EDA CASE STUDY
Lead Scoring Case Study

by

AYUSH and DHEERAJ

About the Problem

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

When people end up landing in the course related website and fill up a form providing their details, they are considered as leads. Now sales team will be making calls to the leads to try talking to them and convert them as paying customers.

X eductaion wants to know the potential leads so that the sales team can proceed calling only the potential leads. This would increase the conversion rate and save time and cost for the sales team.

Solution Strategy.

We will be building a logistic regression model using the given data. The result of the model will show the probability of a lead being converted to a paying customer.

So, the sales team now has the list of leads who have better odds of being converted. Now they can contact only these potential leads and eventually have a higher conversion rate.

About the Data Frame.

We have the dataset of leads from the past with around 9000 data points. This dataset consists of various attributes such as details filled by lead in the form, Total visits made to the website, details given by the sales team after contacting the leads, etc.. We have deleted the data provided by the sales team after making a call to them. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not.

Null values

For the sake of this analysis we have dropped the below 49 variables which had more than 30% missing value.

Lead Source	0.389610
TotalVisits	1.482684
Page Views Per Visit	1.482684
Last Activity	1.114719
Country	26.634199
Specialization	36.580087
How did you hear about X Education	78.463203
What is your current occupation	29.112554
What matters most to you in choosing a course	29.318182
Tags	36.287879
Lead Quality	51.590909
Lead Profile	74.188312
City	39.707792
Asymmetrique Activity Index	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Score	45.649351

dtype: float64

Imputation

imputing TotalVisits, Page Views Per Visit and Last Activity with median :
and Lead Source with mode.

EXPLORATION

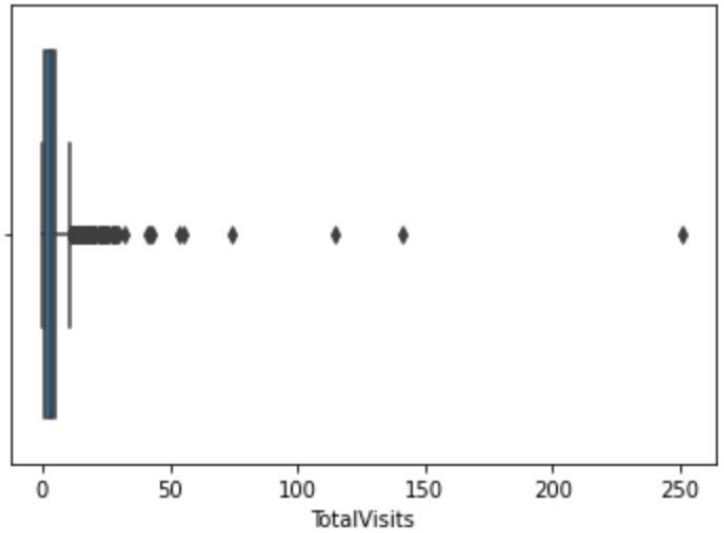
Final stats after data cleaning.

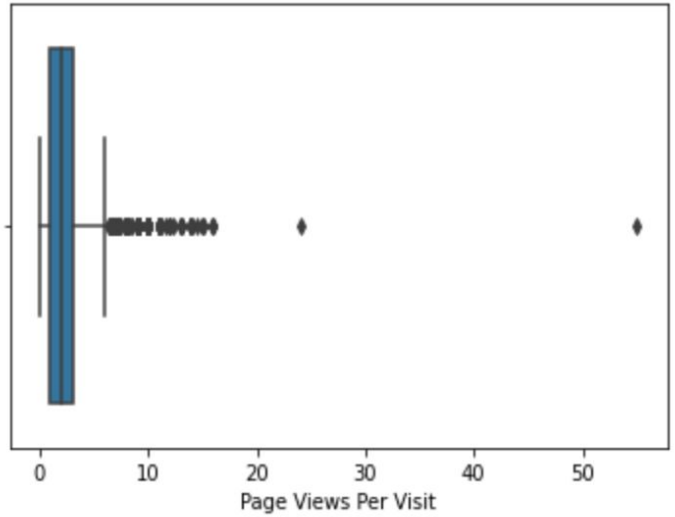
	Lead Origin	Lead Source	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	A free copy of Mastering The Interview	Last Notable Activity
count	9240	9240	9240	9240.000000	9240.000000	9240.000000	9240.000000	9240	9240	9240
unique	5	5	2	NaN	NaN	NaN	NaN	4	2	16
top	Landing Page Submission	Google	No	NaN	NaN	NaN	NaN	Email Opened	No	Modified
freq	4886	2868	8506	NaN	NaN	NaN	NaN	3437	6352	3407
mean	NaN	NaN	NaN	0.385390	3.438636	487.698268	2.357440	NaN	NaN	NaN
std	NaN	NaN	NaN	0.486714	4.819024	548.021466	2.145781	NaN	NaN	NaN
min	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	NaN	NaN	NaN
25%	NaN	NaN	NaN	0.000000	1.000000	12.000000	1.000000	NaN	NaN	NaN
50%	NaN	NaN	NaN	0.000000	3.000000	248.000000	2.000000	NaN	NaN	NaN
75%	NaN	NaN	NaN	1.000000	5.000000	936.000000	3.000000	NaN	NaN	NaN
max	NaN	NaN	NaN	1.000000	251.000000	2272.000000	55.000000	NaN	NaN	NaN

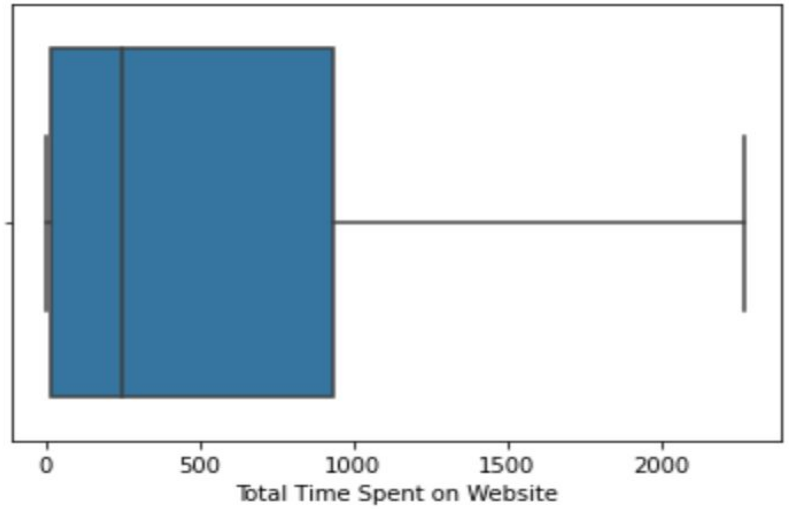
digging into percentiles to check for outliers

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.00	9240.00	9240.00	9240.00
mean	0.39	3.44	487.70	2.36
std	0.49	4.82	548.02	2.15
min	0.00	0.00	0.00	0.00
15%	0.00	0.00	0.00	0.00
35%	0.00	2.00	98.00	1.50
50%	0.00	3.00	248.00	2.00
55%	0.00	3.00	305.00	2.00
75%	1.00	5.00	936.00	3.00
95%	1.00	10.00	1562.00	6.00
max	1.00	251.00	2272.00	55.00

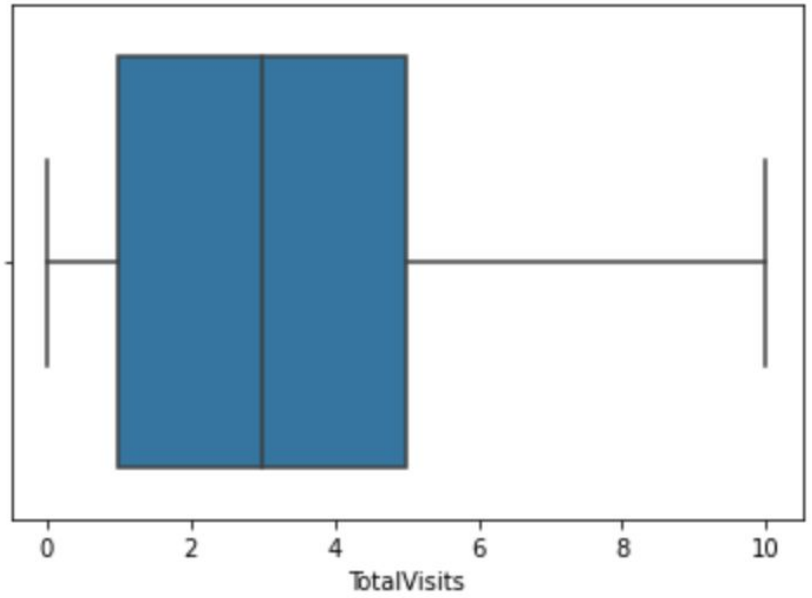
Outlier treatment
before outlier treatment plots

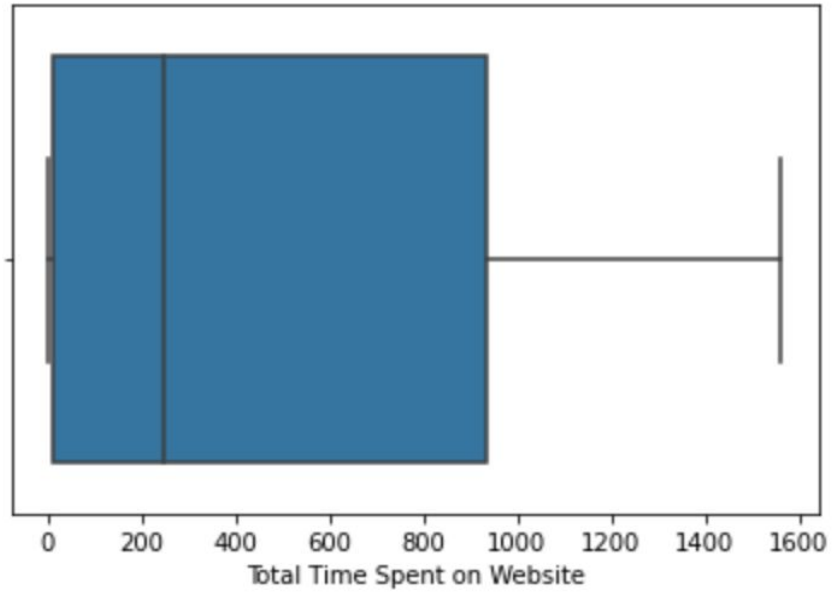


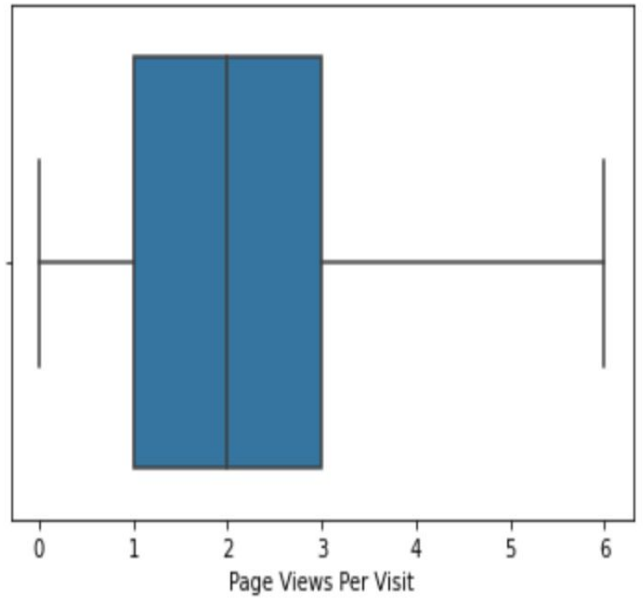




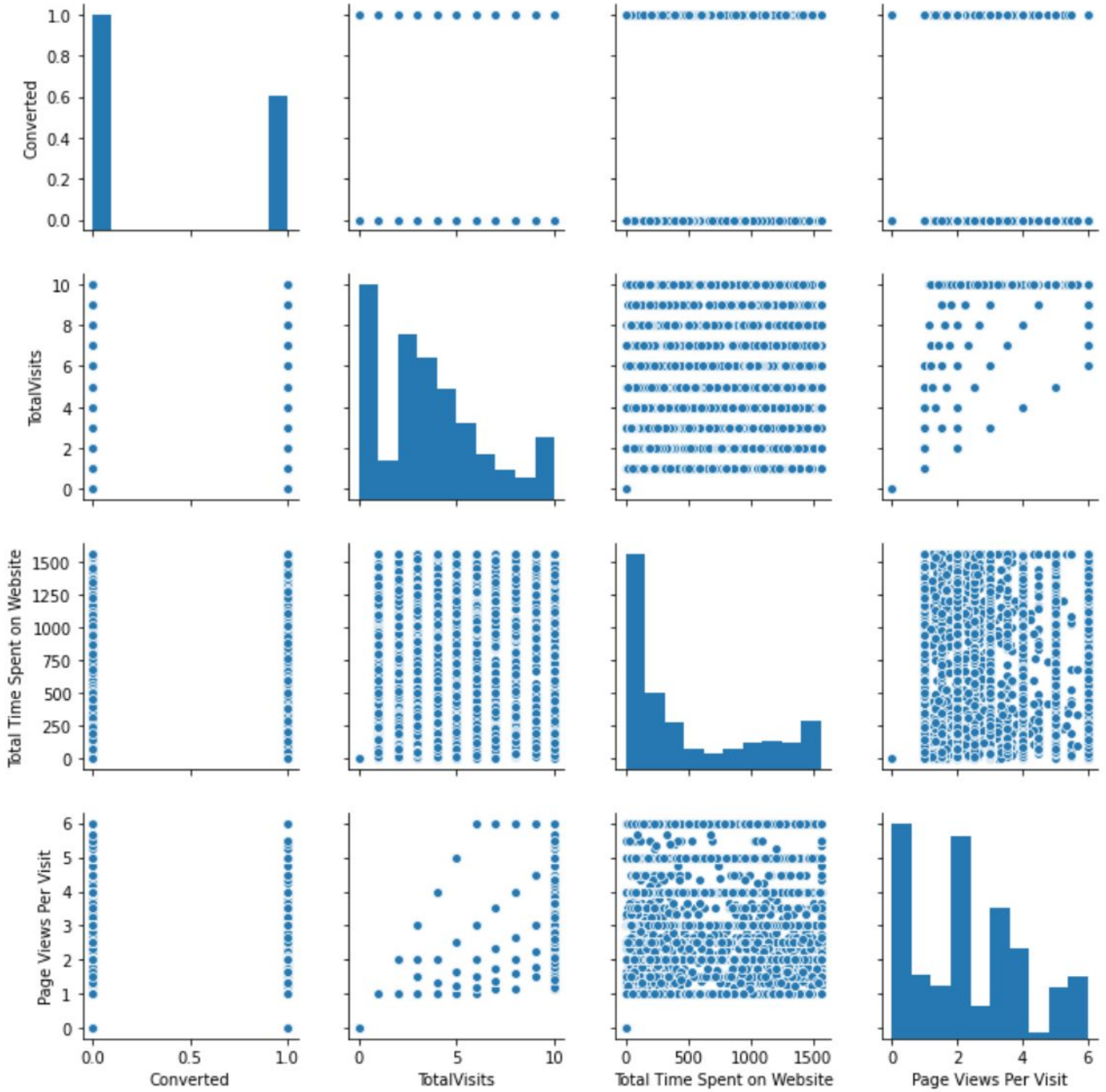
After treatment of outliers capping the upper values to 95% plots



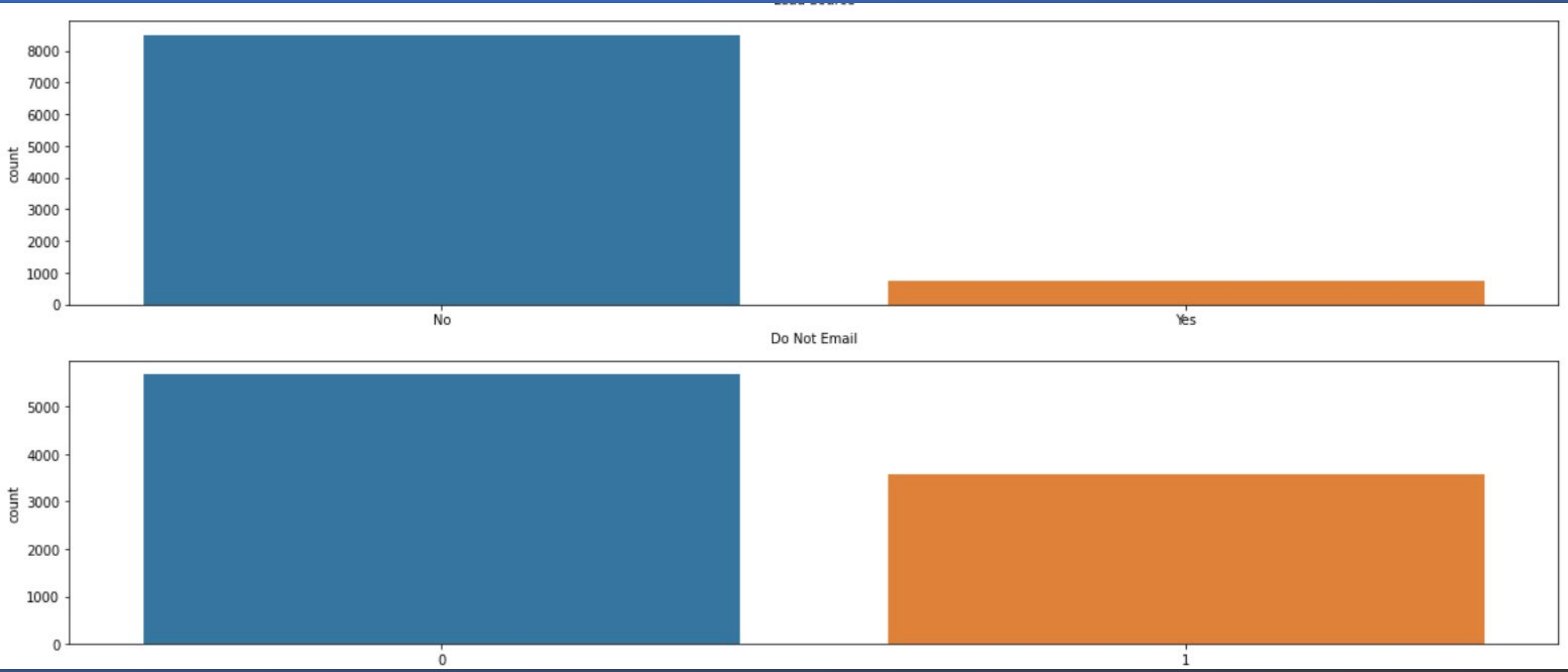


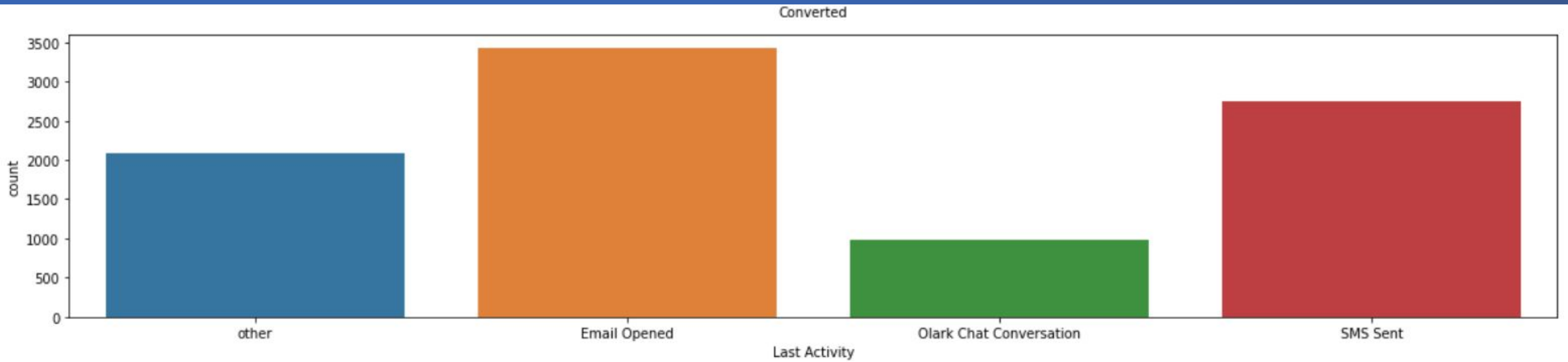


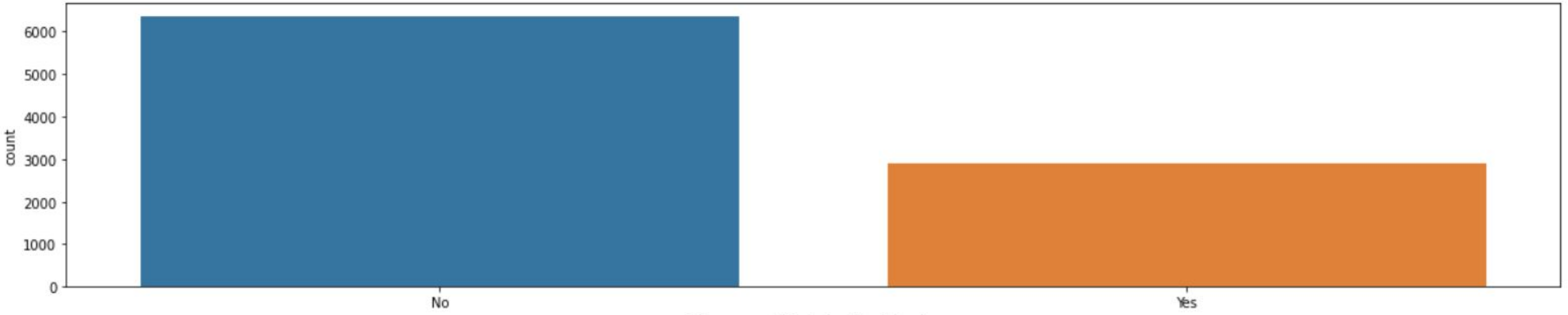
Plots - Pairplot

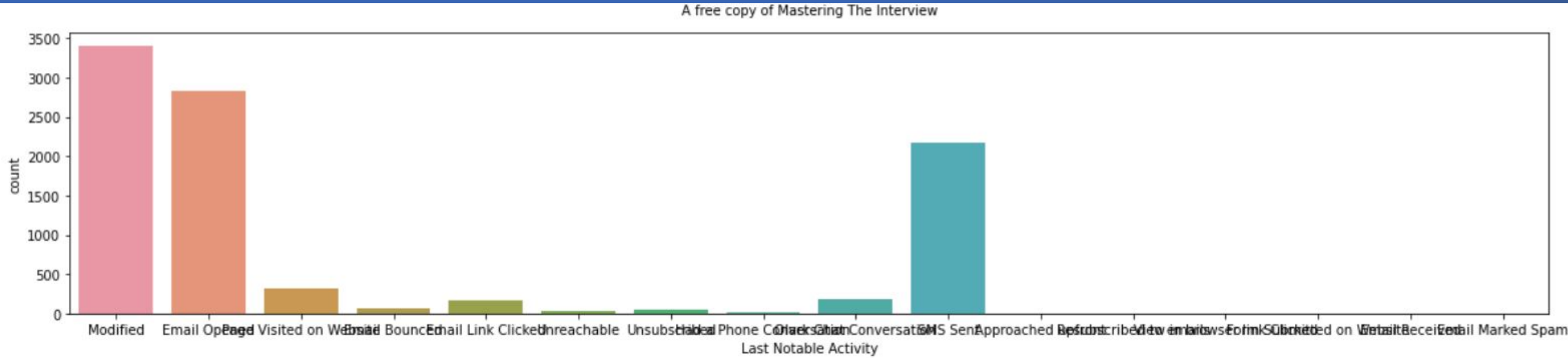


categorical plots

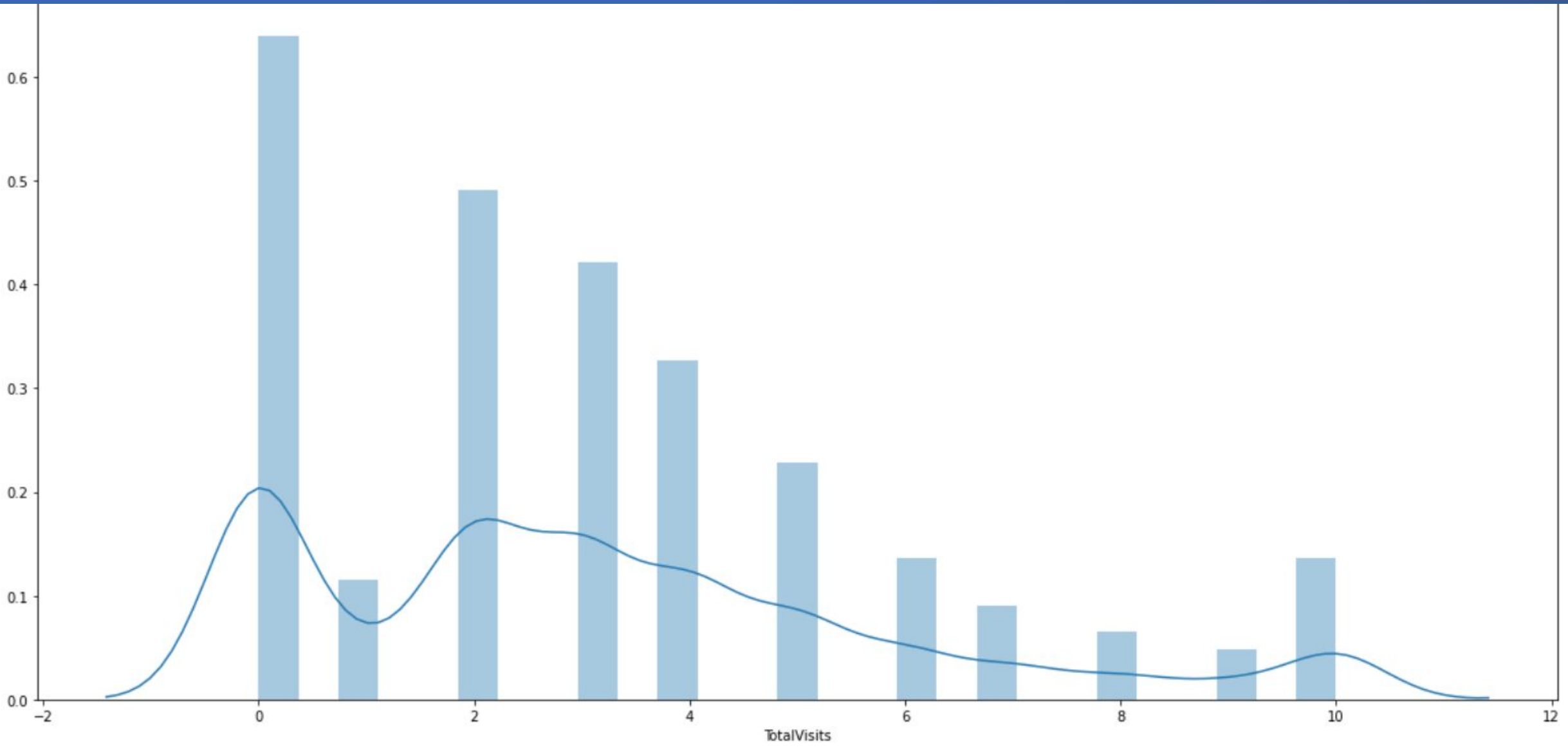


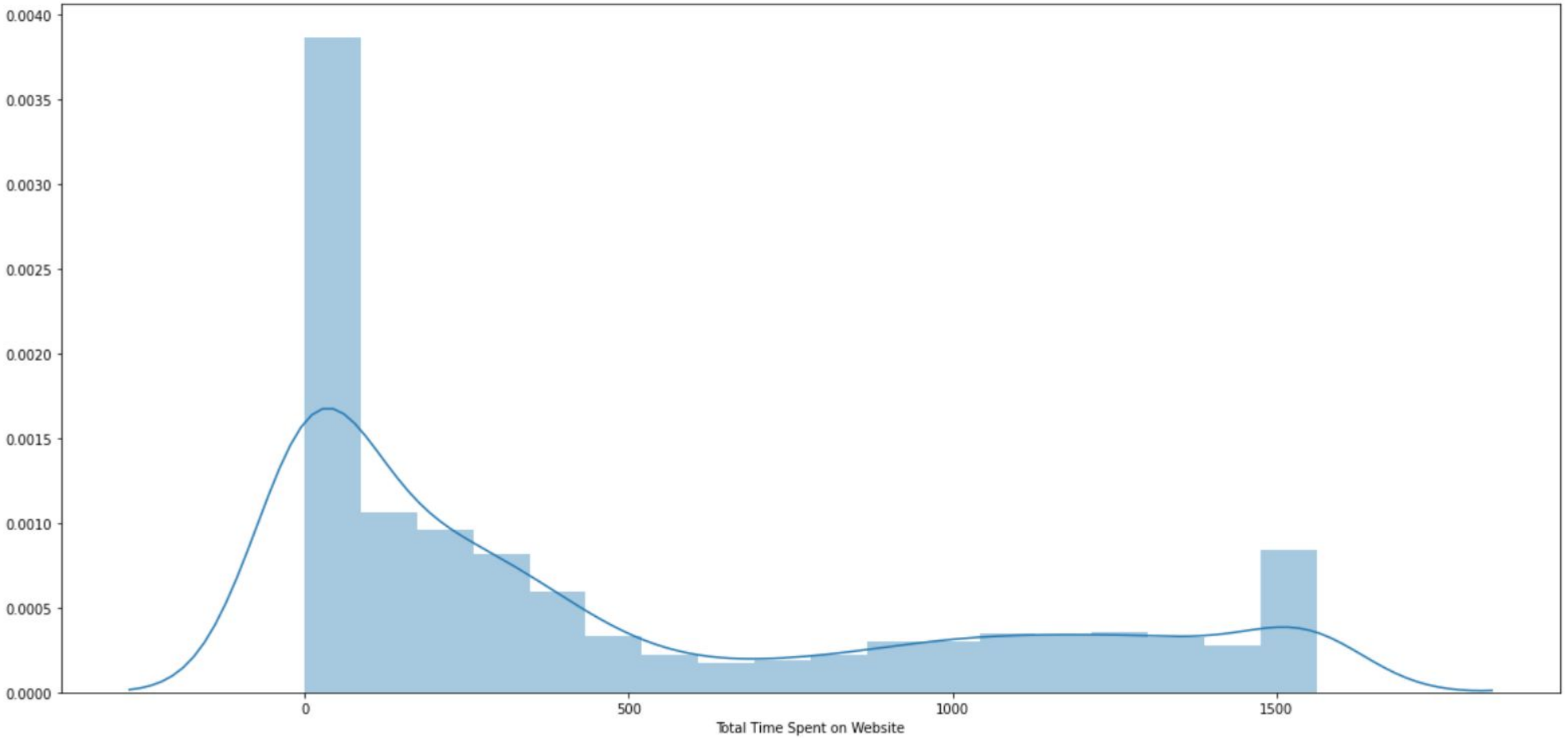

Converted
3500
3000
2500
2000
1500
1000
500
0
count
other
Email Opened
Olark Chat Conversation
SMS Sent
Last Activity

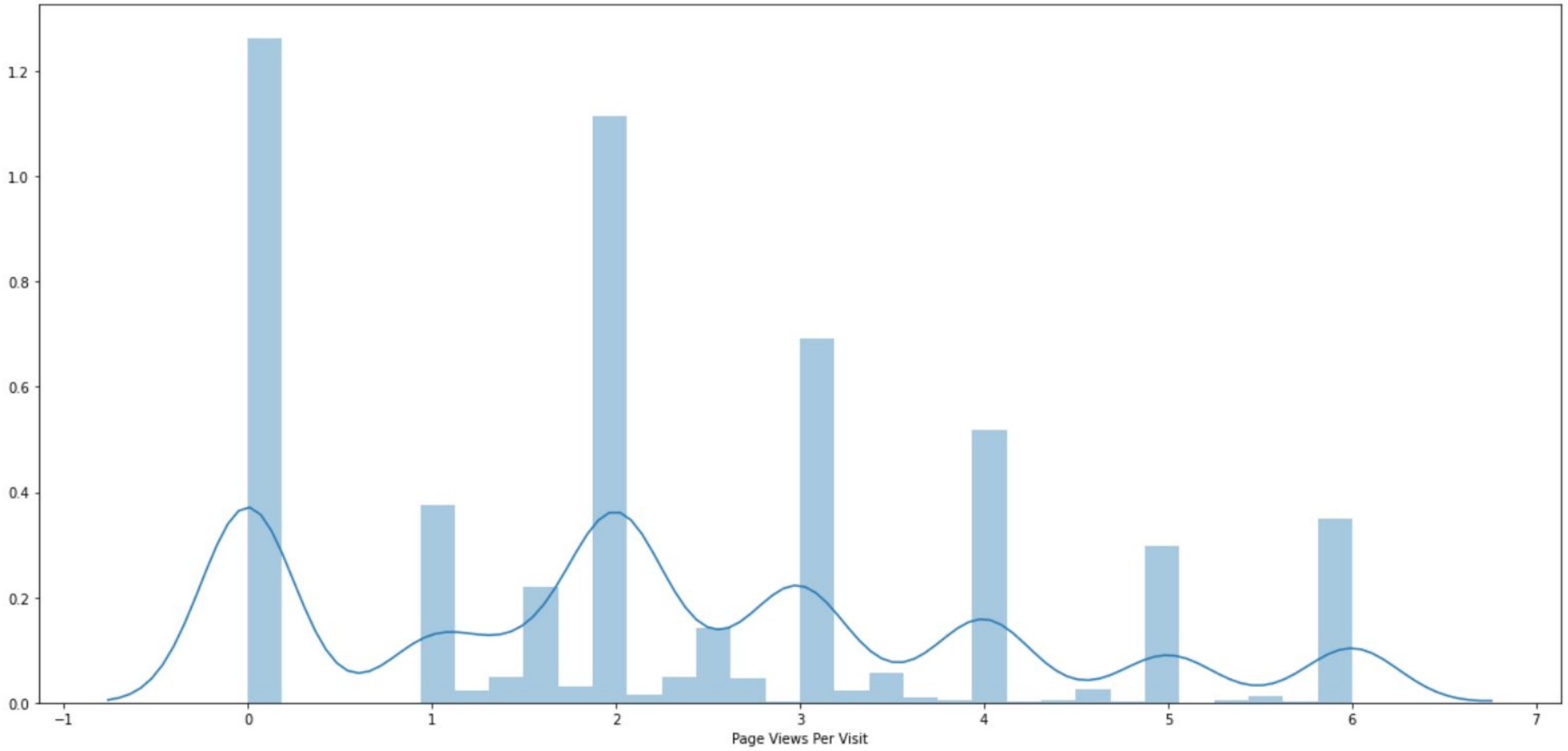

6000
5000
4000
3000
2000
1000
0
count
No
Yes



converted(target) seem well distributed
last notable activity seem sparse maybe we can club these lesser numbered values <br. we already have last activity which has a good distribution and hence last notable activity may not be of that much interest so lets drop last notable activity
the rest looks good and ready for modeling

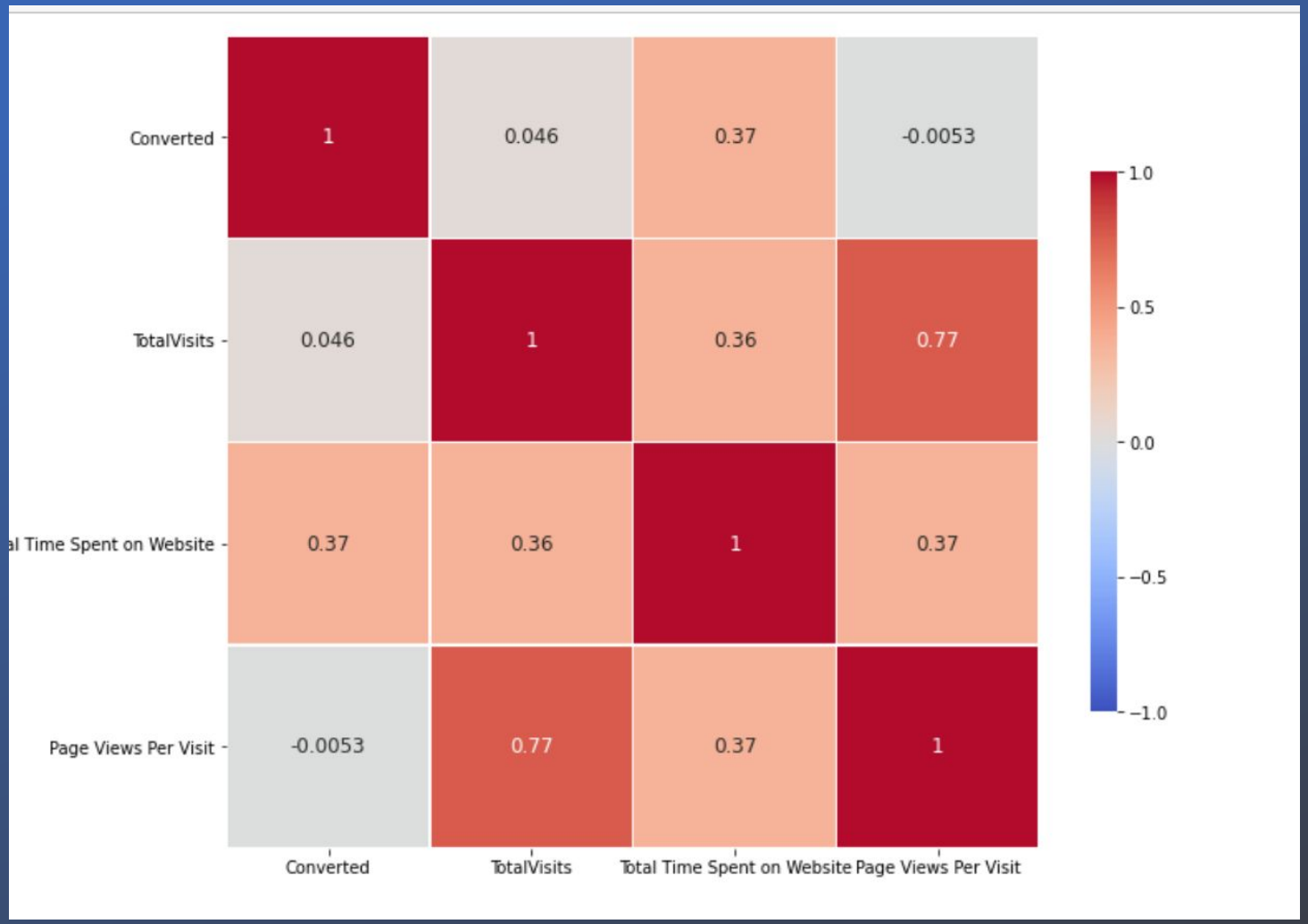

0.6
0.5
0.4
0.3
0.2
0.1
0.0
−2
0
2
4
6
8
10
12
TotalVisits


0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
500
1000
1500
Total Time Spent on Website


1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5
6
7
Page Views Per Visit

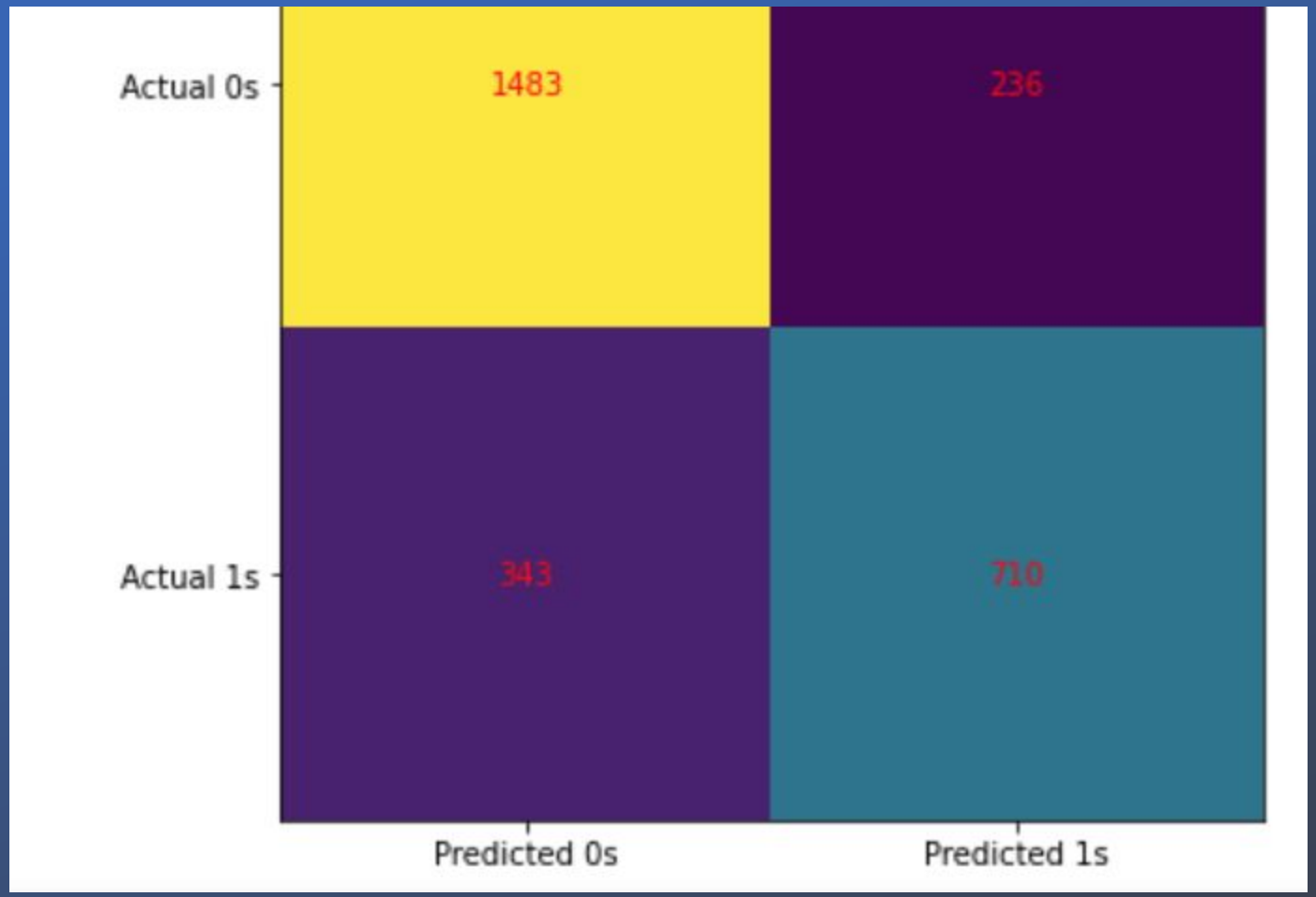
Correlation metrics

total visit and page view per visit seem to be highly correlated the rest look good lets go ahead and model the data



Modeling

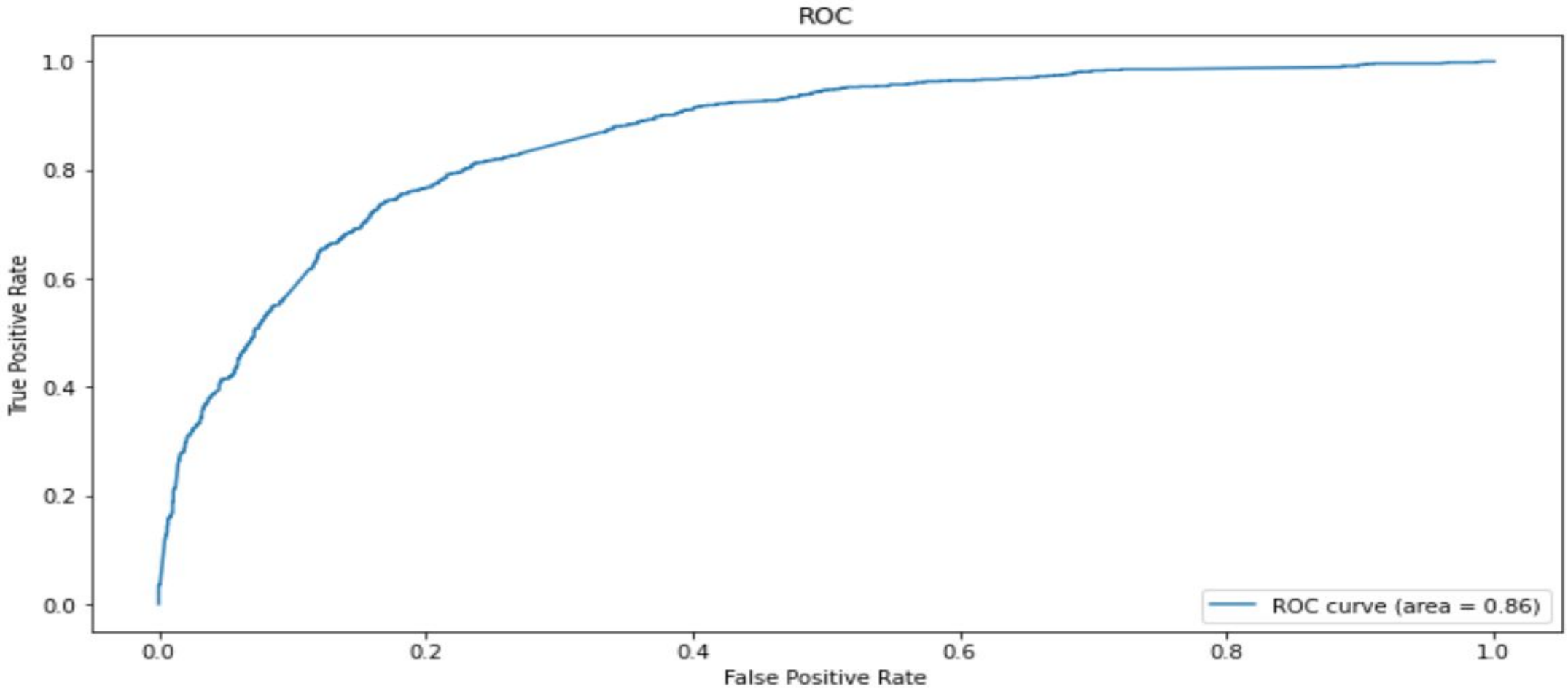
Logistic Regression confusion matrix



	precision	recall	f1-score	support
0	0.81	0.86	0.84	1719
1	0.75	0.67	0.71	1053
accuracy			0.79	2772
macro avg	0.78	0.77	0.77	2772
weighted avg	0.79	0.79	0.79	2772

accuracy of 79 lokks good for first try

Logistic Regression ROC curve



RFE

```
: rfe.support_
```

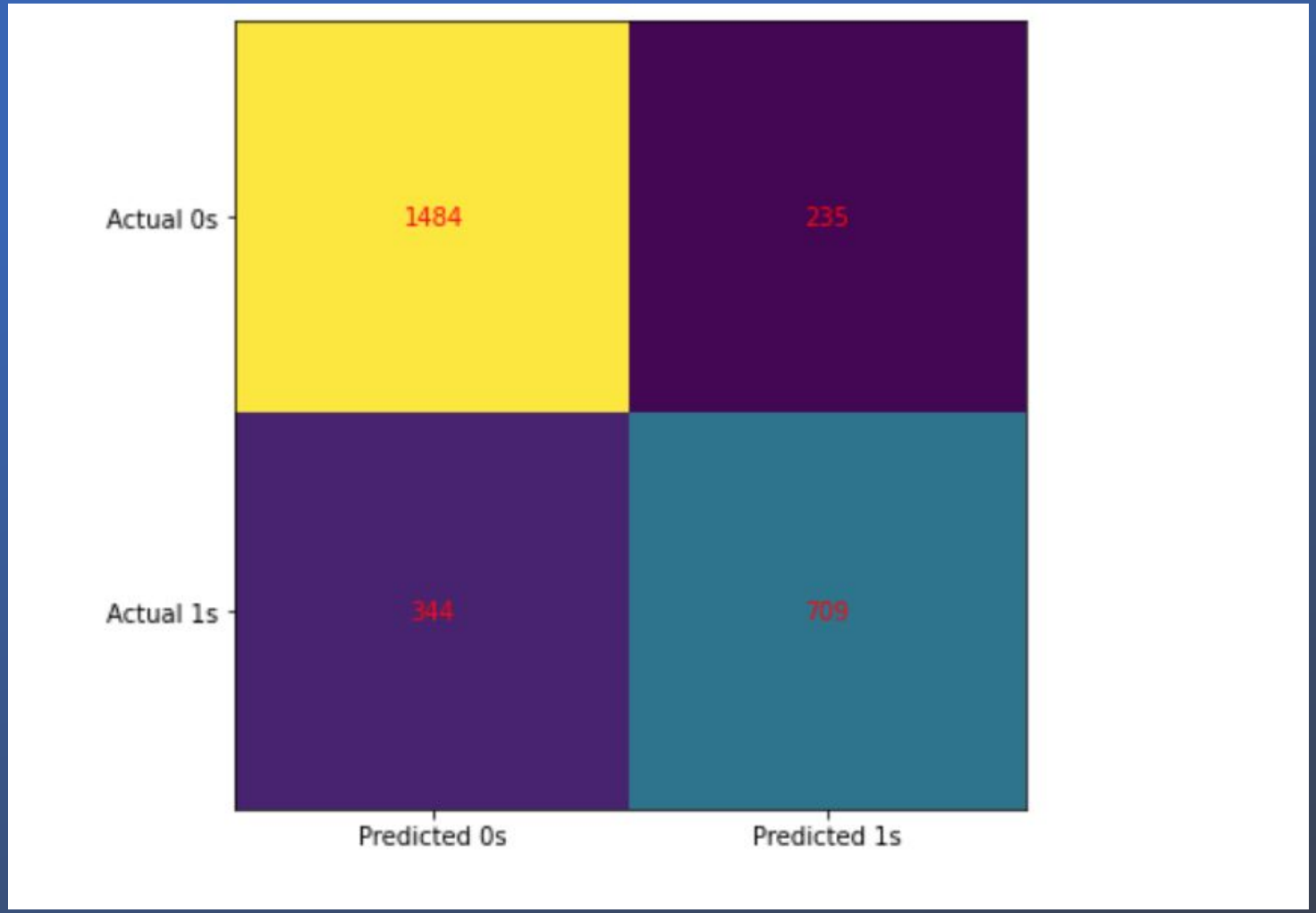
```
: array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
         True,  True,  True,  True,  True,  True,  True])
```

VIF

```
vif
```

	Features	VIF
10	Lead Source_other	6.62
4	Lead Origin_Lead Add Form	6.60
3	Lead Origin_Landing Page Submission	3.79
2	Page Views Per Visit	3.29
15	A free copy of Mastering The Interview_Yes	2.69
0	TotalVisits	2.59
8	Lead Source_Olark Chat	2.10
7	Lead Source_Google	1.79
13	Last Activity_SMS Sent	1.73
14	Last Activity_other	1.68
12	Last Activity_Olark Chat Conversation	1.44
5	Lead Origin_Lead Import	1.41
9	Lead Source_Organic Search	1.34
1	Total Time Spent on Website	1.29
11	Do Not Email_Yes	1.24
6	Lead Origin_Quick Add Form	NaN

new confusion matrix

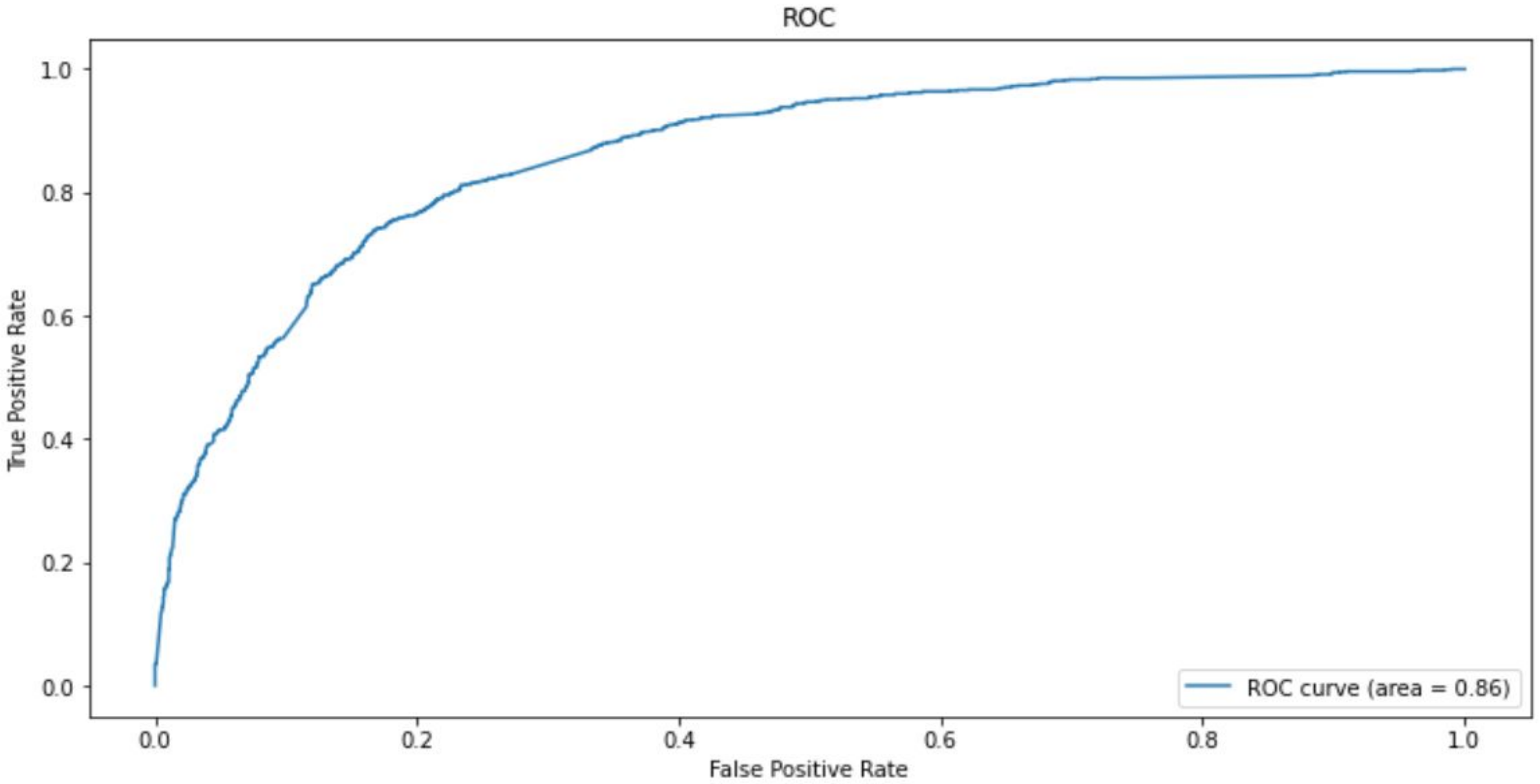


new confusion matrix

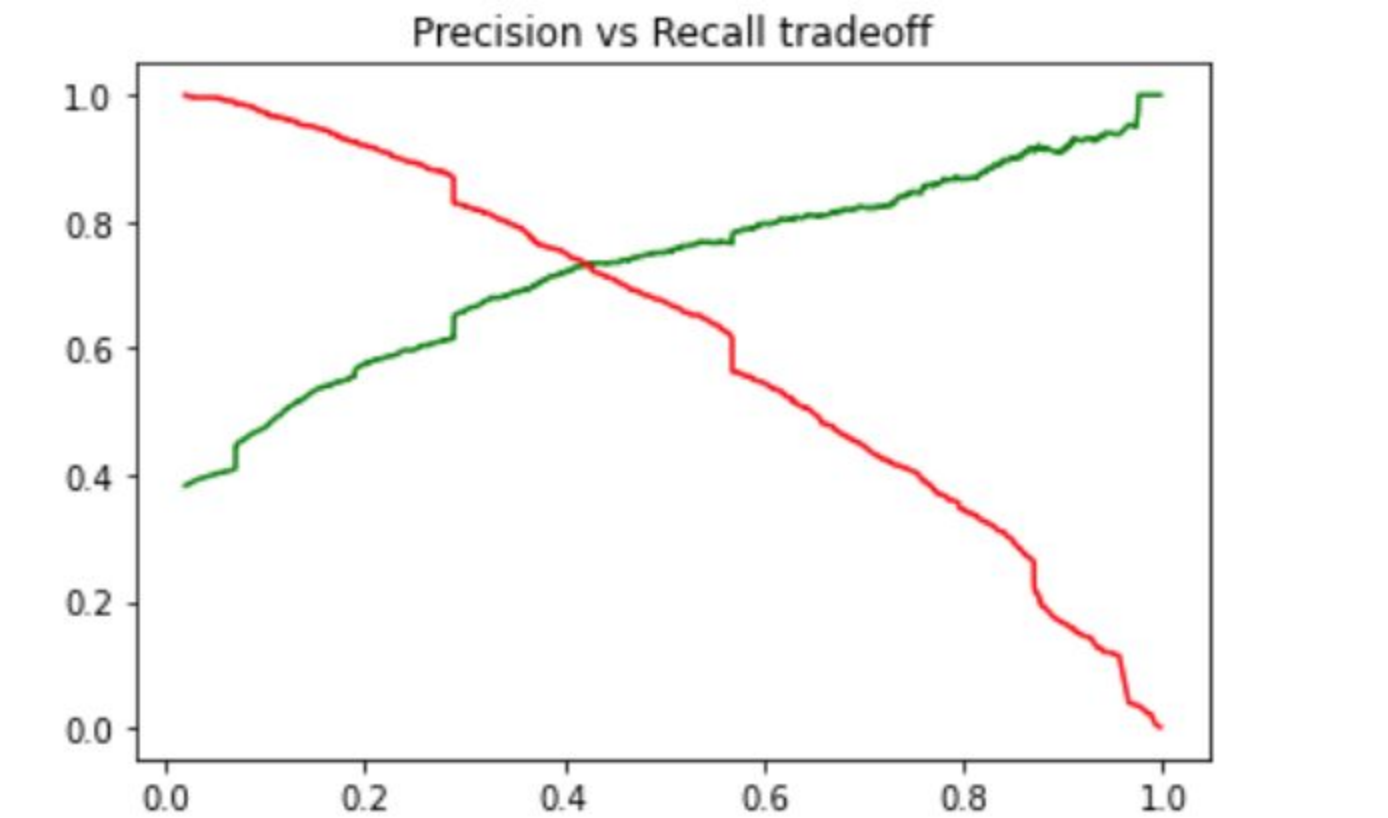
	precision	recall	f1-score	support
0	0.81	0.86	0.84	1719
1	0.75	0.67	0.71	1053
accuracy			0.79	2772
macro avg	0.78	0.77	0.77	2772
weighted avg	0.79	0.79	0.79	2772

accuracy of 79 lokks good for first try

ROC curve


ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Thresholding



Final results after thresholding

	precision	recall	f1-score	support
0	0.81	0.86	0.84	1719
1	0.75	0.67	0.71	1053
accuracy			0.79	2772
macro avg	0.78	0.77	0.77	2772
weighted avg	0.79	0.79	0.79	2772

	index	Converted_Probability	Predicted	real
0	8081	0.981816	1	1
1	4534	0.458544	1	0
2	4110	0.288606	0	0
3	2780	0.069883	0	0
4	2193	0.069883	0	0

THANK you!!!!!