**Advertising Sales Channel Prediction**

**Problem Definition :**

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.

Features:

* TV: advertising dollars spent on TV
* Radio: advertising dollars spent on Radio
* Newspaper: advertising dollars spent on Newspaper

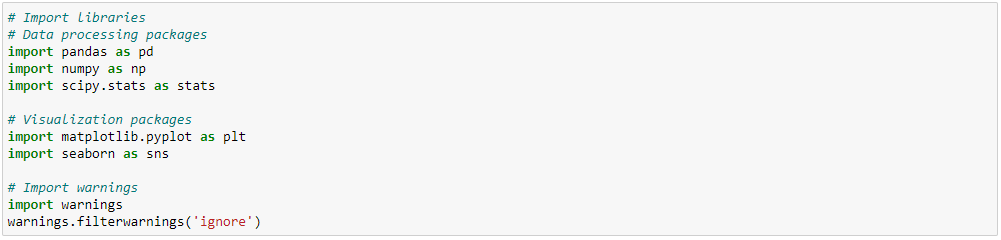
Target:

* Sales budget in thousands of dollars

We are going to build a model which predicts sales based on the money spent on different platforms for marketing and then we will predict the total sales generated from the entire sales channel.

**Importing Python Libraries**

The Python Standard Library is very extensive that contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc. Most of the Python Libraries are written in the C programming language.



**Importing dataset and Identifying target and independent features**



Dataset is having numeric type columns i.e. continuous values in the data. Since, target column is having continuous values in it, this is a regression problem.

Regression is an important machine learning model for these predicting sales kinds of problems, where we can fit a line of high sale and low sale. It required significant amount of data for training and testing of the model.

**Data Analysis :**

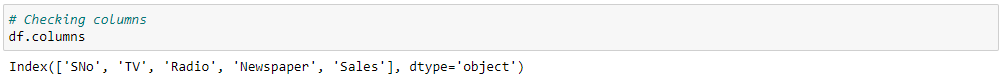
Data Analysis is the process of systematically applying statistical and/or logical techniques to draw inferences for good accuracy of model building.

**Checking the shape of the dataset**



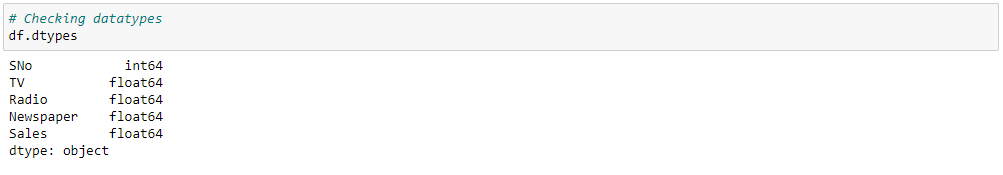
Dataset comprises of 200 observations and 5, out of which 1 is dependent variable and rest 4 are independent variables.

**Checking the column details**



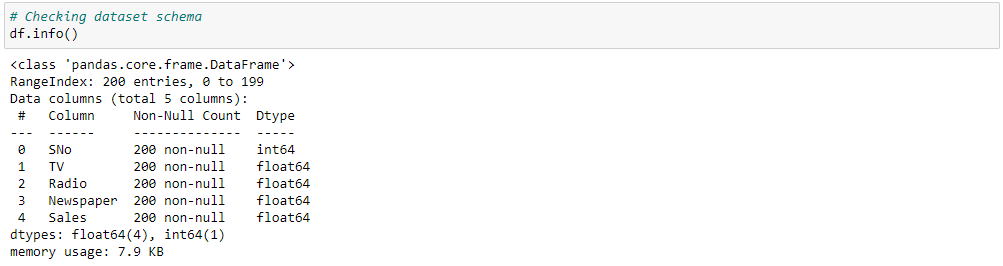
Columns names are displayed where 'Sales' is the target variable and all other are independent variables.

**Checking the data type of the columns to check if feature engineering is required**



All data types are continuous in nature i.e. integer and float.

**Checking Dataset summary with df.info()**

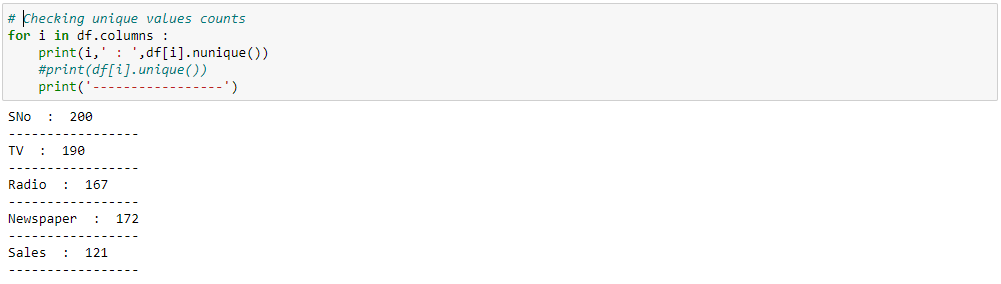


Here, we get the information regarding dataset i.e. RangeIndex, Dtype, Non-null count. There are no nulls in the dataset. Data types also seem good.

**Data Preprocessing :**

Data preprocessing is the process of transforming available data in understandable format to ensure high-quality data before applying machine learning or data mining algorithms. It includes: data cleaning, data integration, data reduction, and data transformation.

**Checking for the unique values in the dataset columns**



Column 'SNo' is having 200 unique values which are equal to total number of rows in the dataset and hence it is not relevant in model training and we can drop it.

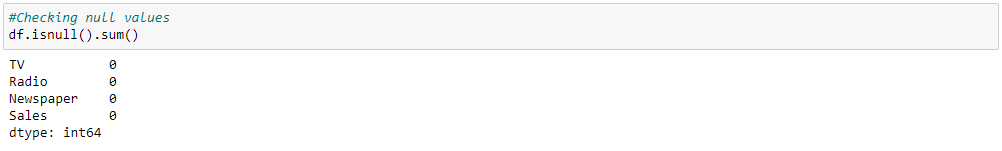


Dropped the irrelevant column to proceed further with the analysis.

**EDA and Visualizations :**

Exploratory Data Analysis refers to the critical process of investigations on data so as to discover patterns, anomalies and to check assumptions with the help of summary statistics and graphical representations to understand the data better.

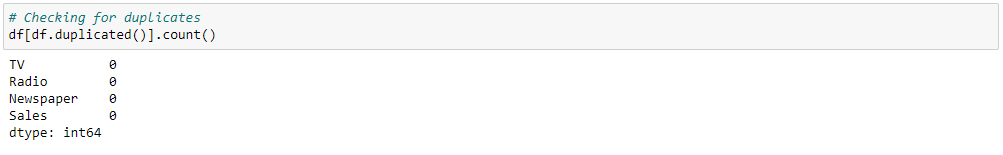
**Checking and visualizing missing values in the dataset using isnull()**





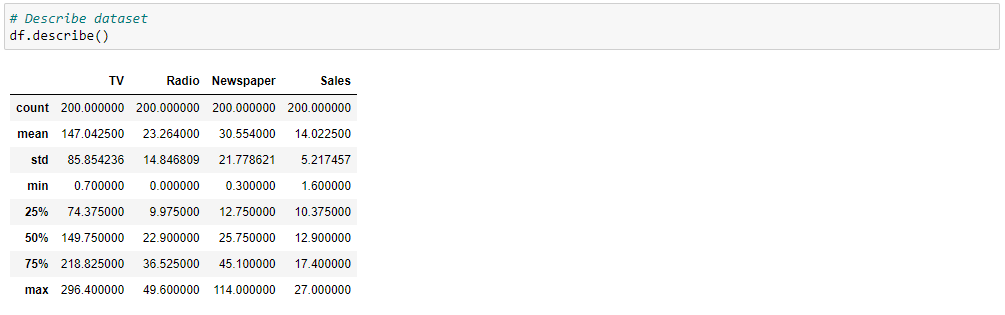
There are no null values in the dataset.

**Checking the presence of duplicates in the dataset**



There are no duplicate rows in the dataset.

**Describing the Aggregated Statistical Figures with describe() function**This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.



Values in column 'TV' ranges from 0.7 to 296.4. Values in column 'Radio' ranges from 0.0 to 49.6. Values in column 'Newspaper' ranges from 0.3 to 114.0. Values in target column 'Sales' ranges from 1.6 to 27.0.

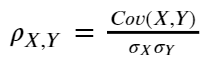
TV is left skewed (mean<median). Radio, Newspaper & Sales are right skewed (mean>median).

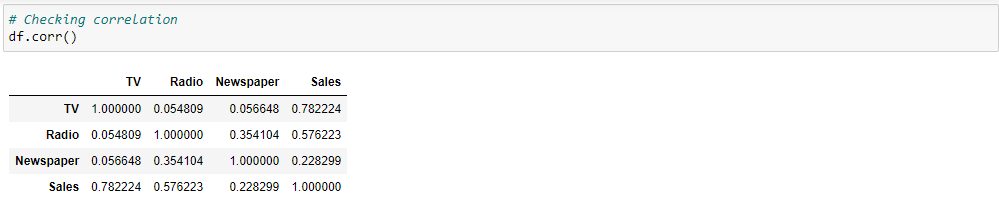
All the columns have high standard deviation except target column. Difference in 75% and max is high in Newspaper hence Outliers might exist.

**Correlation**

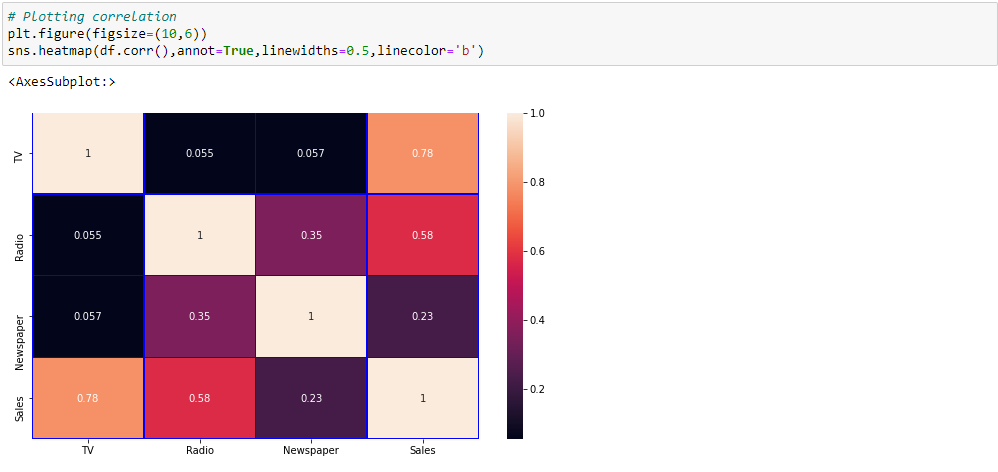
**Correlation** in Statistic is a measure of strength of a relationship between two continuous variables. Pearson Correlation is one of the most used correlations that measures the linear relationship between continuous variables and has a value range of -1 to +1. Towards -1 is strong negative linear relationship, towards +1 is strong positive linear relationship and towards 0 is no correlation that is no linear relationship.

**Formula of correlation :**





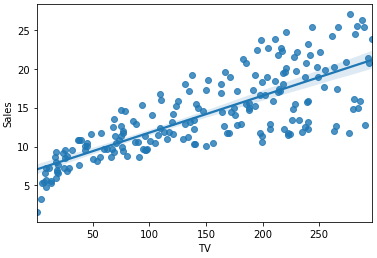
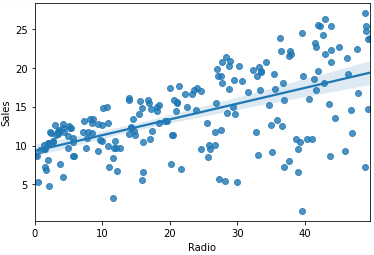
**Heatmap** showing correlation between different variables

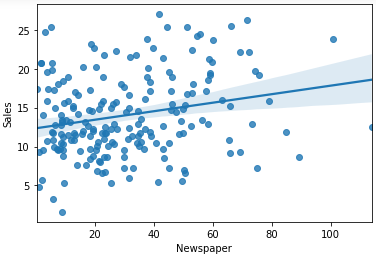
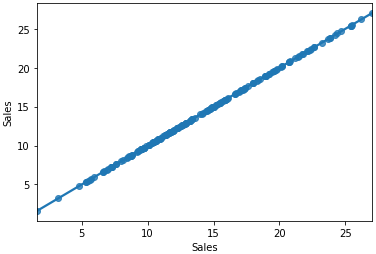


TV and Radio are strongly linearly correlated with target column Sales. Newspaper has a good correlation with target. Radio and Newspaper has good correlation. Multicollinearity might exist.

**Regplot** is used to plot data and a linear regression model fit. It gives a best fit line on which maximum points are lying.

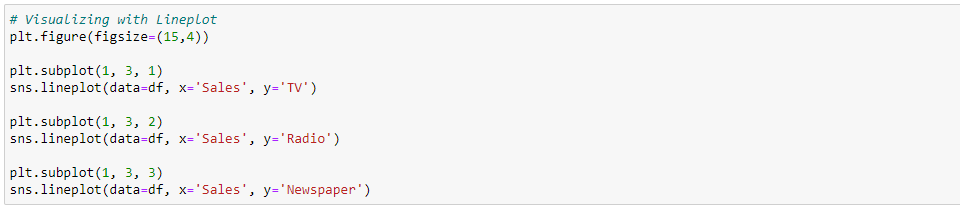


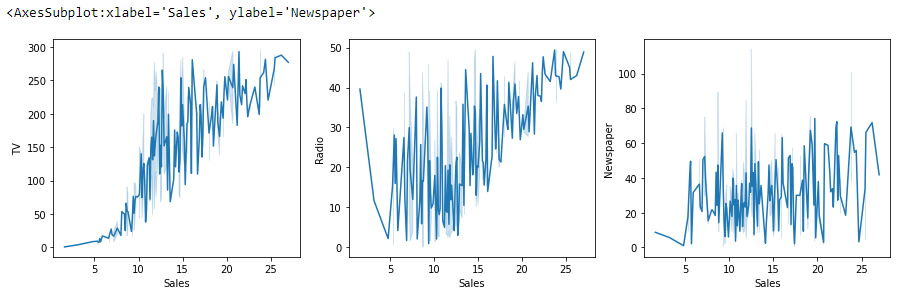
 

Plot showing high linear correlation of TV and Radio with target column Sales. Newspaper also has a good correlation with target.

**Line plots** give annotation to each of the points

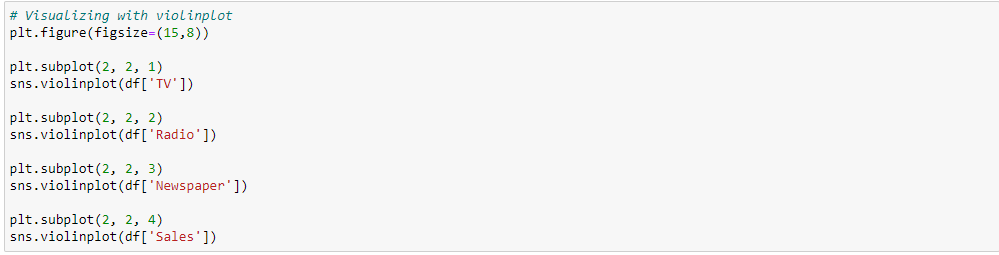


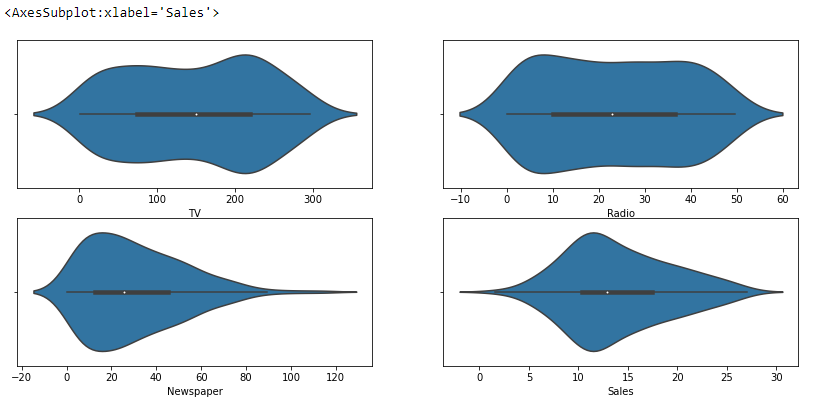


**Observations of Lineplot are as below :**

There is increase in Sales with increase in the money spent on TV for advertising.  
Initial investment in Radio for advertisement is high for making the increase in Sales and there are many ups and downs of money spent on Radio for advertising to have an increase in the Sales.  
Plot is not showing any defined relation of money spent on Newspaper for have an increase in Sales. Here, with some balanced investments on Newspaper advertisement, Sales is growing.

**Violin plots** are used to visualize data distributions and show the probability density of the data at different values. It also shows summary statistics such as mean/median and interquartile ranges.



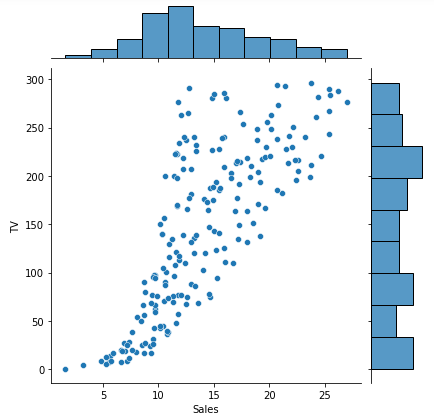
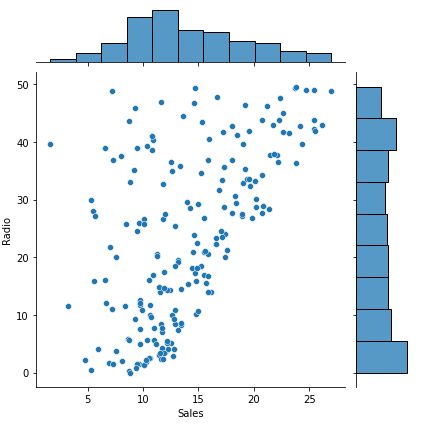


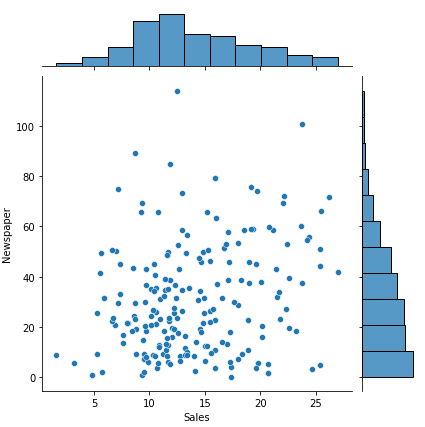
Violinplot observations are as below :

Money spent on TV is highly between 150 to 250. Money spent on Radio is mostly from 4 to 15. Money spent on Newspaper is mostly from 8 to 23. Sale is highly between 9 to 16.

**Joint plot** comprises of three plots out of which, one plot displays a bivariate graph that shows the variation in dependent variable with the independent variable. One of the other two univariate plots is placed horizontally at the top and other at the right margin of the bivariate graph showing the distribution of the independent and dependent variables respectively.



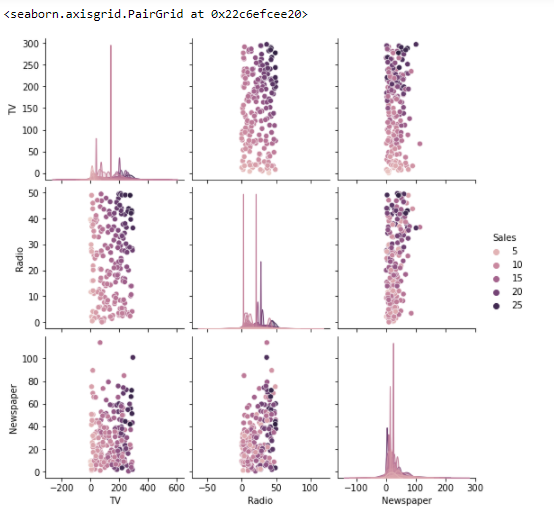
 



Jointplot showing money is highly spent on TV between 150 to 250 with continuous increase in Sales. Highest money spent on Radio was in the initial phase. Initially money spent on Newspaper is higher and high amount on Newspaper advertisement is least paid. Hence, we can say Newspaper is the cheapest medium for advertisement.

**Pair plot** shows pairwise relationships of all the variables in a dataset



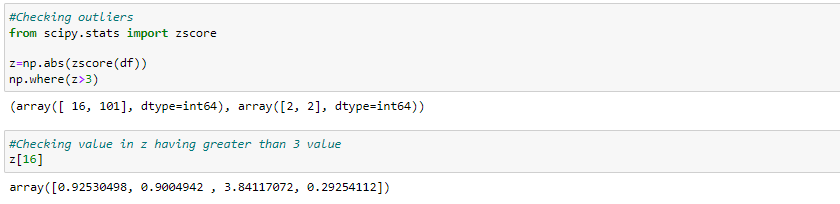


We can see there is overlapping of data in the plot.

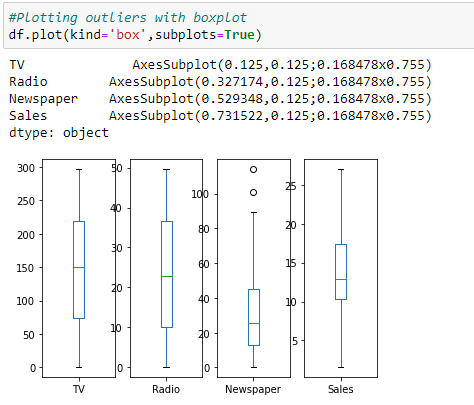
‘Sales’ is highest when money spent on TV is above 150 and on Radio is above 30. Sales is highest when money spent on TV is above 170 and on Newspaper is between 0-100. Sales is highest when money spent on Radio is above 30 and on Newspaper is between 0-100.

**Outliers Detection**

An **outlier** is an observation that lies at an abnormal distance from other values in a sample of a population. **z-score** is a statistical technique through which we can detect the outliers. As per empirical rule, 3 standard deviation is the threshold value.



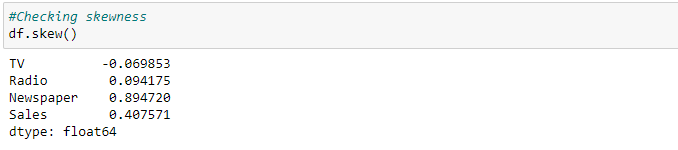
**Detecting outliers using visualization :**

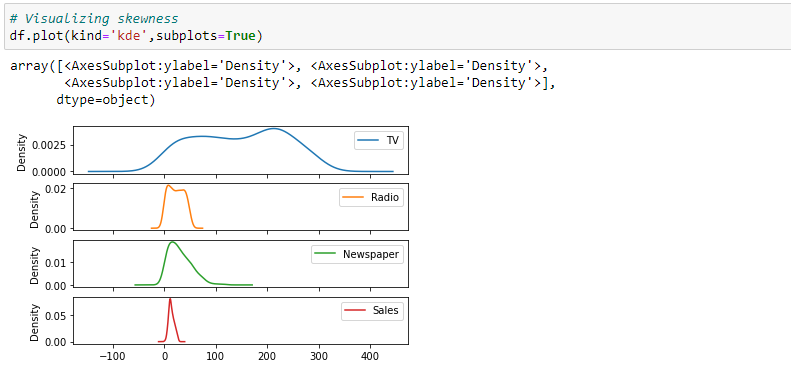


We can see a few outliers present in column Newspaper of the dataset.

**Checking Skewness**

**Skewness** is a measure of the asymmetry of the probability distribution of data variables about its mean.



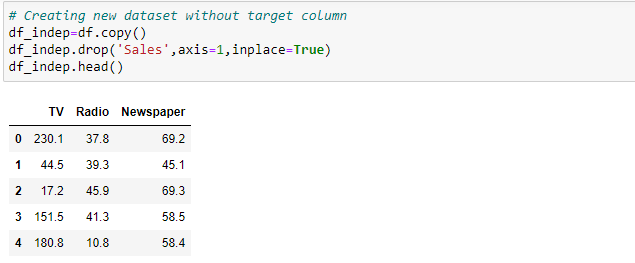


Skewness is present in the dataset in column Newspaper.

**Data Cleaning :**

***Outliers removal and Skewness removal are not performed on target column as it will change the actual target values which will impact the prediction.***

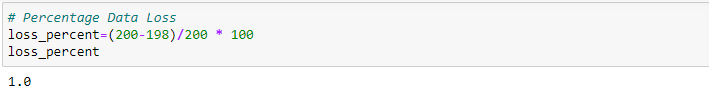
**Dropping the target column**

****

**Removing the outliers**

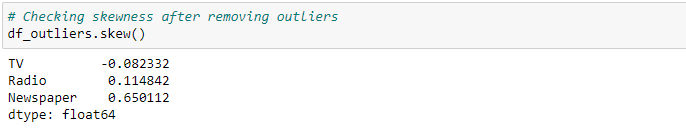


**Checking for the percentage of data lost after removing outliers**

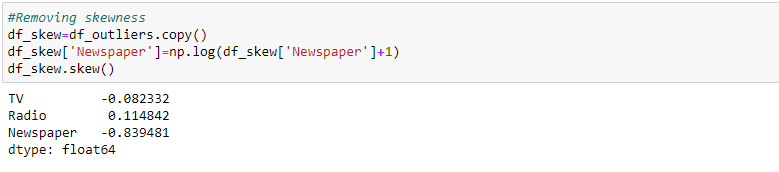


Just 1% data is lost after removing skewness.

**Checking and removing the skewness using Log transformation**

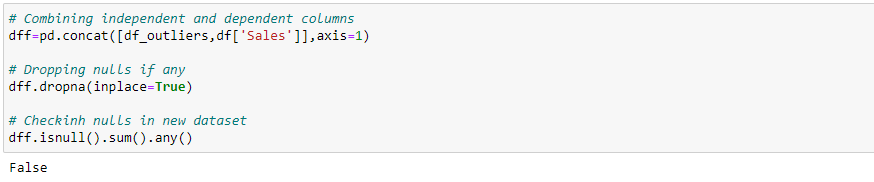


Skewness of Newspaper column has been reduced after removing outliers but needs to be reduced more.



Skewness of Newspaper is increasing on applying log transformation hence based on the kind of distribution; we will consider the current values as skewness threshold.

**Combining independent and target variables for further process**

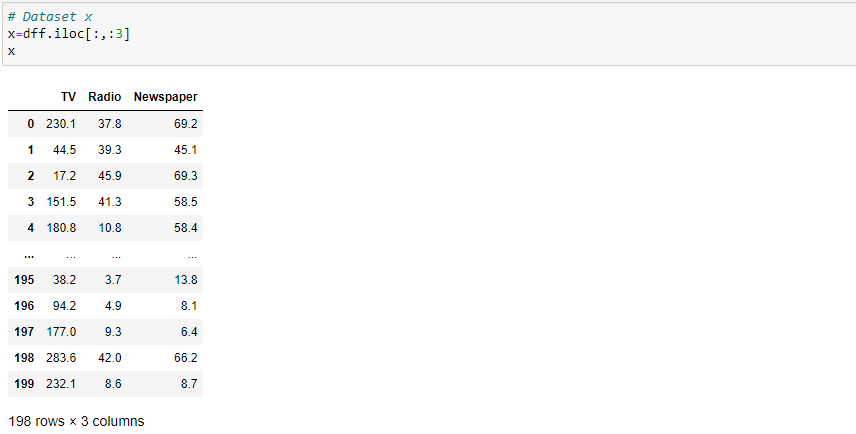


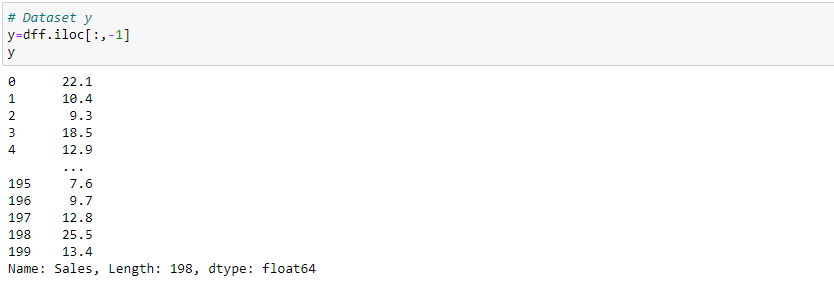
Combined all the columns and removed nulls from the final dataset.

**Building Machine Learning Models :**

Building of machine learning model is based on continuous learning and generalizing from training data, then applying that acquired knowledge to test a new data that it has never seen before to make predictions and fulfill its purpose.

**Separating the columns into x and y**

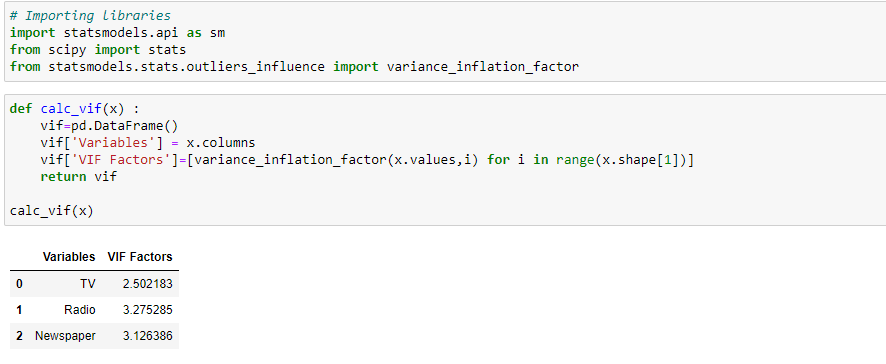




**Multicollinearity with VIF**

**Multicollinearity** is the state wheretwo or more independent variables are highly correlated with each other and therefore the possibility is they are giving a similar kind of information about the variance within the given dataset. Due to this, output is not able to conclude which particular input is contributing high.

**VIF (Variance inflation factor)** is a technique through which we can find out the variables/columns which could be having high multicollinearity and later we can drop those columns to remove multicollinearty.



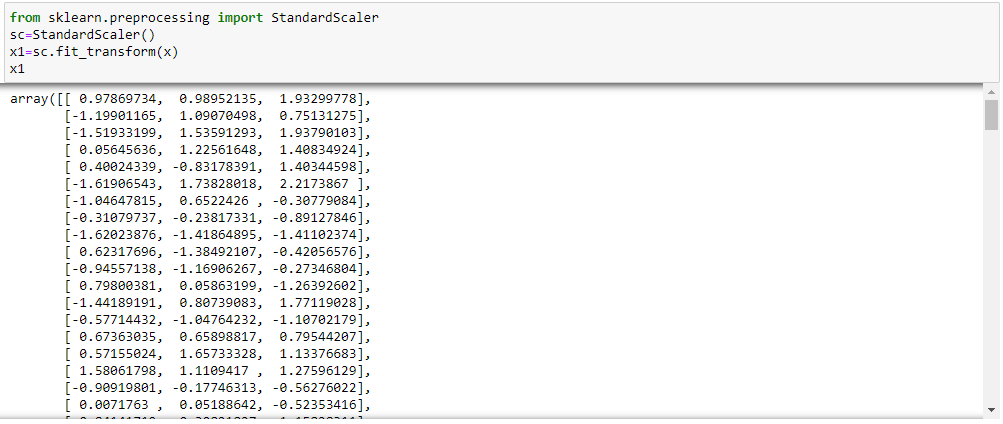
There is no multicollinearity present in the dataset.

**Scaling : Standard Scaler / Min-Max Scaler**

**Feature Scaling** is a technique to standardize the independent features present in the data in a fixed range. If not done, machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values and it will provide the biased results/prediction answers.

The goal is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

**Standard Scaler :**Technique which re-scales the feature values to make the distribution with 0 mean value and variance equals to 1.



Independent variables are scaled / standardized.

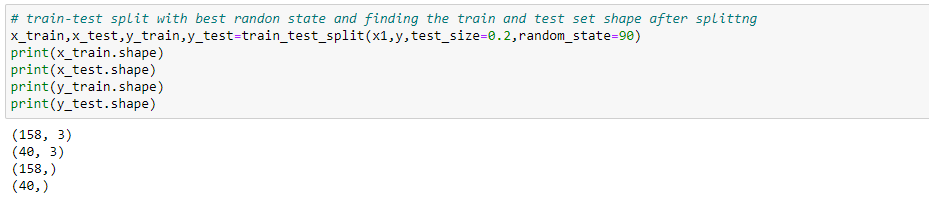
**Find best random state for Train-Test Split**

The **train-test split** is a model selection technique that evaluates the performance of a machine learning algorithm on some particular random state. The procedure involves taking a dataset and dividing it into two subsets train and test sets : Training set is used to train or create a model and testing set is used to test the predictions.



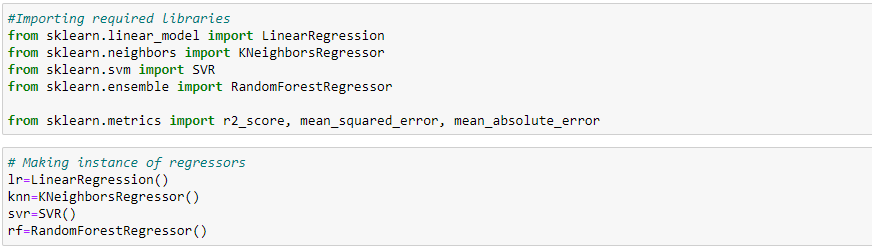
Here, random state=90 is giving a good accuracy. So, let’s take 90 as our best random state.

**Checking the shape of train and test sets**

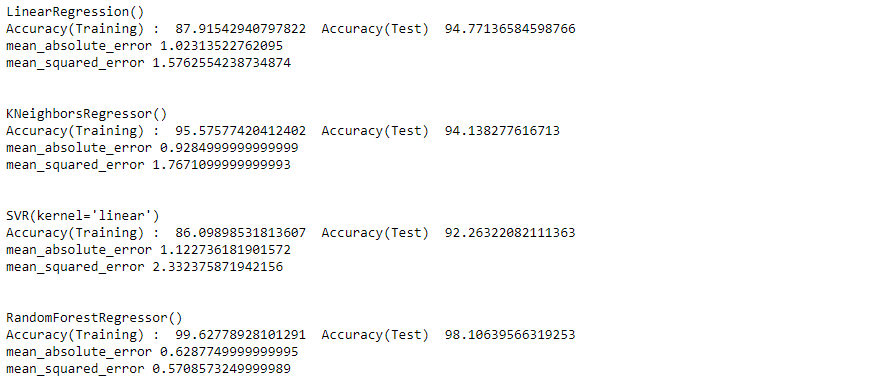


**Finding the best algorithm**

In machine learning/model training, we should use multiple models for training the dataset to find the best one based on evaluation matrix and cross validation score. Here we are using 4 models :   
1. Linear Regression   
2. KneighborsRegressor  
3. SVR (Support Vector Regressor)  
4. RandomForestRegressor.



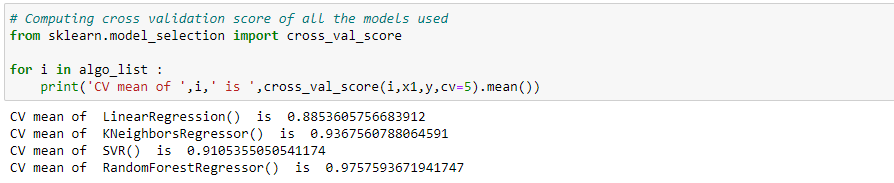


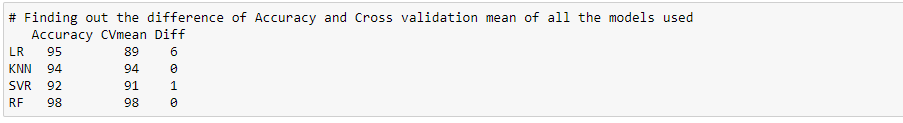


**RandomForestRegressor()** has the highest accuracy and least error hence opting for this model.

**Cross Validation**

**Cross-validation** is a resampling technique used to protect against overfitting in a predictive model. Here, we take a parameter called k that refers to a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. This procedure is often called **k-fold cross-validation**.



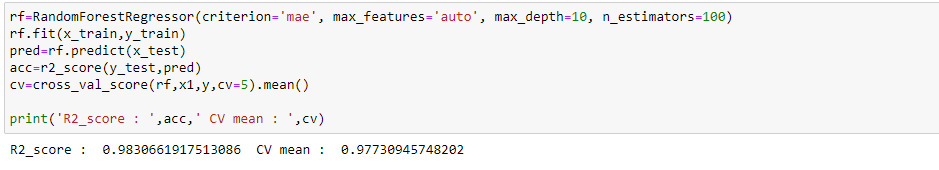


RandomForestRegressor() has the highest accuracy and least difference in accuracy and Cross validation mean. So, let’s select RandomForestRegressor() as the best algorithm.

**Hyperparameter Tuning**

**Hyperparameter** tuning is the process of choosing a set of optimal hyperparameters for a learning algorithm to improve the accuracy and to avoid overfitting or underfitting issues. The ultimate goal is to find an optimal combination of hyperparameters that minimizes the predefined loss function to give better results/accuracy for selected model.





Executed the selected model with hyper tuned parameters for better accuracy.

***We are getting model accuracy as 98% and cross validation score as approx 98% which shows our model is performing good.***

**AUC-ROC Curve**

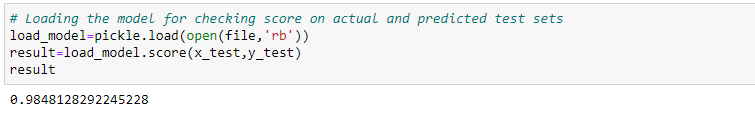
For selected algorithm ROC AUC is not available, since the ROC curve by definition works with a binary classification problem by plotting true positive rate against false positive rate by varying the classification threshold. ROC AUC is only calculable in the case if the algorithm returns a continuous probability value (and only 1 value) for an unseen element.

**Save the model into a pickle file**

**Serialization** is done to make the model platform independent (converted to machine language, byte stream).



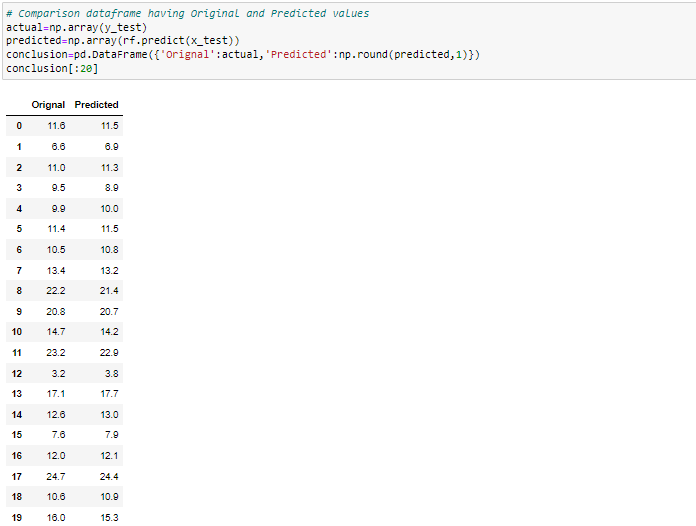
**Loading the model to check the prediction score on test sets**



We are getting very good accuracy on test sets.



**Printing the actual and predicted values in a DataFrame for comparison.**



**Concluding Remarks :**

To summarize, we have performed Randomforest Regression having bestaccuracy and without over-fitting and under-fitting issues on the provided ‘Advertising Sales Channel Prediction’ dataset. Along with which we have covered some basic introductory statistics and a complete procedure to be followed to build a machine learning model. Below mentioned are the various phases that we have gone through, throughout the project -

1. Data Preprocessing
2. Exploratory data analysis and Visualizations
3. Outlier detection and skewness treatment
4. Multicollinearity with VIF
5. Scaling the data — Standard scaler
6. Fitting the machine learning models
7. Cross-validation of the selected model
8. Model hyperparameter tuning
9. Saving the final model and prediction using saved model

This is an example of how to perform and interpret any dataset/ problem using machine learning with all necessary preprocessing and post processing steps.