**Dataset:**

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

**Dependent variables, and Independent Variables:**

The file has information about the Weekly Sales of 45 stores for the year 2010-2012 including the factors affecting Sales. So, here the **Dependent variable** is **weekly Sales** and **Independent variables** are rest of them like **Store, Date, Holidays, Temperature, Fuel Price, CPI, and Unemployment.**

**Understanding data:**

**Data analytics:**

We have started by taking 10 default rows to understand the data and analyzed 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 using describe () method which analyzes 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.

**Exploratory Data Analytics:**

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.

**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.

**Approach/Method:**

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.

**Results of 3 models:**

**Results of Linear Regression:**

Train Accuracy Score: 12.316 %

Test Accuracy Score: 14.427000000000001 %

Mean Absolute Error: 431897.142

Mean Squared Error: 275418661409.513

Root Mean Squared Error: 524803.45

For accuracy we performed cross validation where, Mean Score: 11.916301922610591 %

For better results and to improve accuracy Random Forest Regression is used.

**Results of Random Forest Regressor:**

Accuracy = 93.375 %

Mean Absolute Error = 80335.22335

Mean Squared Error = 21323774910.44586

Root Mean Squared Error = 146026.62398

After applying cross validation, we got **Mean Score: 94.4012682988436 %**

**Results of Decision Tree Regressor:**

Accuracy = 88.67 %

Mean Absolute Error = 102000.52052

Mean Squared Error = 36465398895.8725

Root Mean Squared Error = 190959.15505

After applying cross validation, we got Mean Score: 91.07795261708544 %

**Conclusion:**

On comparing all results, Linear Regression is not an appropriate model to use as accuracy is very low. However, **Random Forest Regression gives accuracy of almost 95%.** So, it is the best model to forecast weekly sales.