**School of Information Studies**

**Syracuse University**

IST 687 Introduction to Data Science

Customer Shopping Preferences Analysis

Group A

Dennis Amirulla, Quintin Covington, Kéli Davis, Darious Dixon

19 Dec 2023

Table of Contents

[1. Introduction 3](#_Toc153883040)

[2. Results Summary 3](#_Toc153883041)

[2.1 Demographic Findings 3](#_Toc153883042)

[2.2 Top and Bottom Performing Categories 3](#_Toc153883043)

[2.3 Revenue Insights by State 3](#_Toc153883044)

[2.4 Preferred Payment Method 4](#_Toc153883045)

[3. Business Objective 4](#_Toc153883046)

[4. Dataset Attributes 4](#_Toc153883047)

[5. Analytical Methodology 5](#_Toc153883048)

[6. Data Munging and Cleaning 5](#_Toc153883049)

[7. Data Analysis 5](#_Toc153883050)

[7.1 Customer Demographic Analysis 6](#_Toc153883051)

[7.2 Overall Customer Behavior Analysis 8](#_Toc153883052)

[7.3 Focused Customer Behavior Analysis 11](#_Toc153883053)

[8. Recommendations & Conclusion 12](#_Toc153883054)

[9. Additional Analysis and Visualizations 13](#_Toc153883055)

[10. References 15](#_Toc153883056)

[Appendix A - R Code 16](#_Toc153883057)

# Introduction

The purpose of this report is to present the findings of the Customer Shopping Preferences dataset analysis. Analysis of customer behavior and purchasing patterns can offer valuable insights for any business. It is important in a competitive environment that companies shift their focus to analytics to better understand their customer’s preferences. This will allow companies to adjust their strategies and adapt to the needs of their customer base.

The dataset provided to our team is a sample of customer transactions from different client store locations across the United States. The data consists of the transaction information from 3,900 customers. The dataset includes but is not limited to the following: customer age, gender, purchase amount, preferred payment methods, items purchased, shopping frequency, and preferred shopping seasons.

With this report, our team aims to provide information and recommendations that will allow the client to align their strategies with customer needs and preferences.

# Results Summary

# 2.1 Demographic Findings

Our team's analysis determined that middle-aged consumers (35-54 years old) represented the largest segment of customer transactions. We also found that 58% of customers are men. Our team recommends the client shift general marketing focus to the middle-aged demographic to increase future sales. Additionally, we also recommend the client develop marketing strategies and increase product offerings specifically to the male middle-aged consumer.

# 2.2 Top and Bottom Performing Categories

Our analysis shows clothing and accessories lines are the top-performing product categories. We recommend inventory management and marketing strategies be tailored to towards these top selling products.

Identifying areas of underperformance can create opportunities for targeted marketing campaigns, product improvements, or strategic adjustments to boost sales. Our analysis found that footwear and outerwear have much lower sales than clothing and accessories. More data and analysis may be required to identify the cause of the underperformance of these items.

# 2.3 Revenue Insights by State

California and Illinois demonstrate both the highest number of purchases and largest sales revenue. Understanding the drivers of success in these two states may allow the client to focus their resources to maintain and potentially expand market share.

Kansas ranks at the bottom in both the number of purchases and generated revenue. Collecting additional data may help identify the cause of the underperformance. Additional analysis may assist the company in developing strategies to address local market dynamics and help with brand penetration.

# 2.4 Preferred Payment Method

Another key finding is the customer base prefers using credit cards as their primary method of payment. This insight is critical for optimizing payment processing systems, ensuring seamless transactions, and exploring partnerships with credit card companies.

# Business Objective

The objective of this project is to provide an in-depth analysis of customer demographics and purchasing behaviors.

Our analysis aims to help the client improve business inventory management, marketing strategies, and sales growth by answering the following questions:

* + Who is our target demographic? By gender, age, and location.
  + Which products are generating the most sales revenue?
  + What states are generating the most sales by products and revenue?
  + Does the season of the year impact customer sales purchases?
  + Does method of payment or frequency of purchase show any usable trends?

# Dataset Attributes

The dataset contains 3,900 observations and 18 attributes for a total of 70,200 “values”. The dataset includes information such as customer age, gender, item purchased, purchase amount, purchase location by state, purchase item details (size, color), payment method, number of previous purchases, etc. Figure 1 summarizes the dataset structure. Table 1 provides a description of each attribute.

Figure 1. Data Structure

A screenshot of a computer code

Description automatically generated

Table 1. Data Summary

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Customer ID | unique identifier for each customer |
| Age | age of the customer |
| Gender | gender of the customer |
| Item Purchased | the item purchased by the customer |
| Category | category of the item purchased |
| Purchase Amount (USD) | the amount of the purchase in USD |
| Location | location where the purchase was made |
| Size | size of the purchased item |
| Color | color of the purchased item |
| Season | season during which the purchase was made |
| Review Rating | rating given by the customer for the purchased item |
| Subscription Status | indicates if the customer has a subscription |
| Shipping Type | type of shipping chosen by the customer |
| Discount Applied | indicates if a discount was applied to the purchase |
| Promo Code Used | indicates if a promo code was used for the purchase |
| Previous Purchases | the total count of transactions concluded by the customer at the store |

# Analytical Methodology

The main tool used to manipulate data, analyze data, and generate the visualizations contained in this report is R. The IDE used is RStudio. Please see Appendix A for the R Code associated with this analysis.

# Data Munging and Cleaning

Base R functions and Tidyverse are used to clean and convert the data into a more useable form. The original dataset is provided in CSV text file. This file is read into RStudio. The CSV file is relatively clean; however, several transformations are needed before analyzing the data. The following functions were performed to clean the data:

* Remove NAs:
* Remove special characters from column names:
* Remove spaces from column names
* Filtering data
* Sorting data

# Data Analysis

The dataset analysis is separated into two sections: Customer Demographic Analysis and Customer Behavior Analysis. The first section of the analysis provides basic customer demographic data such as age and gender. The second section discusses the shopping preferences of the customers.

# 7.1 Customer Demographic Analysis

Understanding customer base demographics is important as it drives actionable insight specifically designed for the target audience. This can create marketing strategies, improve inventory management, and even personalize the customer experience.

Figure 2 shows the gender distribution of the customers. The data indicates that 58.1% of customers are male and 41.9% are female.

Figure 2. Customer Gender Analysis

A green and yellow pie chart

Description automatically generated

The minimum customer age is 18 years old. The maximum customer age is 70 years old. Due to the size of the dataset, customer age has been grouped into six categories: 18-24 years old, 25-34, 35-44, 45-54, 55-64, and 65 years and older. Figure 3 shows the distribution of customers in each age group. The largest age group in the dataset is 45 to 54 at 30%. The age group 35 to 44 years old is the second largest at 27.6%. The two smallest age groups are younger than 25 and older than 65. The average age of a customer is 43 years old.

Figure 3. Customer Age

A pie chart with numbers and a percentage

Description automatically generated

The figure below (Figure 4) demonstrates gender differences across age groups. The distribution of male and female is similar across each age group. The largest age group of either gender ranges between 35 and 54 years old.

Figure 4. Customer Age by Gender

A graph of a graph

Description automatically generated with medium confidence

# 7.2 Overall Customer Behavior Analysis

Understanding customer purchasing behavior is vital as the client can better identify what items are selling best and what locations are most profitable. By understanding these trends, the client can focus their marketing efforts on areas of the country that are struggling or items that are not selling well. Additionally, the client can ensure customer demand is always met by increasing inventory of highest selling stores in each state.

Figure 5 illustrates the number of purchases made in each state. In the dataset, each transaction is linked to a unique customer, therefore, this map also shows the number of customers in each state. The states with the greatest number of customers are Montana, California, Idaho, Illinois, and Alabama. The states with the smallest number of customer transactions are Rhode Island, Kansas, Arizona, New Jersey, and Florida.

A map of the united states

Description automatically generatedFigure 5. Total Number of Purchases per State

Figure 6 provides the total dollars spent per State regardless of the number of customer transactions. The state that spent the most is California, followed by Illinois, New York, Maryland, and Pennsylvania. The states that spent the least are Wyoming, New Hampshire, Wisconsin, South Dakota, and Kansas.

Figure 6. US Customer Total Purchases

A map of the united states

Description automatically generated

Clothing items are the most frequent item bought by customers. This category includes items such as blouses, dresses, pants, hoodies, shirts, and shorts. Accessories are the second most frequently purchased item. This includes belts, jewelry, hats, gloves, backpacks, and purses. The item least purchased is outerwear which includes jackets and coats.

Figure 7. Product Category

A graph of a product category

Description automatically generated

Purchases are most frequently made during the summer season. However, as seen in Figure 8, the purchases made each season are mostly evenly distributed.

Figure 8. Seasonal Purchases

A graph showing the seasons of customer purchase

Description automatically generated

Figure 9 indicates the payment method most used. Credit cards are used most to purchase items. Payment methods such as Venmo and PayPal are also popular. Cash transactions are among the least common payment methods.

Figure 9. Payment Method Preference

A graph of a payment method

Description automatically generated

Figure 10. Frequency of Purchase

A graph of green rectangular bars

Description automatically generated

Figure 10 shows the frequency with which customers make purchases. The most common frequency is monthly and every 3 months.

# 7.3 Focused Customer Behavior Analysis

As seen in Figure 3, customers in the age range of 35-54 years old make up about 60% of total purchases. Our team recommends focusing on this demographic to increase sales. The following analysis is based on the age group 35-54 years old.

The graph (Figure 11) shows the most frequent item purchased by this group is pants. Sandals, jackets, scarves, and jewelry are also purchased more than other items. The graph also shows this group spends the most money on scarves, more than any other item. Other items which bring in the most money are pants, sandals, jackets, and blouses. The least frequently purchased items are jeans and gloves. As expected, the least amount of dollars is spent purchasing these two items.

A graph of blue dots

Description automatically generatedFigure 11. Frequency of Purchase $ Spent per Item

# Recommendations & Conclusion

This analysis provides an understanding of the client’s customer base, product performance, and sales revenue by state. Using these insights, our team has developed the recommendations below.

Inventory Focus

Top Performing Categories: Since both the clothing and accessories lines are leading sales, future strategies should be focused on promoting and expanding products in these categories. We suggest the client ensure there is always stock available for these items and develop targeted promotions to drive even more revenue.

Refresh Underperforming Categories: There are challenges currently with the footwear and outerwear categories. We recommend additional market research to explore the root cause of product underperformance. We also suggest ensuring there is no surplus of warehouse inventory as this may increase operating cost.

Localized Inventory Management: It is important that levels of inventory are managed at the local level to ensure that the company is optimizing overall inventory efficiency.

Marketing Focus

Targeted Marketing Strategies for Key Demographics: 60% of the total sales are generated by customers between the ages of 35-54. 58% of the total customer base are male. Our team suggest the clients’ future marketing efforts be tailored to these specific demographic segments. Targeted marketing campaigns can appeal to this demographic and result in new and repeat customers.

Regional Marking Strategies: Since the company has success in states in California and Illinois, they should consider implementing region-specific marketing strategies. This can include local promotions and partnerships with local celebrities or influencers. They can tailor their marketing messages to be unique based on the characteristics of each location. We suggest the client collect additional data to understand and address the challenges that other states, such as Kansas, may be encountering.

Credit Card Promotions: Since credit cards are the preferred method of payment by the customer base, we suggest marketing strategies include partnerships and promotions with credit card companies. Promotions could include exclusive discounts and loyalty programs. Partnerships with credit card providers can help enhance the overall customer experience.

Based on our recommendations, the company can capitalize on their strengths and improve their weaknesses. These inventory control and marketing strategies are designed to help the company remain competitive in a customer-focused market. By implementing these recommendations, the company will ensure that demand is always met while increasing their customer base and their sales revenue.

# Additional Analysis and Visualizations

Other analyses were also performed. Below are two linear regression models that were created. The first model (Figure A) is the purchase amount compared to the age of the customer. As you can see there are no clear correlations between the two variables as the data is spread evenly across. The second model (Figure B) illustrates the Purchase Amount versus the number of previous purchases made by the customer. Again, there was no correlation found between the two variables as the data was also spread evenly across.

Figure A. Linear Regression Model #1

A black and white text

Description automatically generated

Figure B. Linear Regression Model #2

A screen shot of a screen

Description automatically generated

# References

Saltz, J. S. (2022). *Data Science for Business with R.* Syracuse University.

# Appendix A - R Code

##################################################

#IST 687 Group A

#Final Project

#Customer Shopping Preferences

#####libraries needed#####

library(readxl)

library(tidyverse)

library(ggplot2)

library(RColorBrewer)

library(R.utils)

library(ggplot2)

library(ggmap)

library(jsonlite)

library(RCurl)

library(dplyr)

#####Importing Customer Preference Dataset excel file and put into dataframe named ST (Shopping Trends)

Final\_Project\_Data1 <- read\_excel("~/Final\_Project\_Data1.xlsx")

View(Final\_Project\_Data1)

#Create dataframe ST

ST <- data\_frame(Final\_Project\_Data1)

#####Exploring the data at high level#####

str(ST)

summary(ST)

View(ST)

###################

######Data Cleaning#####

#Check dataframe for NAs

sum(is.na.data.frame(ST))

#No NAs were found

#Check for spaces

colnames(ST) #view column names

#replace space with underscore

colnames(ST) <- gsub(" ","\_",colnames(ST))

head(ST,1) #confirm column names updated

#Column Rename to get rid of characters

ST <- ST %>% rename("Purchase\_Amount" = "Purchase\_Amount\_(USD)")

###################

######Customer Demographics Analysis#####

Age <- ST$Age

Gender <- ST$Gender

Age\_Gender <-data.frame(Age, Gender)

range(Age\_Gender$Age)

mean(Age\_Gender$Age)

#Customer Gender Analysis

Female <- sum(Age\_Gender$Gender == "Female")

Male <- sum(Age\_Gender$Gender == "Male")

Female

Male

nrow(Age\_Gender)

Gender <- c("Female","Male")

Totals <- c(41.9, 58.1)

Genderdf <- data\_frame(Gender,Totals)

GenderAnalysis <- ggplot(Genderdf, aes(x="",y=Gender, fill=Gender)) + geom\_bar(width=1,stat="identity") + coord\_polar("y", start=0) + theme\_void() +

theme(plot.title=element\_text(size=18),legend.text=element\_text(size=12),legend.title=element\_text(size=12)) +

scale\_fill\_manual(values = c("darkgoldenrod2","seagreen")) + ggtitle("Customer by Gender (%)") +

geom\_text(aes(label = Totals), size= 5, position = position\_stack(vjust = 0.5))

GenderAnalysis

#Customer Age Analysis

#Customer Age by Gender Histogram

HistAG <- ggplot(Age\_Gender,aes(x=Age, fill=Gender)) + geom\_histogram(position="stack", binwidth = 3, color="black", alpha=.9) +

labs(title = "Customer Age by Gender Histogram", x = "Age", y = "Frequency") + scale\_fill\_manual(values = c("darkgoldenrod2","seagreen")) +

theme(axis.text=element\_text(size=18),axis.title=element\_text(size=20),plot.title=element\_text(size=18),legend.text=element\_text(size=14),legend.title=element\_text(size=14))

HistAG

#Sampling Distribution for Customer Age and Descriptive Statistics

Age\_Distribution <- replicate(1000,mean(sample(Age\_Gender$Age,size=5,replace=TRUE)))

mean(Age\_Distribution)

summary(Age\_Distribution)

quantile(Age\_Distribution, probs =c(0.25, 0.50, 0.75))

sd(Age\_Distribution)

#Sampling Distribution for Customer Age

Age\_Distribution\_Hist <- hist(replicate(1000,mean(sample(Age\_Gender$Age,size=5,replace=TRUE))), main="Customer Age Normal Distribution", xlab="Average Age")

Age\_Distribution\_Hist

ST18\_24 <- ST %>% filter(between (Age, 18, 24)) %>% arrange(Age, Gender, Category, Item\_Purchased)

ST25\_34 <- ST %>% filter(between (Age, 25, 34)) %>% arrange(Age, Gender, Category, Item\_Purchased)

ST35\_44 <- ST %>% filter(between (Age, 35, 44)) %>% arrange(Age, Gender, Category, Item\_Purchased)

ST45\_54 <- ST %>% filter(between (Age, 45, 54)) %>% arrange(Age, Gender, Category, Item\_Purchased)

ST55\_64 <- ST %>% filter(between (Age, 55, 64)) %>% arrange(Age, Gender, Category, Item\_Purchased)

ST65\_older <- ST %>% filter(between (Age, 65, 100)) %>% arrange(Age, Gender, Category, Item\_Purchased)

#Age Vector Math for report Writeup

Age1 <- (nrow(ST18\_24)/nrow(ST))\*100

Age2 <- (nrow(ST25\_34)/nrow(ST))\*100

Age3 <- (nrow(ST35\_44)/nrow(ST))\*100

Age4 <- (nrow(ST45\_54)/nrow(ST))\*100

Age5 <- (nrow(ST55\_64)/nrow(ST))\*100

Age6 <- (nrow(ST65\_older)/nrow(ST))\*100

Agepercent <- c(Age1, Age2, Age3, Age4, Age5, Age6)

Agepercent <- format(Agepercent,digits = 2, format="f")

Agegroup <- c(1824, 2534, 3544, 4554,5564,65)

Agepie <- data\_frame(Agegroup, Agepercent)

AgeAnalysis <- ggplot(Agepie, aes(x="",y=Agepercent, fill=Agepercent)) + geom\_bar(width=1,stat="identity") + coord\_polar("y", start=0) + theme\_void() +

theme(plot.title=element\_text(size=18),legend.text=element\_text(size=12),legend.title=element\_text(size=12)) + ggtitle("Customer by Age (%)") +

geom\_text(aes(label = Agepercent), size= 6, position = position\_stack(vjust = 0.5)) +

scale\_fill\_manual(values=c("darkgoldenrod2","seagreen","aquamarine3","cadetblue3", "steelblue","royalblue4"), name = "Age Group", labels = c("65+","18-24","55-64","25-34","35-44","45-54"))

AgeAnalysis

#Map of Total Number of Purchases by State

dfPurchaseLocation <- ST %>% group\_by(Location) %>% count(Location)

view(dfPurchaseLocation)

dfPurchaseLocation[which.min(dfPurchaseLocation$n),]

dfPurchaseLocation[which.max(dfPurchaseLocation$n),]

USA <- data.frame(map\_data("state"))

view(USA)

dfPurchaseLocation$Location <- tolower(dfPurchaseLocation$Location)

view(dfPurchaseLocation)

merged\_df2 <- merge (USA, dfPurchaseLocation, all.x = TRUE, by.x = 'region', by.y = 'Location')

view(merged\_df2)

merged\_df2 <- merged\_df2[order( merged\_df2[,4], merged\_df2[,5] ),]

colnames(merged\_df2)[7] = "Location\_Count"

ggplot() + geom\_polygon(data=merged\_df2,aes(x=long, y=lat, group = group, fill=Location\_Count)) +

ggtitle("US Total Purchases per State") + coord\_map() +

theme(axis.text.x=element\_text(size=18),axis.text.y=element\_text(size=18),axis.title=element\_text(size=20),plot.title=element\_text(size=18)) +

labs(fill = "Total Purchases")

###################

######Customer Behavior Analysis#####

#Season of Purchase Breakdown

ST\_Sea <- ggplot(ST) + aes(x=Season) + geom\_bar(color="black", fill = "seagreen") +

ggtitle("Season of Customer Purchase") + xlab("Season") + ylab("Frequency") +

theme(axis.text.x=element\_text(angle=0, vjust=0.6, size=18),axis.text.y=element\_text(size=18),axis.title=element\_text(size=20),plot.title=element\_text(size=18))

ST\_Sea

#Item Category Breakdown

ST\_cat <- ggplot(ST) + aes(x=Category) + geom\_bar(color="black", fill = "seagreen") +

ggtitle("Product Category") + xlab("Product Category") + ylab("Frequency") +

theme(axis.text.x = element\_text(size=18, angle=0, vjust=0.6),axis.text.y=element\_text(size=18),axis.title=element\_text(size=18),plot.title=element\_text(size=18))

ST\_cat

#Payment Method Breakdown

ST\_pay <- ggplot(ST) + aes(x=Payment\_Method) + geom\_bar(color="black", fill = "seagreen") +

ggtitle("Most Common Payment Method") + xlab("Payment Method") + ylab("Frequency") +

theme(axis.text.x=element\_text(angle=0, vjust=0.5, size=18),axis.text.y=element\_text(size=18),axis.title=element\_text(size=20),plot.title=element\_text(size=18))

ST\_pay

#Frequency of Purchase breakdown

ST\_FP <- ggplot(ST) + aes(x=Frequency\_of\_Purchases) + geom\_bar(color="black", fill = "seagreen") +

ggtitle("Purchase Frequency") + xlab("Purchase Frequency") + ylab("Frequency") +

theme(axis.text.x = element\_text(angle=65, vjust=0.6, size=18),axis.text.y=element\_text(size=18),axis.title=element\_text(size=20),plot.title=element\_text(size=18))

ST\_FP

#Map of $ Spent by State

dfPurchase <- ST %>% group\_by(Location) %>% summarise(Purchase\_Total = sum(Purchase\_Amount))

view(dfPurchase)

dfPurchase[which.min(dfPurchase$Purchase\_Total),]

dfPurchase[which.max(dfPurchase$Purchase\_Total),]

USA <- data.frame(map\_data("state"))

view(USA)

dfPurchase$Location <- tolower(dfPurchase$Location)

view(dfPurchase)

merged\_df <- merge (USA, dfPurchase, all.x = TRUE, by.x = 'region', by.y = 'Location')

view(merged\_df)

merged\_df <- merged\_df[order( merged\_df[,4], merged\_df[,5] ),]

ggplot() + geom\_polygon(data=merged\_df,aes(x=long, y=lat, group= group, fill=Purchase\_Total)) +

ggtitle("US Total $ Spent per State") + coord\_map() +

theme(axis.text.x=element\_text(size=18),axis.text.y=element\_text(size=18),axis.title=element\_text(size=20),plot.title=element\_text(size=18)) +

labs(fill = "Total $ Spent")

###################

#####Focused Customer Behavior Analysis######

#Focus analysis on Age group 35-55

#Item Purchase Frequency and $ Spent per Item

ST35\_54 <- ST %>% filter(between (Age, 35, 54))

ST35\_541 <- ST35\_54 %>% group\_by(Item\_Purchased) %>% summarise(sum=sum(Purchase\_Amount))

ST35\_542 <- ST35\_54 %>% group\_by(Item\_Purchased) %>% count(Item\_Purchased)

ST35\_543 <- merge(ST35\_541, ST35\_542, by = "Item\_Purchased")

df\_IP <- ggplot(ST35\_543) + aes(x=Item\_Purchased, y=n, color=sum, size=sum) + geom\_point() + theme(axis.text.x = element\_text(angle=65, vjust=0.6, size=18), axis.text.y=element\_text(size=18),axis.title=element\_text(size=18), plot.title=element\_text(size=20), legend.text=element\_text(size=14), legend.title=element\_text(size=14)) +

ggtitle("Frequency of Purchase and $ Spent per Item") + xlab("Item Purchased") + ylab("Frequency of Item Purchase") +

scale\_color\_continuous(name = "$ Spent") + scale\_size(name = "$ Spent", range = c(3 , 23))

df\_IP

#Linear Regression Models Not Used

ggplot(data = ST) + aes(x = Age, y = Purchase\_Amount) + geom\_point() + geom\_smooth(method = "lm", se = FALSE)

ggplot(data = ST) + aes(x = Previous\_Purchases, y = Purchase\_Amount) + geom\_point() + geom\_smooth(method = "lm", se = FALSE)