ADSC1000

Statistical Data Analysis

Thompson Rivers University

Fall-2023

Walmart Sales Forecasting

Submitted To:

Prof. Sean Hellingman

Submitted By:

Akansha Bhargavi-T00736533

Solomon Maccarthy- T00734513

Viswateja Adothi-T00736529

Table of Contents

Summary

The Walmart Sales Forecasting project is undertaken with the primary objective of predicting future sales. Accurate sales predictions are crucial for effective inventory management and revenue calculation. This report defines the methodology, data analysis, and key findings of the project.

1.Introduction

1.1 Background

Being a giant in the retail space, Walmart faces the challenge of efficiently managing inventory and calculating revenue. Predicting sales is essential for addressing these challenges.

1.2 Objective

This project is to predict sales accurately, enabling an optimized inventory level and enhancing the processes of revenue calculations.

1.3 Significance of Sales Forecasting

The significance of the project is to find insights into behavioral changes with customers base on sales and aiding in decision-making. This presents a proactive plan to predict future demands, determining enough stock levels and increasing revenue.

2.Data Collection

2.1 Data Source

The dataset is sourced from Kaggle, including various relevant features such as Date, Weekly sales, Holidays, Departments, Store type, Temperature, Unemployment, CPI (consumer price index), Store size and Store Id.

2.2 Data Overview & Cleaning

The Walmart dataset has four CSV files – "features", "stores", "train", "test". We merged these data into a single file to integrate this information into a comprehensive dataset. This dataset was pre-processed by replacing NA values with 0 in markdown columns for smoother visualization and converting Boolean values into binary values in Is-Holiday columns.

2.3 Dataset

Dataset Feature Description

This dataset contains three different CSVs. Let's understand a brief description of each feature below

Feature	Description	Variable Type
Store	Stores numbered from 1 to 30	Categorical
Type	Store type has been provided, there are 3 types — A, B and C.	Categorical
Size	Stores size provided	Numerical
Dept	The department numbers.	Categorical
Date	Date of sales	Date
Weekly Sales	Sales for the given department in the given store.	Numerical
Is Holiday	Whether the week is a special holiday week.	Categorical
Temperature	The average temperature in the region.	Numerical
Fuel Price	The cost of fuel in the region.	Numerical
CPI	The consumer price index.	Numerical

Markdown 1-5	Markdowns Anonymized data related to promotional markdowns that Walmart is running.	Numerical
Unemployment	The unemployment rates.	Numerical

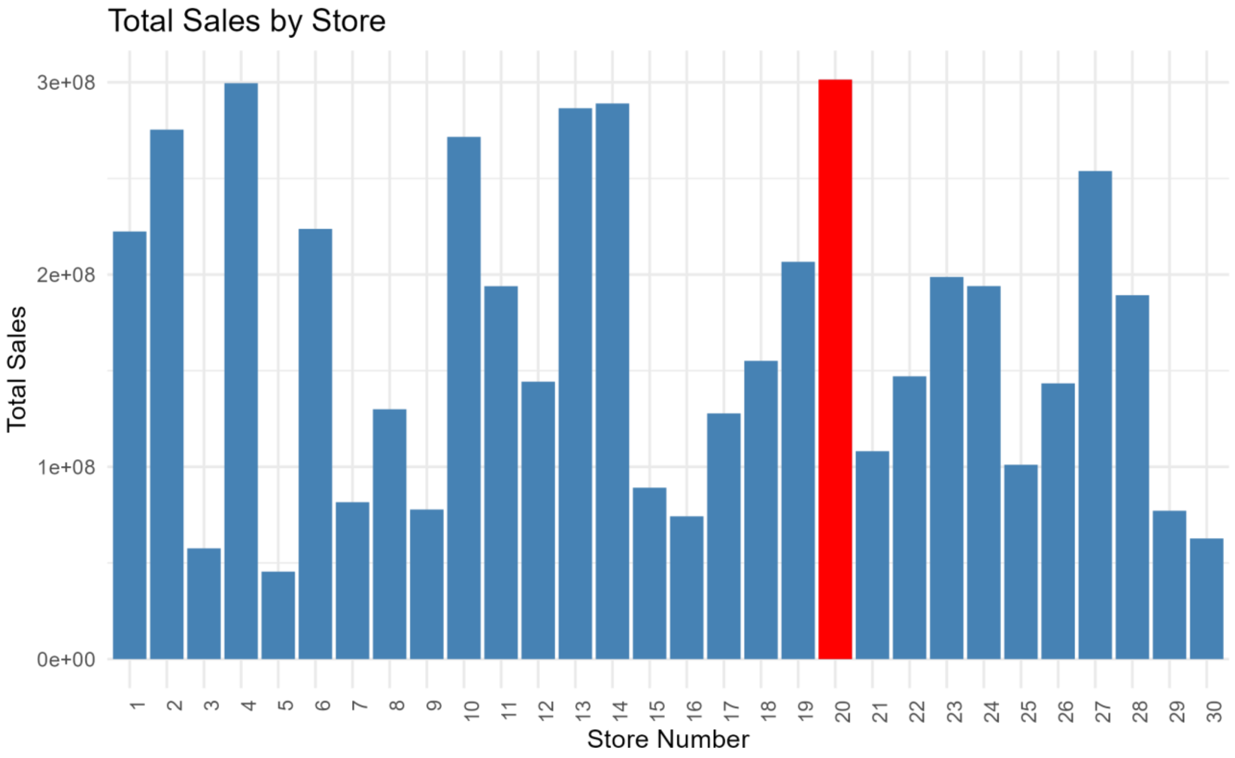
3.Exploratory Data Analysis (EDA)

3.1 Data Visualization

Visualizations, including histograms, scatter plot, bar plot, violin plot, boxplot, line plot, were employed to discover trends and associations within the data.

3.1.1 Total sales by store

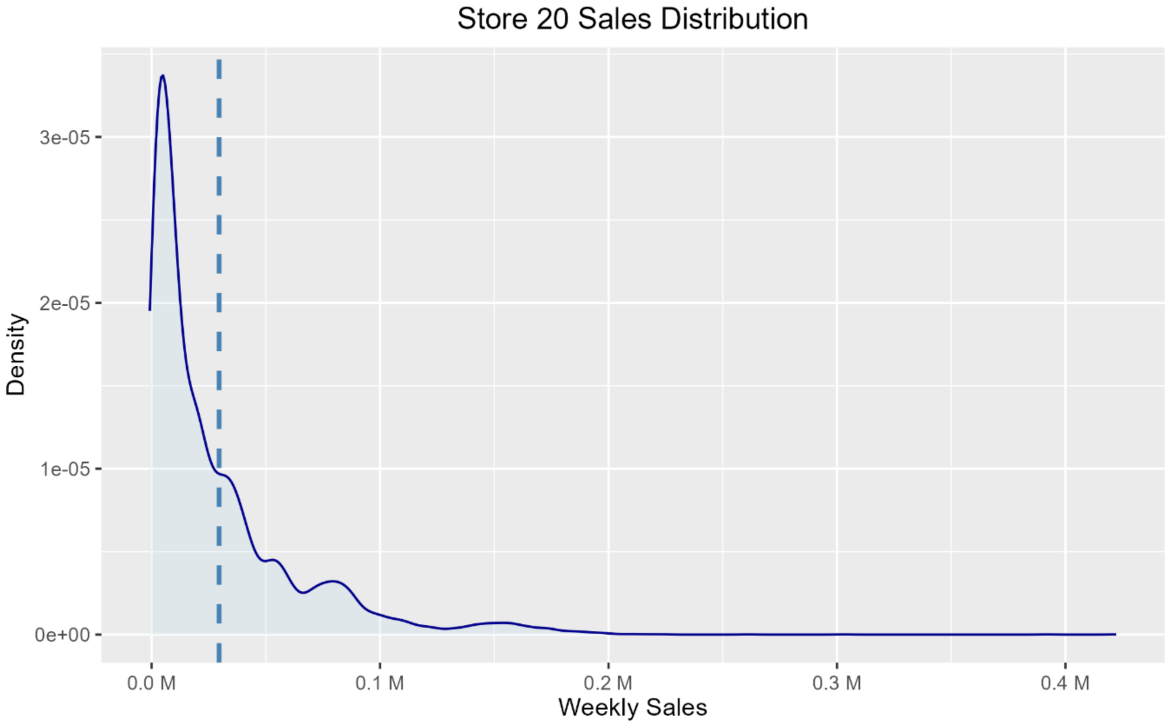
Our aim is to determine which store has the maximum sales and visualize the total sales distribution across different Walmart stores



From the visualization we can state that Store 20 stands out with the highest sales. Stores 2, 4, 10, 13, and 14 follow closely, showing similar sales performance. On the other hand, Store 5 has the lowest sales among the analyzed stores.

3.1.2 Density plot for store 20

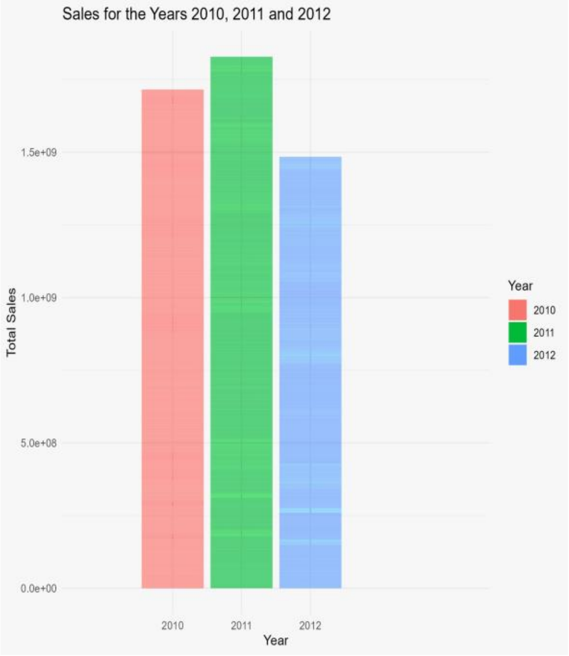
objective is to provide insights into the distribution of weekly sales for Store 20, highlighting central tendencies and identifying any notable patterns in the sales data.



From the visualization we can state that Store 20 sales are right skewed that is, It had very high sales in few weeks which resulted in increasing of Standard deviation.

3.1.3 Year with maximum sales

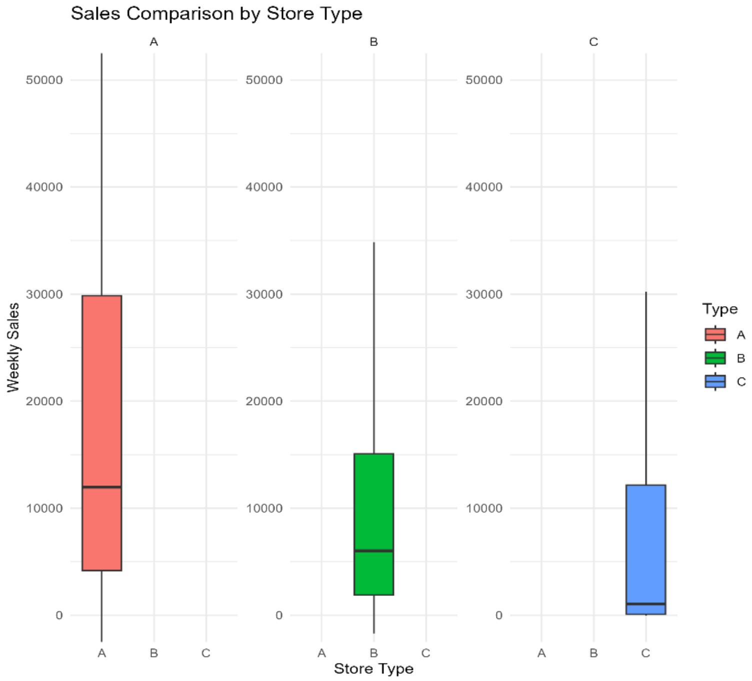
Our objective of this analysis is to identify the year with the maximum sales.



From the above visualization it is evident that the year 2011 has recorded the highest sales among the analyzed years compared to year 2010 and 2012

3.1.4 Store type with maximum and minimum sales

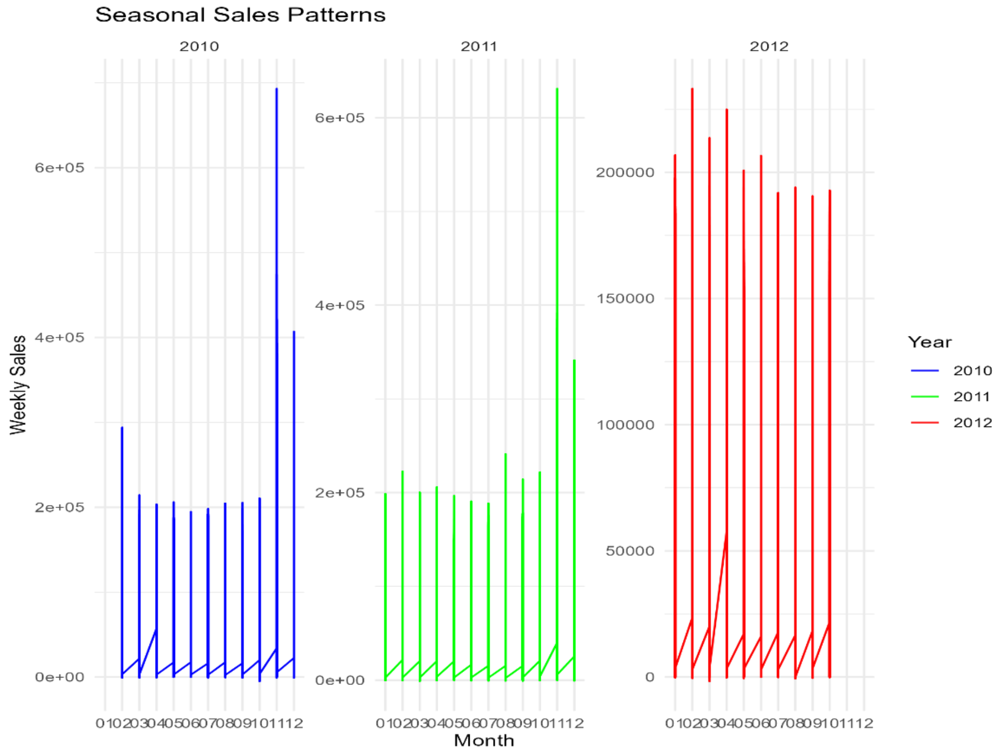
Our objective is to is to determine the store type that exhibits the maximum and minimum sales.



From visualization we can state that Store Type A records the highest sales among the various store types. Conversely, Store Type C is identified as having the minimum sales.

3.1.5 Seasonal pattern in sales based on year

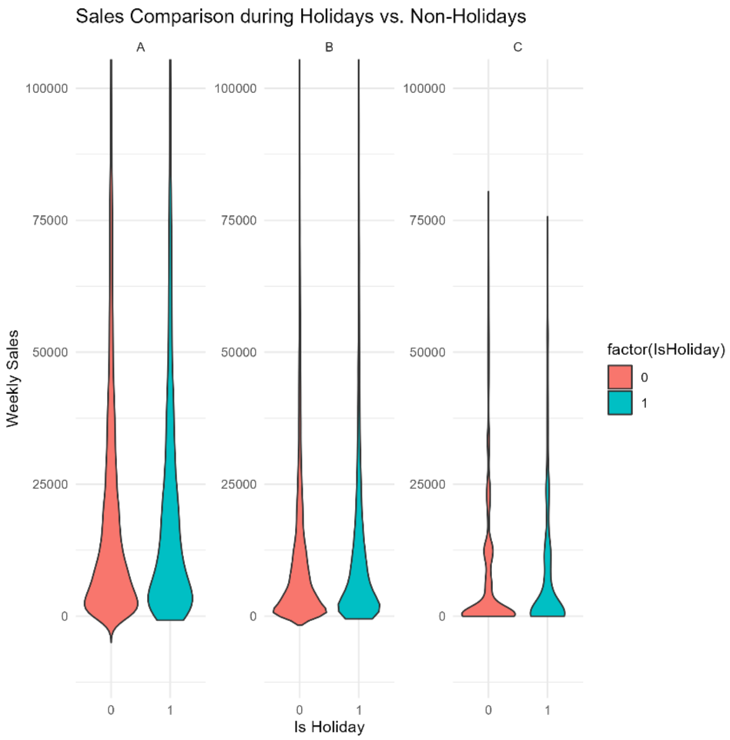
Our objective is to understand how sales vary throughout the year and whether there are recurring patterns that can be identified.



This plot provides a clear view of the sales trends for each year. We can see that in the month of November for 2010 and 2011 and February 2012 has highest sales which are almost similar. By focusing on the fluctuations in sales across different months, we can identify any seasonal patterns. This information is crucial for understanding when Walmart experiences peak sales and when there may be lower sales periods.

3.1.6 Sales during holidays and normal days

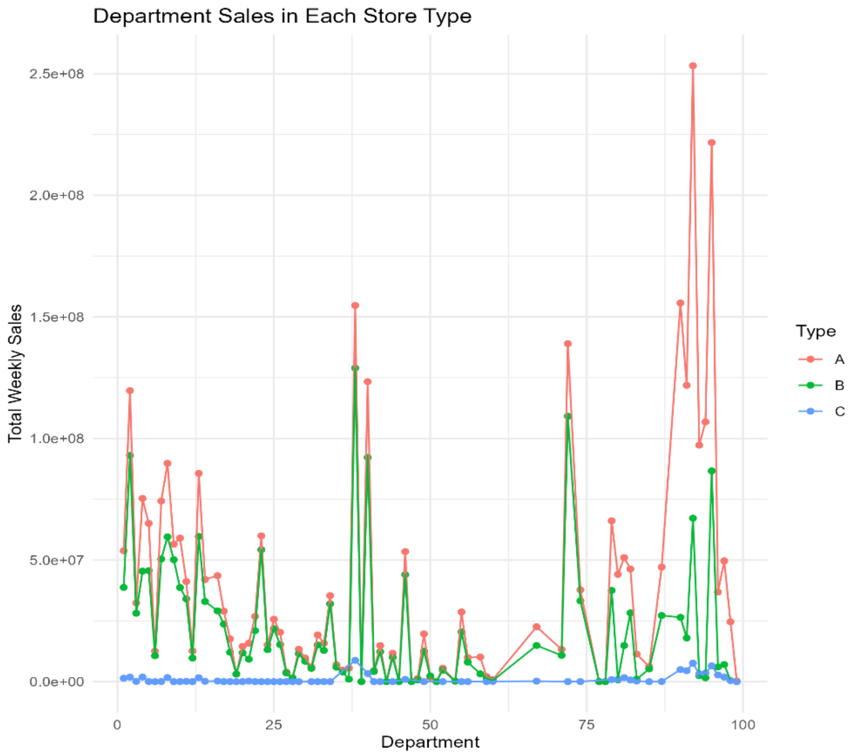
Our objective of this analysis is to compare sales during holidays and non-holidays to understand the impact of festive seasons on retail performance.



Using Violin plot it is evident that sales during holidays significantly outpace those during non-holidays.

3.1.7 Department sales in each store type

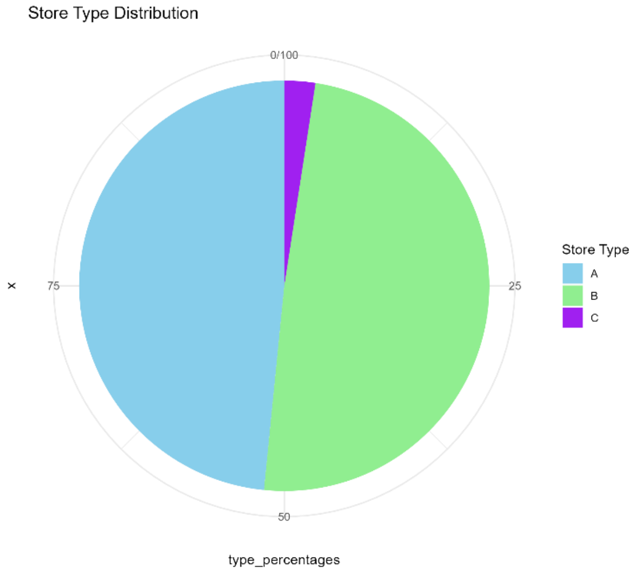
Our primary objective is to gain insights into the distribution of sales across departments within each store type and identify patterns or trends that may inform business decisions.



This scatter plot helps in understanding how sales vary not only between departments but also across different types of stores. By ordering the departments within each store type based on total sales, we can quickly identify the top-performing departments in terms of weekly sales. The use of lines connecting the points allows for the observation of patterns in sales over time for each department within a specific store type.

3.1.8 Pie chart for distribution of store types

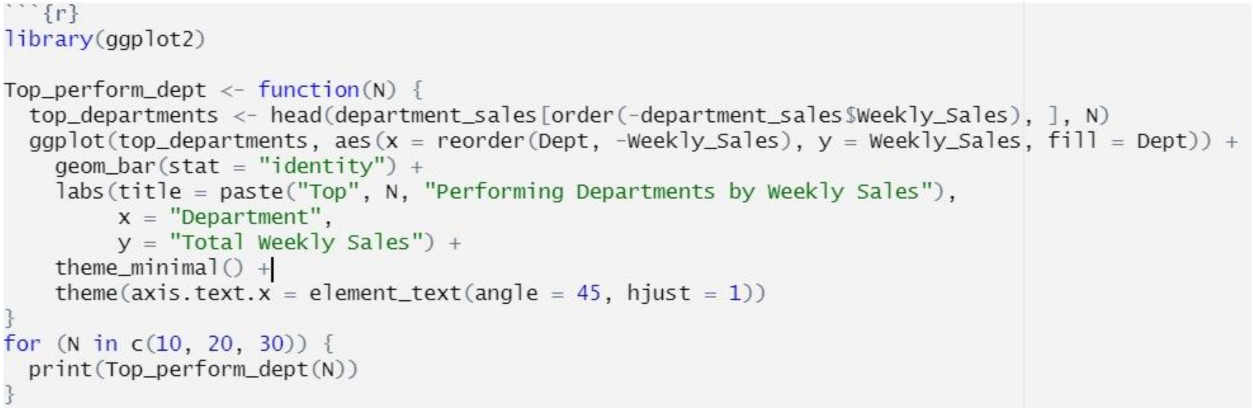
objective of this analysis is to visually represent the distribution of store types within the dataset using a pie chart.



Examining the pie chart representing store type distributions, it is evident that Store Type A and B constitutes the largest share, and Store Type C constitutes lowest share.

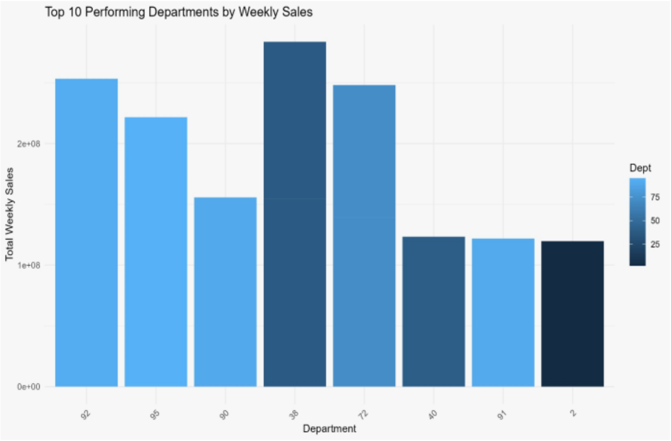
3.1.9 Department sales according to top 10 stores

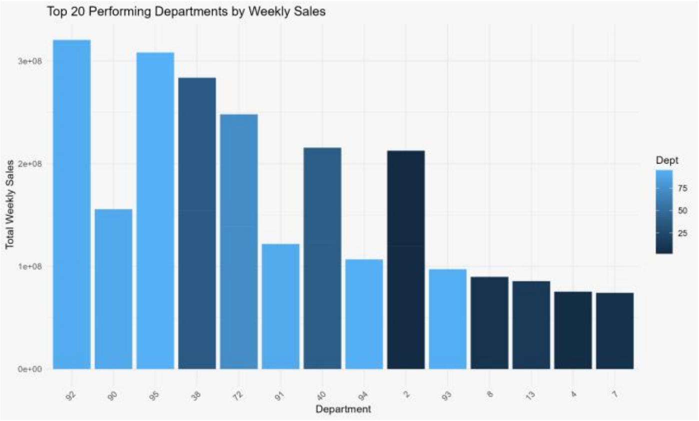
To identify top-performing departments by weekly sales.

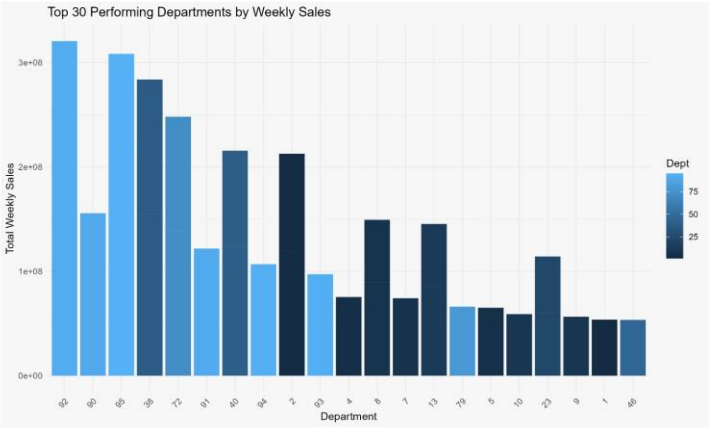


~~~~
```{r}
library(ggplot2)

Top_perform_dept <- function(N) {
  top_departments <- head(department_sales[order(-department_sales$Weekly_Sales), ], N)
  ggplot(top_departments, aes(x = reorder(Dept, -Weekly_Sales), y = Weekly_Sales, fill = Dept)) +
    geom_bar(stat = "identity") +
    labs(title = paste("Top", N, "Performing Departments by Weekly Sales"),
         x = "Department",
         y = "Total Weekly Sales") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}
for (N in c(10, 20, 30)) {
  print(Top_perform_dept(N))
}
~~~~



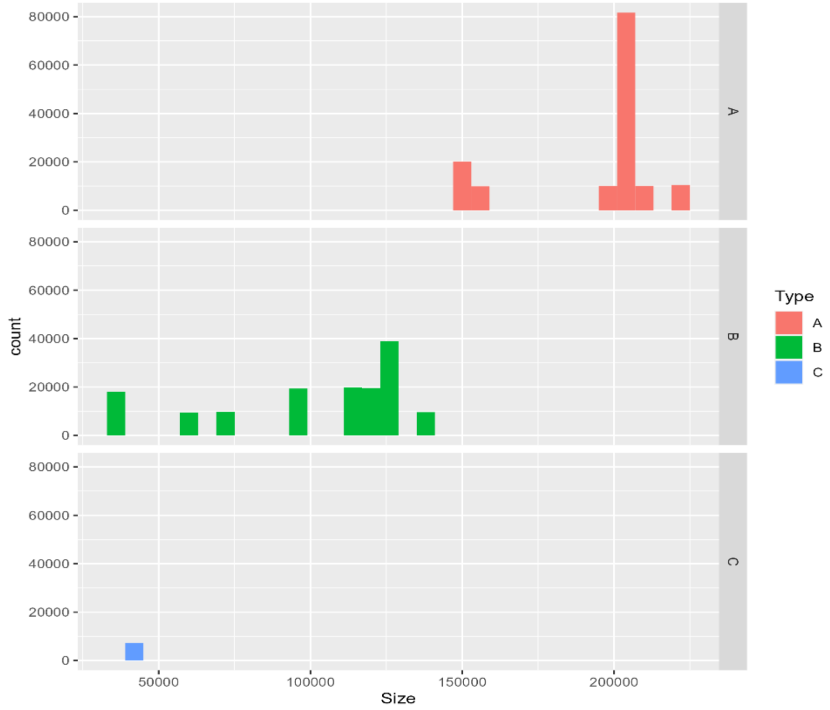




We employed a for loop to iterate through different values of N (in this case, 10, 20, and 30) and created bar plots for each scenario. we visually identify the top-performing departments in terms of weekly sales for different scenarios (top 10, top 20, and top 30).

3.1.10 Distribution of store size by type

Objective of this data visualization using ggplot is to explore and compare the distribution of store sizes across different store types.



Type 'A' is the largest format followed by Type 'B' and 'C'. Also, most of the stores in data belong to 'A' and 'B' type.

4. Results and Findings

- Stores has 3 types as A, B and C according to their sizes. Most of the stores are categorized as A.
- As expected, holiday average sales are higher than normal days.
- November sales are significantly high than other months. This is due seasonal sales like Black Friday, Thanks giving.
- Among the 30 Walmart stores "store 20" has the highest sales and year 2011 records the highest sales.
- Store A has highest sales compared to store B ,C and Department 92 in store A has highest sales
- CPI, temperature, unemployment rate and fuel price have no pattern on weekly sales.

4.1 Conclusion

These findings empower stakeholders with valuable insights for strategic decision-making in inventory management, revenue calculation, and overall business planning. The results contribute to a comprehensive understanding of Walmart's sales dynamics, guiding informed decisions for sustained growth and operational optimization.

5.Challenges and Limitations

5.1 Data Quality Issues

The presence of incomplete or inaccurate data in Walmart dataset for Markdown columns posed a challenge.

5.2 Categorical Variables

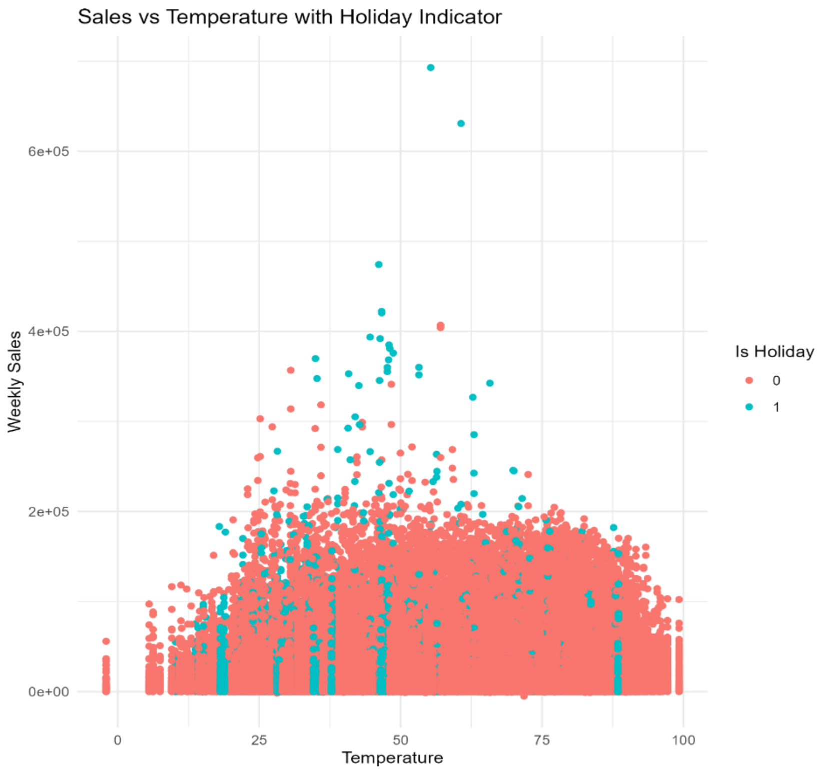
Categorical variables with a high number of unique values in columns like department and store,it lead to overcrowded bar charts, especially when trying to visualize distributions. As the dataset contain around 3 lakh observations we limited the y-axis range to focus on the most relevant portion of the data.

6.Appendix

6.1 Data Visualizations

6.1.1 Sales based on temperature

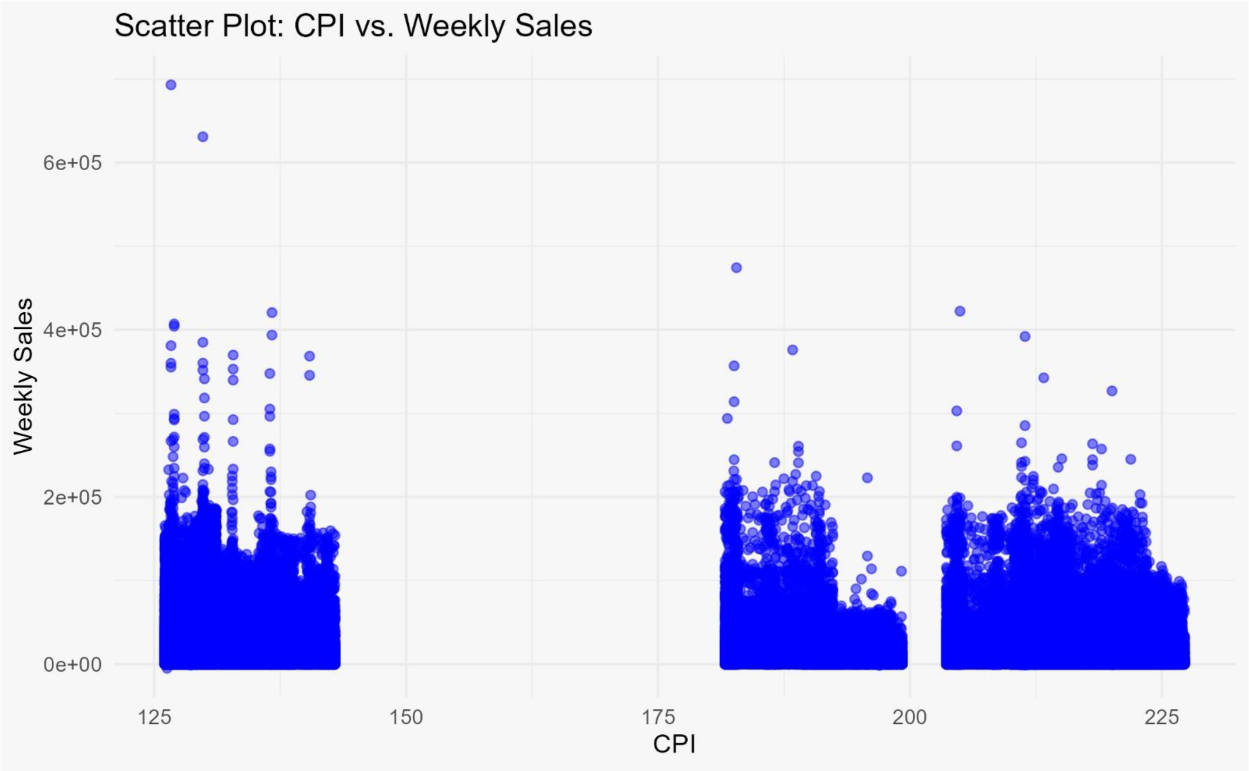
For understanding the relationship between weekly sales and temperature, while also considering the influence of holidays on these sales dynamics.



The scatter plot reveals the distribution of weekly sales in relation to temperature. We can assess whether there is a discernible pattern, By visualizing the graph we can state that there is no relation between temperature and sales.

6.1.2 Relation between CPI and sales

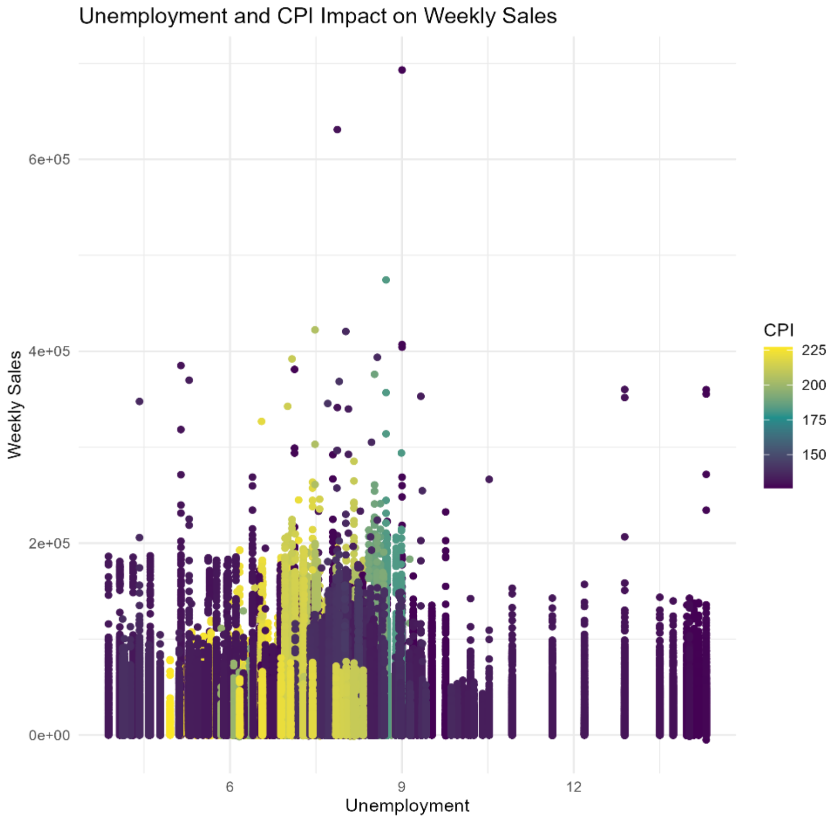
We aim to uncover insights into how changes in the CPI might impact Walmart's sales performance.



If the fitted regression line slopes upward, it indicates a positive correlation, suggesting that as the CPI increases, so do the Weekly Sales, and vice versa. Conversely, a downward slope would imply a negative correlation. Relation is not good between sales and. With increase in price index sales decreases.

6.1.3 Unemployment and CPI on weekly sales

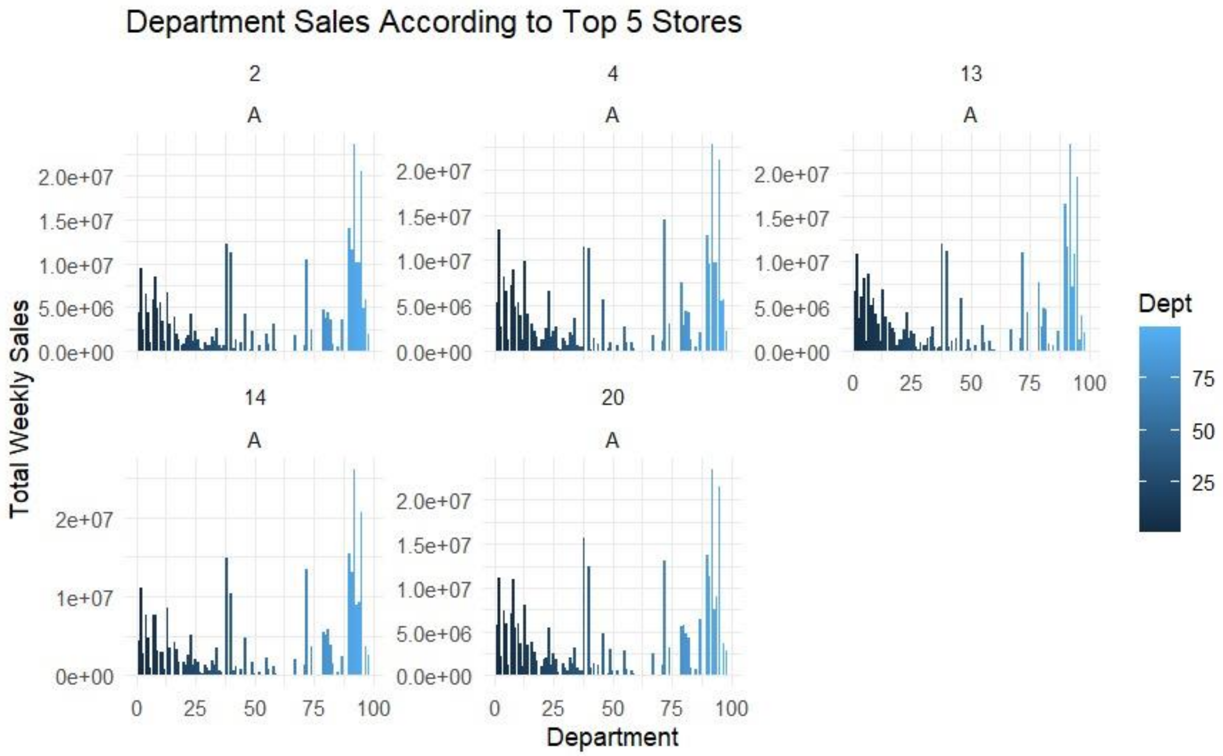
Our objective is to create a visually intuitive representation of the relationships between Unemployment, Weekly Sales, and CPI.



By visualization we can state that unemployment is not effecting sales. We can see slight variation at high values of unemployment.

6.1.4 Department sales according to top 5 stores

Objective of analyzing department sales across the top 5 stores is to gain insights into the distribution of sales among different departments within these stores.



By identifying patterns and opportunities, this information can guide strategic decisions to optimize sales, enhance customer experiences, and ensure efficient inventory management.

6.2 Code

Data_visualization_
project_final_report.

**Data_visualization_
project.rmd**

6.3 References

https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast

We collaborated together to interpret the solutions that can be predicted using Walmart dataset where we divided the work of visualization and documentation equally.