**Advertising Sales Channel Prediction**

**Problem Statement:**

When a company enters a market, the distribution strategy and channel it uses are keys to its success in the market, as well as market know-how and customer knowledge and understanding. Because an effective distribution strategy under efficient supply-chain management opens doors for attaining competitive advantage and strong brand equity in the market, it is a component of the marketing mix that cannot be ignored.

The distribution strategy and the channel design have to be right the first time. The case study of Sales channel includes the detailed study of TV, radio and newspaper channel to predict the total sales generated from all the sales channel.

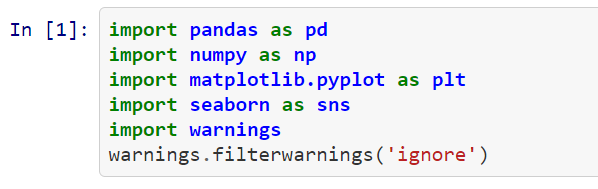
Build a model which predicts sales based on the money spent on different platforms for marketing

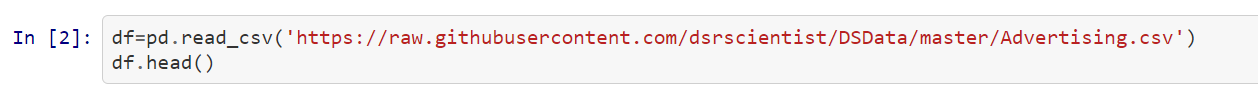
**Dataset Link:**

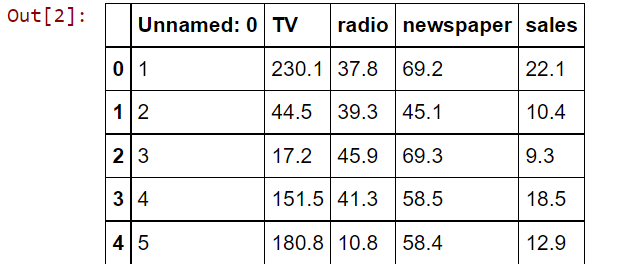
<https://github.com/dsrscientist/DSData/blob/master/Advertising.csv>

**Data Analysis:**

Reading the CSV file and displaying the data:

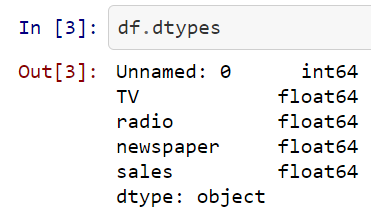




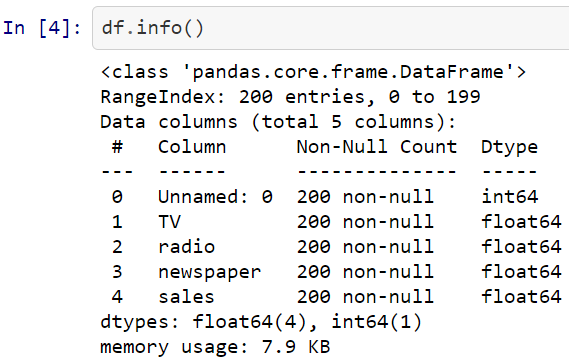


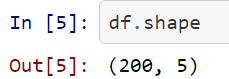
Data Inspection:

Checking for DataTypes



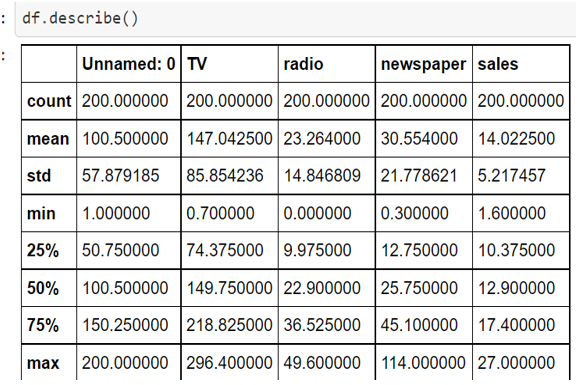
Getting information about the data like null values, data types, columns in the dataset



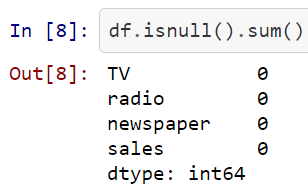


Shows that there are 200 rows and 5 columns

Checking the statistical value for all the columns,

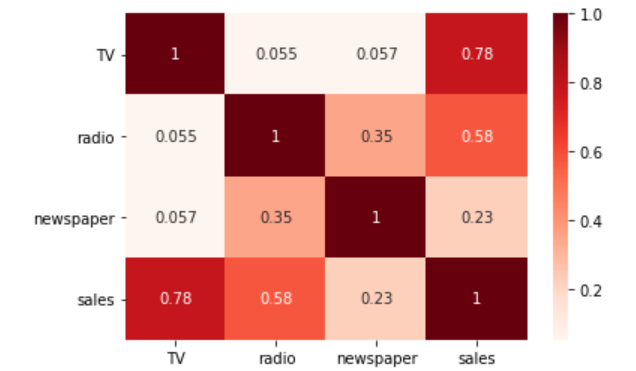


Checking for null values in the dataset



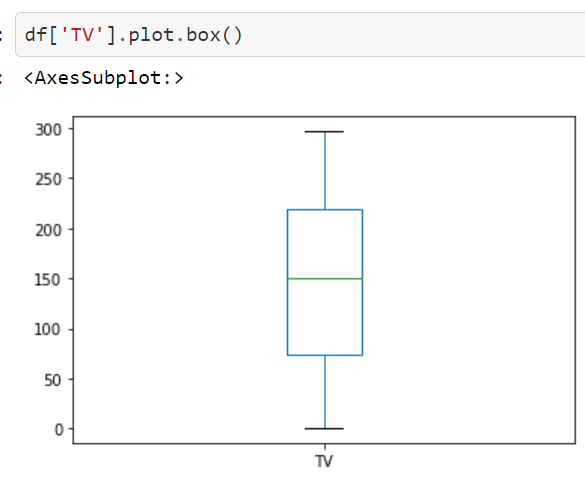
**Exploratory Data Analysis:**

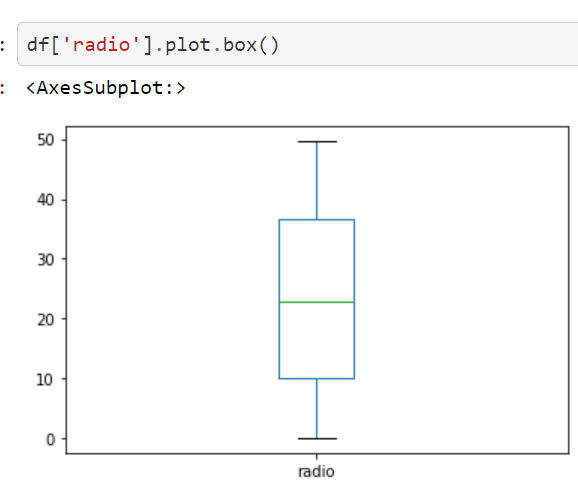


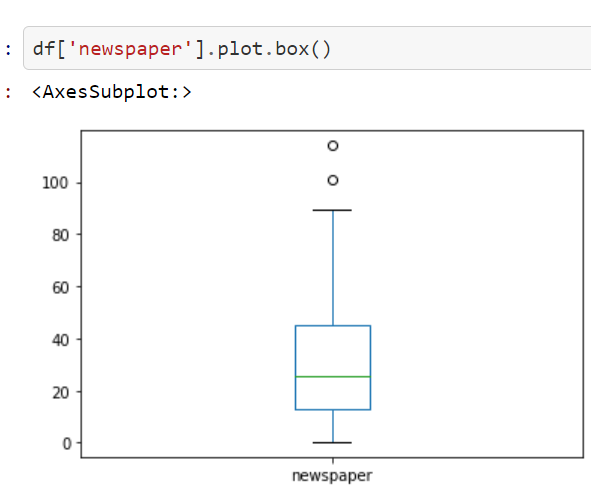


TV and Sales column are highly correlated to each other as compares to other columns

-Univariate Analysis:

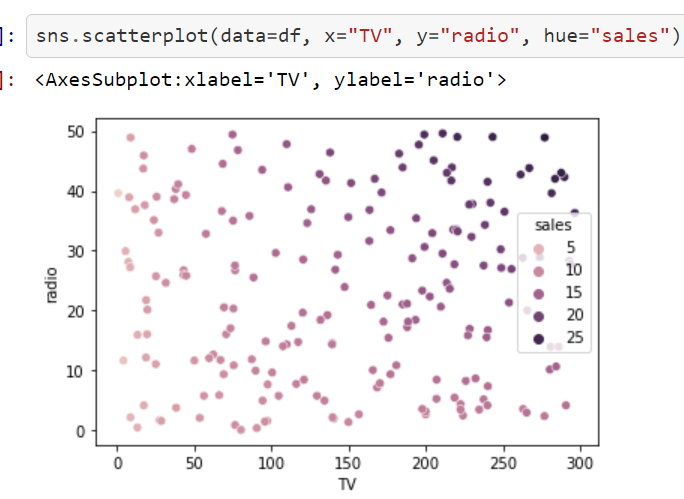




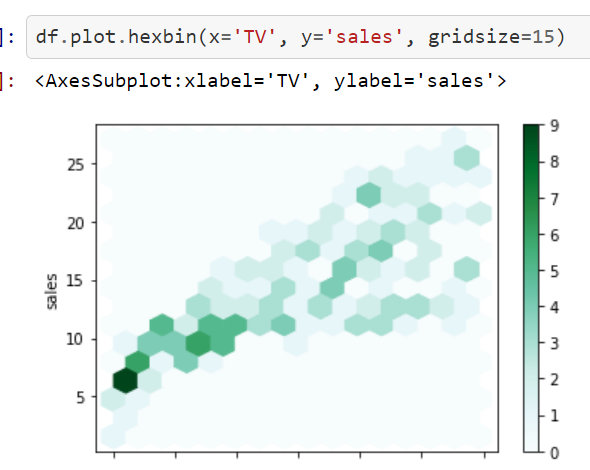


There are no outliers present in TV and Radio columns whereas there are outliers present in Newspaper column

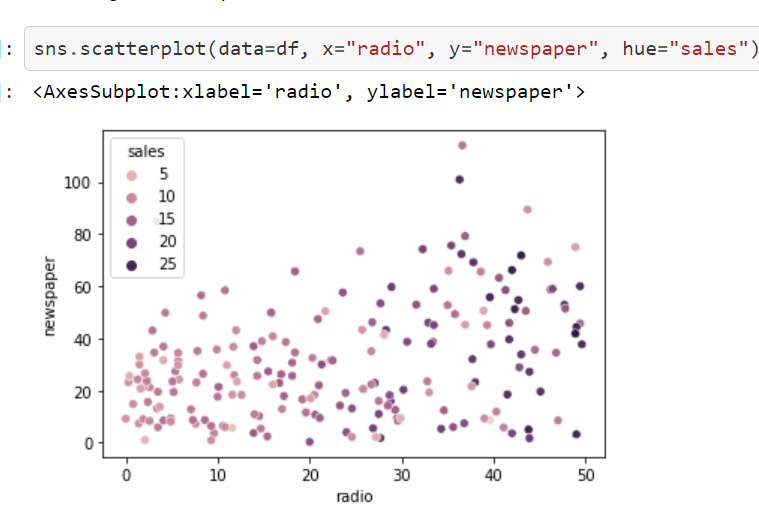
-Bivariate Analysis:



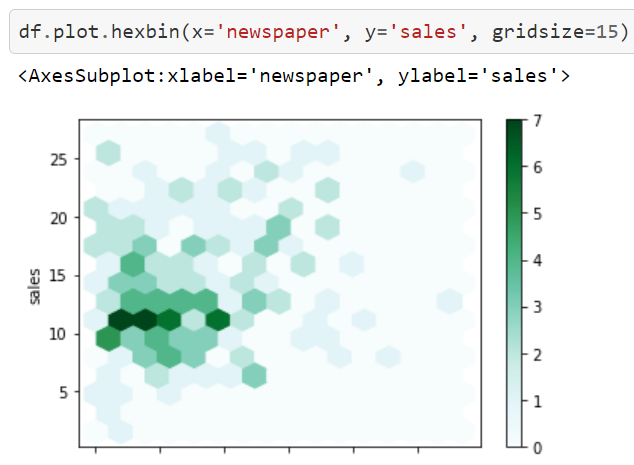
Relationship between "TV" and "radio" columns and data is highly spreaded



Checking relationship between "TV" and "sales" columns

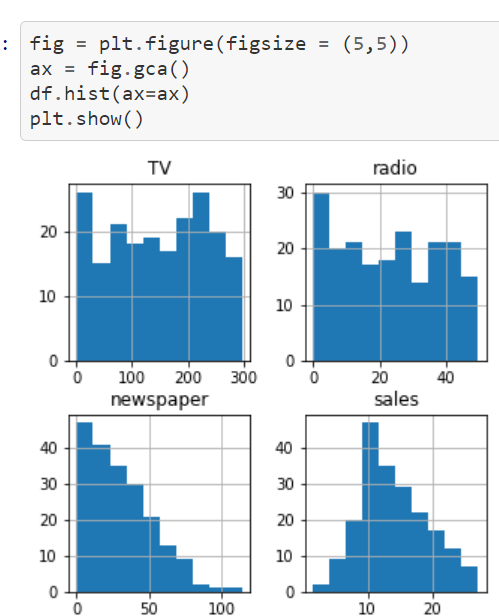


Checking relationship between "radio" and "newspaper" columns and the data is spreading

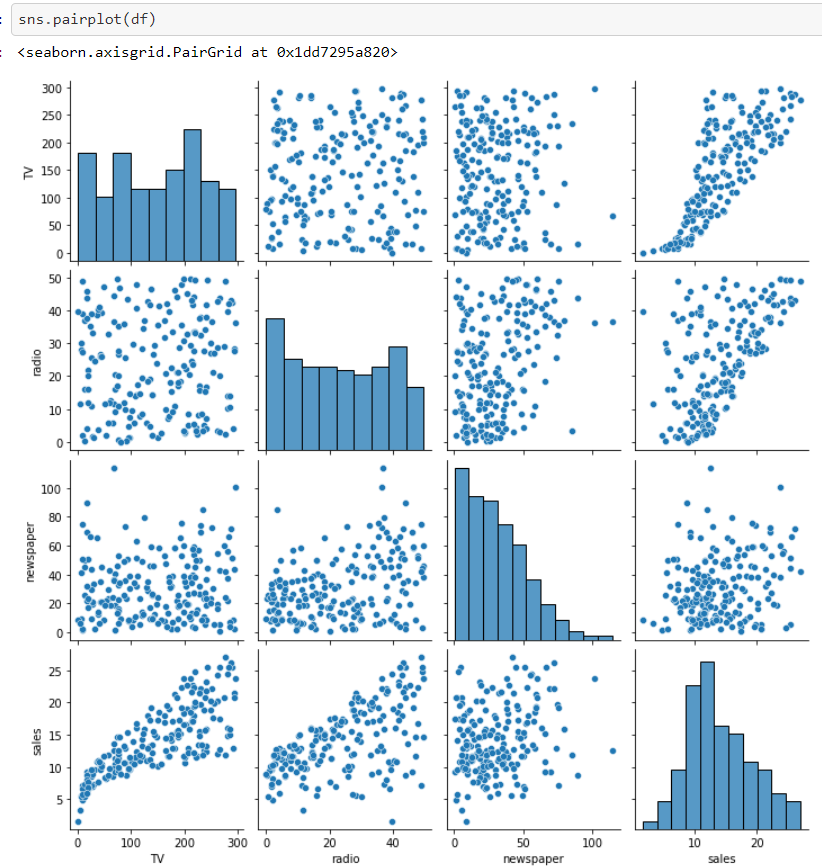


Checking relationship between "sales" and "newspaper" columns and the data is spreading

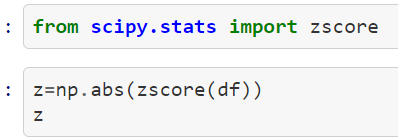
-Multivariate analysis:

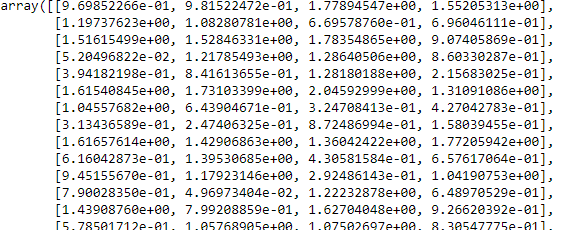


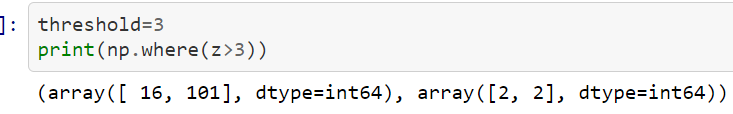
Ploting all columns in one place with the help of histplot



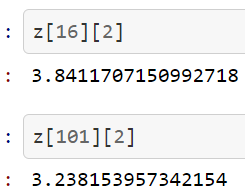
Applying **Zscore** functionality on entire data to remove outliers



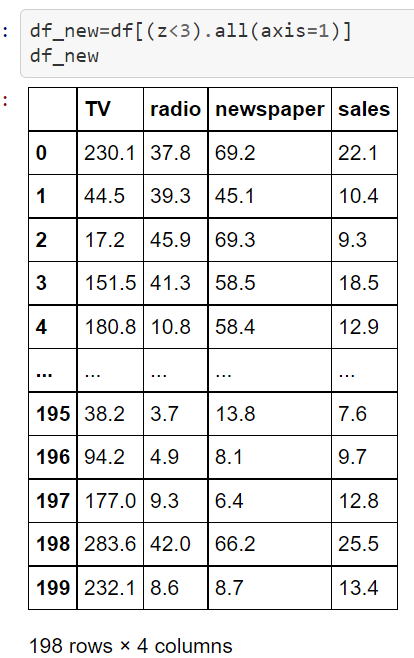




Setting threshold point

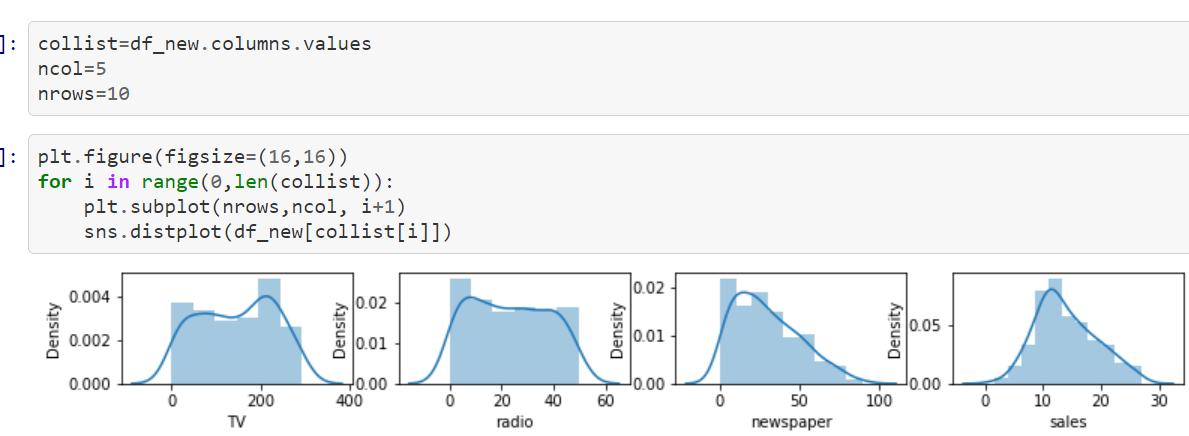


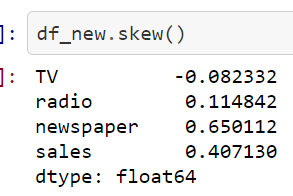
Checking the value whether it is below threshold point or not



With the help of above functionality, we are trying to get all the values that are below the threshold point

-Skewness:

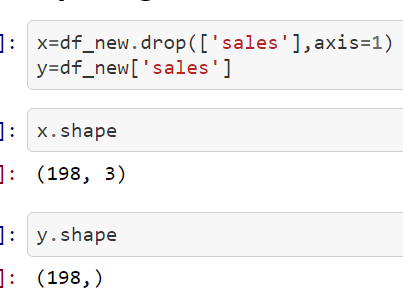




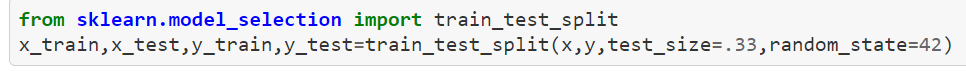
All the variables are already skewed

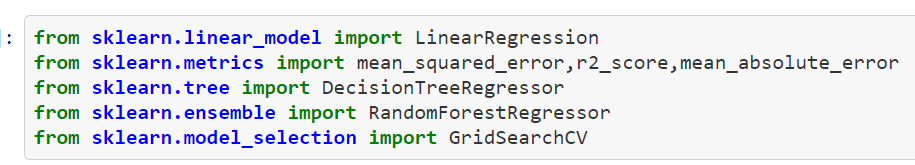
**Building Machine Learning Models**

Splittling X and Y columns. Dropping Sales from X and adding it in Y since it is our Target variable



Splitting X and Y into Train Data and Test Data respectively

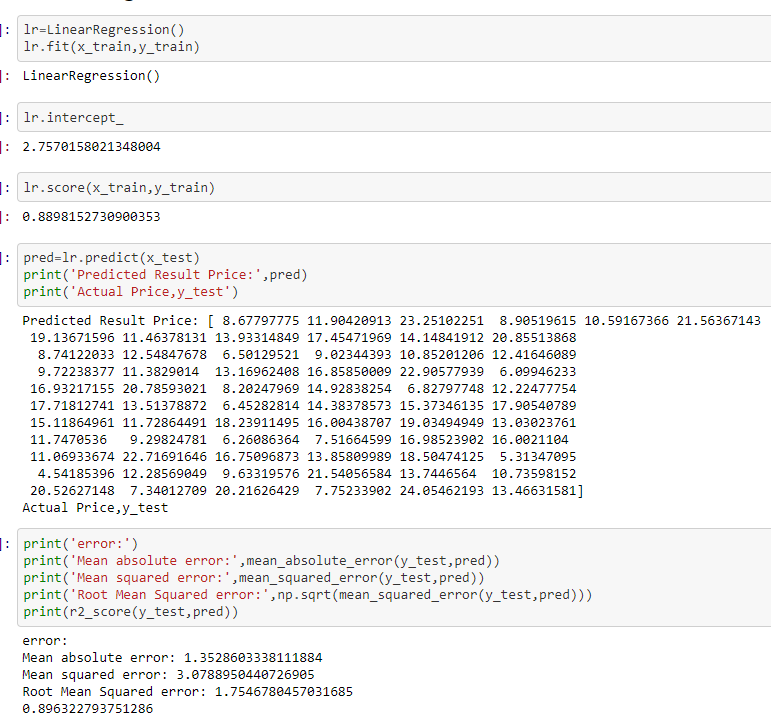




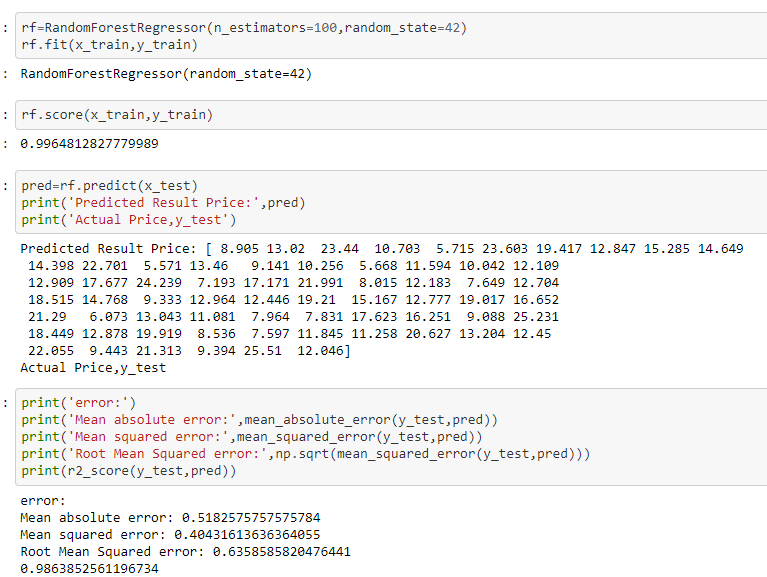
Model Building:

We’ll find out which ML model shows the best accuracy and cross check it doing cross validation and the model which has the least difference between the accuracy and cross validation will be chosen. Hyper parameter tuning will be done further.

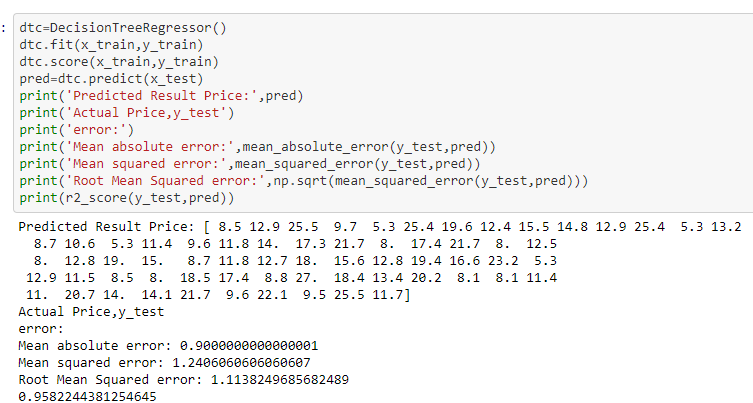
Linear Regression:



Random Forest:

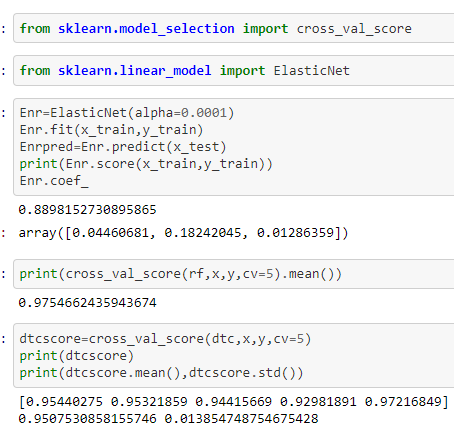


Decision Tree Regressor:

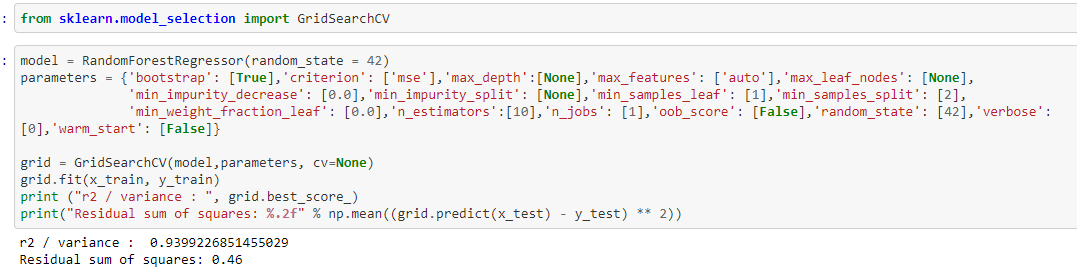


Random Forest seems to be the best model.

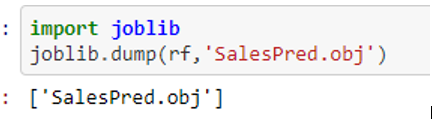
-Cross Validation Score:



-Hyperparameter Tuning:



-Saving the model:



**Concluding Remarks:**

Naturally, there is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features.