

# **Exploratory Data Analysis**

## **Lending Club Case Study**

**By**

**Sirin Shaikh**

**6-Dec-2023**

## Problem Statement:

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 are labelled as 'charged-off' are the 'defaulters'.

Task is to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

Find out the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

Those insights will be utilized for portfolio and risk assessment.

## Catalog

Analysis Approach: .....	3
Data Understanding .....	3
Data Cleaning .....	3
Data Standardization .....	3
Outlier Identification and Treatment .....	3
Statistical Summary of all columns .....	4
Univariate Analysis with plots .....	5
Bivariate Analysis .....	6
Summary of Bivariate Analysis: .....	9
Multivariate Analysis - Correlation matrix .....	10
Recommendation .....	10

# Analysis Approach:

## Data Understanding

Following are work done as part of this task:

- Checking the number of rows and columns
- Find out the column with null values and their percentage count
- Check for duplicate value
- Find out the column with zero values and their percentage count

## Data Cleaning

- Drop the columns where null percent is greater than 60%
- Drop the column with zero values percent is greater than 80%
- Drop additional columns which does play role in analysis like url, title etc
- Change column header name with proper readable names
- Drop all the rows where the column loan\_status has "Current" value because it its ongoing thing and we cannot get insights for out target data ie identifying loan defaulters

## Data Standardization

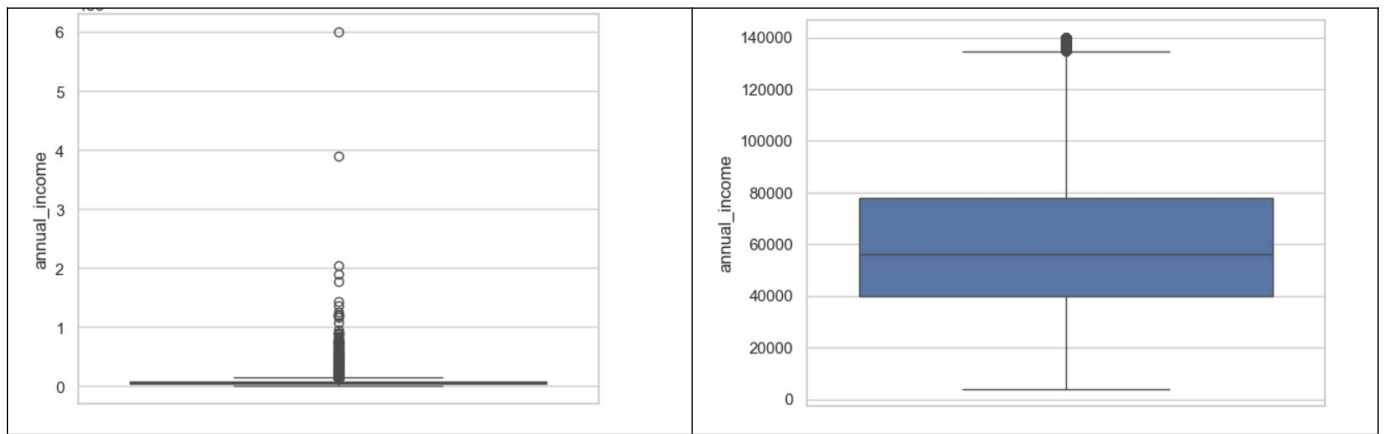
- Check for missing values and replace with mean, mode or median. In this case “**tenure**’ column was having missing values, hence filled all the null values with mode value ie value of 10+ years.
- Removal of months string from values in **loan\_period** cols
- Replace the years/year/+ in **tenure** with blank. Make year < 1 as 0. Make the tenure 10+ as 10
- Bifurcate “Issue\_d” into month and year columns

After doing data understanding, cleaning and standardization we created the target dataframe for analysis namely **loan\_data\_cleaned\_df**.

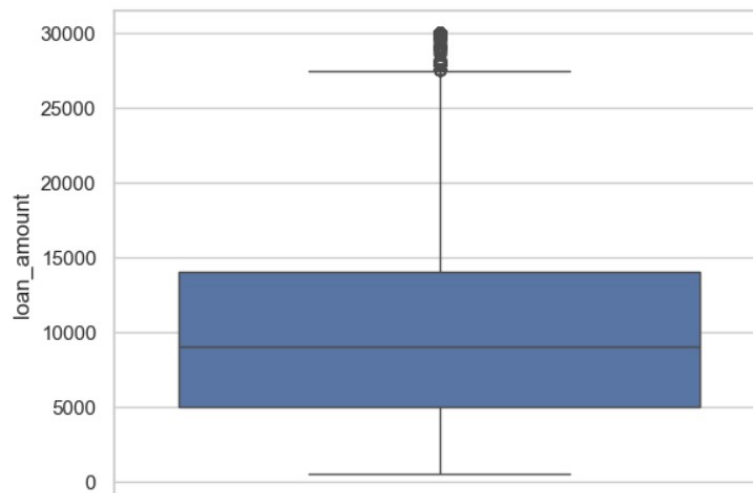
## Outlier Identification and Treatment

- Following numerical columns have been analyzed for outliers
  - **Annual\_income**

Outlier	After removal
---------	---------------



## ■ Loan\_amount



Similar approach has been taken for removing outliers in Loan\_amount. Though we don't see much change, we can live with it as Majority of Values for the loan amount are distributed continuous.

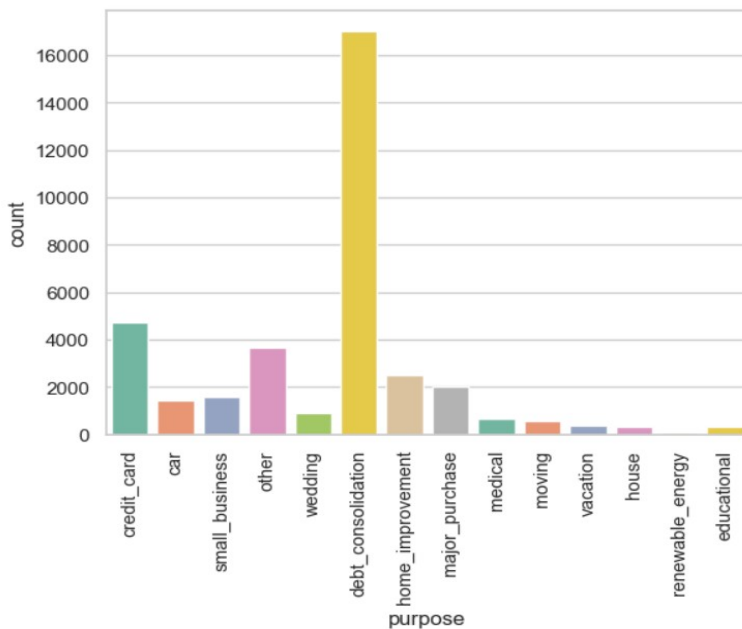
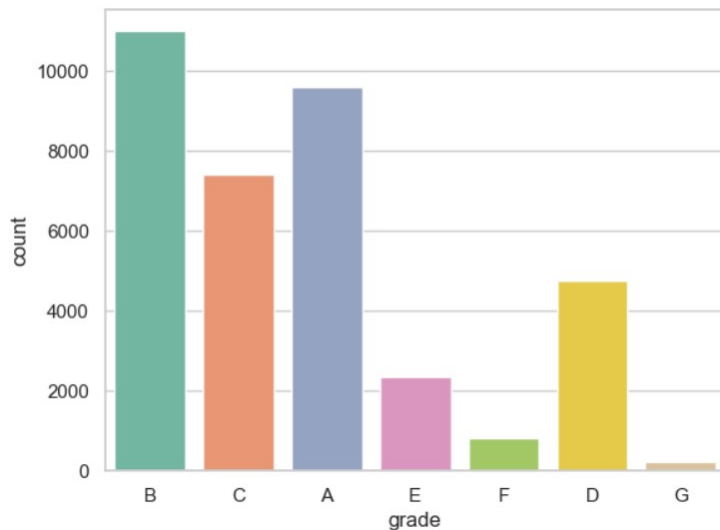
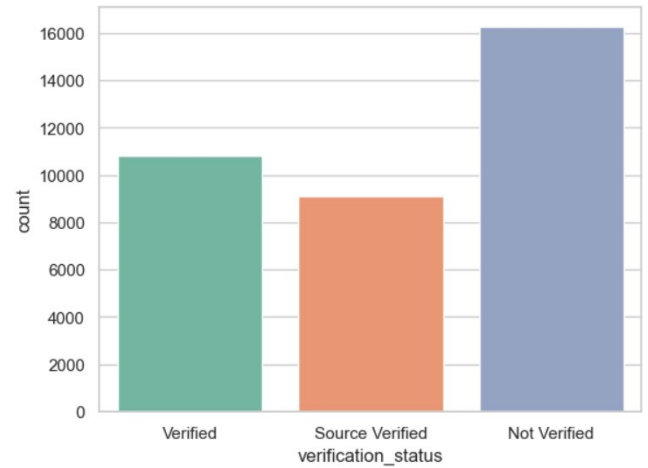
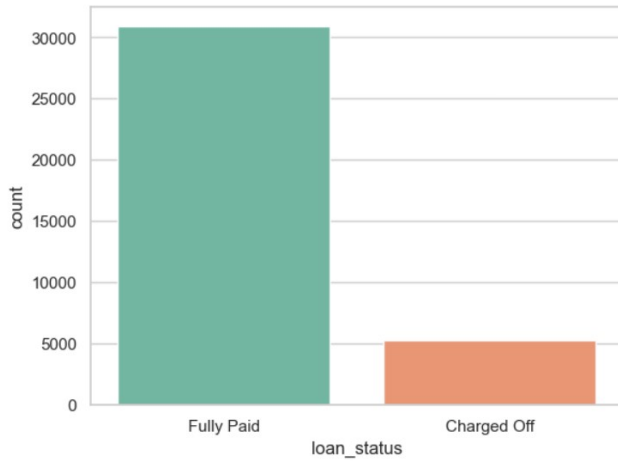
## Statistical Summary of all columns

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
loan_amount	36205.0	NaN	NaN	NaN	10138.444966	6290.365896	500.0	5000.0	9000.0	14000.0	30000.0
loan_period	36205.0	NaN	NaN	NaN	41.70087	10.213904	36.0	36.0	36.0	36.0	60.0
int_rate_percentage	36205.0	NaN	NaN	NaN	11.840919	3.637771	5.42	8.9	11.71	14.27	24.4
installment	36205.0	NaN	NaN	NaN	304.211137	187.342322	15.69	162.25	267.33	402.49	1106.07
grade	36205	7	B	11006	NaN	NaN	NaN	NaN	NaN	NaN	NaN
sub_grade	36205	35	A4	2731	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tenure	36205.0	NaN	NaN	NaN	5.022511	3.592407	0.0	2.0	4.0	9.0	10.0
home_ownership	36205	5	RENT	17912	NaN	NaN	NaN	NaN	NaN	NaN	NaN
annual_income	36205.0	NaN	NaN	NaN	60446.755988	27568.200792	4000.0	40000.0	55200.0	76460.0	140004.0
verification_status	36205	3	Not Verified	16276	NaN	NaN	NaN	NaN	NaN	NaN	NaN
issue_d	36205	55	Dec-11	1925	NaN	NaN	NaN	NaN	NaN	NaN	NaN
loan_status	36205	2	Fully Paid	30915	NaN	NaN	NaN	NaN	NaN	NaN	NaN
purpose	36205	14	debt_consolidation	17030	NaN	NaN	NaN	NaN	NaN	NaN	NaN
state	36205	50	CA	6487	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_account	36205.0	NaN	NaN	NaN	21.556802	11.213071	2.0	13.0	20.0	28.0	90.0
total_payment	36205.0	NaN	NaN	NaN	11125.799557	7827.611744	0.0	5374.35	9262.13	15173.33	48155.65

## ● Insights from summary

- The average interest rate on loan is 11.8% where minimum is 5.43% and maximum is 24.4%
- The average annual income of borrowers is 60k where minimum is 4k and maximum is 140k
- Most of the loan tenure isn of 36 months

## Univariate Analysis with plots



## Summary of Univariate analysis:¶

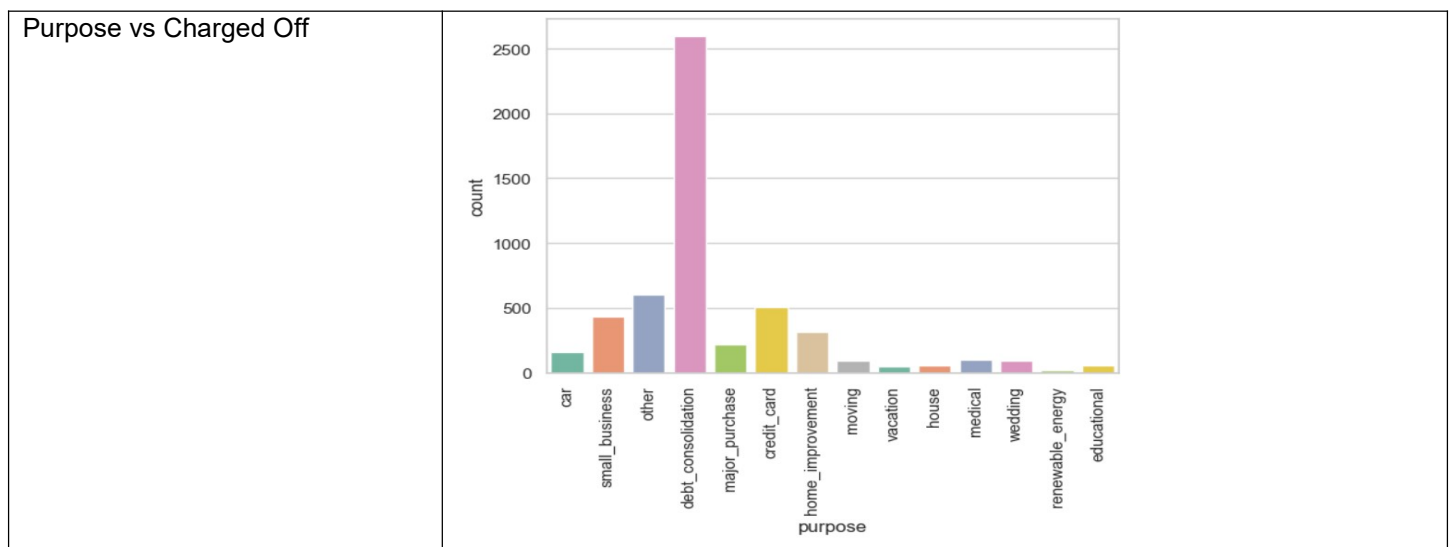
- 14% of loans are charged off.
- Near about 70% of loan were taken for debt consolidation
- Near 40% of loan is not verified.
- Most of the loan belongs to A,B,C grade where C being the highest.
- Most of the loan borrowers have their home either on RENT or Mortgage.

**Note:** loan is referred to the total dataset which contains all the rows containing either Fully Paid or Charged Off loan status.

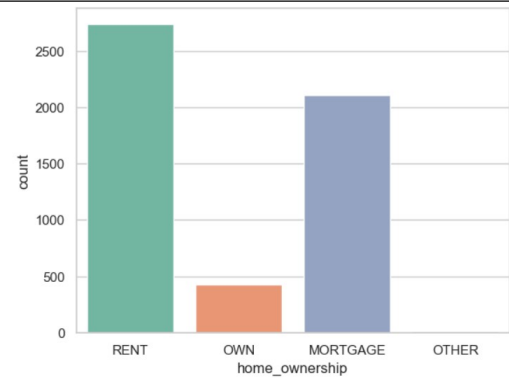
## Bivariate Analysis

Analyzed columns wrt charged off to check the impact:

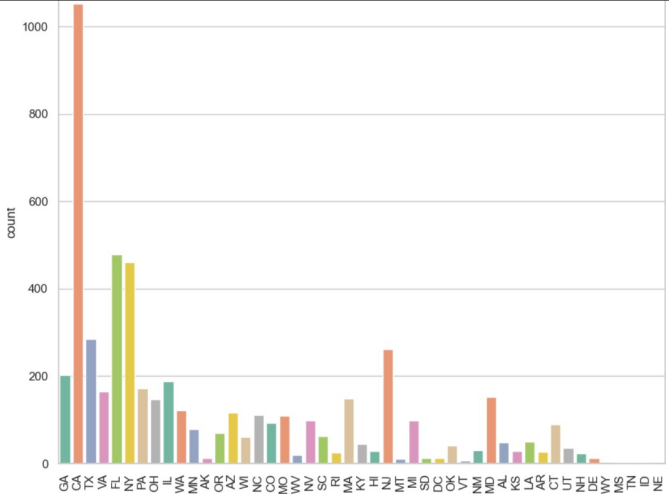
1. Purpose vs Charged Off
2. Home Ownership vs Charged Off
3. State vs Charged Off
4. Verification status vs Charged off
5. Grade vs Charged off
6. Tenure vs Charged off
7. Interest slab vs Charged off
8. Total account slab vs Charged off
9. Income slab vs Charged off
10. Loan tenure vs Charged off



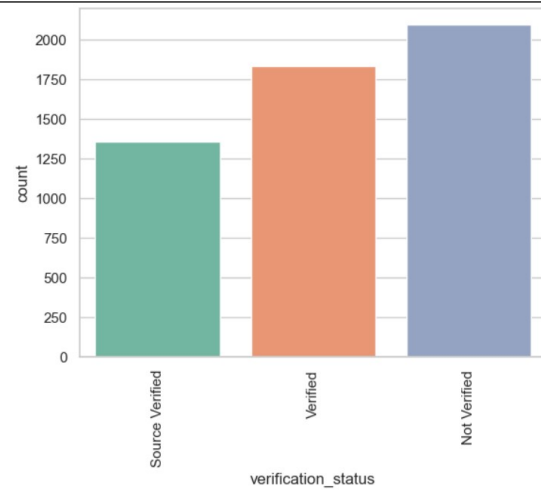
Home Ownership vs Charged Off



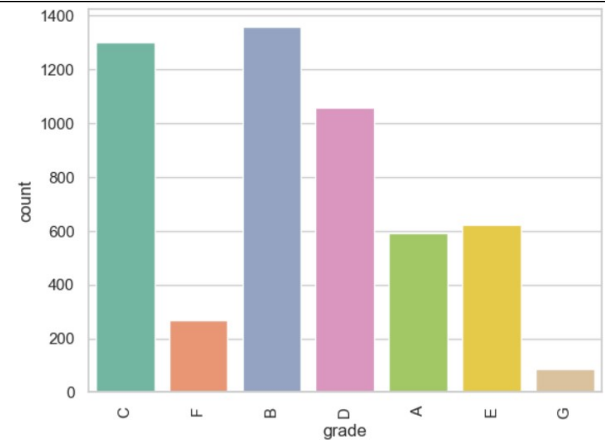
State vs Charged Off

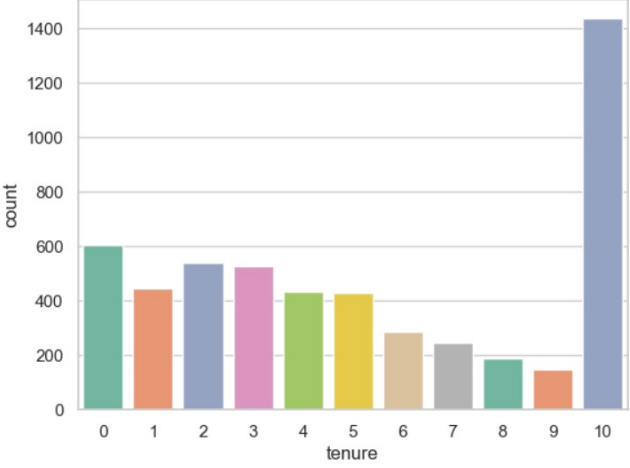
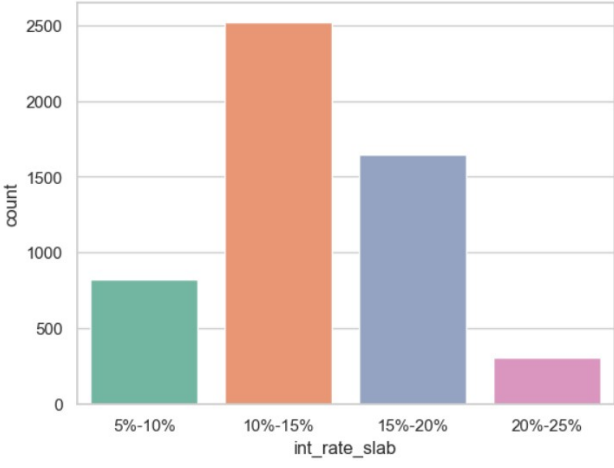
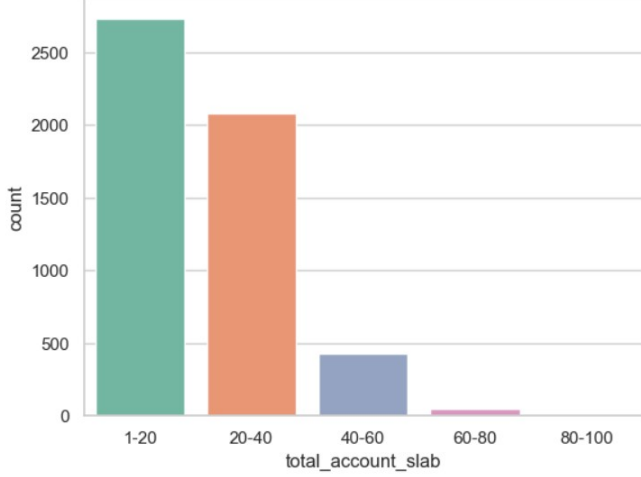
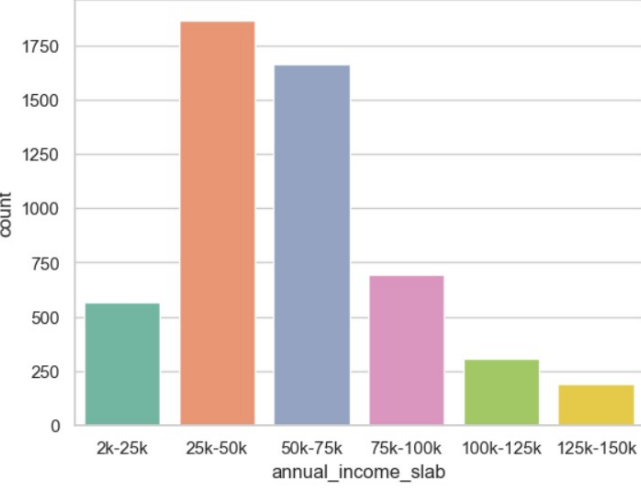


Verification status vs Charged off

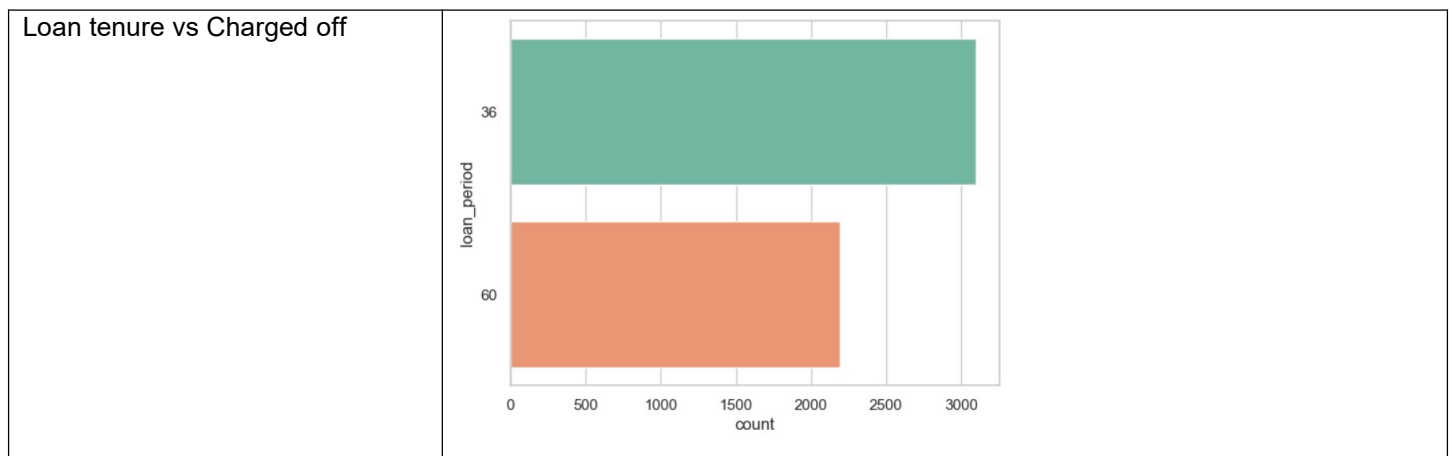


Grade vs Charged off



Tenure vs Charged off	 <table border="1"><thead><tr><th>tenure</th><th>count</th></tr></thead><tbody><tr><td>0</td><td>600</td></tr><tr><td>1</td><td>450</td></tr><tr><td>2</td><td>530</td></tr><tr><td>3</td><td>520</td></tr><tr><td>4</td><td>430</td></tr><tr><td>5</td><td>430</td></tr><tr><td>6</td><td>280</td></tr><tr><td>7</td><td>240</td></tr><tr><td>8</td><td>180</td></tr><tr><td>9</td><td>140</td></tr><tr><td>10</td><td>1420</td></tr></tbody></table>	tenure	count	0	600	1	450	2	530	3	520	4	430	5	430	6	280	7	240	8	180	9	140	10	1420
tenure	count																								
0	600																								
1	450																								
2	530																								
3	520																								
4	430																								
5	430																								
6	280																								
7	240																								
8	180																								
9	140																								
10	1420																								
Interest slab vs Charged off	 <table border="1"><thead><tr><th>int_rate_slab</th><th>count</th></tr></thead><tbody><tr><td>5%-10%</td><td>820</td></tr><tr><td>10%-15%</td><td>2500</td></tr><tr><td>15%-20%</td><td>1650</td></tr><tr><td>20%-25%</td><td>300</td></tr></tbody></table>	int_rate_slab	count	5%-10%	820	10%-15%	2500	15%-20%	1650	20%-25%	300														
int_rate_slab	count																								
5%-10%	820																								
10%-15%	2500																								
15%-20%	1650																								
20%-25%	300																								
Total account slab vs Charged off	 <table border="1"><thead><tr><th>total_account_slab</th><th>count</th></tr></thead><tbody><tr><td>1-20</td><td>2700</td></tr><tr><td>20-40</td><td>2050</td></tr><tr><td>40-60</td><td>420</td></tr><tr><td>60-80</td><td>30</td></tr><tr><td>80-100</td><td>0</td></tr></tbody></table>	total_account_slab	count	1-20	2700	20-40	2050	40-60	420	60-80	30	80-100	0												
total_account_slab	count																								
1-20	2700																								
20-40	2050																								
40-60	420																								
60-80	30																								
80-100	0																								
Income slab vs Charged off	 <table border="1"><thead><tr><th>annual_income_slab</th><th>count</th></tr></thead><tbody><tr><td>2k-25k</td><td>550</td></tr><tr><td>25k-50k</td><td>1850</td></tr><tr><td>50k-75k</td><td>1650</td></tr><tr><td>75k-100k</td><td>680</td></tr><tr><td>100k-125k</td><td>300</td></tr><tr><td>125k-150k</td><td>180</td></tr></tbody></table>	annual_income_slab	count	2k-25k	550	25k-50k	1850	50k-75k	1650	75k-100k	680	100k-125k	300	125k-150k	180										
annual_income_slab	count																								
2k-25k	550																								
25k-50k	1850																								
50k-75k	1650																								
75k-100k	680																								
100k-125k	300																								
125k-150k	180																								



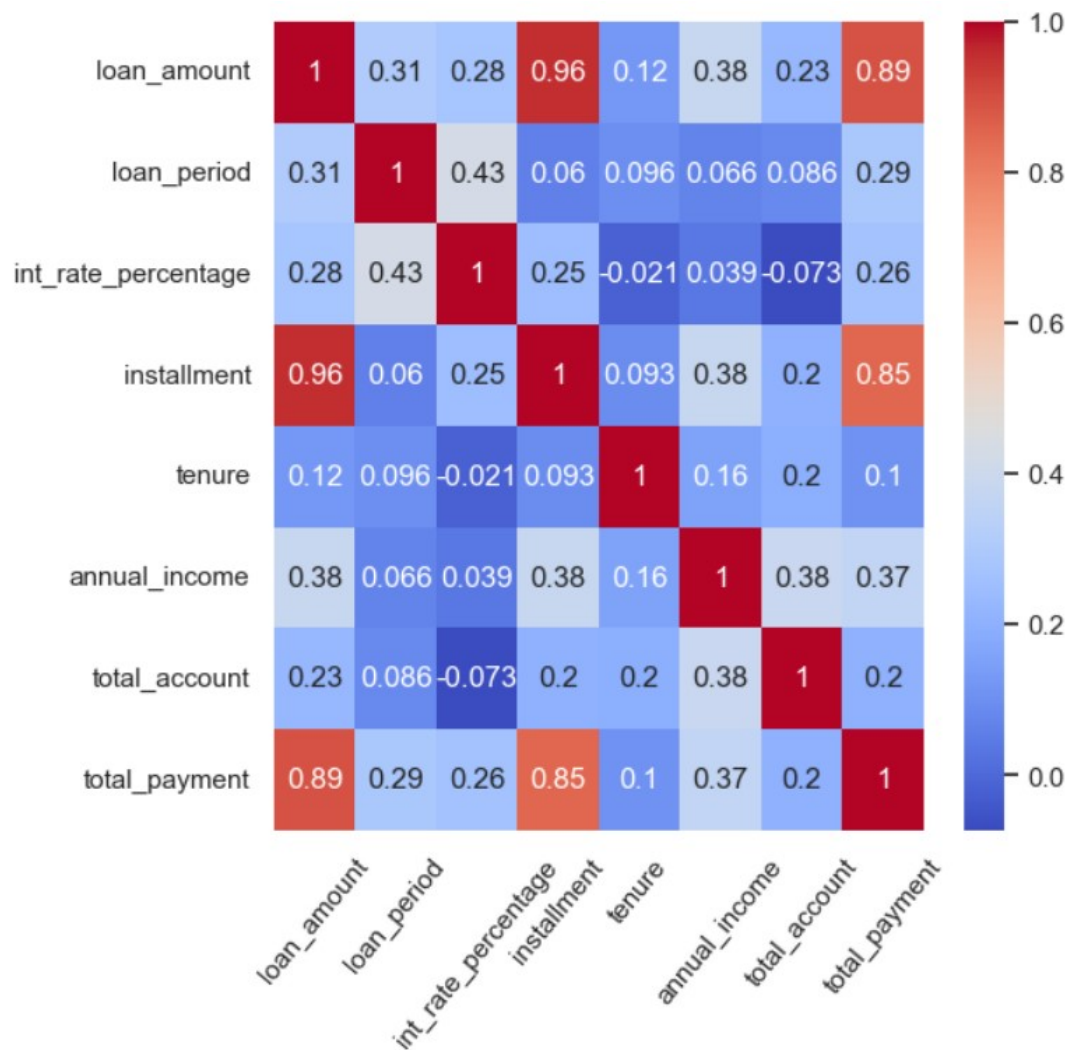


## Summary of Bivariate Analysis:

### Following Insights from each variable with respect to loan charged-off status:

- Borrowers have taken loans to clean their existing debt.
- Borrowers have home ownership as rent.
- Borrowers belong to California state.
- Borrowers were given loans without verification.
- Borrowers have been loan-graded as B.
- Borrowers' tenure in the company is 10 years or more.
- Borrowers who receive interest at the rate of 10-15%.
- Borrowers who have total accounts in the range of 1 - 20.
- Borrowers whose income falls in the range of 25k to 50k.
- Borrowers taking a loan for a tenure of 36 months.
- Maximum borrowers who defaulted belong to loans where the issue year is 2011, and issue months mostly fall under the latter half of the year, peaking in December.

## Multivariate Analysis - Correlation matrix



## Recommendation

- Non verified borrowers shall be verified first before giving a loan
- Loan shall be granted for borrowers whose annual income is greater than 75k
- Loan application to pay the existing debt is high risk
- Loan Interest slab of 10 to 20% is of high disk of defaulters
- High risk borrowers whose tenure is 10 years and greater
- California state applicant are higher probability of defaulter.
- Rented and Mortgage home owners have higher chances to become defaulters.