Data Preparation and Exploratory Data Analysis Report

Index:

Expected result:

- After following the above procedure, we need to summarise all the data in 1 single table by choosing what is relevant for us, keeping in mind that we have shown at least 50% increase when compared to the baseline model during a time period of 7 days.

- The final/base table should contain two columns: customer_id and conversion_probability. And it should only have top 250 customers who have the potential of purchasing the products and have visited the website over the last 48 hours.

1. Data Preparation

Dataset 1: b.csv

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

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

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

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

Dataset 4: s.csv

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.

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

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

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

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

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

2. Exploratory Data analysis of base data

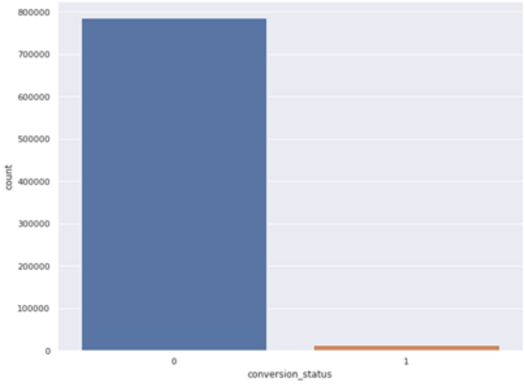
A. FINAL DATASET INFO:

i) Number of rows- 798162
ii) Number of columns- 16

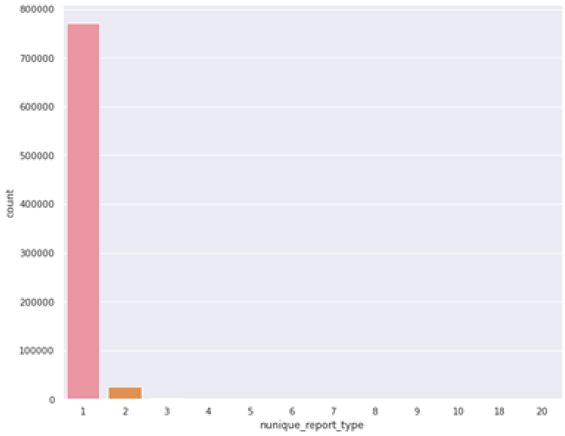
B. Questions explored

a) What is the distribution of status among the 798162 entries?

1.5% of calls to customers got converted to them purchasing premium services.



b) What is the distribution of different report counts that were sold?



c) What are the summary statistics of the amount? Are their outliers present?

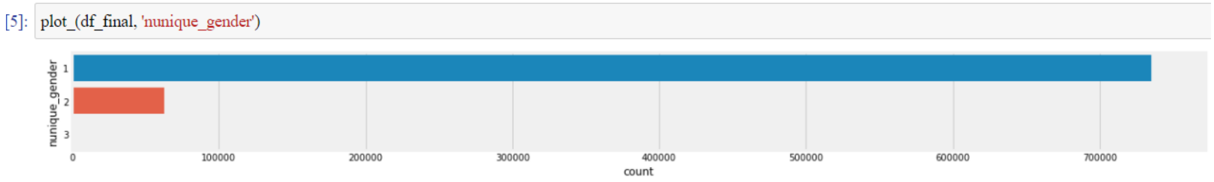
Report Count	Value Count
1	770190
2	25055
3	2521
4	325
5	51
6	10
10	3

- The amount looks heavily skewed to the right which needs further exploration.
- The top 3 most frequent transaction amounts are shown below:

0.0	733795
999.0	23246
-1.0	

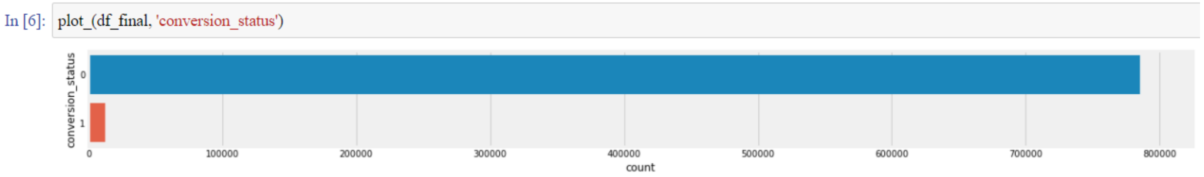
d) Distribution and count plots:

I. across gender



```
[5]: plot_(df_final, 'nunique_gender')
```

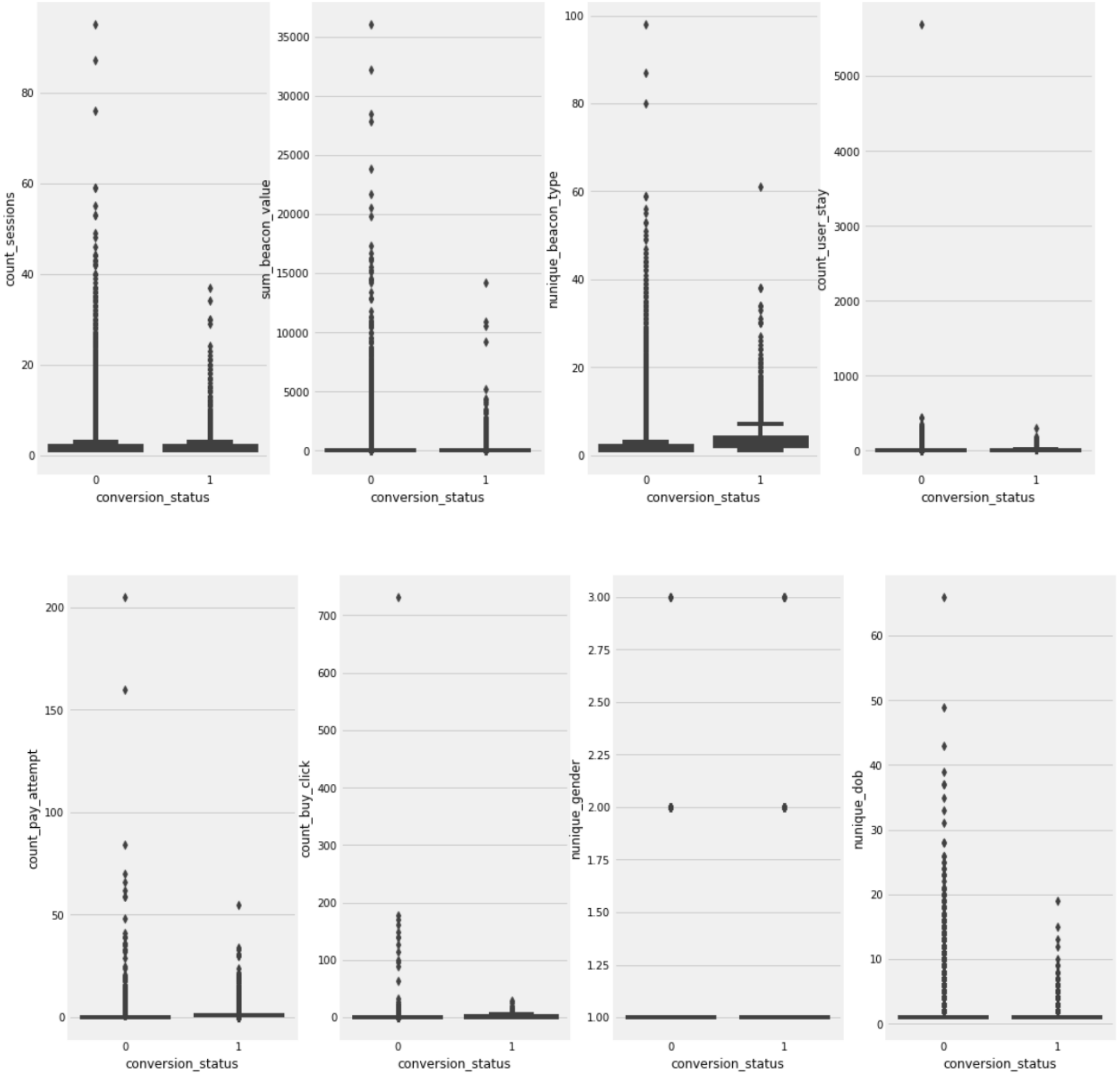
II. across conversation_status



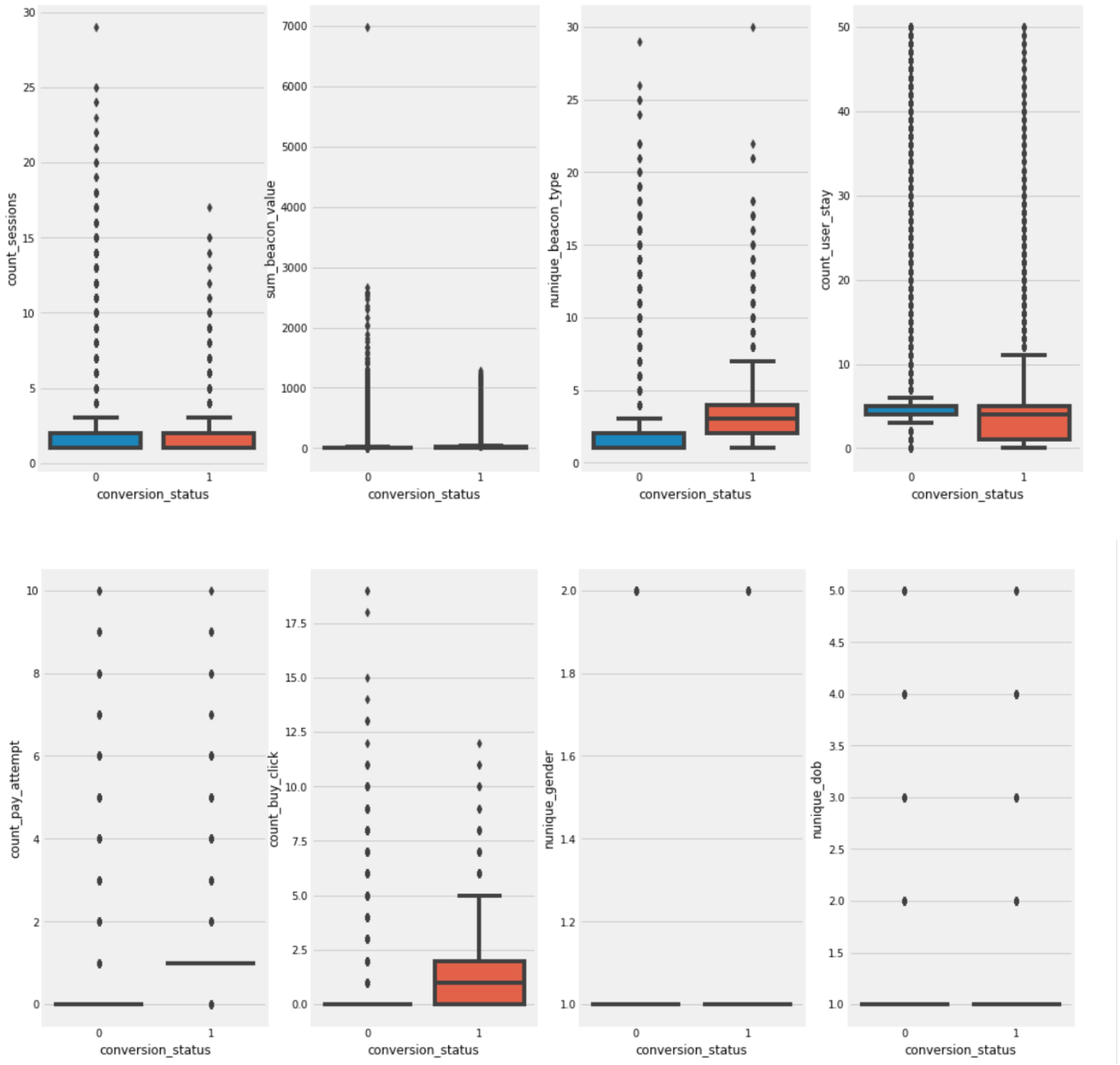
```
In [6]: plot_(df_final, 'conversion_status')
```

e) OUTLIER DETECTION using BOX PLOTS

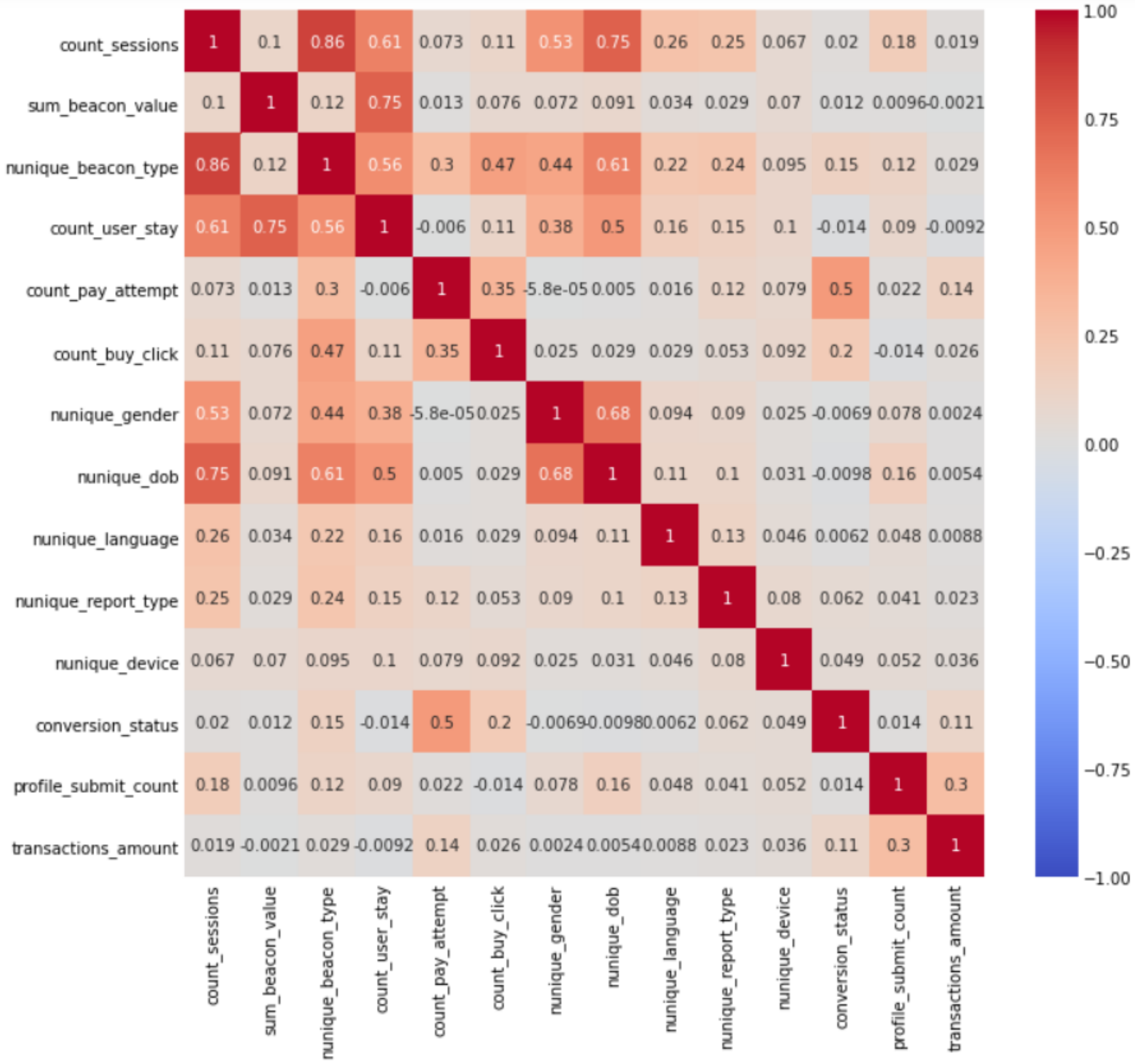
Box plots before the outlier detection



Box plots after outlier detection



f) Correlational analysis



	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	conversion_status	profile_submit_count	transactions_amount
count_sessions	1	0.1	0.86	0.61	0.073	0.11	0.53	0.75	0.26	0.25	0.067	0.02	0.18	0.019
sum_beacon_value	0.1	1	0.12	0.75	0.013	0.076	0.072	0.091	0.034	0.029	0.07	0.012	0.0096	-0.0021
nunique_beacon_type	0.86	0.12	1	0.56	0.3	0.47	0.44	0.61	0.22	0.24	0.095	0.15	0.12	0.029
count_user_stay	0.61	0.75	0.56	1	-0.006	0.11	0.38	0.5	0.16	0.15	0.1	-0.014	0.09	-0.0092
count_pay_attempt	0.073	0.013	0.3	-0.006	1	0.35	-5.8e-05	0.005	0.016	0.12	0.079	0.5	0.022	0.14
count_buy_click	0.11	0.076	0.47	0.11	0.35	1	0.025	0.029	0.029	0.053	0.092	0.2	-0.014	0.026
nunique_gender	0.53	0.072	0.44	0.38	-5.8e-05	0.025	1	0.68	0.094	0.09	0.025	-0.0069	0.078	0.0024
nunique_dob	0.75	0.091	0.61	0.5	0.005	0.029	0.68	1	0.11	0.1	0.031	-0.0098	0.16	0.0054
nunique_language	0.26	0.034	0.22	0.16	0.016	0.029	0.094	0.11	1	0.13	0.046	0.0062	0.048	0.0088
nunique_report_type	0.25	0.029	0.24	0.15	0.12	0.053	0.09	0.1	0.13	1	0.08	0.062	0.041	0.023
nunique_device	0.067	0.07	0.095	0.1	0.079	0.092	0.025	0.031	0.046	0.08	1	0.049	0.052	0.036
conversion_status	0.02	0.012	0.15	-0.014	0.5	0.2	-0.0069	-0.0098	0.0062	0.062	0.049	1	0.014	0.11
profile_submit_count	0.18	0.0096	0.12	0.09	0.022	-0.014	0.078	0.16	0.048	0.041	0.052	0.014	1	0.3
transactions_amount	0.019	-0.0021	0.029	-0.0092	0.14	0.026	0.0024	0.0054	0.0088	0.023	0.036	0.11	0.3	1