In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
df=pd.read_csv("train.csv",low_memory=False)
```

In [3]:
```
df.head()
```

Out[3]:

	ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Accounts
0	0x1602	CUS_0xd40	January	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	3
1	0x1603	CUS_0xd40	February	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	3
2	0x1604	CUS_0xd40	March	Aaron Maashoh	-500	821-00-0265	Scientist	19114.12	NaN	3
3	0x1605	CUS_0xd40	April	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	3
4	0x1606	CUS_0xd40	May	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	3

5 rows × 28 columns

Exploratory Data Analysis

(1) Data understanding.

https://github.com/bozekry/final-data-science-project/blob/main/About%20Dataset.pdf

(2) Checking Information

In [89]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 66229 entries, 0 to 99999
Data columns (total 28 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   ID                        66229 non-null  object 
 1   Customer_ID               66229 non-null  object 
 2   Month                     66229 non-null  object 
 3   Name                      59608 non-null  object 
 4   Age                       66229 non-null  int32  
 5   SSN                       66229 non-null  object 
 6   Occupation                66229 non-null  object 
 7   Annual_Income             66229 non-null  float64
 8   Monthly_Inhand_Salary     56276 non-null  float64
 9   Num_Bank_Accounts         66229 non-null  int64  
 10  Num_Credit_Card           66229 non-null  int64  
 11  Interest_Rate             66229 non-null  int64  
 12  Num_of_Loan               66229 non-null  int32  
 13  Type_of_Loan              57982 non-null  object 
 14  Delay_from_due_date       66229 non-null  int64  
 15  Num_of_Delayed_Payment    61599 non-null  float64
 16  Changed_Credit_Limit      66229 non-null  float64
 17  Num_Credit_Inquiries      64936 non-null  float64
 18  Credit_Mix                66229 non-null  object 
 19  Outstanding_Debt          66229 non-null  float64
 20  Credit_Utilization_Ratio  66229 non-null  float64
 21  Credit_History_Age        60131 non-null  object 
 22  Payment_of_Min_Amount     66229 non-null  object 
 23  Total_EMI_per_month       66229 non-null  float64
 24  Amount_invested_monthly   63301 non-null  float64
 25  Payment_Behaviour         66229 non-null  object 
 26  Monthly_Balance           65651 non-null  float64
 27  Credit_Score              66229 non-null  object 
dtypes: float64(10), int32(2), int64(4), object(12)
memory usage: 14.1+ MB
```

(3) Checking Duplicates

In [91]:
```
df.duplicated().sum()
```

Out[91]:
```
0
```

(4) Univariate analysis for categorical feature to check for error

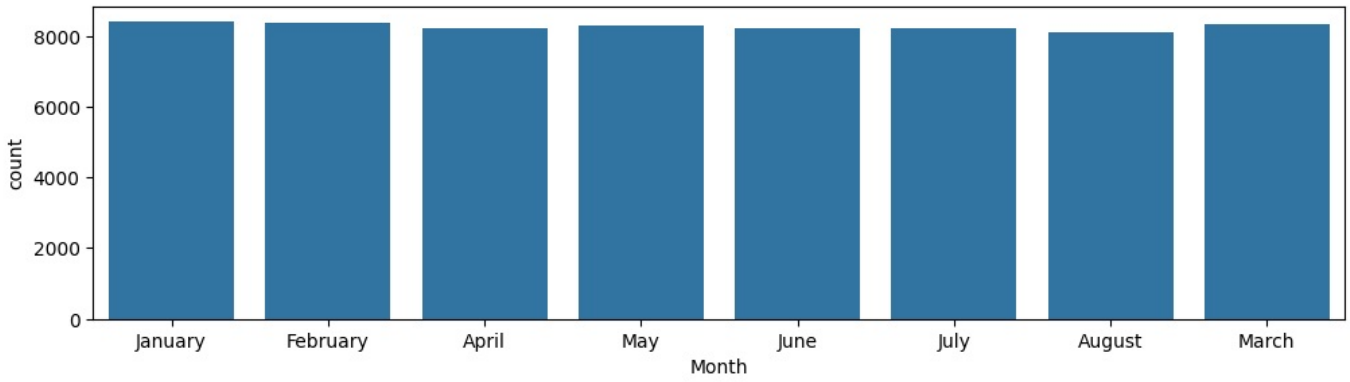
Month

In [92]:
```
df["Month"].value_counts()
```

Out[92]:
```
Month
January     8425
February    8384
March       8343
May         8290
July        8237
April       8222
June        8208
August      8120
Name: count, dtype: int64
```

In [93]:
```
plt.figure(figsize=(12,3))
sns.countplot(data=df,x="Month")
```

Out[93]:
```
<Axes: xlabel='Month', ylabel='count'>
```



Age

In [94]:
```
df["Age"].value_counts()
```

Out[94]:
```
Age
28    2000
38    1996
27    1977
26    1977
41    1970
36    1950
25    1939
44    1935
35    1905
31    1903
43    1883
22    1875
30    1871
45    1869
34    1867
32    1867
39    1865
24    1861
29    1859
20    1825
37    1821
19    1811
23    1806
21    1801
40    1748
33    1735
42    1719
18    1541
46    1181
53    1058
48    1047
15    1042
49    1034
55    1023
54     990
52     988
51     953
17     938
50     928
16     927
47     897
14     776
56     271
Name: count, dtype: int64
```

In [95]:
```
df["Age"].unique()
```

Out[95]:
```
array([23, 28, 34, 54, 55, 21, 31, 33, 30, 24, 44, 45, 40, 41, 32, 35, 36,
       39, 37, 20, 46, 26, 42, 19, 48, 43, 22, 16, 18, 15, 27, 38, 25, 14,
       17, 47, 53, 29, 49, 51, 50, 52, 56])
```

In [96]:
```
df["Age"].nunique()
```

Out[96]:
```
43
```

In [101…
```
df["Age"]
```

Out[101…
```
0        23
1        23
3        23
4        23
5        23
         ..
99994    25
99995    25
99996    25
99998    25
99999    25
Name: Age, Length: 66229, dtype: int32
```

In [102…
```
df["Age"].value_counts()
```

Out[102…
```
Age
28    2000
38    1996
27    1977
26    1977
41    1970
36    1950
25    1939
44    1935
35    1905
31    1903
43    1883
22    1875
30    1871
45    1869
34    1867
32    1867
39    1865
24    1861
29    1859
20    1825
37    1821
19    1811
23    1806
21    1801
40    1748
33    1735
42    1719
18    1541
46    1181
53    1058
48    1047
15    1042
49    1034
55    1023
54     990
52     988
51     953
17     938
50     928
16     927
47     897
14     776
56     271
Name: count, dtype: int64
```

In [103…
```
df["Age"]=df["Age"].astype('int')
```

In [104…
```
df["Age"].dtype
```

Out[104…
```
dtype('int32')
```

Occupation

In [105…
```
df["Occupation"].value_counts()
```

Out[105…
```
Occupation
Lawyer           4741
Architect        4536
Mechanic         4490
Scientist        4462
Engineer         4462
Media_Manager    4449
Journalist       4441
Developer        4436
Accountant       4411
Entrepreneur     4408
Teacher          4362
Doctor           4340
Writer           4323
Manager          4193
Musician         4175
Name: count, dtype: int64
```

In [106…
```
df["Occupation"].unique()
```

Out[106…
```
array(['Scientist', 'Teacher', 'Engineer', 'Entrepreneur', 'Developer',
       'Lawyer', 'Media_Manager', 'Doctor', 'Journalist', 'Manager',
       'Accountant', 'Musician', 'Mechanic', 'Writer', 'Architect'],
      dtype=object)
```

In [107…
```
df[df["Occupation"]=='_______']
```

Out[107…

ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Accounts	...	Credit

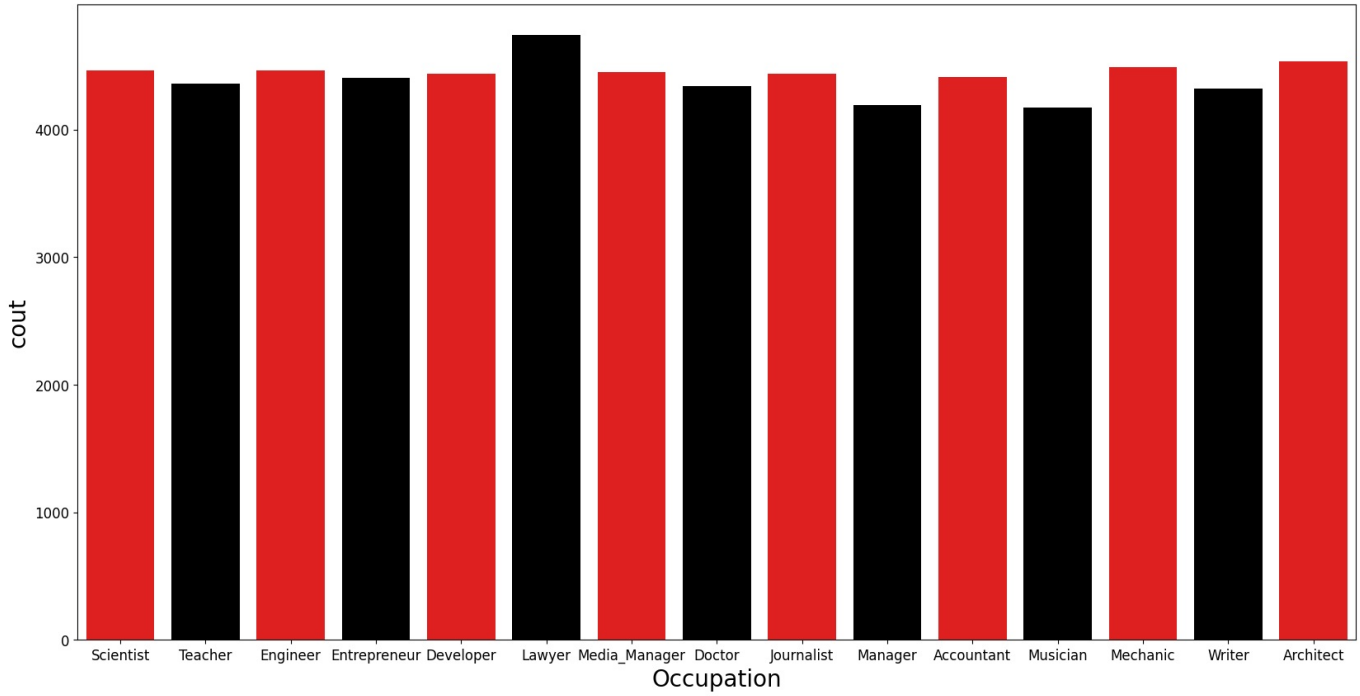
0 rows × 28 columns

In [108…
```
df.drop(df[df["Occupation"]=='_______'].index,inplace=True)
```

In [109…
```
plt.figure(figsize=(20,10))

gfg=sns.countplot(data=df,x="Occupation",palette=["red","black"])

gfg.tick_params(labelsize=12)

gfg.set_xlabel("Occupation",fontsize=20)

gfg.set_ylabel("cout",fontsize=20)
```

```
C:\Users\shiva\AppData\Local\Temp\ipykernel_25140\1206106532.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable
to `hue` and set `legend=False` for the same effect.

  gfg=sns.countplot(data=df,x="Occupation",palette=["red","black"])
C:\Users\shiva\AppData\Local\Temp\ipykernel_25140\1206106532.py:3: UserWarning:
The palette list has fewer values (2) than needed (15) and will cycle, which may produce an uninterpretable plot
.
  gfg=sns.countplot(data=df,x="Occupation",palette=["red","black"])
```

Out[109… `Text(0, 0.5, 'cout')`



Annual_Income

In [110…
```
df["Annual_Income"].unique()
```

Out[110…
```
array([ 19114.12,  34847.84, 143162.64, ...,  37188.1 ,  20002.88,
        39628.99])
```

In [111…
```
df["Annual_Income"]
```

Out[111…
```
0        19114.12
1        19114.12
3        19114.12
4        19114.12
5        19114.12
           ...
99994    39628.99
99995    39628.99
99996    39628.99
99998    39628.99
99999    39628.99
Name: Annual_Income, Length: 66229, dtype: float64
```

In [112…
```
df["Annual_Income"]=df["Annual_Income"].astype("float")
```

In [113…
```
df["Annual_Income"].dtype
```

Out[113…
```
dtype('float64')
```

Num_of_Loan

In [114…
```
df["Num_of_Loan"].unique()
```

Out[114…
```
array([4, 1, 3, 0, 2, 7, 5, 6, 8, 9])
```

In [116…
```
df["Num_of_Loan"]=df["Num_of_Loan"].astype("int")
```

Type_of_Loan

In [117…
```
df["Type_of_Loan"]
```

Out[117…
```
0        Auto Loan, Credit-Builder Loan, Personal Loan,...
1        Auto Loan, Credit-Builder Loan, Personal Loan,...
3        Auto Loan, Credit-Builder Loan, Personal Loan,...
4        Auto Loan, Credit-Builder Loan, Personal Loan,...
5        Auto Loan, Credit-Builder Loan, Personal Loan,...
                              ...
99994                          Auto Loan, and Student Loan
99995                          Auto Loan, and Student Loan
99996                          Auto Loan, and Student Loan
99998                          Auto Loan, and Student Loan
99999                          Auto Loan, and Student Loan
Name: Type_of_Loan, Length: 66229, dtype: object
```

Num_of_Delayed_Payment

In [118…
```
df["Num_of_Delayed_Payment"]
```

Out[118…
```
0        7.0
1        NaN
3        4.0
4        NaN
5        4.0
        ...
99994    6.0
99995    7.0
99996    7.0
99998    NaN
99999    6.0
Name: Num_of_Delayed_Payment, Length: 66229, dtype: float64
```

In [121…
```
df["Num_of_Delayed_Payment"].unique()
```

Out[121…
```
array([ 7.000e+00,        nan,  4.000e+00,  8.000e+00,  6.000e+00,
        1.000e+00,  3.000e+00,  0.000e+00,  5.000e+00,  9.000e+00,
        1.500e+01,  1.700e+01,  2.000e+00,  1.400e+01,  1.100e+01,
        2.000e+01,  2.200e+01,  1.000e+01,  1.300e+01,  1.600e+01,
        1.200e+01,  1.800e+01,  1.900e+01,  2.300e+01,  2.100e+01,
        2.400e+01,  3.318e+03,  3.083e+03,  1.338e+03,  1.830e+02,
       -1.000e+00,  1.106e+03,  2.600e+01,  2.672e+03,  2.008e+03,
        3.478e+03,  2.420e+03,  7.070e+02,  2.500e+01,  7.080e+02,
        3.815e+03,  2.800e+01, -2.000e+00,  1.867e+03,  2.700e+01,
        2.250e+03,  1.463e+03, -3.000e+00,  1.941e+03,  2.628e+03,
        1.320e+02,  3.060e+02,  3.539e+03,  3.684e+03,  1.823e+03,
        1.946e+03,  8.270e+02,  2.297e+03,  2.566e+03,  1.820e+02,
        2.503e+03,  2.812e+03,  1.697e+03,  8.510e+02,  3.905e+03,
        9.230e+02,  8.800e+01,  1.668e+03,  3.253e+03,  2.689e+03,
        3.858e+03,  6.420e+02,  3.457e+03,  2.204e+03,  3.103e+03,
        1.063e+03,  2.569e+03,  2.110e+02,  7.930e+02,  3.484e+03,
        3.491e+03,  3.050e+03,  3.402e+03,  1.718e+03,  3.260e+03,
        3.855e+03,  8.400e+01,  2.311e+03,  3.251e+03,  1.832e+03,
        4.069e+03,  3.010e+03,  7.330e+02,  4.241e+03,  1.660e+02,
        2.461e+03,  1.749e+03,  3.200e+03,  6.630e+02,  2.185e+03,
        4.161e+03,  3.009e+03,  3.590e+02,  2.015e+03,  1.523e+03,
        5.940e+02,  1.199e+03,  1.015e+03,  5.590e+02,  2.165e+03,
        7.790e+02,  1.920e+02,  2.323e+03,  1.538e+03,  3.529e+03,
        3.456e+03,  3.040e+03,  1.014e+03,  3.179e+03,  1.332e+03,
        3.175e+03,  3.112e+03,  8.290e+02,  4.022e+03,  5.310e+02,
        3.092e+03,  2.400e+03,  3.536e+03,  5.440e+02,  1.864e+03,
        2.300e+03,  7.200e+01,  4.970e+02,  3.980e+02,  2.222e+03,
        1.473e+03,  3.043e+03,  4.216e+03,  2.903e+03,  1.323e+03,
```

```
1.328e+03,  3.404e+03,  2.438e+03,  8.090e+02,  1.996e+03,
4.164e+03,  1.370e+03,  1.204e+03,  2.167e+03,  4.011e+03,
2.594e+03,  2.533e+03,  1.663e+03,  1.018e+03,  2.919e+03,
3.316e+03,  2.589e+03,  2.801e+03,  4.266e+03,  1.243e+03,
8.450e+02,  4.107e+03,  2.900e+02,  2.450e+03,  3.738e+03,
1.792e+03,  9.600e+02,  1.706e+03,  3.031e+03,  2.794e+03,
2.219e+03,  4.096e+03,  2.657e+03,  2.938e+03,  4.384e+03,
3.533e+03,  2.677e+03,  2.609e+03,  4.326e+03,  4.211e+03,
8.230e+02,  1.608e+03,  2.860e+03,  4.219e+03,  1.531e+03,
7.420e+02,  5.200e+01,  4.024e+03,  1.673e+03,  4.900e+01,
2.243e+03,  1.685e+03,  3.489e+03,  7.490e+02,  1.164e+03,
2.616e+03,  8.480e+02,  4.134e+03,  1.530e+03,  4.075e+03,
2.697e+03,  2.573e+03,  6.400e+02,  2.585e+03,  2.230e+03,
1.795e+03,  1.180e+03,  1.534e+03,  3.739e+03,  1.849e+03,
4.191e+03,  9.960e+02,  3.720e+02,  6.020e+02,  7.870e+02,
4.135e+03,  1.359e+03,  3.107e+03,  1.263e+03,  7.090e+02,
4.077e+03,  2.943e+03,  2.793e+03,  3.191e+03,  2.317e+03,
2.237e+03,  3.819e+03,  8.470e+02,  1.833e+03,  2.737e+03,
1.192e+03,  1.481e+03,  2.286e+03,  2.730e+02,  3.944e+03,
1.478e+03,  3.749e+03,  2.508e+03,  2.959e+03,  1.300e+02,
2.940e+02,  3.097e+03,  3.511e+03,  4.150e+02,  2.196e+03,
2.149e+03,  1.874e+03,  1.553e+03,  1.222e+03,  2.907e+03,
3.051e+03,  2.314e+03,  1.636e+03,  8.000e+01,  3.708e+03,
1.950e+02,  2.945e+03,  1.911e+03,  3.416e+03,  3.796e+03,
2.255e+03,  9.380e+02,  4.397e+03,  2.148e+03,  3.864e+03,
1.687e+03,  1.034e+03,  2.044e+03,  3.661e+03,  1.211e+03,
2.007e+03,  1.020e+02,  1.891e+03,  3.162e+03,  3.142e+03,
3.881e+03,  2.728e+03,  1.952e+03,  3.840e+03,  3.119e+03,
4.185e+03,  2.954e+03,  6.830e+02,  1.614e+03,  3.447e+03,
1.852e+03,  2.131e+03,  1.900e+03,  1.699e+03,  2.018e+03,
5.080e+02,  5.770e+02,  2.604e+03,  1.411e+03,  2.351e+03,
2.352e+03,  1.191e+03,  9.050e+02,  4.053e+03,  3.869e+03,
9.330e+02,  3.660e+03,  3.300e+03,  3.629e+03,  3.208e+03,
2.142e+03,  2.521e+03,  4.500e+02,  5.830e+02,  8.760e+02,
1.210e+02,  3.919e+03,  2.560e+03,  2.578e+03,  2.060e+03,
1.236e+03,  4.360e+03,  4.172e+03,  3.909e+03,  3.951e+03,
2.712e+03,  2.498e+03,  3.171e+03,  1.750e+03,  1.970e+02,
2.650e+02,  2.397e+03,  4.337e+03,  2.950e+03,  1.859e+03,
1.070e+02,  2.348e+03,  2.810e+03,  2.873e+03,  3.078e+03,
1.278e+03,  3.793e+03,  2.276e+03,  2.879e+03,  2.141e+03,
2.230e+02,  1.862e+03,  2.617e+03,  3.972e+03,  2.334e+03,
2.759e+03,  4.169e+03,  2.280e+03,  2.492e+03,  3.750e+03,
1.825e+03,  3.090e+02,  2.431e+03,  3.099e+03,  2.080e+03,
2.279e+03,  2.666e+03,  1.976e+03,  5.290e+02,  1.985e+03,
3.060e+03,  3.212e+03,  3.790e+03,  1.536e+03,  3.955e+03,
2.324e+03,  2.381e+03,  3.710e+02,  3.880e+03,  2.991e+03,
4.319e+03,  6.620e+02,  4.144e+03,  6.930e+02,  2.006e+03,
3.751e+03,  4.262e+03,  2.913e+03,  3.492e+03,  8.000e+02,
3.766e+03,  1.087e+03,  1.086e+03,  2.216e+03,  3.522e+03,
3.274e+03,  3.488e+03,  2.380e+02,  3.510e+02,  3.706e+03,
4.280e+03,  4.095e+03,  2.926e+03,  1.329e+03,  3.370e+03,
2.429e+03,  1.133e+03,  4.388e+03,  4.282e+03,  4.281e+03,
3.415e+03,  2.001e+03,  4.410e+02,  9.400e+01,  3.499e+03,
3.368e+03,  1.004e+03,  3.946e+03,  2.956e+03,  4.324e+03,
4.113e+03,  1.172e+03,  2.553e+03,  1.765e+03,  3.495e+03,
1.392e+03,  4.239e+03,  6.740e+02,  2.636e+03,  3.722e+03,
4.295e+03,  1.653e+03,  1.325e+03,  1.879e+03,  1.096e+03,
1.735e+03,  3.584e+03,  1.975e+03,  3.827e+03,  2.552e+03,
3.754e+03,  5.320e+02,  9.260e+02,  3.763e+03,  7.780e+02,
2.621e+03,  2.418e+03,  3.926e+03,  3.861e+03,  3.574e+03,
1.750e+02,  1.620e+02,  2.834e+03,  3.765e+03,  3.355e+03,
5.230e+02,  2.274e+03,  1.606e+03,  1.443e+03,  1.354e+03,
1.422e+03,  4.231e+03,  2.278e+03,  1.045e+03,  4.106e+03,
3.155e+03,  6.660e+02,  1.841e+03,  2.076e+03,  2.384e+03,
1.954e+03,  7.190e+02,  4.002e+03,  5.410e+02,  3.894e+03,
1.256e+03,  3.990e+02,  8.600e+01,  1.571e+03,  4.037e+03,
4.160e+02,  4.005e+03,  2.671e+03,  1.150e+03,  2.591e+03,
1.801e+03,  1.775e+03,  2.260e+03,  3.707e+03,  1.820e+03,
1.480e+03,  1.850e+03,  4.300e+02,  1.579e+03,  3.391e+03,
2.385e+03,  3.336e+03,  3.688e+03,  2.210e+02,  2.047e+03])
```

In [122…
```
df["Num_of_Delayed_Payment"]=df["Num_of_Delayed_Payment"].astype("float")
```

Changed_Credit_Limit

In [123…
```
df["Changed_Credit_Limit"]
```

Out[123…
```
0        11.27
1        11.27
3         6.27
4        11.27
5         9.27
         ...
99994     9.50
99995    11.50
99996    11.50
99998    11.50
99999    11.50
Name: Changed_Credit_Limit, Length: 66229, dtype: float64
```

In [124…
```
df[df["Changed_Credit_Limit"]=="_"]
```

Out[124…

ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Accounts	...	Credit

0 rows × 28 columns

In [125…
```
df.drop(df[df["Changed_Credit_Limit"]=="_"].index,inplace=True)
```

In [127…
```
df["Changed_Credit_Limit"]=df["Changed_Credit_Limit"].astype("float")
```

Credit_Mix

In [128…
```
df[df["Credit_Mix"]=='_']
```

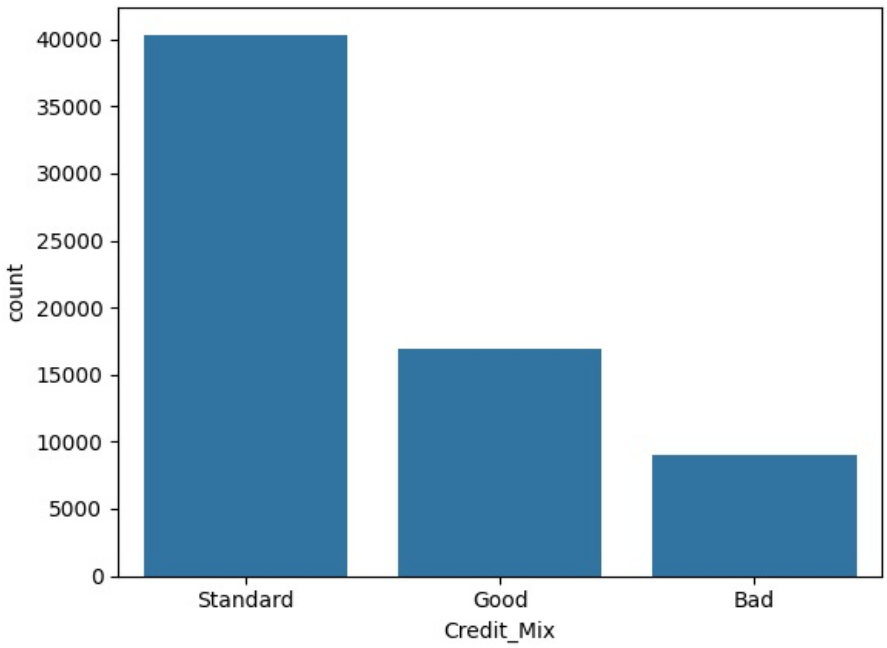
Out[128…

ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Accounts	...	Credit

0 rows × 28 columns

In [129…
```
sns.countplot(data=df,x="Credit_Mix")
```

Out[129…
```
<Axes: xlabel='Credit_Mix', ylabel='count'>
```



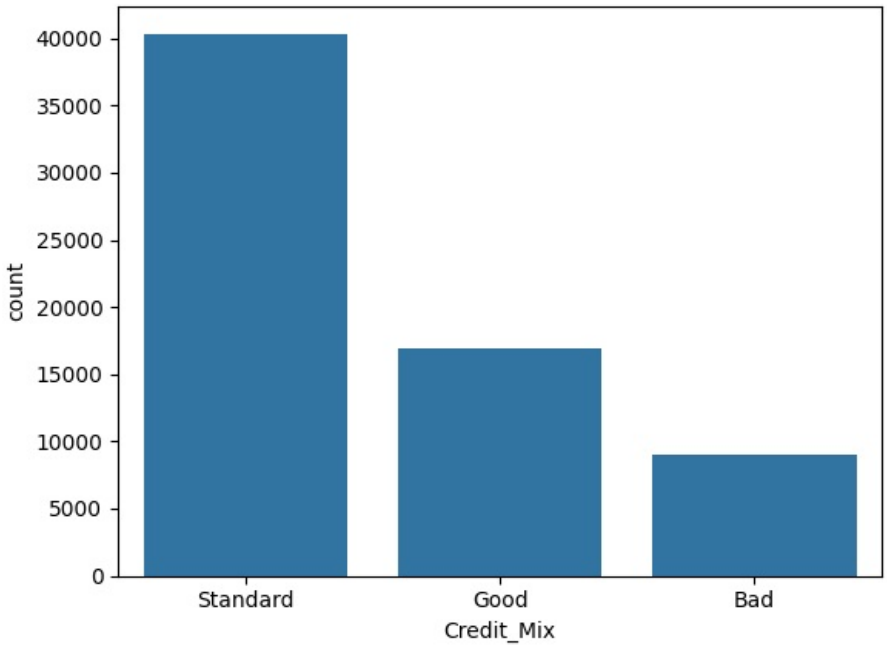
In [130…
```
df["Credit_Mix"]=df["Credit_Mix"].apply(lambda x:x.replace('_',"Standard"))
```

In [131…
```
df["Credit_Mix"].unique()
```

Out[131…
```
array(['Standard', 'Good', 'Bad'], dtype=object)
```

In [132…
```
sns.countplot(data=df,x="Credit_Mix")
```

Out[132…
```
<Axes: xlabel='Credit_Mix', ylabel='count'>
```



Outstanding_Debt

In [145…
```
df["Outstanding_Debt"]=df["Outstanding_Debt"].astype("float")
```

In [147…
```
df["Outstanding_Debt"]
```

Out[147…
```
0        809.0
1        809.0
3        809.0
4        809.0
5        809.0
         ...
99994    502.0
99995    502.0
99996    502.0
99998    502.0
99999    502.0
Name: Outstanding_Debt, Length: 66229, dtype: float64
```

Credit_History_Age

In [148…
```
df["Credit_History_Age"]
```

Out[148…
```
0         22 Years and 1 Months
1                           NaN
3         22 Years and 4 Months
4         22 Years and 5 Months
5         22 Years and 6 Months
                  ...
99994     31 Years and 5 Months
99995     31 Years and 6 Months
99996     31 Years and 7 Months
99998     31 Years and 9 Months
99999    31 Years and 10 Months
Name: Credit_History_Age, Length: 66229, dtype: object
```

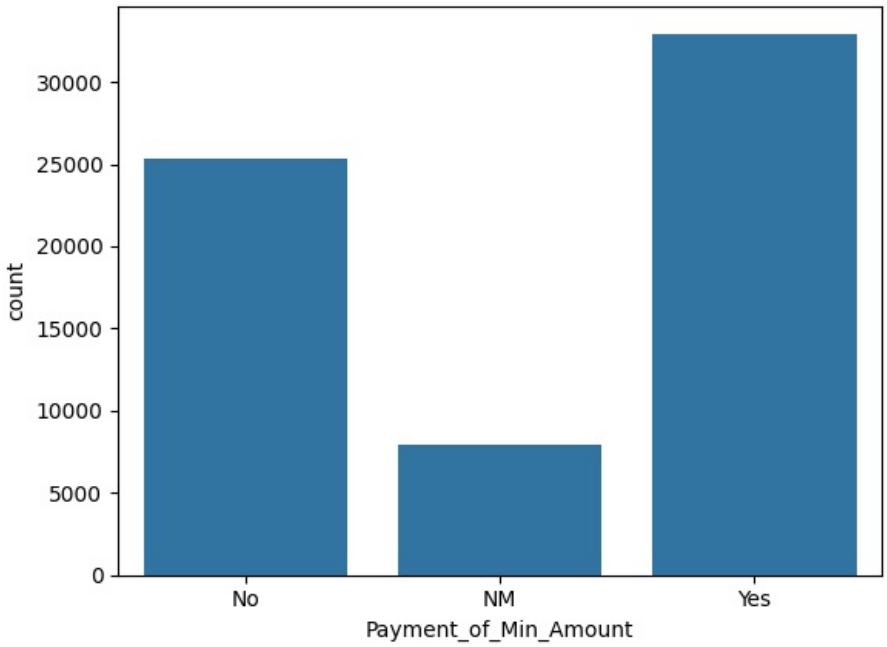
Payment_of_Min_Amount

In [149…
```
df["Payment_of_Min_Amount"].value_counts()
```

Out[149…
```
Payment_of_Min_Amount
Yes    32953
No     25371
NM      7905
Name: count, dtype: int64
```

In [150…
```
sns.countplot(data=df,x="Payment_of_Min_Amount")
```

Out[150…
```
<Axes: xlabel='Payment_of_Min_Amount', ylabel='count'>
```



Amount_invested_monthly

In [151…
```
def Amount_invested_monthly(col):
    if "__" in str(col):
        return str(col).split("__")[1]
    else:
        return str(col)
```

In [152…
```
Amount_invested_monthly('__10000__')
```

Out[152…
```
'10000'
```

In [153…
```
df["Amount_invested_monthly"]=df["Amount_invested_monthly"].apply(Amount_invested_monthly)
```

In [154…
```
df["Amount_invested_monthly"]=df["Amount_invested_monthly"].astype("float")
```

Payment_Behaviour

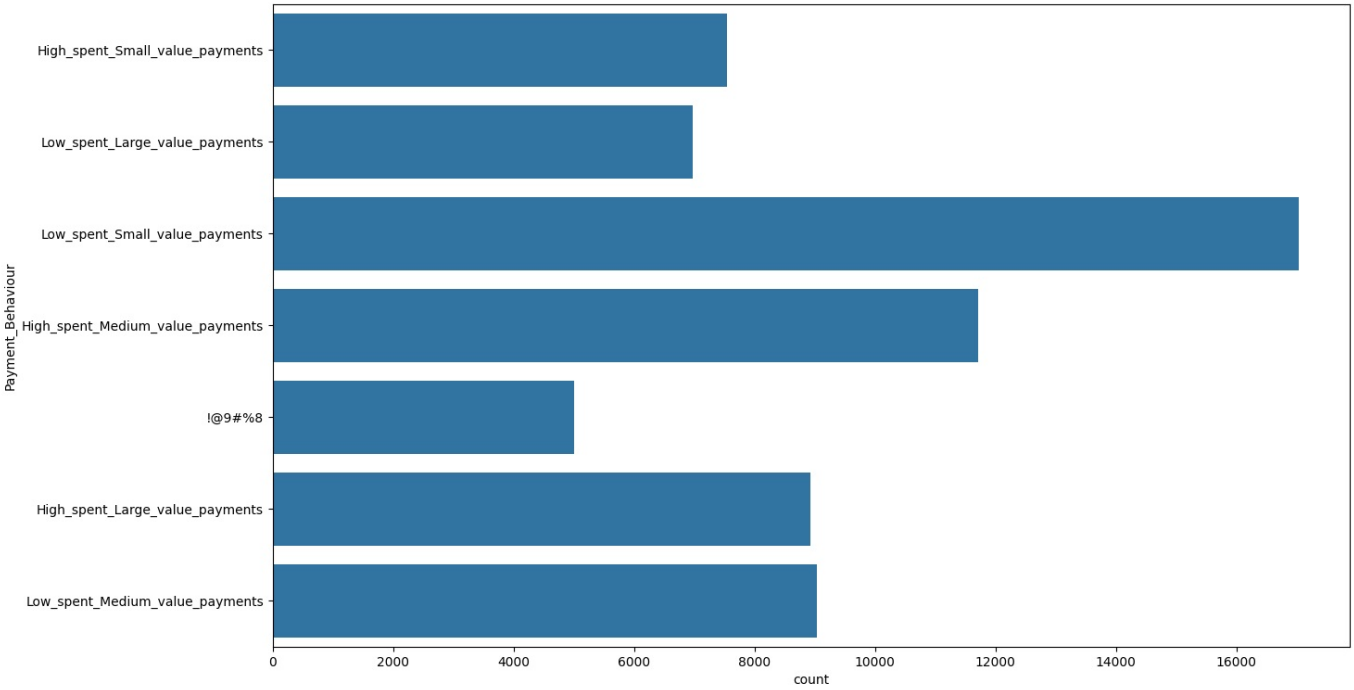
In [155…
```
df["Payment_Behaviour"].value_counts()
```

Out[155…
```
Payment_Behaviour
Low_spent_Small_value_payments      17033
High_spent_Medium_value_payments    11707
Low_spent_Medium_value_payments      9035
High_spent_Large_value_payments      8932
High_spent_Small_value_payments      7539
Low_spent_Large_value_payments       6977
!@9#%8                               5006
Name: count, dtype: int64
```

In [ ]:
```
```

In [156…
```
fig = plt.figure(figsize= (15,9))
sns.countplot(data=df,y="Payment_Behaviour")
```

Out[156…
```
<Axes: xlabel='count', ylabel='Payment_Behaviour'>
```



(5) Check Missing Values

In [164…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 66229 entries, 0 to 99999
Data columns (total 28 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   ID                        66229 non-null  object
 1   Customer_ID               66229 non-null  object
 2   Month                     66229 non-null  object
 3   Name                      59608 non-null  object
 4   Age                       66229 non-null  int32
 5   SSN                       66229 non-null  object
 6   Occupation                66229 non-null  object
 7   Annual_Income             66229 non-null  float64
 8   Monthly_Inhand_Salary     56276 non-null  float64
 9   Num_Bank_Accounts         66229 non-null  int64
 10  Num_Credit_Card           66229 non-null  int64
 11  Interest_Rate             66229 non-null  int64
 12  Num_of_Loan               66229 non-null  int32
 13  Type_of_Loan              57982 non-null  object
 14  Delay_from_due_date       66229 non-null  int64
 15  Num_of_Delayed_Payment    61599 non-null  float64
 16  Changed_Credit_Limit      66229 non-null  float64
 17  Num_Credit_Inquiries      64936 non-null  float64
 18  Credit_Mix                66229 non-null  object
 19  Outstanding_Debt          66229 non-null  float64
 20  Credit_Utilization_Ratio  66229 non-null  float64
 21  Credit_History_Age        60131 non-null  object
 22  Payment_of_Min_Amount     66229 non-null  object
 23  Total_EMI_per_month       66229 non-null  float64
 24  Amount_invested_monthly   63301 non-null  float64
 25  Payment_Behaviour         66229 non-null  object
 26  Monthly_Balance           65651 non-null  float64
 27  Credit_Score              66229 non-null  object
dtypes: float64(10), int32(2), int64(4), object(12)
memory usage: 14.1+ MB
```

In [165…
```
df.isna().sum()
```

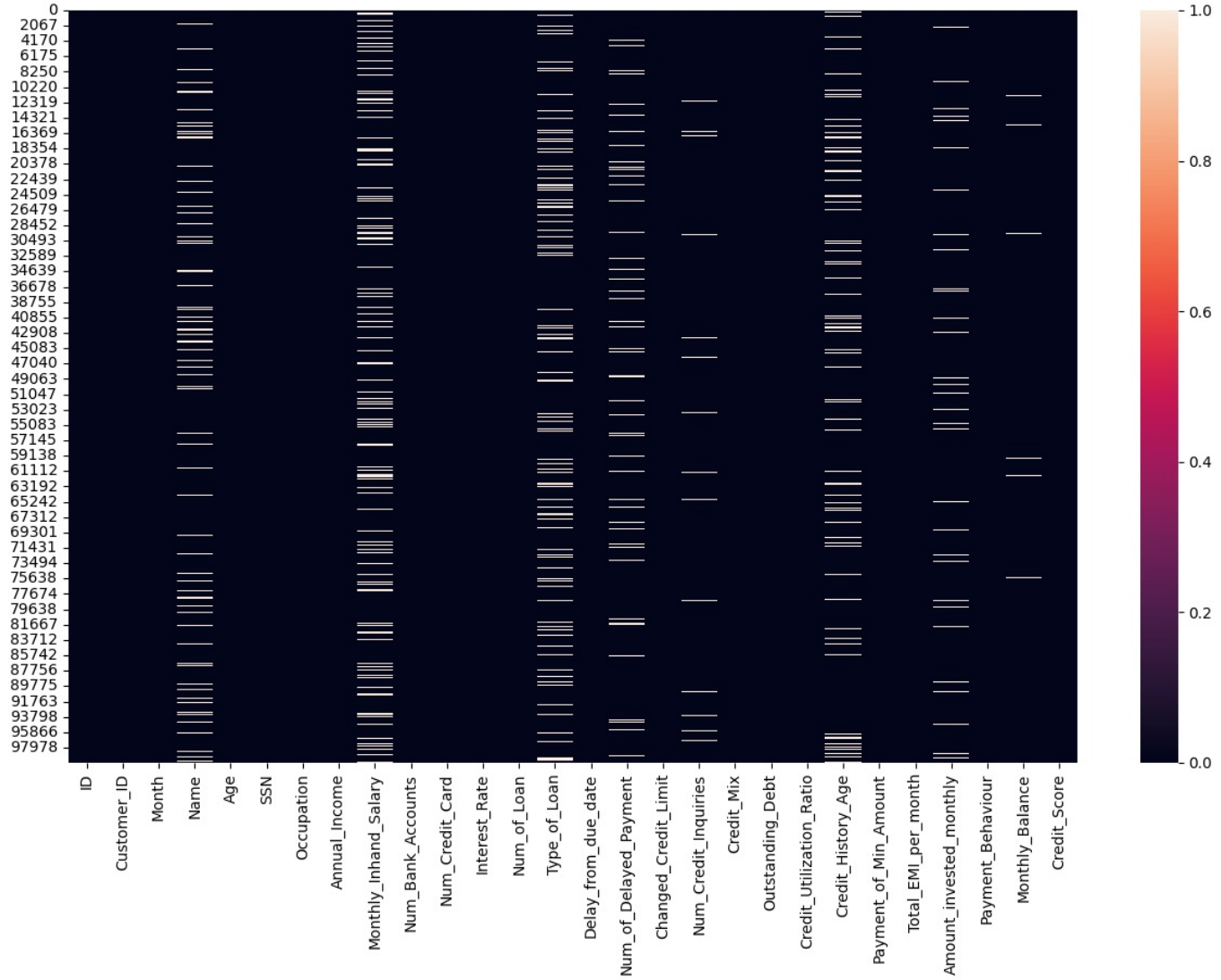
Out[165…
```
ID                          0
Customer_ID                 0
Month                       0
Name                     6621
Age                         0
SSN                         0
Occupation                  0
Annual_Income               0
Monthly_Inhand_Salary    9953
Num_Bank_Accounts           0
Num_Credit_Card             0
Interest_Rate               0
Num_of_Loan                 0
Type_of_Loan             8247
Delay_from_due_date         0
Num_of_Delayed_Payment   4630
Changed_Credit_Limit        0
Num_Credit_Inquiries     1293
Credit_Mix                  0
Outstanding_Debt            0
Credit_Utilization_Ratio    0
Credit_History_Age       6098
Payment_of_Min_Amount       0
Total_EMI_per_month         0
Amount_invested_monthly  2928
Payment_Behaviour           0
Monthly_Balance           578
Credit_Score                0
dtype: int64
```

In [166…
```
df.isna().mean()*100
```

Out[166…
```
ID                        0.000000
Customer_ID               0.000000
Month                     0.000000
Name                      9.997131
Age                       0.000000
SSN                       0.000000
Occupation                0.000000
Annual_Income             0.000000
Monthly_Inhand_Salary    15.028160
Num_Bank_Accounts         0.000000
Num_Credit_Card           0.000000
Interest_Rate             0.000000
Num_of_Loan               0.000000
Type_of_Loan             12.452249
Delay_from_due_date       0.000000
Num_of_Delayed_Payment    6.990895
Changed_Credit_Limit      0.000000
Num_Credit_Inquiries      1.952317
Credit_Mix                0.000000
Outstanding_Debt          0.000000
Credit_Utilization_Ratio  0.000000
Credit_History_Age        9.207447
Payment_of_Min_Amount     0.000000
Total_EMI_per_month       0.000000
Amount_invested_monthly   4.421024
Payment_Behaviour         0.000000
Monthly_Balance           0.872729
Credit_Score              0.000000
dtype: float64
```

In [167…
```
fig = plt.figure(figsize= (15,9))
sns.heatmap(df.isna())
```

Out[167…
```
<Axes: >
```



(6) Detect outliers and split df to outliers and clean then analyze outliers df.

In [168…
```
pip install datasist
```

```
Requirement already satisfied: datasist in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (1.5.3)Note: you may need to restart the kernel to use updated packages.

Requirement already satisfied: pandas in c:\users\shiva\appdata\local\programs\python\python311\lib\site-package
s (from datasist) (2.1.0)
Requirement already satisfied: matplotlib in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pac
kages (from datasist) (3.8.0)
Requirement already satisfied: seaborn in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packag
es (from datasist) (0.13.0)
Requirement already satisfied: numpy in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packages
(from datasist) (1.25.2)
Requirement already satisfied: jupyter in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packag
es (from datasist) (1.0.0)
Requirement already satisfied: scikit-learn in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from datasist) (1.3.2)
Requirement already satisfied: nltk in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packages
(from datasist) (3.8.1)
Requirement already satisfied: Joblib in c:\users\shiva\appdata\local\programs\python\python311\lib\site-package
```

```
s (from datasist) (1.3.2)
Requirement already satisfied: notebook in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from jupyter->datasist) (7.0.2)
Requirement already satisfied: qtconsole in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from jupyter->datasist) (5.4.3)
Requirement already satisfied: jupyter-console in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from jupyter->datasist) (6.6.3)
Requirement already satisfied: nbconvert in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from jupyter->datasist) (7.7.4)
Requirement already satisfied: ipykernel in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from jupyter->datasist) (6.25.1)
Requirement already satisfied: ipywidgets in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pac
kages (from jupyter->datasist) (8.1.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from matplotlib->datasist) (1.1.1)
Requirement already satisfied: cycler>=0.10 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from matplotlib->datasist) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from matplotlib->datasist) (4.43.1)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from matplotlib->datasist) (1.4.5)
Requirement already satisfied: packaging>=20.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from matplotlib->datasist) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-
packages (from matplotlib->datasist) (10.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from matplotlib->datasist) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\shiva\appdata\local\programs\python\python311\li
b\site-packages (from matplotlib->datasist) (2.8.2)
Requirement already satisfied: click in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packages
(from nltk->datasist) (8.1.7)
Requirement already satisfied: regex>=2021.8.3 in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from nltk->datasist) (2023.12.25)
Requirement already satisfied: tqdm in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packages
(from nltk->datasist) (4.66.1)
Requirement already satisfied: pytz>=2020.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from pandas->datasist) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site
-packages (from pandas->datasist) (2023.3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from scikit-learn->datasist) (1.11.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\shiva\appdata\local\programs\python\python311\li
b\site-packages (from scikit-learn->datasist) (3.2.0)
Requirement already satisfied: six>=1.5 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from python-dateutil>=2.7->matplotlib->datasist) (1.16.0)
Requirement already satisfied: colorama in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from click->nltk->datasist) (0.4.6)
Requirement already satisfied: comm>=0.1.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from ipykernel->jupyter->datasist) (0.1.4)
Requirement already satisfied: debugpy>=1.6.5 in c:\users\shiva\appdata\local\programs\python\python311\lib\site
-packages (from ipykernel->jupyter->datasist) (1.6.7.post1)
Requirement already satisfied: ipython>=7.23.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from ipykernel->jupyter->datasist) (8.14.0)
Requirement already satisfied: jupyter-client>=6.1.12 in c:\users\shiva\appdata\local\programs\python\python311\
lib\site-packages (from ipykernel->jupyter->datasist) (8.3.0)
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in c:\users\shiva\appdata\local\programs\python\python
311\lib\site-packages (from ipykernel->jupyter->datasist) (5.3.1)
Requirement already satisfied: matplotlib-inline>=0.1 in c:\users\shiva\appdata\local\programs\python\python311\
lib\site-packages (from ipykernel->jupyter->datasist) (0.1.6)
Requirement already satisfied: nest-asyncio in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from ipykernel->jupyter->datasist) (1.5.7)
Requirement already satisfied: psutil in c:\users\shiva\appdata\local\programs\python\python311\lib\site-package
s (from ipykernel->jupyter->datasist) (5.9.5)
Requirement already satisfied: pyzmq>=20 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from ipykernel->jupyter->datasist) (25.1.1)
Requirement already satisfied: tornado>=6.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from ipykernel->jupyter->datasist) (6.3.3)
Requirement already satisfied: traitlets>=5.4.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from ipykernel->jupyter->datasist) (5.9.0)
Requirement already satisfied: widgetsnbextension~=4.0.7 in c:\users\shiva\appdata\local\programs\python\python3
11\lib\site-packages (from ipywidgets->jupyter->datasist) (4.0.8)
Requirement already satisfied: jupyterlab-widgets~=3.0.7 in c:\users\shiva\appdata\local\programs\python\python3
11\lib\site-packages (from ipywidgets->jupyter->datasist) (3.0.8)
Requirement already satisfied: prompt-toolkit>=3.0.30 in c:\users\shiva\appdata\local\programs\python\python311\
lib\site-packages (from jupyter-console->jupyter->datasist) (3.0.39)
Requirement already satisfied: pygments in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from jupyter-console->jupyter->datasist) (2.16.1)
Requirement already satisfied: beautifulsoup4 in c:\users\shiva\appdata\local\programs\python\python311\lib\site
-packages (from nbconvert->jupyter->datasist) (4.12.2)
Requirement already satisfied: bleach!=5.0.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-
packages (from nbconvert->jupyter->datasist) (6.0.0)
Requirement already satisfied: defusedxml in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pac
kages (from nbconvert->jupyter->datasist) (0.7.1)
```

```
Requirement already satisfied: jinja2>=3.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from nbconvert->jupyter->datasist) (3.1.2)
Requirement already satisfied: jupyterlab-pygments in c:\users\shiva\appdata\local\programs\python\python311\lib
\site-packages (from nbconvert->jupyter->datasist) (0.2.2)
Requirement already satisfied: markupsafe>=2.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from nbconvert->jupyter->datasist) (2.1.3)
Requirement already satisfied: mistune<4,>=2.0.3 in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from nbconvert->jupyter->datasist) (3.0.1)
Requirement already satisfied: nbclient>=0.5.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from nbconvert->jupyter->datasist) (0.8.0)
Requirement already satisfied: nbformat>=5.7 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-
packages (from nbconvert->jupyter->datasist) (5.9.2)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\users\shiva\appdata\local\programs\python\python311\li
b\site-packages (from nbconvert->jupyter->datasist) (1.5.0)
Requirement already satisfied: tinycss2 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from nbconvert->jupyter->datasist) (1.2.1)
Requirement already satisfied: jupyter-server<3,>=2.4.0 in c:\users\shiva\appdata\local\programs\python\python31
1\lib\site-packages (from notebook->jupyter->datasist) (2.7.2)
Requirement already satisfied: jupyterlab-server<3,>=2.22.1 in c:\users\shiva\appdata\local\programs\python\pyth
on311\lib\site-packages (from notebook->jupyter->datasist) (2.24.0)
Requirement already satisfied: jupyterlab<5,>=4.0.2 in c:\users\shiva\appdata\local\programs\python\python311\li
b\site-packages (from notebook->jupyter->datasist) (4.0.5)
Requirement already satisfied: notebook-shim<0.3,>=0.2 in c:\users\shiva\appdata\local\programs\python\python311
\lib\site-packages (from notebook->jupyter->datasist) (0.2.3)
Requirement already satisfied: ipython-genutils in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from qtconsole->jupyter->datasist) (0.2.0)
Requirement already satisfied: qtpy>=2.0.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from qtconsole->jupyter->datasist) (2.3.1)
Requirement already satisfied: webencodings in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from bleach!=5.0.0->nbconvert->jupyter->datasist) (0.5.1)
Requirement already satisfied: backcall in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from ipython>=7.23.1->ipykernel->jupyter->datasist) (0.2.0)
Requirement already satisfied: decorator in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from ipython>=7.23.1->ipykernel->jupyter->datasist) (5.1.1)
Requirement already satisfied: jedi>=0.16 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pac
kages (from ipython>=7.23.1->ipykernel->jupyter->datasist) (0.19.0)
Requirement already satisfied: pickleshare in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from ipython>=7.23.1->ipykernel->jupyter->datasist) (0.7.5)
Requirement already satisfied: stack-data in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pac
kages (from ipython>=7.23.1->ipykernel->jupyter->datasist) (0.6.2)
Requirement already satisfied: platformdirs>=2.5 in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from jupyter-core!=5.0.*,>=4.12->ipykernel->jupyter->datasist) (3.10.0)
Requirement already satisfied: pywin32>=300 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from jupyter-core!=5.0.*,>=4.12->ipykernel->jupyter->datasist) (306)
Requirement already satisfied: anyio>=3.1.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (3.7.1)
Requirement already satisfied: argon2-cffi in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (23.1.0)
Requirement already satisfied: jupyter-events>=0.6.0 in c:\users\shiva\appdata\local\programs\python\python311\l
ib\site-packages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (0.7.0)
Requirement already satisfied: jupyter-server-terminals in c:\users\shiva\appdata\local\programs\python\python31
1\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (0.4.4)
Requirement already satisfied: overrides in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (7.4.0)
Requirement already satisfied: prometheus-client in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (0.17.1)
Requirement already satisfied: pywinpty in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packa
ges (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (2.0.11)
Requirement already satisfied: send2trash>=1.8.2 in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (1.8.2)
Requirement already satisfied: terminado>=0.8.3 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (0.17.1)
Requirement already satisfied: websocket-client in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (1.6.2)
Requirement already satisfied: async-lru>=1.0.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from jupyterlab<5,>=4.0.2->notebook->jupyter->datasist) (2.0.4)
Requirement already satisfied: jupyter-lsp>=2.0.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\
site-packages (from jupyterlab<5,>=4.0.2->notebook->jupyter->datasist) (2.2.0)
Requirement already satisfied: babel>=2.10 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (2.12.1)
Requirement already satisfied: json5>=0.9.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (0.9.14)
Requirement already satisfied: jsonschema>=4.17.3 in c:\users\shiva\appdata\local\programs\python\python311\lib\
site-packages (from jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (4.19.0)
Requirement already satisfied: requests>=2.28 in c:\users\shiva\appdata\local\programs\python\python311\lib\site
-packages (from jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (2.31.0)
Requirement already satisfied: fastjsonschema in c:\users\shiva\appdata\local\programs\python\python311\lib\site
-packages (from nbformat>=5.7->nbconvert->jupyter->datasist) (2.18.0)
Requirement already satisfied: wcwidth in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packag
es (from prompt-toolkit>=3.0.30->jupyter-console->jupyter->datasist) (0.2.6)
Requirement already satisfied: soupsieve>1.2 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-
packages (from beautifulsoup4->nbconvert->jupyter->datasist) (2.4.1)
Requirement already satisfied: idna>=2.8 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
```

```
ages (from anyio>=3.1.0->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (3.4)
Requirement already satisfied: sniffio>=1.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from anyio>=3.1.0->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (1.3.0)
Requirement already satisfied: parso<0.9.0,>=0.8.3 in c:\users\shiva\appdata\local\programs\python\python311\lib
\site-packages (from jedi>=0.16->ipython>=7.23.1->ipykernel->jupyter->datasist) (0.8.3)
Requirement already satisfied: attrs>=22.2.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-
packages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (23.1.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in c:\users\shiva\appdata\local\programs\pyt
hon\python311\lib\site-packages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datas
ist) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in c:\users\shiva\appdata\local\programs\python\python311\lib
\site-packages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site
-packages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (0.9.2)
Requirement already satisfied: python-json-logger>=2.0.4 in c:\users\shiva\appdata\local\programs\python\python3
11\lib\site-packages (from jupyter-events>=0.6.0->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (2.0.7)
Requirement already satisfied: pyyaml>=5.3 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from jupyter-events>=0.6.0->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (6.0.1)
Requirement already satisfied: rfc3339-validator in c:\users\shiva\appdata\local\programs\python\python311\lib\s
ite-packages (from jupyter-events>=0.6.0->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (0.1.4)
Requirement already satisfied: rfc3986-validator>=0.1.1 in c:\users\shiva\appdata\local\programs\python\python31
1\lib\site-packages (from jupyter-events>=0.6.0->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (0.1.1)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\shiva\appdata\local\programs\python\python31
1\lib\site-packages (from requests>=2.28->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (3.2.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\
site-packages (from requests>=2.28->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\shiva\appdata\local\programs\python\python311\lib\
site-packages (from requests>=2.28->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (2023.7.22)
Requirement already satisfied: argon2-cffi-bindings in c:\users\shiva\appdata\local\programs\python\python311\li
b\site-packages (from argon2-cffi->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (21.2.0)
Requirement already satisfied: executing>=1.2.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from stack-data->ipython>=7.23.1->ipykernel->jupyter->datasist) (1.2.0)
Requirement already satisfied: asttokens>=2.1.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from stack-data->ipython>=7.23.1->ipykernel->jupyter->datasist) (2.2.1)
Requirement already satisfied: pure-eval in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from stack-data->ipython>=7.23.1->ipykernel->jupyter->datasist) (0.2.2)
Requirement already satisfied: fqdn in c:\users\shiva\appdata\local\programs\python\python311\lib\site-packages
(from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (1.5.1)
Requirement already satisfied: isoduration in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (20.11.0)
Requirement already satisfied: jsonpointer>1.13 in c:\users\shiva\appdata\local\programs\python\python311\lib\si
te-packages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (2.4)
Requirement already satisfied: uri-template in c:\users\shiva\appdata\local\programs\python\python311\lib\site-p
ackages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (1.3.0)
Requirement already satisfied: webcolors>=1.11 in c:\users\shiva\appdata\local\programs\python\python311\lib\sit
e-packages (from jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (1.13)
Requirement already satisfied: cffi>=1.0.1 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pa
ckages (from argon2-cffi-bindings->argon2-cffi->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist) (1.15.1)
Requirement already satisfied: pycparser in c:\users\shiva\appdata\local\programs\python\python311\lib\site-pack
ages (from cffi>=1.0.1->argon2-cffi-bindings->argon2-cffi->jupyter-server<3,>=2.4.0->notebook->jupyter->datasist
) (2.21)
Requirement already satisfied: arrow>=0.15.0 in c:\users\shiva\appdata\local\programs\python\python311\lib\site-
packages (from isoduration->jsonschema>=4.17.3->jupyterlab-server<3,>=2.22.1->notebook->jupyter->datasist) (1.2.
3)
[notice] A new release of pip is available: 23.1.2 -> 23.3.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [169…
```
from datasist.structdata import detect_outliers
outliers_indices = detect_outliers(df, 0 ,list(df.select_dtypes(exclude="object").columns))
len(outliers_indices)
```

Out[169… 8069

In [ ]:

In [170…
```
df.drop(outliers_indices, inplace=True)
df
```

Out[170…

	ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Acc
0	0x1602	CUS_0xd40	January	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	
1	0x1603	CUS_0xd40	February	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	
3	0x1605	CUS_0xd40	April	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	
4	0x1606	CUS_0xd40	May	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	
5	0x1607	CUS_0xd40	June	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	
...	...	...	...	...	...	...	...	...	...	
99994	0x25fe8	CUS_0x942c	March	Nicks	25	078-73-5990	Mechanic	39628.99	3359.415833	
99995	0x25fe9	CUS_0x942c	April	Nicks	25	078-73-5990	Mechanic	39628.99	3359.415833	
99996	0x25fea	CUS_0x942c	May	Nicks	25	078-73-5990	Mechanic	39628.99	3359.415833	
99998	0x25fec	CUS_0x942c	July	Nicks	25	078-73-5990	Mechanic	39628.99	3359.415833	
99999	0x25fed	CUS_0x942c	August	Nicks	25	078-73-5990	Mechanic	39628.99	3359.415833	

58160 rows × 28 columns

In [171…
```
df.select_dtypes(include="object").columns
```

Out[171…
```
Index(['ID', 'Customer_ID', 'Month', 'Name', 'SSN', 'Occupation',
       'Type_of_Loan', 'Credit_Mix', 'Credit_History_Age',
       'Payment_of_Min_Amount', 'Payment_Behaviour', 'Credit_Score'],
      dtype='object')
```

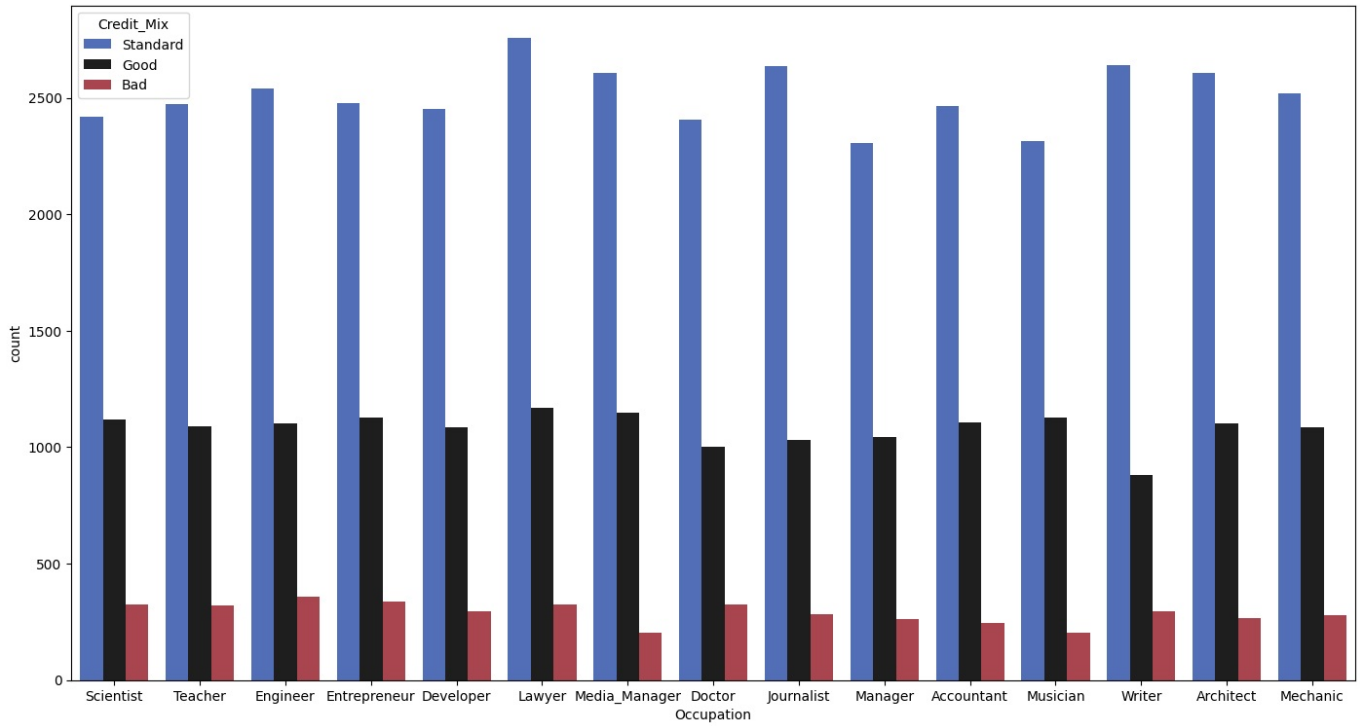
In [172…
```
df.select_dtypes(exclude="object").columns
```

Out[172…
```
Index(['Age', 'Annual_Income', 'Monthly_Inhand_Salary', 'Num_Bank_Accounts',
       'Num_Credit_Card', 'Interest_Rate', 'Num_of_Loan',
       'Delay_from_due_date', 'Num_of_Delayed_Payment', 'Changed_Credit_Limit',
       'Num_Credit_Inquiries', 'Outstanding_Debt', 'Credit_Utilization_Ratio',
       'Total_EMI_per_month', 'Amount_invested_monthly', 'Monthly_Balance'],
      dtype='object')
```

(7) Bivariate Categorical Analysis

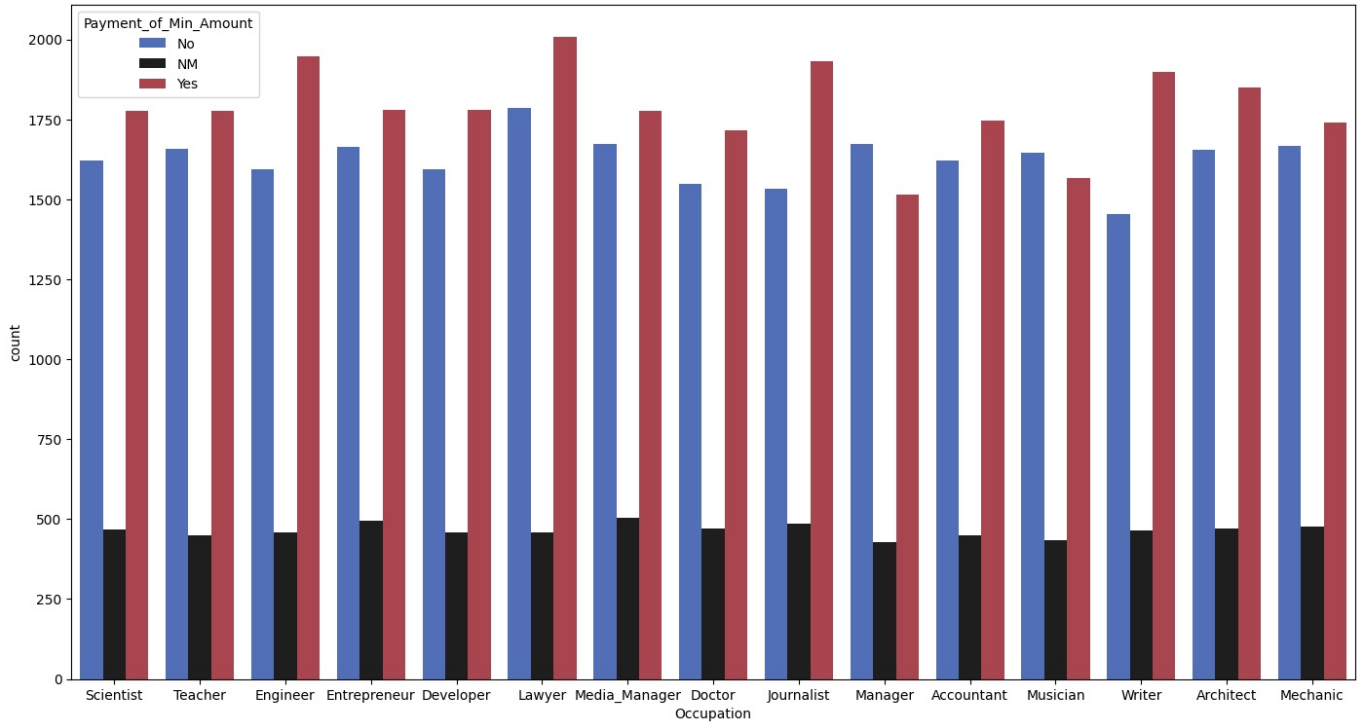
In [173…
```
fig = plt.figure(figsize= (17,9))
sns.countplot(data=df,x="Occupation",hue="Credit_Mix",palette="icefire")
```

Out[173…
```
<Axes: xlabel='Occupation', ylabel='count'>
```



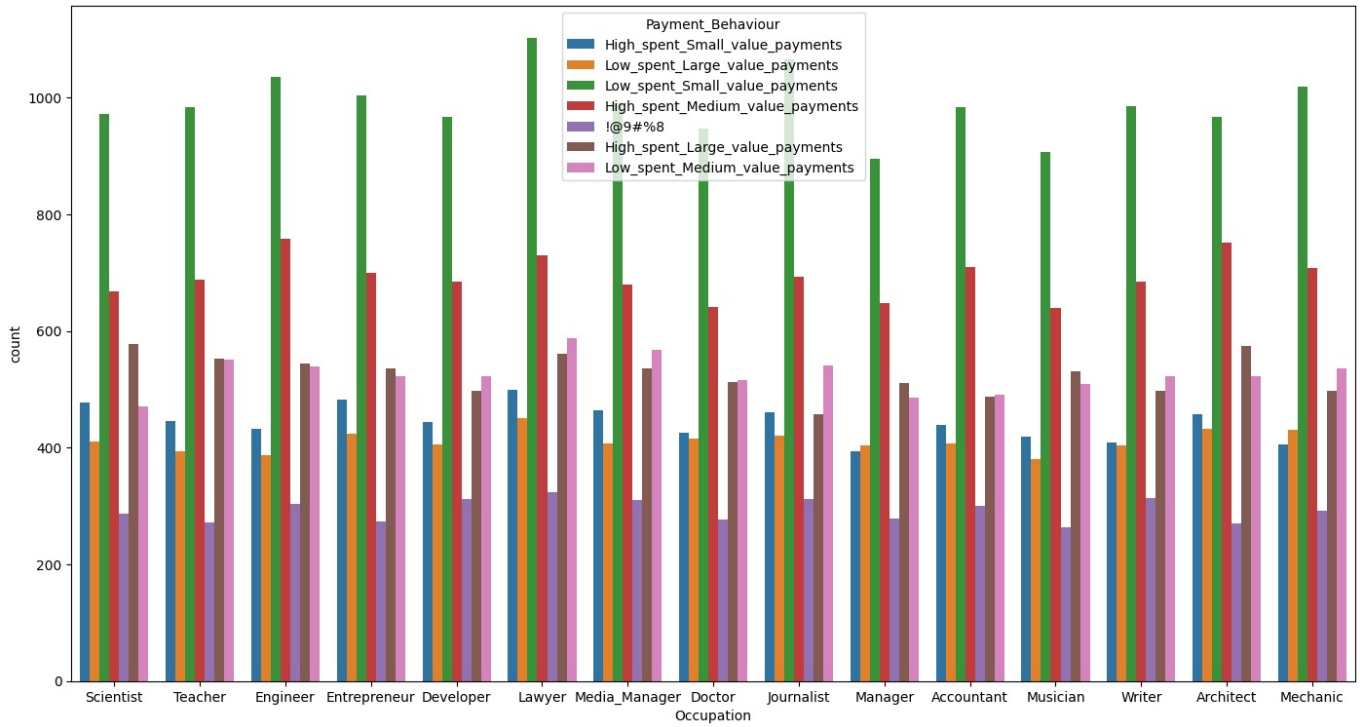
In [174…
```
fig = plt.figure(figsize= (17,9))
sns.countplot(data=df,x="Occupation",hue="Payment_of_Min_Amount",palette="icefire")
```

Out[174… <Axes: xlabel='Occupation', ylabel='count'>



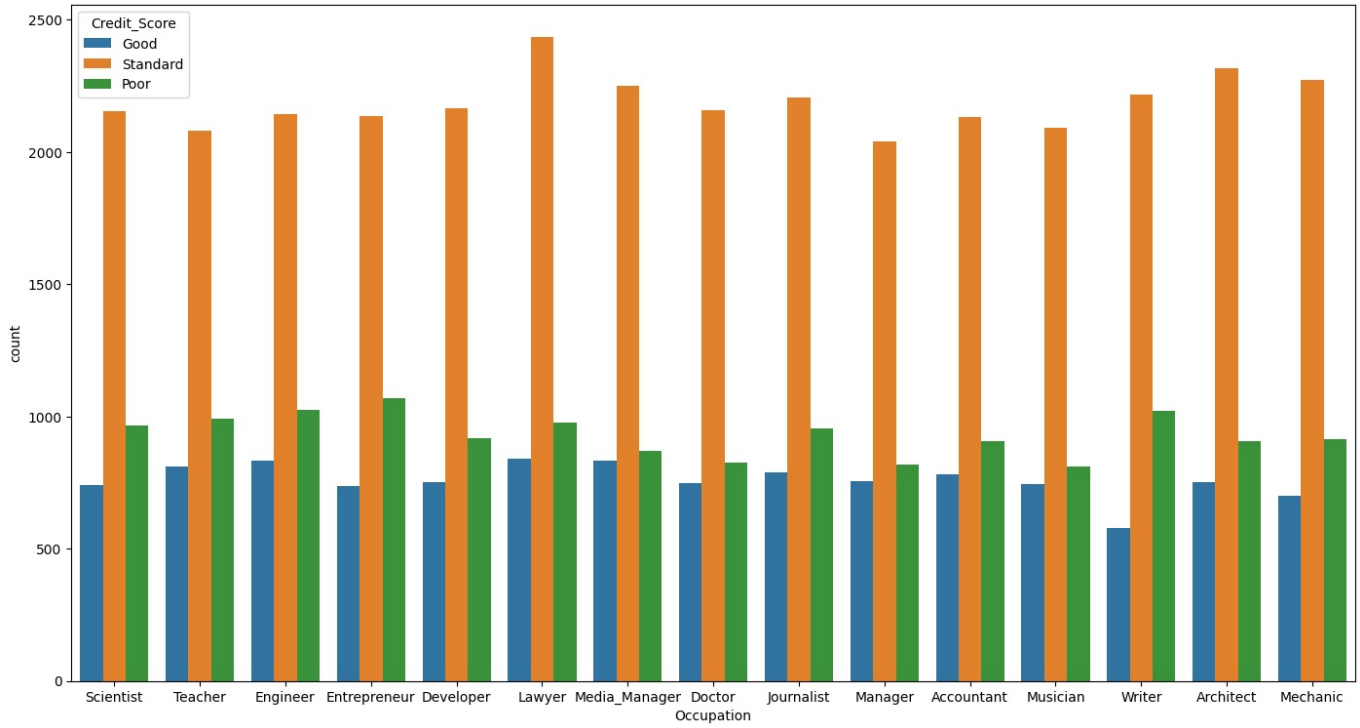
In [175…
```
fig = plt.figure(figsize= (17,9))
sns.countplot(data=df,x="Occupation",hue="Payment_Behaviour")
```

Out[175… <Axes: xlabel='Occupation', ylabel='count'>



In [176…
```
fig = plt.figure(figsize= (17,9))
sns.countplot(data=df,x="Occupation",hue="Credit_Score")
```

Out[176… <Axes: xlabel='Occupation', ylabel='count'>



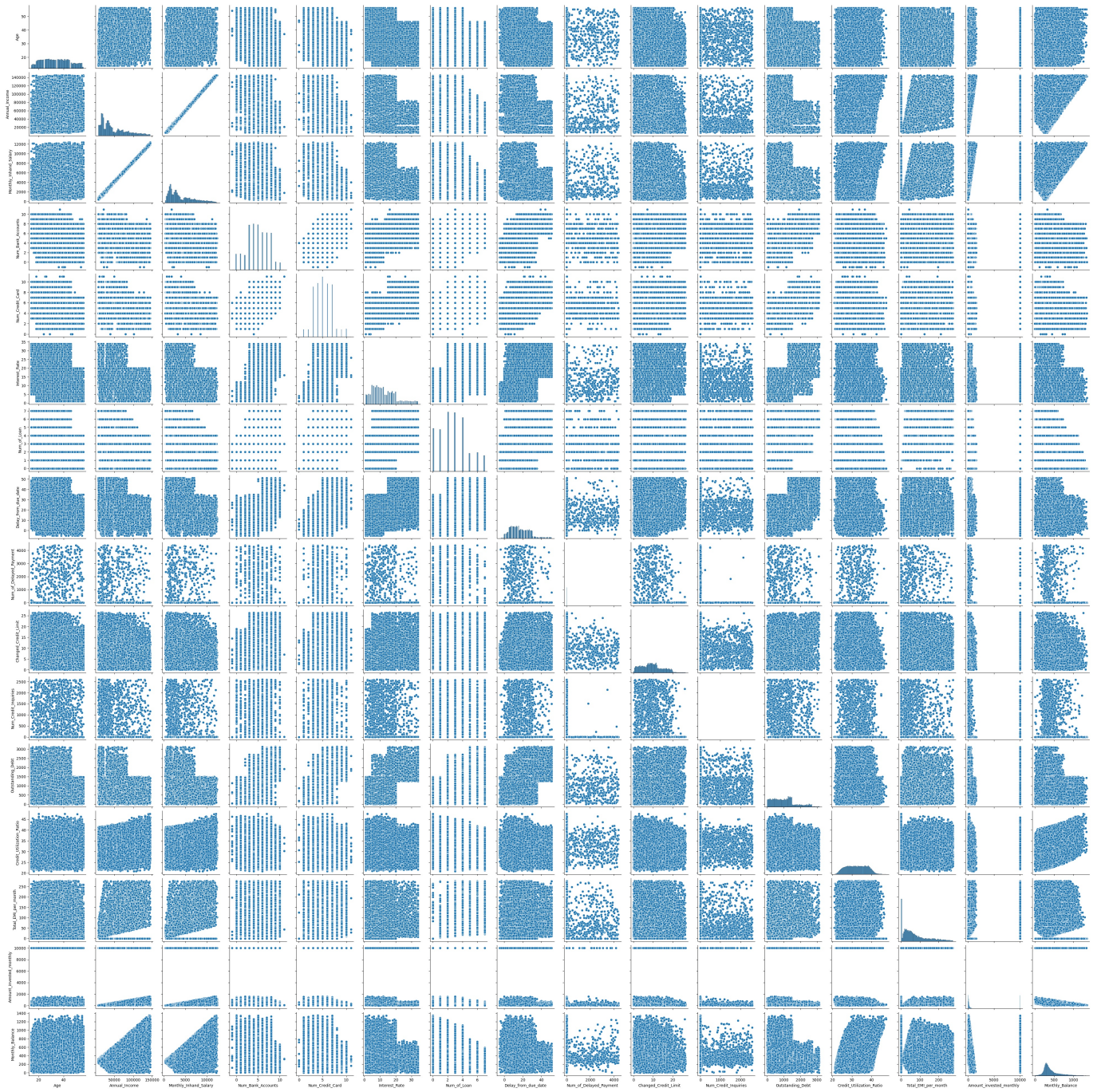
Bivariate Numerical Analysis

In [177…
```
df.select_dtypes(exclude="object").columns
```

Out[177…
```
Index(['Age', 'Annual_Income', 'Monthly_Inhand_Salary', 'Num_Bank_Accounts',
       'Num_Credit_Card', 'Interest_Rate', 'Num_of_Loan',
       'Delay_from_due_date', 'Num_of_Delayed_Payment', 'Changed_Credit_Limit',
       'Num_Credit_Inquiries', 'Outstanding_Debt', 'Credit_Utilization_Ratio',
       'Total_EMI_per_month', 'Amount_invested_monthly', 'Monthly_Balance'],
      dtype='object')
```

In [178…
```
plt.figure
sns.pairplot(data=df)
```

Out[178… <seaborn.axisgrid.PairGrid at 0x1dfaa607d50>



Machine learning

In [183…
```
df.drop(["ID","Customer_ID","Month","Name","SSN","Credit_History_Age","Type_of_Loan"],axis = 1,inplace = True)
```

In [184…
```
x=df.drop(columns="Credit_Score")
y=df["Credit_Score"]
```

In [185…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 58160 entries, 0 to 99999
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       58160 non-null  int32
 1   Occupation                58160 non-null  object
 2   Annual_Income             58160 non-null  float64
 3   Monthly_Inhand_Salary     49443 non-null  float64
 4   Num_Bank_Accounts         58160 non-null  int64
 5   Num_Credit_Card           58160 non-null  int64
 6   Interest_Rate             58160 non-null  int64
 7   Num_of_Loan               58160 non-null  int32
 8   Delay_from_due_date       58160 non-null  int64
 9   Num_of_Delayed_Payment    54096 non-null  float64
 10  Changed_Credit_Limit      58160 non-null  float64
 11  Num_Credit_Inquiries      57021 non-null  float64
 12  Credit_Mix                58160 non-null  object
 13  Outstanding_Debt          58160 non-null  float64
 14  Credit_Utilization_Ratio  58160 non-null  float64
 15  Payment_of_Min_Amount     58160 non-null  object
 16  Total_EMI_per_month       58160 non-null  float64
 17  Amount_invested_monthly   55595 non-null  float64
 18  Payment_Behaviour         58160 non-null  object
 19  Monthly_Balance           57804 non-null  float64
 20  Credit_Score              58160 non-null  object
dtypes: float64(10), int32(2), int64(4), object(5)
memory usage: 9.3+ MB
```

In [186…
```
y.value_counts()* 100 / len(df)
```

Out[186…
```
Credit_Score
Standard    56.394429
Poor        24.019945
Good        19.585626
Name: count, dtype: float64
```

In [187…
```
from sklearn.model_selection import train_test_split
x_train ,x_test ,y_train,y_test = train_test_split(x,y,test_size = 0.2,random_state=42,stratify=y)
```

In [219…
```
numeric_columns = x_train.select_dtypes(exclude='object').columns
numeric_columns
```

Out[219…
```
Index(['Age', 'Annual_Income', 'Monthly_Inhand_Salary', 'Num_Bank_Accounts',
       'Num_Credit_Card', 'Interest_Rate', 'Num_of_Loan',
       'Delay_from_due_date', 'Num_of_Delayed_Payment', 'Changed_Credit_Limit',
       'Num_Credit_Inquiries', 'Outstanding_Debt', 'Credit_Utilization_Ratio',
       'Total_EMI_per_month', 'Amount_invested_monthly', 'Monthly_Balance'],
      dtype='object')
```

In [220…
```
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
numeric_feature = Pipeline(steps=[('handlingmissing',SimpleImputer(strategy='median')),('scaling',StandardScale
```

In [221…
```
cat_columns = x_train.select_dtypes(include='object').columns
cat_columns
```

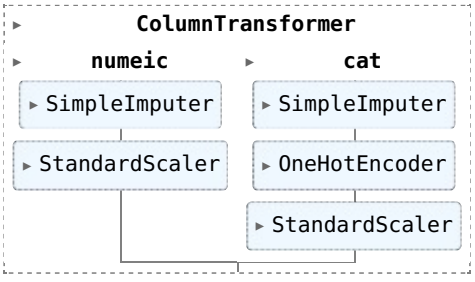
Out[221…
```
Index(['Occupation', 'Credit_Mix', 'Payment_of_Min_Amount',
       'Payment_Behaviour'],
      dtype='object')
```

In [222…
```
from sklearn.preprocessing import OneHotEncoder
cat_feature = Pipeline(steps = [('missing',SimpleImputer(strategy='most_frequent')),('encoding',OneHotEncoder()
```

In [223…
```
from sklearn.compose import ColumnTransformer
processing = ColumnTransformer([('numeic',numeric_feature,numeric_columns),
                                ('cat', cat_feature, cat_columns)])
```

In [224…
```
processing
```

Out[224…

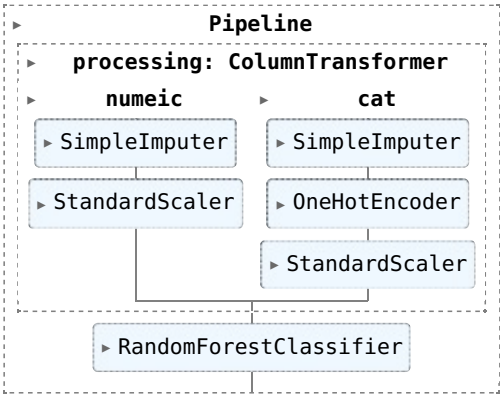


In [225…
```
from sklearn.ensemble import RandomForestClassifier
```

In [226…
```
final_pipe= Pipeline(steps = [('processing',processing),('modeling',RandomForestClassifier())])
```

In [227…
```
final_pipe.fit(x_train,y_train)
```

Out[227…



In [228…
```
from sklearn.metrics import confusion_matrix, classification_report
y_pred =final_pipe.predict(x_test)
```

In [229…
```
print('The confusion matrix is :','\n' , confusion_matrix(y_test,y_pred))
```

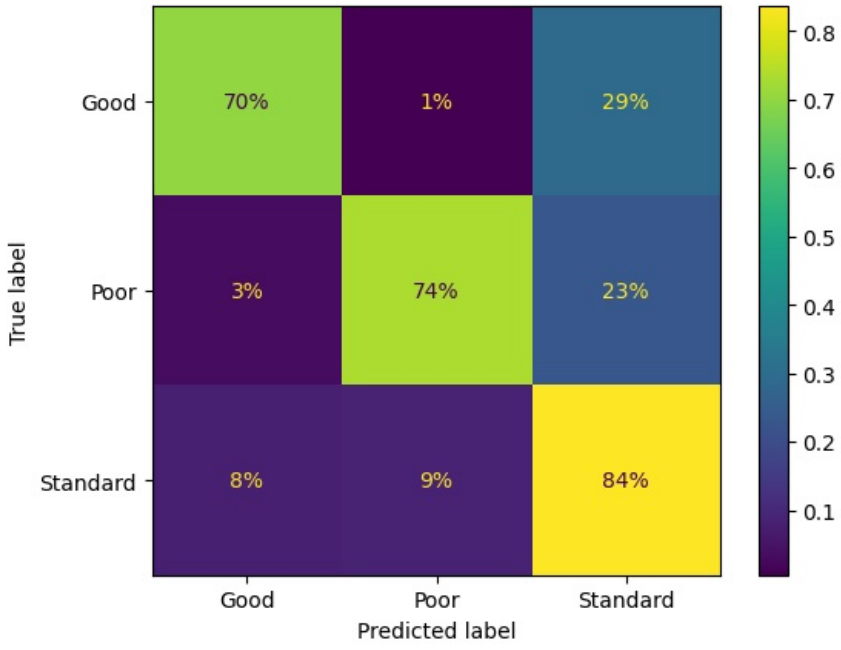
```
The confusion matrix is :
 [[1601   12  665]
 [  91 2054  649]
 [ 515  559 5486]]
```

In [230…
```
print('The classification report is :' ,'\n',classification_report(y_test,y_pred))
```

```
The classification report is :
              precision    recall  f1-score   support

        Good       0.73      0.70      0.71      2278
        Poor       0.78      0.74      0.76      2794
    Standard       0.81      0.84      0.82      6560

    accuracy                           0.79     11632
   macro avg       0.77      0.76      0.76     11632
weighted avg       0.78      0.79      0.79     11632
```

In [236…
```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
ConfusionMatrixDisplay.from_predictions(y_test,y_pred,normalize="true",values_format='.0%')
```

Out[236… `<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1dfe1c7ee50>`



In [ ]: