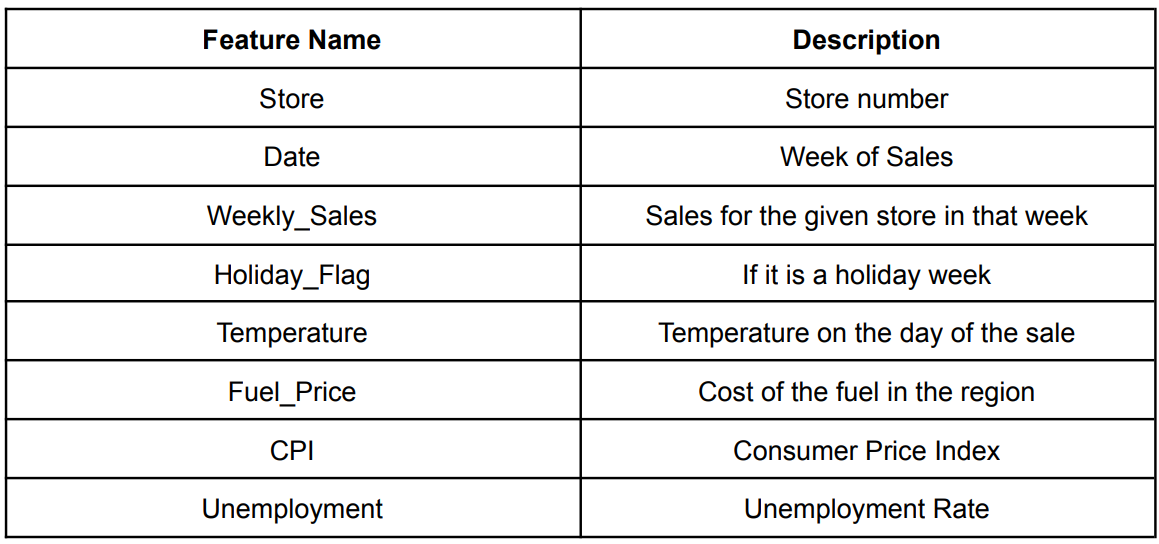
WALMART SALES ANALYSIS

**Problem Statement**

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. Dataset Information: The walmart.csv contains 6435 rows and 8 columns.

**Data Description**



**Project Objective**

**Data Pre-processing Steps and Inspiration**

Load the "walmart.csv" dataset.

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import warnings

warnings.filterwarnings("ignore")

dt=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

.head()

Used to show first 5 rows by default

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.head())

.tail()

Used to show rows by descending order

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.tail())

.shape

Shows the row and column of the dataset.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

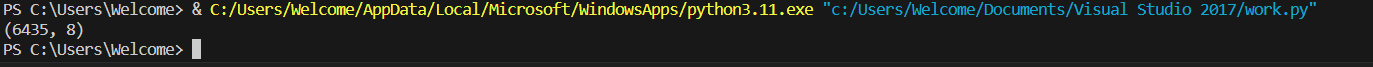
import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.shape)



.columns

Shows each number of columns in the dataset

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

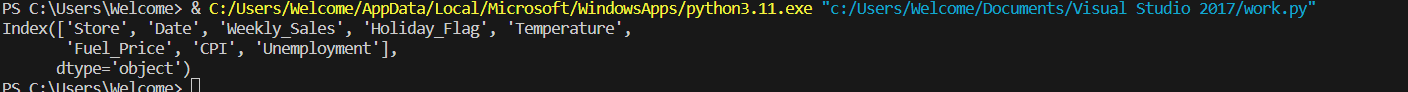
import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.columns)



.dtypes

Data type of each column is displayed by this attribute

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

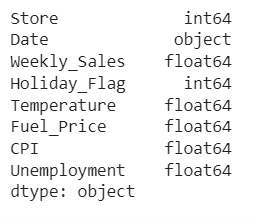
import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.dtypes)



.nunique

Shows unique values of each column

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

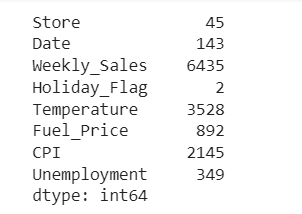
import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.nunique())

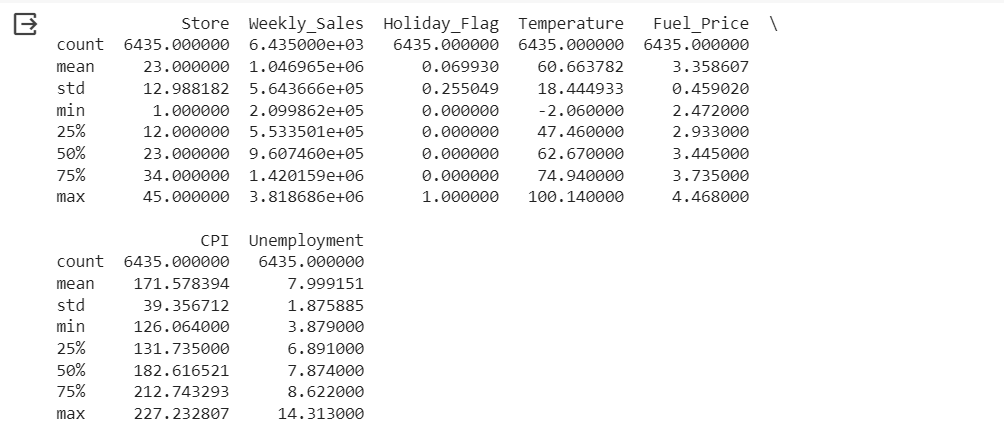


.describe()

It shows the count, mean, median and etc of the dataset

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.describe())

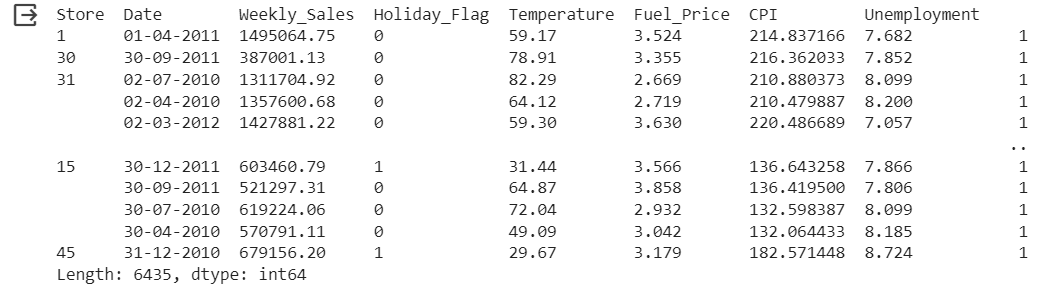


.value\_counts

Shows the unique values with their count.

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.value\_counts())



Checking missing values and null:

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import warnings

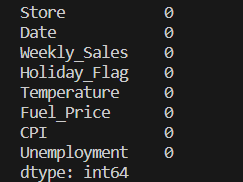
warnings.filterwarnings("ignore")

dt=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(dt.isna().sum())

print(dt.isnull().sum())

Output:



Data exploration and visualization

* Analyzing the relationship between weekly sales and unemployment rate.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

print(df.head())

x=df['Unemployment']

y=df['Weekly\_Sales']

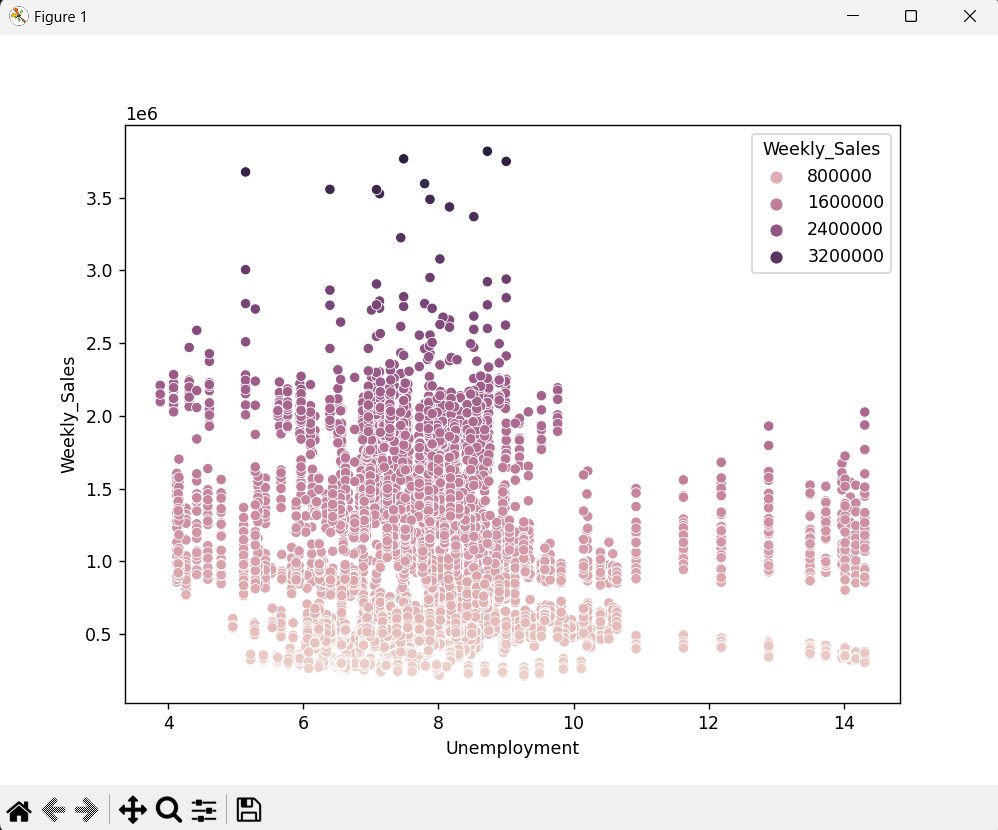
plt.figure(figsize=(8,6))

sns.scatterplot(data=df,x=x,y=y,hue="Weekly\_Sales")

plt.xlabel("Unemployment")

plt.ylabel("Weekly\_Sales")

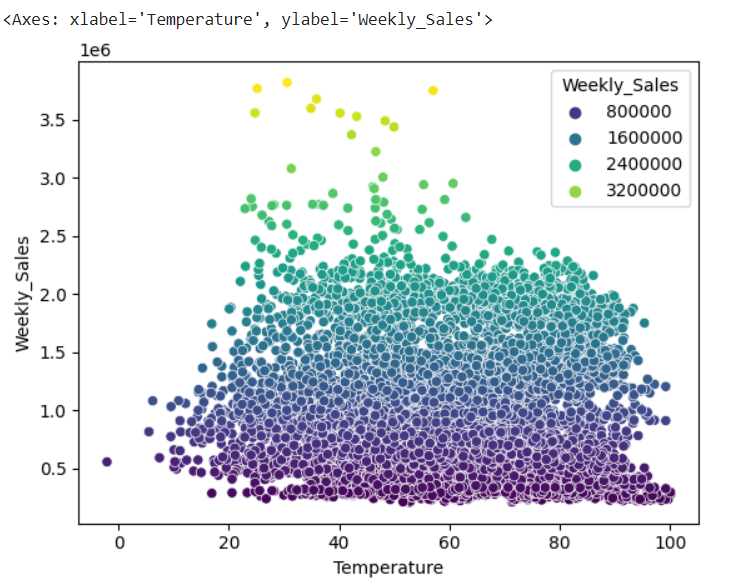
plt.show()



From the plot the sales gradually decreased as the unemployment rate increase.

* Analyzing the relationship between weekly sales and temperature.

sns.scatterplot(data=dt,x=x,y=y,hue='Weekly\_Sales',palette="viridis")



* Analyzing the effect on weekly sales due to CPI.

Seasonal Trend

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df=pd.read\_csv(r'C:\Users\Welcome\OneDrive\Desktop\DS\CAPSTONE PROJECT\WALMART\drive-download-20230709T122717Z-001-1\Walmart DataSet\Walmart DataSet.csv')

df['Date'] = pd.to\_datetime(df['Date'],format='%d-%m-%Y')

df.set\_index('Date', inplace=True)

weekly\_sales = df['Weekly\_Sales'].resample('W').mean()

plt.figure(figsize=(12, 6))

plt.plot(weekly\_sales, color='b', label='Weekly Sales')

plt.xlabel('Date')

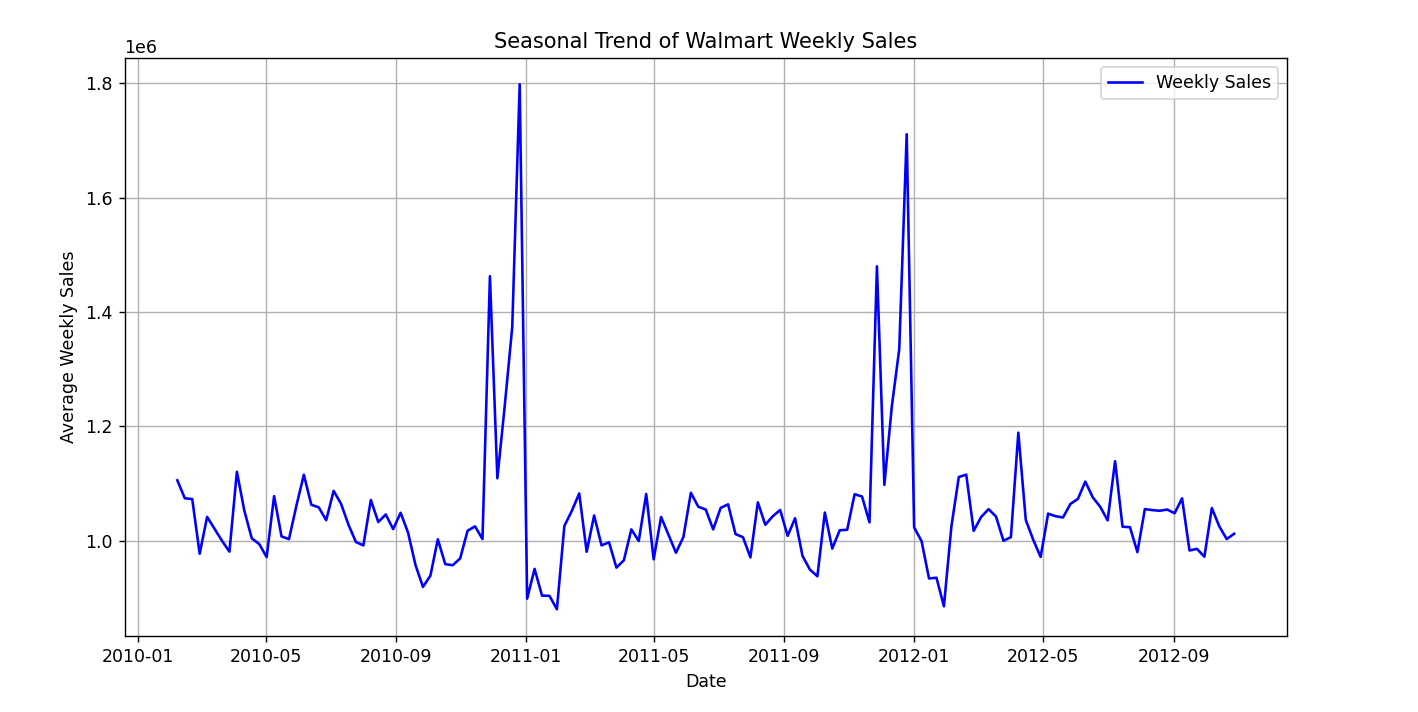
plt.ylabel('Average Weekly Sales')

plt.title('Seasonal Trend of Walmart Weekly Sales')

plt.legend()

plt.grid(True)

plt.show()



From the data visualization it is observed that the sales are high on end of the each year that is December.

**Using ARIMA MODEL:**

Using the ARIMA model to predict the sales based on the past seasonal trend.

**Augmented Dickey-Fuller test :**

df = pd.read\_csv('/content/Walmart DataSet.csv')

df['Date'] = pd.to\_datetime(df['Date'],format='%d-%m-%Y')

df.set\_index('Date', inplace=True)

weekly\_sales = df['Weekly\_Sales'].resample('W').sum()

result = adfuller(weekly\_sales)

print(f'ADF Statistic: {result[0]}')

print(f'p-value: {result[1]}')

if result[1] <= 0.05:

    print("Stationary (Rejecting the null hypothesis)")

else:

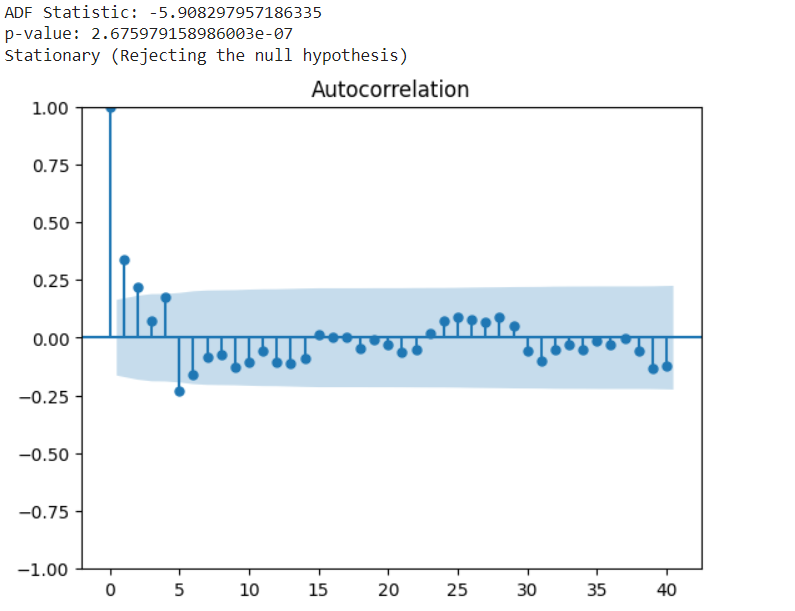
    print("Non-Stationary (Fail to reject the null hypothesis)")

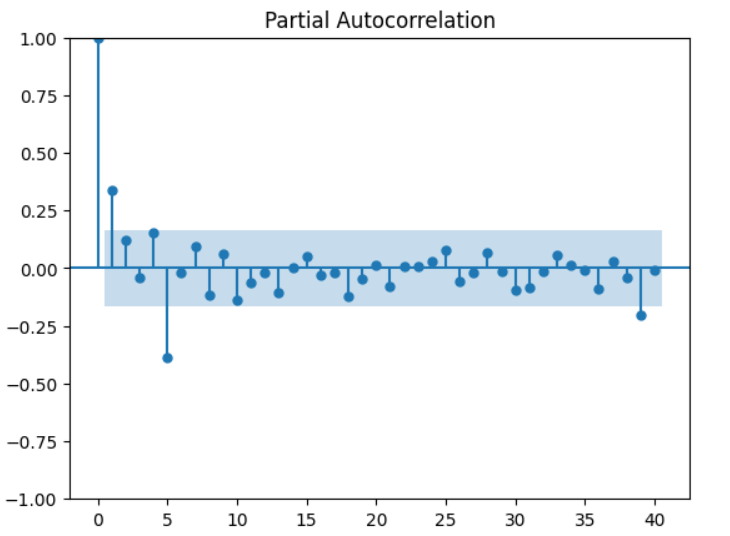
plot\_acf(weekly\_sales, lags=40)

plot\_pacf(weekly\_sales, lags=40)

plt.show()

AUTOCORRELATION:





**ARIMA Model:**

p, d, q = 1, 1, 1

model = ARIMA(weekly\_sales, order=(p, d, q))

results = model.fit()

forecast\_steps = 12

forecast = results.forecast(steps=forecast\_steps)

plt.figure(figsize=(12, 6))

plt.plot(weekly\_sales, label='Historical Sales')

plt.plot(forecast, label='Forecasted Sales', color='red')

plt.xlabel('Date')

plt.ylabel('Weekly Sales')

plt.title('Historical and Forecasted Sales')

plt.legend()

plt.show()

**Conclusion**

The aim of this study was to predict Walmart’s sales based on the available historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores . Sales are relatively higher during holidays than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue. As observed through the exploratory data analysis, store size and holidays have a direct relationship with high Walmart sales. It was also observed that out of all the store types, Type A stores gathered the most sales for Walmart. Pertaining to the specific factors provided in the study (temperature, unemployment, CPI, and fuel price), it was observed that sales do tend to go up slightly during favourable climate conditions as well as when the prices of fuel are adequate. By the observations in the exploratory data analysis, sales also tend to be relatively higher when the unemployment level is lower. Additionally, with the dataset provided for this study, there does not seem to be a relationship between sales and the CPI index. While the interaction effects were tested on a combination of significant variables, a statistically significant relationship was only observed between the independent variables of temperature, CPI and unemployment, and weekly sales (predictor variable). Relationships between independent and target variables were tried to be identified through EDA components like the correlation matrix and scatter plots.