# PACMANN Probability Project - The Effect of Debtors' Financial State on Credit Approval

By: Muhammad Alif Aqsha

# Introduction

A credible credit risk scoring is crucial for financial institutions, especially those who lend money. This system allows them to quantify and estimate the risk that loan applicants carry; whether it is the risk of defaulting or serious late payments such that high-risk applicants may be flagged and given a higher interest rate to account for their risks or be rejected at all.

It is necessary for financial institutions to be able to balance their selection criteria; too restrictive, and they might reject a large number of sound applicants; too lax, and they might allow for a large number of NPLs (non-performing loans).

# Objectives

- Describe the financial characteristics of loan applicants as a whole and their probability distribution
- Identify the financial condition of loan applicants causing them to be flagged (and thus having their credit application rejected)

# Data Description

The data features a target column ‘SeriousDlqin2yrs’ describing whether the applicant had experienced 90 days (or worse) late payments in a period.

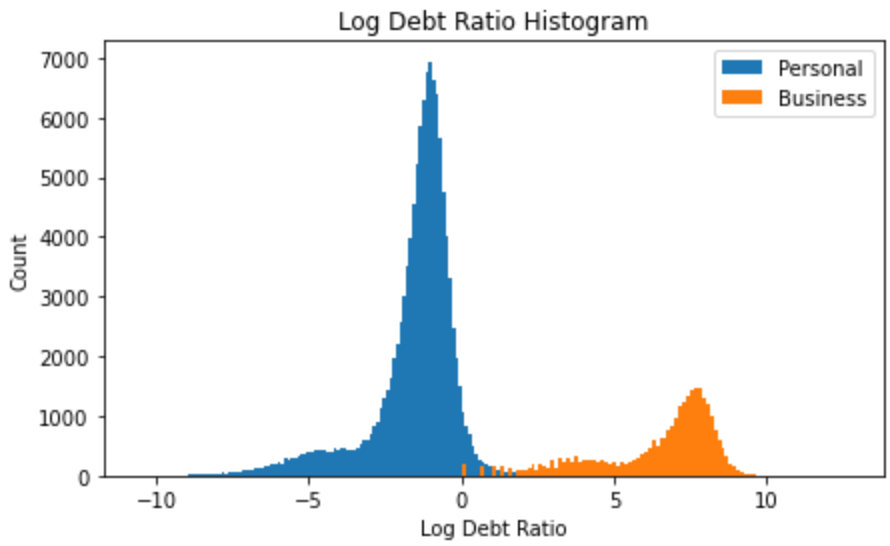
The (financial state) predictor variables are as follows.

| Variable Name | Description | Type |
|---|---|---|
| RevolvingUtilizationOfUnsecuredLines | Total balance on credit cards and personal lines of credit (excluding real estate and installment debt like car loans) divided by the sum of credit limits | float |
| NumberOfTime30-59DaysPastDueNotWorse | Number of times applicants has experienced 30-59 days late payments in the last 2 years (before the ‘SeriousDlqin2yrs’ period) | integer |

| | | |
|---|---|---|
| DebtRatio | Monthly debt payments, alimony, living costs divided by monthy gross income | float |
| MonthlyIncome | Monthly income | float |
| NumberOfOpenCreditLinesAndLoans | Number of open loans (installments like car loans or mortgages) and lines of credit (e.g. credit cards) | integer |
| NumberOfTimes90DaysLate | Number of times applicants have experienced 90 days or more late payments in the last 2 years (before the 'SeriousDlqin2yrs' period) | integer |
| NumberRealEstateLoansOrLines | Number of mortgages and real estate loans (including home equity) lines of credit | integer |
| NumberOfTime60-89DaysPastDueNotWorse | Number of times applicants have experienced 60-89 days late payments in the last 2 years (before the 'SeriousDlqin2yrs' period)<br>. | integer |

# Data Cleaning

The variables which include null values are only MonthlyIncome and NumberOfDependents. For MonthlyIncome, it is very likely that null values represent a different type of loan applicants, most likely business ones. The reason for this is that for these applicants, their DebtRatio's are exact integers (unlike those with non-null MonthlyIncome). Furthermore, their DebtRatio distribution is very different too. As such, we might introduce a new TypeCustomer column derived from MonthlyIncome; the value is 'Personal' if it has a non-null MonthlyIncome, and 'Business' otherwise.




Now, we should identify unusual (potentially mistaken) values from the columns. We expect DebtRatio to range between 0 and 1, but there exists many large values in the data. But this might be explained by two reasons:

- It is normal for businesses to apply for an extremely large sum of loans.
- For personal loans with big DebtRatio, it is likely that they came from student loans (students or recent-graduates are likely to have 0 or very little income).

Then, we also expect RevolvingUtilizationOfUnsecuredLines to range between 0 and 1, but there also exists substantial amount of large values in the data. But this might be explained as follows:

- For values greater than 1 (but not extremely large), it is still possible for borrowers to exceed their credit limits.
- For extremely large ones, it is possible that the borrowers might have closed their credit cards but still have an outstanding invoice. Thus, their credit balance values are large while their credit limits are 0 (perhaps to avoid undefined division, the system computes RevolvingUtilizationOfUnsecuredLines by dividing credit balance with (1 + credit limits).

[Another insight why the system might compute RevolvingUtilizationOfUnsecuredLines by dividing credit balance with (1 + credit limits): after checking with a simple code, all DebtRatio with non-null MonthlyIncome values are a multiple of 1/(1+MonthlyIncome)]

Now, for the age column, there exists a 0-year-old applicant with NumberOfDependents value of 2, which does not make any sense. We impute this value with the median age.

For NumberOfTime30-59DaysPastDueNotWorse column, we found unusually large values (96 & 98 to be exact). There exists a 21-year-old having this column equals 98. It doesn't make any

sense for a 21-year-old to have 98 times of 30-59 late payments (98*30/365 = 8.05 years); it doesn't make sense for a 21-year-old to have started making credit payment from the age 13. Due to this, we impute these large values with the mode (which is 0). The same applies for the column NumberOfTime60-89DaysPastDueNotWorse and NumberOfTimes90DaysLate columns.

# Descriptive Statistics

Below are the descriptive statistics for the variables we are going to analyze.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **NumberOfOpenCreditLinesAndLoans** | 150000.0 | 8.452760 | 5.145951 | 0.0 | 5.000000 | 8.000000 | 11.000000 | 58.0 |
| **NumberRealEstateLoansOrLines** | 150000.0 | 1.018240 | 1.129771 | 0.0 | 0.000000 | 1.000000 | 2.000000 | 54.0 |
| **NumberOfTime30-59DaysPastDueNotWorse** | 150000.0 | 0.245353 | 0.697231 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 13.0 |
| **NumberOfTime60-89DaysPastDueNotWorse** | 150000.0 | 0.064707 | 0.329788 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 11.0 |
| **NumberOfTimes90DaysLate** | 150000.0 | 0.090293 | 0.485107 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 17.0 |
| **RevolvingUtilizationOfUnsecuredLines** | 150000.0 | 6.048438 | 249.755371 | 0.0 | 0.029867 | 0.154181 | 0.559046 | 50708.0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **DebtRatio** | 150000.0 | 353.005076 | 2037.818523 | 0.0 | 0.175074 | 0.366508 | 0.868254 | 329664.0 |
| **MonthlyIncome** | 120269.0 | 6670.221237 | 14384.674215 | 0.0 | 3400.000000 | 5400.000000 | 8249.000000 | 3008750.0 |

Note that there are extreme outliers on all continuous variables (the bottom 3 on the table). Note also that more than 75% of customers have never experienced late payments (30-59, 60-89, or more than 90 days late) in the last 2 years.

Note also that the sample mean and variance ($1.129771^2$ = 1.276245) of the NumberRealEstateLoansOrLines are close to each other. A characteristic of Poisson distribution is that its mean equals its variance, so we may suspect that this variable is Poisson-distributed (perhaps after treating the right-tail outliers).

# Discrete Variables Analysis

## SeriousDlqin2yrs (target variable)

This is the variable representing whether the customer experienced 90 days (or worse) late payments on the current debt (not the last 2 years), and thus being flagged as a bad debtor.

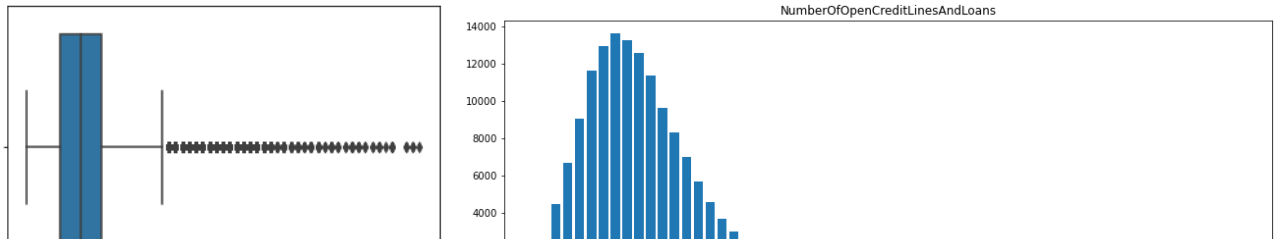
```
flagged = (credit_data["SeriousDlqin2yrs"] == 1)
prob_flagged = flagged.mean()

print(f"Probability of Being Flagged = {prob_flagged}")
```

```
Probability of Being Flagged = 0.06684
```

There are about 6.7% of flagged customers. Because of this, this is a case of an imbalance class of binary variable.

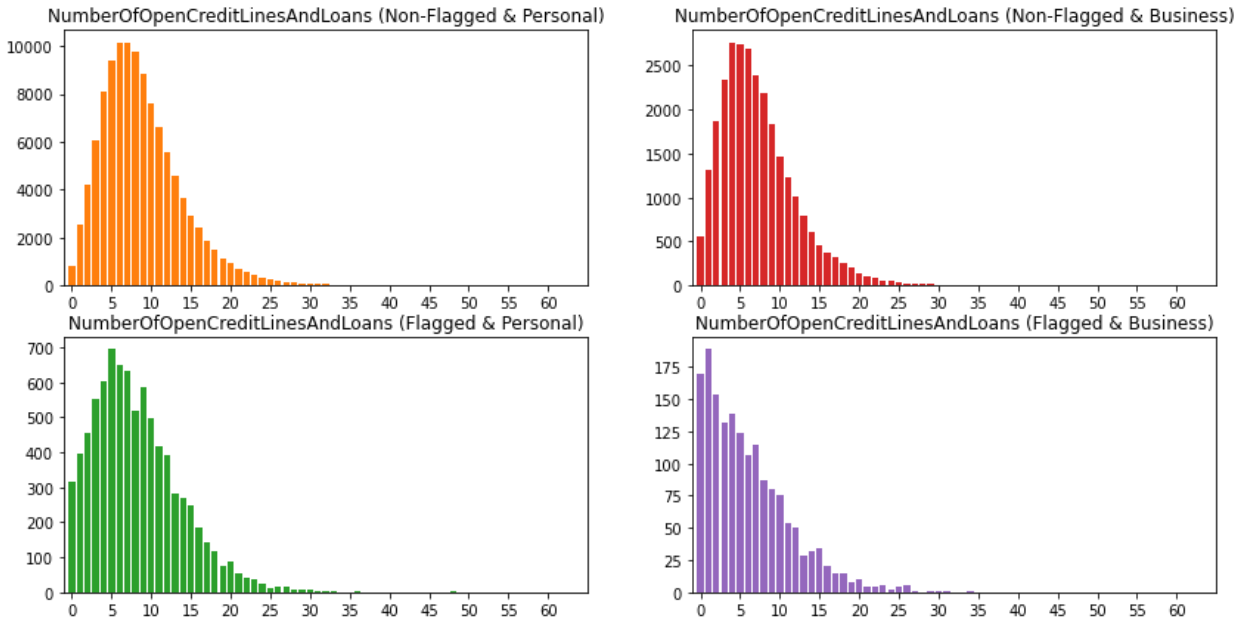
## NumberOfOpenCreditLinesAndLoans

Below are the boxplot and histogram for this variable.



Based on the boxplot, there are some outliers, but they are not clustered far from the upper whisker, so we decide to keep these values unchanged.

We might also obtain the histograms categorized into 4 (a combination from the TypeCustomer and SeriousDlqin2yrs variables).




Visually, there are substantial differences between the histograms of Non-Flagged & Flagged category; the probability mass seems to be shifted to the left for the Flagged category.

By applying Bayes Theorem, the probability of a customer being flagged given 0 credit line or loan is 0.2564, a substantial increase (4 times higher) from the whole proportion of flagged customers (0.067 or 6.7%).

```
Probability of Zero Number of Credit Lines = 0.012586666666666666
Probability of Flagged = 0.06684

Probability of Zero Num of Cred Lines, conditioned on Flagged = 0.04827448633552763
Probability of Flagged, conditioned on Zero Num of Cred Lines = 0.25635593220338987

Probability of Zero Num of Cred Lines, conditioned on Flagged & Personal = 0.037573291851142755
Probability of Flagged, conditioned on Zero Num of Cred Lines & Personal = 0.27162629757785467

Probability of Zero Num of Cred Lines, conditioned on Flagged & Business = 0.1018573996405033
Probability of Flagged, conditioned on Zero Num of Cred Lines & Business = 0.23224043715846998
```

# TypeCustomer

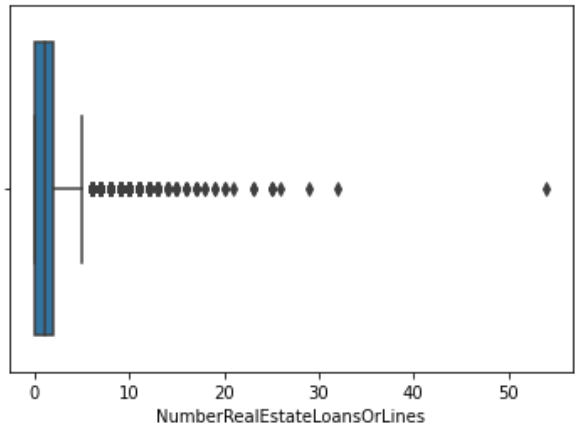
As stated in Data Cleaning section, the TypeCustomer column is derived from MonthlyIncome; if MonthlyIncome is NaN-valued, then the TypeCustomer is 'Business', otherwise 'Personal'.

```
Probability of 'Personal' Customer Type = 0.8017933333333334
Probability of 'Business' Customer Type = 0.19820666666666667
```

Personal customers constitute around 80% total customers, while business customers constitute around 20%.

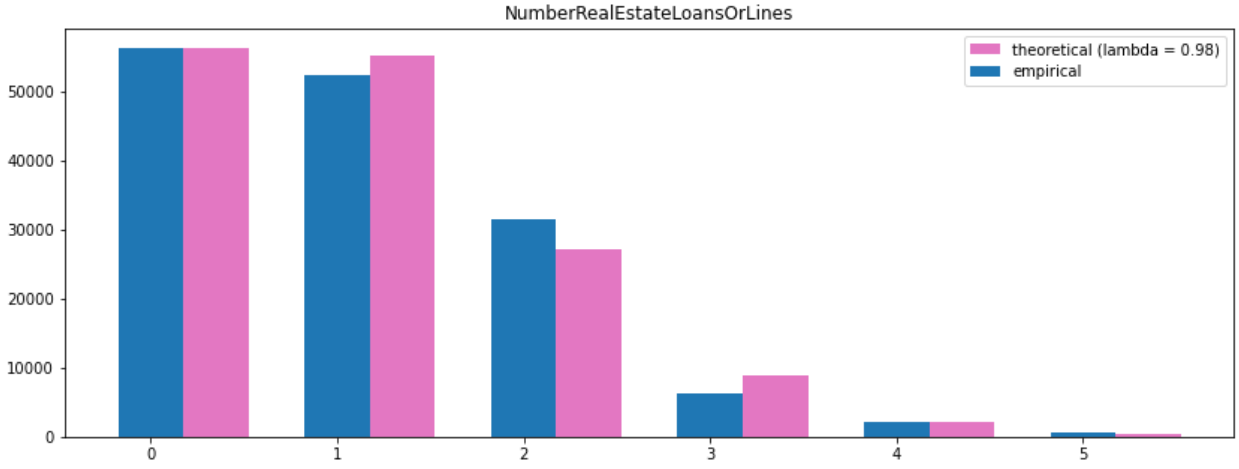
## NumberRealEstateLoansOrLines

Below are the boxplot for this variable




Visually, there are extreme values on the right. As it turns out, there are only 0.53% of customers having more than 5 real estate loans or lines.

After excluding values greater than 5, the mean and variance of this variable gets closer to each other (0.98 and 0.96 respectively), strengthening our guess that this column is Poisson-distributed. The theoretical histogram (for lambda = 0.98 and population size 150,000) also closely matches the empirical histogram.

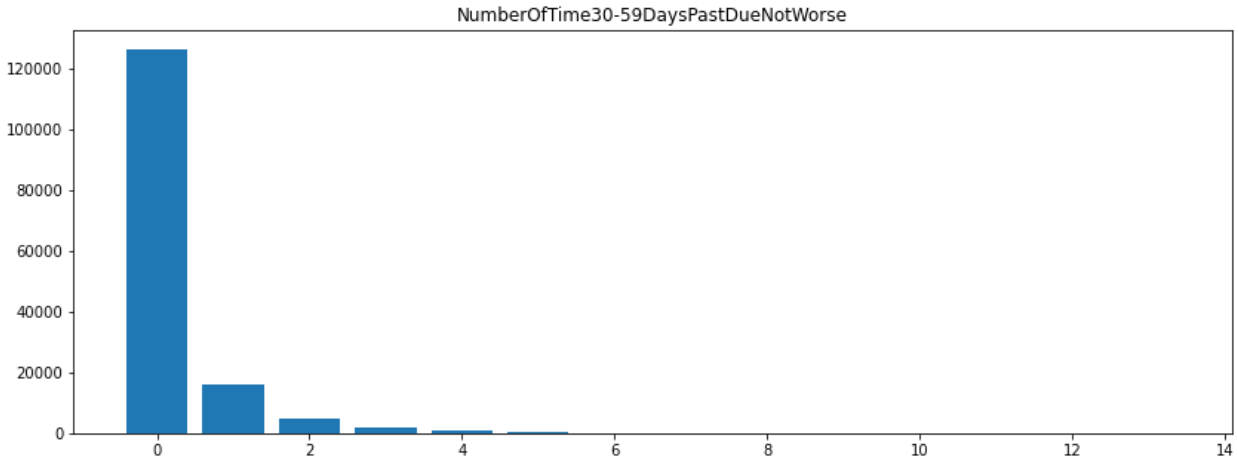


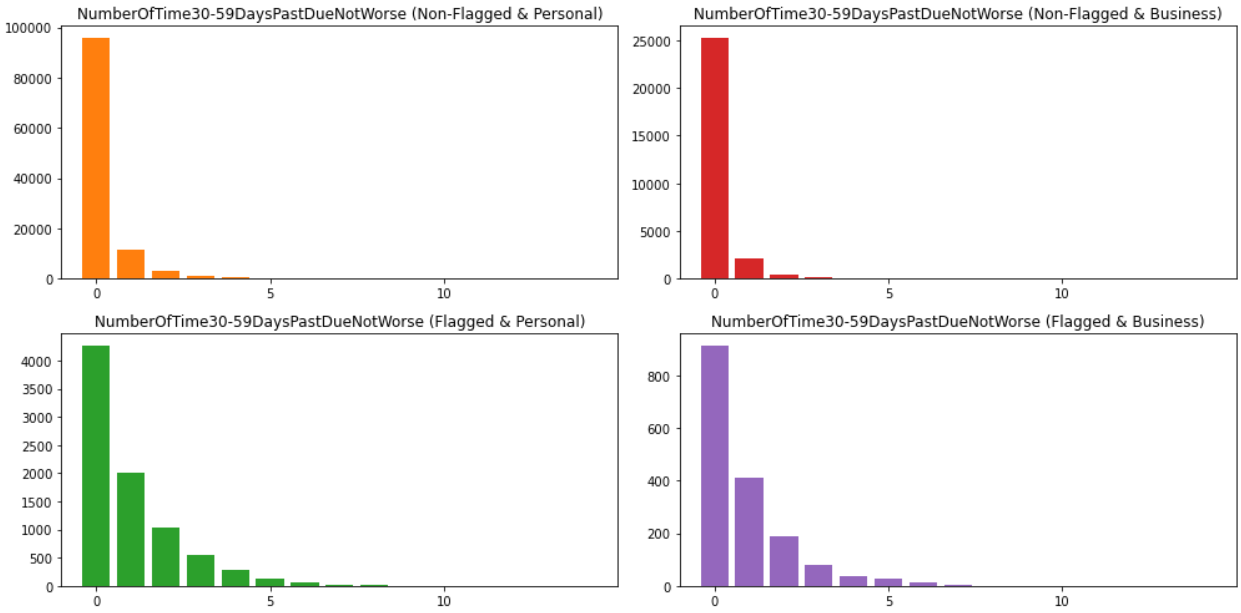
We also observed that conditioned on the extreme value of NumberRealEstateLoansOrLines (>5) and customers from Personal category, the proportion of flagged customers is 19.3%, around 3 times the whole proportion of flagged customers. The proportion of flagged customers (conditioned on >5 NumberRealEstateLoansOrLines and Business type) is 9.7%, only slightly larger than 6.7%.

```
Probability of >5 Number of Real Est. Loans = 0.005286666666666667
Probability of Flagged = 0.06684

Probability of Flagged, conditioned on >5 Num of Real Est. Loans = 0.1790668348045397

Probability of Flagged, conditioned on >5 Num of Real Est. Loans & Personal = 0.19264705882352942

Probability of Flagged, conditioned on >5 Num of Real Est. Loans & Business = 0.09734513274336283
```

## NumberOfTime30-59DaysPastDueNotWorse

Below are the histograms for this variable (as a whole & divided into categories based on customer type and flagged status).





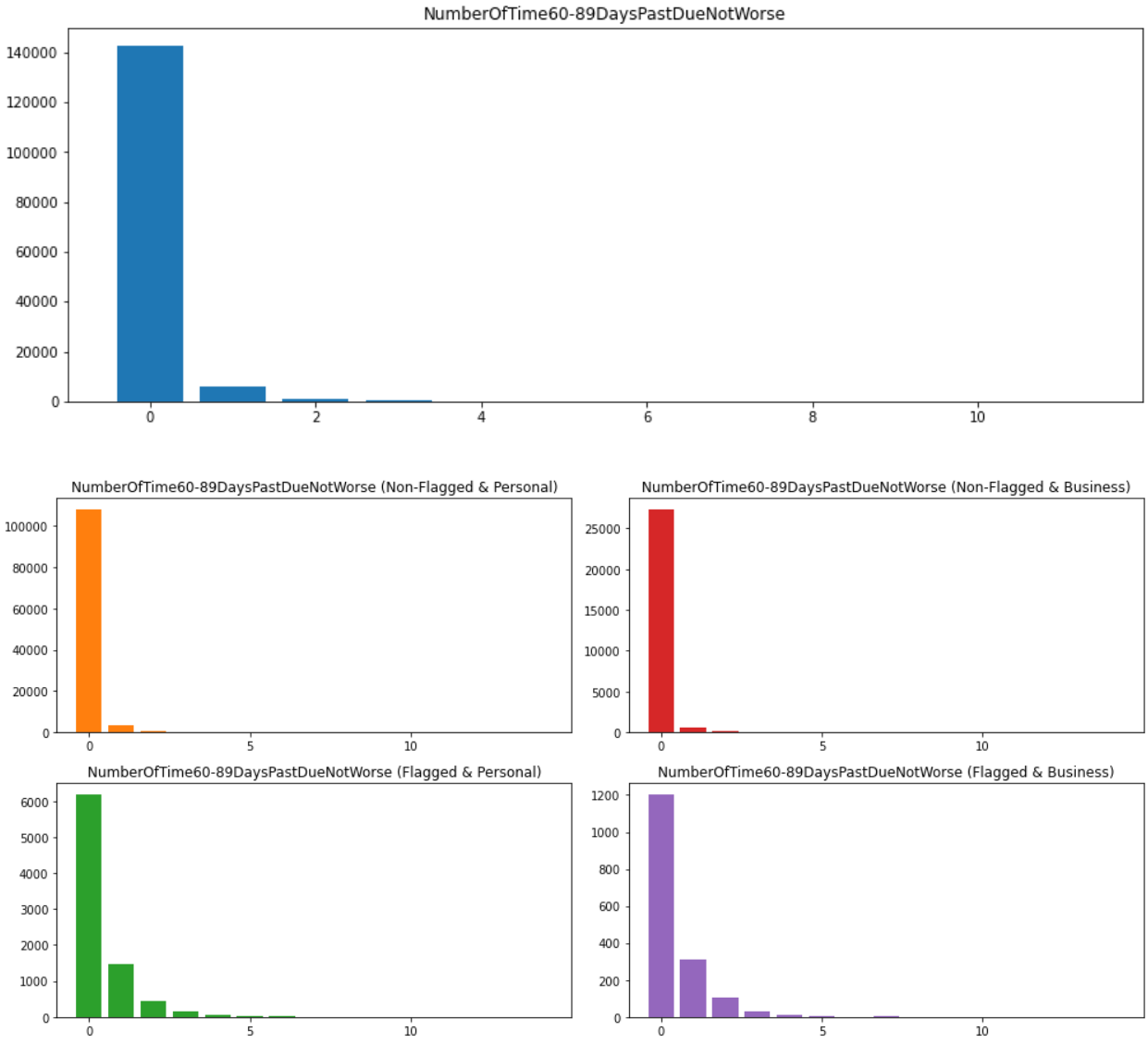

Visually, the NumberOfTime30-59DaysPastDueNotWorse values for non-flagged customers are more concentrated on 0, while the values for flagged ones are more spread out.

We may also compute that the odds a customer will be flagged is more than tripled if they have at least experienced 30-59 days late payments once.

```
Probability of Nonzero Number of 30-59 Days Past Due = 0.15808666666666665
Probability of Flagged = 0.06684

Probability of Flagged, conditioned on Nonzero Num of 30-59 Days Past Due = 0.20402310968666978

Probability of Flagged, conditioned on Nonzero Num of 30-59 Days Past Due & Personal = 0.202620217358
94004

Probability of Flagged, conditioned on Nonzero Num of 30-59 Days Past Due & Business = 0.211959573273
4419
```

## NumberOfTime60-89DaysPastDueNotWorse

Below are the histograms for this variable (as a whole & divided into categories based on customer type and flagged status).



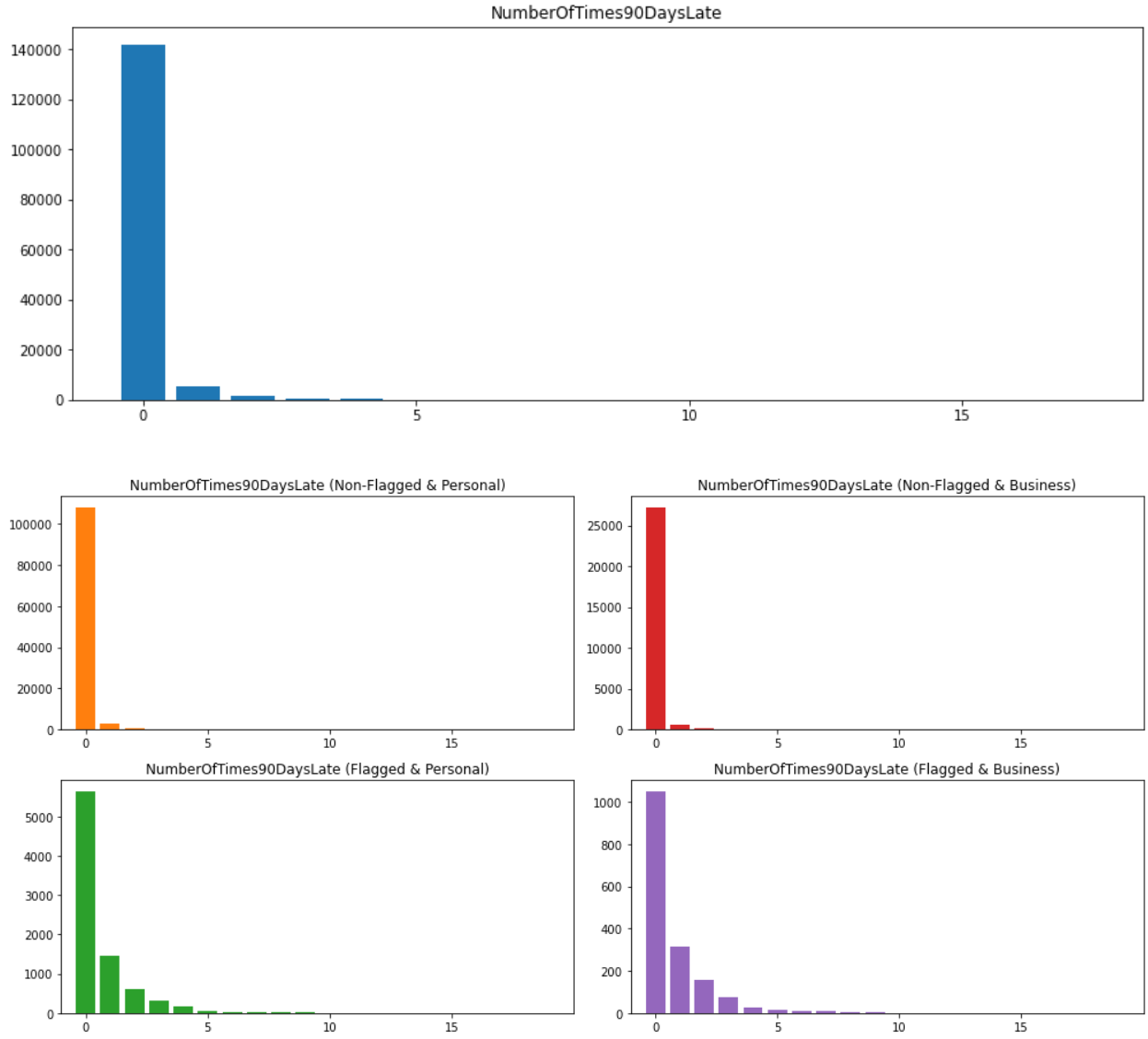

Visually, the NumberOfTime60-89DaysPastDueNotWorse values for non-flagged customers are more concentrated on 0, while the values for flagged ones are more spread out.

We may also compute that the odds a customer will be flagged is increased by a factor of around 6 if they have at least experienced 60-89 days late payments once.

```
Probability of Nonzero Number of 60-89 Days Past Due = 0.0489
Probability of Flagged = 0.06684

Probability of Flagged, conditioned on Nonzero Num of 60-89 Days Past Due = 0.3576005453306067

Probability of Flagged, conditioned on Nonzero Num of 60-89 Days Past Due & Personal = 0.352374734780
4798

Probability of Flagged, conditioned on Nonzero Num of 60-89 Days Past Due & Business = 0.384105960264
9007
```

# NumberOfTimes90DaysLate

Below are the histograms for this variable (as a whole & divided into categories based on customer type and flagged status).




Visually, the NumberOfTime60-89DaysPastDueNotWorse values for non-flagged customers are more concentrated on 0, while the values for flagged ones are more spread out.

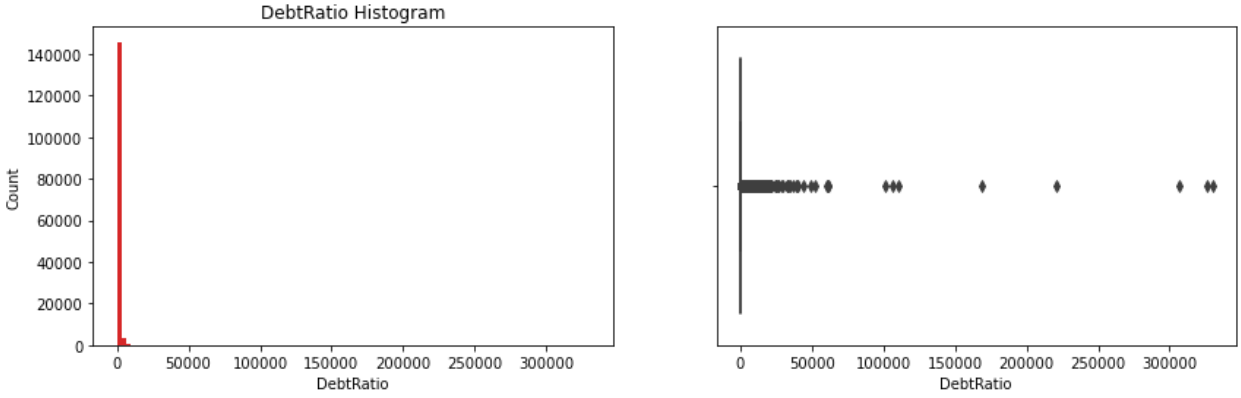
We may also compute that the odds a customer will be flagged is increased by a factor of around 7 if they have at least experienced 60-89 days late payments once.

```
Probability of Nonzero Number of >=90 Days Past Due = 0.05379333333333333
Probability of Flagged = 0.06684

Probability of Flagged, conditioned on Nonzero Num of >= 90 Days Past Due = 0.41207088858594626

Probability of Flagged, conditioned on Nonzero Num of >=90 Days Past Due & Personal = 0.4115858294055
0403

Probability of Flagged, conditioned on Nonzero Num of >= 90 Days Past Due & Business = 0.414209115281
5014
```

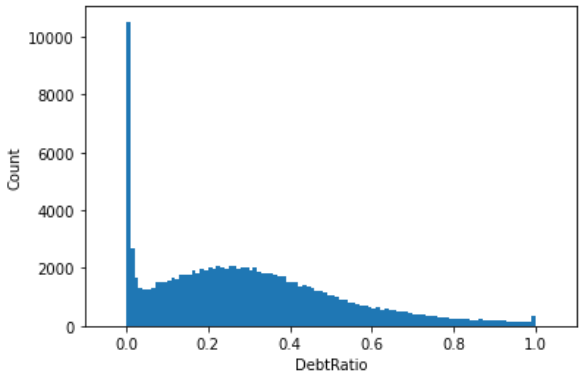
# Continuous Variables Analysis

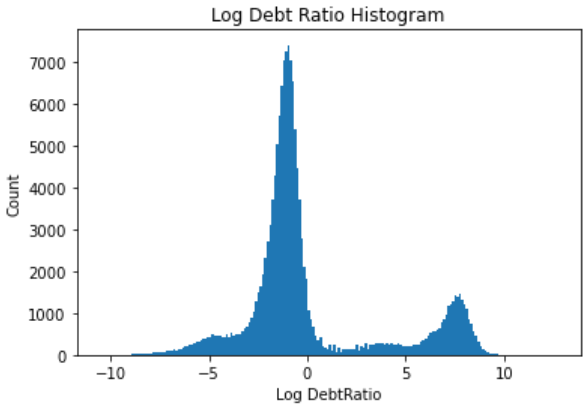
## DebtRatio

Below are the histogram and boxplot for this variable.



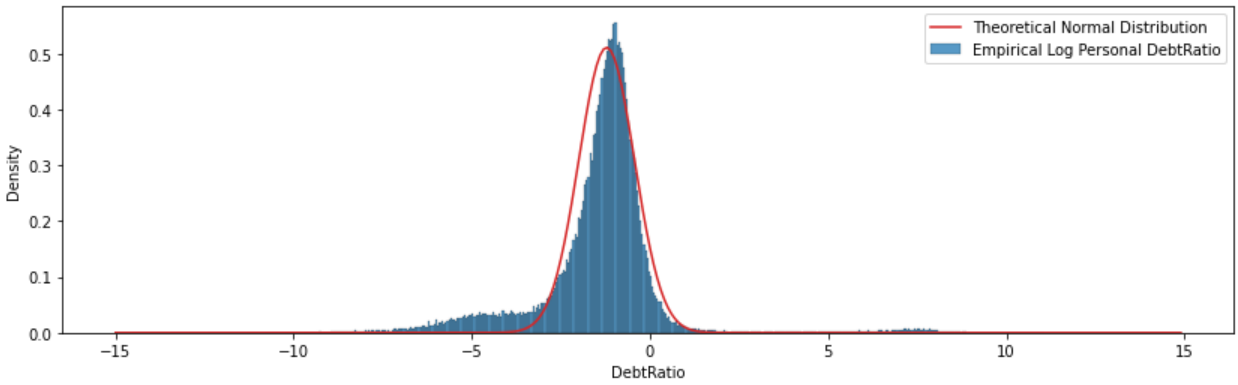

As this column includes extremely large values, we show the histogram in two ways: truncated histogram in the range of normal values (0-1), and the log-transformed histogram (excluding 0 as log 0 is undefined, and this value only accounts for 2.7% customers).



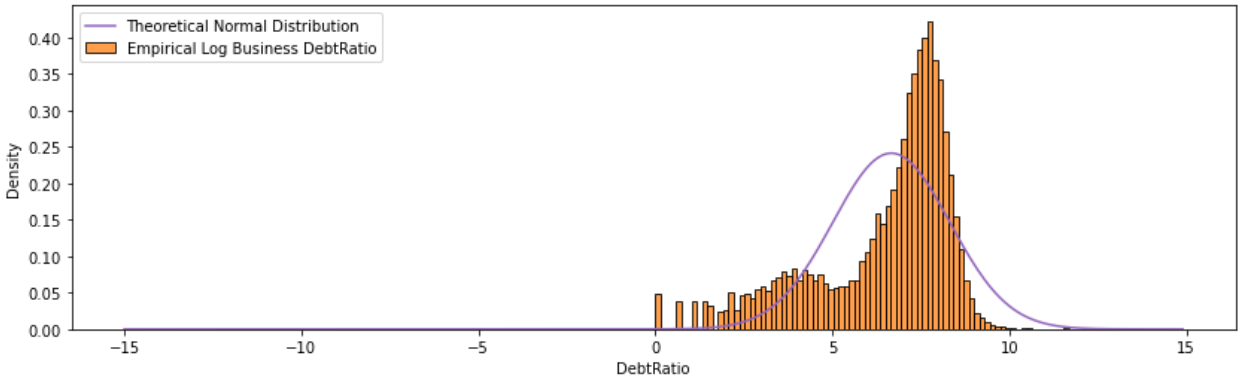




For the log-transformed DebtRatio, the distribution is bimodal as it actually came from 2 populations: customers with undefined MonthlyIncome (or Business type) and those with defined MonthlyIncome (or Personal type) as explained on Data Cleaning section.

For the Personal type, the distribution of the log-transformed DebtRatio is approximately normal (but with extreme values on both sides). To compute a representative mean and variance for these values, we need to exclude the outliers first using the IQR method. From this, we obtain a sample mean of -1.22 and standard deviation of 0.78. The empirical density histogram also closely matches the theoretical normal distribution density, except that the empirical one has a slightly thicker left tail.

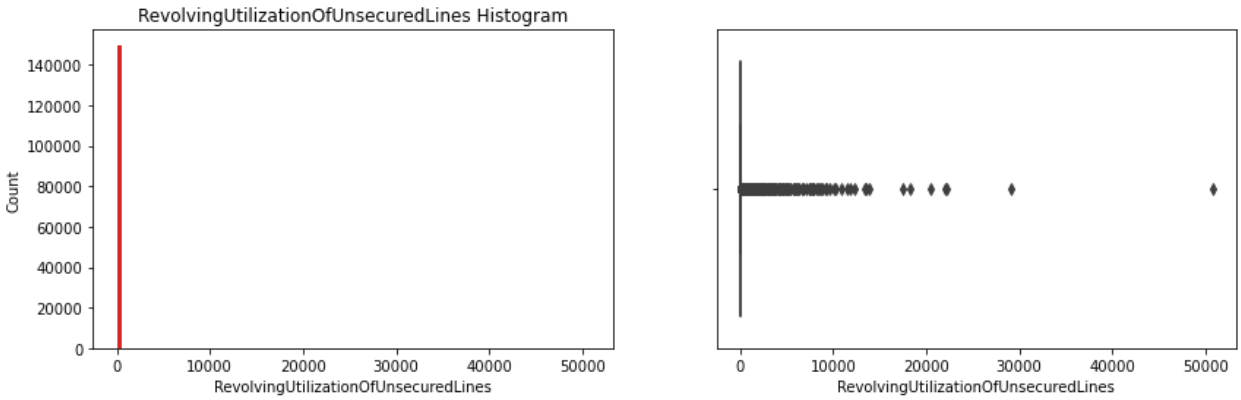



However, if we apply the same method as above for the Business-type log-transformed DebtRatio, the density fit is not very good. This is because for the Business-type, the distribution of the log-transformed DebtRatio has a substantially thicker left-tail. It also seemed that the distribution of the Business-type log-transformed DebtRatio is actually bimodal, but we haven't found a way to separate it into two populations.

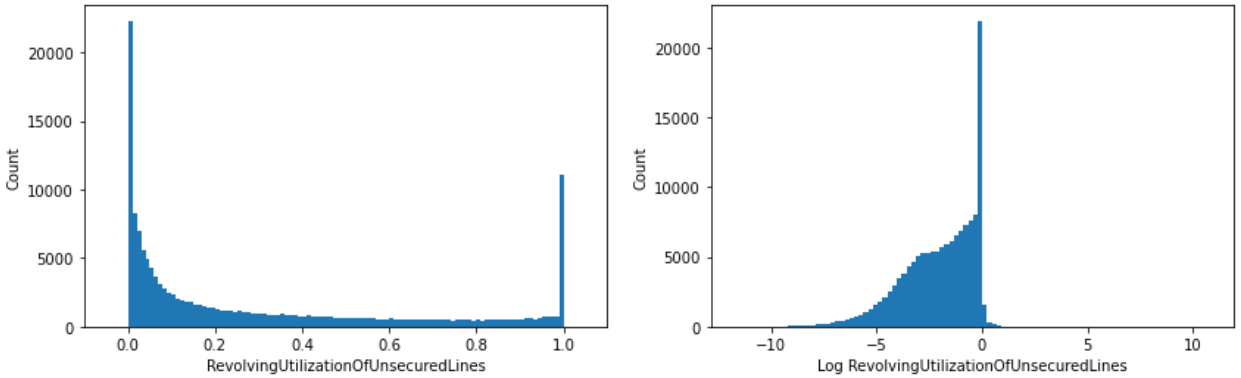



# RevolvingUtilizationOfUnsecuredLines

Below are the boxplot and histogram for RevolvingUtilizationOfUnsecuredLines.




As this column includes extremely large values, we show the histogram in two ways: truncated histogram in the range of normal values (0-1), and the log-transformed histogram (excluding 0 as log 0 is undefined).



From the histograms above, there are spikes around the values 0 and 1. This suggests that there are a substantial number of customers who didn't use (or very little) their credit limits, and customers who, after passing a certain threshold, prefer to (almost or completely) max out their balance.

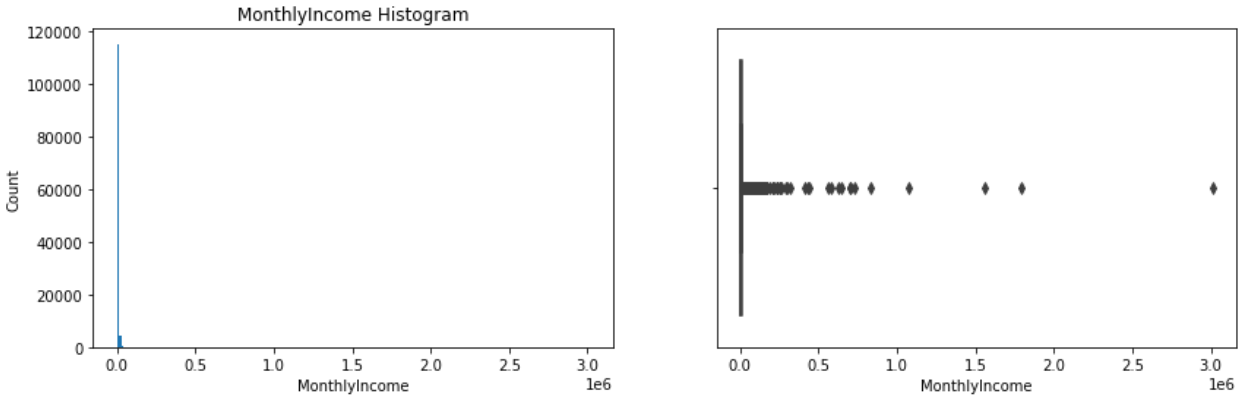
```
Probability of 0 RevolvingUtilizationOfUnsecuredLines: 0.073
Probability of 0.99999-1 RevolvingUtilizationOfUnsecuredLines: 0.068
```

We may also compute that the probability of Flagged customer conditioned on greater-than-1 RevolvingUtilizationOfUnsecuredLines is 0.372 (6 times higher than the initial proportion).

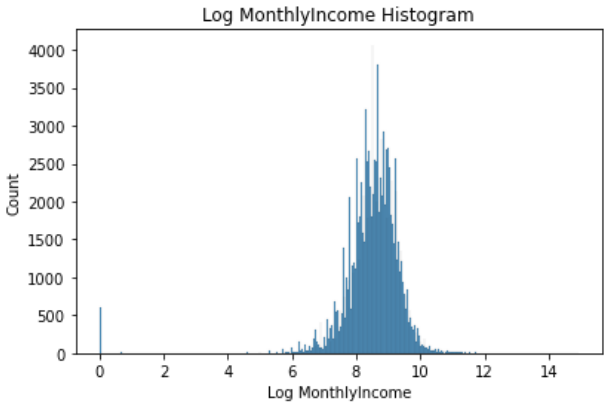
```
Probability of Flagged customer, given greater than 1 RevolvingUtilizationOfUnsecuredLines: 0.372
```

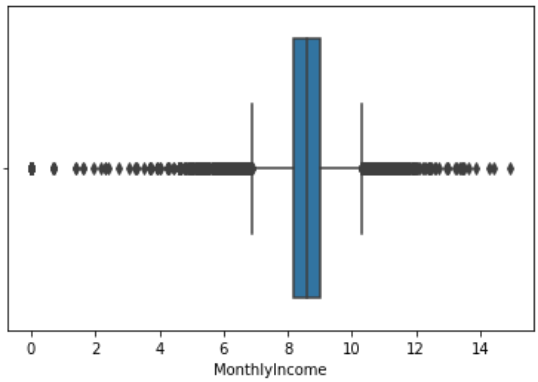
# MonthlyIncome

Below are the histogram and boxplot for MonthlyIncome variable. Note that as discussed in Data Cleaning section, only Personal-type customers have non-null MonthlyIncome.

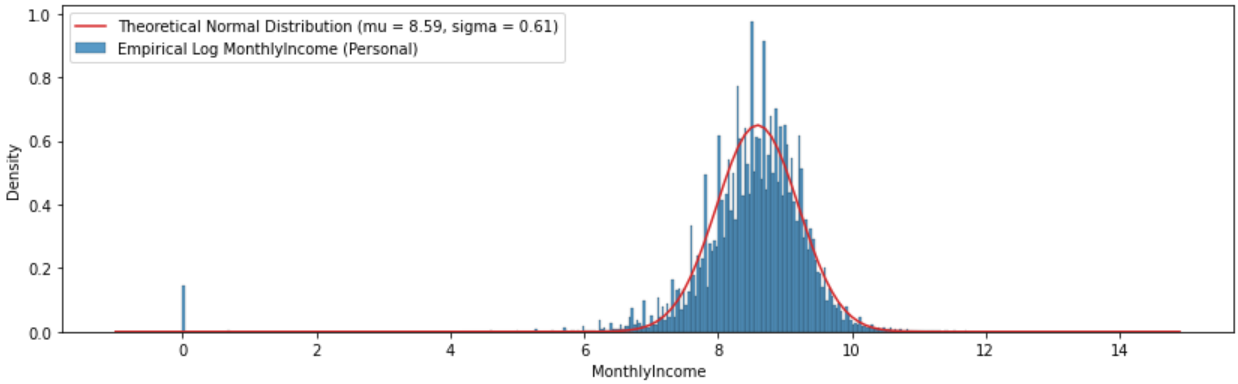



As this column includes extremely large values, we transform these values into logarithmic scale (excluding zero incomes which constitute 1.36% of Personal-type customers).



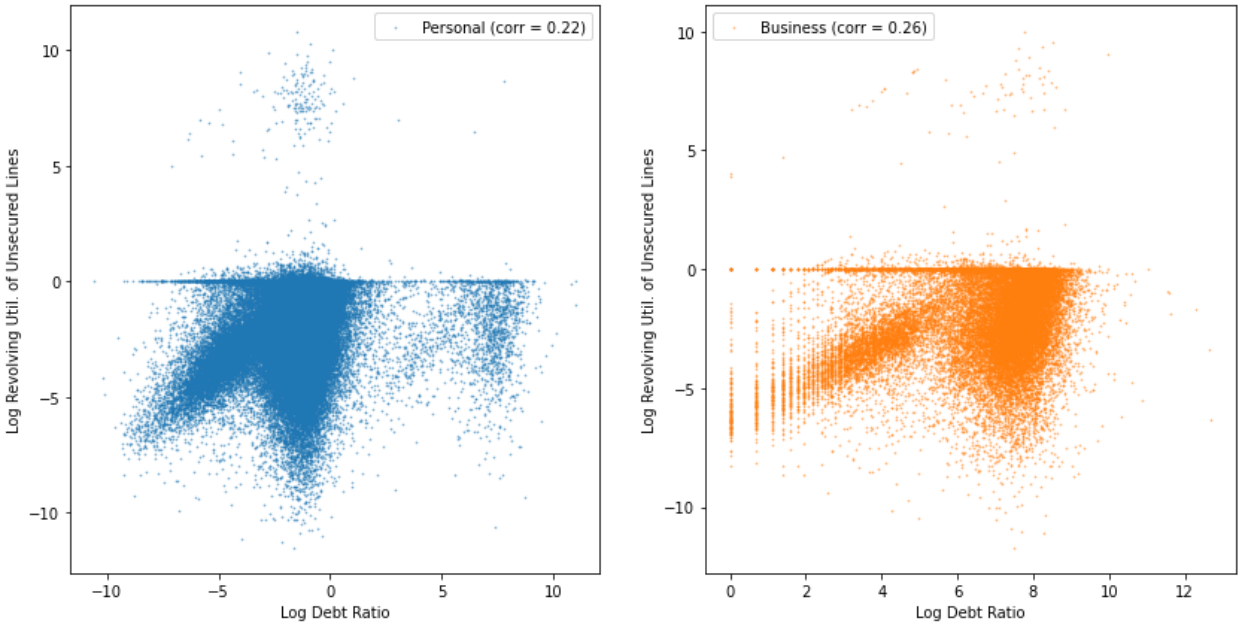

Visually, the log-transformed MonthlyIncome is approximately normally-distributed, except for the fact that there are some spikes (especially on the value 0, which corresponds to the MonthlyIncome value of 1). To obtain a representative sample mean and standard deviation of this log-transformed variable, we first exclude the outliers using the IQR criterion. From this, we obtained the sample mean and standard deviation of 8.59 and 0.61 respectively.




Visually, the theoretical normal distribution more-or-less fits the empirical distribution, except for the spikes part (especially on the value 0).

# Correlation Analysis

Below are the scatterplots for (log-transformed) DebtRatio vs RevolvingUtilizationOfUnsecuredLines for each TypeCategory and their respective correlation coefficient.



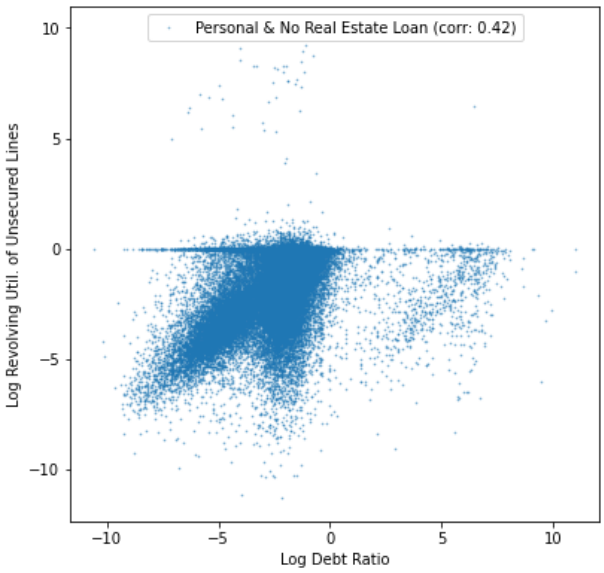

For both TypeCustomer categories, we obtained weak positive correlations. This is because RevolvingUtilizationOfUnsecuredLines is actually the ratio of credit card balances (excluding real estate and other installment-type of loans such as car loans) to their credit limits. On the other hand, DebtRatio is the ratio of TOTAL debts (including no-installment and installment-type of debts) to incomes.

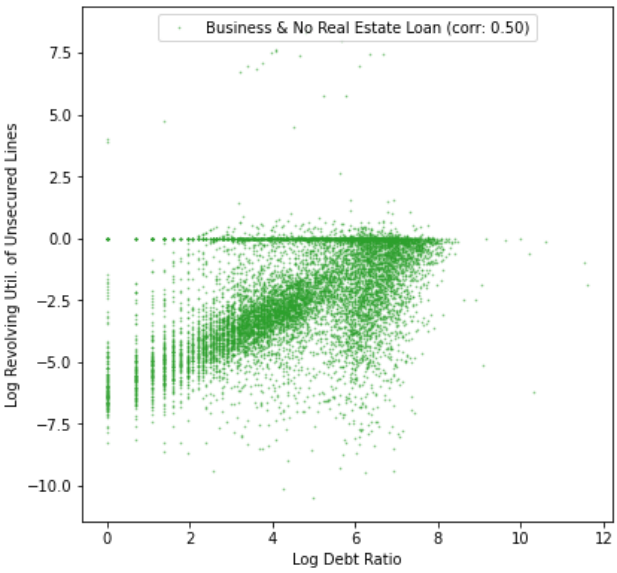
From these definitions, we may guess that for each TypeCustomer categories, there are 2 populations:

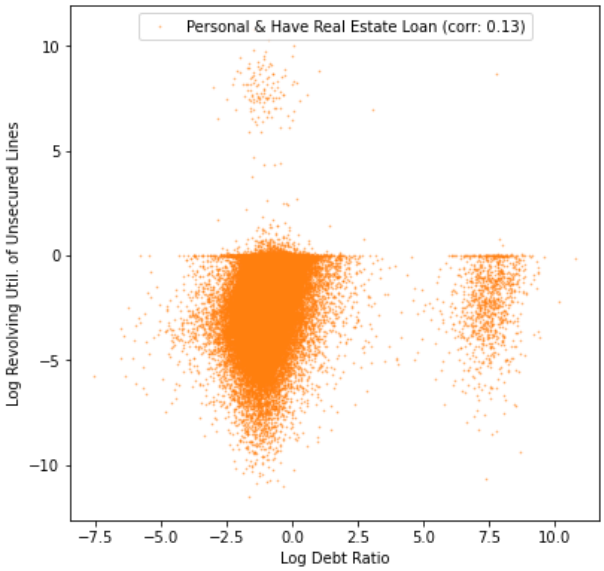
- Customers where no-installment type of debts dominate their total debts
- Customers where installment type of debts (such as real estate or car loans) dominate their total debts

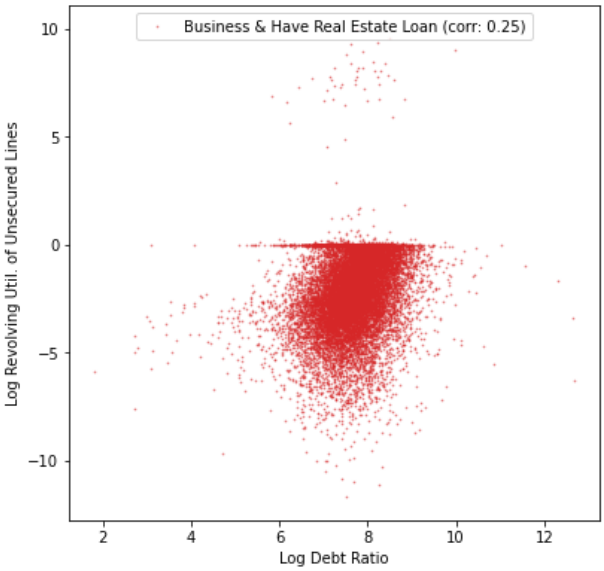
For the first type, the log DebtRatio and log RevolvingUtilizationOfUnsecuredLines are actually positively and linearly related, as the numerators on both variables are proportional to each other (Total debts ~ Credit limits). For the second type, as the installment type of debts (such as real estate or car loans) dominate their total debts, their credit limits are not proportional to their total debts; hence why we can visually observe there are actually 2 lines on the scatter plots (the positively-sloped line and horizontal line).

If we divide further into categories whether the customer has a real estate loan or not, the pattern becomes clearer. The correlation coefficient doubled if we restrict ourselves to customers with no real estate loan (although there still seems to be two lines as we have not account for other type of installment-type loans such as car loans, which is not available from data).



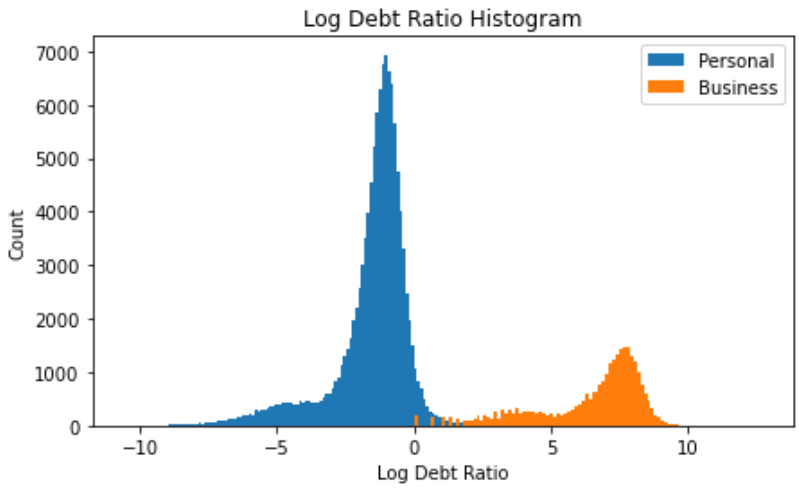




# Hypothesis Testing

Obviously, we don't need to test wheter the population means of (log) DebtRatio for Personal vs Business customer type are equal (as the histogram clearly shows that they are exteremely different).




## Comparing the population mean of (log) DebtRatio for Non-Flagged and Flagged populations

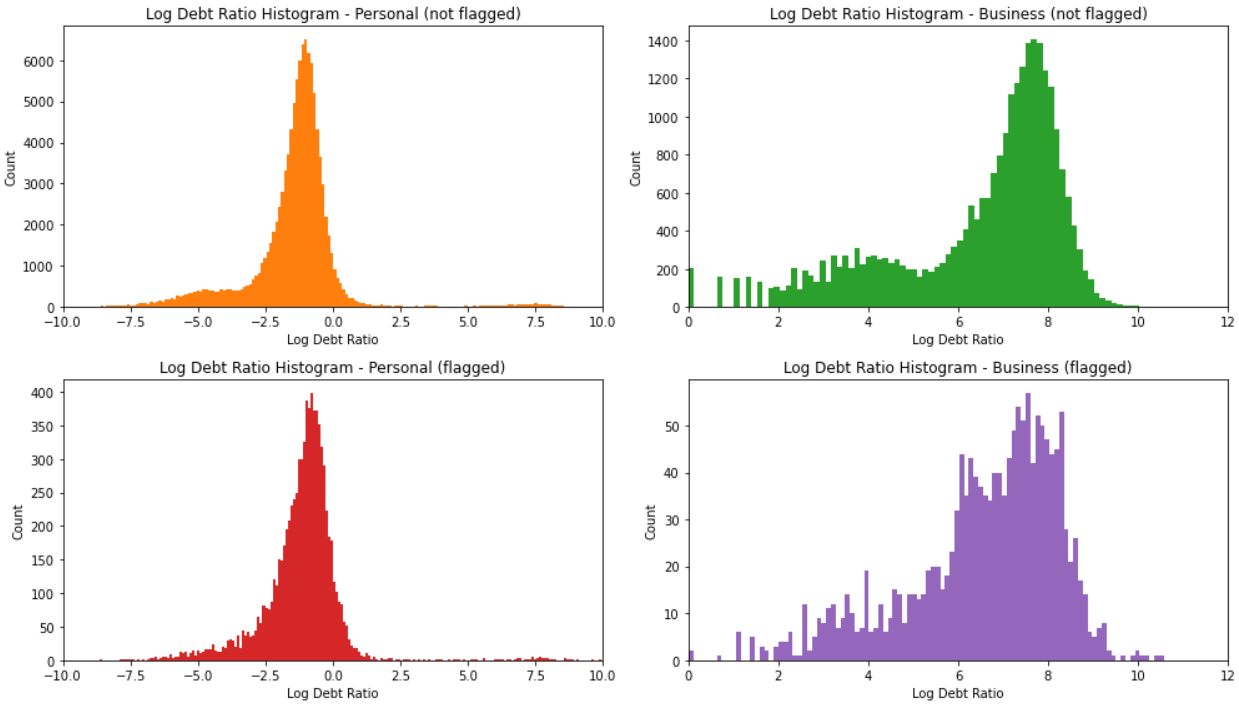
First, we exclude outliers first to compute a representative sample mean and variance for both populations.

Then, we use two-tailed F-test to test whether the variances for Non-Flagged and Flagged populations are the same. If we fail to reject the null hypothesis, we proceed to use a two-means t-test with equal variances; otherwise, we use the unequal one. Note that we approximate the t-distribution with standard normal due to large sample size (and thus large degrees of freedom).

```
Personal TypeCustomer
The null hypothesis (the variance of both populations are the same) is rejected
The null hypothesis (the means of both populations are the same) failed to be rejected

Business TypeCustomer
The null hypothesis (the variance of both populations are the same) is rejected
The null hypothesis (the means of both populations are the same) failed to be rejected
```

For both categories of TypeCustomer, the null hypothesis that the means of (log) DebtRatio from Non-Flagged and Flagged populations are the same failed to be rejected. Note that we rejected the null hypothesis that the variances of (log) DebtRatio from Non-Flagged and Flagged populations are the same.

When we look at the histograms, the results above make sense as there are no discernable differences on the distribution centers across the Non-Flagged and Flagged customers.

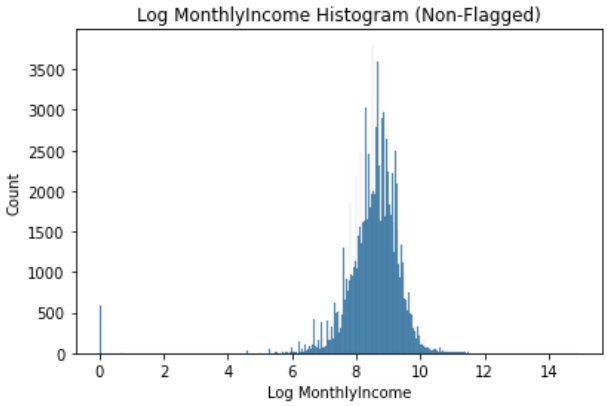


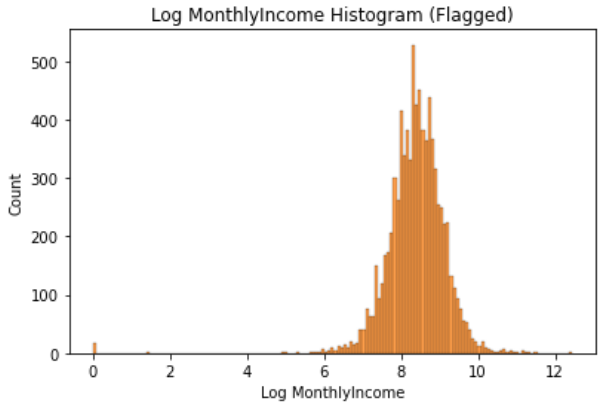

# Comparing the population mean of (log) MonthlyIncome for Non-Flagged vs Flagged populations

To test whether the population mean of (log) Income for Non-Flagged vs Flagged populations are equal, we apply the same method as for the DebtRatio test.

```
The mean of (Log) Monthly Income for Non-Flagged vs Flagged
The null hypothesis (the variance of both populations are the same) failed to be rejected
The null hypothesis (the means of both populations are the same) failed to be rejected
```

We failed to reject both null hypotheses (the variances are equal and the means are equal). When we look at the histograms, these results make sense as there is no (visually) significant difference between both of them.




# Conclusion

- The null-valued MonthlyIncome is due from the differences in customer type (Personal vs Business).
- The number of real estate loans (or lines) is approximately Poisson-distributed.
- Having zero open credit lines and loans increases the odds of being flagged (experiencing current-day 90 days late payments) by 4 times.
- Having experienced 30-59 days late payments in the last 2 years increases the odds of being flagged by 3 times.
- Having experienced 60-89 days late payments in the last 2 years increases the odds of being flagged by 6 times.
- Having experienced >90 days late payments in the last 2 years increases the odds of being flagged by 7 times.
- The log-transformed of debt ratio for Personal-type customers is approximately normally-distributed.
- Having greater than 1 revolving utilization of unsecured lines increases the odds of being flagged by 6 times.
- The log-transformed monthly income is approximately normally-distributed.
- There are two types of customers based on their loans: those whose no-installment type of loans dominated their total debts, and those whose installment-type of loans (e.g. real estate and car loans) dominated their total debts.

- The population means of (log) debt ratio for flagged and unflagged customers may be assumed as equal.
- The population means of (log) monthly income for flagged and non-flagged customers may be assumed as equal.

# Data Source

Give Me Some Credit | Kaggle (cs-training.csv)

# Github Repository

muhammadalifaqsha/creditscore-probabilityproject-pacmann: Python notebook code for Pacmann Probability course project (github.com)