Lending Club case study

(Analysing variables which are inducing charged off(default) from historical data)

Business understanding

- The Lending club company is a financial lending company which forms a bridge between borrower and lenders by it's online platform and gives loan at a lower interest rate.
- Like Any other finance lending company it does not want to have loss due to defaults.
- **What is default ?**
- If a borrower does not pay it's due or term against the loan , or runs away with the sum he has taken then he/she considered as defaulter , in such case financial lender suffers loss.
- **What is the action we can take about the applicants who are likely to default ?**
- We can take some measures against such applicants : 1 such as denying for loan 2 giving loan in a higher interest 3 OR decrease the amount of loan for risky customers
- **The problem statement.**
- The problem statement here is to find out the important variables or attributes of customers and the loan which are inducing the default , which in turn is causing credit loss to the lending company (.i.e. Lending Club) , which in future will direct the decision of the investor whether to invest the loan amount on a particular borrower or not , and there by cutting the financial loss of LC.

What we have done to get the significant variables which induces default through EDA ?

- **Cleaned the data set**
- **Described the variables in business terms and further removed the attributes which are not important to risk analysis**
- **Removed outliers**
- **Did univariate analysis to see further distribution of data across population**
- **Did bivariate analysis to describe the relationship between the *target variable* and *important attributes***
- **Did multivariate analysis to describe the relationship between the *target variable* and *important attributes***
- **Gave summery and recommendation**

Cleaning the dataset.

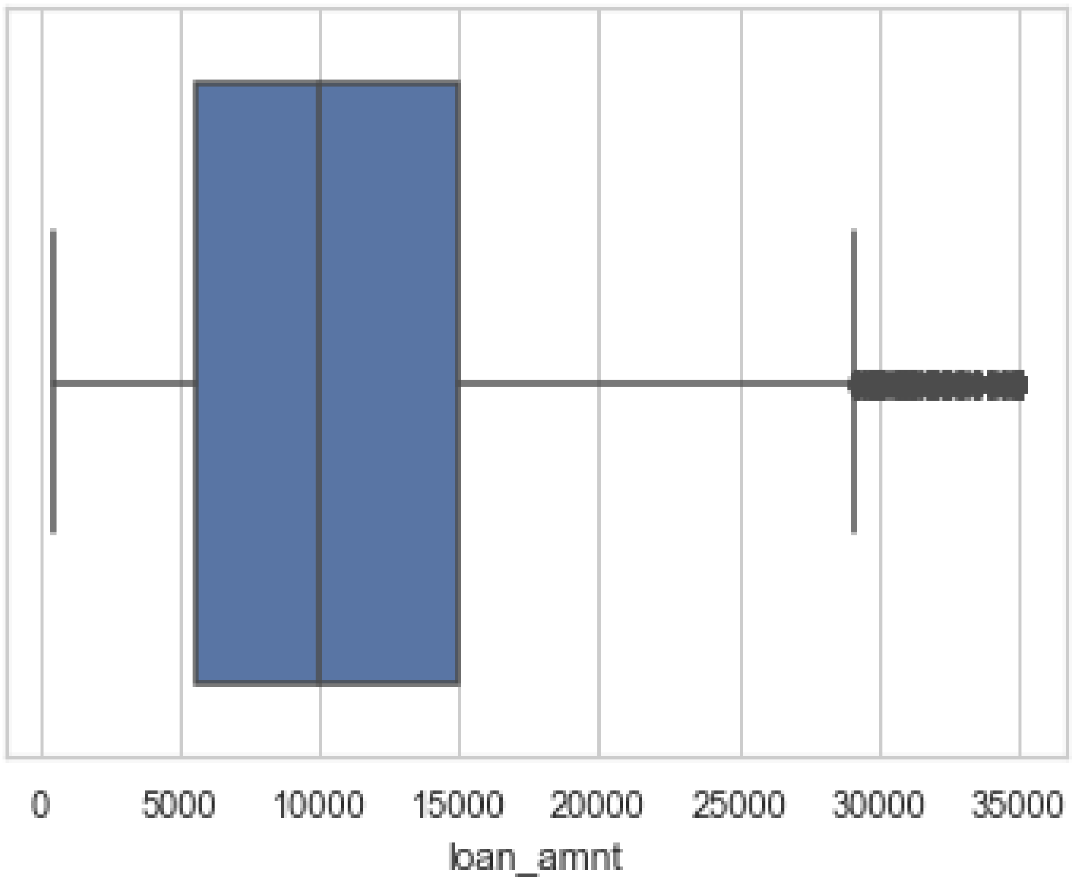
- We have cleaned the data set by dropping the columns , of which ,most of the entries were null.
- We have removed a significant amount of columns which were not contributing to our analysis , which reduced the analysis task easier ,and clean.
- We have also corrected the incorrect datatypes such as objects which should to be date and objects which should be float.

Described the variables in business terms and further removed the attributes which are not important to risk analysis

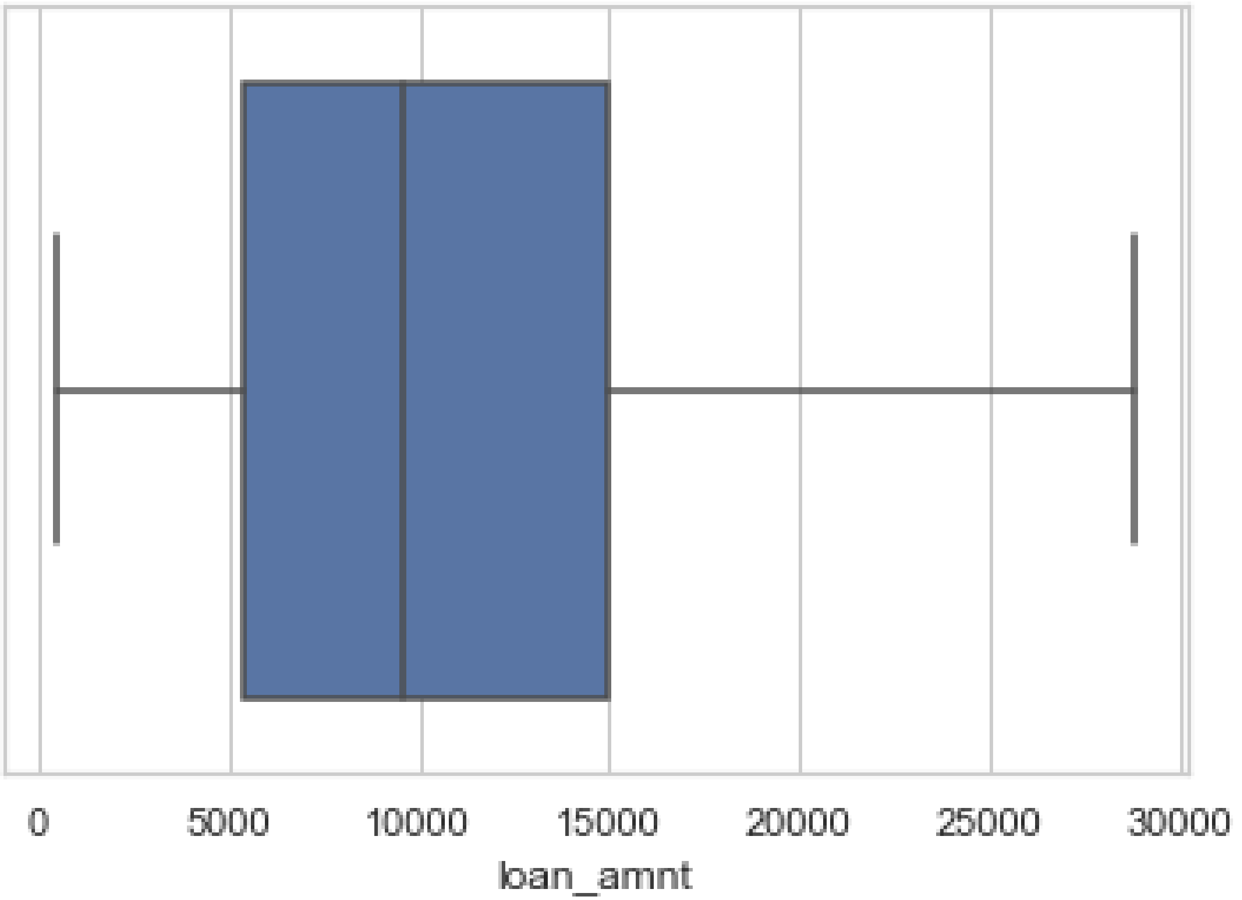
- We have described all the variables and understood the importance by researching , which helped us in identifying important attributes .
- This resulted further removal of unimportant columns and made our data set clean .
- This step also gave us the idea about our target variable which here is ,loan status (Whether the borrower charged off , fully paid or is a currently paying the EMIs)

Removed outliers

- We have used box plots to identifying outliers and then for this analysis we have removed them as they might induce biased business decision .
- For example there were some outliers in the loan amount which might induce the biased ness in our further analysis when we had to check , say ,relation between loan amount and defaults. Which hence needed to be removed .



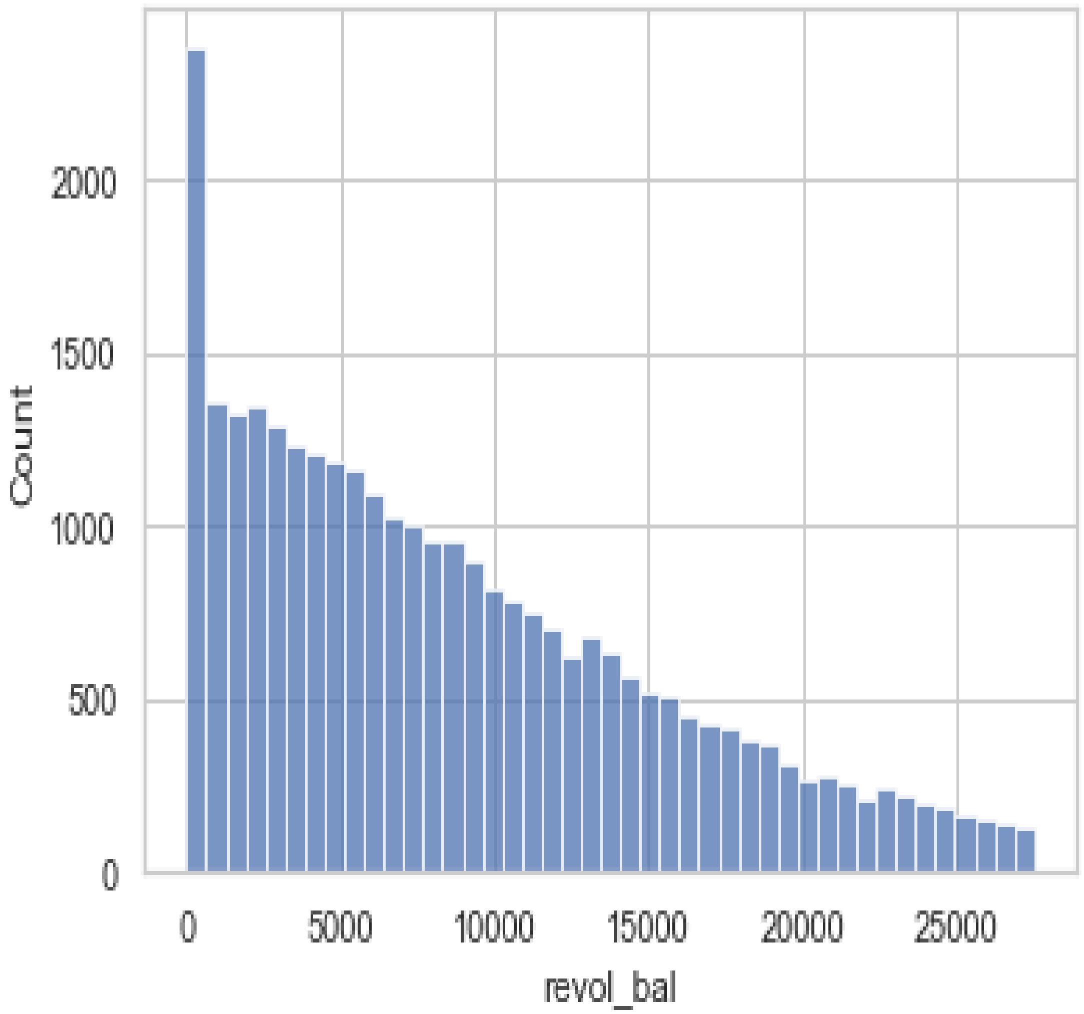
- Hence we have removed the outliers which is basically the population which are lying outside of the higher fence of the boxplot or who are taking more than ~29,000$ loan amount . Which again is very less as ~3.21% from our total population.
- And the result we got is this which removed all the outliers from our loan amount column .



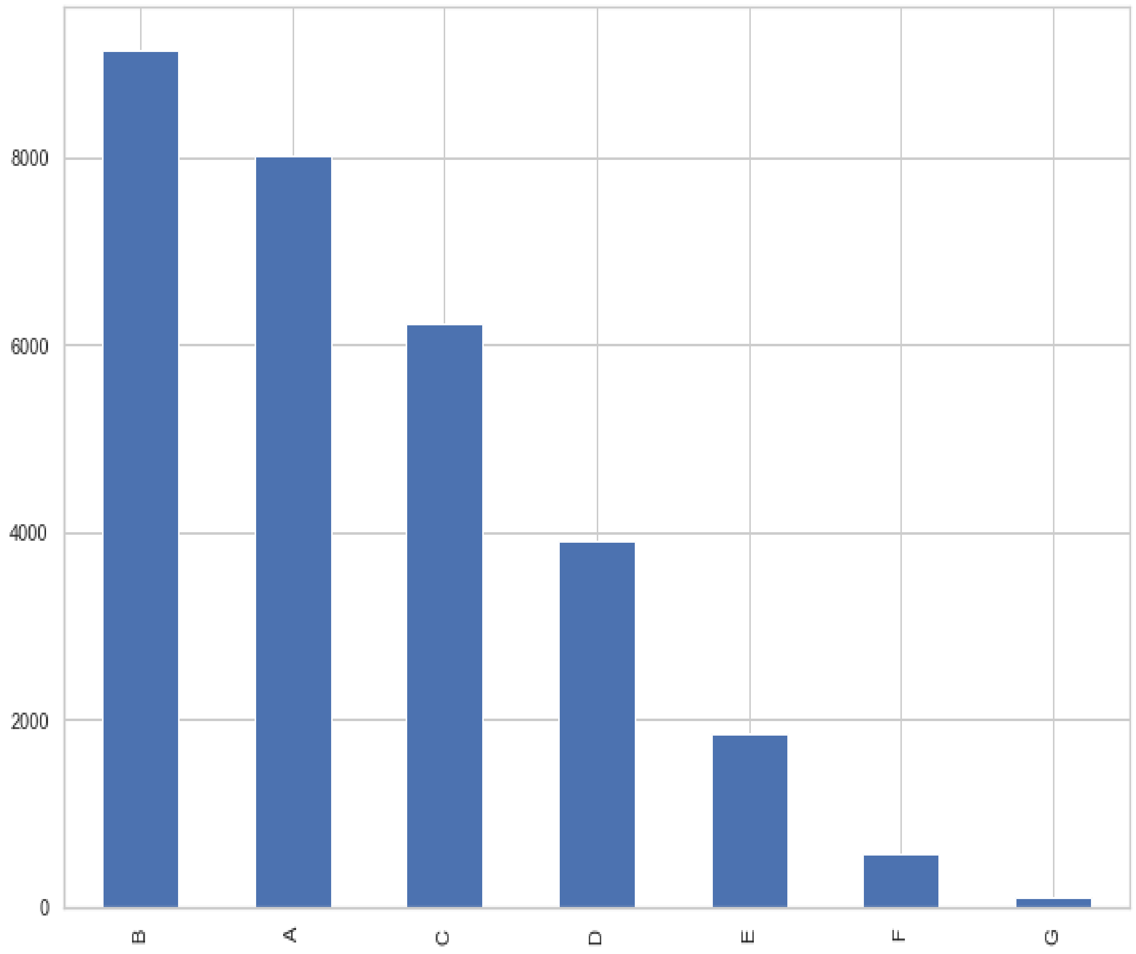
- Similarly we have also removed outliers from all our important variables like -:
- Interest rate -:interest rate on the loan (Removed interest rate greater than ~22)
- Instalments -: removed instalment greater than ~730$.
- Annual income-:removed annual income greater than or equals to ~1,27,000$.
- Open account -:Removed number of open account greater than equals to ~18
- Revolving balance -:Removed revolving balance greater than equals to ~27,500.

Did univariate analysis to see further distribution of data across population

- We have tried to understand the spread of data in all our important loan attributes as well as borrower attributes .
- For numerical variables we have used histograms to define the spread where as for categorical variables we have used bar plot or pie plot to define the spread.
- For example -: for revolving balance (the amount of unpaid credit amount which is getting added to the next months EMI or credit amount) we have the following histogram
- Which clearly defines a diminishing trend
- As most of the population are not using higher revolving balance.



- And for describing spread of categorical variables such as grade which are assigned by the LC to estimate the risk factor , calculated based on the 3rd party bureau data ,we have taken bar chart .
- As follows-: where x axis represents grades and y axis represents the amount of population under that particular grade.
- Also this figure shows us there are most of the population under risk category A and eventually the population is diminishing towards the higher risk categories.



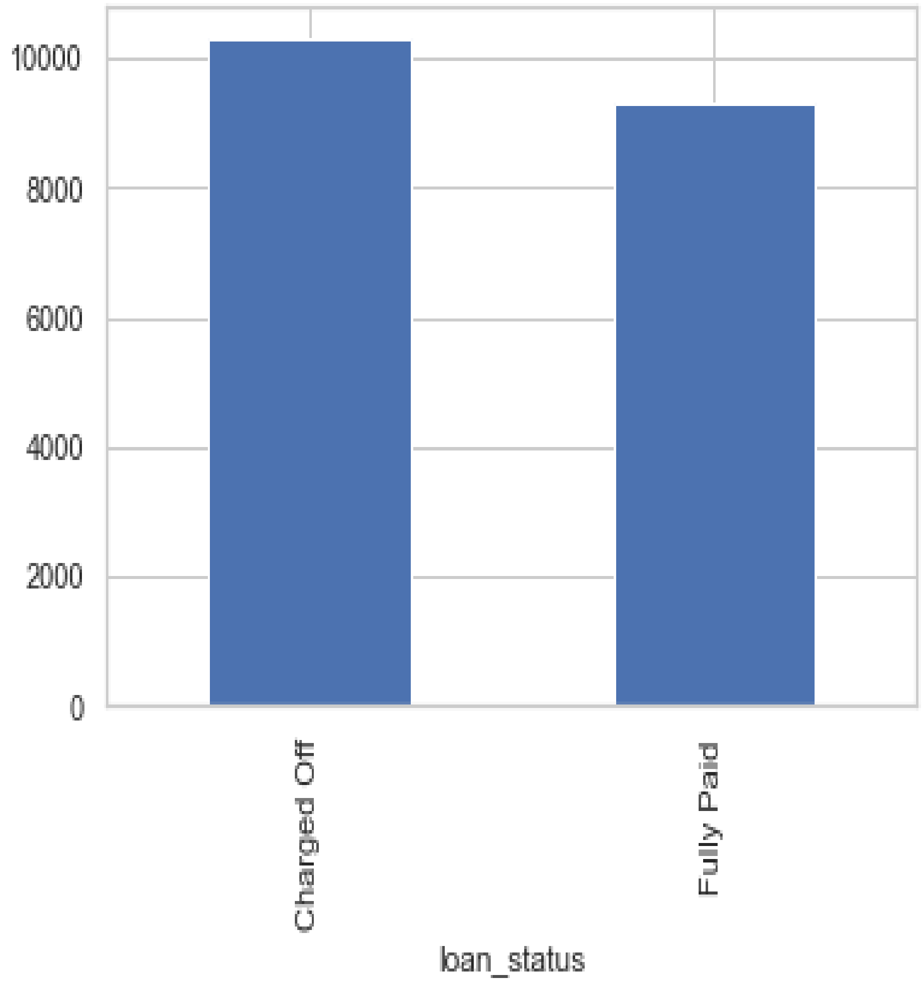
Did bivariate analysis to describe the relationship between the ***target variable*** and ***important attributes***

- Basically to avoid unnecessary analysis we have only tried to describe the relationship between the targeted variable , loan status and other numerical and categorical important borrower and loan attributes.
- And we found out some interesting findings.
- **More higher loan amount might be a causation of charged off.**
- **Defaulters are having less annual income than the fully paid Borrowers.**
- **DTI might be a risk contributing factor.**
- **Revolving balance is positively contributing to the charged off population**
- **Revolving balance utilization is significantly contributing to the charged off population**
- **Grade of the applicant is a major risk contributing factor**
- **Term is also serving as an important indicator of risk .**

Let's see the findings one by one ..

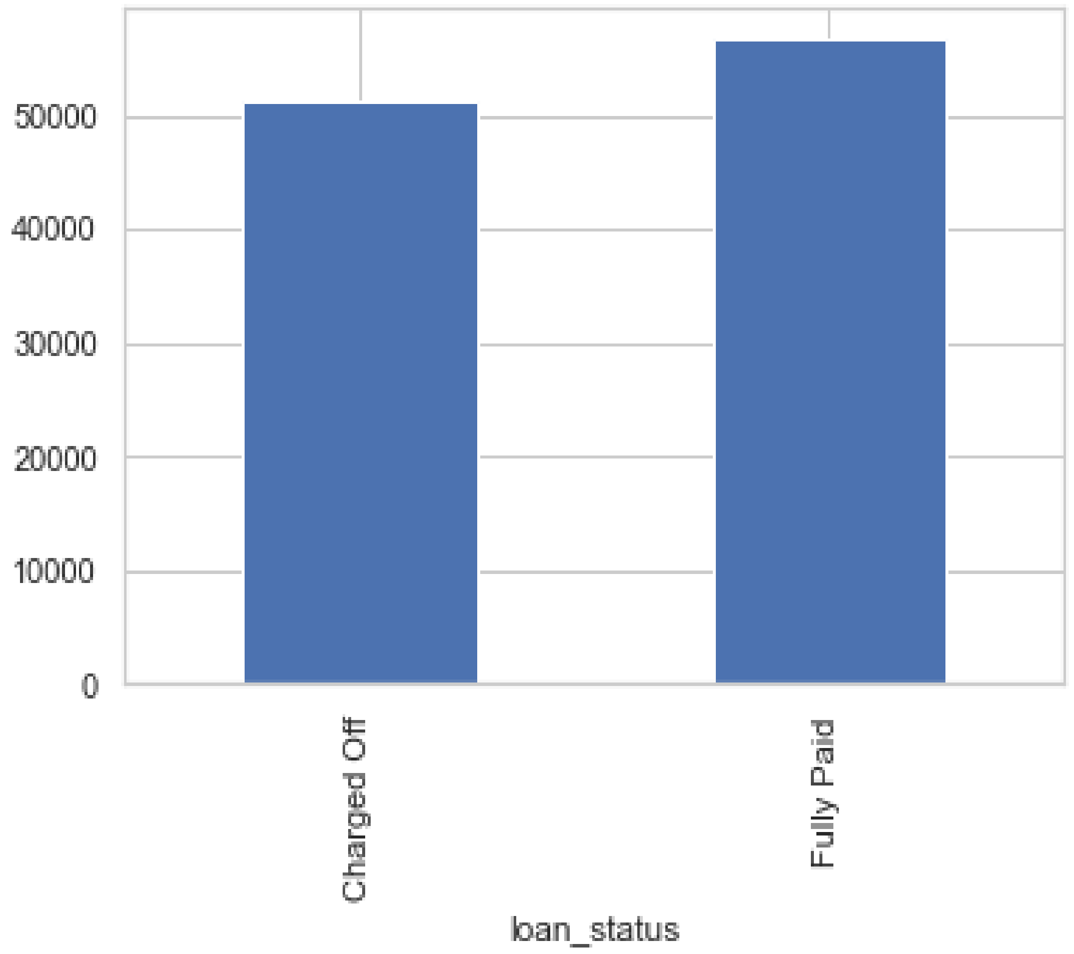
More higher loan amount might be a causation of charged off

- Here the x axis shows loan status where y axis shows loan amount .
- We can clearly see the more loan amount is inducing the higher charged off rate.



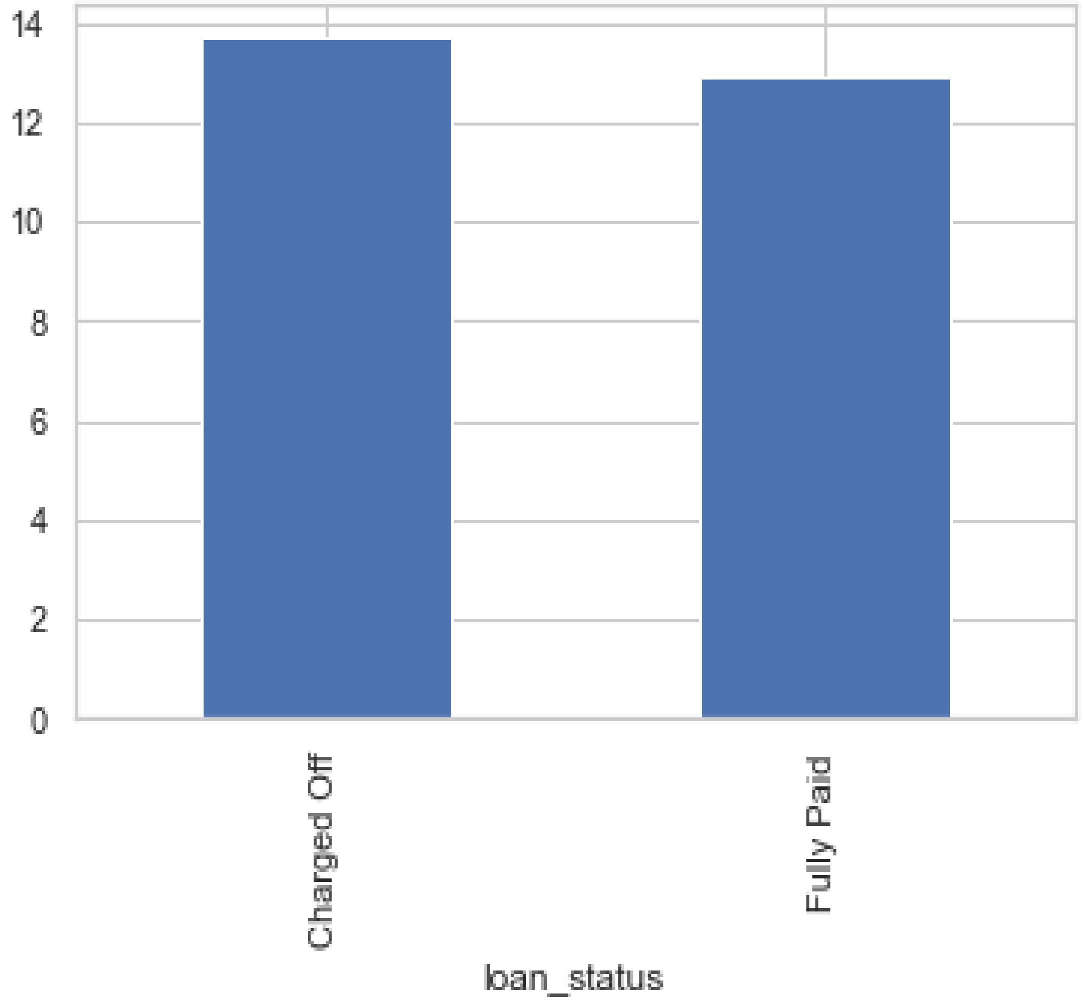
Defaulters are having less annual income than the fully paid Borrowers.

- Here x axis represents the loan status and y axis represents annual income.
- We can clearly see that the annual income of the charged off population is lesser than the fully paid population.



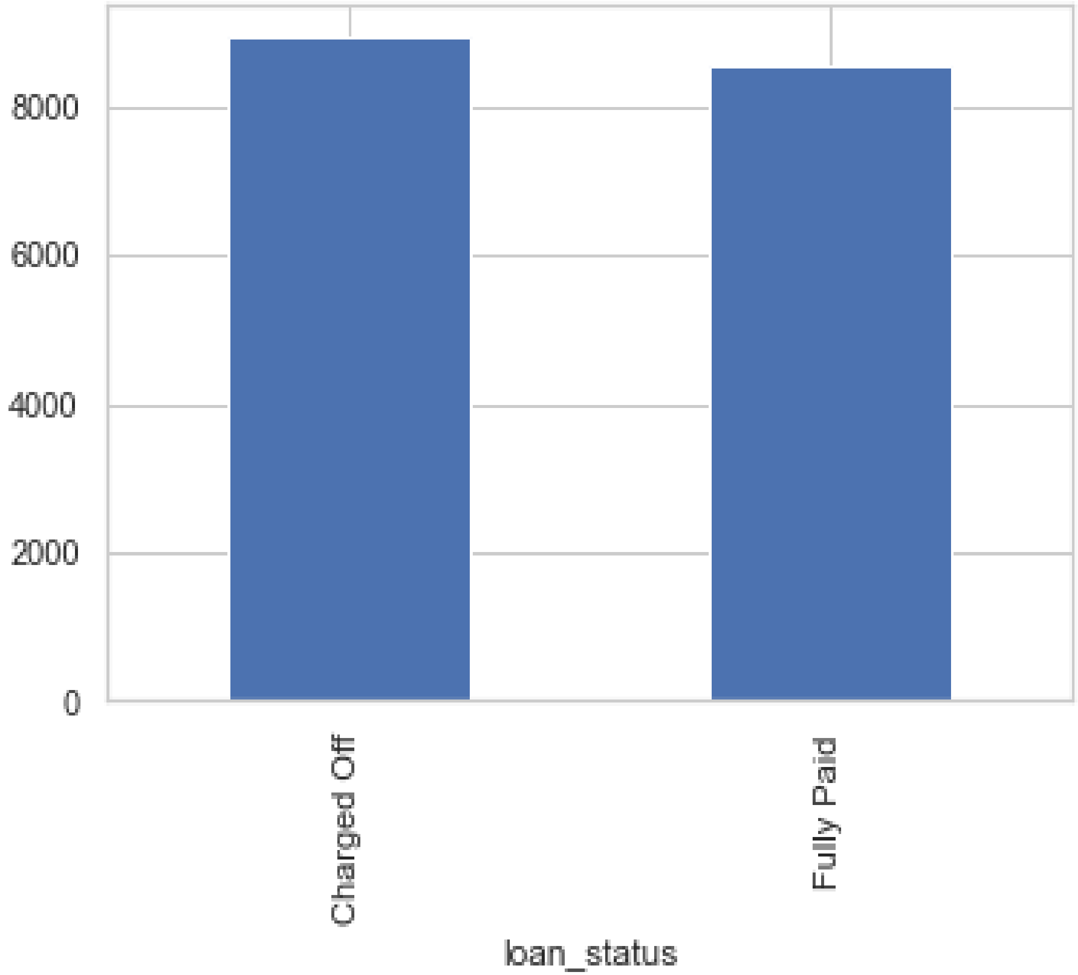
DTI might be a risk contributing factor.

- Here the x axis represents the loan status and y axis represents the DTI.
- DTI(Debt to income ratio) = debt/income .
- We can clearly see the charged off population is having more DTI in comparison to fully paid population . Which says people who have more DTI are likely to default.



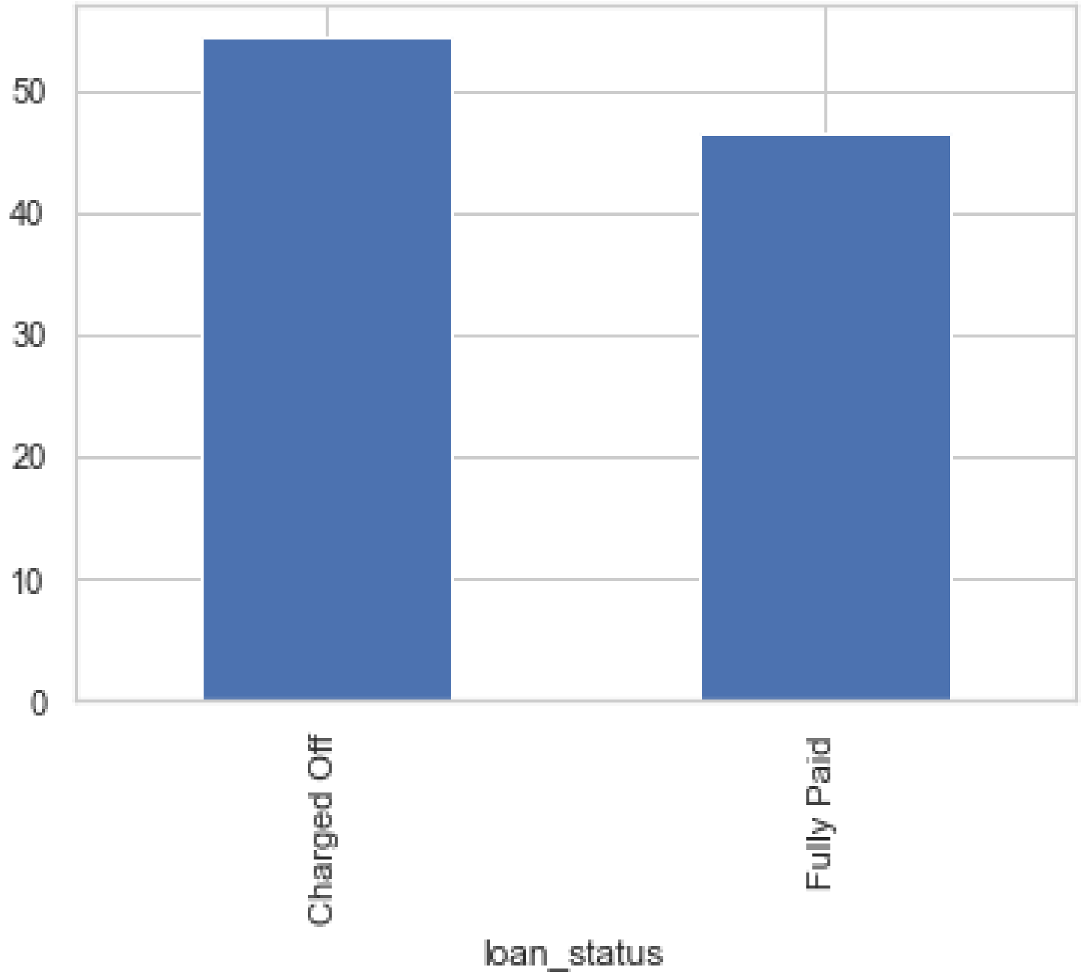
Revolving balance is positively contributing to the charged off population

- Here the x axis represents loan status and y axis represents the revolving balance .
- As we can see the more amount of revolving balance utilization is contributing to charged off population.

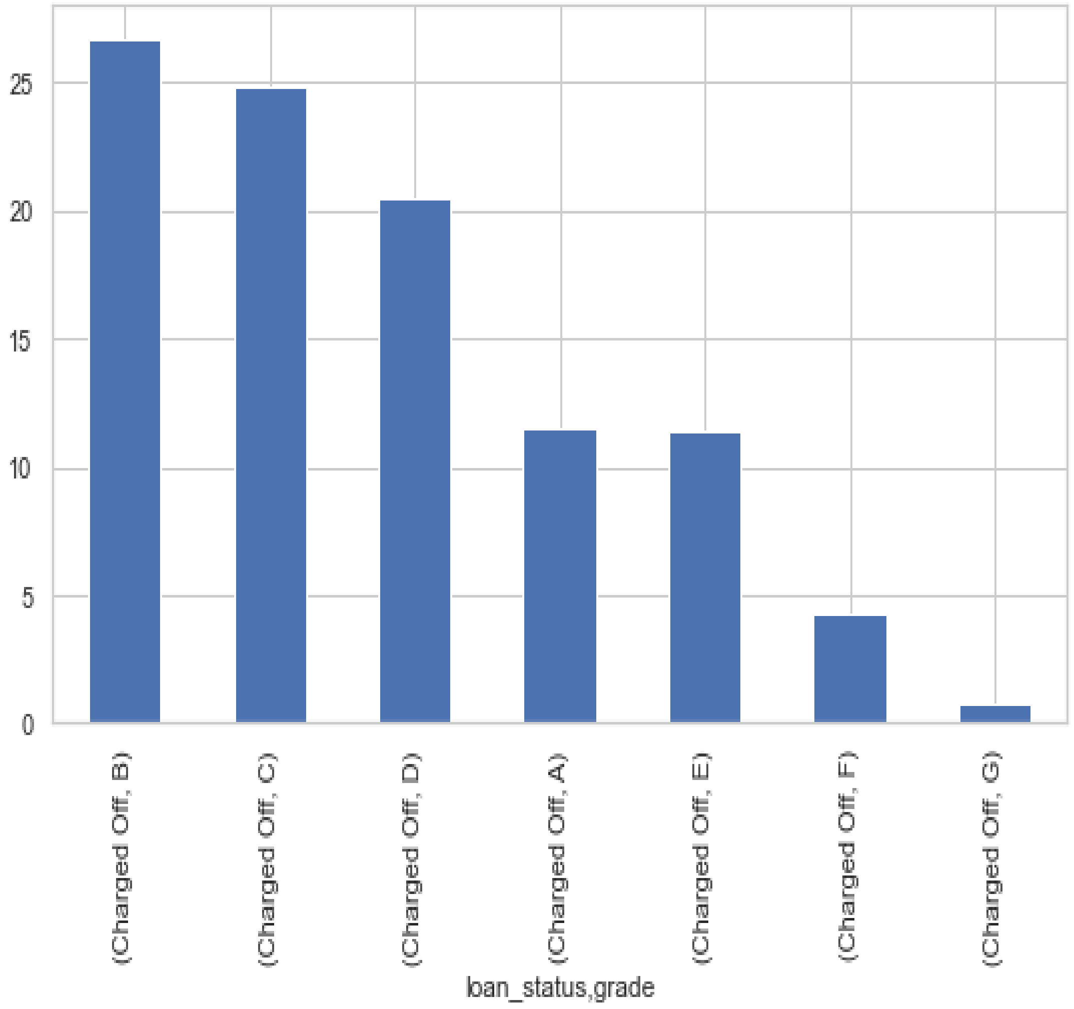


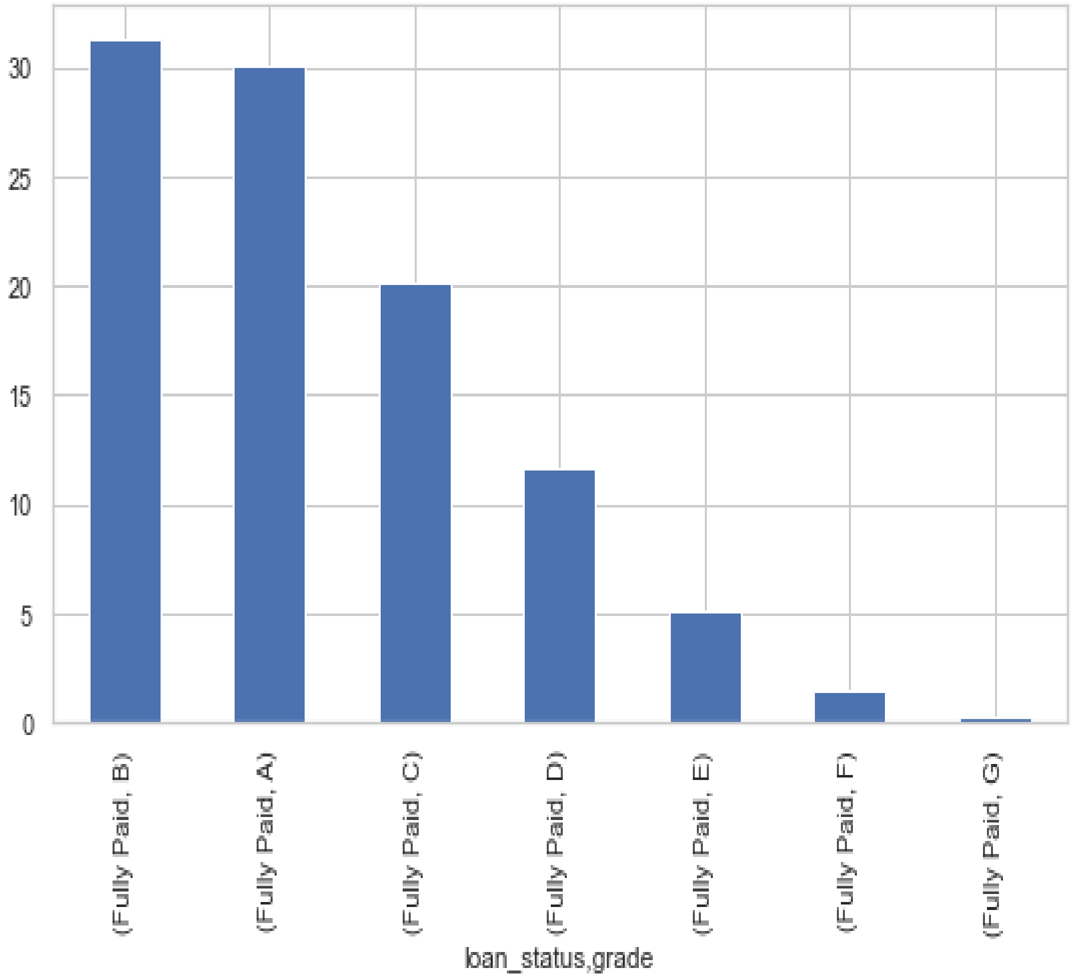
Revolving balance utilization is significantly contributing to the charged off population

- Here the x axis represents the loan status while the y axis represents what percentage of the revolving balance against the EMI amount has been used by the borrower.
- We can clearly see the population who are using more amount of revolving utilization are likely to charged off.



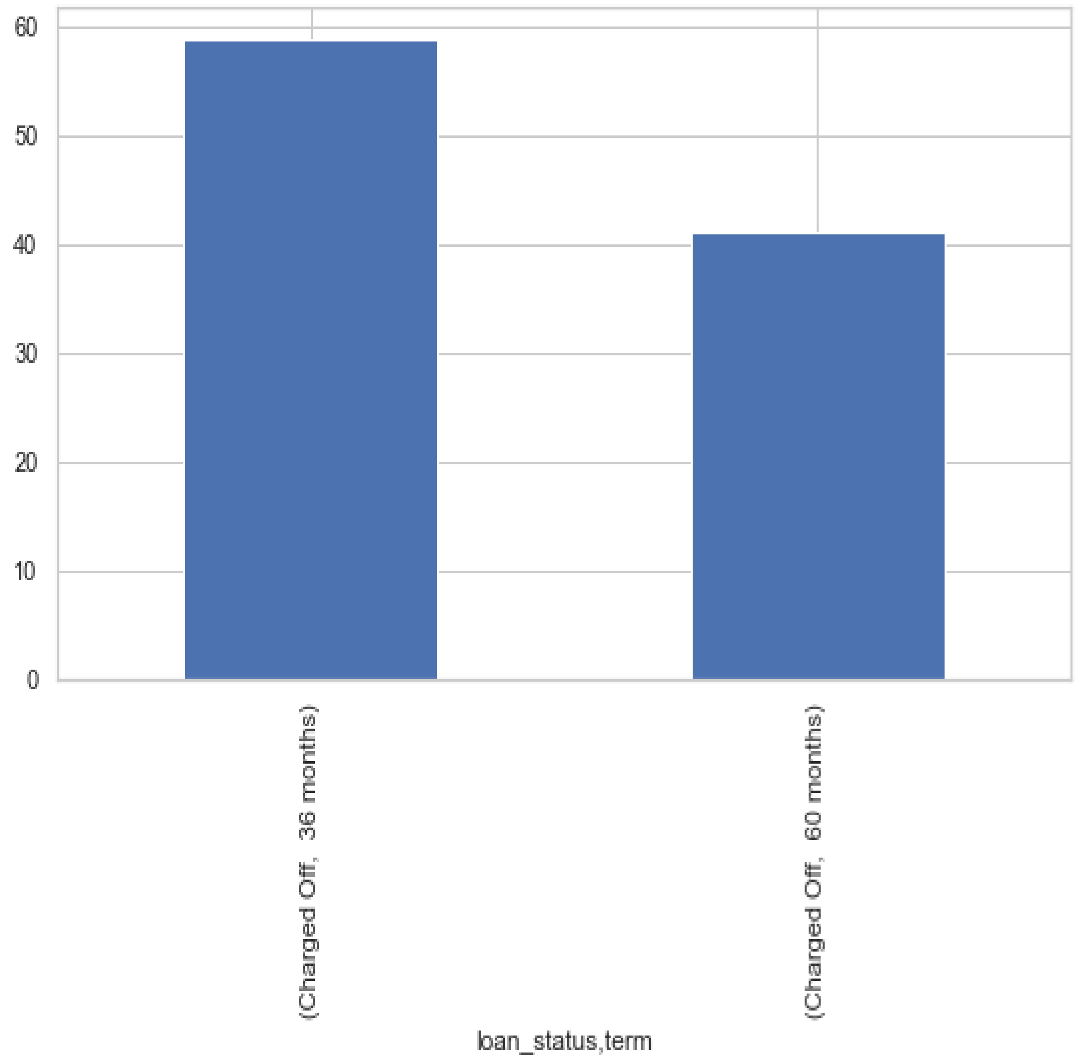
Grade of the applicant is a major risk contributing factor

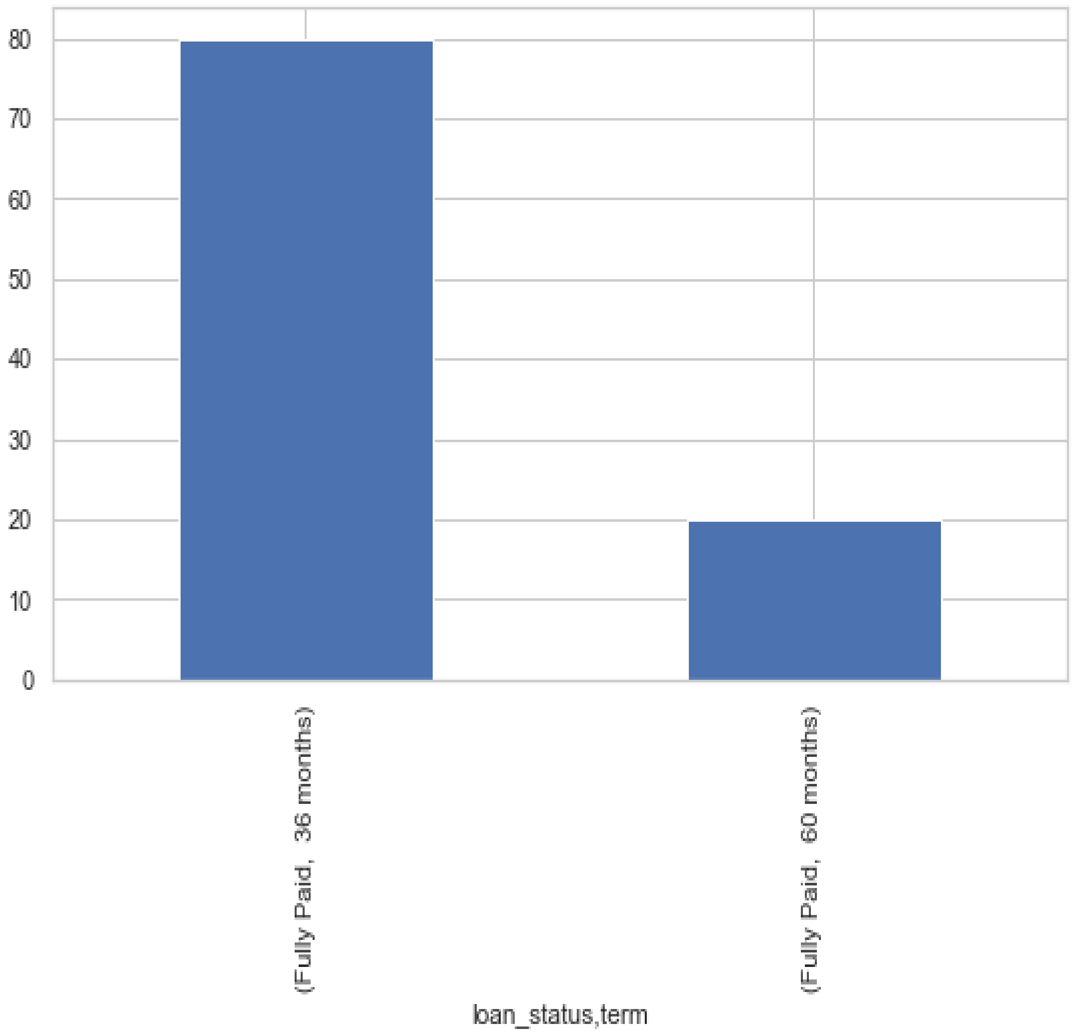




- If we see at the above figure then x represents (loan status, grade) and y represents the percentage of charged off population in the first graph , percentage of fully paid population in the second graph.
- If we compare both then we can see there are more percentage of high risk grade population in the charged off section rather than the fully paid section.
- And also we can see more percentage of lower risk grade population in the fully paid section rather than the charged off section.

Term is also serving as an important indicator of risk .



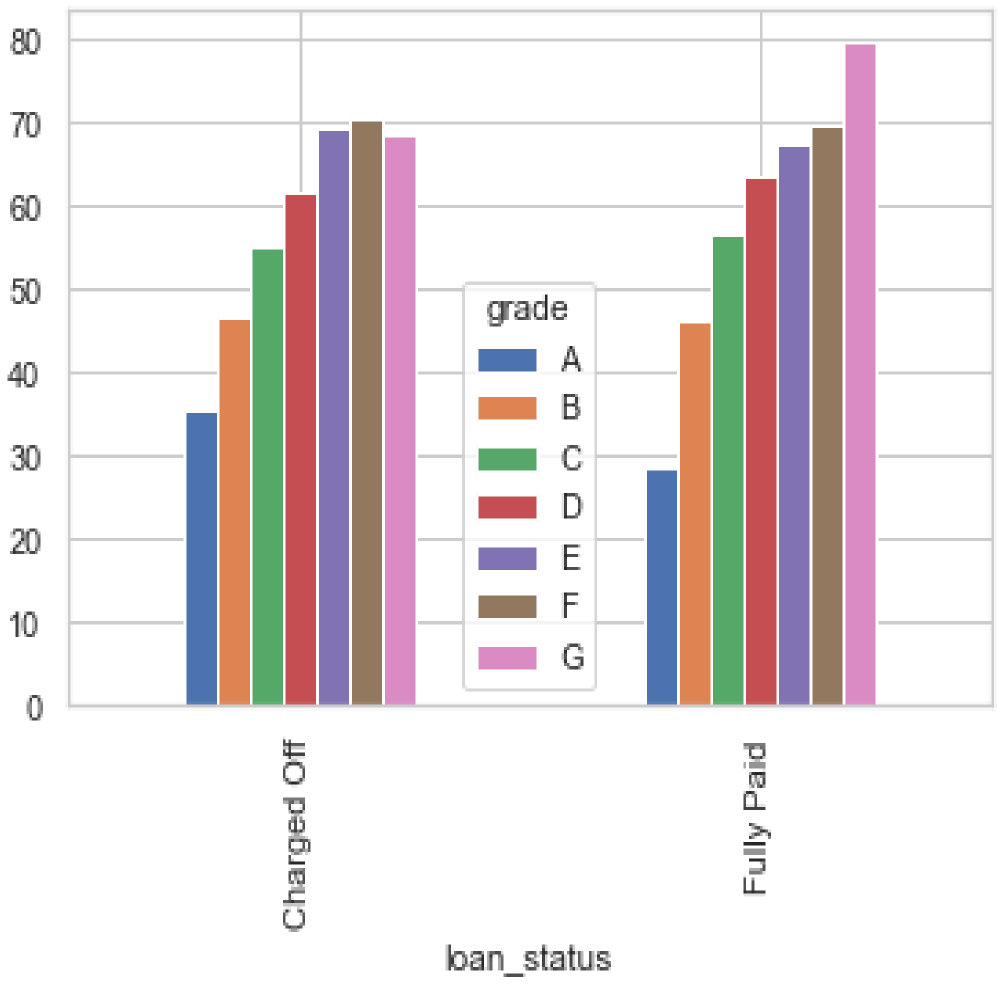


- If we see at the above figure then x represents (loan status , tenure of loan) and y represents the percentage of charged off population in the first graph , percentage of fully paid population in the second graph.
- If we compare both then we can see there are more percentage of population who have opted 36 months tenure in fully paid section in comparison to the charged off section.
- And also we can see there are less percentage of people who have opted for 60 months tenure in fully paid section in comparison to the charged off section
- Hence this indicates the higher the tenure is the more it can contribute to the charged off population.

Did multivariate analysis to describe the relationship between the ***<u>target variable</u>*** and ***<u>important attributes</u>***

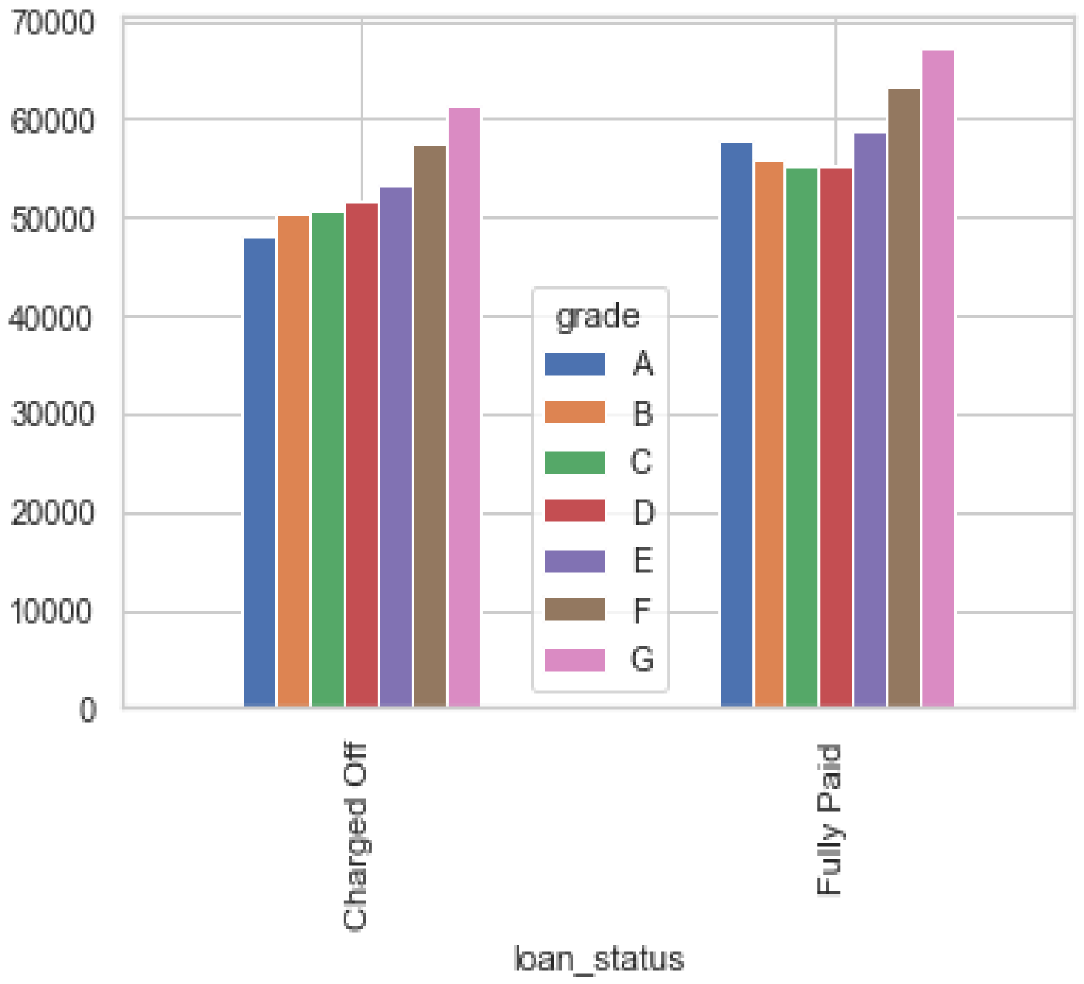
How loan amount and grade combinedly contributing to the charged off population ?

- The right graph describes the relationship between the loan status with grade and loan amount.
- If we see closely then Except the grade A population in all other grades are taking lesser loan amount in fully paid section , in comparison to the charged off section.



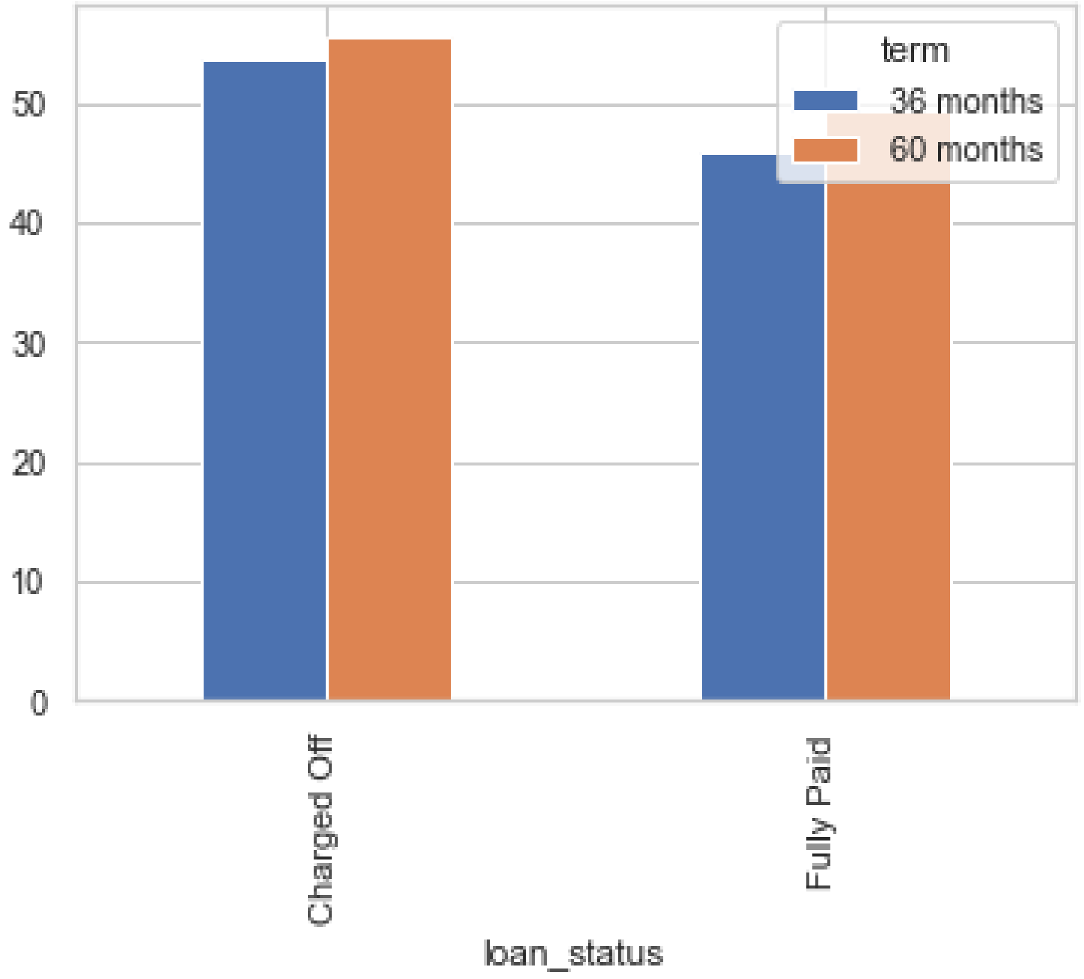
How annual income and risk grade combinedly contributing to charged off population ?

- The right graph describes the relationship between the annual income and grade with the loan status.
- We can see the annual income is higher for all the risk grades in fully paid section in comparison to the charged off section.
- Which clearly indicate the annual income is a great indicator of the default risk.



How revolving utilisation and term combinedly contributing to the charged off population ?

- The right graph is describing that the charged off population across both the tenure have been using more than 50 percent of revolving balance utilisation.
- Which clearly indicates irrespective of term revolving balance utilization is an indicator of higher charged off rate.



- Similarly we have also identified contribution of DTI , and revolving balance combined with term and risk grades across loan status.
- And found out DTI and revolving balance having a positive impact on boosting the charged off population.

Summery

We have found following pointers from our analysis-:

- More higher loan amount is a causation of charged off in case of higher term and Risk grade.
- Defaulters are having less annual income than the fully paid Borrowers , which is boosting default heavily.
- DTI is a risk contributing factor in case of most population.
- Revolving balance is positively contributing to the charged off population.
- Revolving balance utilization is significantly contributing to the charged off population.
- Grade of risk given by lending club is very much likely to be true in case of contribution to charged off as we have seen in our analysis .
- Term is also serving as an important indicator of risk for sure.

Recommendations

We would like to suggest some points to LC based on our analysis-:

- Amount of loan given should be less for the people having higher risk grade
- Amount of loan should be greater only if the customer is in the lesser risk category A and opting for a 36 months term.
- Deny to give loan to the customers who are earning less than 50000 if allowed put a higher interest rate .
- DTI should not be greater than 13(after rounding) in case of 36 months term and should be less than and equal to 14(after rounding) in case of 60 months term .
- incase DTI does not fall in range reject the loan application.
- Do not allow people to utilize more % of revolving balance. Restrict the use of Revolving balance.