INITIAL DATA ANALYSIS (IDA) REPORT

(Submitted by: Sanyam)

Initial Data Analysis(IDA) is a process done on the given dataset to find out some meaningful insights from the data for the purpose of getting a feel of the data for further analysing the data.

For this project we have been provided with 6 datasets:

1. a.csv
2. b.scv
3. c.csv
4. ct.csv
5. s.csv
6. tp.csv

Some initial analysis is done on these datasets in further steps.

"a.csv" dataset

- Contains 998822 rows and 7 columns.

```
             log_time  phone    status  type        product pay_mode  marker
0  2019-05-03 15:57:56    0.0  assigned  1002            NaN      NaN       0
1  2019-05-03 15:57:56    1.0  assigned  1002            NaN      NaN       0
2  2019-05-03 11:34:57    2.0  purchase  1001  In-depth Book    cc-dc       1
3  2019-05-03 11:44:02    3.0  purchase  1001  In-depth Book    cp-nb       1
4  2019-05-03 15:57:56    4.0  assigned  1002            NaN      NaN       0



                  log_time     phone           status  ...  product pay_mode marker
998817  2021-07-31 23:58:52  607730.0         Assigned  ...      NaN      NaN     10
998818  2021-07-31 23:59:06  607730.0  Follow-up later  ...       LI      NaN      1
998819  2021-07-31 23:59:23  607731.0         Assigned  ...      NaN      NaN     10
998820  2021-07-31 23:59:37  607731.0  Follow-up later  ...       LI      NaN      1
998821  2021-07-31 23:59:45  607732.0         Assigned  ...      NaN      NaN     10

[5 rows x 7 columns]
```

- Information regarding each column in the dataset is shown in the following image:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 998822 entries, 0 to 998821
Data columns (total 7 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   log_time  998822 non-null  object
 1   phone     998814 non-null  float64
 2   status    998822 non-null  object
 3   type      998822 non-null  int64
 4   product   330126 non-null  object
 5   pay_mode  223003 non-null  object
 6   marker    998822 non-null  int64
dtypes: float64(1), int64(2), object(4)
memory usage: 53.3+ MB
None
```

- Following are the null values count in each column.

```
log_time         0
phone            8
status           0
type             0
product     668696
pay_mode    775819
marker           0
dtype: int64
```

INFERENCE :-

The dataset contains 7 columns on which valuable data can be extracted::

- Log_time -> contains the timestamp of each call made.
- Phone -> contains the phone numbers of customers.
- Product -> contains the product query type for the customer.
- Pay_mode -> contains the payment mode by customer if he/she purchased the product.

Could not get much information regarding the columns status and marker.

From the given dataset, a customer can be categorised into different types and can be checked for a potential customer from his/her status.

“b.csv” dataset

- Contains 39009332 rows and 5 columns.

```
"C:\Users\sanya\Dekstop\Customer Segmentation Project Packt\venv\Scripts\pytho
   uuid beacon_type  beacon_value             log_date  status
0   0.0   user_stay          26.0  2019-02-26 16:19:08       1
1   0.0   user_stay          32.0  2019-02-26 16:30:08       1
2   1.0   user_stay           1.0  2019-02-26 16:33:39       1
3   2.0   user_stay           1.0  2019-02-26 16:42:00       1
4   3.0   user_stay           1.0  2019-02-26 16:42:00       1



              uuid        beacon_type  ...             log_date status
39009327  10058134.0  pay_button_paypal  ...  2021-07-31 23:59:58      1
39009328  10058148.0          user_stay  ...  2021-07-31 23:59:58      1
39009329  10058129.0          user_stay  ...  2021-07-31 23:59:59      1
39009330  10058118.0          user_stay  ...  2021-07-31 23:59:59      1
39009331  10058149.0          user_stay  ...  2021-07-31 23:59:59      1

[5 rows x 5 columns]
```

- This data consists of beacon data which are uuid, beacon_type, beacon_value, timestamps, status.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39009332 entries, 0 to 39009331
Data columns (total 5 columns):
 #   Column        Dtype
---  ------        -----
 0   uuid          float64
 1   beacon_type   object
 2   beacon_value  float64
 3   log_date      object
 4   status        int64
dtypes: float64(2), int64(1), object(2)
memory usage: 1.5+ GB
None
```

- The null values in each column are as shown.

```
uuid           15
beacon_type     2
beacon_value    2
log_date        0
status          0
dtype: int64
```

INFERENCE :-

From the above dataset, we can extract information regarding some beaconvalues with their log dates and status.

“c.csv” dataset

- Contains 2295101 rows and 5 columns.

```
[2295101 rows x 5 columns]
   id    email  primary_phone secondary_phones  profile_submit_count
0   1   537606           22.0              NaN                   592
1   5  1443908            NaN              NaN                     3
2   6   534973            NaN           588180                     6
3   7  3259797            NaN              NaN                     3
4   8  1701404            NaN              NaN                     5



             id    email  primary_phone secondary_phones  profile_submit_count
2295096  4867864  5554890      4869388.0              NaN                     1
2295097  4867865  5554891      4869389.0              NaN                     1
2295098  4867867  5554892      4869390.0              NaN                     1
2295099  4867879  5554893            NaN              NaN                     1
2295100  4867881  5554894      4869391.0              NaN                     1
```

- Different columns of the dataset and their data types are shown in following image.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2295101 entries, 0 to 2295100
Data columns (total 5 columns):
 #   Column               Dtype
---  ------               -----
 0   id                   int64
 1   email                int64
 2   primary_phone        float64
 3   secondary_phones     object
 4   profile_submit_count  int64
dtypes: float64(1), int64(3), object(1)
memory usage: 87.6+ MB
None
```

- Columns of the dataset contain null values as follows.

```
id                          0
email                       0
primary_phone          793012
secondary_phones      2180343
profile_submit_count        0
dtype: int64
```

INFERENCE :-

This dataset mainly consists of users' data with user_id and their email id's, primary and secondary phone numbers.
Columns email, primary_phone, secondary_phone contain many nan values.

"ct.csv" dataset

- Contains 4174013 rows x 5 columns.

```
[4174013 rows x 5 columns]
   id  cid            timestamp   amount             status
0   4    1  2021-04-26 17:21:24    730.0  PAYMENT_COMPLETED
1   5    5  2021-01-01 05:57:10  17700.0  PAYMENT_COMPLETED
2   6    6  2021-01-01 10:33:22    849.0  PAYMENT_COMPLETED
3   7    7  2021-01-02 06:10:53   1685.0  PAYMENT_COMPLETED
4   8    8  2021-01-02 08:32:43   2000.0  PAYMENT_COMPLETED



             id      cid            timestamp  amount status
4174008  4174055  4867893  2021-08-25 20:10:18     0.0      N
4174009  4174056  2197111  2021-08-25 20:10:18     0.0      N
4174010  4174057  3423129  2021-08-25 20:10:23     0.0      N
4174011  4174058  4867896  2021-08-25 20:10:26     0.0      N
4174012  4174059   653124  2021-08-25 20:10:50     0.0      N
```

- Columns information is shown in following image.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4174013 entries, 0 to 4174012
Data columns (total 5 columns):
 #   Column     Dtype
---  ------     -----
 0   id         int64
 1   cid        int64
 2   timestamp  object
 3   amount     float64
 4   status     object
dtypes: float64(1), int64(2), object(2)
memory usage: 159.2+ MB
None
```

```
id           0
cid          0
timestamp    0
amount       0
status       0
dtype: int64
```

INFERENCE :-

This dataset contains information regarding the transactions made and their amount and timestamps in another column.
There are no null values in any column.

“s.csv” dataset

- Contains 9095602 rows and 10 columns.

```
(9095602, 10)
       uuid     phone  status  ... report_type  device            log_date
0  10058150     145.0       1  ...       LS-MT  mobile  2019-02-26 16:07:25
1         0     145.0       1  ...       LS-MT  mobile  2019-02-26 16:12:08
2         1     145.0       1  ...       LS-MT  mobile  2019-02-26 16:33:00
3  10058153  607734.0       1  ...       LS-MP  mobile  2019-02-26 16:44:19
4        26  607735.0       1  ...       LS-MT  mobile  2019-02-26 16:44:32

[5 rows x 10 columns]



            uuid      phone  status  ... report_type  device            log_date
9095597  19153747  4007596.0       1  ...       LS-MT  mobile  2021-07-31 23:59:45
9095598  19153748   607007.0       1  ...       LS-MT  mobile  2021-07-31 23:59:49
9095599  19153749  4007729.0       1  ...       LS-MT  mobile  2021-07-31 23:59:51
9095600  19153750  4007717.0       1  ...       LS-MT  mobile  2021-07-31 23:59:54
9095601  19153751  4007730.0       1  ...       LS-MT  mobile  2021-07-31 23:59:54

[5 rows x 10 columns]
```

- Column information are given in following images.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9095602 entries, 0 to 9095601
Data columns (total 10 columns):
 #   Column       Dtype
---  ------       -----
 0   uuid         int64
 1   phone        float64
 2   status       int64
 3   gender       object
 4   dob          object
 5   language     object
 6   email        float64
 7   report_type  object
 8   device       object
 9   log_date     object
dtypes: float64(2), int64(2), object(6)
memory usage: 693.9+ MB
None
```

```
uuid              0
phone           977
status            0
gender         4765
dob              20
language        398
email           733
report_type      70
device          187
log_date          0
dtype: int64
```

INFERENCE :-

This dataset contains customer's information such as id, phone number, gender, dob, email id, his device, language of conversation, etc.

Could not get what does report_type, status columns show.

“t.csv” dataset

- Contains 4179024 rows x 4 columns.

```
[4179024 rows x 4 columns]
   ctid  variant language status
0     4  premium      tel    NaN
1     5  premium      eng    NaN
2     6  premium      eng    NaN
3     7  premium      eng    NaN
4     8  premium      eng    NaN



            ctid variant language status
4179019  4174090   basic      eng    NaN
4179020  4174091   basic      eng    NaN
4179021  4174092   basic      eng    NaN
4179022  4174093   basic      tam    NaN
4179023  4174094   basic      eng    NaN
```

- Column information is given in next images

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4179024 entries, 0 to 4179023
Data columns (total 4 columns):
 #   Column    Dtype
---  ------    -----
 0   ctid      int64
 1   variant   object
 2   language  object
 3   status    object
dtypes: int64(1), object(3)
memory usage: 127.5+ MB
None
```

```
ctid              0
variant           0
language       1824
status      4127969
dtype: int64
```

INFERENCE :-

Contains customer information regarding his language and variant of plan he/she has opted for.