EDA - Lending Club Case Study

Pradeep Ravi

Business Objective

This 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.

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

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Breakdown of Objective

- Derive the Target variable to Approve or Reject the loan based on the features
- Identify the key features which can influence the decision on approving or rejecting the loan
- Visualize and demonstrate on how features influence the desired output

EDA Approach

- ❏ Data Understanding
- ❏ Dropping features which have more nulls and single static value
- ❏ Dropping Columns based the Domain Knowledge
- ❏ Missing values Treatment
- ❏ Deriving new column
- ❏ Filtering the rows based on Domain and problem statement understanding
- ❏ Outlier Treatment
- ❏ Univariate, Bivariate, multivariate and Derived Metrics Analysis

Dropping Columns and Rows

- Columns are dropped which are meeting the below criteria
  - 70 % column values are null
  - Loan Post-approval columns - The whole idea is to find who is likely to get defaulted, post-approval columns are not in scope
  - Columns which has only single value

Rows which match the loan_status as ‘Fully Paid’ or ‘Charged Off’ are included, the rest of the rows are filtered

Initial Shape - (39717, 111)

Shape after dropping columns/rows - (38577, 23)

Missing Value Treatment

- Missing values of the Categorical column - emp_length are treated with mode value
- Missing values of the pub_rec_bankruptcies column - emp_length are treated with median value

Data Standardisation and Deriving new metric column

- int_rate - Type is object though the value is numerical, removing % from the data and casting to float will standardise this column
- emp_length - We can convert this as a numerical column by assigning <1 year category to '0' and for the rest of values we can remove the years and + to standardise as int
- issue_d - Derive new columns like year and month for time-series charts/analysis
- revol_util - Type is object though the value is numerical, removing % from the data and casting to float will standardise this column

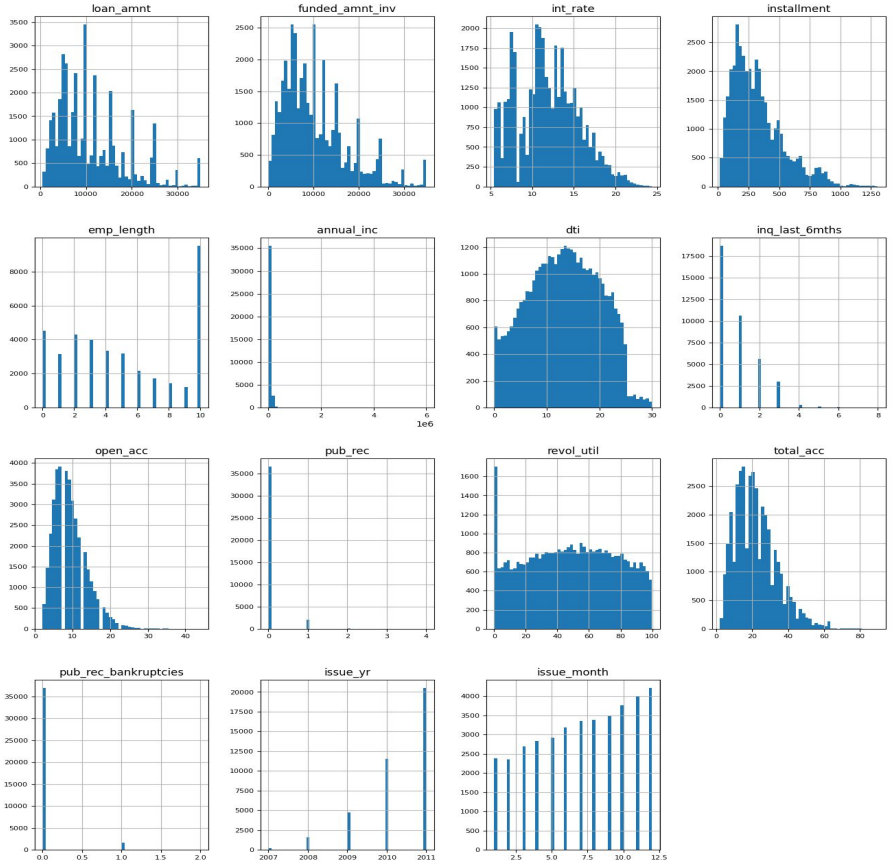
Numerical Distribution

The histogram of the numerical distribution helps us to understand the below inferences,

Columns which are heavily skewed

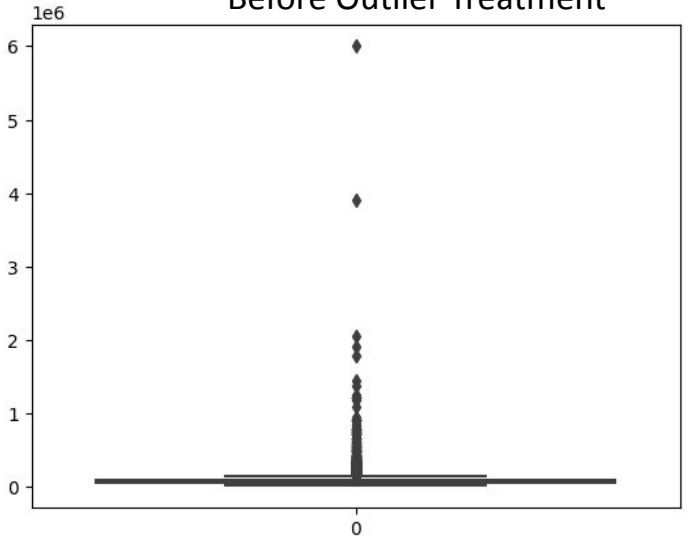
are annual_inc,pub_rec,pub_rec_bankruptcies,

this would need further outlier treatment

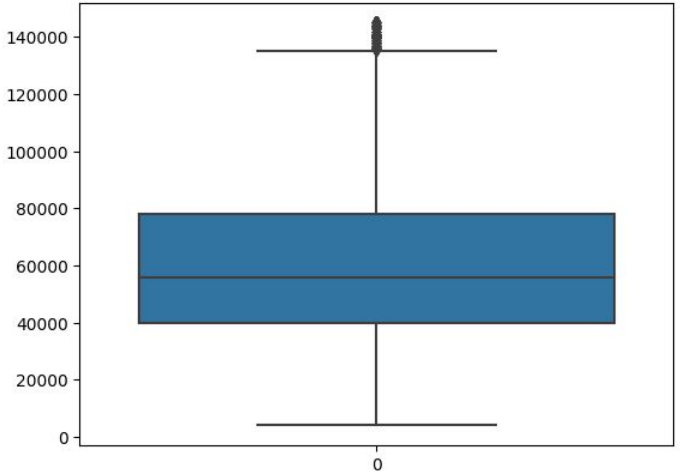


Outlier Treatment - annual_inc

Before Outlier Treatment



After Outlier Treatment

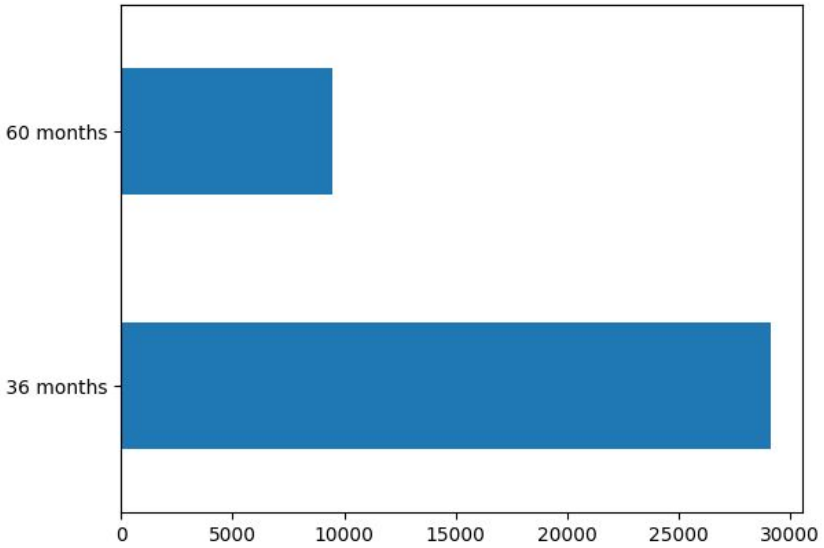


The below steps are followed to remove the outliers

1. Calculated the IQR
2. Used the condition to filter out the values beyond the IQR range
3. Outlier dataframe is created and the same is used to drop the outlier values from the orginal dataframe using drop

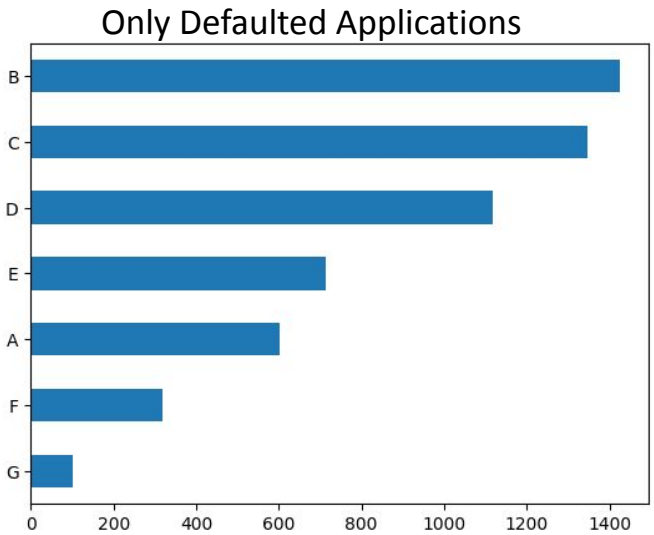
Univariate analysis using Categorical variable

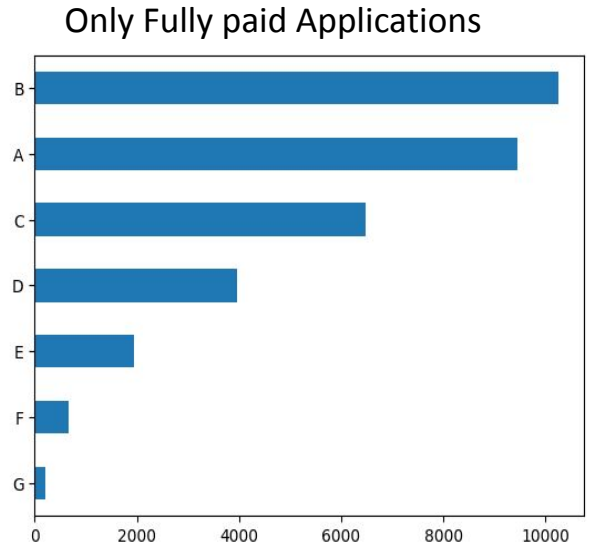
Univariate analysis can help to understand the distribution but it may not help to find the causation in all scenarios - variable used is term



Inference - 36 months more preferred compared to 60 months

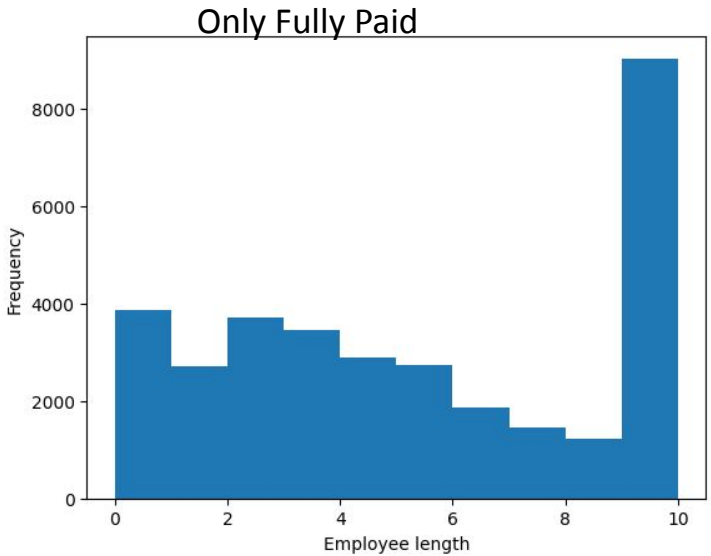
Univariate analysis using Categorical variable(Segmented Analysis - for the column grade)

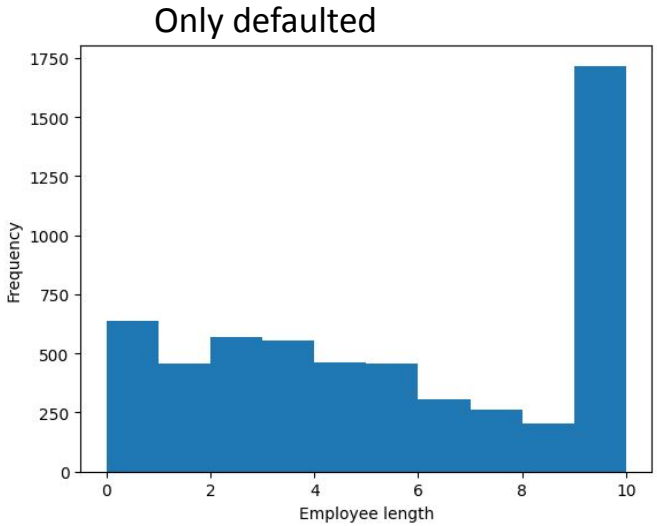




Inference - By comparing the above two charts , its clearly evident , the loan grade which we considered less risky are the one who paid most of the loans and the very same category are the ones who defaulted the loans. So the grade assignment to qualify the application is less riskier didn't helped or not accurate (This is an example of Univariate segmented analysis - Comparing one group with another, in our case defaulted vs fully paid)

Univariate analysis using Categorical variable- Segmented Analysis

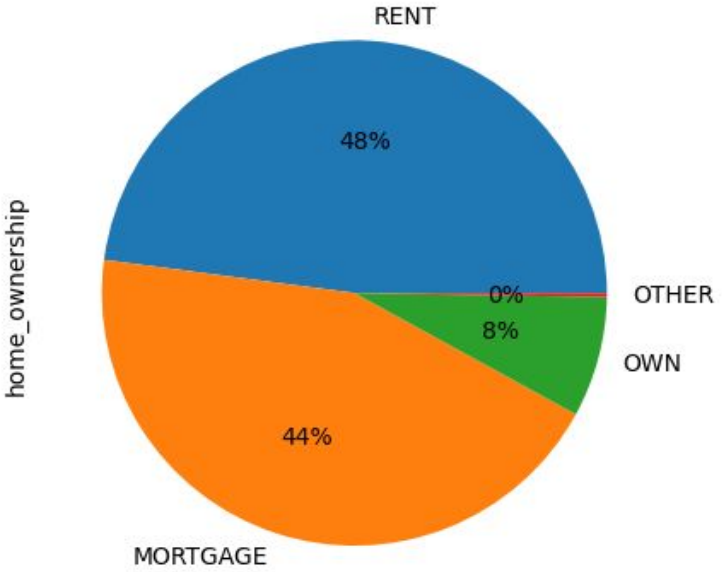




Inference - The experience category of 9-10 are the ones who defaulted most of the loans.By comparing the defaulted vs paid group, in both the cases 9-10 experience category are the ones who paid and who defaulted most . So extra risk mitigation has to be taken by the bank to ensure the loan is approved for this group

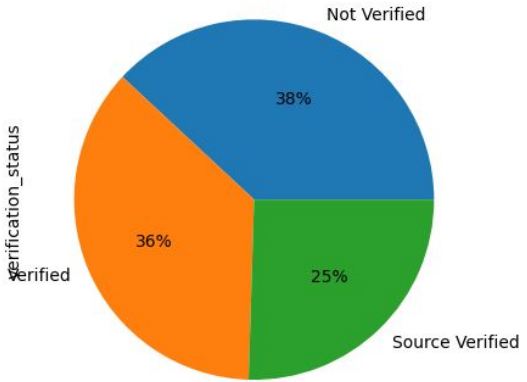
Univariate analysis using Categorical variable

Observation from the pie chart is , the most amount of the loan applications are from the home_membership category of Rent and Mortage

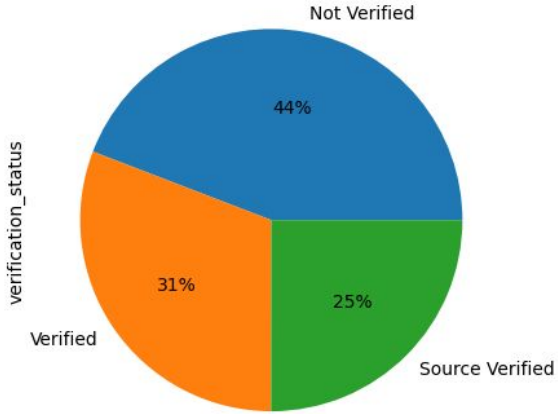


Univariate analysis using Categorical variable

Only defaulted



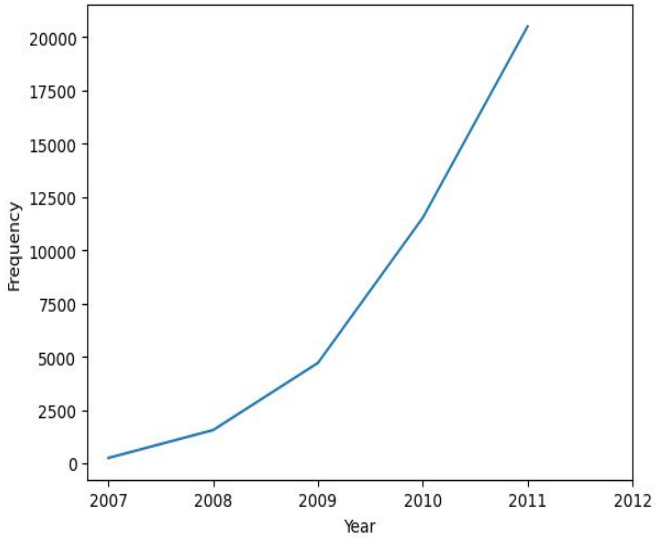
Only fully paid



The inference is regardless of income is verified or not , it did not have any inference over loan getting defaulted or fully paid . Because , in both the case we could see Not-verified is having close to equal share

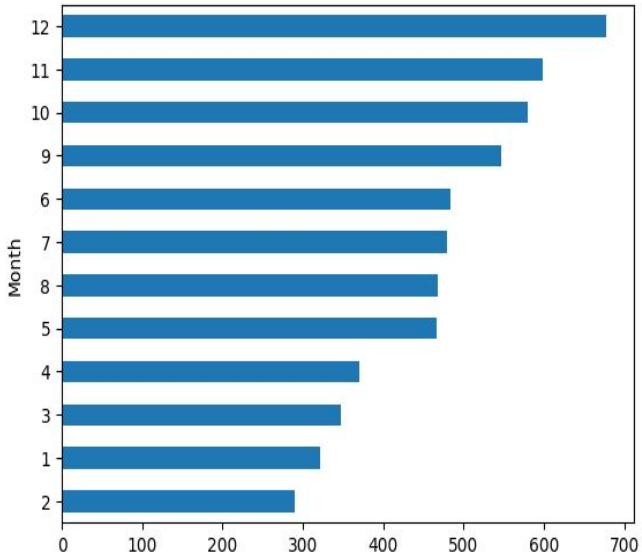
Univariate analysis using derived column

Observation from the line chart is, the number of loans issued increases as the year increases



Univariate analysis using derived column

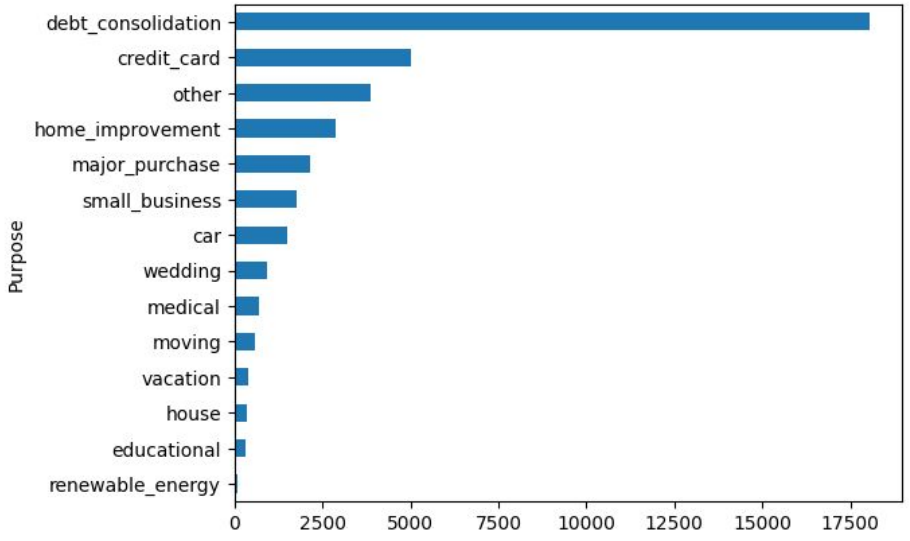
Its interesting to observe the december month has more defaults compared to other months across all years



Univariate analysis - Categorical Column

The top most purpose for the loan application is Debt Consolidation

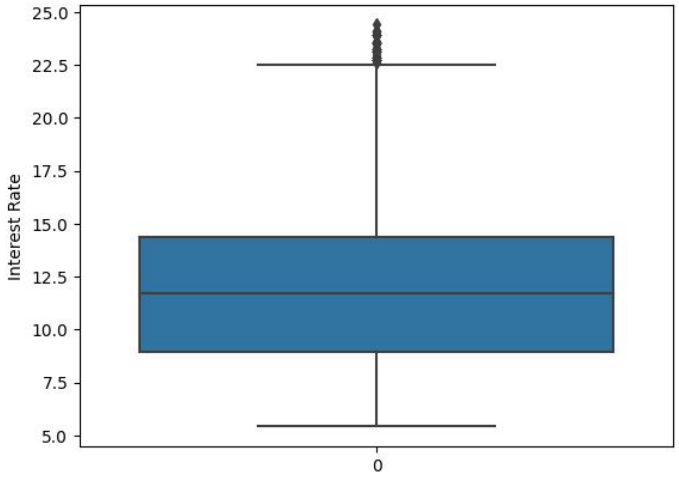
Debt consolidation refers to taking out a new loan or credit card to pay off other existing loans or credit cards. By combining multiple debts into a single, larger loan,



Univariate analysis for Numerical column

The median value of the int_rate is 12.5%

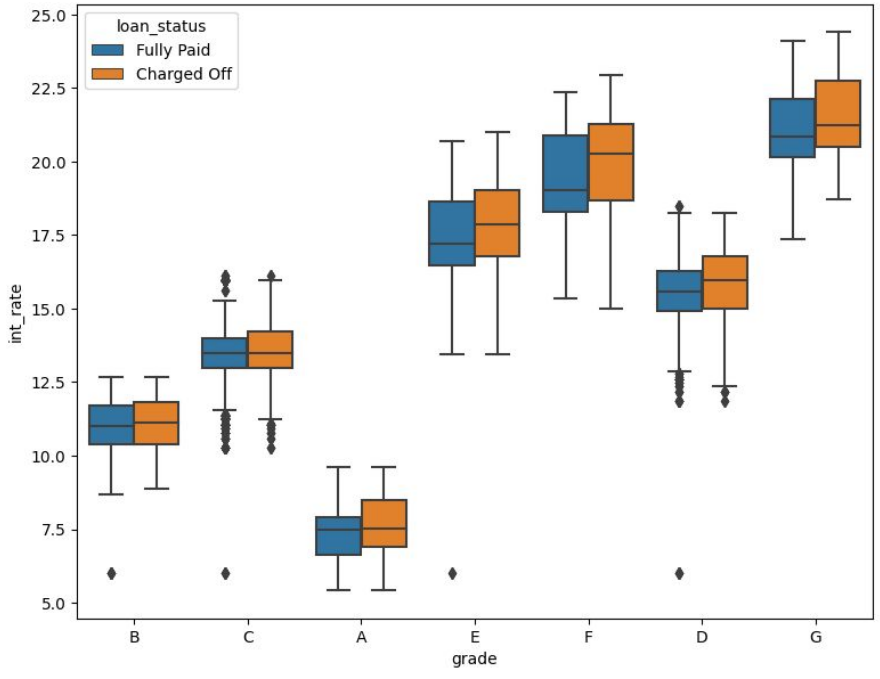
25th and 75th percentile is 8.9 and 14.4 respectively
The box plot helps us to bucket the ranges of int_rate



Bi-Variate Analysis

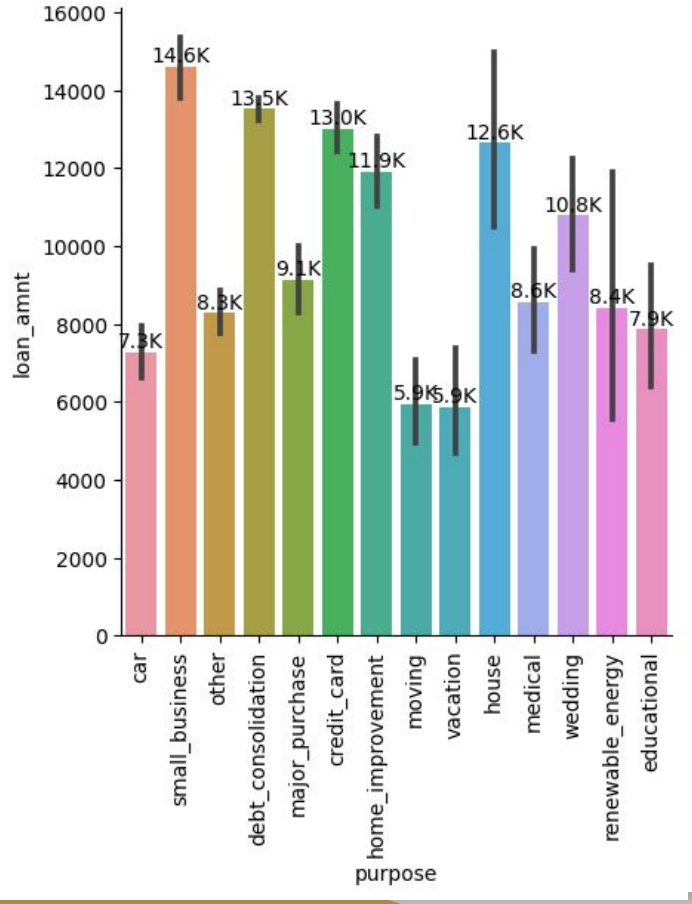
Lets analyse the relationship between grade vs int_rate - Categorical vs Numerical

The Box plot clearly demonstrates , the int_rate is more when the grade is more riskier such as Grade-G and int_rate is less for the safer loan application such as Grade-A . We can also notice, the quantiles (25th,50th,75th) values of charged-off is slightly higher than fully paid for every grade



Bi-Variate Analysis

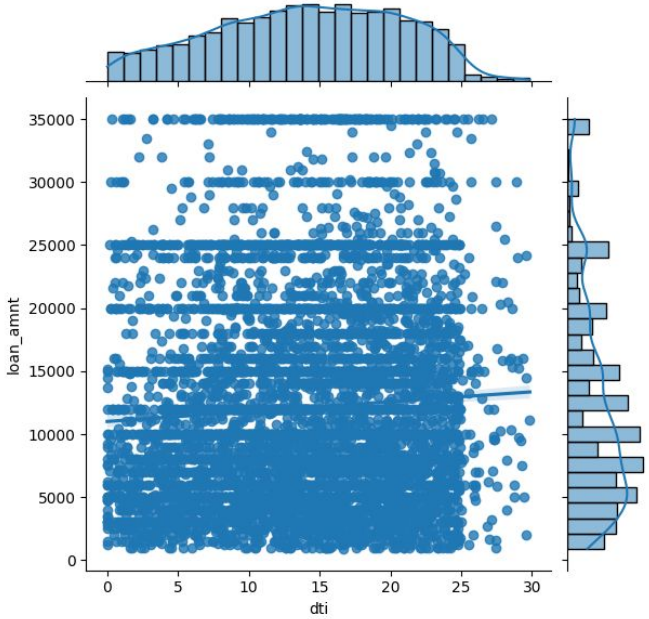
Here the dataset is filtered only with defaulted loans and we could observe the small_business category take high loan and being the top most category of getting defaulted and the second top most is debt_consolidation



Bi-Variate Analysis

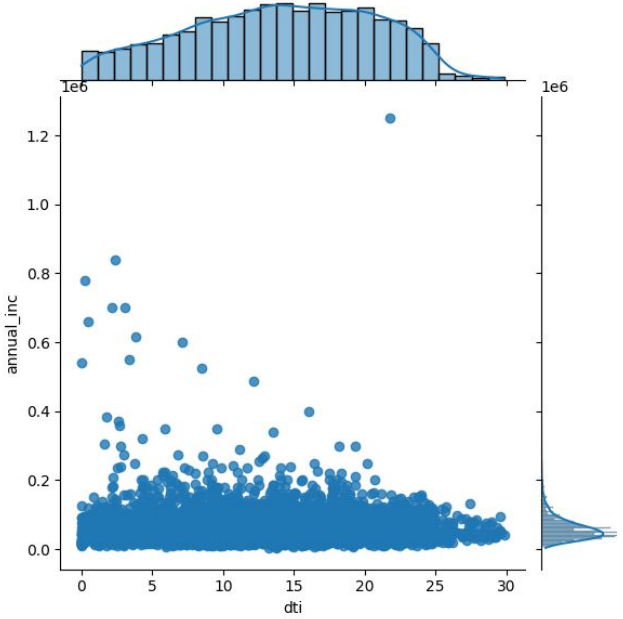
The dti is The debt-to-income (DTI) ratio is the percentage of your gross monthly income that goes to paying your monthly debt payments and is used by lenders to determine your borrowing risk.

The chart represents , DTI is weakly correlated with loan_amnt for the defaulted applications



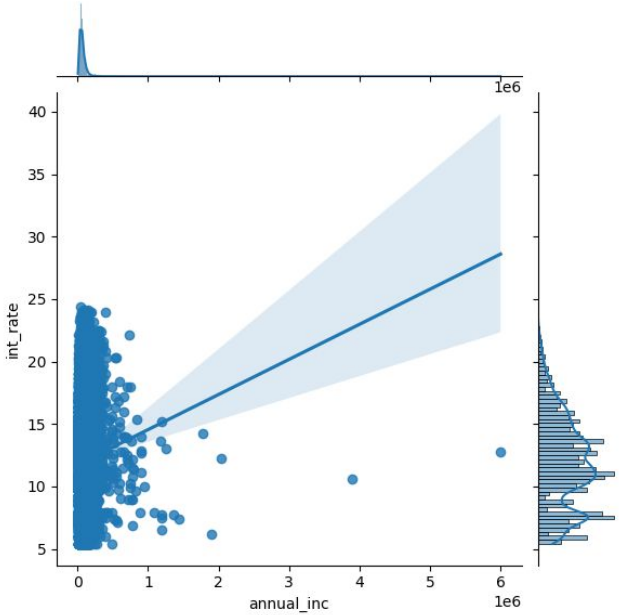
Bi-Variate Analysis

The chart express that the dti is negatively correlated with annual income

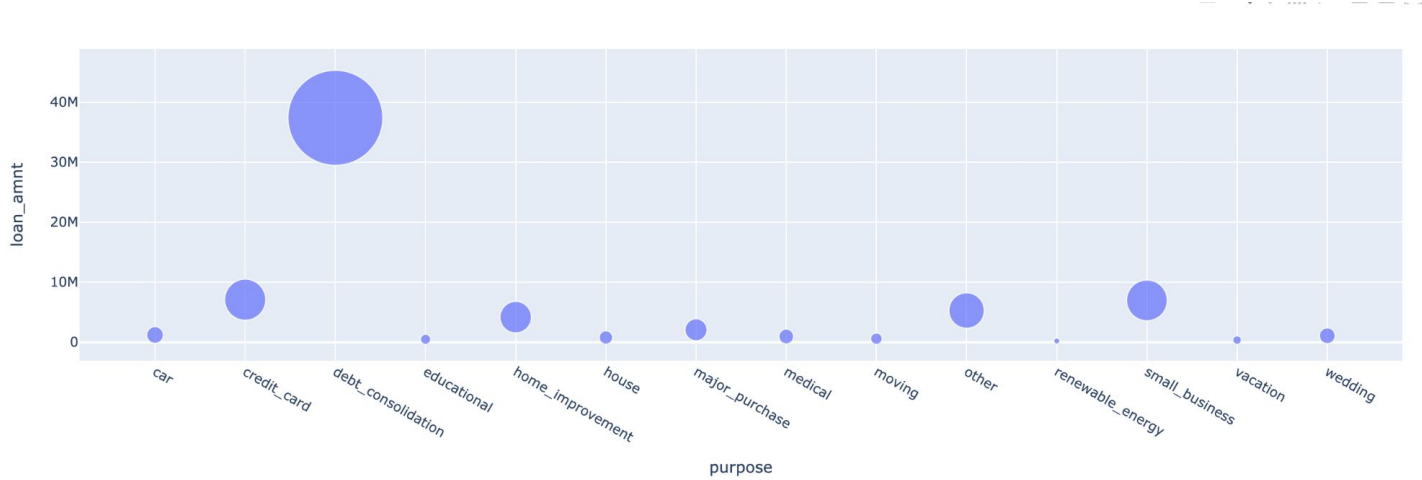


Bi-Variate Analysis

Lets analyse int_rate and annual_inc have any relationship, one of the way to identify is to use jointplot and draw the regression line to check whether the line meets the points , the chart shows there is a weaker correlation between int_rate and annual_inc



Bi-Variate Analysis



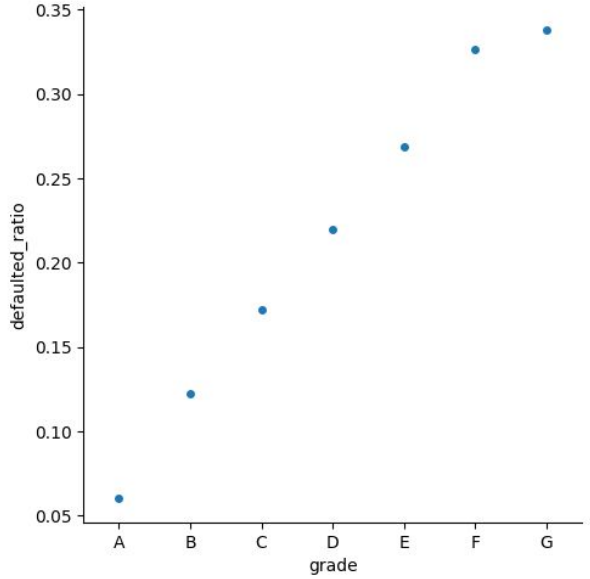
The above bubble chart clearly explains the sum of the loan amount is higher for dept_consolidation and we have only defaulted data points in the dataset

Multivariate Analysis - Derived column

Defaulted_ratio is derived per Grade using the formula - No of Charged Off/Total loan applicants

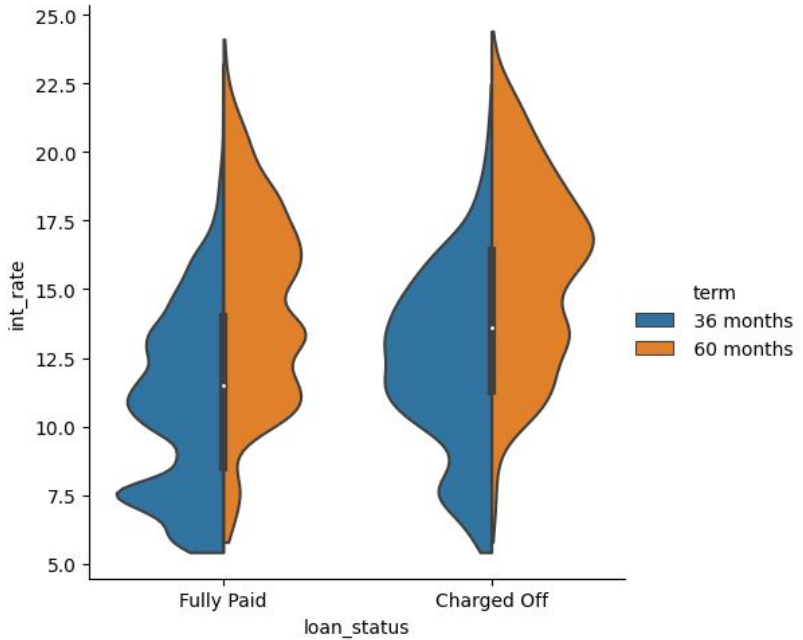
Three columns are used - grade(x-axis),int_rate(as values), loan_status

The ratio analysis for the possibility of getting defaulted based on int_rate and grade concludes, the Grade G, F and E are the top most categories who has higher possibility of getting defaulted

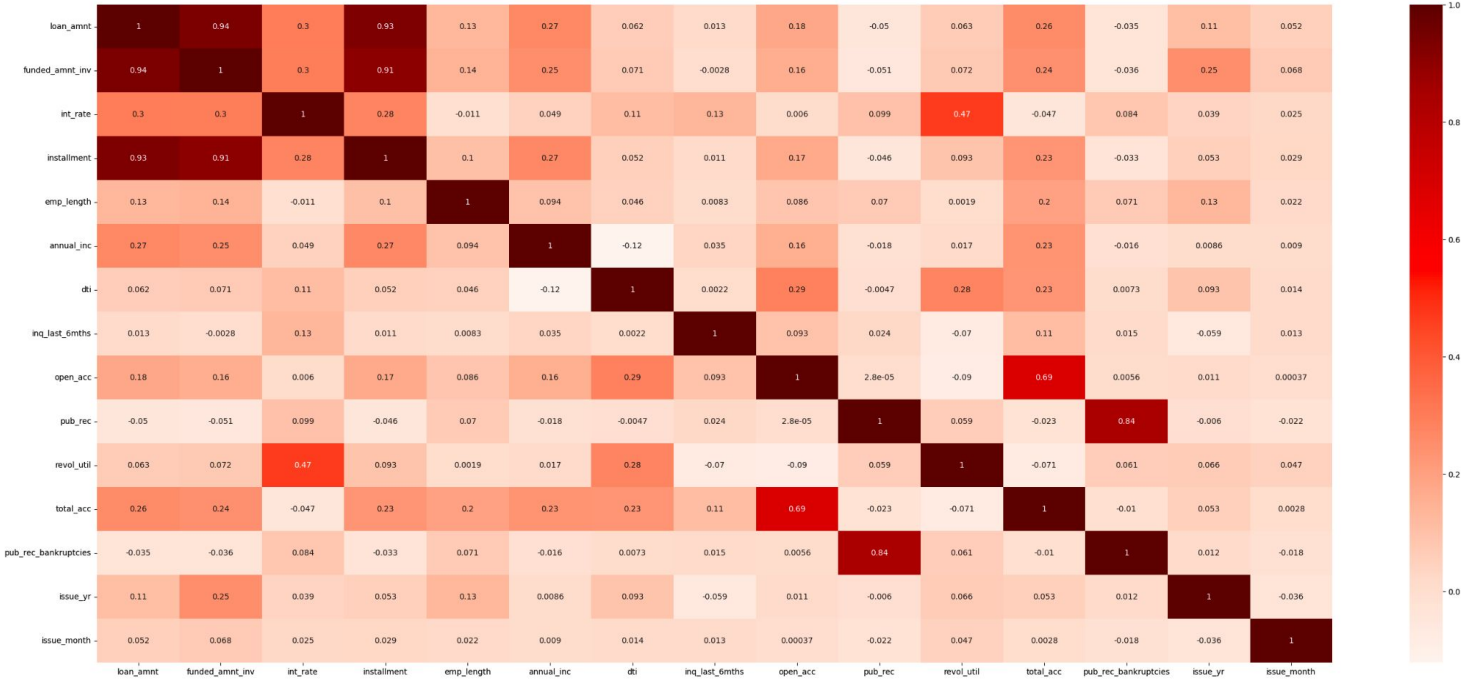


Multivariate Analysis - Segmented Analysis

The violin chart clearly indicates the 60 months term has more possibilities of getting defaulted compared to the 36 months term

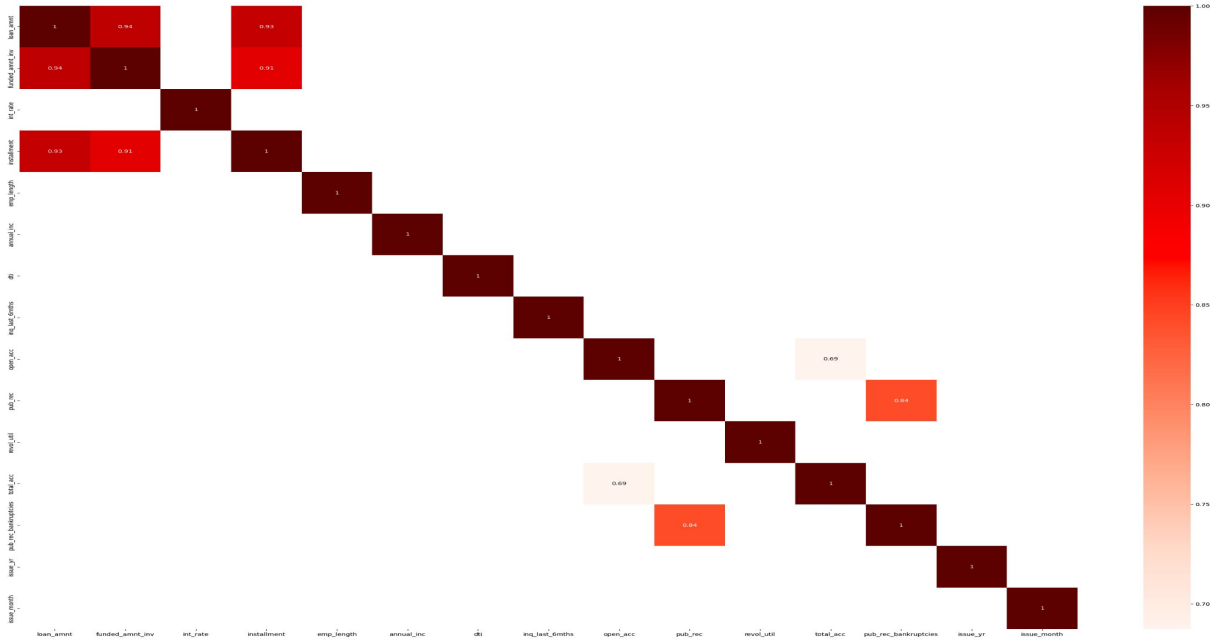


Multivariate Analysis - Correlation Analysis



The heatmap chart shows good number of features with correlation

Multivariate Analysis - Correlation Analysis



Though the heatmap chart shows good number of features with correlation , it looks exhaustive and we dont need features which are less correlated. Lets filter down by selecting the features which are correlated > 50%

conclusion

- The correlation doesn't prove causation (Understanding the domain and logically reasoning would provide more insights)
- The final Golden features in the numerical types are : 1- loan_amnt, 2- funded_amnt_inv, 3- installment, 4- open_acc, 5- pub_rec, 6- total_acc, 7- pub_rec_bankruptcies
- The final Golden features in the Category types are : 1- term 2- grade 3- sub_grade 4- home_ownership 5- verification_status 6- loan_status 7- purpose 8- addr_state 9- earliest_cr_line The other column which has influence over the loan status is emp_length

EDA may not need to be single time activity , the process can be iterative. When we discover more amount of columns are removed and more missing values . The recommended approach is to find a way to bring the data, though its costly , the true insights can be derived only when we have more data with valid values