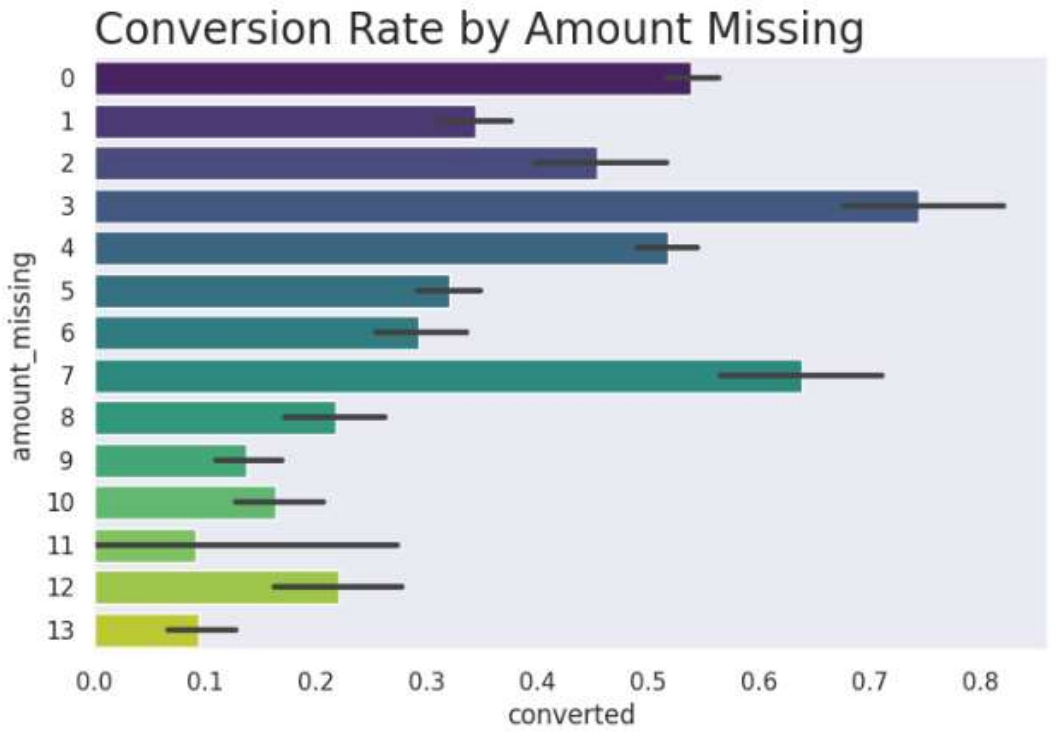
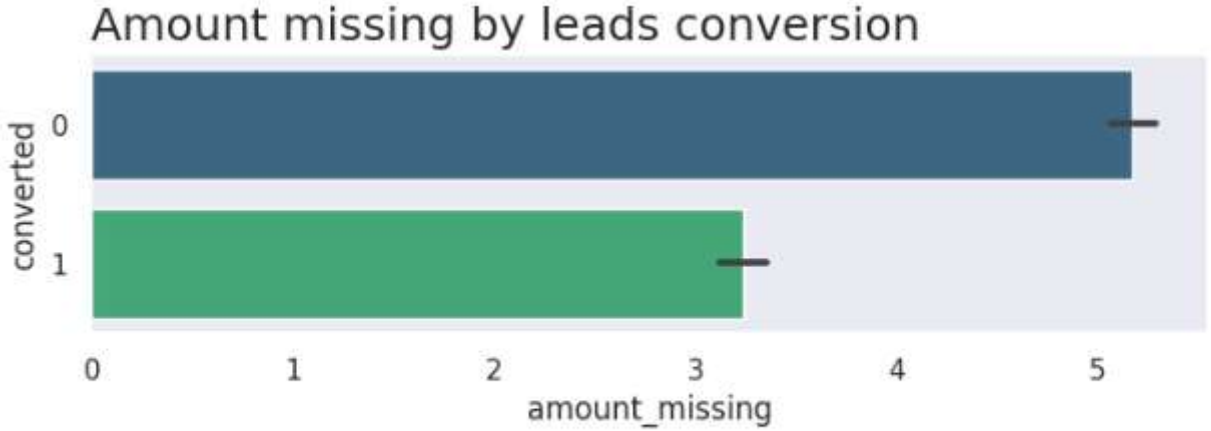
Explore Missing Values

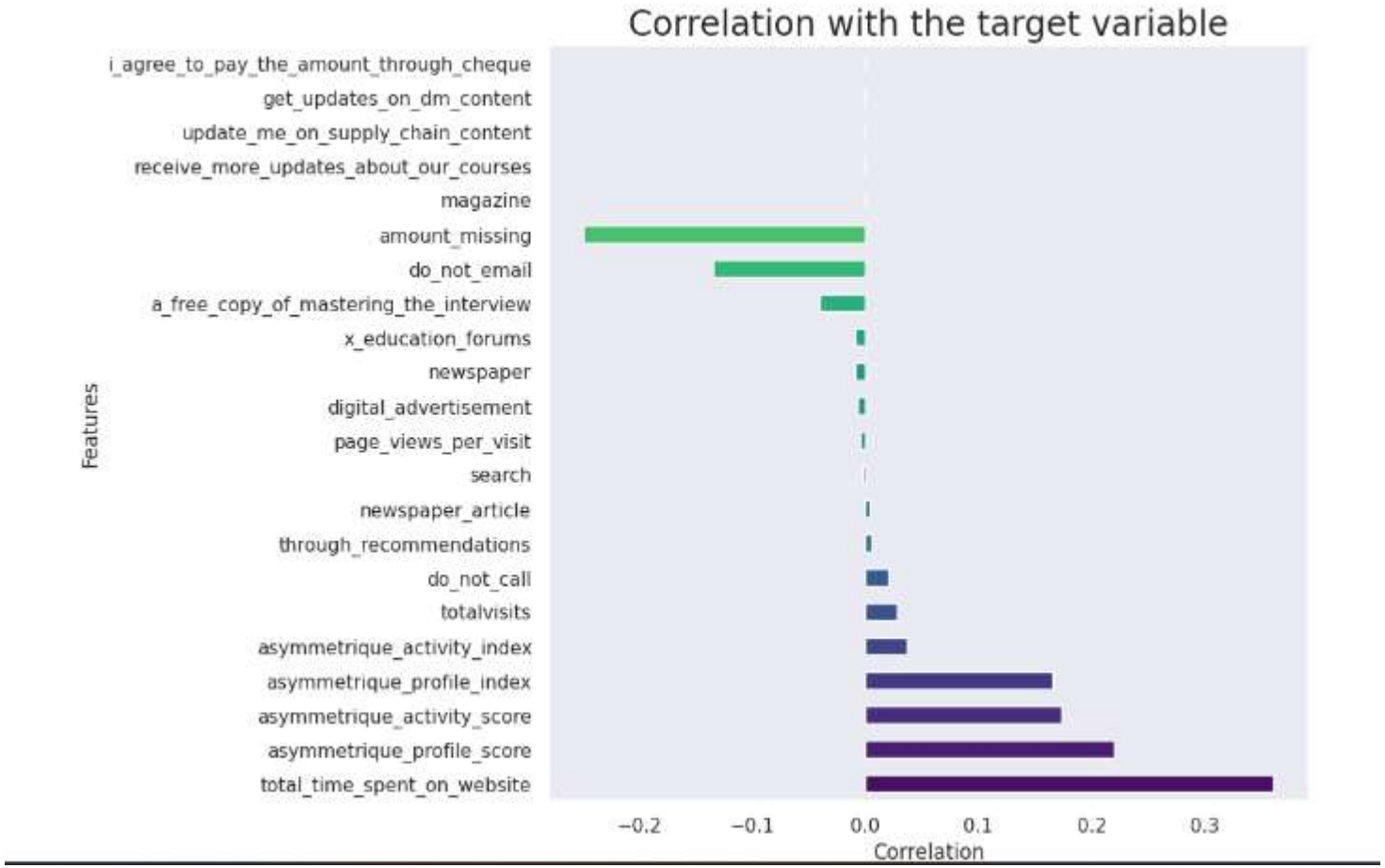
	proportion	amount
lead_quality	51.35	3796
asymmetrique_activity_index	45.48	3362
asymmetrique_profile_score	45.48	3362
asymmetrique_profile_index	45.48	3362
asymmetrique_activity_score	45.48	3362
tags	36.35	2687
lead_profile	29.40	2173
what_matters_most_to_you_in_choosing_a_course	29.40	2173
what_is_your_current_occupation	29.21	2159
country	26.50	1959
how_did_you_hear_about_x_education	23.92	1768
specialization	15.61	1154
city	15.41	1139
page_views_per_visit	1.45	107
totalvisits	1.45	107
last_activity	1.08	80

Missing values in certain columns, often requiring employee input, might stem from uncategorized leads. Streamlining lead management can improve data collection, inform decision-making, and optimize lead conversion strategies. Further investigation is necessary to confirm this hypothesis.

How much of the missing values belong to the same people?



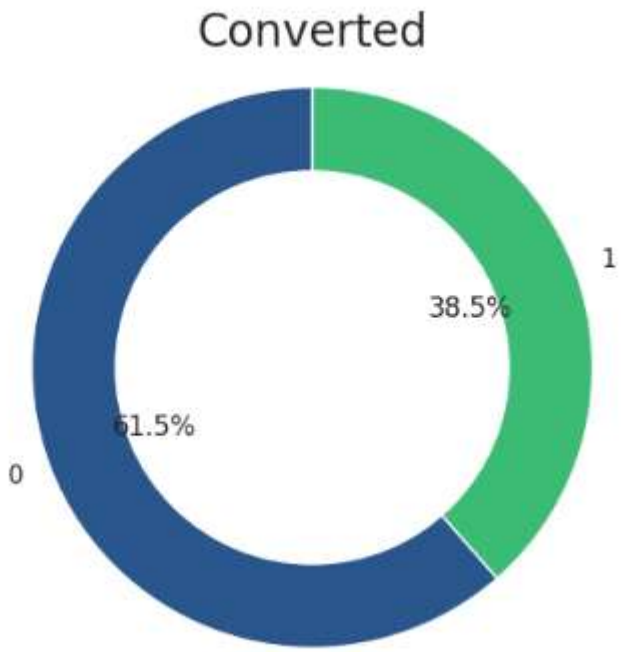




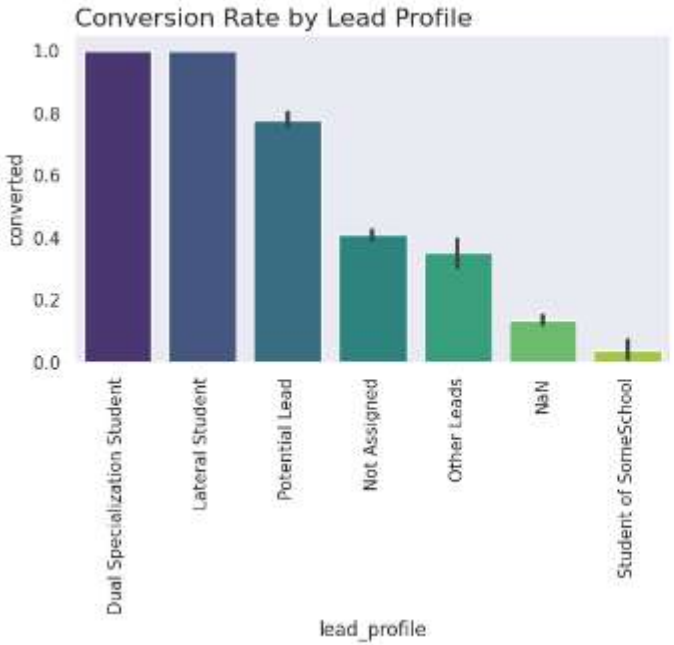
There's a negative correlation between missing lead data and the conversion rate. Higher instances of missing data might signify incomplete or poorly managed lead information, leading to potential difficulties in accurately categorizing and nurturing leads.

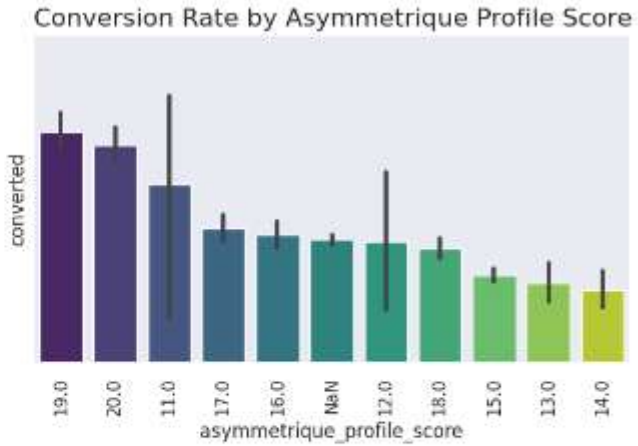
Notably, there's one column, "Total Time Spent On Website," which has a stronger correlation with the target variable than the number of missing values columns. This suggests that it might be a better potential predictor than most of the other features.

Exploratory Data Analysis

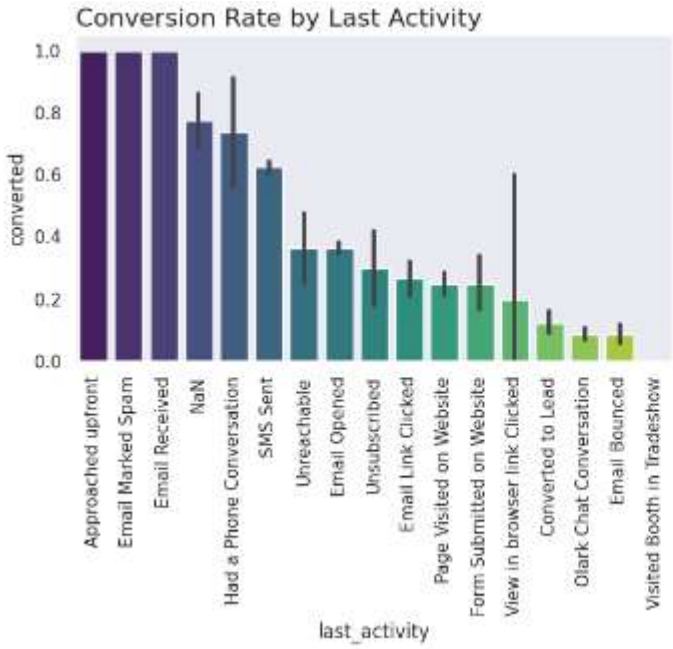


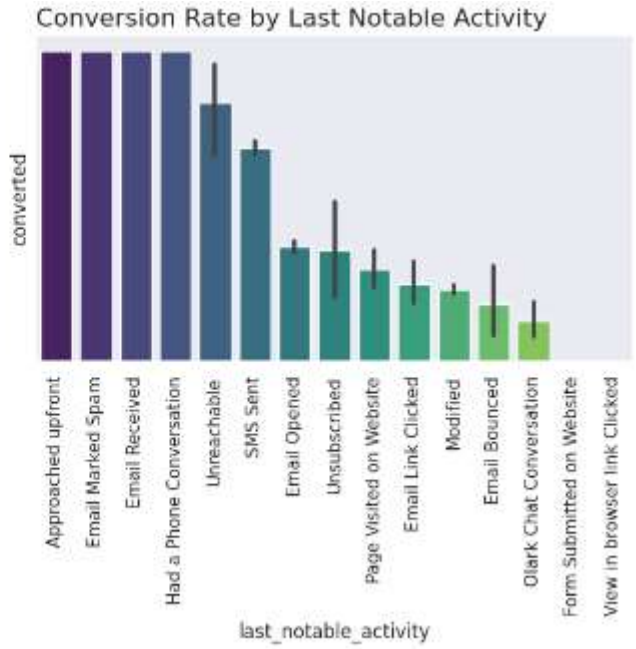
	asymmetrique_activity_index	asymmetrique_profile_index	asymmetrique_activity_score	asymmetrique_pr
asymmetrique_activity_index	1.000000	-0.145399	0.855985	-0.122669
asymmetrique_profile_index	-0.145399	1.000000	-0.145366	0.883177
asymmetrique_activity_score	0.855985	-0.145366	1.000000	-0.114636
asymmetrique_profile_score	-0.122669	0.883177	-0.114636	1.000000

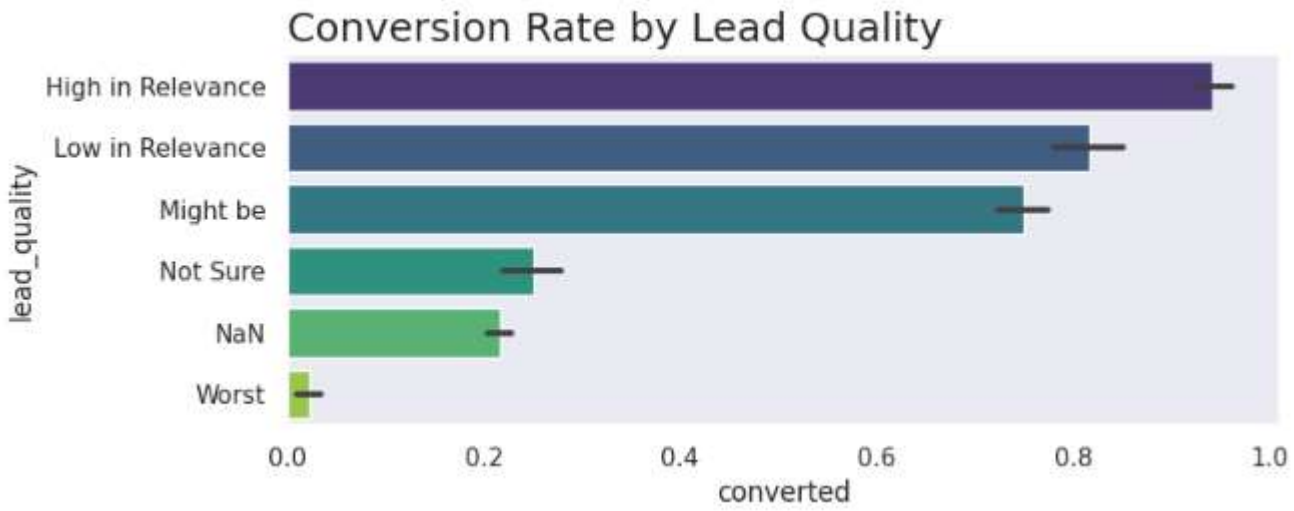


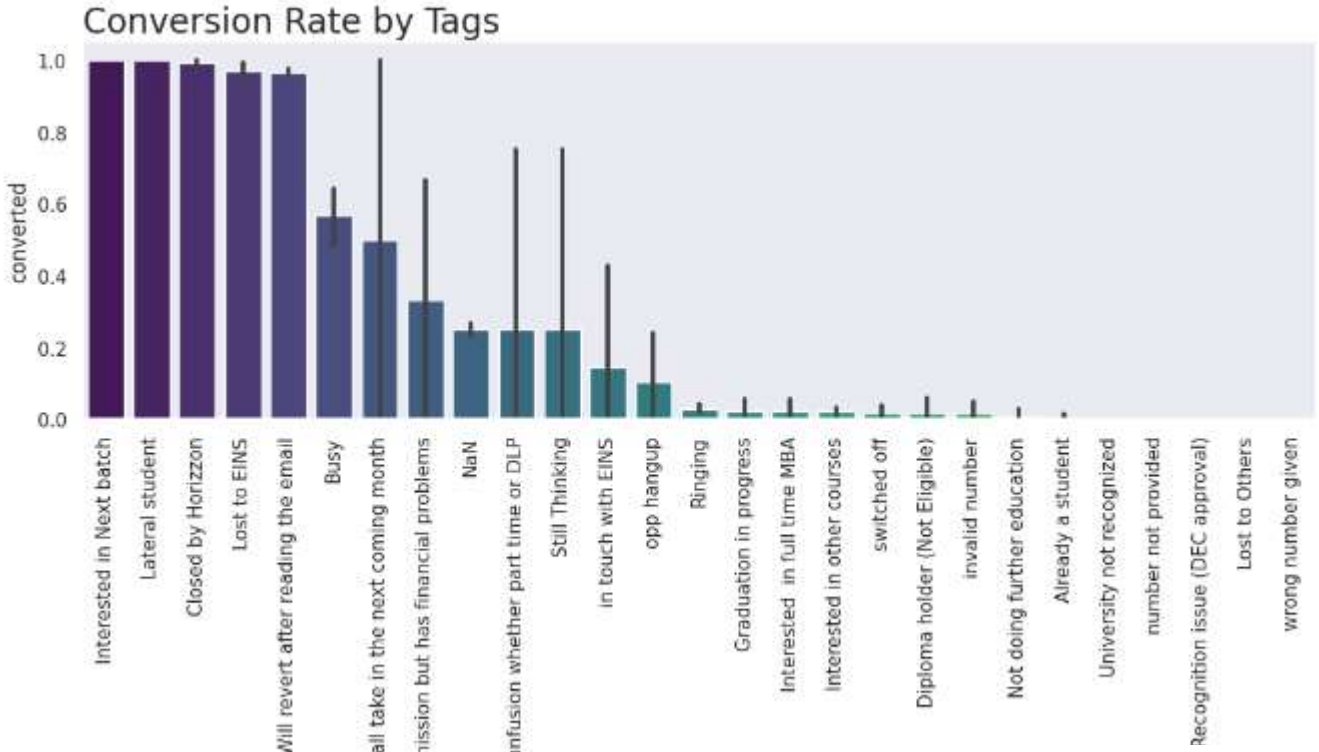


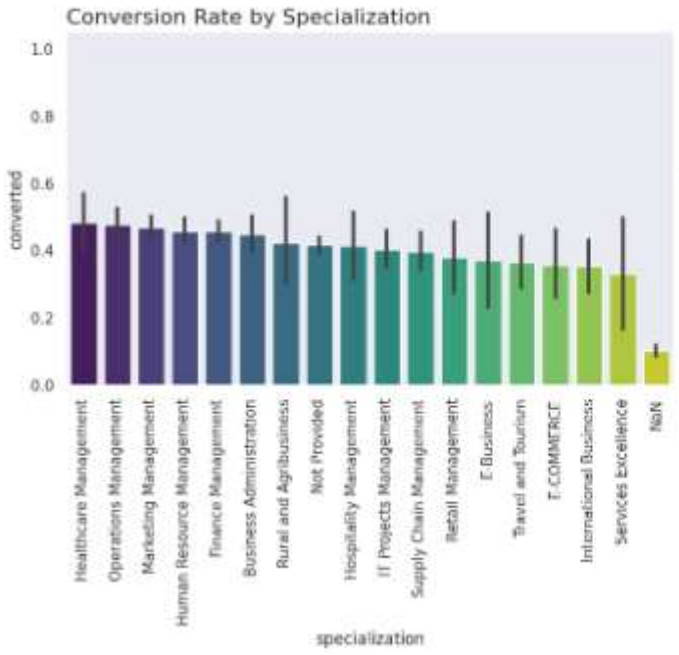
	totalvisits	total_time_spent_on_website	page_views_per_visit	asymmetrique_profile_score	asymmetric
totalvisits	1.000000	0.261952	0.598883	0.129016	-0.061397
total_time_spent_on_website	0.261952	1.000000	0.323684	0.167992	-0.066008
page_views_per_visit	0.598883	0.323684	1.000000	0.165945	-0.171264
asymmetrique_profile_score	0.129016	0.167992	0.165945	1.000000	-0.114636
asymmetrique_activity_score	-0.061397	-0.066008	-0.171264	-0.114636	1.000000

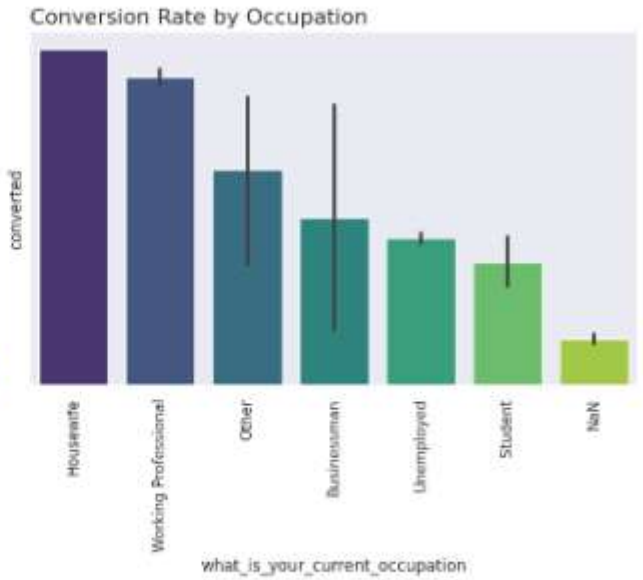

Conversion Rate by Last Activity
converted
1.0
0.8
0.6
0.4
0.2
0.0
Approached upfront
Email Marked Spam
Email Received
NaN
Had a Phone Conversation
SMS Sent
Unreachable
Email Opened
Unsubscribed
Email Link Clicked
Page Visited on Website
Form Submitted on Website
View in browser link Clicked
Converted to Lead
Olark Chat Conversation
Email Bounced
Visited Booth in Tradeshow
last_activity


Conversion Rate by Last Notable Activity
converted
Approached upfront
Email Marked Spam
Email Received
Had a Phone Conversation
Unreachable
SMS Sent
Email Opened
Unsubscribed
Page Visited on Website
Email Link Clicked
Modified
Email Bounced
Olark Chat Conversation
Form Submitted on Website
View in browser link Clicked
last_notable_activity

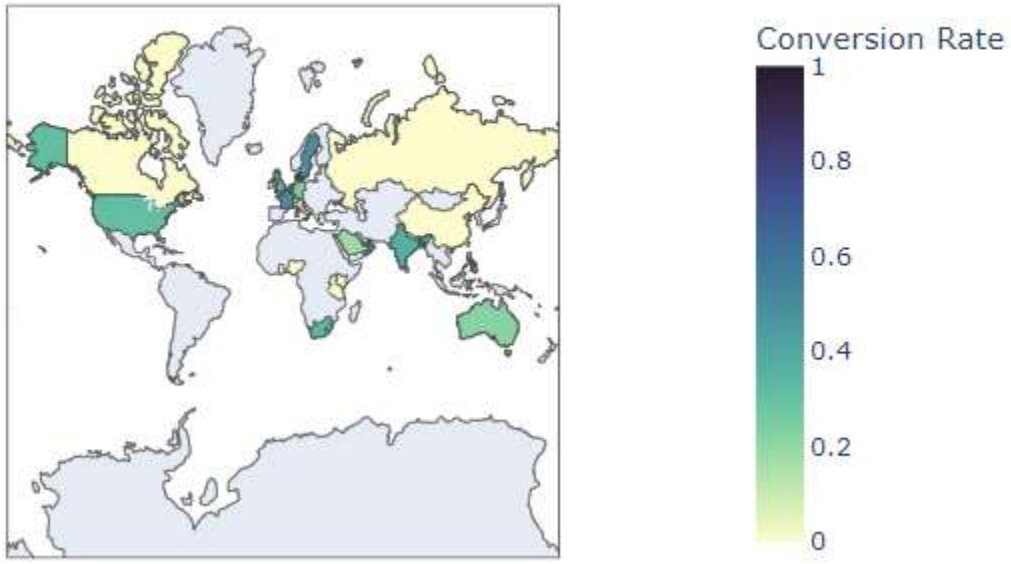

Conversion Rate by Lead Quality
lead_quality
High in Relevance
Low in Relevance
Might be
Not Sure
NaN
Worst
0.0
0.2
0.4
0.6
0.8
1.0
converted

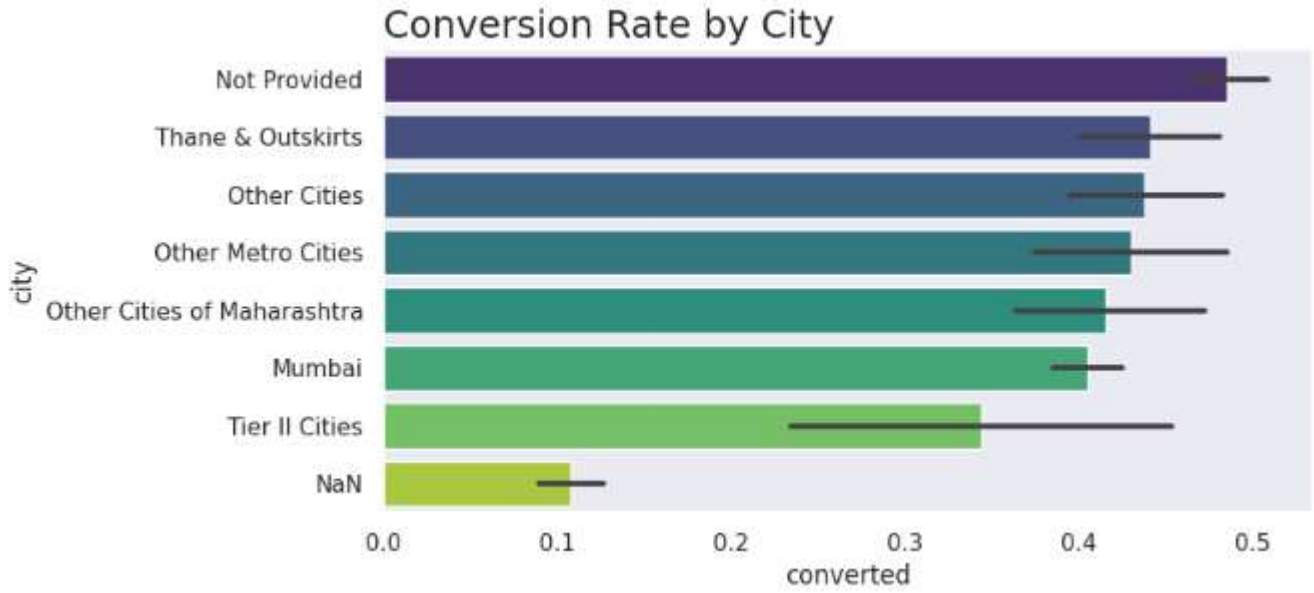

Conversion Rate by Tags
converted
1.0
0.8
0.6
0.4
0.2
0.0
Interested in Next batch
Lateral student
Closed by Horizzon
Lost to EINS
Will revert after reading the email
Busy
Shall take in the next coming month
ake admission but has financial problems
NaN
In confusion whether part time or DLP
Still Thinking
In touch with EINS
opp hangup
Ringing
Graduation in progress
Interested in full time MBA
Interested in other courses
switched off
Diploma holder (Not Eligible)
invalid number
Not doing further education
Already a student
University not recognized
number not provided
Recognition issue (DEC approval)
Lost to Others
wrong number given


Conversion Rate by Specialization
converted
1.0
0.8
0.6
0.4
0.2
0.0
Healthcare Management
Operations Management
Marketing Management
Human Resource Management
Finance Management
Business Administration
Rural and Agribusiness
Not Provided
Hospitality Management
IT Projects Management
Supply Chain Management
Retail Management
E-Business
Travel and Tourism
E-COMMERCE
International Business
Services Excellence
NaN
specialization


Conversion Rate by Occupation
converted
Housewife
Working Professional
Other
Businessman
Unemployed
Student
NaN
what_is_your_current_occupation

Conversion Rate by Country





Cities where country isn't India:

Out[46]:

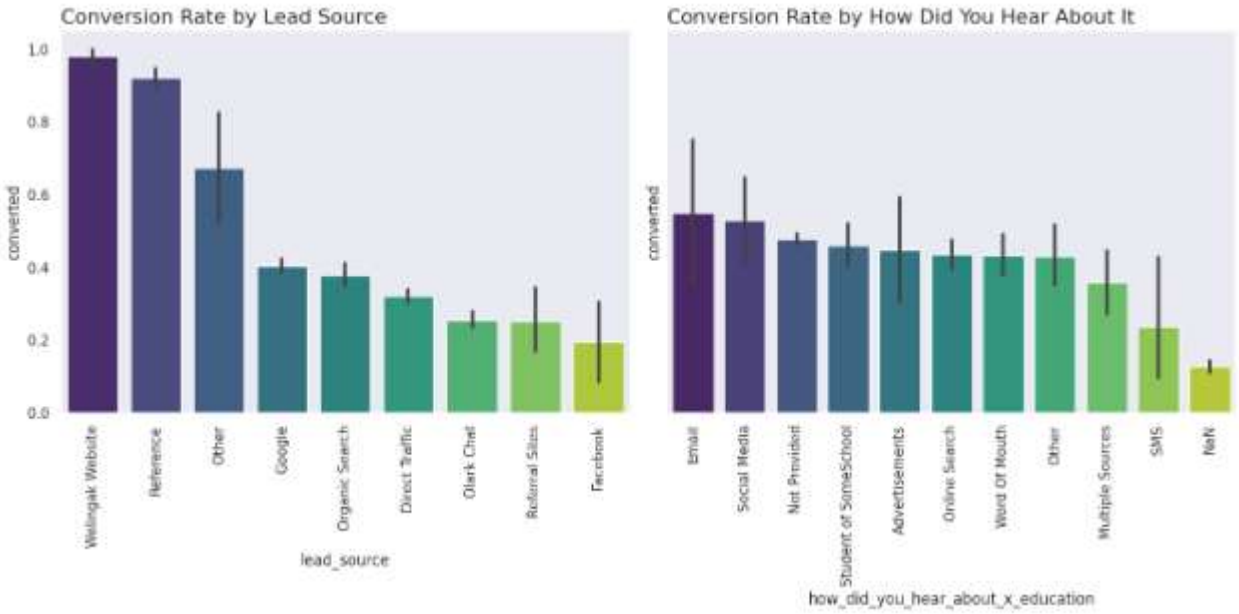
city	
Not Provided	992
NaN	693
Mumbai	244
Other Cities	98
Thane & Outskirts	83
Other Cities of Maharashtra	49
Other Metro Cities	27
Tier II Cities	5

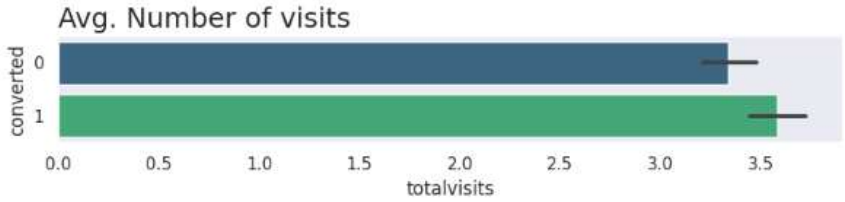
Name: count, dtype: int64

Countries where City es equal to an Indian city:

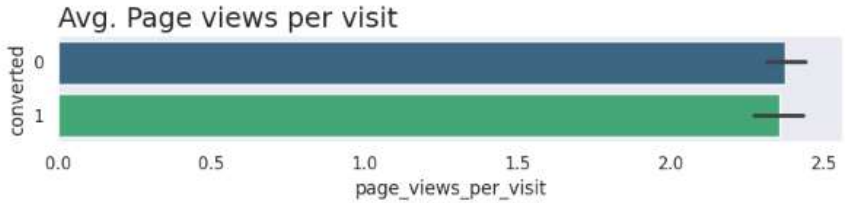
country	
India	3220
NaN	270
United States	32
United Arab Emirates	19
Singapore	11
United Kingdom	9
Saudi Arabia	8
Australia	6
Qatar	5
Bahrain	4
Germany	3
Belgium	2
Canada	2
Netherlands	2
Kuwait	1
France	1
Sweden	1
Malaysia	1
Hong Kong	1
Switzerland	1
Oman	1
China	1

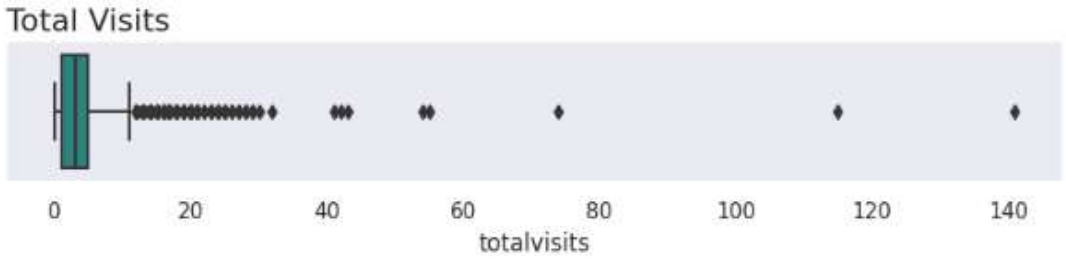
Name: count, dtype: int64

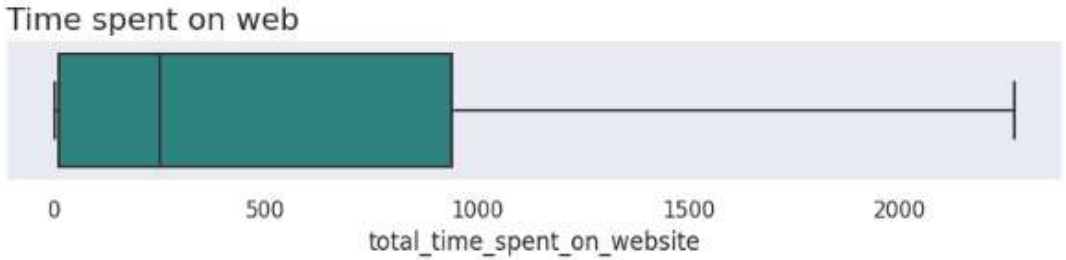


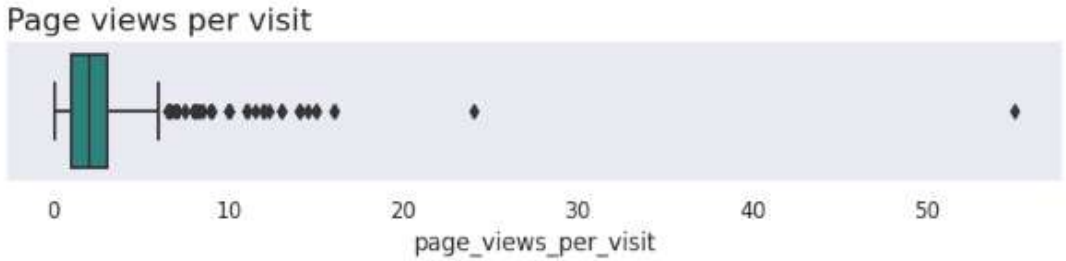

Avg. Number of visits
converted
0
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
totalvisits


Avg. Time spent on website
converted
0
1
0
100
200
300
400
500
600
700
total_time_spent_on_website


Avg. Page views per visit
converted
0
1
0.0
0.5
1.0
1.5
2.0
2.5
page_views_per_visit


Total Visits
0
20
40
60
80
100
120
140
totalvisits


Time spent on web
0
500
1000
1500
2000
total_time_spent_on_website


Page views per visit
0
10
20
30
40
50
page_views_per_visit

---------- Logistic Regression ----------

---------- Cross validation scores:
Scores: [0.91741472 0.91592129 0.91785714 0.91896705 0.93191866]
Mean: 0.920415772237881
Standard deviation: 0.005833839247863671

---------- Scores in the training set:
Precision: 0.9382491827097712
Recall: 0.9066339066339066
F1 score: 0.9221706533380934
ROC - AUC score: 0.9346068498610871

---------- Support Vector Machine ----------

---------- Cross validation scores:
Scores: [0.91838565 0.92184725 0.9341637 0.91954023 0.93848858]
Mean: 0.9264850809036904
Standard deviation: 0.008226619600733434

---------- Scores in the training set:
Precision: 0.9437074220150592
Recall: 0.9238329238329238
F1 score: 0.9336644200070947
ROC - AUC score: 0.9446371310778091

---------- Random Forest ----------

---------- Cross validation scores:
Scores: [0.91921005 0.91974752 0.93027361 0.9215859 0.93960924]
Mean: 0.9260852646703863
Standard deviation: 0.00785009330810528

---------- Scores in the training set:
Precision: 0.9908995449772489
Recall: 0.9936819936819937
F1 score: 0.9922888187872415
ROC - AUC score: 0.9939794516065702
Oob score: 0.9422348484848485

---------- Decission Tree ----------

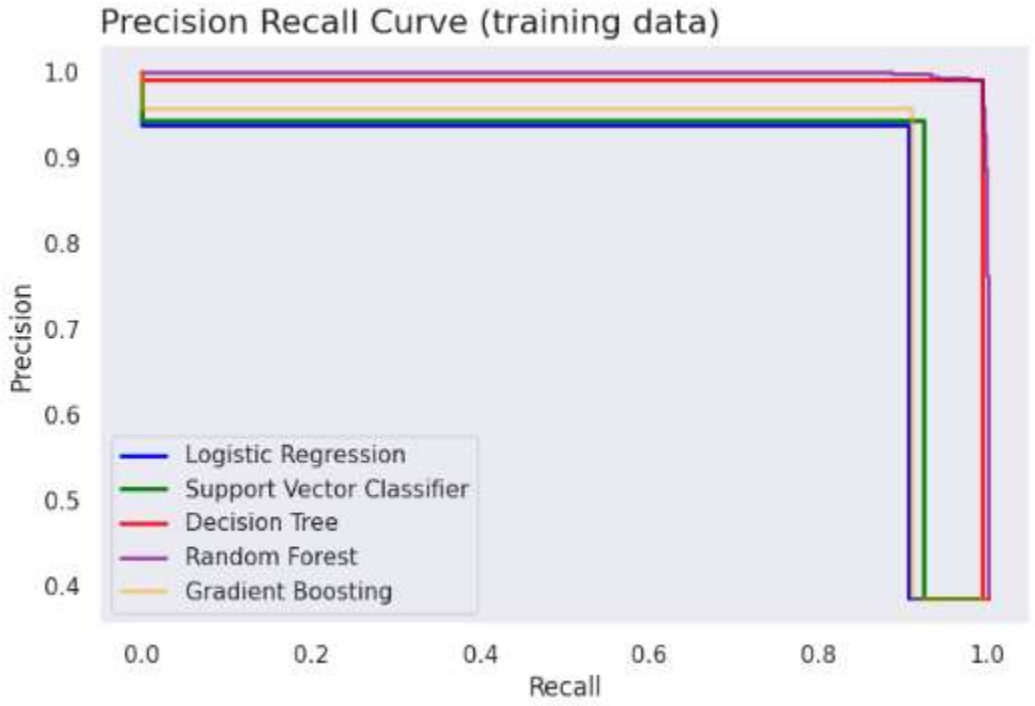
---------- Cross validation scores:
Scores: [0.89492119 0.89612676 0.89806678 0.89137931 0.89837746]
Mean: 0.8957743001598371
Standard deviation: 0.0025377098574763494

---------- Scores in the training set:
Precision: 0.9912434325744308
Recall: 0.9933309933309933
F1 score: 0.9922861150070126
ROC - AUC score: 0.9939140108631633

---------- Gradient Boosting ----------

---------- Cross validation scores:
Scores: [0.91862568 0.92196007 0.9309417 0.92197309 0.94201606]
Mean: 0.927103321202558
Standard deviation: 0.00850606304189926

---------- Scores in the training set:
Precision: 0.9571322985957132
Recall: 0.9090909090909091
F1 score: 0.9324932493249325
ROC - AUC score: 0.9417785604226282



Precision: 0.9402332361516035

Recall: 0.9058988764044944

F1 score: 0.9227467811158798

ROC - AUC score: 0.9349036635543598