Customer Value Analysis

Shresth Sharma

Table of Contents

Value of customers: What makes a customer Valuable?

1. Executive Summary:

- The business problem at hand revolves around understanding customer behaviour.
- Identifying factors that significantly impact customer spending on the 'drinks@home.uk' website.
- Recommend the most suitable marketing strategy to increase profits.

Data Analysis:

- Corrgram and Ggplots are used for **visual data analysis**.
- Linear Regression is used for **statistical data analysis**.

Results:

- **"Seen Voucher" and "Advertisement through an Influencer"** gives the better performing values than other explanatory variables.
- **Option 3:** Spend more money on advertising with an **influencer** is recommended as an optimal solution for Task 2.

2. Introduction to Business Problem:

The report will focus on two main business questions:

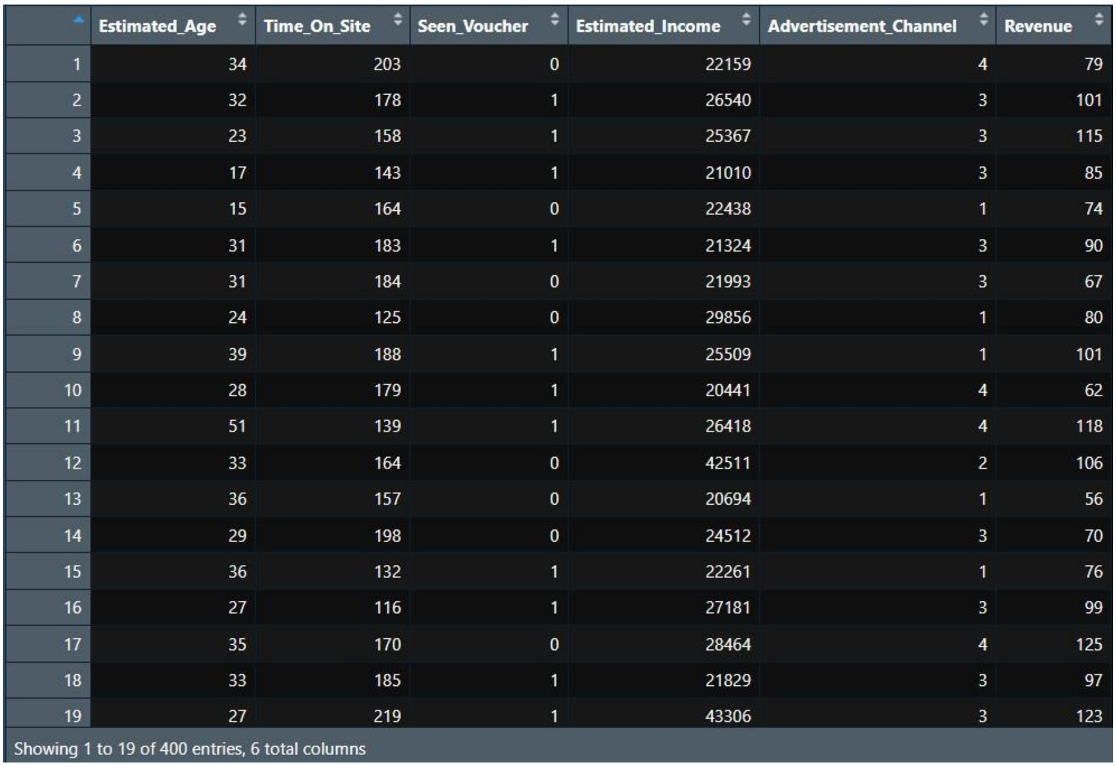
- Determining Factors Influencing Customer Spending:
  1. Analysing the given demographic record of 400 customers
  2. Identifying the behavioural pattern of customers.
  3. Identify the key factors that significantly affect Revenue.
  4. Exploring correlations between variables like age, income, etc.

- Recommendation on Marketing Project:
  1. Evaluating and recommending the most effective marketing project.
  2. Recommendation will be based on insights derived from the analysis, considering factors such as
     - Potential impact on revenue,
     - Alignment with customer behaviour patterns,
     - Profitability.

3. Data-Set Description:

- Variables:
  - Revenue (GBP) – Revenue generated.
  - Estimated Age of Customer
  - Seen Voucher –If the customer has seen any discount voucher popup.
  - Estimated Income of Customers
  - Time on website spent by Customers per week (Seconds).
  - Advertisement Channel – 1 Leaflet, 2 social media, 3 Search Engine, 4 Influencer.

- Format of data of customers:

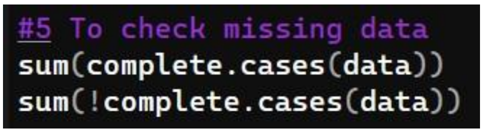


	Estimated_Age	Time_On_Site	Seen_Voucher	Estimated_Income	Advertisement_Channel	Revenue
1	34	203	0	22159	4	79
2	32	178	1	26540	3	101
3	23	158	1	25367	3	115
4	17	143	1	21010	3	85
5	15	164	0	22438	1	74
6	31	183	1	21324	3	90
7	31	184	0	21993	3	67
8	24	125	0	29856	1	80
9	39	188	1	25509	1	101
10	28	179	1	20441	4	62
11	51	139	1	26418	4	118
12	33	164	0	42511	2	106
13	36	157	0	20694	1	56
14	29	198	0	24512	3	70
15	36	132	1	22261	1	76
16	27	116	1	27181	3	99
17	35	170	0	28464	4	125
18	33	185	1	21829	3	97
19	27	219	1	43306	3	123

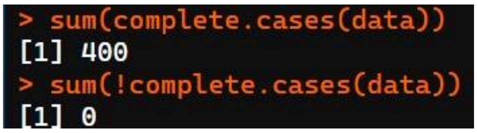
Showing 1 to 19 of 400 entries, 6 total columns

4. Data Preparation:

- The data has no missing values and can be used for **correlational and regression analysis.**

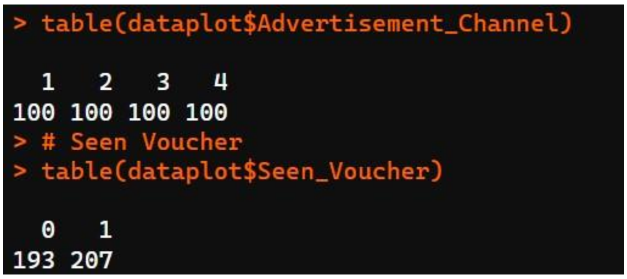


```
#5 To check missing data
sum(complete.cases(data))
sum(!complete.cases(data))
```



```
> sum(complete.cases(data))
[1] 400
> sum(!complete.cases(data))
[1] 0
```

- There are **100** transactions done through each **Advertisement Channel**.



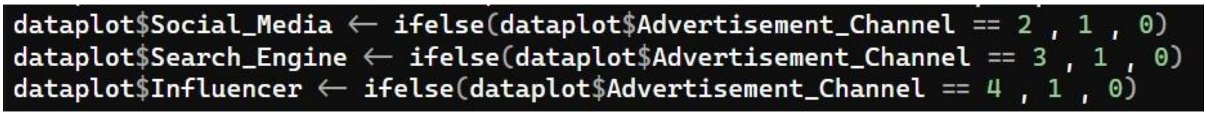
```
> table(dataplot$Advertisement_Channel)

  1   2   3   4 
100 100 100 100 
> # Seen Voucher
> table(dataplot$Seen_Voucher)

  0   1 
193 207 
```

- **193** transactions were done who didn't see any voucher whereas **207** were done who've seen vouchers.

- Average revenue generated by the sample is **80 GBP.**

Dummy Variables: Used for changing categorical values to binary numerical values.

This way we can record the impact of each variable on the Revenue.



```
dataplot$Social_Media <- ifelse(dataplot$Advertisement_Channel == 2 , 1 , 0)
dataplot$Search_Engine <- ifelse(dataplot$Advertisement_Channel == 3 , 1 , 0)
dataplot$Influencer <- ifelse(dataplot$Advertisement_Channel == 4 , 1 , 0)
```

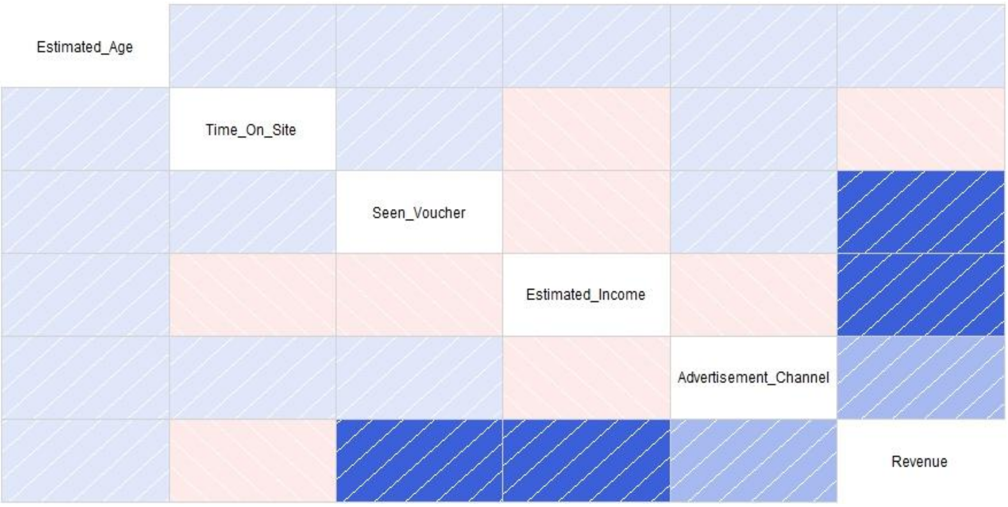
There would be evident behavioural findings as we proceed towards Correlative Analysis.

5. Correlation Analysis:

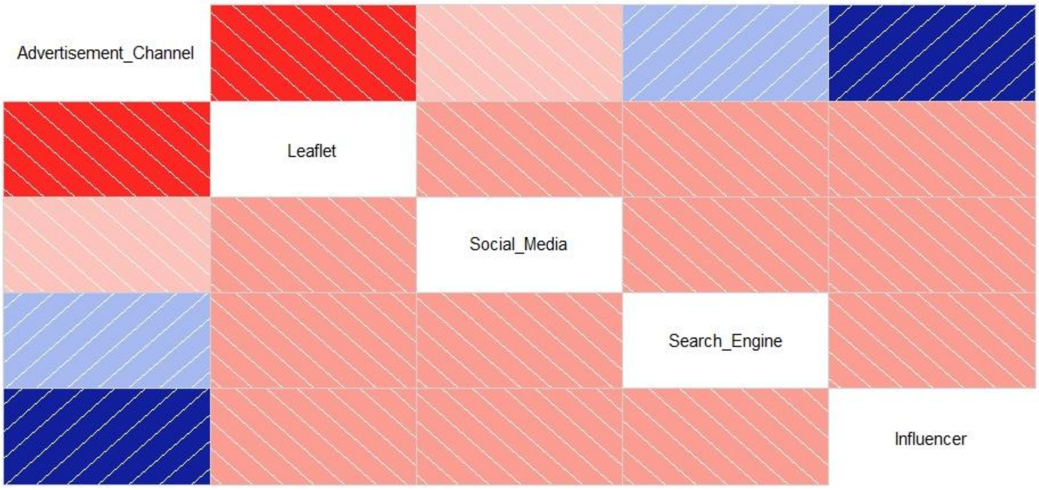
5.1 Using Corrgram:

- Examining the link between one numerical variable and another is typically helpful when using linear regression. This can be done in R.
- For Correlation Analysis we will use "corrgram" Package in R Studio to check if there's any relationship between variables.
- A **positive correlation** means that two values are highly correlated when one is high.
- A **negative correlation** means that a high value corresponds with a low value.
- Dark shade indicates strong correlation, **blue** indicates positive and **red** indicate negative between two variables.

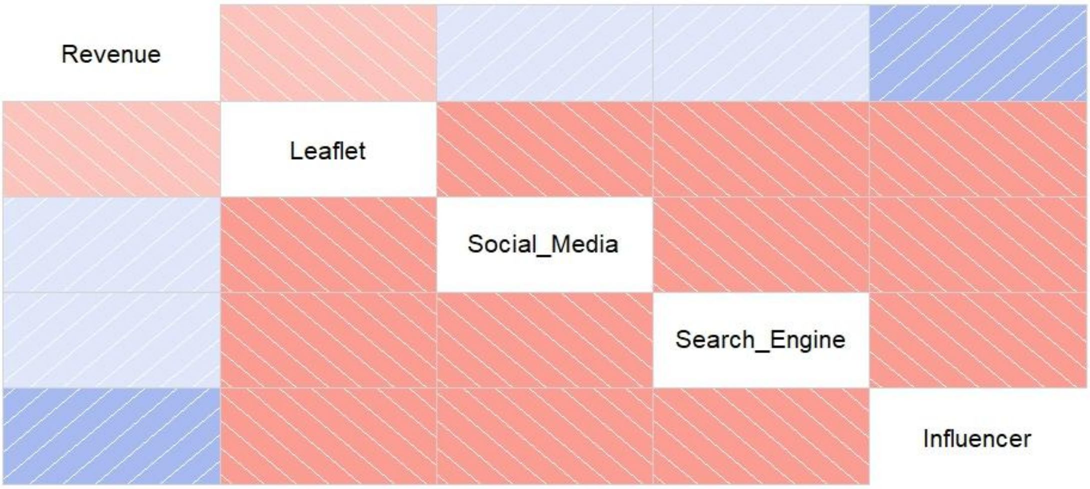
(Taiyun Wei et al 2017)
Corrgram:



- Relationship with Revenue:
  - **Seen voucher** and **Estimated Income** shares **a strong positive** correlation.
  - Advertisement Channel also shares a **positive correlation** but seemingly less impactful.
  - Estimated Age & Time on site seems to have **no significant impact**.



- Relationship with (Advertisement Channel’s Performance):
  - **Influencer and search engine** contributes **positively**.
  - **Social Media** has **no significant** impact.
  - **Leaflet** seemingly leaves a **strong negative impact**.



- Relationship with Revenue:
  - **Influencer** contributes **positively**.
  - **social media** and **search engine** seems to have **no significant impact**.
  - Leaflet seems to share a negative correlation.

With all these interesting insights, we could find some statistical evidence through plots and linear regression analysis.

5.2 Using GGPLOTS:

Assumptions for using **Linear Regression**:

(i) Each value of Xi and of Y is observed without measurement error Poole et al, 1971.

(ii) The relationships between Y and each of the independent variable Xi are linear in the parameters of the specific function.

(iii) The error (distance from the line) of each point should be independent from the error of other points.

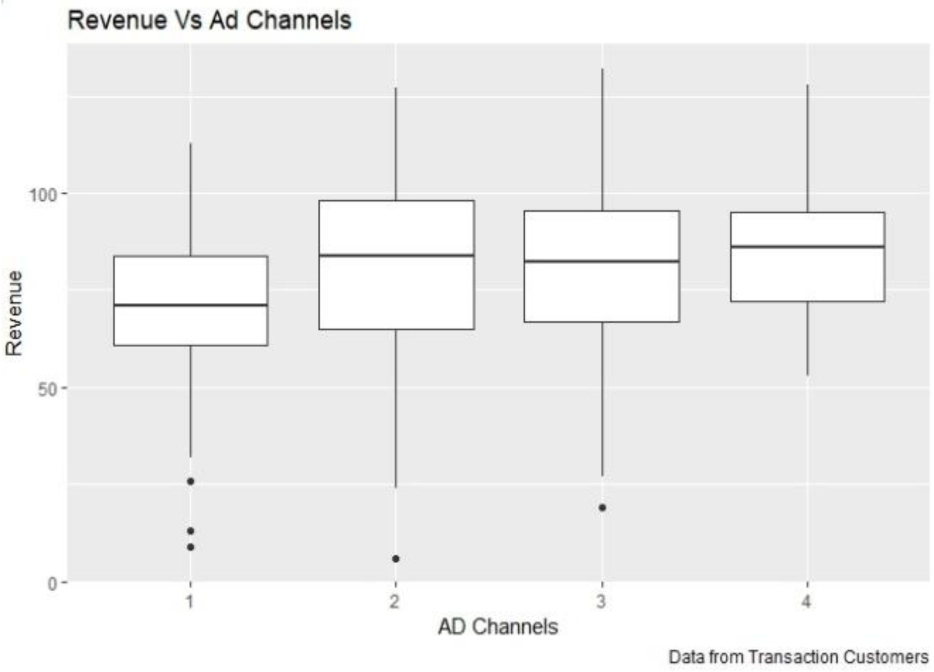
Boxplots:

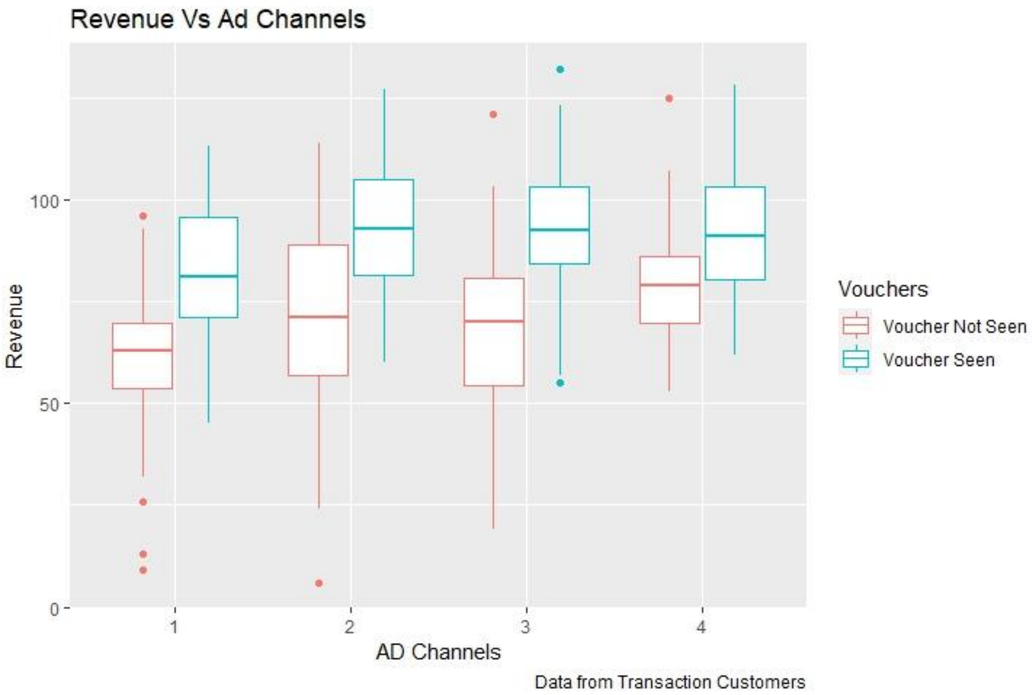
A boxplot aims to summarize a batch of data by displaying several main features. (Michael Frigge et al, 1989)

Box Signifies 50% of the total observation, vertical lines show the 25% on each side of the boxes. Horizontal lines show the median value of the observation. (Heike Hofmann et al, 2017).

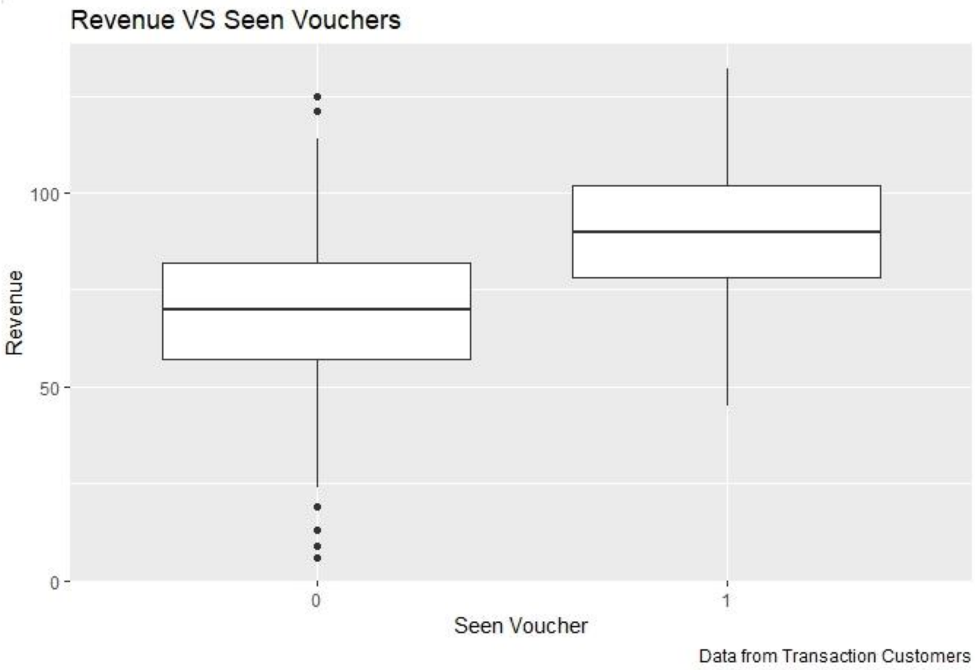
Scatterplots:

A scatter plot uses dots to represent values for two different numeric variables. Scatter plots are used to observe relationships between variables. Friendly, M., & Wainer, H. (2021)

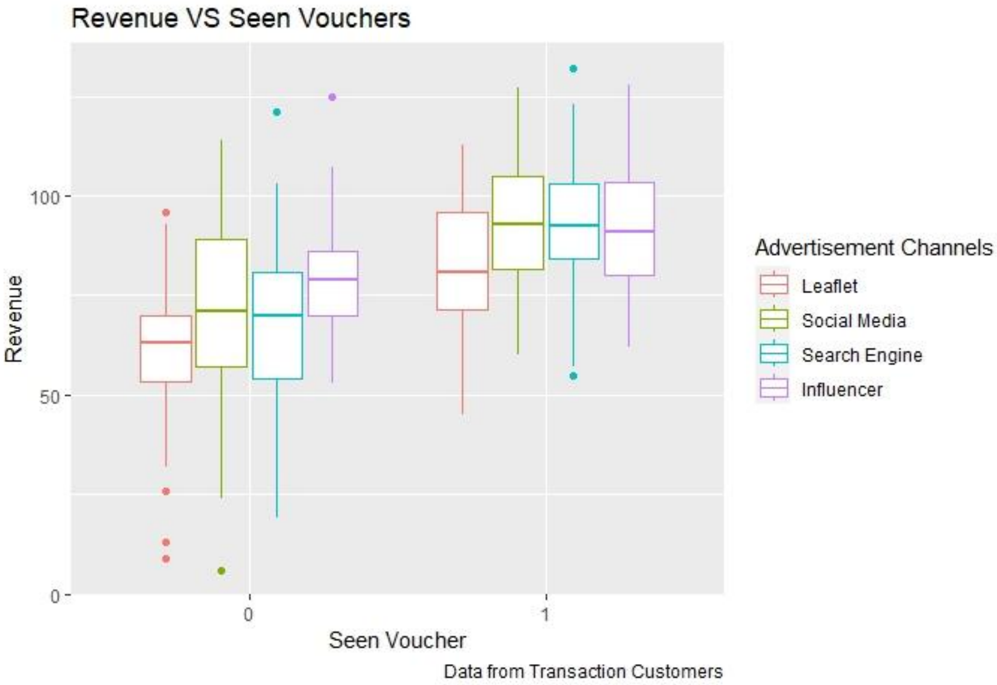




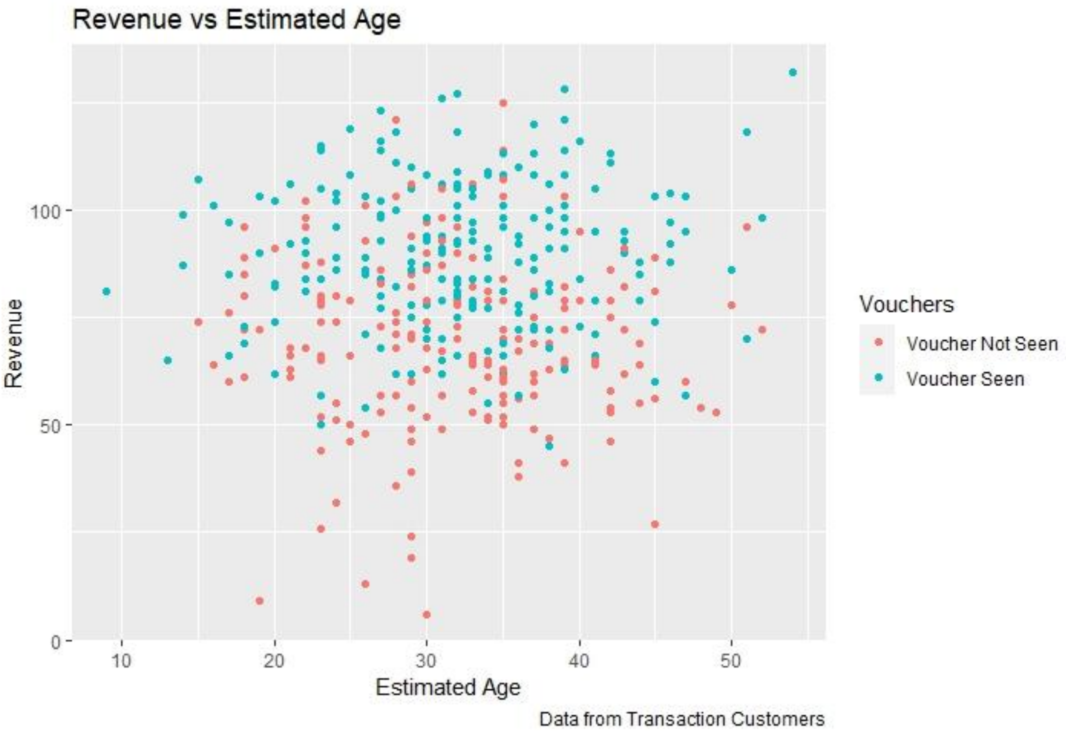
Customers seeing vouchers contributes **more towards revenue irrespective of different Advertisement Channels.**

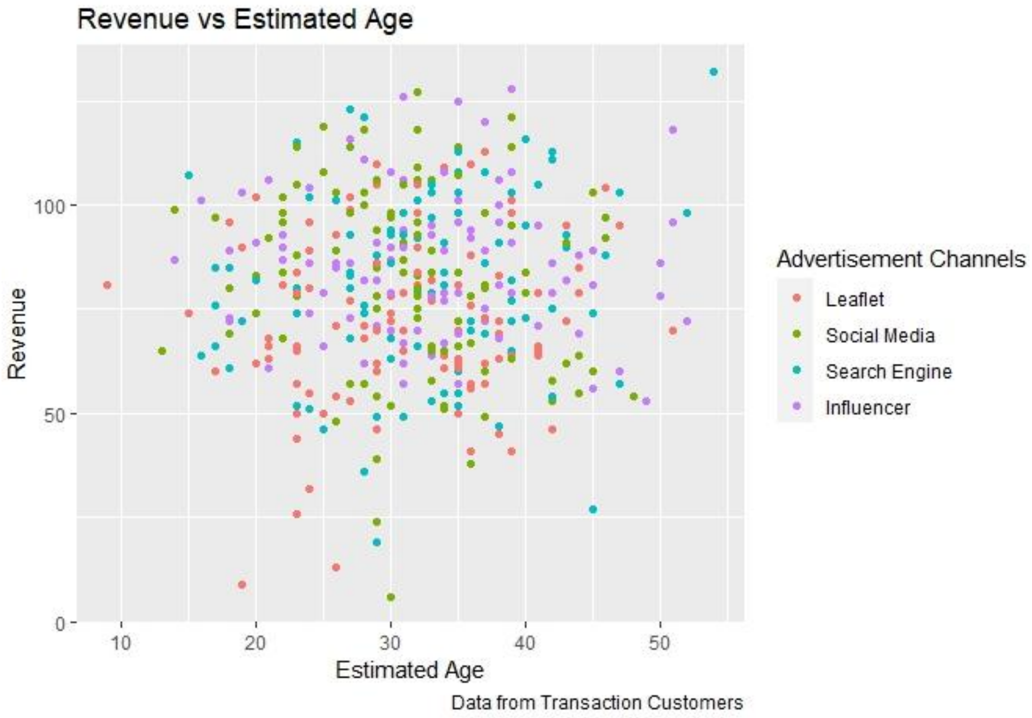


Evidently, **correlation between seen voucher and revenue** can be seen.

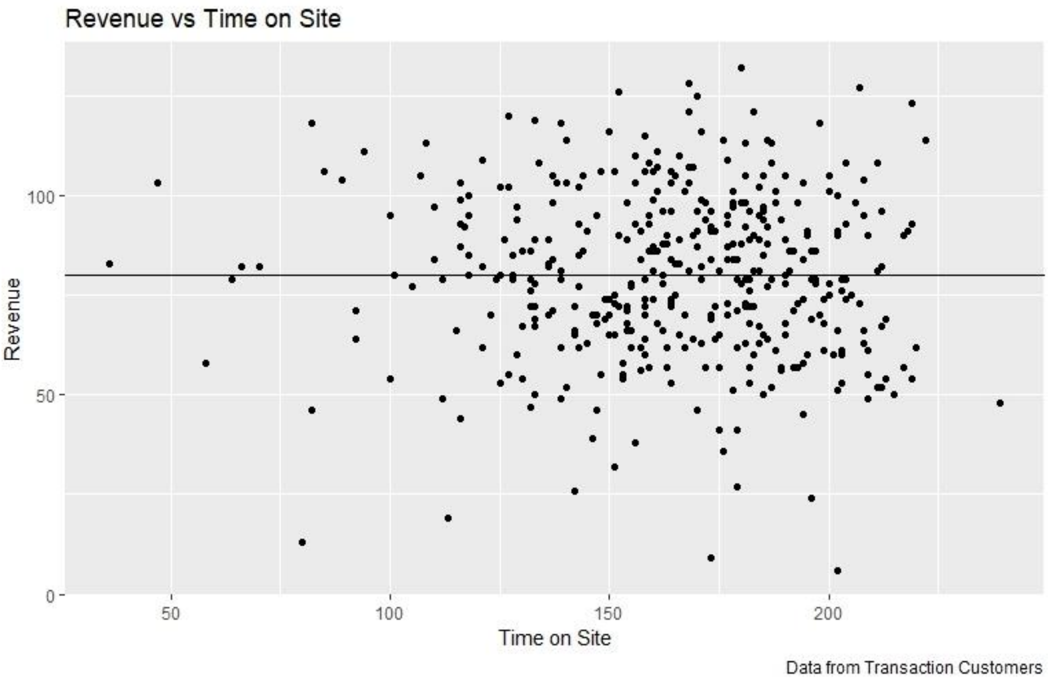


Transactions through **leaflet contributes least towards revenue** in both cases**.**



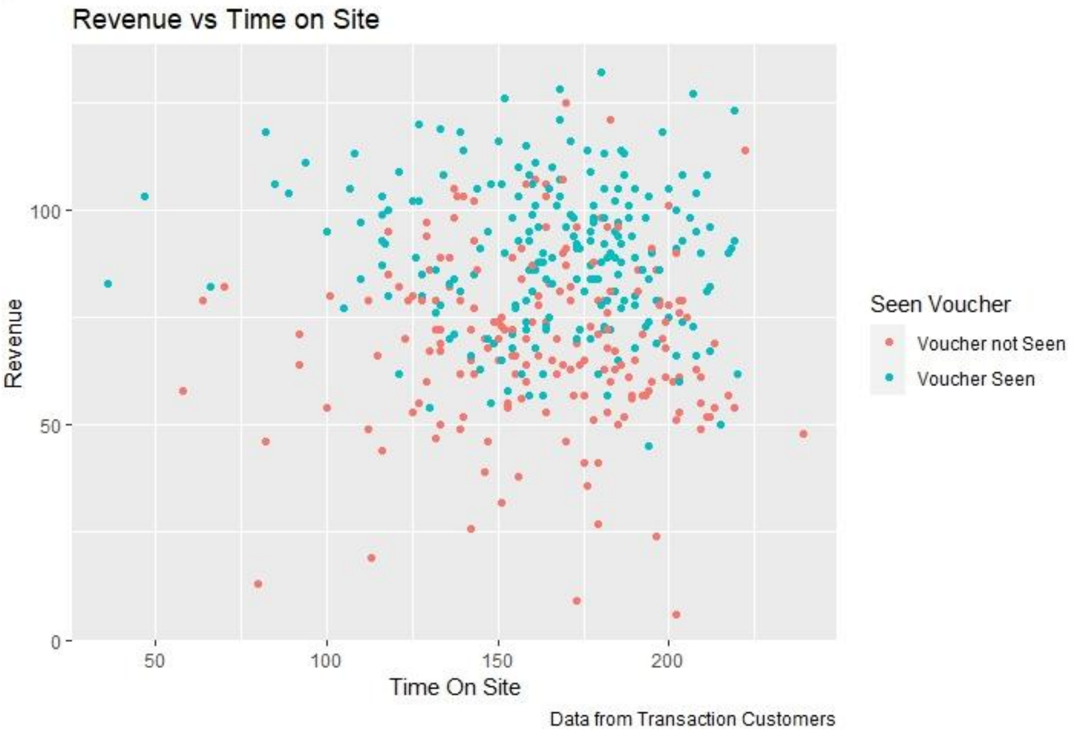


There is **no substantial correlation** between Revenue and Estimated Age irrespective of Seen Vouchers or Advertisement Channels.


Revenue vs Time on Site

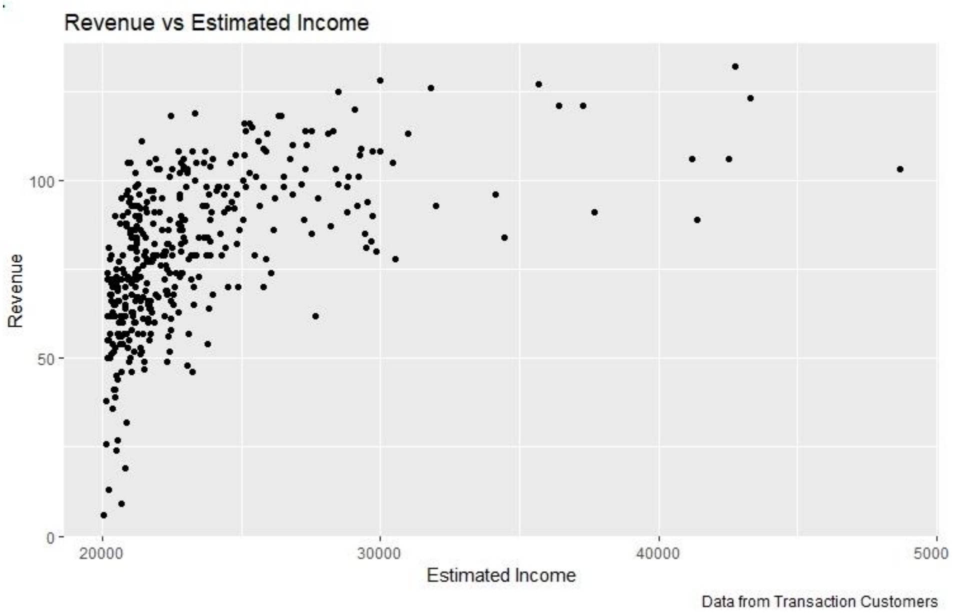
Data from Transaction Customers

Data scattered across takes a shape of cloud, so there is no linear relationship seen.


Revenue vs Time on Site

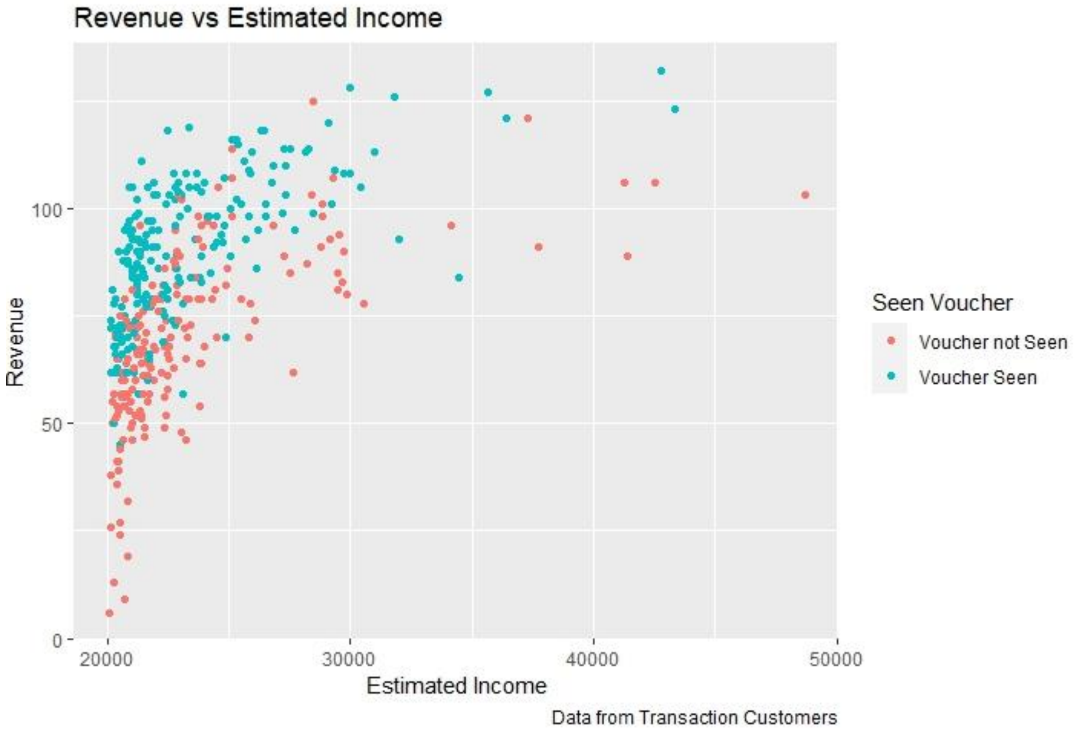
Data from Transaction Customers

Even after adding seen voucher variable it seems there’s not any relationship between Revenue and Time on Site.

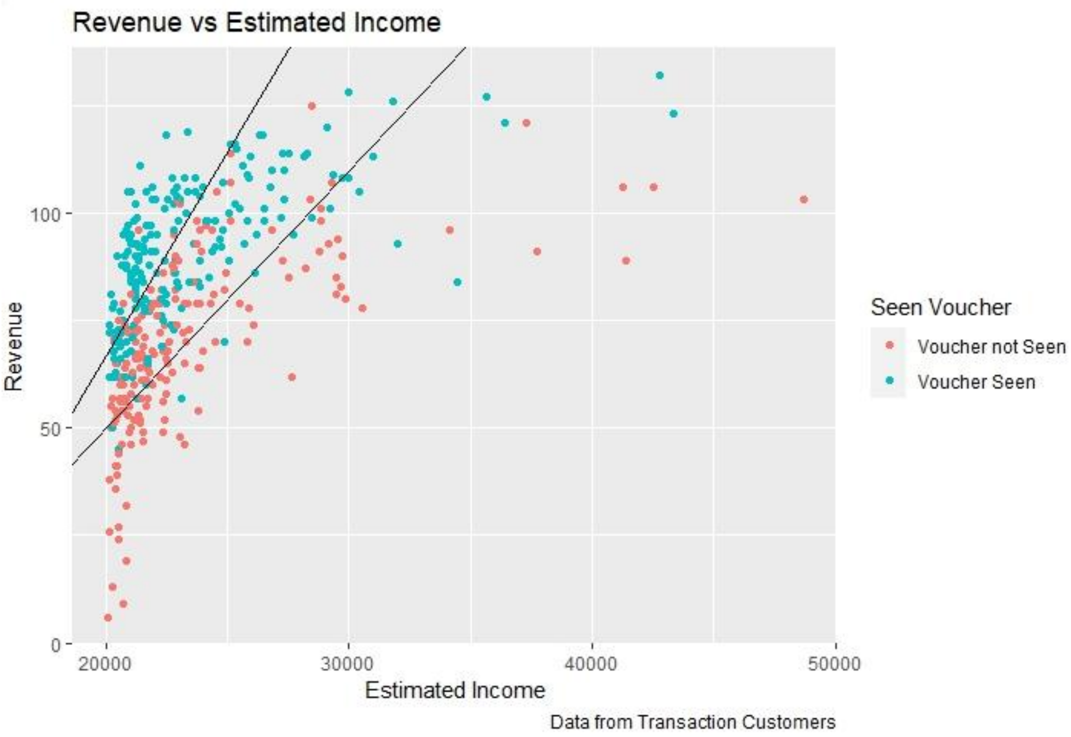


There is a variation in the transactions done by customer related to estimated income.

By adding seen voucher in the scatterplot, we can attain a clear visualization.



The revenue seems to follow a **gradual growth** trend with increase in estimated income.



Revenue shows a significant linear correlation with Estimated Income.

5.3 Summary

Assumption 1 is clearly proven in the scatterplots. The available data is unique and measured accurately as there are no missing values and is scattered evenly.

Assumption 2 is also evident here because a few explanatory variables are sharing a linear relationship with revenue. e.g. Revenue vs Estimated Income.

Assumption 3 largely holds true as the error is quite independent from the error of other points.

6.Regression Analysis:

As all assumptions are satisfied so we can try to fit model

- Nullify the Leaflet column as it **acts as the base** for the others Advertisement Channels.
- Nullify the Advertisement Channel column as we would be checking the impact of every advertising channel over revenue.

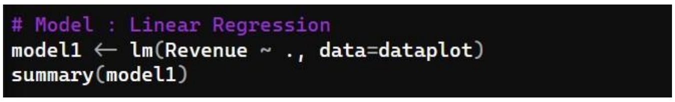
6.1 Multiple Linear Regression:

A statistical technique used to describe relationships between variables by fitting a line to the observed data. It allows you to estimate how a dependent variable change as the independent variable(s) change. (Anscombe, F. J. ,1973)

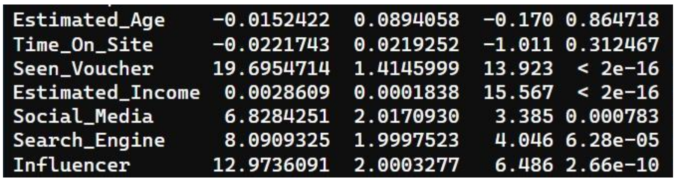
$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

The key is understanding the overall capability of the model as well as understanding how each variable is statistically related to the dependent variable.

6.2 Model 1: Revenue against all explanatory variables.



```
# Model : Linear Regression
model1 <- lm(Revenue ~ ., data=dataplot)
summary(model1)
```



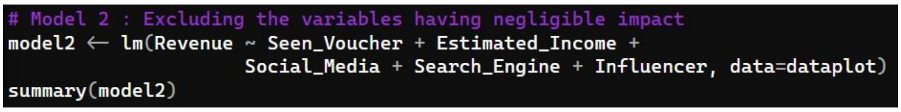
```
Estimated_Age    -0.0152422  0.0894058  -0.170 0.864718
Time_On_Site     -0.0221743  0.0219252  -1.011 0.312467
Seen_Voucher     19.6954714  1.4145999  13.923  < 2e-16
Estimated_Income  0.0028609  0.0001838  15.567  < 2e-16
Social_Media      6.8284251  2.0170930   3.385 0.000783
Search_Engine     8.0909325  1.9997523   4.046 6.28e-05
Influencer       12.9736091  2.0003277   6.486 2.66e-10
```

Model 1 Evaluation:

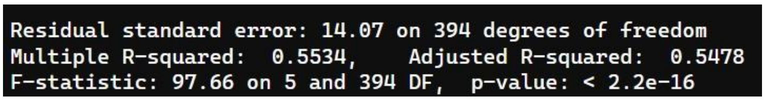
- Expected value for p should be under **0.05** and these variables have **p value > 0.05**. (H. M. James Hung et al 1997)
- As estimated age and time on site has **negligible** effect on revenue.
- So, we can eliminate these and introduce the other explanatory variables to explain the dependent variable in the new model.

By doing this, we make our model more accurate.

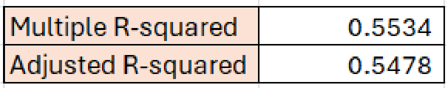
6.3 Model 2:



```
# Model 2 : Excluding the variables having negligible impact
model2 <- lm(Revenue ~ Seen_Voucher + Estimated_Income +
                    Social_Media + Search_Engine + Influencer, data=dataplot)
summary(model2)
```



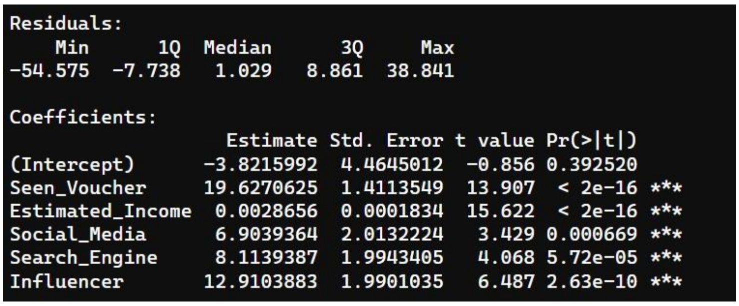
```
Residual standard error: 14.07 on 394 degrees of freedom
Multiple R-squared:  0.5534,    Adjusted R-squared:  0.5478
F-statistic: 97.66 on 5 and 394 DF,  p-value: < 2.2e-16
```



Multiple R-squared	0.5534
Adjusted R-squared	0.5478

As can be seen, both are approximately at 0.55. This mean that this model can explain 55% of variation within Revenue.

6.4 Model 2 Evaluation & Summary of the Model:



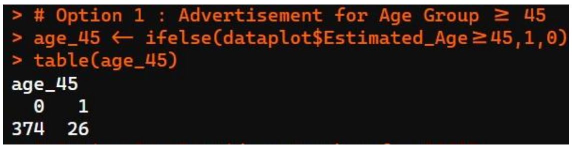
```
Residuals:
    Min      1Q  Median      3Q     Max 
-54.575  -7.738   1.029   8.861  38.841 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -3.8215992  4.4645012  -0.856 0.392520    
Seen_Voucher     19.6270625  1.4113549  13.907  < 2e-16 ***
Estimated_Income  0.0028656  0.0001834  15.622  < 2e-16 ***
Social_Media      6.9039364  2.0132224   3.429 0.000669 ***
Search_Engine     8.1139387  1.9943405   4.068 5.72e-05 ***
Influencer       12.9103883  1.9901035   6.487 2.63e-10 ***
```

- **Seen Voucher is statistically significant** related to revenue. A customer seeing voucher would result in an increase **19.627 GBP** in revenue on average with std error of **1.41**.
- **Influencer** also produces a **significant impact on revenue**. Customers coming in through influencer are predicted to bring an increase of ~**13 GBP** in revenue with a std error of 1.99.
- **Social Media and Search engine** would also be contributing to revenue with an amount of **~7 GBP** and **8.1 GBP** with every customer coming in through them respectively.
- Estimated income has **no quantifiable impact** on Revenue.

7 Result and Recommendation:

Here our aim is to increase the profitability of company.

- **Option 1: Run an advertisement for >45 years old as they are likely to spend more money.**

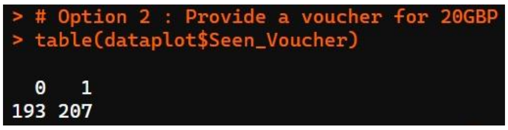


```
> # Option 1 : Advertisement for Age Group ≥ 45
> age_45 ← ifelse(dataplot$Estimated_Age≥45,1,0)
> table(age_45)
age_45
  0   1 
374  26 
```

  - Ratio of consumer aged over 45 is “26:400”.
  - Also, Consumers aged more than 45 years show similar Average Revenue i.e. 79.13 GBP generated by the whole sample i.e. 80 GBP.

  So, probability of creating any impact on revenue by spending money on marketing for people aged over 45 years are **statistically very low.**

- **Option 2: Provide a voucher for 20 GBP off their next orders.**

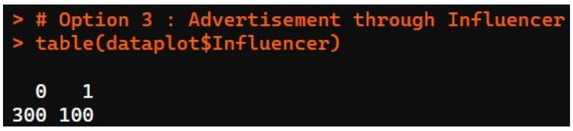


```
> # Option 2 : Provide a voucher for 20GBP
> table(dataplot$Seen_Voucher)

  0   1 
193 207 
```

  - Voucher engagement accounts to 50 % of the total observations.
  - It would have created a big impact as, by model prediction Seen Voucher generates the most revenue per customer i.e. **19.627 GBP.** But providing a coupon of **20 GBP** would put company in slight average loss of **0.373 GBP per customer.**

  Overall, it would not be helpful for profitability.

- **Option 3: Spend more money on advertising with an influencer.**



```
> # Option 3 : Advertisement through Influencer
> table(dataplot$Influencer)

  0   1 
300 100 
```

  - Variable with **second highest** predicted increase in revenue is Advertising through Influencer i.e. 13 GBP
  - Out of 400 people only 100 people came through this advertising medium.
  - Spending more money on the same would ideally get us more and new customers.

On every new customer there would be an increase of revenue by 13 GBP.