Lending Club Case Study

Arsh Mishra
Aviral Joshi

Problem Statement

The Consumer finance company carries the risk when choosing whether to accept or deny loan application based on applicants profile

Two Major risks involved while making decisions:

1. If the applicant is likely to repay the loan , the company will lose business if the loan Is denied.
2. Approval of the loan could result in the loss of business if the applicant is unlikely to repay the loan and is likely to default.

The dataset includes information on previous loan applicants with a particular emphasis on those who defaulted , a crucial component in detecting high risk borrowers.

Here our major aim is to identify the factors that indicates whether the user is likely to default , helping to minimize the total loss.

Main goal is to use EDA techniques to understand the default patterns and key risk indicators.

Exploratory Analysis Approach

Define the Objective

- Understand the business problem to be solved and identifying the key variables required for the analysis.

Understand the Data

- Reviewing the data distribution and the type of columns that we have to deal with.

Data Cleaning

- Handling the missing values , rectify the datatype of each column and identifying outliers

Univariate / Bivariate Analysis

- Analyzing the single feature distribution and understanding the correlation between two columns.

Feature Engineering

- Create new features by combining or transforming existing data.

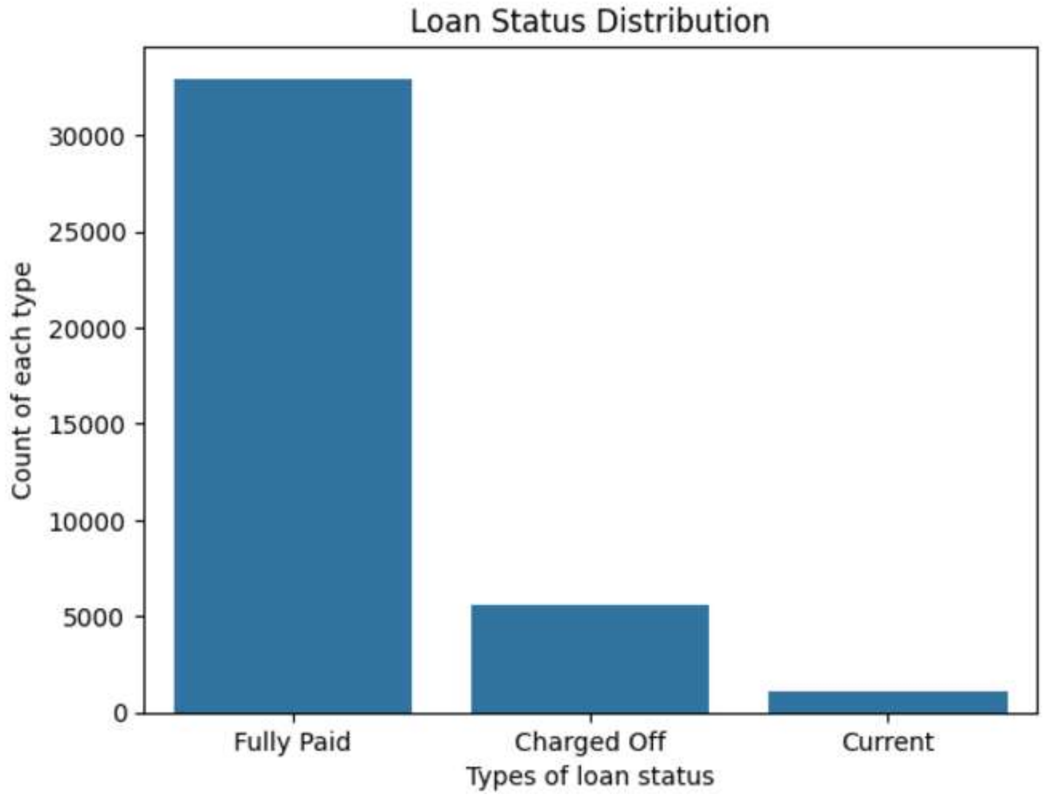
Draw Conclusions

- Summarize key insights from the analysis.

Note: Based on the Charged Off Column, we have added a new binary column called **is_defaulter**, where 1 indicates that the applicant is default and 0 indicates that they are not.

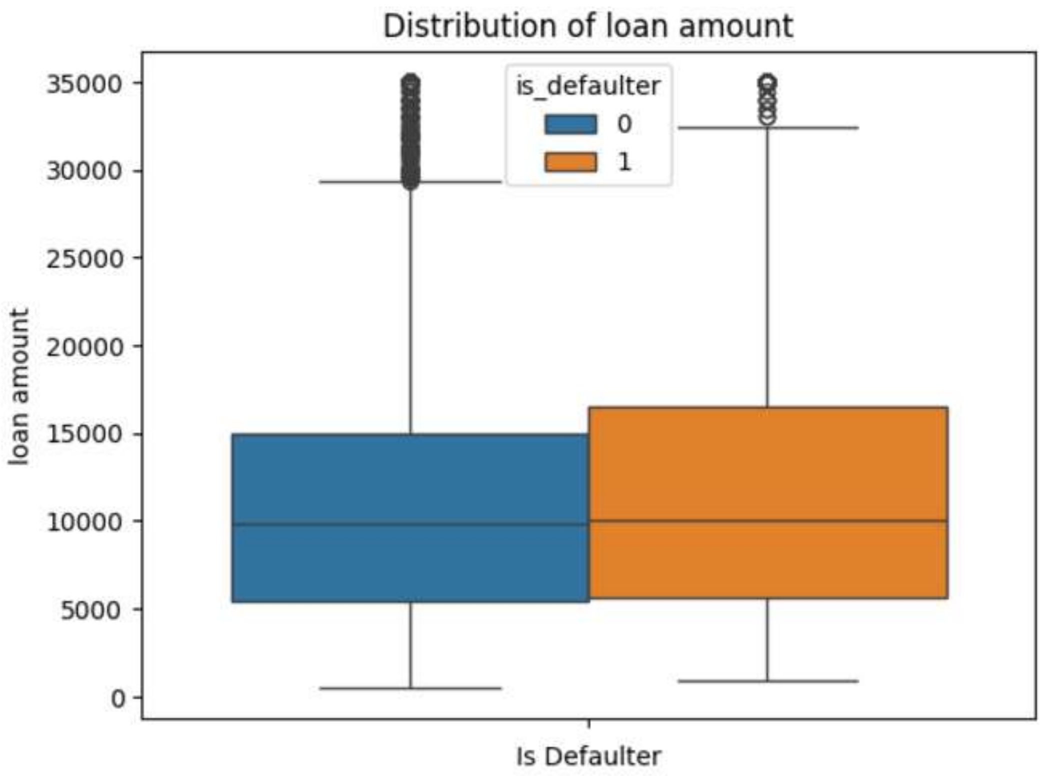
Inference

We have almost 1/8 defaulters out off total data, this shows that defaulters are likely to occur less.

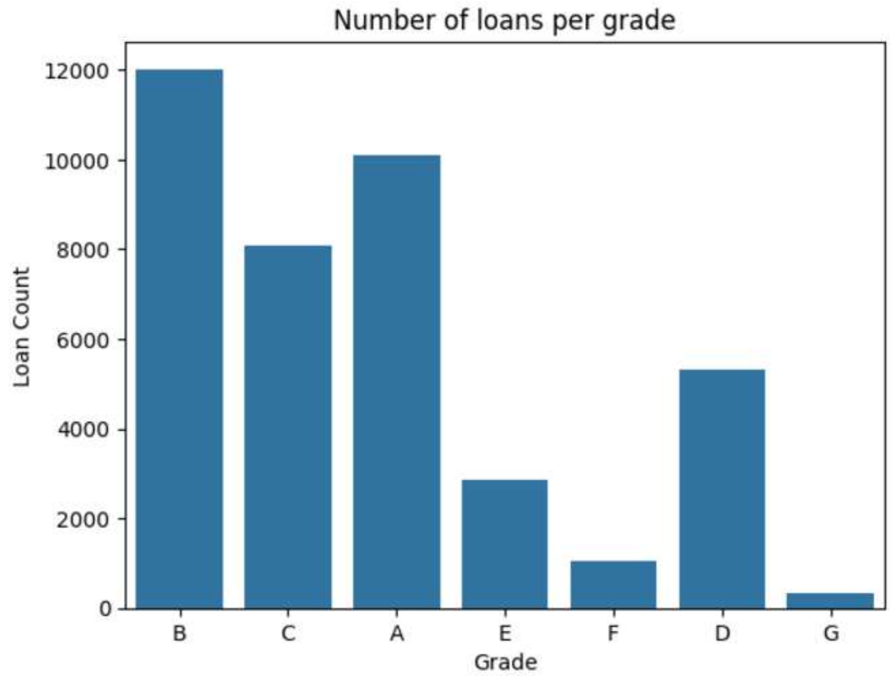


Inference

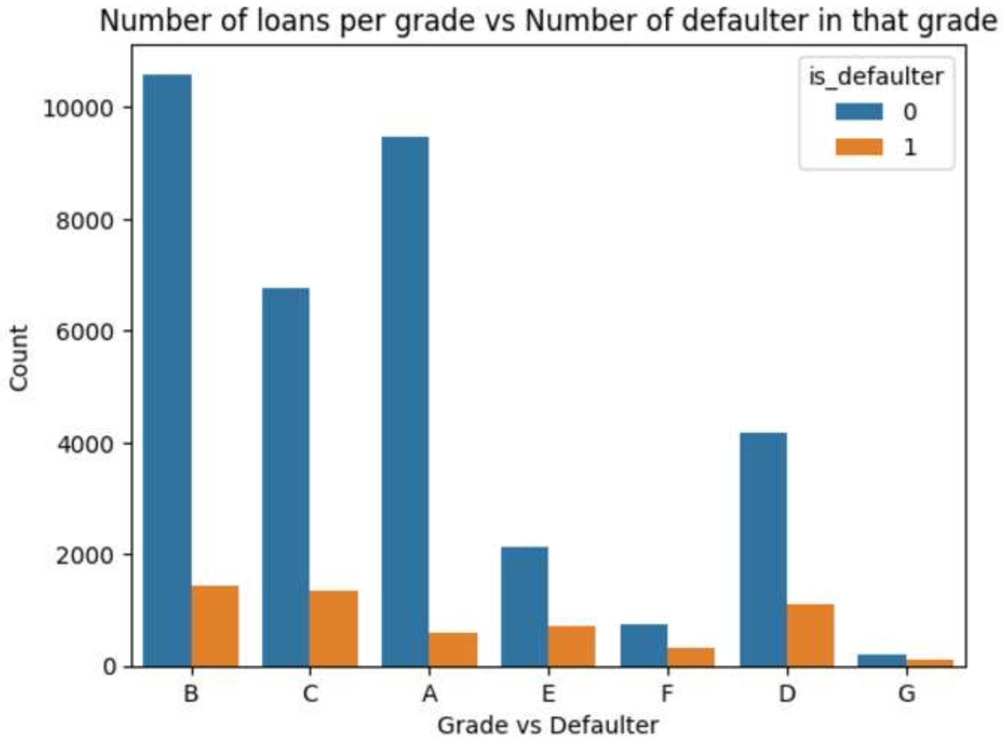
we can see that there are few amount > 29250 that can be called outliers , the borrower taking loan for the amount greater that 29250 is less likely to make it to default.



Inference

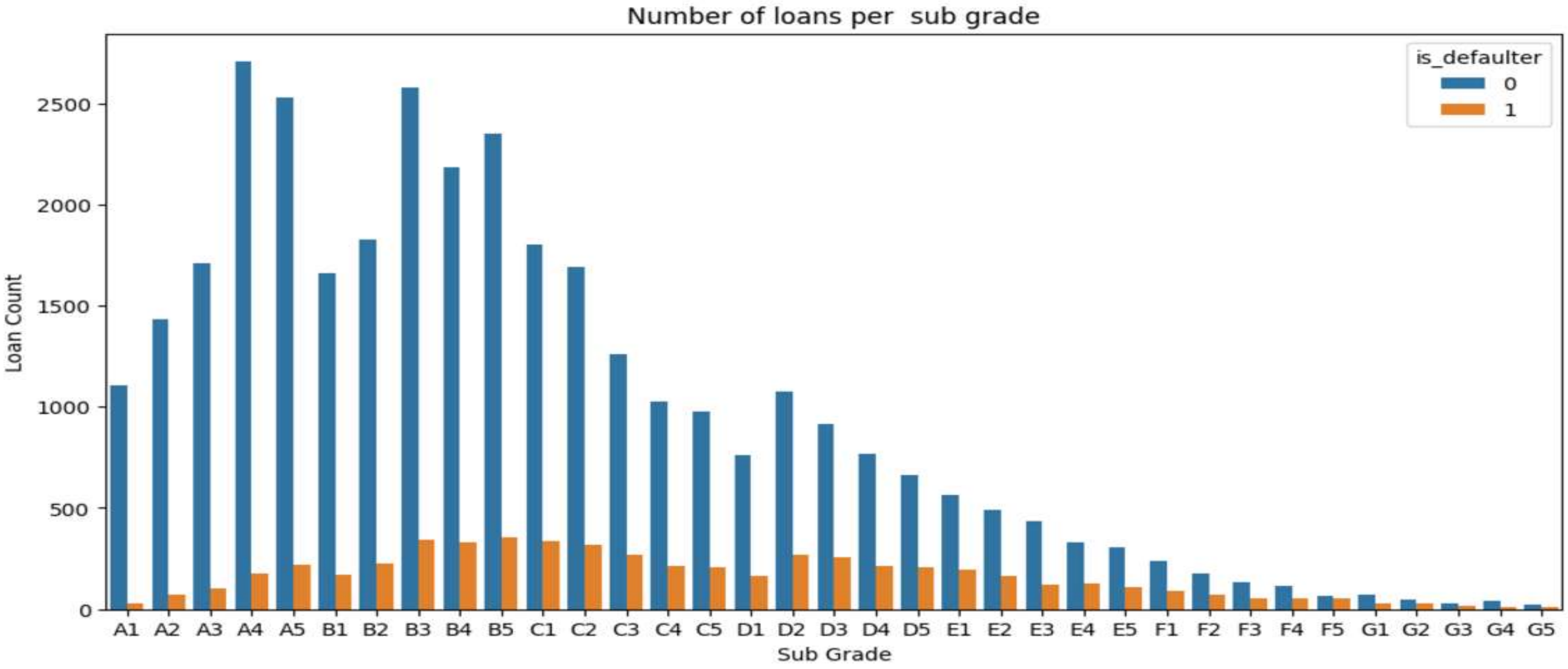


We can see clearly that less loan has been given to grade



we can clearly see that grading system have clearly reduce the defaulter percentage , hence it is having a direct impact on the defaulter .

Inference

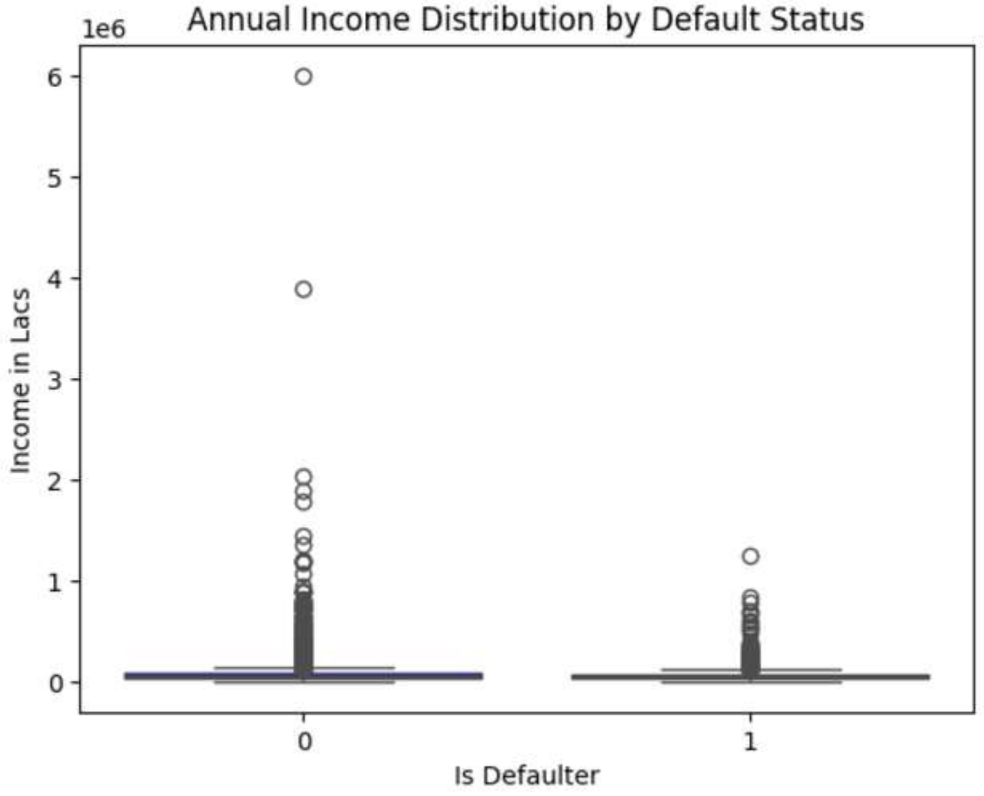


We can observe a trend where the number of loans approved decreases as we move into the riskier category.

Since grading and subgrading have a positive impact, we can deduce that we shouldn't lend money to clients or can increase the interest for the applicants who pose a greater risk.

Inference

Since the ratio of defaulters is declining as salaries rise, it is evident that there is an inverse relationship between salaries and defaulters.

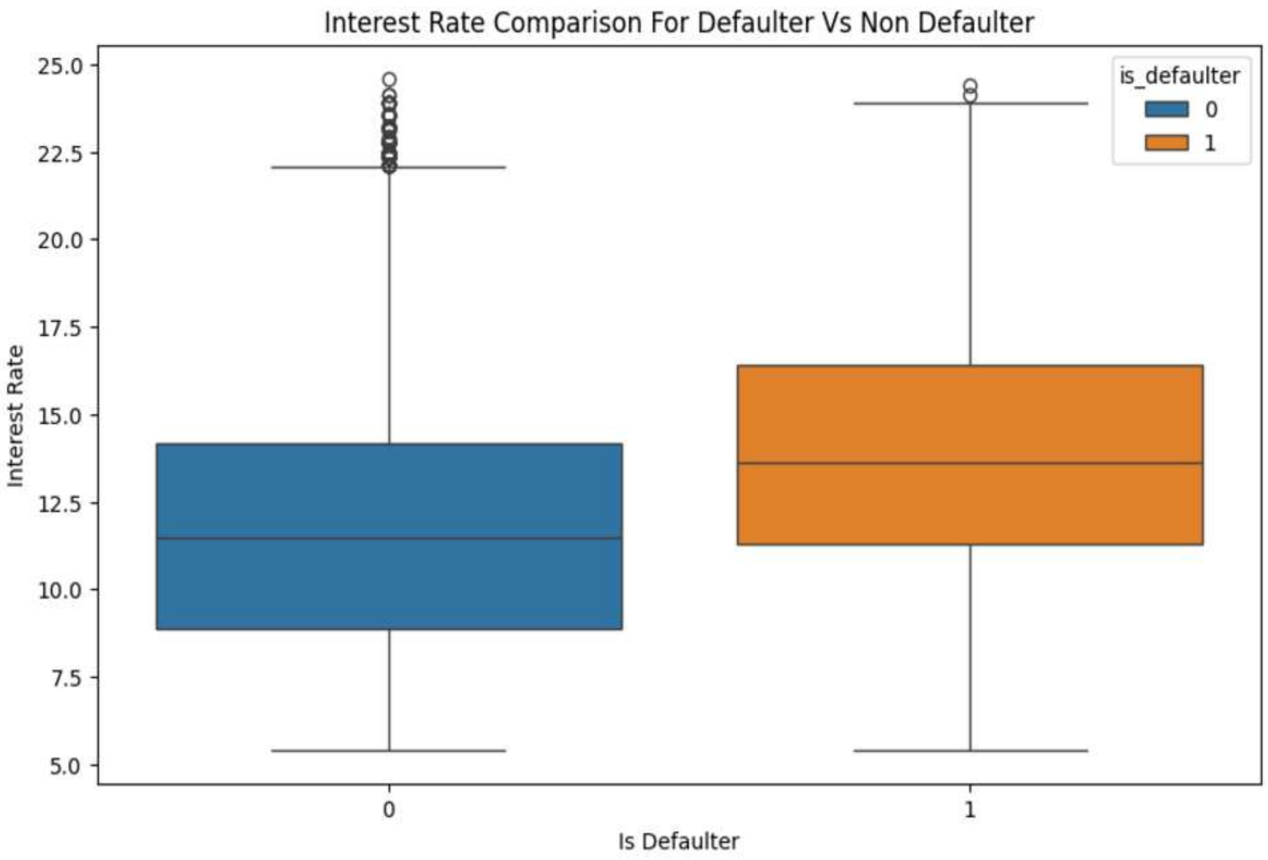


Inference

There is a greater likelihood of default for loans with higher interest rates since the mean interest rate for defaulters (13.82%) is higher than the mean interest rate for non-defaulters (11.72%).

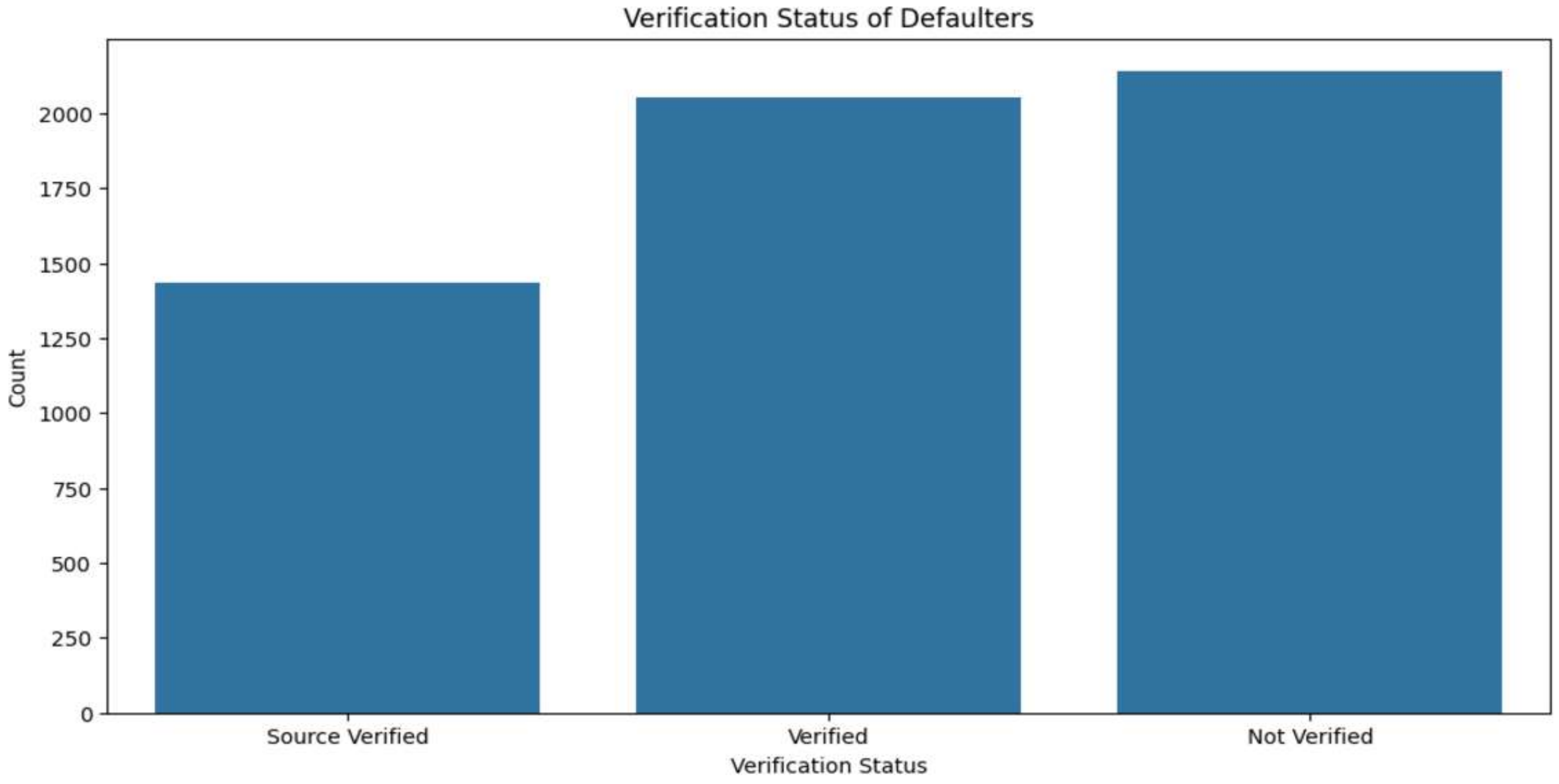
Since applicants with higher risk profiles are given higher interest rates, this could be a powerful signal of default risk.

According to this relationship, loans with interest rates higher than a particular threshold may require greater supervision, which can be helpful for risk assessment.



	count	mean	std	min	25%	50%	75%	max
is_defaulter								
0	34090.0	11.724186	3.652109	5.42	8.88	11.49	14.17	24.59
1	5627.0	13.820432	3.654413	5.42	11.31	13.61	16.40	24.40

Inference



Note: Visual is plotted based on records that have is_defaulter flag as 1.

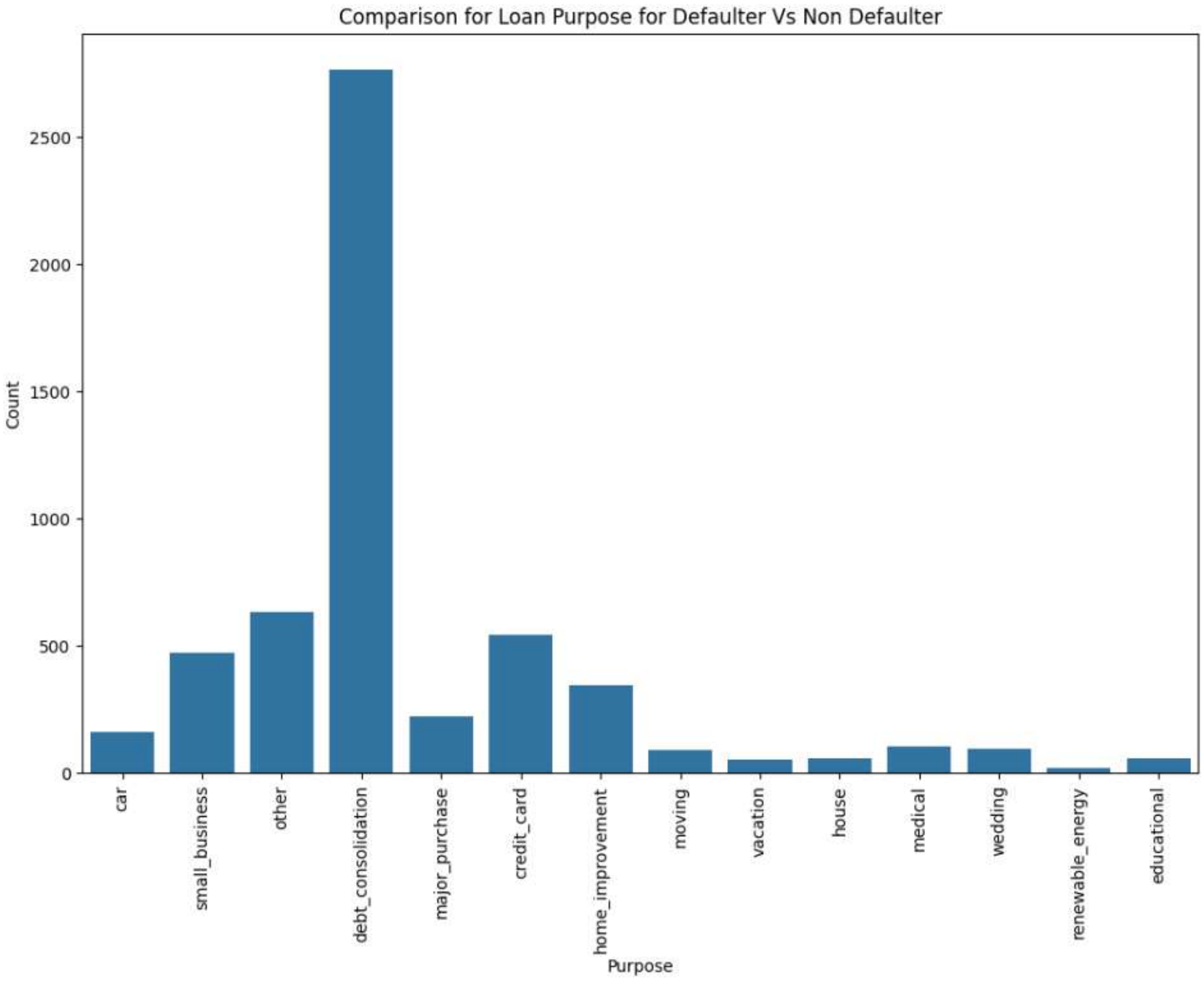
The count of Not Verified Applicants are higher as compared to to the Verified Applicants , hence a watch is required for applicants which are not verified from a good source with proper proof.

Inference

When compared to other categories, the majority of defaulters have occurred in case when the loan was taken out with the intention of debt consolidation. For this reason, we must specifically modify the policies and interest rates.

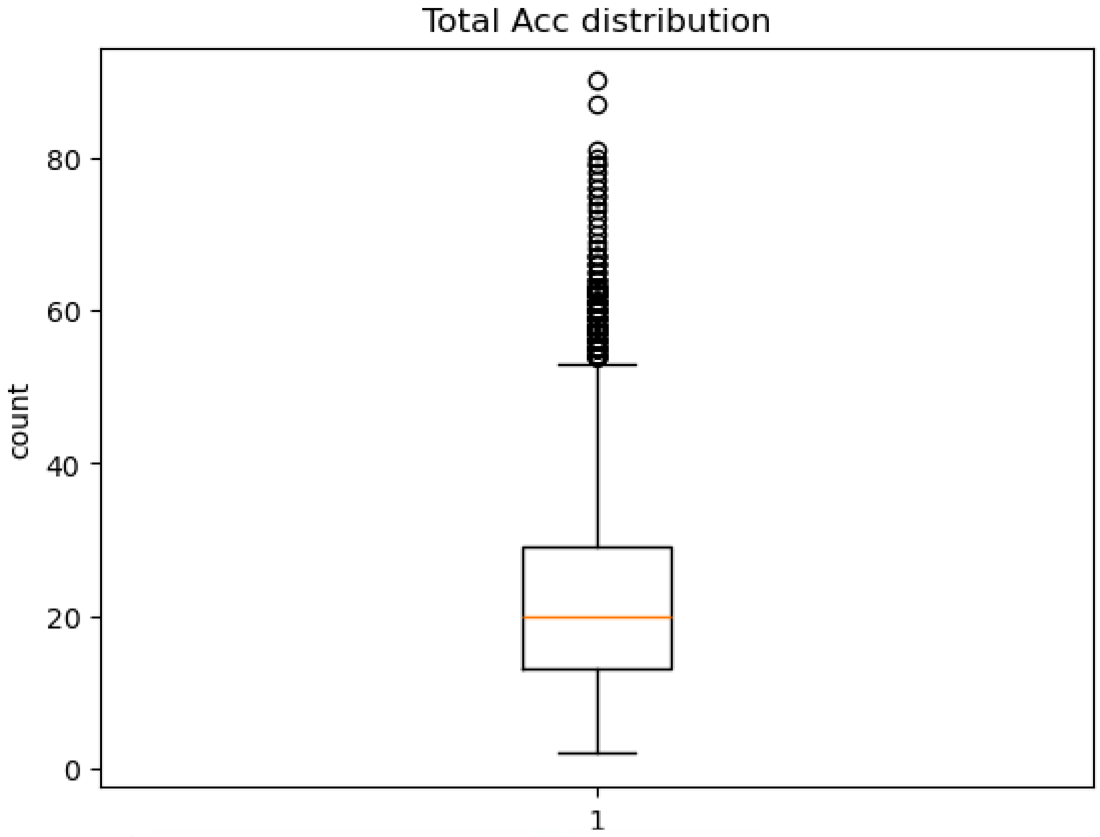
```
# numerical analysis
filtered_loan_df[filtered_loan_df['is_defaulter']==1]['purpose'].value_counts()
```

```
purpose
debt_consolidation    2767
other                  633
credit_card            542
small_business         475
home_improvement       347
major_purchase         222
car                    160
medical                106
wedding                 96
moving                  92
house                   59
educational             56
vacation                53
renewable_energy        19
Name: count, dtype: int64
```



Inference

Given that the number of defaulters varies by only 1%, we can draw the conclusion that anyone can become a defaulter, regardless of the number of credit accounts they own.




```python
# checking the defaulter with salary greater than upperwhisker_value_total_acc
(filtered_loan_df[filtered_loan_df['total_acc'] > upperwhisker_value_total_acc]['is_defaulter'].value_counts() /
len(filtered_loan_df[filtered_loan_df['total_acc'] > upperwhisker_value_total_acc])) *100
```

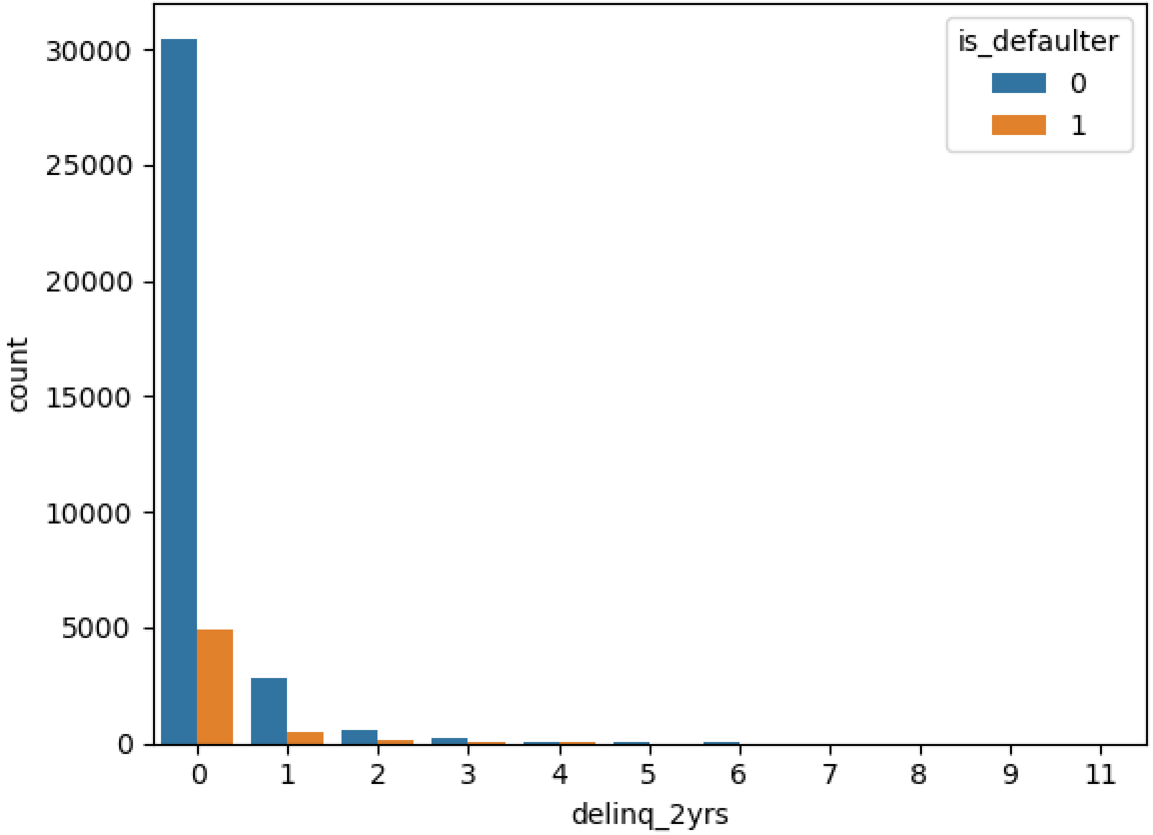
```
is_defaulter
0    86.666667
1    13.333333
Name: count, dtype: float64
```

```python
# checking the defaulter with salary smaller than upperwhisker_value_total_acc
(filtered_loan_df[filtered_loan_df['total_acc'] < upperwhisker_value_total_acc]['is_defaulter'].value_counts() /
len(filtered_loan_df[filtered_loan_df['total_acc'] < upperwhisker_value_total_acc])) *100
```

```
is_defaulter
0    85.810362
1    14.189638
Name: count, dtype: float64
```

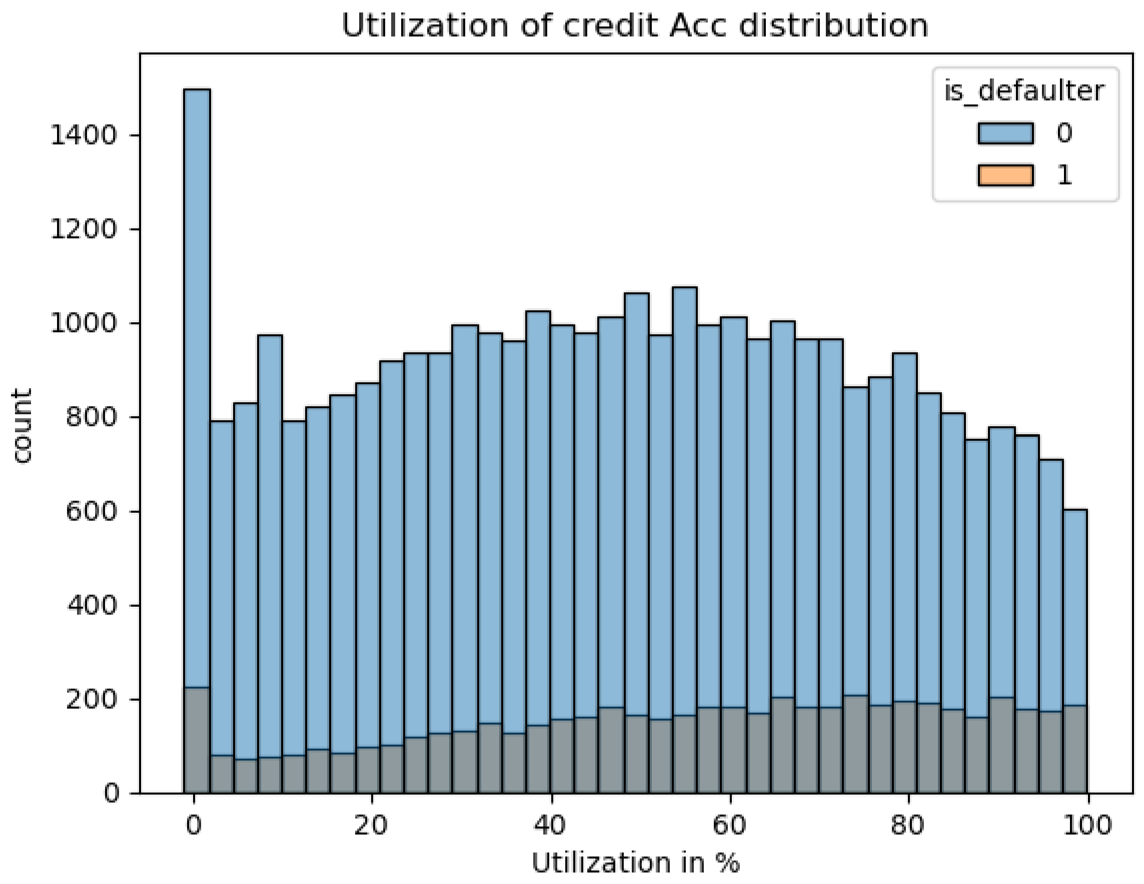
Inference

There are 5627 defaulters, and over 4936 of them have not defaulted in the last two years, so we may assume that they are either recent defaulters or have defaulted in the past.



Inference

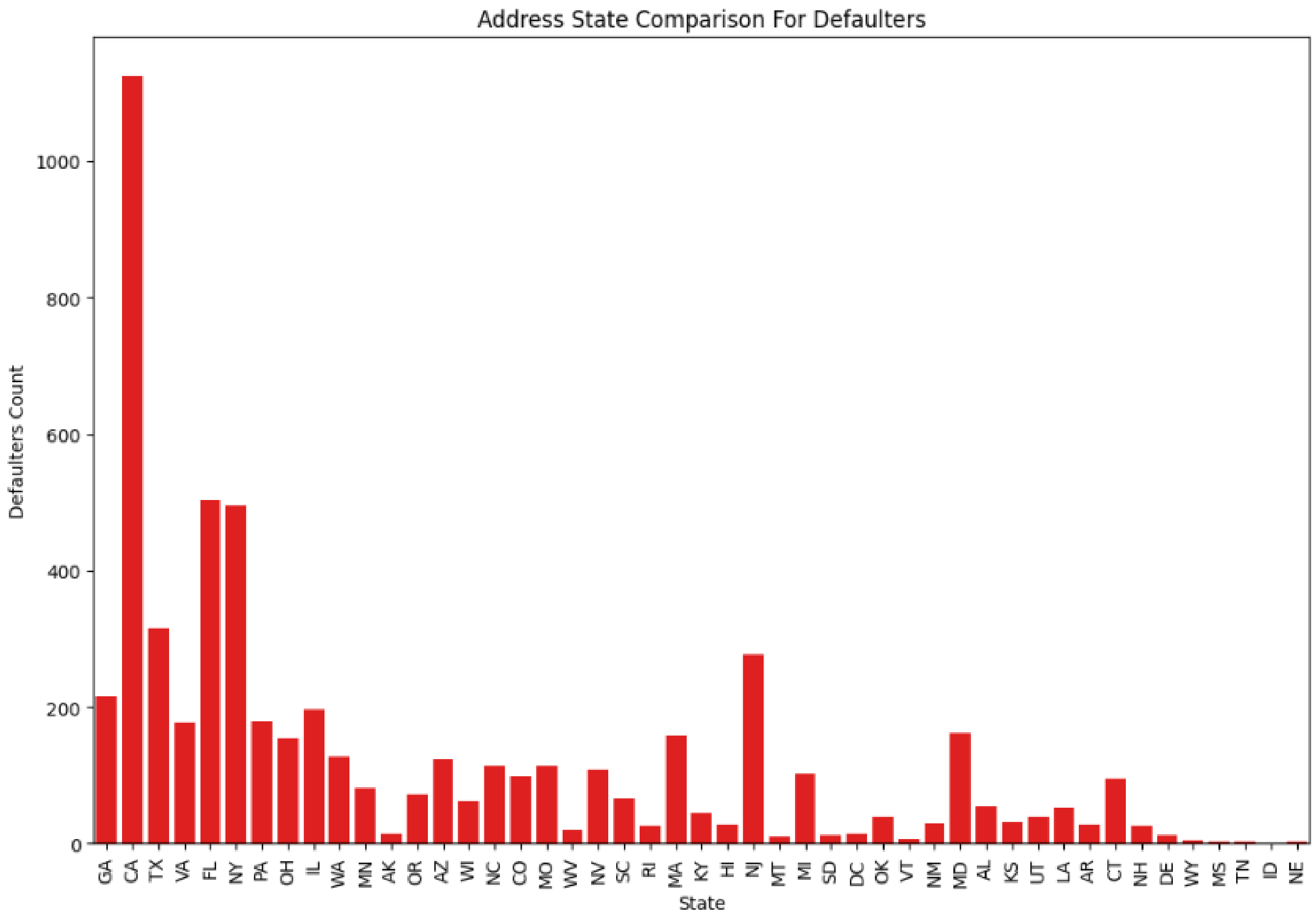
This has little to do with the defaulter, who comes from all ranges. It is evident that defaulters can belong to any category, regardless of how the credit accounts are used.



Inference

Maximum number of defaulters are coming from states CA , FL , NY

```
plt.figure(figsize=[12,8])
plt.title('Address State Comparison For Defaulters')
sns.countplot(data=filtered_loan_df[filtered_loan_df['is_defaulter']==1],x='addr_state',color='red')
plt.xlabel('State')
plt.ylabel('Defaulters Count')
plt.xticks(rotation=90)
plt.show()
```



Result of Analysis

Our research indicates that **borrowers with the following characteristics are significantly more likely to default on their loans**:

1. **Low Income**: Borrowers earning below a defined salary threshold.

2. **Purpose of Loan**: Loans taken specifically for **debt consolidation**.

3. **Loan Amount**: Borrowers seeking loans of **less than $30,000**.

4. **Loan Grade**: Loans with a grade **equal to or above the mid-tier (e.g., E or higher)**.

5. **Interest Rate**: Loans with **interest rates exceeding 15%**, reflecting prior credit challenges.

Key Insights:

•Borrowers in this demographic often face financial strain, leading to higher **default rates**.

•The combination of **high-interest rates** and **low salary** exacerbates their repayment challenges.

•**Debt consolidation** loans, though intended to streamline debt, often reflect existing financial instability.

These findings emphasize the need for **targeted risk assessment strategies** for such profiles to mitigate default risks.