Summer Internship
On
Statistical Analysis of Impact of Social Media Advertising on Sales

Submitted to the Amity University Uttar Pradesh In partial fulfillment of requirements for the award of the Degree of

BACHELOR OF STATISTICS

By

Arpit Saxena
Enrollment No: A4479119002

Under the Supervision of:

Supervisor

Dr. Dheeraj Pawar

Department of Statistics
Amity Institute of Applied Science

Amity Institute of Applied Sciences, Amity University Uttar Pradesh, Sector 125, Noida – 201303 (India)

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF APPLIED SCIENCES

Synopsis of Summer Internship:

Title: Statistical Analysis of Impact of Social Media Advertising on Sales

Name of Guide: Dr. Dheeraj Pawar

Programme:- B.Stats.		**Year/Semester:- 5th semester**	
S.No.	**Enrollment No.**	**Name**	**Signature**
1	**A4479119002**	**Arpit Saxena**	

Summary:- Construction and analyzation of the multiple linear regression model that can predict sales of a particular product in 200 different markets on the basis of its advertising budgets for youtube, facebook and newspaper.

Schedule of work completion:- 19 May 2021 - 29 June 2021

Signature of Student

Signature of Guide

Signature of Programme Leader

Approval by Board of Faculty

Member	**Signature**	**Remark (Approved / Not Approved)**

DECLARATION

I, Arpit Saxena, student of Bachelor Of Statistics hereby declare that the Summer Internship project titled “Statistical Analysis of Impact of Social Media Advertising on Sales” which is submitted by me to Department of Statistics, Amity Institute of Applied Sciences, Amity University, Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of Bachelor Of Statistics, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Noida

Date: 29 June 2021

Arpit Saxena

CERTIFICATE

On the basis of declaration submitted by Arpit Saxena ,student of Bachelor Of Statistics,I hereby certify that the Summer Internship project titled "Statistical Analysis of Impact of Social Media Advertising on Sales" which is submitted to Department of Statistics,Amity Institute of Applied Sciences, Amity University, Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of Bachelor Of Statistics, is an original contribution with existing knowledge and faithful record of work carried out by him under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Noida

Date: 29 June 2021

Dr. Dheeraj Pawar

Department of Statistics
Amity Institute of Applied Sciences
Amity University, Uttar Pradesh, Noida

ACKNOWLEDGEMENT

I, Arpit Saxena, would like to express my deep sense of gratitude towards my faculty guide Dr. Dheeraj Pawar, for guiding me from the inception till the completion of the project. The experience of working under my faculty guide has been a value addition to the learning during my course of bachelors of statistics. I would like to express my heartfelt gratitude for Dr. Dheeraj Pawar with whose guidance and valuable suggestions I was able to maximize on the learning curve during the completion of my project. His timely responses to any issues that came along and his promptness helped me to successfully complete the project.

I am thankful to him that he provided me an opportunity to work under his guidance and sharing his valuable experience

Also, I am thankful to Amity Institute of Applied Science (AIAS) and Department of Statistics for providing me with such an opportunity to gain analytical knowledge and skills, imparted as a part of curriculum.

Arpit Saxena

CONTENTS

Abstract

This project report includes the construction of a multiple linear regression (MLR) model that can predict sales of a particular product in 200 different markets on the basis of its advertising budgets on 3 different social media channels. The different social media channels used for the advertisements are youtube, facebook and newspaper. Also, it includes the effects of these advertising budgets of individual media channel on the total sales of the product. There are few assumptions which should be set up before constructing a MLR model. Null hypothesis and alternate hypothesis should be set up before proceeding to the construction part of the model. In this project, construction of model is done on MS- Excel using data analysis tool Pak. There are 3 parts of the regression results which will indicate that the sales of the product and the social media advertising budgets holds a regression relationship. Further, it includes analysis on all assumptions set earlier for checking the model adequacy. All the assumptions will hold true for the model which brings the conclusion that the MLR model is adequate for predicting sales of that product using the social media advertising budgets.

Introduction

In today's world, social media becomes an important platform for marketers to advertise their products and generate sales. It plays a vital role in the life of marketers in building a hub where they can socially connect to the people who are in need of the required products with the help of advertisements to increase sales. There are several social media platforms such as Facebook, Youtube, Instagram, Twitter, etc which are much efficient for marketers and are considered as an important hub for marketing. Different social media have differet effects on the sales of the product.

In this project, we will study the effect of 3 different advertising social media channels on sales of a particular product. So, at first, a dataset was collected which gave details on the advertising social media budgets for a particular product for 3 different sites and the sales of that product in 200 different markets. The budgets are in thousand of dollars and the sales are in thousand of units. The different social media sites used for the advertisements are youtube, facebook and newspaper.

The dataset is collected from secondary resources. The dataset for the study is as follows:

	Advertising Budgets			
S.No.	**Youtube**	**Facebook**	**Newspaper**	**Sales**
1	276.12	45.36	83.04	26.52
2	53.4	47.16	54.12	12.48
3	20.64	55.08	83.16	11.16
4	181.8	49.56	70.20	22.20
5	216.96	12.96	70.08	15.48
6	10.44	58.68	90.00	8.64
7	69.00	39.96	28.20	14.16
8	144.24	23.52	13.92	15.84
9	10.32	2.52	1.20	5.76
10	239.76	3.12	25.44	12.72
11	79.32	6.96	29.04	10.32
12	257.64	28.8	107.00	20.88
13	28.56	42.12	79.08	11.04
14	117.00	9.12	8.64	11.64
15	244.92	39.48	55.20	22.80
16	234.48	57.24	91.00	26.88
17	81.36	43.92	136.80	15.00
18	337.68	47.52	66.96	29.28
19	83.04	24.6	21.96	13.56
20	176.76	28.68	22.92	17.52
21	262.08	33.24	64.08	21.60
22	284.88	6.12	28.20	15.00
23	15.84	19.08	59.52	6.72

24	273.96	20.28	31.44	18.60
25	74.76	15.12	21.96	11.64
26	315.48	4.20	23.40	14.40
27	171.48	35.16	15.12	18.00
28	288.12	20.04	27.48	19.08
29	298.56	32.52	27.48	22.68
30	84.72	19.20	48.96	12.60
31	351.48	33.96	51.84	25.68
32	135.48	20.88	46.32	14.28
33	116.64	1.80	2.64	11.52
34	318.72	24.00	107.28	20.88
35	114.84	1.68	8.88	11.40
36	348.84	4.92	13.92	15.36
37	320.28	52.56	6.00	30.48
38	89.64	59.28	54.84	17.64
39	51.72	32.04	42.12	12.12
40	273.60	45.24	38.40	25.80
41	243.00	26.76	90	19.92
42	212.40	40.08	46.44	20.52
43	352.32	33.24	156	24.84
44	248.28	10.08	31.68	15.48
45	30.12	30.84	51.96	10.20
46	210.12	27.00	60.60	17.88
47	107.64	11.88	42.84	12.72
48	287.88	49.80	22.20	27.84
49	272.64	18.96	59.88	17.76
50	80.28	14.04	44.16	11.64
51	239.76	3.72	41.52	13.68
52	120.48	11.52	4.32	12.84
53	259.68	50.04	47.52	27.12
54	219.12	55.44	70.44	25.44
55	315.24	34.56	19.08	24.24
56	238.68	59.28	72.00	28.44
57	8.76	33.72	49.68	6.60
58	163.44	23.04	23.5	15.84
59	252.96	59.52	45.24	28.56
60	252.84	35.40	11.16	22.08
61	64.20	2.40	25.68	9.72
62	313.56	51.24	65.64	29.04
63	287.16	18.60	32.76	18.84
64	123.24	35.52	10.08	16.80
65	157.32	51.36	34.68	21.60
66	82.80	11.16	1.08	11.16

67	37.80	29.52	2.64	11.40
68	167.16	17.40	12.24	16.08
69	284.88	33.00	13.20	22.68
70	260.16	52.68	32.64	26.76
71	238.92	36.72	46.44	21.96
72	131.76	17.16	38.04	14.88
73	32.16	39.60	23.16	10.56
74	155.28	6.84	37.56	13.20
75	256.08	29.52	96	20.40
76	20.28	52.44	107.28	10.44
77	33.00	1.92	24.84	8.28
78	144.60	34.20	42.72	17.04
79	6.48	35.88	11.28	6.36
80	139.20	9.24	27.72	13.20
81	91.68	32.04	26.76	14.16
82	287.76	4.92	44.28	14.76
83	90.36	24.36	39.00	13.56
84	82.08	53.40	42.72	16.32
85	256.20	51.60	40.56	26.04
86	231.84	22.08	60.6	18.24
87	91.56	33.00	19.20	14.40
88	132.84	48.72	75.84	19.20
89	105.96	30.60	88.08	15.48
90	131.76	57.36	90	20.04
91	161.16	5.88	11.16	13.44
92	34.32	1.80	39.60	8.76
93	261.24	40.20	70.80	23.28
94	301.08	43.80	86.76	26.64
95	128.88	16.80	13.08	13.80
96	195.96	37.92	63.48	20.28
97	237.12	4.20	7.08	14.04
98	221.88	25.20	26.40	18.60
99	347.64	50.76	61.44	30.48
100	162.24	50.04	55.08	20.64
101	266.88	5.16	59.76	14.04
102	355.68	43.56	155	28.56
103	336.24	12.12	162	17.76
104	225.48	20.64	21.48	17.64
105	285.84	41.16	6.36	24.84
106	165.48	55.68	70.80	23.04
107	30.00	13.20	35.64	8.64
108	108.48	0.36	27.84	10.44
109	15.72	0.48	30.72	6.36

110	306.48	32.28	6.60	23.76
111	270.96	9.84	67.80	16.08
112	290.04	45.60	27.84	26.16
113	210.84	18.48	39.6	16.92
114	251.52	24.72	12.84	19.08
115	93.84	56.16	41.40	17.52
116	90.12	42.00	63.24	15.12
117	167.04	17.16	30.72	14.64
118	91.68	0.96	17.76	11.28
119	150.84	44.28	75	19.08
120	23.28	19.20	26.76	7.92
121	169.56	32.16	55.44	18.60
122	22.56	26.04	60.48	8.40
123	268.80	2.88	18.72	13.92
124	147.72	41.52	14.88	18.24
125	275.40	38.76	89.04	23.64
126	104.64	14.16	31.08	12.72
127	9.36	46.68	60.72	7.92
128	96.24	0.00	11.04	10.56
129	264.36	58.80	3.84	29.64
130	71.52	14.40	51.72	11.64
131	0.84	47.52	23.5	1.92
132	318.24	3.48	51.60	15.24
133	10.08	32.64	2.52	6.84
134	263.76	40.20	54.12	23.52
135	44.28	46.32	78.72	12.96
136	57.96	56.40	10.20	13.92
137	30.72	46.80	11.16	11.40
138	328.44	34.68	71.64	24.96
139	51.60	31.08	24.60	11.52
140	221.88	52.68	2.04	24.84
141	88.08	20.40	15.48	13.08
142	232.44	42.48	90.72	23.04
143	264.6	39.84	45.48	24.12
144	125.52	6.84	41.28	12.48
145	115.44	17.76	46.68	13.68
146	168.36	2.28	10.80	12.36
147	288.12	8.76	23.5	15.84
148	291.84	58.80	53.16	30.48
149	45.60	48.36	14.28	13.08
150	53.64	30.96	24.72	12.12
151	336.84	16.68	78.00	19.32
152	145.20	10.08	58.44	13.92

153	237.12	27.96	90	19.92
154	205.56	47.64	45.24	22.80
155	225.36	25.32	11.40	18.72
156	4.92	13.92	6.84	3.84
157	112.68	52.20	60.60	18.36
158	179.76	1.56	29.16	12.12
159	14.04	44.28	54.24	8.76
160	158.04	22.08	41.52	15.48
161	207.00	21.72	36.84	17.28
162	102.84	42.96	59.16	15.96
163	226.08	21.72	30.72	17.88
164	196.20	44.16	8.88	21.60
165	140.64	17.64	1.11	14.28
166	281.40	4.08	101.76	14.28
167	21.48	45.12	25.92	9.60
168	248.16	6.24	23.28	14.64
169	258.48	28.32	96	20.52
170	341.16	12.72	59.88	18.00
171	60.00	13.92	22.08	10.08
172	197.4	25.08	42.72	17.40
173	23.52	24.12	20.40	9.12
174	202.08	8.52	15.36	14.04
175	266.88	4.08	15.72	13.80
176	332.28	58.68	50.16	32.40
177	298.08	36.24	24.36	24.24
178	204.24	9.36	42.24	14.04
179	332.04	2.76	28.44	14.16
180	198.72	12.00	13.92	15.12
181	187.92	3.12	9.96	12.60
182	262.20	6.48	32.88	14.64
183	67.44	6.84	35.64	10.44
184	345.12	51.60	86.16	31.44
185	304.56	25.56	100.00	21.12
186	246.00	54.12	23.52	27.12
187	167.40	2.52	31.92	12.36
188	229.32	34.44	21.84	20.76
189	343.20	16.68	75.84	19.08
190	22.44	14.52	28.08	8.04
191	47.40	49.32	6.96	12.96
192	90.60	12.96	7.20	11.88
193	20.64	4.92	37.92	7.08
194	200.16	50.40	4.32	23.52
195	179.64	42.72	7.20	20.76

196	45.84	4.44	16.56	9.12
197	113.04	5.88	9.72	11.64
198	212.40	11.16	13.92	15.36
199	340.32	50.40	79.44	30.60
200	278.52	10.32	27	16.08

Table-1

Objectives of the Project

This project consists of 5 main objectives which are as follows:

- To determine whether there is a statistically significant relationship or not.
- To assess how the advertising budgets of three individual social media channels effect the total sales of the product.
- To develop a model that can predict sales on the basis of three social media budgets.
- To understand how well sales are predicted by the regression equation.
- To throw light on the contribution of the 3 social media sites(youtube, facebook, newspaper) to the prediction.

Setting up the Hypothesis

In the dataset, it can be observed that the dependent variable is sales and the independent variables are Youtube, Facebook and Newspaper. The multiple linear regression model for three independent variables is:

$\hat{Y} = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+ \varepsilon$

Where

X_1 = Advertising budget for Youtube

X_2 = Advertising budget for Facebook

X_3 = Advertising budget for Newspaper

$\hat{Y}$ = Predicted sales (in thousands of units)

β_i = Regression coefficient

ε = Residual

Pursuant to the purpose of this project, the null hypothesis is as follows:

H_o: There is no regression relationship between advertising budgets and sales of the product i.e. $\beta_1 = \beta_2 = \beta_3 = 0$

H_1: There is a regression relationship between advertising budgets and sales of the product i.e. at least one of the β is not equal to 0.

Methodology

Now to understand how advertising budgets of social media sites effects the overall sales in the market, we need to find a relationship between advertising budgets and sales. So, as there are 3 independent variables and 1 dependent variable, the statistical tool used in this project would be multiple linear regression. And also, as there are large number of observations, it would be convenient to use MS-Excel. So, by using the regression tool in the data analysis tool Pak in MS-Excel, we will perform multiple linear regression analysis. Further, we will perform model adequacy analysis by checking whether all assumptions holds true or not. If the model holds true for all the assumptions, we will summarize and interpret the results. And at last, we will finalize the MLR model.

Multiple Linear Regression

Regression Analysis is considered as the method used in statistics to predict future trends or values with the help of some relevant data. In this method, we determine the relationship between two or more variables and thus observe how a change in one variable affects another variable like in a cause and effect relationship which gives rise to the following variables:

- **Dependent variable:** It is a variable which we use to predict in multiple regression analysis and is also often called as explained or response variable.
- **Independent variable:** It is a variable which is used for prediction in multiple regression analsysi and is also often called as explanatory or regressor variable.

In general, if we want to study the relationship between one independent variable and dependent variable then we use linear regression but if we want to understand the relationship of two or more than two independent variable then we use multiple linear regression or simply multiple regression. It is an extension of linear regression which is used by many analysts to estimate the outcome of a dependent variable.

Now, if we have n independent variables, then a multiple linear regression will take the form as follows:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+ \ldots + \beta_nX_n+ \varepsilon$$

where:

- Y: dependent variable
- X_n: n^{th} independent variables
- β_0: y-intercept (constant term)
- β_n: slope coefficients for each independent variable

- ε: The regression residual(error term)

Assumptions in the model

The following assumptions must be considered when MLR model is constructed-

1. There should be a linear relationship between the dependent and independent variables.

2. Residuals should give the value of its mean as zero or close to zero.

3. Perfect multicollinearity is absent i.e. there is no or very weak correlation between the independent variables.

4. The residuals have constant variance i.e. homoscedasticity is present.

5. The residuals are normally distributed.

6. Count of independent variables should be lesser than the count of observations.

7. Outliers are absent.

Construction of Regression Model

Now, using the regression tool in the data analysis tool Pak in MS-Excel, we will perform multiple linear regression analysis. First, we will input all the values of sales in Y range column then input all the values of advertisings budgets in X range column. Now, keeping the confidence level at 95% (as $\alpha = 0.05$), we will select okay and then proceed further to the model adequacy analysis.

Model Adequacy Analysis

Now for model adequacy, we need to satisfy the model assumptions set earlier.

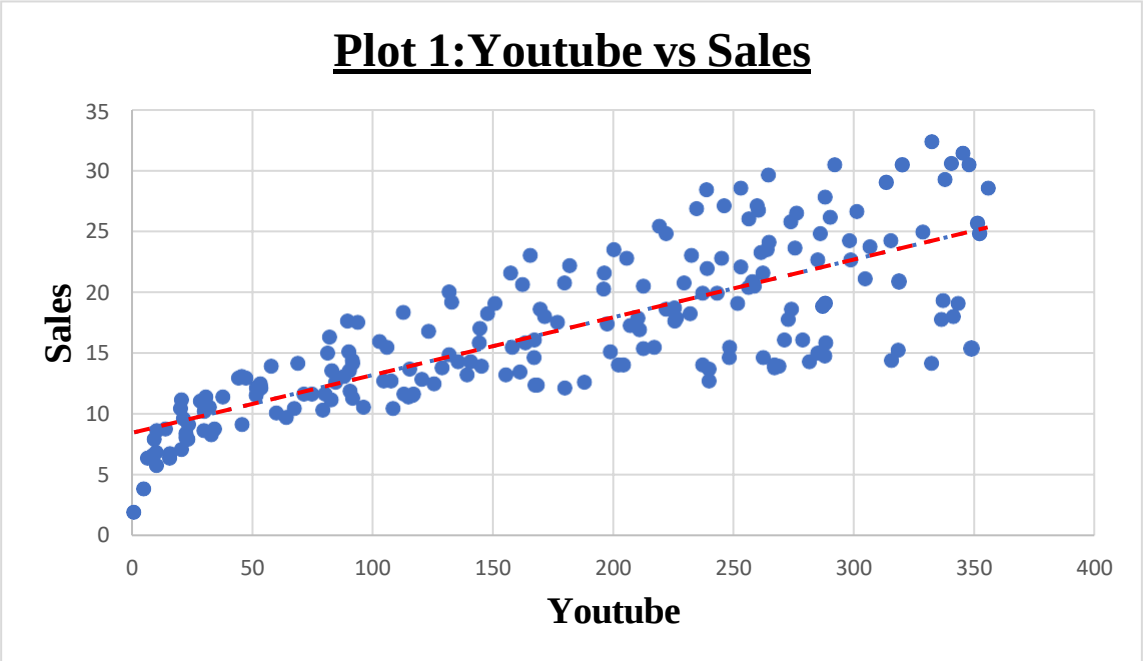
- **To check for linear relationship between dependent and independent variable.**

  The simplest way to detect the linearity between dependent and independent variable is to create a three scatter plots of the three independent variables(youtube, facebook, newspaper) vs dependent variable(sales). Now, check if the trendline shows an upper

trend or downward trend and how close the data points are to the trendline. Also, check the value of the correlation coefficient.

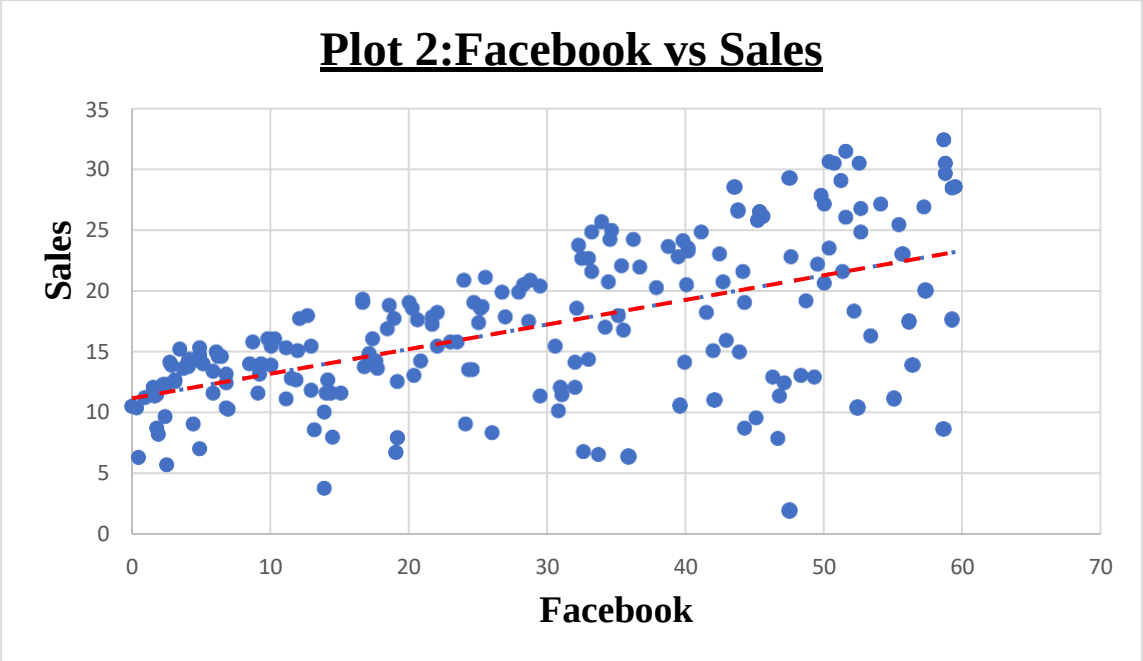
- If it shows upper trend ,it will indicate a positive relationship which means increase in independent variable will also results in the increase in dependent variable as well but if it shows downward trend, it will indicate a negative relationship which means increase in independent variable will result in decrease in the dependent variable.
- If many data points are close to the trendline, it will indicate linear relationship otherwise it will be a nonlinear relationship.
- More the value of correlation coefficient is away from zero, stronger the linearity between two variables.

Now, let's see the scatter plots according to our dataset:



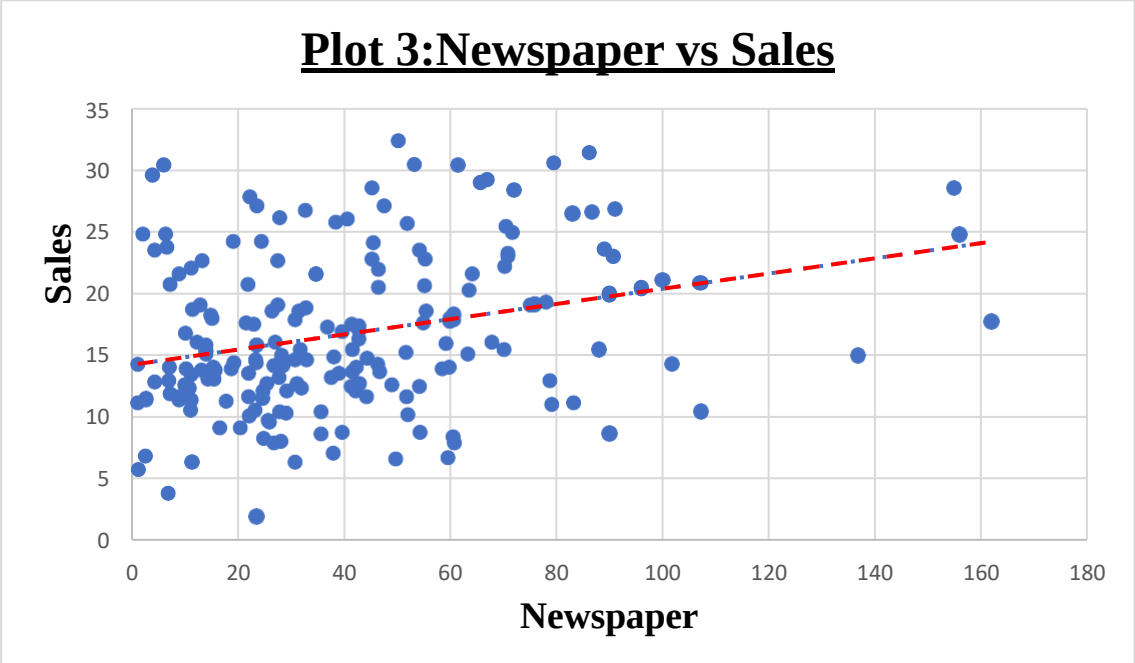
In the first scatter plot , we observe an upper trend and the value of many data points are close to the trendline. Also, the value of the correlation coefficient between the Advertising budget of Youtube and sales is 0.78222.

Therefore, the advertising budgets for youtube has a strong positive linear relationship with the sales.



In the second scatter plot , we observe an upper trend and the value of some data points are close to the trendline. Also, the value of the correlation coefficient between the Advertising budget of Facebook and sales is 0.57609.

Therefore, the advertising budgets for facebook has a moderate positive linear relationship with the sales.



In the third scatter plot , we observe a slightly upper trend and the value of many data points are far away from the trendline. Also, the value of the correlation coefficient between the advertising budget of newspaper and sales is 0.30383.

Therefore, the advertising budgets for newspaper has a weak positive linear relationship with the sales.

Hence, 1st assumption holds true.

- **To check for mean of residuals.**

The residuals are defined as the difference between the actual sales and predicted sales. It shows how far away the actual sales are from the predicted sales. In order to check this assumption, we must check whether the mean of the residuals are zero or close to zero which will indicate a good fit of the model. Let's check it in our dataset.

Observation	Sales	Predicted Sales	Residuals
1	26.52	24.39675092	2.123249084
2	12.48	14.68651028	-2.206510283
3	11.16	14.38242836	-3.222428363
4	22.20	20.9543352	1.245664796
5	15.48	15.52781189	-0.047811894
6	8.64	14.52999628	-5.889996284
7	14.16	14.29788709	-0.137887091
8	15.84	14.77778988	1.062210125
9	5.76	4.627523932	1.132476068
10	12.72	15.16336386	-2.443363859
11	10.32	8.400038327	1.919961673
12	20.88	20.08633302	0.793666981
13	11.04	12.2929364	-1.252936397
14	11.64	10.78689136	0.853108639
15	22.80	22.10537611	0.694623891
16	26.88	24.66742342	2.212576579
17	15.00	14.48495404	0.515045964
18	29.28	27.84904582	1.430954179
19	13.56	12.05290399	1.507096005
20	17.52	17.19155143	0.328448572
21	21.60	21.60535015	-0.005350155
22	15.00	17.81280558	-2.81280558
23	6.72	7.461708025	-0.741708025
24	18.60	20.00320961	-1.40320961
25	11.64	9.837890249	1.802109751
26	14.40	18.91717771	-4.517177711

27	18.00	18.27928808	-0.279288076
28	19.08	20.65786268	-1.577862677
29	22.68	23.55241171	-0.872411713
30	12.60	10.80233997	1.797660033
31	25.68	26.03453559	-0.354535585
32	14.28	13.51678103	0.763218966
33	11.52	9.421006312	2.098993688
34	20.88	21.99936456	-1.119364562
35	11.40	9.247863928	2.152136072
36	15.36	20.70918721	-5.349187213
37	30.48	28.65894988	1.821050118
38	17.64	18.70457048	-1.064570476
39	12.12	11.81742271	0.302577294
40	25.80	24.73002217	1.069977834
41	19.92	19.19170931	0.72829069
42	20.52	20.80078065	-0.280780655
43	24.84	24.82935458	0.010645416
44	15.48	16.83695533	-1.356955335
45	10.20	10.47621582	-0.276215818
46	17.88	18.0199028	-0.139902798
47	12.72	10.52109071	2.198909289
48	27.84	26.44657332	1.393426677
49	17.76	19.38521958	-1.625219577
50	11.64	9.65074566	1.98925434
51	13.68	15.10853229	-1.428532289
52	12.84	11.45789323	1.382106772
53	27.12	24.91188781	2.208112185
54	25.44	23.82339243	1.616607572
55	24.24	24.81150462	-0.571504619
56	28.44	25.458228	2.981771999
57	6.60	10.06225386	-3.462253856
58	15.84	15.47698249	0.363017512
59	28.56	26.45305329	2.106946707
60	22.08	22.15381555	-0.073815546
61	9.72	6.851960239	2.868039761
62	29.04	27.45858162	1.58141838
63	18.84	20.27923031	-1.439230315
64	16.80	16.15734716	0.642652844
65	21.60	20.53946473	1.060535273
66	11.16	9.669309383	1.490690617
67	11.40	11.10221173	0.297788268
68	16.08	14.68103184	1.398968162
69	22.68	23.1599735	-0.4799735
70	26.76	25.6016613	1.158338701
71	21.96	21.38642111	0.573578893
72	14.88	12.71354947	2.166450528

73	10.56	12.56749457	-2.007494574
74	13.20	11.82135521	1.378644791
75	20.40	20.2694286	0.130571399
76	10.44	13.60018777	-3.160187768
77	8.28	5.316302281	2.963697719
78	17.04	16.55022897	0.489771033
79	6.36	10.78053024	-4.420530244
80	13.20	11.64068548	1.559314519
81	14.16	13.8400033	0.319996701
82	14.76	17.54458795	-2.784587947
83	13.56	12.16640534	1.393594658
84	16.32	17.34649628	-1.026496284
85	26.04	25.12488774	0.915112265
86	18.24	18.08108065	0.158919347
87	14.40	14.09992835	0.300071645
88	19.20	18.45394918	0.746050821
89	15.48	13.57590671	1.90409329
90	20.04	19.92099877	0.119001234
91	13.44	12.18984951	1.250150495
92	8.76	5.197941715	3.562058285
93	23.28	22.83826935	0.441730647
94	26.64	25.21772994	1.422270063
95	13.80	12.77490946	1.025090536
96	20.28	19.43801127	0.841988727
97	14.04	15.44378401	-1.403784006
98	18.60	18.58267578	0.017324225
99	30.48	28.99646092	1.483539079
100	20.64	20.29717799	0.342822011
101	14.04	16.45496575	-2.414965746
102	28.56	26.98814672	1.571853277
103	17.76	19.94110235	-2.181102353
104	17.64	17.92230922	-0.282309221
105	24.84	24.85215837	-0.012158375
106	23.04	21.36969262	1.670307375
107	8.64	7.239192539	1.400807461
108	10.44	8.495929764	1.944070236
109	6.36	4.171832395	2.188167605
110	23.76	24.09622775	-0.336227747
111	16.08	17.46278261	-1.382782608
112	26.16	25.67661791	0.483382087
113	16.92	16.63184441	0.288155587
114	19.08	20.01325926	-0.933259262
115	17.52	18.44046493	-0.920464935
116	15.12	15.30263149	-0.182631493
117	14.64	14.43302322	0.206976781
118	11.28	7.936890495	3.343109505

119	19.08	18.44356893	0.636431066
120	7.92	8.178734899	-0.258734899
121	18.60	17.18306059	1.416939415
122	8.40	9.107563769	-0.707563769
123	13.92	16.53976346	-2.619763461
124	18.24	18.40365142	-0.163651416
125	23.64	23.0257374	0.614262598
126	12.72	10.94632793	1.773672069
127	7.92	12.47437718	-4.554377184
128	10.56	8.035120167	2.524879833
129	29.64	27.28392314	2.35607686
130	11.64	9.232346815	2.407653185
131	1.92	12.63497851	-10.71497851
132	15.24	18.60740553	-3.367405532
133	6.84	10.41567988	-3.575679882
134	23.52	23.13254251	0.387457487
135	12.96	13.8389294	-0.8789294
136	13.92	17.1481508	-3.228150799
137	11.40	14.01746313	-2.617463127
138	24.96	24.89119573	0.068804272
139	11.52	11.81246862	-0.292468624
140	24.84	24.14497197	0.69502803
141	13.08	11.54558574	1.534414256
142	23.04	21.72669658	1.313303421
143	24.12	23.19383533	0.926164671
144	12.48	10.39696307	2.083036931
145	13.68	11.97819788	1.701802124
146	12.36	11.83390967	0.526090329
147	15.84	18.52299255	-2.682992552
148	30.48	28.03937197	2.44062803
149	13.08	14.97790219	-1.897902192
150	12.12	11.88296836	0.237031639
151	19.32	21.74050694	-2.420506945
152	13.92	11.75603814	2.163961856
153	19.92	19.14968339	0.770316609
154	22.80	21.95433027	0.845669731
155	18.72	18.92695682	-0.206956822
156	3.84	6.516644036	-2.676644036
157	18.36	18.34916078	0.010839221
158	12.12	12.03062856	0.089371436
159	8.76	12.29771618	-3.537716184
160	15.48	14.84918005	0.630819951
161	17.28	17.10777206	0.172227943
162	15.96	16.1231527	-0.1631527
163	17.88	18.06062565	-0.180625646
164	21.60	21.23293054	0.367069464

165	14.28	13.611321	0.668678999
166	14.28	16.47653657	-2.196536571
167	9.60	13.10659154	-3.506591536
168	14.64	16.17936675	-1.539366751
169	20.52	20.14950859	0.370491406
170	18.00	21.36952337	-3.369523374
171	10.08	8.918135904	1.161864096
172	17.40	17.24712748	0.152872516
173	9.12	9.206981637	-0.086981637
174	14.04	14.55907645	-0.519076446
175	13.80	16.71385572	-2.913855725
176	32.40	29.92996951	2.470030488
177	24.24	24.28116387	-0.041163868
178	14.04	14.53647881	-0.496478813
179	14.16	19.35640611	-5.196406109
180	15.12	15.08965548	0.03034452
181	12.60	12.91519602	-0.315196023
182	14.64	16.77718453	-2.137184526
183	10.44	7.753991007	2.686008993
184	31.44	28.77899506	2.661004941
185	21.12	21.71875374	-0.598753743
186	27.12	25.31741519	1.802584807
187	12.36	11.6114387	0.748561303
188	20.76	20.76066678	-0.000666784
189	19.08	22.0593982	-2.979398203
190	8.04	7.22237043	0.81762957
191	12.96	15.32462993	-2.364629932
192	11.88	10.31475874	1.565241262
193	7.08	5.18133205	1.89866795
194	23.52	22.66995849	0.850041511
195	20.76	20.20219062	0.557809381
196	9.12	6.48806558	2.63193442
197	11.64	9.965809007	1.674190993
198	15.36	15.5641457	-0.2041457
199	30.60	28.39532466	2.204675335
200	16.08	18.34019338	-2.260193377

Table-2

Sum of Residuals = -3.37508E-13

Mean of Residuals = -3.37508E-13/200 = -1.68754E-15

Therefore, we can observe that the mean of residuals is very close to zero.

Hence, 2nd assumption holds true.

- **To check for multicollinearity.**

Since we verified the linear relationships in the first assumption with the help of Scatter Plots, there can be an indication that it exhibits multicollinearity which is a big problem for our model.

Multicollinearity arises to the extent when one independent variable is perfectly or highly correlated to another variable. To identify multicollinearity, we will observe the correlation matrix:

	Youtube	**Facebook**	**Newspaper**
Youtube	1		
Facebook	0.0546261	1	
Newspaper	0.2589744	0.2872942	1

Table-3

Hence, we can observe that the correlation between Facebook and Newspaper is 0.2873 which is considered as high. However this is an casual observation.

A more suitable way to check multicollinearity is with the help of Variance Inflation Factor(VIF).

VIF tells us how inflated the estimated regression coefficents are. It is also the way to exactly express the same information found in the coefficient of multiple correlation.

We compute VIF for each indendenpent variable with the help of formula:

$$\text{VIF} = \frac{1}{1-R_k^2}$$

Where R_k = the correlation coefficient for each independent variable on other Independent variables

Now, if VIF is equal to 1, then variables are not correlated to each other , if VIF lies in the range of 1 to 5 , then variables are moderately correlated to each other but if VIF lies in the range of 5 to 10, then variables are highly correlated to each other (which will be a major problem to our model).

Results of VIF for each of the follwing independent variable on other independent variables are as follows:

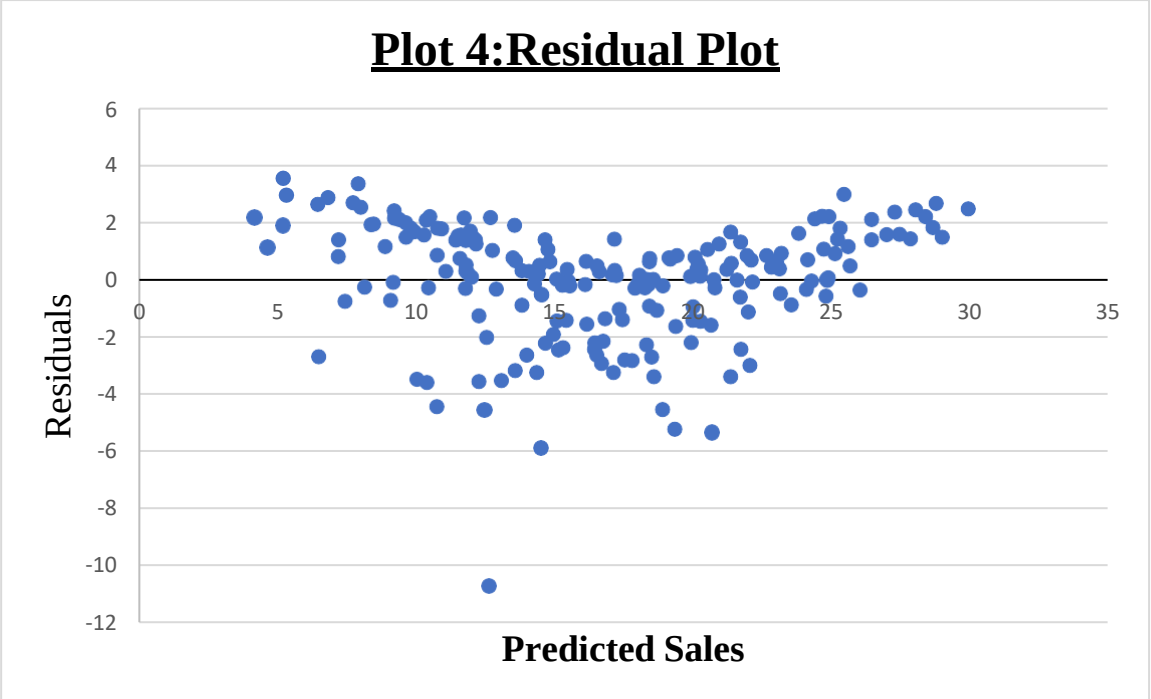
Independent variable	**VIF**
$\boldsymbol{X_1}$	1.07238
$\boldsymbol{X_2}$	1.09046
$\boldsymbol{X_3}$	1.16536

Table-4

Although the chance that an independent variable is already explained by another independent variable increases with the increase in values of VIF but in the following table of VIF, we observe that the values of all VIF of each independent variables on other independent variables is around 1 so it will not create problem to our model. Hence, 3rd assumption holds true.

- **To check for homoscedasticity.**

When the residuals have equal or almost equal variance across the regression line, then it is called homoscedasticity. When homoscedasticity is absent, then the regression coefficients loses their efficiency. A casual effective way to see that a model is homoscedastic or heteroscedastic is the residual plot. In this plot, residuals are plotted on y-axis and predicted sales are plotted on x-axis. If the plot will tend to spread out with the increase in values of predicted sales (i.e. funnel shape) then the homoscedasticity is absent.



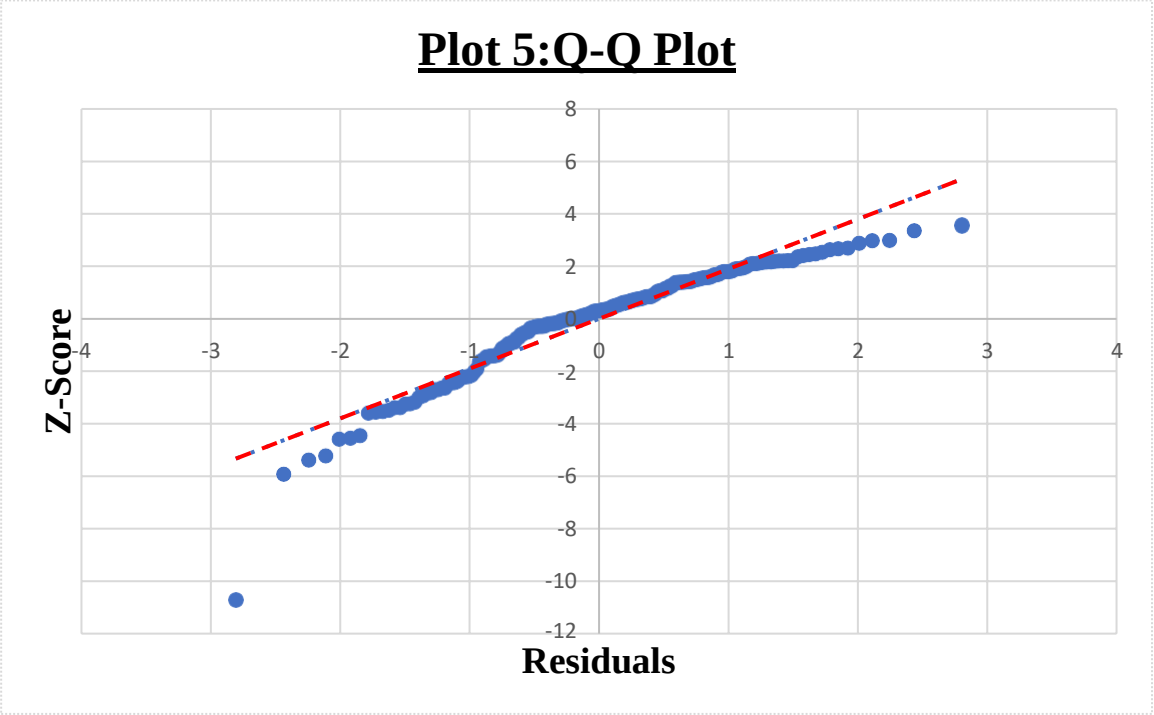
Here, the negative values for the residual means that the prediction was too high and the positive values means that the prediction was too low while values at 0 means that the prediction was exactly correct. As the above plot is not showing any spread of

residual values with increasing x-axis values, we can conclude that the model is homoscedastic.

Hence, 4th assumption holds true.

- **To check for normality.**

Now, to check whether the residuals are normally distributed or not, we will use Q-Q plot which helps us to verify this assumption. Q-Q plot are made using the residuals which are first sorted in ascending order with the Z-scores of each observation in the model.



Here, we can see that the data appears to be a rough straight line. Although it is also observed that the beginning and end observations deviate from the trendline but those deviations are small so it cannot be a cause to concern. So, the residuals are normally distributed.

Hence. 5th assumption holds true.

- **To check whether the count of independent variables is less than the number of observations.**

  Our model should contain less independent variables than the total number of observations. Here, our model contains 200 observations and 3 independent variables which is less than the total number of observations.

  Hence, 6th assumption hold true.

- **To check for outliers.**

  We can observe that the outliers are absent in the above scatter plots.

  Hence, last 7th assumption also holds true.

  Therefore, the regression model is adequate as all of the assumptions holds true for this model. So, now we can finalize our regression results.

Interpretation of the Regression Analysis

Regression Results are summarized below:

1) The first task is to check the significance of the model.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	7017.051394	2339.017131	585.020821	1.665E-97
Residual	196	783.642806	3.998177582		
Total	199	7800.6942			

Table-5

ANOVA divides the sum of squares into separate parts that provides details regarding the amount of variability within the MLR model.

The smaller the residual sum of squares compared with the total sum of squares, the better the regression model fits the data. The overall significance of the model is tested by F-statistic. The p-value or significance F value gives an idea about the reliability of our results. The F statistic of 585.0208 (p-value or significance F = 1.665E-97) implies that the regression as a whole is statistically significant.

2) The second task is to check how three advertising budgets effect the total sales of the product.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.67363006	0.35907050	10.2309435	5.8658E-20	2.9654923	4.3817678
Youtube	0.04653618	0.0014247	32.6629245	3.15893E-81	0.0437264	0.049346
Facebook	0.19300571	0.00830705	23.2339618	3.30659E-58	0.176623	0.2093884
Newspaper	-0.01061162	0.0049595	-2.13963114	0.033621141	-0.020393	-0.000831

Table-6

Here, the test of significance of individual social media channels is measured by T-statistic. We can see that all the three social media channels are statistically significant at 5% level of significance since p-value of all three channels is less than 0.05. Here, the lower 95% and upper 95% are the lower and upper limits of the confidence interval respectively. Therefore, the null hypothesis is rejected, so there is a regression relationship between advertising budgets and sales of the product as none of the β's are equal to zero.

3) The third task is to define a regression equation which predicts sales on the basis of three social media channels. Now, since our model has three independent variables and one dependent variable.

So, our output comes out to be :

	Coefficients
Intercept	3.673630064
Youtube	0.046536185
Facebook	0.19300571
Newspaper	-0.010611629

Table-7

So, our final regression model is specified as follows:

$\hat{Y} = 3.674 + 0.046X_1 + 0.193X_2 - 0.011X_3$

Where $\hat{Y}$ = Predicted sales(in thousands of units)

X_1 = Advertising Budget for youtube

X_2 = Advertising Budget for facebook

X_3 = Advertising Budget for newspaper

4) The fourth task is to assess how well sales are affected are predicted by the regression equation i.e. how well model fits the data.

Regression Statistics	
Multiple R	0.948441835
R Square	0.899541914
Adjusted R Square	0.898004291
Standard Error	1.999544343
Observations	200

Table-8

Multiple R: It measures the strength of the MLR model by providing the correlation coefficient between the variables. The value of Multiple R ranges from -1 to 1. Here, 0.948 value of multiple R tells that the variables have a strong positive relationship.

R Square: It is the Coefficient of Determination, which shows how many points fall on the regression line. Goodness of fit can be checked using R square.

Adjusted R Square: It is the adjusted value of R-Square for the number of independent variables in the model. Here the value of Adjusted R-Square is 0.8980 which suggests it is a good fit as about 89.80% of the variation in sales is explained by the three independent variables combined.

Standard error: The lesser the standard error, the more certain we can be about our MLR model. It is another goodness of fit measure that shows the precision of our regression analysis. It is an absolute measure that tells about the average distance that the data points fall from the regression line.

5) The fifth task is to determine the contribution of three social media channels which is attained by the equation:

Sales= 3.674 + 0.046***Youtube** + 0.193***Facebook** - 0.011***Newspaper**

Since the three independent variables are statistically significant. Therefore, for every increase in youtube and facebook budgets will lead to increase of sales by 0.046 and 0.193 respectively but for every increase in newspaper budgets will lead to decrease in sales by 0.011.

Conclusion

Therefore, there is a multiple linear regression relationship between the advertising budgets for 3 different social media channels and sales of the product. Also, it is analysed that the youtube and facebook advertising budgets has greater effect on sales of the product than the advertising budgets of newspaper. So, the regression equation $\hat{Y}= 3.674 + 0.046X_1 +$

$0.193X_2 - 0.011X_3$ is a good and adequate fitted model which can predict sales on the basis of the advertising budgets of 3 different social media channels.

References

1. https://www.kaggle.com/
2. An Introduction to Statistical Learning – By Gareth James
3. Fundamentals of Applied Statistics – By Gupta and Kapoor
4. www.dimensionless.in/multiple-linear-regression-assumptions
5. https://stattrek.com/multiple-regression/regression-coefficients.aspx