# Lead Score Case Study

SYED SAIFULLAH TARIQUE & PARTEEK KAUSHIK

# Problem Statement

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.

When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.

# Problem Statement

Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Although X Education gets a lot of leads, its lead conversion rate is very poor.

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.


# Lead Conversion Process

# Goals

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. s. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

# Goals

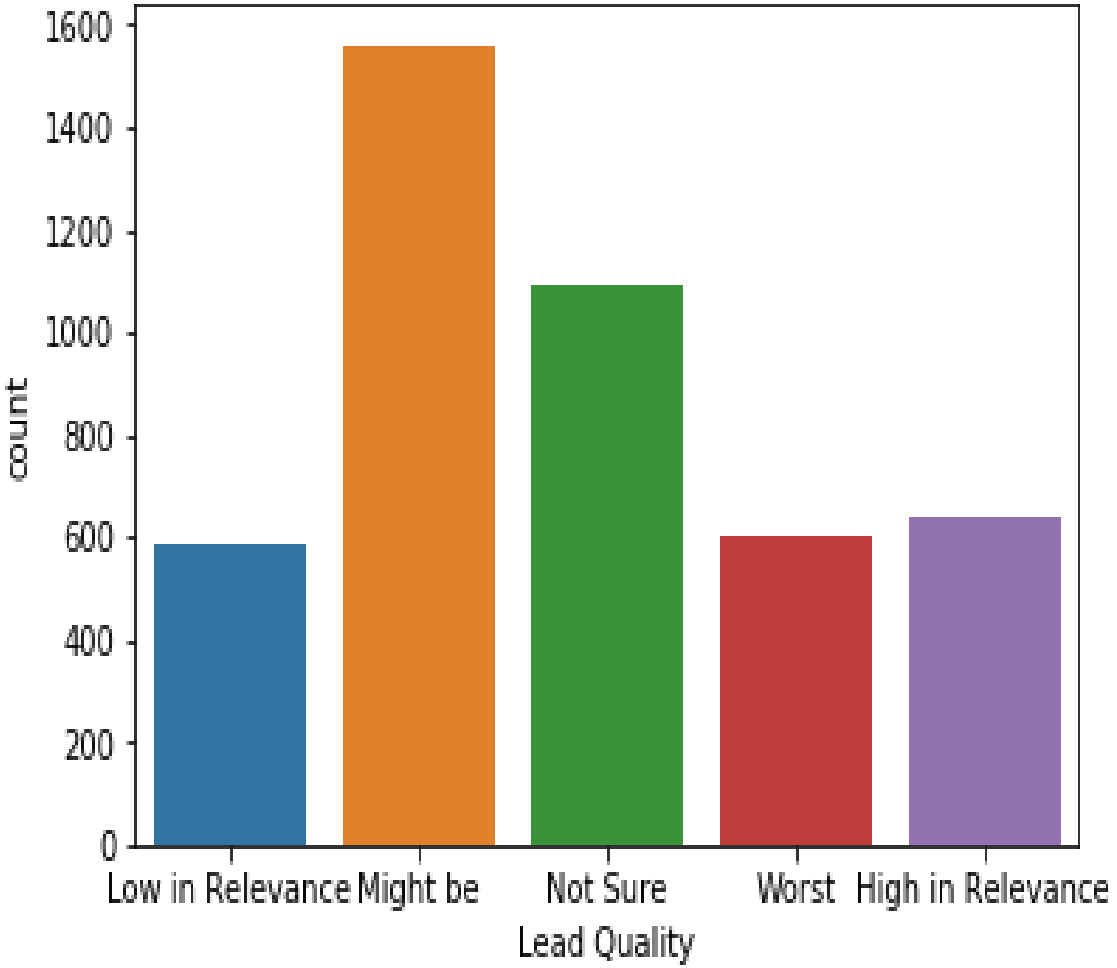
X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO has given a ballpark of the target lead conversion rate to be around 80%.

# Lead Quality Variable

1. Here looking at the graph we could say that the null values in the variable 'Lead Quality' can be put in the category of 'Not Sure' since the variable 'Lead Quality' indicates quality of the leads depending upon the data and the intuition. Since we don't know the quality for the leads for the rows having null values so we assign them to the categories 'Not Sure'.

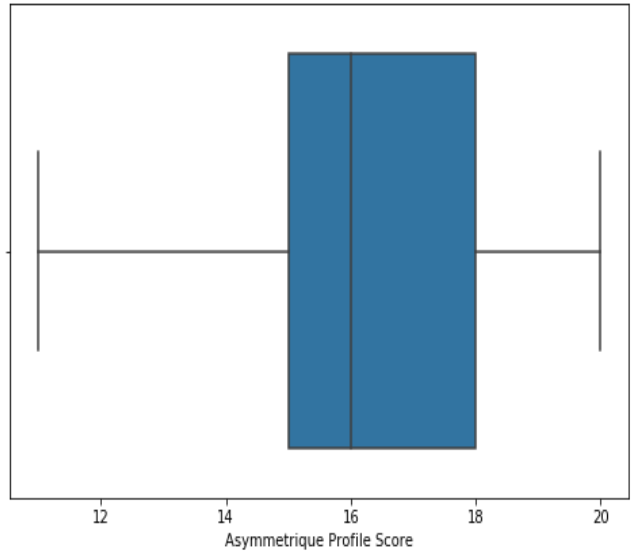


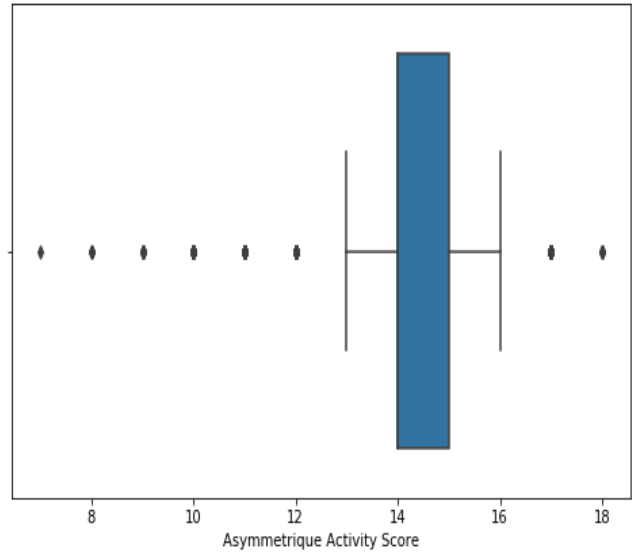
# Variables Plots

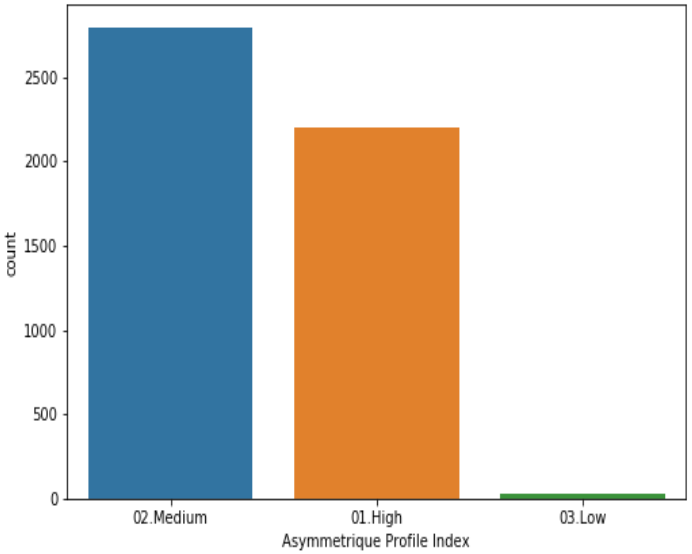
In all the above four variables there are around 45% missing values.

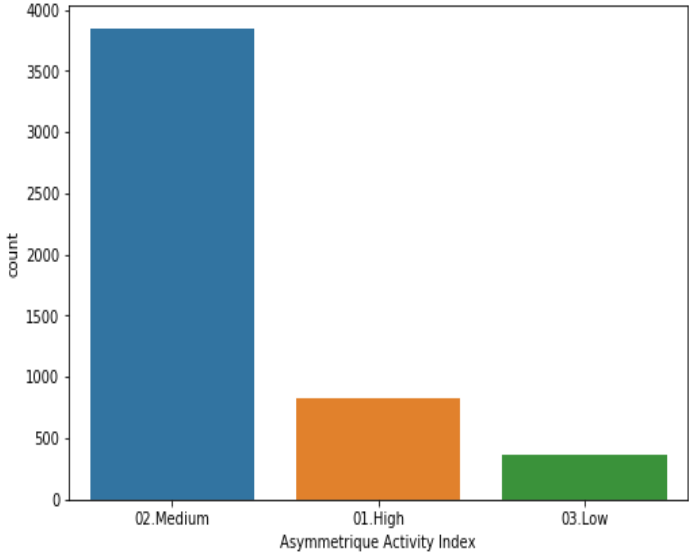
For the categorical variables there aren't any suitable substitute for the null value and for the numerical variables, 45% missing values are a huge amount of number to replace with any value.

So we will drop these columns.



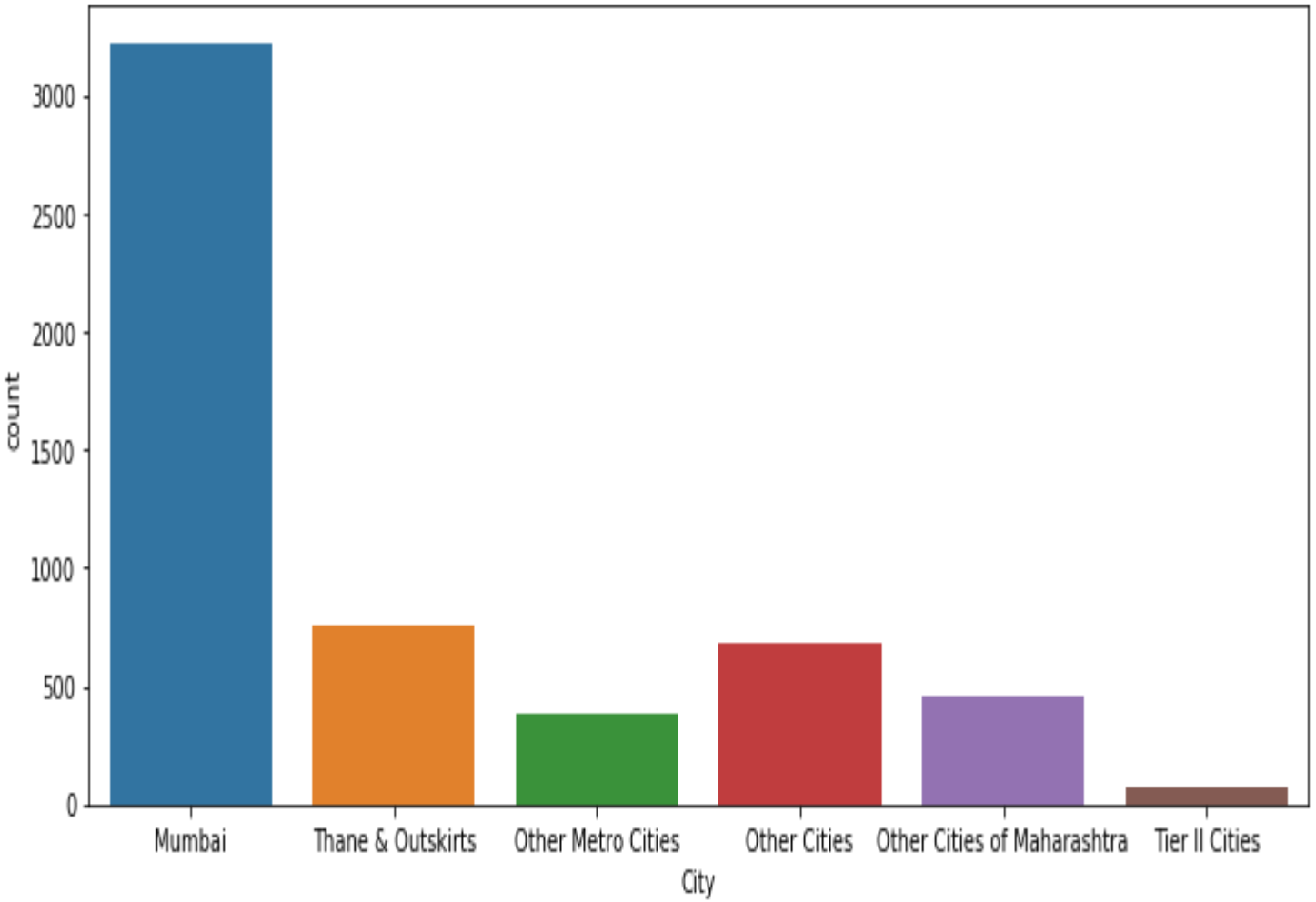



# City Variable Plot

In variable 'City' we could see that the category 'Mumbai' is maximum(around 60%). And is far ahead of any other categories.

We will impute the null value with the most frequent category 'Mumbai‘.

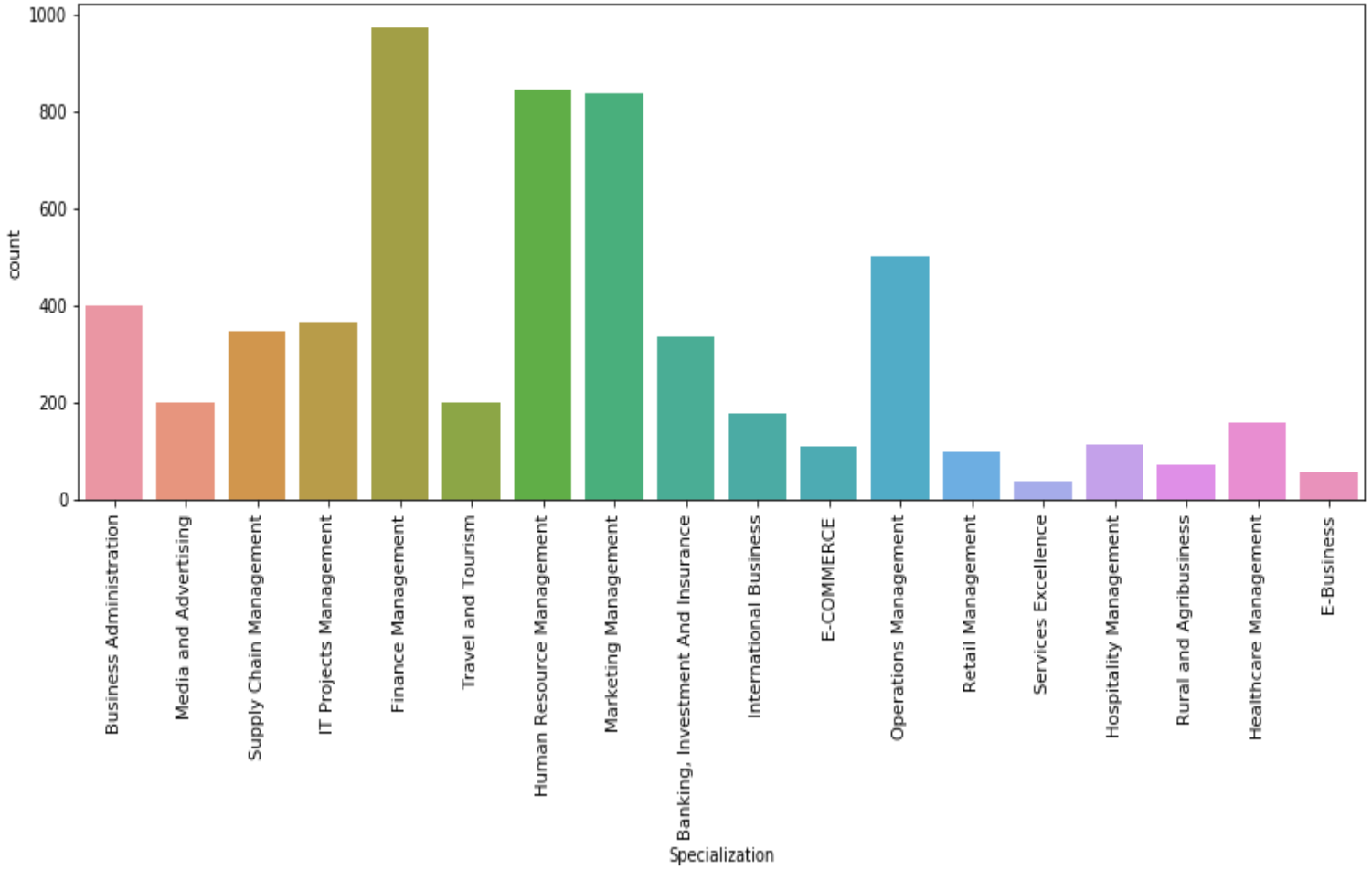


# Specialization Variable Plot

This category Specialization tells us about the industry customer is specialized

Here we could see that the category 'Financial Management' has the maximum number of customer followed by 'Human Resource Management' and 'Marketing Management'.

Here to impute null value we will create a new column called 'Other'. Since people who doesn't fill this variable are generally those who do not find their profession in the list.

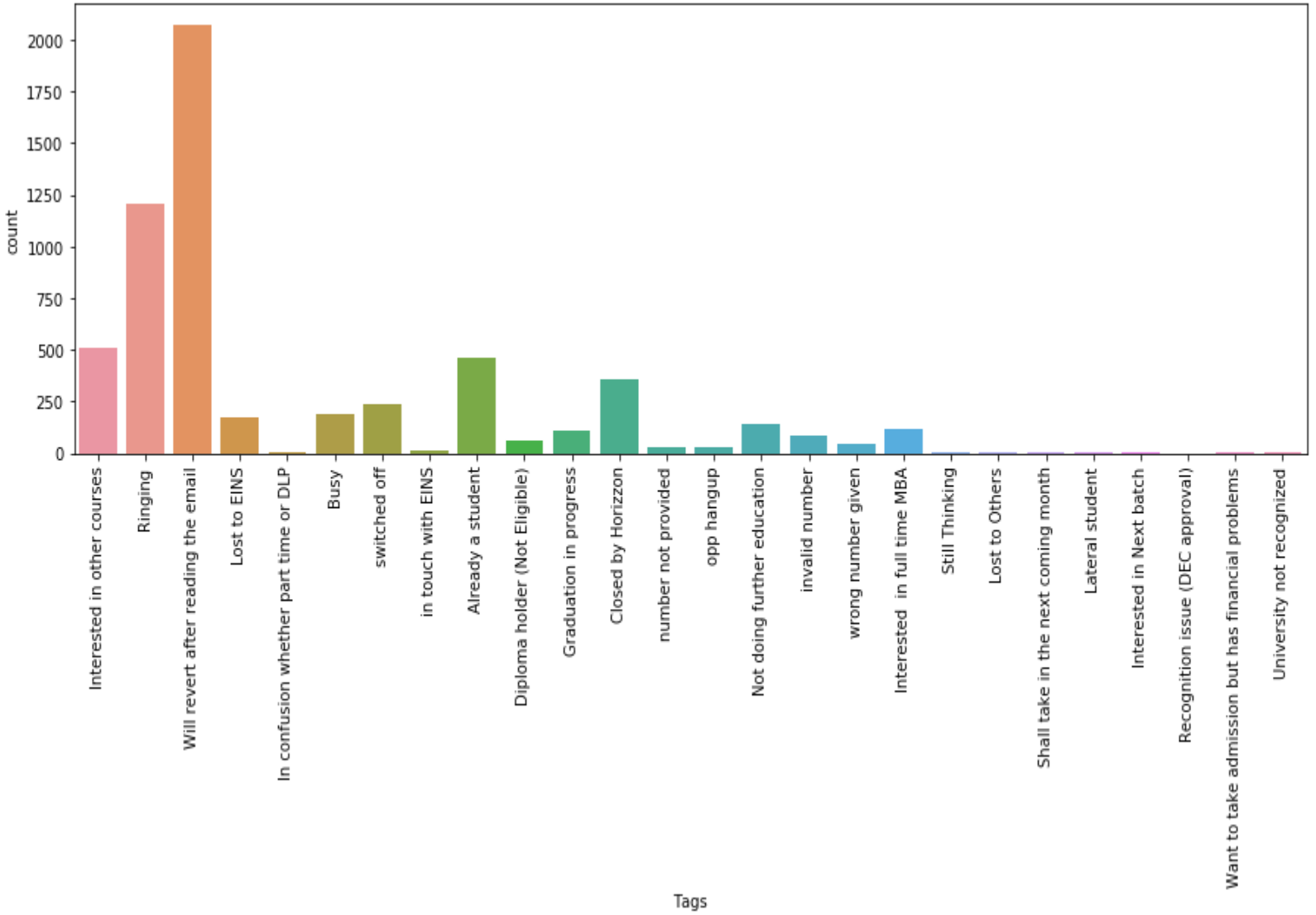


# Tags Variable Plot

This variable tells us about the current status of the customer.

Here we could see that around 40% of the leads are tagged as 'Will revert after reading the email'.

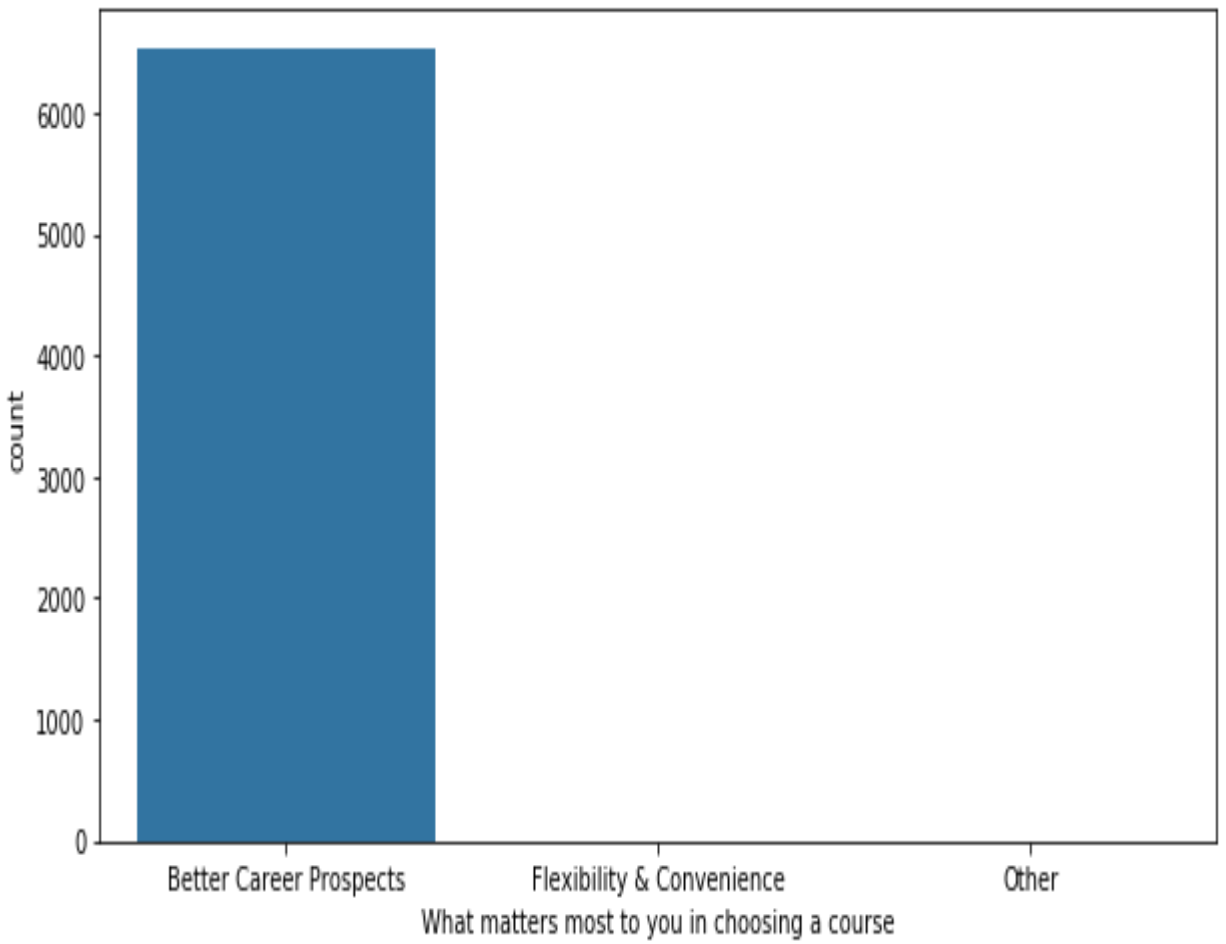
Here we will impute the missing value with the category having the maximum value.



# What matters most to you in choosing a course

In this column you could see that almost all the leads are to the category 'Better Career Prospect'

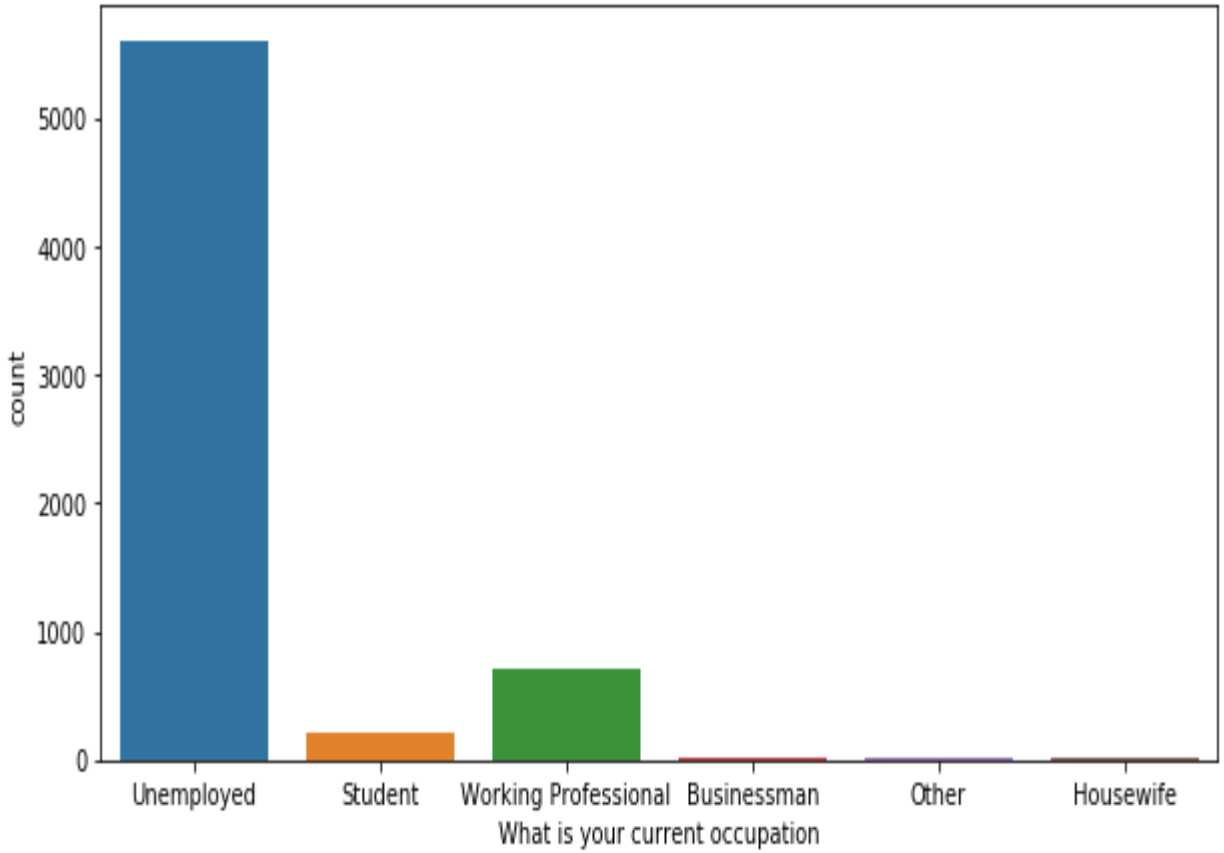
Here we will impute the missing value with the category 'Better Career Prospect‘.



# What is your current occupation variable Plot

Here you could see that more than 80% of the leads are in the category 'Unemployed'

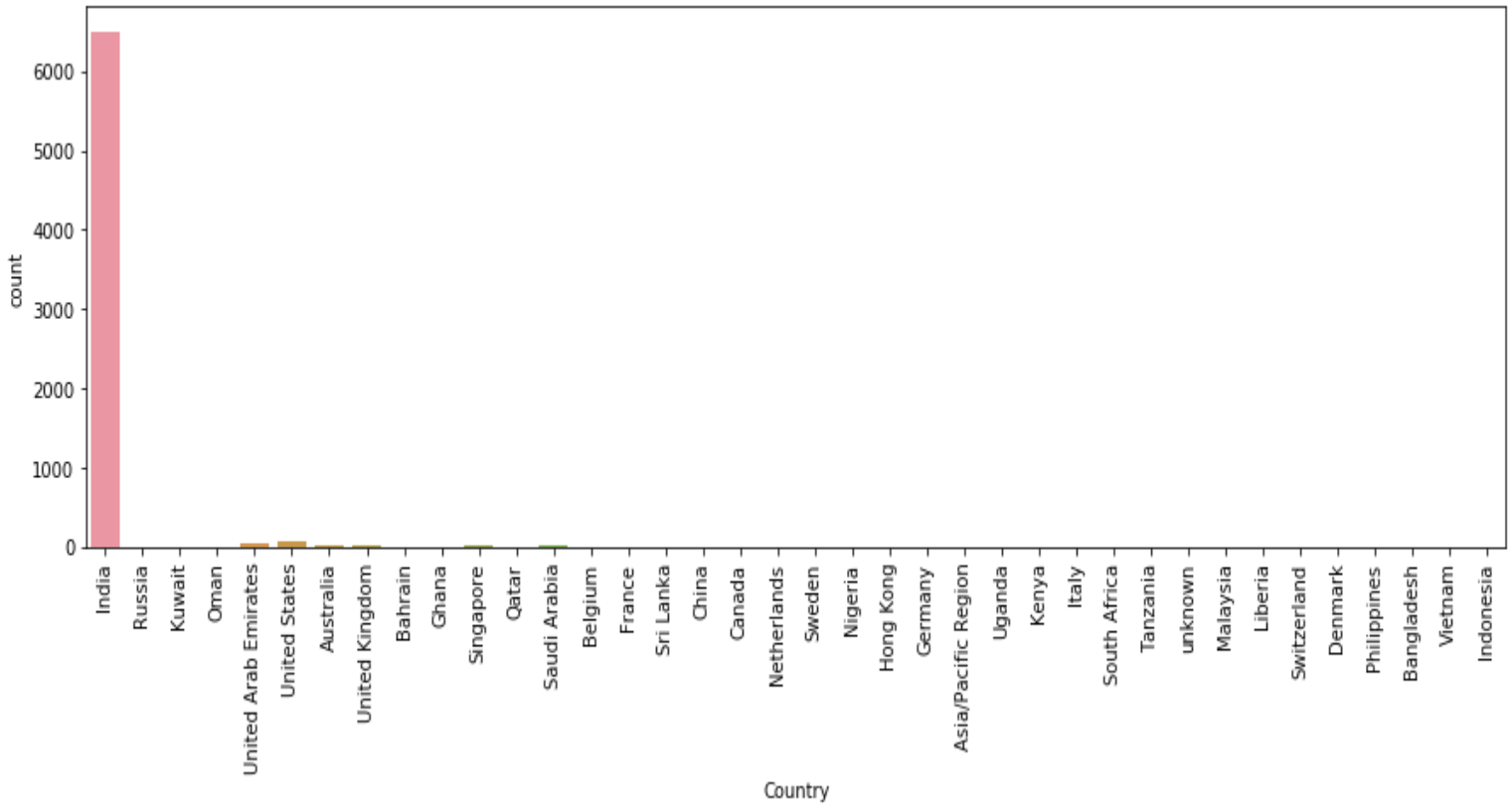
We will impute the missing value with the category 'Unemployed'



# Country Variable Plot

Here you could see that almost all the leads are from category 'India'.

We will impute the missing value with the category 'India'.

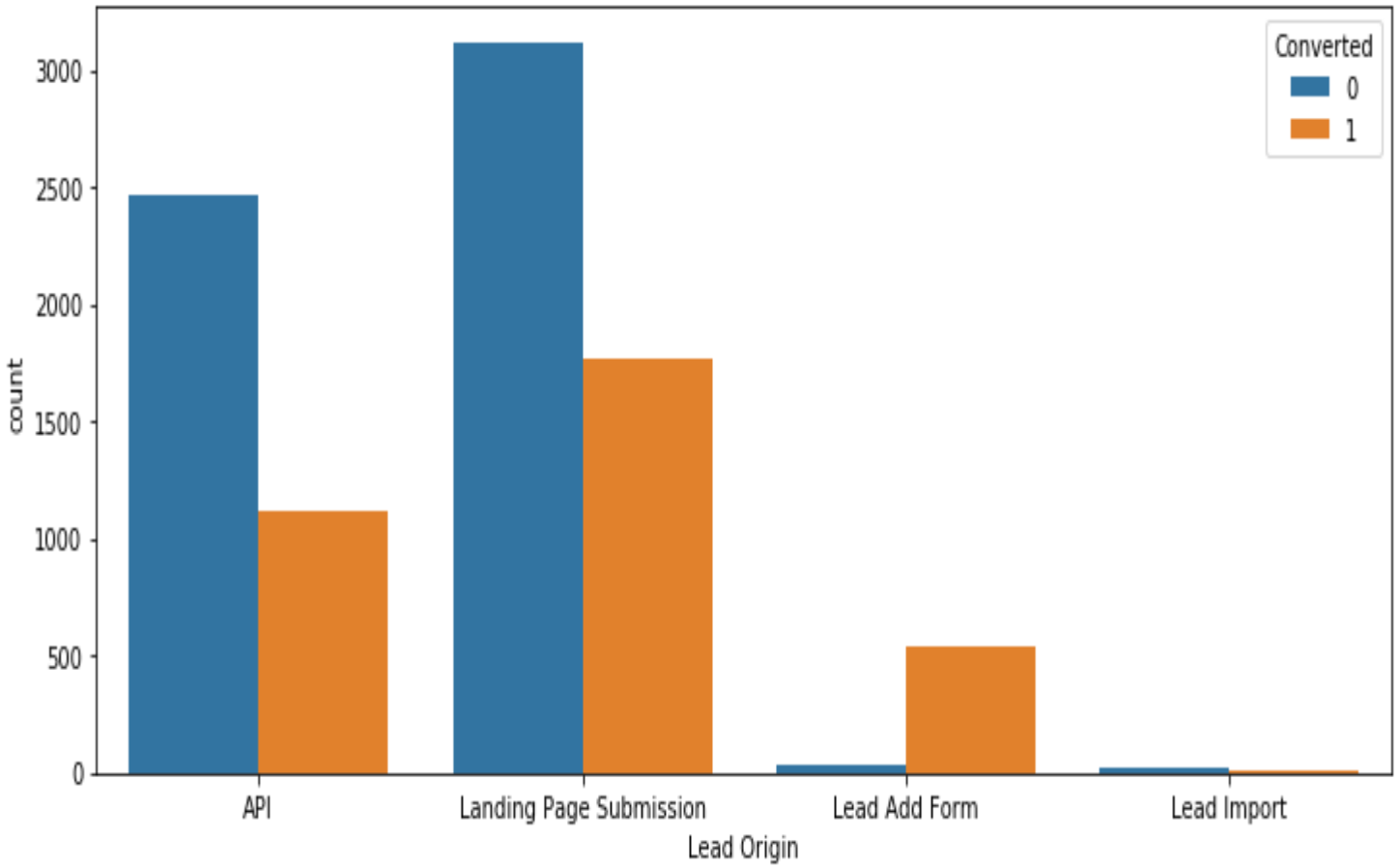


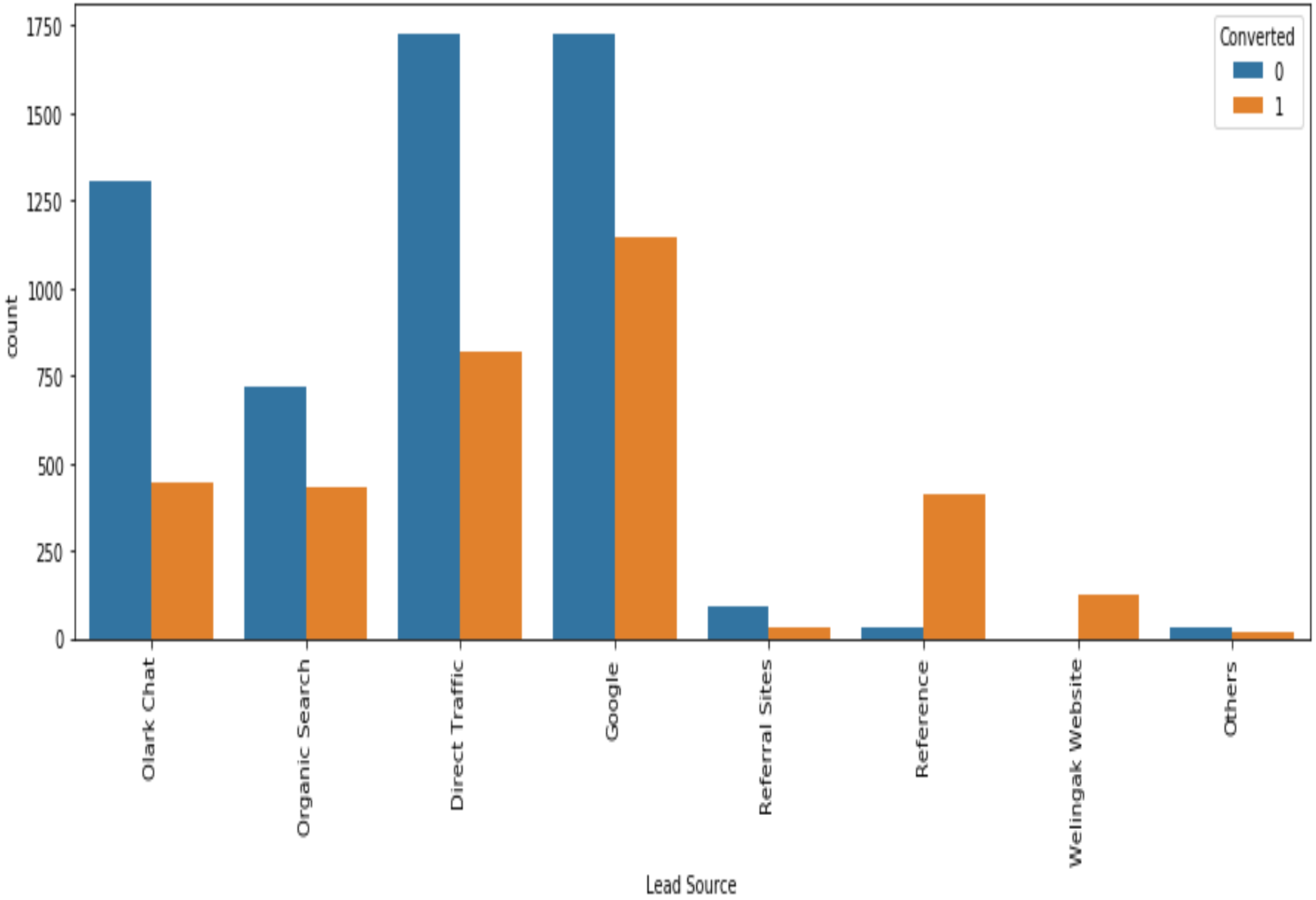
# Lead Origin Plot

Maximum number of leads are of category 'Landing Page Submission' followed by category 'API'.

In both the category 'Landing Page Submission' and 'API' the conversion is less than 40%.

Category 'Lead Add Form' has conversion around 90%, but the number of leads are very less.



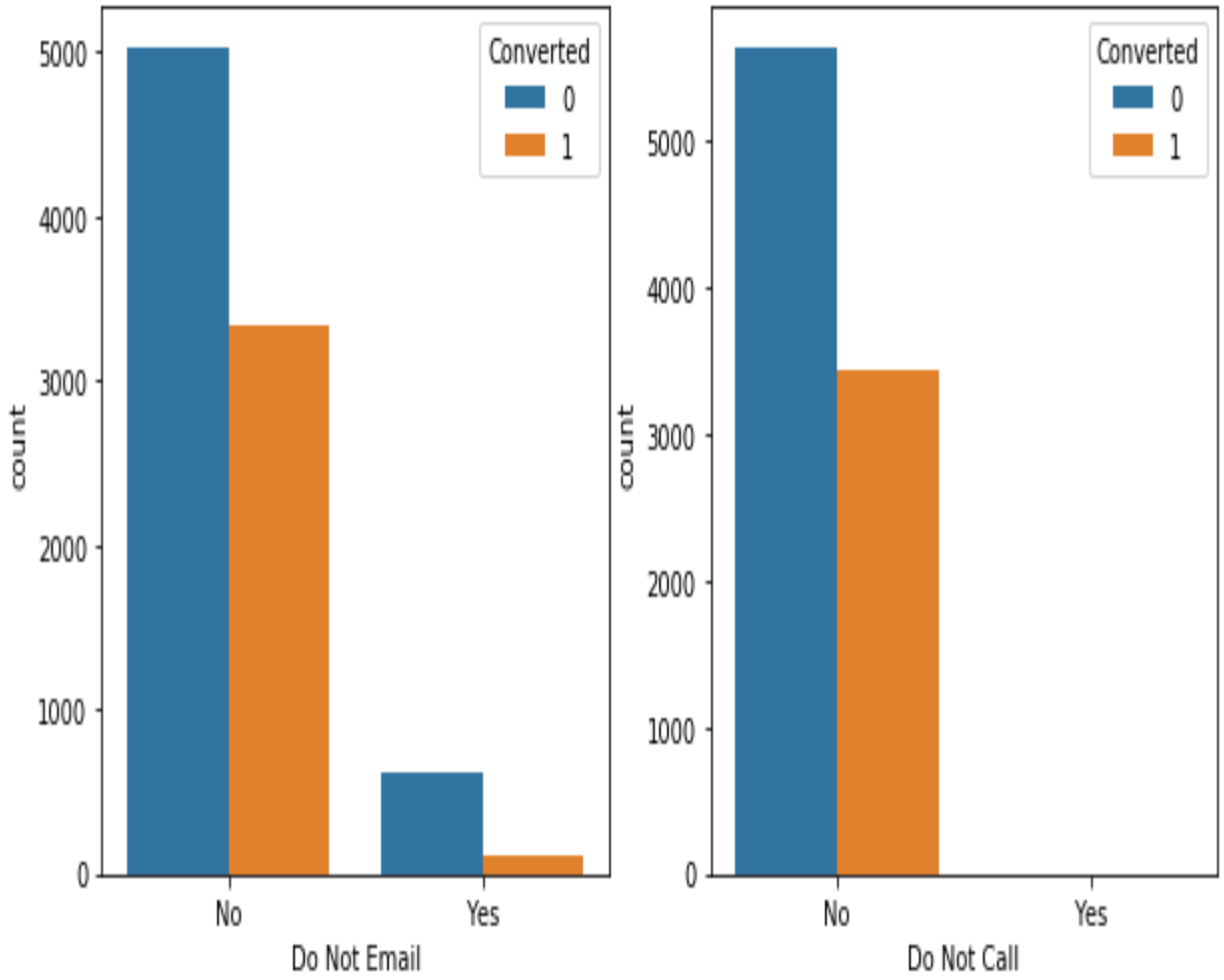



# Conversion from Lead Sources

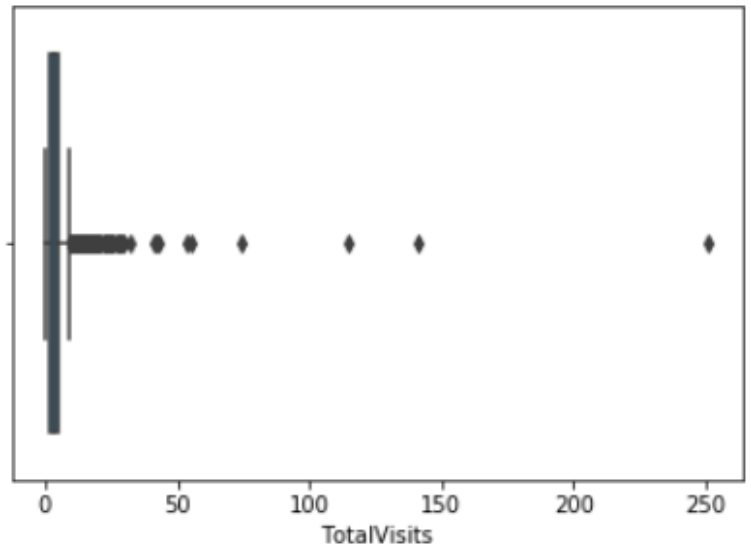
Maximum number of leads are of category 'Google' followed by category 'Direct Traffic' and 'Olark Chat'.

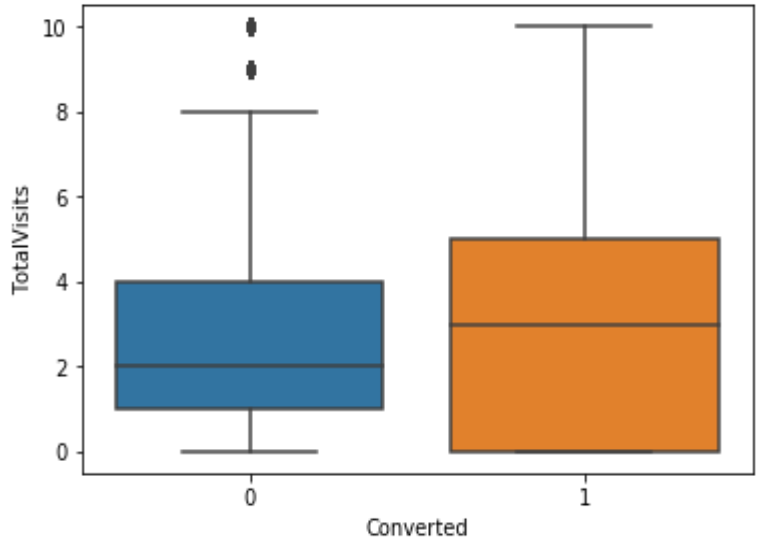
But the conversion rate in the these above categories are less than 40%.

Category 'Reference' and 'Welingak Website' has the conversion rate of around 90%, but total count of leads are very small for these categories.



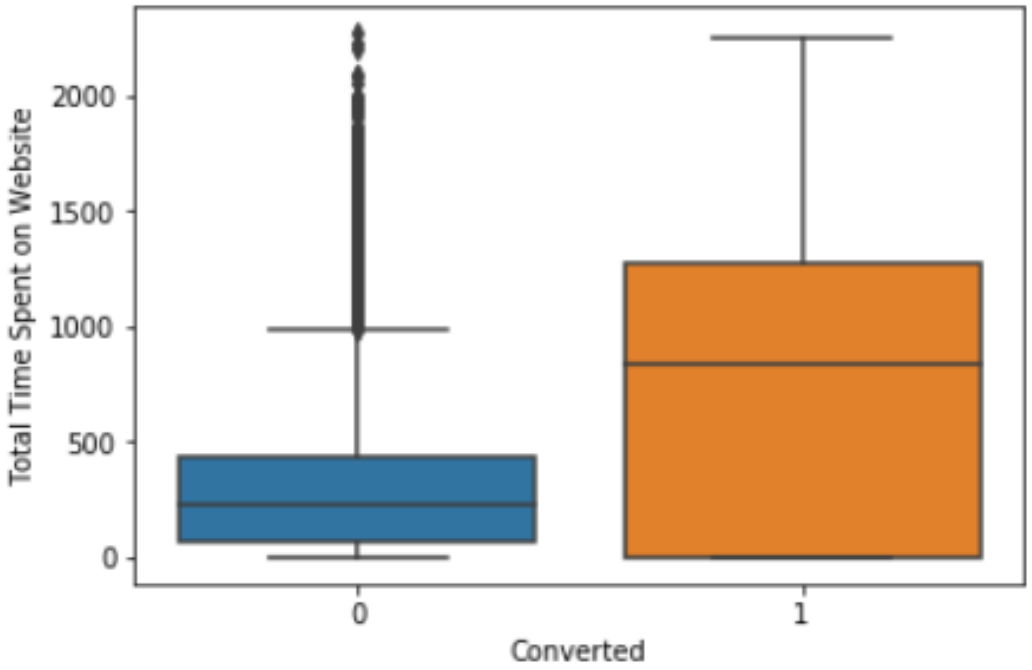

# Conversion of Don’t Email & Don’t Call

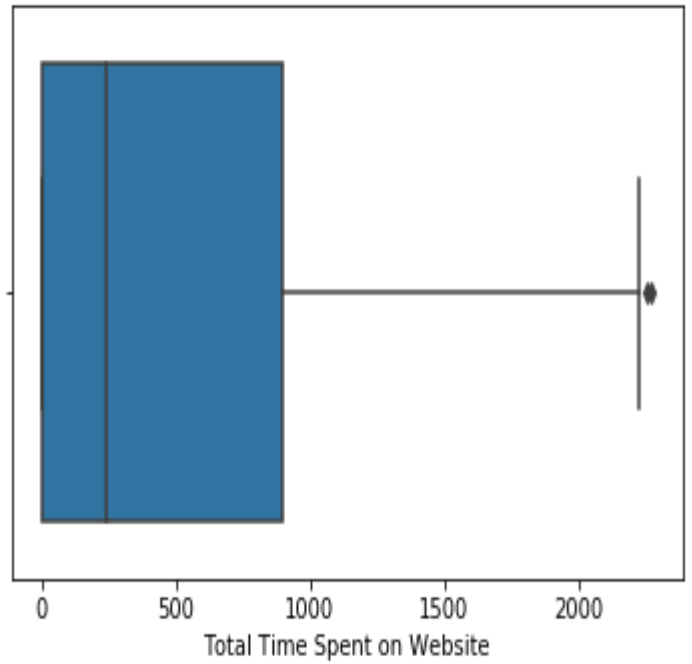


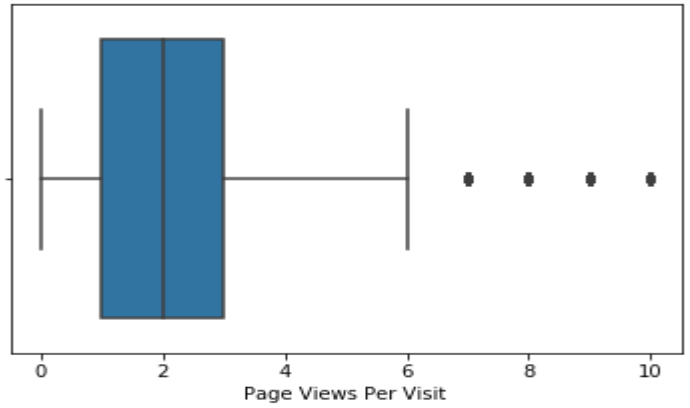

# Total Visits Vs Conversion

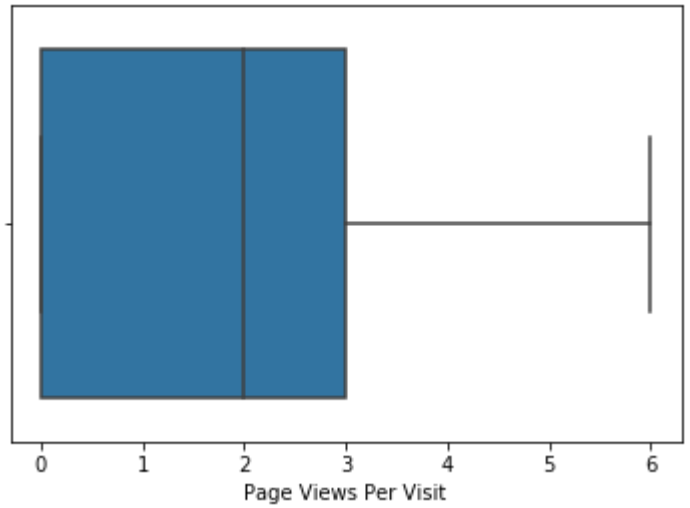
1. From the above graph we could see that median of the Converted are little higher than that of Not Converted.

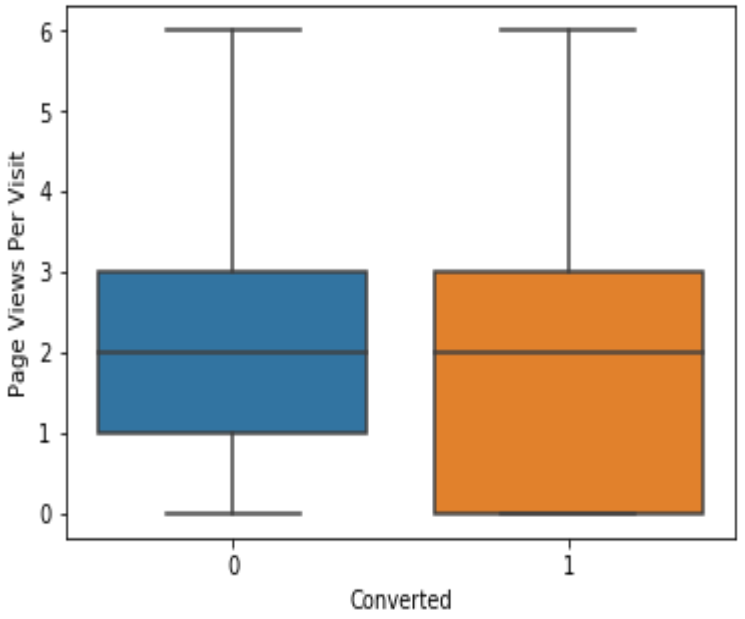



# Conversion based on Total Time Spent on Website

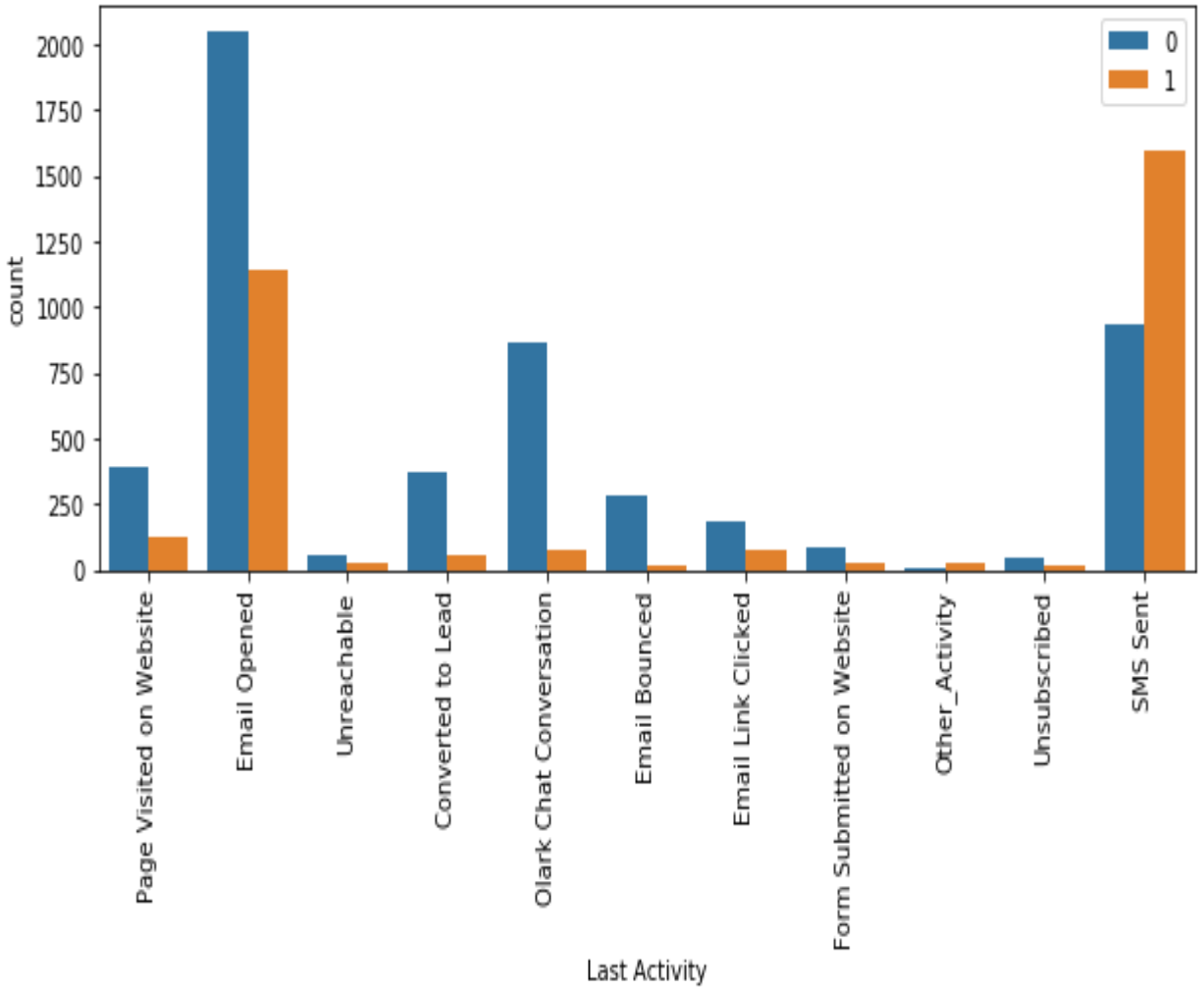


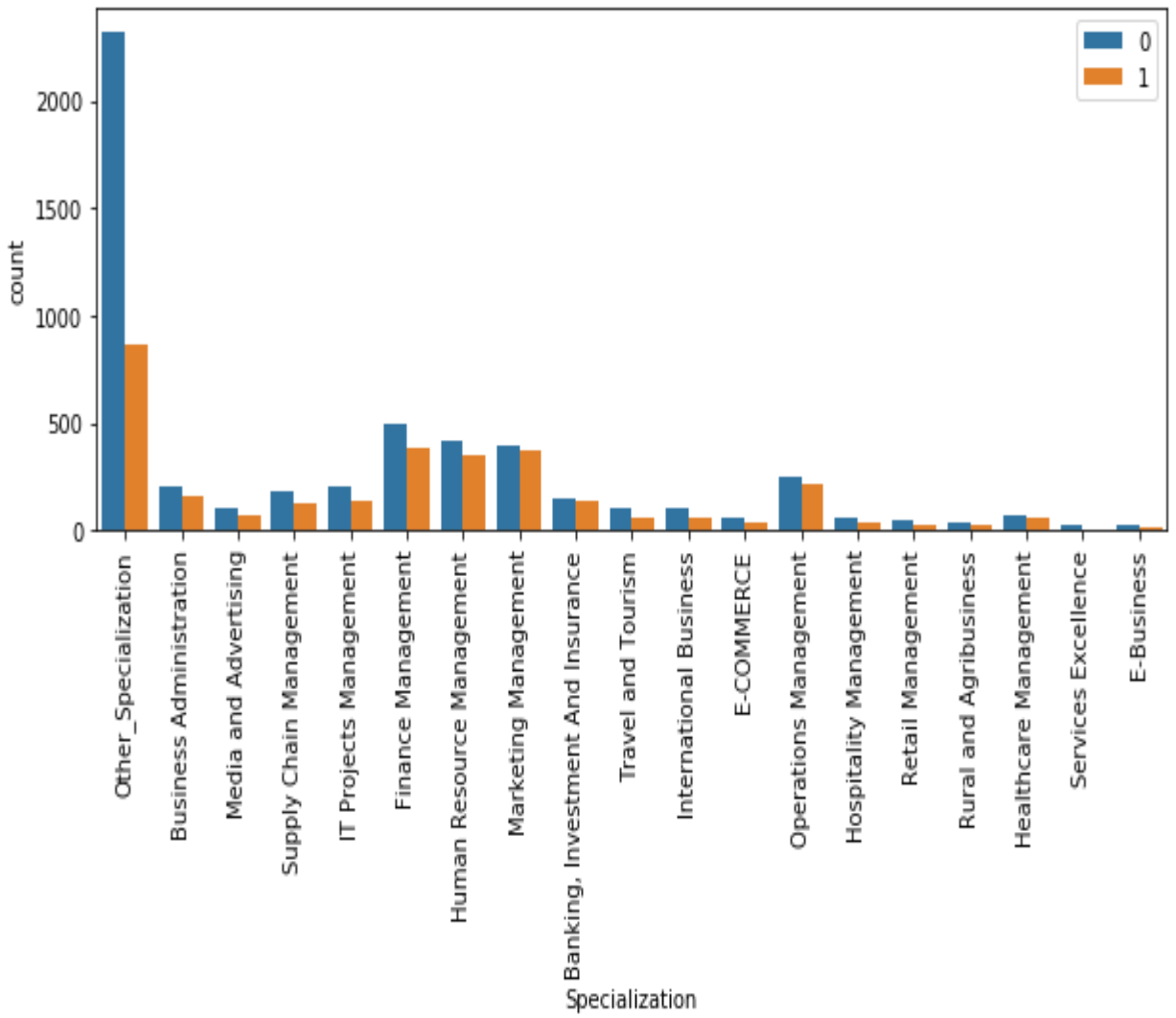

# Conversion based on Page Views Per Visits

1. The median of the variable 'Page Views Per Visit' for both the converted and non converted leads are same.



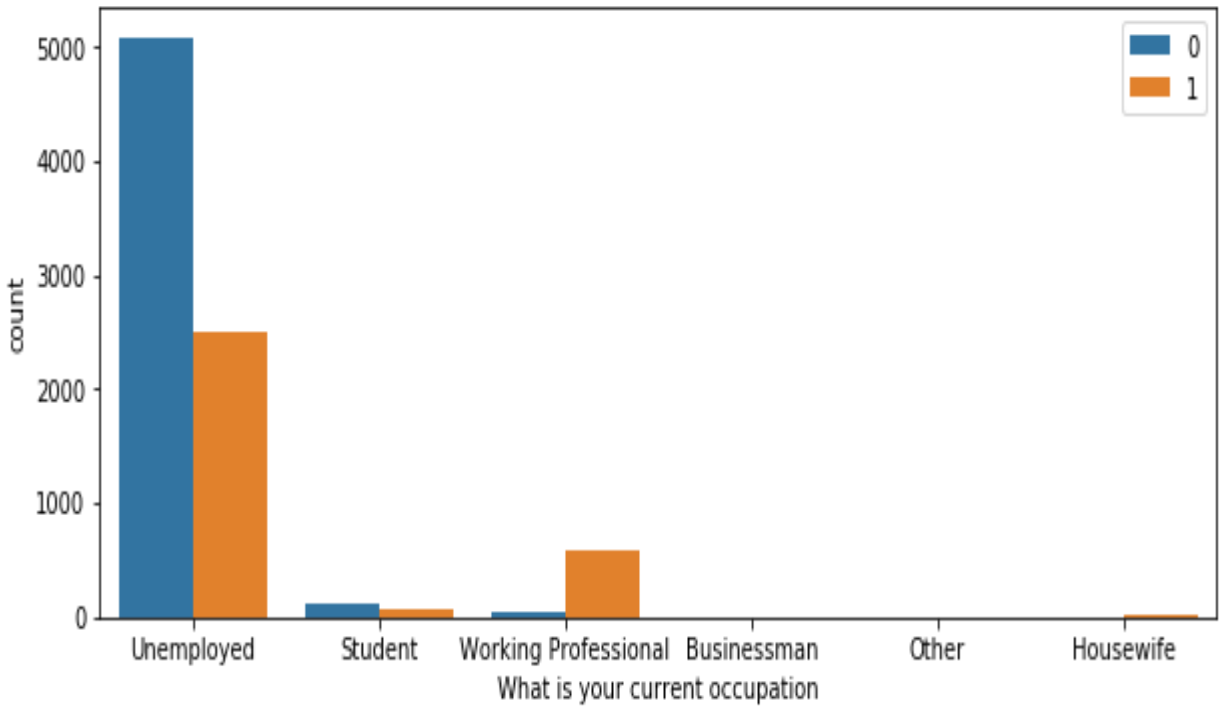

# Last Activity - Converted

1. - Most of the leads have Email Opened as Last Categories.
2. - Maximum Conversion is for the last activity being SMS Sent.




# Specialization Vs Conversion

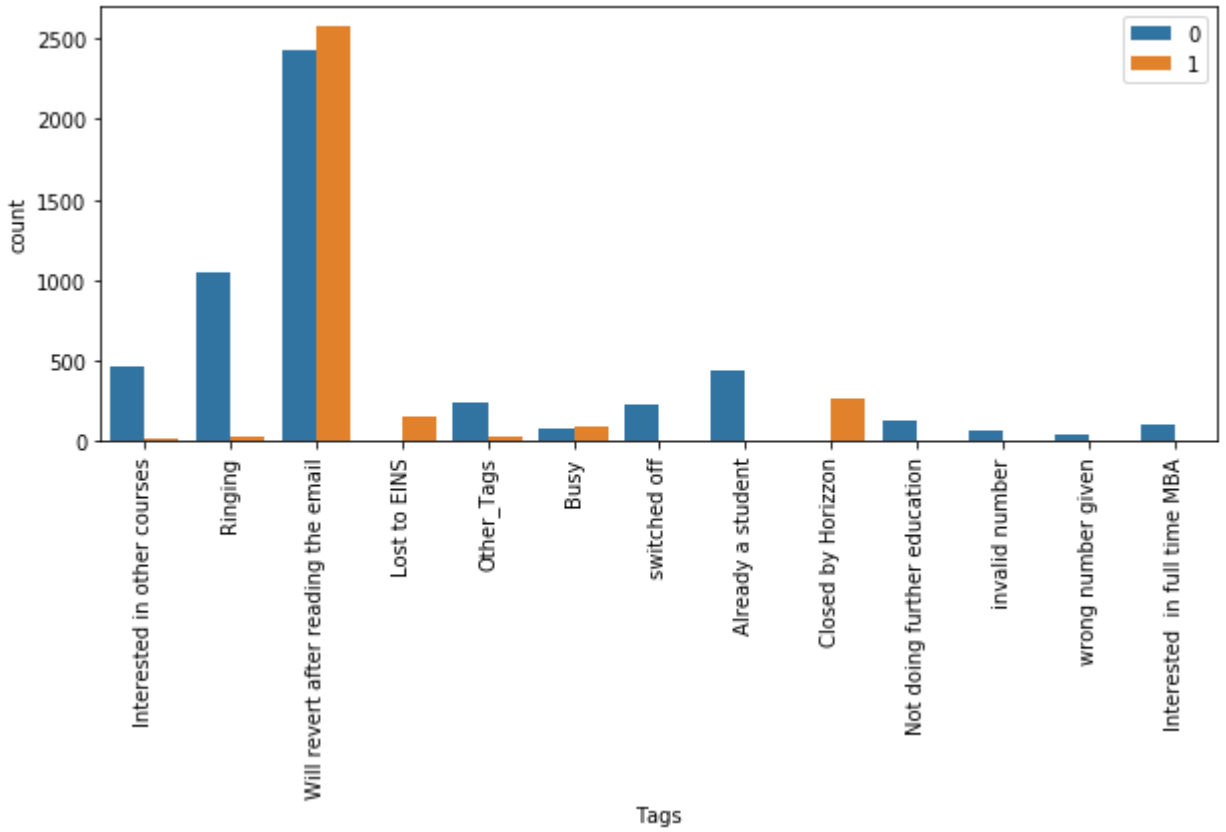
- Here we could see that count for the category 'Other_Specialization' is maximum but the for the variable Specialization we should be focusing more on the variable having most conversion rate.




# Conversion Vs Current Occupation

From the above graph we could see that the maximum leads are Unemployed.

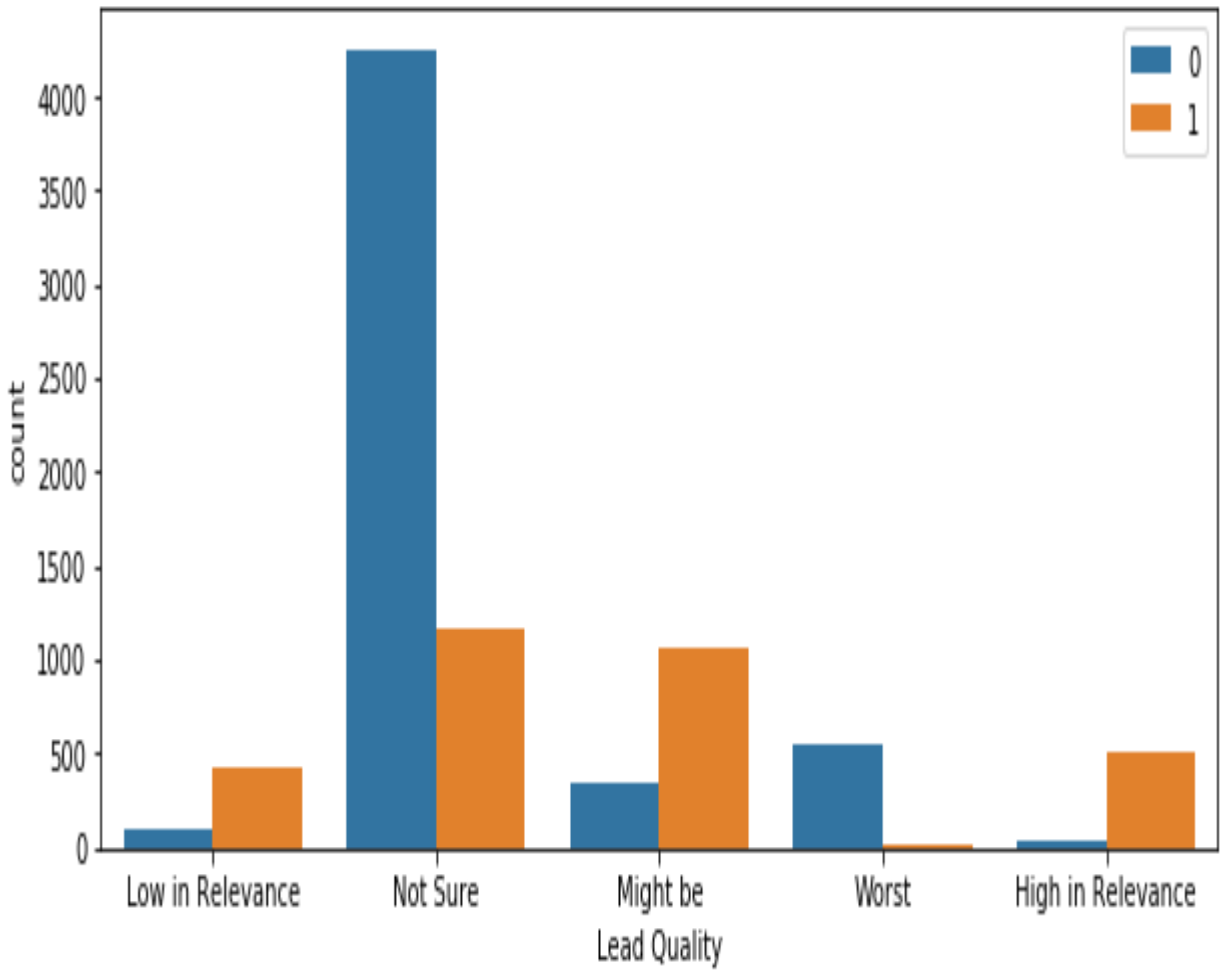
But Maximum rate of conversion is for the Working Professional.




# Tags Vs Conversion

Maximum number of leads are of those who said that they will revert after reading the email and those are also maximum number of converted leads.

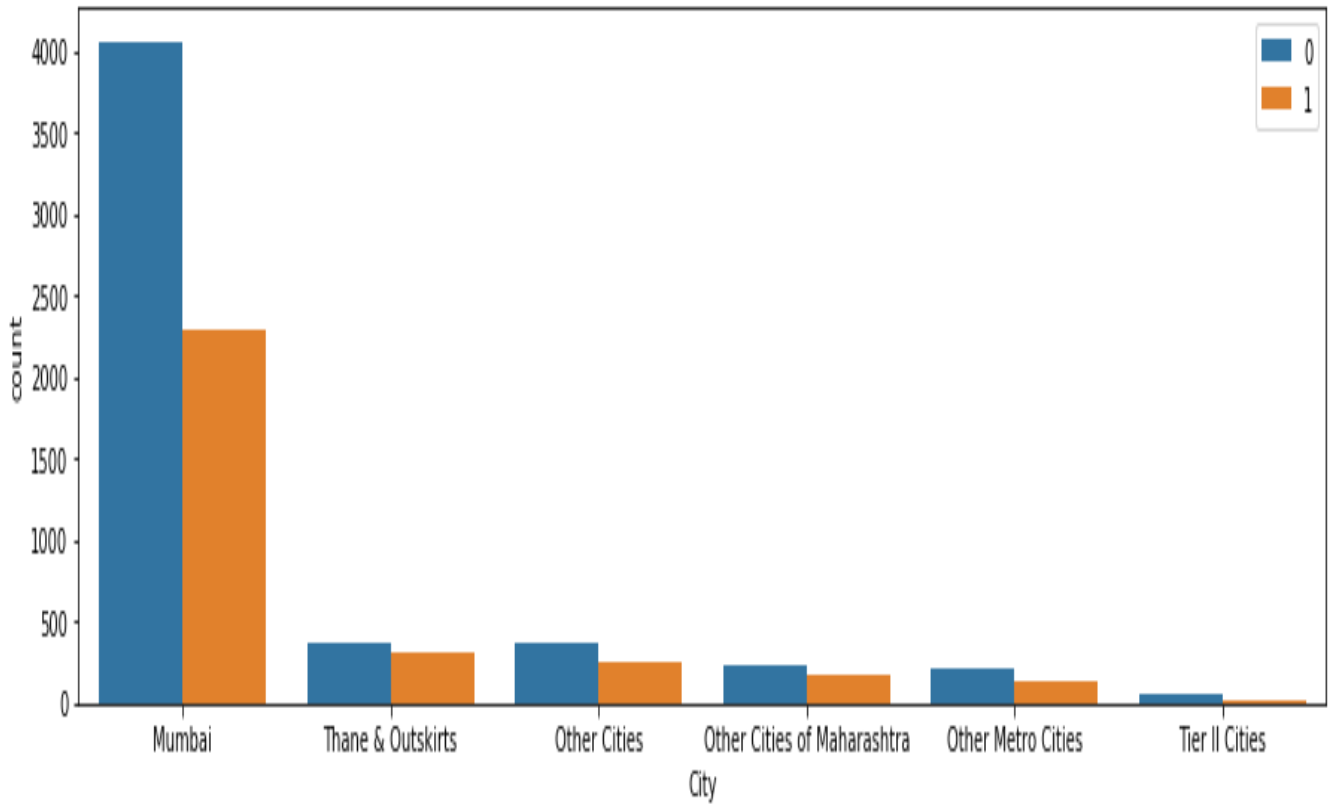
Maximum percentage of conversion is for the category 'Closed by Horizon' and 'Lost to EINS'.




# Lead Quality Vs Conversion

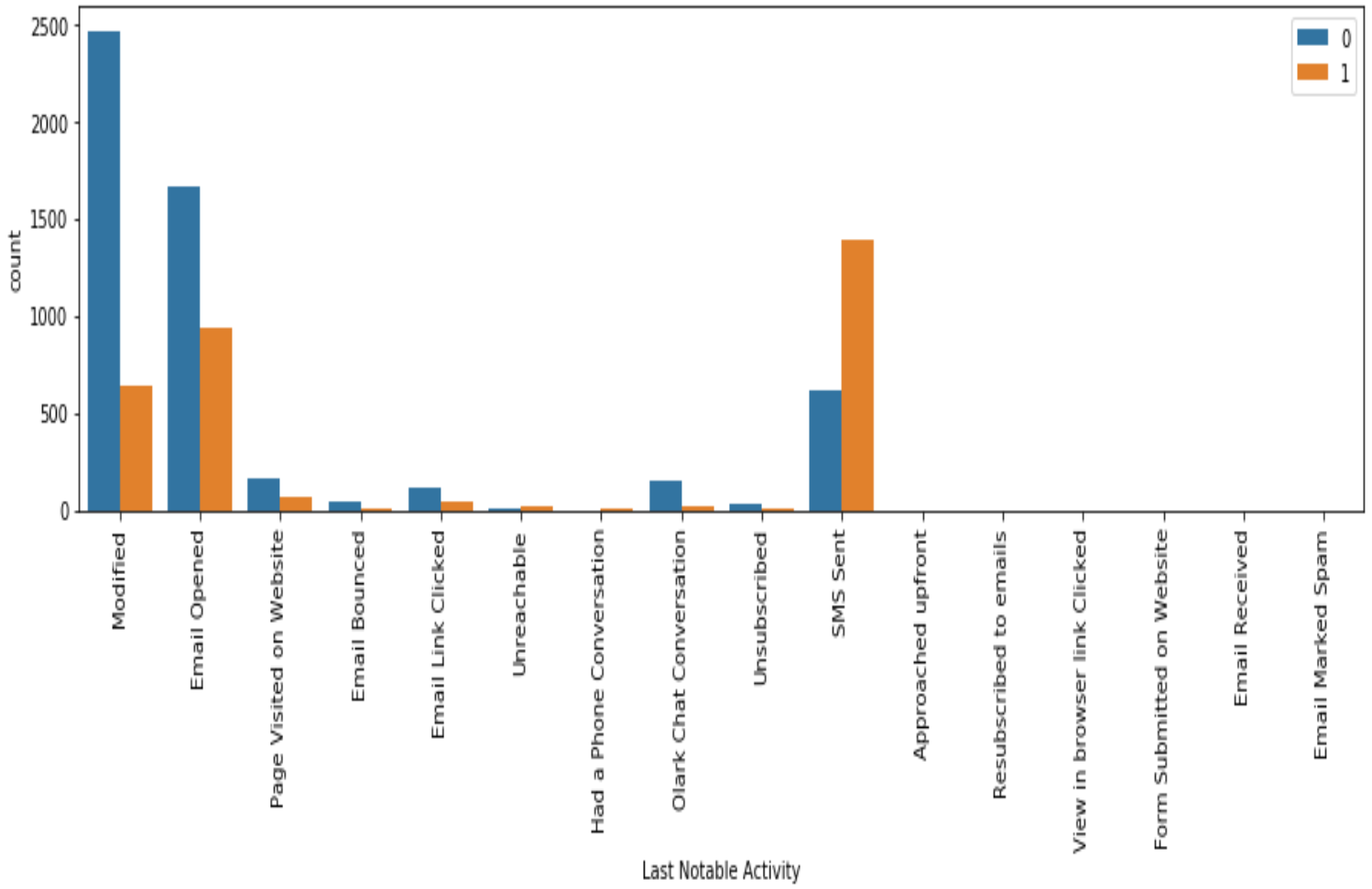
In this variable you could see that the maximum counts of lead is where Lead Quality is Not Sure. But its conversion percentage is very low.

Conversion rate of Category High in Relevance is highest followed by Might be. But the count of these categories are small.




# Conversion in Cities

1. Most leads are from Mumbai with conversion rate around 30%.



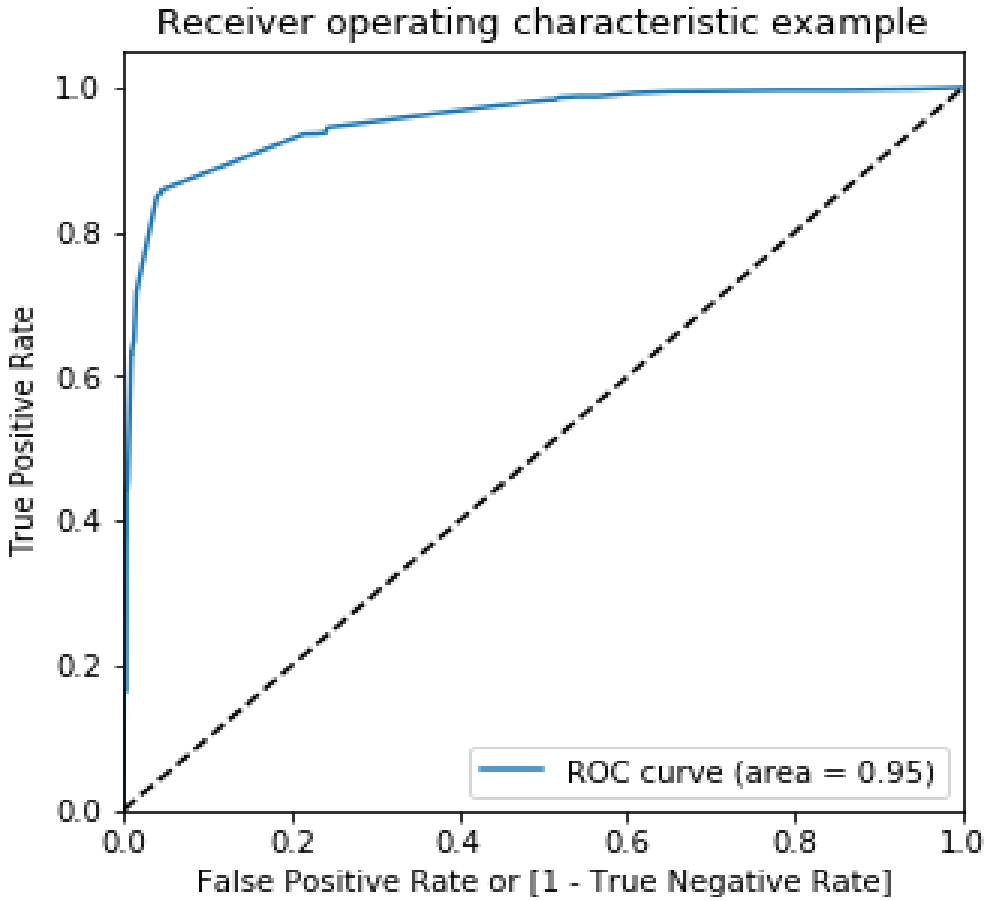

# Last Notable Activity Leading into Conversion

Here we could see that the maximum leads are of 'Modified', but its conversion rate is very small.

The conversion rate of 'SMS Sent' category is maximum.

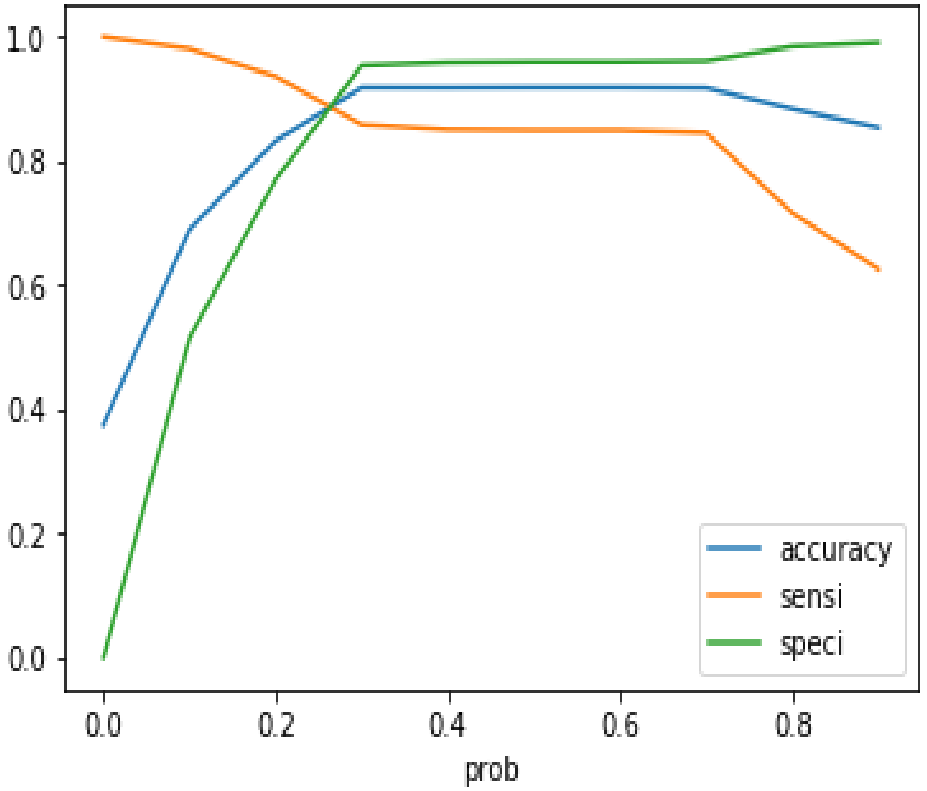
- From the above Univariate analysis we are able to find out few variables which are not helping in deciding whether the lead will convert or not.

From the above Univariate analysis we are able to find out few variables which are not helping in deciding whether the lead will convert or not.



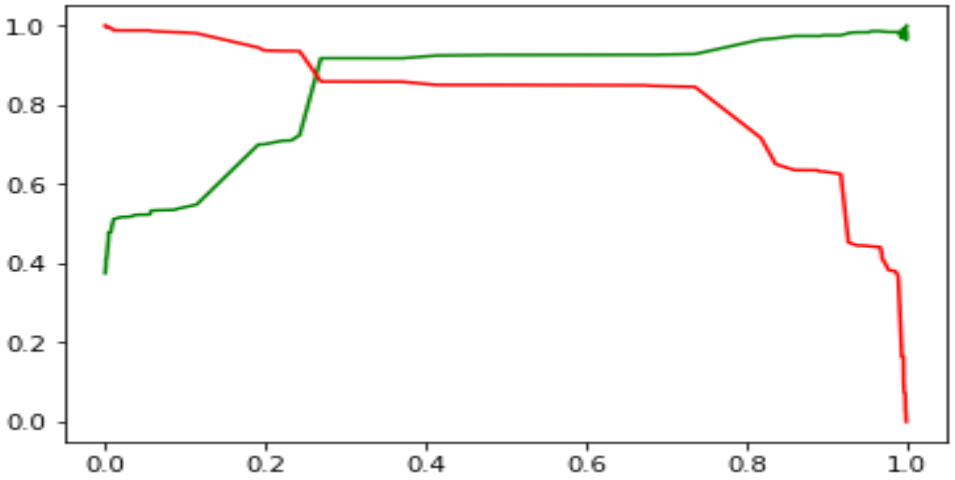

# ROC Curve

- - Optimal Threshold




# Accuracy, Specificity & Sensitivity

- From graph we could say that optimum point to take as cutoff is approx 0.25




# Precision & Recall Thresholds

# Key Values

OVERALL ACCURACY - .91

SENSITIVITY OF OUR LOGISTIC REGRESSION MODEL - .86

SPECIFICITY - .95

FALSE POSITIVE RATE - PREDICTING CHURN WHEN CUSTOMER DOES NOT HAVE CHURNED - .046

POSITIVE PREDICTIVE VALUE - .917

NEGATIVE PREDICTIVE VALUE - .918

PRECISION - .917

RECALL - .8493


THANK YOU