# EDA Report

In Exploratory Data Analysis, we perform a detailed analysis so that any underlying patterns could be deduced.However, during EDA there are no modifications done to the Datasets.

EDA is the step that precedes Data Preprocessing.

I have used the functions **describe(), isnull(), unique(), nunique(), shape** for **non-graphical** analysis and **countplot(), hist()** for **graphical** analysis.All these functions have been applied to all the datasets individually and the **results** have been attached in the form of **screenshots** along with some **text**.

Note:

1) **Count plots** are used for graphical analysis of **categorical** data.
2) **Histograms** are mostly used for graphical analysis of **numerical** data.

---

**DATASET** **ANALYSIS**

**a.csv**: The dataset consists of 7 variables.These are categorized into numerical and categorical based on the data types.

**Numerical** variables include the columns: log_time, phone, type and marker.**Categorical** variables are: status, product, pay_mode.

The statistics of dataset “a.csv” are displayed using the **describe()** method.Screenshot of the same.

```
>>> data=pd.read_csv('a.csv')
>>> data.shape
(998822, 7)
>>> data.describe()
                phone           type         marker
count  998814.000000  998822.000000  998822.000000
mean   260397.183060    1244.267067       1.623012
std    187624.002369     457.458217       2.865491
min         0.000000    1001.000000     -99.000000
25%     89754.250000    1001.000000       1.000000
50%    244574.500000    1002.000000       1.000000
75%    423747.000000    1002.000000       1.000000
max    607732.000000    2209.000000      10.000000
```

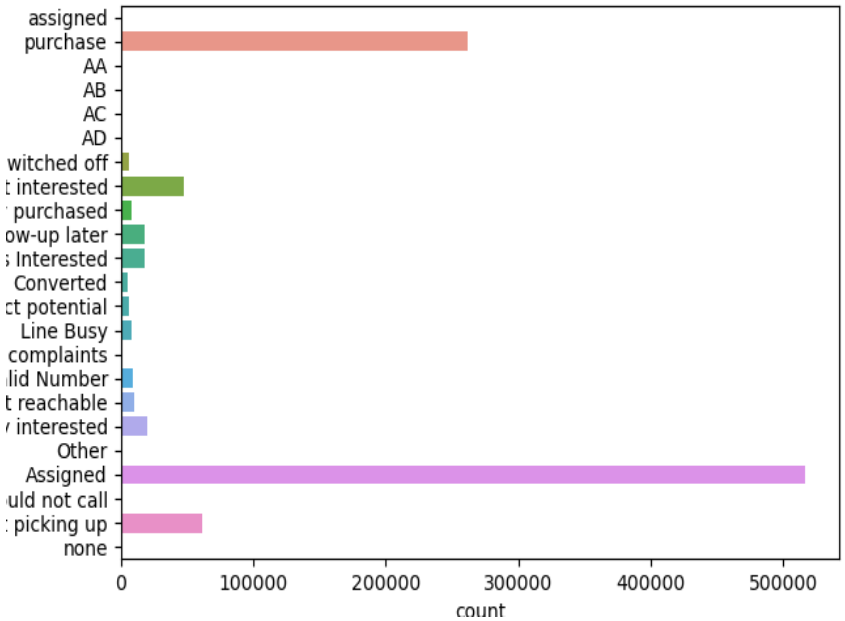
Null values are displayed using the **isnull()** method as shown below.

```
>>> data.apply(lambda x: sum(x.isnull()),axis=0)
log_time         0
phone            8
status           0
type             0
product     668696
pay_mode    775819
marker           0
dtype: int64
```

The functions **nunique()** and **unique()** output the number of unique values and array of unique values respectively.These operations are performed on individual **columns** of a given dataset.

```
>>> data['status'].nunique()
23
>>> data['status'].unique()
array(['assigned', 'purchase', 'AA', 'AB', 'AC', 'AD', 'Switched off',
       'Not interested', 'Already purchased', 'Follow-up later',
       'User is Interested', 'Converted', 'New product potential',
       'Line Busy', 'Has complaints', 'Invalid Number', 'Not reachable',
       'Partially interested', 'Other', 'Assigned', 'Could not call',
       'Not picking up', 'none'], dtype=object)
```

The image below describes the **countplot** plotted with the "**status**" variable on y axis and readings on x axis.



**b.csv**: The dataset consists of 5 variables.These are categorized into numerical and categorical based on the data types.

**Numerical** variables include the columns: uuid, beacon_value and status.**Categorical** variables are: log_date and beacon_type.

The statistics of dataset “b.csv” are displayed using the **describe()** method.Screenshot of the same.

```
>>> df=pd.read_csv('b.csv')
>>> df.shape
(39009332, 5)
>>> df.describe()
               uuid  beacon_value      status
count  3.900932e+07  3.900933e+07  39009332.0
mean   4.922558e+06  5.610764e+00         1.0
std    2.985628e+06  1.102048e+01         0.0
min    0.000000e+00  1.000000e+00         1.0
25%    2.232751e+06  1.000000e+00         1.0
50%    4.901919e+06  2.000000e+00         1.0
75%    7.526520e+06  4.000000e+00         1.0
max    1.005815e+07  9.990000e+02         1.0
```

Null values are displayed using the **isnull()** method as shown below.

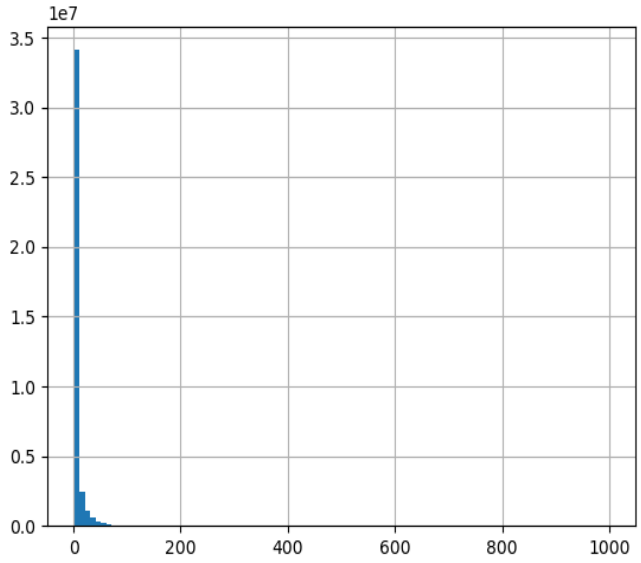
```
>>> df.apply(lambda x : sum(x.isnull()),axis=0)
uuid           15
beacon_type     2
beacon_value    2
log_date        0
status          0
dtype: int64
```

The **nunique()** and **unique()** output:

```
>>> df['uuid'].nunique()
10058149
>>> df['uuid'].unique()
array([0.0000000e+00, 1.0000000e+00, 2.0000000e+00, ..., 1.0058147e+07,
       1.0058148e+07, 1.0058149e+07])
```

As for the graphical uni-variate analysis of “b.csv”, the plot was slightly overwhelming due to the size of our dataset.We have considered  the “**beacon_value**” for the **histogram**.

```
>>> df['beacon_value'].hist(bins=300)
<AxesSubplot:>
>>> plt.show()
>>> df['beacon_value'].hist(bins=100)
<AxesSubplot:>
>>> plt.show()
```



**c.csv**: The dataset consists of 5 variables.These are categorized into numerical and categorical based on the data types.

**Numerical** variables include the columns: id, email, primary_phone, secondary_phones and profile_submit_count (all of them).

The statistics of dataset “c.csv” are displayed using the **describe()** method.Screenshot of the same.

```
>>> df=pd.read_csv('c.csv')
>>> df.shape
(2295101, 5)
>>> df.describe()
                 id         email  primary_phone  profile_submit_count
count  2.295101e+06  2.295101e+06   1.502089e+06          2.295101e+06
mean   2.594772e+06  3.393918e+06   2.920364e+06          2.927417e+00
std    1.435092e+06  1.281969e+06   1.560694e+06          1.084325e+01
min    1.000000e+00  0.000000e+00   2.000000e+00          1.000000e+00
25%    1.285110e+06  2.580305e+06   1.799900e+06          2.000000e+00
50%    2.830071e+06  3.156446e+06   3.497062e+06          2.000000e+00
75%    3.868155e+06  4.638336e+06   3.926352e+06          3.000000e+00
max    4.867881e+06  5.554894e+06   4.869391e+06          9.842000e+03
```

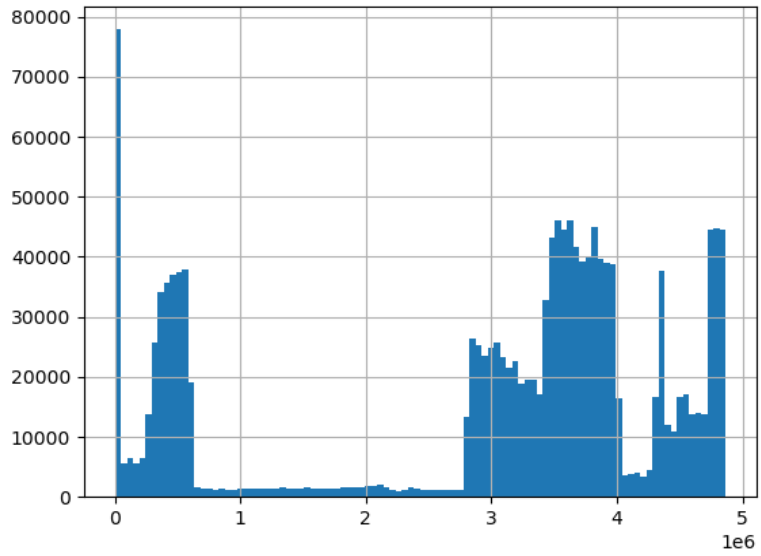
Null values are displayed using the **isnull()** method as shown below.

```
>>> df.apply(lambda x : sum(x.isnull()),axis=0)
id                         0
email                      0
primary_phone         793012
secondary_phones     2180343
profile_submit_count       0
dtype: int64
```

The functions **nunique()** and **unique()** output:

```
>>> df['email'].nunique()
2295101
>>> df['email'].unique()
array([ 537606, 1443908,  534973, ..., 5554892, 5554893, 5554894])
```

We have considered the "**profile_submit_count**" variable for this **histogram**.



**ct.csv**: The dataset consists of 5 variables.These are categorized into numerical and categorical based on the data types.

**Numerical** variables include the columns: cid, id and amount.**Categorical** variables are: status and timestamp.

The statistics of dataset “ct.csv” are displayed using the **describe()** method.Screenshot of the same.

```
>>> df=pd.read_csv('ct.csv')
>>> df.shape
(4174013, 5)
>>> df.describe()
                 id           cid        amount
count  4.174013e+06  4.174013e+06  4.174013e+06
mean   2.087048e+06  1.767756e+06  3.592679e+01
std    1.204940e+06  1.375565e+06  3.090433e+02
min    4.000000e+00  1.000000e+00  0.000000e+00
25%    1.043544e+06  5.536010e+05  0.000000e+00
50%    2.087049e+06  1.517702e+06  0.000000e+00
75%    3.130556e+06  2.748210e+06  0.000000e+00
max    4.174059e+06  4.867896e+06  1.301137e+05
```

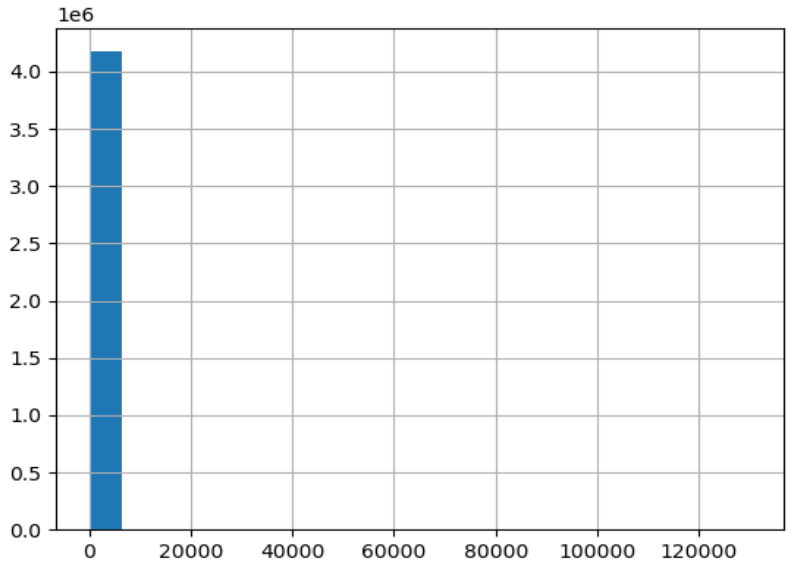
Null values are displayed using the **isnull()** method as shown below.

```
>>> df.apply(lambda x : sum(x.isnull()),axis=0)
id           0
cid          0
timestamp    0
amount       0
status       0
dtype: int64
```

The functions **nunique()** and **unique()** output:

```
>>> df['amount'].nunique()
1094
>>> df['amount'].unique()
array([  730.  , 17700.  ,   849.  , ...,   159.2 , 10389.  ,   179.24])
```

We have considered the “**amount**” variable for this **histogram**.



**s.csv**: The dataset consists of 10 variables.These are categorized into numerical and categorical based on the data types.

**Numerical** variables include the columns: uuid, log_date, phone and status.**Categorical** variables are: gender, dob, language, email, report_type and device.

The statistics of dataset “s.csv” are displayed using the **describe()** method.Screenshot of the same below.

```
>>> df=pd.read_csv('s.csv')
>>> df.describe()
               uuid         phone     status         email
count  9.095602e+06  9.094625e+06  9095602.0  9.094869e+06
mean   6.474746e+06  1.415220e+06        1.0  1.240107e+06
std    4.422480e+06  1.215990e+06        0.0  9.714803e+05
min    0.000000e+00  0.000000e+00        1.0  0.000000e+00
25%    2.790156e+06  2.984330e+05        1.0  3.941350e+05
50%    6.237638e+06  1.125327e+06        1.0  1.030454e+06
75%    9.048801e+06  2.339591e+06        1.0  2.018648e+06
max    1.915375e+07  4.007730e+06        1.0  3.259793e+06
>>> df.shape
(9095602, 10)
```

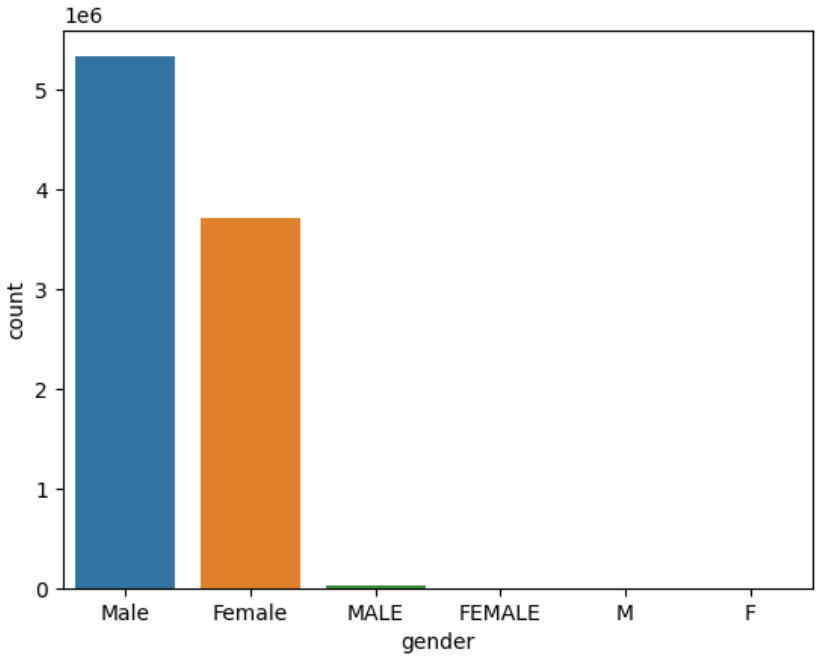
Null values are displayed using the **isnull()** method as shown below.

```
>>> df.apply(lambda x : sum(x.isnull()),axis=0)
uuid             0
phone          977
status           0
gender        4765
dob             20
language       398
email          733
report_type     70
device         187
log_date         0
dtype: int64
```

The functions **nunique()** and **unique()** output:

```
>>> df['status'].value_counts()
1    9095602
Name: status, dtype: int64
>>> df['status'].nunique()
1
>>> df['status'].unique()
array([1])
```

For the **countplot** of s.csv, we have taken the "**gender**" as the x-variable and readings on the y-axis as shown here.



**tp.csv**: The dataset consists of 4 variables.These are categorized into numerical and categorical based on the data types.

**Numerical** variables include the column: ctid.**Categorical** variables are: status, variant and language.

The statistics of dataset “tp.csv” are displayed using the **describe()** method.Screenshot of the same below.

```
>>> df=pd.read_csv('tp.csv')
>>> df.shape
(4179024, 4)
>>> df.describe()
               ctid
count  4.179024e+06
mean   2.086676e+06
std    1.204867e+06
min    4.000000e+00
25%    1.043103e+06
50%    2.086494e+06
75%    3.129968e+06
max    4.174094e+06
```

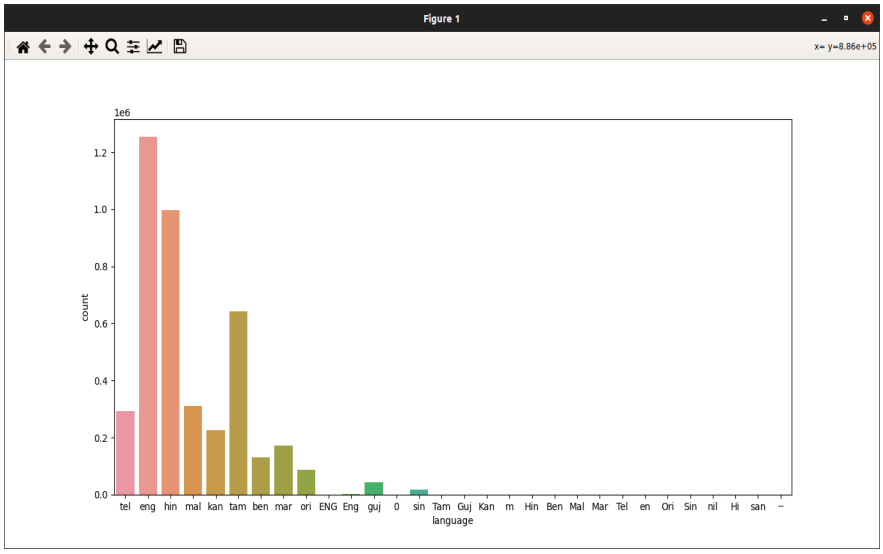
Null values are displayed using the **isnull()** method as shown below.

```
>>> df.apply(lambda x : sum(x.isnull()),axis=0)
ctid              0
variant           0
language       1824
status      4127969
dtype: int64
```

The functions **nunique()** and **unique()** output:

```
>>> df['language'].nunique()
30
>>> df['language'].unique()
array(['tel', 'eng', 'hin', 'mal', 'kan', 'tam', 'ben', 'mar', nan, 'ori',
       'ENG', 'Eng', 'guj', '0', 'sin', 'Tam', 'Guj', 'Kan', 'm', 'Hin',
       'Ben', 'Mal', 'Mar', 'Tel', 'en', 'Ori', 'Sin', 'nil', 'Hi', 'san',
       '--'], dtype=object)
```

For the **countplot** of tp.csv, we have taken the "**language**" as the x-variable and readings on the y-axis as shown below.




**Ravindra.**