Lead Scoring Case Study - Logistic Regression

By

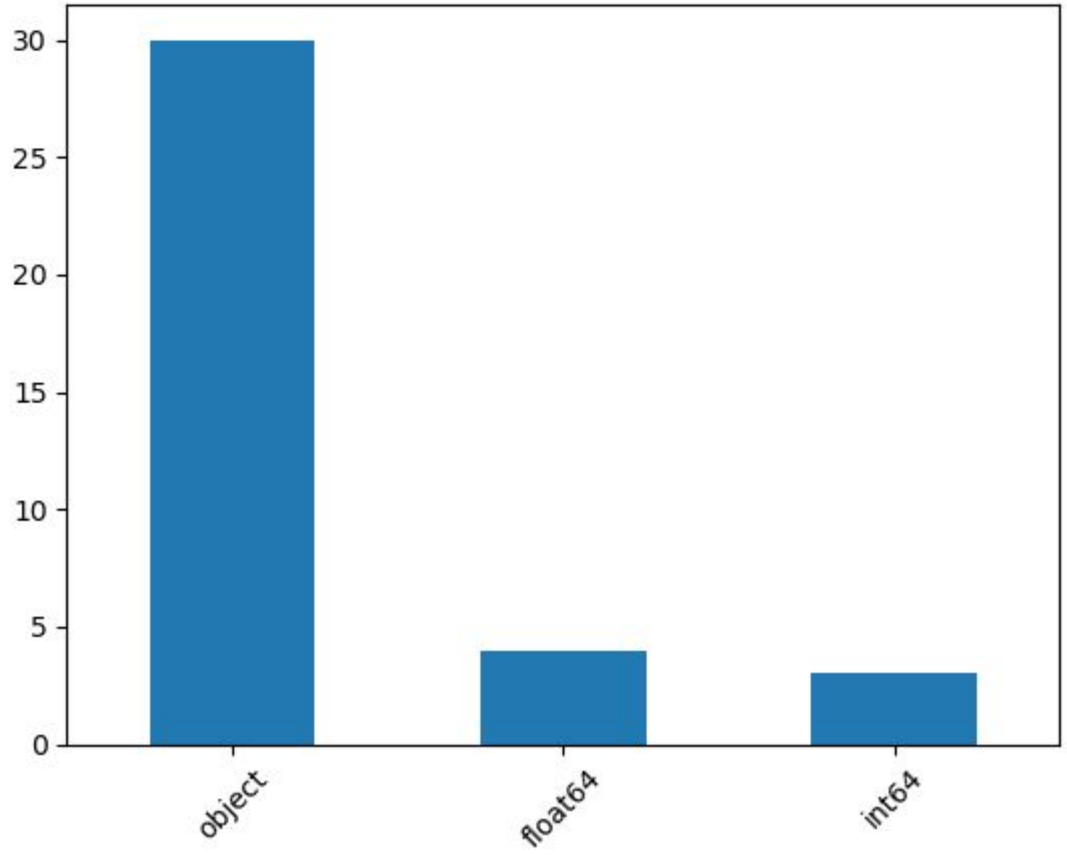
MOUNICA NAGULAPALLI

BHARATHNAN KUMARAN

SNEHA PADALA

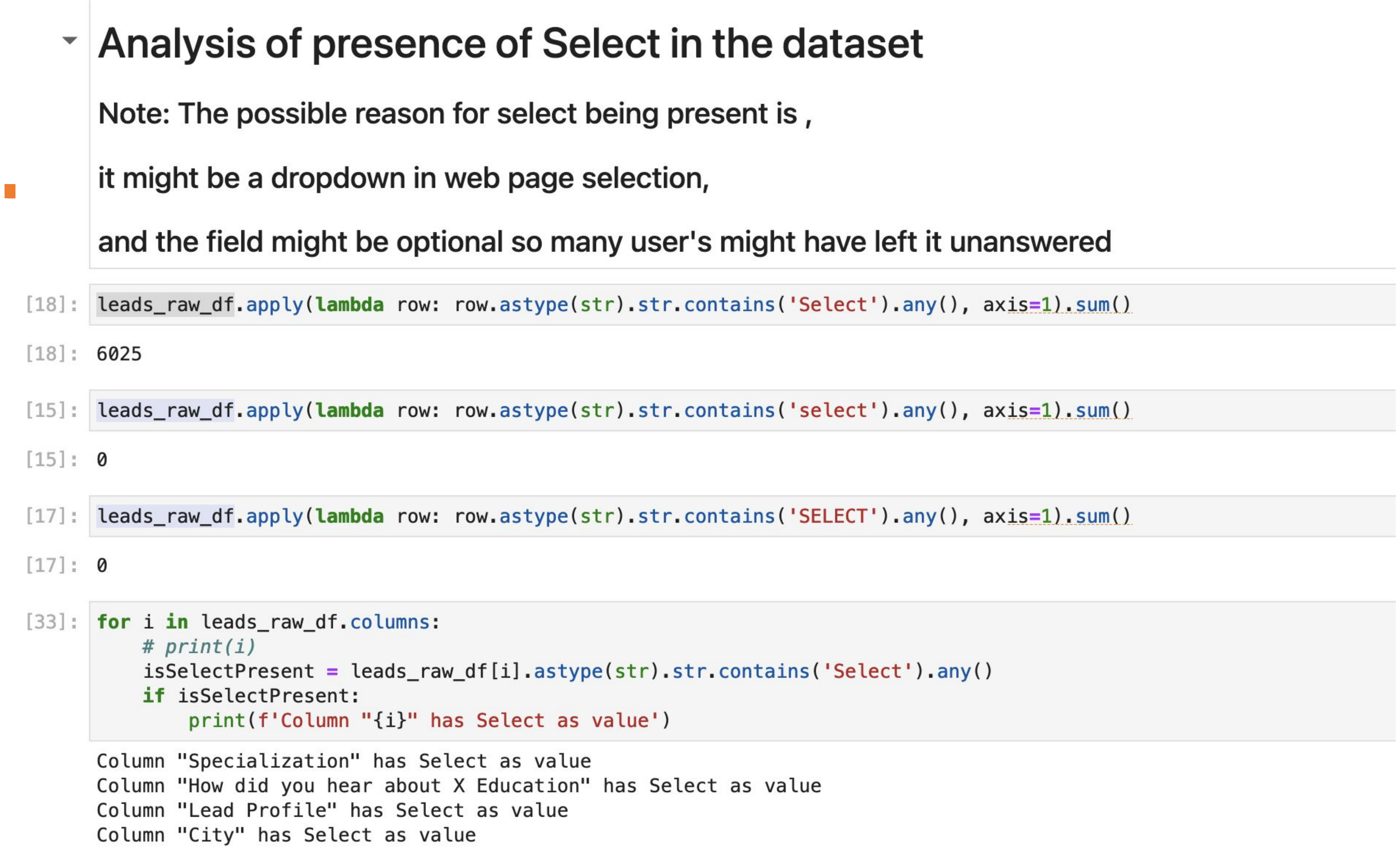
16th Dec 2023

Analyzing datatypes of application attributes



Below is the distribution of datatypes across different columns

float64	4
int64	3
object	30



Analysis of presence of Select in the dataset

Note: The possible reason for select being present is ,

it might be a dropdown in web page selection,

and the field might be optional so many user's might have left it unanswered

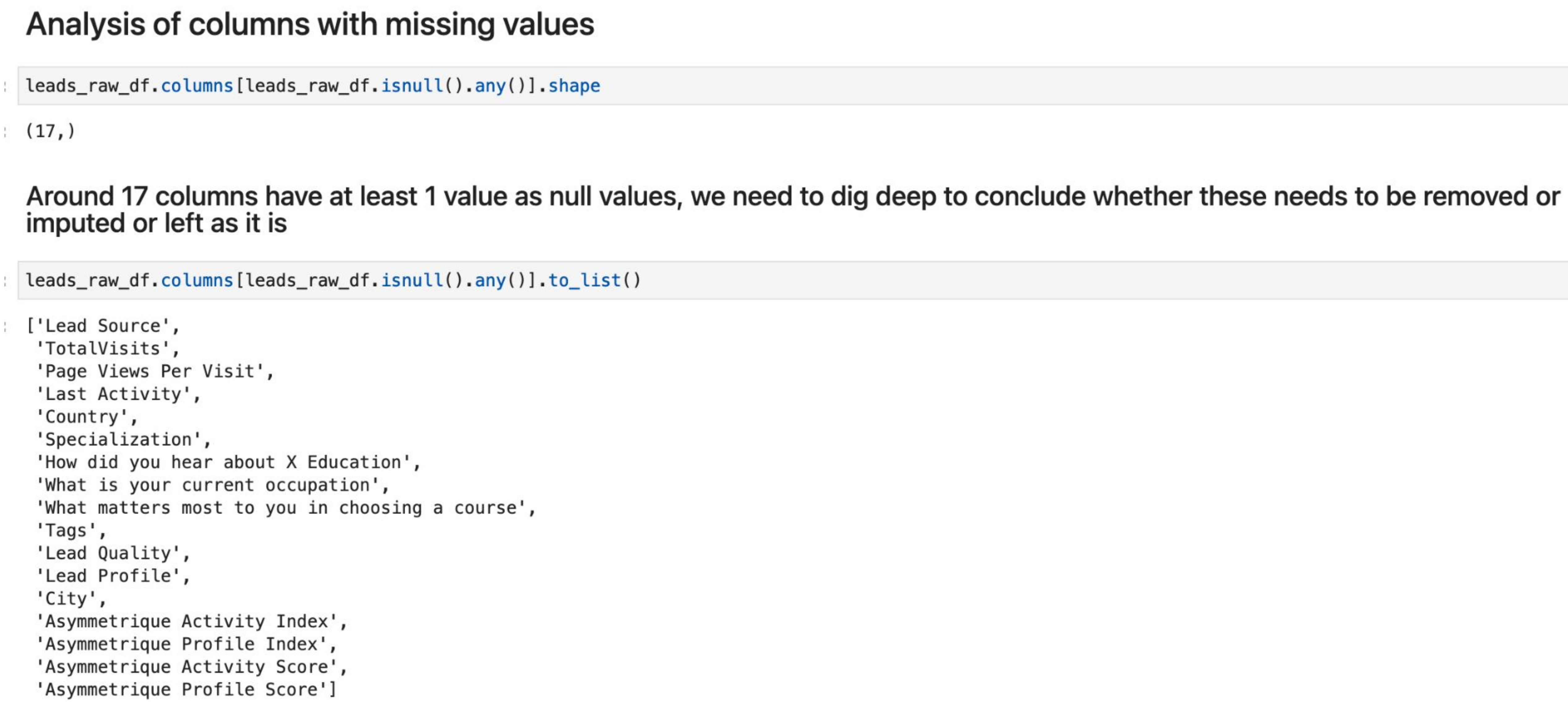
```
[18]: leads_raw_df.apply(lambda row: row.astype(str).str.contains('Select').any(), axis=1).sum()

[18]: 6025
```

```
[15]: leads_raw_df.apply(lambda row: row.astype(str).str.contains('select').any(), axis=1).sum()

[15]: 0
```

```
[17]: leads_raw_df.apply(lambda row: row.astype(str).str.contains('SELECT').any(), axis=1).sum()

[17]: 0
```

```
[33]: for i in leads_raw_df.columns:
          # print(i)
          isSelectPresent = leads_raw_df[i].astype(str).str.contains('Select').any()
          if isSelectPresent:
              print(f'Column "{i}" has Select as value')
```

```
Column "Specialization" has Select as value
Column "How did you hear about X Education" has Select as value
Column "Lead Profile" has Select as value
Column "City" has Select as value
```



Analysis of columns with missing values

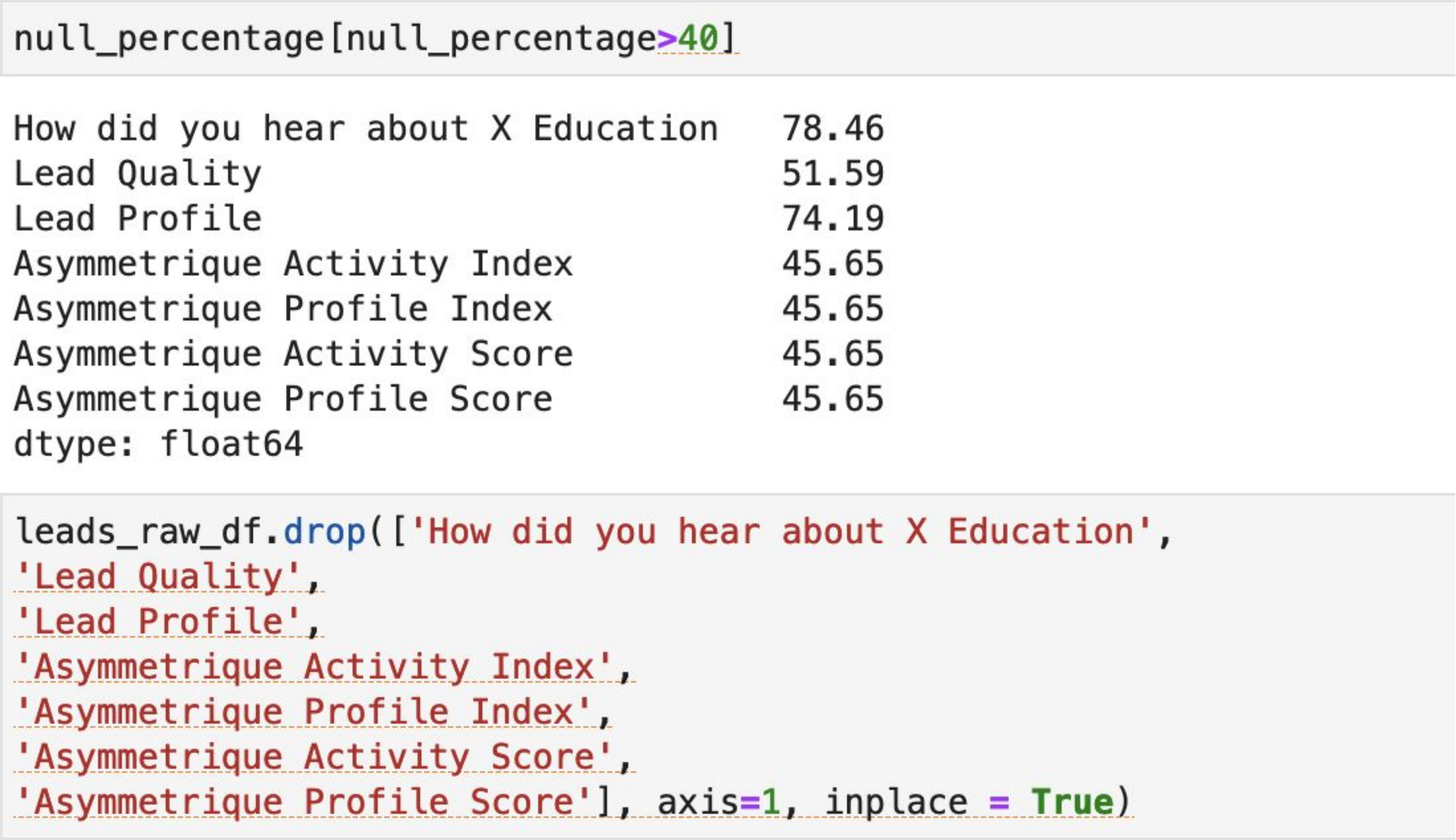
```
leads_raw_df.columns[leads_raw_df.isnull().any()].shape
```

```
(17,)
```

Around 17 columns have at least 1 value as null values, we need to dig deep to conclude whether these needs to be removed or imputed or left as it is

```
leads_raw_df.columns[leads_raw_df.isnull().any()].to_list()
```

```
['Lead Source',
 'TotalVisits',
 'Page Views Per Visit',
 'Last Activity',
 'Country',
 'Specialization',
 'How did you hear about X Education',
 'What is your current occupation',
 'What matters most to you in choosing a course',
 'Tags',
 'Lead Quality',
 'Lead Profile',
 'City',
 'Asymmetrique Activity Index',
 'Asymmetrique Profile Index',
 'Asymmetrique Activity Score',
 'Asymmetrique Profile Score']
```

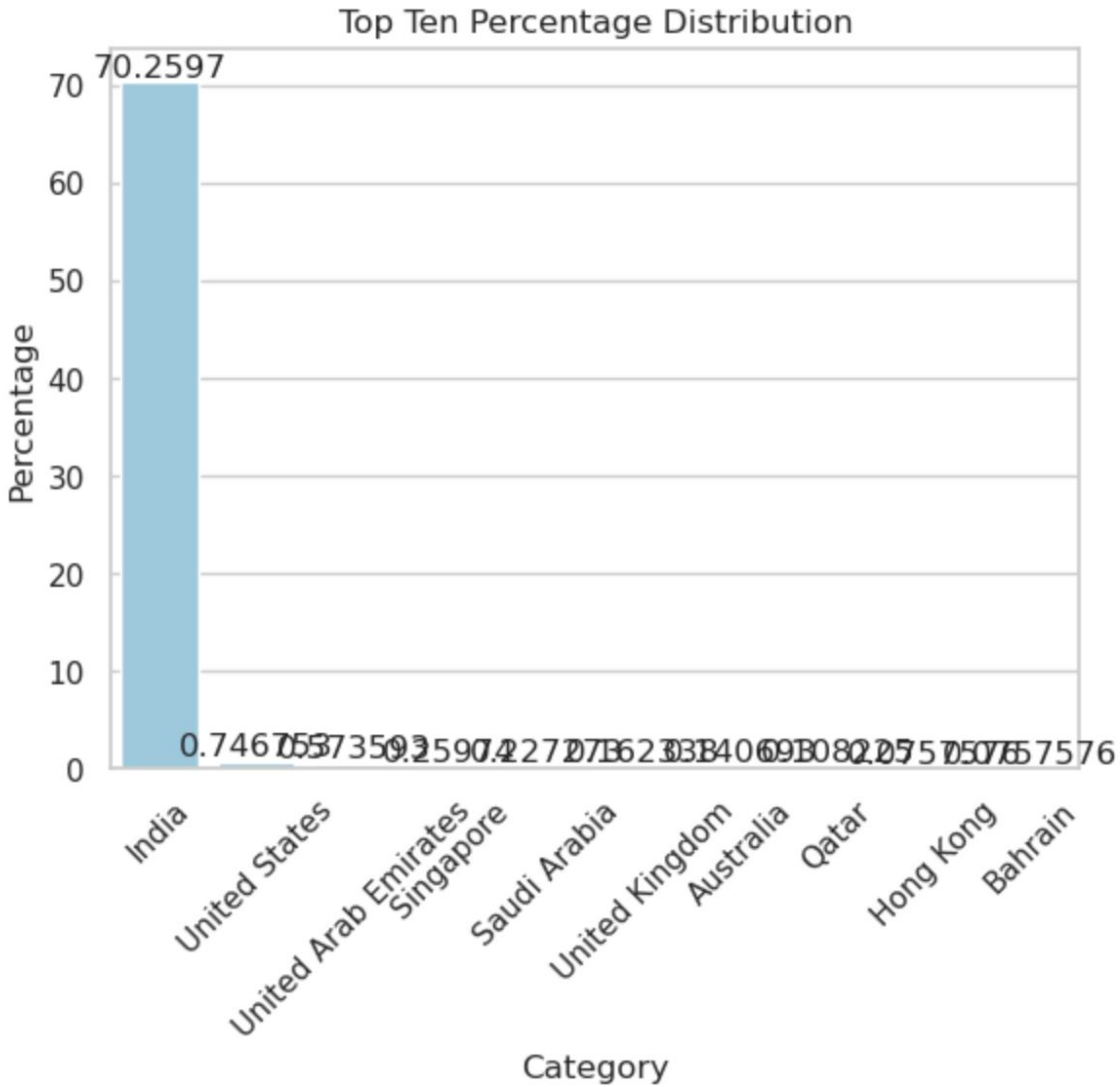


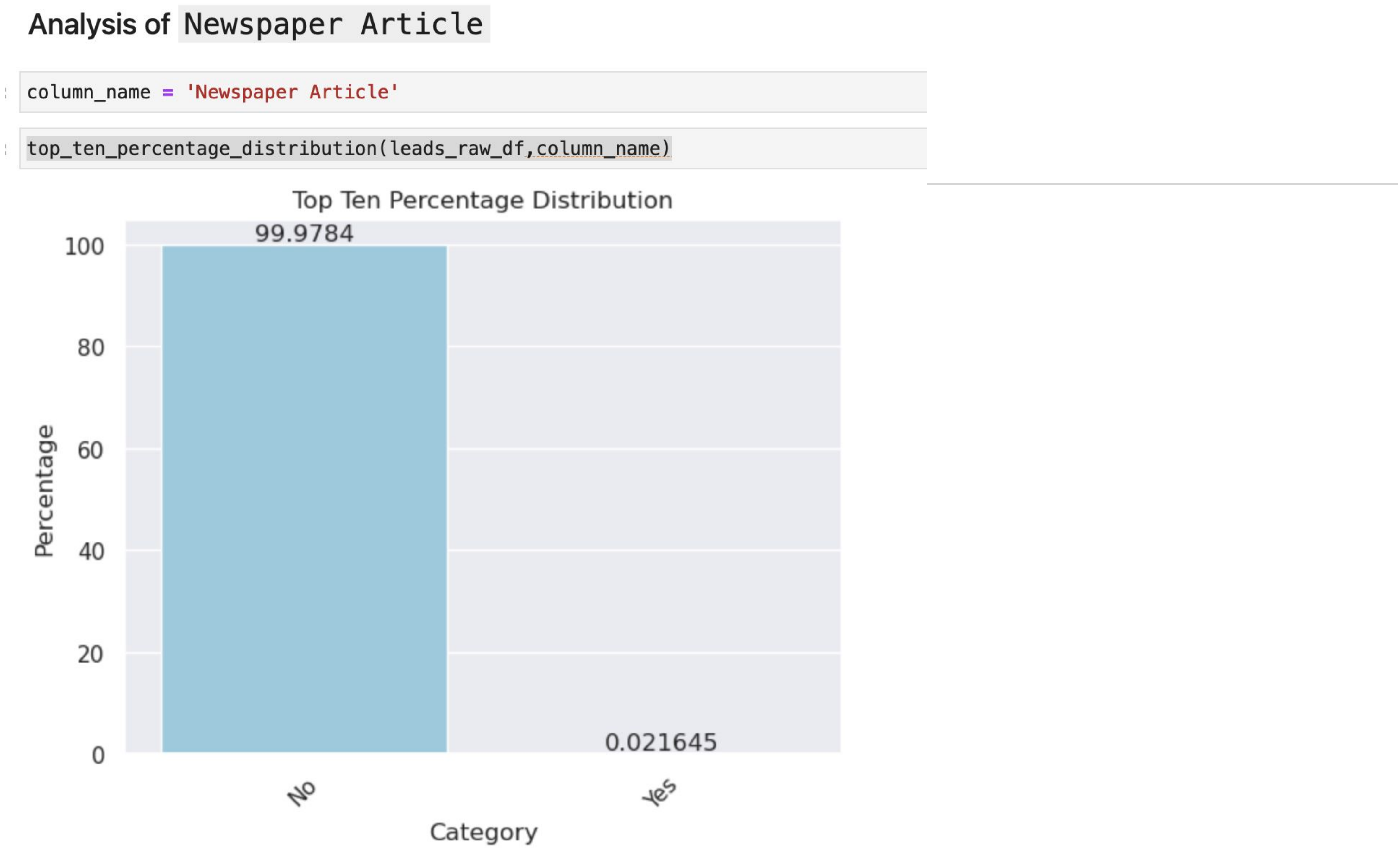
```
null_percentage[null_percentage>40]
```

```
How did you hear about X Education   78.46
Lead Quality                          51.59
Lead Profile                          74.19
Asymmetrique Activity Index           45.65
Asymmetrique Profile Index            45.65
Asymmetrique Activity Score           45.65
Asymmetrique Profile Score            45.65
dtype: float64
```

```
leads_raw_df.drop(['How did you hear about X Education',
'Lead Quality',
'Lead Profile',
'Asymmetrique Activity Index',
'Asymmetrique Profile Index',
'Asymmetrique Activity Score',
'Asymmetrique Profile Score'], axis=1, inplace = True)
```

```
[231]: top_ten_percentage_distribution(leads_raw_df,'Country')
```

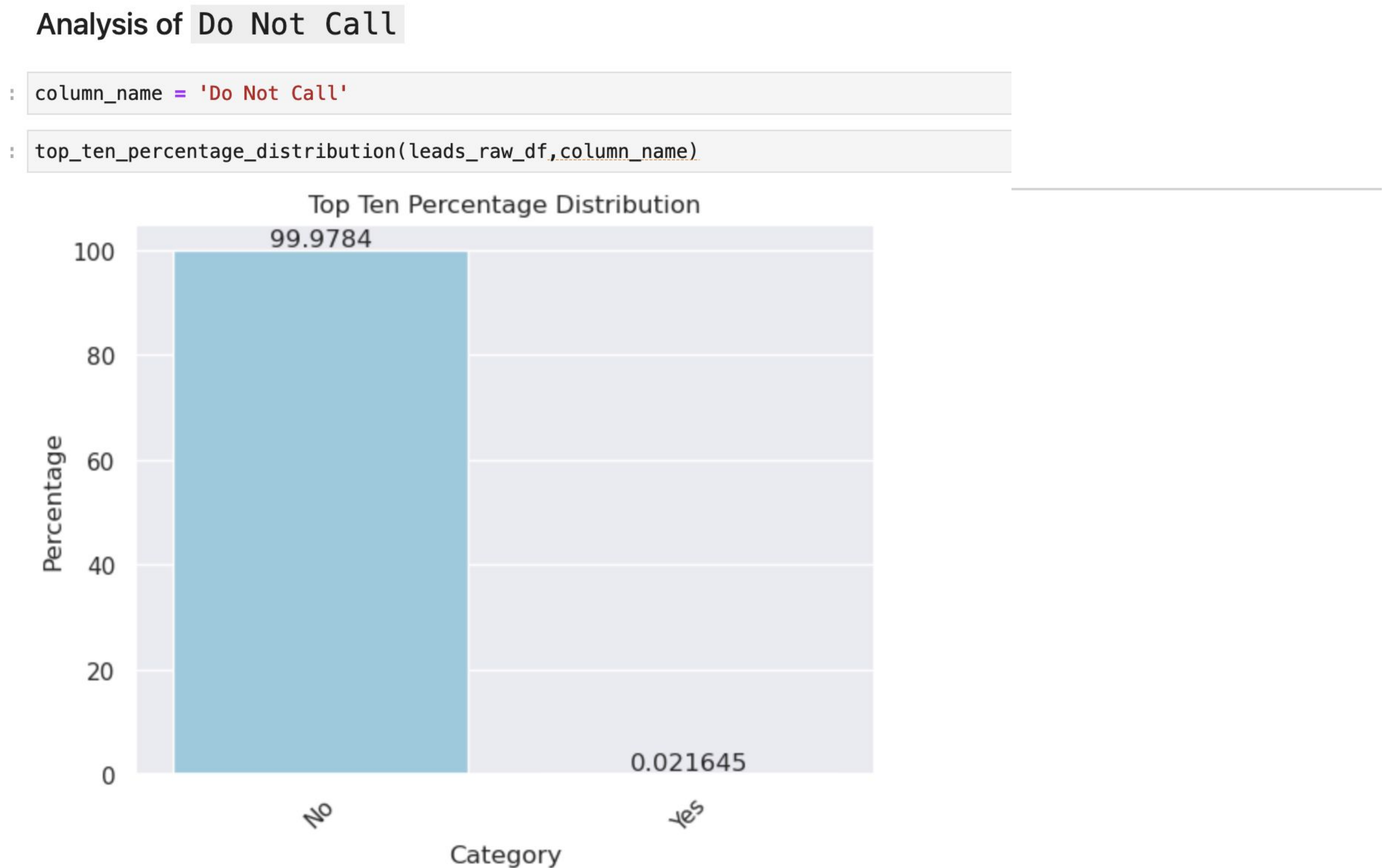




Analysis of `Newspaper Article`

```
column_name = 'Newspaper Article'
```

```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



Analysis of `Do Not Call`

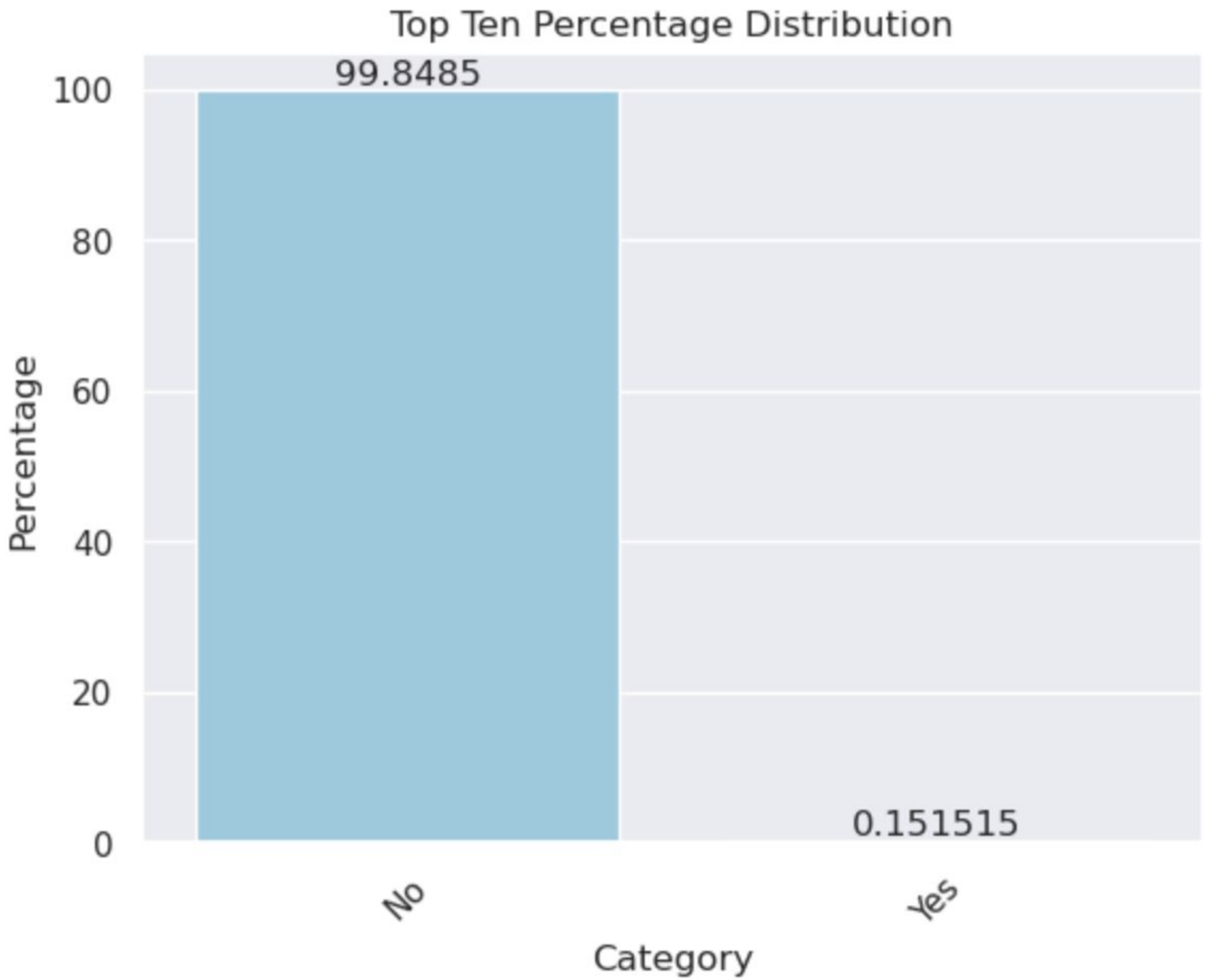
```
column_name = 'Do Not Call'
```

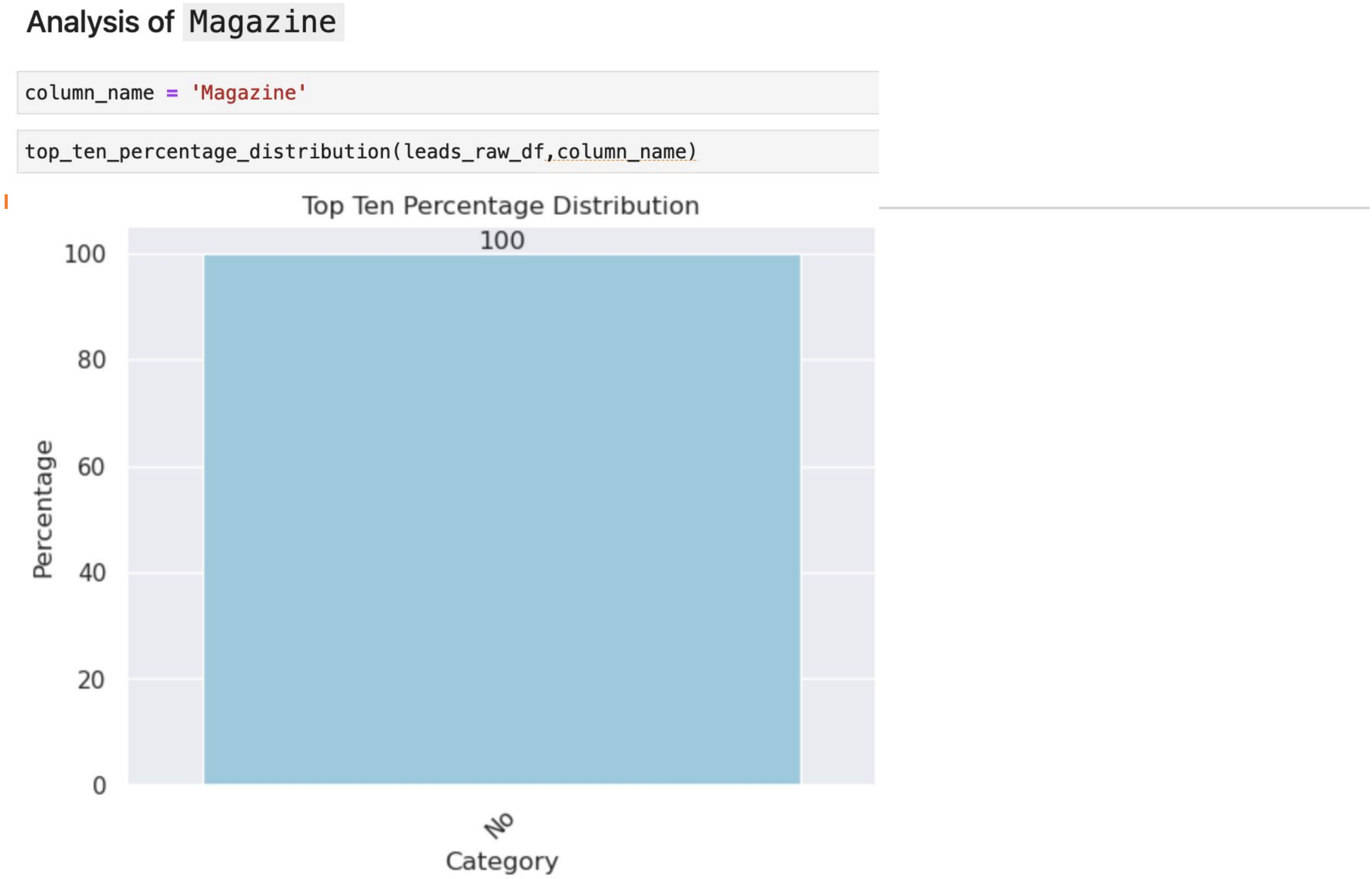
```
top_ten_percentage_distribution(leads_raw_df,column_name)
```

Analysis of `Search`

```
column_name = 'Search'
```

```
top_ten_percentage_distribution(leads_raw_df,column_name)
```

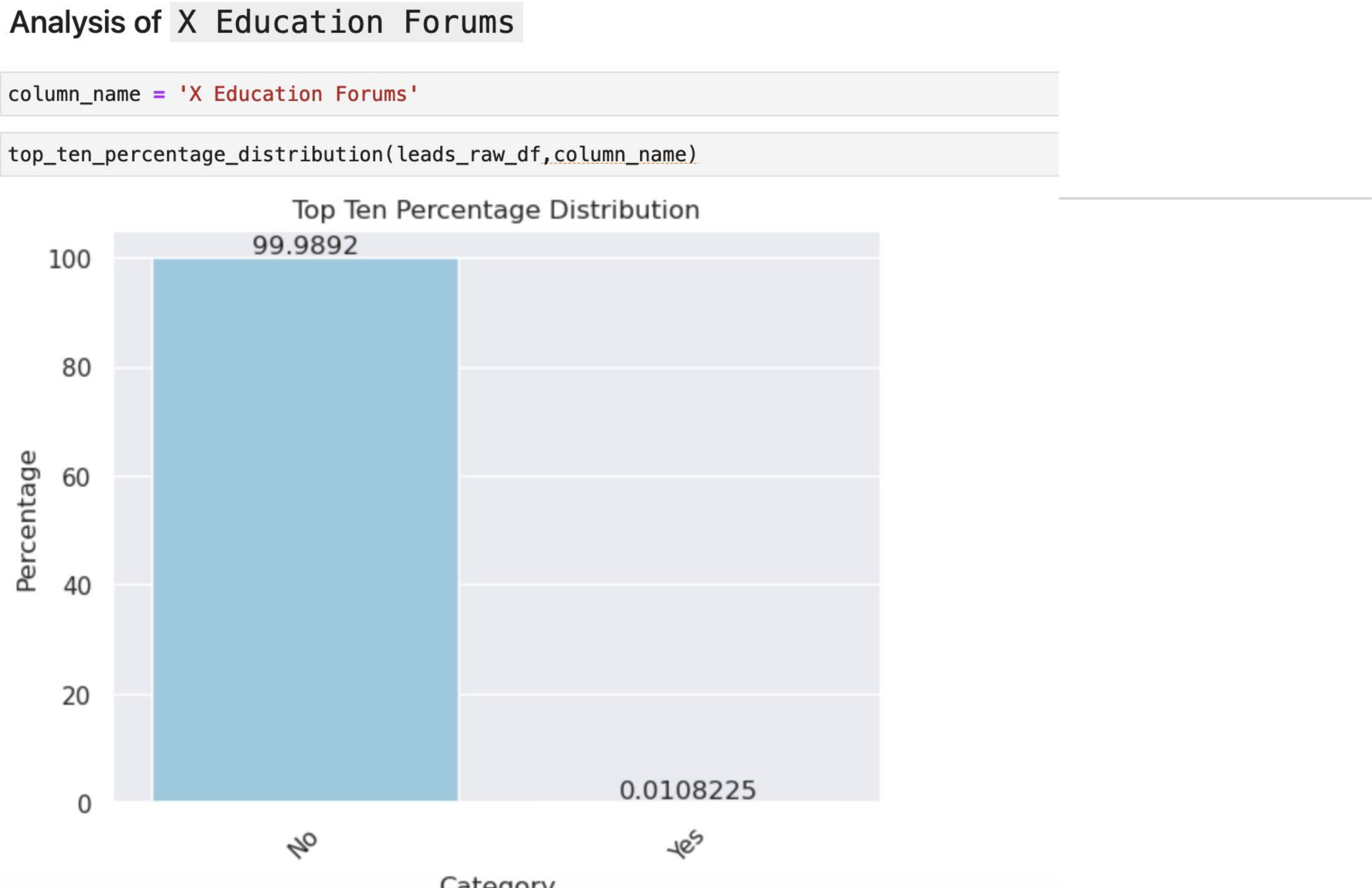




Analysis of `Magazine`

```
column_name = 'Magazine'
```

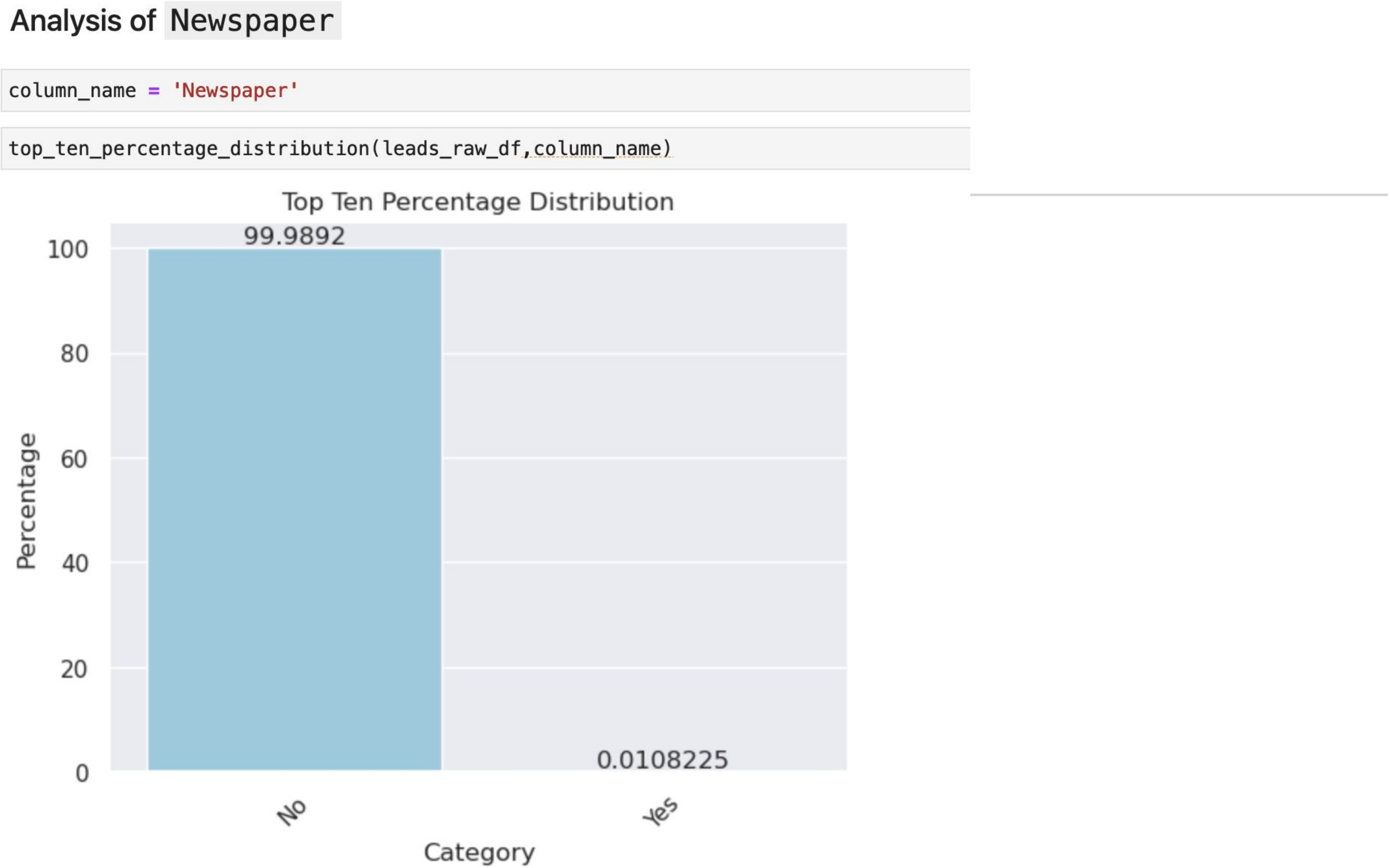
```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



Analysis of `X Education Forums`

```
column_name = 'X Education Forums'
```

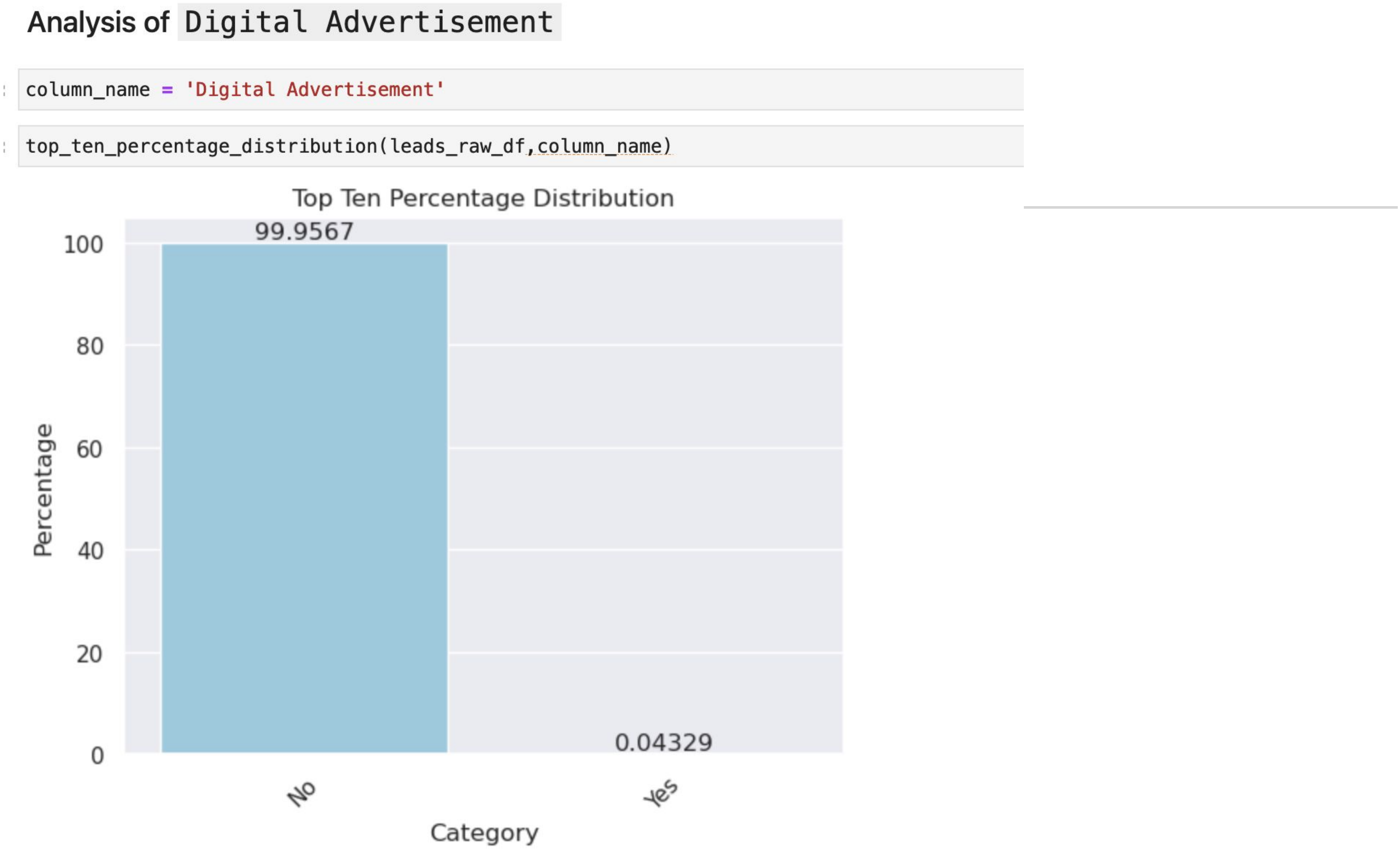
```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



Analysis of `Newspaper`

```
column_name = 'Newspaper'
```

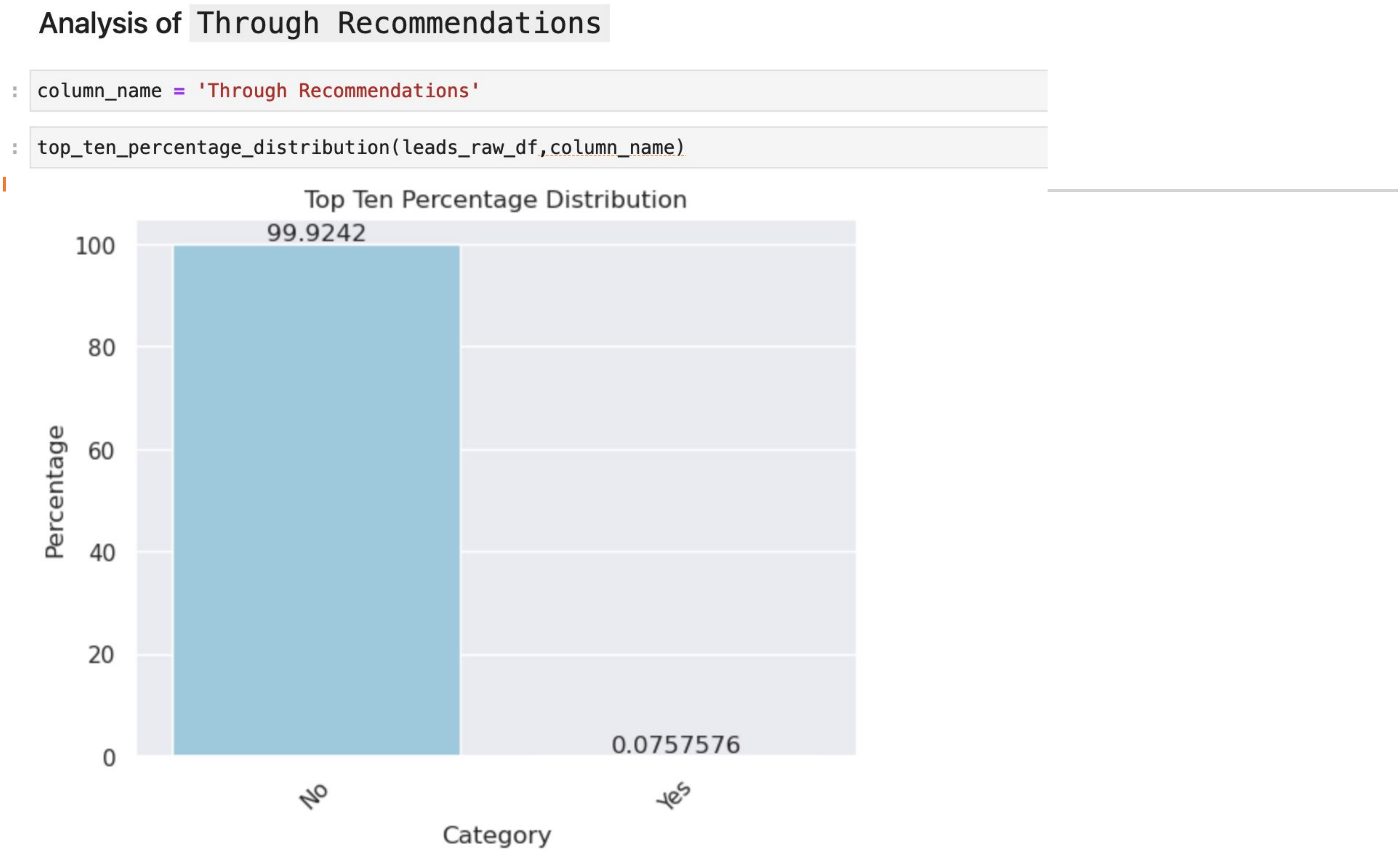
```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



Analysis of `Digital Advertisement`

```
column_name = 'Digital Advertisement'
```

```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



Analysis of `Through Recommendations`

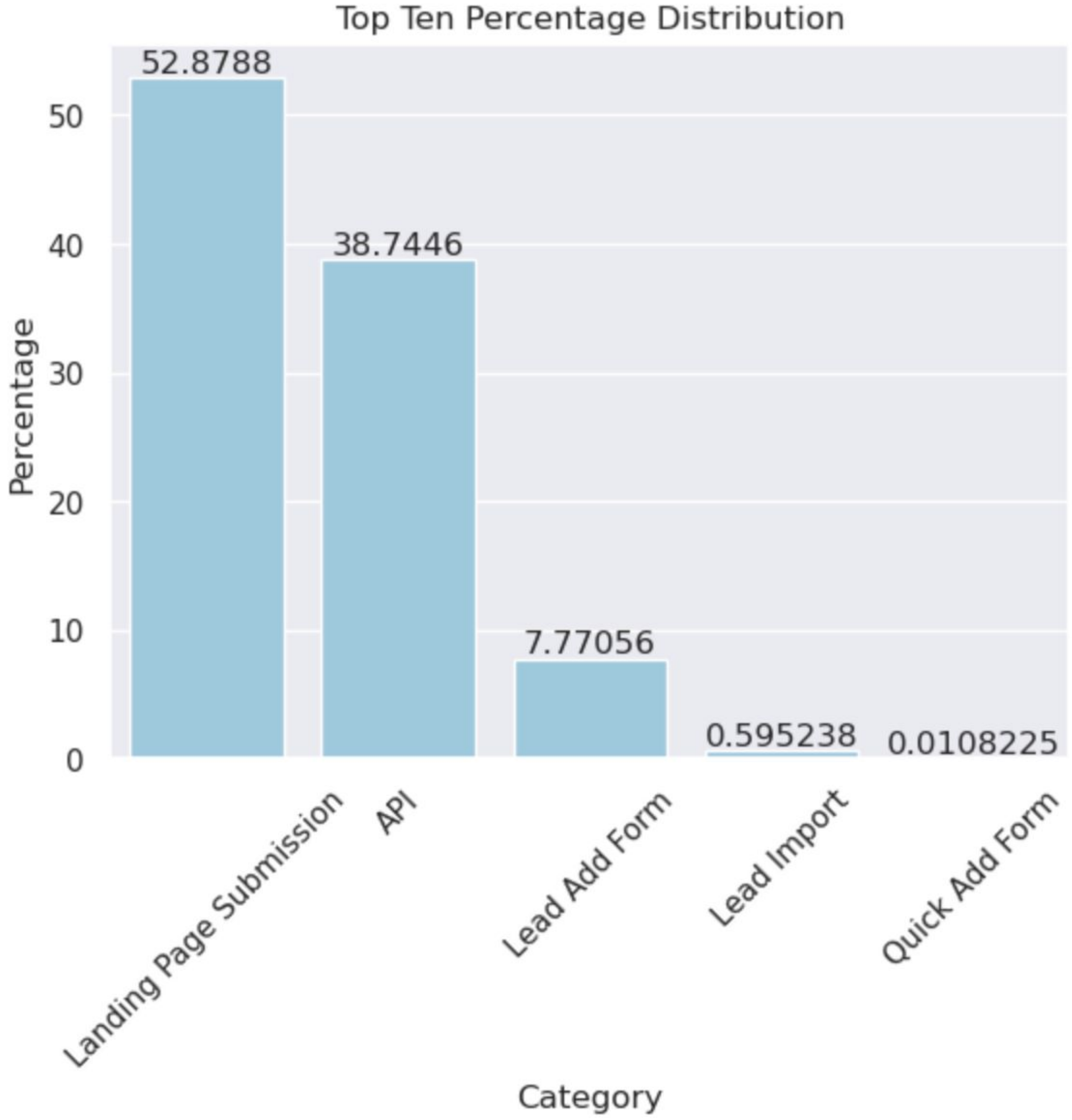
```
column_name = 'Through Recommendations'
```

```
top_ten_percentage_distribution(leads_raw_df,column_name)
```

Analysis of `Lead Origin`

```
column_name = 'Lead Origin'
```

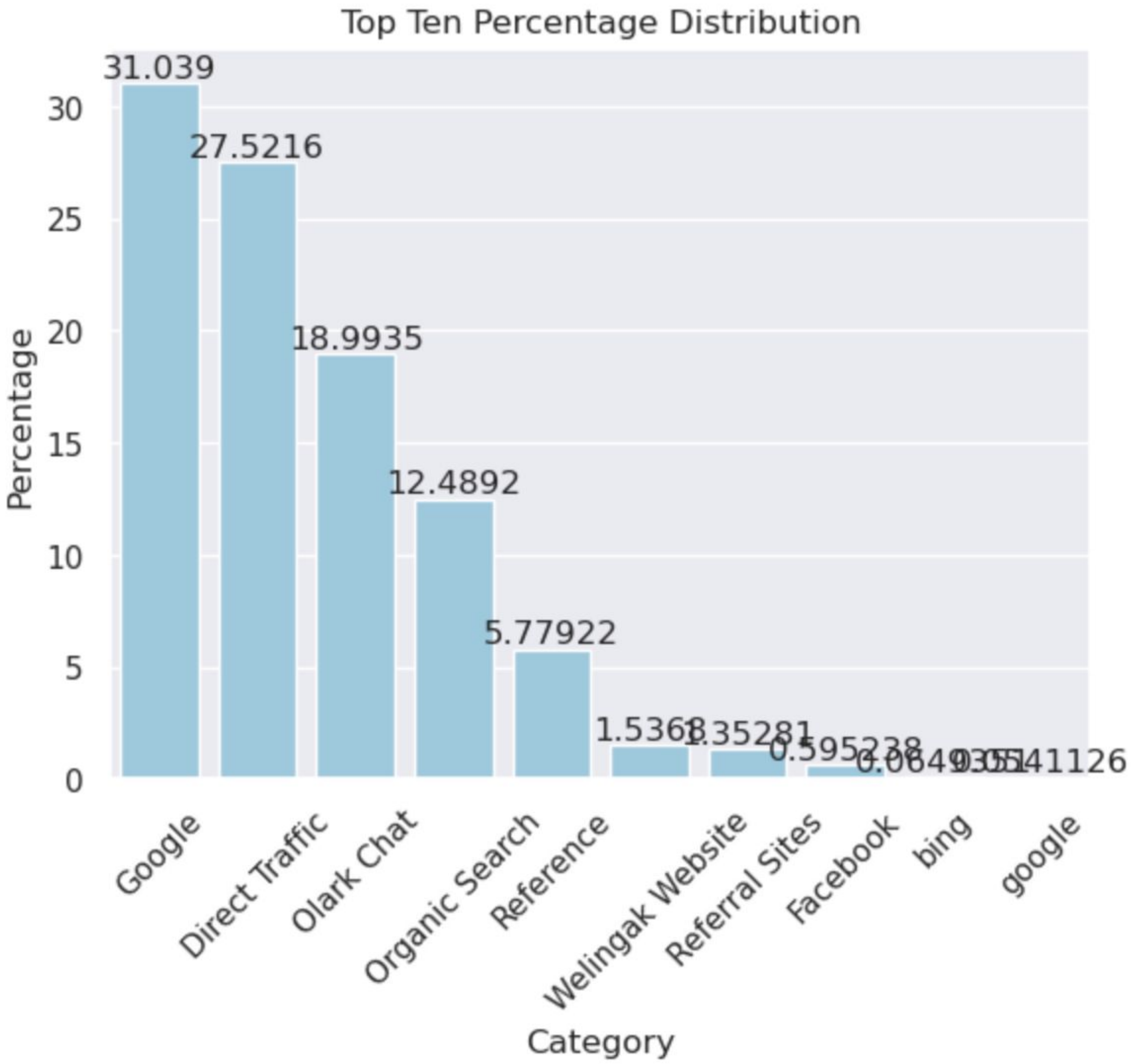
```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



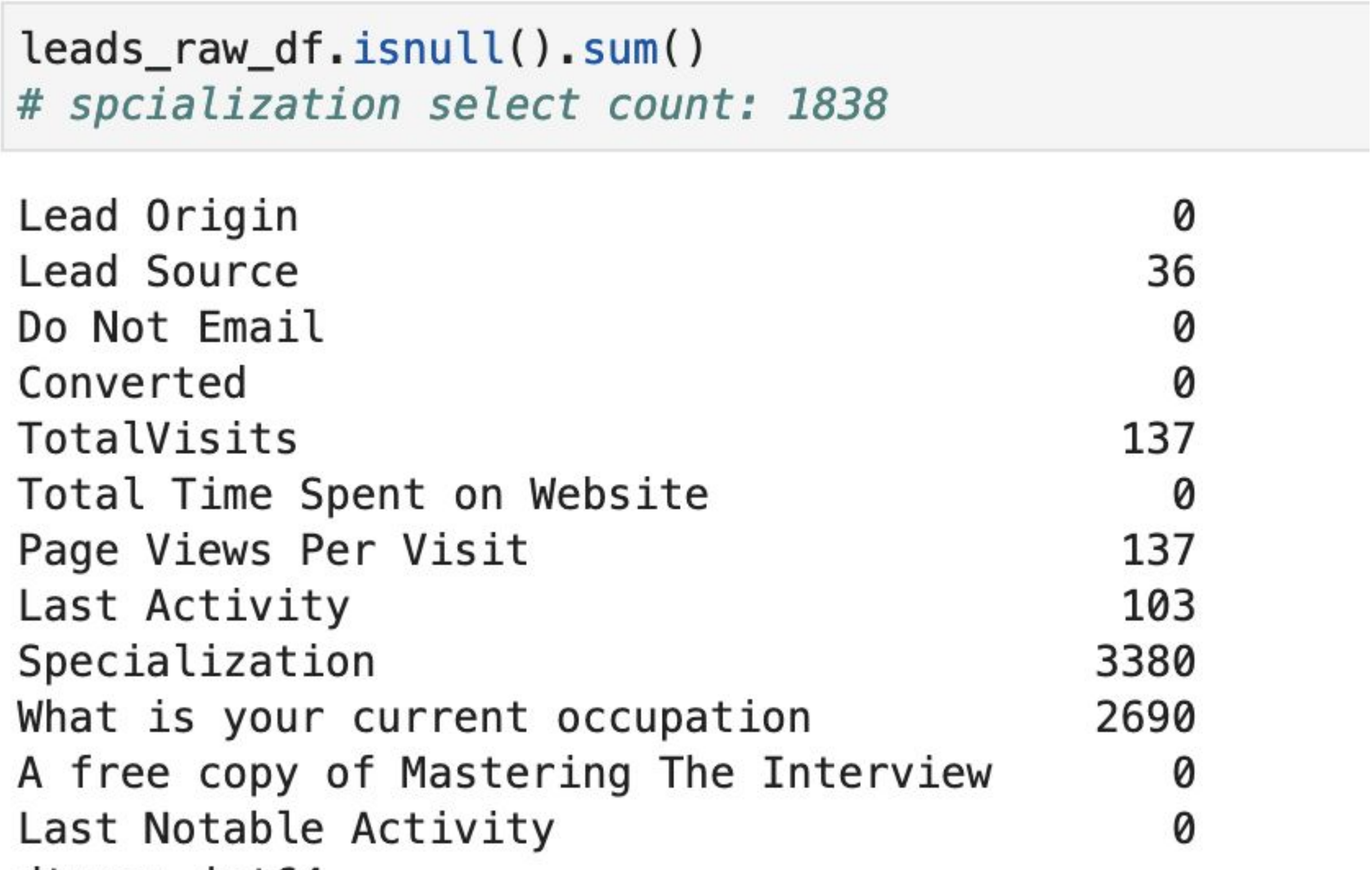
Analysis of `Lead Source`

```
column_name = 'Lead Source'
```

```
top_ten_percentage_distribution(leads_raw_df,column_name)
```



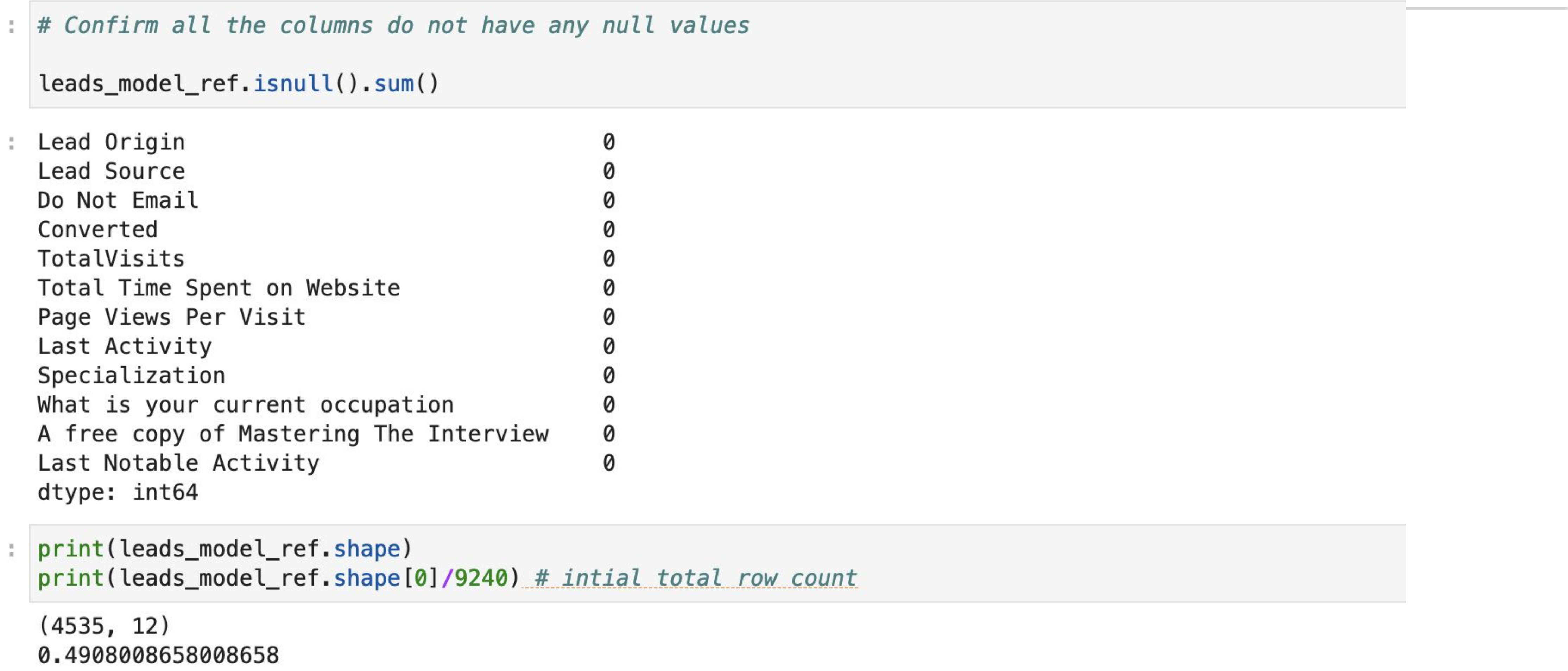
Analyzing Null Values (including `Select`)



```
leads_raw_df.isnull().sum()
# spcialization select count: 1838
```

```
Lead Origin                                 0
Lead Source                                36
Do Not Email                                0
Converted                                   0
TotalVisits                               137
Total Time Spent on Website                 0
Page Views Per Visit                      137
Last Activity                             103
Specialization                           3380
What is your current occupation          2690
A free copy of Mastering The Interview      0
Last Notable Activity                       0
```

Post removal of Null value columsn and Rows



```
# Confirm all the columns do not have any null values

leads_model_ref.isnull().sum()
```

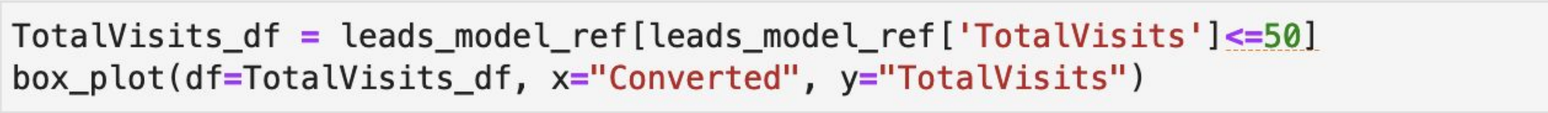
```
Lead Origin                               0
Lead Source                               0
Do Not Email                              0
Converted                                 0
TotalVisits                               0
Total Time Spent on Website               0
Page Views Per Visit                      0
Last Activity                             0
Specialization                            0
What is your current occupation           0
A free copy of Mastering The Interview    0
Last Notable Activity                     0
dtype: int64
```

```
print(leads_model_ref.shape)
print(leads_model_ref.shape[0]/9240) # intial total row count
```

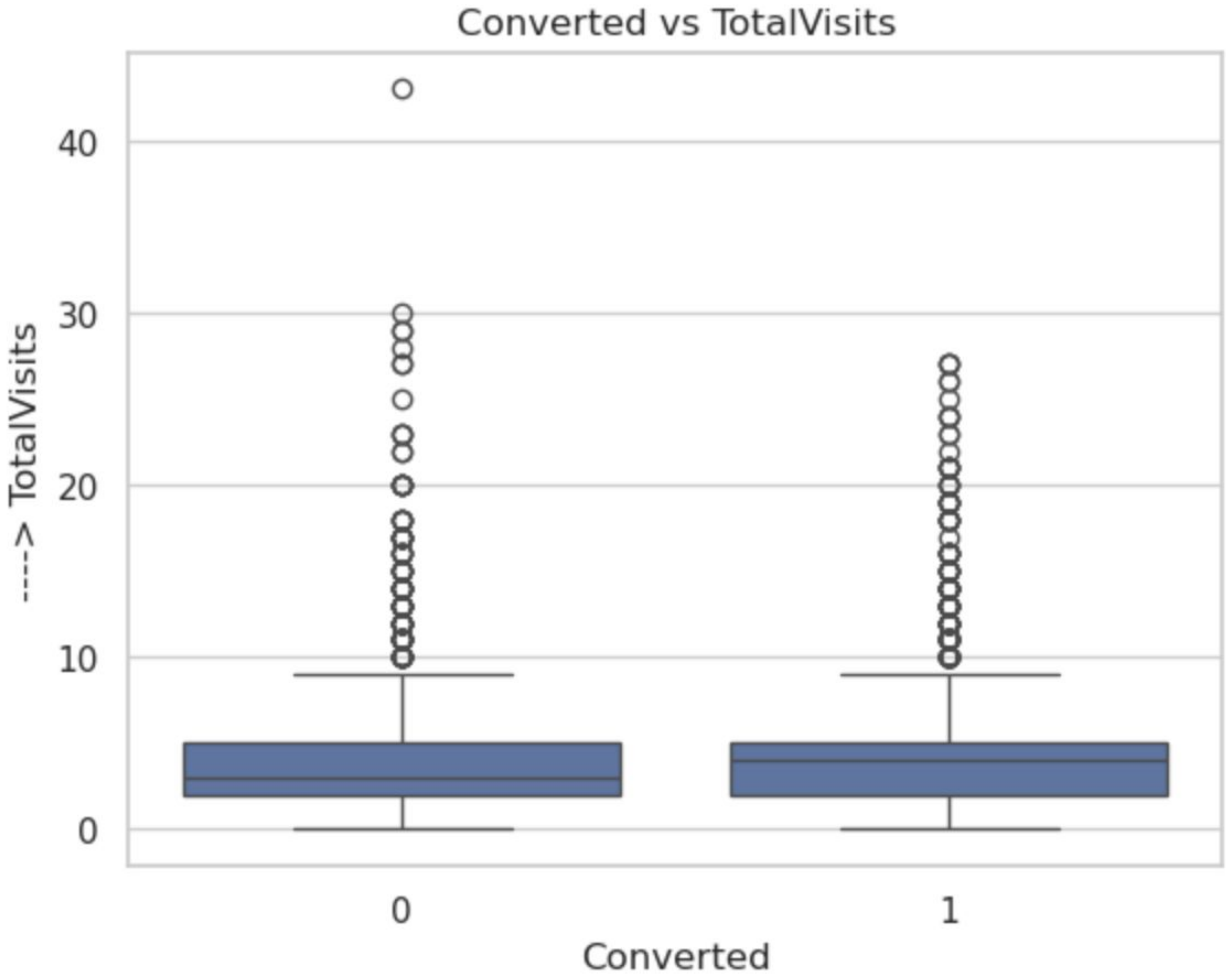
```
(4535, 12)
0.4908008658008658
```

- We still have around 50% of the rows , thought this value is not great, this data has the accurate data and cleaned up

TotalVisist analysis

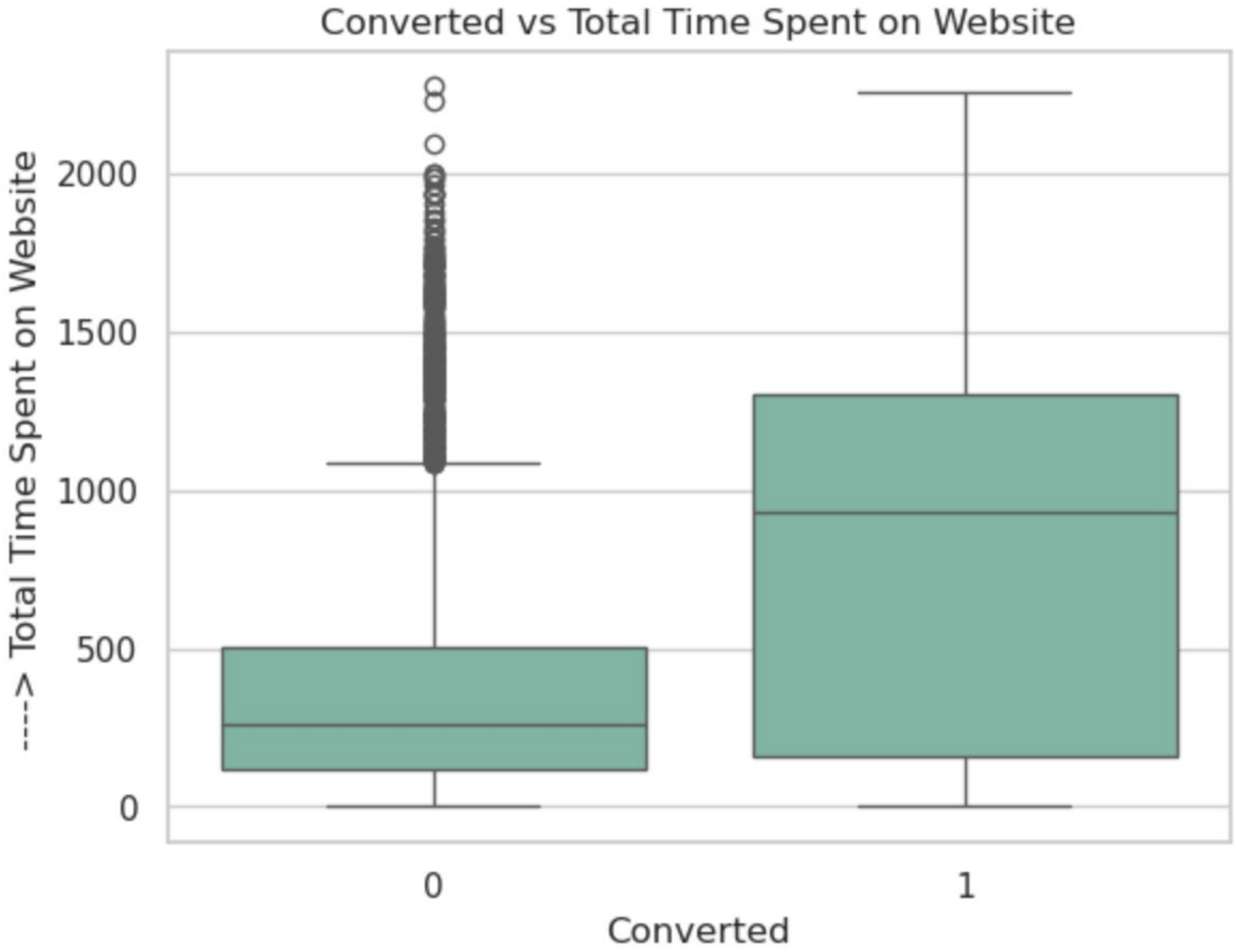


```
TotalVisits_df = leads_model_ref[leads_model_ref['TotalVisits']<=50]
box_plot(df=TotalVisits_df, x="Converted", y="TotalVisits")
```

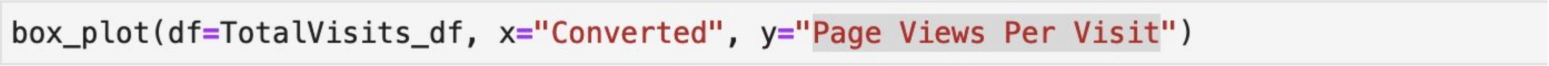


Total Time Spent on Website analysis

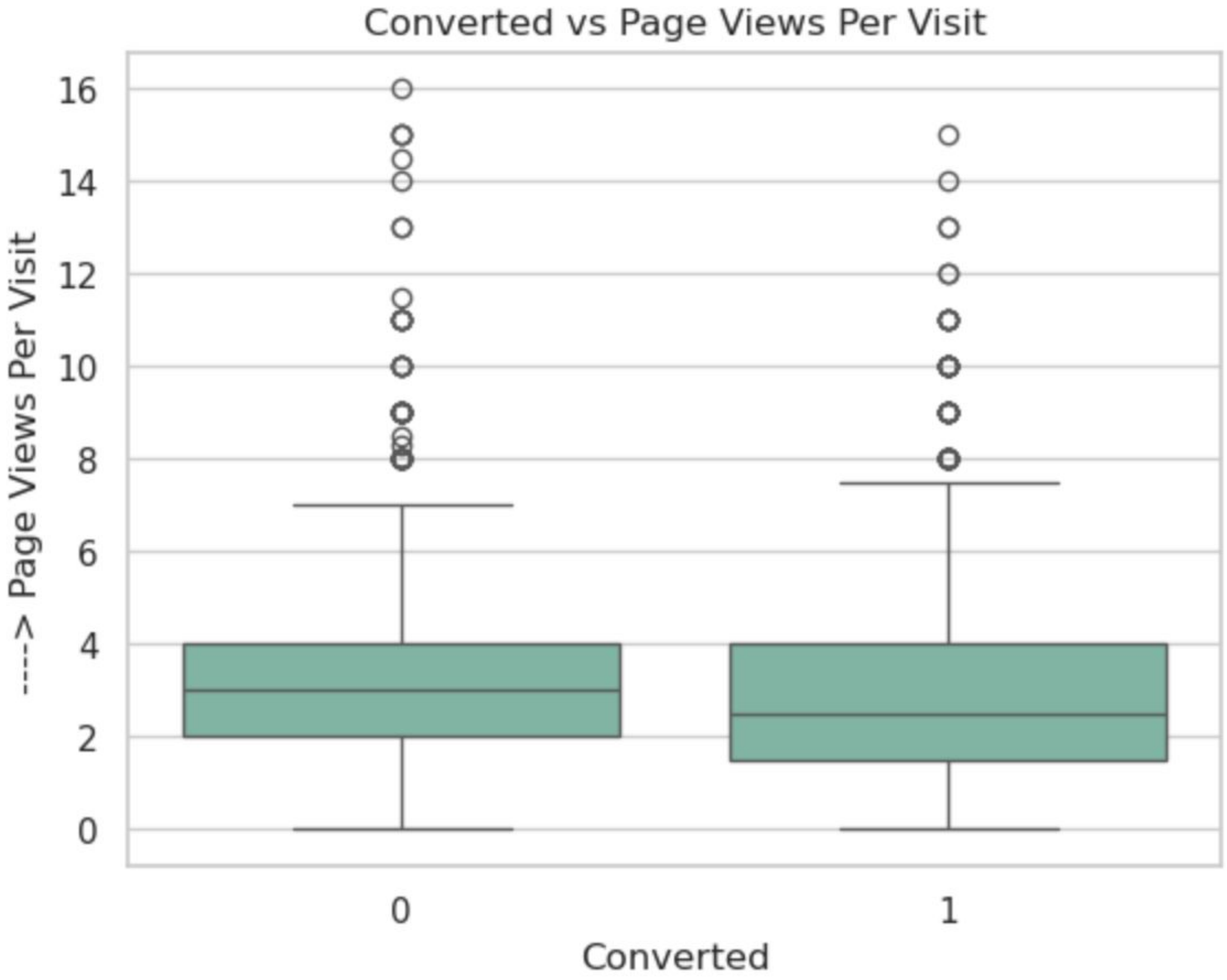
```
box_plot(df=TotalVisits_df, x="Converted", y="Total Time Spent on Website")
```



Page Views Per Visit analysis



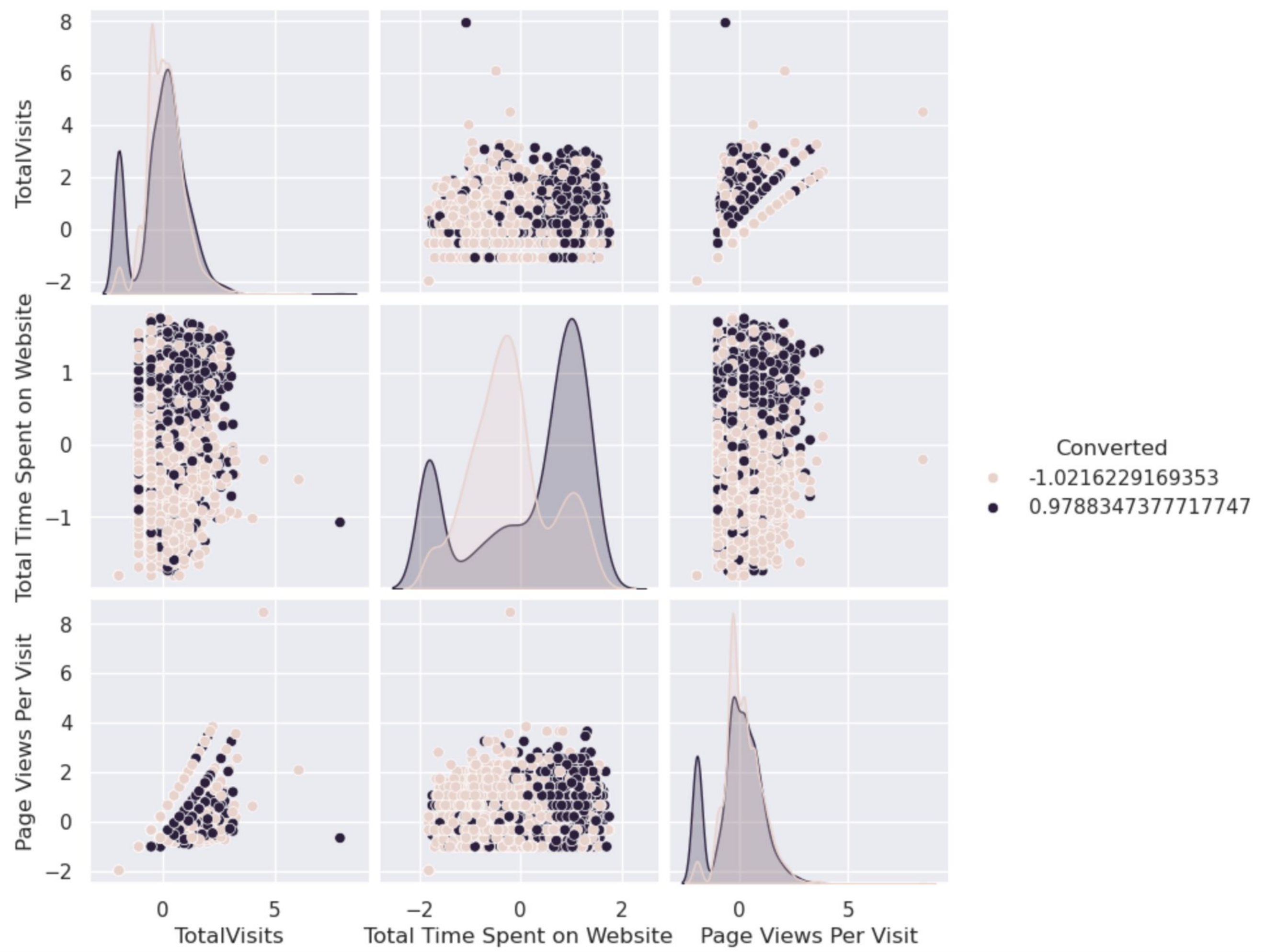
```
box_plot(df=TotalVisits_df, x="Converted", y="Page Views Per Visit")
```



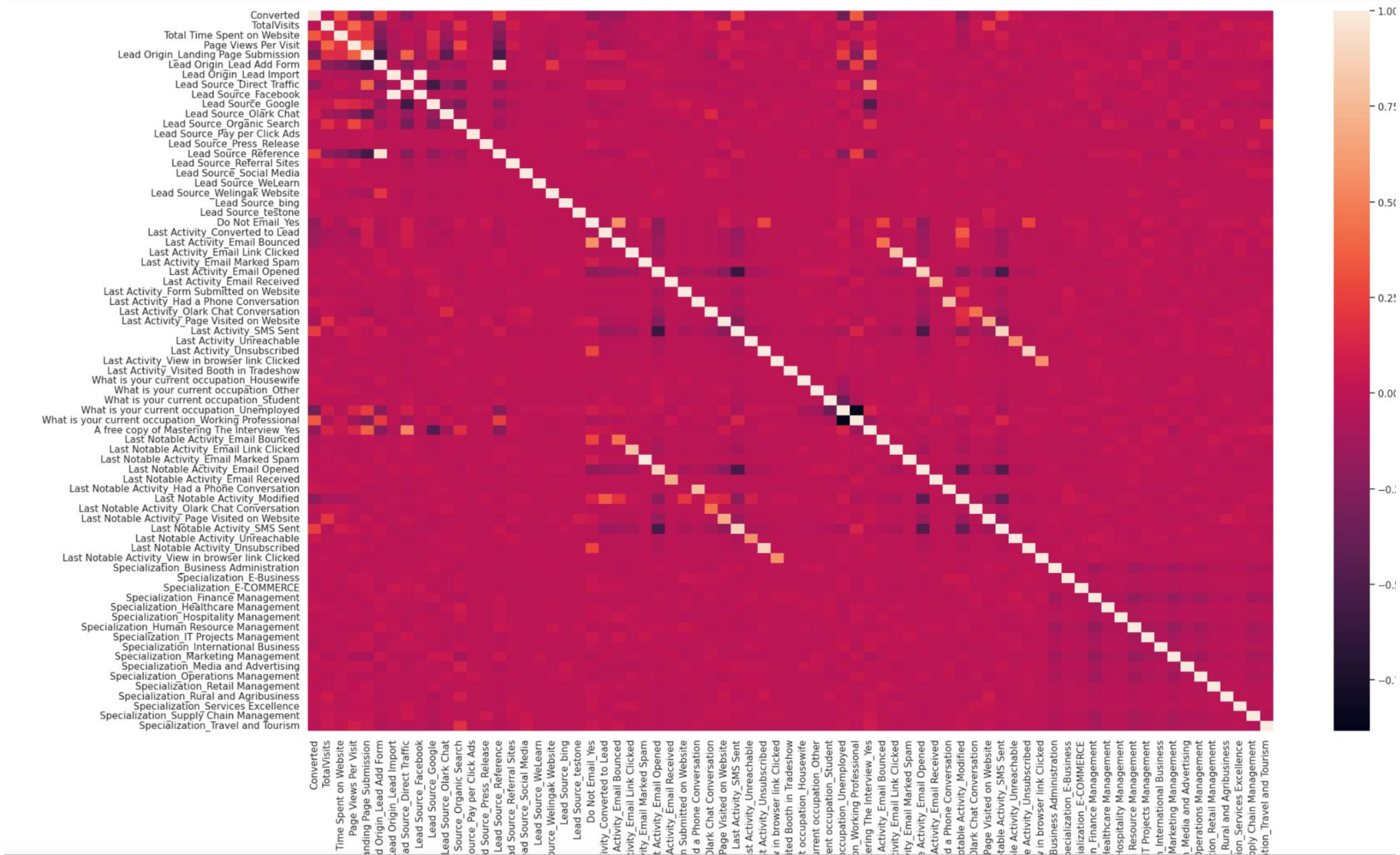
`Page Views Per Visit`

- The more the pages a user vistis per visit the more chance the user can convert to leads

Analyzing PowerTransformer Pair Plot



Analyzing Heatmap



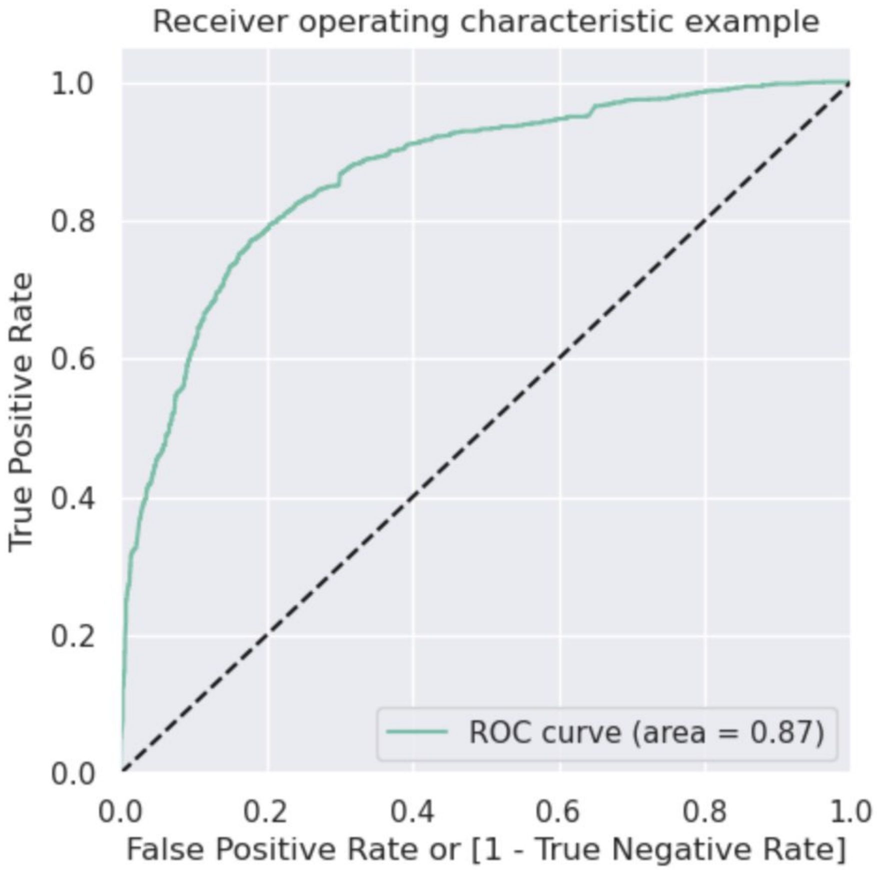
RFE Columns Removed

TotalVisits
Total Time Spent on Website
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Source_Direct Traffic
Lead Source_Organic Search
Lead Source_Reference
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Converted to Lead
Last Activity_Email Bounced
Last Activity_Had a Phone Conversation
Last Activity_Olark Chat Conversation
Last Activity_SMS Sent
What is your current occupation_Housewife
What is your current occupation_Unemployed
What is your current occupation_Working Professional
Last Notable Activity_Email Bounced
Last Notable Activity_Had a Phone Conversation
Last Notable Activity_Unreachable

Manual analysis Columns Removed in Model

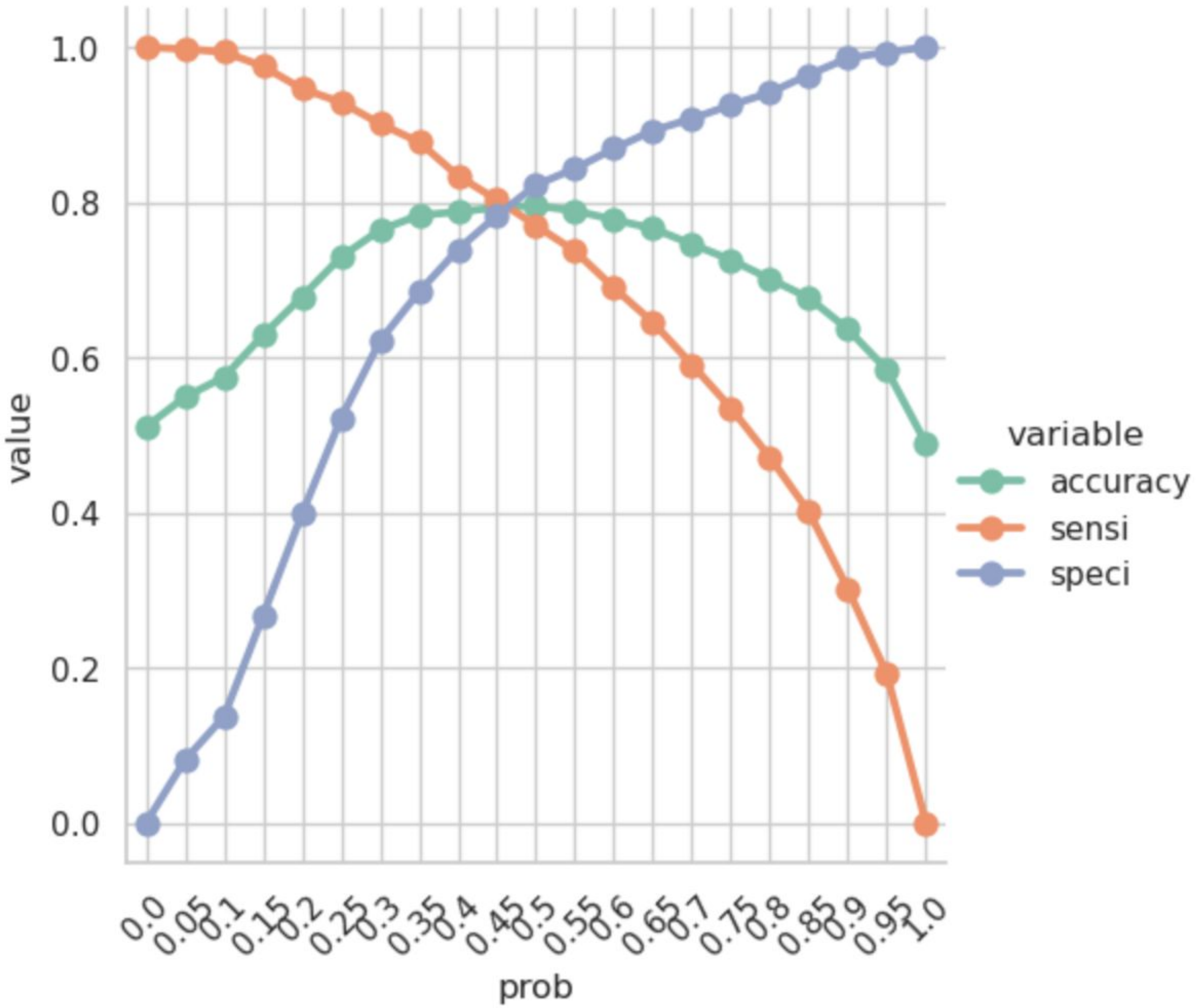
Lead Source_Reference
Lead Source_Welingak Website
What is your current occupation_Housewife
Last Notable Activity_Had a Phone Conversation
Lead Origin_Landing Page Submission

ROC Curve

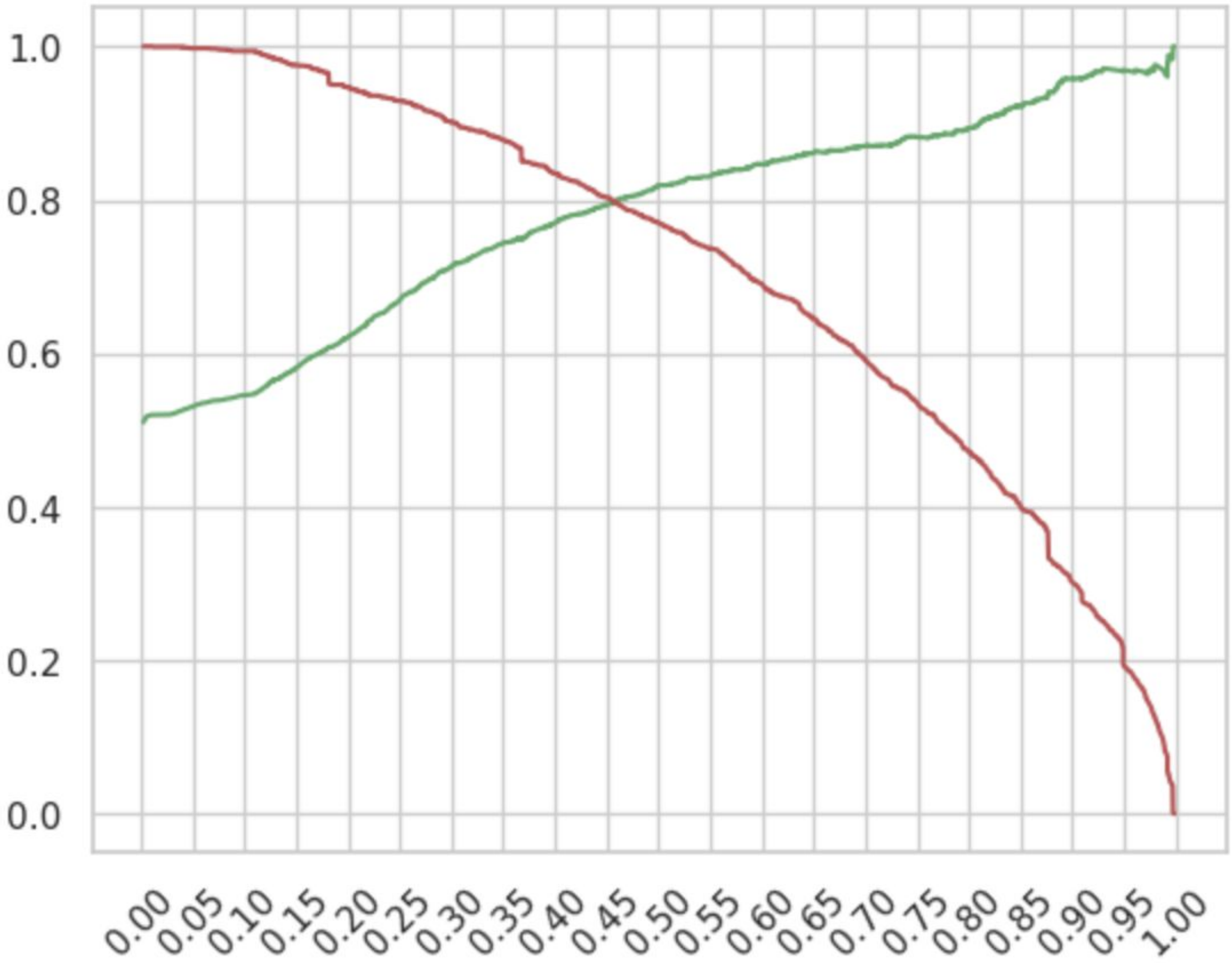


- Area under curve or RoC is `0.87` - very good optimal value

Predicted Probability vs Model Metrics



Predicted Probability vs Precision Recall Curve



Final Metrics

This recall cutoff point of 0.45 is optimal considering
* Accuracy: 78 (No major difference between Initial model and recall)
* Sensitivity: 79 (No major difference between Initial model and recall)
* Specificity: 77 (No major difference between Initial model and recall)

End Of Report
Thank you.