Lending Club Case Study

DRIVING FACTORS BEHIND LOAN DEFAULT

Ajay Lamkhade
Nachiket Sali

Objectives

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit 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. In other words, borrowers who default cause the largest amount of loss to the lenders.

Objective is to identify the risky loan applicants at the time of loan application so that 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, to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment. And thus minimize the risk of losing money while lending to customers.

Solution (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations

Problem solving methodology

Data Analysis

Data Analysis

Data Analysis

Data Understanding

Working with the Data Dictionary and getting knowledge of all the columns and their domain specific uses

Data Cleaning

Removing the null valued columns, unnecessary variables and checking the null value percentage and removing the respective rows.

Univariate Analysis

Analysing each column, plotting the distributions of each column.

Segmented Univariate Analysis

Analysing the continuous data columns with respect to the categorical column

Bivariate Analysis

Analysing the two variable behaviour like term and loan status with respect to loan amount.

Recommendations

Analysing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

Data Understanding

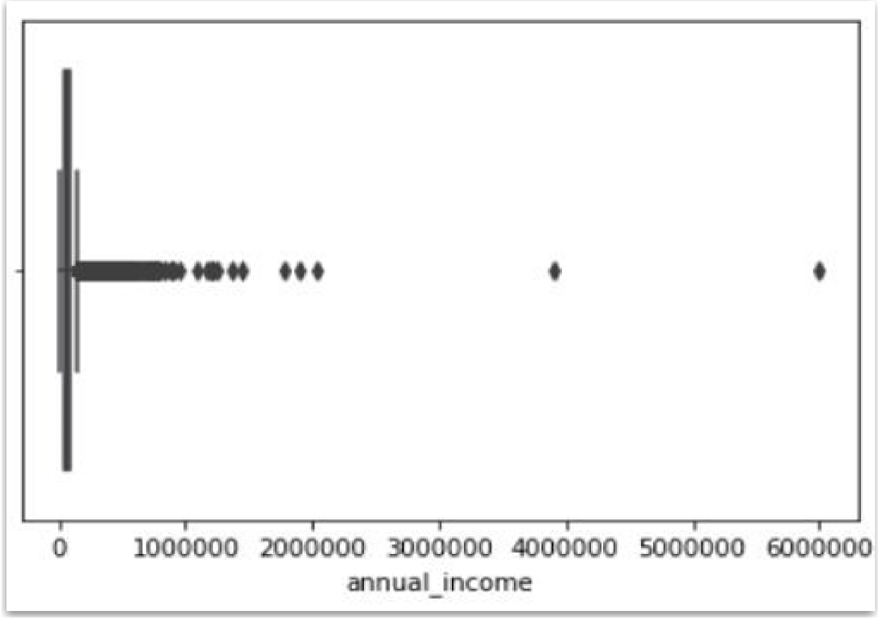
- Total records - 39717
- Attributes - 111
- Data types – Float64,object type, int64

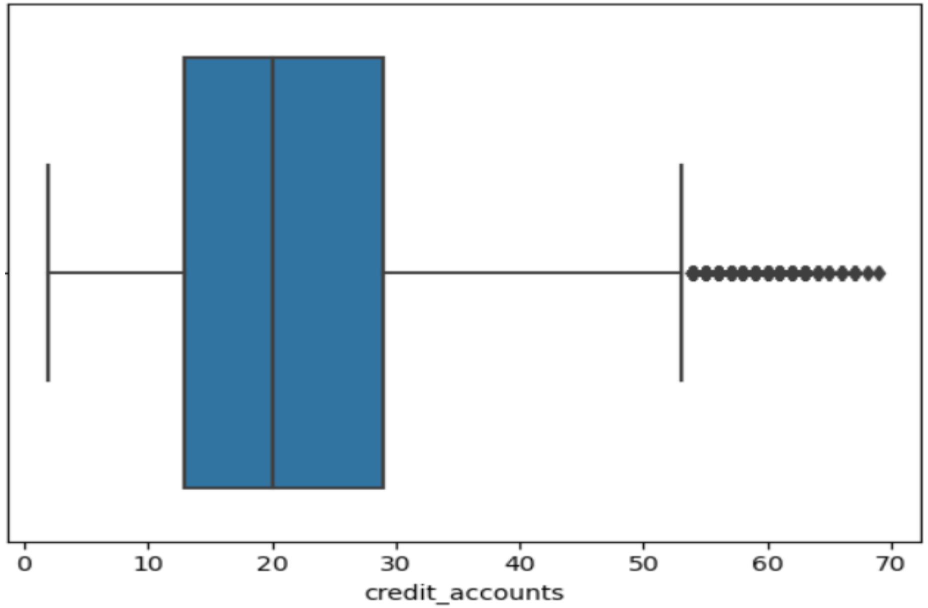
Data Cleaning

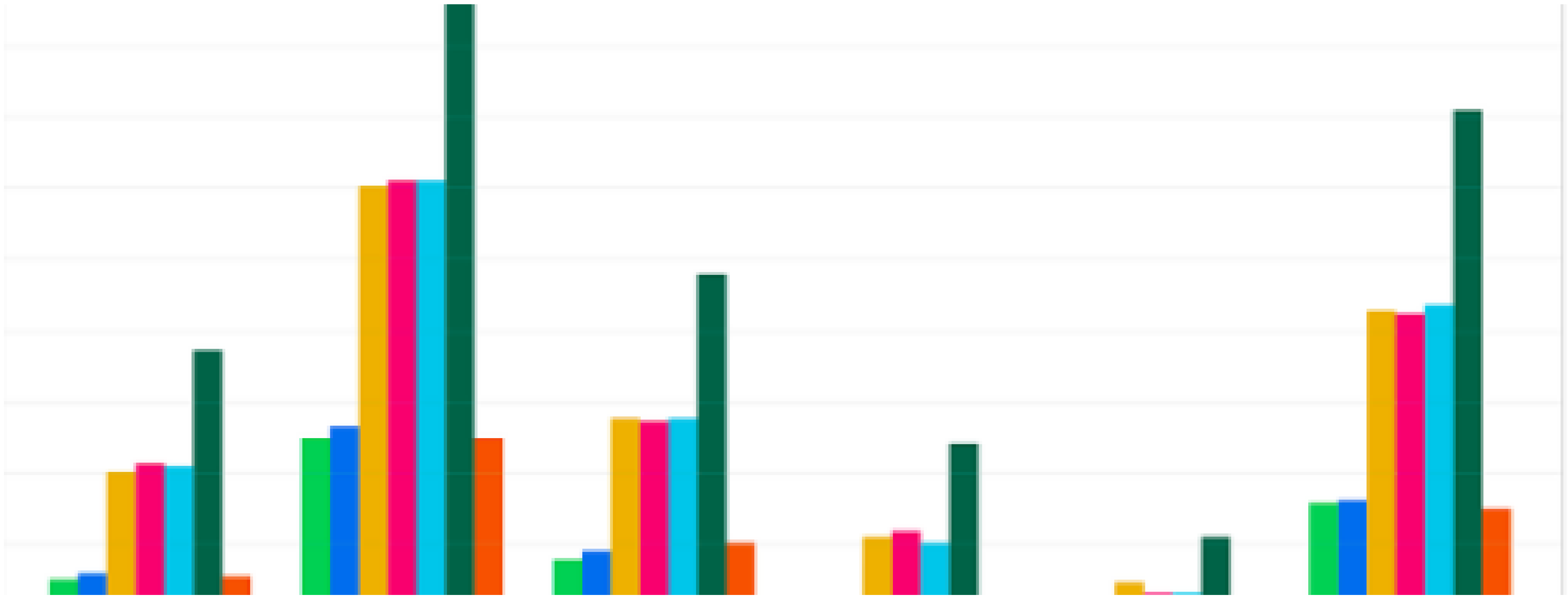
- 54 Columns with All values as null are dropped
- 4 Columns with 25% values as null are dropped
- 9 Columns having same value for all records are dropped
- 16 Irrelevant columns based on observation of rows are dropped
- Renamed few columns for better understanding.
- NAN and Null values are replaced by appropriate value.
- Dropping rows where loan status is current

Fix Invalid Values

Fixed outliers using box plot for *annual_income* and *credit_accounts*





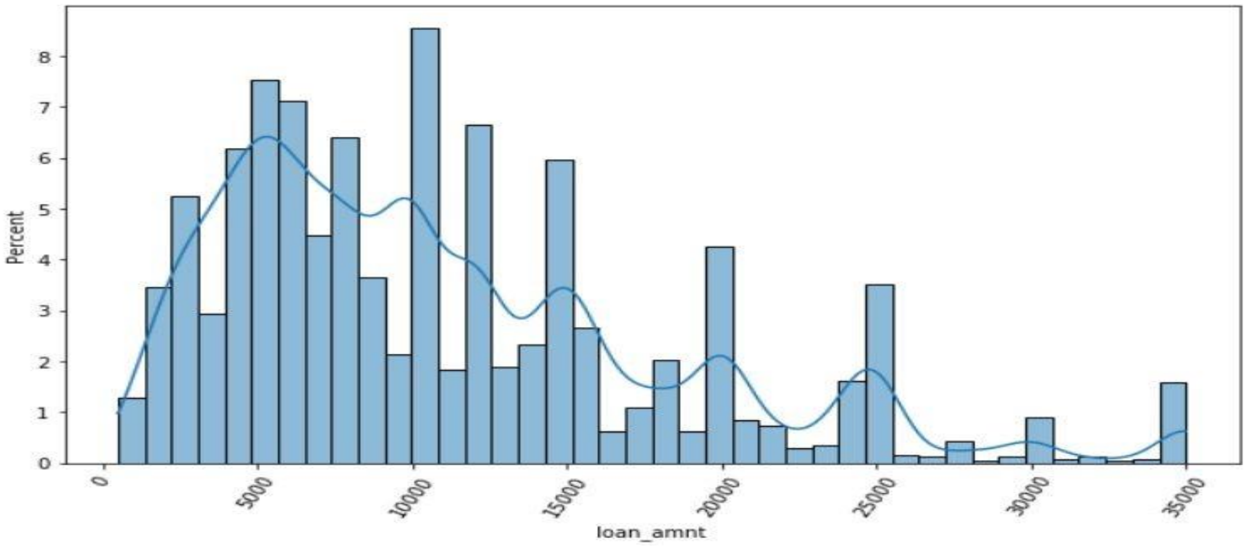


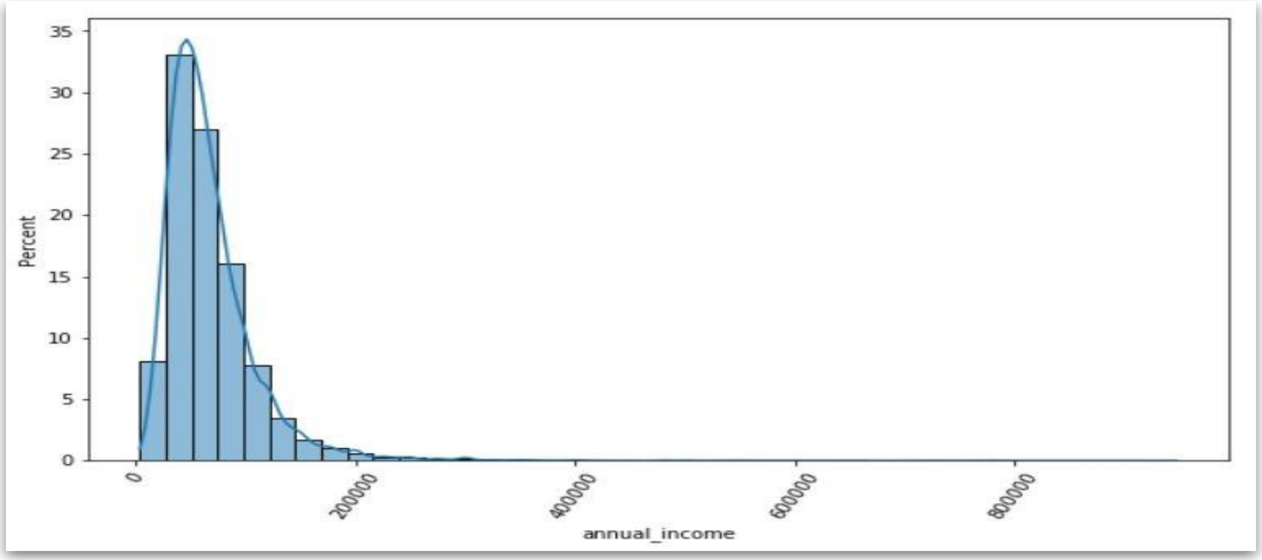
Univariate Analysis

Numerical Columns

- loan_amnt : Loans are taken usually in multiple of 5k and maximum request are of loan upto 15k which is 75 Percentile
- funded_amnt : Loan amount and funded amount looks identical on both box plot and histogram implaying mostly full funding is done for loan amount
- funded_amnt_by_investor : same as funded_amnt
- int_rate : Interest rate has spike at 7.5 and the mostly given at 10 to 15 percentage
- installment : These are evenly distributed
- annual_income : Annual income is mostly distributed in 10k to 50k
- debt_to_income_ration : It can be seen clearly that if debt to income ratio follows normal curve uptil 25 then it dip's suddenly
- 30_day_delinq_2yrs & pub_rec : Nothing significant

  can be deduced out of this variable
- inq_last_6mths : We have maximum 0 records
- credit_accounts : Most people have 10 to 30 credit accounts
- job_experiance : 10+ years have maximum entries since it covers larger experience data set also indicating larger age group





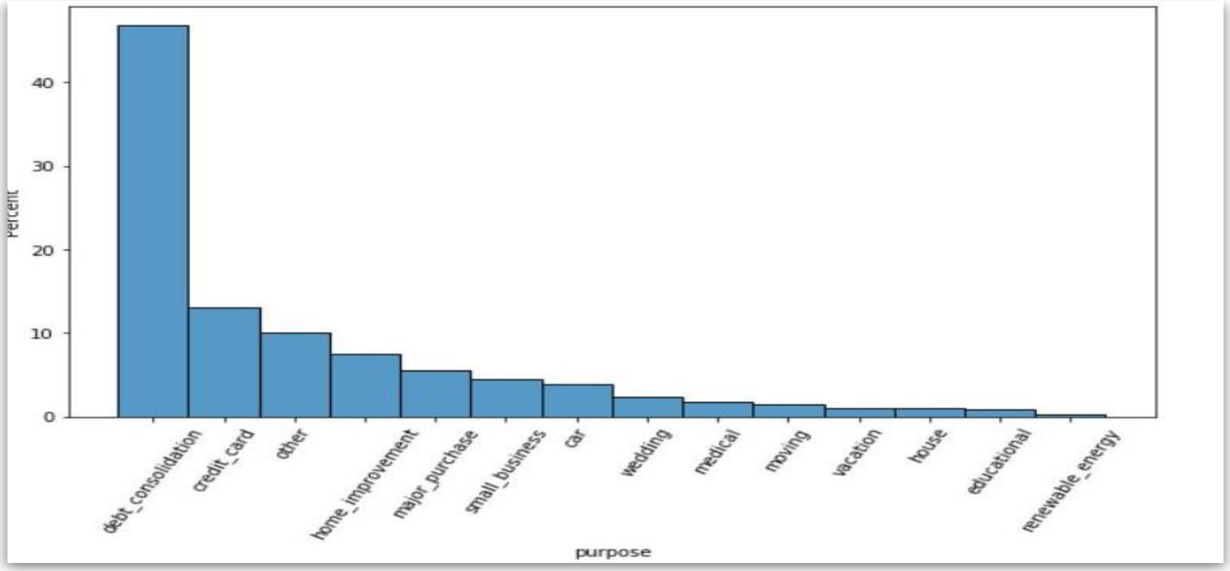
Categorical Columns

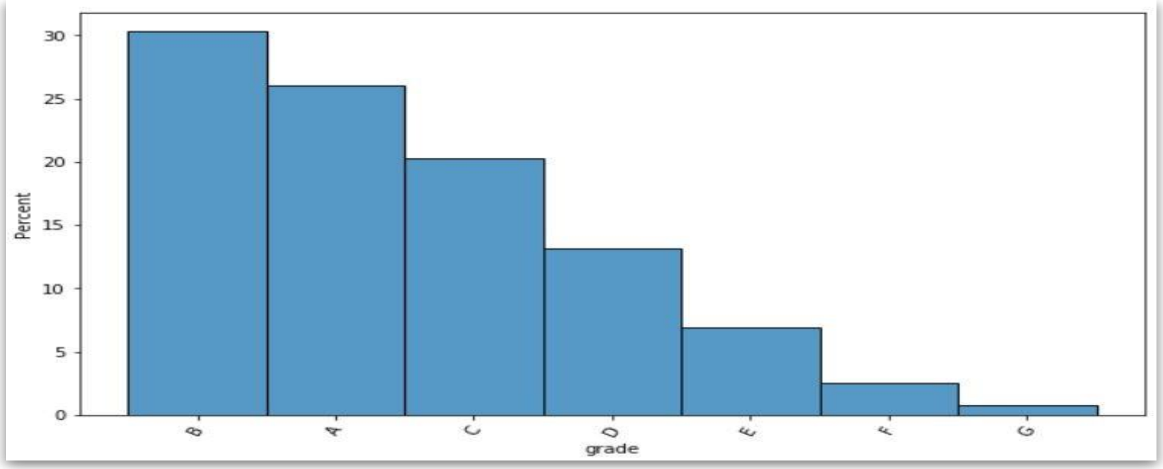
- term : Most people prefer 36 months duration
- grade : Maximum loans are from B, A, C and D grade
- sub_grade : maximum loans are from B grade,lower A and upper C sub grades
- emp_length : we have seen it in numerical columns that maximum loans are from 10 years + experience people
- home_ownership : majority of people live in rented

  or mortgage house
- loan_status : We can be seen that ~85% of loans are fully paid
- purpose : It can be seen that maximum loan is for

  debt_consolidation, which seems fishy, we will investigate further
- addr_state : It seems company is more active in

  CA,NY and FL states





Segmented - Numerical Columns

- loan_amnt : it can be seen that loan above 15k are more prone to default than loan below 15k
- funded_amnt : it is same as loan amount in terms

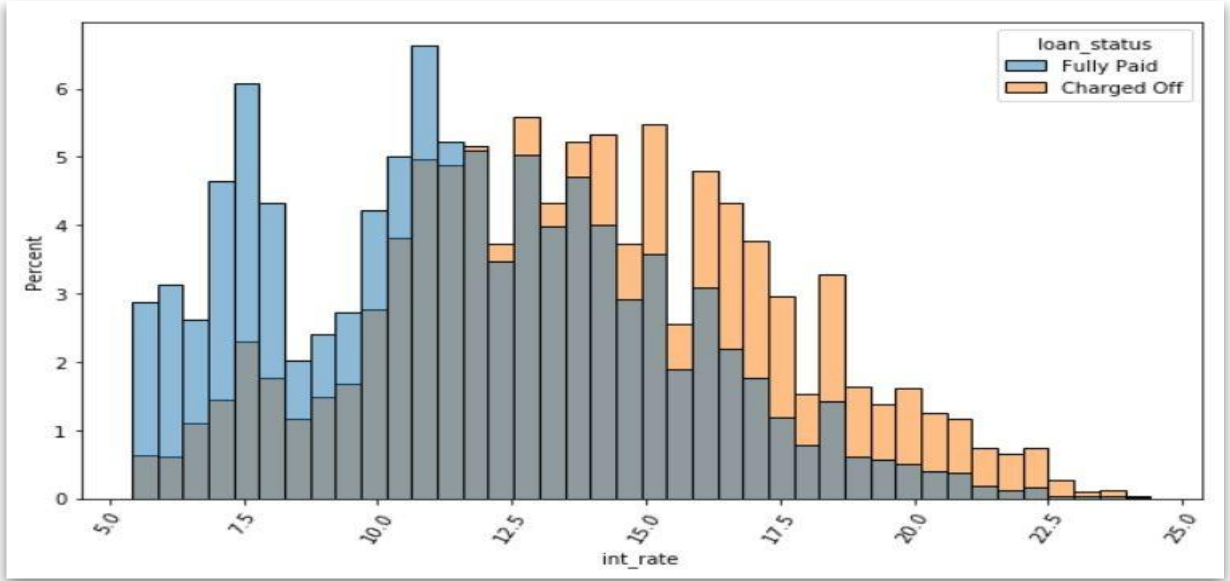
  of default
- funded_amnt_by_investor : same as funded_amnt
- int_rate : it can be clearly seen that interest rate more than 12.5 is more prone to default
- installment : we can see 100 to 300 number of

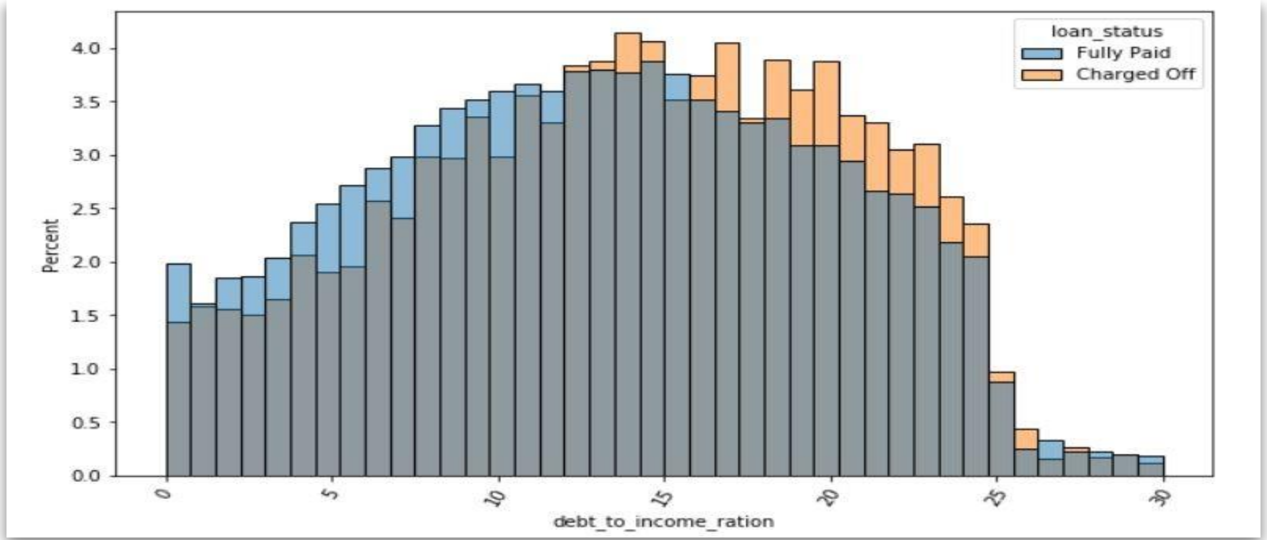
  installments are less prone to default
- annual_income : loans given to 0 to 10k annual income are prone to default
- debt_to_income_ration : It can be seen clearly that

  if debt to income ratio is greater than 12 then loan are more prone to default
- 30_day_delinq_2yrs & pub_rec : Nothing significant

  can be deduced out of this variable
- inq_last_6mths : lesser or 0 inquiry means better for full paid
- credit_accounts : 15 or lesser credit accounts are

  more prone to default
- job_experiance : Surprisingly lesser job experience applicants are less prone to default





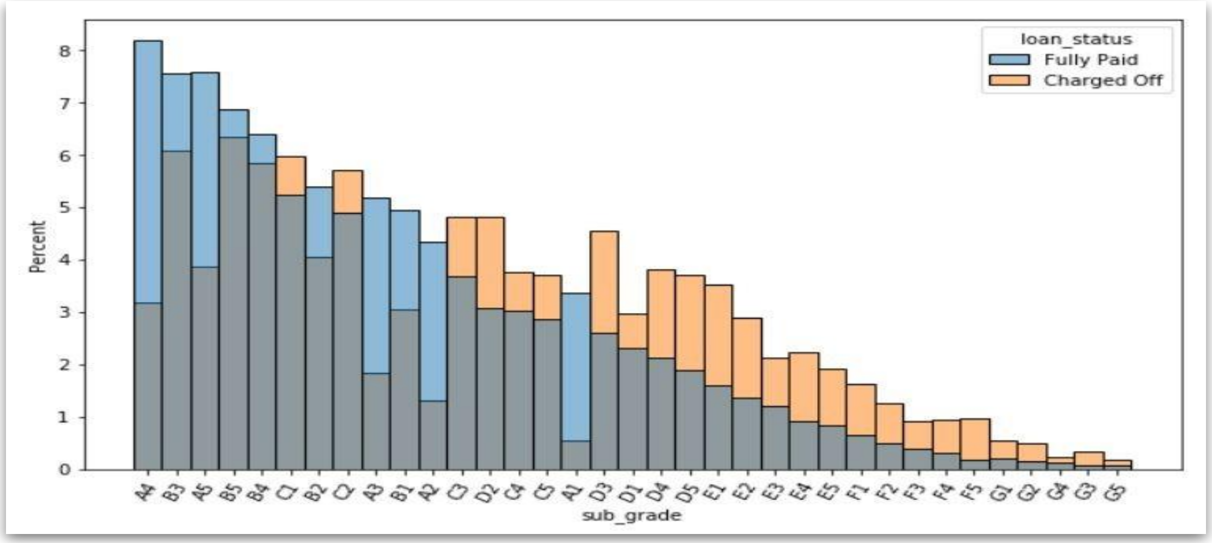
Segmented - Categorical Columns

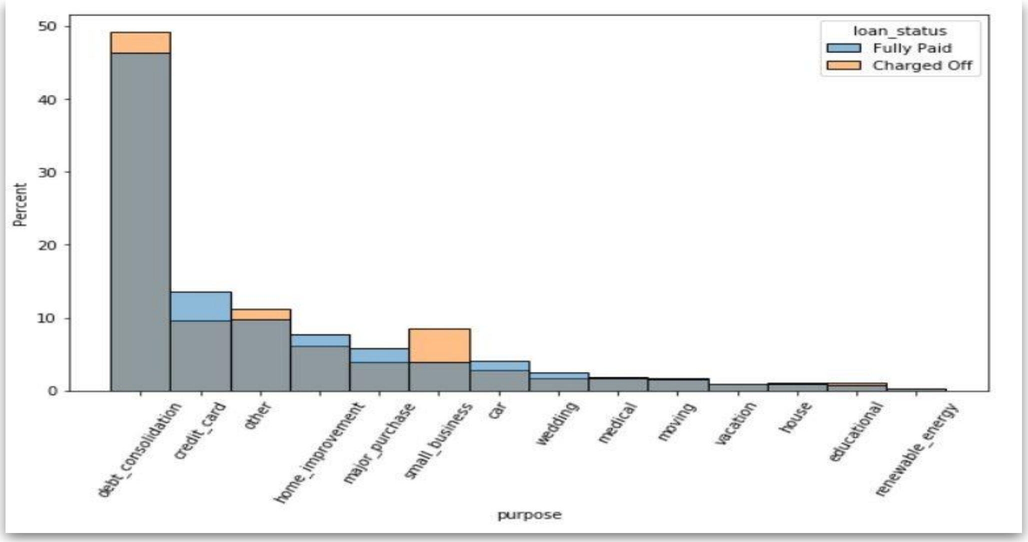
- term : 60 Months term is more prone to default
- grade : lower grade from C are more prone to default than A and B grade
- sub_grade : higher grade means lower default
- emp_length : it is surprising that lower years of experience has lesser default
- home_ownership : people staying on rent are more

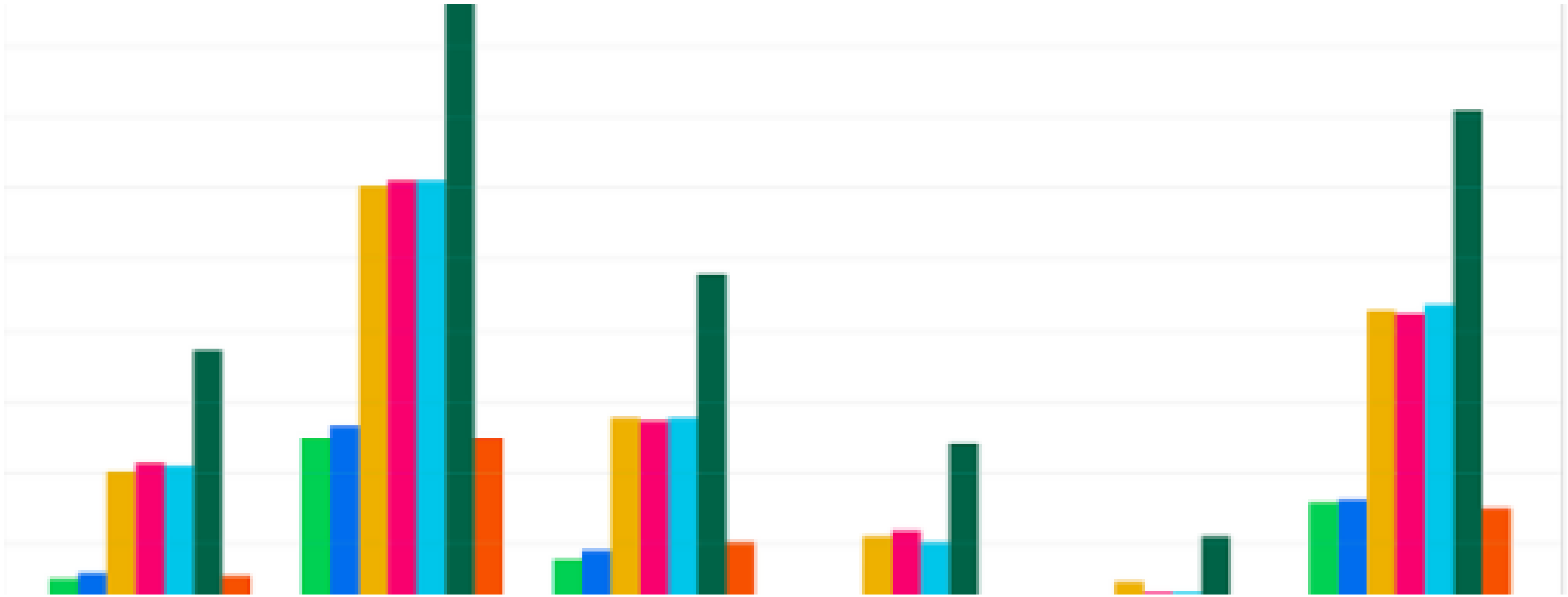
  prone to defaulting the loan than others
- verification_status : surprisingly Not verified status are less prone to default than verified
- purpose : Small business and debt consolidation

  are more prone to default
- addr_state : CA, FL and NV are more prone to default than other states
- pub_rec_bankruptcies : any record or unidentified

  records are more prone to default than 0 records







Bivariate Analysis

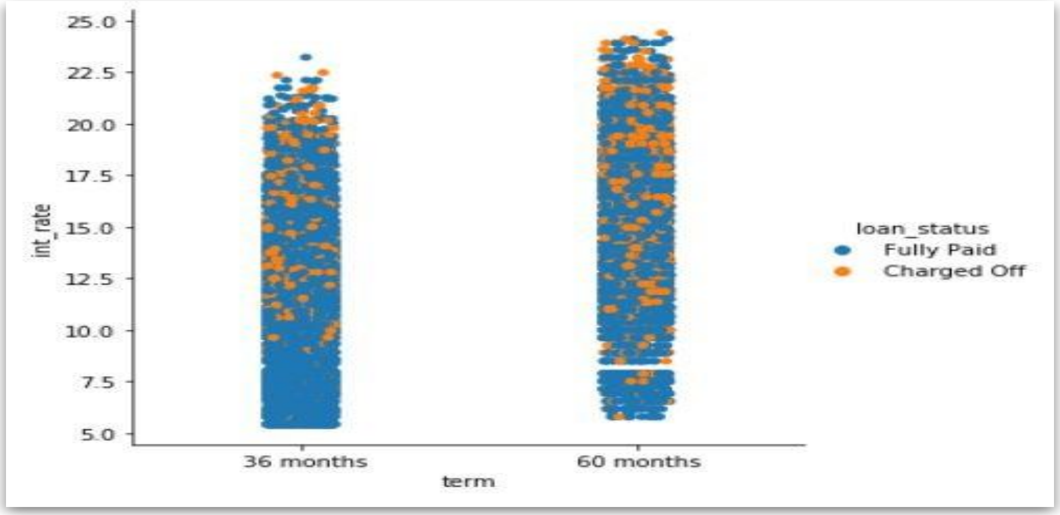
Numerical Columns

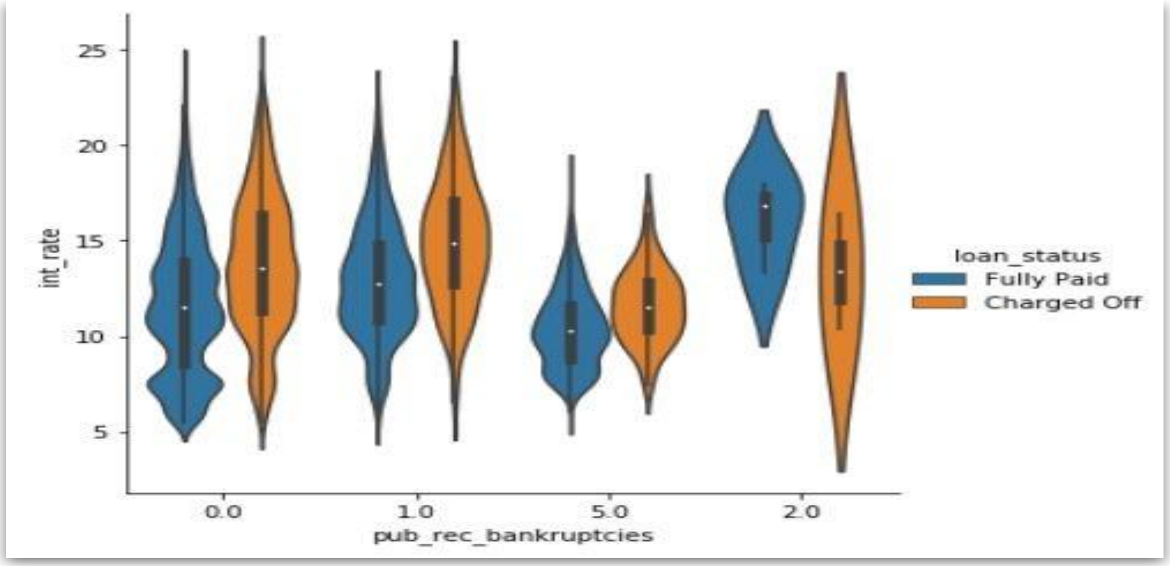
- There exist strong correlation between loan_amount/funded_amnt/funded_amnt_inv as mostly all loans are funded
- Higher amount means higher installments which can be seen here with strong correlation.
- Relatively positive correlation between loan_amnt and annual_income can be see which is obvious.
- We can observe negative correlation between

  annual_inc and dti
- higher loan amount attract higher interest rates that can be seen here

- term vs int_rate : We can see that lower term and lower interest rate has higher full payment status
- grade vs loan_amnt and int_rate : higher interest rate and lower grades are prone to default
- home_ownership vs loan_amnt, int_rate and credit_accounts : home ownership of mortgage type with higher loan mount is prone to default, also interest rate is not driving factor in this category and credit_account does not have any relation with home ownership
- verification_status vs job_experiance : verified loan with lesser job experience are less prone to default
- purpose vs loan_status : small business as purpose has highest default rate
- emp_length vs int_rate : there is no relation between emp_length vs int_rate all emp_length data are equally distributed
- pub_rec_bankruptcies vs int_rate : people with pub_rec_bankruptcies has higher chances of default irrespective of interest rate

Categorical Columns





Recommendation

Recommendation

- **grade** -Grades given by LC are C,D,E,F or G
- **term** -Loan tenure selected is 60 months
- **home_ownership** -Home ownership is Rent
- verification_status -Verification status is Verified
- **purpose** -Purpose of the loan is Small Business or Debt Consolidation
- **add_state** -States are one of the following CA, FL or NV
- **pub_rec_bankruptcies** -Any public record of bankruptcies
- **loan_amnt** -Loan amount is 15K or more
- **int_rate** -Interest rate is 12.5 or more
- **annual_income** -Annual income is 10K or less than 10K
- **debt_to_income_ration** -DTI is 12 or more
- **inq_last_6mths** -2 or more inquiries in last 6 months
- **credit_accounts** -Credit accounts are 15 or less
- **job_experiance** -Job experience is 5 or more

We cannot simply consider one or two variable to predict the default, we have defined 14 parameters which needs to be considered while giving loan.

In an a application if more than 5 matching parameters are there then we should avoid giving loan for such applications

Thank You

Technologies used and versions

- pandas 1.3.5
- numpy 1.21.6
- matplotlib 3.1.1
- seaborn 0.12.2

Reference

https://stackoverflow.com/questions/51070985/find-out-the-percentage-of- missing-values-in-each-column-in-the-given-dataset

https://www.geeksforgeeks.org/select-columns-with-specific-data-types-in- pandas-dataframe/

https://www.geeksforgeeks.org/how-to-set-a-seaborn-chart-figure-size

https://stackoverflow.com/questions/63373194/how-to-plot-percentage-wit h-seaborn-distplot-histplot-displot

https://stackoverflow.com/questions/56942670/first-and-last-row-cut-in-hal f-of-heatmap-plot

https://www.geeksforgeeks.org/multi-plot-grid-in-seaborn/

https://seaborn.pydata.org/generated/seaborn.histplot.html

https://seaborn.pydata.org/generated/seaborn.pairplot.html