# Executive Report

Nicholas Lewis, Sam Sheth, and Fareena Imamat

5/7/2020

## Overview

There is always a risk for credit card companies that a particular client defaults on his or her payments. In this study, we examine past bills, past payments, and select demographic characteristics for 24,518 clients. Our goal is to determine the best model that uses these variables to predict the likelihood that clients in a separate dataset will default on payments in the upcoming month, October 2005. We begin by processing and cleaning the data and performing exploratory data analysis in order to get a better sense of the data and examine relationships between variables. Then, we use our cleaned dataset to begin building models. We consider a variety of models with varying degrees of accuracy in predicting defaults, and we ultimately present the following four within this report: logistic regression, linear discriminant analysis (LDA), principal components analysis in LDA, and random forests. Our best performing model is the random forests model, and we are confident that it will be of significant value to credit card companies as they learn more about their clients and work to develop the best possible experience for them.

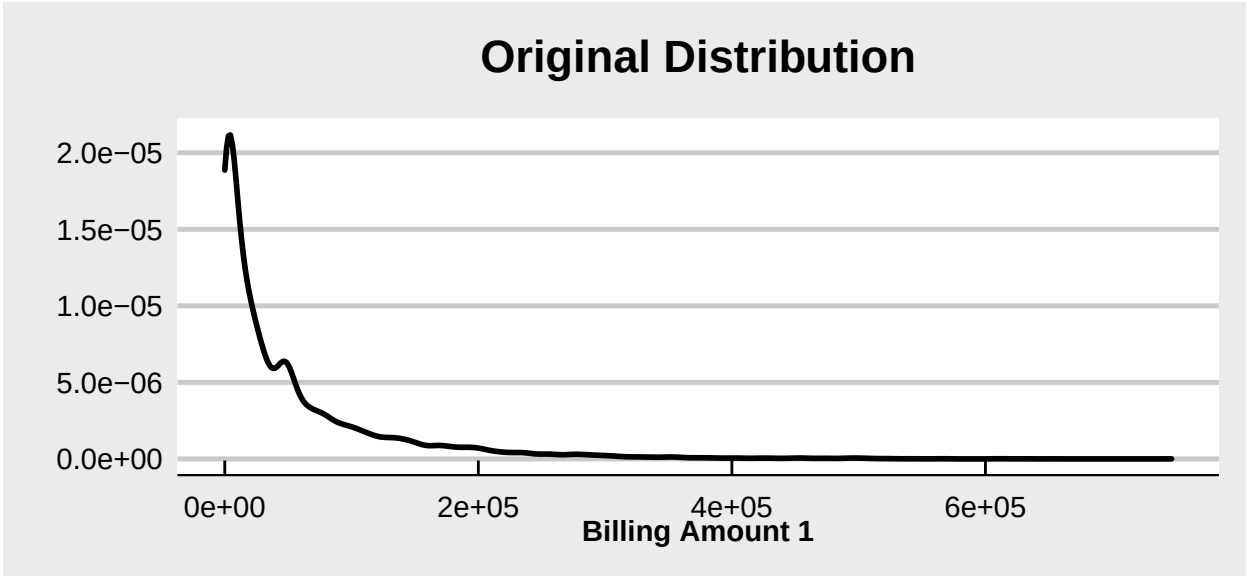
## Data Processing

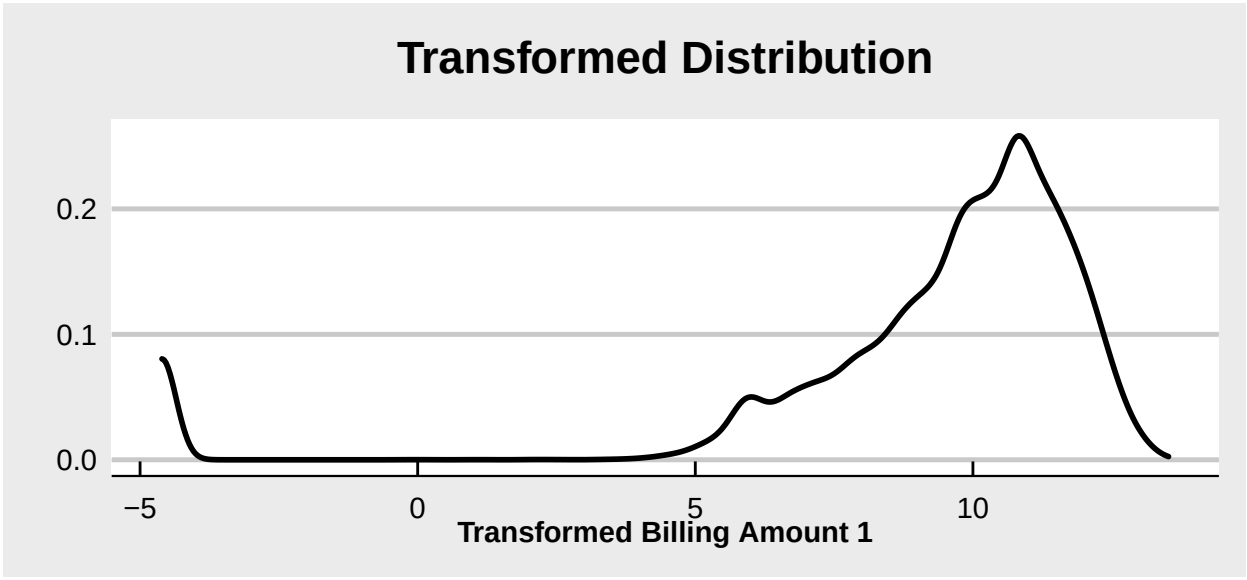
We first partitioned the data into two distinct sets: a training set and a testing set. The training data was used for all EDA and model building. Reported performance measures were calculated on the testing data.

Exploratory Data Analysis (EDA) and initial modelling attempts quickly illustrate the need for new features and transformation of existing features. First, simple density plots of the features 'bill_amt1', ..., 'bill_amt6', and 'pay_amt1', ..., 'pay_amt6' as well as age and balance limit are clearly not normally distributed. Logarithmic transformations are also taken in order to remedy this issue. There are several cases in which this transformation is problematic, however, since both the billing features and payment features contain zero values and bill amounts may also be negative. A slightly more complex version of this transformation is included below.

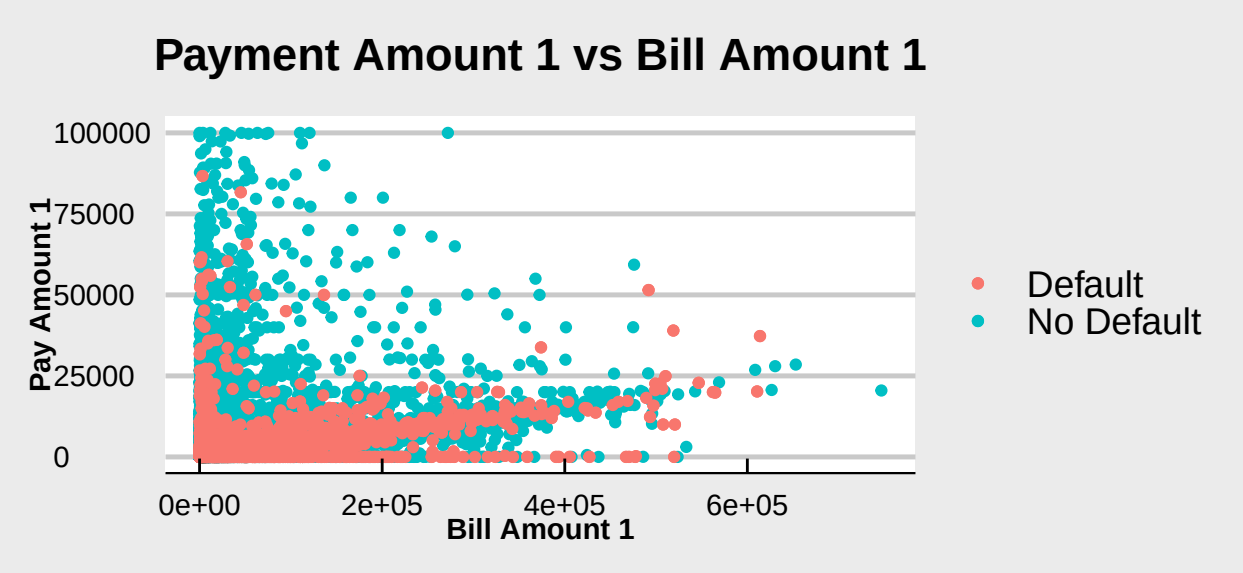
$$F(x) = \begin{cases} \log(x) & \text{if } x > 0 \\ \log .01 & \text{if } x = 0 \\ -\log|x| & \text{if } x < 0 \end{cases} \tag{1}$$

The transformation is illustrated graphically below for 'bill_amt1'. All plots illustrating this transformation are included in Appendix -. It far from a perfect normal distribution even after the transformation, which may present issues with modelling techniques that assume normality.




Next, we create variables denoting the percent paid for each month. Intuitively, this ratio may be more influential than the billing and payment amount are separately. The plot below, as well as the rest of the family of plots included in Appendix B, illustrates that there is some sort of relationship between these two, but that observations who do not default do not adhere to this relationship very strongly. The noise extends above the top of this graph for those who do not default.

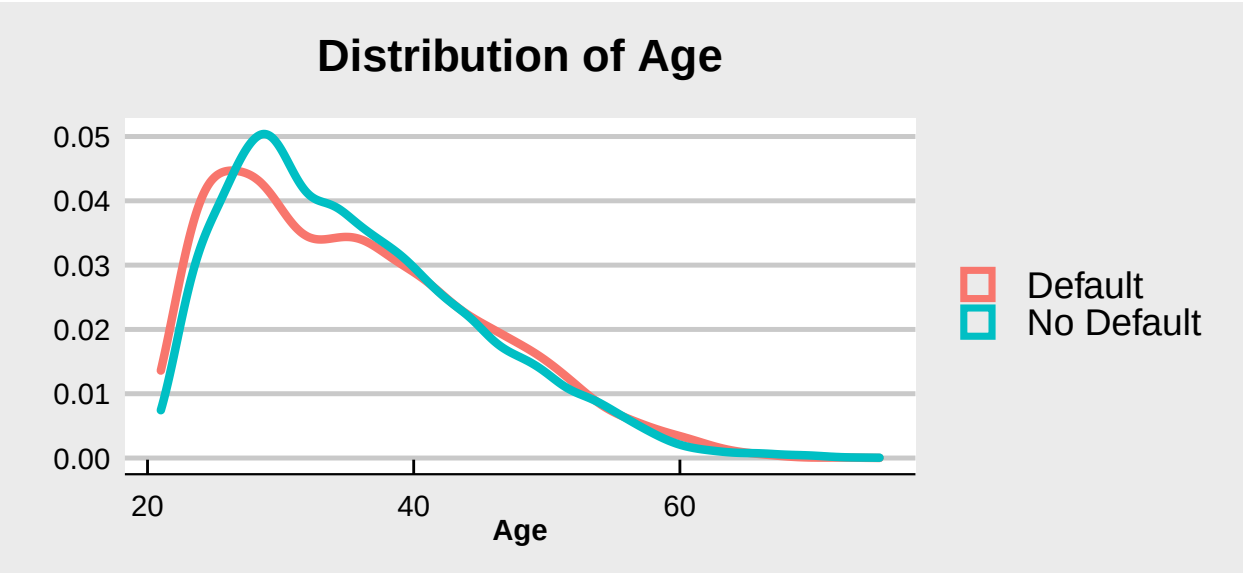



Again, observations which have a transformed billing amount of zero present a challenge. This is dealt with through the logic below.

$$F(x) = \begin{cases} \frac{\log(\text{Amount Paid})}{\log(\text{Amount Billed})} & \text{if } x = 0 \\ \frac{\log(\text{Amount Paid})}{\log(.01)} & \text{otherwise} \end{cases} \tag{2}$$

Extra features are also created for the log transformation of the mean billing amount over the six months and the log transformation of the mean payment amount. This is done to provide a more concise representation of the billing and payment info if needed. The logic for the transformation of negative and zero amounts applies to this transformation as well.

Then, looking at the density plots of age by the value of the default value shows that there are age domains that are more likely to contain defaulting observations than others. To potentially reinforce these distinctions, we create indicator variables for the inclusion of an observation in the ranges [0, 27), [27, 40), [40, 55), and [55, inf).



We use a similar methodology to create indicator variables for high and low pay (divided at \$2980) and for high and low billing limits (divided at $\$1.25 * 10^5$). Similar plots for these variables are included in Appendix C.

## Modeling

### Logistic Regression

### Linear Discriminant Analysis (LDA)

### Quadratic Discriminant Analysis (QDA)

### LDA with Principal Component Analysis (PCA)

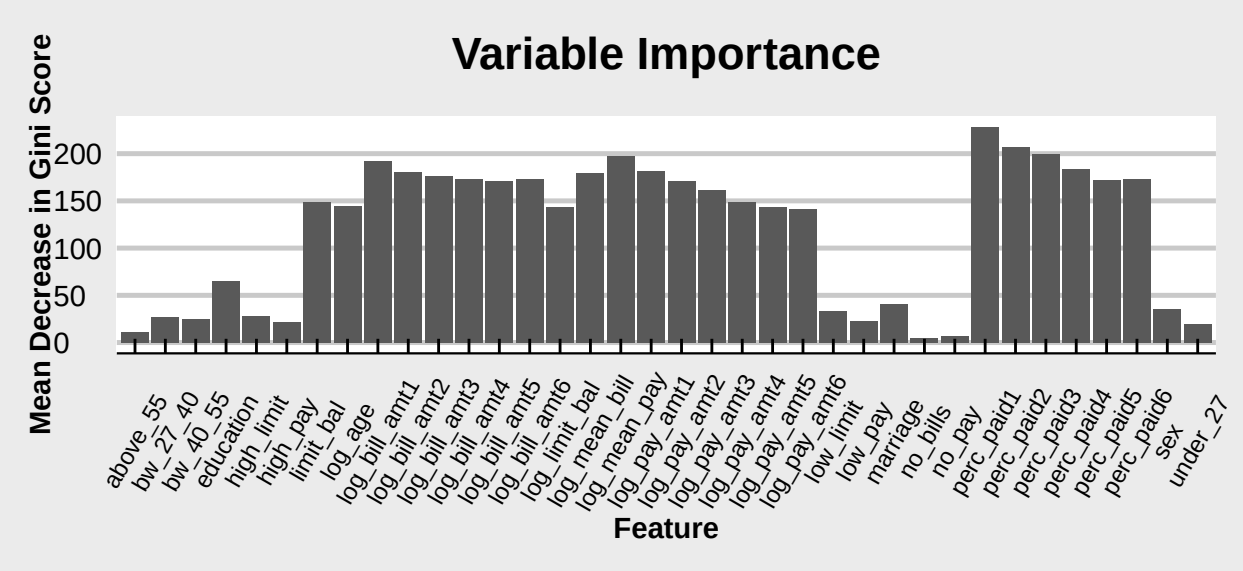
We attempt to create an LDA model that utilizes Principal Component Analysis (PCA). We perform PCA on the `bill_amt` family and `pay_amt` family, as originally presented in the data. Because the best subset selection process in the previous LDA model did not select any of the indicator variables created during the data processing stage, we build this LDA model on the PCA components and the original qualitative variables `age`, `sex`, and `education`.

This model produces predictions on the test data with a log loss score of 0.4993. However, a closer look at the results reveals that all observations are predicted to not default. This is the same behavior as the trivial model. Therefore, although the log loss is better than the trivial model, it is only because the probabilities are smaller on average. In other words, based off this model we are more confident that all observations will not default. In some sense, that actually makes this model less useful than the trivial model.

### Random Forest

We first used the `train()` function supplied by the `caret` package to obtain the optimal values for the number of features considered at each split of each tree `m` and the number of trees `B`. This process was fairly computationally expensive, but produced a final model with $m = 2$ and $B = 500$.

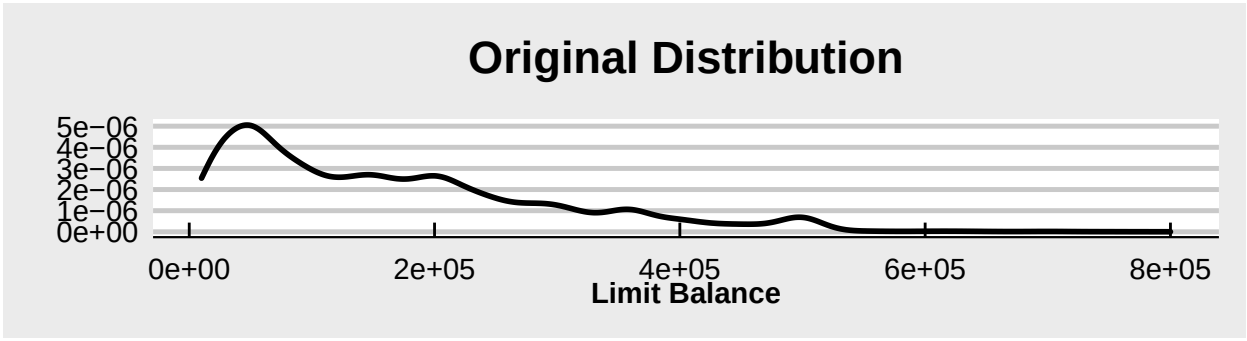
At this point, the log loss is fairly impressive, 0.4607 to be exact. However, we see that many of the variables created during the data processes are not important to the model (as displayed below). A low importance means that the feature was rarely used to make decisions in the forest because it was not a powerful predictor. However, because of the random nature, these low importance features will still be chosen at times. Removing them may allow more predictive features to be used more often, effectively increasing the power of our model.

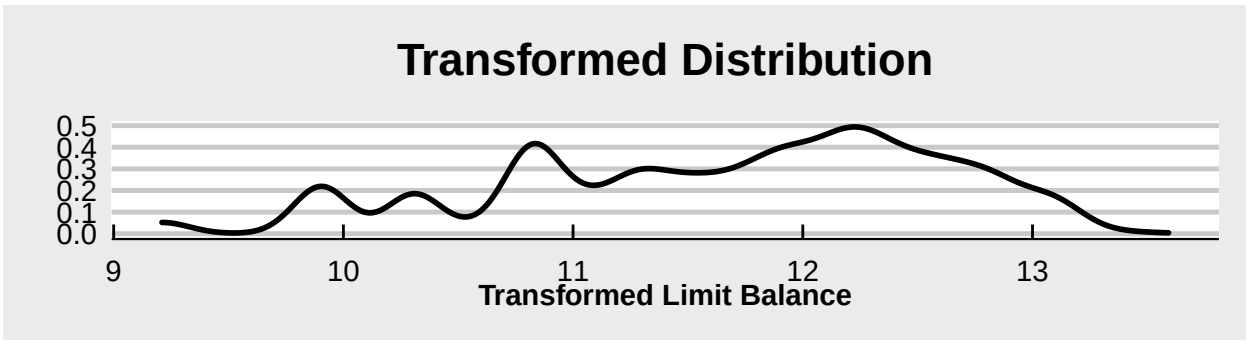


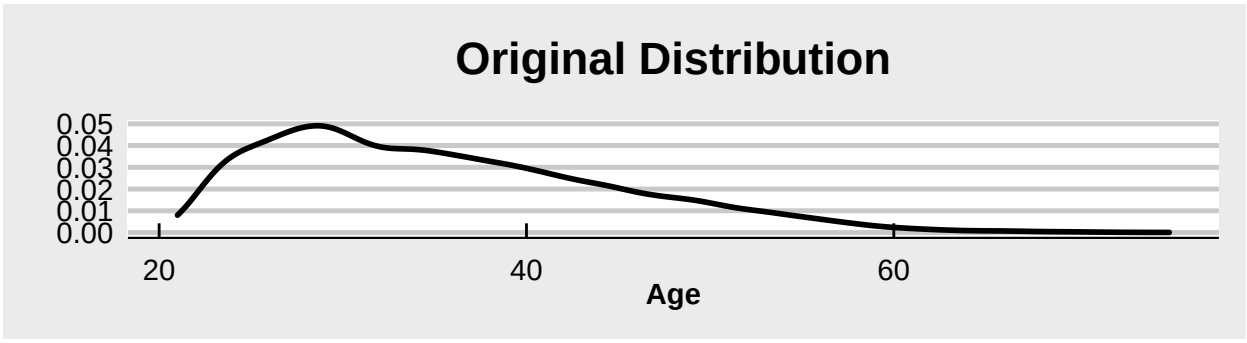

We remove the lowest variables, rebuild the model with $m = 2$ and $B = 500$. By doing this, we now obtain a log loss score of 0.4548.

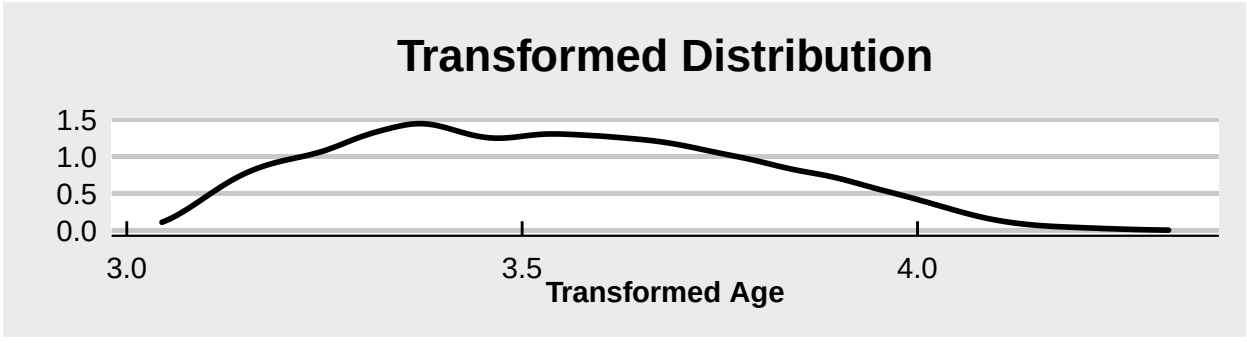
## Appendix A

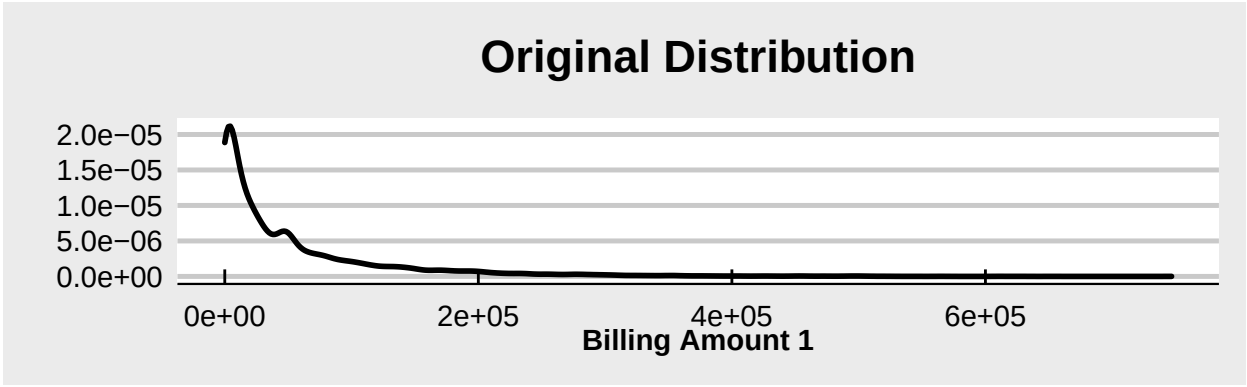
The following are plots of the variables transformed according to equation (1) before and after the transformation.

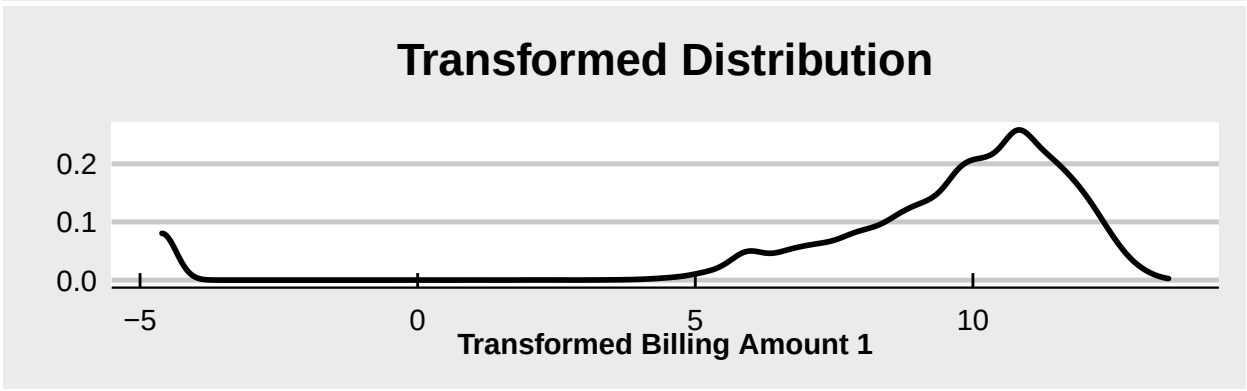


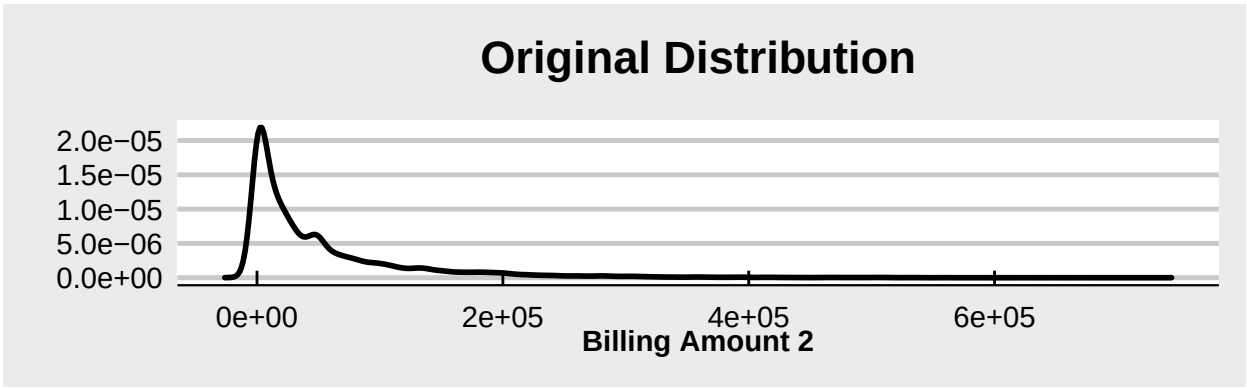



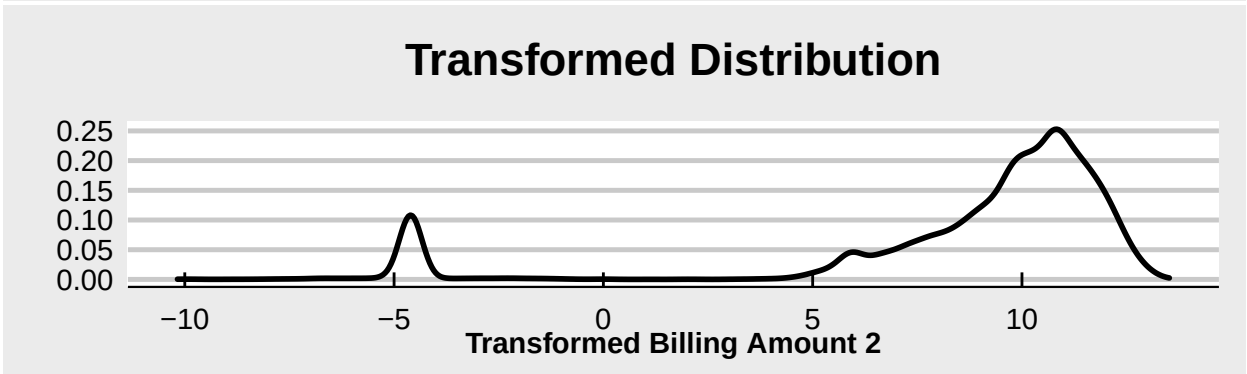



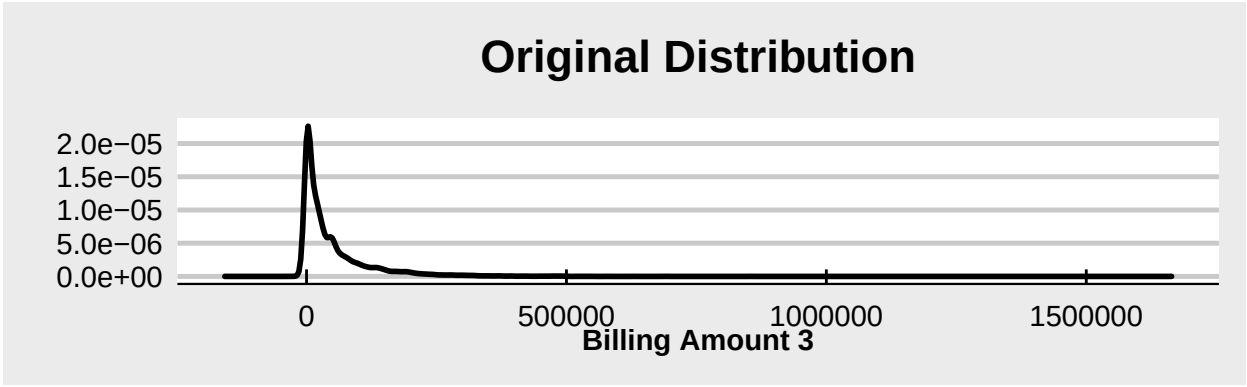



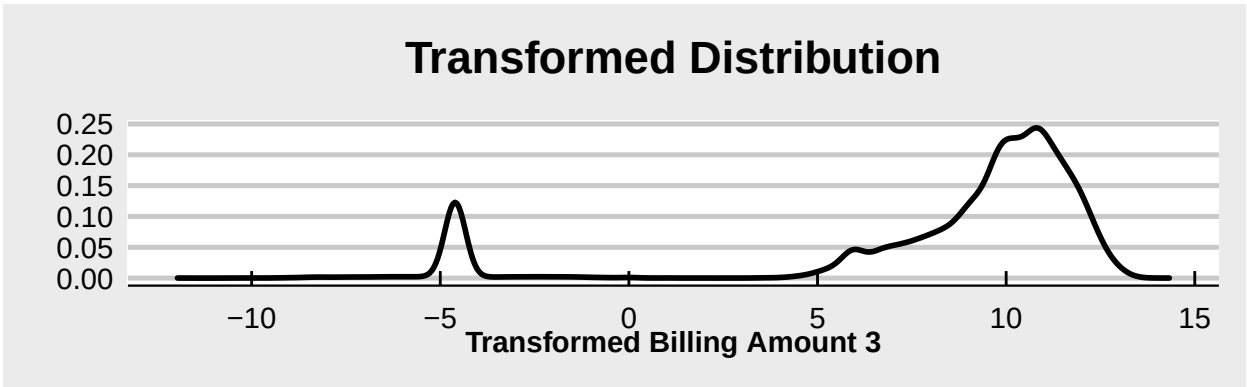

Original Distribution
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
0e+00
2e+05
4e+05
6e+05
Billing Amount 1

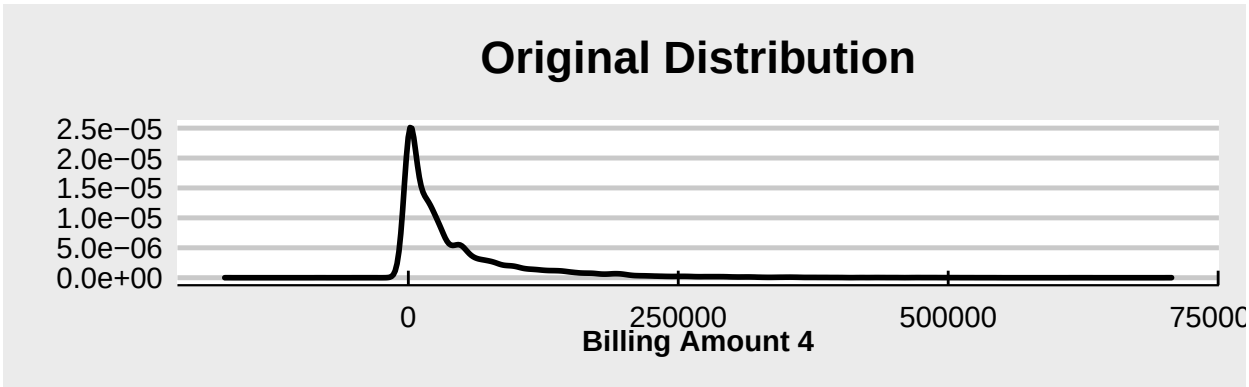

Transformed Distribution
0.2
0.1
0.0
−5
0
5
10
Transformed Billing Amount 1

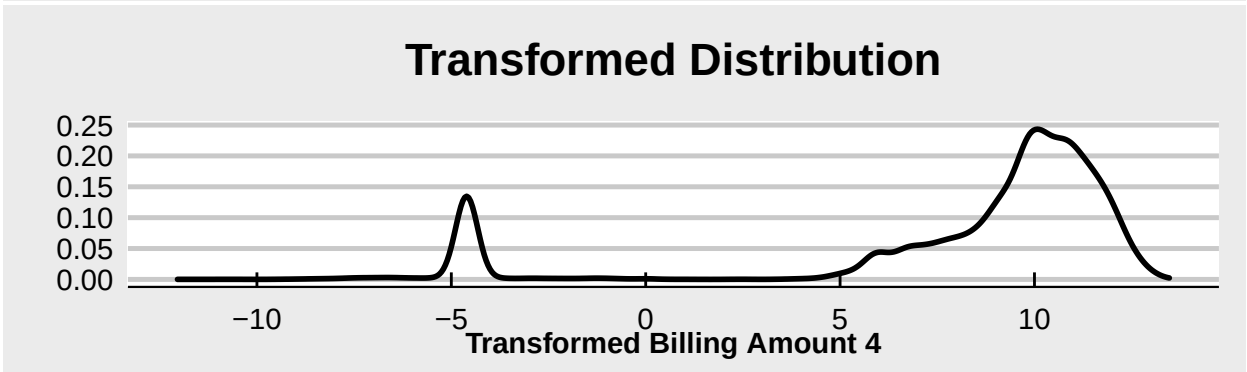

Original Distribution
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
0e+00
2e+05
4e+05
6e+05
Billing Amount 2

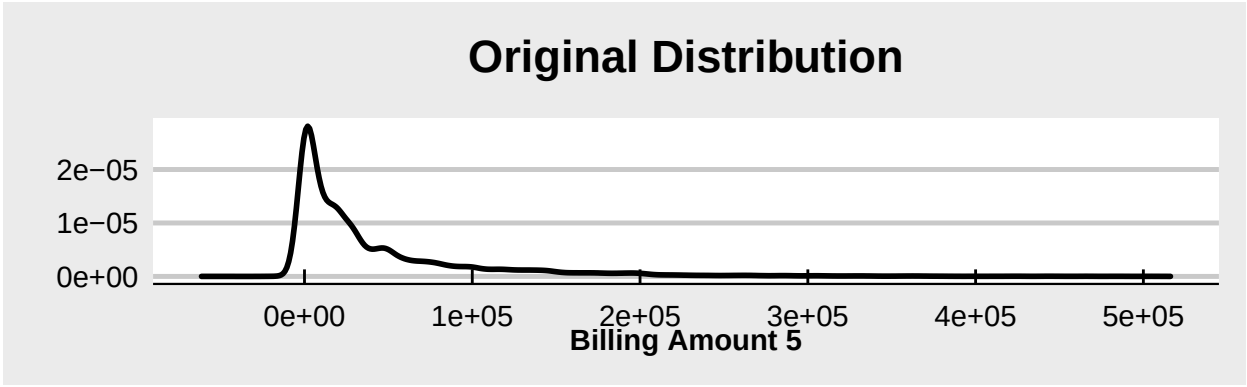

Transformed Distribution
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
Transformed Billing Amount 2

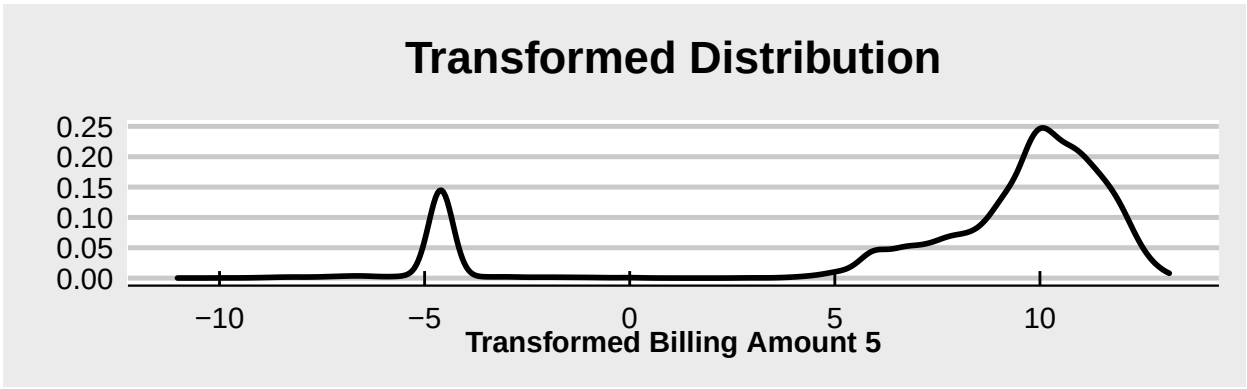

Original Distribution
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
0
500000
1000000
1500000
Billing Amount 3

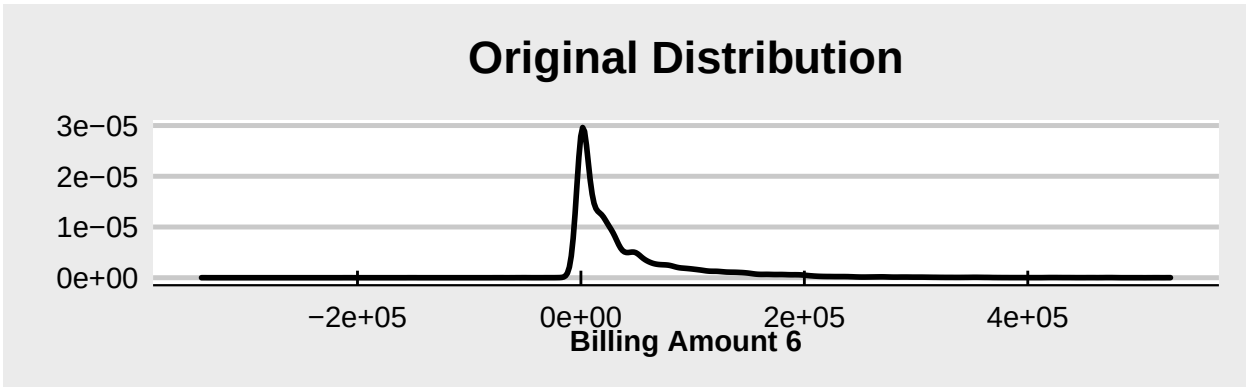

Transformed Distribution
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
15
Transformed Billing Amount 3

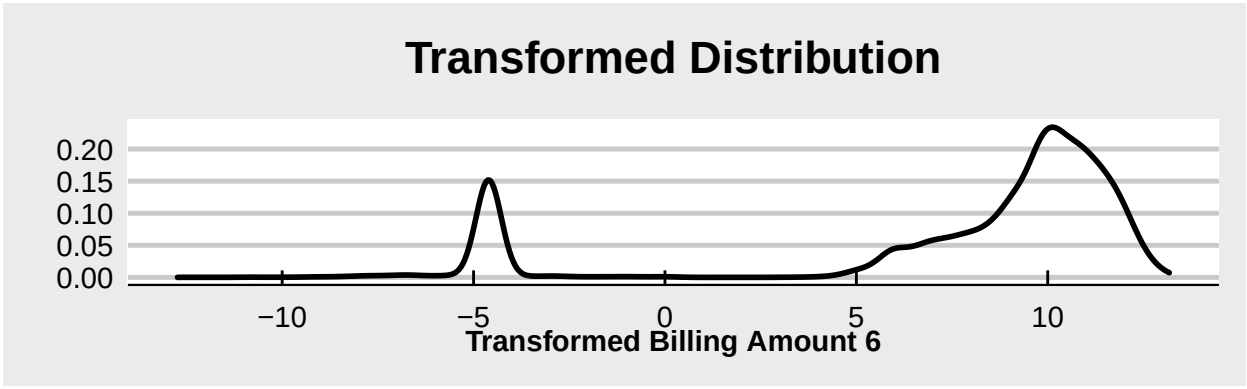

Original Distribution
2.5e−05
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
0
250000
500000
75000
Billing Amount 4

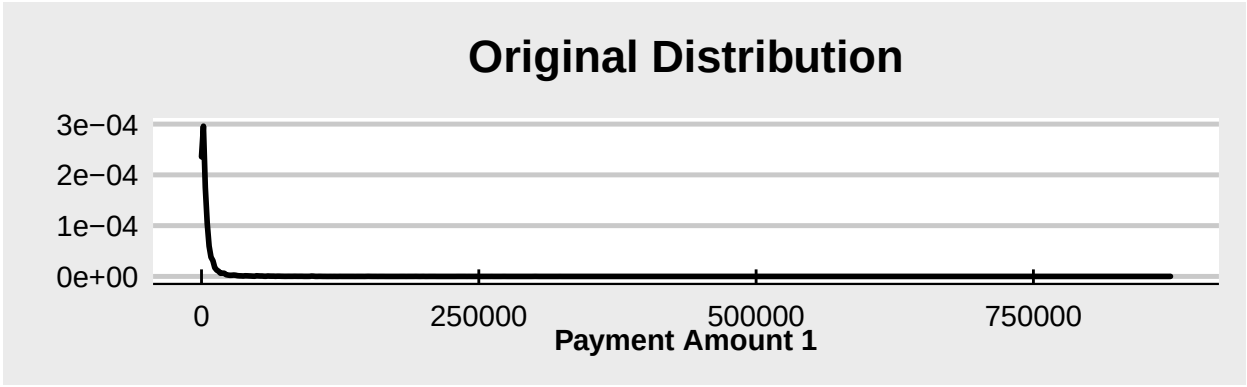

Transformed Distribution
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
Transformed Billing Amount 4

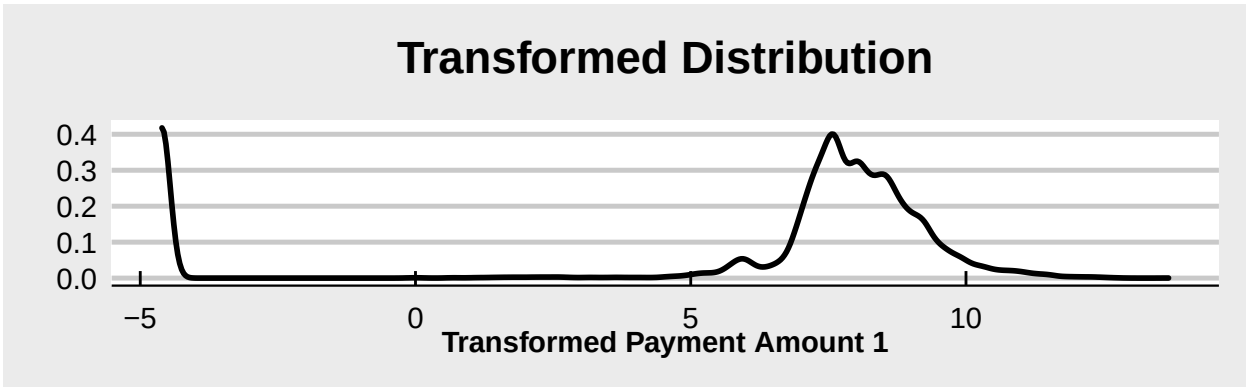

Original Distribution
2e−05
1e−05
0e+00
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Billing Amount 5

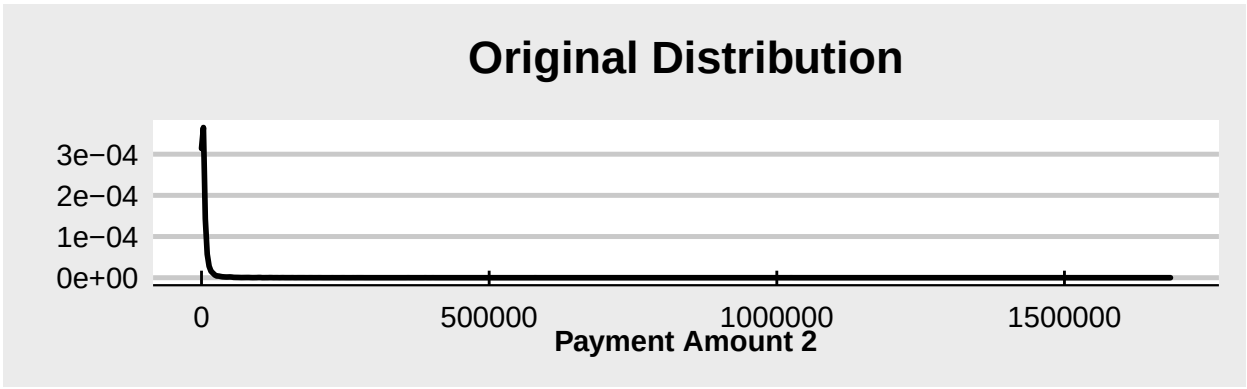

Transformed Distribution
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
Transformed Billing Amount 5

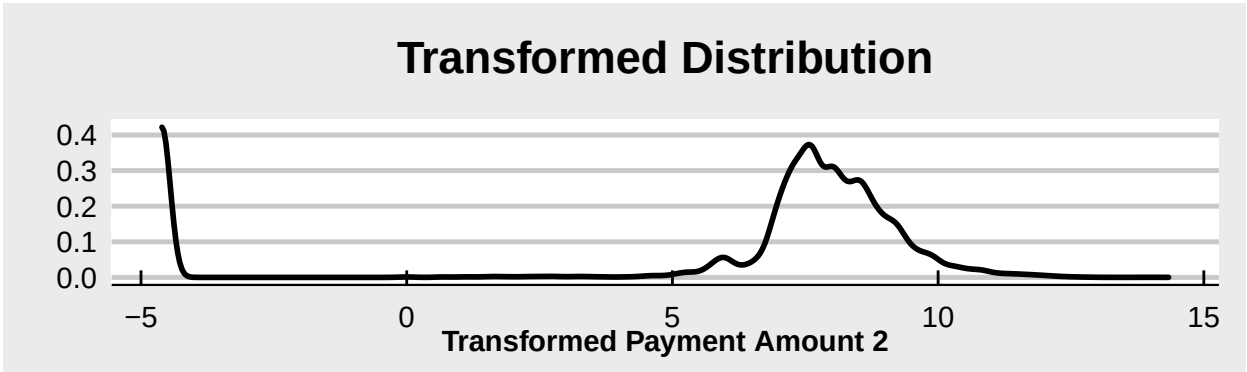

Original Distribution
3e−05
2e−05
1e−05
0e+00
−2e+05
0e+00
2e+05
4e+05
Billing Amount 6

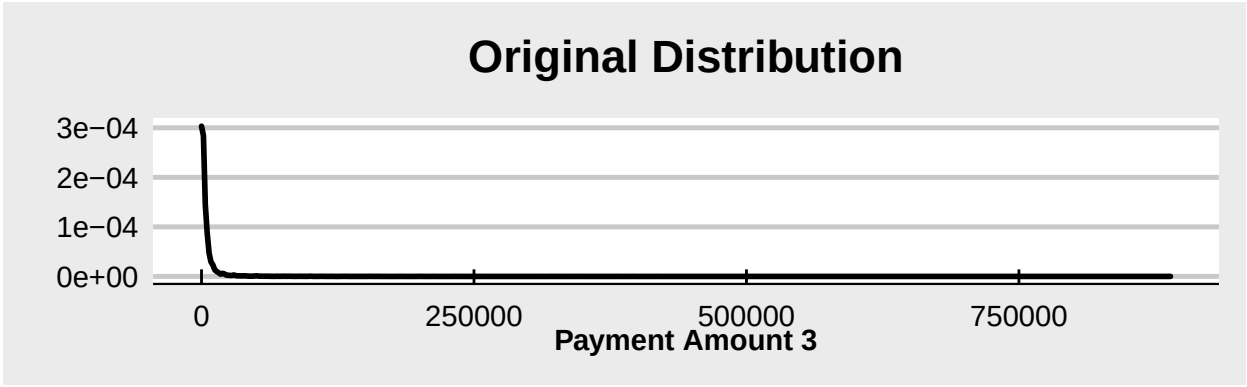

Transformed Distribution
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
Transformed Billing Amount 6

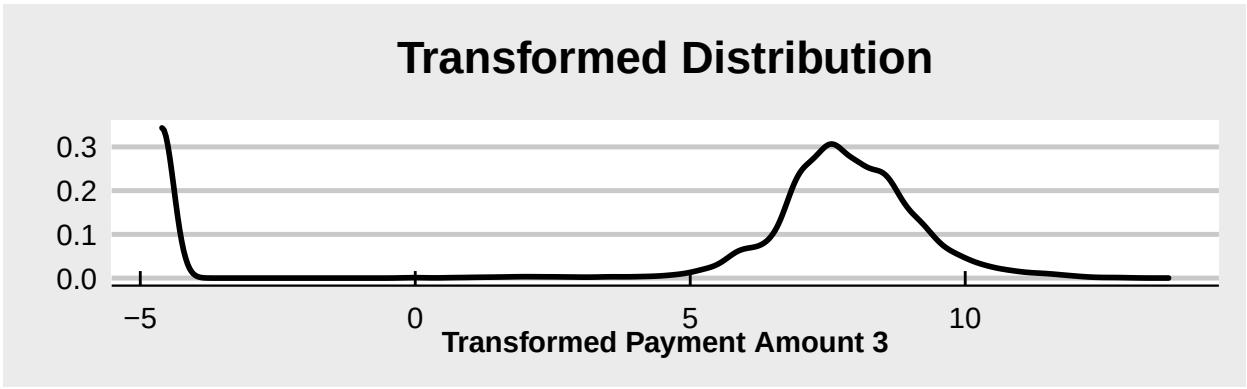

Original Distribution
3e−04
2e−04
1e−04
0e+00
0
250000
500000
750000
Payment Amount 1

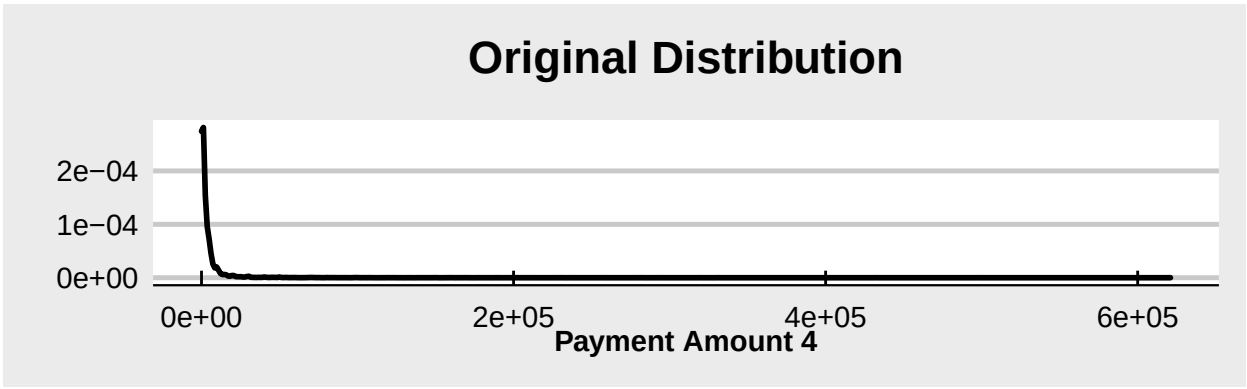

Transformed Distribution
0.4
0.3
0.2
0.1
0.0
−5
0
5
10
Transformed Payment Amount 1

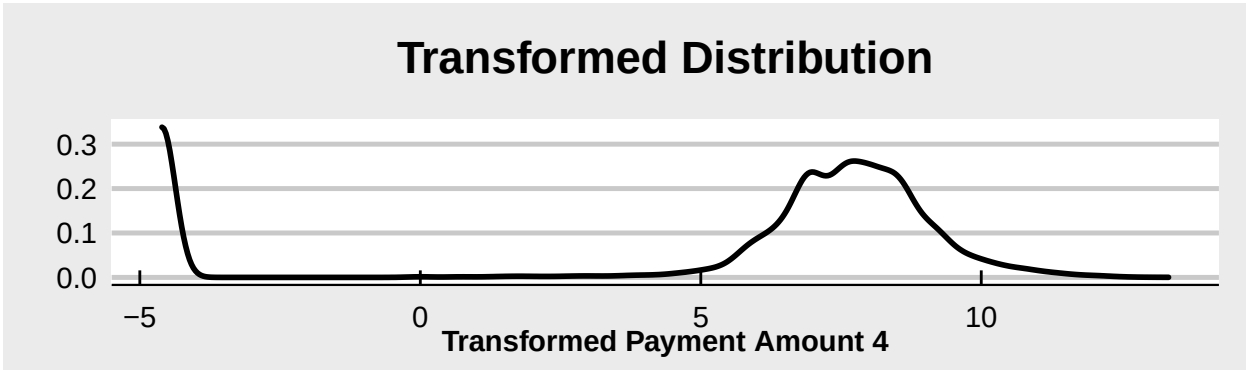

Original Distribution
3e−04
2e−04
1e−04
0e+00
0
500000
1000000
1500000
Payment Amount 2

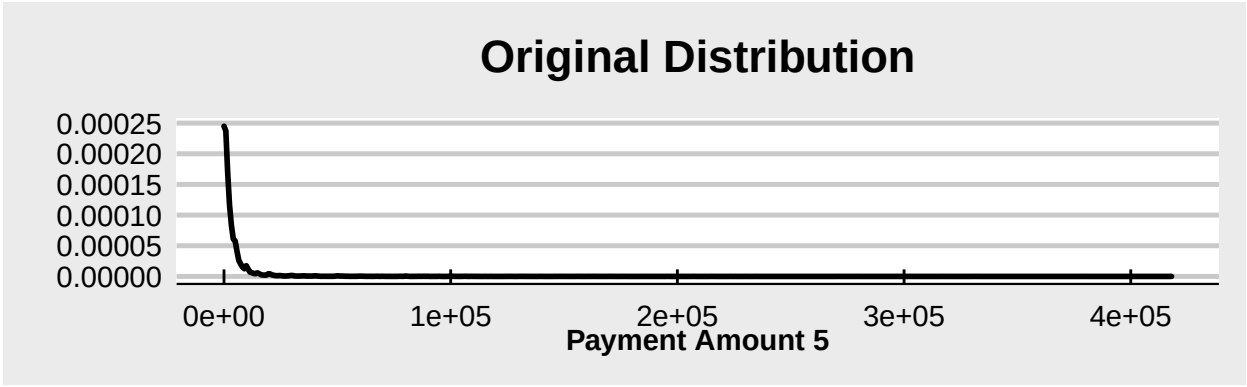

Transformed Distribution
0.4
0.3
0.2
0.1
0.0
−5
0
5
10
15
Transformed Payment Amount 2

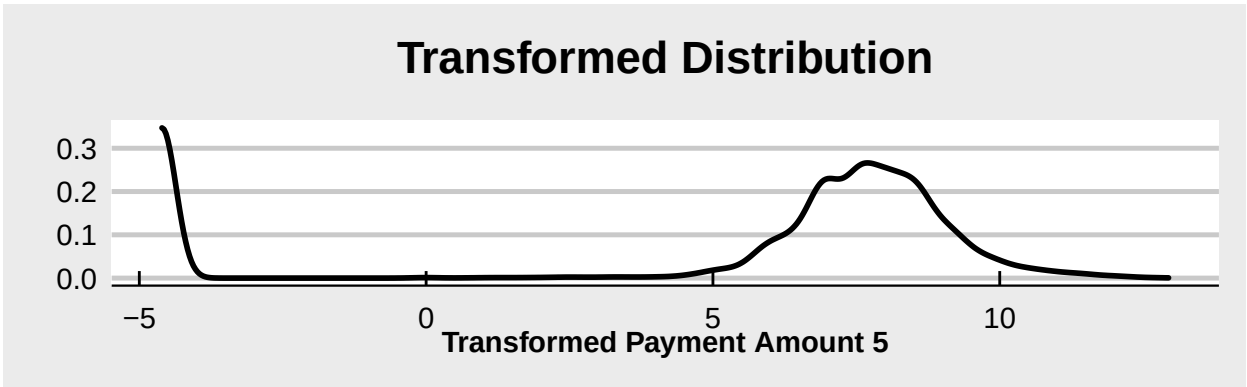

Original Distribution
3e−04
2e−04
1e−04
0e+00
0
250000
500000
750000
Payment Amount 3

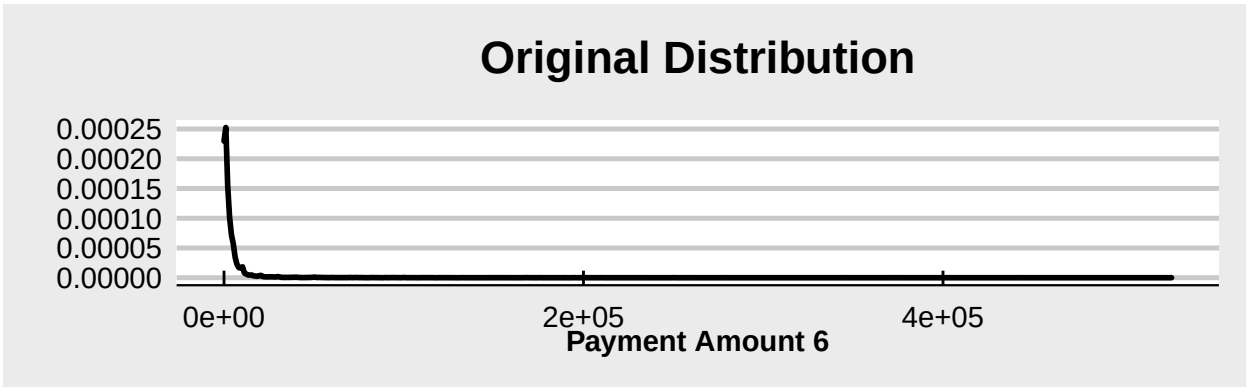

Transformed Distribution
0.3
0.2
0.1
0.0
−5
0
5
10
Transformed Payment Amount 3

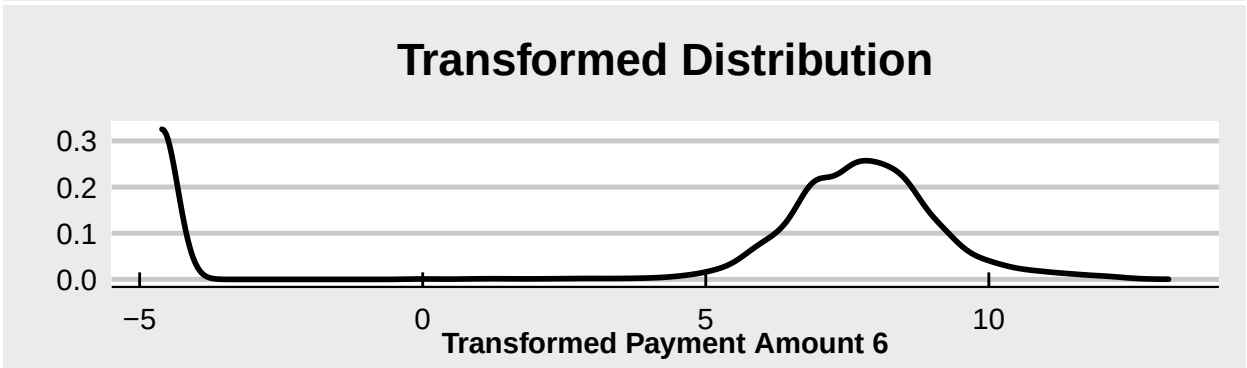

Original Distribution
2e−04
1e−04
0e+00
0e+00
2e+05
4e+05
6e+05
Payment Amount 4


Transformed Distribution
0.3
0.2
0.1
0.0
−5
0
5
10
Transformed Payment Amount 4


Original Distribution
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0e+00
1e+05
2e+05
3e+05
4e+05
Payment Amount 5

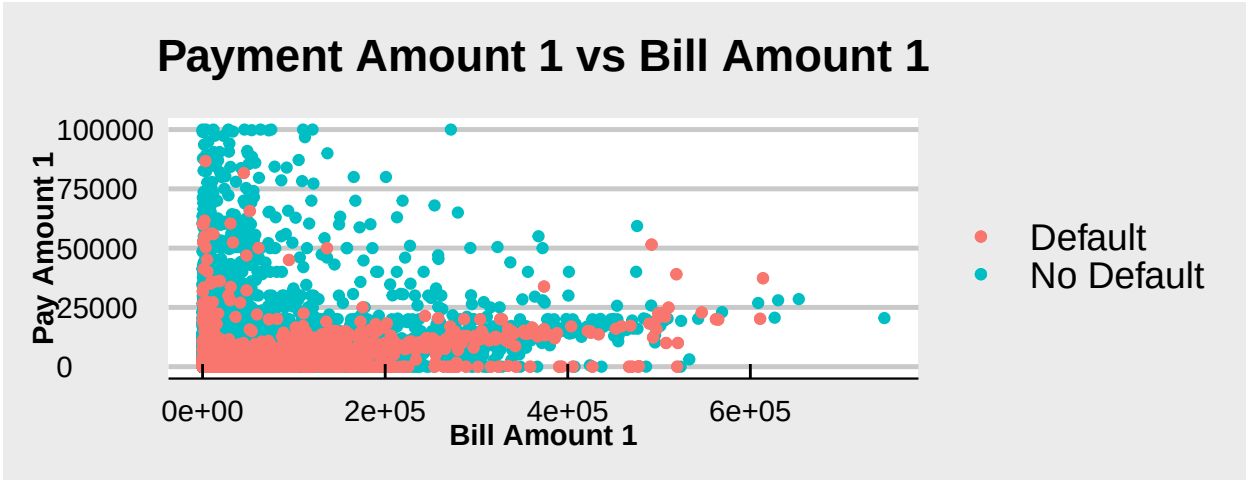

Transformed Distribution
0.3
0.2
0.1
0.0
−5
0
5
10
Transformed Payment Amount 5

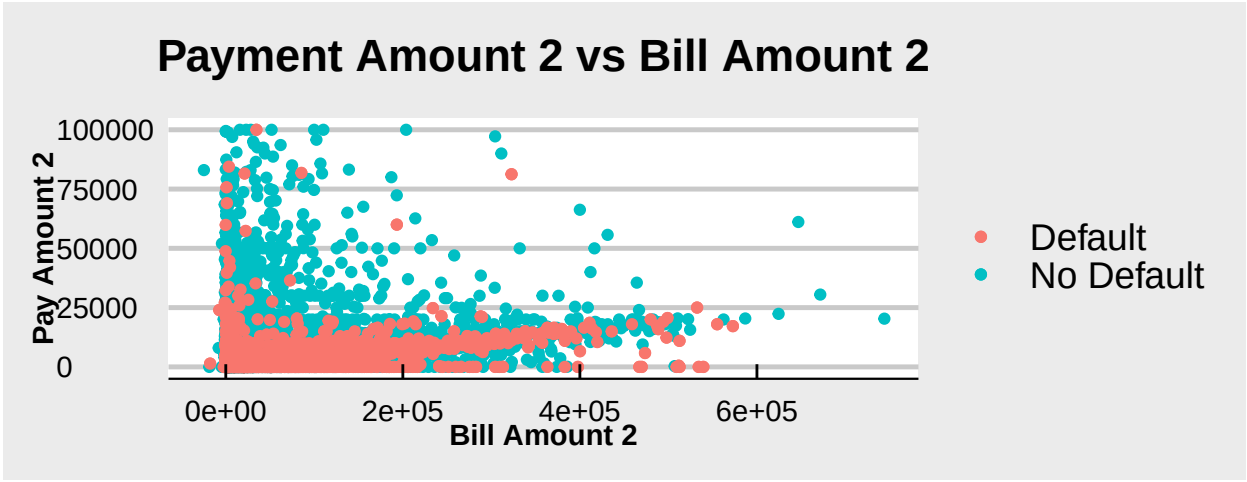

Original Distribution
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0e+00
2e+05
4e+05
Payment Amount 6

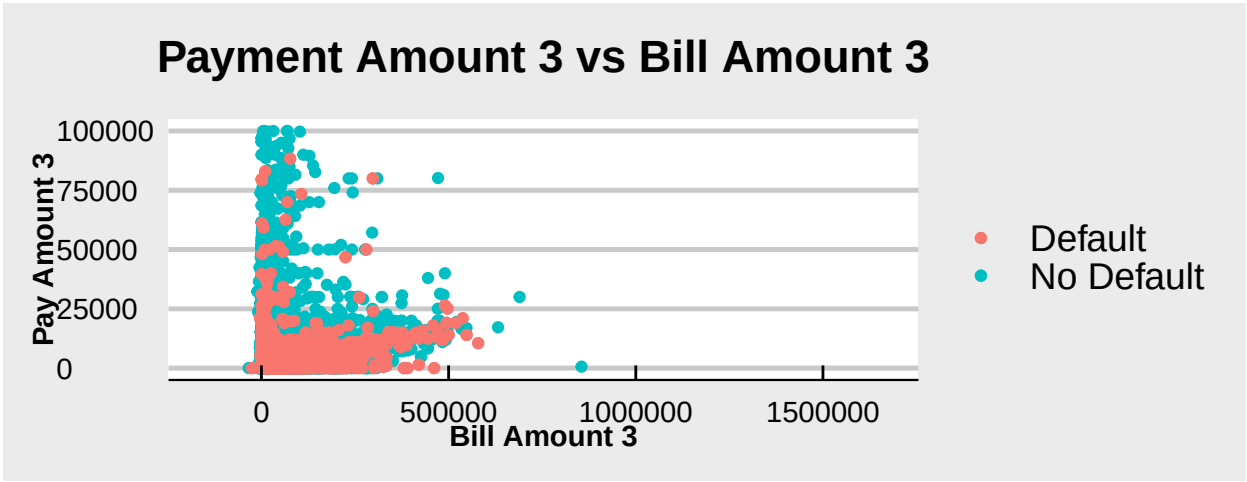

Transformed Distribution
0.3
0.2
0.1
0.0
−5
0
5
10
Transformed Payment Amount 6

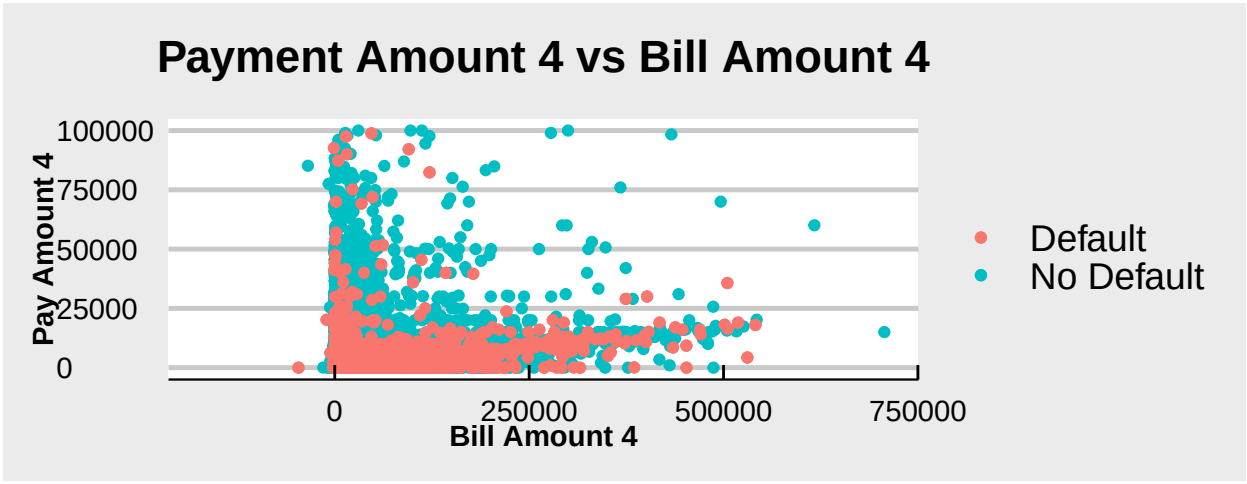
## Appendix B

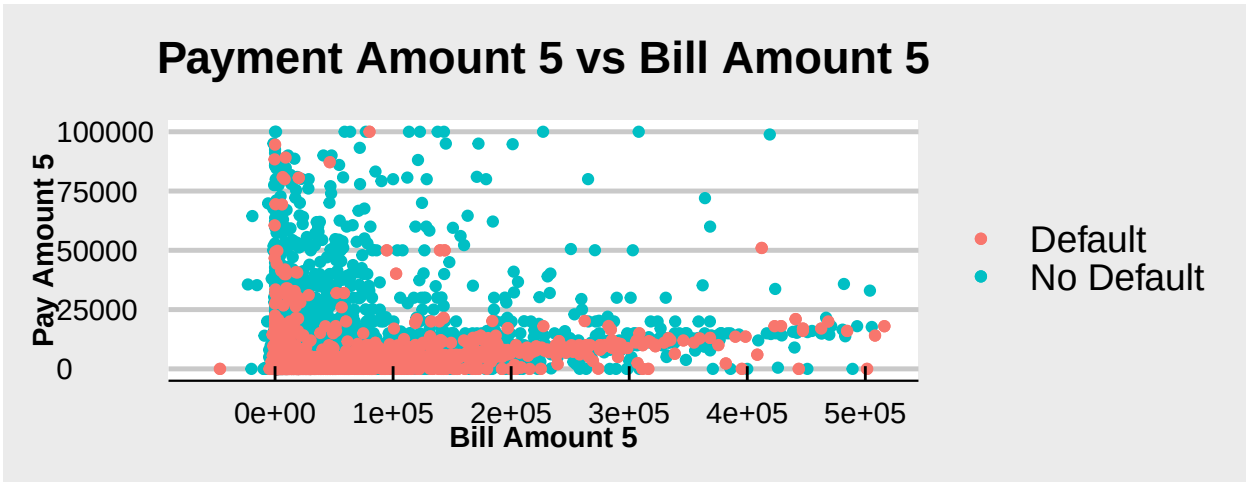
Below are the family of graphs visualizing the relationship between the transformed bill amounts and payment amounts.

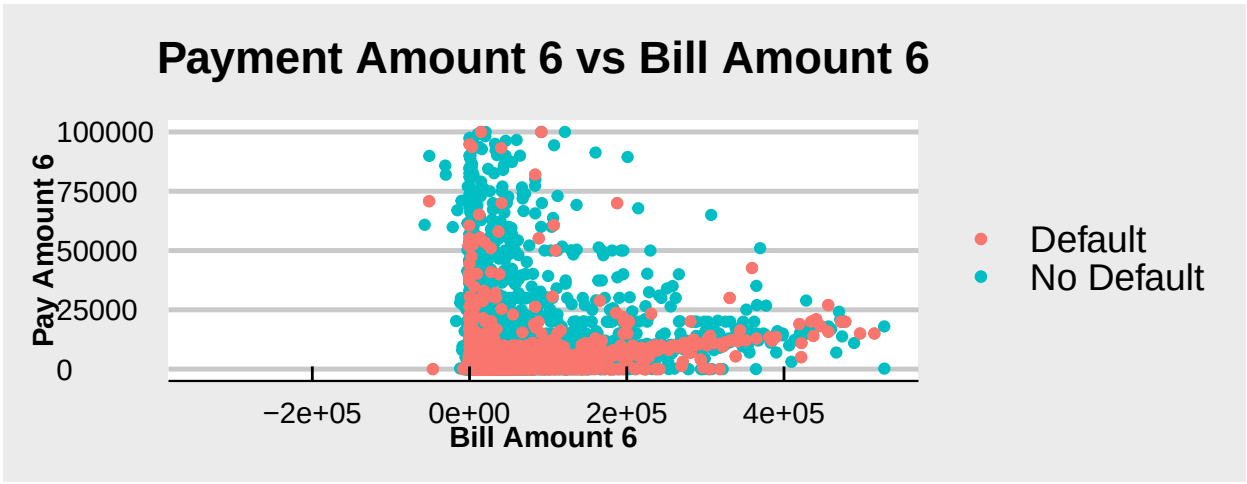



Payment Amount 4 vs Bill Amount 4
Pay Amount 4
100000
75000
50000
25000
0
0
250000
500000
750000
Bill Amount 4
Default
No Default

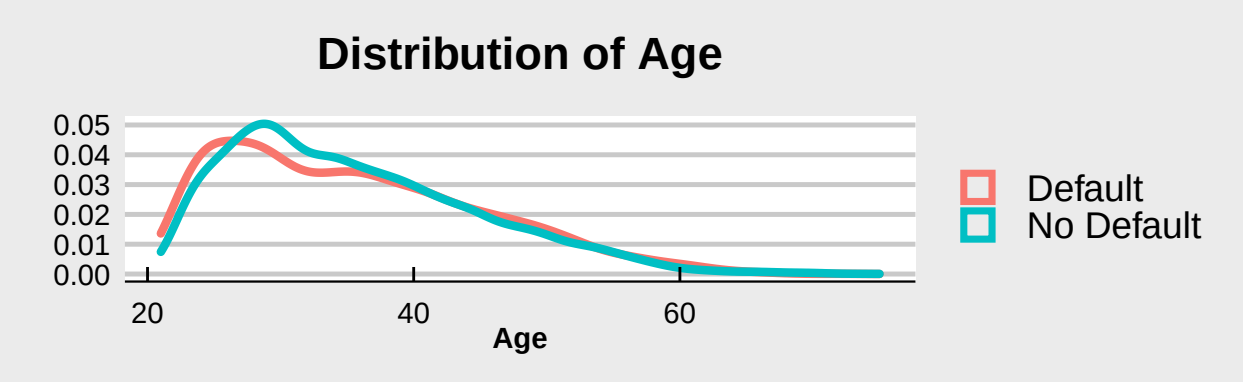

Payment Amount 5 vs Bill Amount 5
Pay Amount 5
100000
75000
50000
25000
0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Bill Amount 5
Default
No Default


Payment Amount 6 vs Bill Amount 6
Pay Amount 6
100000
75000
50000
25000
0
−2e+05
0e+00
2e+05
4e+05
Bill Amount 6
Default
No Default

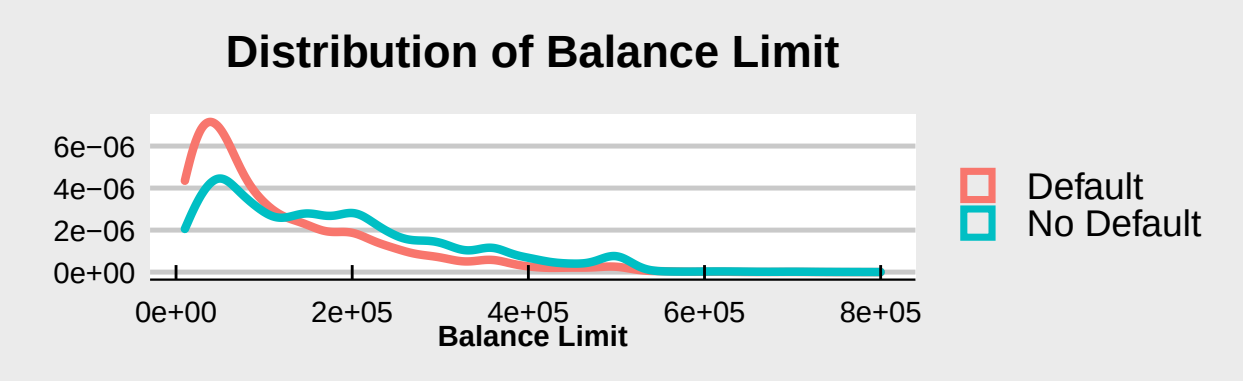
## Appendix C

The following plots support our choices to make indicator variables for certain variables.

Based on the graph below, we made indicators for the ranges [0, 27), [27, 40), [40, 55), and [55, $\infty$).




Based on the graph below, we made indicators for the ranges of [0, $\$1.25 * 10^5$) and [ $\$1.25 * 10^5$, $\infty$).




Based on the graph below, we made indicators for the ranges of [0,8) and [8,$\infty$)