GUVI ZEN FINAL PROJECT - 3
SALES FORECASTING PERFORMANCE AND KEY INSIGHTS

NAME: **ASTER NATHAN**
BATCH: **DM3**

CONTENTS

1. INTRODUCTION:

Sales forecasting is the process of predicting future sales for a business or organization. It is an important tool that helps businesses plan and prepare for the future by estimating the volume of goods or services that they are likely to sell over a given period of time.

Sales forecasting is critical for businesses of all sizes and types. It provides valuable information for budgeting, resource allocation, and decision-making. Accurate sales forecasting helps businesses ensure that they have the right amount of inventory on hand to meet demand, and that they are properly staffed and equipped to handle future sales volume. Overall, sales forecasting is a critical tool for businesses that want to plan for the future and achieve their financial goals. By accurately predicting sales trends, businesses can make better decisions, optimize their operations, and maximize their profitability.

2. OBJECTIVE OF THE PROJECT:

The objective of the project is to predict the weekly sales value of a store based on various features.

3. DATA:

i. SOURCE:

The Dataset was provided by the Guvi mentors.

ii. SIZE:

Contains 421570 RECORDS & 16 FEATURES

Records were collected between the timeframe of 2010-02-05 to 2012-10-26

iii. ATTRIBUTES **:**

- Store:
  Gives us information on the number of stores, there are in total 45 different stores in the dataset.
- Dept:
  Department number of particular store that each store consists
- Date:
  The date when the input was recorded
- Weekly_Sales:
  The value of the sale of a particular week
- IsHoliday:
  Whether the day was a Holiday or Not a Holiday
- Temperature:
  The temperature recorded on the specific day
- Fuel_Price:

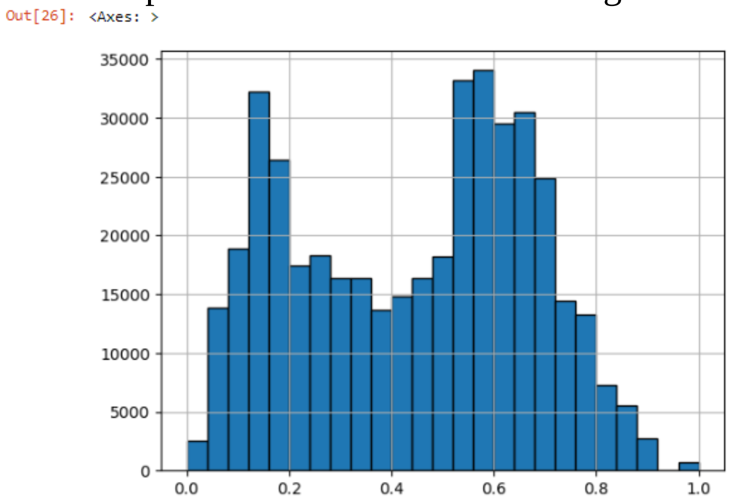
The Fuel Price on the specific day

- MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5:
  Every markdown consists of different schemes applied by the department to increase the sale
- CPI:
  CPI stands for Consumer price index and is a measure of inflation that tracks the changes in the prices of a basket of goods and services purchased by households.
- Unemployment:
  Unemployment refers to the number of employees per square area
- Type:
  Type refers to the store type divided into three categories (A, B and C)
- Size:
  Size is the total area of the store

4. FEATURE ENGINEERING:

i. Removing outliers from column weekly_sales using **z-score** method

ii. Column fuel price has been normalizes using a user defined **min-max()** function



This figure gives us an insight on fuel price column after normalizing

iii. Removed null values using **fillna()** function by replacing it with **mean()** of the c olumn

iv. Conversion of string values into numeric:-columns ‘Type’ was initially a column consisting of 3 string values(‘A’,’B’,’C’) which was later converted into numeri c(1,2,3)using **.replace()** function

5. FEATURE SELECTION:

i. Only features :-Store,dept,markdown1,markdown3,markdown5 & size were taken further in model development as rest of the features were less correlated with dependent variable (weeekly_sales).

6. MODELING:-

- from sklearn.tree import DecisionTreeRegressor
- We have used decision tree model for the model training of train set data
- This gives an accuracy of **89.76%**
- Score of mean_absolute_error = 1865.6381
- Score of mean_squared_error = 5088.6526
- Labeled data=’Weekly_Sales’

```
[55]: #SCATTER PLOT BETWEEN OBSERVED AND PREDICTED VALUES OF WEEKLY SALES FROM decisionTree REGRESSOR
      plt.figure(figsize = (8,5))
      sns.scatterplot(x = y_test, y = y_pred)
```

```
[55]: <Axes: xlabel='Weekly_Sales'>
```

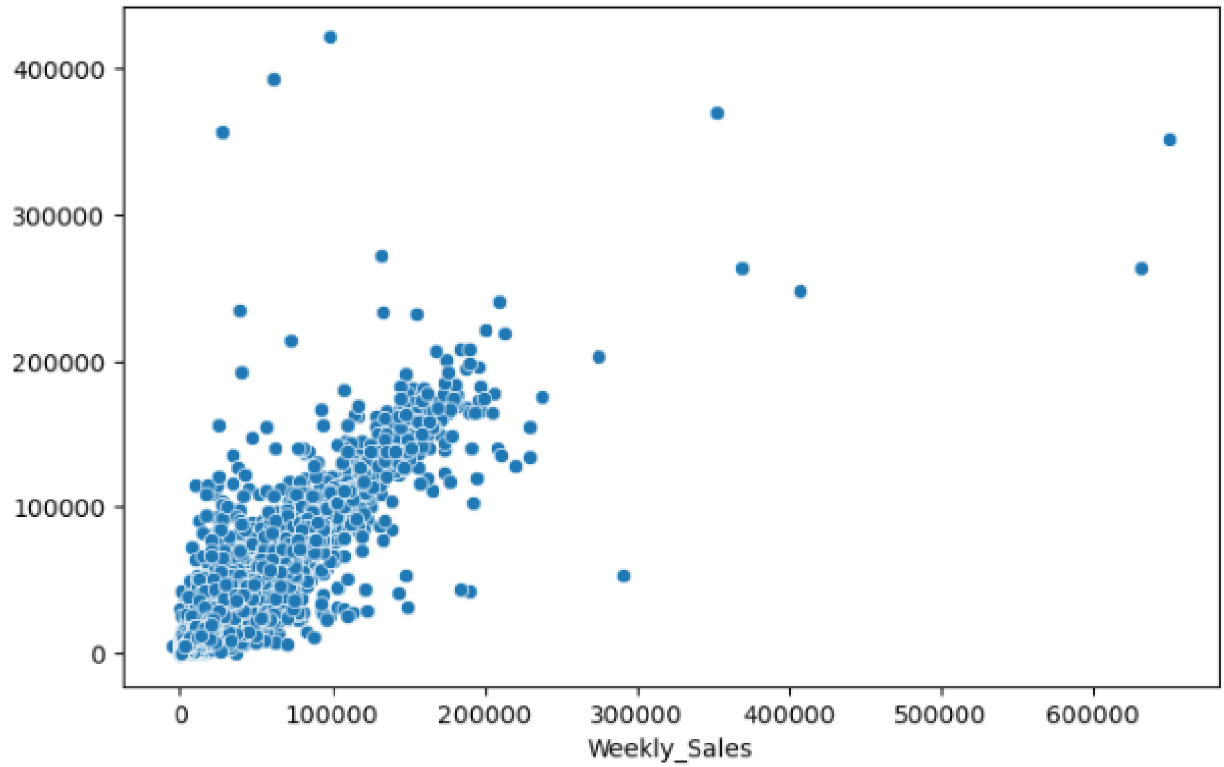
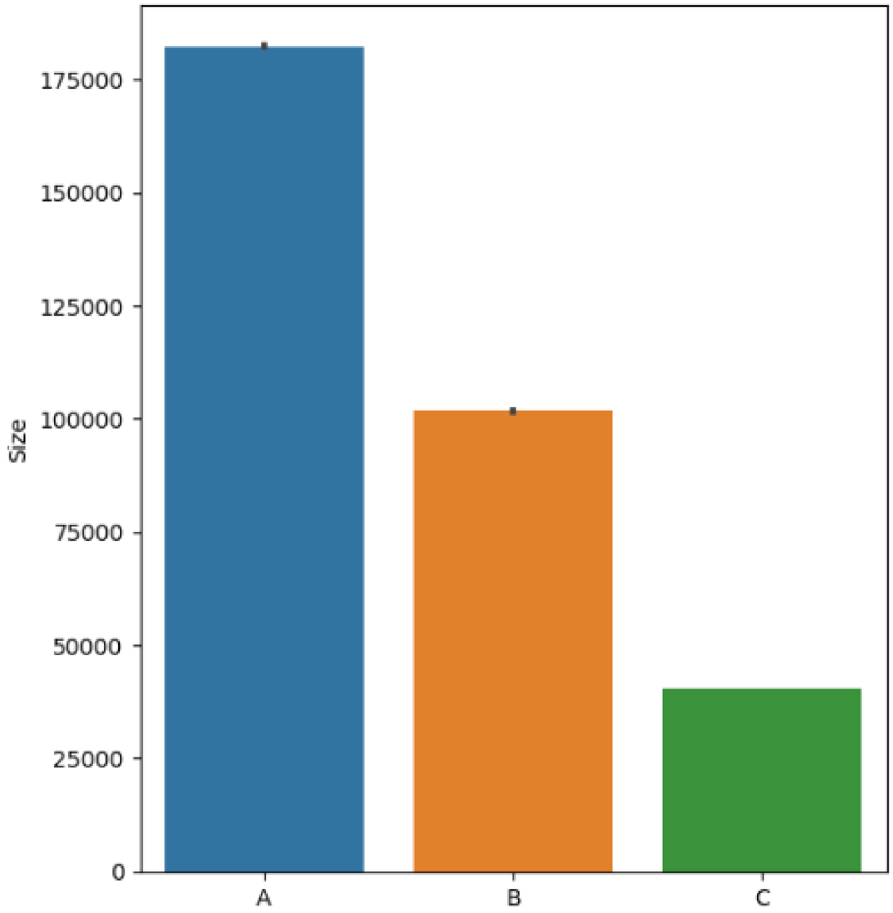


Figure above shows the distribution of data points over the line of decision tree

7. EDA:

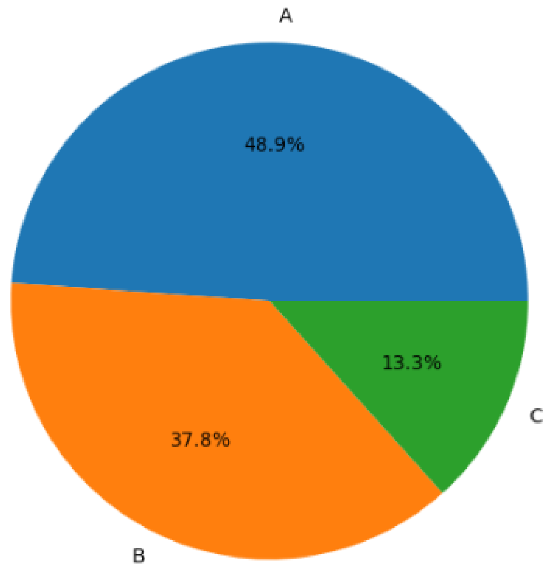
]: <Axes: xlabel='Type', ylabel='Size'>



- The graph is a barplot used to plot the count between types of stores avai lable in dataset and their sizes
- It is imported from the seaborn library
- The small line in every bar depicts the outliers in the dataset.

```
print(temp)
plt.figure(figsize = (8,6))
plt.pie(temp['Store'],labels = temp['Type'],autopct='%1.1f%%')
plt.show()
```

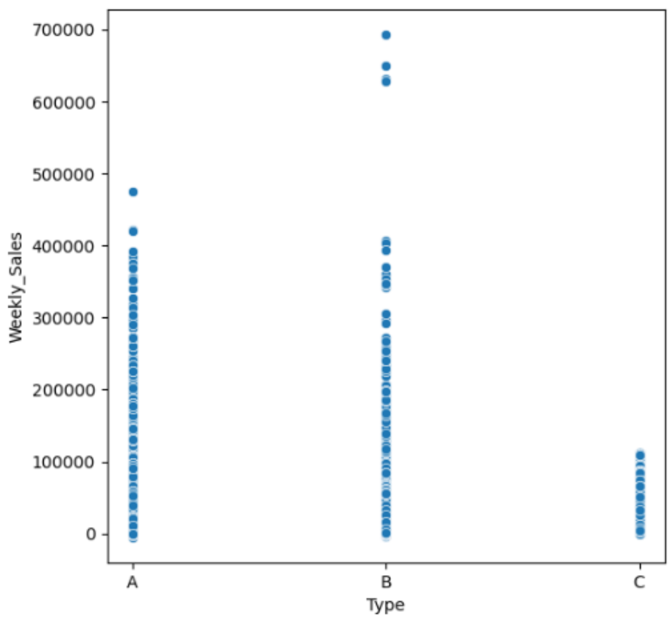
```
  Type  Store
0    A     22
1    B     17
2    C      6
```



- The pie chart above is used to show the type of stores (A, B, C) and how many percentages of stores belongs to the three types
- The Pie chart is imported from the matplotlib library

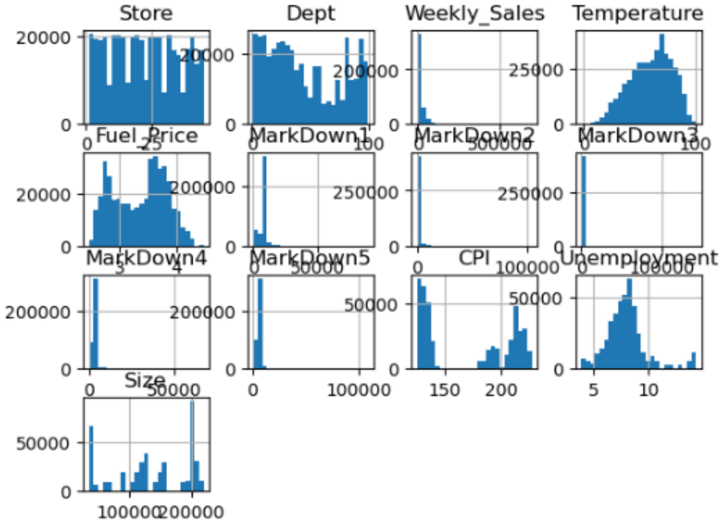
```
In [17]: #DISTRIBUTION OF WEEKLY SALES BASED ON STORE TYPE
plt.figure(figsize = (6,6))
sns.scatterplot(x ='Type',y = 'Weekly_Sales',data = df)

Out[17]: <Axes: xlabel='Type', ylabel='Weekly_Sales'>
```

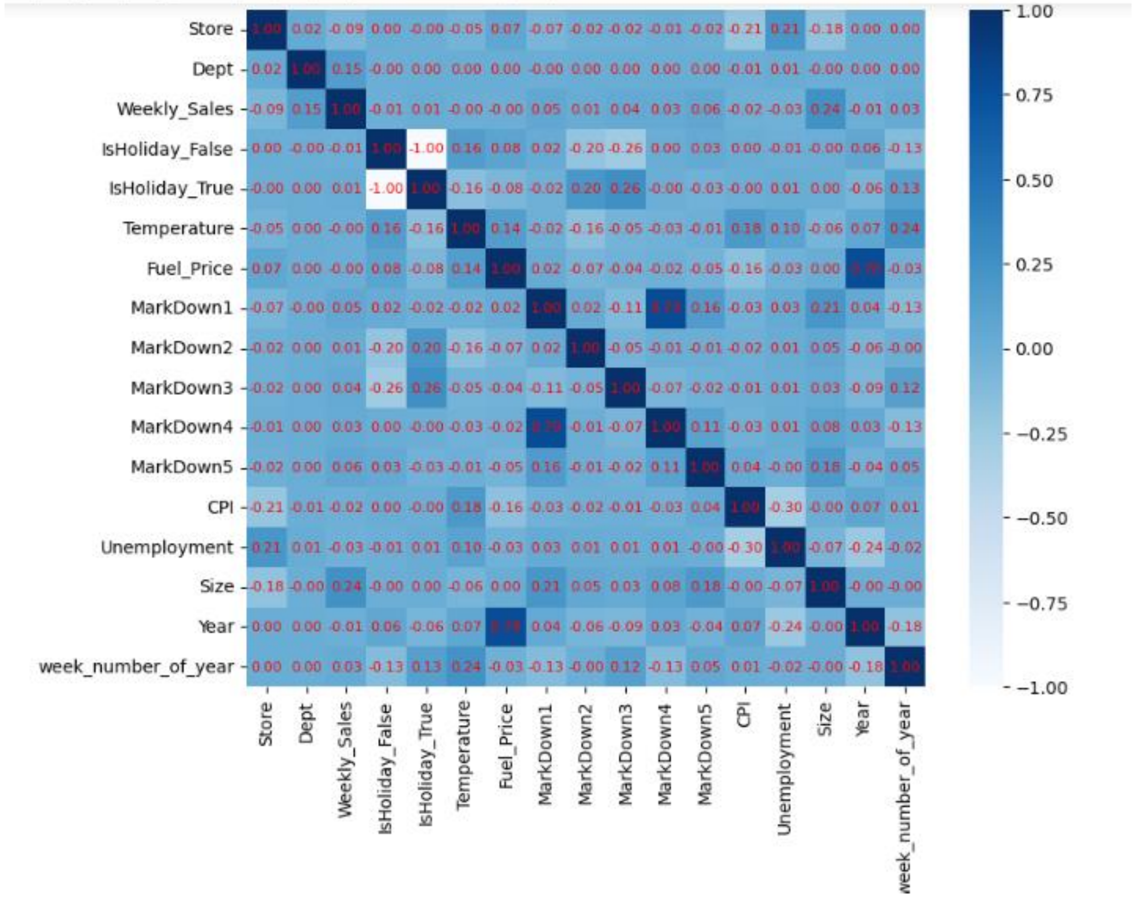


- The above graph is a scatter plot used to show the distribution of data-points according to type of store for weekly sales.
- The plot is imported from seaborn library

```
[<Axes: title={'center': 'Fuel_Price'}>,
 <Axes: title={'center': 'MarkDown1'}>,
 <Axes: title={'center': 'MarkDown2'}>,
 <Axes: title={'center': 'MarkDown3'}>],
[<Axes: title={'center': 'MarkDown4'}>,
 <Axes: title={'center': 'MarkDown5'}>,
 <Axes: title={'center': 'CPI'}>,
 <Axes: title={'center': 'Unemployment'}>],
[<Axes: title={'center': 'Size'}>, <Axes: >, <Axes: >, <Axes: >]],
dtype=object)
```



- The figure above is a histogram showing distribution of data-points of each colu mn over the mean line



- The heatmap displays a correlation matrix which shows correlation coefficient be tween dependent and the 1 independent variable.

8. **CONCLUSION**:

Predicted the weekly sales of the store after giving the attributes as input