LENDING CLUB CASE STUDY

Group Facilitator: Soumit Sarkar

Team Member: Harsha Billalli

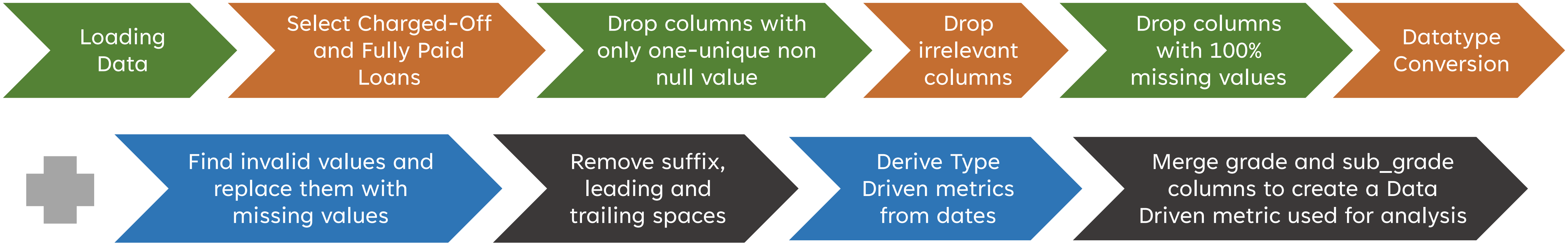
OBJECTIVE

- Upon receiving a loan application, the business must decide whether to approve the loan based on the applicant's profile. Granting the loan could result in a loss of money for the company if the borrower is not likely to repay the debt, or if default is probable.
- The purpose of the analysis is to identify the crucial variables that are solid indicators of loan defaults so that those making decisions can minimize the likelihood of default risk by utilizing them to accept or reject loan applications.

ABOUT DATASET

- The data contains details about previous loan applicants, including whether they had defaulted on their debts, were in the process of making loan payments (i.e. their loan tenure was not yet over), or had paid off their debts in full.
- Fully Paid, Charged Off, and Current are the three loan statuses associated with the dataset of sanctioned loans. For applicants who had their loan requests denied by the company, there is no transaction history included in the dataset.

DATA CLEANING



Loading Data

Select Charged-Off and Fully Paid Loans

Drop columns with only one-unique non null value

Drop irrelevant columns

Drop columns with 100% missing values

Datatype Conversion

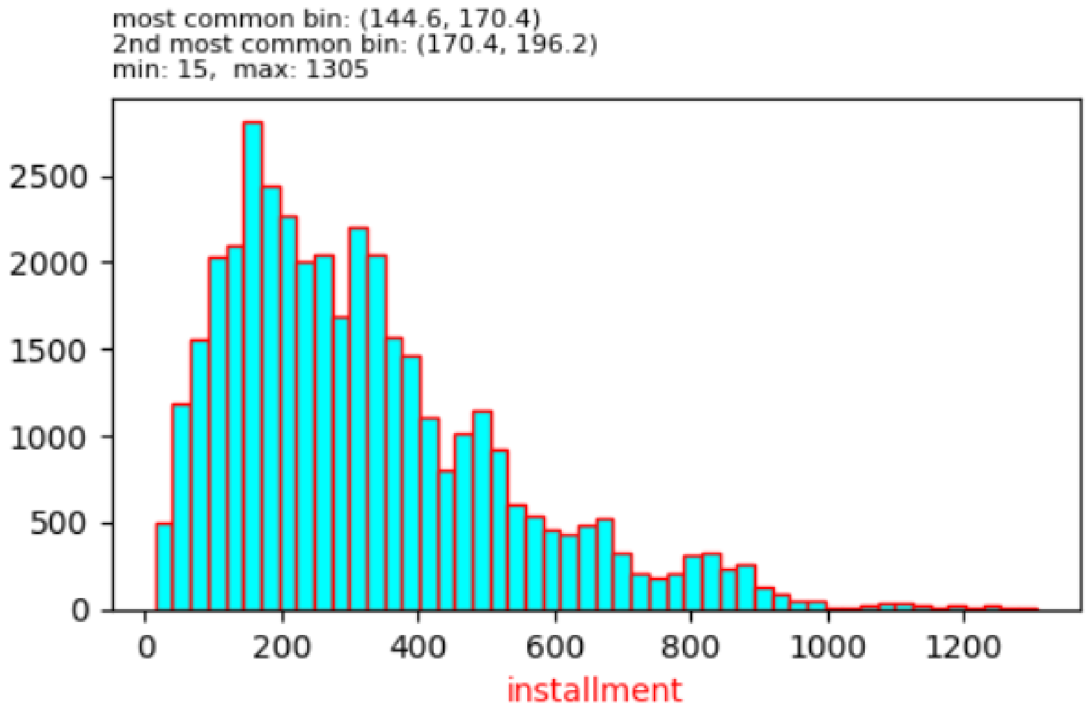
Find invalid values and replace them with missing values

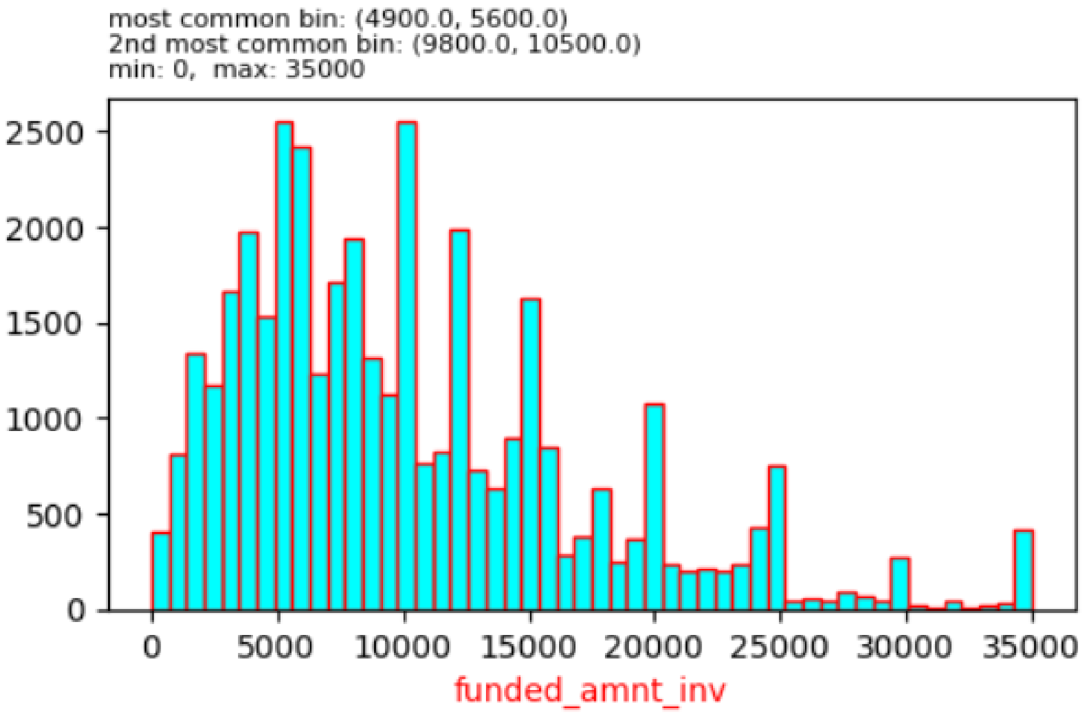
Remove suffix, leading and trailing spaces

Derive Type Driven metrics from dates

Merge grade and sub_grade columns to create a Data Driven metric used for analysis

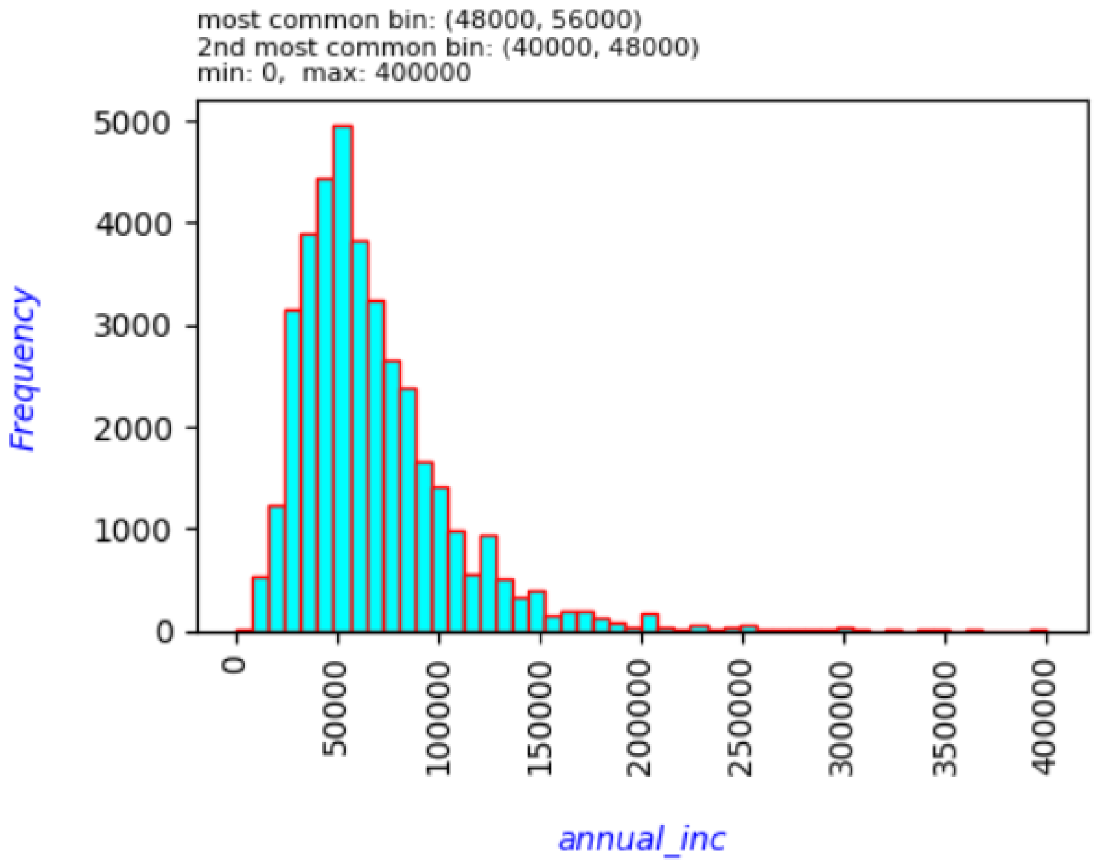
UNIVARIATE ANALYSIS ON QUANTITATIVE VARIABLES



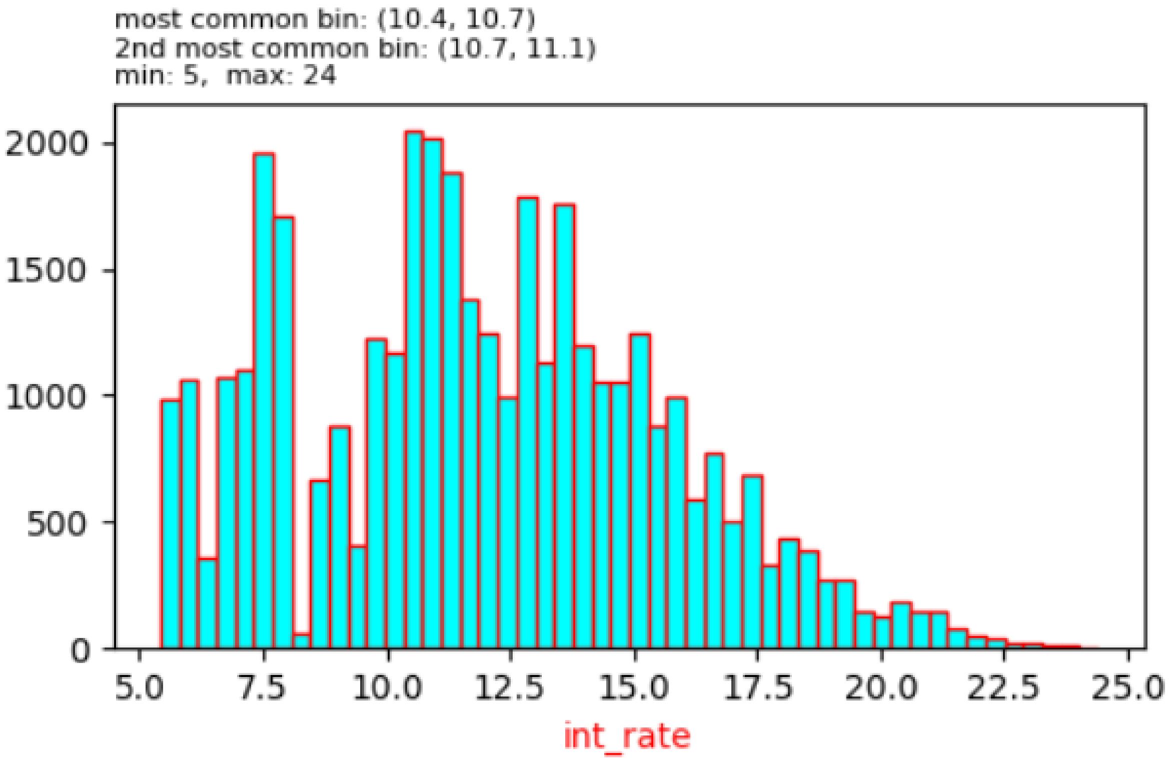


- The visuals display the min and max values for each variable as well as the most common bins, or places where the majority of similar values fall.
- The variables contain a few extreme values that contribute to the outliers.
- The majority of borrower **installments** have a value of less than 400.
- **funded_amnt_inv** - The majority of applicants are granted loan between 4,500 and 10,000.

UNIVARIATE ANALYSIS ON QUANTITATIVE VARIABLES



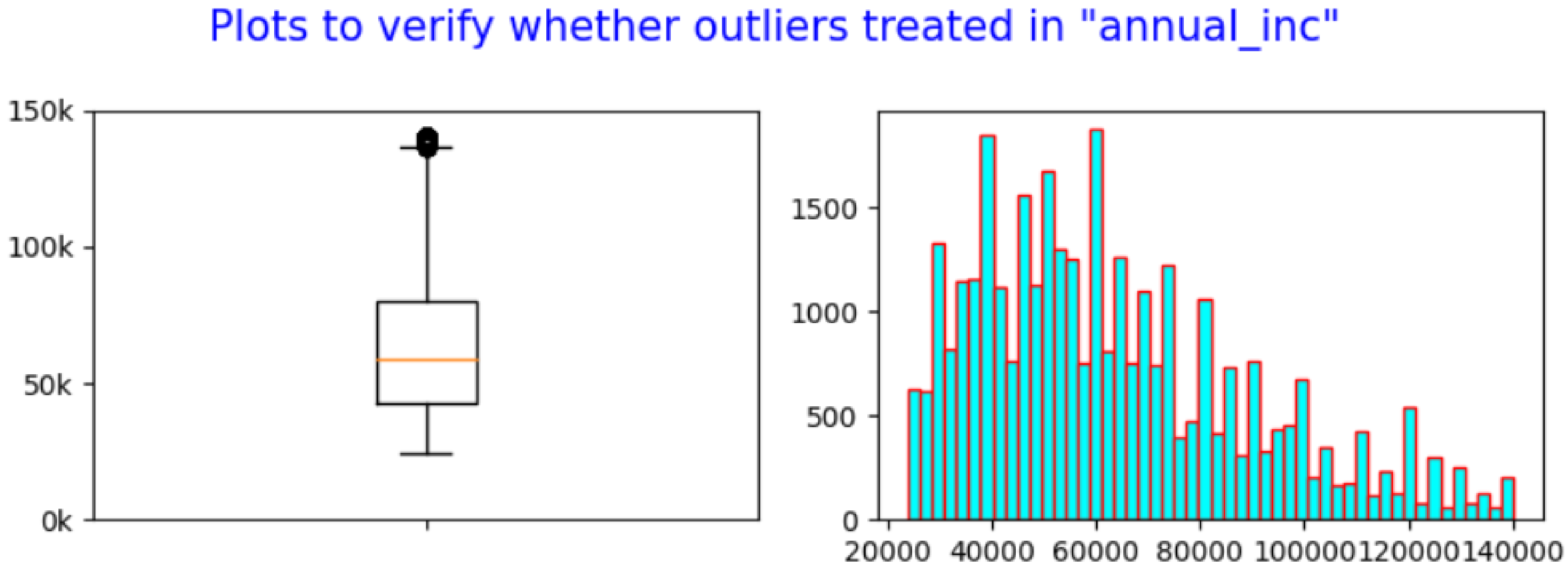
Most of the debtors make between 25,000 and 85,000 annually. The lender appears to get loan requests from low salaried individuals. It is heavily skewed; outliers are present.



Usually, **int_rate** is offered in the range of 6-8% or 10-15%. May be the case that the interest rate is higher for the borrowers with higher debt-to-income ratios. Perhaps the interest rate fluctuates periodically depending on the purpose of loan. It has extreme values indicating the presence of outliers.

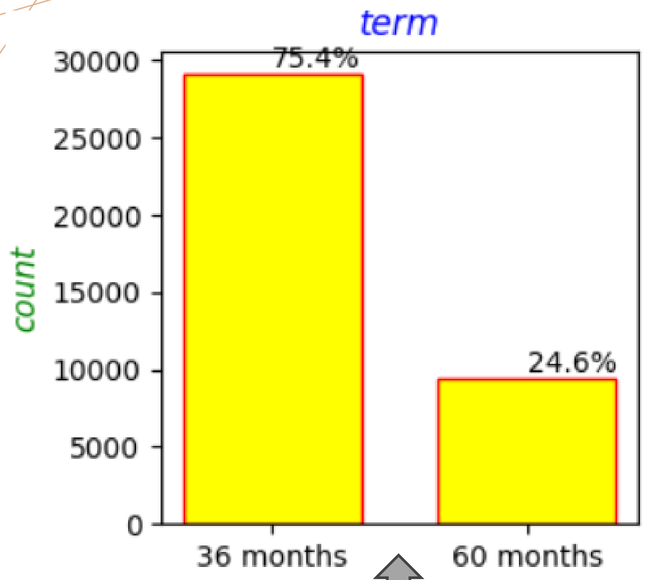
DETECT AND TREAT OUTLIERS

- **Let's impute the outliers as missing values for this case study.** Later we can decide how to remove these outliers before feeding the data to the model
- **Strategy to detect outliers**: Detecting outliers on the basis of the interquartile difference (q3-q1), may lead to removing large number of them, losing vital information from the dataset. Instead, determine the growth percentage of the consecutive quantiles to detect outliers.
  - We can compare the rate at which the values grow at the lower percentiles with those in the higher percentiles and if the higher percentile values grow much more than the lower percentile values, it will indicate presence of outliers
  - Example (for annual_inc, the outliers were successfully imputed with missing values):



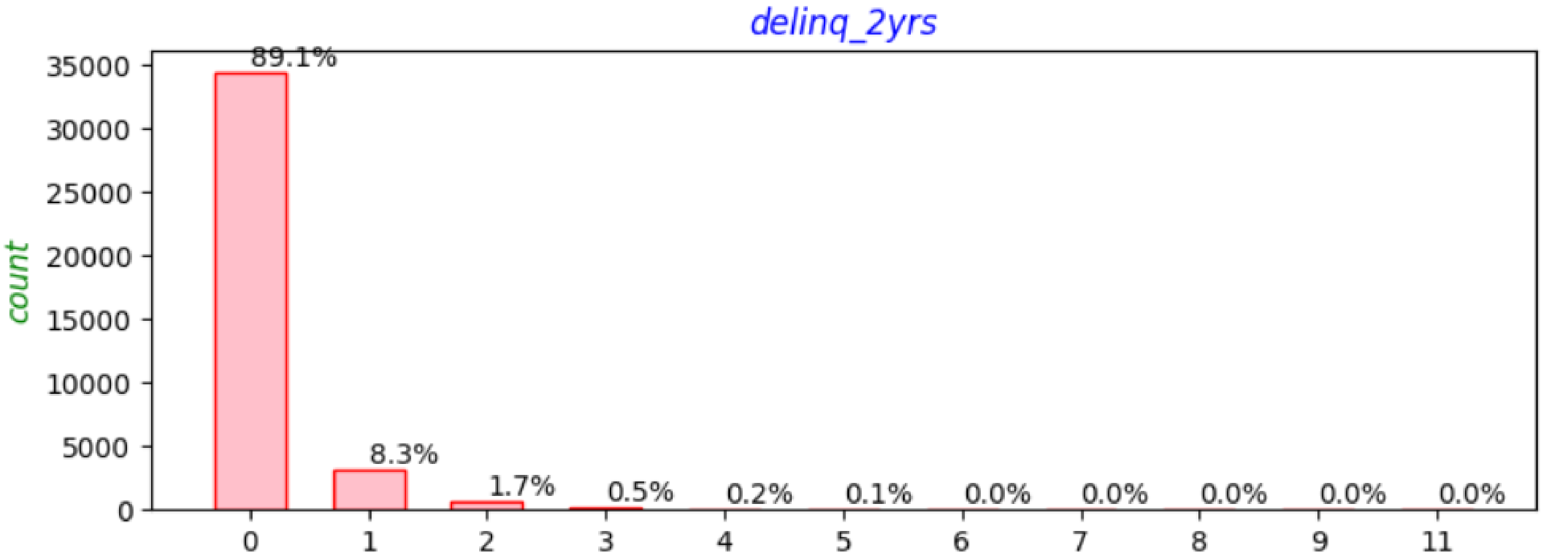
Plots to verify whether outliers treated in "annual_inc"

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES

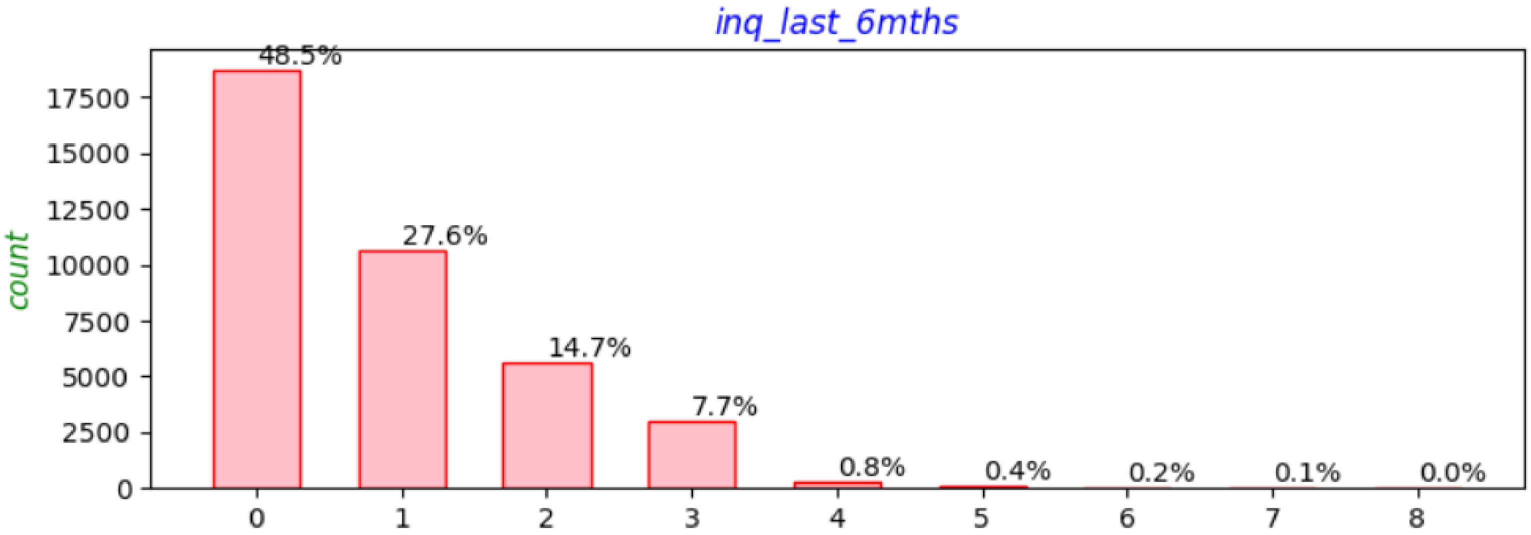


The maximum number of loans has a 36-month loan tenure (around 75%), which is more than twice as many as loans with a 60-month tenure.

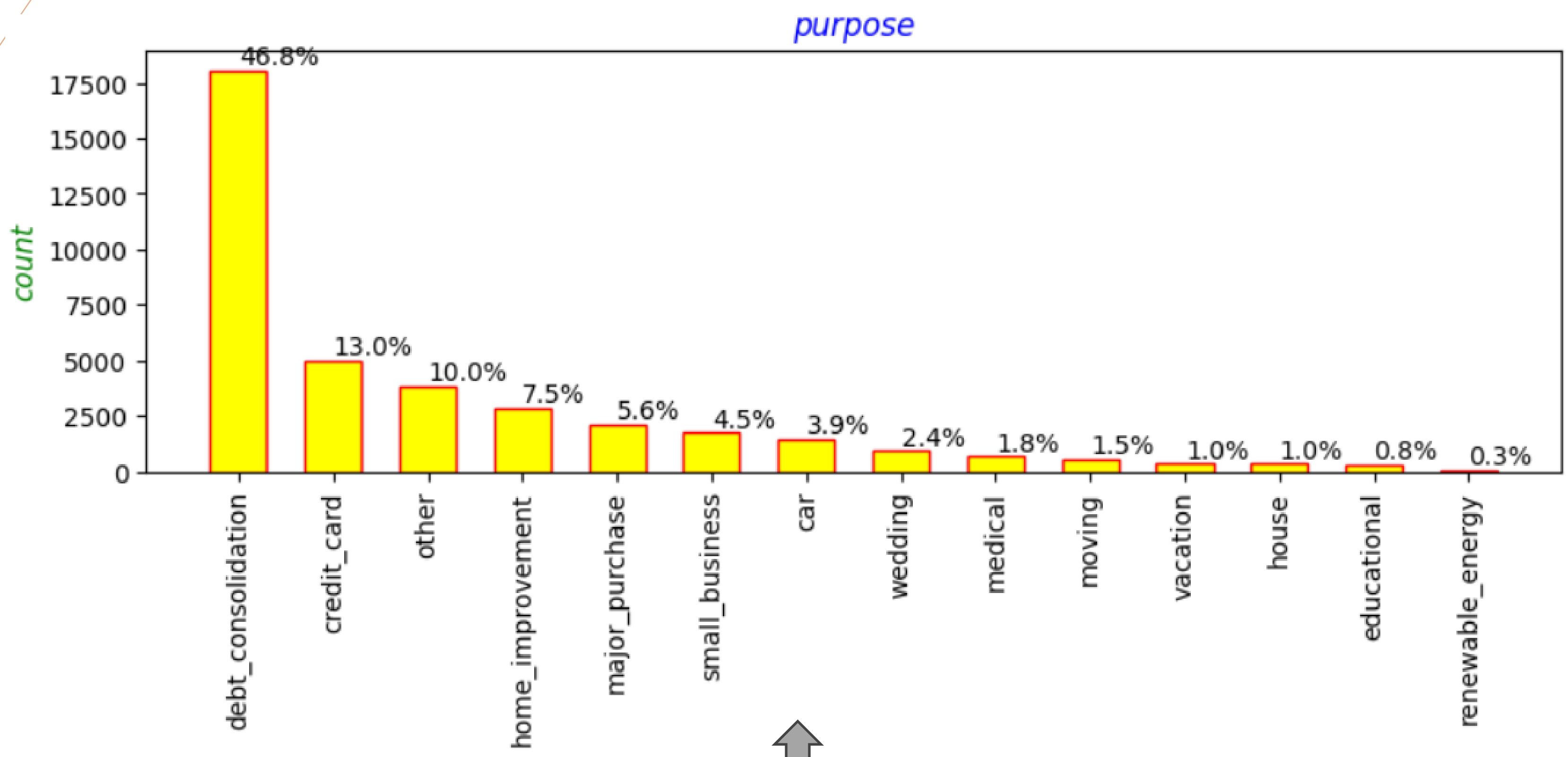
Around 90% of the borrowers have not miss their repayments in the last two years. Lender ensures lower risk of default



Borrowers who tend to apply for loans more frequently are also less likely to get approved for loans.

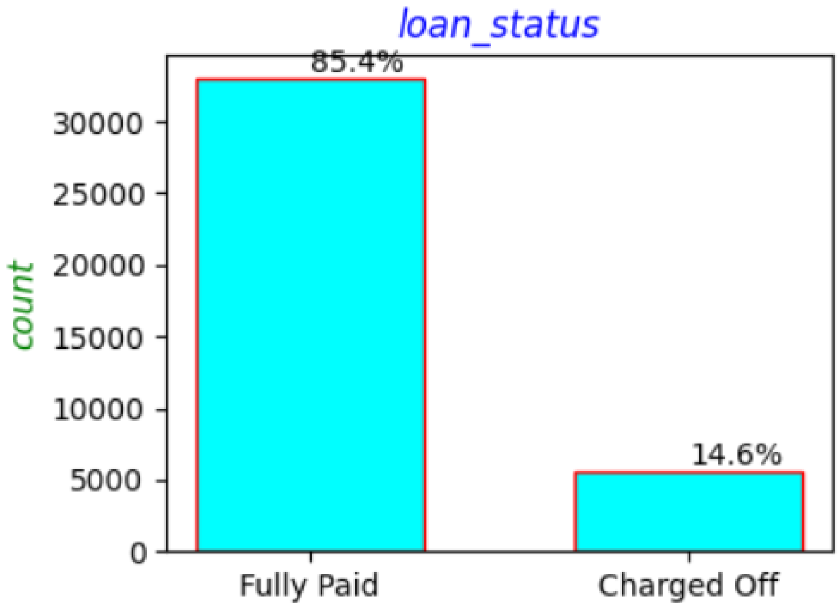


UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES

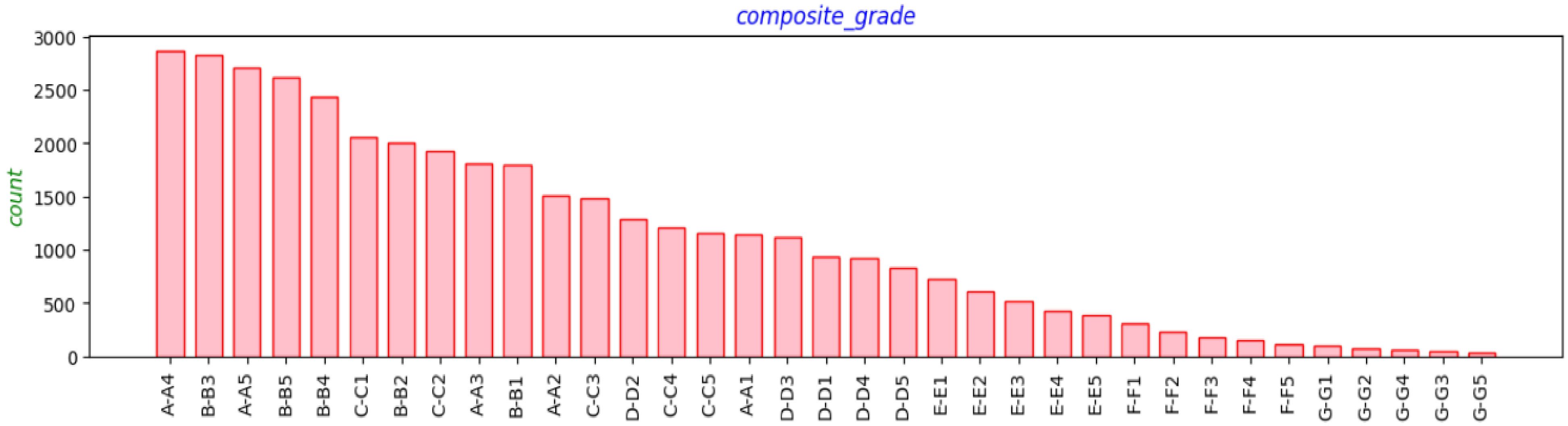


Among the loan accounts, the primary purpose is debt consolidation, which is contributing nearly 47% of the all the loan purposes, followed by credit card, which contributes to only 13%. Lender is taking high risk by approving loan for debt consolidation.

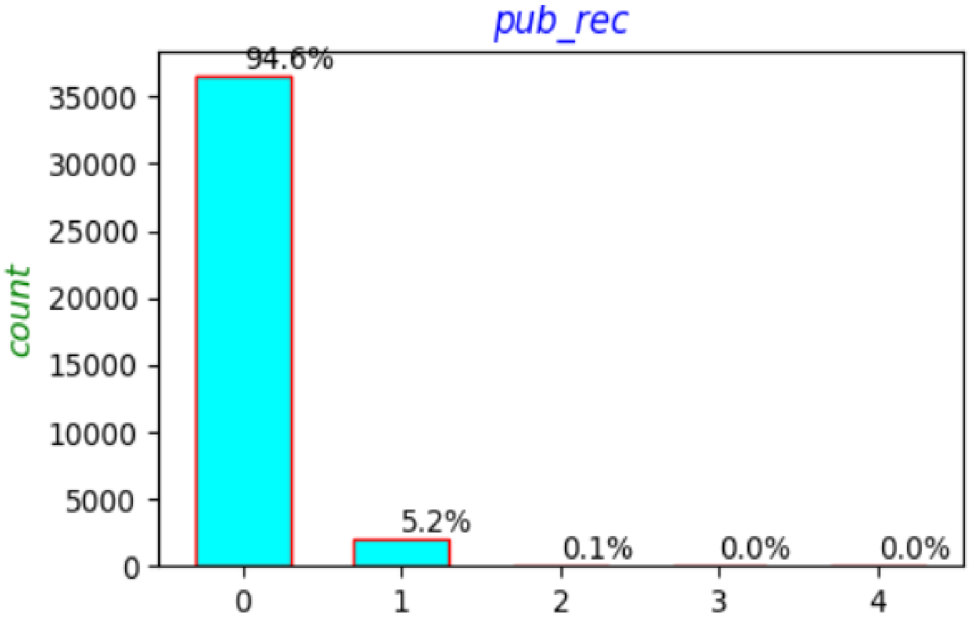
The number of fully paid loans exceeds the number of charged-off loans by nearly six times.



UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES

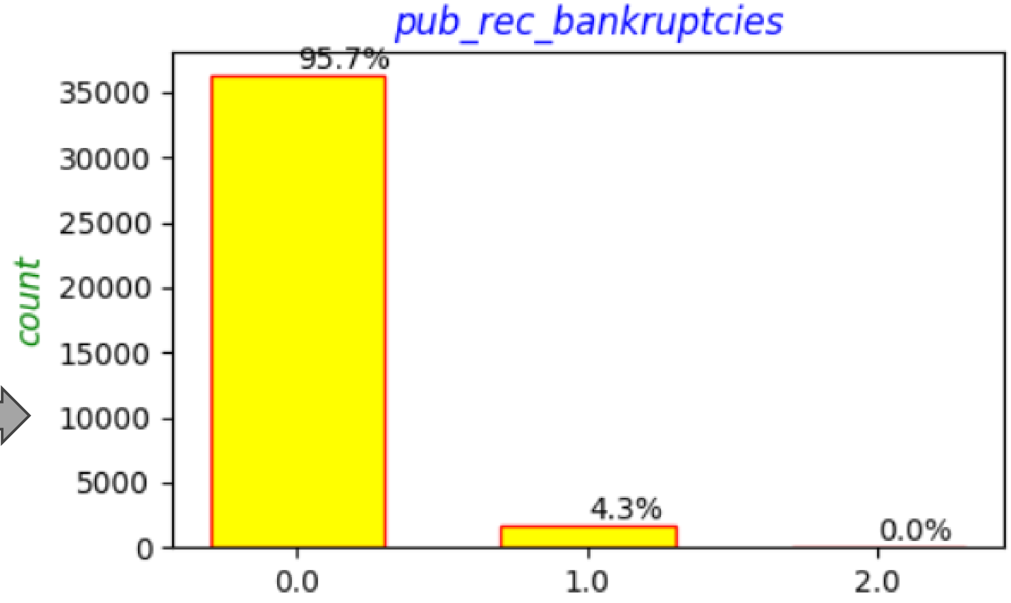


Majority of the borrowers have **composite grades** (grade + sub_grade) as A-A4, B-B3 and A-A5. Higher graded applicants are more likely to get loans from the lender

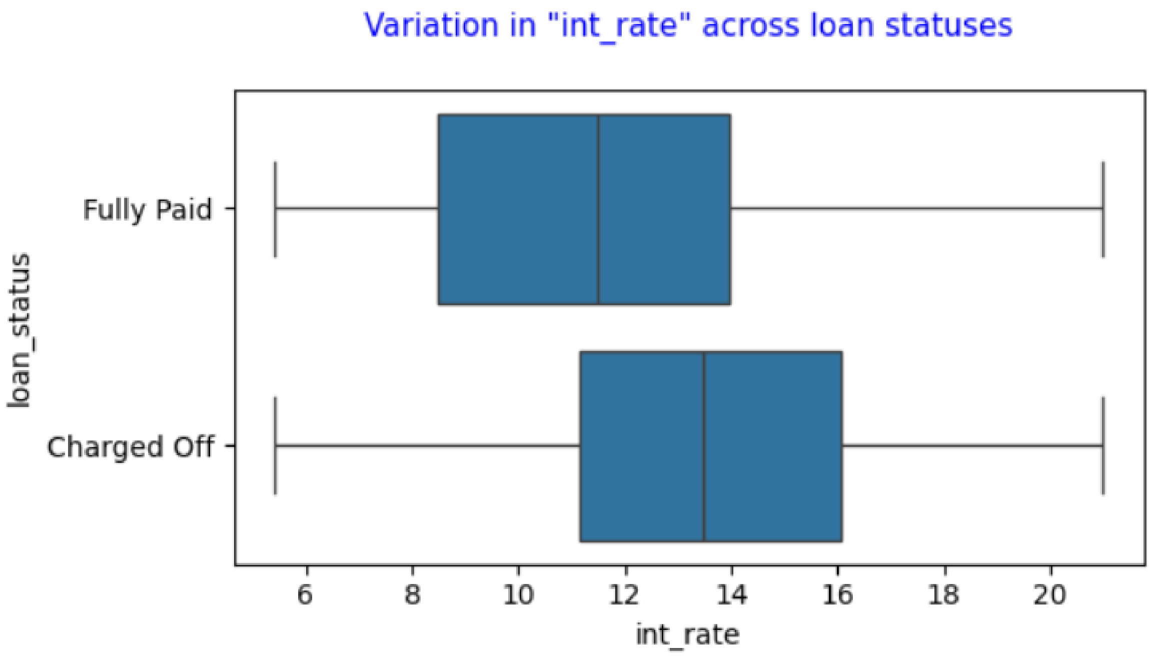


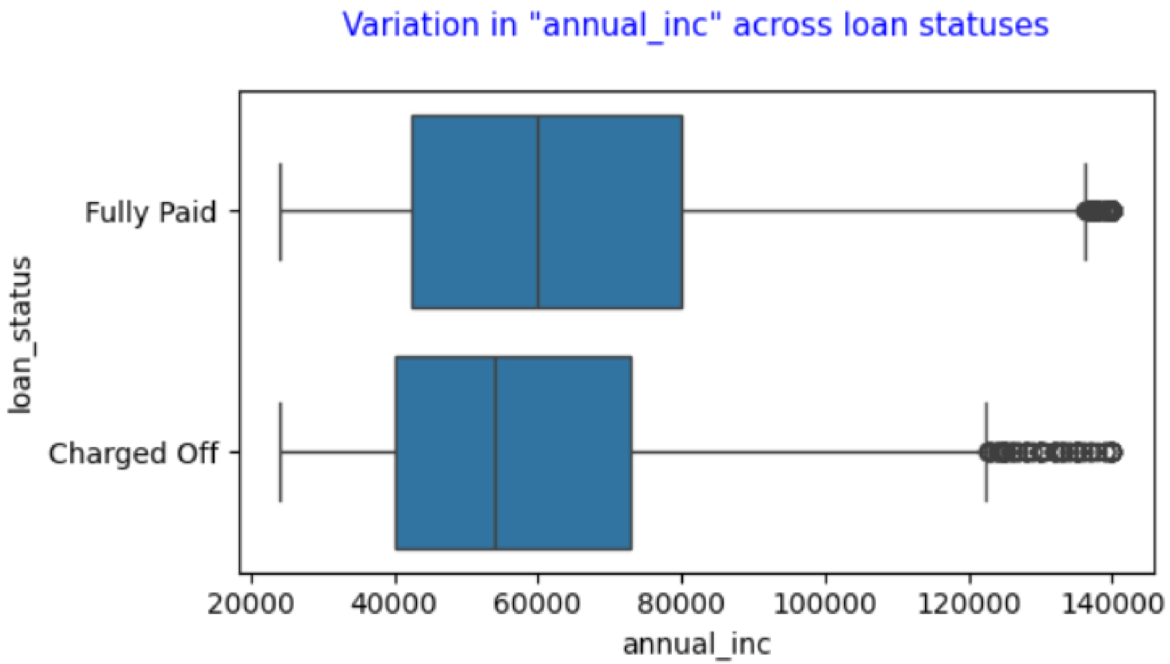
Less publicly available negative history makes an applicant more likely to be approved for a loan from the lender.

Approximately 95% of the borrowers have never filed for bankruptcy, making them a desirable option for the lender.



SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF NUMERICAL VARIABLES ON LOAN STATUSES

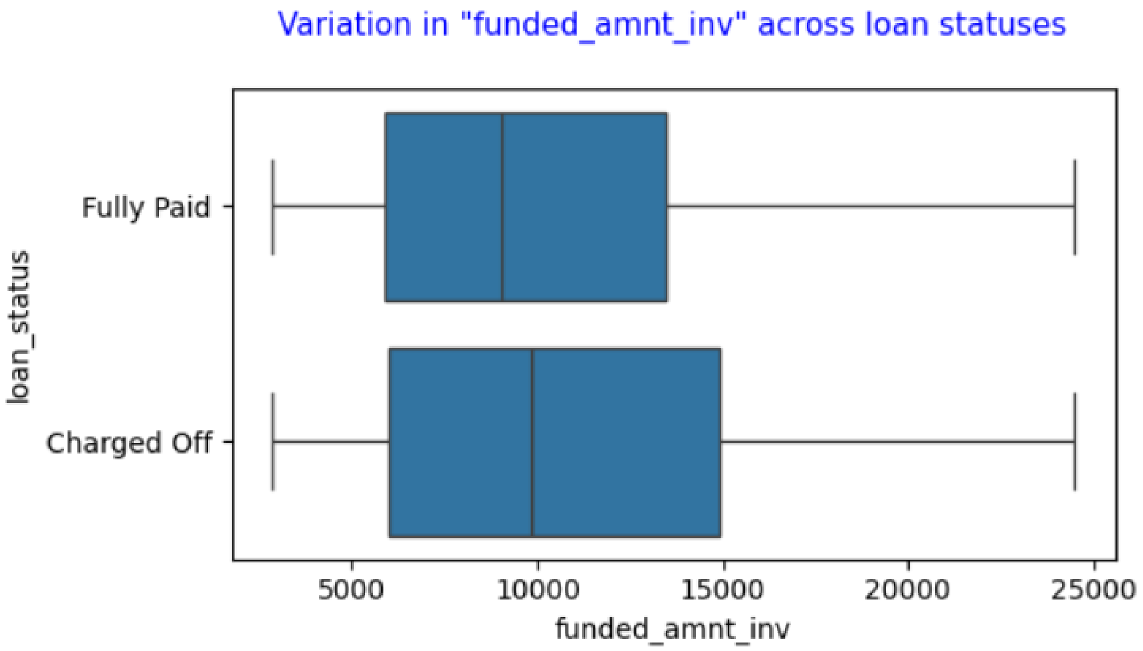


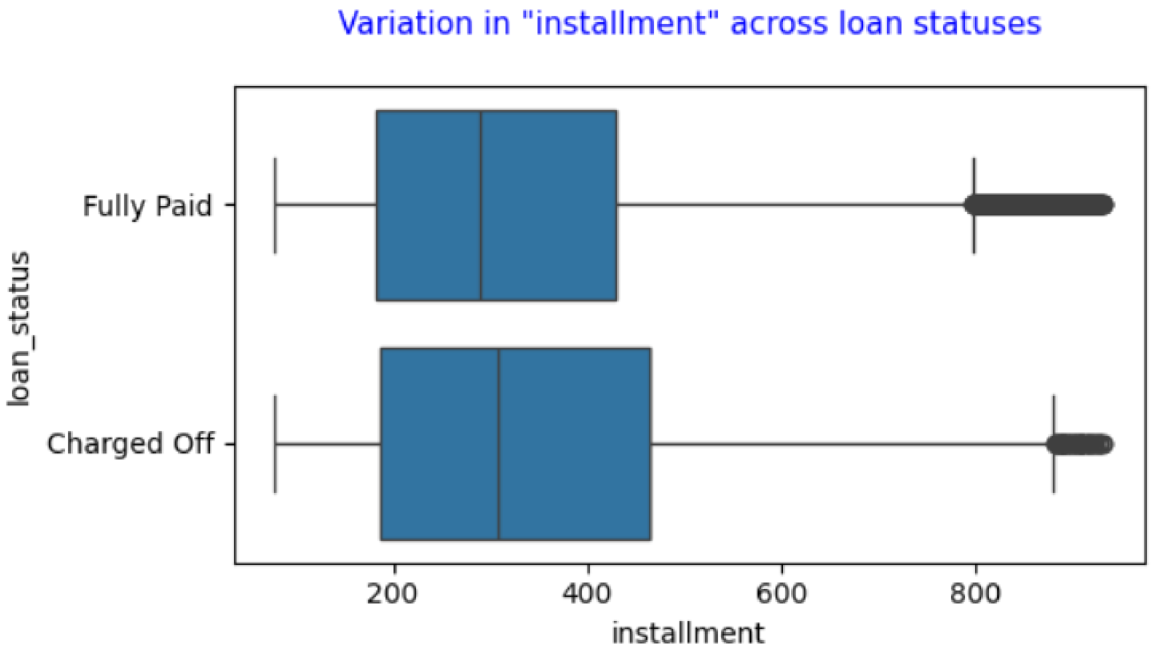


The interest rates of fully paid loans and defaulted loans differ noticeably, with the defaulted loan interest rate being greater on average. The riskier the borrower, the higher the loan interest rates.

- The annual income of loan defaulters at 75th percentile is lower than those who have fully paid back their loans.
- This shows that the borrowers with higher income are less likely to default.

SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF NUMERICAL VARIABLES ON LOAN STATUSES





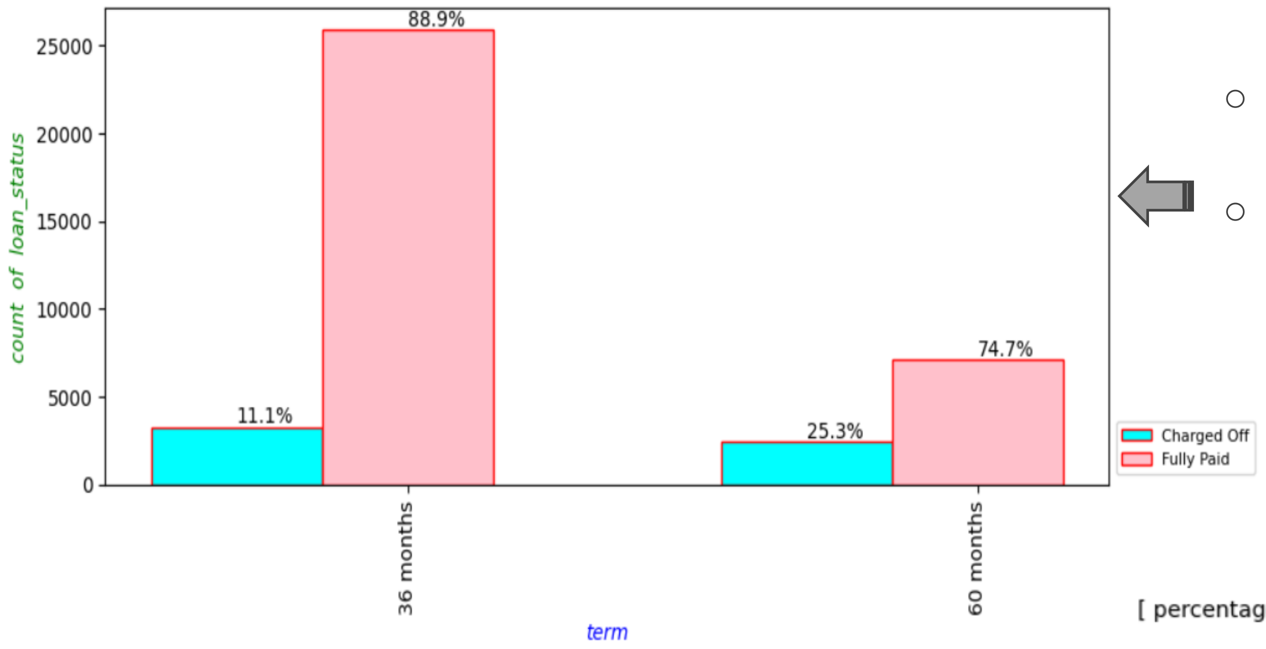
- The 75th percentile value of the funded amount from investors is higher in the defaulted loans compared to the loans which are already paid in full. This shows that large amount of loans are more likely to default
- it is evident that the funded amount from investors varies widely across the loan statuses. This led to a tiny difference in those funded amount averages among loan statuses, which may be caused by randomness.

- The 75th percentile installment is higher in charged-off loan account, indicating that loans with larger installments are more likely to default.
- Installment variable showing similar behavior as the funded amount from investor, indicating **high correlation** between them

SEGMENTED UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES TO ASSESS THEIR IMPACT ON LOAN STATUSES

Variability of "term" in relation to "loan_status"

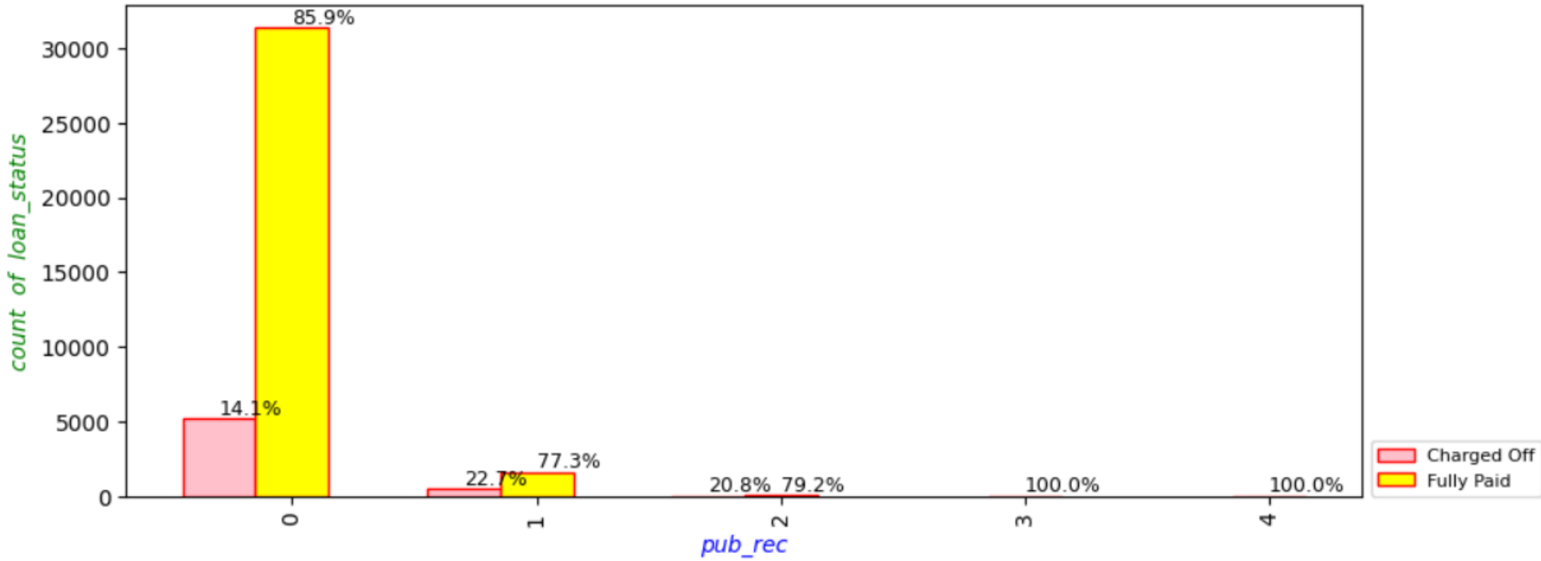
[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "term"]



- Borrowers are more likely to repay their loans successfully within 36-month timeframe.
- The percentage of loans that default within a 36-month period is less than half of the percentage of loans that default during a 60-month period.

Variability of "pub_rec" in relation to "loan_status"

[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "pub_rec"]

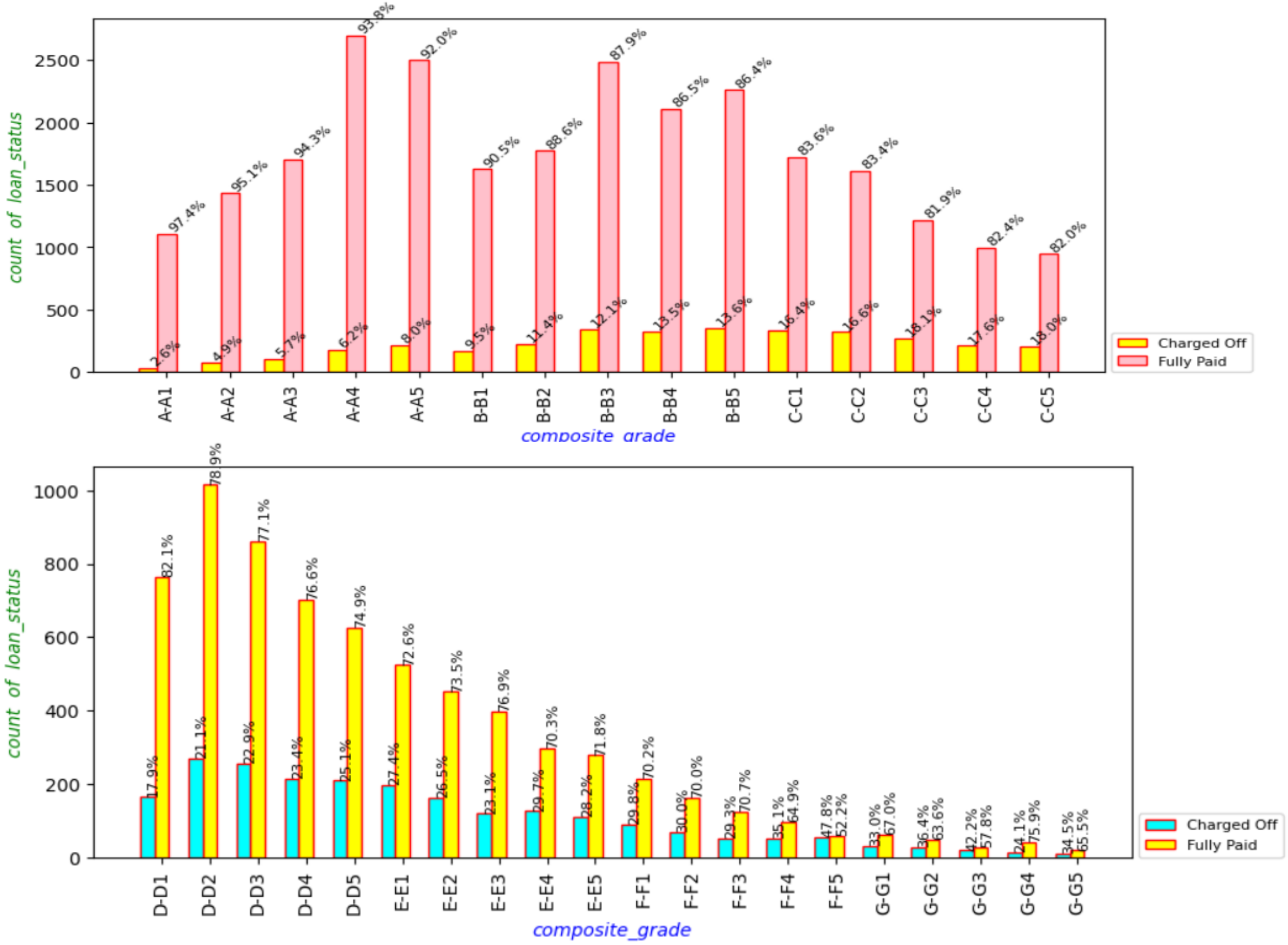


There is a greater percentage of defaults among borrowers who have at least one derogatory past than among those who have none, suggesting that borrowers with less recorded derogatories are probably going to repay their loans.

SEGMENTED UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES TO ASSESS THEIR IMPACT ON LOAN STATUSES

Variability of "composite_grade" in relation to "loan_status"

[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "composite_grade"]

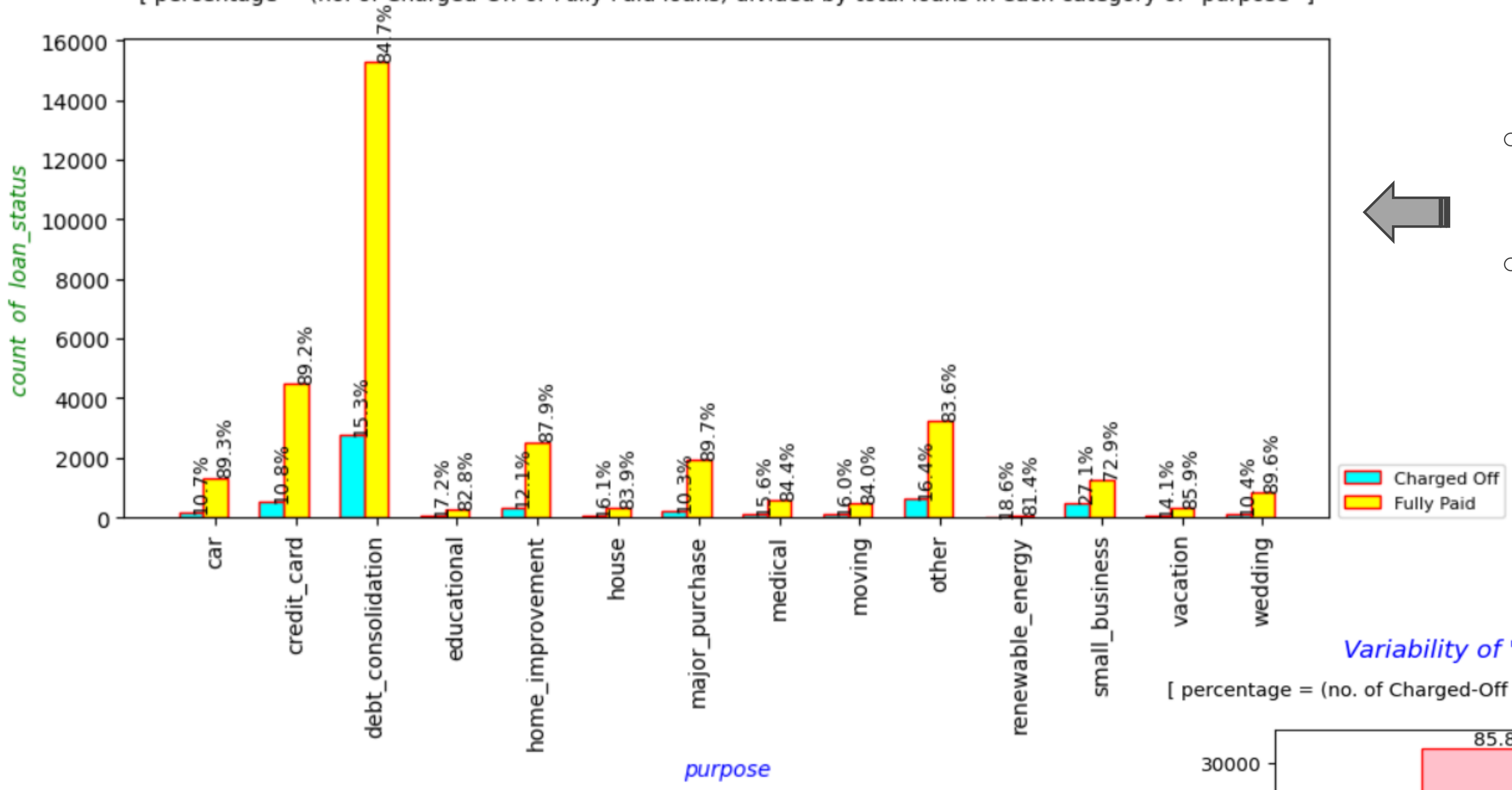


- We see a very less proportion of loan defaults among high sub-graded borrowers (such as A-A1, A-A2, A-A3) because they are more likely to repay their debts than low composite-graded borrowers.
- Thus, it appears from the data that borrowers with lower composite-grades have a lower likelihood of repaying their debts.

SEGMENTED UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES TO ASSESS THEIR IMPACT ON LOAN STATUSES

Variability of "purpose" in relation to "loan_status"

[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "purpose"]

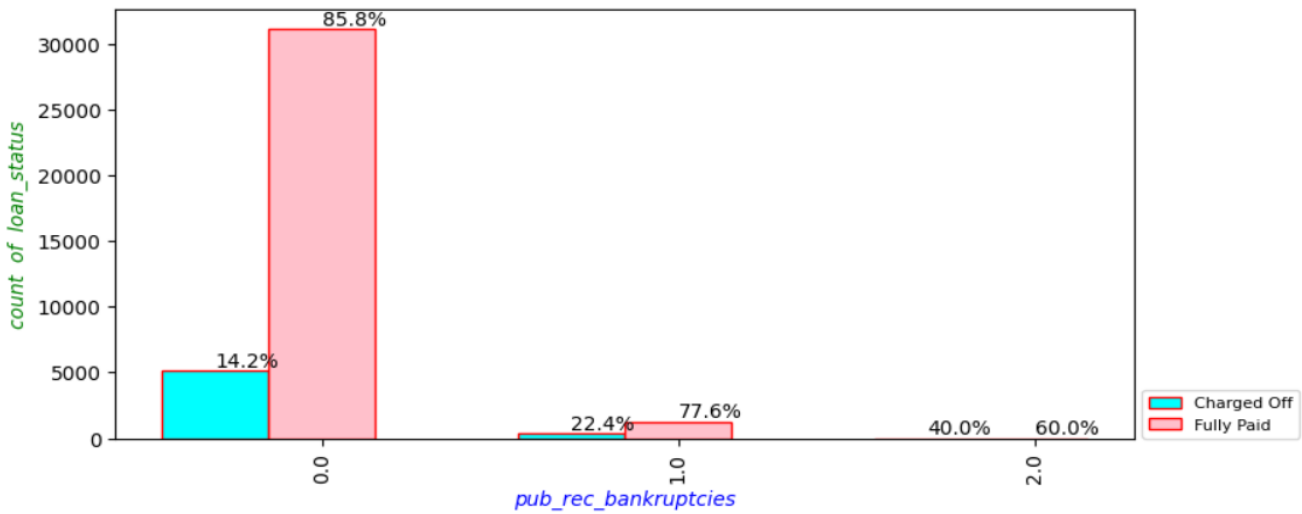


- The percentage of defaulters is higher when their purpose of taking loans is either Debt Consolidation, Educational, House, Medical, Moving, Other, Renewable Energy, Vacation.
- The majority of defaults occur when borrowers take **Small Business** loans, which makes up about 27% of all borrowers who take out these kinds of loans.

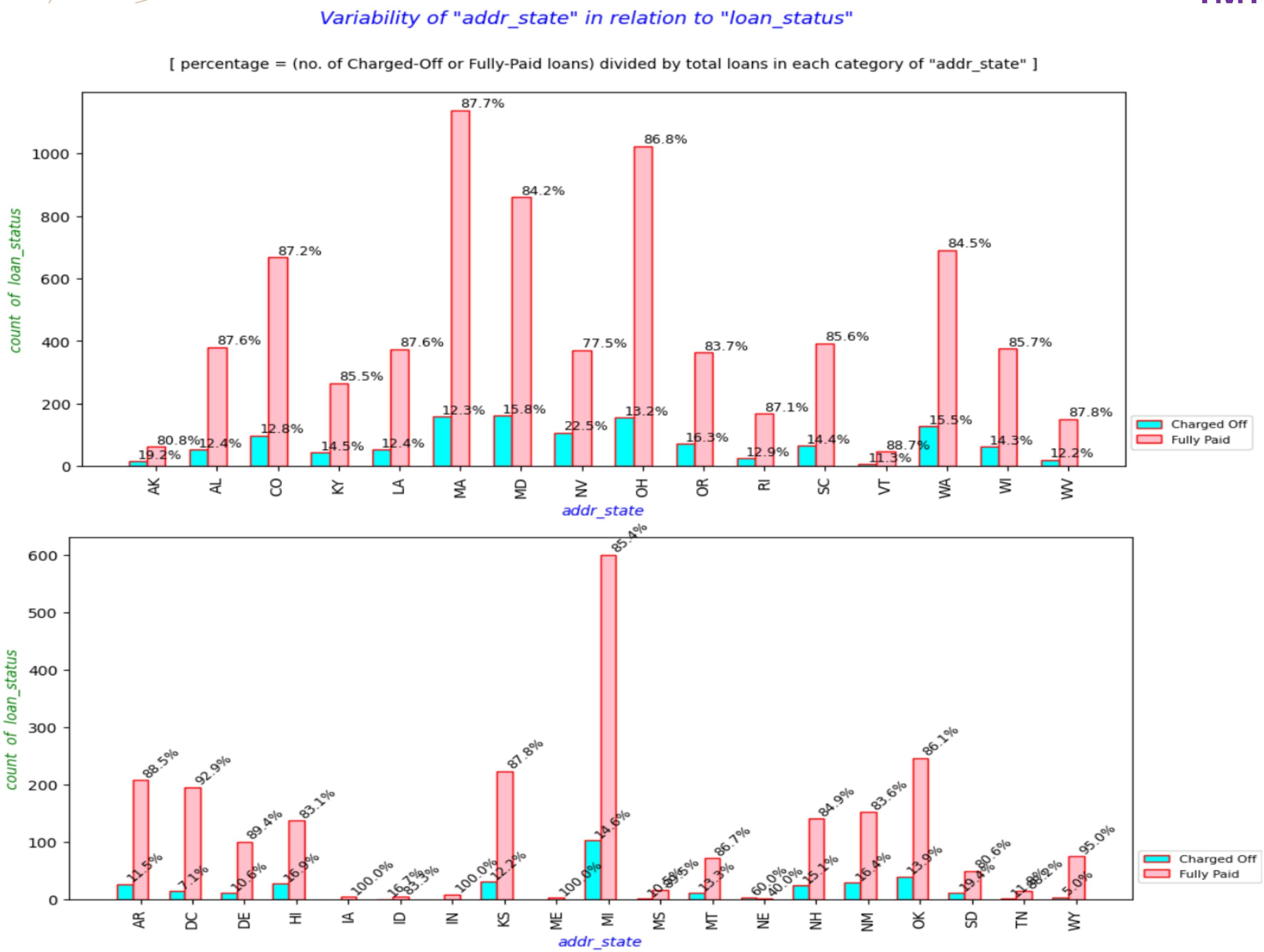
- We observe a lower percentage of loan defaults among borrowers who have never reported their bankruptcies, compared to borrowers who have reported bankruptcy atleast once.
- Thus, the data suggest that borrowers have lesser likelihood of repaying their loans who have reported bankruptcies once or twice earlier.

Variability of "pub_rec_bankruptcies" in relation to "loan_status"

[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "pub_rec_bankruptcies"]



SEGMENTED UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES TO ASSESS THEIR IMPACT ON LOAN STATUSES



Variability of "addr_state" in relation to "loan_status"

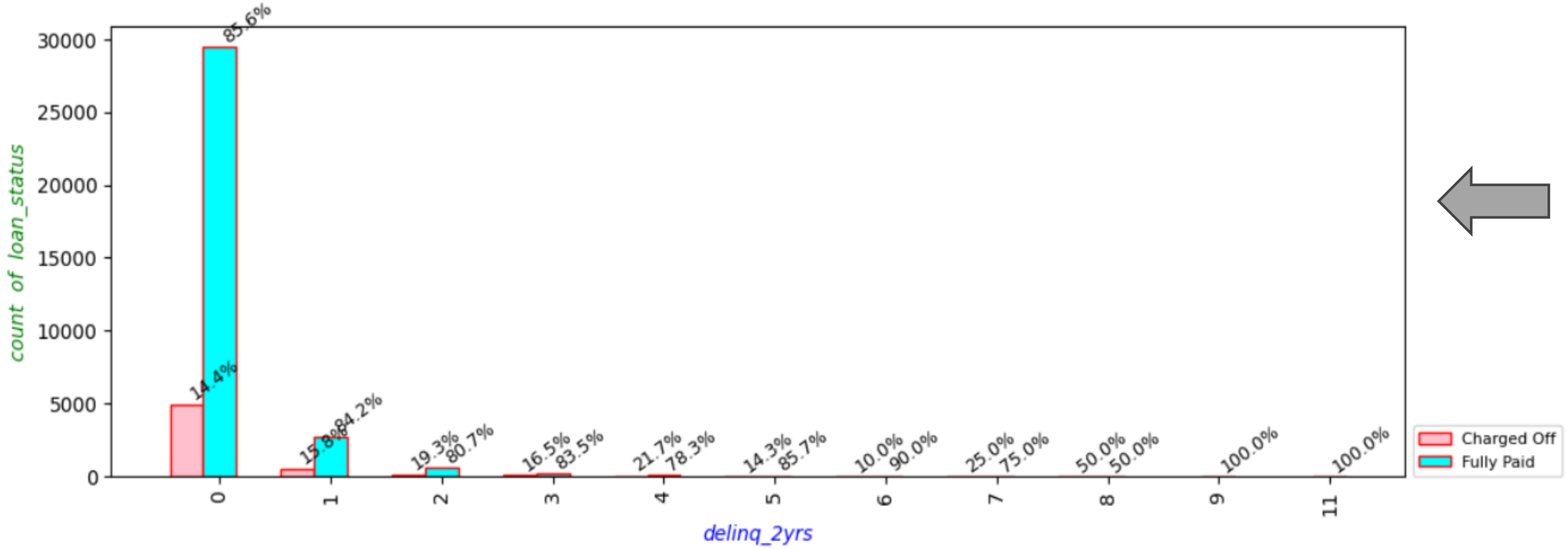
[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "addr_state"]

- ○ The observation suggests that there is a significant difference in the percentage of defaults among borrowers from different states.
- ○ From the state **ME**, all borrowers were in default. 60% of debtors in, NE have defaulted. Of the borrowers in **NV**, 23% have defaulted. The percentage of defaulters varies primarily between 10% and 20% among the US states.

SEGMENTED UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES TO ASSESS THEIR IMPACT ON LOAN STATUSES

Variability of "delinq_2yrs" in relation to "loan_status"

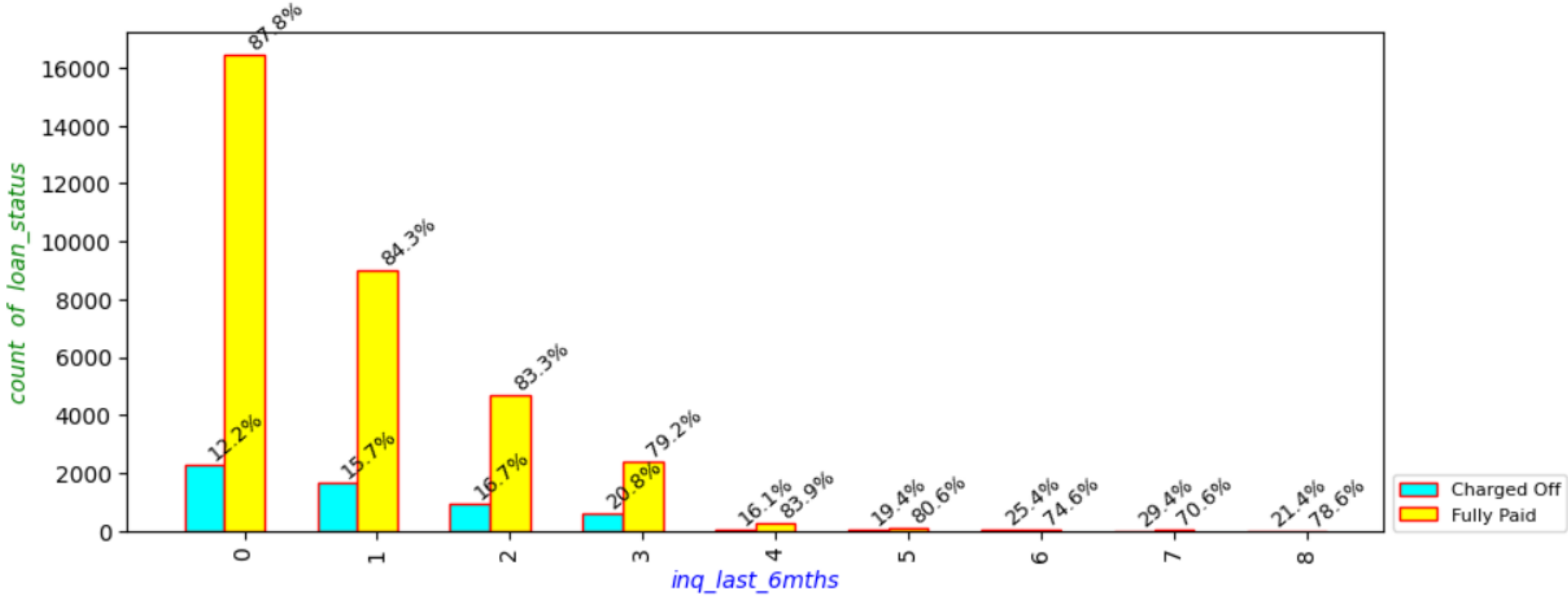
[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "delinq_2yrs"]



Borrowers who have not missed their loan repayments in the last two years are more likely to make loan repayments than those who have missed payments at least once in the previous two years and are therefore more likely to default. Thus, borrowers' loan default rates will increase in proportion to their delinquency.

Variability of "inq_last_6mths" in relation to "loan_status"

[percentage = (no. of Charged-Off or Fully-Paid loans) divided by total loans in each category of "inq_last_6mths"]

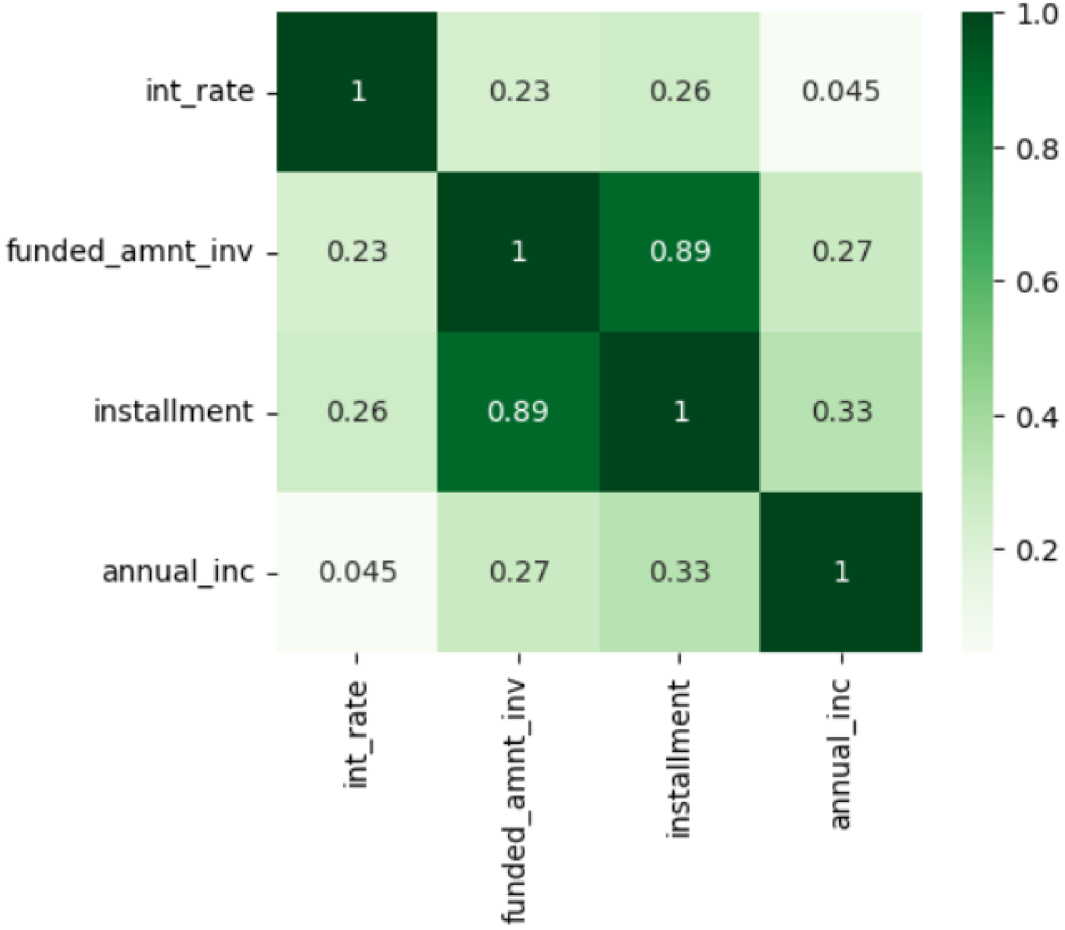


The likelihood of defaulting appears to be lower for borrowers who have not inquired about loans in the last six months than for those who did.

BIVARIATE ANALYSIS ON NUMERICAL VARIABLES

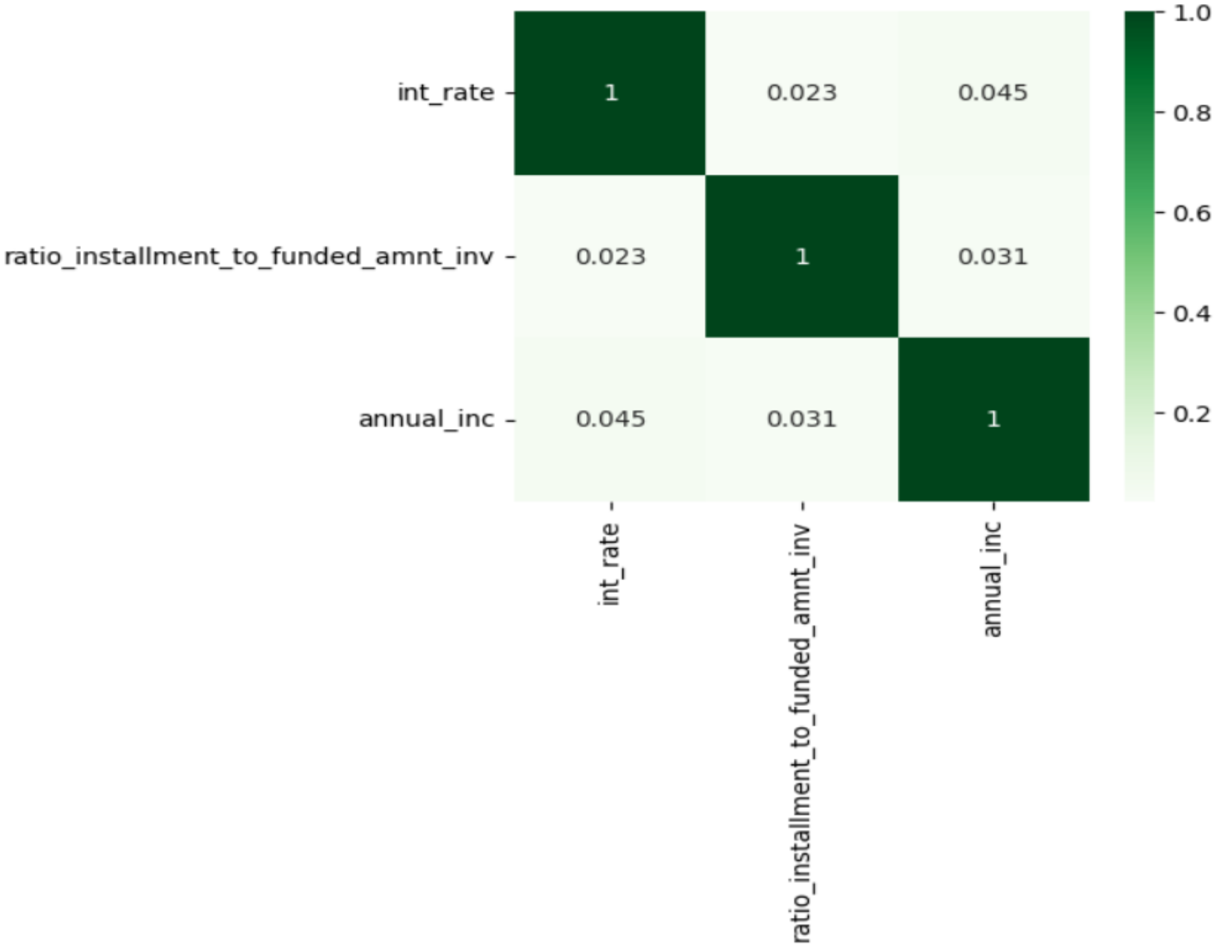
installment and **funded_amnt_inv** are highly correlated variables.

Correlation between numerical variables
(with higher potential to identify loan defaulters)



Derived a Data Driven Metric as the ratio of the loan installment to the amount that is funded by the investor (funded_amnt_inv) for each borrower to take out the correlation effect which will prevent the model from overfitting.

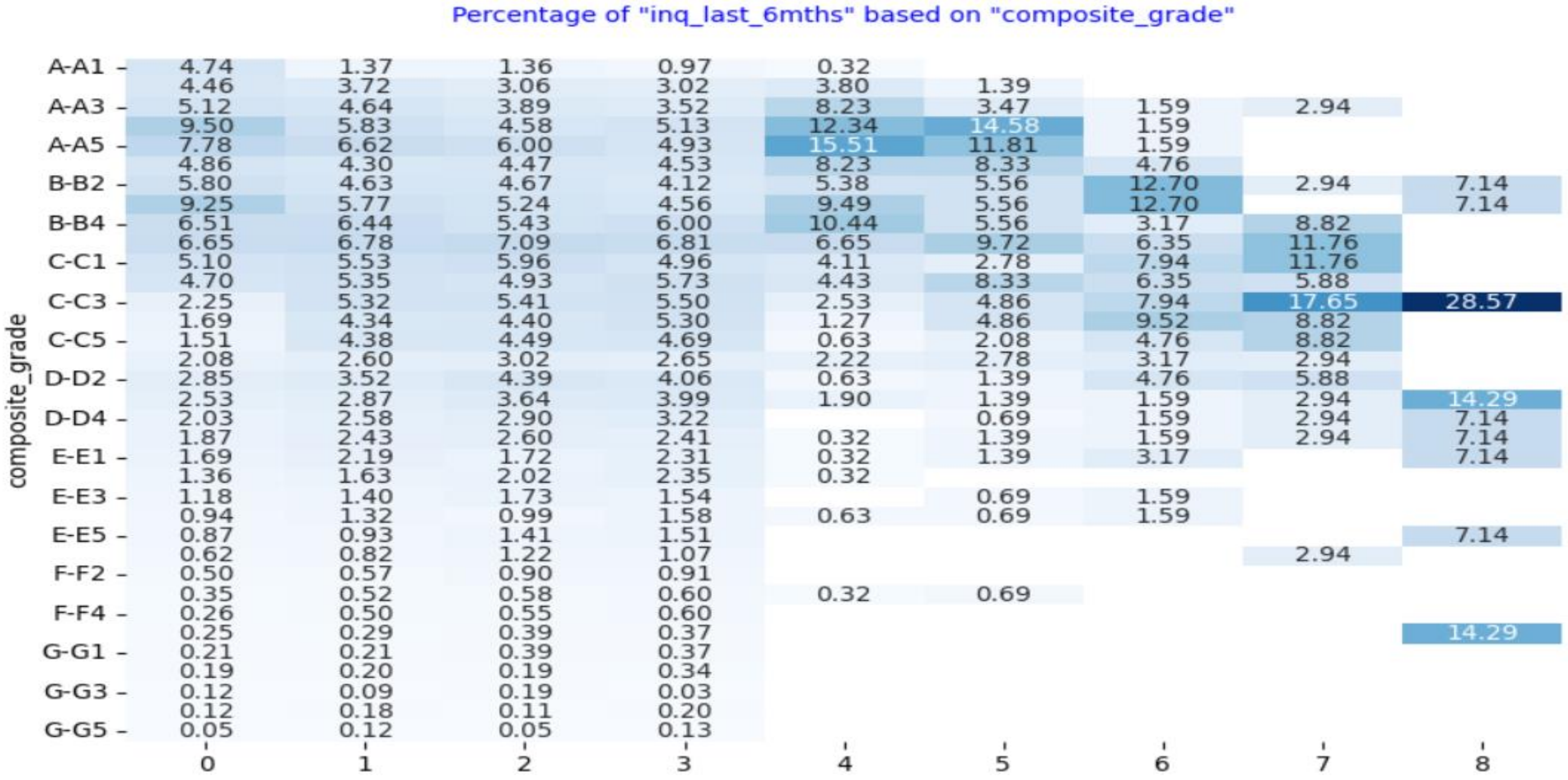
Correlation between numerical variables
(with higher potential to identify loan defaulters)



BIVARIATE ANALYSIS ON CATEGORICAL VARIABLES

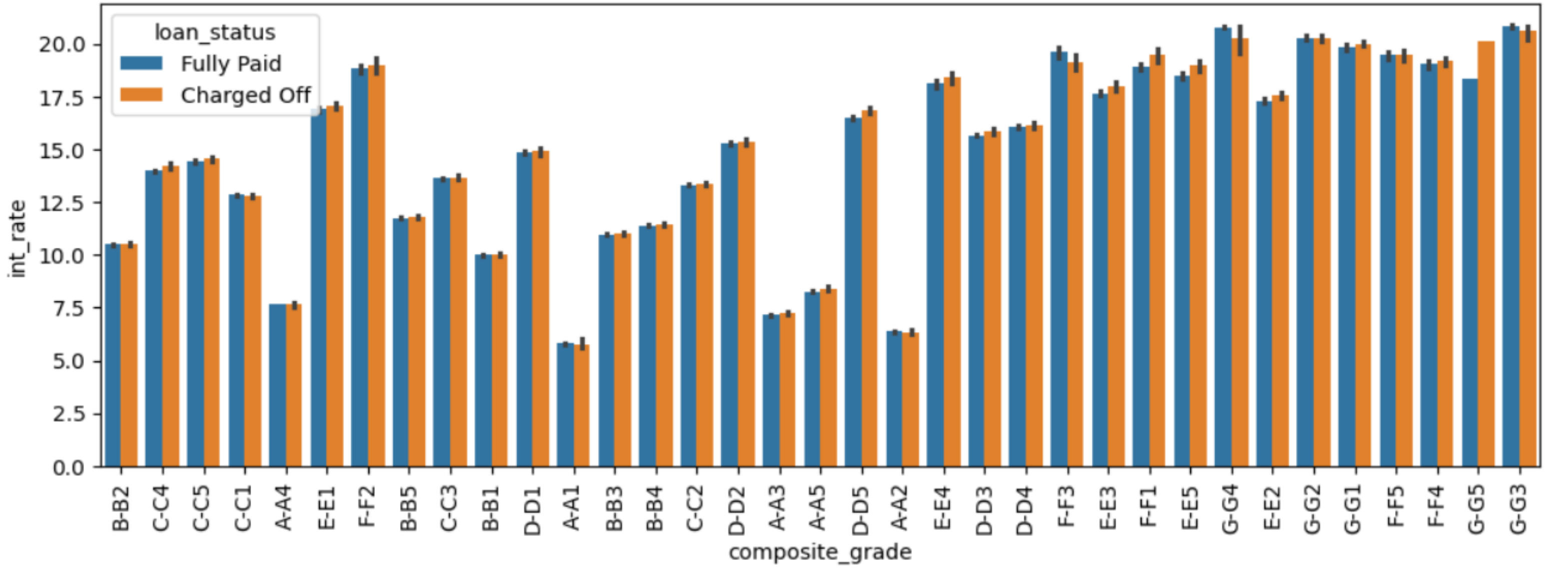
- inq_last_6mths and composite_grade are highly correlated.

- It appears that there is a correlation between these variables because the percentage of borrowers who are at the higher composite grades than D-D2 differ significantly among the number of the inquires they make in past 6 months.

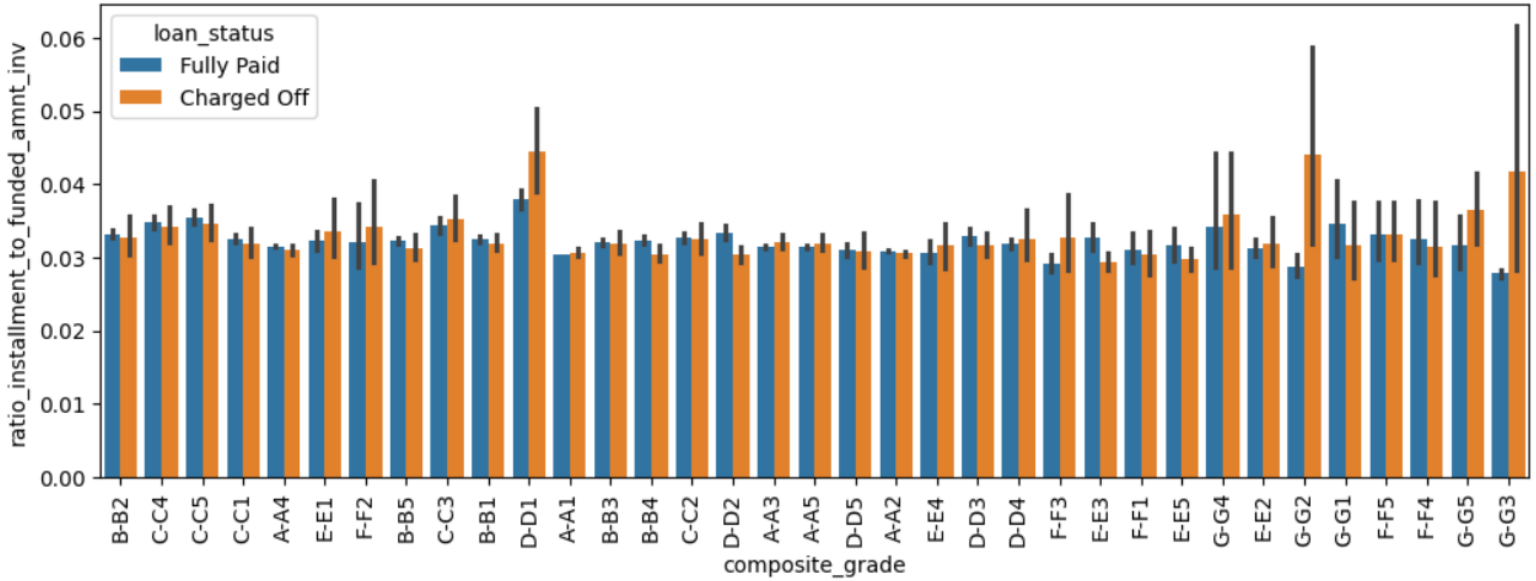


Percentage of "inq_last_6mths" based on "composite_grade"

BIVARIATE ANALYSIS TO ASSESS THE RELATIONSHIP BETWEEN CATEGORICAL AND NUMERICAL VARIABLES



- The interest is higher for borrowers who have lower composite grades (eg: G-G3, G-G5, G-G4, F-F3 etc.)
- Although loan interest rates fluctuate over the composite grades, they are less likely to differ between loan statuses for each composite grade. Hence we can conclude that int_rate is highly correlated to composite_grade of the borrowers.



- In addition to showing minimal variation among composite grades, the installment to investor financed amount ratio is also less likely to vary between loan statuses for the majority of composite grades except G-G2, G-G3 & D-D1.
- Hence, ratio_installment_to_funded_amnt_inv is highly correlated to composite_grade of the borrowers.

TOP INFLUENCERS

int_rate
term
composite_grade (merging grade and sub_grade variables)
ratio_installment_to_funded_amnt_inv (ratio of Installment to funded_amnt_inv as these are correlated)
annual_inc
purpose
addr_state
delinq_2yrs
inq_last_6mths
pub_rec
pub_rec_bankruptcies
open_acc (*moderate impact*)

CORRELATED VARIABLES

Installment and funded_amnt_inv
composite_grade and term
composite_grade and last_inq_6mths
term and last_inq_6mths
composite_grade and int_rate
ratio_installment_to_funded_amnt_inv and composite_grade
annual_inc and composite_grade
term and purpose (*moderately correlated*)
term and delinq_2yrs (*moderately correlated*)