Exploratory Data Analysis

Submitted by: Reet

Index:

About exploratory data analysis

Steps for Exploratory data analysis

1. Variable Identification: Numerical/Categorical
2. Summary of Numerical Variables
3. Statistical Summary of dataset
4. Graphical analysis
5. Null values and unique variables
6. Summary or conclusion from the data

Analysing the csv files

A. a.csv

- Variable Identification: log_time, phone, status, type, product, pay_mode, marker
  Categorical: status, product, pay_mode
  Numerical: type, marker
  Neither numerical or categorical: phone, log_time

- Summary of Numerical Variables: to create an efficient algorithm, we should try to convert all categorical data i.e. object data type. I converted log_time into date_time.

```
log_time     object
phone       float64
status       object
type          int64
product      object
pay_mode     object
marker        int64
dtype: object
```

- Statistical summary of dataset:

```
(998822, 7)
                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
```

```
In [36]: data.nunique()
Out[36]: log_time    559772
         phone       607732
         status          23
         type            21
         product        125
         pay_mode        62
         marker           6
         dtype: int64

In [37]: data.isnull().sum()
Out[37]: log_time         0
         phone            8
         status           0
         type             0
         product     668696
         pay_mode    775819
         marker           0
         dtype: int64
```

null values and unique values

- Summary and conclusion from data:
    1. What exactly is in the pay_mode column?
    2. Not interested users can be removed from the database?
    3. Already purchased can also be a new product potential
    4. Type has no further requirements
    5. Status need some modifications to simplify the unique values
    6. Not interested customers should be dropped

B) b.csv

- Variable Identification:

Numerical: uuid, beacon_value, status
Categorical
Neither of them: beacon_type, log_date

- Summary of numeric variables:

```
[5 rows x 5 columns]
(39009332, 5)
Index(['uuid', 'beacon_type', 'beacon_value', 'log_date', 'status'], dtype='object')
              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
```

Size and statistics of the data

Summary and conclusion from the data:

1. Status column has no null values, any point in keeping it?
2. There are many unique values in beacon_type, what does each of them signify?

C) c.csv

Variable Identification:
Numerical: id, profile_submit_count, submit_count
Neither numerical or categorical: email, primary_phone, secondary_phone

Many null values in phone numbers
Null values

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

Statiscal analysis of data

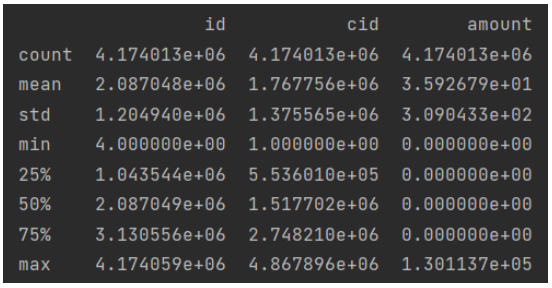
```
                 id         email  primary_phone  profile_submit_count
count  2.295101e+06  2.295101e+06   1.502089e+06          2.295101e+06
mean   2.594772e+06  3.393918e+06   2.920364e+06          2.927417e+00
std    1.435092e+06  1.281969e+06   1.560694e+06          1.084325e+01
min    1.000000e+00  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
```

D) ct.csv

Variable identification:
Numerical: amount, id, cid
Categorical: status
Time and date: timestamp

We can simply drop the rows with non confirmed payment status and and then drop the status column itself.
We can also visualise this dataset better if we combine it with a.csv
The amount paid is also not relevant to our work, so we can also drop the column

Statistical summary of the data

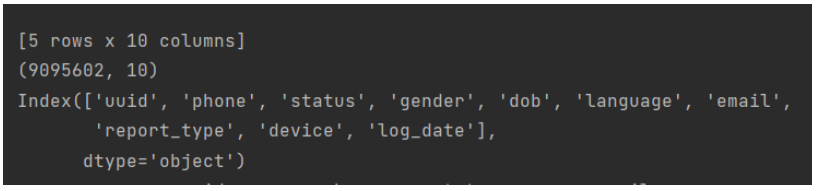


	id	cid	amount
count	4.174013e+06	4.174013e+06	4.174013e+06
mean	2.087048e+06	1.767756e+06	3.592679e+01
std	1.204940e+06	1.375565e+06	3.090433e+02
min	4.000000e+00	1.000000e+00	0.000000e+00
25%	1.043544e+06	5.536010e+05	0.000000e+00
50%	2.087049e+06	1.517702e+06	0.000000e+00
75%	3.130556e+06	2.748210e+06	0.000000e+00
max	4.174059e+06	4.867896e+06	1.301137e+05

E) s.csv

Variable Identification:
Categorical: gender, language, report_type, device
Numerical: uuid, status
Neither: phone, dob, email, log_date

Size and variable type



```
[5 rows x 10 columns]
(9095602, 10)
Index(['uuid', 'phone', 'status', 'gender', 'dob', 'language', 'email',
       'report_type', 'device', 'log_date'],
      dtype='object')
```

Required to plot a histogram to identify whether male or female buys the product more.
And which language is used the most

.

	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

Statiscal analysis of data

```
uuid           9095602
phone          3399997
status               1
gender               6
dob              36934
language            17
email          3259793
report_type         81
device               5
log_date       8285461
dtype: int64
uuid              0
phone           977
status            0
gender         4765
dob              20
language        398
email           733
report_type      70
device          187
log_date          0
dtype: int64

Process finished with exit code 0
```

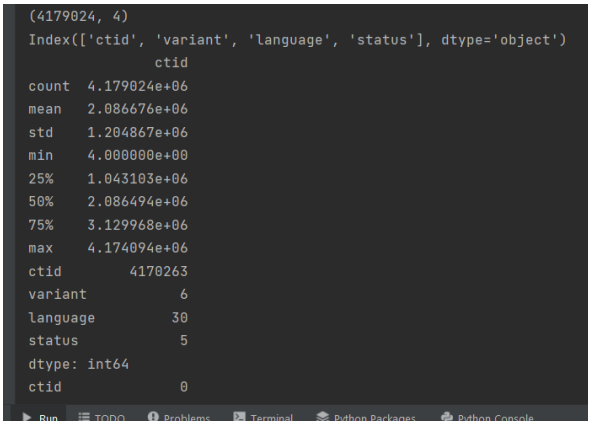
Unique and null values

e) tp.csv

Variable identification:
Numerical: ctid, status
Categorical: variant, language

Need to plot a graph to check which variant is subscribed more

T and ct can be combined for better results



size and statistical analysis of data