Predicting Loan Status - Exploratory Data Analysis

Rita Miller

04/03/2024

## Load Packages

library(tidyverse)  
library(e1071)

## Load Data

dataset <- read.csv("loans50k.csv")

#Explore Data  
head(df) #view first 6 rows

##   
## 1 function (x, df1, df2, ncp, log = FALSE)   
## 2 {   
## 3 if (missing(ncp))   
## 4 .Call(C\_df, x, df1, df2, log)   
## 5 else .Call(C\_dnf, x, df1, df2, ncp, log)  
## 6 }

#summary(df) #summary of data set  
#dim(df) #dimensions 50k rows 32 cols  
#View(df)#quick view of the entire data set

## Introduction

The purpose of this study is to predict which customer is more likely to default on their banking loan. The dataframe has 50,000 observations and 32 variables with 7,581 loans in bad status and 2,7074 in good status. In this investigation, we will prepare, clean, explore, and transform the data. Variables related to good and bad loans will be included, while variables that are not useful as predictors will be excluded.

## Preparing and Cleaning Data

We will keep loans that were fully paid, charged off and default in our data.

## [1] "Charged Off" "Fully Paid" "Default"

## [1] 34655 32

## [1] "factor"

### Variable elimination

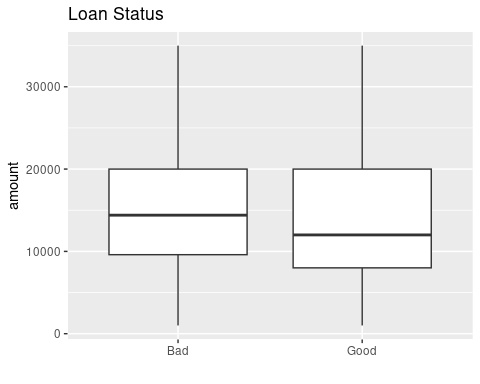
We will remove the variables LoanID and employment as they are irrelevant.

### Feature engineering and NA values

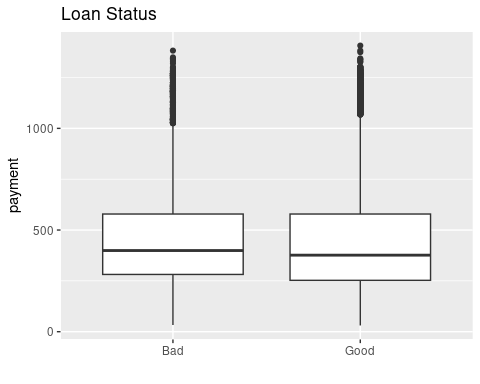
Looking at the dataset we see a well distributed dataset, so feature engineering is unnecessary.

## Exploring and Transforming the Data

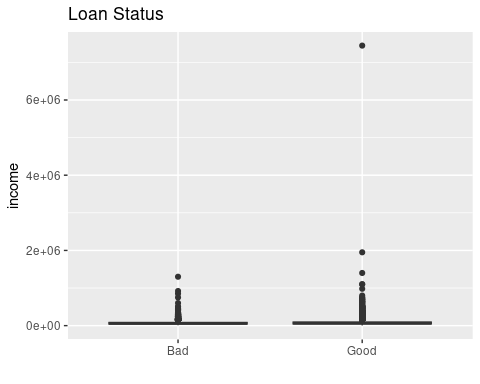
The following charts show the distributions of some of the quantitative predictor variables, to see if if the variables are distributed differently for good and bad loans.



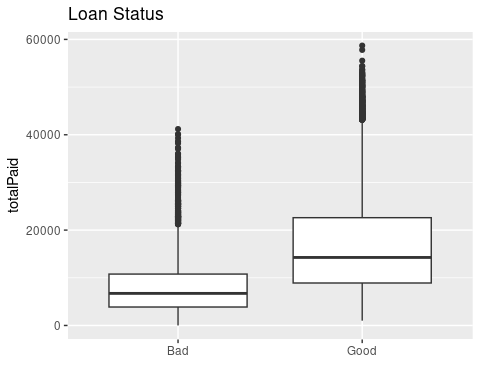
The distribution of good and bad loans are moderately skewed to the right with no apparent outliers. Will replace this predictor in the data frame with transformed values using logarithms.



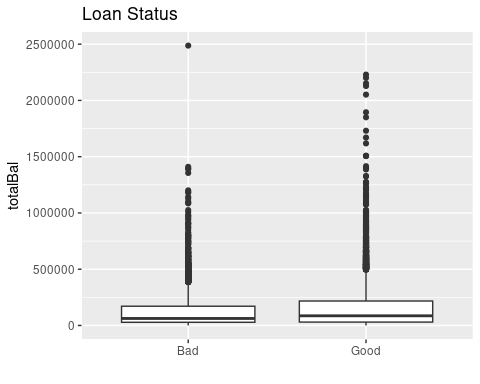
The distribution of good and bad loans are moderately skewed to the right and some outliers are apparent. Will replace this predictor in the data frame with transformed values using logarithms.



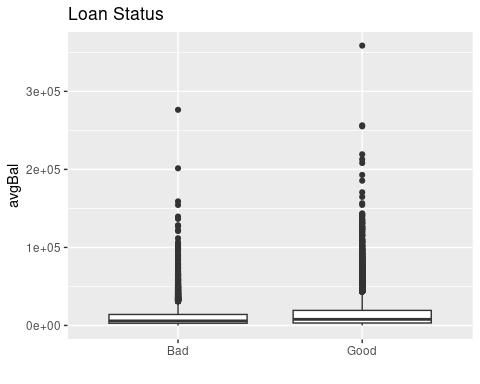
The distribution of good and bad loans are strongly skewed to the right with outliers. Will replace this predictor in the data frame with transformed values using logarithms.



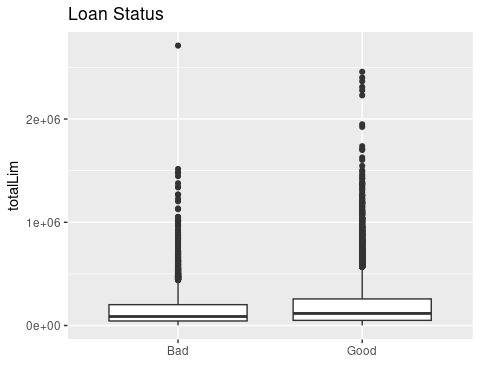
The distribution of good and bad loans are strongly skewed to the right with outliers. Will replace this predictor in the data frame with transformed values using logarithms.



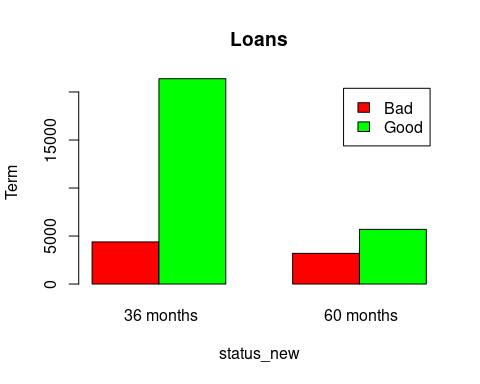
The distribution of good and bad loans are strongly skewed to the right with outliers. Will replace this predictor in the data frame with transformed values using logarithms.

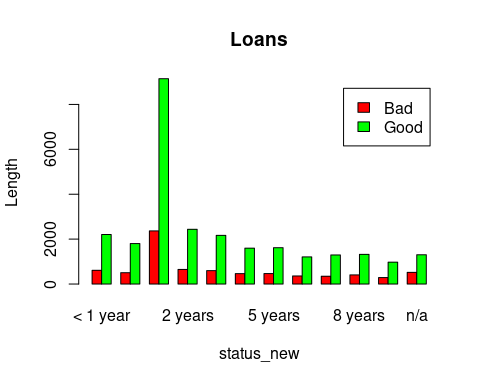
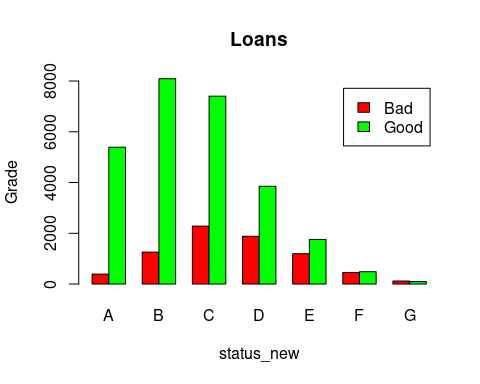


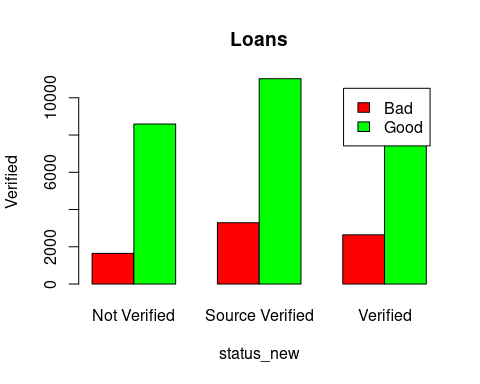
The distribution of good and bad loans are strongly skewed to the right with outliers. Will replace this predictor in the data frame with transformed values using logarithms.

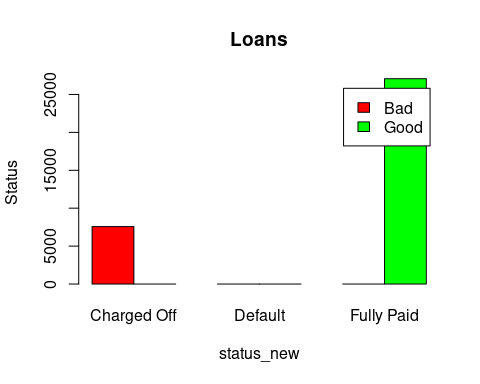


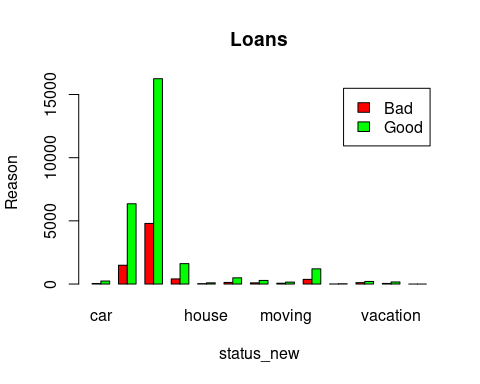
The distribution of good and bad loans are strongly skewed to the right with outliers. Will replace this predictor in the data frame with transformed values using logarithms.

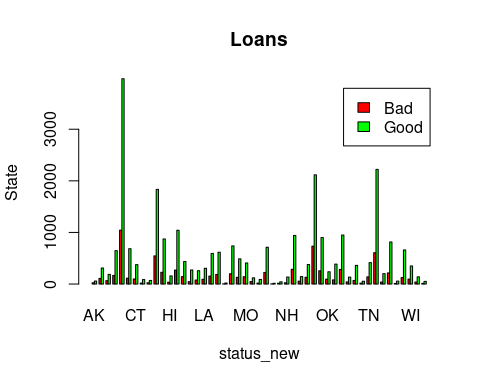
 The distribution of the term for most loans were in good status at 36 months.

 The distribution of the length of loans were approximately 2 years and most were in good status.  The distribution of home loans were mostly for mortgages and most were in good status.  The distribution of the grade for most loans was a B and most were in good status.

 The distribution of verified for most loans was Source Verified and most were in good status.

 The distribution of the status of loans were fully paid and in good status.

 The distribution of the reason for loans appears to be mostly for home improvement and most were in good status.

 The largest distribution of loans by state were in Arizona and most were in good standing.