Big Data Mart Sales Problem



Rapid increase in Self Selected Store / Super Market / Modern Trade across the regions has brought lot of importance in data to ensure the right product is available to right audience with considering the demographics in place.

Data helps stores to manage their space / sales effective and efficiently with maintaining the appropriate inventory and placing the fast-selling SKU’s (Stock Keeping Units) at POS (Point of Sales).

These format stores make every aspect like Promotions, Signages, Product Placements as by keeping the consumer behaviour in mind with help of data captured from the previous purchases. Every customer data is used to send customised message / alert with regarding to sales.

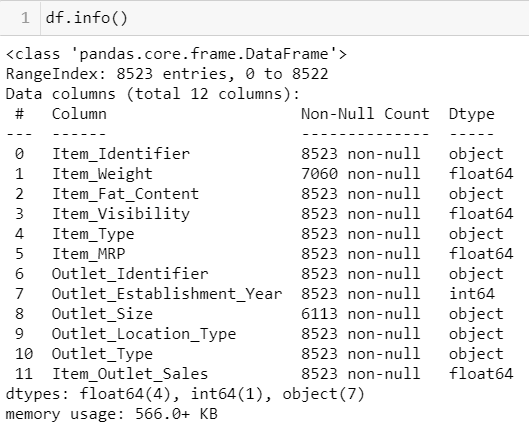
Coming to our dataset Big Mart Data, the data scientists at Big Mart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

Let us find few attributes of the dataset:

* Item\_Identifier: Unique product ID
* Item\_Weight: Weight of product
* Item\_Fat\_Content: Whether the product is low fat or not.
* Item\_Visibility: The % of total display area of all products in a store allocated to the product.
* Item\_Type: The category to which the product belongs.
* Item\_MRP: Maximum Retail Price (list price) of the product
* Outlet\_Identifier: Unique store ID
* Outlet\_Establishment\_Year: The year in which store was established.
* Outlet\_Size: The size of the store in terms of ground area covered.
* Outlet\_Location\_Type: The type of city in which the store is located.
* Outlet\_Type: Whether the outlet is just a grocery store or some sort of supermarket.
* Item\_Outlet\_Sales: Sales of the product in the particulat store. This is the outcome variable to be predicted.

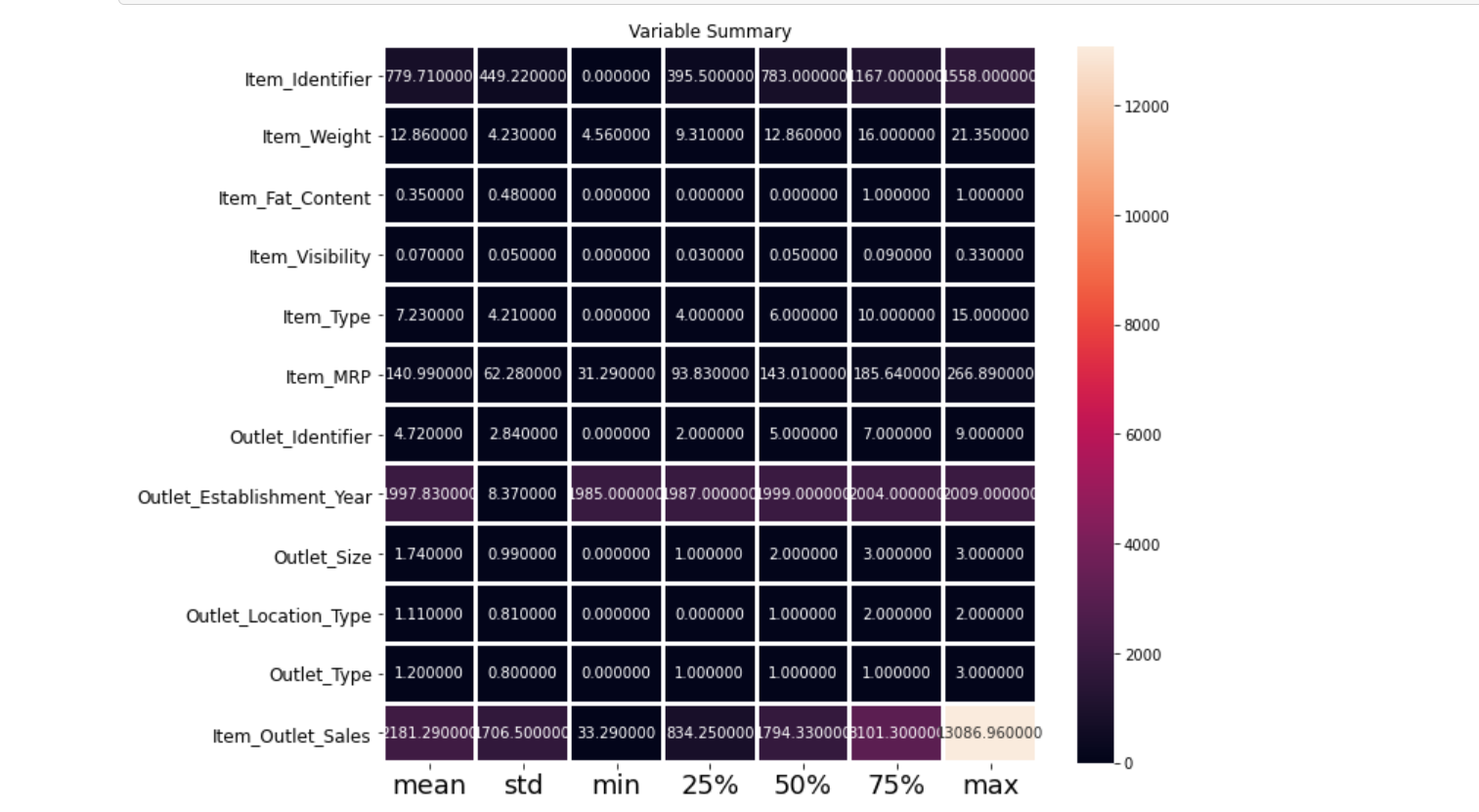
**Data captured in DataSet:**

We can find 8523 observation and 12 feature.

Target variable is Item\_Outlet\_Sales

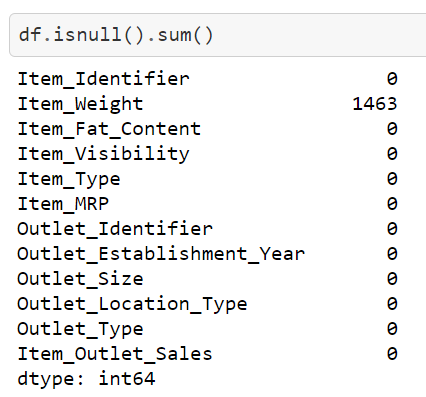
**Import Libraries:**

**Understand the Dataset:**

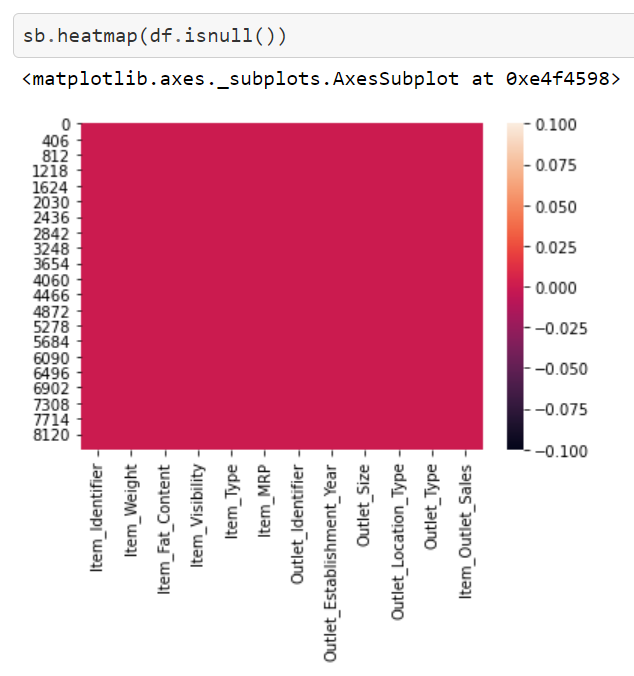
****

Mean, Max, Min of every variable is caputured. Most of the Mean and 50% values are similar.

**Checking the Null Values:**



Item\_Weight is having the 1463 as Null Values.

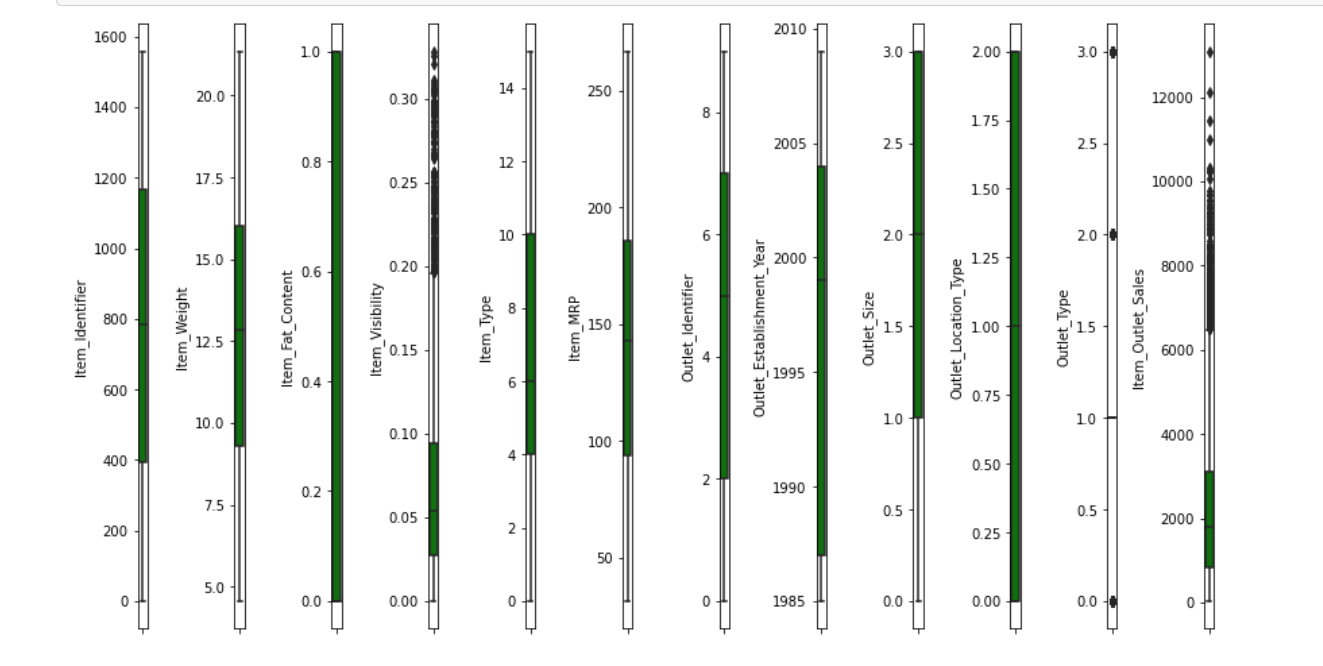
**After treating the dataset to remove Null Values**

**Observing the Correlation:**

****

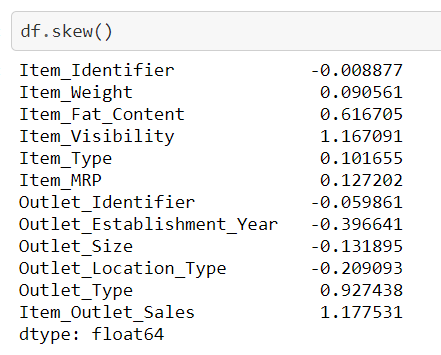
Sales having correlation with Outlet Type, Item MRP, Outlet Identifier, Outlet Location

**Checking the Outlier:**

****

Item Visibility, Outlet Type and Outlet Sales are having outliers.

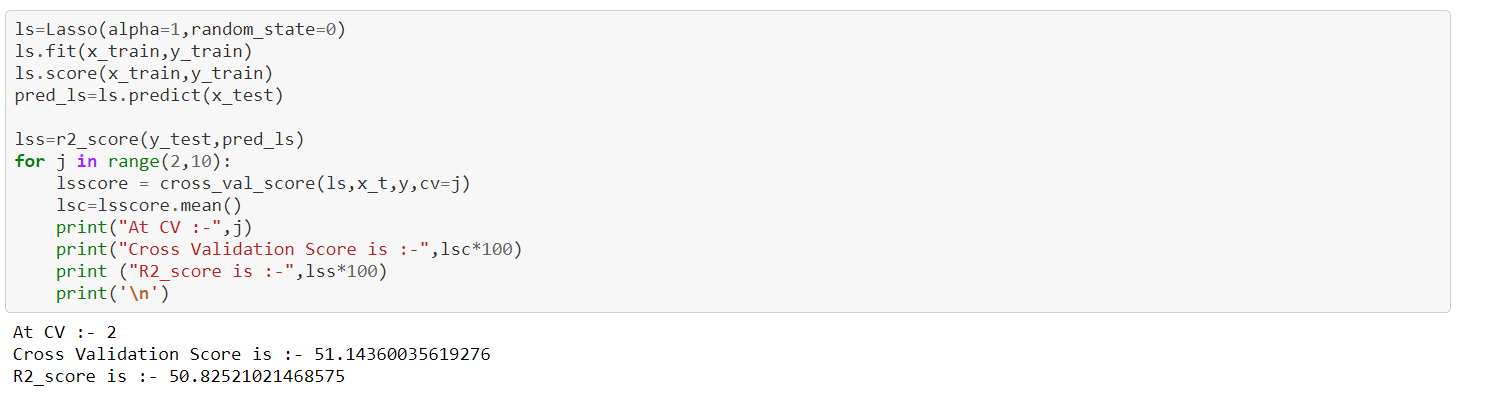
**Identifying the Skewness:**

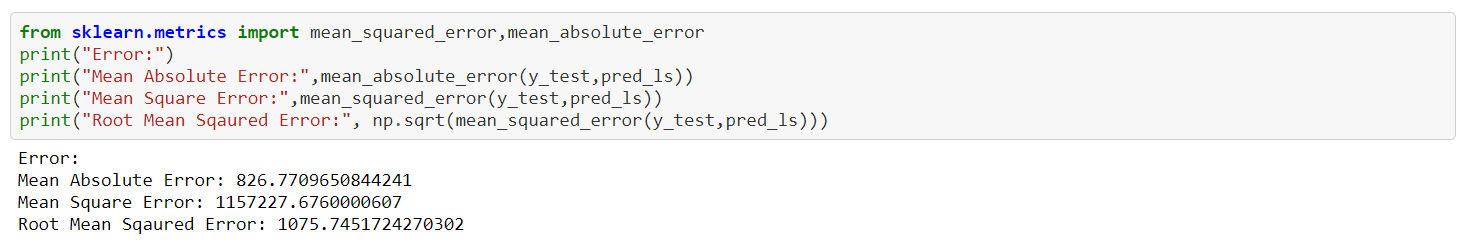
****

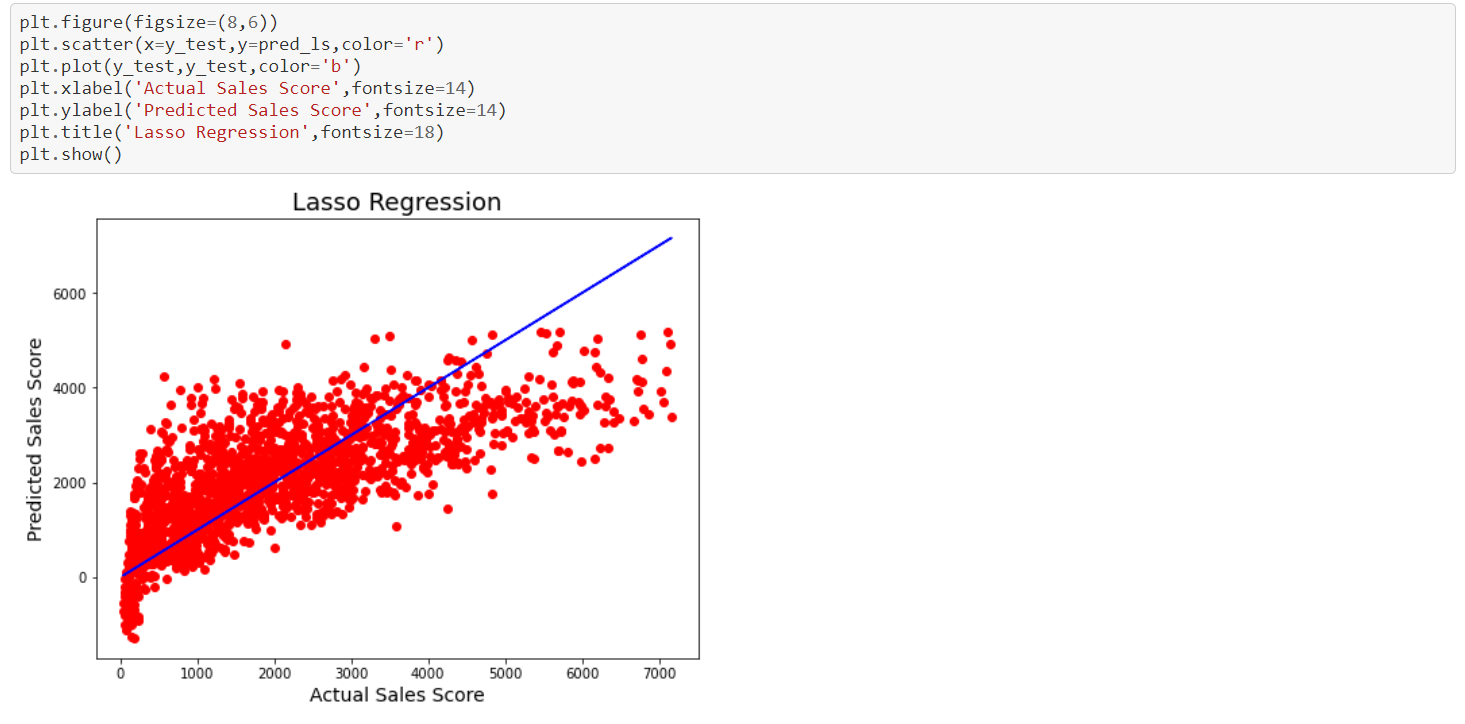
Columns having high skewness are: Item\_Visibility Outlet\_Type Item\_Outlet\_Sales

**Using Regularisation Models To Find Out Predictions:**



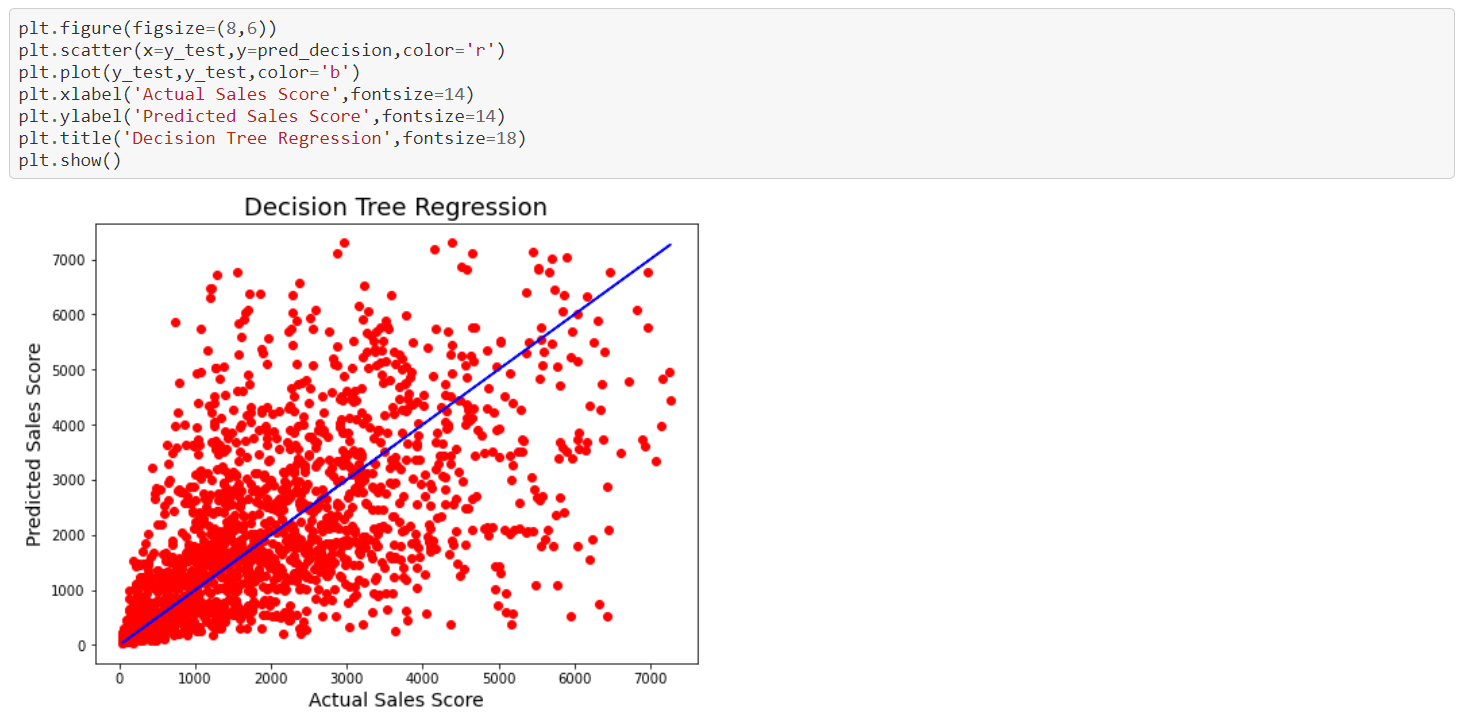




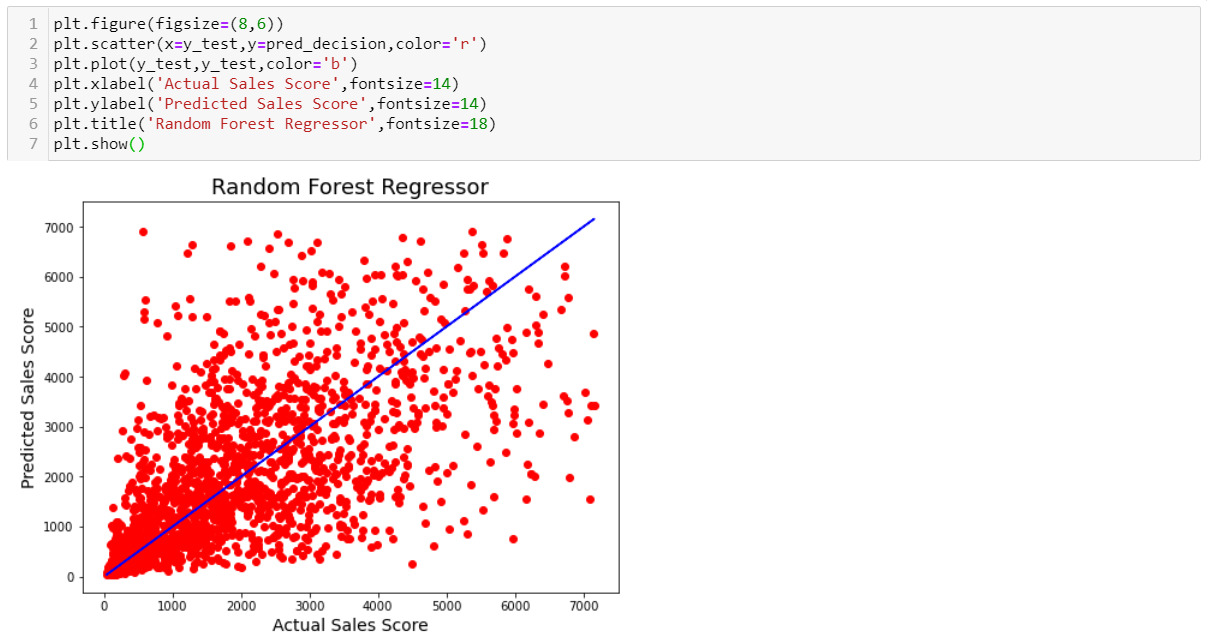


Most of the data points are away from the predicted line. 50% Accuracy can’t be the best model to confirm Lasso as best model to work.

**Ensemblier Techniques**

1. **Decision Tree Regressor**

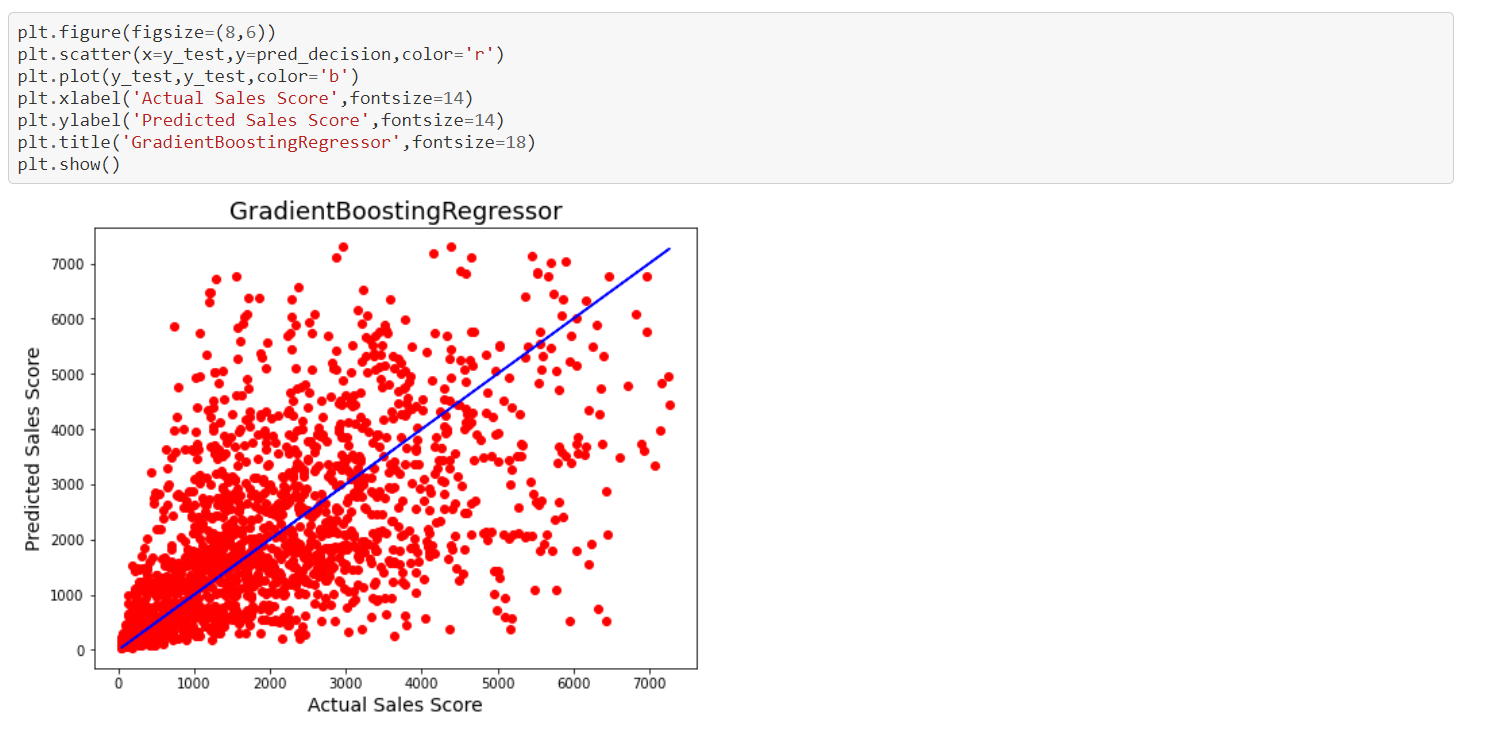
Even after trying decision model we can find data points are far away from predicted line. R2 Score and CV score are not satisfactory.

Using Random Forest Regression

Data Points in Random Forest are also far from each other. That is Cross Validation and r2 Score are less compare with Lasso Regressior.

Gradient Booster Regressor:





Even in Gradient Boosting Regressor we can find r2 score and Cross Validation Score less when compare with Lasso Regression.

**Testing The Dataset with Best Model**

