EDA - Gramener Case Study Presentation

SUBMISSION

Group Name: Synergy

1. D Mruthyunjaya Kumar (Facilitator)
2. Dharmanandana Reddy Pothula
3. Ashwin Suresh
4. Manohar Shanmugasundaram

EDA - Gramener Case Study Analysis

Background

The company 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.

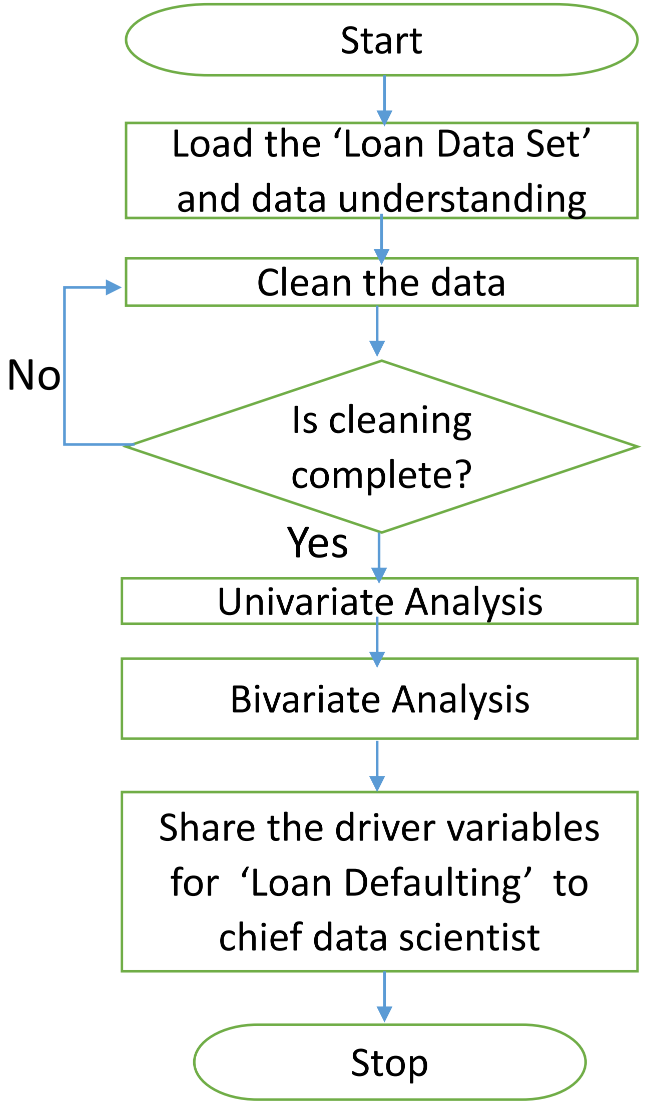
Business Objectives

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The 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’.

Our aim on this case study is to understand the driving factors (or driver variables) behind loan default and identify risky loan applications using EDA. The company can utilize this knowledge for its portfolio and risk assessment so that such loans can be reduced thereby cutting down the amount of credit loss.

Problem solving methodology

The following flowchart provides the steps carried out for this analysis:



Start

Load the 'Loan Data Set' and data understanding

Clean the data

Is cleaning complete?

No

Yes

Univariate Analysis

Bivariate Analysis

Share the driver variables for 'Loan Defaulting' to chief data scientist

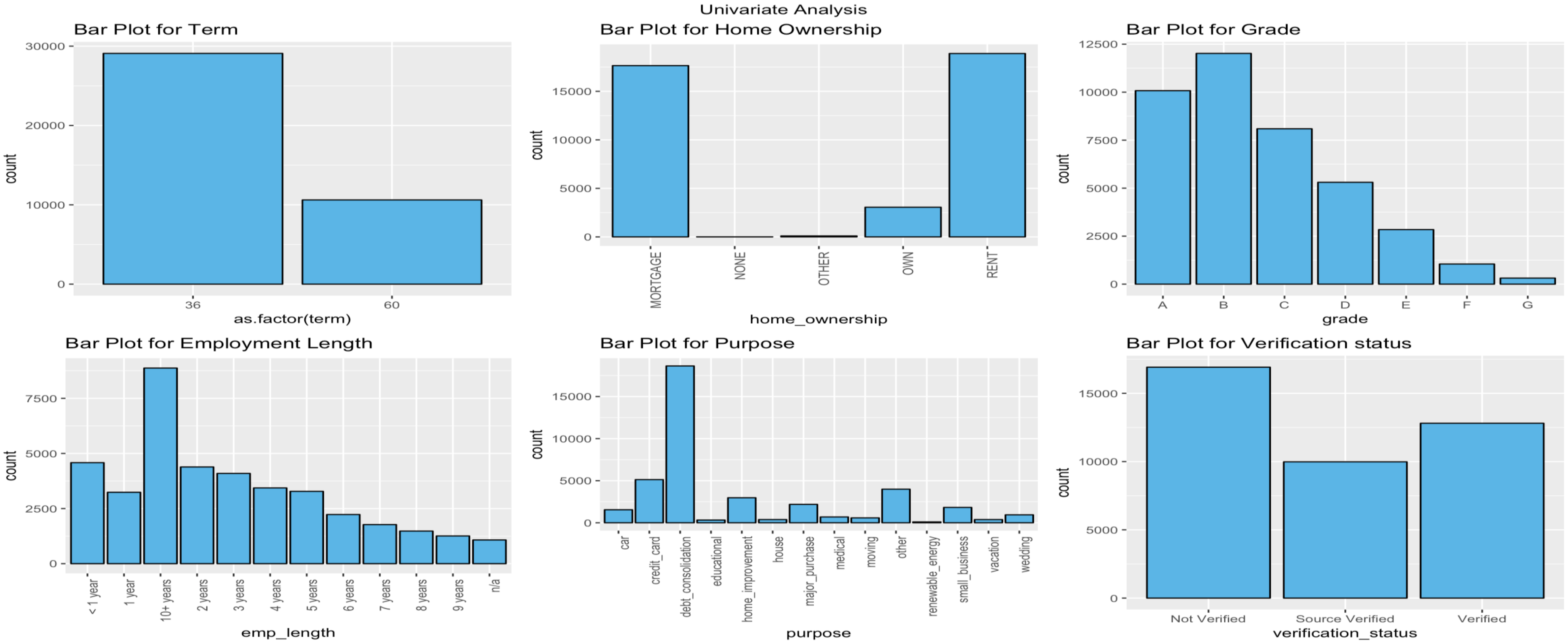
Stop

Data Cleaning

- Ignored all the columns which are only have 'NA' values.
- Identified and removed the near zero variance columns.
- Ignored all the columns which are related to customer payments (since these details will not help for this analysis).
- Ignored the other fields like zip_code, emp_title, URL, etc, as these not related to this analysis.
- Removed additional string or special character values in some of the variables (term, int_rates, etc), so that it can be used for analysis like grouping, etc.
- Removed outliers in the variable like annual income (using box plots).

Univariate Analysis

Univariate analysis is done on the variables after data cleaning to see the insight of the variable. Following are some of the plots identified during the univariate and segmented univariate analysis.



Univariate Analysis

Bar Plot for Term

Bar Plot for Home Ownership

Bar Plot for Grade

Bar Plot for Employment Length

Bar Plot for Purpose

Bar Plot for Verification status

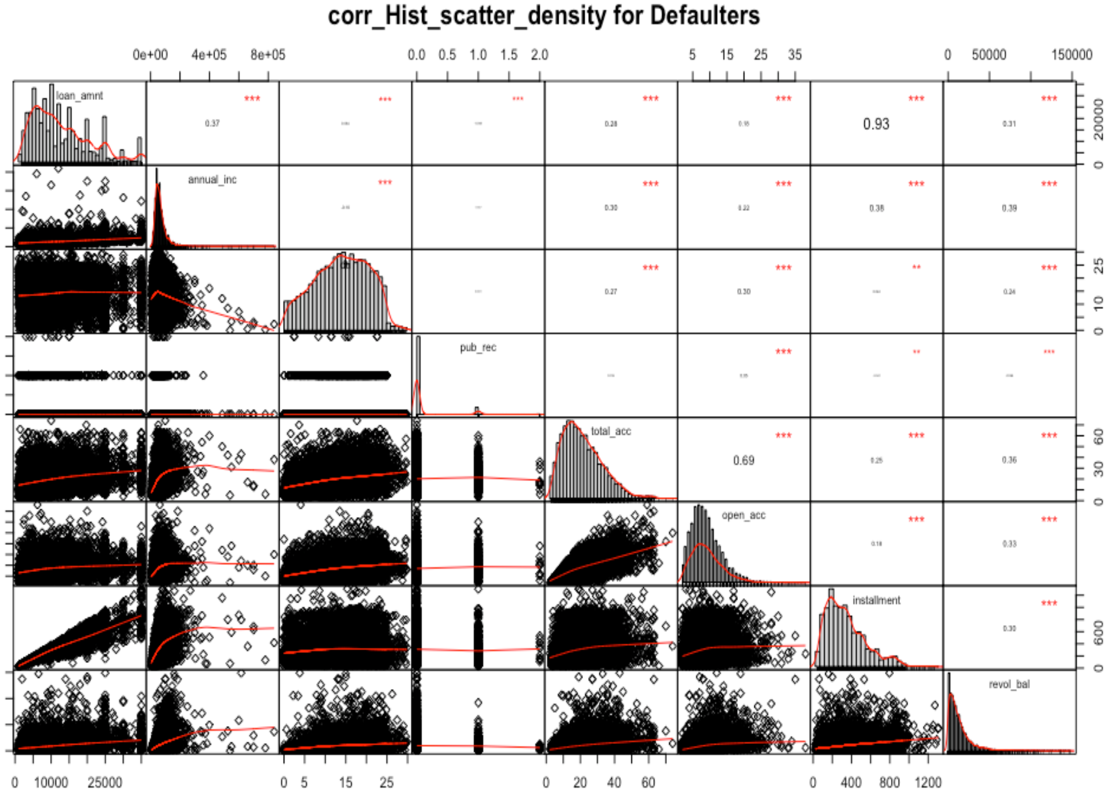
Derived Metrics Analysis

Derived Metrics analysis is done on the variables to get better insights on the variables. Below the derived variables as part of the analysis.

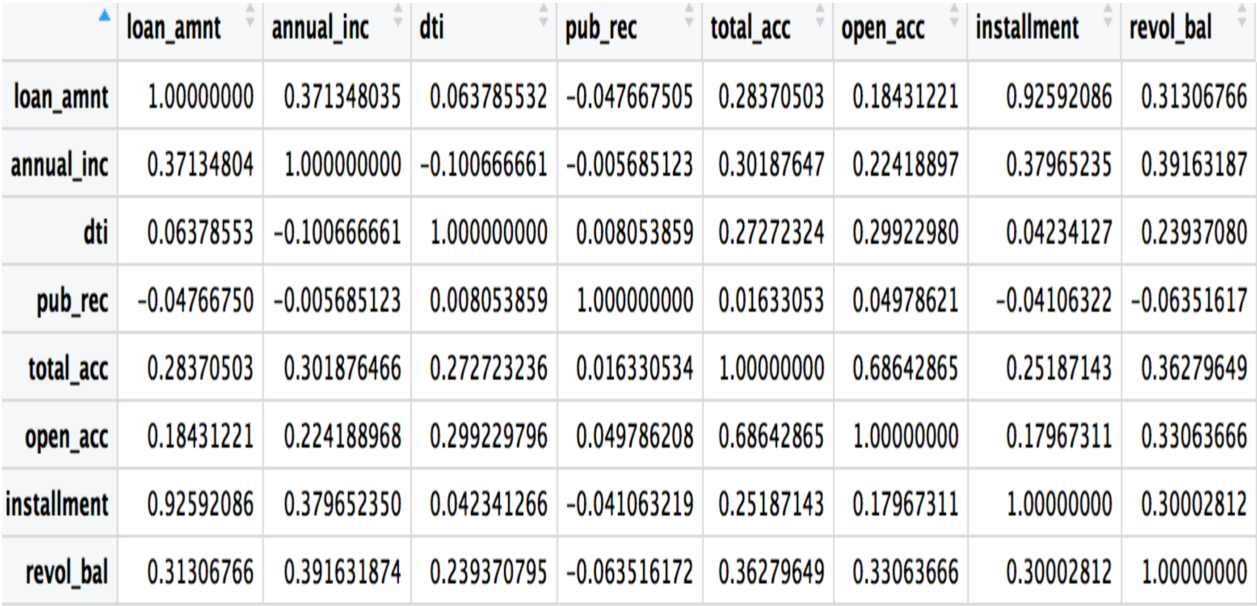
- Default (where the loan is default or not)
- int_rate_grp (Interest rate group, where bins are created from the int_rate variable for analysis)
- annual_inc_grp (Annual income group, where bins are created from the annual_inc variable for analysis)
- Installment_grp (bins formed from the installment)
- dti_grp (bins formed from the dti variable)
- revol_util_grp (bins formed from the revol_util variable)

Correlation Analysis on numeric data for defaulters

The below plot represents the correlation matrix with the correlation coefficient and density between the variables. Also the below table provides correlation matrix with the correlation values across each numeric variables, this provides insight about the correlation between the variables.



Correlation Matrix



	loan_amnt	annual_inc	dti	pub_rec	total_acc	open_acc	installment	revol_bal
loan_amnt	1.00000000	0.371348035	0.063785532	-0.047667505	0.28370503	0.18431221	0.92592086	0.31306766
annual_inc	0.37134804	1.000000000	-0.100666661	-0.005685123	0.30187647	0.22418897	0.37965235	0.39163187
dti	0.06378553	-0.100666661	1.000000000	0.008053859	0.27272324	0.29922980	0.04234127	0.23937080
pub_rec	-0.04766750	-0.005685123	0.008053859	1.000000000	0.01633053	0.04978621	-0.04106322	-0.06351617
total_acc	0.28370503	0.301876466	0.272723236	0.016330534	1.00000000	0.68642865	0.25187143	0.36279649
open_acc	0.18431221	0.224188968	0.299229796	0.049786208	0.68642865	1.00000000	0.17967311	0.33063666
installment	0.92592086	0.379652350	0.042341266	-0.041063219	0.25187143	0.17967311	1.00000000	0.30002812
revol_bal	0.31306766	0.391631874	0.239370795	-0.063516172	0.36279649	0.33063666	0.30002812	1.00000000

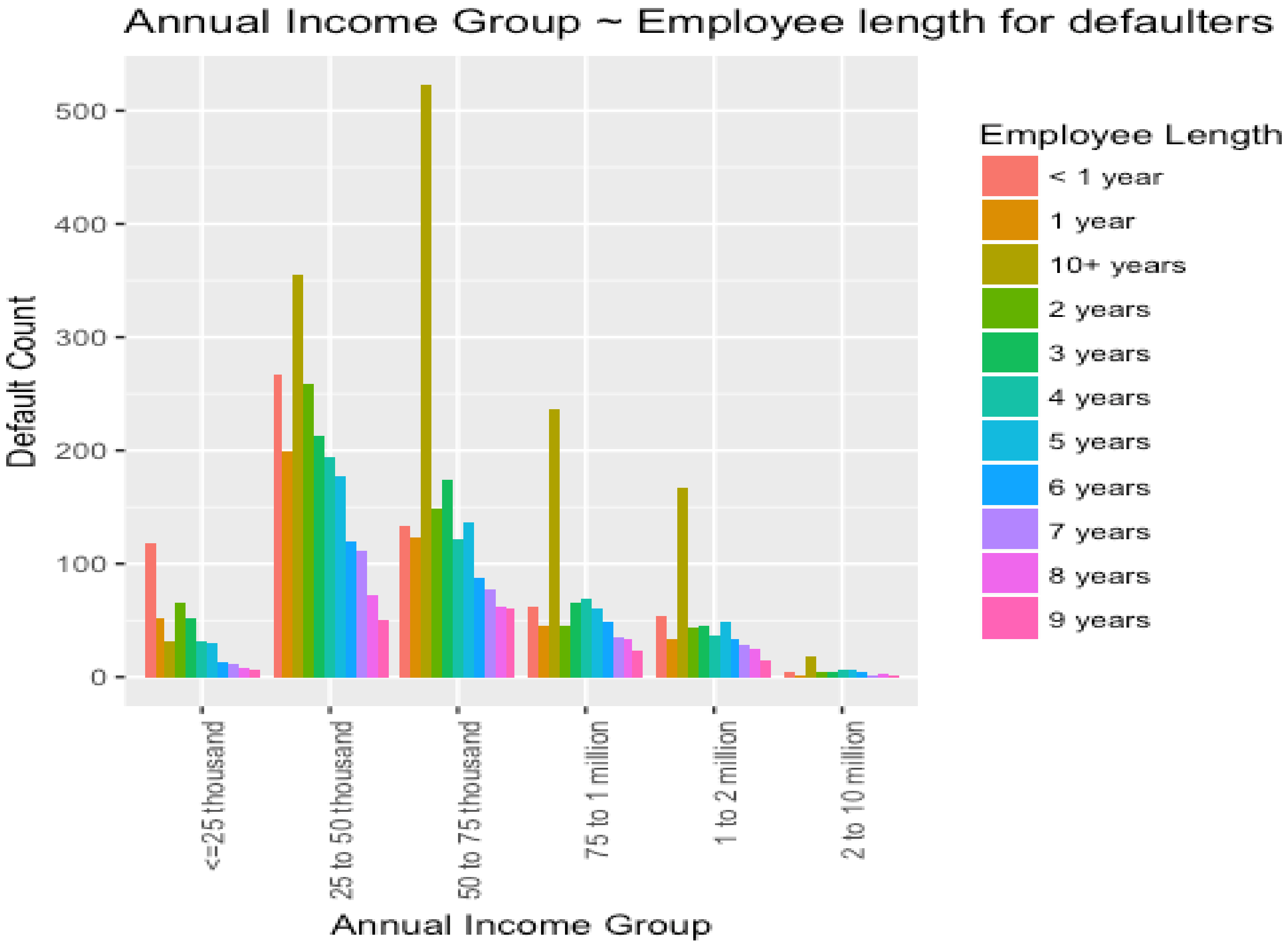
Bivariate Analysis

Bivariate analysis is done on the variables after univariate analysis are complete. This analysis give more insights about the variables with relation with another variable.

- Bivariate analysis on the categorical variables
- Bivariate analysis on the continuous variables (using correlation, etc.)

The next slides will show some of the plots generated as part of the bivariate analysis.

Annual Income and Employment length Analysis

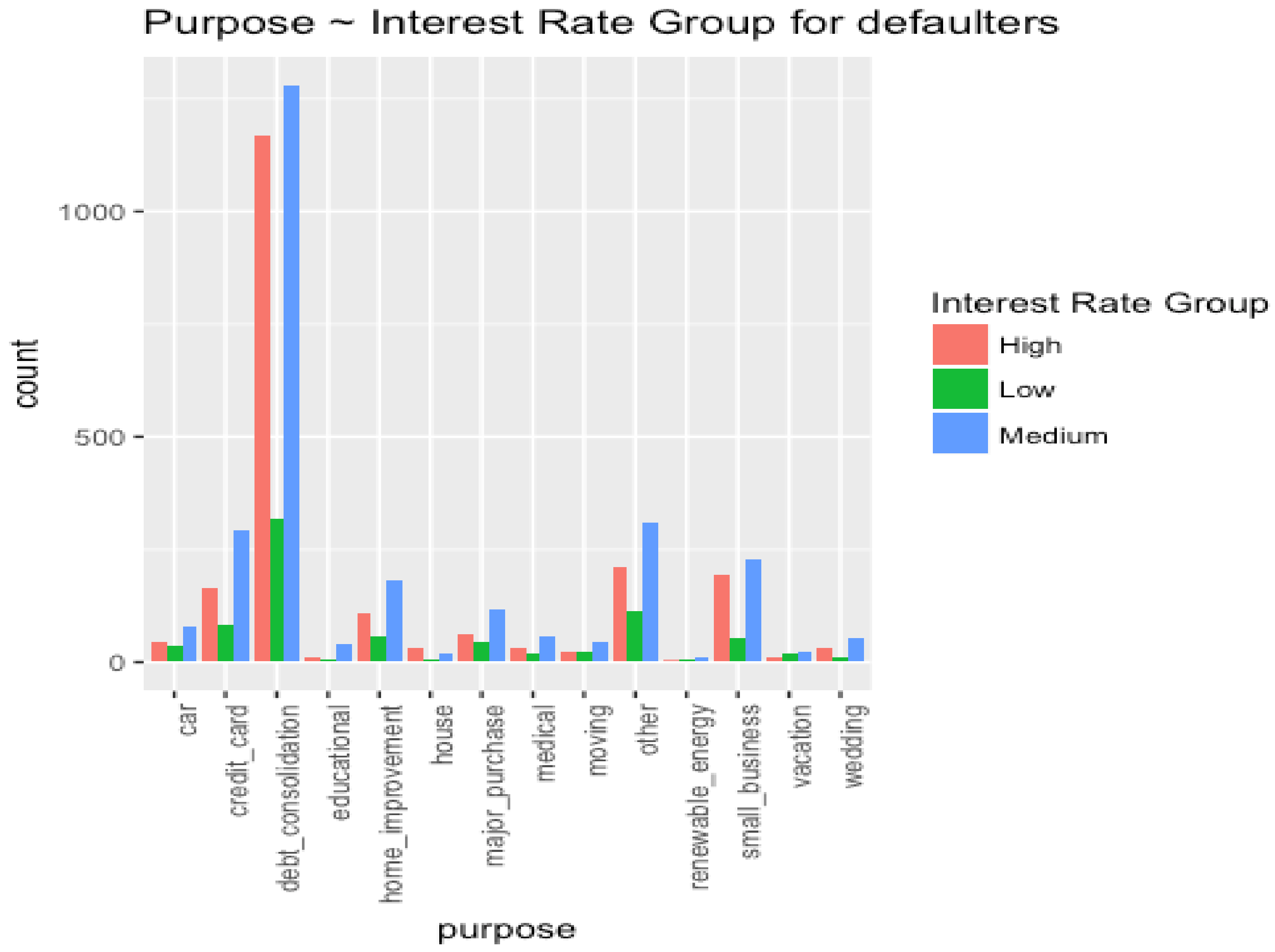


For this analysis the annual income is grouped into multiple bins for defaulters.

From this plot analysis, we could see the Annual income of **25 to 50 thousand & 50 to 70 thousand** are the major driver for default identification.

Also the employee length of **0 to 5 years and 10+ years** are have more default records and this also another driver for default identification.

Loan Purpose and Interest Rate Analysis



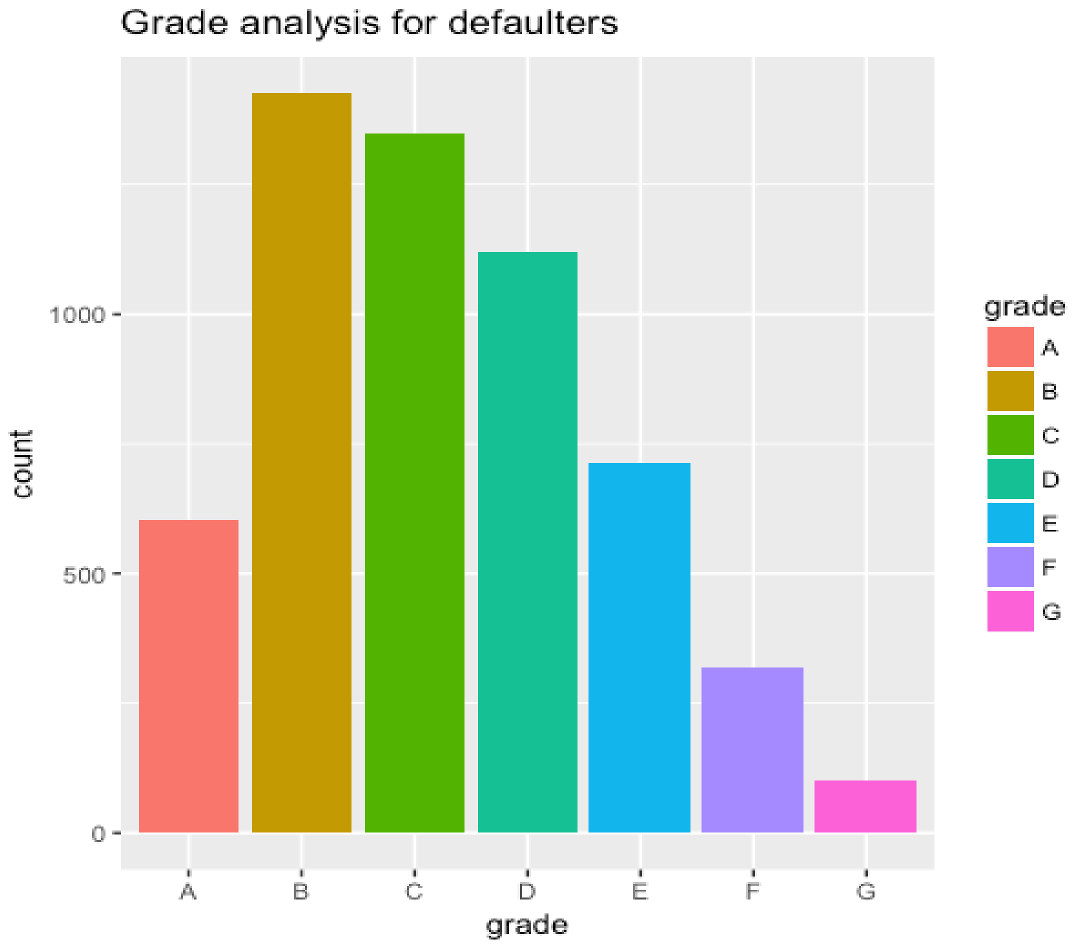
Interest Rate Group

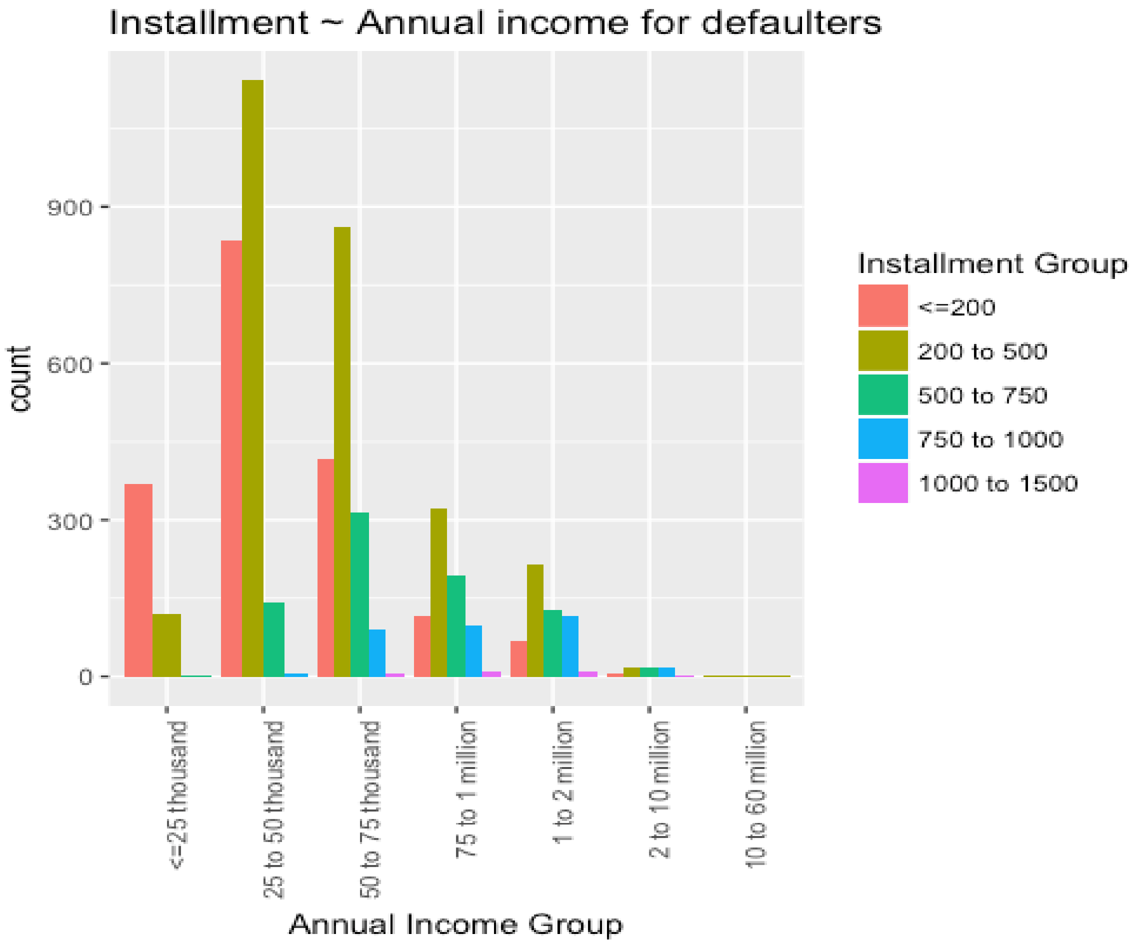
Group	Interest rate
Low	int_rate < 10
Medium	int_rate >=10 and < 15
High	int_rate >= 15

In this plot, we could see that the loan purpose with the following **'debt_consolidation', 'credit_card', 'other' and 'small_business'** are higher cause for defaults.

Additionally the **interest rate**, which are **'High' (>=15) and 'Medium'(>=10 and <15)** are high drivers for default.

Loan Grade and Installment Analysis

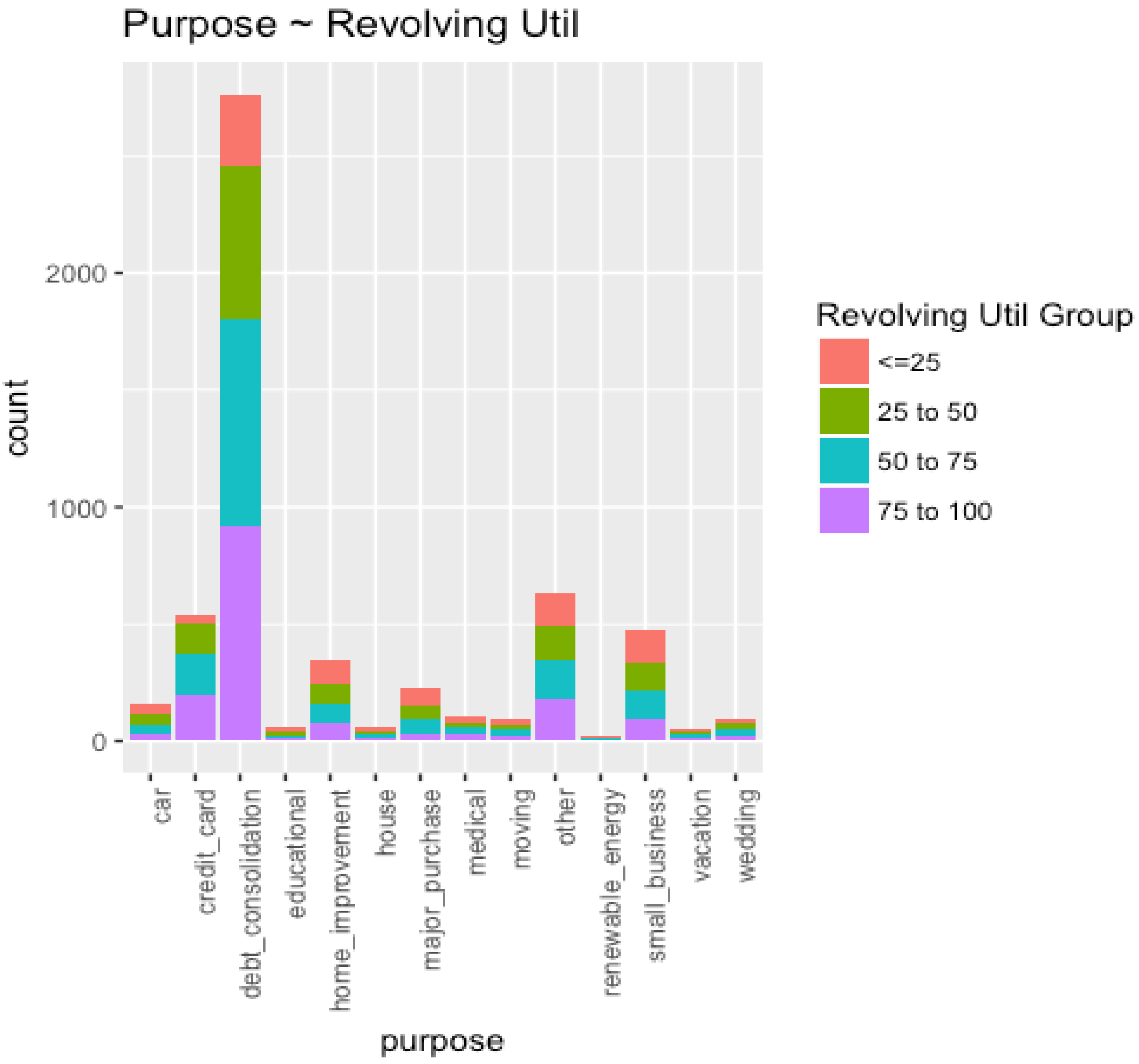




This above plot clearly shows the Loan grades **‘B’, ‘C’ and ‘D’** are high drivers for default.

This above plot clearly shows the installments from **<= 500 (groups <=200 and 200 to 500)**, have a major influence on the number of defaulters.

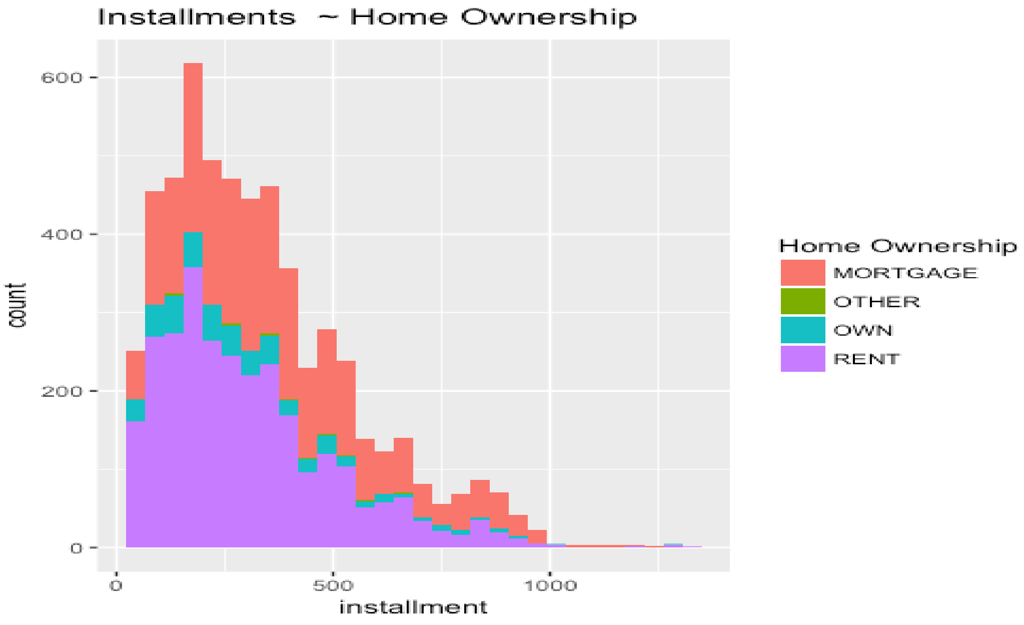
Revolving Utilisation Variable Analysis

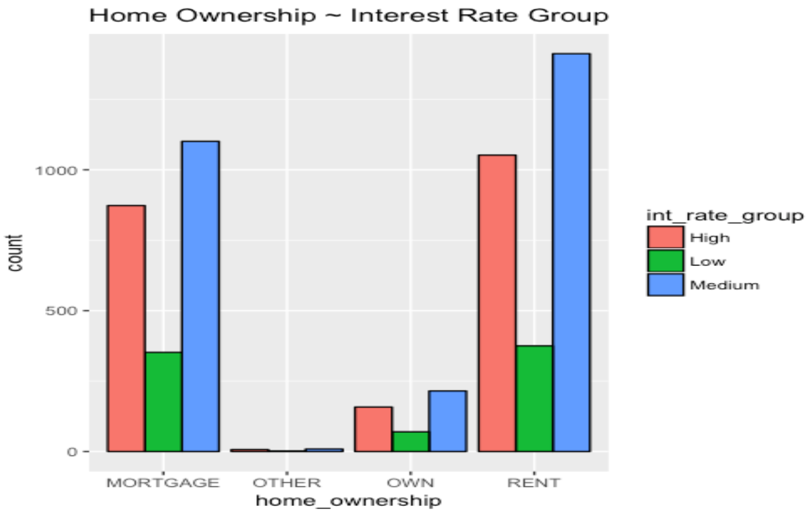


This plot analysis provides the insight that the revolving utilisation credit values impacts majorly on the following loan purposes ’debt_consolidation’, ‘credit_card’, ‘other’ and ‘small_business’ in the previous plots.

This variable (**revol_util**) along with the purpose could be used as identifier for identifying the defaulters during the loan application.

Home Ownership Variable Analysis



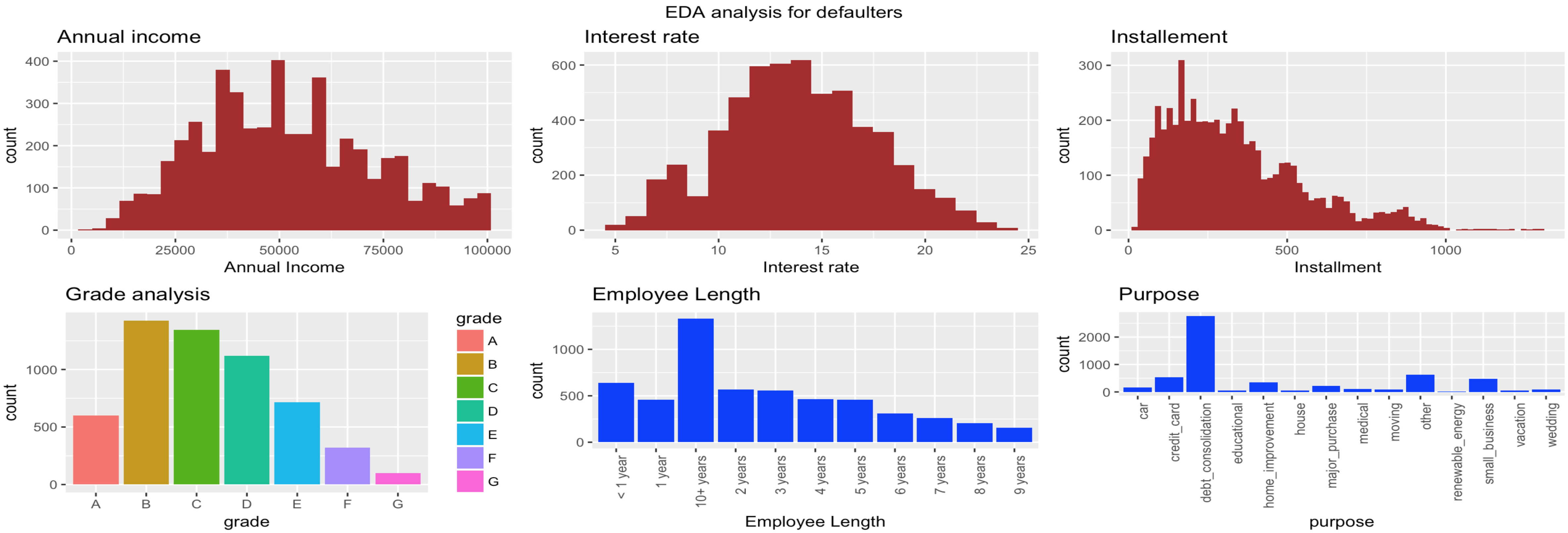


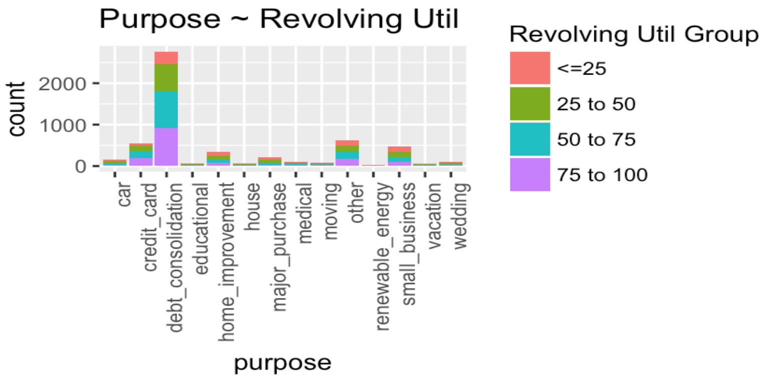
This plot analysis provides the insight that the Home Ownership impacts majorly on the installments.

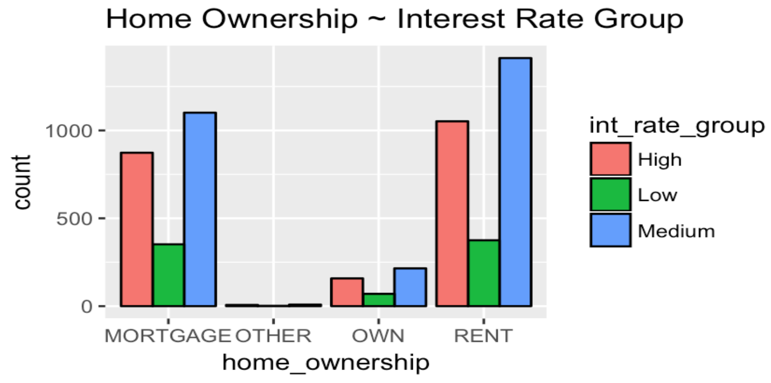
Close to 58% of people who have monthly installments in the range of 200 – 400 are most likely to default
Among them ~40% have mortgaged their home and ~50% lives in the rented flats

This variable (**home_ownership**) along with the installment and interest could be used as identifier for identifying the defaulters during the loan application.

Plots for driver variable analysis for defaulters







This above plots provides the individual analysis of the driver variables. The following are the driver variables identified as part of this analysis **annual_inc, emp_length, int_rate, installment, grade, purpose, revol_util and home ownership** can be used for identifying / predicting the defaulters during the loan application.

Recommendations to identify the defaulters (driver variables)

Below are the variables from this analysis, that can be used for identifying the loans which are risk (i.e. higher possibility of defaulters). Hence these variables combinations can be checked during loan application to decide on approving or rejecting the application.

Variable	Values to consider for defaulting	Comment
Annual Income (annual_inc)	>= 25000 and <= 75000	Annual income between 25000 and 75000 are more likely to become default.
Interest rate (int_rate)	>=10 & <=18	Majority of defaults are between interest rates 10% and 18%.
Grade	B, C & D	The customers with these grades are having a higher rate of defaulters compared to other grades.
Installment	>= 100 and <= 400	Customers with the monthly installment of between 100 and 400 are high rate of defaulters.
Employment length (emp_length)	0 to 5 years and 10+ years	Customers with work experience of 0 to 5 years and 10+ have high rate of defaulters.
Purpose	‘debt_consolidation’, ‘credit_card’, ‘other’ and ‘small_business’	These loan purposes are the major elements for defaulters.
Revolving utilization (revol_util)	<= 50	The revolving utilization (<= 50) along with the loan purposes mentioned above are contributing to majority of the defaults.
Home Ownership	Mortgage, Rent	Close to 58% of people who have monthly installments in the range of 200 – 400 are most likely to default. Among them ~40% have mortgaged their home and ~50% lives in the rented flats