STAT-670: Exploratory Data Analysis

Team:
Komaragiri Usha Bhanu
Henish Shah

WALMART SALES ANALYSIS

Abstract:

In the twenty-first century, technological advancement is reaching new heights, with massive amounts of data to be processed and researched in order to derive new insights from the data. Retailers now require a 360-degree perspective of their customers and their purchasing patterns in order to maintain their competitive edge in the market. Based on these insights, they launch various promotions and offers to boost sales and accomplish marketing goals. In this project, we explore a dataset of one of the world's largest retailers Walmart Inc., to establish a relationship between customer purchasing patterns based on various parameters and the overall sales earned.

Data:

For the purpose of this project, the dataset of the Walmart Store sales from Feb 2010 to Nov 2012 was collected from Kaggle. The dataset comprises of the weekly sales data collected from 45 different Walmart stores within the country. It also consists of some additional features that might be essential in charting out the customer purchasing patterns. Table 1 describes the various features present within our dataset.

Feature	Description
Store	Store number (1-45)
Date	Week of the sales
Weekly Sales	Sales for the given store
Holiday Flag	Holiday week -1 Non-holiday week - 0
Temperature	Average Temperature during the week
Fuel Price	Cost of fuel in the region
CPI	Prevailing consumer price index
Unemployment	Prevailing unemployment rate

Table 1: Feature Description

Thus, our dataset comprises of a dependent variable (Weekly Sales) and other seven independent variables (Store, Date, Holiday Flag, Temperature, Fuel Price, CPI, Unemployment).

In order to check the extent of the dependance of our target variable 'Weekly Sales' on the independent continuous variables in our dataset, we plot a '*Correlation plot*'.

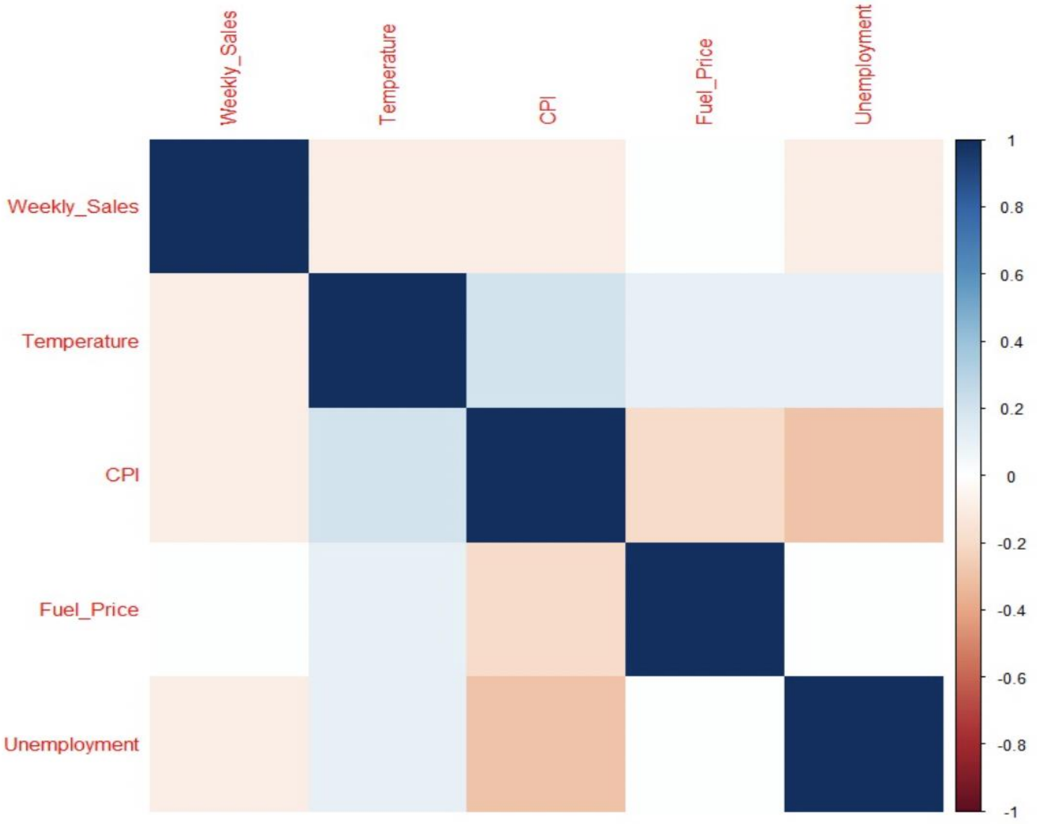


Figure 1: Correlation plot

The values 1 (dark blue shade), -1(red shade) and 0(white shade) describe highest positive, negative and no correlation respectively. Upon observing the first row/first column of the Correlation plot, we notice that all our continuous independent variables are only slightly negatively correlated if not correlated at all to the target variable ‘Weekly Sales’. Hence, this signifies the absence of a direct linear relationship between these variables to that of ‘Weekly Sales’.

We add new features like month and occasion as part of our feature engineering process.

Feature	Description
month	Month of the sale
Occasion	Type of Holiday

Table 2: Feature description

(*Note*: The value of the ‘Occasion’ feature is marked as "Not a holiday" if there is no holiday event)

In our study, we specifically look at four holidays.

- Super Bowl (February)
- Labour Day (September)
- Thanksgiving (November)
- Christmas (December)

Research Questions:

Based on the preliminary analysis of our data, we try to answer the following questions as part of a business problem statement:

- Which holiday week generates the most or least sales?
- Is there a positive impact on overall sales during the holidays?
- How does the weekly sales pattern vary by stores around different holidays?
- What type of relationship do the other continuous variables have with weekly sales?
- Which are the most important features that drive sales?

Significance:

The study findings can be informative in the following ways:

- Consumer demands are properly represented.
- Inventory Management Enhancement.
- Identifying peak and off-peak periods to determine how much staff to hire and optimize marketing tactics.

Analysis:

In light of our previous observation from the correlation plot, we start our analysis surrounding the "Date" feature which may highlight the seasonal fluctuations in sales. In order to gain a holistic picture of the trends, we consider looking at the monthly data of Average Weekly sales across all the years [Fig 2].

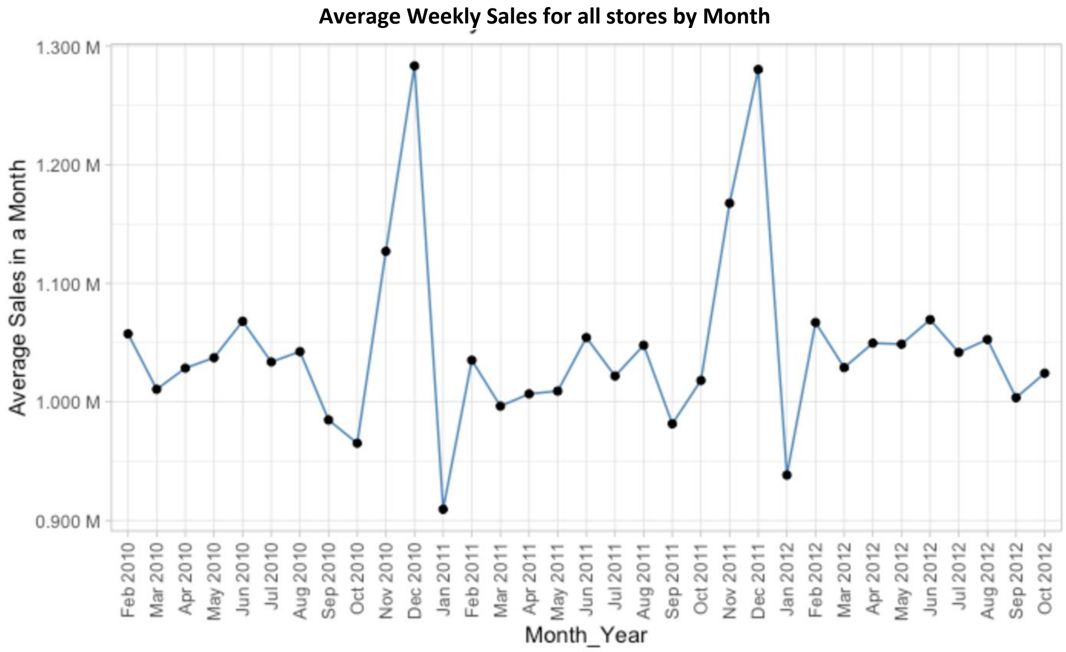


Fig 2: Average Weekly Sales by Month

We observe that during the months of November and December, there is a noticeable increase in sales while January has the lowest performance.

Since mean is highly prone to outliers and sudden changes in the positive or negative values, we also look at the boxplot for Weekly Sales [Fig 3] which gives us an estimation of the distribution of the sales values for each month.

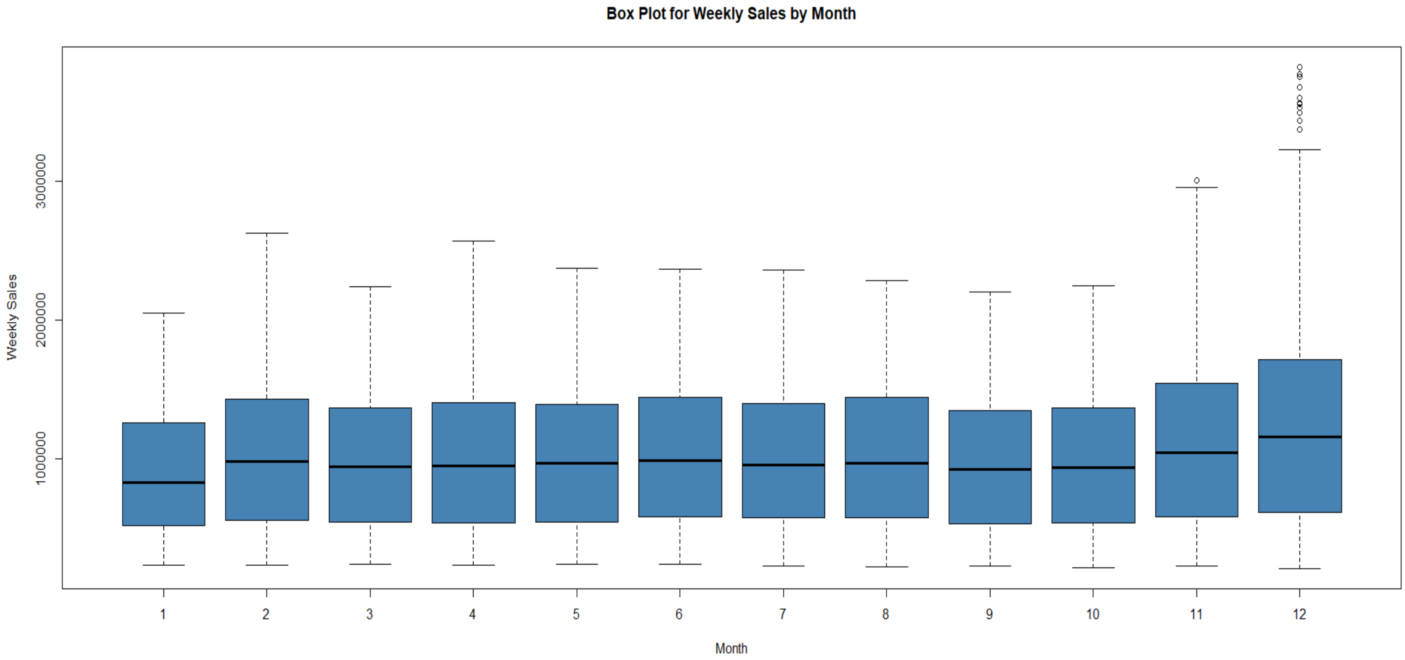


Figure 3: Boxplot for Weekly Sales by Month

Contrary to what we saw in [Fig 2] where there was a clear spike in the Average Weekly Sales in the months of November and December, the IQR and median values for each month in [Fig 3] are comparable.

This implies that a single high-performing holiday week is sufficient enough to lead a sharp rise of approximately $100,000 in November and $2,00,000 in December [Fig 2] in the Average Weekly Sales.

In order to gain further insight into the influence of 'Holidays' on sales, we plot [Fig 4] specifically to see if the impact on sales is on, during, or after the holiday.

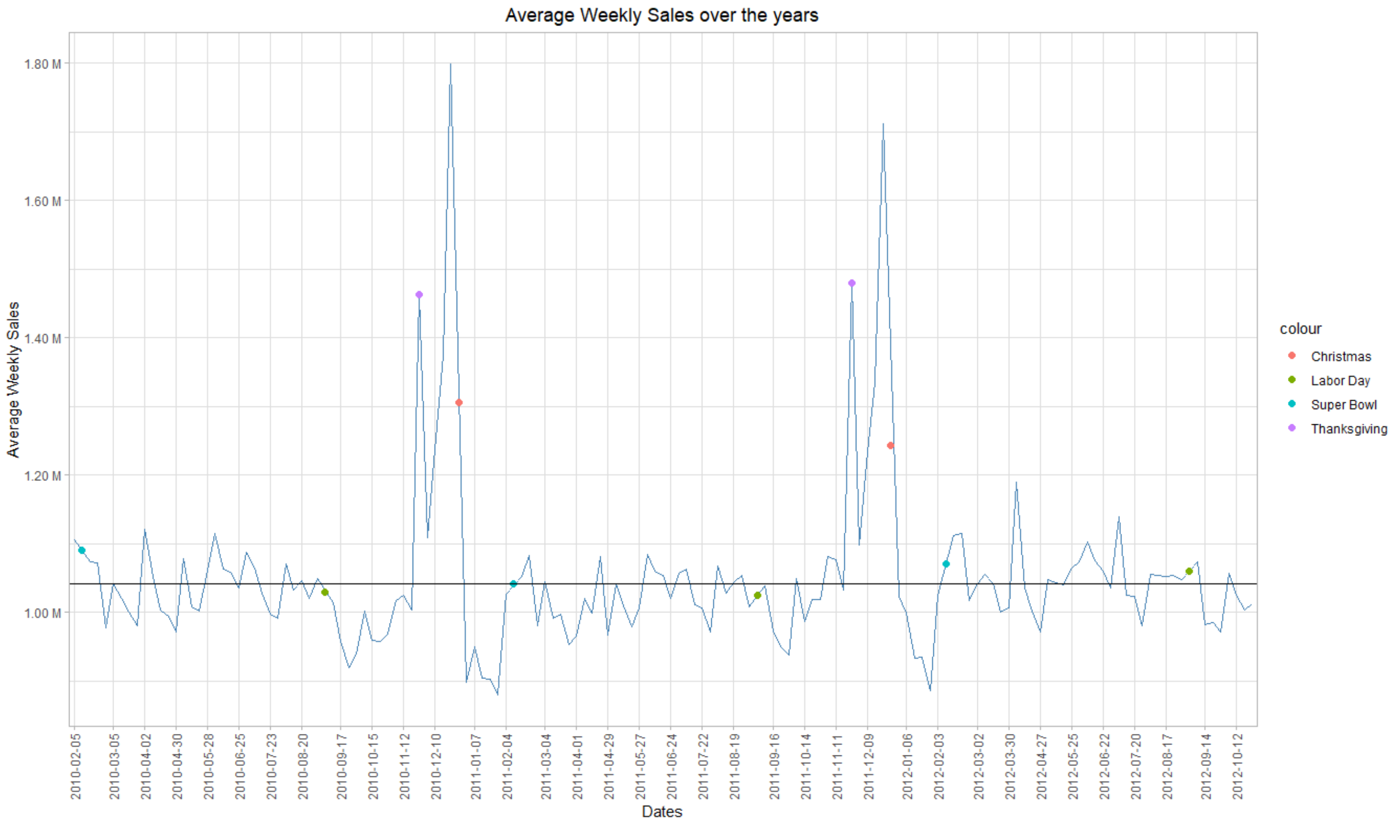


Figure 4: Average weekly sales over the years

Thus, for each holiday we observe a unique pattern in sales. For Christmas, highest amount of sales is earned in the weeks before the holiday. As for Thanksgiving, we observe highest sales during the Thanksgiving week. Super Bowl is usually preceded and followed by increased sales while Labour Day is followed by a dip in sales for every year.

As we know that the data represents 45 different Walmart stores across the country, we assume that the revenue generated by each store is a function of the location it is situated in. Thus, based on the highest sales generated before, on or after the holiday, we investigate the Average Sales of each store during those weeks.

Average Weekly Sales during Holiday season by Stores

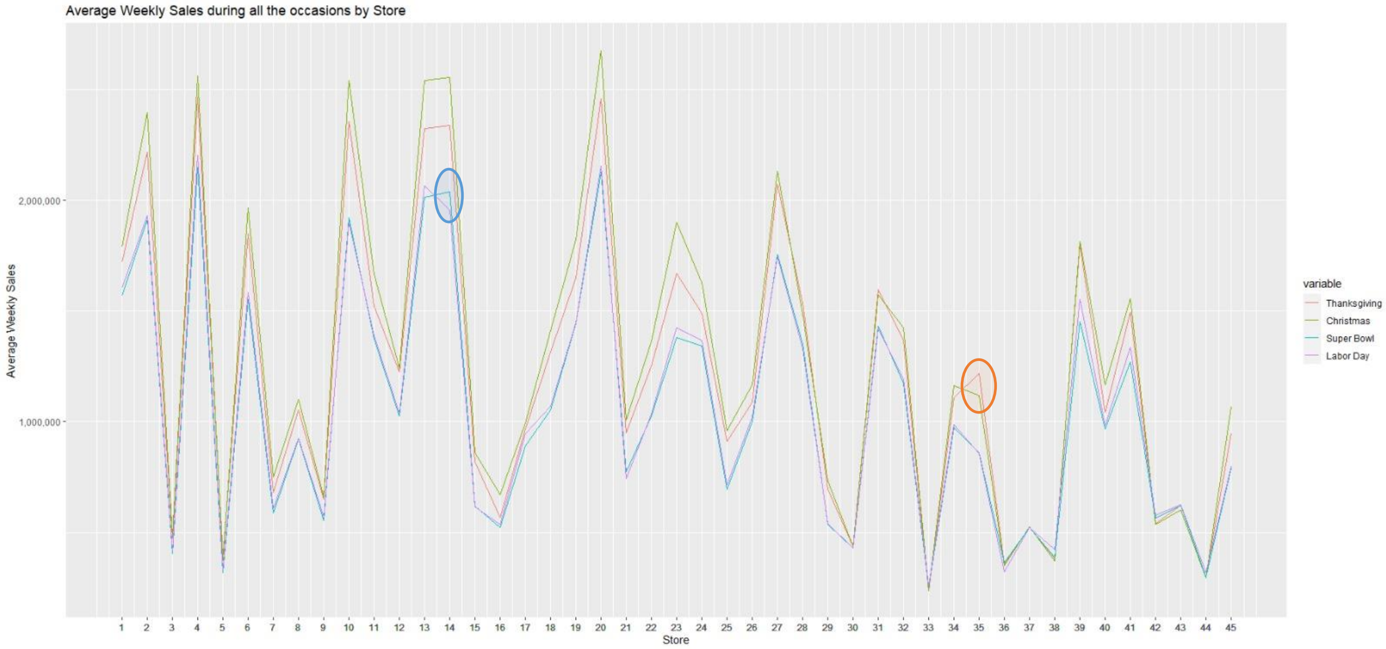


Figure 5: Sales of different stores during Holiday Weeks

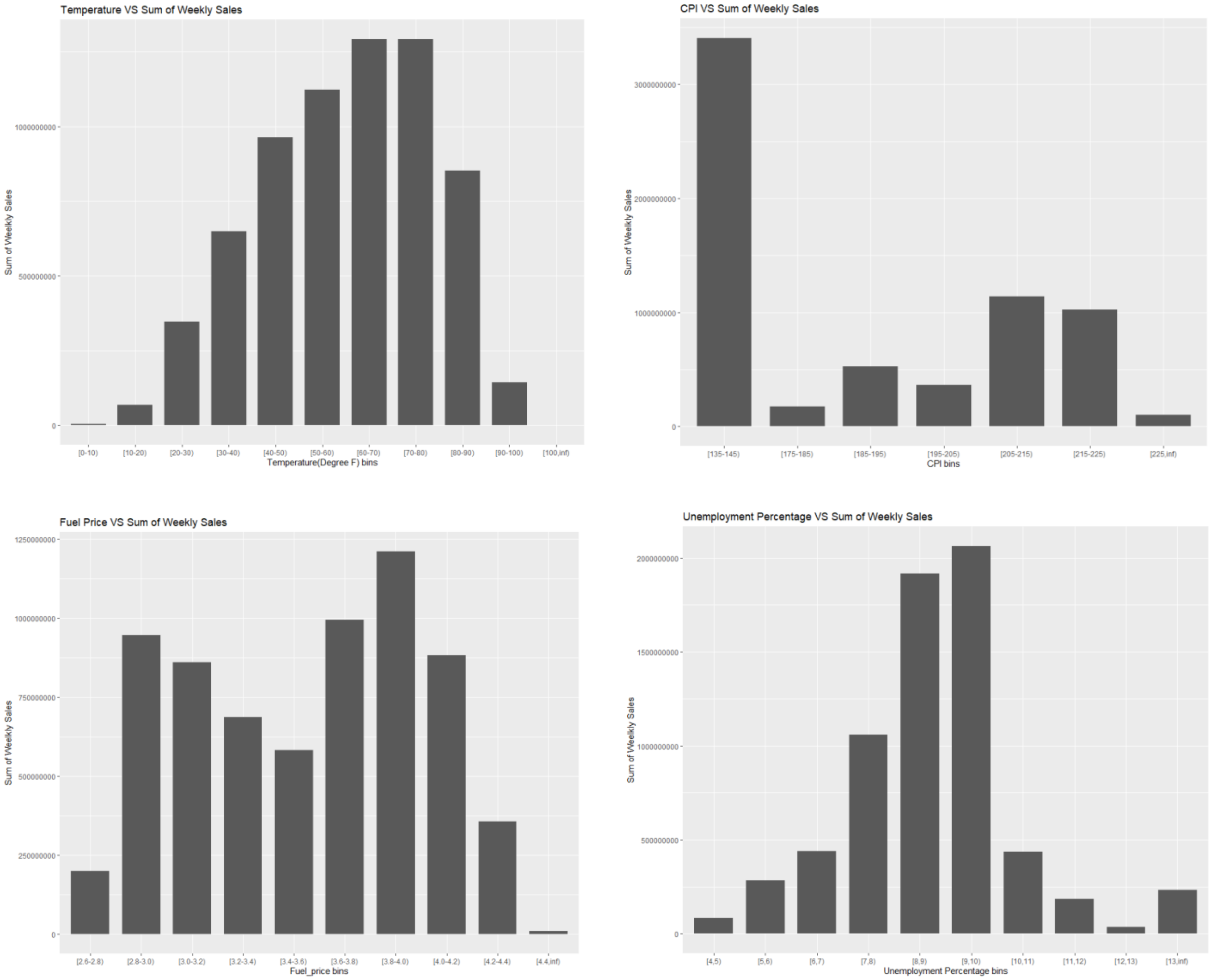
○-Thanksgiving sales cross Christmas sales
○-Super bowl sales cross Labor day sales

For most part of the plot, we see a similar trend for the stores for all the holidays. However, for store 35 we notice that the sales in Thanksgiving are higher than that of Christmas. Similarly, for Store 14, Super Bowl generated more revenue than Labour Day which is contrary to the trend. It's theoretically possible that this variation is related to differences in locality/region, but it certainly demands further investigation in the future.

We now look at the continuous variables that had a lower correlation with the target variable 'Weekly Sales'. We create 10 bins for each variable and plot the Sum of the Weekly Sales generated for each of the bins. The plots in the [Fig 6] are as follows:

- Temperature **VS** Sum of Weekly Sales [Top Left]
- CPI **VS** Sum of Weekly Sales [Top Right]
- Fuel Price **VS** Sum of Weekly Sales [Bottom Left]
- Unemployment **VS** Sum of Weekly Sales [Bottom Right]

Non-linear relationship of Continuous Variables with Weekly Sales



Temperature VS Sum of Weekly Sales

CPI VS Sum of Weekly Sales

Fuel Price VS Sum of Weekly Sales

Unemployment Percentage VS Sum of Weekly Sales

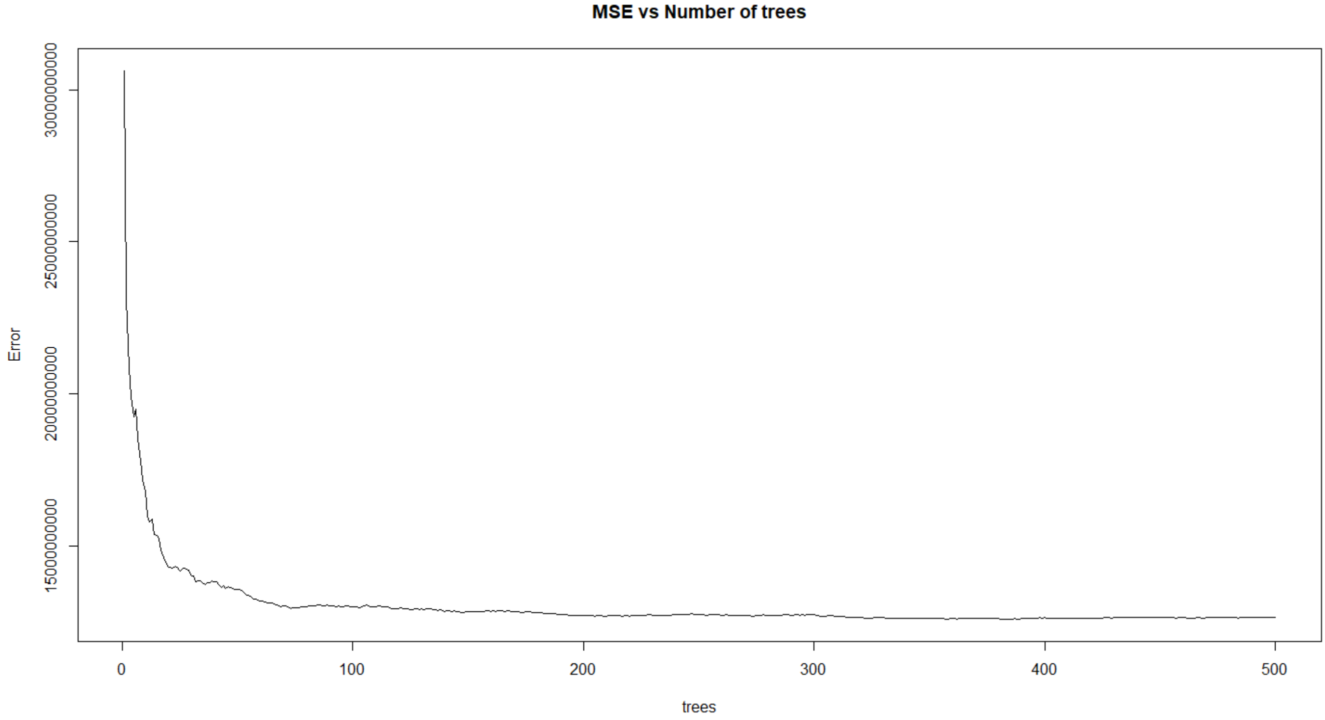
Figure 6: Continuous Variables VS Sum of Weekly Sales

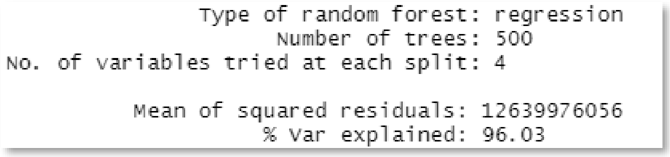
From the plots in [Fig 6] we can clearly see that the continuous variables have a non-linear relationship with the Weekly Sales. In order to check whether these non-linear relationships help us to explain the target variable Weekly Sales, we leverage the Random Forest model using all the features to check the fit of the target variable.

Non-linear correlations between input features and the target variable can be found using Random Forest. As, an ensemble of decision trees is referred to as random forest, its inner working is highly similar to that of a Decision tree. A Decision Tree may be conceived as a series of if-else conditions. A Random Forest is made up of several trees that have been built in a "random" approach. Each tree generates its own unique prediction. The average of these predictions is then used to give a single solution.

Fitting the Random Forest Model:

After fitting the Random Forest model to our data, we observe that the optimum number of trees are set to 500 at which we get the lowest MSE value. The RMSE value at 500 Decision trees is 112,427.





```
              Type of random forest: regression
                    Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 12639976056
                    % Var explained: 96.03
```

Figure 7

Finally, we look at the top features within our model by Mean Decrease Accuracy and Mean Decrease Gini [Fig 8]. The above parameters are explained as follows:

- Mean Decrease Accuracy (%IncMSE) - This shows how much our model accuracy decreases if we leave out that variable.

- Mean Decrease Gini (IncNodePurity) - This is a measure of variable importance based on the Gini impurity index used for the calculating the splits in trees.

The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable to our model.

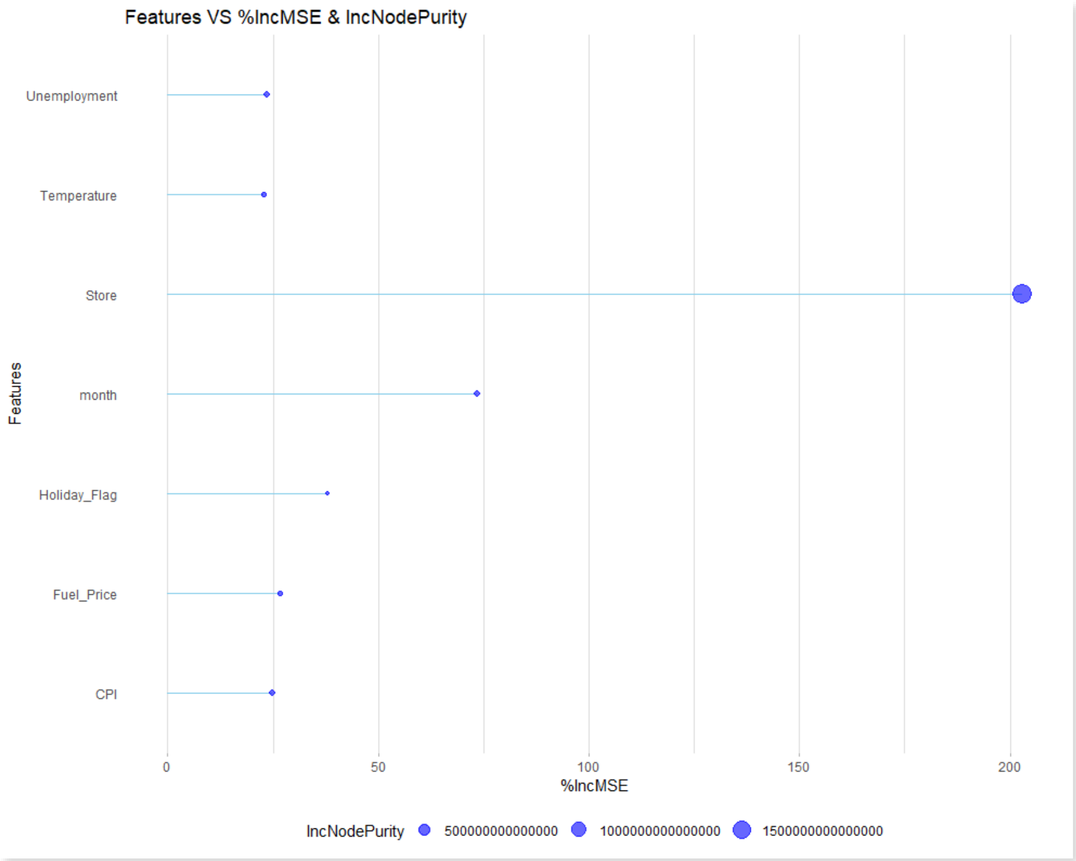


Fig 8

Thus, our model clearly states that the topmost feature is the 'Store' followed by 'month', while all our continuous variables have a lower IncMSE score of around 25 which suggests that they do not influence the target variable to a great extent.

Notable Observations and Conclusion:

- Throughout the years, we've seen that the highest sales occurred in December, while the lowest being in January.
- The total sales for the month of June were higher than those for Labour Day and the Super Bowl. This might be due to the various promotional offers and sales during the summer.
- Among the holidays, the most sales were made at Christmas, while Labour Day had the least influence.
- With just a couple of exceptions, we found that most of the stores have similar trends for each holiday which could be due to the monolithic culture of the country.
- Finally, after fitting the model we observe that the most important features are Store and month, which suggests that sales are mostly influenced by the location of the store and the time of the year.

Limitations:

- There were missing values for January 2010, November 2012 and December 2012, in the dataset which also affected the analysis to some extent.
- The continuous features in the dataset had very less correlation to the Weekly Sales and thus, did not add to explanation of the sales trends.
- The locations of the stores were not mentioned in the dataset which further limited the scope of the analysis.
- Additionally, some features like discounts or promotional offers could potentially help us in getting a better understanding of the customer behaviour.