Lending Club Case Study

Analysis by

Bhaskar M

Exploratory Data Analysis

- EDA helps to manipulate data sources to get the answers, making it easier for Analyst to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

- In this EDA, Loan data set is analyzed to get a basic understanding of risk analytics in banking and financial services and to help the business to minimize the risk of losing money while lending to customers.

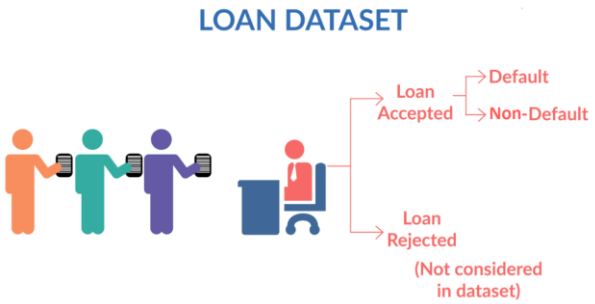
Understanding the Business concept

The given **consumer finance company** is specializes in lending various types of loans to urban customers. The company receives a loan application and to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

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

The data set given contains the information about the past loan applicants and whether they 'defaulted' or not, with the data available we need to identify patterns which indicate if a person is likely to default, which may 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.

Loan Lending concept



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

Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

I. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
II. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
III. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Objectives

Business objective

The aim is to identify the ‘risky’ applicants so that such loans can be reduced thereby cutting down the amount of credit loss. Lending loans to ‘risky’ applicants is the largest source of financial 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.

EDA objective

From the data availability, we need to find the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default is identified and the company can utilize this knowledge for its portfolio and risk assessment.

Data Handling and Cleaning

The given data set consists of about 40,000 records with 111 columns in it.

For better understanding of data we use only the selected columns to find the driving factor of default

Removing columns:

- There are about 56 column having more than 80% of missing values. So these column must be removed.
- After removing the 80% missing columns the desc and mths_since_last_delinq has more number of missing values

  mths_since_last_dealing has about 64% of missing data, by comparing with other columns it has more missing data so removed the column

  Desc column consists of data in improper format(text as well as character) not suitable for analysis.
- **Unique column check** – There are about 9 column with 1 unique values, so including/excluding will create no impact for these columns, for proper dataset we remove those column as well.
- Since our analysis will be at issue of loan , we need to identify and remove the Customer behavioural columns because only the old customer with transaction records have these data and will be less in number, if we included those columns then it may affect the new applicant for loan.

  The below listed column are identified as the Customer behavioural column.

  ['delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'open_acc', 'pub_rec','revol_bal', 'revol_util', 'total_acc', 'out_prncp', 'out_prncp_inv','total_pymnt', 'total_pymnt_inv',

  'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee','last_pymnt_d', 'last_pymnt_amnt', 'last_credit_pull_d']
- Some the columns like, id, member_id which are unique and not used for any analysis. The funded_amnt_inv is the amount that is issued after loan processing , Zip code is unique for each area so may not impact with analysis. Installemnt will be calculated after loan approval so can be removed, Funded amount is almost similar as the loan amount

  unused_columns = ['id','member_id','funded_amnt_inv', 'zip_code', 'url','installment',' funded_amnt']

Data Handling and Cleaning

Dropping the emp_title ,title column

- The Emp_title column consists of the different job role of the employee which is distinct , not uded for analysis
- The Title column consists of the purpose of the loan with explanation, it has so many distinct values, cannot be used for analysis.

Removing the current loan records

In the loan status column we have value as ‘Current’ which indicates that the loan is not yet completed because in future it may be ‘default’ or ‘fully paid’, so cannot be considered and removed from data set.

Imputing the missing values to column

Emp_length – There are about 1033 missing values and it is replaced by mode value(categorical variable)

pub_rec_bankruptcies – There are about 697 missing values and it is replaced by the median value

Removing the unwanted character from the column values.

int_rate – In the interest rate column the % value is removed and the column became numerical and used for detail analysis

term - The term column has string value ‘month’ need to remove for better analysis

Column Conversion

issue_d – The column is converted into date time and 2 separate column month and year is created for in depth analysis

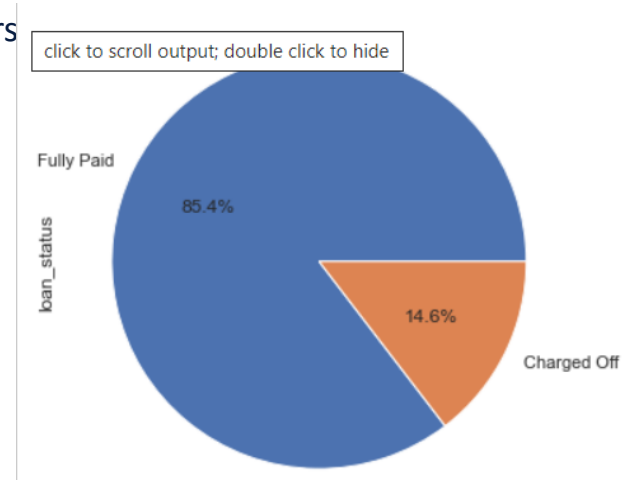
emp_length - The emp_length is the object column gives us the experience of the employee, with the help of data dictionary it is converted to numeric values for in depth analysis

Univariate Analysis

Int_rate – This is evident from the box_plot the median value of interest rate is around 12% and max value is 24% and some outlier as well

Annual Income - From the box plot it is identified there are many outliers and can be treated. Some of the income may be greater than 10,00,000 and there loan amount is around 50,000. There may/may not be an issue but for this analysis removing the records which has income>10,00,000

There are more outliers after removing the income>10,00,000 so assigning the IQR upper bound value to outliers



Loan_amnt - From the loan amount box plot, it is identified that the almost most of the members has the median loan amount less than 10000 and the max amount is 35000

Loan_status - From the pie chart, it is clear that almost 85% of the loan are fully paid and 14% is charged off.

Emp_length - It is identified that the person with more than 10+ experience has brought more loans followed by the least experience. from there the curve starts decling till the one with 9 yrs experience

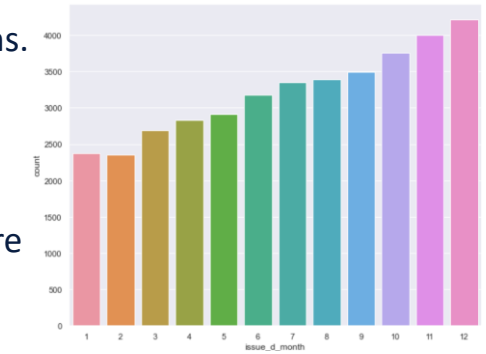
issue_d_year - It is identified there is increase in issuing loans over the year , directly proportional

home_ownership -The more loans are issued for the person who lives in Rent or mortage

verification_status

Purpose - Most of the persons falls under debt consolidation clearly shows the single person brought more loans.

Grade - Most of the loans are assigne under A,B by Lending club clearly shows the loans are issued in large number for low risk applicants



issue_d_month - it is clear from the bar chart the loans are issues mostly on the Nov,Dec. May be more loans are issued to achieve the year end target.

Univariate Analysis

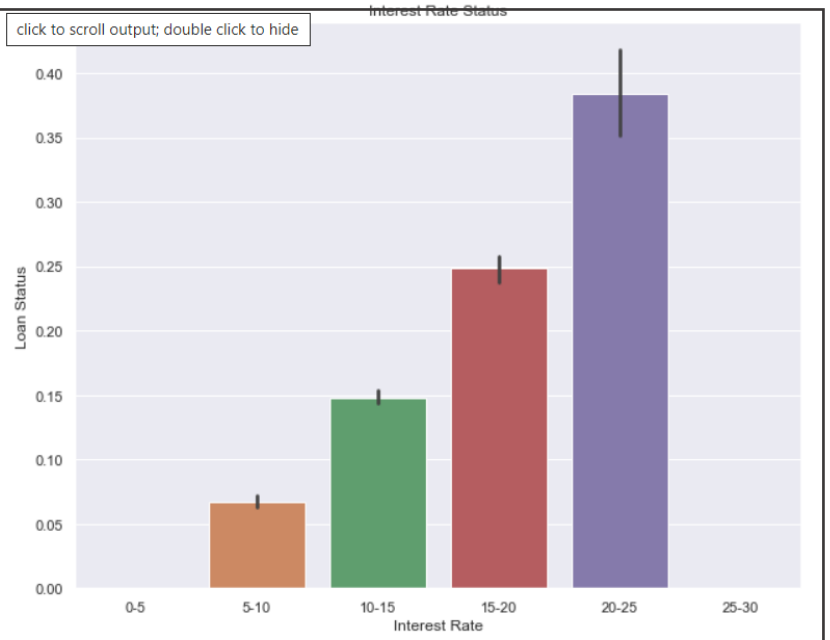
Interest Rate Vs Default Loan Status

Dividing the interest rate column in to 6 parts based on min,median, max value and the graph is

Plot against the default loan status. It is clear from the graph that loan with high interest rate

Caused for more default rate

High interest rates defaults more



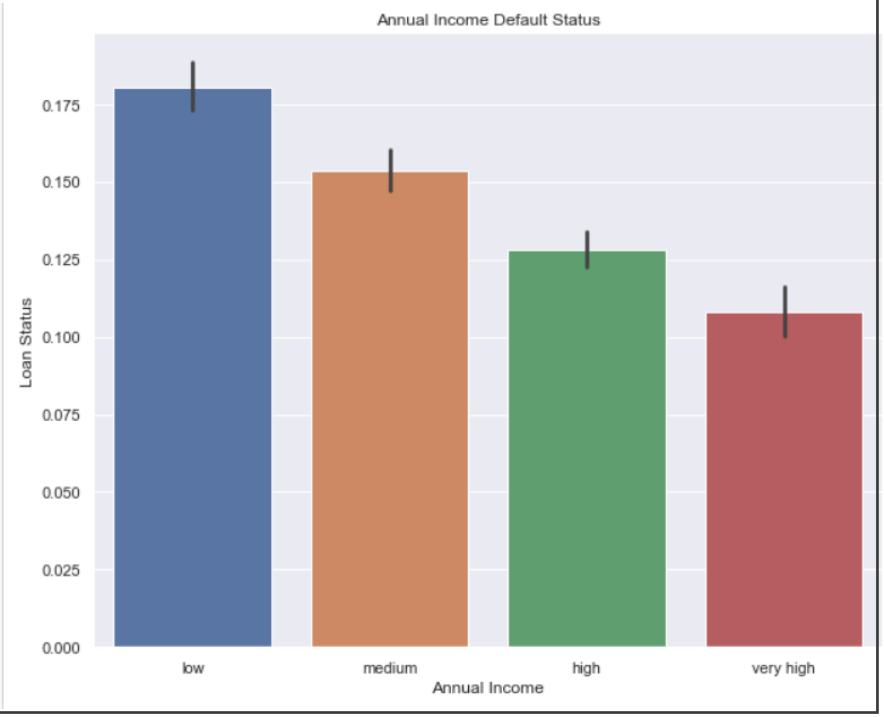
Annual Income Vs Default Loan Status

The Annaul Income divided into 4 parts based on min, 25%, Median, 50%

and max values as low, medium,high, very high for easy analysis.

Plot against the default loan status. It is clear from the graph that one with **low income are**

High chance of defaults followed by medium income .



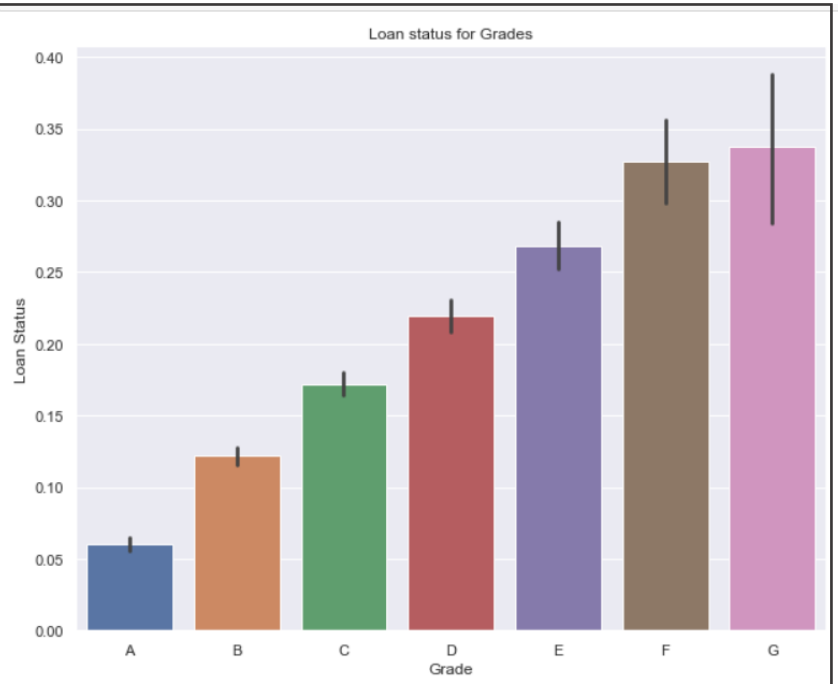
Univariate Analysis

Grades Vs Default Loan Status

The Grades are assigned by LC to identify the risk status ,as already knows

most of the loan are issued for the low risk clients. This shows more clear that the as

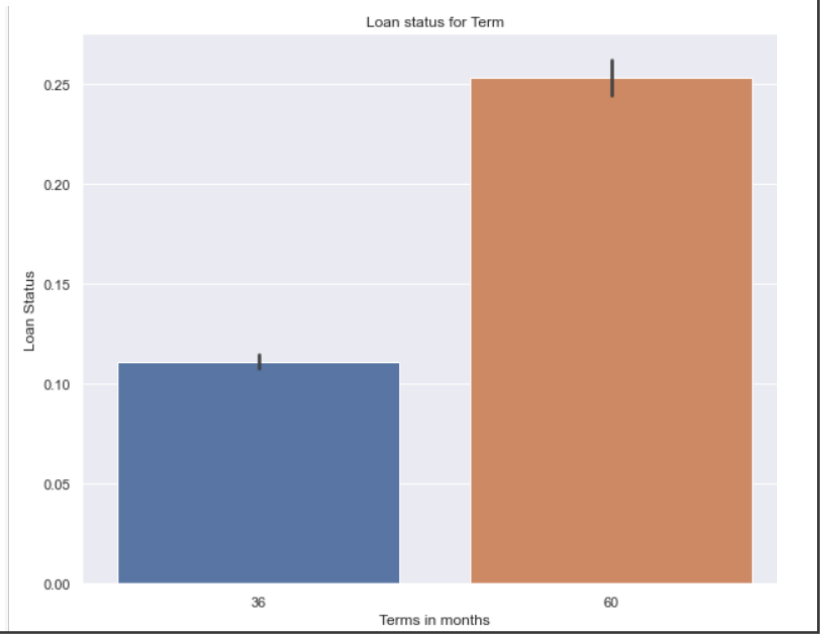
Grade increases the default rate also increases



Term Vs Default Loan Status

There are 2 terms of loan 36 and 60. Almost 75% loans are issued for 36 month term period

and most of those are fully paid, **increase in term period increasing the default rate.**



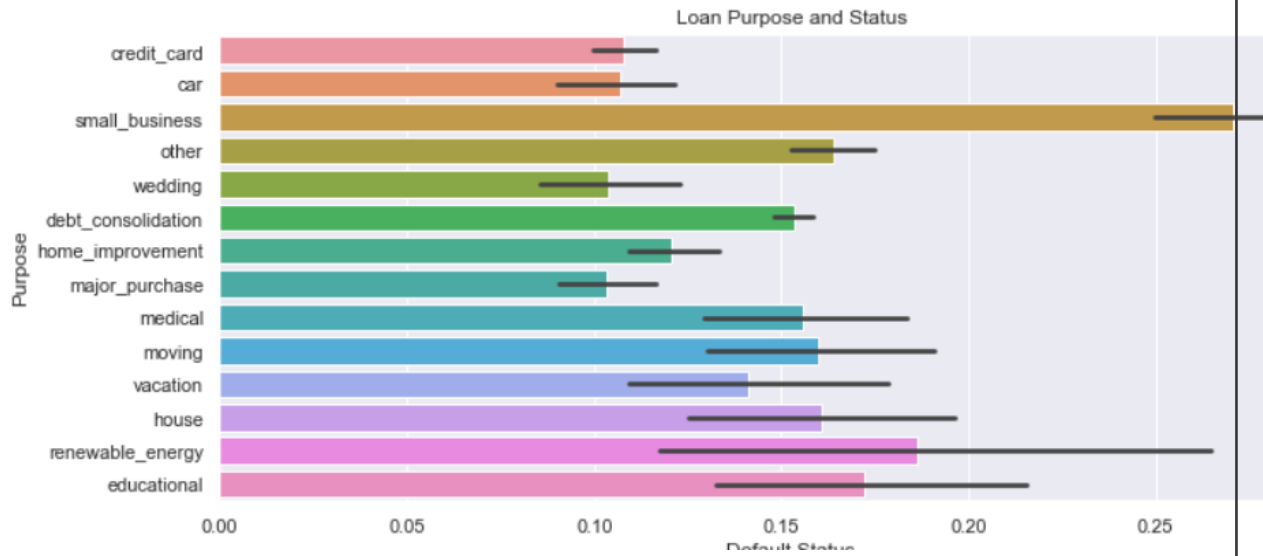
Univariate Analysis

Loan Puposes Vs Default Loan Status

Since the most number of loan was issued for Debt consolidation ,

From the plot it is evident that '**Small Business**' purpose loan is mostly

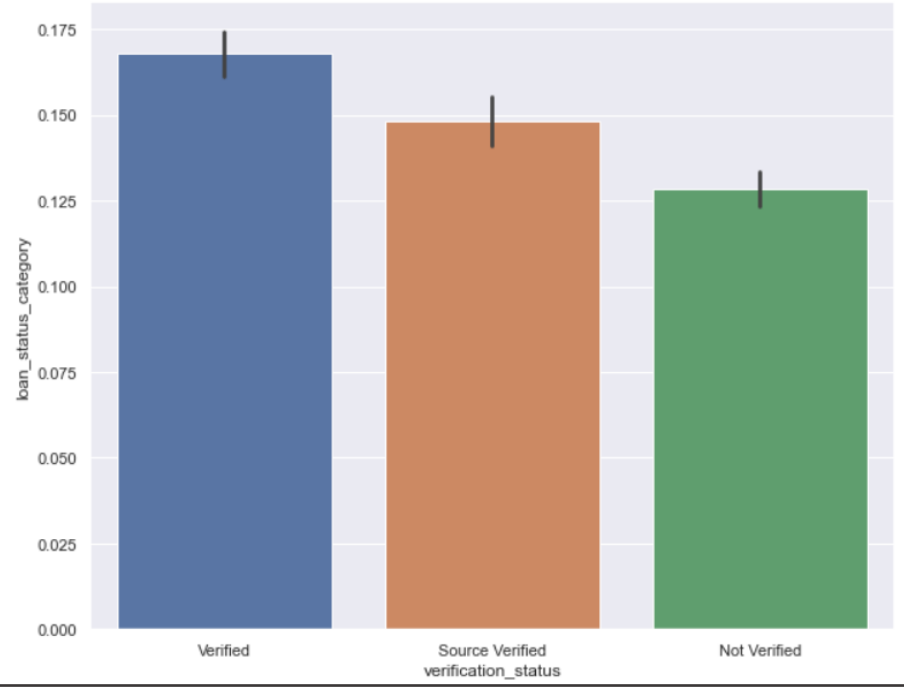
Likely to be defaulted followed by the renewable energy and other.



Verification status Vs Default Loan Status

The Most loan was given to Non verified followed over other 2 status.

But from the graph it is shown an opposite trend like the most of the **verified loans are defaulted**

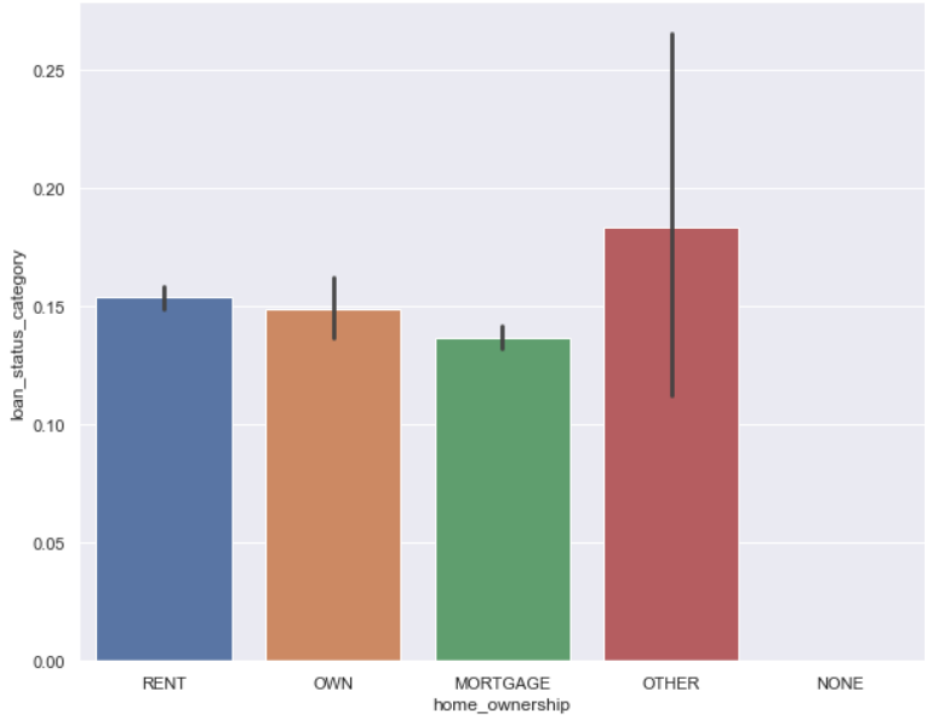


Univariate Analysis

Home Ownership Vs Default Loan Status

The Home ownership against the loan follows the similar pattern.

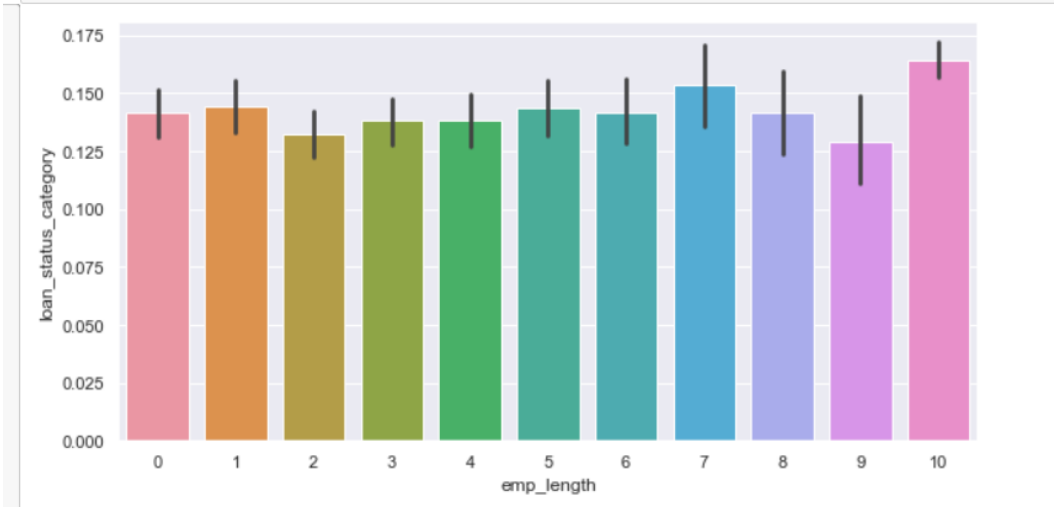
No inference.



Employee Experience Vs Default Loan Status

This shows the normal trend , **no inference** found.

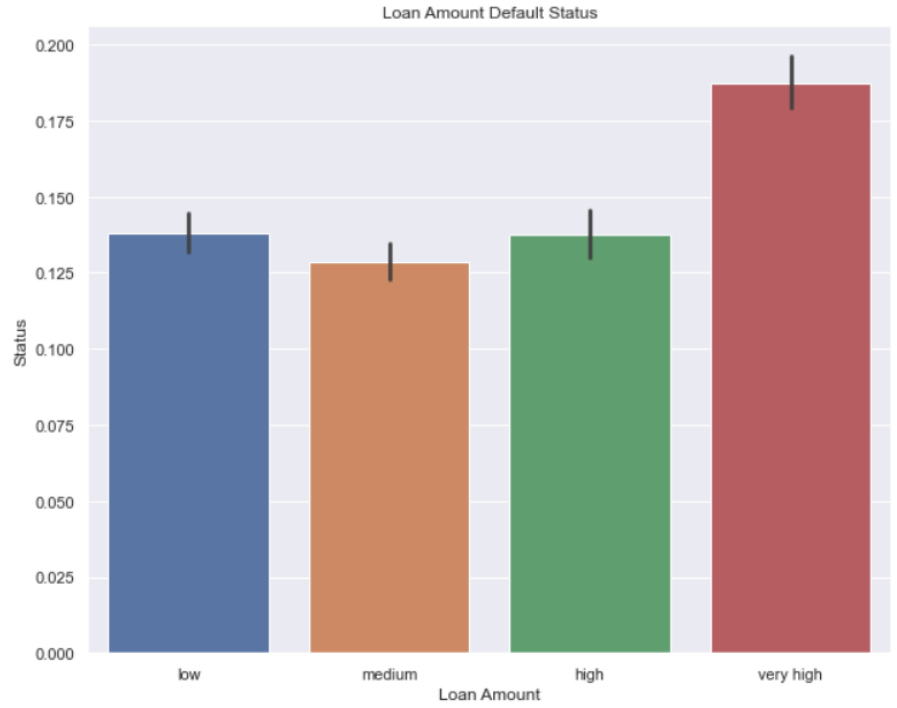
The most number of loans are issued to the employee experience over 10 years so it default rate is high but not very high compared to other.



Univariate Analysis

Loan Amount Vs Default Loan Status

The loan amount is classified to low, medium, high, very high based on the 25%, median, 50% ,75% for easy analysis. It is clearly evident from the graph that **very high loan amount causes more default**



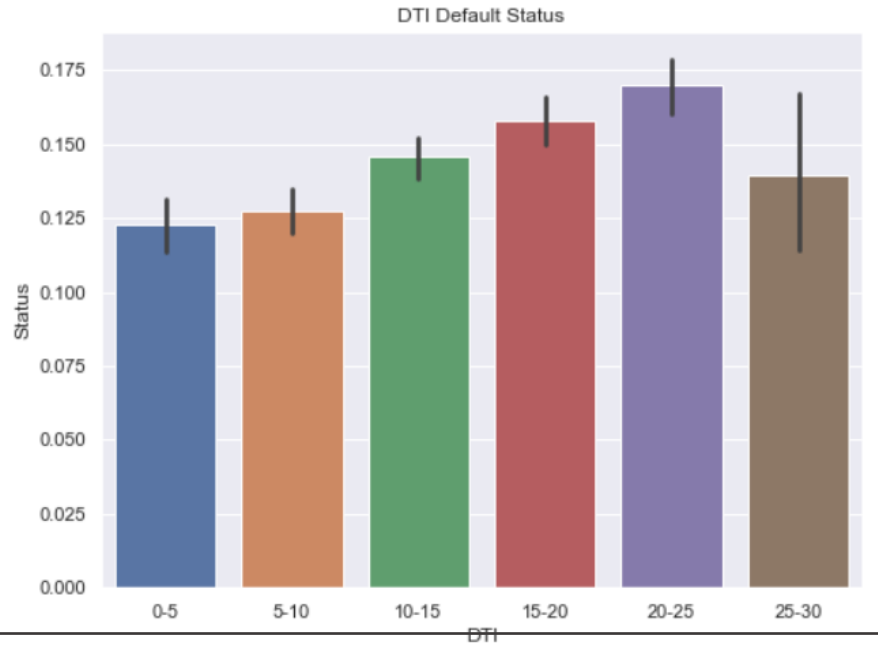
DTI Vs Default Loan Status

The DTI is divided to 6 bins as the min value is 0 and max valye is less than 30.

The more loans are issued to the person with DTI between 10 – 20.

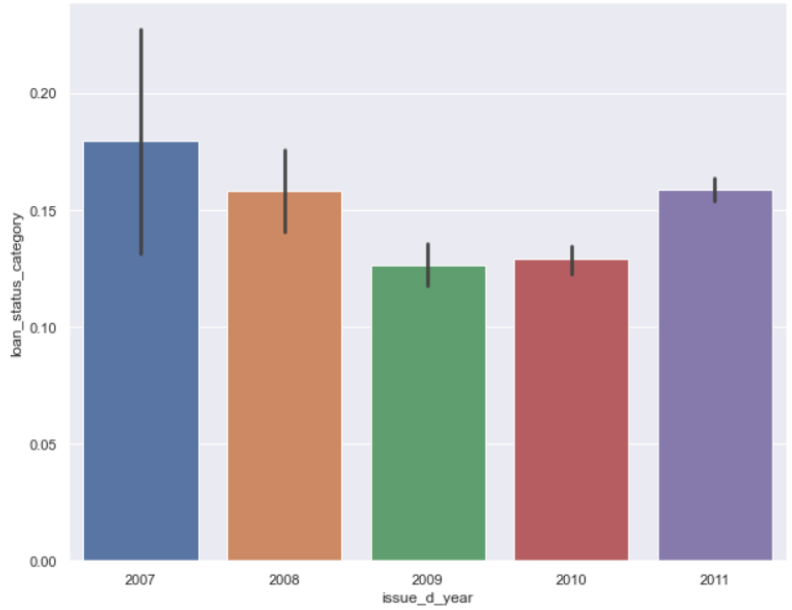
Only 623 applicants are comes under the 25-30 ratio, and also shows high default rate Compared to other DTI bins.

It is clearly evident that the **Default rate increases with increases in DTI**



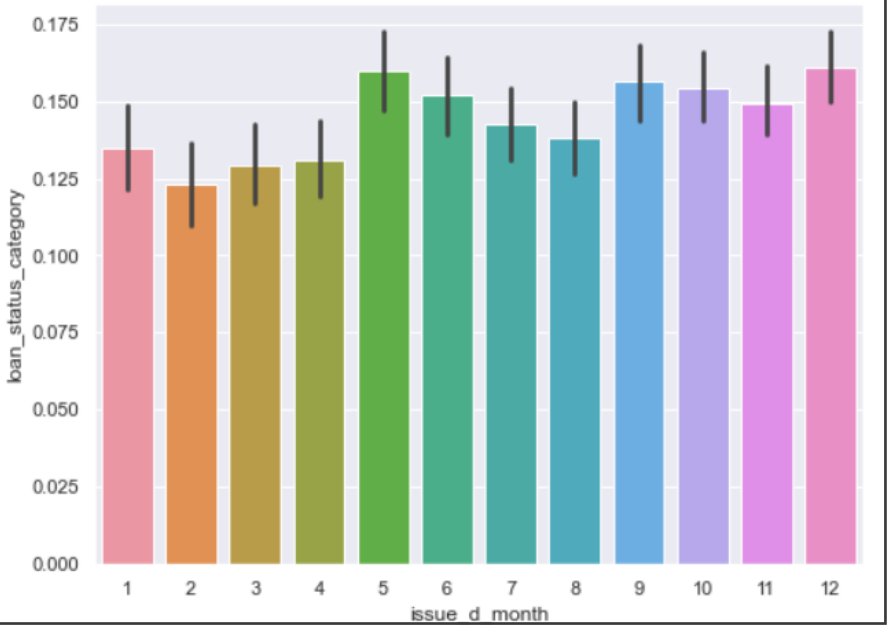
Univariate Analysis

Loan Year Vs Default Loan Status



Here from the graph it shows the default rate is high in 2007 , but very less number of loans Are issued in the year compared to others but the default is high because the lending of Loan start from the year**. No much inference.**The number of loans increased over the years

Loan Month Vs Default Loan Status



From the graph, it is evident that most default rate at the Dec month followed by the May month. There is almost regular pattern of default rates over the month

No inference found.

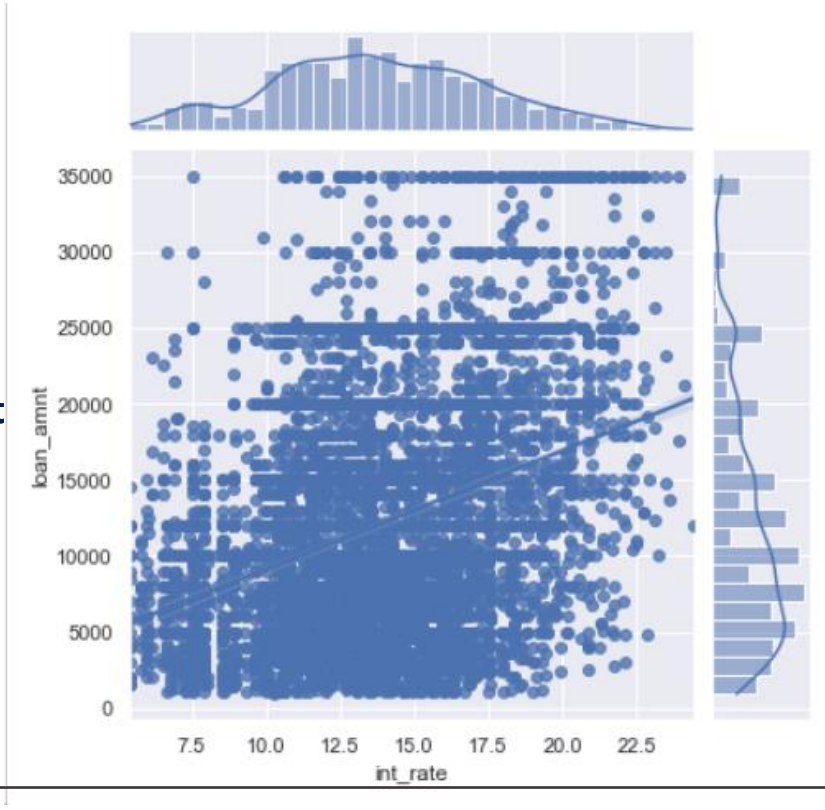
Bivariate analysis

Verifciation status vs Loan amount

From the graph it is identified that the high loan amount was given only for the '**verified**' status applicant and **the high loan amount causes the high default rates.**

Loan amount Vs Interest rate

From the scatter plot it is identified that there is **low positive co-relation** between the Loan amount and interest rate. Ideally the **loan amount increases with increase in interest Rate.**

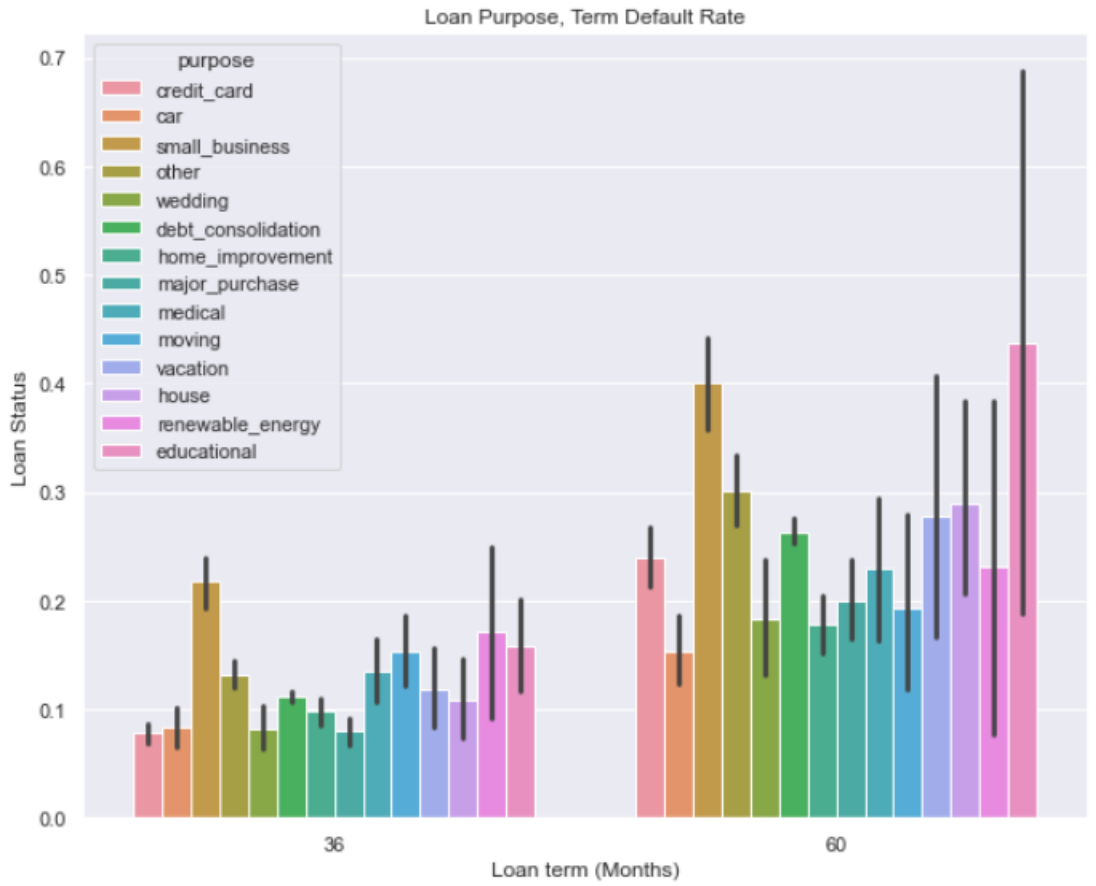


Bivariate analysis

Loan Purpose Vs Term for default rates

In the both the terms (36,60) the Loan purposes of the **'Small Bussiness' is defaulted highly.**

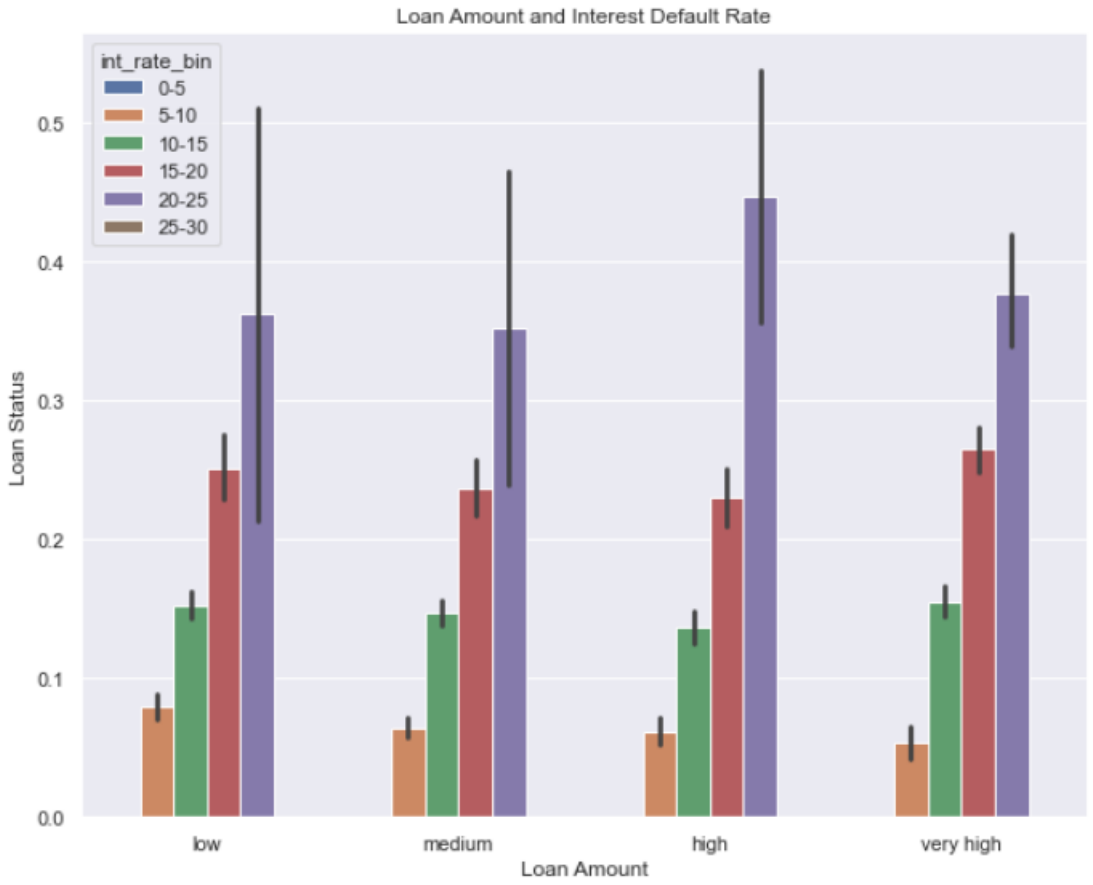
But the educational loan is defaulted high in 60 months but in 36 months the default rate is low.



Bivariate analysis

Loan Amount Vs Interest for default rates

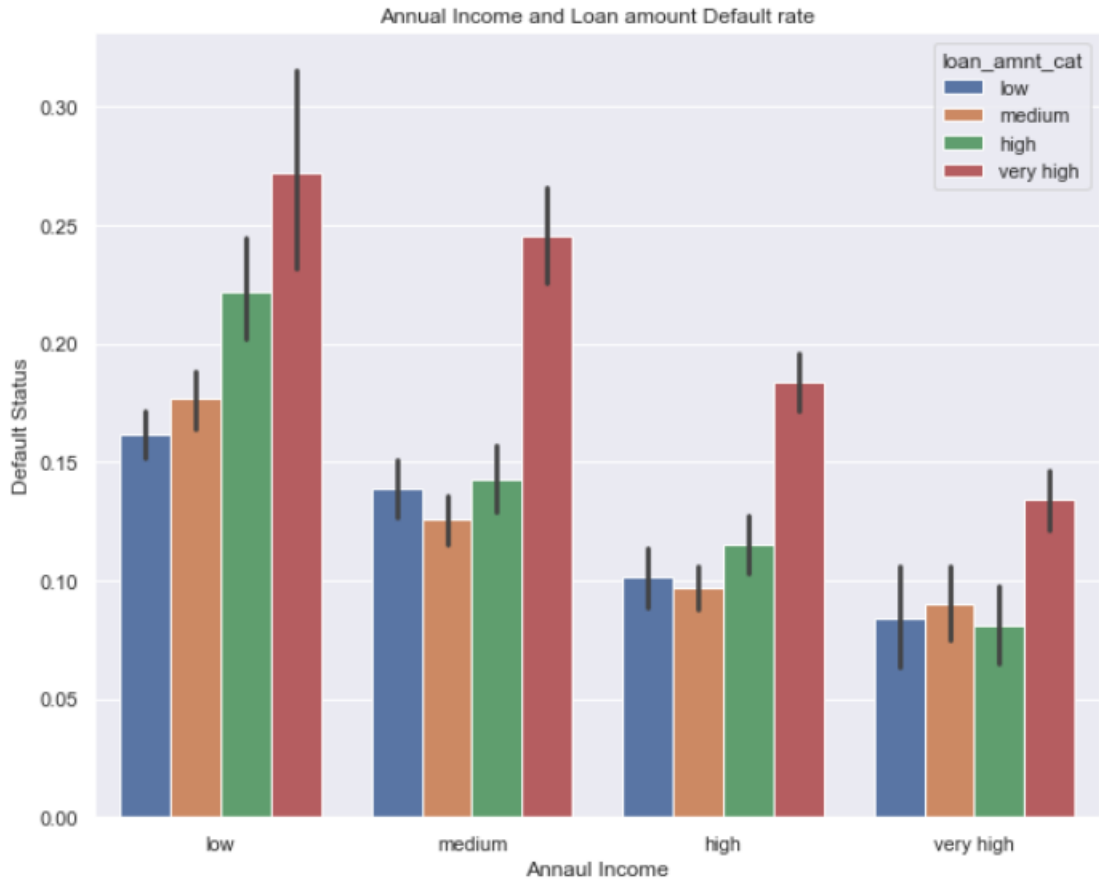
From the graph it is identified clearly that **the loan amount with high interest rate are likely to be defaulted** more.



Bivariate analysis

Loan Amount Vs Annual Income for default rates

From the graph it is identified that **higher Loan amount is more likely to be defaulted across the all the category of annual income** (Low, medium,high, very high) and it follows the trend like the defaulters are more in the Low income category compared to overall.



Observations and Recommondations

1. It is observed clearly that the loan amount with high interest are likely to be defaulted.
2. The loan issued to ‘Small Businness’ is highly defaulted compared to other purposes.
3. The DTI ratio is increased with increase in default rate.
4. Higher the loan amount across all the category of the Annual income is likely to be default and particularly higher default rate in low income category.
5. It is observed that the most of the ‘Verified’ status of the loan are getting defaulted.
6. Most of the loans with the period of 60 months are likely to be defaulted.
7. The Grade ‘G’ is more likely to be defaulted. It follows a linear pattern as grade increase default rate increases.