Lending Club Case Study

Shashi Khurana

Using EDA to find out how consumer attributes and loan attributes influence the tendency of default.

Problem Statement

Money Lending business makes money by lending loans to its customer. Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

This case study aims at identifying such applicants, discover various insights using the data provided, unfold key scenarios so that lending company can identify such risky loan applicants and reduce their credit loss.

Using EDA, We will discover some **driving factors (or driver variables)** behind loan default; the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Solution Approach – EDA

Check the data carefully, analyze each column to find what data can contribute for this analysis.

Clean/Curate the data – Many column contain null values or single values only, they cannot contribute for analysis, eliminate such columns.

For other cases impute the value correctly. For example, check missing values and see how can that missing value be replaced, mode or median. For example, replace null for emp_length with the mode for emp_length. Eliminate the records where Income details is outlier.

Correct the data in various cases for example Employment Length (10+ as 10, <1 as 0 etc.) removing % from interest, correcting revol_util and addressing other missing subset of records like annual income outliers

Do the univariate analysis to find how the interest rate is distributed, find the loan status for the dataset provided, home distribution for total population, loan amount for the total population. Find other variables like Grade, sub grade, Terms month, revol_util, dti, address (state), loan purpose.

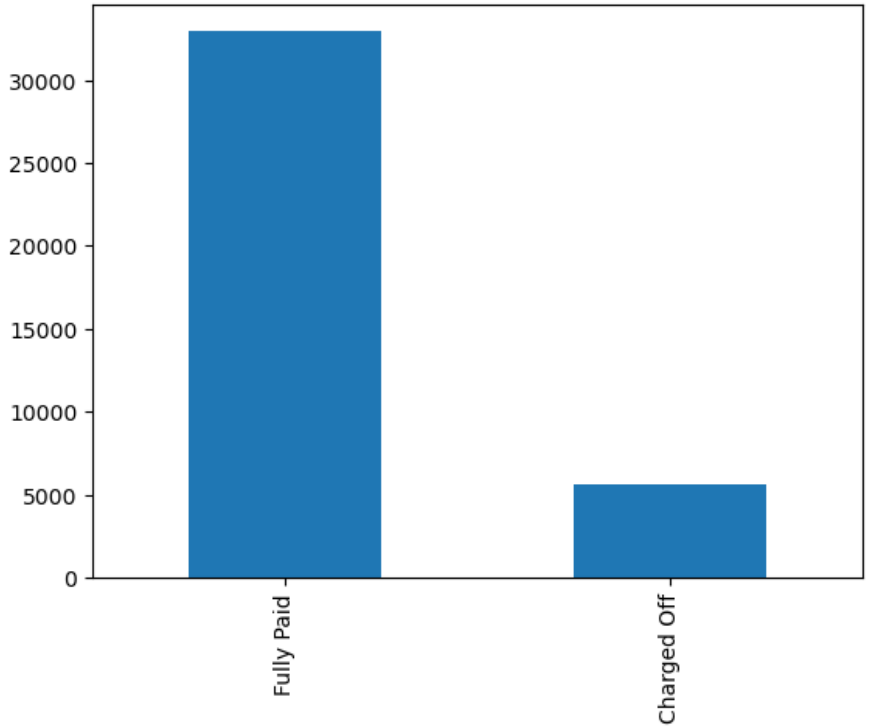
Ordered categorical variables often produce valuable patterns that help us derive information. It is important to understand data and sometimes derived metrics make more sense instead of data itself. Keep the inter quartile value for annual income will give the right results.

Bivariate analysis :Risk analysis by comparing defaulters' data or its percentage based on their home ownership, annual Income, Loan purpose, term, Verification status, interest rate, debt to income ratio.

Data Understanding

Loan Status – Fully Paid Vs Charged Off

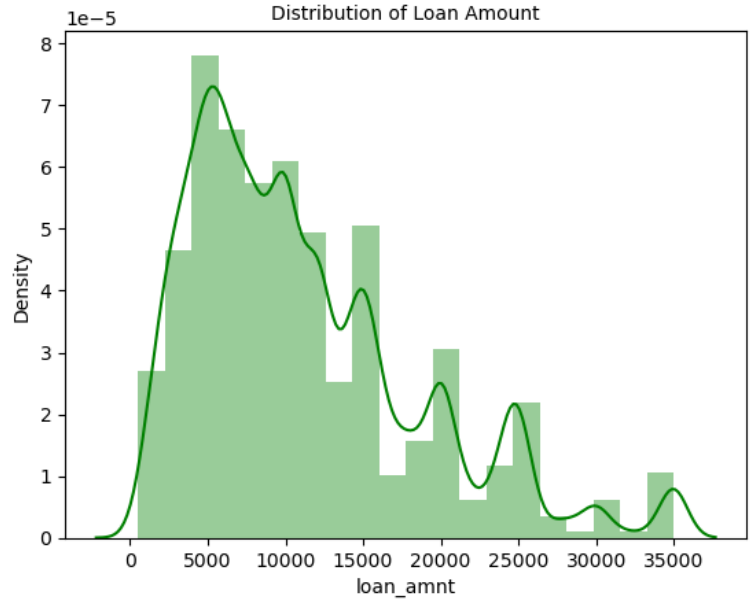
- Fully paid population is way more than Charged off Population.
- Loan status for current is not considered in the dataset



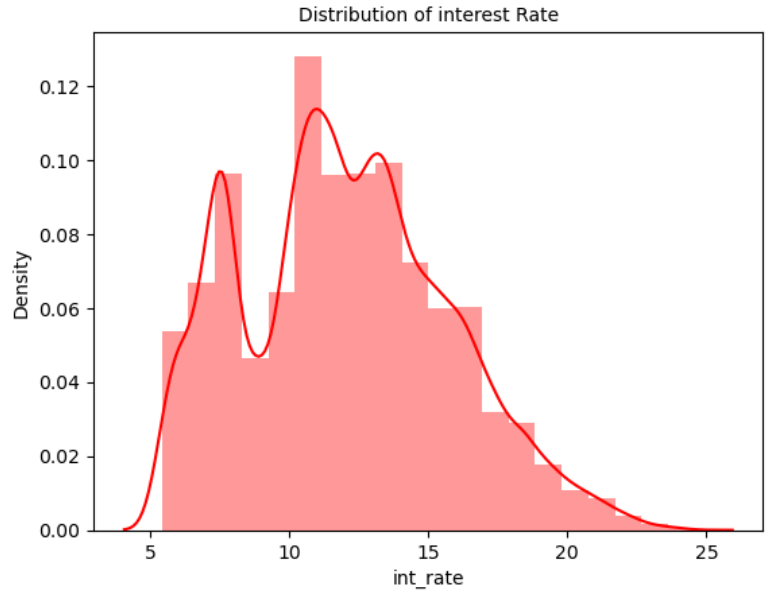
Data understanding

Distribution of Loan Amount & Interest Rate

- Loan amount that has been disbursed highest is in range of 5000-15000



- Interest Rate starts at approx. 6% & vary little over 21 %

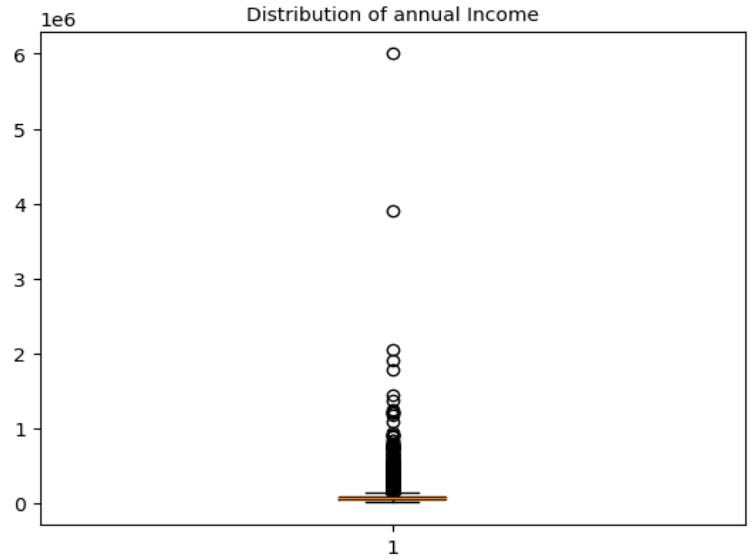


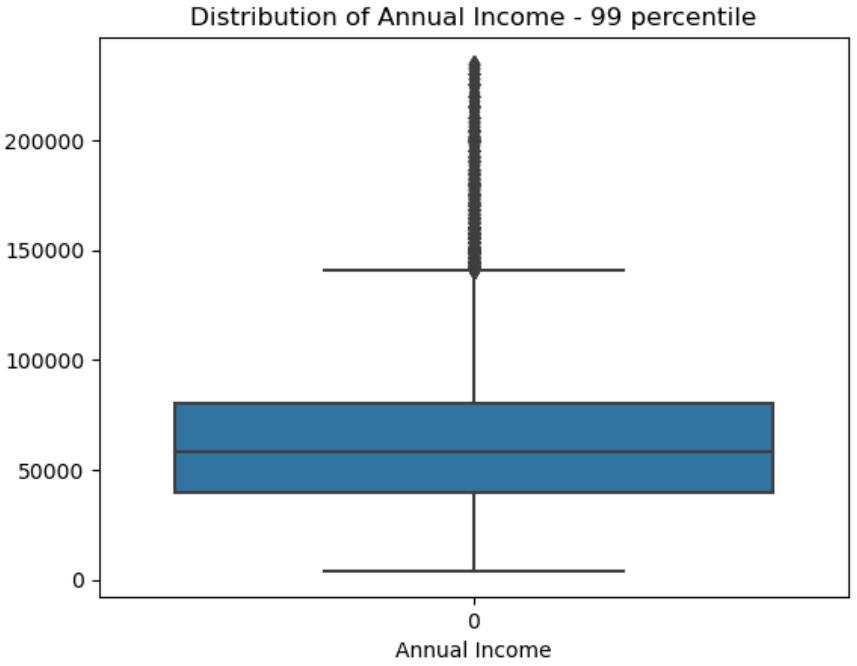
Data Understanding

Annual Income

- Annual income overall – Boxplot chart predicts outliers for annual income.

- After adjusting data for annual income. Dropping outliers and keeping the data to 99 percentile Median income lies around 65K



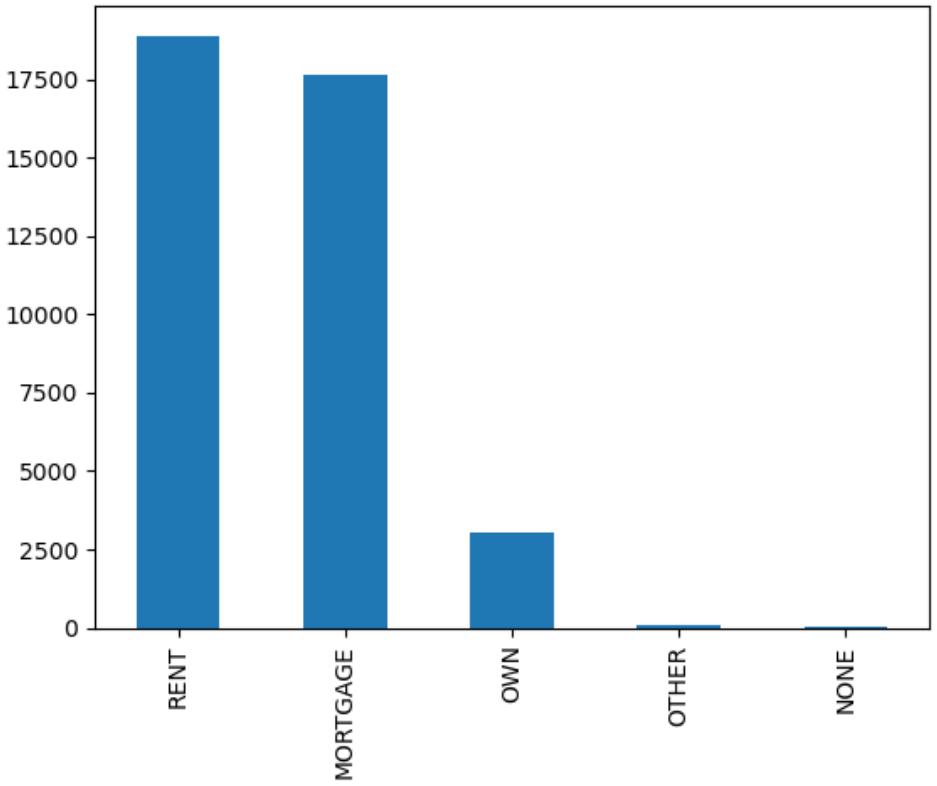


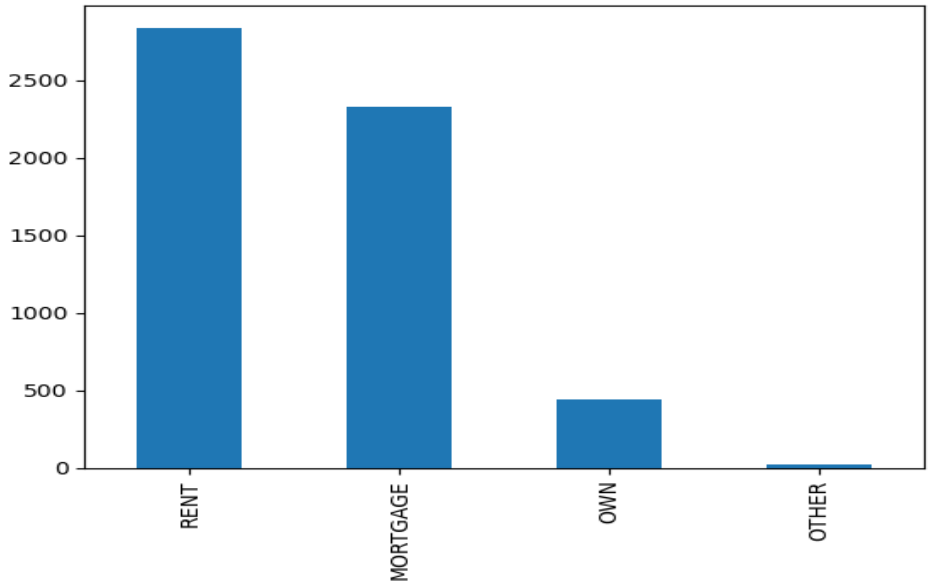
Data Understanding

Home Ownership

- Home Ownership overall has higher in **Rental Category.**

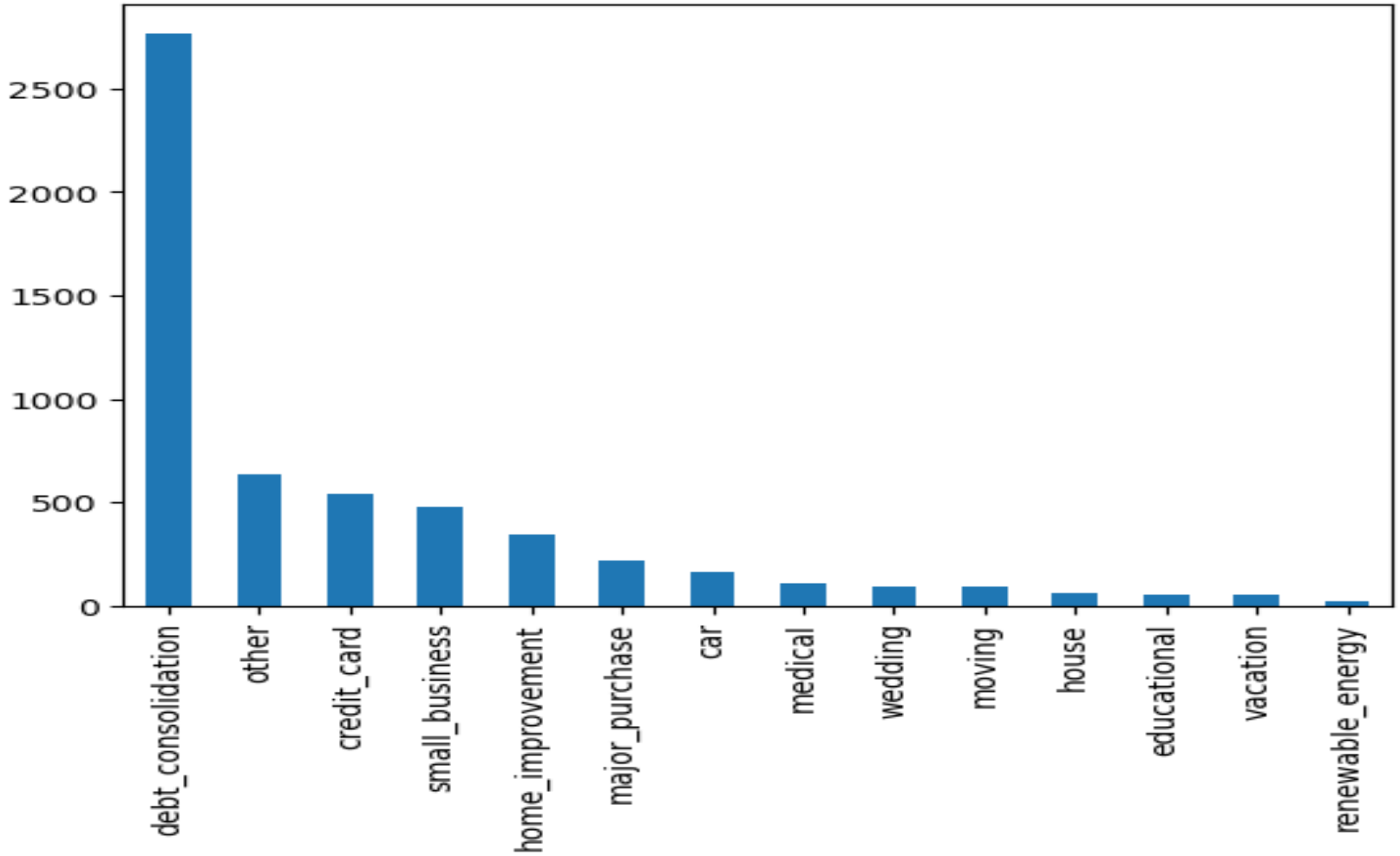
- For Defaulter population trend is also similar Rent followed by Mortgage and then own.
- Chart below refers for smaller population that defaults.





Univariate Analysis –
Business Purpose

Business purpose - debt consolidation tops as the major reason for loan default among population who defaulted.



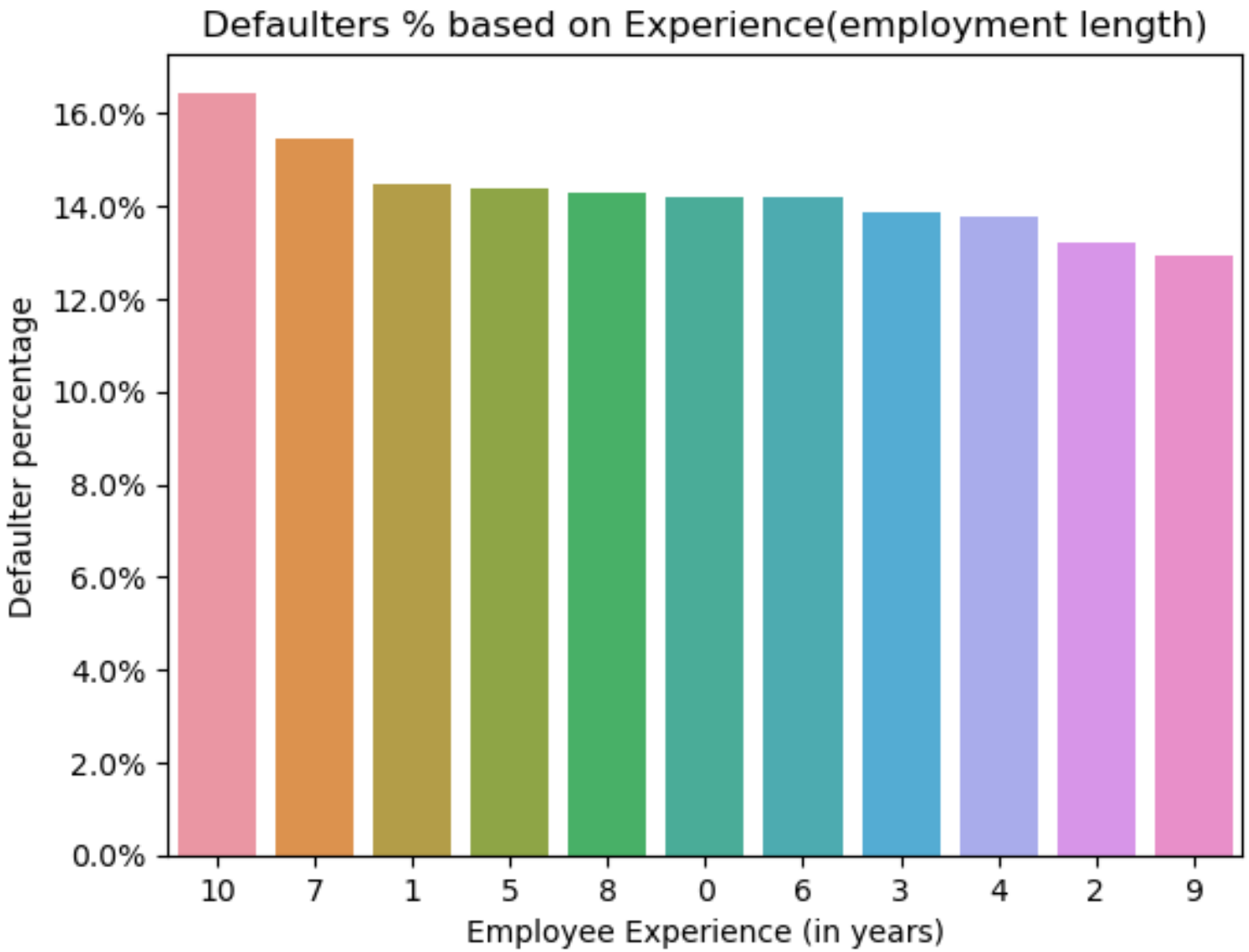
Univariate Analysis

- Employment Experience

Default percentage w.r.t Employee Experience

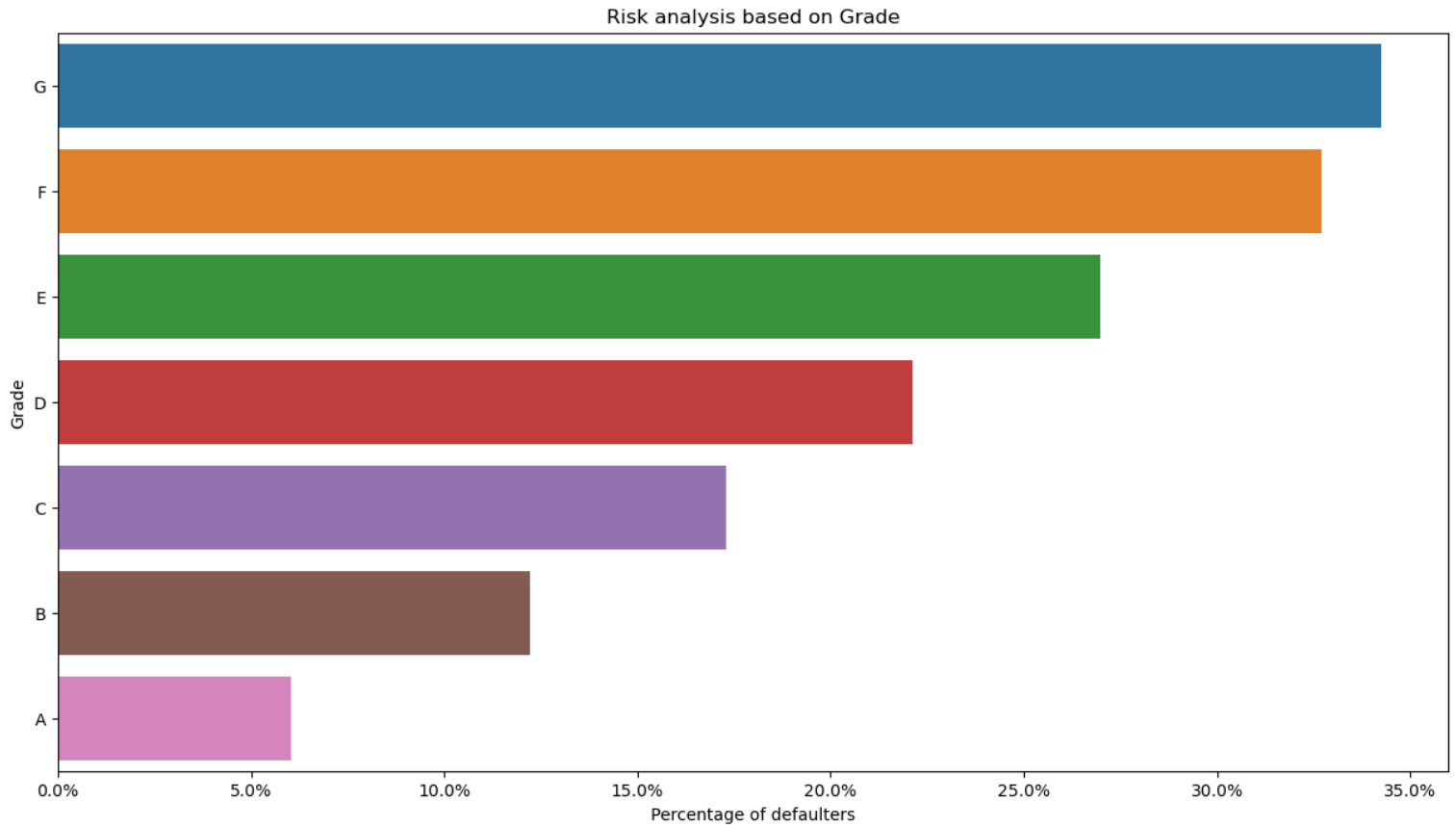
Candidates with 10 years experience band has highest defaulter percentage.

Defaulter percentage is calculated as percent of defaulters in specific experience band w.r.t total population in that experience band.



Univariate Analysis

People tend to less default for Grade A vs Grade G

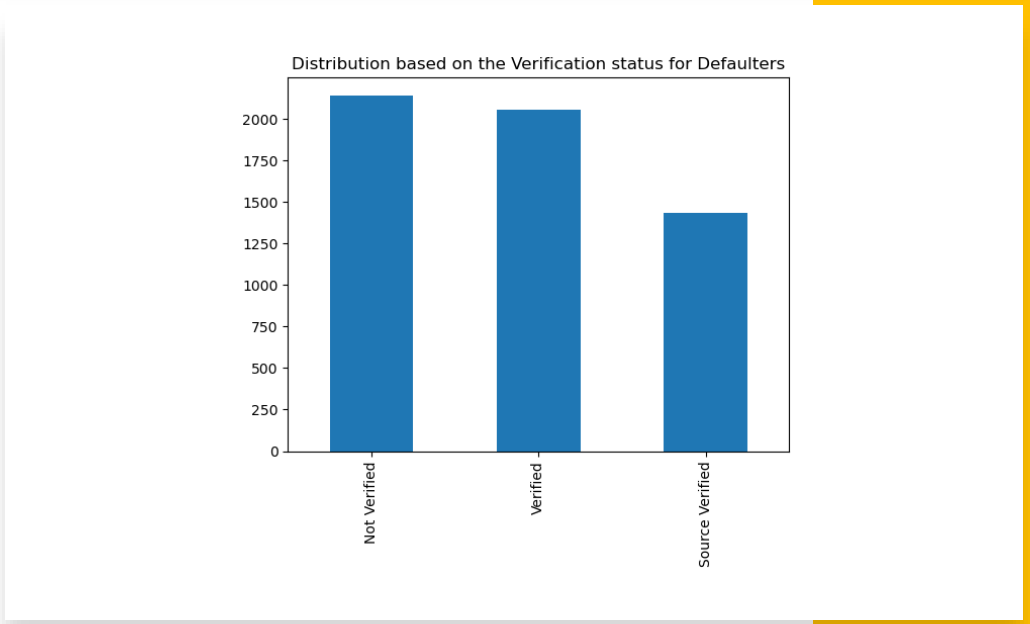


Univariate Analysis

Verification Status

Not Verified is more likely to default.

If verification status is source verified, they are less prone to default.



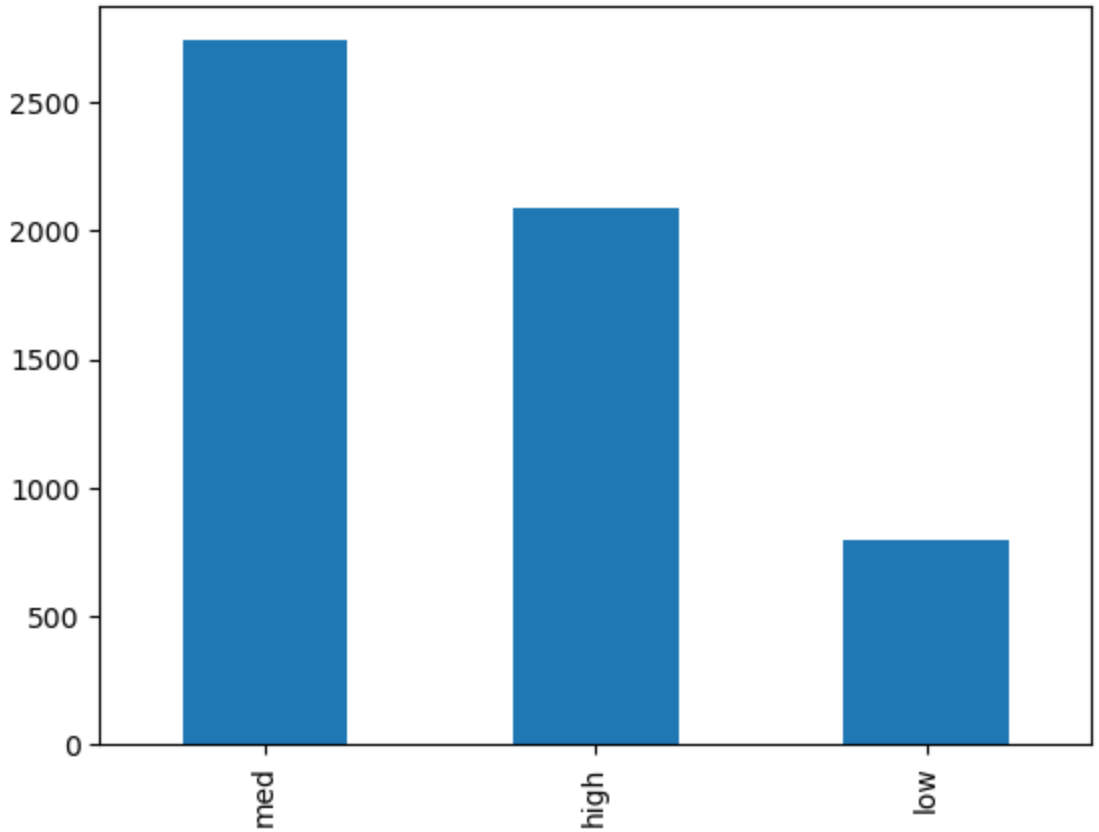
Derived Metrics

Interest Rate Bucket

Interest over 10% is more likely to default compared to low interest rates.

Derived variable is categorized as low for interest rate <10% , medium between 10 % - 15% high for over 15 % and

```
inp3['int_rate_buc'] = inp3['interestrate'].astype(float).apply(lambda x:  'low' if 0 < x < 10.0 else 'med' if 10.0 <= x < 15.0 else 'high')
```



Derived Metrics Analysis

Annual Income Group

Middle Income Group pays high interest compared to other income group for complete population.

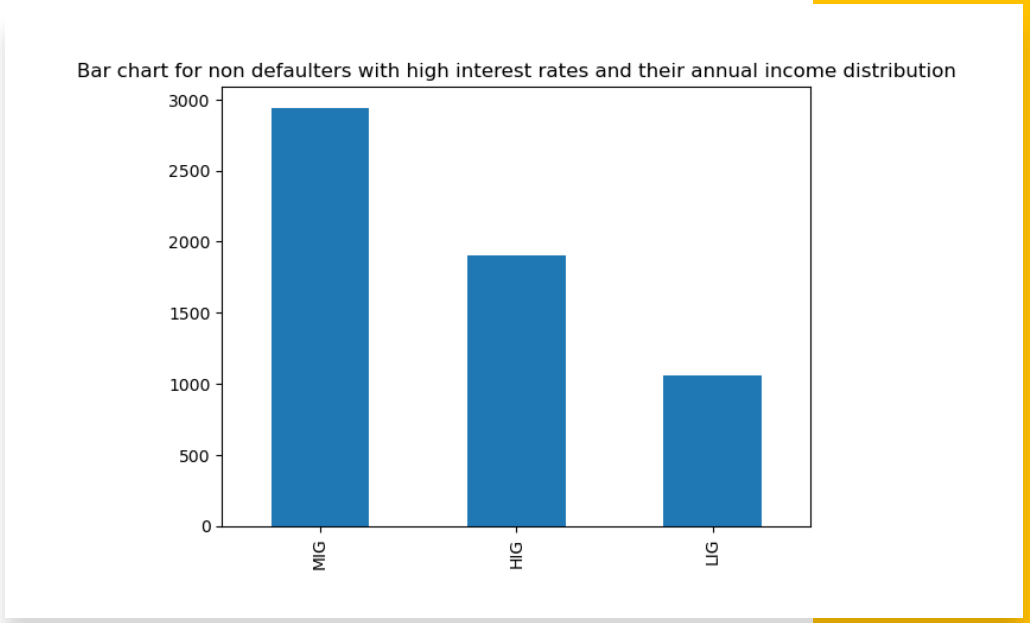
HIG is less likely to default.

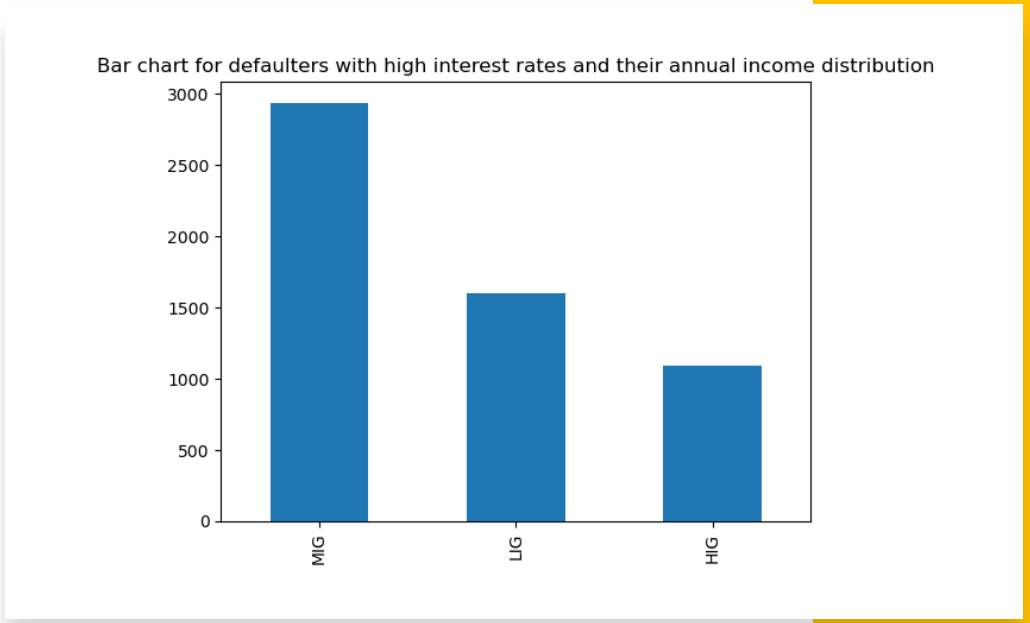
Create 3 Buckets of Income Group (Low, Medium, High)

Annual income group for candidates less than 25 quantile **- LIG**
Between 25 quantile and 75 quantile - **MIG**

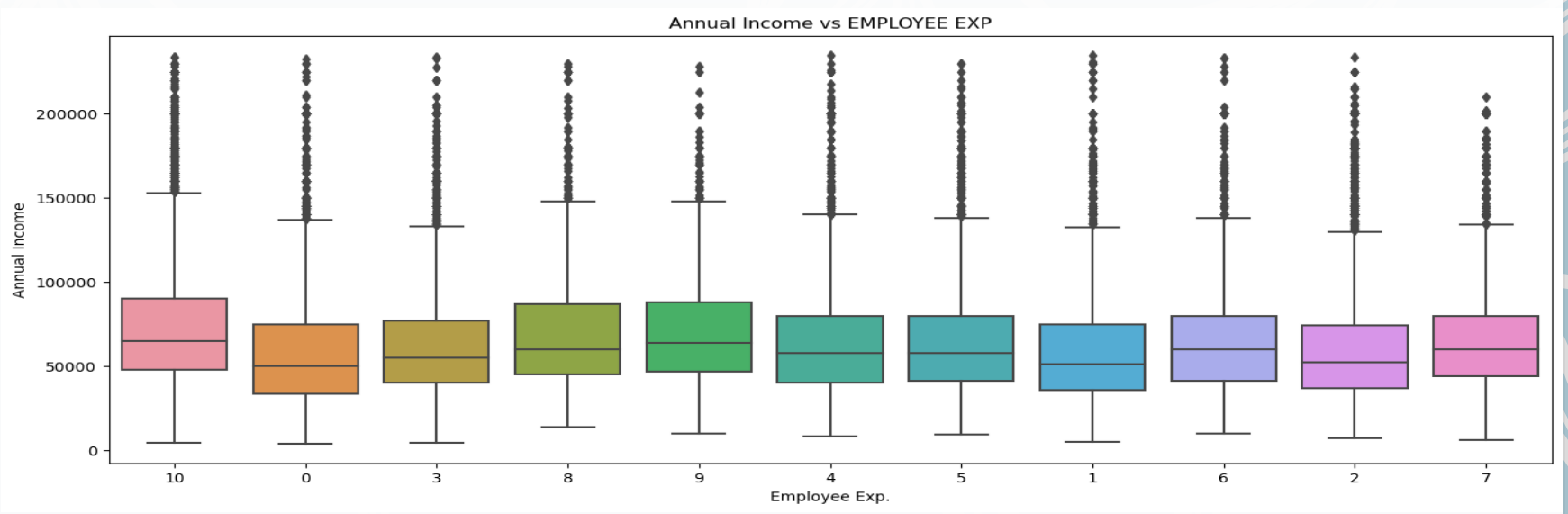
Remaining as - **HIG**

```
inp3['incomegroup'] = inp3['annual_inc'].apply(lambda x:  'LIG' if 0 < x < 40000.0 else 'MIG' if 4000.00 <= x < 82000.0 else 'HIG')
```

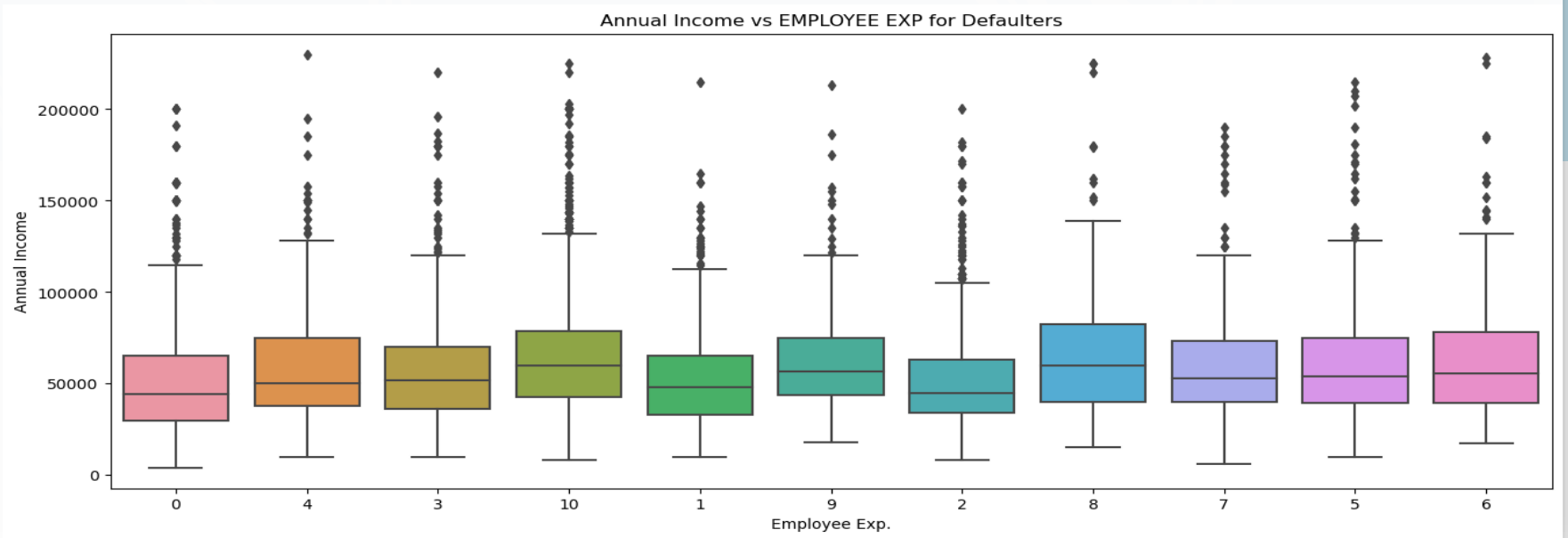




Bivariate Analysis



Annual Income
&
Emp. Experience

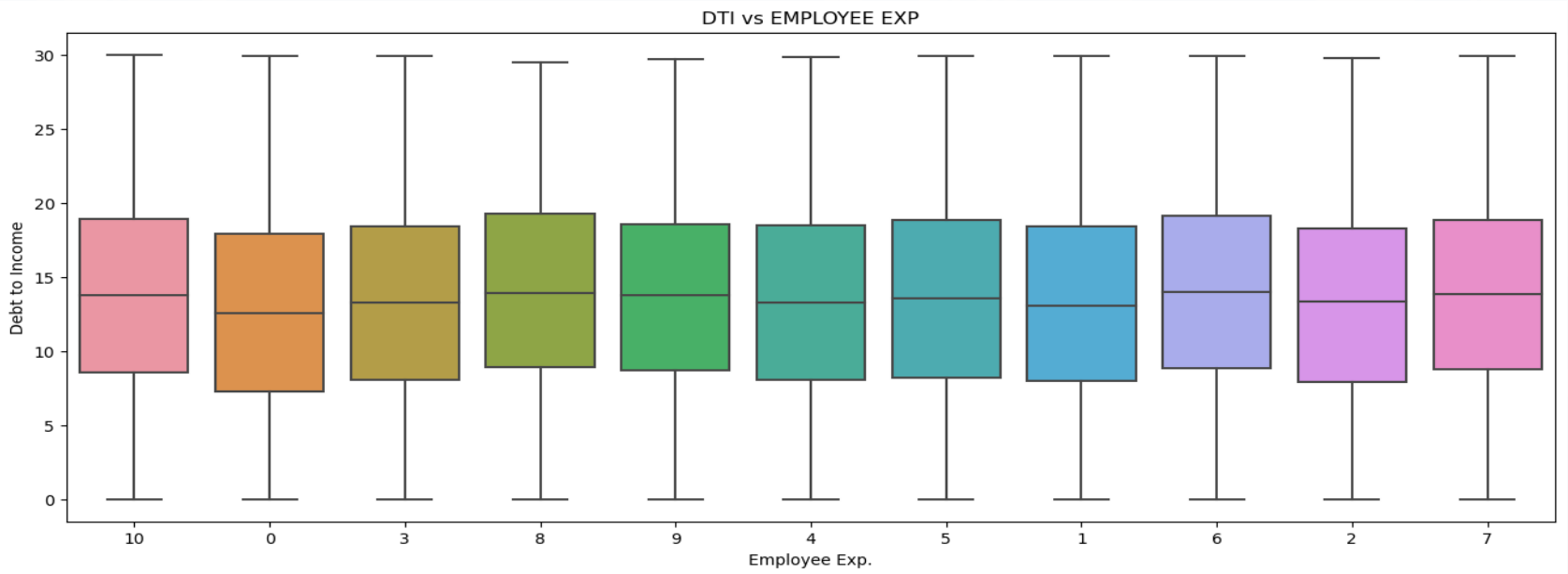


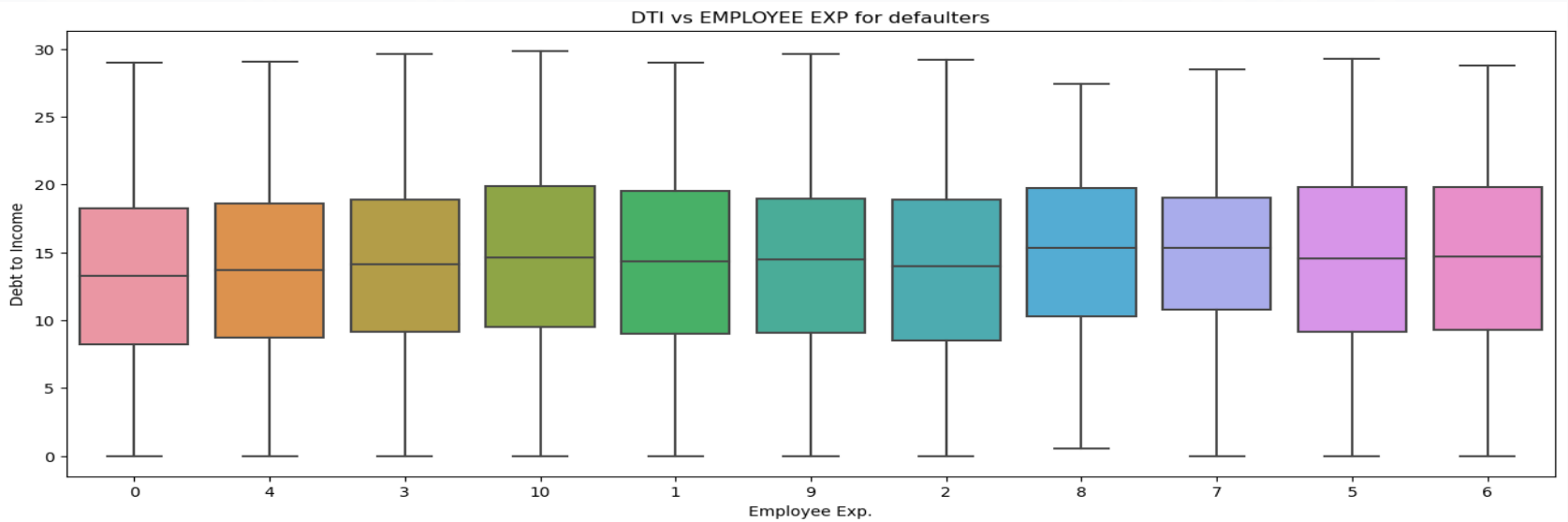
Candidates with 8 years of experience have more annual income compared to 10 years in case of defaulters however overall population 10 years of experience has higher annual income than lower experience.

Bivariate Analysis

Debt to Income
Vs.
Emp. Experience

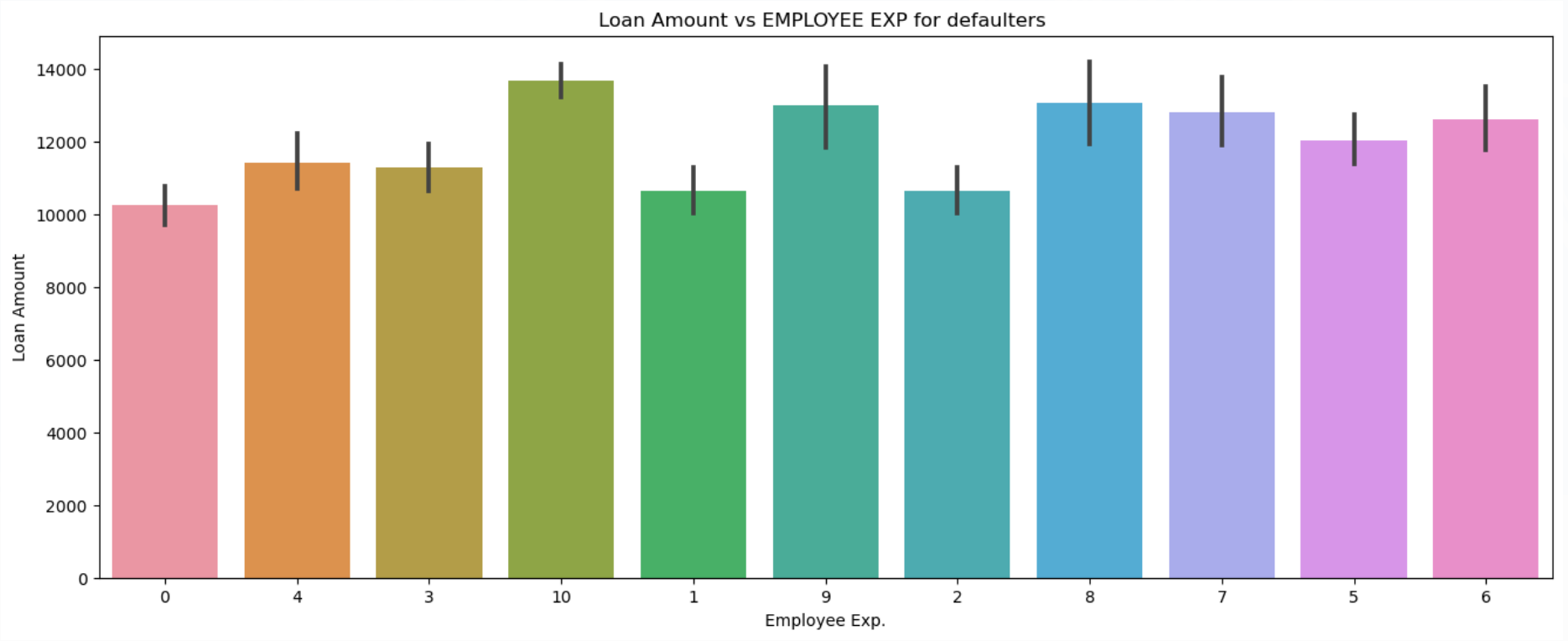
10 years or more experience have more debt-to-Income ratio compared to other experience range. It implies if 10 or above experience has higher dti(means more debt compared to their income) will default comparatively higher.





Bivariate Analysis

Loan amount & Experience

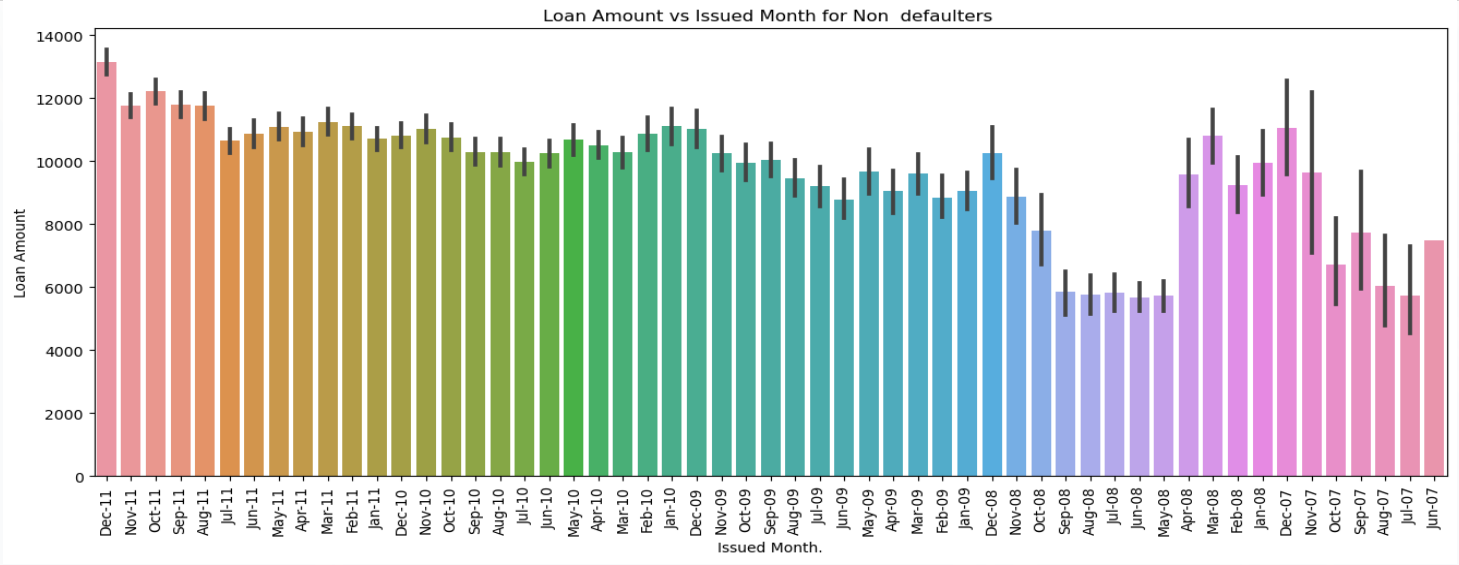


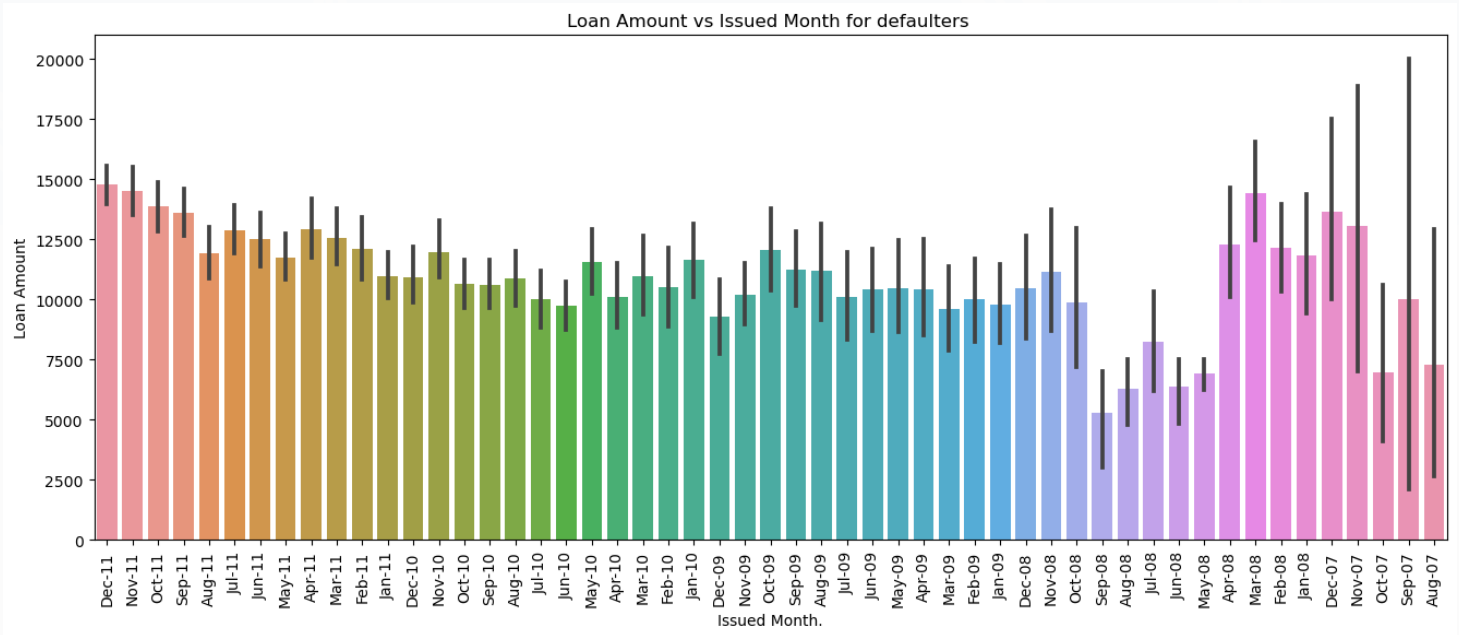
In Defaulters Category 10 years and above experience takes higher loan compared to other experience range. This may act as red flag when combining other factors like, purpose, loan amount, dti.

Business Analysis

Loan amount & Issued Month

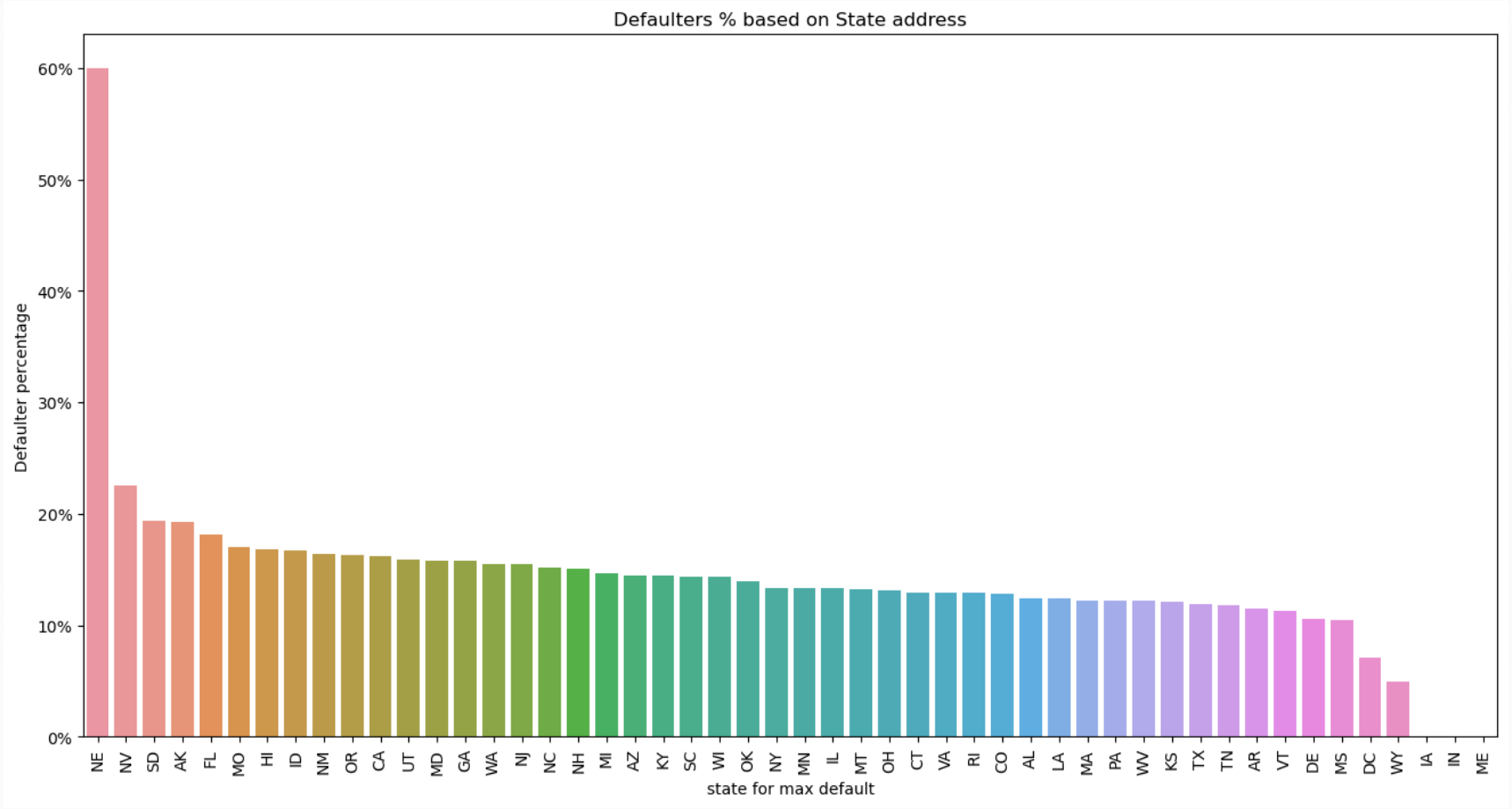
December is the month when Loan amount is highest or issued the most. Defaulter population is also following the same trend. This may lead to another interesting fact that people tend to get loan more in month of dec and verification is not done accurately due to holiday and may contribute to loan default.





Bivariate Analysis

State Address and Defaulter percentage



Defaulter percentage is higher in NE compared to other states.

Defaulter percentage is calculated by people defaulting to the not defaulting population for a particular state.

Conclusion

- Fully paid population is way more than Charged off Population.
- Home Ownership overall has higher in Rental Category.
- Business purpose - debt consolidation tops as the major reason for loan default among population who defaulted.
- If Verification Status is Not Verified; it is more likely to default. compare to verification status of source verified, they are less prone to default.
- Interest rate over 10% is more likely to default compared to low interest rates.
- Middle Income Group(MIG) pays high interest compared to other income group for complete population.
- HIG is less likely to default.
- Candidates with 8 years of experience have more annual income compared to 10 years in case of defaulters however overall population 10 years of experience has higher annual income than lower experience. This gives interesting fact that if experience is more and comparatively their annual income is lower compared to lower experience band, it may be indicator for default.
- 10 years or more experience have more debt-to-Income ratio compared to other experience range. It implies if 10 or above experience has higher dti(means more debt compared to their income) will default compared to other lower experience or lower dti.
- In Defaulters Category 10 years and above experience takes higher loan compared to other experience range. This may act as red flag when combining other factors like, purpose, loan amount, dti.