# Mortgage Group
# Team 4

Tong Chen (chent17@rpi.edu) ,Donghuai Li (lid20@rpi.edu), Han Zhang (zhangh28@rpi.edu)

## Project goal

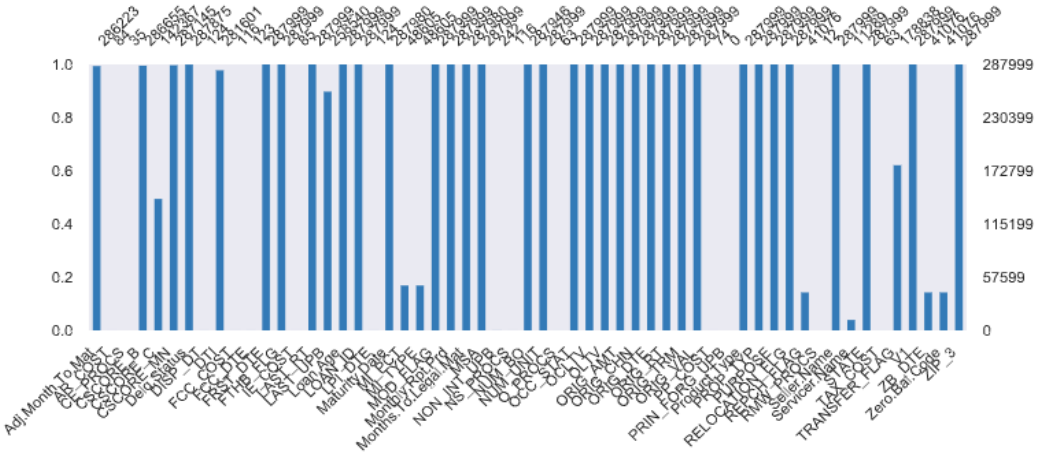
For project 2, we concentrated on using unsupervised learning to get a deeper insight into the Mortgage dataset. Besides, we also further some unclear points left in project 1, to make our project more comprehensive. In the beginning, we first identify groups and profiles of single-family loan borrowers who share similar characteristics and recognize group patterns based on the mortgage data from 2000 to 2017. Since the results given by unsupervised learning are not defined in advance, we would then analyze the economic significance of specific mortgage groups characteristics, to dig out the economic meaning behind the clusters. With significant customer groups being segmented, we will furtherly generate models to predict mortgage default rate based on these groups. Targeting at specific characteristics, we are expected to improve our model with better performance.

## Data introduction

To support its credit risk sharing program, Fannie Mae has released an extensive data set since 2013, which provides an in-depth understanding of the credit performance of a part of Fannie Mae's single-family book of business.
The data set now contains nearly 22 million records, providing detailed information on monthly loan levels, and is designed to help investors better understand the credit performance of some single-family loans owned or guaranteed by Fannie Mae.

The attributes of a loan divided into two groups: Acquisition file and Performance file. We combined two files together based on loan id and then checked variables.

The mortgage dataset consists of 57 variables in total, including 33 numeric variables, 19 categorical variables, 4 Bool variables.  First, we check the missing value of each variable. The count of the missing value of each variable are shown below:




From the graph we can see many variables have ‘NA’s. We cleaned this data before we use unsupervised models.

## Data processing

First, we selected 1 row of records from each loan. And because some performance columns are only populated at the end of the loan, we chose the latest record and analyzed the special variables. For example, the ‘Zero balance code’ indicates the reason the mortgage loan's balance was reduced to zero. Besides we made dummy variables from the different codes, we also found that the Zero balance code has a relationship with the delinquency status. This helps us define the default loan more precisely.

Second, we continually tried to extract information from text variables: ‘Sellers’ name’ and ‘Servicers’ name’. The original text is hard to use in our model, the content is sparse and has errors. There are some similar seller names that should be treated as the same seller, so we choose the first 7 characters as a new seller indicator. Then based on different company's frequency of occurrence, we assigned four groups to all of the data: ‘Big company’, ‘Medium company’, ‘Small company’, ‘Other’. There is a small proportion of data that has value on ‘Servicers’ name’, we made dummy variables on it to indicate whether the loan has a servicer.

Then, we processed variables that contain high percentages of missing value. Instead of dropping them, we chose to fill the missing value or generate data-existing indicator variables. For example, the flag variable ‘MI_PCT’ which means mortgage insurance percentage. The data of this variable ranged from 2% to 50%. Those loans have ‘NA’ in this variable we fill 0 into it as a reasonable value. We decide to only extract year and month from date information now and drop the following variables: 'FRST_DTE',

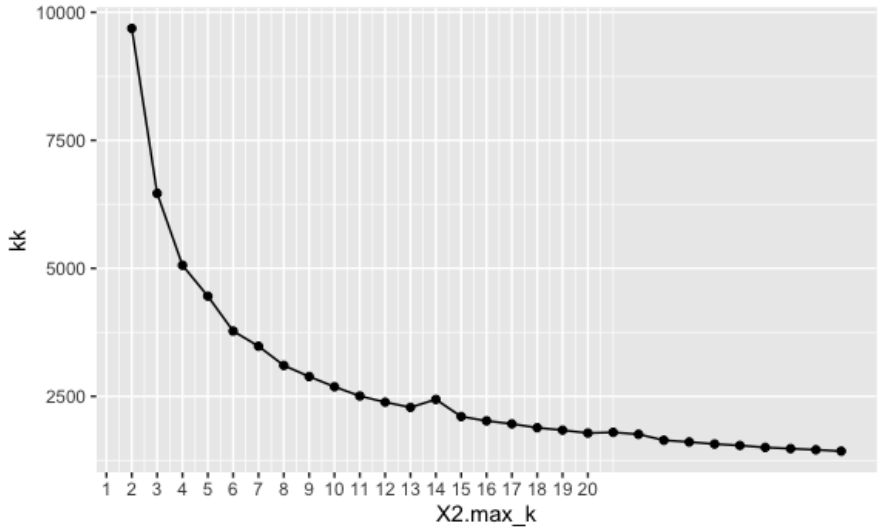
'Monthly.Rpt.Prd', 'Maturity.Date', 'ORIG_DTE'. We also drop some meaningless variables:'Unnamed: 0', 'V1', 'Product.Type' (have the same data in all rows).

Finally, we generate default indicators from the ‘Delinquency status’, we chose those loans which continually have delinquent status for over 6 month as default loans.

## K-mean Clustering

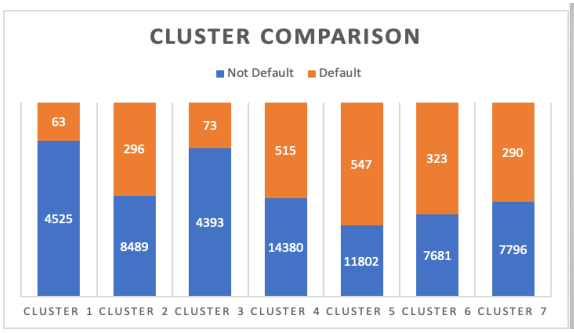
We use the K-mean clustering to identify similar mortgage loan borrowers in the subgroups. The algorithm behind the K-mean clustering includes four steps. First, it will randomly select the centers of the clusters. Second, the distance between each point and the center point will be calculated. Third, the data point will be assigned to the cluster which has the smallest distance. Last, the new cluster center point will be recalculated.

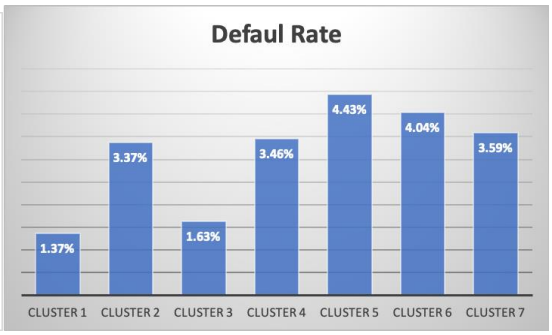
Before creating the clustering model, we found the optimal K values, that is optimal numbers of the clusters, through the elbow method. The idea of the elbow method is to run the clustering model through the assigned K values, which is 20 in our work, and calculate the sum squared error for each K value. The elbow method graph of our mortgage data is in the below. The K =7 is the optimal number of clusters because it captured 85.2% variances for the clustering model. The bend in the graph shows extra clusters beyond the 7 add little values for our model.




We added a new column which can indicate the individual group number and then subsetted seven data frames by their group numbers. We were interested about the mortgage defaults within each group. Therefore, we build two histograms to reveal the

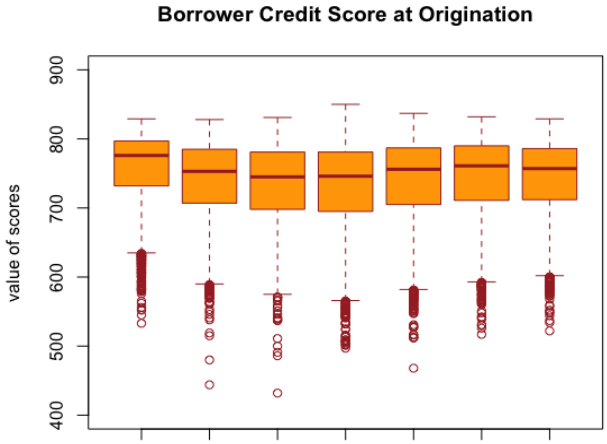
default counts and not default counts, and the default rate for each cluster. Here are the histogram graphs. We found that cluster 1 has the smallest number of defaults and the lowest default rate, but cluster 5 has the largest number of defaults and highest default rate.



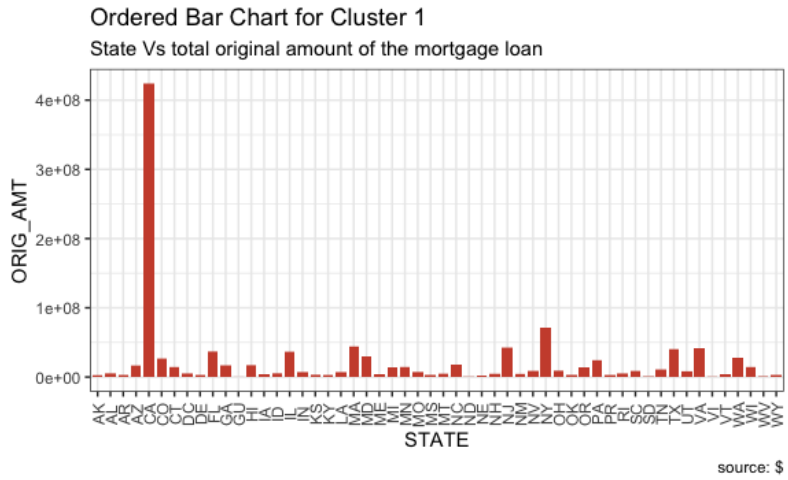

## Clustering Analysis

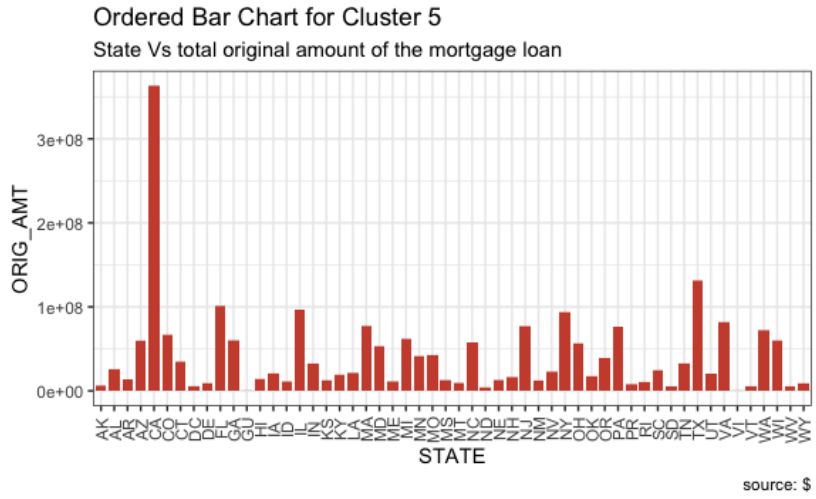
Then we focused our analysis on generating insights about what factors caused or associated with the highest or lowest default rates for cluster 5 and cluster 1. Here is a box plot which shows the borrowers credit scores at mortgage origination. We can see that the borrowers in cluster 1 have the highest credit scores in terms of Q1 to Q3 quarter values. Therefore, the good credit performance may explain why people in the cluster 1 generally have low default rates.



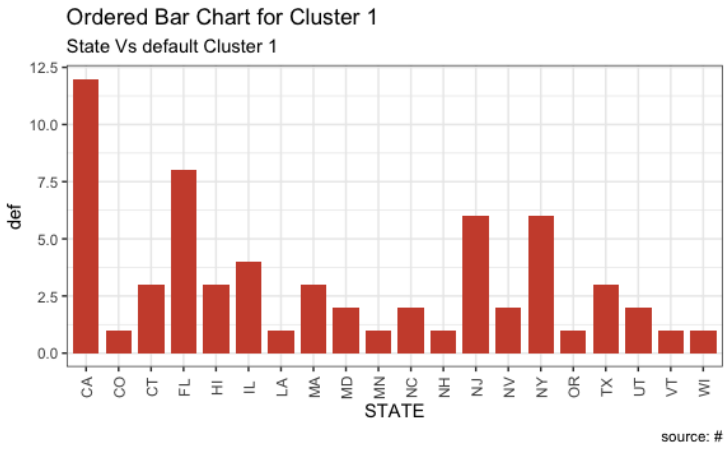

Borrower Credit Score from Cluster 1 to Cluster 7

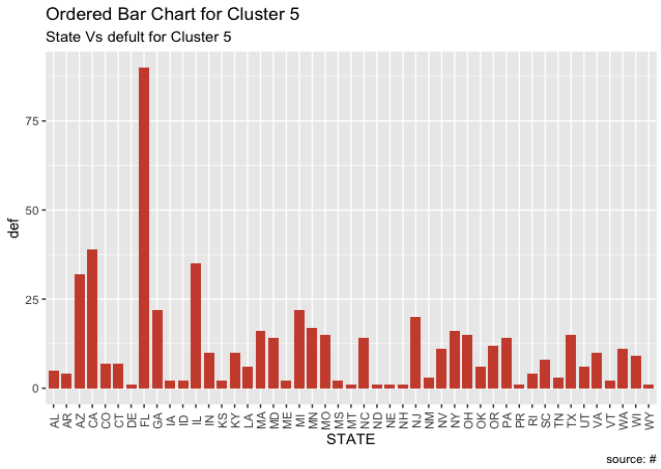
Besides, we also analyzed the total mortgage loan amounts distributions for each state in the US. Here are the histograms of cluster 1 and 5. The California state has the largest amount of mortgage loan for both cluster 1 and 5. But Texas and Florida clearly have more loans in cluster 5 than 1.



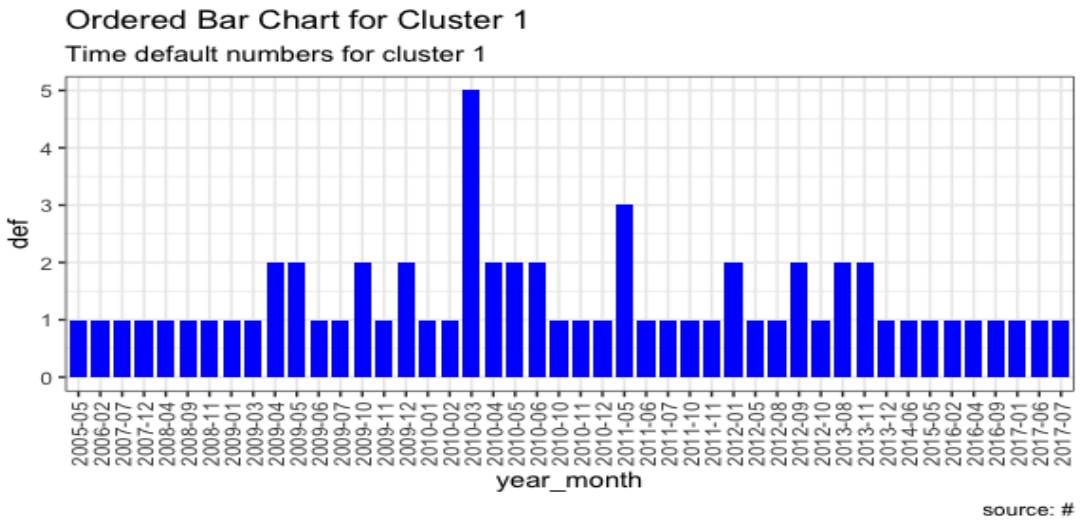



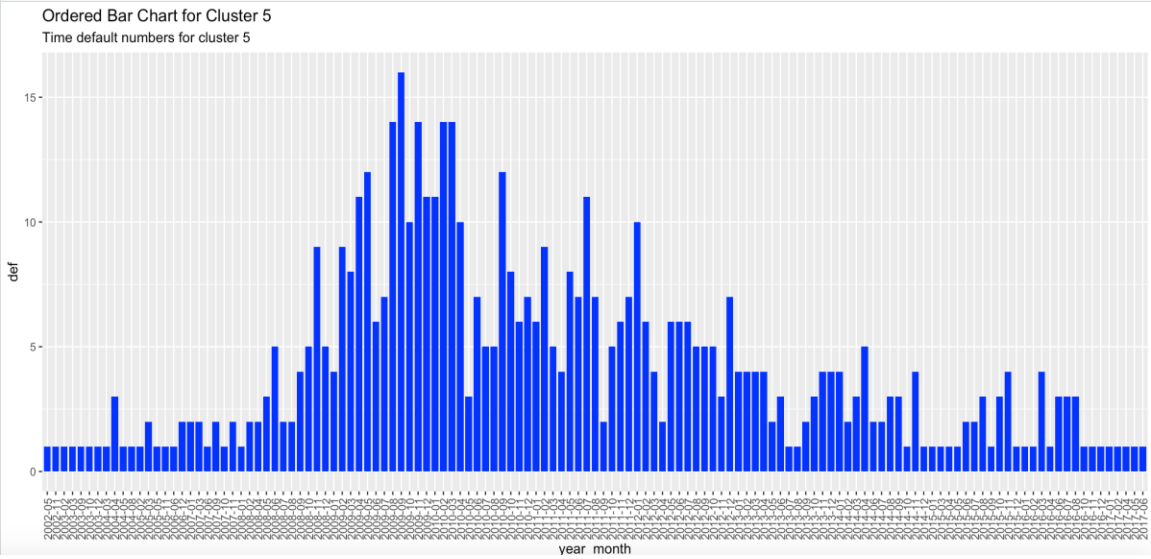

The mortgage default numbers of each state are presented below. While the California state has the highest mortgage default counts in the cluster 1, the Florida state has remarkably high default counts in the cluster 5.



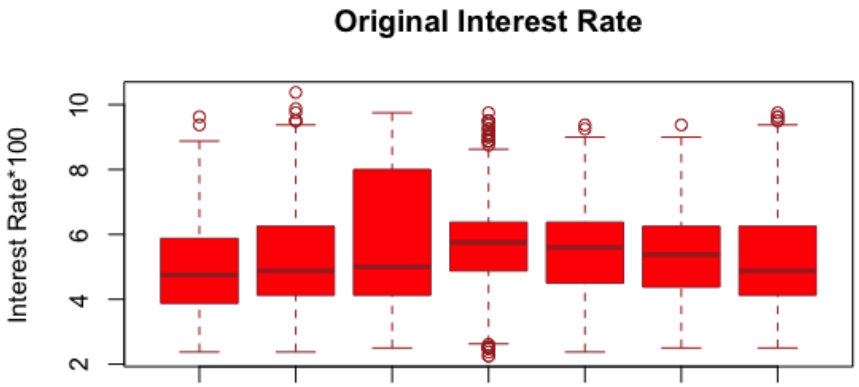



Next, we analyzed the default counts of cluster 1 and 5 from 2000 to 2017. The histograms are below. Cluster 1's default counts span from May 2005 to July 2017. The maximum number occurred in March 2010, but most of the months have only 0 or 1 default cases. However, cluster 5 has more default cases between 2008 and 2012. The peak is in September 2009. The distribution of mortgage default counts in the cluster 5 is matching with the Credit Crisis and Economic Crisis started from 2007 and 2008 respectively.






The original interest rate of cluster 1 and 7 is the last graph. We can see that the cluster 1 on the most left has the lowest Q1 - Q3 values, which might explain why the borrowers in cluster 1 have the lowest default rate. Cluster 3 has the widest interquartile range.Cluster 4 has most outliers and the most narrow interquartile range. The median is also the highest in the cluster 4.



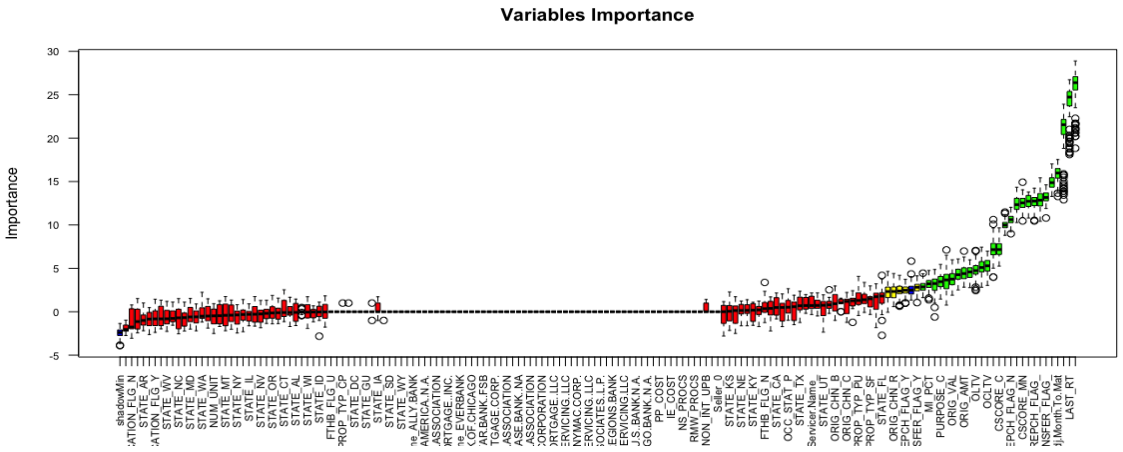

Original Interest Rate from Cluster 1 to Cluster 7

In general, we can conclude that the characteristics of borrowers in cluster 1 include low original mortgage interest rate, high credit scores, and low default rates during the Credit Crisis and Economic Crisis. Borrowers in cluster 5 have relatively low credit scores, relatively high-interest rates, and high default rates between 2008 and 2012. Also, many mortgage default borrowers in cluster 5 came from Florida State.

## Feature selection

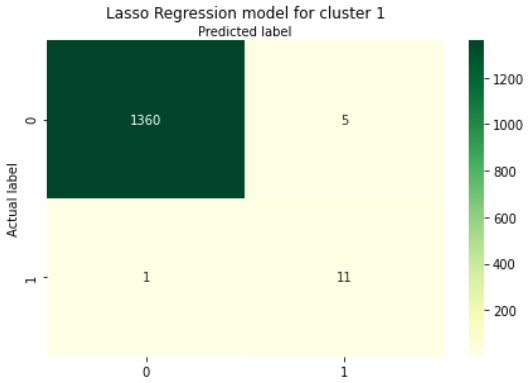
Before building other models, we used the Boruta Algorithm to help us reduce our data dimensionalities by selecting important features related to the outcome variable. The Boruta Algorithm is built around the random forest classification. It firstly copied our data and then shuffled the data into shadow values. Then, it trained our model through Random Forest Classifier. Last, the algorithm will check the Z scores of each feature and compare them with the best shadow features' Z scores. If the feature's Z score is larger, the algorithm will classify it as important features. The outcome plot and mean importance table are in the below. We can see that the top important features include: Current Interest Rate (LAST_RT), Adjusted Months To Maturity(Adj.Month.To.Mat), Servicing Activity Indicator (TRANSFER_FLAG), Repurchase Make Whole Proceeds Flag(REPCH_FLAG), Co-Borrower Credit Score at Origination(CSCORE_C), Original Loan-to- Value (LTV), ORIGINAL HOME VALUE (ORIG_VAL), The original amount of the mortgage load(ORIG_AMT), and Cash-out Loan Purpose (PURPOSE_C). We will use those features to create our models for cluster 1 and cluster 5.

| | Mean Importance | Decision |
|---|---|---|
| LAST_RT | 25.75262 | Confirmed |
| Adj.Month.To.Mat | 24.09637 | Confirmed |
| TRANSFER_FLAG | 20.61492 | Confirmed |
| REPCH_FLAG | 15.85642 | Confirmed |
| CSCORE_C | 14.86933 | Confirmed |
| OLTV | 13.17333 | Confirmed |

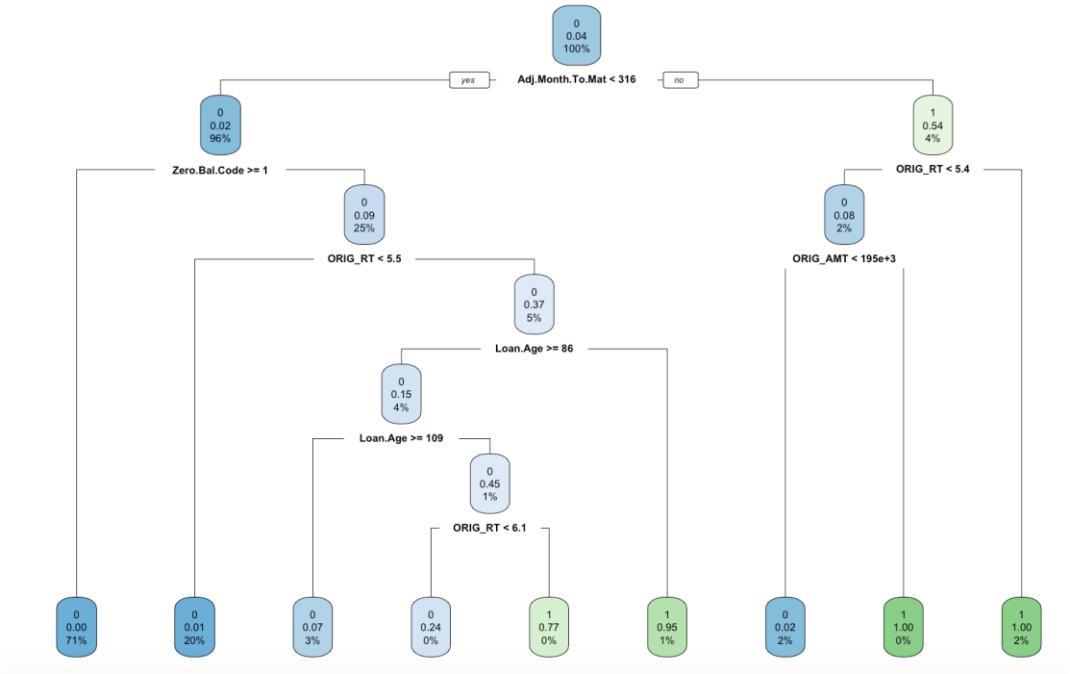



## Lasso Regression and Decision Tree Models based on Clustering

We then subsetted the cluster 1 data and created a Lasso Regression model to find the precision rate. The Precision rate is defined by using the number of true positive to divide by the sum number of true positive and false positive. Before creating the model, we splitted the data by 70% of training data and 30% of testing data. Then we calculated the best lambda through the Area Under Curve. After that, we used the best lambda to create a lasso regression model. The confusion matrix below showed our model prediction outcomes. Then prediction accuracy of testing data is 99.56%. The Precision rate is 68.75%.




For the cluster 5, we built a Decision Tree model to find the prediction accuracy and precision rate. The data was also splitted  by 70% of training and 30% of testing. The resulting tree graph and the confusion matrix table are in the below. By looking at the map, we can see the borrowers in the cluster 5 tended to be default when loan age was more than 86 months, original interest rate was more than 6.1%, and the original loan amount was larger than the threshold.




The prediction accuracy of the decision tree model is 99.13. The precision rate is 94.55%. Therefore, our model has a strong capacity to label the borrowers with default payment in the cluster5.


## Future and Implementation

Based on our work, we propose several fields to refine our project:

1. Time series analysis
   We chose the latest record of each field to form the data set. But we can do more on it. The relationship between loan age and its start year or month is not explored yet.

2. Check the trade-off of bias/variance in each cluster's model
   In the selection of best k-means, we choose an empirical k. However, we can calculate the different k leads to different biases and variance in all the clusters and choose the best k.