Centre for Development of Advanced Computing, Mumbai

Report on

Sales Forecasting on Walmart Dataset

PG-DBDA March 2022

Submitted by:

Project Team 8

Yogita Khatavkar
Anushka Umbre
Pritam Powar
Shreya Singh
Yukti Pant

Mr. Prathmesh Shivaji Dalve
Project Guide

1 INTRODUCTION

“Walmart is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores”
The company has 45 stores across the United States. Every year the company runs several promotional markdown events before prominent holidays like Super Bowl, Labor Day, Thanksgiving, and Christmas to increase sales.
One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are salesdata available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Technology used - Python (ML), AWS, Spark, Data Visualization (Power BI)

1.1Project Goals and Background:

Walmart runs several promotional markdown sales throughout the year on days immediately following the prominent holidays in the United States; it becomes crucial for the organization to determine the impact of these promotional offerings on weekly sales to drive resources towards such key strategic initiatives. It is also essential for Walmart to understand user requirements and user buying patterns to create higher customer retention, increasing their demand adding to their profits. The findings from this study can help the organization understand market conditions at various times of the year and allocate resources according to regional demand and profitability.
The analysis for this study has been done using Python (ML), Big Data (Spark and Hive), AWS and Data Visualization (Power BI) on the dataset. The modelling, as well as the exploratory data analysis for the research, have been performed Python, aggregation and querying will be performed using Hive and the final dashboard has been created using Power BI.

1.2 Dataset:

Conventional retail stores still play a prominent role in a world dominated by Ecommerce. Retail is the process of selling consumer goods or services to customers through multiple channels of distribution to earn a profit. From groceries to clothing to electronics, customers keep flooding the gates of retail stores to satisfy their needs. As time has passed, retailers havehad to evolve in order to keep up with changes in demands and the ever-changing mindset of customers. One such retail industry juggernaut that has kept up with the demands of customersas well changed the face of the retail industry for the better is Walmart Inc.
Walmart Inc is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores, headquartered in Bentonville, Arkansas. They have many stores across the globe and it is the largest retail company by revenue.
We have historical sales data for 45 Walmart stores located in different regions. Eachstore contains a number of departments. Apart from these, weekly data of Fuel price, Holiday, Temperature with some other features are also present in the data set.

In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.
Data consists of 421570 records of weekly sales from stores spanning between '05-Feb- 2010' to '26-Oct-2012'. This comprises of 143 Weeks of sales data.
The datasets contain historical sales data for 45 Walmart stores in the United States along with store information and regional activity from 2/5/2010 to 11/1/2012. There are 3csv files: stores, sales, and features.
The variables are described below:

stores.csv
This file contains information about 45 Walmart stores and includes the following fields:

Store – the store number

Type – the type of store (A, B, or C)

Size – the size of store in square feet

sales.csv

This file contains 421,570 Walmart’s historical sales records from February 5th, 2010 toNovember 1st, 2012 and includes the following fields:

Store – the store number

Dept – the department number

Date – last day of the week

Weekly_Sales

Weekly sales for the given department in the given store

Negative if returns exceed sales

Positive if sales exceed returns

IsHoliday –
True if special holiday falls within the week; otherwise, False

features.csv

This file contains 8,190 records related to the store, department, and regional activity for

the given dates and includes the following fields:

Store – the store number

Date – last day of the week

Temperature – average temperature in the region in Fahrenheit

Fuel_Price – weekly average fuel price (USD)

MarkDown1-5 – anonymized data related to promotional markdowns that Walmartis running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time.

CPI – consumer price index

Unemployment – weekly average unemployment rate

IsHoliday – True if holiday falls within the week. False if holiday does not fall withinthe week.

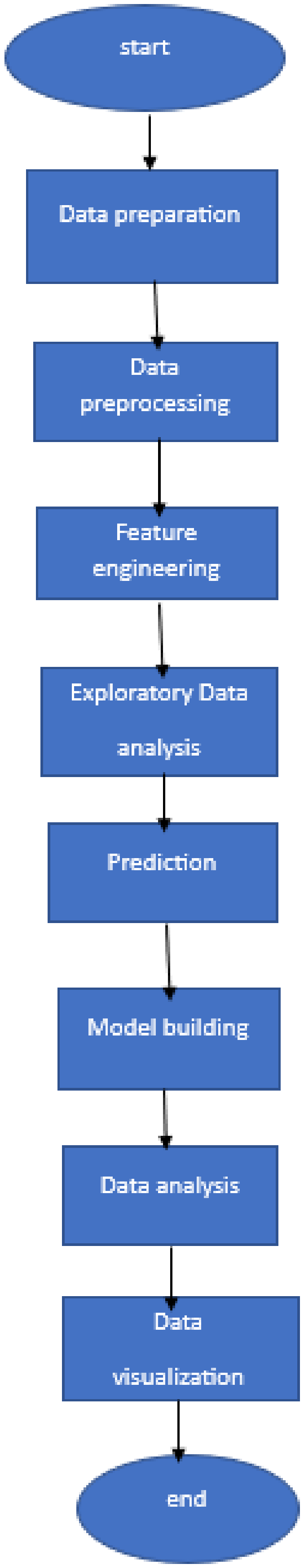
2 PROBLEM STATEMENT

There are many seasons that sales are significantly higher or lower than averages. If the company does not know about these seasons, it can lose too much money. Predicting future sales is one of the most crucial plans for a company. Sales forecasting gives an idea to the company for arranging stocks, calculating revenue, and deciding to make a new investment. Another advantage of knowing future sales is that achieving predetermined targets from the beginning of the seasons can have a positive effect on stock prices and investors' perceptions. Also, not reaching the projected target could significantly damage stock prices, conversely. And, it will bea big problem especially for Walmart as a big company.

1.2 Objective:

Our aim in this project is to build a model which predicts sales of the stores. With this model, Walmart authorities can decide their future plans which is very important for arranging stocks, calculating revenue and deciding to make new investment or not.
This project studies Walmart’s historical sales data for 45 stores in the United States to assist Walmart’s management team in the decision-making process by:
Performing exploratory data analysis and time series analysis of Walmart’s sales data.
Identifying the factors that impact sales.
Developing machine learning models to forecast Walmart’s future sales.

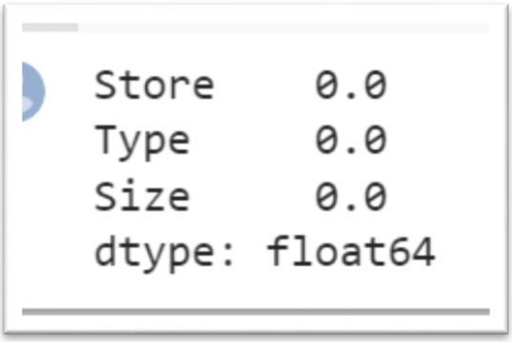
3 WORK FLOW


start
Data preparation
Data preprocessing
Feature engineering
Exploratory Data analysis
Prediction
Model building
Data analysis
Data visualization
end

4 DATA PREPARATION

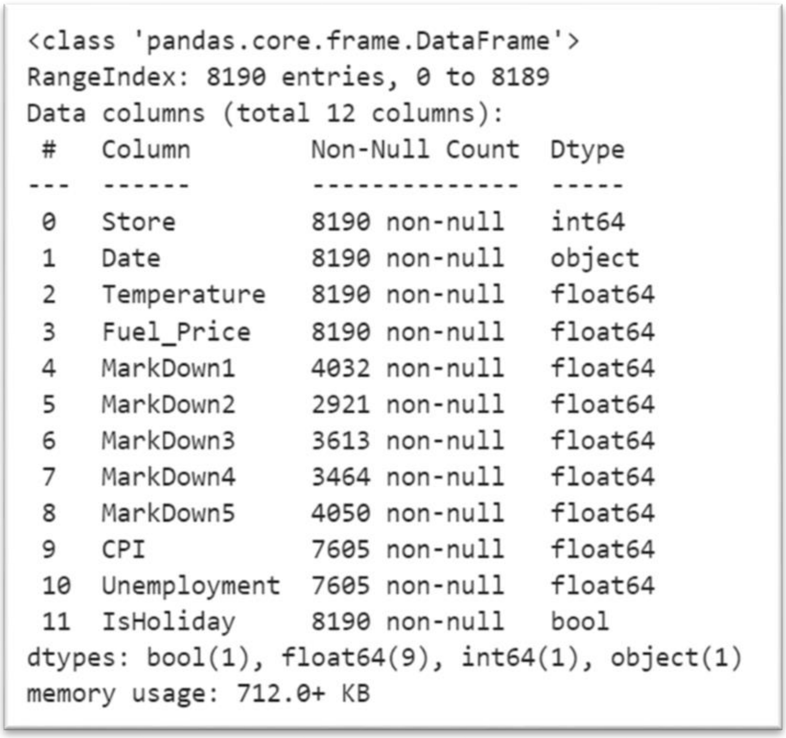
Missing Values:

There are no missing values in store data.



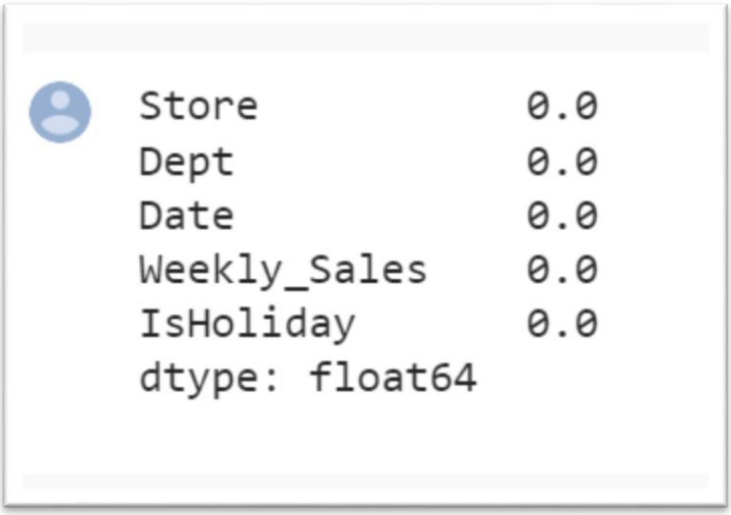
```
Store     0.0
Type      0.0
Size      0.0
dtype: float64
```

There are seven columns that contain missing values including MarkDown 1-5, CPI, and Unemployment. Since most missing values exist because there was no information available at a specific time, fields containing missing values are left as ‘NA’. The table below lists columns that have missing values along with statistics:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8190 entries, 0 to 8189
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         8190 non-null   int64
 1   Date          8190 non-null   object
 2   Temperature   8190 non-null   float64
 3   Fuel_Price    8190 non-null   float64
 4   MarkDown1     4032 non-null   float64
 5   MarkDown2     2921 non-null   float64
 6   MarkDown3     3613 non-null   float64
 7   MarkDown4     3464 non-null   float64
 8   MarkDown5     4050 non-null   float64
 9   CPI           7605 non-null   float64
 10  Unemployment  7605 non-null   float64
 11  IsHoliday     8190 non-null   bool
dtypes: bool(1), float64(9), int64(1), object(1)
memory usage: 712.0+ KB
```

There are no missing values in sales data.



```
Store           0.0
Dept            0.0
Date            0.0
Weekly_Sales    0.0
IsHoliday       0.0
dtype: float64
```

New Columns:

The following variables have been added to the dataset:

Total_MarkDown = sum of MarkDown1-5

Year = year extracted from Date

Month = month extracted from Date

Week = week of year extracted from Date

Outliers:

There are 9 columns that have outliers: Num of Depts, Dept Weekly Sales (Thousand), Temperature, MarkDown1-5, and Unemployment. Since outliers may contain important information, no outliers have been removed from the table.

One-Hot Encoding:

One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assigna binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.
We have done one-hot encoding by using get_dummies function of pandas library. In our dataset, we applied one-hot encoding on following categorical columns:
Store
Department
Type

Data Normalization:

It is a scaling technique method in which data points are shifted and rescaled so that they end up ina range of 0 to 1. It is also known as min-max scaling.
We have normalized our dataset by using MinMaxScaler function from sklearn library on following columns:

- Weekly_Sales
- Size
- Temperature
- Fuel_Price
- CPI
- Unemployment
- Total_MarkDown

Data Splitting into Training, Testing:

The main difference between training data and testing data is that training data is the subsetof original data that is used to train the machine learning model, whereas testing data is used to checkthe accuracy of the model. The training dataset is generally larger in size compared to the testing dataset.
One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model. A better option is to split our data into two parts: first one for training our machine learning model, and second one for testing our model. Split the dataset into two pieces: a training set and a testing set. Train the model on the training set. Test the model on the testing set, and evaluate how well our model did.
Here we use train test split function to split or data set we have split 20% data as test data andremaining 80% as training data

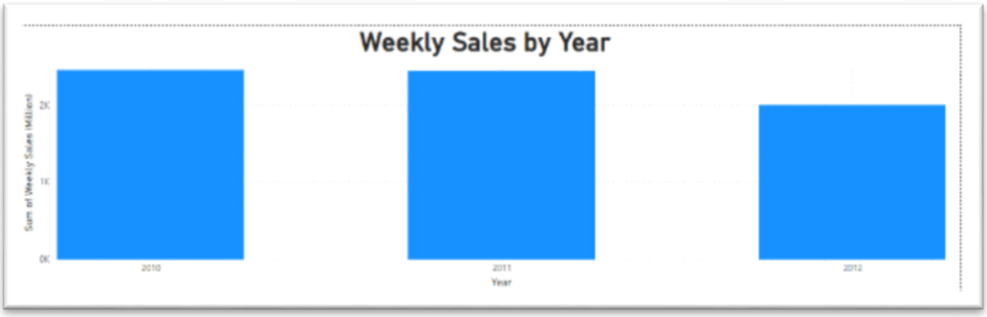
Advantages of train/test split:

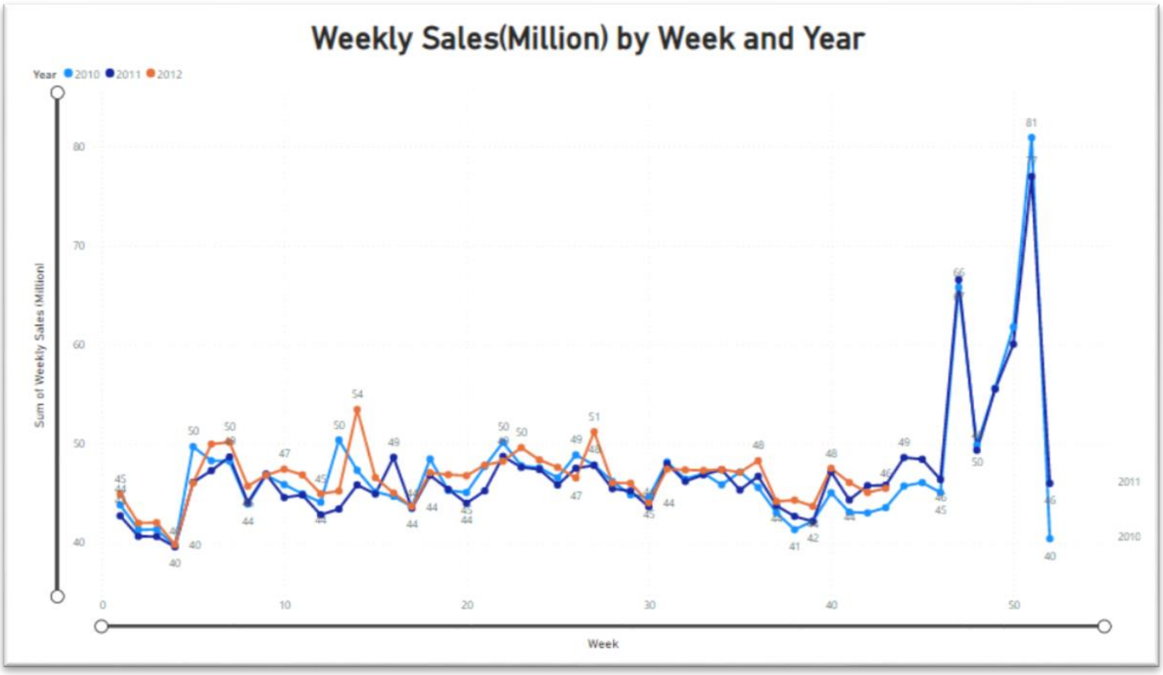
- Model can be trained and tested on different data than the one used for training.
- Response values are known for the test dataset, hence predictions can be evaluated
- Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

5 EXPLORATORY DATA ANALYSIS

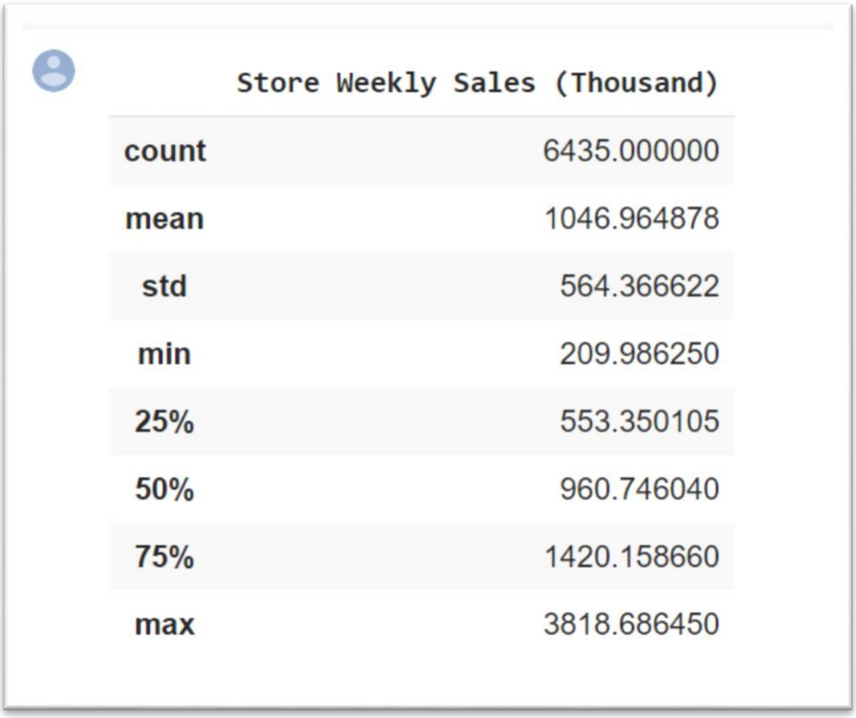
Bivariate Data Analysis:

The strip plot below shows Walmart’s yearly sales by store from 2010 to 2012. Walmart’s sales seem to be low in 2010 and 2012 because we are missing sales data for the first month of 2010 and the last two months of 2012.



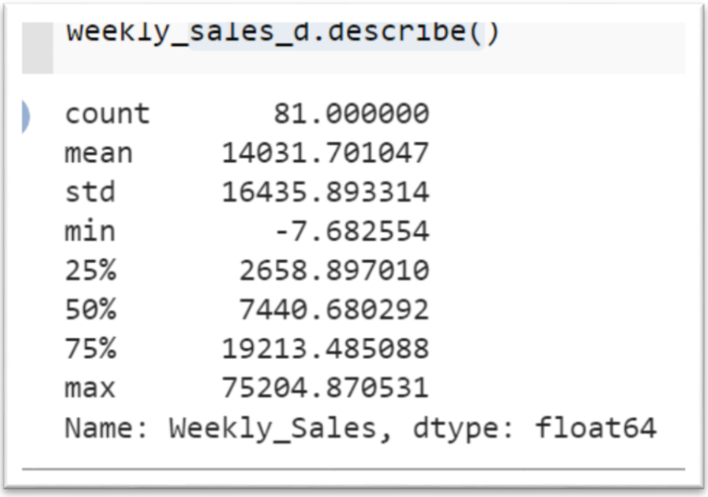


From the graph, it is seen that 2011 has lower sales than 2010 generally. When we look at the mean sales it is seen that 2010 has higher values, but 2012 has no information about Novemberand December which have higher sales. Despite of 2012 has no last two months sales, it's mean is near to 2010. Most probably, it will take the first place if we get 2012 results and add them.
Walmart’s store weekly sales ranges from $209 thousand to $3.8 million. About half of the 45 Walmart stores have weekly sales greater than or equal to $960 thousand.



	Store Weekly Sales (Thousand)
count	6435.000000
mean	1046.964878
std	564.366622
min	209.986250
25%	553.350105
50%	960.746040
75%	1420.158660
max	3818.686450

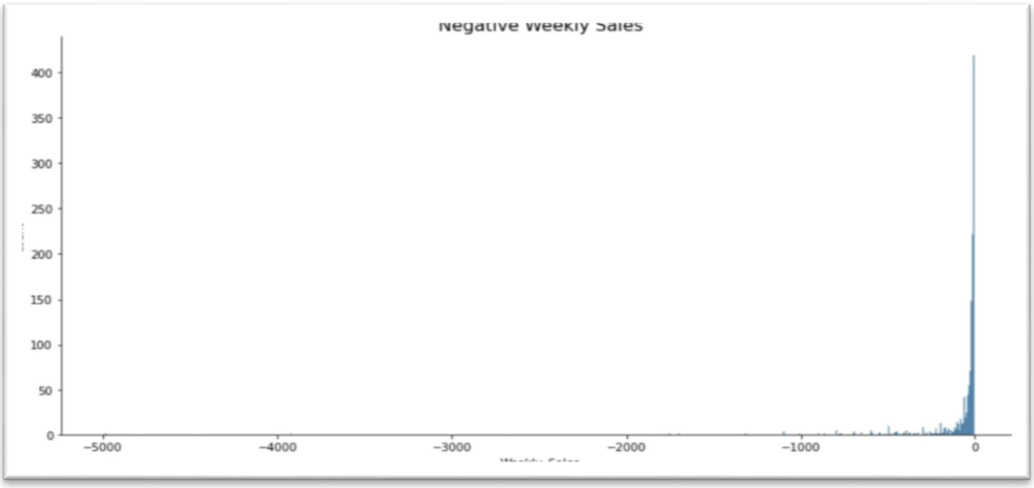
The department's weekly average sales range from -$7,682 to $75204,870. The negative values in weekly sales indicates returns exceed sales in the department store. About half of Walmart's department stores have sales greater than or equal to $7440,682.



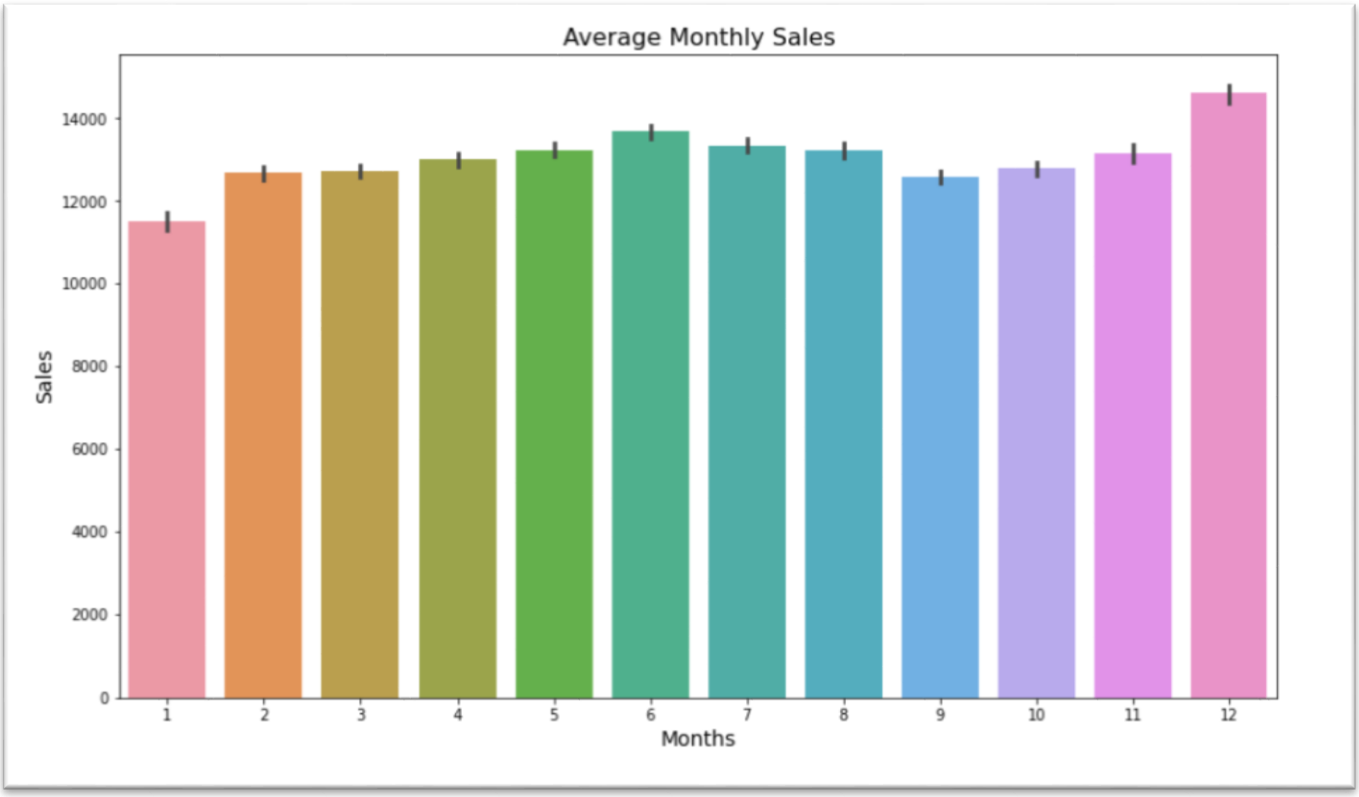
```
weekly_sales_d.describe()

count       81.000000
mean     14031.701047
std      16435.893314
min         -7.682554
25%       2658.897010
50%       7440.680292
75%      19213.485088
max      75204.870531
Name: Weekly_Sales, dtype: float64
```

Negative Weekly Sales:

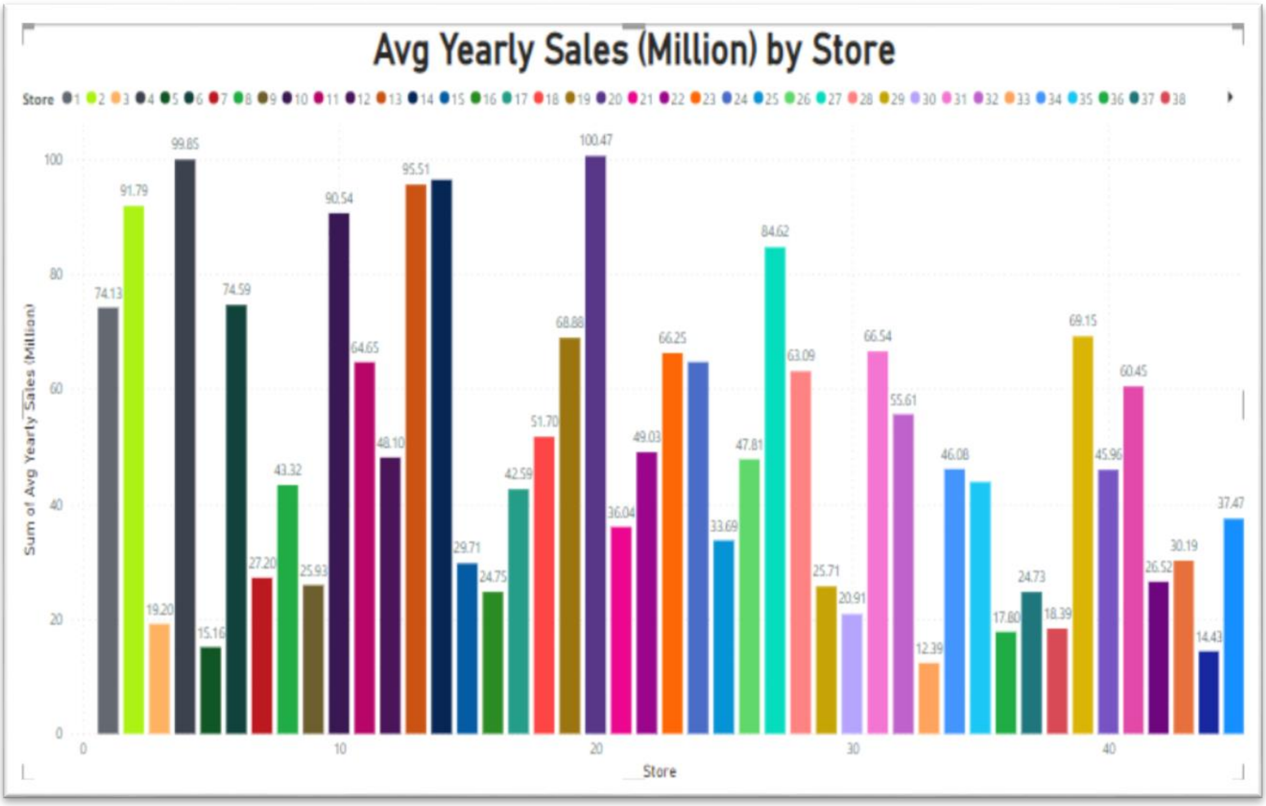


In Below Graph, we can see average monthly sales are highest in December and lowest in January. When we look at the graph above, the best sales are in December and November, as expected. The highest values are belonging to Thanksgiving holiday but when we take average it is obvious that December has the best value.

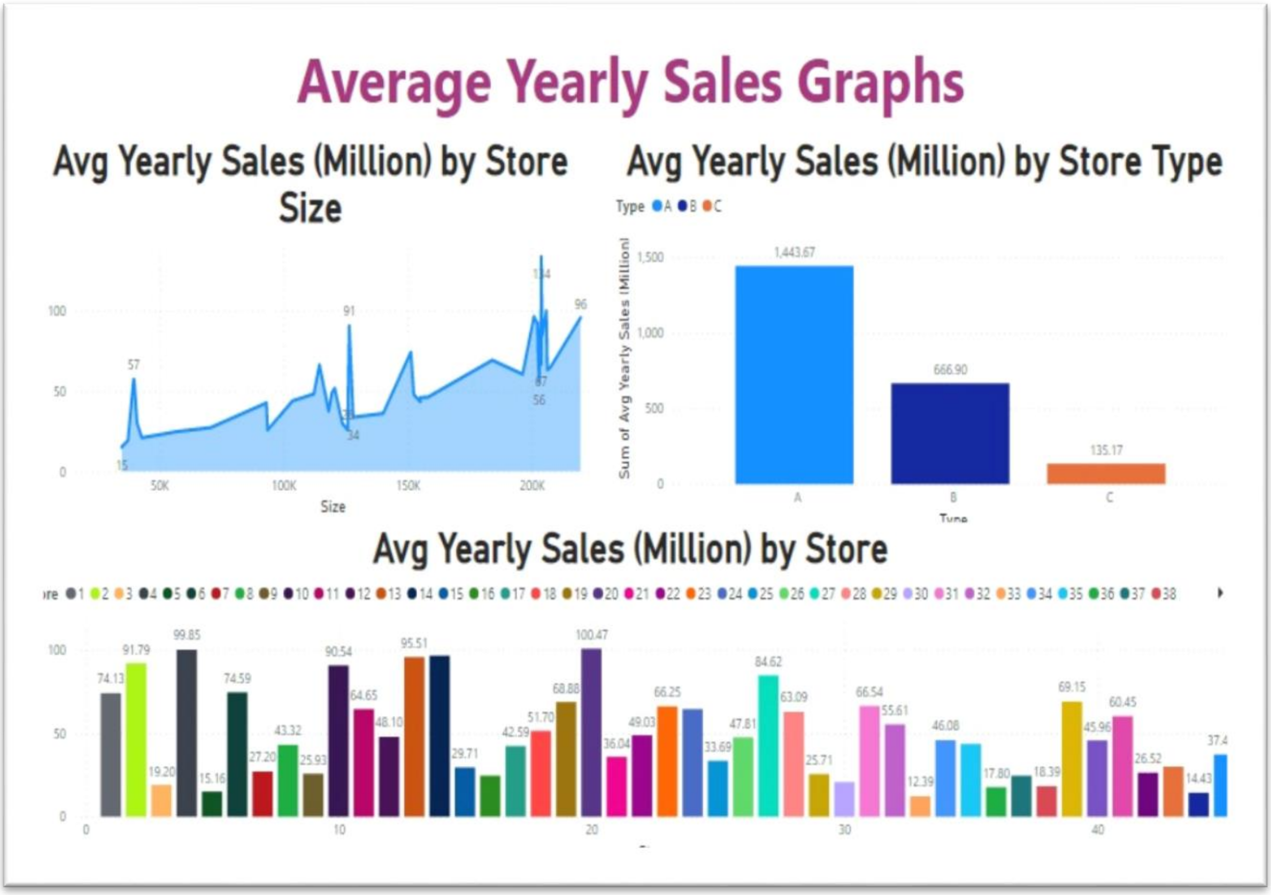


In Below figure, we have average sales per store in which we can see top 5 stores are 20,14,4,2,13 having maximum sales and stores 3,5,29,33,44 have lowest sales.

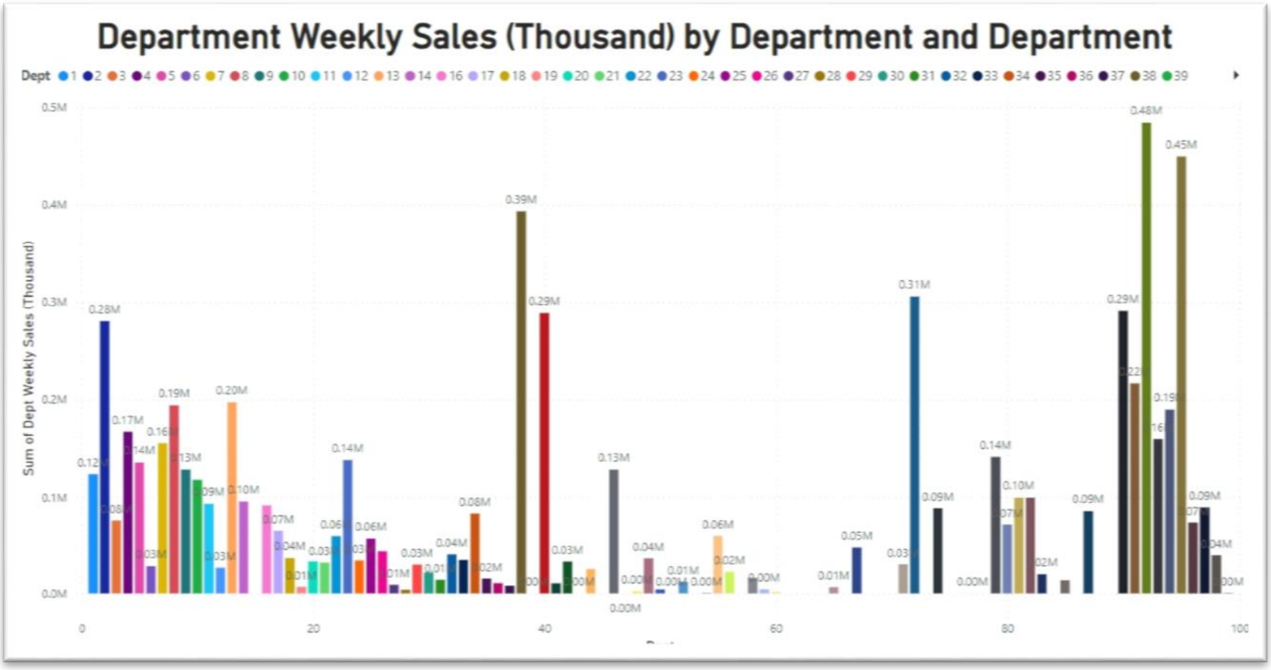
Avg Yearly Sales (Million) by Store



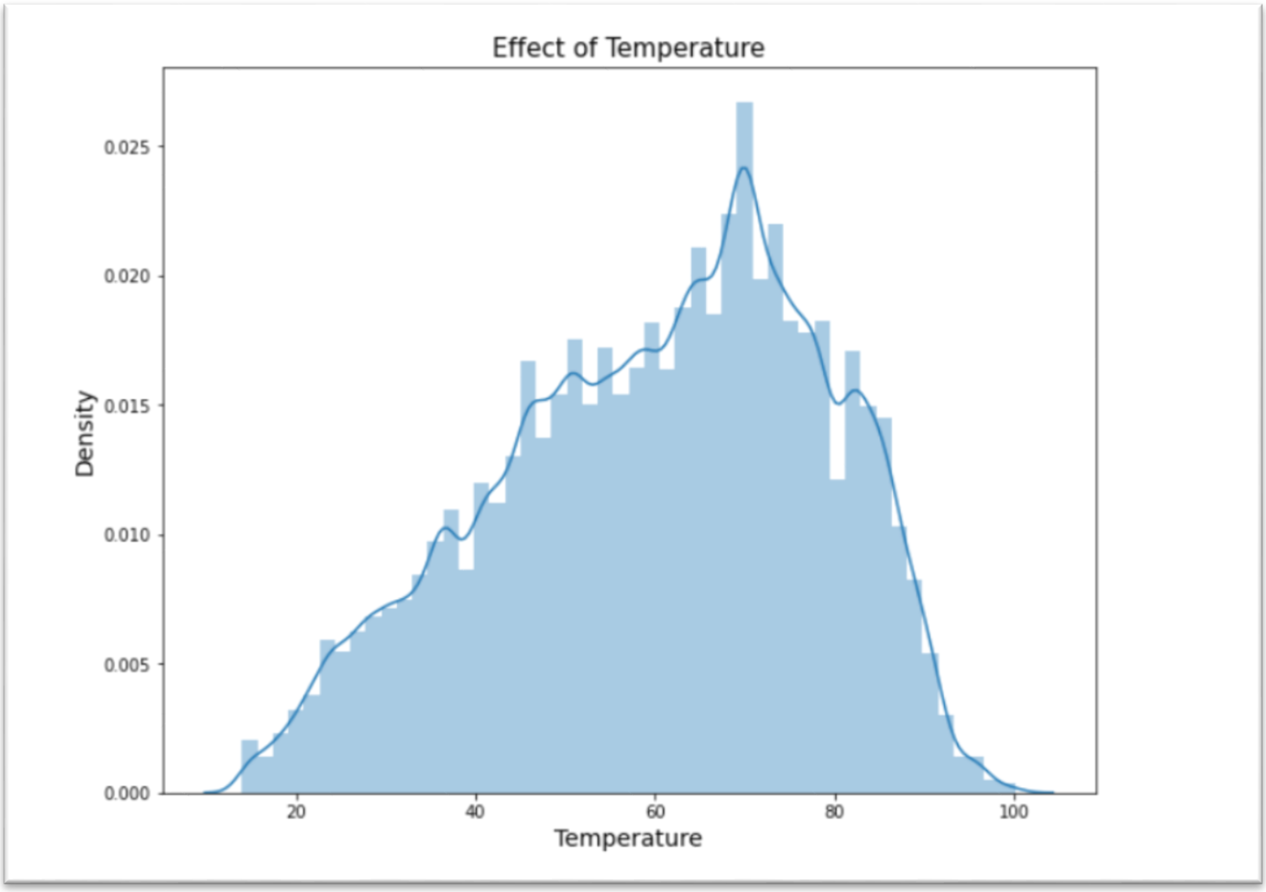
Average Yearly Sales



Weekly Sales by Department:

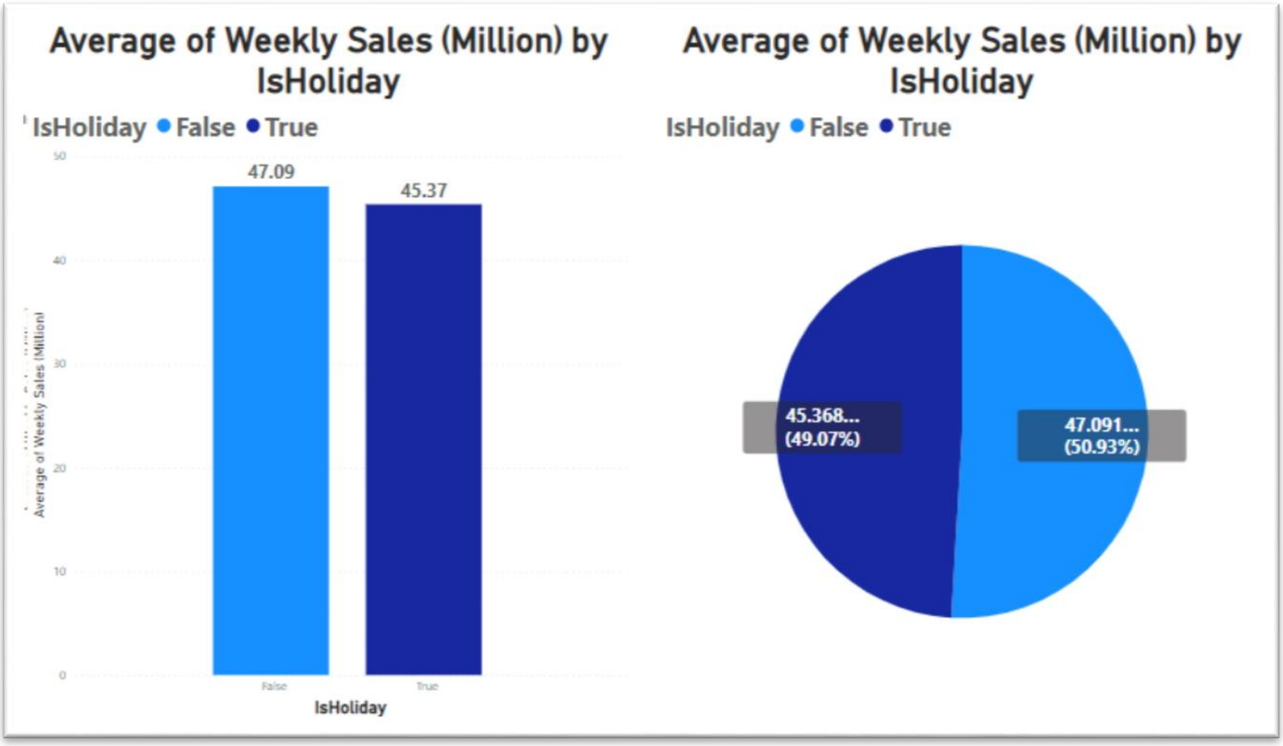


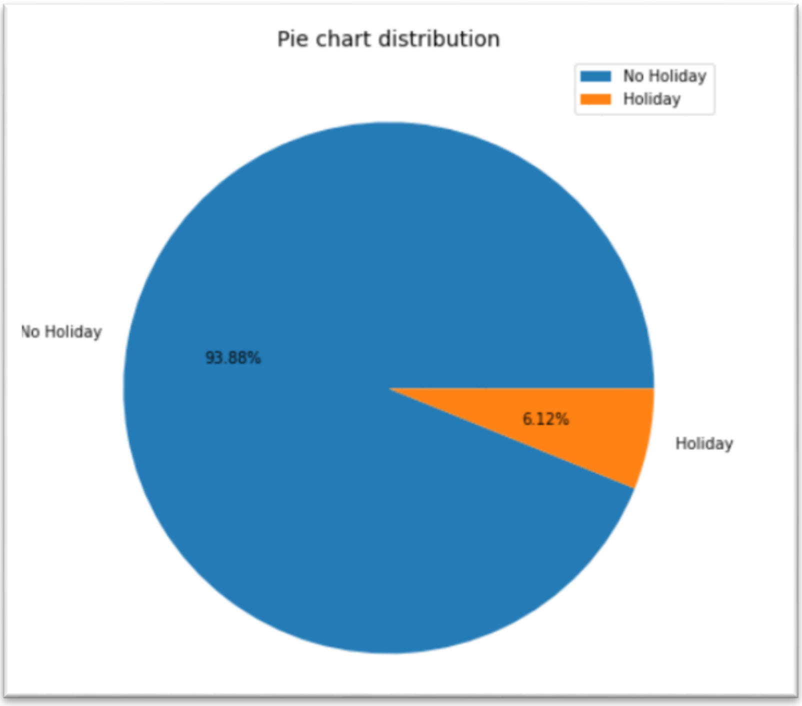
Temperature Density Plot



In this Chart, we can see that most of the sales are likely to be between 40-90º Fahrenheit. And maximum sales are near 70º Fahrenheit.

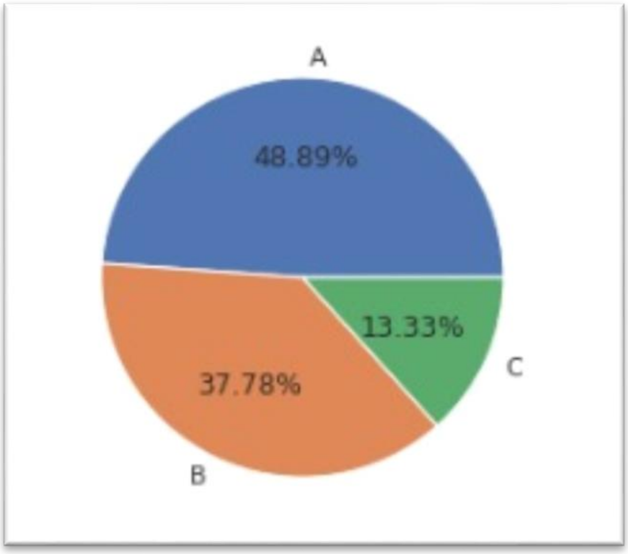
Sales by Holiday

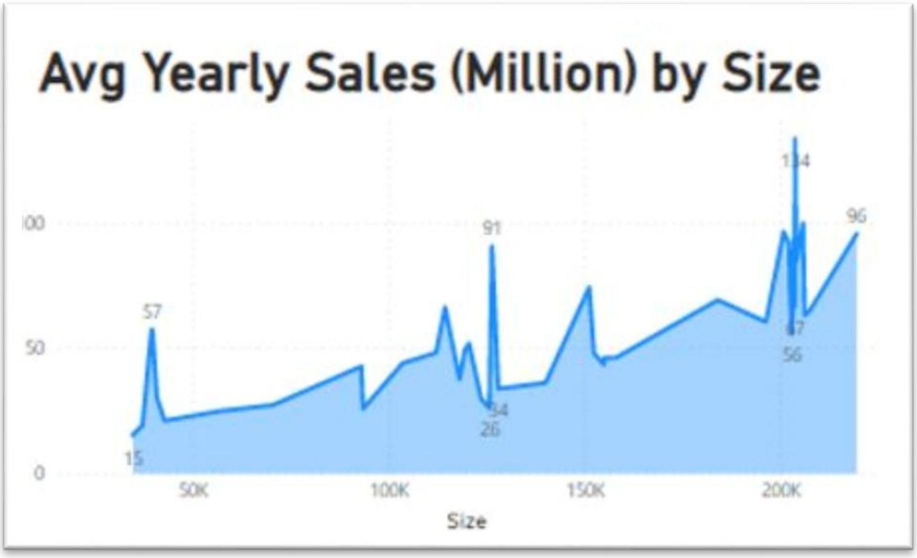




Weeks with holiday are 6.12% and Weeks without holiday 93.88%. Still holiday average sales are almost equal to sales on non-holiday weeks.

Walmart stores are classified into 3 types: A (22 stores), B (17 stores), and C (6 stores).



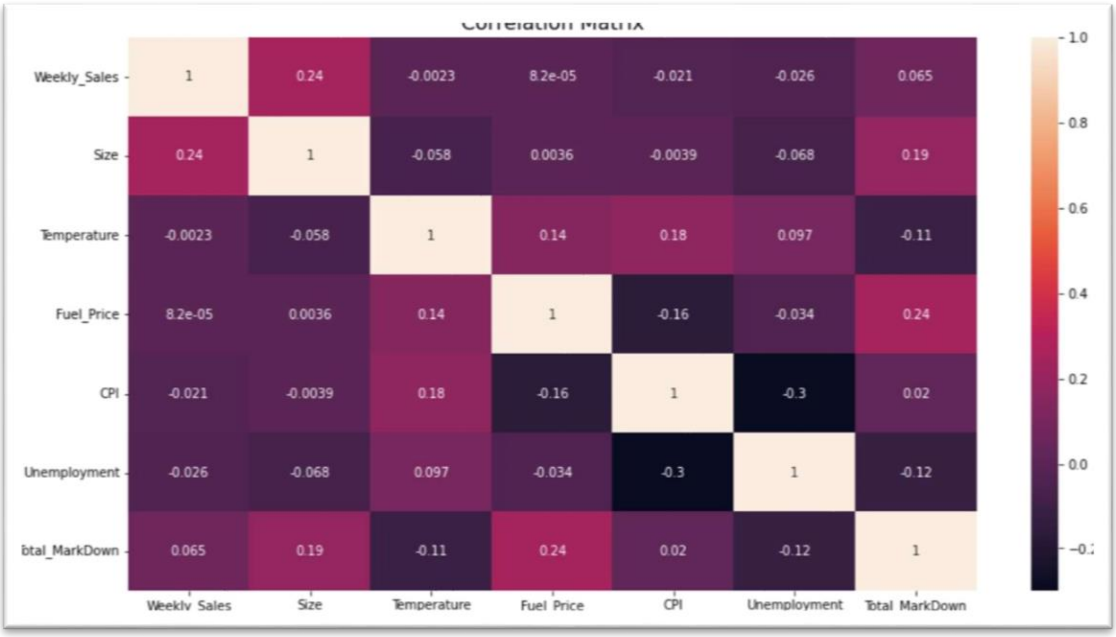


Most type A stores seems to be in a group that have highest average yearly sales. Most type C stores appears to be in a group that have lowest average yearly sales. Most type B stores’ average yearly sales are higher than type C stores’ average yearly sales and lower than type A stores’ average yearly sales.

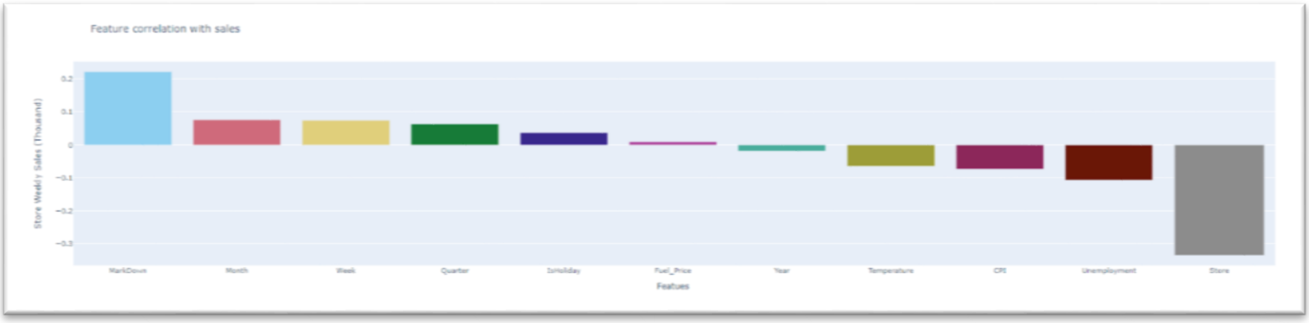
Multivariate Data Analysis:

The heatmap bellow shows that there is not a strong correlation between the following variables:

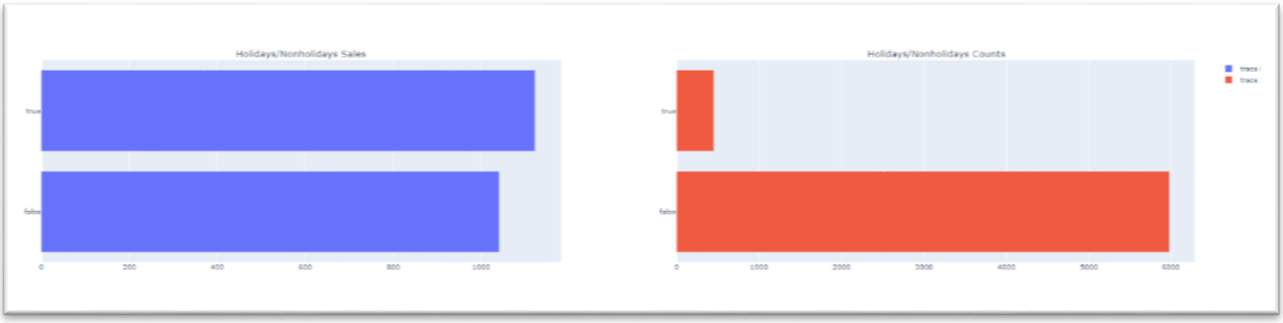
- Store weekly sales and temperature
- Store weekly sales and fuel price
- Store weekly sales and CPI
- Store weekly sales and unemployment
- Store weekly sales and weekly markdown



Correlation of features with weekly sales



Markdown, Date (Month, week, Quarter) and Isholiday, fuel price has positive correlation with sales. Year, Temperature, CPI, Unemployment and store has negative correlation with weekly sales.



Only 7 percent of the weeks in the data are the holiday weeks Despite being the less percentage of holiday weeks the sales in the holiday's week are on the average higher than in the non-holiday weeks.

Conclusion On EDA:

Walmart’s stores are classified into three types: A, B, and C. Type A stores often have high sales, big store’s sizes, large number of departments, and large markdown values. Type C stores often have low sales, small store sizes, small number of departments, and small.
markdown values. Walmart’s sales are often at peak during the week of Thanksgiving and three weeks after Thanksgiving. External factors like temperature, fuel price, consumer price index, and unemployment rate do not have significant impact on Walmart’s sales. Promotional
markdown events before holidays seem to increase Walmart’s sales except for Christmas.

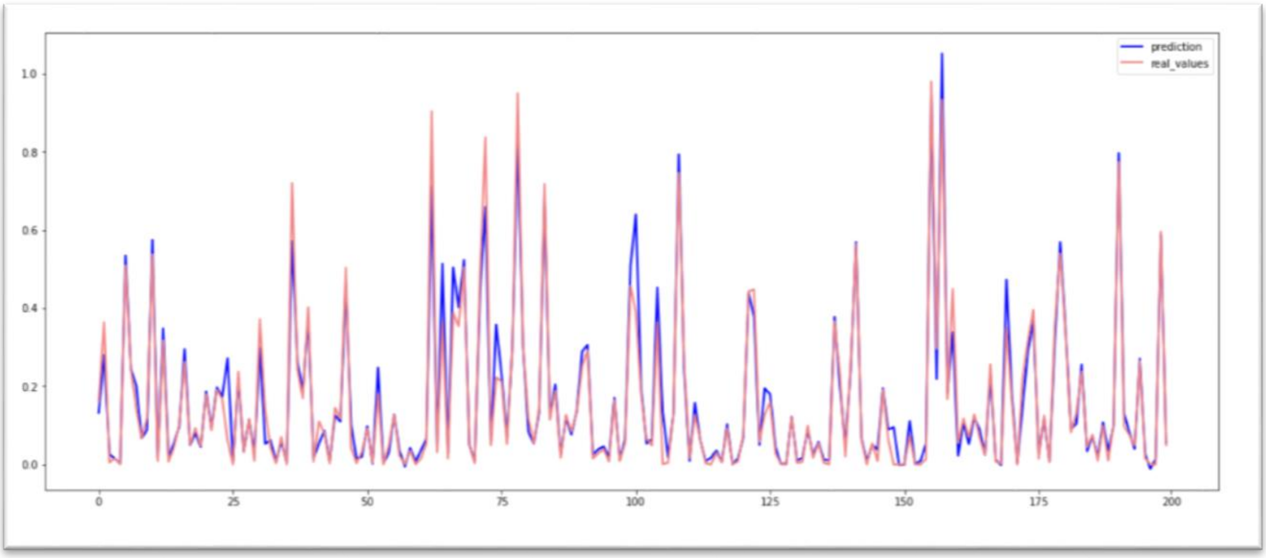
6 FUTURE IMPROVEMENTS

More data with no missing values can be useful and will increase the model accuracy.
More detailed feature engineering and feature selection can be done.
More data can be found to observe holiday effects on sales and different holidays will be added like Easter, Halloween and Come Back to School times.
Markdown effects on model can be improved according to department sales. Different models can be built for special stores or departments.
Market basket analysis can be done to find higher demand items of departments.

7 MODEL TESTING

1. Linear Regression:

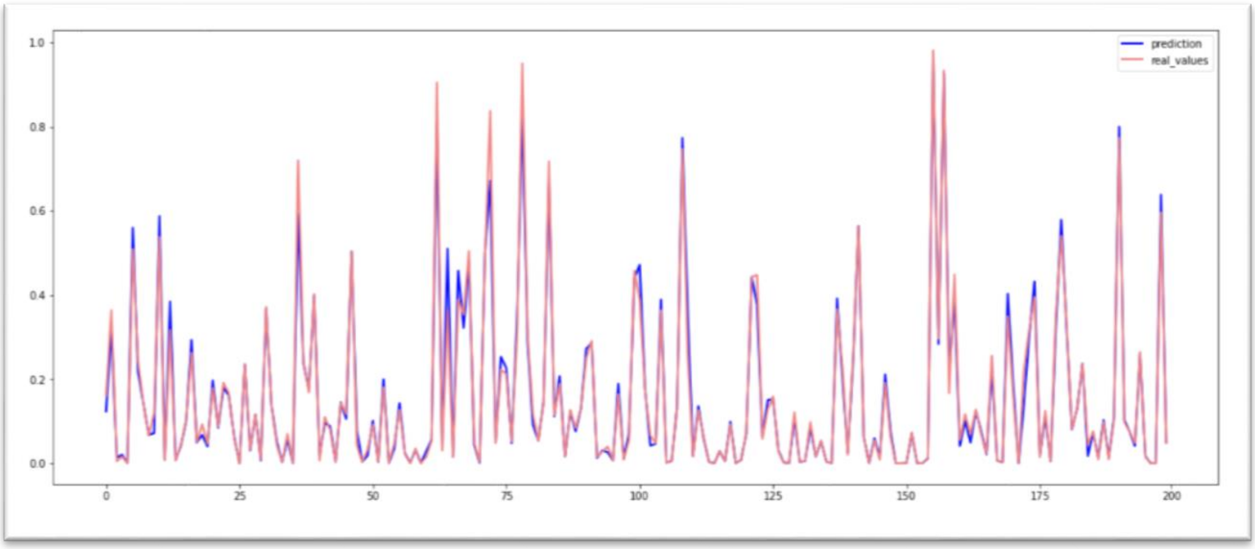
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.



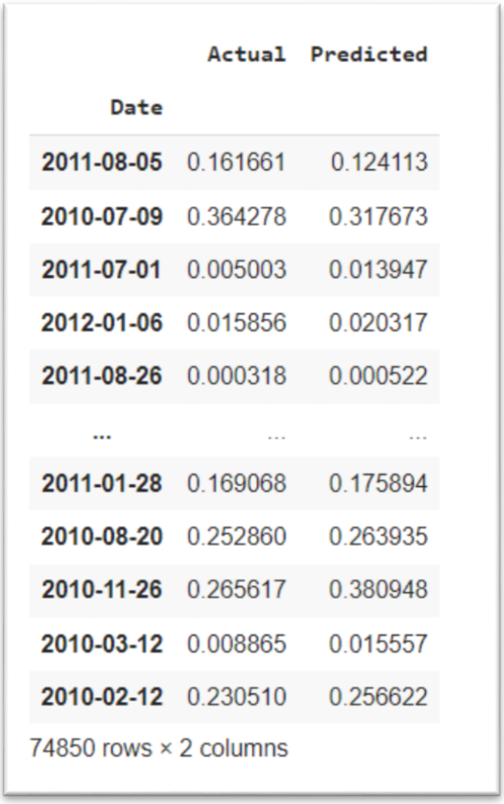
In this linear regression model, we got an accuracy is equal to 92.28%
MAE 0.03005771492151418
MSE 0.0034851431916206573
RMSE 0.05903510135182845
R2 0.9228079866096734

2. Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



In this model we got an accuracy is equal to 97.90%
MAE 0.015497700214673305
MSE 0.0009441195875861165
RMSE 0.030726529052044205
R2 0.9790892409415295

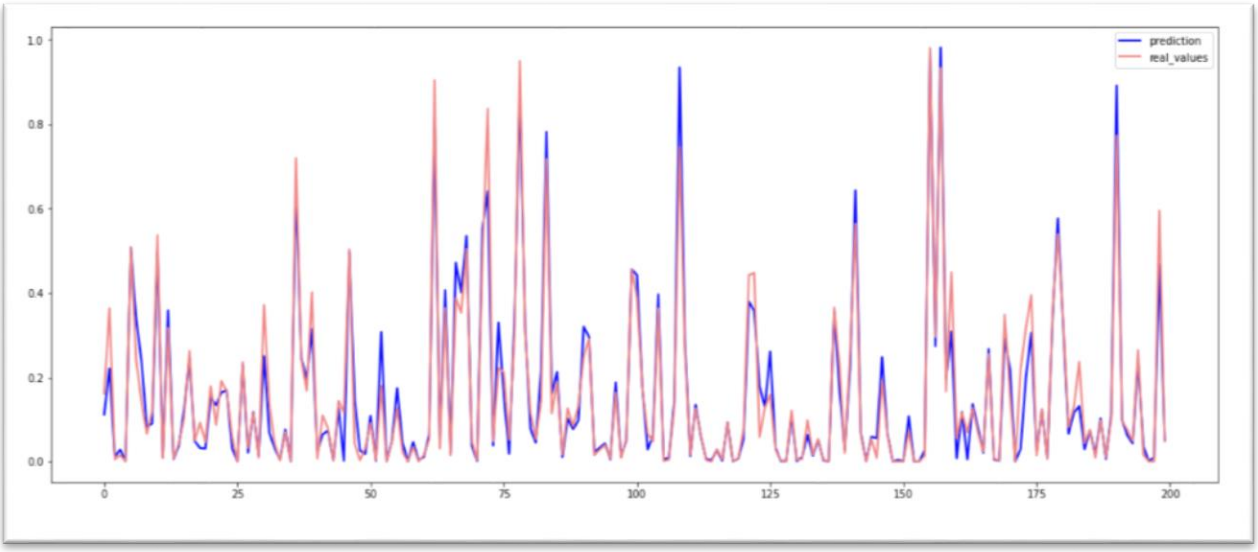


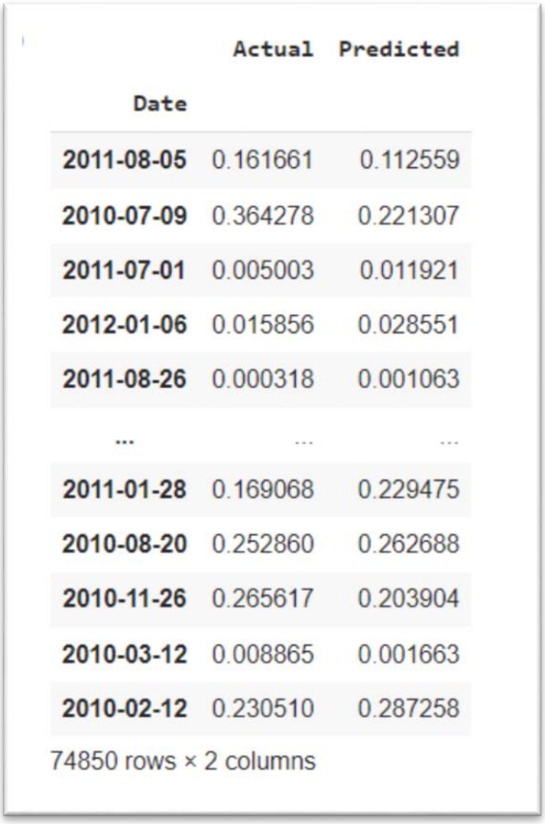
Date	Actual	Predicted
2011-08-05	0.161661	0.124113
2010-07-09	0.364278	0.317673
2011-07-01	0.005003	0.013947
2012-01-06	0.015856	0.020317
2011-08-26	0.000318	0.000522
...	...	...
2011-01-28	0.169068	0.175894
2010-08-20	0.252860	0.263935
2010-11-26	0.265617	0.380948
2010-03-12	0.008865	0.015557
2010-02-12	0.230510	0.256622

74850 rows × 2 columns

3. K Neighbors Regressor Model:

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimizes the mean-squared error.





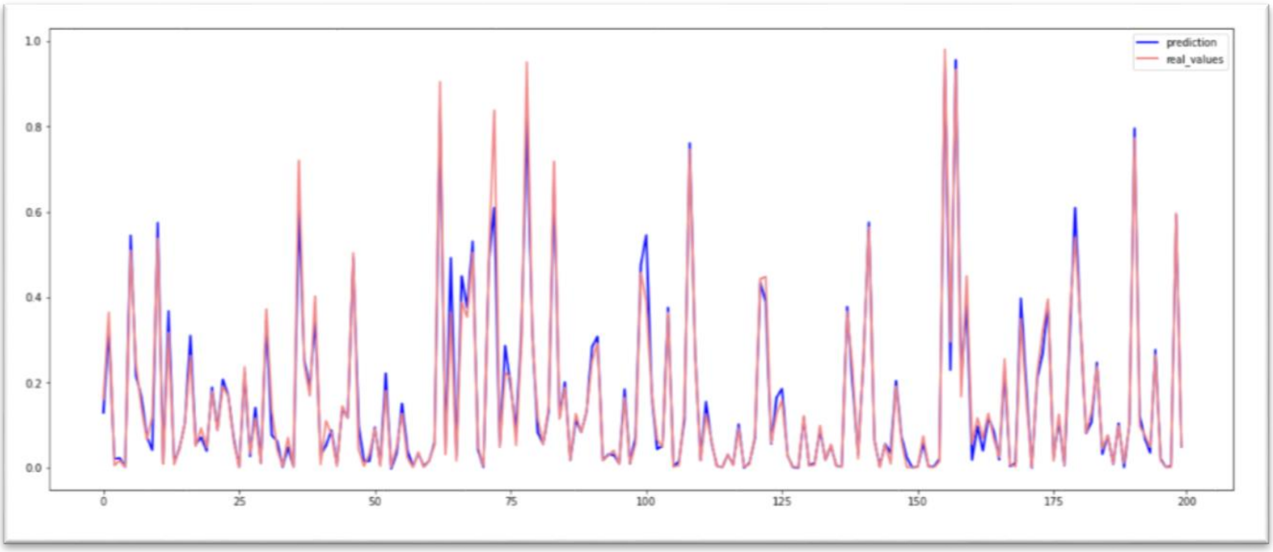
	Actual	Predicted
Date		
2011-08-05	0.161661	0.112559
2010-07-09	0.364278	0.221307
2011-07-01	0.005003	0.011921
2012-01-06	0.015856	0.028551
2011-08-26	0.000318	0.001063
...	...	...
2011-01-28	0.169068	0.229475
2010-08-20	0.252860	0.262688
2010-11-26	0.265617	0.203904
2010-03-12	0.008865	0.001663
2010-02-12	0.230510	0.287258

74850 rows × 2 columns

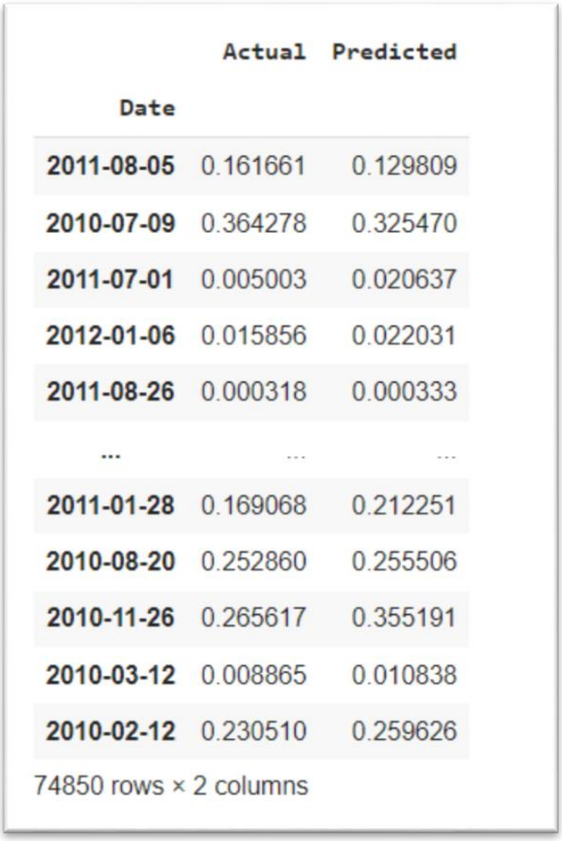
In this model we got an accuracy is equal to 91.97%
MAE 0.033122157844495987
MSE 0.003624289652612284
RMSE 0.060202073490971085
R2 0.9199211027805663

4. XGboost Model:

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.



In this model we got an accuracy is equal to 97.28%
MAE 0.019863883311422855
MSE 0.0012266346907320747
RMSE 0.03502334493922696
R2 0.9728314959830722

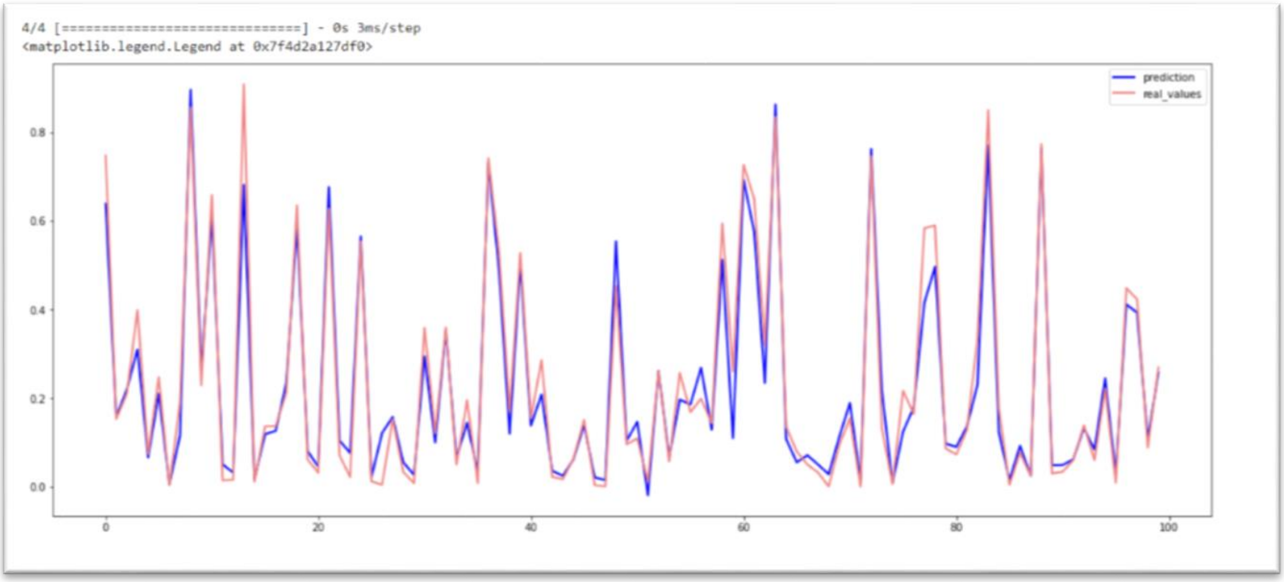


	Actual	Predicted
Date		
2011-08-05	0.161661	0.129809
2010-07-09	0.364278	0.325470
2011-07-01	0.005003	0.020637
2012-01-06	0.015856	0.022031
2011-08-26	0.000318	0.000333
...	...	...
2011-01-28	0.169068	0.212251
2010-08-20	0.252860	0.255506
2010-11-26	0.265617	0.355191
2010-03-12	0.008865	0.010838
2010-02-12	0.230510	0.259626

74850 rows × 2 columns

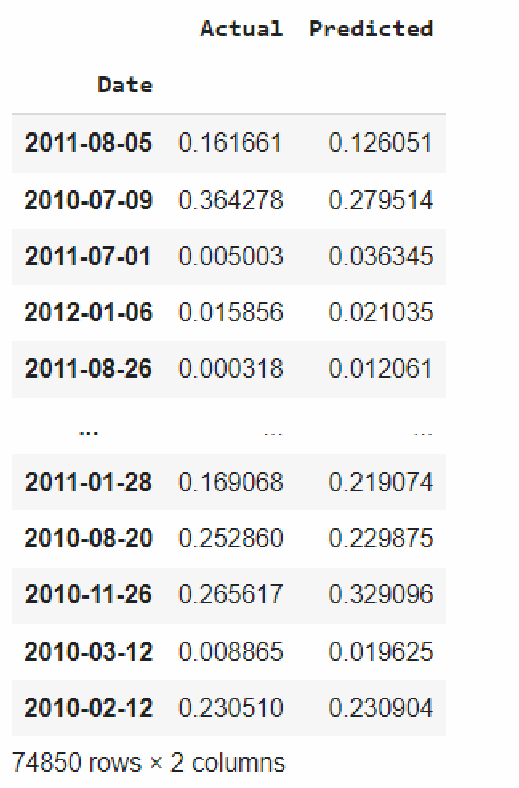
5. Custom Deep Learning Neural Network:

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.



```
4/4 [==============================] - 0s 3ms/step
<matplotlib.legend.Legend at 0x7f4d2a127df0>
```

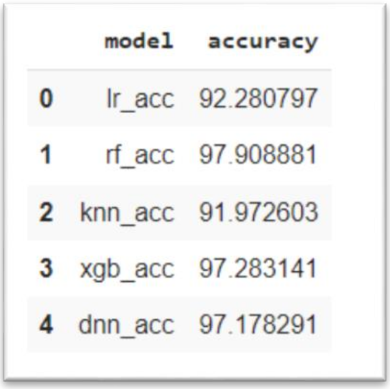
In this model we got an accuracy is equal to 97.17%
MAE 0.034005930661177824
MSE 0.003993838791591234
RMSE 0.0631968258031306
R2 0.9119278490383779



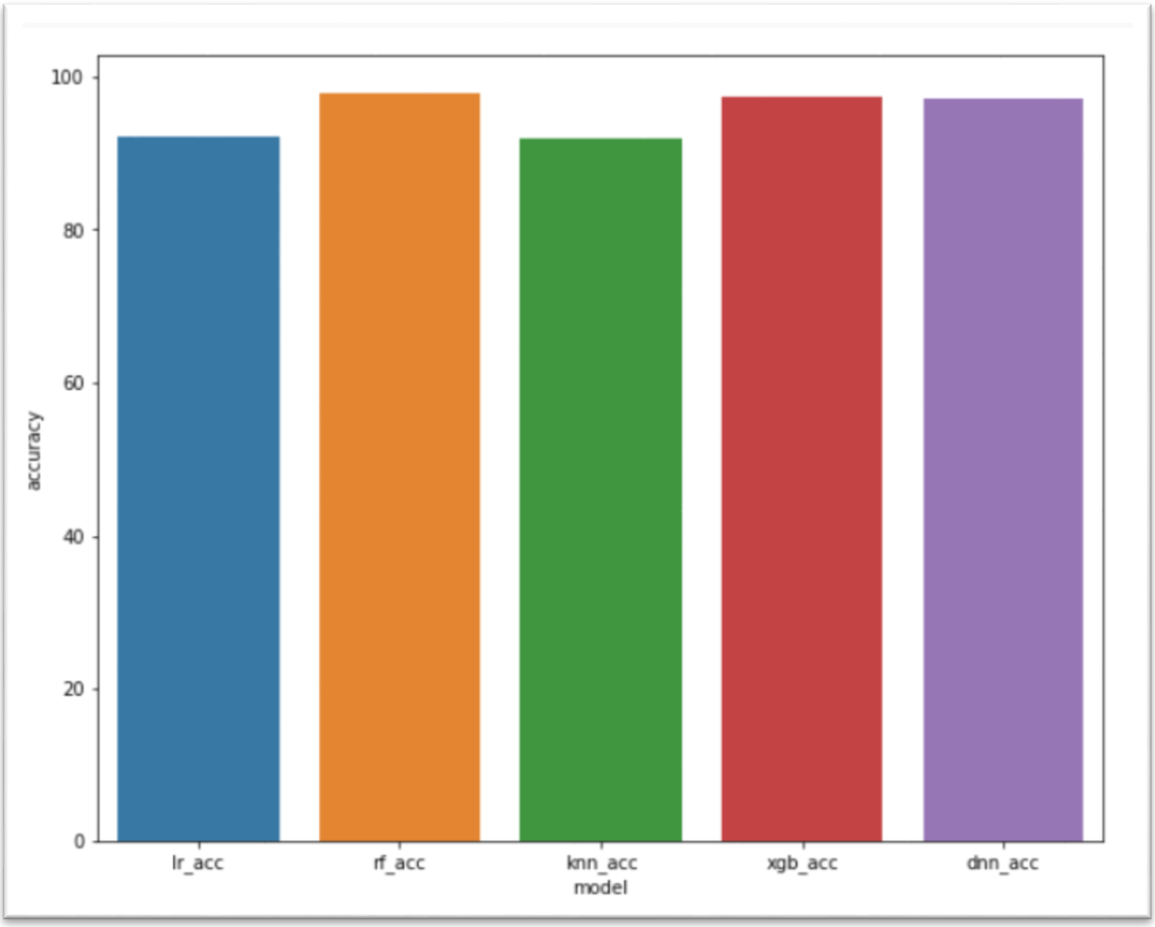
Date	Actual	Predicted
2011-08-05	0.161661	0.126051
2010-07-09	0.364278	0.279514
2011-07-01	0.005003	0.036345
2012-01-06	0.015856	0.021035
2011-08-26	0.000318	0.012061
...	...	...
2011-01-28	0.169068	0.219074
2010-08-20	0.252860	0.229875
2010-11-26	0.265617	0.329096
2010-03-12	0.008865	0.019625
2010-02-12	0.230510	0.230904

74850 rows × 2 columns

8 COMPARING MODELS



	model	accuracy
0	lr_acc	92.280797
1	rf_acc	97.908881
2	knn_acc	91.972603
3	xgb_acc	97.283141
4	dnn_acc	97.178291

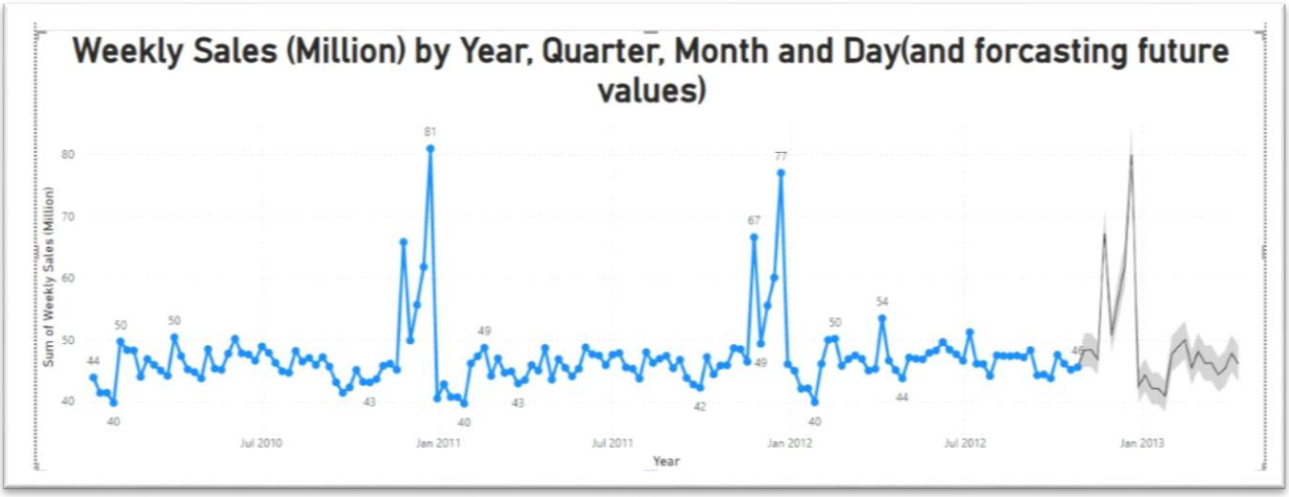


From the above figures we can see that the **random forest regression** model has the highest accuracy Which is **97.90 %.** So, we will use random forest regression model for sales prediction of this Walmart datasets other models also perform well as their accuracies are:

- Linear regression = 97.28 %
- KNN Regression = 91.97 %
- XGBoost regression = 97.28 %
- Deep Learning Neural Network=97.17%

Sales Forecast using Power BI:

Forecasting in Power View is based on an established suite of methods for time series prediction called exponential smoothing. The exponential smoothing method has a good track record in both academia and business, and has the advantage that it suppresses noise, or unwanted variation that can distort the model, while efficiently capturing trends. Power View uses the appropriate model automatically when we start a forecast for our line chart, based on an analysis of the historical data. The graph plotted below is between weekly sales and year. Here we are forecasting for next 6 months (from 4 October 2012 to 2 April 2013)



9 CONCLUSION

In conclusion, we find that our regression equation is quite accurate (97.17% accuracy) in predicting the weekly sales with Random Forest Regressor. Walmart can use it to forecast the sales better. They need to overhaul the markdowns that are given currently as they are not having the intended impact on sales. They need to focus on the year-end inventory as December month plays a crucial part in predicting sales.

REFERENCES

Dataset link:
https://www.kaggle.com/datasets/divyajeetthakur/walmart-sales-prediction

Models:

Linear Regression:
https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least- squares

Random Forest Regression:
https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized- trees

KNN regression:
https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors- regression

XGBoost regression:
https://xgboost.readthedocs.io/en/latest/index.html

Deep Learning Neural Network:
https://www.tensorflow.org/tutorials/keras/regression