Title

- **Prediction**
- Image Preprocessing
- Text Preprocessing
- Recommendation

Predicton

ABOUT

EDA Prediction

Exploratory Data Analysis (EDA)

EDA, or Exploratory Data Analysis, is a critical phase in the data analysis process that involves visually and statistically exploring and summarizing key c
trends within a dataset. The primary objectives of EDA are to uncover insights, identify relationships, and gain an understanding of the structure and di
process is crucial for informing subsequent steps in the data analysis pipeline, such as feature engineering, modeling, and hypothesis testing.

Dataset Information

Shape of Dataset

No.of.Rows: 100000

No.of.Columns: 46

In this 100000 rows there are 90793 duplicates, for model building we can delete all duplicates from the dataset. So that classification will perform better

Aggregate Information

From this table we can take the mean, min, max, etc... for a

After Removing all Duplicates:

No.of.Rows: 9207

No.of.Columns: 46

	count_session	count_hit	totals_newVisits	geoNe
count	9,207	9,207	9,207	
mean	19.908	2,489.7939	0.0124	
std	21.059	3,399.7782	0.1106	
min	1	2	0	
25%	7	525	0	
50%	14	1,347	0	
75%	26	3,149	0	
max	270	48,744	1	

Data Preprocessing

Null Percentage

From this table there is no empty values. So we don't need to make any changes

Column_Name	Null_Percentage
count_session	0
latest_medium	0
time_latest_visit	0
avg_visit_time	0
days_since_last_visit	0
days_since_first_visit	0
visits_per_day	0
bounce_rate	0
earliest_source	0
latest_source	0

Sparsity Data

In this sparsity table the columns having more than 50 percent of zero values we can delete that columns

Column_name	Zero_Percentage
youtube	100
days_since_last_visit	100
bounces	99.84
totals_newVisits	98.76
bounce_rate	95.83
has_converted	58.37
avg_session_time_page	58.26
historic_session_page	58.26
transactionRevenue	48.13
latest_isTrueDirect	14.96

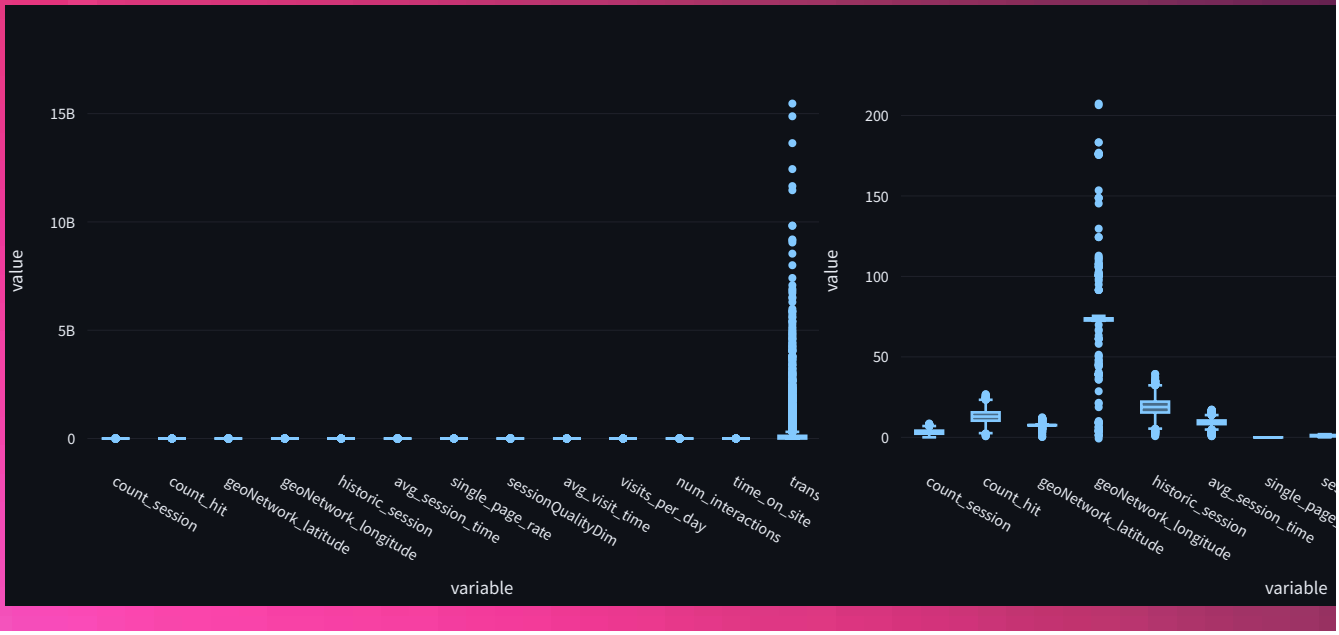
Note: The column has_conv
we should not make any cha
numeric values.

Column_name	Zero
has_converted	
transactionRevenue	
count_session	
count_hit	
geoNetwork_latitude	
geoNetwork_longitude	
historic_session	
avg_session_time	
single_page_rate	
sessionQualityDim	

Outlier Detection:

It's an unusually extreme value that lies outside the typical range of values in a dataset. Identifying outliers is important in machine learning

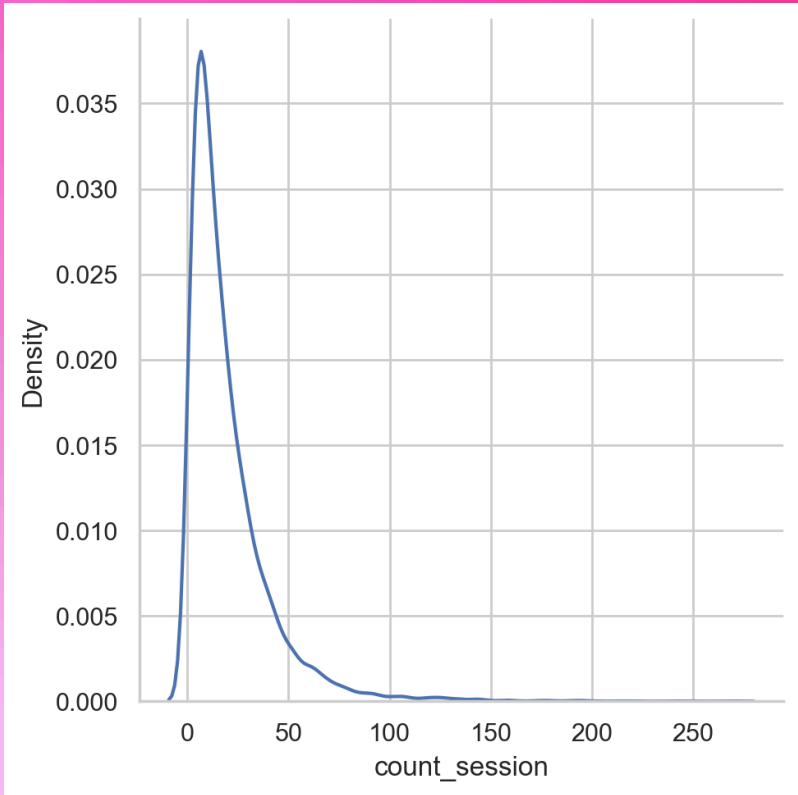
We can treat the outliers by changing the values using boxc
outlier the mean,median and mode values will be in the pl

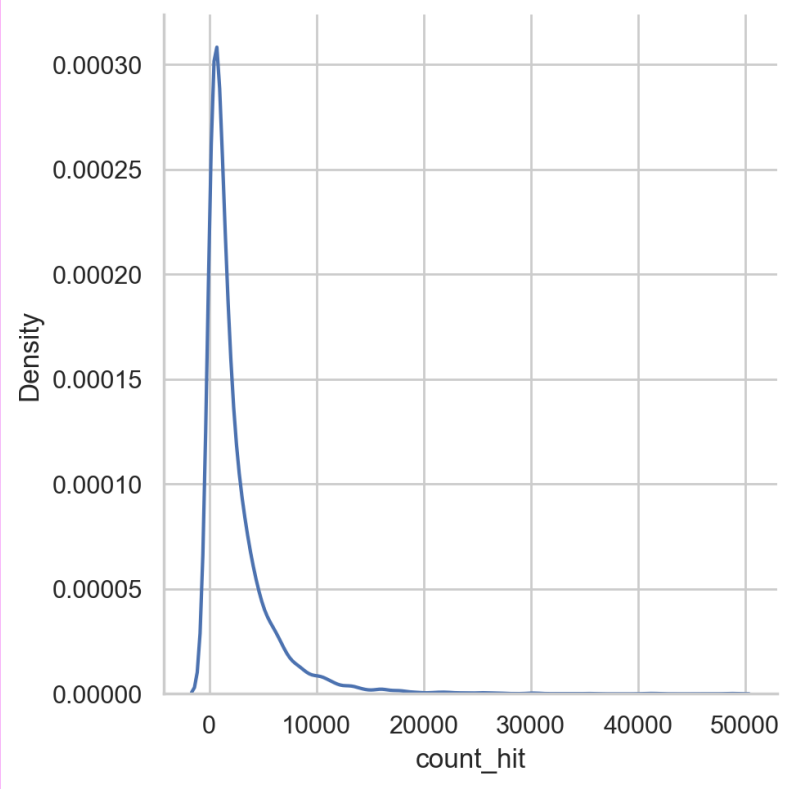


Distributation Curve:

This graphs are ploted to show the distribution of values for induvial columns. Most of the columns are Right Skewed this is not normaly distributed

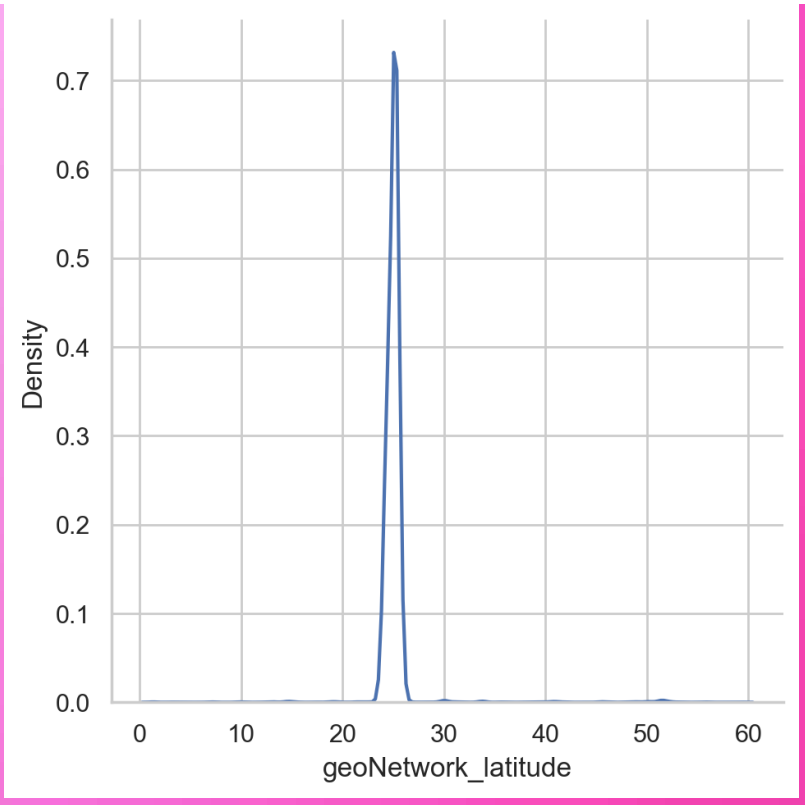
View Distribution curve

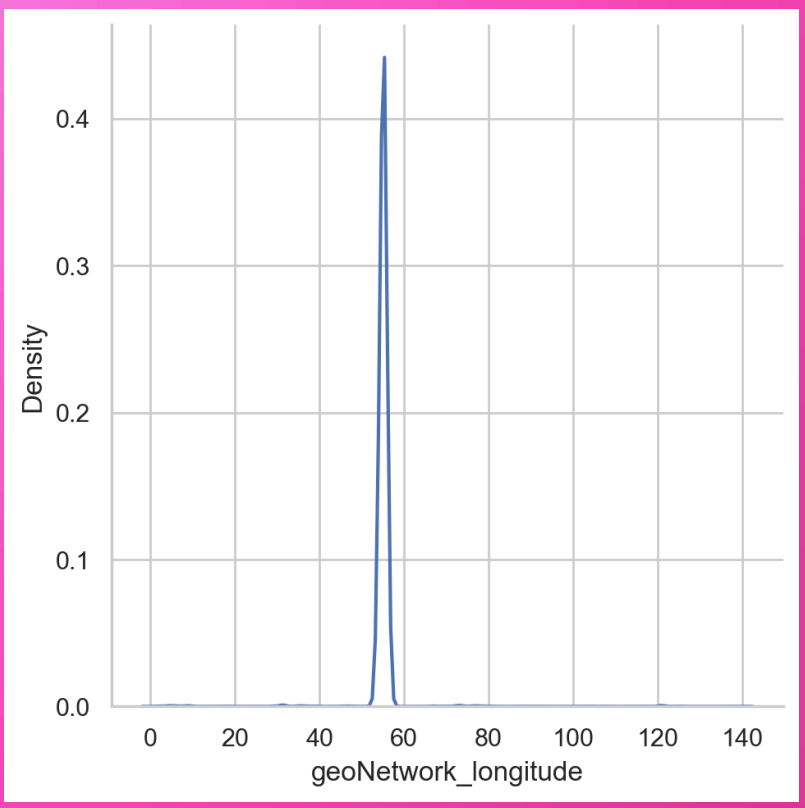


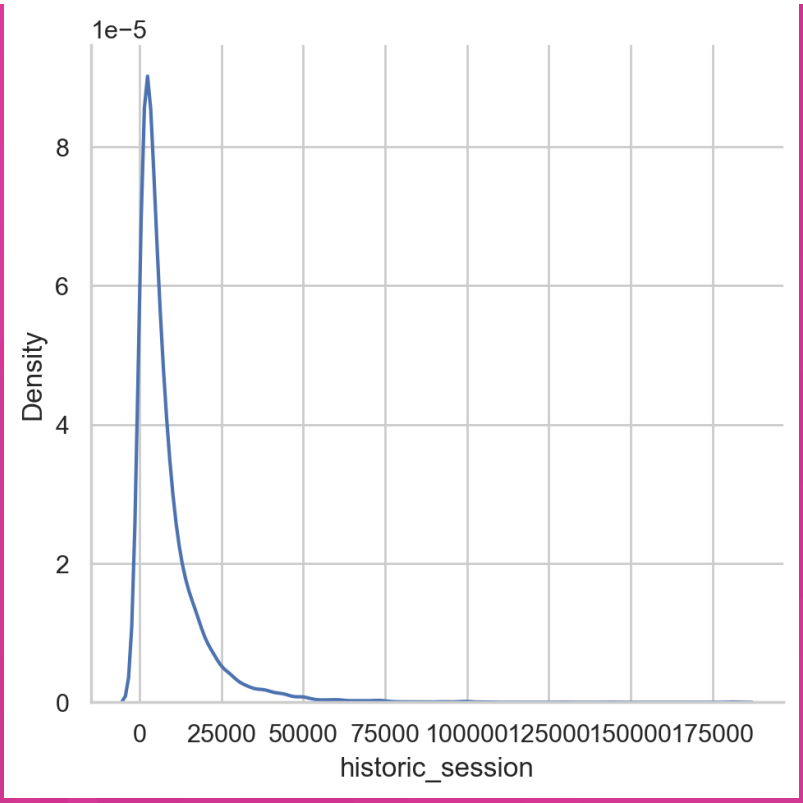


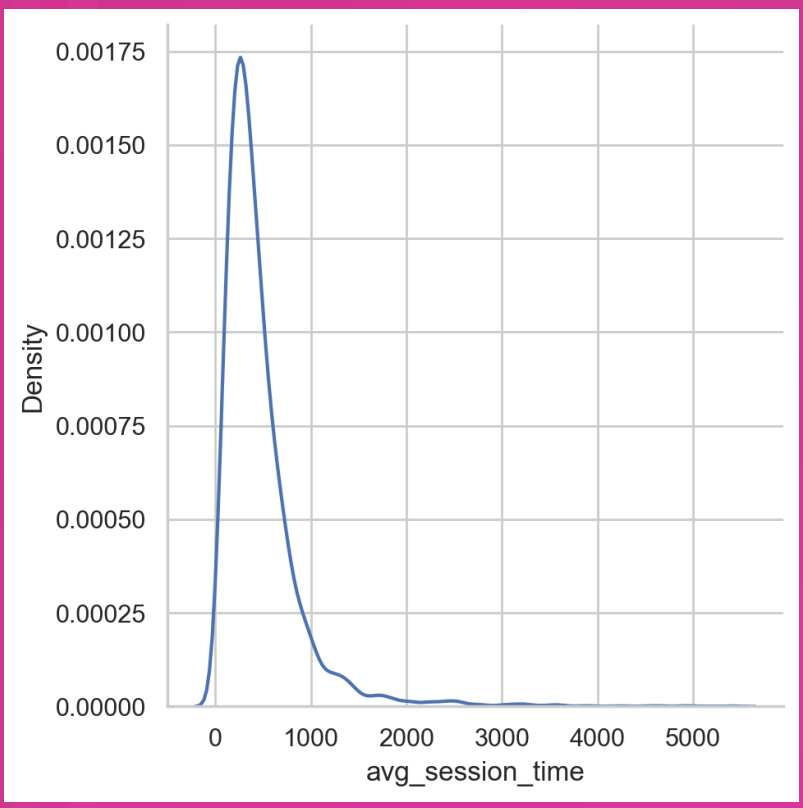
Here we used boxcox and log1p method to make right skev distribution

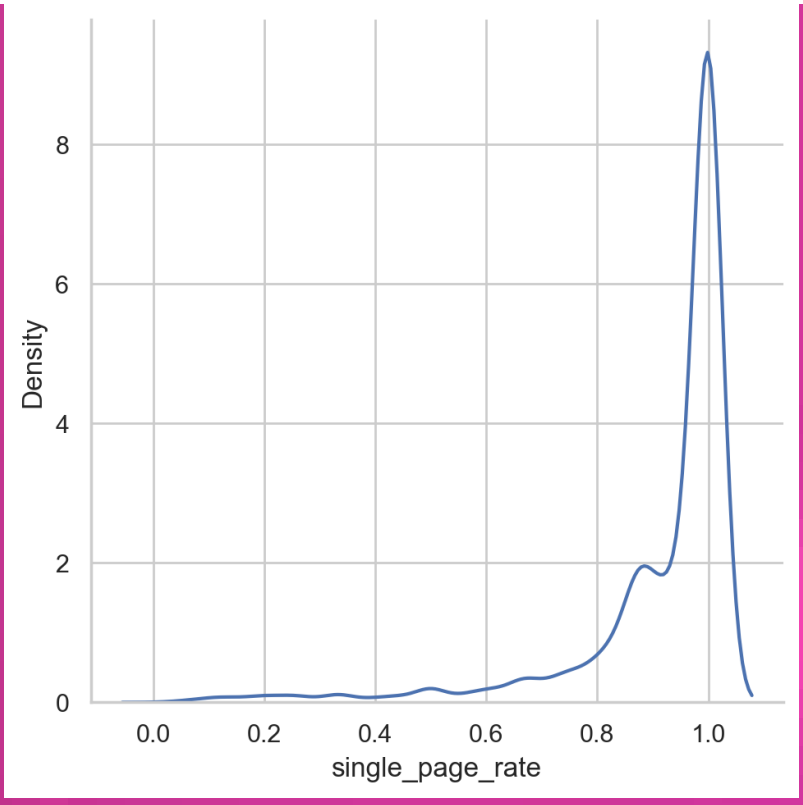
View Distribution curve

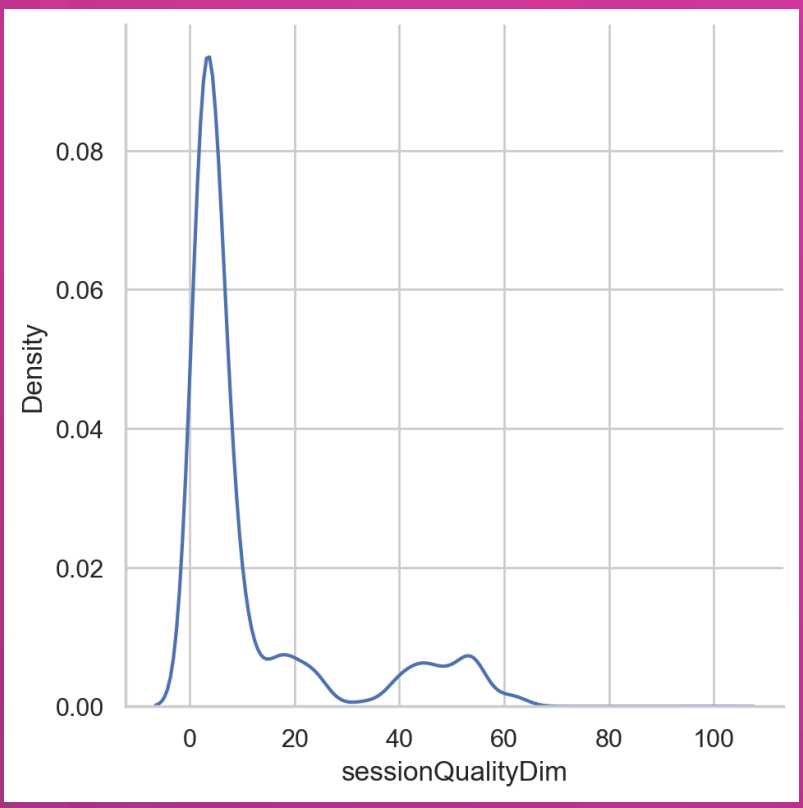

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
0
10
20
30
40
50
60
geoNetwork_latitude

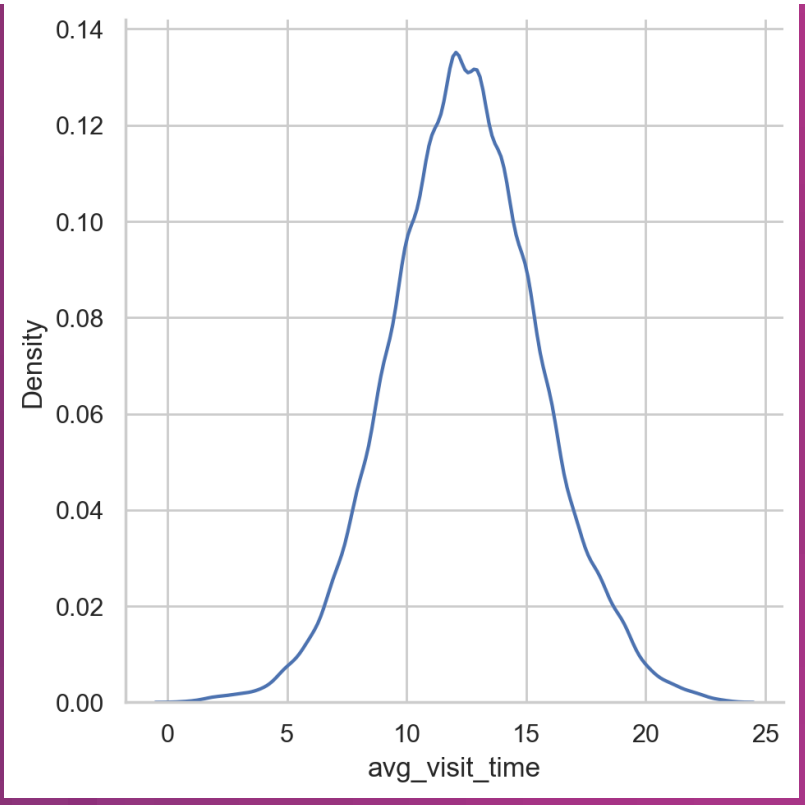

0.4
0.3
0.2
0.1
0.0
Density
0
20
40
60
80
100
120
140
geoNetwork_longitude

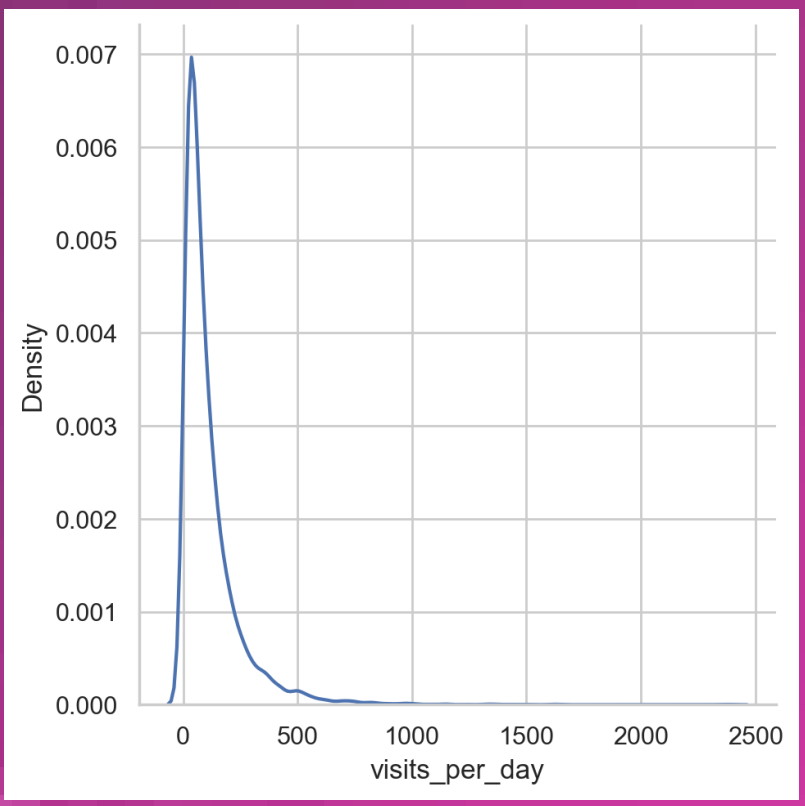

1e−5
8
6
4
2
0
Density
0
25000
50000
75000
100000
125000
150000
175000
historic_session

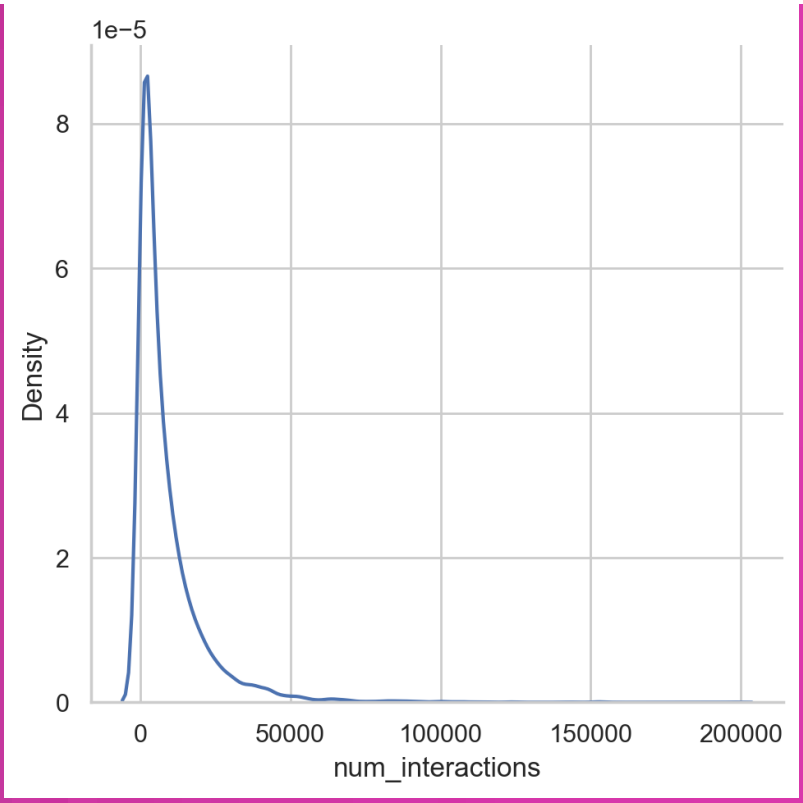

0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
Density
0
1000
2000
3000
4000
5000
avg_session_time

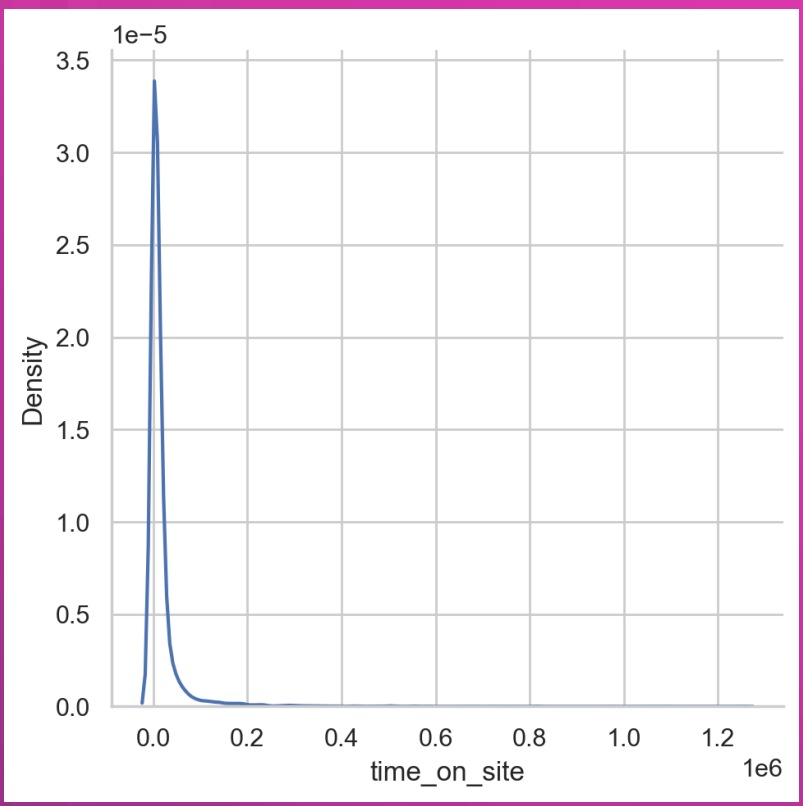

Density
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
single_page_rate

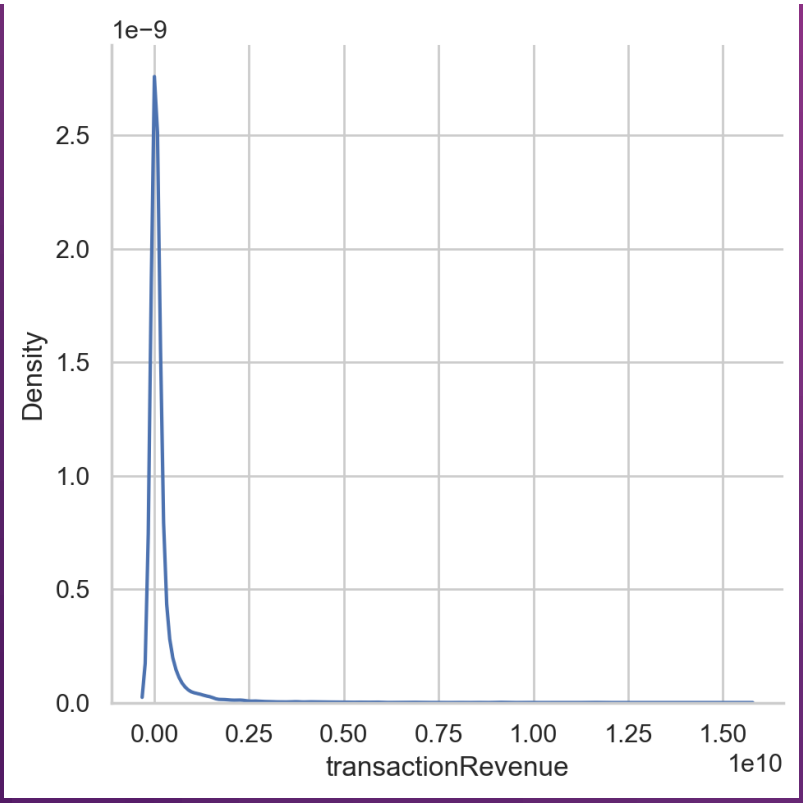

Density
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
sessionQualityDim

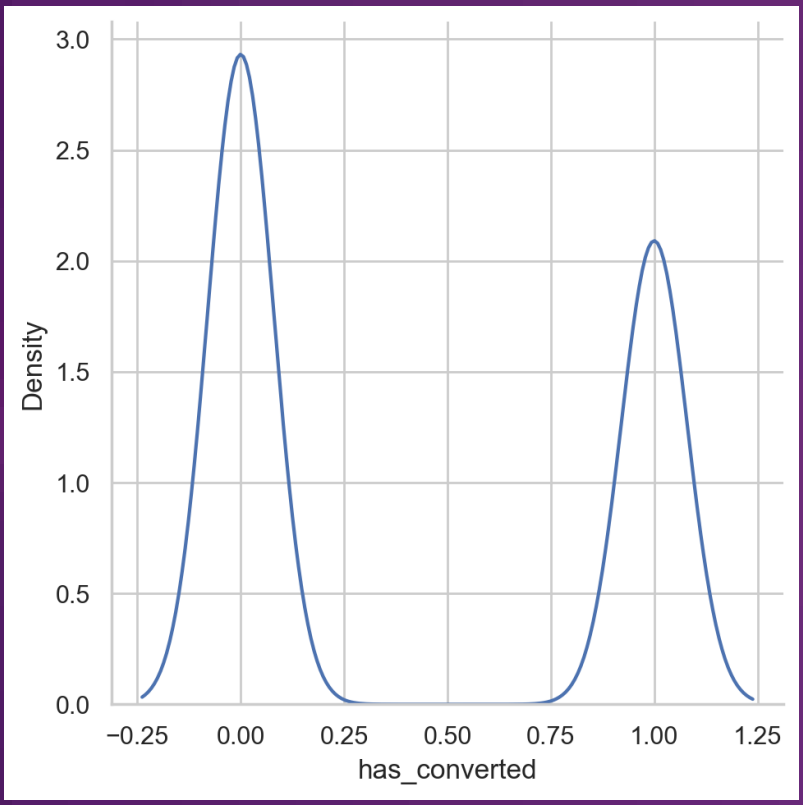

0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
0
5
10
15
20
25
avg_visit_time


0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
500
1000
1500
2000
2500
visits_per_day


1e−5
8
6
4
2
0
Density
0
50000
100000
150000
200000
num_interactions


1e−5
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6
time_on_site

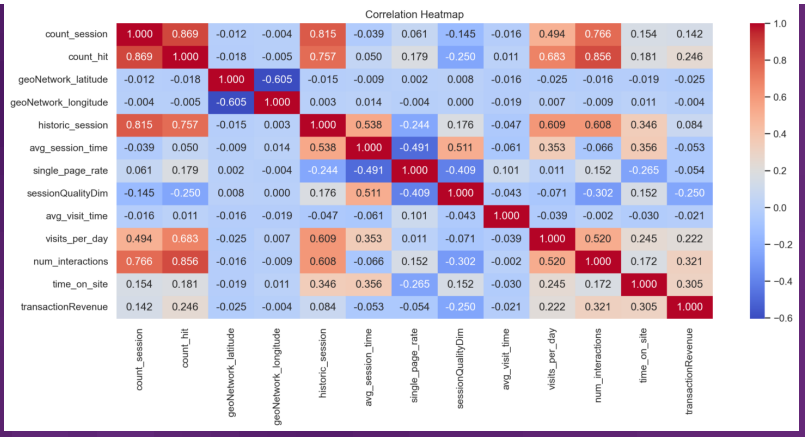




Correlation Heatmap:

Correlation heatmap for all continous variables that quantifies the degree to which two variables are related

From the previous correlation map the columns **count sess historic_session, single_page_rate** are having highest cor columns



Feature Importance:

	Columns	Feature_percentage
8	transactionRevenue	60.8309
3	sessionQualityDim	12.7368
6	num_interactions	6.272
7	time_on_site	6.2651
2	avg_session_time	6.1817
5	visits_per_day	4.3661
4	avg_visit_time	1.9718
0	geoNetwork_latitude	0.7012
1	geoNetwork_longitude	0.6743

In this Feature Importance we can understand the importa dataset.

We can remove the sessionqualityDim, geonetwork latitud avg_"visit_time