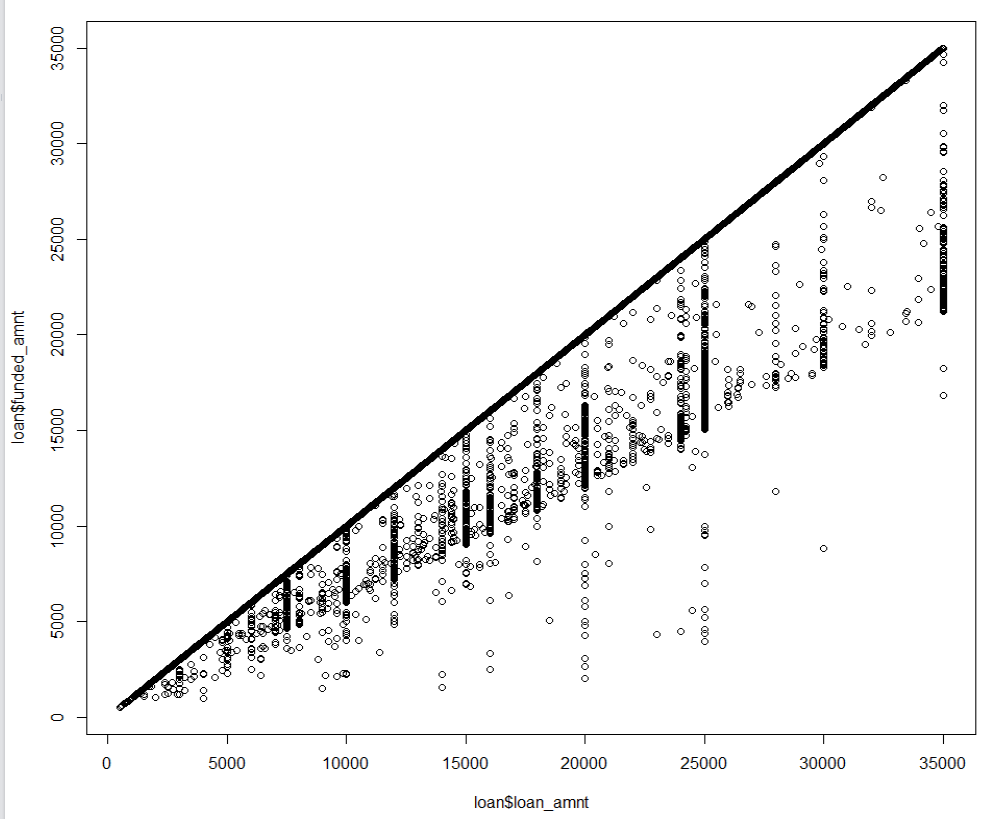
STAT 5291 Final Project Question Proposals

Daniel Wang

1. From the plot we can see that the loan amount that customers requested is almost a straight line with funded amount. It means that most of the customers got the funds finally. Yet, there are some customers did not get enough amount of funds as they requested and some customers were even rejected.

We would like to figure out what are the reason that cause those customers did not get enough funds, and the reason why they were denied.



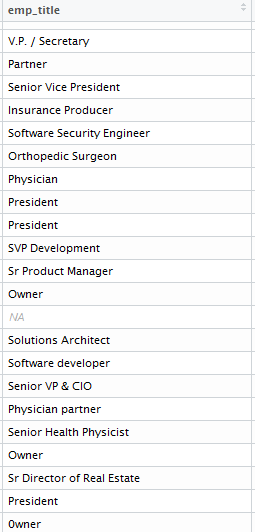
1. We want to predict the grade (ie. A, B, C, D, E) that the customers would receive as they applied for the funds, and the interest rate that corresponding with the grade. Briefly review the data, we could guess that the grade might be hugely influenced by several variables, such as loan\_amnt, emp\_title, emp\_length, home\_ownership, annual\_inc, and purpose, etc.

We want to figure out what are the relationship between those variables.

The histogram of grade. Obviously, not a normal distribution. Most of the customers obtained grade “B” and “C”.

1. Some of the variables are characters, such as emp\_title (VP, partner, analyst), purpose (credit card, home improvement, debt consolidation). If we want to figure out the relationship of these character variables with the loan final result, we need to quantify these variables.

An easiest way to do it is to simply put them into different level, such as partner is level 1, VP is level 2, analyst is level 3. However, since different company’s has different standard of title, this method may have huge bias.



1. Making a multiple linear regression to predict funded amount using int. rate, installment, grade. I think it is an easy way to briefly check their relationship. Further investigate is needed. We may also need to use more complicated model such as logistic regression, decision trees, etc.

