Lending Club Case Study

SUBMITTED BY:

ATUL CHAUHAN

SOWMYA RENUKESHWARA

BUSINESS UNDERSTANDING

- The following analysis is 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 must decide 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 provided data contains the 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.
- In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

BUSINESS UNDERSTANDING

- When a person applies for a loan, there are two types of decisions that could be taken by the company:
  - 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
  - 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)

BUSINESS OBJECTIVES

- Lending club data containing information about past loan applicants and whether they ‘defaulted’ or not are provided.
- Apply Exploratory data analysis on the data provided
- Find the key variables that influences a customer from being ‘default’
- Identification of patterns between variables which are indicators of customer being ‘default’
- Recommendations from the analysis to reduce default

APPROACH

Data Cleaning

- Missing Values Treatment
- Remove data that does not add any value to analysis
- Removing Customer Behaviour Variables
- Fixing Datatypes
- Removing Outliers

Data Analysis

- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis
- Inference/Conclusions

Recommendations

- How to reduce customers from being ‘default’

DATA CLEANING

- **Missing Value Treatment (Columns):**
  - All columns with 100% missing values were removed.
  - Then we removed the columns with maximum number of missing values.
    - 'next_pymnt_d' - 97% missing values
    - 'mths_since_last_record' - 93% missing values
    - 'mths_since_last_delinq' - 64% missing values
    - 'desc' - 32% missing values

      Note: We have removed the 'desc' column as well, even though it has only 32% missing values. This is because, there is not much analysis can be done on this column as this column just tells us the description of the loan.
  - Columns with constant values as below removed.
    - 'collections_12_mths_ex_med', 'chargeoff_within_12_mths', 'tax_liens', 'delinq_amnt', 'acc_now_delinq', 'application_type', 'policy_code', 'initial_list_status', 'pymnt_plan'
  - Other columns which did not add value to the analysis like url, zip_code and other 20 'customer behavior variables' were removed.

DATA CLEANING

- **Missing Value Treatment (Rows):**
  - Checked for the rows with 100% missing values but didn't find any.
  - Checked for duplicate values but didn't find any.
  - Then we removed the rows with missing values as below,
    - 'emp_title'
    - 'emp_length'
    - 'pub_rec_bankruptcies'
    - 'last_pymnt_d'
    - 'revol_util'
    - 'title'

Note: We had seen that the missing values were very less as compared to the length of data that we had. Even if we removed the rows with these missing values, it wouldn't make much difference to our analysis of the data. Hence, we removed these rows one by one, next.

DATA CLEANING

- **Fixing Datatypes:**
  - We also converted the data types of few columns to help in our analysis Eg:'term', 'emp_length'

- **Removing Outliers:**
  - Outliers were removed from the column 'annual_inc'

- **Data imputation:**
  - No imputation was performed in the data analysis

DATA ANALYSIS

After having addressed the data quality issues, we now move on to the analysis part. Our analysis is divided into 03 parts namely:

- ***Univariate Analysis*:** Under univariate analysis, we analyzed single individual columns and derived some really important insights which helped further in our analysis. We checked various categorical and numerical variables with the help of suitable plots like histograms, boxplots etc.

- ***Segmented Univariate Analysis*:** Here, we segmented our variables into the cases where the loan has been fully paid and the cases where the loan has been defaulted. The analysis was aided by plotting appropriate plots.

- ***Bivariate Analysis*:** Lastly, we checked the correlation between various variables in order to arrive at the most important combinations along with making business and analytical sense.

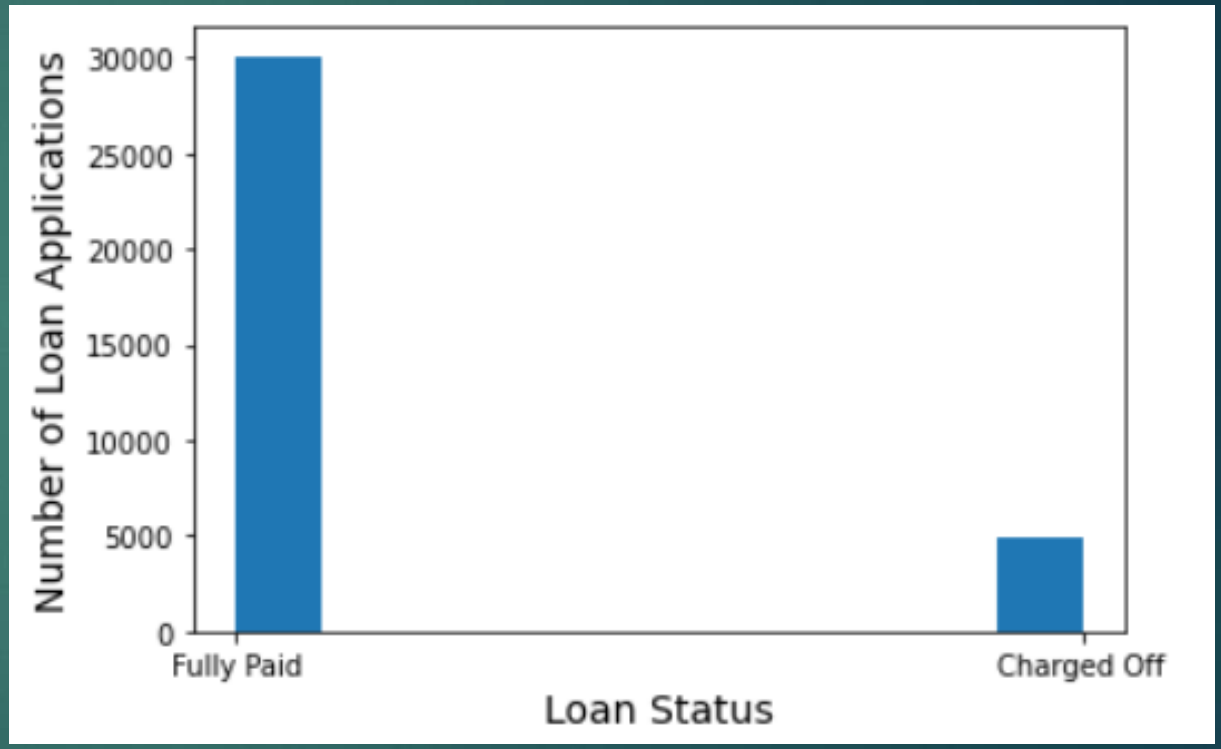
UNIVARIATE ANALYSIS

- **We performed univariate analysis on the below columns and derived certain insights:**
  - **Categorical Variables:**
    - 'loan_status' column
    - Derived 'issue_month_year' column
    - 'Grade' column
    - 'emp_length' column
  - **Numerical Variables:**
    - 'int_rate' column
    - 'loan_amnt' column
    - 'dti' column

UNIVARIATE ANALYSIS – CATEGORICAL VARIABLE

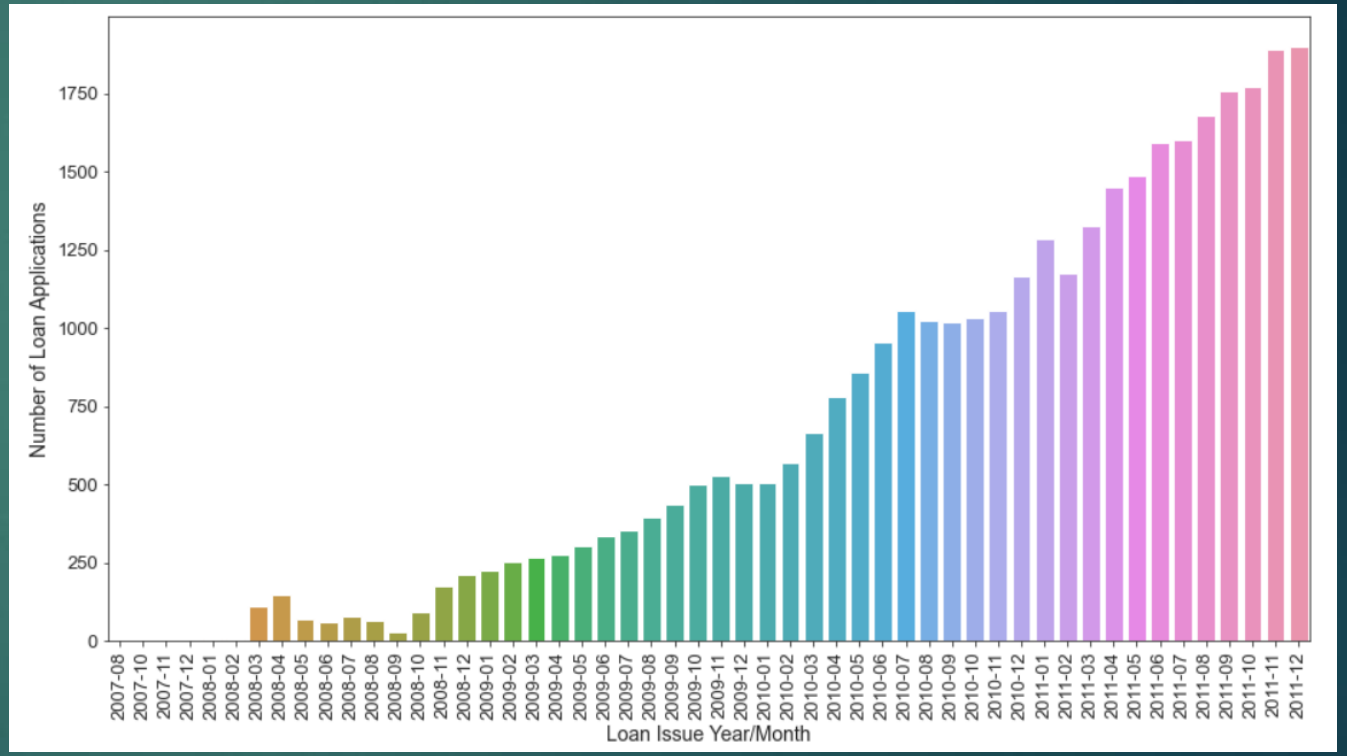
- **'loan_status'** column - The ones marked as 'Current' are neither fully paid nor defaulted and will not add value to the analysis and hence removed

  After removal, upon analysis, it found that the charged off loans are only 14% while most of the loans are fully paid



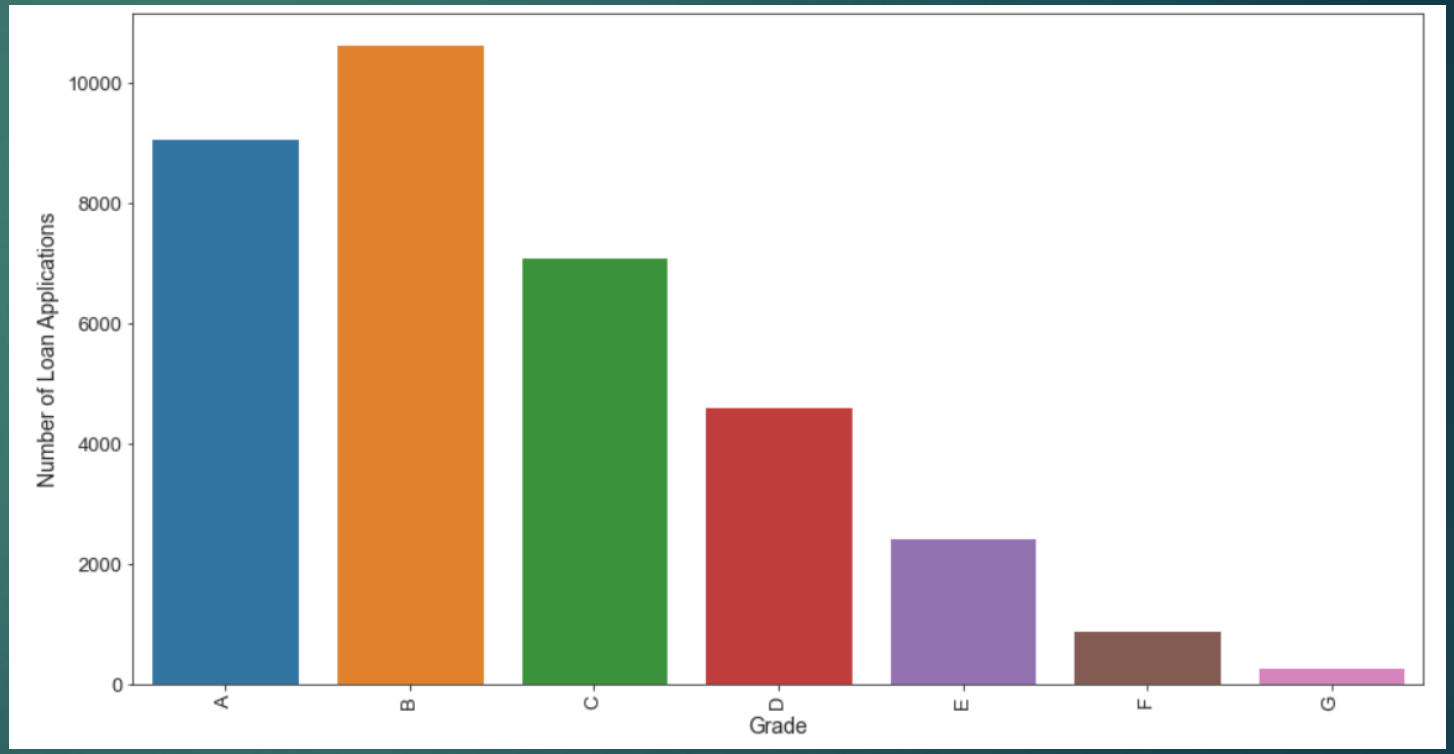
UNIVARIATE ANALYSIS – CATEGORICAL VARIABLE

- '**issue_month_year' column** - The loan applications keep increasing each month starting year 2007. Though, there is a dip in the loan applications from the end of second quarter of 2008 , all the way to third quarter. However, it gradually increases in the fourth quarter



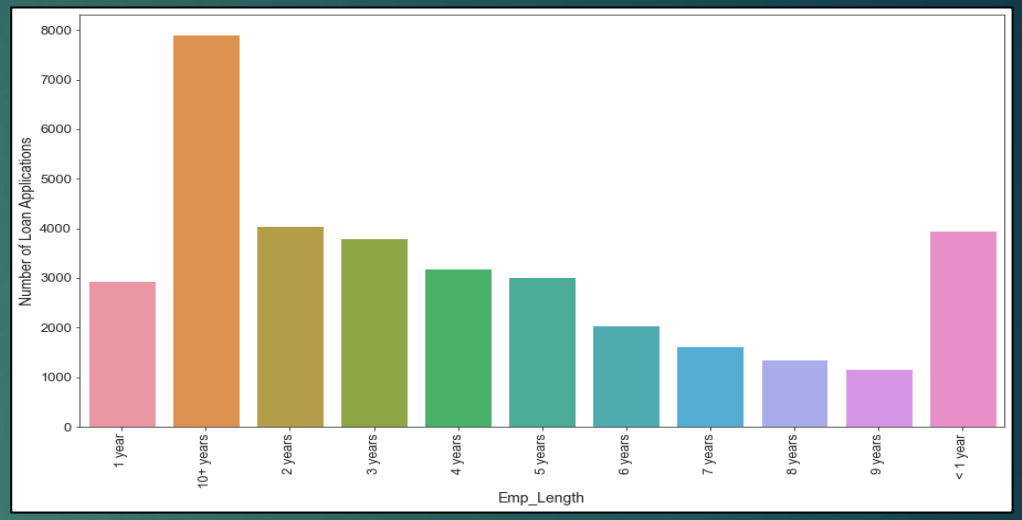
UNIVARIATE ANALYSIS – CATEGORICAL VARIABLE

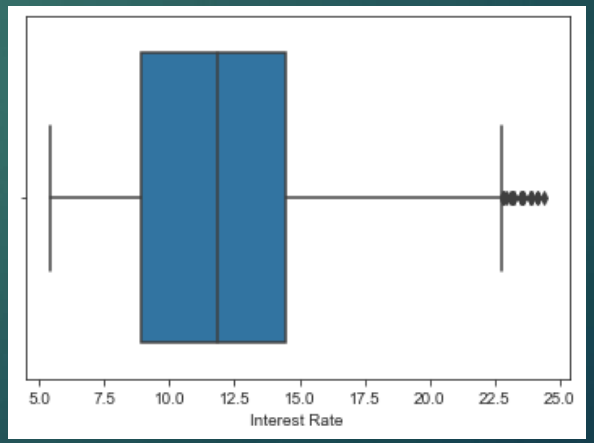
- **'Grade'** column - we can infer that most of the loans are graded as 'B'.



UNIVARIATE ANALYSIS – NUMERIC VARIABLE

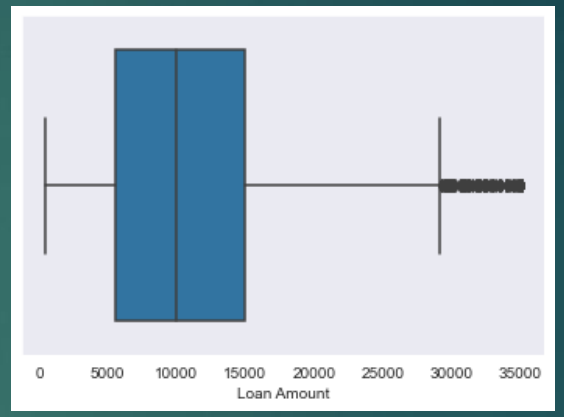
- 'emp_length' column – We can infer most of the loan applicants have experience of 10 years or more. Also, the employees with an experience of 1 year or less are the second highest loan applicants
- 'int_rate' column – We can infer most of the loans have interest rates between 8 and 15

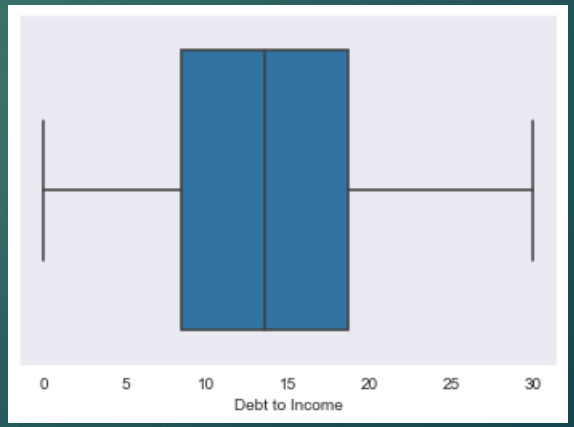




UNIVARIATE ANALYSIS – NUMERIC VARIABLE

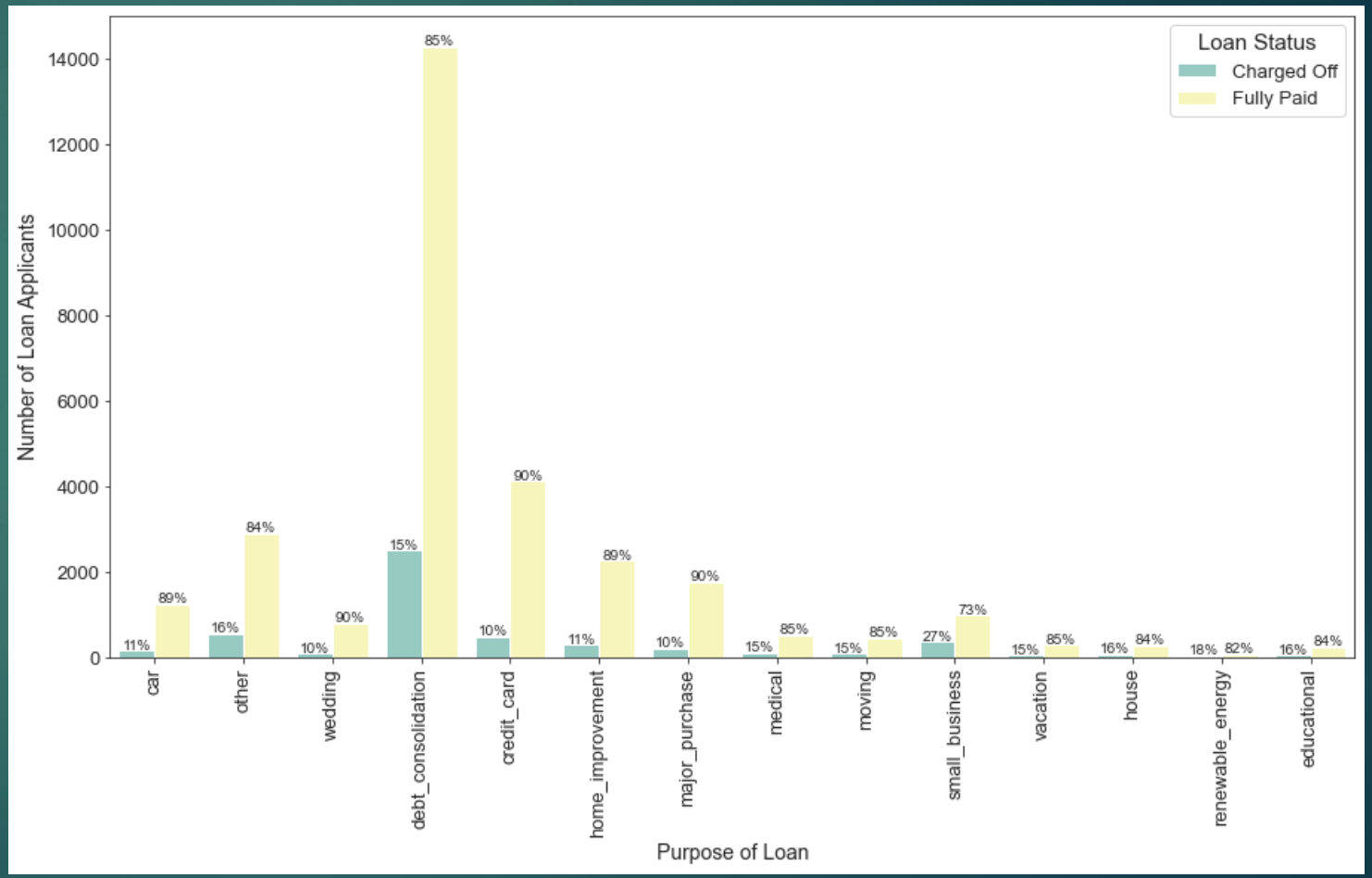
- 'loan_amnt' column - maximum loan applications are for the amount between 5000 and 15000
- 'Debt to Income' column - most of the loans have dti between 8 and 19



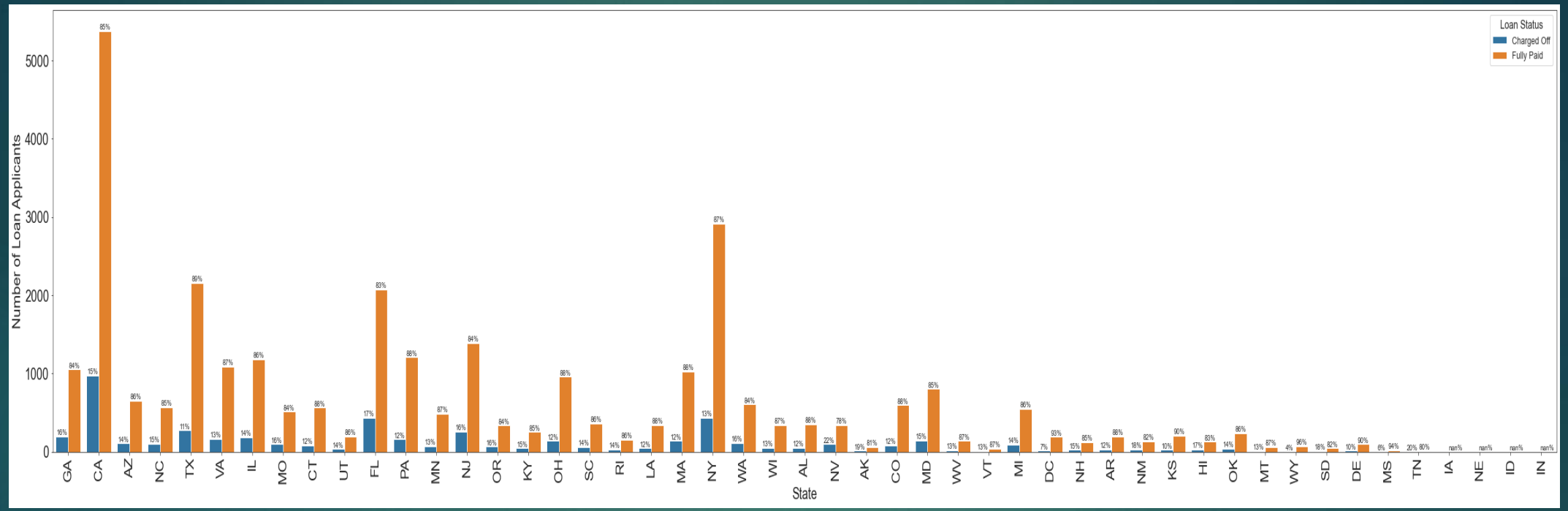


SEGMENTED UNIVARIATE ANALYSIS

- **'purpose' column:** the most no of loan applications are for 'debt consolidation' purpose. However, the default cases are only 15%.
- However, loan_applications for small business have most percentage of default cases, i.e. 27%.
- We may conclude if purpose is 'business', more likely to default than any other purpose.
- One more purpose worth noting here is 'renewable_energy' which is 18%



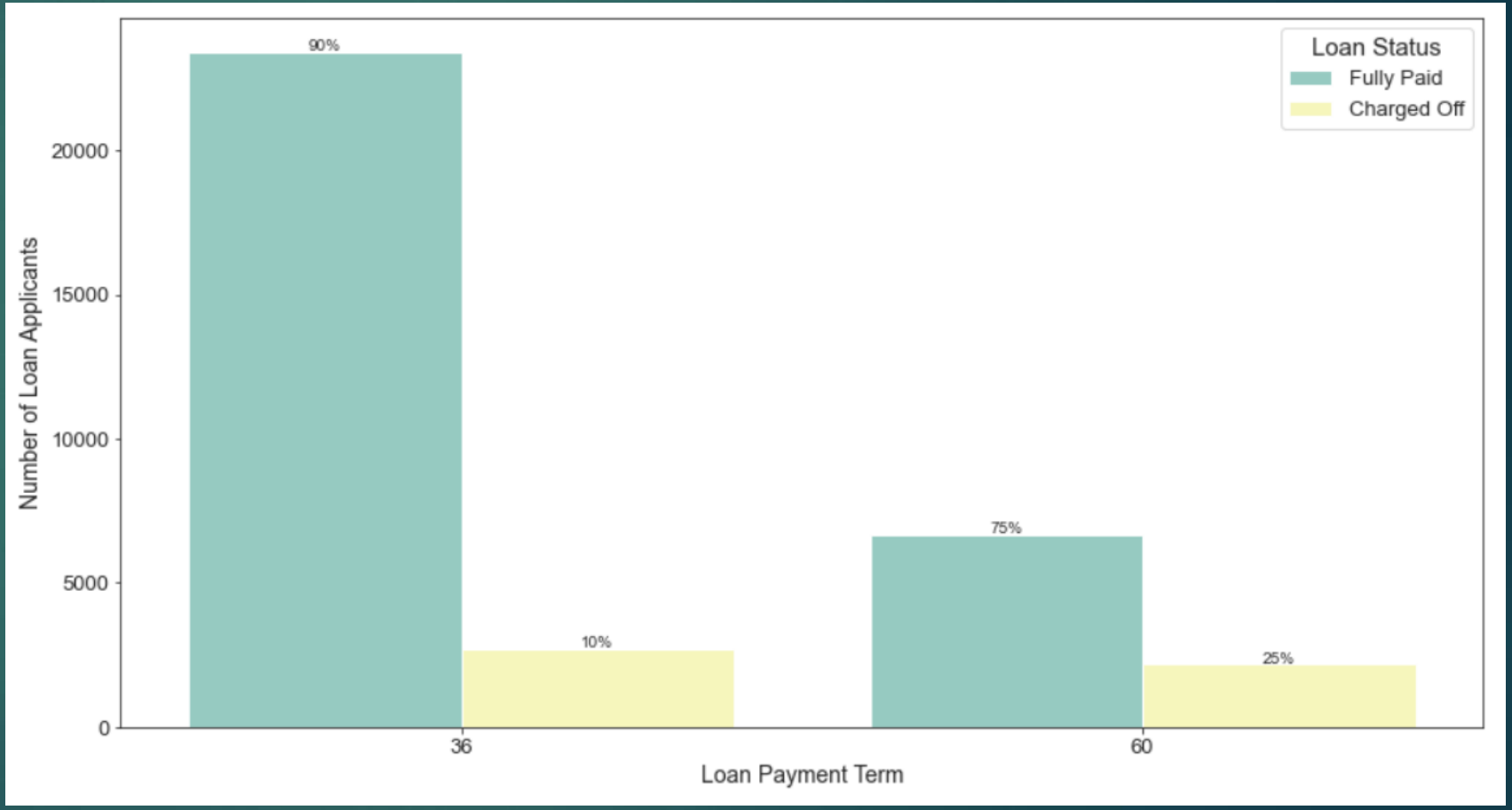
SEGMENTED UNIVARIATE ANALYSIS



- In **addr_state** variable, the highest number of loan applications are received from CA, NY, TX, and FL. However, when we check the default ratio for these, we see that for CA, it is 15%, NY is only.13%, TX is just 11%, while FL is the highest of the four, being 17%.
- The highest percentage of default cases out of all the states are seen in NV which is 22%. Also, we can see TN which has second highest percentage of default cases, 20% followed by AK, 19% and SD & NM each having 18%
- Also, we can see that the states 'IA', 'ID', 'NE' and 'IN' have 0% of loan default cases.

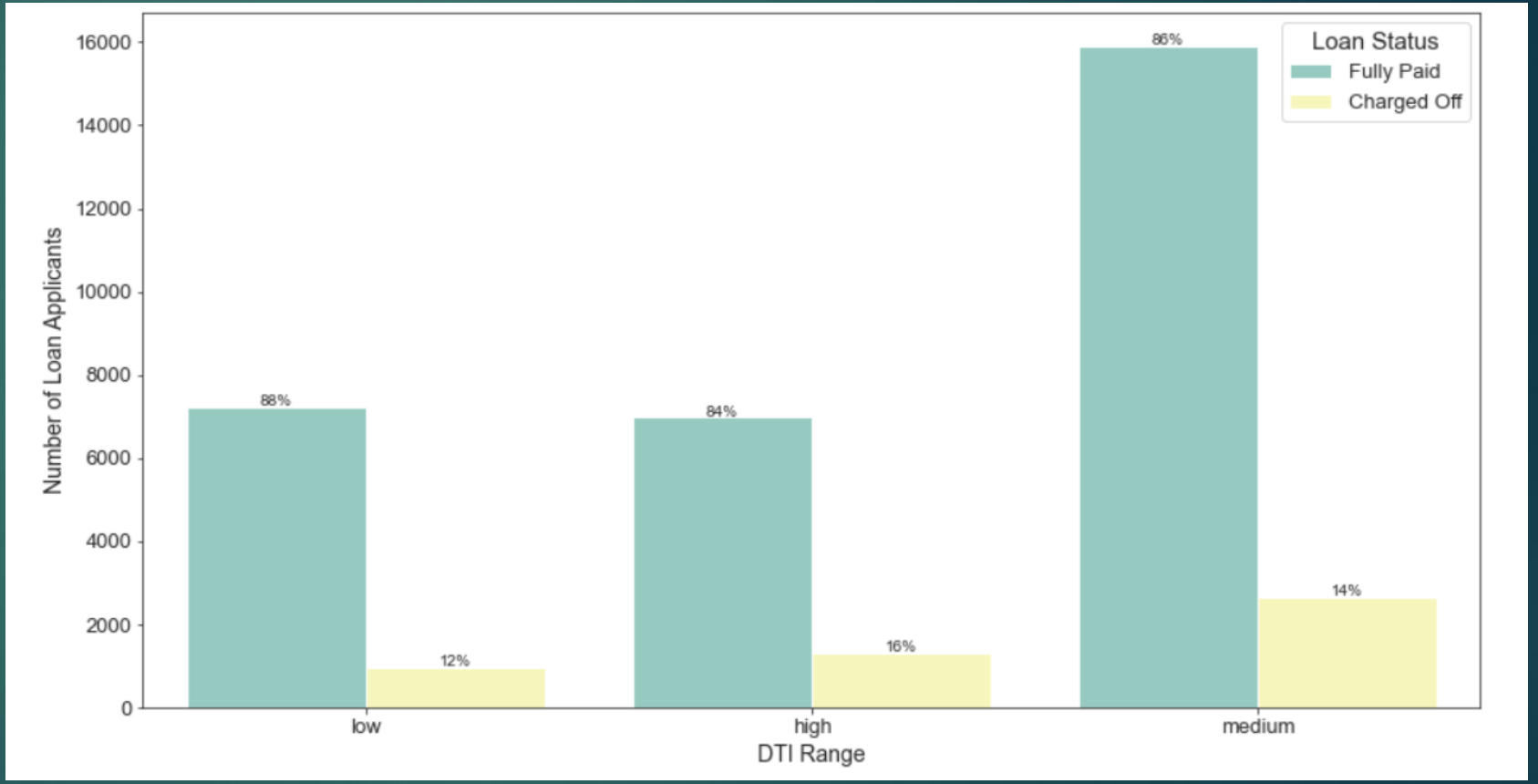
SEGMENTED UNIVARIATE ANALYSIS

- From the **'term'** column, we can deduce that the loan applications with 60-month term have higher percentage of loan default cases - 25%. This can be strong indicator which can help us tell if the loan may default or not.



SEGMENTED UNIVARIATE ANALYSIS

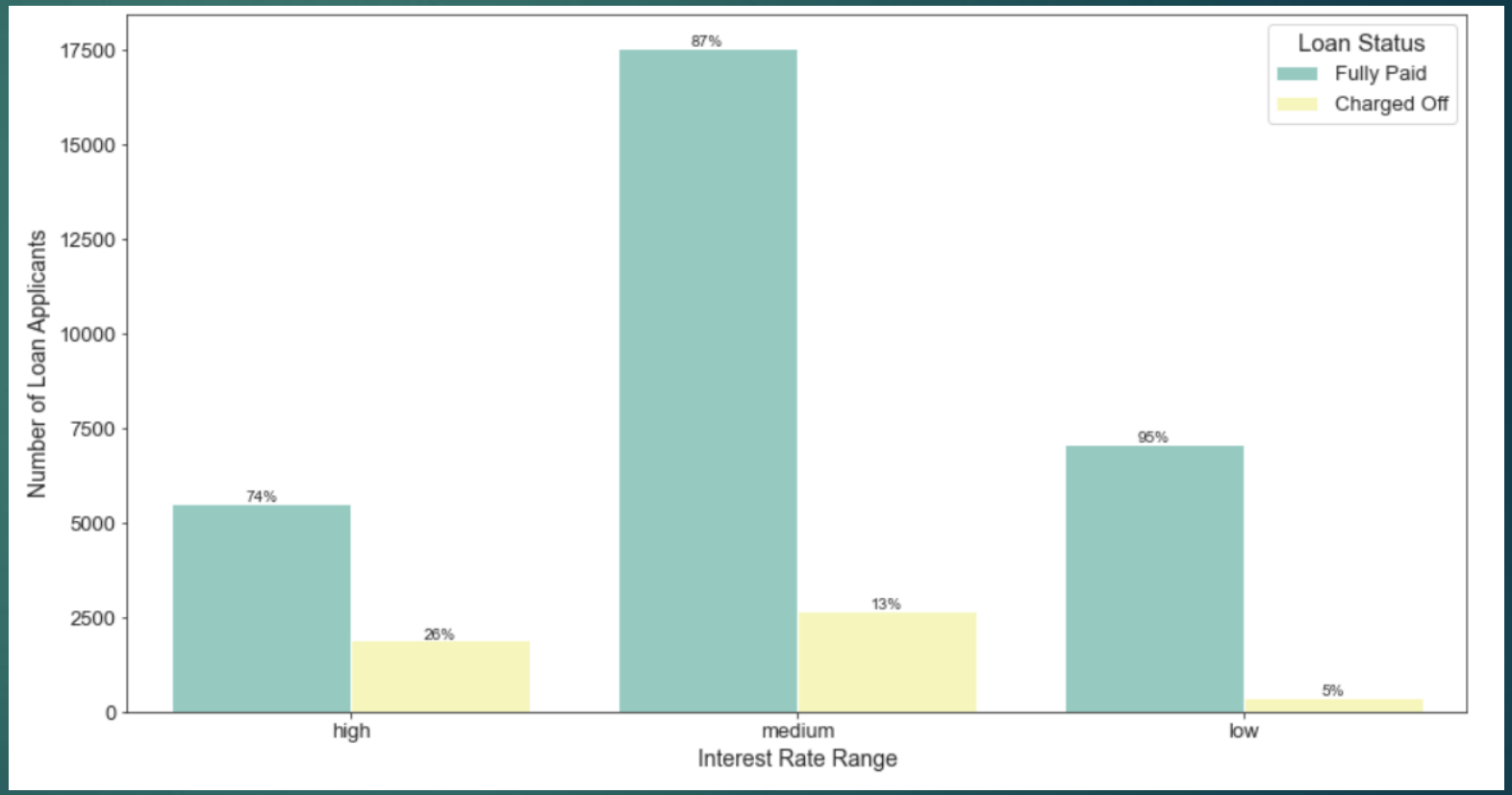
- For the **'dti'** column, we see that the maximum loan applications have dti between 8 and 19 i.e. medium range. However, the highest percentage of loan default cases is within high range, meaning, the loan applications where the dti is greater than 19, there are higher chances of the loan getting defaulted.



SEGMENTED UNIVARIATE ANALYSIS

- For the **‘int_rate’** column, we see that the maximum loan applications have interest rates between 8 and 15 i.e. medium range. However, the loan applications with interest rate under 'medium' range have only 13% of loan default cases. If we take close look at other two categories, we see that 'low' category of interest rates has the least ratio of loan default cases. The 'high' category has the highest percentage of loan default cases, i.e. 26%.

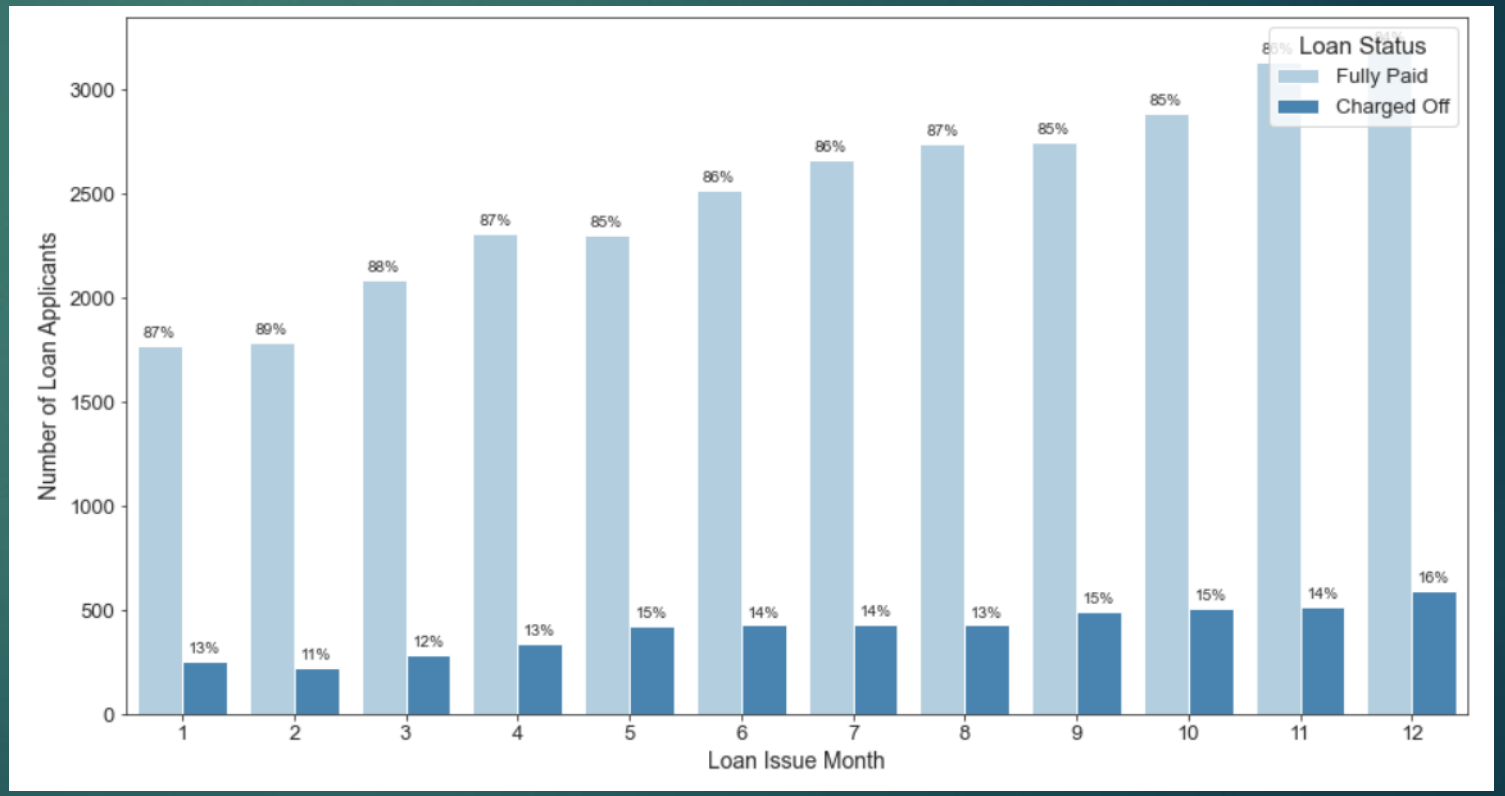
  This is a really strong indicator that tells us whether the loan will be defaulted or not.



SEGMENTED UNIVARIATE ANALYSIS

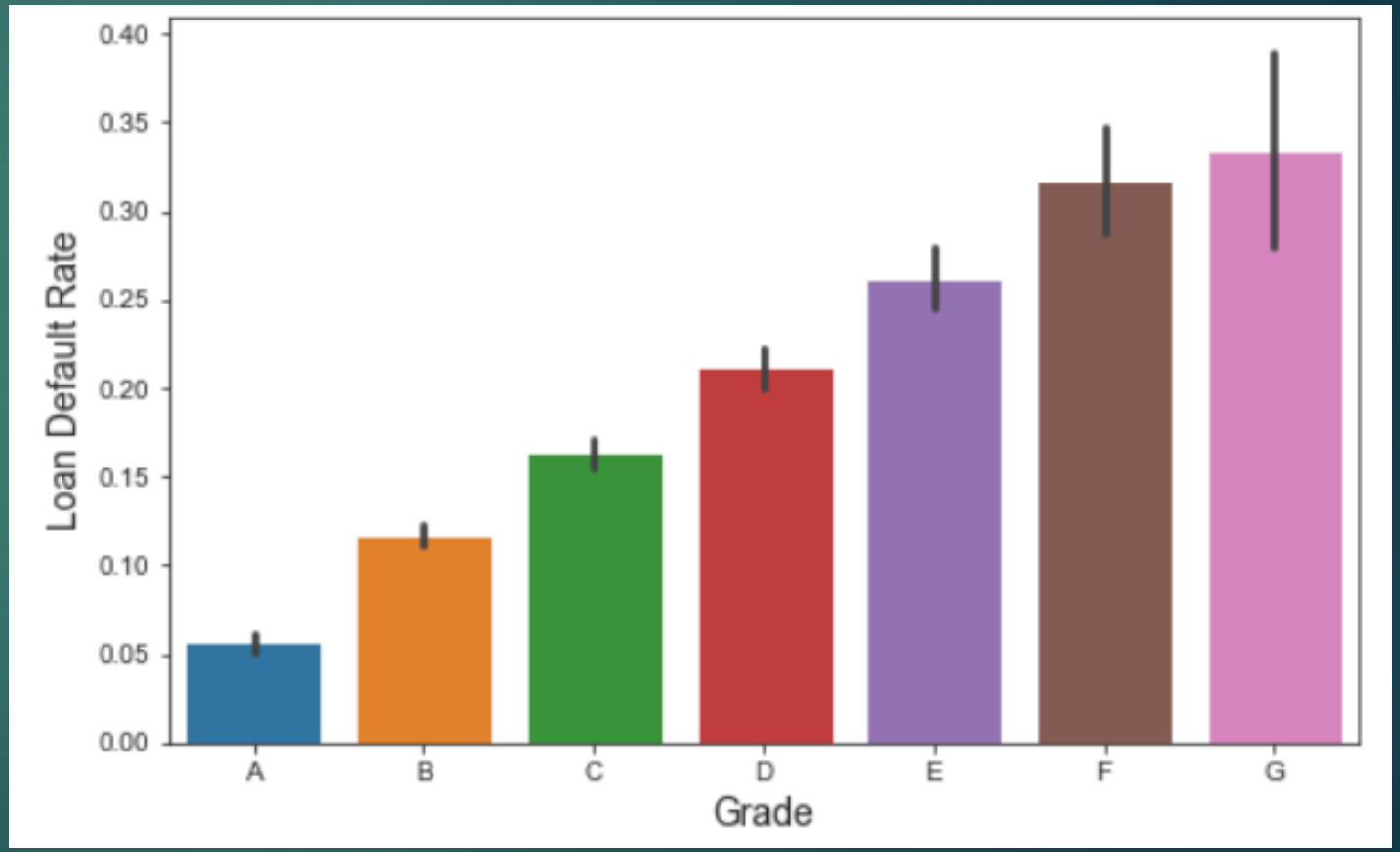
- **Loan_issue_month:** As per the graph it is quite visible that december has the highest number of loan approvals in a particular year. This may be becasue the organisation has to achieve targets in december due to year end.

  Also, we see here that december also has the highest percentage of loan default cases than any other month. This may be because that these loan applications are just verified without income source verification. (as we saw in our analysis of the verification_status column, that the highest percentage of loan default cases occur when the loan is verified but without income source verification).



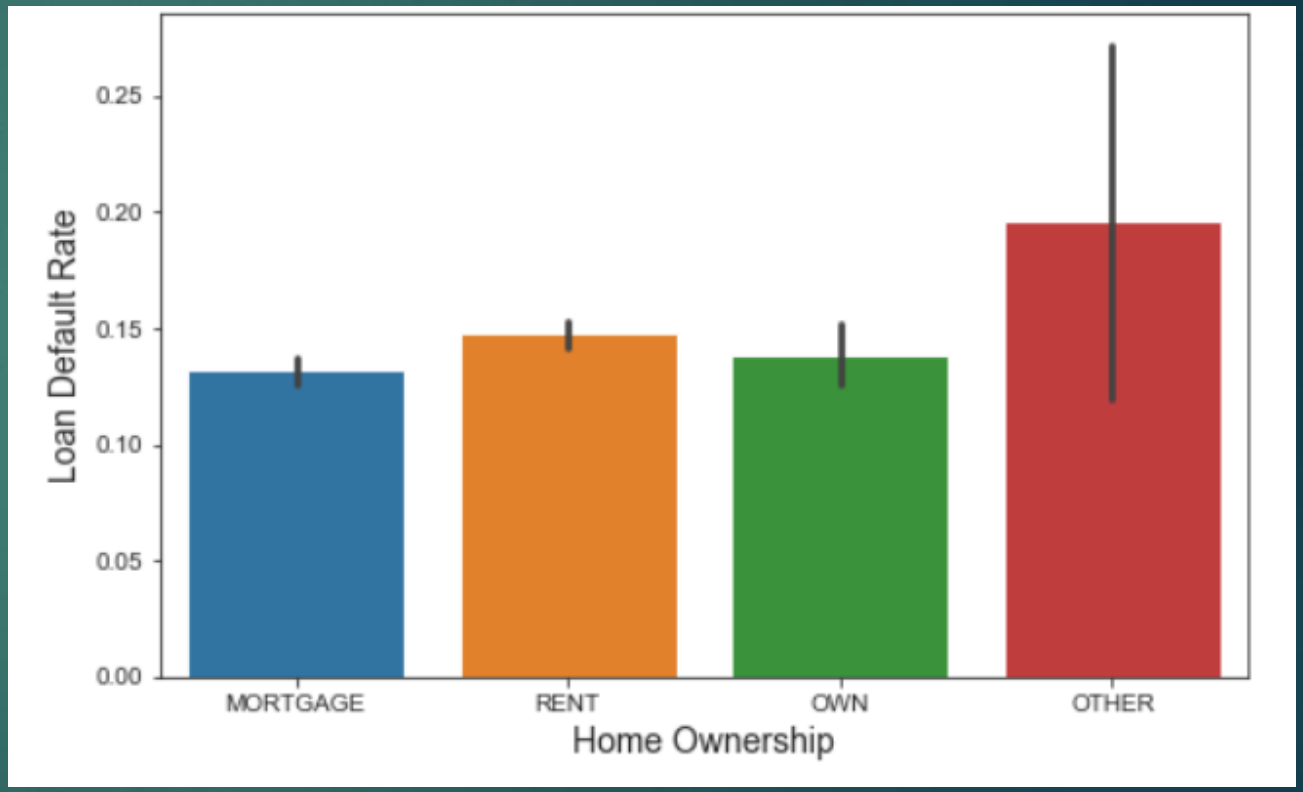
BIVARIATE ANALYSIS

- **grade:** Grade is a very good indicator of default cases. As we move towards right from A to G, the number of total loan applications decrease.
- However, with this, the percentage of charged off cases increase.
- The grade 'A' has the least default ratio whereas the grade 'G has the highest default ratio of 33%.
- Thus, we can further conclude that if the loan application is allotted a grade 'G' it is more likely to be a default case than any other grade.



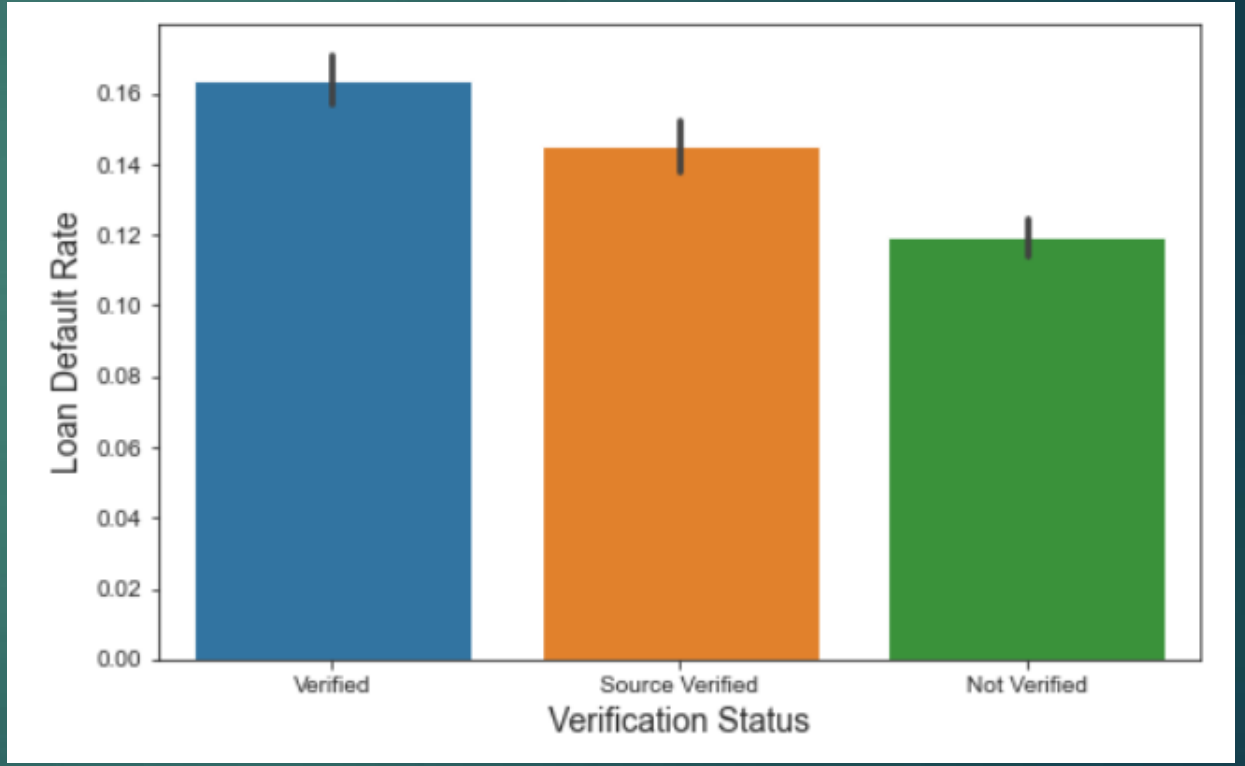
BIVARIATE ANALYSIS

- **home_ownership:** "Other" has the highest percentage of default cases followed by the "Rent" category. The loan applicants with category "Mortgage" have the highest loan repayment ratio. This may be because the properties of the applicants are at stake due to which they tend to repay the loan.



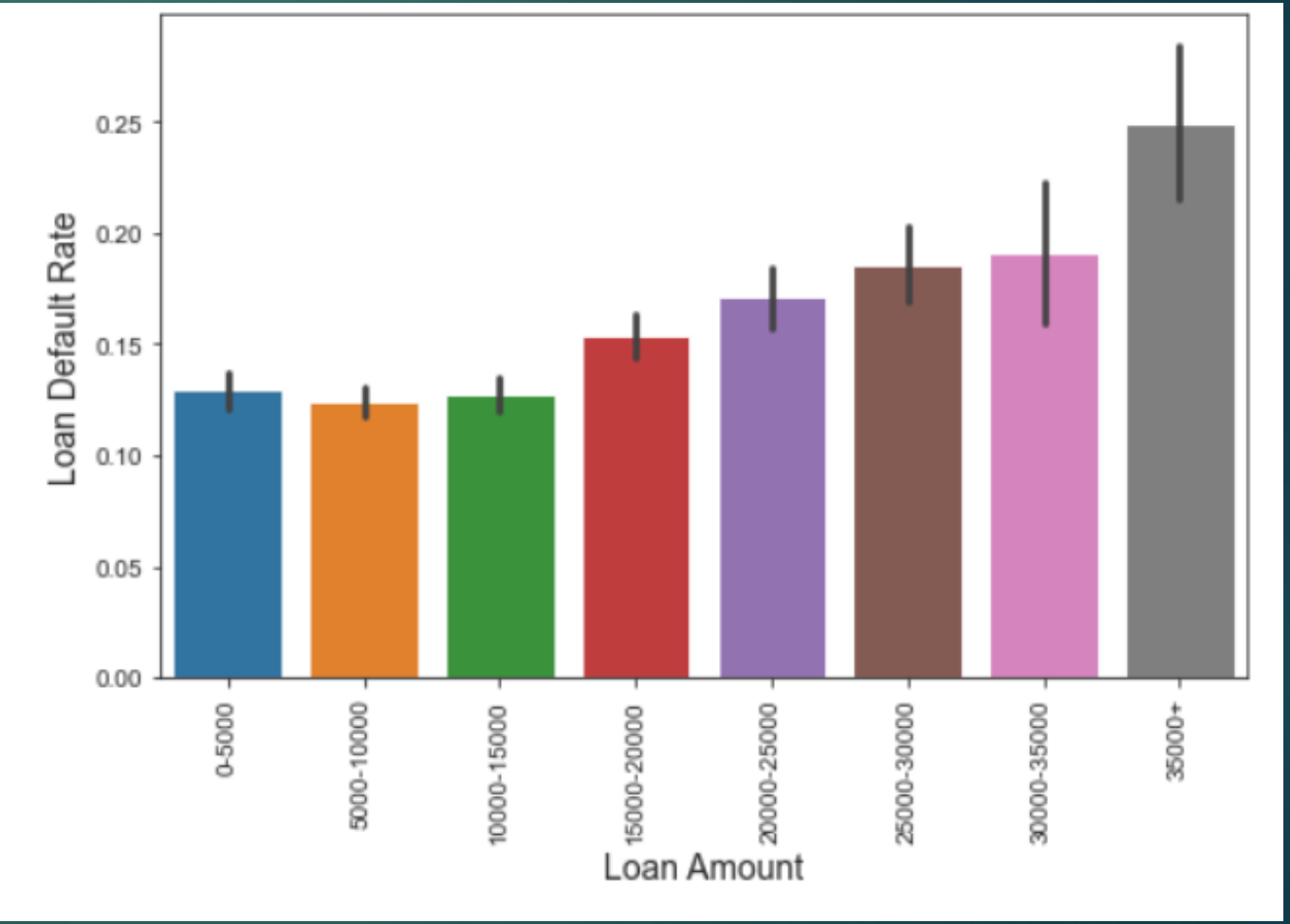
BIVARIATE ANALYSIS

- **Verification_status:** The loans which are verified by LC but the income source is not verified have 16% chances that it will be a default case.
- The loans with source verified have 15% default ratio. However, it is very surprising to note here that the loan cases which are not at all verified have only 12% default ratio.



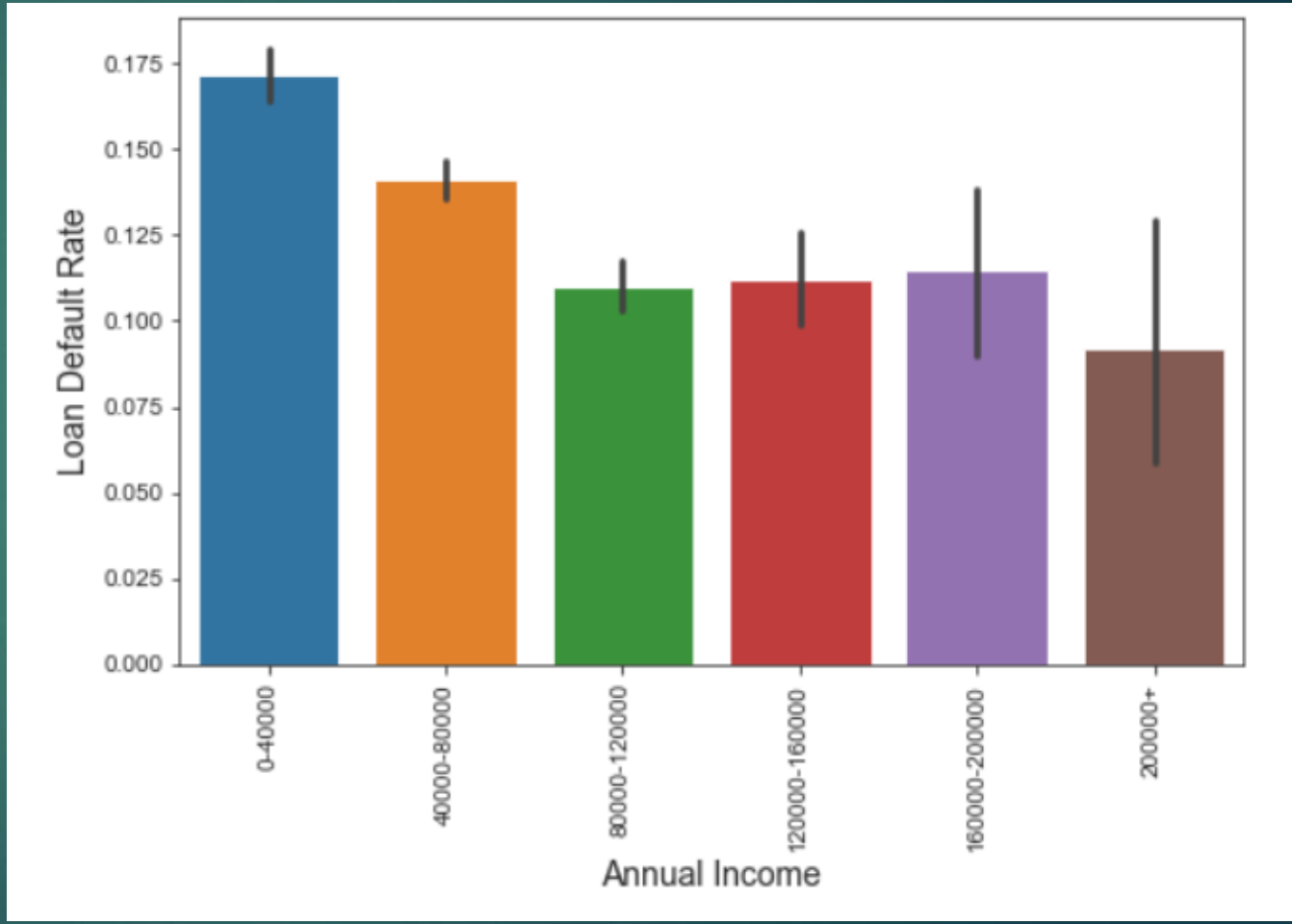
BIVARIATE ANALYSIS

- **Loan_amount:** Looking at the trend between the loan status and loan amount, we can infer that as the loan amount is increasing, the percentage of default cases is also increasing.



BIVARIATE ANALYSIS

- **Annual Income:** We see that the loans where the applicant's annual income is lower, have the highest chances of loan defaults.



CONCLUSIONS

- Applicants whose annual income is lower, have the highest chances of loan defaults.
- Loan applications for ‘small business’ and ‘renewable_energy’ as purpose tend to have the most percentage of default cases.
- Loans with higher grades have highest percentage of loan defaults.
- Higher interest rates mean high chances of defaults.
- Similarly, higher Debt to income ratio means higher default cases.
- Loans of higher values have the highest percentage of default cases as well.
- Loans with loan term as 60 months have higher chances of default cases as compared to 30 months.
- We also saw that the loans which are verified only (and not income source verified) do have higher chances of loan defaults.
- One more point worth noting is that the loan applications in December have higher percentage of loan defaults. Again, this can be linked to verification of the loan applications. The loans are approved without proper verification since the organization wants to achieve their year end targets.
- We also saw that state can also be one deciding factor – States like ‘NV’, ‘TN’, ‘AK, ‘SD & ‘NM’ have higher default percentages.

RECOMMENDATIONS

- Additional verification checks should be put in place to avoid defaults for the mentioned points.
- Higher interest rates should be levied on the loan applicants with lower annual income.
- Lower loan amount should be sanctioned to the applicants with purpose “small_business” or “renewable _energy”.
- Proper income source should be verified before sanctioning any loans in order to avoid any default cases in future.
- Loan tenure should be kept as low as possible.
- Higher interest rates should be levied for states such as ‘NV’, ‘TN’, ‘AK, ‘SD & ‘NM’ .

THANK YOU