Lending Club Case Study

Submitted by-
Santanu Mahata

Steps

Data Understanding

Data Cleaning

Data Visualization

Conclusion

Data Understanding:

- From the loan data set we are going to analyze the attributes which leds to the tendency of default
- The total number of different features or attributes are 111
- Total nos. Of entries are 39717
- To avoid the default case it has to pre analyzed or predicted before approving the loan to avoid credit loss and it also has to keep in mind that no eligible candidates are debarred from getting the loan which will affect the business aspect of the company.
- Our point of interest is the persons whose loan status is defaulted to analyze the sublime reason of default.

Data Understanding:

- The Id and member_Id columns are unique and have no such importance in our data analysis.
- Out of all the columns the following columns seem to have impact on the tendency of default.
- loan_amnt','funded_amnt','funded_amnt_inv','term','int_rate','grade','sub_grade','emp_title','emp_length','home_ownership','annual_inc','verification_status','loan_status','purpose','dti'
- Some data are of object datatype line interest rate which need to be converted into numeric type for analyzing the data.

Data Cleaning:

- As the following columns are of our interest we will only keep these columns and neglect the others for better visualization.
- loan_amnt','funded_amnt','funded_amnt_inv','term','int_rate','grade','sub_grade','emp_title','emp_length','home_ownership','annual_inc','verification_status','loan_status','purpose','dti'
- After extracting the data our next goal is to look the data set which leads to the default cases. For that we have filtered out the entries in which the loan status is Charged Off.
- We come up with 15 columns and 5627 entries

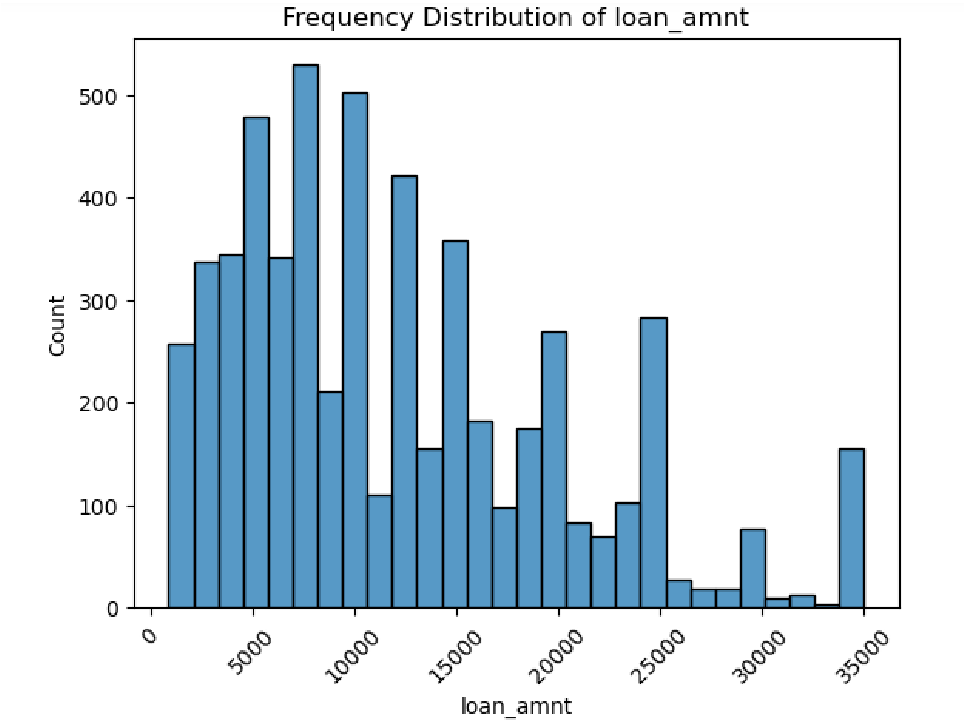
Data Cleaning:

- Next step is to check the data type of each columns. The interest rate column is of object data type which is converted to numeric data type for analysis.
- There are some data missing in the ‘emp_title’ data column. This may be the cases where the applicants are unemplyed or the data may be missing. To analyze it lets look into the emp_length column. The entries of emp_title in which emp_length is the cases where it is highly probable that the aplicants are unemployed. So we have replaced those entries by unemployed. We have found 216 of such entries.
- The interest rate column has % sign in the entries and data type is object. We removed the % sign and converted the data type of 'int_rate' column from object to float for better analysis.

Data Visualisation:

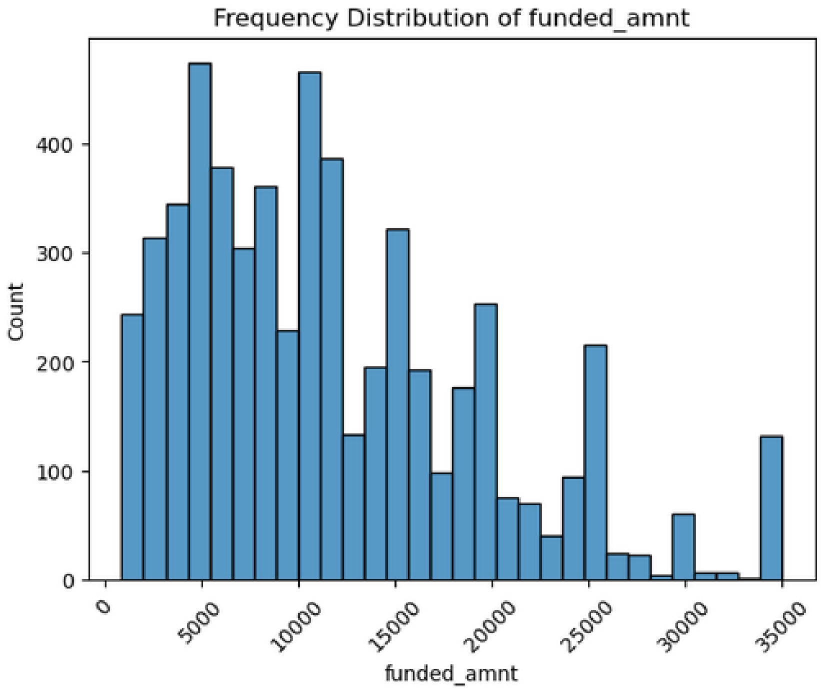
- For visualising the data for which most of the default cases has been arised, we have segregated the sorted columns in to two parts i.e. Categorical Columns and Numerical Columns.
- The entries in the categorical column are 'term','grade','sub_grade','home_ownership','verification_status','purpose', 'emp_title','emp_length'
- The entries in the numerical column are 'loan_amnt','funded_amnt','funded_amnt_inv','int_rate','annual_inc','dti'
- We have seperately done univariate analysis for each of the categorical and numerical columns. Lets observe the plots one by one.

Data Visualisation:



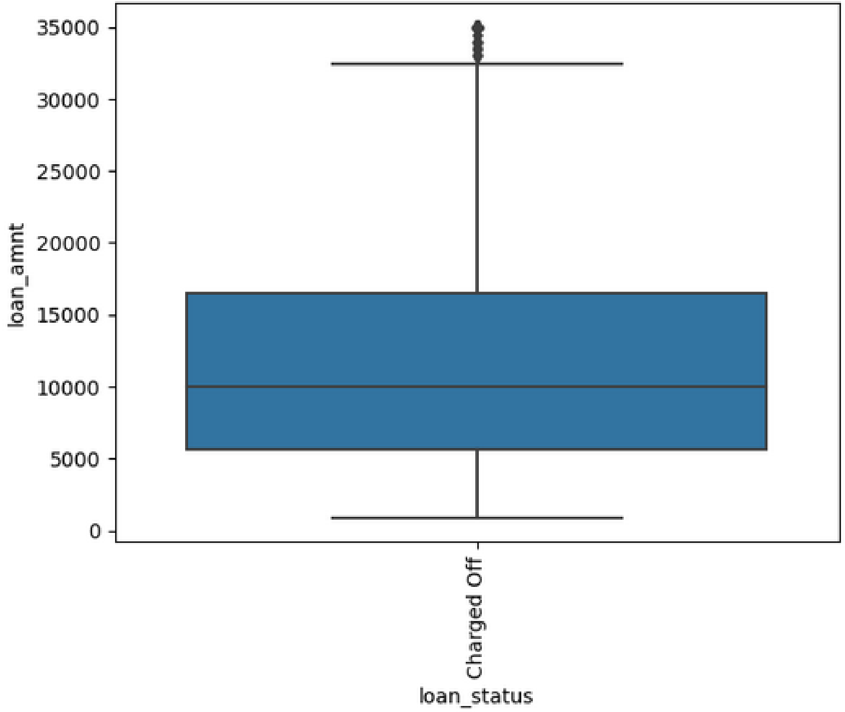
- The loan_amnt histogram shows a positively skewed nature.
- Most default cases has been arised for loan amount less than 25000

Data Visualisation:



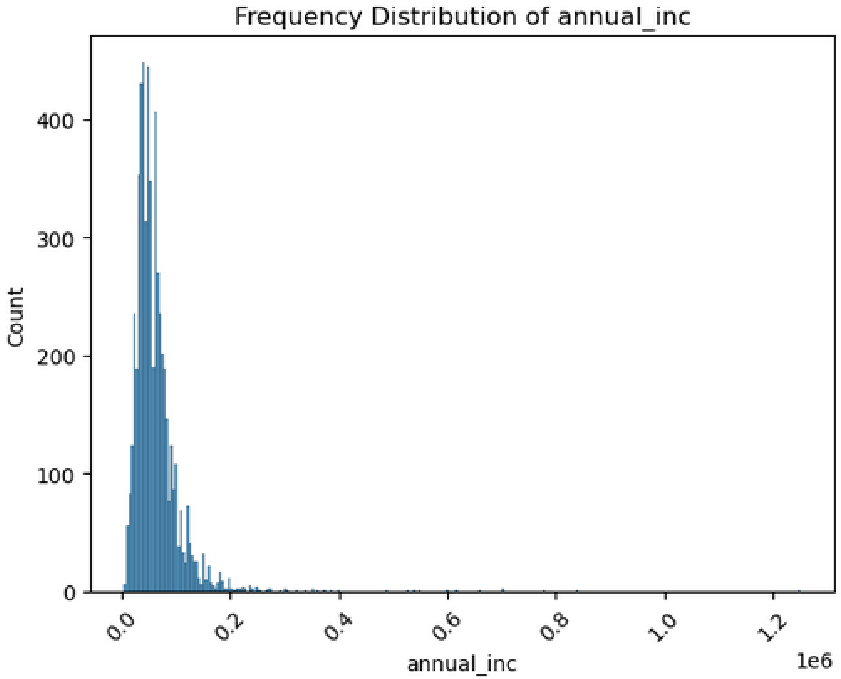
- Like the loan_amnt the funded_amnt histogram also shows a positively skewed nature.
- Most default cases has been arised for loan amount less than 25000.

Data Visualisation:



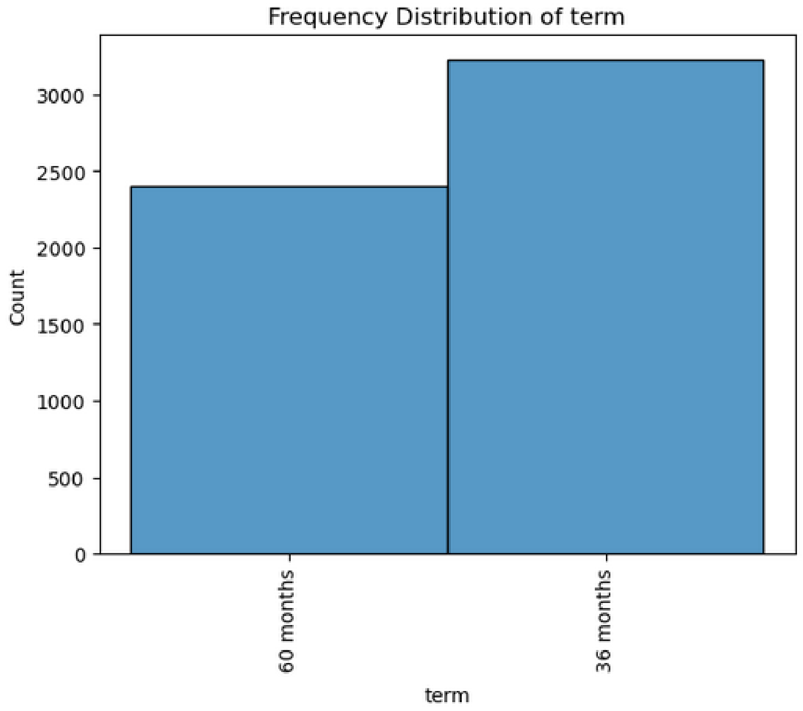
- The loan_amnt box plot shows that 75 percentile of default cases are of the applicants who have applied for a loan less than 20000
- Among which 50 percent cases are of loan amount 10000 or less.

Data Visualisation:



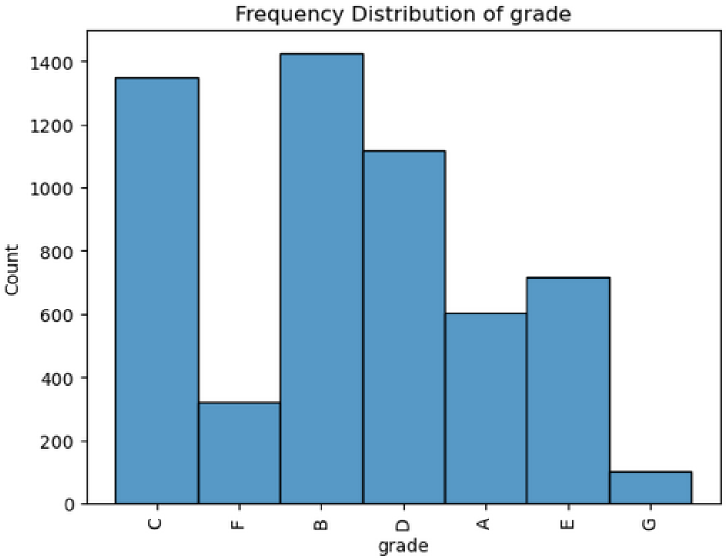
- From the Annual income histplot it has been cleared that the most default cases are of applicants with family income less than 1 lakhs.

Data Visualisation:

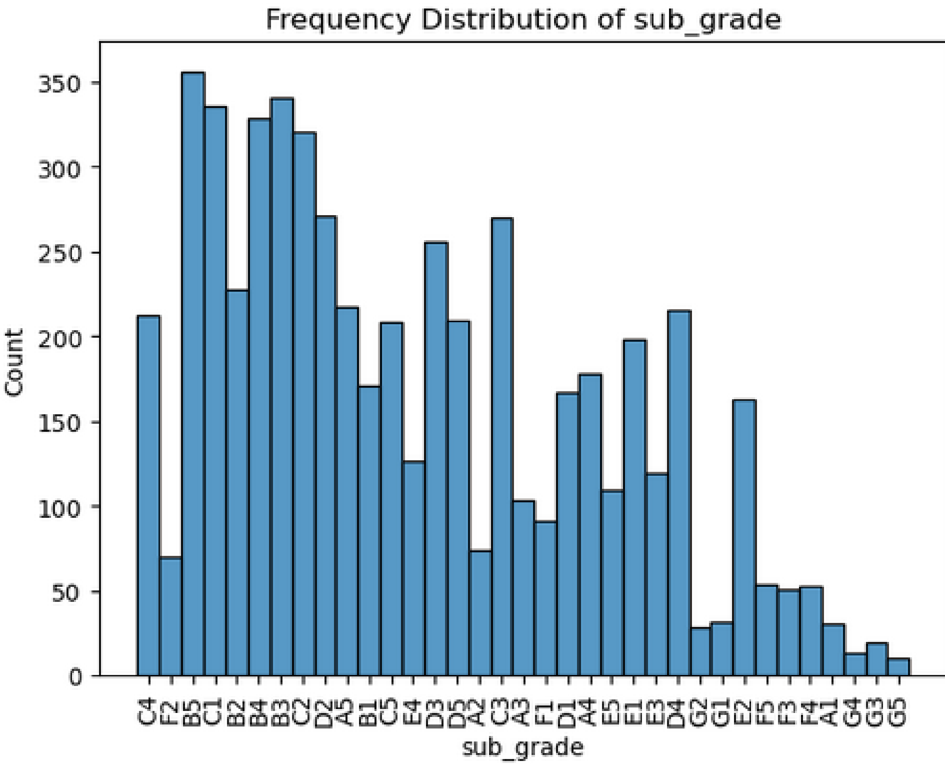


- More default cases has been arised for short term loan.

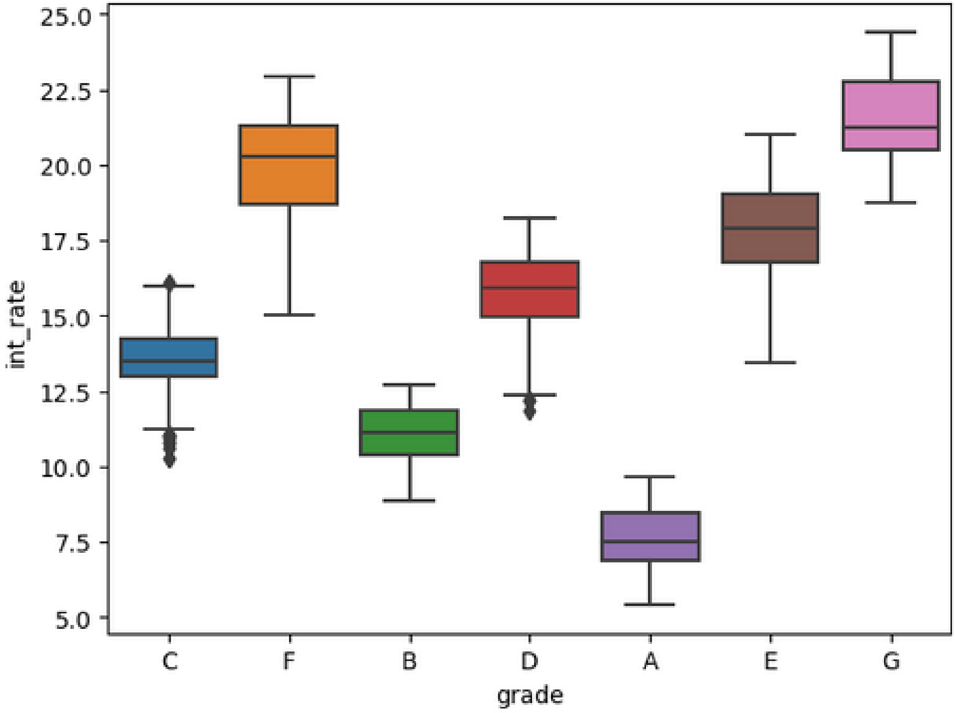
Data Visualisation:



- The most default cases are from group A,B,C,D and E
- Among which B3,B4,B5,C1 and C2 have the maximun default

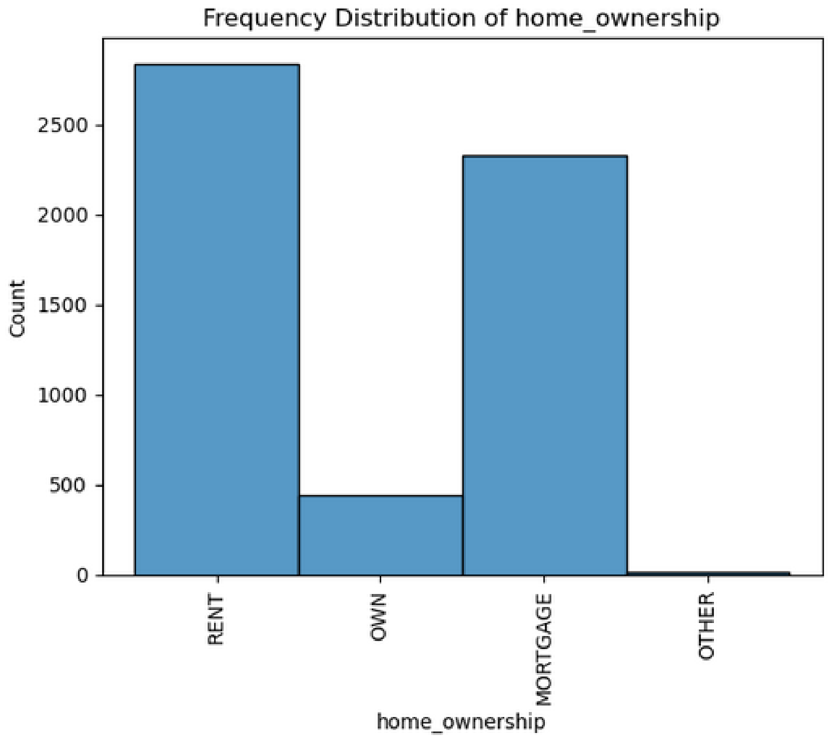


Data Visualisation:



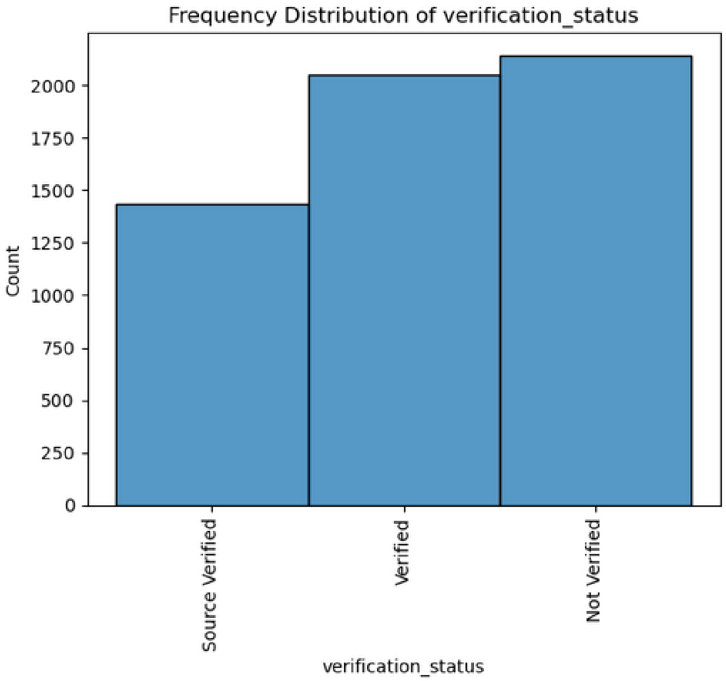
- In the above slide we have seen that the most default cases are from the group A,B,C,D and E
- The interest rate of C,D and E is comparatively higher than that of A,B
- F and G has the highest interest rate

Data Visualisation:



- From home ownership histogram plot it is cleared that the applicants who are living in a rent house or in mortage are liable to pay their rent on time so it is more probable that they are failing to repay the loan.
- The graph shows this observation. The count of defaulter with home ownership Rent or mortage is more than that of applicants with own accomodation.

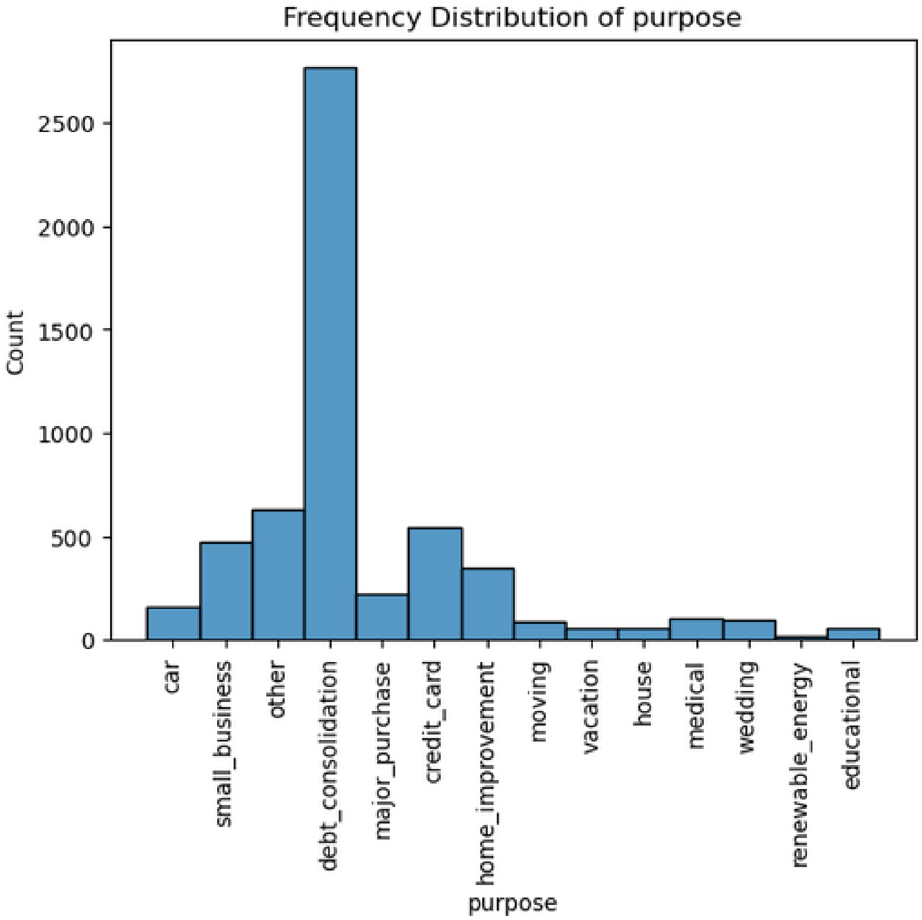
Data Visualisation:



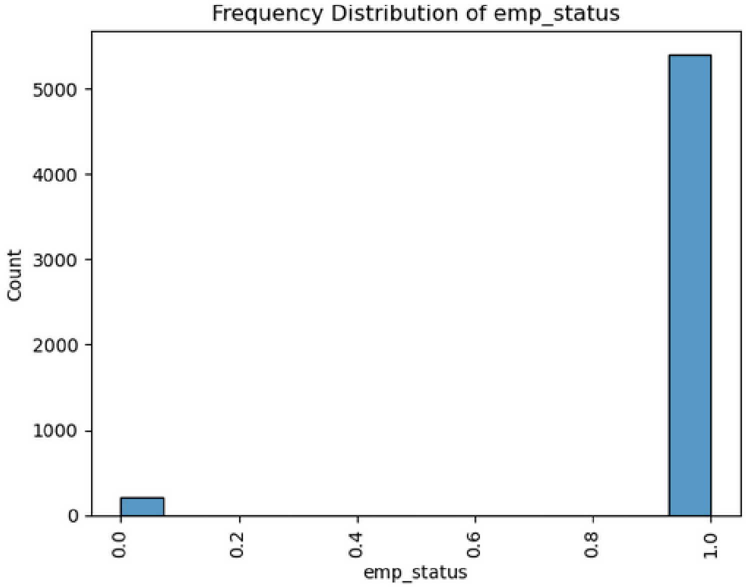
- The default cases for Not verified applicant is maximum.

Data Visualisation:

- The most default cases are of the type debt_consolidation.
- The applicants were already in debt so there is a high probability that he/ she will fail to repay the amount taken to clear the debt.

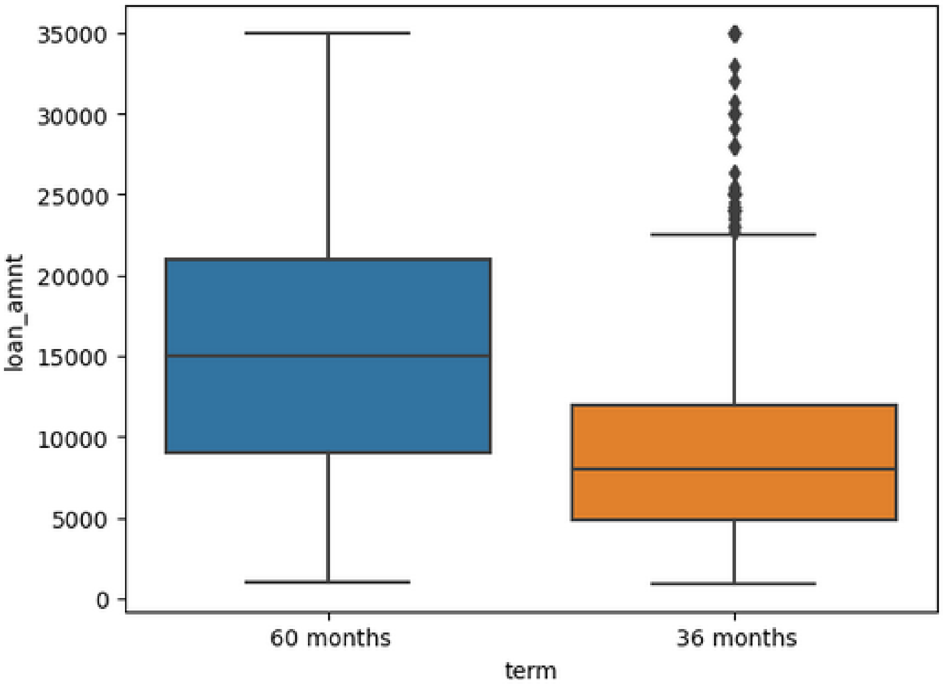


Data Visualisation:



- Most of the defaulters are employed but few defaulters are also there who are unemployed that we have estimated from the emp_status and emp_length data.
- There are 216 such cases where the applicants are unemployed and hence couldn't repay the amount.

Data Visualisation:



- From this data we can see that there are many outlieres in the box plot of loan_amnt vs term.
- From the box plot it can be concluded that the more default cases has been arised when comparatively high loan has been taken for short term.

Conclusion:

The lending club has to consider the following observations while deciding the loan aproval.

- Most default cases are for loan amount less than 25000 and 50 percentile defaulter have taken loan of less than 10000 i.e these applicants are of financially weaker section hence can’t able to repay this amount.
- From the income plot also it is clear that the maximum defaulters having anual income<1 lakh rupee.
- Max default cases has been arised when comparatively high loan has been taken for short term i.e. 36 months
- The most default cases are from group A,B,C,D and E. Among which B3,B4,B5,C1 and C2 have the maximum default.

Conclusion:

- The applicants who are living in a rent house or in mortage are failing to repay the loan.
- Maximum defaulters are generated due to failing in verifcation of applicant data.
- Applicants who are taking loan to repay their debt are mostly failing to repay the loan.
- The interest rate of C,D and E is comparatively higher than that of A,B which may be the reason of unable to repay the loan. F and G has the less default rate but highest interest rate compare to other group.