

**Capstone**

**Project**

**Project-1: Walmart Store**

Nadeem Irfan

21-September-2023

Table of content

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons For Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same

10) Future Possibilities of the Project

11) Conclusion

12) References

13) Appendix

Problem Statement

Walmart has multiple outlets across the country, which is facing issues in managing the inventory, to match the demand with respect to supply. This project undertakes to review the sales records from the store and to provide useful insights to the company and also forecast the sales for the each store for the next twelve weeks.

Project Objective

Walmart has multiple outlets across the country, are facing issues in inventory management. The task is to come up with useful insights using the provided data and make prediction models to forecast the sales for the next twelve weeks.

Data Description

The available dataset contains 6,435 records (rows) and eight features (columns) as shown in the table below.

|  |  |
| --- | --- |
| **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 |

Description of the data,

From the given data of the company, it is observed that there are six thousand four hundred and thirty five (6,435) records and seven features captured weekly. Further Data description is as follows:

1. Stores: there are 45 stores and for each store there are 143 entries of:
   1. Starting date of the week when the sales happen,
   2. Total sales for a given week,
   3. Holiday flag for the week (1 if there is holiday in the week or else 0),
   4. Temperature on the day of the sales,
   5. Cost of the fuel in the location of the store,
   6. Consumer price index of the given week
   7. Rate of the unemployment for the given week of the record.

Data Pre-processing Steps and Inspiration

The pre-processing of the data includes the following steps:

1. Step 1: Load the data.
2. Step 2: Perform exploratory data analysis (EDA).
   * 1. Confirming number of records in the data and how they are distributed.
     2. Check data type.
     3. Check for missing data, invalid entries, and duplicate records.
     4. Examine the correlation of the independent features with the target (weekly\_sales) variable.
     5. Check the outliers that are known to distort prediction and forecast.
3. Step 3: Model prediction, two approaches:
4. Time series model (ARIMA).
5. Linear Regression model(s).
6. Step 4: Data forecasting.
7. Step 5: Compare results from different models.

Exploratory Data Analysis

**Correlation of features with the target variable in entire sales data:**

There are 5 features (Holiday\_Flag, Temperature, Fuel\_Price, CPI, and Unemployment). Following is the approach followed to analyse each of these features dependency over weekly\_sales.

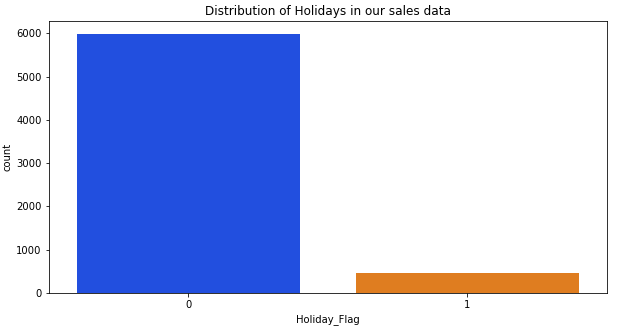
For a given feature

1. Take the frequency-distribution of a given feature and get the inference form that distribution
2. Checking the majority range of a feature, in our sales data (by using box-plot) and also seeing, if there are any outliers in it.
3. Plotting the behaviour of Weekly\_Sales over a given feature and getting inference from it.

Following is the correlation of different features with weekly\_sales and its inferences:

1. Holiday Flag:

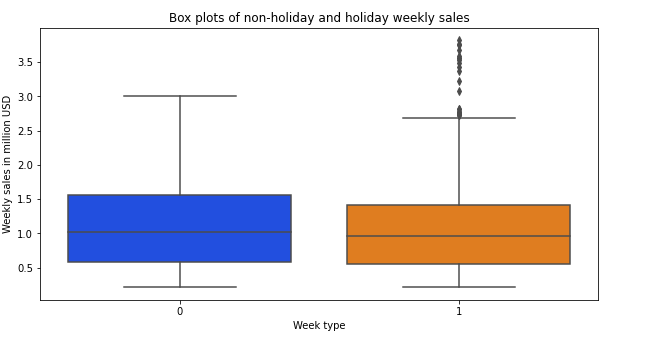
*Frequency Distribution:*



For 45 stores, Around 450 days are Holidays and 5985 days are not holidays in our sales data.

It means around 7% of our days are holidays for each store.

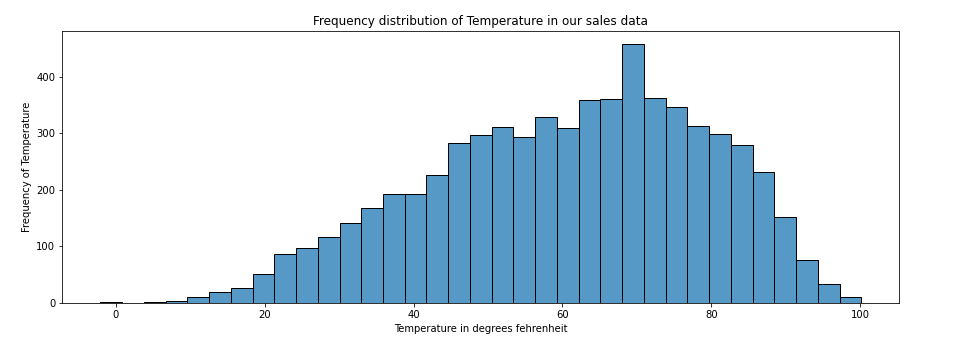
*Plot Weekly Sales for different range of Holiday Flag*

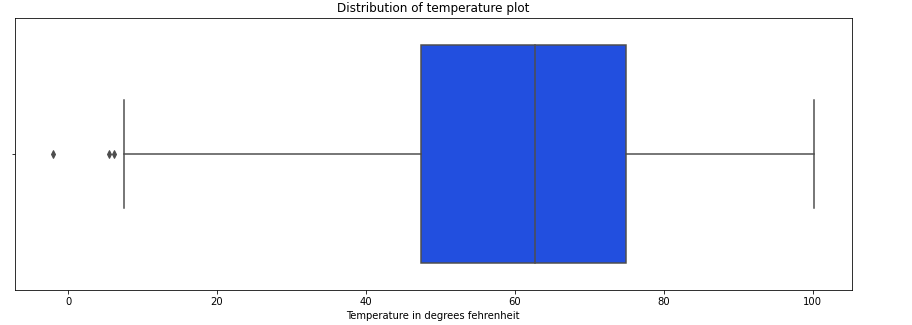


We can see that both holiday and non-holiday weekly sales have similar spread. However, the bigger sales happen during the holiday weeks.

ii) Temperature:

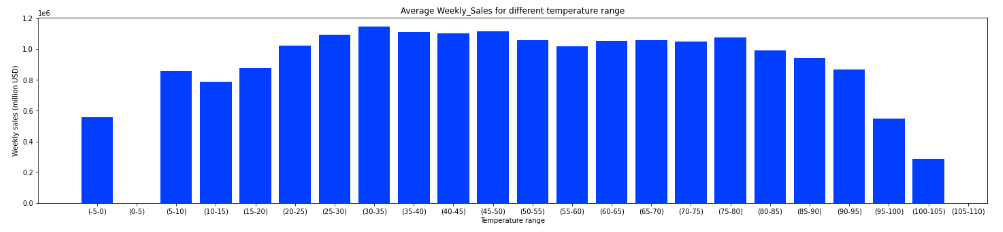
*Frequency Distribution:*





Majority of the temperature lies on 45 to 75 degrees Fehrenheit, with mean temperature value as 60

*Plot Weekly Sales for different range of Temperature*



For temperature range of 20-80 (which is the majority frequency of temperature in our sales data), the sales is around 1.0 to 1.1 M$.

The exception cases are when the temperature is in extreme ranges

Only for

'(-5,0)': 0.8M$ ,

'(80-85)': 0.99M$,

'(85-90)': 0.94M$,

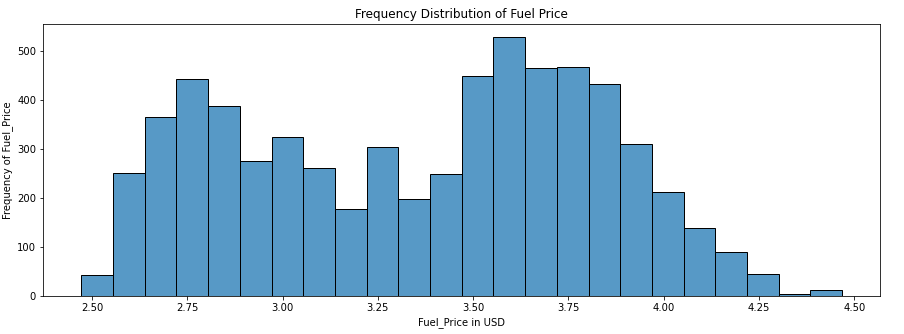
'(90-95)': 0.86M$,

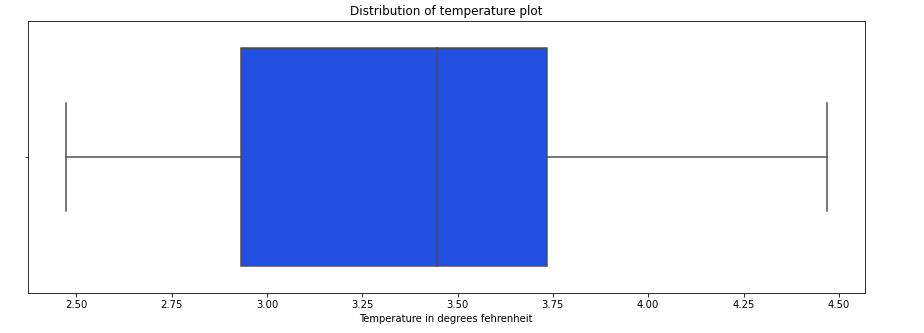
'(95-100)': 0.54M$,

'(100-105)': 0.28M$,

iii) Fuel Price:

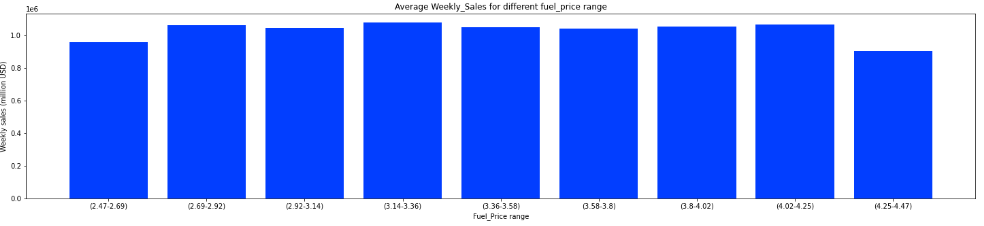
*Frequency Distribution:*





Majority of the Fuel\_Price lays around 3.0 to 3.75

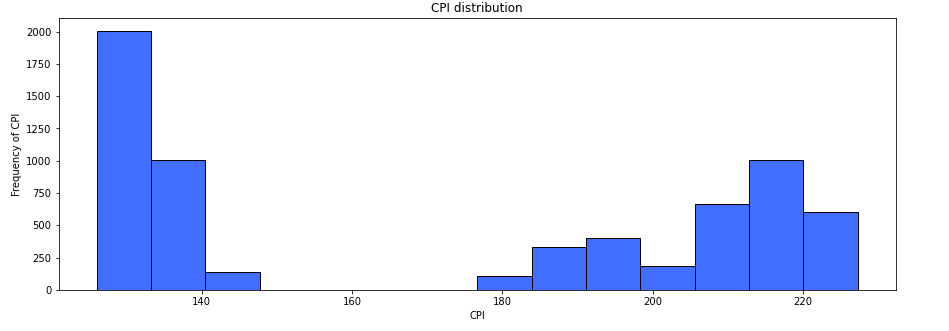
*Plot Weekly Sales for different range of Fuel\_Price*

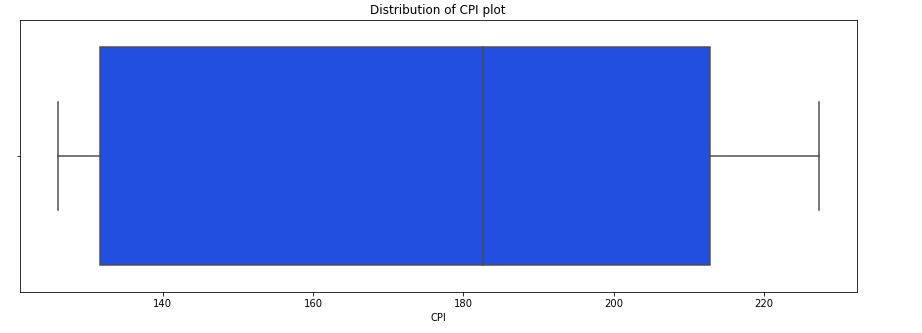


There is not much variation in weekly\_sales over Fuel\_Price range it ranges around (1.04-1.08) M$. But in optimum temperature conditions the sales are at higher when compared to extreme conditions.

iv) CPI (Consumer Price Index):

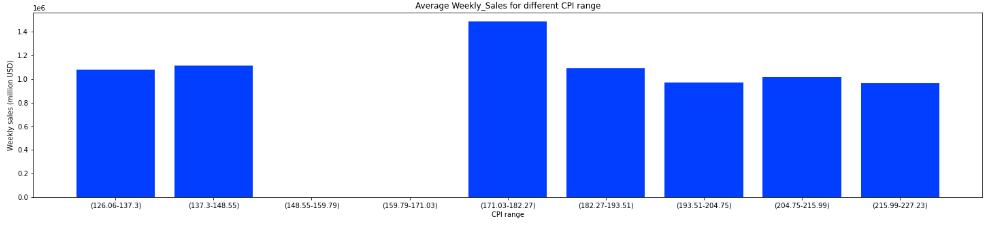
*Frequency Distribution:*



**

Majority of CPI distribution lies in 130 to 140 (approx.) in our sales data, However there is few distribution of CPI around 180-220 (few in count) . With mean value as 171.5

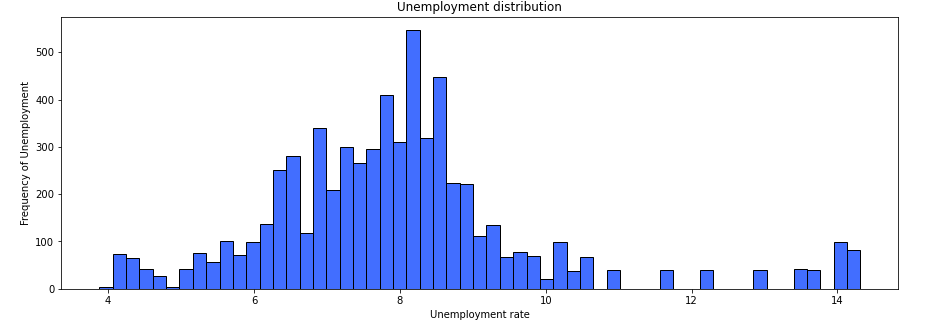
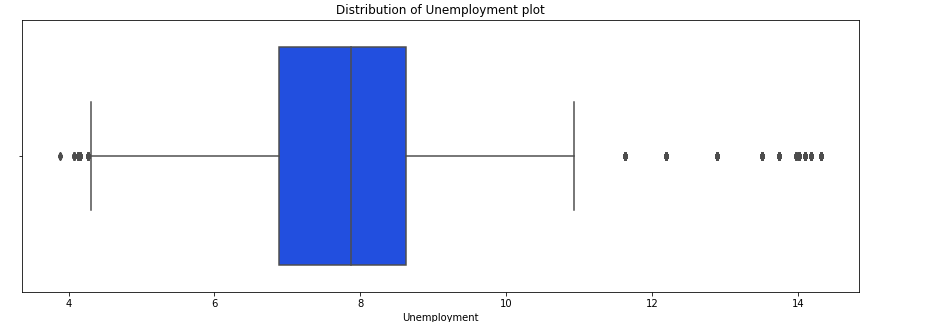
*Plot Weekly Sales for different range of CPI*



Only in case of CPI range (approx.) (171-182) Weekly\_sales is highest with 1.48M USD, for rest of the cases weekly sales is round 1M USD. The least is weekly\_sales value is from (215-227) as 0.96M USD

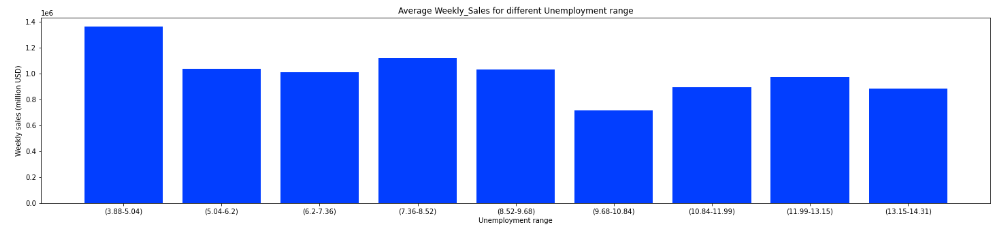
v) Unemployment:

*Frequency Distribution:*

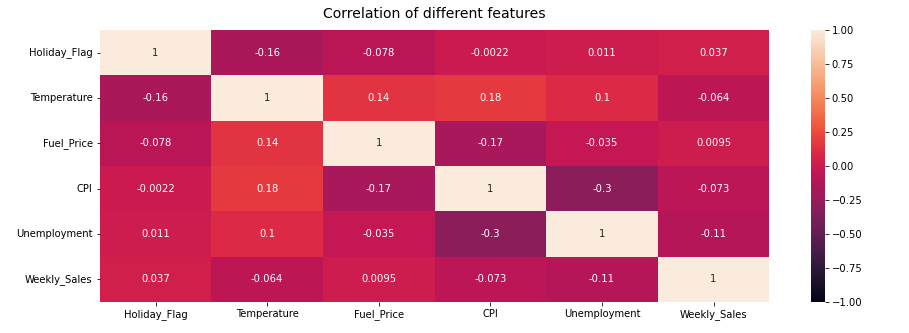
 

Majority of the Unemployment distribution lies in 6.9 to 8.6 with mean value as 8.0 (approx.)

*Plot Weekly Sales for different range of Unemployment*

Here majority of the Unemployment Distribution count in our sales data lays around 8.0, and For lower unemployment rate, there is higher sales (1.4 M USD) when compared to higher unemployment rate(1.02 M USD for 9.0 unemployment and 0.7 M USD for 10.0 unemployment).

Overall Correlation of different features



Correlation of each feature with ‘weekly sales’ are

Holiday\_Flag 0.036891

Temperature -0.063810

Fuel\_Price 0.009464

CPI -0.072634

Unemployment -0.106176

**Conclusion of correlation of features over weekly sales:**

After correlation of overall weekly\_sales we can say that, Holiday\_Flag Has less significance very weekly\_sales, but we saw that for Holidays, the weekly\_sales grows significantly high Temperature Here Weekly\_Sales decreases over increase in temperature Fuel\_Price It has Very less significance on weekly\_sales , however weekly\_sales increases upon increase in fuel\_price CPI Here Weekly\_Sales decreases over increase in temperature Unemployment : It has relatively higher significance (when compared to other features) on weekly\_sales and it decreases upon increase in unemployment

**Conclusion of each Feature dependency on Weekly\_Sales**

**Holiday\_Flag** : -- Around 450 days are Holidays and 5985 days are not holidays in our sales data. Which means around 7% of our days are holidays.

-- We can see that both holiday and non-holiday weekly sales have similar spread. However, the bigger sales happen during the holiday weeks.

**Temperature** : -- Majority of the temperature lies on 45 to 75 degrees Fehrenheit, with mean temperature value as 60 -- for temperature range of 20-80 (which is the majority frequency of temperature in our sales data), the sales is around 1.0 to 1.1 M$. The exception cases are when the temperature is in extreme ranges only for

'(-5,0)': 0.8M USD , '(80-85)': 0.99M USD, '(85-90)': 0.94M USD, '(90-95)': 0.86M USD, '(95-100)': 0.54M USD, '(100-105)': 0.28M USD,

**Fuel\_Price** : -- Majority of the Fuel\_Price lies around 3.0 to 3.75 -- There is not much variation in weekly\_sales over Fuel\_Price range it ranges around (1.04-1.08)M$. BUt in optimum temperature conditions the sales are at higher when compared to extreme conditions.

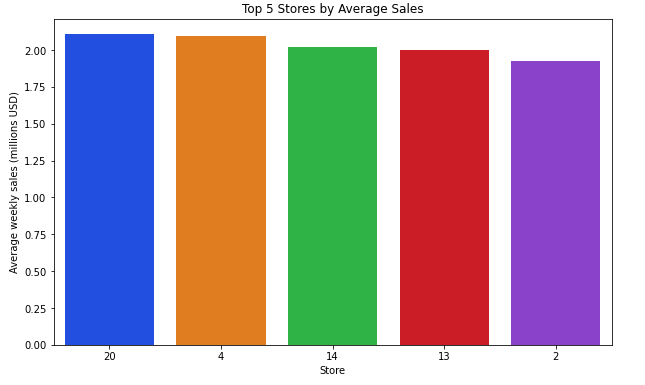
**CPI** : -- Majority of CPI distribution lies in 130 to 140 (approx.) in our sales data, However there is few distribution of CPI around 180-220 (few in count) . With mean value as 171.5 -- Only in case of CPI range (approx.) (171-182) Weekly\_sales is highest with 1.48M USD, for rest of the cases weekly sales is round 1M USD. The least is weekly\_sales value is from (215-227) as 0.96M USD

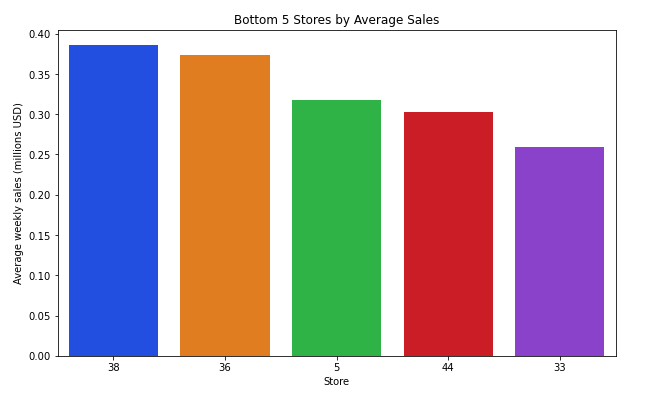
**Unemployment** : -- Majority of the Unemployment distribution lies in 6.9 to 8.6 with mean value as 8.0 (approx.) -- Here majority of the Unemployment Distribution count in our sales data ies around 8.0, and For lower unemployment rate, there is higher sales (1.4M USD) when compared to higher unemployment rate(1.02M USD for 9.0 unemployment and 0.7M USD for 10.0 unemployment)

Exploratory Data Analysis

**Stores that has the highest and lowest average revenues over the years**

Identifying the top performing and low performing stores or products in sales analysis can be useful for a variety of purposes. By analysing the sales data for different stores, businesses can identify opportunities for growth, understand customer preferences, optimise inventory levels, and identify potential problems or areas for improvement. Understanding the performance of different stores can inform product development and marketing efforts, as well as help businesses allocate resources more effectively and make more informed business decisions.

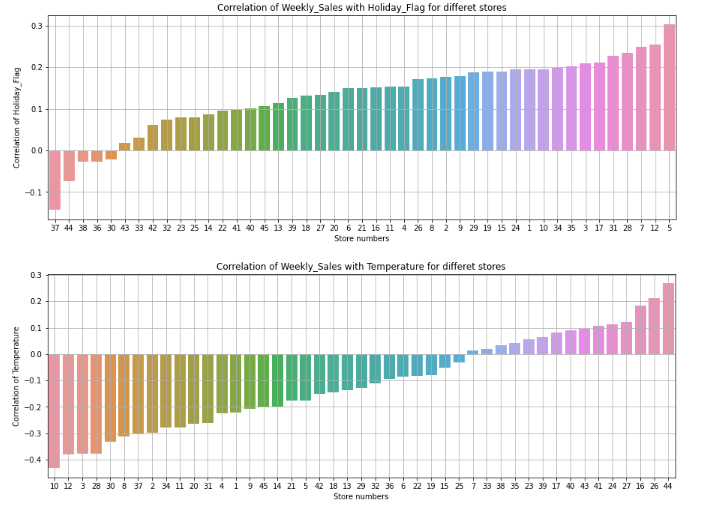
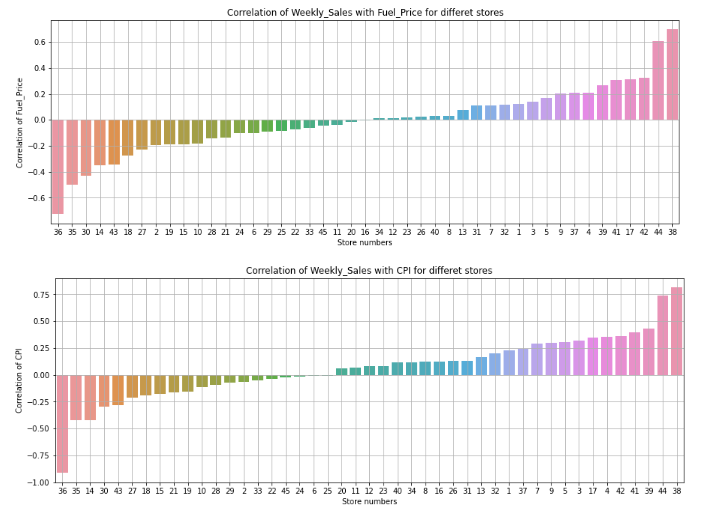
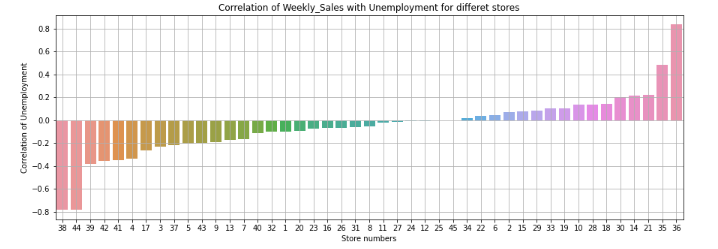




The graphs show that the top performing stores have relatively stable sales with an average of around $2 million USD. Store 20 appears to be the top performer among these stores, with relatively little variation in sales compared to the other top performers.

On the other hand, the lowest performing stores have higher variations in sales, with the highest sales at around $0.38 million USD. This suggests that there may be more variability in the sales performance of these stores.

**Correlation of features with the target variable for different stores:**

Insights for each store

Correlation of features with the weekly sales for different stores is taken. And following are the inference of sales of different stores with respect to 5 features.

**Holiday:**

For store 37, 44, 28, 36, 30 there is a negative impact of sales when there are holidays, but the 41 stores the sales are high in holidays.

**Temperature:**

For 15 stores, the sales are high in summer season, but for the rest of the stores the sales are high in winter season

**Fuel Price:**

For 20 stores the sales are low, when the fuel prices are high, but for the rest of 25 stores the sales are high when the fuel prices are high.

**CPI:**

Most of stores are not almost affected by CPI. But for store number 36, 35, 14 the stores there sales are less when there is high unemployment, and for store number 44 and 38 the CPI is high when the sales are high.

**Unemployment:**

For 25 stores the sales are low, when the unemployment rate is high. But for the rest of the sales are high when there is more unemployment.

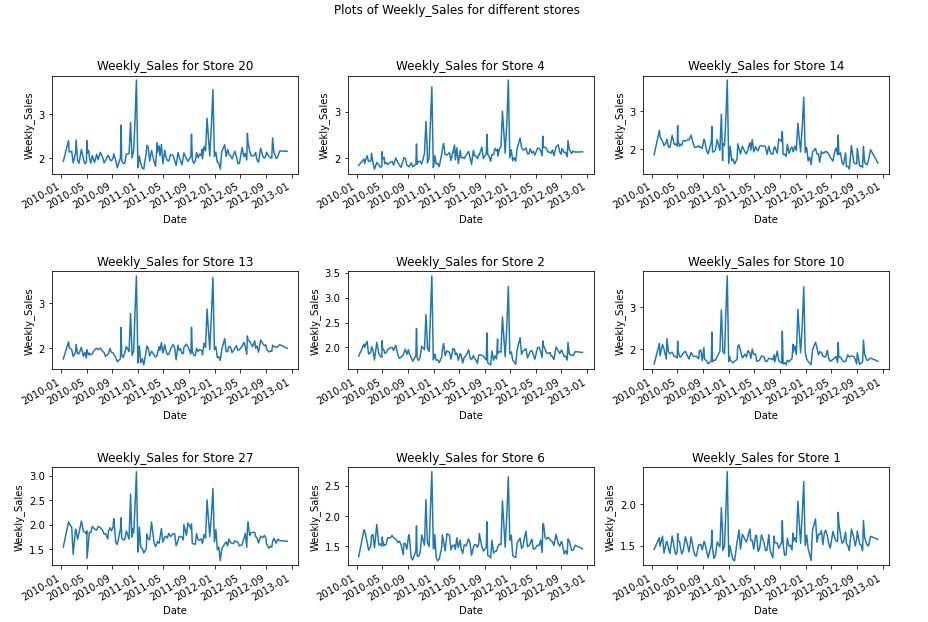
Most of stores are not almost affected by unemployment. But for store number 38, 44 the stores there sales are less when there is high unemployment, and for store number 35 and 36 the unemployment is high when the sales are high.

Temperature effect on weekly sales

Evaluation on how changes in temperature affect the weekly sales. The weekly Sales revenue is presented below.

|  |  |  |
| --- | --- | --- |
| Store | Outlook – Recommendation(s) | |
| 1 |  | Most Sales in summer warm weather months – must shore up inventory for summer |
| 2 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 3 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 4 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 5 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 6 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 7 | Most Sales in Winter warm weather months – must shore up inventory for winter |
| 8 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 9 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 10 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 11 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 12 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 13 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 14 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 15 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 16 | Most Sales in winter warm weather months – must shore up inventory for winter |
| 17 | Most Sales in winter warm weather months – must shore up inventory for winter |
| 18 | Most Sales in summer warm weather months (Fall & Summer) – must shore up inventory for warm weather |
| 19 | Most Sales in summer warm weather months (Fall) – must shore up inventory for summer |
| 20 | Most Sales in summer warm weather months (summer) – must shore up inventory for summer |
| 21 | Most Sales in summer warm weather months (summer) – must shore up inventory for summer |
| 22 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 23 | Most Sales in summer warm weather months – must shore up inventory for summer |
| 24 | Most Sales in summer warm weather months (Fall & Summer) – must shore up inventory for warm weather |
| 25 | * Mild sales in very cold weather * Significant sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for cold and warm weather |
| 26 | * Most sales in very cold weather * Significant sales in cold and warm weather * No sales in warm weather * Must shore up inventory for cold and warm weather |
| 27 | * Mild sales in very cold weather * Moderate sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for hot weather |
| 28 | * No sales in very cold weather * Moderate sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for hot weather |
| 29 | * Mild sales in very cold weather * Significant sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for hot weather |
| 30 | * No sales in very cold weather * Mild sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for hot weather |
| 31 |
| 32 | * Moderate sales in very cold weather and warm weather * Most sales in hot weather * Must shore up inventory for hot weather |
| 33 | * No sales in very cold weather * Mild sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for hot weather |
| 34 | * Mild to Moderate sales in very cold weather * Moderate sales in warm and warm weather * Most sales in hot weather * Must shore up inventory for hot weather |
| 35 |
| 36 | * No sales in very cold weather * Mild sales in cold and warm weather * Most sales in warm weather * Must shore up inventory for hot weather |
| 37 |
| 38 |
| 39 |
| 40 | * Significant sales in very cold weather * Moderate sales in winter and warm weather * Most sales in warm weather (early summer) * Must shore up inventory for cold and warm weather |
| 41 | * Moderate sales in hot and warm weather * Most sales in cold weather * Must shore up inventory for cold and warm weather |
| 42 | * No sales in very cold weather * Moderate sales in warm weather * Most sales in hot weather (summer) * Must shore up inventory for hot weather |
| 43 |
| 44 | * Mild sales in very cold weather * Moderate sales in warm weather and early winter * Most sales in hot weather * Must shore up inventory for hot weather |
| 45 |

Model Evaluation and Techniques



**Model Selection:**

Examination of the plot of the target feature, weekly\_sales (as shown above) shows continuously time varying data.

A Time Series (TS) model (ARIMA, SARIMA and SARIMAX) will be employed for the predictions and forecast. Attempt will also be made to use Linear-Regression models (Gradiant\_Boosting, Linear Regression, Random\_Forest) for prediction and compare the result with the TS prediction.

1. **The ARIMA model:**

Autoregressive Integration Moving Average (ARIMA) is defined as a statistical analysis model that uses time series data to either better understand the data set or predict future trends [[1]](#Reference_1)

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 term and parameters are constant,
        3. Historic time points dictate behaviour of present time points,
        4. Time series is stationary

1. **Regression models:**
2. **Gradient Boosting:** Gradient boosting stands out for its prediction speed and accuracy, particularly with large and complex datasets [[4]](#Reference_4). The algorithm has produced the best results from kaggle competitions and machine learning solutions for business. In machine learning algorithm, two types of errors, (ie) bias error and variance error. Gradient boosting algorithm is based on minimizing the *bias error* of the model. The gradient boosting algorithm is based on building models sequentially where the subsequent models try to reduce the errors of the previous models. The subsequent models are built on the errors or residuals of the previous models. The process is repeated until there is no more significant change on the errors.
3. **Linear Regression:** is a basic predictive analysis technique that se historical data to predict an output variable [[2]](#Reference_2). It is a popular algorithm employed to predict continuous (dependant) variables such as price, based on their correlation with other independent variables. It is based on the following:
4. **Linear Relationship:** The relationship between the independent and dependent variables should be linear.
5. **Multivariate Normal:** Allthe variables together should be multivariate normal, which means that each variable separately has to be univariate normal means, a bell-shaped curve.
6. **No Multicollinearity:**  There is little or no multicollinearity in the data which means that the independent variables should have minimal correlation with each other.
7. **No Autocorrelation:** There is a little or no autocorrelation in the data where the data values of the same columns are related to each other.
8. **Random Forest:**  Random Forest is acommonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. It ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems [[3]](#Reference_3).

Model Evaluation and Techniques

**Model Evaluation**

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

1. Load necessary libraries
2. Load the dataset
3. Perform the exploratory data analysis (EDA) of the dataset
   1. Find the shape of the data
   2. Check for invalid and null entries
   3. Explore data description
   4. Examine the correlation of the independent variables to the target (weekly\_sales) variable
   5. Line plot of the effect of the independent variables on the target variable
   6. Box plot of the features to identify the outliers
4. Model Prediction
5. Forecast

Model Evaluation and Techniques

**Model Design:**

It was observed from ‘Correlation of features with target variable for different stores’ that the EDA that the effect of the independent features (unemployment, CPI, Temperature, Fuel\_Price, Holiday\_flag) on the target variable, weekly\_sales differ greatly by the store. For example, as shown above, the effect of unemployment varies by the store where it appears to have positive effect on some and negative effect 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 prediction by the stores as the single prediction for all the stores may not be reasonable given the peculiar conditions prevalent in each region of the store.

For simplicity and ease of presentation, I have also decided to limit any prediction for the five stores with the highest weekly\_sales revenue. That notwithstanding, the model could always be used to provide predictions for each of the store.

Model Evaluation and Technique

**Model Approach:**

1. **TS model, ARIMA:**

* The first step of this model is to check the stationarity of the dataset.
* Next is to find the best ARIMA order of the model
* Using the best ARIMA order, make predictions for the selected stores.
* Forecast using SERIMAX.
  + - Detrend the dataset if necessary.
    - Using SERIMAX estimate a 12-weeks weekly sales forecast

1. **Regression model:**

Regression model, Gradient\_Boosting, LinearRegression and Random\_Forest models were also used for the prediction. The best of the three predictions will then be compared to the predictions by ARIMA model predictions.

Inference from the Project

**Model result**

1. ARIMA model:

**a. Predictions:**

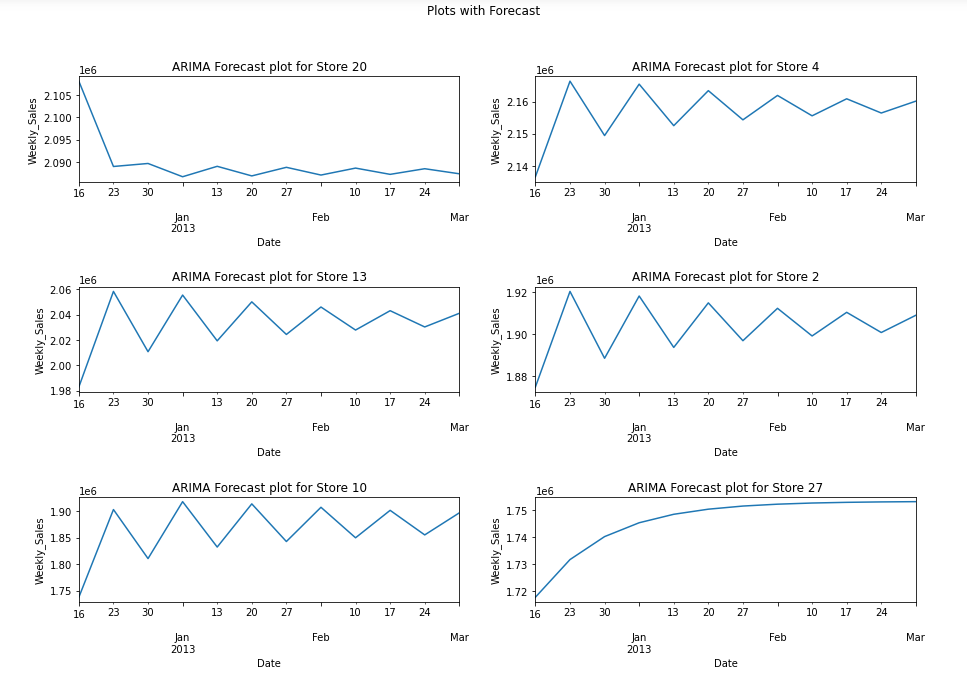
Predictions were performed for six stores (store: 20, 4, 13, 2, 10 and 27 in order of decreasing weekly\_sales revenue). The predictions results are summarized in the Table and graphs below:

****



**b. Forecast:**

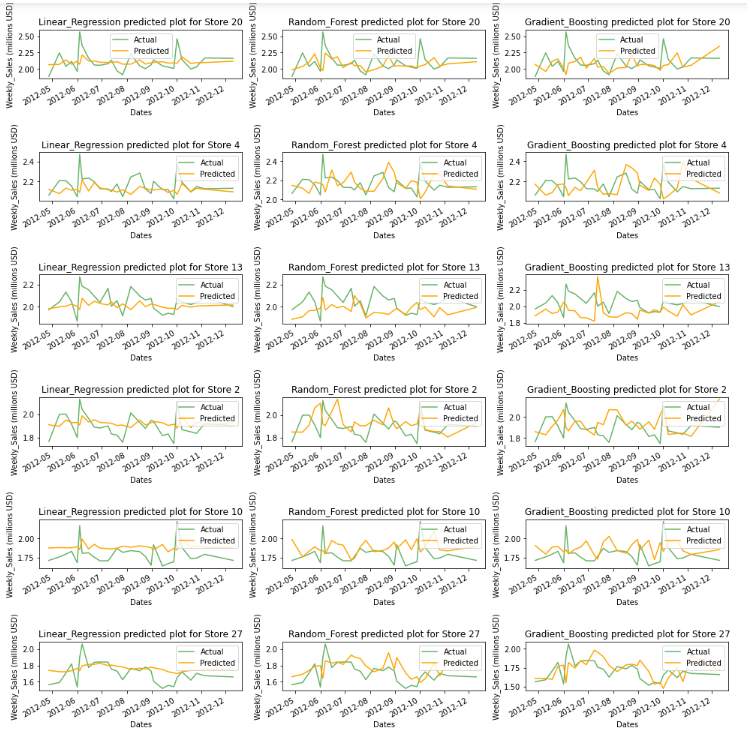
Following areforecast, for 6 different stores, and pattern looks quite similar for them. However, there is change in the weekly sales attained for different stores for given week.



Inference from the project

1. **Regression model:**

The prediction results from the three chosen regression models: Gradient\_Boosting, LinearRegression, and Random Forest are summarized in the chart shown:



Inference from the Project

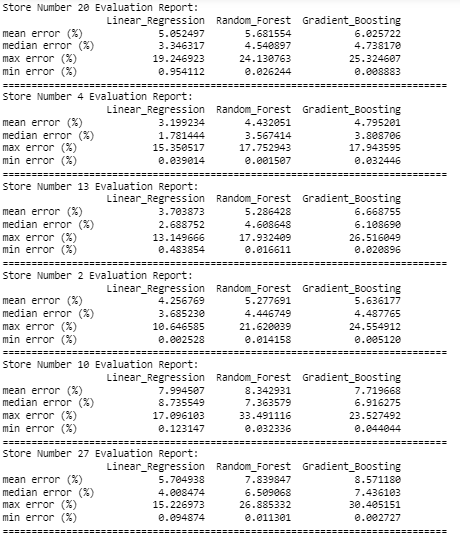
**Model Evaluation:**

**1. ARIMA/SERIMAX models:**

The model predictions for the selected stores were okay showing variability of the weekly sales inline with the sales history.

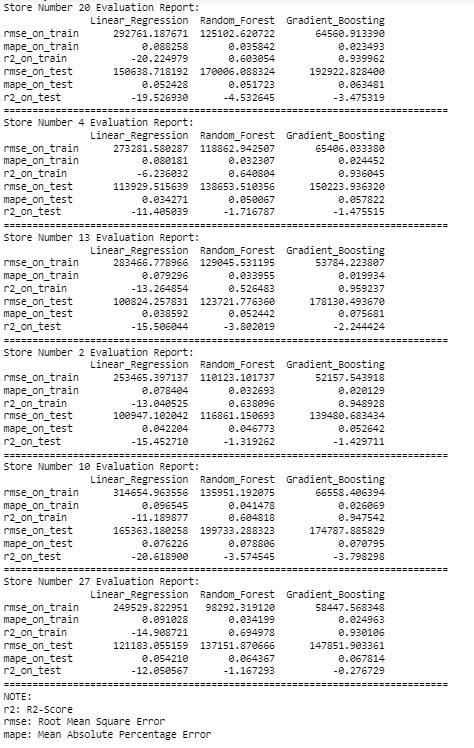
**2. The Regression model:**

The regression model result is summarized below (Table & chart):

****

As seen, the mean percentage error is from 3.1% to 7.9% for all the models which is within acceptable range. As seen in the prediction report Table, the results from the three models are very comparable.

The summary of evaluation report (r2, root\_mean\_square\_errors and mean\_absolute\_percentage\_errors) is present in the table below.



It is noted that, the r2-Score (for both train and test data), which is a measure of the goodness to fit of the model to the data is negative for all the models. However the mean\_square\_errors and mean\_absolute\_percentage\_errors (for both train and test data) which is a measure of accuracy of the model is very good for all the models.

**Comparing the Models:**

Since Gradient\_Boosting has minimum rmse value it is the best among the three Regression models. Following the comparison of prediction of ARIMA and Gradient\_boosting model.

|  |  |
| --- | --- |
| Prediction of ARIMA | Prediction of Gradient Boosting |
|  |  |

**Future Possibilities:**

The future of machine learning is exceptionally exciting. At present almost, every common domain is powered by machine learning applications.

By enabling enterprises to better understanding both the customers’ and business functionality behaviour, Machine Learning has enabled companies to offer better/ targeted customer services leading to more loyal customers and ultimately improved sales revenue.

**Conclusion**

The project undertook the study of a retail company with 45 outlets stores. Some of the important findings from the report include the followings:

1. Sales revenue projections for the next 12 weeks are down for most of the stores.
2. Some of the stores have very week or no sales activities, during some period of year.
3. To improve sales revenue, the following steps are recommended:
   1. Concrete efforts by the company to find out through local market surveys and past sales records what products are in high demand by the local population at any given period of the year and make efforts to republish those stocks.
   2. Create increased local awareness of the products on offer at each store through commercial outreach: social media, television commercials and print media that could help improve sales.
   3. Have detailed records of inventory of the items on offer at each store indicating amount and dates if sold as it is needed for effective inventory tracking.
   4. Explore other services options that have worked well for similar companies, such as same-day or next-day home delivery.

It may just be that some stores may just have to be wound up if sales revenue does not improve.

**References:**

1. Autoregressive Integrated moving average defined: autoregresor Integrator moving average (ARIMA) definition ([Investopedia link](https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp)).
2. Introduction to Linear regression in Python: ([Towards data-science link](https://towardsdatascience.com/linear-regression-explained-89cc3886ab48#:~:text=Linear%20regression%20is%20probably%20the,simplicity%2C%20speed%2C%20and%20interpretability.))
3. Random forest : ([Random forest ‘Towards Data-science link’](https://towardsdatascience.com/understanding-random-forest-58381e0602d2))
4. Gradient Boosting: ([Gradient Boosting ‘Towards Data-science link’](https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab))

**Appendix**

There is Code file named ‘Capstone project for walmart.ipynb’ is attached to the zip file, as a reference, to see the functions used to evaluate the result in this report.