Statistics & EDA

Case Study

Problem Inspection

The given data is of a loan providing company whose purpose for sharing the data was to predict whether the loan given is going to be paid off or being ‘default’. The main objective of the analysis was to determine the conditions and situations that leads to an applicant being charged off or default. The dataset had initially 111 columns with 39716 entries.

Data Cleaning

A lot of these columns are empty so they need to be removed. Taking a threshold of 50%, any column with more the 50% of its rows empty or ‘NAN’ will be removed. This drastically filters the column to 54, which is still a high number.

The column ‘desc’ holds the description of the purpose of the loan as result is not important and can be removed. Some of the columns hold values which reference values that are collected after the loan is sanctioned as a result are little to no use of the analysis and can be dropped.

This leaves us with 28 columns, i.e.

id	38642 non-null int64
member_id	38642 non-null int64
loan_amnt	38642 non-null int64
funded_amnt	38642 non-null int64
funded_amnt_inv	38642 non-null float64
term	38642 non-null object
int_rate	38642 non-null float64

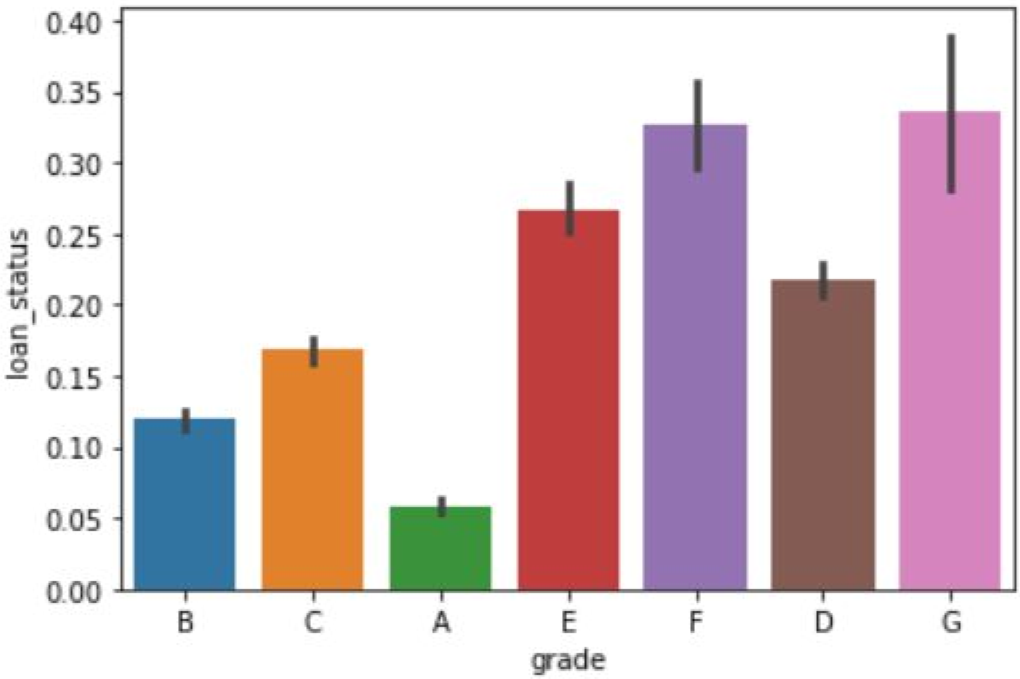
installment	38642 non-null float64
grade	38642 non-null object
sub_grade	38642 non-null object
emp_title	37202 non-null object
emp_length	38642 non-null int64
home_ownership	38642 non-null object
annual_inc	38642 non-null float64
verification_status	38642 non-null object
issue_d	38642 non-null object
loan_status	38642 non-null object
pymnt_plan	38642 non-null object
purpose	38642 non-null object
dti	38642 non-null float64
initial_list_status	38642 non-null object
collections_12_mths_ex_med	38586 non-null float64
policy_code	38642 non-null int64
acc_now_delinq	38642 non-null int64
chargeoff_within_12_mths	38586 non-null float64
delinq_amnt	38642 non-null int64
pub_rec_bankruptcies	37945 non-null float64
tax_liens	38603 non-null float64

Data Analysis

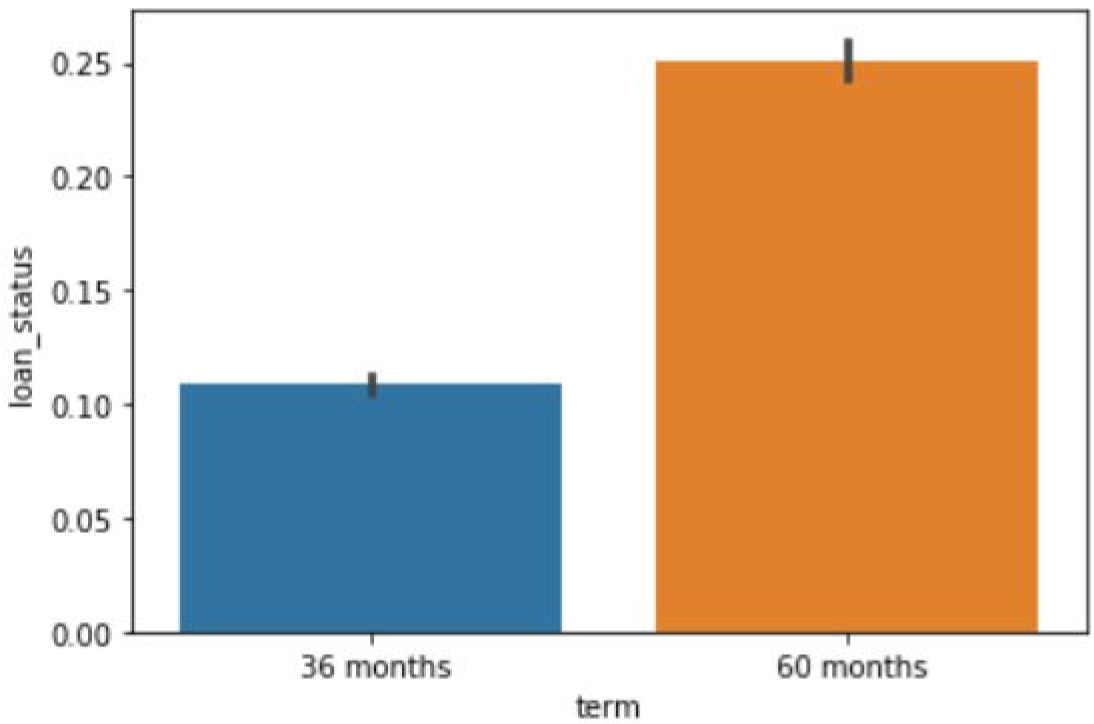
Now, of the 28 columns we need to find the ones which affect the ‘loan_status’ columns. We’ll do this by comparing it with other columns and by analyzing each of these columns on their own.

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

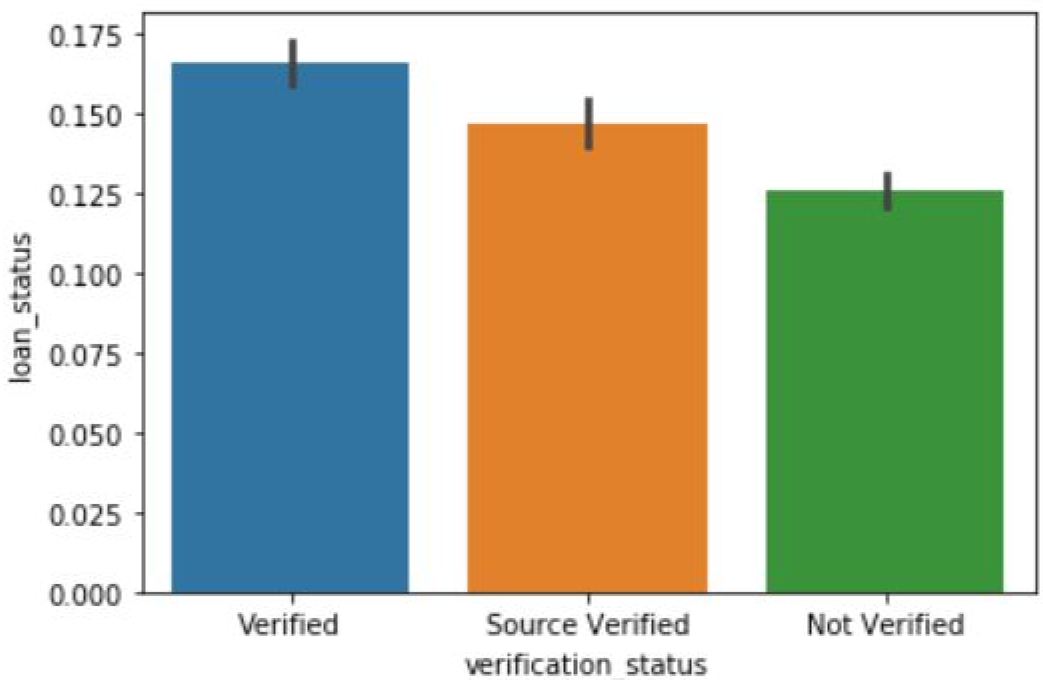
We will plot them against 'loan_status' column.



Clearly the risk of loan increases as we go from grade A to F, which is in compliance with the LC guidelines of assigning the grade.

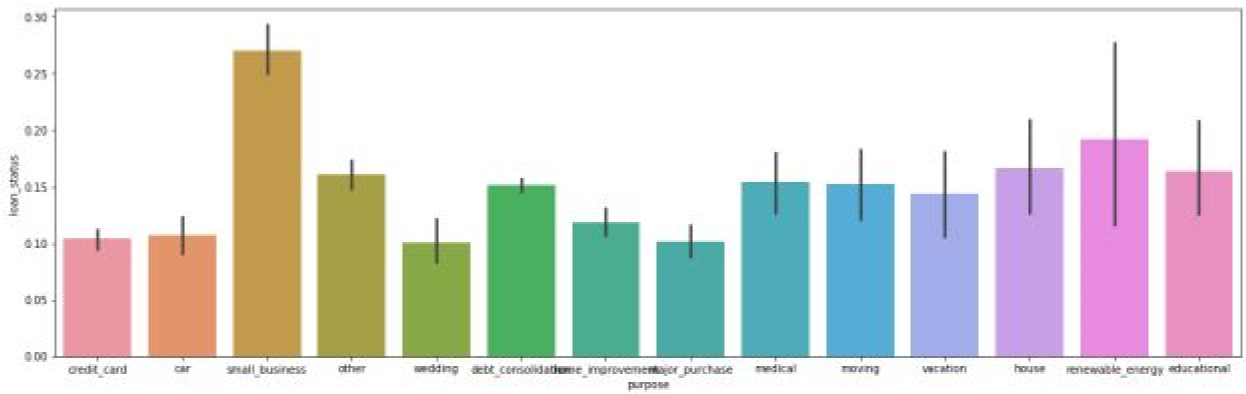


This shows that loans of longer term tend to default more then short term loans.

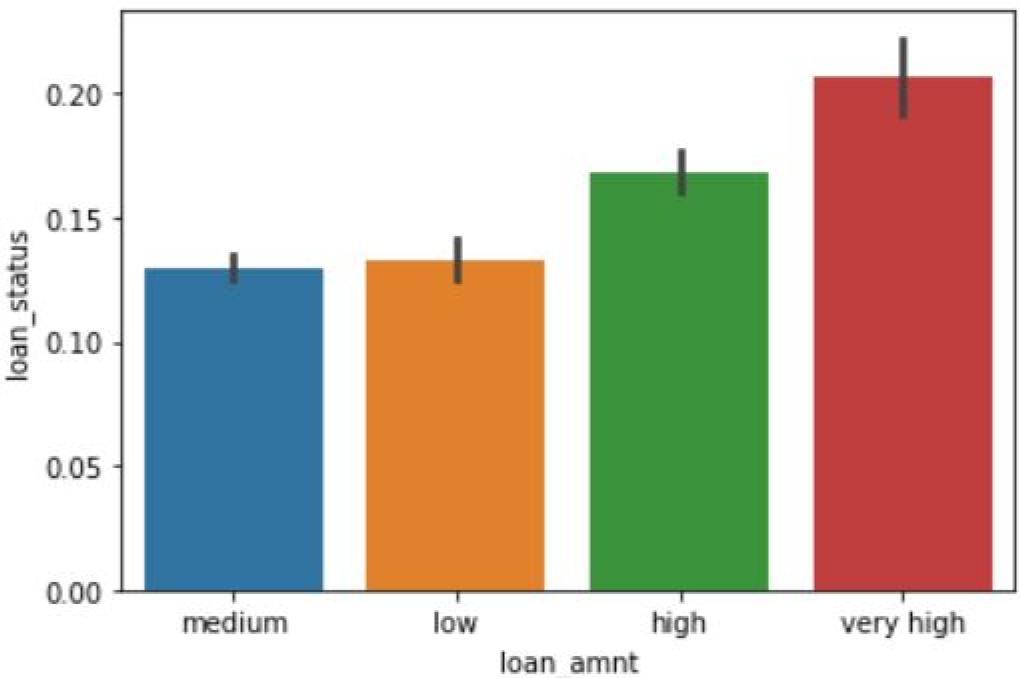


For some reason verified loans tend to default more than non-verified ones.

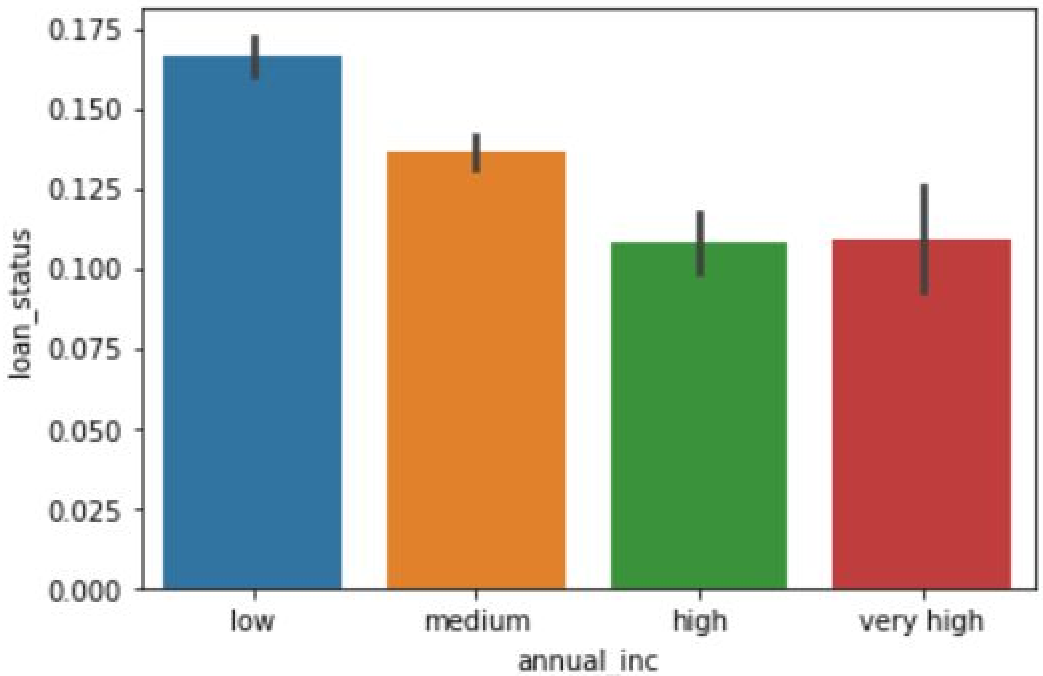
Plotting the purpose of loans shows that small business loans default more than any other category.



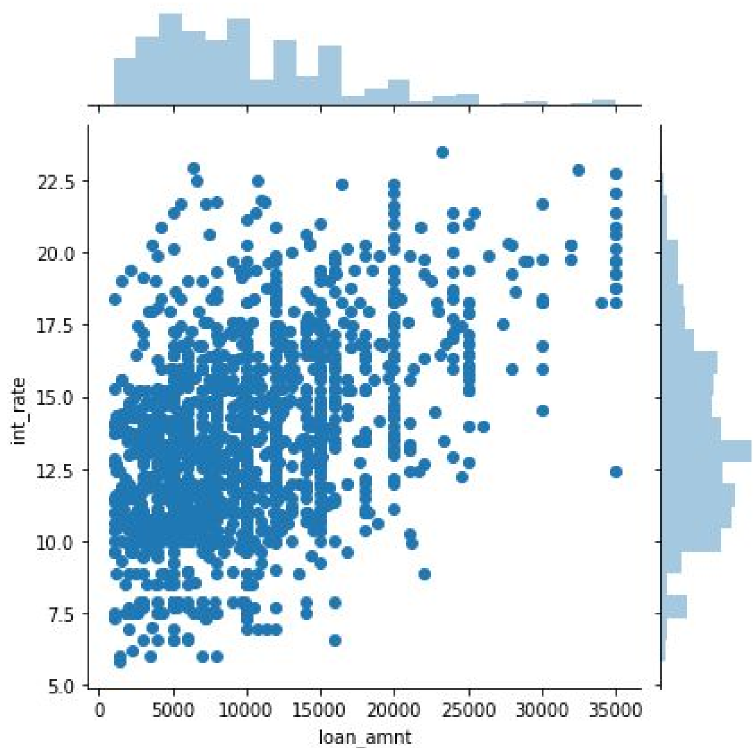
After analyzing categorical variables let's now move on to continuous variables. We will bin these variables into different categories to plot them better.



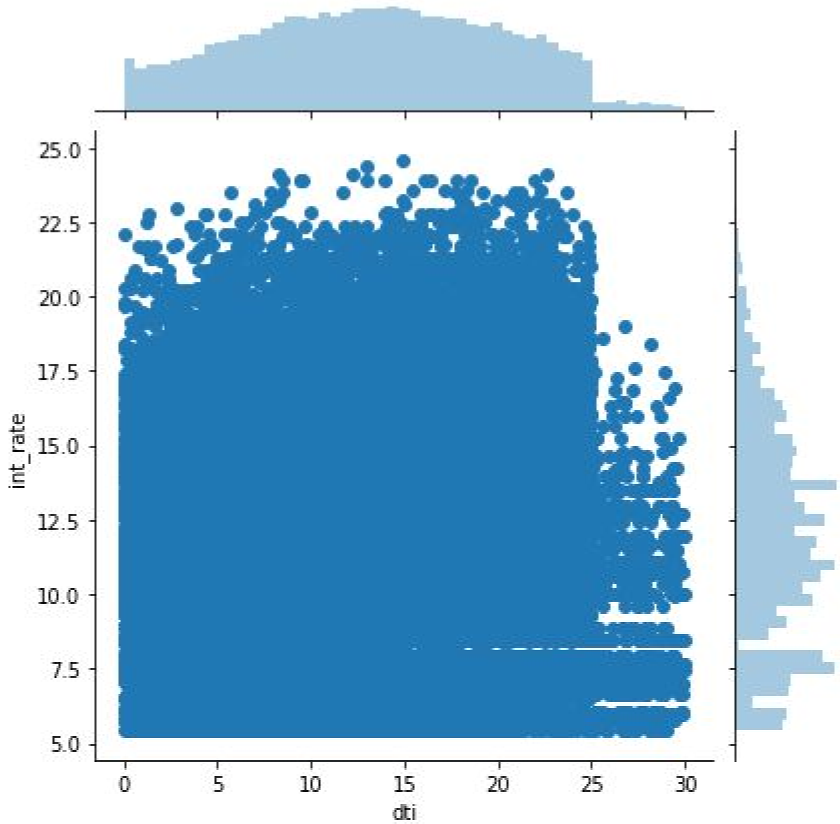
Loan amount shows that high amount loans tend to default most.



Annual income seems to inversely affect the default rate. Which is quite obvious.



High value loans, as well as low interest loans have been extended to those with prior public derogatory records.This practice can be stopped to improve business metrics.



Higher interest rates should be charged for higher dti, but we see spread across all values.

Conclusion

1. Stop – approving loans where amount/income is higher than 30%.

2. Reduce – number of approvals where purpose is small business.

3. Stop – approving high-value loans when revolving line utilization rate greater than 75%.

4. Stop – approving loans to people with prior bad record. Or at least stop approving high-value loans.

5. Start – charging higher interest rates for loans with dti greater than 20.