EXPLORATORY DATA ANALYSIS (EDA) REPORT

Submitted by: Sanyam Goel

Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

In our case, I have found counts of different categorical data and also been able to make some visual data analysis using pandas and matplotlib.

Dataset “a.csv”

This dataset contains 7 variables(features) which are of 2 types, i.e., categorical data and numeric data.

- Categorical Data - status, pay_mode, product, marker, type
- Numeric Data - log_time, phone

Unique values in ‘status’ columns are:

```
['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']
```

```
Assigned                 516435
purchase                 261558
Not picking up            61454
Not interested            47341
Partially interested      19780
Follow-up later           18244
User is Interested        17756
Not reachable              9774
Invalid Number             9327
Line Busy                  8628
Already purchased          8123
Switched off               6316
New product potential      5938
Converted                  5579
Other                      1655
Could not call              632
assigned                    252
Has complaints               22
AC                            2
AB                            2
AA                            2
AD                            1
none                          1
Name: status, dtype: int64
```

Unique values in ‘type” columns are:

```
[1002 1001 1003 1005 1004 1006 2005 2209 2001 2102 2103 2104 2206 2207
 2208 2010 2011 2012 2013 2105 2106]
```

```
1002    516685
1001    261752
2106     61455
2005     47341
2208     19780
2206     18242
2207     17753
2102      9774
2001      9327
2103      8627
2010      7927
2104      6315
2209      5937
2011      5578
2013      1656
2105       632
2012        21
1005         8
1003         7
1004         4
1006         1
Name: type, dtype: int64
```

Unique values in 'pay_mode' columns are:

```
[nan 'cc-dc' 'cp-nb' 'PayPal' 'ptm' 'upi' 'paypal' 'wallet' 'UPI' 'WL'
 'PTM' 'gbp' 'RY' 'paycancel_cc-dc_HDFC' 'rupay' 'RU' 'MN2942952864BEB'
 'AVAC125921Dfc' 'MN294346739b3aa' 'INVC1557ZTXO' 'USD' 'usd'
 'MN294496345f5db' 'gpay' 'MN294568094c4ab' 'rz' 'MN294589903D2FB'
 'MN29467013811BB' 'MN294695472D3CD' 'MN2946970583D2d' 'AVFD1359951acb5'
 'MN294783007CC1b' 'no_mode' 'jpay' 'MN294802793d1F3' 'MN294811647ac31'
 'MN294854885AC1c' 'GPAY' 'MN294955924cd1A' 'paytm' 'MN295018279AA1e'
 'CP-NB' 'CC-DC' 'paycancel_WL_PAYU' 'PhonePe' 'PHONEPE' 'CC' 'RZ' 'bank'
 'paycancel_CC_HDFC' 'invoice' 'NB' 'paycancel_NB_INST'
 'paycancel_NB_HDFC' 'paycancel_CC_INST' 'paycancel_NB_CITRUS' 'PTMMini'
 'paycancel_CC_RZ' 'paycancel_CC_PAYU' 'paycancel_NB_RZ'
 'paycancel_NB_PAYU' 'paycancel_UPI_RZ' 'PYPL']
```

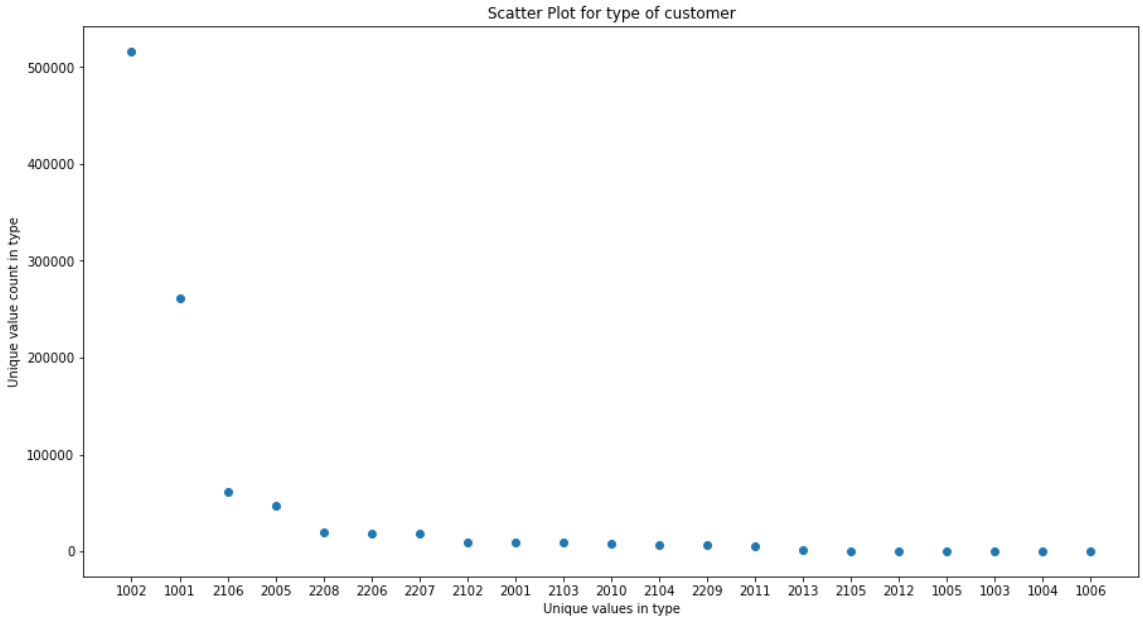
```
cc-dc              79030
cp-nb              57755
gpay               18825
UPI                12459
ptm                12332
                   ...
MN294955924cd1A        1
MN2942952864BEB        1
MN295018279AA1e        1
CP-NB                  1
PYPL                   1
Name: pay_mode, Length: 62, dtype: int64
```

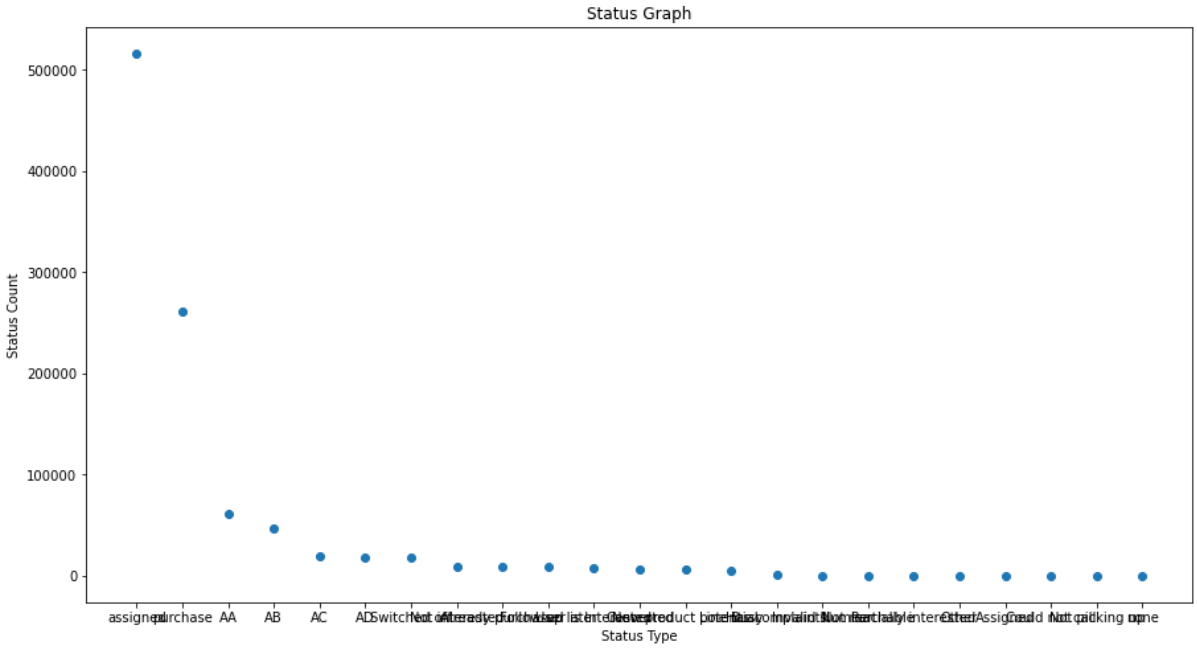
Unique values in 'marker' columns are:

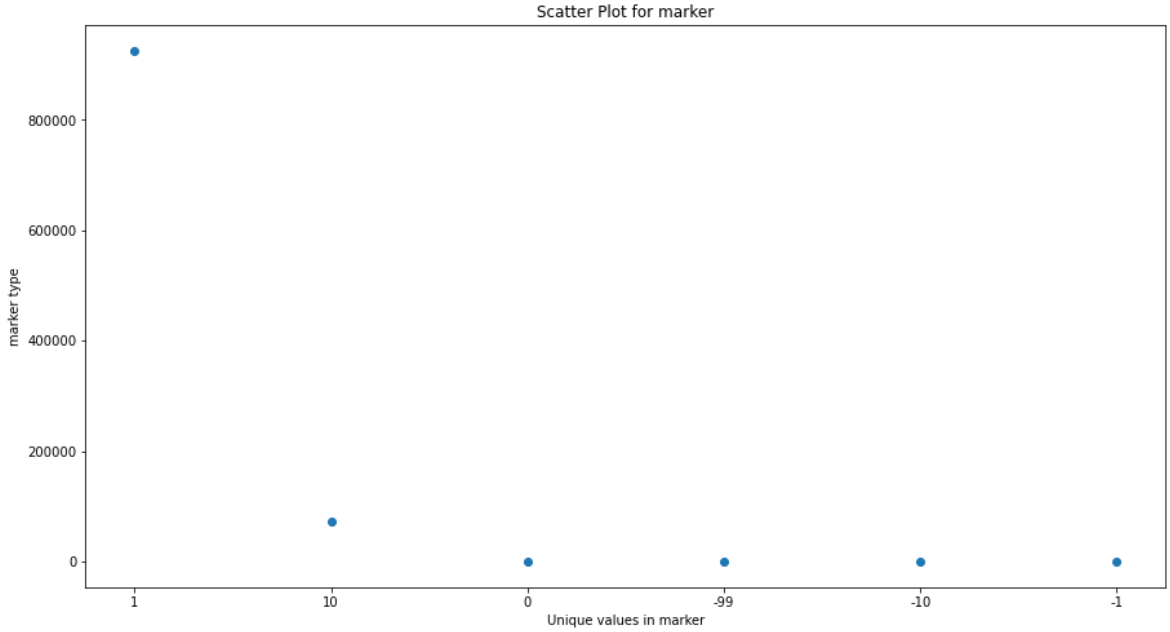
```
[  0   1 -99  -1 -10  10]

 1     925749
 10     72274
 0        488
-99       273
-10        36
-1          2
Name: marker, dtype: int64
```

From the above dataset, some of the visualizations that can be made are:-







Dataset “b.csv”

This dataset contains 5 variables, i.e., uuid, beacon_type, beacon_value, log_data, status.

- Categorical Data - beacon_type, status
- Numeric Data - uuid, log_date, beacon_value

The ‘status’ column contains just one value of 1 for all rows without any null values.

The data summary can be described using describe() to get some of the understanding from data.

```
In [114]: df.describe()
```

Out[114]:

	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

The unique values present in the ‘beacon_type’ column are:

```
array(['user_stay', 'bottom_banner', 'buy_button_FH', 'pay_button_cc-dc',
       'pay_button_cp-nb', 'pay_button_paypal', 'buy_button_PP',
       'masked_content', 'buy_button_top', 'buy_button_autoPopupE',
       'buy_button_autopopup90p', 'pay_button_upi', 'pay_button_ptm',
       'pay_button_wallet', 'buy_button_InReportMaskToast',
       'buy_button_AdBannerRight', 'purchased in-depth,AVOL3218525210124',
       'pay_button_gbp', 'buy_button_CR', 'pay_button_undefined',
       'pay_button_PP', 'pay_fail', 'buy_button_undefined',
       'pay_button_usd', 'pay_button_gpay', 'mask_area', 'top_banner',
       'pay_button_rz', 'pay_button_PayPal', 'pay_button_RZ',
       'pay_button_PTM', 'pay_button_IM', 'pay_button_WL',
       'pay_button_NB', 'pay_button_CC', 'pay_button_PYPL',
       'pay_button_RU', 'pay_button_RY', 'pay_button_UPI', 'pay_button_',
       'pay_button_ebs', 'pay_button_EBSG', 'pay_button_jpay',
       'pay_button_GPAY', 'buy_button_WL', 'pay_button_PAYU',
       'pay_button_rupay', 'pay_button_avupi', 'pay_button_juspay',
       'pay_button_ebsg', 'pay_button_CTRS', 'pay_button_hdfc',
       'pay_button_USD', 'pay_button_paytm', 'pay_button_EUR',
       'pay_button_PTMMini', 'social_proof_banner', 'pay_button_phonepe',
       'pay_button_PhonePe', 'pay_button_HDFC', 'pay_button_EBS',
       'pay_button_payu', '6', nan, 'pay_button_NA', 'pay_button_invoice',
       'pay_button_bank'], dtype=object)
```

Dataset “c.csv”

This dataset consists of 5 variables only in the form of numerical data.
Different columns of this dataset contain information regarding users about their email, phone numbers, etc.

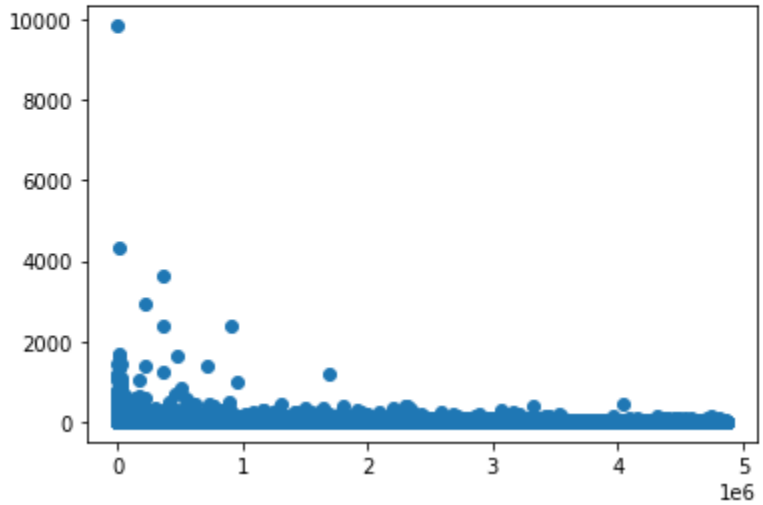
Null values in the data are:

```
In [35]: df.isnull().sum()

Out[35]: id                          0
         email                       0
         primary_phone          793012
         secondary_phones      2180343
         profile_submit_count        0
         dtype: int64
```

Visualizations that can be made on this dataset are between user id (x axia) and their profile_submit_count (y-axis).

```
In [34]: l=df['id'].to_list()
         l1=df['profile_submit_count'].to_list()
         plt.scatter(l, l1)
         plt.show()
```



Dataset “ct.csv”

This dataset contains 5 variables.

- Categorical data - status
- Numerical data - id, cid, timestamp, amount

The ‘status’ column contains status of the payment process as:

```
df['status'].unique()
```

```
array(['PAYMENT_COMPLETED', 'N', 'PROCESSED', 'INITIATED', 'PDF_ERROR',
       'SUSPENDED', 'ROLLED_BACK', 'TOPROCESS', 'PAYMENT_FAILED', 'Y'],
      dtype=object)
```

Counts of different categories in ‘status’ column are:

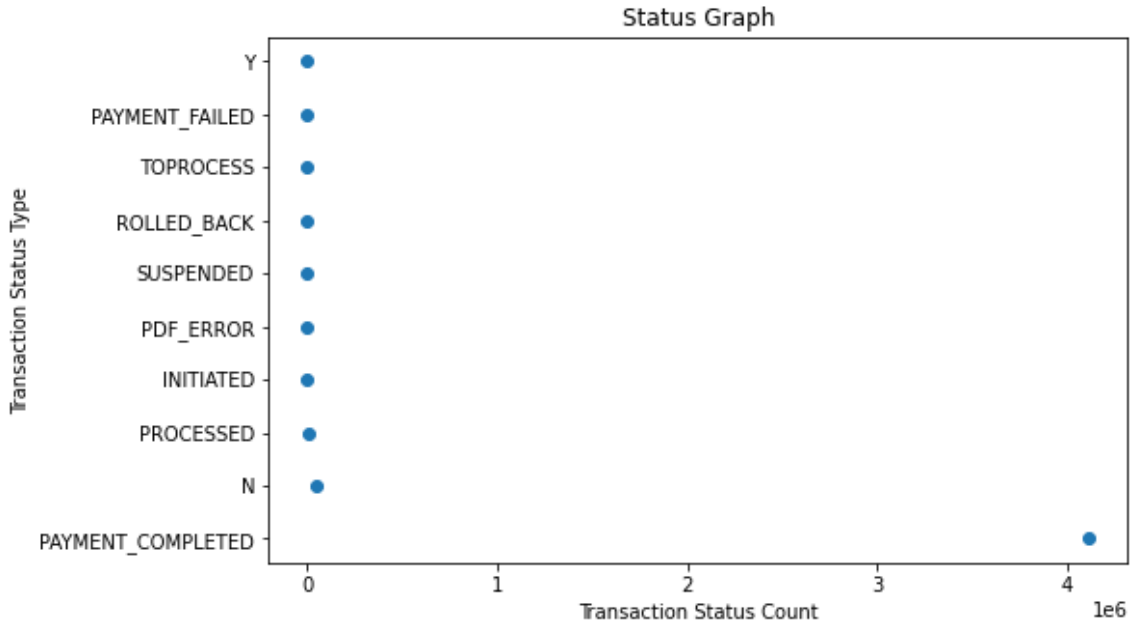
```
N                  4115487
PROCESSED            48328
PAYMENT_COMPLETED     9226
SUSPENDED              526
INITIATED              279
ROLLED_BACK             85
PDF_ERROR               55
PAYMENT_FAILED          16
TOPROCESS               10
Y                        1
Name: status, dtype: int64
```

Statistical Summary about the data is:

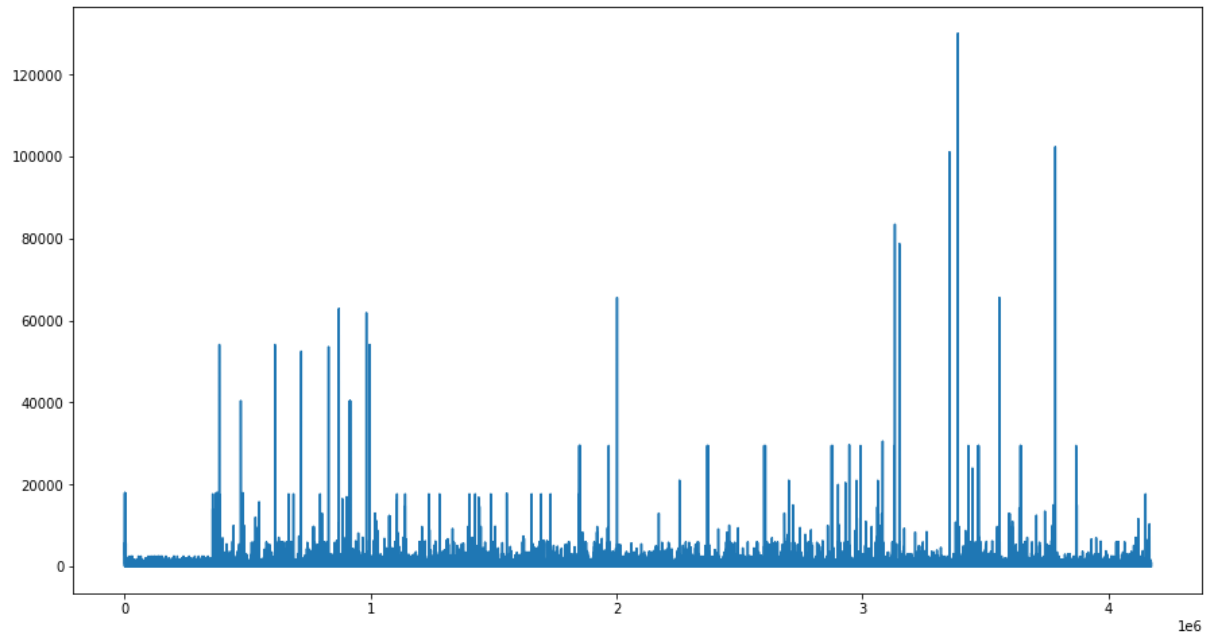
	id	cid	amount
count	4.174013e+06	4.174013e+06	4.174013e+06
mean	2.087048e+06	1.767756e+06	3.592679e+01
std	1.204940e+06	1.375565e+06	3.090433e+02
min	4.000000e+00	1.000000e+00	0.000000e+00
25%	1.043544e+06	5.536010e+05	0.000000e+00
50%	2.087049e+06	1.517702e+06	0.000000e+00
75%	3.130556e+06	2.748210e+06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

Visualizations that can be made from this data are:

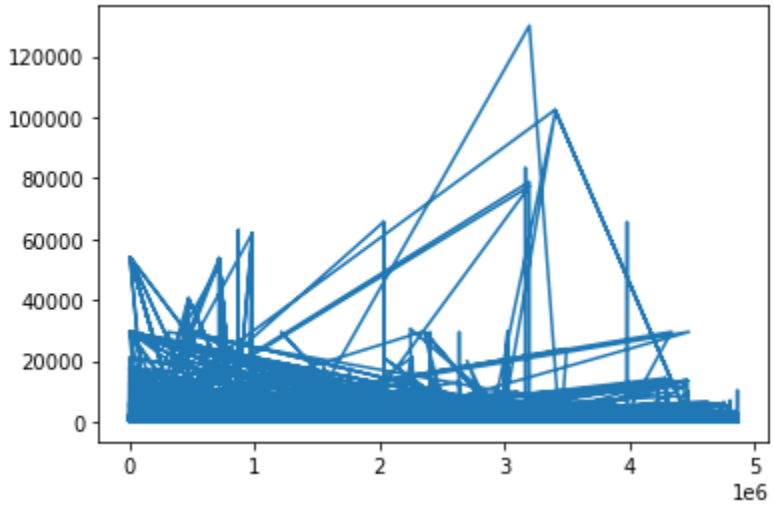
Scatter graph for categorical data in ‘status’ column.



Plot for id vs amount is:



Plot for cid vs amount is:



Dataset “s.csv”

This dataset contains 10 variables.

- Categorical Data - language, status, report_type, device

Categories and their counts in each of categorical variable are:

‘gender’ :

```
Male      5336491
Female    3722945
MALE        31319
M              45
FEMALE         20
F              17
Name: gender, dtype: int64
```

‘language’ :

```
TAM      2842756
ENG      2096781
HIN      1399821
KAN       666861
MAL       644473
TEL       635283
MAR       304390
BEN       234800
ORI       218320
GUJ        44658
SIN         7048
H              4
NIL            3
HINDI          2
SAN            2
E              1
M              1
Name: language, dtype: int64
```

‘status’ :

```
1    9095602
Name: status, dtype: int64
```

‘report_type’ :

```
LS-MT          7908843
LS-MP           680885
LS-CR           243535
LS-FH            40034
LS-SC            34223
                 ...
1029-ASK-1           1
1029-GEN-P           1
MPYS                 1
CMYM                 1
1029-TEL-C           1
Name: report_type, Length: 81, dtype: int64
```

‘device’ :

```
mobile     8165158
pc          914283
desktop      15513
MOBILE         275
PC             186
Name: device, dtype: int64
```

Dataset “tp.csv”

Contains only categorical data with ctid as numerical variables.

Categories and their counts in each of categorical variable are:

‘variant’ :

```
basic                    4008071
premium                   168432
premiumplus                 2134
premiumpluscolor             370
premiumplusconsultancy        10
premiumplusleather             7
Name: variant, dtype: int64
```

‘language’ :

```
eng    1253528
hin     995944
tam     643472
mal     309591
tel     293184
kan     225973
mar     171512
ben     130427
ori      87416
guj      42696
sin      18163
Eng       2679
Hin        799
Tam        459
Mal        336
Tel        291
Kan        205
Mar        177
Ben        138
Ori         44
Guj         44
Sin         34
ENG         33
en          18
0           14
nil         12
san          5
m            2
Hi           2
--           2
Name: language, dtype: int64
```

‘status’ :

```
PROCESSED    50085
SUSPENDED      577
INITIATED      253
ROLL_BACK       85
PDF_ERROR       55
Name: status, dtype: int64
```