Capstone 1 : Data Story
Matthew Gudorf

1. General exploration and data visualization

The dataset containing loan information that I am using (the Lending Club loan dataset from (https://www.kaggle.com/wordsforthewise/lending-club), contains categorical variables in the form of strings, discrete numerical variables, and continuous numerical variables. The data can also be segmented into information describing the borrower (geographical location, income, etc.) and information about the loan (principal amount, interest rate, etc.). The goal of this document is to develop an understanding of the data. There are also features that are time dependent and can be naturally visualized as a time series. One example of a time dependent feature is the date when each loan was issued. By aggregating by month, we can see that there is an upwards trend in the number of issued loans. As the number of loans issued increases, it becomes more important to prevent loan defaults, as the amount of capital they represent also increases (assuming a constant proportion of issued loans default over time).

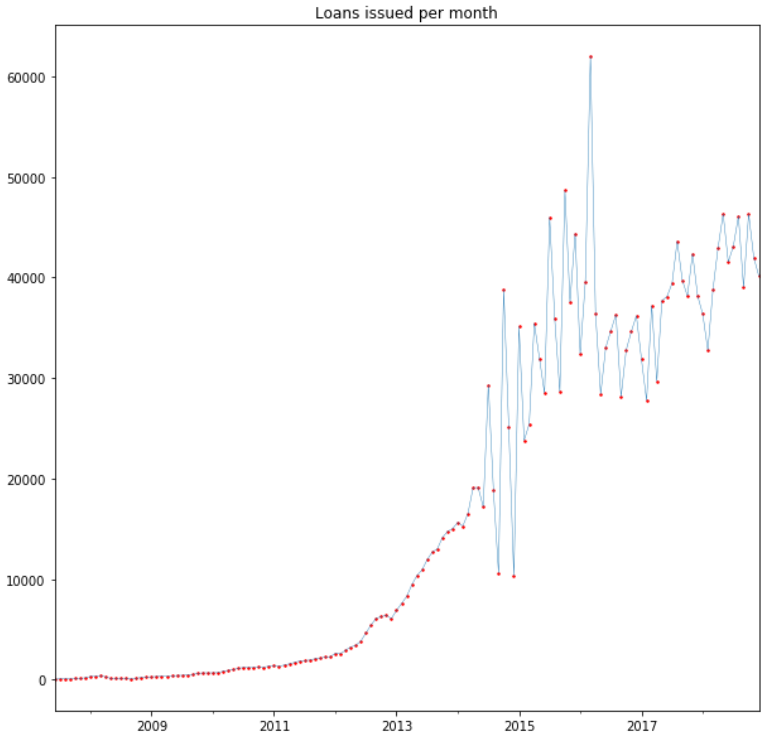

Figure 1 : Loans issued per month over time.

This is only a figure to give a sense of how the loans have grown up to the present moment. Aggregating by the (top three) statuses of the loan shows the time distribution of when fully paid, current and charged off loans. All of the loans have terms of three or five years so it's understandable that the approximate maximum of the fully paid loans and charged off loans occurs in the same range, three to five years ago. (The most current date in the data set is December 1st 2018).

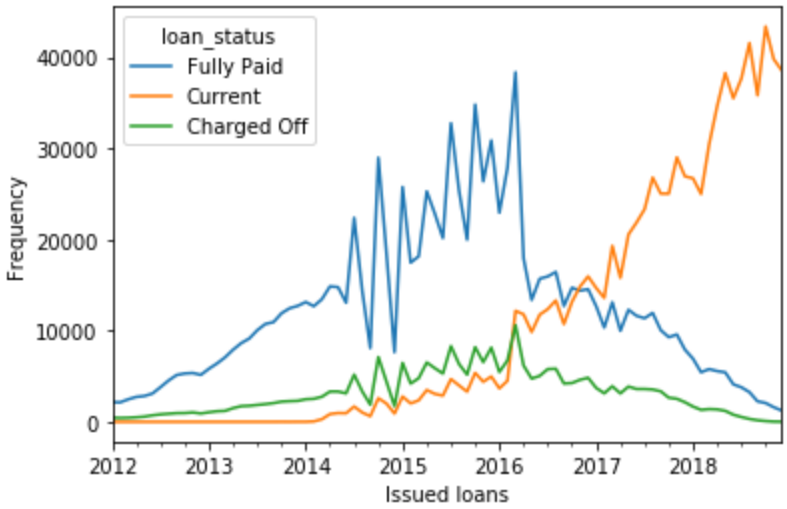

Figure 2: Loan status (three main categories) time series

Continuing with the data exploration, a unique and interesting quantity is the distribution of loan amounts; it really shows the affect of human psychology by how the distribution manifests.

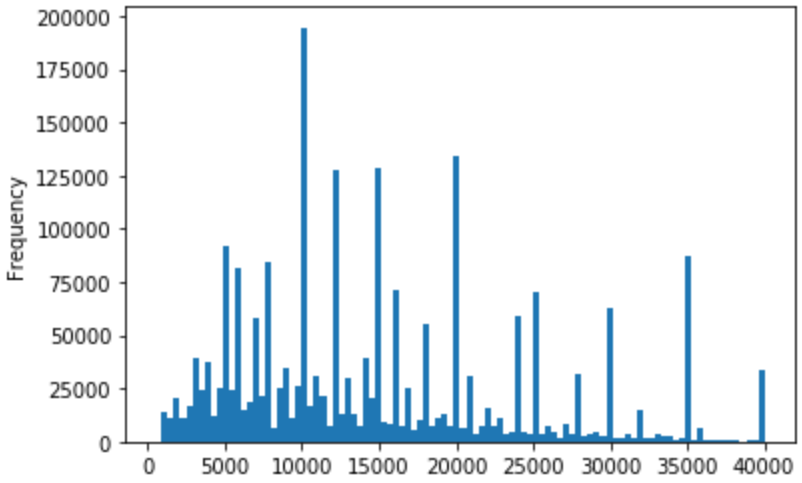

Figure 3 : Distribution of loan principal amounts

Why is this distribution so strange? As can be seen, the distribution seems to be clustered around "pretty" numbers. This includes round numbers like multiples of five thousand and ten thousand. To show this explicitly, let's look at the top three most frequent loan values.

Amount	Count
10000	187236
20000	131006
15000	123226

Table 1 : Top three most common principal loan amounts

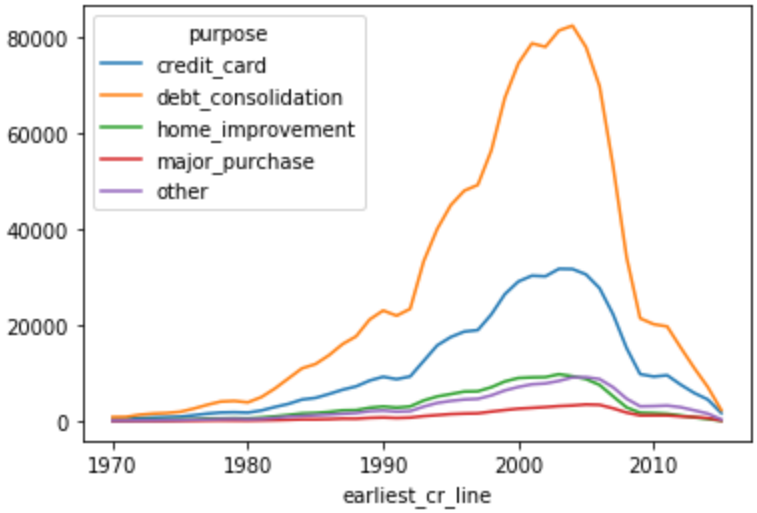

Figure 4 : Loan purposes over time

There doesn't seem to be any distinctive behavior between loan purposes, this would have been evidenced by difference in the mean of each the time series. The goal of stratifying the earliest credit line by the purpose of the loan was to see if different age demographics use loans for different purposes. This would be measured by differences in the mean between these stratified distributions, as I believe it is a fair assumption that everyone gets their first credit line (earliest) around the same age.

For the categorical variables the frequency of each category can be displayed in a histogram. Other features were investigated such as the distribution of utilized credit, the distribution of grades and subgrades of loans, the geographical distribution by zip code of the borrower, the date of the earliest credit line of the borrower.

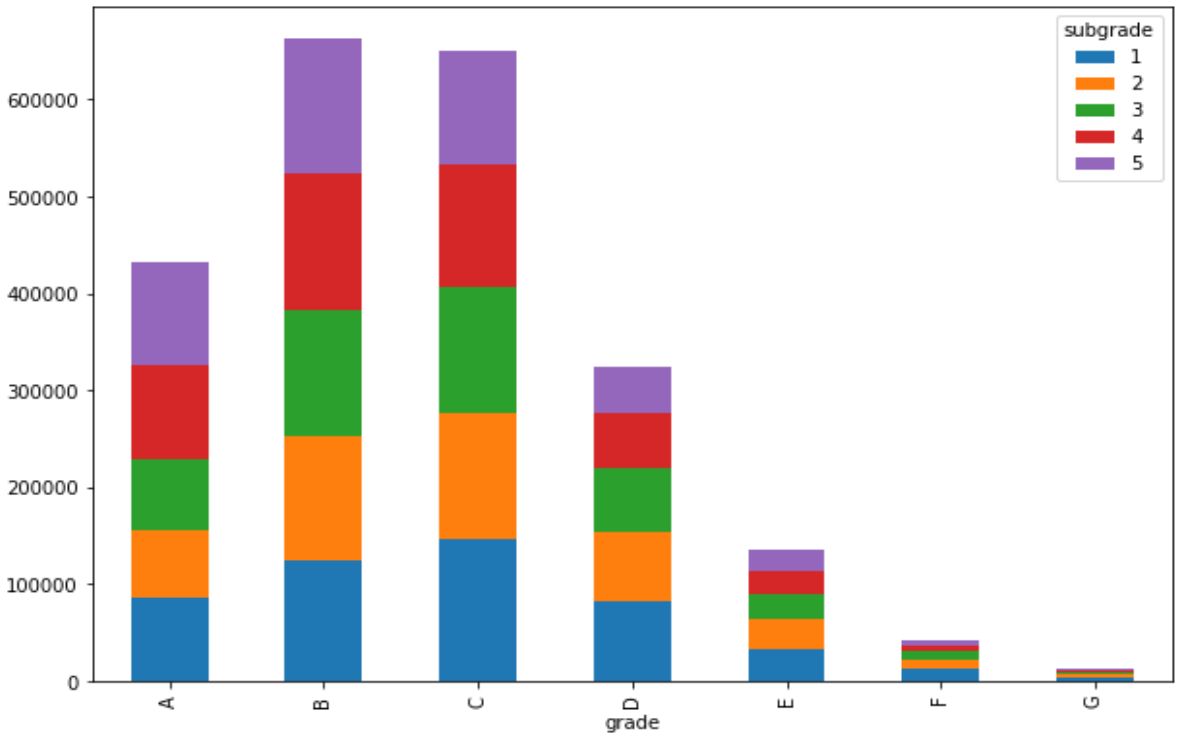

Figure 5 : Distribution of loans by grade and subgrade.

2. Investigation of significant subsets of borrowers.

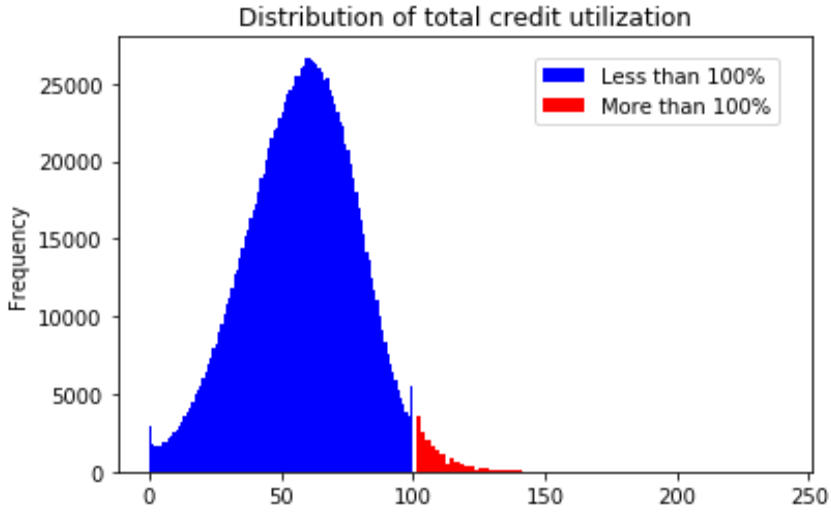

Figure 6 : Total credit utilization distribution, values over 100% represent borrowers past their credit limits.

For the percentage of utilized credit it can be seen that there is a small portion of the dataset which exceeds the theoretical one-hundred percent threshold. The reason for these values is unknown but it made me ask the question, can more targeted models be made based on different subsets of the borrower population? To investigate, I first looked towards the difference between those above and below 100% credit utilization. For example, the loan amounts and average current balances for each group were:

The average **balance** for borrowers utilizing **more than 100%** of their credit: $182575.92
The average **balance** for borrowers utilizing **less than 100%** of their credit: $143809.02

The average **loan amount** for borrowers utilizing **more than 100%** of their credit: $11729.86
The average **loan amount** for borrowers utilizing **less than 100%** of their credit: $15274.88

The average current balance is actually greater for people utilizing less of their credit, This is not indicative of not paying off loans; in fact, people that use less credit on average have loans with higher principal amounts. To get a sense as to whether the separation by credit utilization is useful for partitioning the set of borrowers, we can look at the differences in summary statistics between the two populations. The quantity that stood out to me is the total balance excluding mortgage. The averages for each sample population were computed to be approximately 40000 dollars; but it seems that this value nearly doubles when excluding

mortgage balances. It could be that perhaps there is a difference in home ownership status between the two groups.

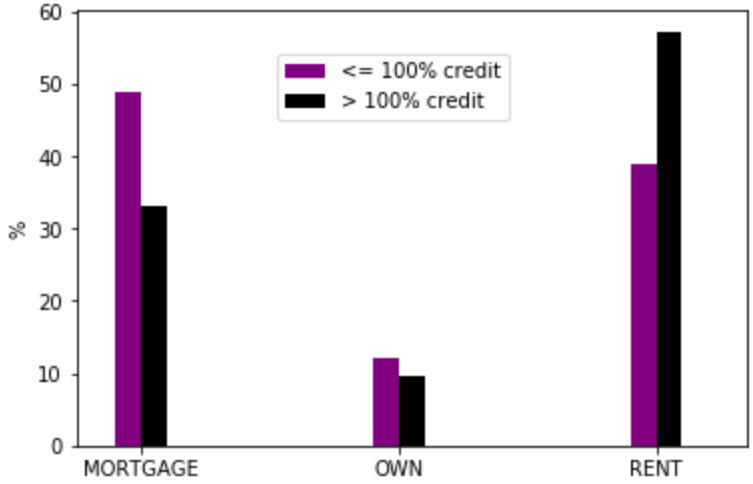

Figure 7 : Home ownership status broken up by credit utilization percentage.

While this bar plot does not prove anything it does seem to imply that there is a difference in housing status between the two groups. It seems that a majority of the borrowers over their credit limits are renters while the other borrowers are more likely to either own or mortgage a home. The question I wanted to answer was this: can home ownership status be used as a distinguishing factor in any other manner? To explore this line of thinking, let's look at the total current balances of all borrowers.

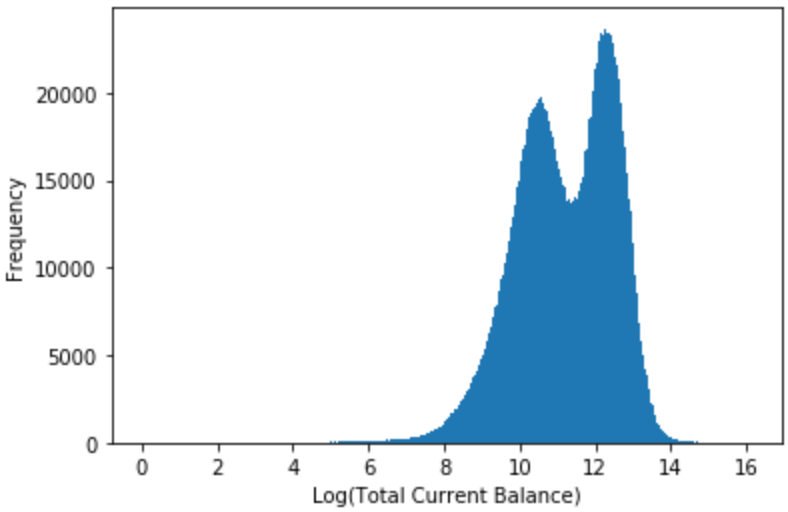

Figure 8: Distribution of the logarithm of total current balances. Note the bi-modal structure.

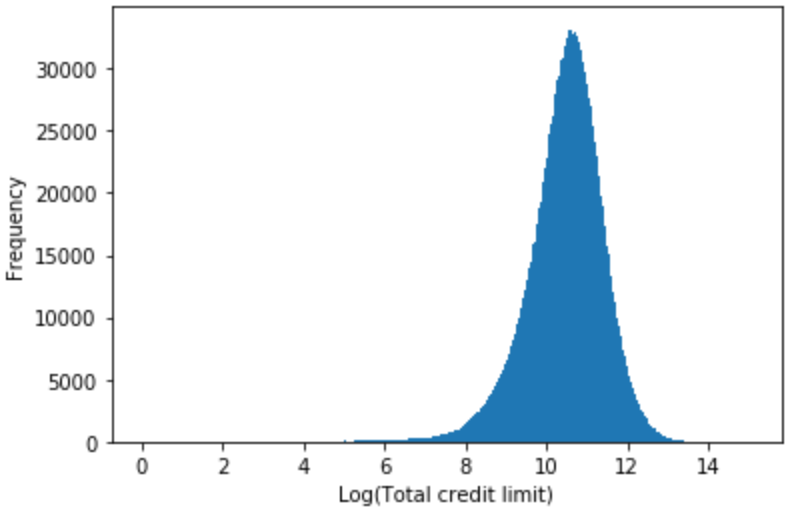

Figure 9 : Distribution of the logarithm of total current balance, excluding mortgages.

I would have guessed that the total current balance of all accounts would have a normal distribution. Plotting the histogram, however, shows another story as can be seen in figure 8. My goal is to explore where these two peaks are originating from I then plot
the total balance excluding mortgages.

The effect of home ownership status seems immediate as the two distributions are entirely different. To investigate the reason for the bimodal distribution I split the total current balance data in half (via the median value) and then create a bar plot for the home ownership status for each half. The stratification is much more drastic than was the case with credit utilization. distribution. For an example on how to interpret: mortgages account for nearly 80% of the accounts above the median total current balance, as indicated by the green bar reaching a value ~80%.

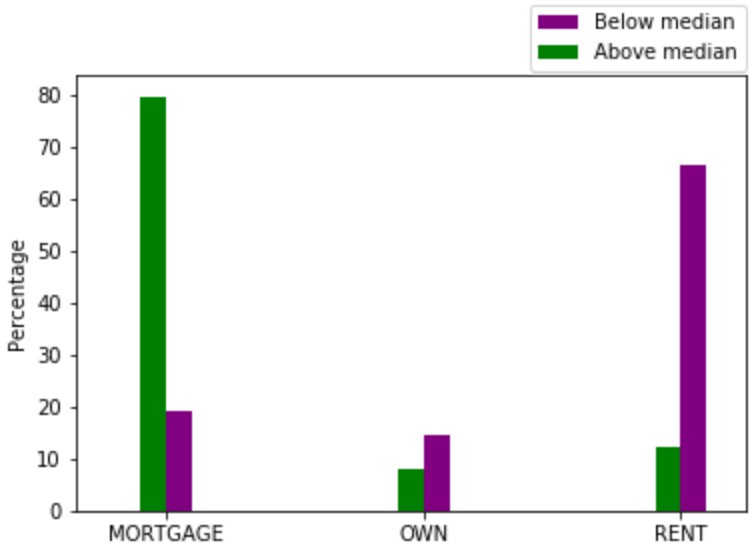

Figure 10: Percentage of each home ownership category for each half of the total

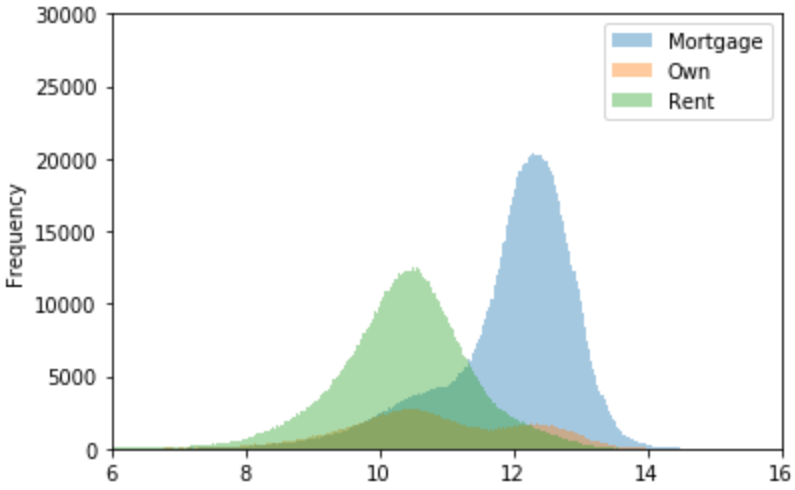

Figure 11 : Total current balance distribution stratified by home ownership category.

This is visualized in figure 11 by plotting the distributions of (log) total current balance with respect to each category. Own and mortgage and perhaps the distribution for those who have mortgages still look like they're bimodal, but the total current balance distribution for renters seems to be unimodal. This almost seems like it's hinting at the possibility that there are hidden populations of customers. The idea I have in my head at least is that perhaps there is a useful way of subdividing or partitioning the data such that these subsets could each be treated as their own population; that is to say, each would have a separate model instead of there being

single models for loan outcome prediction and recovered capital regression. The problem is that the next step forward, breaking down the mortgage and own categories is not obvious, even after repeating the same steps. So, in summary, it might be useful to further investigate or subdivide the borrowers into distinct populations, but up until now I have not found an obvious means of doing so.

3. Exploration of loan outcomes

4.

To continue, the main goal of this project is to produce a model which accurately predicts the outcome of a loan. A bad outcome is defined as when either a borrower is late on payments or the borrower has charged off the loan. A good outcome is when the loan is paid in full. This "good" and "bad" dichotomy is an artificial construction resulting from aggregation by loan status. This frames our problem as a binary classification problem which can be modeled by logistic regression and random forest classification. With these methods and other considerations, a model for loan status prediction can be formulated.

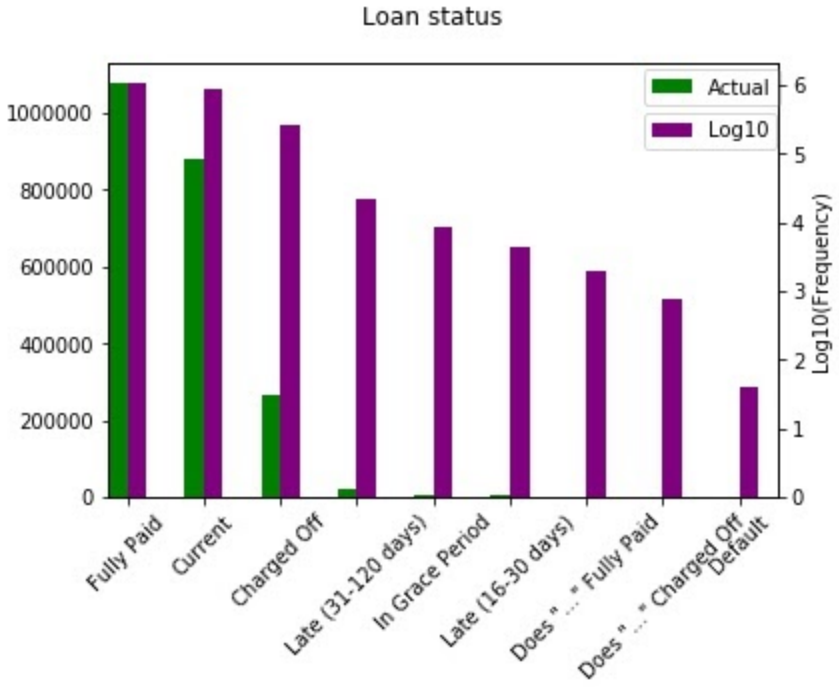

Figure 12 : Loan status distribution (and logarithm for scale purposes)

As we can see by this figure 12, the vast majority of loans fall into three categories: "Fully paid", "Current", and "Charged Off". Because of the time dependence of the problem, the quantity being modeled is modified to respect this distribution. The modification is to predict the outcome of loans by their maturity date. Therefore we can prune the loans which have not matured yet. In addition, to make the problem a relatively balanced, binary classification problem the only categories that are retained are "Fully paid" and "Charged Off". The binary distribution plot shows the proportion of loans with bad outcomes.

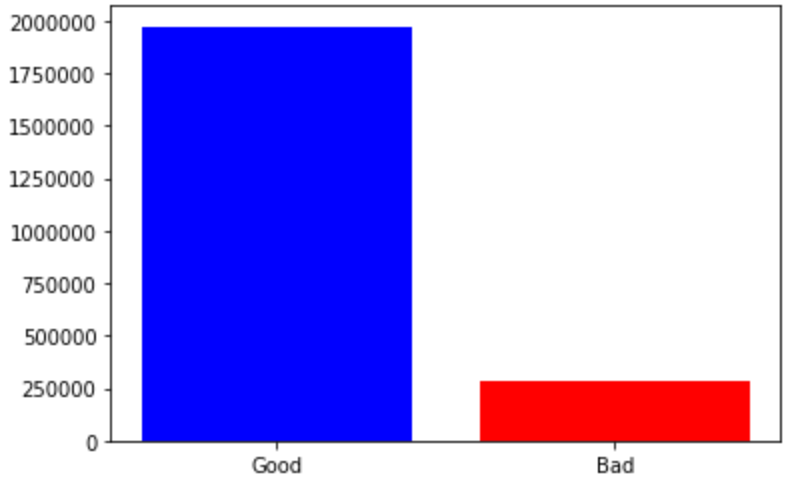

Figure 13: Loan statuses binned as either “good” or “bad”.

5. Exploration of loan recoveries

What can be done with loans of bad standing? Instead of the money being lost to the ether, financial institutions will naturally attempt to recover capital of charged off loans. This amount of capital recovered can be modeled as a continuous variable via regression techniques such as linear regression and many others. This variable is important because it is closely associated with loans of bad status, and provides a recommendation or course of action for loans that have become charged off or delinquent. Unsurprisingly the recovery amount is correlated to loan status but this is a misleading quantity as loans of good status do not need recovery to begin with. Because of this, I postulate that it may be wise to completely filter out loans of good status before performing any modeling. The main concern is the order of these time dependent variables in regards to training and testing data sets for the models as well as for cross validation. If future data is accidentally used in the training then the predictions are essentially worthless. This goes down all the way to the level of normalization; that is, if normalizing the data then one should be sure to only normalize with respect to the training set.

How should these be accounted for in the modeling process? For the loan classification process, all time dependent quantities need to be removed from the model training process for

the loan classification problem, as the model is attempting to predict a decision made at a certain point in time. The model for the recovery amounts is agnostic of time as it merely wants to predict the amount that can be recovered going forward; it is not a time sensitive decision; therefore, no considerations need to be made for the second stage of the capital recovery model.

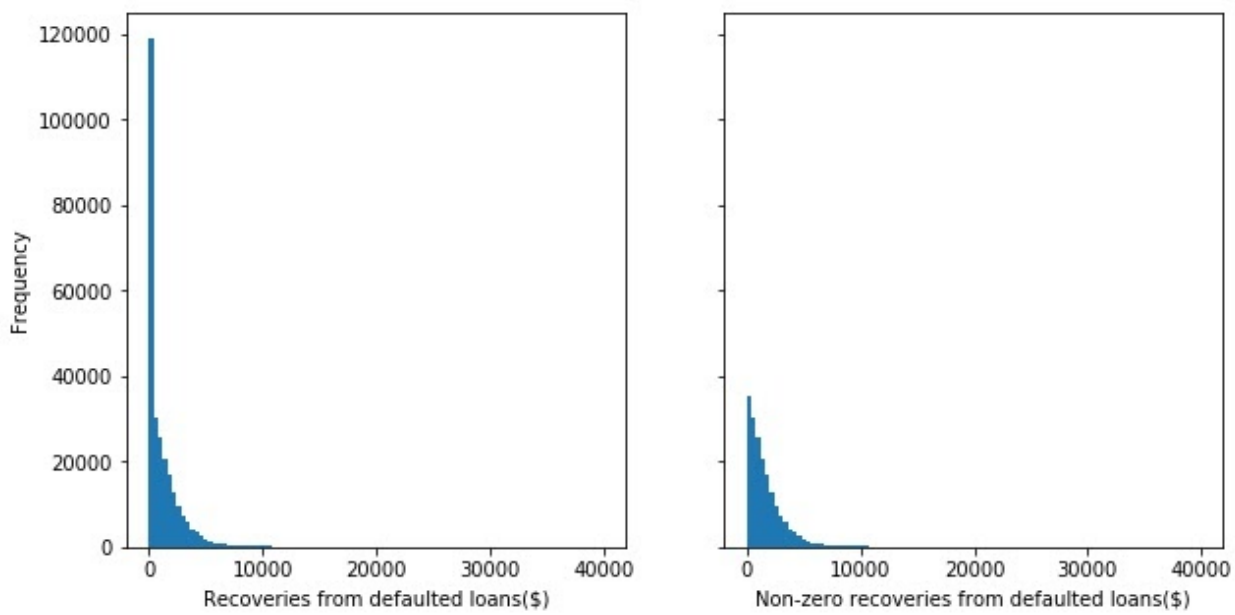

Figure 14: Distribution of recoveries from defaulted loans, with and without zero recoveries.

First, look at the distribution of the would-be dependent variable. It seems that there are still a large number of charged-off loans that have had no money recovered from them; part of the motivation for this stage of the project. Nearly a third of all defaulted loans have had zero capital recovered. In the modeling stage of this project, I explore what happens when both including and excluding these 0 value from the models.