

# Recent Customer Shopping Trends

Ava Cascario, Sydney Holt, and Kacie Rohn

2024-12-03

```
# Load necessary packages ----
library(ggplot2)
library(dplyr)
library(kableExtra)
library(knitr)
library(tinytex)
# Load Data ----
shopping_trends_raw <- read.csv(
  file = "shopping_trends.csv",
  header = TRUE,
  sep = ",",
)
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"
  )

type_reviews <- shopping_trends_clean %>%
```

```

select(
  gender,
  review_rating
) %>%
group_by(
  gender,
  review_rating
)
ggplot(
  data = type_reviews,
  aes(
    x = gender,
    y = review_rating
  )
) +
  geom_boxplot()
shopping_summary <- lm(formula = purchase_amount_usd ~ category, data = shopping_trends_clean)
shopping_summary_model <- summary(shopping_summary)

shopping_summary_model$coefficients %>%
  kable() %>%
  kableExtra::kable_classic()
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),
    medianAbsoluteDeviation = mad(purchase_amount_usd),
    sampleArithmeticMean = mean(purchase_amount_usd),
    sampleArithmeticSD = sd(purchase_amount_usd)
  )

shopping_summary %>%
  kable() %>%
  kableExtra::kable_classic()
# 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"
)
# 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"
)
# New Column Added for Region
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")
# Visualization of Data by Region Summarized by 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(),
  panel.grid.minor = 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)
)
}

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)
```

## 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