LENDING CLUB CASE STUDY

EXECUTIVE PG PROGRAMME IN MACHINE LEARNING & AI - JANUARY 2022

Rahul Kunjumon
Rashmi Kumbugowdana

PROBLEM STATEMENT

We are working for a consumer finance company which specialises 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

Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If we can identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants' using EDA is the aim of this case study.

OBJECTIVE

We have been provided with a dataset which 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.

PROCEDURE

- **Understanding the Data**

We used Microsoft Excel to view the data, understand various fields available. We were aided by a Data Dictionary file which had one liners on the column data.

We used certain youtube videos to understand the lending domain and various analysis done in this domain

- **Cleaning the Data**

Loan.csv file given to us had a raw data which is difficult to process at the first instant. Hence, it had to be cleaned to perform the data analysis.

We used Python libraries Pandas to import the file and perform the cleansing. Various actions performed were :

- Removing Columns with more than 60 % Null values
- Removing columns with just single unique data as it won't return any valid information as all rows have same value
- Cleaning rows having Null Values
- Necessary columns with null values were assigned appropriate values such that they don't return errors during Analysis
- Convert columns with dates to python readable format
- Convert columns with strings like %, months etc. to integer format.(E.g., 8 Months to 8, 90% to 90)
- Convert values with decimal to its nearest 2 places

PROCEDURE CONT'D

- **Data Analysis**
- **Univariate Analysis :**

1. Analysis on Loan Status
2. Analysis on Funded Amount over Charged off Loans
3. Analysis on Interest Rate over Charged off Loans
4. Analysis on Income over Charged off Loans
5. Analysis on DTI (Debt To Income Ratio) over Charged off Loans
6. Analysis on Loan Term over Charged off Loans
7. Analysis on Employment Length over Charged off Loans
8. Analysis on House Ownership over Charged off Loans
9. Analysis on Installment Size over Charged off Loans
10. Analysis on Verification Status over Charged off Loans

PROCEDURE CONT'D

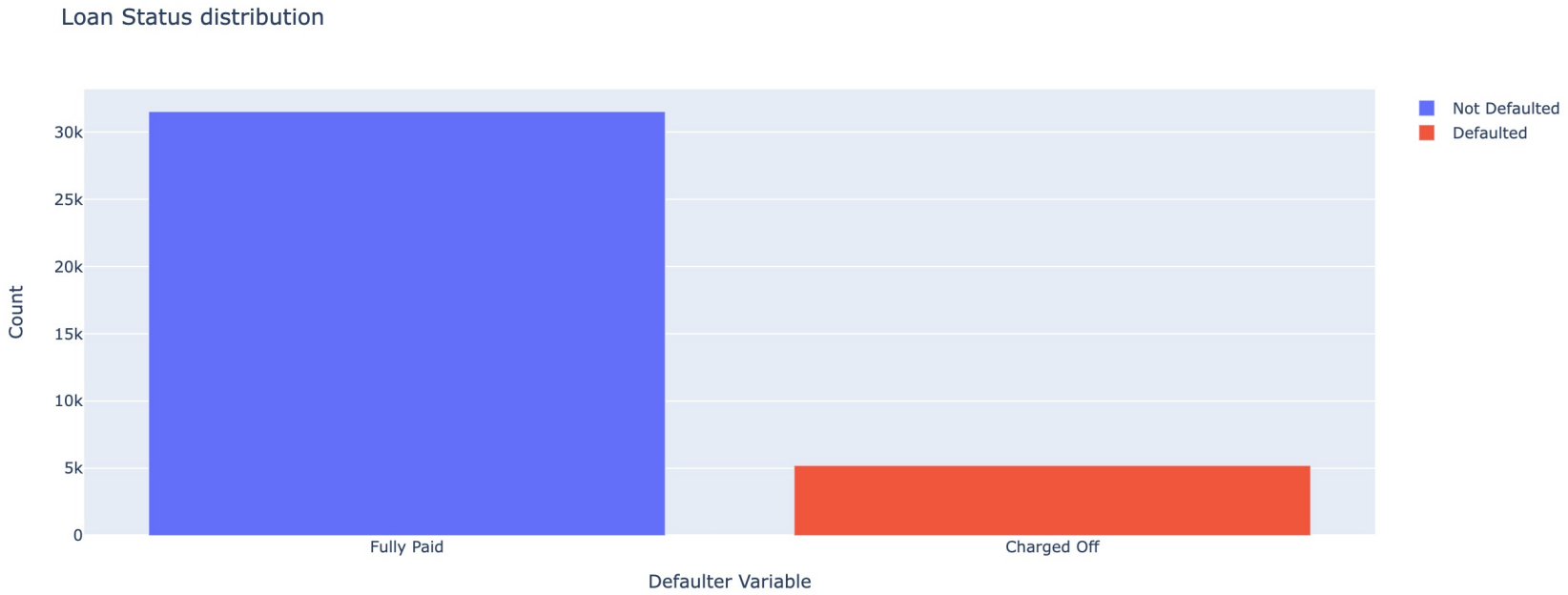
- **Data Analysis**

- **Bivariate Analysis :**

1. Analysis on Funded Loan over Annual Income
2. Analysis on Funded Loan over Grade
3. Analysis on Funded Loan over Loan Status and Verification Status
4. Analysis on Funded Loan over Grade and Term

UNIVARIATE ANALYSIS

- Analysis on Loan Status

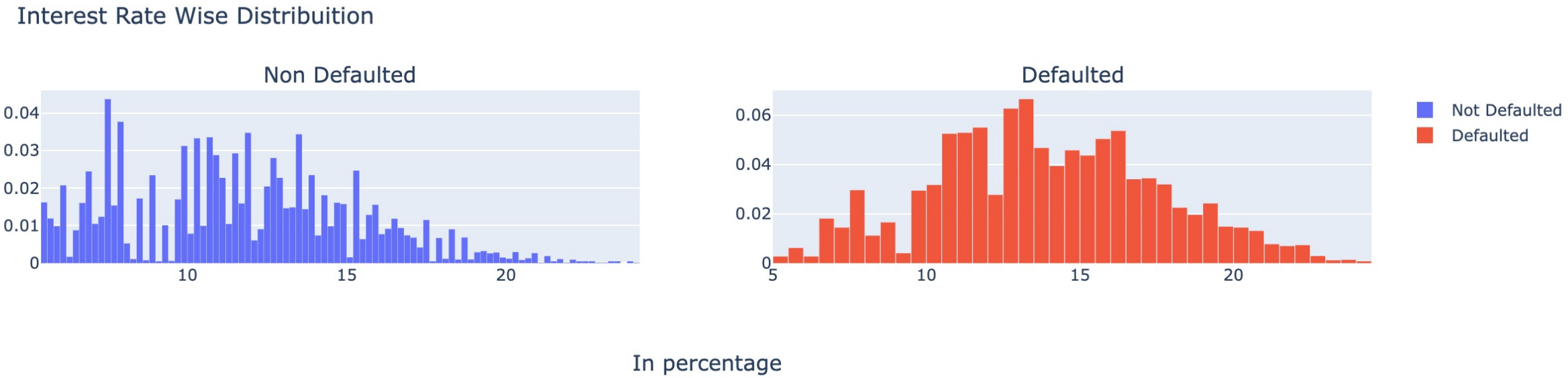


From the above graph, its evident that around 16 percent of the loan is charged off when compared with the fully paid Loans. Rise in this percentage over 20, would lead to drastic loss as per Industry standards

Analysis on Funded Amount over Charged off Loans

We Observe That most loan charged off are in the range of 6k to 16k. We must stricten the scrutiny of loans where request is in this particular range

Analysis on Interest Rate over Charged off Loans

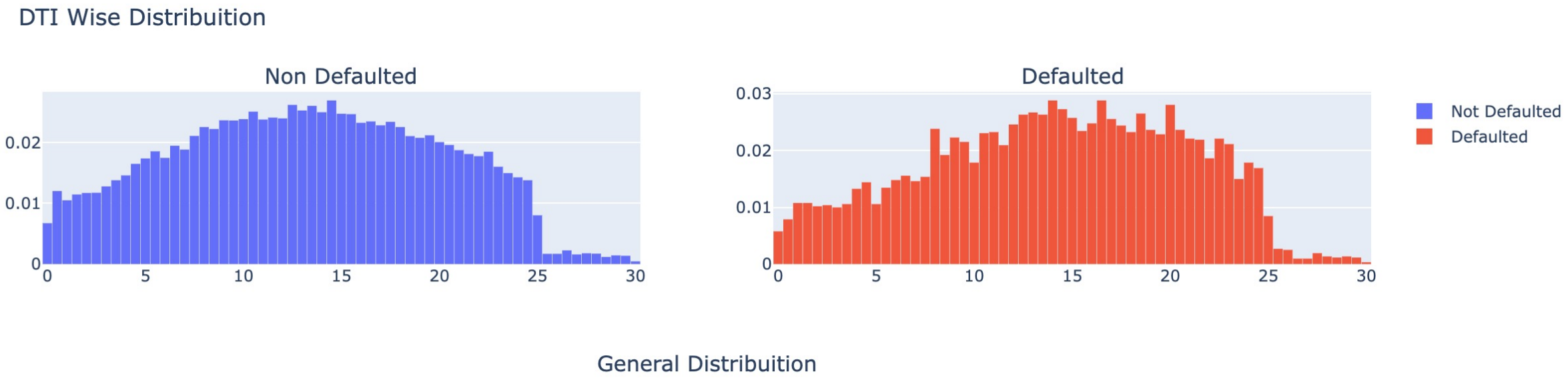


It is clearly evident that for interest below 10%, defaulting is low. It rises steep for interest above 10-16%.
It is advisable to provide short term loan at lower interest rate than higher interest rate loans

Analysis on Income over Charged off Loans

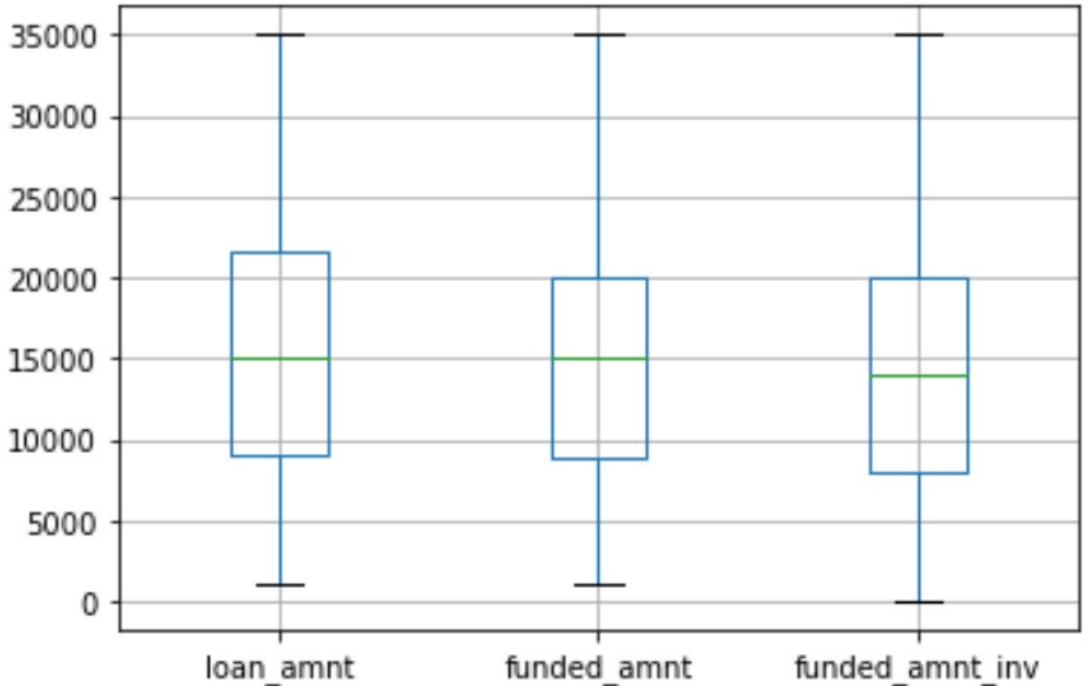
We observe that borrowers with income ranging from 30k – 50k seems to have higher default rates. This needs to be evaluated as it can be due to higher number of loans been taken by this income group. This shall be evaluated in Bivariate Analysis

Analysis on DTI (Debt To Income Ratio) over Charged off Loans



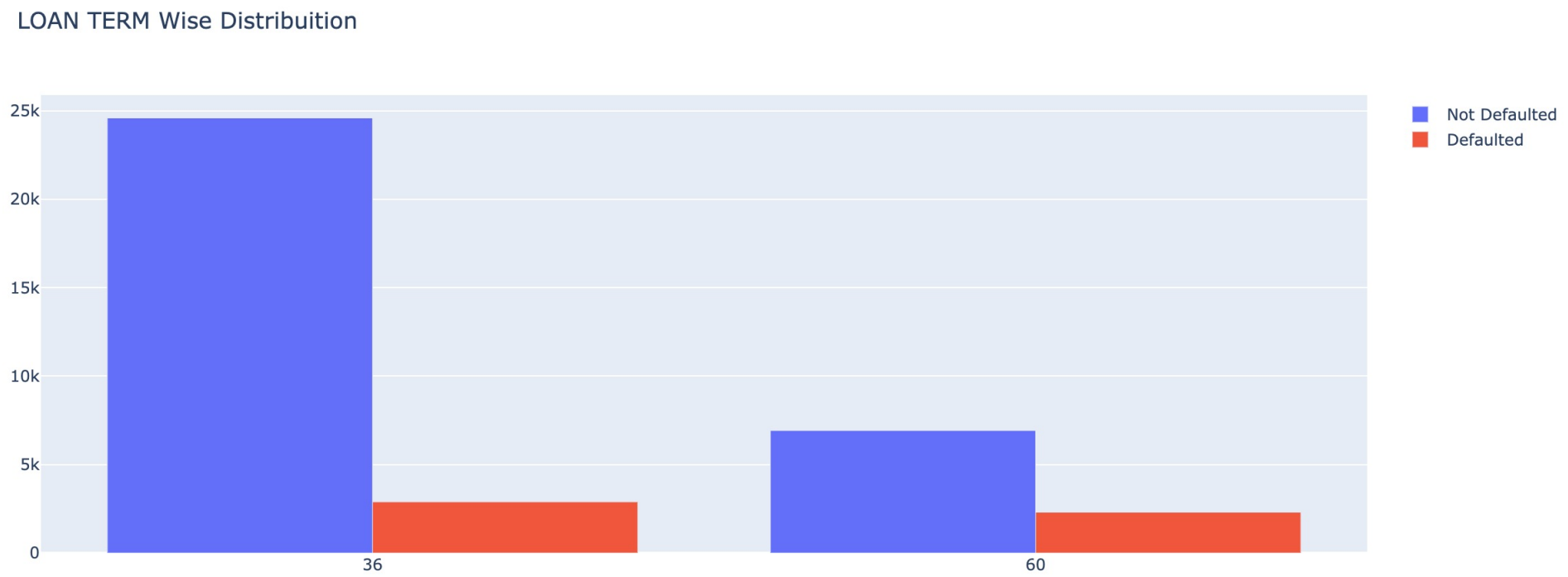
The graph says very clearly that higher the DTI, higher is the chance to default. Loan approval needs to be scrutinized for DTI higher that 10.

Analysis on Eligible and Verified on Loan amount applied and approved



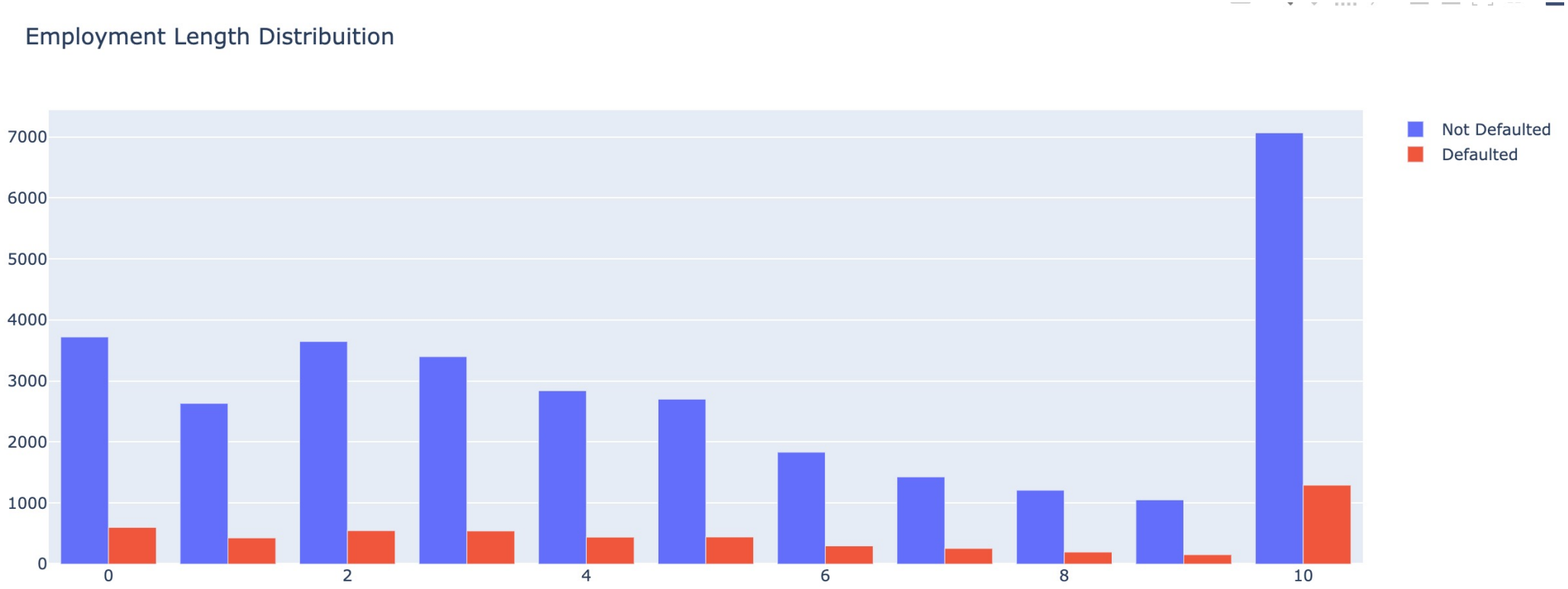
Its observed that among the eligible and verified applicants the 50th percentile is about 12k and mostly getting approved.
Beyond that until 20K a bit lesser will be approved

Analysis on Loan Term over Charged off Loans



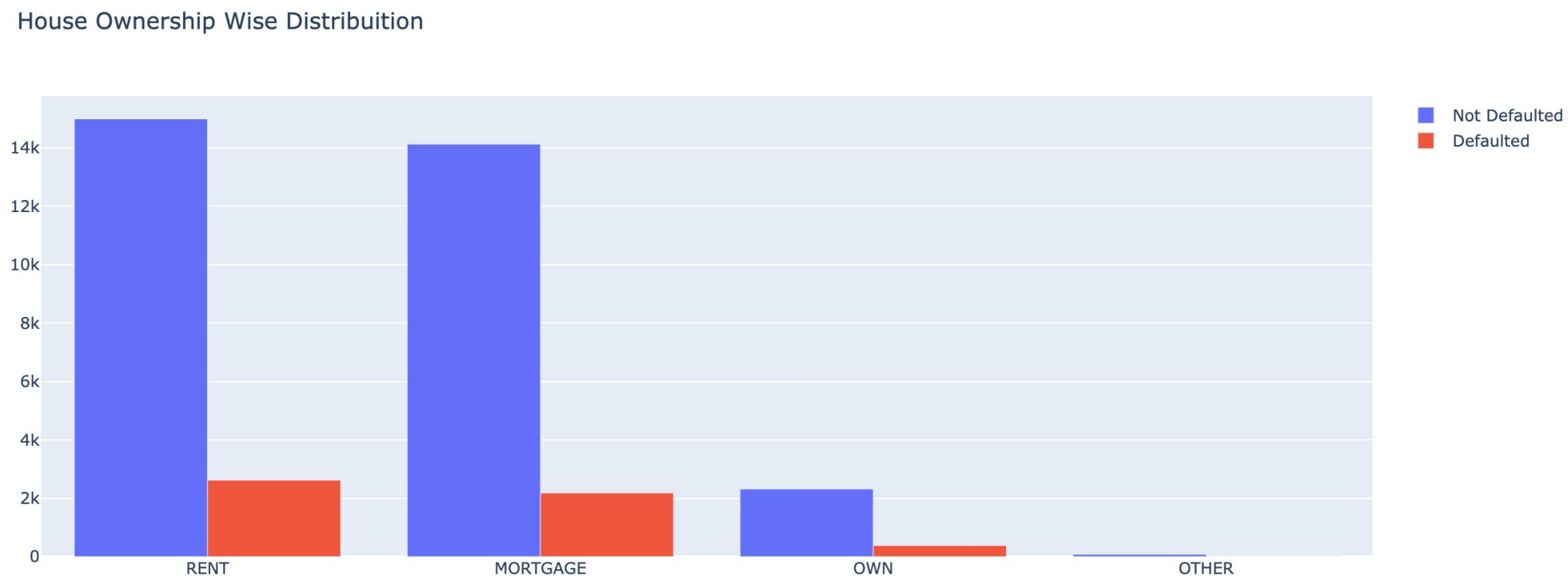
Most of the loans taken were of 36 months duration. But the ratio of Default to full payment is higher for 60 months loan duration (ie, around 33 percent) which is high.
Loan with higher duration needs to be scrutinized to see the long-term capability of the borrower

Analysis on Employment Length over Charged off Loans



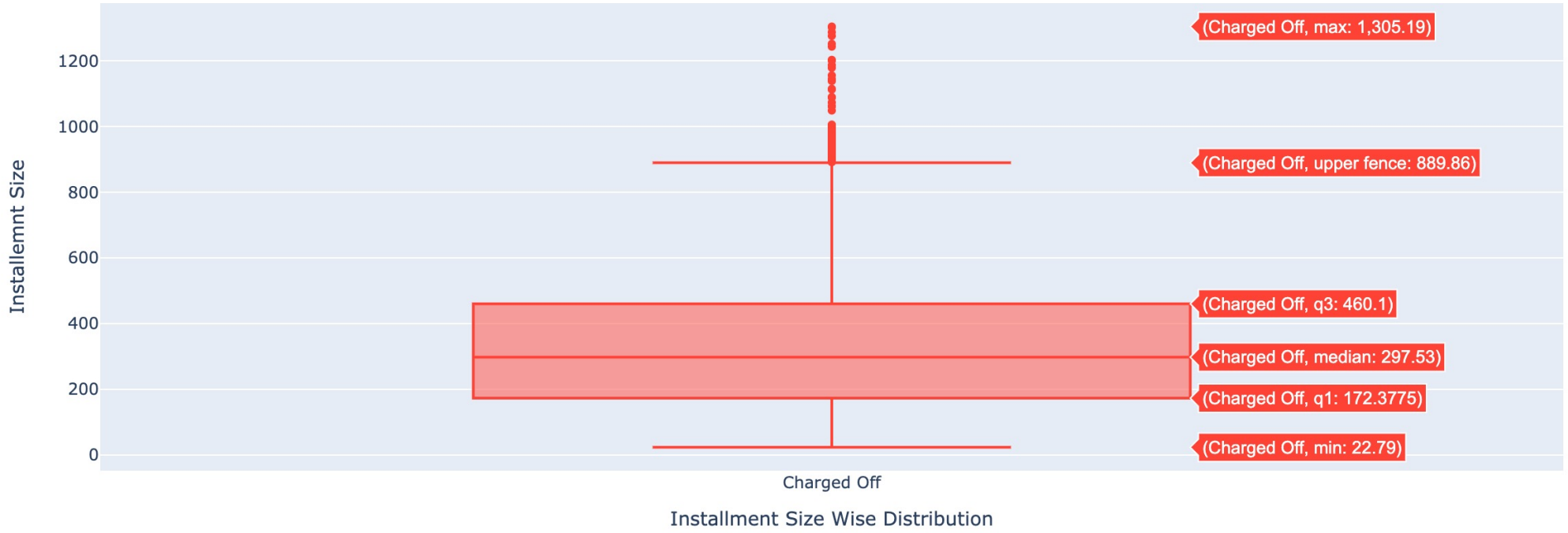
We observe that most loans are taken by borrowers with 10+ years of employment, but, there is no strong ratio which describes them to be defaulted. Hence, we don't have any significant conclusions on the same

Analysis on House Ownership over Charged off Loans



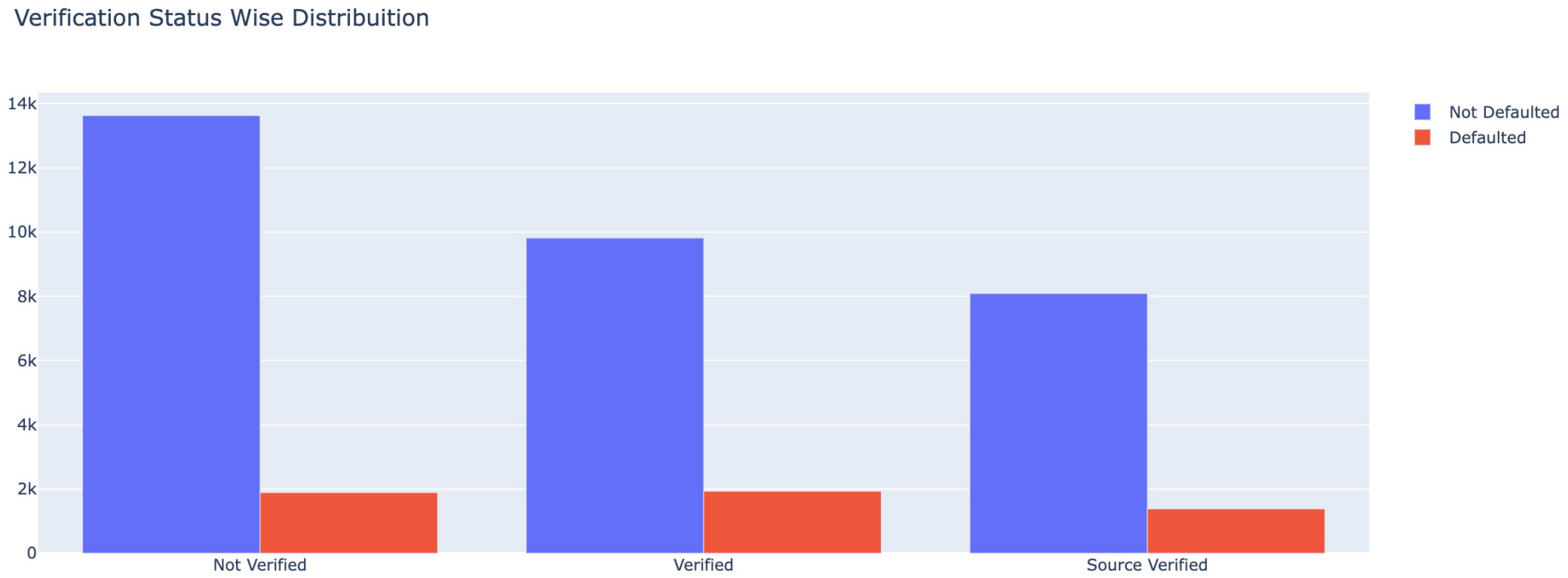
Borrowers living in Rented or a Mortgaged house, tends to default 15 percent more than those who have own home. Its not advisable to give loans to Borrowers living in Rented or a Mortgaged house without scrutiny of their purchasing power.

Analysis on Installment Size over Charged off Loans



Borrowers with installment size of 200 to 400 tends to default more. But this can be attributed to higher quantity of loans been taken in this range. But we also see a trend where higher the Installment size, higher is the chance to default

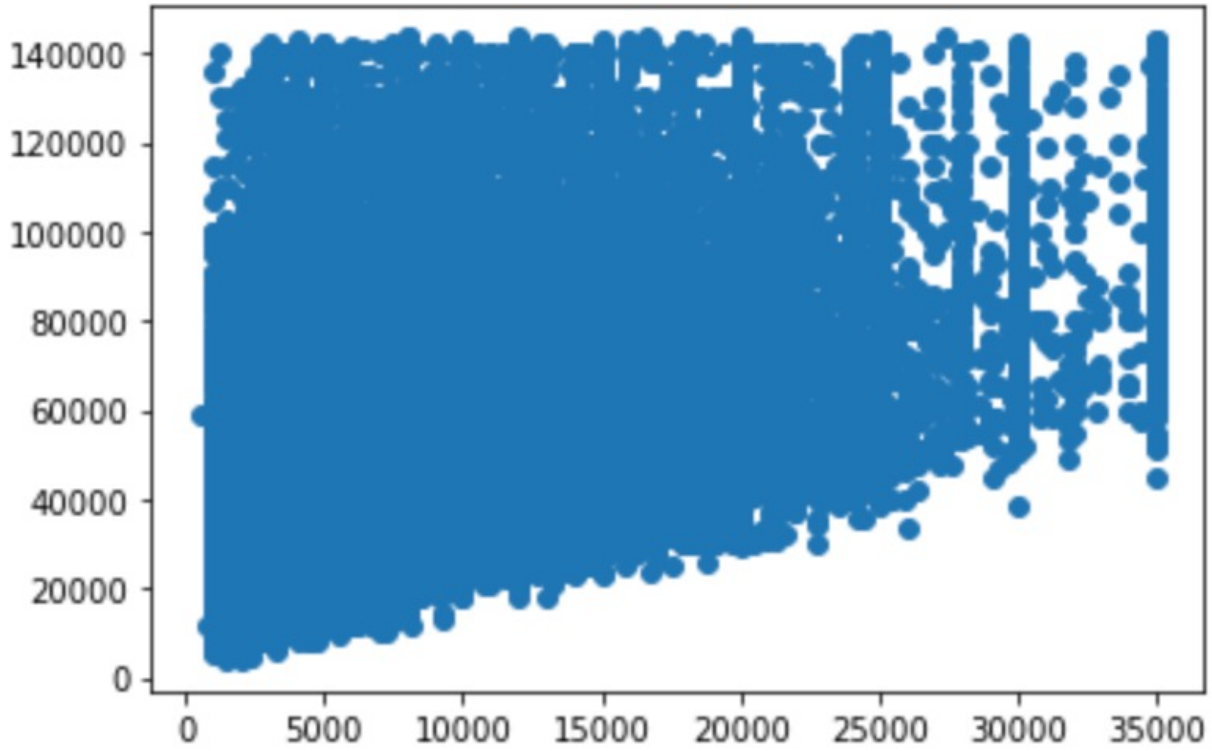
Analysis on Verification Status over Charged off Loans



Surprisingly, Verified loans are defaulting slightly higher than Not verified loans. This needs to be validated with further analysis.

BIVARIATE ANALYSIS

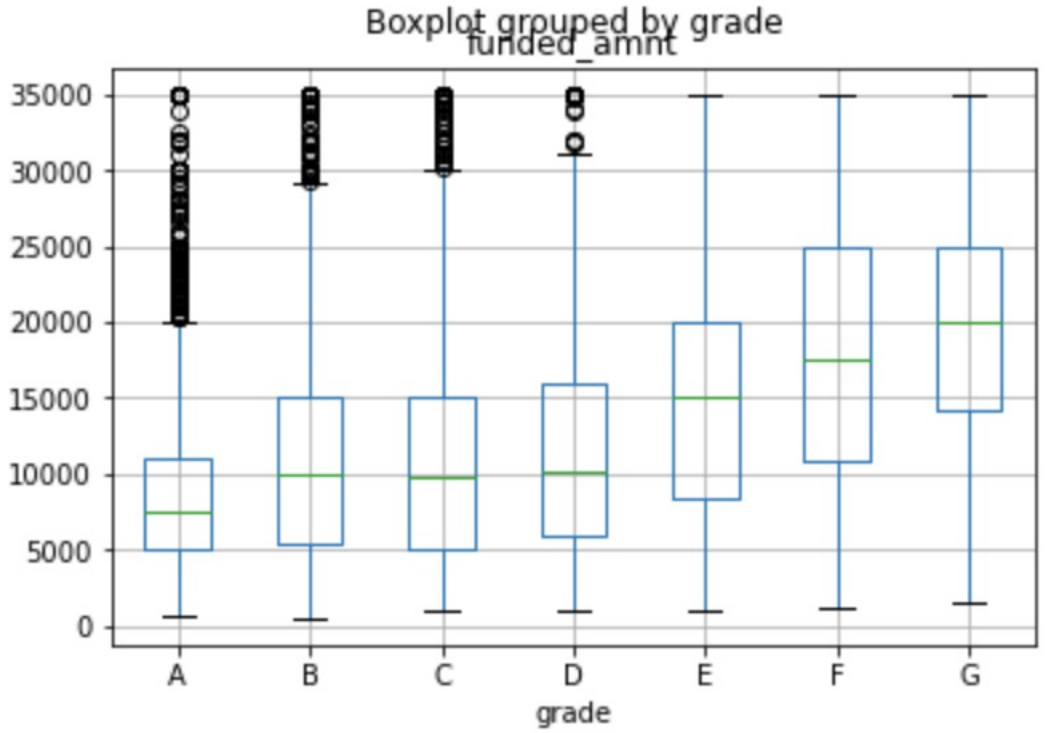
Analysis on Funded Loan over Annual Income



X Axis : Funded Loan
Y Axis : Annual Income

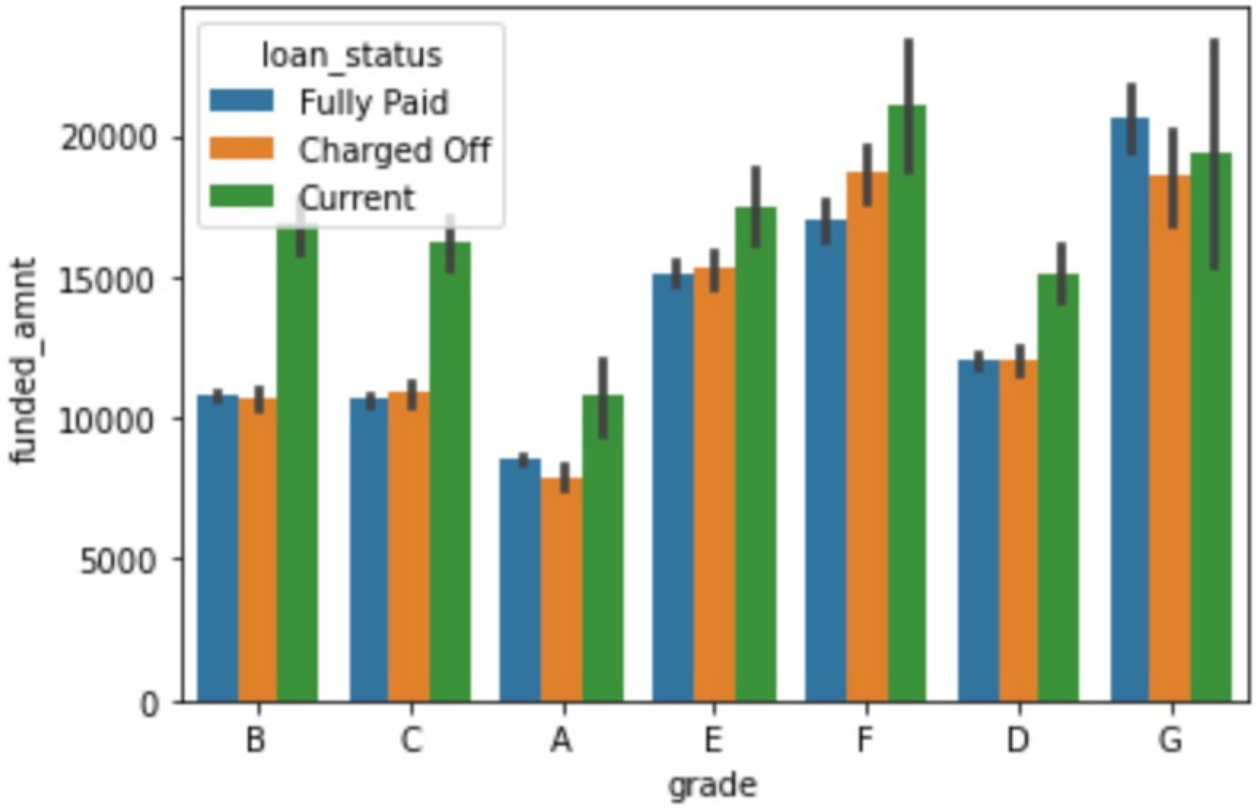
We Observe that lower income group people tends to take more loans as the scatter plot is very dense here. With further analysis, we have observed that loans were given to people where loan amount is higher than the annual income. Which adds to the risk factor and must be avoided

Analysis on Funded Loan over Grade



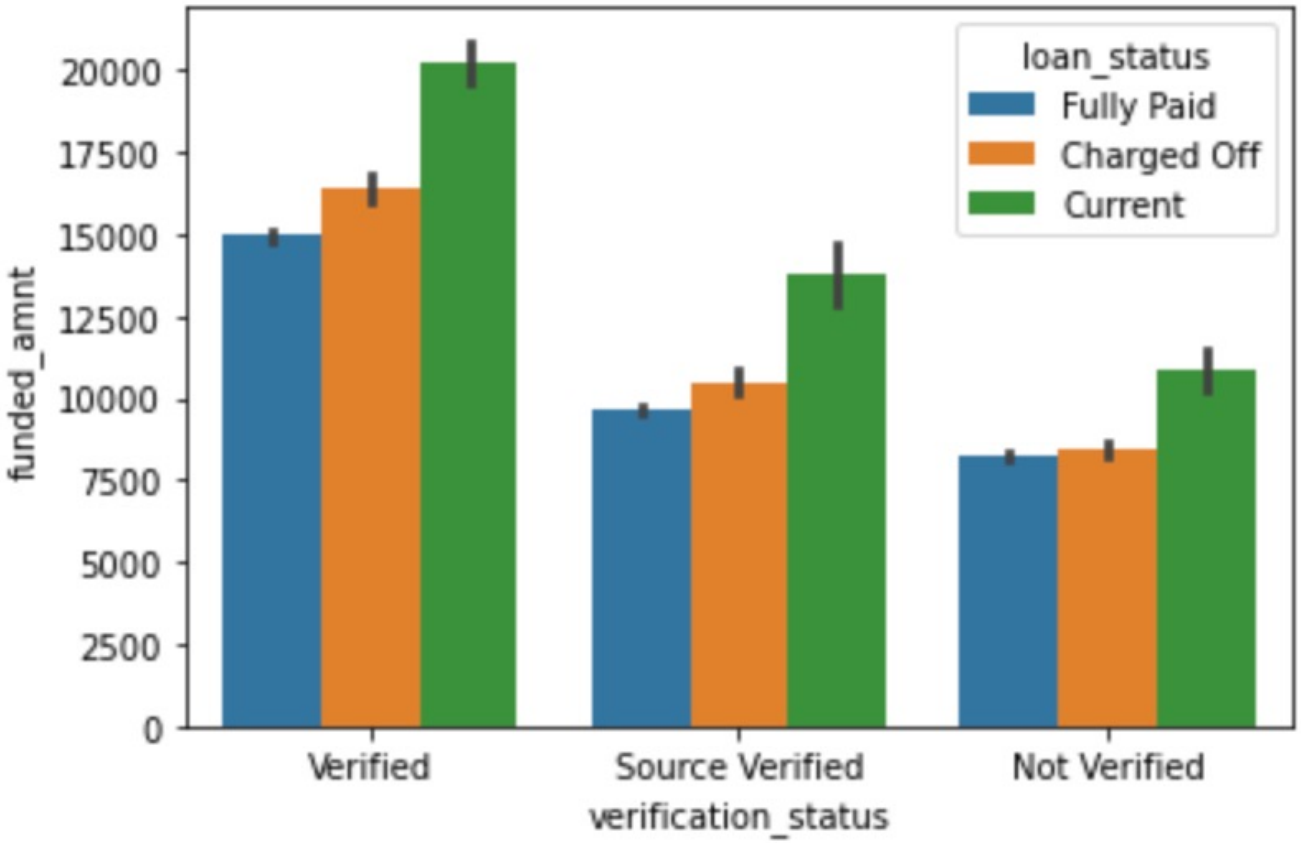
It is observed that most loans are taken by borrowers with lower grade. We would be further analyzing how much do they default in the further analysis.

Analysis on Funded Loan over Grade and Loan Status



As we see in the graph, lower grade borrowers tends to take higher loans, as well as tends to default.
Lending needs to be limited to employees with lower grades

Analysis on Funded Loan over Loan Status and Verification Status



We had observed a reverse trend where verified Borrowers were defaulting more than the Not Verfied borrowers. This graph clearly tells that it's attributed to the higher quantity of loans being verified ones. No relevant conclusions are derived hence.