Customer Shopping Data Analysis

Jai Patel

2024-10-20

Abstract

This report analyzes customer shopping behavior using a dataset of transaction records from various shopping malls. The purpose of this study is to investigate how money, purchase categories, and consumer demographics relate to one another. According to the findings, older consumers are more likely to purchase shoes, while younger consumers are more likely to favor electronics.

Introduction

Retailers must understand customer buying habits in order to customize their marketing campaigns and raise client satisfaction levels. This study looks into the demographics and buying habits of Turkish shoppers in malls. The objectives include identifying the main factors that influence purchasing decisions and assessing how demographic characteristics affect product categories.

Data

The dataset used for this analysis has 66936 records, containing many variables. The variables are `invoice_no`, `customer_id`, `gender`, `age`, `category`, `quantity`, `price`, `payment_method`, `invoice_date`, and `shopping_mall`(see Aslan [1]). The information in this dataset contains shopping information from 10 different shopping malls between 2021 and 2023 based in Istanbul. The dataset is representative of shopping behaviors of many malls and includes limitations such as potential biases in seasonal shopping trends and customer reporting.

```
shopping_data_Istan <- read.csv("../customer_shopping_data.csv")

summary(shopping_data_Istan)
```

```
##   invoice_no        customer_id           gender               age        
##  Length:66936       Length:66936       Length:66936       Min.   :18.00  
##  Class :character   Class :character   Class :character   1st Qu.:30.00  
##  Mode  :character   Mode  :character   Mode  :character   Median :43.00  
##                                                           Mean   :43.42  
##                                                           3rd Qu.:56.00  
##                                                           Max.   :69.00  
##    category            quantity         price         payment_method    
##  Length:66936       Min.   :1.000   Min.   :   5.23   Length:66936      
##  Class :character   1st Qu.:2.000   1st Qu.:  45.45   Class :character
```

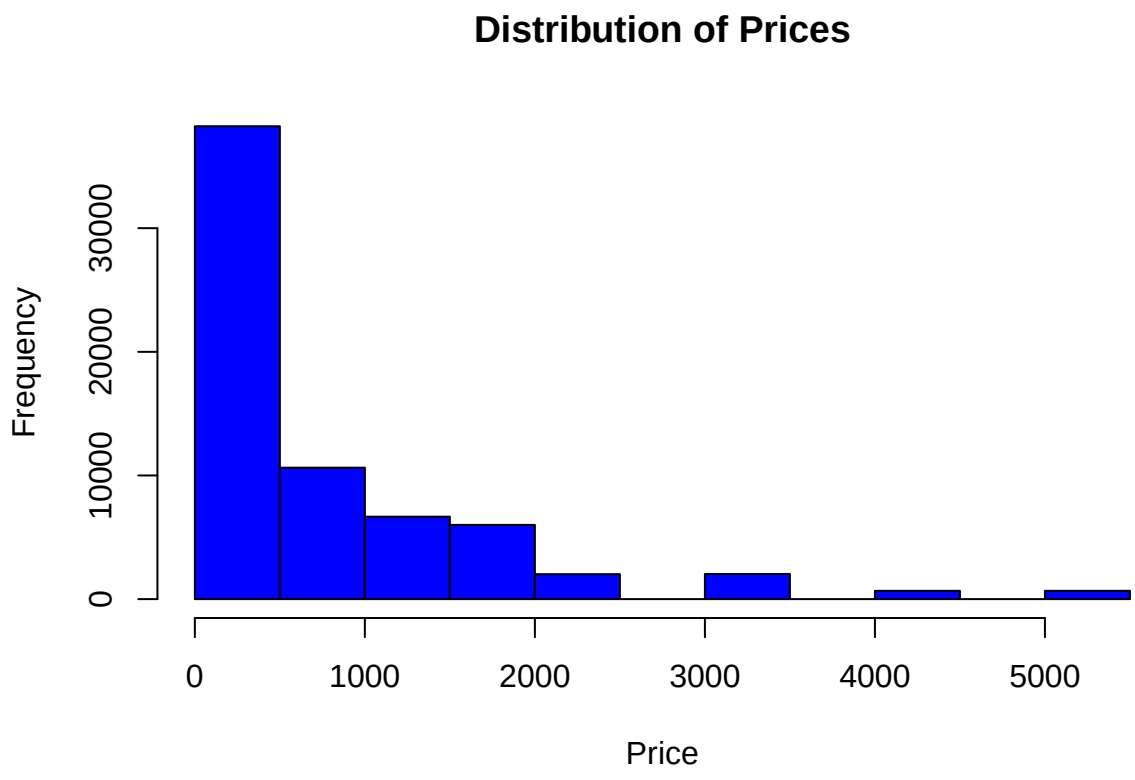
```
##  Mode  :character   Median :3.000   Median : 203.30   Mode  :character
##                     Mean   :3.005   Mean   : 688.10
##                     3rd Qu.:4.000   3rd Qu.:1200.32
##                     Max.   :5.000   Max.   :5250.00
##  invoice_date       shopping_mall
##  Length:66936       Length:66936
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

The above data is showing basic information of the categories for which we can potentially see a relationship between.

Analysis

Distribution of Prices

```
hist(shopping_data_Istan$price, breaks = seq(0, 5500, by = 500),
    main = "Distribution of Prices", xlab = "Price", ylab = "Frequency",
    col = "blue", )
```



```
Standard_d = sd(shopping_data_Istan$price, na.rm = TRUE)
```

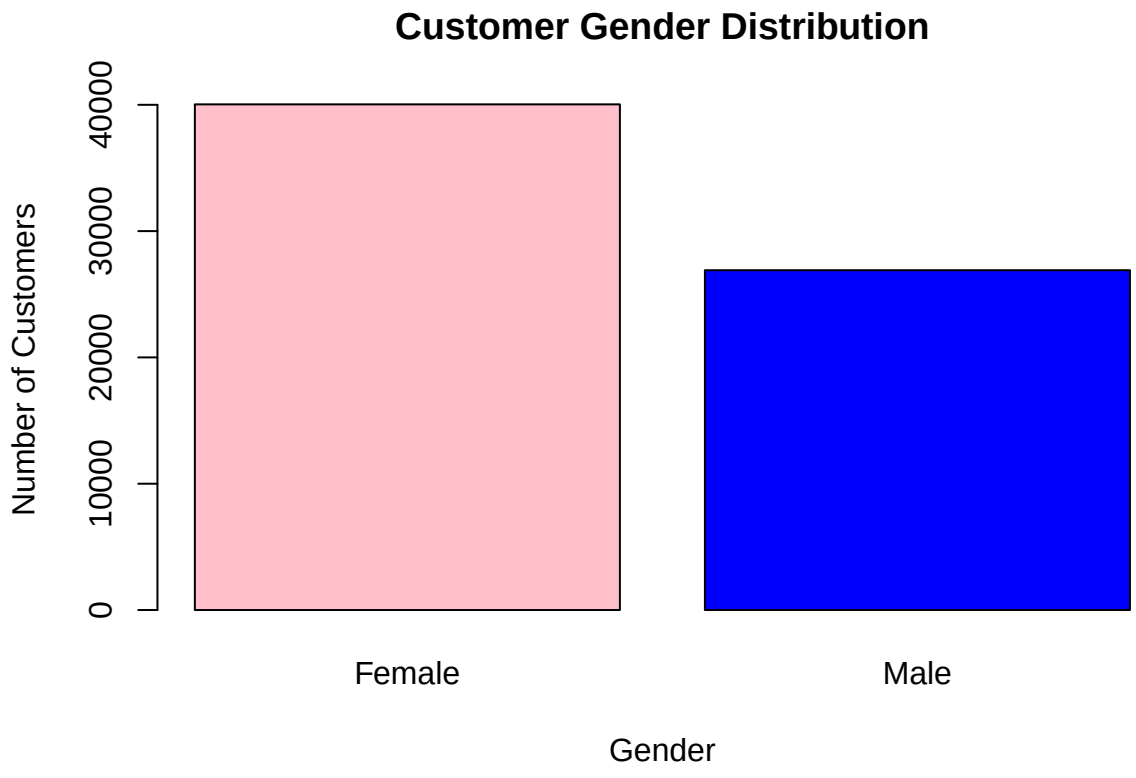
Result

This Histogram (see R Core Team [4]) is showing us that the highest price is 5250 and the minimum is 5.23. The highest frequency in prices in between 0 - 500 Turkish Liras. Also, the standard deviation of the price is 939.9185568(see Posit team [3]). This is around 250 Turkish Liras more than the mean which is showing that the value of prices are spread out over a wider range which is what the histogram is showing.

Gender Distribution

```
gender_dist = table(shopping_data_Istan$gender)

barplot(gender_dist, main = "Customer Gender Distribution", xlab = "Gender",
    ylab = "Number of Customers", col = c("pink", "blue"), )
```




```
femalesTotal <- sum(shopping_data_Istan$gender == "Female")
malesTotal <- sum(shopping_data_Istan$gender == "Male")
```

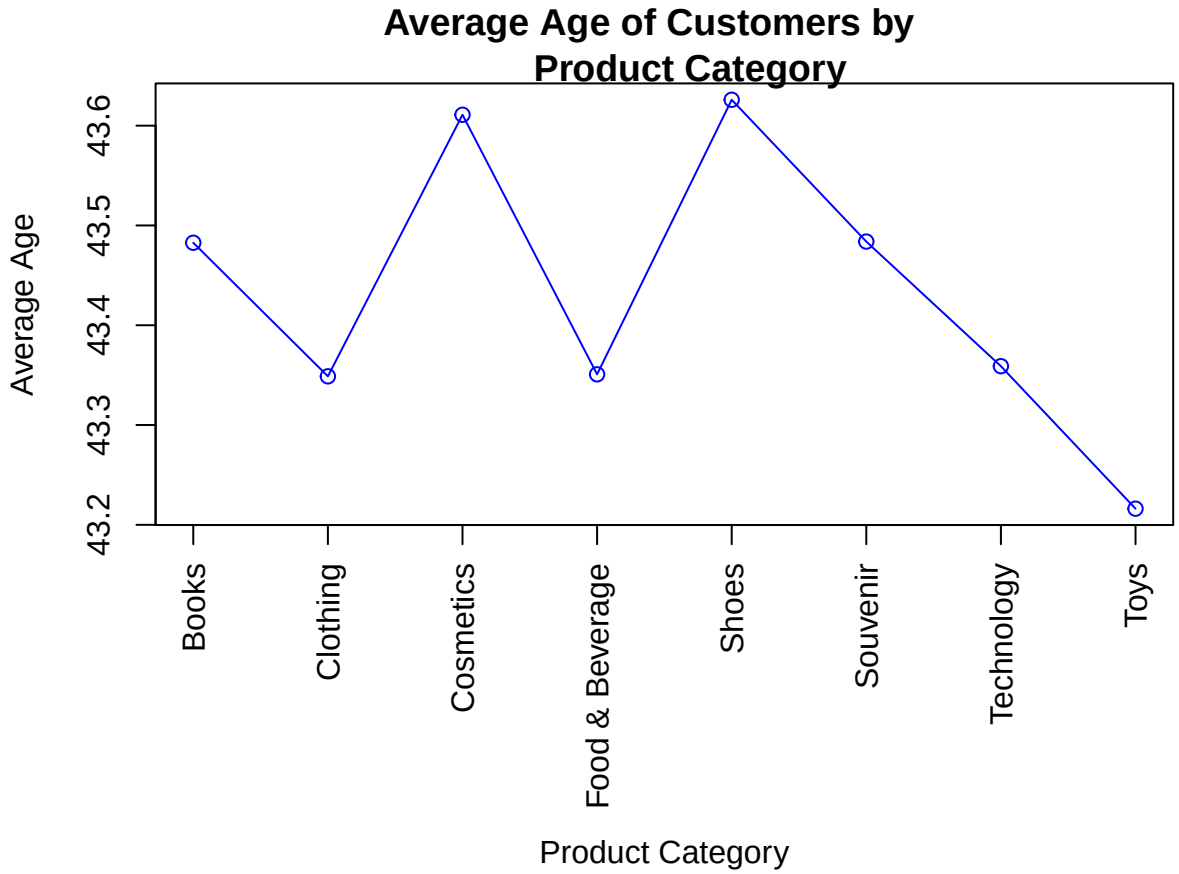
Result

The barplot above is showing that there are 40035 females total and 26901 males total which means that in Istanbul, there are more females than males shopping retail. This data is showing that that means when going out for retail shopping, females are more likely to go out than men.

Average Age by Product Distribution

```
ages = aggregate(age ~ category, data = shopping_data_Istan,
    FUN = mean, drop = FALSE)
par(mar = c(9, 4, 2, 2))
plot(ages$age, type = "o", xaxt = "n", main = "Average Age of Customers by
     Product Category",
    xlab = " ", ylab = "Average Age", col = "blue")

axis(1, at = 1:length(ages$category), labels = ages$category,
    las = 2)
mtext("Product Category", side = 1, line = 8)
```



Result

The plot above is showing the relationship between the average age of customers and the products they buy. According to this, if you are a older consumer, then you are most likely to buy cosmetics or shoes. However, if you are a younger consumer, then you are most likely to buy toys or technology.

Research Table

This table (see Le Tan et al. [2]) shows information of the demographics of a surveyed population, showing factors such as gender, age, occupation and income. In the frequency of the gender in this table, you can see that the females have a higher frequency than the males just like earlier in our analysis. To visit the table you can click here.

Conclusion

This analysis reveals important new information about Istanbul consumers' purchasing habits. The results show that most shoppers are women and that older consumers tend to buy different categories than younger consumers. Furthermore, the connection between age and different spending categories shows the importance of focused marketing campaigns. One future reference could be how shoppers change during seasonal times and this could help store owners prepare for those times. Overall, a more thorough look of the relationships among gender, age, and spending categories can improve stores' ability to maximize their choices and effectively interact with consumers.

References

[1] Mehmet Tahir Aslan. *Customer Shopping Dataset - Retail Sales Data*. https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset/data. 2022.

[2] Trinh Le Tan et al. *Research on factors affecting customers' shopping behavior on e-commerce exchanges during the covid-19 pandemic*. Oct. 2021. DOI: 10.47747/ijbme.v2i4.440.

[3] Posit team. *RStudio: Integrated Development Environment for R*. Posit Software, PBC. Boston, MA, 2024. URL: http://www.posit.co/.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.