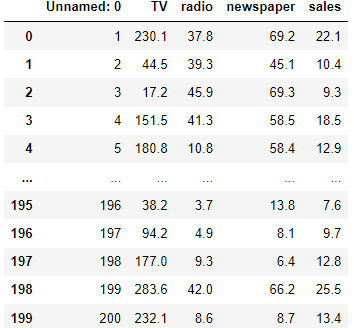
**ADVERTISING SALES PROJECT**

**PROBLEM DEFINATION:-**

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.

Below is the snapshot of the dataset;



Where unnamed column act as an identity for the entries present in the data.

TV – this input variables represents the influence of Television ads on sales.

Radio – this variable represents the influence of radio advertising on sales

Newspaper – it represents the influence of newspaper ads on sales.

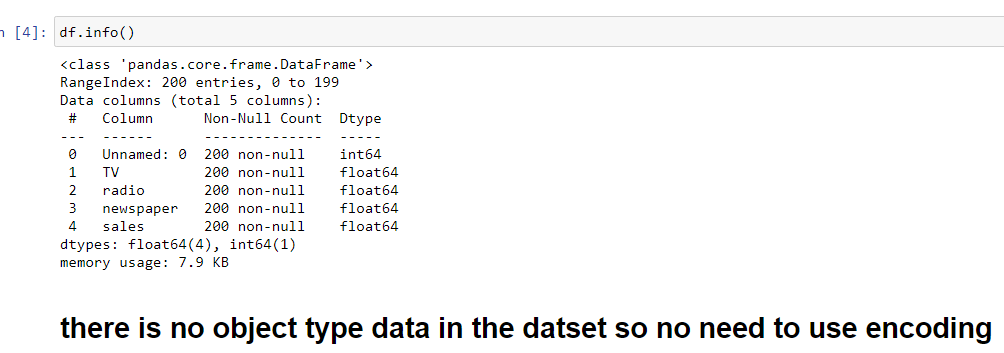
Sales – it represents the total sales generated.

We need to predict the total sales generated using the input variables given and help develop better advertising strategies to generate more sales for the particular product and generate more profit by making better decisions based on the analysis.

Sales is our dependent variable.

**DATA ANALYSIS :**

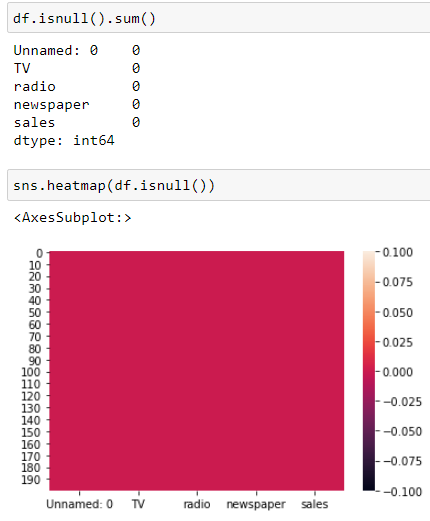
We checked the information the dataset and its data types. Below is the snapshot :



What we saw that all the variables are in numerical form and there is no categorical data present in the data.

This data contains 200 rows and 5 columns.

Below is the snapshot of the heatmap of the nan values if present

****

What we saw is that there are no nan values present in the dataset and our data is clean as of now. We can move forward and do some EDA on the data to check correlation of variables with the sales.

**EDA AND CONCLUDING REMARKS :**

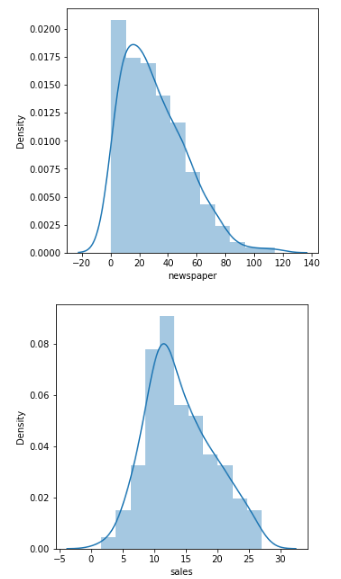
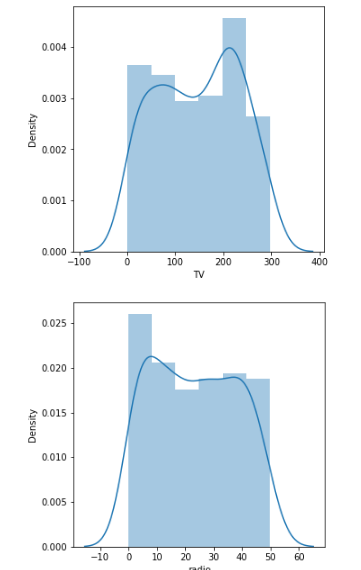
We decide to drop unnamed column as it only gives the identity to the particular entry.

We will be performing some eda and visualization and give our observations to the analysis we have done.

We will try to find out the relations between the input variables with the output variables which is sales.

Below are snapshots of distribution plot to check if there is any biasing or skewness present in the data.

Graph :-

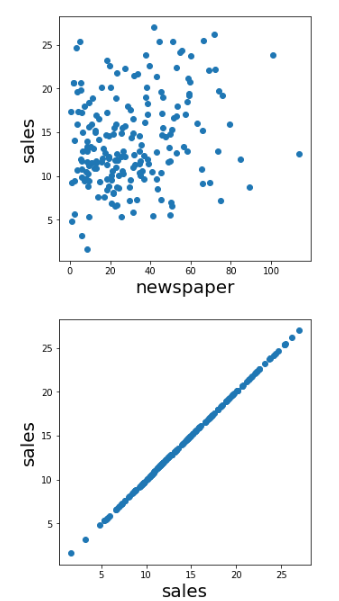
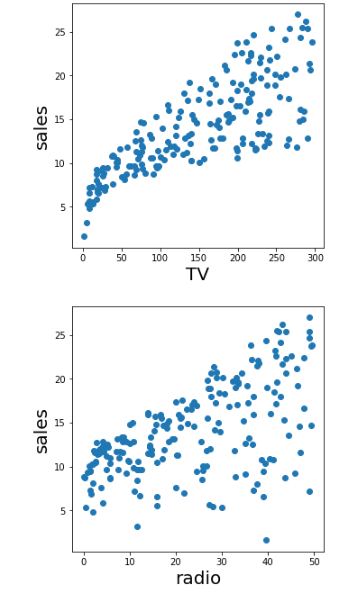


**Observation:**

We can see that there is skewness present in the newspaper variable.

Tv and radio shows some multimodal distribution in the data.

We will be plotting some scatter plot to check the relation of input variables with the sales column.



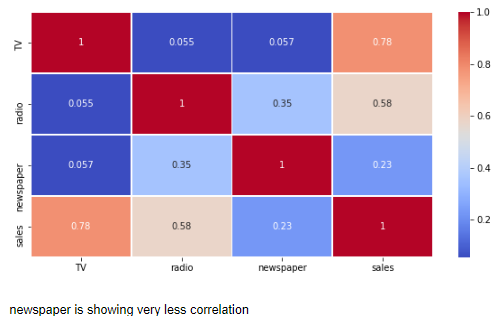
from above we can conclude that :-

tv ads have a highest positive impact on sales

radio has quite positive response on sales

newspaper is not showing any good relation to sales where the point are all scattered.

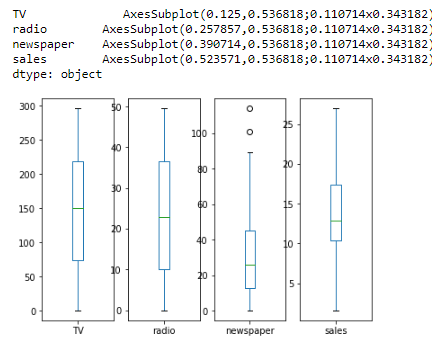
We will check the correlation between the variables with each other using heatmap to know which variables are highly correlated or positively correlated with the output variables.



What we saw is that the television is highly correlated with the sales and and the least correlated variables is newspaper.

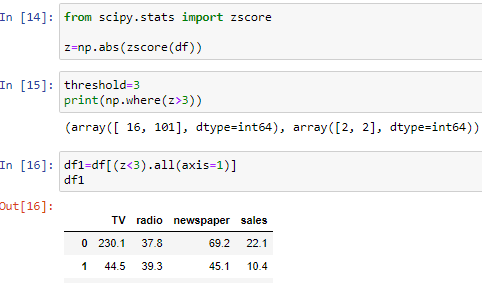
Checking outliers and removing them before building our prediction models.

We used box plot to check the outliers.



What we observed from the above boxplot is that newspaper variable shows some signs of outliers present in that column.

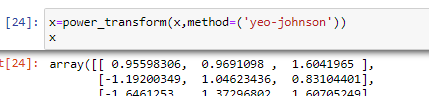
We used zscore method to remove the outliers.



**PRE-PROCESSING**

We split the data into X and Y where x hold all the input variables and y holds sales variable.

Power transform method was used to remove the skewness and biasing from the data so that the model could perform accurately and give better predictions.



We used standard scales to scale the data and make it better so that it will help in making better and accurate predictions.

**BUILDING MACHINE LEARNING MODELS**

Machine Learning analyzes large chunks of data automatically. Machine Learning basically automates the process of Data Analysis and makes data-informed predictions in real-time without any human intervention. A Data Model is built automatically and further trained to make real-time predictions. This is where the Machine Learning Algorithms are used in the Data Science Lifecycle.

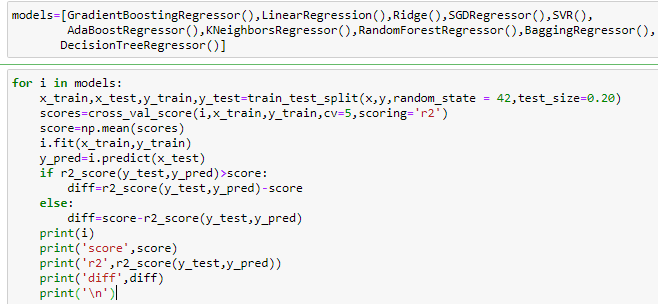
As our sales variable is a continuous data it clearly indicates that it is a regression problem and we will be using regression models to predict the data for the sales variable.

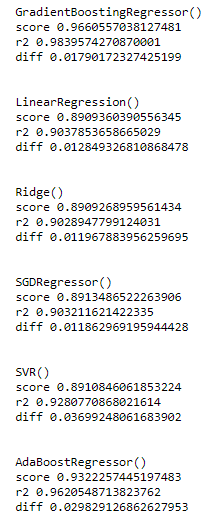
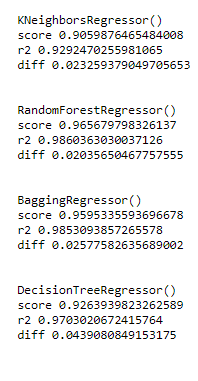
We need to first train the data with 80% of the data set and then the model will learn from the patterns of the dataset and it will help the model to predict the data on test data which is other 20% of the dataset.

We will be splitting the data of x and y into xtrain, ytrain and xtest,ytest with the help of train test split.

We included cross val score and cv set to 5 and checking the mean score. Random state is set to 85, we will be checking the r2 score and difference between the r2 score and cross val mean score. Where diff represents the difference in cross\_val mean score and r2\_score and check the difference. We can’t choose the model if the difference is high as it leads to overfitting and underfitting problems.

Below are the images of the results that we got after using machine learning models



What we observed is that :-

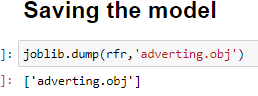
Random forest and gradient boosting regressor gave us the accuracy of mor that 98% and with least difference between the cross\_val mean score and r2 score.

We will try to hyper tune these models and check if we can improve the accuracy of these models.

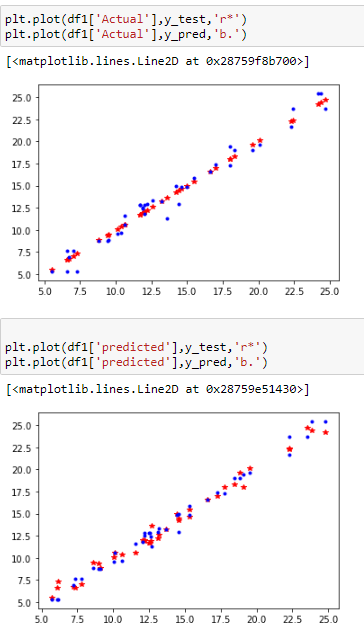
Below is the image of the hyper tuning of the models like gradient boosting regressor and random forest regressor.



After hyper tuning we got the best model with more than 98% accuracy was random forest and we choose to save the model for predicting sales variable for this data.



We saved the actual data and predicted data in a dataframe and also tried to plot this data and check how the actual data and predicted data are close to each other.



From the above image we can say that the data very similar to each other and that our model is giving us very good results.

**CONCLUDING REMARKS**

We had to predict the sales for the advertising data which was a very intresting data to work on.

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.

Using the data science we were able to analyze some important feature from the data like we got to know that people get more attracted to television ads compared to other advertising channels.

Newspaper and radio have less influence compared to television when it comes to advertising and sales.

After building the model we were able to achieve the accuracy of more than 98% accuracy and our predicted data was very similar to actual data.