**ASSIGNMENT 1**

**SUBMITTED BY**

**ASHOK BHATRAJU – 670248723**

**ARCHANA SINGH – 668528470**

**NIKITA BAWANE 661069000**

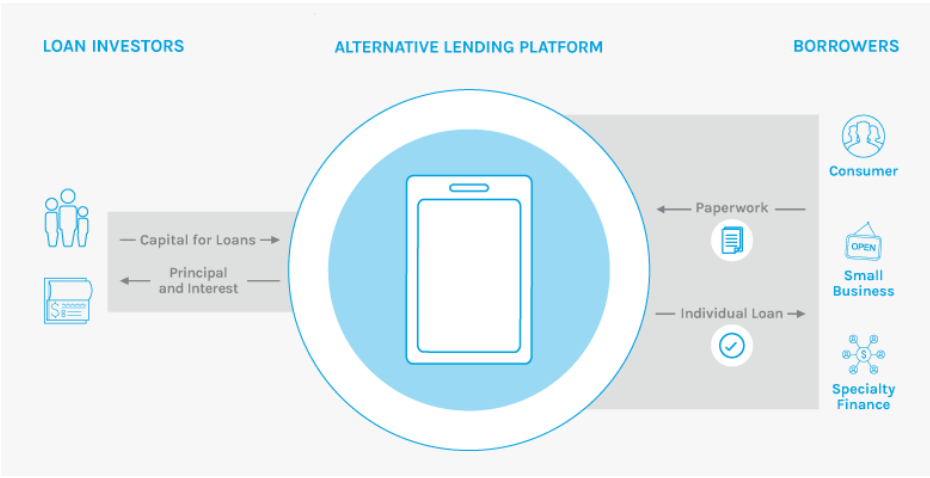
1. **Describe the business model for Lending Club. Consider the stakeholders and their roles, and what advantages Lending Club offers. How does the platform make money?**

Lending club is a peer-to-peer lending platform. It enables borrowers to create unsecured personal loans ranging between $1000 to $40000. Each borrower is categorized into different grades(A-F) based on their credit score, credit history, desired loan amount and the borrower’s [debt-to-income ratio](https://en.wikipedia.org/wiki/Debt-to-income_ratio), Lending Club determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees.

Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. The loans can only be chosen at the interest rates assigned by Lending Club, but investors can decide how much to fund each borrower, with the minimum investment of $25 per note.

Investors make money from interest. Rates vary from 6.03% to 26.06%, depending on the credit grade assigned to the loan request. The grades assigned to these requests range alphabetically from A to G, with A being the highest-grade, lowest-interest loan. Each of these letter grades has five finer-grain sub-grades, numbered 1 to 5, with 1 being the highest sub-grade.

**Stakeholders:**

* Borrowers – Borrows money for various personal requirements.
* Lending club platform – rates the borrowers based on the available information. Shares this information with investors to help them with their investment.
* Investors – selects borrowers and invests based on their risk appetite.

**Advantages Lending club offer:**

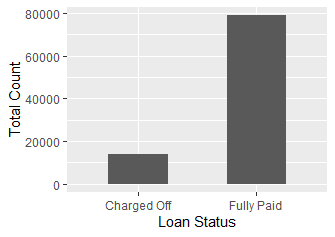
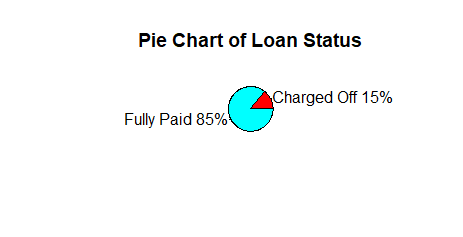
* Complete transparency about the borrowers.
* Competitive interest rates for both borrowers and investors
* Short term loan duration
* Flexible repayment options
* Better interest rates compared to majority of the lending options

**Lending club source of Money:**

Lending Club makes money by charging borrowers an [origination fee](https://en.wikipedia.org/wiki/Origination_fee) and investors a service fee. The size of the origination fee depends on the credit grade and ranges to be 1.1%-5.0% of the loan amount. The size of the service fee is 1% on all amounts the borrower pays.

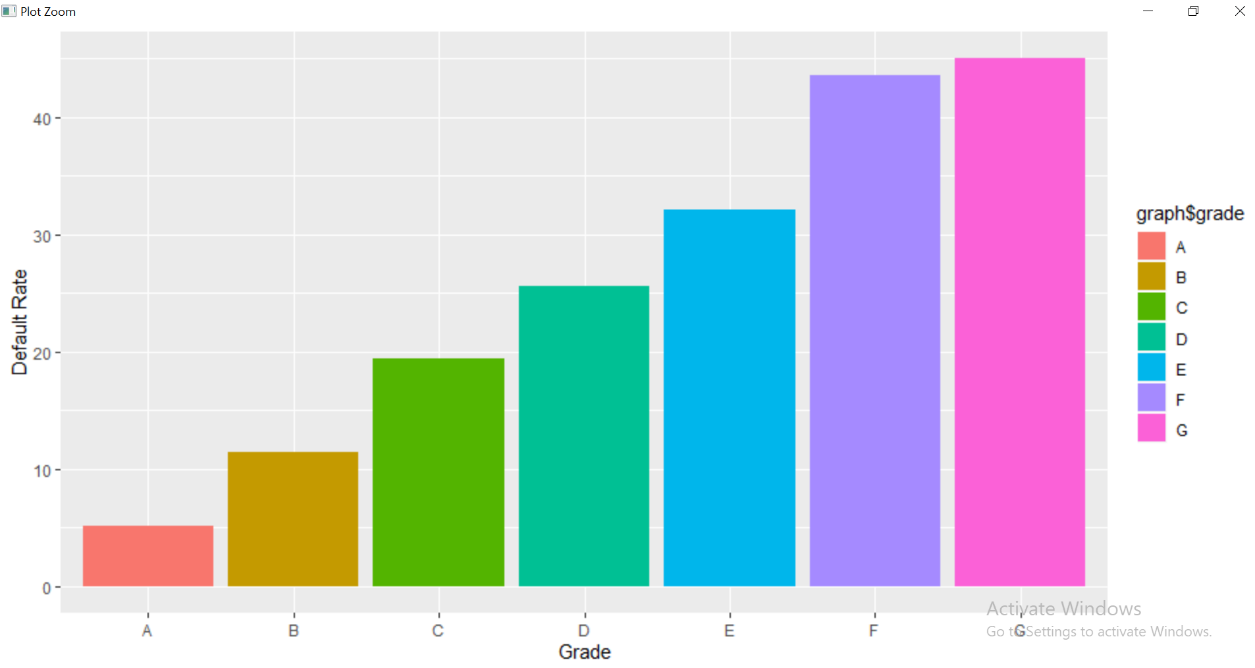
**Source:** Lending Club Wikipedia

1. **Data exploration (a)**

**i) What is the proportion of defaults (‘charged off’ vs ‘fully paid’ loans) in the data? How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why**?

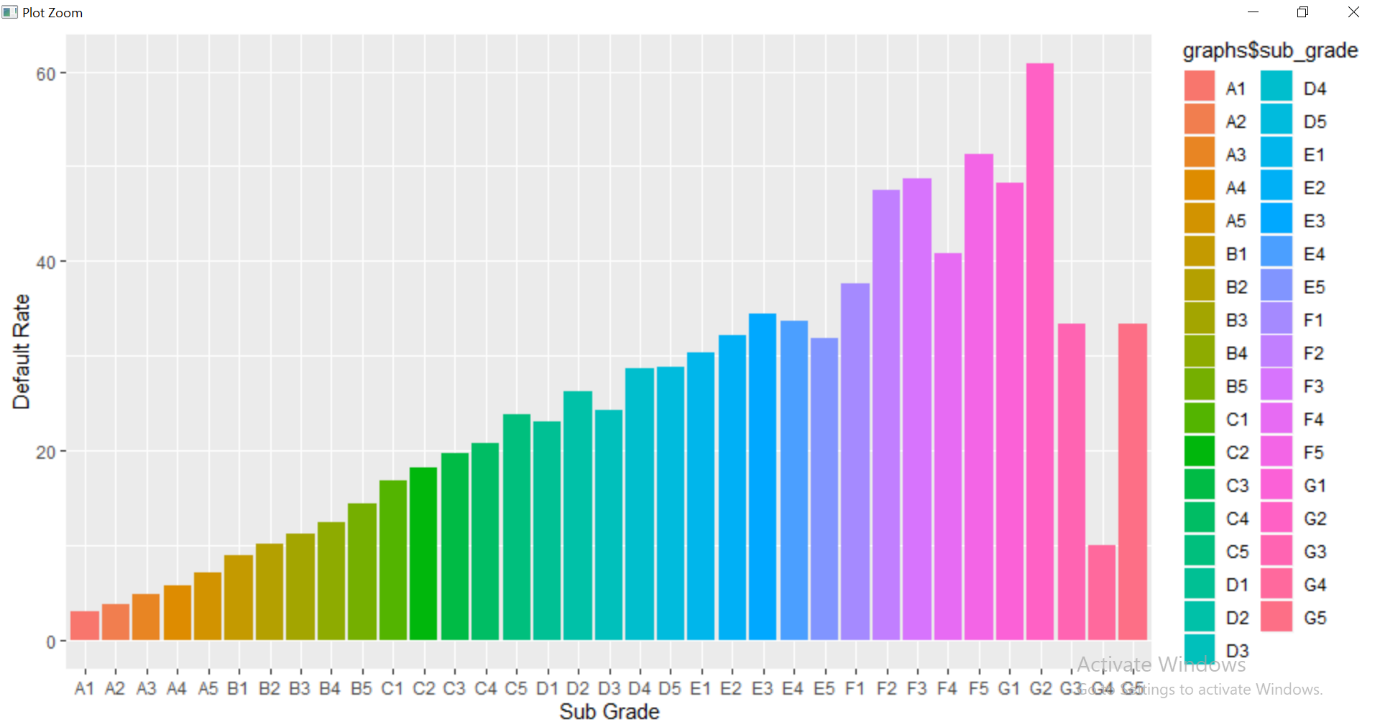
**Interpretation:**

From plots above its clear that major part of borrowers paid off their loans.



**Interpretation:**

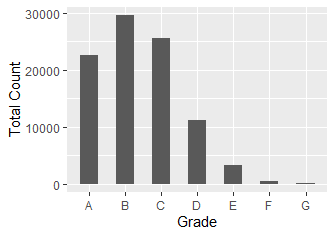
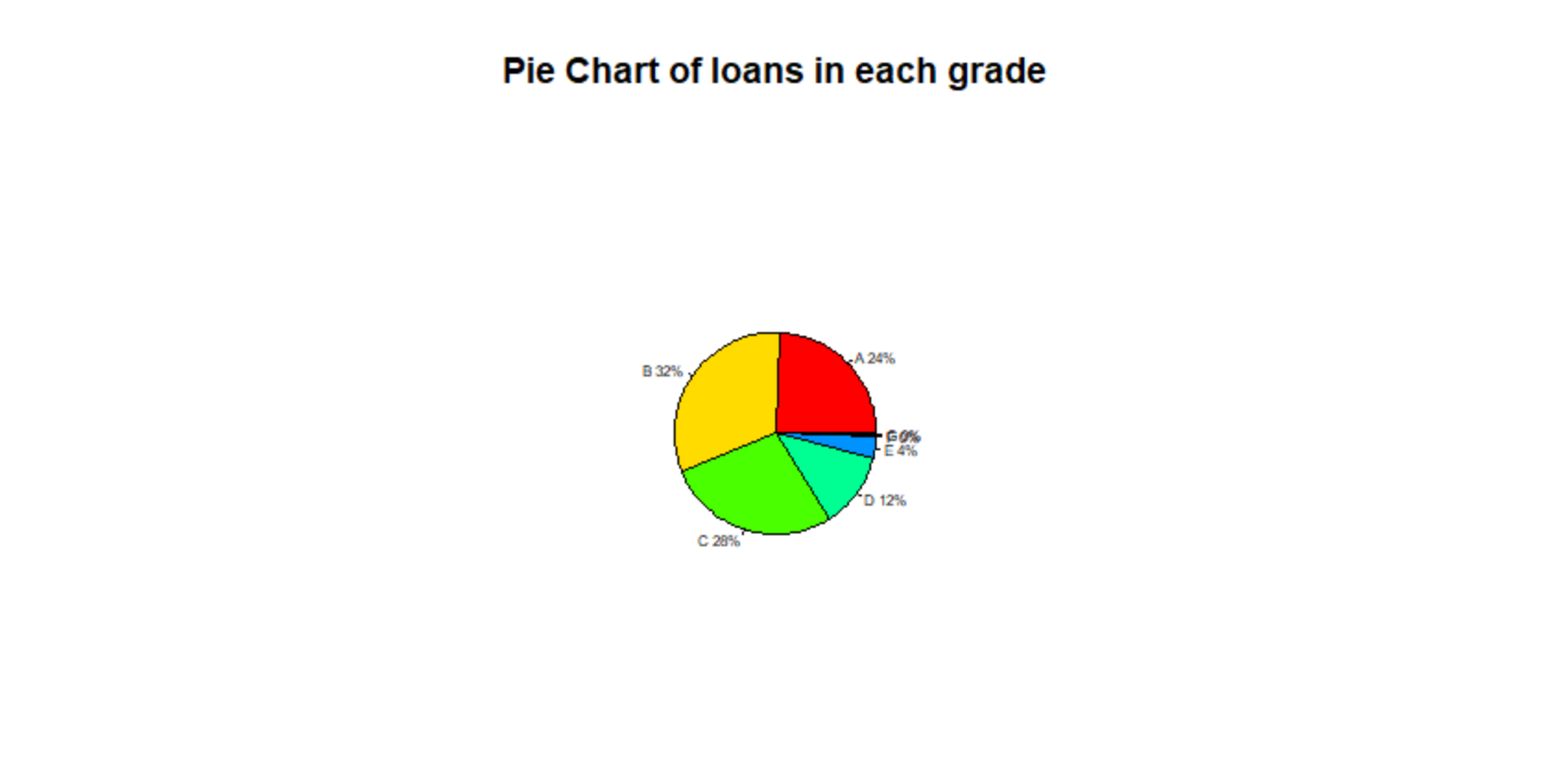
Default rate increase from A to G. This is what we expected, as the interest rate and risk of investing increases from A to G.



**Interpretation:**

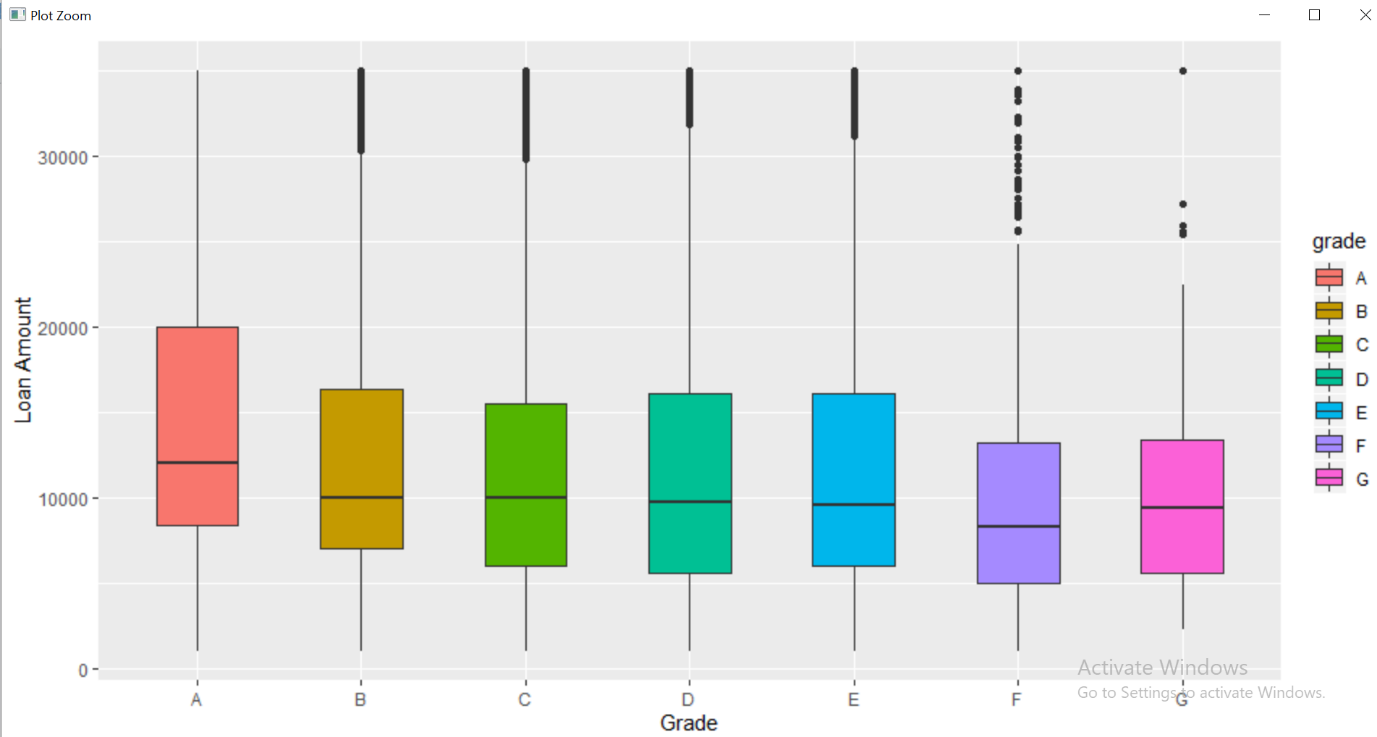
As the earlier graph shown, default rate does increase from A to G. The pattern is also similar with the Sub grades. In sub grades of A, B and C, default rate increases from 1 to 5. In D the default rate of D3 is lower than D2, in E the default rate increases from E1 to E3 and decreases from there on till E5. In F, the default rate increases from F1 to F5 with F4 not following the pattern. In G the default rate is higher for G1 and G2 when compared to G3, G4 and G5.

Apart from G, default rate pattern is exactly what we expected it would be.

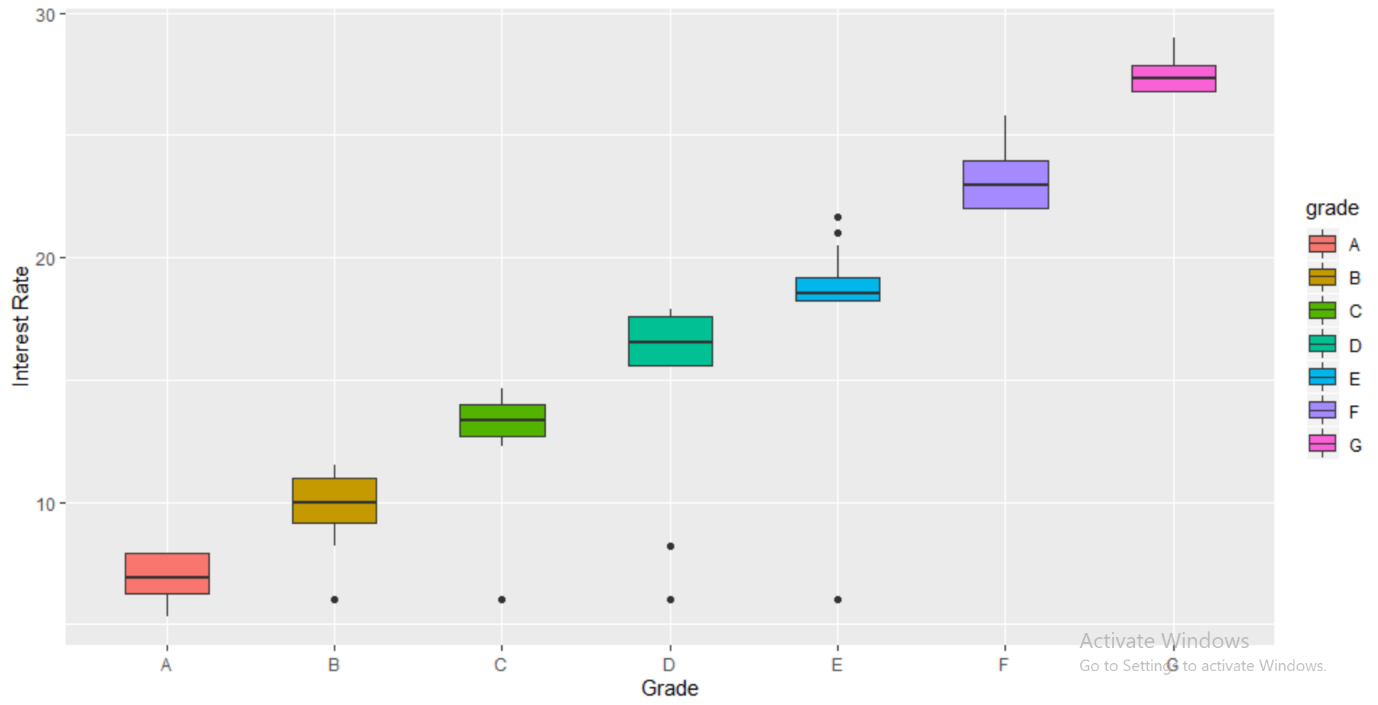
**(ii) How many loans are there in each grade? And do loan amounts vary by grade? Does interest rate for loans vary with grade, subgrade? And is this what you expect, and why?**

**Interpretation:**

Count wise grade B has the highest number of loans in our given data and G has the least.



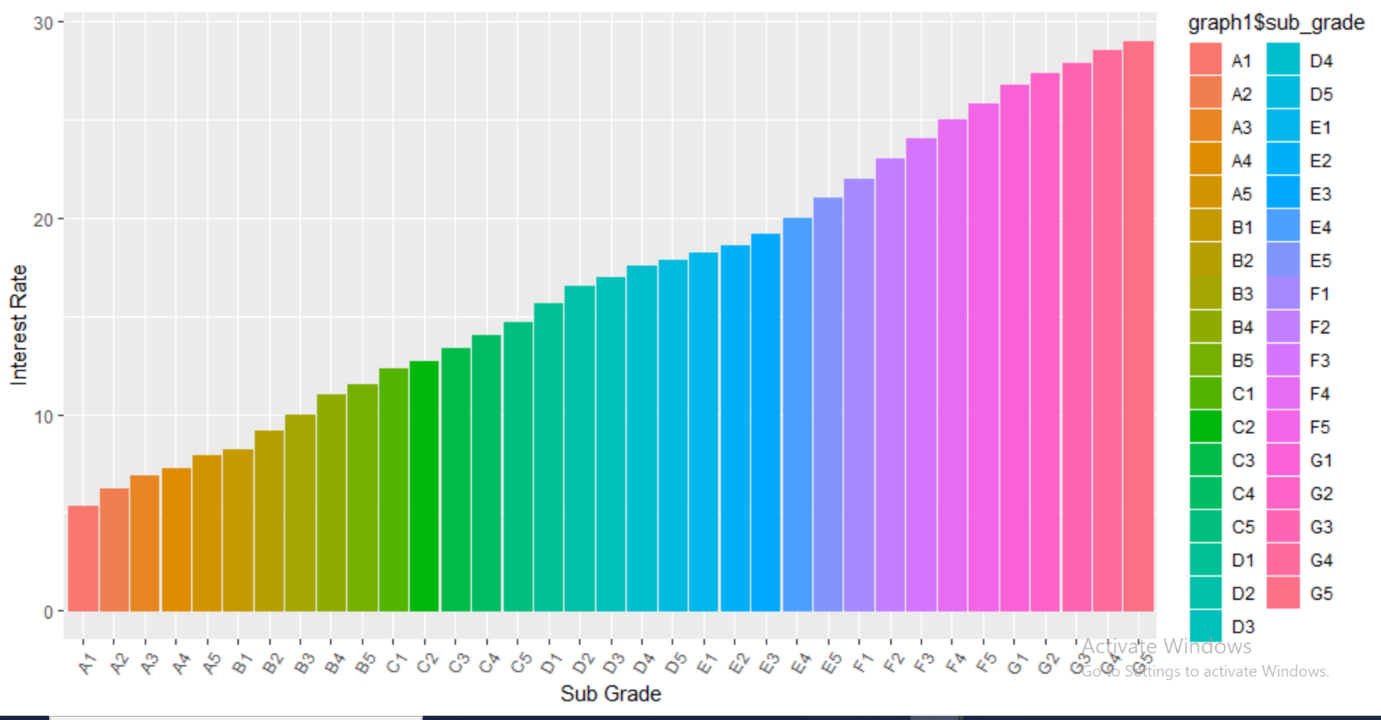
**Interpretation:**

Loan amount does vary from grade to grade. It is highest for grade A and lowest for grade F. The range of loan amount is max – 12000 and min – 8000. So, the difference between grades is not that significant.

**Interpretation:**

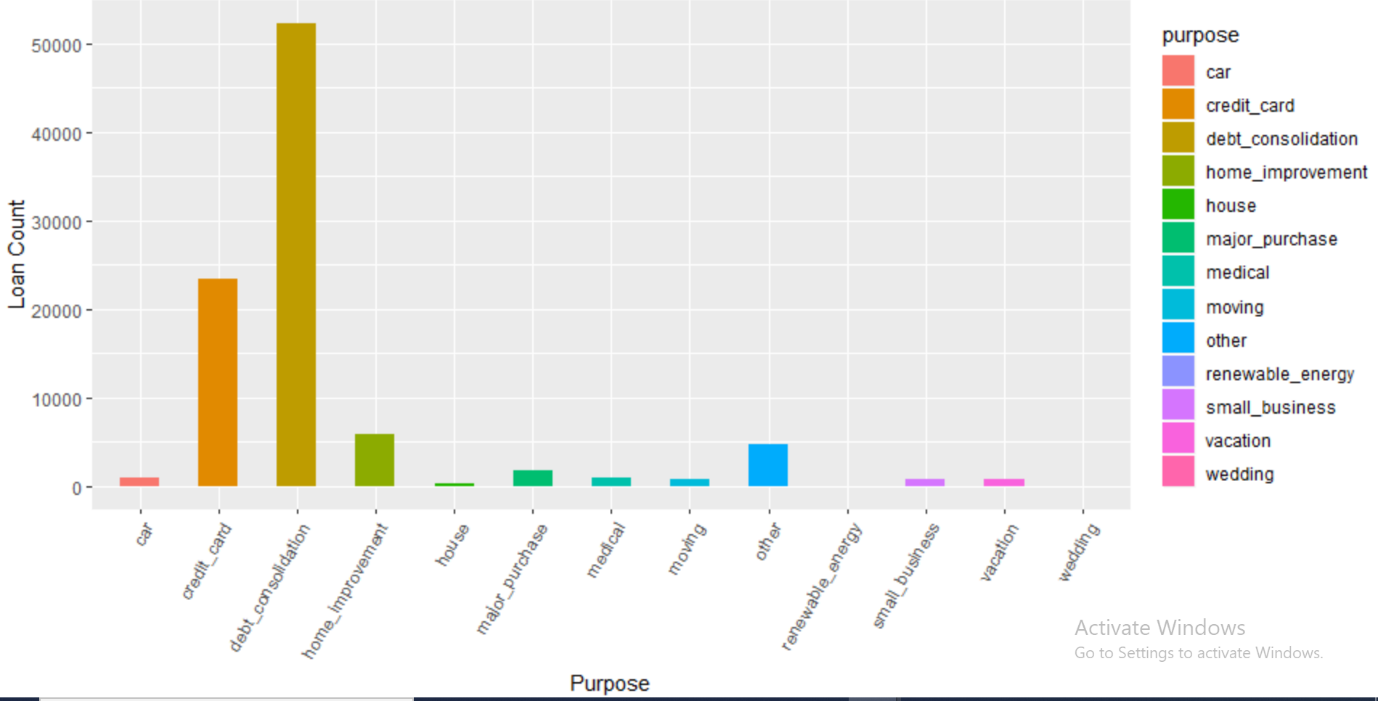
Interest rate increases from A to G. This is what we expected, as the risk of investing increases from A to G the interest rate increases.

Risk – low credit scores, low annual incomes, other debt obligations and various other parameters that define the grade of the borrower



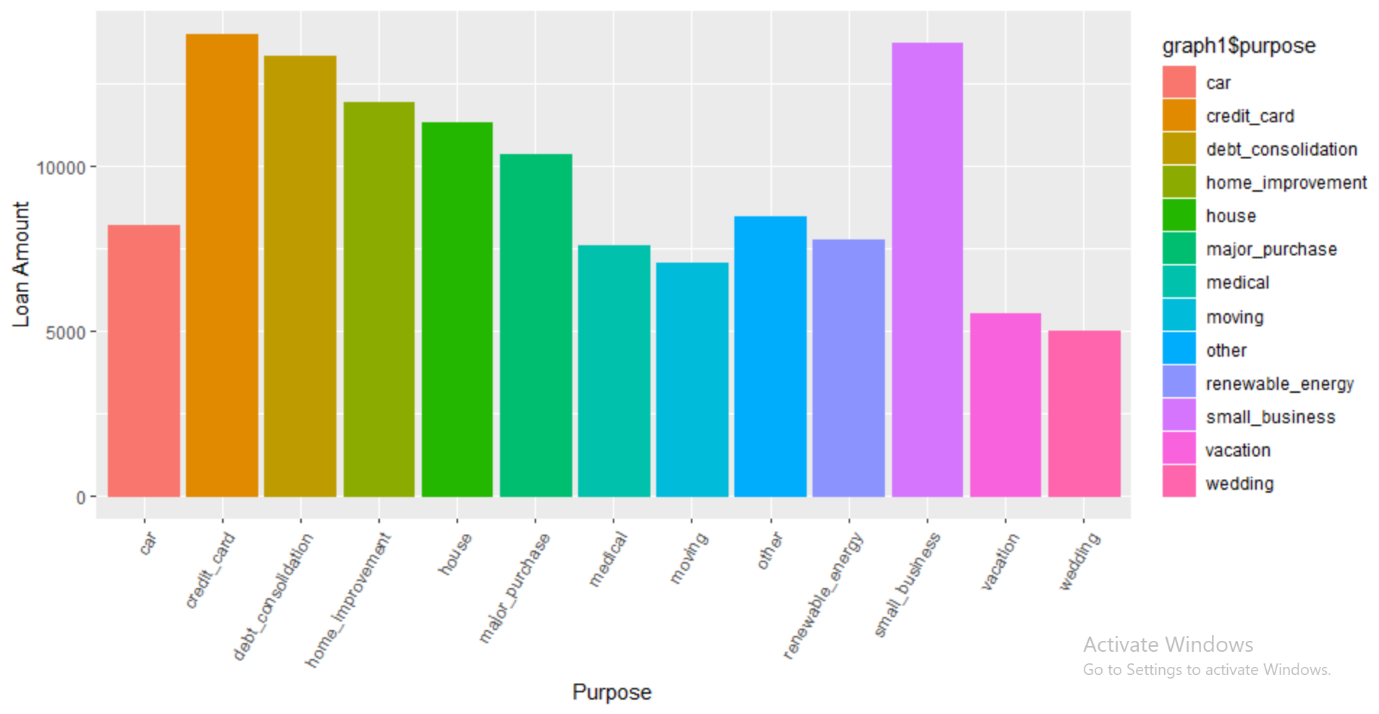
**Interpretation:**

Interest rate increases from 1 to 5 in sub grades of all the grades. This is what we expected as the risk of investment increases from 1 to 5 in sub grades.

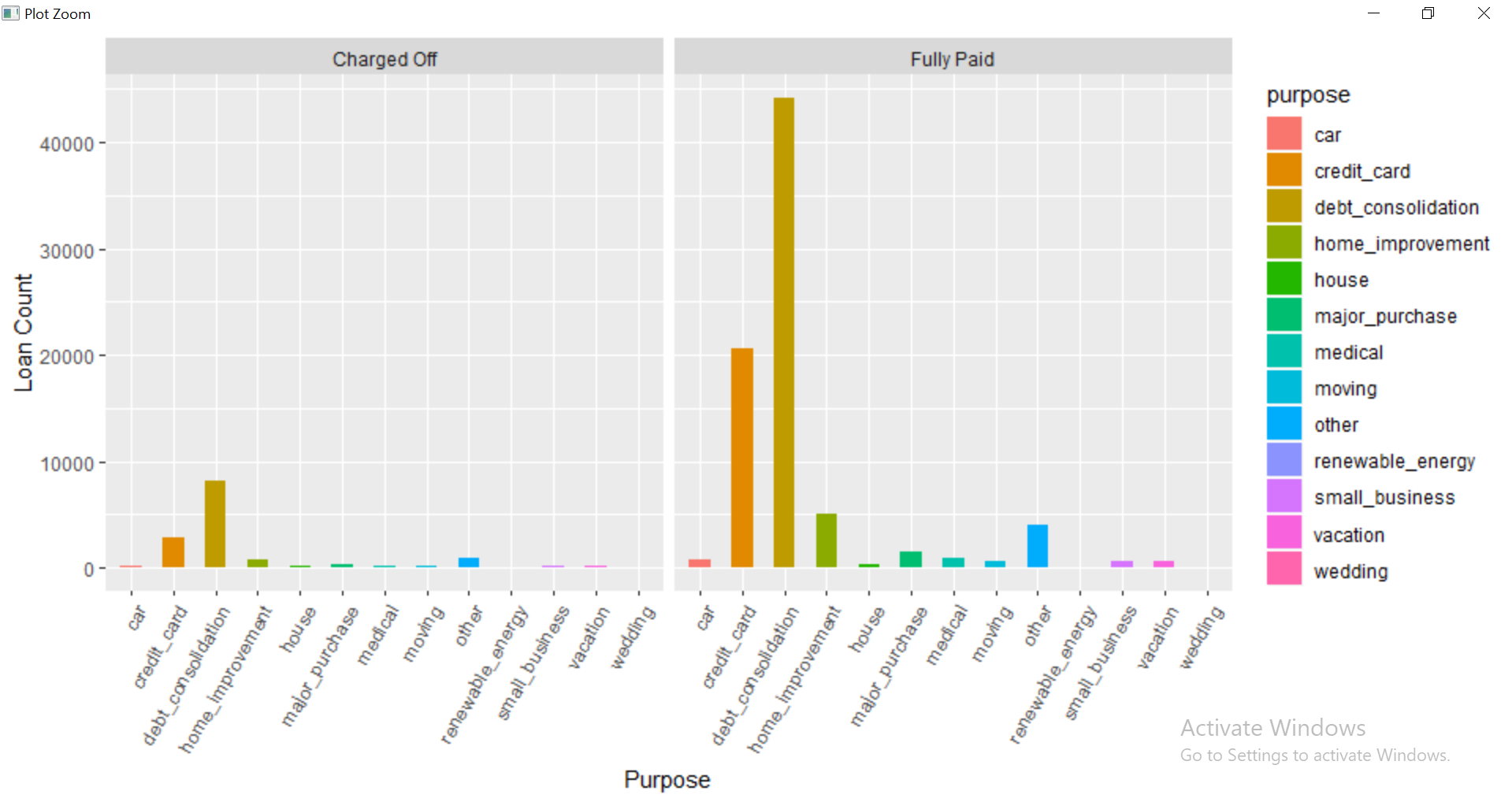
**iii) What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? And within grade? Do defaults vary by purpose?**

**Interpretation:**

As we can clearly see that the major reason people borrow loans is for **debt consolidation** and **credit card.**

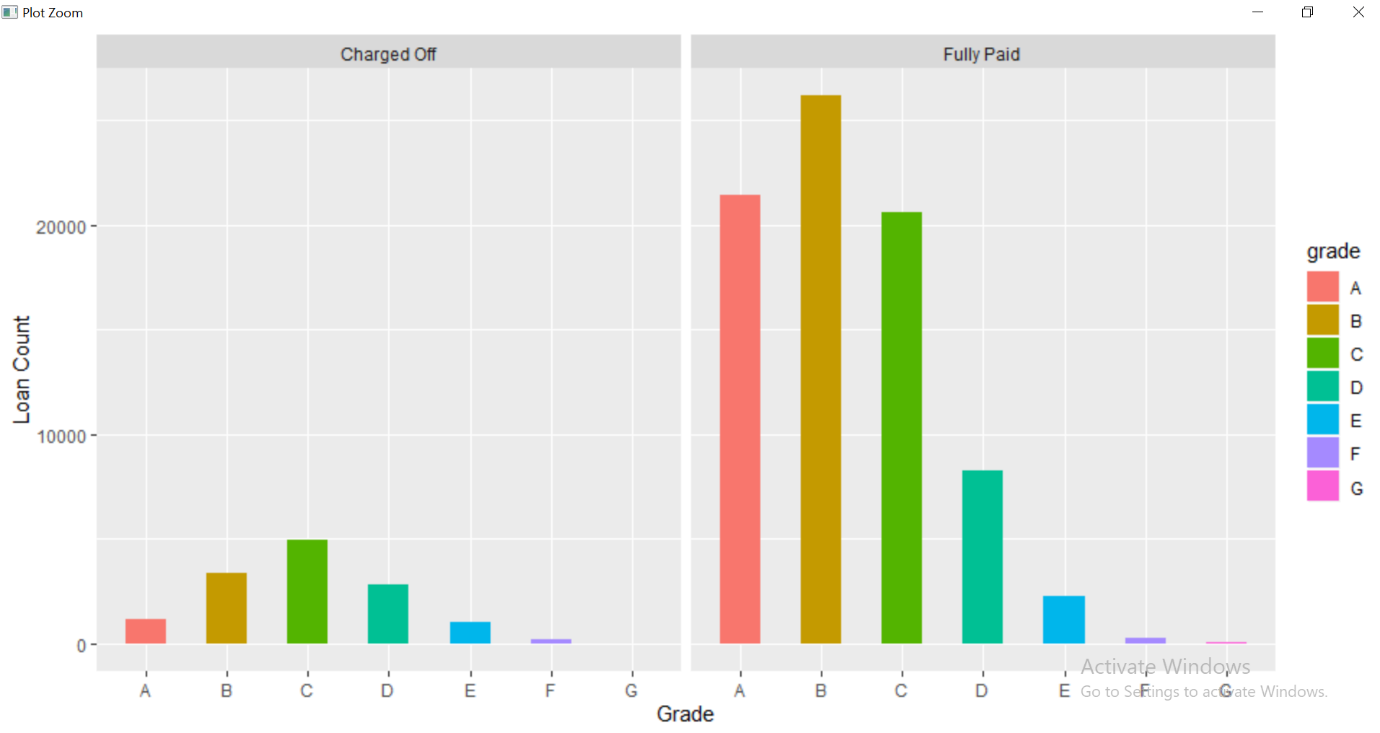
**Interpretation**:

Purpose wise Credit card, small business and debt consolidation are among the top 3 reasons that borrow highest loan amount. Wedding and vacation are the purposes with lowest loan amount. The max amount – 18000 and min amount – 5000.

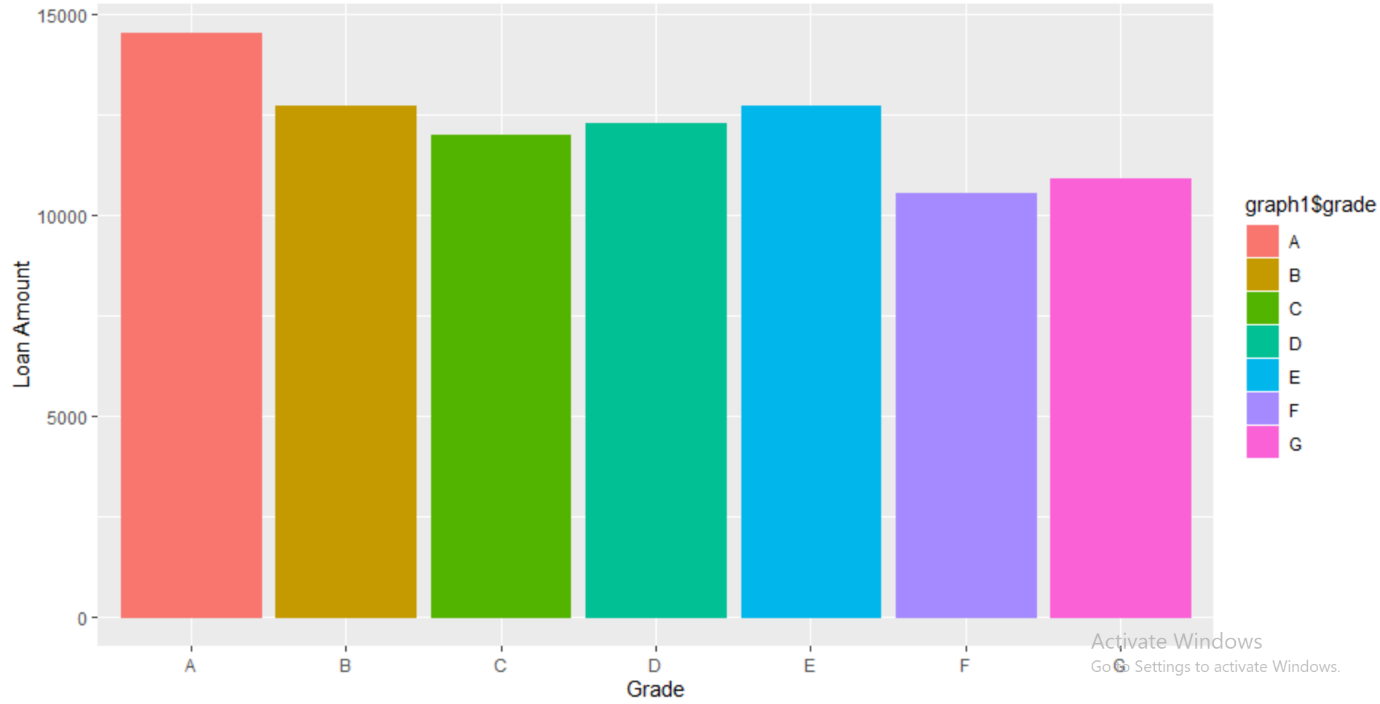


**Interpretation:**

Loan status wise charged off and fully paid is highest for debt consolidation and credit card. The reason can be, there are the reasons for which the number of loan borrowings are highest in number. Defaults doesn’t vary by purpose, the proportion of loans borrowed for a particular purpose and defaults is similar.

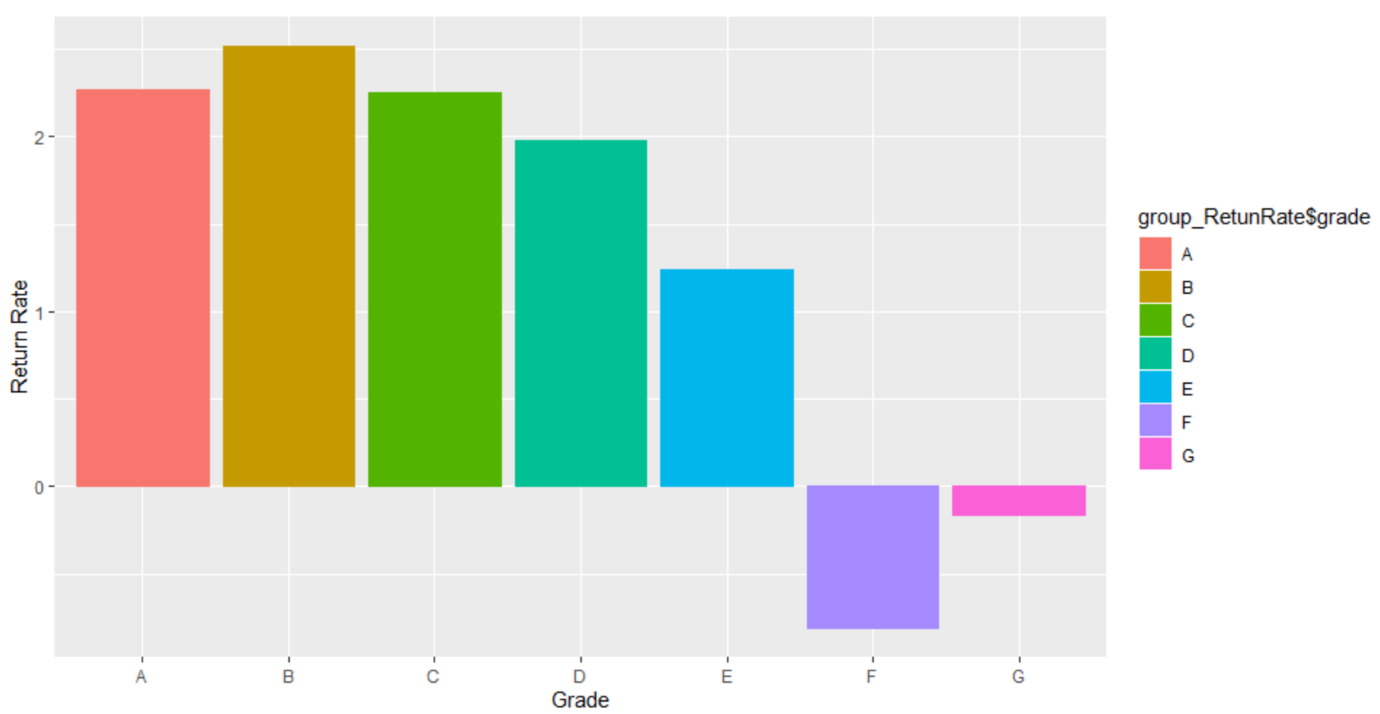


**Interpretation:**

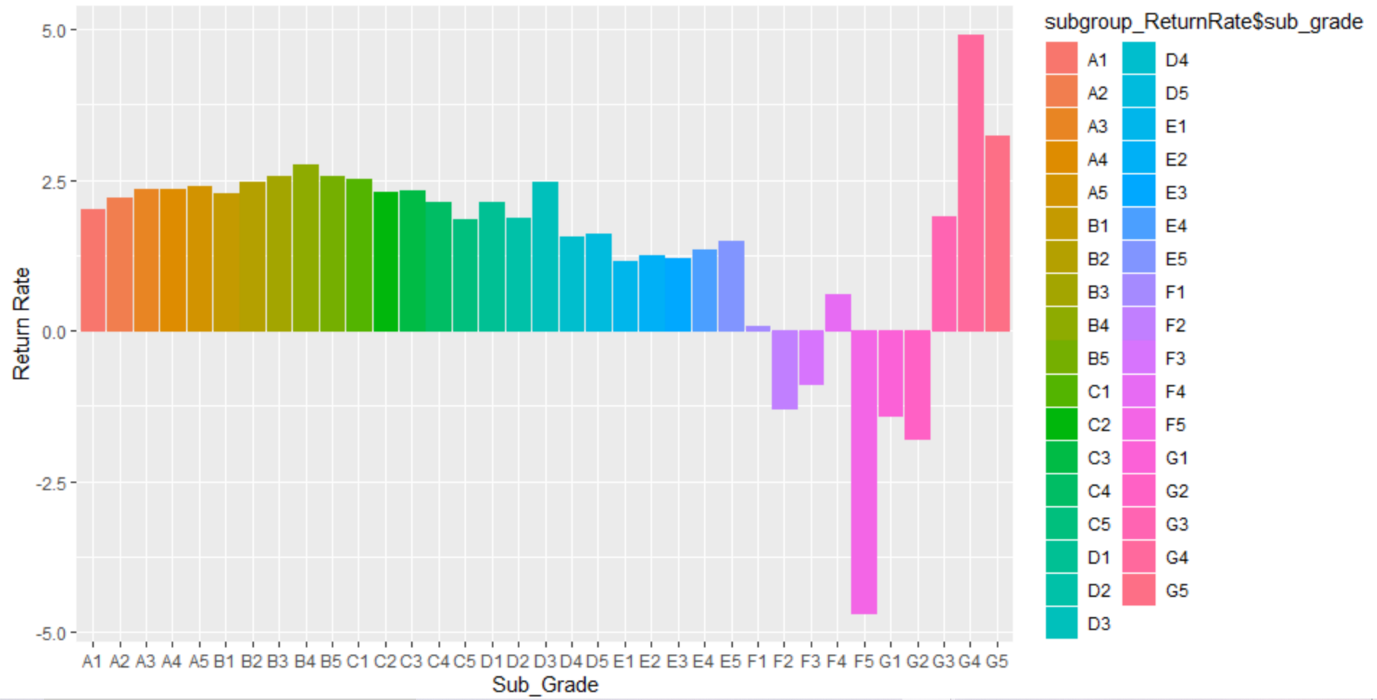
Grade wise the loan count is highest in grade C for Charged off and lowest for grade G. Grade wise the loan count is highest in grade B for fully paid and lowest for G. So, it does change by loan status.

**Interpretation:**

Loan amount does vary from grade to grade. It is highest for grade A and lowest for grade F. the maximum amount is around 14500 and lowest amount is around 11000.

**(iv) Calculate the annual return. Show how you calculate the percentage annual return. 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. 3 If you wanted to invest in loans based on this data exploration, which loans would you invest in?**

**Interpretation:**

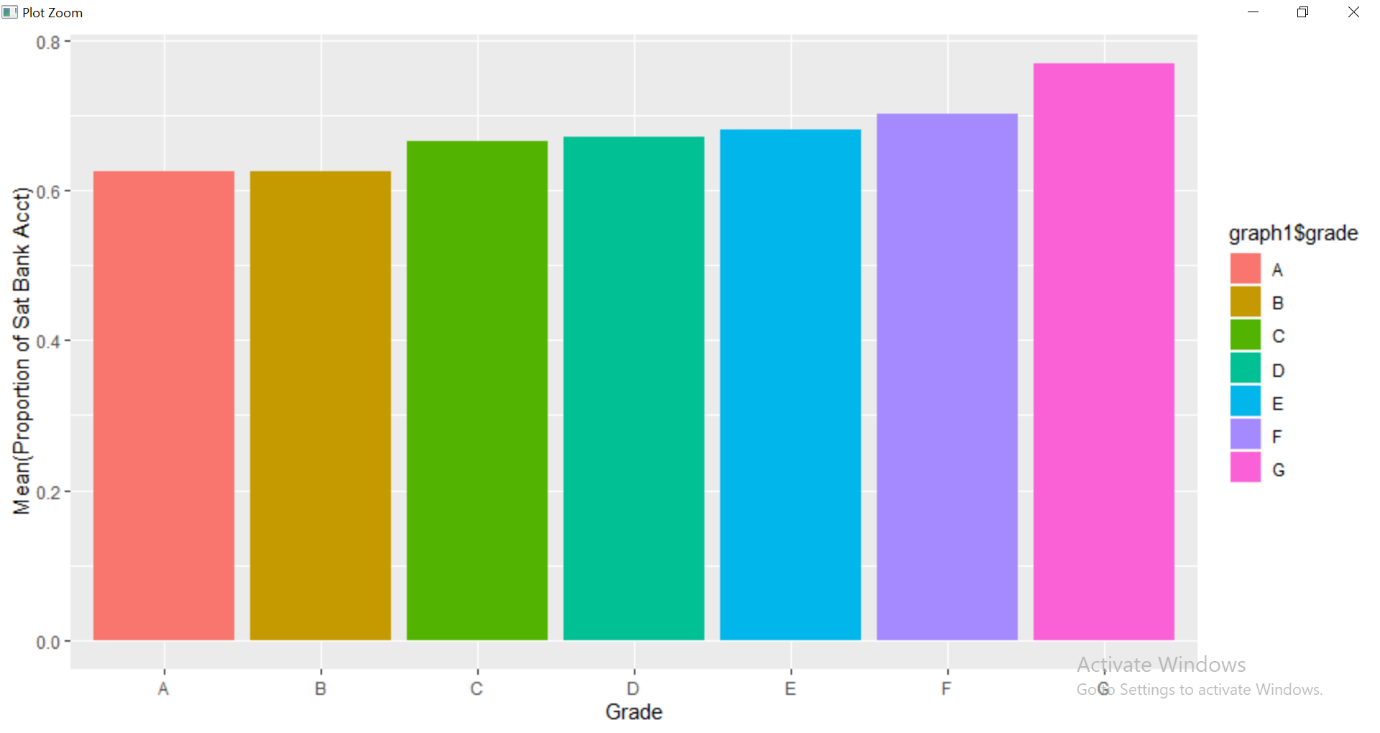
Annual Return rate is highest for grade B and lowest for grade F. average return rate is 2% and above for grades A, B, C and D. It is in negatives for F and G.

**Interpretation:**

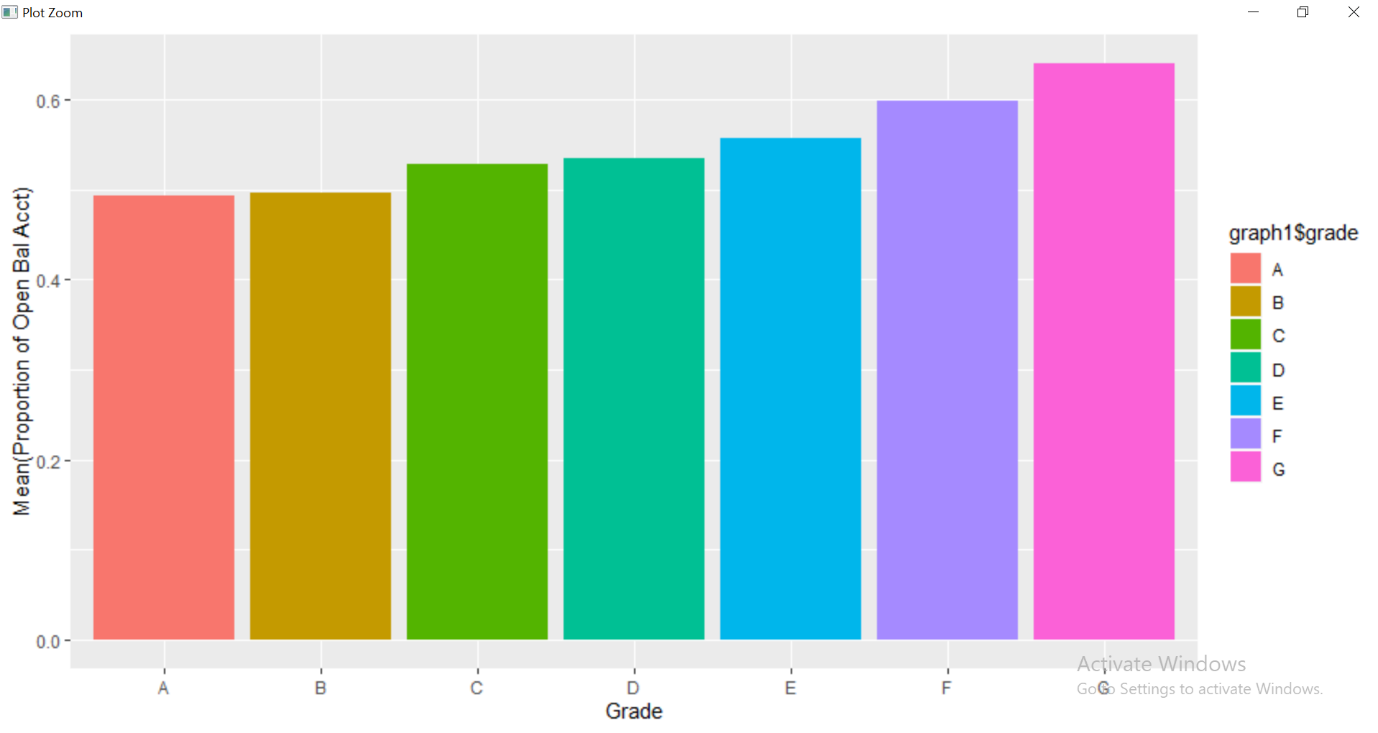
Average return rate is highest for G4 and lowest for F5 in sub grades. It is in negatives for entire F sub grades and G1, G2.

**Based on the average return, I will invest in B (grade wise) and G4(sub-grade), they have the highest return rate of 2.5% and 4.9% respectively.**

**(v) Generate some new derived attributes which you think may be useful for predicting default. and explain what these are.**

**Interpretation:**

Number of bankcard accounts and Number of satisfactory bankcard accounts are used to create an attribute proportion of satisfactory bankcard accounts. It is created by number of satisfactory bankcard accounts/number of bankcard accounts.

**Interpretation:**

The total number of credit lines currently in the borrower's credit file and number of open credit lines in the borrower's credit file are used to create proportion of open credit lines. It is created by number of open credit lines/number of credit lines.

1. **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). For example, it has been noted that the FICO scores on loan applicants are updated periodically, and the data can carry thus FICO scores from after the loan issue date. So, even though FICO score can be useful, the values in the data may not be usable. Identify and explain which variables you will exclude from the model.**
2. First, we start with removing all the attributes which have MORE than 60% NA values.
3. Data Leakage and Irrelevant Data – Below are the list of attributes which either cause data leakage or is irrelevant to the prediction of defaults.

|  |  |
| --- | --- |
| **Attributes Removed** | **Reason** |
| addr\_state | Irrelevant to the prediction model |
| collection\_recovery\_fee | changes if any payment is done |
| earliest\_cr\_line | Irrelevant to the credit worthiness |
| emp\_title | Irrelevant to the credit worthiness |
| fico\_range\_high | Causes data leakage |
| fico\_range\_low | Causes data leakage |
| funded\_amnt\_inv | Causes data leakage |
| initial\_list\_status | Irrelevant to the credit worthiness |
| inq\_last\_6mths | can change if there are any additional inquiries made over time |
| issue\_d | Irrelevant to the credit worthiness |
| last\_credit\_pull\_d | changes over time, does not help in default prediction |
| last\_fico\_range\_high | Causes data leakage |
| last\_fico\_range\_low | Causes data leakage |
| last\_pymnt\_amnt | Causes data leakage |
| last\_pymnt\_d | Causes data leakage |
| mo\_sin\_old\_il\_acct | changes over time, does not help in default prediction |
| mo\_sin\_old\_rev\_tl\_op | changes over time, does not help in default prediction |
| num\_bc\_sats | changes over time, does not help in default prediction |
| num\_bc\_tl | changes over time, does not help in default prediction |
| num\_tl\_op\_past\_12m | changes over time, does not help in default prediction |
| open\_acc | changes over time, does not help in default prediction |
| percent\_bc\_gt\_75 | changes over time, does not help in default prediction |
| propOpenAcct | changes over time, does not help in default prediction |
| propSatBC | changes over time, does not help in default prediction |
| recoveries | changes over time, does not help in default prediction |
| returnRateAnnual | changes over time, does not help in default prediction |
| revol\_bal | varies depending on borrowers payments to credit lines changes over time, does not help in default prediction |
| revol\_util | changes depending on borrowers expenses changes over time, does not help in default prediction |
| total\_acc | changes over time, does not help in default prediction |
| total\_pymnt | changes over time, does not help in default prediction |
| total\_pymnt\_inv | changes over time, does not help in default prediction |
| total\_rec\_int | changes over time, does not help in default prediction |
| total\_rec\_prncp | changes over time, does not help in default prediction |
| zip\_code | Irrelevant to the credit worthiness |
| term | an option to increase term depending on situation (36 to 60) |
| Pymnt\_plan | Varies from person to person |
| out\_prncp | Leaks data over time |
| out\_prncp\_inv | Leaks data over time |
| total\_rec\_late\_fee | Changes if any payment is made |
| collections\_12\_mths\_ex\_med | Changes over time |
| policy\_code | Leaks data from future |
| application\_type | Changes over time |
| acc\_now\_delinq | Changes over time |
| tot\_coll\_amt | Changes if any new account is opened |
| chargeoff\_within\_12\_mths | Changes if any payment is made towards charge off |
| delinq\_amnt | Changes over time |
| num\_tl\_120dpd\_2m | Changes from time to time |
| num\_tl\_30dpd | Changes from time to time |
| num\_tl\_90g\_dpd\_24m | Changes from time to time |
| pct\_tl\_nvr\_dlq | Leaks data from future |
| tax\_liens | Changes if any tax payments is done |
| hardship\_flag | Leaks data from future |
| debt\_settlement\_flag | Changes over time |

Non Zero Variance

An interesting aspect of this dataset is that many variables have extremely low variances. This means that there is very little information in these variables because they mostly consist of a single value(Appendix No.1). In general, near-zero variance vectors are discarded because they do not provide much information for the learning algorithm. However, that's not to say you should always throw away all data with near-zero variance (Appendix No.2).

To identify these types of predictors, the following two metrics can be calculated:

* the frequency of the most prevalent value over the second most frequent value (called the “frequency ratio’’), which would be near one for well-behaved predictors and very large for highly-unbalanced data and
* The “percent of unique values’’ is the number of unique values divided by the total number of samples (times 100) that approaches zero as the granularity of the data increases

If the frequency ratio is greater than a pre-specified threshold and the unique value percentage is less than a threshold, we might consider a predictor to be near zero-variance.(Appendix No.4)

Below are the attributes we received from Non Zero Variance function [nearZeroVar(NonNaData)] in package “caret” in R–

[1] "term"

[2] "pymnt\_plan"

[3] " out\_prncp "

[4] "out\_prncp\_inv"

[5] " total\_rec\_late\_fee "

[6] "collections\_12\_mths\_ex\_med"

[7] "policy\_code"

[8] " application\_type "

[9] "acc\_now\_delinq"

[10] "tot\_coll\_amt"

[11] "chargeoff\_within\_12\_mths"

[12] "delinq\_amnt"

[13] "num\_tl\_120dpd\_2m"

[14] "num\_tl\_30dpd"

[15] "num\_tl\_90g\_dpd\_24m"

[16] "pct\_tl\_nvr\_dlq"

[17] "tax\_liens"

[18] "hardship\_flag"

[19] "debt\_settlement\_flag"

In general, near-zero variance vectors are discarded because they do not provide much information for the learning algorithm. However, that's not to say you should always throw away all data with near-zero variance (Appendix No.2).

However, from the list above we retain “acc\_now\_delinq” and “delinq\_amnt” as these variables might be helpful in the prediction model.

4. Develop decision tree models to predict default.

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

We have considered the proportion of 70:30 and 50:50 to have a good number of models to choose from. The preferred proportion is 70% for the training set and 30% for the test. The idea is that more training data is a good thing because it makes the classification model better.

Also, the prediction model performed better with 70:30 than for 50:50.

(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?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Training and Test Dataset Split** | **CP Value** | **Min.Split** | **Accuracy- Training Dataset** | **Accuracy - Test Dataset** | **Precision - Train Dataset** | **Precision - Test Dataset** | **AUC** |
| Entropy | 50:50:00 | 0.0004735 | 20 | 85.63% | 85.12% | 85.69% | 85.45% | 65.68% |
| "Information" |
| Gini | 50:50:00 | 0.0004735 | 20 | 85.68% | 85.23% | 85.81% | 85.50% | 64.53% |
| Entropy | 70:30:00 | 0.0004735 | 20 | 85.57% | 84.92% | 85.67% | 84.86% | 65.91% |
| "Information" |
| Gini | 70:30:00 | 0.0004735 | 20 | 85.66% | 84.89% | 73.21% | 71.21% | 64.95% |

The following parameters have been changed and experimented with-

1. Model Type – Gini and Information
2. Training and Test Dataset Split (50/50 and 70/30)

The following parameters will help us evaluate as to which is the best prediction model:

1. Accuracy
2. Precision
3. AUC

(c) Identify the best tree model. Why do you consider it best? Describe this model – in terms of complexity (size). Examine variable importance. Briefly describe how variable importance is obtained in your best model.

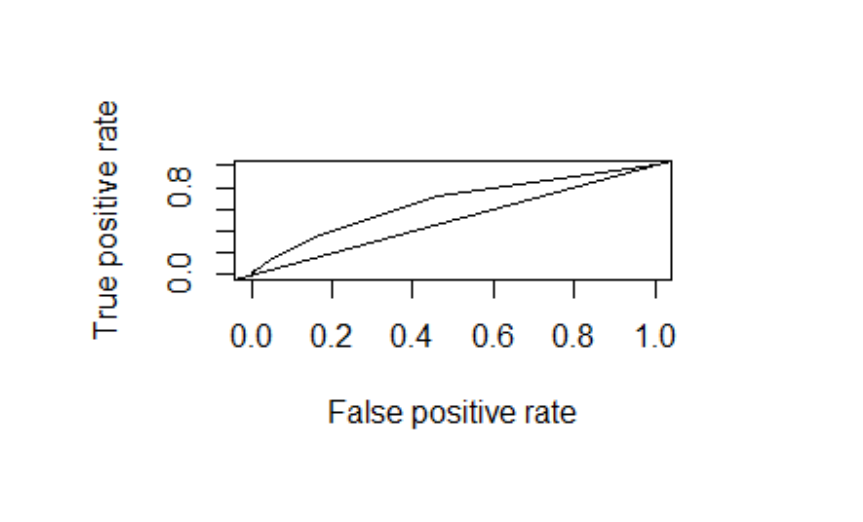
The THIRD decision tree gives us the Accuracy of 85.57%, Precision of 85.67% and AUC of 65.91%, which is the best model. Even though the difference in the accuracy of the 1st model is lesser than the 3rd, the AUC value for the 3rd is better, which makes the 3rd model the best out of the four models.

The number of variables used to make the decision tree is 46. The variable importance can be determined data.frame$variable.importance. The important variables were -

1. annual\_inc
2. avg\_cur\_bal
3. bc\_open\_to\_buy
4. dti
5. int\_rate
6. mo\_sin\_rcnt\_rev\_tl\_op
7. mort\_acc

ROC Curve-

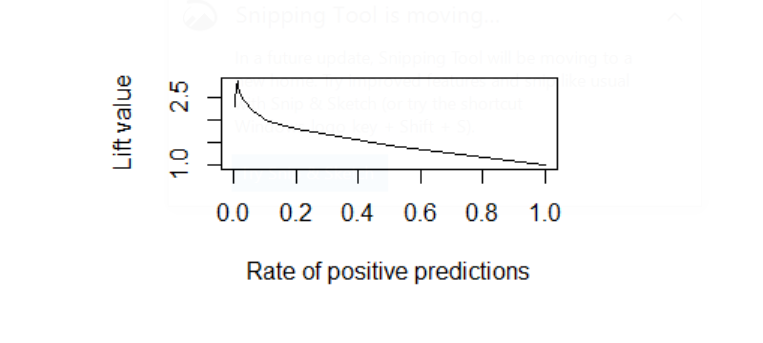
The AUC (Area under the Curve) is 65.91%. A model is considered good if the AUC is between 0.6 and 1. Model 3 has a fair AUC value.



Lift Graph –

**Lift** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. The greater the area between the lift curve and the baseline, the better the model.

The lift curve output for Model 3 shows that the model is good.



**Appendix-**

1. <https://campus.datacamp.com/courses/machine-learning-toolbox/preprocessing-your-data?ex=13>
2. <https://www.kaggle.com/c/digit-recognizer/discussion/38768>
3. <https://rdrr.io/cran/caret/man/nearZeroVar.html>
4. <https://topepo.github.io/caret/pre-processing.html#nzv>
5. <http://information-gain.blogspot.com/2012/07/why-split-data-in-ratio-7030.html>
6. <http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html>