Capstone Project

Table of Contents

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. This project undertakes to review the sales records from stores to provide useful insights using the data and make prediction models to forecast the sales for next 12 weeks.

Problem Objective

The retail store with multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. This task is to come up with useful insights using the data and make prediction models to forecast the sales for next 12 weeks.

Data Description

The available dataset contains 6435 records and 8 features.

Feature Name	**Description**
Store	Store number
Date	Week of Sales
Weekly_Sales	Sales for the given store in that week
Holiday_Flag	If it is a holiday week
Temperature	Temperature on the day of the sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

Data description, various insights from the data.

From the given dataset of the company, it is observed that the data consists of six thousand four hundred and thirty-five (6,435) records with seven features as follows:

1. Stores: there are 45 stores and each store has 143 recorded entries of:
   a. Date of record(weekly),
   b. Total sales record for the week,
   c. Holiday flag for the week (1 or 0)
   d. Temperature: average temperature recorded during the week,
   e. Fuel price: average fuel price for the week,
   f. CPI: average Consumer price index for the week,
   g. Unemployment: rate of unemployment for the week of record

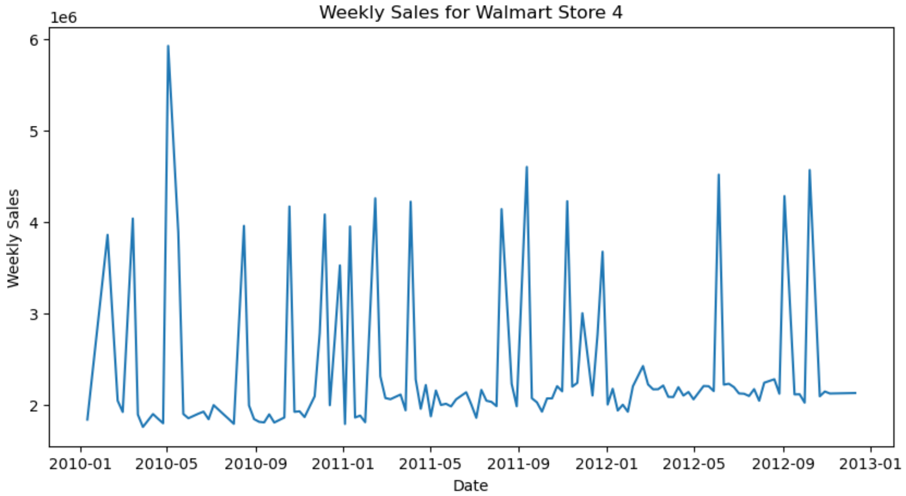
Data preprocessing Steps and Inspiration

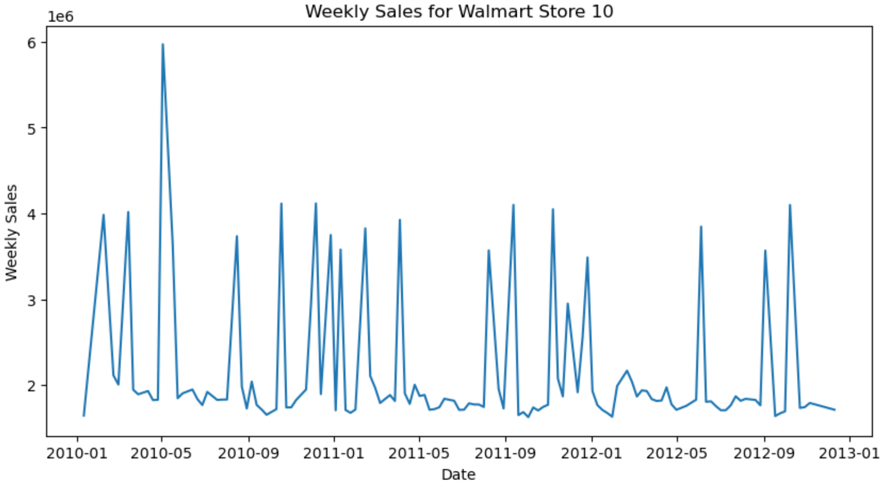
The preprocessing of the data included the following steps:

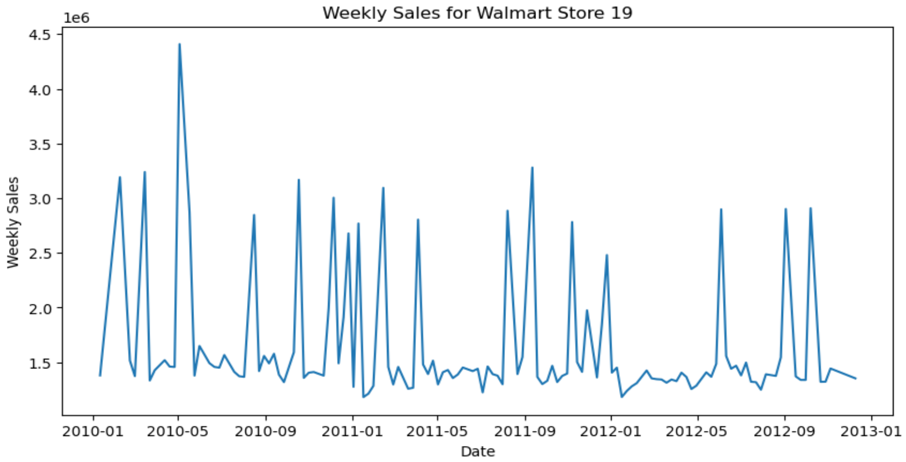
1. Step.1: Load data
2. Step.2: Perform Exploratory Data Analysis
   a. Confirm number of records in the data how they are distributed
   b. Check data types
   c. Check for missing data, invalid entries, duplicates
   d. Examine the correlation of the independent features with the target (Weekly_sales) variable.
   e. Check for outliers that are known to distort predictions and forecasts
3. Step.3: Model Predictions, two approaches:
   a. Time series model (ARIMA)
   b. Linear regression model
4. Step.4: Forecast
5. Step.5: Compare results from different models

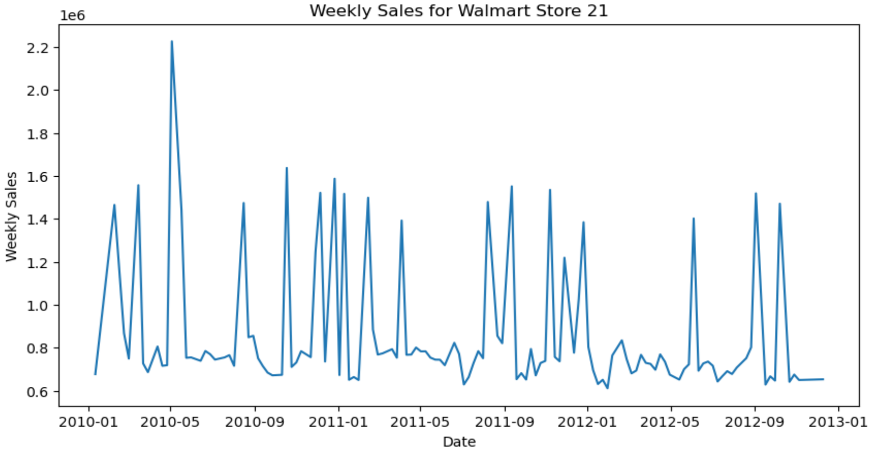
Model Evaluation and Technique

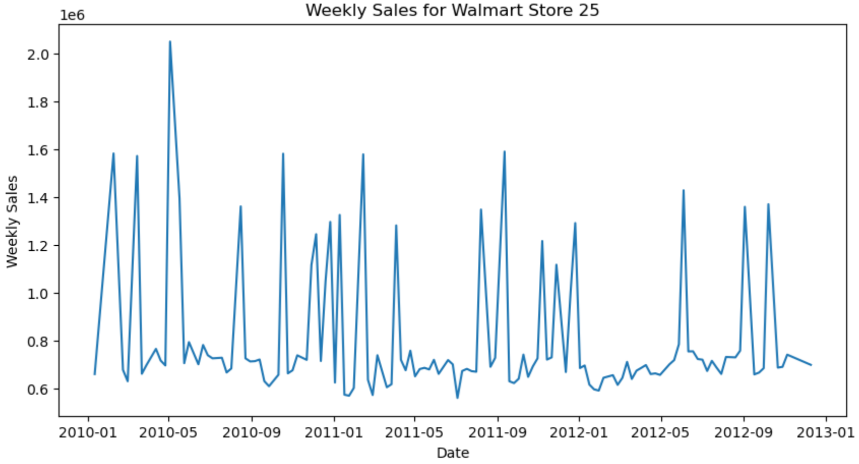
Plot of weekly sales against time for selected stores

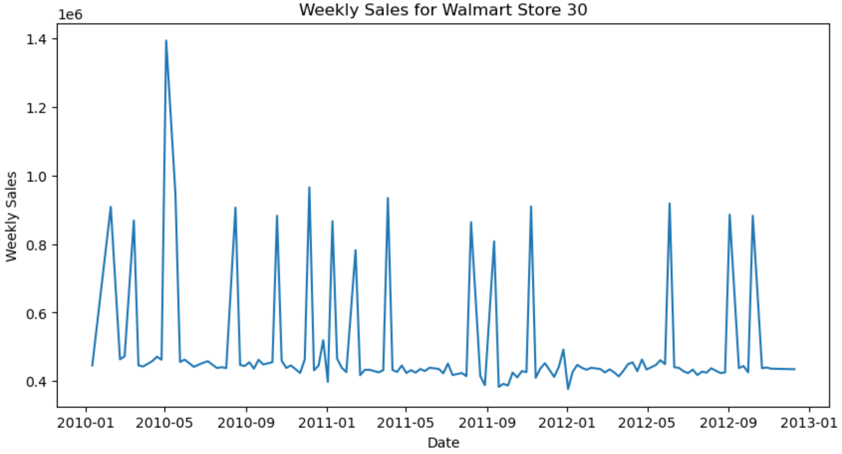


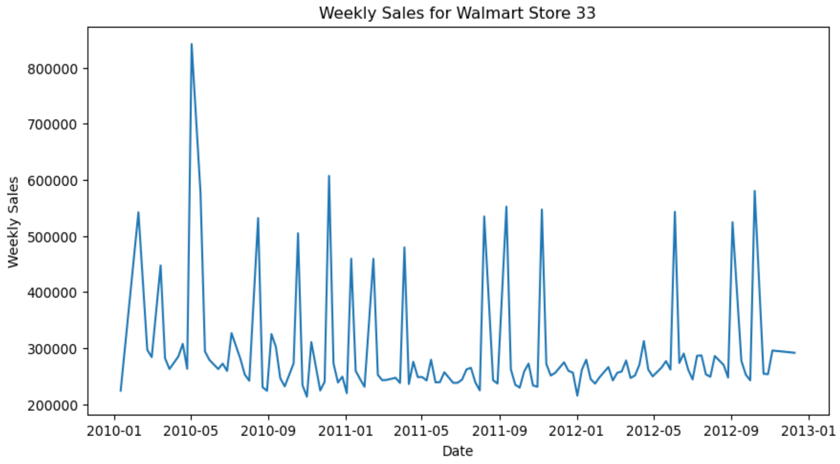


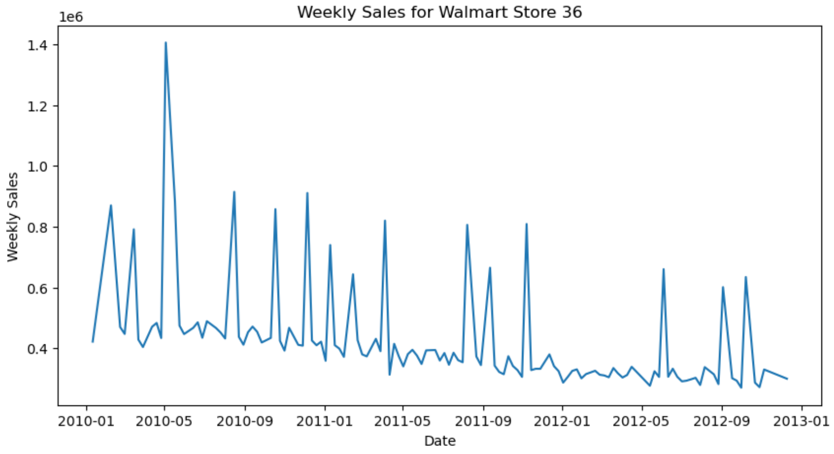

Weekly Sales for Walmart Store 19


Weekly Sales for Walmart Store 21


Weekly Sales for Walmart Store 25


Weekly Sales for Walmart Store 30





Model Selection:

Examination of the plot of the target feature, weekly_Sales (as shown above) shows a continuously time varying data.

A time series model (ARIMA, SERIMA, SERIMAX) will be employed for the predictions and forecast. Attempt will also be made to use Linear regression models (gradient_boosting, Linear regression, Random Forest) for prediction and compare the results with the time series predictions

1. The ARIMA model:
   Autoregressive Integrated Moving Average (ARIMA) is defined as a statistical analysis model that use time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values.

   ARIMA model is based on a number of assumptions including:
      1. Data does not contain anomalies,
      2. Model parameters and error term is constant,
      3. Historic timepoints dictate behaviour of present timepoints,
      4. Time series is stationary
2. Regression Models:
   a. Gradient_boosting: Gradient boosting stands out for its prediction speed and accuracy, particularly with large and complex dataset. In machine learning algorithm, two types of errors, otherwise called loss functions, are encountered bias error and variance error. Gradient boosting algorithm is based on minimizing the bias error or the loss function of the

model. The gradient boosting algorithm is based on building models sequentially where the subsequent models are built on the errors or residuals of the previous model. The process is repeated until there is no more significant change on the error.

b. Linear Regression is a basic predictive analytics technique that uses historical data to predict an output variable. It is a popular algorithm employed to predict continuous (dependent) variables such as price, based on their correlation with other independent variables. It is based on the following assumptions:
   i. Linear Relationship: The relationship between the independent and dependent variables should be linear.
   ii. Multivariate Normal: All the variables together should be multivariate normal, which means that each variable separately bias to be univariate normal means, a bell-shaped curve,
   iii. No Multicollinearity: There is little or no multicollinearity in the data which means that the independent variables should not have minimal correlation with each other.

iv. No Autocorrelation: There is a little or no autocorrelation in the data where the data values of the same column are related to each other.

v. Homoscedasticity: There should be homoscedasticity or “same variance” across regression lines. In other words, residuals are equal across regression line.

c. Random Forest: Random forest is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption as it handles both classification and regression problems.

Model Evaluation:

The following techniques and steps were involved in the evolution of model

a. load necessary libraries
b. load the data set
c. perform exploratory data analysis (EDA) on the data set
    i. find the shape are size of the data

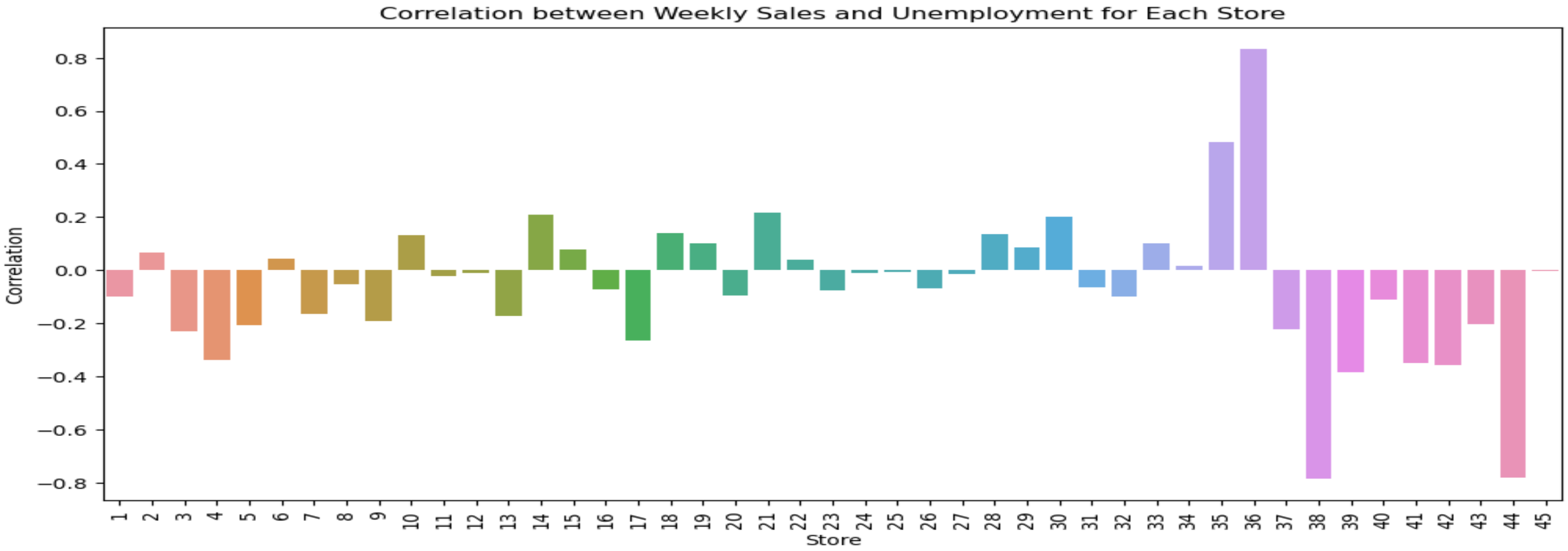
    ii. check for invalid and null entries
    iii. explore data description
    iv. examine the correlations of the independent variable to the target variable (weekly sales)
    v. line plot of the effects of the independent variables on the target variable
    vi. box plot of the features to identify outliers

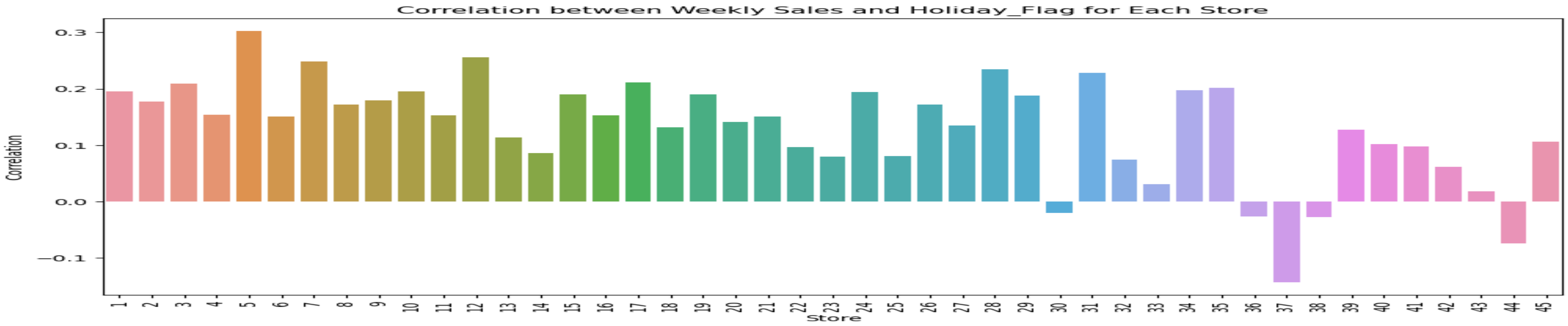
d. model prediction

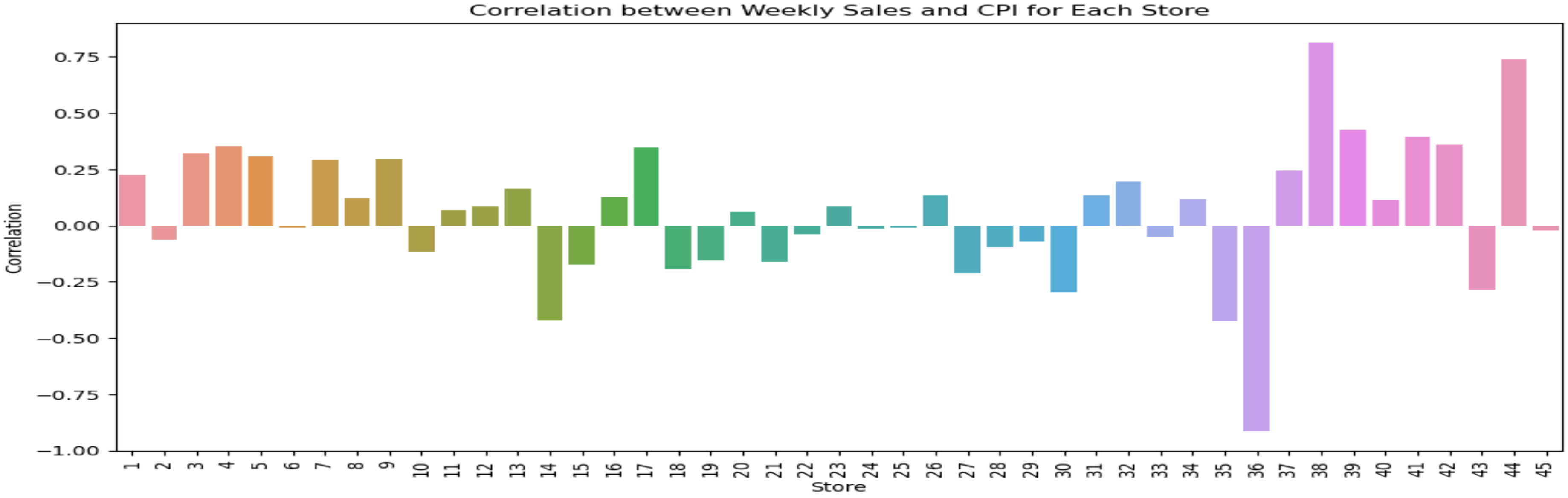
e. forecast

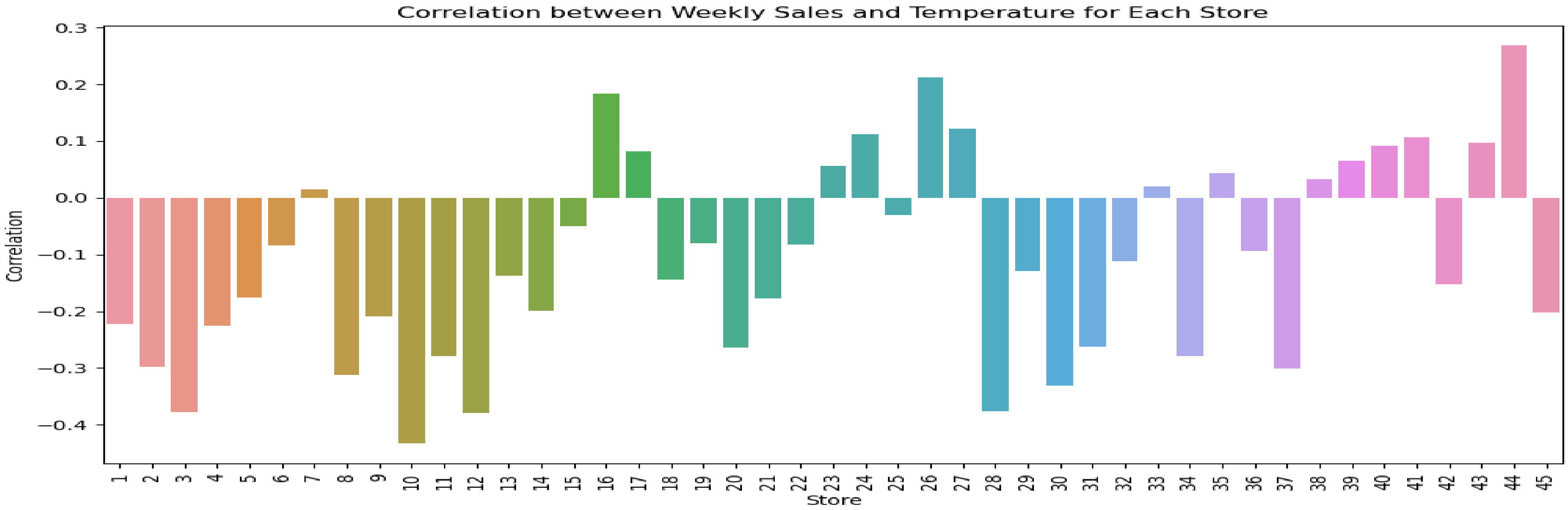
Model Evaluation and technique

Model design:


Correlation between Weekly Sales and Unemployment for Each Store
Correlation
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
Store


Correlation between Weekly Sales and Holiday_Flag for Each Store
Correlation
0.3
0.2
0.1
0.0
−0.1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
Store


Correlation between Weekly Sales and CPI for Each Store
Correlation
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
Store



It was observed from EDA that the effects of the independent features (Unemployment, Temperature, Holiday Flag and CPI) on the target variable weekly sales differ greatly the store. For example, as shown above, the effects of unemployment vary by stores whereas it appears to have positive effects on some and negative effects on others. The same is also true for temperature, CPI and to some extent, the Holiday flag.

Premised on the findings, the decision was taken to handle the model predictions by the stores as a single prediction for all the stores may not be reasonable given the peculiar conditions prevalent in each region of the stores.

Model Approach:

1. TS Model, ARIMA.
    - The first step for this model is to check the stationarity of the dataset (p-value less than 0.05)
    - Next is to find the best ARIMA order for the model
    - Using the best ARIMA order, make predictions for the selected stores.
        - estimate a 12 – weeks weekly sales forecast
2. Regression Models:
    - Regression model, Linear Regression and Random_Forescast models were also used for the prediction. The best of the three predictions will then be compared to the predictions by ARIMA model predictions.

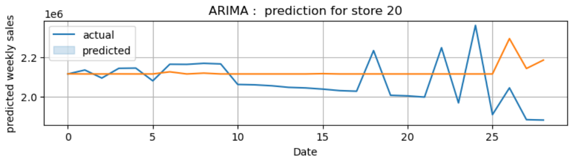
Inferences from the project

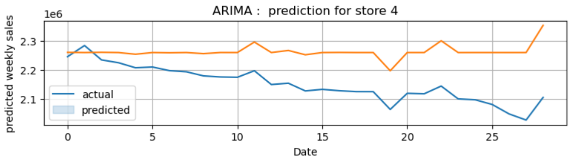
Model Results:

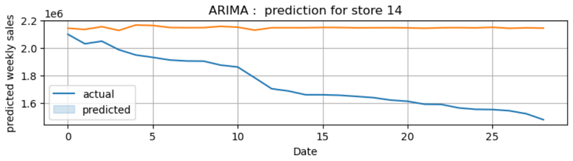
1. ARIMA model:

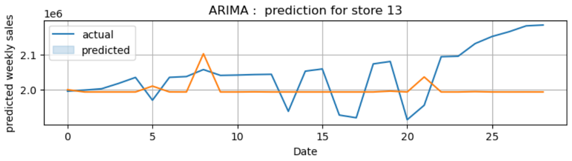
a. Predictions: Predictions were performed for six stores (stores20,4,13,2,10 and 27 in order of decreasing Weekly_Sales sales revenue). The predictions results are summarized in the Table and graphs below:

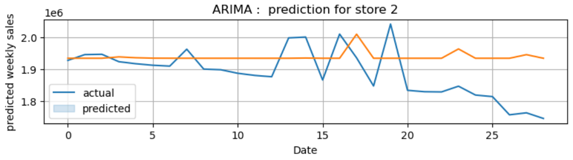
	Store_20	Store_4	Store_13	Store_2	Store_10	Store_27
Median error (%)	3.42	5.83	2.90	3.28	9.35	5.14
Mean error (%)	4.80	5.37	3.57	3.94	9.18	6.94

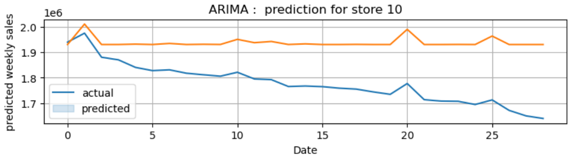

ARIMA : prediction for store 20
1e6
predicted weekly sales
2.2
2.0
actual
predicted
0
5
10
15
20
25
Date

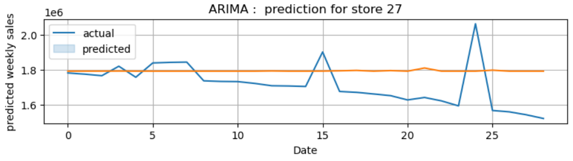

ARIMA : prediction for store 4
1e6
predicted weekly sales
2.3
2.2
2.1
actual
predicted
0
5
10
15
20
25
Date

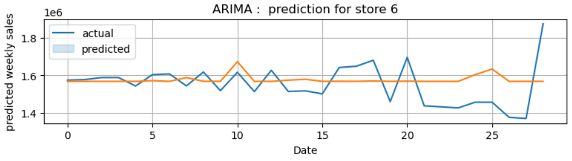

ARIMA : prediction for store 14
1e6
predicted weekly sales
2.2
2.0
1.8
1.6
actual
predicted
0
5
10
15
20
25
Date

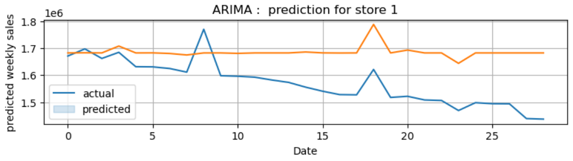

ARIMA : prediction for store 13
1e6
predicted weekly sales
2.1
2.0
actual
predicted
0
5
10
15
20
25
Date

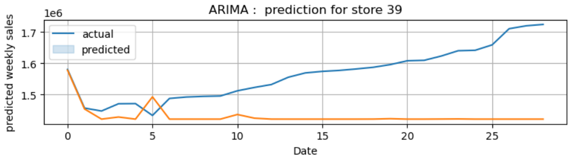

ARIMA : prediction for store 2
1e6
predicted weekly sales
2.0
1.9
1.8
actual
predicted
0
5
10
15
20
25
Date


ARIMA : prediction for store 10
1e6
predicted weekly sales
2.0
1.9
1.8
1.7
actual
predicted
0
5
10
15
20
25
Date


ARIMA : prediction for store 27
1e6
predicted weekly sales
2.0
1.8
1.6
actual
predicted
0
5
10
15
20
25
Date

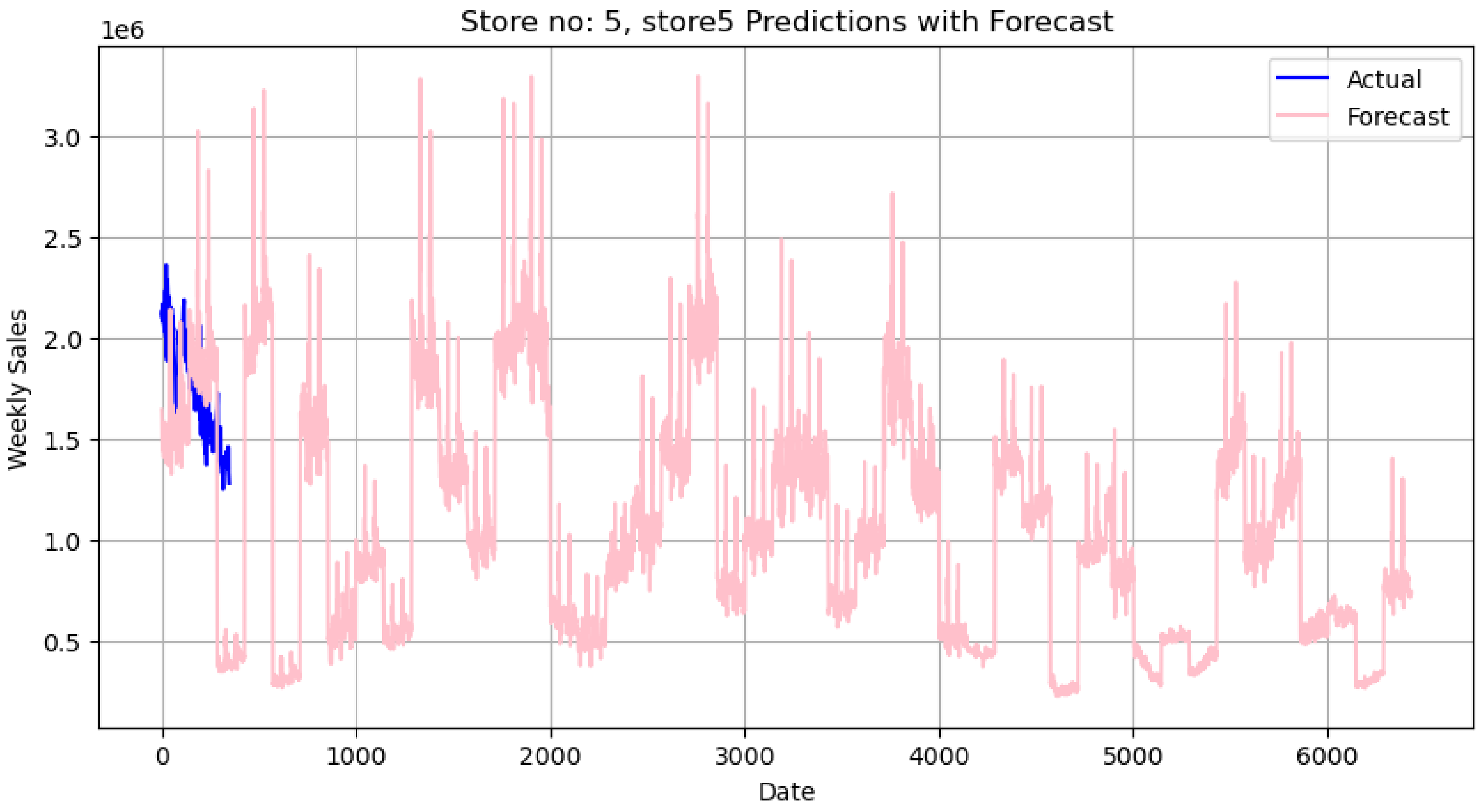

ARIMA : prediction for store 6
1e6
predicted weekly sales
1.8
1.6
1.4
actual
predicted
0
5
10
15
20
25
Date


ARIMA : prediction for store 1
1e6
predicted weekly sales
1.8
1.7
1.6
1.5
actual
predicted
0
5
10
15
20
25
Date


ARIMA : prediction for store 39
1e6
predicted weekly sales
1.7
1.6
1.5
actual
predicted
0
5
10
15
20
25
Date

b.Forecast:

The projected sales outlook for the next 12 weeks is it down for all the stores studied



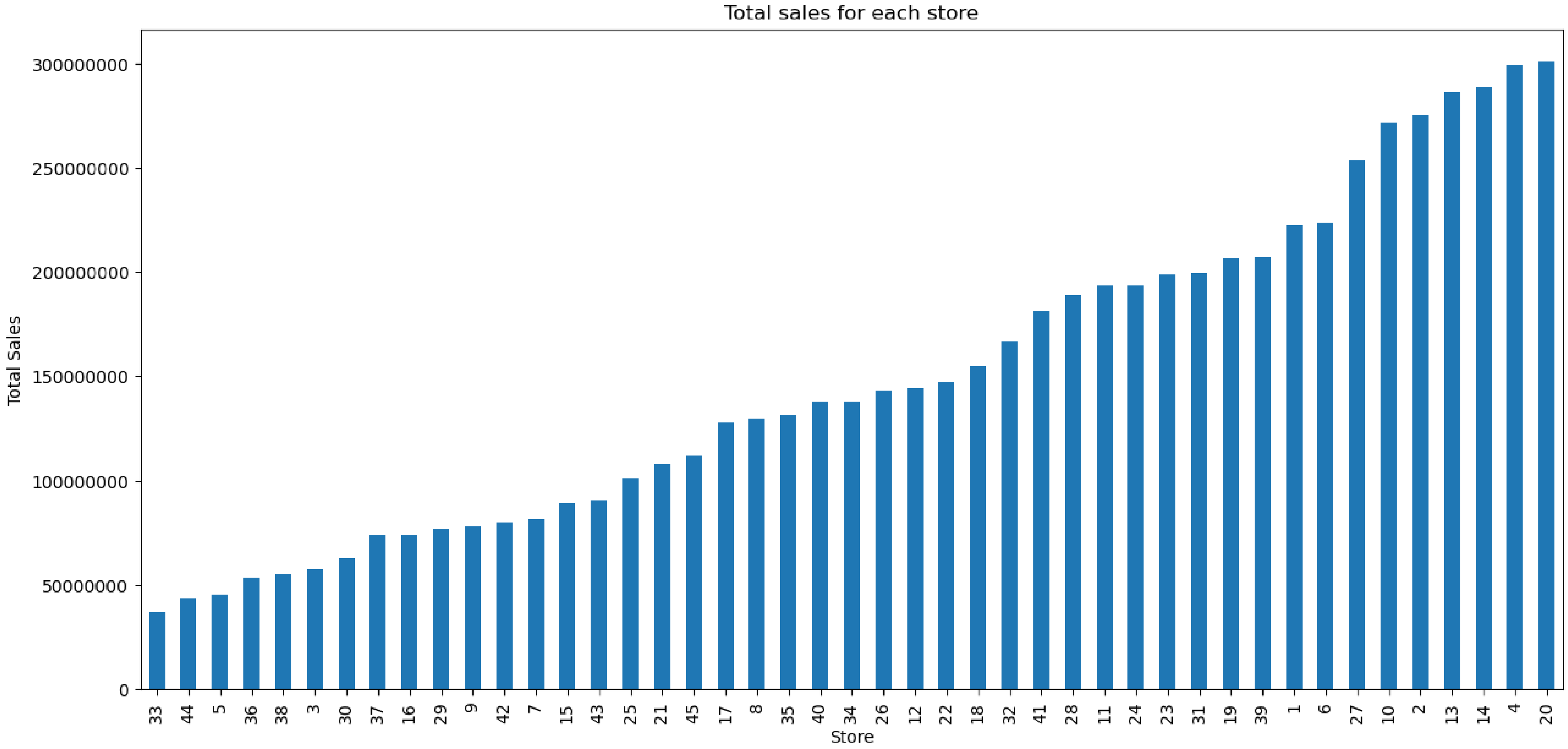
Inferences from the project

Model evolution:

ARIMA models

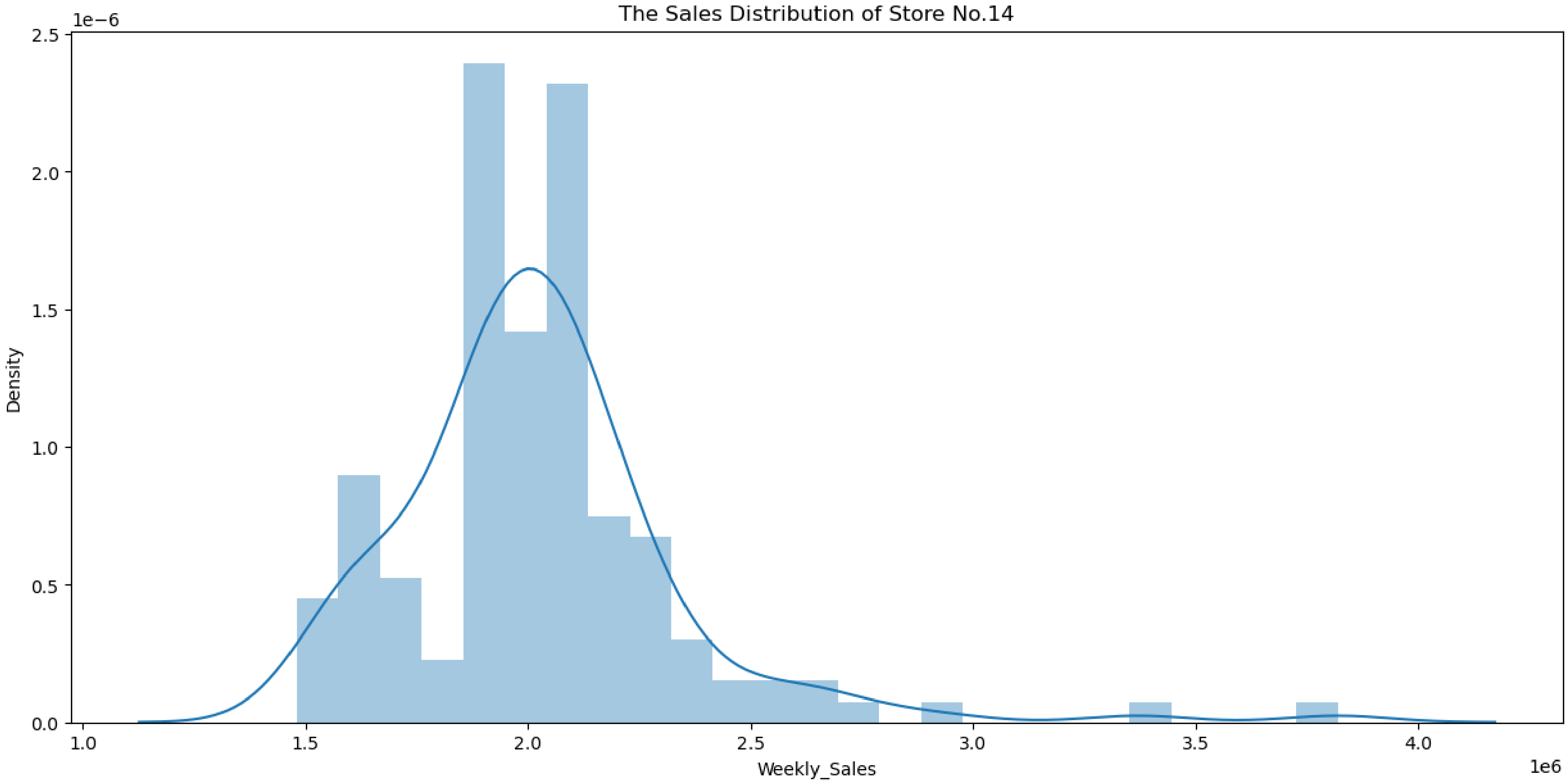
The model predictions for the selected stores were okay.

The regression models



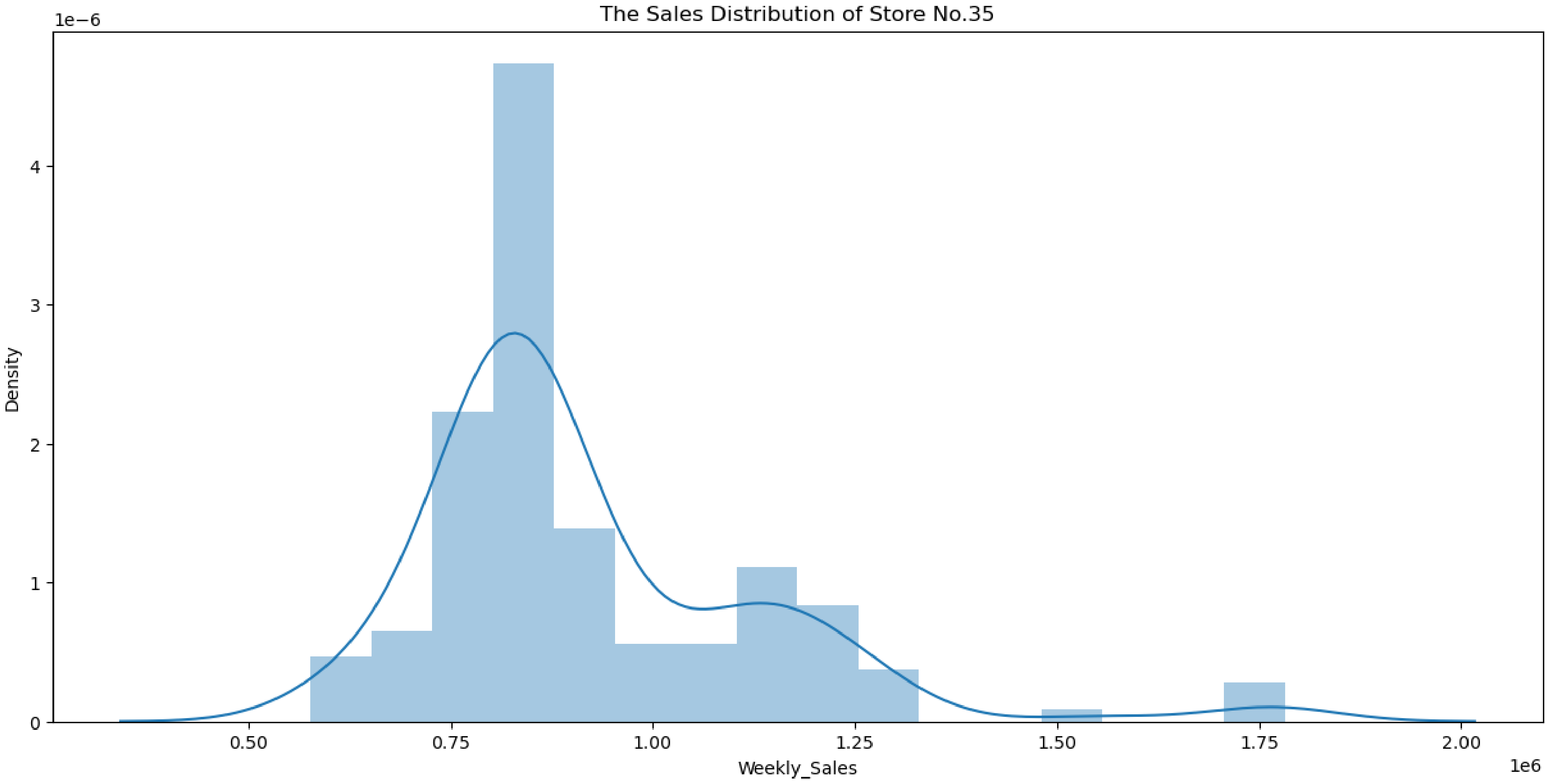
from the above graph , store number 20 has maximum weekly sales , store number 33 has minimum sales

The store 14 has maximum standard deviation, means sales vary alot from std

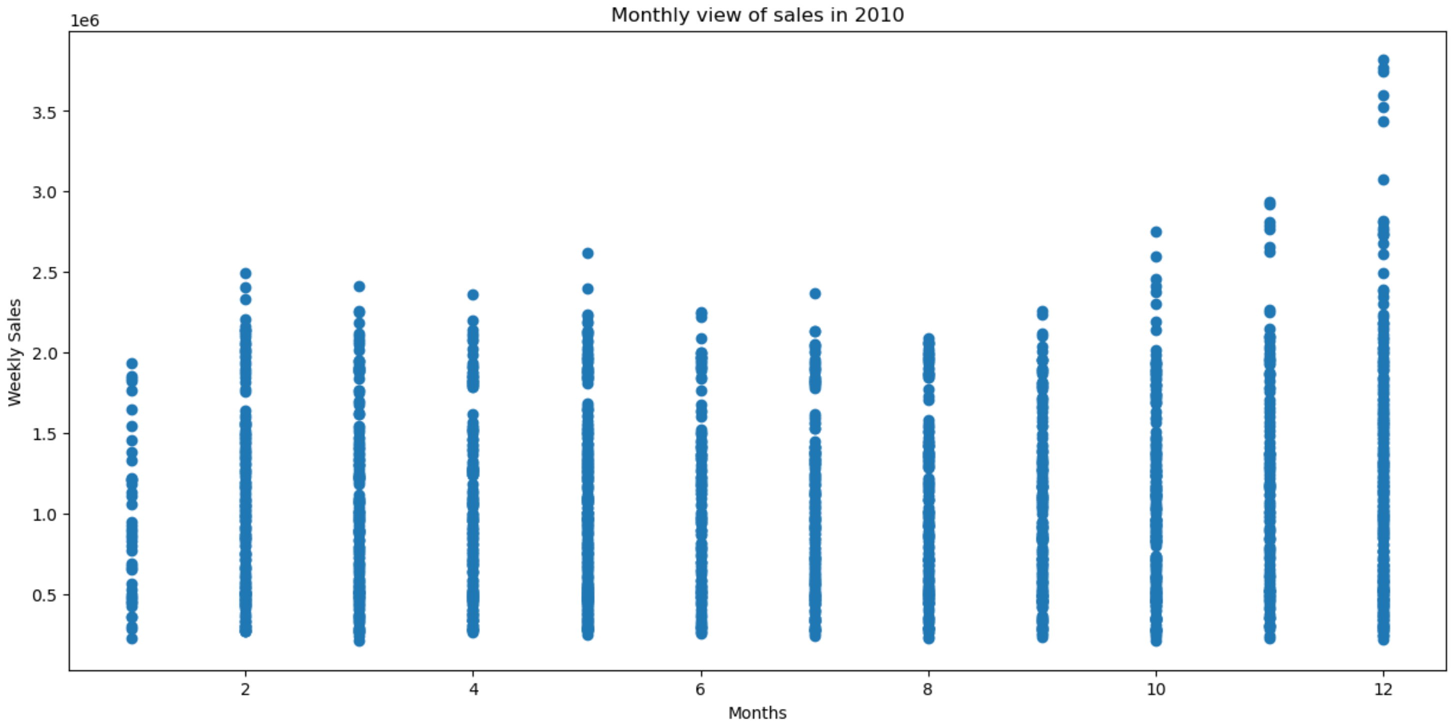


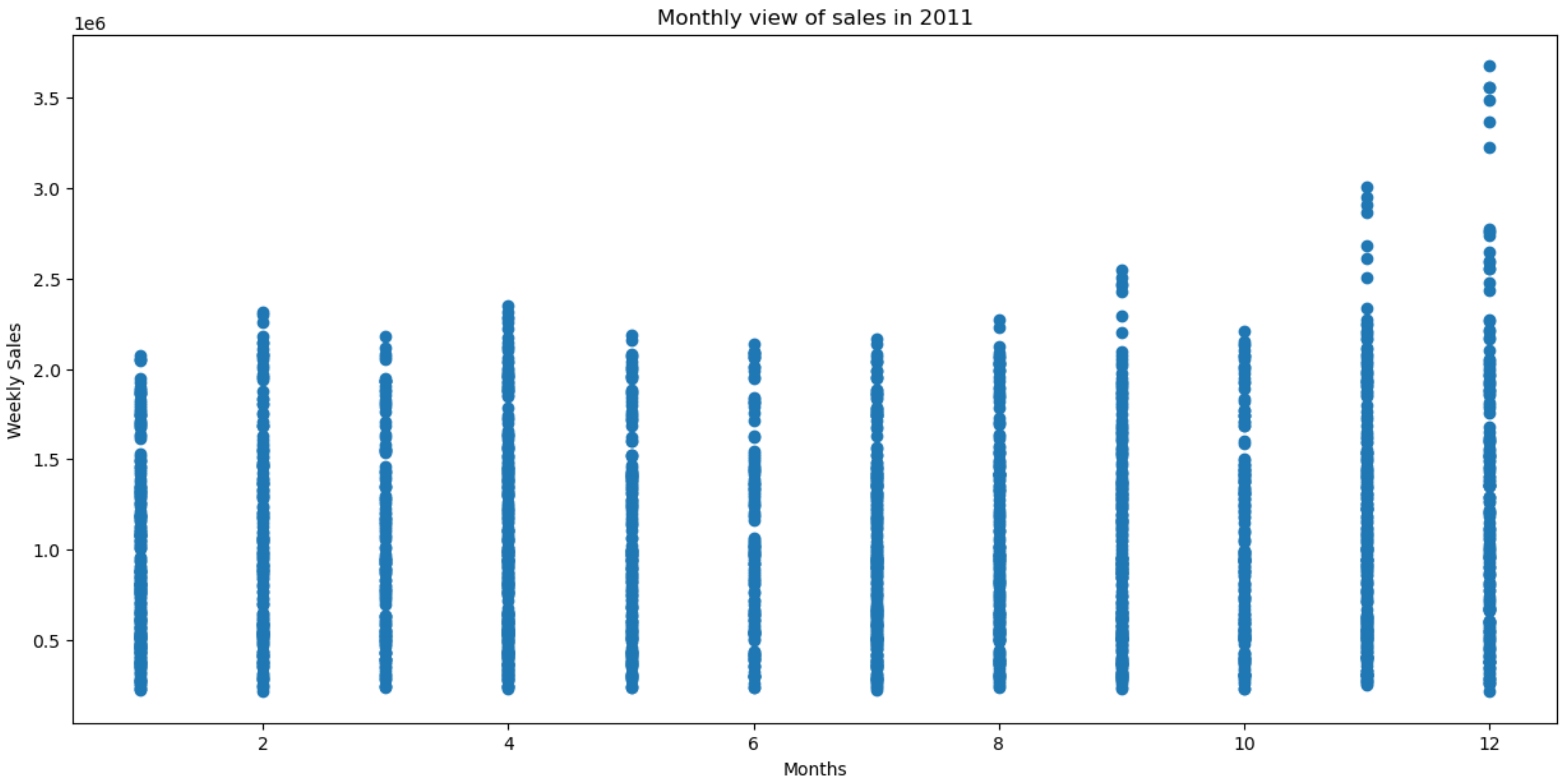
the store which has maximum coefficient of mean to standard deviation is store number 35

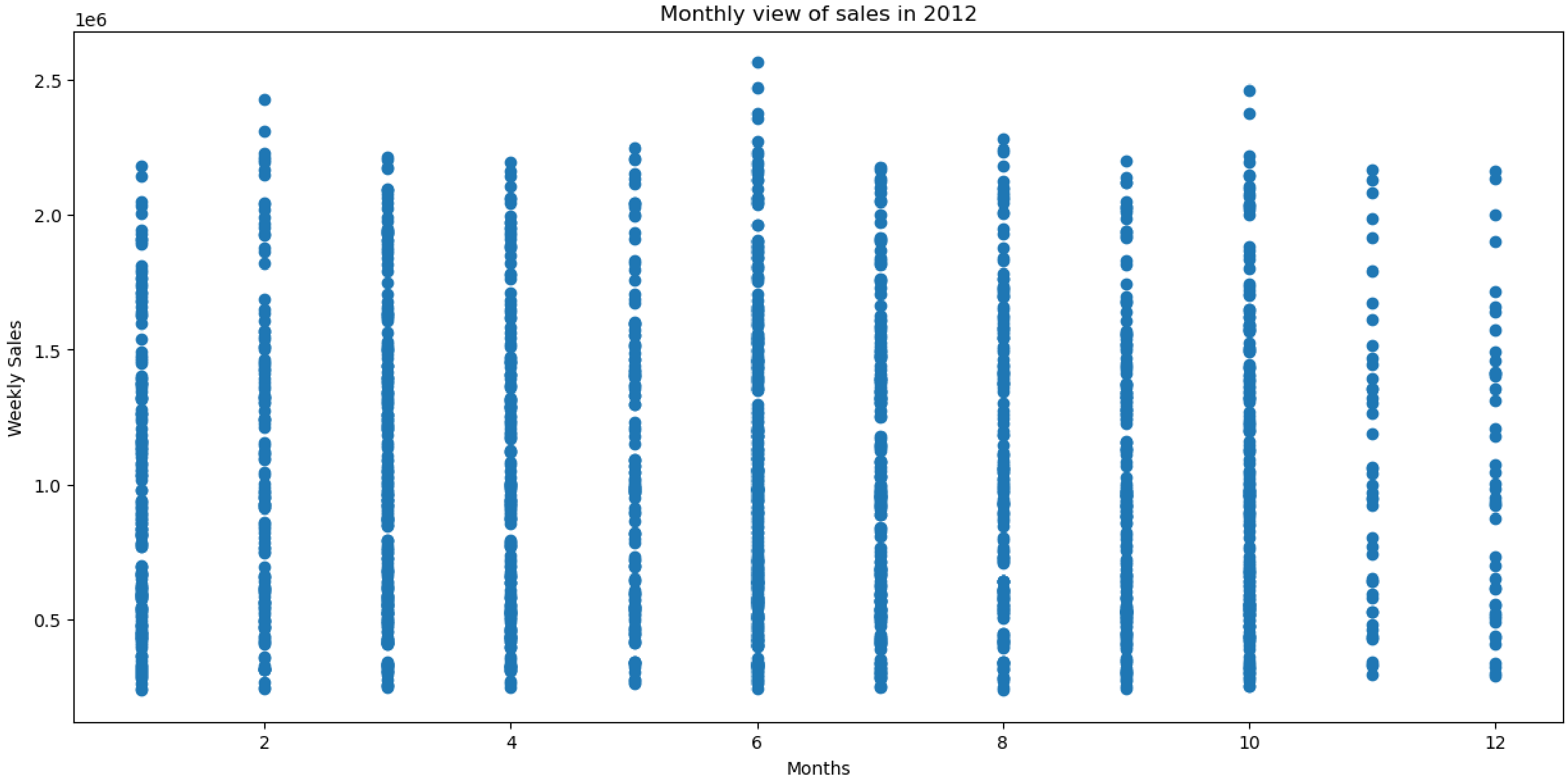
Distribution of store 35 has maximum coefficient of mean to standard deviation


The Sales Distribution of Store No.35
1e−6
Density
0
1
2
3
4
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e6
Weekly_Sales

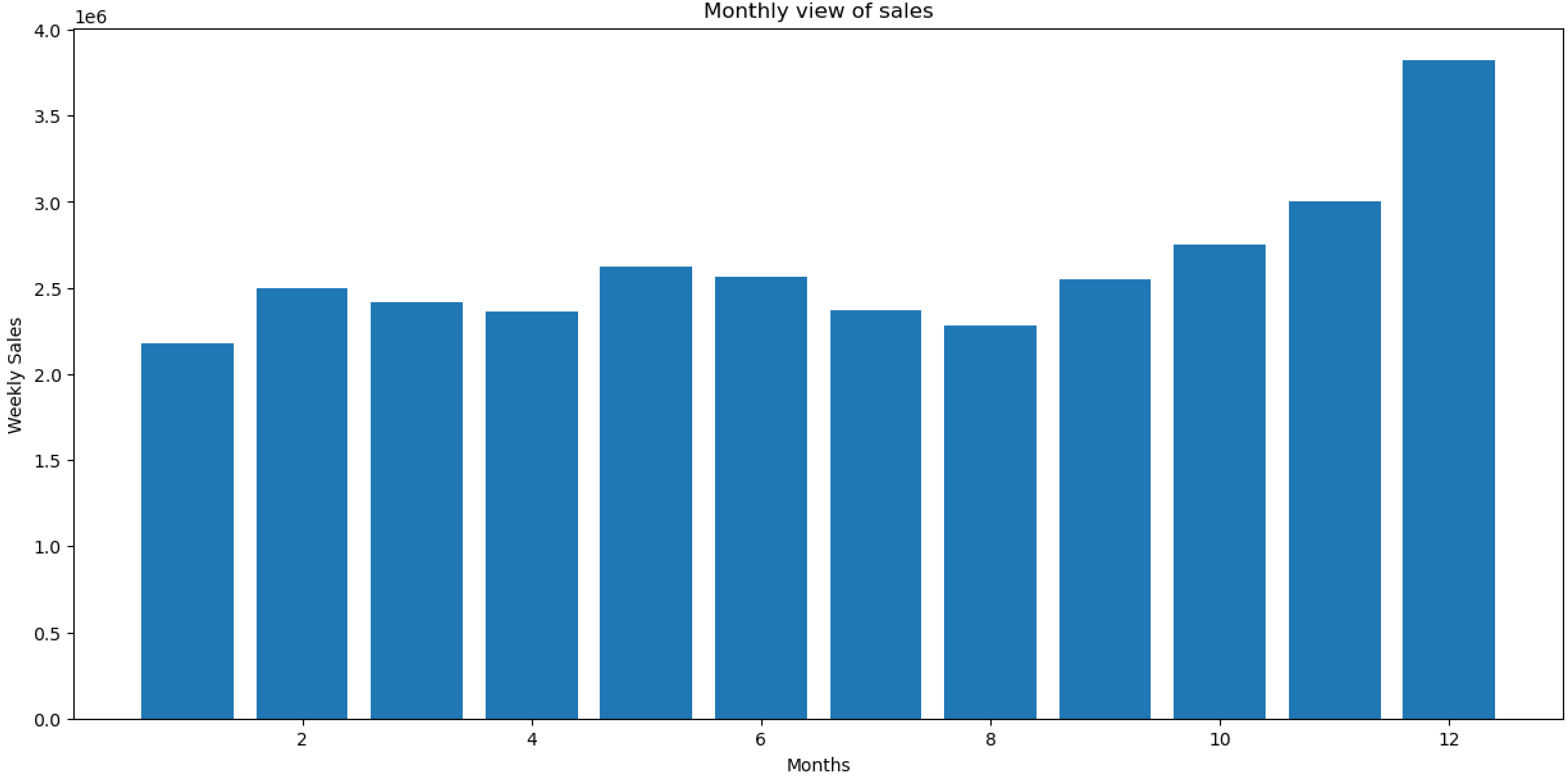
a monthly and semester view of sales in units and give insights:



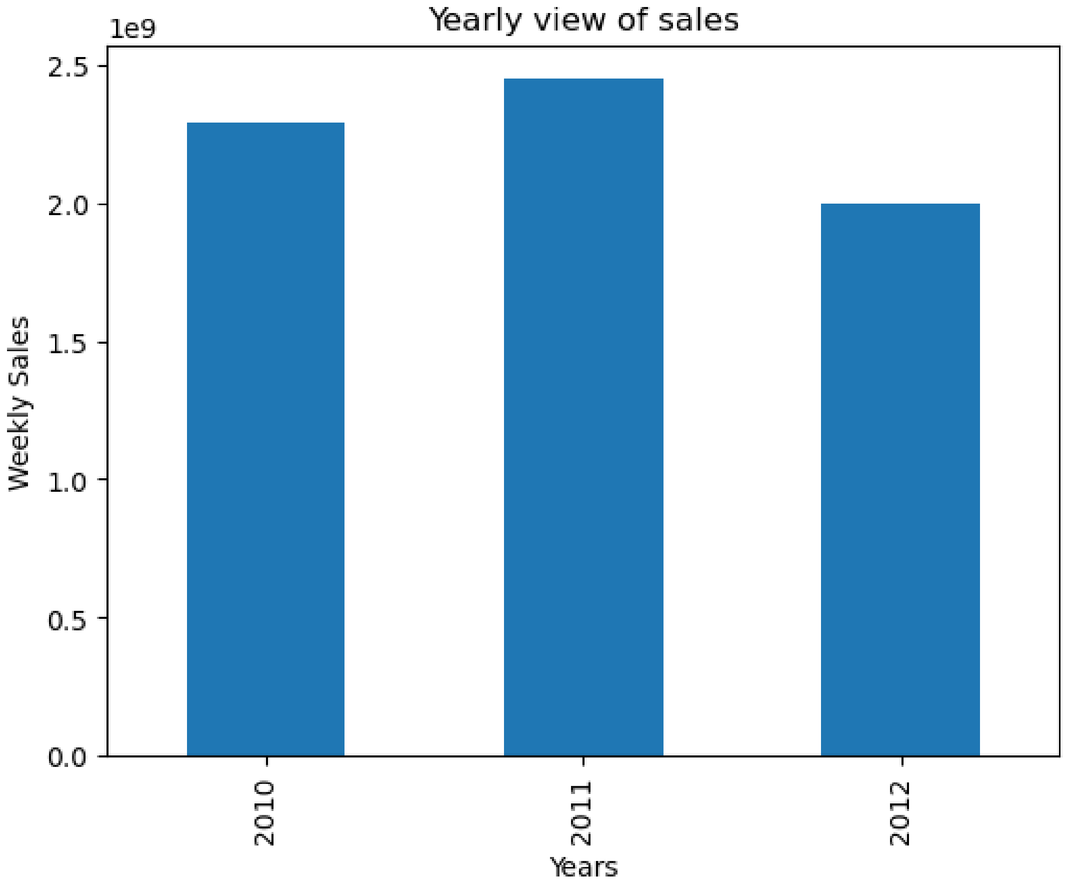

Monthly view of sales in 2011
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Weekly Sales
2
4
6
8
10
12
Months


Monthly view of sales in 2012
1e6
2.5
2.0
1.5
1.0
0.5
Weekly Sales
2
4
6
8
10
12
Months

Overall Monthly Sales



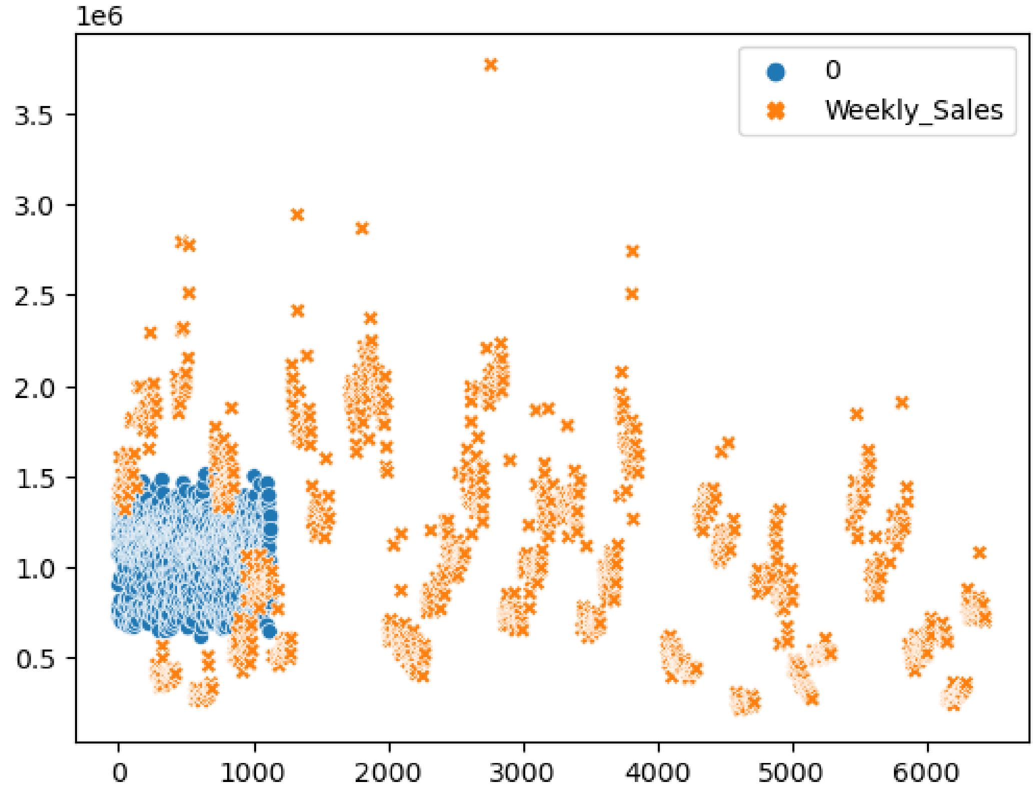
Yearly Sales



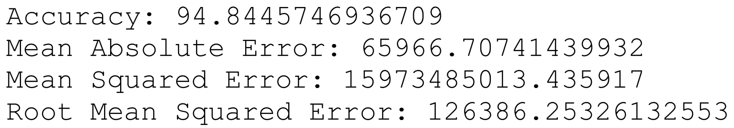
Here, overall monthly sales are higher in the month of December while the yearly sales in the year 2011 are the highest.

Linear Regression:

```
Accuracy: 12.16413503459849
Mean Absolute Error: 428333.41990933055
Mean Squared Error: 260710817219.49414
Root Mean Squared Error: 510598.4892452132
```



Random Forest Regressor:



```
Accuracy: 94.8445746936709
Mean Absolute Error: 65966.70741439932
Mean Squared Error: 15973485013.435917
Root Mean Squared Error: 126386.25326132553
```

Here, Linear Regression is not an appropriate model to use which is clear from its low accuracy. However, Random Forest Regression gives accuracy of over 95%, so, it is the best model to forecast demand.