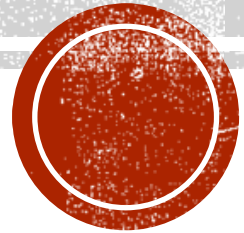
Upgrad Project

Statistics & EDA

By :- Anupam Mishra



PROBLEM STATEMENT :-

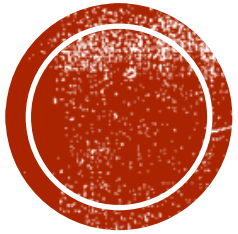
The Given data has information on the loan application. The data given below contains the information about past loan applicants and whether they 'defaulted' or not.

The main objective of the analysis was to determine the conditions and situations that leads to an applicant being charged off or default.

For this task, we need to predict the "loan_status" column of the dataset which specifies the status of the loan .

The datasethad initially 111 columns with 39716 entries.

BUSSINESS UNDERSTANDING



Consumer finance company-largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.

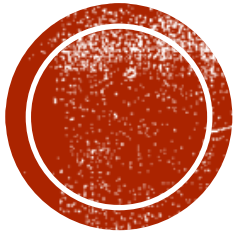
2 risks associated with banks decision to approve loans.

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

2. not likely to repay the loan, then approving the loan may lead to a

financial loss for the company

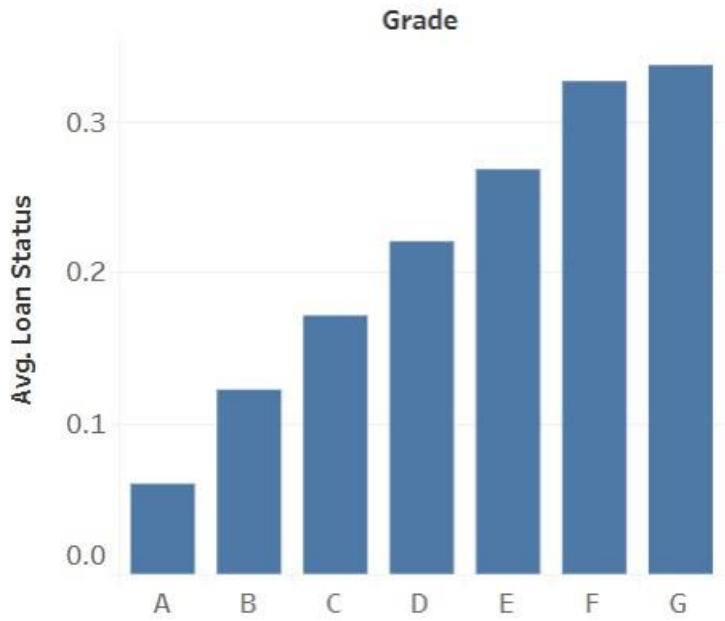
ANALISIS PROCEDURE



This task will be done by using univariate and bivariate analysis of different columns of the dataset.

So, naturally the first step would be to reduce these to a sizable quantity.

OBSERVATION FROM ANALISIS



Now, of the 28 columns we need to find the ones which affect the target variable 'loan_status'. We'll do this by comparing it with other columns and by analyzing each of these columns on their own.

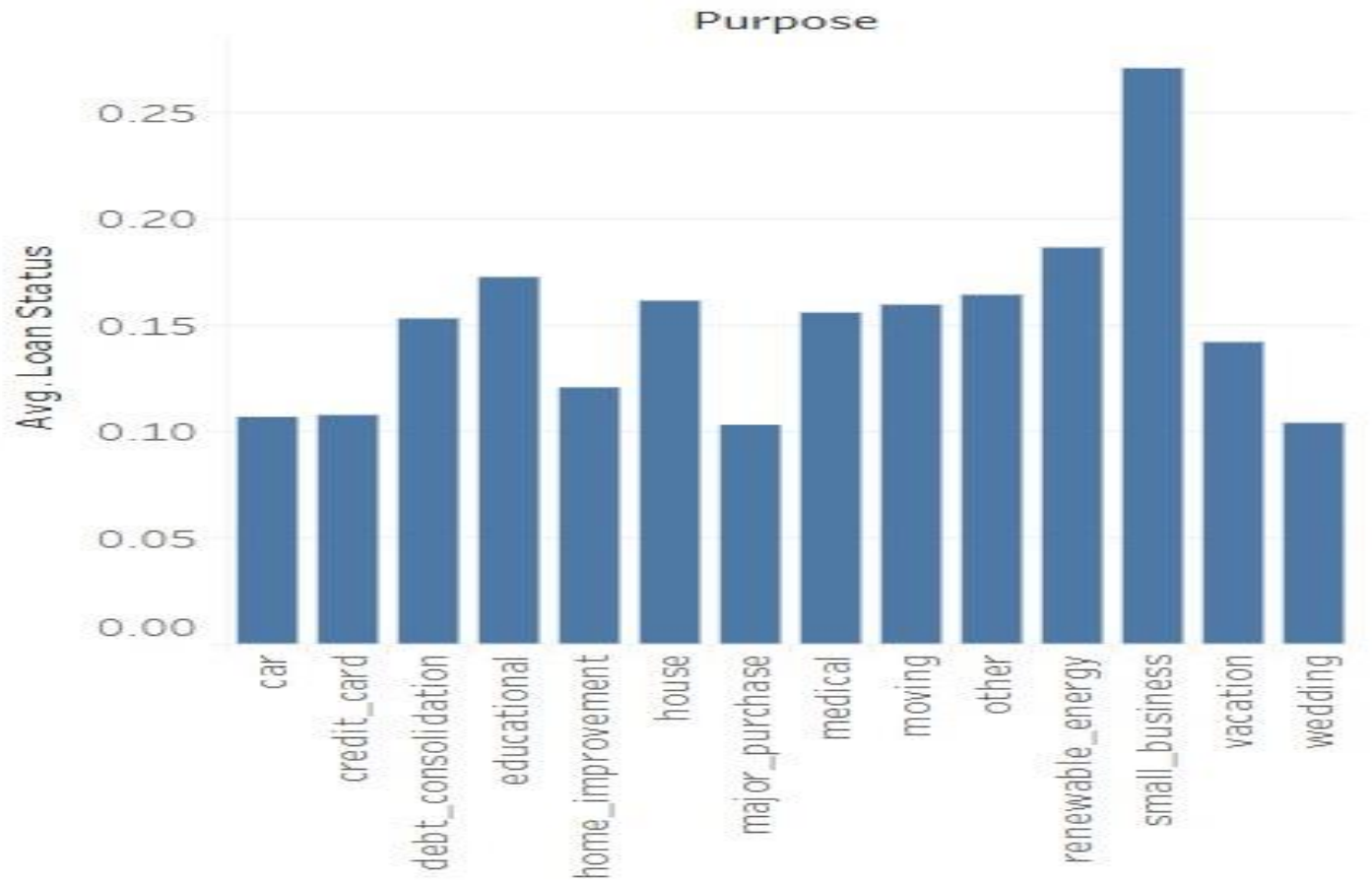
On analyzing our 'loan_status' column we find that the **overall default rate is up to 14%**

To start things off, let's look at all the categorical columns first.

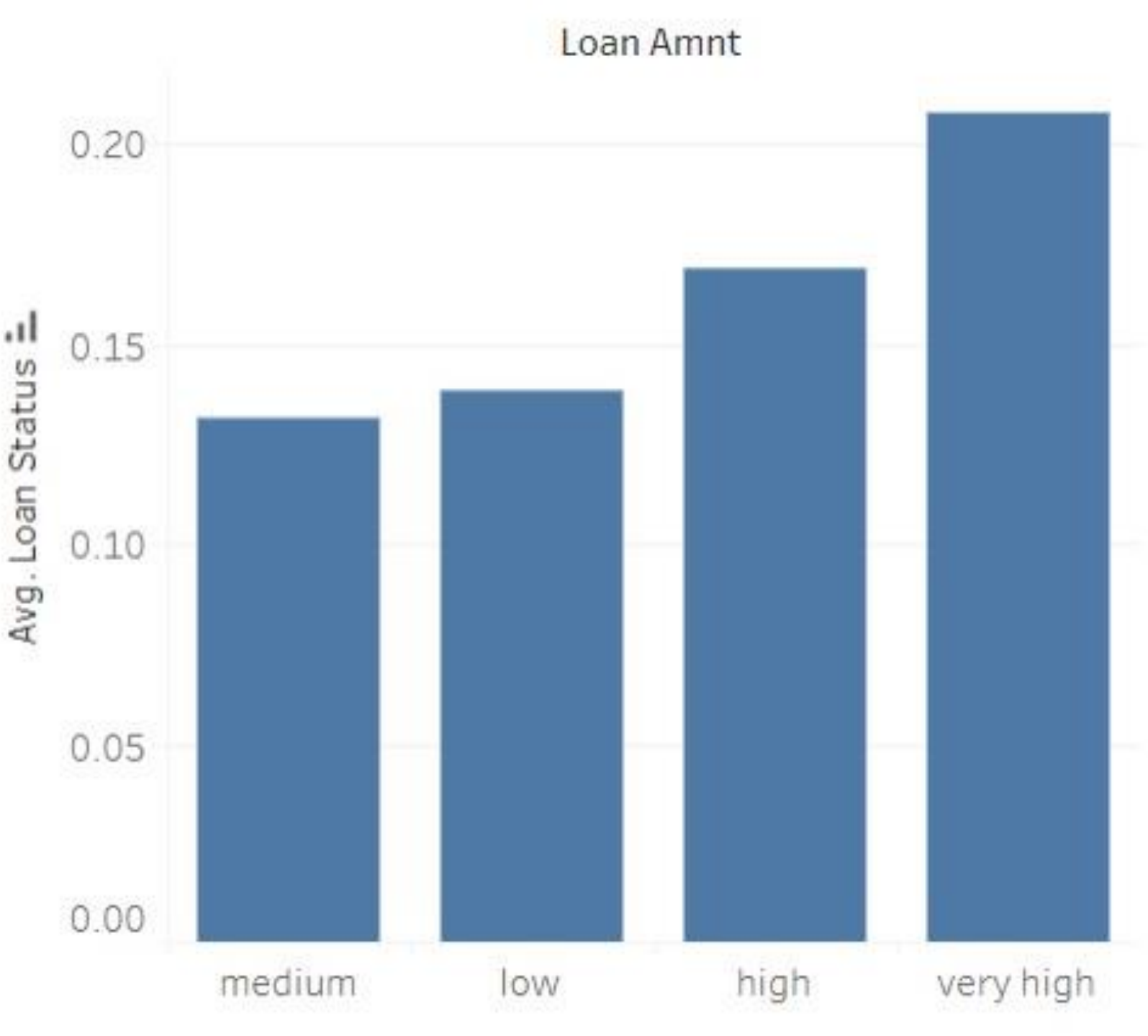
We will plot them against our target variable 'loan_status'.

It can be clearly seen that the risk of loan increases as we go from grade A to G, which is expected because of LC guidelines of assigning the grade.

From this it can be observed that loans of 60 months term tend to default more than 36 months term loans.

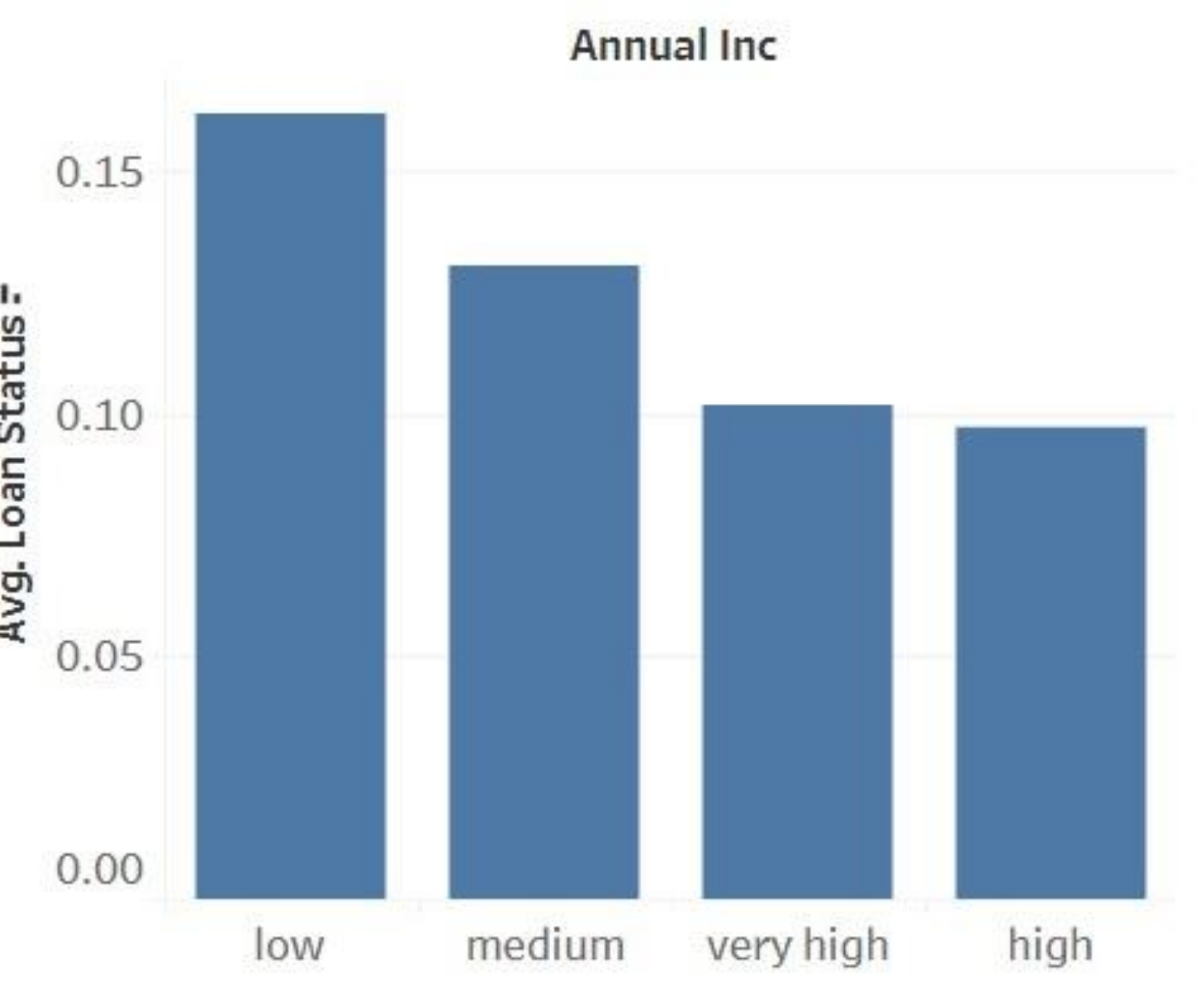


Plotting the purpose of loans shows that small business, debt consolidation, educational and renewable energy loans default more than any other category.

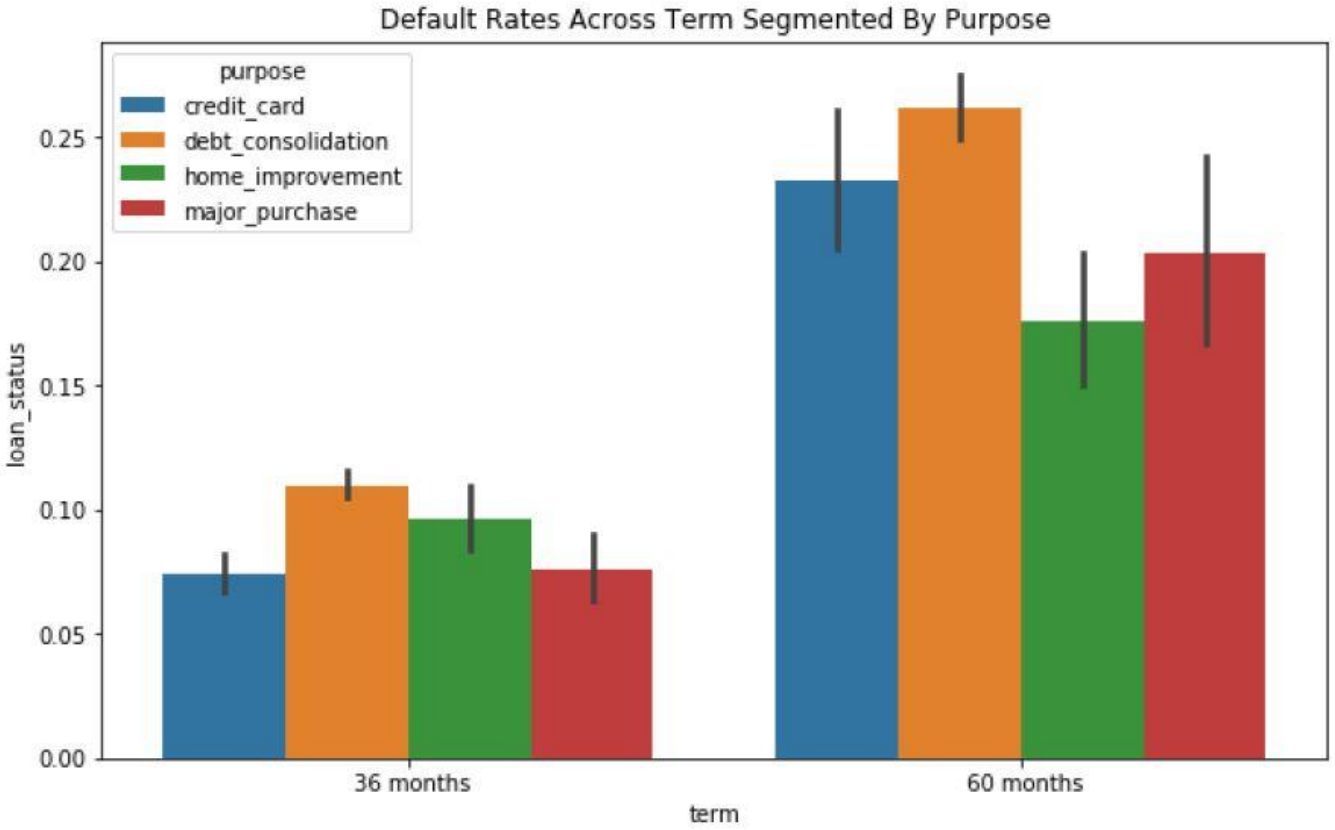


After analyzing categorical variables let's now move on to continuous variables. We will bin these variables into different categories to plot them better.

Loan amount shows that as loan amount increases loan tend to default more.



Annual income seems to inversely affect the default rate. Which is an expected result.

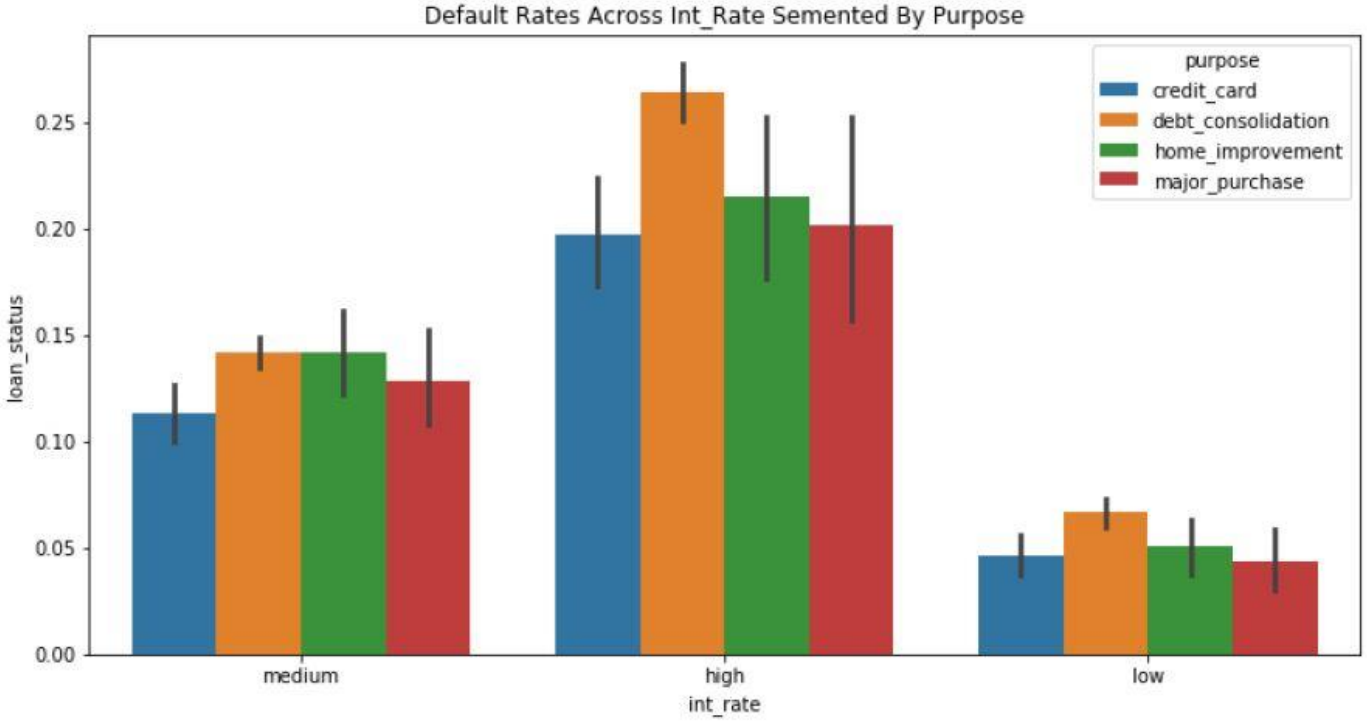


Now let's do some segmented univariate analysis.

Let's analyze various categories with 'loan_status', segmenting it with purpose

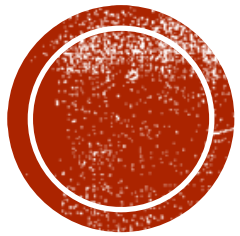
First let's compare it with term.

Interesting,



From above plots, a general trend can be observed that the debt consolidation loans have higher default rates in almost every other category.

CONCLUSION



Segmented univariate analysis doesn't yield much insight but it can be seen that debt consolidation loans tends to default frequently.

The factors to consider while giving loans are:

- Annual Income
- Term
- Interest Rate
- Grade
- Loan Amount
- Purpose