Connecting the dots

Understanding loan applicant information collected by US credit bureau

Yuki MATSUNO

Based on Give Me Some Credit competition from Kaggle

Objective

- Data contains loan applicant information collected by a US credit bureau.
- Each row represents a loan application & info gathered on the applicant at the time of the application.
- The target is to understand the correlation between person who experienced financial distress in two years & the other features .

SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans
1	0.766127	45	2	0.802982	9120.0	13
0	0.957151	40	0	0.121876	2600.0	4
0	0.658180	38	1	0.085113	3042.0	2
0	0.233810	30	0	0.036050	3300.0	5
0	0.907239	49	1	0.024926	63588.0	7

Understanding the Data

- Data has 12 columns. Info on how columns are correlated is not provided & have to be inferred by data scientist.This dataset contains 150,000 data points.
- Column 1: 'Unnamed: 0' does not contain any information but index & hence can be dropped.
- **Handling null values**: Most columns have values, but some columns like 'MonthlyIncome' and 'NumberOfDependents' have no value ('null').

  Example:

  - 19.6 % of MonthlyIncome is null.
  - 2.6% of NumberOfDependents is null

- **Handling duplicate rows**: 0.41% (609 rows/ 150,000) Most of duplicated data points have no value ("NaN") in "MonthlyIncome" column. These data points seem less trustworthy (perhaps data entry errors). These columns are dropped from the dataset.
- **Handling outliers**: There are many big outliers in certain columns (discussed in detail later)
- **Identifying correlation among columns**: Some columns are correlated in initial data

Handling null values

Column Name	Non-Null Count	Dtype
SeriousDlqin2yrs	150,000 non-null	int64
RevolvingUtilizationOfUnsecuredLines	150,000 non-null	float64
age	150,000 non-null	int64
NumberOfTime30to59DaysPastDueNotWorse	150,000 non-null	int64
DebtRatio	150,000 non-null	float64
MonthlyIncome	120,269 non-null	float64
NumberOfOpenCreditLinesAndLoans	150,000 non-null	int64
NumberOfTimes90DaysLate	150,000 non-null	int64
NumberRealEstateLoansOrLines	150,000 non-null	int64
NumberOfTime60to89DaysPastDueNotWorse	150,000 non-null	int64
NumberOfDependents	146,076 non-null	float64

Handling Duplicates

	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30to59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCredit
1669	0	1.0	29	0	0.0	NaN	
7823	0	1.0	29	0	0.0	NaN	
7920	0	1.0	22	0	0.0	820.0	
8840	0	1.0	23	0	0.0	820.0	
10869	0	1.0	73	0	0.0	NaN	
14067	0	0.0	45	0	0.0	NaN	
14465	0	1.0	23	0	0.0	NaN	
14874	0	1.0	71	0	0.0	NaN	
15346	0	0.0	48	0	0.0	NaN	
15544	0	1.0	37	0	0.0	NaN	

Handling Outliers

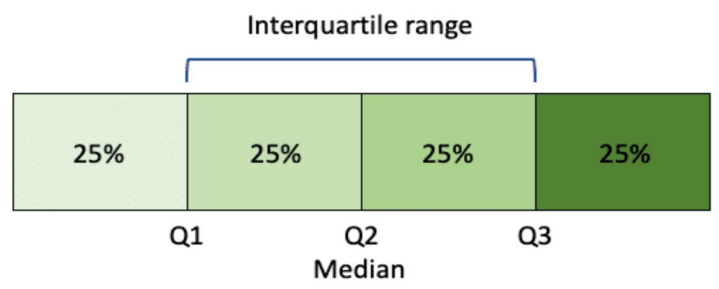
- The dataset has a lot of big outliers.
- **Method**: Replace outliers with median
- Median is used because big outliers affect the mean too much. Hence median is used as measure of central tendency.
- Outliers formula:

  IQR = Q3 - Q1

  Lower outlier: Q1 - 1.5 * IQR

  Upper outlier: Q3 + 1.5 * IQR

  IQR: first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.



Identifying Correlation among columns

When identifying correlations across columns, I found that many columns are strongly correlated. Examples

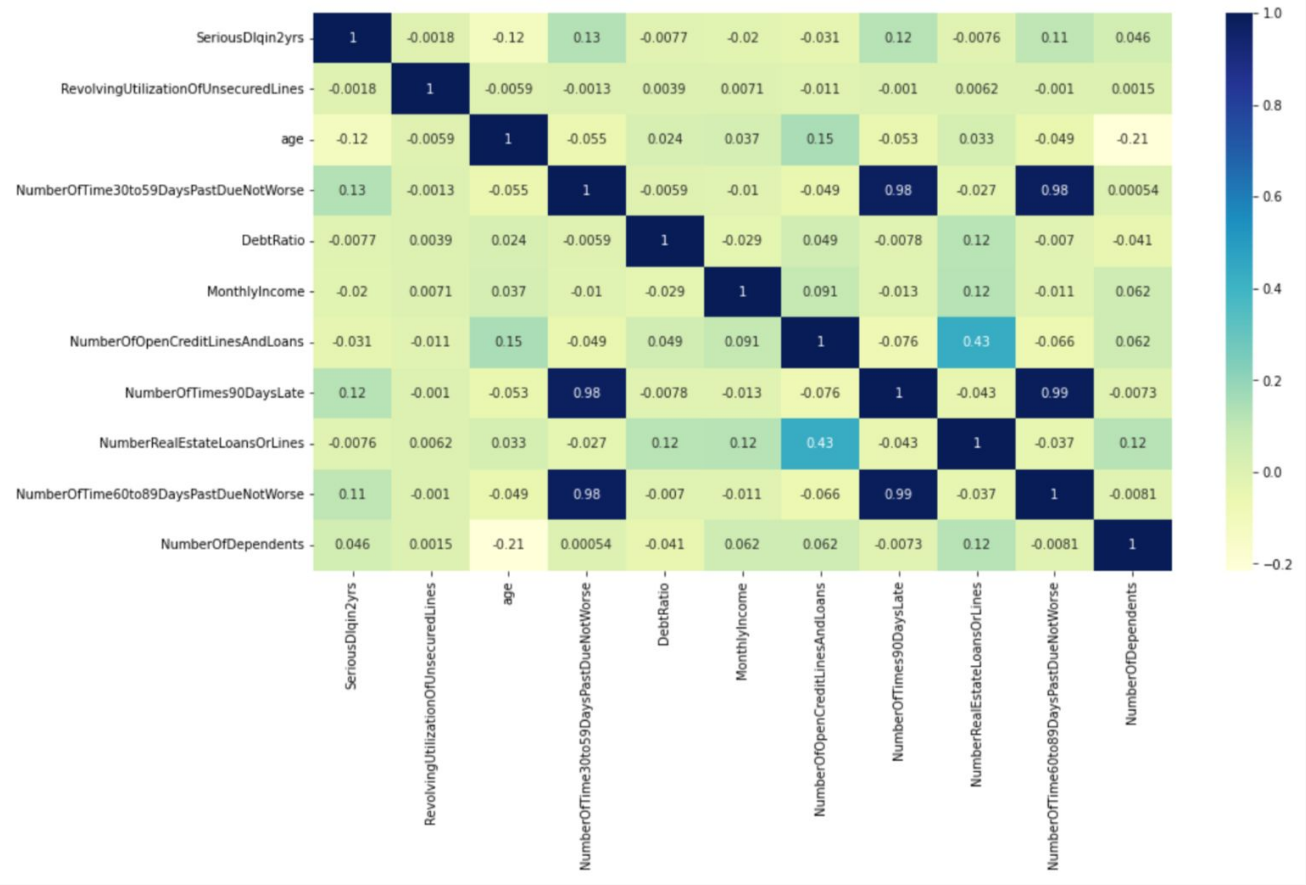
Column 1	Column 2	Correlation (%)
NumberOfTimes90DaysLate	NumberOfTime60to89DaysPastDueNotWorse	**99.1**
NumberOfTime30to59DaysPastDueNotWorse	NumberOfTime60to89DaysPastDueNotWorse	**98.5**
NumberOfTime30to59DaysPastDueNotWorse	NumberOfTimes90DaysLate	**98.0**

Findings

- 'NumberOfTimes90DaysLate', 'NumberOfTime30-59DaysPastDueNotWorse' and 'NumberOfTime60to89DaysPastDueNotWorse' have strong correlation.
- Though they are correlated strongly, the meaning of these three columns are different. So, these columns need to be kept and further analysis needs to be done.

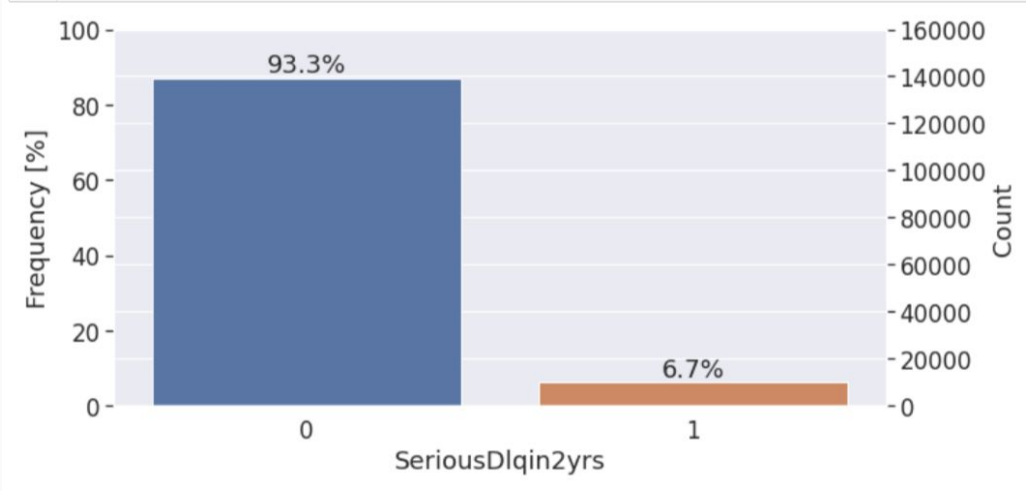
Visualizing Correlation Heatmap

For better visualization, please refer to Notebook.



Target feature

- Feature “SeriousDlqin2yrs” represents person who experiences 90 days past due delinquency or worse.
- Most data points in this dataset is about somebody who did not experience financial distress in two years.
  - 0 (no): 93.3% (139,382)
  - 1 (yes): 6.7% (10,009)
- The target column in this dataset is imbalanced and need to be balanced for further analysis.

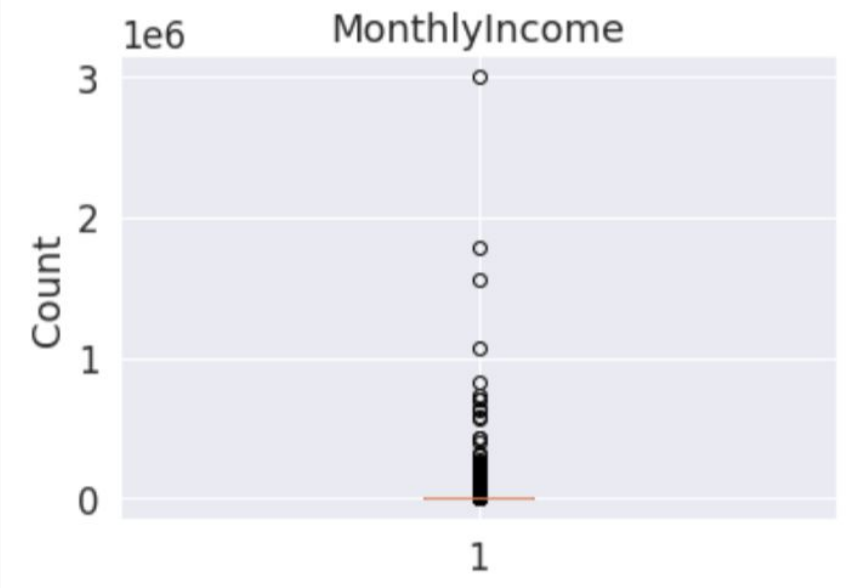


Understanding Each Column

MonthlyIncome

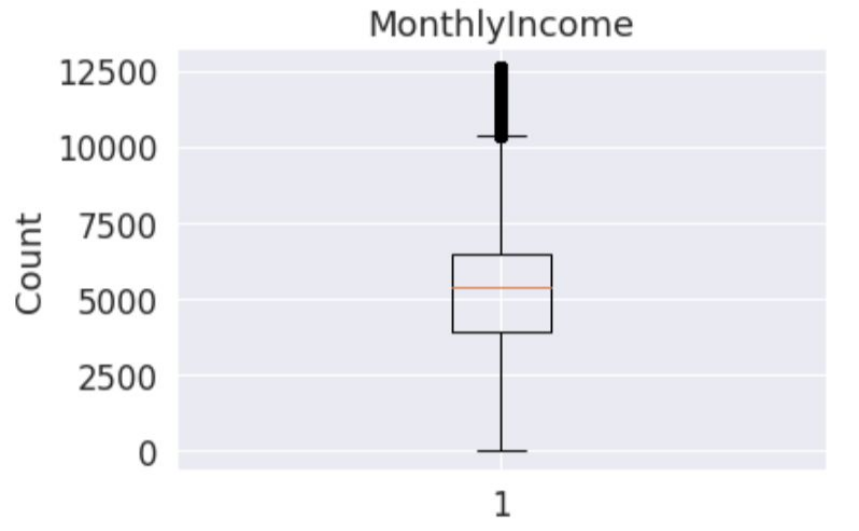
Monthly income ($)
Noise: NaN, outliers

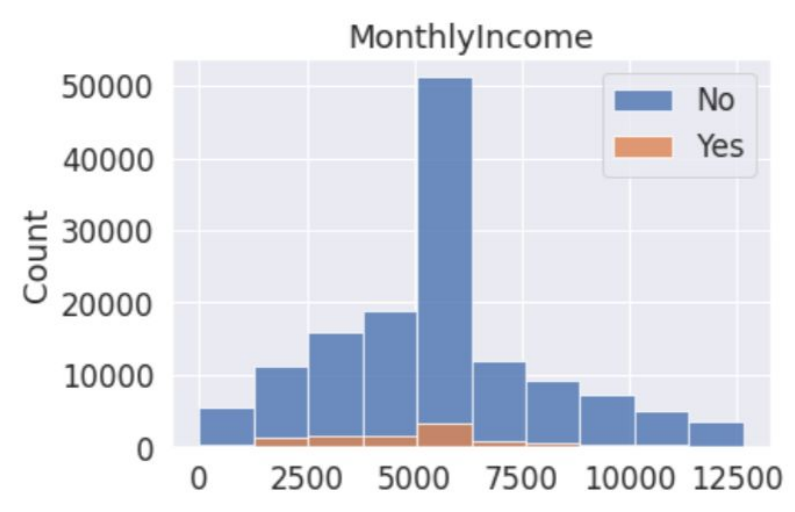
count	1.201700e+05
mean	6.675098e+03
std	1.438958e+04
min	0.000000e+00
25%	3.400000e+03
50%	5.400000e+03
75%	8.250000e+03
max	3.008750e+06



- There are big outliers. Mean is affected by outliers and tend to higher than median.
- I replaced NaN and outliers with median.

After replacing outliers with median

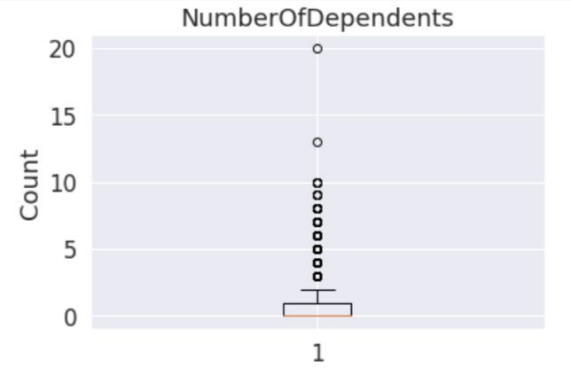




- There are outliers in upper range.
- MonthlyIncome range is wide.

Number Of Dependents

count	145563.000000
mean	0.759863
std	1.116141
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	20.000000

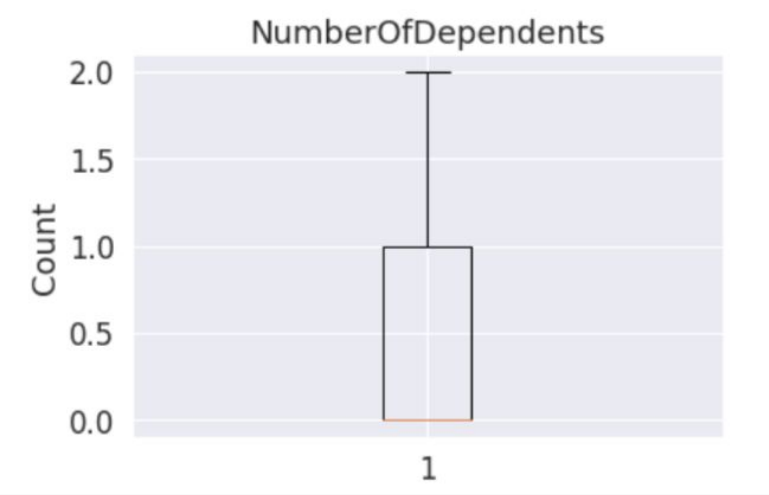


Number of dependents in family excluding applicant (spouse, children, etc...)
Noise: NaN, outliers

- In this dataset, people have less dependents than average, mostly 0, but max is 20.
- Average number of own children under 18 in families with children in the United States in 2020: 1.93 people*
- I did not see anything particular in NaN columns. Replace NaN using median since there are big outliers.

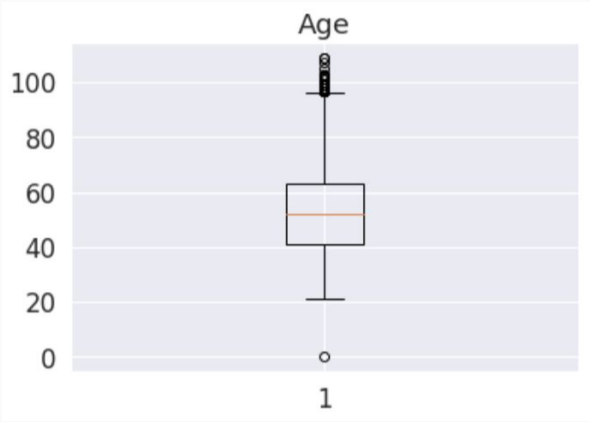
*https://www.statista.com/statistics/718084/average-number-of-own-children-per-family/

After replacing outliers with median



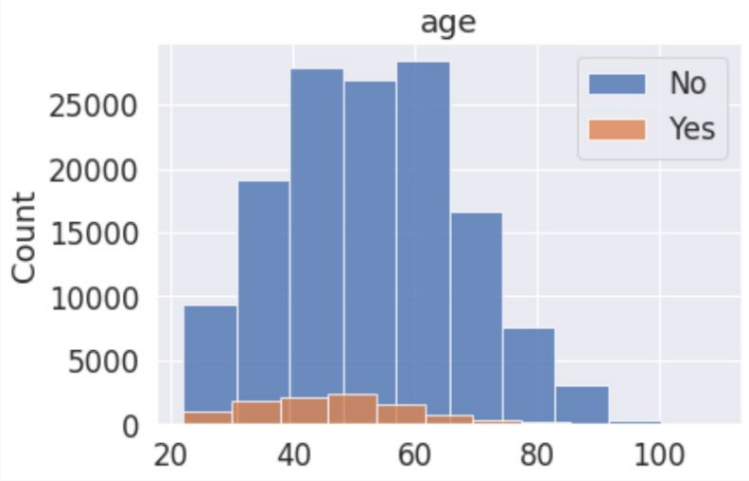
Age

Age of borrower in years



Age	Count
0	1
22	162
23	368
24	592
25	783

After replacing an outlier with median

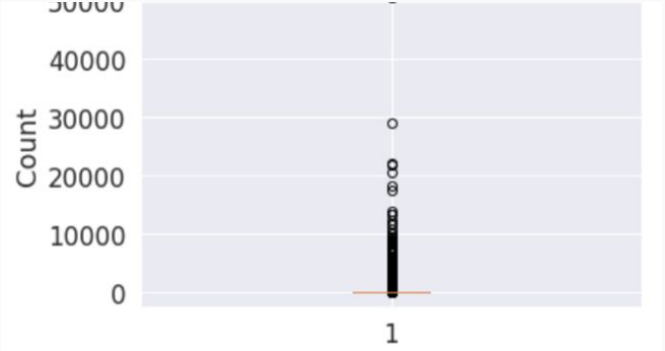


- Age is between 0 and 109.
- There is one outlier age "0". Except this outlier, the youngest applicant is aged to "22".
- It can assumed that age under 22 is an outlier / input errors. I replace it with "22".

Revolving Utilization Of Unsecured Lines

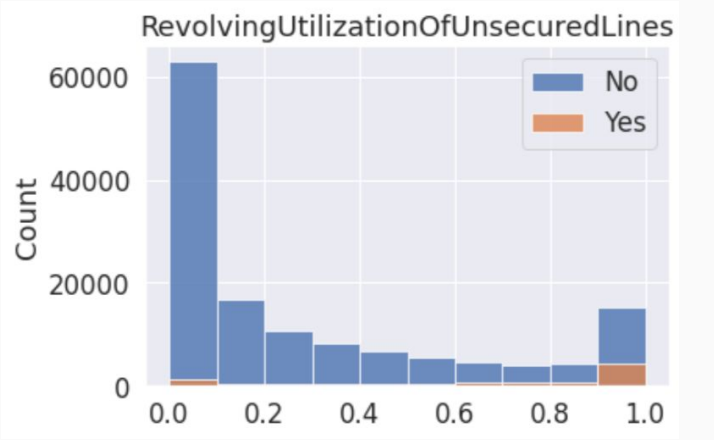
Total balance on credit cards and personal lines of credit except real estate and installment debt (e.g. car loans) divided by the sum of credit limits.

count	149391.000000
mean	6.071087
std	250.263672
min	0.000000
25%	0.030132
50%	0.154235
75%	0.556494
max	50708.000000



- There are many big outliers.
- I assume generally the input should be between 0 and 1 because 75 percentile is under 0.55 (that is < 1).
- I replaced data points above 1 with 1

After replacing data points above 1 with 1



Number Of Time 30 to 59 Days Past Due Not Worse

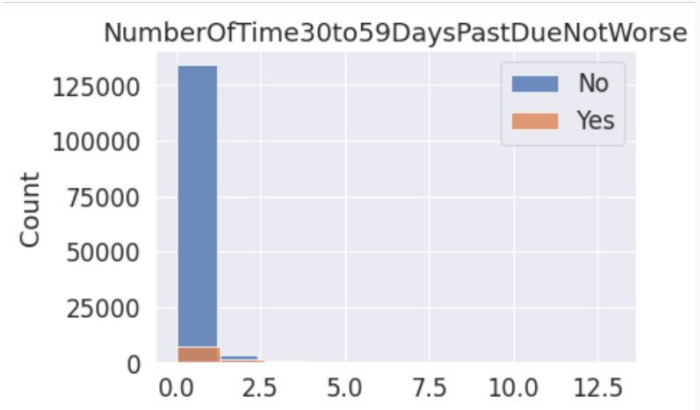
Number of times borrower has been 30-59 days past due but no worse in the last 2 years

- Data points are between 0 and 13.
- Outliers are "98" and "96". Replace them with median.

Data points & count

0	125453
1	16033
2	4598
3	1754
4	747
5	342
98	220
6	140
7	54
8	25
9	12
96	5
10	4
12	2
13	1
11	1

After replacing outliers ‘96’ and ‘98’ with median



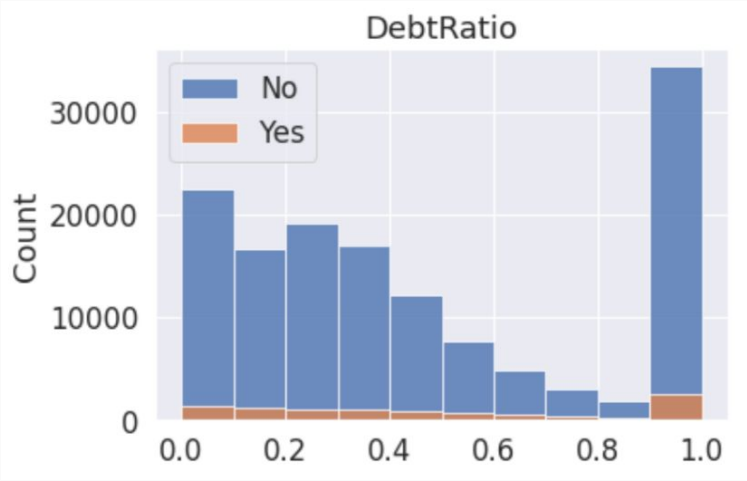
DebtRatio

Monthly debt payments, alimony, and living costs divided by monthly gross income

count	149391.000000
mean	354.436740
std	2041.843455
min	0.000000
25%	0.177441
50%	0.368234
75%	0.875279
max	329664.000000

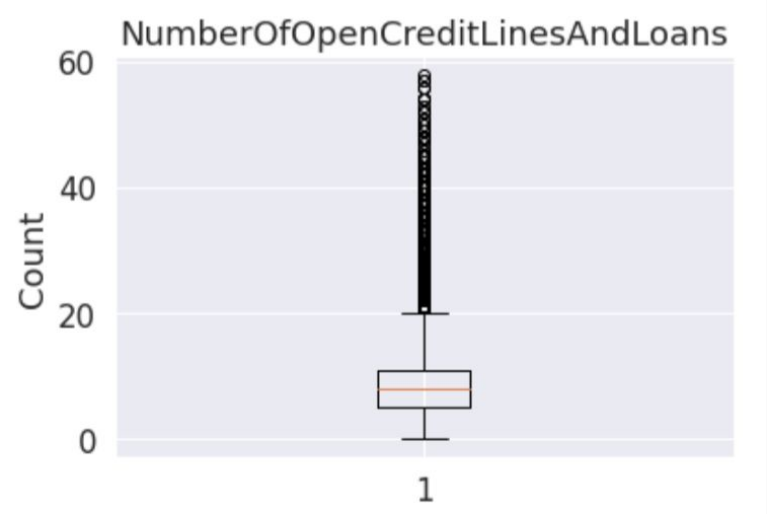
- Most data points are under 1. Replace value above 1.0 with 1.0 to group them as people who have more debt than income.

After replacing data points above 1 with 1



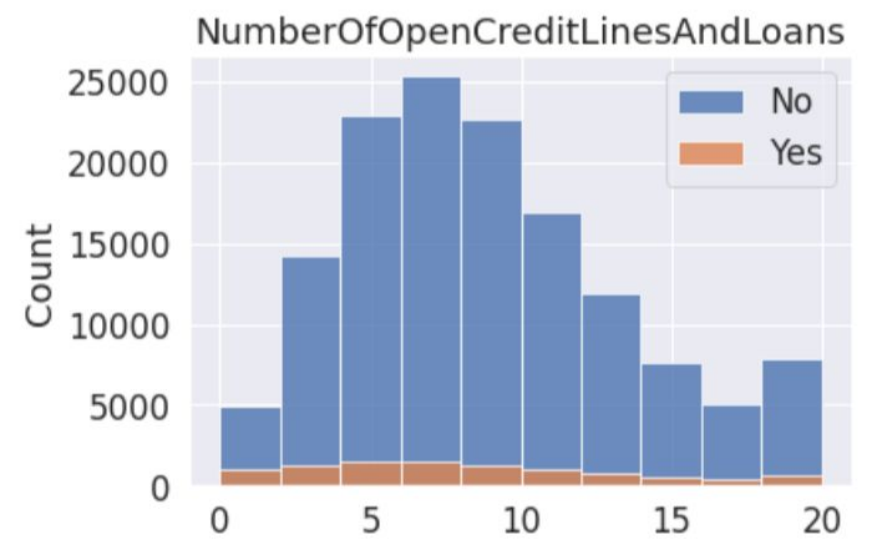
Number Of Open Credit Lines And Loans

Number of open loans (e.g. car loan, mortgage) and lines of credit (e.g. credit cards)



There is no limit for this feature. Replace points outliers, above 20, with 20.

After replacing data points above 1 with 1



Number Of Times 90 Days Late

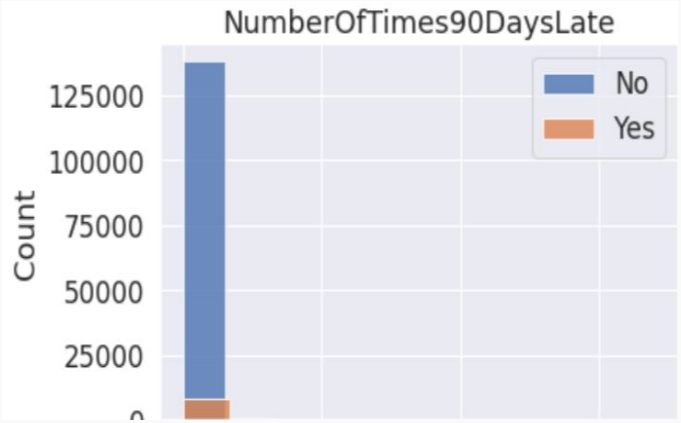
Number of times borrower has been 90 days or more past due

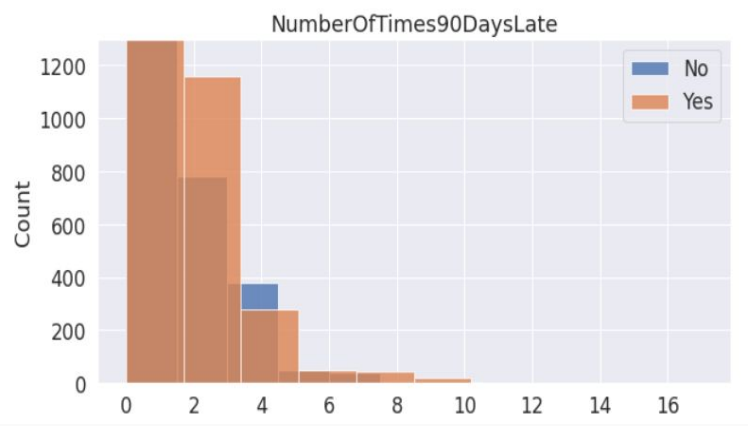
- Data points are between 0 and 17.
- Outliers are "98" and "96". Replace them with median.

Data points & count

0	141108
1	5232
2	1555
3	667
4	291
98	220
5	131
6	80
7	38
8	21
9	19
10	8
96	5
11	5
13	4
15	2
14	2

After replacing "98" and "96" with median





Number Real Estate Loans Or Lines

Number of mortgage and real estate loans including home equity lines of credit*

count	149391.000000
mean	1.022391
std	1.130196
min	0.000000
25%	0.000000
50%	1.000000
75%	2.000000
max	54.000000

- Most data points are between 0 and 2
- Replace outliers with median

After replacing outliers with median

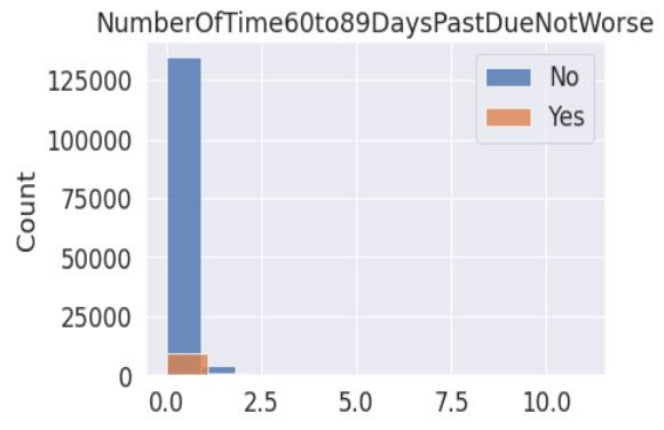
* A home equity line of credit (HELOC) is a line of credit secured by your home that gives you a revolving credit line to use for large expenses or to consolidate higher-interest rate debt on other loansFootnote1 such as credit cards.

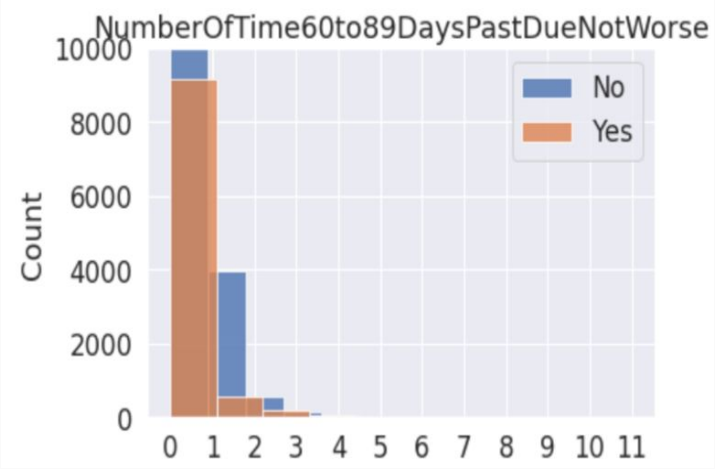
Number Of Time 60 to 89 Days Past Due Not Worse

Number of times borrower has been 60-89 days past due but no worse in the last 2 years

0	141831
1	5731
2	1118
3	318
98	220
4	105
5	34
6	16
7	9
96	5
8	2
11	1
9	1

After replacing outliers with median



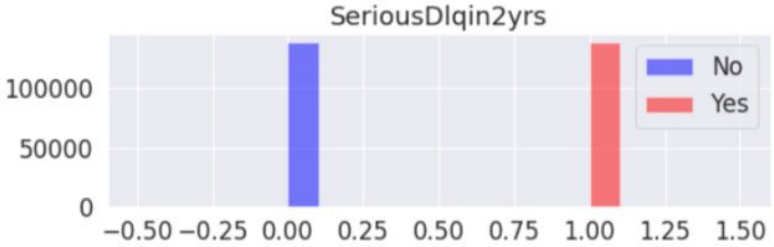


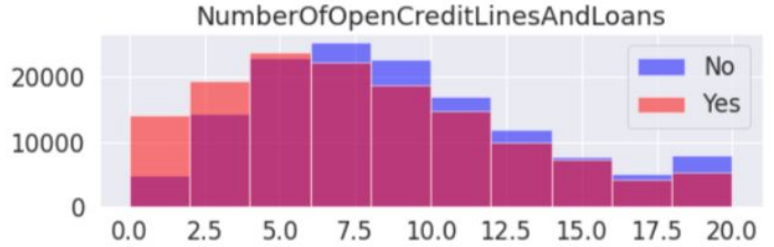
- Data points are between 0 and 11.
- Outliers are "98" and "96". Replace them with median.
- Most Data points are between 0 and 3.

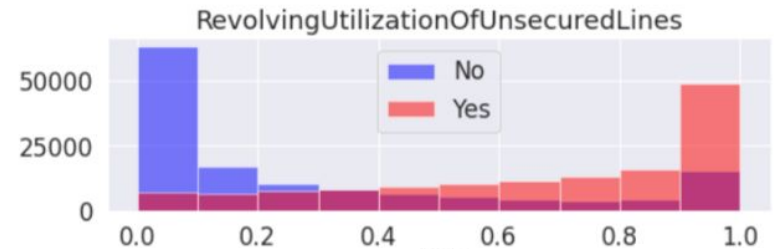
Resampling

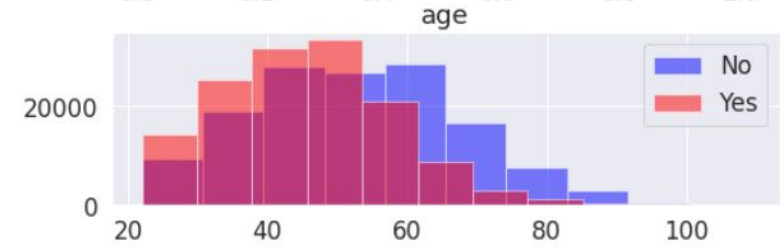
Resampled data to rebalance the target feature, ‘SeriousDlqin2yrs’, to make the correlation between the target and the other features clearer.

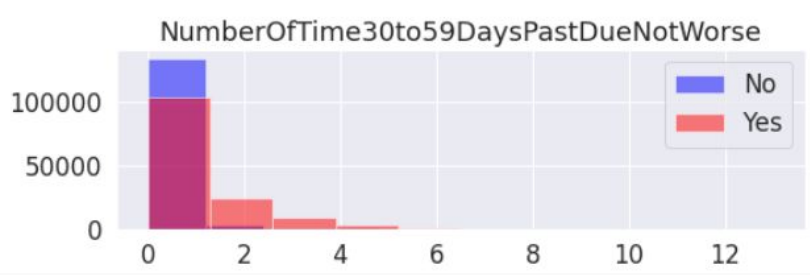
,

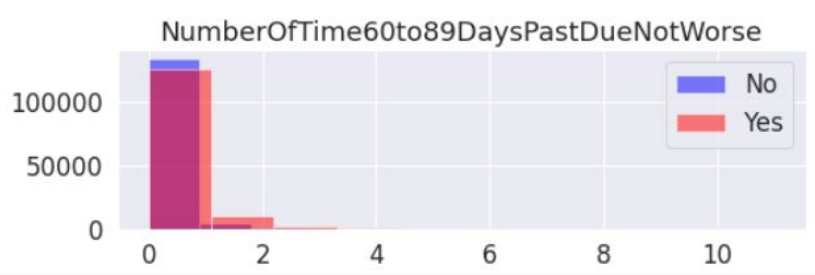




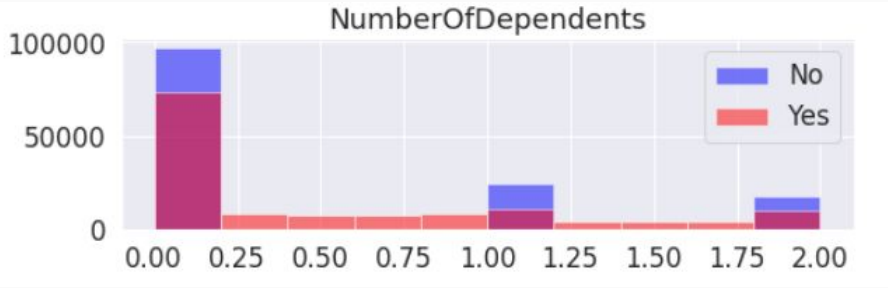


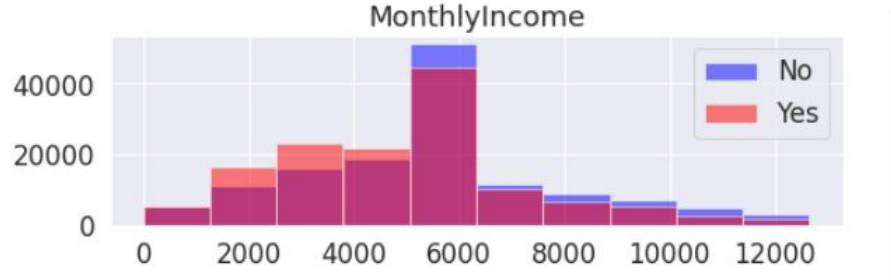



NumberOfTime30to59DaysPastDueNotWorse
No
Yes
100000
50000
0
0
2
4
6
8
10
12


NumberOfTime60to89DaysPastDueNotWorse
No
Yes
100000
50000
0
0
2
4
6
8
10


DebtRatio
No
Yes
20000
0
0.0
0.2
0.4
0.6
0.8
1.0


NumberOfDependents
No
Yes
100000
50000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


MonthlyIncome
No
Yes
40000
20000
0
0
2000
4000
6000
8000
10000
12000

Findings

1. Correlation among columns

Column 1	Column 2	Correlation (%)
SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	51.8
NumberOfOpenCreditLinesAndLoans	NumberRealEstateLoansOrLines	47.7
SeriousDlqin2yrs	NumberOfTime30to59DaysPastDueNotWorse	37.1

After handling outlier and resampling, column 'SeriousDlqin2yrs' has some correlation with 'RevolvingUtilizationOfUnsecuredLines'.

2. As 'RevolvingUtilizationOfUnsecuredLines' increases, the number of ‘yes’ for 'SeriousDlqin2yrs' (person who experienced 90 days past due delinquency or worse) increases clearly, while the number of ‘no’ decreases

Findings

3. From the relation between ‘age’ and ‘SeriousDlqin2yrs’, people who experienced 90 days past due delinquency or worse tend to belong in the younger generation.

4. When ‘NumberOfOpenCreditLinesAndLoans’ is between from 0 to 4, the number of ‘yes’ of ‘SeriousDlqin2yrs’ exceeds the number of ‘no’, but after 5, the number of ‘no’ exceeds the number of ‘yes’.

5. If column `number of times 90 days late` is more than 1, column `SeriousDlqin2yrs` (person experienced 90 days past due delinquency or worse) tends to be ‘yes’. It is more than the number of ‘no’.

6. As the number of ‘NumberRealEstateLoansOrLines’ increases, the percentage of 'No' for ‘SeriousDlqin2yrs’ increases.

7. As the number of ‘NumberOfTime30to59DaysPastDueNotWorse’ increases, the ratio of ‘yes’ of ‘SeriousDlqin2yrs’ to NO's increases.

Summary

- ‘MonthlyIncome’ and ‘NumberOfDependents’ have data points without values (‘null’)
- There are duplicated data
- All features have outliers to be handled
- Some columns are strongly correlated in initial data
- After handling all duplicated data, null and outliers, though the initial strong correlation is not observed, there are characteristic correlations such as:
    - People who experience 90 days past due delinquency or worse tend to have the less balance rate of the credit cards and personal credit lines against credit limit.
    - People who experience 90 days past due delinquency or worse tend to be the younger generation.
    - People who have more mortgage and real estate loans are unlikely to face 90 days past due delinquency or worse.

Thank you!