Institut Mines-Télécom Business School

MSc in Management of Innovation in the Digital Economy

Online Sales Data

Name: Kayinat Naveed

Professor: Anahid BAUER

Course: Application of Statistics

23 February 2024
Évry, France

Table of Contents

Introduction

Online shopping has become increasingly popular in recent years, offering convenience and accessibility to consumers worldwide. Analyzing online sales data is crucial for businesses in the digital age as it provides valuable insights into consumer behavior, market trends, and competitive dynamics. Online sales data offers a comprehensive view of customer preferences, buying patterns, and purchasing habits. By analyzing this data, businesses can gain insights into what products customers are interested in, how they shop, and what influences their purchasing decisions. This understanding is essential for tailoring marketing strategies, product offerings, and customer experiences to meet the needs and preferences of target audiences. (Source: Li, X., & Wu, L. (2019). "Consumer Behavior Analysis in Online Shopping Environments: A Review.")

By analyzing which gender shops more and spends more on online shopping helps businesses optimize their product assortments to cater to the preferences and needs of different gender segments. By identifying top-selling products and categories among male and female shoppers, businesses can adjust their inventory levels, introduce gender-specific product lines, and capitalize on market opportunities. This data-driven approach to product assortment optimization improves customer satisfaction and drives sales growth. (Source: Sundar, B., et al. (2020). "Analyzing Online Shopping Behavior to Improve Product Assortment Strategies.")

Research Questions

In this analysis, we explore:

- A dataset of online sales data to understand the spending patterns.
- Preferences of customers, with a focus on gender differences.

Dataset Overview

Our dataset comprises 1116 entries with various attributes related to online sales transactions, including Class Order ID, Customer Name, Order Date, State, City, Gender, Amount, Profit, Quantity, Category, Subcategory, Payment Mode, and ID. We've cleaned the data by removing NA values and unnecessary columns to focus on relevant variables for our analysis.

Gender Spending Patterns

In our analysis, we initially assumed that females spend more money and order more products from online stores. However, our analysis revealed surprising results: males were found to spend more money and order more products compared to females.

Descriptive statistics revealed that the mean amount spent by males was $292.59, while for females it was $304.09.

Gender Order Frequency

After our initial analysis we found that males spend more money and order more products as compared to females. Then we create a frequency table which shows that males exhibit a higher frequency of online shopping orders. We found that out of the 1,116 entries, 58.4% were male (level 0) and 41.6% were female (level 1). This distribution indicates that there were more male customers in our dataset compared to females.

level	freq	perc	cumfreq	cumperc
0	652	58.4%	652	58.4%
1	464	41.6%	1'116	100.0%

Product Category Preferences

When examining product category preferences by gender, we observe interesting insights:

- **Clothing:** Males and females both show a preference for clothing, with 419 orders placed by males and 289 orders placed by females.
- **Electronics:** Although males and females both purchase electronics, the frequency is slightly higher among males (130 orders) compared to females (103 orders).
- **Furniture:** Both genders also purchase furniture items, with 103 orders placed by males and 72 orders placed by females.

	Online_Sales$Gender		
Online_Sales$Category	0	1	Row Total
Clothing	419	289	708
Electronics	130	103	233
Furniture	103	72	175
Column Total	652	464	1116

Descriptive Statistics

To see the summary of our dataset we perform descriptive statistics. Which provides a summary of the sales amount variable (denoted as "Amount") across the entire dataset and broken down by gender groups.

Across the entire dataset (all genders combined), the mean sales amount is $297.37, with a standard deviation of $464.19, indicating a wide variation in sales amounts.

```
> describe(Online_Sales$Amount)
   vars    n   mean     sd median trimmed    mad min  max range skew kurtosis   se
X1    1 1116 297.37 464.19  125.5  197.29 135.66   4 5729  5725 4.24    29.54 13.9
```

When examining sales amounts by gender groups:

Group 0 (males): The mean sales amount is $292.59, with a standard deviation of $487.93.

Group 1 (females): The mean sales amount is $304.09, with a standard deviation of $429.04.

```
 Descriptive statistics by group
group: 0
   vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 652 292.59 487.93  125.5  188.04 131.95   4 5729  5725  4.9    37.08 19.11
------------------------------------------------------------------------------
group: 1
   vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 464 304.09 429.04  125.5  210.44 138.62   6 3151  3145 2.79    10.31 19.92
```

T-TEST

We perform t-tests to statistically compare the spending and ordering behavior between males and females in your online sales data and determine if there was a significant difference between the two groups. And the results show that:

The p-value for the t-test comparing the amount spent by males and females is 0.6834. Since this p-value is greater than the significance level (typically 0.05), we fail to reject the null hypothesis. Therefore, there is no significant difference in the amount spent between males and females.

```
        Two Sample t-test

data:  amount_male and amount_female
t = -0.4079, df = 1114, p-value = 0.6834
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -66.84238  43.83371
sample estimates:
mean of x mean of y
 292.5905  304.0948

> confidence_interval <- T_Test_spending$conf.int
> confidence_interval
[1] -66.84238  43.83371
attr(,"conf.level")
[1] 0.95
```

The p-value for the t-test comparing the quantity ordered by males and females is 0.8912. Again, since this p-value is greater than the significance level, we fail to reject the null hypothesis. Therefore, there is no significant difference in the quantity ordered between males and females.

```
        Two Sample t-test

data:  order_male and order_female
t = -0.13678, df = 1114, p-value = 0.8912
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2803849  0.2438395
sample estimates:
mean of x mean of y
 3.779141  3.797414

> confidence_interval <- T_Test_Ordering$conf.int
> confidence_interval
[1] -0.2803849  0.2438395
attr(,"conf.level")
[1] 0.95
> |
```

Ordinary Least Squares (OLS)

Additionally, we conduct two separate Ordinary Least Squares (OLS) regression models. This model is conducted to analyze the relationship between gender and two key variables: amount spent (Amount) and quantity ordered (Quantity) in an online sales dataset.

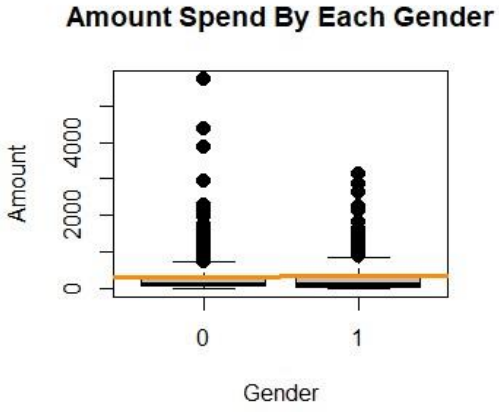
The first OLS regression model analyzed the relationship between gender and the amount spent by customers. The model yielded an R-squared value of 0.0001493, indicating that gender explains only a very small proportion of the variance in the amount spent. The coefficient for the Gender variable was not statistically significant (t = 0.408, p = 0.683), suggesting that there is no significant difference in the amount spent between males and females.

```
Call:
lm(formula = Amount ~ Gender, data = Online_Sales)

Residuals:
   Min     1Q Median     3Q    Max
-298.1 -247.6 -170.6   41.5 5436.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   292.59      18.19  16.089   <2e-16 ***
Gender1        11.50      28.20   0.408    0.683
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 464.4 on 1114 degrees of freedom
Multiple R-squared:  0.0001493, Adjusted R-squared:  -0.00074
F-statistic: 0.1664 on 1 and 1114 DF,  p-value: 0.6834
```



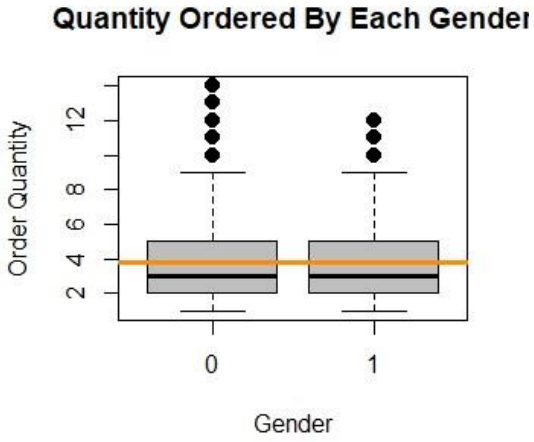
The second OLS regression model examined the relationship between gender and the quantity of products ordered. Similar to the amount spent model, the R-squared value for this model was very low (1.679e-05), indicating that gender explains almost no variance in the quantity ordered. The coefficient for the Gender variable was also not statistically significant (t = 0.137, p = 0.891), suggesting that there is no significant difference in the quantity of products ordered between males and females.

```
Call:
lm(formula = Quantity ~ Gender, data = Online_Sales)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7974 -1.7791 -0.7791  1.2209 10.2209

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.77914    0.08614  43.873   <2e-16 ***
Gender1      0.01827    0.13359   0.137    0.891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.199 on 1114 degrees of freedom
Multiple R-squared:  1.679e-05, Adjusted R-squared:  -0.0008809
F-statistic: 0.01871 on 1 and 1114 DF,  p-value: 0.8912
```



Conclusion

Based on the analysis, there is no significant difference between genders in terms of spending and ordering behavior in online shopping. Contrary to the initial assumption, males spend more money and order more products compared to females. This finding may have implications for marketing strategies and targeted advertising in online retail. Further research may be warranted to understand the underlying factors influencing purchasing decisions among different genders.