Name: Vanshita Desai (23060641055) &

Shruti Rane (23060641076)

**LINEAR MODEL**

**MULTIPLE LINEAR REGRESSION**

**TOPIC: MARKETING OUTCOME PREDICTION**

**Introduction:**

In today's competitive business landscape, companies invest significant resources in marketing campaigns to attract and retain customers. However, determining the effectiveness of these marketing strategies and optimizing their allocation of resources is a complex task. Traditional approaches often rely on intuition or historical data without leveraging the full potential of advanced analytics.

**Problem Statement:**

A Company XYZ aims to promote their flagship product, X, through various advertising channels, as a part of their marketing strategy. However, determining the effectiveness of each marketing channel on product sales is essential for optimizing their advertising budget and maximizing return on investment (ROI).

**Objective:**

The objective of this project is to develop a predictive model that can accurately estimate the impact of various marketing strategies on key performance of metrics such as sales. By leveraging multiple linear regression analysis, we aim to identify the most influential factors and quantify their effects on the outcome variable.

**About Dataset:**

The data used consist of the following variables, which were used as platforms for running advertisement.

* TV promotion budget (in million)
* Social Media promotion budget (in million)
* Radio promotion budget (in million)
* Influencer: Whether the promotion collaborate with Mega, Macro, Nano, Micro influencer
* Sales (in million)

Here, Independent variables (**X**) = TV, Social Media, Radio, Influencer

Dependent variable (**Y**) = Sales

**Link to the dataset:**

<https://www.kaggle.com/datasets/harrimansaragih/dummy-advertising-and-sales-data>

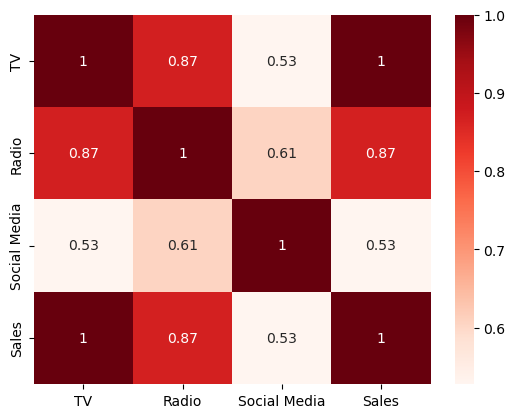
**Link to Code file:**

<https://colab.research.google.com/drive/1qKiJk68NuO-iDZD9-rhBFdE83WmsbdsQ?authuser=7>

**Analysis and Inference:**

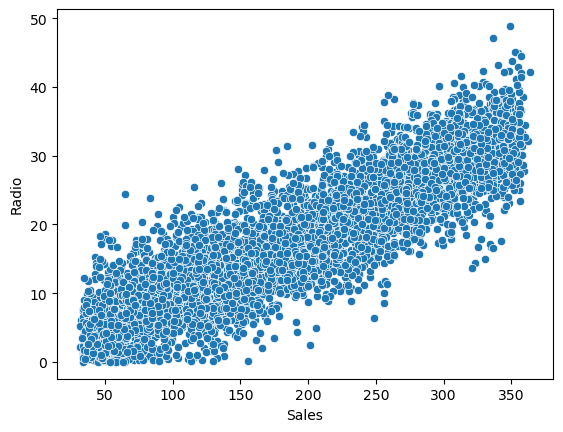
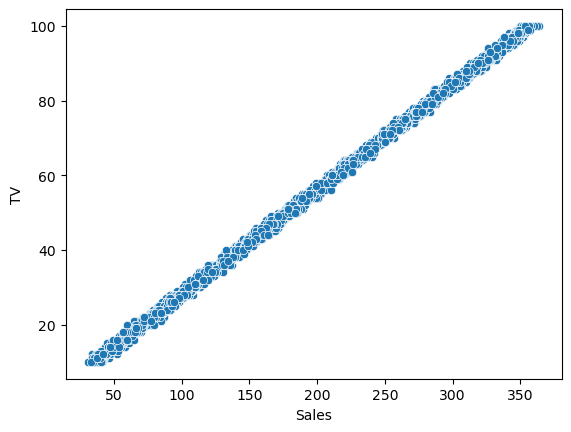
1. **Data Preprocessing**: Cleaned the dataset, handled missing values, encoded categorical variables and scaled numerical features as for Influencer column Nano=0, Micro=1, Mega=2, Macro=3.
2. **Exploratory Data Analysis**: Analyzed the relationships between predictor variables and the outcome variable through visualizations and statistical summaries by

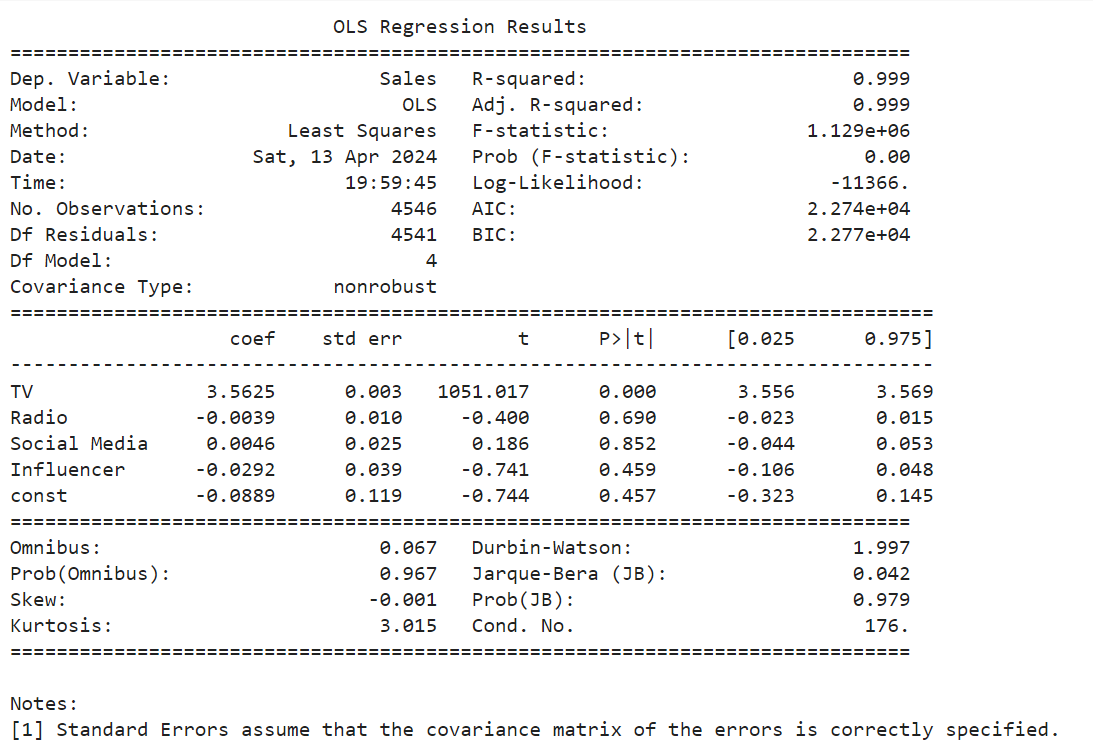
* Finding correlation between the independent variables.



From the plot we can infer that the correlation between TV and Sales (0.9994) as well as the correlation between Radio and Sales (0.868638) are strong.

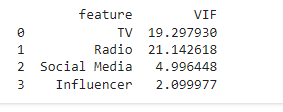
The scatterplot for TV and Radio with Sales are as follows:



* **Model fitting**:
* **Statistical Summary:**

From the summary above we can see that value (0.999) which explains that 99% of the variation is explained through the model’s input and value (0.999) are high indicating that the model is a good fit.

* **Checking multicollinearity**



**Variance Inflation factor (VIF)** is a measure used to assess multicollinearity in a linear regression model. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other. Mathematically, it can be expressed

Typically, a VIF above 5 or 10 is considered problematic and indicates a significant level of multicollinearity.

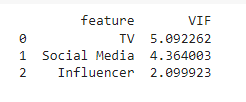
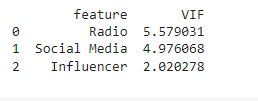
From our data we can see that high **multicollinearity is** **present**.

1. **Feature Selection:**

Feature selection techniques aim to identify the subset of predictor variables that are most relevant for predicting the target variable while reducing redundancy and overfitting.

Since there is presence of high multicollinearity we try and improve the regression model by removing either TV or Radio and then checking the value for VIF

**By removing the variable i.e Radio By removing the variable i.e TV**

   
Since both the models have **VIF <10**, we can use both of them further in our analysis.

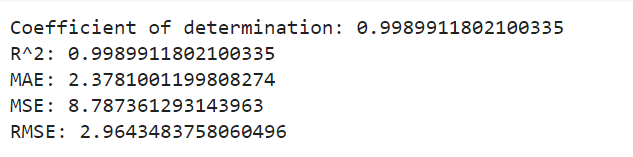
4. **Model Evaluation:**

After visualizing the data and confirming the correlation among them we proceed with splitting the data into Train data and Test data to build the regression model for prediction. For this we assume a value of 0.8, i.e 80% of our data will be divided into train data and 20% of the data will be divided into test data for both Model. Out of 4546 observation, 3636 observation are in train data and 910 are in test data.

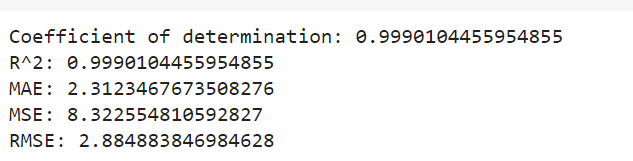
**Original model:**

**Here, x1= ‘TV’, x2=’Radio’ ,x3=’Social Media’, x4=’Influencer’**

Summary for train data

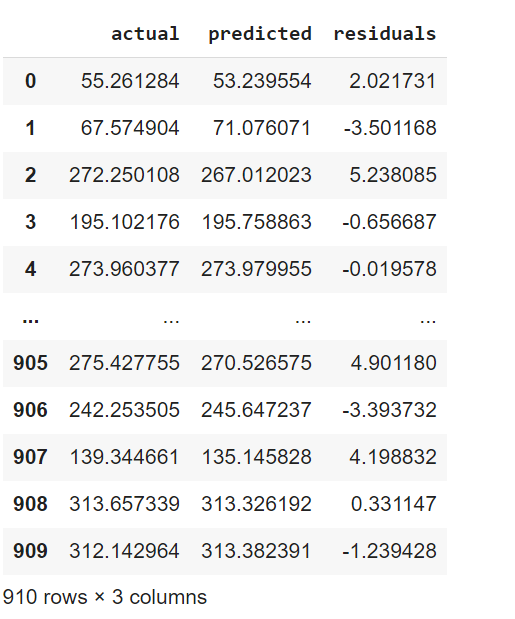


Summary for test data,

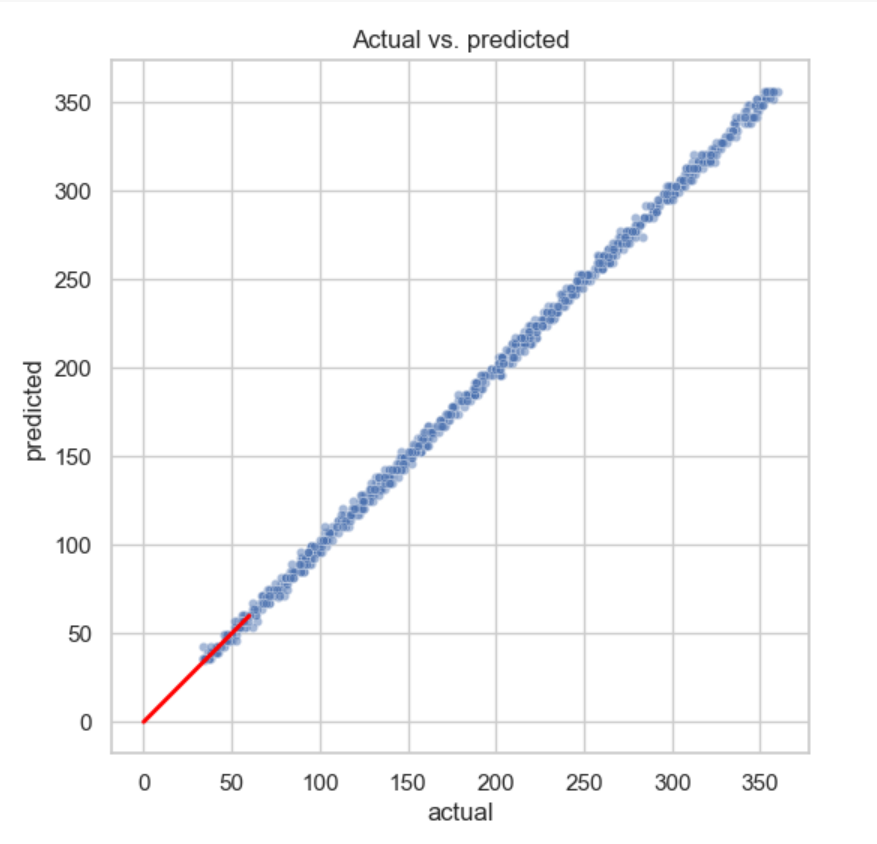


Here,  **= 0.99** and **Mean Square Error value is 8.32255** indicates the average squared difference between the observed and predicted values by a model, with lower values indicating better model performance and closer alignment between predictions and actual observations.

**Actual, predicted and residual values for the model:**



**Plot of Actual vs Predicted values,**



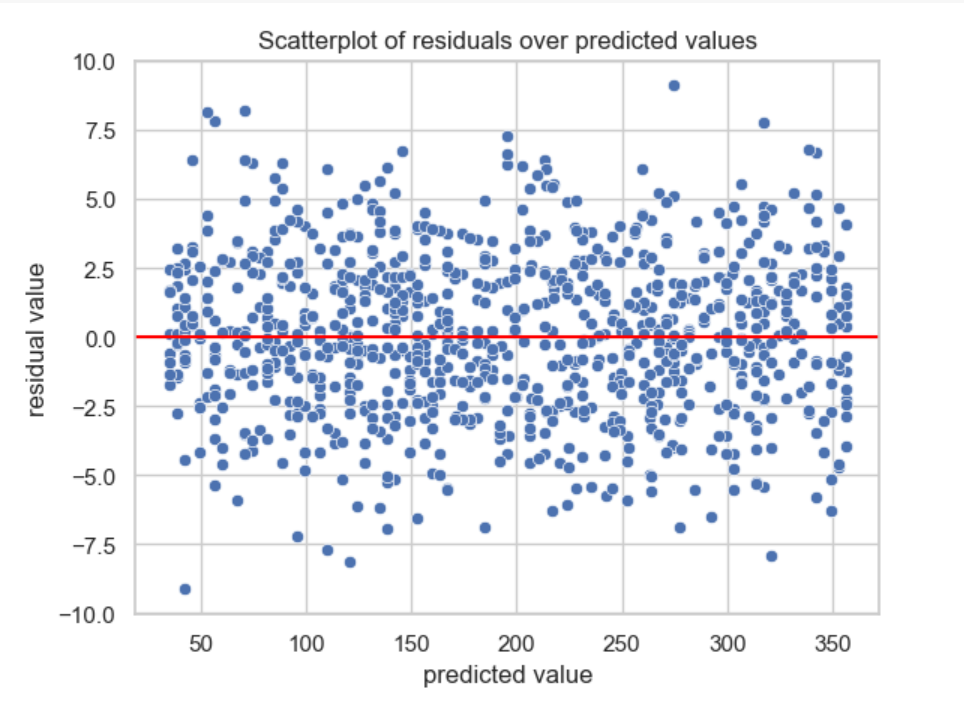
**Since the actual versus predicted values forms a perfect linear relationship, it indicates that the regression model's predictions are accurate and unbiased.**

The slope of the linear relationship is equal to 1, indicating that the model's predictions are on par with the actual values without any systematic errors.

The linear relationship between actual and predicted values provides confidence in using the regression model for inference and prediction tasks.

The model can be considered highly reliable and suitable for making predictions on new data within the same domain.

**Plot for Predicted value vs Residual values**



Here we can observe that the residuals are randomly scattered around zero, with no discernible pattern or trend, it suggests that the model is capturing all the available information in the data.

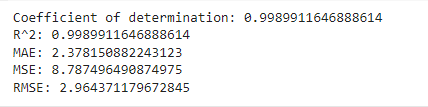
Random scatter indicates that the model's predictions are unbiased, and the residuals have no systematic relationship with the predicted values.

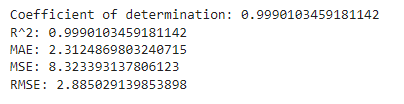
Since the residual are randomly scattered the model is appropriate for the data and that the assumptions of the regression analysis are likely met.

**For model 1:**

**Here, x1= ‘TV’,,x3=’Social Media’, x4=’Influencer’**

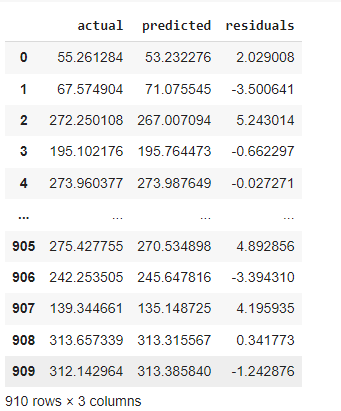
Summary for train data,



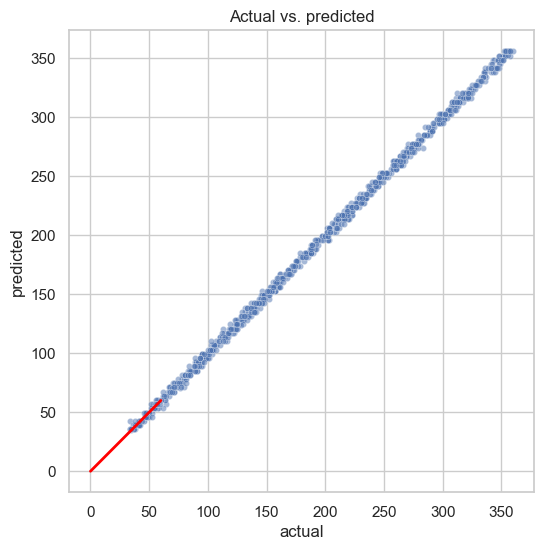
 Summary for test data,

Here,  **= 0.99** and **Mean Square Error value is 8.32339** indicates the average squared difference between the observed and predicted values by a model, with lower values indicating better model performance and closer alignment between predictions and actual observations.

**Actual, predicted and residual values for the model:**



**Plot of Actual vs Predicted values**,



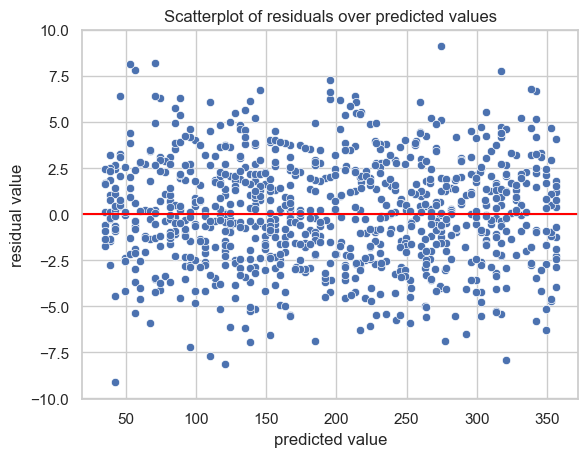
**Since the actual versus predicted values forms a perfect linear relationship, it indicates that the regression model's predictions are accurate and unbiased.**

The slope of the linear relationship is equal to 1, indicating that the model's predictions are on par with the actual values without any systematic errors.

The linear relationship between actual and predicted values provides confidence in using the regression model for inference and prediction tasks.

The model can be considered highly reliable and suitable for making predictions on new data within the same domain.

**Plot for Predicted value vs Residual values**



Here we can observe that the residuals are randomly scattered around zero, with no discernible pattern or trend, it suggests that the model is capturing all the available information in the data.

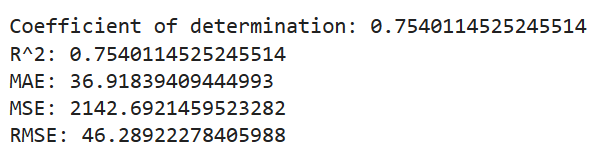
Random scatter indicates that the model's predictions are unbiased, and the residuals have no systematic relationship with the predicted values.

Since the residual are randomly scattered the model is appropriate for the data and that the assumptions of the regression analysis are likely met.

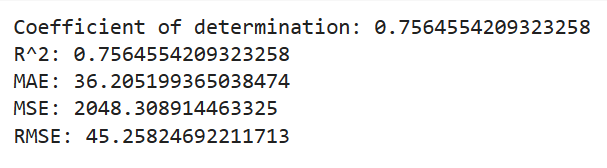
**For Model 2:**

**Here, x2=’Radio’ ,x3=’Social Media’, x4=’Influencer’**

Summary for train data,

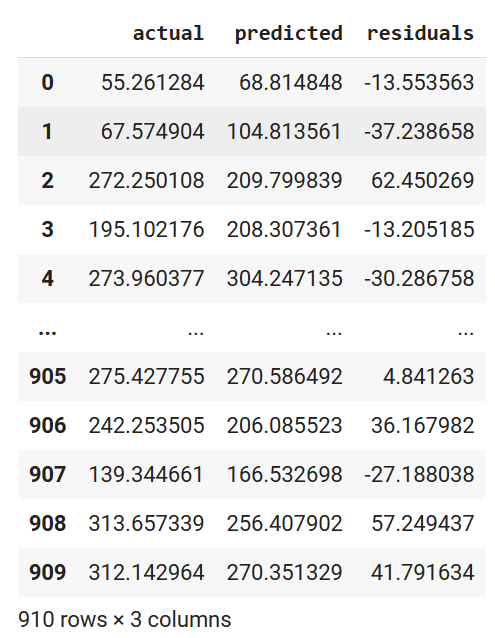


Summary for test data,

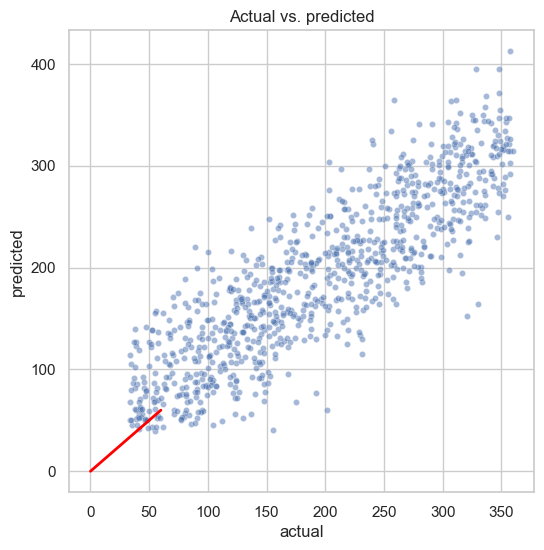


Here the  **value is 0.756455** and **Mean Square Error value is 2048.30** indicates the average squared difference between the observed and predicted values by the model, with higher values suggesting poorer model performance and greater deviation between predictions and actual observations.

**Actual, predicted and residual values for the model:**



**Plot of Actual vs Predicted values,**

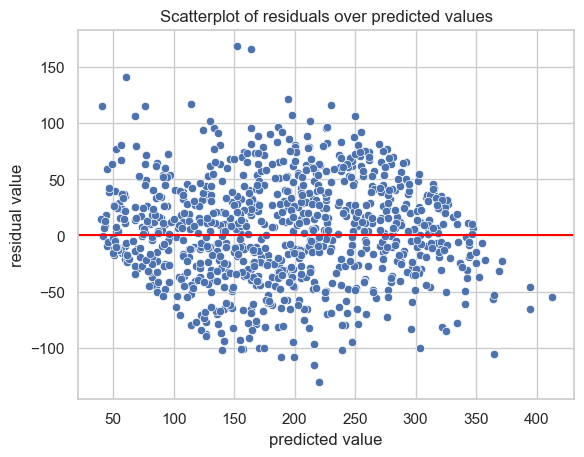


**The plot shows that the relationship between the actual and predicted values is not linear. This indicates that the model's predictions deviate from the actual values in a non-linear manner.**

We can also conclude that due to lack of linearity the model is mis-specified.

It might not include the appropriate variables or may not capture the true relationship between the variables adequately.

**Plot for Predicted value vs Residual values**



We can observe a **bow shape pattern** indicating that the spread of residuals changes systematically as the predicted values increase or decrease.

The bow shape suggests heteroscedasticity, which means that the variance of the residuals is not constant across the range of predicted values.

Model's errors exhibit different levels of variability at different levels of the predicted values.

**Conclusion**:

From the output and analysis of both the models, we can conclude that the Model 1 which includes TV, Social Media and Influencer performs well against the Model 2 which includes Radio, Social Media and Influencer.

Model 1 can be considered for decision making.

Also ,according to the requirement of the company, Model 2 can be used with further more improvements and analysis, or the original Model can also be used with all the variables i.e TV, Radio, Social Media and Influencer. The results of the original Model and Model 1 are quite similar.

Therefore, depending upon the requirement, the Model can selected.