1) Problem Definition:-

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

For this problem we are taking advertising dataset given in ISLR and analyses the relationship between 'TV advertising' and 'sales' using regression models. I this dataset the features are TV, Radio and Newspaper marketing spend in thousands of dollars.

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

Features:

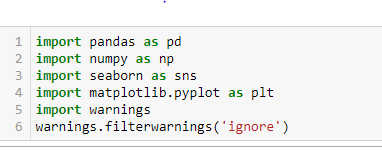
* TV: advertising dollars spent on TV for a single product in a given market (in thousands of dollars)
* Radio: advertising dollars spent on Radio
* Newspaper: advertising dollars spent on Newspaper

Target:

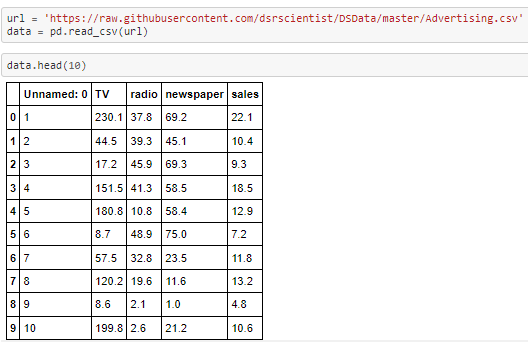
* Sales budget in thousands of dollars

2) Data Analysis:-

**Let’s start with importing the required modules**

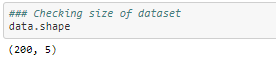


**Importing dataset and see what is inside the data set by seeing the first five values of dataset by head () command.**



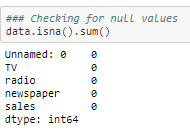
Here Unnamed: 0 column is serial number we can delete that column.TV, Radio, and Newspaper are columns which has advertising investment detail. This are our features. Sales column contain total sales data for company which is our label. The data is in continuous data form.

**Checking size of our dataset using .shape command.**



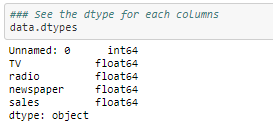
The shape of the data (200, 5) which shows there are 200 Rows and 5 columns in the data.

**Checking for null values.**



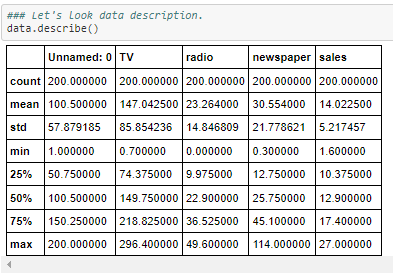
There is no null value in our dataset.

**Checking data type for each columns**



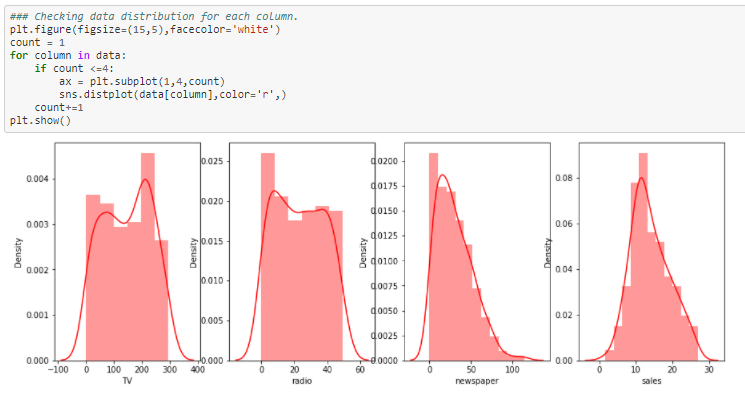
We can see d-type for all the columns in our dataset are in floats type.

**Data Description and Exploratory Visualizations**

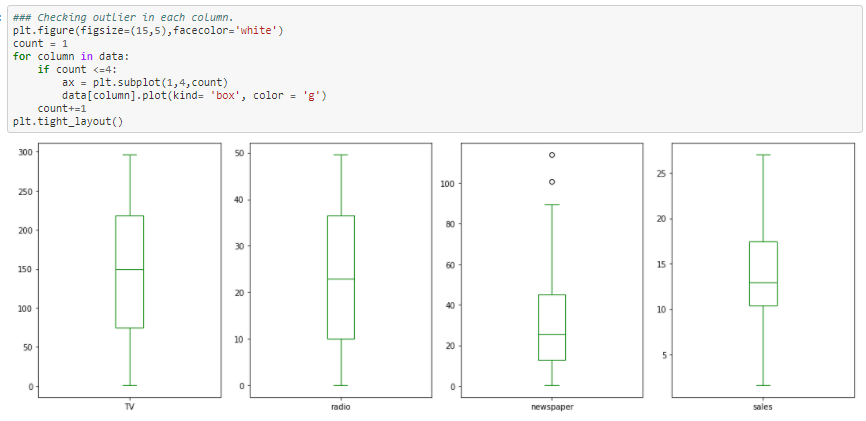


As we seen the first column 'Unnamed: 0’ contains counting of data points hence we can remove this column. The Count for each data point in each column is equal means there is no data lose in any column. The difference between mean and std for each column is fine. Difference between min value, max value, and Quartiles are seems fine for all the columns. Only for newspaper the difference between Q3 and max is more will check with that.

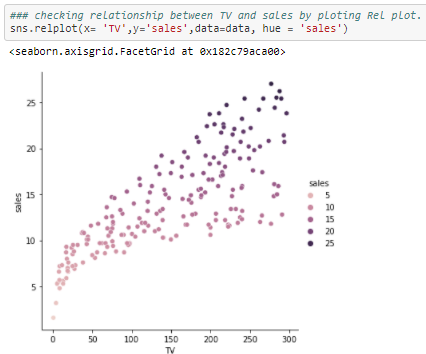
**Let’s plot distribution graph to check how data distributed in our dataset.**

Seems like data is distributed normally. There is slight right skewness in newspaper column.

**As we have skewness in our dataset let’s check for Outlier as well by plotting Box plot**.

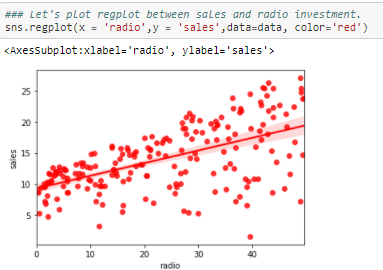
We can see only 2 outlier in Newspaper column.

**Plotting scatter plot between TV feature and sales target**



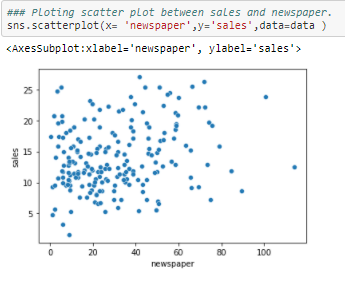
There is strong relationship between TV investment and Sales. As we increases investment in tv our sales also increases*.*

**Plotting regression graph between radio feature and sales target**



The radio investment is also showing some linear relationship with sales but the data points are far away from best fit line if we increases investment in radio.

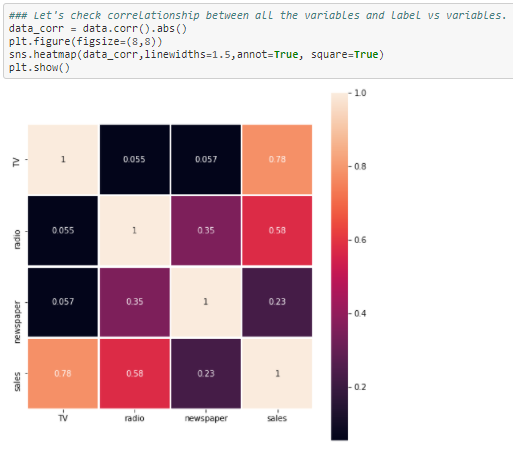
**Plotting Scatter graph between radio feature and sales target**



We can see there is no relationship between newspaper and sales we can say Investment in newspaper does not show any relationship with sales.

**Correlation**

Let’s take a look at some of most significant correlations. It is worth remembering that correlation coefficients only measure linear correlations.



As shown in above heat map, TV and radio shows 78% and 58% relationship with Sales. Newspaper shows 23% relationship with Sales. No correlation found between variables.

3) EDA Concluding Remarks:-

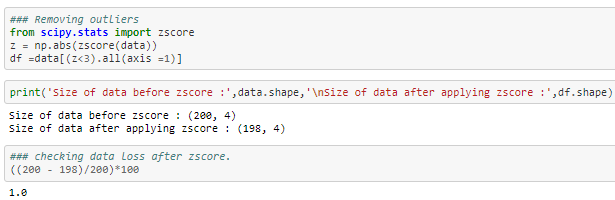
* There is no null values in our dataset
* Found that data in newspaper column is slightly right skewed.
* We observed that newspaper column contain 2 outliers.
* TV and radio shows 78% and 58% relationship with Sales. Newspaper shows 23% relationship with Sales. No correlation found between variables.

4) Pre-processing Pipeline:-

We undertake data pre-processing steps to prepare the datasets for Machine Learning algorithm implementation.

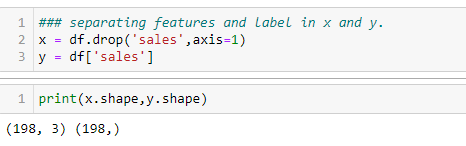
**Applying z-score to remove outliers.**

A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of [standard deviations](https://www.investopedia.com/terms/s/standarddeviation.asp) from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.



After removing outlier we always check for data loss in this problem data lost only 1% data is fine.

**Separating Independent variables and Target variable**

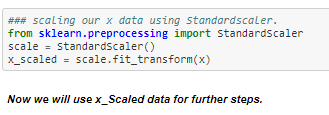


In X we stored Features and in Y we store our Target variable.

**Standardization**

Standard Scalar follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance

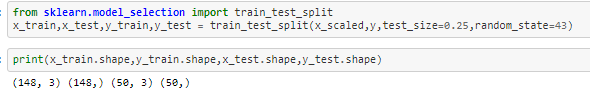
Before implementing models we need to normalize our features to remove Biasness. For that we use Scaling technique.



**Train - Test Split**

Prior to implementation or applying any Machine Learning algorithms, we must decouple training and testing data from our master dataset

Importing Train – test split from Sk-learn library:



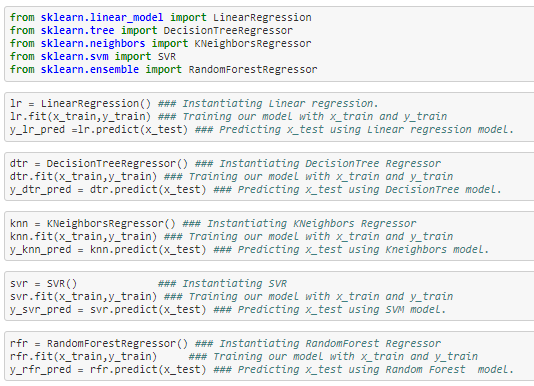
We have divided data into training and testing for model building we will use 75% data for training and use 25% for testing purpose. After checking size of our training and testing data.

5) Building Machine Learning Models:-

In [statistical modeling](https://en.wikipedia.org/wiki/Statistical_model), regression analysis is a set of statistical processes for [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the relationships between a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) (often called the 'outcome' or 'response' variable) and one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable) (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is [linear regression](https://en.wikipedia.org/wiki/Linear_regression), in which one finds the line (or a more complex [linear combination](https://en.wikipedia.org/wiki/Linear_combination)) that most closely fits the data according to a specific mathematical criterion.

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

**Let’s Import regression models, instantiating them then Train them on training data and then testing them by predicting test data.**



**After predicting test data let’s check evolution metrics and R2 score to check model accuracy.**

It is necessary to obtain the accuracy on training data, but it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.

So to build and deploy a generalized model we require to evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result.

* Mean Absolute Error (MAE)

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

* Mean Squared Error (MSE)

MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

* Root Mean Squared Error (RMSE)

As RMSE is clear by the name itself, that it is a simple square root of mean squared error.

**R Squared (R2)**

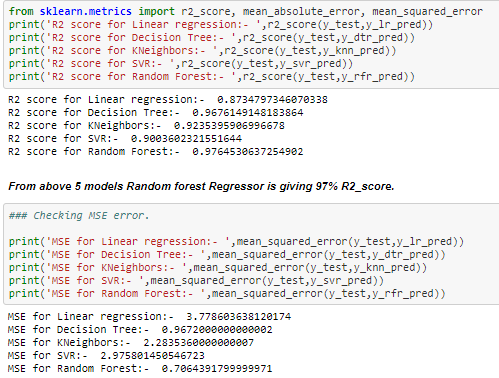
R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides. The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically R2 squared calculates how must regression line is better than a mean line.

Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

**In our model we are using R2 score for accuracy check and Mean squared error as evolution metrics**

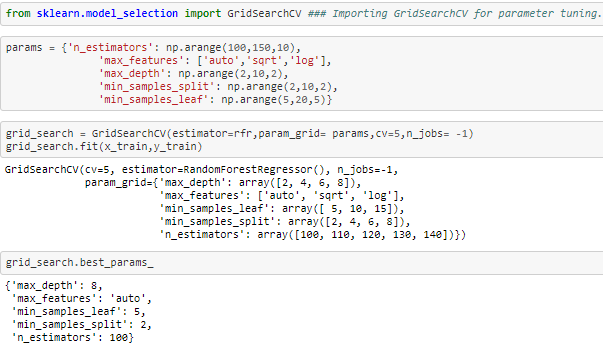


Out of all 5 model we found that Random forest model is giving best R2 score 97% and also making least Mean squared error 0.7%.

**Hyper parameter tuning using Grid Search CV**

Hyper parameters tuning is crucial as they control the overall behavior of a machine learning model. Every machine learning models will have different hyper parameters that can be set. A hyper parameter is a parameter whose value is set before the learning process begins. Grid search is arguably the most basic hyper parameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyper parameter values provided, evaluating each model, and selecting the architecture which produces the best results**.**

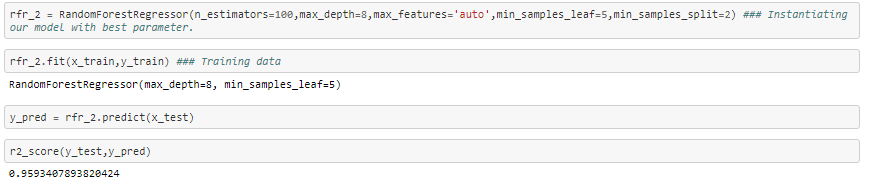
**Let’s import Grid search CV from SK-learn library**



We got the best parameters by using grid search CV.

**Instantiating our model**

We get the best parameters by Grid search CV. Now let’s instantiating our model again with parameters calculated by Hyper parameter



After parameter tuning on Random forest its giving is 95% and MSE 0.7%.

We didn't see any increasing in R2 score and MSE percentage. Additionally, it is not even guaranteed to find the best solution, often aliasing over the best configuration. After using Grid search CV.

6) Concluding Remarks:-

Based on the all the graphs and plots we come to an conclusion that not all input features are essential and affect the data, for example from the scatter plot against Sales and newspaper. We see that as we increases investment in newspaper there is no drastically increment in sales. So this feature is not so essential as compared to others like TV and radio, so we can drop this feature while feature selection.

For predicting sales, we have implemented multiple algorithms, namely:

1. Linear Regression
2. Decision Tree Regressor
3. K-Neighbors Regressor
4. SVR
5. Random Forest Regressor

We were able to achieve maximum accuracy using linear regression of 87%. Decision Tree Regressor giving an accuracy of 96%. K-Neighbors Regressor has an accuracy of 92%. SVR giving accuracy of 90%. And Random Forest Regressor as 97%.