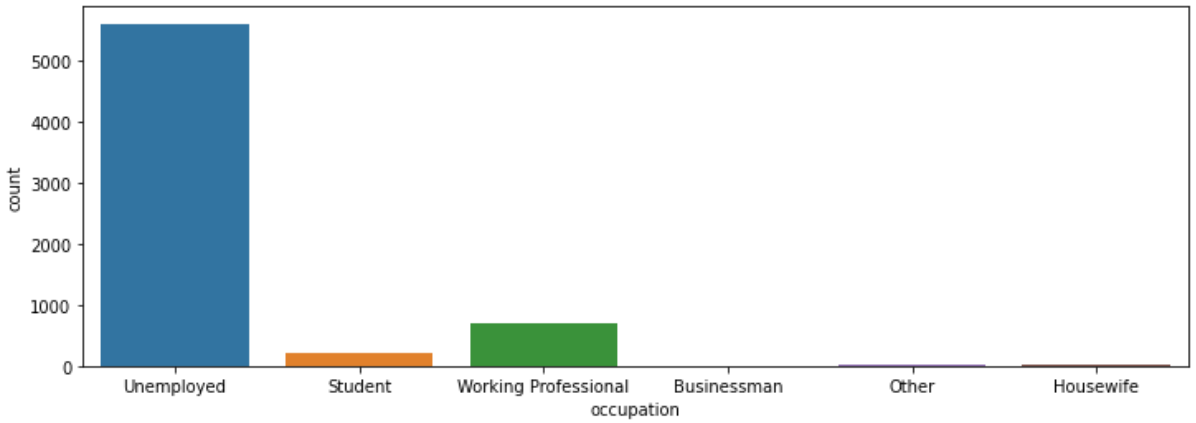
Team Members-12

1. Mohd Afraaz Firoz Khan Roll no: 220940325042
2.Shubham Bharat Chaudhari Roll no:220940325069
3.Mandar Manish Ghaisas Roll no:220940325041
4.Aishwarya Devdas Bhalbhar Roll no:220940325005
5.Rahul Vinayrao Joshi Roll no:220940325053

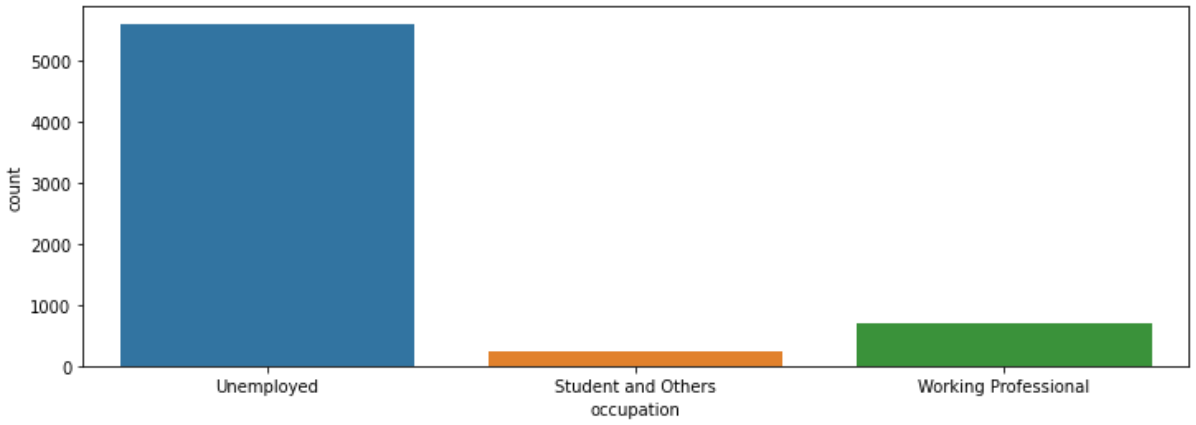
Problem Statement-

- **Cleaning The Data:**
- Importing necessary libraries
- Upload the csv file and then read the Dataset.
- Finding Statistical information about the data
- The Column Prospect id seems to be unimportant [Claim]
- Columns like Do Not Email, Do not Call, Get updates on DM Content, etc seems to be categorical as yes or no
- Some values seems to be rubbish such as 'select' in some columns, These acts as Null values
- There are 37 columns and 9240 Rows in the dataset.
- Making a copy of dataframe in order to restore it incase of an emergency
- Finding the information of the data
- Null values are present in the dataset
- Data type of columns are object integer and float
- Finding Statistical information about the data and then we check for Null values
- Checking column:
- Here we are done Claim checking
- After that we change nomenclature
- After we check Renaming the Columns
-
-
- The Column Prospect id seems to be unimportant
- In Observation we have seen that the Prospect is an unimportant column, Decided to drop the column
- Then we Successfully dropped Prospect Id
-
-
- Then after which we are Renaming the columns
- Then Selecting all the non-numeric columns i.e the columns with data type as 'Object'
- After which we are Handling the "Select = Null" claim
- Then Finding out the columns that have "Select" as a value in it
- We got the Observation that These are the columns containing "Select" as a value
- Then replacing all the values
- **Handling Null Values:**
- Checking for percent of null values in the dataframe
- after analysing, we decided to drop the columns having Null values greater than 40%

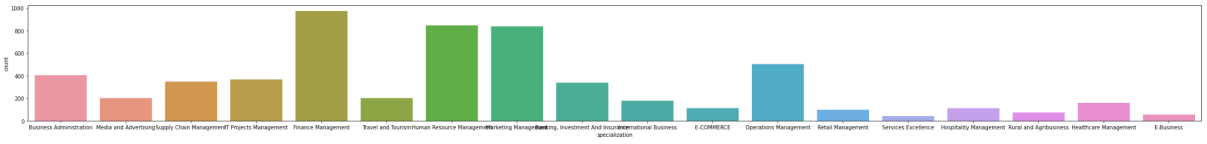
- Now checking the remaining columns having Null values, One by one
- From the observation we can see that the data is completely Skewed toward 'India'.
- Therefore we will drop the 'Country' column
- Checking for the course_selection_reason column
- From the observation it is observed that the data is skewed towards
- 'Better Career Prospects' & NaN is upto 30%
- It is better to drop this column
- Working on 'occupation' column



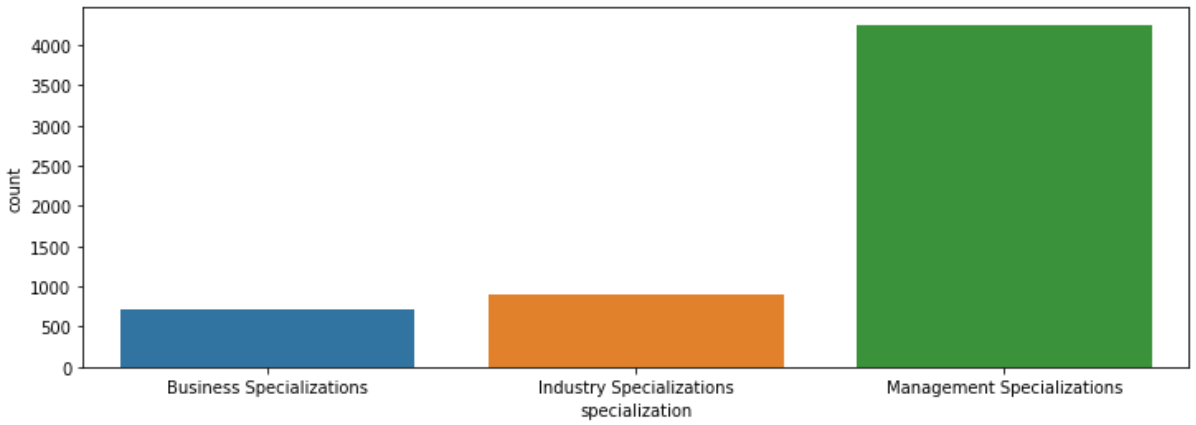
- **Observation:**
- It is observed that the data is highly skewed toward 'Unemployed' &
- The Null values is almost 30%, therefore it is better to drop the column



- **Working on Specialisation Column**

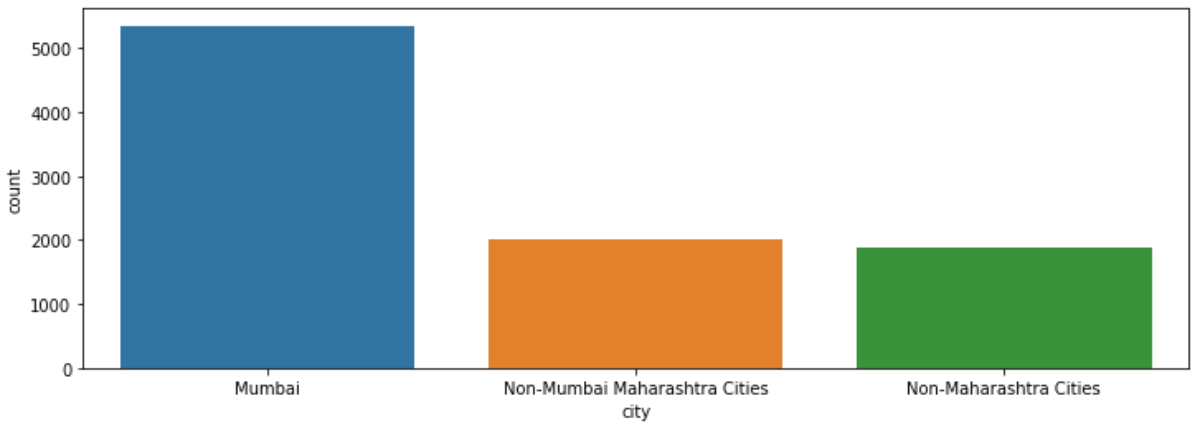


- For specialisation, we will first combine categories based on the course type,
- and then impute proportionally to maintain the distribution and not introduce bias
- First, Categorizing based on Management
- Categorising on Business Courses
- Categorising on Industry Courses



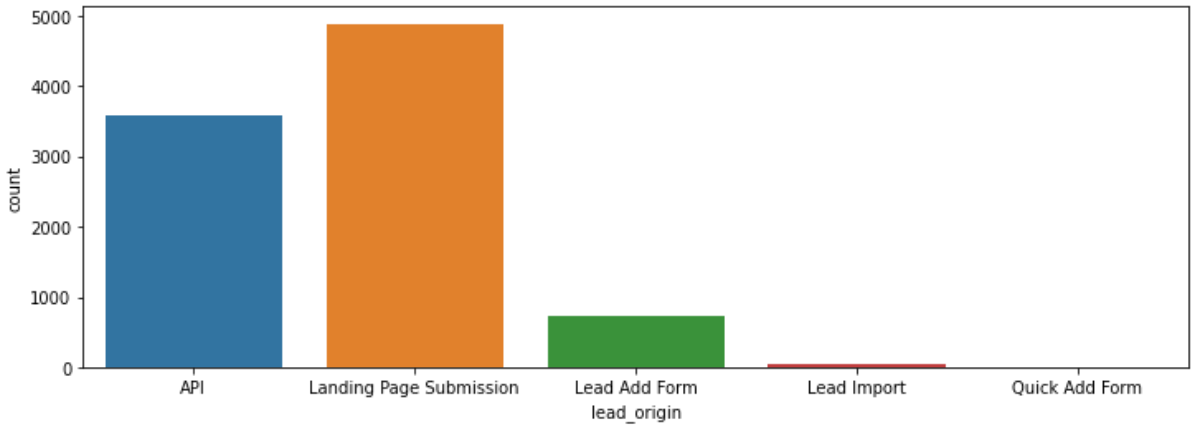
- **Observation:**
- From above plot it can be concluded that maximum count is of Management Specialization course whereas least count is for Business Specialization courses
- Null values are eliminated in specialisation column
- Working on City Column

- categorise all non-mumbai, but Maharashtra cities
- categorise all other cities

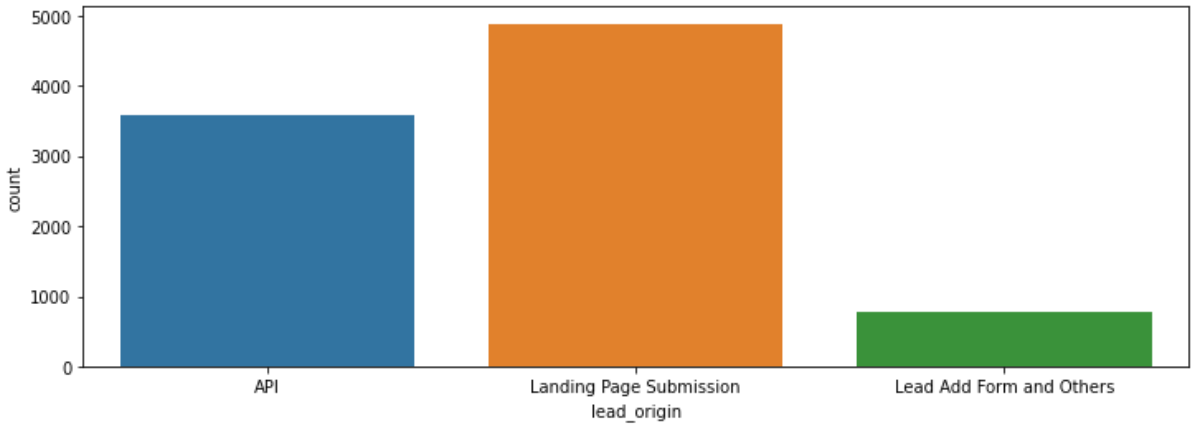


- determine unique values for all object data type columns
- As can be seen from the above output, the categorical columns (i.e. number of unique values > 2) are:
- lead_origin

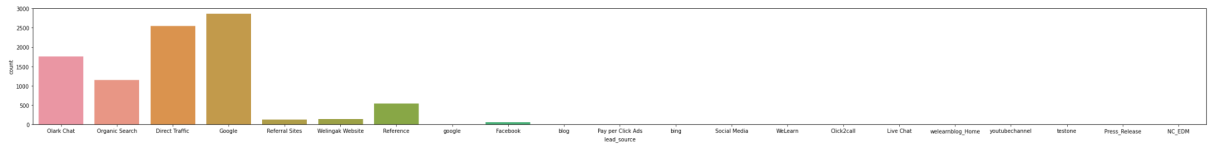
- lead_source
- **Working on lead_origin column**



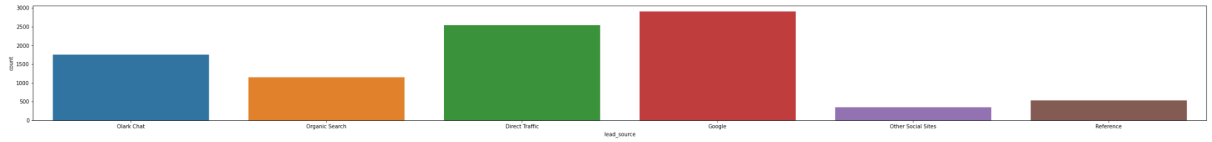
- Combining the columns with lower percentage
- No Null values present



- Working with the 'lead_source' column

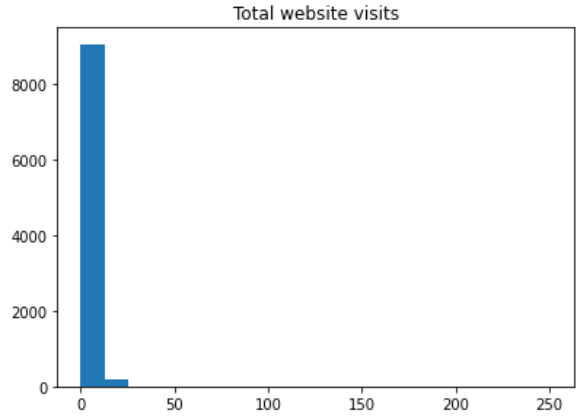


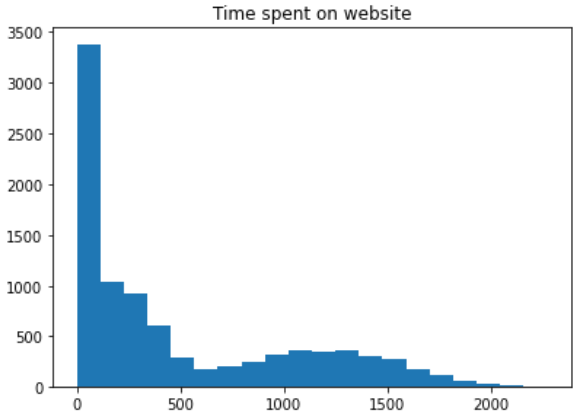
- Checking for the mode value from the data
- Imputing missing values with mode if the data that is mode.

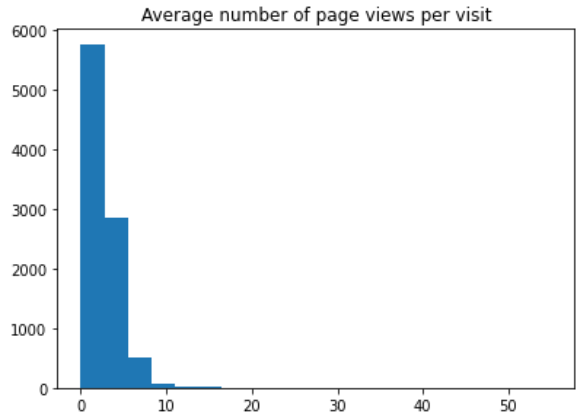


- **Handling Binary Column:**
-
- **Observation:**
- Columns having 1 value can be dropped.
- Dropping Binary values having Biassed data
-
- **Handling numerical Columns:**
- Checking for the Null Values
- Handling missing values of total_visits
- Handling missing values of page_views_per_visit

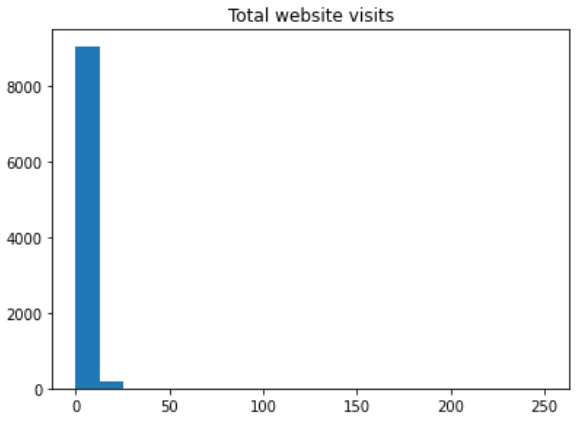
- Successfully handle all the missing values
- **Observation :**
-
- Looking at both the box plots and the statistics, there are upper bound outliers in both total_visits and page_views_per_visit columns.

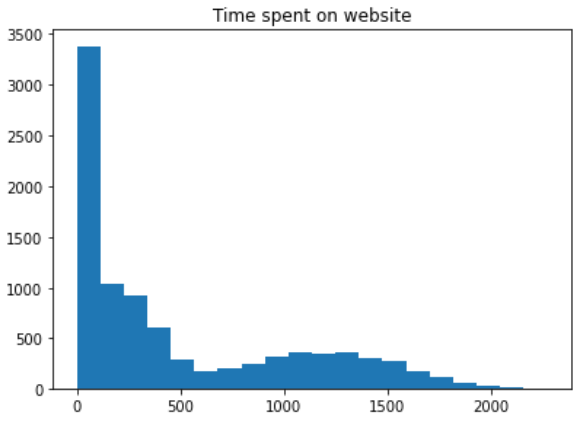


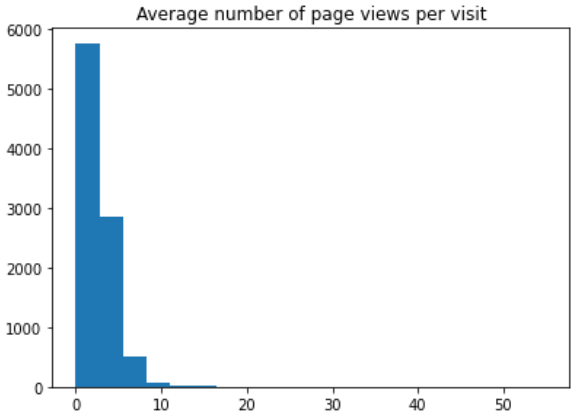


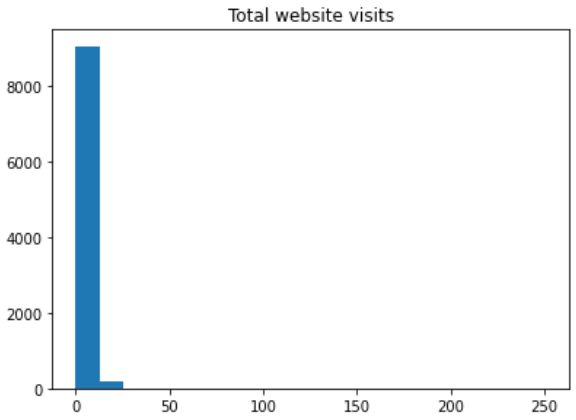


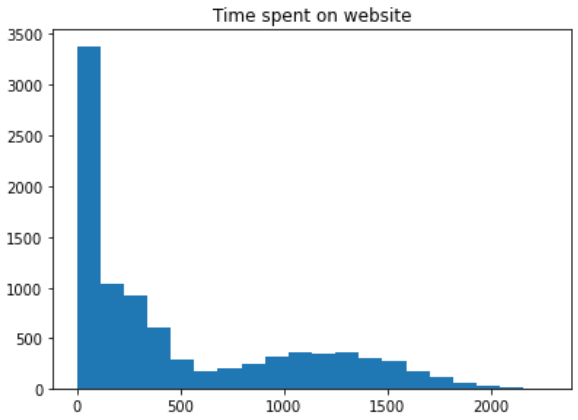
-
- **Observation:**
-
- The total website visit is maximum in the range of 0 to 25
- The maximum time spent on website is highest in the range from 0 -100 which is gradually decreasing with time
- The Average number of page visit per day is max at 0 whereas there is a half way drop in the range approx between 2-3 & than gradually decreasing
-
- **Handling Categorical Data:**
- determine unique values
- Considering columns having 2 unique values
- Creating dummy values for categorical data

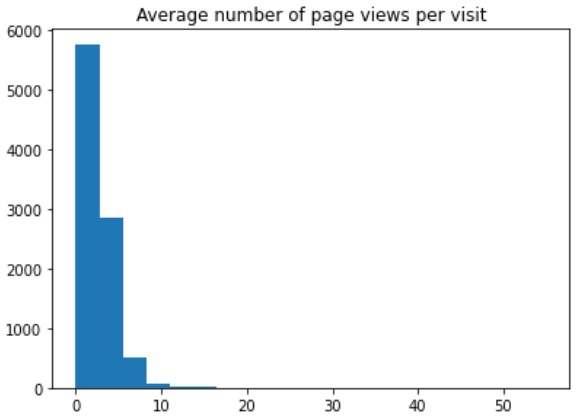
- 





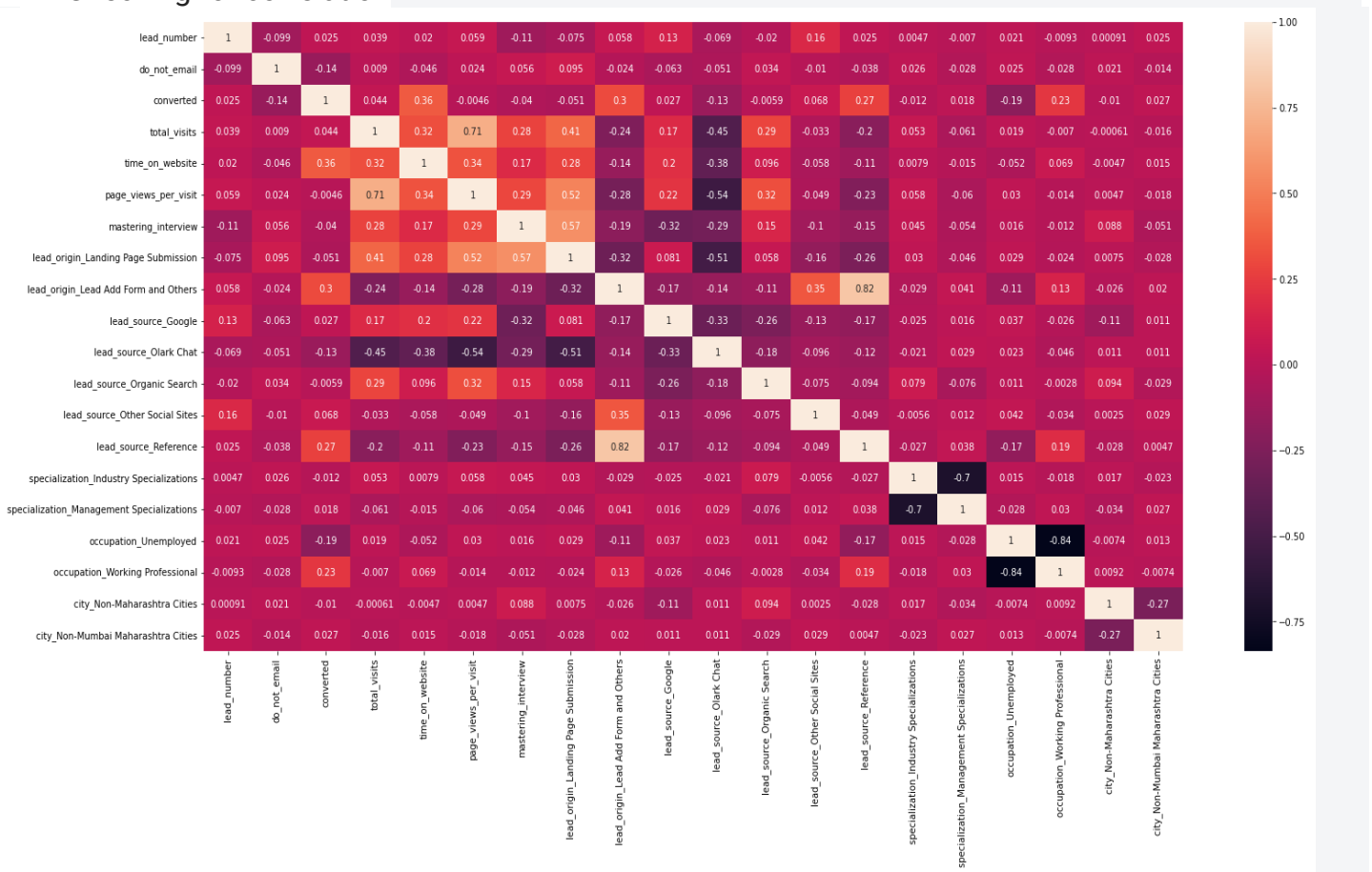
- 



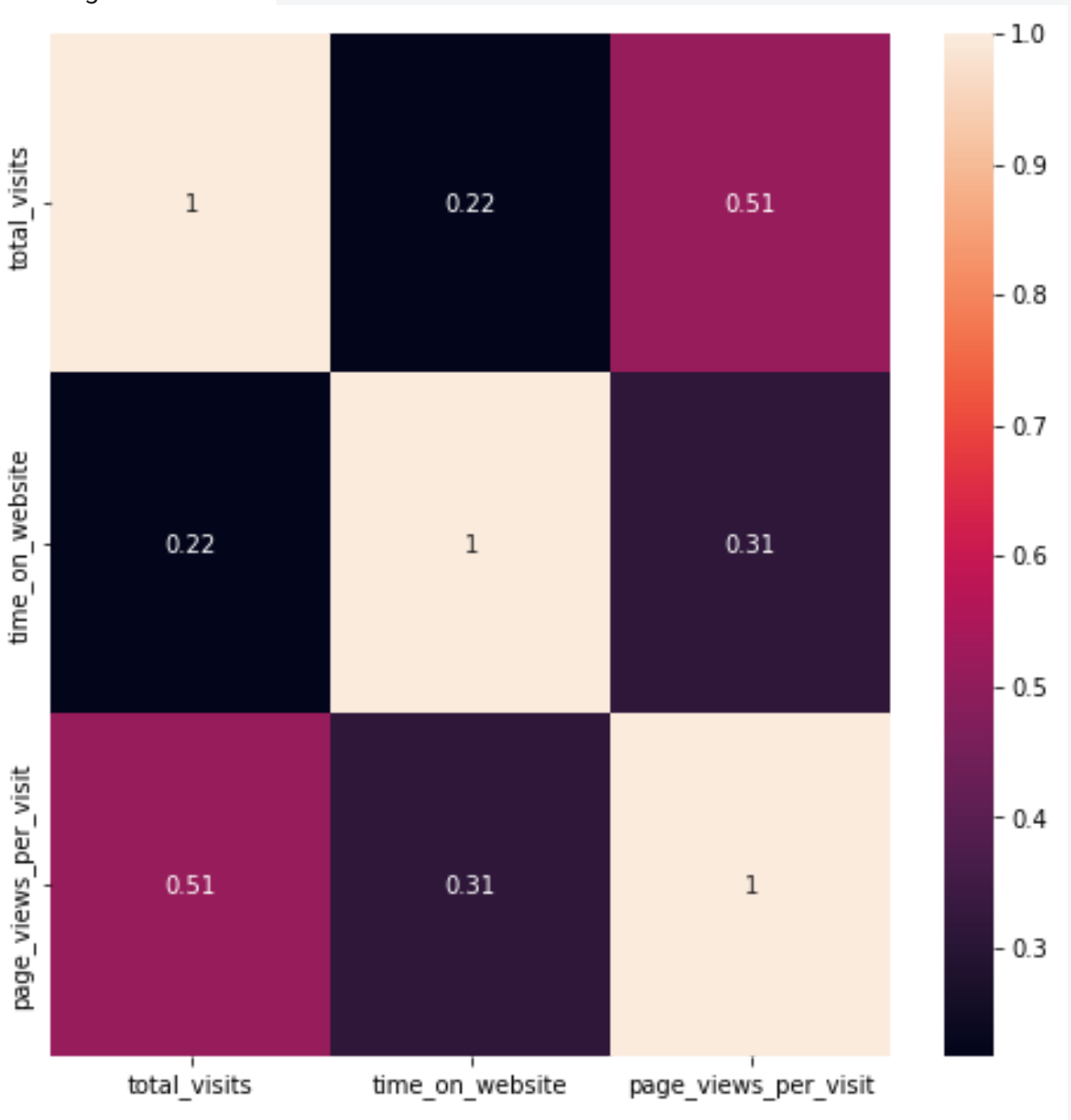


- Checking for Outliers at various level
-
- **Train Test Split:**

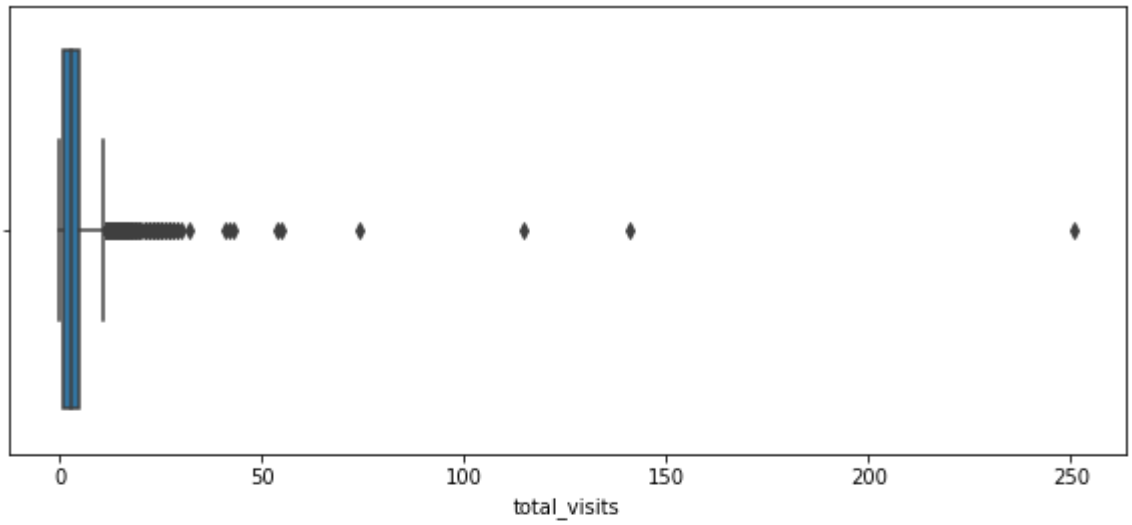
- Checking for correlation

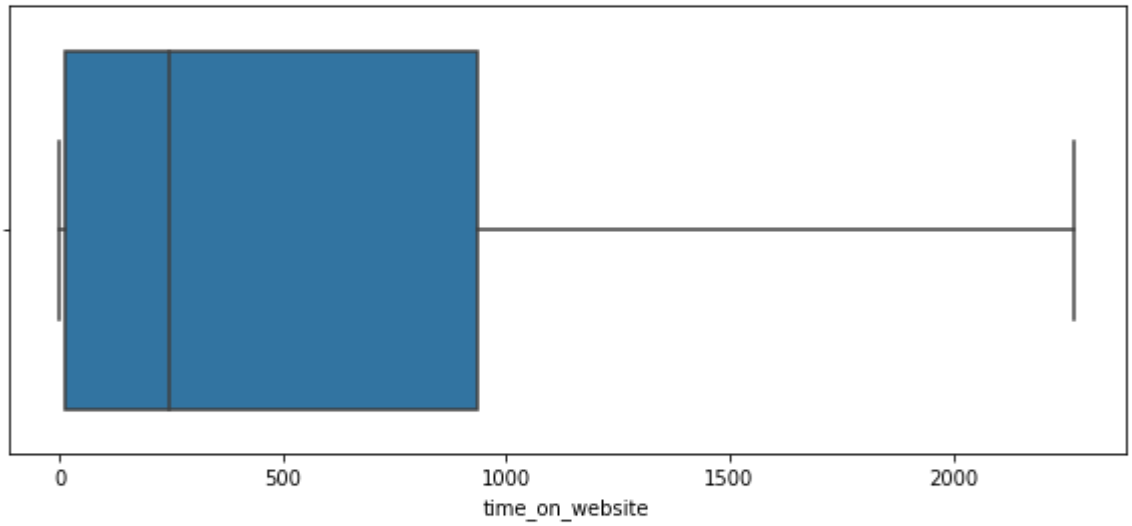


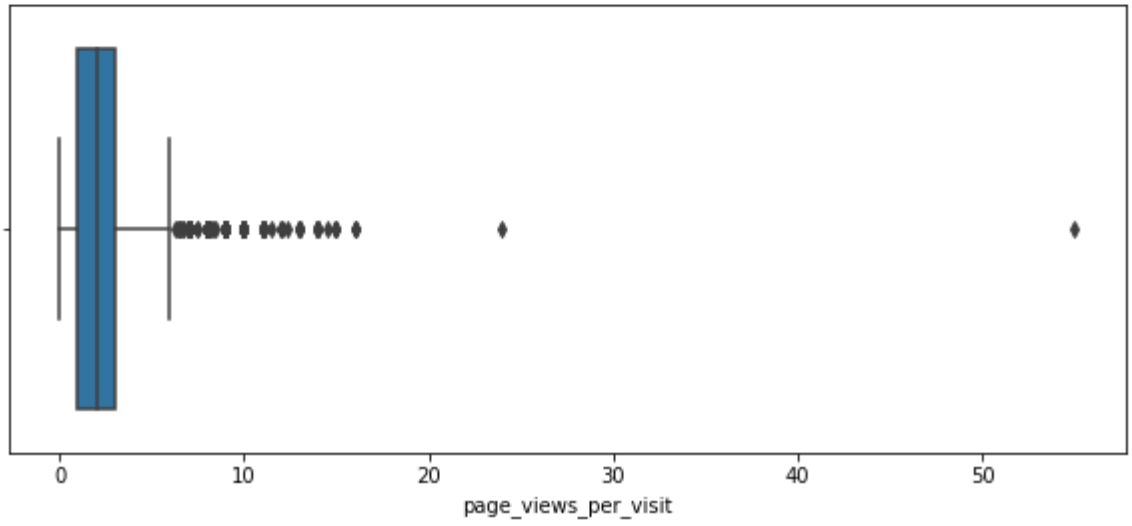
- **EDA [Exploratory Data Analysis]:**
- Checking for correlation



- Observation:
- No Major correlation







-
-
-

1)Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

From Correlation matrix, it is defined that

1)Time_on_website originally called as "total_time_spent_on_website"

-> The Correlation Matrix defines that the Time_on_website has 36% Positive correlation with converted, which is the target column.

2)Lead_origin_Landing Page Submission originally called lead_origin

-> From the Correlation matrix we can check that, Lead_origin_Landing Page Submission 3% Positive of correlation with the target column

3)lead_source_Reference originally called as lead_source

-> Correlation Matrix shows the correlation of lead_source of 27% Positive with converted[Target column]

4)occupation_Working Professional originally known as occupation

-> occupation_Working Professional has 23% Positive correlation with converted [Target column]

5)occupation_unemployed

-> occupation_unemployed has a negative 19% correlation with converted which is the target column.

2. What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

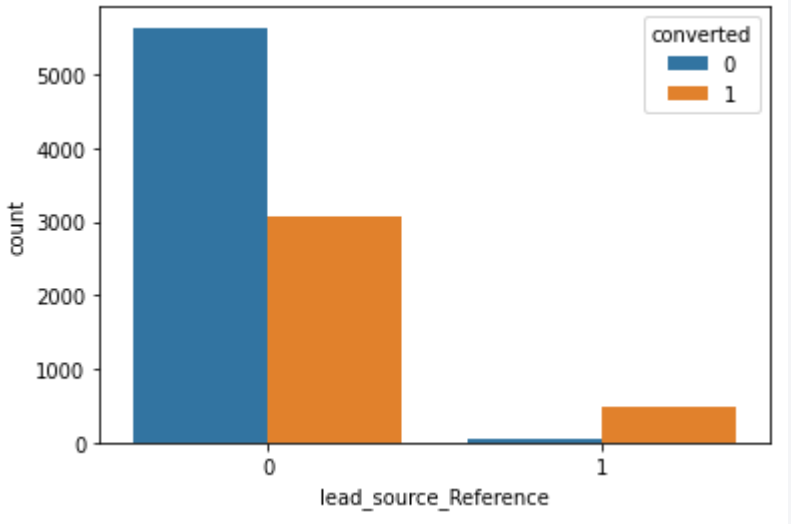
Ans: From the Correlation matrix, it is understood that,

1)Lead Source_Reference

2)lead_origin_Lead Add Form and Others Originally called as Lead Source_Social Media

3)Lead Source_Olark Chat

3.X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as many of such people as possible. Suggest a good strategy they should employ at this stage.

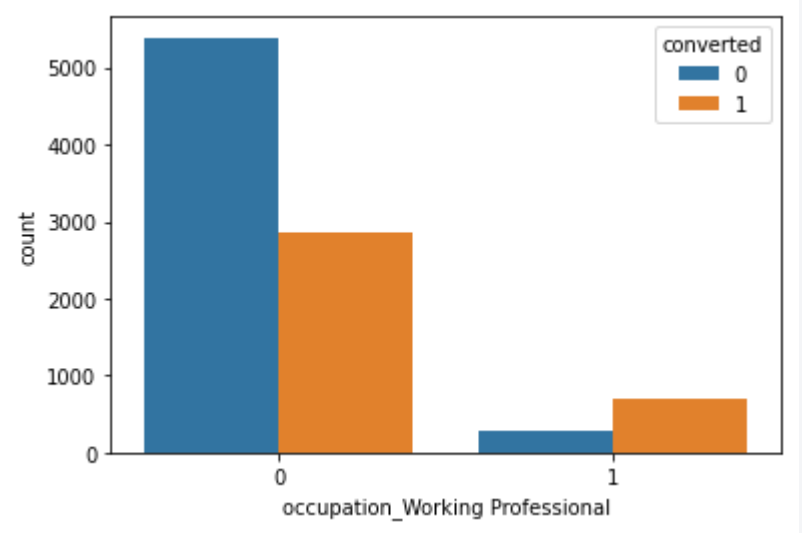


```
pd.crosstab(df.lead_source_Reference, df.converted)
```

converted	0	1
lead_source_Reference		
0	5635	3071
1	44	490

1)The Above countplot shows that, the lead source coming from reference has more probability of conversion, thus company can focus on people whose lead source is reference

2)The correlation matrix shows that people spending more time on website [column-total_time_spent_on_website] have higher rate of conversion, thus company should focus on people spending more time on website



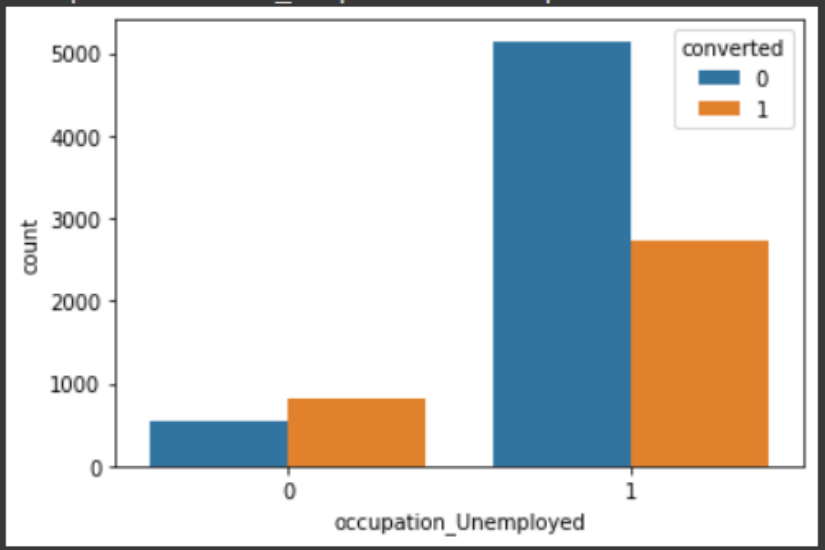
from above plot we can say that,

3)People having a working profession as Profession have higher probability of getting converted, The company can focus on people having working profession as Professional.

4. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimise the rate of useless phone calls. Suggest a strategy they should employ at this stage.

```
sns.countplot(df['occupation_Unemployed'], hue=df['converted'])
```

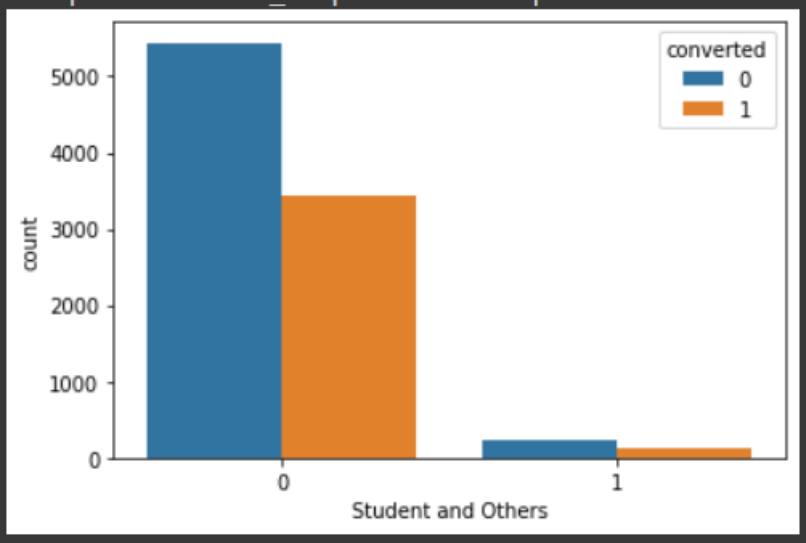
```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pas
  warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7fd4cbb8bdc0>
```



At a time when company wants to minimise the rate of useless phone calls, the team may not focus on people having occupation as unemployed. Because from the above plot we can understand that, the occupation_umemployed people are less converted

```
sns.countplot(occupation_analysis['Student and Others'], hue=df['converted'])
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass t
  warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7fd4cba36e20>
```



The people being Students and others have less probability of converting, lokks like they are already studying and not want to enroll in courses specially designed for the working professionals, so at time of minimizing the rate of useless phone calls, such people are ignored

Conclusion:

1)Number of Unemployed people were more

2)Maximum no. of specialisation was in Finance_Management

3)Maximum no. of people were from Mumbai

4)Maximum lead origin was from Landing Page Submission

5)Maximum lead_source was from Google with approx. 31%

6)Toptal_visit & page_views_per_visit had maximum outliers

7)Total time spent on website has maximum variance in data