**Project Title:** Big Data Sales Prediction.

**Background:**

Let’s understand Why, What and How in Linear Regression model. Firstly, Why Linear Regression? This regression model draws the relationship between the dependent and independent variables, where dependent variable is otherwise known as predictor variable and independent variable is known as response variable. Now let’s investigate What is Linear Regression? This is a statistical analysis that attempts to show the relationship between two variables. How Linear Regression works? With the data provided by the dataset, we draw a plot between independent and dependent variable and spot a mean value for those values. It selects the best fit line that passes through the mean of the data. But as we know there will be multiple lines that passes through that mean point. To solve this scenario, we move the line until we have least square distance from all the data points. To analyze this technique there few steps, that need to be followed:

* Generating inputs using .CSV file.
* Importing the libraries.
* Splitting the dataset into train and test.
* Applying regression on independent variables.
* Validating the model.

**Introduction:**

The necessity of this project is to explain the applicability of the **linear regression model** within a Big Data Sales dataset. This dataset consists of a data of 1559 products across 10 stores in different cities. This data contains of training and testing datasets in which, there are 12 attributes and 8524 instances and 11 attributes and 5782 instances respectively. The following is the description of attributes in training and testing datasets models:

|  |  |  |
| --- | --- | --- |
| **SNO** | **Variable** | **Description** |
| 1 | Item\_Identifier | Unique product ID |
| 2 | Item\_Weight | Weight of product |
| 3 | Item\_Fat\_Content | Whether the product is low fat or not |
| 4 | Item\_Visibility | The % of total display area of all products in a store allocated to the particular product |
| 5 | Item\_Type | The category to which the product belongs |
| 6 | Item\_MRP | Maximum Retail Price (list price) of the product |
| 7 | Outlet\_Identifier | Unique store ID |
| 8 | Outlet\_Establishment\_Year | The year in which store was established |
| 9 | Outlet\_Size | The size of the store in terms of ground area covered |
| 10 | Outlet\_Location\_Type | The type of city in which the store is located |
| 11 | Outlet\_Type | Whether the outlet is just a grocery store or some sort of supermarket |
| 12 | Item\_Outlet\_Sales | Sales of the product in the particular store. This is the outcome variable to be predicted. |

So, our main aim is to predict data of missing attribute in testing dataset i.e., **“Item\_Outlet\_Sales”** with the help of training dataset. As, we all know the data in the sales sector is extremely huge to handle. This regression analysis will promote us to predict the future and help in developing the market sales value to meet the customer satisfaction. This prediction techniques helps the companies to evolve the product efficiently and get back to the market with the better version of it. To predict this Item\_Outlet\_Sales we used Linear Regression Model.

**Challenges Experienced and How these were resolved:**

1. **Data Cleaning:**

Data cleaning mainly involves identifying and removing noisy and inconsistent data in order to improve the quality of the data. The main reason for noisy and inconsistent data is during the initial stage, entering incorrect and missing data at the data entry level. This cleaning task is vital, because to take the best possible decision in business implementation. To lie about statistics is very easy task but to get the true data is the most challenging one. We have resolved this noisy data by using Replace Command in R.

1. **Connection:**

To observe all the attributes and draw some connections between those attributes to identify the predict variable and response variable. In this case the predict variable is “Item\_Outlet\_Sales” and it depends on the response variables like Item\_Identifier and Outlet\_Identifier which are described as the unique product ID. That means these two independent variables are responsible to predict the values for Item\_Outlet\_Sales. This conclusion is defined by observing the data as the same item is available in different outlets with different price. So, this Item\_Outlet\_Sales depend on the Item\_Identifier and Outlet\_Identifier. To get the values we considered Item\_Identifier + Outlet\_Identifier which leads to Item\_Outlet\_Sales.

**Implementation:**

My observation on the datasets both on training and testing data sets, the training data set contains sales values whereas in testing dataset we don’t have sales values. To predict that, as we draw the connection between attributes as mention in the above topic i.e.,” connection”. To get the predicted values for Item\_Outlet\_Sales. We need to implement the following steps:

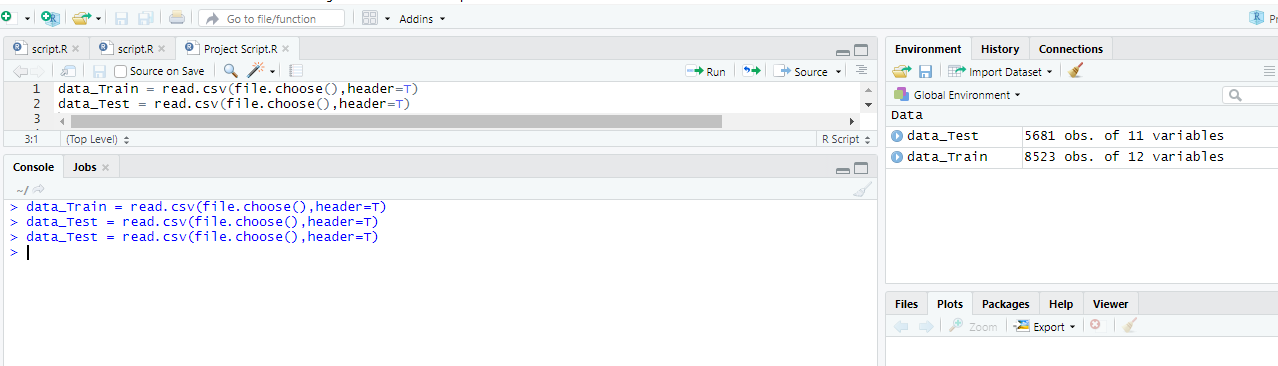
1. Loading the training dataset file into R studio by using “read.csv” R command.
2. Loading the testing dataset file into R studio by using “read.csv” R command.
3. As there are some missing values in both test and train datasets, we used “replace” R command to replace missing values.
4. Making scatter plot on how sales price depends on Item\_Type related.
5. Creating a Linear Regression model to get Item\_Outlet\_Sales based on Item\_Identifier and Outlet\_Identifier.
6. Display the summary of model we created.
7. Plot the model to display the graphical representation
8. Predict the Item\_Outlet\_Sales for test dataset using predict function. It uses sales values from training dataset, and it does predict the Item\_Outlet\_Sales for test dataset.
9. Display the values what we have predicted. It does give three values i.e., approx. value, minimum value and maximum value.
10. Write the values of Item\_Outlet\_Sales that we predicted into .CSV file.

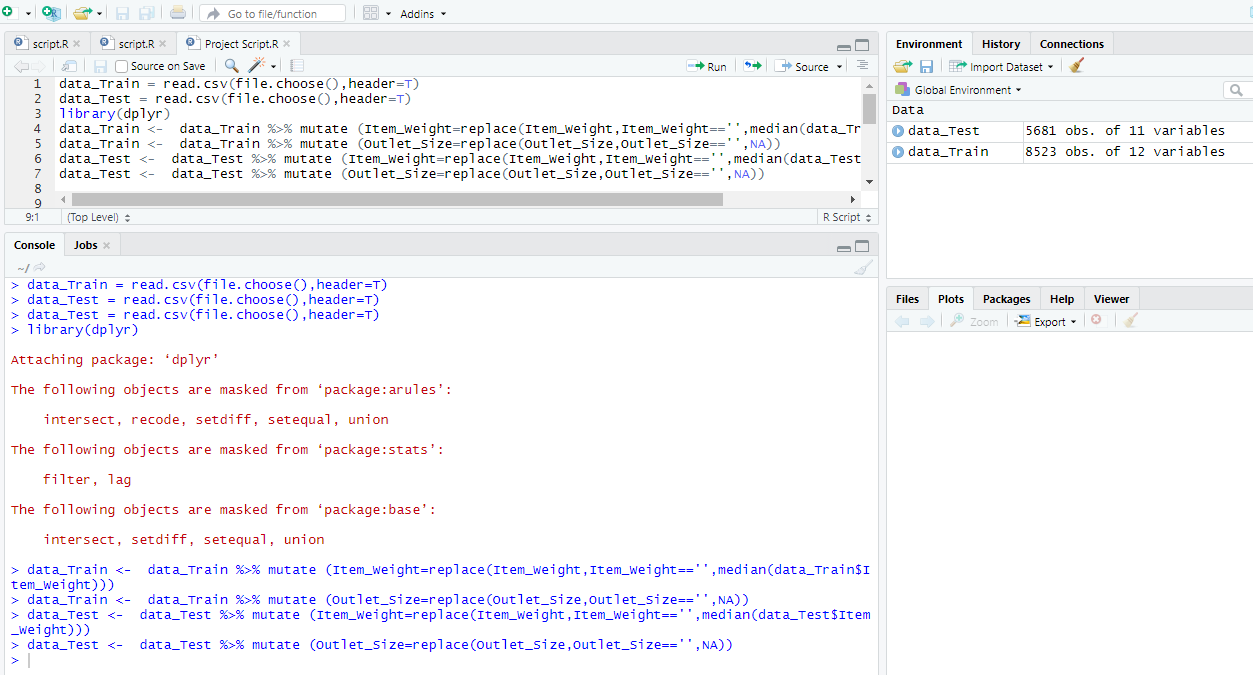
**Results:**

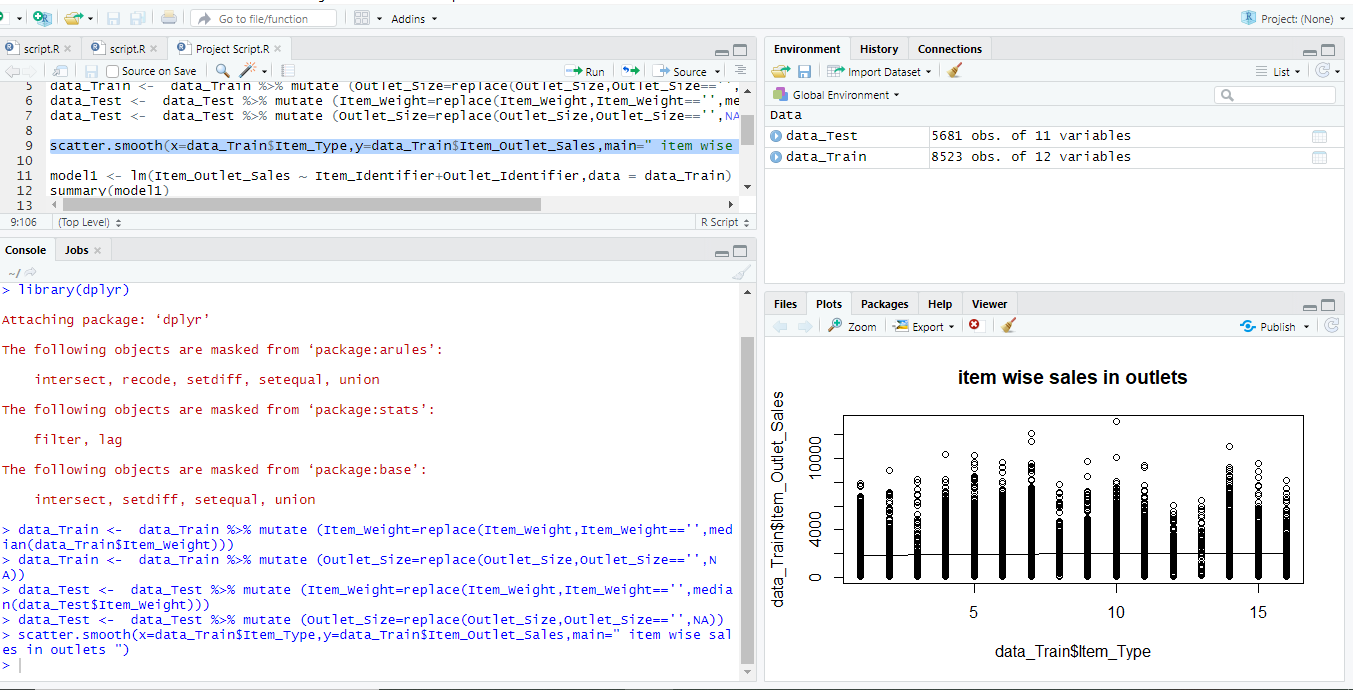
Finally, by implementing the above steps we obtain the of values of Item\_Outlet\_Sales in testing dataset.

**Step 1**:

Loaded training and Test data into R studio.

**Step 2**:

As there are some missing values in both test and train data sets, we used below commands to replace missing values.

**Step 3**: Make a scatter plot on how sales price depends and Item\_Type related

**References:**

1. <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/>
2. <https://towardsdatascience.com/car-selection-and-sales-day-prediction-8c4a474f9dca>
3. <https://sebastianraschka.com/blog/2016/model-evaluation-selection-part1.html>
4. <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>
5. <http://onlinestatbook.com/2/regression/intro.html>
6. <https://www.sisense.com/glossary/data-exploration/>
7. <https://www.youtube.com/watch?v=2Sb1Gvo5si8>
8. <https://www.youtube.com/watch?v=zPG4NjIkCjc>