Lending Club Case Study

Exploratory Data Analysis

By

Rakesh Kumar Padhiary

Himanshu Rana

Summary Of Contents

- Problem Statement & Objective
- Data Summary & Highlights
- Data Cleaning
- Data Conversion, Handling Missing Values & Derived Columns
- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis
- Correlation Analysis
- Conclusions

Problem Statement & Objective

- In a certain consumer finance company which specializes in lending various types of loans to urban customers receives loan application and it has to make a decision for loan approval based on applicant's profile.
- The bank has to deal with two type of risks associated to these applications

  A. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

  B. 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

- Based on the above risks, the bank would take two below decisions

  A. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

  Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

  Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

  Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

  B. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.).Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Objective *: Use EDA to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.*

Data Summary & Highlights

- The loan.csv file contains 39717 rows and 111 columns consisting of loan attributes and customer attributes
- Out of 111 columns, 87 are numerical and 24 are Object types

Data Cleaning

- There are no header/footer or no summary rows (Total/Sub-Total rows etc.) present in the dataframe
- There were 1140 rows where loan_status = 'Current' and were removed from analysis as their loan tenure is not yet completed
- The number of duplicate rows count was 0
- There were 55 columns having all the row values as null/NA. Removed those from the analysis.
- Deleted columns having greater than 30 percent null values as per the industry practice(desc, mths_since_last_delinq, mths_since_last_record columns)
- Dropped columns with only one unique value as they don't contribute to analysis.
- Out of the 3 columns ['member_id', 'url','id'] , deleted the first two columns and kept only 'id' columns for analysis. These are the type of columns where all the values are unique in nature. As doing analysis on these values wont result into any specific patterns.
- Deleted columns irrelevant to loan approval process (post-approval behavioral columns).
- Deleted columns ['title' and 'emp_title'] containing so much of textual values
- Limited the analysis to Group level only and deleted the sub group level to make analysis more easier
- Removed the rows having all the column values as blank/null. Though in this analysis we have 0 such rows

Data Conversion, Handling Missing Values & Derived Columns

Data Conversion

- Converted data types of int_rate, term, loan_amnt, funded_amnt, and issue_d to required data types needed for analysis

Handling Missing Values

- Removed the rows for emp_length and pub_rec_bankruptcies columns where the missing values where identified. In this case we choose to remove instead of imputing for the reason as follws: emp_length column is already a categorical column and imputing the missing values with 'Mode' of the column may lead to biasing
- pub_rec_bankruptcies, even though it is an integer column, but since the number of unique values are very less and repeating, we can consider it as a categorical column and again imputing the values will lead to biasing to certain extent.

Derived Columns

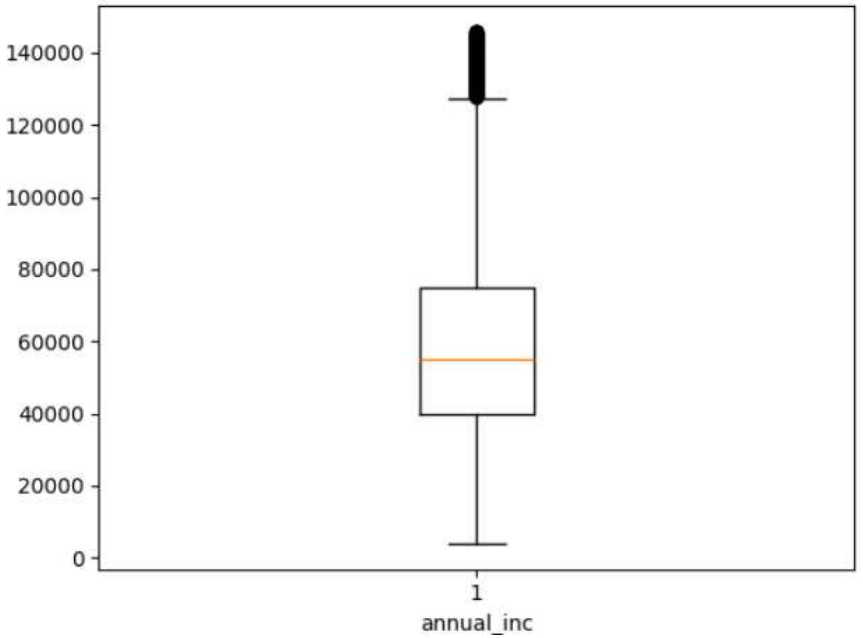
- Derived issue_yr and issue_mon from issue_d. columns for better insights.

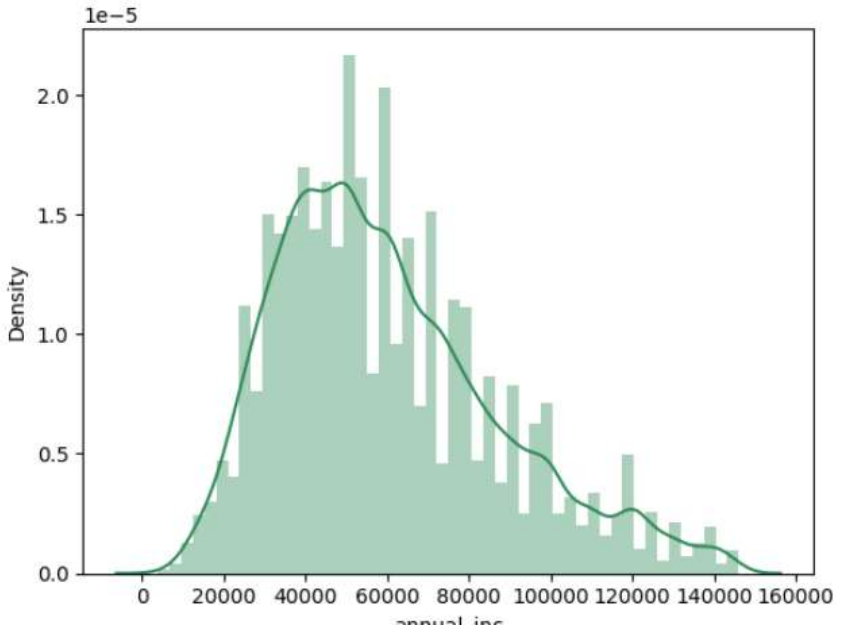
After doing these above steps ,we had 36847 rows and 21 columns available for the next set of analysis

Univariate Analysis

- Removed the outliers as required before analysis from numerical columns

Analysis on Annual Inc Column

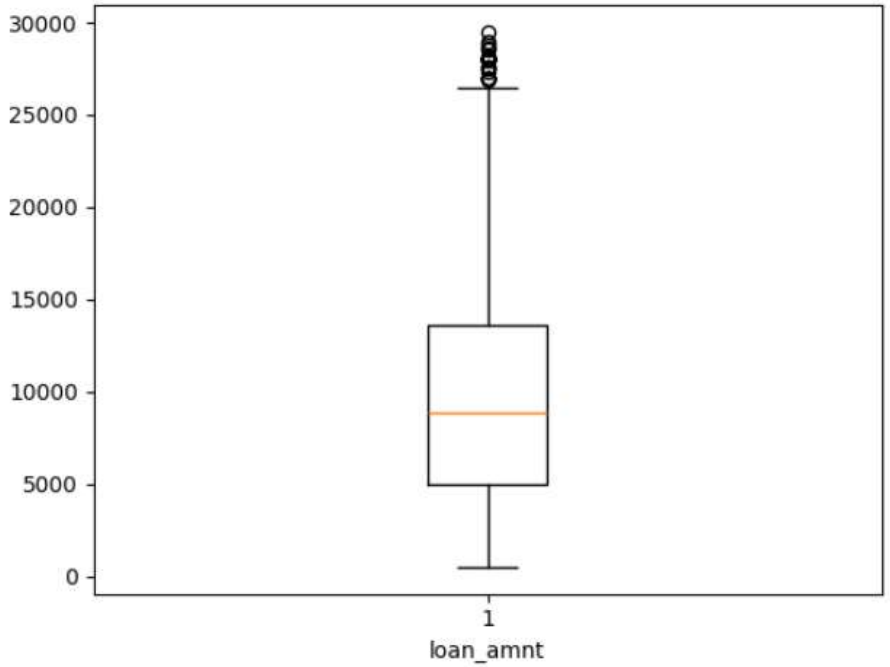


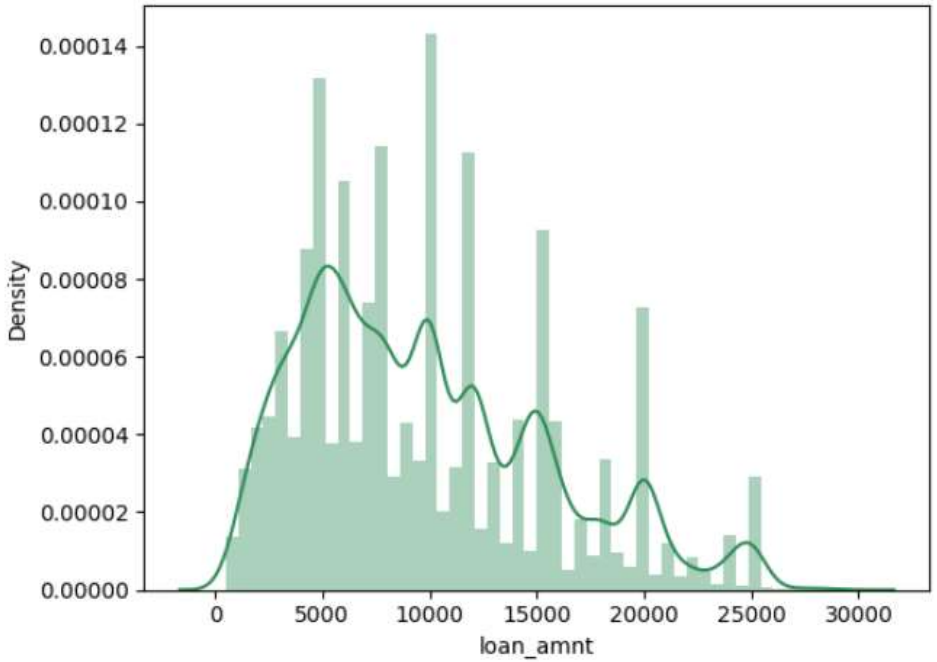


1. The Annual income of most if applicants lies between 40k-75k
2. The median annual income is 55K

Univariate Analysis

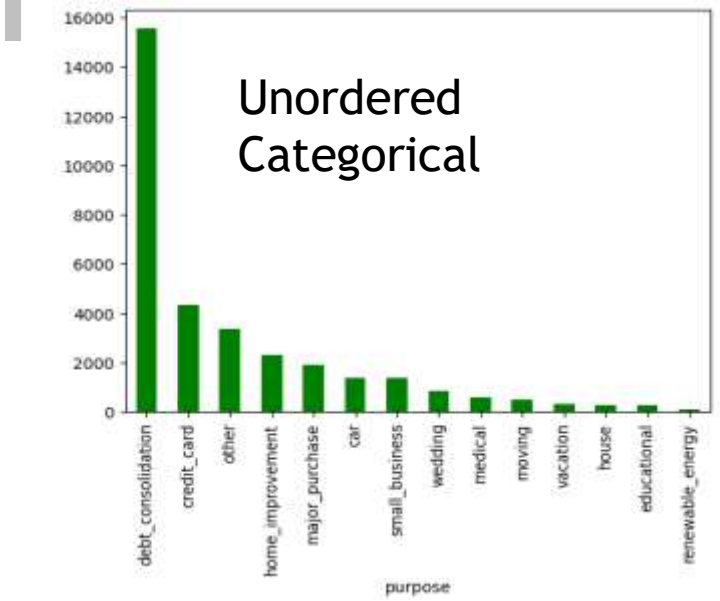
Analysis on Loan amt Column



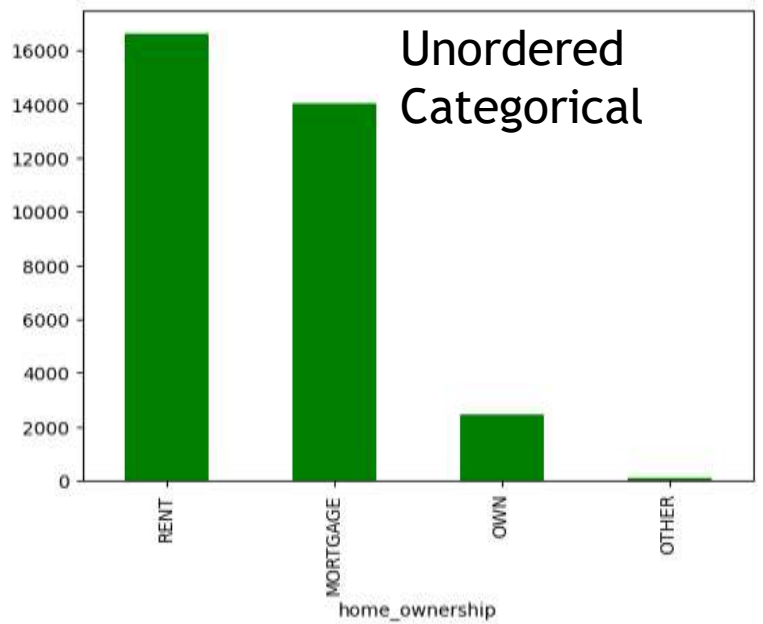


1. Most of the loan amount applied was in the range of 5k-14k
2. Actual Max Loan amount applied was around 27k.

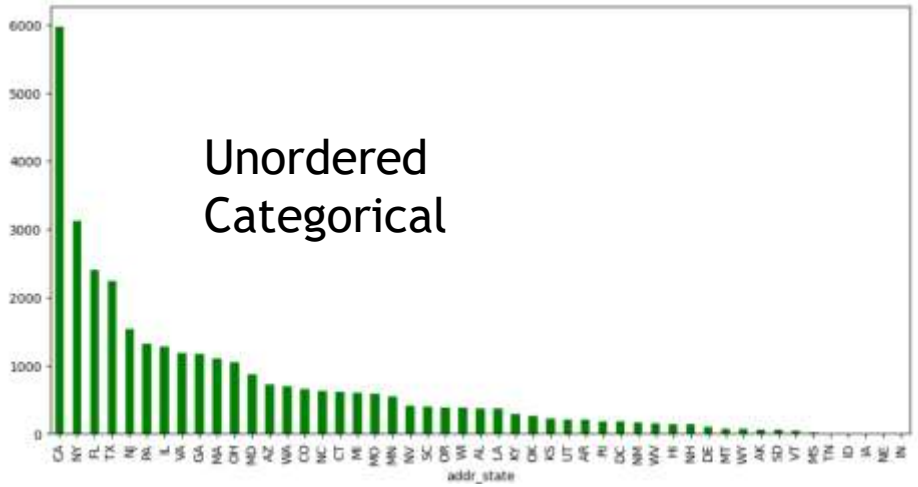
Univariate Analysis Categorical Variable(Ordered & Un Ordered)



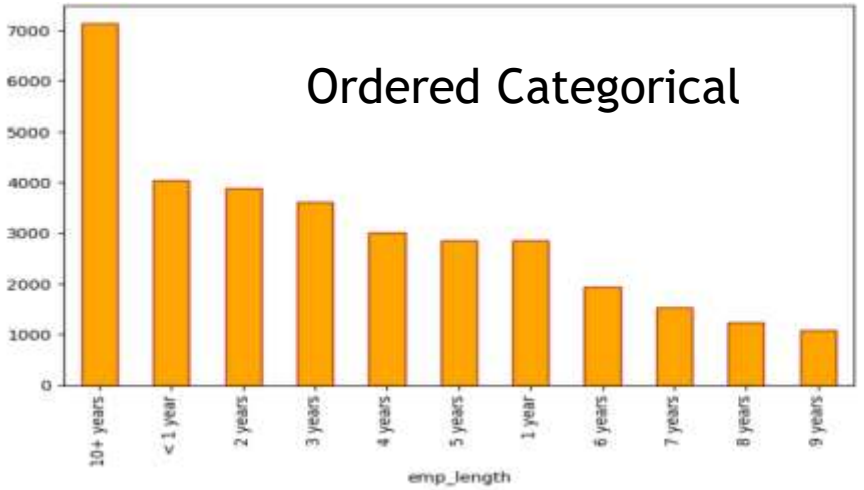
Most of the loan applicants are for debt consolidations.



Maximum of loan applicants are either living on Rent or on Mortgage

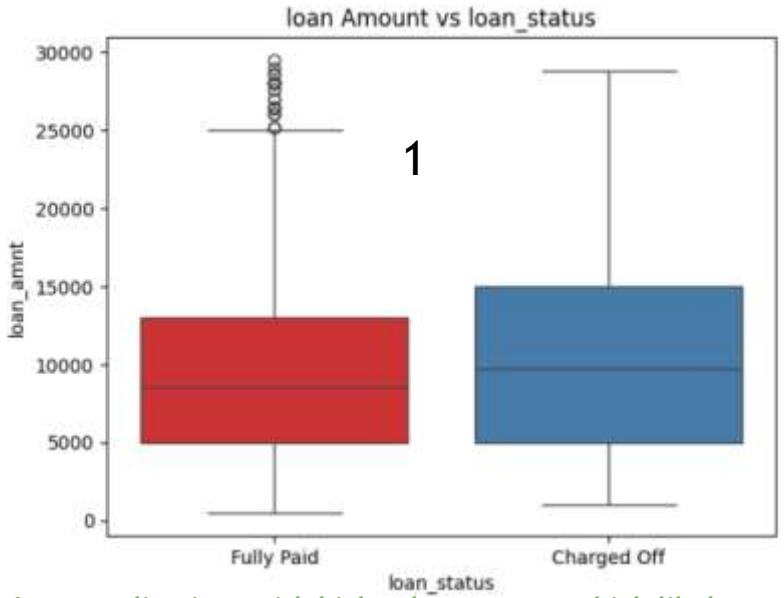


Majority of the Loan applicants are from CA(State)

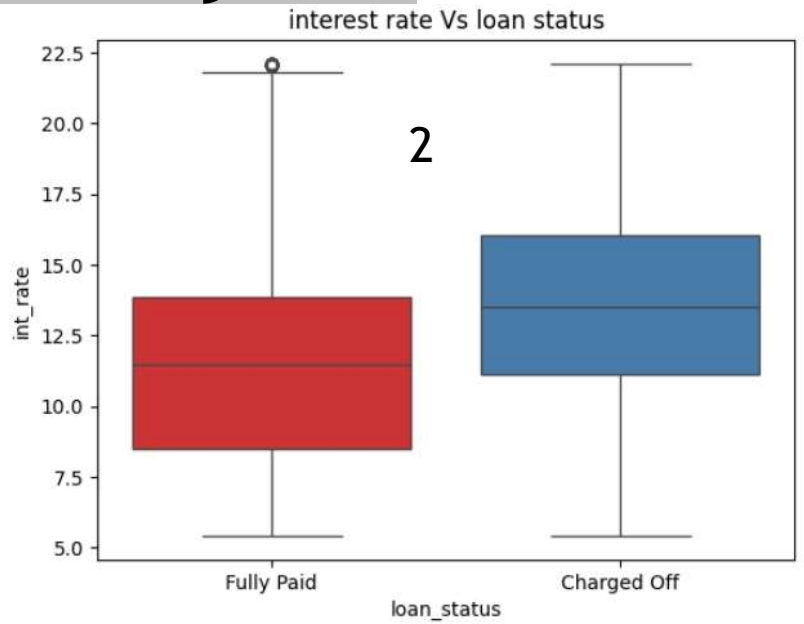


Most of the applications are having 10+ yrs of Exp

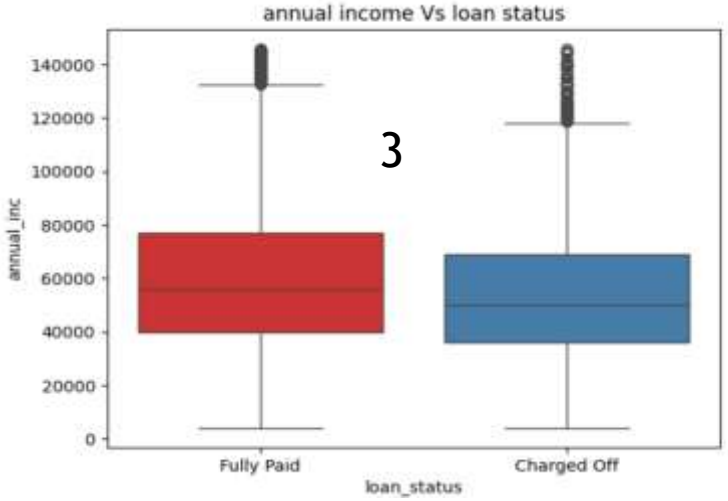
Segmented Univariate Analysis



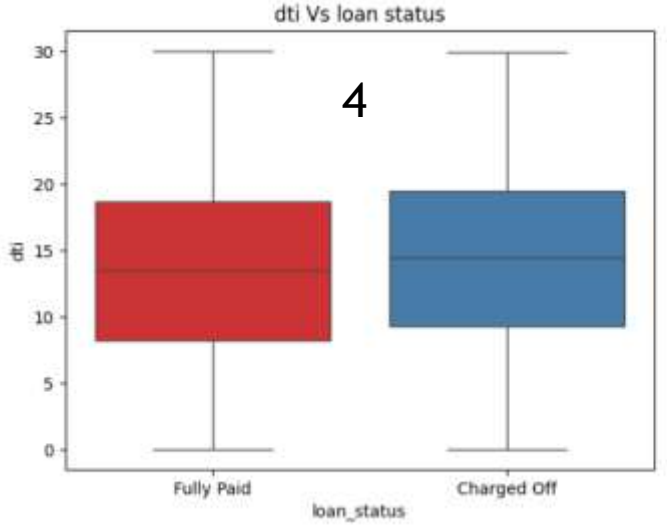
1. Loan applications with higher loan amounts high likely to get charged off.



2. Obvious observation is greater the interest rate more the chances of defaulting the loan



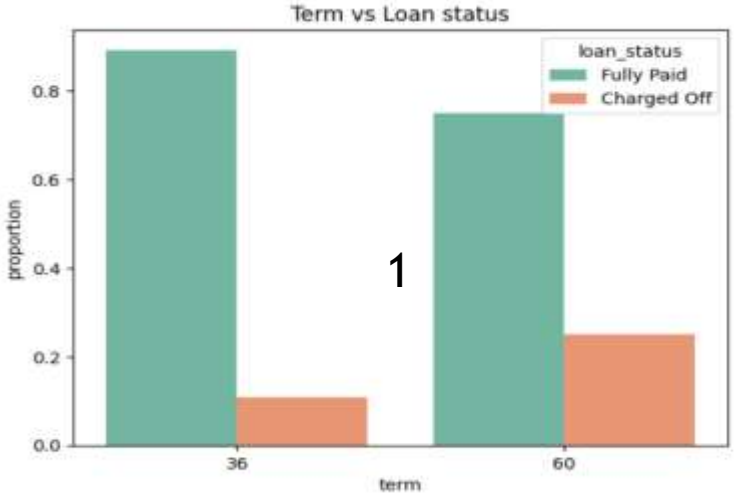
3. With Higher annual income borrower are more likely to fully pay the loan



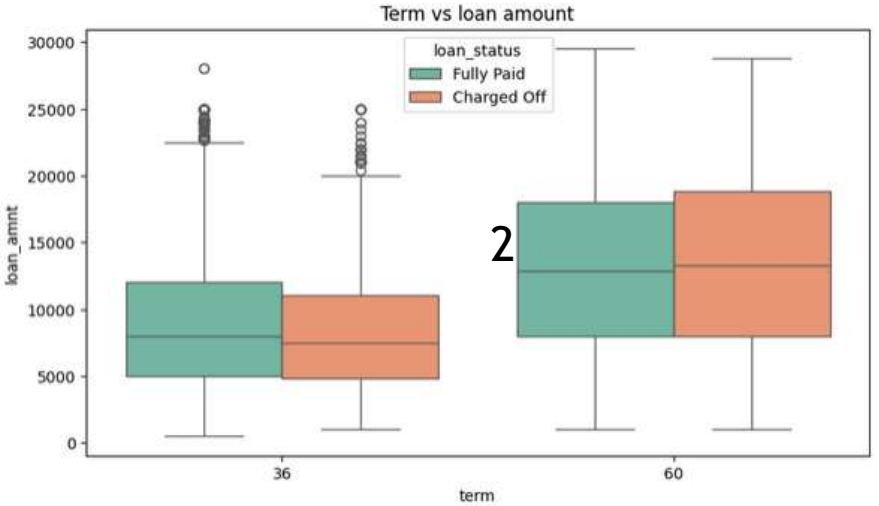
4. Borrowers with higher dti have more probability to default the loan

Bivariate Analysis

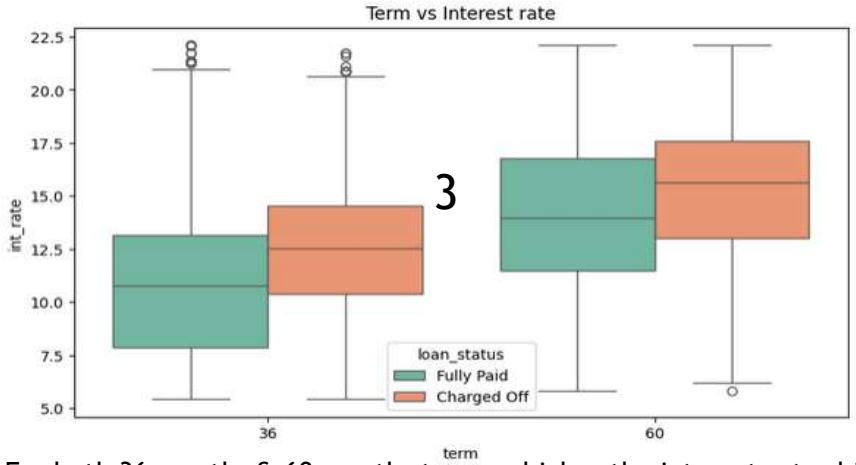
- On term column



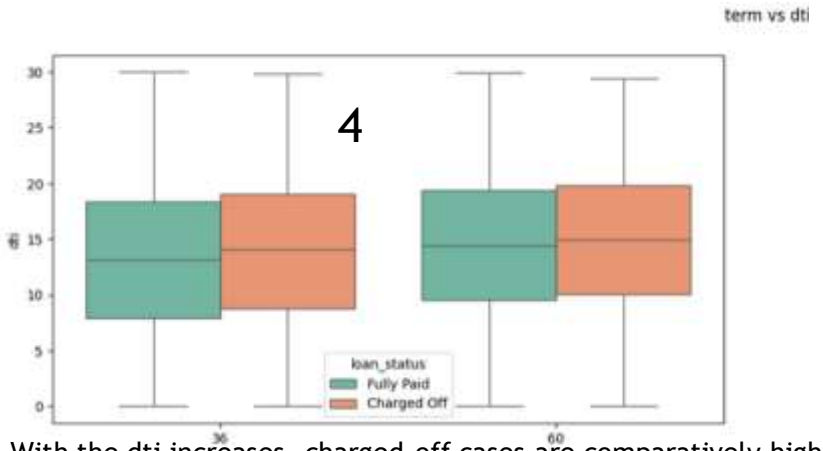
1. More proportion of borrowers defaulted loan in 60 months term compared to 36 months term.
Fully Paid proportion is higher in case of 36 months term.



2. The spread of 'Fully Paid' and 'Charged off' are almost equal for both 36 months & 60 months tenure
Loan amount is not a deciding factor for defaults in both 36 and 60 months



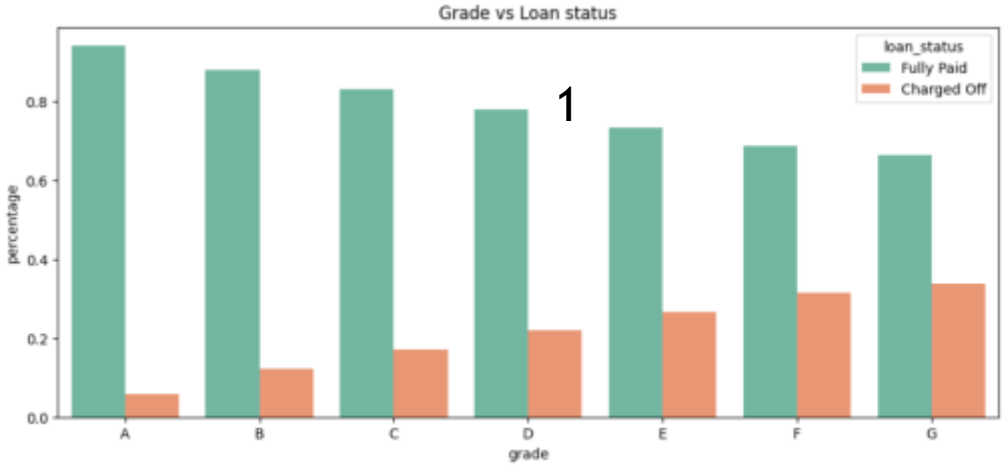
3. For both 36 months & 60 months tenure, higher the interest rate, higher chance of default



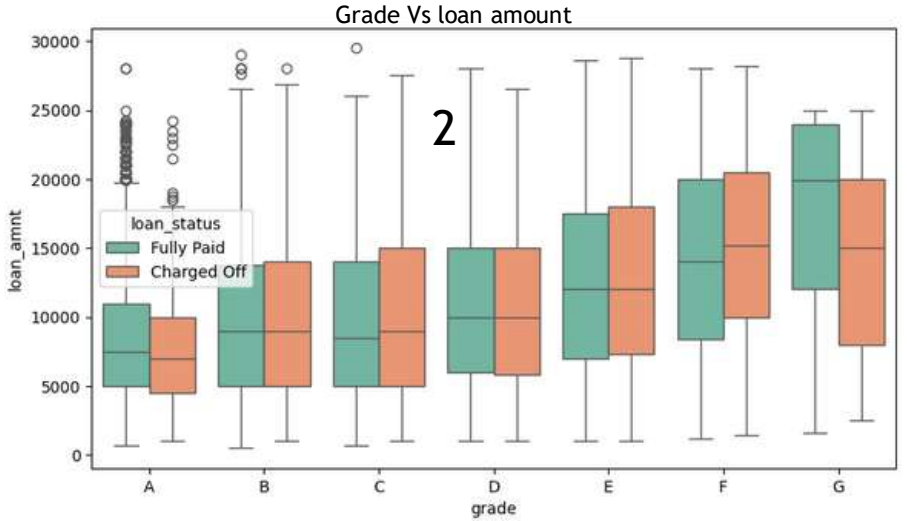
4. With the dti increases, charged-off cases are comparatively higher than fully-paid cases, for both 36 & 60 months tenures

Bivariate Analysis

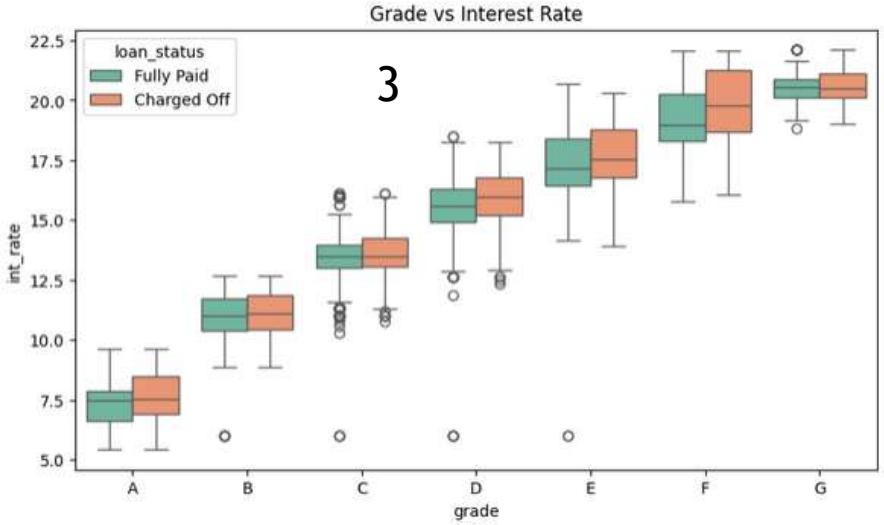
- On grade column



1. Interest Rates are Higher as Grades are Lowering (A to G).
The interest rates are higher for Higher tenure loans (60 months).



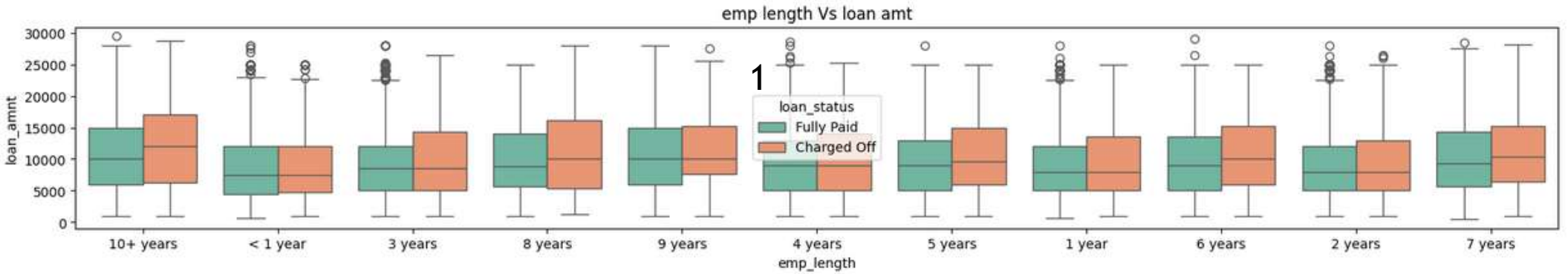
2. Borrowers opting for lower grade loans (e.g. F & G) with loan high amount are more likely to default.



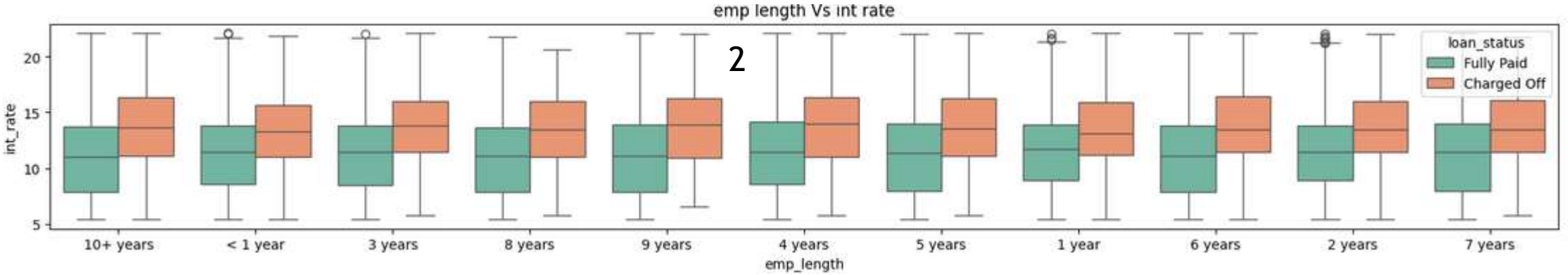
3. With the decrease in grade leads to increase in interest rate, and the borrowers are more exposed to default the loan.

Bivariate Analysis

- On emp length column



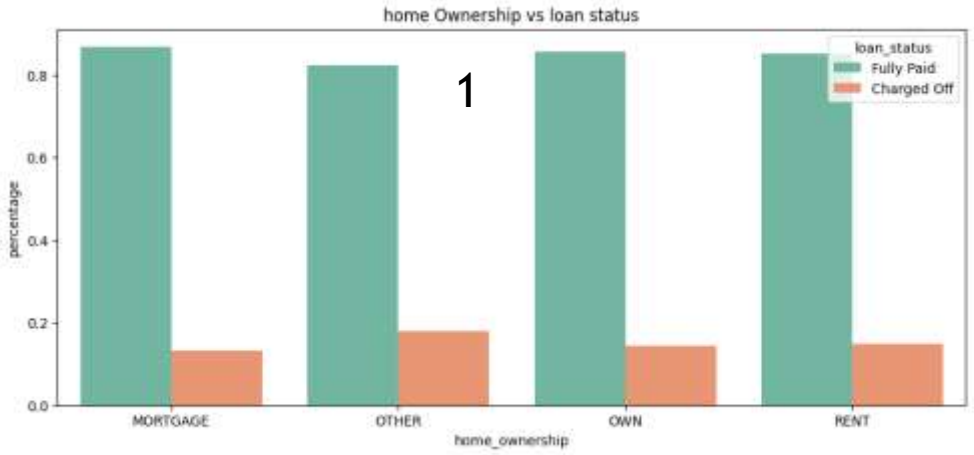
1. Borrowers having employment lengths opted for more loan amounts and are more like to default.



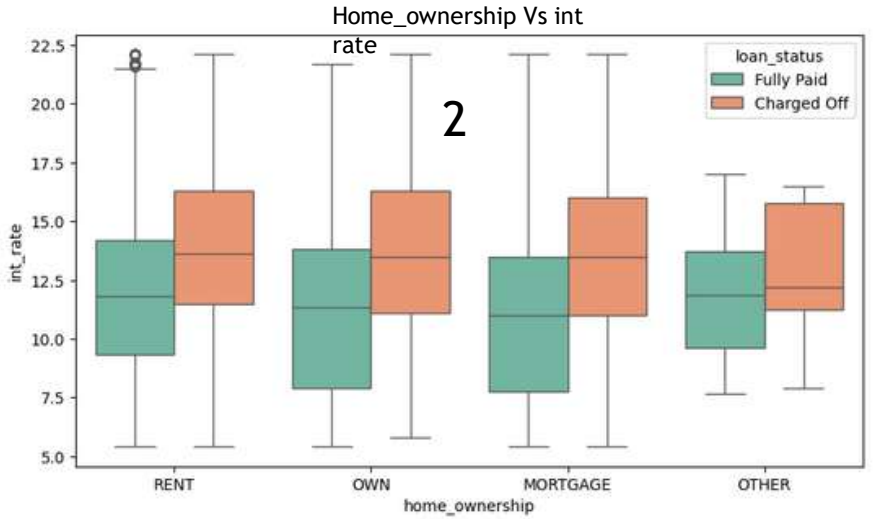
2. Irrespective of employment length, loans with more interest rate likely to be defaulted more.

Bivariate Analysis

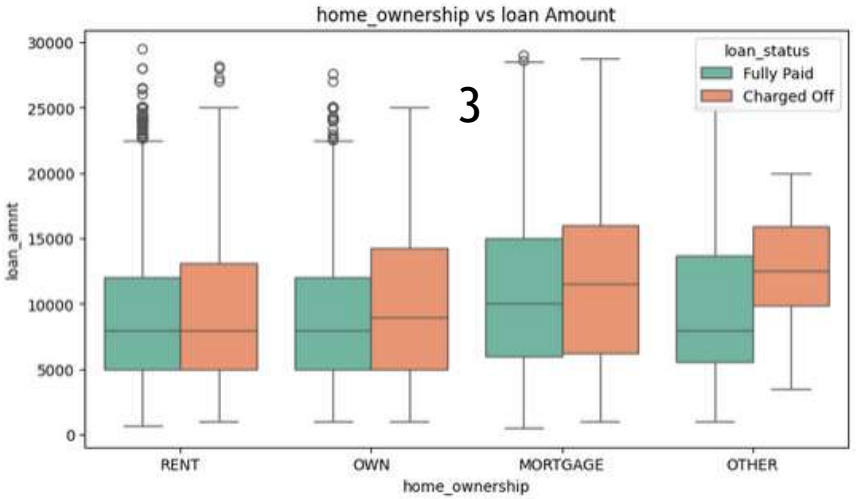
- On Home ownership Column



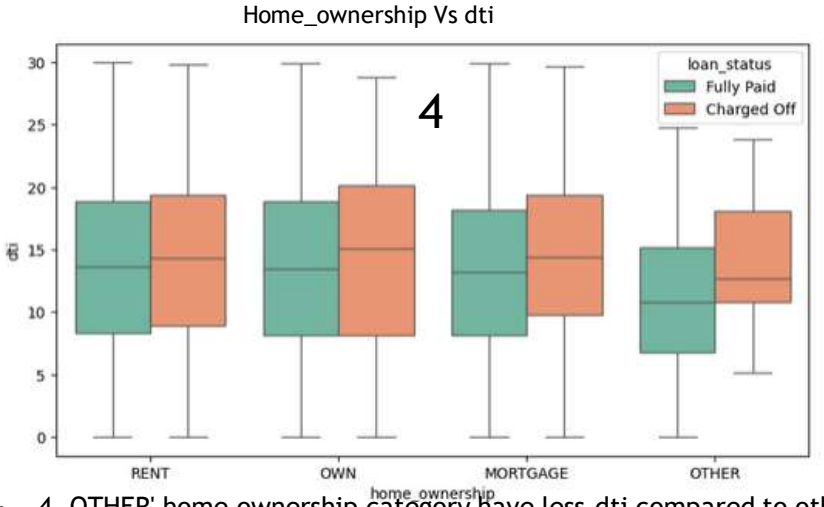
1. A slightly high percentage of defaults being identified for 'OTHER' home ownership category.



2. Irrespective of home ownership criteria, the charged off rate is high when the interest rate is high



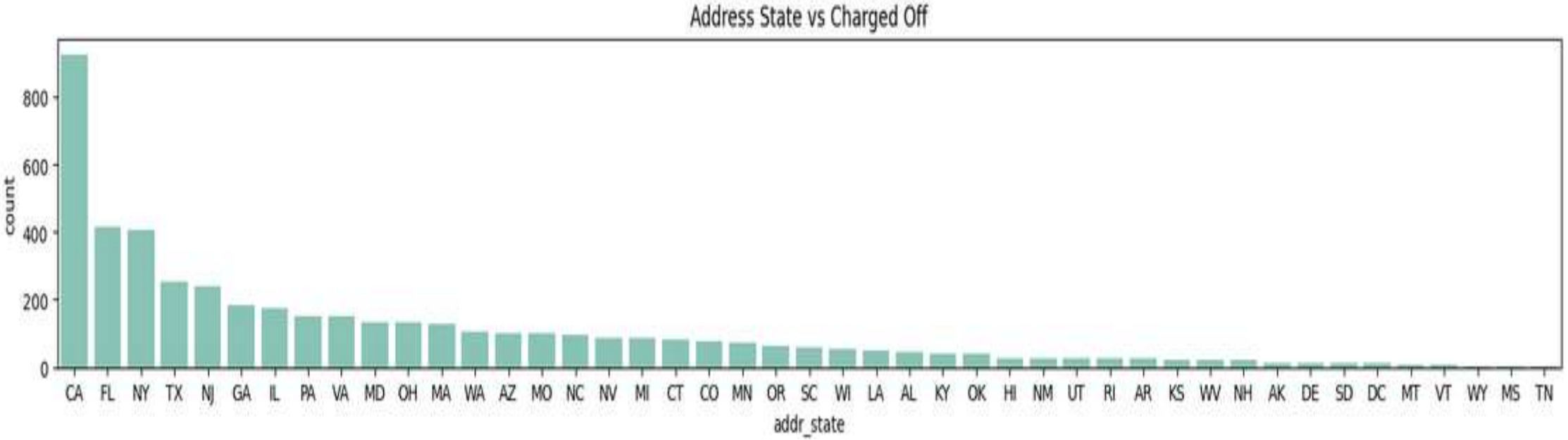
3. Irrespective of home ownership criteria, higher the loan amount leads higher chances of default.



4. OTHER' home ownership category have less dti compared to other categories. and has less spread out of the fully paid and charged off

Bivariate Analysis

- On Address state column

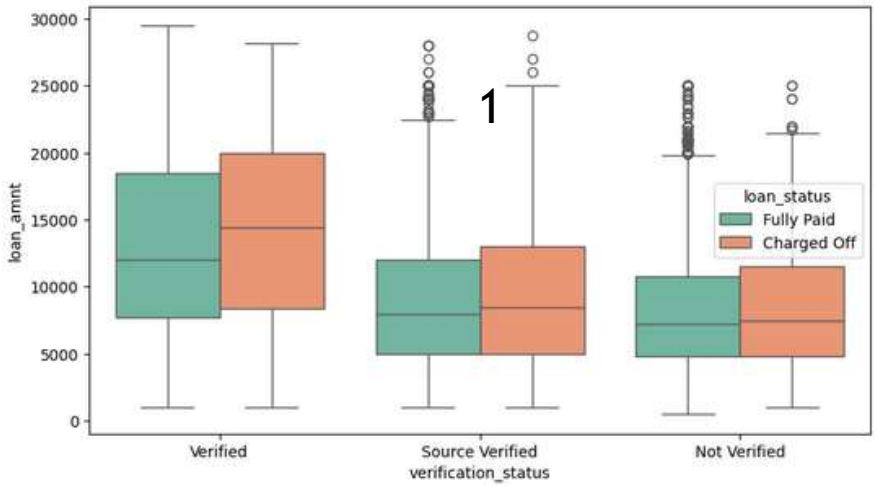


Top 3 states where more borrowers defaulted are CA , FL and NY states

Bivariate Analysis

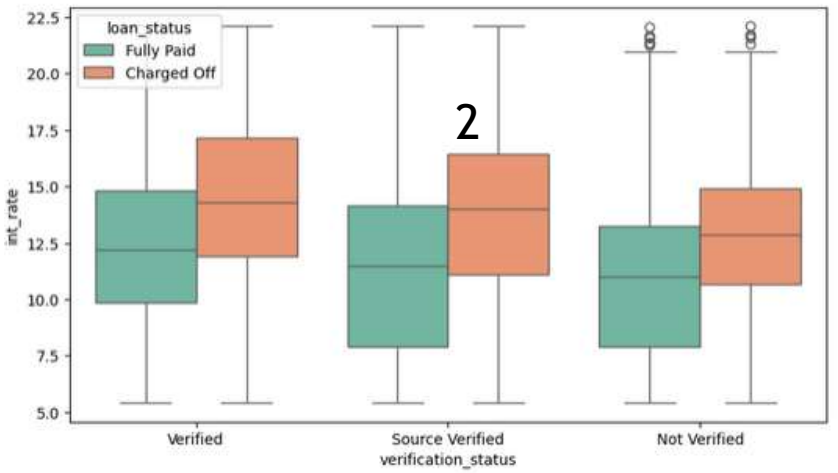
- On Verification status Column

Verification status Vs loan amt



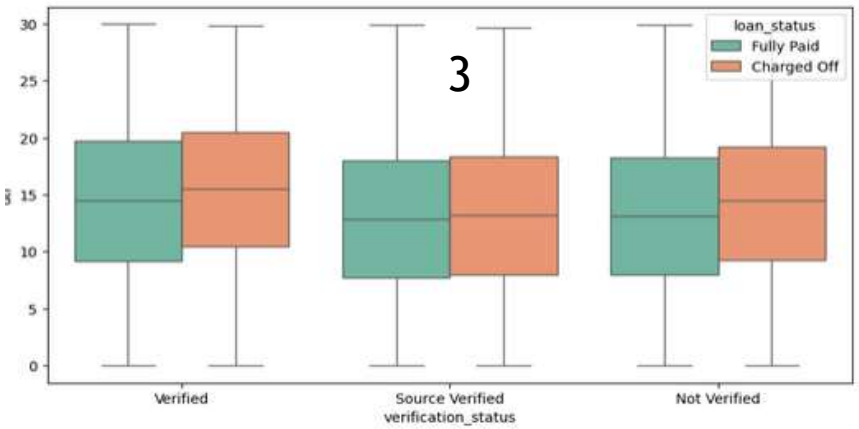
1. Across all the categories of verification status, more the loan amount, higher chances of getting charged-off

Verification status Vs int



2. Irrespective of verification status, higher the interest rate, higher are the chances of getting charged-off.

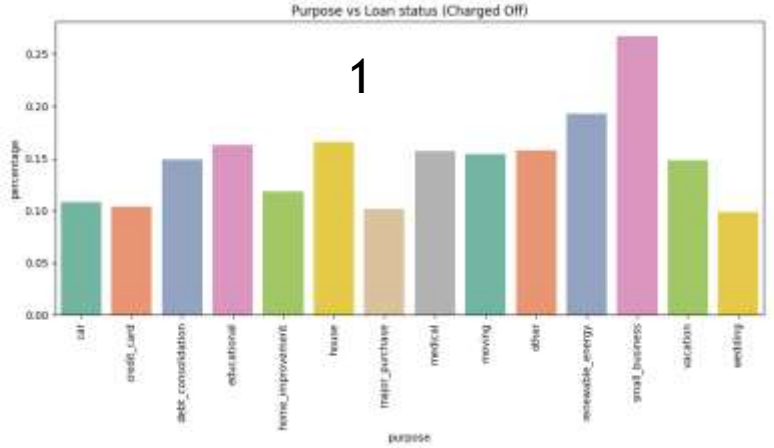
Verification status Vs dti



3. Irrespective of verification status, higher the dti value leads to higher charged-off

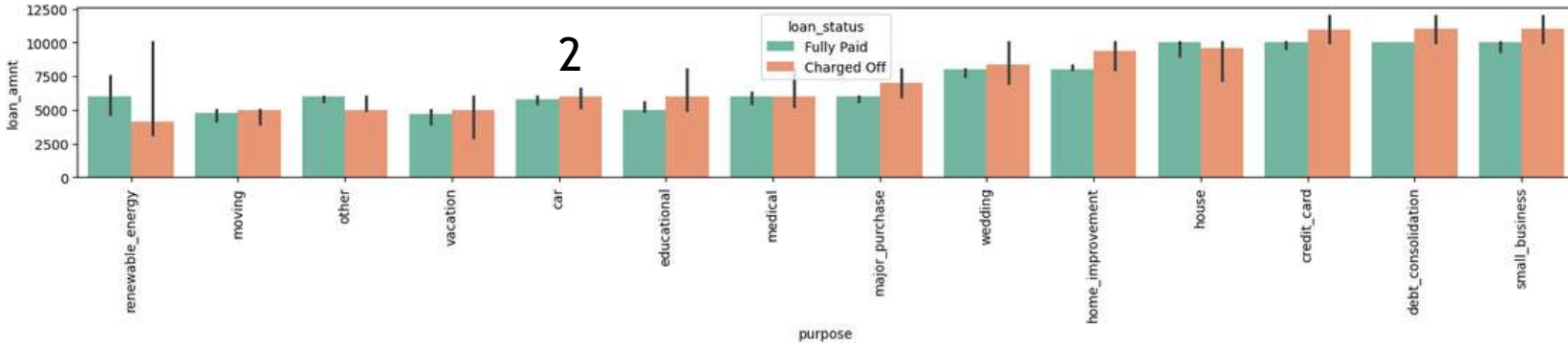
Bivariate Analysis

- On Purpose Column



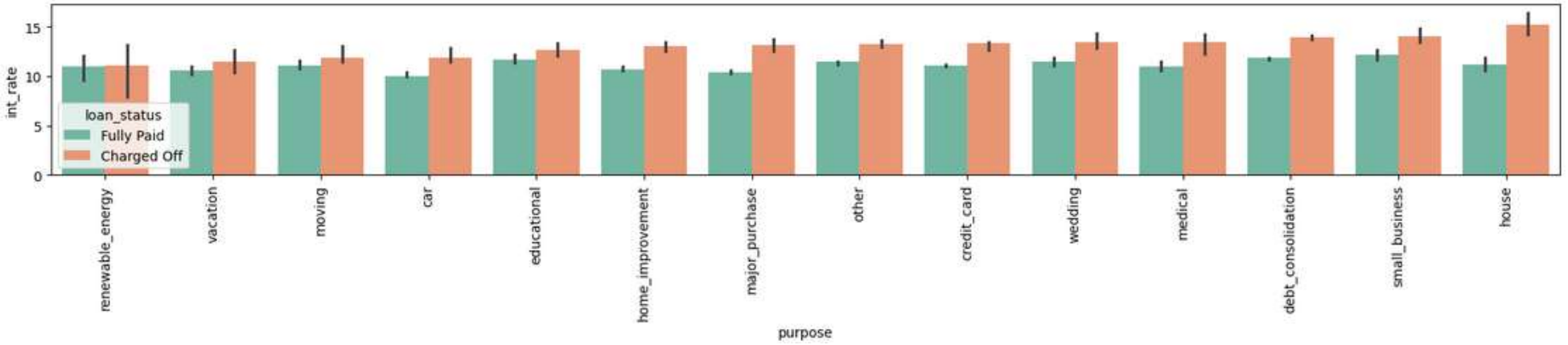
1. 'small business' purpose has the highest percentage charge off

Purpose Vs loan amount



2. Borrowers taking higher loan amount for credit_card, small_business and debt_consolidation purposes have higher default rate

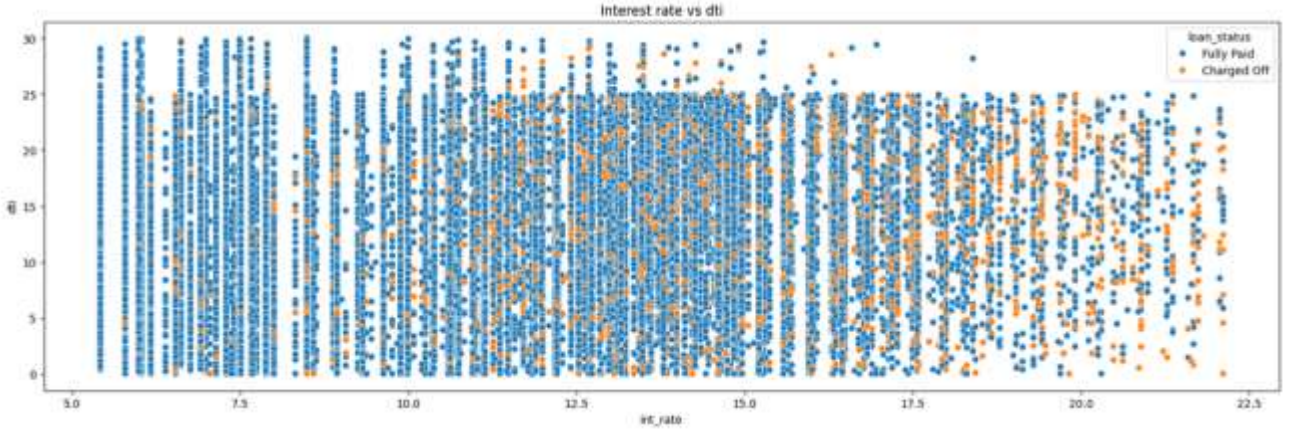
Purpose Vs int rate



3. Home loans with high interest rates are mostly defaulted. Small business and debt consolidation also have similar observation.

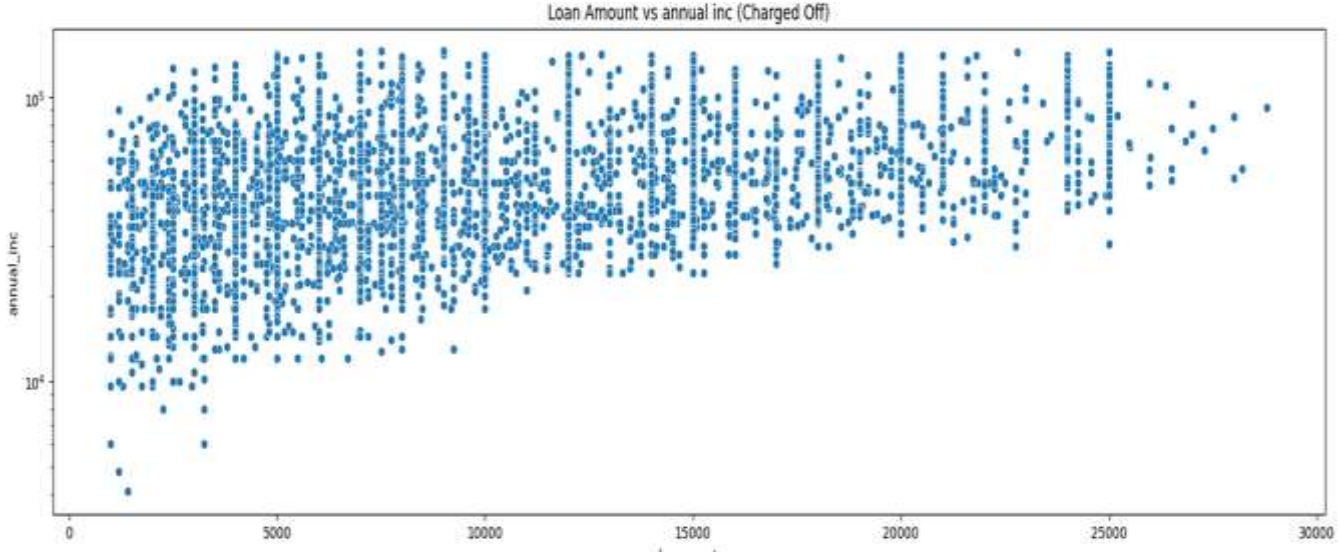
Bivariate Analysis(Numerical Vs Numerical)

- **Dti Vs Int rate**



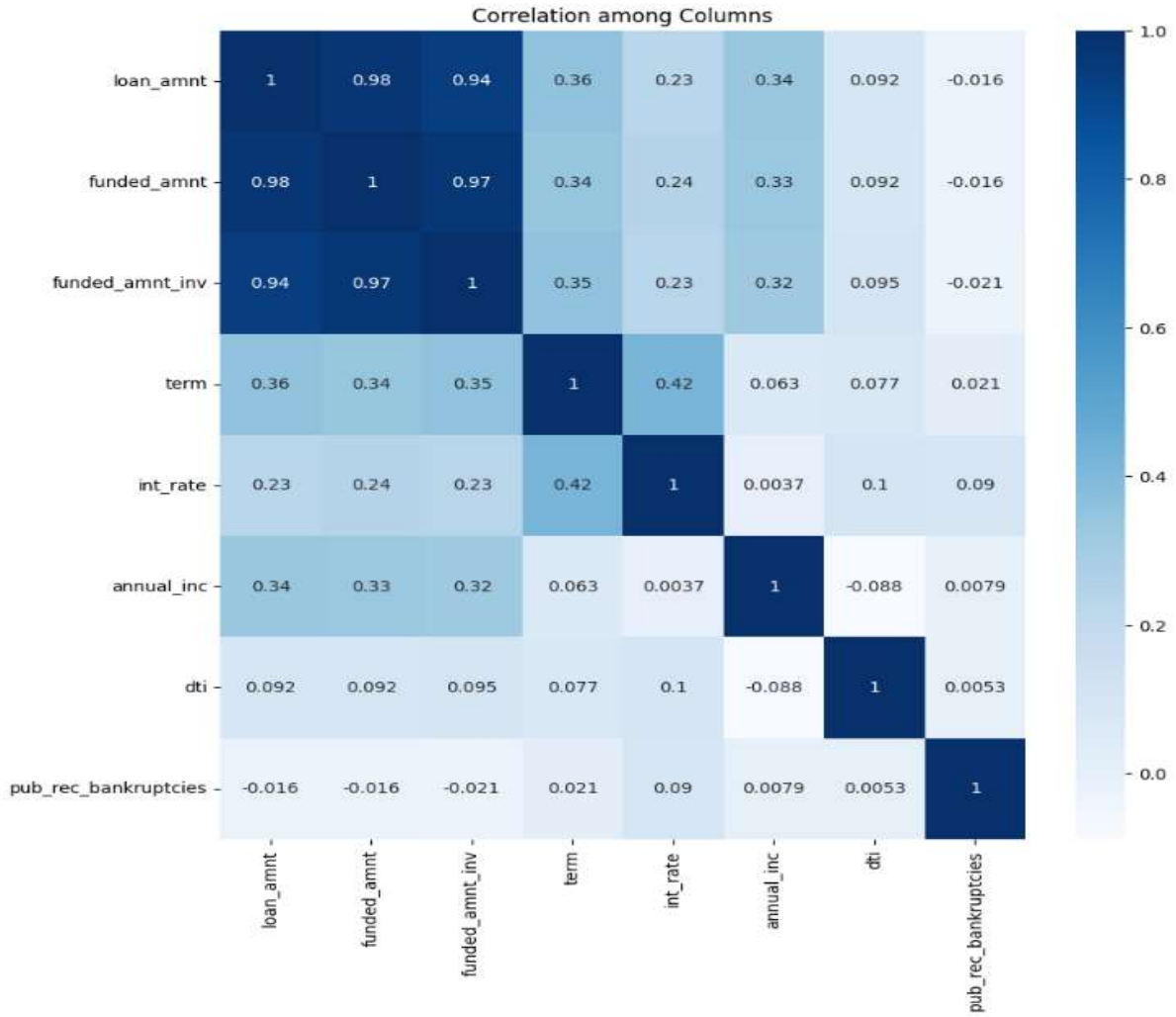
Values are spread all across, but we can see one thing here that irrespective of DTI, when the interest rates are high charged off loans are also high.

Annual Inc(Log Scale) Vs Loan amt



Since, we could not get enough actionable insights for 'Loan amount vs annual income' across loan status, we considered the logarithmic scales on both the axe for both plots.
As the annual income increases, and also the loan amount, the borrowers are more likely to fully pay the loan or less likely to get charged off

Correlation Analysis



1. Negative Correlation:
    1. loan_amnt has negative correlation with pub_rec_bankrupticies
    2. funded_amnt has negative correlation with pub_rec_bankrupticies
    3. annual income has a negative correlation with dti
2. Moderate Correlation:
    1. loan_amnt has moderate correlation with term
    2. term has moderate correlation with int_rate
3. Strong Correlation:
    1. loan_amnt has strong correlation with funded_amnt
    2. funded_amnt_inv has strong correlation with funded_amnt

Conclusions

1. The applicants are defaulted when they are given high interest rate loans over a period of 60 months and also have higher side debt to income ratios.

2. Borrowers applying for low grade loans with high loan amount and high interest are surely going to default. The company should refrain from providing such loans.

3. Borrowers with high employment lengths going for high loan amounts are more likely to default.

4. Top 3 states for default are CA, FL and NY. For these states the company should have tighter process on approving the loans

5. Applicants opted for small business purpose with high interest loans are the most prone to default

6. For homeownership as OTHER Category with high interest rate, higher loan amount are the most to get defaulted