Lending Club Case Study

BY RITIK BOBADE

Problem Statement

- You work for a **consumer finance company** which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are 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

- The data given contains information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Steps for Analysis

1. Reading the data set.
2. Data Cleaning
3. Data Standardization
4. Outlier Treatment
5. Univariate Analysis
6. Bivariate Analysis

Data Cleaning

- In data cleaning, we first check if there are any null values present in the dataset.
- After checking we found that 54 columns that have null values in them so it is of no use in the analysis so we dropped these columns from the dataset.
- Further, there were some columns which were having just a single value in them so these columns could not contribute anything to the analysis so we dropped these columns too.
- Now in the left column, some columns are seen after approval of loan and as we doing these studies to help before approving the loan so we can neglect such columns and some of them don't contribute much in the analysis dropped these types of columns.

E.g. delinq_2yrs, revol_bal , out_prncp , total_pymnt , total_rec_prncp , total_rec_int

Data Standardization

- There are some columns that can be use in analysis with categorical variables.
- Some of these columns are not in numerical format and have some extra characters in them that we don’t need so we removed these extra characters and converted the data to numeric format.

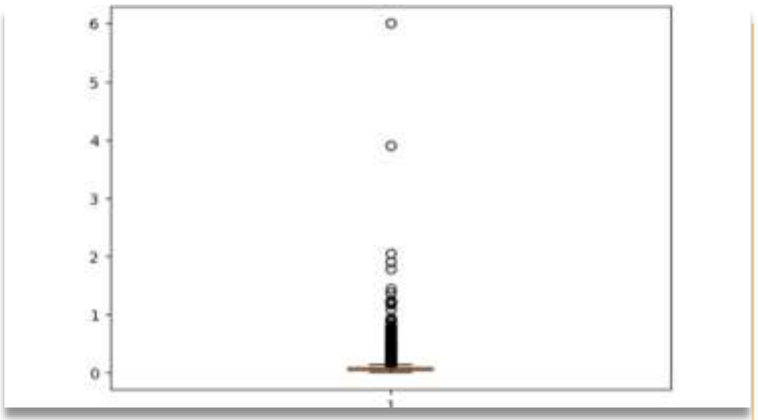
e.g interest rate, revol util,etc.

- Now there may be some missing values in the data set but that column can not be dropped so we need to fill in the missing values.
- There are 3 ways to fill in the missing values mean, mode, and standard deviation.
- We had to check for all three and after checking we found out that mode frequency in much Higer than the most frequent value.
- So we can fill the missing values using mode.

Outlier Treatment

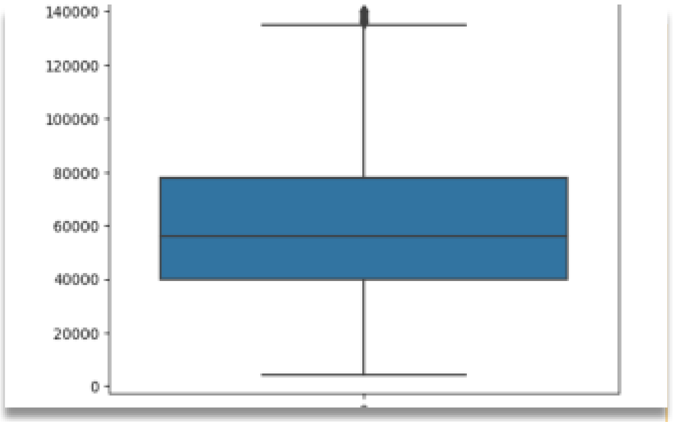
- An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.
- We can check if there are outliers are present we can plot a box plot for that particular column where outliers maybe present and that column is significantly important.
- We will plotted the box plot for annual income , dti, loan amount.

Box plot for Outliers



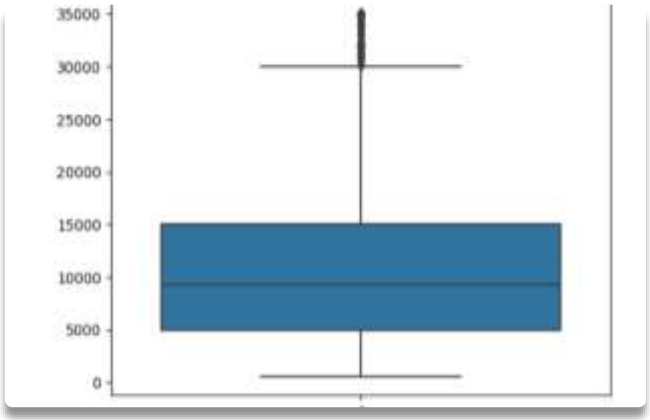
Annual Income

We can clearly see there are outliers present after 95 percentile. So we will have to remove the values after 95 percentile.



Dti

There are outliers present but the data is continuous so don't need to worry it won't affect the analysis.

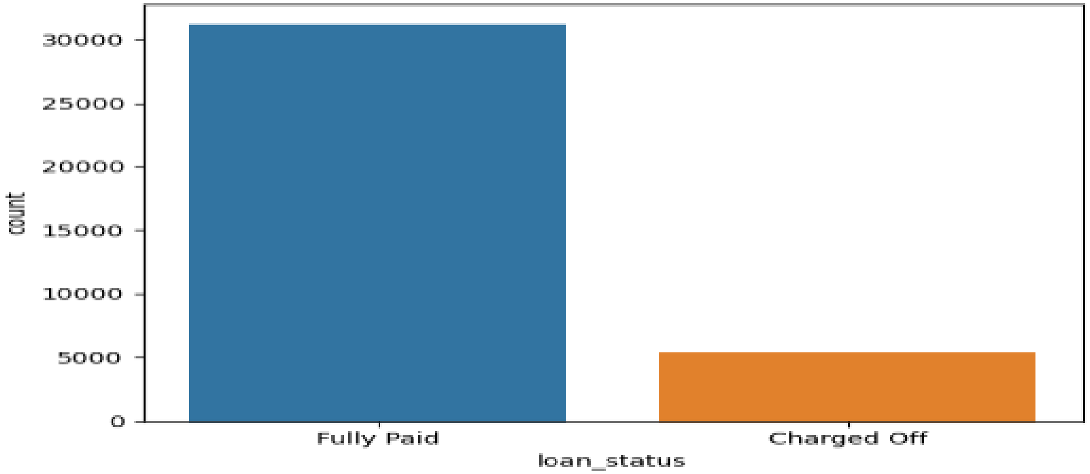


Loan Amount

There are outliers present but the data is continuous so don't need to worry it won't affect the analysis.

Univariate Analysis

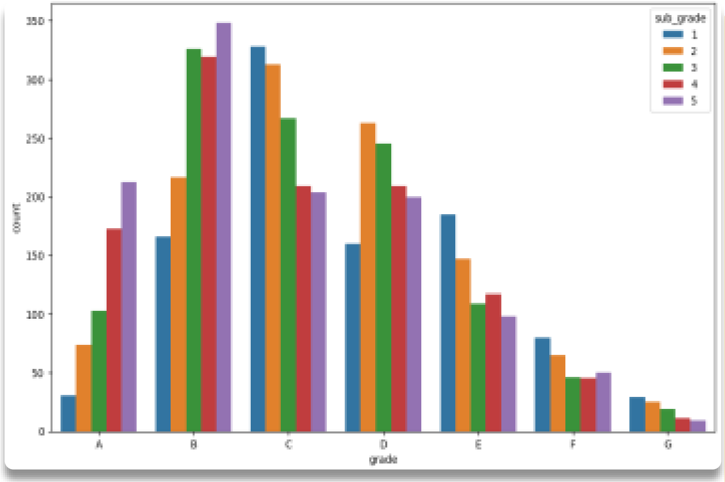
- In this analysis we will analyze loan status column .
- ‘loan_status ’ column has three entries in it i.e. Fully Paid,Current and Charged-off.
- So current loan is something we can not do anything with it, as it is in progress not completed nor set as charged off i.e. default.
- We compared fully paid and charged off loan by plotting a count plot.



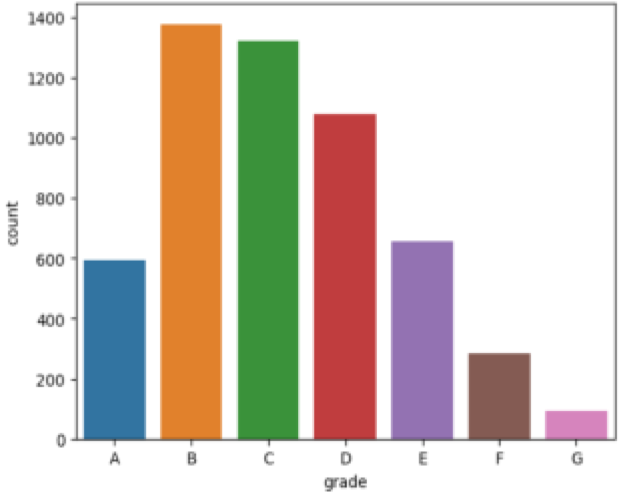
Bivariate Analysis
Visualization of Categorical Variables

- From the above univariate analysis, it is clear that a fully paid loan has a higher value and its charge of loan is less in number.
- So we focused on charged loans and compared it with different columns present in the dataset.
- We plotted a graph comparing each column separately with loan status i.e. charged off loans

The preceding examination regarding charged off loans for each variable indicates the following: There is a higher likelihood of defaulting when

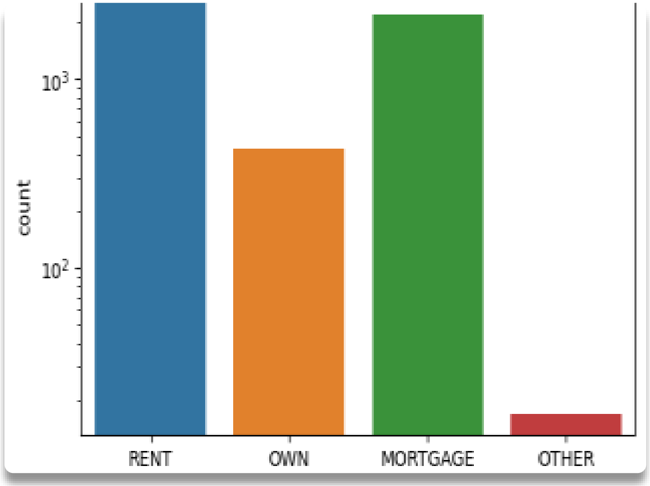


Sub-grade vs Charged off loans



Grade vs Charged off loans

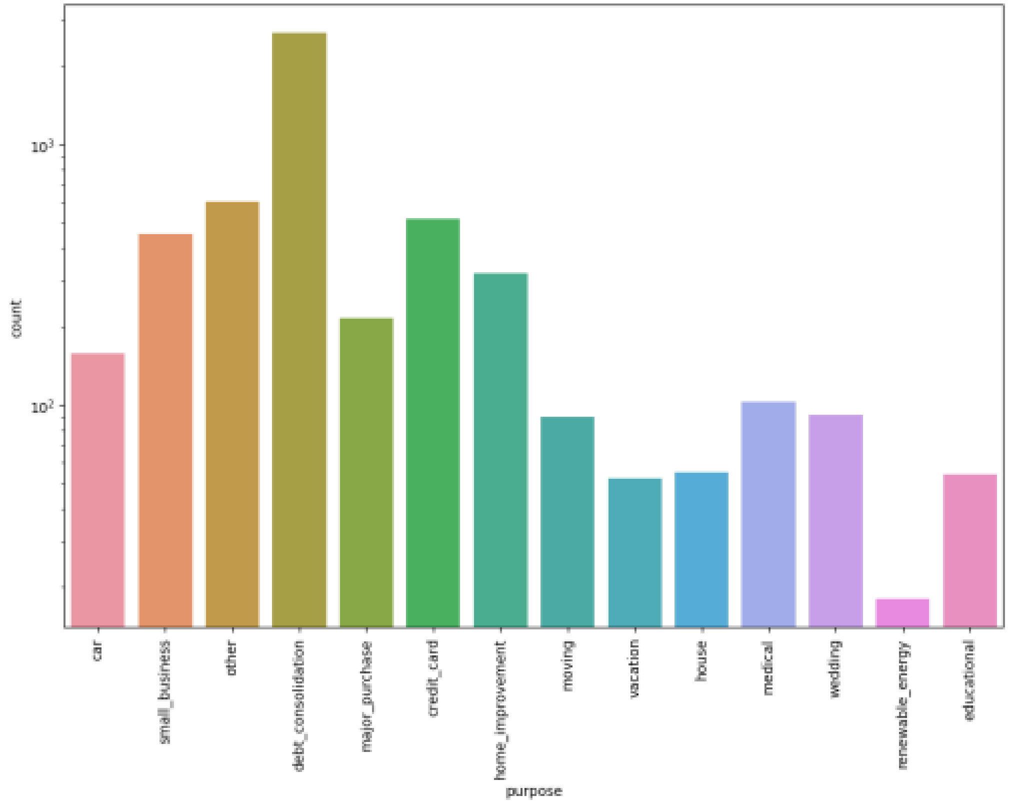
Applicants who has Grade B



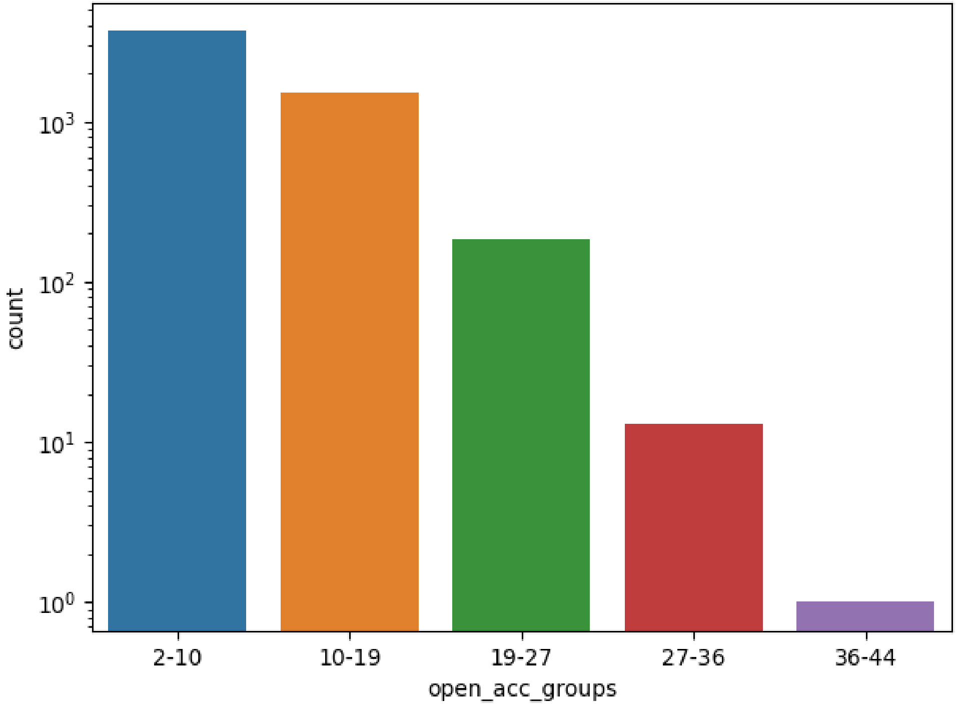
Home ownership vs charged off loans

When the applicant has home_ownership as 'RENT'.

The preceding examination regarding charged off loans for each variable indicates the following: There is a higher likelihood of defaulting when

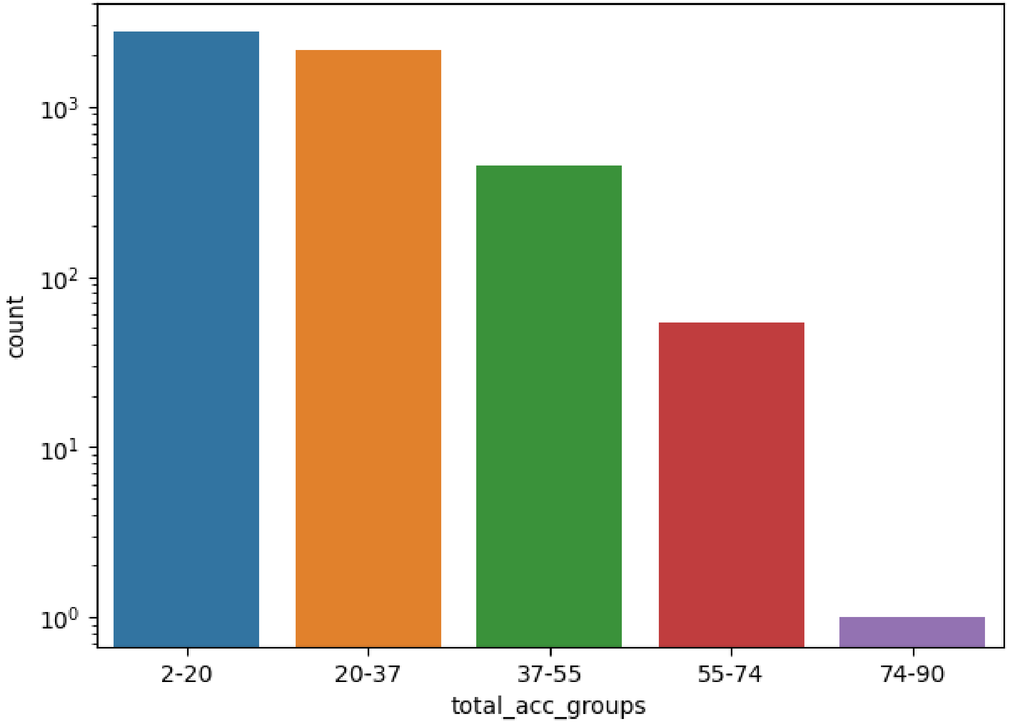


Applicant takes loan for the purpose of 'debt_consolidation'.

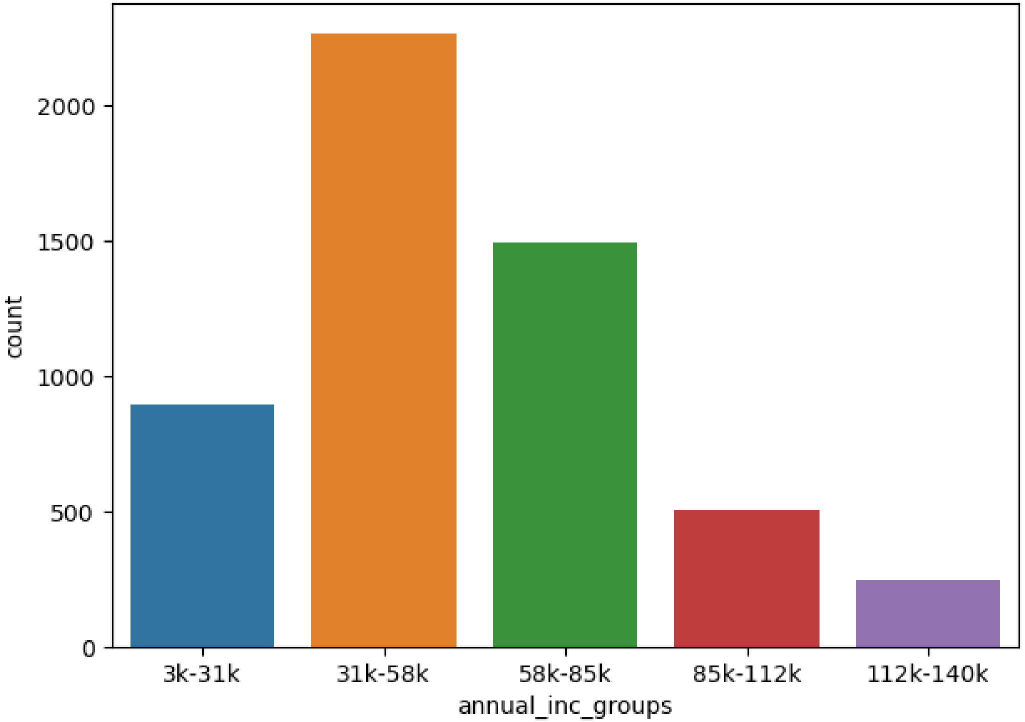


When the applicant has home_ownership as 'RENT'.

The preceding examination regarding charged off loans for each variable indicates the following: There is a higher likelihood of defaulting when

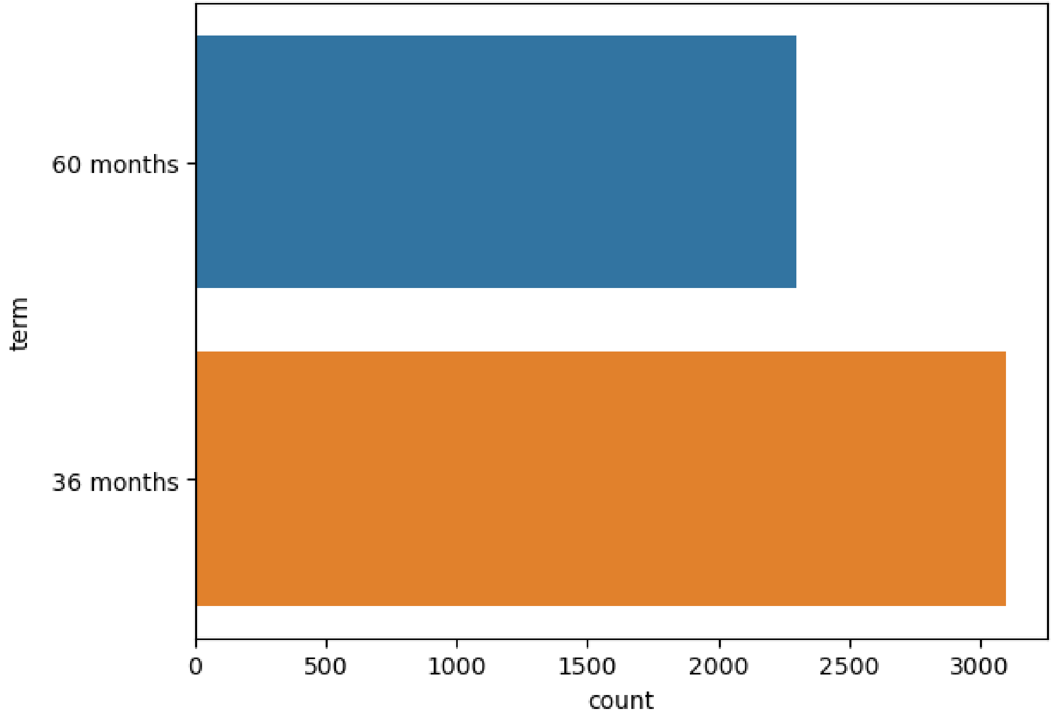


Applicant's who has “total_acc” 2 to 20

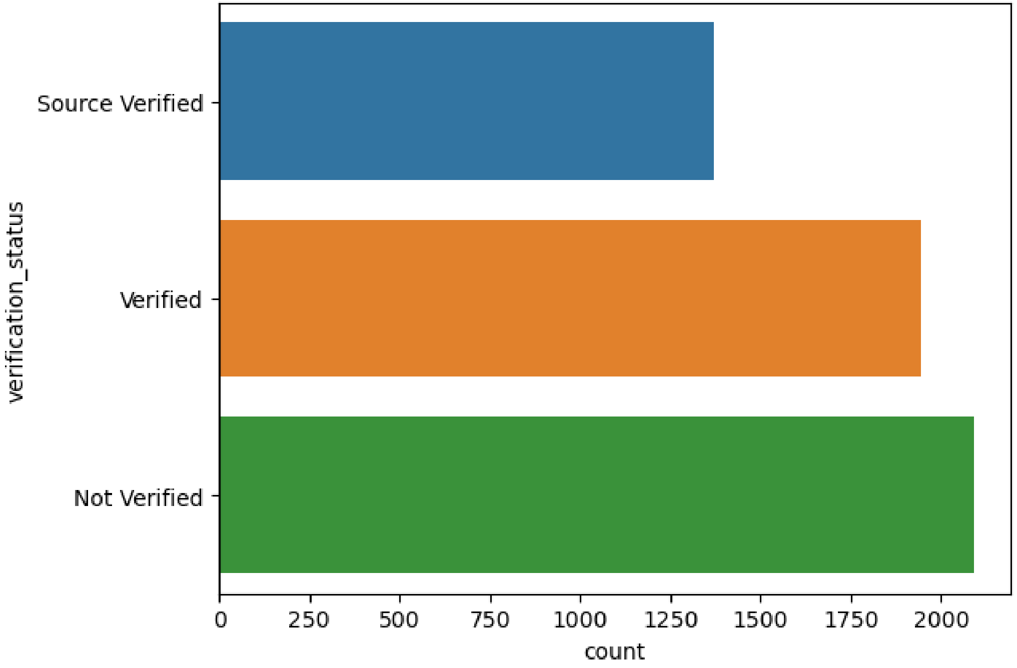


Applicants who has income between 31k to 58k.

The preceding examination regarding charged off loans for each variable indicates the following: There is a higher likelihood of defaulting when

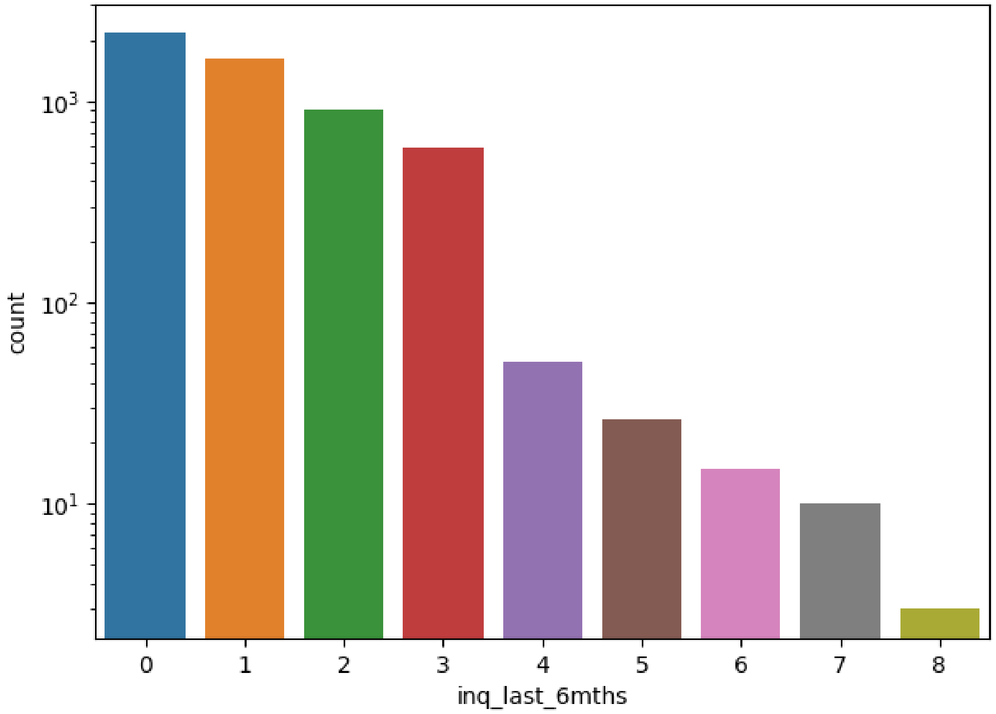


Applicants who has loan term as 36 Months.

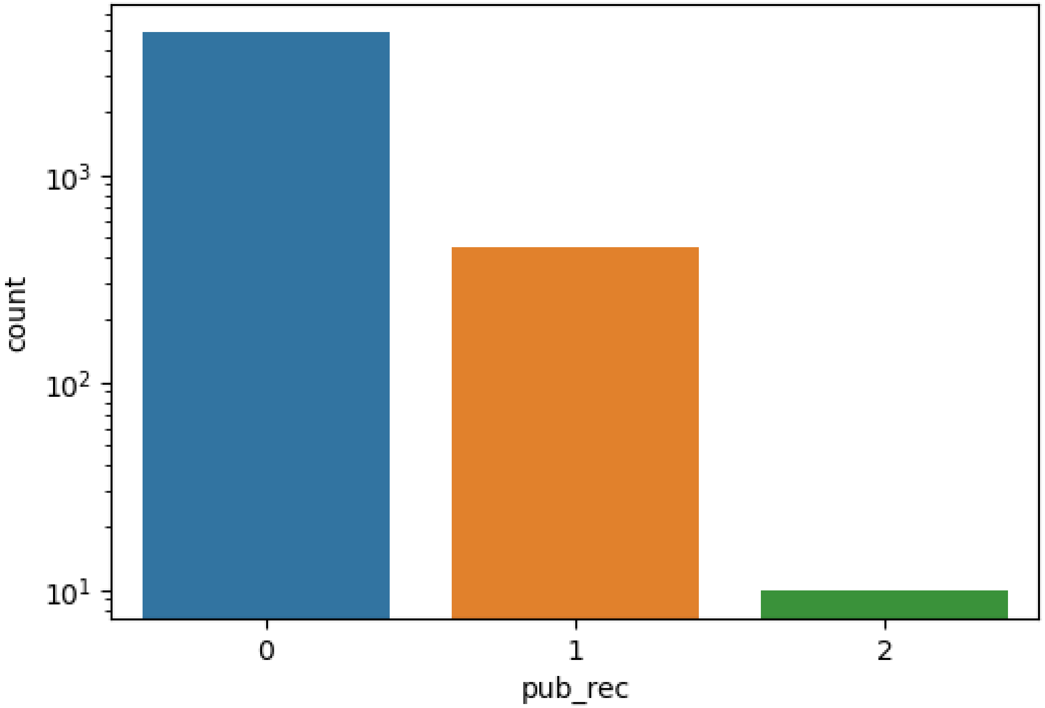


Applicants who has verification_status as 'Not verified'.

The preceding examination regarding charged off loans for each variable indicates the following: There is a higher likelihood of defaulting when

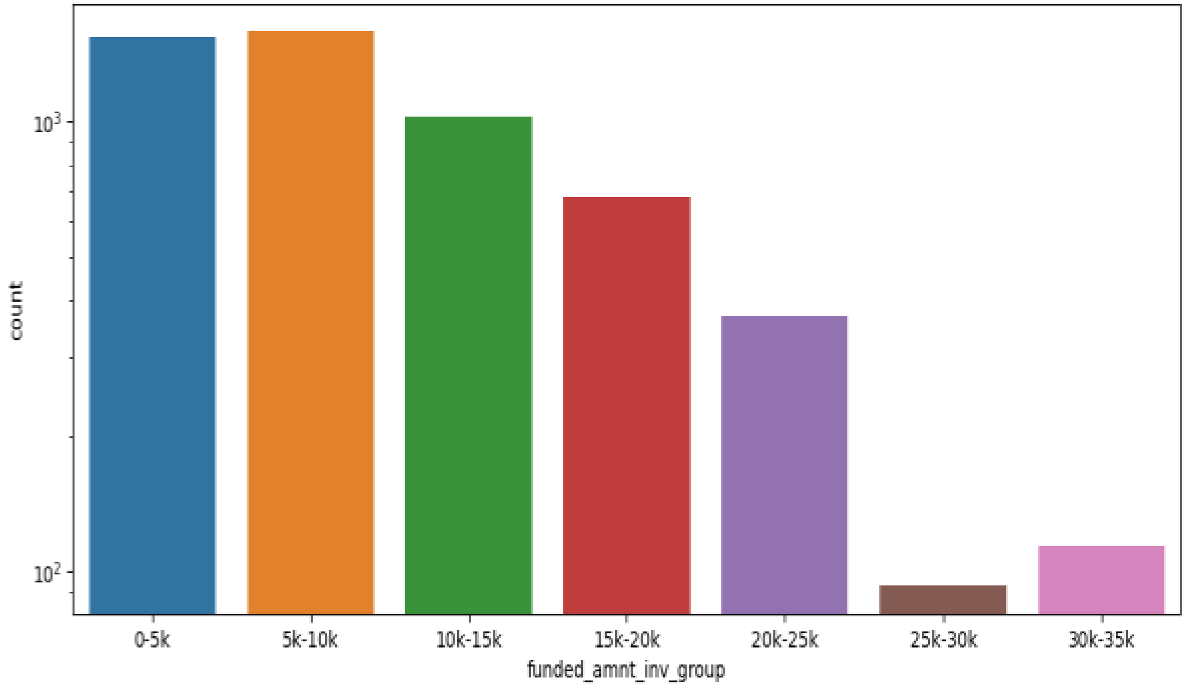


Applicants who has number of enquires in last 6 months as 0.

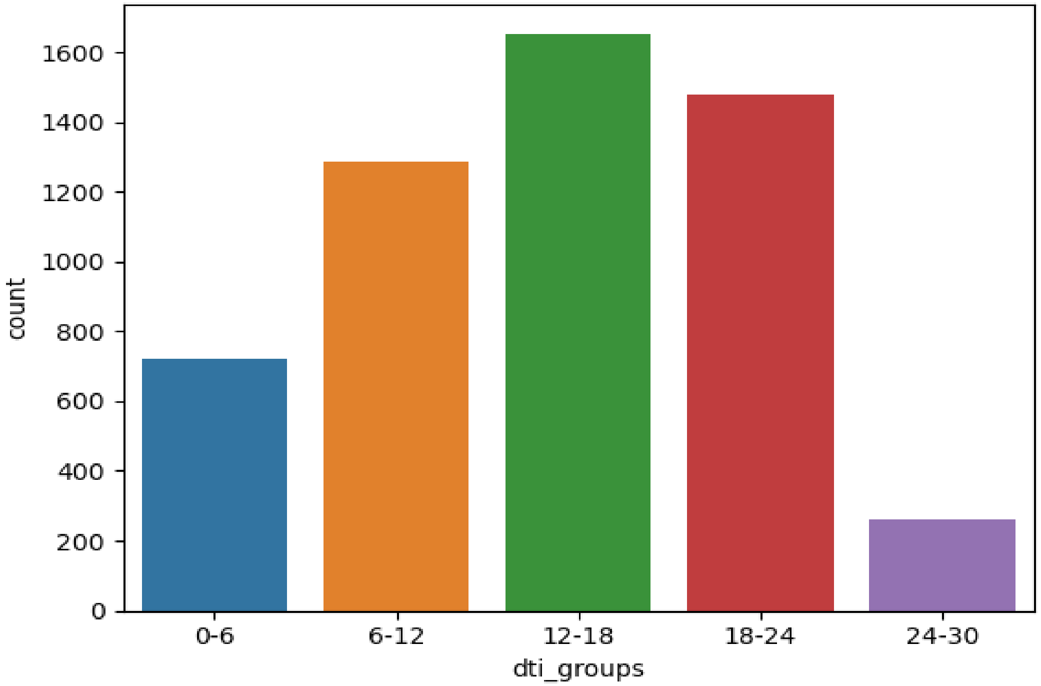


Applicants who has number of derogatory public records is 0.

The preceding examination regarding charged off loans for each variable indicates the following: There is a higher likelihood of defaulting when



When funded amount by investor is between 5000-10000.

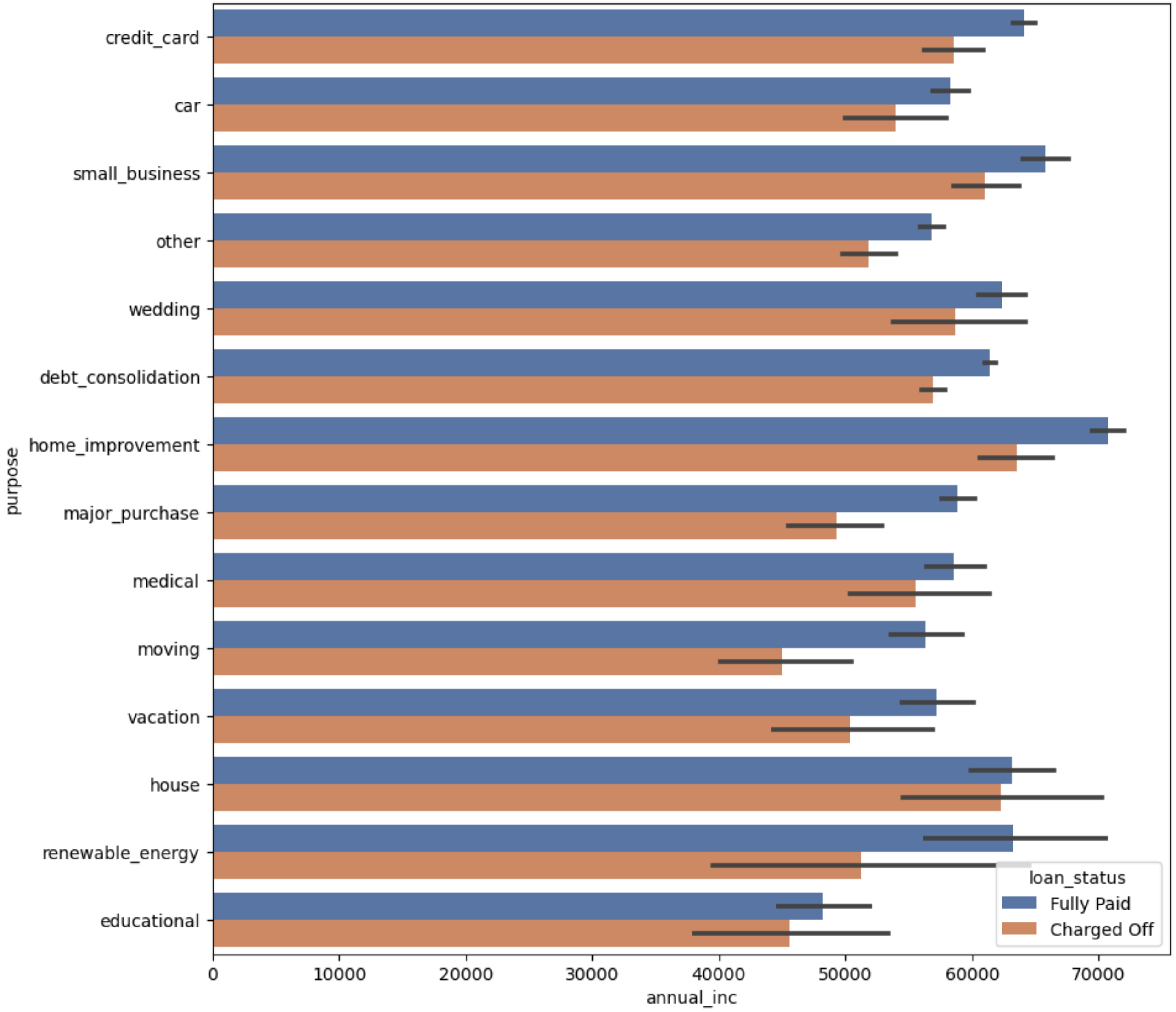


Applicants who has dti between 12-18.

Annual income vs Loan Purpose

While "debt consolidation" has the highest number of loan applications and defaults, it is not associated with the highest annual income among applicants.

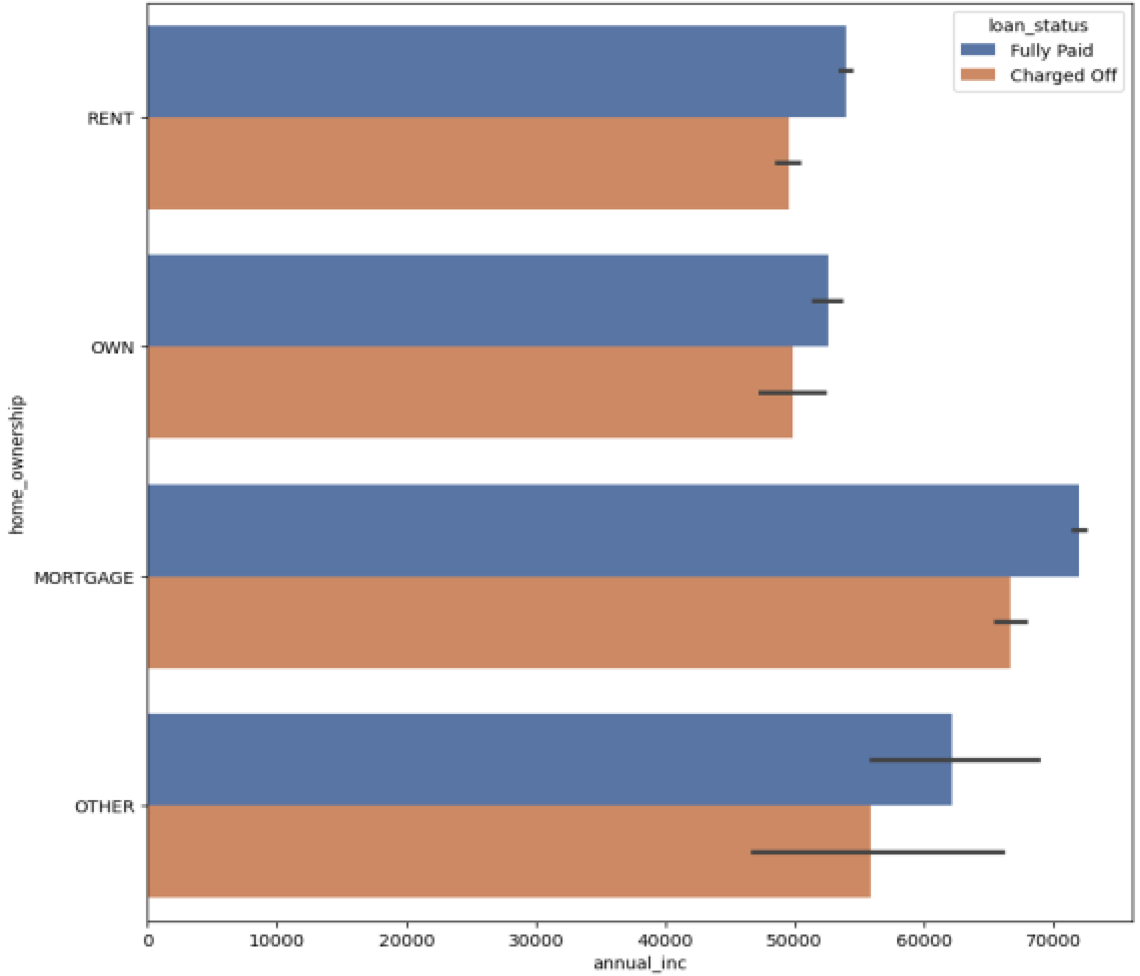
Individuals earning a higher income tend to seek loans primarily for purposes related to "home improvement," "housing," "renewable energy," and "small businesses."



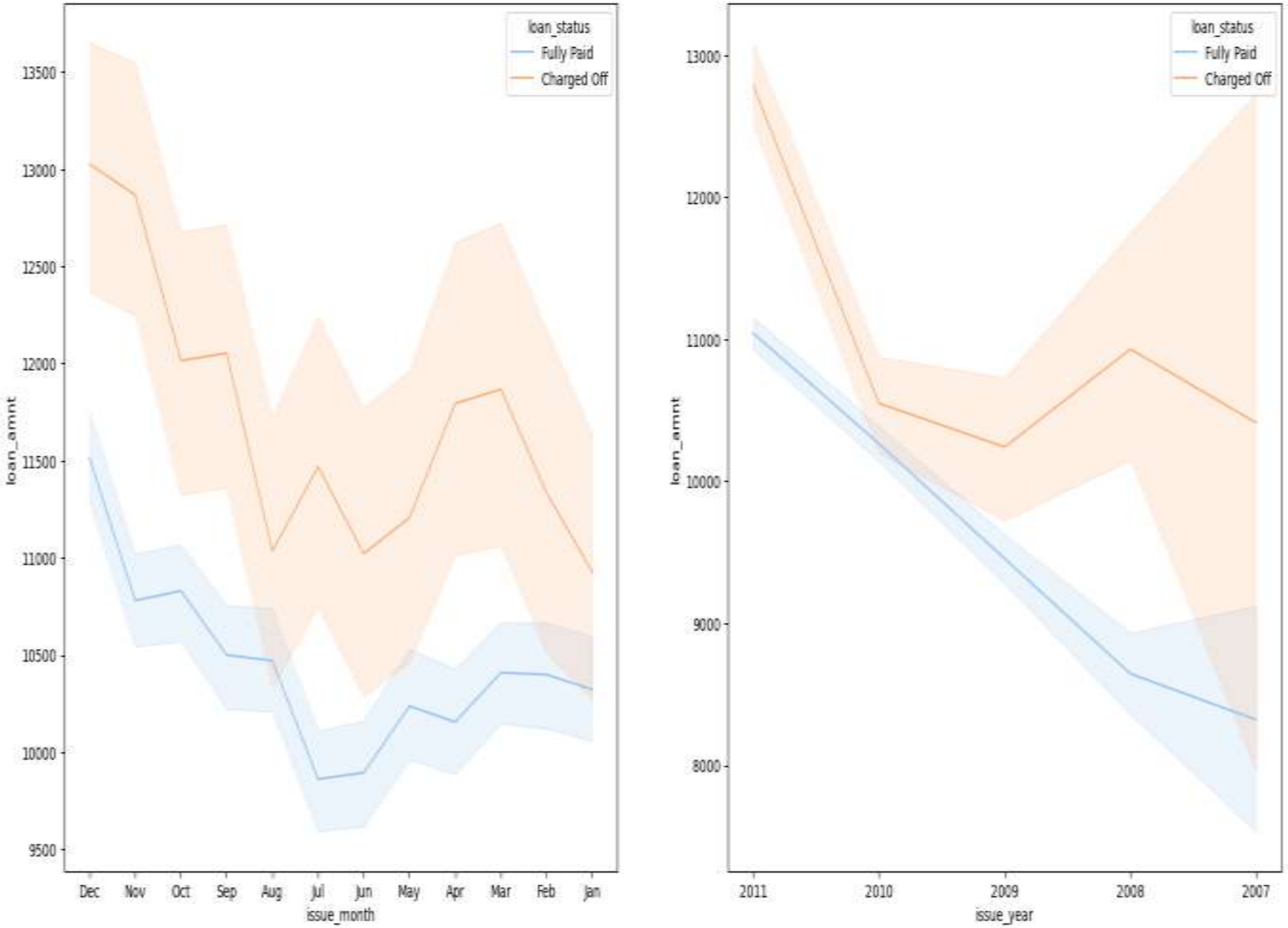
Annual Income vs Home Ownership

The preceding examination regarding charged off loans with different variables. There is a higher likelihood of defaulting when

Individuals seeking a loan for 'home improvement ' with an annual income in the range of 60,000 to 70,000.



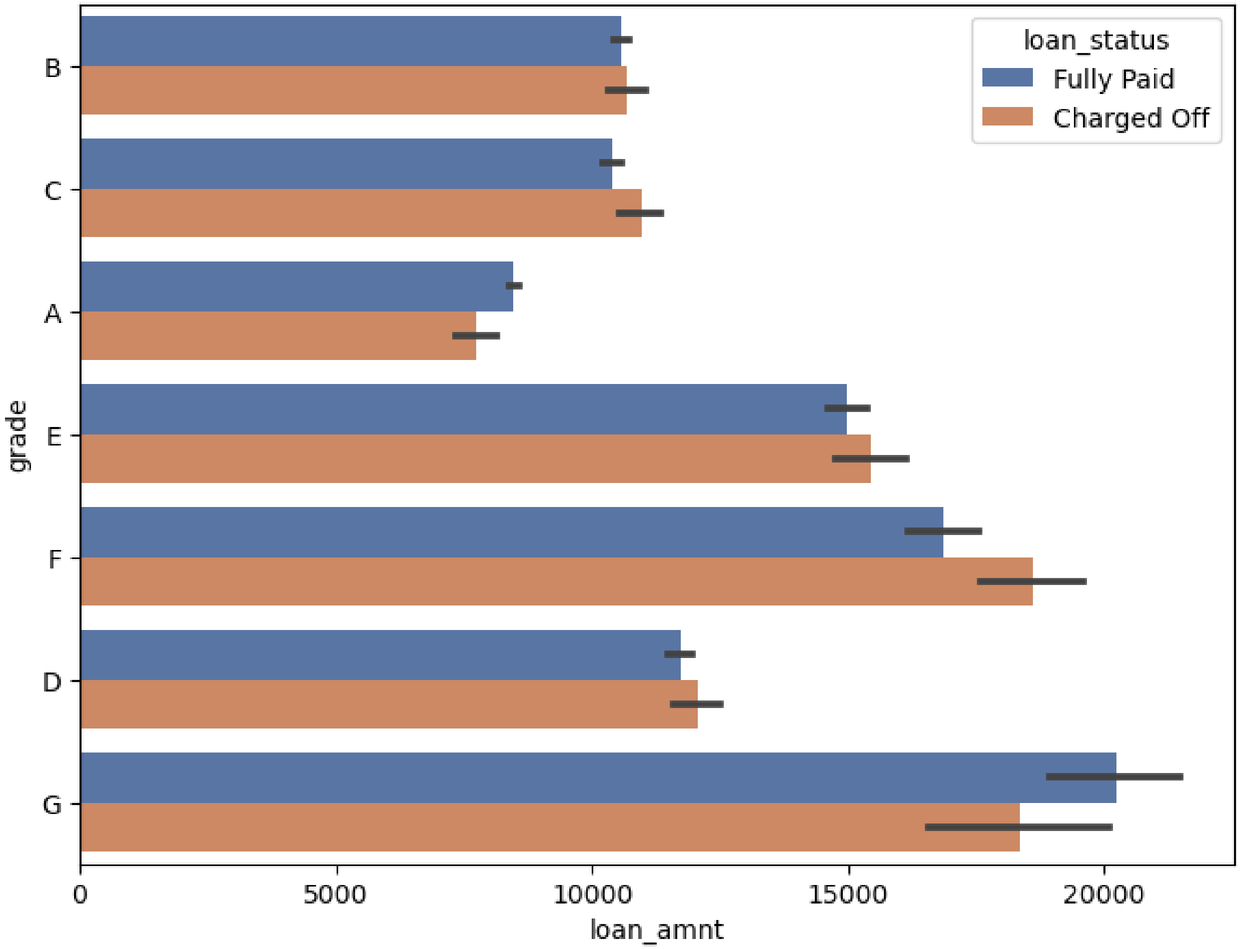
Loan amount vs Month issues and year issued



Loan Amount vs Grade

The preceding examination regarding charged off loans with different variables. There is a higher likelihood of defaulting when

When the grade is F and the loan amount falls within the range of 15,000 to 20,000.

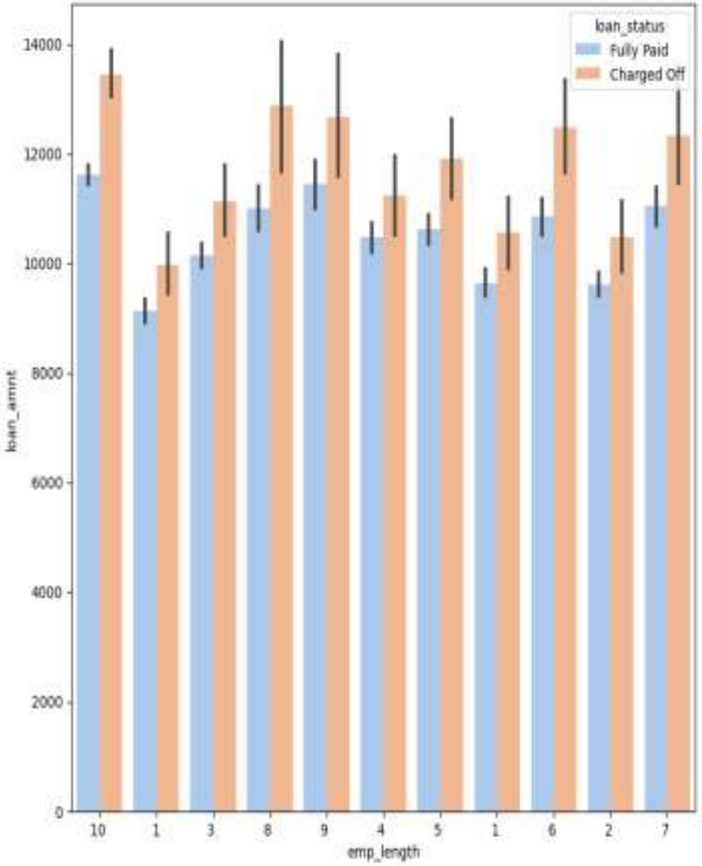


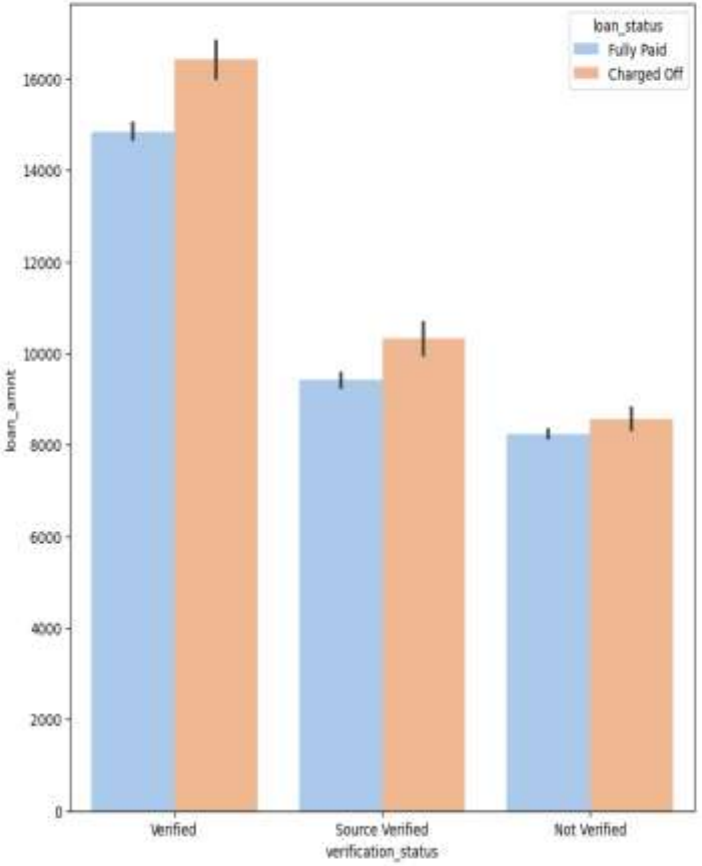
Loan Amount vs Emp Length

The preceding examination regarding charged off loans with different variables. There is a higher likelihood of defaulting when

Approval of loan seems higher for employees with longer work history.

- Analyzing the verification status data reveals a correlation between verified loan applications and higher loan amounts, suggesting that firms may prioritize verifying loans with greater values initially.

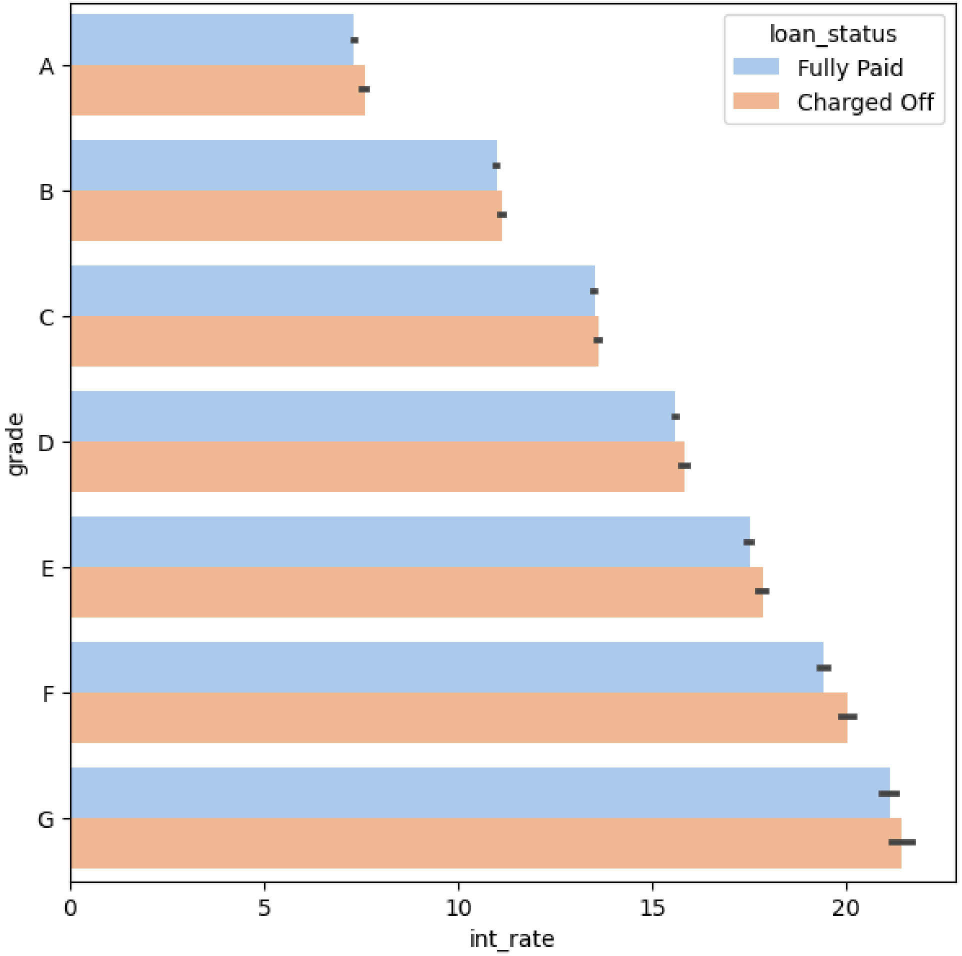




Grade vs Int rate

The preceding examination regarding charged off loans with different variables. There is a higher likelihood of defaulting when

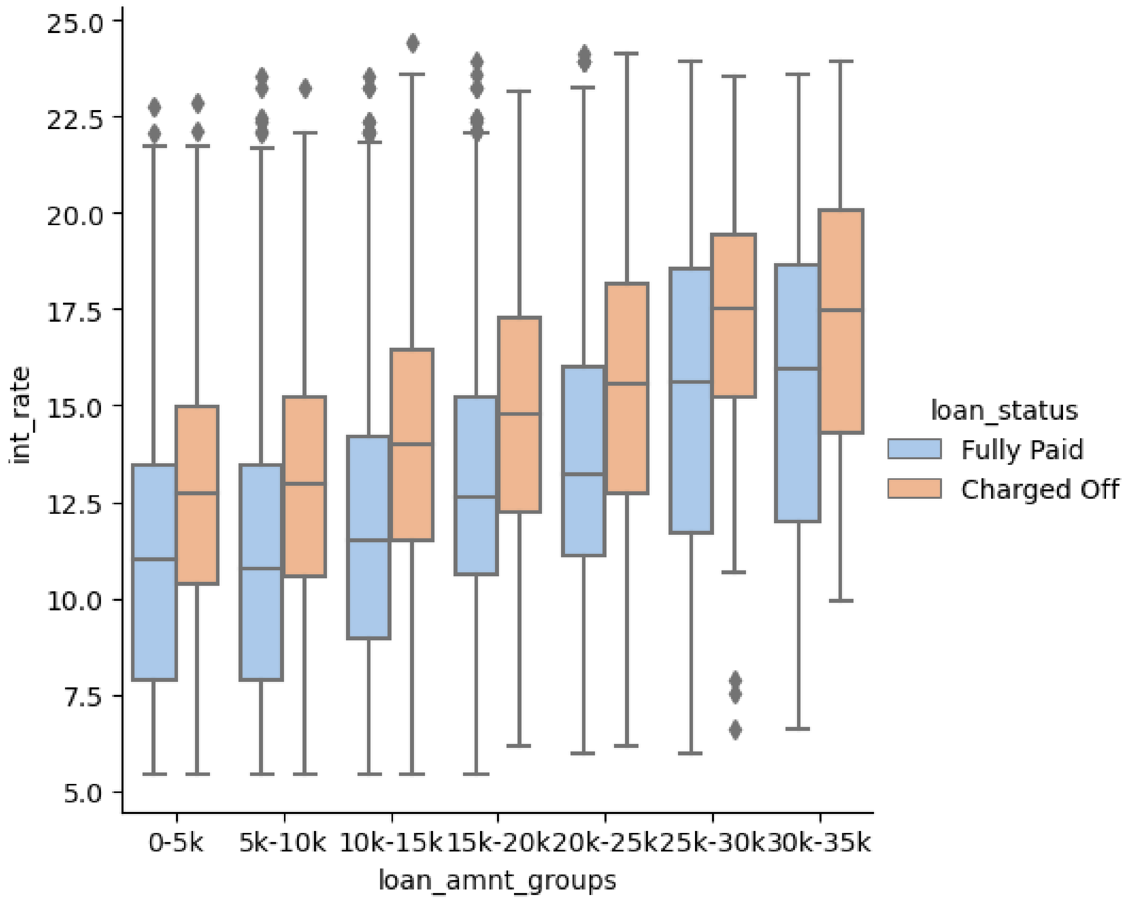
For a grade of G and an interest rate exceeding 20%.



Loan Amount vs Interest rate

The interest rate on charged-off loans is considerably higher than that on fully paid loans across all loan amount categories.

- This is a important factor.



Results

1. **Risk Factors:**
   - Applicants with a loan purpose of 'debt consolidation' and 'home improvement' may have an increased risk of default.
2. **Extended Loan Term:**
   - Longer loan terms, such as 36 months, may be associated with higher default risks.
3. **Marginal Income Ranges:**
   - Applicants with income ranging from 31k to 80k may face an elevated default risk, particularly if it's towards the lower end of this range.
4. **Home Ownership Types:**
   - Those who rent ('RENT') or have a mortgage ('MORTGAGE') may be more susceptible to defaulting.

Results

5. **Credit Grade Variation:**
   - Applicants with credit grades 'B', 'F', or 'G' might be at a higher risk of loan default, depending on the specific grade.
6. **Income and Loan Amount Correlation:**
   - Borrowers with certain income and loan amount combinations could be at greater risk of default.
7. **Debt-to-Income Ratio (DTI):**
   - A DTI between 12-18 may signify a higher likelihood of defaulting.
8. **Higher Interest Rates:**
   - Applicants facing interest rates above a certain threshold, such as 20-24%, could be at a heightened risk of default.