Lending Club Case Study

1) Shruthip Venkatesh
2) Prerna Shukla

ML-C54

Introduction

- Lending club company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
- Lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). There are two types of risks associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- Borrowers who default cause the largest amount of loss to the lenders.
- **Objective:**
  - The aim is to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default using EDA.

Methodology to find the driver variables

Data Sourcing and Understanding → Data Cleaning and Manipulation → Univariate Analysis → Segmented Univariate Analysis → Bivariate Analysis → Multivariate Analysis → Recommendations

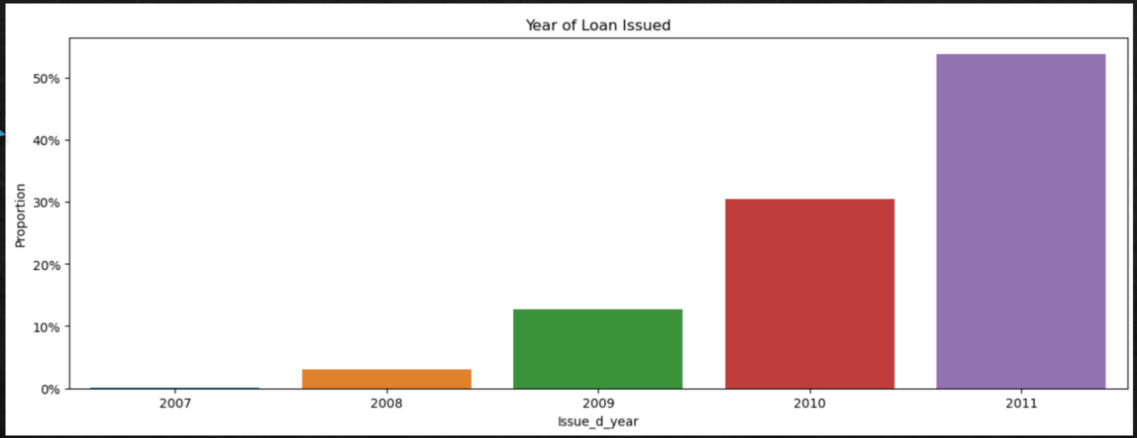
Note : Only the important points and the observations are highlighted in these slides. For detailed information, kindly have a look at the python notebook available here.

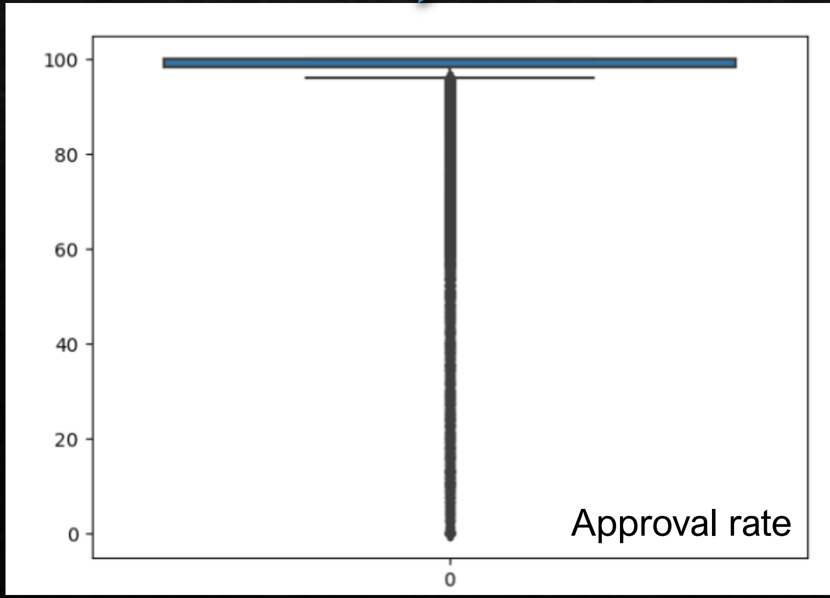
Points to remember

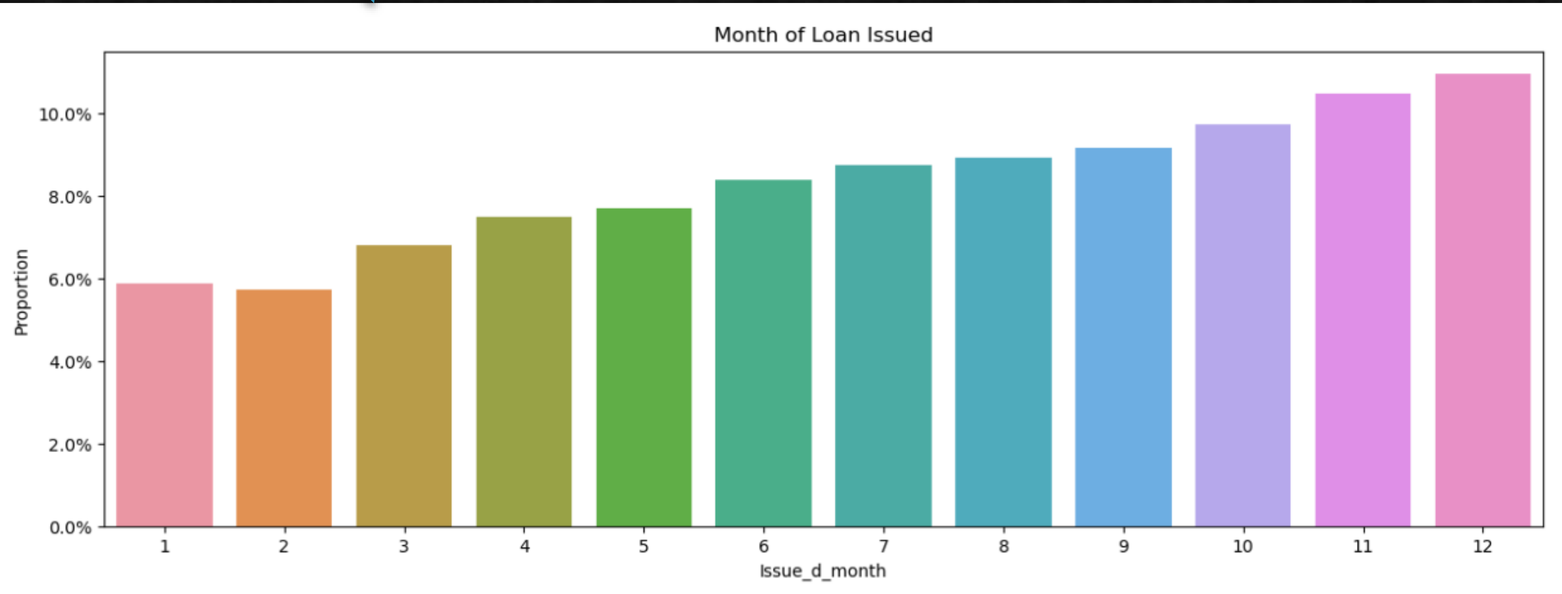
- Records of people having ‘current loan’ status are not considered as we are interested in finding the factors for defaulters.
- Meaning for 'verification_status_joint’ column in loan data was not mentioned in the dictionary excel sheet. But we found a similar column 'verified_status_joint’ meaning in excel sheet. We assume these two columns to be the same and considering the appropriate meaning in the loan data.
- After data cleaning, there were just **29 columns** from 111 columns and reduced around 2k records.
- Derived year and month from issued date and year and month from earliest credit line date.
- Derived approved loan amount as a percentage of loan amount funded by investors.
- Derived new columns based on binning few existing columns.
- Outlier treatment is done post univariate analysis.

Trend of borrowers (from Univariate analysis)

- The company has significantly issued more loans every year and more towards the end of the year.
- The company has approved most of the loans at the requested amount.

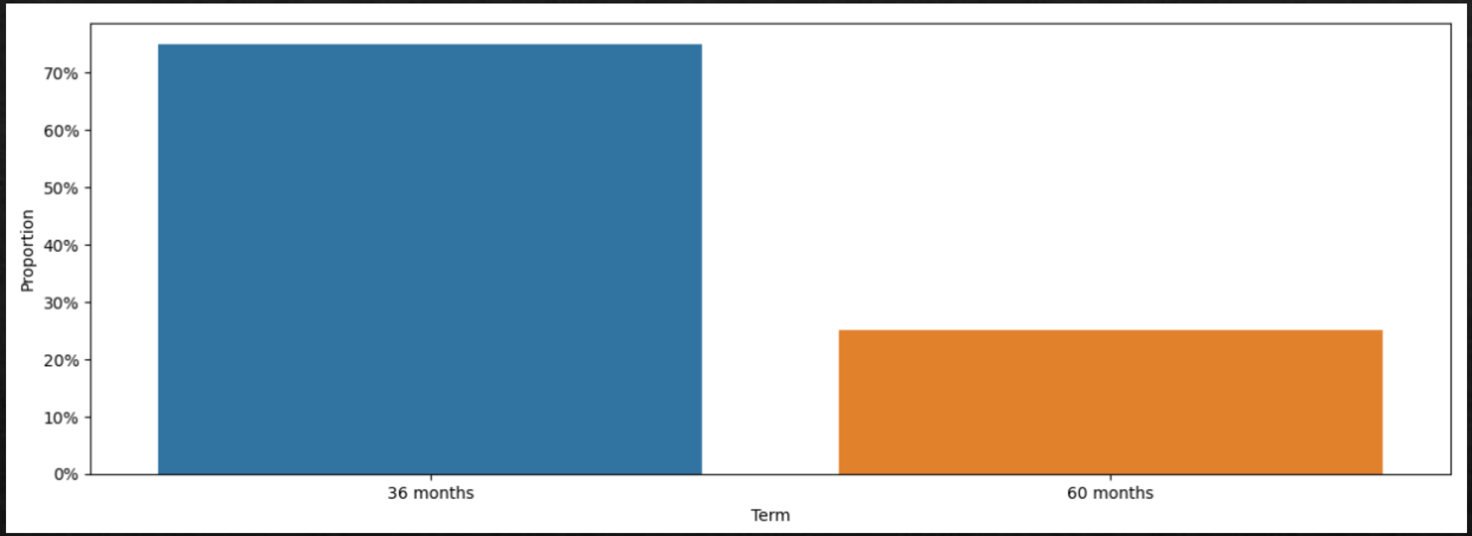




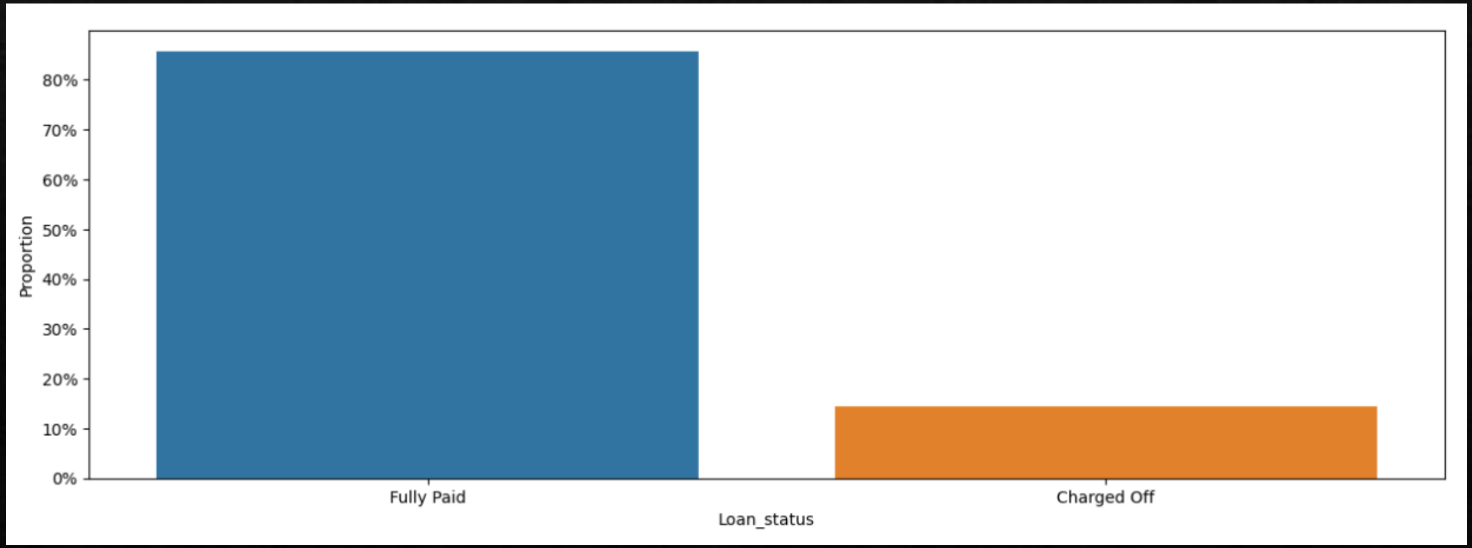


Trend of borrowers (from Univariate analysis)

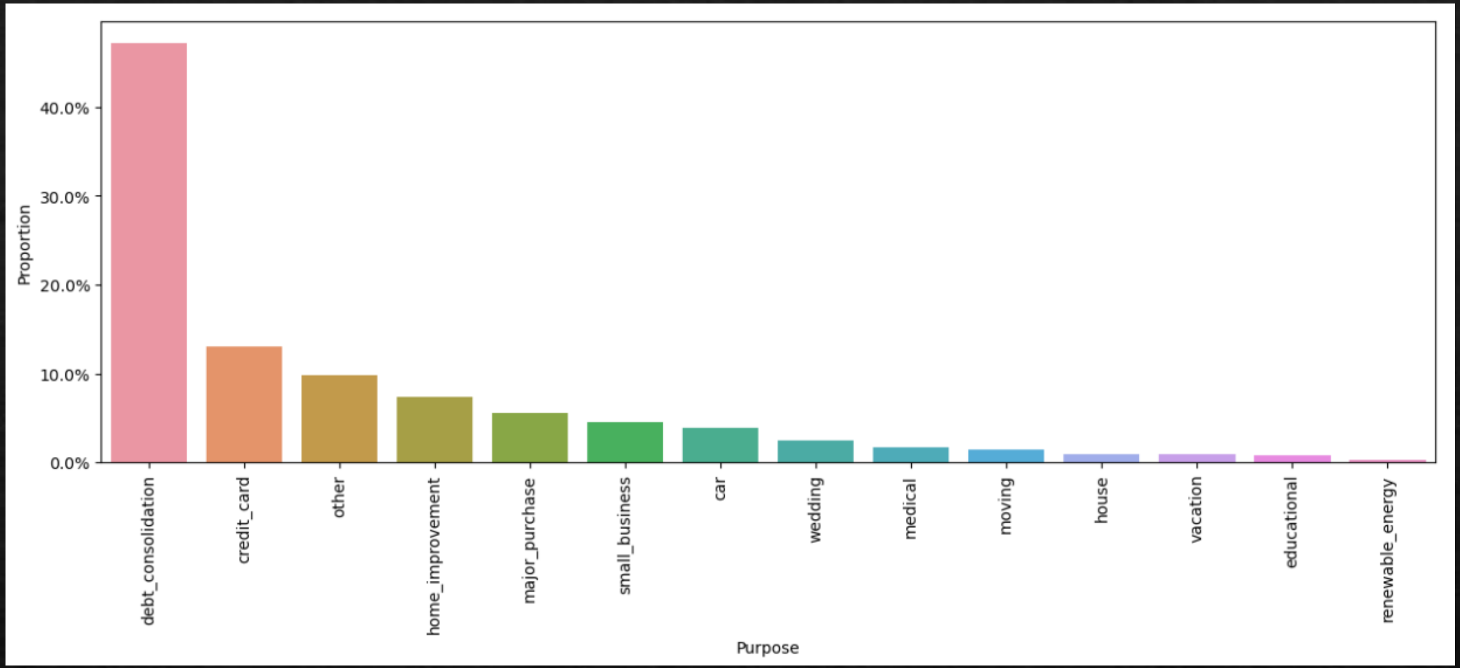
- Around **75%** of borrowers have taken loans with small term, 36 months



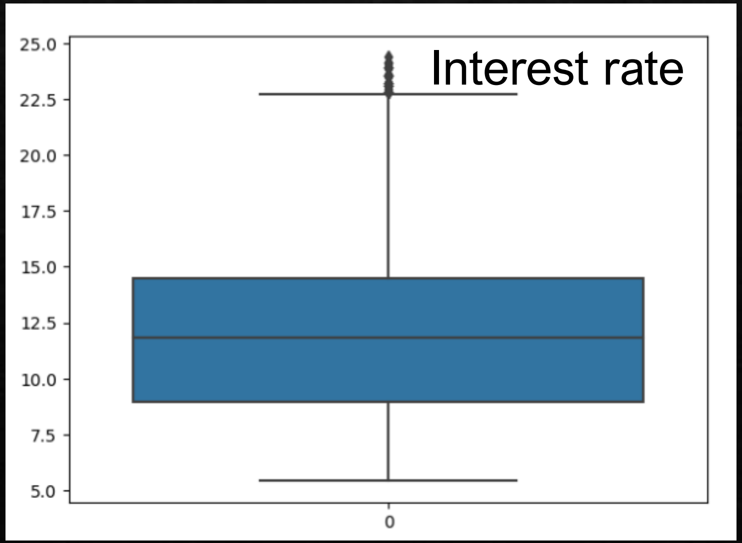
- Around **85%** of borrowers have fully paid the loans.



Trend of borrowers (from Univariate analysis)



- Debt consolidation and credit card contribute to the majority(60%) purpose of the loan taken.



- Most of the loans have an interest rates from 8.9% - 14.5% with an average of around 12%.

Other Highlights from Univariate analysis

- 75% of borrowers have requested for smaller loan values of 15000 and below.
- Majority of borrowers belong to grade B (30%) and A (25%).
- Borrowers with more experience 10+ years and less experience < 4 years contribute to 50% of the total borrowers. This behavior may be because more experienced people need loan for meeting their bigger responsibilities like education, marriage of kids etc and less experienced people earn less which makes them to take loan for buying their needs like a vehicle, mobile phones, education loan etc.
- 90% of borrowers are with home ownership as RENT and MORTGAGE.
- Most of the borrowers are from California and New York city.
- 75% of borrowers earn an annual income of less than or equal to 83000.

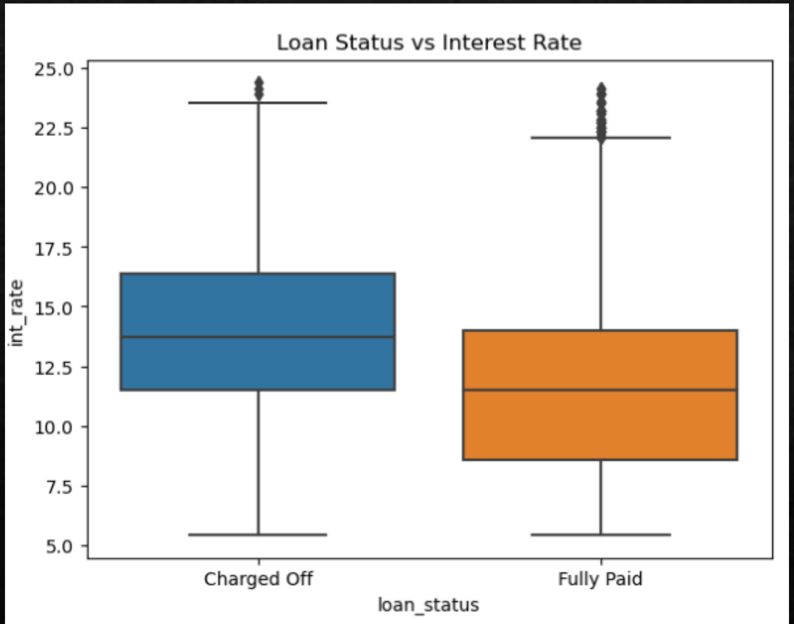
Post Univariate analysis:

- Outliers were found in annual income, open credit lines and total credit lines. These outliers were removed which removed around 1k records of borrowers.

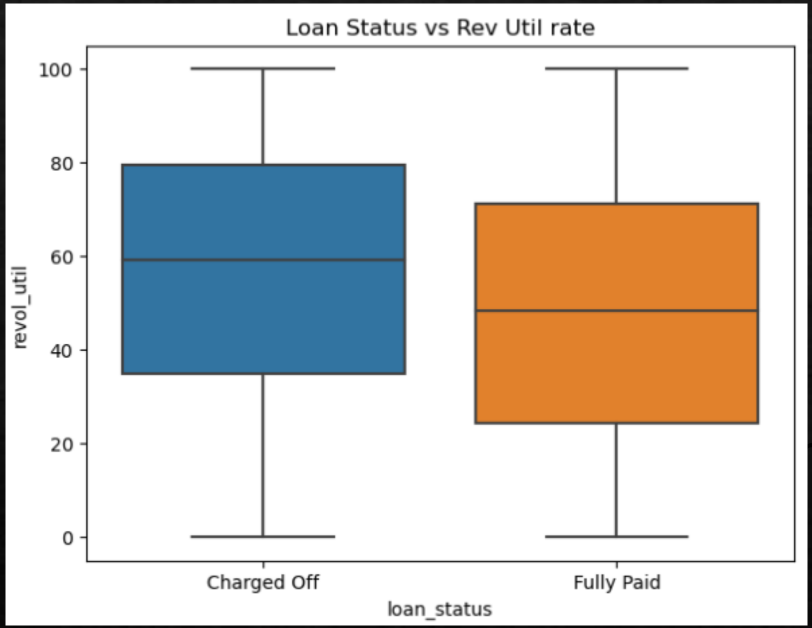
- We are ignoring few columns like sub grade, zip code and title as they have alternative columns and also they have wide range of categories and analyzing that many categories will not provide much insight.

Trend of defaulters (from Segmented Univariate analysis)

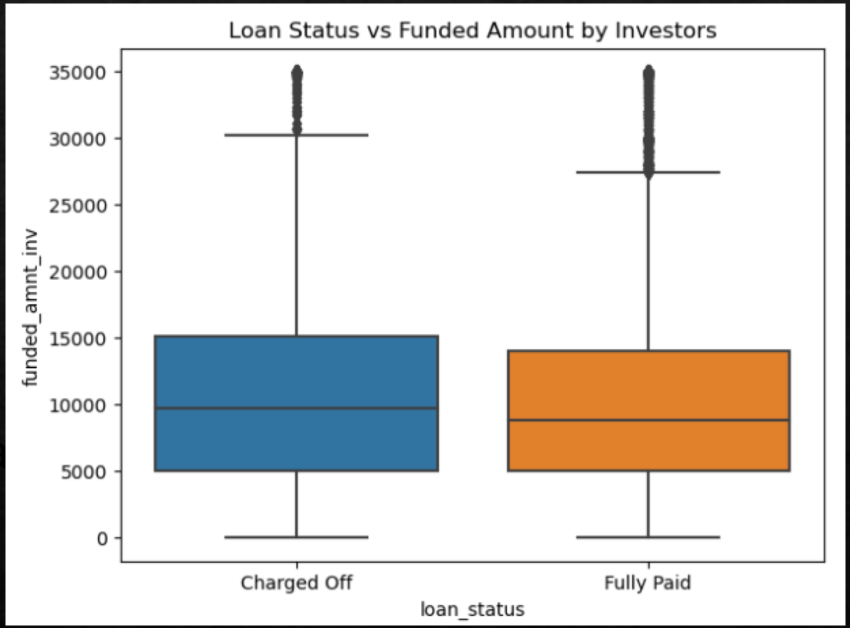
- Defaulters are more at higher interest rates.



- Borrowers with high revolving utilization rate tend to be defaulters.



- Loan amount might influence defaulters. Higher amounts might be tough to be paid.

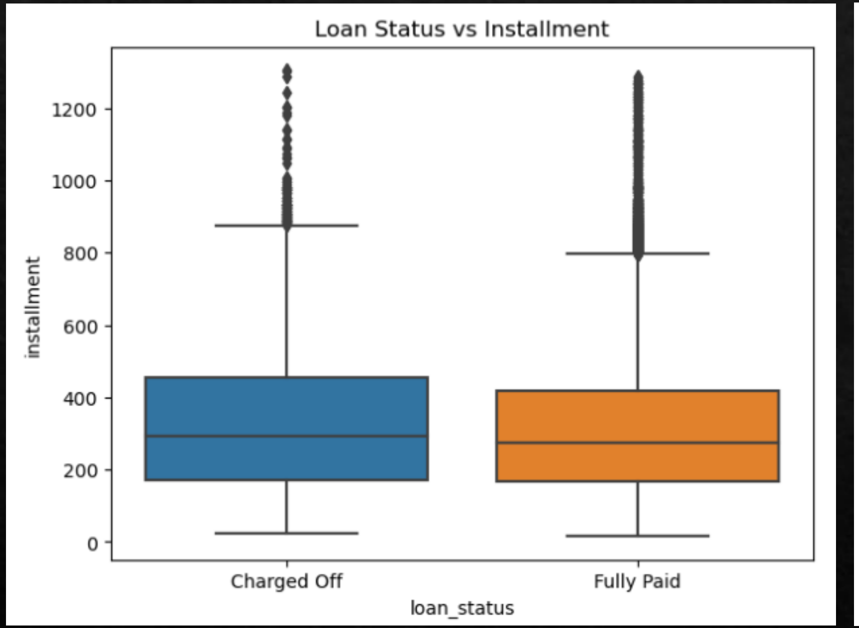


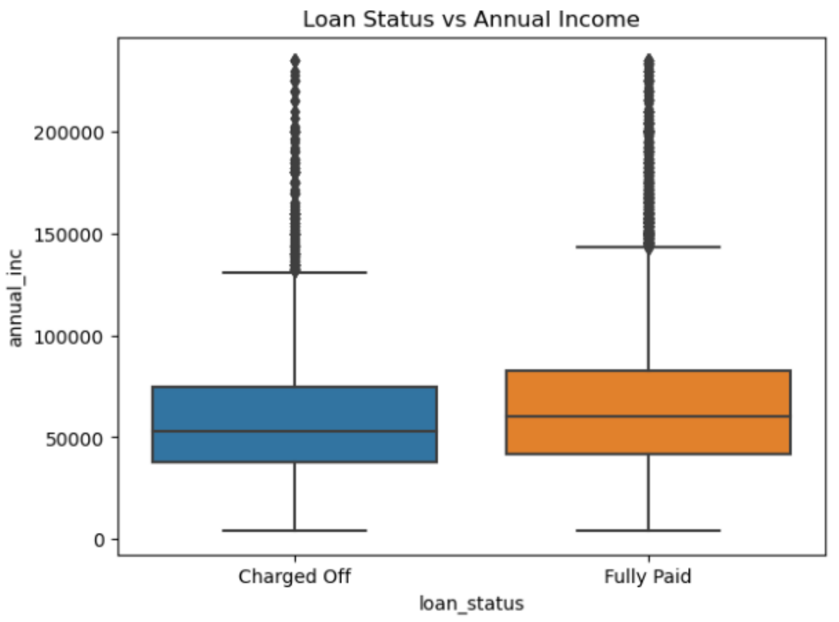
Trend of defaulters (from Segmented Univariate analysis)

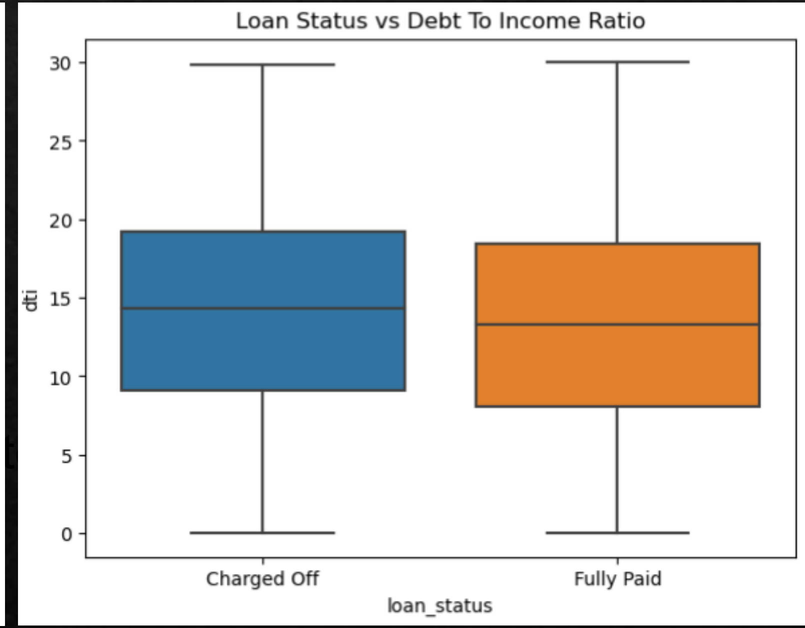
- Installment might influence defaulters.

- Annual income might influence defaulters. Less salary might be a reason they could not pay

- Debt to Income ratio might influence defaulters.







Other Highlights from Segmented Univariate analysis

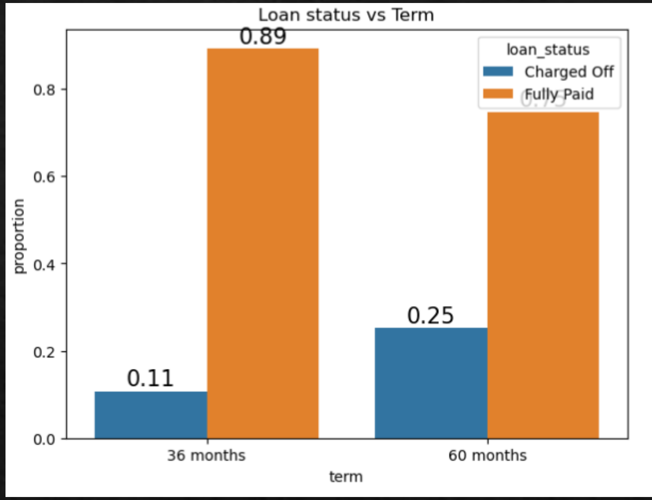
- Borrowers from lower grades tend to request for higher loan amount.
- Borrowers from OTHER home ownership category have lesser approval rate.
- Most of the borrowers from lower grade have higher interest rate and higher revolving utilization rate.
- Long term loans have high interest rate.
- High salaried borrowers take long term loans, it might be because they are taking bigger loan amounts.
- Borrowers in MORTGAGE and OWN have lesser interest rate, this could be because of the security.

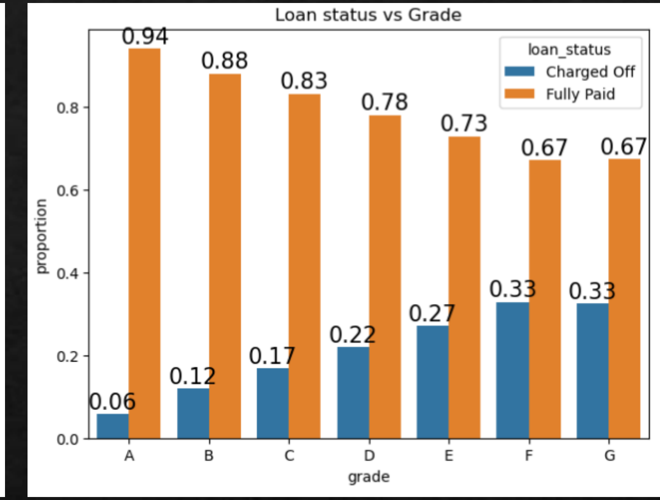
Post Univariate analysis:

- We found few columns (approved loan amount percentage, delinquency since 2 years, inquires in past 6 months, open credit lines, public derogatory, public recorded bankruptcies, revolving balance) do not have any impact on analyzing defaulter pattern.
- We can also see that loan amount, funded amount and funded amount by investors are all redundant. It is worth to just analyze any one of these columns, and let us choose funded amount by investor as it is the real value funded by investors.
- We can also see that loan issued month and earliest credit line month do not have any pattern with any column.

Trend of defaulters (from Bivariate analysis)

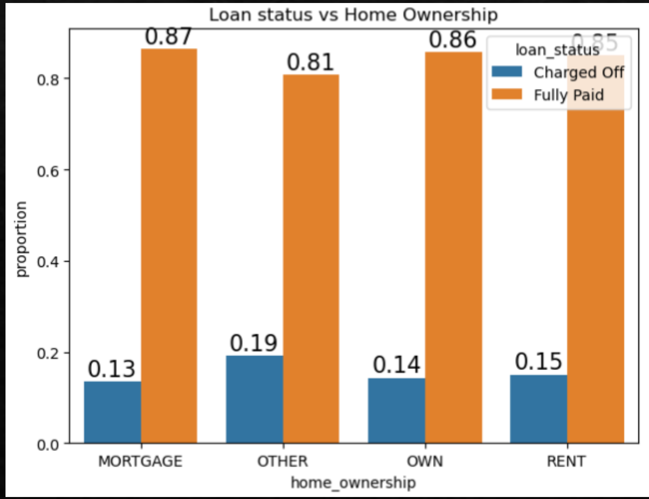
- Defaulters are more for long term loan.

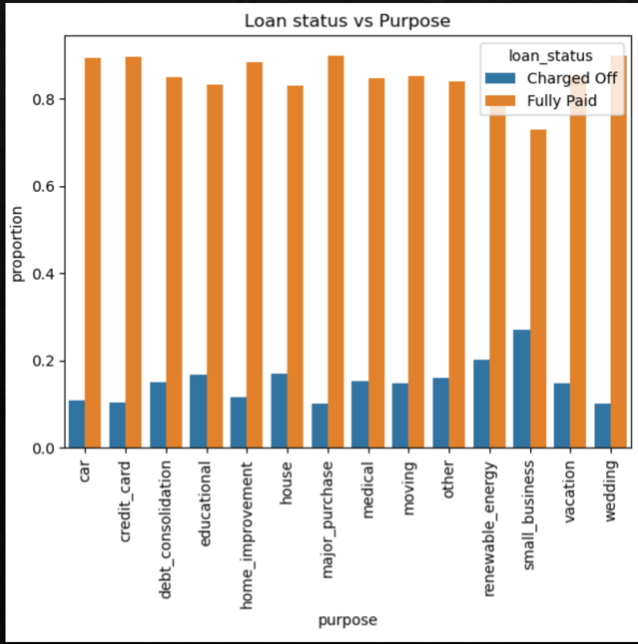




- As the grade decreases, the proportion of defaulters increases.

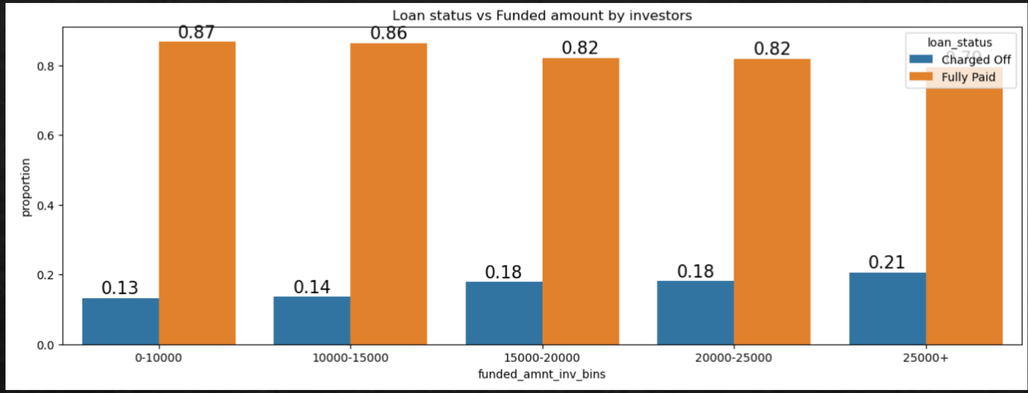
- OTHER category home ownership has more defaulters, and there is a pattern observed.



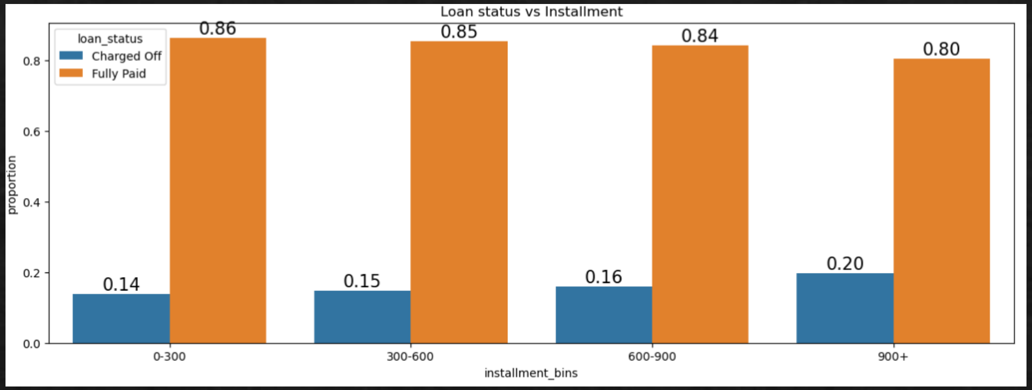


- Purpose has a clear difference in the proportion of defaulters. Loan taken for small business have the most defaulters, loan taken for wedding have the least defaulters

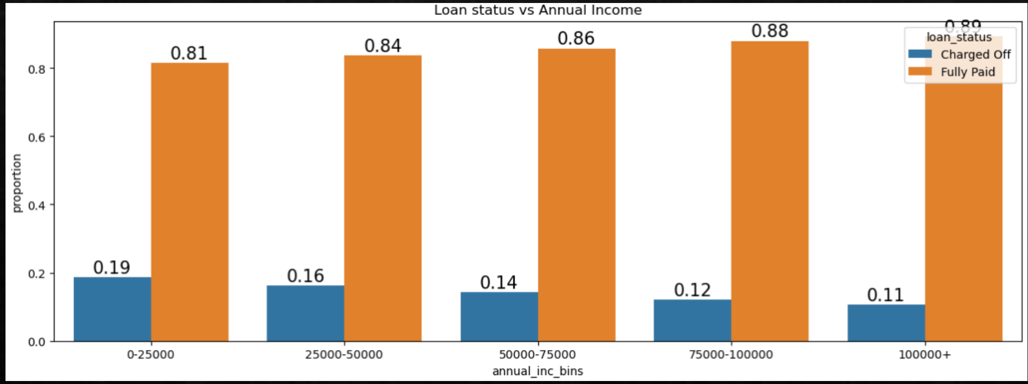
Trend of defaulters (from Bivariate analysis)



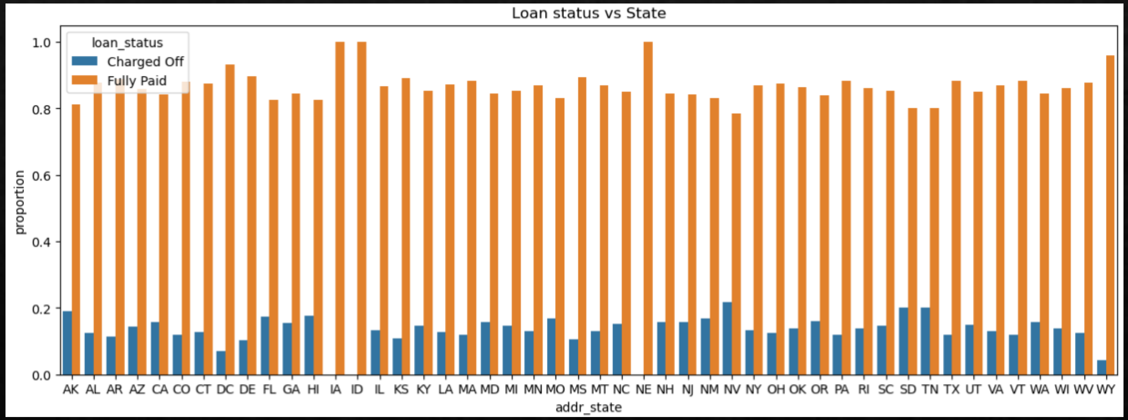
- We can clearly see that as the loan amount increases, the proportion of defaulters increases.



- We can clearly see that as the installment increases, the proportion of defaulters increases.

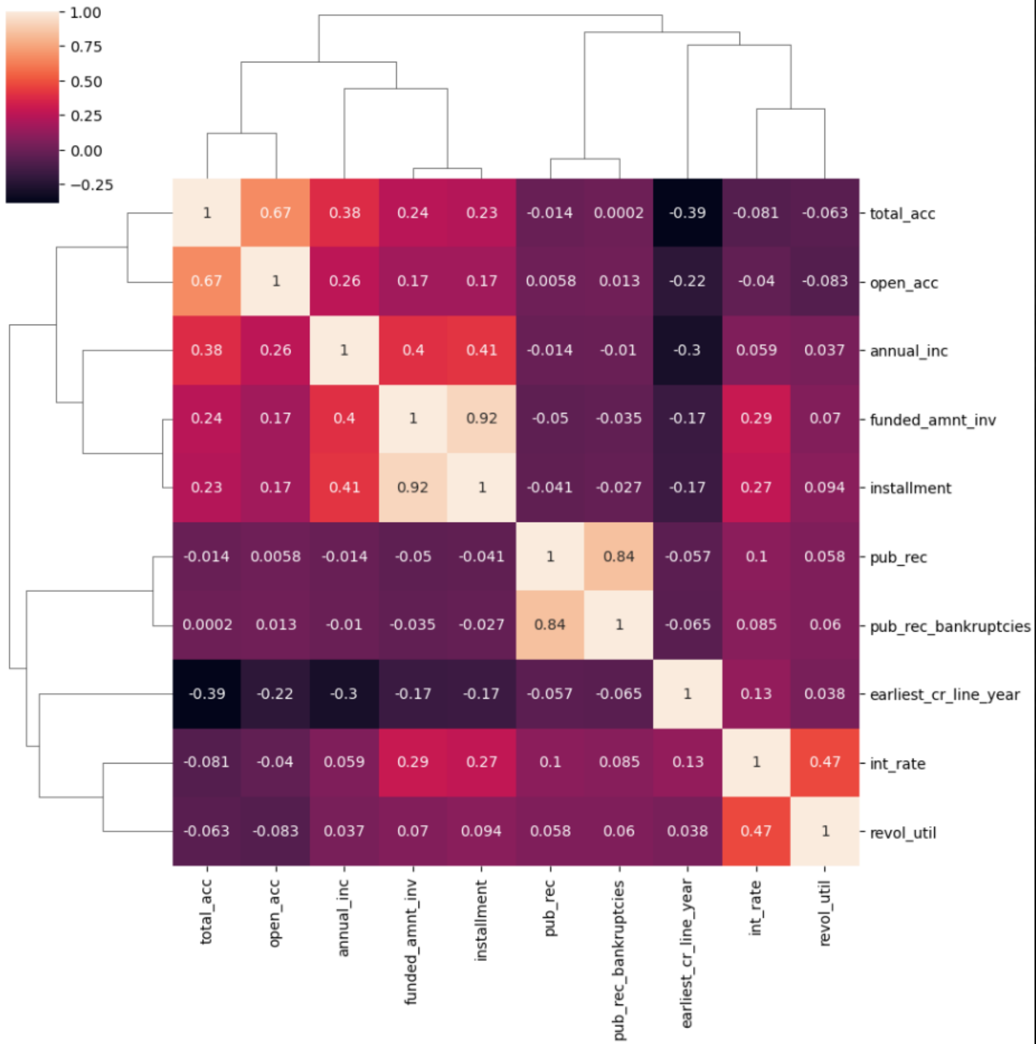


- We can clearly see that people earning low salary find it difficult to pay the loan and become defaulters. As the income increases, the proportion of defaulters decreases.



- Though there is not a definitive pattern to comment, we can see that the demography has some influence on defaulters.

Highlights from Bivariate analysis



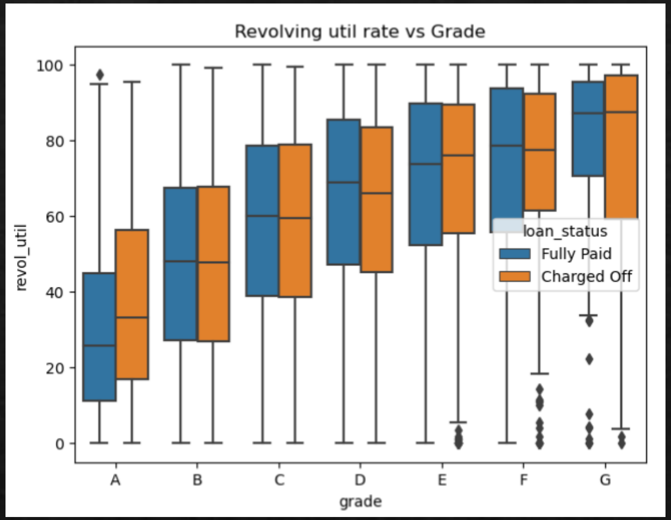
- More defaulters are seen from verified borrowers than not verified borrowers. But this does not imply that we should stop verification to avoid defaulters.
- People from NV have more defaulters whereas those from IA,ID and NE have no defaulters. Though there is not a definitive pattern to comment, we can see that the demography has some influence on defaulters.

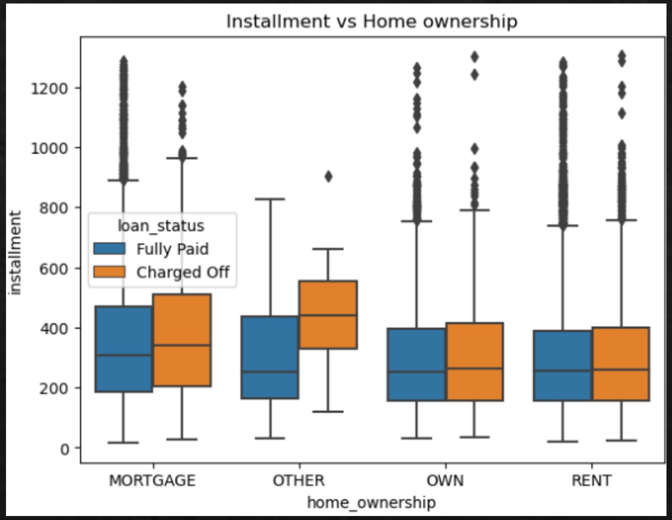
From the heat map, we could see

- Funded amount by investor is highly related with installment of borrowers.
- Public derogatory is highly related with public recorded bankruptcies.
- Total credit lines is related open credit lines.
- Interest rate is related with revolving utilization rate.
- Total credit lines is inversely related with earliest credit line.
- Annual income is inversely related with earliest credit line.

Trend of defaulters (from Multivariate analysis)

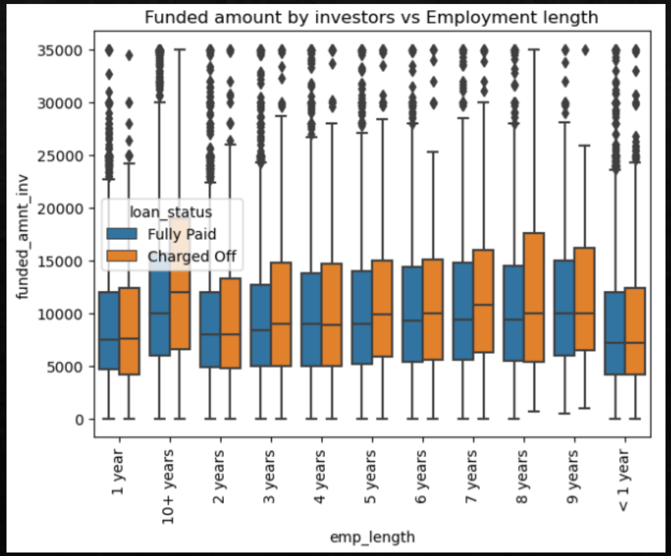
- For lowest grade G, there are more defaulters when revolving utilization rate is around 60+.

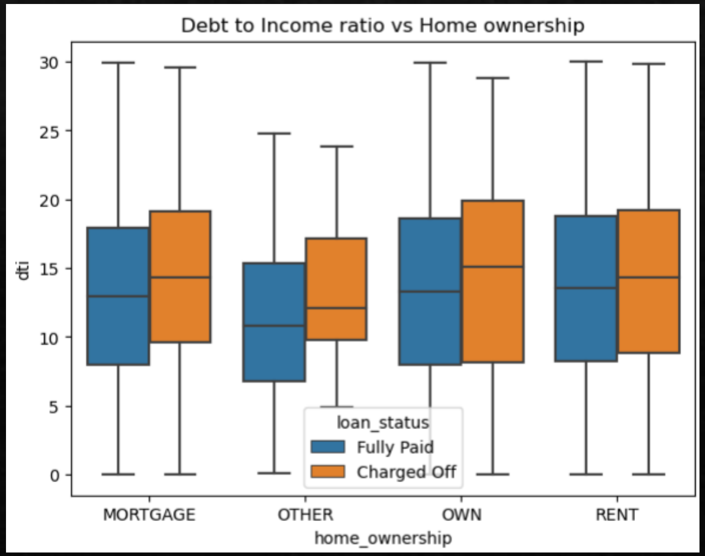




- Borrowers with OTHER home ownership, having higher installments tend to be defaulters.

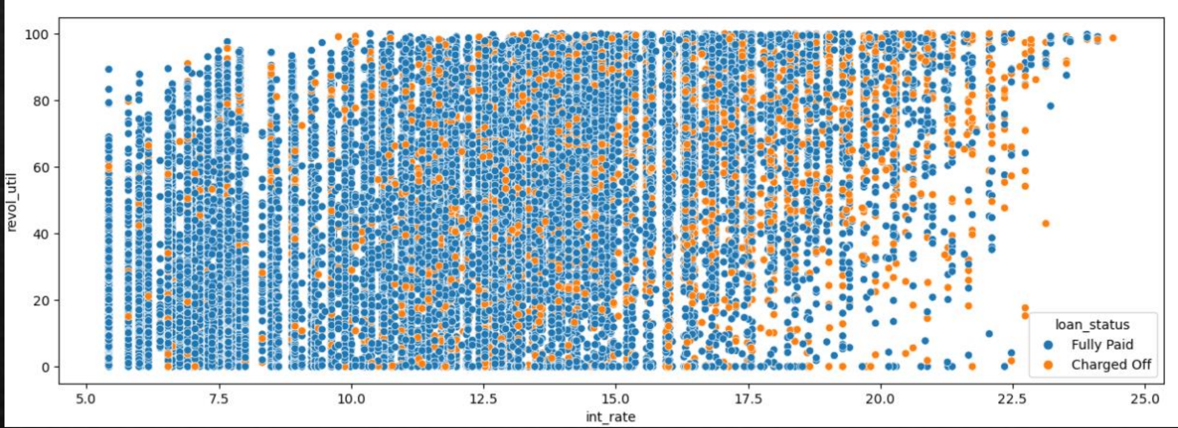
- As the experience increases, as the loan amount increases, there is a little increase in defaulters



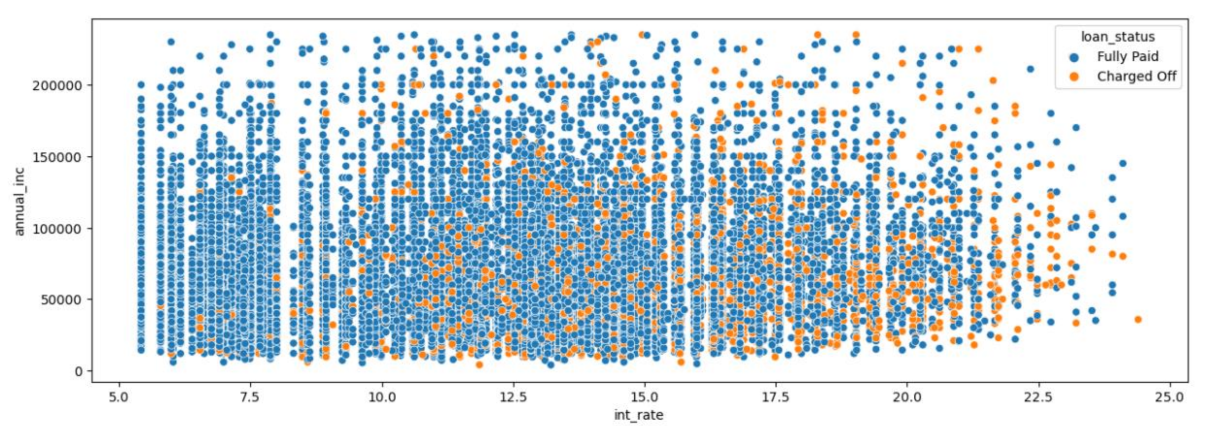


- People with MORTGAGE and OTHER ownership, at higher debt to income ratio, there is a little more defaulters.

Trend of defaulters (from Multivariate analysis)



- At higher interest rate, borrowers with high revolving utilization rate tend to be defaulters.



- When interest rate is more for low salaried people, they tend to default.

Other highlights,

- It is seen that interest rate, annual income and revolving utilization rate have high impact on defaulters irrespective of other columns.
- When income is less than 100000 and interest rate is more than 15%, there seems to be more defaulters.
- When open credit lines and total credit lines are high, they tend to default.
- When annual income is more, the requested loan amount is more.
- Borrowers with more derogatory tend to bankrupt who in turn tend to be defaulters.
- Borrowers having less total credit lines in recent years from 2000 tend to be defaulters.

Driving factors for defaulters

Driving Factor	Impact Level	How does it impact
Interest rate	High	Defaulters increase with high interest rate
Revolving utilization rate	High	Defaulters increase with high revolving utilization rate
Term	High	Lending long term loans increase defaulters
Grade	High	Defaulters are more as the grade decreases (High A to low G)
Purpose	High	Borrowers taking loan for small business tend to be defaulters followed by debt consolidation and credit card purposes
Funded amount by investors	High	Defaulters increase as the loan amount increases
Annual Income	High	When borrower earn low salary, they tend to default
Installment	Medium	When borrower installment is high, they tend to default
State	Medium	People from NV have more defaulters whereas those from IA,ID and NE have no defaulters. There is not a definitive pattern, we can see that the demography has some influence on defaulters.
Employment Length	Low	As the experience increases, as the loan amount increases, there is a little increase in defaulters
Home Ownership	Low	OTHER home ownership, having higher installments tend to be defaulters. MORTGAGE and OTHER ownership, at higher debt to income ratio, there is a little more defaulters.

Recommendations for Lending Club Company

To assess if the applicant might default

- Look into the grade of the applicant.
- Validate the purpose of the loan.
- Look into annual income, installment and revolving utilization rate.
- Look into the status of defaulters from the applicant’s state.

To avoid defaulters / To minimize the credit loss (either by rejecting or taking some measures)

- Based on the assessment and driving factors, control the percentage of loan approved.
- Reduce the loan approved for small business as they might not have high profit and tend to default.
- Promote short term loans by attractive interest rates and Reduce the interest rate for long term loans.
- Reduce the interest rate to below 15% for people with less than 100000 as their annual income.
- Reduce the loan amount approved for experienced people, 10+ years.
- For people with higher revolving utilization rate, either reduce/reject the loan amount.
- For people with OTHER home ownership, if they have high installments, either reduce/reject the loan amount.
- Reject loans to people with more derogatory who might tend to bankrupt.

Appendix

Data Sourcing and Understanding

- The loan details dataset is provided with around 40k records of borrowers(rows) and around 117 attributes (columns) for each borrower is provided in a csv file.
- The meaning of each column is detailed in data dictionary excel file.
- We found meaning of few columns mentioned in the excel was not available in the loan data.
- We found that meaning for 'verification_status_joint' column in loan data was not mentioned in the excel sheet. But we found a similar column 'verified_status_joint' meaning in excel sheet. We assume these two columns to be the same and considering the appropriate meaning in the loan data.

Data Cleaning and Manipulation

- Removed **54 columns** with complete empty data.
- Removed **4 columns** with more than 30% empty data.
- Removed **2 columns** which had less empty data, but they were the payment dates and it does not add any value to our analysis.
- There were **7 columns** which had **less than 3%** of empty data. Removing such columns will not be appropriate and imputing that much of data would affect our analysis. So, we removed those rows which had empty data in any one of these 7 columns.
- Removed **15 columns** which had more than **75%** of unique values. These columns will not help in understanding a pattern as it has too many unique values.
- Removed **6 columns** which just detailed information of the loan like principal collected, outstanding principal and post charge off fees which do not influence the analysis.
- After cleaning, there were just **29 columns** from 111 columns and reduced around 2k records.