EDA-Lending Club Case Study

Settu Murugan

Sid Sinari

Problem Statement

- You work for a **consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has 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 in "loan.csv" contains information about past loan applicants and whether they 'defaulted' or not. The aim is 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,

Objectives:

- **Using Exploratory Data Analysis we need to understand how consumer attributes and loan attributes influence the tendency of default.**
- When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    2. **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'.
    3. **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
2. **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)

Loan Data:

- Shared loan.csv file contains 39717 Rows and 111 columns.
- Data Dictionary provided to explain the details of 111 columns
- This file contains several details about the loan like, amount, interest rate, distribution date, terms and etc..
- This file contains several details about customers like, ID, Address, grade, annual income, employment length, monthly debt paying ratio and etc..

1. Data Cleaning

- The first part of EDA is Data cleaning. Looking at the head and tail of data frame Looks like there are no header and footer present. Also no summary columns available in the data frame.
- After checking for duplicate, could not find any duplicate rows.
- Found 54 null columns.

['mths_since_last_major_derog', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit']

- Deleted all the above columns having Null values as it is not useful for analysis.
- Deleted the columns having unique values since it may not be useful for analysis (*id, member_id, url, chargeoff_within_12_mths*)
- Now Delete all the columns which are having all the values as same.. also may not be usefull for further analysis(*pymnt_plan','initial_list_status', 'collections_12_mths_ex_med*).
- Some columns having same values across all loan Ids, hence may not be useful in the analysis. Hence removing them *(sub_grade','total_rec_late_fee','last_pymnt_amnt','last_pymnt_d','next_pymnt_d', 'last_credit_pull_d', 'policy_code','application_type','acc_now_delinq','delinq_amnt','tax_liens','earliest_cr_line')*
- *#Now Lets drop all columns having more than 40% values are null (mths_since_last_delinq', 'mths_since_last_record)*
- *Some columns values are grouped together for better analysis example., Purpose column has home_improvement , house both are having the same meaning in the purpose. So treating them as same value.*
- *Now we do not need a title or description of the loan as it is already categorized properly with "purpose" (desc','title','emp_title). Hence delete it.*

2. Data Standardization

- We need to trim some characters or strings in the values for proper analysis ex, remove months in term, % in *int_rate and revol_util.*
- ***emp_length*** has 2.706650% values as empty. Covert the column dtype to float, find the median(4.0 here) and fill the missing values with it. Similarly do it for other colums (***revol_util, pub_rec_bankruptcies***).
- converting Columns to proper data type. *loan_amnt','funded_amnt','int_rate*','*revol_util* to float , *term* to int *and issue_d to* datetime.

Segmentation

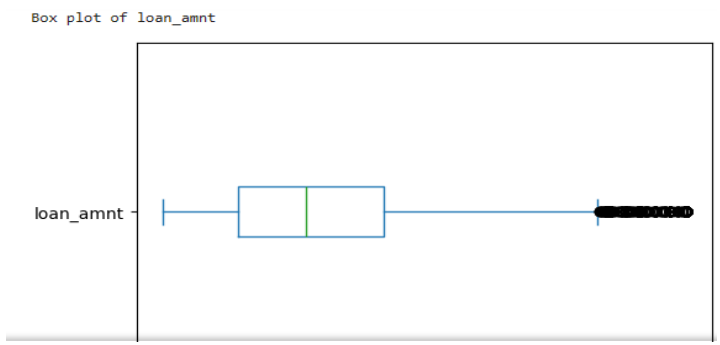
- Segmentation done based on numerical and categorical variables
- numerical_column=

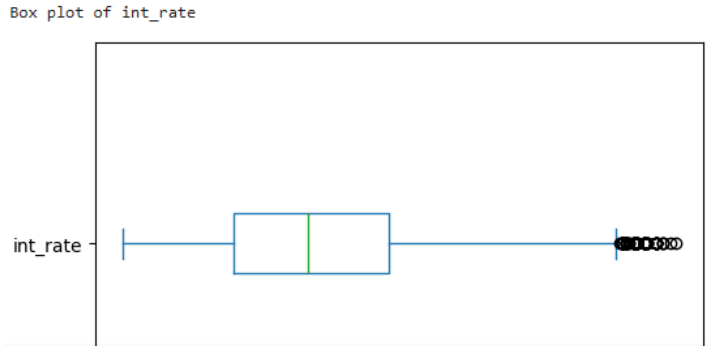
  ['loan_amnt','funded_amnt','funded_amnt_inv','term','int_rate','installment','emp_length','annual_inc','dti','delinq_2yrs','inq_last_6mths','open_acc','pub_rec','total_acc','pub_rec_bankruptcies']
- categorical_column =

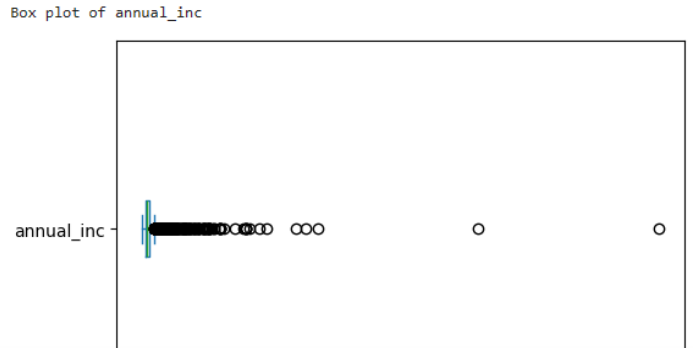
  ['grade','home_ownership','verification_status','issue_d','loan_status','purpose','addr_state']
- *Followingly we do univariate analysis. We draw the box plot to find out the outliers and remove the outliers.*

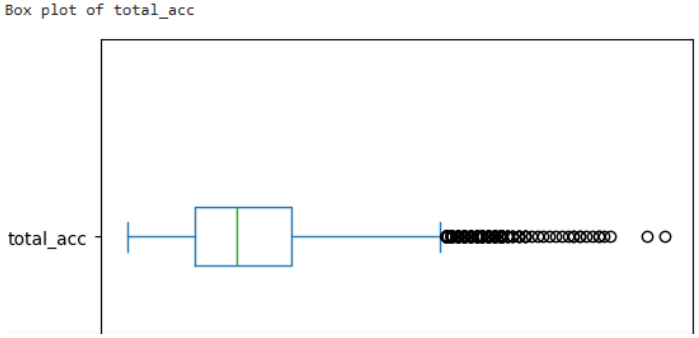
3.Univariate Analysis

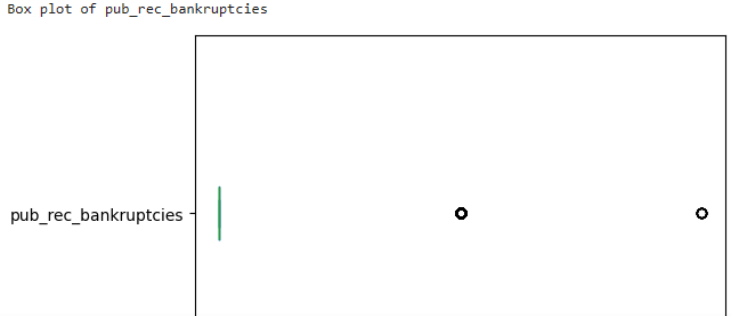
- Univariate Analysis on Numerical Colums: Draw Box plot for numerical columns to understand the outliers values .











- **Outliers removed using following formula**

Q1 = df[col].quantile(0.25)

Q3 = df[col].quantile(0.75)

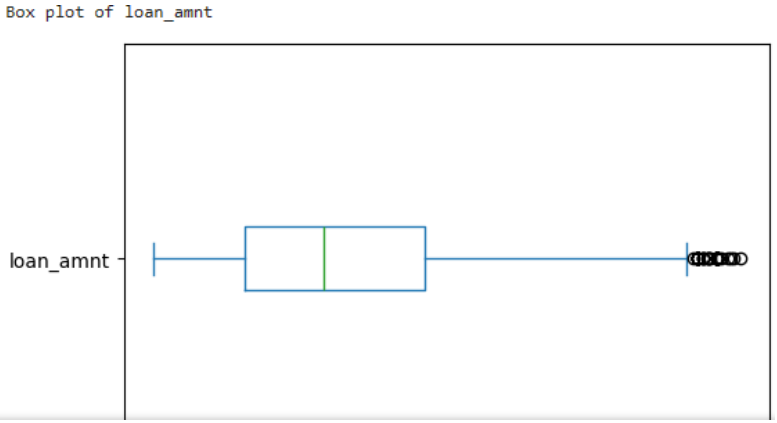
IQR = Q3 - Q1

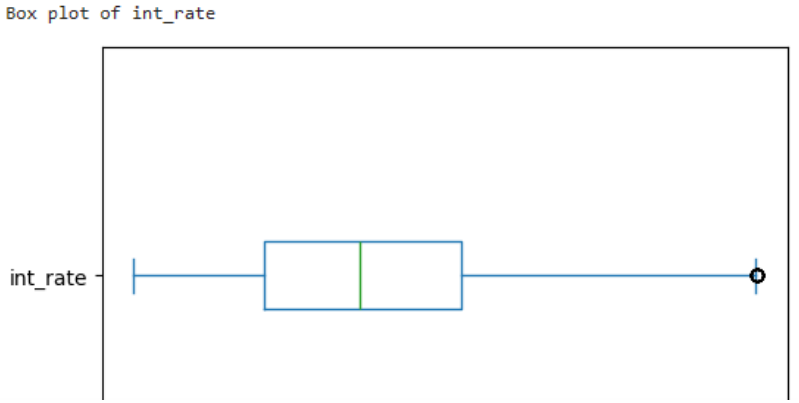
Upper_Boundary = Q3 + threshold * IQR

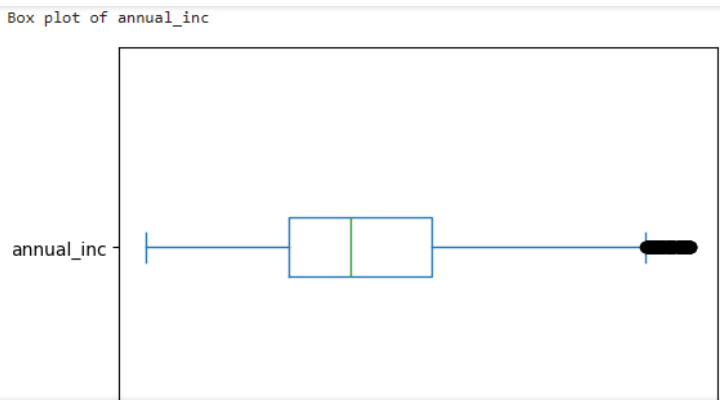
Lower_Boundary = Q1 - threshold * IQR

df = df [(df[col] >= Lower_Boundary) & (df[col] <= Upper_Boundary)]

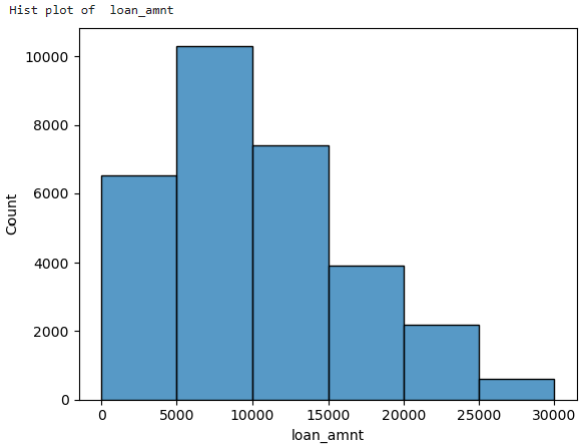
After removing outliers…





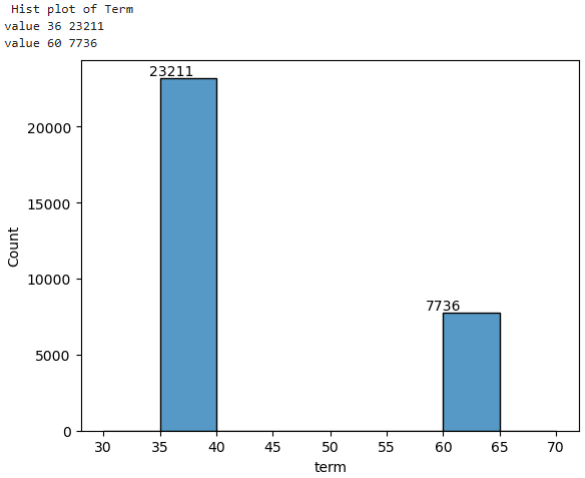


Distribution of loan amount



- Max number customer got loan from 5000 to 15000.
- Min is 500 and Max is 29100.

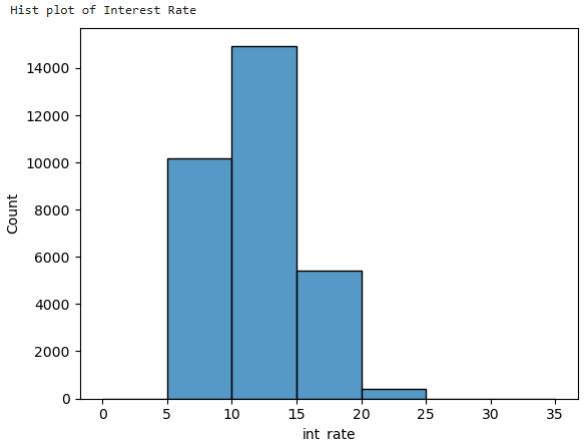
Loan Term



Only two terms of loans

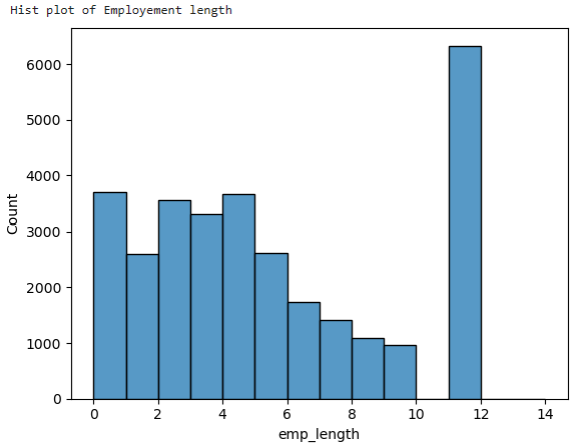
- 36 months and 60 months

Interest rate



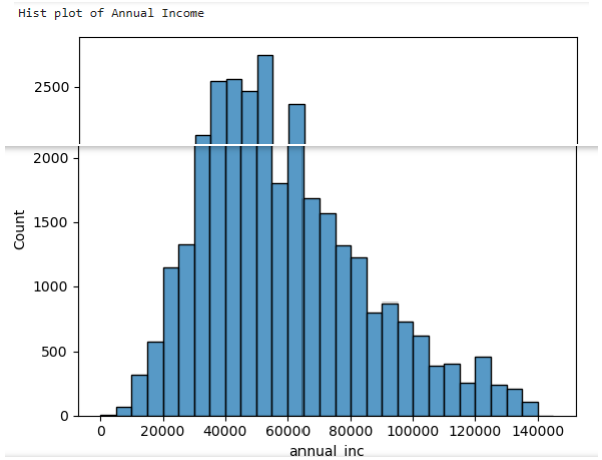
Most of the loan distributed with interest rate ranging from 10 to 15

Employment length



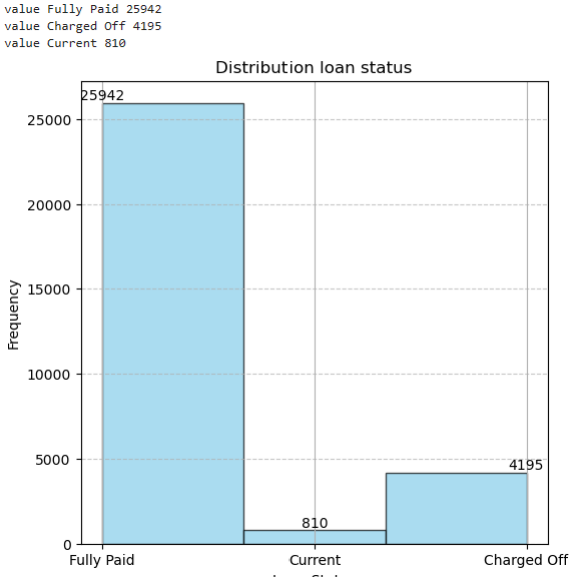
Max, Customer's employment length
Varies from 1 year to 6 years

Annual income distribution

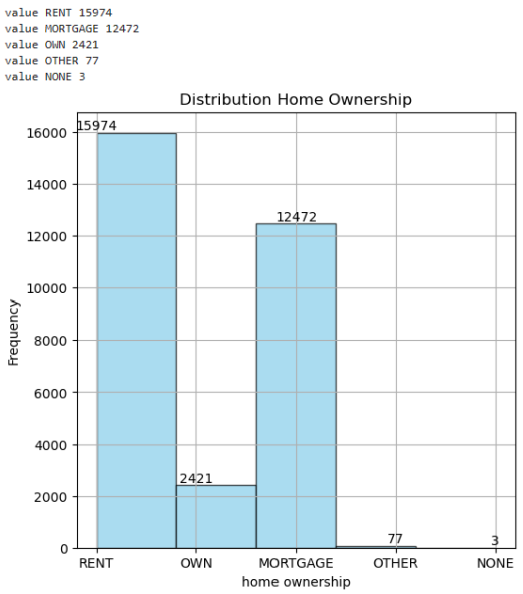


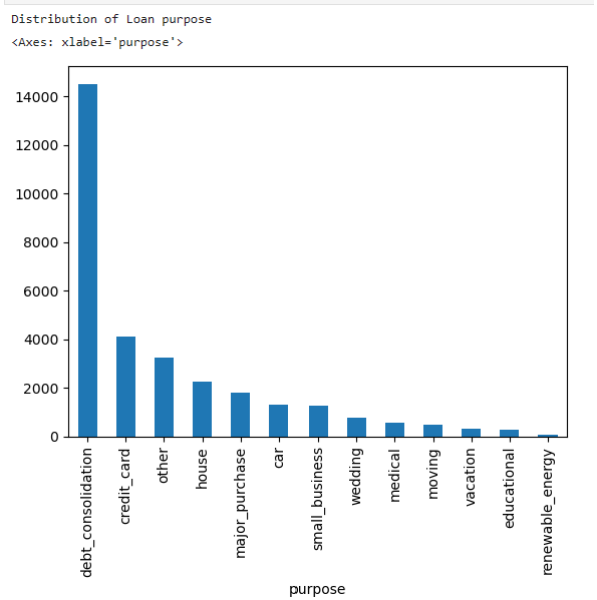
- Min income 4k, Max 137k.
- Most of the them in range of 20k to 85k

Un ordered Categorical – Univariate Analysis



Un ordered Categorical – Univariate Analysis





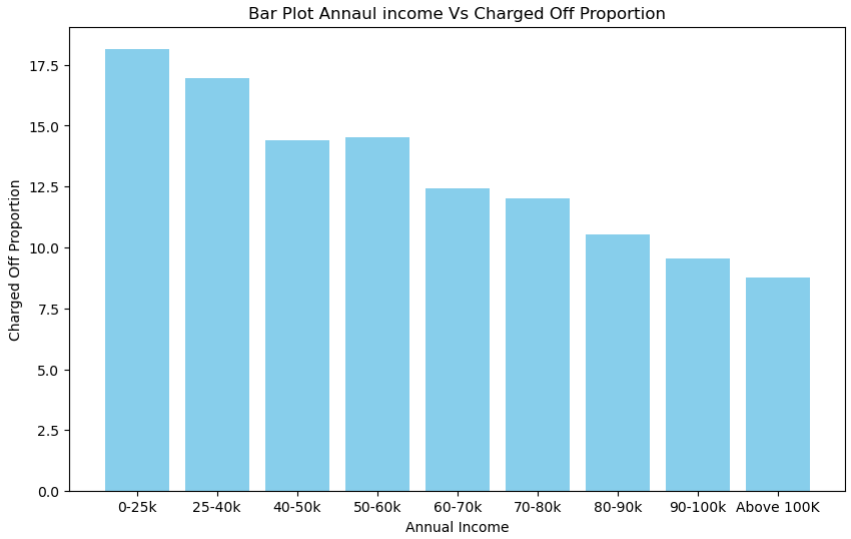
Observations:

- Most of the customers fully paid their loan.
- Less than 5% of customers had charged off.
- Most of the customers living in Rent and Mortgaged houses
- Maximum loan is taken for the purpose of debt consolidation.

4.Bivariate Analysis – Relationship between two variables

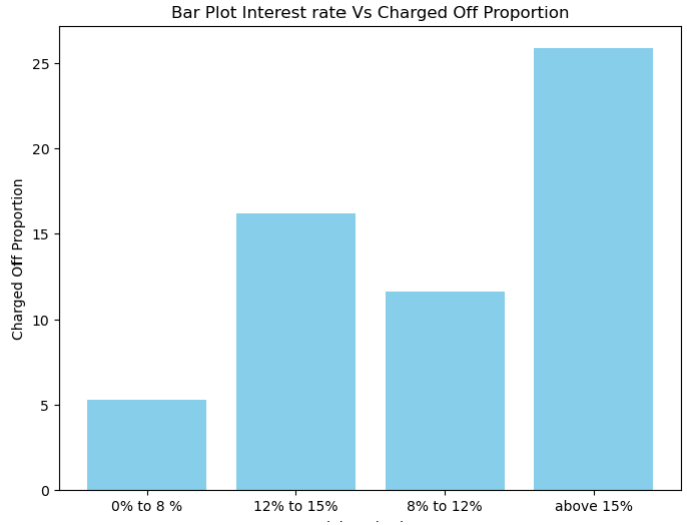
- Annual income, loan amount, interest rate are continuous variables. So for better analysis lets bucketize this values and create different data set to get better insights.

Annual Income vs Charged Off Proportion



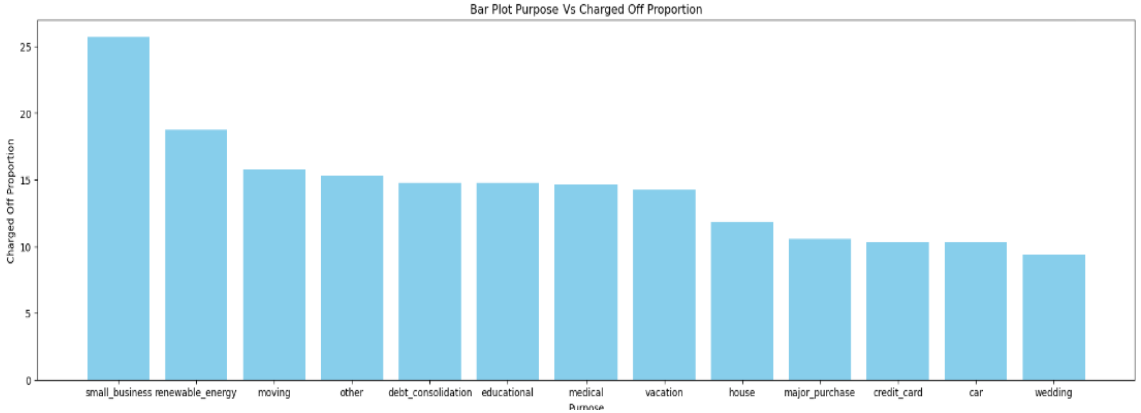
- Customers who get less salary likely to default the loan.
- 35% of customers who defaulted having salary less than 40k

Interest rate vs Charged Off Proportion



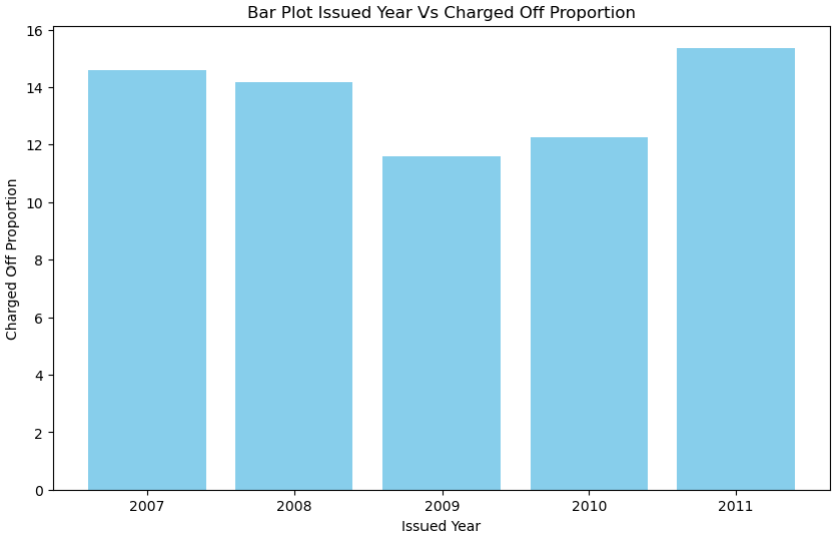
- Higher interest rate loans - high percentage of charged off
- Int rate greater 12% having higher percentage of charged off

Purpose vs Charged off proportion



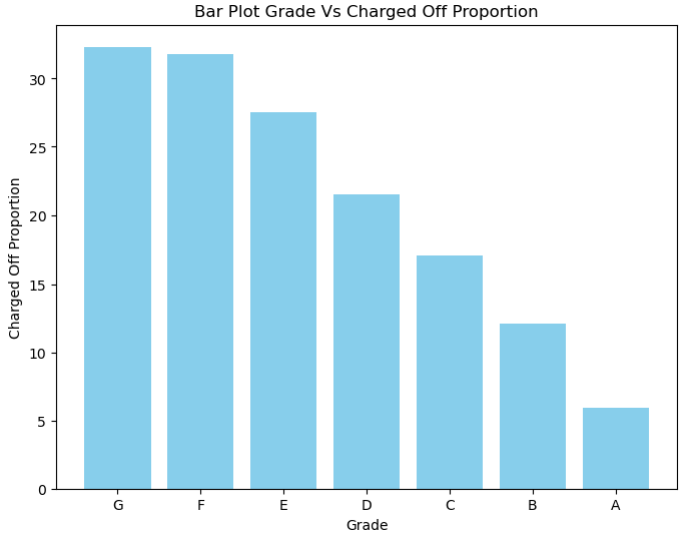
Loan given for Small business and renewable energy having highest percentage of charged off

Loan issued year Vs Charged off

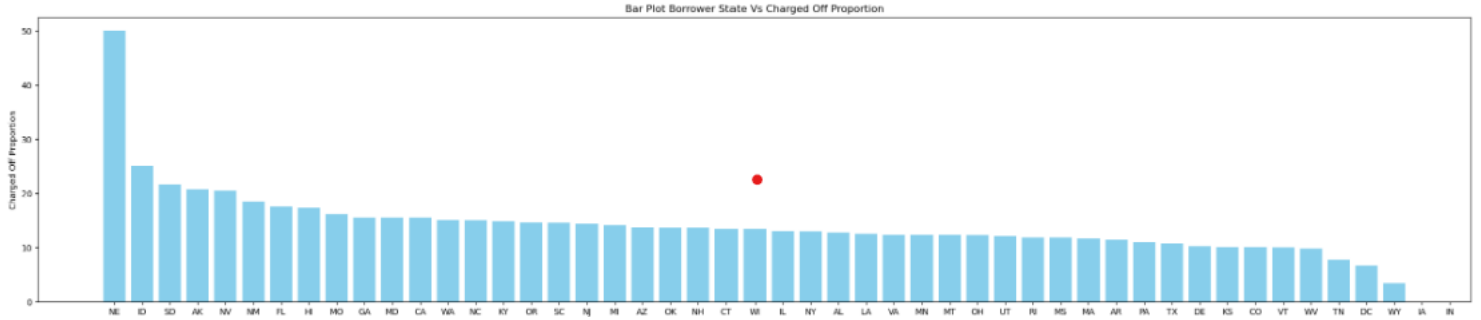


Loan issued in 2011 having higher charged off rate.

Customers grade vs Charged off proportion

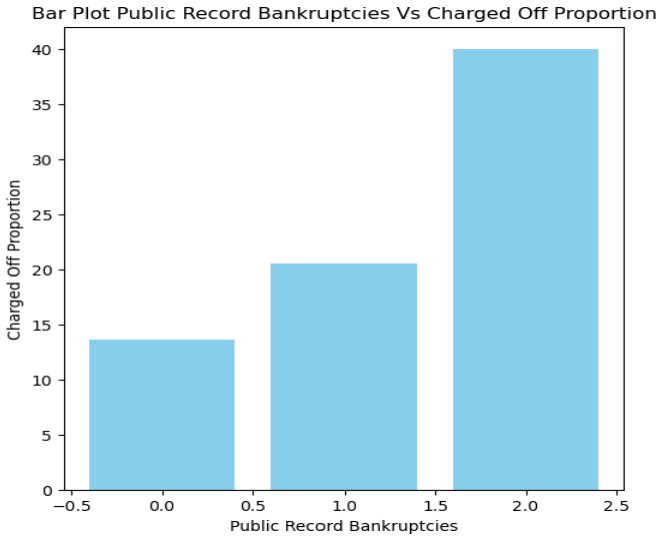


Customers State vs Charged off proportion



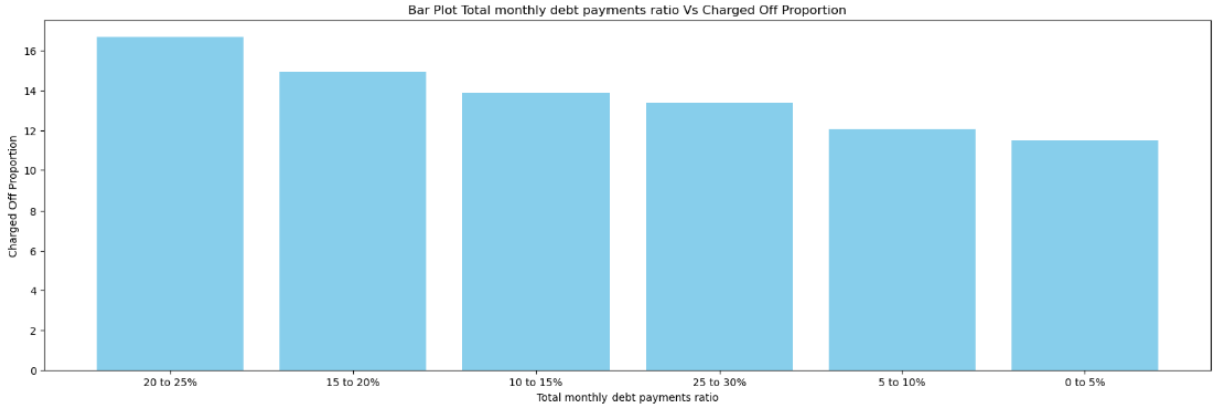
- G and F grade employees having higher charged off proportion
- Customer who belong to NE state having higher charged off proportion

Public Record Bankruptcies VS Charged off proportion



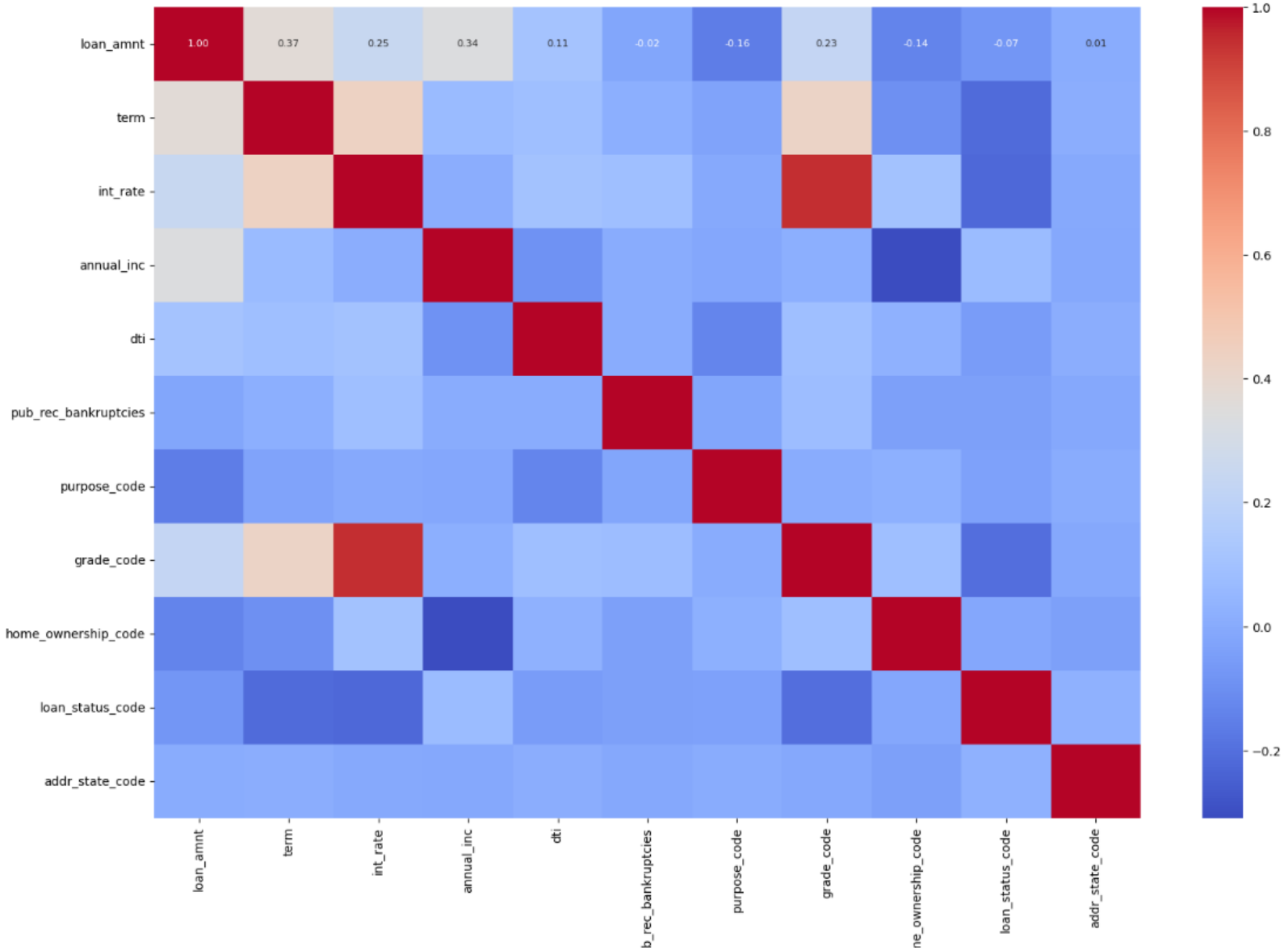
Customers having more public record bankruptcies having higher charged off proportion.

Monthly debt payments ratio VS Charged off proportion



Customers having monthly debt payment ratio Greater than 10% having higher charged off proportion

Correlation:



Positive Correlation

- Loan amount and term having positive correlation
- Annual income and loan amount having positive correlation
- Interest rate and term having positive correlation

Negative Correlation:

- Loan amount and loan purpose having negative correlation
- Annual income and home ownership having strong negative correlation.

Summary:

- Lower the income group having higher percentage of charged off.
- Interest rate above 12% is having higher percentage of Charged off
- Small business and renewable energy loans having higher percentage of charged off.
- The loan distributed in 2011 and Dec month having higher percentage of charged off.
- G and F grade customers having higher percentage of charged off.
- Customers who are having higher number of public record bankruptcies having higher percentage of charged off.
- Customers having above 15% total montly monthly debts payment having higher percentage of charged off.