Lending Club Case Study

Members:-
Kumar Abhishek
Clara Regalado

Problem Statement & Dataset

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 company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

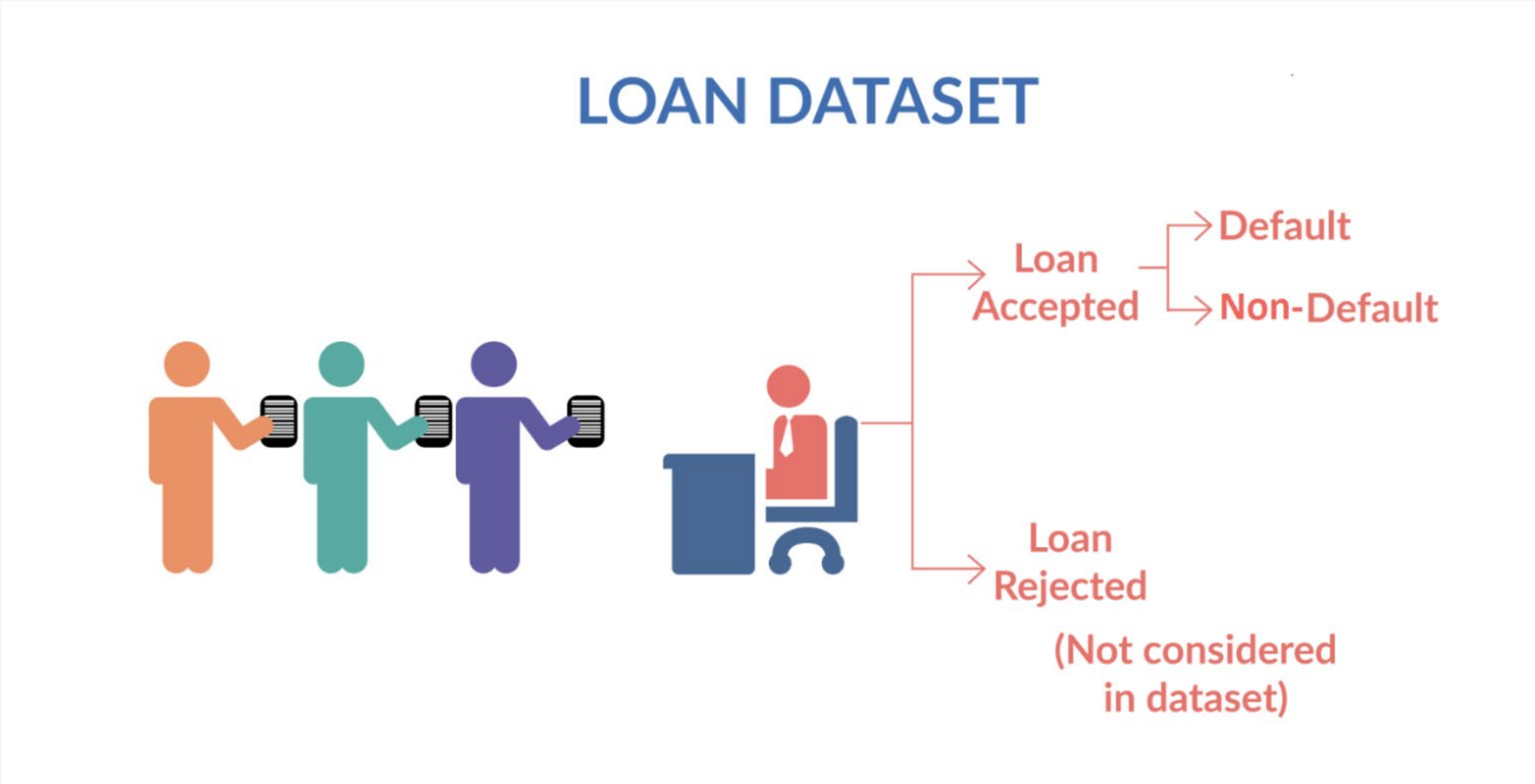


Figure 1. Loan Data Set

Business Understanding

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    - Fully paid:
    - Current:
    - Charged-off:

2. **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.).

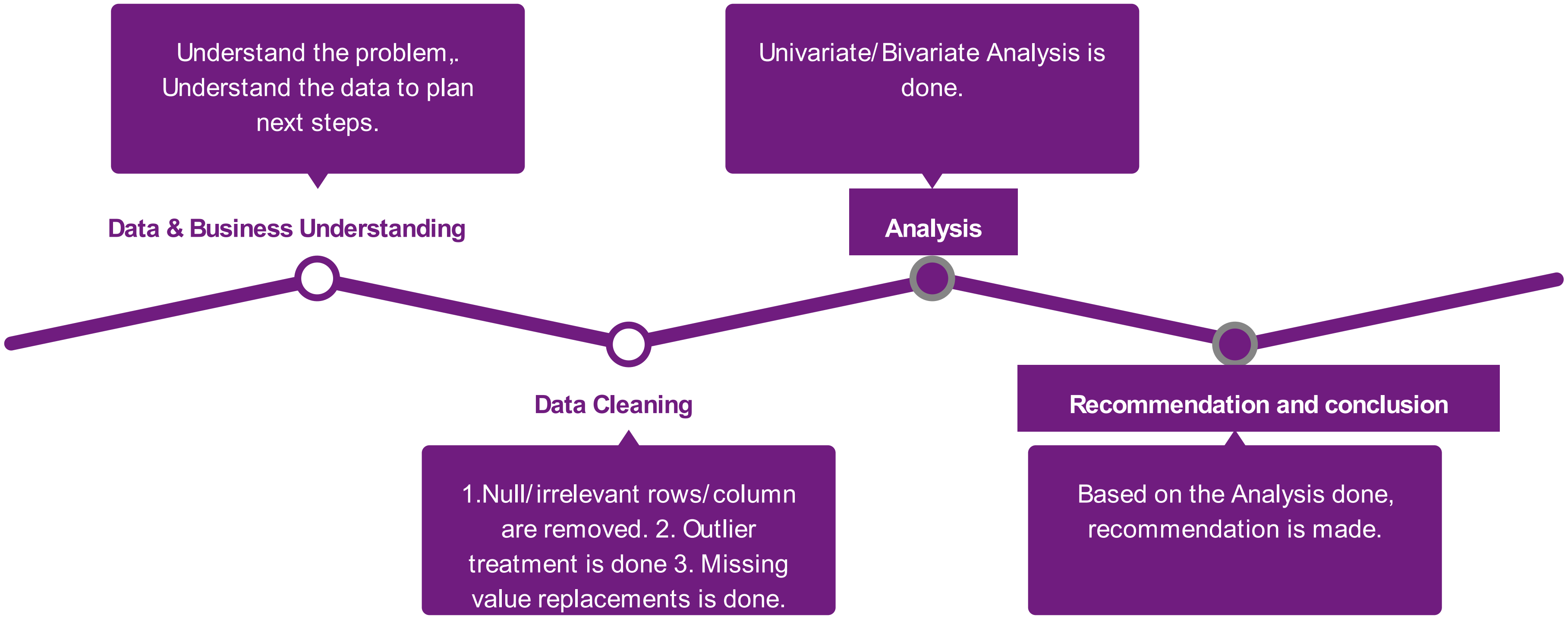
Type of Risks

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is **likely to repay the loan**, then **not approving the loan** results in a **loss of business** to the company

2. 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**

Methodology



Understanding the Data

After we understood the problem statement, It was important to understand the data.

1. We went through the each column of data_dictionary.xlsx to understand each feature variable.

2. This helped us to understand what to expect from data for the case study.

3.Once this was done, next step was data cleaning.

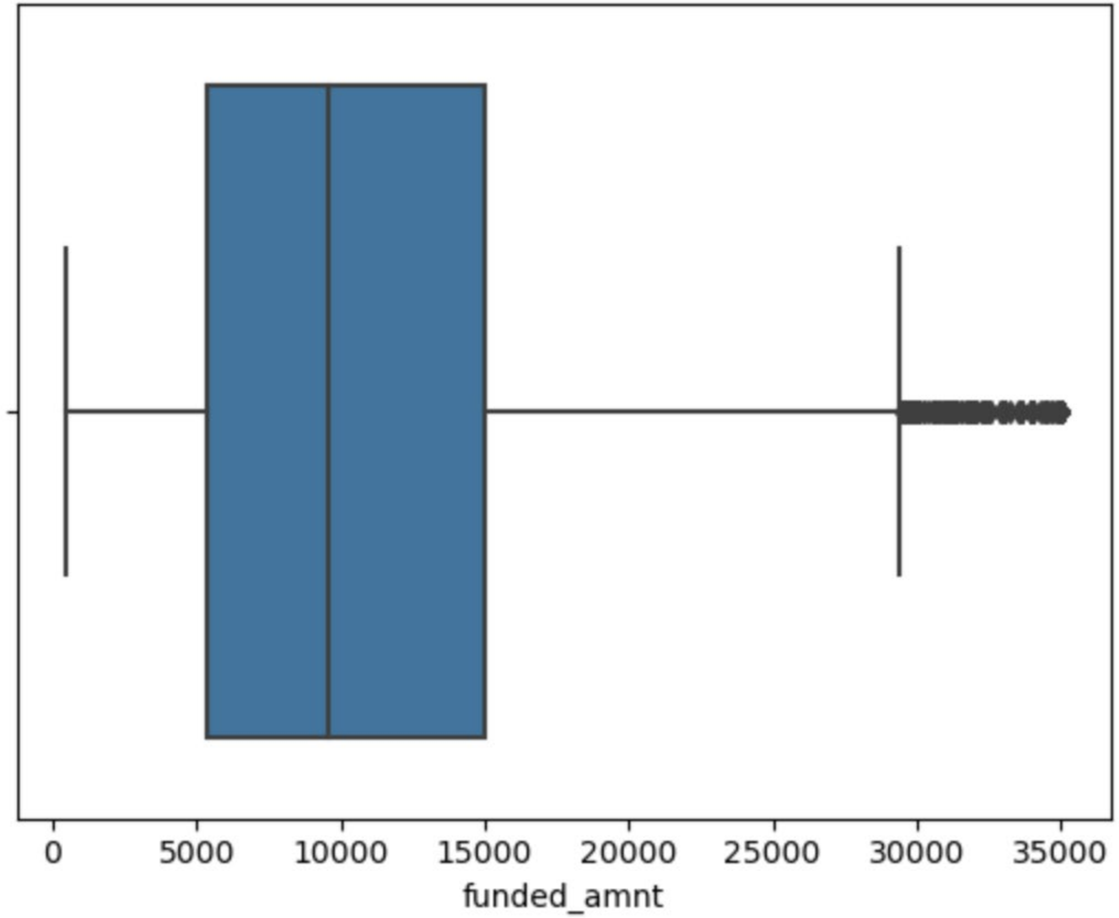
Data Cleaning

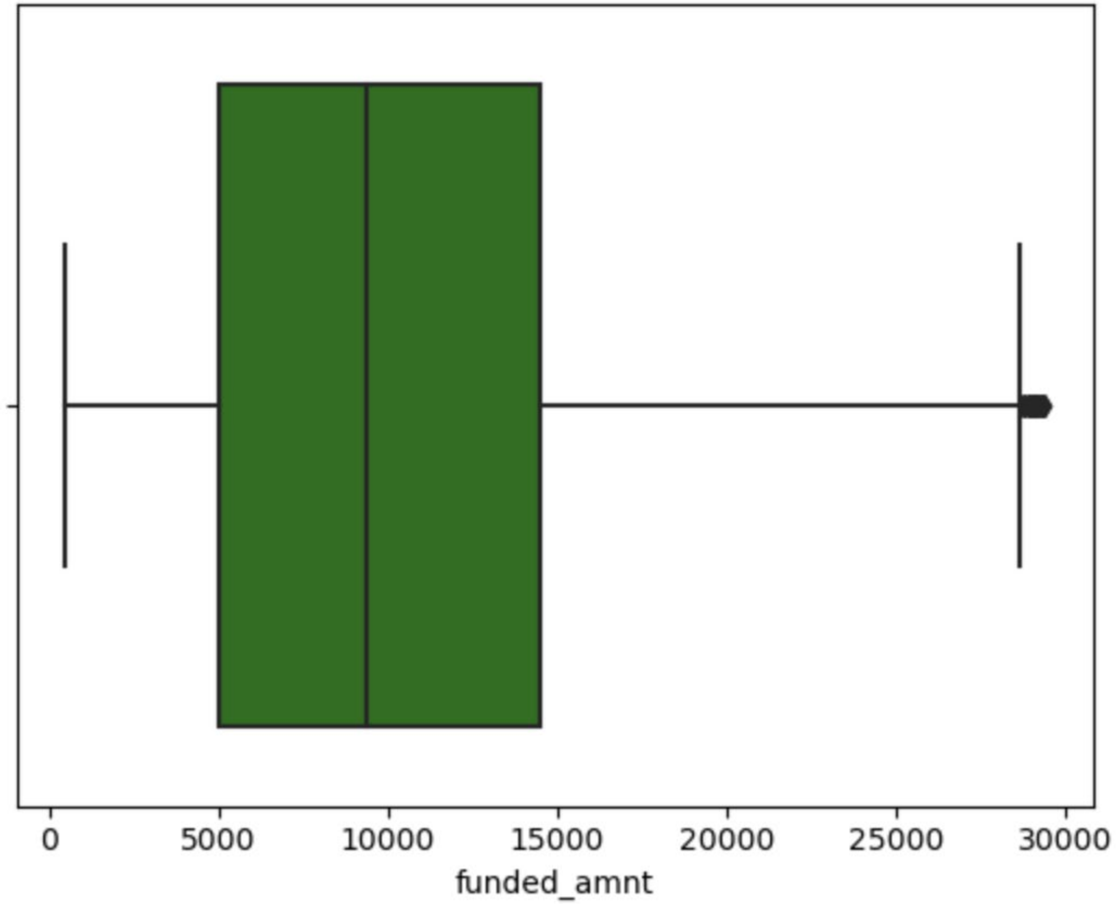
1.Dropping Column (More than 40% null values for column was dropped)
2.All same value columns were dropped.
3.All different value columns like id/url were dropped.
4.Columns which was related to post-approval eg revol_bal were dropped
5.Columns with was not relevant to problem statement like description were dropped.
6. Rows with more than 60% null values were dropped
7.Outlier Treatment => **values < 25th percentile - 1.5 * IQR and > 75th %ile + 1.5IQR** were removed

Example of outlier Treatment(funded_amnt)

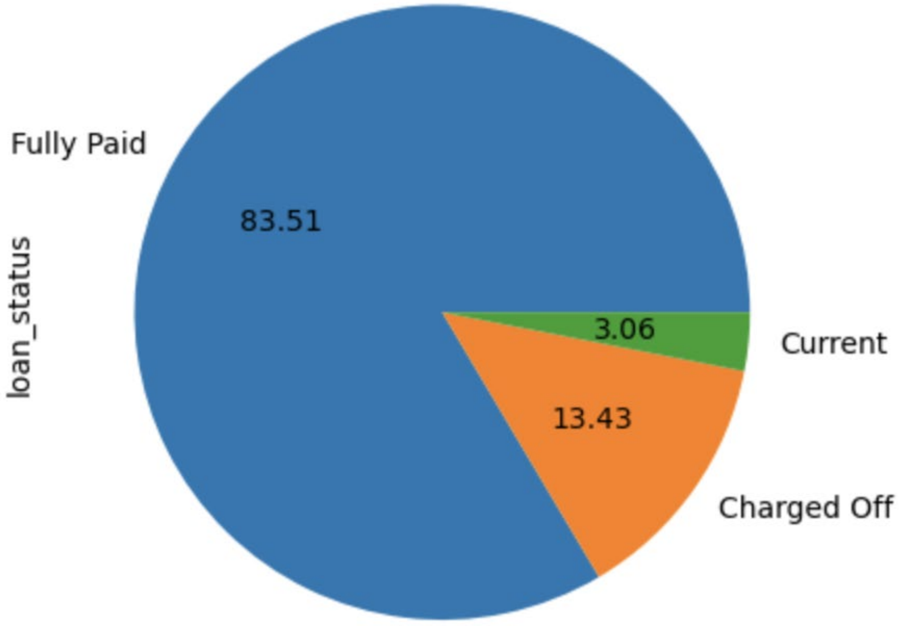
Before

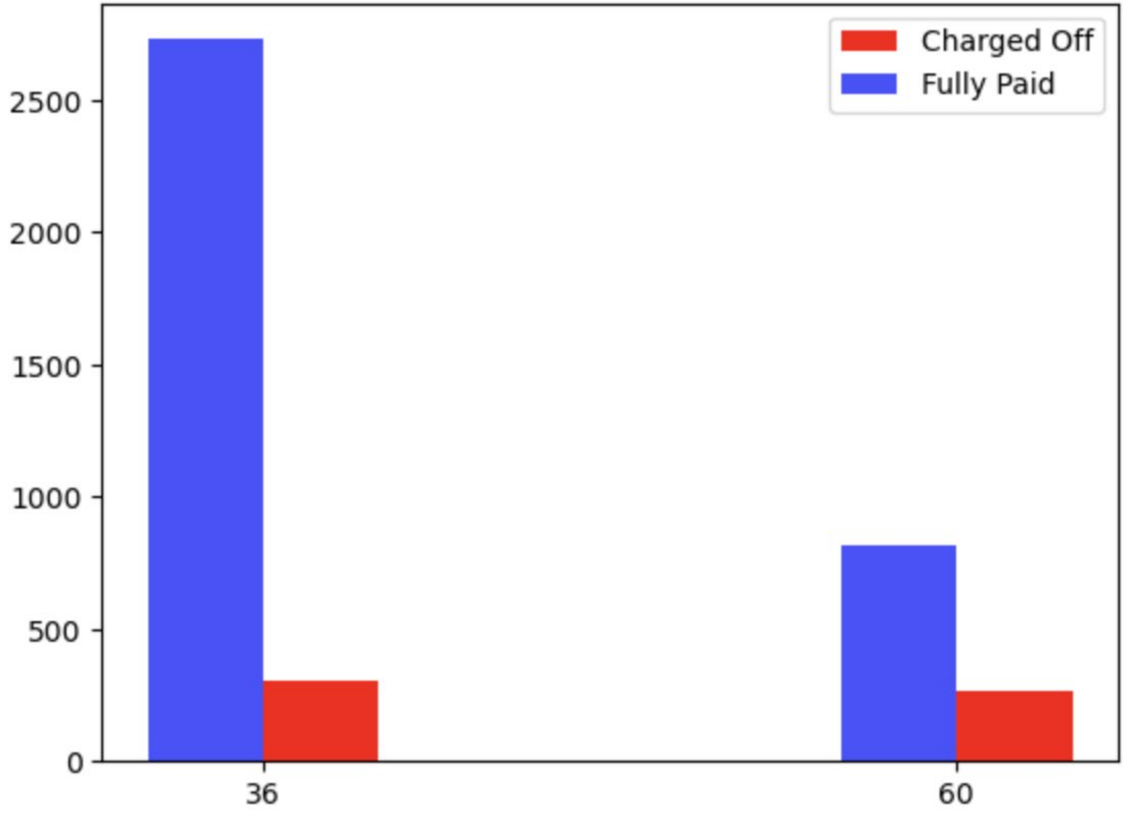
After





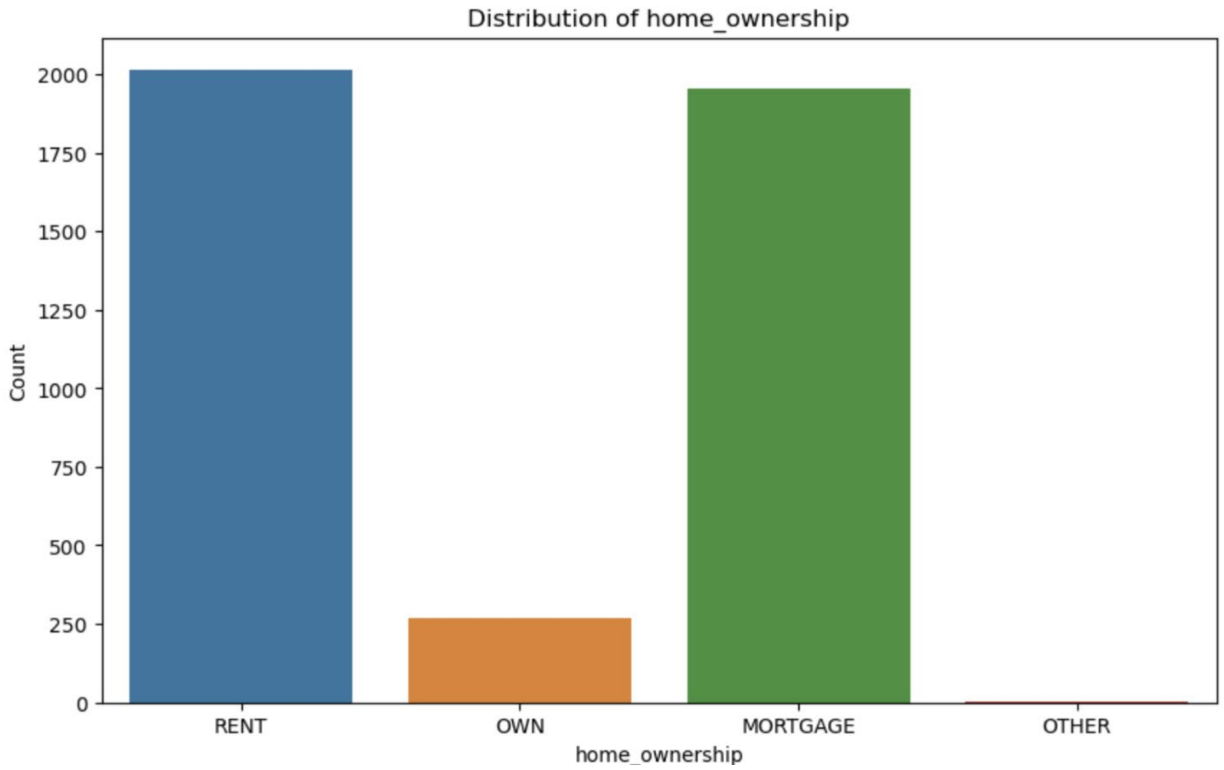
Univariate Analysis

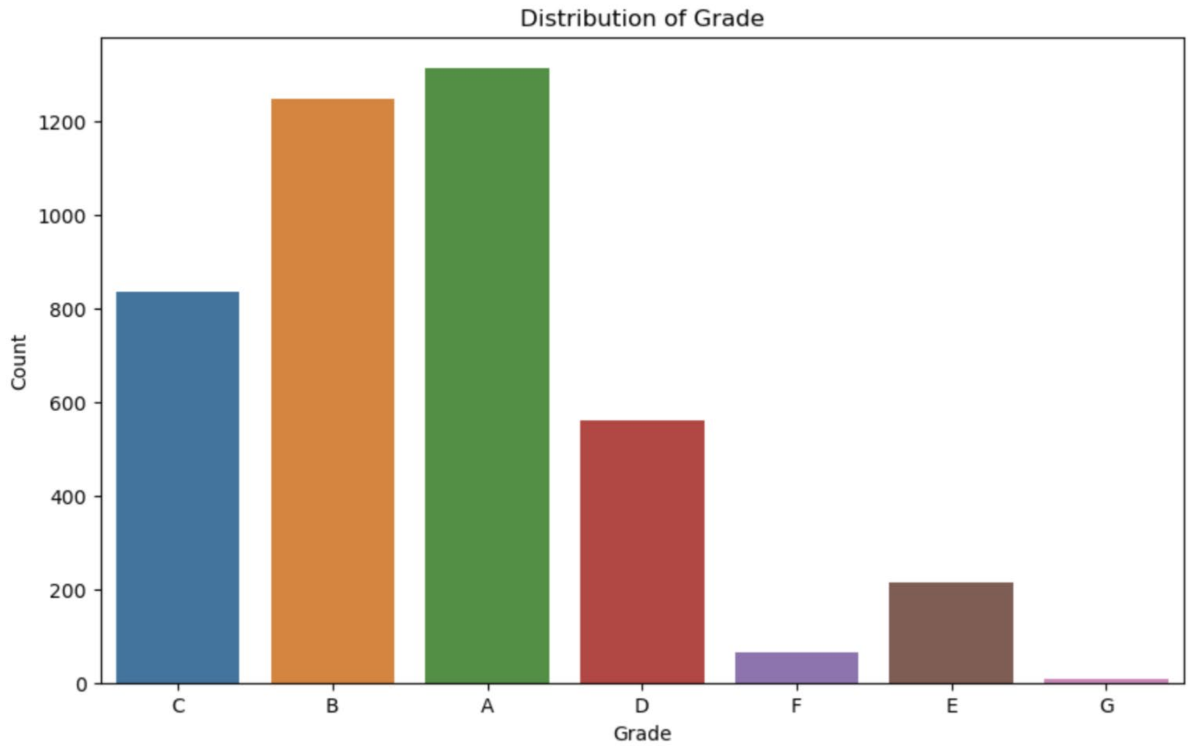




- A total of ~13.6% customers in total are defaulters
- 73% customers went for shorter term loan of 36 months term while 27% went for 60 months term loan
- More defaulters with 60 months term loan as they took bigger loan and had hard time returning it

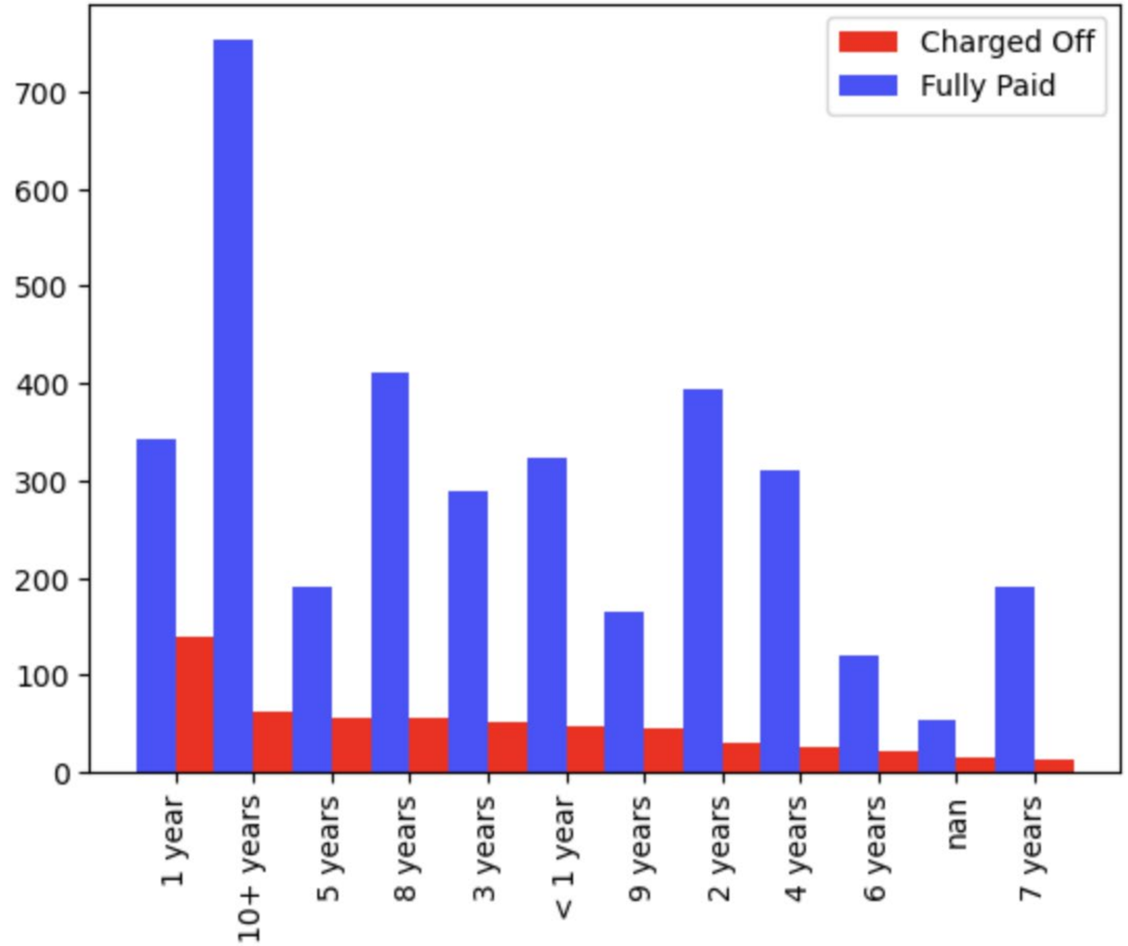
Univariate Analysis

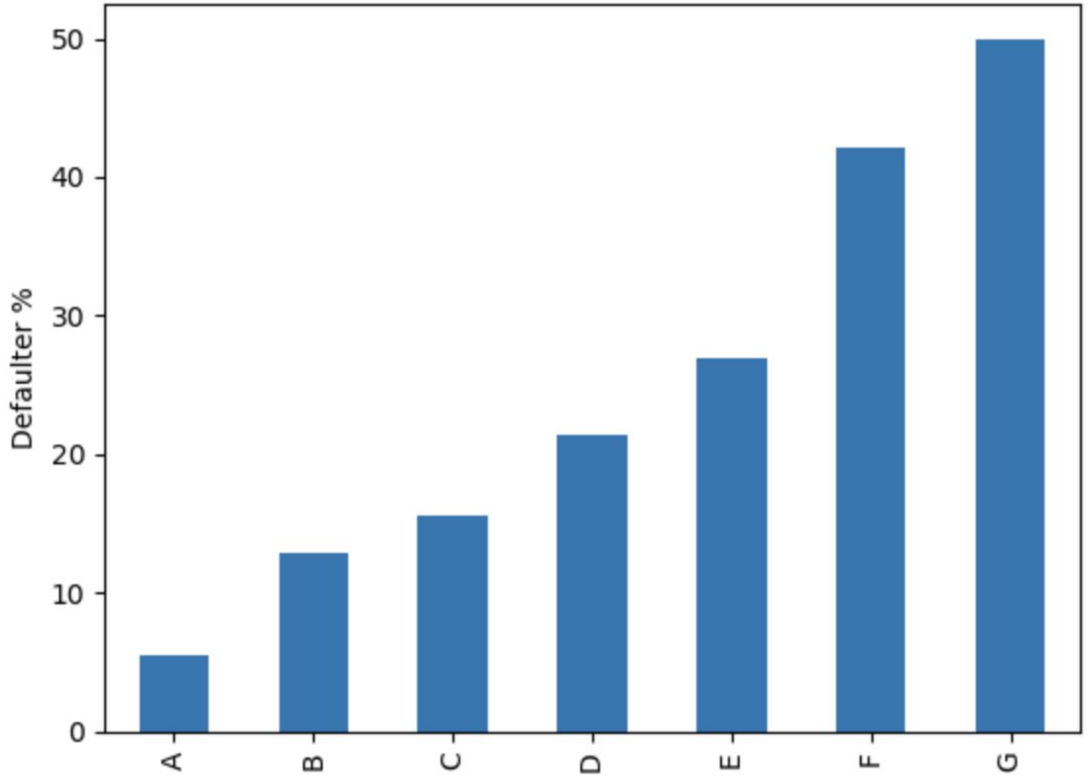




- The majority of borrowers don’t have their own house.
- Grades are an important attribute to tell the probability of defaulting the loan.
- Trend shows A,B.C loan grades had the highest distribution

Segmented Univariate Analysis

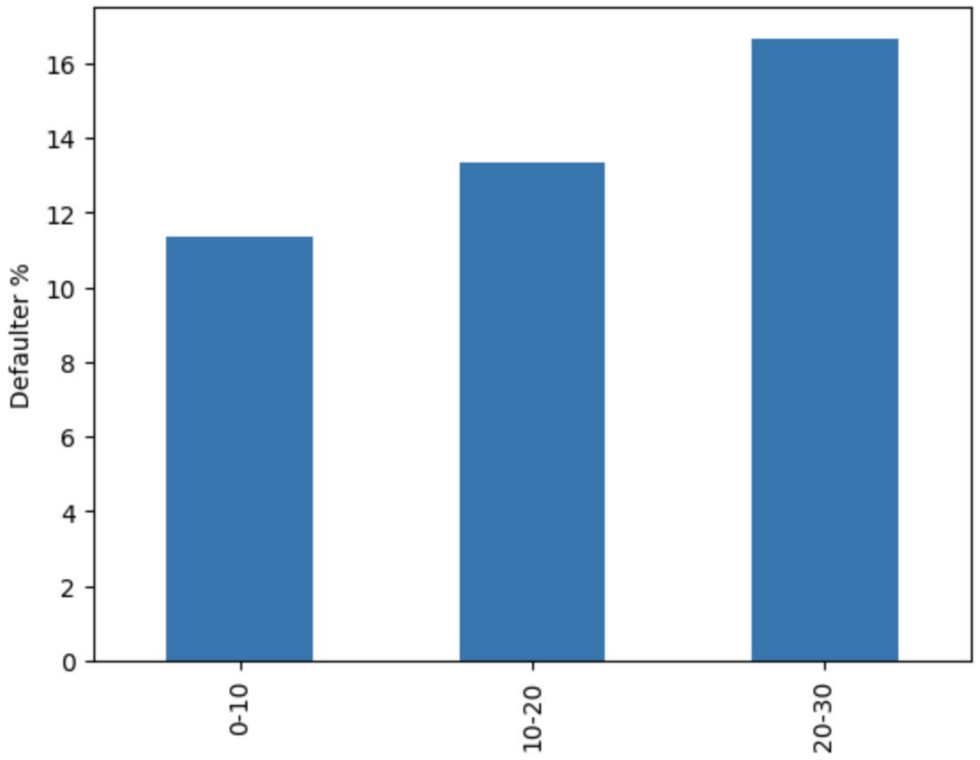


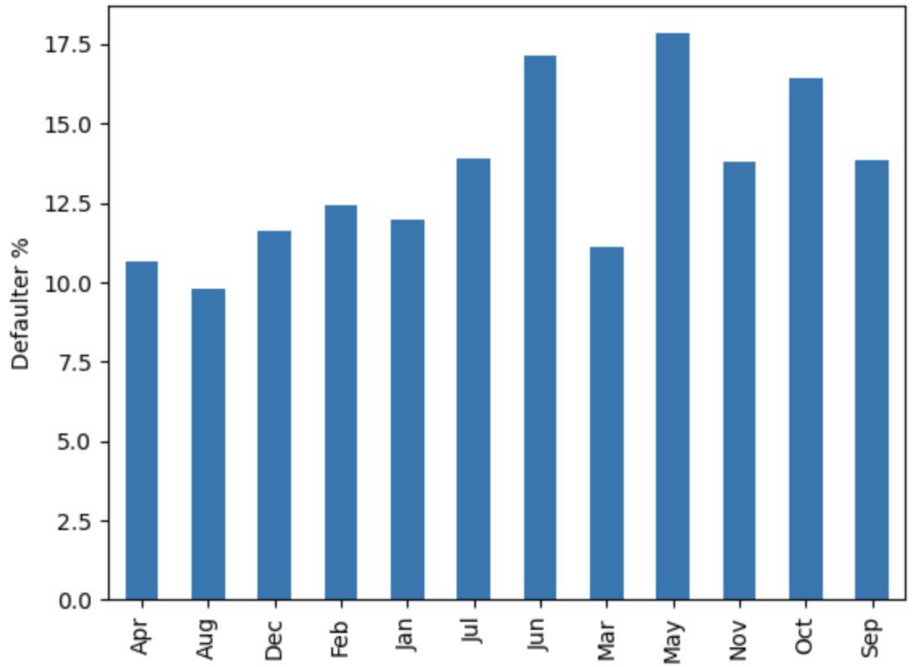


Majority of people with 10 + years employment status has paid their loan on time

Loan grade 'G' loans has 50% chance of defaulting while Loan Grade 'A' loans has just 5%

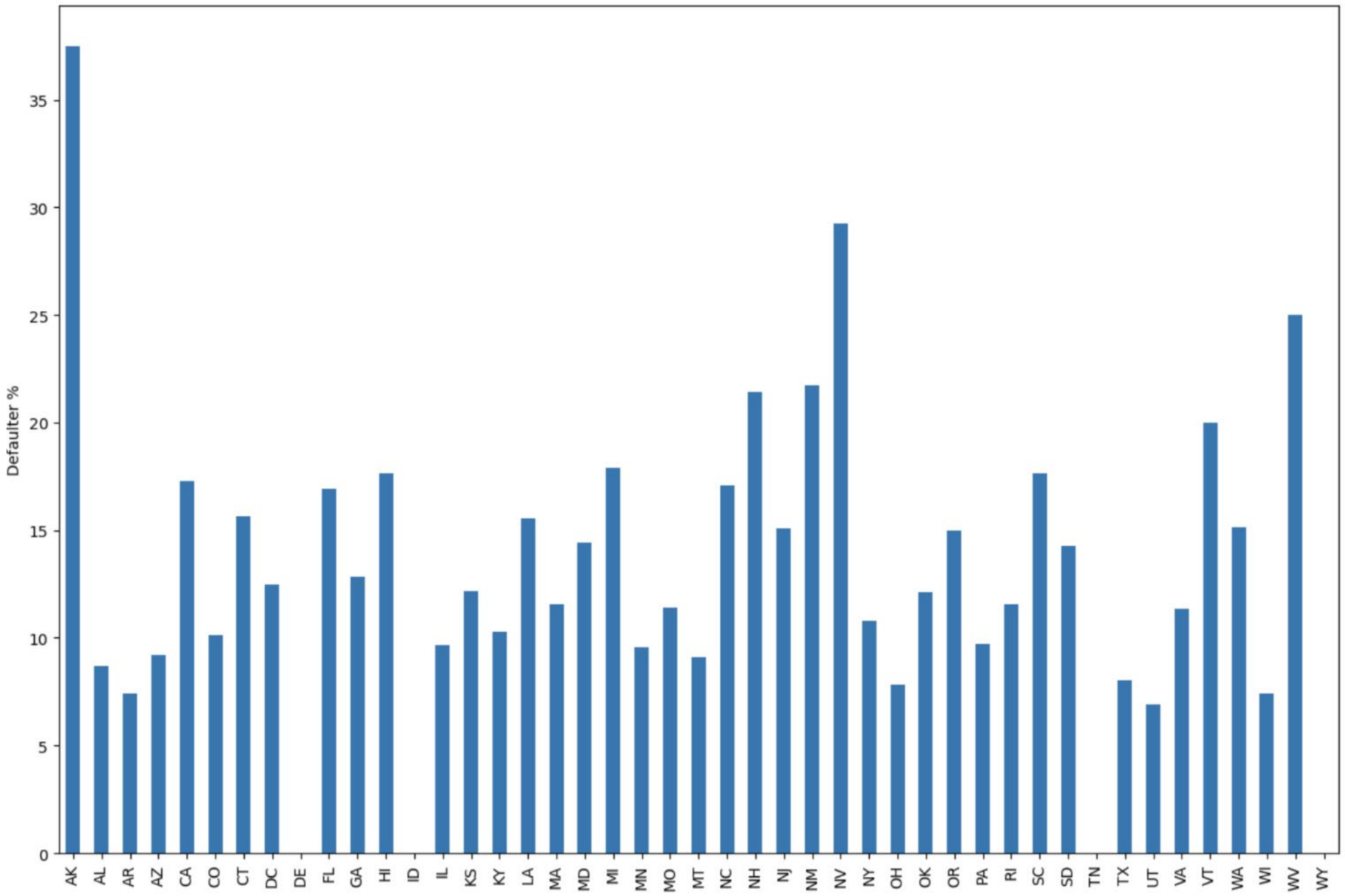
Segmented Univariate Analysis





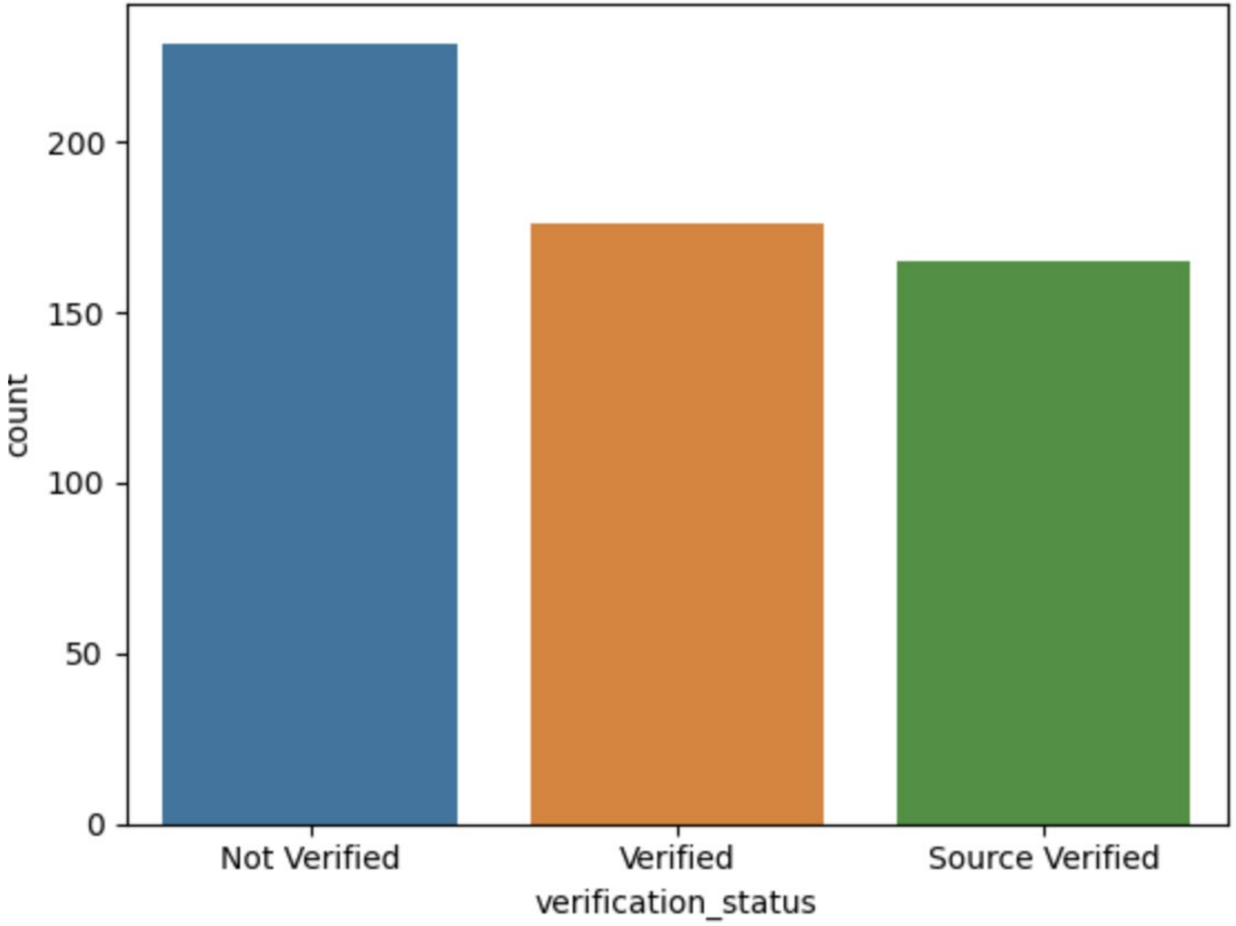
- Higher dti means higher risk to default. dti with value 20-30 has more 16% risk to default
- Loans issued in May, June and october interestingly has higher defaulters

Segmented Univariate Analysis



- Alaska has the highest default % which makes sense as Alaska has lower job opportunities
- DE,WY,TN,ID are clean states with no/low defaulters

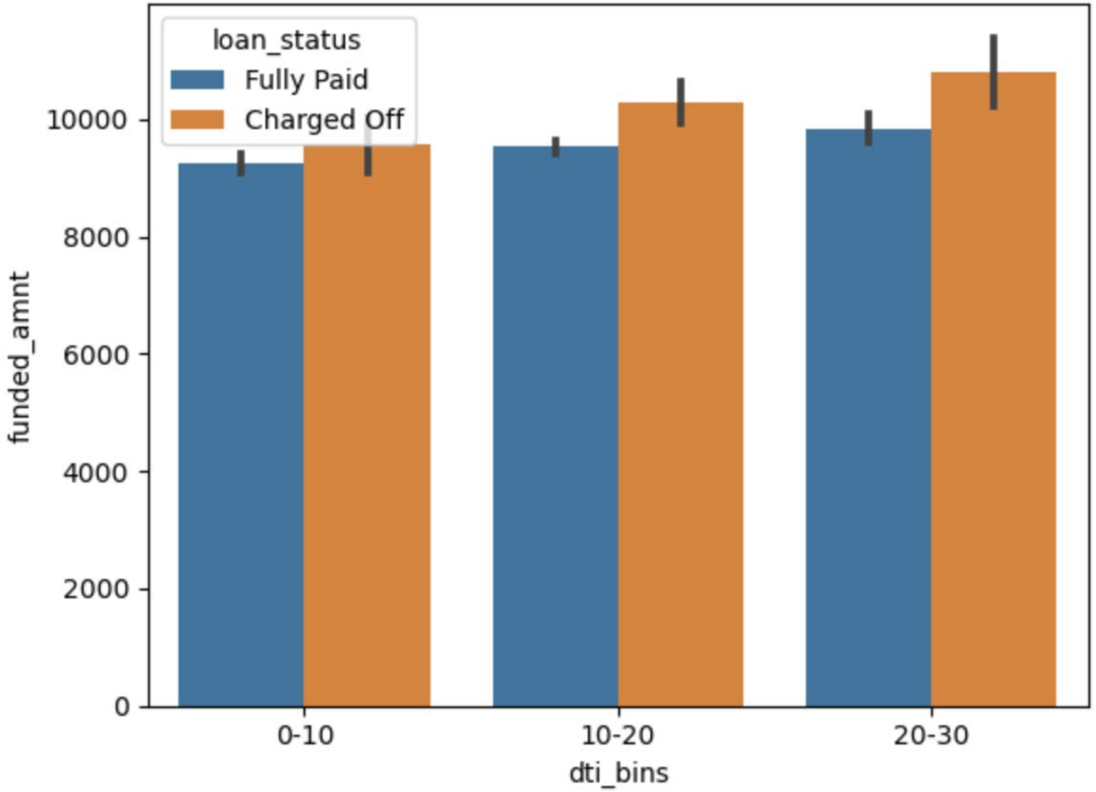
Segmented Univariate Analysis

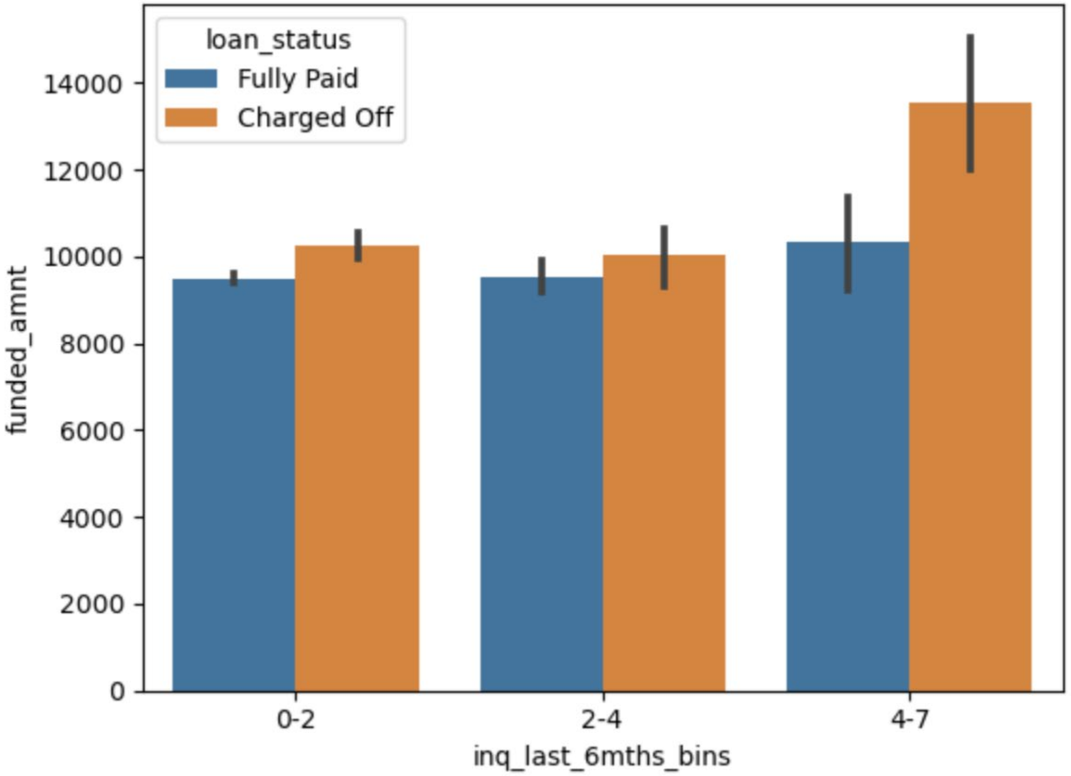


- Count of Defaulters were more unverified compared to verified/source verified

Bivariate Analysis

UpGrad

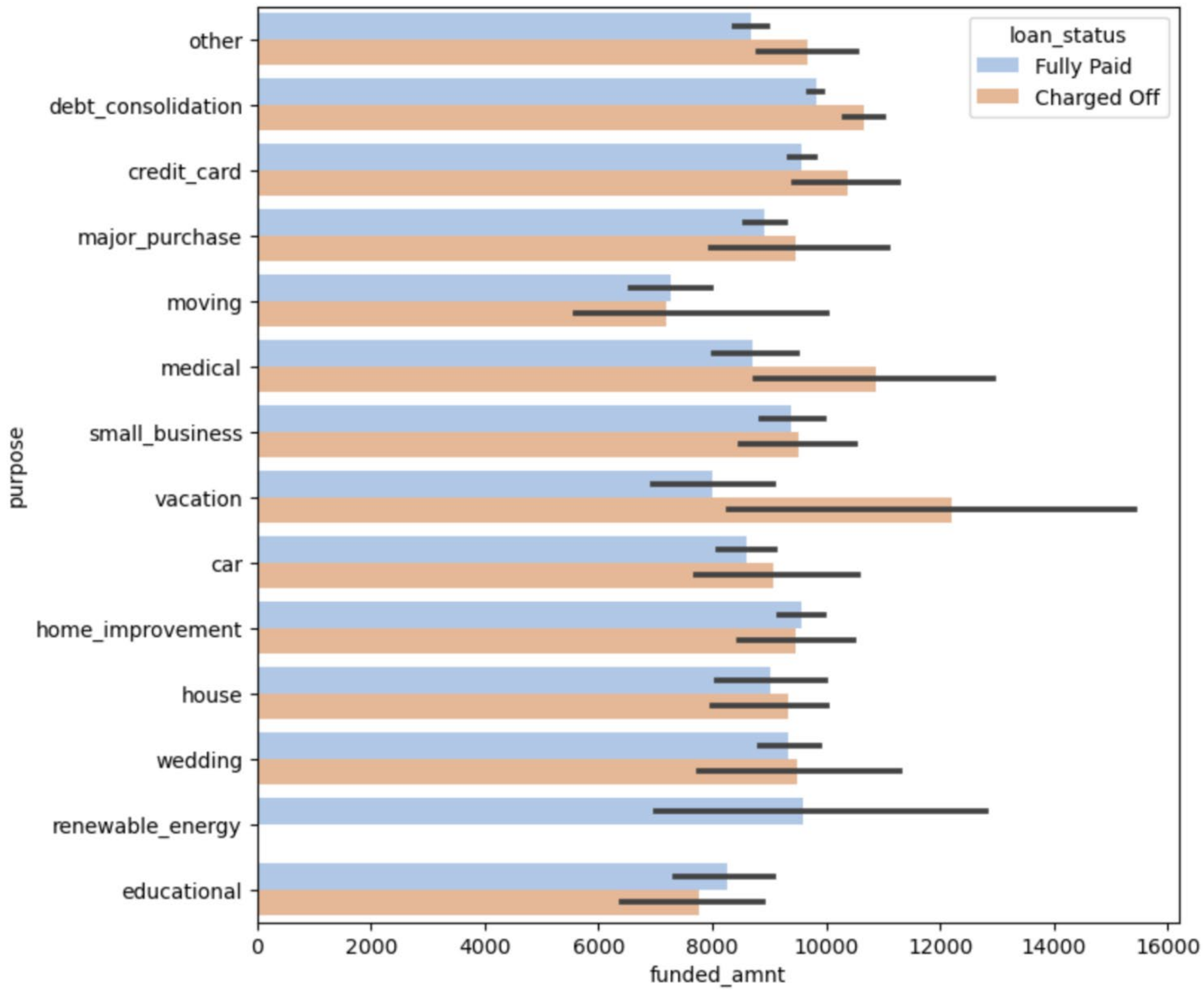




- Higher DTI and funded amount greater than 10k led to more defaulters
- 4-7 inquiries in last 6 months and higher loan/funded amount(> 12k) leading to more defaulters

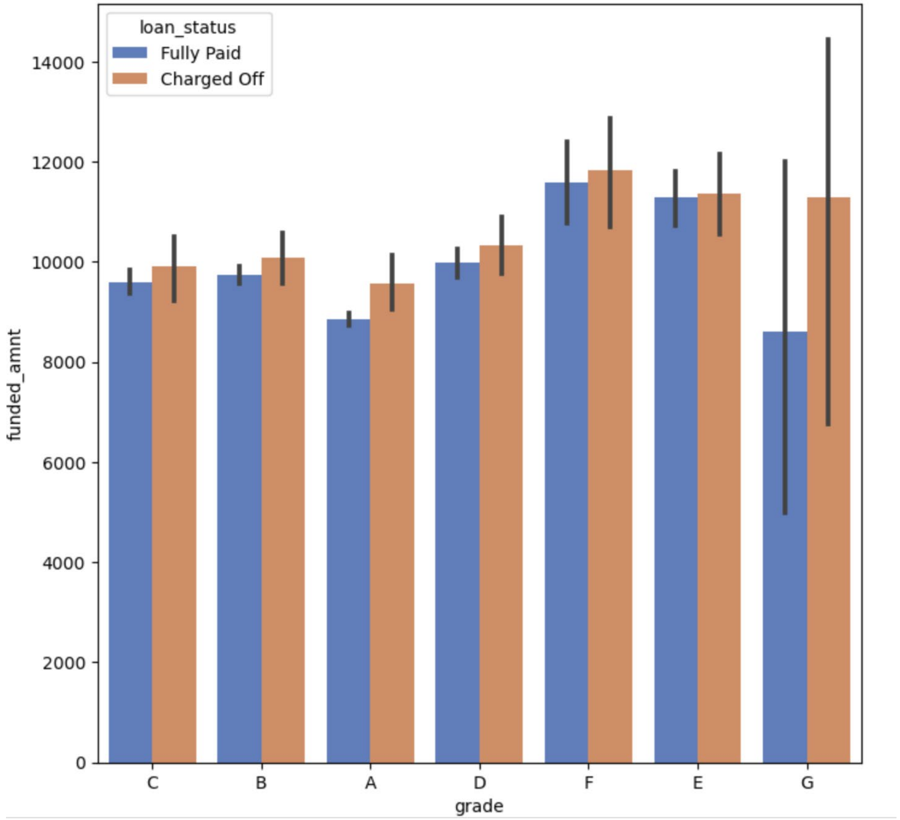
Bivariate Analysis

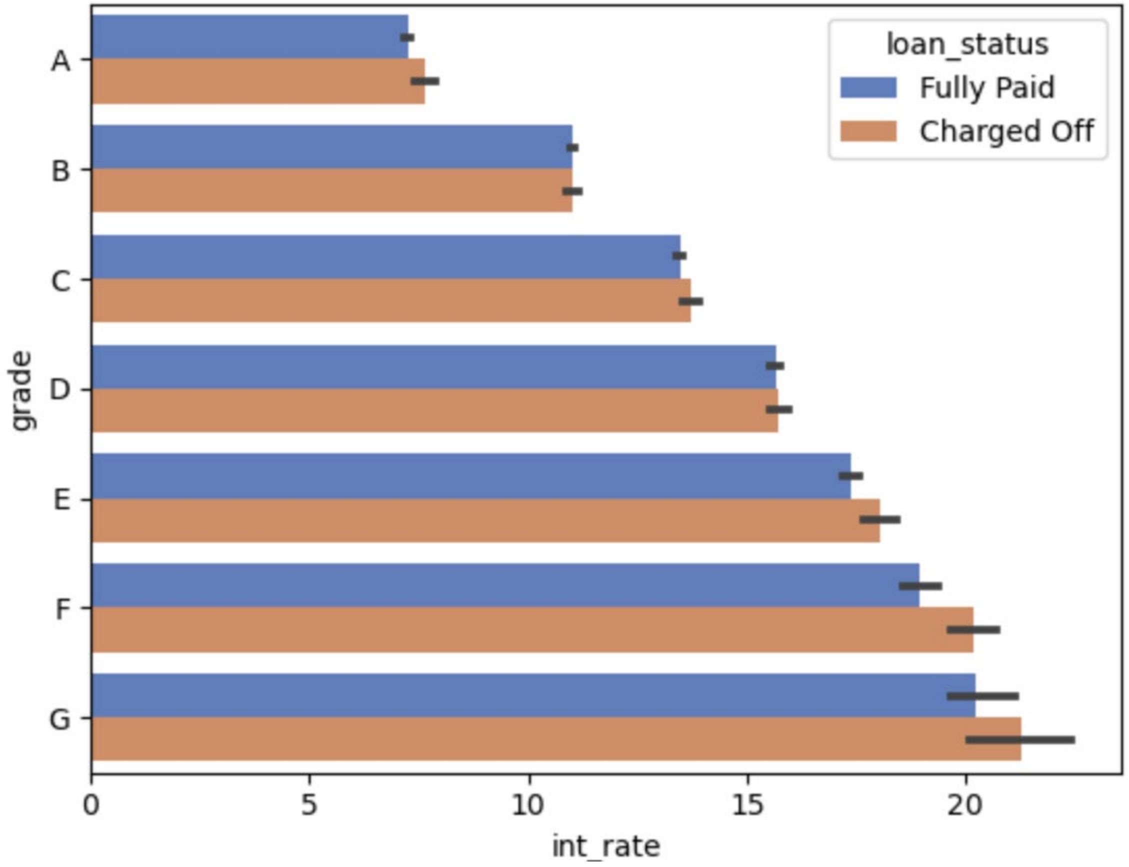
UpGrad



- Loan/Funded amount > 12k and pupose of loan is 'vacation' is more likely to default

Bivariate Analysis

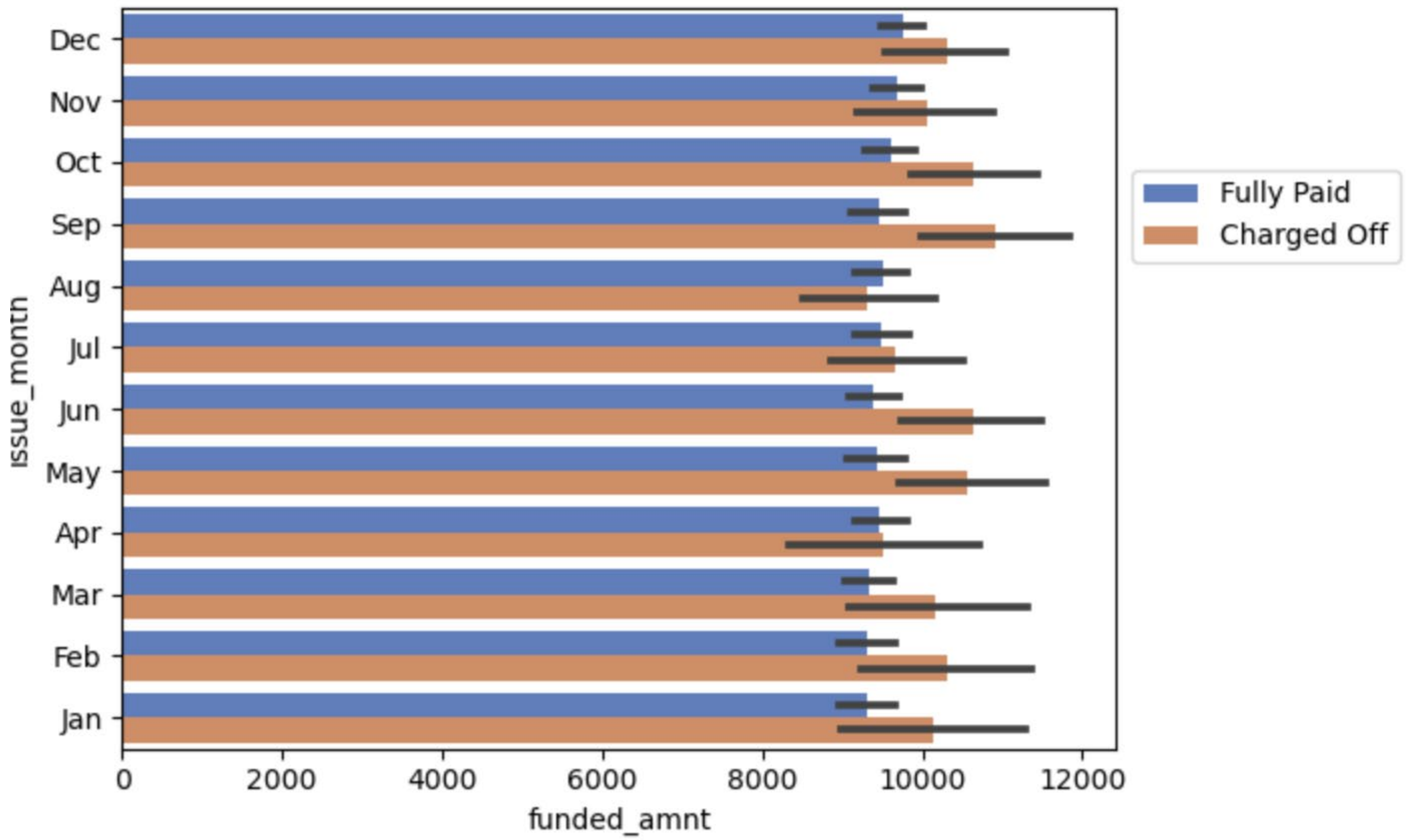




Grade 'G' and loan >10k is more likely to default

Grade 'G' and interest rate >20 is more likely to default

Bivariate Analysis



May/june/sep/oct with 10k+ funded or loan amount is more likely to default

Summary & Insights

Univariate Analysis

- A total of ~13.6% customers in total are defaulters
- 73% customers went for shorter term loan of 36months term while 27% went for 60 months term loan
- Interest rate is more dense between 10-15%
- 10+ years experienced people are more in number who needs loan
- Majority of borrowers don't have their own house
- Majority of borrowers are not verified
- Majority of borrowers income is less than 70k
- Most of the borrowers didn't have public bankruptcies record
- Most loans were issued in Q4

Summary & Insights

Segmented Univariate Analysis

- Most of the loans were taken with the purpose of "Debt Consolidation"
- The chances to default for 60 months term loan is higher compared to 36 months loan.
- People with lower salary is more likely to default. Trend shows people with salary between 40k-70k has around 15% chance to default while the one in higher income group of 70k-85k has 11% chance
- 10+ years experience people fully paid the loan on time.
- People on rent and mortgage are more likely to default when compared to the one with own house
- 20% of people with at least 1 public record bankruptcies couldn't pay the loan while 13% with no public record bankruptcies ended up defaulting
- Trend shows people with unverified income source are more likely to default.
- Loan grade 'G' loans has 50% chance of defaulting while Loan Grade 'A' loans has just 5%
- Higher dti means higher risk to default. dti with value 20-30 has more 16% risk to default¶
- If the enquiry in last 6 months is higher than 2 is more likely to default
- Loans issued in May, June and october interestingly has higher defaulters
- Alaska has the highest default % which makes sense as Alaska has lower job opportunities

Summary & Insights

Bivariate Analysis

- Funded amount was higher for people who defaulted in every income group
- Higher DTI and funded amount greater than 10k led to more defaulters
- 4-7 inquiries in last 6 months and higher loan/funded amount(> 12k) leading to more defaulters
- Customers with interest rate of 20% + and loan/funded amount of 10k+ has more chance to default.
- Loan/Funded amount > 12k and pupose of loan is 'vacation' is more likely to default
- Grade 'G' and loan >10k is more likely to default
- Grade 'G' and interest rate >20 is more likely to default
- DTI with 20-30 and int_rate 12-14% are more likely to default
- 4 years employment length with 10k+ loan/funded amount is more likely to default
- May/june/sep/oct with 10k+ funded or loan amount is more likely to default

Recommendations & conclusion

Major factors/features that can be used to predict chance of defaulting.

- Verification Status
- home_ownership
- Grades
- Annual income
- DTI
- Employment length
- addr_state

Model

- This is a classic case of logistic regression. It's a binary classification problem, A model can be developed using a sigmoid function to predict the probability. A probability > 0.5 can be predicted as defaulter while probabilty < 0.5 can be predicted as non-defaulter.
- Alternatively, a neural network can be with RELU activation functional on the hidden layer and and a sigmoid on the outer layer