Lending Club Case Study

Group Members:

Karthika K Namboothiri

Kiran Hunagund

Course: ML-66 : upGrad and IIITB Machine Learning & AI Program - June 2024

Introduction

- This project mainly focusses on Exploratory Data Analysis in the banking field.
- EDA is done 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.
- 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.
- The objective of analysis is to use the information about past loan applicants and find whether they ‘defaulted’ or not.

➨ **Dataset**

- The dataset used is named as loan.csv.
- It contains details about the customer id, employment details, address, loan amount availed, interest rate, whether the customers have defaulted or not etc.
- This dataset can be used to get the patterns required to see how a customer has defaulted.

Problem Solving Pipeline

Data Understanding → **Data Cleaning** → **Data Analysis** → **Conclusion**

- Understanding the dataset provided.
- Studying the columns, the datatypes of each column, values present int hem, NULL values present etc

- Removing columns having all NULL values.
- Removing columns having NULL values>30%.
- Replacing the values in numerical type columns with median or mean and character type columns with mode.
- Creating derived columns from existing columns.
- Changing the datatypes of some columns.

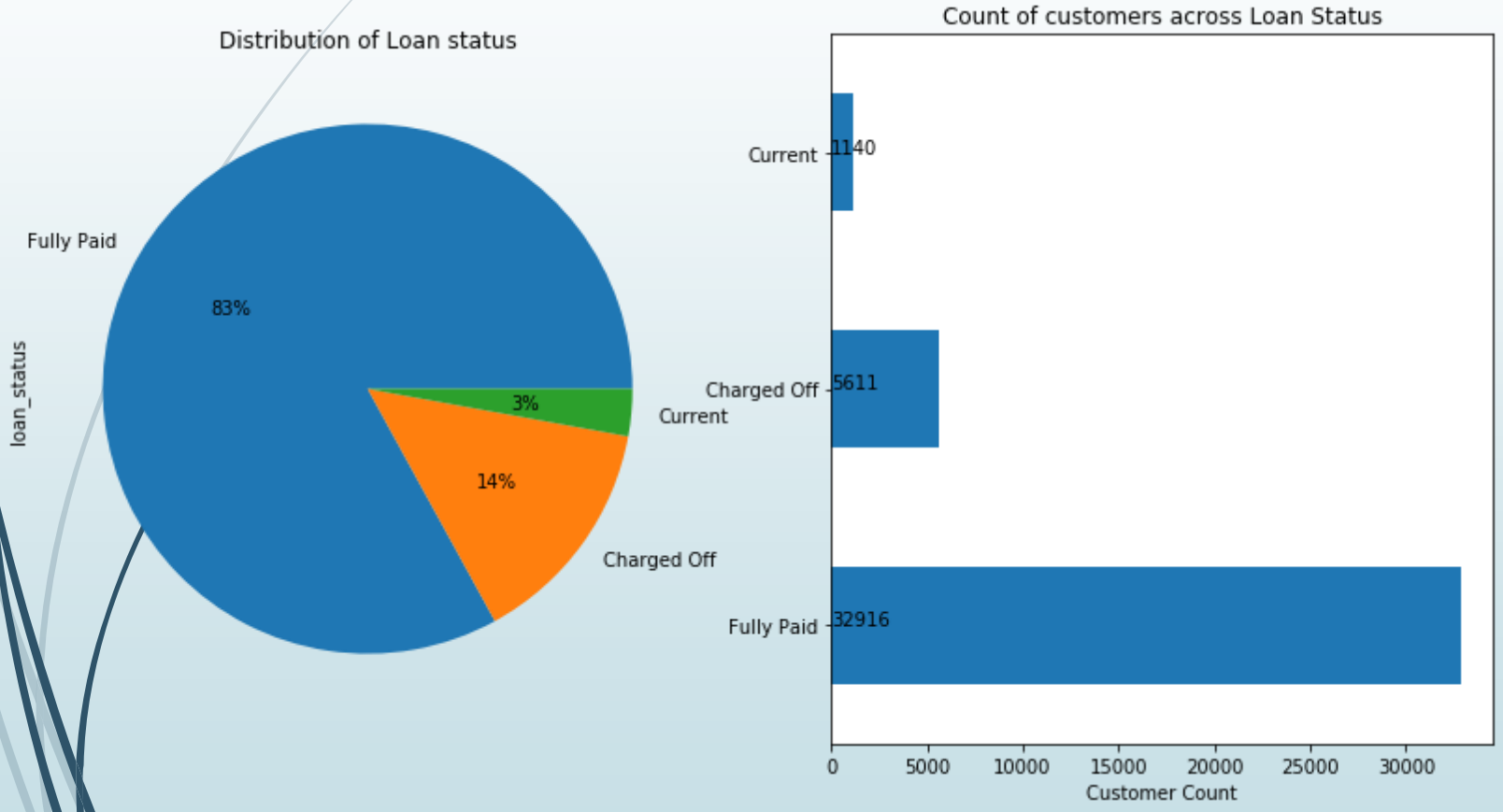
- Univariate analysis and bivariate analysis on the cleaned dataset.
- Univariate analysis involves taking each column into consideration and analyzing them.
- Bivariate analysis involves using more than 1 column at a time for analysis.

- Drawing conclusions from our analysis and finding the patterns with which it can be determined whether customers can get defaulted or not.

Data Understanding and Cleaning

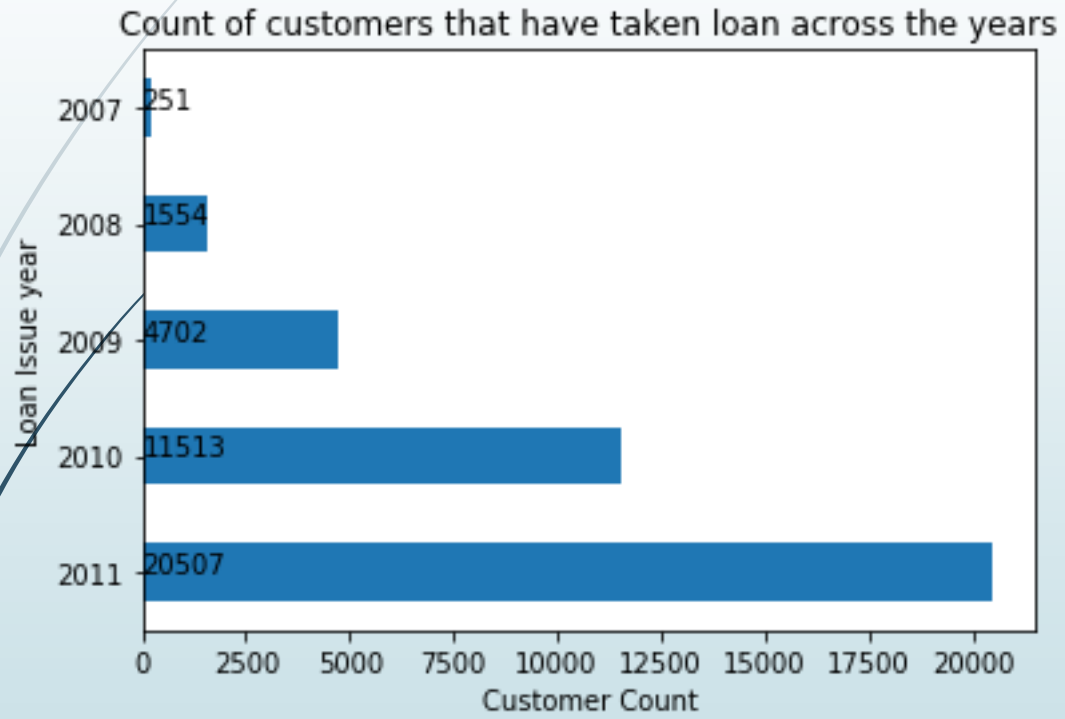
- The given dataset is named as loan.csv.
- It contains details about customers lie their their personal details , the loan amount availed, interest rate, term of loan and whether they have defaulted or not.
- Data is cleaned by first calculating the percentage of null values in each column.
- Columns having null value percent greater than 30% are removed.
- The columns for which the null percent is less than 30%, for columns of numerical datatype, the missing values are replaced with median or mean and for the character type columns, they are replaced with mode.
- The datatypes of columns like interest rate are changed to numerical type for analysis.
- The required columns for analysis are selected from the loan dataset for further analysis.

Analysis



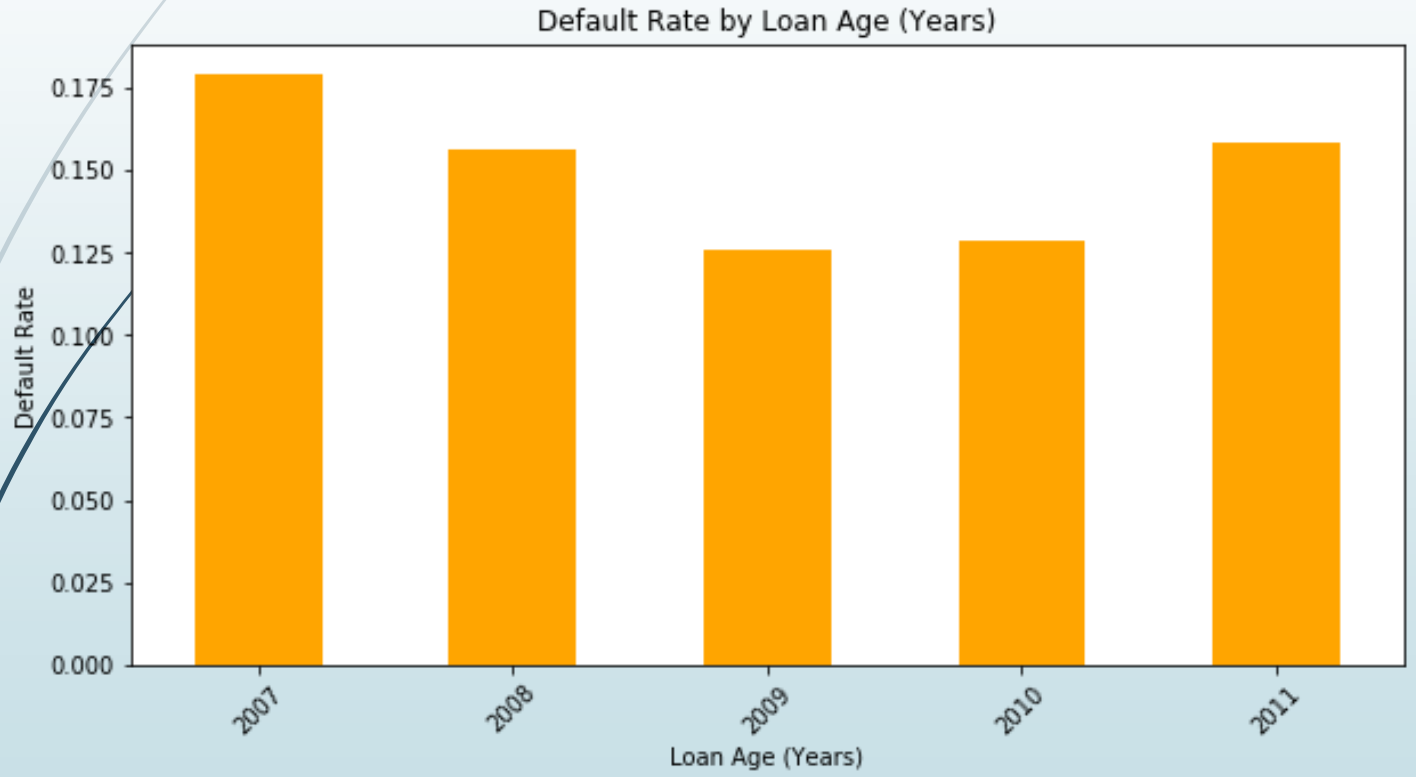
- Out of the whole customer base, 14 % have defaulted, which is equal to 5611 customers.

Analysis



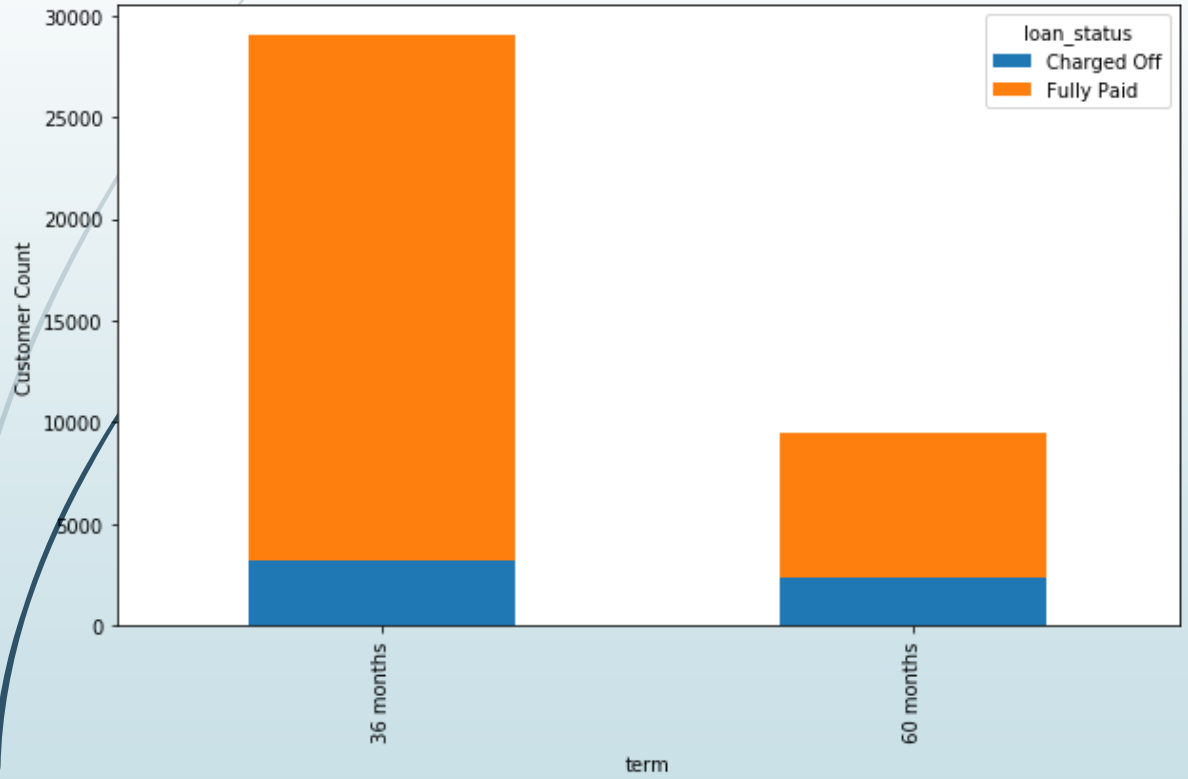
- The bank has given more number of loans since 2007, 2011 being the year with the highest number of loans provided.

Analysis



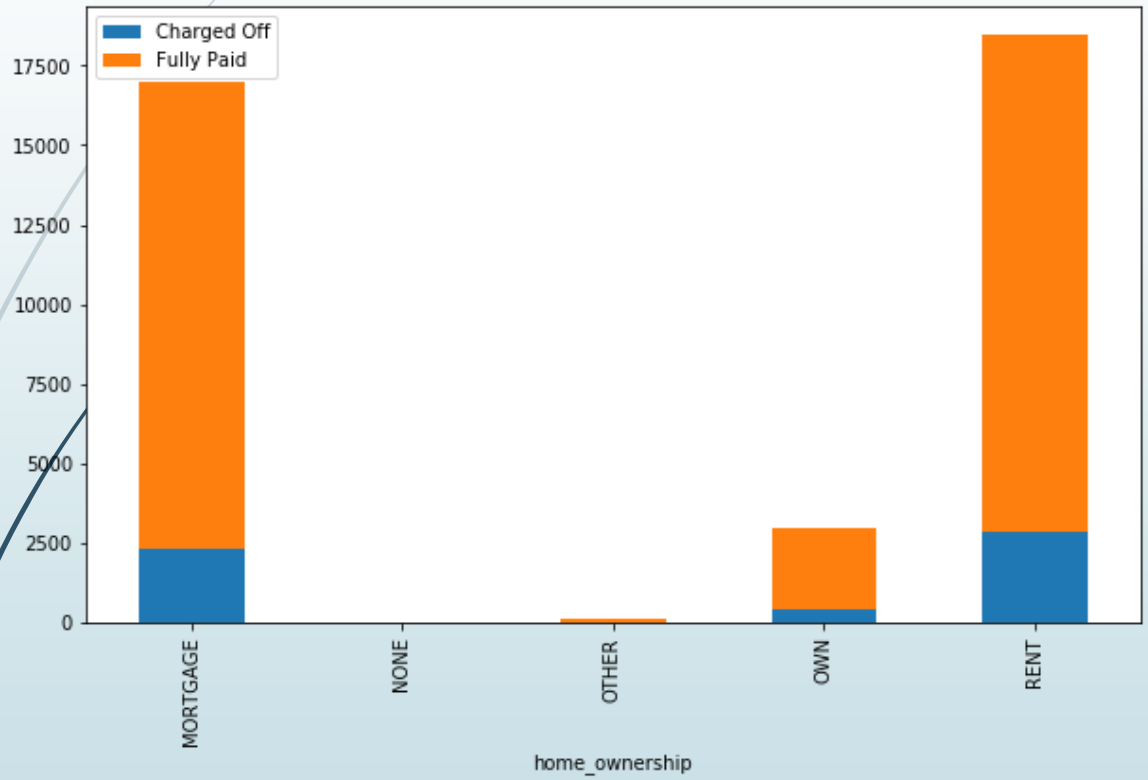
- The default rate is also highest in 2011, the same year in which the number of loans given was also highest.

Analysis



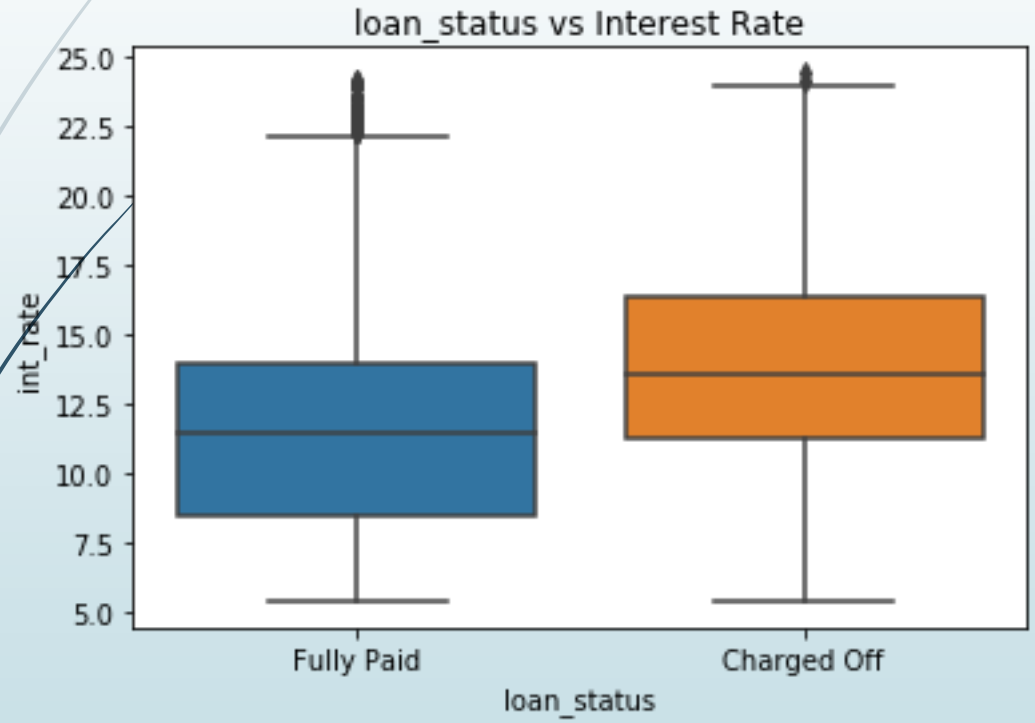
- Majority of the customers have taken loan on a 3 year term.

Analysis



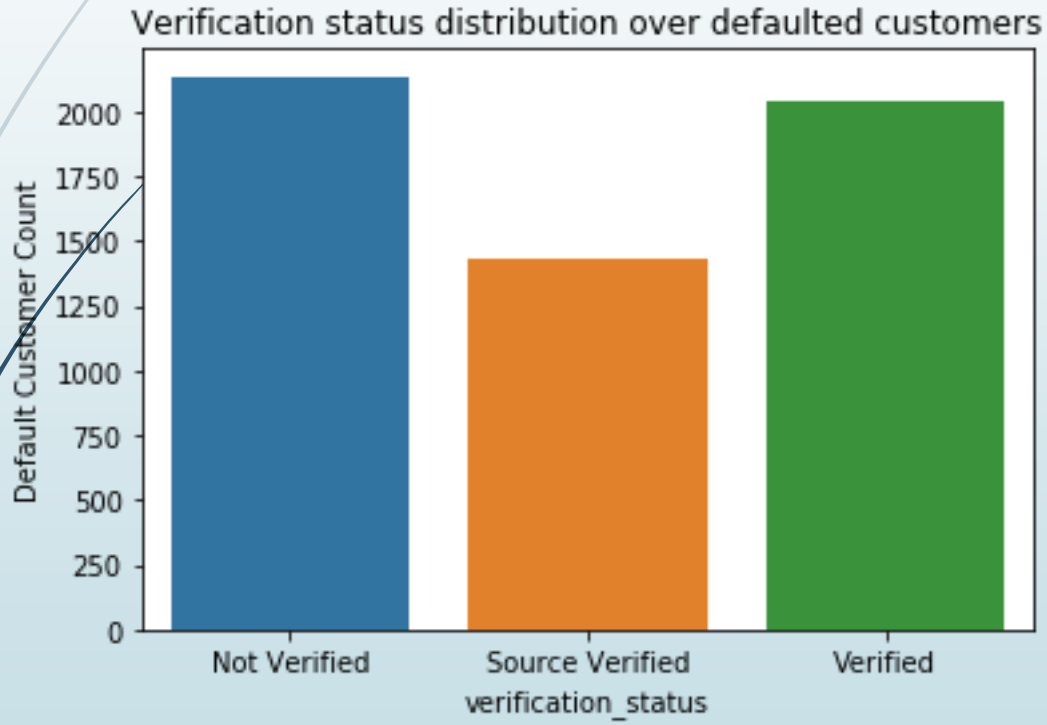
- The defaulted customers either have their home on mortgage or rent.

Analysis



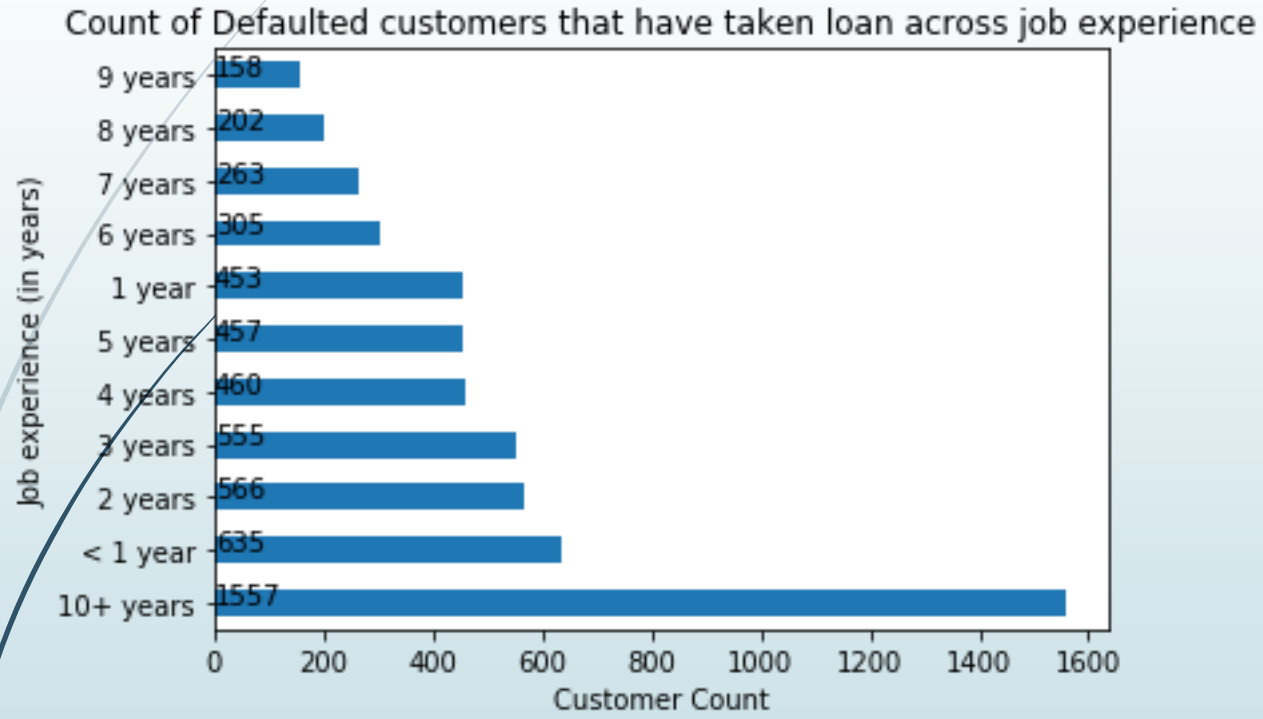
- The interest rate charged on defaulted customers is higher than the fully paid customers as they are riskier in nature.

Analysis



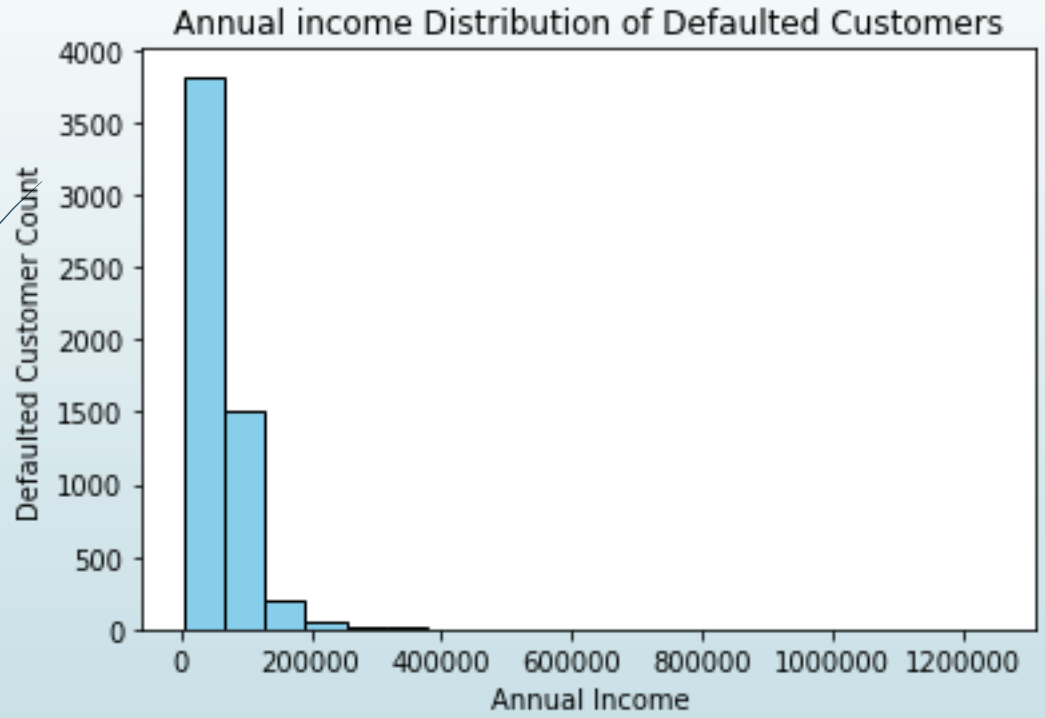
Majority of defaulted customers do not have their income or income source verified.

Analysis



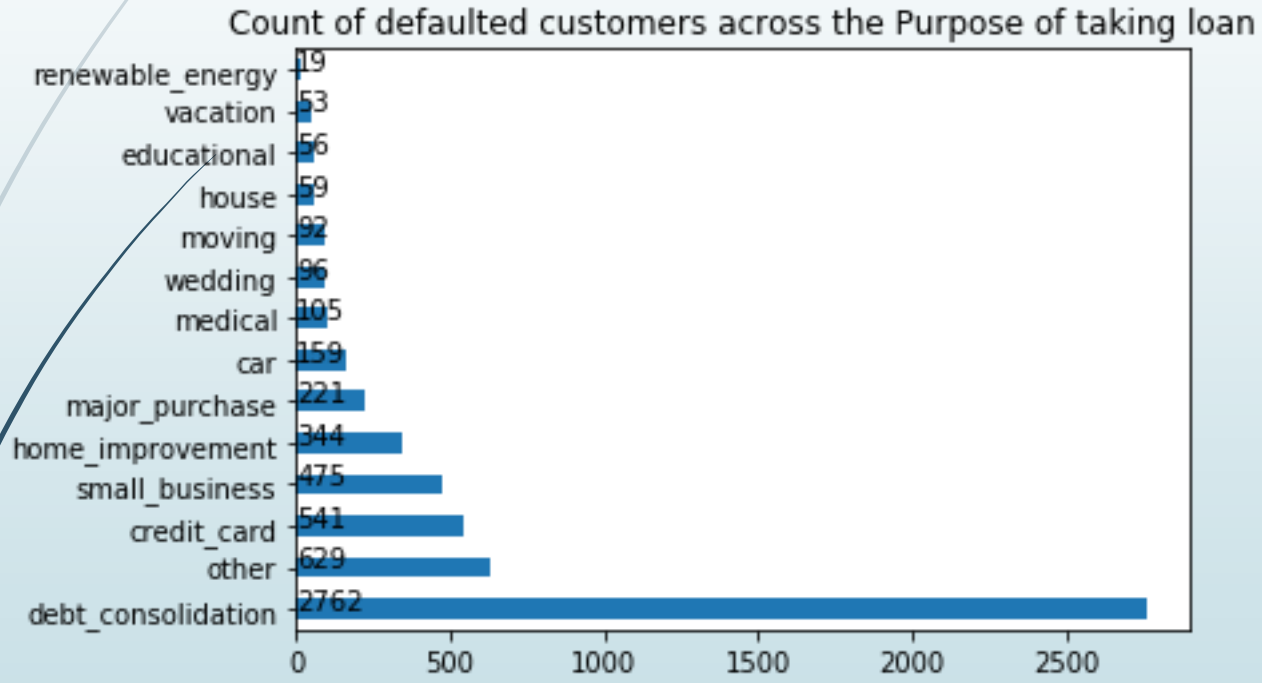
- Defaulted customers with job experience less than 10 years and greater than 1 year have the highest number of loans.
- This can be for personal purpose like home loan or such for customers with 10+ years of experience and for customers with <1 year of experience, it could be for any startup funding.

Analysis



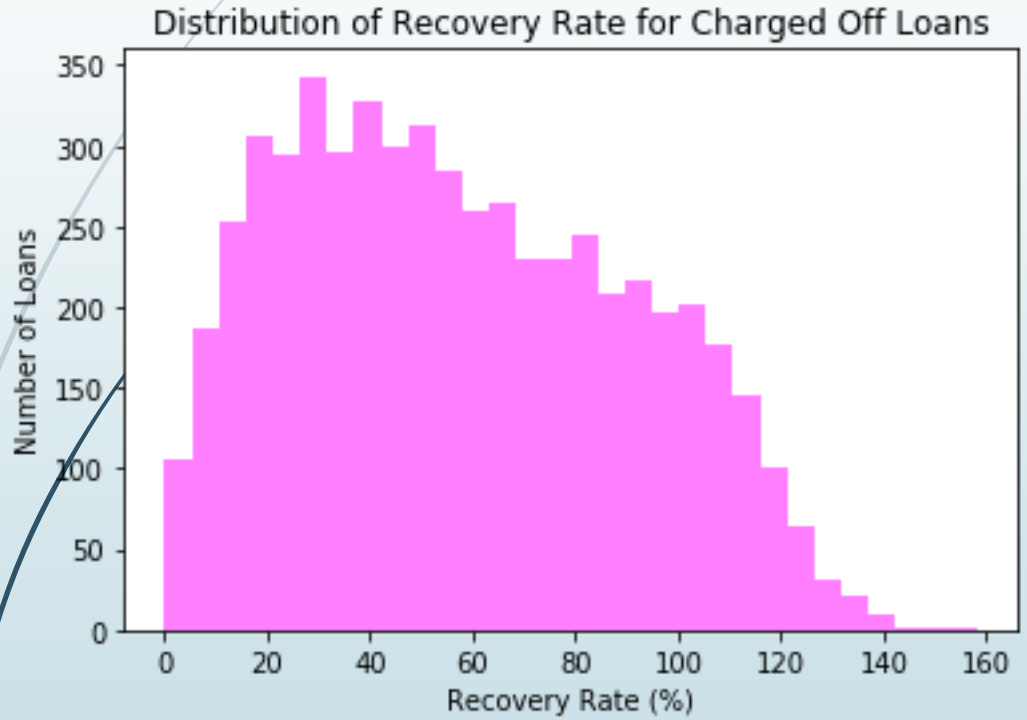
- Majority of the defaulted customers earn less than 200000.

Analysis



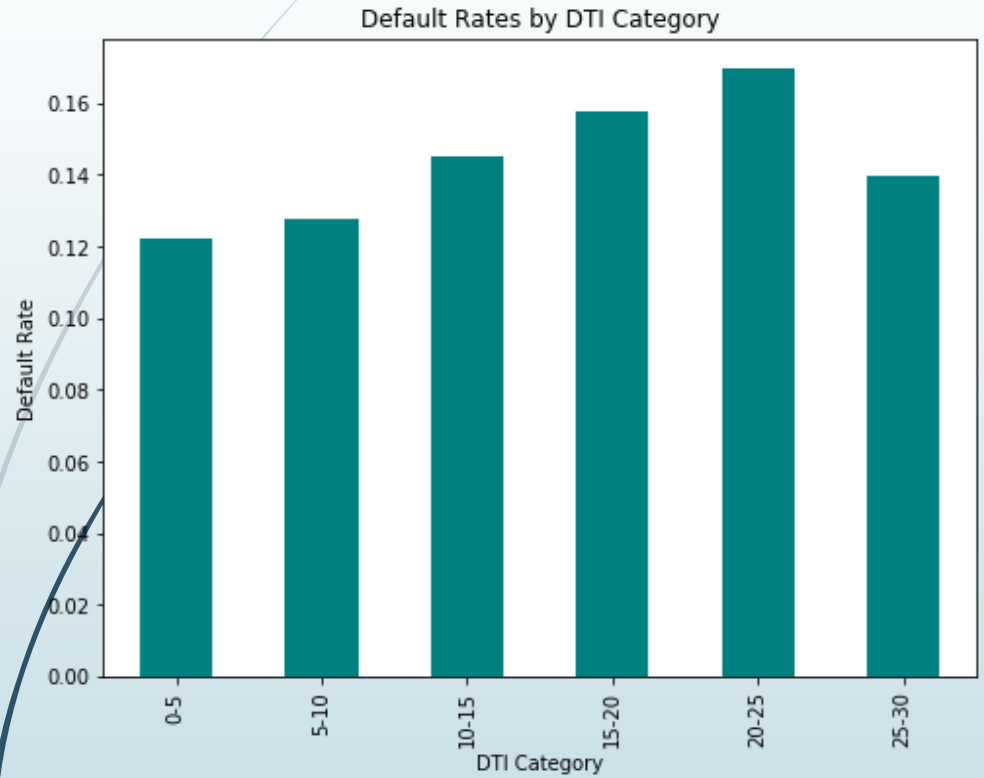
- About half of the defaulted customers take new loan to pay for older loans.

Analysis



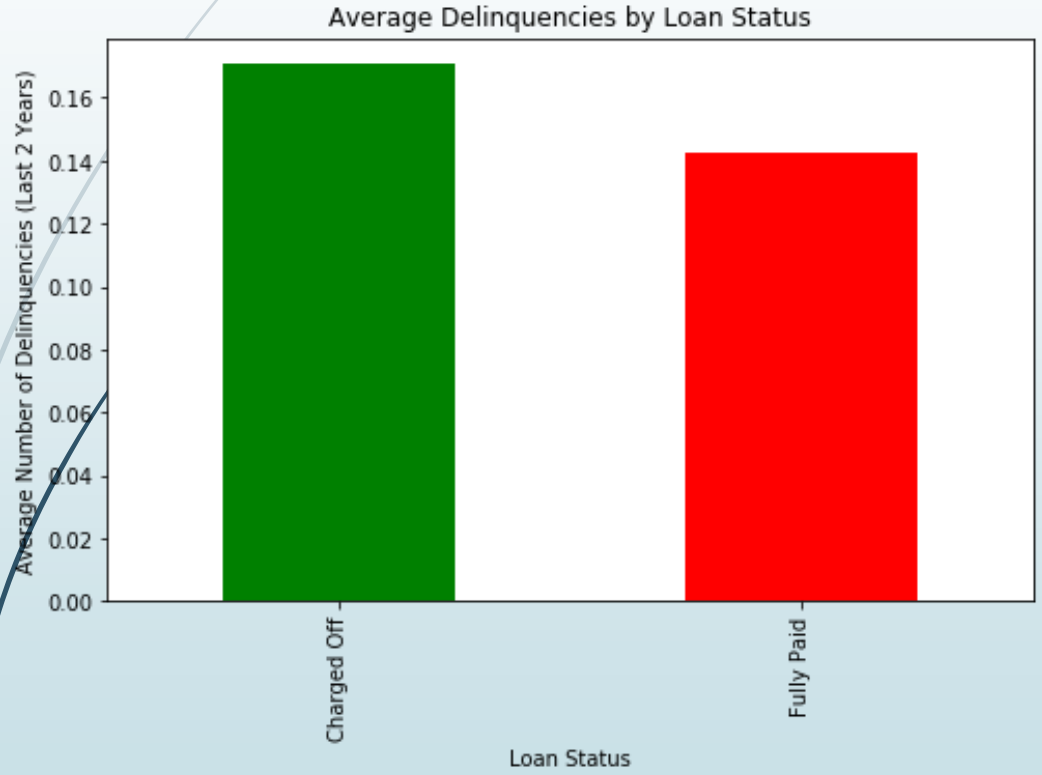
- Peak Recovery Rate: The histogram peaks around the 20%-40% range. This suggests that for most charged-off loans, lenders are recovering about 20%-40% of the original loan amount.
- Higher Recovery Rates: A smaller number of loans show recovery rates higher than 100%, which indicates that in some cases, the amount recovered exceeded the original loan amount. This could be due to additional fees or penalties collected during recovery.
- Skewed Distribution: The distribution is right-skewed, with most recovery rates falling on the lower end of the spectrum. This indicates that while some loans do recover close to or more than the loan amount, the majority recover significantly less.

Analysis



Higher DTI ratios may indicate a higher risk of default.

Analysis



There's a general upward trend in the proportion of charged-off loans as the number of delinquencies increases. This indicates a strong correlation between delinquencies and the likelihood of a loan defaulting.

Conclusion

- More stringent verification processes need to be placed in the system as majority of the defaulted customers do not have their income or income source verified.
- Customers who have their house on mortgage or are staying on rent should be charged higher rates of interest.
- Customers having job experience less than 1 year or above 10 years should be charged higher rates of interest as they are more prone to default.
- Customers who are taking loan for debt consolidation should be charged higher rate of interest.
- Higher DTI ratios suggest higher risk of defaulting so such customers should be recognized and thoroughly verified.
- Customers with higher number of delinquencies are at a higher risk of defaulting. So lenders should be careful while providing loan to such customers.