APPLIED STATISTICS REPORT ON RETAIL AND E-COMMERCE SALES DATA

Muahmmad Ammar Shahid, ID: 23103156 [*1]

Word Count: 2196

1. Executive Summary

The report discusses the importance of sales forecasting in the retail sector and presents an analysis of the Walmart sales dataset. It explores various statistical methods, including Exploratory Data Analysis (EDA), Correlation Analysis, Regression Analysis, and Time Series Analysis, to gain insights into factors affecting sales. Key findings include a positive correlation between temperature and sales, an inverse correlation between unemployment and sales, and the presence of seasonality in sales data, notably a spike during the Christmas season. The report concludes with recommendations for retailers, such as running promotions before Christmas, offering summer discounts, optimizing inventory levels, and considering price adjustments based on unemployment and CPI trends to boost sales and profitability.

2. Introduction

2.1. Problem Domain

Sales forecasting holds immense importance for both e-commerce and in-store retailers. Accurate predictions are essential for efficient inventory management, enabling retailers to strike the right balance between overstocking and understocking, optimizing resources, and planning marketing strategies. Reliable sales forecasts also play a crucial role in supply chain management, helping suppliers and manufacturers plan production schedules and inventory levels. Ultimately, accurate sales forecasting is instrumental in enhancing the overall customer experience, ensuring products are available when and where customers need them, thus driving revenue and customer satisfaction in the highly competitive retail landscape.

This report will discuss and explore the different factors that can affect sales by investigating the Walmart sales dataset 4; moreover, it will uncover the complex patterns behind sales by using four well-known statistical methods, explained in Section 3. These methods will help us describe the data thoroughly and make predictions. In the results section, the findings and their implications will be discussed. The goal of this report is to illuminate the various factors influencing sales at Walmart, contributing valuable insights to the study of retail sales.

2.2. Statistical Questions

Statistical questions in retail are crucial for demand forecasting, pricing, customer segmentation, inventory management, and supply chain optimization, all leading to improved profitability and competitiveness.

1. Is there any relation between fuel price and customer price index?
2. Does running different promotions have anything to do with inflation or unemployment?
3. Do customers tend to buy more in holidays?
4. Does the unemployment rate have any effect on sales?
5. Do Weather conditions have any effect on the sales?
6. Which store have the highest sales?

3. Methodology

3.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) extracts insights from data without formal models. It focuses on central tendency (mean, mode, median), spread (standard deviation, variance), distribution shape, and outliers. The primary goal is understanding data before formal modeling or hypothesis testing (Sahoo et al., 2019). This report will employ essential visualization techniques like histograms and box plots. Histograms will be used to depict data distribution and skewness, while box plots will provide a graphical summary that effectively highlights data outliers. These tools collectively enhance data analysis and presentation.

[1]M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: <muhammad.shahid3@mail.bcu.ac.uk>.

Modern Optimisation Report, CMP 7213 - Modern Optimisation.

Table 1. Dataset variables

Variable	Statistical Type
Store	Discrete
Date	Categorial (Ordinal)
Temperature	Continues
Fuel-price	Continues
Unemployment	Ration (continues)
CPI	Ration (continues)
Holiday-Flag	Nominal
Weekly-sale	continues

3.2. Correlation Analysis

A correlation coefficient quantifies the connection between two variables, whether quantitative or categorical, revealing the relationship's strength and direction. When variables are correlated, they change together, either positively or negatively. It's a common statistical tool in this study to assess relationships between variables and understand their strength (Alsaqr, 2021). A correlation matrix, displaying Pearson's Correlation Coefficient, will be used to examine linear relationships and trends among the attributes, enabling the analysis of correlations between variables.

3.3. Regression Analysis

This report focuses on analyzing changes in weekly sales, treated as the dependent variable, within the context of a multiple linear regression model. This report aims to predict (dependent variable) using various predictors (independent variables). By leveraging the information gathered from the data, this approach helps to forecast potential outcomes based on the influence of these independent variables.

3.4. Time Series Analysis

Time series analysis (TSA) is a statistical technique for studying data collected or recorded over time, such as stock prices, weather patterns, or sales figures. It helps identify underlying trends, patterns, and seasonality, facilitating forecasting and decision-making. Key components include data decomposition, trend and seasonality detection, and model fitting, making it valuable for understanding and predicting time-dependent phenomena. This report will conduct TSA to find trends in sales over time.

4. Data-set

This report uses the "Walmart sales forecasting" (WSF) dataset from Kaggle. Walmart is a multinational retail corporation based in the United States. As of January 2023, it had 10,623 retail stores throughout the world. The WSF dataset was released by Walmart in 2014 for a competition where Walmart aims to predict department-wide sales for 45 stores using historical sales data (from 05,02,2010 to 26,10,2012).

The provided data contains information about sales in 45 different Walmart stores located in various regions. It includes details on department-level sales within these stores. Additionally, the dataset accounts for promotional markdown events that precede significant holidays, such as the Super Bowl, Labor Day, Thanksgiving, and Christmas. These holiday weeks are given higher importance in the evaluation process, emphasizing the need to accurately model the impact of markdowns on holiday sales even when complete historical data may be lacking. However, the scope of this report is not limited to sales forecasting; this study will conduct descriptive analysis to get useful insights.

The original dataset consisted of 6436 records, and a different file with stores information only. Features along with the target variable will be used in this report. 1 shows the variables and their types. **Store** represents the different stores, **CPI** represents Customer price index (inflation or average change of price over time), **Unemployment** represents the unemployment rate, **Is-holiday** indicates holiday or not, **Temperature** represents the temperature of the day, and **Weekly-sales** represents the amount of money earned by the store in the previous week. **Markdown** columns represent Walmart's promotional markdown data, available from November 2011 onward, contain missing values marked as "NA," and it is not consistently available for all stores. In terms of values in markdown, it is assumed that it represents the spending.

4.1. Preprocessing

As mentioned above there is very little data about Markdowns and contains so many null values, To gain insights, missing values (NA) in the promotional markdown data are treated as indicating no ongoing promotional campaign at that time, and they are replaced with 0. Additionally, for a different perspective, rows containing NA values are excluded to assess their impact on the analysis. The rest of the data does not have any missing values. Due to huge missing values, Markdowns would not be used except to see their Correlation with CPI, and Temperature.

5. Results and Discussion

1. **Exploratory Data Analysis**

   Figure 1 shows the summary of the data in detail. The minimum sale is 209986 and the maximum sale is 3818686 Notably median is lower than the mean which shows that its distribution is positively skewed. Similarly, the minimum temperature is -2.06 and the maximum is 100.14. The highest fuel price in the data is 4.468.

Store	Date	Weekly_Sales	Holiday_Flag
Min. : 1	Length:6435	Min. : 209986	Min. :0.00000
1st Qu.:12	Class :character	1st Qu.: 553350	1st Qu.:0.00000
Median :23	Mode :character	Median : 960746	Median :0.00000
Mean :23		Mean :1046965	Mean :0.06993
3rd Qu.:34		3rd Qu.:1420159	3rd Qu.:0.00000
Max. :45		Max. :3818686	Max. :1.00000

Temperature	Fuel_Price	CPI	Unemployment
Min. : -2.06	Min. :2.472	Min. :126.1	Min. : 3.879
1st Qu.: 47.46	1st Qu.:2.933	1st Qu.:131.7	1st Qu.: 6.891
Median : 62.67	Median :3.445	Median :182.6	Median : 7.874
Mean : 60.66	Mean :3.359	Mean :171.6	Mean : 7.999
3rd Qu.: 74.94	3rd Qu.:3.735	3rd Qu.:212.7	3rd Qu.: 8.622
Max. :100.14	Max. :4.468	Max. :227.2	Max. :14.313

Figure 1. Data Summary

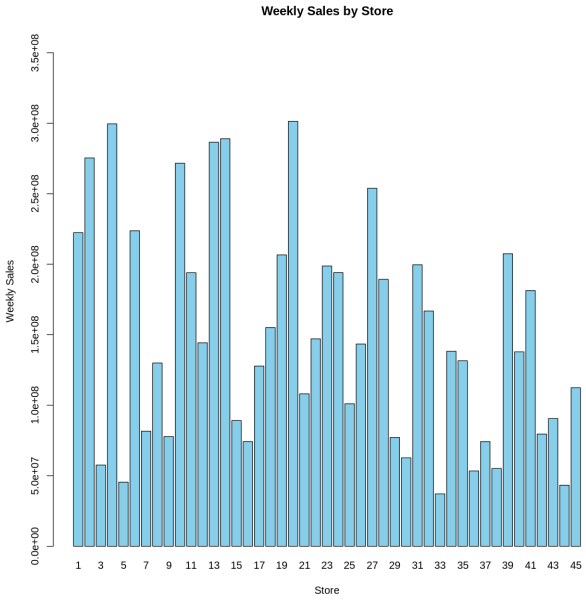

Figure 2. Weekly sales with respect to Stores

In figure 2 all the stores are listed on the x-axis with correspondence to Weekly sales on the y-axis. It can be clearly seen that Stores 4 and 20 have the highest sales; likewise, store 33 have the lowest sale.

Figure 3 shows a right skewed histogram. The right-skewed histogram of "Weekly-Sales" concerning the "Holiday Flag" indicates that most weeks have relatively modest sales, typical for non-holiday periods. However, holiday weeks exhibit a rightward tail, signifying that some weeks experience exceptionally high sales. These outliers may be due to holiday promotions or special events. This skewness implies that the mean sales during holidays are higher than during non-holidays. Understanding this distribution is crucial for optimizing holiday sales strategies and capitalizing on

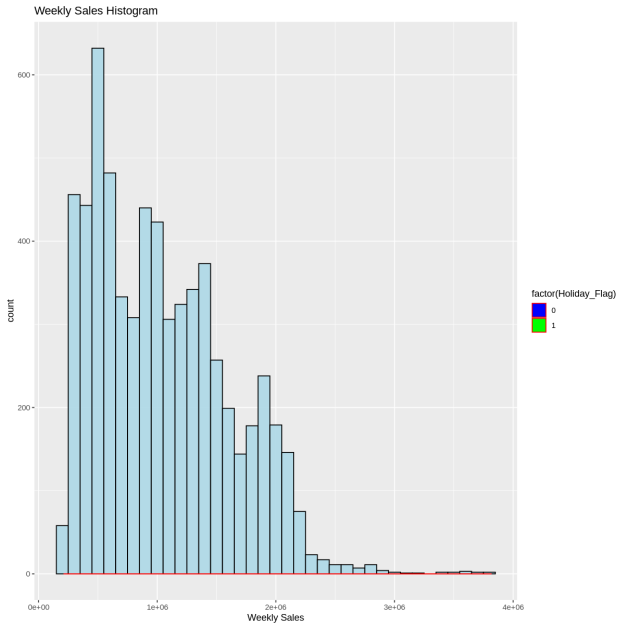

Figure 3. Weekly sales with histogram

high-demand periods.

Usually in the summer, people have holidays and tend to go out more which can result in more spending. Figure 4 is evident about this data that as the temperature increases people shop more.

Similarly, Figure 5 shows the relationship between two continuous variables i-e Unemployment and Weekly sales. It can be inferred from the data that as the unemployment rate rises, there is a subtle reduction in sales. This trend may be attributed to decreased consumer spending resulting from financial constraints caused by unemployment.

In the context of Figure 6, the box plot for Fuel Price exhibits a relatively narrow interquartile range (IQR) and lacks any outlier, indicating a relatively stable distribution of data. Conversely, the Unemployment box plot displays outliers on both ends, suggesting the presence of extreme values. Furthermore, the Temperature box plot reveals only two outliers, indicating a moderate degree of variation. Notably, the CPI stands out with a larger IQR compared to the other variables, signifying greater dispersion in its distribution. These statistical observations provide insights into the data's distribution and the presence of potential anomalies.

2. **Correlation Analysis**

In the analysis, missing values were handled in the promotional markdown data by treating them as an absence of ongoing promotions, and replacing them with

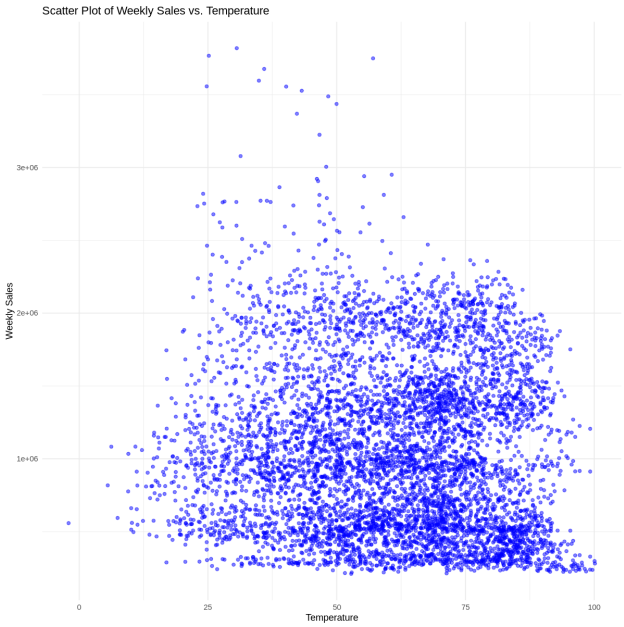

Figure 4. Weekly sales vs Temperature

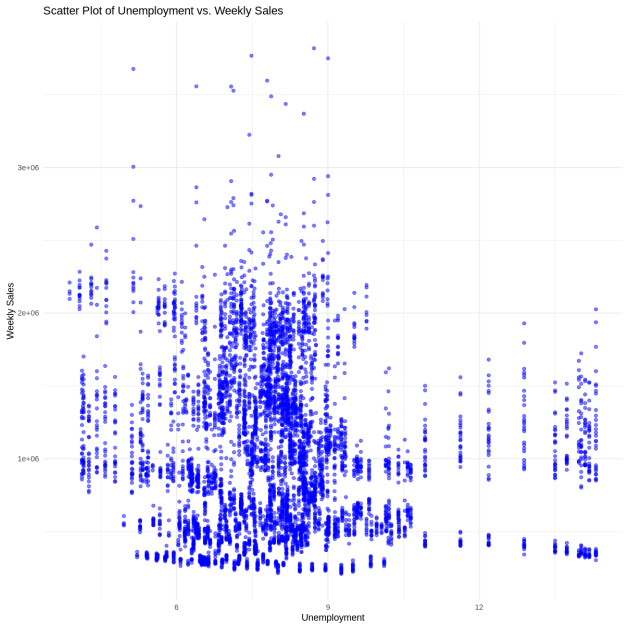

Figure 5. Weekly sales vs unemployment

zeros. Alternatively, another analysis was conducted by excluding rows containing NA values. The results revealed no significant distinction in the outcomes between these two approaches. This suggests that the presence or absence of promotional markdowns had minimal impact on the variables of interest, namely CPI, holiday, and temperature, underscoring their independence from the markdown data.

In Figure 7, a notable finding emerges: a strong correlation between Markdown1 and Markdown4. This suggests a robust association, indicating that these specific promotional events tend to co-occur and share equal significance. Conversely, both CPI and Temperature exhibit an absence of meaningful relationships with any of the Markdowns, implying their limited impact on promotional campaigns.

Additionally, it is noteworthy that Temperature displays a moderate negative association with Markdown2, suggesting a reluctance for this event to coincide during the summer season. This nuanced understanding of the data enriches our insights into the intricate dynamics between variables and their potential implications on promotional strategies.

Figure 8 shows symmetric correlations,values above and below the diagonal are perfectly symmetrical. Correlations (e.g., Cor(X, Y) and Cor(Y, X)) are similar. Although relationships exist in the correlation matrix, none are strong. This emphasizes the lack of robust linear associations, prompting exploration of other influencing factors.

A subtle positive link exists between CPI and temperature. Studies have shown that high temperature has effect on prices it could be due to Global warming fuels extreme weather, affecting food prices. Carbon-neutral goals may raise consumer prices. Higher temperatures hinder productivity and growth, limiting monetary policy options Additionally, CPI and unemployment share a moderate inverse correlation, aligning with the Phillips Curve theory. When unemployment is low, higher demand due to increased employment can drive up prices, contributing to inflation, and vice

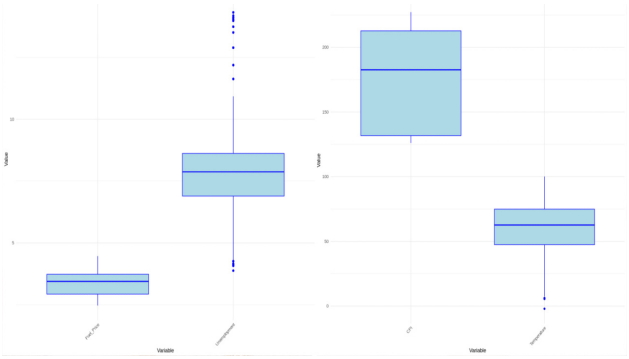

Figure 6. Box Plots

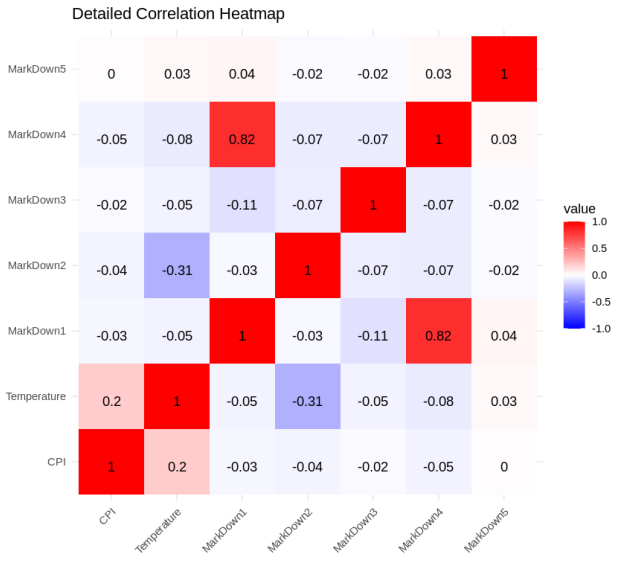

Figure 7. Markdowns Correlation Heatmap

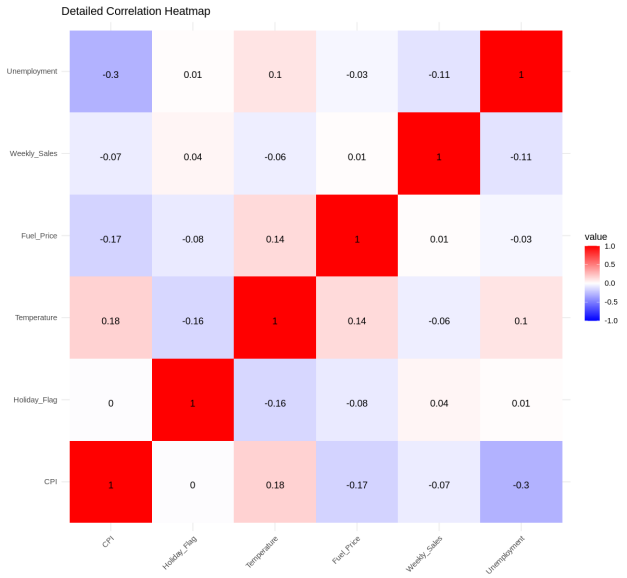

Figure 8. Correlation Heat map

versa.

.

3. **Regression Analysis** Figure 9 shows the summary of the linear model. In this multiple linear regression analysis, several key findings emerge. CPI and Unemployment display significant negative relationships with the dependent variable, with coefficients of -1,598.9 and -41,552.3, respectively. These results indicate that higher CPI and lower unemployment are associated with decreased values of the dependent variable. Additionally, Holiday-Flag has a moderate positive impact, as reflected by its coefficient of 74,891.7. The model explains approximately 2.54 percent of the variance, with an overall F-statistic of 33.57, signifying its statistical significance.

4. **Time series Analysis**

   The given ARIMA model summary in Figure 10, represents a seasonal time series model tailored to analyze the given dataset, with a yearly pattern. The coefficients indicate the strength and direction of relationships between past observations and the current value, with both autoregressive (AR) and moving average (MA) terms incorporated. Notably, a seasonal MA term with a 12-month lag (sma1) suggests a yearly seasonality factor.

   Regarding training set error measures, the model exhibits characteristics such as a Mean Absolute Error (MAE) of 86,283.68 and a Root Mean Squared Error (RMSE) of 160,172, which reflect the model's performance on the training data. The ACF1 statistic, nearly zero, suggests minimal autocorrelation in the residuals. This ARIMA model can be valuable for forecasting and understanding the underlying patterns within the time series data, particularly when dealing with seasonality and autoregressive effects.

   In addition to the previously discussed analysis, the accompanying Figure 11 provides further support for the seasonality factor present in the data. The visual representation highlights a notable spike in sales during the month of December, which corresponds to the holiday season, particularly Christmas. This surge in sales underscores a well-established consumer behavior pattern, where people tend to make more purchases than usual during this festive period. This insight can be valuable for businesses in planning inventory, promotions, and marketing strategies to capitalize on these seasonal trends and optimize sales.

6. Conclusion and Recommendations

The conducted detailed analysis of the data was able to answer all Statistical questions asked in 2.2. Weather conditions exhibit a positive correlation with sales, while no substantial evidence supports Fuel Price's impact on sales. Nevertheless, unemployment demonstrates a negative correlation with CPI, indicating higher sales during lower un-

```
lm(formula = Weekly_Sales ~ Temperature + CPI + Unemployment +
    Fuel_Price + Holiday_Flag, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1022429  -478555  -117266   397246  2800620

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1726523.4    79763.5  21.646  < 2e-16 ***
Temperature     -724.2      400.5  -1.808  0.07060 .
CPI            -1598.9      195.1  -8.194 3.02e-16 ***
Unemployment  -41552.3     3972.7 -10.460  < 2e-16 ***
Fuel_Price    -10167.9    15762.8  -0.645  0.51891
Holiday_Flag   74891.7    27639.3   2.710  0.00675 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 557400 on 6429 degrees of freedom
Multiple R-squared:  0.02544,   Adjusted R-squared:  0.02469
F-statistic: 33.57 on 5 and 6429 DF,  p-value: < 2.2e-16
```

Figure 9. Linear model summary

```
Series: my_time_series
ARIMA(4,1,5)(0,0,1)[12]

Coefficients:
          ar1      ar2      ar3     ar4      ma1     ma2      ma3     ma4
      -0.1621  -0.2035  -0.0759  0.1104  -0.2256  0.0747  -0.1503  0.0872
s.e.   0.0482   0.0469   0.0380  0.0383   0.0473  0.0377   0.0325  0.0394
          ma5     sma1
      -0.2798  -0.0276
s.e.   0.0226   0.0128

sigma^2 = 2.57e+10:  log likelihood = -86235.47
AIC=172492.9   AICc=172493   BIC=172567.4

Training set error measures:
                    ME   RMSE      MAE       MPE     MAPE      MASE
Training set -368.1103 160172 86283.68 -1.615024 8.326868 0.5108322
                     ACF1
Training set 4.301187e-05
```

Figure 10. ARIMA model summary

employment. Holiday periods witness increased shopping activity. CPI and Temperature display negligible associations with Markdowns. Moreover, by considering the following factors Company can boost their sales and enhance the customer experience.

1. Running promotions before the Christmas season can result in higher sales
2. Summer Discounts can boost the sales as customers go out more and there is also a slight rise in sales during this period
3. Maintaining adequate seasonal inventory levels can positively impact sales, ensuring products are available when customer demand peaks.
4. Conversely, reducing inventory during non-seasonal periods can lower transportation and storage expenses, optimizing cost efficiency.

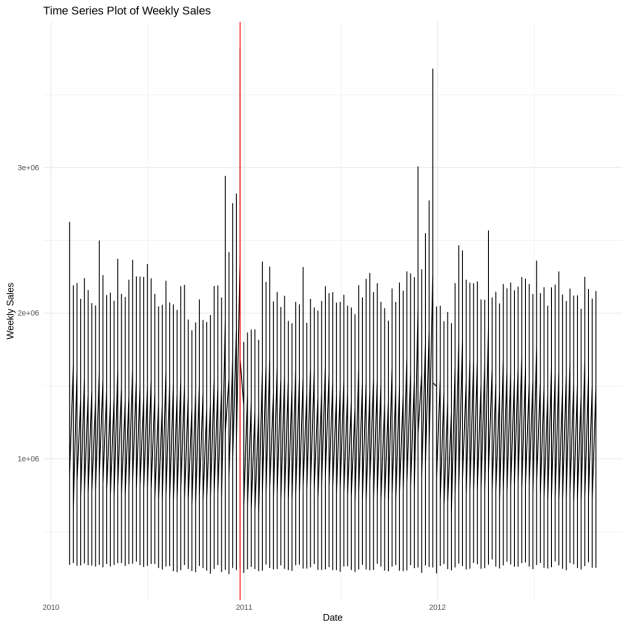

Figure 11. Time series data of Sales

5. To boost profits, consider slight price increases during low unemployment when consumer demand is high. Data also shows a correlation between low unemployment and rising CPI

References

Alsaqr, A. M. Remarks on the use of pearson's and spearman's correlation coefficients in assessing relationships in ophthalmic data. *African Vision and Eye Health*, 80 (1):10, 2021.

Sahoo, K., Samal, A. K., Pramanik, J., and Pani, S. K. Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, 8(12):4727–4735, 2019.