Contents

0. Objective

a) To gain understanding and find cues from the **6 csv files** containing our data.
b) To formulate assumptions and hypothesis for our modelling.
c) To check the quality of data and decide the kind of pre-processing required to make the data model ready.

1. Outline

- Exploring the raw data.
- Transforming the data appropriately and redoing the EDA.
- Summarizing the results and key insights.

2. Analysis of Tables

Part-1: Exploring as is

i) a

- The table has 7 features and 998822 rows. 4 of the features are categorial.
- Value counts of **status**, **type** and **marker** columns are shown alongside.

```
 1     925749
 10     72274
 0        488
-99       273
-10        36
-1          2
Name: marker, dtype: int64
```

```
Assigned                516435
purchase                261558
Not picking up           61454
Not interested           47341
Partially interested     19780
Follow-up later          18244
User is Interested       17756
Not reachable             9774
Invalid Number            9327
Line Busy                 8628
Already purchased         8123
Switched off              6316
New product potential     5938
Converted                 5579
Other                     1655
Could not call             632
assigned                   252
Has complaints              22
AC                           2
AB                           2
AA                           2
AD                           1
none                         1
Name: status, dtype: int64
```

```
1002    516685
1001    261752
2106     61455
2005     47341
2208     19780
2206     18242
2207     17753
2102      9774
2001      9327
2103      8627
2010      7927
2104      6315
2209      5937
2011      5578
2013      1656
2105       632
2012        21
1005         8
1003         7
1004         4
1006         1
Name: type, dtype: int64
```

- Log time to be explored in the next section.
- We have *607732 unique phone numbers* and **998822** total entries indicating that may be people have used the same number to generate reports which could be a potential lead.
- The **feature** status has 23 categories and it looks like that the data was manually collected by the sales team after contacting the potential lead.
- **Product** and **pay_mode** have majority missing values which could be because these entries have values **only when the customer gets converted**.

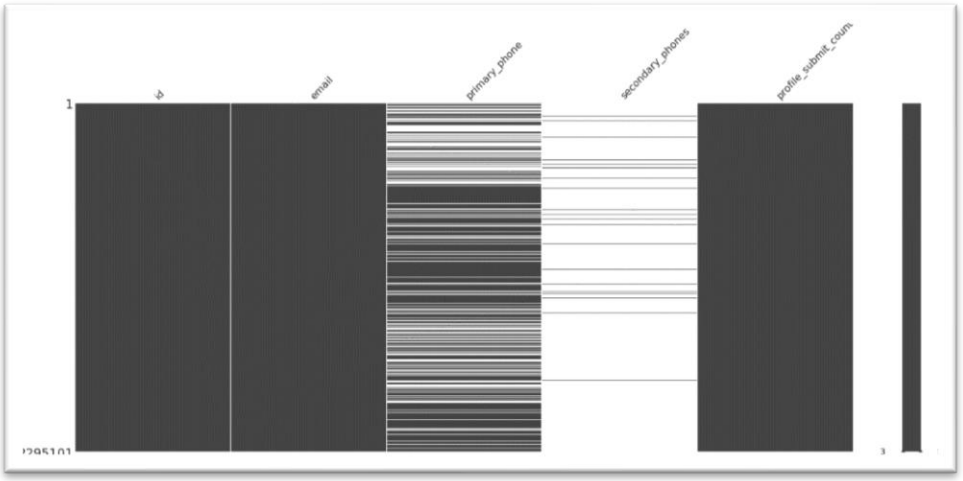
ii) b

- The table has 5 features and 10058164 entries.
- **Beacon_type** and **beacon_value** seem categorical but need to be explored further.
- The **beacon_value** description and the top 20 **beacon_type** value counts are shown alongside.
- Log date to be explored in the next section.

```
count   39009330.000000
mean            5.610764
std            11.020478
min             1.000000
25%             1.000000
50%             2.000000
75%             4.000000
max           999.000000
Name: beacon_value, dtype: float64
```

```
user_stay              33602564
masked_content          2075807
bottom_banner           1767163
buy_button_top           666880
pay_button_gpay          191147
pay_fail                 168369
buy_button_FH            133763
pay_button_cc-dc          99153
pay_button_cp-nb          56579
pay_button_upi            56436
pay_button_ptm            52459
buy_button_PP             36704
pay_button_CC             19748
pay_button_paypal         17948
buy_button_CR              9164
pay_button_PhonePe         8327
pay_button_NB              8172
pay_button_UPI             8045
pay_button_PTM             7615
pay_button_GPAY            6693
Name: beacon_type, dtype: int64
```

iii) c



- The table has 5 features and 2295101 entries.
- The missing number matrix is shown alongside. **Primary phone** and **secondary phones** have the highest percentage of null values.
- Profile submit count is a numerical variable with 413 unique entries top 20 of which are shown alongside.

```
2     1087209
1      540209
3      262788
4      157448
5       71038
6       48311
7       28264
8       21162
9       13745
10      10986
11       7967
12       6344
13       4943
14       4286
15       3267
16       2913
17       2266
18       2049
19       1743
20       1518
Name: profile_submit_count, dtype: int64
```

iv) ct

- The table has 5 features and 4174013 entries.
- **cid** represents the id of 1633595 customers.
- This table seems to hold a part of the **transaction data** of the customers who made a purchase.
- The descriptive statistics of customers who made a **purchase > 0** and the different categories of the payment status is shown alongside.

```
count   175996.000000
mean       852.058584
std       1252.887682
min          0.330000
25%        520.000000
50%        999.000000
75%        999.000000
max     130113.740000
Name: amount, dtype: float64
```

```
N                   4115487
PROCESSED             48328
PAYMENT_COMPLETED      9226
SUSPENDED               526
INITIATED               279
ROLLED_BACK              85
PDF_ERROR                55
PAYMENT_FAILED           16
TOPROCESS                10
Y                         1
Name: status, dtype: int64
```

v) S

- The table has 10 features and 9095602 entries.
- **Uuid** seems to represent the unique id of the customers.
- About 59% population in this table is Male.
- The majority of people seeking services either spoke at least one of Tamil or English.
- The top 20 report types are shown alongside.
- Majority using the website use it on a mobile platform.

```
LS-MT    7908843
LS-MP     680885
LS-CR     243535
LS-FH      40034
LS-SC      34223
LS-WP      31226
LS-CP      28584
LS-FS      24479
MR         23670
LS-CU      20169
CCMB       14954
LS-WL      12188
INDP        7466
LI          5937
STNR        4536
YG          3112
RKTR        2671
CP          1440
JPTR        1411
ST           876
Name: report_type, dtype: int64
```

```
Male      0.587019
Female    0.409527
MALE      0.003445
M         0.000005
FEMALE    0.000002
F         0.000002
Name: gender, dtype: float64
```

```
TAM      2842756
ENG      2096781
HIN      1399821
KAN       666861
MAL       644473
TEL       635283
MAR       304390
BEN       234800
ORI       218320
GUJ        44658
SIN         7048
H              4
NIL            3
HINDI          2
SAN            2
E              1
M              1
Name: language, dtype: int64
```

vi) Tp

- The table has 5 features and 4170263 entries.
- Majority of customers were interested in the **basic** service report.
- In this table, majority requested for the service in either English or Hindi.

```
basic                    4008071
premium                   168432
premiumplus                 2134
premiumpluscolor             370
premiumplusconsultancy        10
premiumplusleather             7
Name: variant, dtype: int64
```

Part-2: Exploration after cleaning data

i) a

- Converted **status, type** and **marker** into categorical type because they had limited unique values.
- Created a new column called Date which is a **data_time** type obtained by changing the data type of **log_time**.
- Because phone number column was almost empty, I dropped it.
- The feature **product** also has 67% missing values but it may be important for our analysis. Therefore, it needs to be consulted upon before making any Imputation decision.

Questions: -

- What kind of a product/s does this table deal with?
- What is type and marker?

ii) b

- Dropped all the rows containing missing values.
- Converted **uuid** and **beacon_value** to int64 and float16 respectively.

Questions: -

- What does beacon collect?
- What is status with respect to this table?

s

iii) c

- Dropped primary_phone and secondary_phone numbers because majority was missing data.
- Final table looks like this:

Questions:

- What is profile submit count?

	id	email	profile_submit_count
0	1	537606	592
1	5	1443908	3
2	6	534973	6
3	7	3259797	3

iv) ct

- The cid column seems to represent the customer id.
- This table is fairly consistent with respect to null values.
- This table most likely contains the transaction data for a service and must be linked to that.

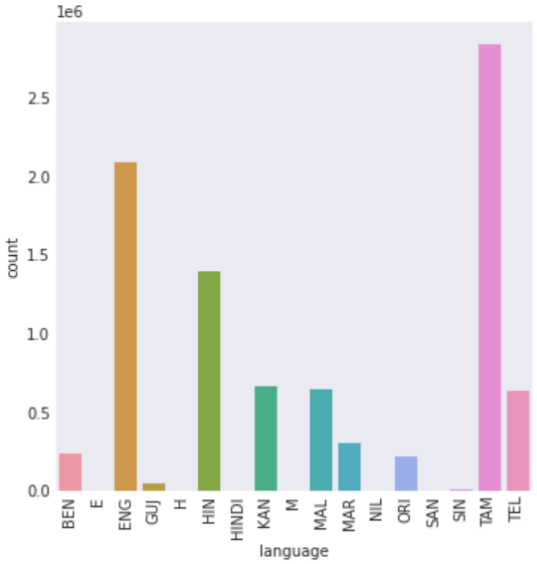
	id	cid	timestamp	amount	status
9	13	13	2021-01-04 09:16:32	2500.000000	N
10	14	13	2021-01-04 09:33:17	2500.000000	N
12	16	16	2021-01-04 09:44:14	1800.000000	N
25	29	29	2021-01-05 20:35:05	3540.000000	N

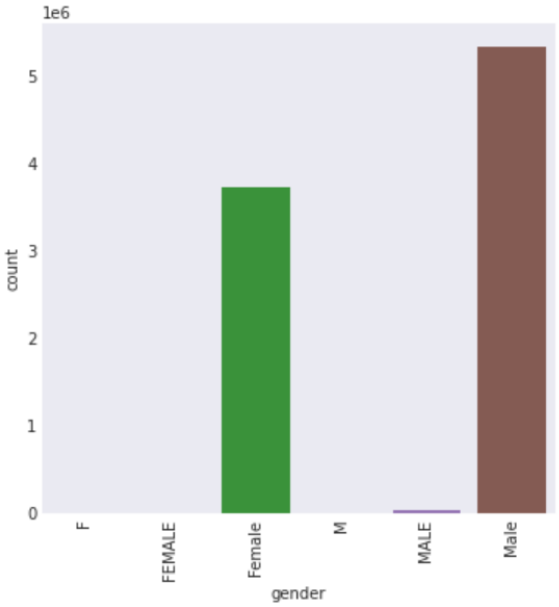
Questions:

- Which table is the Ct table linked to?
- The table alongside shows the amount need to be paid by those customers whose status was ‘N’. Does this mean they have denied or that the shall pay later?

v) s

- There are very few entries with null values (**at most 0.05%**). Therefore, we can safely drop them.
- Features: **status, gender, language and device** were converted into category types.
- 6112 date of birth entries have a value 000000 which is a potential null value. Because there is no clear strategy to impute the birth entries, we can simply remove these entries.





- Majority of subscribers were male.
- The two most widely used languages were Tamil and English.

Questions:

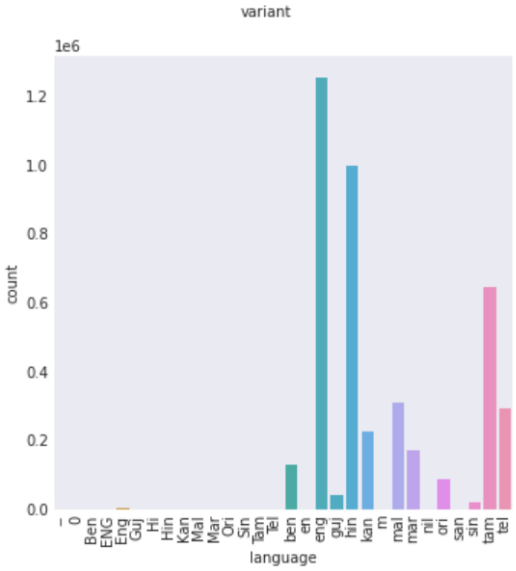
- Are the phone numbers truly non-unique for so many customers?
- Was dropping the 6112 entries with 000000 null values alright or should we explore some other techniques?
- Is this the data of all those customers who have subscribed to a report? Is this table linked to any other reflecting the payment status?

vi) tp

- About 98.8% of the status column has null values.
- Variant and language of the subscription were converted into categorical type.
- In this table, English and Hindi were the two most used language.
- The most common variant, about 96%, was ‘basic’.

Questions:

- Which table is this table linked to?
- When we say variants, what variants are we referring to?



Part-3: Insights and Summary

1. ***The presence of different ‘status’ values in different tables indicate that there may be multiple products we are dealing with.***

2. ***The product information is contained in table ‘a’ as shown alongside.***

In-depth	130692
LI	47037
LS Mini Horoscope	20152
Marriage Horoscope	17736
CCMB Combo	17454
CCMB	16456
In-depth Horoscope	15851
Yearly Horoscope	10248
Horoscope Compatibility	8195
OTH	7739
Saturn Transit Predicitons	7697
Jupiter Transit Predicitons	3232
Career and Business Horoscope	3206
Super Horoscope	3129
MR	2935

Name: product, dtype: int64

3. ***The table ‘ct’ contains the transaction data of the purchase made by the customer.***

4. ***For a specific product from table ‘a’, there seem to be several report types in table ‘s’ in a specific language. (shown in the second table)***

LS-MT	7905428
LS-MP	680756
LS-CR	243499
LS-FH	40034
LS-SC	34217
LS-WP	31224
LS-CP	28578
LS-FS	24479
MR	23651
LS-CU	17499
CCMB	14942
LS-WL	12173
INDP	7462
LI	5859
STNR	4536

Name: report_type, dtype: int64

5. ***Finally, the table ‘tp’ contains the variant of the report that the customer is willing to purchase. Tables ‘tp’ and ‘s’ appear to be linked.***

6. ***“In-depth LS-MT basic variant in English” is the most popular product among others.***