***Sales Forecasting Performance Insights***

**INTRODUCTION:**

Forecasting means to predict the future. Forecasting is used to predict future conditions and making plans accordingly. In our daily life, we are using a weather forecast and plan our day activity accordingly. Forecasting is used in many businesses. Here we will learn Sales Forecasting using Dataset using Machine Learning in Python.

**PROBLEM STATEMENT:**

The main problem that I am assigned with is that I have to predict the sales given the data-set. As I can understand from the problem itself is that it is a regression problem. That we have to use regression models in-order to predict the sales from the data-set.

**STEPS INVOLVED IN SALES FORECASTING:**

we use the dataset of Walmart sales to forecast future sales using machine learning in Python. Linear regression use to forecast sales. Numpy, Pandas, Sklearn, Scipy, Seaborn Python libraries used in this program. We implement in three steps first to import libraries second by using that libraries prepare data and third forecast.

* Get the data from .csv file.
* Merging the data.
* Analyse the data.
* Manipulating the data.

**GET DATA FROM .CSV FILE:**

**Stores.csv:** This file contains anonymized information about the 45 stores, indicating the type and size of store.

**train.csv:** This is the historical training data, which covers to 2010–02–05 to 2012–11–01.

**test.csv:** This file is identical to train.csv, except we have withheld the weekly sales. You must predict the sales for each triplet of store, department, and date in this file.

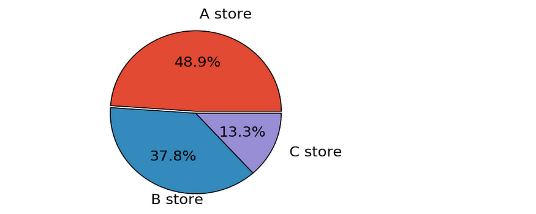
**features.csv:** This file contains additional data related to the store, department, and regional activity for the given dates.

For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13  
Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13  
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13  
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

**EXPLORATORY DATA ANALYSIS:**

After going through the data, I decided to analyse the data at first. The main perspective of the analysis is that Sales can be responsive to time factor and space factor Store’s sales records are the aggregation of each department Date variable can be split into y/m/w/d variables. Day variable can provide much information on sales. Outside data such as national holiday of US will be combined to add information. As we already know that in machine learning problem solving part, we have to try all the possible ways to solve the problem whereas 90 percent of our experiments can be in vain. But we have to still try hard for the approximate solution.

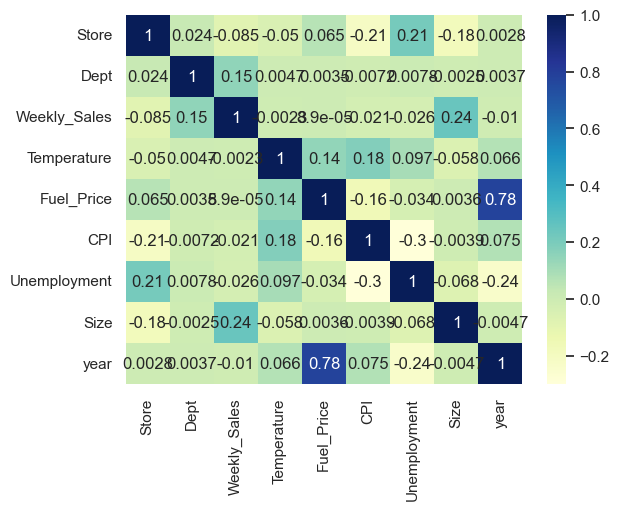


What we can see in this pie-chart is that 48.9% of the sales belongs to A store about 37.8% belongs to the B store and 13.3 % of the sales belong to C store.

In order to understand the size of the type of the sales in each of the types A, Band C, I have used box-plot.

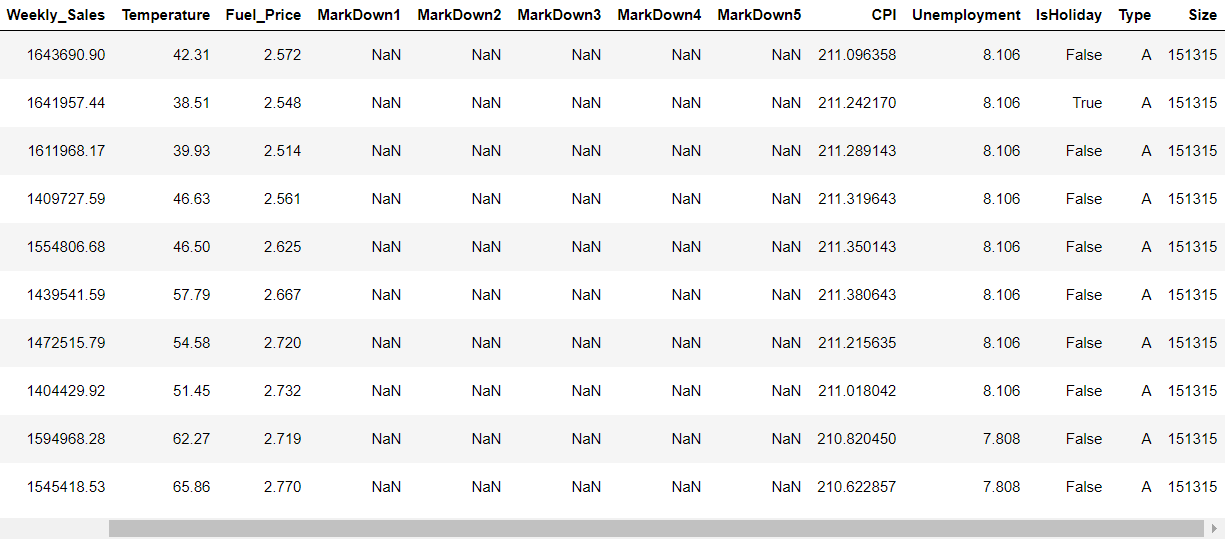
**DATA CORRELATION:**

Method Data correlation is a method that helps to predict one attribute from another attribute and is used as a basic quantity in many modelling techniques. If one feature increases, the correlation will be positive, so the other feature increases as well and negative if one feature increases there will be a reduction in another. If there is no relation between any two attributes then it is said to be no correlation. If there is a linear relationship between the constant variables then the Pearson correlation coefficient is used. If there is a non-linear relation between the constant variables then the Spearman correlation coefficient is used.



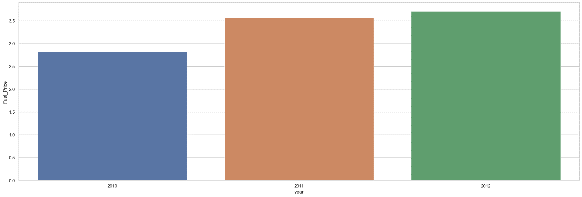
**MERGING DATA:**

Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary when you have raw data stored in multiple files, worksheets, or data tables



**ANALYSING THE DATA:**

Data analysis is a technique that typically involves multiple activities such as gathering, cleaning, and organizing the data. These processes, which usually include data analysis software, are necessary to prepare the data for business purposes. Data analysis is also known as data analytics described as the science of analysing raw data to draw informed conclusions based on the data.



**FILL MISSING VALUES:**

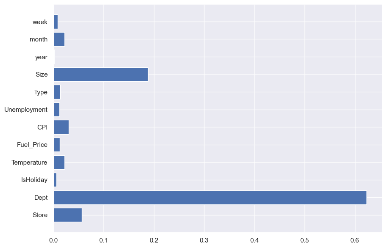
Due to various reasons, missing values commonly appear in raw data. To check how many missing values in each column

There are many approaches to dealing with missing values, including:

* Drop columns with missing values
* Imputation, filling in missing values with some number, such as the mean value along each column.
* An extension to Imputation. Imputation is the standard approach. However, imputed values may be above or below their actual values. To help model make a better prediction, we can add a column to tag the missing value as True or False.

**RANDOM FOREST REGRESSOR:**

The box plot in Figure 4.10 shows the accuracy score (ACC) obtained by Random Forest Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum accuracy score of 91.35 percent, the middle quartile



**XGBOOST:**

This algorithm only works with the quantitative variable. It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed.

**CONCLUSION:**

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both the revenue and the income. Different types of Machine Learning techniques such as Support Vector Regression, Gradient Boosting Regression, Simple Linear Regression, and Random Forest Regression have been evaluated on food sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy, mean absolute error, and max error, the Random Forest Regression is found to be the appropriate algorithm according to the collected data and thus fulfilling the aim of this thesis.