The Effects of Family Clothing Store Sales on Women’s Apparel Sales: A Statistical Investigation

2023-10-28

# Intro

This analysis explores the relationship between family and women’s clothing sales. The study aims to determine whether family clothing sales can serve as a predictor of women’s clothing sales.

Alternative Hypothsis: Family clothing sales will predict womens clothing sales. Null Hypothesis: There will be no association between family and womens clothing sales.

### Libraries

These are the libraries I used for the project.

library(ggplot2)  
library(readxl)  
library(knitr)  
library(ggplot2)  
library(ggrepel)

### Where I Got the Data From

In the link below you can see where I got the data from.

Original <- read\_excel("~/Documents/Data/Retail\_Salse/mrtssales92-present.xlsx")  
  
## https://www.census.gov/retail/sales.html

# Cleaning the data

In this block of code I am doing data selction, column labeling, and tabylar presentation. In the first like fo code, I am narrowing down the dataset by selecting specific rows (from 6-110) and columns (from 2 to 10) to create a new dataset called Modify. This helps to focus on the relevant subset of data for further analysis.

In the second, line of code, I am assigning meaningful column names to the modify dataset. The names represent different months in the year 2023, making the data more interpretable.

In the final line of code, I’m further refining the data set by selecting only the rows that are needed (rows 37 and 38), which could represent specific categories or information of interest.

Finally, the kable function is used to present the final dataset in a formatted tabular view, making it easier to read and analyze the selected data.

Modify<-Original[6:110,2:10]  
  
colnames(Modify)<-c("Category" ,"Jan\_2023","Feb\_2023","March\_2023","April\_2023","May\_2023", "June\_2023","July\_2023", "Aug\_2023")  
  
Final<-Modify[37:38,]

kable(Final)

| Category | Jan\_2023 | Feb\_2023 | March\_2023 | April\_2023 | May\_2023 | June\_2023 | July\_2023 | Aug\_2023 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Women’s clothing stores | 2514 | 2735 | 3408 | 3349 | 3663 | 3138 | 3243 | 3251 |
| Family clothing stores | 7786 | 7999 | 10226 | 10291 | 10733 | 10434 | 10992 | 11552 |

In the step below, I am transposing the data using the t() function to reoganize it. Transposing the data changes the orientation of the dataset, making it suitable for analysis by switching rows to columns and columns to rows. This transformation allows for a more structured and convient format for further data analysis

Final1<-t(Final)

Due to the transformation in the step above, the column names needed to be adjusted to reflect the new arrangement in the code below. Specifically, I renamed the columns as ‘WomenClothingStores’ and ‘FamilyClothingStores’ to better describe the data. Next, I further refined the dataset by selecting rows 2 to 9, capturing the relevant information. Finally, I converted the modified data into a data frame, making it ready for analysis and visualization.

colnames(Final1)<-c("WomenClothingStores","FamilyClothingStores")  
Final1<-Final1[2:9,]  
Final1<<- as.data.frame(Final1)

In this step, I converted the values of ‘WomenClothingStores’ and ‘FamilyClothingStores’ to numeric data types. This transformation is necessary to prepare the data for parametric analysis, enabling the use of statistical tests and techniques that require numeric inputs.

Final1$WomenClothingStores<-as.numeric(Final1$WomenClothingStores)  
Final1$FamilyClothingStores<-as.numeric(Final1$FamilyClothingStores)

The table below illustrates the outcomes of the data wrangling process, showing the transformed and refined data ready for analysis.

kable(Final1)

|  | WomenClothingStores | FamilyClothingStores |
| --- | --- | --- |
| Jan\_2023 | 2514 | 7786 |
| Feb\_2023 | 2735 | 7999 |
| March\_2023 | 3408 | 10226 |
| April\_2023 | 3349 | 10291 |
| May\_2023 | 3663 | 10733 |
| June\_2023 | 3138 | 10434 |
| July\_2023 | 3243 | 10992 |
| Aug\_2023 | 3251 | 11552 |

# Analysis

### Correlation

The correlation analysis indicates a highly significant and strong association between ‘WomenClothingStores’ and ‘FamilyClothingStores,’ with a correlation coefficient of approximately 0.835. This suggests a positive relationship between the two variables, indicating that changes in women’s clothing store sales are closely related to changes in family clothing store sales.

cor(Final1$WomenClothingStores, Final1$FamilyClothingStores)

## [1] 0.8351892

## Correlation chart https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html

### Linear Regression

In this model, I am predicting women’s clothing store sales (WomenClothingStores) based on family clothing store sales (FamilyClothingStores). The regression model assess how effectively variations in family clothing store sales can explain or predict changes in women’s clothing store sales.

RetailModelFamilyPredicitngWomen <- lm(Final1$WomenClothingStores ~ Final1$FamilyClothingStores)

# Linear Model

For my model the F-statistic is 13.84, and the associated p-value is 0.009854, which is less than the significance level of 0.05, indicating that the model is statistically significant. This result suggests that the model is statistically significant, indicating a meaningful relationship between family clothing store sales and women’s clothing store sales. However, we would need to check if the regression model meets assumptions and the data fits the model properly.

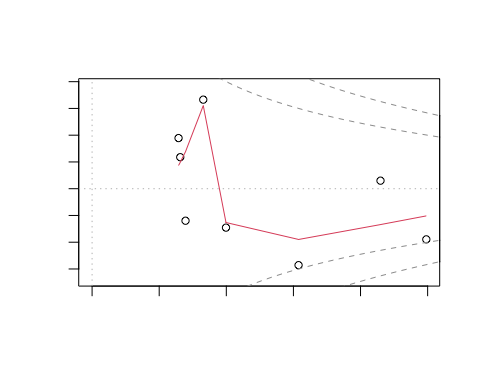
summary(RetailModelFamilyPredicitngWomen )

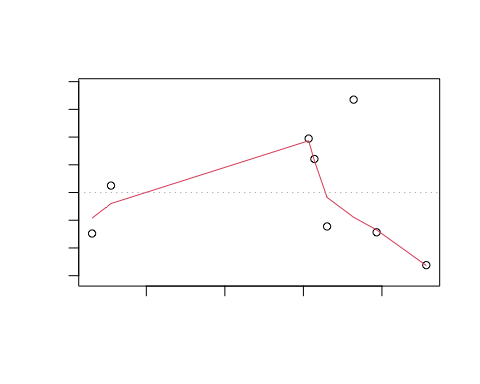
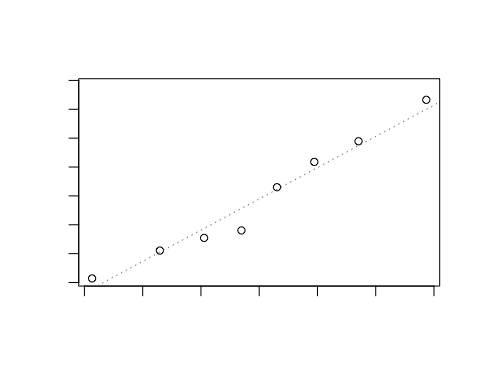
##   
## Call:  
## lm(formula = Final1$WomenClothingStores ~ Final1$FamilyClothingStores)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -262.04 -144.56 -48.67 139.39 335.07   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 902.06539 612.67557 1.472 0.19135   
## Final1$FamilyClothingStores 0.22602 0.06076 3.720 0.00985 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 220.4 on 6 degrees of freedom  
## Multiple R-squared: 0.6975, Adjusted R-squared: 0.6471   
## F-statistic: 13.84 on 1 and 6 DF, p-value: 0.009854

# Spearman

Based on the plots, we can see that they violete the assumptions. For example, in the first plot there is curvature associated with the residual or error terms. The second plot the data the data looks fairly normal, with some divergence. However, the third and the fourth plot there is curvature which indicates non-linearity. Therefore, the data does not fit a linear regresion model and a different tool would be needed to better fit the model. Thus, the analysis of the spearman correlation is used.

Plot<-plot(RetailModelFamilyPredicitngWomen)

A graph with a line drawn on it

Description automatically generated

The Spearman’s rank correlation coefficient (rho) is a measure of the strength and direction of a monotonic relationship between two variables. Here’s the interpretation of the output you provided:

Correlation Coefficient (rho): The calculated Spearman’s rank correlation coefficient (rho) is approximately 0.4524.Interpretation: The value of rho is positive, indicating a positive relationship between WomenClothingStores and FamilyClothingStores. As one variable increases, the other tends to increase as well.

Test Statistic (S): The test statistic (S) is 46. Interpretation: This statistic is used in the calculation of the p-value. It is not directly interpretable on its own.

P-value: The p-value associated with the test is 0.2675. Interpretation: The p-value is relatively high (greater than the commonly used significance level of 0.05), suggesting that there is not enough evidence to reject the null hypothesis. The null hypothesis is that there is no significant correlation between WomenClothingStores and FamilyClothingStores. In other words, the observed correlation could plausibly occur by random chance.

Alternative Hypothesis: The alternative hypothesis is that the true Spearman’s rank correlation (rho) is not equal to 0. Interpretation: The analysis is testing whether there is a significant non-zero correlation between the two variables. Sample Estimate: The sample estimate of rho is 0.4524. Interpretation: This is the estimated population value of the correlation coefficient based on the sample data. In this case, the positive value suggests a moderate positive relationship between the two variables.

In summary, based on the given output, there is a positive correlation between WomenClothingStores and FamilyClothingStores, but the evidence for this correlation is not statistically significant at the 0.05 significance level (p-value > 0.05).

corr <- cor.test(x=Final1$WomenClothingStores, y=Final1$FamilyClothingStores, method = 'spearman')  
  
corr

##   
## Spearman's rank correlation rho  
##   
## data: Final1$WomenClothingStores and Final1$FamilyClothingStores  
## S = 46, p-value = 0.2675  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.452381