Data Preparation and Exploratory Data Analysis Report

---

1. Data Preparation 2
Dataset 1: b.csv 2
Dataset 2: c.csv 3
Dataset 3: ct.csv 3
Dataset 4: s.csv 4
Dataset 5: tp.csv 5
Merged data set: c.csv, ct.csv and tp.csv 6
Merged data set: b.csv and s.csv 6
Final merged data set: base dataset 7
Splitting base data: 9

Value counts: 12

2. Exploratory Data analysis of base data 13
Distribution across various features 13
i) Conversion_status: 1.5% of calls got converted to them purchasing premium services. 13
ii) nunique_report_type 13
iii) nunique_gender 13
iv) conversation_status 14
OUTLIER DETECTION using BOX PLOTS 14
Box plots before the outlier detection 14
Box plots after outlier detection 15
Correlational analysis 16
Resampling of data 17

1. Data Preparation

Dataset 1: b.csv

Dimensions: 6970265 rows x 4 columns

Column	Column type	Null count	Unique count
uuid	float64	2	1794005
Beacon ype	str	2	42
Beacon value	float64	2	277
Log date	str	0	92

Changes made to the data while data preparation:

1. Changing the data types.
2. Extracting the dates from the log date.
3. Remove decimal parts of uuid.
4. Group by date and uuid.
5. Aggregate by number of unique values for beacon types.

Dimensions: 6970263 rows x 4 columns

Column	Column type	Null count	Unique count
uuid	str	0	1794005
Beacon type	str	0	42
Beacon value	int64	0	277
Log date	str	0	92

Dataset 2: c.csv

Dimensions: 2295101 rows x 5 columns

Column	Column type	Null count	Unique count
id	int64	0	2295101
email	int64	0	2295101
Primary phone	float64	793012	1348348
Secondary phones	float	2180343	97881
Profile submit count	int64	0	413

Dataset 3: ct.csv

Dimensions: 3304478 rows x 5 columns

Column	Column type	Null count	Unique count
id	int64	0	3304478
cid	int64	0	1325451
timestamp	str	0	92
amount	float64	0	878
status	str	0	10

Dataset 4: s.csv

Dimensions: 9095602 rows x 10 columns

Column	Column type	Null count	Unique count
uuid	int64	0	9095602
phone	float64	977	3399997
status	int64	0	1
gender	str	4765	6
dob	str	20	36934
language	str	398	17
email	float64	733	3259793
Report type	str	70	81
device	str	187	5
Log date	str	0	8285461

Changes made to the data while data preparation:

1. Drop status and log date.
2. Extract the date.
3. Fix language codes.
4. Fix device code.
5. Change to appropriate data types.

Dimensions: 9088534 rows x 9 columns

Column	Column type	Null count	Unique count
uuid	str	0	9088534
phone	str	0	3398850
gender	str	0	6
dob	str	0	36751
language	str	0	17
email	str	0	3258713
Report type	str	0	78
device	str	0	5
Log date	datetime.date	0	827

Dataset 5: tp.csv

Dimensions: 4179025 rows x 4 columns

Column	Column type	Null count	Unique count	Null percent
ctid	int64	0	4170263	0
variant	str	0	6	0
language	str	1824	30	0.043647
status	float	4127969	5	98.778303

Changes made to the data while data preparation:

1. Drop status.
2. Change data types.
3. Fix the language codes.
4. Drop the duplicates.

Data was cleaned after merging.

Csv files merged:

- c.csv, ct.csv, tp.csv are merged into a single csv file.
- b.csv and s.csv are merged into another single csv file.

Merged data set: c.csv, ct.csv and tp.csv

Dimensions: 3308922 rows x 9 columns

Column	Column type	Null count	Unique count
date	str	0	92
email	int64	0	1325200
Conversation status	str	0	2
Profile submit count	int64	0	413
Transactions amount	float64	0	2320

Merged data set: b.csv and s.csv

Dimensions: 1604430 rows x 14 columns

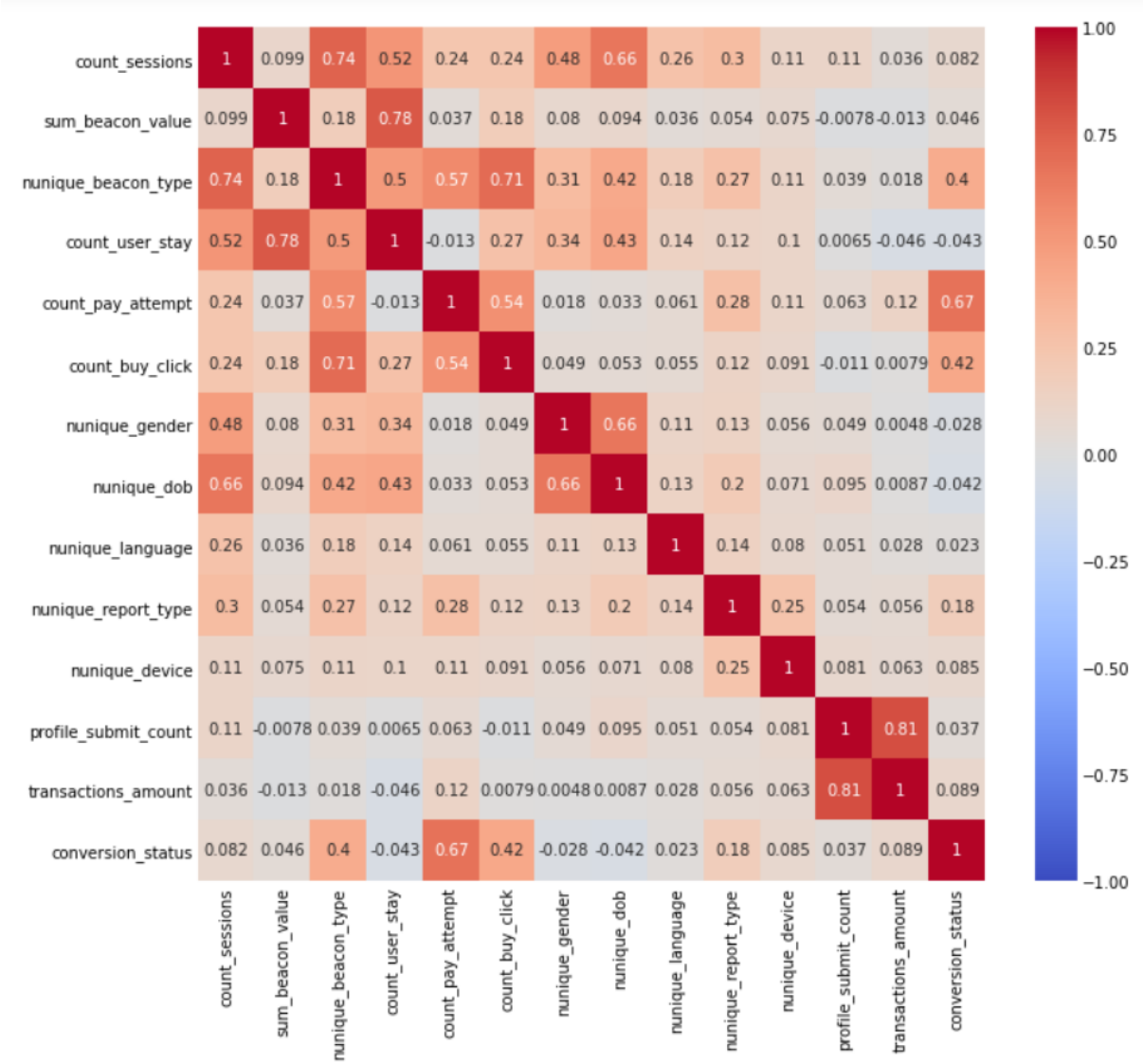
Column	Column type	Null count	Unique count
date	str	0	92
email	int64	0	824412
Count sessions	int64	0	56
Sum beacon value	int64	0	2549
Count user stay	int64	0	237
Count pay attempt	int64	0	45
Count buy click	int64	0	42
Nunique gender	int64	0	3
Nunique dob	int64	0	42
Nunique language	int64	0	8
Nunique report type	int64	0	13
Nunique device	int64	0	5

Final merged data set: base dataset

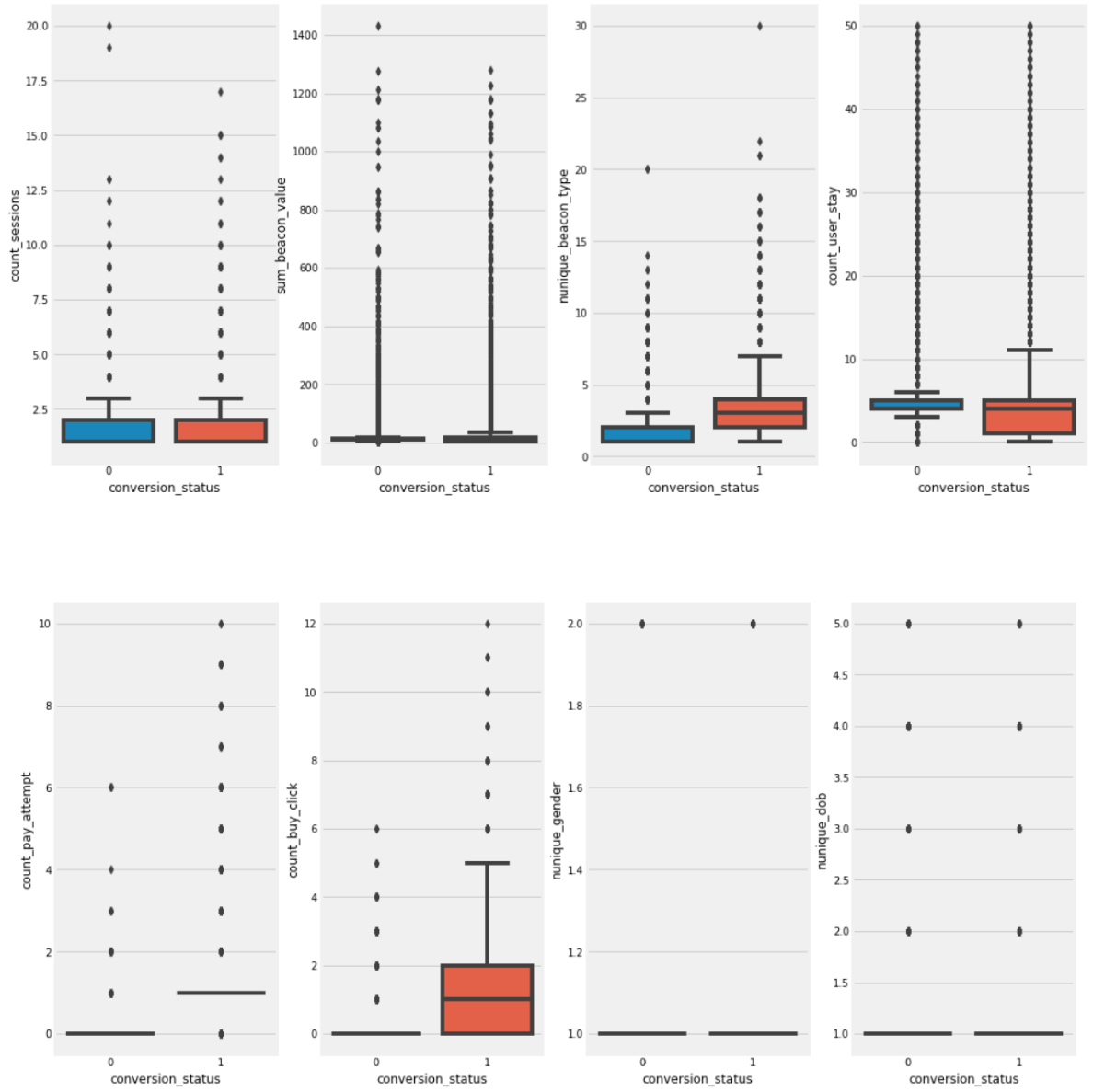
Dimensions: 10642016 rows x 16 columns

Column	Column type	Null count	Unique count
date	str	0	92
email	int64	0	643967
Count sessions	int64	0	52
Sum beacon value	int64	0	2257
Nunique beacon type	int64	0	58
Count user stay	int64	0	228
Count pay attempt	int64	0	44
Count buy click	int64	0	40
Nunique gender	int64	0	3
Nunique dob	int64	0	35
Nunique language	int64	0	8
Nunique report type	int64	0	12
Nunique device	int64	0	4
Conversation status	int64	0	2
Profile submit count	int64	0	360
Transactions amount	float64	0	1288

Correlational analysis:



Box plots after outlier detection



Splitting base data:

Splitting base data into two parts:

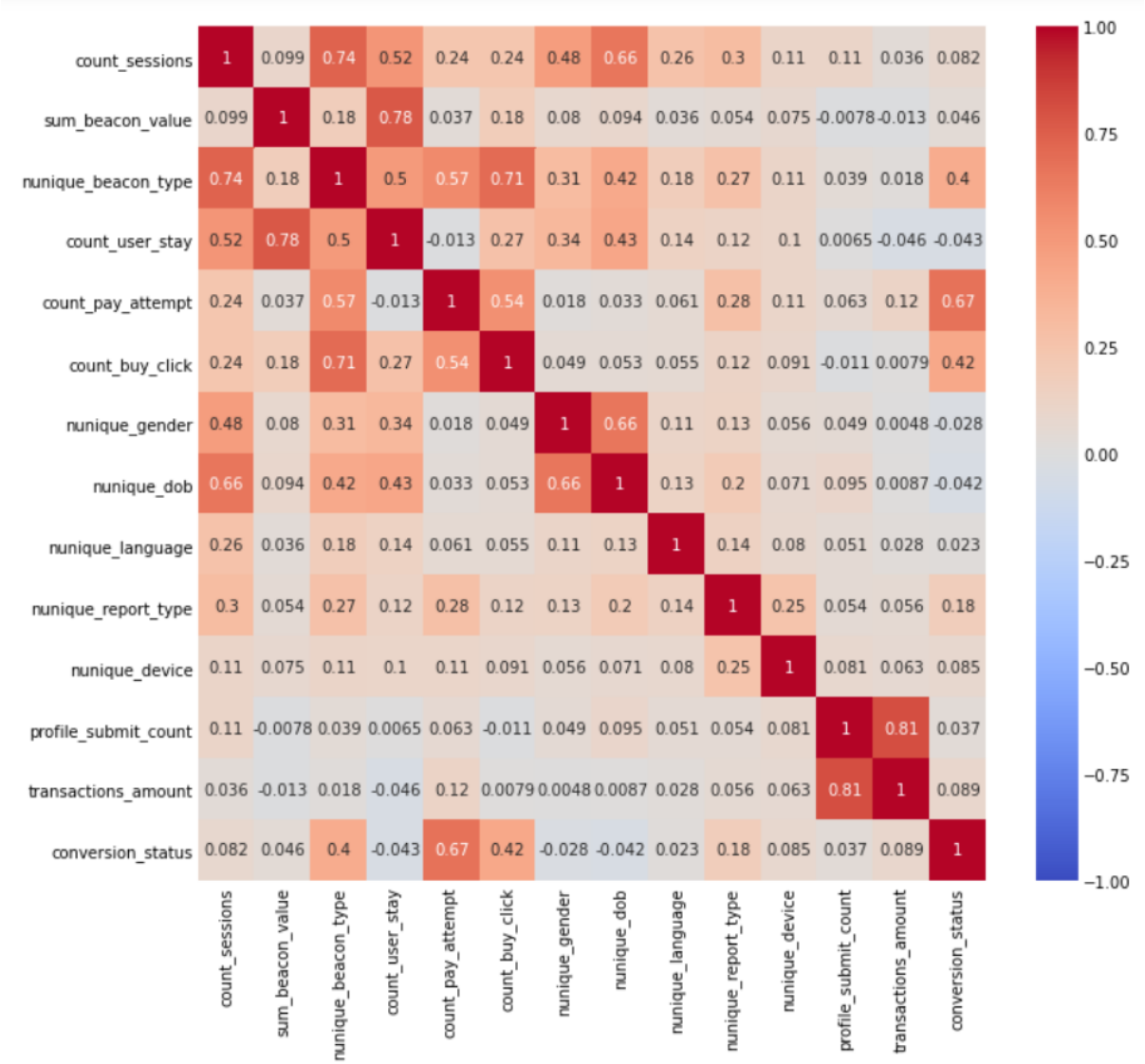
1. base_data_dev
2. base_data_ops

1. base_data_dev

Dimensions: 798162 rows x 16 columns

Column	Column type	Null Count	Unique count
date	int64	0	92
email	int64	0	643967
Count sessions	int64	0	52
Sum beacon value	int64	0	2257
Nunique beacon type	int64	0	58
Count user stay	int64	0	228
Count pay attempt	int64	0	44
Count buy click	int64	0	40
Nunique gender	int64	0	3
Nunique dob	int64	0	35
Nunique language	int64	0	8
Nunique report type	int64	0	12
Nunique device	int64	0	4
Conversion status	int64	0	2
Profile submit count	int64	0	360
Transactions amount	float64	0	1288

Value counts:
1 12624
0 785538
Name: conversion_status, dtype: int64


count_sessions 1 0.099 0.74 0.52 0.24 0.24 0.48 0.66 0.26 0.3 0.11 0.11 0.036 0.082
sum_beacon_value 0.099 1 0.18 0.78 0.037 0.18 0.08 0.094 0.036 0.054 0.075 -0.0078 -0.013 0.046
nunique_beacon_type 0.74 0.18 1 0.5 0.57 0.71 0.31 0.42 0.18 0.27 0.11 0.039 0.018 0.4
count_user_stay 0.52 0.78 0.5 1 -0.013 0.27 0.34 0.43 0.14 0.12 0.1 0.0065 -0.046 -0.043
count_pay_attempt 0.24 0.037 0.57 -0.013 1 0.54 0.018 0.033 0.061 0.28 0.11 0.063 0.12 0.67
count_buy_click 0.24 0.18 0.71 0.27 0.54 1 0.049 0.053 0.055 0.12 0.091 -0.011 0.0079 0.42
nunique_gender 0.48 0.08 0.31 0.34 0.018 0.049 1 0.66 0.11 0.13 0.056 0.049 0.0048 -0.028
nunique_dob 0.66 0.094 0.42 0.43 0.033 0.053 0.66 1 0.13 0.2 0.071 0.095 0.0087 -0.042
nunique_language 0.26 0.036 0.18 0.14 0.061 0.055 0.11 0.13 1 0.14 0.08 0.051 0.028 0.023
nunique_report_type 0.3 0.054 0.27 0.12 0.28 0.12 0.13 0.2 0.14 1 0.25 0.054 0.056 0.18
nunique_device 0.11 0.075 0.11 0.1 0.11 0.091 0.056 0.071 0.08 0.25 1 0.081 0.063 0.085
profile_submit_count 0.11 -0.0078 0.039 0.0065 0.063 -0.011 0.049 0.095 0.051 0.054 0.081 1 0.81 0.037
transactions_amount 0.036 -0.013 0.018 -0.046 0.12 0.0079 0.0048 0.0087 0.028 0.056 0.063 0.81 1 0.089
conversion_status 0.082 0.046 0.4 -0.043 0.67 0.42 -0.028 -0.042 0.023 0.18 0.085 0.037 0.089 1
count_sessions
sum_beacon_value
nunique_beacon_type
count_user_stay
count_pay_attempt
count_buy_click
nunique_gender
nunique_dob
nunique_language
nunique_report_type
nunique_device
profile_submit_count
transactions_amount
conversion_status
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

2. base_data_ops

Dimensions: 266054 rows x 16 columns

Column	Column type	Null Count	Unique count
date	int64	0	92
email	int64	0	239453
Count sessions	int64	0	46
Sum beacon value	int64	0	1387
Nunique beacon type	int64	0	46
Count user stay	int64	0	171
Count pay attempt	int64	0	24
Count buy click	int64	0	20
Nunique gender	int64	0	3
Nunique dob	int64	0	31
Nunique language	int64	0	8
Nunique report type	int64	0	8
Nunique device	int64	0	4
Conversion status	int64	0	2
Profile submit count	int64	0	349
Transactions amount	float64	0	817

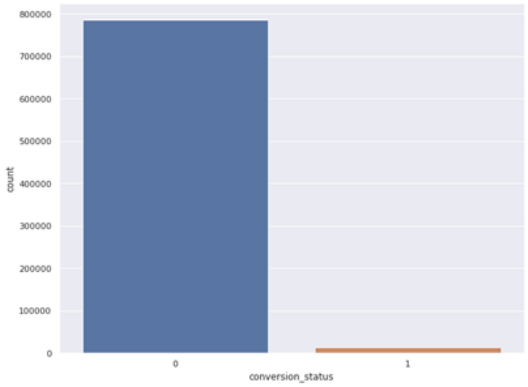
Value counts:

1 4208
0 261846
Name: conversion_status, dtype: int64

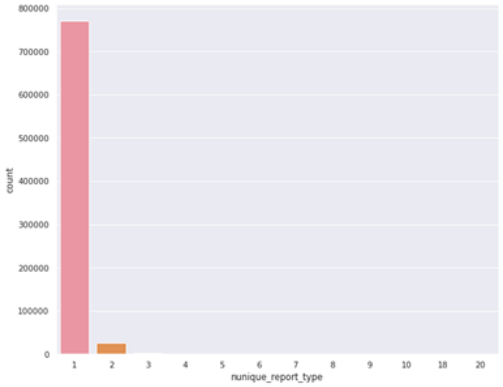
2. Exploratory Data analysis of base data

- Distribution across various features

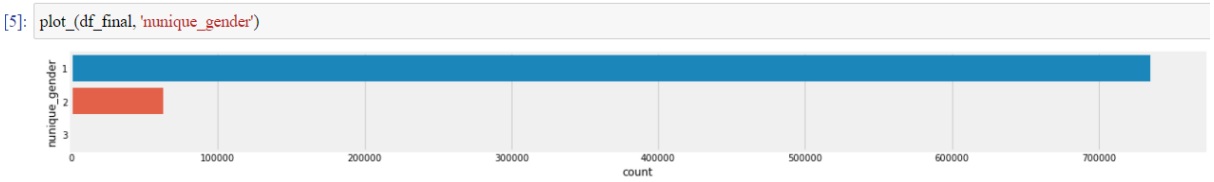
i) Conversion_status: 1.5% of calls got converted to them purchasing premium services.



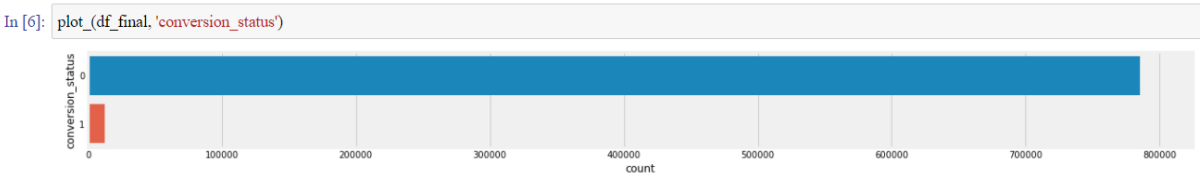
ii) nunique_report_type



iii) nunique_gender

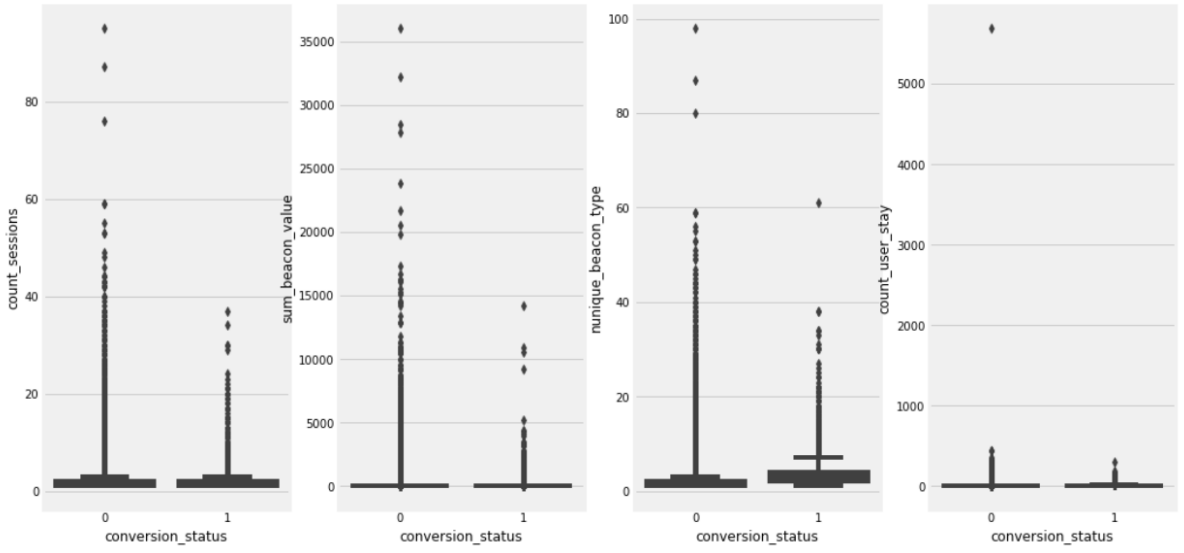


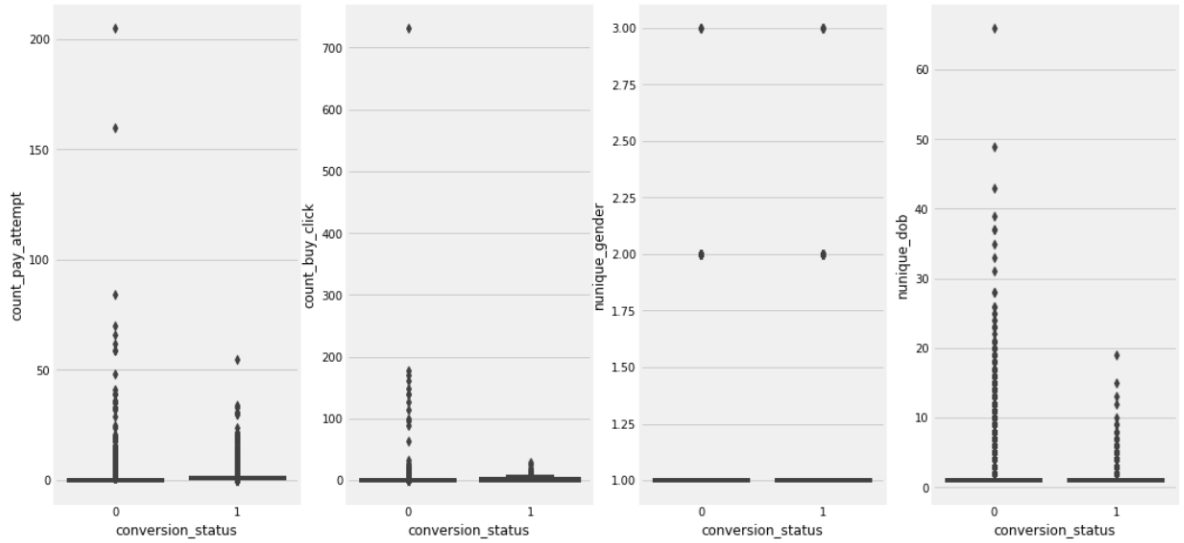
iv) conversation_status



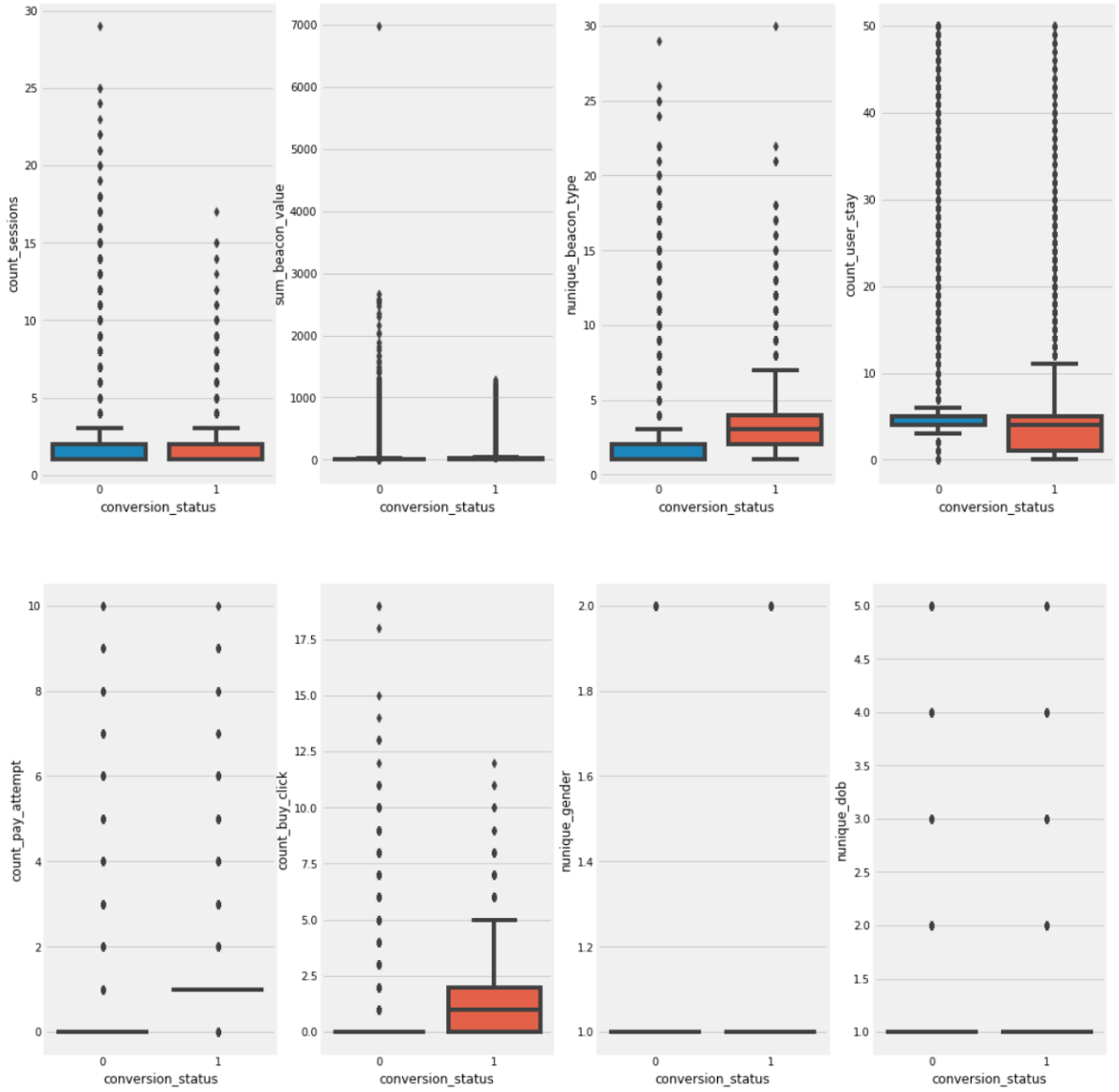
- OUTLIER DETECTION using BOX PLOTS

Box plots before the outlier detection

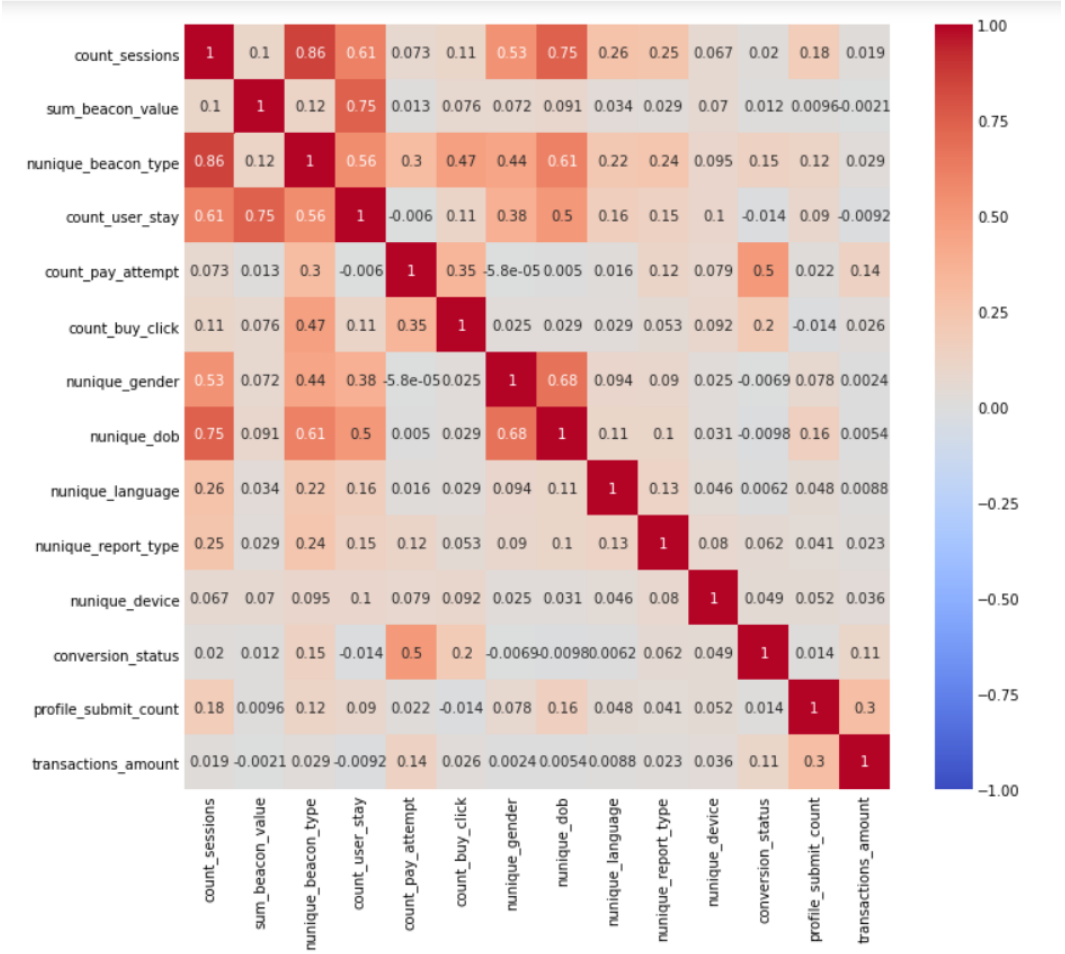




Box plots after outlier detection



- Correlational analysis



- Resampling of data

Dimensions: 26303 rows x 16 columns

Column	Column type	Null Count	Unique count
date	str	0	92
email	int64	0	24541
Count sessions	int64	0	17
Sum beacon value	int64	0	447
Nunique beacon type	int 64	0	22
Count user stay	int64	0	51
Count pay attempt	int64	0	11
Count buy click	int64	0	13
Nunique gender	int64	0	2
Nunique dob	int64	0	5
Nunique language	int64	0	2
Nunique report type	int64	0	2
Nunique device	int64	0	3
Profile submit count	int64	0	240
Transactional amount	float64	0	979
Conversion status	int64	0	2

Value counts
0 13619
1 12234
Name: conversion_status, dtype: int64

Correlational Analysis