Lending Club Case Study

Abhimanu Pandey

Shen Shaji

Index

- Problem statement
- Analysis approach
- Result of Univariate analysis
- Result of Bivariate analysis
- Visualizations
- Recommendations
- Conclusion

Problem statement

In this case study we are provided with the data of customers who has taken loan from the firm. A lot of features regarding the customer and other aspects are provided as a dataset. The dataset is about the past customers, so it included the information whether the customer has completed his payment, or still paying, or defaulted in the loan payment. Our objective is to find what all features/variables are relevant in deciding whether a future applicant is likely to default, so that we can give the information about a future applicant to the investors.

Analysis approach

- Data understanding
  - Analyzing the variables and understanding their meaning
- Data cleaning
  - Treating missing values
  - Treating outliers
  - Typecasting
  - String to numeric conversion and striping characters
- Data Analysis
  - Univariate analysis
  - Segmented univariate analysis
  - Bivariate analysis
- Recommendations
- Conclusion
- Future scope

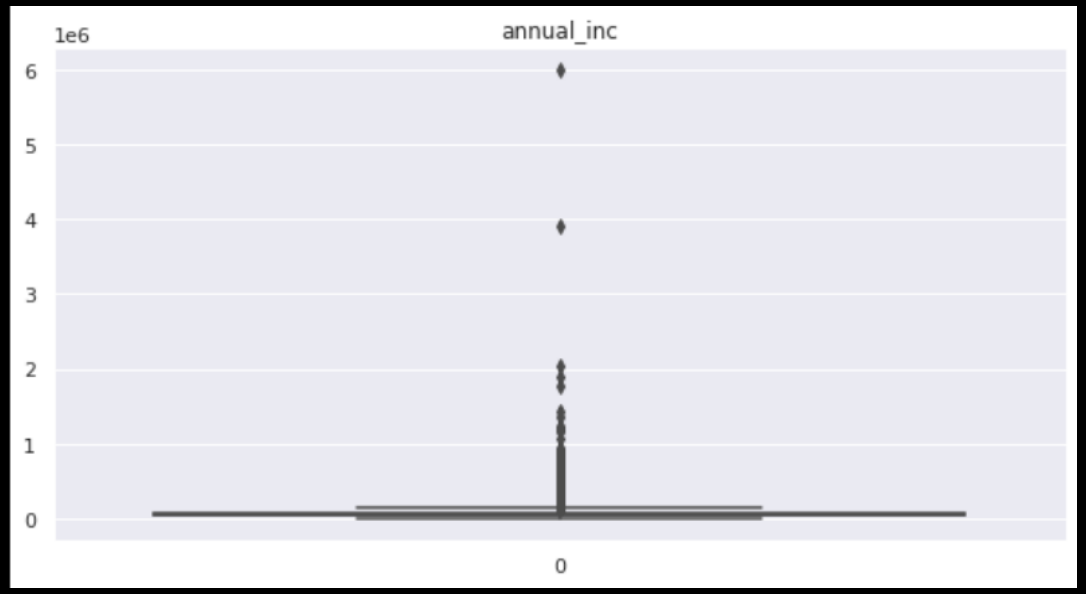
Assumptions

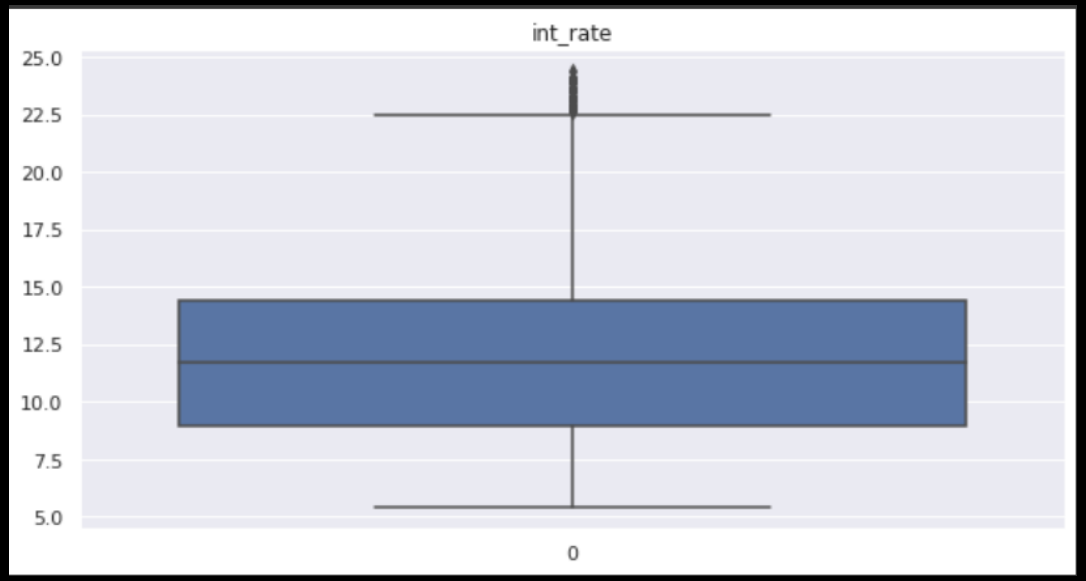
- As we do not know whether the users with 'Current' loan status are going to be defaulted or not, we exclude those datapoints from the dataset.
- We treat those who has less than 1 year of work experience as 1 year experienced and those with more than 10 years of experience as 10 years experienced.
- Filling the missing values for ‘public record bankruptcies' as '0' assuming that the missing value users do not have any bankruptcies.
- Filling the missing values of work experience' with its average of all the values. Since there are no outliers in the work experience feature, we are using average instead of the 50^{th} percentile value.
- Since 'loan amount', 'funded amount' and 'funded amount by the investors' are highly correlated we are considering 'funded amount'. In the ' 'funded amount by the investors’ feature there are zero values for which the ‘loan status’ feature shows either fully paid or charged off. So we are not taking that column among the three variables.
- Since the variable 'annual income' has a lot of outliers we are considering the values below the 95^{th} percentile.

Result - Univariate analysis

- The data distribution structure of loans among each variable is analyzed
- Annual income and interest rate variables had a lot of outliers, we eliminated the outliers as they may result in a biased analysis.
- The customers from each category are analyzed and their distribution among each values is identified. For example, we understood that the customers are majorly from the state 'CA'.

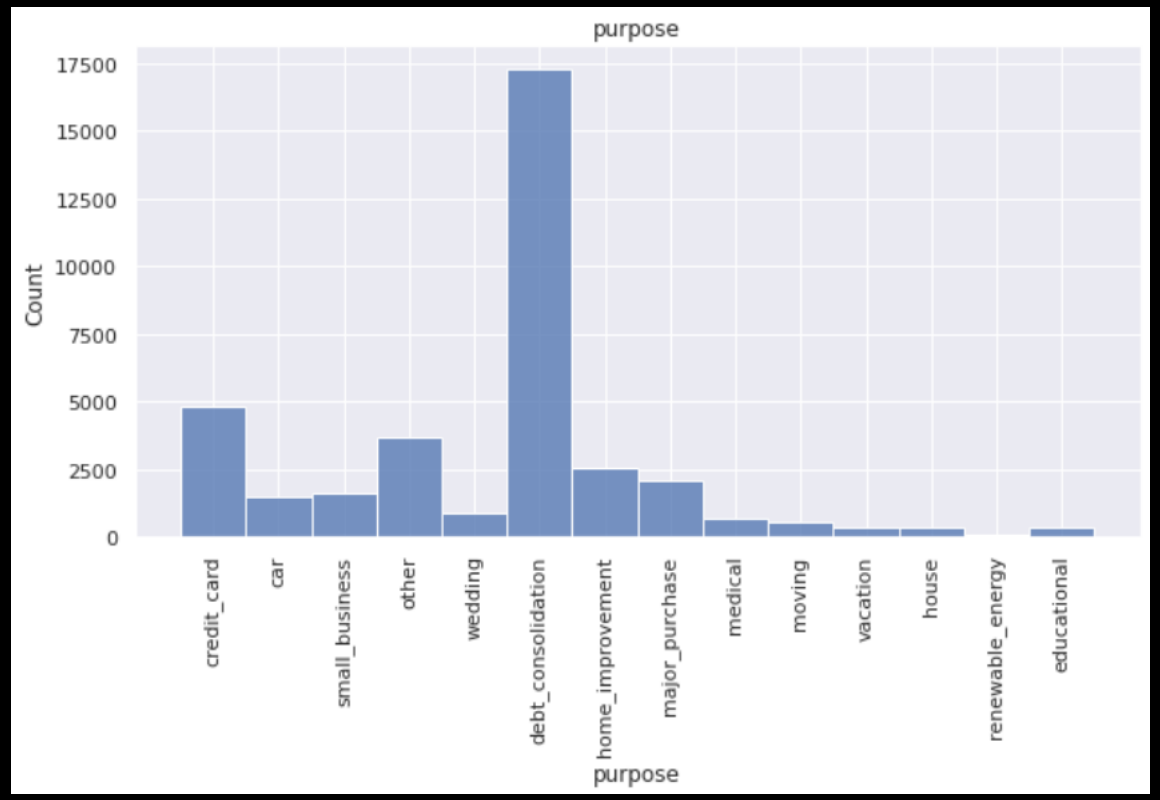
Visualizations - univariate analysis





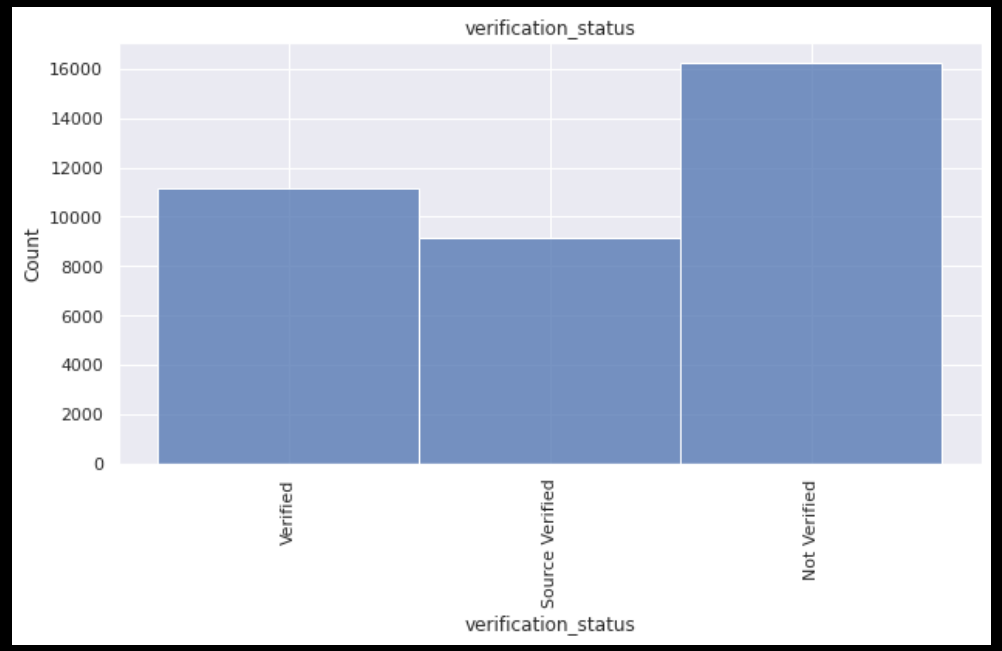
Annual income and interest rate variables have significant amount of outliers

Visualizations - univariate analysis



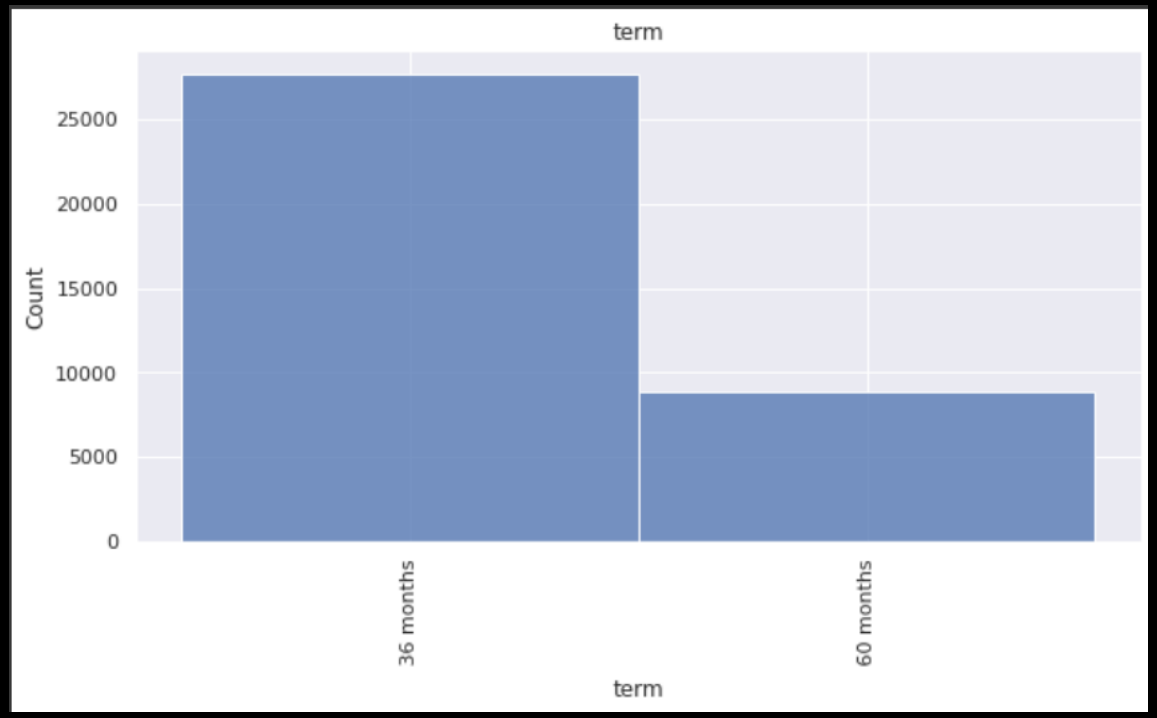
Majority of the borrowers have the purpose of debt consolidation

Visualizations - univariate analysis



The combined verified customers are more than those who are not verified

Visualizations - univariate analysis

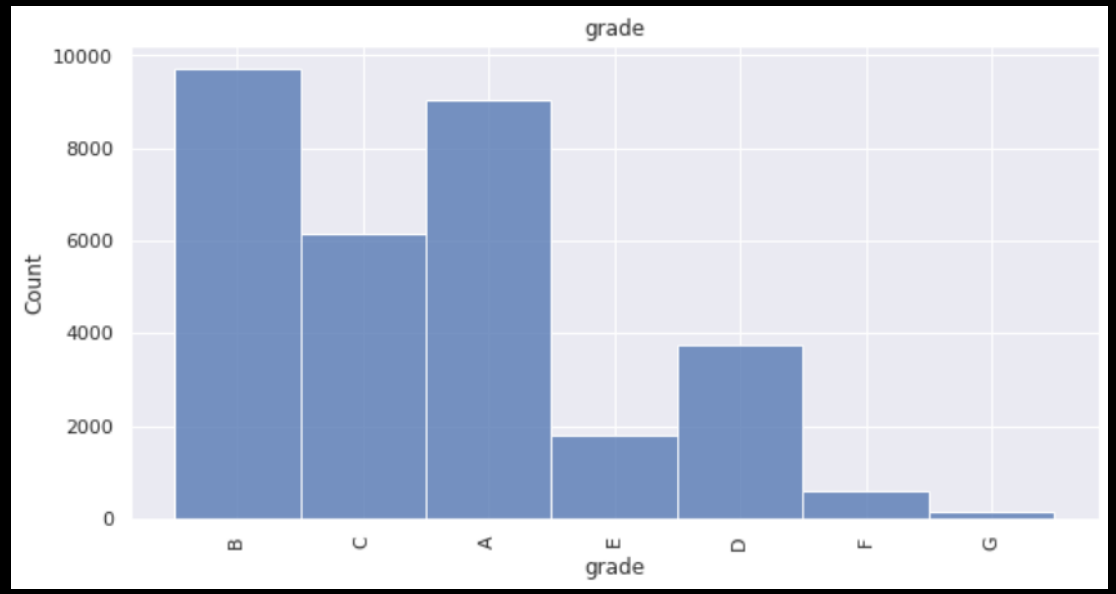


Major share of customers take 36 month term loan

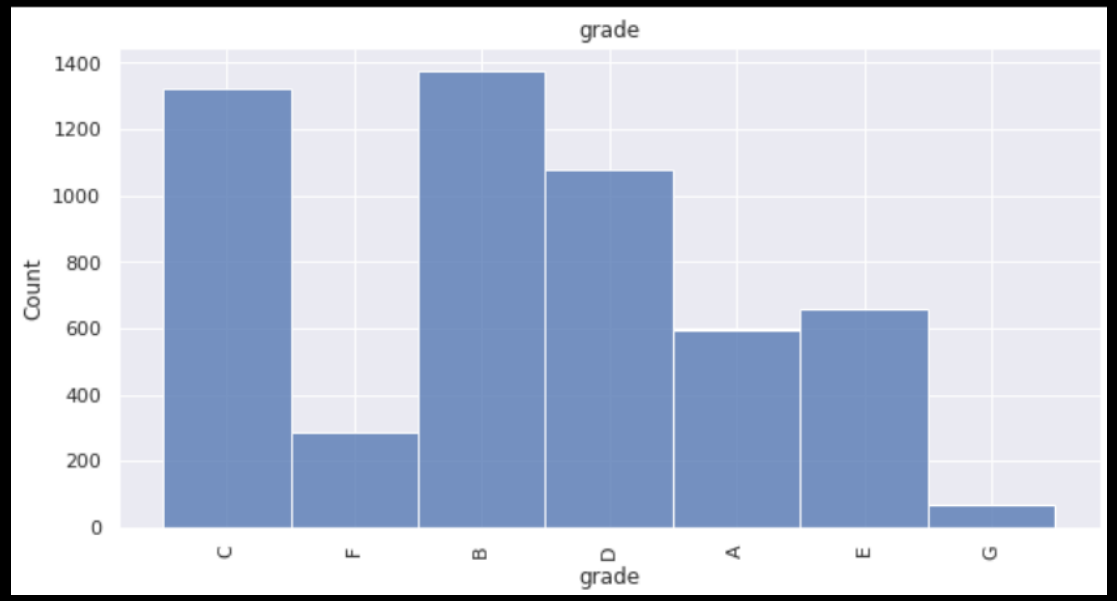
Result - Segmented univariate analysis

- A and B grade customers fall in fully paid loan status more than the others but B and C grade customers fall in charged off loan status more than the others
- Debt consolidation is the major purpose among borrowers irrespective of their loan status
- People from the state 'CA' have more borrowers irrespective of their loan status
- 36 month term has more customers in charged off status but not that significant change compared to the fully paid customers

Visualizations – segmented univariate analysis

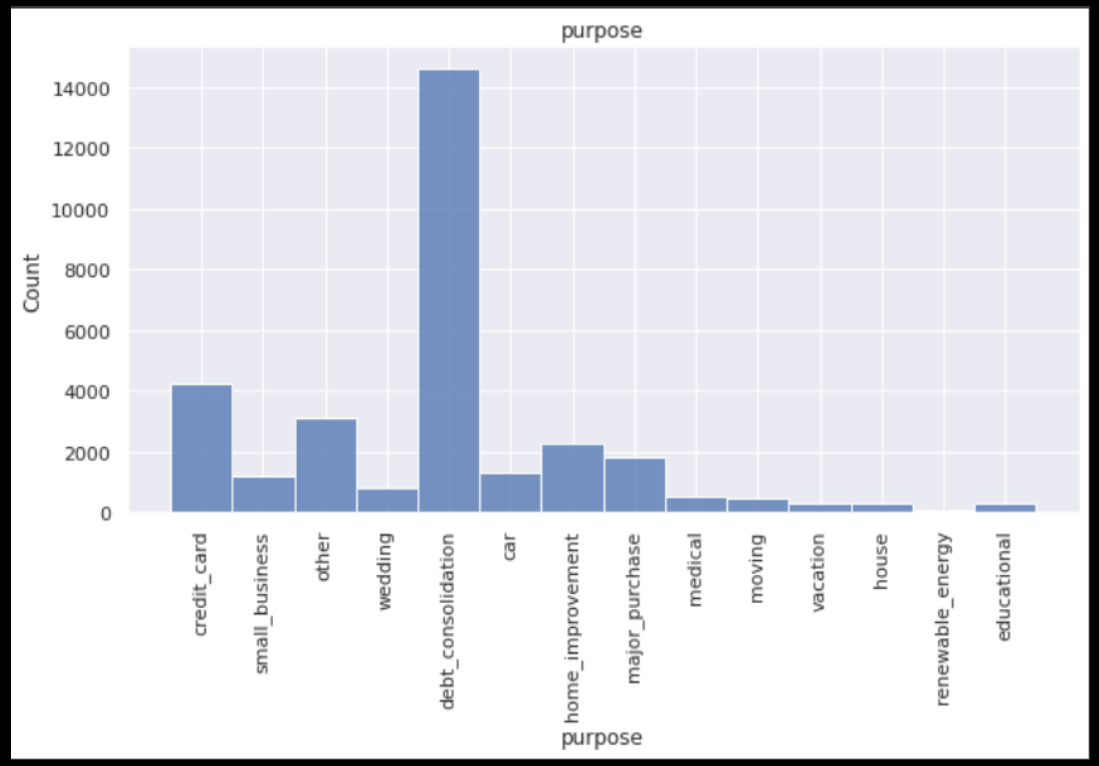


Customers who are fully paid

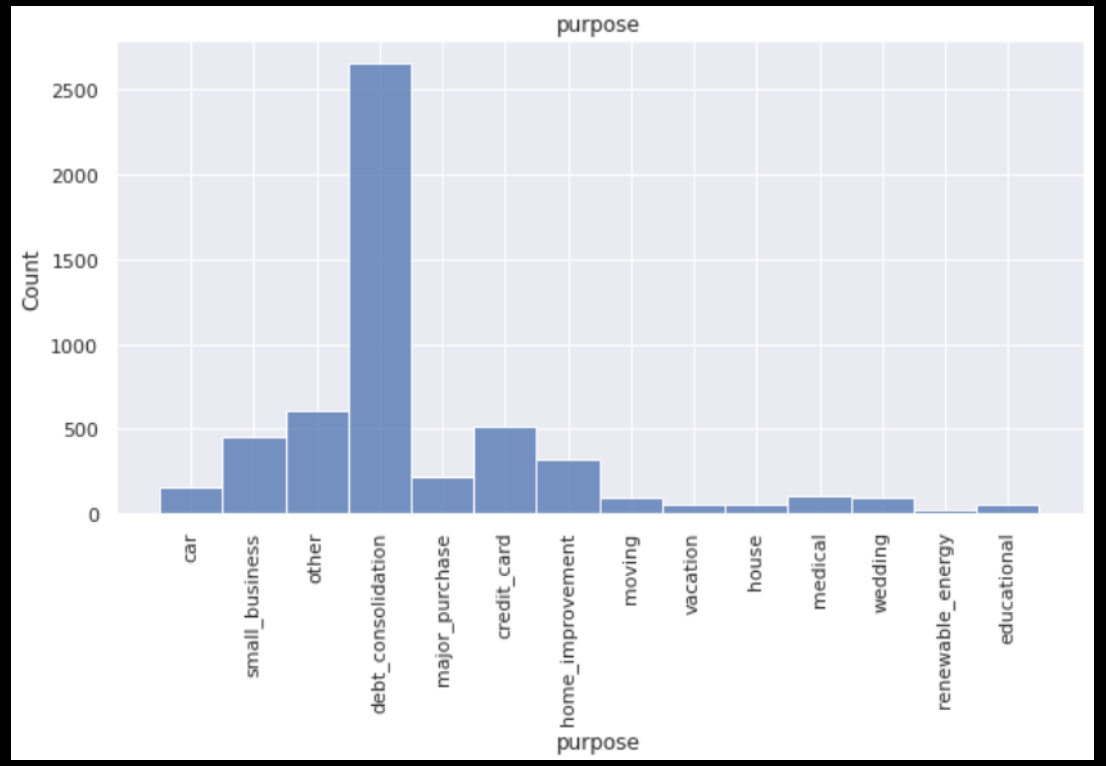


Customers who are defaulted

Visualizations – segmented univariate analysis

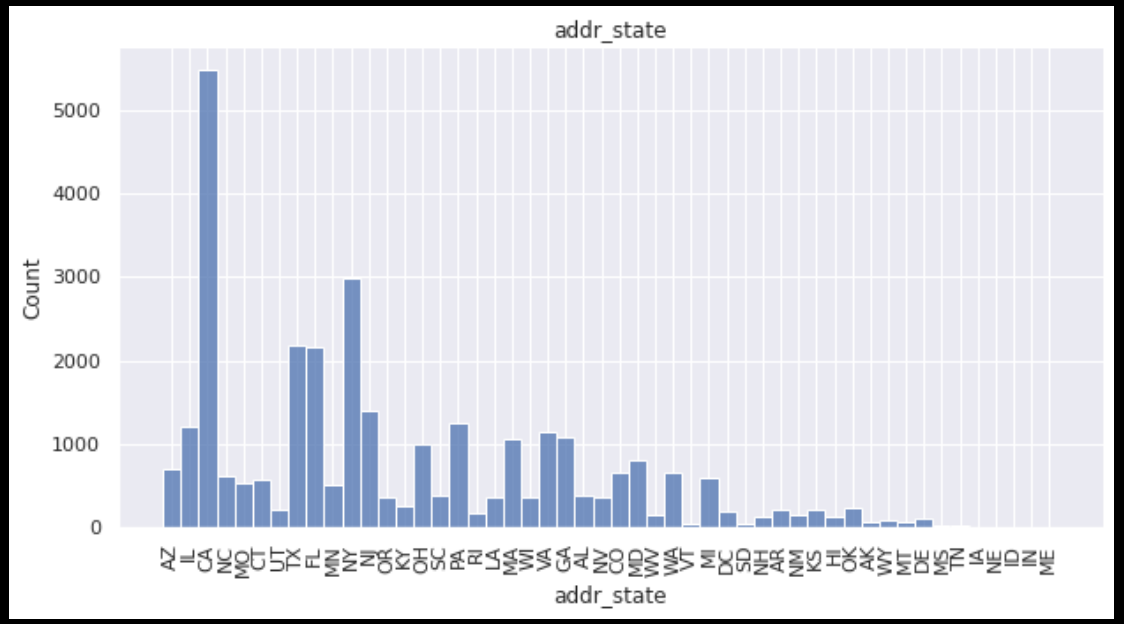


Customers who are fully paid

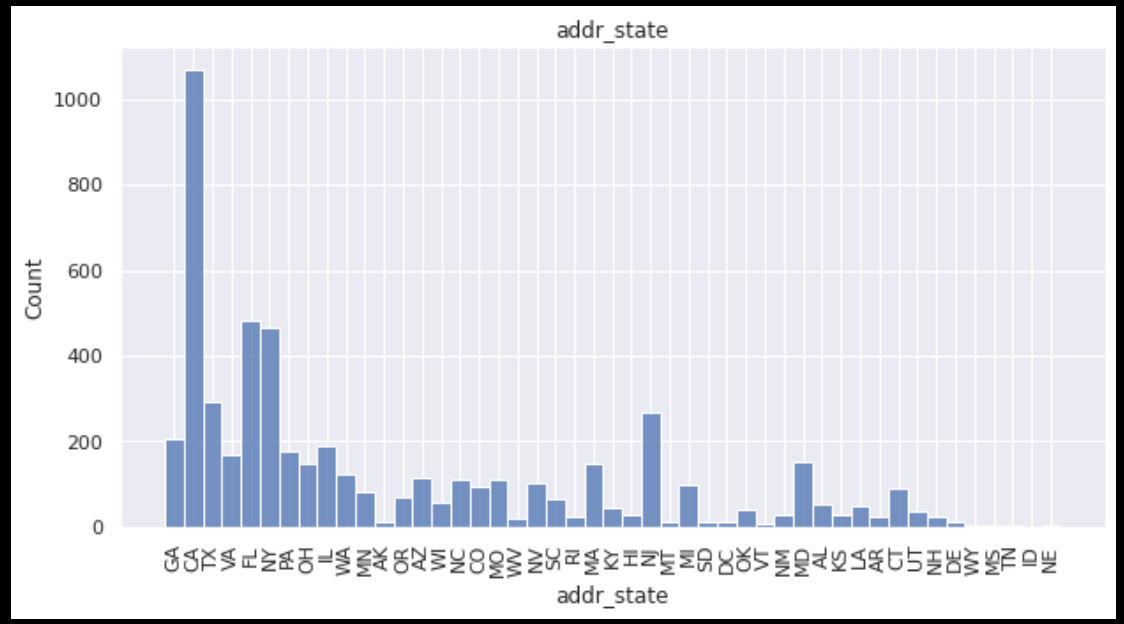


Customers who are defaulted

Visualizations – segmented univariate analysis

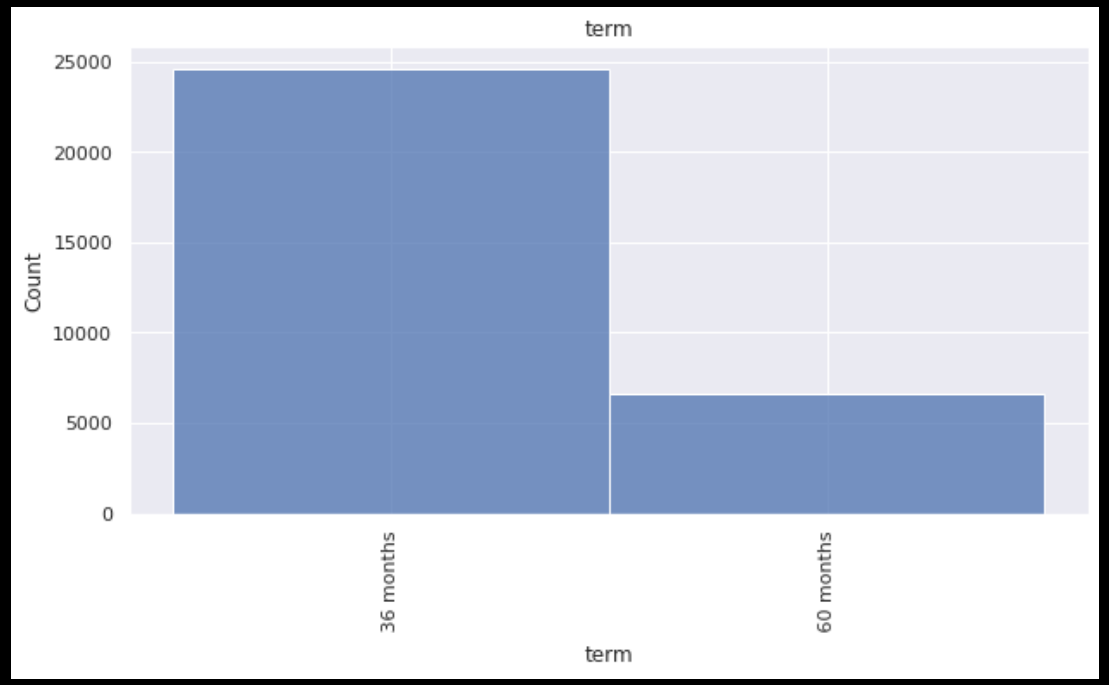


Customers who are fully paid

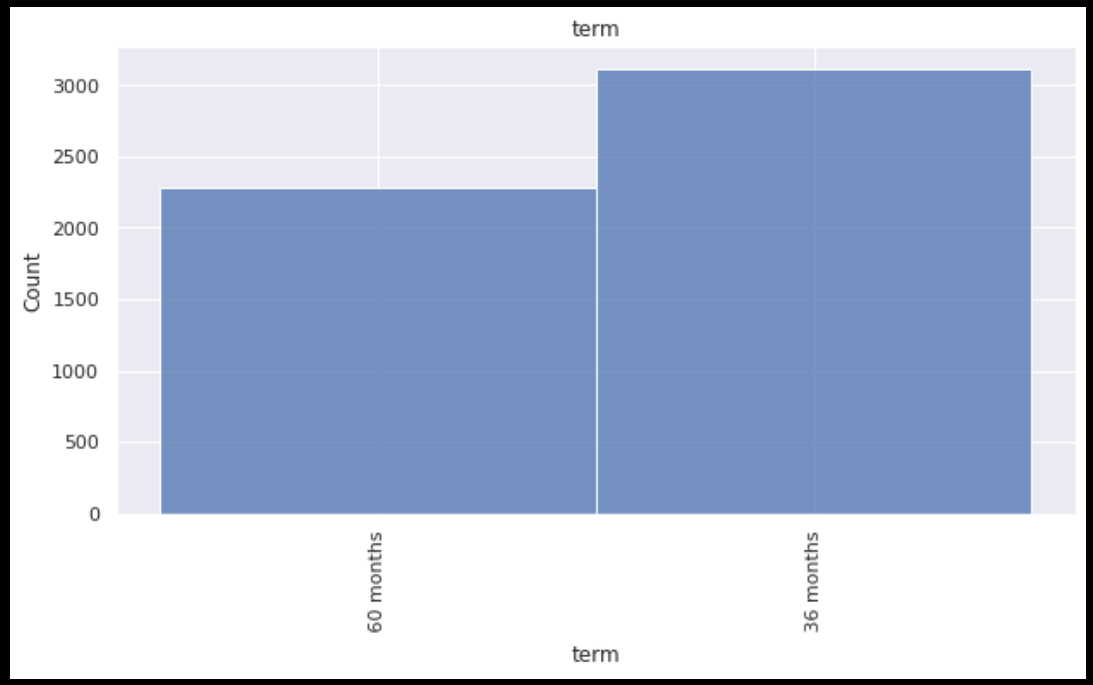


Customers who are defaulted

Visualizations – segmented univariate analysis



Customers who are fully paid

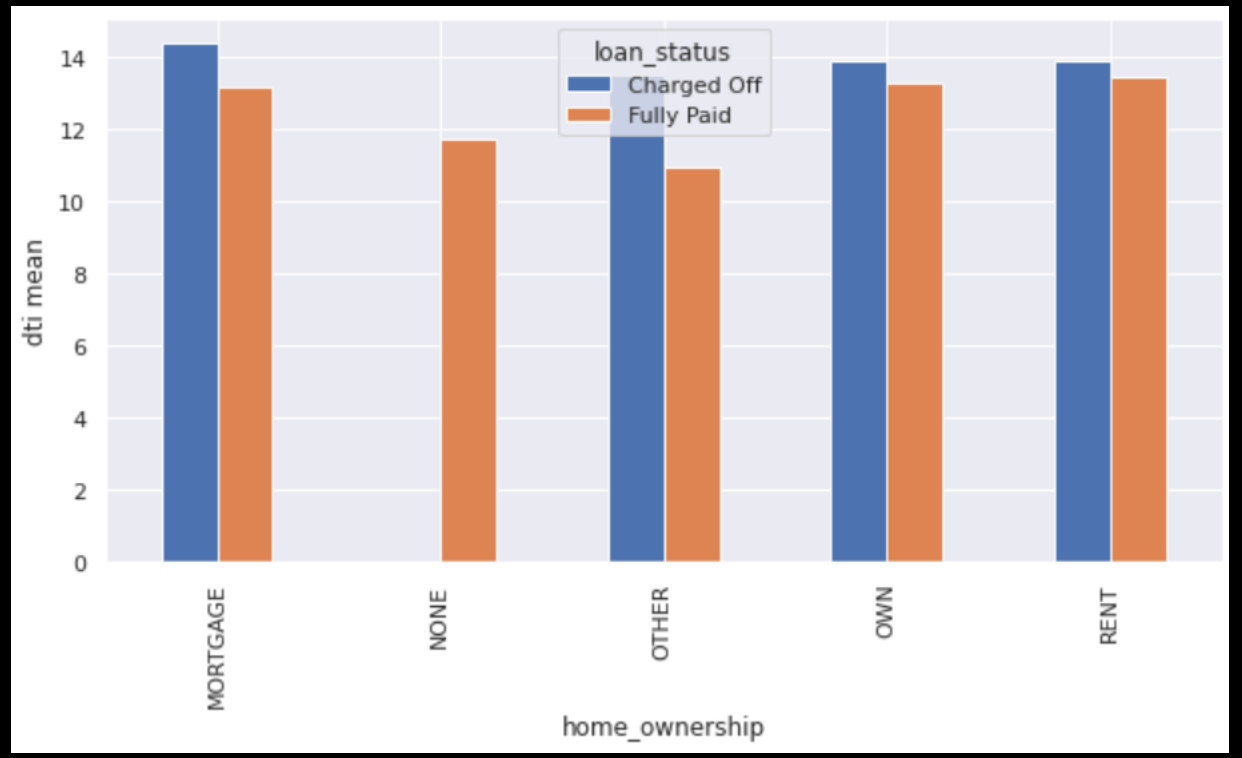


Customers who are defaulted

Result - Bivariate analysis

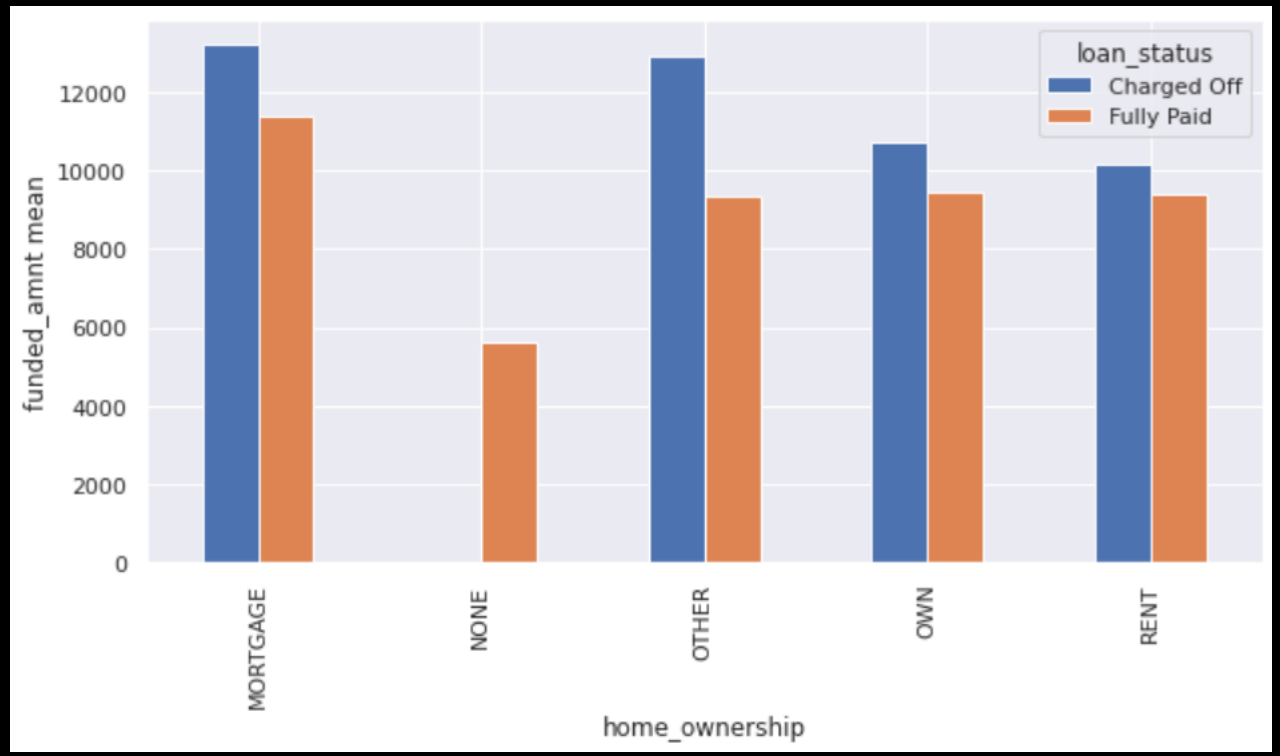
From the visualizations, we see that the mean value of funded_amnt and dti for users with 'Charged Off' loan_status is significantly higher than Fully paid customers for home_ownership type as 'mortgage'. So we can suggest to reduce the loan amount of customers with home_ownership type as 'mortgage'.

Visualizations - bivariate analysis



This bar chart represents the average value of ‘debt to income ratio’ for each ‘home ownership’ type aggregated based on the ‘loan status’

Visualizations - bivariate analysis



This bar chart represents the average value of ‘funded amount’ for each ‘home ownership’ type aggregated based on the ‘loan status’

Recommendations

- Reduce the funded amount for users with home ownership as ‘mortgage’.
- We recommend to deny loan if debt to income ratio is very high for the customers having home ownership as ‘mortgage’.

Conclusion

- From the above plots we see that the mean value of funded amount and ‘debt to income ratio’ for users with 'Charged Off' loan status is significantly higher than ‘Fully paid’ customers for home ownership type as 'mortgage'. So we suggest to reduce the loan amount of customers with home ownership type as 'mortgage’.
- ***Funded amount, Annual income, Debt to income ratio, Home ownership, Purpose of loan, Grade***, are the important features those helps in identifying a potential loss of business or financial loss for the lending club firm from a future borrower.

Thank you