**Predicting Advertising vs Sales using Python**

**By: PREETI SINGH**

Problem Statement

Sales Channel Prediction Case Study

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.

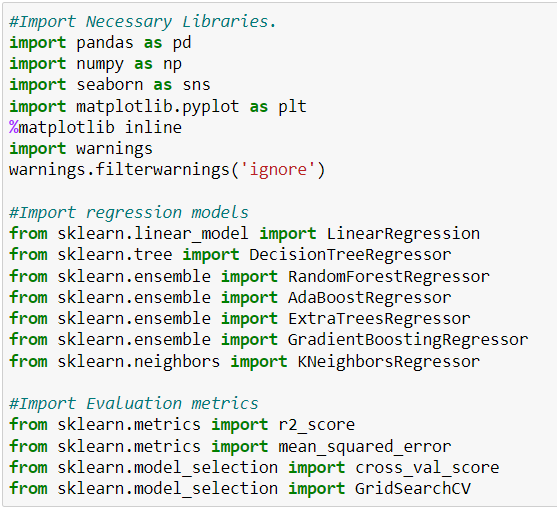
The distribution strategy and the channel design have to be right the first time. The case study of Sales channel includes the detailed study of TV, radio and newspaper channel. The predict the total sales generated from all the sales channel.

Problem Solving Description

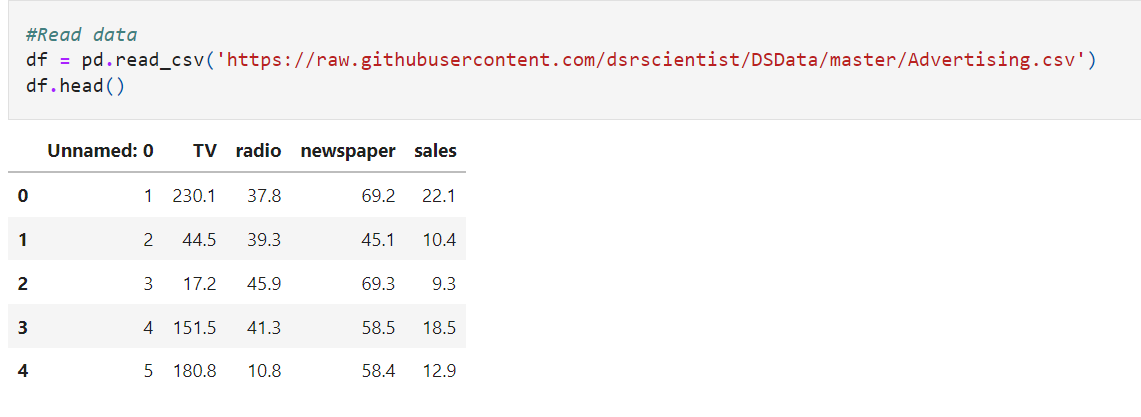
As per the problem statement I'll have to predict the sales hence 'Sales' is my target variable and rest all the features are my independent variable. In sales column I am observing that the numbers are in continuous so it is a regression problem. I will be building some regression models and will compare them according to their metrics, then I am going to choose the best model using evaluation metrics to predict Sales count. For complete codes and execution, you can visit the following link:

https://github.com/preeti0292/Practice-Projects/blob/main/Advertising%20Sales%20Channel%20Prediction.ipynb

# Import Necessary Libraries



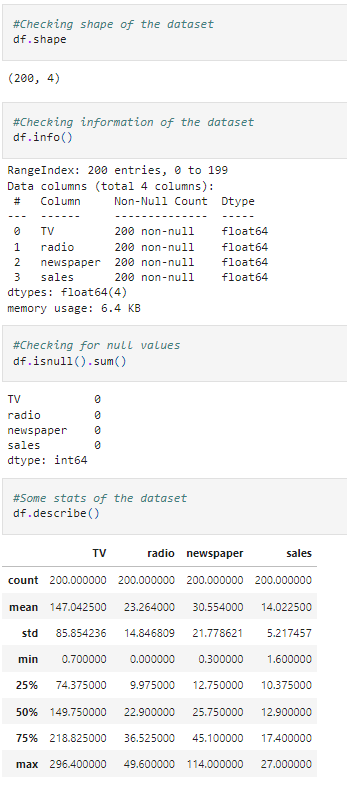
# Reading Dataset

****

# Feature Description

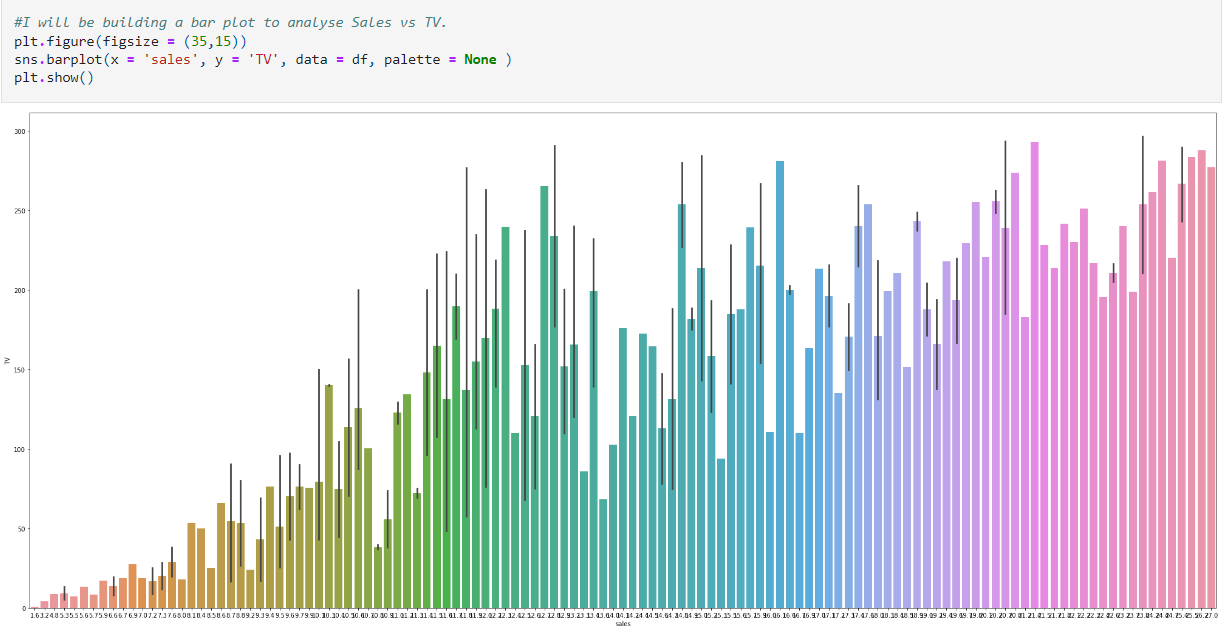
* Unnamed:0 - does not contribute in sales will remove it.
* TV - It refers to the sales count generated from TV advertisement.
* Radio - Here radio refers to the sale count generated from radio advertisement.
* Newspaper - Newspaper column shows the sales count generated from newspaper sales.
* Sales - Sales column refers to the total sales of TV, radio and newspaper.

# Stats of the Dataset



In this dataset we have 200 rows and 4 columns, no object data only float. No Null values are present in the dataset.

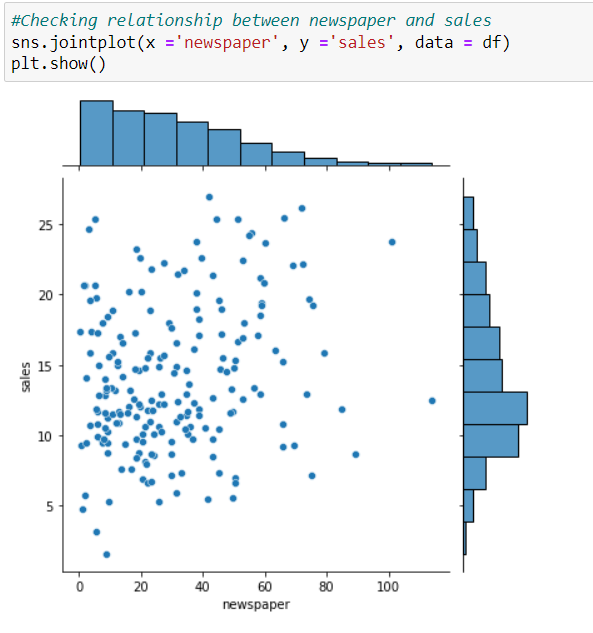
# Data Visualization and Pre-processing

****

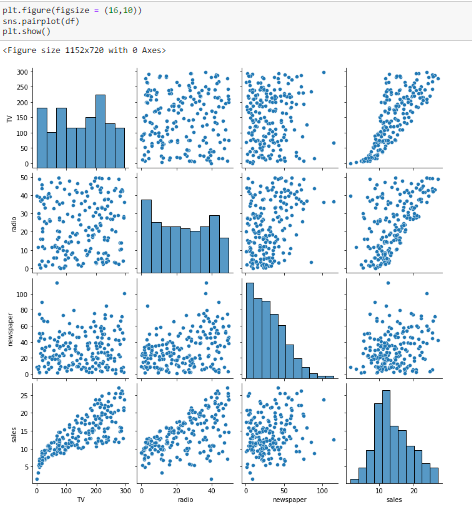
From the above bar plot, I can see strong relationship between TV and Sales, Sales are surely increasing because of TV advertisement.

# 

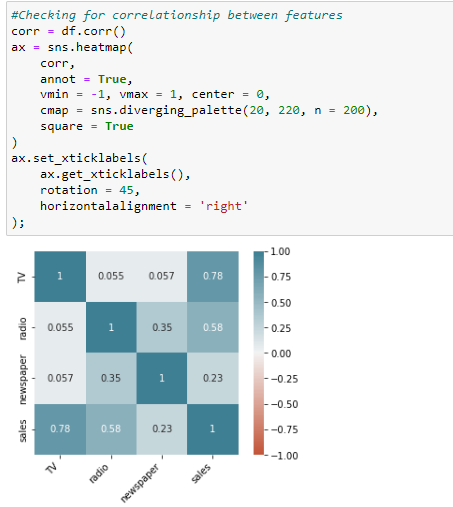
Seeing good relationship between radio and sales though this relationship is not as strong as with TV but still I can observe some trend. I can say radio advertisements are surely helping in increasing sales



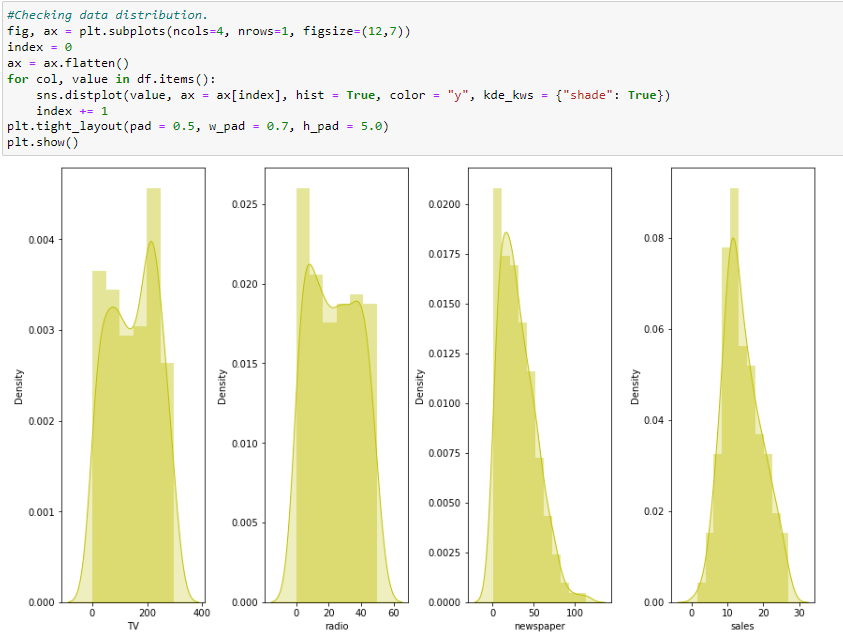
In the above scatter plot, I can't see any trend. Newspaper advertisements are not really helping in sales.



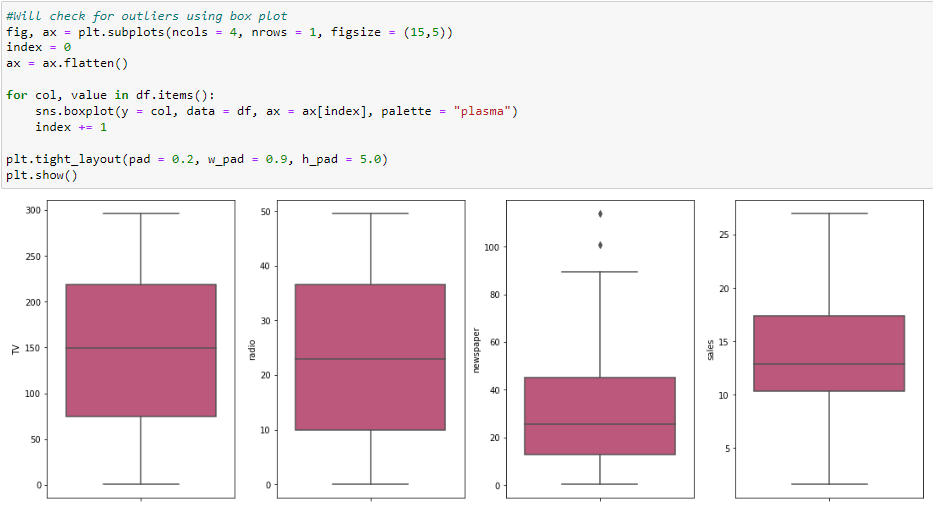
From the above pair plot, I can say that, apart from newspaper rest two features have good relationship with sales.

****

* TV has strong relationship with Sales.
* Radio has good relationship with Sales and newspaper.
* Newspaper and sales relationship is okay not as strong as TV and Radio.



Distribution looks good for all the features. However, seeing some skewness in newspaper. Data looks pretty clean next will check for outliers.

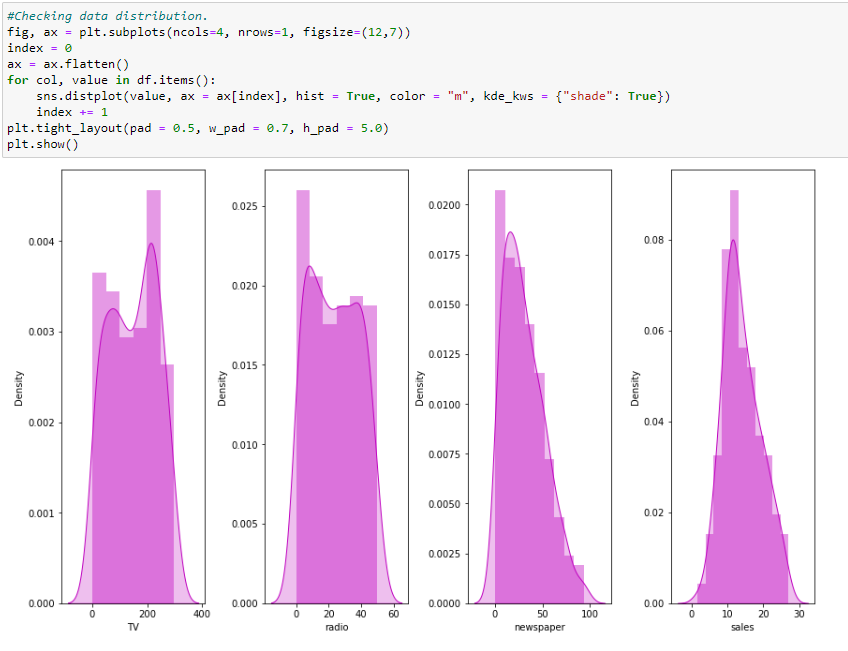


All the features look good excluding newspaper. In newspaper I am seeing some outliers are present on the higher side, I will be using Inter Quantile Range method to identify those outliers and then will remove them.

# Removing Outliers using IQR Method



Outliers have been removed successfully next I am going to check distribution again.

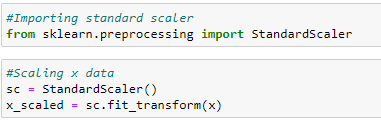


Distribution looks good. Next, I am going to split data into X and Y variable.

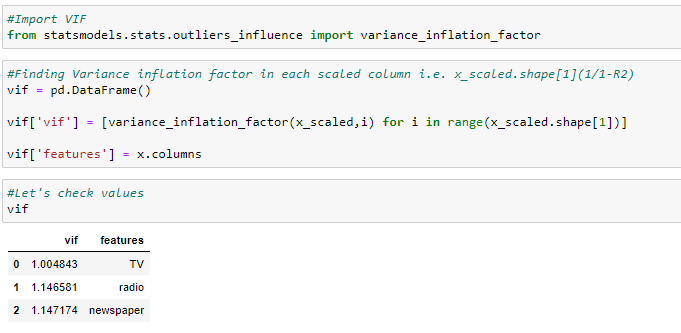
# Splitting Dataset into X and Y variable

# 

# Scaling X Data:

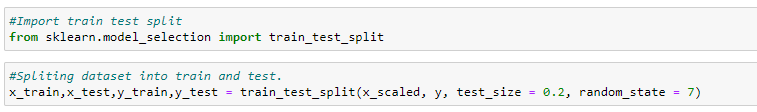


# Checking Multicollinearity using VIF:

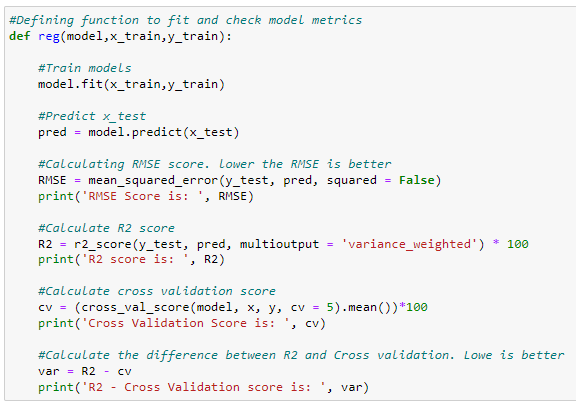


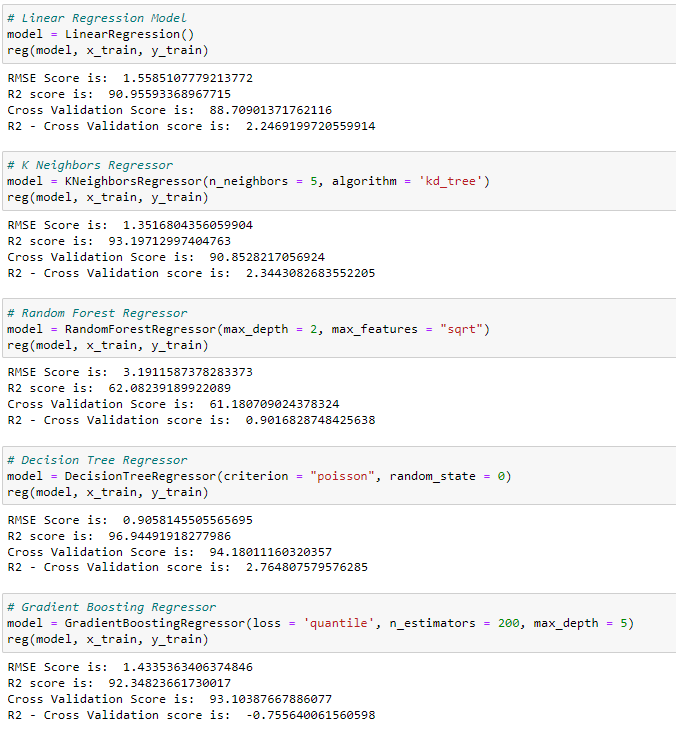
No feature has VIF value more than 5 so there is no multicollinearity issue.

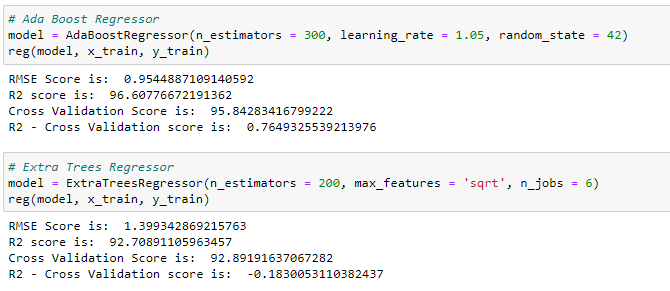
# Splitting Data into Train and Test:



# Model Building:

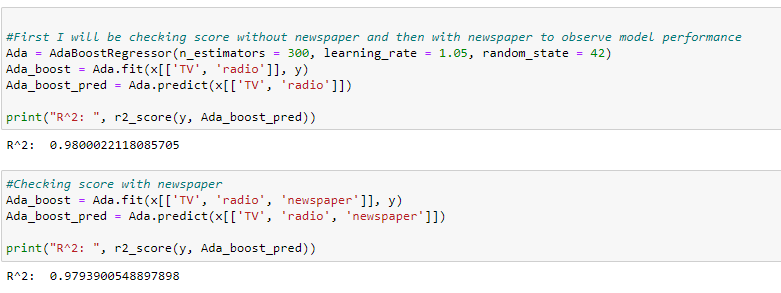




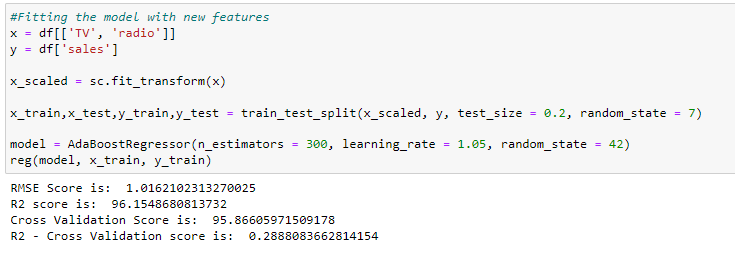


After observing metrics for all the models, I can conclude that Ada Boost Regressor is giving the best result hence I will be using Ada Boost Regressor to predict sales.

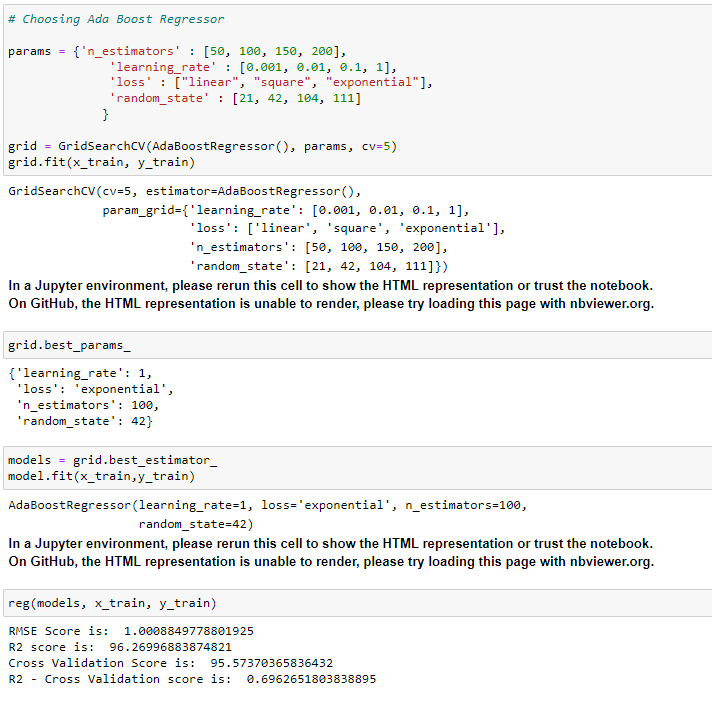
# Feature Selection for the Best Model:



After adding newspaper, I can see model performance have decreased hence I will be removing 'newspaper' from the dataset.

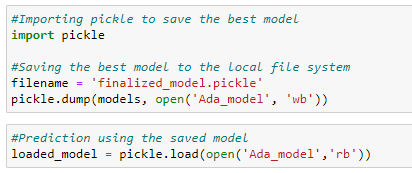


# Hyper Parameter Tuning:

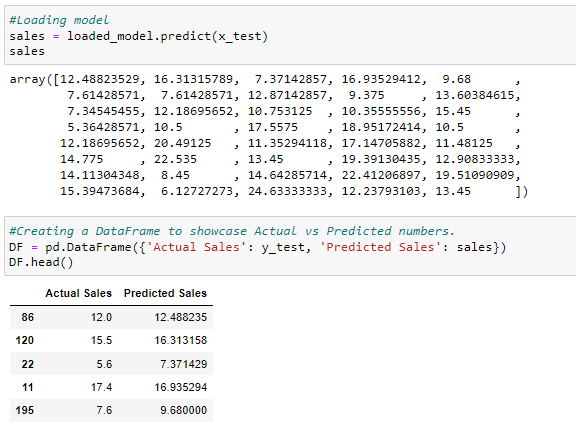


I have tuned the final model but score has come down by some points no issues still this is the best performing model so far hence I will be using it for further prediction.

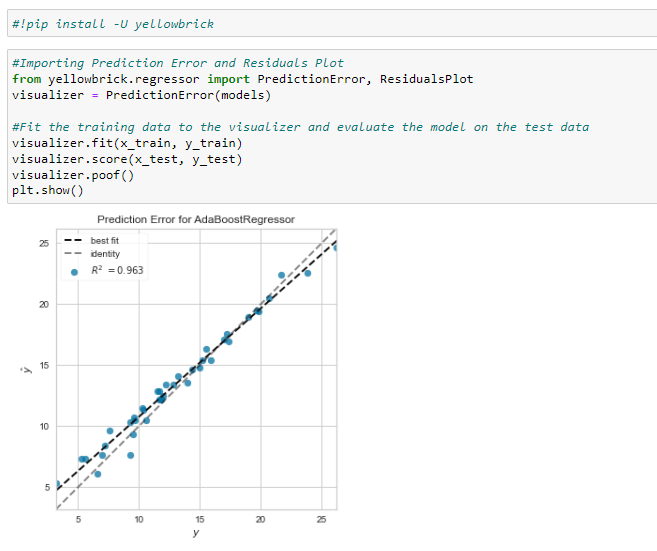
# Saving The Best Model:

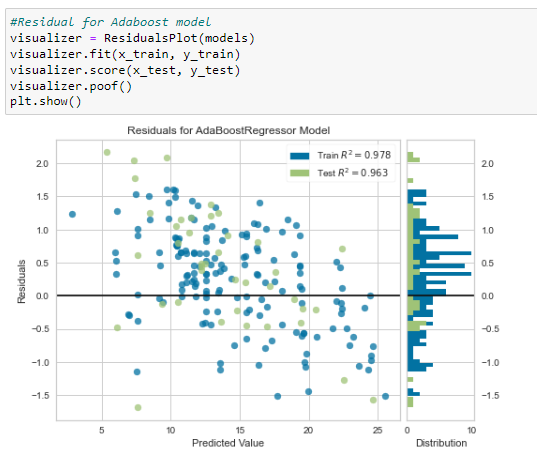


# Making Prediction Using Saved Model:



# Visual Error Residual:





# Conclusion:

To predict which advertising platform is giving the best sales performance I have first followed some EDA steps and visualizations. Checked correlation between features and removed outliers using Inter Quantile Range Method. We should always try to reduce skewness from our dataset as much as possible or else our prediction can go to wrong direction. Next, I divided the dataset into X and Y variable which is X = features which will help in prediction and Y = target column which is are target variable. Next, I have scaled the x dataset using Standard Scaler. Standard scaler helps to scale the values so that it’s easy for our model to understand. I have checked for Multicollinearity problem using Variance Inflation Factor which helps us to understand if any feature is correlated to other feature, we should always solve multicollinearity issue before building our model or else our model can overfit. However, in this dataset we do not have multicollinearity problem hence, we can go ahead and start building our model. I have built 7 models and AdaBoost Regressor has given the best performance where RMSE Score is: 0.9544887109140592, R2 Score is 96% and Cross Validation score is 95%. Then I did feature selection for my best model and removed newspaper. Hypertuned the parameters using GridSearchCV however my R2 score for AdaBoost model was still 96% hence saved the original model using pickle and later I have loaded the saved model and made prediction. My model did a good job in terms of prediction as the training R2 score was 97% and test R2 score was 96% which indicates no overfitting and I have built a good model.