Report for Walmart Sales Data Analysis and modelling

Project Description:

- The project deals with sales data from 45 Walmart stores.
- The primary challenge is to accurately predict demand and avoid stockouts.
- Economic factors like CPI, Unemployment Index, and others are considered.

Analysis Tasks:

1. Identify the store with the maximum sales.
2. Identify the store with the maximum standard deviation in sales and calculate the coefficient of mean to standard deviation.
3. Determine stores with good quarterly growth rates in Q3'2012.
4. Identify holidays that have higher sales than the mean sales in the nonholiday season.
5. Provide a monthly and semester view of sales and provide insights.

Statistical Model:

- The project mentions building a prediction model for Store 1 using linear regression.
- Hypotheses are formulated regarding the impact of CPI, unemployment, and fuel prices on sales.
- Model selection is based on accuracy.

Project Execution:

- The project begins by loading necessary libraries (e.g., dplyr, ggplot2, lubridate) and reading the dataset.

- Data exploration is performed, including checking dimensions, class, structure, and summary statistics.
- Various analysis tasks are executed, including finding the store with maximum sales, maximum standard deviation, good quarterly growth rates, and holidays with high sales.
- Monthly and semester views of sales are visualized and insights are provided.
- The project moves on to building a statistical model using linear regression.
- Outliers are detected and removed using bivariate box plots.
- Correlation analysis and correlation matrix visualization are performed.
- Dummy variables are created for categorical features.
- Finally, a linear regression model is created and evaluated on the training and test datasets.

The project combines data exploration, data pre-processing, statistical analysis, and machine learning to provide insights and predictions for Walmart store sales. Further details on the dataset, model accuracy, and specific insights are not provided in this code snippet.

Dataset link: - https://www.kaggle.com/datasets/surajjjjjjjjj/dataset-forwalmart-stores

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	1	05-02-2010	1643691	0	42.31	2.572	211.0964	8.106
2	1	12-02-2010	1641957	1	38.51	2.548	211.2422	8.106
3	1	19-02-2010	1611968	0	39.93	2.514	211.2891	8.106
4	1	26-02-2010	1409728	0	46.63	2.561	211.3196	8.106
5	1	05-03-2010	1554807	0	46.50	2.625	211.3501	8.106
6	1	12-03-2010	1439542	0	57.79	2.667	211.3806	8.106
7	1	19-03-2010	1472516	0	54.58	2.720	211.2156	8.106
8	1	26-03-2010	1404430	0	51.45	2.732	211.0180	8.106
9	1	02-04-2010	1594968	0	62.27	2.719	210.8204	7.808
10	1	09-04-2010	1545419	0	65.86	2.770	210.6229	7.808
11	1	16-04-2010	1466058	0	66.32	2.808	210.4887	7.808

Analysis Tasks:

1. Store with Maximum Sales

- The store with the maximum sales is Store 20, with total sales of $301.39 million.

Code:

```
#Aggregating data by 'Store' and Finding sum of 'Weekly_Sales'

Store_Sales<- aggregate(Weekly_Sales ~ Store, data = data1, sum)

#Changing column name of sales  colnames(Store_Sales)[2]

<- "Total_Sales_by_Store"

#Finding out Store with highest Sales

Store_Sales <-arrange(Store_Sales, desc(Total_Sales_by_Store)) #Arranged Stores based on Sales in descending order

Store_Sales[1,] #Choosing the first store that comes in this order

#Printing the output print(paste('Store no.',

Store_Sales[1,]$Store,

        'has the maximum sales and the value is = ', Store_Sales[1,]$Total_Sales_by_Store))
```

Output:

```
> #Printing the output
> print(paste('Store no.', Store_Sales[1,]$Store,
+             'has the maximum sales and the value is = ', Store_Sales[1,]$Total_Sales_by_Store))
[1] "Store no. 20 has the maximum sales and the value is =  301397792.46"
>
```

2. Store with Maximum Standard Deviation

- Store 14 has the maximum standard deviation of sales, which is 317,569.95, with a coefficient of variation (CV) of 15.714%.

3. Stores with Good Quarterly Growth Rate in Q3’2012

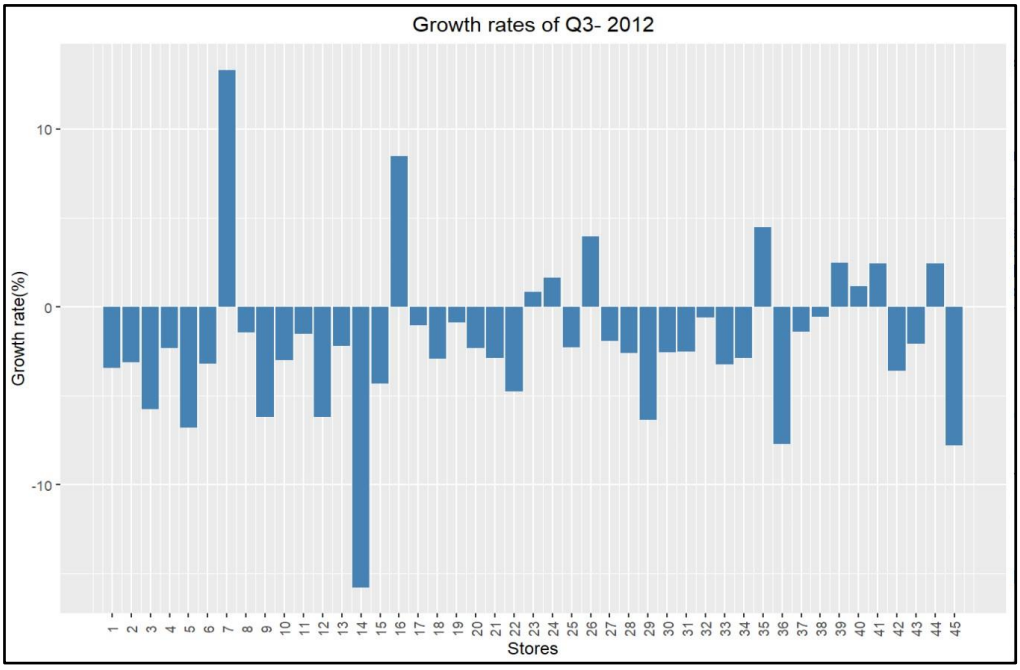
Store 7 has the highest growth rate of 13.33% in Q3'2012.

Other stores with positive growth rates include Store 16, Store 35, and 7 more stores.

-Store 14 has highest negative growth rate.

Output:

	Store	Q2_2012_Sales_by_Store	Q3_2012_Sales_by_Store	Growth_Rate
5	39	20214128	20715116	2.4784040
10	23	18488883	18641489	0.8253951
8	24	17684219	17976378	1.6520877
6	41	17659943	18093844	2.4569801
4	26	13155336	13675692	3.9554775
9	40	12727738	12873195	1.1428413
3	35	10838313	11322421	4.4666372
1	7	7290859	8262787	13.3307760
2	16	6564336	7121542	8.4883781
7	44	4306406	4411251	2.4346377



4.

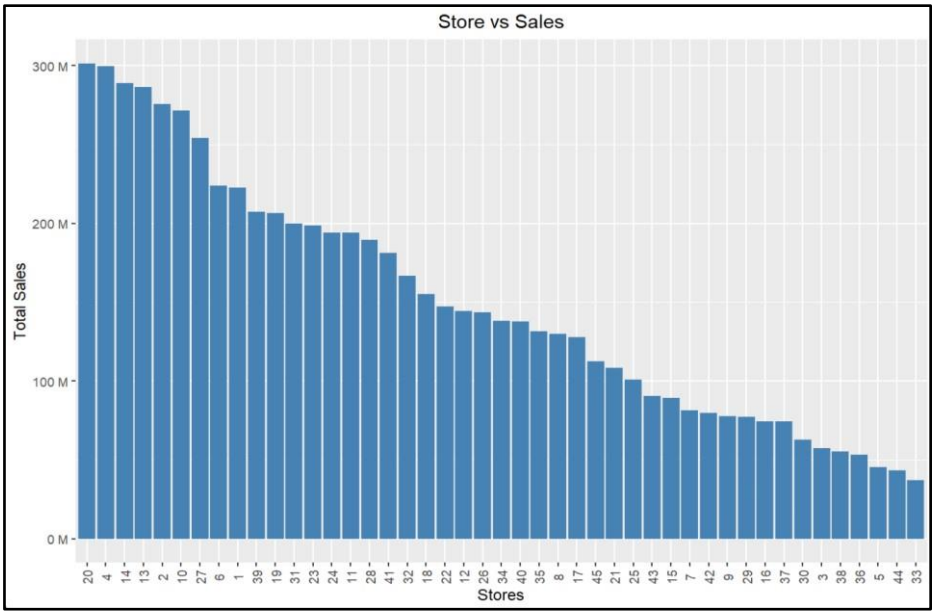
Plotting Store vs TotalSales

Code:

```
options(repr.plot.width = 14, repr.plot.height = 8)

a<-ggplot(data=Store_Sales, aes(x=Store, y=Total_Sales_by_Store)) +
geom_bar(stat="identity",fill="steelblue") +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5, hjust=0.5))+
scale_x_discrete(breaks = data1$Store)+
  scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6))+
ggtitle('Store vs Sales')+
  theme(plot.title = element_text(hjust =
0.5))+   xlab("Stores") + ylab("Total Sales") a
```

output:



Insights:

Store 20 has the maximum sales and the value is 301397792.46 (i.e., 301.39M)

Store 4 is the second largest store in terms of sales and the value is 299543953 (i.e., 299.54M)

Store 33 has the least sales 37160222 (i.e., 37.16M)

5.

Holidays with Higher Sales than Mean Sales in Non-Holiday Season

-Christmas and Labour Day has negative impact on sales where as Thanks giving and Super Bowl has positive impact on sales

- Holidays such as Super Bowl, Thanksgiving, and Labor Day have higher sales than the mean sales during non-holiday seasons.

Output:

	Events	Mean_Sales_by_Event_Type
1	Christmas	960833.1
2	Labour Day	1042427.3
3	No_Holiday	1041256.4
4	Super Bowl	1079128.0
5	Thanksgiving	1471273.4

	Date	mean(Weekly_Sales)	higher_than_non_holiday[,"mean(Weekly_Sales)"]
1	07-09-2012	1074001.3	TRUE
2	09-09-2011	1039182.8	FALSE
3	10-02-2012	1111320.2	TRUE
4	10-09-2010	1014097.7	FALSE
5	11-02-2011	1051915.4	TRUE
6	12-02-2010	1074148.4	TRUE
7	25-11-2011	1479857.9	TRUE
8	26-11-2010	1462689.0	TRUE
9	30-12-2011	1023165.8	FALSE
10	31-12-2010	898500.4	FALSE

Insights:

Super Bowl, Thanks giving, Labour Day has higher sales than mean sales of a non-Holiday and creating positive impact on sales.
9th Sept 2011, 10th Sept 2010 ,30th Dec 2011, 31st Dec 2010 were the dates which created negative impact on sales
All the dates related to Christmas have low sales than mean, whereas all the dates related to Super Bowl, Thanks giving have high sales than mean.

6.

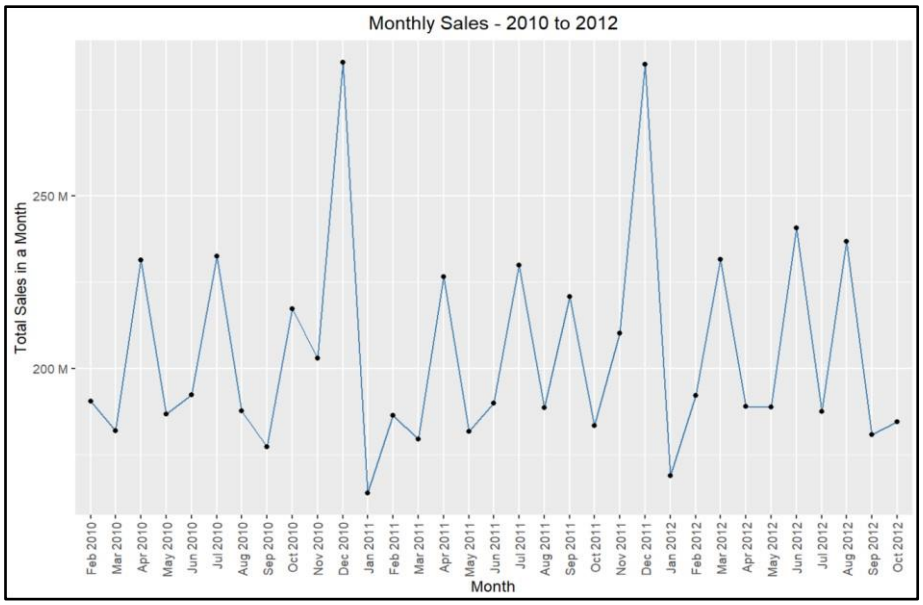
It is interesting to note that Labour Day has overall positive impact on sales inspite of having two days below mean value.

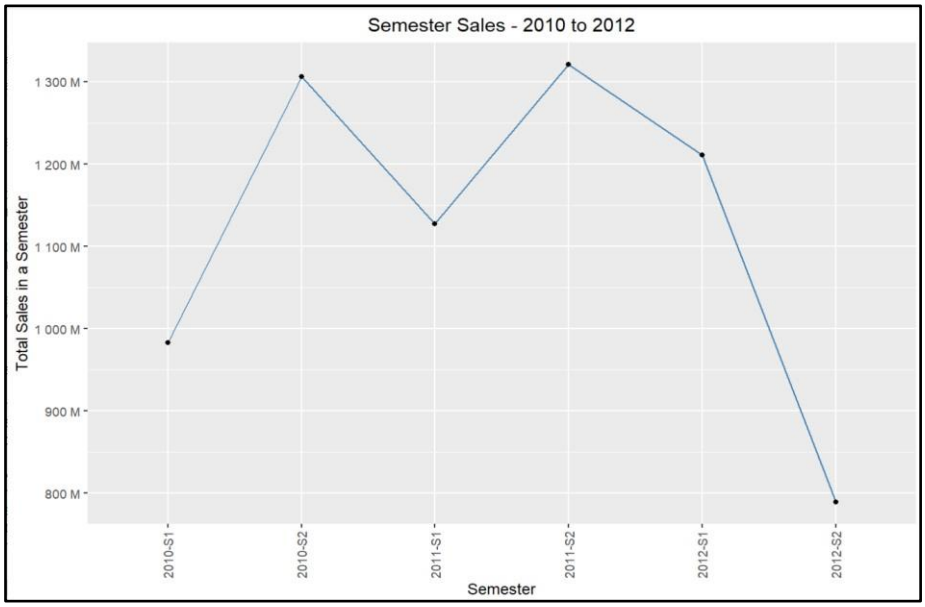
Monthly and Semester Sales View

- The monthly sales view shows fluctuations in sales over time, with December having the highest sales.

- The semester sales view reveals that sales are generally higher in the second semester of each year.

Output:





7.

Insights:

The sales are highest in December and Lowest in January

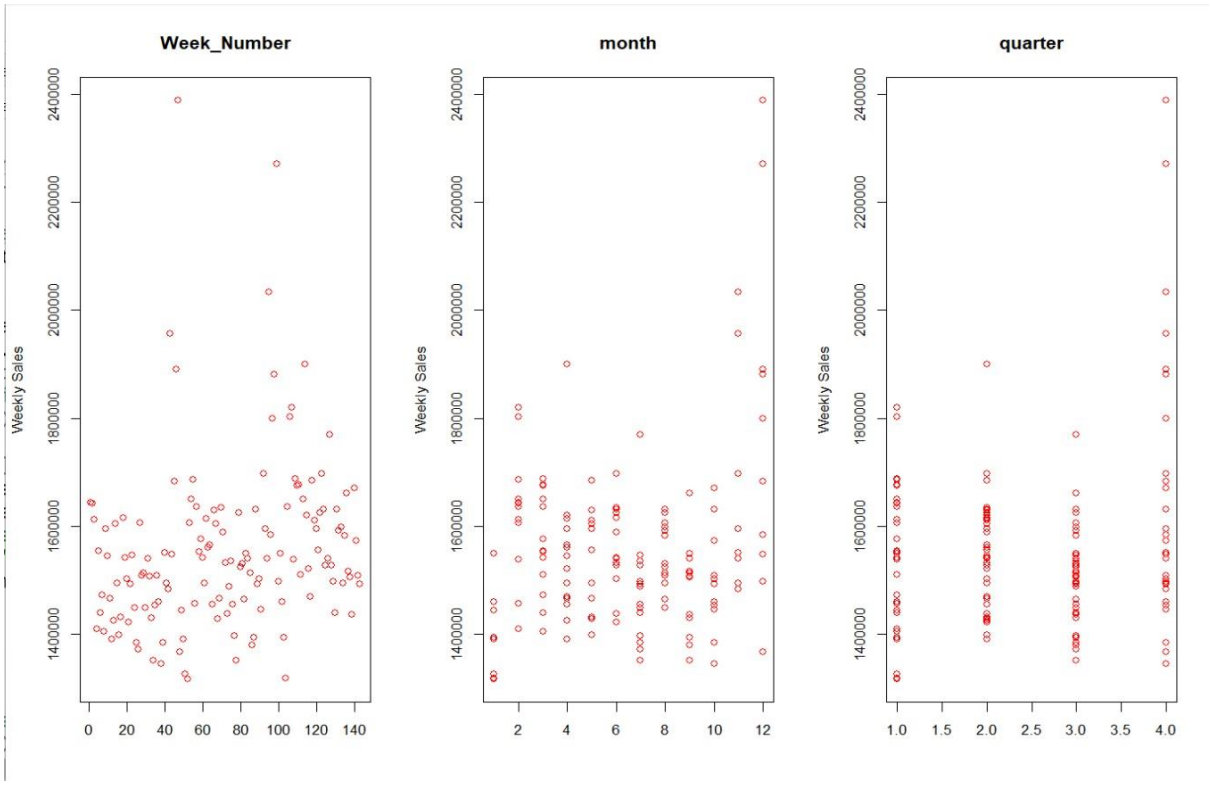
The sales are higher in second semester of every year

The plot shows a drop in S2-2012 & S1-2010. It is due to absence of Jan data in S1-2010 & Nov-Dec 2012 data in S2-2012.

Statistical Model:

7. Linear Regression Model for Store 1

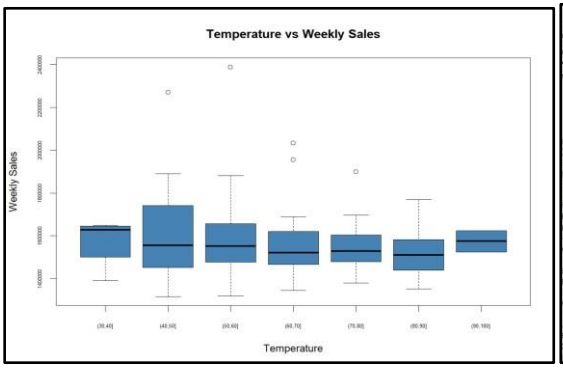
- A linear regression model was built for Store 1 to forecast demand.
- Variables such as date, week number, fuel price, CPI, and unemployment were considered.
- The model's accuracy was assessed to select the best-performing model.

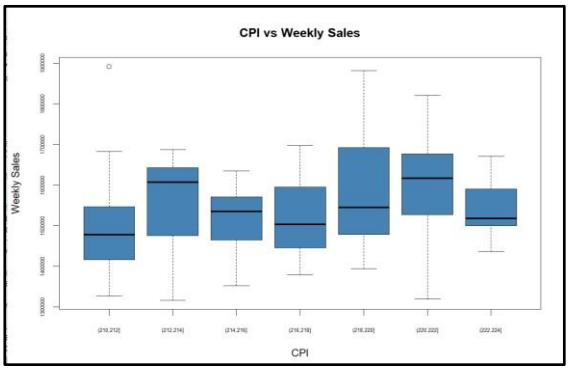


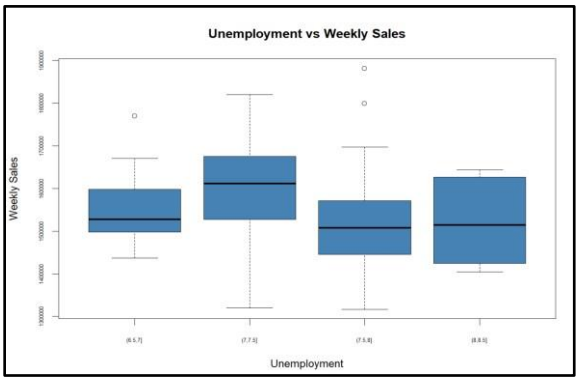
Detection and removal of outliers using Bi Variate Box Plots:

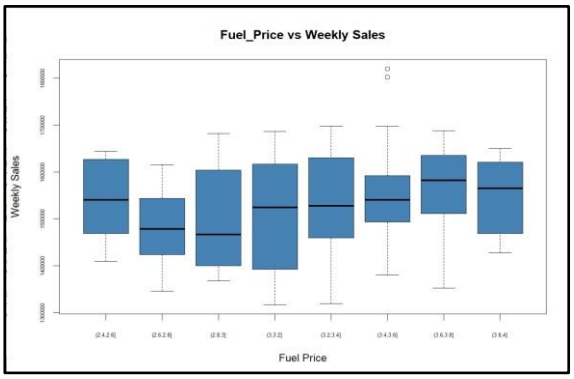
8. Outlier Detection

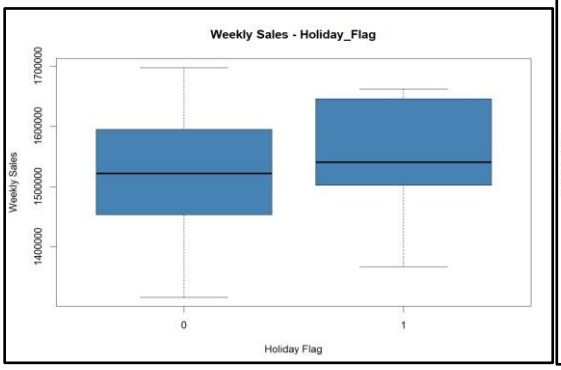
- Outliers in various parameters (temperature, CPI, unemployment, fuel price, month, quarter) were detected using box plots.

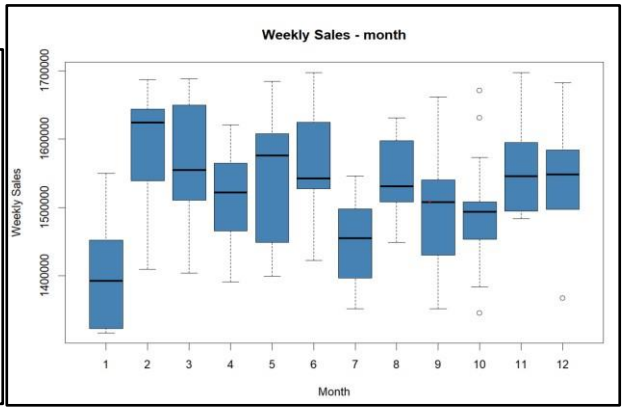

Temperature vs Weekly Sales
Weekly Sales
Temperature

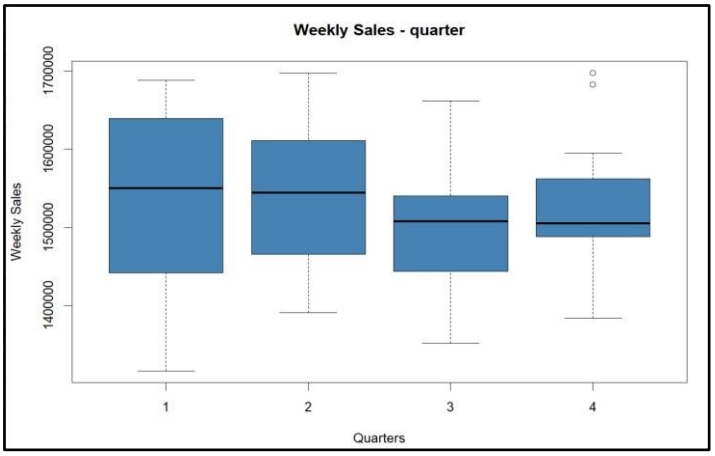

CPI vs Weekly Sales
Weekly Sales
CPI


Unemployment vs Weekly Sales
Weekly Sales
Unemployment


Fuel_Price vs Weekly Sales
Weekly Sales
Fuel Price


Weekly Sales - Holiday_Flag
Weekly Sales
1400000
1500000
1600000
1700000
0
1
Holiday Flag


Weekly Sales - month
Weekly Sales
1400000
1500000
1600000
1700000
1
2
3
4
5
6
7
8
9
10
11
12
Month


Weekly Sales - quarter
Weekly Sales
1400000
1500000
1600000
1700000
1
2
3
4
Quarters

9. Outlier Treatment

- Outliers were removed from the dataset to improve model performance.

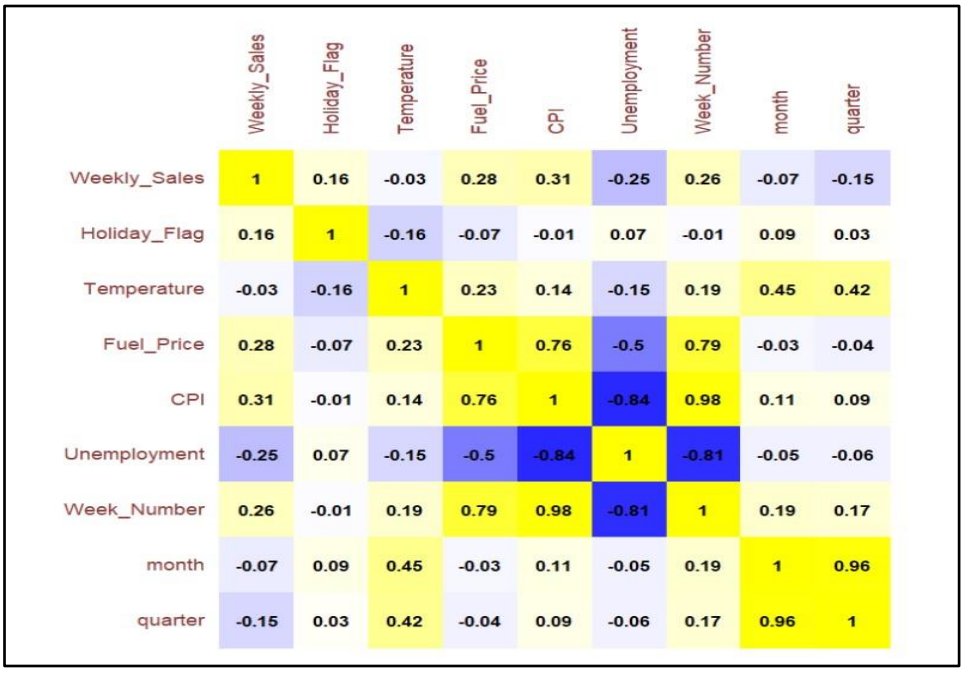
Temperature 5 outliers
CPI Outlier treatment 1 outliers
Unemployment 3 outliers
Fuel price 2 outliers
Holiday Flag - No outliers
Month - 4 outliers
Quarter - 2 outliers

Correlation Analysis:

10.Correlation Matrix and Correlation Plot

- Correlation analysis revealed low correlations between temperature and weekly sales.
- Categorical variables such as month, quarter, and holiday flag showed low correlations with weekly sales.

Output:



	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Week_Number	month	quarter
Weekly_Sales	1	0.16	-0.03	0.28	0.31	-0.25	0.26	-0.07	-0.15
Holiday_Flag	0.16	1	-0.16	-0.07	-0.01	0.07	-0.01	0.09	0.03
Temperature	-0.03	-0.16	1	0.23	0.14	-0.15	0.19	0.45	0.42
Fuel_Price	0.28	-0.07	0.23	1	0.76	-0.5	0.79	-0.03	-0.04
CPI	0.31	-0.01	0.14	0.76	1	-0.84	0.98	0.11	0.09
Unemployment	-0.25	0.07	-0.15	-0.5	-0.84	1	-0.81	-0.05	-0.06
Week_Number	0.26	-0.01	0.19	0.79	0.98	-0.81	1	0.19	0.17
month	-0.07	0.09	0.45	-0.03	0.11	-0.05	0.19	1	0.96
quarter	-0.15	0.03	0.42	-0.04	0.09	-0.06	0.17	0.96	1

Insights:

Observed very low correlation between Temp and Weekly Sales – So can omit Temperature

Observed low correlation between month, quarter, Holiday Flag with Weekly_Sales may be due to considering categorical variables as continuous variables

Statistical Model Development:

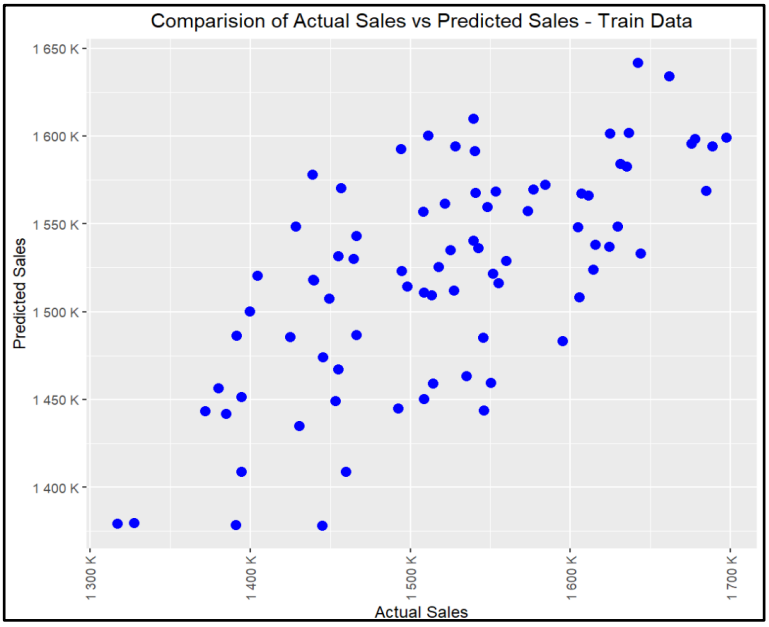
11.Liner Regression Model

- A linear regression model was developed using selected parameters.

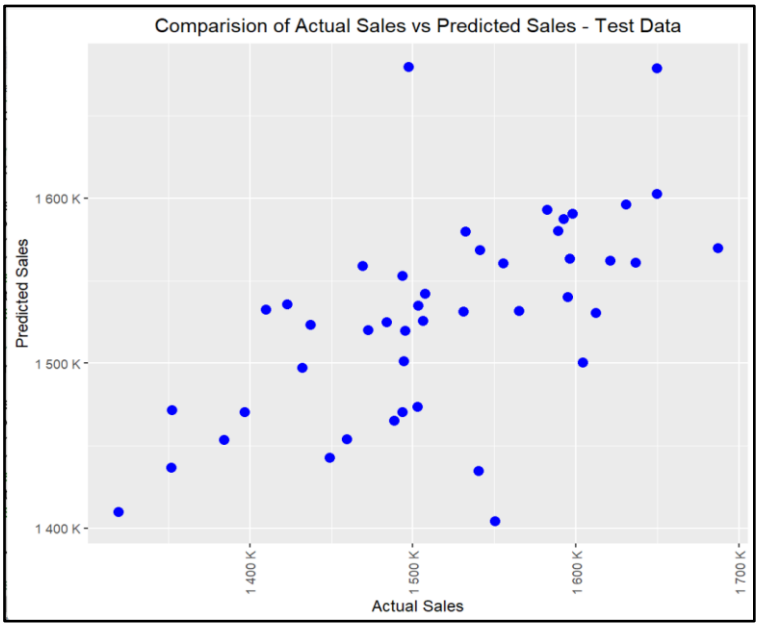
- The dataset was split into training and test sets for model evaluation.

```
options(repr.plot.width = 10, repr.plot.height = 10)
```

For Train dataset Output:



For Test dataset Output:



Parameters to validate the accuracy of the model and improvise.

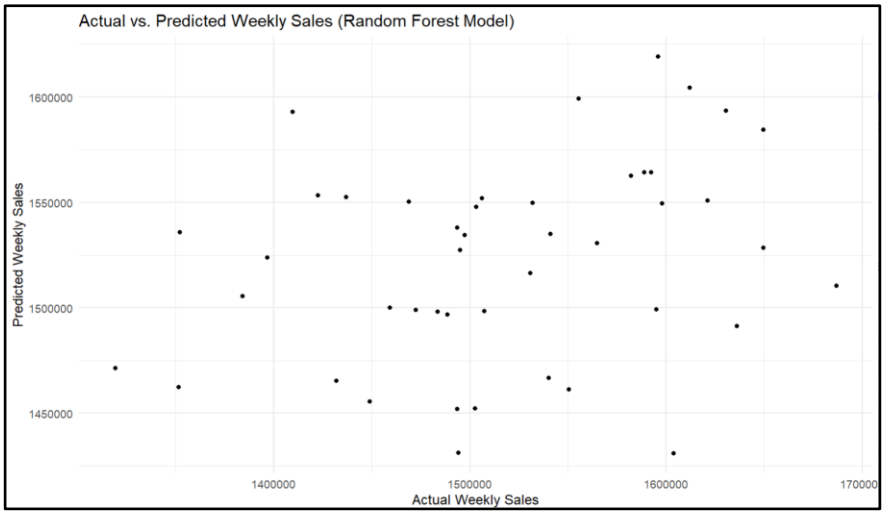
MAPE(y_pred_test,testSet$Weekly_Sales)

RMSE(y_pred_test,testSet$Weekly_Sales)

Output:

```
> ### Parameters to validate the accuracy of the model and improvise.
> MAPE(y_pred_test,testSet$Weekly_Sales)
[1] 0.04486674
> RMSE(y_pred_test,testSet$Weekly_Sales)
[1] 83085.05
```

12. Random Forest Model:




```
> print(paste("Random Forest MAPE:", rf_mape))
[1] "Random Forest MAPE: 0.04483934451462"
> print(paste("Random Forest RMSE:", rf_rmse))
[1] "Random Forest RMSE: 85496.251985434"
```

Conclusion:

Overall, these two algorithms, Linear Regression and Random Forest, are applied to the same dataset to predict Weekly Sales and are evaluated using MAPE and RMSE to gauge their performance.

Random Forest has a slightly higher MAPE (0.0466) compared to Linear Regression (0.0449).

Lower MAPE indicates a better percentage accuracy in predicting sales. Therefore, Linear Regression has a slightly better performance in this regard.

Random Forest has a slightly lower RMSE (82445.91) compared to Linear Regression (83085.05).

Lower RMSE indicates better accuracy in predicting the actual sales values. In this case, Random Forest has a slightly better RMSE. The differences in MAPE and RMSE between the two models are relatively small.

But here our main concern is percentage accuracy MAPE(Mean Absolute Percentage Error) that is predicting sales rather than predicting actual sales, Linear Regression performs slightly better.

Predicting sales is crucial for strategic planning, inventory management, resource allocation, and overall business decision-making.

Name: Suraj Shete
Gmail: 02surajshete@gmail.com
MyLinkedInProfile
MyGithub

THANKYOU