CREDIT ONE

PREPARE AND EXPLORE DATA -PYTHON

SUGITHA DEVARAJAN

OVERALL PROBLEM

1. *Increase in customer default rates - This is bad for Credit One since we approve the customers for loans in the first place.*
2. *Revenue and customer loss for clients and, eventually, loss of clients for Credit One*

Investigative Questions:

1. *How do you ensure that customers can/will pay their loans? Can we do this?*

SOME LESSONS LEARNT:

1. *We cannot control customer spending habits*
2. *We cannot always go from what we find in our analysis to the underlying "why"*
3. *We must on the problem(s) we can solve: What attributes in the data can we deem to be statistically significant to the problem at hand?*
4. *What concrete information can we derive from the data we have?*
5. *What proven methods can we use to uncover more information and why?*

IDENTIFY WHICH CUSTOMER ATTRIBUTES RELATE SIGNIFICANTLY TO DEFAULT RATE

STEPS TAKEN TO ANALYZE AND PERFORM DATA CLEANUP:

1. *Imported required libraries – pandas, numpy, matplotlib, pandas_profiling etc.*
2. *Imported the data from SQL*
3. *Initial analysis showed (data.head()) that the data need clean up*
4. *Removed Header (invalid row) and the second row was the actual header.*
5. *Found 202 rows of duplicate – removed duplicates*
6. *Dropped ID column since that will cause unnecessary interference in visualization and when pivoting/grouping.*
7. *With data.info found the data was not all integer, some where object. Decided to use **label encoder** to make all column values integer.*

DATA EXPLORATION TO IDENTIFY SIGNIFICANT ATTRIBUTES TO DEFAULT RATE

METHODS USED:

- *PANDAS PROFILING*
- *SEABORN CHARTS – PAIRPLOT, CATPLOT, FACET GRID ETC*
- *BOX PLOTS*
- *HISTOGRAM*
- *GROUPING*
- *FILTER*
- *PIVOT*
- *DISCRETIZATION*

Data.info after cleanup

```
In [40]: #All data is now integer - perfect for ML
         data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 0 to 30200
Data columns (total 24 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   LIMIT_BAL                   30000 non-null  int64
 1   SEX                         30000 non-null  int64
 2   EDUCATION                   30000 non-null  int64
 3   MARRIAGE                    30000 non-null  int64
 4   AGE                         30000 non-null  int64
 5   PAY_0                       30000 non-null  int64
 6   PAY_2                       30000 non-null  int64
 7   PAY_3                       30000 non-null  int64
 8   PAY_4                       30000 non-null  int64
 9   PAY_5                       30000 non-null  int64
 10  PAY_6                       30000 non-null  int64
 11  BILL_AMT1                   30000 non-null  int64
 12  BILL_AMT2                   30000 non-null  int64
 13  BILL_AMT3                   30000 non-null  int64
 14  BILL_AMT4                   30000 non-null  int64
 15  BILL_AMT5                   30000 non-null  int64
 16  BILL_AMT6                   30000 non-null  int64
 17  PAY_AMT1                    30000 non-null  int64
 18  PAY_AMT2                    30000 non-null  int64
 19  PAY_AMT3                    30000 non-null  int64
 20  PAY_AMT4                    30000 non-null  int64
 21  PAY_AMT5                    30000 non-null  int64
 22  PAY_AMT6                    30000 non-null  int64
 23  default payment next month  30000 non-null  int64
dtypes: int64(24)
memory usage: 5.7 MB
```

Attribute info after label encoding:

Attribute Information

Attribute	Description
SEX	1 is Male and 0 is Female
EDUCATION	0 is graduate School, 3 is University , 1 is High School and 2 is others
MARRIAGE	1 is Married , 2 is single, 3 is divorce , and 0 is others
Pay_0 to Pay_6	History of Past Payment from April 2005 to September 2005. Value meaning: -2 = No consumption, -1= Paid in full, 0 = The use of revolving credit, 1 = Payment delay for one month , 2 = Payment delay for two month etc...9 = Payment delay for 9 months
Bill_AMT	Amount of Bill Statement
Pay_AMT	Amount of previous Payment
Defaulted	1 = not defaulted and 0 = defaulted

💡 Artificial neural network is the only one that can accurately estimate the real probability of default

PANDAS PROFILING:

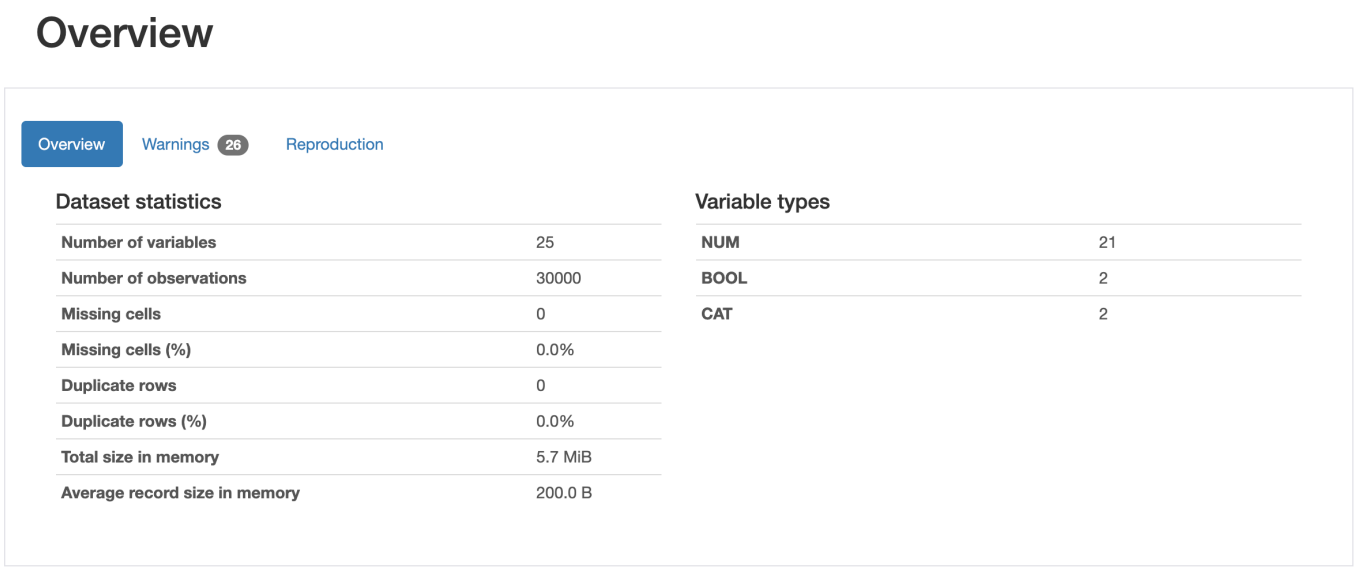
Generates profile reports from a panda Data Frame. The pandas df.describe() function is great but a little basic for serious exploratory data analysis. pandas_profiling extends the pandas Data Frame with df. profile_report () for quick data analysis.

For each column the following statistics - if relevant for the column type - are presented in an interactive HTML report:

- **Type inference**: detect the types of columns in a Data frame.
- **Essentials**: type, unique values, missing values
- **Quantile statistics** like minimum value, Q1, median, Q3, maximum, range, interquartile range
- **Descriptive statistics** like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- **Most frequent values**
- **Histograms**
- **Correlations** highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
- **Missing values** matrix, count, heatmap and dendrogram of missing values
- **Duplicate rows** Lists the most occurring duplicate rows
- **Text analyses** learn about categories (Uppercase, Space), scripts (Latin, Cyrillic) and blocks (ASCII) of text data

Some important observations of credit one data from pandas profiling:

- There are no missing values or duplicates
- Variables types are numerical, Boolean (Sex and default), and categorical (education and marriage).
- Limit balance – min is 10,000 and max is 1,000,000
- Past history gives some valid data with high percentage of zeros. Zeros means use of revolving credit.
- Bill amount seems to have high correlations

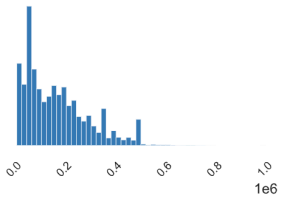


LIMIT_BAL

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	81
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	167484.3227
Minimum	10000
Maximum	1000000
Zeros	0
Zeros (%)	0.0%
Memory size	234.4 KiB



Toggle details

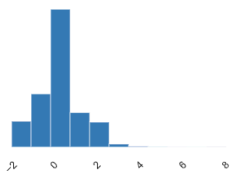
PAY_0

Real number (ℝ)

ZEROS

Distinct	11
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	-0.0167
Minimum	-2
Maximum	8
Zeros	14737
Zeros (%)	49.1%
Memory size	234.4 KiB



Toggle details

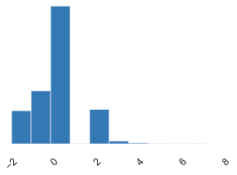
PAY_2

Real number (ℝ)

ZEROS

Distinct	11
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	-0.1337666667
Minimum	-2
Maximum	8
Zeros	15730
Zeros (%)	52.4%
Memory size	234.4 KiB



Toggle details

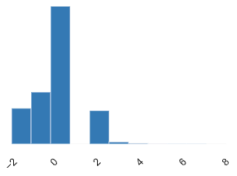
PAY_3

Real number (ℝ)

ZEROS

Distinct	11
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	-0.1662
Minimum	-2
Maximum	8
Zeros	15764
Zeros (%)	52.5%
Memory size	234.4 KiB



Toggle details

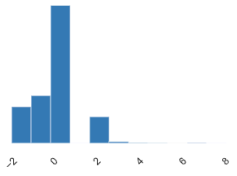
PAY_4

Real number (ℝ)

ZEROS

Distinct	11
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	-0.2206666667
Minimum	-2
Maximum	8
Zeros	16455
Zeros (%)	54.9%
Memory size	234.4 KiB



Toggle details

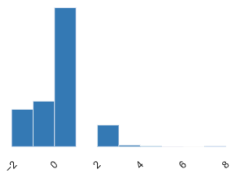
PAY_5

Real number (ℝ)

ZEROS

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	-0.2662
Minimum	-2
Maximum	8
Zeros	16947
Zeros (%)	56.5%
Memory size	234.4 KiB



Toggle details

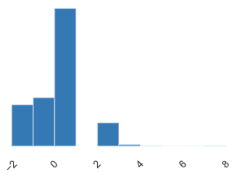
PAY_6

Real number (ℝ)

ZEROS

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	-0.2911
Minimum	-2
Maximum	8
Zeros	16286
Zeros (%)	54.3%
Memory size	234.4 KiB



Toggle details

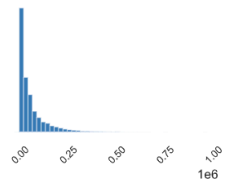
BILL_AMT1

Real number (ℝ)

HIGH CORRELATION
ZEROS

Distinct	22723
Distinct (%)	75.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	51223.3309
Minimum	-165580
Maximum	964511
Zeros	2008
Zeros (%)	6.7%
Memory size	234.4 KiB



Toggle details

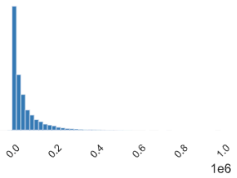
BILL_AMT2

Real number (ℝ)

HIGH CORRELATION
ZEROS

Distinct	22346
Distinct (%)	74.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	49179.07517
Minimum	-69777
Maximum	983931
Zeros	2506
Zeros (%)	8.4%
Memory size	234.4 KiB



Toggle details

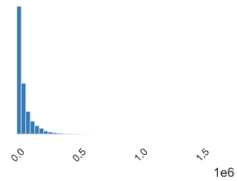
BILL_AMT3

Real number (ℝ)

HIGH CORRELATION
ZEROS

Distinct	22026
Distinct (%)	73.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	47013.1548
Minimum	-157264
Maximum	1664089
Zeros	2870
Zeros (%)	9.6%
Memory size	234.4 KiB



Toggle details

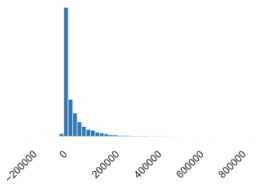
BILL_AMT4

Real number (ℝ)

HIGH CORRELATION
ZEROS

Distinct	21548
Distinct (%)	71.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	43262.94897
Minimum	-170000
Maximum	891586
Zeros	3195
Zeros (%)	10.7%
Memory size	234.4 KiB



Toggle details

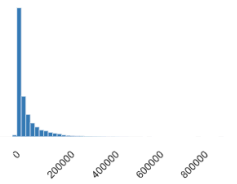
BILL_AMT5

Real number (ℝ)

HIGH CORRELATION
ZEROS

Distinct	21010
Distinct (%)	70.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	40311.40097
Minimum	-81334
Maximum	927171
Zeros	3506
Zeros (%)	11.7%
Memory size	234.4 KiB



Toggle details

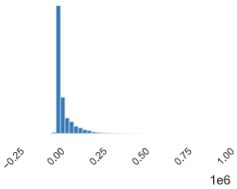
BILL_AMT6

Real number (ℝ)

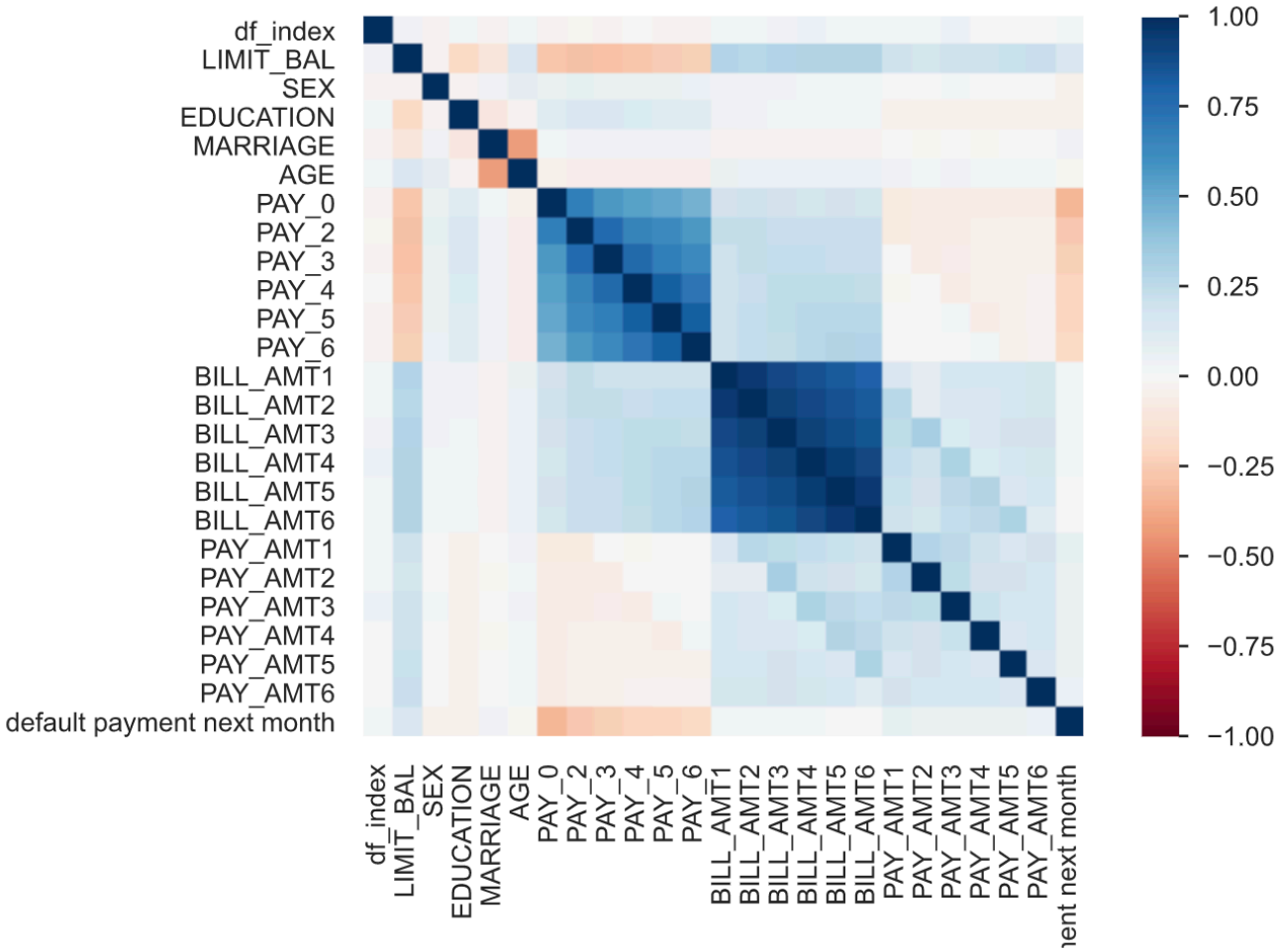
HIGH CORRELATION
ZEROS

Distinct	20604
Distinct (%)	68.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	38871.7604
Minimum	-339603
Maximum	961664
Zeros	4020
Zeros (%)	13.4%
Memory size	234.4 KiB

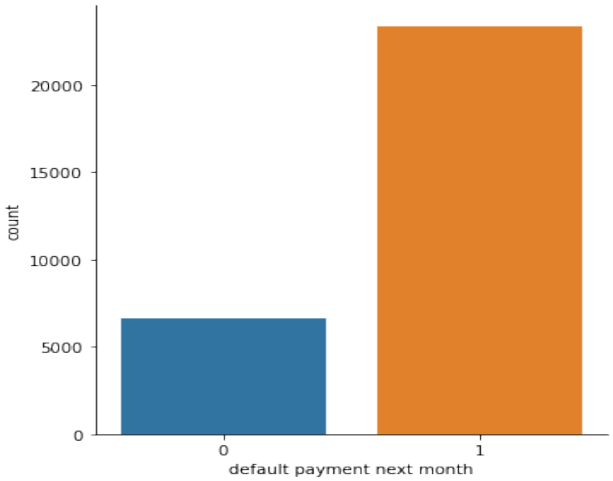


Toggle details

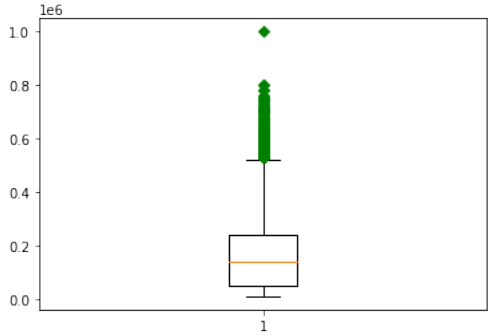


Other Observations from EDA:

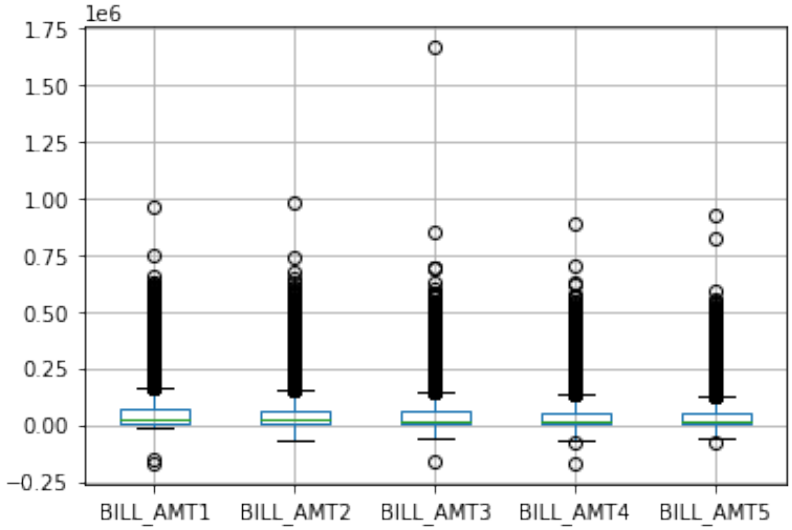
1. We have less data on default



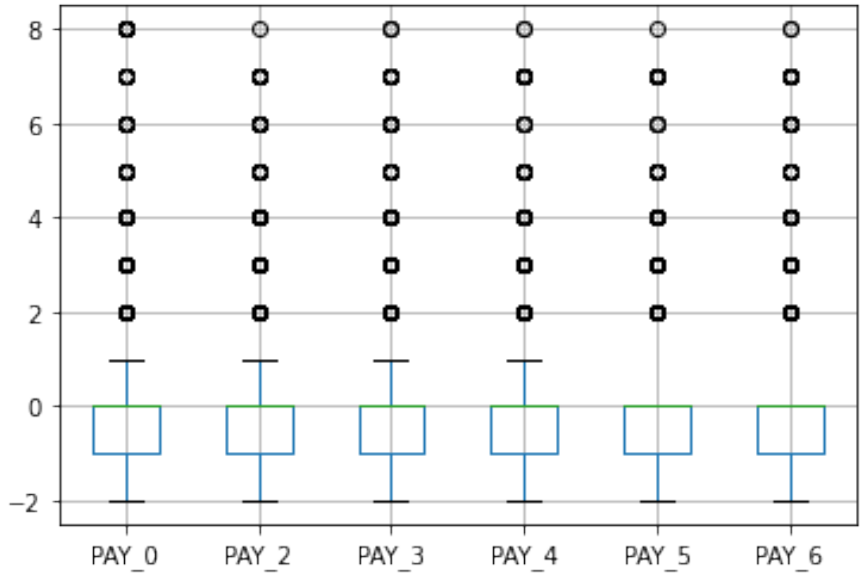
2. Found some anomalies in Limit_Bal using box plot after $600K



3. Bill amount has negative value.

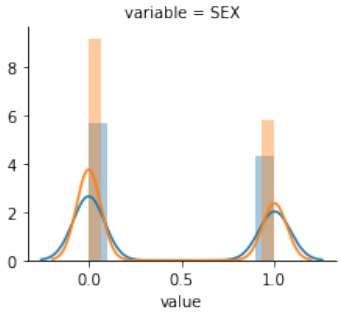


4. Pay_X seems to be important with values greater than 0 which can help predict default.



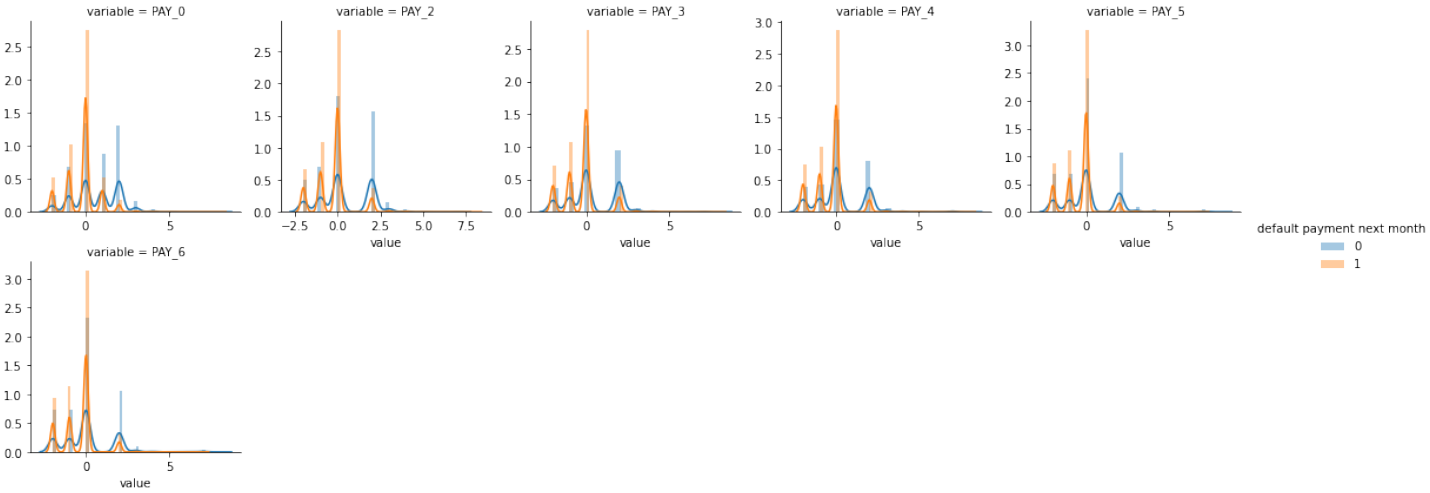
5. Facet Grid - which is one of the most interesting functions in the Seaborn library!
   It allows you to visualize data sets with lots of columns especially categorical columns

   - Which sex defaults more?
     Not defaulted are more by females than males



- Limit_Bal
  Anything over 750K to 1M is not defaulted.
  All default is in the lower limit balance which is less than 250K

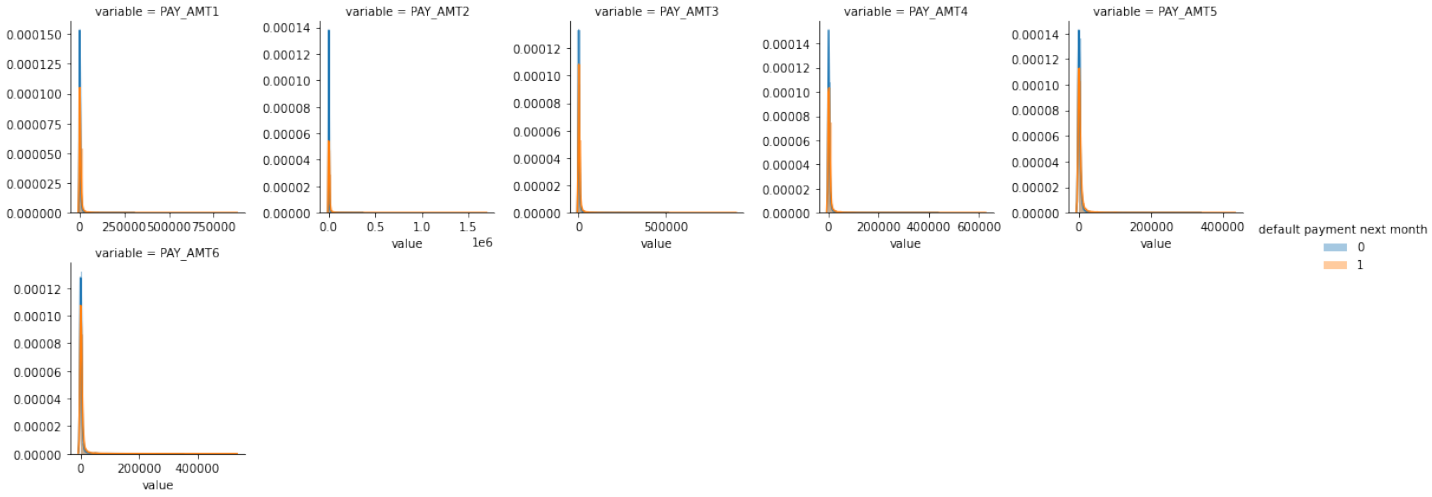
- PAY_X
  Not default is seen more when the Pay_X is greater than 0 meaning who missed payments.
  Month over month it has slowly increased if you see from April to September



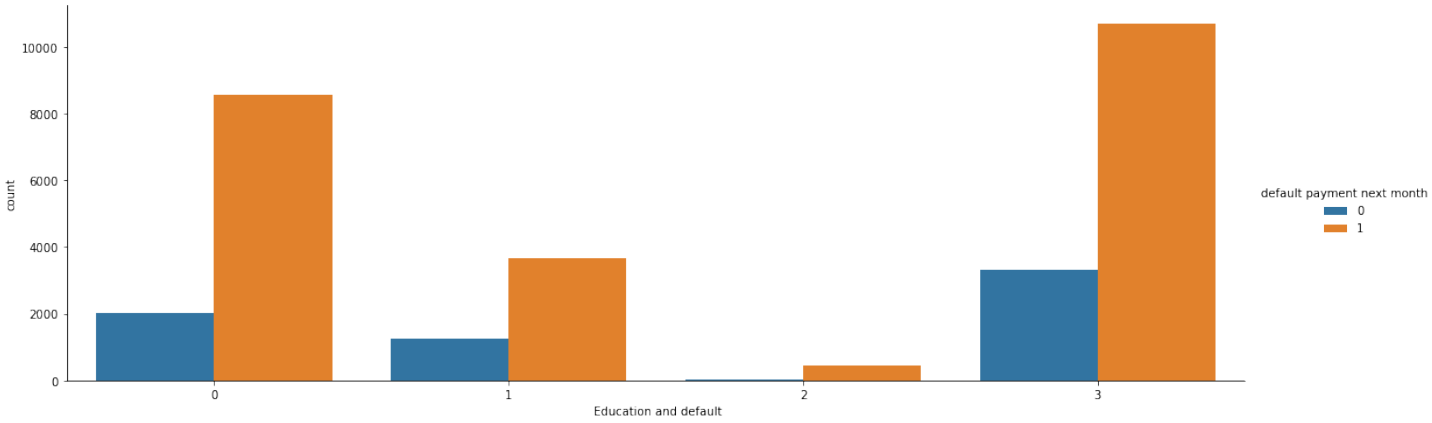
- AGE
  Age 30 to 40 are high in not defaulting

- PAY_AMTX

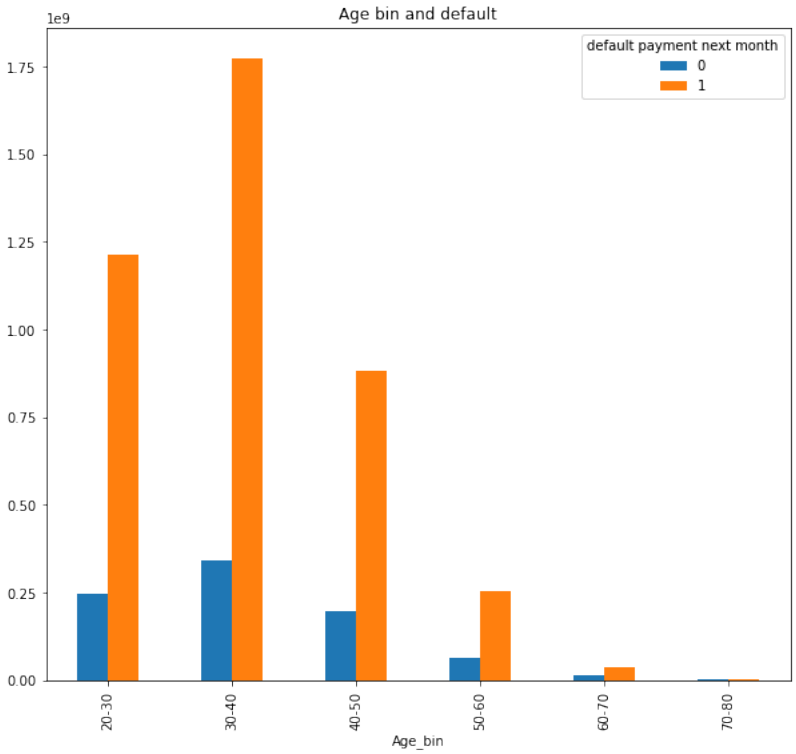
Pay_AMT zeros are in more % than other values and tends to default.



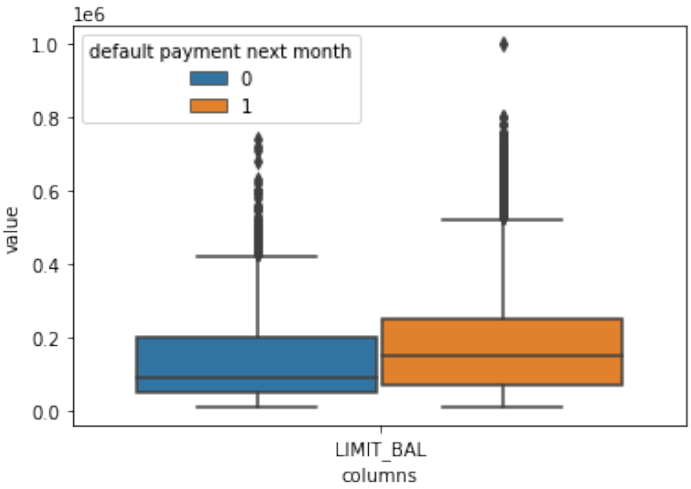
6. With catplot we can confidently say that when customers are well educated then they don't default.
   (0 is graduate School, 3 is University, 1 is High School and 2 is others)



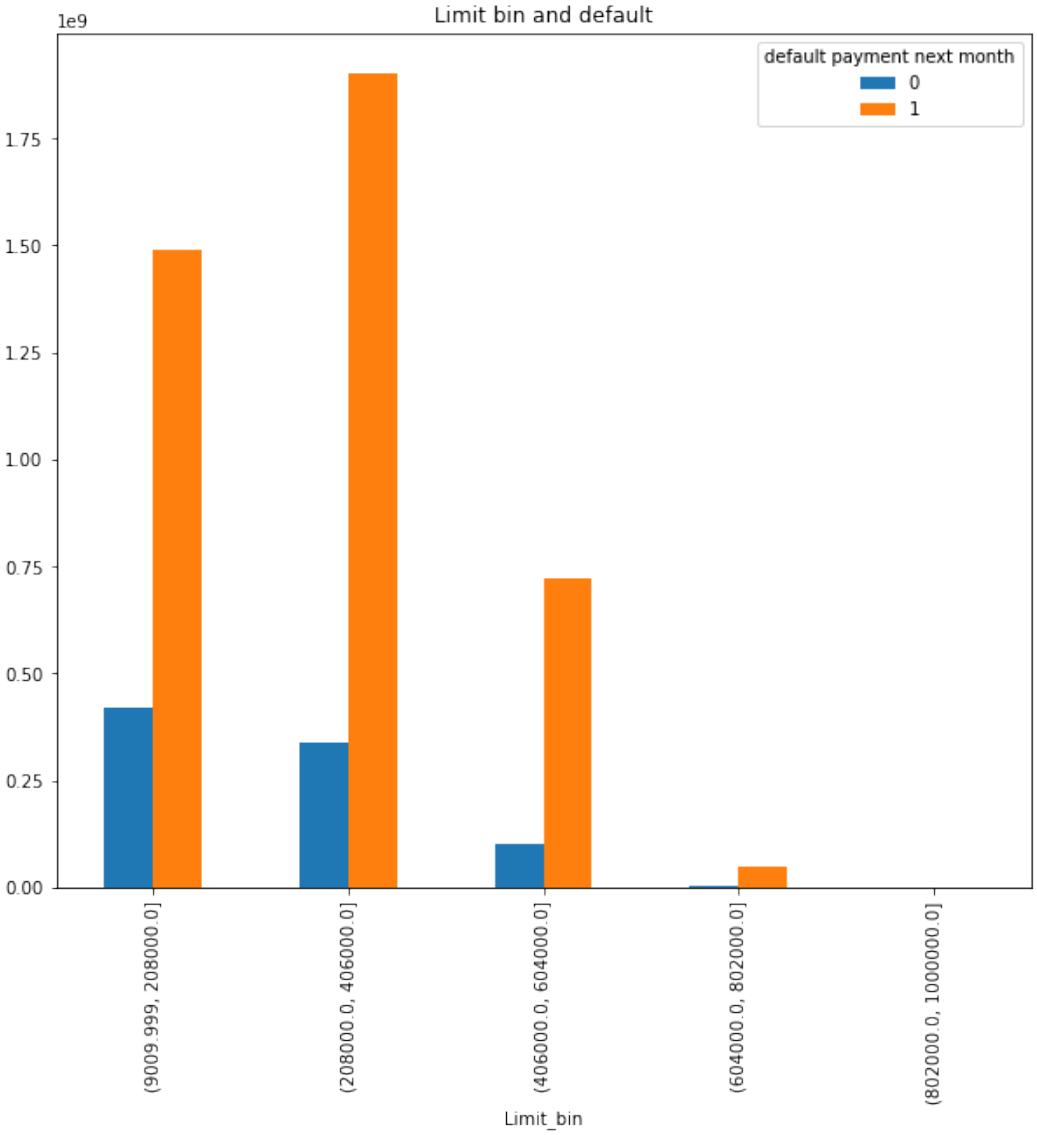
7. Age Discretization – When age is divided into bins I see the default and not default behave very similar.



8. Limit_Bal have more defaults in the lower values.



9. Limit_Bal discretization – Again this says the same thing as box plot above where the default is more in the lower limit balance.



Report Focus:

Did you learn anything of potential business value from this analysis?

Yes, I found that even though correlations are week among variables there are some with high correlation like the Bill_Amt which need to be used in my next step of data modeling to predict. I also found Limit_BAL and PAY_X are other variables which will help me predict in my data modeling.

What are the main lessons you've learned from this experience?

I learnt about pandas profiling and Facet grid along with pd.melt. I also explored about correlation interpretation where I found about the significance test. I need to understand more about the hypothesis test.
https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8

What recommendations would you give based on your findings?

To predict the default customers, we need to model using the following feature variables LIMIT_BAL, PAY-X, BILL_AMTX, and PAY_AMTX