Project Write-up:

Introduction:
In this comprehensive retail sales analysis project, we delve into the wealth of historical sales data from 45 Walmart stores with a goal to predict sales and demand accurately. Walmart, a retail giant in the United States, is confronted with the challenge of unforeseen demand fluctuations and occasional stockouts. The critical issue at hand is the selection of the most appropriate machine learning algorithm to tackle this challenge. The ideal algorithm must encompass a broad spectrum of variables, including economic conditions (CPI, Unemployment Index), climatic factors (temperature), and special events such as holidays and promotional markdowns.

Data Exploration and Analysis:

Store with Maximum Sales:
Our exploration commences by identifying the store that has achieved the maximum sales over the duration of our dataset. This store, Store 20, stands out with the highest total sales of $301397792.46, demonstrating its substantial contribution to the overall sales.

Store with Maximum Standard Deviation:
Diving deeper into the data, we uncover interesting insights into the variability of sales across different stores. Store 14 exhibits the highest standard deviation of $317569.95, signifying that its sales figures vary considerably over time. Further, the coefficient of mean to standard deviation for Store 14 denoted as 15.71, highlights the inherent inconsistency in sales performance, providing an essential metric for decision-makers to consider.

Stores with Good Quarterly Growth Rate in Q3’2012:
A critical facet of our analysis focuses on identifying the stores that experienced noteworthy growth during the third quarter of 2012. Store 4 emerges as the top performer during this period, exhibiting substantial sales growth. This insight into quarterly performance aids in recognizing the stores that have successfully capitalized on specific market conditions.

Holidays with Higher Sales than Non-Holiday Season:
Holidays play a pivotal role in influencing sales, and we've determined which holidays have a notably positive impact on sales when compared to the mean sales during non-holiday weeks across all stores. Super Bowl, Labour day and Thanksgiving were found to drive higher sales, highlighting the importance of planning promotions and stock levels around these occasions.However, Christmas showed lower weekly sales that non-holiday weeks.

Monthly and Semester View of Sales:
Our analysis provides a detailed breakdown of sales on both a monthly and semester basis, enabling a comprehensive understanding of the sales trends across different units (stores) and at the aggregate level. This information is invaluable for understanding seasonality and planning inventory as well as marketing strategies. The results showed that sales peak at the end of the year (Nov, Dec) as well as during summer months (May, Jun, Jul). Semester wise, the second semester of the year shows a slight increase over the first half.

Statistical Model for Store 1:

Linear Regression and Random Forest Models:
In our quest for more accurate sales forecasts, we conducted an in-depth analysis of Store 1. We initially employed a linear regression model to predict demand, leveraging factors such as the date (transformed into days since the earliest date in the dataset). Additionally, we examined whether key economic indicators, including the Consumer Price Index (CPI), unemployment rate, and fuel prices, had a discernible impact on sales. This transformation of dates into days led to the creation of a new variable, adding another layer of granularity to our analysis.

In addition to linear regression, we also explored the application of a Random Forest model to capture complex interactions and patterns within the data. The Random Forest model takes into account multiple decision trees, offering a more robust and flexible approach to modeling sales.

Model Selection and Evaluation:
To ensure the effectiveness of our models, we assessed both linear regression and Random Forest models based on their predictive accuracy. This evaluation process aimed to determine which model best captured the relationship between our chosen variables and sales. The ultimate selection was based on the model's ability to provide the most accurate sales forecasts for Store 1, a crucial step in enhancing stock management and planning.

Model Performance:
Upon evaluation, the linear regression model yielded an accuracy of 15.21%, indicating that it did not provide a sufficiently accurate representation of the sales data for Store 1. In contrast, the Random Forest model demonstrated remarkable accuracy, achieving an impressive 93% predictive capability. This significant difference in accuracy between the two models led us to select the Random Forest model as the more suitable choice for forecasting sales in Store 1.

Conclusion:
This project offers a comprehensive analysis of Walmart's sales dataset, revealing insights crucial for decision-making and strategy development. By identifying top-performing stores, recognizing holidays with the greatest impact on sales, and ultimately selecting the Random Forest model for Store 1, we have contributed to Walmart's ongoing efforts to enhance sales forecasting, optimize stock management, and deliver better outcomes for both the business and its customers. The findings from this analysis serve as a solid foundation for data-driven decisions, providing a clear roadmap for future retail success.