Logo

Description automatically generatedGroup Project Report

Data Mining Business Analytics

Prof. Dr. Gruetzemacher

Barton School of Business

**WALMART SALES ANALYSIS AND PREDICTION**

*Based on data set from Kaggle*

Group 5

Mandadi Vaishnavi

Harsha Avvari

Sashikanth Unnava

Aakrithi Adhikari

TABLE OF CONTENT

|  |  |
| --- | --- |
| List of figures | 3 |
| Introduction | 4 |
| Method | 5 |
| Data pre-processing | 6 |
| Methodology | 6 |
| Evaluation | 7 |
| Result | 11 |
| Conclusion | 15 |
| Reference | 16 |

List of figures

Figure 1: Data set from Kaggle……………….……………………………………………….………….5

Figure 2: Descriptive Analysis……………………………………………...…………………..………...6

Figure 3: Graphs between year, store, holiday flag, Temperature, CPI, Fuel price, Unemployment with weekly sales………………………………………………………………………...………………..…………….7

Figure 4: weekly sales…………………...………………………………………………………………...8

Figure 5: Month-wise sales…………………………………………………..……..………………..…….9

Figure 6: Yearly sales………………………………………………...…………………………….……..10

Figure 7: Train and Test data………………………………………………..................………….……....10

Figure 8: Code for Linear Regression…………………………………………...……...…………….……11

Figure 9: Plot for Linear Regression……………………………………………………....………………..12

Figure 10: Code for Random Forests…………………………………………...…………………………..12

Figure 11: Plot for Random Forests………………………………...…………………………………..…..13

Figure 12: Code for Decision Trees…………………………………………………………....…………...14

Figure 13: Plot for Decision Trees …………………………………………...…........................………......15

**INTRODUCTION:**

Walmart is the world’s largest company by revenue with about US$570 billion in annual revenue and largest private employer in the world with 2.2 million employees. More than 10,000 stores are in 24 different countries. Walmart views this data as their most asset since it enables them to forecast future sales and customer behaviour, as well as develop strategies for making money and competing with other businesses. The variety of goods provided by Walmart, which promises "everyday low pricing" to its customers, brings in close to $500 billion in revenue annually, making it imperative for the business to employ sophisticated ways to predict future sales and ensuing profits. Walmart, the largest company in the world by revenue, sells a wide range of goods, including groceries, home furnishings, body care items, electronics, clothing, etc. It also generates a lot of consumer data, which it uses to forecast consumer purchasing trends, future sales, and promotional plans as well as to develop new and cutting-edge in-store technologies.

In recent years, the company has faced challenges from online competitors and changes in consumer behaviour. As a result, it is important for Walmart to closely monitor its sales performance and to develop accurate sales predictions to stay competitive in the marketplace. In this analysis, we will examine Walmart's sales data and use it to develop a model for predicting future sales. This will help the company to make informed decisions about its operations and strategy. This dataset was taken from Kaggle. This dataset consists of the data of 45 stores across the states from 2010 to 2012.

Within the Dataset file, following fields are present:

1. Store - the store number

2. Date - the week of sales

3. Weekly\_Sales - sales for the given store

4. Holiday\_Flag - whether the week is a holiday week 1 – Holiday week 0 – non-holiday week

5. Temperature - Temperature on the day of sale

6. Fuel\_Price - Cost of fuel in the region

7. CPI – Prevailing consumer price index

8. Unemployment - Prevailing unemployment rate

Holiday Events in the Week

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12

Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12

Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12

Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12

**METHOD**

The following section briefly describes the data set used to develop the model, the process of pre-processing the data, and the methodology used to build and identify the most accurate model.

Table

Description automatically generated

Figure1: Data set from Kaggle

There are total 6435 rows and 8 columns. Our Dependent variable is weekly Sales and Independent variables are rest of them like Store, Date, Holidays, Temperature, Fuel Price, CPI, and Unemployment.

We have started by taking 10 default rows to understand the data and analysed the information of each row and column to make necessary changes. Then eventually data is extracted and description of data from data frame is collected like count, mean, std, minimum value, 25% percentile\*, 50% percentile\*, 75% percentile\*, maximum value for all the independent variables which analyses both numeric and object series, the Data Frame column sets of mixed data types, statistical summary of the Series and Data Frame and number of missing values in the data set.

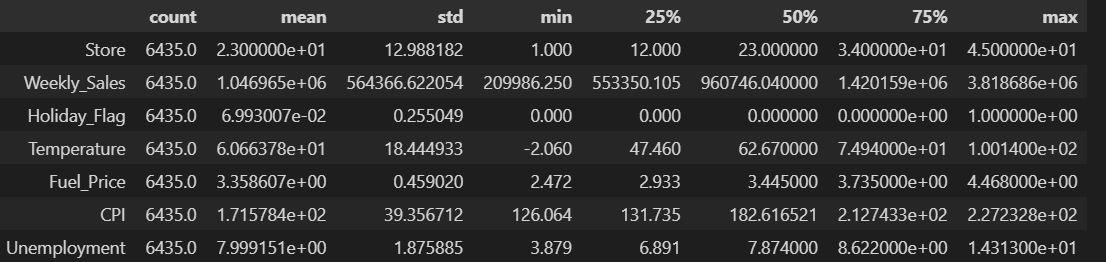


Figure 2: Descriptive Analysis

For the data set, to better understand and come up with a reliable analysis. We came up with different regression analyses to get more insights into the data set. For this data set, the regression analysis we have chosen are Linear regression, Random Forest, and Decision trees. We considered comparing all three R squares and mean square error and get the results that better suits for the data set.

**DATA PRE-PROCESSING:**

We have run the data in python to check whether the data have any errors or missing values, but we found nothing. To choose a better ML model, Dataset is cleaned and made few changes like for date (Month, year, day, and month year value) for accurate results. Every independent variable is plotted with weekly sales, monthly sales, yearly sales and observed variations accordingly.

Along with them subplots are made which gave us understanding of sales are better with visuals using bar graphs, histograms, and boxplots by grouping them. By doing exploratory analysis we observed sales are high at the end of each year and 2011 has more sales than 2010 and 2012 because of various factors affecting them.

**METHODOLOGY:**

We have segmented our project into 3 important parts such as Firstly, we compared some of our independent variables with the dependent variable individually with the visual graphs which were mentioned as figures so that we could have a good glance at how much of those independent variables affects the weekly sales (Dependent variable).

In the second part we would be observing the weekly sales, month wise sales and yearly sales.

In the third part we train and test the data to find the best model which would be suitable for the future predictions.

**EVALUATION:**

PART 1

Analysis few of our most popular independent variables with the dependent variable by representing the graphs below.

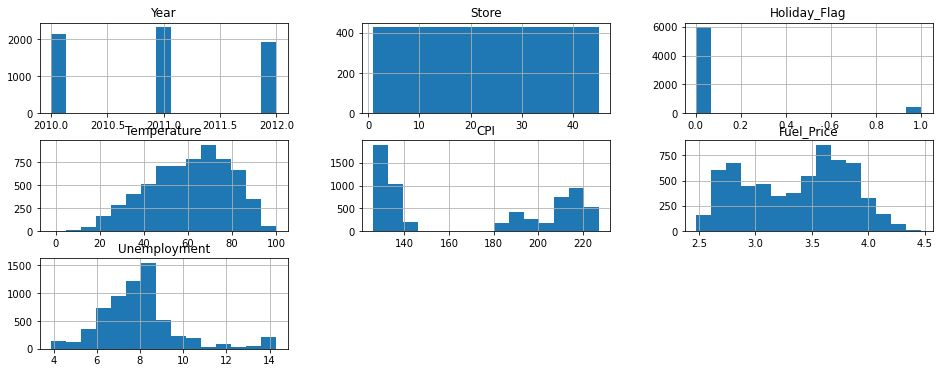


Figure 3: Graphs between year, store, holiday flag, Temperature, CPI, Fuel price, Unemployment with weekly sales.

From Figure 3 we observed that

* 2011 has more sales compared to 2010 and 2012
* Weekly sales are almost equal all over the 45 stores
* Sales are at peak with holiday flag 0 (non-holiday weeks)
* Temperature had a great effect on sales, like when temperature is around 70 degrees then weekly sales are more comparatively.
* Through Customer Price index, we have observed that with average change in price weekly sales have decreased eventually with fluctuations in between.
* With fuel price there is variation in weekly sales
* More unemployment led to less weekly sales

PART 2:

In part 2 From figure 4, we have observed every 6-month sales from January 2010- December 2012 with month year value. Here we can observe that sales are at its peak in the month December because of holiday season.

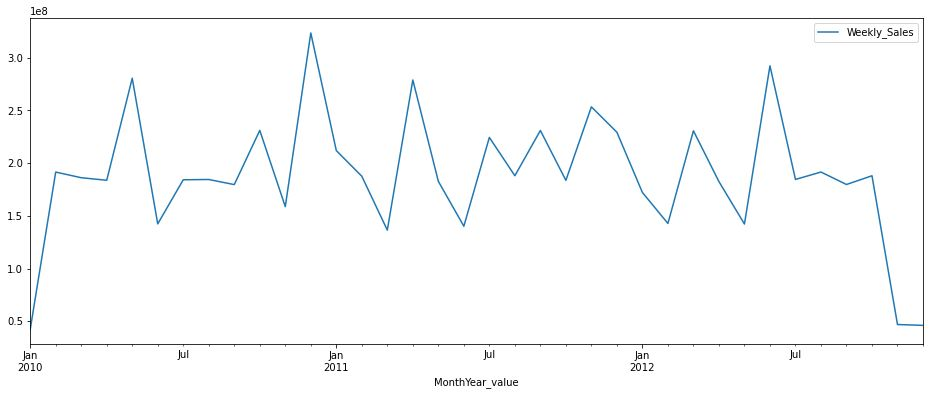


Figure 4: weekly sales

From figure 5, we observed month wise sales with weekly sales over a period of 12 months they are more in the month of December.

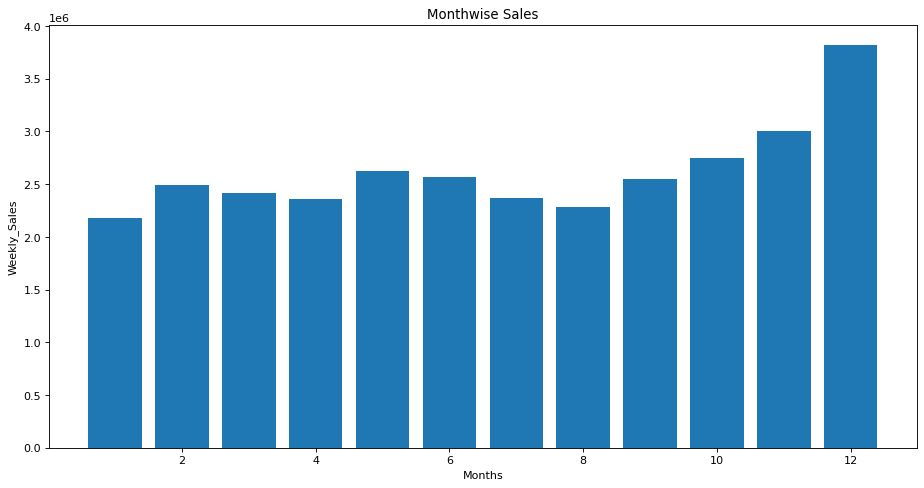


Figure 5: Month-wise sales

From figure 6. On comparing all years, we found that 2011 has great sales compared to 2010 and 2012.Chart, bar chart

Description automatically generated

Figure 6: Yearly sales

PART 3:

Data Splitting:

The data frame is divided for training and fitting the model using the training and test sets. The training data is contained in x\_train and y\_train, while the data for testing is in x\_test and y\_test. Here, we are going to take test\_size=0.2 which means that, approximately 20 percent of samples will be assigned to the test data, and the remaining 80 percent will be assigned to the training data where our train sets are larger than the test sets.

Text

Description automatically generated

Figure 7: Train and test of the data

**RESULTS:**

Here, we will be using 3 different algorithms to understand which model to use to predict the weekly sales that fits our data.

Since, our data set is not a big and predictor variable and a dependent variable related to each other in a linear fashion we used linear regression. Variables are related linearly for forecasting the effect of other factors on sales.

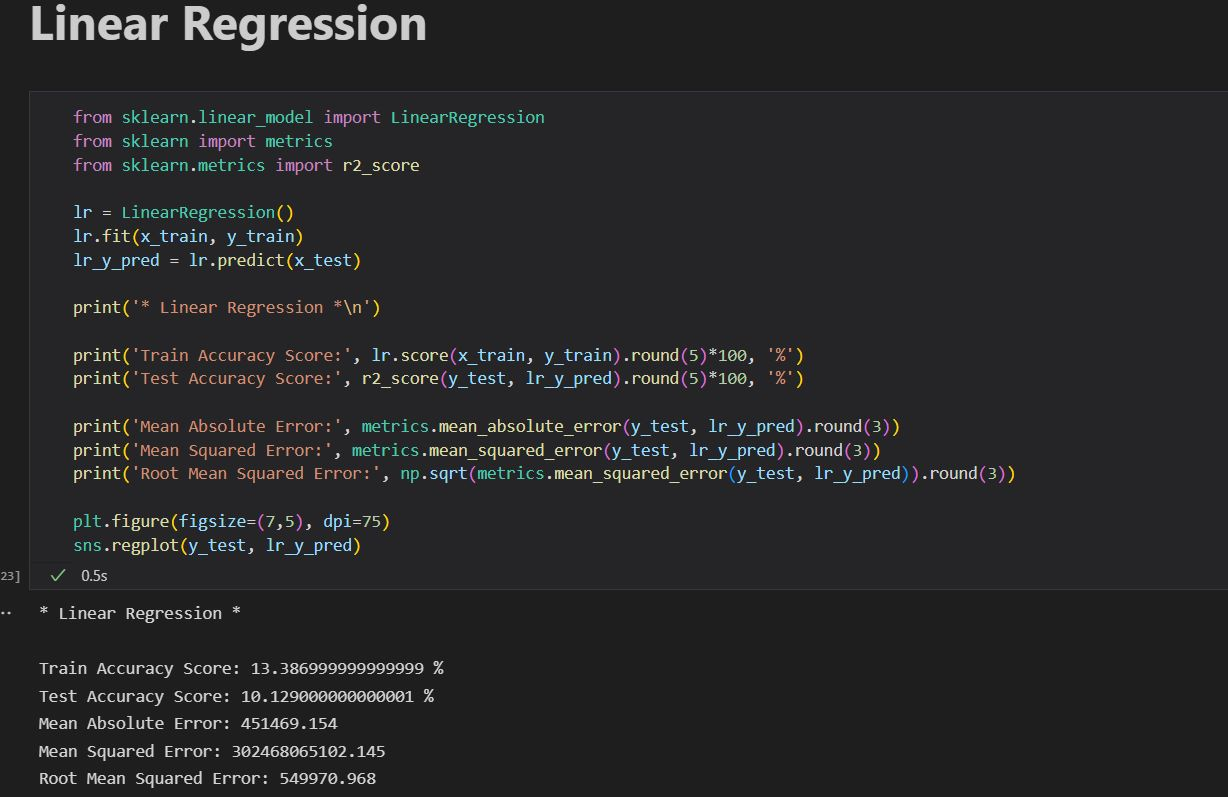


Figure 8: code for Linear Regression

Results of Linear Regression:

Mean Absolute Error: 451469.154

Mean Squared Error: 302468065102.145

Root Mean Squared Error: 549970.968

Mean Score: 12.917730538523609 %

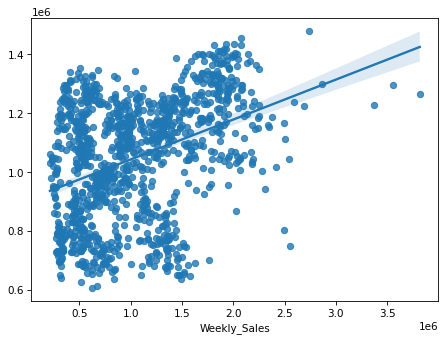


Figure 9: Plot for Linear Regressor model

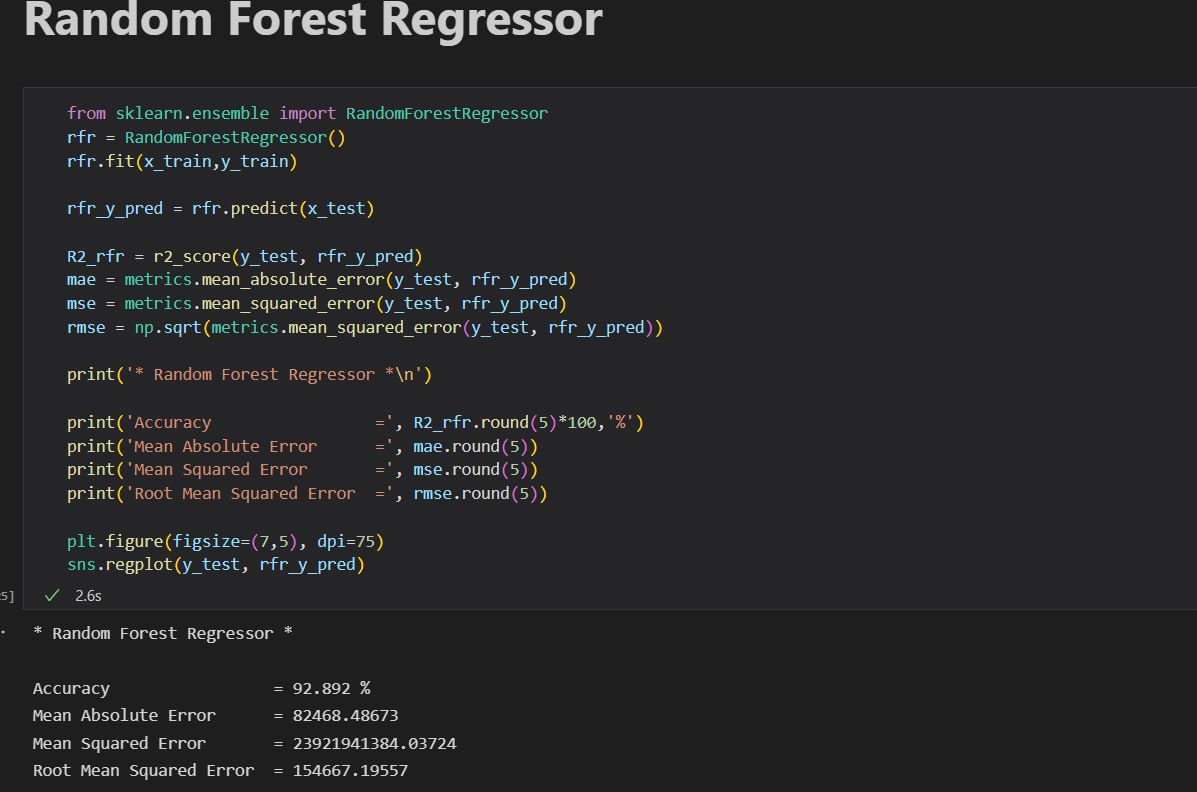


Figure 10: Code for Random Forest Regressor

The Random Forest Regressor model performs better than any other model to date, with a 94.77% accuracy rate of r square and a mean squared error of just 23.92

Finding: We see that the Random Forest Model that we used outperformed the rest of the models with the highest accuracy of 95% and comparatively the root mean square error is low.

we could say that the random forest model is the better fit for the data that we have chosen.

Results of Random Forest Regressor:

Mean Absolute Error = 82468.48673

Mean Squared Error = 23921941384.03724

Root Mean Squared Error = 154667.19557

Mean Score: 94.77539357782486 %

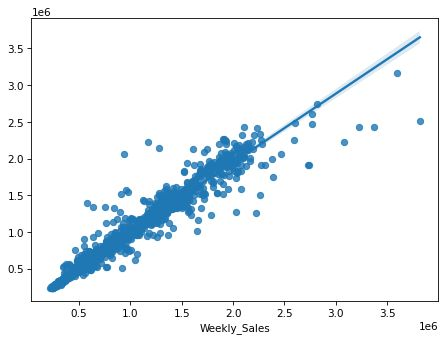


Figure 11: Plot for Random Forest Regressor

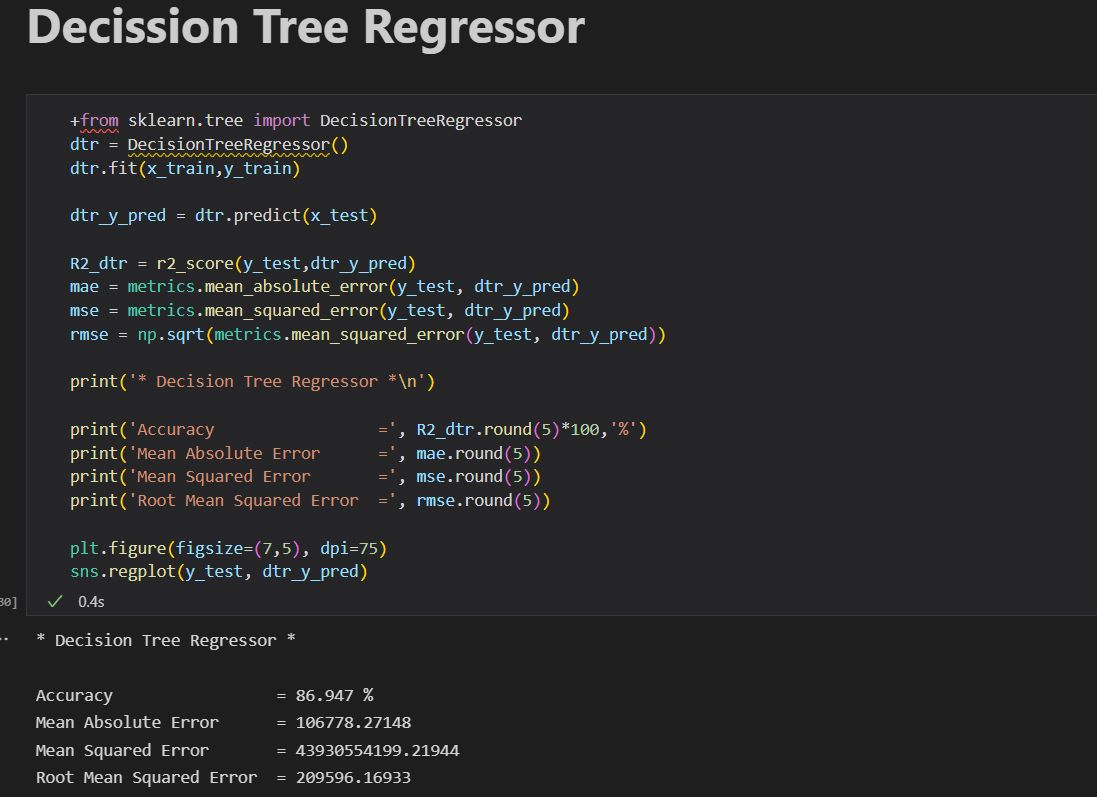


Figure 12: Code for Decision Tree Regressor

Decision tree model performs well with 92% accuracy of r square and with the mean square error of 43930554199.21944

Results of Decision Tree Regression:

Mean Absolute Error = 106778.27148

Mean Squared Error = 43930554199.21944

Root Mean Squared Error = 209596.16933

Mean Score: 92.56085281008289 %

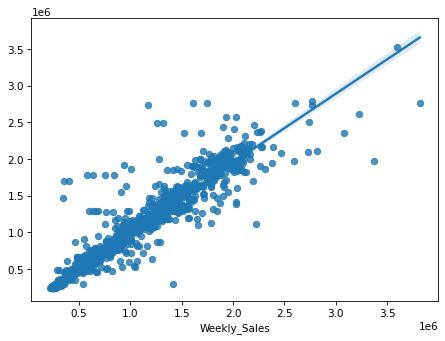


Figure 13: Plot for Decision Tree Regressor

**CONCLUSION:**

This project sets out to determine that running a different regression model with 6435 rows and 8 columns. By taking dependent variable as weekly sales and the rest of all the variables as independent variables. We have segmented our project into 3 major parts as we investigated the popular independent variables which effects the weekly sales. We figured that the 2011 has more sales compared to 2010 & 2012 and are almost equal all over the 45 stores, with holiday flag 0 (non-holiday weeks) when temperature is around 70 degrees. Also CPI, fuel price there is variation in weekly sales (Figure 3)

The results of the initial data analysis of weekly sales were presented in this project. Also included here is a discussion of how the machine learning method may be used to assess the existing data on forecast weekly sales. To better foretell future data, we include the Linear Regression, Decision Tree Model and Random Forest Model to find which model suits the best for the data, along with other forms of data analytics, such as exploratory data analysis. The models (fig 8, 10, 12) associated with machine learning was also described in depth in this study. This study also includes information on the commercial implications of data mining for weekly sales of Walmart. This paper details how to use Python to build linear regression, decision tree and random forest code.

**REFERENCE:**

Textbook: Analytics, Data Science, & Artificial Intelligence 11th ed., Sharda, Delen and Turban.

https://www.kaggle.com/datasets/varsharam/walmart-sales-dataset-of-45stores