



# Lending Club Case Study

## - EDA

---

***Group Members:***

Rajani Nagaraju

Sriram Dhayal

# Introduction



Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface. Like most other lending companies, 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. 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 one is able to 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.

Main objective is to identify driving factors behind loan default using EDA, which can be used by lending company for risk assessment when a loan application is received.

# Approach



---

## 01. Data Cleaning

Handling null values, outliers, invalid values. Split/merge columns. Removing unwanted columns, null rows, duplicate rows. Standardizing values.

---

## 02. Univariate Analysis

Analyzing each column by plotting the distribution of categorical variables and quantitative analysis on numeric variables.

---

## 03. Segmented Univariate Analysis

Analyzing distribution of single continuous variable across various segments.

---

## 04. Bivariate Analysis

Analyzing relationship between two variables.

---

## 05. Inference/ Recommendations


Analyzing all plots and graphs and provide recommendations to reduce the credit loss from loan defaulters.

# Data Cleaning

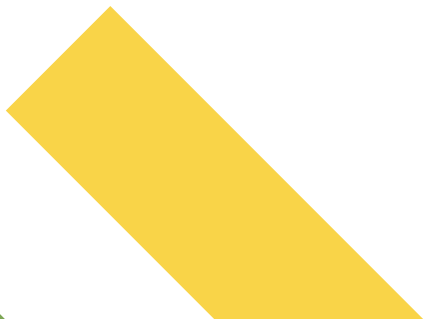
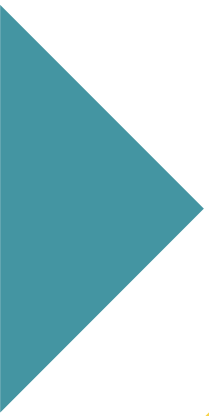


## Actions performed:

1. Dropped the columns with more than 80% missing values - 56 columns
2. Below columns are imputed with default values:
  - `mths_since_last_delinq` - is the number of months since the borrower's last delinquency. Imputed this value to 0 for missing values.
  - `revol_util` - For majority of records where revolving line utilization rate is missing, we see that revolving balance is 0, which ideally means `revol_util` is 0%. So impute these records with value 0.
3. Delete rows where data is missing:
  - `emp_length` - since this is a probable driving factor, not imputing the missing value. Instead, delete the rows where `emp_length` is missing
4. No action taken:
  - `pub_rec_bankruptcies` - is the number of public record bankruptcies. Since value 0 in this column indicates no bankruptcies, we cannot impute the value 0. Leave this column as-is.
  - `last_pymnt_d` - For all the records where last payment date is not recorded, we see that it is the defaulter who has not paid the amount. This is valid and hence not imputing data OR removing null value records
  - Below columns has single value for all non-null records. So, for current analysis these columns will not have any impact on target variable. Hence ignoring the missing values in these columns - `collections_12_mths_ex_med`, `chargeoff_within_12_mths`, `tax_liens`, `title`, `last_credit_pull`

- 
5. No duplicate records found.
  6. Remove unnecessary columns - 14 columns dropped.
  7. Column split: issue\_d is converted to date format and subdivided into issue\_month and issue\_year.
  8. Standardizing values:
    - int\_rate - remove % and convert to numeric value
    - revol\_util - remove % and convert to numeric value
    - emp\_length - As per the data description, emp\_length should be 0 for less than 1year and 10 for 10 years and above. Hence replaced <0 with 0 and >10 with value 10.
  9. Removing outliers on Categorical variables:
    - purpose, home\_ownership - remove rows with less than 1% of values
  10. Removing outliers in numerical variables - annual\_inc
  11. Numerical values as categorical variables - bins are created for few continuous variables which seems to have impact on target variable - annual\_inc\_bin, total\_pymnt\_bin, revol\_bal\_bin, revol\_util\_bin, int\_rate\_bin, dti\_bin, funded\_amnt\_bin, loss\_amnt etc.,


# Univariate Analysis



## Observations:

1. Defaulter rate is almost 15% while the fully paid applicants are 85%.
2. We could see more number of borrowers taking loan for the purpose of debt consolidation and paying credit card bills.
3. Loan approved to the borrowers in the category E,F and G grades are very less.
4. Lending club has expanded year on year. Every year the number of loans applicants are almost doubled.
5. Applicants living in Rented and mortgaged house tend to take more loan than the ones living in own house.
6. Almost 43% of the loans falls under 'Not verified' category where in annual income of the borrower is not verified.
7. Undoubtedly, loan is given only for borrower with no derogatory public records and no public record bankruptcies. Only very few exceptions can be seen.
8. Number of loans taken by the applicants with 10+ years of employment length is more than other categories.
9. Most of the borrowers would prefer 36 months loan period (75%) rather than 60 months (25%)
10. Most of the borrowers annual income is in the range of 20k to 80 k




- 
5. Most of the borrowers have taken loan between 5000 to 10000
  6. Amount distribution looks very similar across loan\_amnt, funded\_amnt and funded\_amnt\_inv. This indicates that requested loan amount is sanctioned 100% in most of the cases.
  7. Most of the interest rate lies between 9% to 14%
  8. DTI of most of the borrowers are in the range 10 - 20

# Segmented Univariate Analysis

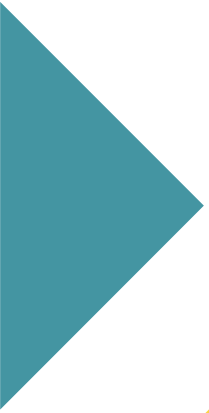


## Observations:

1. Loan applicants applying loan for 60 months are likely to default more than the one taking loan for 36 months
2. Higher the Grade, higher the rate of defaulters
3. The loan applicants who have been verified are defaulting more than the applicants who are not verified.
4. Top 5 states with Maximum defaulters are : CA, NY, FL, TX and NJ
5. The loan applicants who live in a rented or mortgaged house are more likely to default
6. The loan applicants have increased steadily from 2007 to 2011 showcasing positive trend in the upcoming years.
7. Though we see an exponential increase in number of defaulters over the years, when analyzed in terms of percentage distribution, we see that there is no huge variation in defaulter % with the years
8. Defaulter rate is high in the last quarter of the year
9. Most number of loan applicants are 10 or more years of experience. They also are the ones who are most likely to default. However, when analyzed in terms of percentage distribution, we see that there is no huge variation in defaulter % with the employment length
10. We see that Defaulter rate is high for higher loan amount.

- 
11. Debt consolidation is the category where maximum loans are issued and people have defaulted the most in the same category. However, when analyzed in terms of percentage we see most of the defaulters are under the category small\_business, medical and debt\_consolidation; Less defaulters under the purpose category of car and wedding.
  12. Majority of the loan applicants who charged off, reported an annual income between 30k to 60k.
  13. Majority of loan applicants who charged off has Debt-to-Income (DTI) ratio in the range 8-23.
  14. We see that defaulters rate increases with the rise in interest rate.
  15. We see that higher the revolving line utilization rate, higher the chance of defaulting.
  16. Loan applicants with higher grade has been given the loan with high interest rates.
  17. Interest Rate is less for 36 months when compared to 60 months.

# Bivariate Analysis



## Observations:

1. We could see steady increase in the amount loss because of loan defaults from 2007 to 2010 but peaking very high in 2011.
2. Higher the Grade of the loan applicant, higher is the interest rate
3. Majority of the defaulters are with loan amount < 15k and with interest rate in the range 10 to 20%
4. For the applicants with public derogatory records, loan provided for higher amount is mostly resulted in defaults.
5. Higher interest rate should be charged for higher DTI. However we could see that data is spread all across.
6. Interest Rate is high if Revolving line utilization rate is high.
7. Applicants with high annual income is approved with high amount loans.
8. Total credit revolving balance is found to be high for applicants with high annual income.

# Inferences

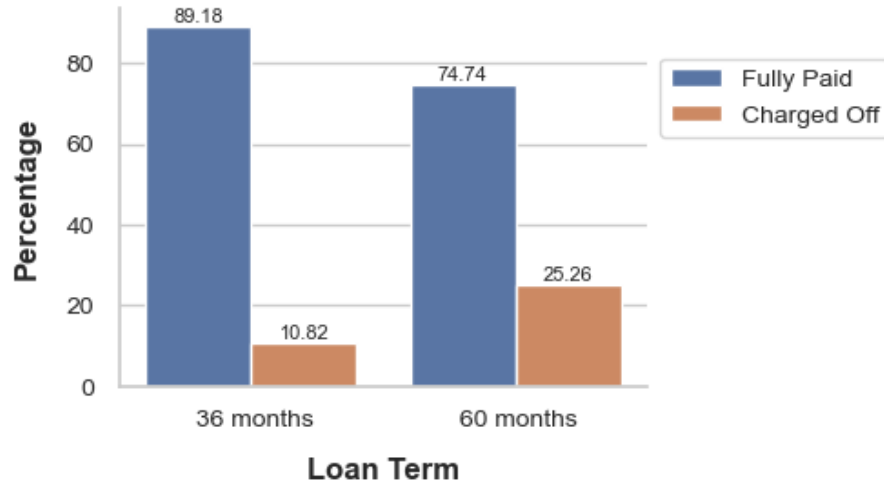


## Inference 1:

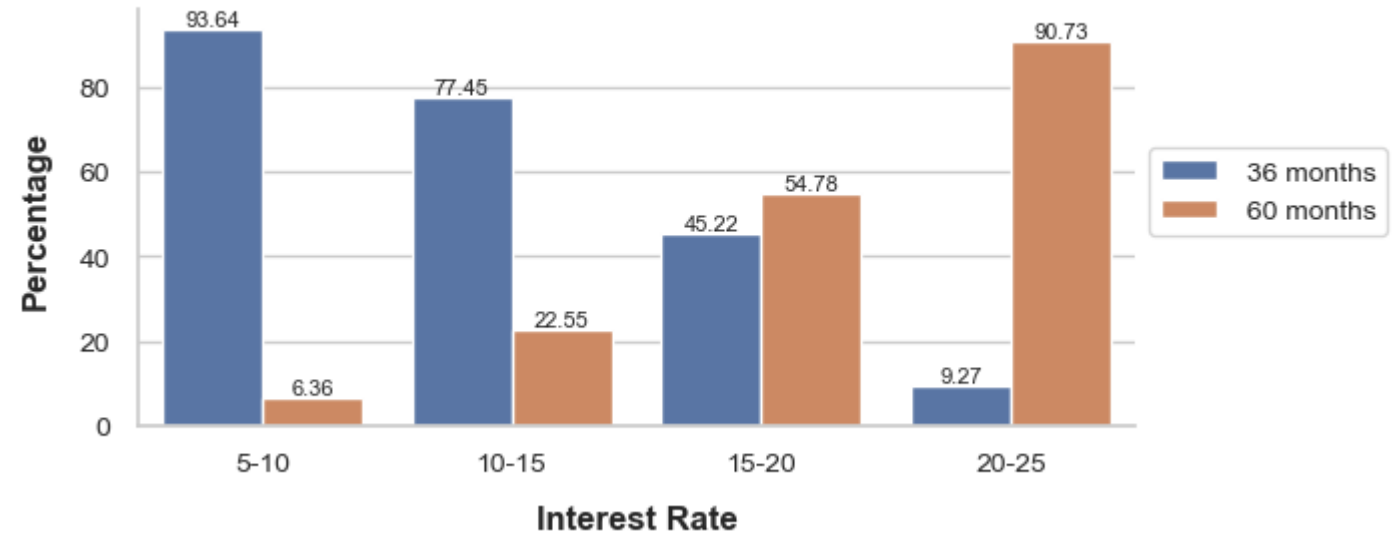
Borrowers who have chosen long term repayment like 60 months are more likely to default than for short term repayments, as the interest rate is also high.

Lending company should consider evaluating the risk associated with longer-term loans and potentially either limit the maximum term or adjust interest rates accordingly.

Loan Status Distribution w.r.t Loan Term



Interest Rate Distribution w.r.t Term

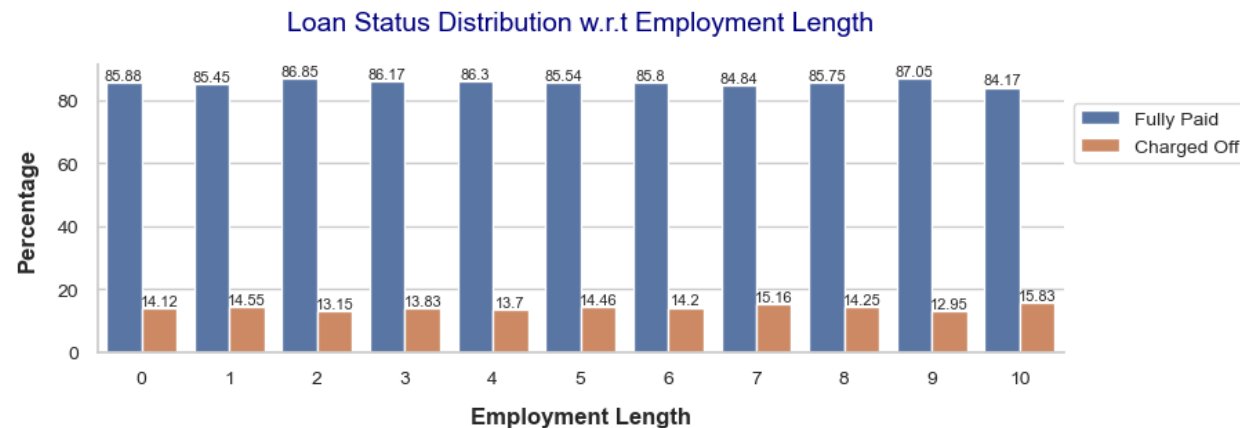
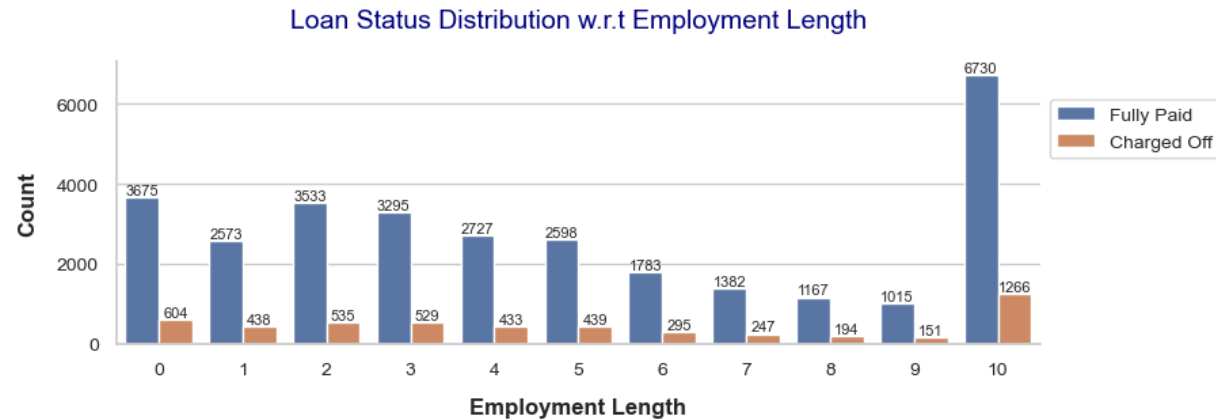




## Inference 2:

Applicants who had been employed for more than 10 years, accounted for the highest number of "Charged off" loans. However, when analyzed in terms of percentage distribution, we see that there is no huge variation in defaulter % with the employment length.

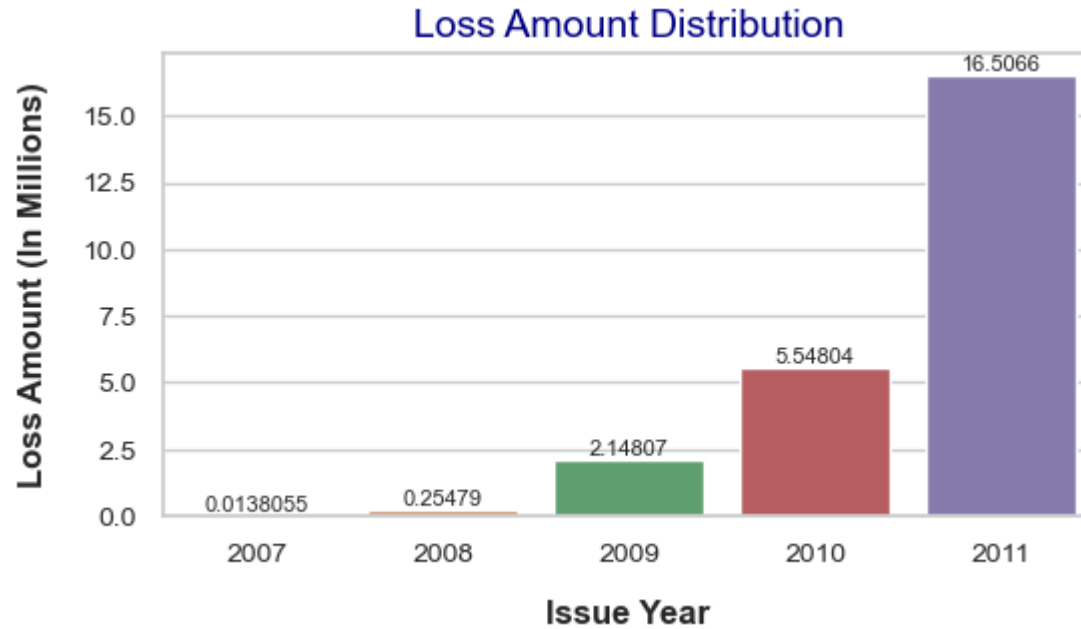
This indicates that long-term employment history does not necessarily guarantee successful loan repayment.



### Inference 3:

We could see steady increase in the credit loss because of loan defaults from 2007 to 2010 but peaking very high in 2011.

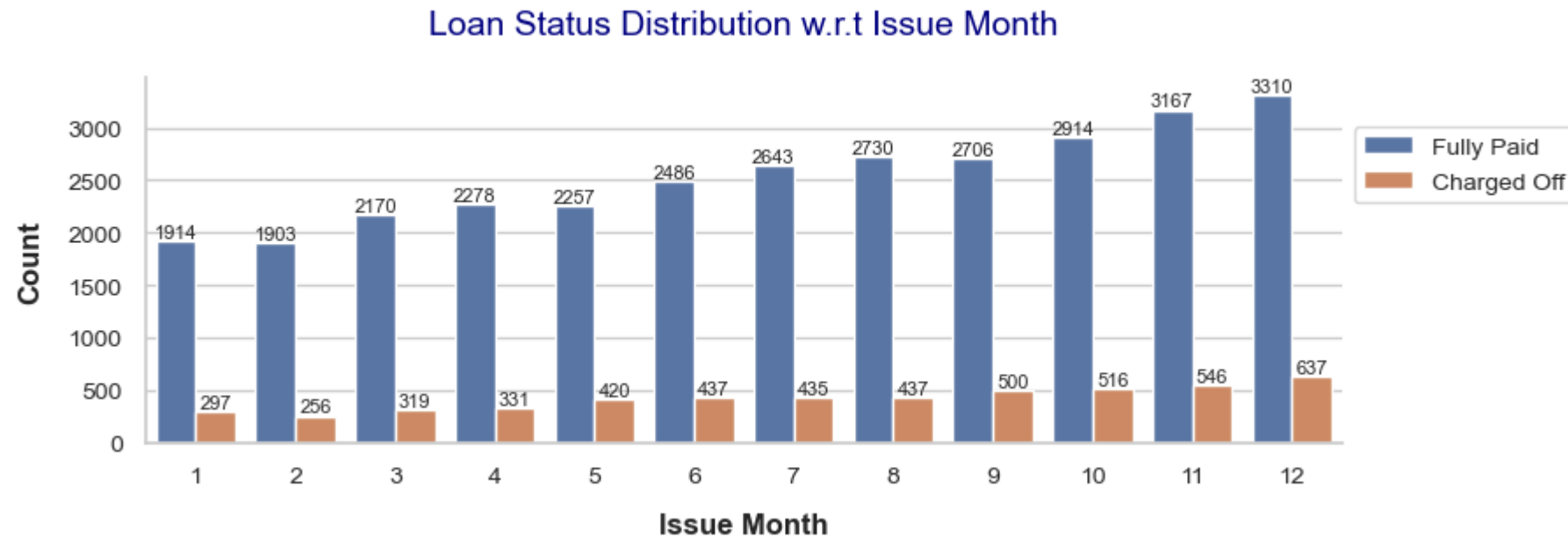
This could be indicative of economic or financial challenges during that year or defining the start of new trend.



## Inference 4:

"Charged off" loans were predominantly taken during the last quarter, primarily in December.

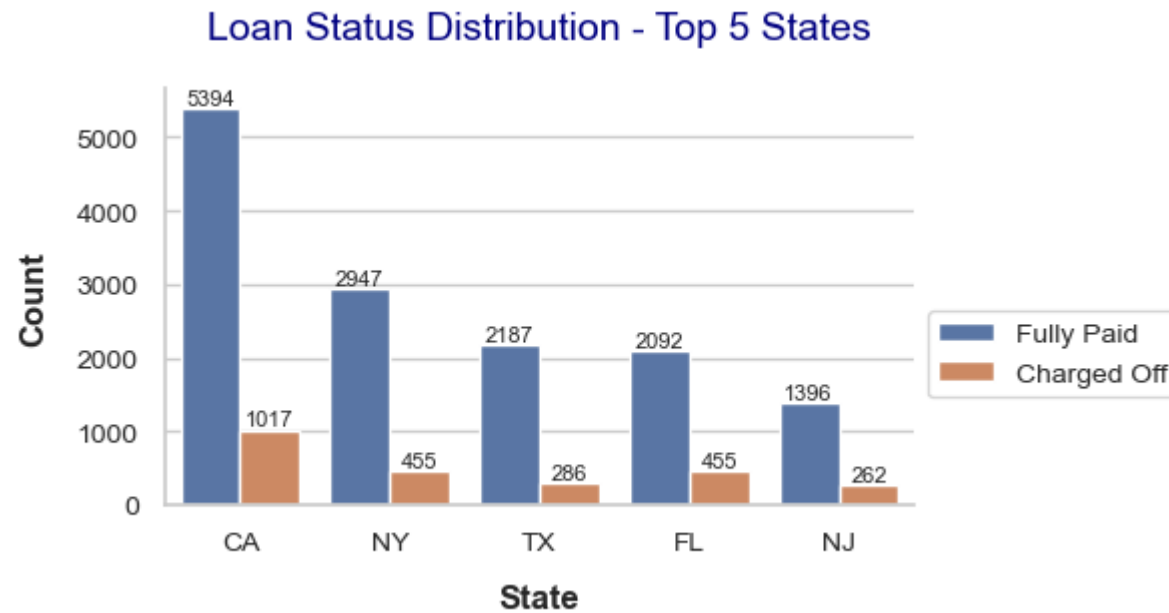
This peak in loan applications during the holiday season might suggest that financial pressures during the holidays might have contributed to loan defaults.



## Inference 5:

Top 5 states with Maximum defaulters are : CA, NY, FL, TX and NJ.

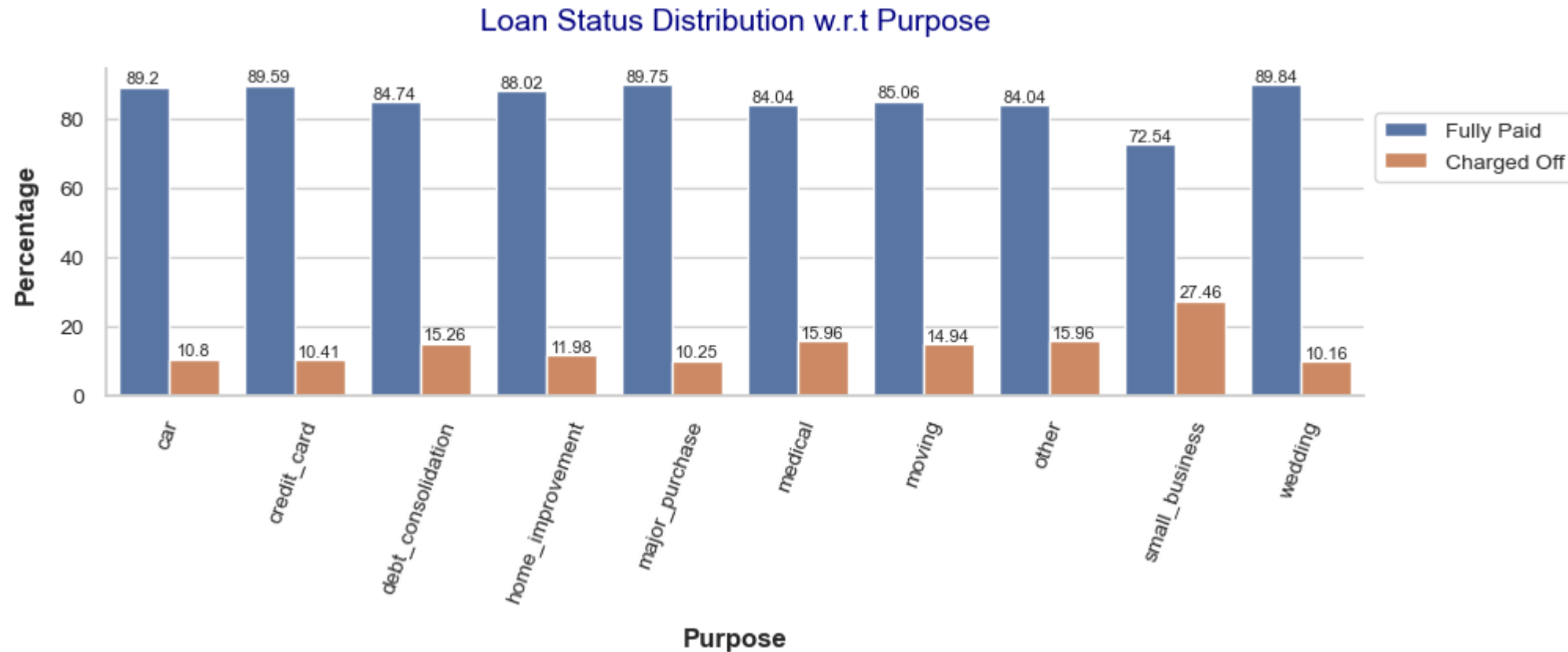
Company should monitor regional risk trends and adjust lending strategies or rates accordingly in these areas.



## Inference 6:

'Small\_business', 'medical' and 'debt consolidation' are the loan purpose categories where rate of defaulters is high.

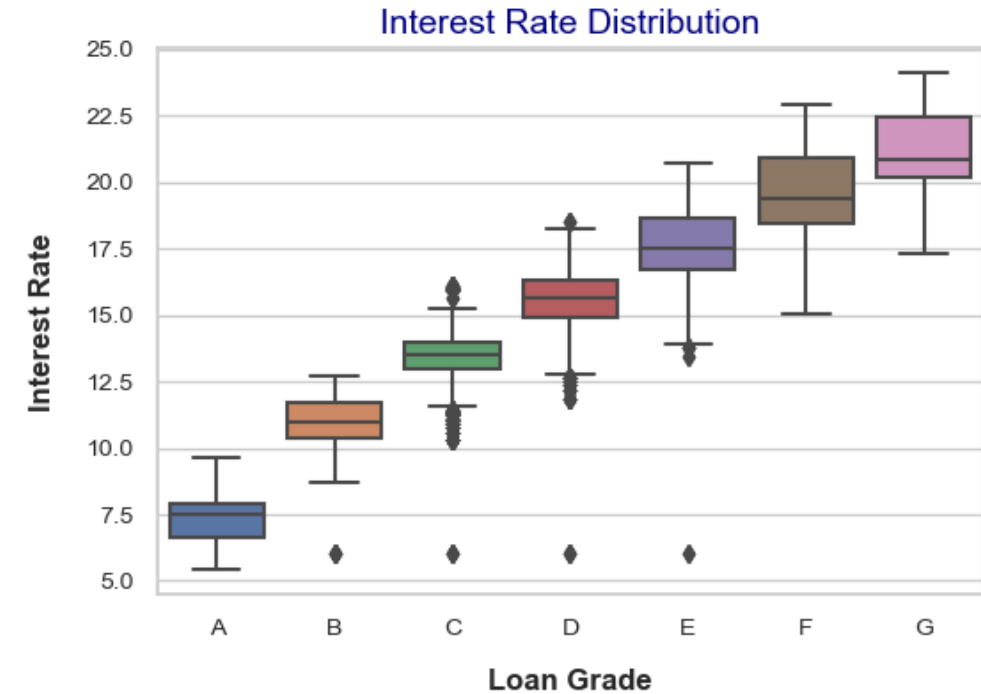
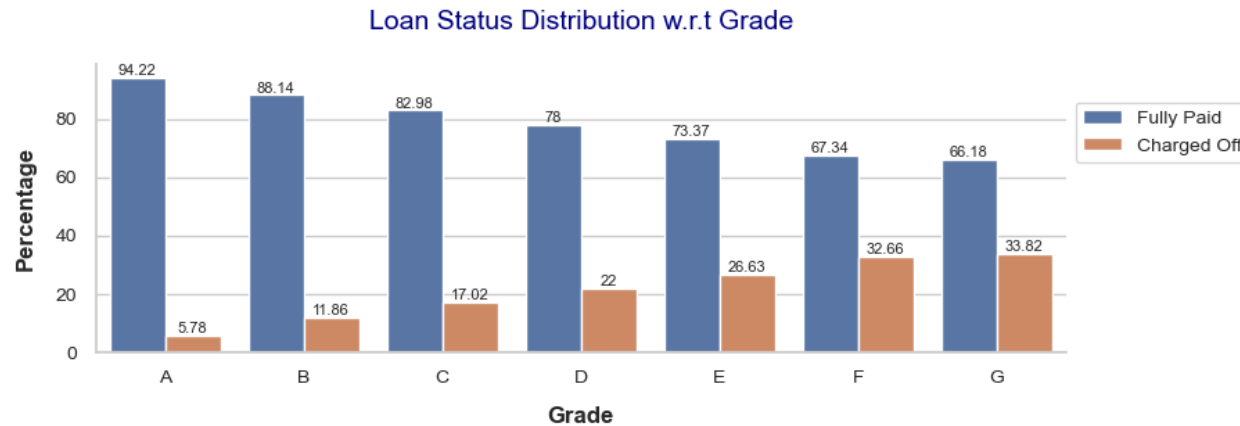
If the applicant is applying loan for one of these categories, lending company should scrutinize such applications as the probability of repayment is less.



## Inference 7:

Higher the grade, we see high is the defaulter rate.

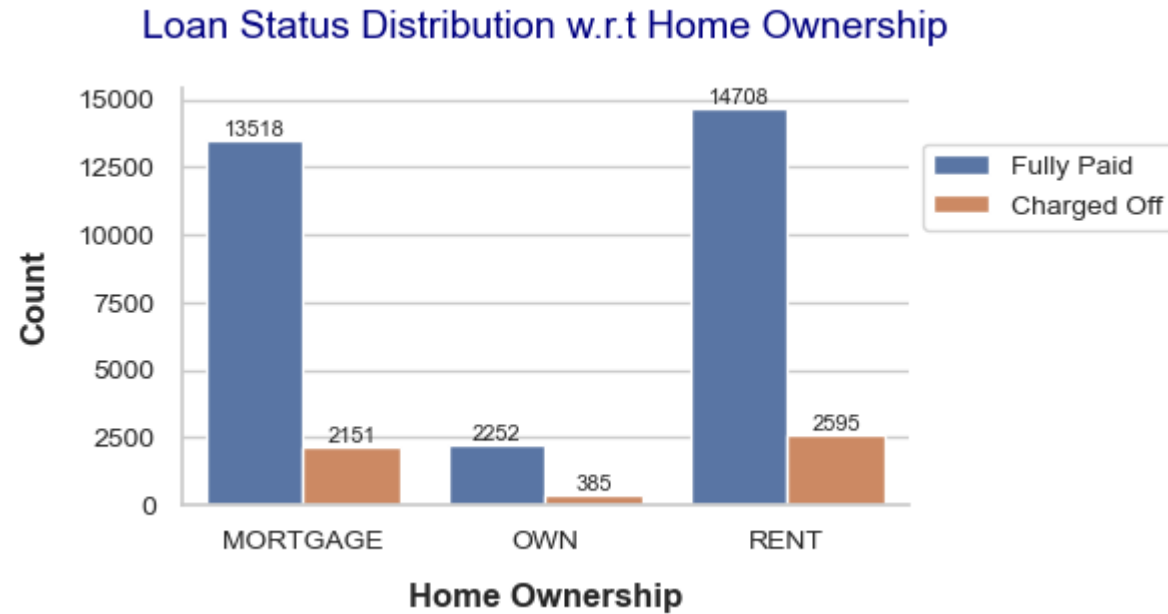
Applicants with higher credit grade faced challenges in repaying their loans as the interest rate is high for higher grades. Lending company should reconsider on interest rate and have better evaluation metrics on loan repayment for these high risk candidates.



## Inference 8:

Majority of the "Charged off" loan participants, lived in rented and mortgaged houses.

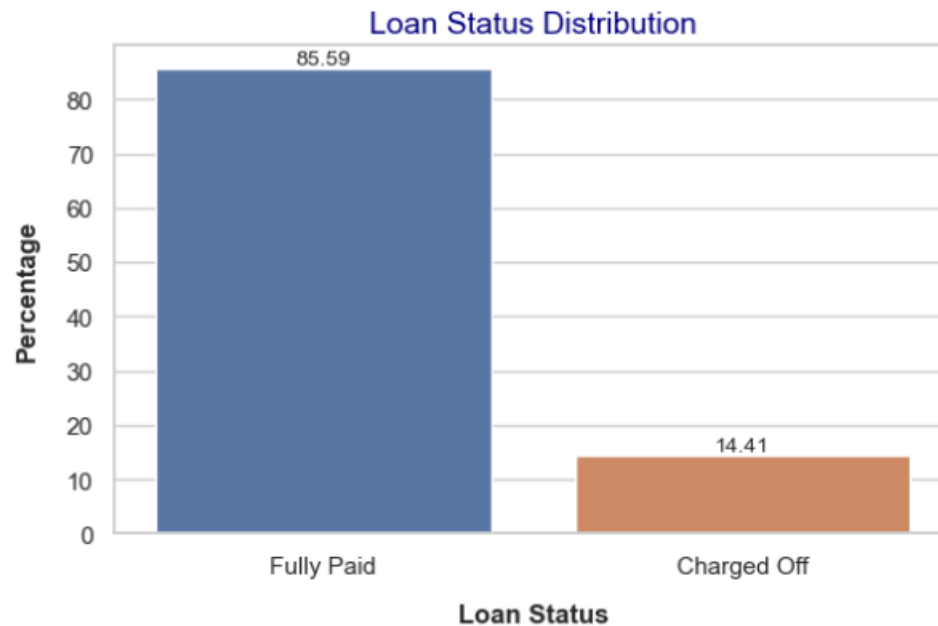
Lending company must assess the financial stability of these applicants, as they may be more susceptible to economic fluctuations.



## Inference 9:

Significant number of loan participants, around 15% were loan defaulters who were unable to clear their loans.

Lending company should enhance risk assessment practices and stricter credit checks. They can also offer financial education and support services to help borrowers manage their finances and improve loan repayment outcomes.

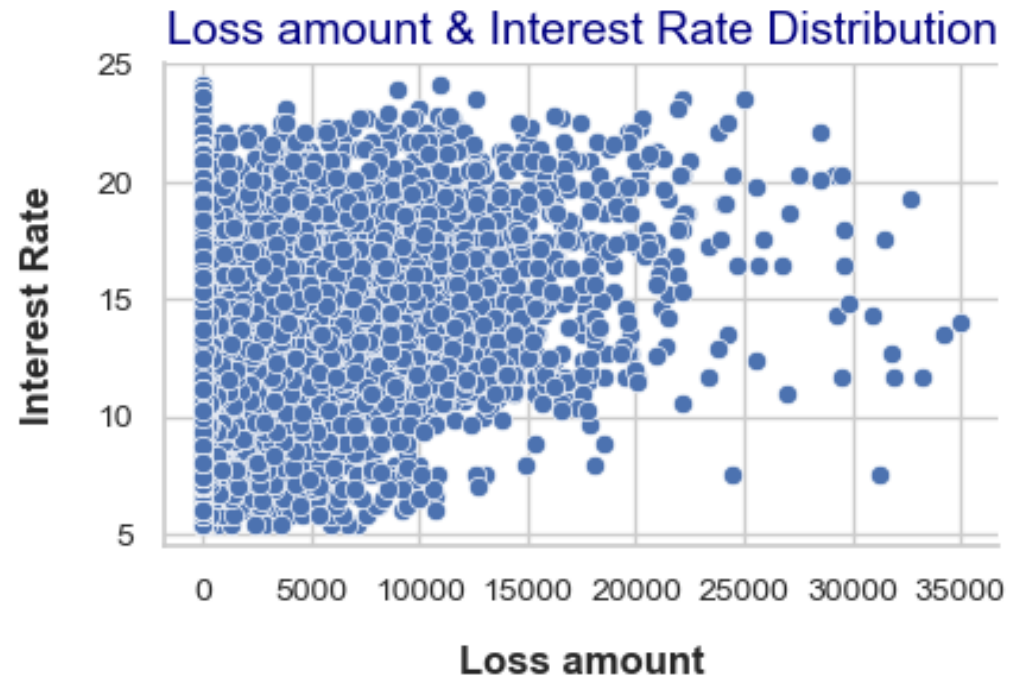




## Inference 10:

Majority of the defaulters are with loan amount < 15k and with interest rate in the range 10 to 20%.

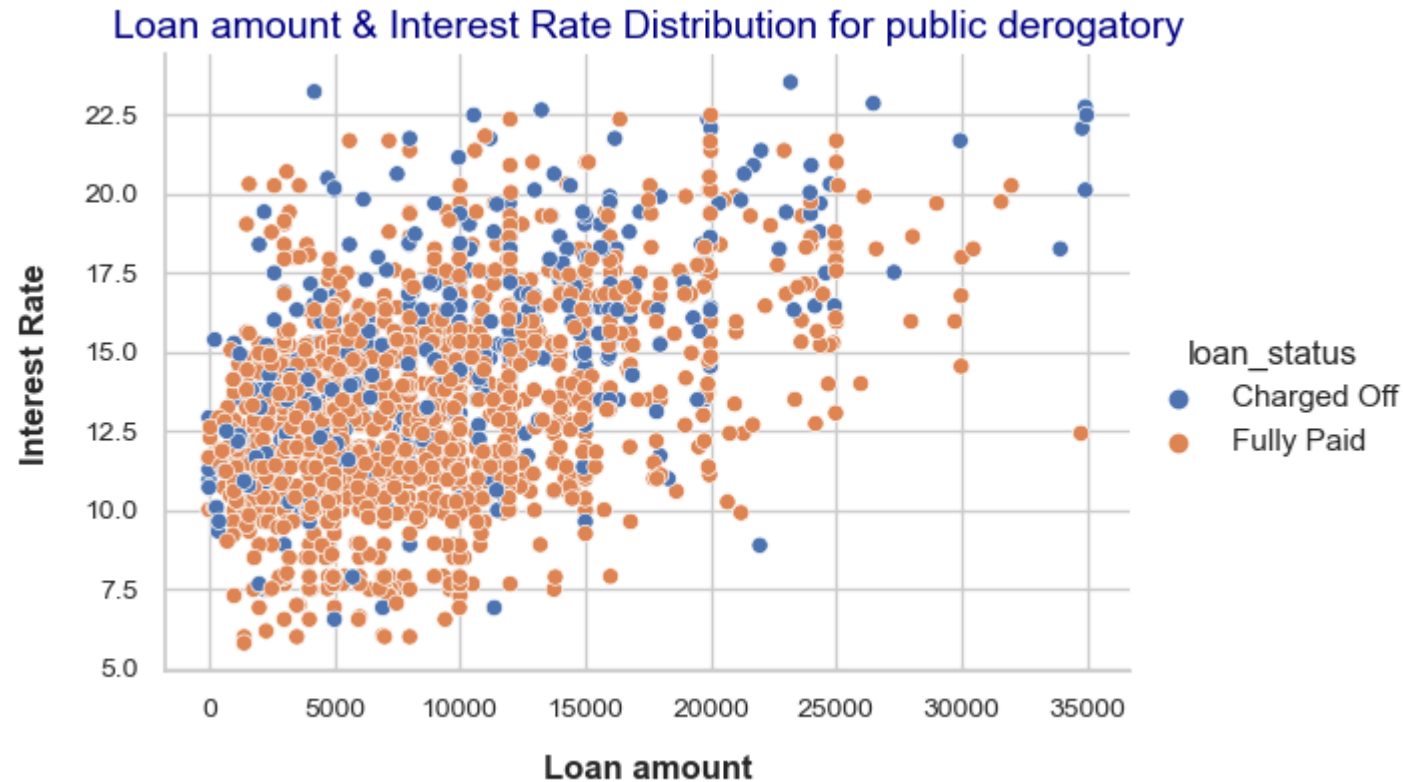
The company should review its interest rate determination process and consider adjusting rates based on DTI ratios to better align with the borrower's ability to repay.



## Inference 11:

For the applicants with public derogatory records, loan provided with higher amount is mostly resulted in defaults.

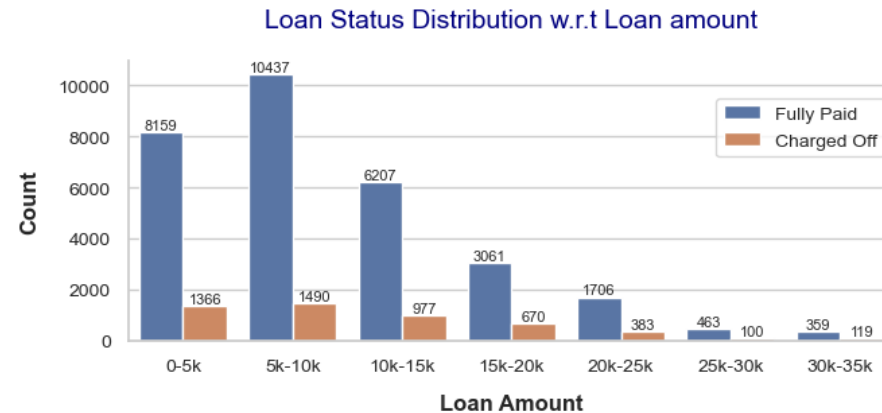
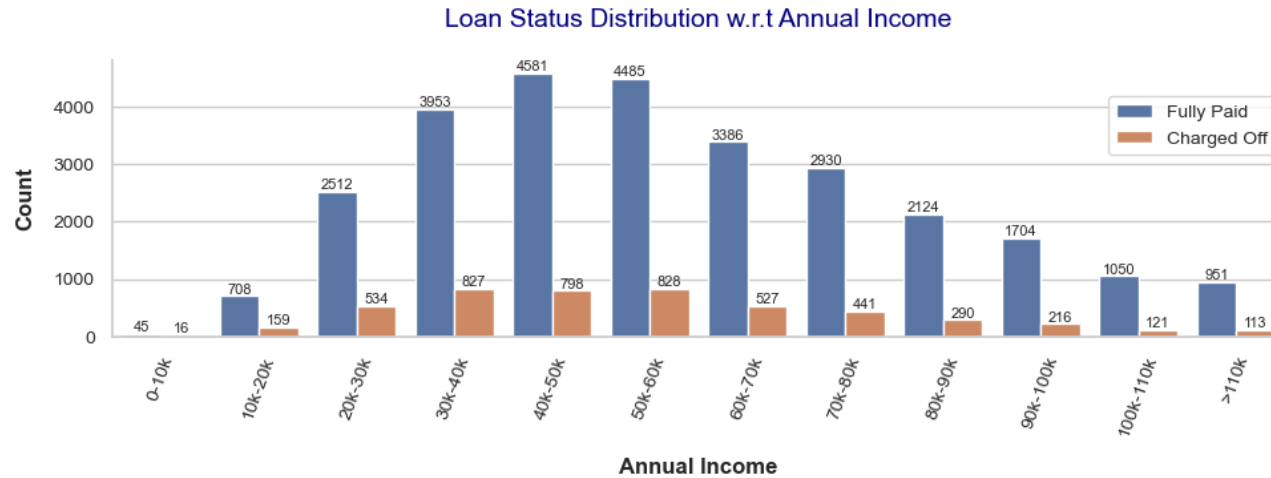
The company can mitigate this risk by conducting more thorough assessments for these applicants and potentially capping loan amounts for higher-risk applicants.



## Inference 12:

Majority of the loan applicants who charged off, reported an annual income between 30k to 60k and for loan amount <15k.

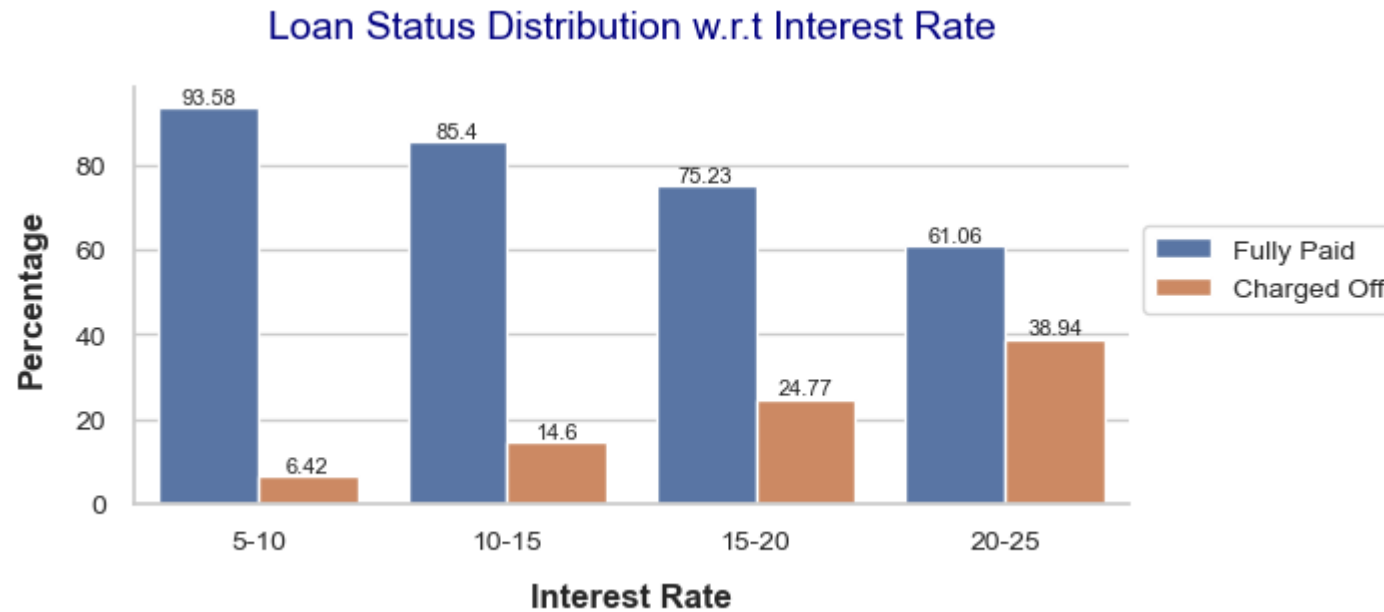
The lending company should exercise caution when lending to low income individuals. They should implement rigorous income verification and assess repayment capacity more thoroughly for applicants in this income bracket.



### Inference 13:

Among loan participants who charged off ,considerable portion belonged to the interest rate bucket of 15%-25%.

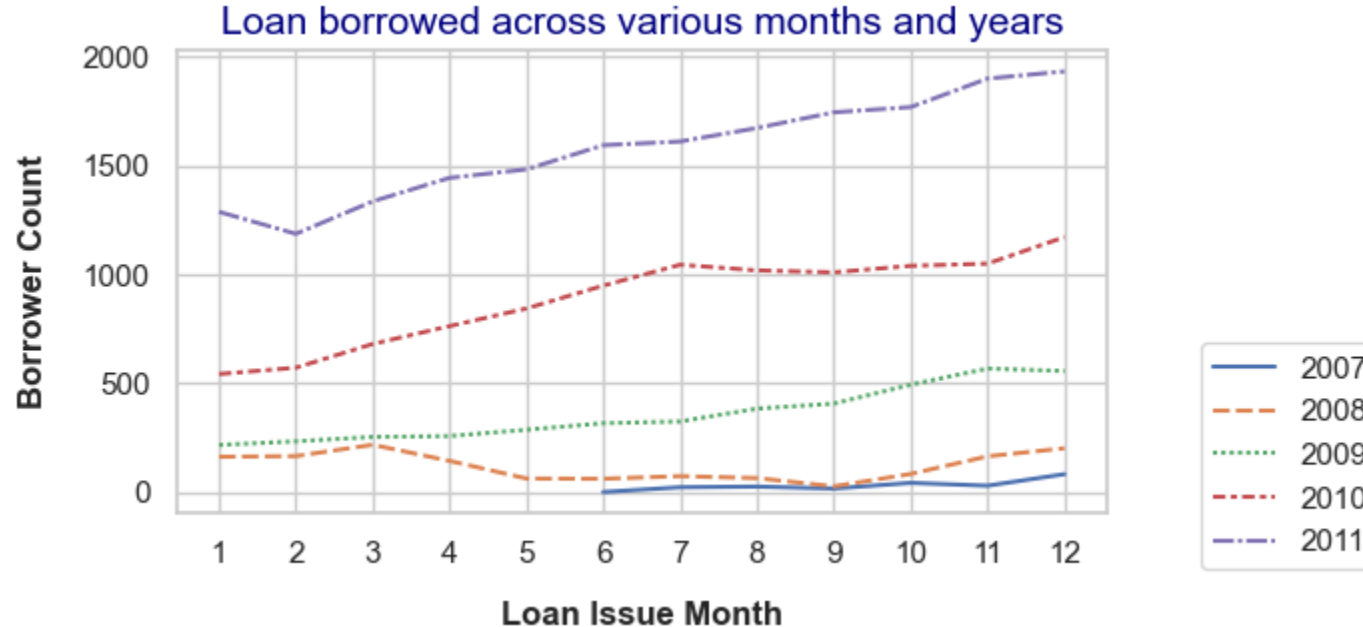
To reduce the risk of default, the lending company can consider offering loans at lower interest rates when possible.



## Inference 14:

The steady increase in the number of loan applicants from 2007 to 2011 indicates growth in the market.

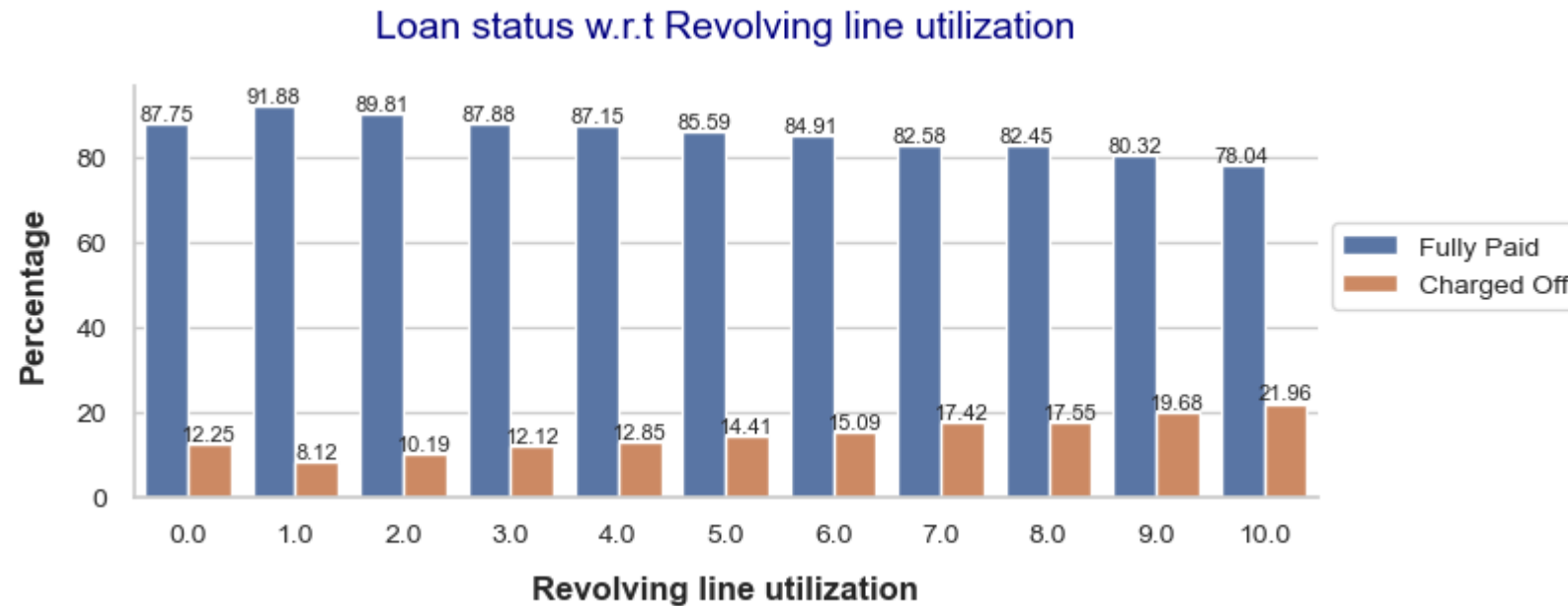
The company can capitalize on this trend by maintaining a competitive edge in the industry while keeping better risk management practices..



## Inference 15:

Higher the revolving line utilization rate, higher the chance of defaulting.

Lending company should consider revolving line utilization when approving loan and also recalibrate on interest rates.

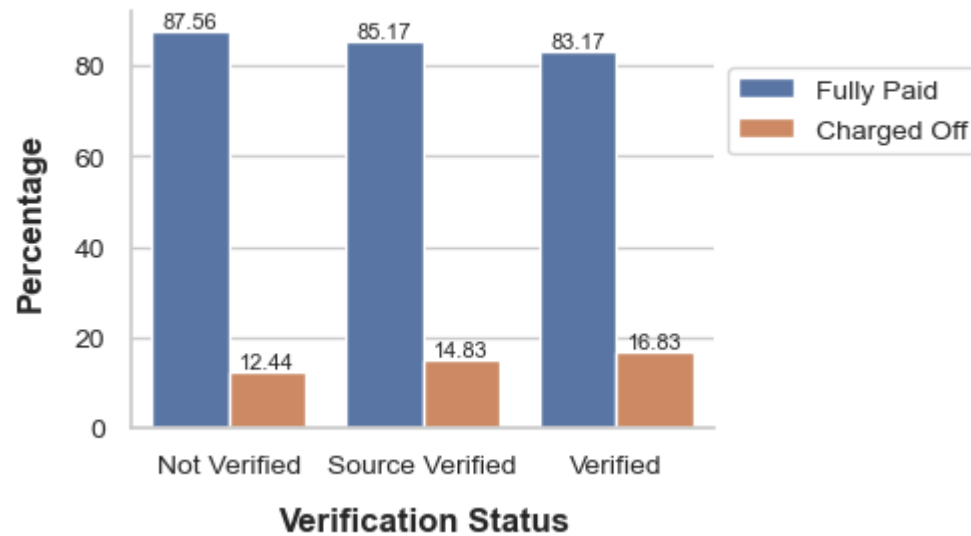


## Inference 16:

Verified loan applicants are defaulting more than those who are not verified.

The company should review its verification process to ensure it effectively assesses applicant creditworthiness and consider improvements or adjustments.

Loan Status Distribution w.r.t Income Verification





# Thank you

---

Rajani Nagaraju

Sriram Dayal