

Table of Contents

0. What is the objective of Initial Data Analysis?	1
1. Performing IDA on each Table.	2
i) Table a	2
ii) Table b	2
iii) Table c	3
iv) Table ct	3
v) Table s.....	4
vi) Table tp.....	4
2. Features of each table together	5

Phase 1 – Initial Data Analysis

0. What is the objective of Initial Data Analysis?

a) Uncover underlying structure of the dataset.

- i. Check Quality of the Data using Descriptive summary Statistics
- ii. Check the number of rows, columns, the size of the dataset.
- iii. Check for missing values, unique values and the data type of each feature.

b) Detect outliers and anomalies.

- i. Check for outliers within the dataset.
- ii. Check the Normality of the dataset

c) Treatment of problems (typically through transformations or imputations).

In this step, we fix all the problems we have identified so far. The problem could be dealing with null values or it could be transformation of non-normal variables.

For the initial IDA, we will skip this steps b and c.

1. Performing IDA on each Table.

i) Table a

File size: - 61MB

```

      column      column_type  null_count  unique_count  null_percent  unique_values
0  log_time      <class 'str'>         0        559772      0.000000  [2019-05-03 15:57:56, 2019-05-03 11:34:57, 201...
1   phone  <class 'numpy.float64'>         8        607732      0.000801  [0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...
2   status      <class 'str'>         0         23      0.000000  [assigned, purchase, AA, AB, AC, AD, Switched ...
3    type  <class 'numpy.int64'>         0         21      0.000000  [1002, 1001, 1003, 1005, 1004, 1006, 2005, 220...
4  product      <class 'float'>    668696        125    66.948465  [nan, In-depth Book, In-depth, MR, LI, OTH, CC...
5  pay_mode      <class 'float'>    775819         62    77.673399  [nan, cc-dc, cp-nb, PayPal, ptm, upi, paypal, ...
6   marker  <class 'numpy.int64'>         0          6      0.000000  [0, 1, -99, -1, -10, 10]
The columns are: Index(['log_time', 'phone', 'status', 'type', 'product', 'pay_mode', 'marker'], dtype='object')
Rows: 998822      Columns: 7

```

Key Takeaways: -

- The table has 7 features and 998822 entries.
- Two features **product** and **pay_mode** have a high percentage of missing values indicating that there may be some inconsistencies with data collection. Need to be explored further.
- Need to confirm with the data source what “**status**” is.

ii) Table b

File size: - 1.9GB

```

      column      column_type  null_count  unique_count  null_percent  unique_values
0      uuid  <class 'numpy.float64'>        15    10058149      0.000038  [0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...
1  beacon_type      <class 'str'>         2         66      0.000005  [user_stay, bottom_banner, buy_button_FH, pay...
2  beacon_value  <class 'numpy.float64'>         2         724      0.000005  [26.0, 32.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0...
3   log_date      <class 'str'>         0    27352274      0.000000  [2019-02-26 16:19:08, 2019-02-26 16:30:08, 201...
4    status  <class 'numpy.int64'>         0          1      0.000000  [1]
The columns are: Index(['uuid', 'beacon_type', 'beacon_value', 'log_date', 'status'], dtype='object')
Rows: 39009332      Columns: 5

```

Key Takeaways: -

- The table has 5 features and 39009332 entries.
- Data in this table seems fairly consistent with respect to missing values.
- The table seems to be a collection of beacon data.

iii) Table c

File size: - 75MB

```

      column      column_type  null_count  unique_count  null_percent  unique_values
0         id  <class 'numpy.int64'>         0        2295101         0.000000  [1, 5, 6, 7, 8, 9, 10, 11, 13, 16, 17, 18, 20,...
1        email  <class 'numpy.int64'>         0        2295101         0.000000  [537606, 1443908, 534973, 3259797, 1701404, 11...
2   primary_phone  <class 'numpy.float64'>    793012    1348348    34.552379  [22.0, nan, 153435.0, 3475327.0, 171697.0, 47....
3  secondary_phones      <class 'float'>    2180343      97881    94.999871  [nan, 588180, 1370498, 3843741, 66569, 3843743...
4  profile_submit_count  <class 'numpy.int64'>         0         413         0.000000  [592, 3, 6, 5, 2, 10, 72, 7, 20, 12, 9842, 4, ...
The columns are: Index(['id', 'email', 'primary_phone', 'secondary_phones',
                        'profile_submit_count'],
                        dtype='object')
Rows: 2295101  Columns: 5

```

Key Takeaways: -

- The table has 5 features and 2295101 entries.
- **secondary_phones** feature is almost entirely null which means it may need to be discarded.
- **Primary_phone** has about 35% missing values which needs to be explored further.

iv) Table ct

File size: - 204MB

```

      column      column_type  null_count  unique_count  null_percent  unique_values
0         id  <class 'numpy.int64'>         0        4174013         0.0  [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...
1         cid  <class 'numpy.int64'>         0        1633595         0.0  [1, 5, 6, 7, 8, 9, 10, 11, 13, 16, 17, 18, 20,...
2   timestamp      <class 'str'>         0        3240843         0.0  [2021-04-26 17:21:24, 2021-01-01 05:57:10, 202...
3    amount  <class 'numpy.float64'>         0         1094         0.0  [730.0, 17700.0, 849.0, 1685.0, 2000.0, 1000.0...
4    status      <class 'str'>         0          10         0.0  [PAYMENT_COMPLETED, N, PROCESSED, INITIATED, P...
The columns are: Index(['id', 'cid', 'timestamp', 'amount', 'status'], dtype='object')
Rows: 4174013  Columns: 5

```

Key Takeaways: -

- The table has 5 features and 4174013 entries.
- The data is fairly consistent with respect to null values.
- The status seems like the **purchase status** of potential customers who may have agreed for paid services.

v) Table s

File size: - 851MB

```

column      column_type  null_count  unique_count  null_percent  unique_values
0  uuid      <class 'numpy.int64'>    0      9095602      0.000000  [10058150, 0, 1, 10058153, 26, 27, 28, 29, 30,...
1  phone     <class 'numpy.float64'>    977      3399997      0.010741  [145.0, 607734.0, 607735.0, 607736.0, 607737.0...
2  status    <class 'numpy.int64'>    0      1      0.000000  [1]
3  gender     <class 'str'>      4765      6      0.052388  [Male, Female, nan, MALE, FEMALE, M, F]
4  dob        <class 'str'>      20      36934      0.000220  [00000000, 1967-06-30, 1980-12-16, 1994-07-06,...
5  language   <class 'str'>      398      17      0.004376  [TAM, TEL, ENG, HIN, KAN, ORI, MAL, BEN, MAR, ...
6  email      <class 'numpy.float64'>    733      3259793      0.008059  [0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...
7  report_type <class 'str'>      70      81      0.000770  [LS-MT, LS-MP, LS-CP, LS-WP, LS-PJ, LS-CR, LA-...
8  device     <class 'str'>      187      5      0.002056  [mobile, pc, desktop, PC, MOBILE, nan]
9  log_date   <class 'str'>      0      8285461      0.000000  [2019-02-26 16:07:25, 2019-02-26 16:12:08, 201...
The columns are: Index(['uuid', 'phone', 'status', 'gender', 'dob', 'language', 'email',
                        'report_type', 'device', 'log_date'],
                        dtype='object')
Rows: 9095602      Columns: 10

```

Key Takeaways: -

- The table has 10 features and 9095602 entries.
- The data seems fairly consistent with respect to null values.
- It contains personal information and the type of report the client has opted for.
- The table seems like the record of customers who agreed for paid services.

vi) Table tp

File size: - 107MB

```

column      column_type  null_count  unique_count  null_percent  unique_values
0  ctid      <class 'numpy.int64'>    0      4170263      0.000000  [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...
1  variant    <class 'str'>      0      6      0.000000  [premium, basic, premiumplus, premiumpluscolor...
2  language   <class 'str'>      1824      30      0.043647  [tel, eng, hin, mal, kan, tam, ben, mar, nan, ...
3  status     <class 'float'>      4127969      5      98.778303  [nan, PROCESSED, INITIATED, PDF_ERROR, SUSPEND...
The columns are: Index(['ctid', 'variant', 'language', 'status'], dtype='object')
Rows: 4179024      Columns: 4

```

Key Takeaways: -

- The table has 4 features and 4179024 entries.
- 98.8% of the status data is missing indicating data inconsistencies. Needs to be explored further.

2. Features of each table together

- The table alongside shows all the *features that are common between different combinations* of files.
- This could be useful information especially when we later try to join the tables to create a single Data Frame.

files	Common features
{'a', 'b'}	{'status'}
{'a', 'ct'}	{'status'}
{'a', 's'}	{'status', 'phone'}
{'a', 'tp'}	{'status'}
{'b', 'ct'}	{'status'}
{'b', 's'}	{'log_date', 'status', 'uuid'}
{'tp', 'b'}	{'status'}
{'ct', 'c'}	{'id'}
{'s', 'c'}	{'email'}
{'ct', 's'}	{'status'}
{'tp', 'ct'}	{'status'}
{'tp', 's'}	{'status', 'language'}