Initial Data Analysis (IDA) Report

Initial Data Analysis is the process of having a basic understanding of the datasets.It helps in getting an idea of the data being handled.Once we get this right, it forms the foundation to various other analyses done during later stages.

Number of datasets = 6
Datasets include:

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

Note:

(**.csv**) is a file format in which data is stored in the form of **Comma Separated Values**. It typically contains unstructured data.Mainly used for data sharing in various organizations.

My approach:

I have loaded the above mentioned Datasets and used various **Pandas Functions** for analysing the rows, columns and datatypes.

I have attached the screenshots of the same here.

1) **a.csv**: This dataset has 998822 rows, 7 columns and the data types include int, float and objects.

```
>>> data=pd.read_csv('a.csv')
>>> count_row=data.shape[0]
>>> print(count_row)
998822
>>> count_col=data.shape[1]
>>> print(count_col)
7
>>> print(data.dtypes)
log_time     object
phone       float64
status       object
type          int64
product      object
pay_mode     object
marker        int64
dtype: object
```

2) **b.csv** : This dataset has 39009332 rows, 5 columns and the data types include int, float and objects.

```
>>> data=pd.read_csv('b.csv')
>>> count_row=data.shape[0]
>>> print(count_row)
39009332
>>> count_col=data.shape[1]
>>> print(count_col)
5
>>> print(data.dtypes)
uuid            float64
beacon_type      object
beacon_value    float64
log_date         object
status            int64
dtype: object
```

3) **c.csv**:This dataset has 2295101 rows, 5 columns and the data types include int, float and objects.

```
>>> data=pd.read_csv('c.csv')
>>> count_row=data.shape[0]
>>> print(count_row)
2295101
>>> count_col=data.shape[1]
>>> print(count_col)
5
>>> print(data.dtypes)
id                       int64
email                    int64
primary_phone          float64
secondary_phones        object
profile_submit_count     int64
dtype: object
```

4) **ct.csv**:This dataset has 4174013 rows, 5 columns and the data types include int, float and objects.

```
>>> data=pd.read_csv('ct.csv')
>>> count_row=data.shape[0]
>>> print(count_row)
4174013
>>> count_col=data.shape[1]
>>> print(count_col)
5
>>> print(data.dtypes)
id             int64
cid            int64
timestamp     object
amount       float64
status        object
dtype: object
```

5) **tp.csv** :This dataset has 4179024 rows, 4 columns and the data types include int and objects.

```
>>> data=pd.read_csv('tp.csv')
>>> count_row=data.shape[0]
>>> print(count_row)
4179024
>>> count_col=data.shape[1]
>>> print(count_col)
4
>>> print(data.dtypes)
ctid         int64
variant     object
language    object
status      object
dtype: object
```

6) **s.csv** :This dataset has 9095602 rows, 10 columns and the data types include int, float and objects.

```
>>> data=pd.read_csv('s.csv')
>>> count_row=data.shape[0]
>>> print(count_row)
9095602
>>> count_col=data.shape[1]
>>> print(count_col)
10
>>> print(data.dtypes)
uuid             int64
phone          float64
status           int64
gender          object
dob             object
language        object
email          float64
report_type     object
device          object
log_date        object
dtype: object
```

Ravindra.