Walmart Sales

Exploratory Data Analysis

Wiktoria Wróbel

Table of Contents

1 Introduction

1.1 About dataset

Walmart Inc. is a global retail powerhouse, renowned for its extensive network of hypermarkets, discount department stores, and grocery outlets. During the analysis of a Kaggle dataset, valuable insights can be uncovered regarding the variables that influence weekly sales. As a result, data-driven strategies can be developed to support the optimization process and enhance business decision-making.

The dataset consists of over 6,000 records and contains 8 variables, which are described in detail below. There are no missing values in the dataset.

Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	05-02-2010	1643691	0	42.31	2.572	211.0964	8.106
1	12-02-2010	1641957	1	38.51	2.548	211.2422	8.106
1	19-02-2010	1611968	0	39.93	2.514	211.2891	8.106
1	26-02-2010	1409728	0	46.63	2.561	211.3196	8.106
1	05-03-2010	1554807	0	46.50	2.625	211.3501	8.106
1	12-03-2010	1439542	0	57.79	2.667	211.3806	8.106

Table 1: *Dataset preview*

1.2 Features description

Feature	Description
Store	Store number
Date	Sales week date
Weekly_Sales	Sales (in USD)
Holiday_Flag	Presence of a holiday
Temperature	Temperature (in Fahrenheit)
Fuel_Price	Fuel price in the region (in USD)
CPI	Consumer Price Index
Unemployment	Unemployment rate

Table 2: *Description of dataset features*

1.2.1 Store

Each store has its own number (1-45), which appears in the dataset 143 times, representing the number of weeks for which data was collected. Unfortunately, there is a lack of information about the order of the stores. On one hand, the number may be just the ID of the store randomly chosen for a particular department. On the other hand, the store number may be associated with its location.

1.2.2 Date

Data were collected continuously on a weekly basis from February 5, 2010, to October 26, 2012, across all 45 locations, with the reported day set as Friday. The data are provided in dd-mm-YYYY format. Data from all stores were collected during the same period of time and on the same days, which will be highly valuable for further analysis.

1.2.3 Weekly Sales

Sales were measured at each store on a weekly basis. Basic descriptive statistics for weekly sales are presented below.

mean	sd	median	min	max	skew	kurtosis
1046965	564366.6	960746	209986.2	3818686	0.6680502	0.0512188

Table 3: *Basic descriptive statistics for weekly sales*

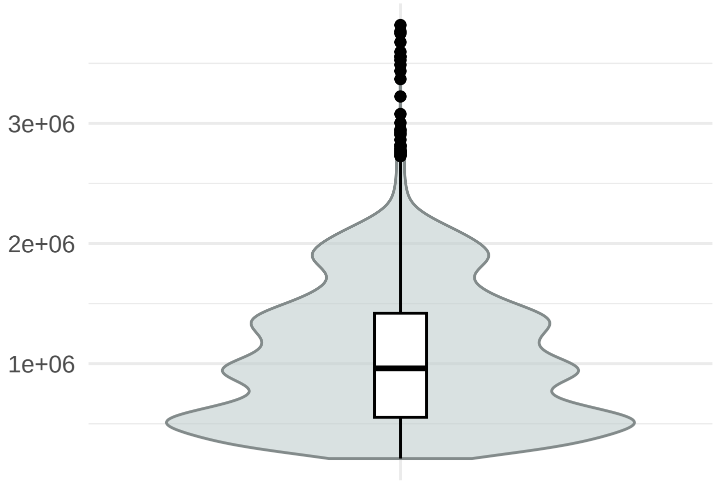

Figure 1: *Weekly sales density distribution*

The violin plot displays a right-skewed distribution, which is further confirmed by the positive skew value. The median is located near the center of the box plot, and the kurtosis is close to zero. However, there are multiple outliers above the upper whisker.

To provide clearer insights into weekly sales, the average sales (the total sales divided by the number of stores) are presented below. Nevertheless, a more complex analysis of weekly sales over time is provided in Chapter 3.

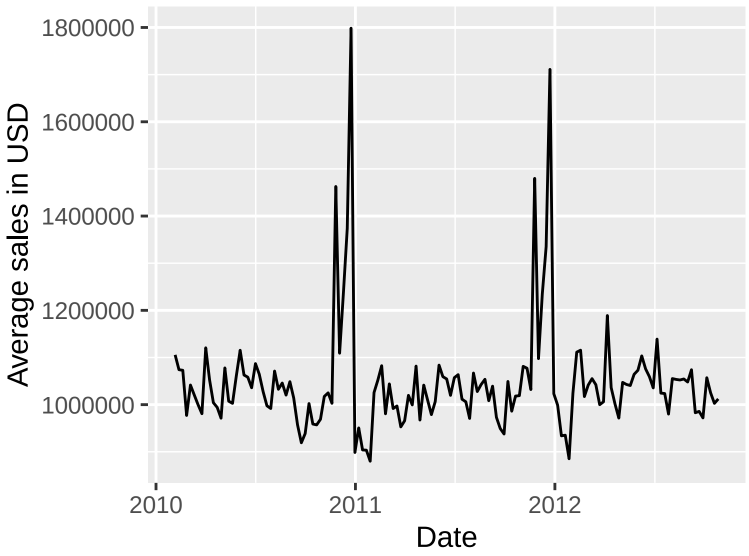

Figure 2: *Average weekly sales in USD*

The average sales remain relatively stable throughout the year. Nonetheless, two noticeable spikes occur at the end of each year. Seasonal peaks may be attributed to the Christmas period, which likely results in higher demand of goods.

1.2.4 Holiday Flag

The column contains two values: 0 indicates the absence of holidays during the week, while 1 signifies that holidays fall within that week.

The presence of holiday per shop	Frequency
0	133
1	10

Table 4: *Frequency of holidays*

With an occurrence of just 6.99%, holiday days make up the minority of the weeks. Holidays refer to the celebrations such as Super Bowl, Labor Day or Christmas.

Date	Average_Weekly_Sales	Holidays
2010-02-12	1074148.4	Super Bowl
2010-09-10	1014097.7	Labor Day, Back-To-School
2010-11-26	1462689.0	Thanksgiving, Black Friday
2010-12-31	898500.4	Christmas
2011-02-11	1051915.4	Super Bowl
2011-09-09	1039182.8	Labor Day, Back-To-School
2011-11-25	1479857.9	Thanksgiving, Black Friday
2011-12-30	1023165.8	Christmas
2012-02-10	1111320.2	Super Bowl
2012-09-07	1074001.3	Labor Day, Back-To-School

Table 5: *Presence of holidays*

Without a doubt, the week of Thanksgiving and Black Friday generates the highest sales among all holidays. Furthermore, all holiday weeks produce greater income than the median of weekly sales.

1.2.5 Temperature

The Fahrenheit temperature has been changed to Celsius for easier understanding.

mean	sd	median	min	max	skew	kurtosis
15.92432	10.24718	17.03889	-18.92222	37.85556	-0.3366106	-0.6139989

Table 6: *Basic descriptive statistics for temperature*

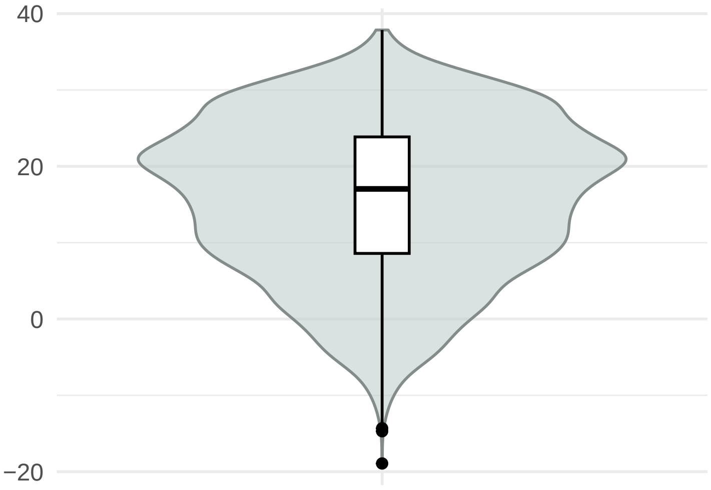

Figure 3: *Temperature density distribution*

The median temperature is slightly above 17°C, indicating a slightly left-skewed distribution. The negative kurtosis suggests a flatter shape compared to a normal distribution. There are a few outliers below the lower whisker, indicating some exceptionally low temperatures. The temperature range is relatively wide, with values spanning from -18.9°C to 37.86°C.

1.2.6 Fuel price

mean	sd	median	min	max	skew	kurtosis
3.358607	0.4590197	3.445	2.472	4.468	-0.0961135	-1.177962

Table 7: *Basic descriptive statistics for fuel price*

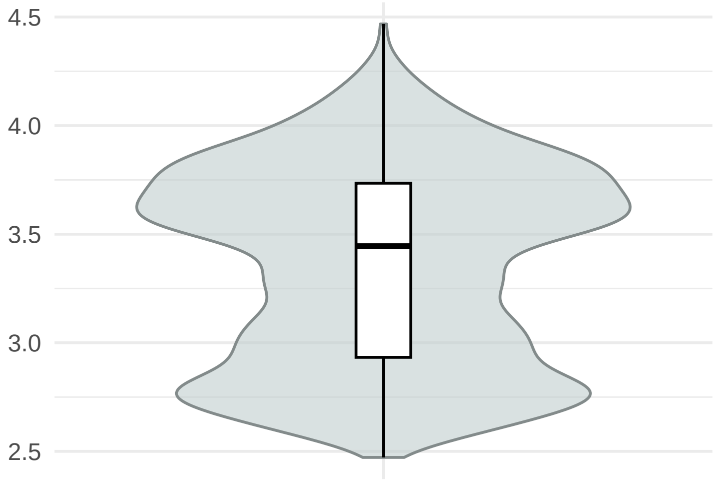

Figure 4: *Fuel price density distribution*

The median closely aligns with the mean, suggesting a nearly symmetric distribution. However, the violin plot reveals two distinct density peaks, indicating potential bimodality in the fuel price distribution. Additionally, there are no outliers present.

1.2.7 CPI

mean	sd	median	min	max	skew	kurtosis
171.5784	39.35671	182.6165	126.064	227.2328	0.0634623	-1.839677

Table 8: *Basic descriptive statistics for CPI*

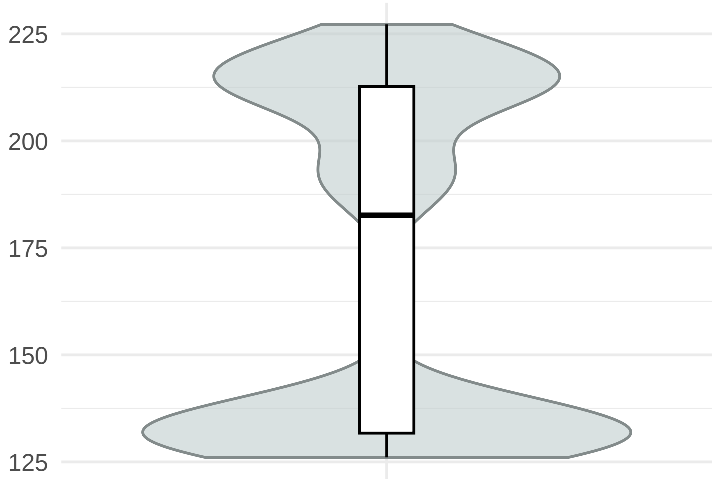

Figure 5: *CPI density distribution*

The violin plot reveals a bimodal distribution, indicating two distinct clusters of CPI values rather than a single central peak. The density is higher around the upper and lower extremes. Negative kurtosis suggests a distribution that is spread out and less peaked. However, the violin plot also shows specific areas of higher concentration. Additionally, the relatively high standard deviation reflects a wide dispersion of values, while the median being slightly above the mean indicates a slight right skew in the data.

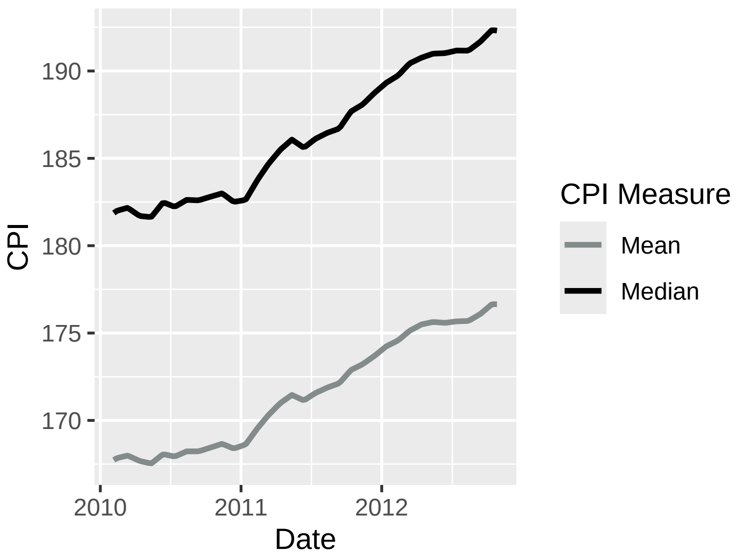

Figure 6: *CPI in time*

Over time, the CPI demonstrates a consistent upward trajectory, indicating a gradual rise in prices. This persistent increase suggests a long-term trend of inflation during the observed period.

1.2.8 Unemployment

mean	sd	median	min	max	skew	kurtosis
7.999151	1.875885	7.874	3.879	14.313	1.18759	2.634977

Table 9: *Basic descriptive statistics for unemployment*

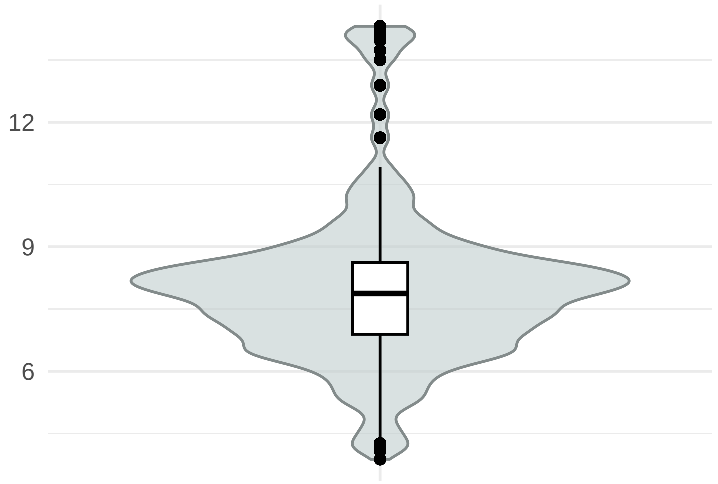

Figure 7: *Unemployment density distribution*

The distribution is right-skewed, with multiple outliers present both below the lower whisker and above the upper whisker. The high kurtosis indicates a more peaked distribution with pronounced tails. Despite this, the median and mean remain relatively close to each other, suggesting a balanced central tendency.

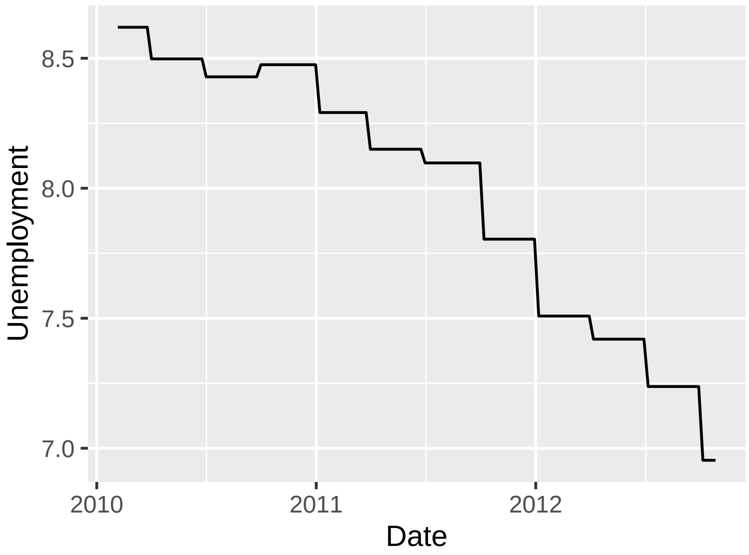

Figure 8: *Unemployment in time*

During the investigated period, the unemployment rate exhibited a downward trend, continuing throughout the entire timeframe.

1.3 Exploratory Data Analysis

1.3.1 Evaluating the impact of holidays on weekly sales

Holiday_Flag	mean	sd	min	max
0	1041256	558957.4	209986.2	3818686
1	1122888	627684.9	215359.2	3004702

Table 10: *Basic descriptive statistics for weekly sales by holidays*

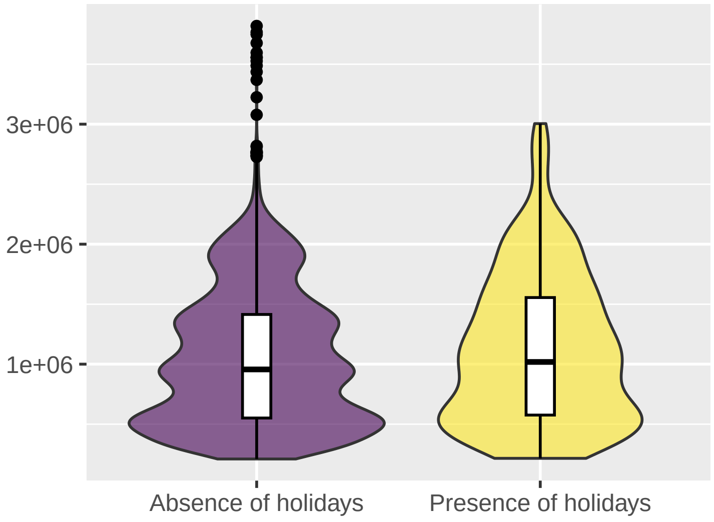

Figure 9: *Weekly sales group by presence of holidays*

On average, weekly sales are higher during holidays. The violin plot indicates that while non-holiday sales experience more extreme outliers, holiday sales are more consistently elevated. Although peak sales occasionally occur outside of holiday periods, holidays provide a more predictable increase in sales.

1.3.2 Connection between the number of the store and CPI

CPI may be determined by the distribution of shops. Before analysing potential impact of that features in weekly sales, valuable insight of the correlation between that two variables may be highly rewarded.

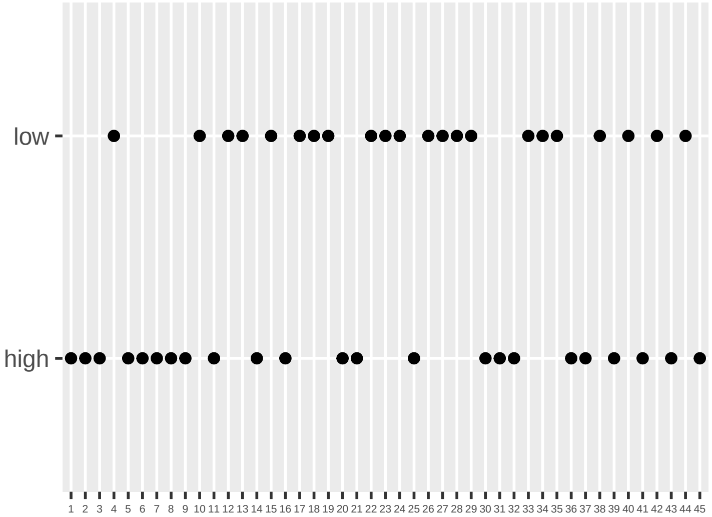

Figure 10: *Weekly sales group by presence of holidays*

The occurrence of the CPI may show some association with the number of stores, especially considering the lower CPI values observed in the central range. However, due to the limited information on how the stores are numbered, it becomes challenging to draw strong conclusions about their potential impact on one another. With this lack of detailed data, there is insufficient evidence to confidently claim a significant relationship between the two features.

1.3.3 Examining the relationship between CPI and weekly sales

Due to the specific distribution of CPI, categorizing it into two distinct groups (lower and higher) could be valuable for further investigating potential differences in weekly sales. This approach would facilitate a more focused analysis, helping to uncover any significant variations between the groups.

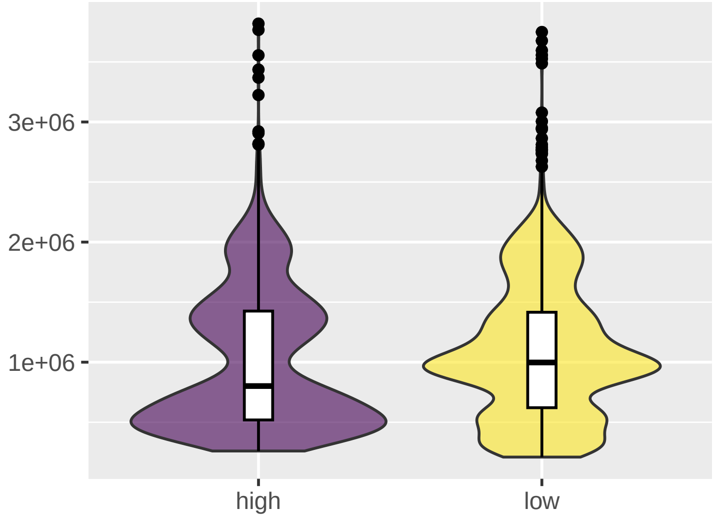

Figure 11: *Weekly sales group by CPI*

Both violin plots exhibit outliers above the upper whisker and display right-skewed distributions. The key difference lies in their distribution—while the high CPI group tends to have a lower median, the low CPI group appears more centered.

The Spearman correlation coefficient -0.05504 indicates little to no correlation between CPI and Weekly Sales.

1.3.4 Determining the relationship between fuel price and weekly sales.

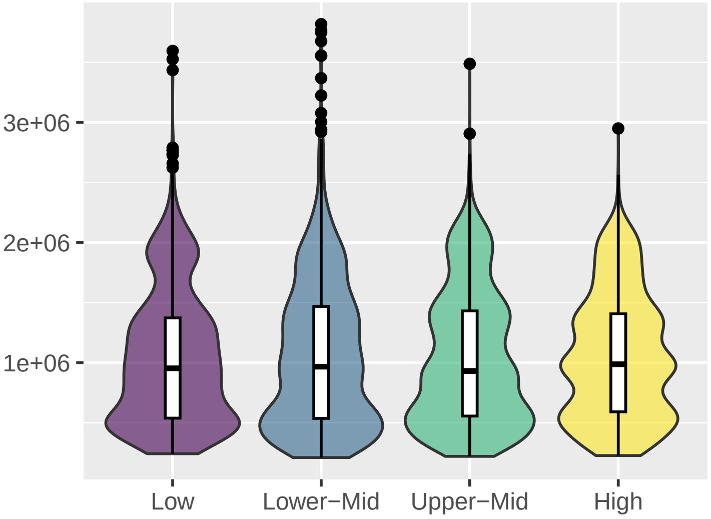

Figure 12: *Weekly sales group by fuel prices groups*

Dunn Test

$\mathcal{H}_0$: No significant difference in the distribution amont the groups,
$\mathcal{H}_1$: Significant difference in the distribution among the groups.

group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
Low	Lower-Mid	1612	1606	0.42950	0.66756	1.00000	ns
Low	Upper-Mid	1612	1615	0.07730	0.93839	1.00000	ns
Low	High	1612	1602	1.38970	0.16462	0.98773	ns
Lower-Mid	Upper-Mid	1606	1615	-0.35248	0.72448	1.00000	ns
Lower-Mid	High	1606	1602	0.95957	0.33727	1.00000	ns
Upper-Mid	High	1615	1602	1.31316	0.18913	0.98773	ns

Table 11: *Dunn test for weekly sales divided into fuel price groups*

There is insufficient evidence to reject the null hypothesis in any case. At the 0.05 significance level, no significant differences were observed between the groups.

1.3.5 Analyzing temperature trends and weekly sales

Using all data, Spearman correlation coefficient -0.0709622 suggests that there is little to no relationship between Temperature and Weekly Sales.

Temperature is categorized based on its quantiles. This approach allows for a clearer understanding of the distribution and variation within different temperature ranges.

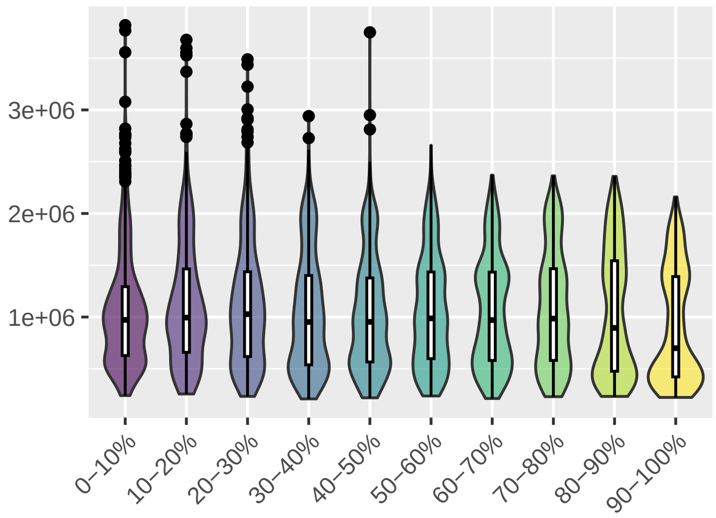

Figure 13: *Weekly sales group by temperature groups*

Outliers tend to occur in colder weather. However, this may be due to the increase in sales during the winter months, as shown earlier. Additionally, the distribution appears to be more concentrated in lower temperatures, whereas a bimodal pattern emerges as temperatures rise.

Dunn Test

$\mathcal{H}_0$: No significant difference in the distribution among the groups,
$\mathcal{H}_1$: Significant difference in the distribution among the groups.

group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
0-10%	90-100%	644	644	-5.67998	0.00000	0.00000	****
10-20%	80-90%	643	643	-3.26623	0.00109	0.03924	*
10-20%	90-100%	643	644	-7.17482	0.00000	0.00000	****
20-30%	90-100%	645	644	-6.79640	0.00000	0.00000	****
30-40%	90-100%	643	644	-4.72534	0.00000	0.00009	****
40-50%	90-100%	643	644	-4.55848	0.00001	0.00020	***
50-60%	90-100%	643	644	-5.70006	0.00000	0.00000	****
60-70%	90-100%	643	644	-5.62955	0.00000	0.00000	****
70-80%	90-100%	644	644	-6.29545	0.00000	0.00000	****
80-90%	90-100%	643	644	-3.90732	0.00009	0.00345	**

Table 12: *Dunn test for weekly sales divided into temperature groups*

Only groups with significant differences are presented. The results show that the majority of varying groups are those between cold and warm temperatures. As mentioned before, this may be due to increased sales during the colder months.

1.3.6 Exploring unemployment and Weekly Sales Patterns

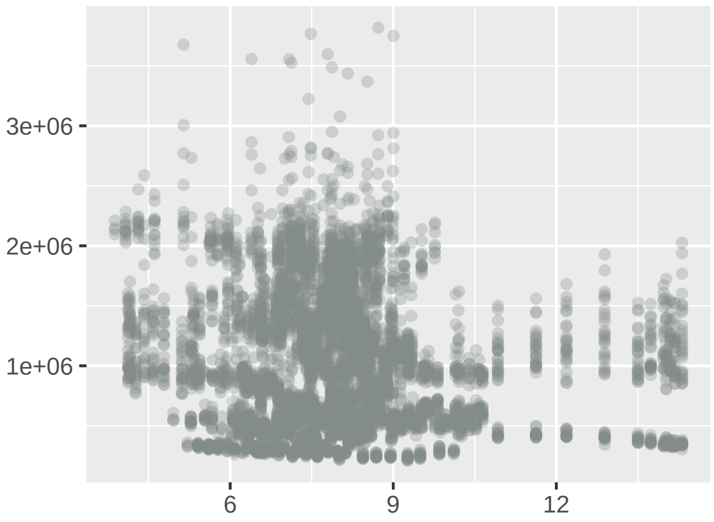

Figure 14: *Unemployment and weekly sales*

Spearman Correlation

The Spearman rate at -0.0623538 suggest, that there is no much correlation between the unemployment and weekly sales. That is reinforced with the plot above.

1.4 Seasonality

While investigating the seasonality and trend, the mean of weekly sales from 45 stores was provided to all data.

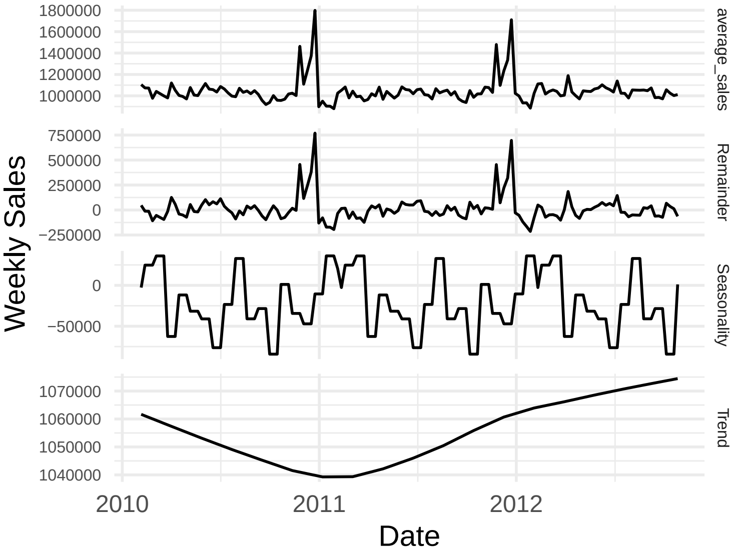

Figure 15: *Weekly sales decomposition*

As mentioned before, large spikes occur at the end of a year, which likely correspond to holiday shopping periods. Slight downward trend from 2010 to 2011, after that the trend starts increasing. Additionally, the clear seasonal pattern repeats consistently over time. Peaks suggest periodic variations in sales, related to shopping cycles, such as holidays, Christmas or back-to-school season.

1.5 Conclusion

1.5.1 Main inference

The analysis revealed that Walmart's sales exhibit clear seasonal patterns. The highest sales spikes were recorded during holiday periods, particularly in the weeks of Black Friday and Christmas.

Average sales during holiday weeks are significantly higher than in regular weeks. The biggest increase in sales is observed around Thanksgiving and Black Friday, indicating that consumers take advantage of promotions and increase their spending during this period.

The analysis found no strong correlation between CPI or fuel price, suggesting local economic conditions may not dictate sales performance.

1.5.2 Potential strategy

1. Walmart should increase inventory for high-demand items during these periods.
2. Since holiday weeks show a predictable increase in sales, Walmart could experiment with targeted discounts before major holidays to drive even higher traffic.
3. Since fuel prices and unemployment rates show little correlation with weekly sales, Walmart should not over-prioritize these factors in pricing or inventory decisions. Instead, focus on holidays and seasonality.