Lead Scoring Case Study

Presented By

Vibhor Srivastava
Divya Voddaboina
Vikas Kharwal

PROBLEM STATEMENT

- X Educationsells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also knownas ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go upas the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE

- ► X education wants to know most promising leads.
- ► For that they want to build a Model which identifies the hot leads.
- ► Deployment of the model for the future use.

SOLUTION METHODOLOGY

Data cleaning and data manipulation.

1.Check and handle duplicate data.

2.Check and handle NA values and missing values.

3.Drop columns, if it contains a large number of missing values and are not useful for the analysis.

4.Imputation of the values, if necessary.

5.Check and handle outliers in data.

City: City has 39.71 % missing values. Imputing missing values with Mumbai will make the data more skewed. Skewness will later cause bias in the model. Hence City column can be dropped.

Specialization: Specialization has 36.58 % missing values. The specialization selected is evenly distributed. Hence imputation or dropping is not a good choice. We need to create additional category called 'Others'.

Tags: Tags has 36.29 % missing values. Tags are assigned to customers indicating the current status of the lead. Since this is current status, this column will not be useful for modeling. Hence it can be dropped.

What matters most to you in choosing a course: This variable has 29.32 % missing values. 99.95% customers have selected 'better career prospects'. This is massively skewed and will not provide any insight.

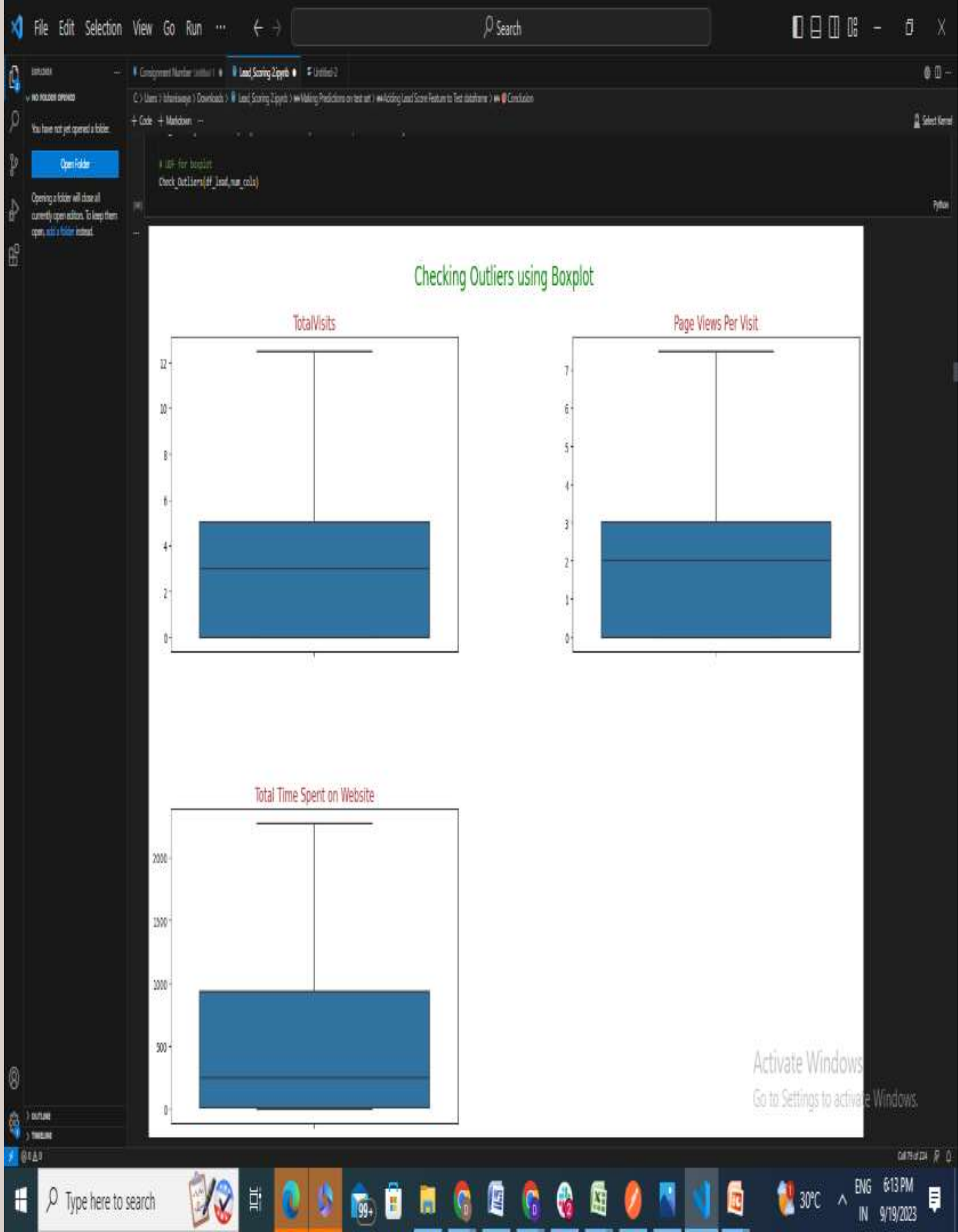
What is your current occupation: We can impute the missing values with 'Unemployed' as it has the most values. This seems to be a important variable from business context, since X Education sells online courses and unemployed people might take this course to increase their chances of getting employed.

Country: X Education sells online courses and appx 96% of the customers are from India. Does not make business sense right now to impute missing values with India. Hence `Country column can be dropped.

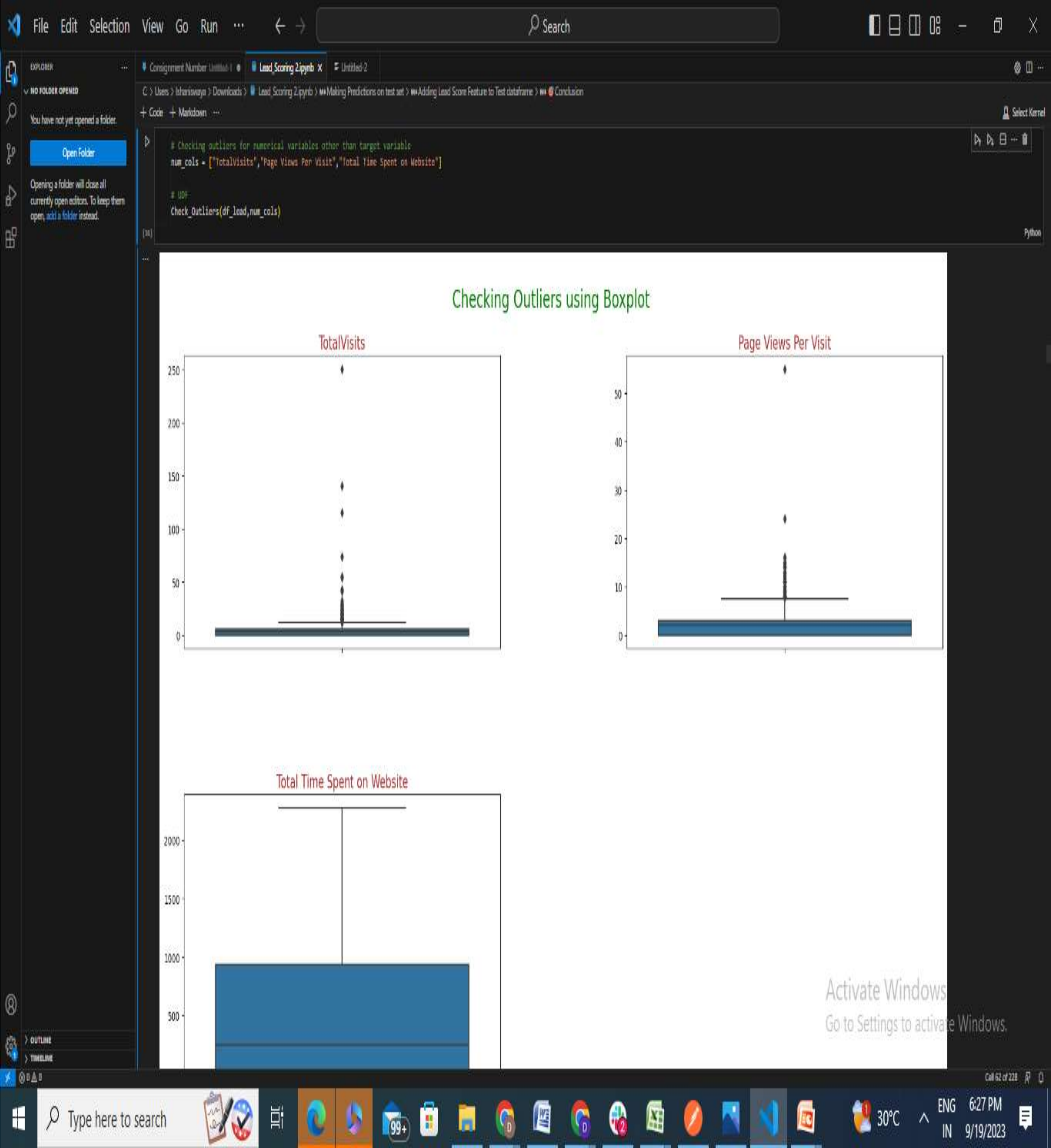
Last Activity: "Email Opened" is having highest number of values and overall missing values in this column is just 1.11%, hence we will impute the missing values with label 'Email Opened'.

Lead Source: "Google" is having highest number of occurences and overall nulls in this column is just 0.39%, hence we will impute the missing values with label 'Google‘

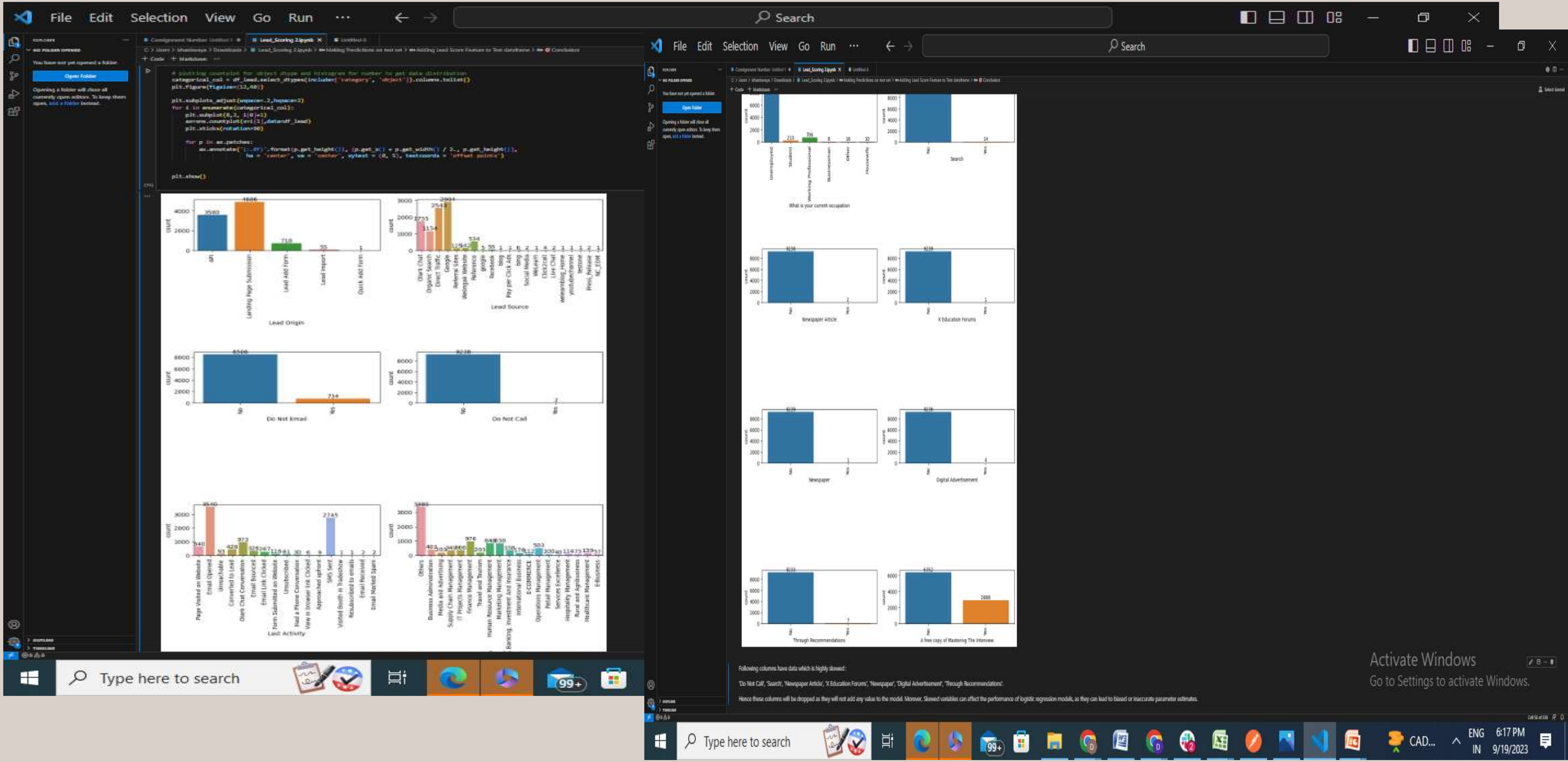
Dropping the columns ('City','Tag','Country','What matters most to you in choosing a course')


File Edit Selection View Go Run ···
Search
Consignment Number Untitled-1
Lead Scoring 2.ipynb
Untitled-2
C > Users > Ishaniswaya > Downloads > Lead Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code + Markdown
Select Kernel
Python
Checking Outliers using Boxplot
TotalVisits
Page Views Per Visit
Total Time Spent on Website
Activate Windows
Go to Settings to activate Windows.
Type here to search
30°C
ENG IN
6:13 PM
9/19/2023

```
# UDF for boxplot
Check_Outliers(df_lead,num_cols)
```


File Edit Selection View Go Run ···
Search
Consignment Number Untitled-1
Lead Scoring 2.ipynb
Untitled-2
C > Users > Ishaniswaya > Downloads > Lead Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code + Markdown
Select Kernel
Python
Checking Outliers using Boxplot
TotalVisits
Page Views Per Visit
Total Time Spent on Website
Activate Windows
Go to Settings to activate Windows.
Type here to search
30°C
ENG IN
6:27 PM
9/19/2023

```
# Checking outliers for numerical variables other than target variable
num_cols = ["TotalVisits","Page Views Per Visit","Total Time Spent on Website"]

# UDF
Check_Outliers(df_lead,num_cols)
```



```
# plotting countplot for object dtype and histogram for number to get data distribution
categorical_col = df_lead.select_dtypes(include=['category', 'object']).columns.tolist()
plt.figure(figsize=(12,40))

plt.subplots_adjust(wspace=.2,hspace=2)
for i in enumerate(categorical_col):
    plt.subplot(8,2, i[0]+1)
    axes=sns.countplot(x=i[1],data=df_lead)
    plt.xticks(rotation=90)

    for p in ax.patches:
        ax.annotate('{:.0f}'.format(p.get_height()), (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha = 'center', va = 'center', xytext = (0, 5), textcoords = 'offset points')


plt.show()
```

Following columns have data which is highly skewed :

'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations'.

Hence these columns will be dropped as they will not add any value to the model. Morever, Skewed variables can affect the performance of logistic regression models, as they can lead to biased or inaccurate parameter estimates.

Exploratory Data Analysis (EDA)

1. Univariate data analysis: value count, distribution of variables, etc.

2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

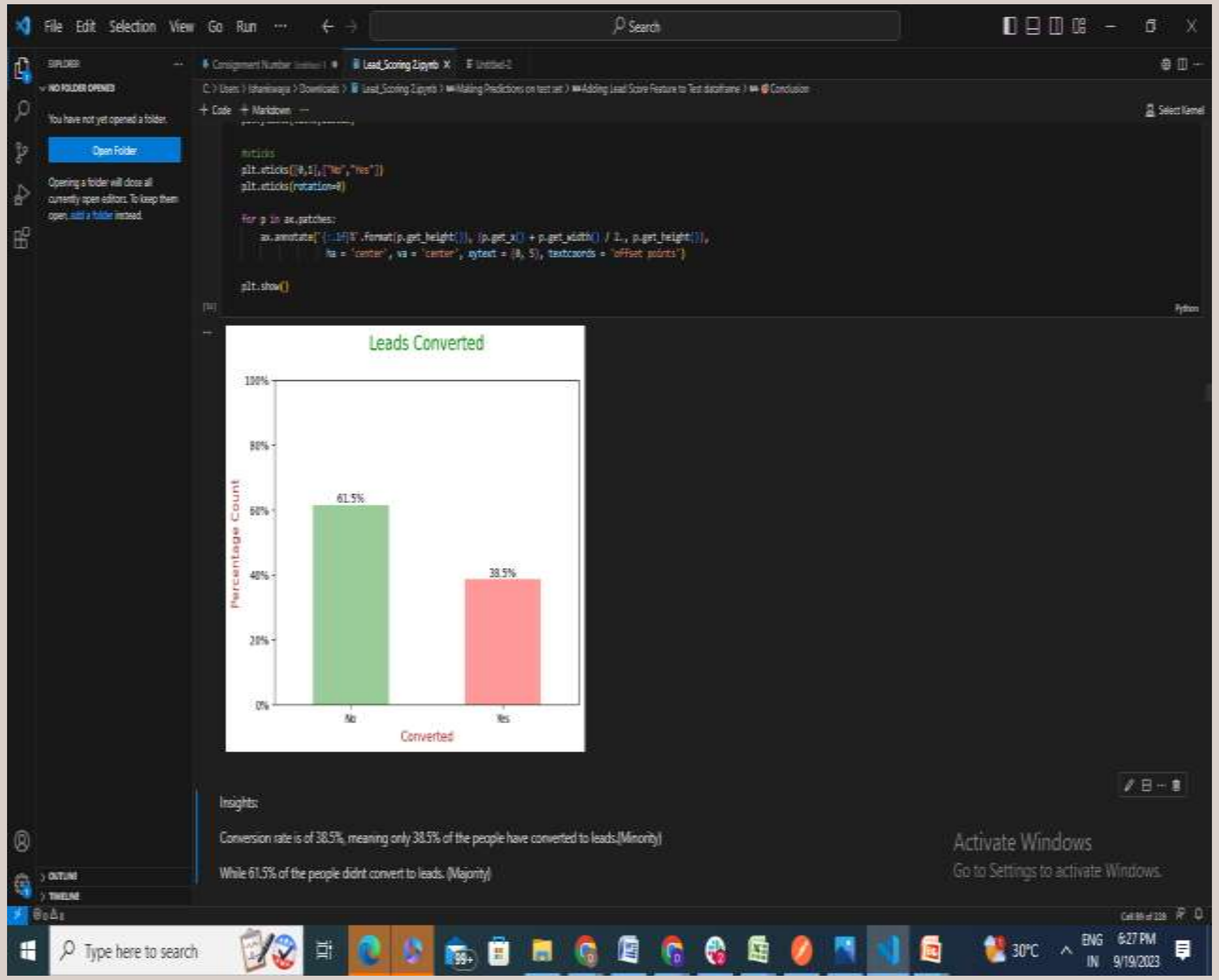
3. Feature Scaling & Dummy variables and encoding of the data.

4. Classification technique: logistic regression is used for model making and prediction.

5. Validation of the model.

6. Model presentation

7. Conclusions and Recommendation

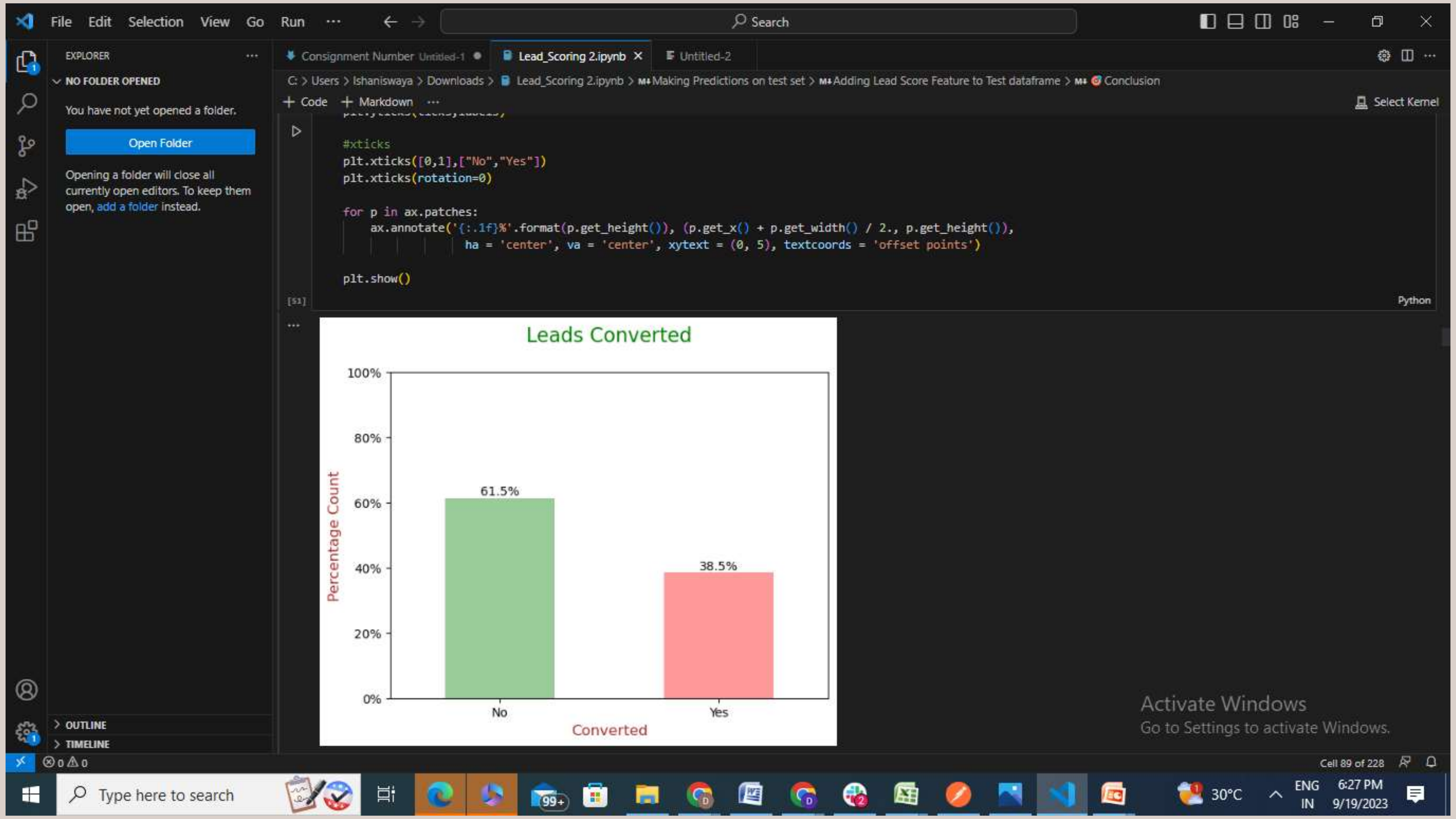


```
#xticks
plt.xticks([0,1],["No","Yes"])
plt.xticks(rotation=0)

for p in ax.patches:
    ax.annotate('{:.1f}%'.format(p.get_height()), (p.get_x() + p.get_width() / 2., p.get_height()),
                ha = 'center', va = 'center', xytext = (0, 5), textcoords = 'offset points')

plt.show()
```

Insights:

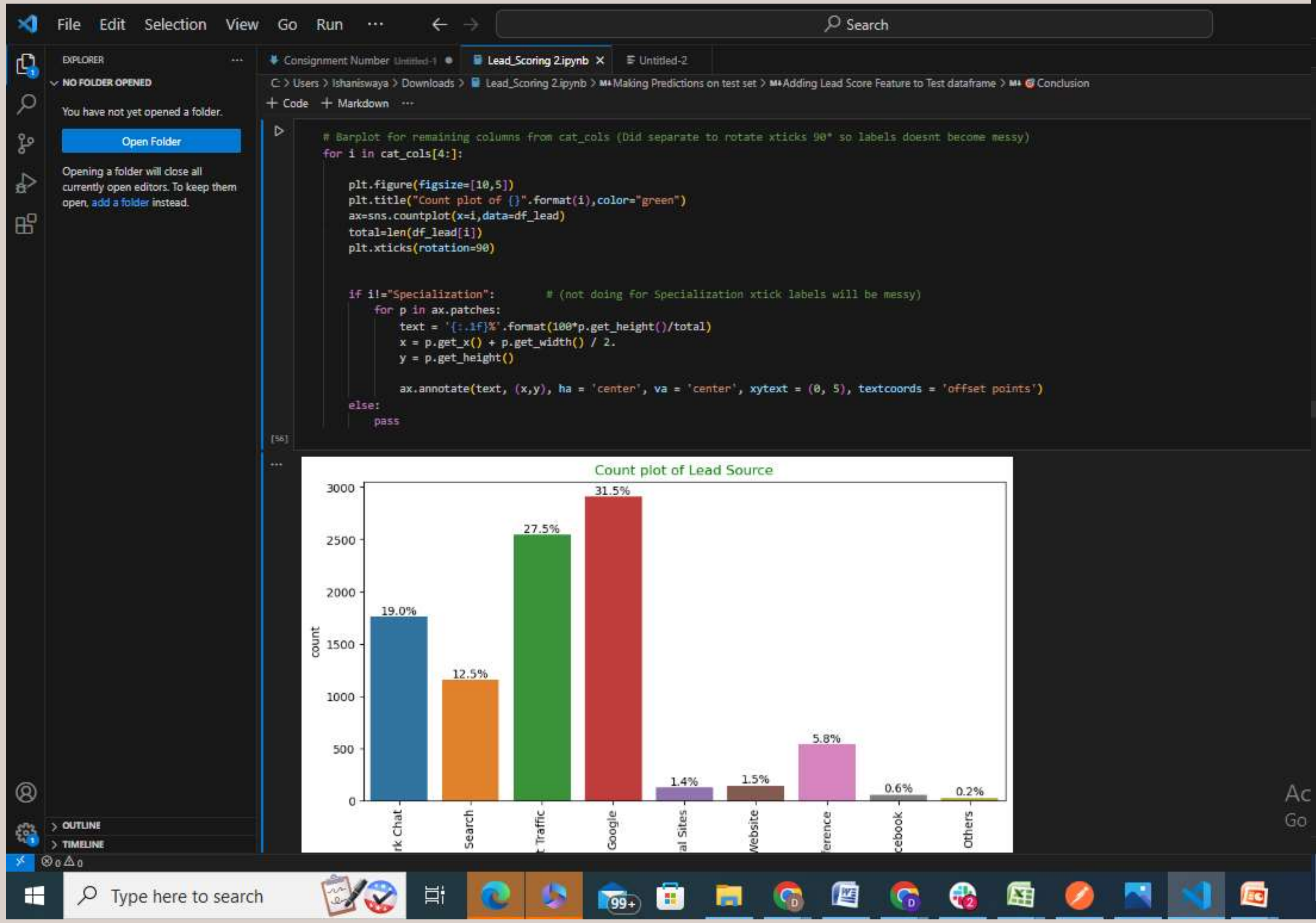
Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

While 61.5% of the people didnt convert to leads. (Majority)



```
#xticks
plt.xticks([0,1],["No","Yes"])
plt.xticks(rotation=0)

for p in ax.patches:
    ax.annotate('{:.1f}%'.format(p.get_height()), (p.get_x() + p.get_width() / 2., p.get_height()),
                ha = 'center', va = 'center', xytext = (0, 5), textcoords = 'offset points')

plt.show()
```

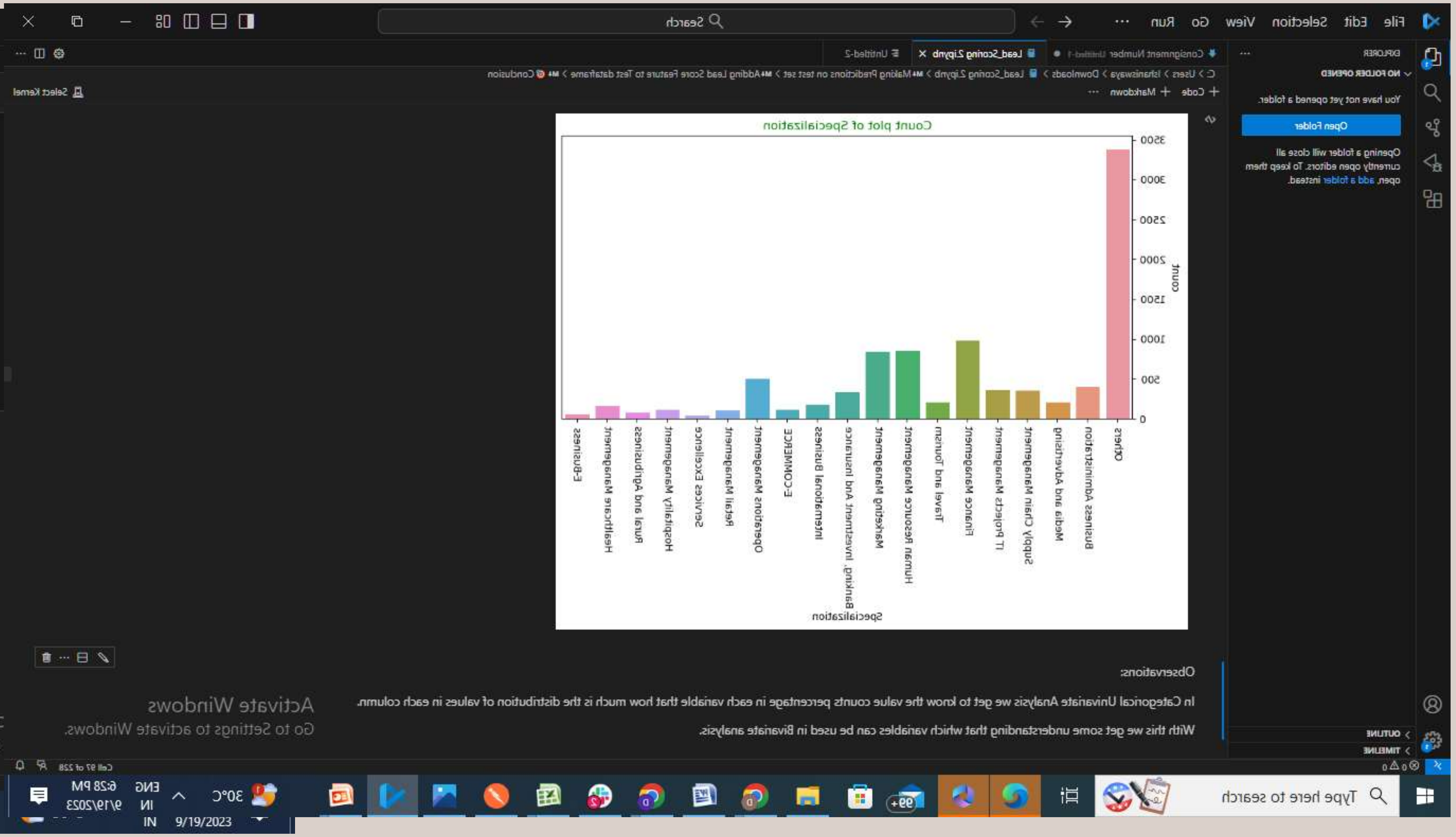
File
Edit
Selection
View
Go
Run
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code
+ Markdown
Select Kernel
[51]
Python
Leads Converted
100%
80%
60%
40%
20%
0%
61.5%
38.5%
Percentage Count
No
Yes
Converted
Activate Windows
Go to Settings to activate Windows.
OUTLINE
TIMELINE
Cell 89 of 228
Type here to search
30°C
ENG IN
6:27 PM
9/19/2023



File Edit Selection View Go Run ···
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
OUTLINE
TIMELINE
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code + Markdown
Type here to search

```python
# Barplot for remaining columns from cat_cols (Did separate to rotate xticks 90* so labels doesnt become messy)
for i in cat_cols[4:]:

    plt.figure(figsize=[10,5])
    plt.title("Count plot of {}".format(i),color="green")
    ax=sns.countplot(x=i,data=df_lead)
    total=len(df_lead[i])
    plt.xticks(rotation=90)


    if i!="Specialization":        # (not doing for Specialization xtick labels will be messy)
        for p in ax.patches:
            text = '{:.1f}%'.format(100*p.get_height()/total)
            x = p.get_x() + p.get_width() / 2.
            y = p.get_height()

            ax.annotate(text, (x,y), ha = 'center', va = 'center', xytext = (0, 5), textcoords = 'offset points')
    else:
        pass
```

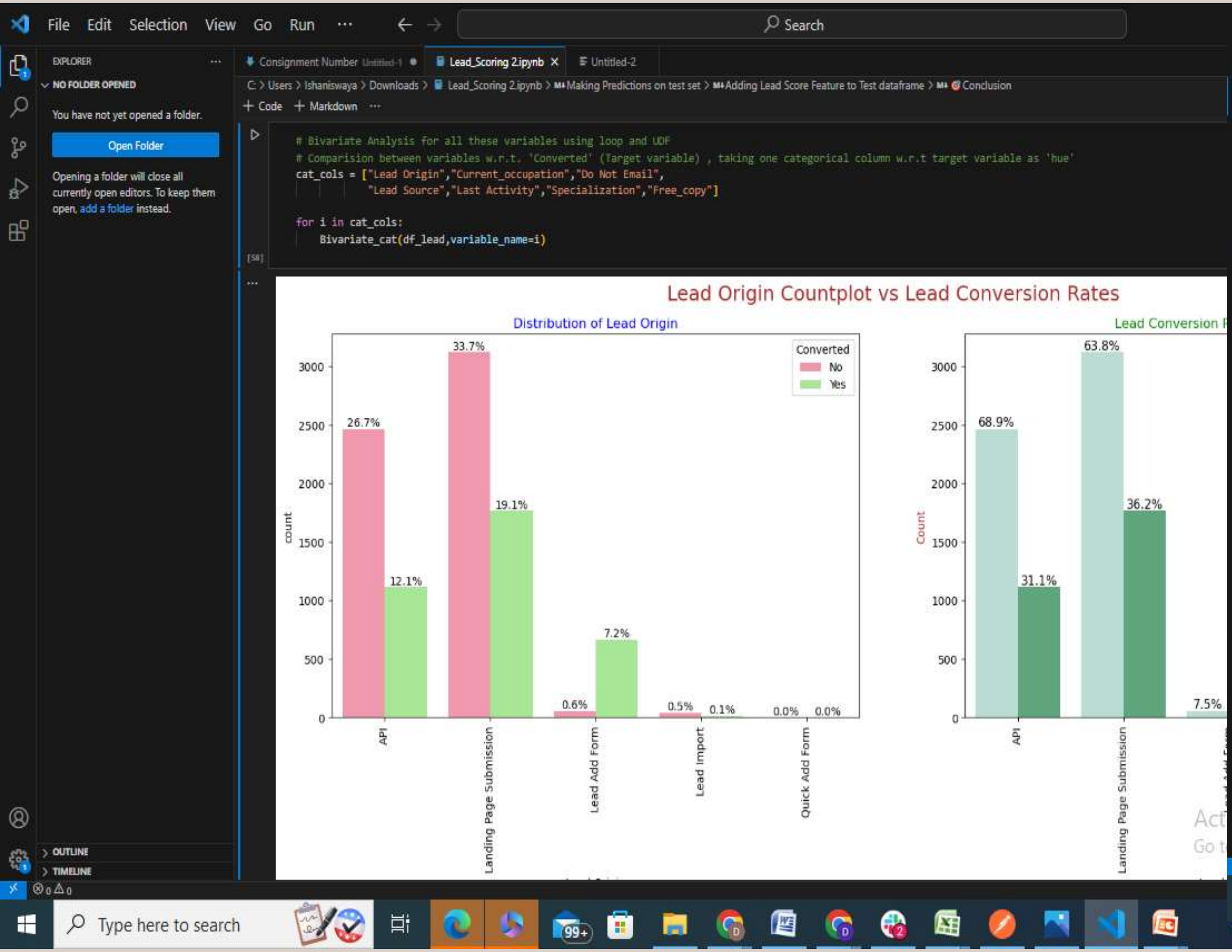
[56]



Observations:

In Categorical Univariate Analysis we get to know the value counts percentage in each variable that how much is the distribution of values in each column.

With this we get some understanding that which variables can be used in Bivariate analysis.

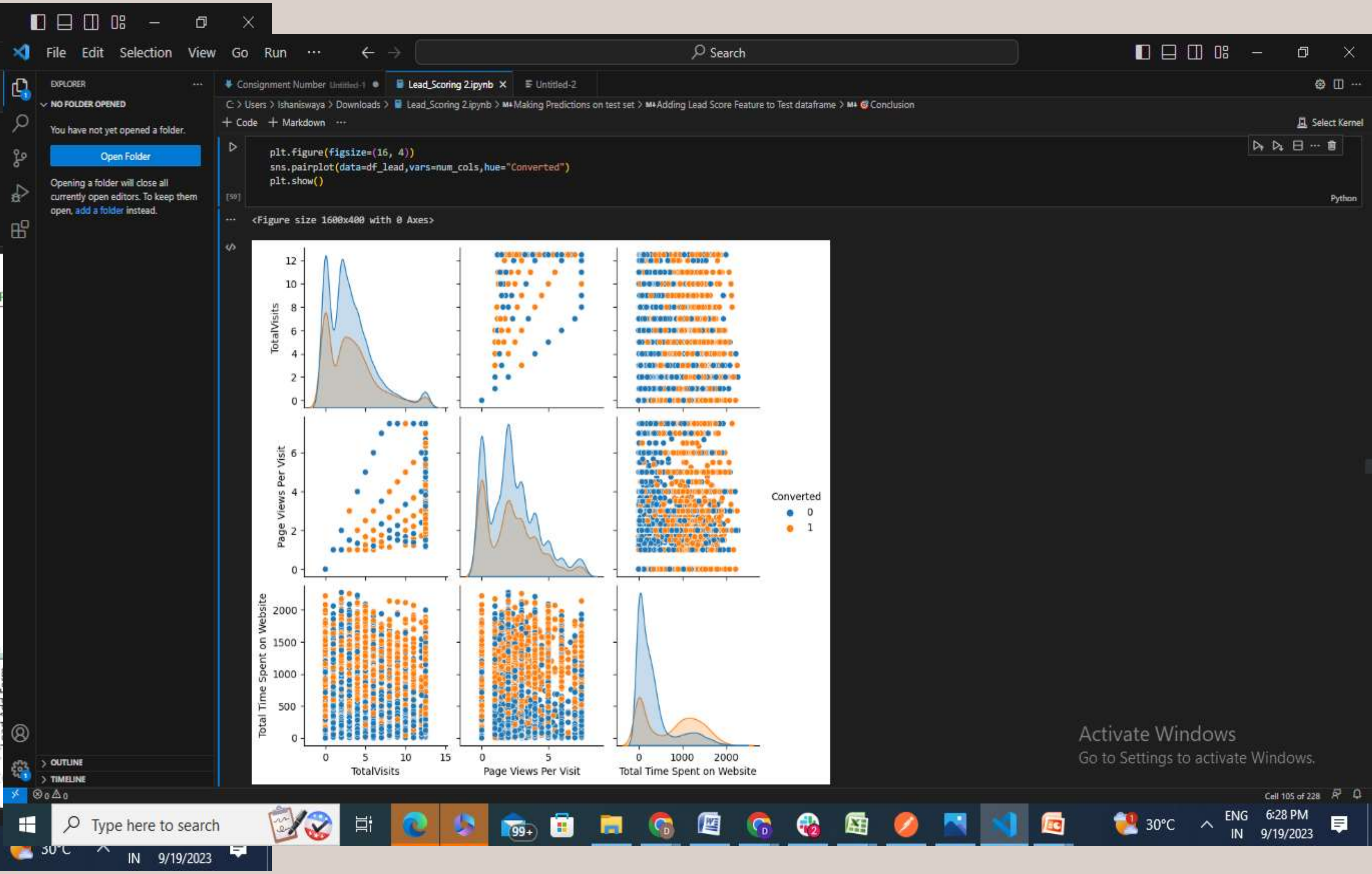


File Edit Selection View Go Run
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code + Markdown
OUTLINE
TIMELINE
Type here to search

```
# Bivariate Analysis for all these variables using loop and UDF
# Comparision between variables w.r.t. 'Converted' (Target variable) , taking one categorical column w.r.t target variable as 'hue'
cat_cols = ["Lead Origin","Current_occupation","Do Not Email",
            "Lead Source","Last Activity","Specialization","Free_copy"]

for i in cat_cols:
    Bivariate_cat(df_lead,variable_name=i)
```

[58]

Lead Origin Countplot vs Lead Conversion Rates
Distribution of Lead Origin
Converted
No
Yes
count
0
500
1000
1500
2000
2500
3000
26.7%
12.1%
33.7%
19.1%
0.6%
7.2%
0.5%
0.1%
0.0%
0.0%
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Conversion R
Count
68.9%
31.1%
63.8%
36.2%
7.5%
API
Landing Page Submission
30°C
IN 9/19/2023



File Edit Selection View Go Run
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code + Markdown
Select Kernel
Python
OUTLINE
TIMELINE
Cell 105 of 228
Type here to search
30°C
ENG IN
6:28 PM
9/19/2023

```
plt.figure(figsize=(16, 4))
sns.pairplot(data=df_lead,vars=num_cols,hue="Converted")
plt.show()
```

[59]

```
<Figure size 1600x400 with 0 Axes>
```

TotalVisits
Page Views Per Visit
Total Time Spent on Website
0
2
4
6
8
10
12
0
5
10
15
1000
1500
2000
Converted
0
1
Activate Windows
Go to Settings to activate Windows.

DATA MANIPULATION

- Total Number of Rows=37,Total Number of Columns =9240.

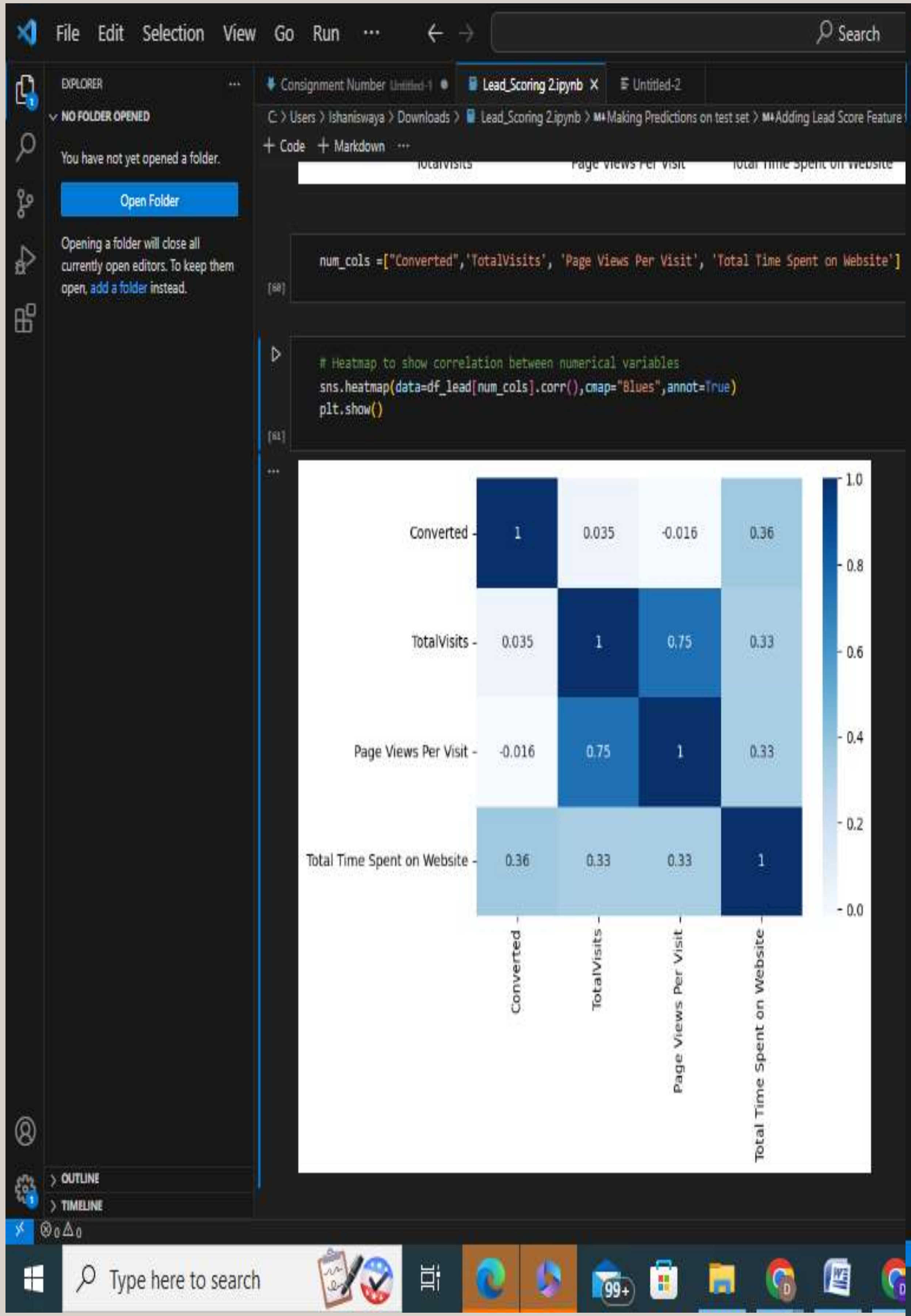
- Single value features like“Magazine”, “ReceiveMoreUpdates About Our Courses”, “Update my supply”

- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.

- Removing the“ProspectID” and “Lead Number” which are not necessary for the analysis.

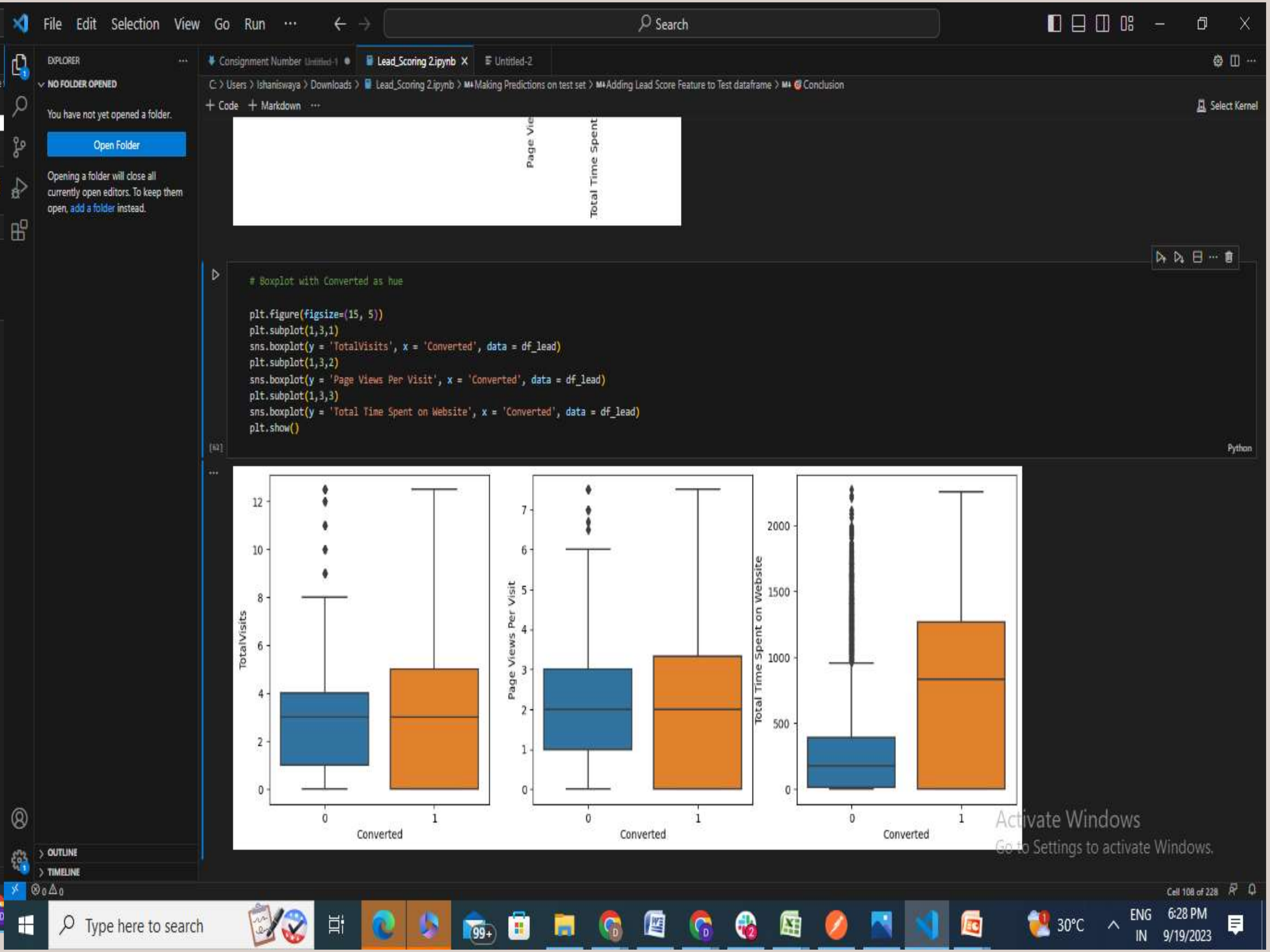
- After checking for the value counts for some of the object type variables, we find some of the features which have enough variance, which have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper, Article”, “XEducation Forums”, “Newspaper”, “DigitalAdvertisement” etc.

- Dropping the column shaving more than 35% as missing values such as ‘How did you hear about X Education’ and‘Lead Profile’.

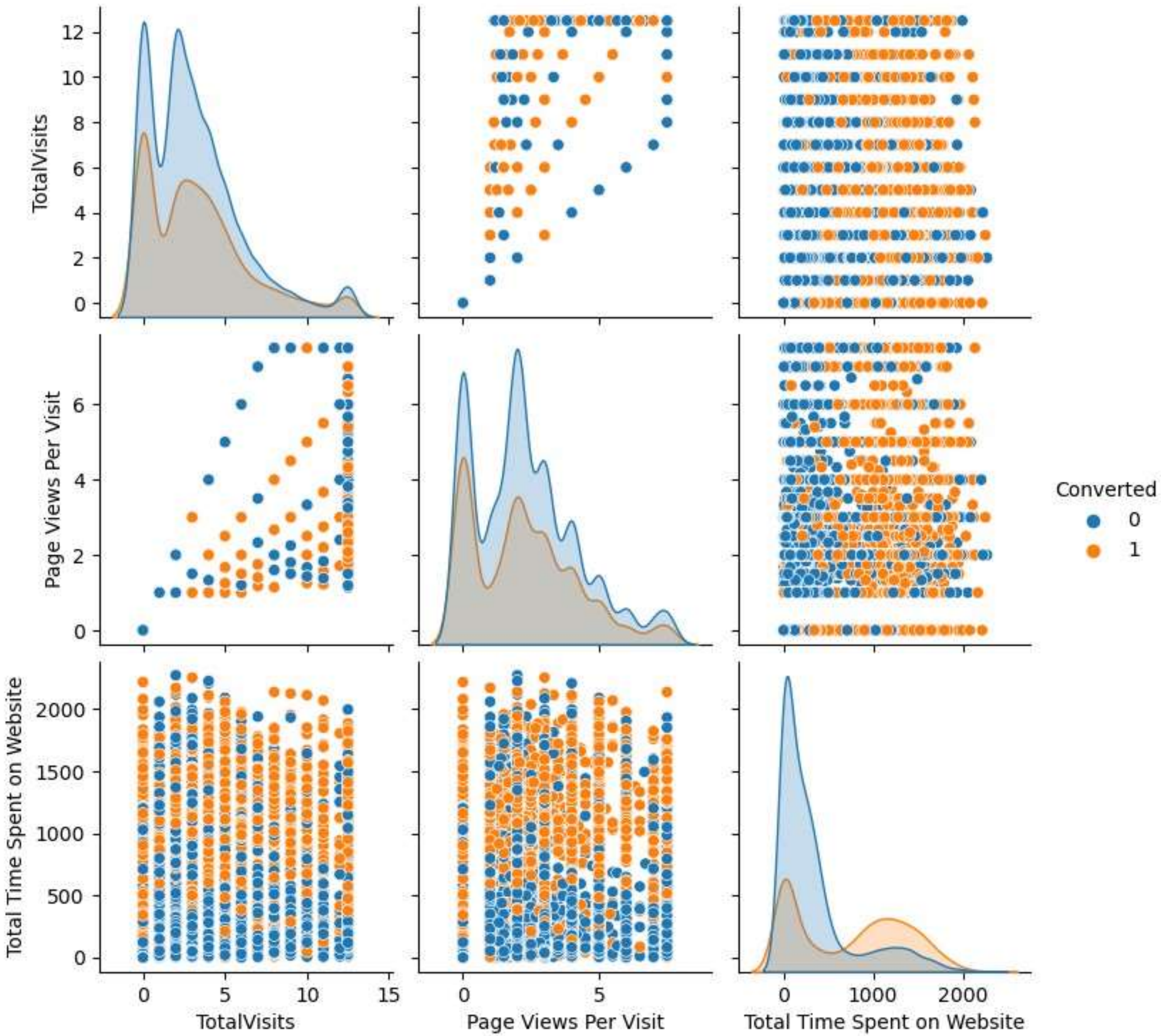

File Edit Selection View Go Run ···
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature
+ Code + Markdown
TotalVisits
Page Views Per Visit
Total Time Spent on Website
[60]
[61]
Converted
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1 0.035 -0.016 0.36
0.035 1 0.75 0.33
-0.016 0.75 1 0.33
0.36 0.33 0.33 1
1.0
0.8
0.6
0.4
0.2
0.0
OUTLINE
TIMELINE
Type here to search

```
num_cols =["Converted",'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']
```

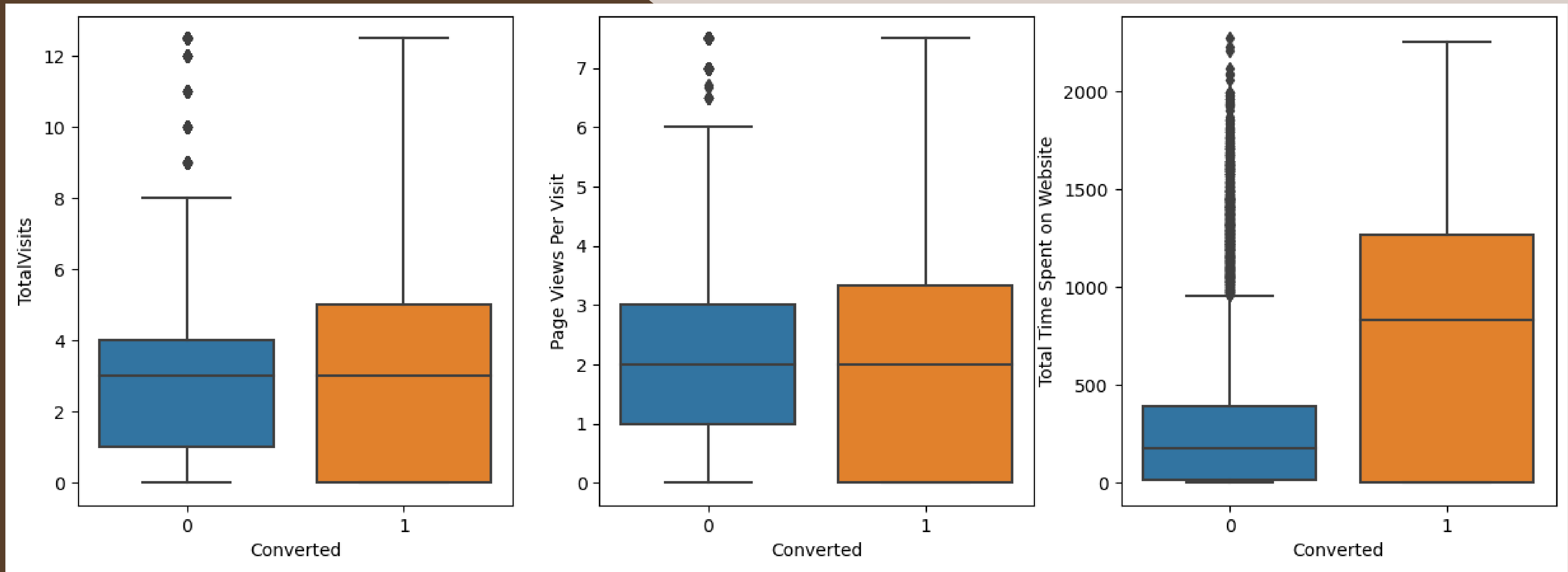
```
# Heatmap to show correlation between numerical variables
sns.heatmap(data=df_lead[num_cols].corr(),cmap="Blues",annot=True)
plt.show()
```


File Edit Selection View Go Run ···
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > Making Predictions on test set > Adding Lead Score Feature to Test dataframe > Conclusion
+ Code + Markdown
Select Kernel
Page Vie
Total Time Spent
[62]
Python
TotalVisits
12
10
8
6
4
2
0
Page Views Per Visit
7
6
5
4
3
2
1
0
Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted
Activate Windows
Go to Settings to activate Windows.
OUTLINE
TIMELINE
Cell 108 of 228
Type here to search
30°C
ENG
IN
6:28 PM
9/19/2023

```
# Boxplot with Converted as hue

plt.figure(figsize=(15, 5))
plt.subplot(1,3,1)
sns.boxplot(y = 'TotalVisits', x = 'Converted', data = df_lead)
plt.subplot(1,3,2)
sns.boxplot(y = 'Page Views Per Visit', x = 'Converted', data = df_lead)
plt.subplot(1,3,3)
sns.boxplot(y = 'Total Time Spent on Website', x = 'Converted', data = df_lead)
plt.show()
```

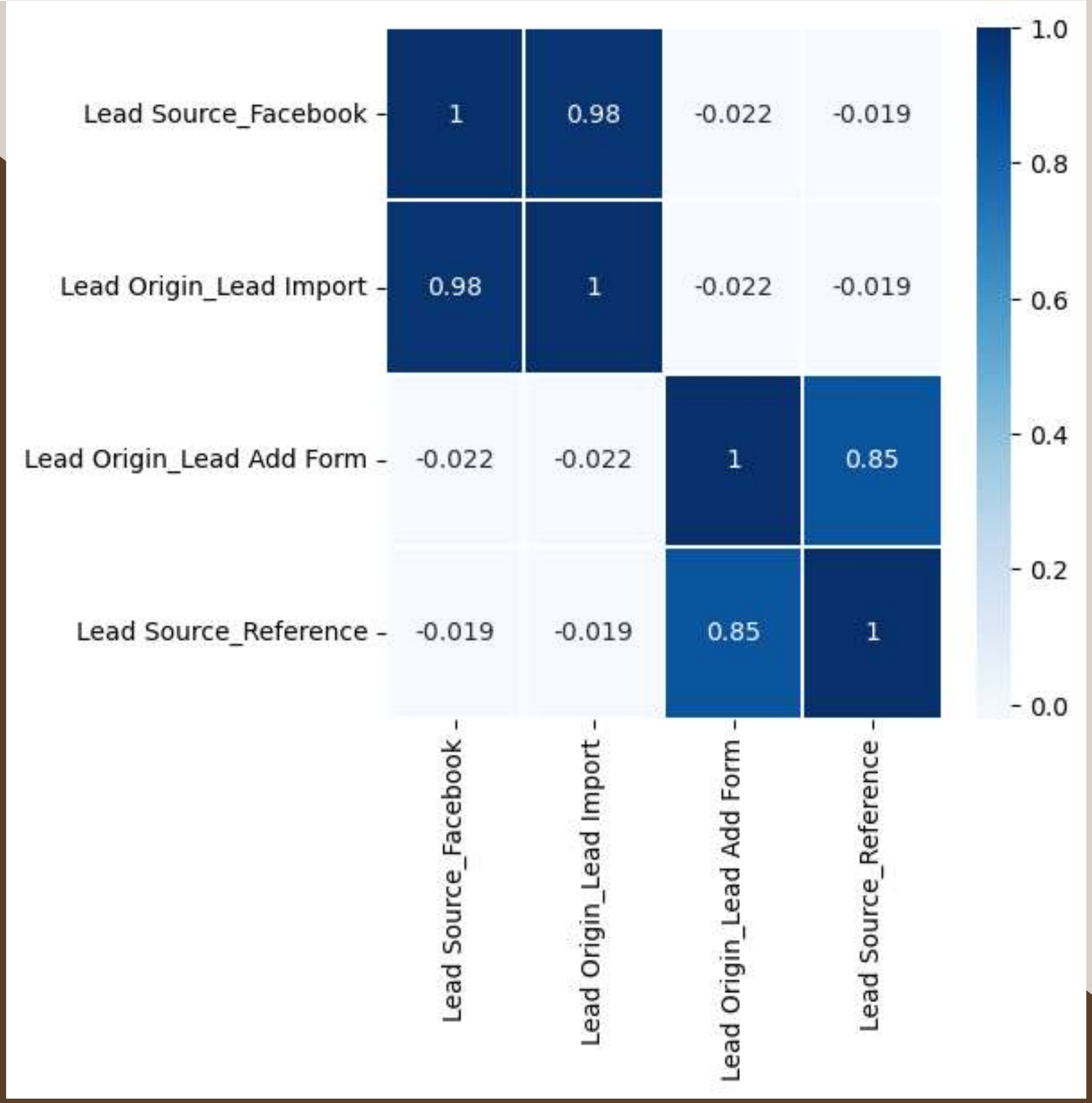


EXPLORATORY DATA ANALYSIS (EDA)



BOX PLOT

HEAT MAP



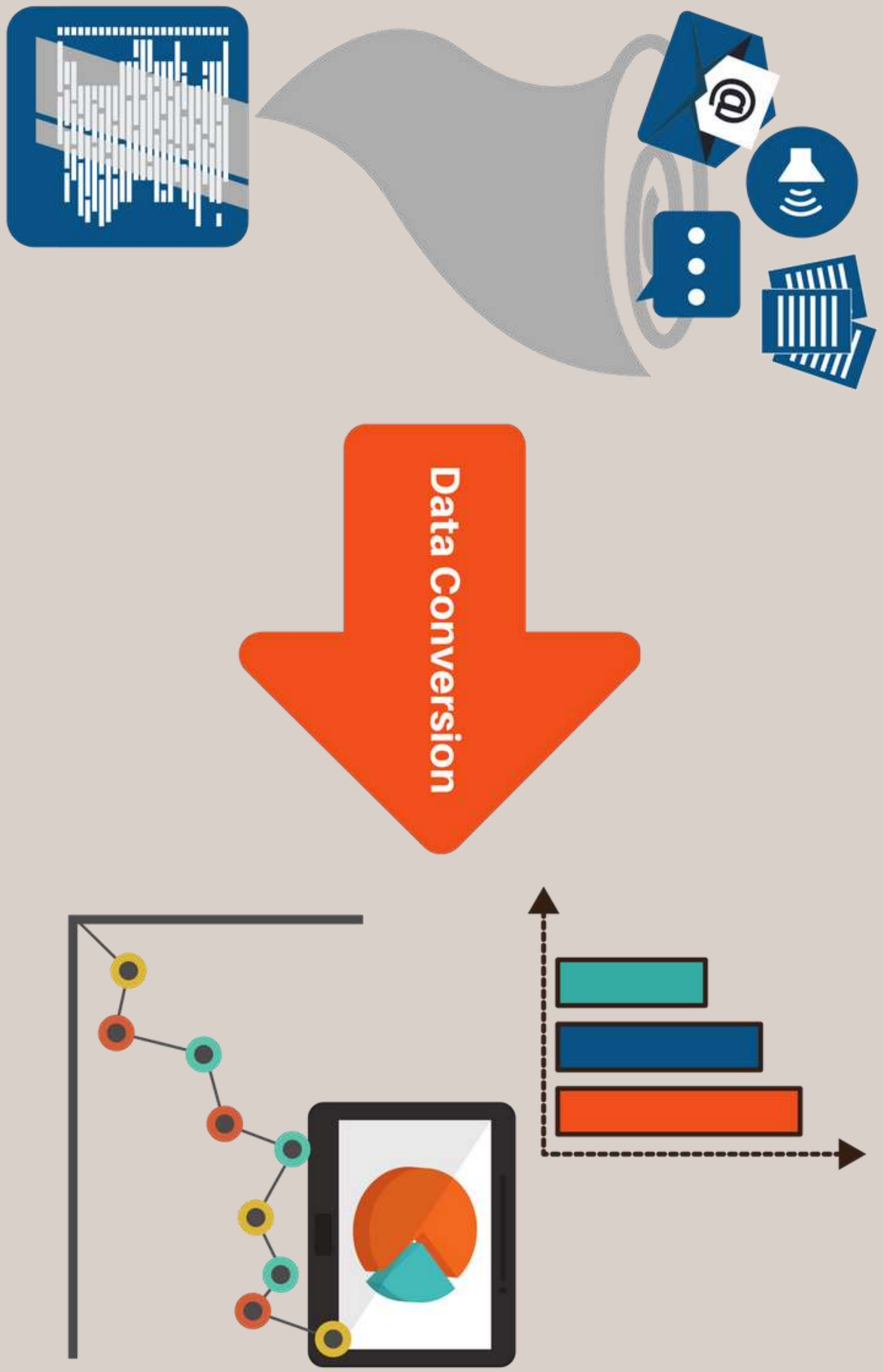
DATA CONVERSION

► Numerical Variables are normalized

► Dummy Variables are created for object type variables

► Total Rows forAnalysis: 9240

► Total Columns for Analysis: 37



MODEL BUILDING

- Splitting the Data into Training and Testing Sets

- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

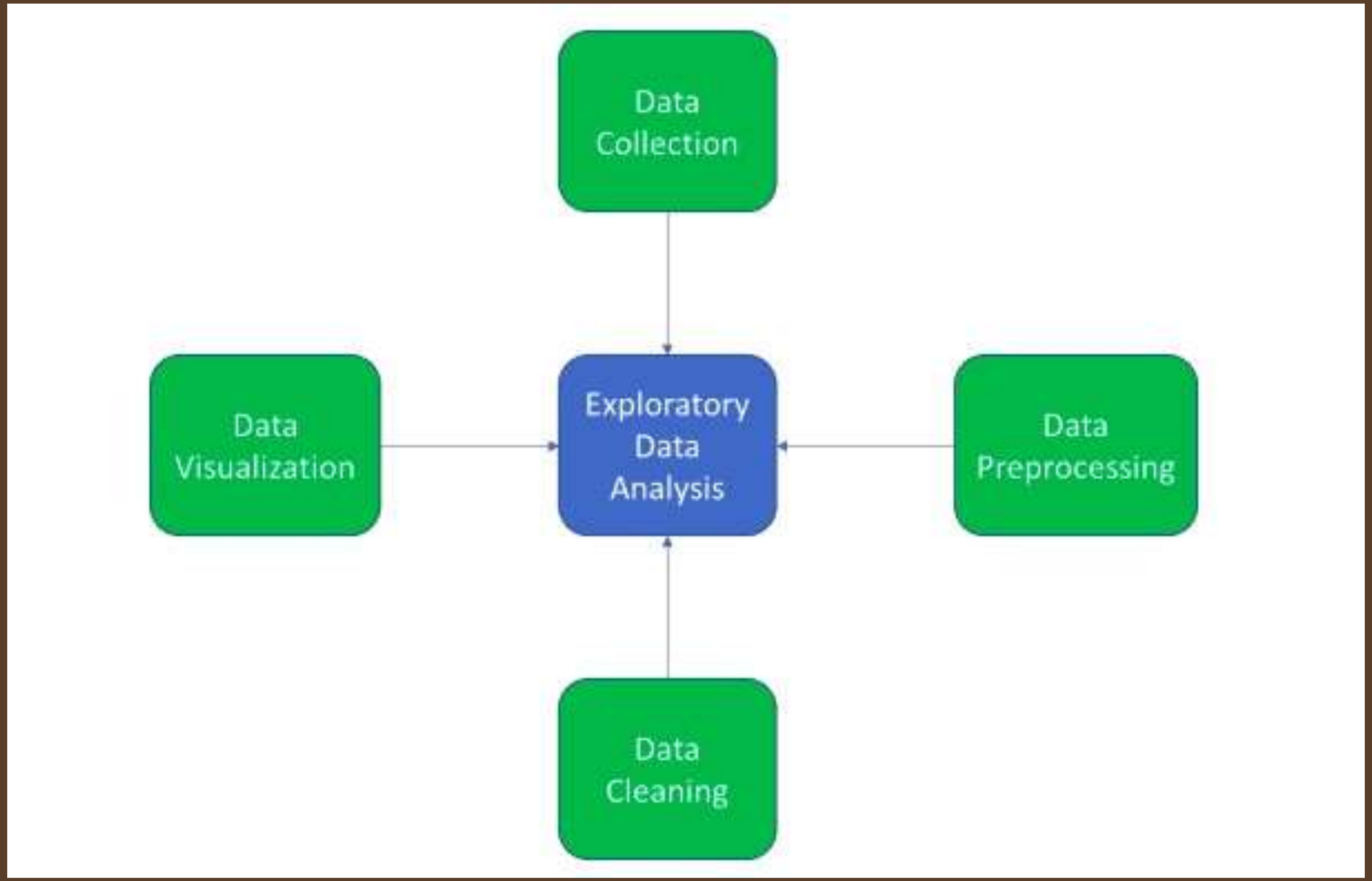
- Use RFE for Feature Selection

- Running RFE with 15 variables as output

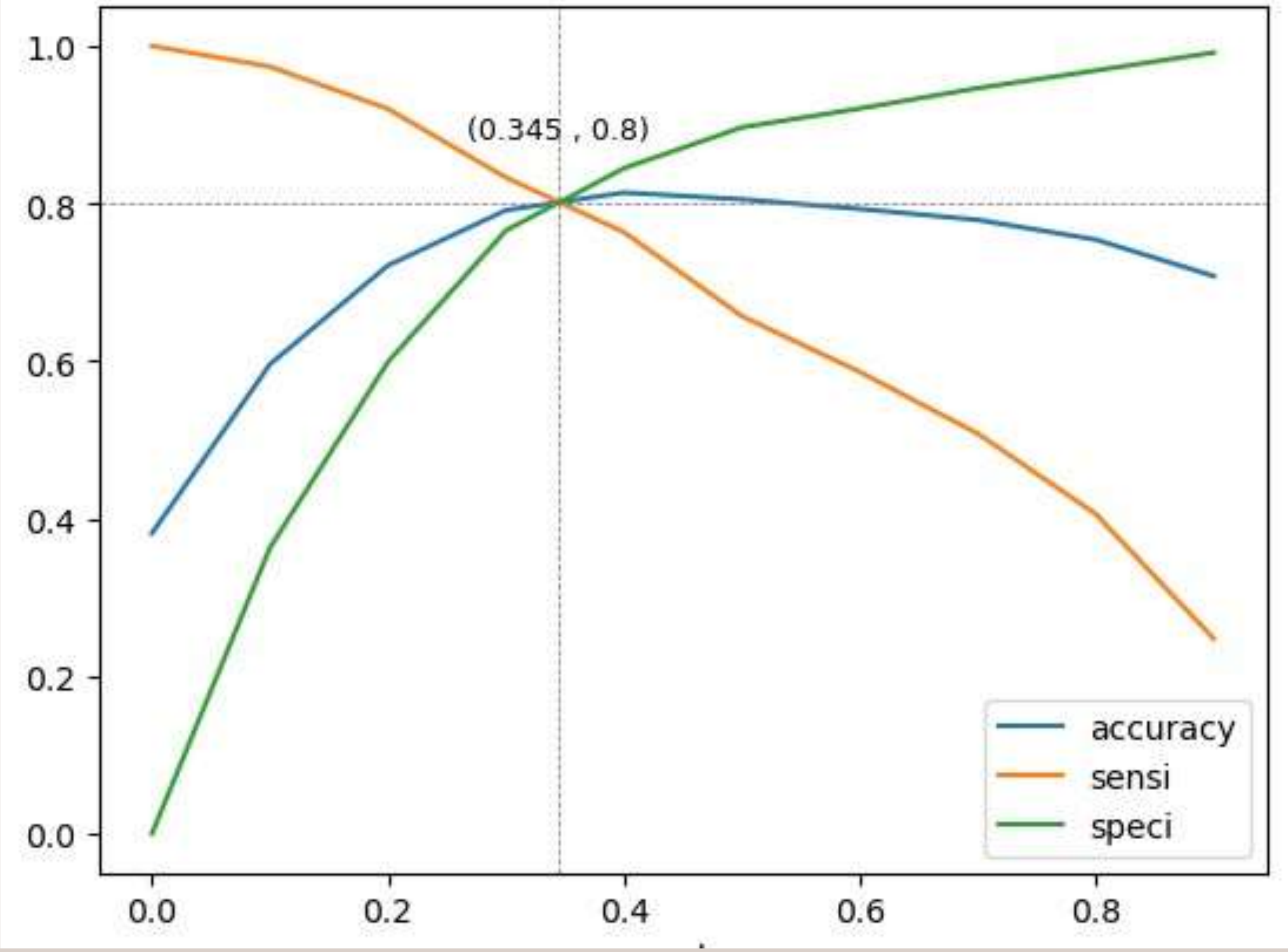
- Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5

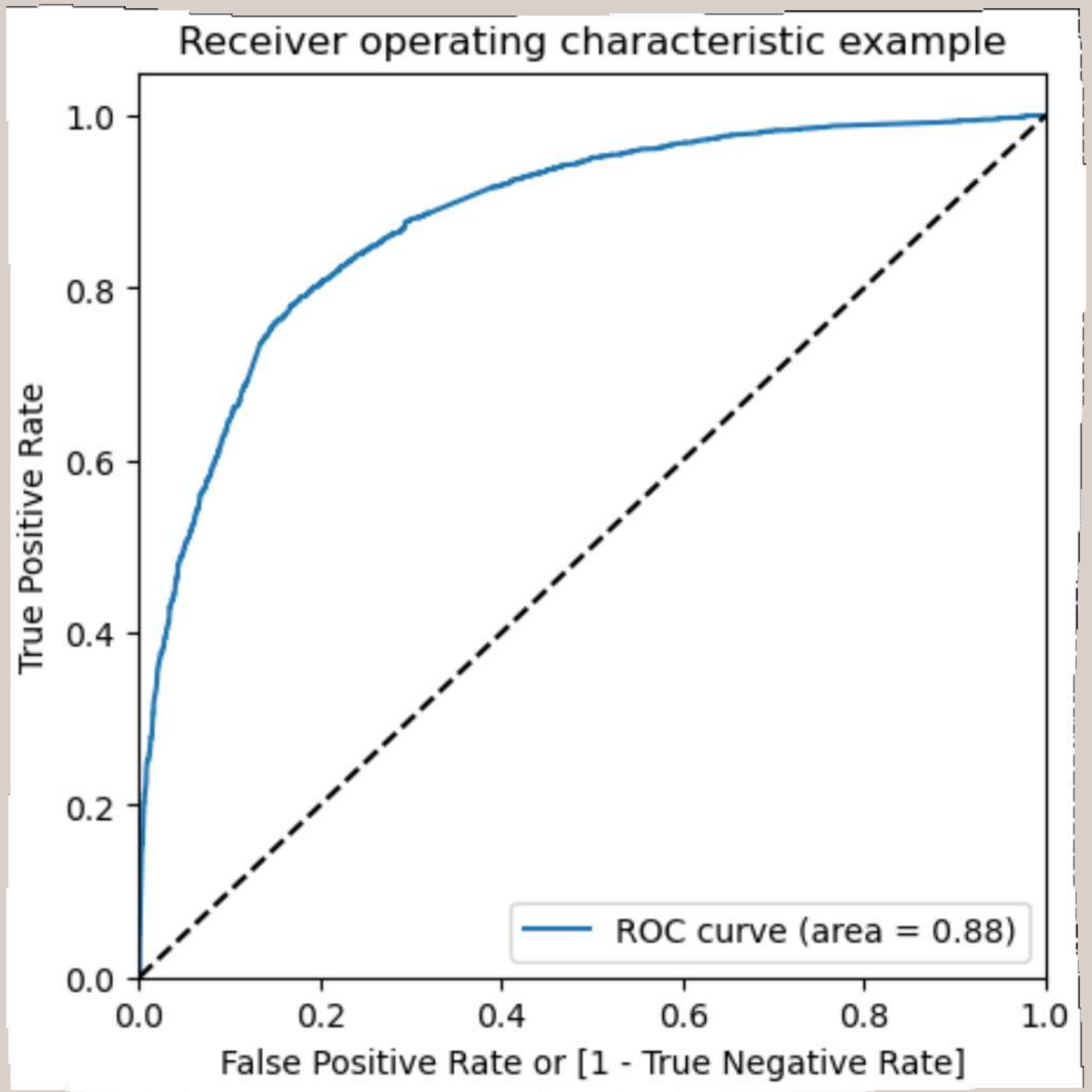
- Predictions on test data set

- Overall accuracy 81%



ROC Curve



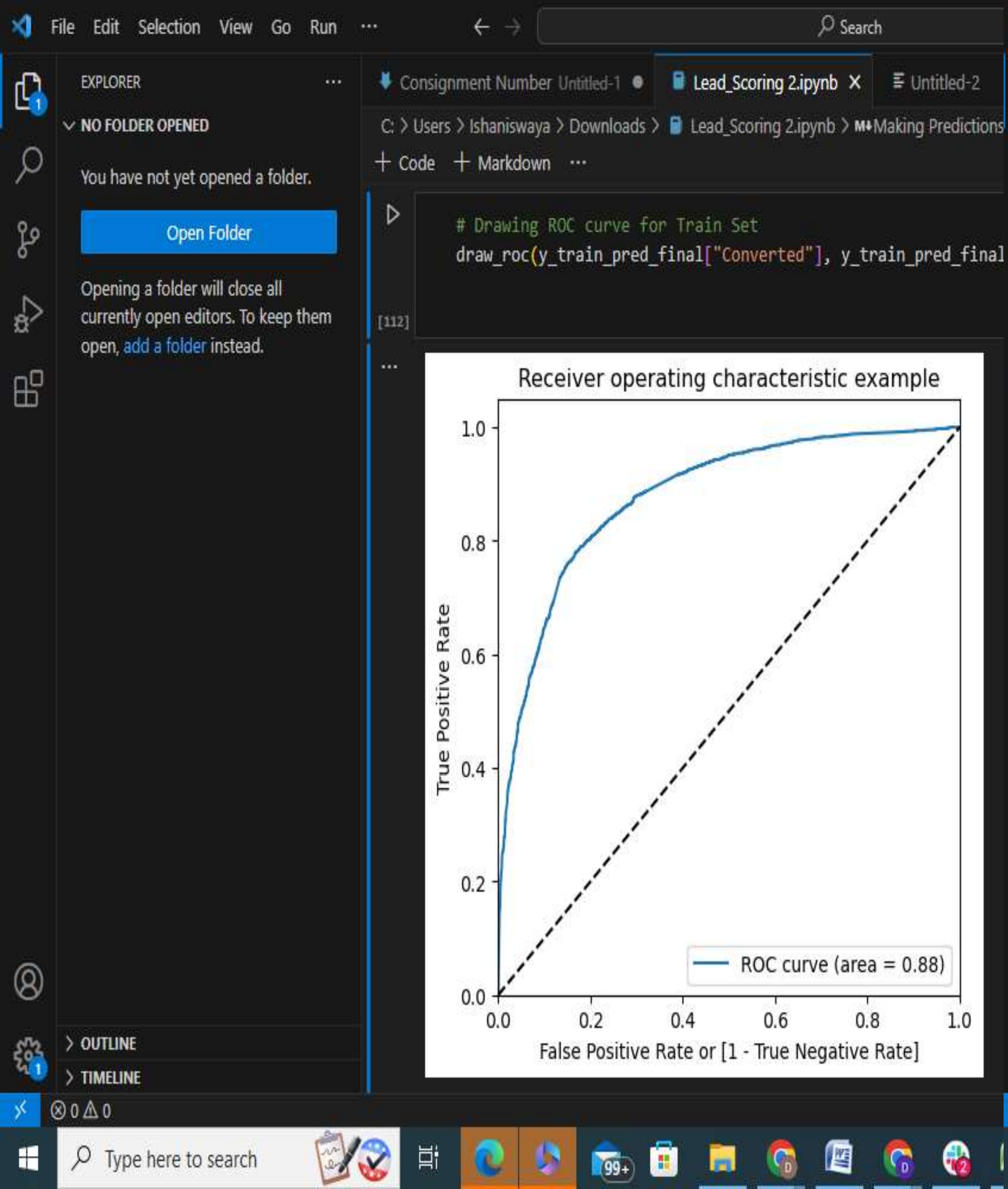


- Finding Optimal Cut off Point

- Optimal cut-off probability is that

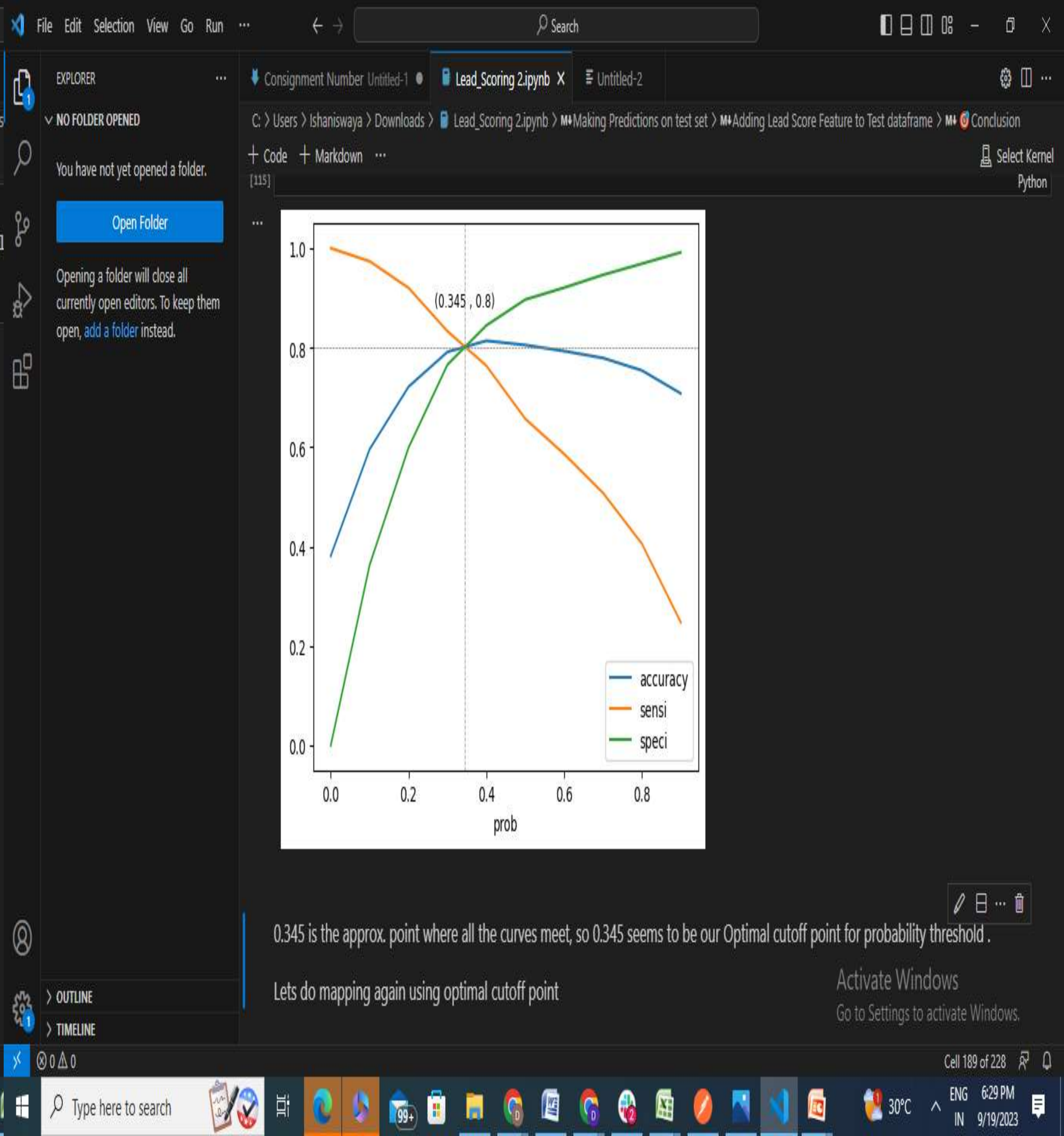
- Probability where we get balanced sensitivity and specificity.

- From the second graph it is visible that the optimal cut off is at 0.35.



File Edit Selection View Go Run ···
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
OUTLINE
TIMELINE
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > M↓Making Predictions
+ Code + Markdown ···
[112]
Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)
Type here to search

```
# Drawing ROC curve for Train Set
draw_roc(y_train_pred_final["Converted"], y_train_pred_final
```



File Edit Selection View Go Run ···
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
OUTLINE
TIMELINE
Consignment Number Untitled-1
Lead_Scoring 2.ipynb
Untitled-2
C: > Users > Ishaniswaya > Downloads > Lead_Scoring 2.ipynb > M↓Making Predictions on test set > M↓Adding Lead Score Feature to Test dataframe > M↓ Conclusion
+ Code + Markdown ···
Select Kernel
Python
[115]
(0.345 , 0.8)
1.0
0.8
0.6
0.4
0.2
0.0
prob
accuracy
sensi
speci
Activate Windows
Go to Settings to activate Windows.
Cell 189 of 228
Type here to search
30°C
ENG IN
6:29 PM
9/19/2023

0.345 is the approx. point where all the curves meet, so 0.345 seems to be our Optimal cutoff point for probability threshold .

Lets do mapping again using optimal cutoff point

PREDICTION ON TEST SET

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.

- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.

- After this we did model evaluation i.e. finding the accuracy, precision, and recall.

- The accuracy score we found was 0.82, precision 0.75, and recall 0.75 approximately.

- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.

- This also shows that our model is stable with good accuracy and recall/sensitivity.

- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spent on the Website.
- Total number of visits.
- When the lead source was:
  Google
  Direct traffic
  Organic search
  Welingak website
- When the last activity was:
  SMS
  Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

THANK YOU