



# LEAD SCORING CASE STUDY

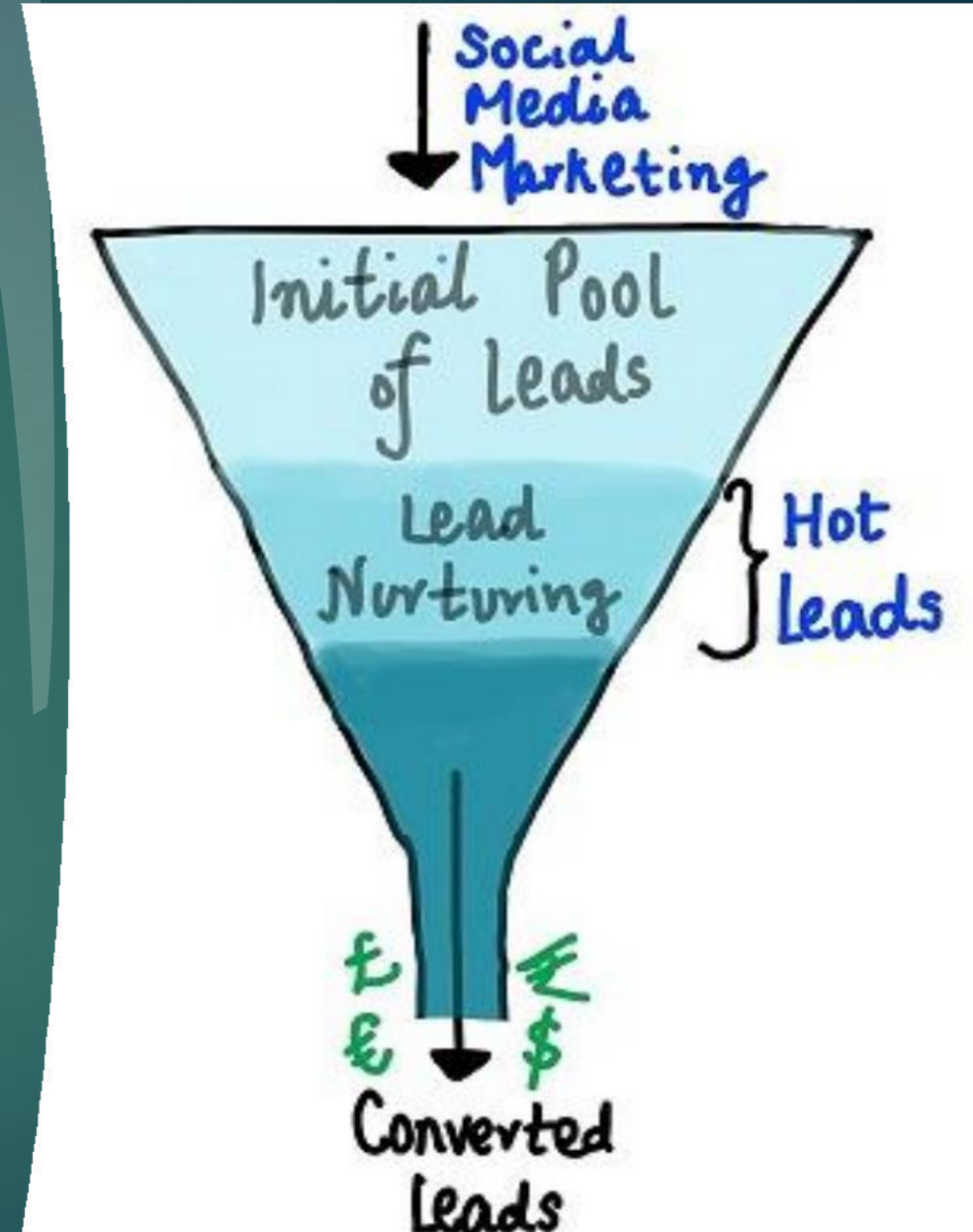
WITH LOGISTIC REGRESSION

BY- TAMANNA, SAKSHI RAI  
RAVI VANANEE

,

## Problem statement

- ▶ NAME OF COMPANY- X EDUCATION. The company markets and sells online course to the on several platforms to the industry professionals like on the several search engine such as google.
- ▶ The interest is to not only focus on the hot, i.e, our potential leads but our lead to the conversion, the leads that comes to the company data due to their referrals and several other sources like browsing.
- ▶ As lead conversion rate is 30% at x education.



# BUSINESS GOAL

- ▶ The company requires build a model where we assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- ▶ The model to be built in lead conversion rate should be around 80% or more.

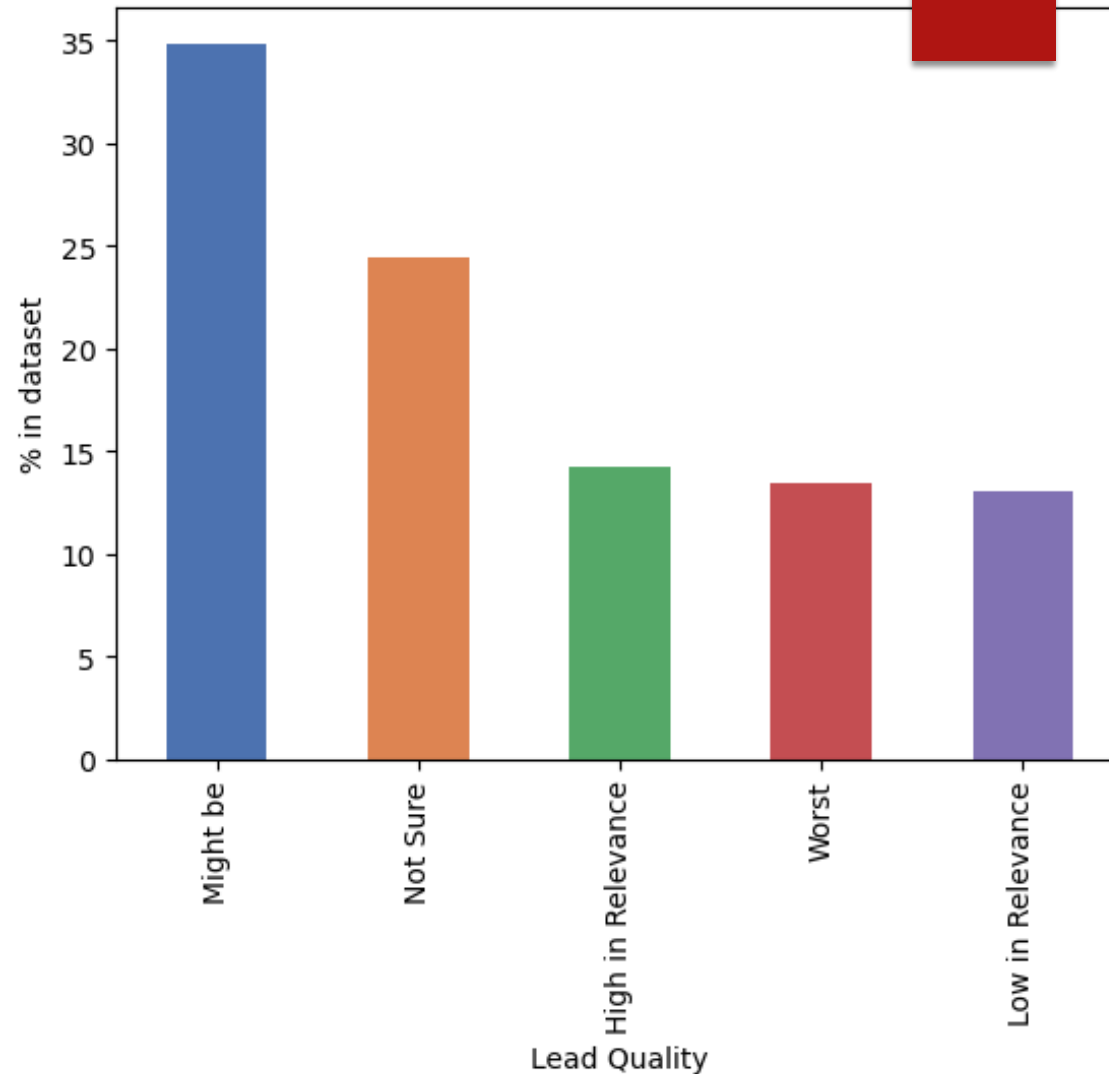
# Approach:

- ▶ Import data.
- ▶ Clean and prepare the data for further analysis.
- ▶ EDA for most helpful attributes for conversion.
- ▶ Scaling features
- ▶ Data preparation for model building
- ▶ Assigning lead score for each lead
- ▶ Model test on train set
- ▶ Model evaluation
- ▶ Model accuracy measurement

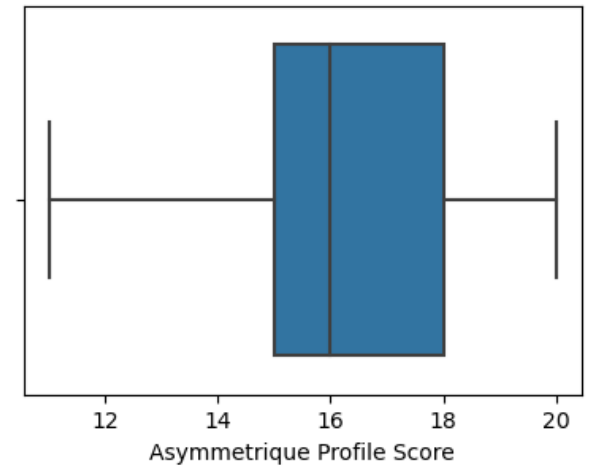
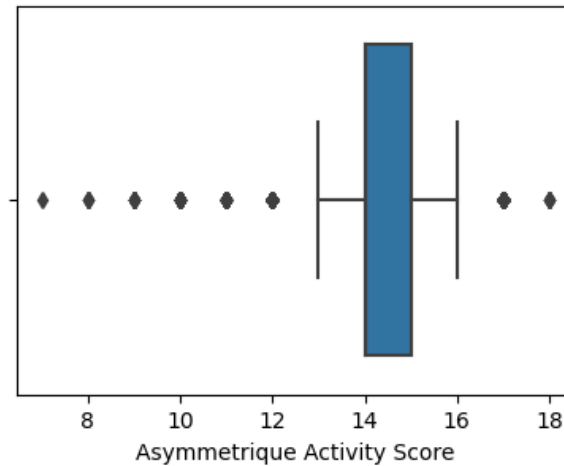
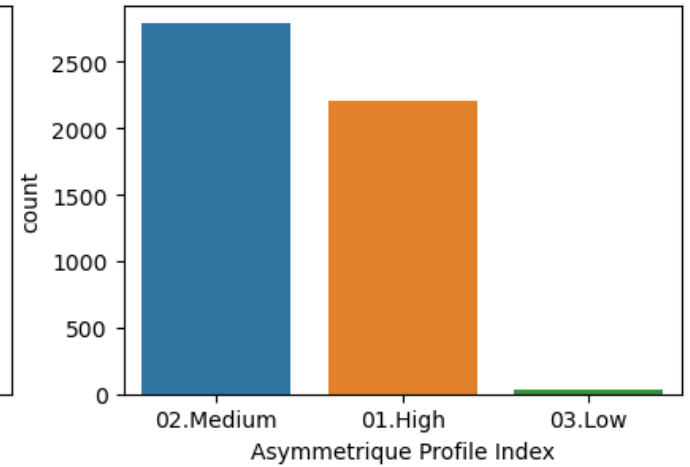
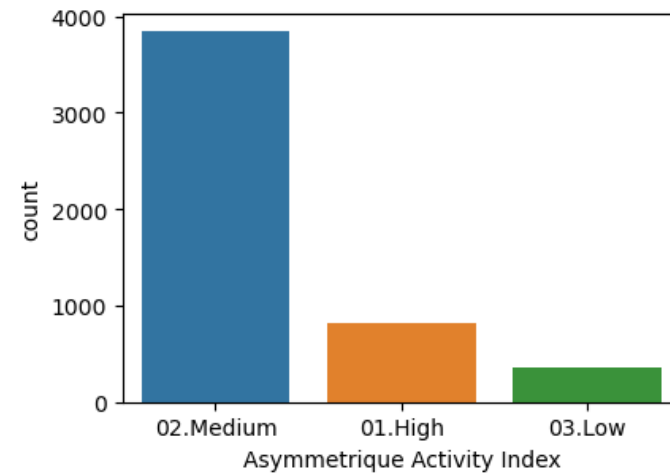


# EDA ANALYSIS:

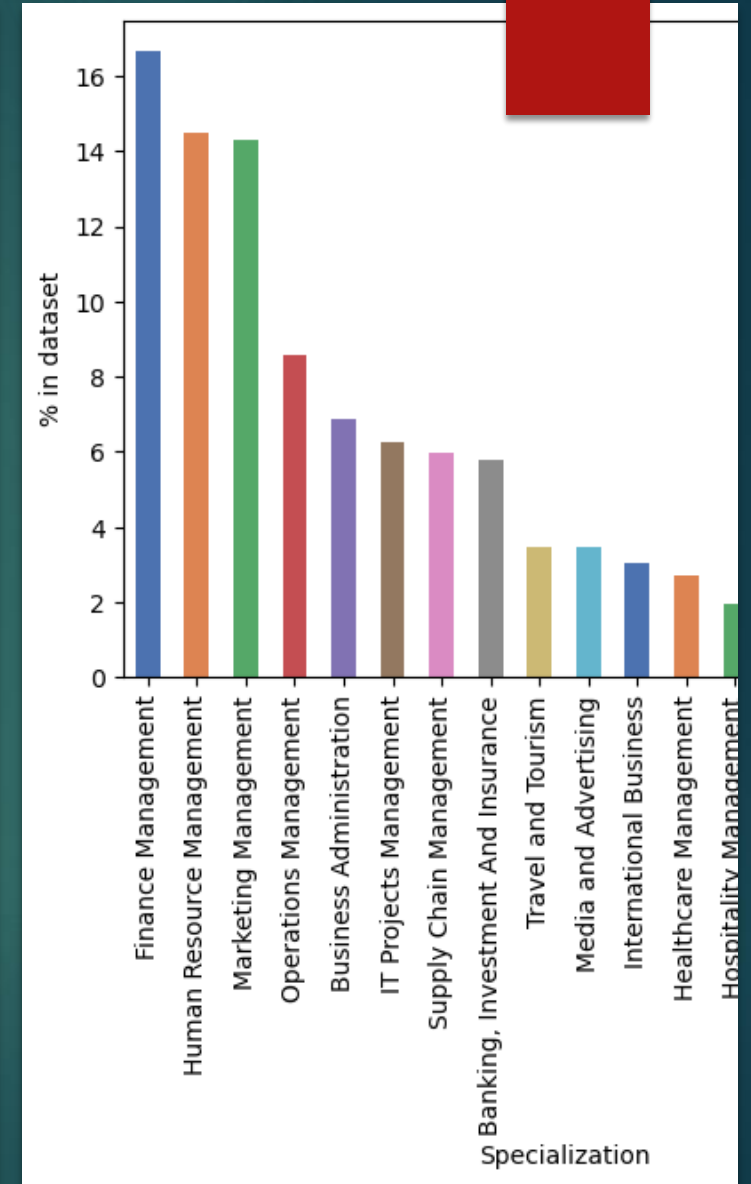
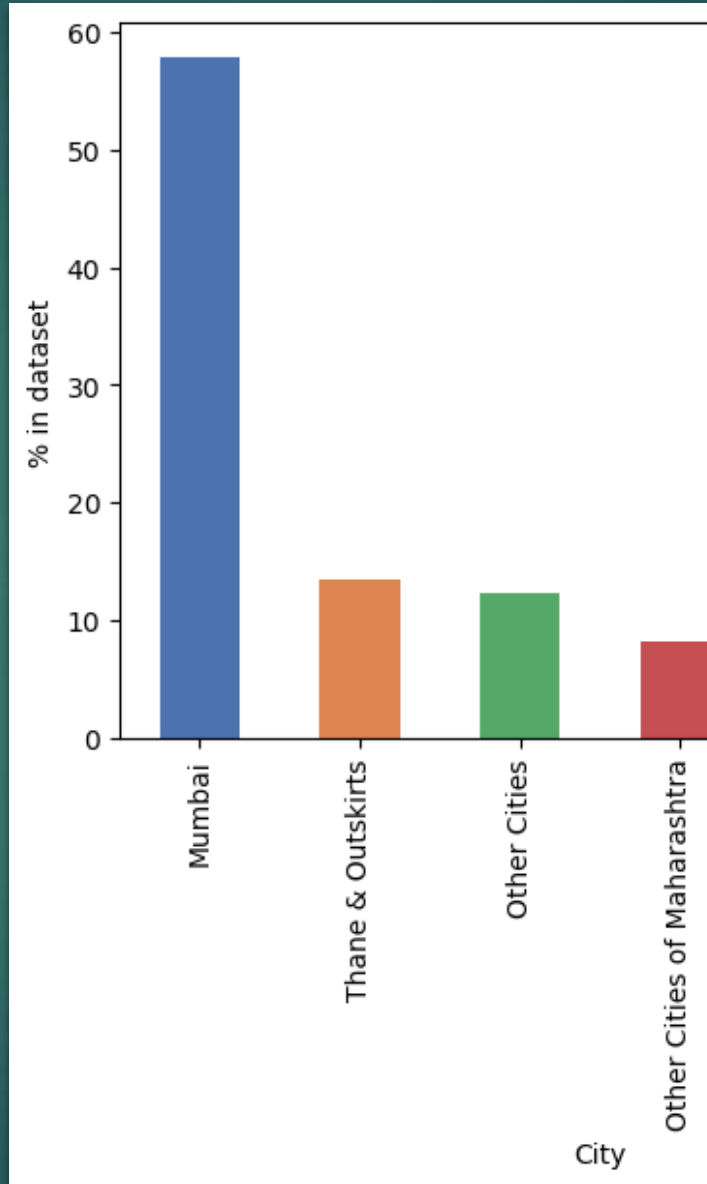
- ▶ Null values in the 'Lead Quality' column can be imputed with the value 'Not Sure' as
- ▶ we can assume that not filling in a column means the employee does not know or is not sure about the option.



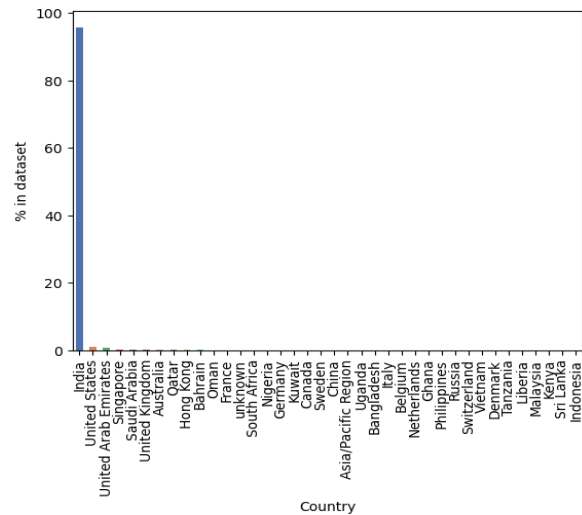
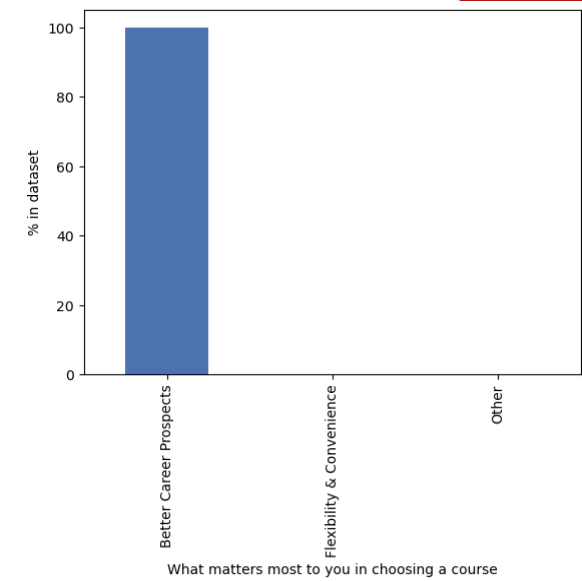
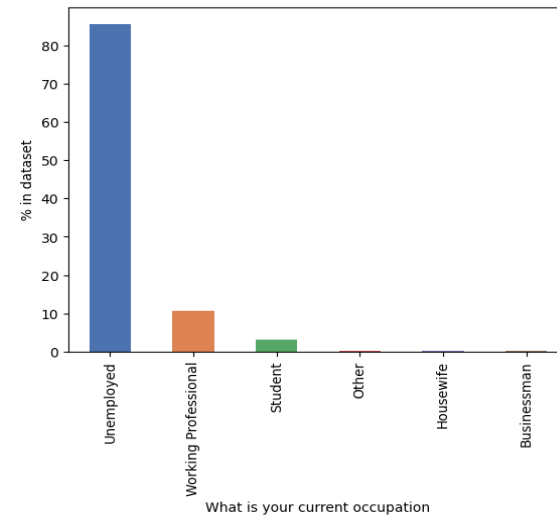
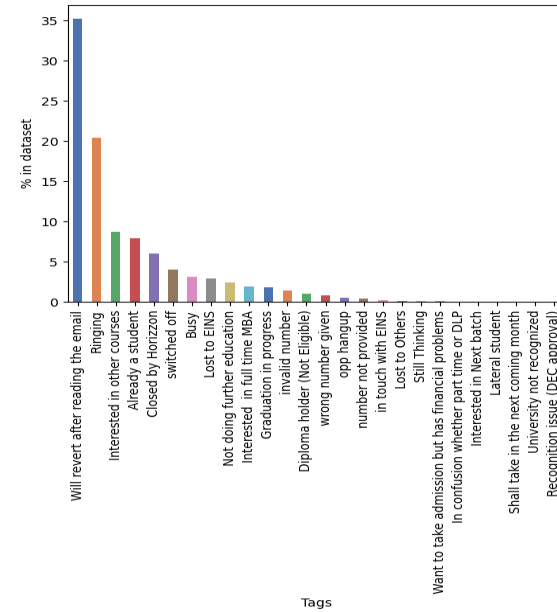
- ▶ These four variables have more than 45% missing values and it can be seen from the plots
- ▶ that there is a lot of variation in them. So, it's not a good idea to impute 45% of the data.
- ▶ Even if we impute with mean/median for numerical variables, these values will not have any significant importance in the model.
- ▶ We'll have to drop these variables.



- ▶ Here Around 60% of the City values are Mumbai
- ▶ There are a lot of different specializations and it's not accurate to directly impute with the mean.

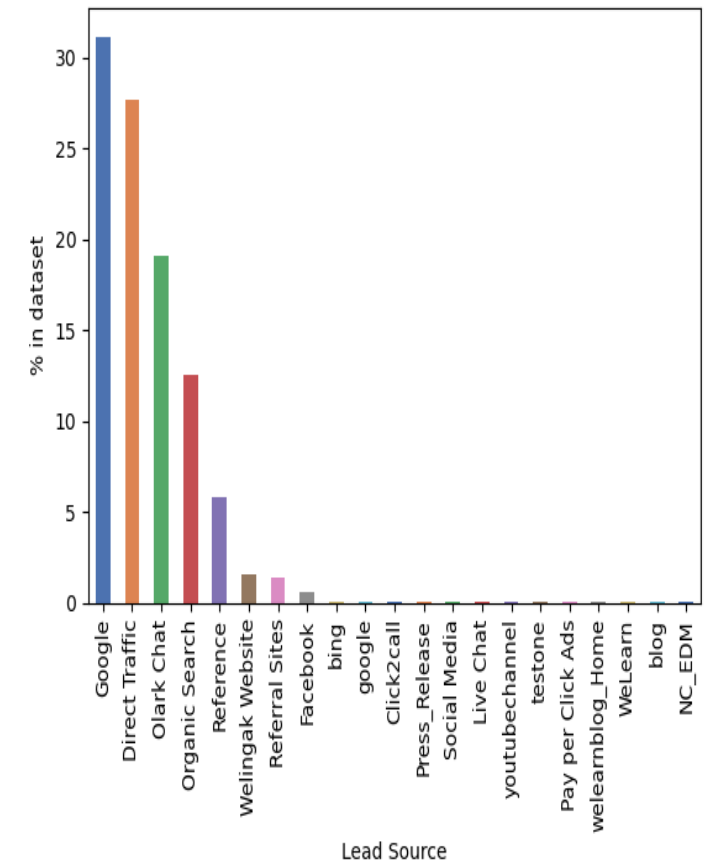
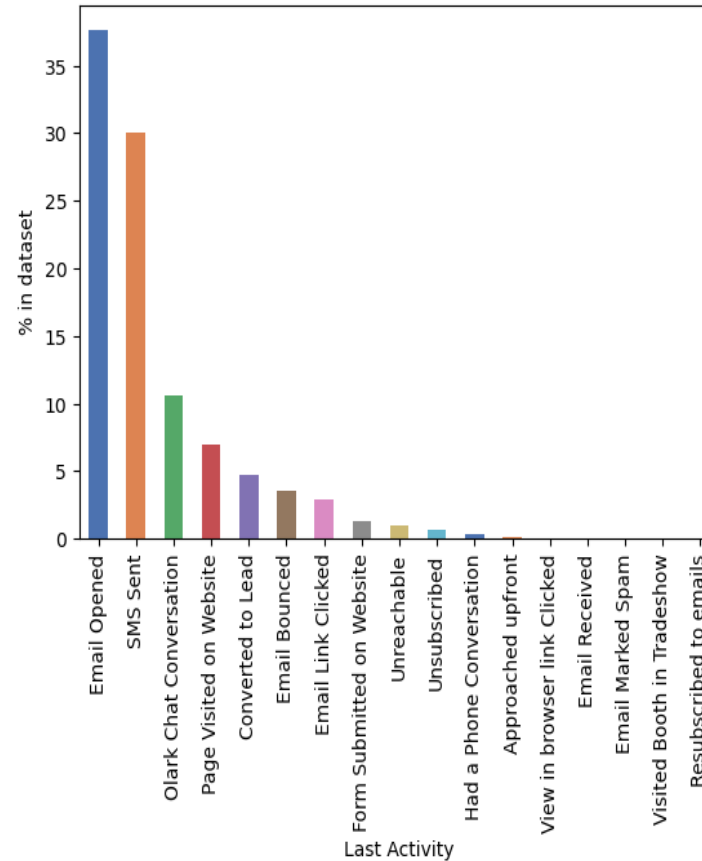


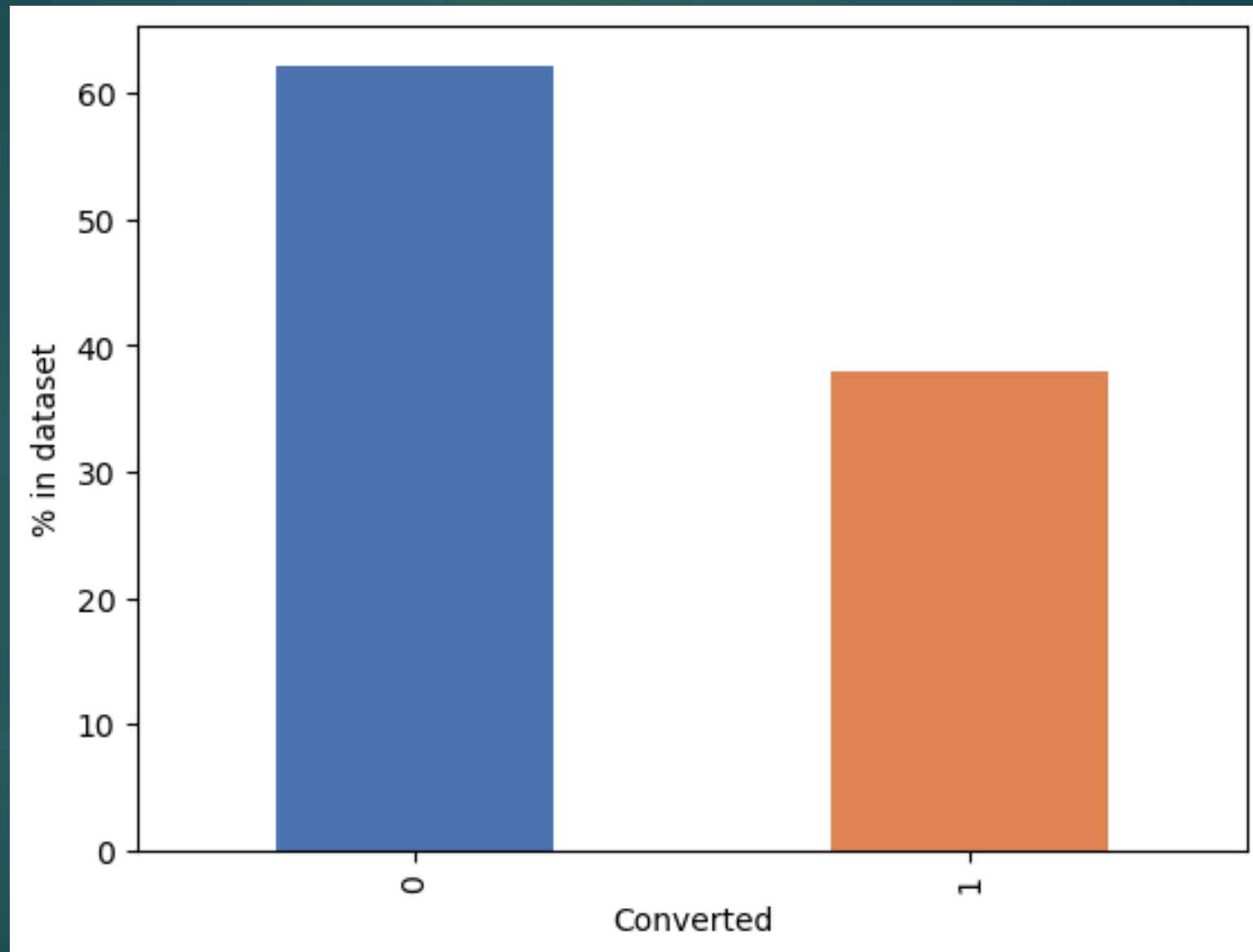
- ▶ In all these categorical variables, one value is clearly more frequent than all others. So it makes sense to impute with the most frequent values.



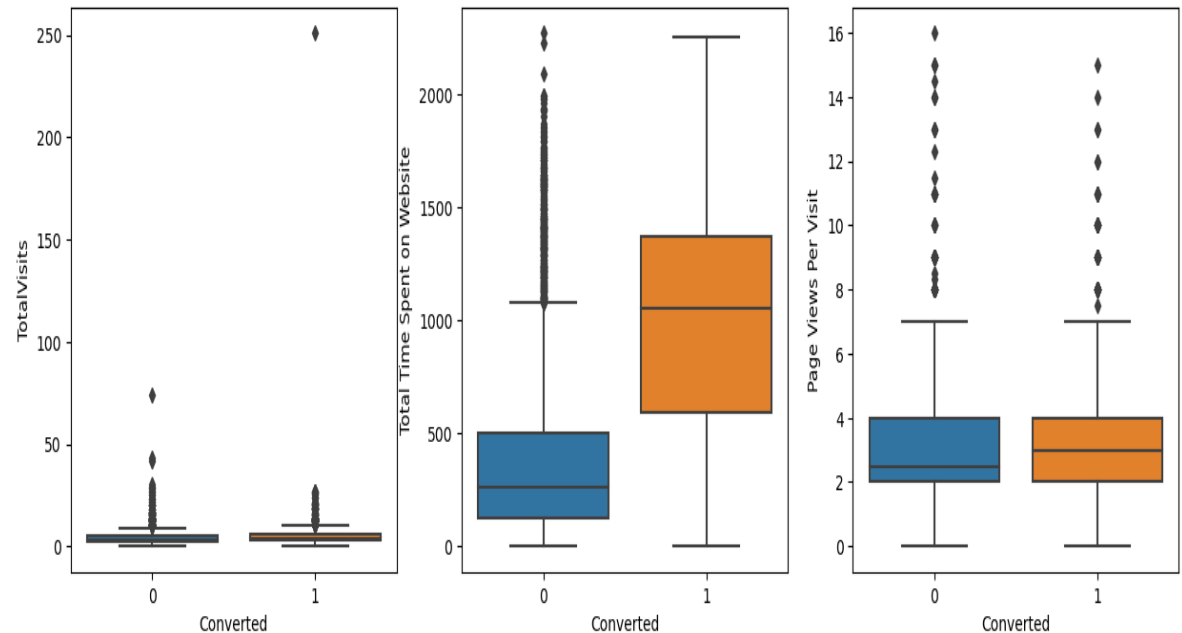


- ▶ In these categorical variables, imputing with the most frequent value is not accurate as the next most frequent value has similar frequency. Also, as these variables have very little missing values, it is better to drop the rows containing these missing values. Hence, we'll drop the rows containing any missing values for above four variables.

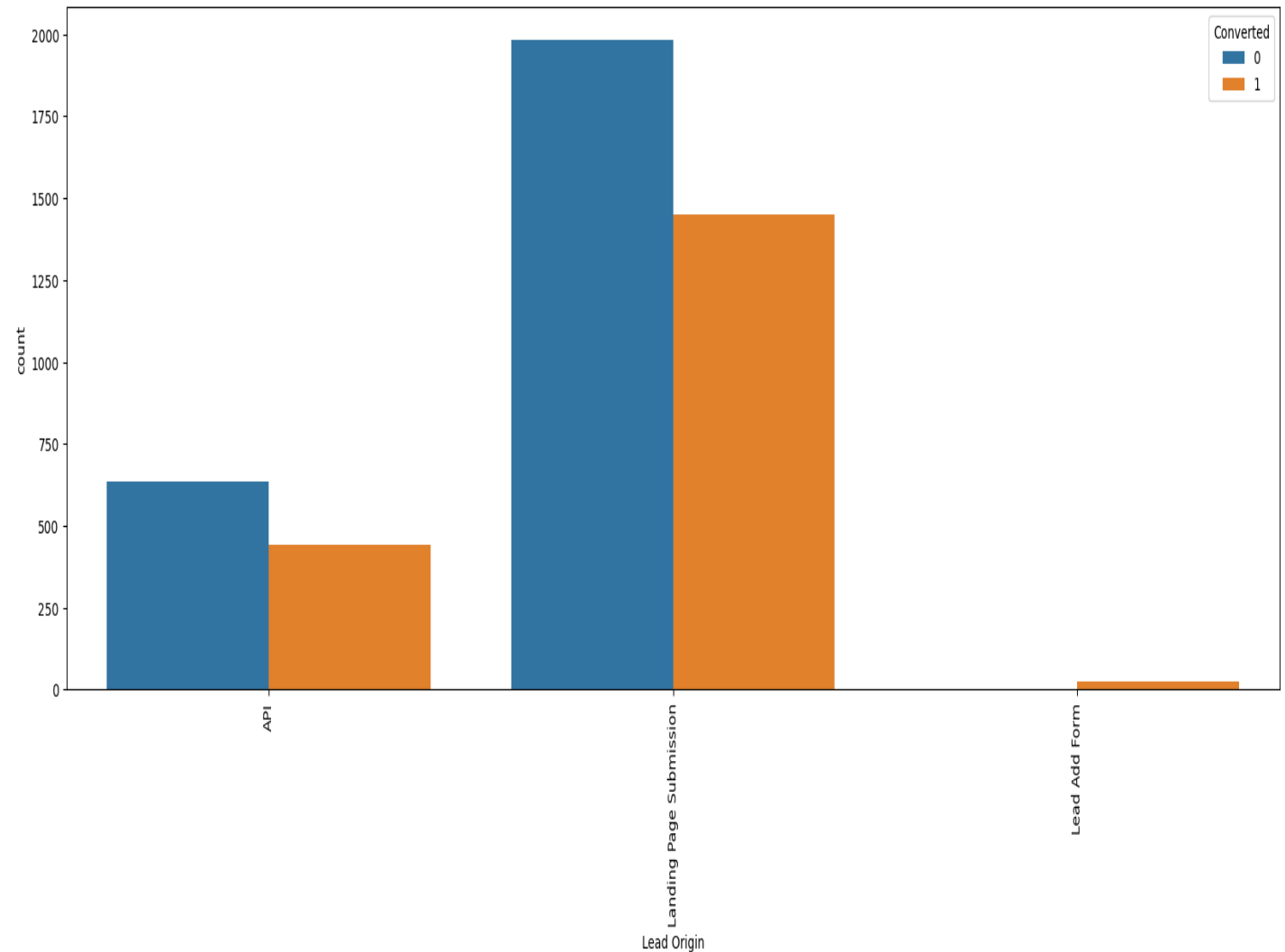




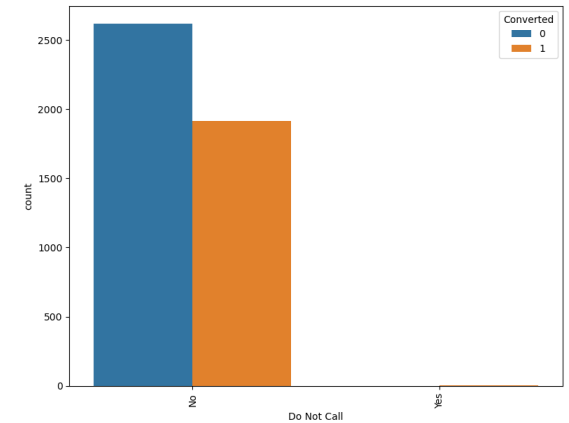
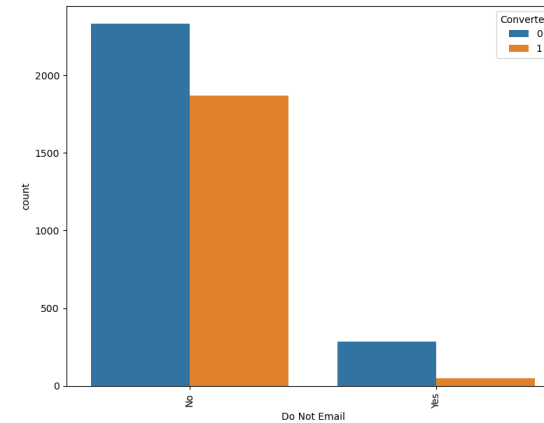
- ▶ Observations:
- ▶ 'Total Visits' has same median values for both outputs of leads. No conclusion can be drawn from this.
- ▶ People spending more time on the website are more likely to be converted. This is also aligned with our general knowledge.
- ▶ 'Page Views Per Visit' also has same median values for both outputs of leads. Hence, inconclusive.



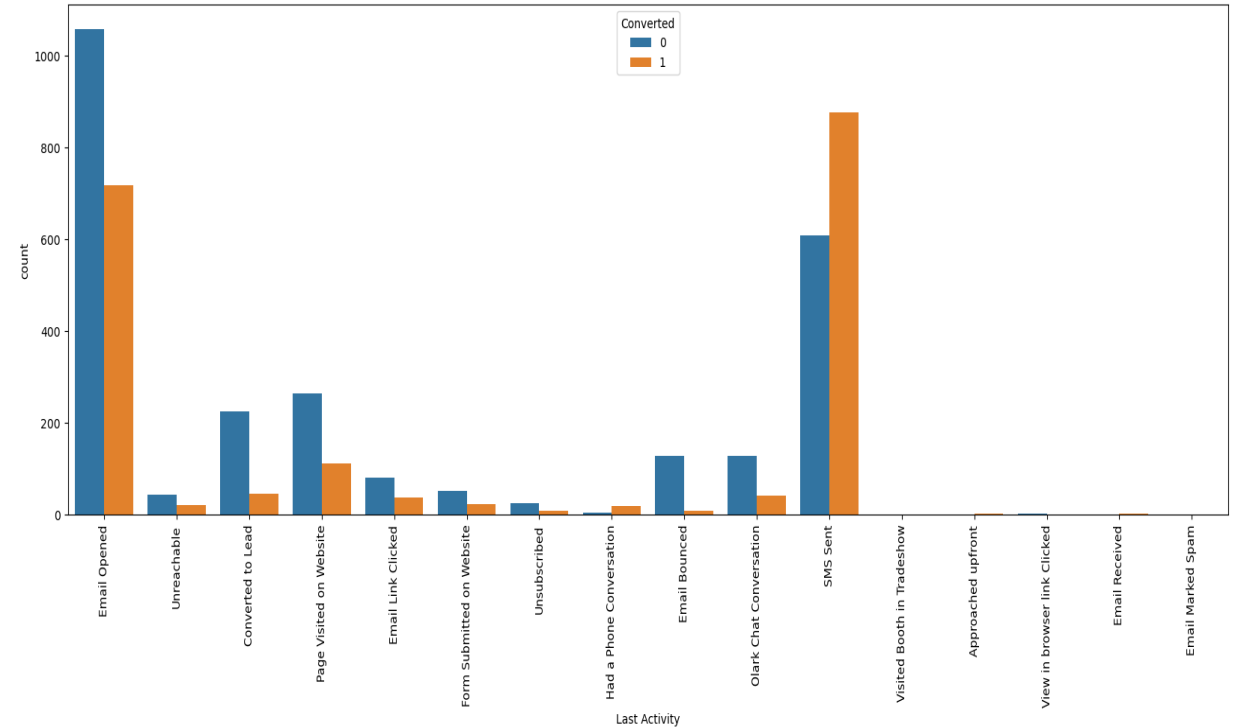
- ▶ Observations for Lead Origin :
- ▶ 'API' and 'Landing Page Submission' generate the most leads but have less conversion rates of around 30%. Whereas,
- ▶ 'Lead Add Form' generates less leads but conversion rate is great.
- ▶ We should try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'. 'Lead Import' does not seem very significant.



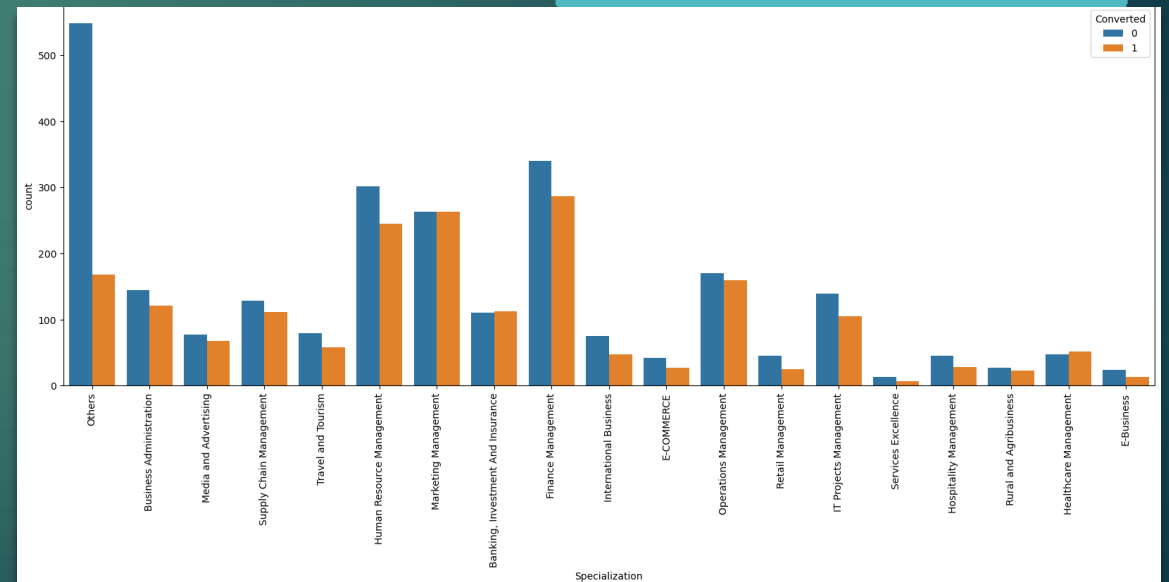
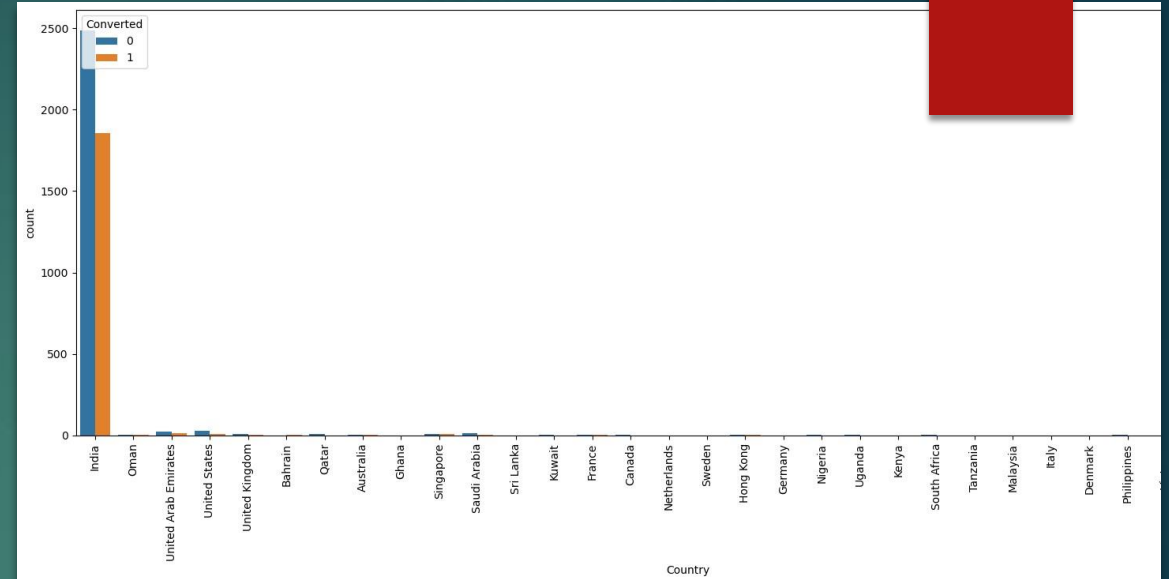
- ▶ Observations for Do Not Email and Do Not Call:
- ▶ As one can expect, most of the responses are 'No' for both the variables which generated most of the leads



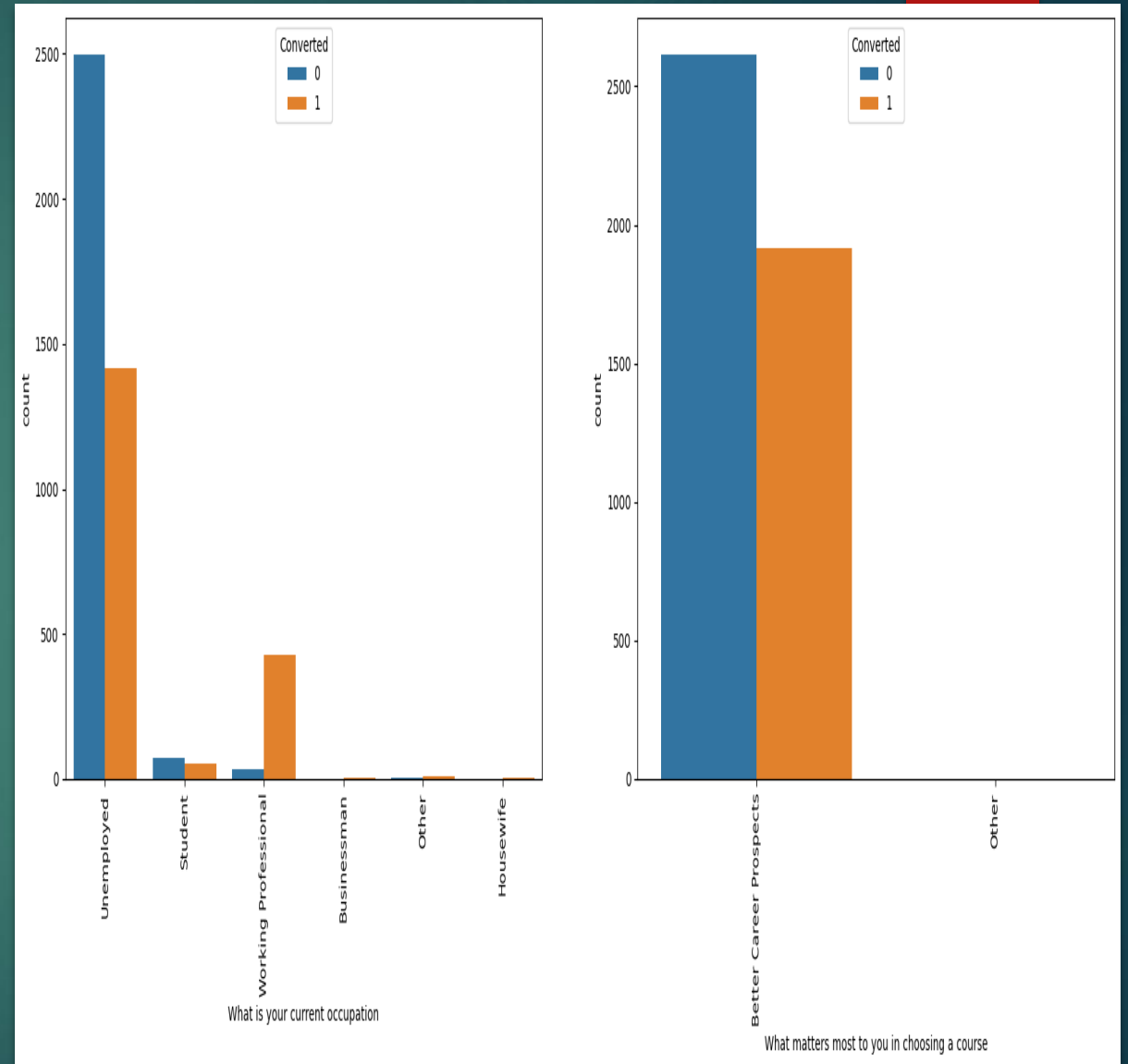
- ▶ Observations for Last Activity :
- ▶ Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.
- ▶ Categories after the 'SMS Sent' have almost negligible effect. We can aggregate them all in one single category.



- ▶ Observations for Country :
- ▶ Most of the responses are for India. Others are not significant.
- ▶ Observations for Specialization :
- ▶ Conversion rates are mostly similar across different specializations.

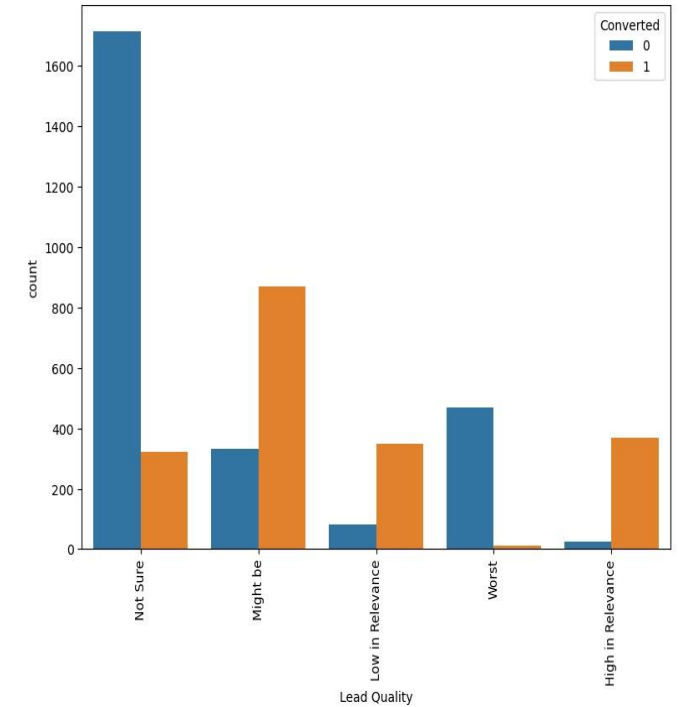
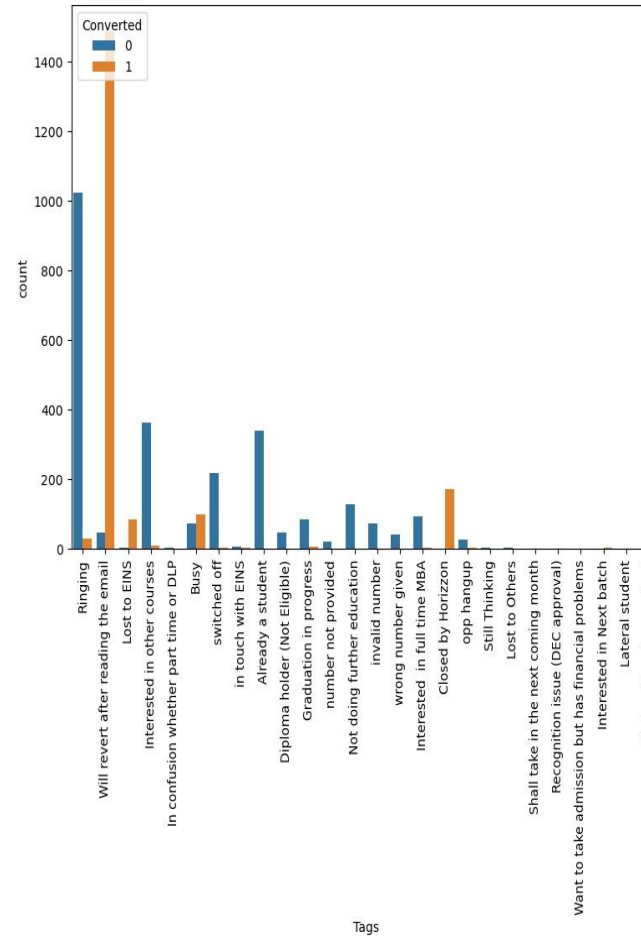


- ▶ Observations for What is your current occupation vs What matters most to you in choosing a course :
- ▶ The highest conversion rate is for 'Working Professional'.
- ▶ High number of leads are generated for 'Unemployed' but conversion rate is low.
- ▶ Variable 'What matters most to you in choosing a course' has only one category with significant count.





- ▶ Observations for Tags and Lead Quality:
- ▶ In Tags, categories after 'Interested in full time MBA' have very few leads generated, so we can combine them into one single category.
- ▶ Most leads generated and the highest conversion rate are both attributed to the tag 'Will revert after reading the email'.
- ▶ In Lead quality, as expected, 'Might be' as the highest conversion rate while 'Worst' has the lowest.



- Observations for Update me on Supply Chain Content, Get updates on DM Content, City, I agree to pay the amount through cheque, A free copy of Mastering The Interview, and Last Notable Activity :

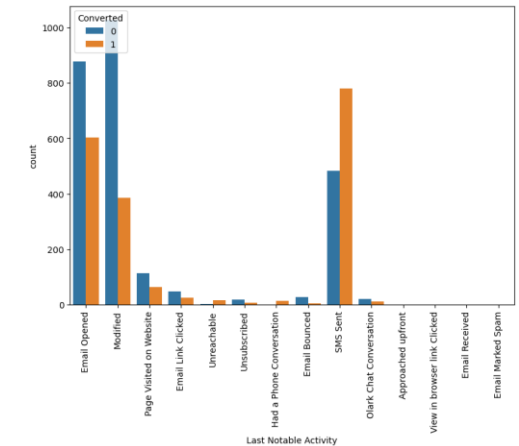
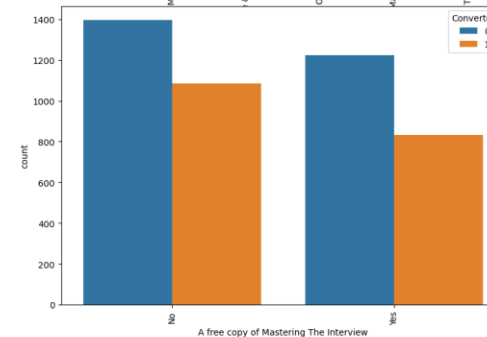
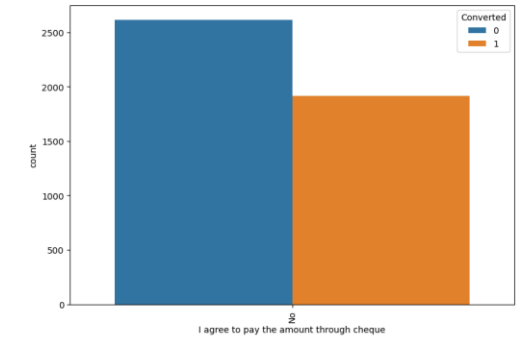
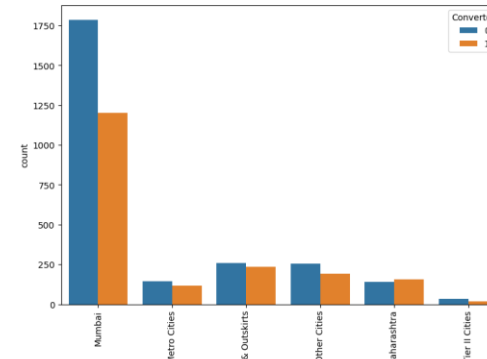
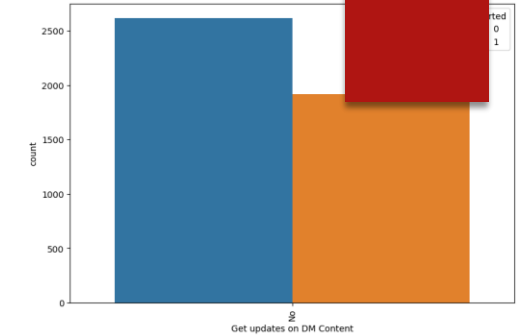
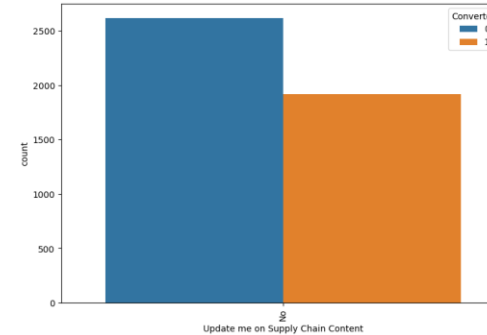
- Most of these variables are insignificant in analysis as many of them only have one significant category 'NO'.

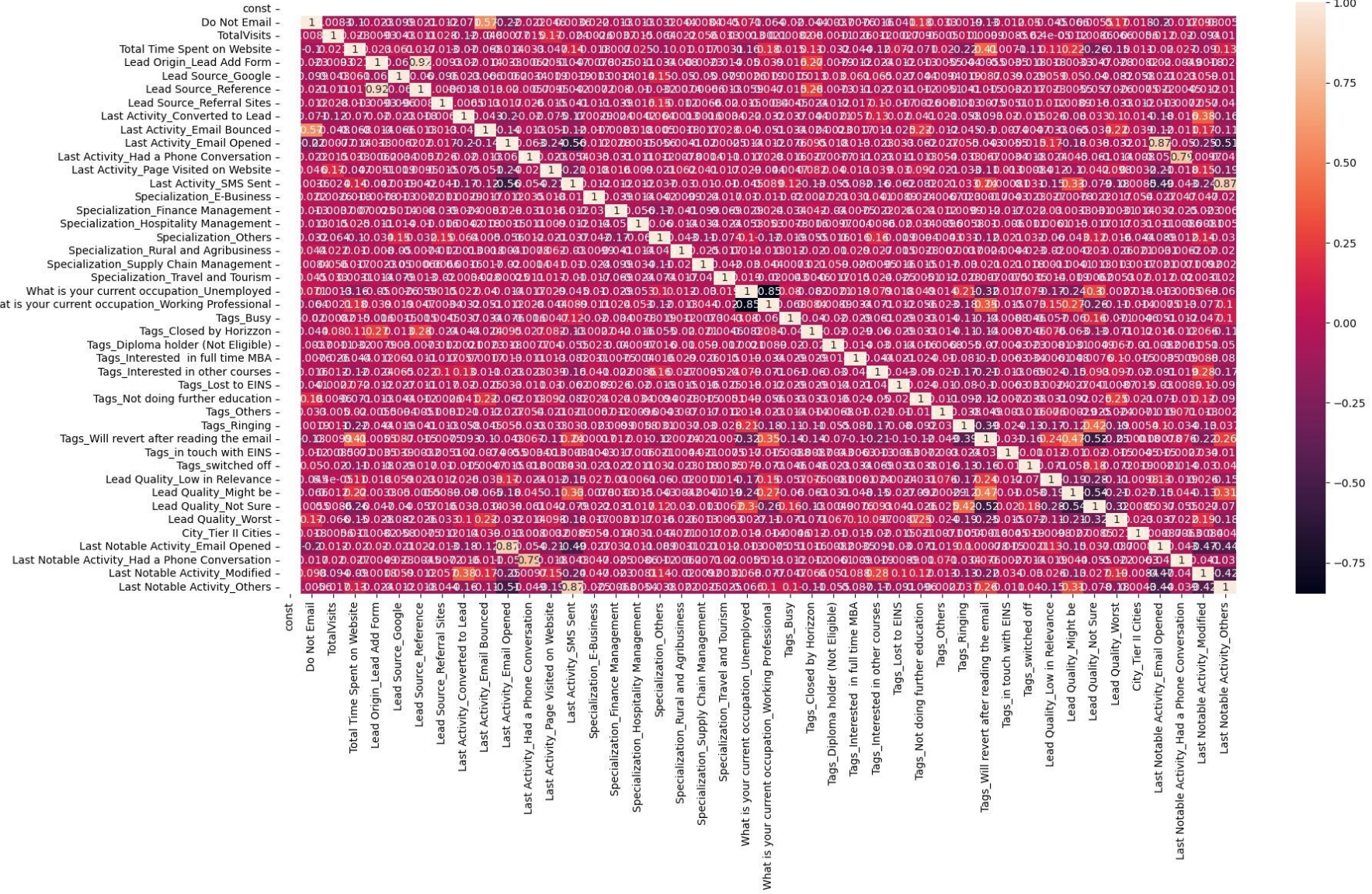
- In City, most of the leads are generated for 'Mumbai'.

- In 'A free copy of Mastering The Interview', both categories have similar conversion rates.


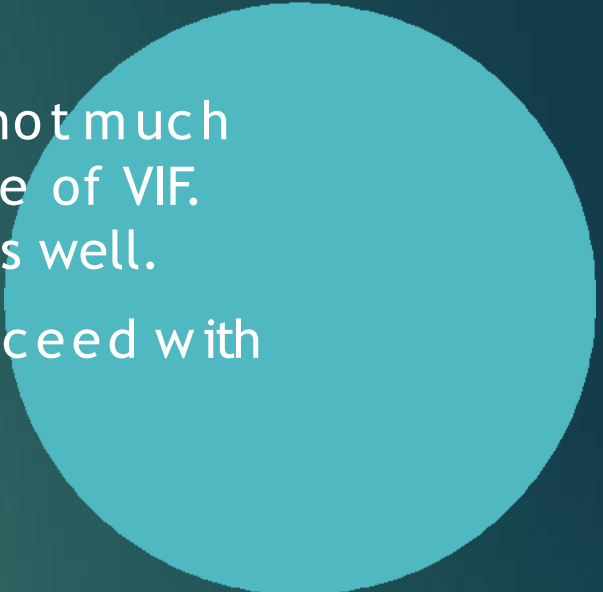
- In 'Last Notable Activity', we can combine categories after 'SMS Sent' similar to the variable 'Last Activity'.

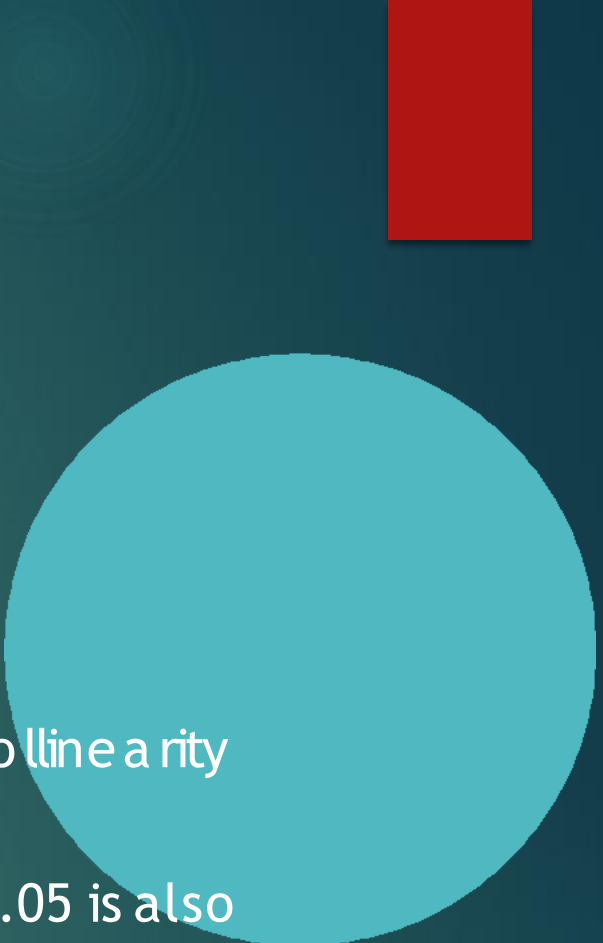
- It has most generated leads for the category 'Modified' while most conversion rate for 'SMS Sent' activity.







- 
- 
- ▶ From VIF values and heat maps, we can see that there is not much multicollinearity present. All variables have a good value of VIF. These features seem important from the business aspect as well.
  - ▶ So we need not drop any more variables and we can proceed with making predictions using this model only.

- 
- ▶ This is our final model:
  - ▶ All p-values are very close to zero.
  - ▶ VIFs for all features are very low. There is hardly any multicollinearity present.
  - ▶ Training accuracy of 91.95% at a probability threshold of 0.05 is also very good.