Loan Analysis - Part 1

Matt Rock

3/8/2020

Section 2: Introduction

In the aftermath of the 2008 recession, the question everyone asked the banking industry was “How did you miss this financial collapse? You had all the information, but ignored it.” The pressure to create more mortgage-backed tranches meant housing lenders were actually competing to lend to individuals with riskier credit, leading to a recession that still continues to have repercussions.

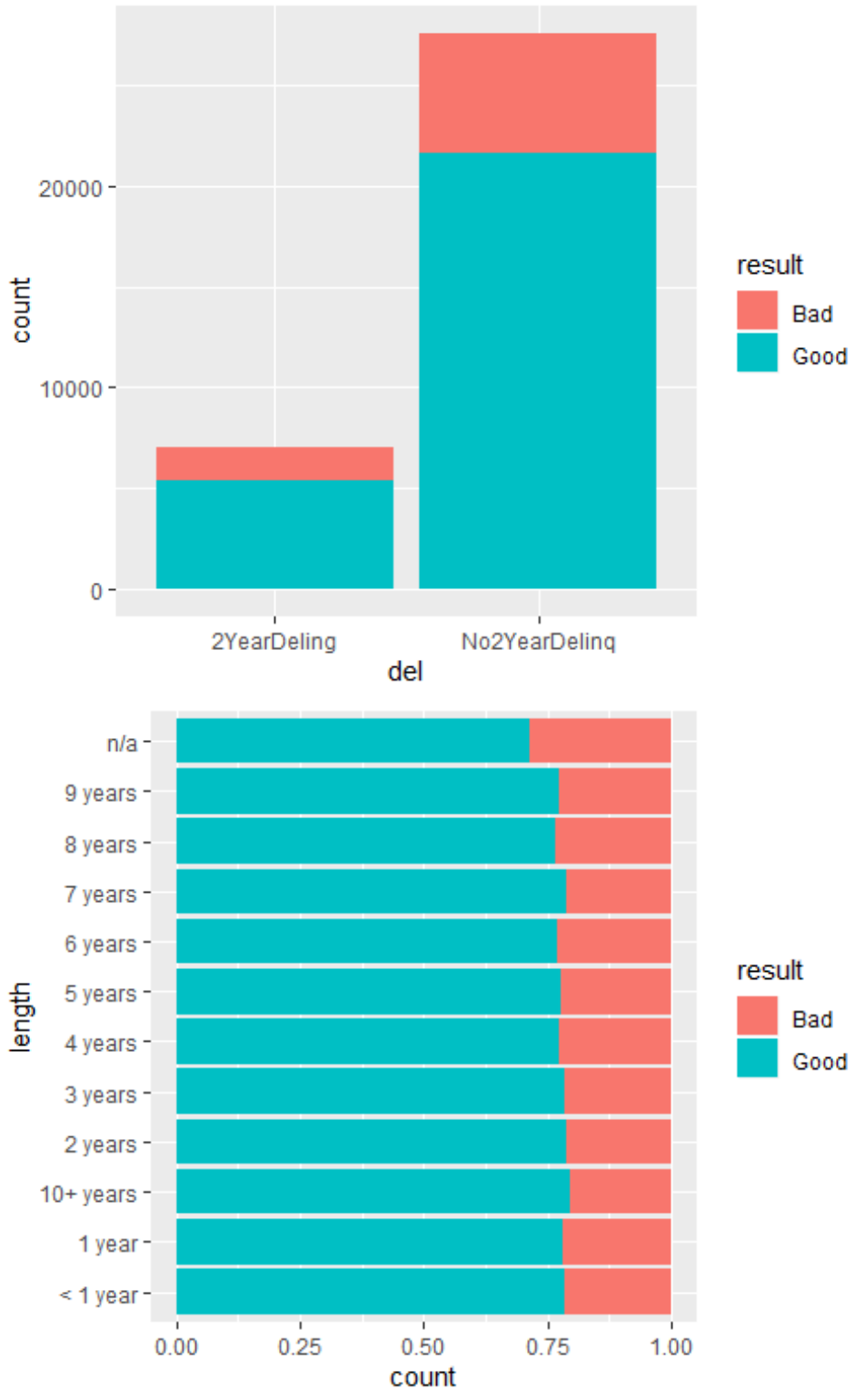
Now, with the spectre of financial ruin still lingering, banks need to protect themselves from bad risks of all kinds. The purpose of this project is to assist banks in being able to protect their investment portfolio with safe investments.

```
## Warning: package 'GGally' was built under R version 3.6.3
```

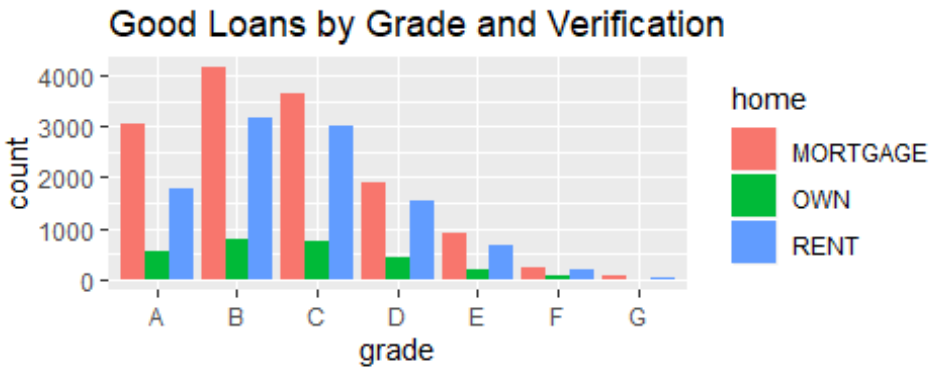
Section 3: Preparing and Cleaning the Data

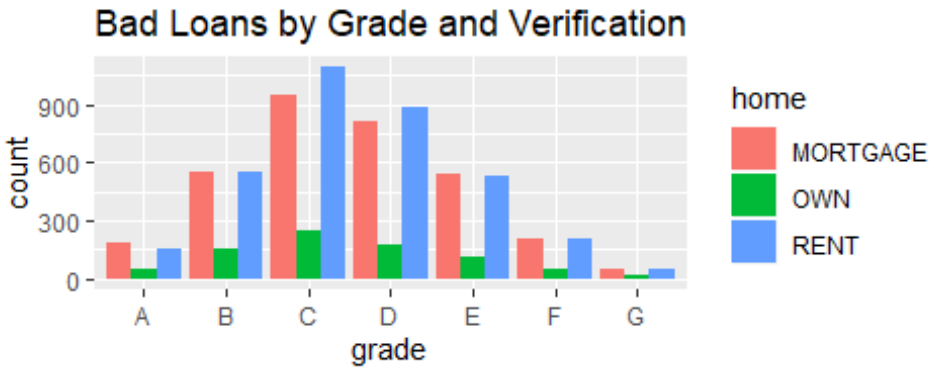
The initial data set had 50,000 observations and 32 variables. Selecting only on the relevant loan status left 34,655 data points. The initial variable trimming removed loanID (randomized number), employment (covered by job length and salary), and state (with 50 states, random chance would lead to false positives), as well as anything learned after the loan, like totalPaid.

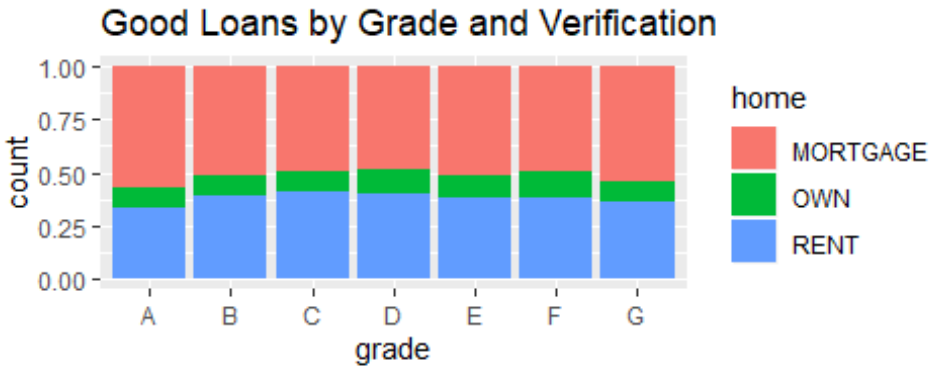
Some variables involved loan approvals more than loan results, and were removed. The majority of loans were given to people with no late payments in the last 2 years, but a loan was more likely to default if it was issued to someone without a late payment than at least one. Length of employment, public record hits, and 6 month inquiries also had no real clear bearing on loan quality

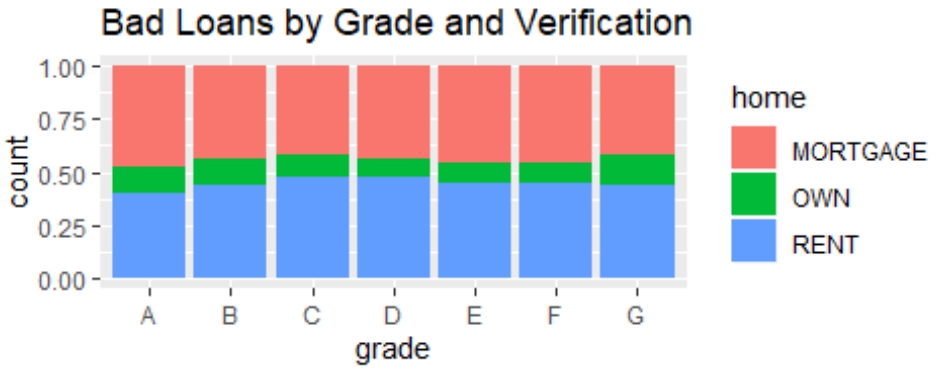


On a grade-by-grade basis, loans to people with a mortgage were more successful than those to renters.









T-tests showed total accounts, total revenue balance and total installment account limits to have possible average values too similar to be much help in predicting good or bad loans.

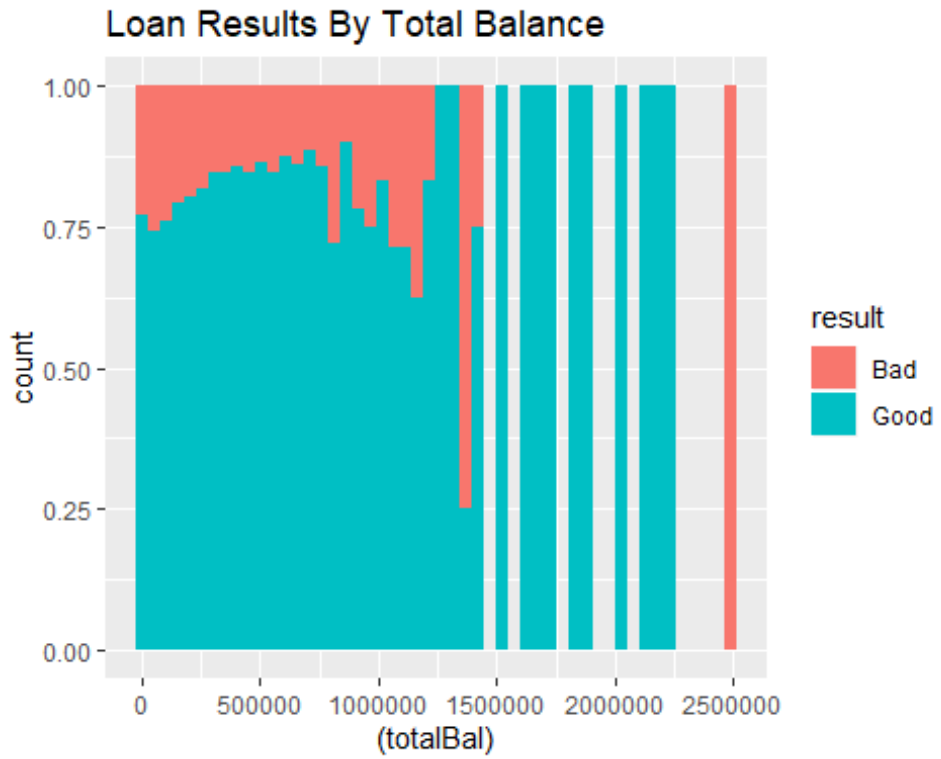
```
## [1] 0.1917847
```

[1] 0.6039935

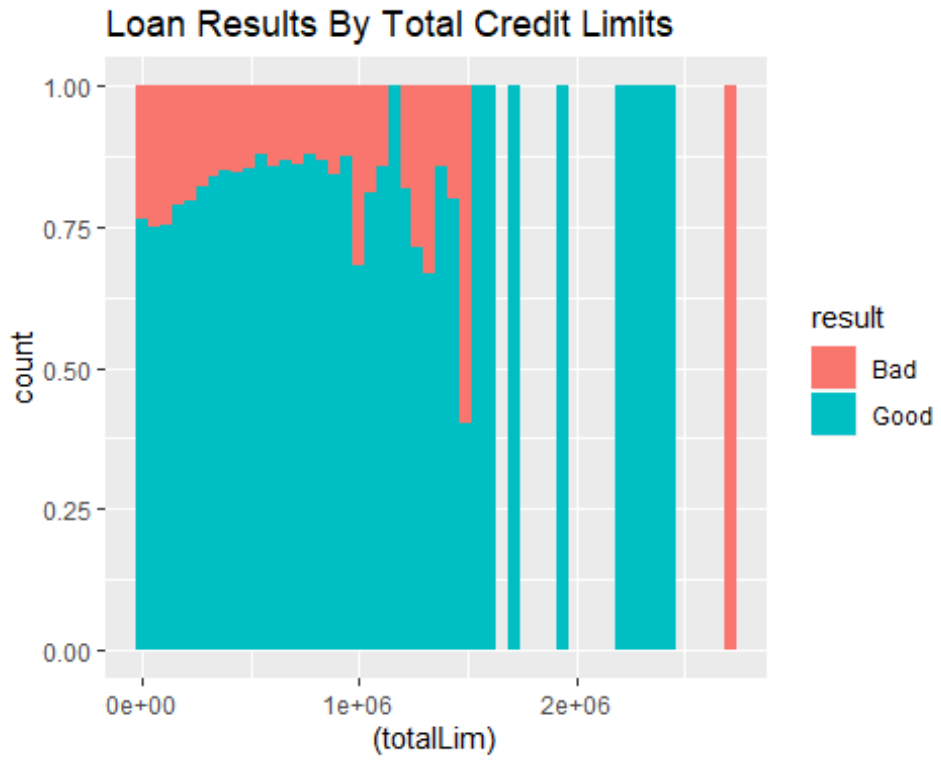
[1] 0.7386903

Looking at credit limits and balance, the difference showed a greater degree of loan success or failure than either one individually, leading to a replacement variable of totalLim - totalBal = waterLevel.

Warning: Removed 20 rows containing missing values (geom_bar).



Warning: Removed 24 rows containing missing values (geom_bar).



Warning: Removed 46 rows containing missing values (geom_bar).