Predicting Walmart Sales

Neetu Rasinger Babu, Sri Mounika Jammalamadaka, Nikhil Gudur, Priya Khandelwal
Department of Applied Data Science, San Jose State University
DATA 230: Data Visualization
Dr. Guannan Liu

Abstract

The study discusses Walmart's challenge in effectively forecasting sales and demand due to sales fluctuation caused by numerous occasions and holidays. The current machine learning algorithm has proven ineffective, leading to supply shortages brought on by unanticipated demand. Despite limited historical data, the main challenge is calculating the impact of markdowns during holidays by creating a Machine learning (ML) system that can effectively estimate demand to address this problem. The data set includes historical sales information from 45 Walmart locations across several geographies including store counts, sales by week, holiday flags, CPI, etc covering February 5, 2010, to November 1, 2012, sourced from Kaggle. The study involves dataset understanding, cleaning, regression model development to forecast sales based on variables, and model evaluation using metrics like R2 and MSE. This research aims to offer insightful analysis and practical suggestions to improve Walmart's sales forecasting accuracy.
In this study, four machine learning models have been implemented which are Linear Regression, Random Forest, Gradient Boost, and XG boost. To conclude this study it found that the XGboost model performed best with a .97 R^2 score.

Introduction

Forecasting sales is a vital task for companies looking to streamline their processes and successfully satisfy customer requests in the dynamic retail industry. One of the biggest and most powerful merchants in the world, Walmart, constantly struggles to forecast sales correctly throughout its extensive network of locations. The goal of this project is to use machine learning and sophisticated data analytics to create a reliable sales forecast model for Walmart.
Walmart works in a dynamic market that is shaped by several variables, including customer behavior, seasonal patterns, and economic situations. Precise sales forecasts allow the business to maximize stock levels, simplify the supply chain, and improve overall productivity. This project aims to give Walmart a sophisticated tool to predict sales patterns by utilizing data science, enabling proactive resource allocation and decision-making.
The project's main goal is to gain a thorough grasp of the variables that affect Walmart's sales, such as past sales information, marketing campaigns, macroeconomic indicators, and seasonal patterns. Create and put into use a predictive model that can accurately estimate sales while accounting for the intricate interactions between various factors. To guarantee the model's dependability and suitability for a variety of retail settings and product categories, assess and

adjust it. Give Walmart useful information and suggestions based on sales forecasts to help with strategic choices like inventory control and marketing strategy. In the end, this project will improve Walmart's capacity to manage the intricacies of the retail landscape and sustain a competitive edge in the market by providing it with an advanced sales forecast tool. The initiative aims to provide significant insights that can influence Walmart's and the retail industry's future sales forecasting by combining data science and retail knowledge.

Data Process

DataSet

The datasets utilized in this study were obtained from a prior Walmart-hosted Kaggle competition, accessible at:
https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/data
This historical sales information relates to 45 Walmart locations spread across various areas. Every store has several departments. The stores.csv file contains anonymized information about the 45 stores, indicating the type and size of the store. The files contain information from 2010-02-05 to 2012-11-01, including the store number, department number, date, weekly sales for the designated department at the given store, and whether the week is a special holiday. Determining whether weekly store sales are higher due to variations in the weather, fuel prices, holidays, markdowns, unemployment rates, and consumer price indexes is another important component of this research. The data for these factors is included in the file "features.csv," which is used in the analysis to examine their effects on sales performances. Figure 1 displays the column names and the data types of each dataset.

Information on the four major holidays—Christmas, Thanksgiving, Labor Day, and Super Bowl—is also included in the data. The list of holidays in the dataset is displayed in the table below.

Holidays	Date 1	Date 2	Date 3	Date 4
Super Bowl	12-Feb-10	11-Feb-11	10-Feb-12	8-Feb-13
Labor Day	10-Sep-10	9-Sep-11	7-Sep-12	6-Sep-13
Thanksgiving	26-Nov-10	25-Nov-11	23-Nov-12	29-Nov-13
Christmas	31-Dec-10	30-Dec-11	28-Dec-12	27-Dec-13

Figure 1

Stores Data

```
Store      int64
Type      object
Size       int64
dtype: object
```

Data Shape (45, 3)

Departments Data

```
Store             int64
Dept              int64
Date             object
Weekly_Sales    float64
IsHoliday          bool
dtype: object
```

Data Shape (421570, 5)

Sales data

```
Store             int64
Date             object
Temperature     float64
Fuel_Price      float64
MarkDown1       float64
MarkDown2       float64
MarkDown3       float64
MarkDown4       float64
MarkDown5       float64
CPI             float64
Unemployment    float64
IsHoliday          bool
dtype: object
```

Data Shape (8190, 12)

Data Preparation

To understand and analyze the data more efficiently it is important to merge all files data into one data frame. This would make it easier for us to comprehend the general data's trends and patterns. This also helps to extract meaningful insights by merging relevant data points. In this study datasets are merged using common columns and the information is viewed on the whole. The combined dataset has 16 columns and 421570 rows. Following that, duplicate and missing values in the combined dataset are examined. The missing values in the dataset are displayed in Figure 2. The dataset does contain zero duplicate values. Inspections have also been done on the datasets' unique values and datatypes which can be seen in Figure 3.

Figure 2

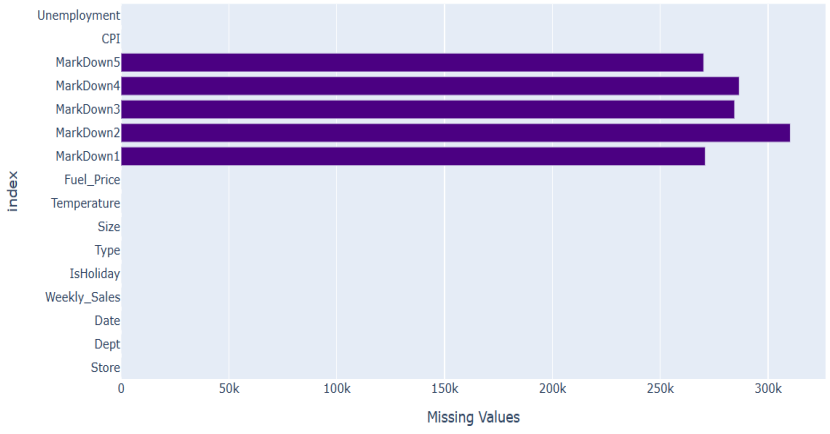


Figure 3

Unique values:

```
Store              45
Dept               81
Date              143
Weekly_Sales    359464
IsHoliday            2
Type                 3
Size                40
Temperature       3528
Fuel_Price         892
MarkDown1         2277
MarkDown2         1499
MarkDown3         1662
MarkDown4         1944
MarkDown5         2293
CPI               2145
Unemployment       349
dtype: int64
```

Data Merging

Using the merge function, the three data frames are combined into a single data frame for additional analysis. Initially, the department and store data are combined, and then the sales data is added to this combined data frame. The shape of the combined data frame consists of 421570 rows and 16 columns.

Explore Descriptive Statistics

One of the most important steps in comprehending a dataset's properties is investigating descriptive statistics. Important characteristics like distribution, variability, and central tendency are summed up by descriptive statistics. Figure 4 describes the statistics of the merged dataset.

Figure 4

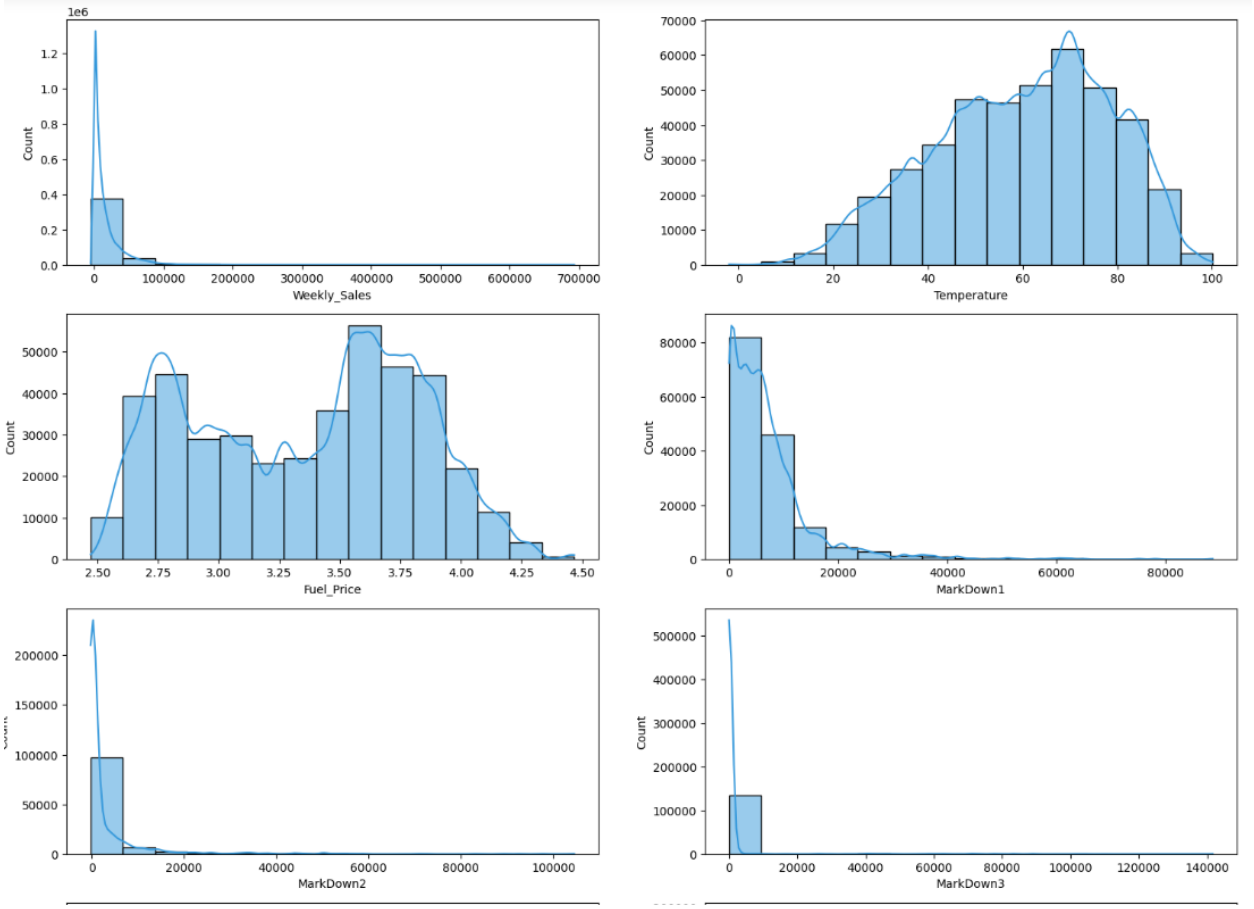
	Weekly_Sales	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	Unemployment	CPI
count	421570.000000	421570.000000	421570.000000	150681.000000	111248.000000	137091.000000	134967.000000	151432.000000	421570.000000	421570.000000
mean	15981.258123	60.090059	3.361027	7246.420196	3334.628621	1439.421384	3383.168256	4628.975079	7.960289	171.201947
std	22711.183519	18.447931	0.458515	8291.221345	9475.357325	9623.078290	6292.384031	5962.887455	1.863296	39.159276
min	-4988.940000	-2.060000	2.472000	0.270000	-265.760000	-29.100000	0.220000	135.160000	3.879000	126.064000
25%	2079.650000	46.680000	2.933000	2240.270000	41.600000	5.080000	504.220000	1878.440000	6.891000	132.022667
50%	7612.030000	62.090000	3.452000	5347.450000	192.000000	24.600000	1481.310000	3359.450000	7.866000	182.318780
75%	20205.852500	74.280000	3.738000	9210.900000	1926.940000	103.990000	3595.040000	5563.800000	8.572000	212.416993
max	693099.360000	100.140000	4.468000	88646.760000	104519.540000	141630.610000	67474.850000	108519.280000	14.313000	227.232807

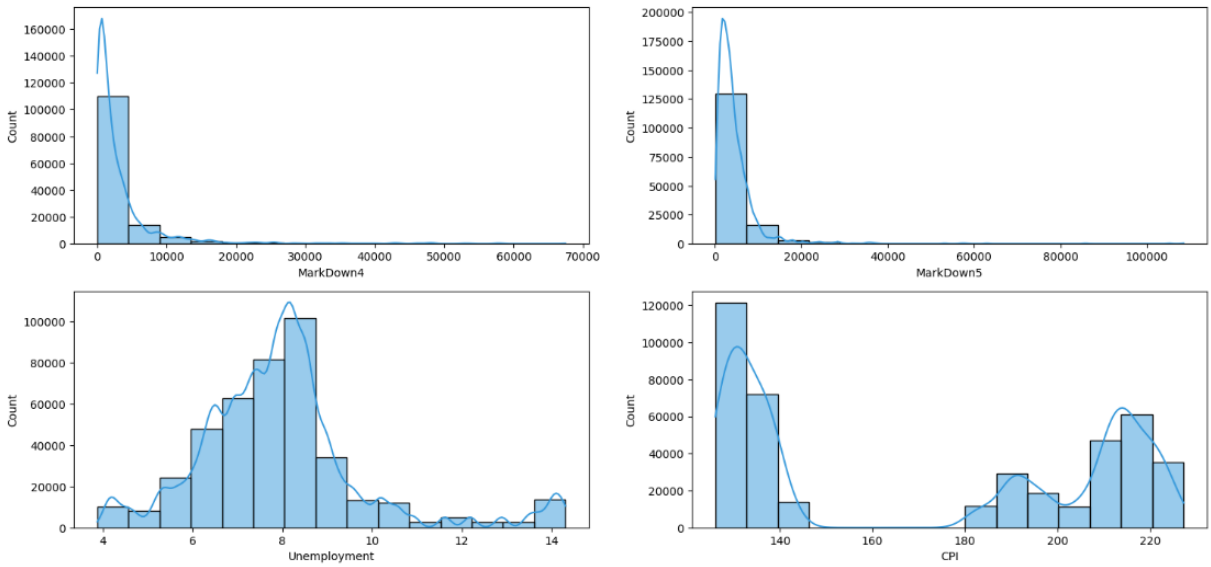
Exploratory Data Analysis

Checking the distribution of data plays an important role in analyzing the dataset correctly and performing the analysis accordingly. A histogram has been chosen to analyze the distribution of all numerical features, which can be seen in Figure 5. It can be observed that the distribution of Weekly_Sales is right-skewed, which is typical as weekly sales may exhibit higher values during certain periods. Temperature and Unemployment follow a normal distribution pattern. In contrast, CPI and Fuel_Price display a bimodal distribution, suggesting the presence of two distinct peaks in their respective datasets. Additionally, the distribution of each markdown column is unimodal.

Figure 5

Histogram for numerical columns





The count of the categorical columns is visualized by plotting a count plot on the categorical features. Figure 7, 8, and 9 shows that non-holiday days, which make up 93% of the total, far outweigh holiday days. This prevalence is seen as typical and is consistent with expectations. Interestingly, Store A is the most common store, indicating that it is well-liked or frequently appears in the dataset. Moreover, 2011 is the year that occurred the most frequently, suggesting that weekly sales measurements were concentrated at this time. Likewise, the most frequent months are April and July, indicating a concentration of sales activity during these particular months. Furthermore, the fact that Friday is mentioned again in every row emphasizes how frequently it occurs, suggesting that there is a pattern in the dataset connected to this specific day of the week.

Figure 7

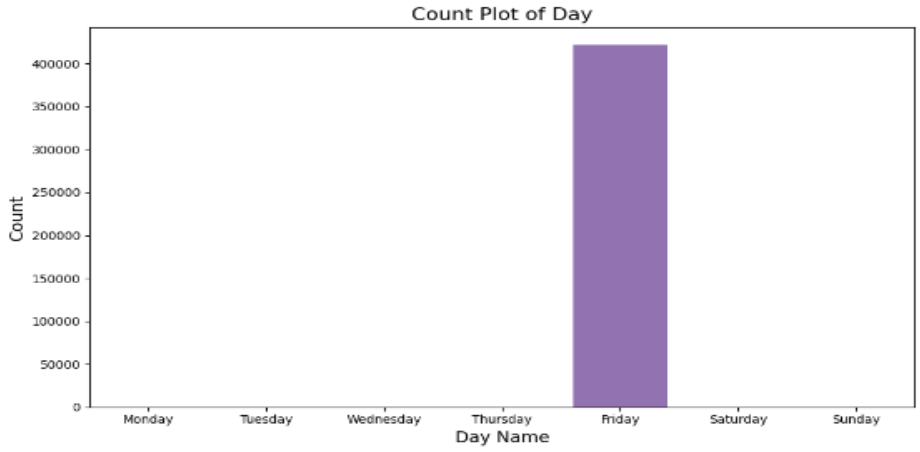


Figure 8

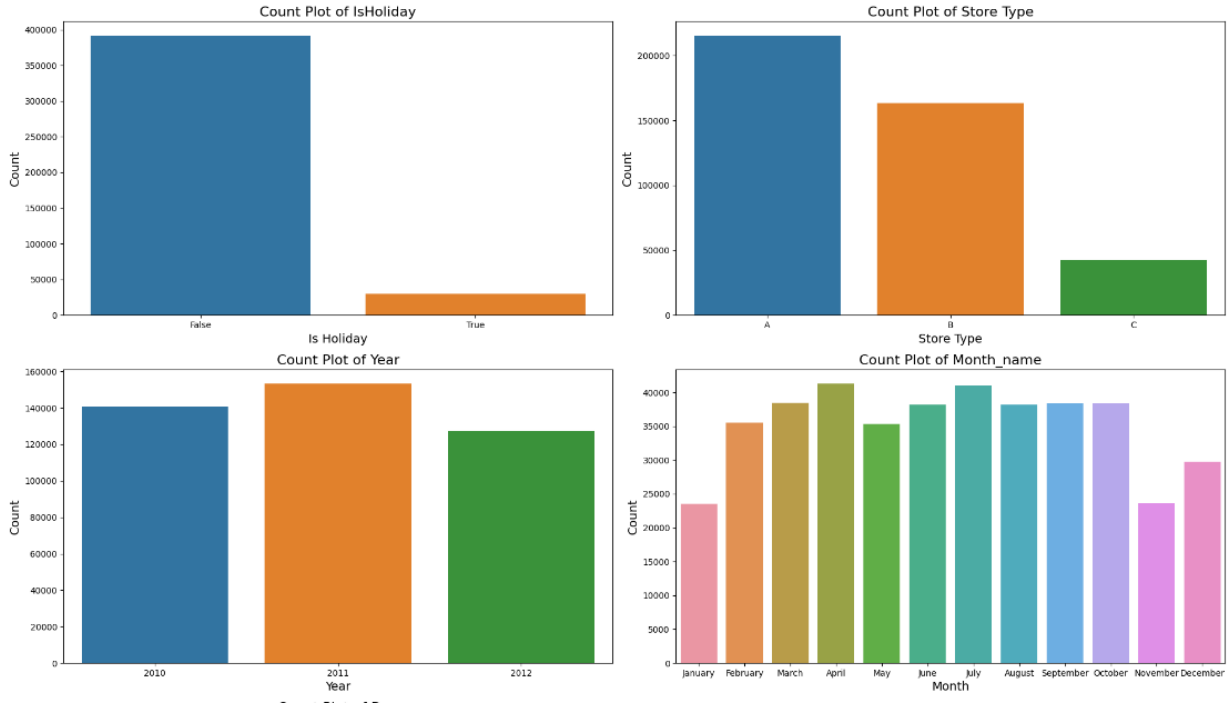
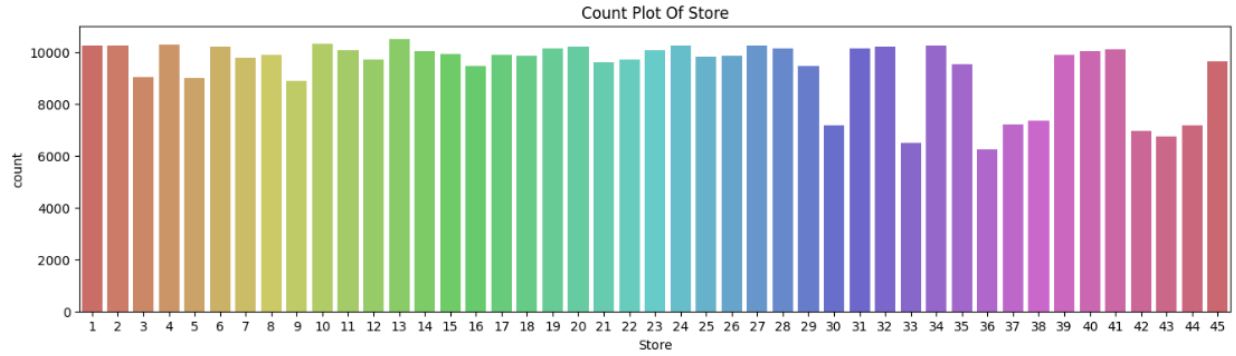


Figure 9



Analyzed sales concerning store type

Determining the typical sales for every Walmart location is also vital. Figure 10 makes it evident that, in comparison to the other two stores, the largest store, Type "A," has the most sales. Additionally, figure 11 shows a direct relationship between a store's physical size and sales numbers. Except for a small exception, the tendency generally shows that sales increase in tandem with shop size. This conclusion is consistent with the expectations of the Walmart

business, which states that larger stores typically have more varied product offers, higher foot traffic, and possibly higher sales volumes.

Figure 10

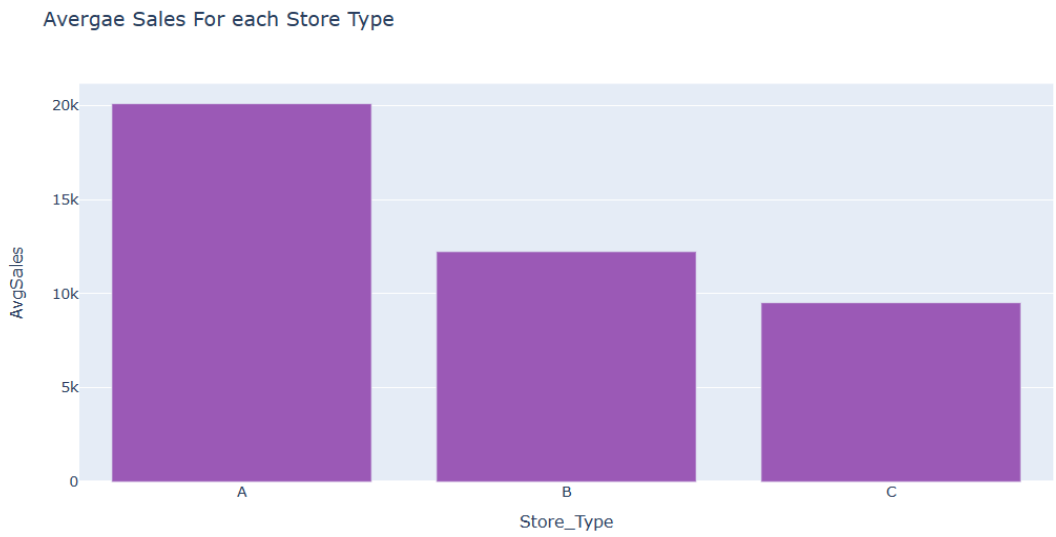
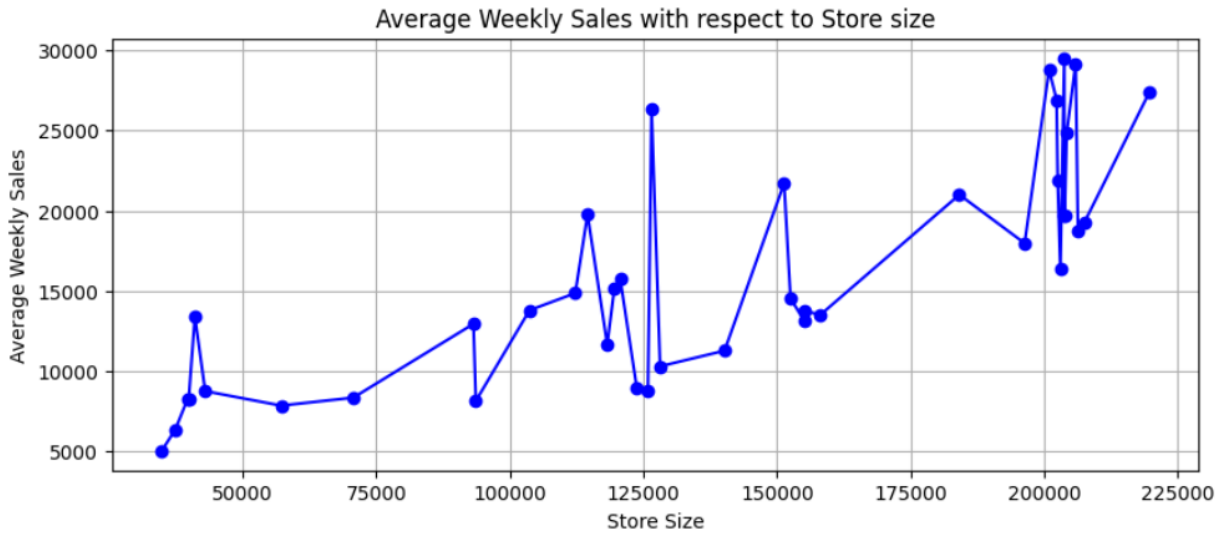


Figure 11

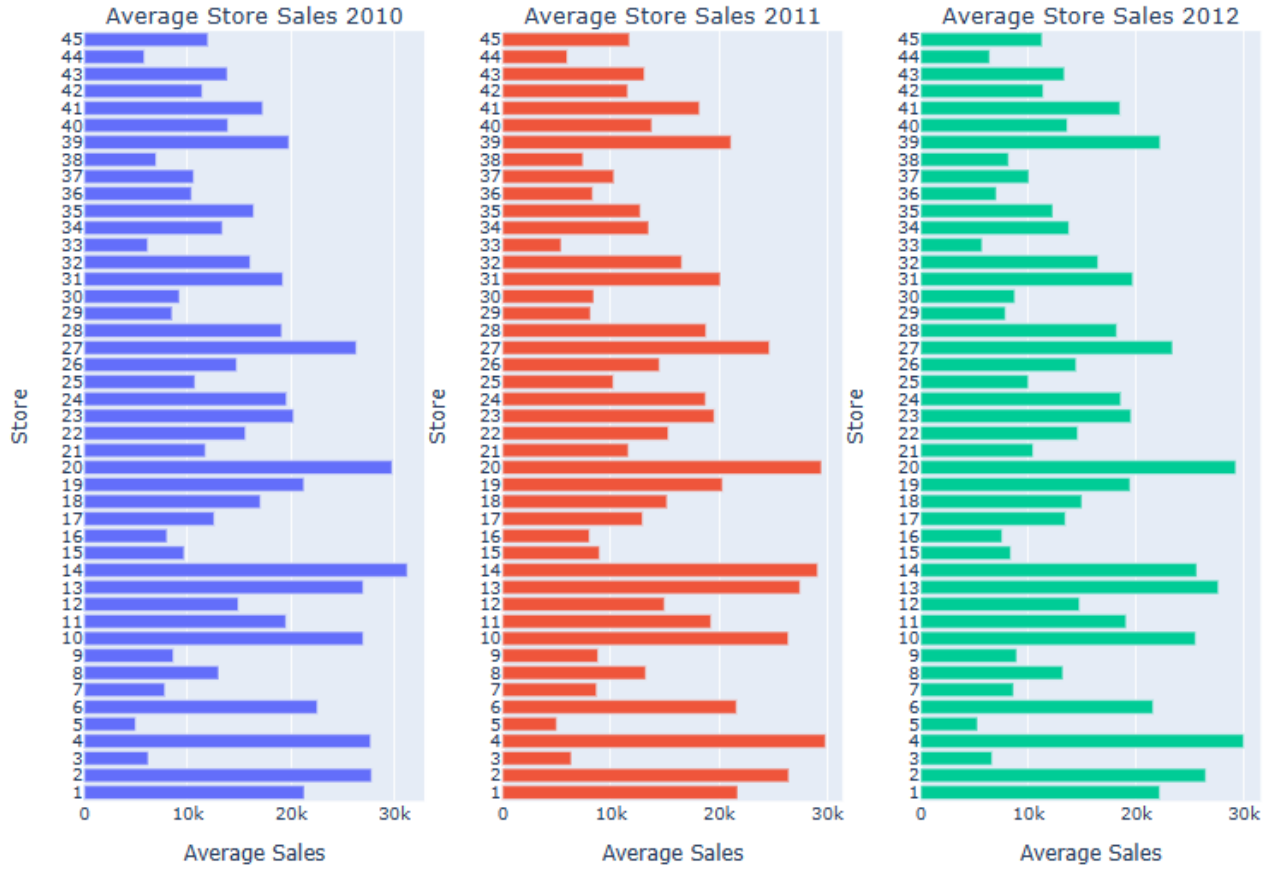


Analyzed sales concerning stores

Since store sales vary depending on the nature and size of the store, the overall trend over the last three years has remained the same. From Figure 12 it is evident that stores 2, 4, 13, 14, and 20 stood out as consistently high performers, demonstrating their continued performance over time. This pattern of recurrence implies that some establishments have strategic advantages

that contribute to their extraordinary sales, which offers insightful information for improving retail operations and tactics.

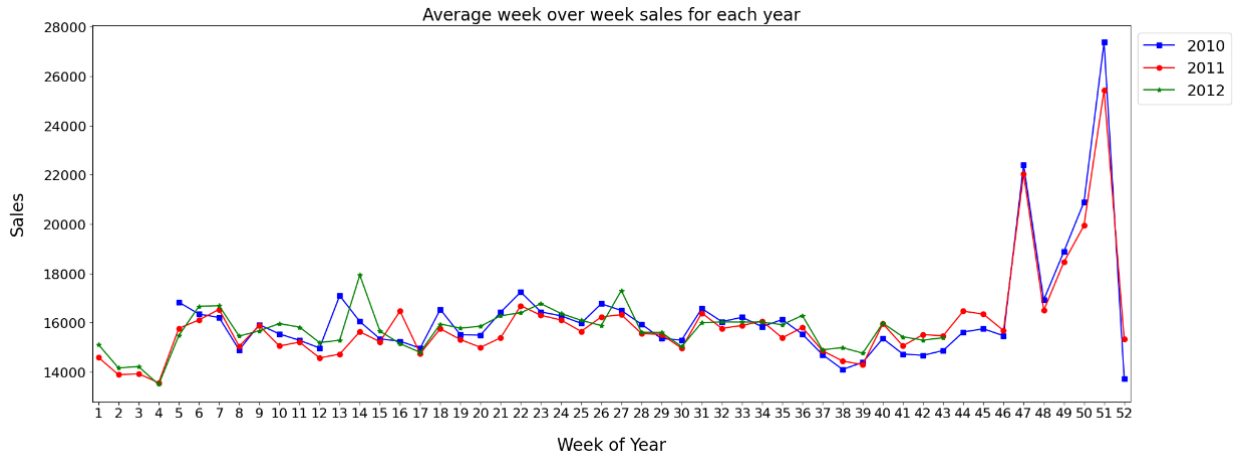
Figure 12



Average Weekly Sales per Year

From Figure 13, it is observed that for the years 2010 and 2011, the sales data analysis shows clear weekly sales increases during particular holiday seasons. Notably, the week before Thanksgiving and the week before Christmas typically had the greatest sales results, indicating that these festive seasons saw an increase in consumer activity. But in 2012, there was a notable exception: week number 14 had the highest sales even though it wasn't directly related to a holiday or other special occasion.

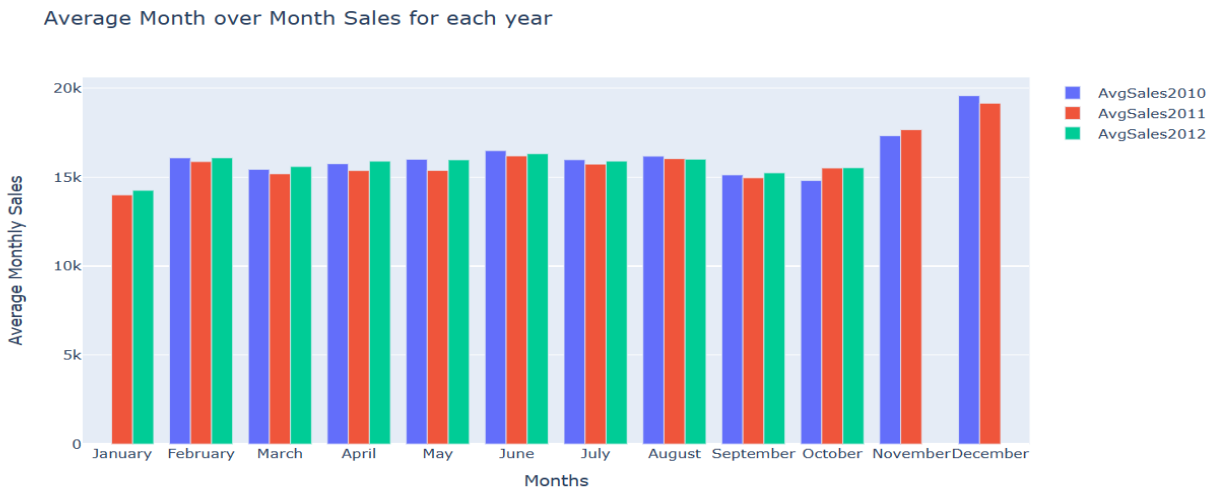
Figure 13



Average Monthly Sales per Year

According to the sales research conducted for all three years, Figure 14 shows that January consistently has the lowest sales in both 2011 and 2012. Since weekly sales data for January 2010 are not accessible, a thorough analysis of this particular month is not possible. Nonetheless, a comparatively steady pattern is noted from February to October, with weekly sales for each of the three years averaging 15,000. The holiday season brings about the most noticeable upturn in November and December, which also happen to be the months with the biggest sales in 2010 and 2011. However, a direct comparison is not possible due to the lack of sales data for December 2012, as the data is not provided.

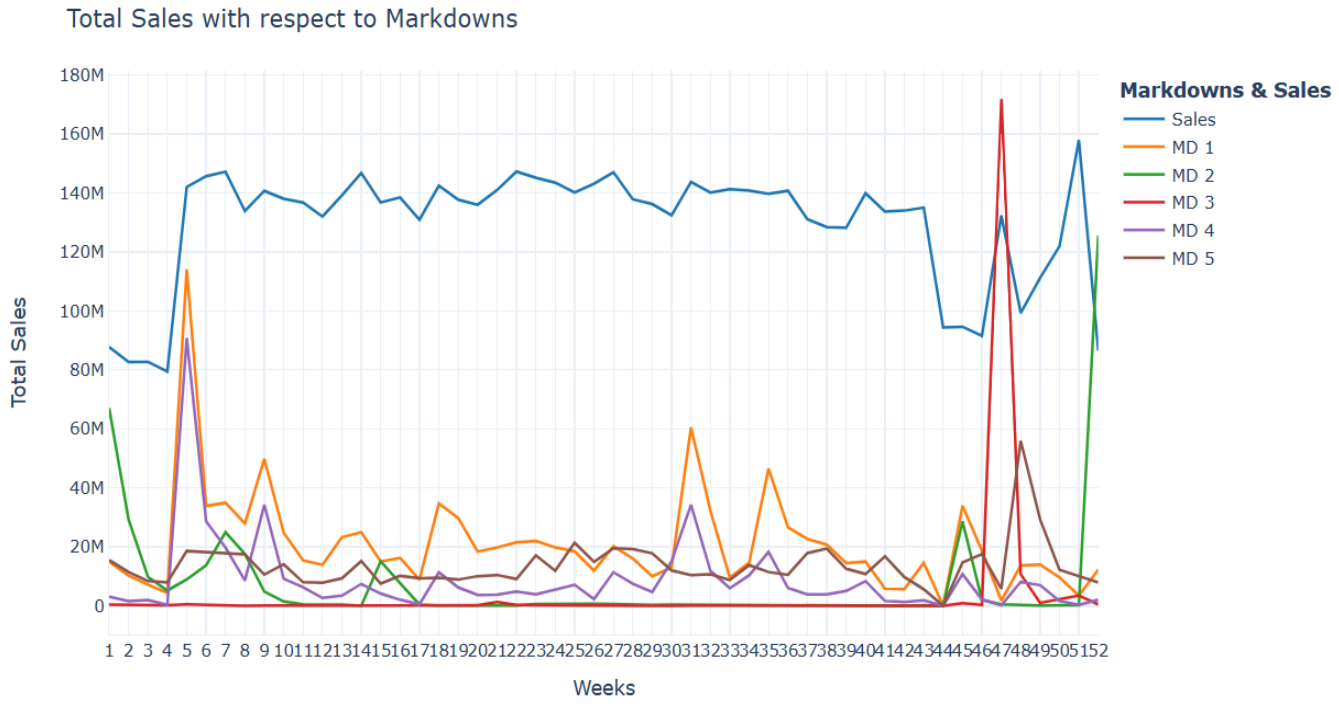
Figure 14



Analyzed sales for markdowns

Markdowns are seasonal promotional activities that are conducted in different stores and are not available all the time. Figure 15 demonstrates that they have a significant role in stabilizing sales levels and stimulating the recovery of sales, especially at the beginning and end of the year. This calculated use of markdowns highlights how important it is to time promotions to coincide with particular time windows to optimize their impact on consumer behaviour.

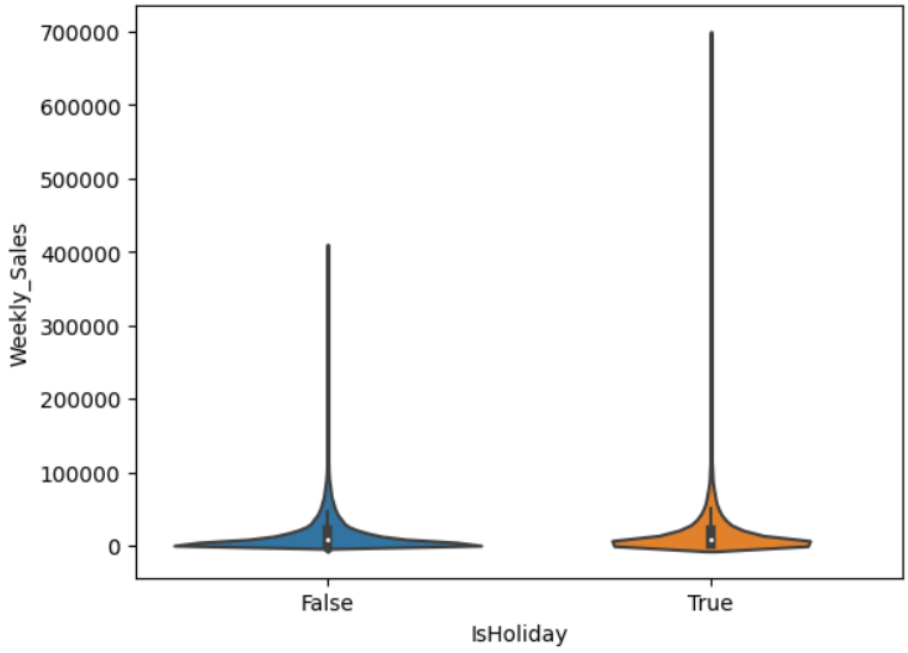
Figure 15



Identifying sales concerning Holidays

In the dataset, weekly Walmart sales statistics for different times of the year are supplied. This includes sales data for Thanksgiving, Christmas, and other holiday periods. To determine whether the holiday season brings in more sales, it was imperative to compare the sales growth between typical weeks and the holiday season. Holiday weeks are rare, but when they do happen, they have a big impact on sales. Even though holiday weeks make up only 7% of the dataset, they regularly have greater average sales than non-holiday weeks. The difference emphasizes how much holidays affect consumers' purchasing decisions. Figure 16 shows the violin plot for sales comparison during holidays and non-holidays.

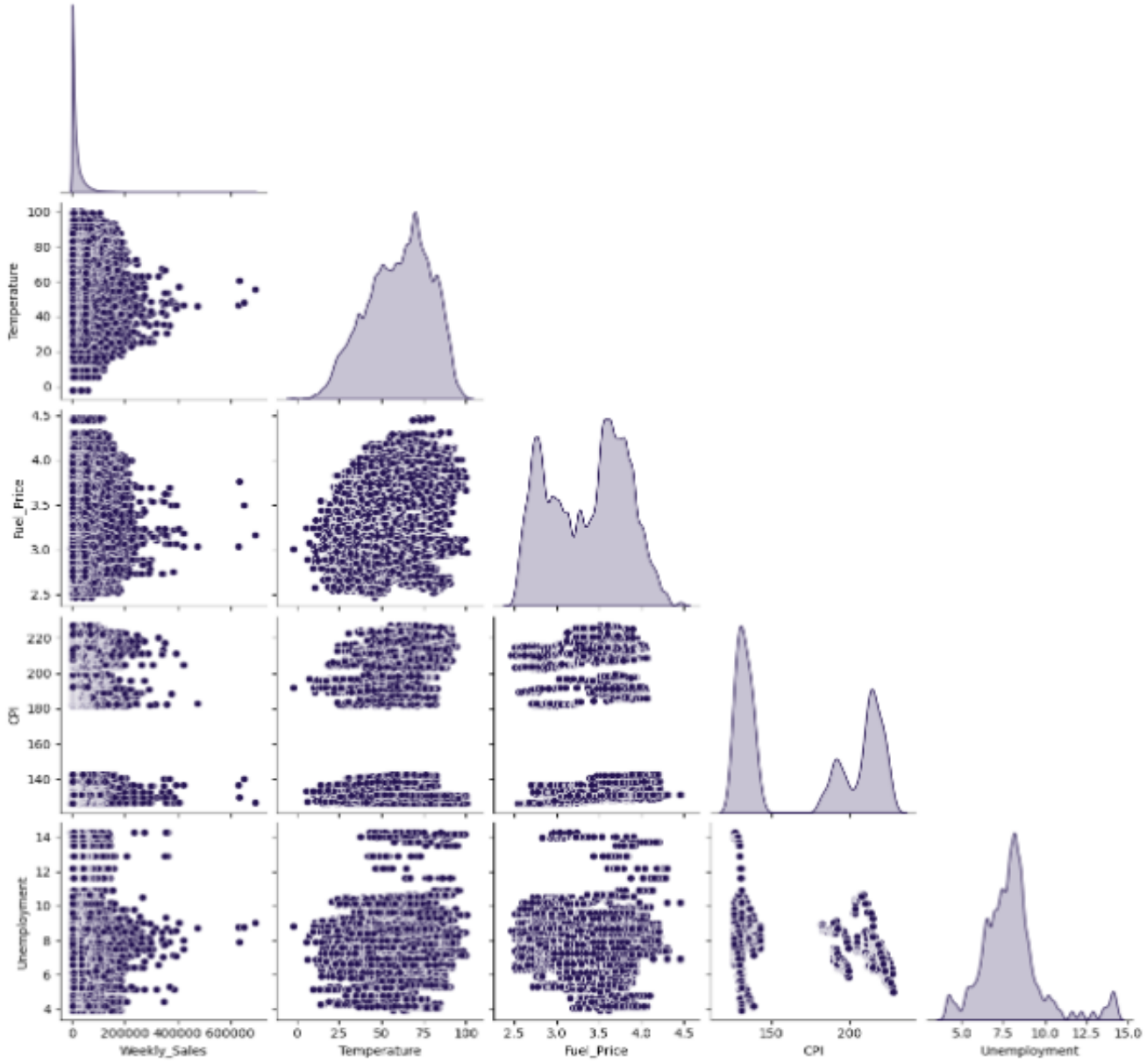
Figure 16



The Impact of Temperature, Fuel Price, CPI, Unemployment and Sales

Sales in the retail industry are known to be significantly impacted by the weather. Although warmer temperatures stimulate sales, excessively hot, cold, or both tend to discourage people from going outside to make purchases. Pleasant weather is thought to boost sales, as evidenced by the pair plot, which shows that most store types have their peak sales between 40 and 80 degrees Fahrenheit. When temperatures are extremely high or low, sales are often lower, but they appear to be sufficiently high when the weather is good. Sales appear to be higher when fuel prices are between 2.75 and 3.75 dollars, but they appear to decline when fuel prices are higher than $4.25. The idea that cheaper fuel prices lead to more sales is supported by certain findings, even though there isn't a clear trend to support this. CPI is a metric used to evaluate price fluctuations related to each person's cost of living. While there appears to be no discernible correlation between the change in CPI and weekly sales for Walmart shops, the pair plot shows three distinct clusters around various CPI ranges. An insignificant point to note is the substantial volume of sales when the CPI is at a mere 140. When the unemployment index rises above 11, there appears to be a noticeable decrease in sales. The store's highest sales history is observed between an 8 and 10 unemployment index. Figure 17 shows the pair plot for the same.

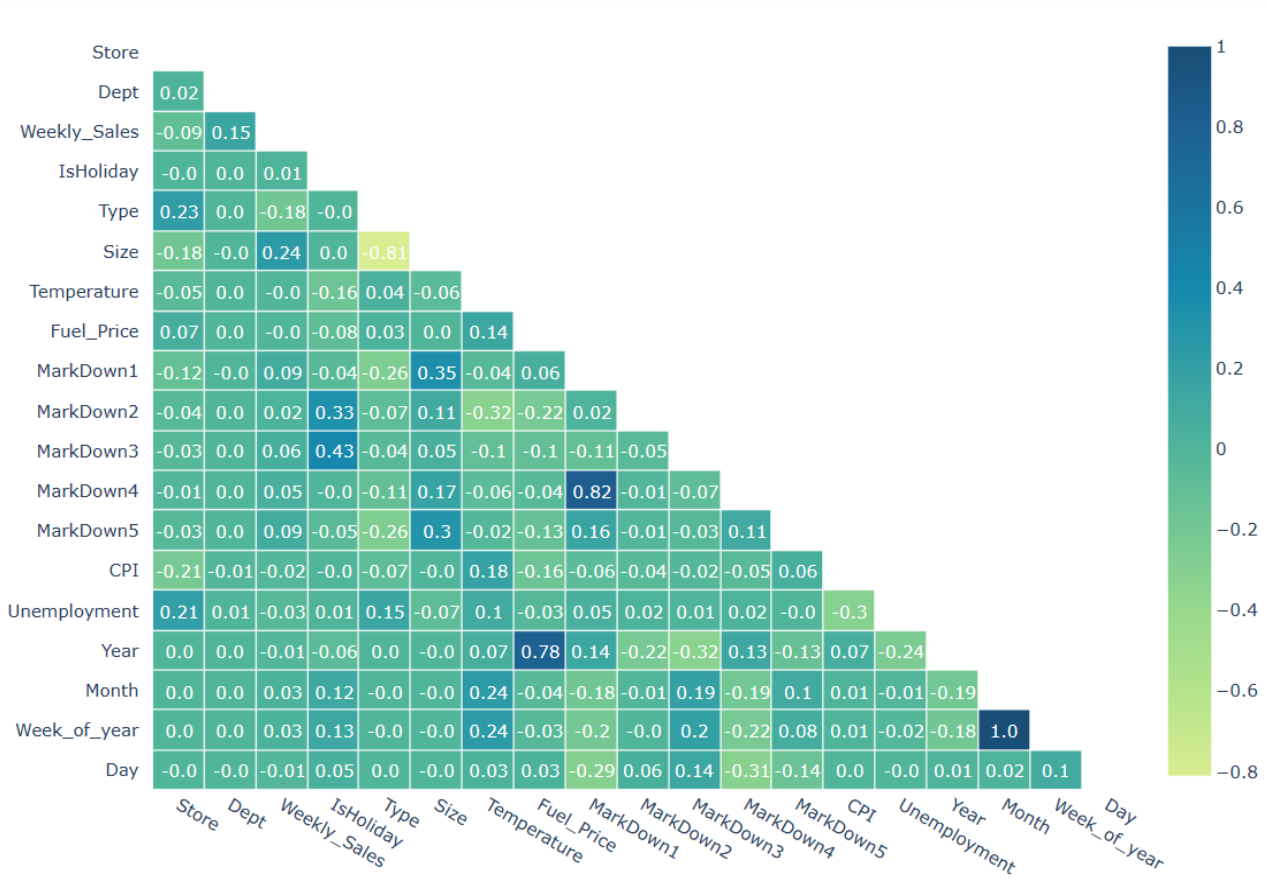
Figure 17



Correlation matrix

A correlation matrix describes the numerical correlation between the different columns in a dataset. A moderate correlation is observed between weekly sales and store size, type and department. There appears to be a negative link between weekly sales and temperature, unemployment, CPI, and fuel price. Markdowns 1–5 are not as significant a factor in the study because they do not appear to have a clear association with weekly sales. Figure 18 shows the correlation matrix.

Figure 18



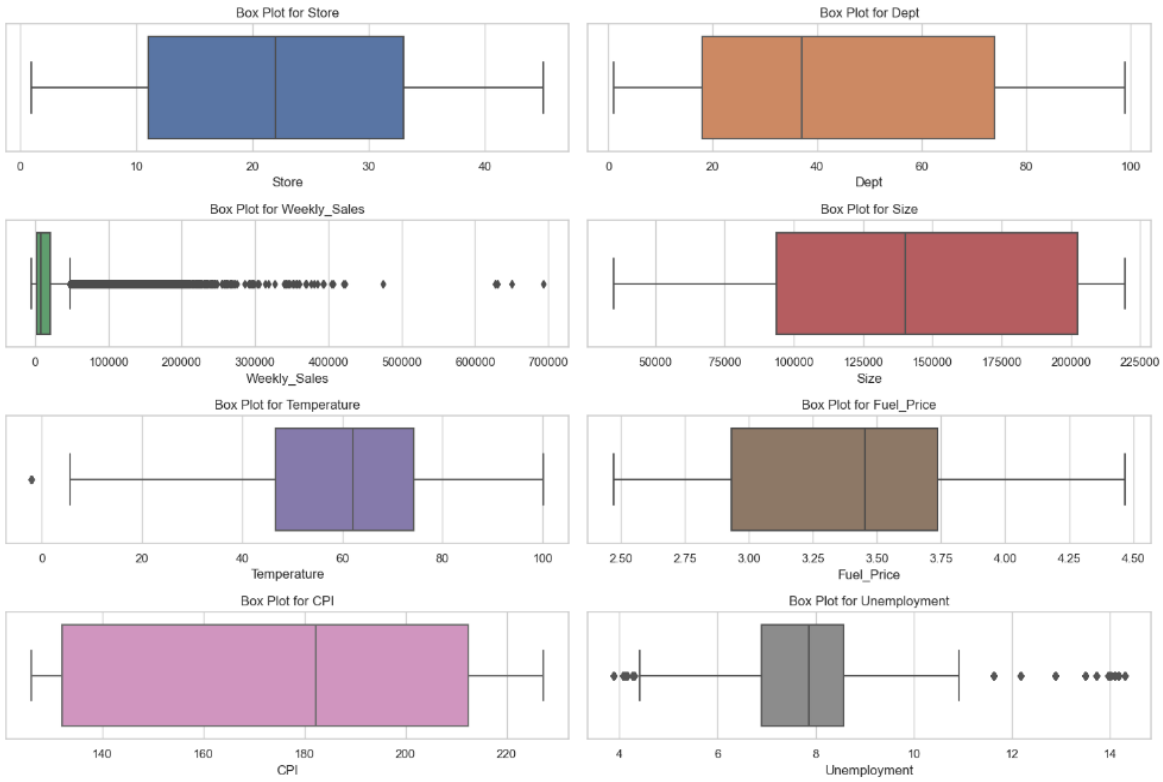
Feature Engineering

To analyze and gain a deeper grasp of the data, feature engineering is done before exploratory data analysis (EDA). This method is carried out to improve understanding of the data's underlying patterns. Six new columns—Year, Month, Month_name, Week_of_year, Day, and Day_Name—are extracted from the dataset "Date" column. In-depth investigation during the EDA phase is made possible by these extracted attributes, which also allow for identifying trends that were not immediately apparent in the original dataset. The dataset, which has 22 columns and 421,570 rows after feature engineering, provides the basis for thoroughly examining Walmart's sales patterns.

Data Preprocessing

Encoding "IsHoliday" as True as 1 and False as 0 is part of the categorical data transformation process. Likewise, the values in the "Type" column, A, B, and C, are encoded as 1, 2, and 3, correspondingly. To tackle the issue of over 65% missing values and an unclear association between markdowns and weekly sales as seen in the heatmap, the markdown columns are eliminated. The previously extracted "Year," "Month," "Month_name," "Week_of_Year," "Day_name," and "Day" columns have been removed, leaving only the "Week of Year" column, as the primary goal of this study is to accurately anticipate weekly sales for various Walmart stores. After a careful analysis of the data, anomalies were identified in the Weekly Sales, Temperature, and Unemployment columns. To guarantee the reliability of future analysis, outliers found in these columns are removed. Because outliers can have a substantial impact on how findings are viewed, treating outliers is essential to improving the accuracy and reliability of statistical measurements. Figure 19 displays the identification of outliers in the dataset's columns. After performing the standardization, the data split in the ratio of 80:20 which is train and test respectively.

Figure 19



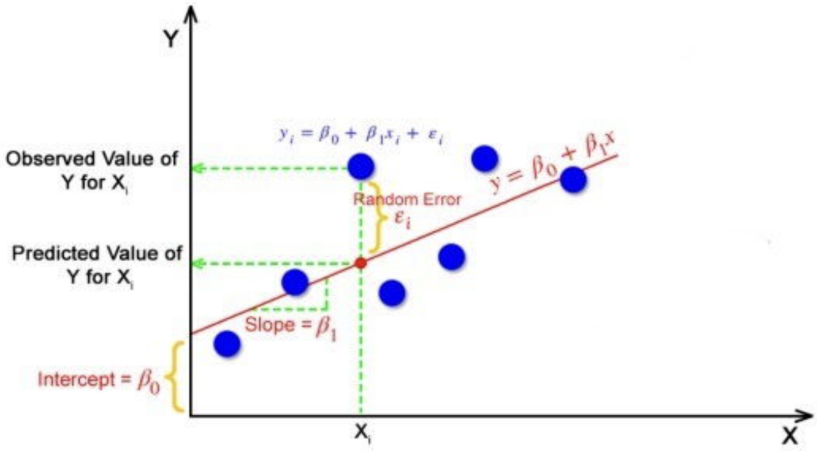
Modelling

Building algorithms that can recognize patterns in data and forecast or decide, without explicit programming. We have implemented and compared the results of four models to get predictions.

Linear Regression

A statistical technique for simulating the connection between a dependent variable and one or more independent variables is called linear regression. Finding the best-fitting linear connection to explain the data is the aim of linear regression. Whereas multiple linear regression involves two or more independent variables, basic linear regression just involves one. A linear relationship can be seen in Figure 20.

Figure 20



$$Y = mx + b \quad (1)$$

Where Y is the dependent variable, x is the independent variable m is the slope of the line, and b is the intercept of y which is the value of y when x = 0.

$$Y = b_0+b_1x_1+b_2x_2+ \ldots\ldots\ldots b_nx_n \quad (2)$$

eq(2) illustrates the multiple linear regression.

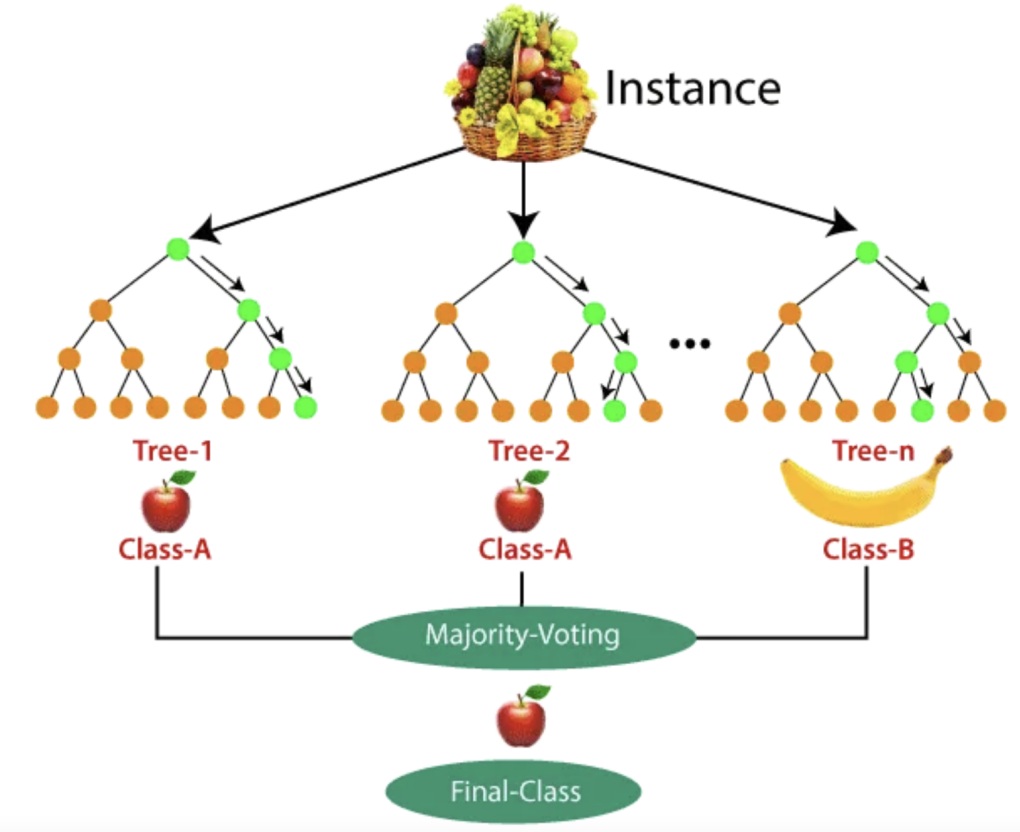
Where Y is the dependent variable and b_0, b_1, b_2…, b_n is the coefficient representing the change in Y. Predicting sales, stock prices, home values, and many other applications where there is a linear relationship between the variables are common uses for linear regression.

Random Forest

A machine learning model known as a "random forest" makes its final prediction by averaging the predictions of many decision trees that were trained on a dataset with a real tree structure, which can be seen in Figure 21. A supervised machine-learning approach for

classification and regression problems is called a decision tree. The data is recursively divided into subsets according to the most important feature at each stage, creating a structure similar to a tree, with the ultimate choice or output represented by the leaves. Decision trees are especially helpful since they are simple to visualize and comprehend.

Figure 21



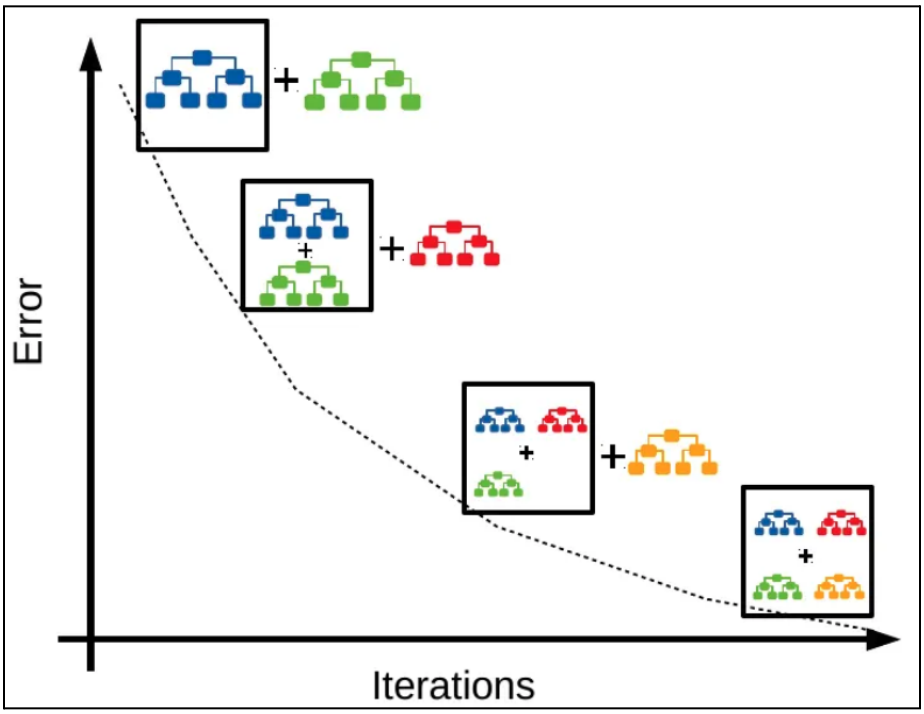
Two important strategies are used during Random Forest construction to increase the ensemble's resilience and variety. First, multiple decision trees are constructed using Bootstrap Sampling, which involves continually selecting random samples with replacements from the training set. By guaranteeing that every tree is trained on a different sample of the data, a method called bagging adds variety to the individual models. Second, by only taking into account a random subset of characteristics at each split in each decision tree, feature randomization introduces another layer of unpredictability. This deliberate tree decorrelation reduces the likelihood of overfitting and increases the ensemble's overall resilience. Finally, the Random Forest uses voting in classification tasks or averaging in regression tasks to blend the outputs of its

constituent trees when generating predictions. By using the strengths of each individual tree, this collective technique produces an accurate and strong ensemble model.

Gradient Boosting Machine

One kind of ensemble approach that combines several weak models to improve overall performance is called gradient boosting. For tabular datasets, one of the most often used machine learning techniques is gradient boosting. With its exceptional usability, it can handle missing values, outliers, and large cardinality categorical values on your features without the need for further handling. It is strong enough to identify any nonlinear connection between your model goal and features. The approach creates an ensemble of weak learners, usually decision trees, during the Gradient Boosting model-building process to iteratively increase prediction accuracy, which is illustrated in Figure 22.

Figure 22



It starts with a basic baseline model, such as a single decision tree. The first tree's predictions are integrated with the first model's once it has been trained on the data. Subsequently, the algorithm concentrates on error correction by computing the discrepancy between the combined forecasts and the actual target values. Then, a fresh tree is trained to anticipate these mistakes. This sequence of additions keeps on, with each new tree trying to fix the mistakes the previous ensemble committed. The optimization technique used in this iterative procedure is where the

phrase "gradient boosting" originates. More specifically, the technique progressively minimizes the total error by using gradient descent to find the best path for updating the model at each iteration. The learning rate is a crucial element in this process that regulates each tree's contribution to the ensemble. The learning rate affects the model's resilience and convergence by scaling each new tree's predictions before adding them. Although resilience is improved by a lower learning rate, more trees are needed for the best results. An algorithmic halting criterion is satisfied when a predetermined number of trees are constructed or a predetermined level of performance is reached. This methodical process produces a strong ensemble model that can recognize complex patterns in the data and make precise predictions.

XGBoost Algorithm

One very effective and adaptable machine learning version of the gradient boosting approach is called eXtreme Gradient Boosting, or XGBoost. This method, created by Tianqi Chen, is well-liked for a range of applications, including machine learning contests, due to its exceptional speed and performance. Fundamentally, XGBoost uses decision trees as base learners, adding these trees to the ensemble one after the other to fix mistakes committed by the previous model. To manage model complexity and avoid overfitting, the method optimizes an objective function that combines regularization terms with a loss function that measures the prediction error. L1 and L2 regularization are features of XGBoost that enable fine-tuning of model complexity. It helps with feature selection and analyzing variable contributions by providing feature significance ratings. The algorithm's efficiency and scalability are enhanced by tree pruning, parallel and distributed computing capabilities, and integrated support for managing missing information. Additionally, XGBoost makes early training stoppage and cross-validation easier, which increases the model's resilience. XGBoost's adjustable hyperparameters enable practitioners to fine-tune the algorithm for their unique datasets and attain peak predicting performance. With support for several programming languages and smooth integration into well-known machine learning frameworks, XGBoost is easily available and extensively utilized within the data science community.

Evaluation

A critical phase in the model creation process is evaluating machine learning models to determine their effectiveness and capacity for generalization. Depending on the kind of problem,

several metrics and approaches (such as regression, classification, etc.) might be used. The following are some typical techniques for evaluating models that have been used:

MSE

Regression model performance is commonly assessed using the Mean Squared Error (MSE) measure. The average squared difference between the actual and anticipated values is what's measured. The following is the formula for mean squared error:

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Where

n is the number of observations in the dataset

Y_i is the actual target value for the i_{th} observation

$\widehat{Y}_i$ is the predicted target value for the i_{th} observation

Better model performance in regression tasks is shown by lower Mean Squared Error (MSE) values. The MSE is a metric that indicates how well the model's predictions match the actual data. It is computed as the average squared differences between the anticipated and actual values. A lower MSE indicates better accuracy and precision since there is, on average, less variation between the model's predictions and the actual data points.

It is important to acknowledge the susceptibility of Mean Squared Error (MSE) to outliers. Since squaring the differences is a requirement of MSE.

MAE

A metric that is often employed to evaluate the effectiveness of regression models is Mean Absolute Error (MAE). The average absolute difference between the actual and anticipated values is what's measured. The following is the formula for mean absolute error.

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

Where

n is the number of observations in the dataset

Y_i is the actual target value for the i_{th} observation

x_i is the predicted target value for the i_{th} observation

A lower number in the MAE implies superior model performance, meaning that the model's predictions are generally closer to the actual values. When outliers are an issue or when a clearer interpretation of the average error is required, MAE is very helpful.

R2 Score

The coefficient of determination, or R-squared (R2) score, is a statistic used to assess how well a regression model fits its data. It measures the percentage of the dependent variable's volatility that can be predicted based on the independent factors. There are 0 to 1 possible R2 scores.

R2 = 0 shows that all of the target variable's variability cannot be explained by the model.

R2 = 1 shows that all of the target variable's variability can be explained by the model.

The following is the formula for the R^2 score

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}}$$

Where:

The Sum of Squared Residuals (SSR) is the sum of the squared differences between the actual and predicted values. Total Sum of Squares (SST) is the sum of the squared differences between the actual values and the mean of the actual values.

Performance Comparison

We trained the models with default parameters and the results can be seen in Table 1 where we can observe that the best-performing model is XGBoost with a .97 R^2 score and Linear Regression is the least-performing model with a .15 R^2 score. The Random Forest model also performed well with the .95 R^2 score.

Table 1

Model	MAE	MSE	R2 Score
Linear Regression	421599.43	257002398384.23	.154
Random Forest	65999.91	14080985096.68	.953

Gradient Boosting Machine	123013.92	27562364717.25	.909
XGBoost Algorithm	**51626.36**	**7189589388.47**	**.976**

Hyperparameters are external machine learning model configuration settings that need to be established before training; they are not learnt from the data. Optimizing the performance of a model requires fine-tuning these hyperparameters. We are using Grid Search, which evaluates model performance for every combination while doing an exhaustive search throughout a given hyperparameter space. Achieving the best possible model performance requires effective hyperparameter adjustment. Striking a balance between taking advantage of promising areas and examining a wide range of hyperparameter values is necessary.

Table 2 illustrates the performance of the model after tuning the hyperparameter with the Gridsearch algorithm and providing the best hyperparameter value. The following table illustrates that XGBoost is still the best-performing model with a .976 R2 score, but Gradient boost improved performance and provided almost equal performance to XGBoost

Table 2

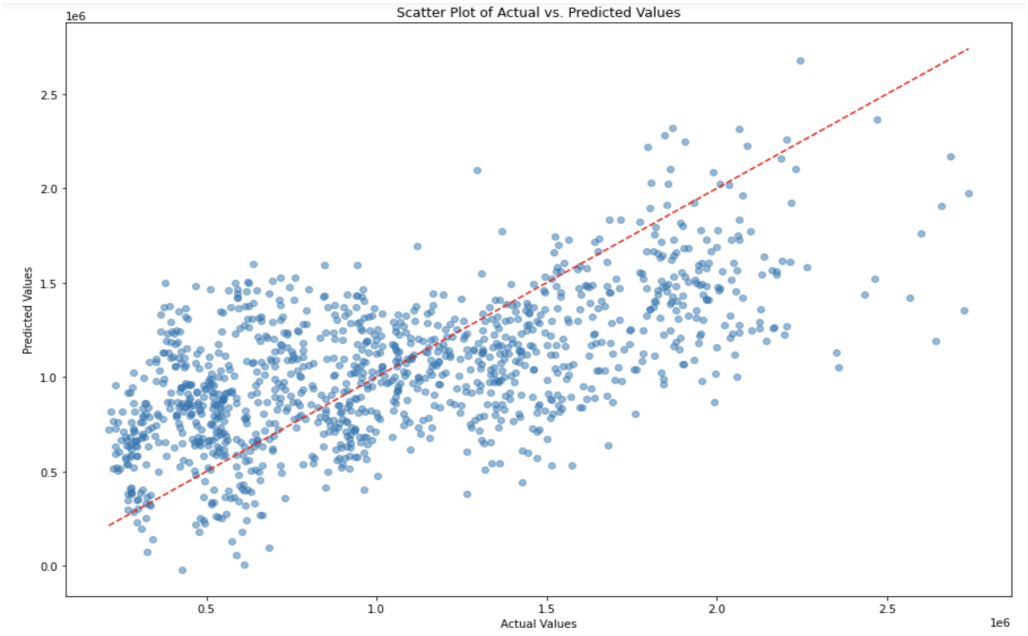
Model	Hyperparameter	Value Used	Best Hyperparameter value	R2 score
Linear Regression	Polynomialfeatures__ degree	2,3,4	3	.388
Random Forest	n_jobs	-1	-1	.954
	max_depth	np.arange(2,15)	14	
	n_estimators	np.arange(25,101, 25)	50	

Gradient Boost	n_jobs	-1	-1	.974
	max_depth	np.arange(2,15)	10	
	n_estimators	np.arange(25,101, 25)	60	
XG Boost	n_jobs	-1	-1	**.976**
	max_depth	np.arange(2,15)	5	
	n_estimators	np.arange(25,101, 25)	180	

A scatter plot comparing the actual value with the predicted value for given models:

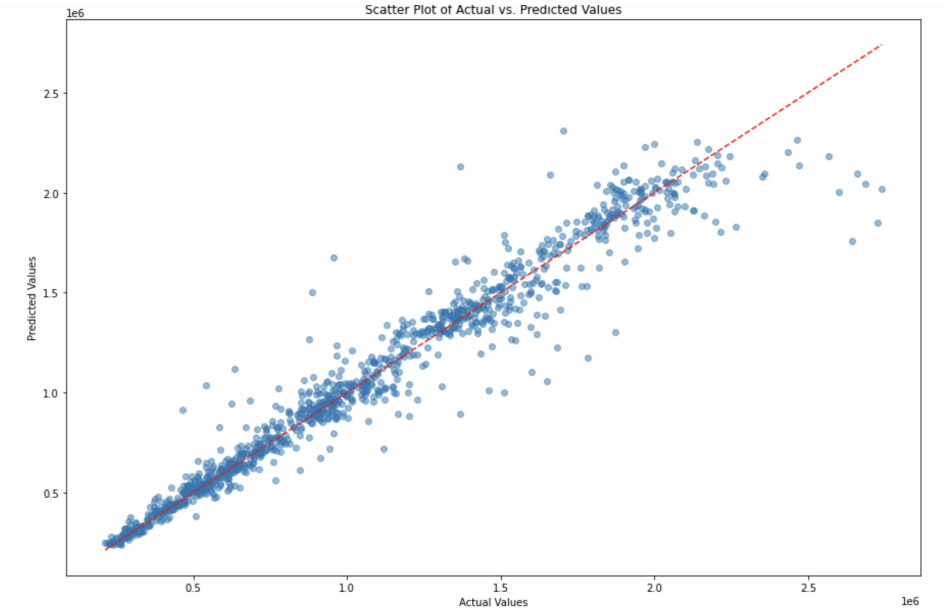
Linear Regression. Figure 23 illustrates the comparison between actual and predicted values where it can be seen that values are not following the regression line and predicted values are not similar to actual values. That defines the low R^2 score.

Figure 23



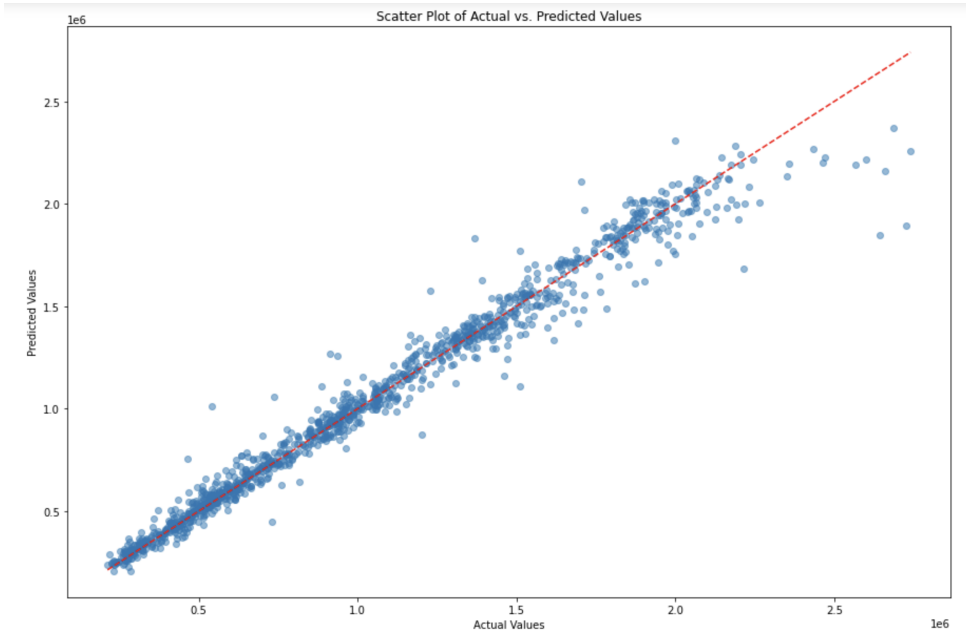
Random Forest. Figure 24 defines the comparison between actual and predicted values where it can be seen that most of the values are following the regression line and predicted values are quite similar to actual values. That defines the high R^2 score for Random Forest.

Figure 24



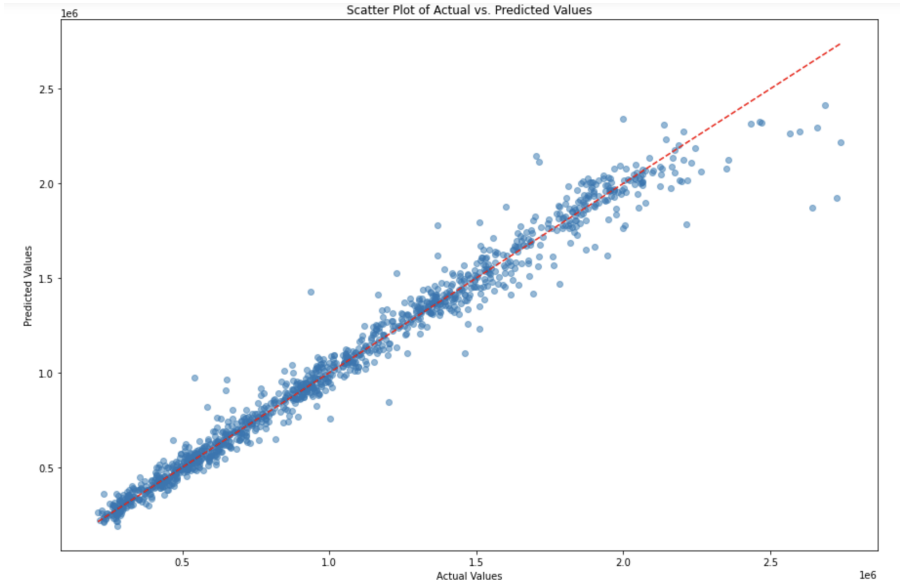
Boosting Algorithm. Figure 25 illustrates the comparison between actual and predicted values where most of the values follow the regression line and predicted values are quite similar to actual values. That defines the high R^2 score for Boosting Algorithm.

Figure 25



XGBoost. Figure 26 shows the comparison between the actual and predicted values and most of the values follow the regression line. That defines the highest R^2 score for the XGBoost Algorithm.

Figure 26



Discussion

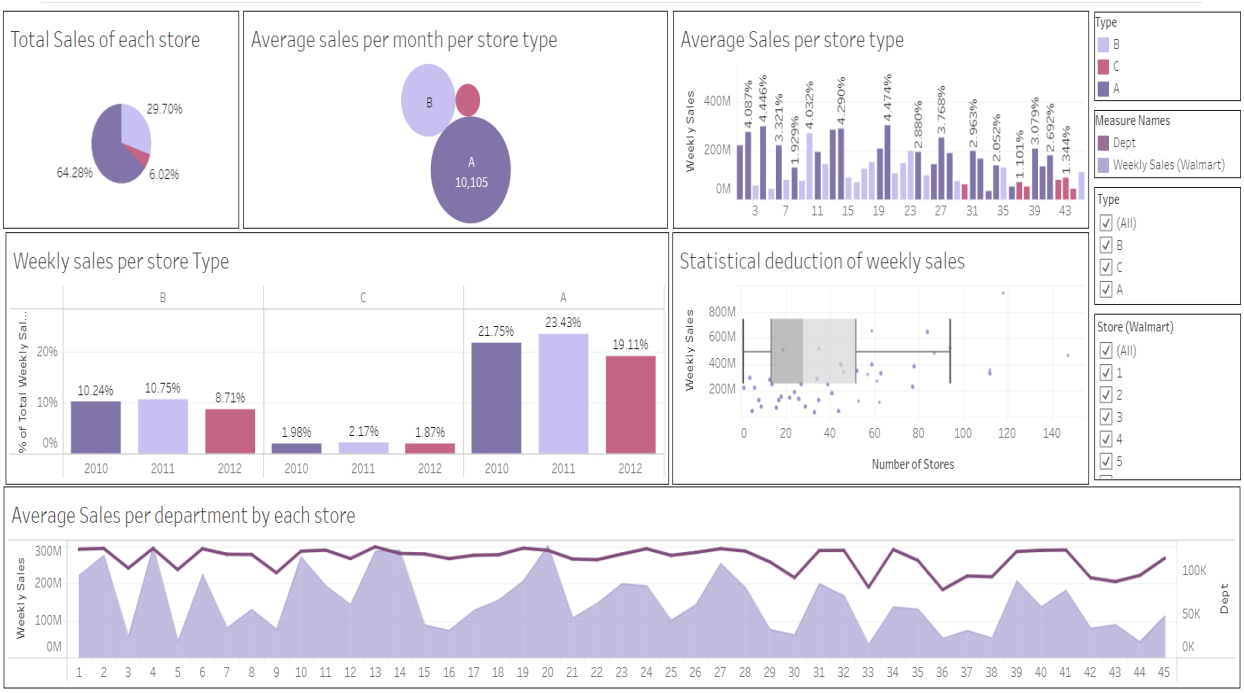
A statistical technique called linear regression is used to identify and measure the connection between one or more independent variables and a dependent variable. By fitting a linear equation to the observed data, this approach predicts the characteristics of the dependent variable from the independent variables. It's crucial to remember, though, that linear regression has its limits. It may be susceptible to data outliers, which means that extreme values may have an outsized impact on the model's predictions. Furthermore, if regularization isn't done correctly, linear regression can lead to overfitting of the training set, which makes it difficult for the model to generalize to new, untested data even while it performs well on the training set. Regularization techniques are employed to address this issue and enhance the model's ability to make accurate predictions on new data.

Several decision trees are used in the ensemble approach known as random forest to improve forecast accuracy. Interestingly, random forests are good at managing complicated nonlinear interactions in the data and are resilient to outliers. Because of their adaptability, they may be used for a wide range of predictive modelling tasks, successfully reducing the influence of extreme values and producing accurate forecasts across a variety of datasets.

Gradient Boosting Machine is an ensemble technique that builds decision trees one after the other, fixing mistakes from the previous tree. A very accurate model is produced as an outcome of this successive learning process. GBM differs from random forest in that it manages complex nonlinear interactions more skillfully and has less tendency to overfit. An improved version of GBM called XGBoost uses sophisticated regularization and tree-splitting methods. Thanks to its great accuracy and efficiency, XGBoost has gained popularity in a variety of machine-learning applications. Hence, it outperformed in the prediction of sales with the Walmart dataset.

Tableau Dashboard

Because of Tableau dashboards' remarkable data display and analytical capabilities, they are frequently used. These dashboards, which are highly regarded for their interactive features, enable users to analyze data, apply filters, and obtain a more profound understanding. The platform's wide range of visualization choices, which include maps and charts, make it easier to depict complicated facts in a clear and visually appealing way. Figure 27 shows a glimpse of the dashboard we created. To view the interactive dashboard, click here.



Conclusion

This study assessed many machine learning techniques to improve the prediction accuracy of sales data for 45 Walmart locations. After comparing the methods for linear

regression, random forest, gradient boosting machines, and XGBoost XGBoost and Gradient Boost turned out to be the best options, obtaining the lowest error rates and the greatest R2 score of .97. Especially, due to intrinsic regularization, ensemble approaches prevented overfitting and efficiently handled nonlinear sales trends. While regular retraining of models with fresh sales data would assure maintained accuracy, more investigation into feature relevance and their consequences might yield deeper business insights. These estimates might reach their full economic usefulness if they are integrated into supply chain and inventory management systems.