**IDS 572**

**Assignment 1**

**Lending Club**

**Harika Lakhinena**

**Roshan Dhanasiri**

**Sathwik Maddi**

**Part A**

**1. Describe the business model for online lending platforms like Lending Club. Consider the stakeholders and their roles, and what advantages Lending Club offers. What is the attraction for investors? How does the platform make money? (Not more than 1.5 pages, single-spaced, 11 pt font.**

*Business Model:*

Matchmaking Business Model

The business model used by Lending Club is a matchmaking business model; where you identify two or more customer groups and bring them together in your marketplace. This platform connects borrowers looking for unsecured loans with the investors (Lenders) who want to get high returns on their investments.

Peer-to-peer Lending is a system where investors can provide money for the borrowers at a certain interest rate, where both the parties meet their needs. (on agreed conditions)

The entire loan process may be completed online. The organization gathers and analyzes data from the borrower, such as his credit score and annual income. This information is delivered to all investors who are interested in making a financial investment. Following an examination of the borrower's information, the investor determines whether to invest totally, partially, or not at all. The borrower receives the final response, after which he may decide whether or not to proceed with the transaction.

Lending Club has a huge cost advantage over traditional banks. Lending Club's cost savings are passed on to borrowers with the finest credit histories, resulting in reduced interest rates. Ideally Lending Club passes the risk of default loans to the Lenders.

*Advantages of Lending Club:*

LendingClub saves money by using technology to run its online credit marketplace at a lower cost than traditional lending programs, passing the savings on to borrowers in the form of cheaper rates and providing investors with the opportunity to make a profit.

Now that COVID has hastened Americans' migration to digital banking, the customer benefits of this purchase are especially evident. LendingClub, as the only full-spectrum fintech marketplace bank, will be able to leverage our technology and data-driven platform to provide new products and services to our millions of members that will help them both pay less when they borrow and pay less when they pay back their loans.

*Advantages to Customers/Borrowers:*

• One of the major advantages is convenience. Borrowers can shop through various loans and interest

rates by sitting at home

• Borrowers enjoy a lower rate of interest as compared to traditional lending systems/banks

• The process is simplified, secure and transparent

• It’s not an institution like a bank. All the operations are online, and savings are passed on to the

stakeholder's

*Advantages to Creditors/Investors:*

• Lenders or investors enjoy higher and faster yield

• It becomes a new asset class, which enables diversification, for the investors

• It is also transparent as all the information related to loans is available for investors

**How do platforms make money?**

Lending club collects fees from both borrowers and investors. If you are an investor, the following fees apply a service fee of one percent (1%) of the amount of any borrower payments received by the payment due date or during applicable grace periods.

Depending on the loan grade and term, borrowers pay a one-time origination charge ranging from 1.11 percent to 5% of the total loan amount.

In the banking industry, efficiency is assessed by the operating ratio, which is defined as marketing costs divided by the total number of loans outstanding. This normally equates to roughly 5% to 7%. As a result, a $100 loan will cost the bank $5 to $7

When borrowers miss payments and loans become late, LendingClub uses best practices from the banking industry to bring delinquent loans back to “current” status.

Lending Club, on the other hand, has an operating ratio of just 2%, which means that issuing a $100 loan costs the company only $2 since they operate fully online, they save on the number of employees and do not incur any location cost.

**2. Your team’s ultimate goal is to help a client determine whether s/he should invest in p2p loans. What is the final decision that you will help the client make? What is the objective, and how will you evaluate ‘better’ vs ‘worse’ decisions? What is the goal of predictive models for this? What will be the potential target variables?**

Our final recommendation to the customer is to invest in loans that will be repaid and will not default. We can determine whether decisions are better or worse by examining loan grade, interest rates, loan amount, loan status, dti, yearly income, employment duration, purpose, state address, and actual return and identifying patterns that forecast where defaults occur. For instance, avoiding investing in a purpose that is prone to default is a wiser choice. Predictive models are used to determine which circumstances result in defaults, and how to avoid them. Actual return, actual term, average yearly return, and recoveries may all be seen as possible goal factors that aid in making more informed selections.

**3. Data exploration**

**(a) Take a look at the data attributes. How would you categorize these attributes, in broad terms, considering what they pertain to? What are attribute types - which are numeric, categorical, and date variables? What do you think will be the important attributes to consider for your decision task? Which attributes do you think will help determine performance?**

* *Categories of Data Attributes according to what they are related to:*

Borrower characteristics - employer title, homeownership, employment length, annual income, state address, zip code

Loan characteristics - purpose, loan amount, funded amount

Lending Club Platform decisions - grade, subgrade, interest rate, verification status

Loan performance: loan status, installment, dti, revol balance, last payment, total payment

* *Types of Data attributes:*

Qualitative - employer title, homeownership, purpose, grade, subgrade, loan status, verification status,

Quantitative - employment length, annual income, loan amount, funded amount, interest rate, installment, dti, revol balance, last payment, total payment

* Important attributes for decision making: annual income, total payment, dti, grade, subgrade
* Attributes that will help determine performance: funded amount, total payment, actual return, average yearly return

**(b) How will you calculate performance (returns) from a loan? There are multiple ways for calculating this. Outline two ways to calculate returns based on the data attributes; what are their advantages and disadvantages**

**A.** One of the methods for calculating performance returns from a loan is Annual Return.

#Annualized return %: ( (lcdf$total\_pymnt - lcdf$funded\_amnt)/ lcdf$funded\_amnt ) \*(12/36)\*100

lcdf$annRet <- ((lcdf$total\_pymnt -lcdf$funded\_amnt)/lcdf$funded\_amnt)\*(12/36)\*100

Annual return is calculated using a formula where total amount is subtracted by funded amount and then it is divided by funded amount which is multiplied by 33%. This yields a fraction for the annualized return.

Another method for calculating performance returns on a loan is actual return

# Actual return: (lcdf$total\_pymnt - lcdf$funded\_amnt)

lcdf$actualReturn <- lcdf$total\_pymnt -lcdf$funded\_amnt

Actual Return is calculated using a formula where total payment is subtracted from the funded amount which gives us the return value of the particular loan.

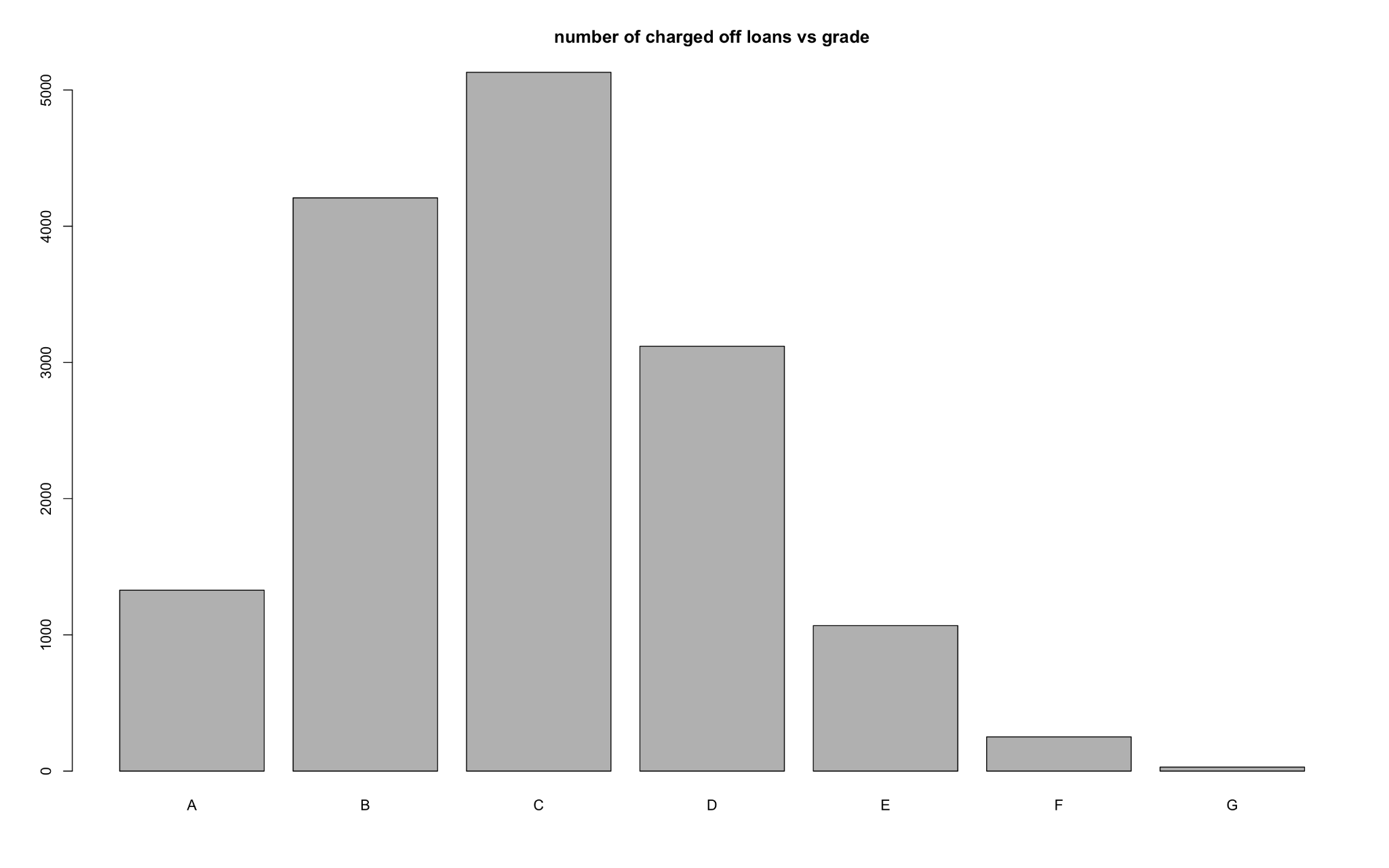
*Advantages and Disadvantages:*

Byusing Actual Return , we can determine the exact value of the return of a loan after the entire loan is paid off. By using this value we can determine the exact amount of profit earned.

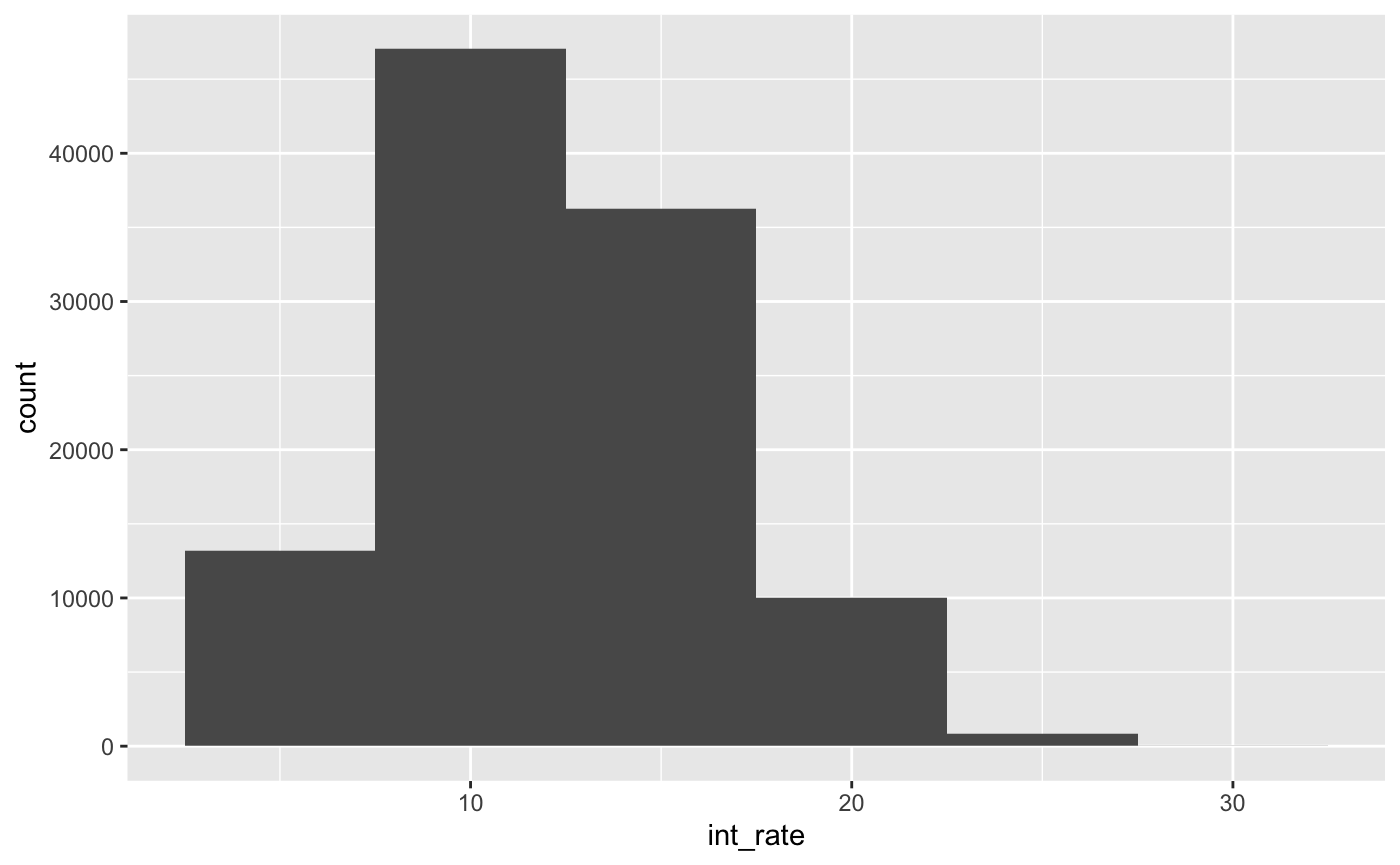
By using Annual return, we can determine the annual return value of a loan which will help investors determine the health of a loan and make a decision to recover based on this value.

An actual return refers to the actual gain or loss an investor experiences on an investment or in a portfolio.

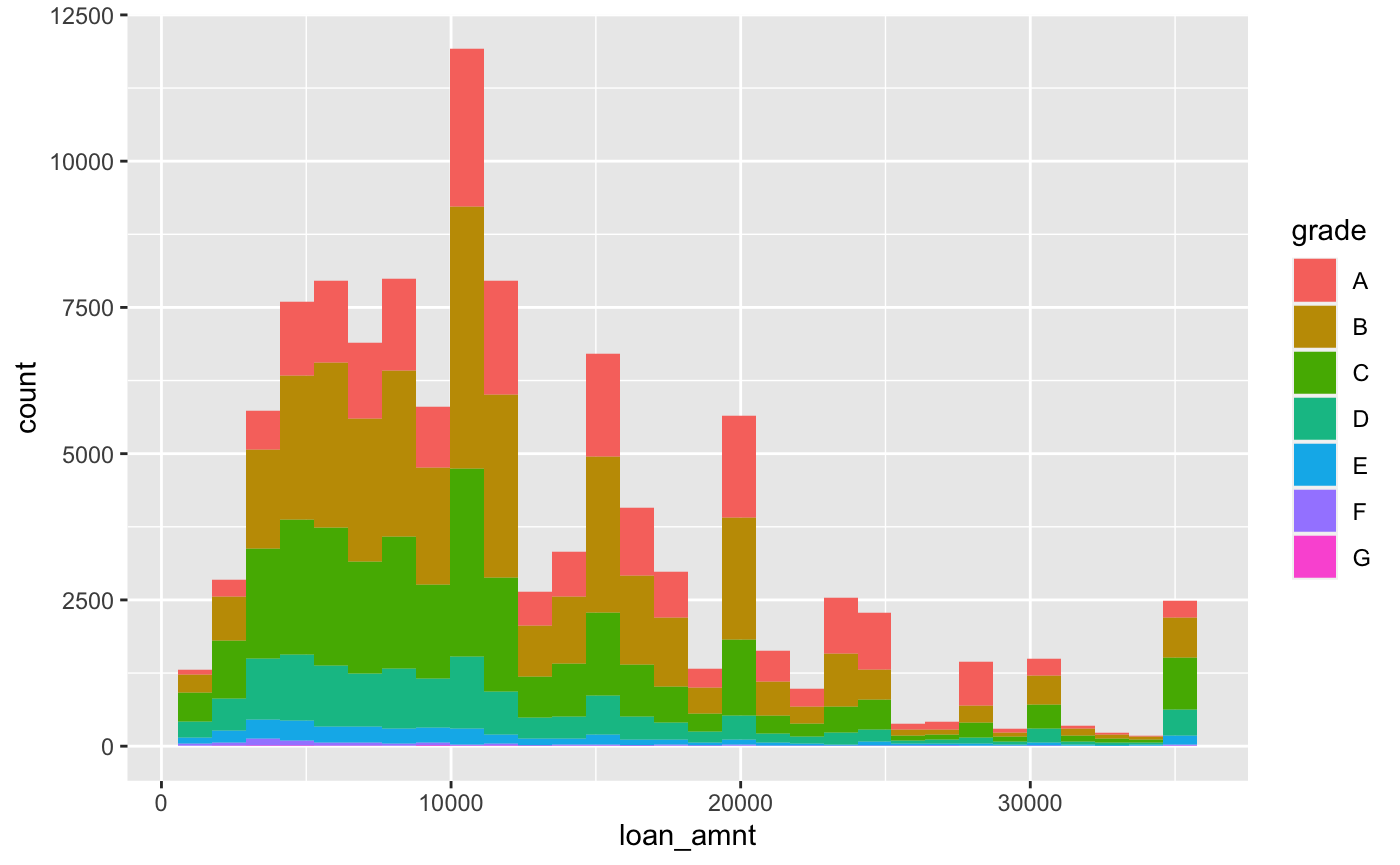
**3 (c) Examine the attributes which you think will be useful in your analyses and modeling. Obtain data descriptions and develop some plots to visualize the data. Summarize your observations (you answer should be more than just the figures and plots – what is the ‘story’ from your initial observations)**

****

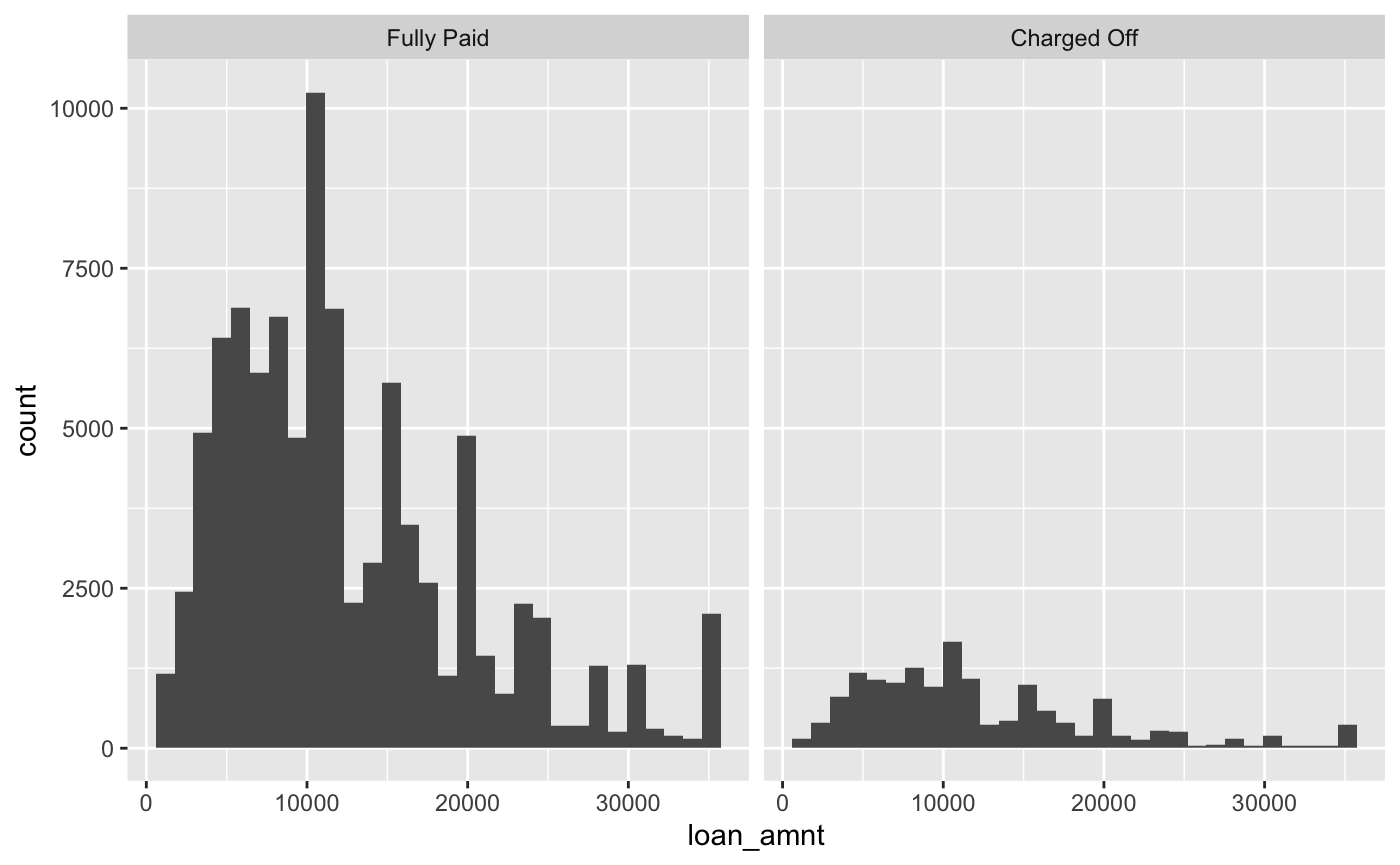
Charged Off loans increase from grade A to C and then decrease to G. This might be because safest loans are anyway paid and riskier loans aren’t granted easily, so a low number of defaults.

​​

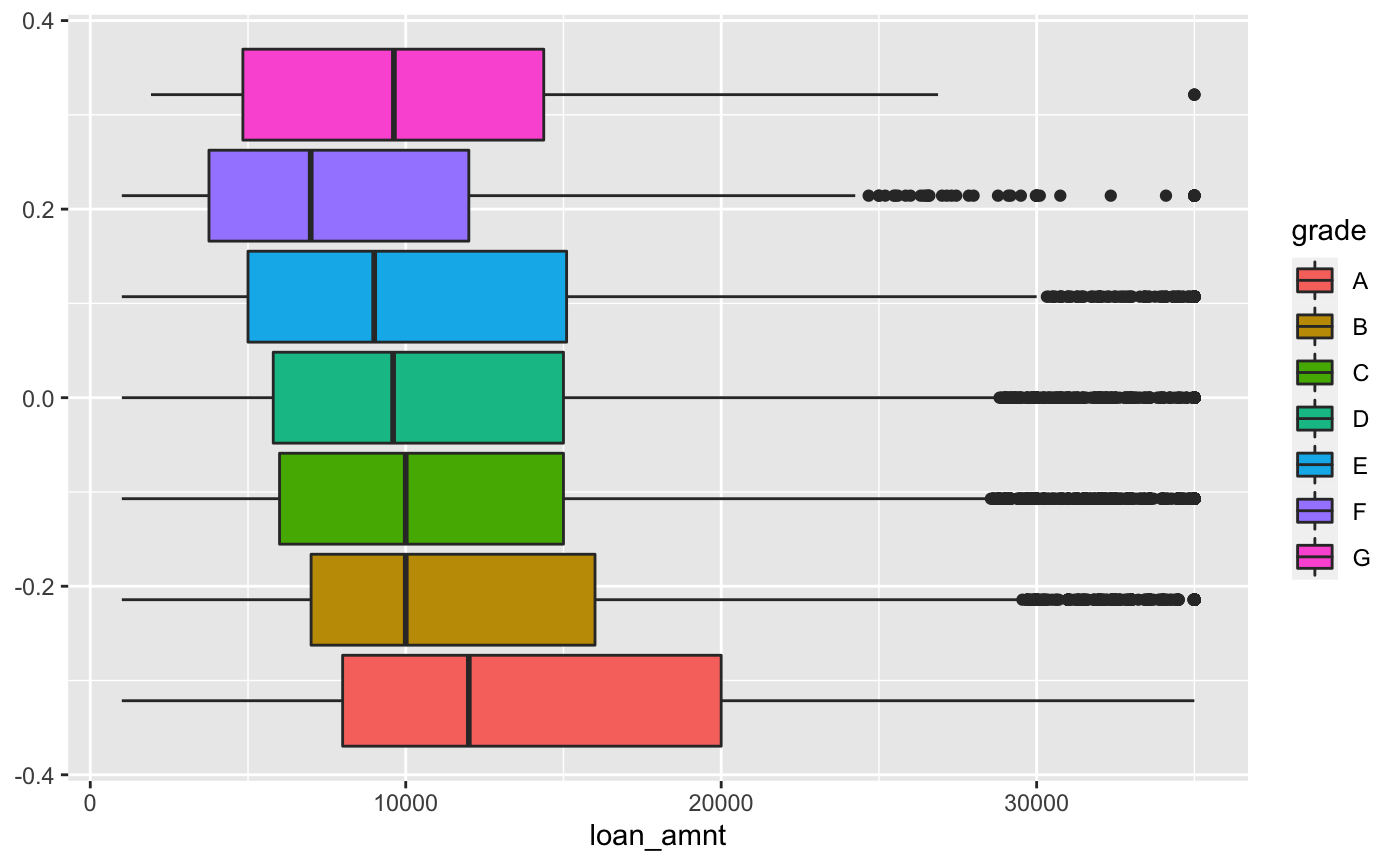
The number of loans issued at lower and higher interest rates is less compared to median interest rates. Higher interest rates are riskier loans, and lower interest rates give less returns. So, lenders might prefer to invest more in loans with median range interest rates



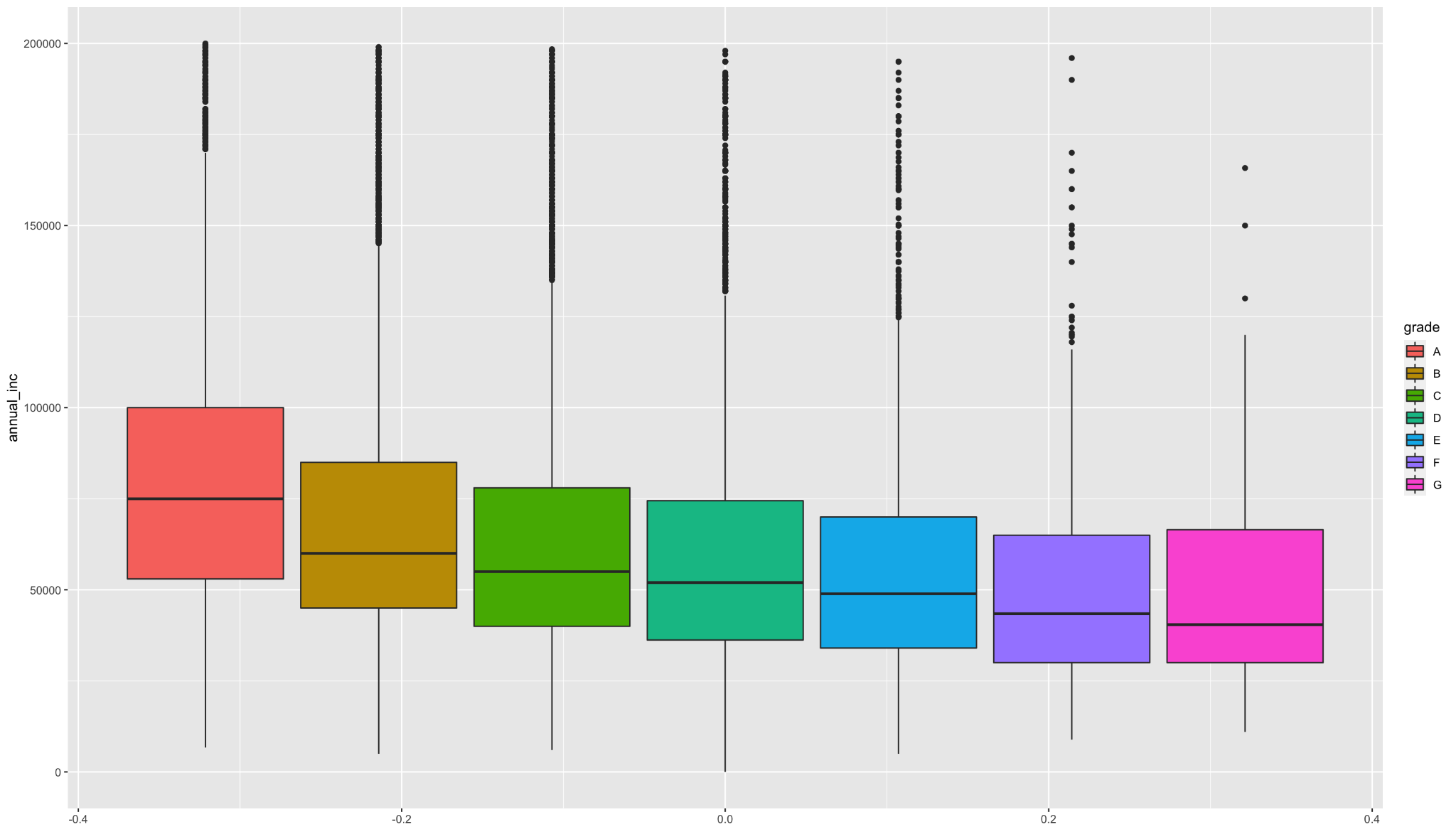
The number of loans issued as per different loan amounts increases and decreases. The returns increase with loan amount so lenders might prefer to invest in the median range (around $10,000), because the returns are better than smaller amounts and the chances of being paid back is better than higher loan amounts. But there’s an unusual pattern where the number of loans issued for loan amount $40,000 are more and larger, and part of them belong to grades B and C.



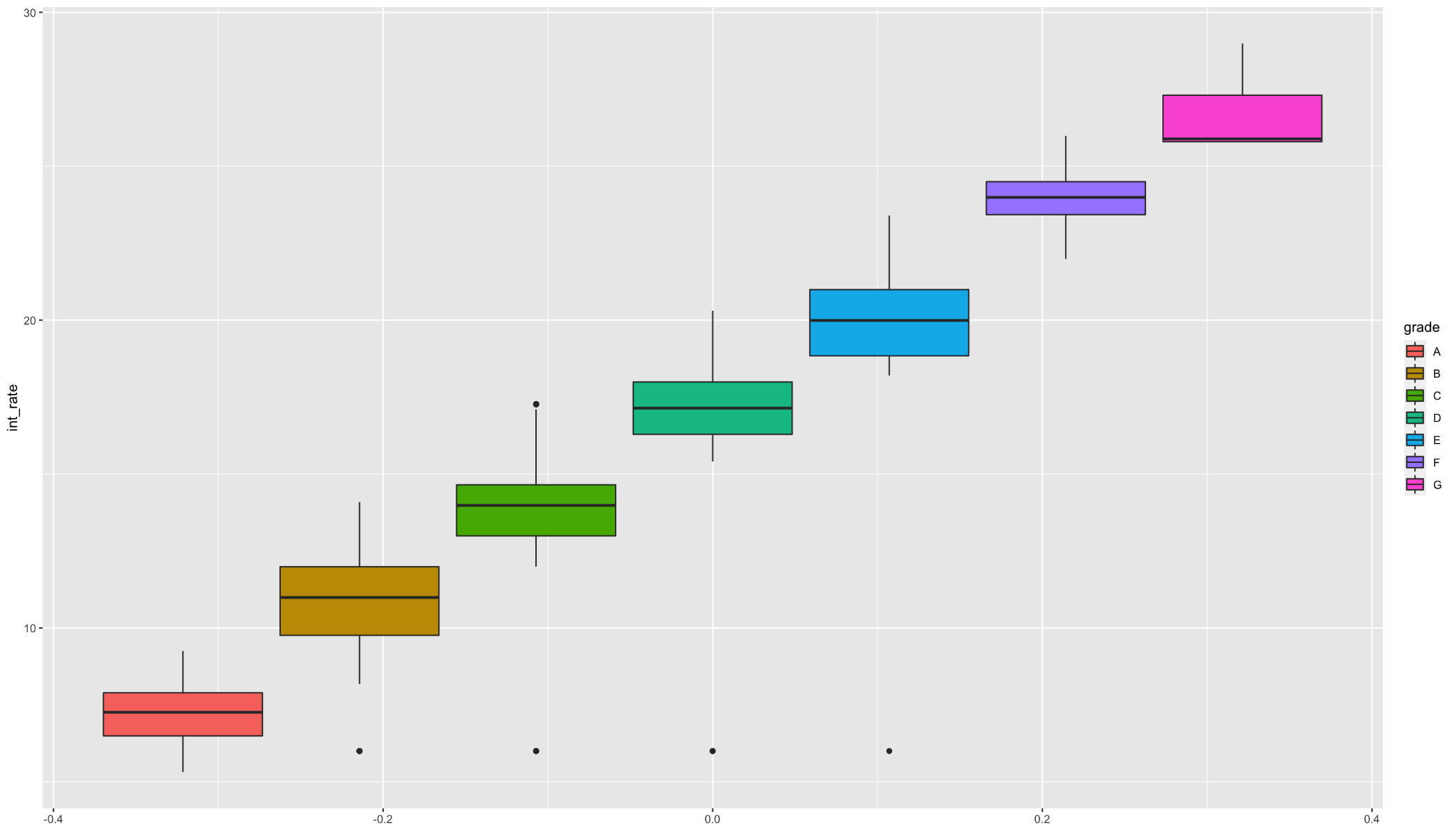
The number of loans that are fully paid and defaulted wrt loan amount follow a similar pattern, they are proportional.



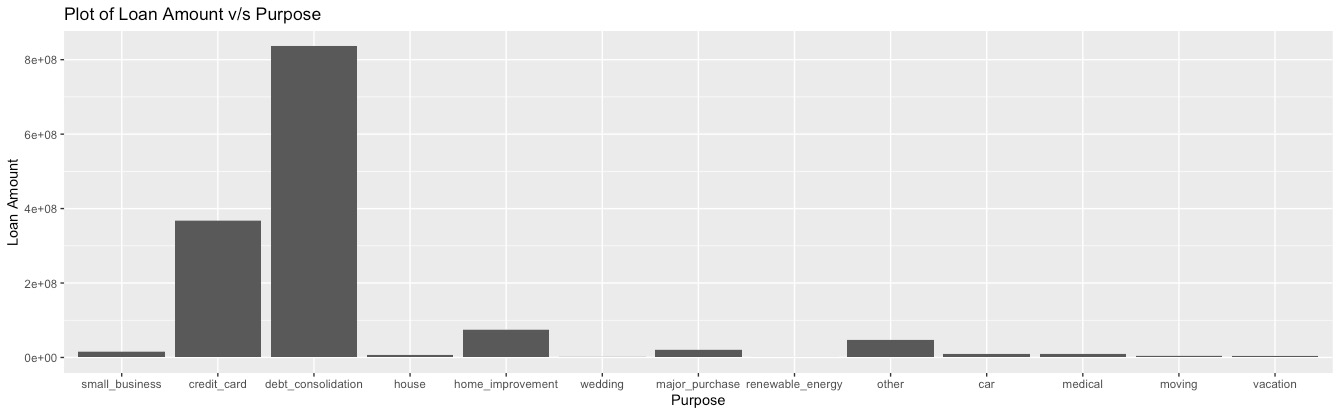
The average loan amount issued for grade A (safest grade) is more, then decreases in Grade B and slightly lowers to Grade F. Unusual pattern is seen for Grade G where the average loan amount increases.

****

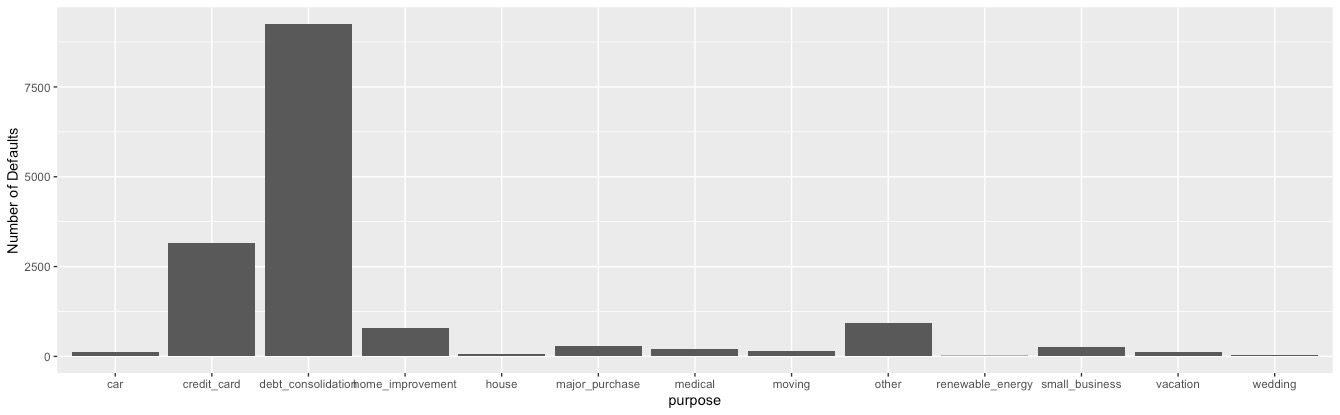
People with higher average annual income tend to pay off loans and their loans are graded safest. As the average annual income decreases, the grade of loans become riskier.



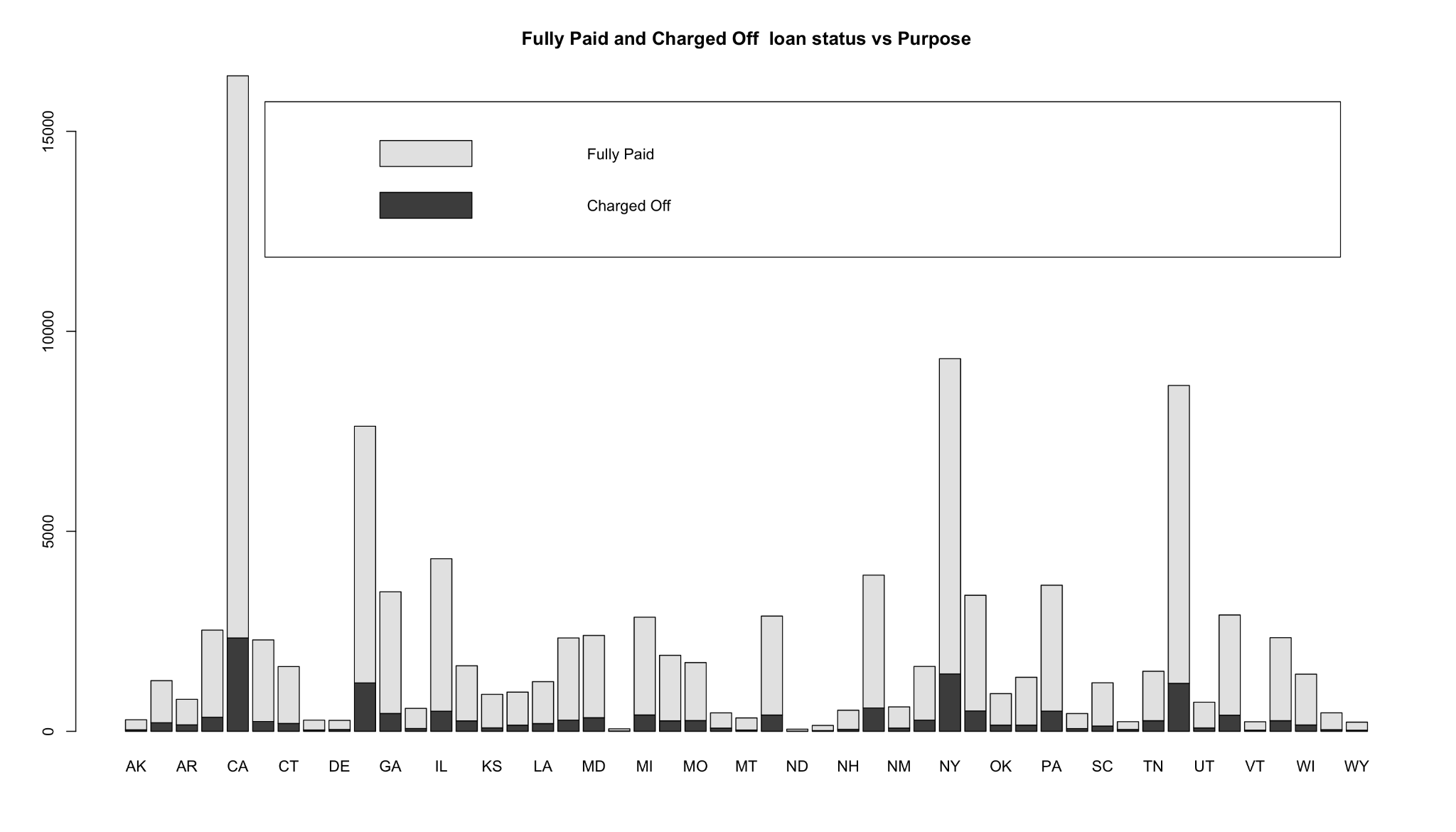
The interest rate of loans increases as the grades of loans become riskier. This is an expected pattern because if loans are riskier the platform assigns higher interest rates.

****

From the above bar chart, we can see the number of loans taken for each purpose. The highest number of loans are taken for debt consolidation, followed by credit card and home improvement.

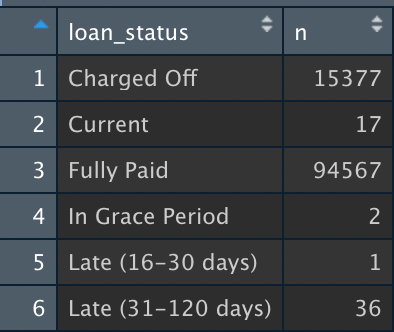
****

Since the number of loans taken for debt consolidation are more, even the defaults are more. Relatively, credit card and home improvement loans. But comparing both the barcharts of loan amount and defaults vs purpose, small business, major purchase, medical, vacation, and other purposes seem to have an increase in defaults pattern wrt loan amount pattern.

****

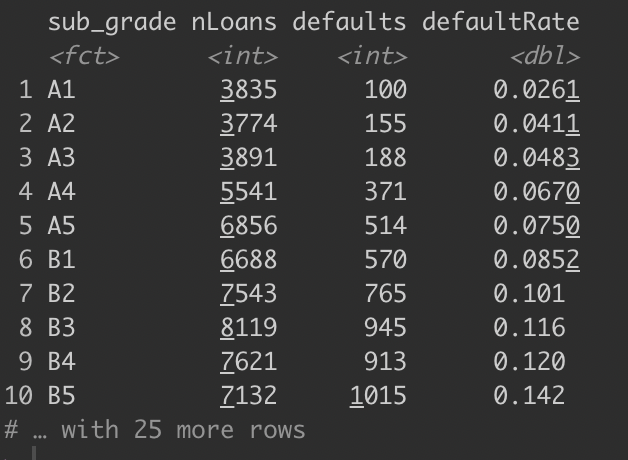
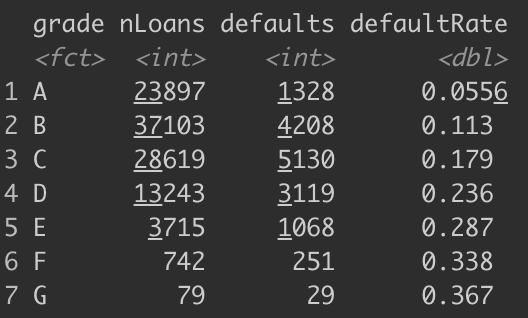
Barplot shows the states which have comparatively lower proportion of default loans, which are WV, ND, NH, MT, KS. States which have higher proportions of default loans are AR, SD, TN, NV, AL, MS.

**3d i) What are the values for loan\_status? Are there values other than “fully paid”, “charged off”? We want to restrict attention to “fully paid” and “charged off” loans, so other values should be removed.  
What is the proportion of defaults (‘charged off’ vs ‘fully paid’ loans) in the data?  
How does the default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?**



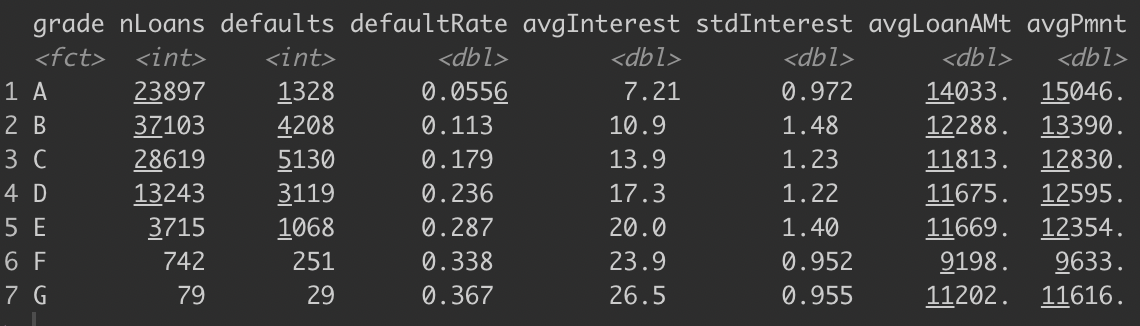
The values of loan status other than fully paid and charged off are Current, In grace period, late (16-30 days), Late (31-120 days).

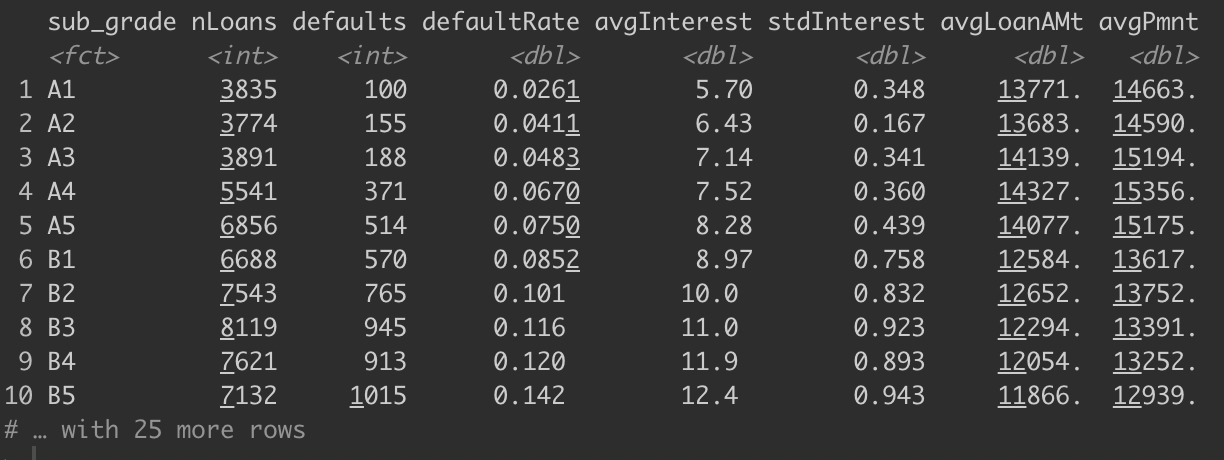
Proportion of defaults (charged off vs fully paid loans) is: 15377: 94567 = 1: 6.15

****

Default rate increases with riskier grades because number of defaults increases more than increase in number of loans

**ii) How many loans are there in each grade? And do loan amounts vary by grade? Does interest rate for loans vary with grade, subgrade? Look at the average, standard-deviation, min and max of interest rate by grade and subgrade. Is this what you expect, and why?**



Number of loans issued in each grade decreases from safe to risky loans, because the total number of investments in safe loans are more and risky loans are less.

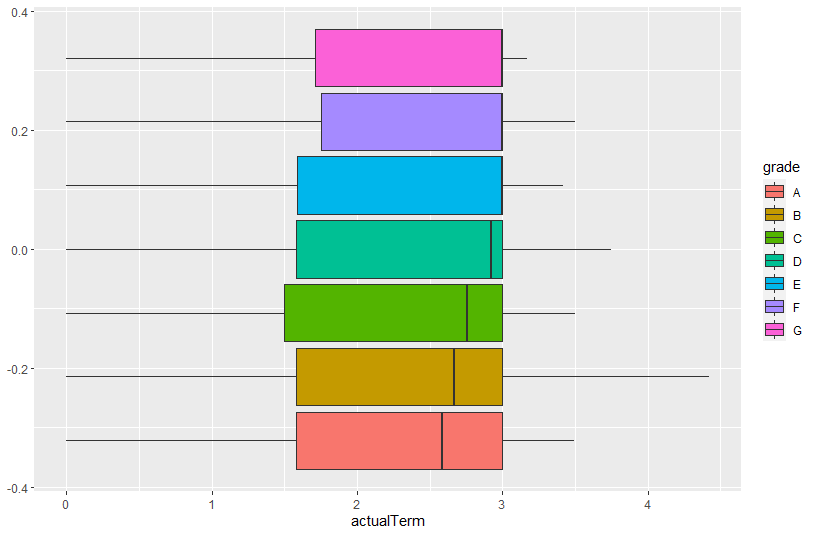
Loan amounts decrease until grade F and then increase for grade G, which is abnormal.

The average of interest rate by both grade and subgrade, increases from safe loans to risky loans. This is a seen pattern because sub grades are just deeper classifications in grades.

Standard deviation increases and then decreases in both grades and sub grades. This may be because interest rates vary more for medium risk loans, compared to safer grade loans where interest rates are certainly less and for risky loans, interest rates are high.

The minimum interest rates increase from grade A to B and constant from B through E, and there’s a sudden surge in interest rates for F and G.

The maximum interest rate also increases gradually from A to G grade and A1 to G5 subgrade.**iii) For loans which are fully paid back, how does the time-to-full-payoff vary? For this, calculate the ‘actual term’ (issue-date to last-payment-date) for all loans. How does this actual term vary by loan grade (a box-plot can help visualize this)?**

****

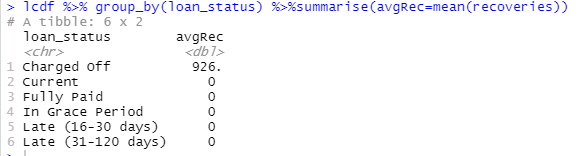
For all the grades people having 3 years actual period lie in the 75th percentile whereas the median lies between 2.5 and 3.

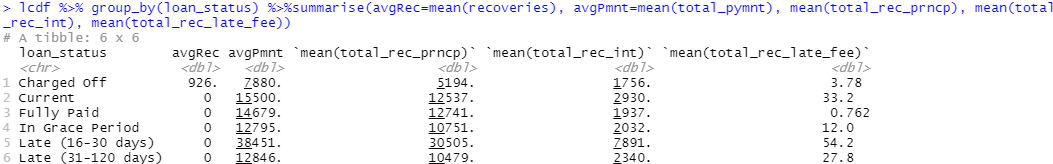
**(iv) What is 'recoveries'? Can we assume that recoveries are only for Charged\_off loans? The data has multiple attributes on recoveries – what is the total amount of recoveries? For charged-off loans, does total\_pymnt include recoveries?**

The recovery rate is the extent to which principal and accrued interest on defaulted debt can be recovered. (Source: <https://www.investopedia.com/terms/r/recovery-rate.asp>)

Debt recovery is when a loan—such as a credit card balance—continues to go unpaid, and a creditor hires a third party, known as a collection service, to focus on collecting the money.

**Source:** <https://www.debt.org/advice/recovery/#:~:text=Debt%20recovery%20is%20when%20a,correlated%20to%20your%20credit%20score.>





**v) Calculate the annual return. Show how you calculate the percentage annual return. Is there any return from loans which are ‘charged off’? Explain. How does return from charged -off loans vary by loan grade? Compare the average return values with the average interest\_rate on loans – do you notice any differences, and how do you explain this? How do returns vary by grade, and by sub-grade. If you wanted to invest in loans based on this data exploration, which loans would you invest in?**

The annual return is calculated by calculating the difference of total payment received by the investor and the amount he funded in proportion to the funded amount and then is multiplied by the actual term.

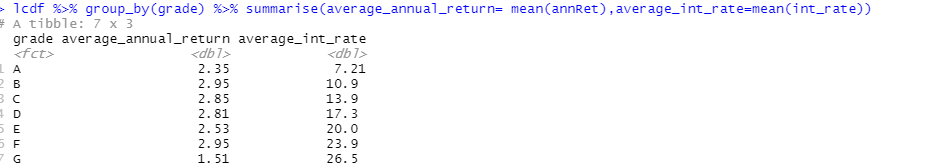
The actual term is the difference from the issue date to the last payment date.

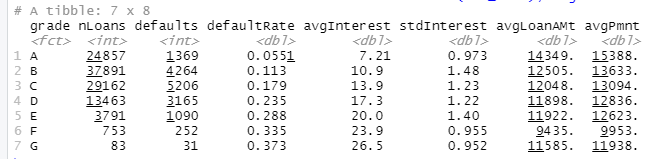
Percentage annual return = 100\*((Payment – Funded Amount)/(Funded Amount))\*(1/Actual Term)

Returns from ‘charged off’ loans are negative as expected. As loan grade changes, i.e., as it becomes riskier, the returns become more negative.

Average return values are less than the average interest rates. There are multiple reasons behind this. One of the reasons is that the loans are completed before their original term. Other reasons include adjustments in return rate by Lending Club based on future charged off rate and the service fee(1%) that the lender pays.

If I wanted to invest in loans, I would diversify my investments across loans from different grades based on average returns and default rate

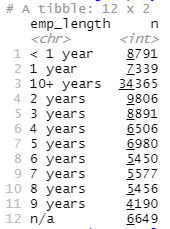


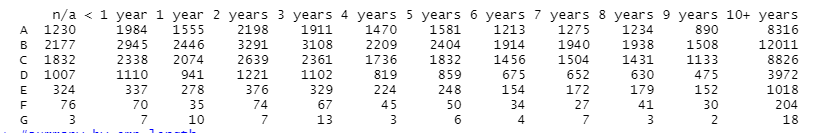


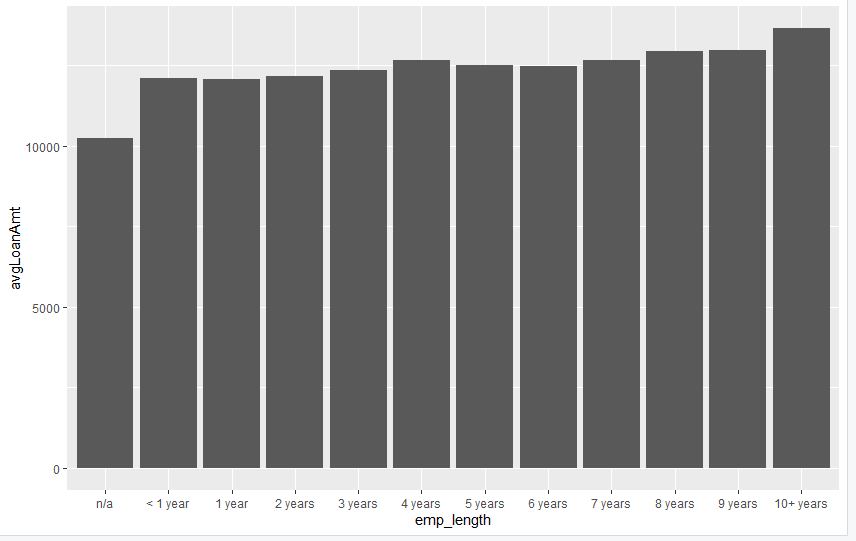
**(vi) What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? Do loan amounts vary by purpose? Do defaults vary by purpose? Does loan-grade assigned by Lending Club vary by purpose?**

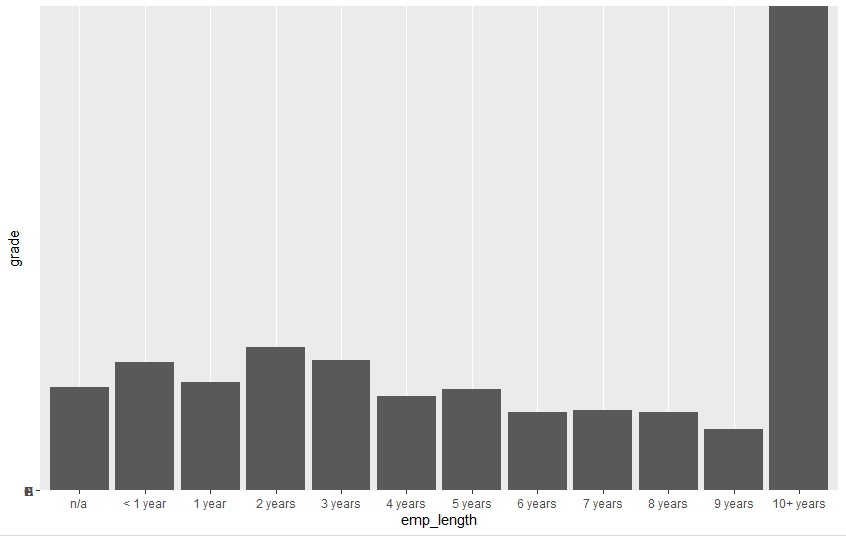
Debt consolidation and credit cards comprise ~82% of the total loans. There’s quite a big spread for average loan amounts by purpose. Debt consolidation, credit card, house, and small business have the high loan amounts(~$13k) being sanctioned as they require quite a big capital and on the other side of the spectrum is vacation which has the lowest average loan amount(~$5.6k) Default rate also has varied quite a bit where at one end is a small business with highest default rate at around 22% which is possible since they are susceptible to collapse and on the other end car, credit has lowest default rate at around 11%.

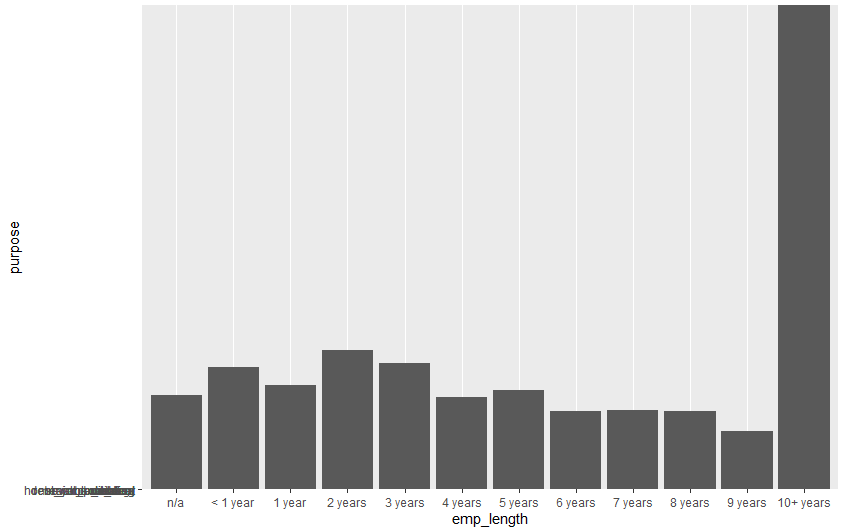
**(vii) Consider some borrower characteristics like employment-length, annual-income, fico-scores (low, high). How do these relate to loan attributes like, for example, loan\_amout, loan\_status, grade, purpose, actual return, etc.**

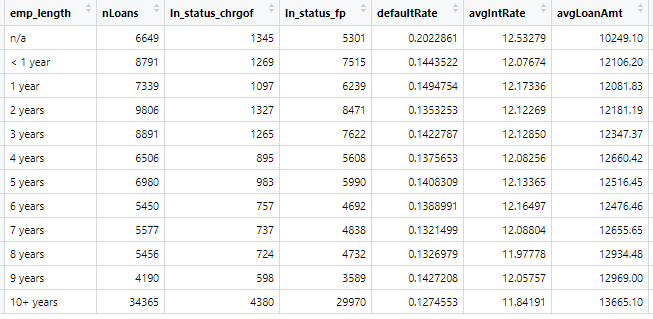
****

****

****

****

****

****

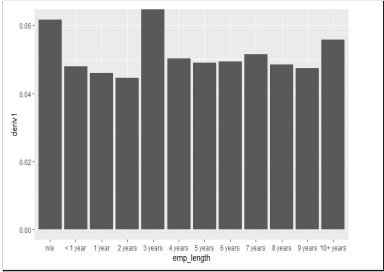
**(viii) Generate some (at least 3) new derived attributes which you think may be useful for predicting default. and explain what these are. For these, do an analysis as in the questions above (as reasonable based on the derived variables).**

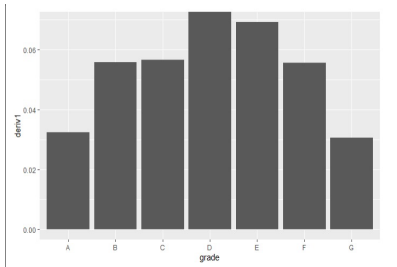
The attributes derived are:

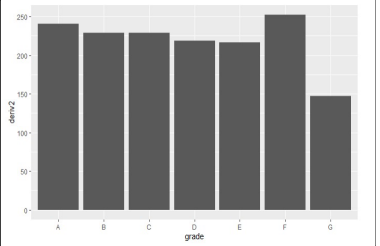
a) num\_acc\_ratio - It is the ratio between the number of accounts having dues for over 120 days (num\_accts\_ever\_120\_pd) and number of currently active accounts (num\_actv\_rev\_tl). This will help in understanding and figuring out account information of applicants that allows finding if the applicant often has dues in their accounts.

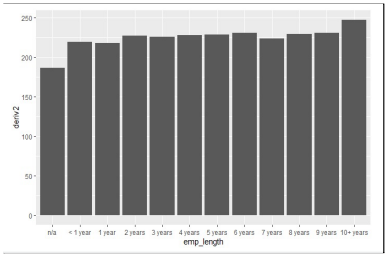
b) inc\_instal\_ratio - It calculates the ratio between the annual income of applicants and the value of the installments they are going to pay. This will help in understanding which applicants have a higher share ratio of income and installment amount.

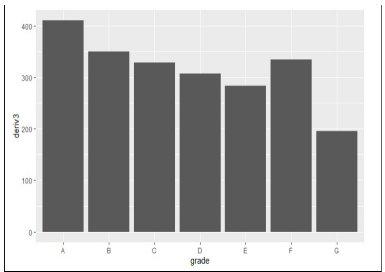
c) available\_bal\_instal\_ratio - This ratio is one between the difference of total balance in the applicant’s account from the revolving balance of the person and the installment.

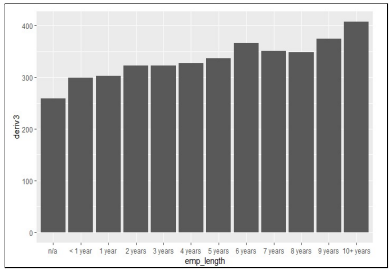


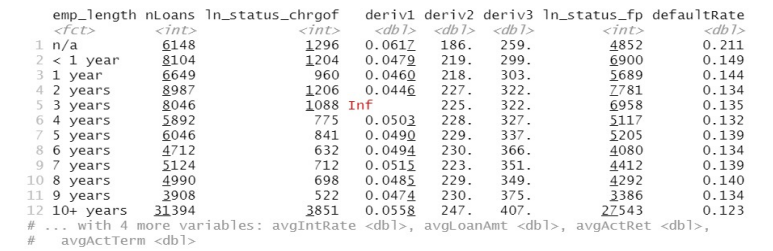












**3d2** *Summary:*

|  |  |
| --- | --- |
| From Grade A to G | Number of Charged Off loans ↑ and ↓  Average loan amount ↓ and ↑  Annual income ↓  Interest rate ↑  Default rate ↑  Actual term ↑ |

From our preliminary analysis, we observed that lenders tend to invest money in grades B and C, median range interest rates and median range loan amount.

Attributes like annual income, employment length, purpose, region, dti can be good indicators to minimize charged off loans.

**3.e) Are there missing values? What is the proportion of missing values in different variables? Explain how you will handle missing values for different variables. You should consider what the variable is about, and what missing values may arise from – for example, a variable monthsSinceLastDeliquency may have no value for someone who has not yet had a delinquency; what is a sensible value to replace the missing values in this case? Are there some variables you will exclude from your model due to missing values?**

Yes, there are many rows with missing values in the data set. Missing values may be because of many reasons, some maybe due to human error or maybe just because they were literally not available.

There are certain columns with more than 60% of the data values missing. We are eliminating these columns because it does not make sense to impute them because the significance of the columns may be lost, and the output may be something totally unexpected.

There are different ways in which we impute data, like mean imputation, in which we substitute the missing data by the mean of the remaining variables. Mean imputation does not generally preserve the relationship among variables.

Another way is regression imputation in which we can impute by predicting the missing values from one or more non missing values. The disadvantage however is that it might still lead to biased parameter estimates.

So, in general there is no particular method for imputation. It all is subjective to the kind of values we are imputing.

So, there were 59 variables which had 60% of rows with NA values and hence it was viable to remove those from our data set.

In a particular case a variable named “months since last Delinquency” has missing values.NO what we can do here is check the average of this variable per grade and then impute the missing values in the respective grade instead of taking average of entire column because its values will be generally high for lower grades compared to higher grade.

**4. Consider the potential for data leakage. You do not want to include variables in your model which may not be available when applying the model; that is, some data may not be available for new loans before they are funded. Leakage may also arise from variables in the data which may have been updated during the loan period (ie., after the loan is funded). Identify and explain which variables you will exclude from the model for leakage considerations, and explain why.**

In statistics and machine learning, leakage (also known as data leakage or target leakage) is the use of information in the model training process which would not be expected to be available at prediction time, causing the predictive scores (metrics) to overestimate the model's utility when run in a production environment.

[Source: <https://en.wikipedia.org/wiki/Leakage_(machine_learning)>]

|  |  |
| --- | --- |
| Leakage Variables | funded\_amnt\_inv, total\_pymnt, |
|  | total\_pymnt\_inv, total\_rec\_prncp, total\_rec\_int |
| Variables which are not useful |  |
|  |  |
|  | funded\_amnt\_inv, term, emp\_title, |
|  | pymnt\_plan,hardship\_flag, title, zip\_code, title, |
|  | zip\_code, out\_prncp, out\_prncp\_inv, |
|  | total\_pymnt, total\_pymnt\_inv, total\_rec\_prncp, |
|  | total\_rec\_int, total\_rec\_late\_fee, recoveries, |
|  | collection\_recovery\_fee, last\_pymnt\_d, |
|  | last\_pymnt\_amnt, last\_credit\_pull\_d, |
|  | policy\_code,annRet,annual\_return,pct\_annual\_r |
|  | eturn,collections\_12\_mths\_ex\_med,inq\_last\_6m |
|  | ths,actualTerm,installment, |
|  | emp\_length,verification\_status,num\_tl\_30dpd,ac |
|  | c\_now\_delinq,chargeoff\_within\_12\_mths,num\_tl |
|  | \_90g\_dpd\_24m,delinq\_amnt,tax\_liens,pub\_rec,d |
|  | elinq\_2yrs,initial\_list\_status,tot\_coll\_amt,num\_a |
|  | ccts\_ever\_120\_pd,mths\_since\_last\_delinq,mths\_ |
|  | since\_recent\_inq,percent\_bc\_gt\_75,debt\_settle |
|  | ment\_flag,earliest\_cr\_line,pub\_rec\_bankruptcies |
|  | ,application\_type,openTototal,creditPerRevolv\_a |
|  | cc |
|  |  |

|  |  |
| --- | --- |
| *Variables with missing values:* |  |
| **Variables** | **Proportions (%)** |
| mths\_since\_last\_delinq | 48.24 |
| revol\_util | 0.05 |
| bc\_open\_to\_buy | 0.96 |
| bc\_util | 1.4 |
| mo\_sin\_old\_il\_acct | 3.54 |
| mnths\_since\_recent\_bc | 0.94 |
| mnths\_since\_recent\_inq | 11.16 |
| num\_tl\_120dpd\_2m | 5.87 |

|  |  |
| --- | --- |
| percent\_bc\_gt\_75 | 1.07 |

**Variables which have more than 60% missing values:**

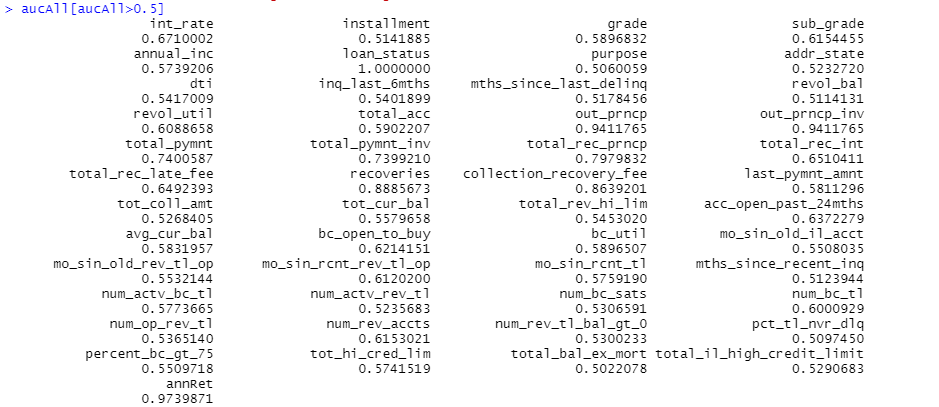
|  |  |  |  |
| --- | --- | --- | --- |
| **1** | id | **29** | sec\_app\_earliest\_cr\_line |
| **2** | member\_id | **30** | sec\_app\_inq\_last\_6mths |
| **3** | url | **31** | sec\_app\_mort\_acc |
| **4** | desc | **32** | sec\_app\_open\_acc |
| **5** | mths\_since\_last\_record | **33** | sec\_app\_revol\_util |
| **6** | last\_pymnt\_d | **34** | sec\_app\_open\_act\_il |
| **7** | next\_pymnt\_d | **35** | sec\_app\_num\_rev\_accts |
| **8** | mths\_since\_last\_major\_derog | **36** | sec\_app\_chargeoff\_within\_12\_mths |
| **9** | annual\_inc\_joint | **37** | sec\_app\_collections\_12\_mths\_ex\_med |
| **10** | dti\_joint | **38** | sec\_app\_mths\_since\_last\_major\_derog |
| **11** | verification\_status\_joint | **39** | hardship\_type |
| **12** | open\_acc\_6m | **40** | hardship\_reason |
| **13** | open\_act\_il | **41** | hardship\_status |
| **14** | open\_il\_12m | **42** | deferral\_term |
| **15** | open\_il\_24m | **43** | hardship\_amount |
| **16** | mths\_since\_rcnt\_il | **44** | hardship\_start\_date |
| **17** | total\_bal\_il | **45** | hardship\_end\_date |
| **18** | il\_util | **46** | payment\_plan\_start\_date |
| **19** | open\_rv\_12m | **47** | hardship\_length |
| **20** | open\_rv\_24m | **48** | hardship\_dpd |
| **21** | max\_bal\_bc | **49** | hardship\_loan\_status |
| **22** | all\_util | **50** | orig\_projected\_additional\_accrued\_interest |
| **51** | hardship\_payoff\_balance\_amount |
|  |  | **52** | hardship\_last\_payment\_amount |
| **23** | inq\_fi | **53** | debt\_settlement\_flag\_date |
| **24** | total\_cu\_tl | **54** | settlement\_status |
| **25** | inq\_last\_12m | **55** | settlement\_date |
| **26** | mths\_since\_recent\_bc\_dlq | **56** | settlement\_amount |
| **27** | mths\_since\_recent\_revol\_delinq | **57** | settlement\_percentage |
| **28** | revol\_bal\_joint | **58** | settlement\_term |

**5. Do a univariate analysis to determine which variables (from amongst those you decide to consider for the next stage prediction task) will be individually useful for predicting the**

**dependent variable (loan\_status). For this, you need a measure of relationship between the dependent variable and each of the potential predictor variables. Given loan-status as a binary dependent variable, which measure will you use? From your analyses using this measure, which variables do you think will be useful for predicting loan\_status?**

**(Note – if certain variables on their own are highly predictive of the outcome, it is good to ask if this variable has a leakage issue).**

In this case, we'll use AUC (area under the curve) as the ROC graph's under-the-curve measure. We can also use accuracy as a metric; however, accuracy is less derivable when data is skewed to one datapoint of the answer variable. For example, 88 percent of the loans in this set are paid in full, so if we guess that all of the loans will be paid in full, we'll be 88 percent correct, which isn't bad but isn't particularly useful.



We have used AUC measure to identify the relationship between predictor (independent variable)

(Numeric variable) and dependent variable.

annualRet gives 0.98 AUC

And for the following variable the AUC is more 0.75

total\_rec\_prncp

last\_pymnt\_amnt

total\_pymnt\_inv

total\_pymnt

But these variables have the issue of data leakage as they were generated after the loan was

granted/funded or closed.

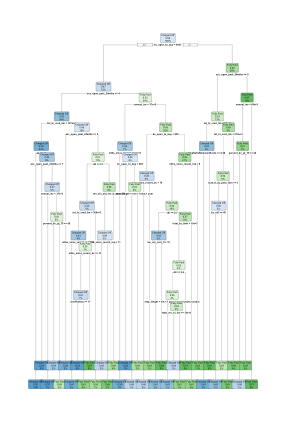
Rest of the variables’ AUC lies around 0.5

**Part-B**

**6. Develop decision tree models to predict default.**

**(a) Split the data into training and validation sets. What proportions do you consider, why?**

Data partitioning is a crucial aspect of data mining. Normally, data sets are divided into two categories: training and testing. The proportions we're looking at are 75% for the training set and 25% for the testing set. We're considering giving the majority of the data to training sets so that the model may learn all of the different types of possible patterns for defining the problem.



**(b) Train decision tree models (use both rpart, c50)**

**[If something looks too good, it may be due to leakage – make sure you address this]**

**What parameters do you experiment with, and what performance do you obtain (on training and validation sets)? Clearly tabulate your results and briefly describe your findings.**

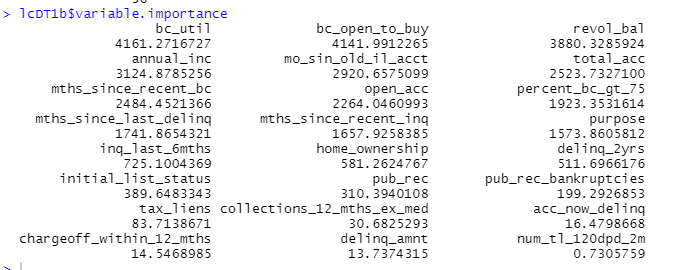
**How do you evaluate performance – which measure do you consider, and why?**

For the Decision Tree Model, we first removed the variables with 100% missing values, then divided the data 70:30 into training and validation. Then, using Training Data, we created a Decision Tree. We found that the actual term, actual return where 2 variables which had a very high importance so we removed them.

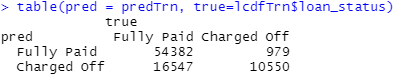
The model performance:

**–rpart:**

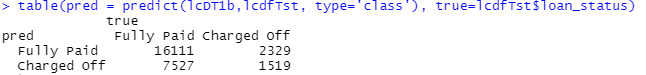
**Variable importance**

****

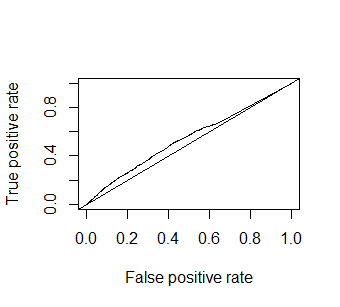
Confusion matrix for training data:



Confusion matrix for Test data: Accuracy 64.10%

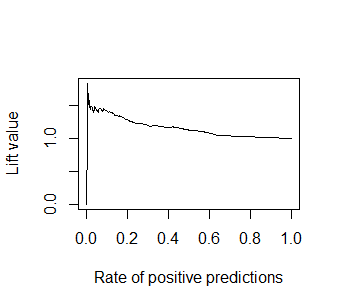


ROC Curve:



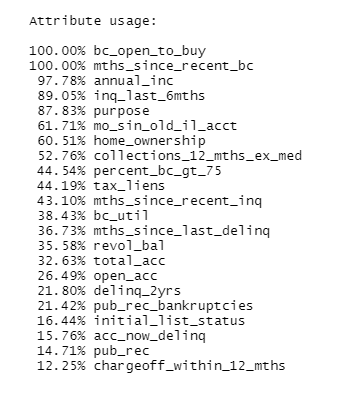
AUC is 0.5440

LIFT Curve:

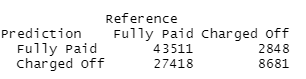


**C50:**

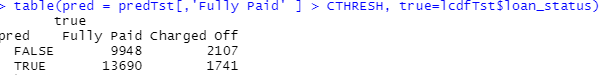
Variable Importance:



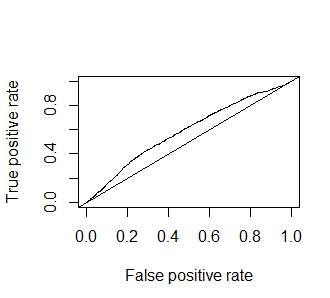
Confusion Matrix for training data:



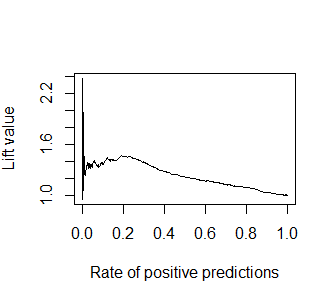
Confusion Matrix for Test Data: 57.44%



**ROC Curve**

****

**Lift Curve:**

****

**(c) Identify the best tree model. Why do you consider it best?**

**Describe this model – in terms of complexity (size). Examine variable importance.**

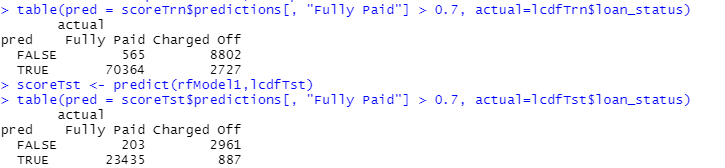
**How does this relate to your uni-variate analysis in Question 4 above? Briefly describe how variable importance is obtained (the process used in decision trees).**

When compared to c50, the Decision Tree model appears to be a better model because its accuracy is higher, and when comparing the confusion matrix, the Decision Tree model has less misclassifications of both Fully Paid and Charged Off.

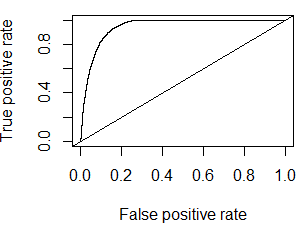
When we look at the table of variable importance, we can see that

we’ll see that bc\_util, bc\_open\_to\_buy and revol\_bal has higher significance as compared to the other variables and it’s backed by our analysis.

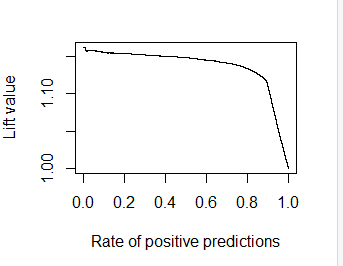
**7. Develop a random forest model. (Note the ‘ranger’ library can give faster computations) What parameters do you experiment with, and does this affect performance? Describe the best model in terms of number of trees, performance, variable importance. Compare the performance of random forest and best decision tree model from the previous question. Do you find the importance of variables to be similar/different? Which model would you prefer, and why?**

****

**ROC Curve:**

****

**AUC: 0.9470**

****

**8. The purpose of the model is to help make investment decisions on loans. How will you evaluate the models on this business objective? Consider a simplified scenario - for example, that you have $100 to invest in each loan, based on the model’s prediction. So, you will invest in all loans that are predicted to be ‘Fully Paid’. Key questions here are: how much, on average, can you expect to earn after 3 years from a loan that is paid off, and what is your potential loss from a loan that has to be charged off?**

**One can consider the average interest rate on loans for expected profit – is this a good estimate of your profit from a loan? For example, suppose the average int\_rate in the data is 12%; so, after 3 years, the $100 will be worth (100 + 3\*12) = 136, i.e a profit of $36. Now, is 12% a reasonable value to expect – what is the return you calculate from the data? Explain what value of profit you use.**

For the loans that are fully paid, the company on average makes a profit of 15.21% per year.

Considering we have invested $100 in paid loans; we will be earning $15.21 a year making the combined profit for three years to be $45.63.

While in the case of charged-off loans, the company incurs a loss of 8.22% per year, so the combined loss for three years will be 24.66%

Considering we have $100 in charged-off loans, we incur a loss of $24.66 over three years.

**For a loan that is charged off, will the loss be the entire invested amount of $100? The data shows that such loans do show some partial returned amount. Looking at the returned amount for charged off loans, what proportion of invested amount can you expect to recover? Is this overly optimistic? Explain which value of loss you use. (15377)**

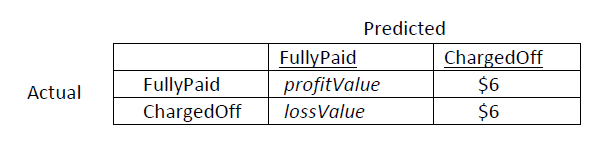
No, for a charged-off loan, the loss isn’t the entire invested amount. As considered in question, for a loan amount of $100, the loss isn’t the entire $100. Every borrower does repay some amount of the entire borrowed amount. This results in a reduction of the total loss incurred by LendingClub.

Upon detailed analyses of the data, it can be observed that a large chunk of the borrowed amount is repaid by the borrowers. To calculate the loss incurred by charged-off loans, we can find the total amount paid by the borrowers (total\_pymnt) and subtract it from the total borrowed amount (funded\_amnt). Then we can find the percentage of the same to understand the proportion of the recovered amount of the charged-off loans.

The results obtained cannot be called overly optimistic as it includes the average recovered amount from a dataset of 15,377 charged-off loans of the database. Considering the massive sample size and the process of calculating the amount, the data cannot be termed as overly optimistic.

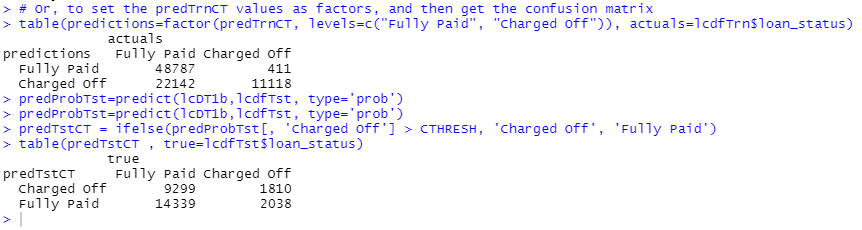
The loss value we can consider is the difference between the total amount paid by the borrowers (total\_pymnt) and the total borrowed amount (funded\_amnt).

**You should also consider the alternate option of investing in, say in bank CDs (certificate of deposit); let’s assume that this provides an interest rate of 2%. Then, if you invest $100, you will receive $106 after 3 years (not considering reinvestments, etc), for a profit of $6. Considering a confusion matrix, we can then have profit/loss amounts with each cell, as follows:**

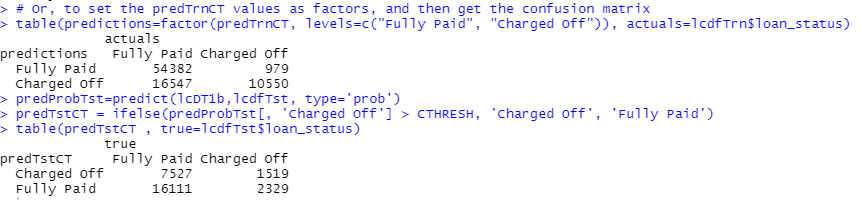
****

**(a) Compare the performance of your models from Questions 6, 7 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Evaluate different thresholds and analyze performance. Which model do you think will be best, and why.**

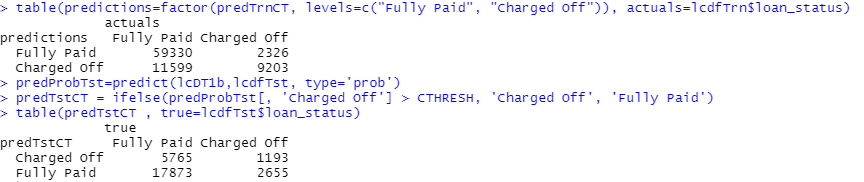
**Decision Tree:**

****

**0.5**

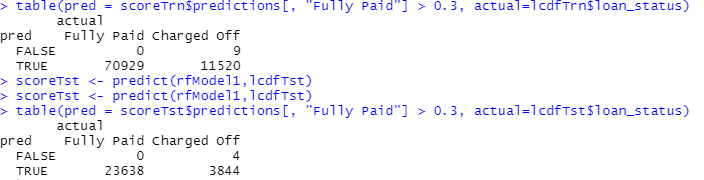
****

**0.7**

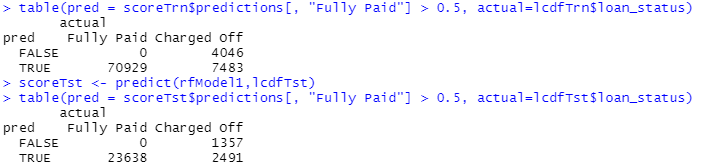
****

**Random Forest:**

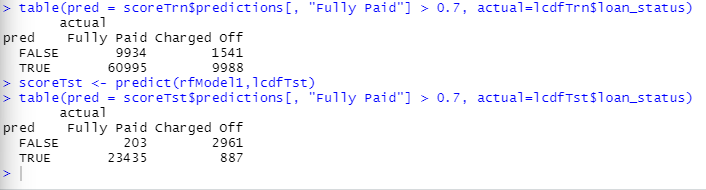
**0.3**

****

**0.5**

****

**0.7**

****

From the above images we can see that the random forest model with 0.7 as threshold is performing better than the same model with different threshold and the decision tree models.

The Misclassification rate with threshold = 0.7 is

(887+203)/27500 = 0.0396 = 3.96%

**(b) Another approach to determining the optimal threshold for implementing the model is to directly consider how the model will be used – you can order the loans in descending order of prob(fully-paid). Then, you can consider starting with the loans which are most likely to be fully-paid and go down this list till the point where overall profits begin to decline (as discussed in class). Conduct an analysis to determine what threshold/cutoff value of prob(fully-paid) you will use and what is the total profit from different models. Also compare the total profits from using a model to that from investing in the safe CDs. Explain your analyses and calculations. Which model do you find to be best and why. And how does this compare with what you found to be best in part (a) above.**

The total profits obtained from using a model is greater than investing in the safe CDs. The profit that we would have obtained from the safe CDs can be calculated by multiplying the rate of interest (5%) to the duration (3 years) with the total amount in hand.

If we consider the total amount to be $10000.

In the case of the CD, where the rate of interest per annum is 5% and duration is 3 years as in the case with our model.

The total profit through this method will be:

Profit earned from CD = 10000 \* 3 \* 0.05

= 1500

Based on our analysis, the Random Forest model is the best model with threshold = 0.7 and number of trees = 200 with the accuracy of ~94%.