Recent Customer Shopping Trends

Ava Cascario, Sydney Holt, and Kacie Rohn

2024-12-18

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 Exploration

Before creating any visualizations, we created a table of several descriptive statistics of customers' purchase amounts by the category of the item they purchased. Table 1 shows a summary table including *count*, *minimum*, *Q1*, *median*, *Q3*, *maximum*, *median absolute deviation*, *mean*, and *standard deviation* of item category and purchase amount. Each item in the original data set was categorizes into one of the four categories. We can utilize this table to not only see simple statistics, but also the spread and deviations of the data. We were interested in gaining a better understanding of customer spending trends, and therefore we chose this summary table.

Table 1: Summary Statistics on Purchase Amount by Category

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

From the table we can make numerous conclusions. Looking at the *count* column, we can see that the Clothing category has the highest *count*, followed by Accessories, Footwear, and Outerwear respectively. This means that out of the entire data set, items that fall under the clothing category are the most popular and most frequently bought. The *minimum* and *maximum* price for each item are the same, 20 and 100 respectively. This could mean that the original data set only chose to collect information on items within that price range. We found it interesting that while Accessories, Clothing, and Footwear all have the same median of 60, Outerwear has a lower median of just 54.5. This could mean that the middle price of outerwear is generally lower than the other categories, which was surprising as we expected heavy items like boots and coats to be more expensive. This can also be shown through the Outerwear *mean* which is 57.17284, slightly lower than the other categories.

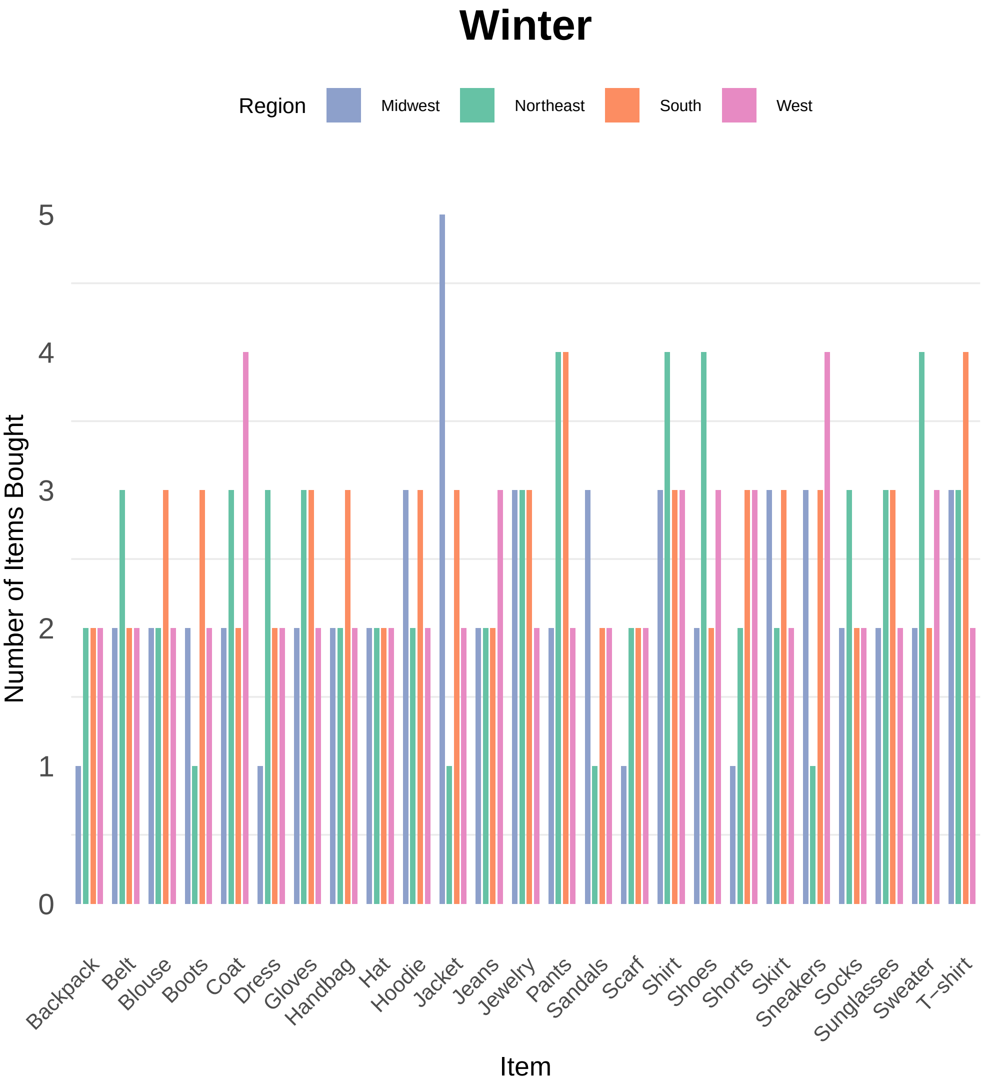
Overall, by looking at the table and each statistic, we can learn about different features and make different conclusions about our data set. Therefore, we introduced the table at the beginning of our data visualizations to give a general understanding of our data.

Data Visualizations

The first question that we chose to explore was how different seasons and geographical locations play a part in what items customers are buying. By separating each of the fifty states into their geographical regions, Midwest, Northeast, South, and West, we were able to condense the data and present the information in an easier-to-understand format.

If we compare all four graphs to one another, we are able to see how item popularity changes by both season and region, allowing us to gain a better understanding on clothing trends depending on time of year and location. We can also see what items are bought least and most frequently.

Figure 1: Popularity of Items in Winter by Region



As we looked at Figure 1, we noted a spike in backpack purchases in the Western region. We assumed that this may have been a cause of back-to-school sales. We also expected to see more spikes in things like t-shirts, skirts, sunglasses, and sandals in the South region due to high temperatures. However, there was not much of an increase in number of these items bought at all. The most striking trend we noticed, was a huge increase in coat purchases in the both the Western and Southern regions This contradicts our hypothesis that coats would be purchased much less frequently due to warm temperatures. The Midwest best represents what we expected to see: a spike in sandals, shorts, sneakers, and t-shirts.

Figure 2: Popularity of Items in Spring by Region

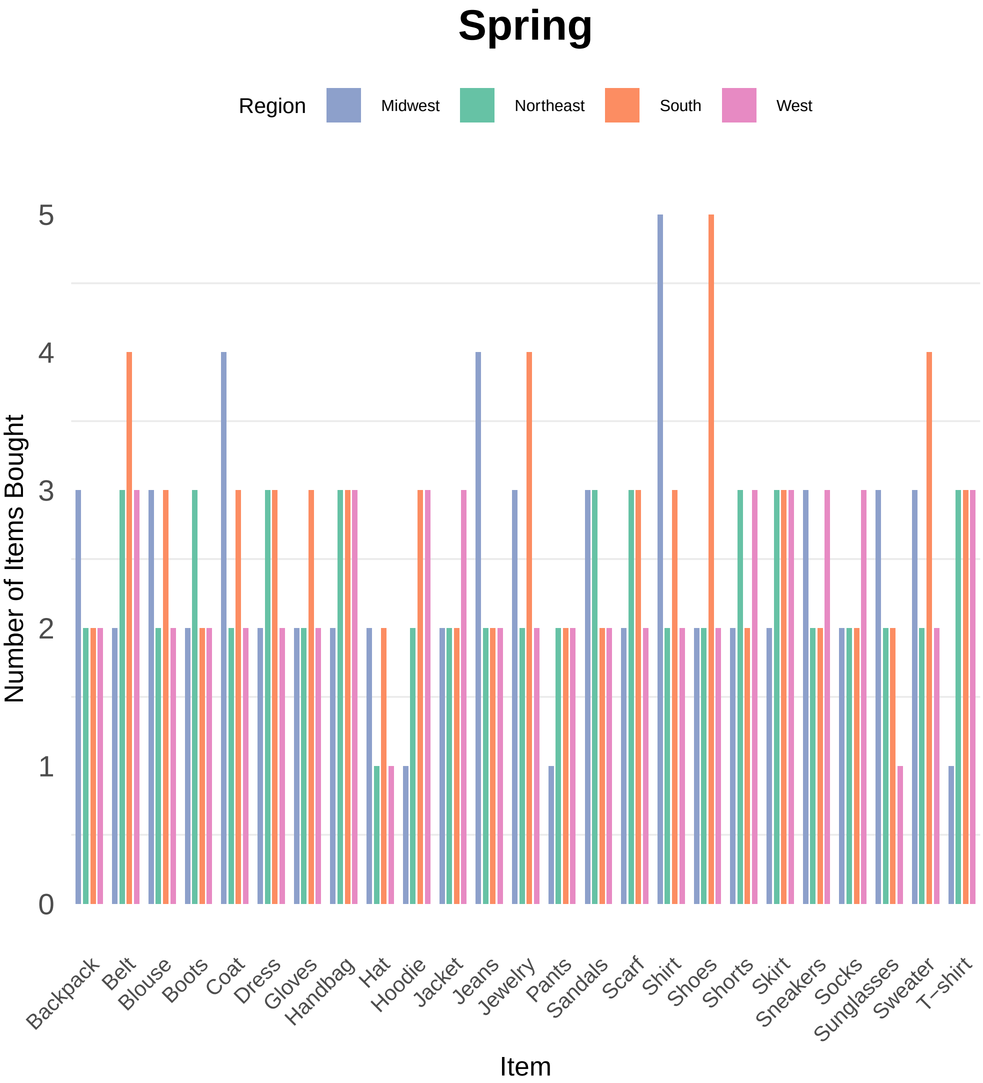
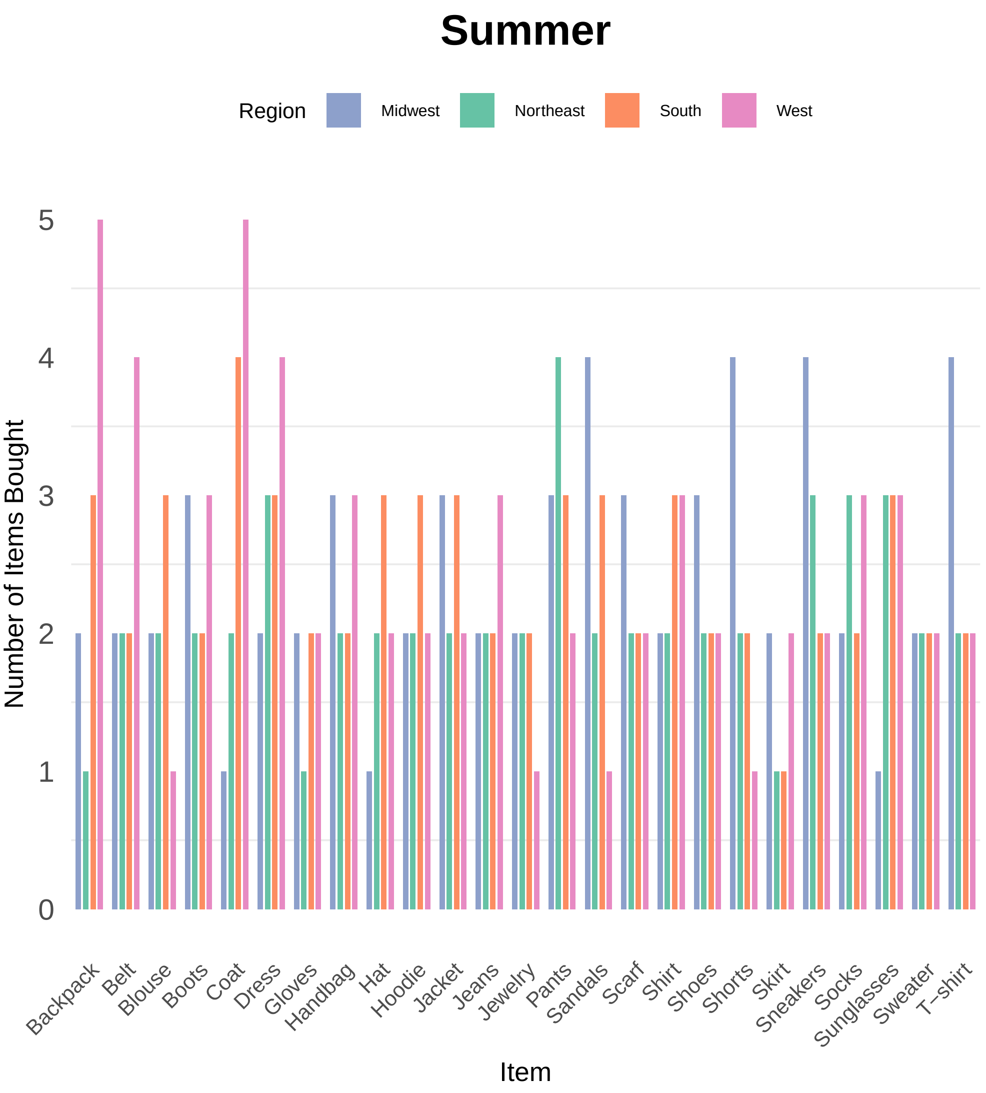


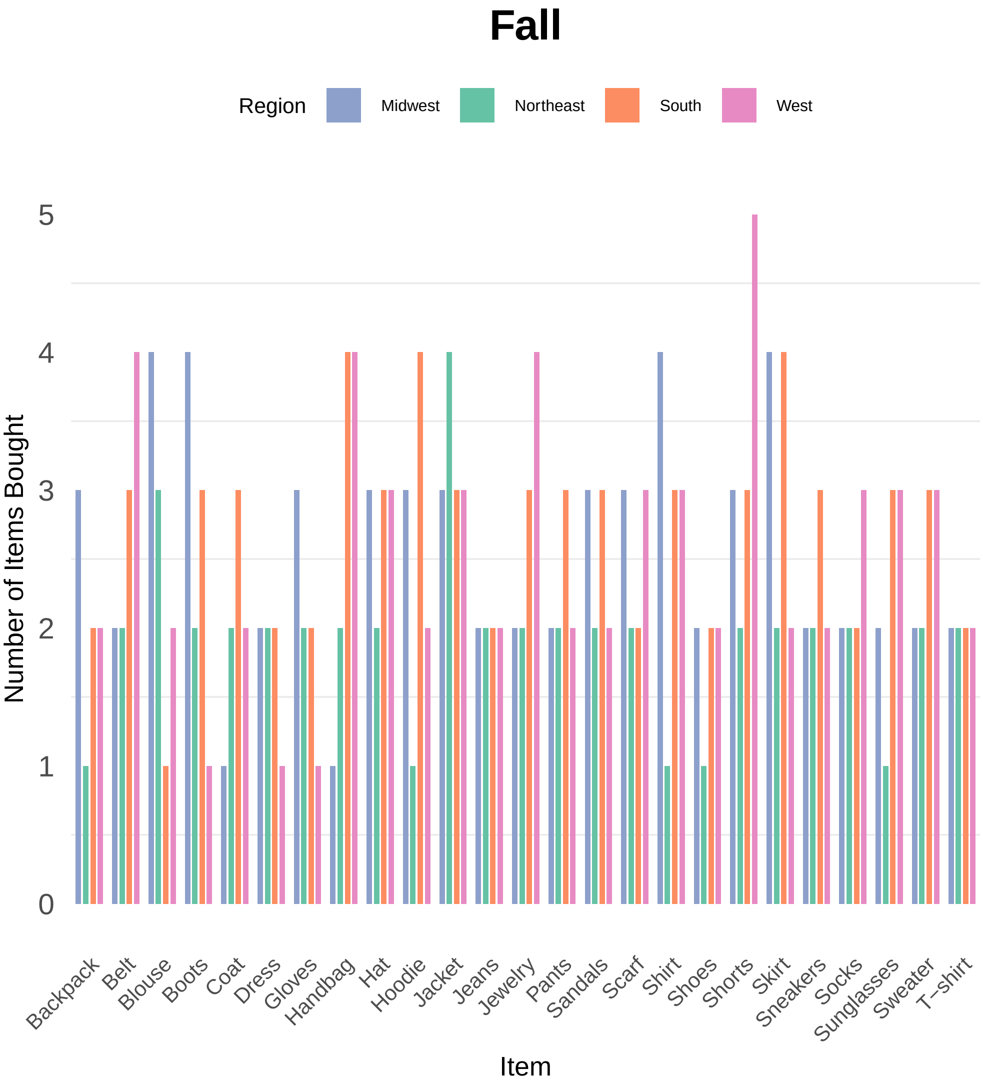
Figure 2 shows a spike in popularity of belts, jewelry, shoes, and sweaters in the South. This could be due to many factors, but we could conclude that people in the south tend to purchase new shoes right before summer. We were surprised to see a spike in sweater purchases, as the weather in the south tends to warm up in the spring. We can see that each region has different purchasing patterns in the spring; however, we noticed a low number of hat purchases across the board. The longer that we studied the graph, the more differences we noticed in purchasing patterns, both expected and surprising.

Figure 3: Popularity of Items in Summer by Region



Moving on to Figure 3, we were very surprised to see such a large number of shorts purchased in the fall, for temperatures typically begin to cool down at that point in the year. However, we concluded that this also could be due to back-to-school season and temperatures remaining warmer for longer due to climate change. We have noticed throughout each figure thus far that the typical number of items bought from each region is 2, which we found interesting. This could be due simply to the nature of our data set.

Figure 4: Popularity of Items in Fall by Region



Finally, looking at Figure 4, we see that in the Midwest there is a large number of jackets bought in Winter. This is consistent with our original predictions, as temperatures begin to cool down especially in the Midwest. However, we did think there would be a higher number of items such as boots, gloves, scarves, and coats.

Overall, we gained a lot of information looking at each region and have a better understanding of how location and time of year affects customer purchases. This allowed us to also look at our data set from a new angle, and the results we got were very interesting!

After seeing what everyone around the world typically buys, we wanted to explore whether their use of promo codes has an effect on their spending amounts. Our dataset displayed customers responses as to whether or not they used a promo code on their purchases as “Yes” or “No.”

Figure 5: Distribution of Spending With and Without a Promo Code

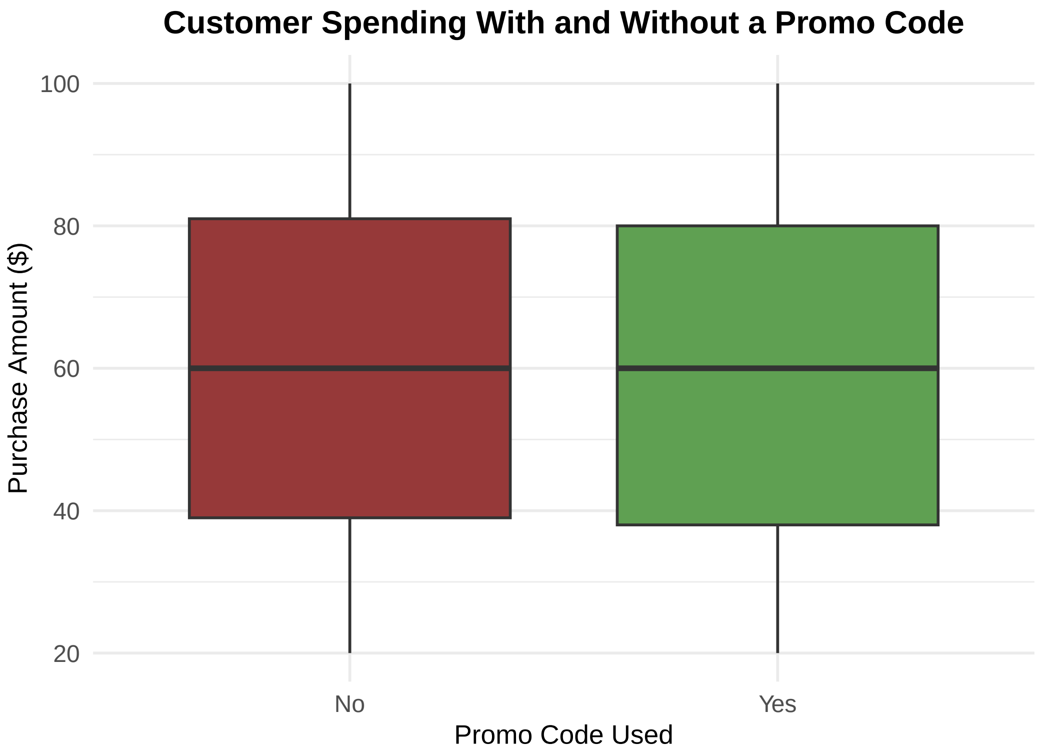
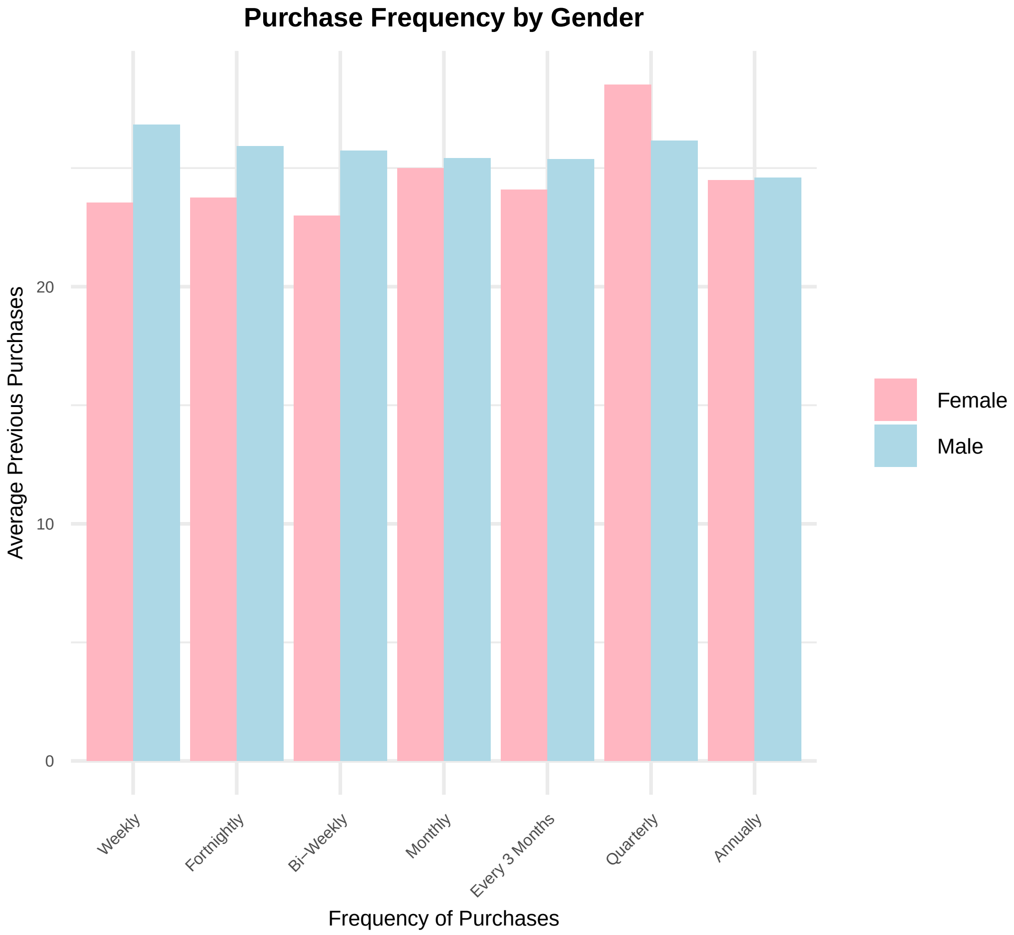


Figure 5 shows the distribution of customer spending by customers who use promo codes and those who do not. Their distributions are compared in terms of their purchase amounts. The box plot clearly conveys the spread and variation in spending for each group, with the box representing the spread between *Q1* and *Q2* and the *median* marked inside each box.

There are obvious confounding variables that can influence the price of items purchased by customers, such as the item itself, the location of the purchase, or the quality of the product, promo codes are expected to lessen the price of an item. Consequently, we assumed there would be a apparent difference among customers who did use a promo code versus those who did not. However, from Figure 5, we can see that there is very little variation, with the *median* purchase amount for both groups being about $60 and the overall distribution of spending looking quite similar. Both the green box (for promo code users) and the red box (for non-promo code users) are almost aligned, indicating that the average spending amounts are very comparable. There is only slight variation shown between the quartiles of each group. While the upper quartile of the group that used a promo code was at about $80, that of the group that did not use a promo code was only slightly higher. Overall, we assumed the variation among groups in Figure 5 would be much greater, and we can conclude that the use of promo codes may not substantially alter customer spending habits.

The next thing we chose to visualize was the distribution of purchase frequency by gender and the average number of previous purchases. We created a bar chart to effectively compare how often customers shop based on both their gender and how much they have shopped in the past for each frequency group: Weekly, Fortnightly, Bi-weekly, Monthly, Every 3 Months, Quarterly, and Annually.

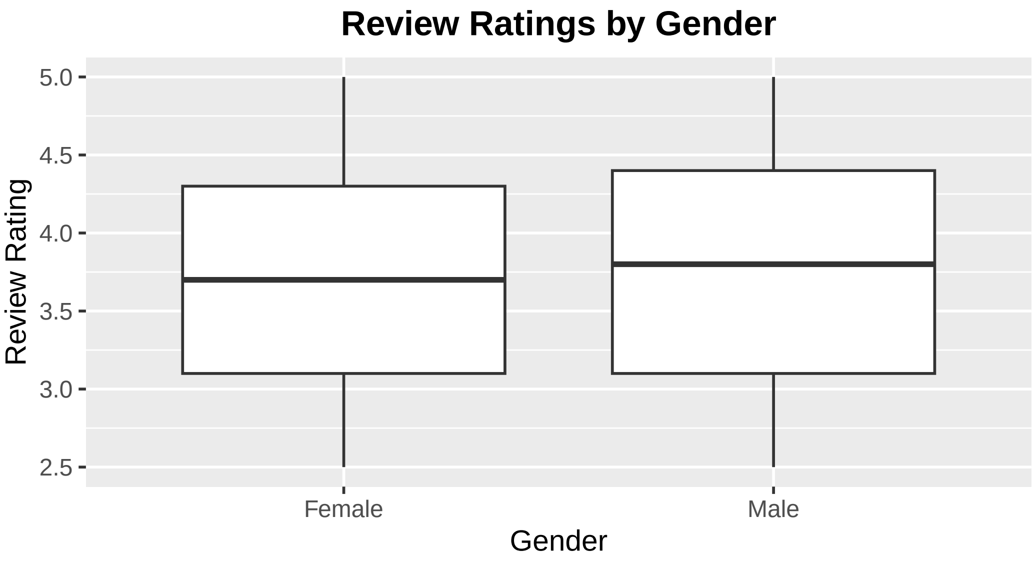
Figure 6: Distribution of Purchase Frequency by Gender



Looking at Figure 6, we see that there is little variation in the shopping frequency between genders. Both males and females tend to have similar shopping patterns, regardless of the frequency with which they shop. Although, after taking a closer look at our dataset, we saw a large variation among the previous purchases of customers, so it would make sense that the bars all appear to be around about 25 previous purchases per gender and frequency pair. The main observation we made from this figure was the spike in previous purchases among females who shop quarterly, or four times a year. It was interesting to see that this combination would have an effect, although it was minimal, on how often these certain females shopped in the past. Our assumption was that the customers that shopped the most frequently would have the highest number of previous purchases. Our assumption that there would be a large variation among male and female shopping trends was also disproven by this visual.

Despite slight differences like this in some categories, the gender-based distinctions are minimal. Thus, we can conclude from Figure 6 that shopping frequency and the number of previous purchases are generally consistent between genders, without any significant disparities.

Figure 7: Distribution of Review Ratings by Gender



Lastly, Figure 7 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 7 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 7 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.

Conclusion

Our analysis of recent customer shopping trends opened our eyes to our misconceptions regarding shopping. While we were all familiar and interested in the topic of shopping before, we now know so much more about the online shopping trends of those around us. It was especially interesting for us to see the tendencies of customers all over the world. Visualizing and understanding trends like the ones we uncovered is not only interesting, but it can be a useful tool for others to visualize. For example, it would be beneficial for businesses to understand how frequency and demographics intersect in order to effectively meet the needs of their customers for the growth of their businesses. Overall, our findings underscore the importance of data exploration for the enhancement of decision-making and simply the general knowledge of the public.

Works Cited

Disease Control, Centers for, and Prevention. *Geographic Regions*. 2024, https://www.cdc.gov/nchs/hus/sources-definitions/geographic-region.htm#:~:text=Geographic%20region,Ohio%2C%20South%20Dakota%2C%20and%20Wisconsin.

Mohit, Bhadra. "Customer Shopping (Latest Trends) Dataset." *Kaggle*, Nov. 2024, https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset.

Code Appendix

```
# Load necessary packages ----
library(ggplot2)
library(dplyr)
library(knitr)
library(tinytex)
library(tidyverse)
# Tidyverse style guide used

# 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"
  )

# Make summary table 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),
    Q1 = quantile(purchase_amount_usd, 0.75),
    max = max(purchase_amount_usd),
    mad = mad(purchase_amount_usd),
    mean = mean(purchase_amount_usd),
    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 ----
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",
```

```r
  "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(
      season_name
    ),
    x = "Item",
    y = "Number of Items Bought",
    fill = "Region"
  ) +
  scale_fill_manual(
    values = c(
      "Northeast" = "#66C2A5",
      "South" = "#FC8D62",
      "Midwest" = "#8DA0CB",
```

```r
      "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)
  )
}

winter_plot <- plot_items_by_season(winter_data, "Winter")
print(winter_plot)

spring_plot <- plot_items_by_season(spring_data, "Spring")
print(spring_plot)

summer_plot <- plot_items_by_season(summer_data, "Summer")
print(summer_plot)

fall_plot <- plot_items_by_season(fall_data, "Fall")
print(fall_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,
    table =
  )
```

```
) +
  geom_boxplot(
   outlier.color = "black",
   outlier.size = 2
  ) +
  labs(
    x = "Promo Code Used",
    y = "Purchase Amount ($)",
    title = "Customer Spending With and Without a Promo Code"
  ) +
  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)
  )

# 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 != "NA")
# Visualize average previous purchases by gender and frequency of purchases ----
ggplot(
  summary_by_gender_freq_filtered,
  aes(
    x = factor(
      frequency_of_purchases,
      levels = c(
        "Weekly",
        "Fortnightly",
        "Bi-Weekly",
        "Monthly",
        "Every 3 Months",
        "Quarterly",
        "Annually"
      )
    ),
    y = average_previous_purchases,
    fill = gender
  )
) +
  geom_bar(
    stat = "identity",
    position = position_dodge()
  ) +
  labs(
    x = "Frequency of Purchases",
    y = "Average Previous Purchases",
    title = "Purchase Frequency by Gender"
  ) +
  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)
  )
```

```
# Group data by gender and review rating ----
type_reviews <- shopping_trends_clean %>%
  select(
    gender,
    review_rating
  ) %>%
  group_by(
    gender,
    review_rating
  )

# Make 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 = "Review Ratings by Gender"
  ) +
  theme(
    plot.title = element_text(
      hjust = 0.5,
      face = "bold"
    )
  )
```