EDA CASE STUDY

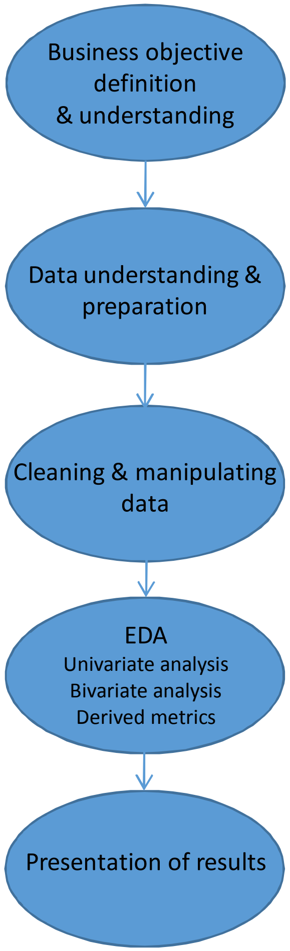
SUBMISSION

Sunil Appanaboyina

Abstract

- The aim of this study is to identify patterns which indicate if a person is likely to default his/her loan.
- The insights gained can 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.
- The objective is to use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.
- By identifying the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- EDA done in this study
  - Univariate analysis
  - Bivariate analysis
  - Derived metrics

Problem Solving Methodology



Business objective definition
& understanding

Data understanding & preparation

Cleaning & manipulating data

EDA
Univariate analysis
Bivariate analysis
Derived metrics

Presentation of results

Data Understanding & Preparation

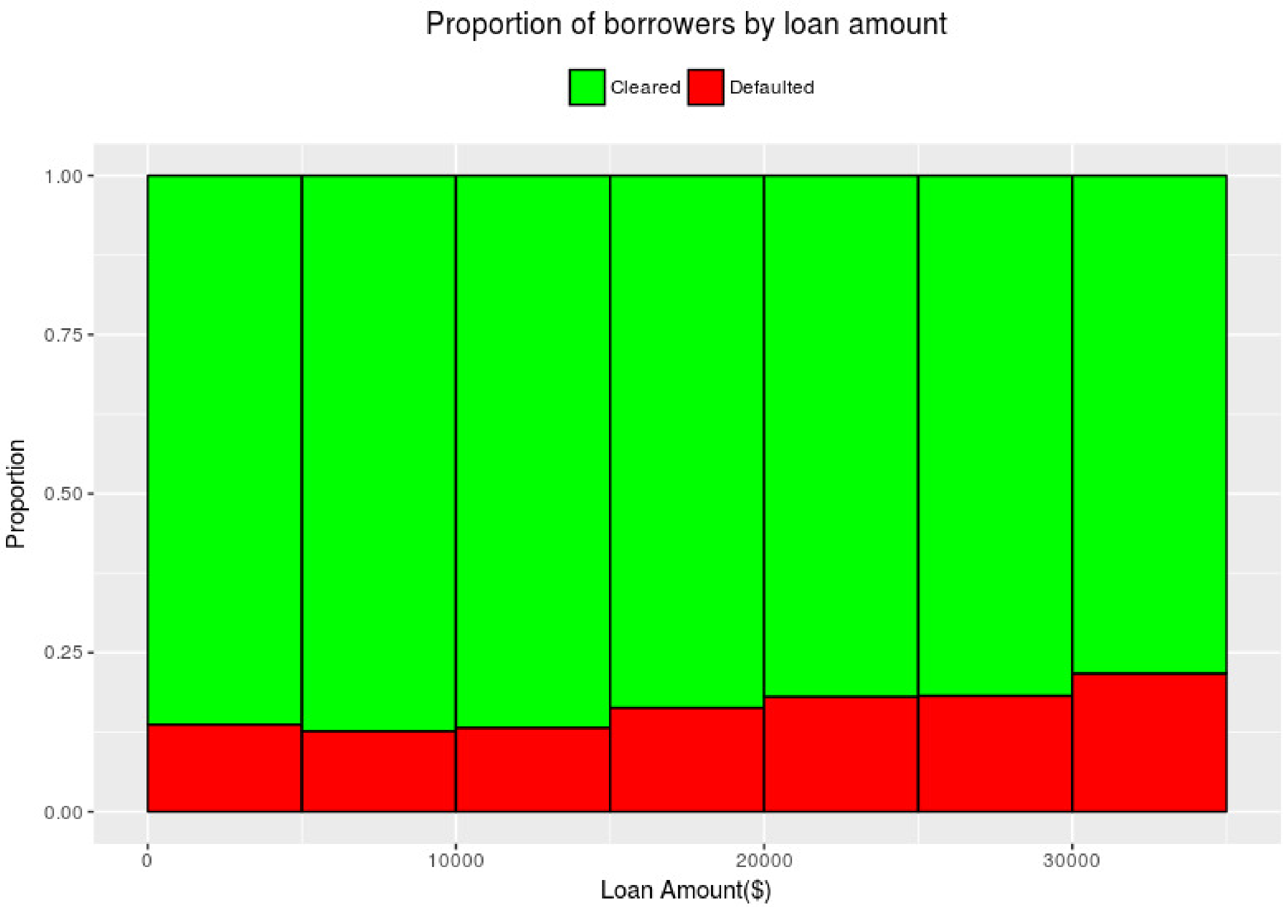
- Removed 54 columns which had all NA's
- Removed 6 columns which had same value throught the column
- Columns with missing values were identified and were dealt when the analysis was done
- Outlier values were also identified and required steps were taken during analysis
- From the remaining data after careful consideration 22 variables are slected for analysis.

Data Cleaning & Manipulation

- Removed “%” from the int_rate values
- Replaced the NA values with the median in pub_rec_bankruptices
- Removed “%” from the revol_util values and imputed NA values with median

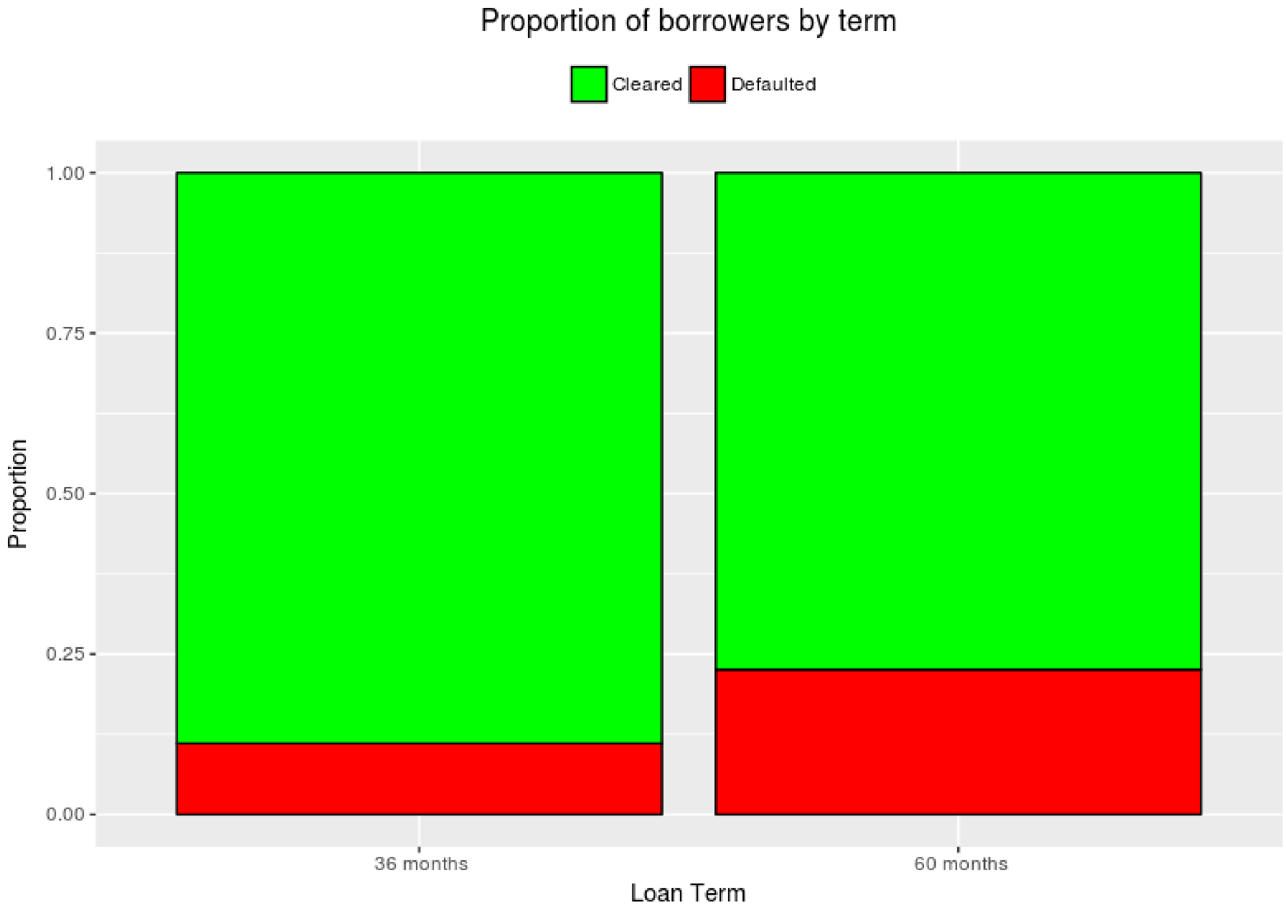
Data Analysis

Univariate Analysis

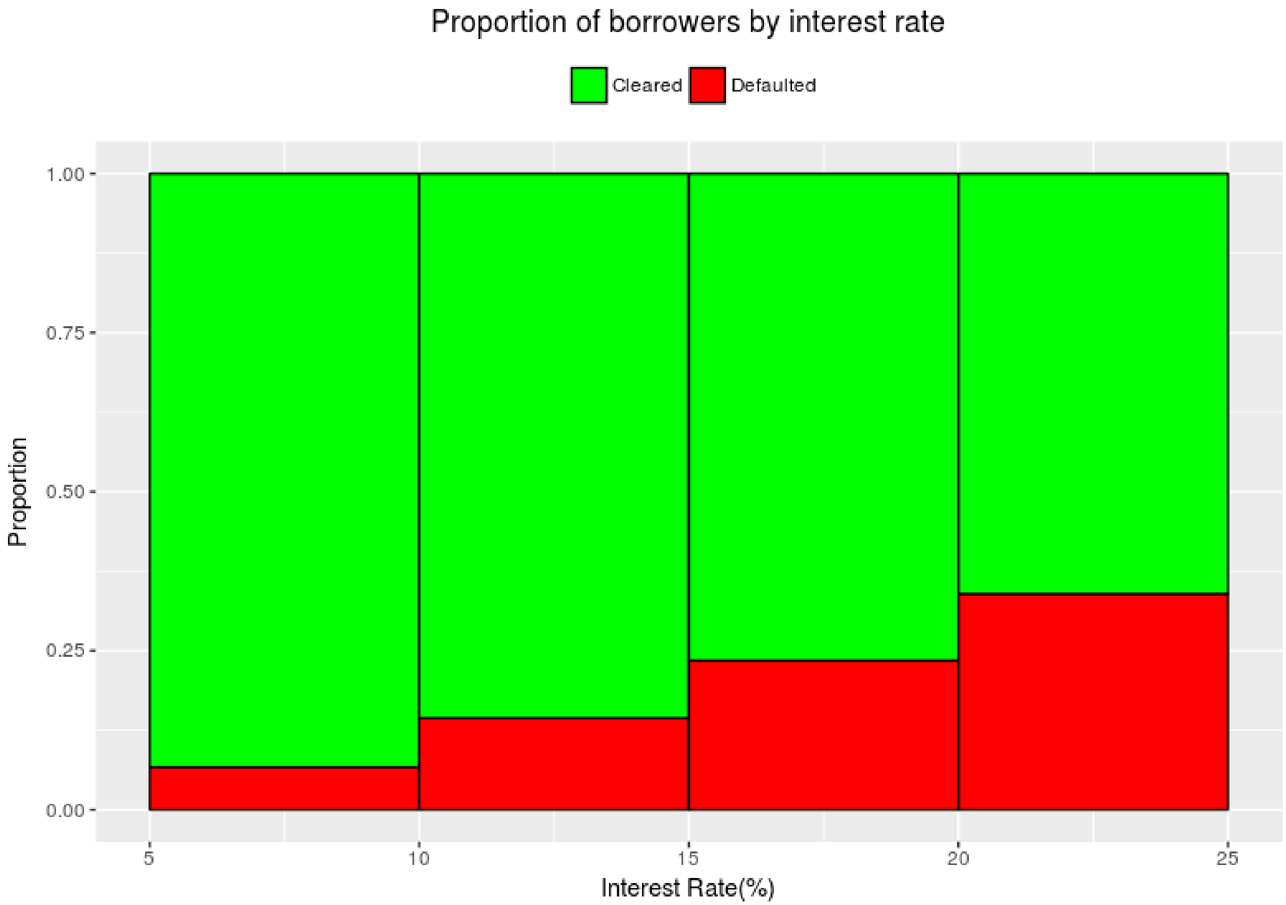


With increase in loan amount the proportion of defaulted loans in increasing.

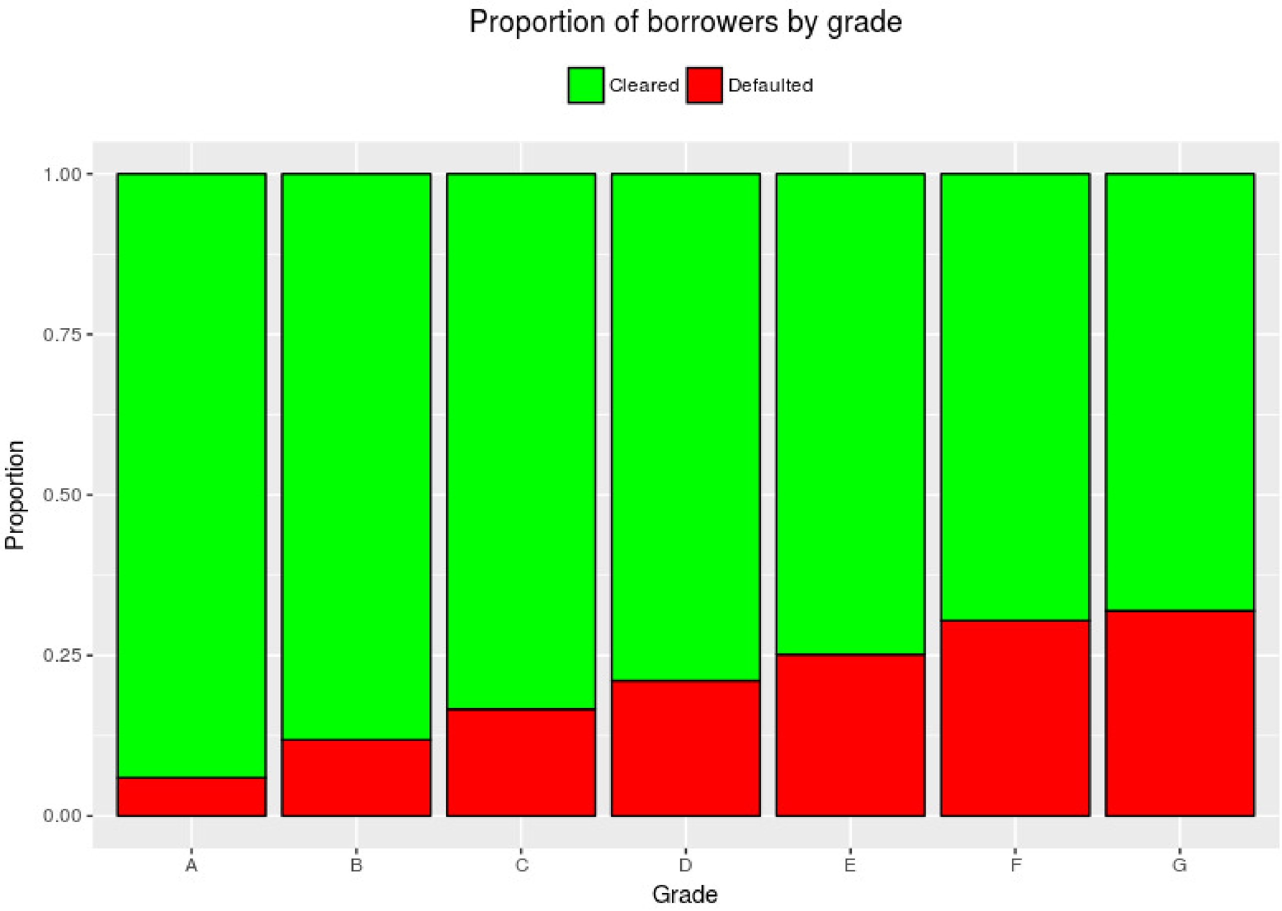
HENCE LOAN AMOUNT IS AN IMPORTANT DRIVER VARIABLE



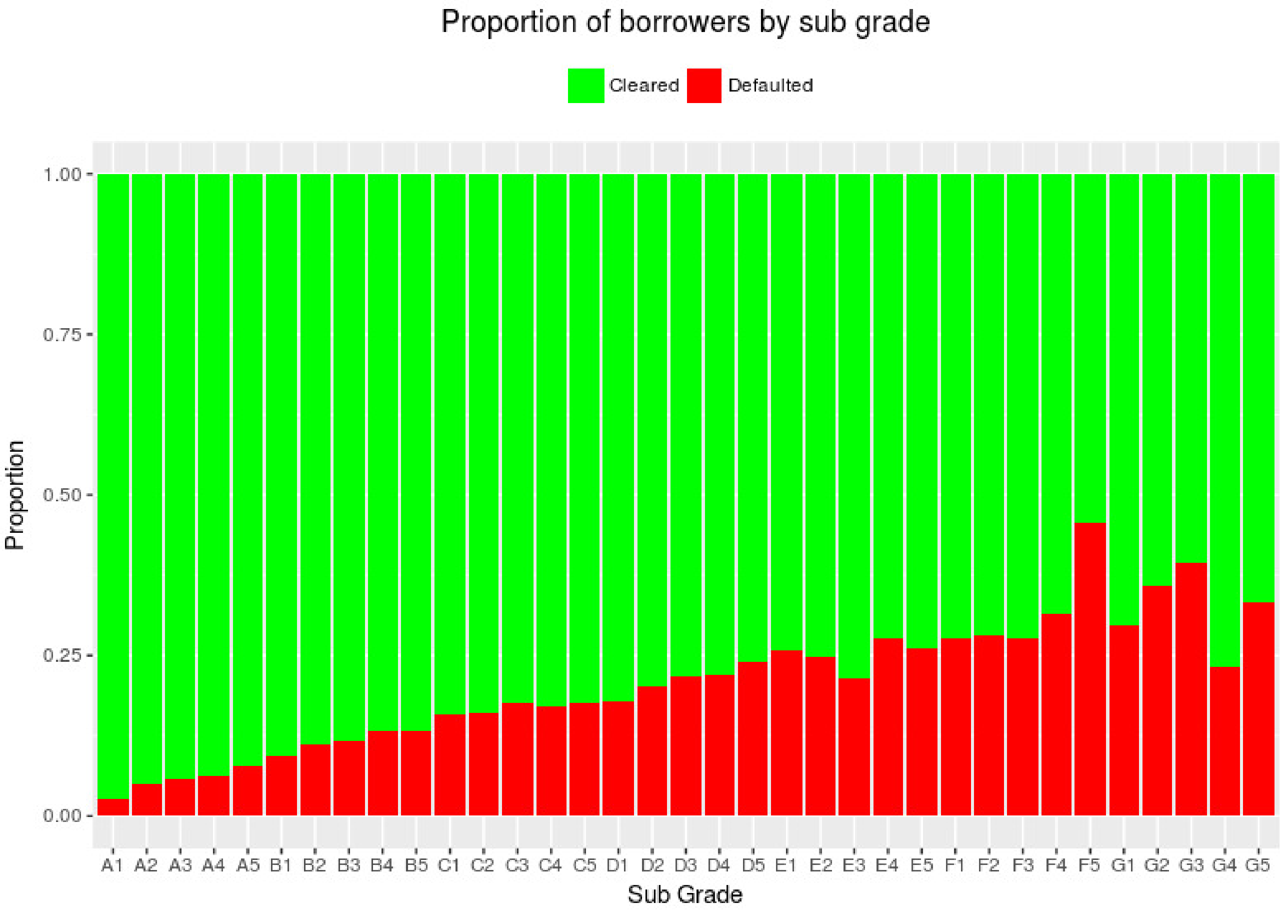
Higher the term higher the is the proportion of defaulted loans.
HENCE LOAN TERM IS AN IMPORTANT DRIVER VARIABLE.



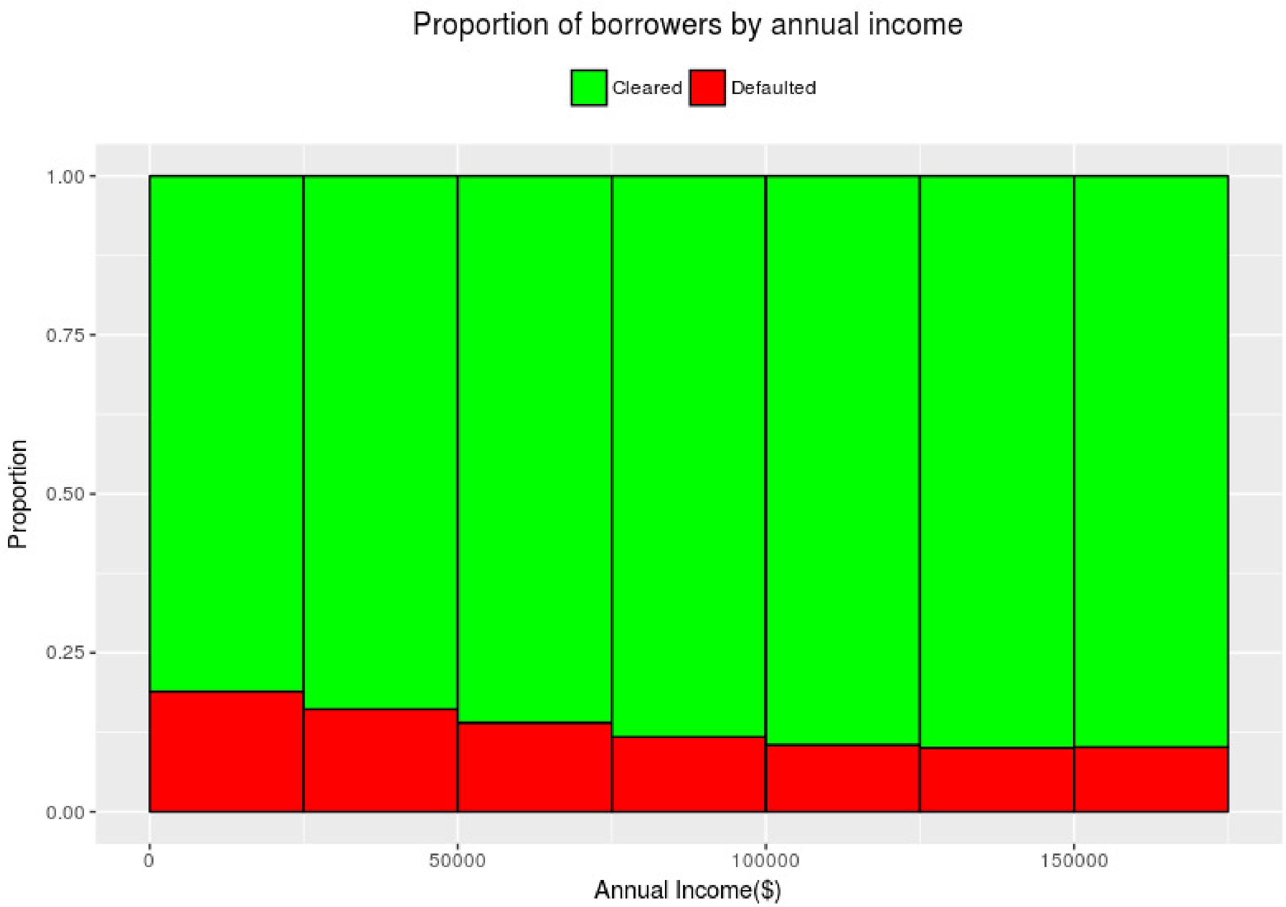
With increase in interest rate the proportion of defaulted loans in increasing.
HENCE INTEREST RATE IS AN IMPORTANT DRIVER VARIABLE.



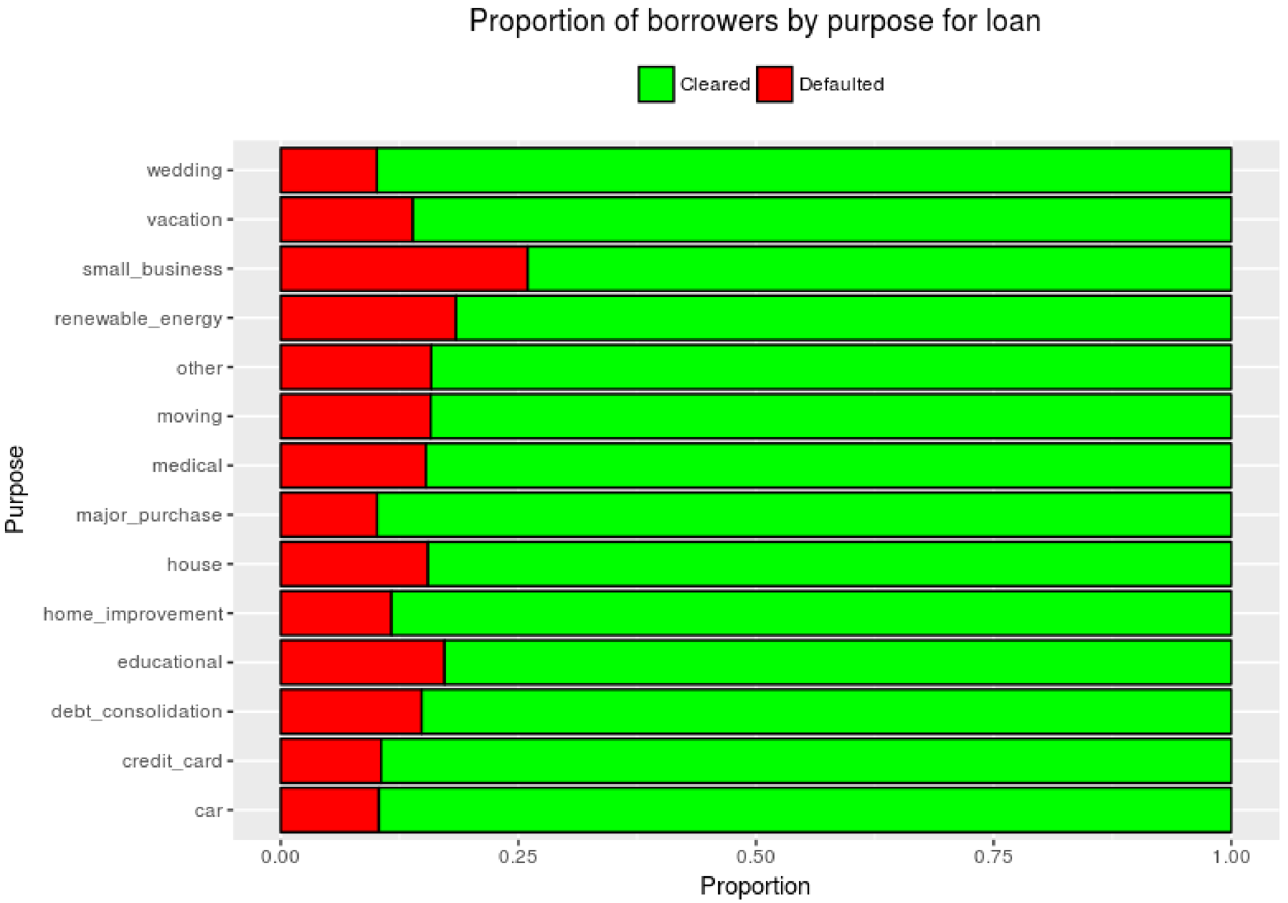
With increase in grade the proportion of defaulted loans in increasing.
HENCE GRADE IS AN IMPORTANT DRIVER VARIABLE.



SUB GRADE IS AN IMPORTANT DRIVER VARIABLE.

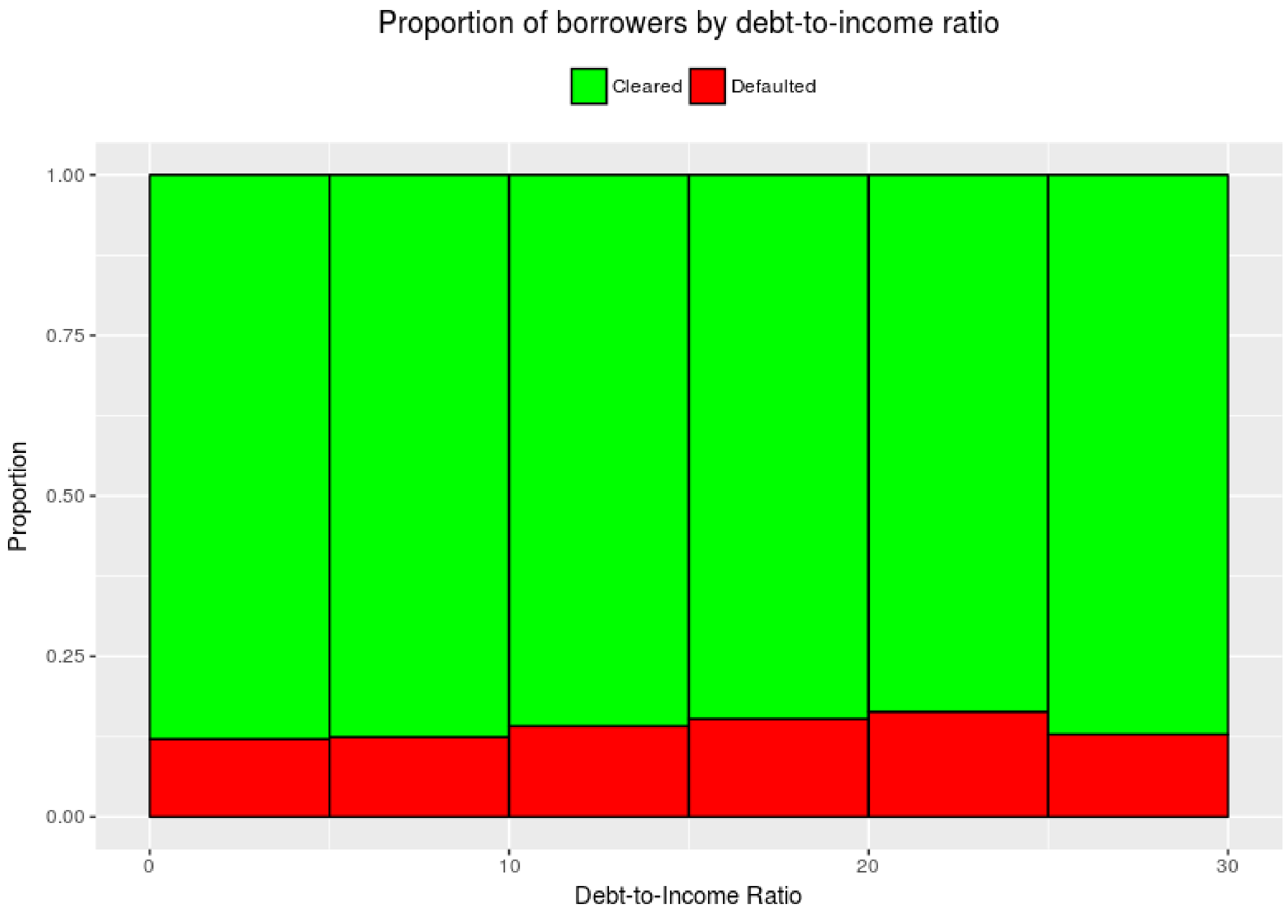


Lower the annual income higher the proportion of defaults.
The proportion of defaults decreases as income increases.
HENCE ANNUAL INCOME IS AN IMPORTANT DRIVER VARIABLE.

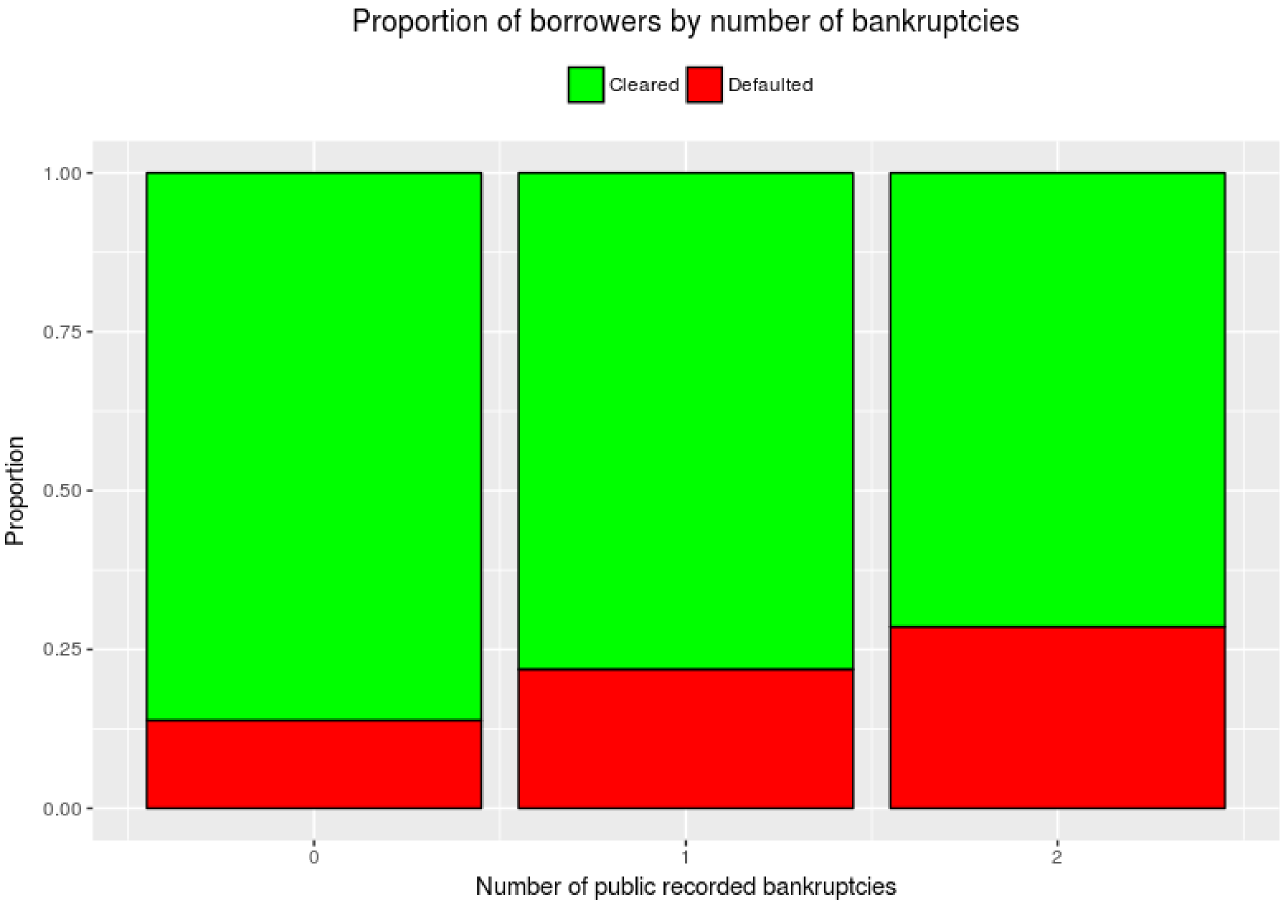


Small business has the highest proportion of defaults. Followed by renewable energy and educational purpose. It makes sense that if the borrower's small business is not doing well then it is difficult to repay the loan.

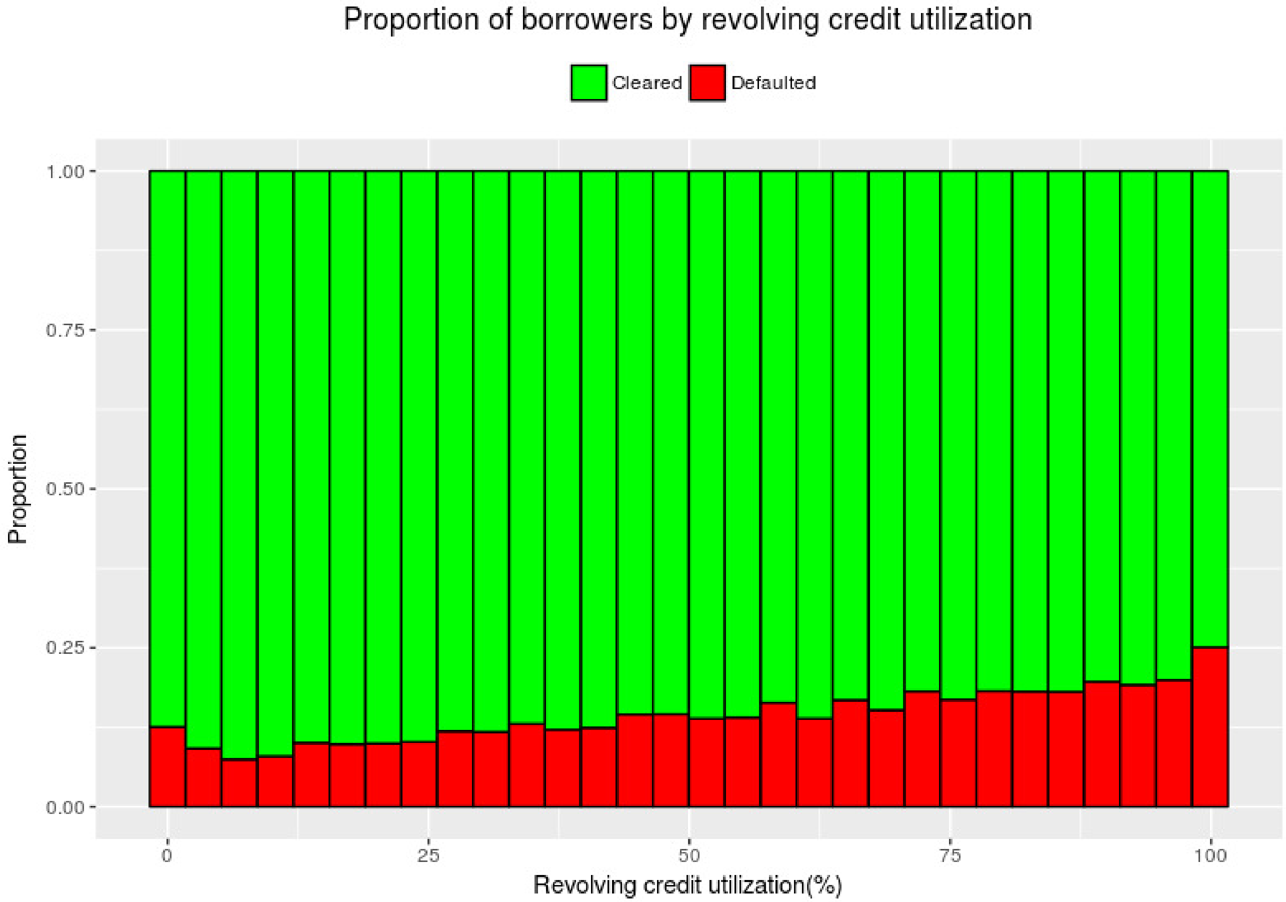
PURPOSE IS AN IMPORTANT DRIVER VARIABLE



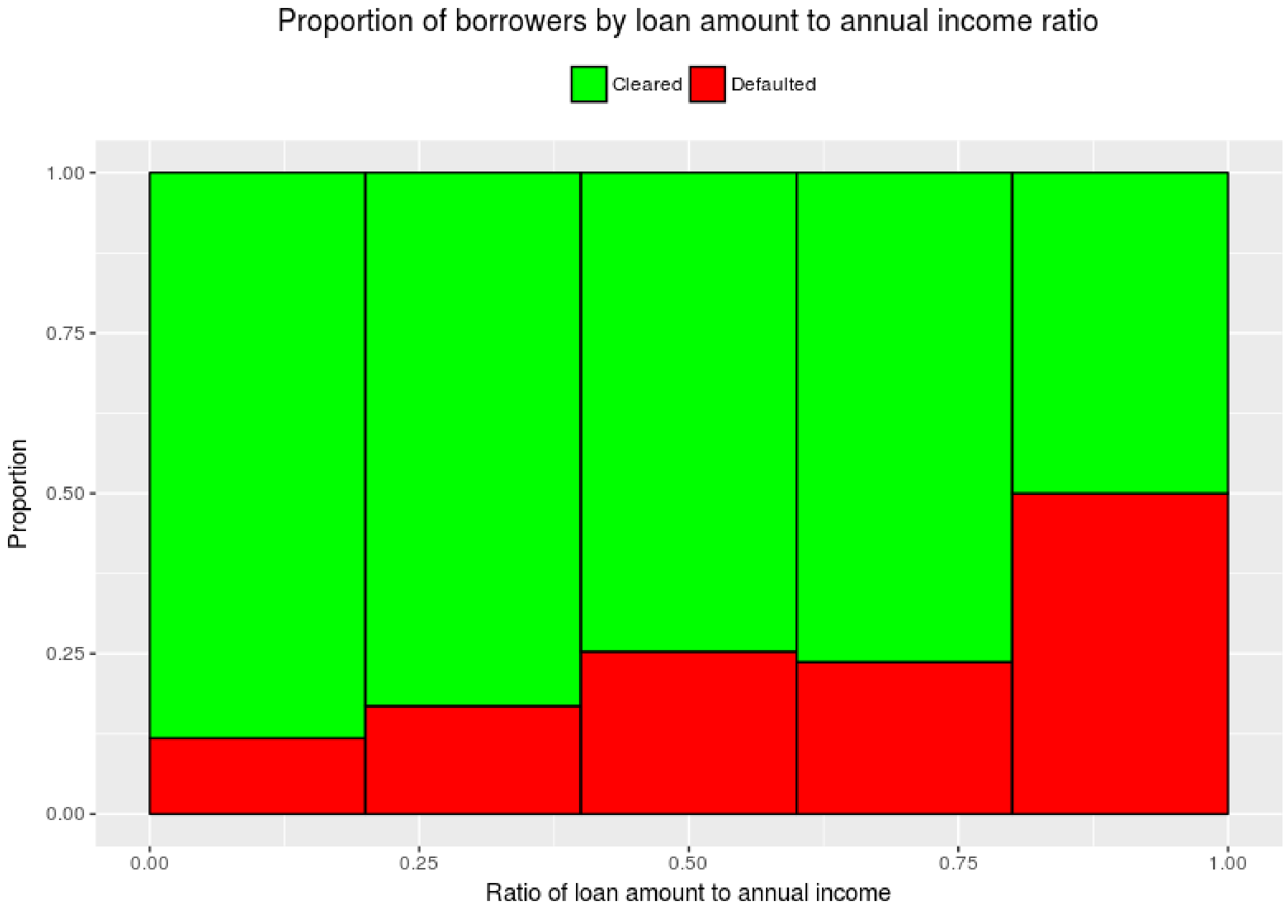
There is a steady increase in the proportion of defaulters as dti increases from 0-25%. There is a decrease at 25-30%. Maybe thats because of stricter standards being followed to approve loans to borrowers with high dti. Nevertheless dti is an important factor. HENCE DEBT-TO-INCOME RATIO IS AN IMPORTANT VARIABLE.



The proportion of defaulters increases with increasing number of bankruptcies.
HENCE NUMBER OF PUBLIC RECORD BANKRUPTCIES IS AN IMPORTANT DRIVER VARIABLE



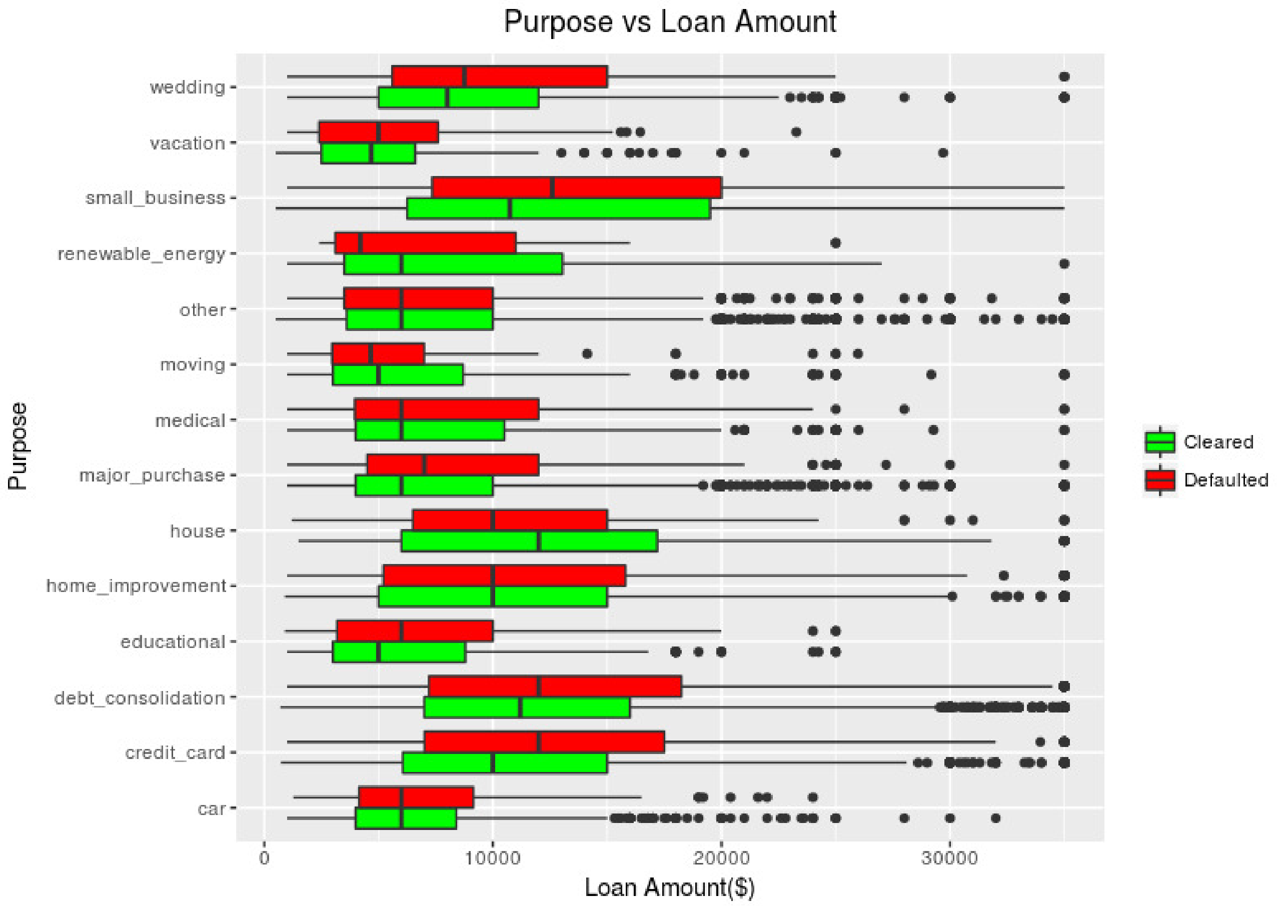
The proportion of defaulters increases with increasing revolving credit utilization.
HENCE REVOLVING CREDIT UTILIZATION IS AN IMPORTANT DRIVER VARIABLE



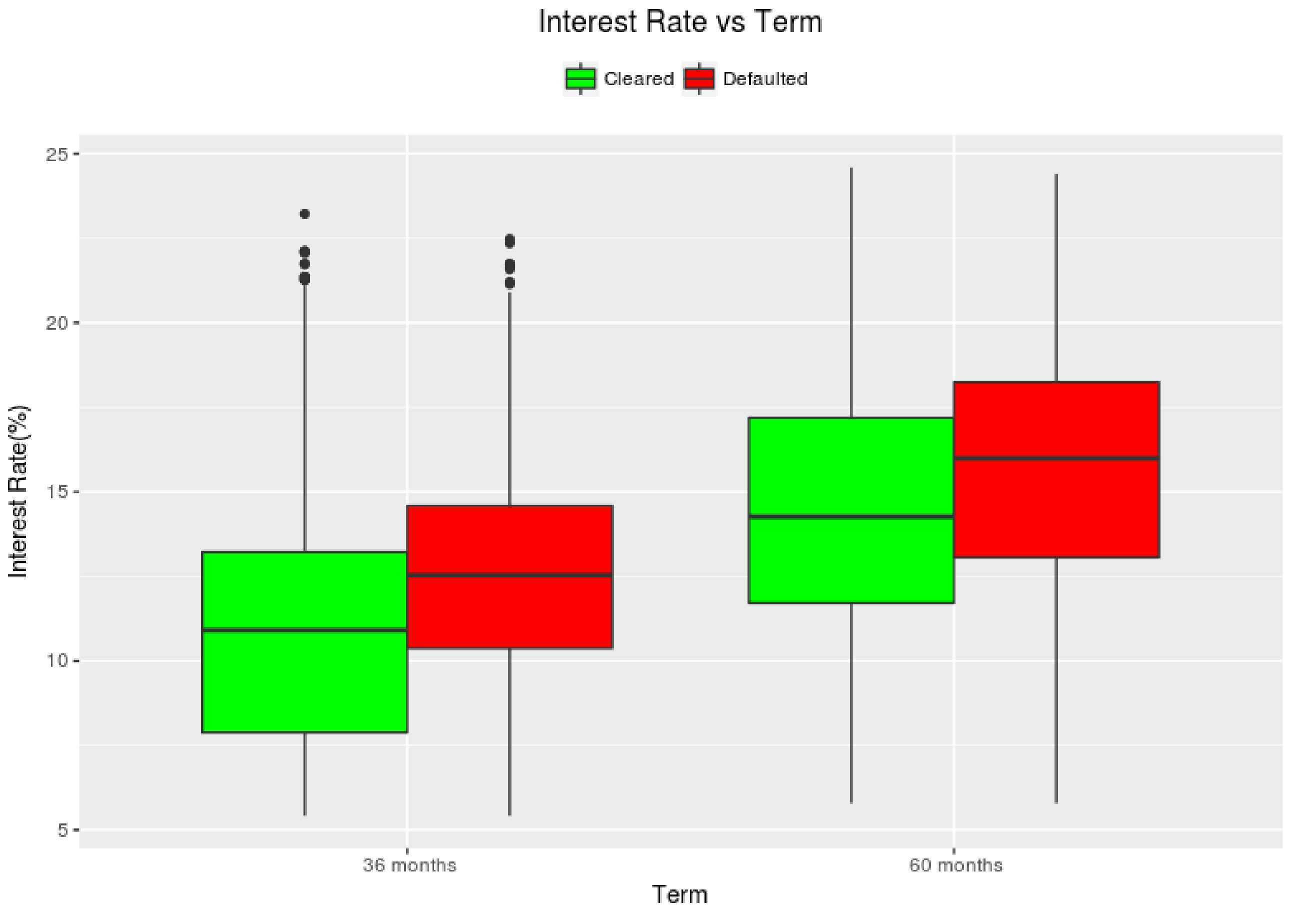
Higher the ratio, higher is the proportion of defaulters.
HENCE LOAN AMOUNT TO ANNUAL INCOME IS AN IMPORTANT DRIVER VARIABLE.

Data Analysis

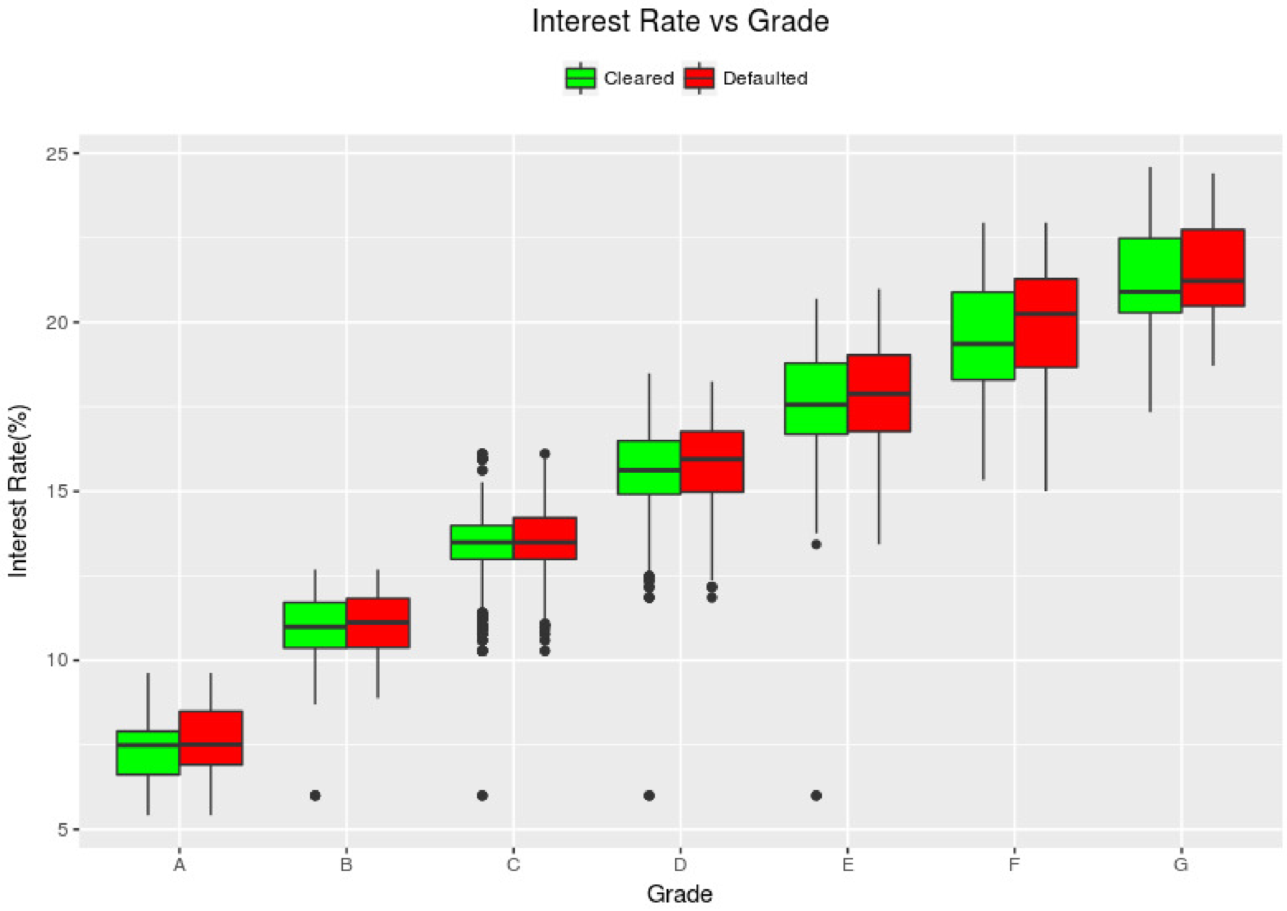
Bivariate Analysis



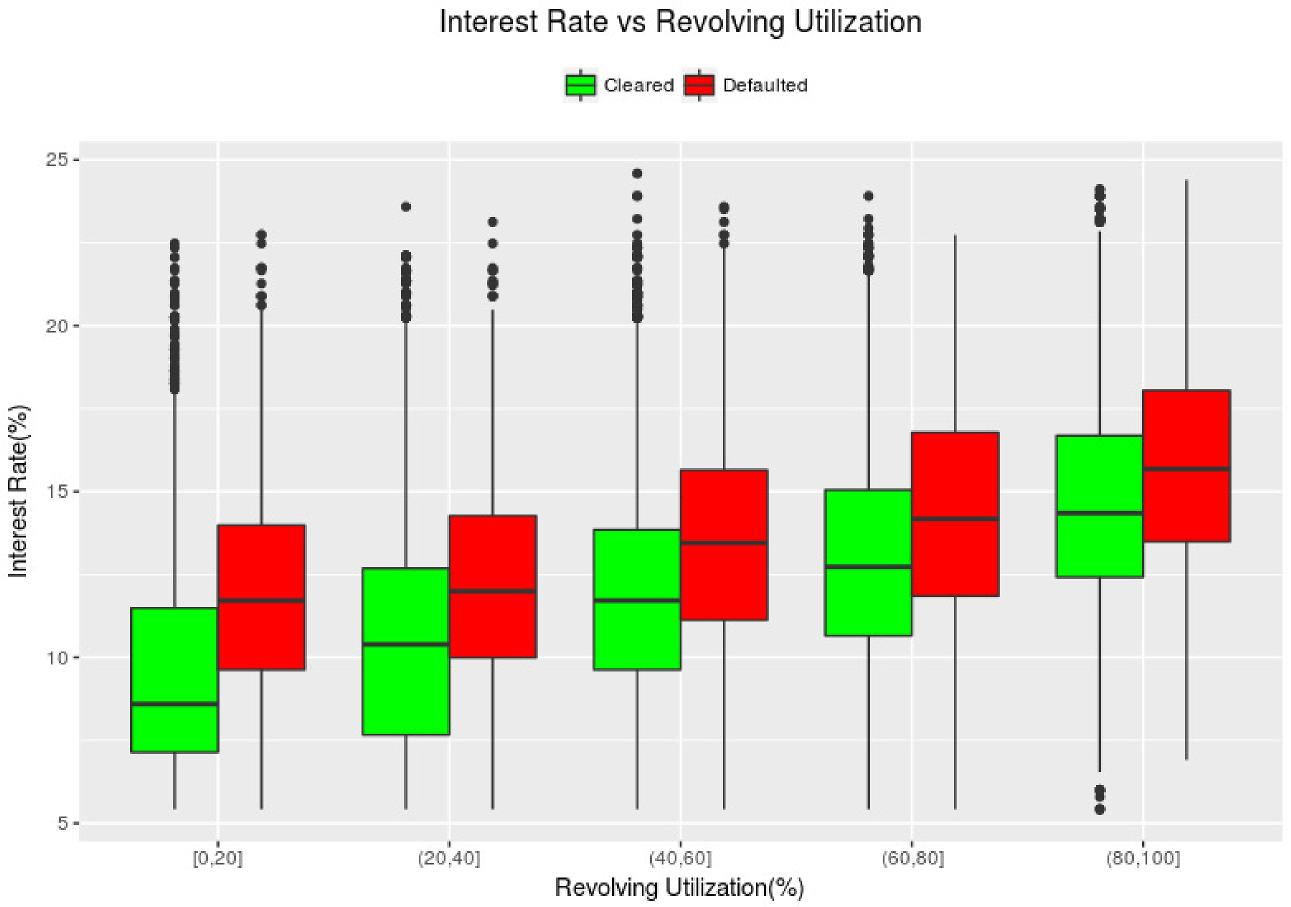
Small business had the highest median loan amount for defaulters. For most of the reasons (purpose) the median loan amount was the same or higher for defaulted compared to cleared loans.



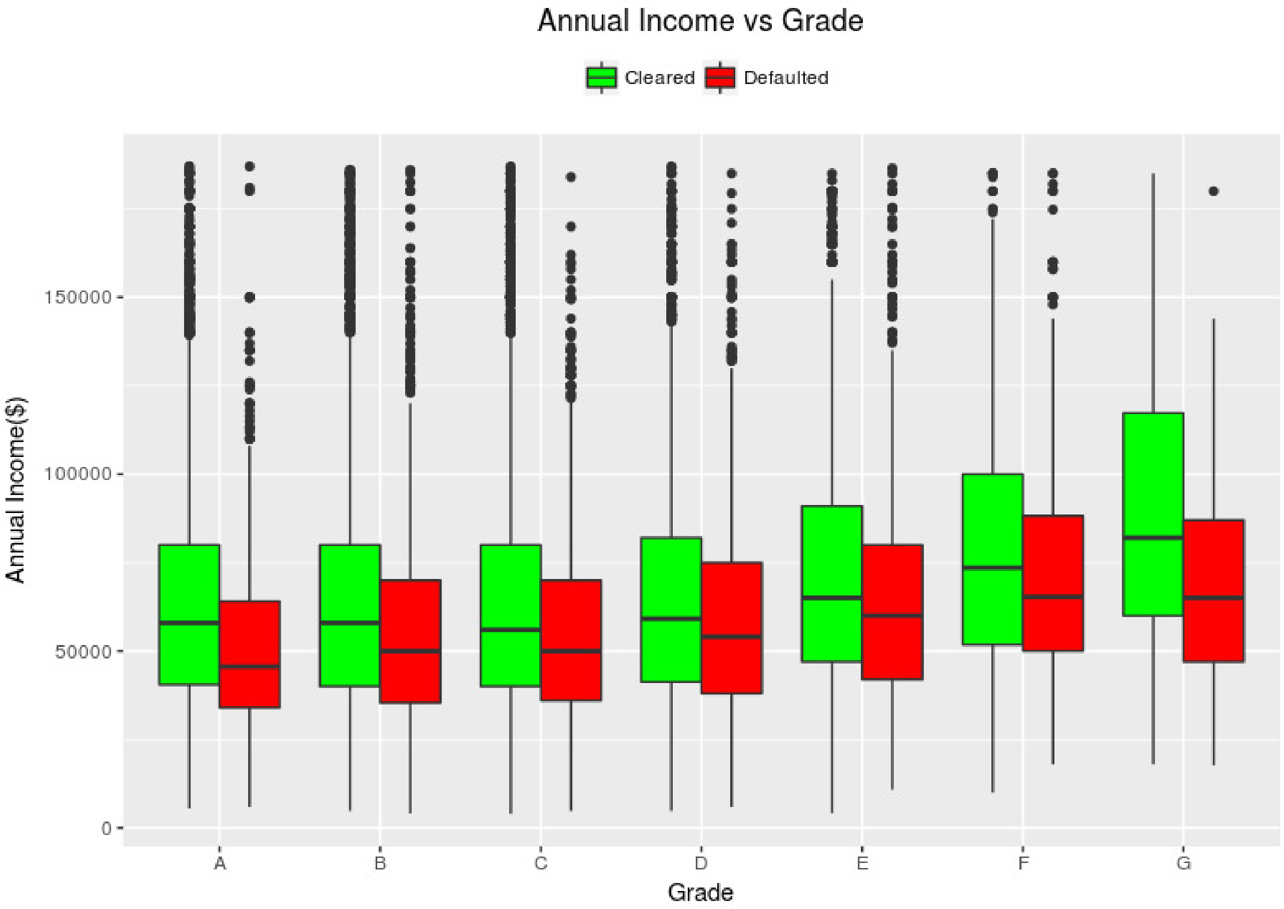
The median interest rate was higher for 60 months duration loans compared to 36 months. And the median interest rate was higher for defaulted loans compared to cleared loans.



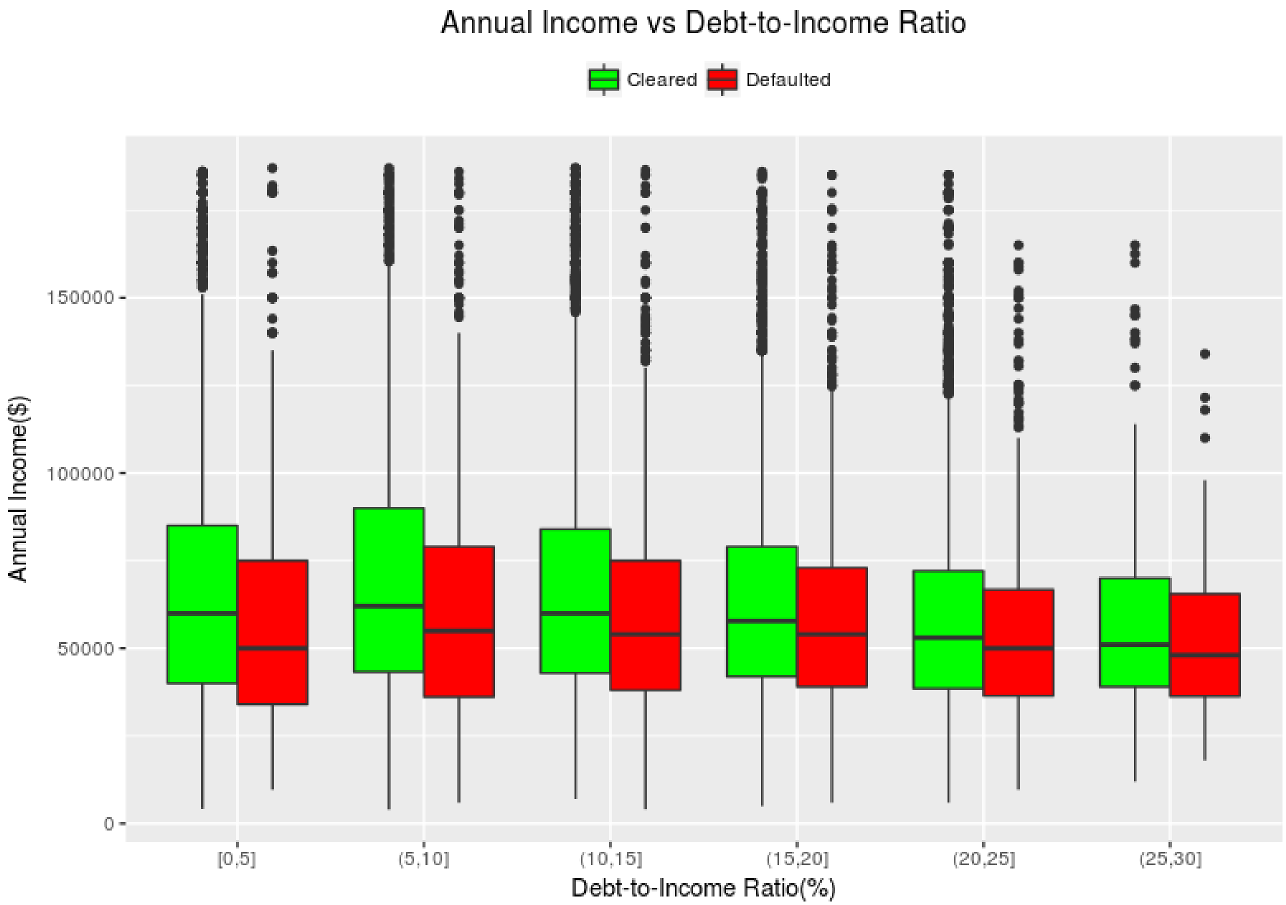
The rate of interest increases as grade increases. For each grade the median interest rate for defaulted loans was higher compared to cleared loans.



The interest rate increases as revolving utilization increases. The median interest rate for defaulted loans was higher compared to cleared loans for all the bins.



For every grade the defaulters had lower median annual income compared to those who cleared their loans.



The median income of defaulters is lower compared to that of cleared loans.

Conclusions

- **Interest rate (or grade) and annual income variables provide the most predictive power for determining potential defaulters.**
- IMPORTANT DRIVER VARIABLES:
  - int_rate - Loan interest rate
  - grade - Assigned loan grade corresponding to interest rate based on borrower's credit history
  - sub_grade - Assigned loan sub grade corresponding to grade based on borrower's credit history
  - dti - Debt-to-income ratio
  - loan_amnt - Loan amount
  - annual_inc - Annual income
  - term - Term of the loan (36 or 60 months)
  - purpose - Reason for borrowing money
  - pub_rec_bankruptcies - Number of bankruptcies on public record
  - revol_util - Amount of credit the borrower is using relative to all available revolving credit.
  - loan_amount_by_annual_inc - Loan amount to annual income ratio