previous_purchases),
    .groups = "drop"
  )
# Filter out N/A frequency of purchases values ----
summary_by_gender_freq_filtered <- summary_by_gender_freq %>%
  filter(frequency_of_purchases != "n/a")
# Visualize average previous purchases by gender and frequency of purchases ----
ggplot(
```

```r
  summary_by_gender_freq_filtered,
  aes(
    x = frequency_of_purchases,
    y = average_previous_purchases,
    fill = gender
  )
) +
  geom_bar(
    stat = "identity",
    position = position_dodge()
  ) +
  labs(
    title = "Average Previous Purchases by Gender and Frequency of Purchases",
    x = "Frequency of Purchases",
    y = "Average Previous Purchases"
  ) +
  scale_fill_manual(
    values = c(
      "Female" = "#FFB6C1",
      "Male" = "#ADD8E6"
    )
  ) +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(
      angle = 45,
      hjust = 1,
      size = 6
    ),
    axis.text.y = element_text(size = 6),
    legend.title = element_blank(),
    legend.text = element_text(size = 8),
    plot.title = element_text(
      size = 10,
      hjust = 0.5,
      face = "bold"
    ),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8)
  )
```


Average Previous Purchases by Gender and Frequency of Purchases
Average Previous Purchases
0
10
20
Annually
Bi−Weekly
Every 3 Months
Fortnightly
Monthly
Quarterly
Weekly
Frequency of Purchases
Female
Male