Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard

By:

Divya Kumari

(May 2023)

Introduction

1.1 Project Goals and Background

The 21st century has seen an outburst of data that is being generated as a result of the continuous use of growing technology. Retail giants like Walmart consider this data as their biggest asset as this helps them predict future sales and customers and helps them lay out plans to generate profits and compete with other organizations. Walmart is an American multinational retail corporation that has almost 11,000 stores in over 27 countries, employing over 2.2 million associates (Wikipedia, n.d.). Catering to their customers with the promise of 'everyday low prices', the range of products sold by Walmart draws its yearly revenue to almost 500 billion dollars thus making it extremely crucial for the company to utilize extensive techniques to forecast future sales and consequent profits. The world's largest company by revenue, Walmart, sells everything from groceries, home furnishings, body care products to electronics, clothing, etc. and generates a large amount of consumer data that it utilizes to predict customer buying patterns, future sales, and promotional plans and creating new and innovative in-store technologies. The employment of modern technological approaches is crucial for the organization to survive in today's cutting-edge global market and create products and services that distinguish them from its competitors. The main focus of this research is 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 under study. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue. 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. Additionally, the application of big data analytics will help analyze past data efficiently to generate insights and observations and help identify stores that might be at risk, help predict as well as increase future sales and profits and evaluate if the organization is on the right track. The analysis for this study has been done using SQL, R, Python, and Power BI on the dataset provided by Walmart Recruiting on Kaggle ("Walmart Recruiting - Store Sales Forecasting," 2014). The modeling, as well as the exploratory data analysis for the research, have been performed in R and Python, aggregation and querying will be 8 performed using SQL and the final dashboard has been created using Power BI.

Project Deliverables

The main objective of this study is to predict weekly sales for Walmart stores and create a Power BI dashboard that tracks the final predicted sales until 2013 through interactive and immersive visualizations. The conclusion section highlights the findings from the exploratory data analysis as well as from the models implemented as part of this study. The dashboard created compares the findings from the Exploratory Data Analysis with the findings from the dashboard. The user of the dashboard can filter data based on stores, departments, store type, store size, week of the year, holiday versus nonholiday sales, etc.

Tools and Technologies Applied

The analysis for this study have been performed using some main tools: R, Python, and Power BI. The models and Exploratory Data Analysis have been executed using development tools like R Studio and PyCharm. Several packages have been used to perform the initial and final outcome EDA for the analysis. For the initial EDA, a combination of R and Python libraries like inspectdf, ggplot2, plotly, caret, matplotlib, seaborn, etc have been implemented. Packages like numpy, pandas, tidyverse, etc. have been used for data wrangling and manipulation. For the models that have been created, several packages like 'scikit-learn', 'xgboost', etc have been applied.

Purpose Statement

The purpose of this study is to predict the weekly sales for Walmart based on available historical data (collected between 2010 to 2013) from 45 stores located in different regions around the country. Each store contains a number of departments and the main deliverable is to predict the weekly sales for all such departments. The data has been collected from Kaggle and contains the weekly sales for 45 stores, the size and type of store, department information for each of those stores, the amount of weekly sales, and whether the week is a holiday week or not. There is additional information in the dataset about the factors that might influence the sales of a particular week. Factors like Consumer Price Index (CPI), temperature, fuel price, promotional markdowns for the week, and unemployment rate have been recorded for each week to try and understand if there is a correlation between the sales of each week and their determinant factors.

Correlation testing has been performed to understand if there is a correlation between the individual factors and weekly sales and whether such factors have any impact on sales made by Walmart. This study also includes an extensive exploratory data analysis on the provided Walmart dataset to understand the following:

- Identifying store as well as department-wide sales in Walmart
- Identifying sales based on store size and type
- Identifying how much sales increase during holidays

• Correlation between the different factors that affect sales
• Average sales per year
• Weekly sales as per region temperature, CPI, fuel price, unemployment Apart from

identifying these direct relationships between independent and dependent variables, some interaction effects have also been studied as part of the Multiple Linear Regression model to understand if a certain combination of the factors under study can directly impact the weekly sales for Walmart. After employing different algorithms to predict future sales and correlation between factors for the retail store, a dashboard that tracks the above-mentioned outcomes has been created (in Power BI) and also includes the new predictions to collectively visualize the outcomes of this research and present them to amateur users more effectively

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. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

Dataset Information:
The walmart.csv contains 6435 rows and 8 columns.

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

Exploratory Data Analysis

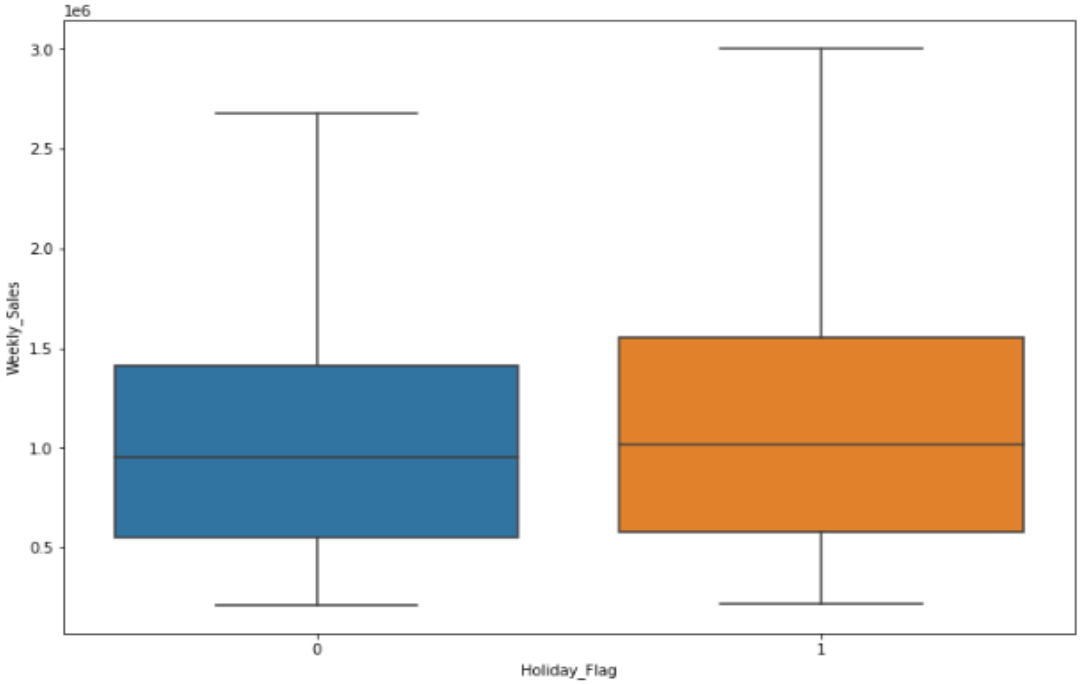
It is crucial to have an in-depth understanding of the dataset that is used in this analysis to understand the models that would give the most accurate prediction. Several times there are underlying patterns or trends in the data that would not be identified as easily, hence the need for an extensive exploratory data analysis. This thorough examination is necessary to understand the underlying structure of the dataset and to draw conclusions or insight about the validity of our analysis. The study is going to begin with a brief analysis of the available dataset to get a sense of the main characteristics and components that are relevant to the research. An exploratory data analysis is crucial to this study considering the numerous attributes that are a part of the dataset that will be essential when trying to draw insights and making predictions. As part of the exploratory data analysis, several visualizations have been created that will help us understand what it is that we are trying to achieve and to keep in mind the various attributes that we can use to improve results. The EDA is like a primary investigation and tries to look at the relationships and nature of the different columns available to us. As part of this, the 'inspectdf' package (Ellis, 2019) and the 'glimpse' package (Sullivan, 2019) have been used and implemented in R that will answer questions related to the number and nature of columns and rows in the dataset, missing values, distribution of numeric and categorical variables, correlation coefficients, etc. Several other packages like 'ggplot2', 'matplotlib', 'seaborn', and 'plotly' have also been used in this study to create visualizations that provide information about weekly sales by store and department, weekly sales on holidays versus on normal days, weekly sales based on region, store type and store size, average sales per year, change in sales as a result of factors like CPI, fuel price, temperature, and unemployment, etc in the form of heatmaps, correlation matrix (Kedia et al., 2013), histograms, scatterplots and several more. These visualizations are accompanied by brief descriptions that will discuss the findings and scope for potential modeling that will be performed in the next stages of this project.

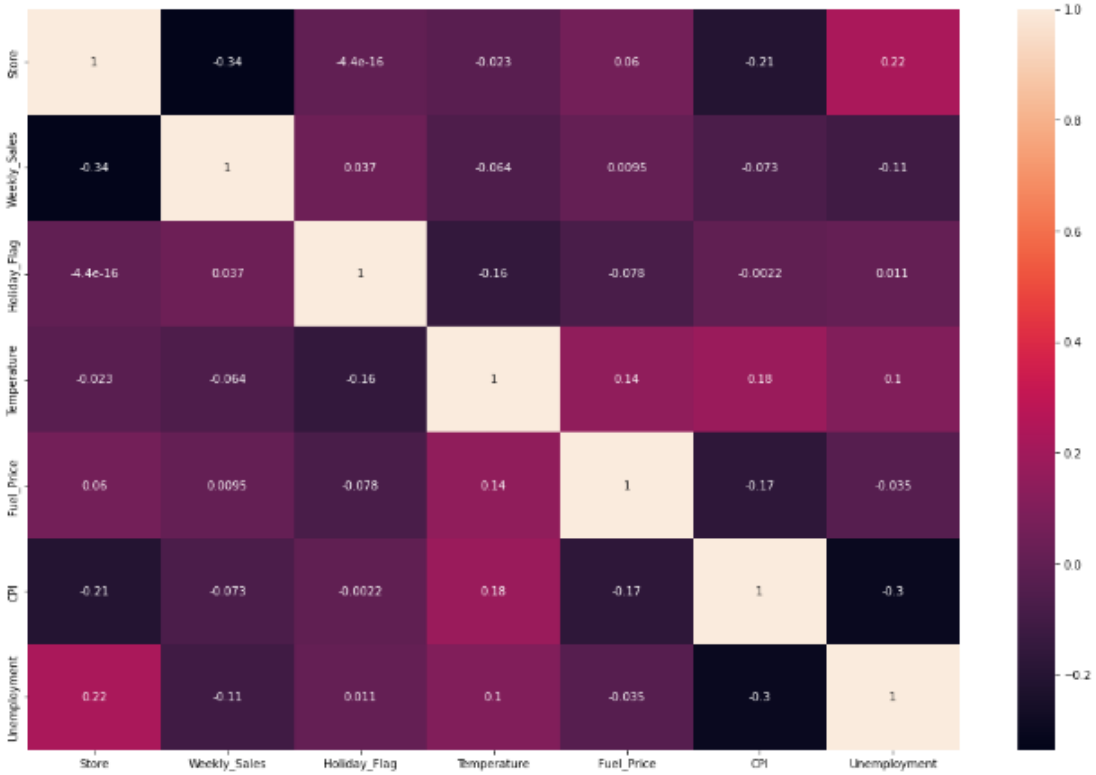
```
1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   int64
 1   Date          6435 non-null   object
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: float64(5), int64(2), object(1)
memory usage: 402.3+ KB
```

In [9]:
```
1 import matplotlib.pyplot as plt
2 import seaborn as sns
3 plt.figure(figsize = (12,8))
4 sns.boxplot(x = 'Holiday_Flag', y ='Weekly_Sales', data = df, showfliers = F
```

Out[9]: <AxesSubplot:xlabel='Holiday_Flag', ylabel='Weekly_Sales'>





1. There are 6435 rows and 8 columns in this dataset.
2. Except date(object) all are numerical datatype.(int and float)
3. There is no null value in the dataset.
4. there is no major impact observed of holidays on weekly sales figures.
5. weekly sales have the highest correlation of 0.34 with the store number, which is in line with our previous findings. The remaining features are mostly uncorrelated with each other
6. This data contains weekly sales figures from Jan 2010 to Dec 2012

The next function from the package looks at the distribution of the numeric variables using histograms created from identical bins. Considering that the features dataset has the most numeric variables, that is the only one that will be looked at in detail. According to the package website, 'The hist column is a list whose elements are tibbles each containing the relative frequencies of bins for each feature. These tibbles are used to generate the histograms when showplot = 'TRUE'. The histograms are represented through heat plot comparisons using Fisher's exact test to highlight the significance of values within the column; the higher the significance, the redder the data label (Rushworth, n.d.).

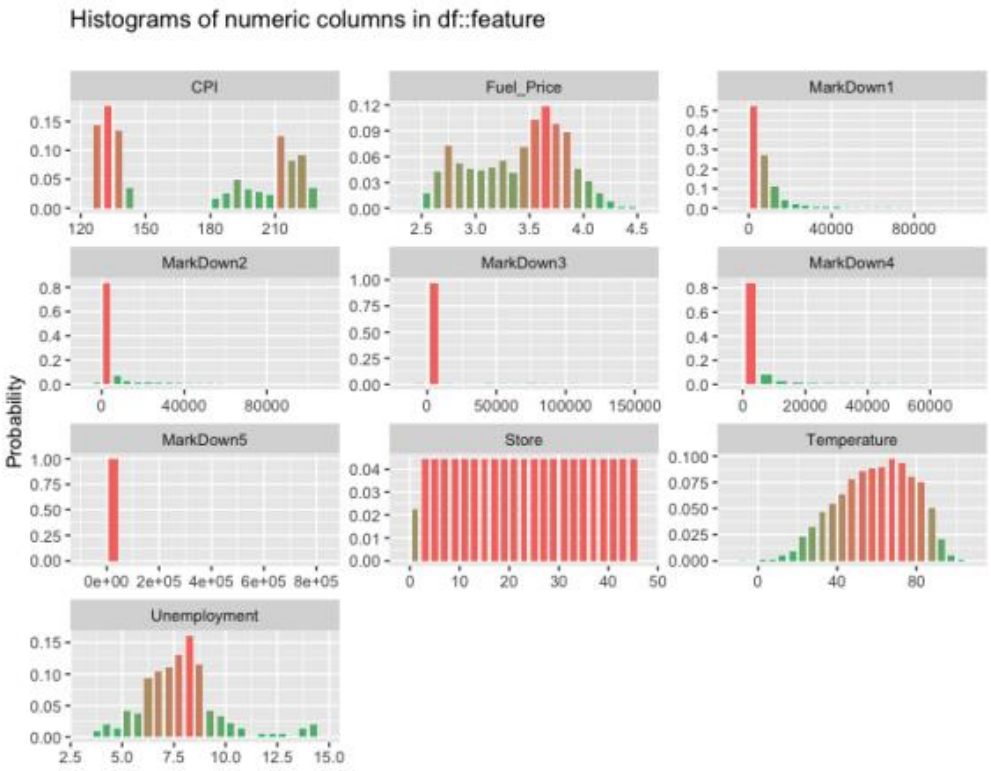
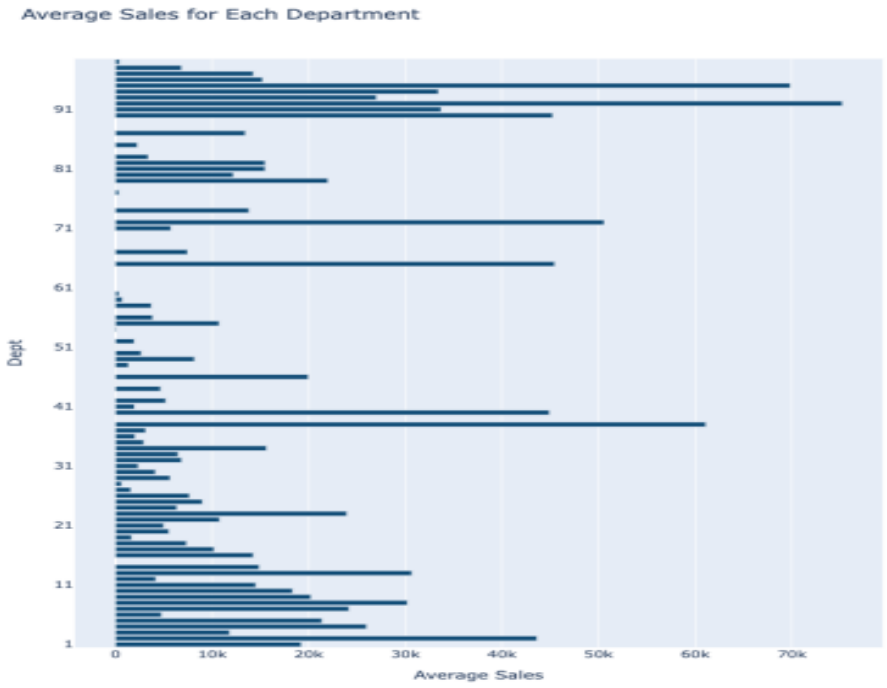


Figure 5. Distribution of Numerical attributes in the Features dataset

Identifying Department-wide Sales:

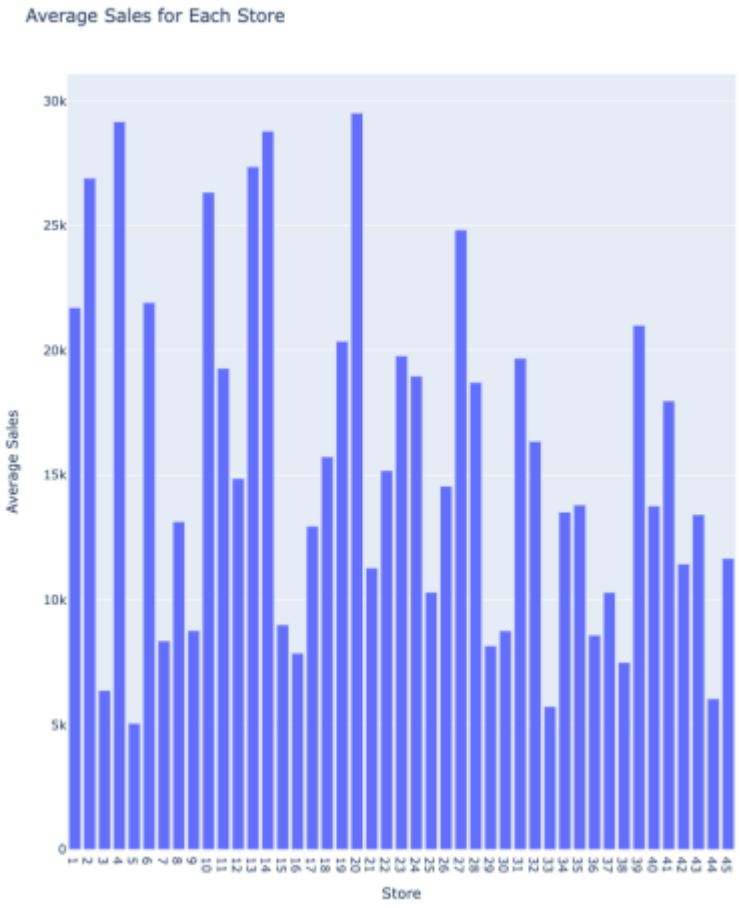
The next step is looking at what departments for each of these stores have the highest average sales. From the image depicted below, it can be identified that Department number 92 has the highest number of average sales.



After looking at the average sales for the stores and departments, it is also imperative to look at the breakdown of sales for both the stores as well as departments. This 22 breakdown will include looking at average sales for each year, week over week sales, monthly sales, week of year sales, etc. Each of these will throw more light on customers' buying patterns over different periods. It will also help in evaluating whether sales go up during certain time periods as compared to normal average sales. I will also look at the average sales for different stores associated with the three store types and evaluate which store number has the highest sales.

Identifying Average Store Sales

There are several stores associated with the three store types and it is crucial to look at the average sales for each of these stores. In total there are 45 stores presented in the Walmart dataset and from the image below it can be concluded that store numbers 4, 14, and 20 have the highest average sales. It should also be noted that there is a very high difference between the average sales for each of the stores; while some stores record huge sales, some others lack vastly in the area. This could be dependent on factors like the kinds of products sold by the store, the geographic location, temperature, unemployment in the vicinity, etc. Some further study reveals that all three of these stores belong to the Store Type A that gather the largest sales out of all three store types.



Identifying Monthly Sales for Each Year

With the holiday information provided in the original dataset, it is known that the major holidays fall at the end of the year. The graph below clearly depicts that the months of November and December recorded the highest average sales for 2010 and 2011. The dataset provided by Walmart contained no weekly sales information for the last two months of the year 2012, hence no conclusion can be drawn for that year. This graph also shows that the month of January tends to have the lowest average sales in the whole year.

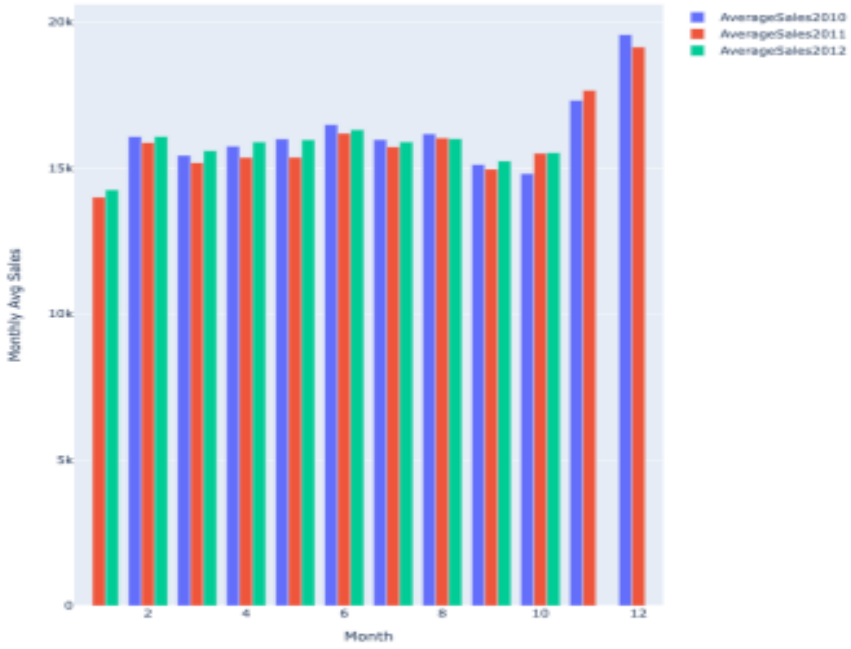


Figure 13. Overall Monthly Sales

Identifying Week Over Week Sales for Each Year

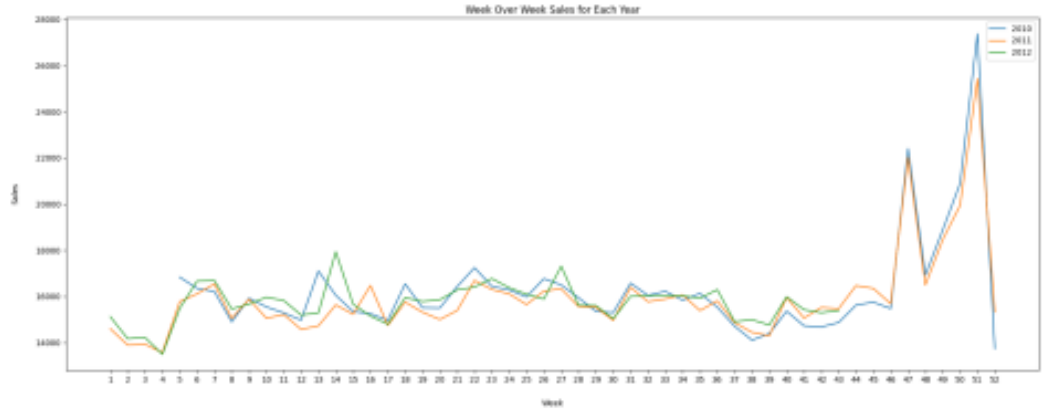


Figure 14. Week Over Week Sales

The week over week overview again helps us in understanding if there is an increase in sales during holiday weeks each year, i.e. the weeks of Thanksgiving, Christmas, Labor 27 Day, etc. There is an evident hike in sales in weeks 47 and 51 that correspond to Thanksgiving and Christmas respectively, proving again that sales rise during the holiday season. Due to the insufficiency of data for the year 2012, these conclusions have only been made based on the data available from 2010 and 2011. This graph also tells that there is a distinguished pattern of decline immediately following Christmas and New Year's. After studying the overall sales summaries of different components of the Walmart dataset, this report will now throw light upon the effect of different factors (such as holidays, markdowns, CPIs, unemployment, etc.) on the weekly sales. It has always been an integral part of this study to understand the effect that these factors have on Walmart's sales performance. I will also create several visualizations that shed light on the difference in Walmart store sales on holidays versus non-holiday days, the impact of store size and type on weekly sales, and finally create a correlation matrix to examine the correlation between the many factors included in the study.

Data Cleaning and Preprocessing

The data contains 421,570 rows, with some store-specific departments missing a few too many weeks of sales. As observed in Figure 4, some columns in the features dataset contain missing values, however, after the features dataset is merged with the training dataset, the only missing values that exist are in the Markdown columns (as shown in figure 23). After the extensive EDA, it was determined that these five markdown files, with missing values, have barely any correlation to the weekly sales forWalmart, hence these five columns have been eliminated from the subsequent training and testing dataset. Because the source already provides training and testing datasets, there is no need to create them for our study. Because the main focus of this study is to accurately predict weekly sales for different Walmart stores,

the previously modified ‘Date’, ‘Month’, ‘Quarter’, and ‘Day’ columns have been dropped and only the ‘Week of Year’ column has been used in the upcoming models.

Model Selection and Implementation

Trying to find and implement the most effective model is the biggest challenge of this study. Selecting a model will depend solely on the kind of data available and the analysis that has to be performed on the data (UNSW, 2020). Several models have been studied as part of this study that were selected based on different aspects of our dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, the following four machine learning models have been used:

- KNN
- Decision Tree
- Gradient Boosting Machine
- Random Forest

Developing the ML models

Since we already have the target variable with us, that means it is a supervised machine learning.

Dependent variable is a regression hence here we are going to use

KNN, Random forest, Decision tree, XGBoost, and will check the accuracy in all cases.

In this project, we will use the coefficient of determination (accuracy score or r-squared score), mean absolute error (MAE), and root mean squared error (RMSE) scores to compare and evaluate the performance of the ML models.

KNN:

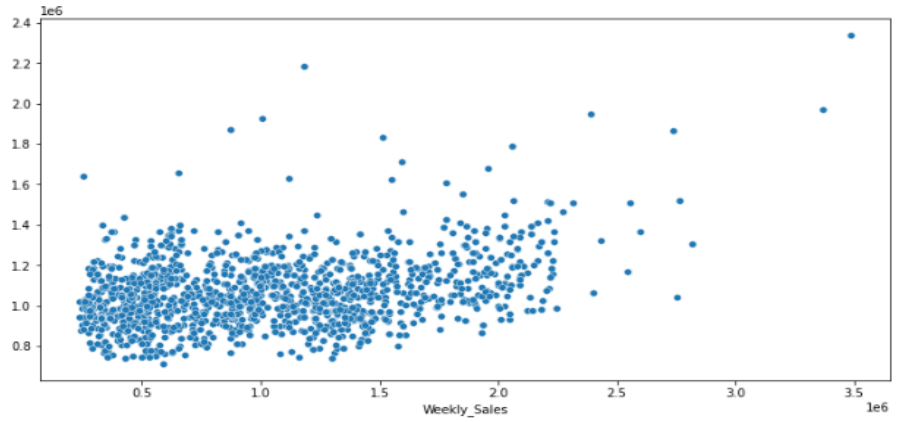
In [20]:
```
from sklearn.neighbors import KNeighborsRegressor
knn_regressor = KNeighborsRegressor(n_neighbors = 20, n_jobs = 4)
knn_regressor.fit(x_train, y_train)
y_pred = knn_regressor.predict(x_test)

print(f'MAE is - {metrics.mean_absolute_error(y_test, y_pred)}')
print(f'RMSE is - {np.sqrt(metrics.mean_squared_error(y_test, y_pred))}')
print(f'Accuracy Score is - {knn_regressor.score(x_test, y_test)}')
```

```
MAE is - 438002.2962498057
RMSE is - 516668.4662414762
Accuracy Score is - 0.10985696142125201
```

In [21]:
```
plt.figure(figsize = (12,6))
sns.scatterplot(x = y_test, y = y_pred)
```

Out[21]: <AxesSubplot:xlabel='Weekly_Sales'>



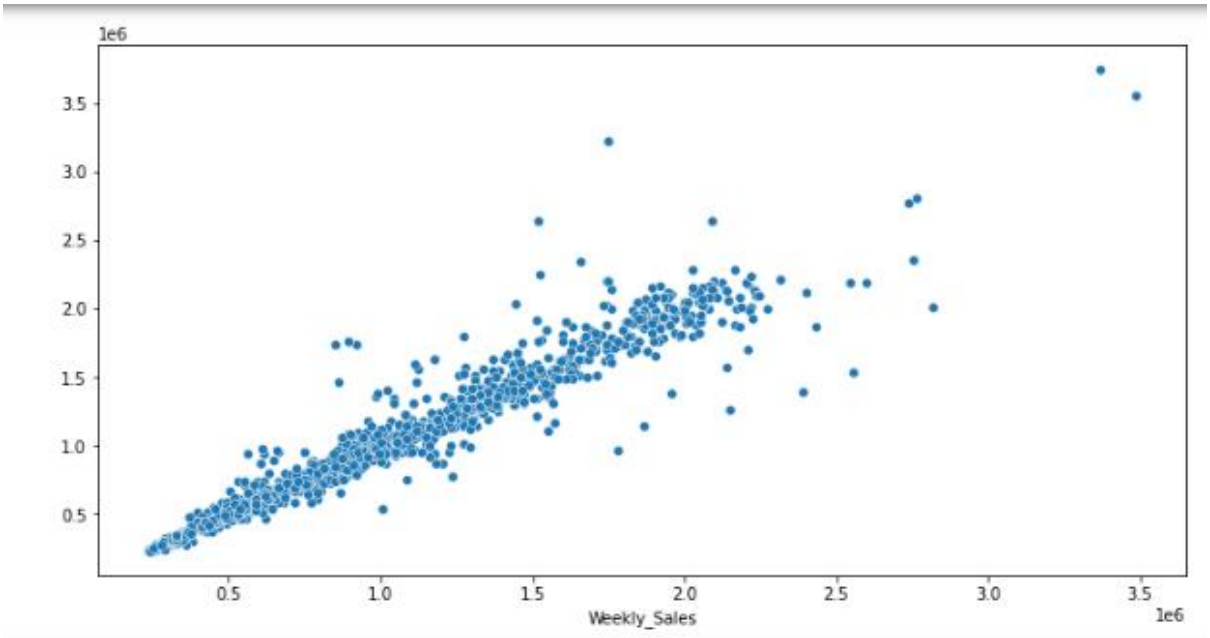
As we can see in the above figure, predicted and observed values have low correlation, and points in the plot are spread out. Let's train a Decision Tree Regressor to check whether we get any improvement in the r-squared score or not.

Decision Tree:

In [27]:
```
from sklearn.tree import DecisionTreeRegressor
decision_tree_regressor = DecisionTreeRegressor(random_state = 1234)
decision_tree_regressor.fit(x_train, y_train)
y_pred = decision_tree_regressor.predict(x_test)

print(f'MAE is - {metrics.mean_absolute_error(y_test, y_pred)}')
print(f'RMSE is - {np.sqrt(metrics.mean_squared_error(y_test, y_pred))}')
print(f'Accuracy Score is - {decision_tree_regressor.score(x_test, y_test)}'
```

```
MAE is - 76189.61445998446
RMSE is - 144077.89930172815
Accuracy Score is - 0.9307801549215738
```



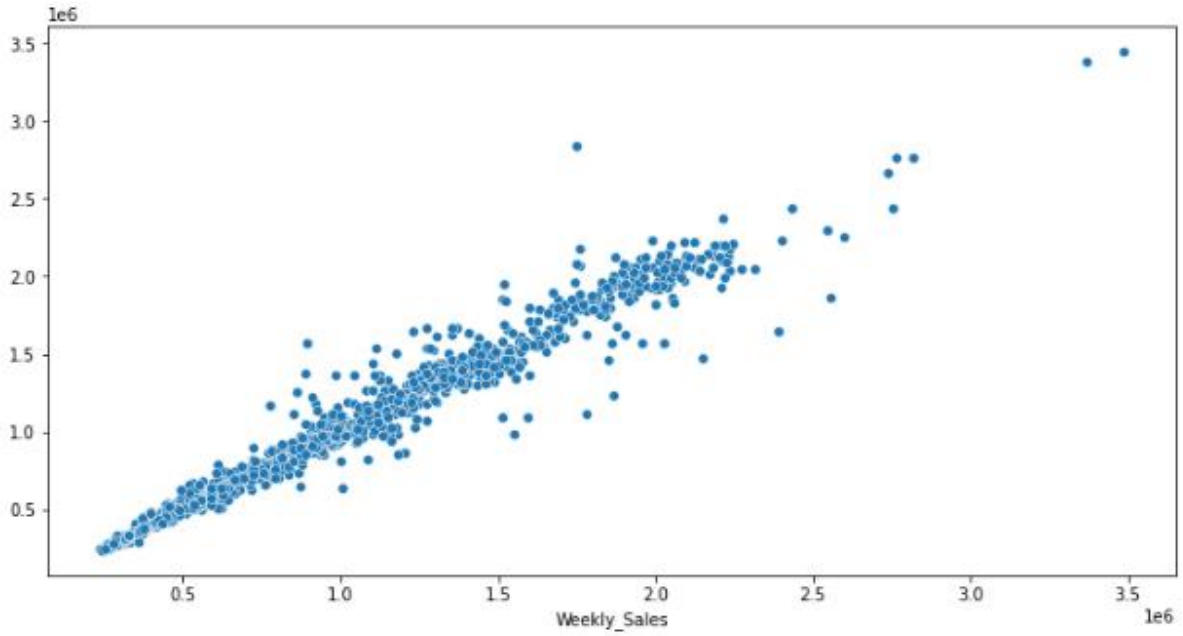
As we can see in the above figure, with Decision Tree Regressor, we get a huge improvement in both RMSE and r-squared score. We get an r-squared score of 0.93. Let's train a Random Forest Regressor and check whether we get any further improvement.

Random Forest

In [29]:
```
rf_regressor = RandomForestRegressor(n_estimators = 400, max_depth = 15, ran
rf_regressor.fit(x_train, y_train)
y_pred = rf_regressor.predict(x_test)

print(f'MAE is - {metrics.mean_absolute_error(y_test, y_pred)}')
print(f'RMSE is - {np.sqrt(metrics.mean_squared_error(y_test, y_pred))}')
print(f'Accuracy Score is - {rf_regressor.score(x_test, y_test)}')
```

```
MAE is - 60127.249368319586
RMSE is - 106926.15242423194
Accuracy Score is - 0.9618755342161982
```



We were able to gain slight improvement in RMSE and r-squared score. With Random Forest, we got an r-squared score of 0.96. Let’s build an XGBoost Regressor in the next step.

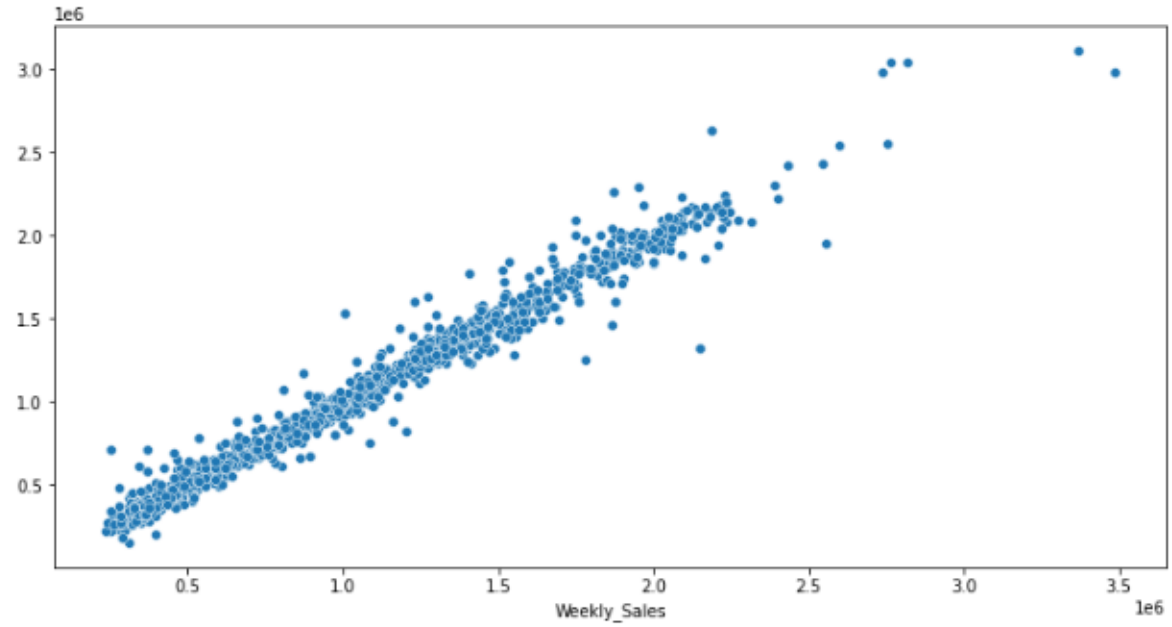
XGBoost

In [24]:
```
xgb_regressor = XGBRegressor(objective = 'reg:linear', n_estimators = 500, m
xgb_regressor.fit(x_train, y_train)
y_pred = xgb_regressor.predict(x_test)

print(f'MAE is - {metrics.mean_absolute_error(y_test, y_pred)}')
print(f'RMSE is - {np.sqrt(metrics.mean_squared_error(y_test, y_pred))}')
print(f'Accuracy Score is - {xgb_regressor.score(x_test, y_test)}')
```

```
[12:03:35] WARNING: C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling
-group-i-07593ffd91cd9da33-1\xgboost\xgboost-ci-windows\src\objective\regressio
n_obj.cu:213: reg:linear is now deprecated in favor of reg:squarederror.
MAE is - 54673.96059052059
RMSE is - 87678.0252890862
Accuracy Score is - 0.9743659393565842
```

```
<AxesSubplot:xlabel='Weekly_Sales'>
```



With XGBoost, we received the highest r-squared score and lowest RMSE. We got an r-squared score of 0.97. Let's plot the scatter plot between observed and predicted values of weekly sales in the test data. As you can see, points in the plot are very tightly distributed and observed and predicted

values of weekly sales have a strong correlation.We examined the Walmart store's sales forecasting dataset by applying various statistical and visualization techniques.

We trained and developed four ML models. We also concluded that for this problem, XGBoost Regressor works best with the accuracy of 0.974

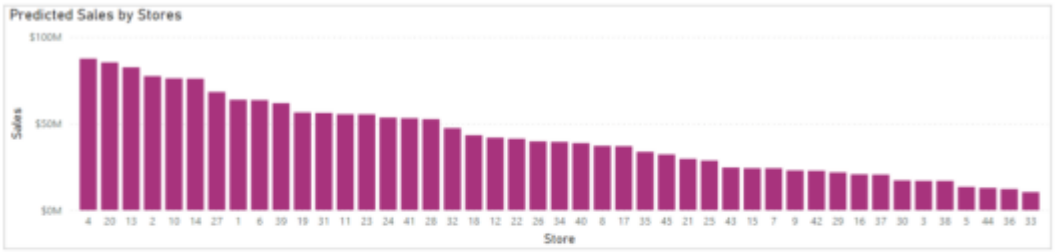
Building the Power BI Dashhoard

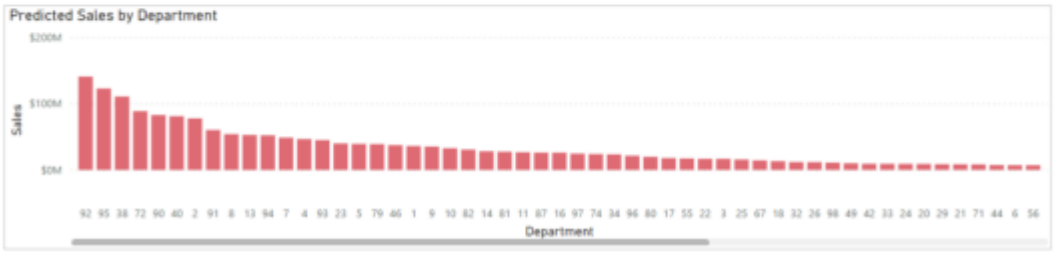
As an end product, this Power BI dashboard is going to serve as the final product of this research. The dashboard contains detailed information about the original data related to the 45Walmart stores as well as displays their respective predicted weekly sales. Most of the explorations that have been performed as part of the EDA will be included in this dashboard in the form of a story and users can filter data based on their requirements in the dashboard.

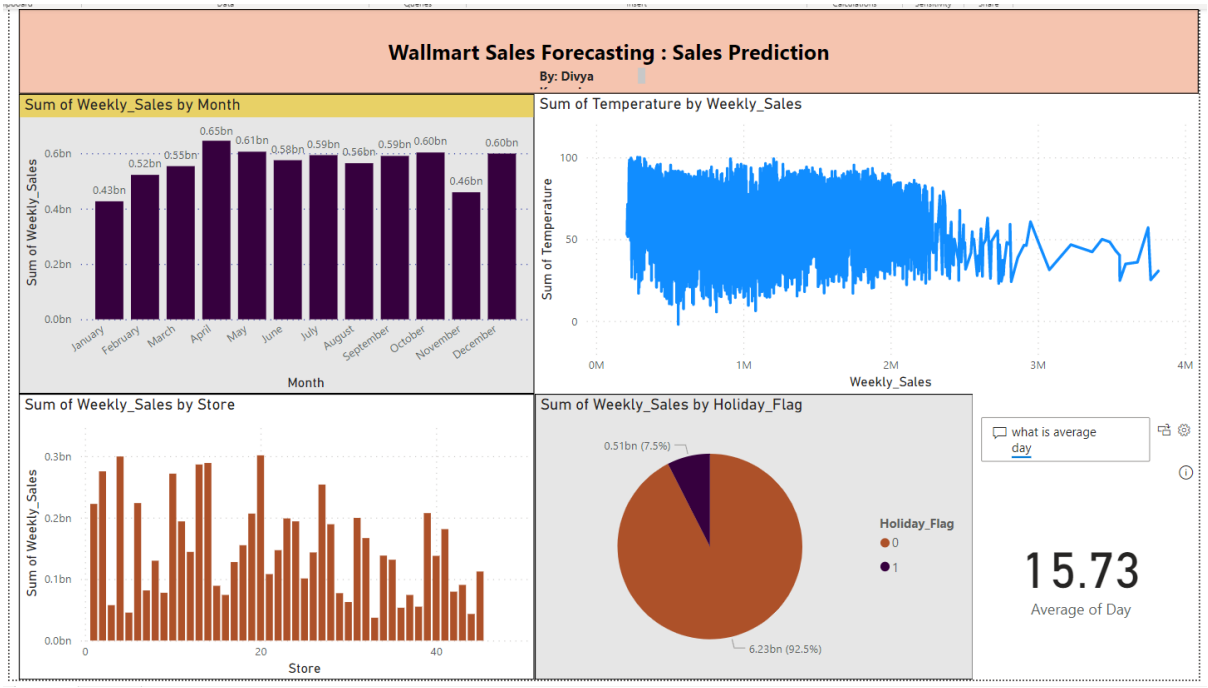
After the final predicted weekly sales are exported in the 'sampleSubmissionFinal' file, the id column is split to separate the store, department, and date information into different columns through Power BI data transformations (as shown in the figures below)

Id	Weekly_Sales
1_1_2012-11-02	30676.883
1_1_2012-11-09	17655.227

Figure 40. Power BI Dashboard View


Predicted Sales by Stores
$100M
$50M
$0M
Sales
Store


Predicted Sales by Department
$200M
$100M
$0M
Sales
Department


Wallmart Sales Forecasting : Sales Prediction
By: Divya
Sum of Weekly_Sales by Month
0.6bn
0.4bn
0.2bn
0.0bn
Sum of Weekly_Sales
0.43bn
0.52bn
0.55bn
0.65bn
0.61bn
0.58bn
0.59bn
0.56bn
0.59bn
0.60bn
0.46bn
0.60bn
January
February
March
April
May
June
July
August
September
October
November
December
Month
Sum of Temperature by Weekly_Sales
100
50
0
Sum of Temperature
0M
1M
2M
3M
4M
Weekly_Sales
Sum of Weekly_Sales by Store
0.3bn
0.2bn
0.1bn
0.0bn
Sum of Weekly_Sales
0
20
40
Store
Sum of Weekly_Sales by Holiday_Flag
0.51bn (7.5%)
6.23bn (92.5%)
Holiday_Flag
0
1
what is average day
15.73
Average of Day

5 Conclusion

Overall Results

The main purpose of this study was to predictWalmart'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 under study. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving 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 highWalmart sales. It was also observed that out of all the store types, Type A stores gathered the most sales forWalmart. Additionally, departments 92, 95, 38, and 72 accumulate the most sales for Walmart stores across all three store types; for all of the 45 stores, the presence of these departments in a store ensures higher sales. 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 favorable climate conditions as well as when the prices of fuel are adequate. However, it is difficult to make a strong claim about this assumption considering the limited scope of the training dataset provided as part of this study. 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. Again, it is hard to make a substantial claim about these findings without the presence of a larger training dataset with additional information available.

Limitations

A huge constraint of this study is the lack of sales history data available for analysis. The data for the analysis only comes from a limited number of Walmart stores between the years 2010 and 2013. Because of this limited past history data, models cannot be trained as efficiently to give accurate results and predictions. Because of this lack of availability, it is harder to train and tune models as an over-constrained model might reduce the accuracy of the model. An appropriate amount of training data is required to efficiently train the model and draw useful insights. Additionally, the models created have been developed based on certain preset assumptions and business conditions; it is harder to predict the effects of certain economic, political, or social policies on the sales recorded by the organization. Also, it is tough to predict how the consumer buying behavior changes over the years or how the policies laid down by the management might affect the company's revenue; these factors can have a direct impact on Walmart sales and it is necessary to constantly study the market trends and compare them with existing performance to create better policies and techniques for increased profits.