In [1]:
```
import pandas as pd
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
```

In [2]:
```
import os
os.listdir(os.getcwd())
```

Out[2]:
```
['.ipynb_checkpoints',
 'Data Dictionary.xlsx',
 'data.xlsx',
 'Loan Default Predictor Project.ipynb']
```

In [3]:
```
master_data = pd.read_excel('data.xlsx')
master_data.head(2)
```

Out[3]:

	UniqueID	disbursed_amount	asset_cost	ltv	branch_id	supplier_id	manufacturer_id	Curren
0	420825	50578	58400	89.55	67	22807	45	
1	417566	53278	61360	89.63	67	22807	45	

In [4]:
```
master_data.shape
```

Out[4]:
```
(233154, 41)
```

In [5]:
```
master_data.dtypes
```

Out[5]:

```
UniqueID                              int64
disbursed_amount                      int64
asset_cost                            int64
ltv                                 float64
branch_id                             int64
supplier_id                           int64
manufacturer_id                       int64
Current_pincode_ID                    int64
Date.of.Birth                datetime64[ns]
Employment.Type                      object
DisbursalDate                datetime64[ns]
State_ID                              int64
Employee_code_ID                      int64
MobileNo_Avl_Flag                     int64
Aadhar_flag                           int64
PAN_flag                              int64
VoterID_flag                          int64
Driving_flag                          int64
Passport_flag                         int64
PERFORM_CNS.SCORE                     int64
PERFORM_CNS.SCORE.DESCRIPTION        object
PRI.NO.OF.ACCTS                       int64
PRI.ACTIVE.ACCTS                      int64
PRI.OVERDUE.ACCTS                     int64
PRI.CURRENT.BALANCE                   int64
PRI.SANCTIONED.AMOUNT                 int64
PRI.DISBURSED.AMOUNT                  int64
SEC.NO.OF.ACCTS                       int64
SEC.ACTIVE.ACCTS                      int64
SEC.OVERDUE.ACCTS                     int64
SEC.CURRENT.BALANCE                   int64
SEC.SANCTIONED.AMOUNT                 int64
SEC.DISBURSED.AMOUNT                  int64
PRIMARY.INSTAL.AMT                    int64
SEC.INSTAL.AMT                        int64
NEW.ACCTS.IN.LAST.SIX.MONTHS          int64
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   int64
AVERAGE.ACCT.AGE                     object
CREDIT.HISTORY.LENGTH                object
NO.OF_INQUIRIES                       int64
loan_default                          int64
dtype: object
```

In [6]:

```
master_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 233154 entries, 0 to 233153
Data columns (total 41 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   UniqueID                              233154 non-null  int64
 1   disbursed_amount                      233154 non-null  int64
 2   asset_cost                            233154 non-null  int64
 3   ltv                                   233154 non-null  float64
 4   branch_id                             233154 non-null  int64
 5   supplier_id                           233154 non-null  int64
 6   manufacturer_id                       233154 non-null  int64
 7   Current_pincode_ID                    233154 non-null  int64
 8   Date.of.Birth                         233154 non-null  datetime64[ns]
 9   Employment.Type                       225493 non-null  object
 10  DisbursalDate                         233154 non-null  datetime64[ns]
 11  State_ID                              233154 non-null  int64
 12  Employee_code_ID                      233154 non-null  int64
 13  MobileNo_Avl_Flag                     233154 non-null  int64
 14  Aadhar_flag                           233154 non-null  int64
 15  PAN_flag                              233154 non-null  int64
 16  VoterID_flag                          233154 non-null  int64
 17  Driving_flag                          233154 non-null  int64
 18  Passport_flag                         233154 non-null  int64
 19  PERFORM_CNS.SCORE                     233154 non-null  int64
 20  PERFORM_CNS.SCORE.DESCRIPTION         233154 non-null  object
 21  PRI.NO.OF.ACCTS                       233154 non-null  int64
 22  PRI.ACTIVE.ACCTS                      233154 non-null  int64
 23  PRI.OVERDUE.ACCTS                     233154 non-null  int64
 24  PRI.CURRENT.BALANCE                   233154 non-null  int64
 25  PRI.SANCTIONED.AMOUNT                 233154 non-null  int64
 26  PRI.DISBURSED.AMOUNT                  233154 non-null  int64
 27  SEC.NO.OF.ACCTS                       233154 non-null  int64
 28  SEC.ACTIVE.ACCTS                      233154 non-null  int64
 29  SEC.OVERDUE.ACCTS                     233154 non-null  int64
 30  SEC.CURRENT.BALANCE                   233154 non-null  int64
 31  SEC.SANCTIONED.AMOUNT                 233154 non-null  int64
 32  SEC.DISBURSED.AMOUNT                  233154 non-null  int64
 33  PRIMARY.INSTAL.AMT                    233154 non-null  int64
 34  SEC.INSTAL.AMT                        233154 non-null  int64
 35  NEW.ACCTS.IN.LAST.SIX.MONTHS          233154 non-null  int64
 36  DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   233154 non-null  int64
 37  AVERAGE.ACCT.AGE                      233154 non-null  object
 38  CREDIT.HISTORY.LENGTH                 233154 non-null  object
 39  NO.OF_INQUIRIES                       233154 non-null  int64
 40  loan_default                          233154 non-null  int64
dtypes: datetime64[ns](2), float64(1), int64(34), object(4)
memory usage: 72.9+ MB
```

In [7]:
```
master_data.nunique()
```

Out[7]:

UniqueID	233154
disbursed_amount	24565
asset_cost	46252
ltv	6579
branch_id	82
supplier_id	2953
manufacturer_id	11
Current_pincode_ID	6698
Date.of.Birth	15433
Employment.Type	2
DisbursalDate	84
State_ID	22
Employee_code_ID	3270
MobileNo_Avl_Flag	1
Aadhar_flag	2
PAN_flag	2
VoterID_flag	2
Driving_flag	2
Passport_flag	2
PERFORM_CNS.SCORE	573
PERFORM_CNS.SCORE.DESCRIPTION	20
PRI.NO.OF.ACCTS	108
PRI.ACTIVE.ACCTS	40
PRI.OVERDUE.ACCTS	22
PRI.CURRENT.BALANCE	71341
PRI.SANCTIONED.AMOUNT	44390
PRI.DISBURSED.AMOUNT	47909
SEC.NO.OF.ACCTS	37
SEC.ACTIVE.ACCTS	23
SEC.OVERDUE.ACCTS	9
SEC.CURRENT.BALANCE	3246
SEC.SANCTIONED.AMOUNT	2223
SEC.DISBURSED.AMOUNT	2553
PRIMARY.INSTAL.AMT	28067
SEC.INSTAL.AMT	1918
NEW.ACCTS.IN.LAST.SIX.MONTHS	26
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	14
AVERAGE.ACCT.AGE	192
CREDIT.HISTORY.LENGTH	294
NO.OF_INQUIRIES	25
loan_default	2

dtype: int64

In [8]:
```
master_data.isnull().sum()
```

Out[8]:
```
UniqueID                              0
disbursed_amount                      0
asset_cost                            0
ltv                                   0
branch_id                             0
supplier_id                           0
manufacturer_id                       0
Current_pincode_ID                    0
Date.of.Birth                         0
Employment.Type                    7661
DisbursalDate                         0
State_ID                              0
Employee_code_ID                      0
MobileNo_Avl_Flag                     0
Aadhar_flag                           0
PAN_flag                              0
VoterID_flag                          0
Driving_flag                          0
Passport_flag                         0
PERFORM_CNS.SCORE                     0
PERFORM_CNS.SCORE.DESCRIPTION         0
PRI.NO.OF.ACCTS                       0
PRI.ACTIVE.ACCTS                      0
PRI.OVERDUE.ACCTS                     0
PRI.CURRENT.BALANCE                   0
PRI.SANCTIONED.AMOUNT                 0
PRI.DISBURSED.AMOUNT                  0
SEC.NO.OF.ACCTS                       0
SEC.ACTIVE.ACCTS                      0
SEC.OVERDUE.ACCTS                     0
SEC.CURRENT.BALANCE                   0
SEC.SANCTIONED.AMOUNT                 0
SEC.DISBURSED.AMOUNT                  0
PRIMARY.INSTAL.AMT                    0
SEC.INSTAL.AMT                        0
NEW.ACCTS.IN.LAST.SIX.MONTHS          0
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   0
AVERAGE.ACCT.AGE                      0
CREDIT.HISTORY.LENGTH                 0
NO.OF_INQUIRIES                       0
loan_default                          0
dtype: int64
```

In [9]:
```
master_data.duplicated().sum()
```

Out[9]:
```
0
```

In [11]:
```
master_data['Employment.Type'].fillna('Self employed', inplace=True)
master_data.isnull().sum()
```

Out[11]:
```
UniqueID                              0
disbursed_amount                      0
asset_cost                            0
ltv                                   0
branch_id                             0
supplier_id                           0
manufacturer_id                       0
Current_pincode_ID                    0
Date.of.Birth                         0
Employment.Type                       0
DisbursalDate                         0
State_ID                              0
Employee_code_ID                      0
MobileNo_Avl_Flag                     0
Aadhar_flag                           0
PAN_flag                              0
VoterID_flag                          0
Driving_flag                          0
Passport_flag                         0
PERFORM_CNS.SCORE                     0
PERFORM_CNS.SCORE.DESCRIPTION         0
PRI.NO.OF.ACCTS                       0
PRI.ACTIVE.ACCTS                      0
PRI.OVERDUE.ACCTS                     0
PRI.CURRENT.BALANCE                   0
PRI.SANCTIONED.AMOUNT                 0
PRI.DISBURSED.AMOUNT                  0
SEC.NO.OF.ACCTS                       0
SEC.ACTIVE.ACCTS                      0
SEC.OVERDUE.ACCTS                     0
SEC.CURRENT.BALANCE                   0
SEC.SANCTIONED.AMOUNT                 0
SEC.DISBURSED.AMOUNT                  0
PRIMARY.INSTAL.AMT                    0
SEC.INSTAL.AMT                        0
NEW.ACCTS.IN.LAST.SIX.MONTHS          0
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   0
AVERAGE.ACCT.AGE                      0
CREDIT.HISTORY.LENGTH                 0
NO.OF_INQUIRIES                       0
loan_default                          0
dtype: int64
```

In [12]:
```
master_data.columns
```

Out[12]:
```
Index(['UniqueID', 'disbursed_amount', 'asset_cost', 'ltv', 'branch_id',
       'supplier_id', 'manufacturer_id', 'Current_pincode_ID', 'Date.of.Birth',
       'Employment.Type', 'DisbursalDate', 'State_ID', 'Employee_code_ID',
       'MobileNo_Avl_Flag', 'Aadhar_flag', 'PAN_flag', 'VoterID_flag',
       'Driving_flag', 'Passport_flag', 'PERFORM_CNS.SCORE',
       'PERFORM_CNS.SCORE.DESCRIPTION', 'PRI.NO.OF.ACCTS', 'PRI.ACTIVE.ACCTS',
       'PRI.OVERDUE.ACCTS', 'PRI.CURRENT.BALANCE', 'PRI.SANCTIONED.AMOUNT',
       'PRI.DISBURSED.AMOUNT', 'SEC.NO.OF.ACCTS', 'SEC.ACTIVE.ACCTS',
       'SEC.OVERDUE.ACCTS', 'SEC.CURRENT.BALANCE', 'SEC.SANCTIONED.AMOUNT',
       'SEC.DISBURSED.AMOUNT', 'PRIMARY.INSTAL.AMT', 'SEC.INSTAL.AMT',
       'NEW.ACCTS.IN.LAST.SIX.MONTHS', 'DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS',
       'AVERAGE.ACCT.AGE', 'CREDIT.HISTORY.LENGTH', 'NO.OF_INQUIRIES',
       'loan_default'],
      dtype='object')
```

In [13]:
```
new_col=[]
for col_name in master_data.columns:
    new_col.append(str(col_name.replace('.', '_')))
print(new_col)
```

```
['UniqueID', 'disbursed_amount', 'asset_cost', 'ltv', 'branch_id', 'supplier_id',
'manufacturer_id', 'Current_pincode_ID', 'Date_of_Birth', 'Employment_Type', 'Disb
ursalDate', 'State_ID', 'Employee_code_ID', 'MobileNo_Avl_Flag', 'Aadhar_flag', 'P
AN_flag', 'VoterID_flag', 'Driving_flag', 'Passport_flag', 'PERFORM_CNS_SCORE', 'P
ERFORM_CNS_SCORE_DESCRIPTION', 'PRI_NO_OF_ACCTS', 'PRI_ACTIVE_ACCTS', 'PRI_OVERDUE
_ACCTS', 'PRI_CURRENT_BALANCE', 'PRI_SANCTIONED_AMOUNT', 'PRI_DISBURSED_AMOUNT',
'SEC_NO_OF_ACCTS', 'SEC_ACTIVE_ACCTS', 'SEC_OVERDUE_ACCTS', 'SEC_CURRENT_BALANCE',
'SEC_SANCTIONED_AMOUNT', 'SEC_DISBURSED_AMOUNT', 'PRIMARY_INSTAL_AMT', 'SEC_INSTAL
_AMT', 'NEW_ACCTS_IN_LAST_SIX_MONTHS', 'DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS', 'AVE
RAGE_ACCT_AGE', 'CREDIT_HISTORY_LENGTH', 'NO_OF_INQUIRIES', 'loan_default']
```

In [14]:
```
master_data.columns = new_col
master_data.columns
```

Out[14]:
```
Index(['UniqueID', 'disbursed_amount', 'asset_cost', 'ltv', 'branch_id',
       'supplier_id', 'manufacturer_id', 'Current_pincode_ID', 'Date_of_Birth',
       'Employment_Type', 'DisbursalDate', 'State_ID', 'Employee_code_ID',
       'MobileNo_Avl_Flag', 'Aadhar_flag', 'PAN_flag', 'VoterID_flag',
       'Driving_flag', 'Passport_flag', 'PERFORM_CNS_SCORE',
       'PERFORM_CNS_SCORE_DESCRIPTION', 'PRI_NO_OF_ACCTS', 'PRI_ACTIVE_ACCTS',
       'PRI_OVERDUE_ACCTS', 'PRI_CURRENT_BALANCE', 'PRI_SANCTIONED_AMOUNT',
       'PRI_DISBURSED_AMOUNT', 'SEC_NO_OF_ACCTS', 'SEC_ACTIVE_ACCTS',
       'SEC_OVERDUE_ACCTS', 'SEC_CURRENT_BALANCE', 'SEC_SANCTIONED_AMOUNT',
       'SEC_DISBURSED_AMOUNT', 'PRIMARY_INSTAL_AMT', 'SEC_INSTAL_AMT',
       'NEW_ACCTS_IN_LAST_SIX_MONTHS', 'DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS',
       'AVERAGE_ACCT_AGE', 'CREDIT_HISTORY_LENGTH', 'NO_OF_INQUIRIES',
       'loan_default'],
      dtype='object')
```

In [15]:
```
master_data.Employment_Type.value_counts()
```

Out[15]:
```
Self employed    135296
Salaried          97858
Name: Employment_Type, dtype: int64
```

In [16]:
```
master_data.describe()
```

Out[16]:

	UniqueID	disbursed_amount	asset_cost	ltv	branch_id	supplier
count	233154.000000	233154.000000	2.331540e+05	233154.000000	233154.000000	233154.000
mean	535917.573376	54356.993528	7.586507e+04	74.746530	72.936094	19638.635
std	68315.693711	12971.314171	1.894478e+04	11.456636	69.834995	3491.9495
min	417428.000000	13320.000000	3.700000e+04	10.030000	1.000000	10524.000
25%	476786.250000	47145.000000	6.571700e+04	68.880000	14.000000	16535.000
50%	535978.500000	53803.000000	7.094600e+04	76.800000	61.000000	20333.000
75%	595039.750000	60413.000000	7.920175e+04	83.670000	130.000000	23000.000
max	671084.000000	990572.000000	1.628992e+06	95.000000	261.000000	24803.000

In [17]:
```
master_data.describe()['loan_default']
```

Out[17]:
```
count    233154.000000
mean          0.217071
std           0.412252
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: loan_default, dtype: float64
```

In [18]:
```
master_data['loan_default'].value_counts(normalize=True)*100
```

Out[18]:
```
0    78.292888
1    21.707112
Name: loan_default, dtype: float64
```

In [19]:
```
variable_list = ['Employment_Type', 'State_ID', 'branch_id']

for i in variable_list:
    display(master_data.groupby(i)[['loan_default']].mean().sort_values('loan_defa
```

Employment_Type	loan_default
Salaried	0.203458
Self employed	0.226917

State_ID	loan_default
22	0.118421
20	0.172973
10	0.175589
1	0.177149
19	0.180676
16	0.183613
3	0.186191
21	0.198718
5	0.198782
7	0.201739
11	0.204285
6	0.205641
4	0.207845
15	0.211527
9	0.217950
18	0.220067
8	0.229485
17	0.245803
12	0.265558
2	0.271394
14	0.275866
13	0.306587

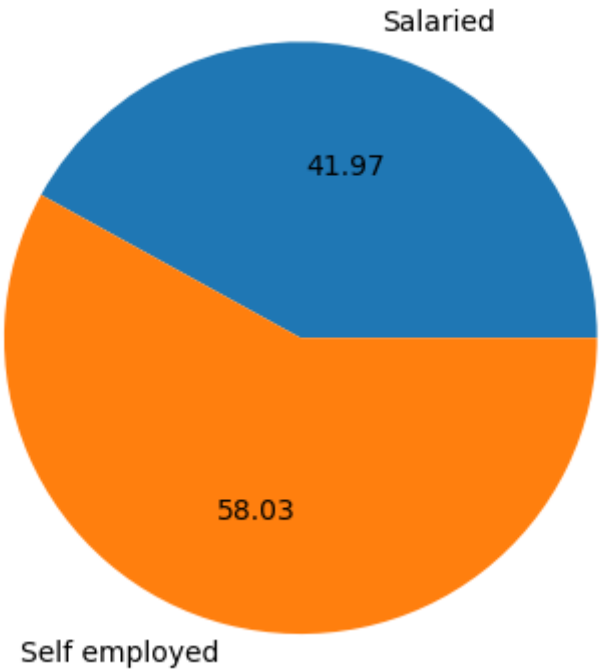
branch_id	loan_default
152	0.133387
8	0.136999
17	0.147414
1	0.149413
100	0.157100
19	0.159215
104	0.159297
142	0.167019
162	0.167513
66	0.168790
3	0.174865
15	0.179832
82	0.180676
135	0.181392
42	0.181651
34	0.182063
2	0.186863
160	0.187161
103	0.187468
63	0.189355
70	0.190666
9	0.191851
77	0.193080
121	0.193439
67	0.194032
207	0.194485
20	0.195216
48	0.197037
68	0.197926
257	0.198248
84	0.198718
130	0.201123
73	0.203576
7	0.203600
258	0.205882

branch_id	loan_default
72	0.207883
138	0.207950
11	0.209055
62	0.211288
29	0.212298
43	0.214041
250	0.216070
255	0.216364
159	0.219083
136	0.219711
79	0.220041
202	0.220110
261	0.221591
5	0.222066
165	0.222331
111	0.224719
61	0.226457
249	0.227273
13	0.227793
259	0.228324
18	0.232313
101	0.233696
64	0.238761
76	0.239766
217	0.245902
248	0.247686
85	0.256397
14	0.256739
260	0.263441
69	0.265432
120	0.265558
147	0.271394
74	0.271818
158	0.275362
10	0.276848

branch_id	loan_default
35	0.278499
16	0.280699
65	0.281847
105	0.282468
146	0.282552
117	0.283154
153	0.286127
78	0.291476
36	0.296762
97	0.313625
254	0.324308
251	0.343913

In [21]:
```
master_data.groupby('Employment_Type').size().plot(kind='pie', autopct ='%.2f')
```

Out[21]: <Axes: >



In [23]:
```
master_data['Person_Age'] = 2022 - master_data['Date_of_Birth'].dt.year
master_data[['Person_Age']].head()
```

Out[23]:

	Person_Age
0	38
1	37
2	45
3	34
4	28

In [24]:
```
master_data.groupby(pd.cut(master_data['Person_Age'], 5))['loan_default'].mean()
```

Out[24]:
```
Person_Age
(21.949, 32.2]    0.235205
(32.2, 42.4]      0.214122
(42.4, 52.6]      0.202379
(52.6, 62.8]      0.196550
(62.8, 73.0]      0.161969
Name: loan_default, dtype: float64
```

In [27]:
```
id_col=['Aadhar_flag', 'PAN_flag', 'VoterID_flag',
       'Driving_flag', 'Passport_flag']
for i in id_col:
    print("The number of people use the ID", i, ":", master_data[i].sum())
```

```
The number of people use the ID Aadhar_flag : 195924
The number of people use the ID PAN_flag : 17621
The number of people use the ID VoterID_flag : 33794
The number of people use the ID Driving_flag : 5419
The number of people use the ID Passport_flag : 496
```

In [26]:
```
master_data['AVERAGE_ACCT_AGE'].value_counts(normalize=True)
```

Out[26]:

0yrs 0mon	0.511992
0yrs 6mon	0.025854
0yrs 7mon	0.023015
0yrs 11mon	0.022462
0yrs 10mon	0.022058
1yrs 0mon	0.021578
0yrs 9mon	0.021522
0yrs 8mon	0.020982
1yrs 1mon	0.019150
0yrs 5mon	0.018674
0yrs 4mon	0.018310
1yrs 2mon	0.016903
1yrs 3mon	0.015844
0yrs 3mon	0.015312
1yrs 4mon	0.013493
1yrs 5mon	0.012781
0yrs 2mon	0.012580
1yrs 6mon	0.011932
1yrs 7mon	0.010624
1yrs 8mon	0.009504
0yrs 1mon	0.009466
1yrs 9mon	0.008737
1yrs 10mon	0.008484
2yrs 0mon	0.008419
2yrs 1mon	0.008342
1yrs 11mon	0.008162
2yrs 2mon	0.006339
2yrs 3mon	0.005447
2yrs 4mon	0.005323
2yrs 6mon	0.004997
2yrs 5mon	0.004838
2yrs 7mon	0.004529
2yrs 8mon	0.004049
2yrs 9mon	0.003959
2yrs 11mon	0.003714
2yrs 10mon	0.003689
3yrs 0mon	0.003581
3yrs 1mon	0.003303
3yrs 2mon	0.002929
3yrs 3mon	0.002681
3yrs 4mon	0.002428
3yrs 5mon	0.002196
3yrs 6mon	0.002192
3yrs 10mon	0.001939
3yrs 7mon	0.001887
3yrs 8mon	0.001879
3yrs 9mon	0.001733
3yrs 11mon	0.001656
4yrs 0mon	0.001634
4yrs 1mon	0.001394
4yrs 2mon	0.001257
4yrs 6mon	0.001192
4yrs 5mon	0.001072
4yrs 3mon	0.001047
4yrs 7mon	0.001034
4yrs 4mon	0.001004
4yrs 9mon	0.000896
5yrs 1mon	0.000854
4yrs 8mon	0.000845
4yrs 10mon	0.000832
4yrs 11mon	0.000802
5yrs 0mon	0.000789
5yrs 2mon	0.000729
5yrs 4mon	0.000575

```
5yrs 3mon      0.000566
5yrs 6mon      0.000558
5yrs 5mon      0.000558
5yrs 7mon      0.000549
5yrs 9mon      0.000463
5yrs 8mon      0.000446
5yrs 10mon     0.000399
5yrs 11mon     0.000377
6yrs 2mon      0.000360
6yrs 0mon      0.000352
6yrs 4mon      0.000347
6yrs 6mon      0.000339
6yrs 1mon      0.000322
6yrs 3mon      0.000305
6yrs 8mon      0.000287
6yrs 5mon      0.000274
7yrs 1mon      0.000262
6yrs 7mon      0.000262
6yrs 9mon      0.000240
7yrs 0mon      0.000223
6yrs 11mon     0.000193
7yrs 3mon      0.000193
6yrs 10mon     0.000193
7yrs 6mon      0.000189
7yrs 5mon      0.000167
7yrs 2mon      0.000163
7yrs 8mon      0.000150
7yrs 4mon      0.000146
7yrs 11mon     0.000142
8yrs 7mon      0.000129
7yrs 7mon      0.000129
7yrs 9mon      0.000124
8yrs 1mon      0.000124
8yrs 0mon      0.000120
8yrs 6mon      0.000120
7yrs 10mon     0.000116
8yrs 5mon      0.000094
9yrs 7mon      0.000090
8yrs 9mon      0.000077
8yrs 3mon      0.000077
9yrs 9mon      0.000069
9yrs 11mon     0.000069
10yrs 11mon    0.000069
8yrs 2mon      0.000069
9yrs 5mon      0.000064
8yrs 4mon      0.000064
10yrs 2mon     0.000064
9yrs 2mon      0.000060
11yrs 3mon     0.000056
10yrs 0mon     0.000056
12yrs 3mon     0.000051
10yrs 9mon     0.000051
11yrs 2mon     0.000051
8yrs 10mon     0.000051
11yrs 7mon     0.000051
10yrs 8mon     0.000051
10yrs 10mon    0.000051
9yrs 4mon      0.000051
10yrs 7mon     0.000047
11yrs 1mon     0.000047
9yrs 6mon      0.000047
8yrs 8mon      0.000043
11yrs 4mon     0.000043
10yrs 3mon     0.000043
```

9yrs 3mon	0.000043
9yrs 1mon	0.000043
9yrs 10mon	0.000039
11yrs 10mon	0.000039
9yrs 0mon	0.000034
12yrs 6mon	0.000034
8yrs 11mon	0.000034
9yrs 8mon	0.000034
11yrs 0mon	0.000034
10yrs 4mon	0.000030
13yrs 6mon	0.000030
11yrs 9mon	0.000030
12yrs 1mon	0.000030
12yrs 0mon	0.000030
10yrs 1mon	0.000030
10yrs 6mon	0.000030
10yrs 5mon	0.000030
11yrs 8mon	0.000026
11yrs 5mon	0.000021
13yrs 0mon	0.000021
11yrs 6mon	0.000021
13yrs 1mon	0.000021
13yrs 5mon	0.000021
12yrs 10mon	0.000017
12yrs 9mon	0.000017
12yrs 4mon	0.000017
11yrs 11mon	0.000017
13yrs 8mon	0.000013
15yrs 5mon	0.000013
12yrs 7mon	0.000013
12yrs 11mon	0.000013
14yrs 0mon	0.000013
14yrs 8mon	0.000013
12yrs 5mon	0.000013
14yrs 3mon	0.000009
17yrs 10mon	0.000009
13yrs 7mon	0.000009
13yrs 10mon	0.000009
13yrs 3mon	0.000009
12yrs 2mon	0.000009
15yrs 0mon	0.000009
13yrs 4mon	0.000009
14yrs 7mon	0.000009
18yrs 11mon	0.000004
30yrs 9mon	0.000004
14yrs 6mon	0.000004
14yrs 11mon	0.000004
16yrs 5mon	0.000004
16yrs 0mon	0.000004
14yrs 5mon	0.000004
14yrs 2mon	0.000004
22yrs 6mon	0.000004
13yrs 11mon	0.000004
24yrs 4mon	0.000004
16yrs 11mon	0.000004
16yrs 3mon	0.000004
14yrs 1mon	0.000004
15yrs 4mon	0.000004
13yrs 2mon	0.000004
15yrs 3mon	0.000004
15yrs 11mon	0.000004
16yrs 7mon	0.000004
15yrs 8mon	0.000004

```
15yrs 2mon     0.000004
Name: AVERAGE_ACCT_AGE, dtype: float64
```

In [29]:
```
#To group the account age on year basis
master_data['ACCT_Age_bracket'] = master_data['AVERAGE_ACCT_AGE'].apply(lambda x: 
master_data.head(1)
```

Out[29]:

	UniqueID	disbursed_amount	asset_cost	ltv	branch_id	supplier_id	manufacturer_id	Curren
0	420825	50578	58400	89.55	67	22807	45	

In [33]:
```
master_data.groupby('ACCT_Age_bracket')['loan_default'].mean().to_frame().sort_valu
```

Out[33]:

ACCT_Age_bracket	loan_default
11yrs	0.254902
0yrs	0.222816
9yrs	0.220000
7yrs	0.209850
8yrs	0.205128
3yrs	0.204772
2yrs	0.203518
1yrs	0.202701
15yrs	0.200000
4yrs	0.197824
12yrs	0.192982
5yrs	0.188750
10yrs	0.178295
6yrs	0.169136
14yrs	0.133333
13yrs	0.114286
24yrs	0.000000
30yrs	0.000000
18yrs	0.000000
17yrs	0.000000
16yrs	0.000000
22yrs	0.000000

In [35]:
```
master_data.groupby(pd.cut(master_data['PRI_NO_OF_ACCTS'], 5))['loan_default'].mea
```

Out[35]:
```
PRI_NO_OF_ACCTS
(-0.453, 90.6]    0.217063
(90.6, 181.2]     0.291667
(181.2, 271.8]    0.000000
(271.8, 362.4]    0.000000
(362.4, 453.0]    1.000000
Name: loan_default, dtype: float64
```

In [36]:
```
master_data.groupby(pd.cut(master_data['SEC_NO_OF_ACCTS'], 5))['loan_default'].mea
```

Out[36]:
```
SEC_NO_OF_ACCTS
(-0.052, 10.4]    0.217089
(10.4, 20.8]      0.223214
(20.8, 31.2]      0.045455
(31.2, 41.6]      0.166667
(41.6, 52.0]      0.000000
Name: loan_default, dtype: float64
```

In [37]:
```
master_data.groupby(pd.cut(master_data['NO_OF_INQUIRIES'], 2))['loan_default'].mea
```

Out[37]:
```
NO_OF_INQUIRIES
(-0.036, 18.0]    0.217073
(18.0, 36.0]      0.181818
Name: loan_default, dtype: float64
```

In [39]:
```
master_data.groupby(pd.cut(master_data['NEW_ACCTS_IN_LAST_SIX_MONTHS'], 2))[['loan_
```

Out[39]:

NEW_ACCTS_IN_LAST_SIX_MONTHS	loan_default
(-0.035, 17.5]	0.217070
(17.5, 35.0]	0.230769

In [40]:
```
master_data.groupby(pd.cut(master_data['DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS'], 2))
```

Out[40]:

DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS	loan_default
(-0.02, 10.0]	0.217065
(10.0, 20.0]	0.428571

In [41]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

master_data.columns
```

Out[41]:
```
Index(['UniqueID', 'disbursed_amount', 'asset_cost', 'ltv', 'branch_id',
       'supplier_id', 'manufacturer_id', 'Current_pincode_ID', 'Date_of_Birth',
       'Employment_Type', 'DisbursalDate', 'State_ID', 'Employee_code_ID',
       'MobileNo_Avl_Flag', 'Aadhar_flag', 'PAN_flag', 'VoterID_flag',
       'Driving_flag', 'Passport_flag', 'PERFORM_CNS_SCORE',
       'PERFORM_CNS_SCORE_DESCRIPTION', 'PRI_NO_OF_ACCTS', 'PRI_ACTIVE_ACCTS',
       'PRI_OVERDUE_ACCTS', 'PRI_CURRENT_BALANCE', 'PRI_SANCTIONED_AMOUNT',
       'PRI_DISBURSED_AMOUNT', 'SEC_NO_OF_ACCTS', 'SEC_ACTIVE_ACCTS',
       'SEC_OVERDUE_ACCTS', 'SEC_CURRENT_BALANCE', 'SEC_SANCTIONED_AMOUNT',
       'SEC_DISBURSED_AMOUNT', 'PRIMARY_INSTAL_AMT', 'SEC_INSTAL_AMT',
       'NEW_ACCTS_IN_LAST_SIX_MONTHS', 'DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS',
       'AVERAGE_ACCT_AGE', 'CREDIT_HISTORY_LENGTH', 'NO_OF_INQUIRIES',
       'loan_default', 'Person_Age', 'ACCT_Age_bracket'],
      dtype='object')
```

In [42]:
```
X = master_data[['disbursed_amount', 'asset_cost', 'ltv', 'branch_id',
       'supplier_id', 'manufacturer_id', 'Current_pincode_ID', 'DELINQUENT_ACCTS_I
       'PRI_OVERDUE_ACCTS', 'PRI_CURRENT_BALANCE', 'PRI_SANCTIONED_AMOUNT',
       'PRI_DISBURSED_AMOUNT', 'SEC_NO_OF_ACCTS', 'SEC_ACTIVE_ACCTS',
       'SEC_OVERDUE_ACCTS', 'SEC_CURRENT_BALANCE', 'SEC_SANCTIONED_AMOUNT',
       'SEC_DISBURSED_AMOUNT', 'PRIMARY_INSTAL_AMT', 'SEC_INSTAL_AMT']]
y = master_data['loan_default']
```

In [43]:
```
y.value_counts()
```

Out[43]:
```
0    182543
1     50611
Name: loan_default, dtype: int64
```

In [44]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_
```

In [46]:
```
model = LogisticRegression()
model.fit(X_train, y_train)
```

```
C:\Users\najmi\anaconda3\Lib\site-packages\sklearn\linear_model\_logistic.py:460:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[46]:
```
▾ LogisticRegression

LogisticRegression()
```

In [47]:
```
model.score(X_test, y_test)
```

Out[47]:
```
0.78738698553758
```

In [48]:
```
y_pred = model.predict(X_test)
```

In [50]:
```
from sklearn.metrics import confusion_matrix, classification_report
```

In [52]:
```
confusion_matrix(y_test, y_pred)
```

Out[52]:
```
array([[45895,     4],
       [12389,     1]], dtype=int64)
```

In [53]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.79      1.00      0.88     45899
           1       0.20      0.00      0.00     12390

    accuracy                           0.79     58289
   macro avg       0.49      0.50      0.44     58289
weighted avg       0.66      0.79      0.69     58289
```

In [ ]: