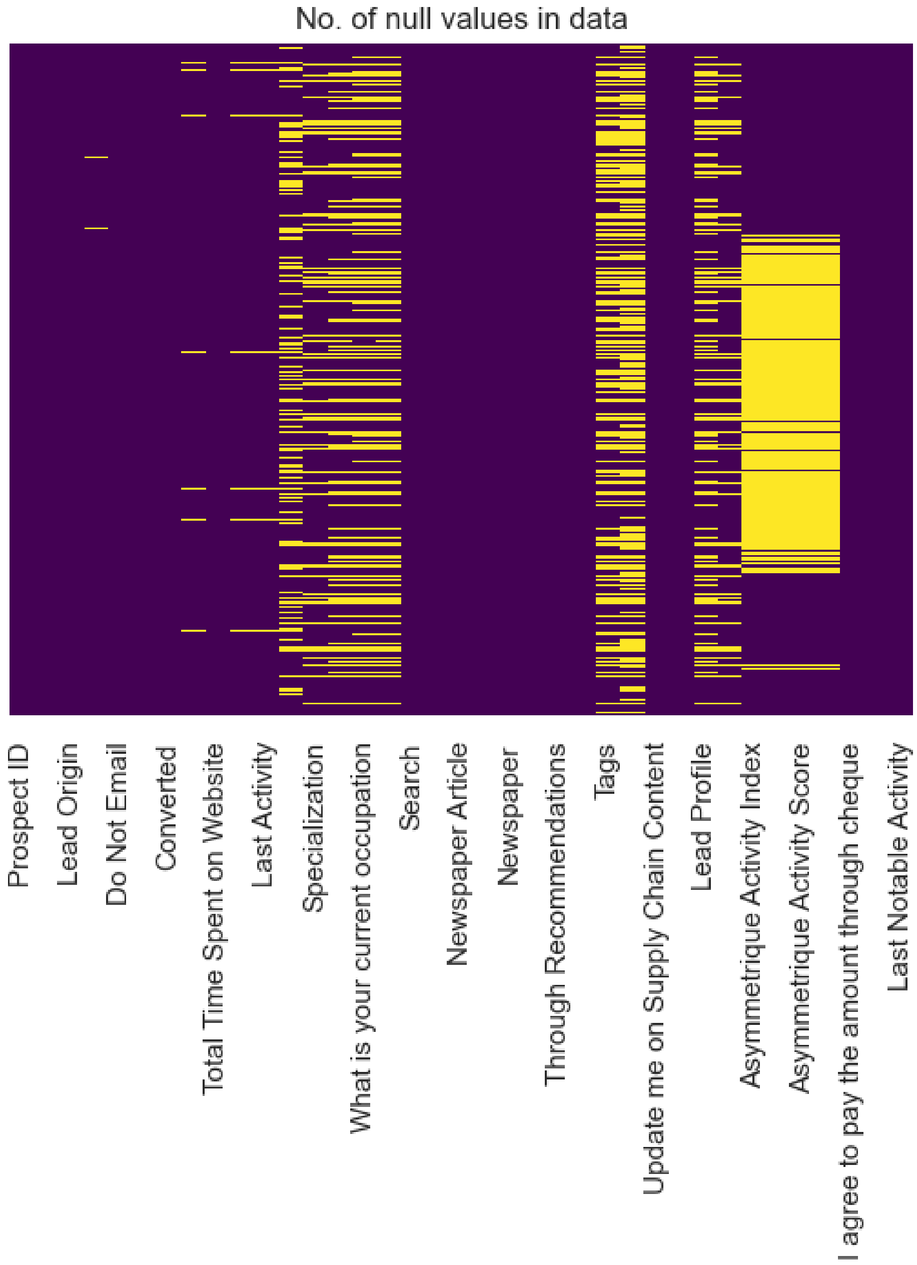
Leads Conversion Using Logistic Regression

Steps to Perform.

1. 2. Basic Information.
2. Handling Missing Values
3. EDA : Exploratory Data Analysis
4. Outlier Detection and Capping.
5. Features Selection Based on correlation and Select K method.
6. Model Selection and Training.

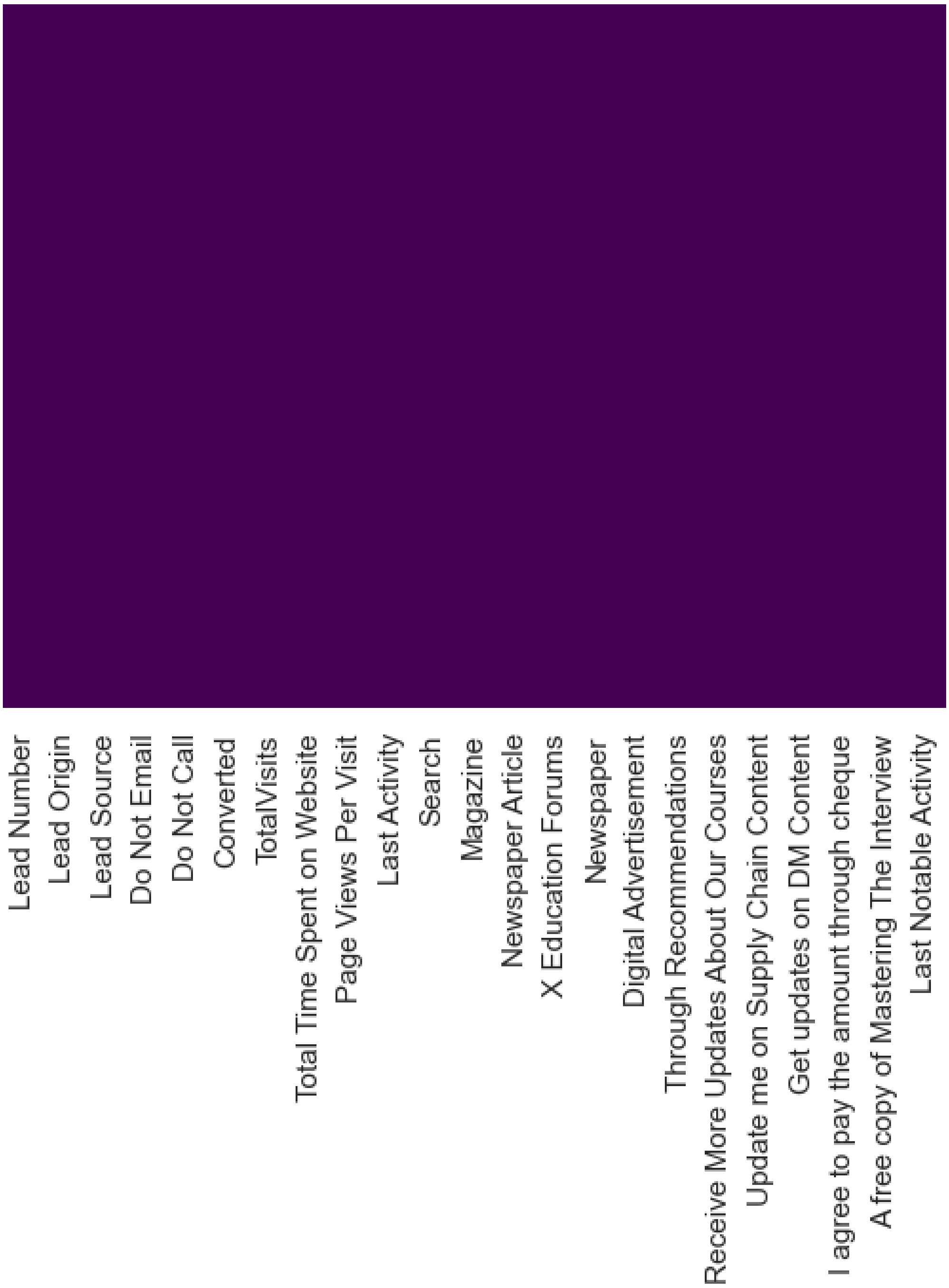
Basic Data Manipulation.



Here Yellow lines show the missing values higher the yello lines, higher the missing values.

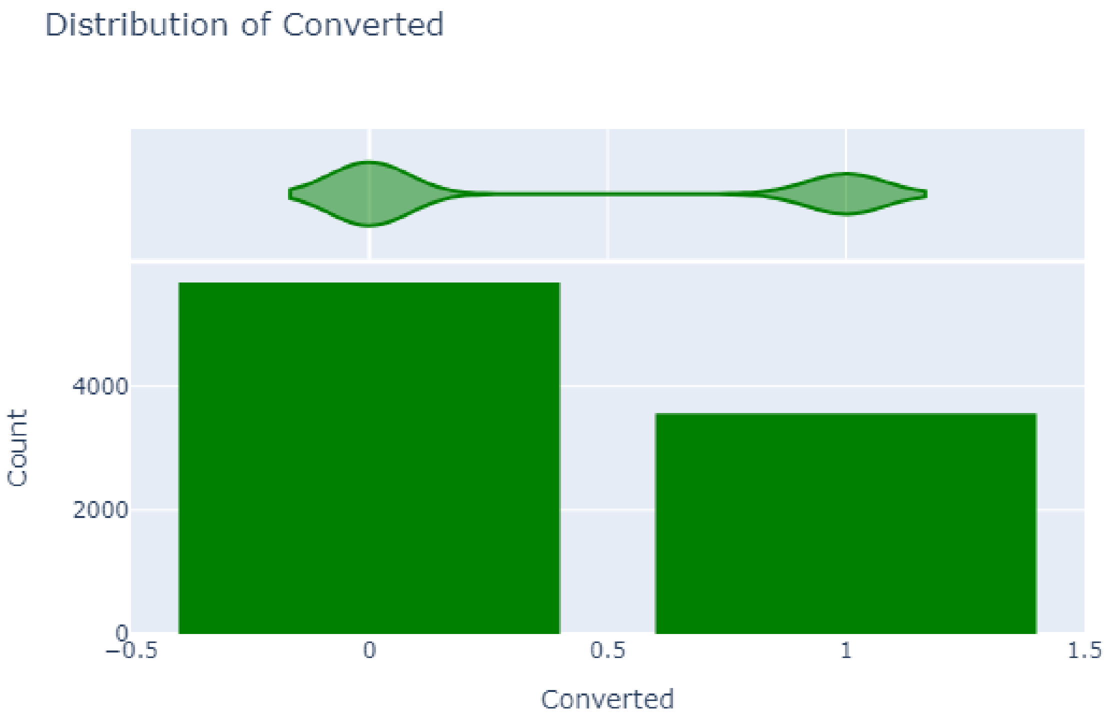
Handling Missing Values

No. of null values in data

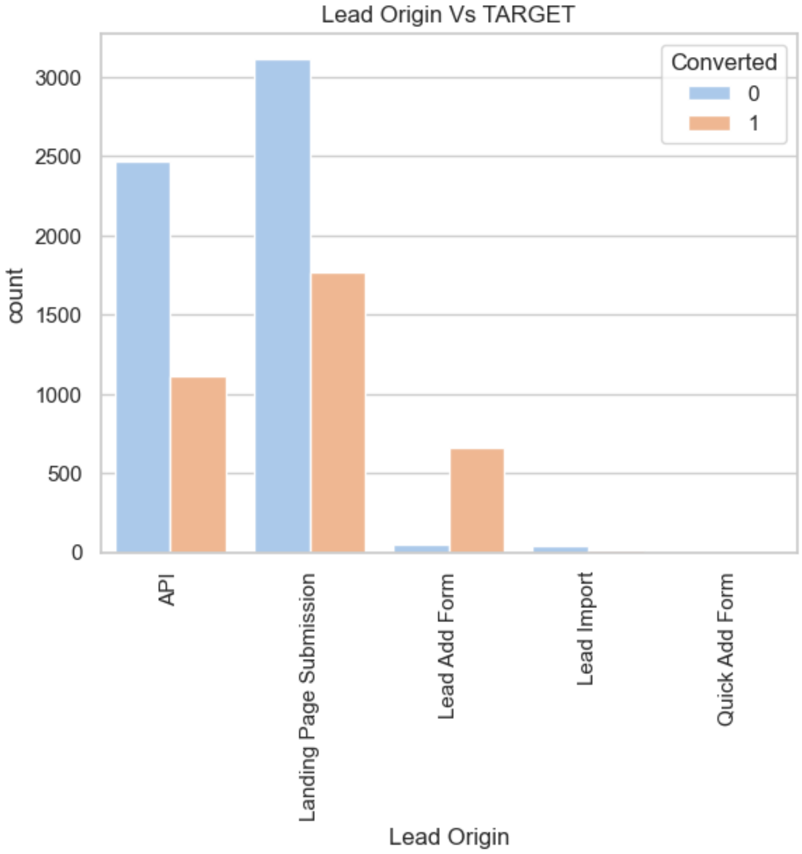


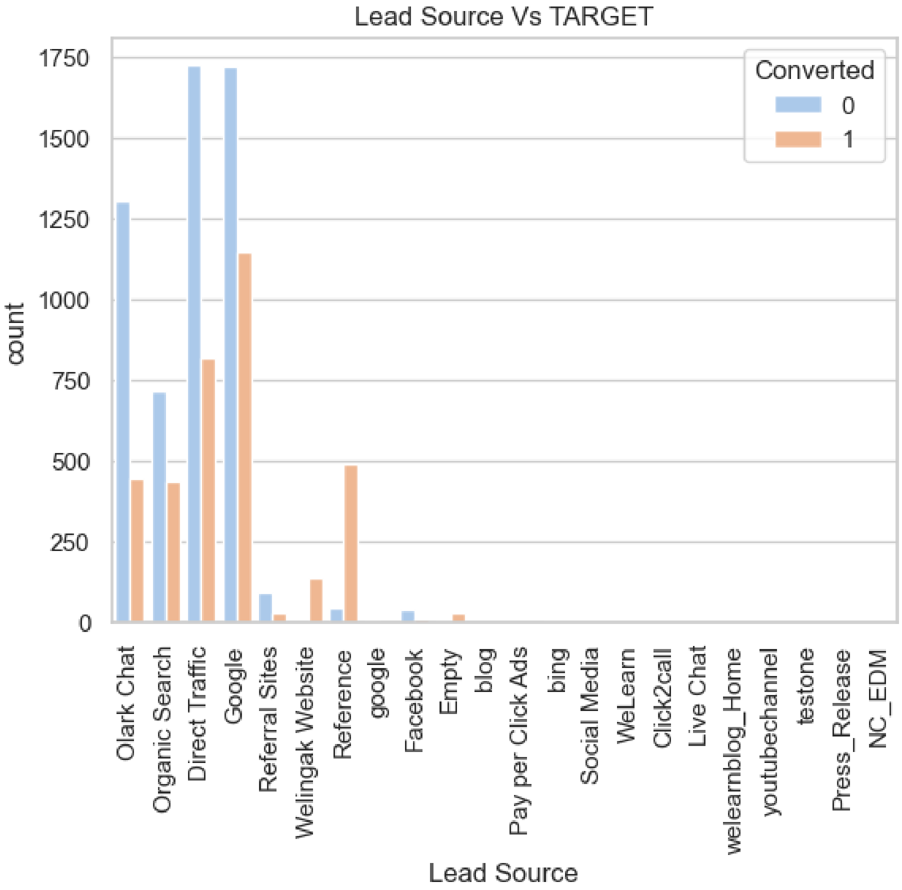
All the null null imputed. for categorical col we replace the values with new category and for numerical col we replaced with mean value.

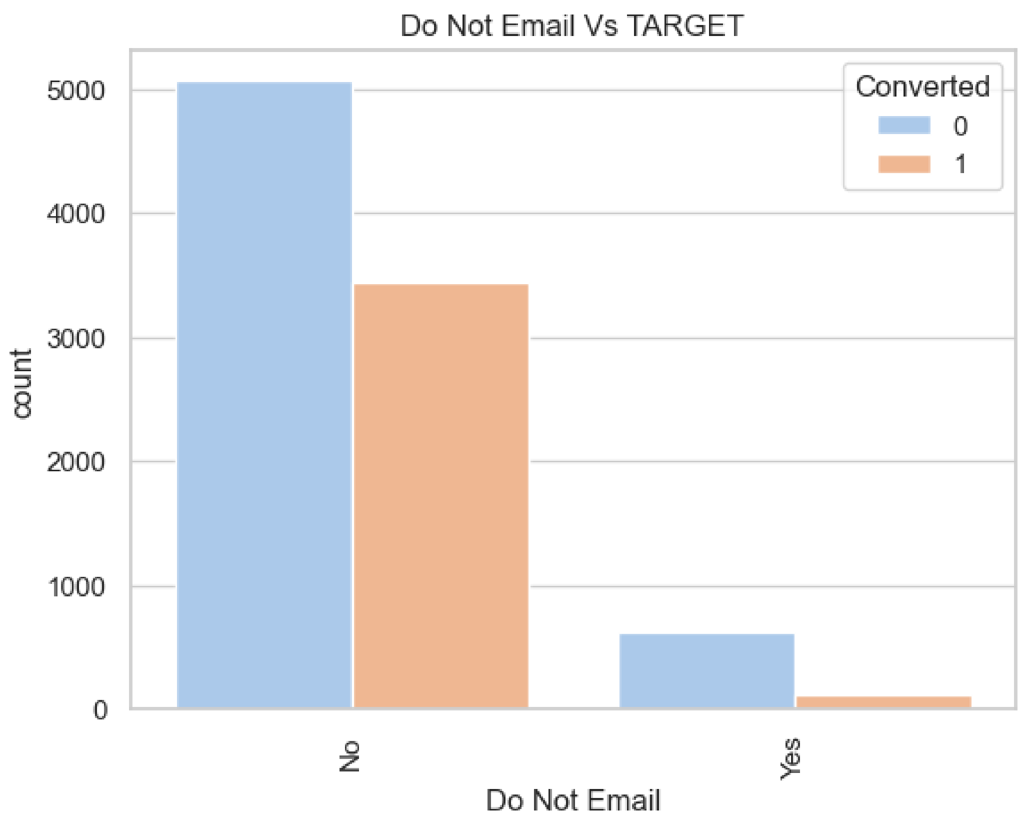
EDA: Exploratory Data Analysis

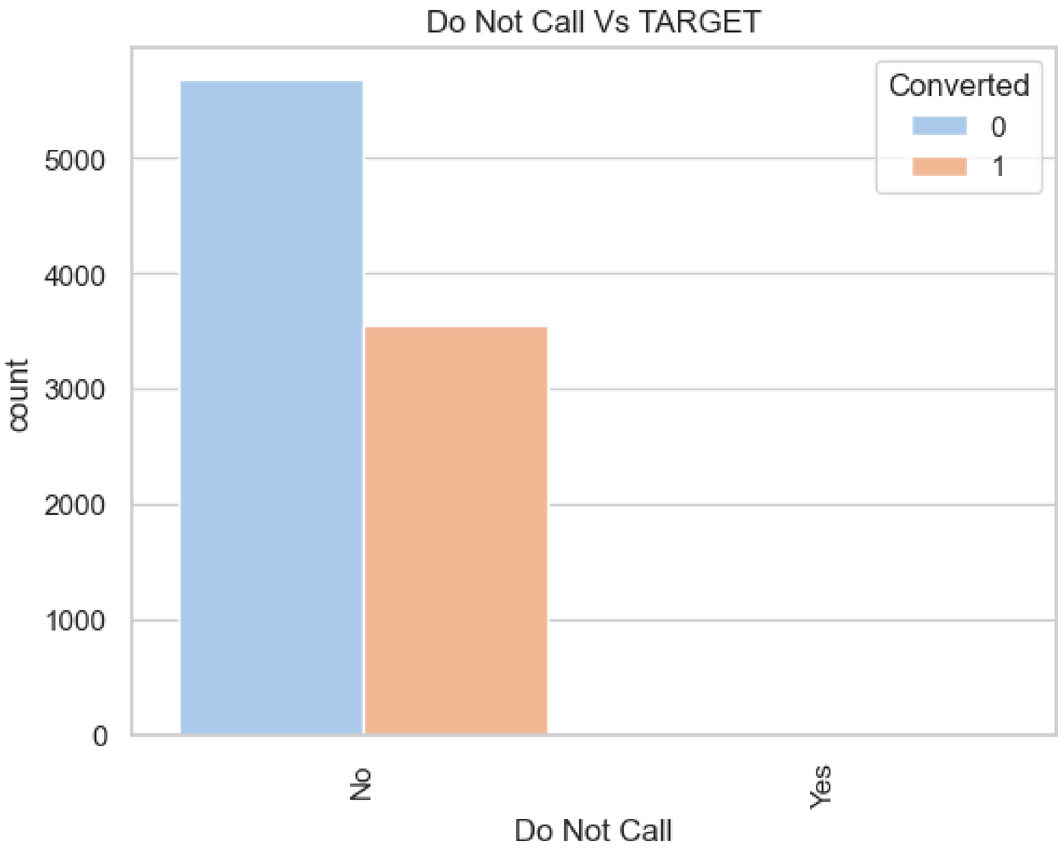


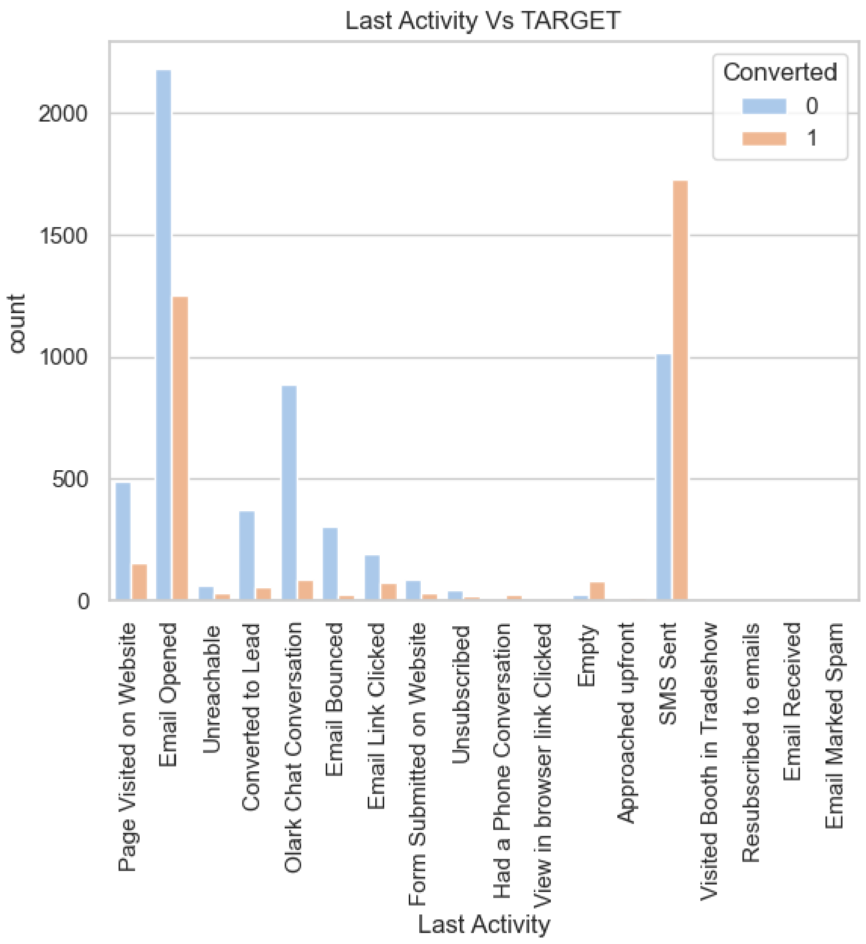
As we earlier predicted that Target column has higher number of 0 values then 1. We have more negative examples in our training then positive.

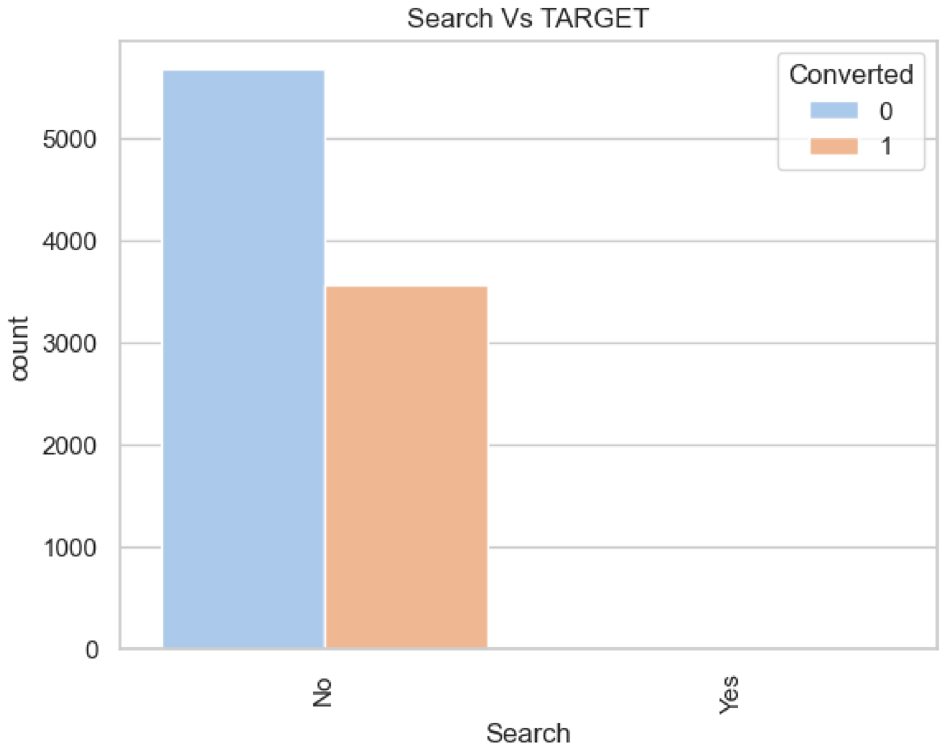

Lead Origin Vs TARGET
count
3000
2500
2000
1500
1000
500
0
Converted
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

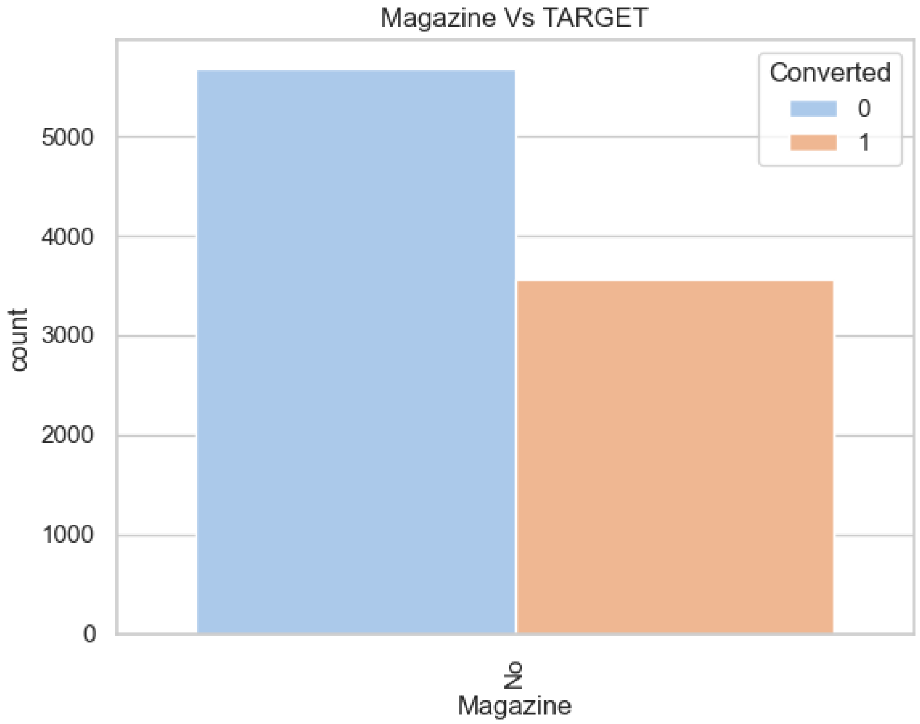

Lead Source Vs TARGET
count
1750
1500
1250
1000
750
500
250
0
Converted
0
1
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
Empty
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

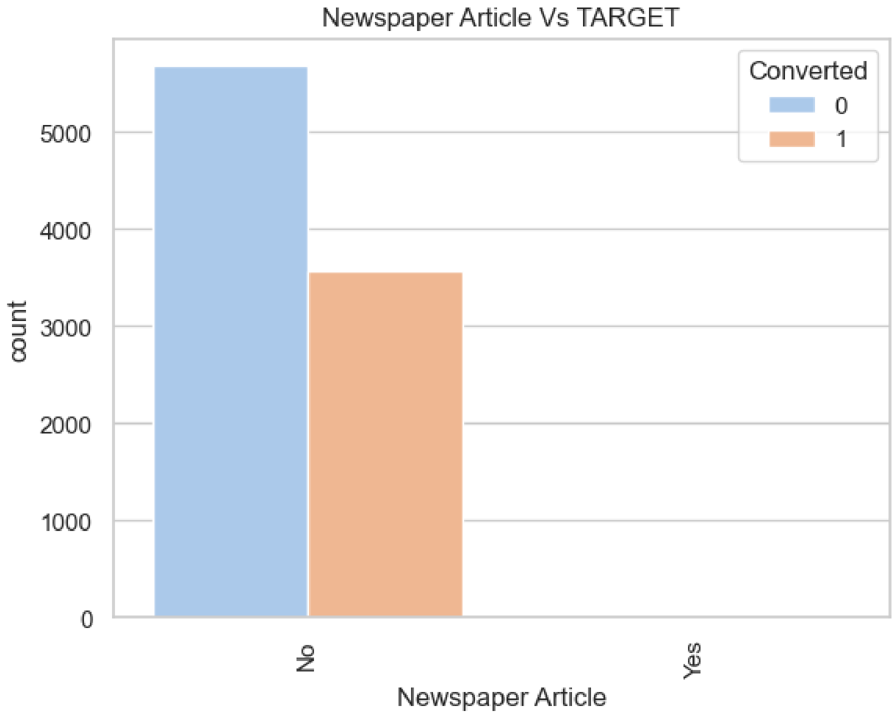

Do Not Email Vs TARGET
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

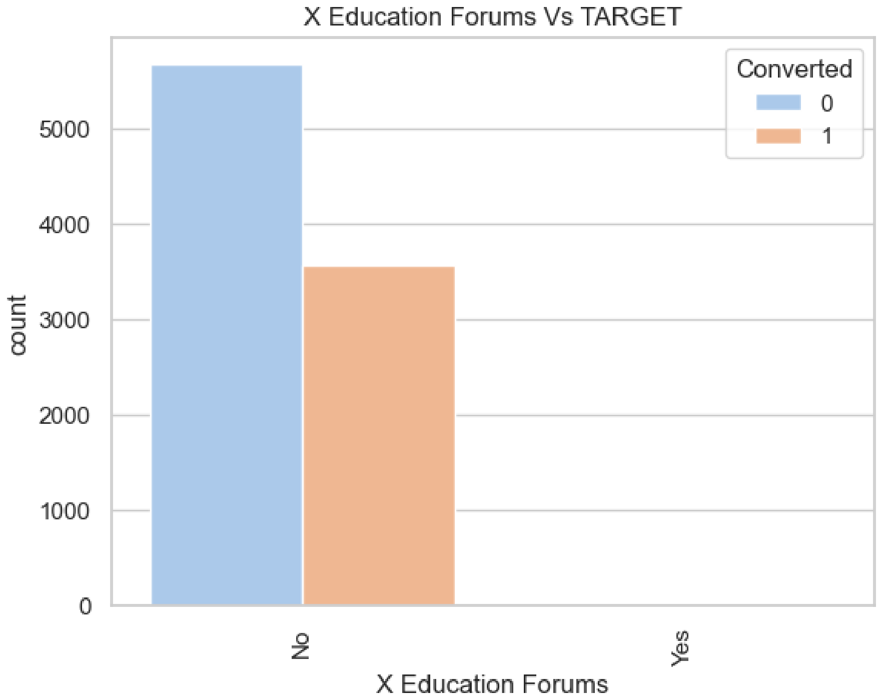

Do Not Call Vs TARGET
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call
Converted
0
1

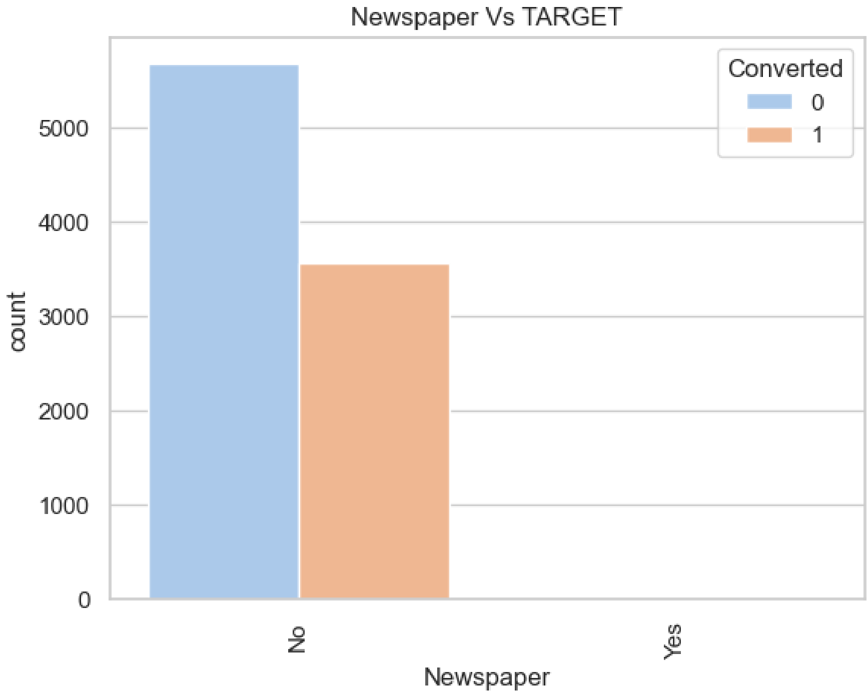

Last Activity Vs TARGET
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Empty
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity
Converted
0
1

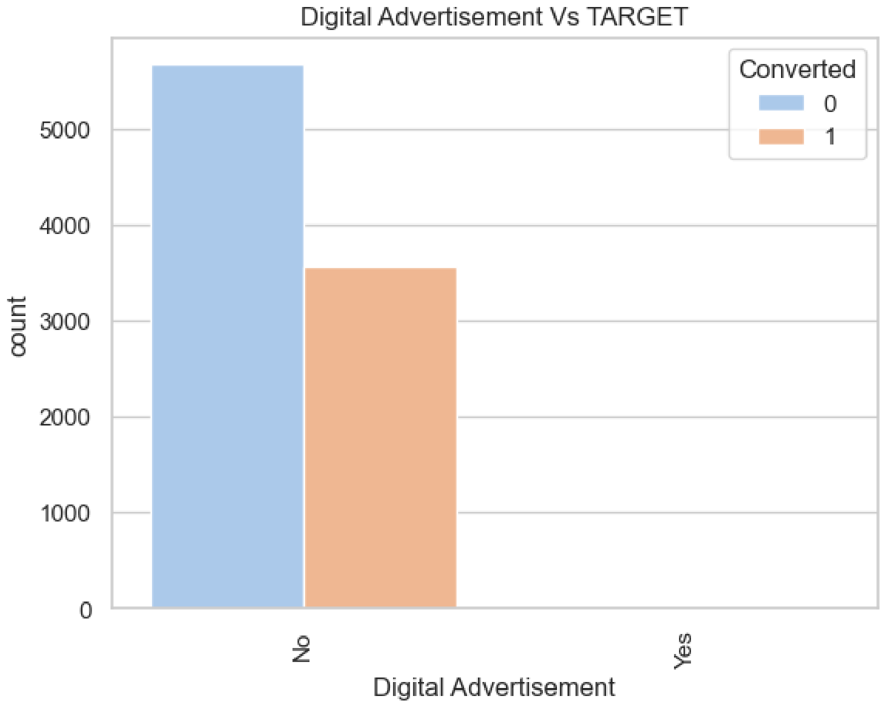

Search Vs TARGET
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Search

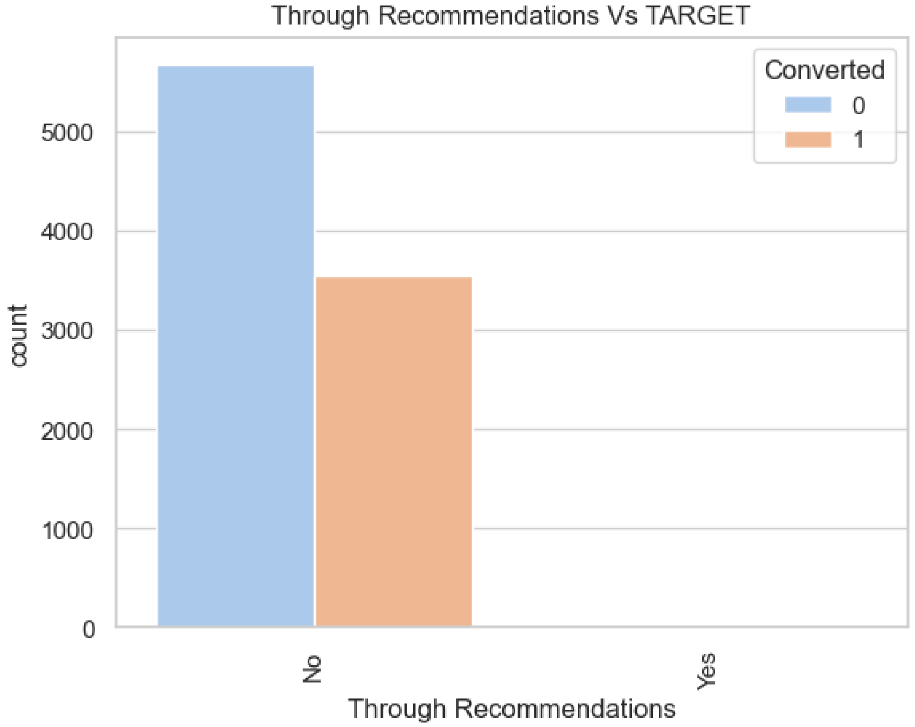

Magazine Vs TARGET
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Magazine

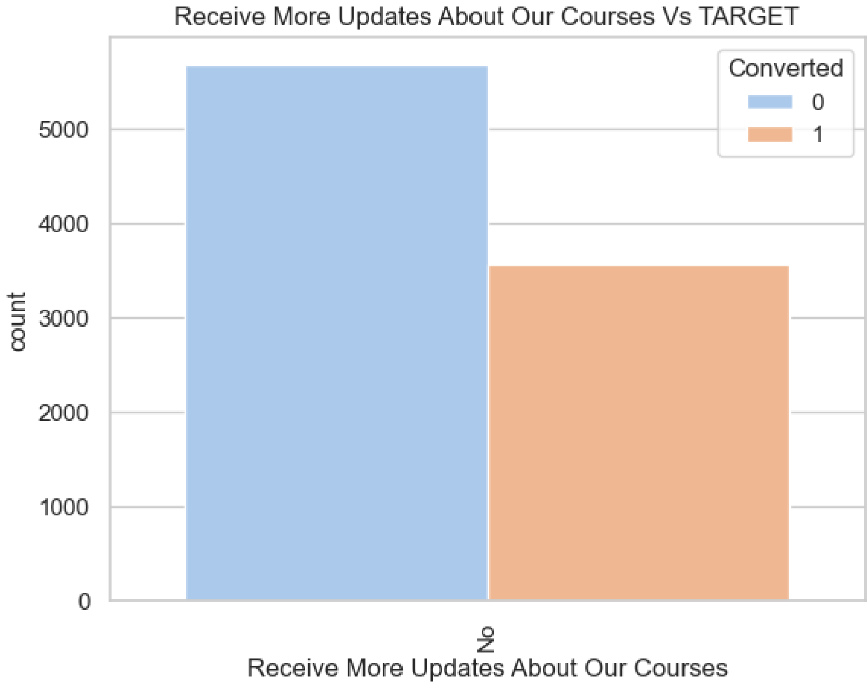

Newspaper Article Vs TARGET
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Newspaper Article

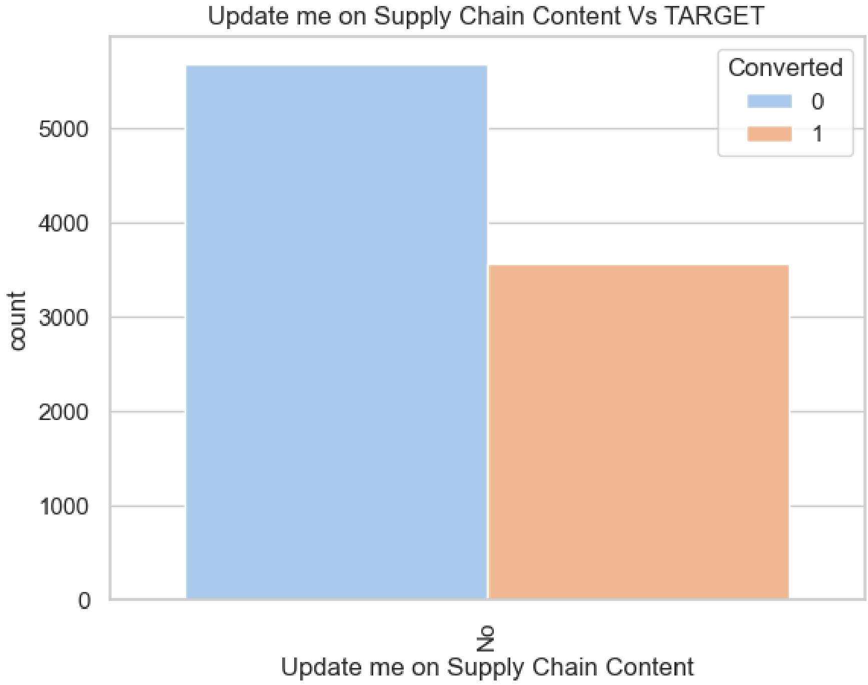

X Education Forums Vs TARGET
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
X Education Forums

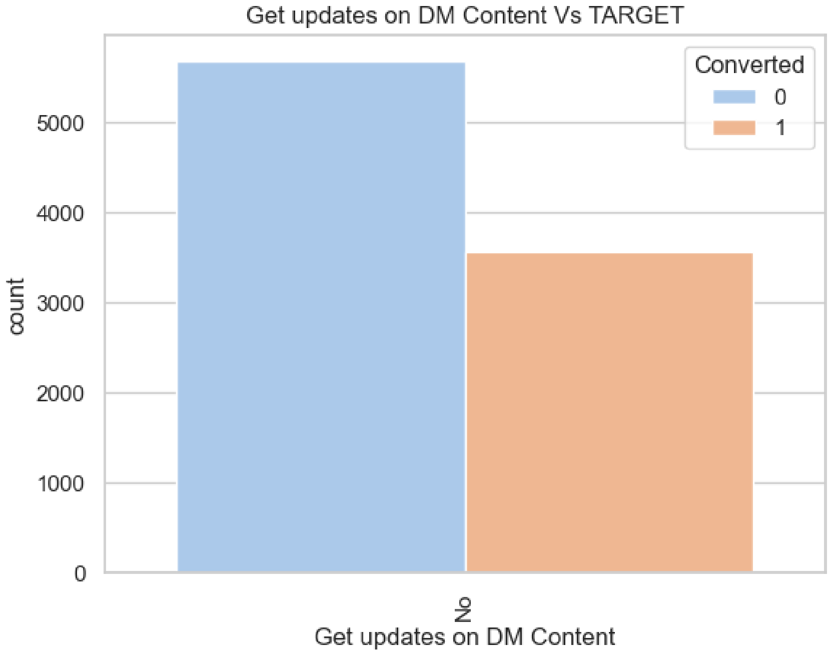

Newspaper Vs TARGET
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Newspaper

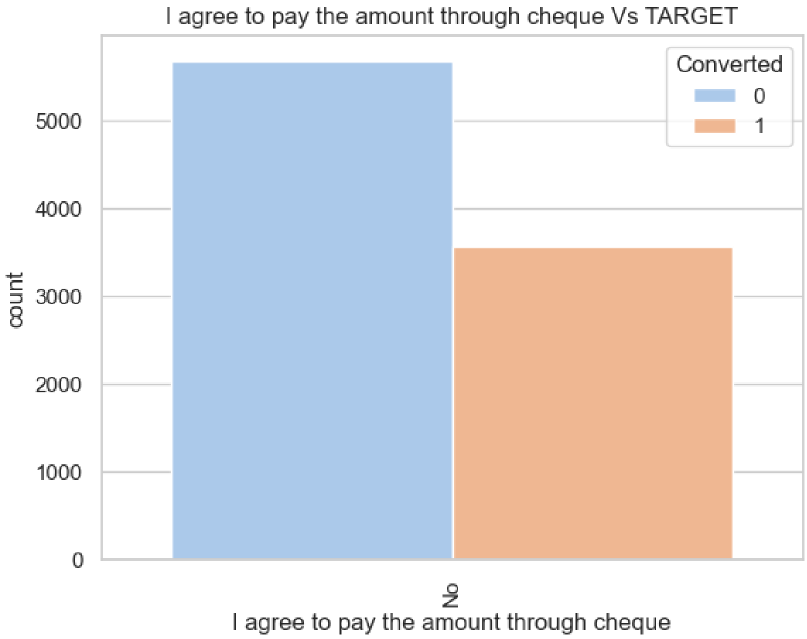

Digital Advertisement Vs TARGET
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Digital Advertisement

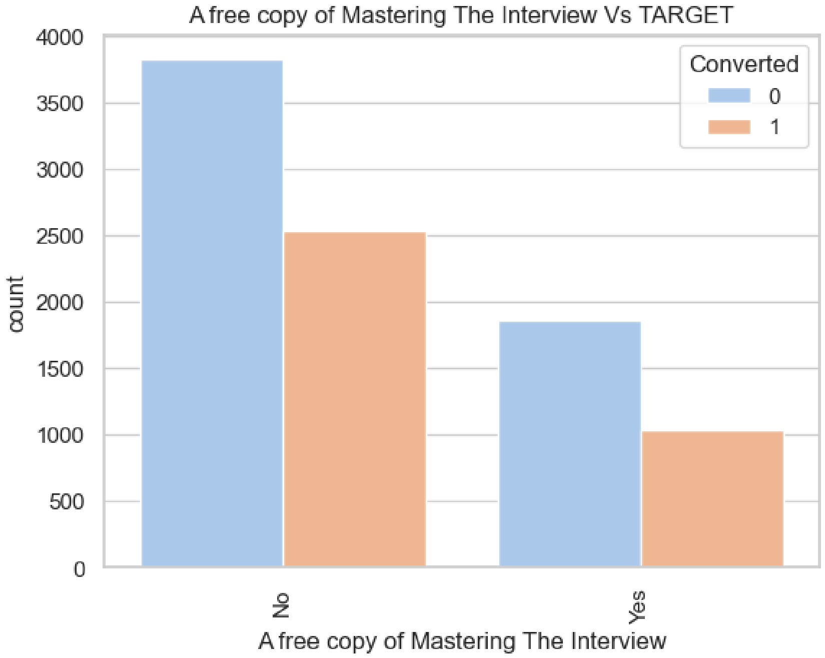

Through Recommendations Vs TARGET
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Through Recommendations

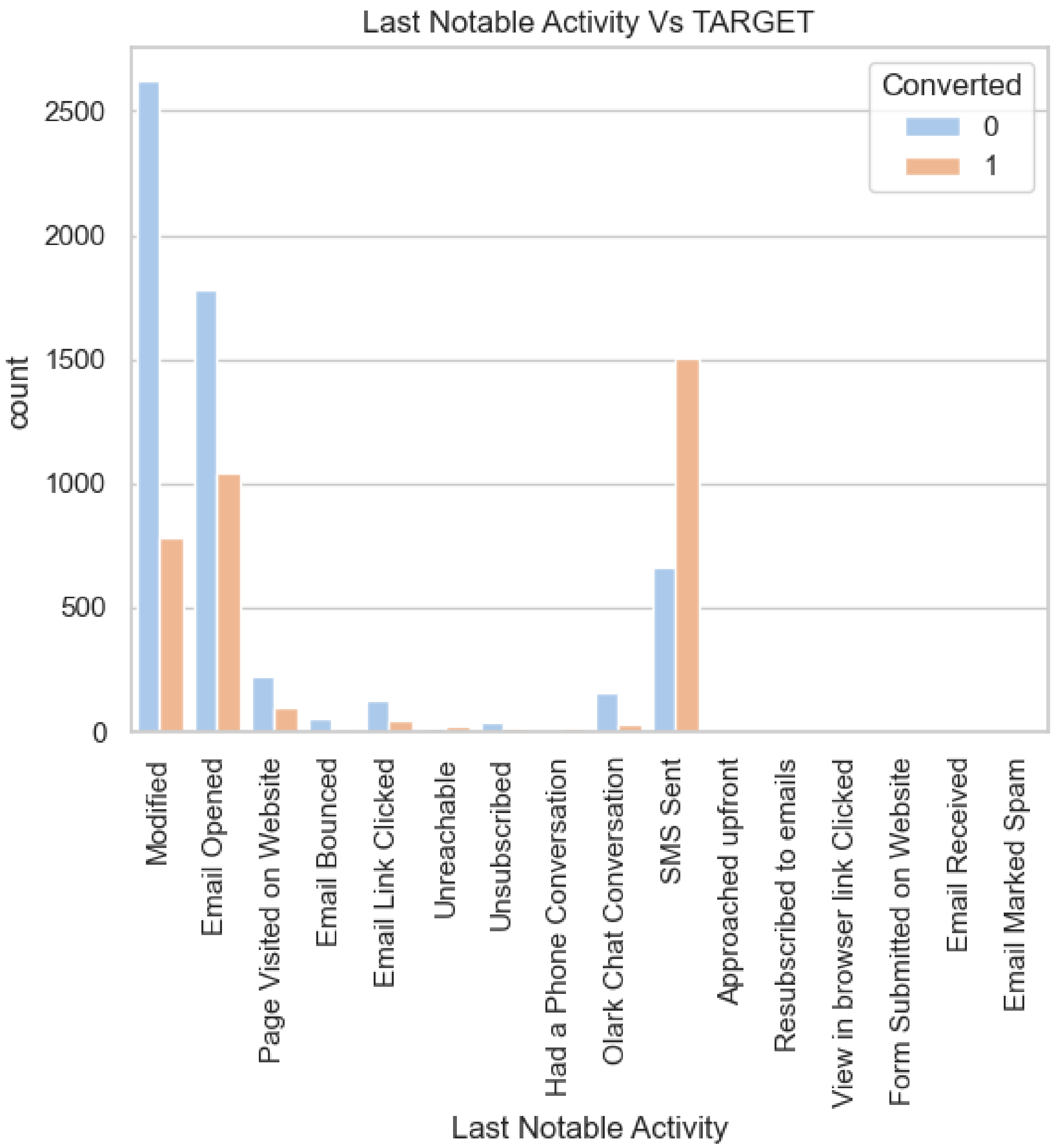

Receive More Updates About Our Courses Vs TARGET
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Receive More Updates About Our Courses


Update me on Supply Chain Content Vs TARGET
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Update me on Supply Chain Content


Get updates on DM Content Vs TARGET
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Get updates on DM Content


I agree to pay the amount through cheque Vs TARGET
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
I agree to pay the amount through cheque


A free copy of Mastering The Interview Vs TARGET
4000
3500
3000
2500
2000
1500
1000
500
0
count
Converted
0
1
No
Yes
A free copy of Mastering The Interview



Observations

1. Here some categories have higher number of value then others in a single column.

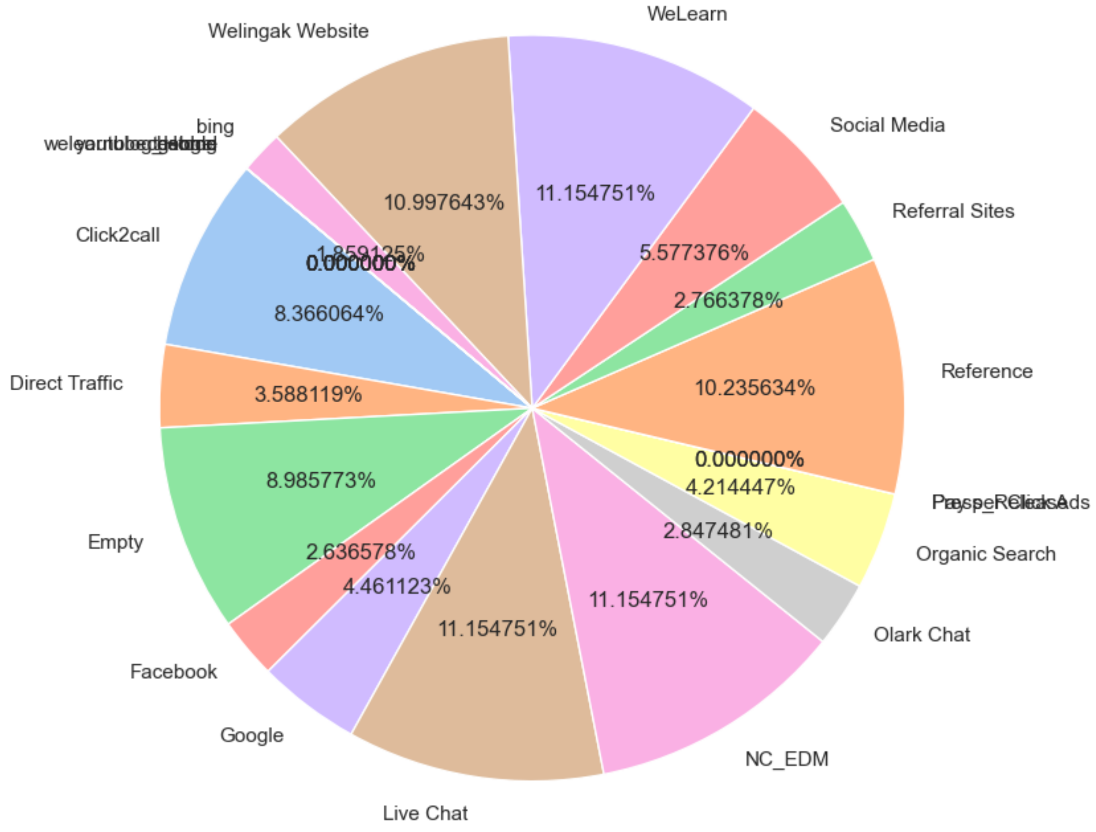
2. higher number categories supports to not conversion and a smaller number of categories support conversion.

3. Like in Last Notable Activity if we sent the SMS then chance of conversion is higher.

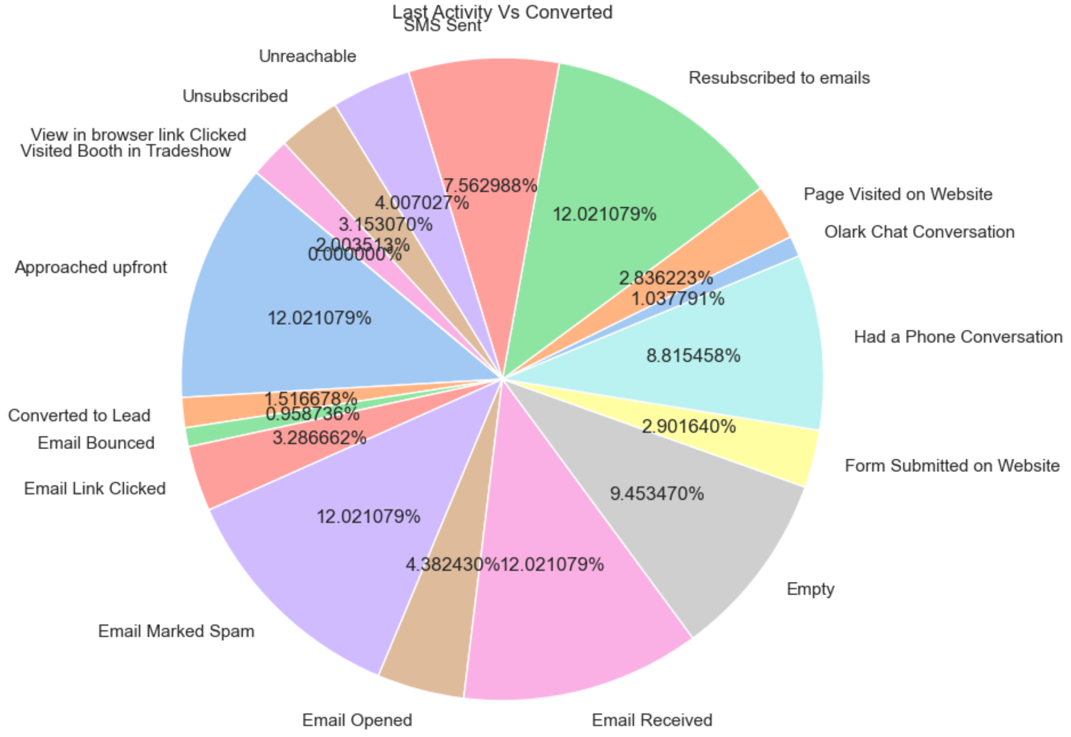
4. Also if a source is references then there is higher chance of conversion of that person.

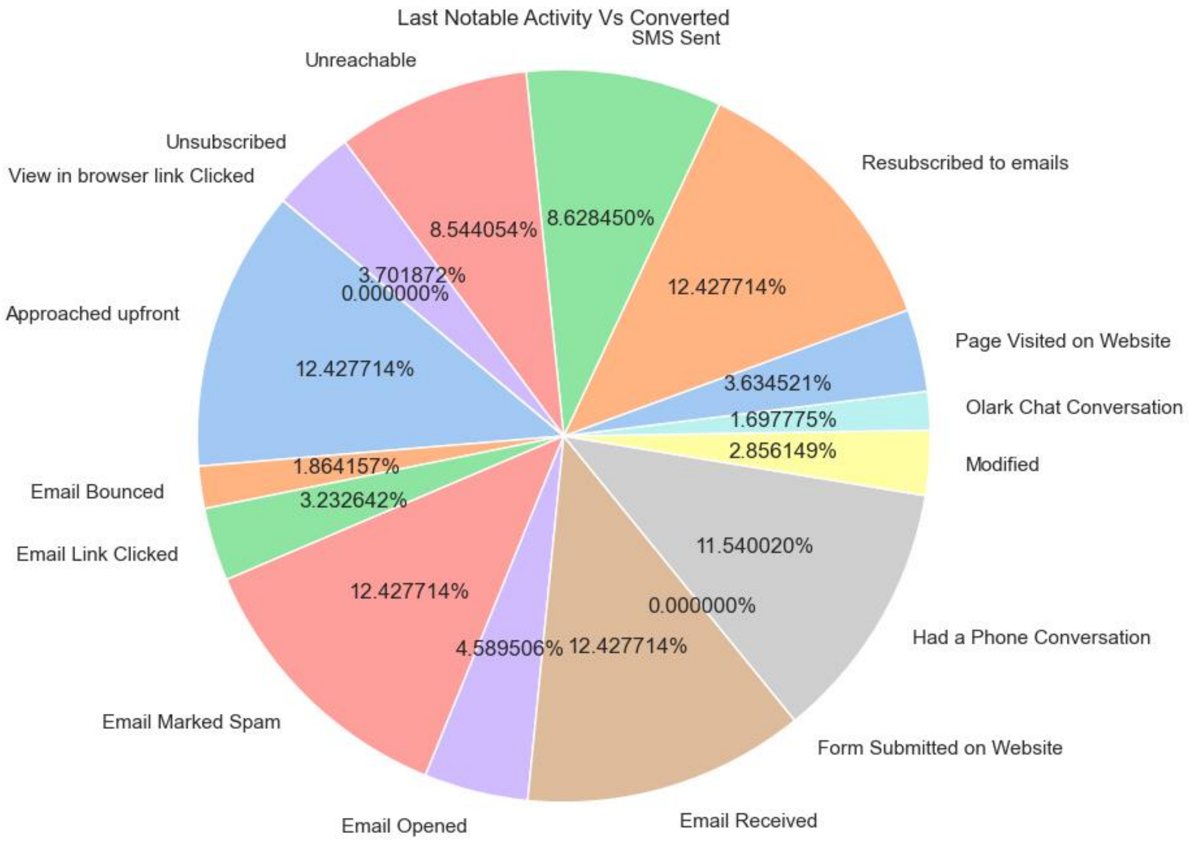
5. Olark chat conversion has higher negative impact on the conversion. So, we need the inchance the bot response.

Lead Source Vs Converted



Last Activity Vs Converted



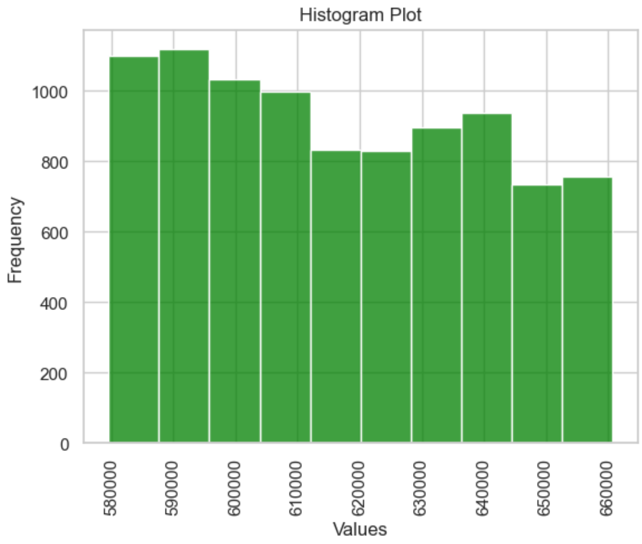


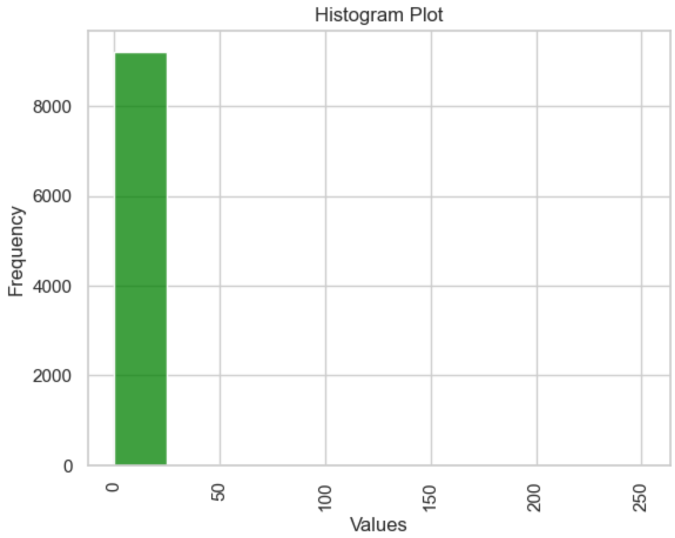
Observation

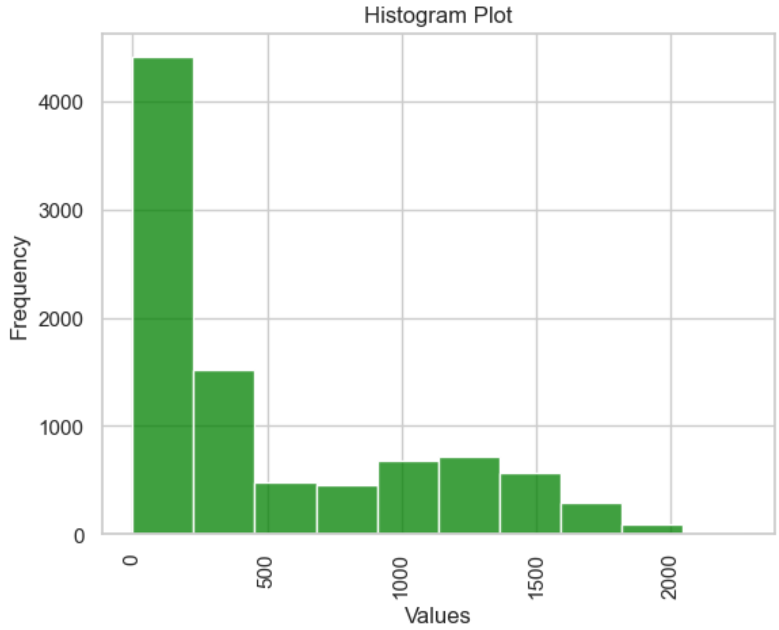
1. As we can see that some the categorical columns contribute 100% to not conversion.

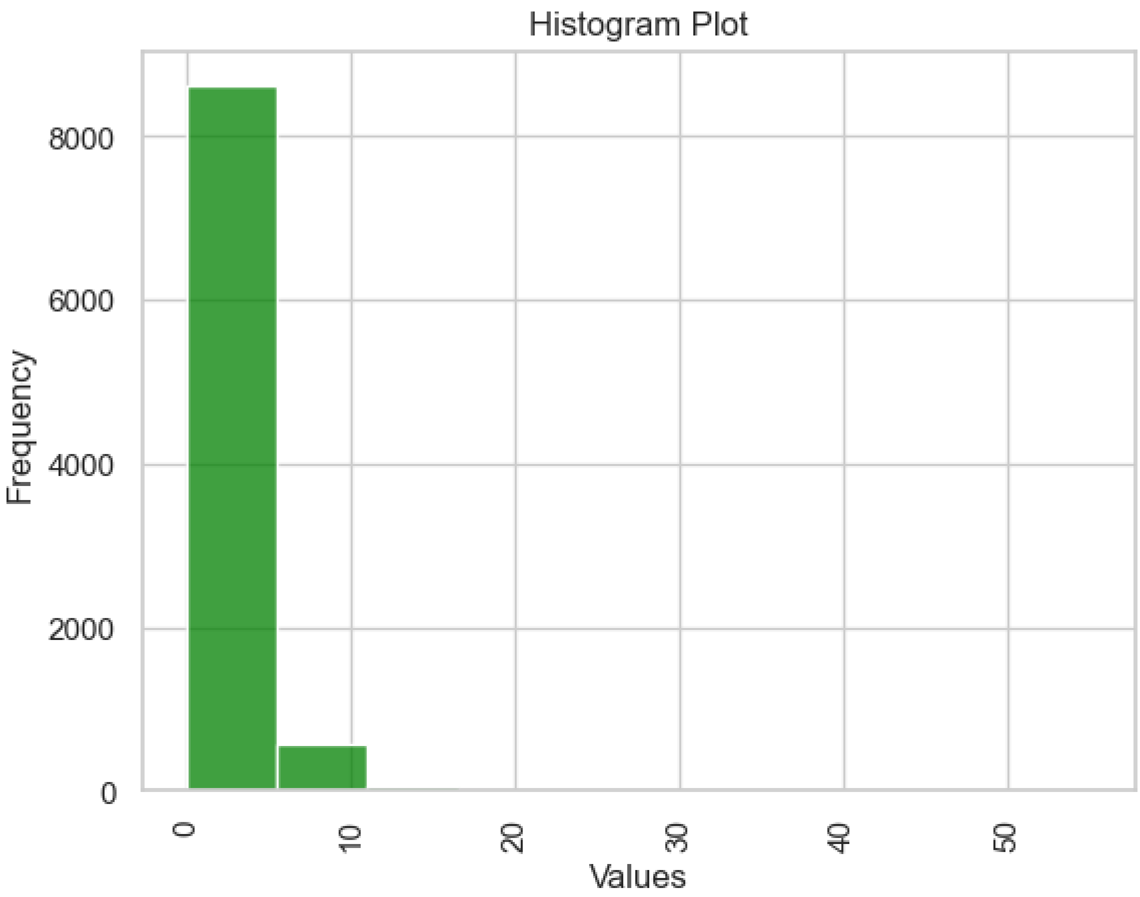
2. Also some of the cat cols like Do not email. Do not call has higher number of negative corelation to conversion.

3. Also references or new paper articles are the effective sources till now.


Histogram Plot
Frequency
0
200
400
600
800
1000
580000
590000
600000
610000
620000
630000
640000
650000
660000
Values

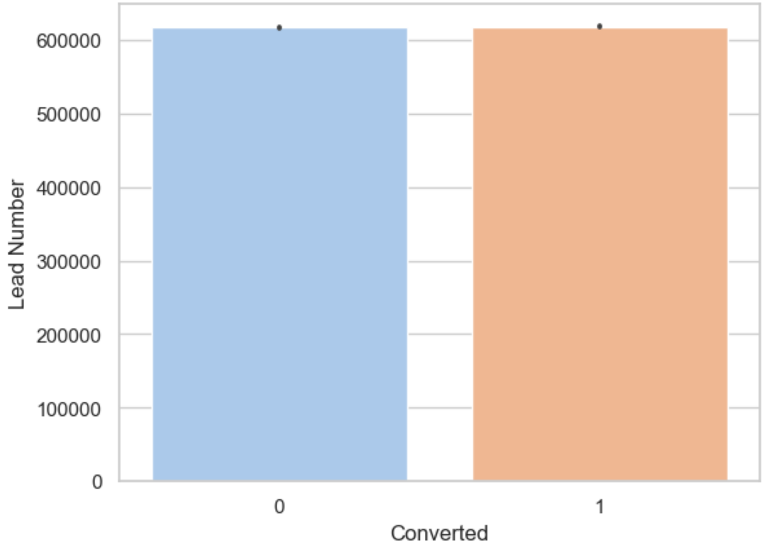

Histogram Plot
Frequency
0
2000
4000
6000
8000
0
50
100
150
200
250
Values


Histogram Plot
Frequency
0
1000
2000
3000
4000
0
500
1000
1500
2000
Values

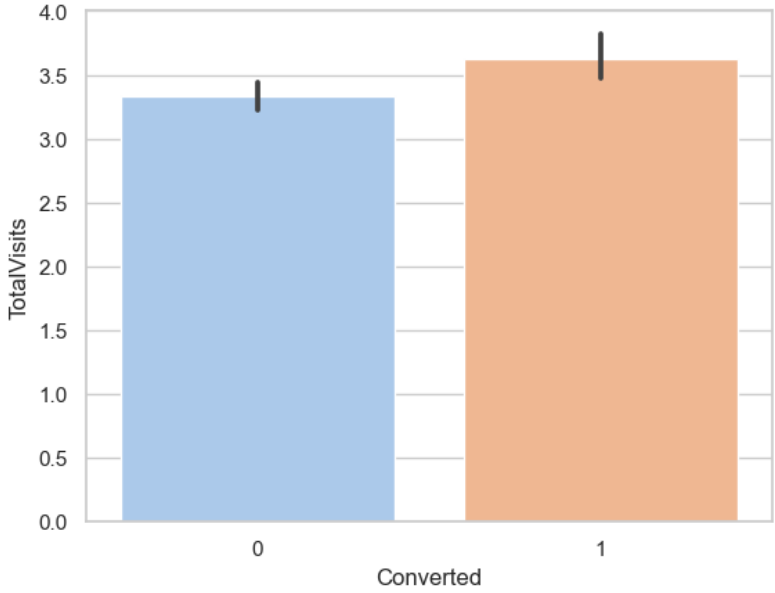


Distribution of these continuous numerical columns are not normalized these are left skewed. Also Some of them have outliers because of these we can see the bar at one side and one single value on other side.

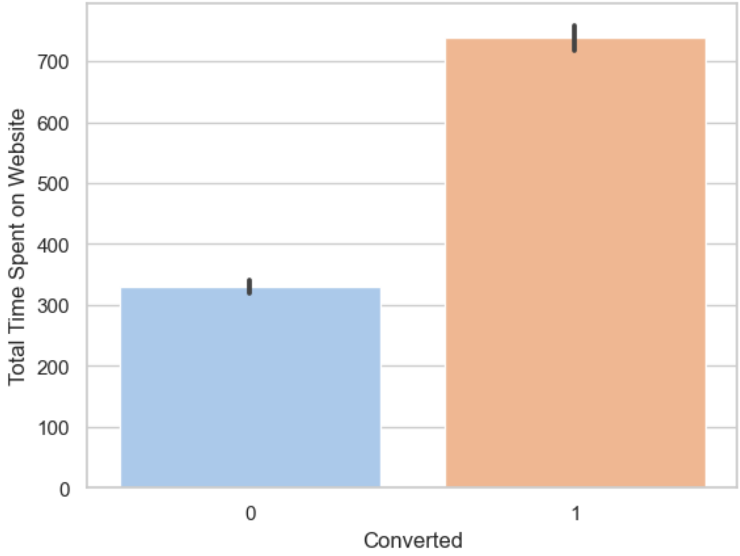
Bar Plot of Lead Number vs Converted

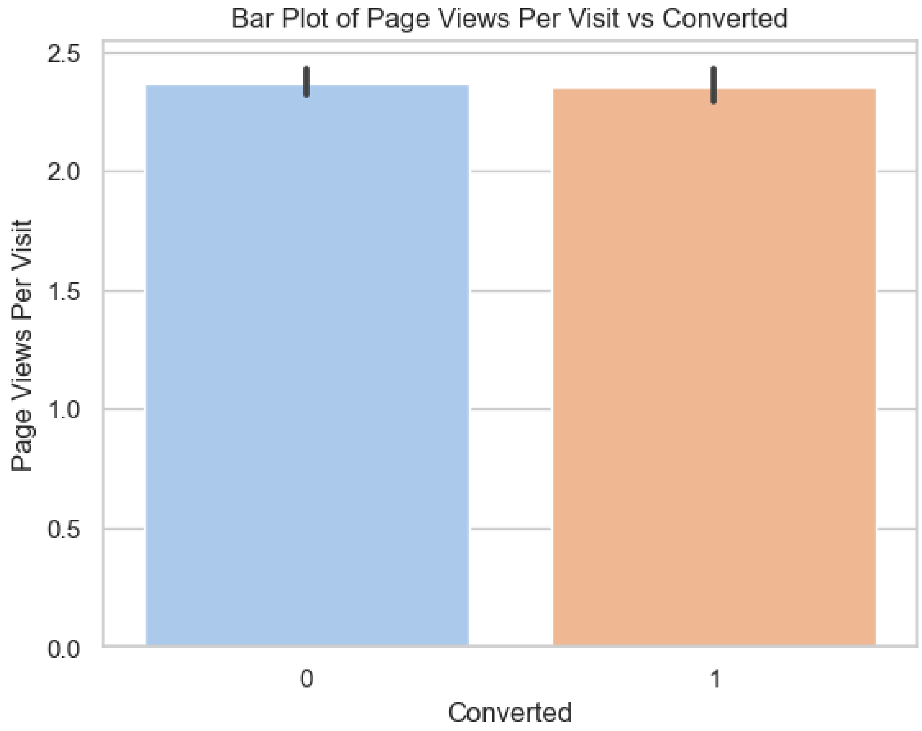


Bar Plot of TotalVisits vs Converted



Bar Plot of Total Time Spent on Website vs Converted



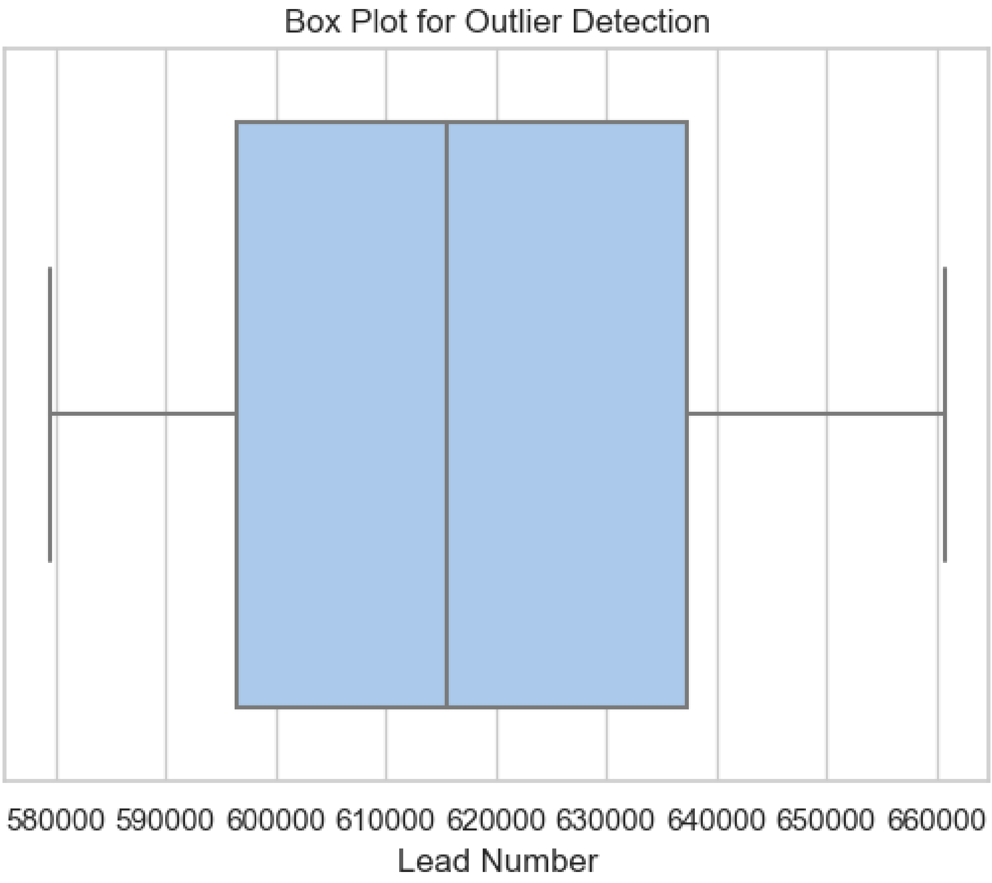


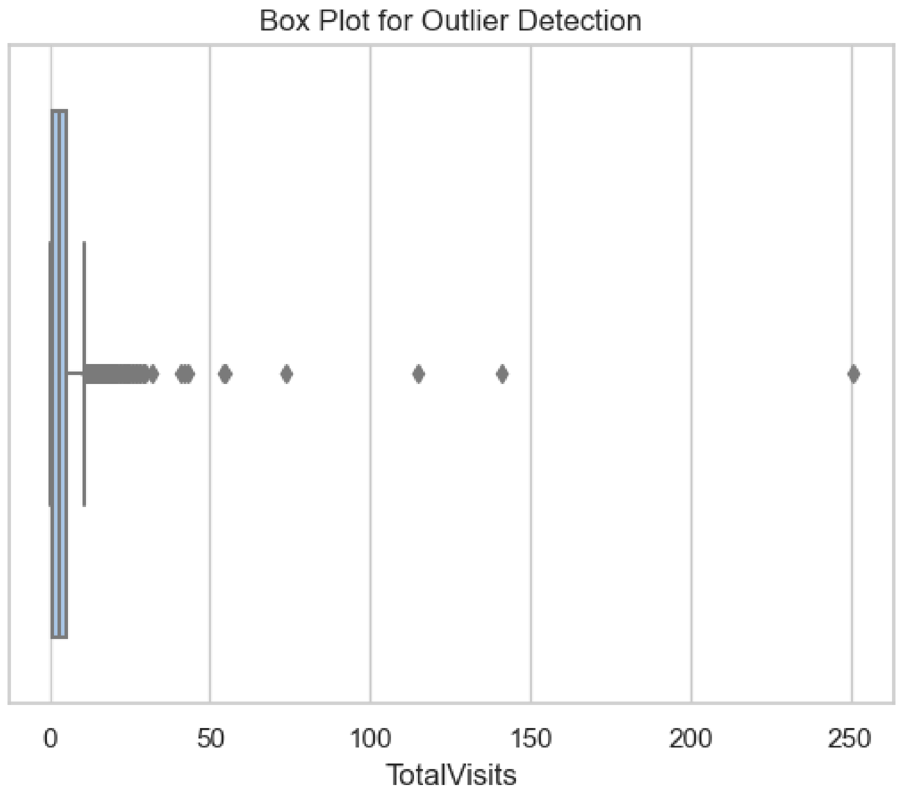
Observations

1. Here we can see that time spent and TotalVisit on website has higher potitive impact on conversion

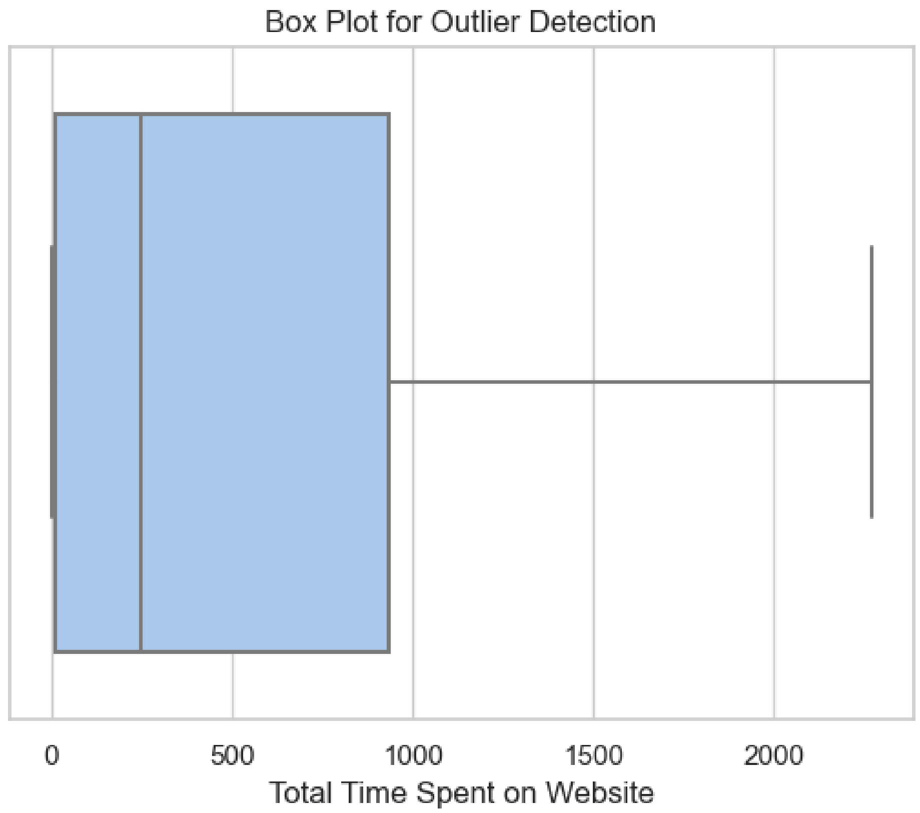
2. And page views has netural impact on the conversion.

Outlier detection

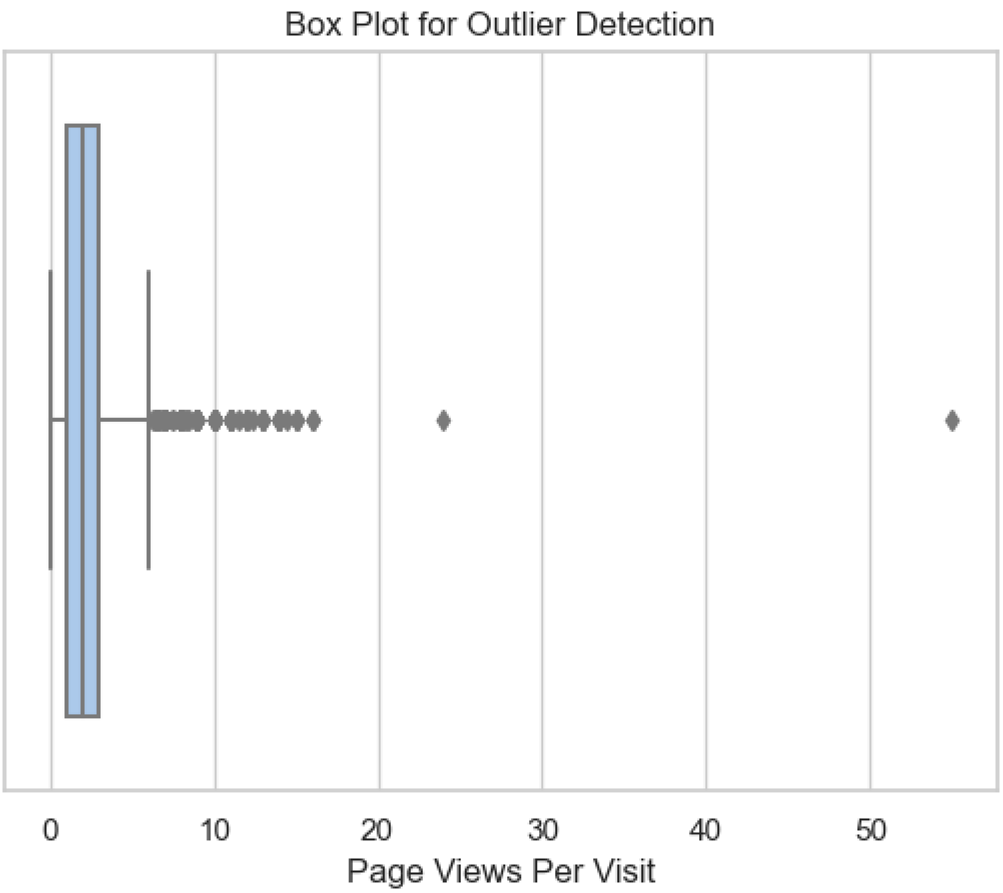




Box Plot for Outlier Detection



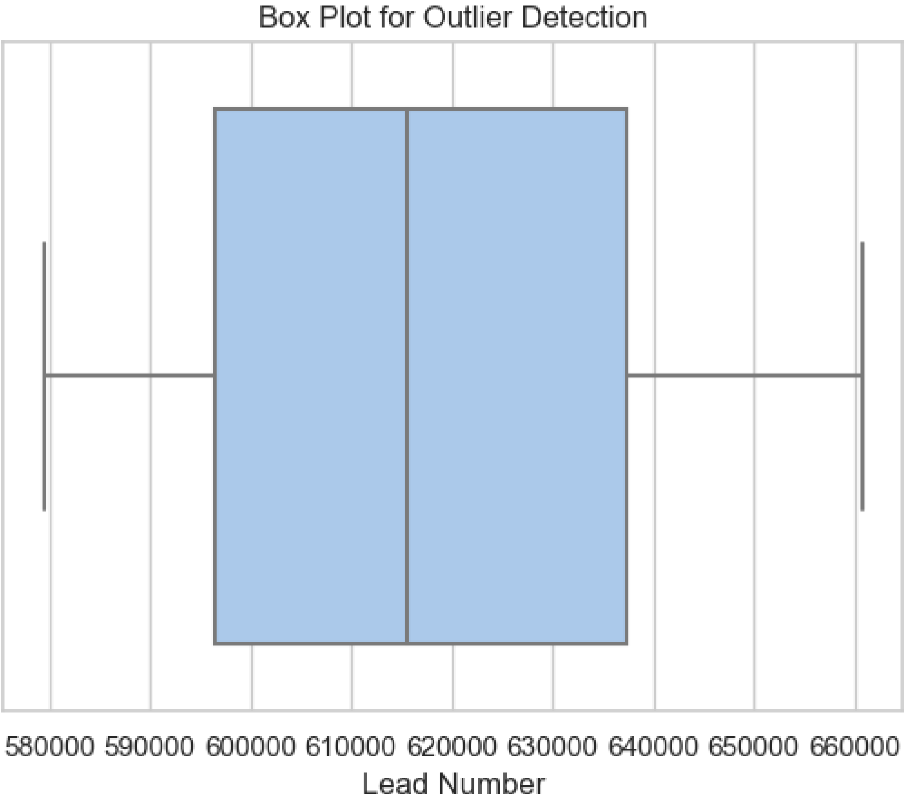
Box Plot for Outlier Detection

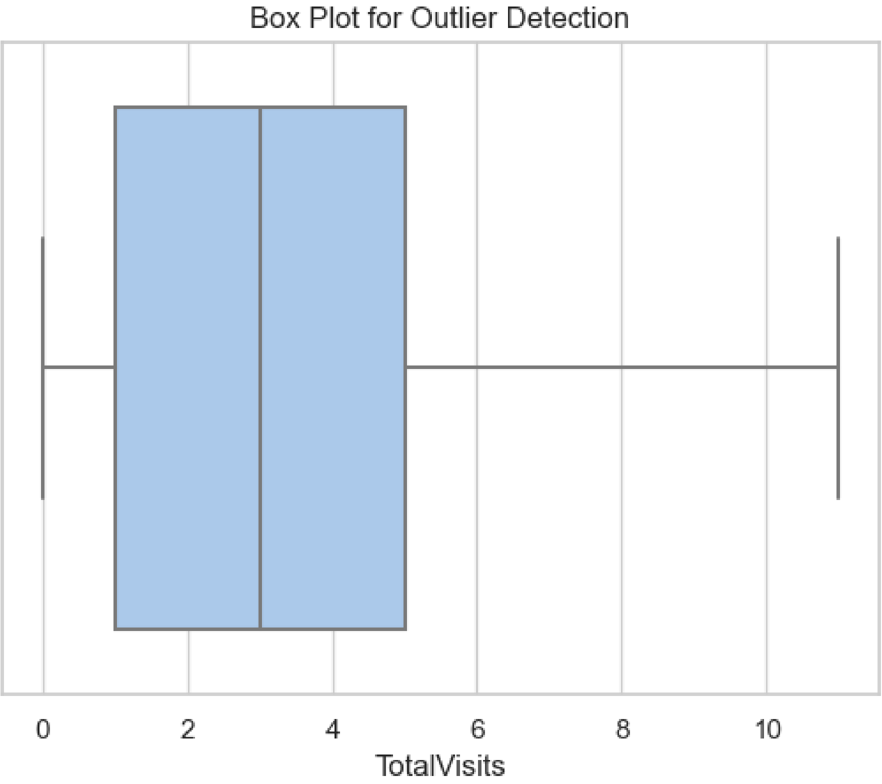


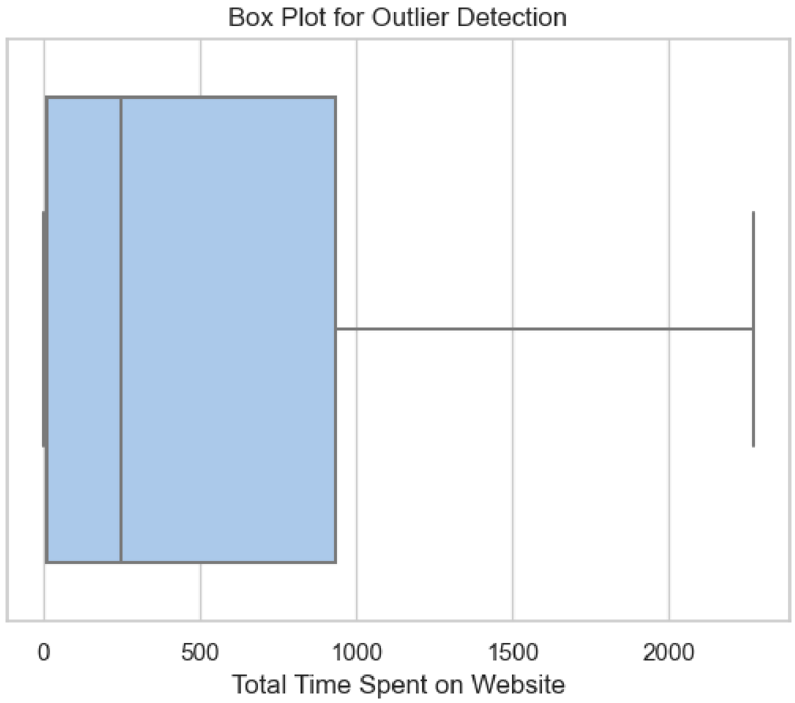
Observations

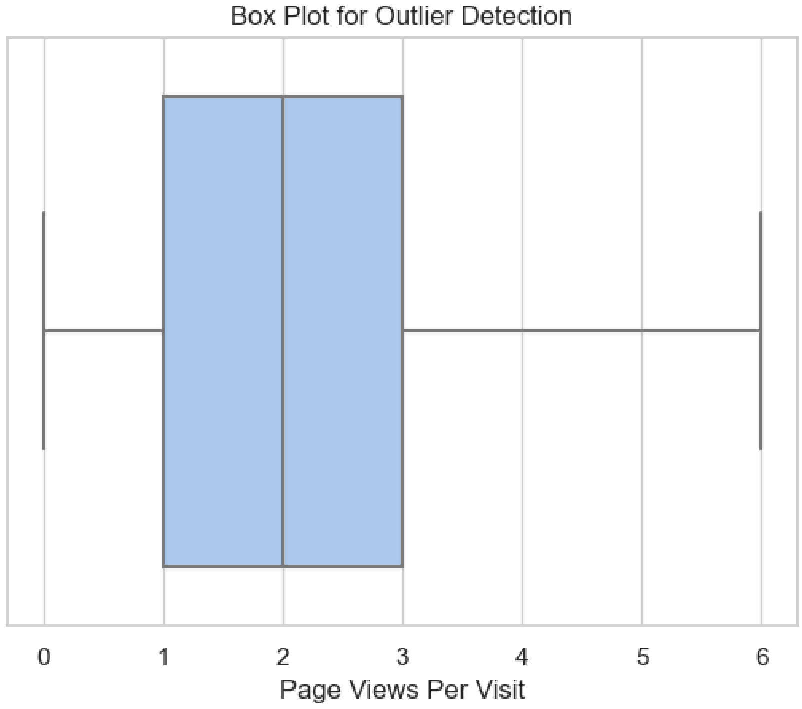
1. In total visits and page views per visit has outliers.

2. Removing outliers does not solve the problem we need to cap the outlier that means if some day total visit per page goes higher than a threshold then we will cap this with thresh value for better results from the model.


Box Plot for Outlier Detection
580000
590000
600000
610000
620000
630000
640000
650000
660000
Lead Number

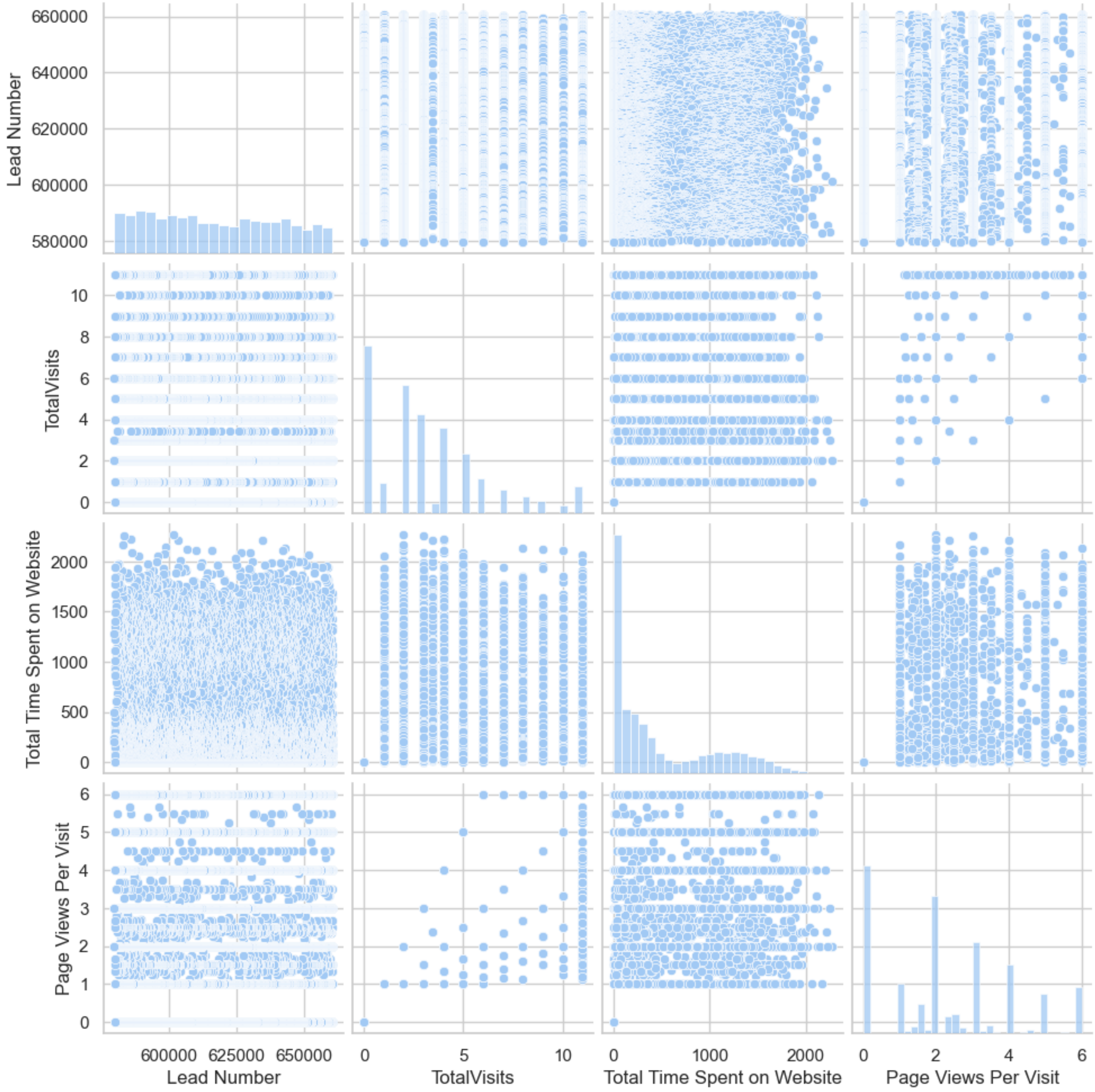

Box Plot for Outlier Detection
0
2
4
6
8
10
TotalVisits





Observations

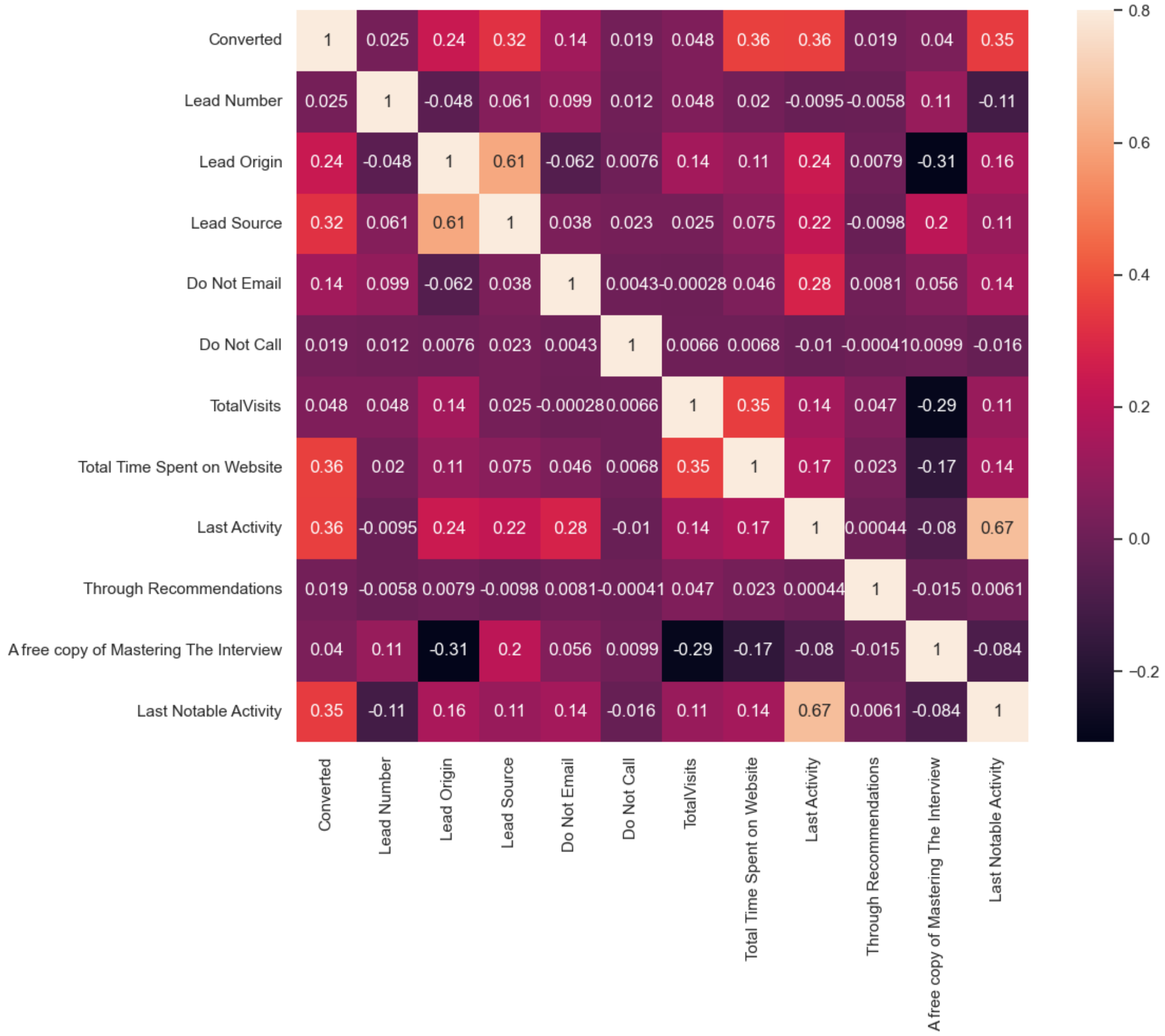
Now there is not outliers, we cap the outliers using z-index method. where we cap the outlier between IQR range.



Observation

In the pair plot we can see the scatter plots between multiple independent numerical variables that follows some trend or has some correlation between them.

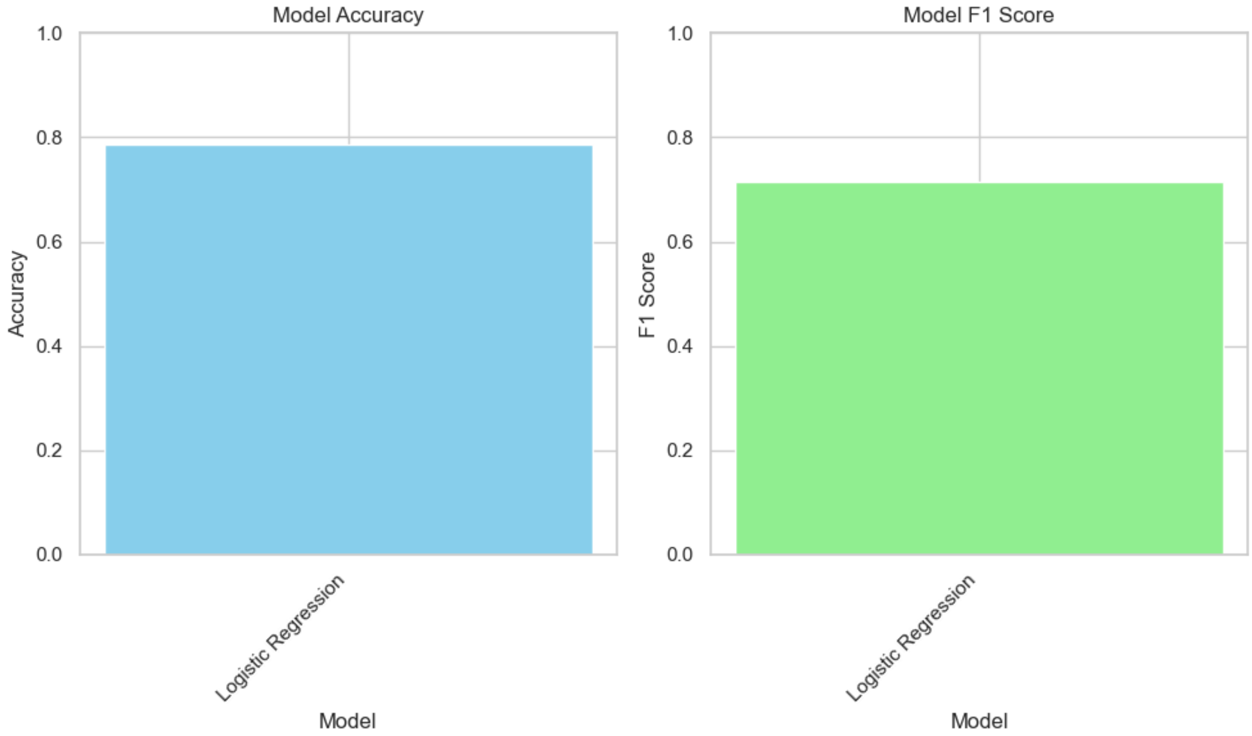
Features Selections or Features Engineering.



Observations

Heat maps showcase the same that we have earlier discussed.

Modelling



Observations

1. Model achived accuracy of 72% and it successfully classify 1445 in positive and 730 in negative side.

2. F-Score is 71 which is good.

3. In the next step we will implemnt hyper parameter tunning for further increasing the accuracy of the model