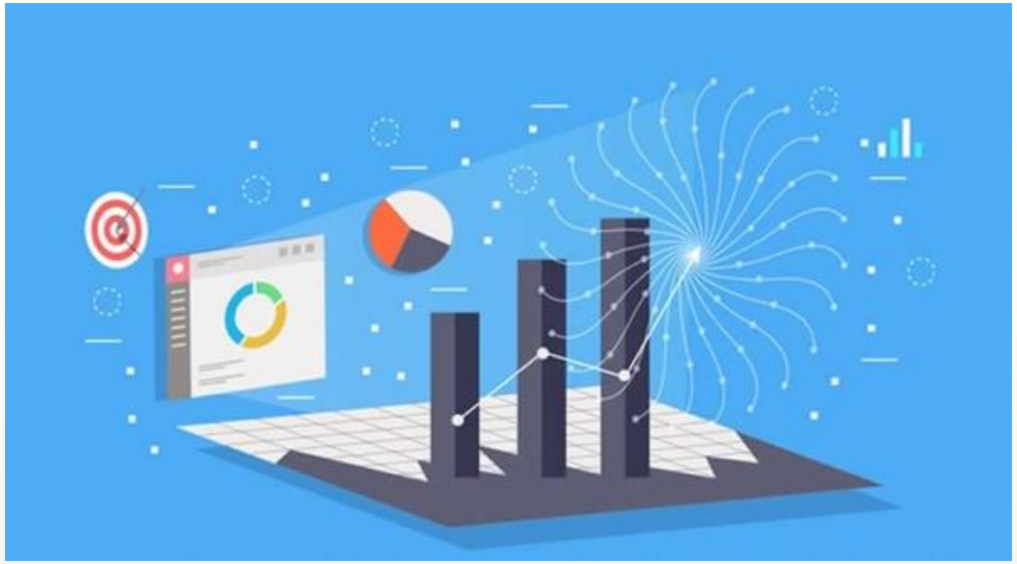


EDA Lending Club Case Study

Presented by – Gopasana Surya Kanth and Swapnil Rodge

Batch – UpGrad C71

Index

- Problem Statement
- Data Understanding
- Univariate Analysis
- Quantitative Variable Analysis
- Unordered Categorical Variable Analysis
- Ordered Categorical Variable Analysis
- Univariate Segmented Analysis
- Bivariate Analysis
- Correlation
- Conclusions/Recommendations

Problem Statement

Problem

You work for a consumer finance company which specializes 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

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'.

The core objective of the exercise is to help the company minimize the credit loss.

Objective

The goal is 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 using the given [dataset](./data/loan.csv), 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 utilize this knowledge for its portfolio and risk assessment.

Data Understanding

- loan.csv contains the complete loan data for all loans issued through the time period 2007 t0 2011
- It contains 39717 rows and 111 columns
- There are 2 types of attributes: Loan Attributes and Customer Attributes
- **Leading Attribute**
  Loan Status - Key Leading Attribute (loan_status). The column has three distinct values

  - Fully-Paid - The customer has successfully paid the loan
  - Charged-Off - The customer is "Charged-Off" ir has "Defaulted“
  - Current - These customers, the loan is currently in progress and cannot contribute to conclusive evidence if the customer will default of pay in future

For the given case study, "Current" status rows will be ignored

Data Cleaning

- No header, footers, summary or Total rows found.
- There were 1140 rows present of loan_status='current' which has been deleted as loan_status ='current' does not participate in analysis.
- No duplicates rows found.
- There were 55 columns which is having all the rows values as null/blank and doesn't participate in analyze has been removed.
- 'url' and 'member_id' is unique in nature and has been deleted. Have kept 'id' for future purpose analyse.
- 'desc' and 'title' text/description values and doesn't participate has been dropped from analysis.
- Limiting our analysis till 'Group' level only hence sub group has been dropped.
- Using domain knowledge, behavioral data is captured and hence will not available during the loan approval and doesn't participate in analysis. 21 behavioral data columns has deleted.
- 8 columns whose values were 1, and is uniqueness in nature has been dropped from analysis.
- There were two columns which is having more that 50% of data as na has been removed.

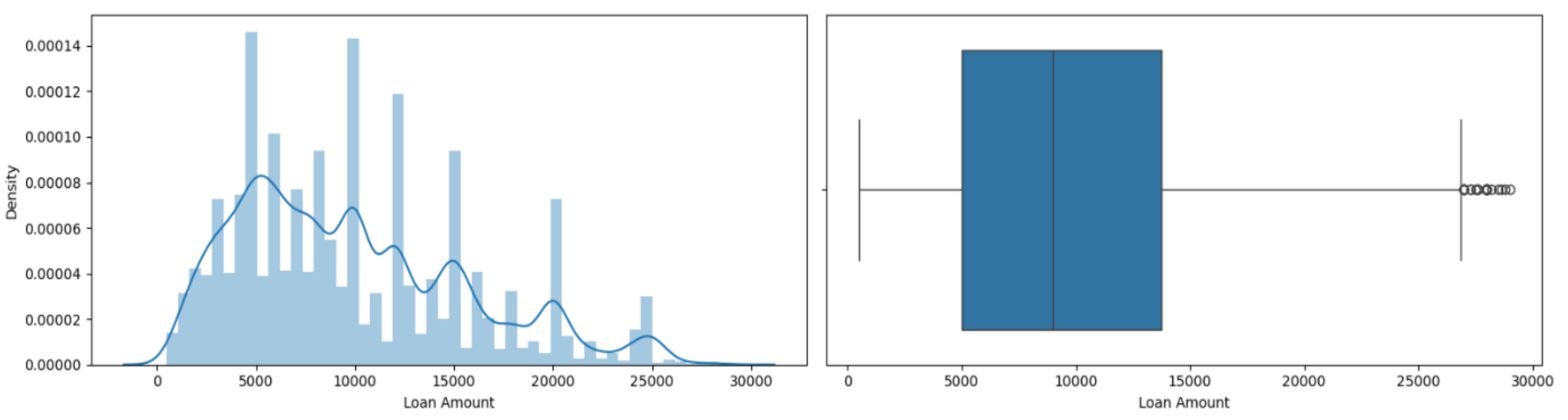
Data Conversions vs Derived Columns

- Additional string value has been trimmed from ‘term’ column and has been converted to int data types.
- ‘int_rate’ has been converted from string to int. Additional ‘%’ has been trimmed.
- Column ‘loan_funded_amnt’ and ‘funded_amnt’ converted to float.
- loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', ‘dti' columns valued rounded off to two decimal points.
- issue_d has been converted to datatype.
- Creating a derived columns for ‘issue_year’ and ‘issue_month ’ from ‘issue_d’ which will be using for further analysis.
- ‘loan_amnt_b’, ‘annual_inc_b’, ‘int_rate_b, and ‘dti_b’ derived columns(multiple bucket kind of data from continuous data) has been created for better analysis.

Univariate Analysis

Quantitative Variable Analysis

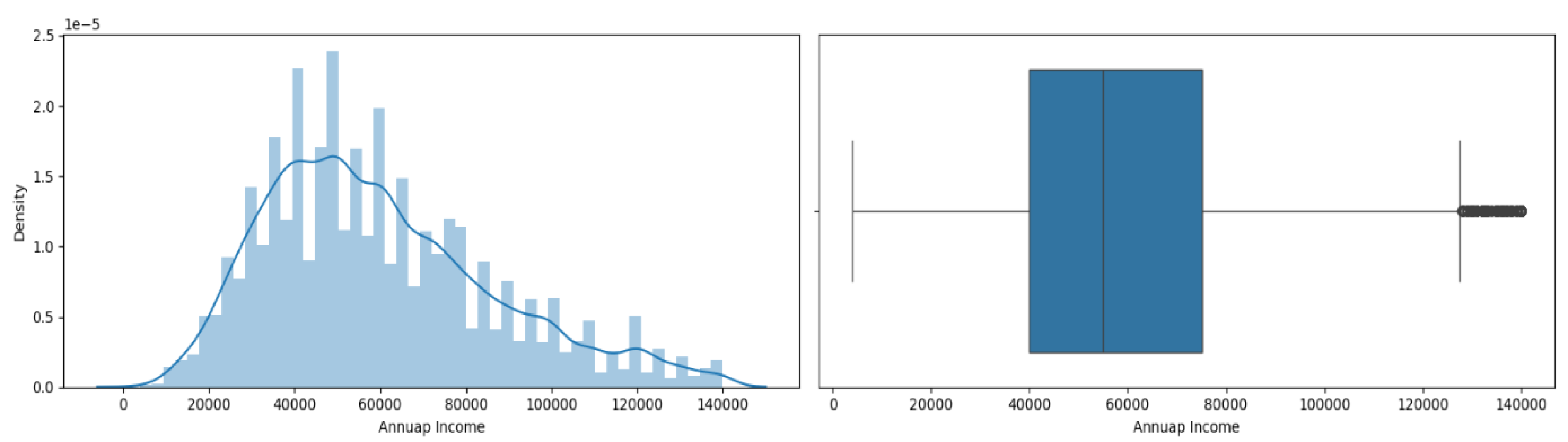
Loan Amount



Observation:

# Most of the loan amount applied was in the range of 5k-14k.
# Max Loan amount applied was ~27k

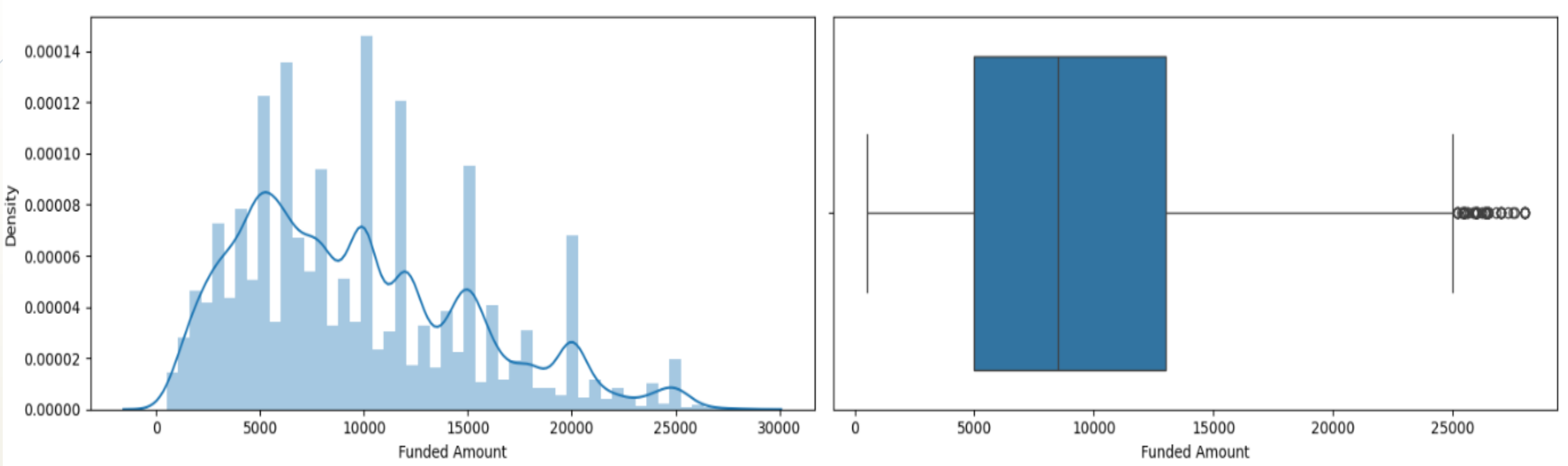
- Annual Income



Observation:

# The Annual income of most of applicants lies between 40k-75k

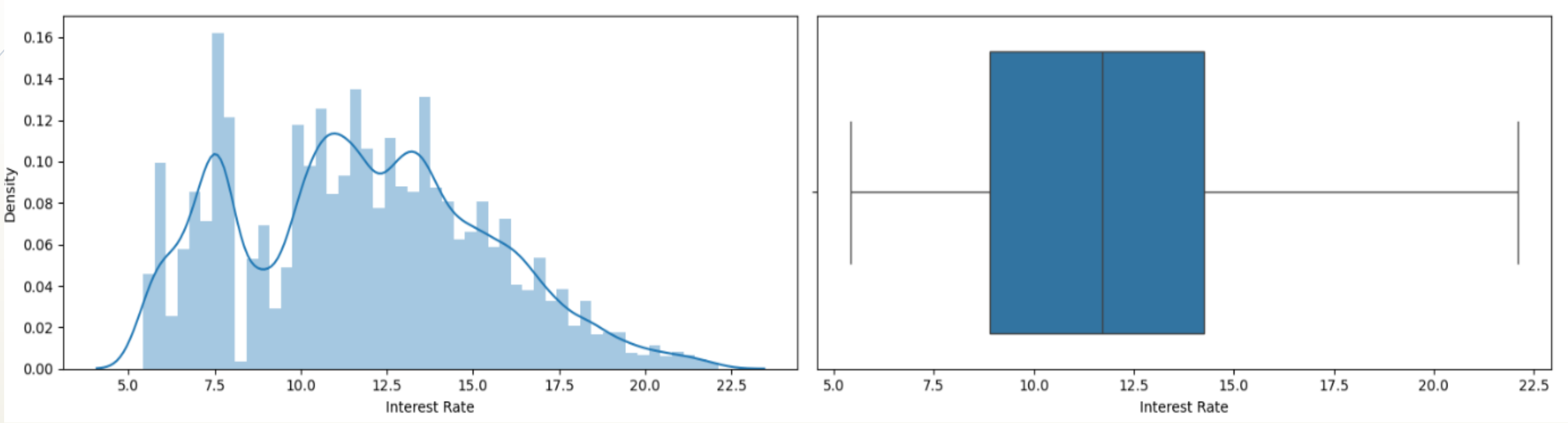
Funded Amount



Observation:

# Majority of the funded_amnt is in the range of 5K to 13K

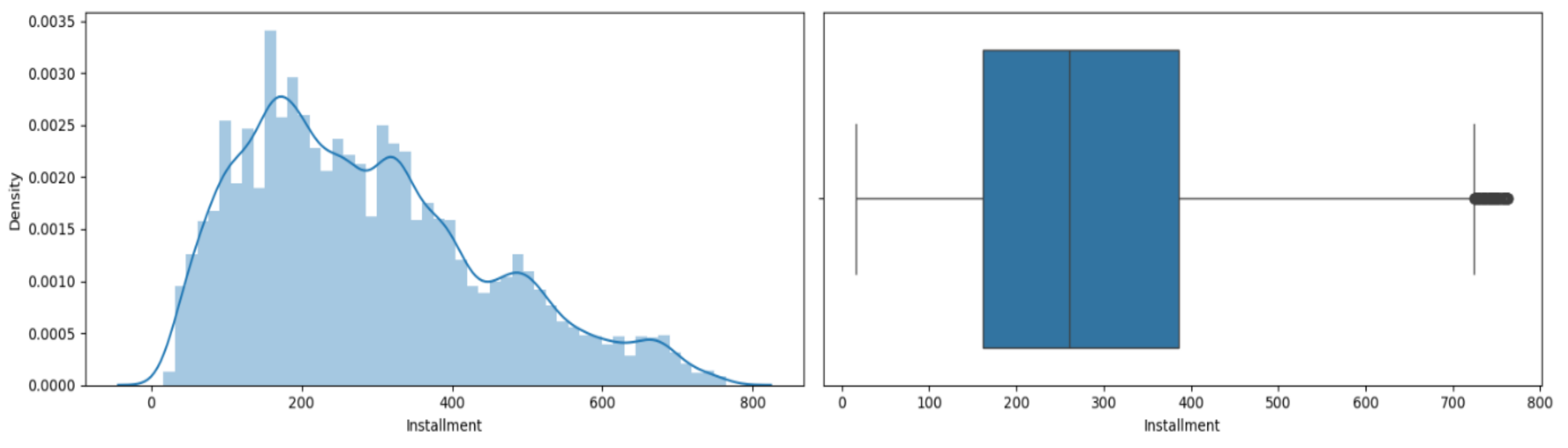
Interest Rate



Observation:

# Most of the applicant's rate of interest is between in the range of 8%-14%
# Average Rate of interest of rate is 11.7 %

Installment

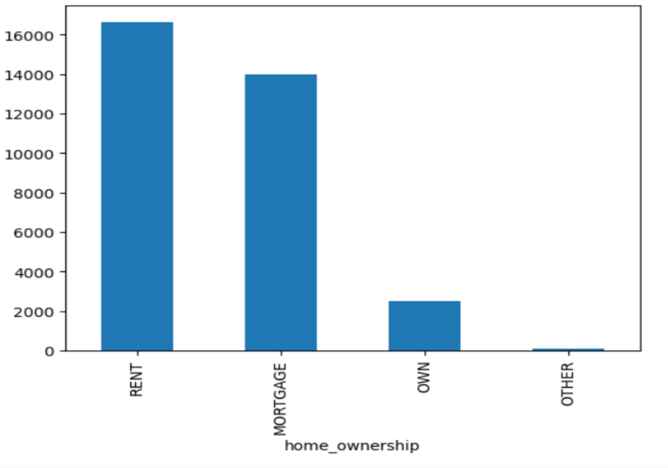


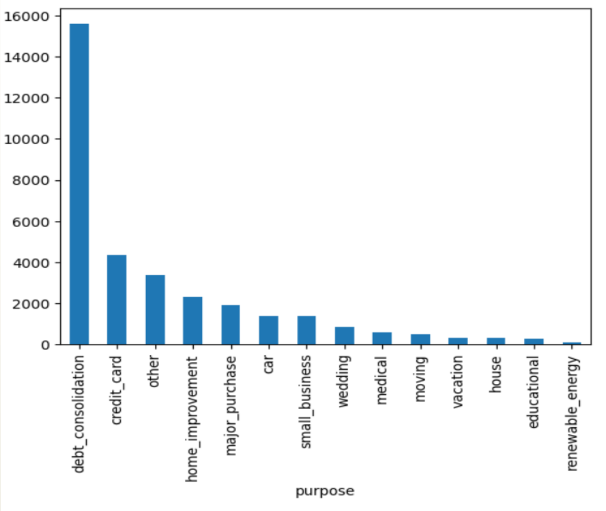
Observation:

#Majority of the installment is in the range of 20 to 400 going at the max to 763

Univariate Analysis

Unordered Categorical Variable Analysis



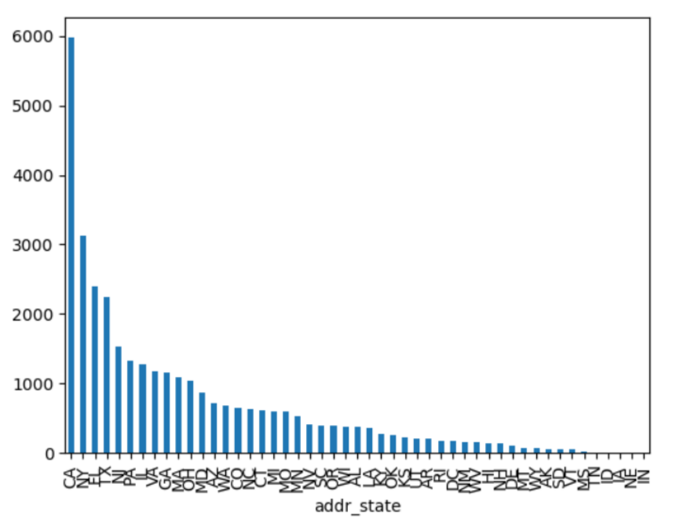


Observation:

#Majority of the home owner status are in status of RENT and MORTGAGE

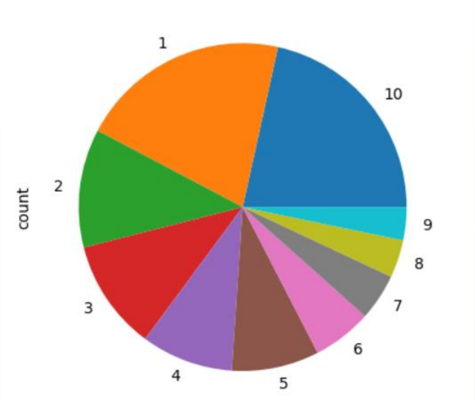
# Majority of loan application are in the category of debt_consolidation

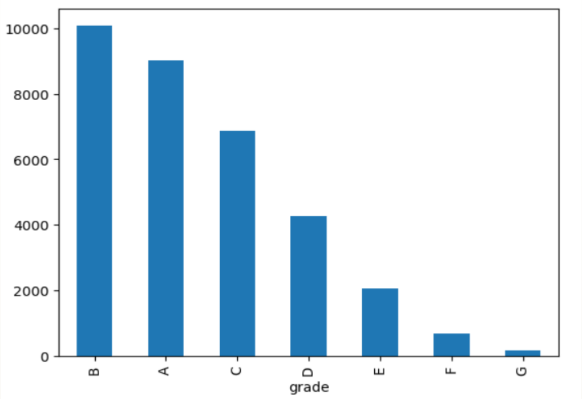
# CA state has the maximum amount of loan applications

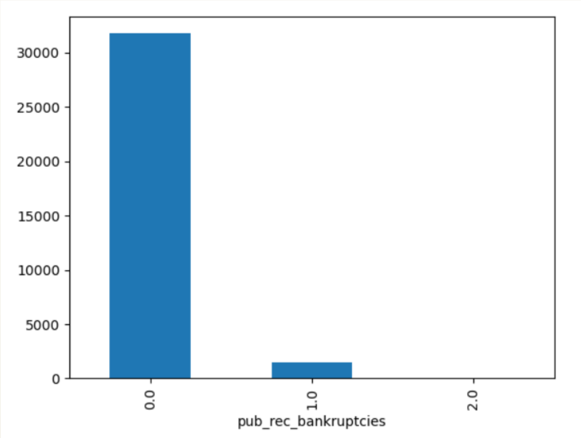


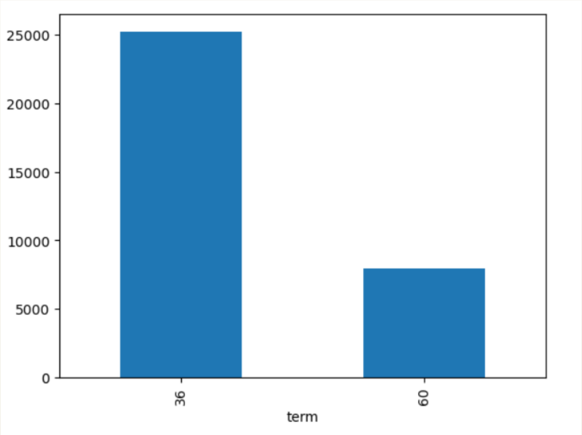
Univariate Analysis

Ordered Categorical Variable Analysis









Observation:

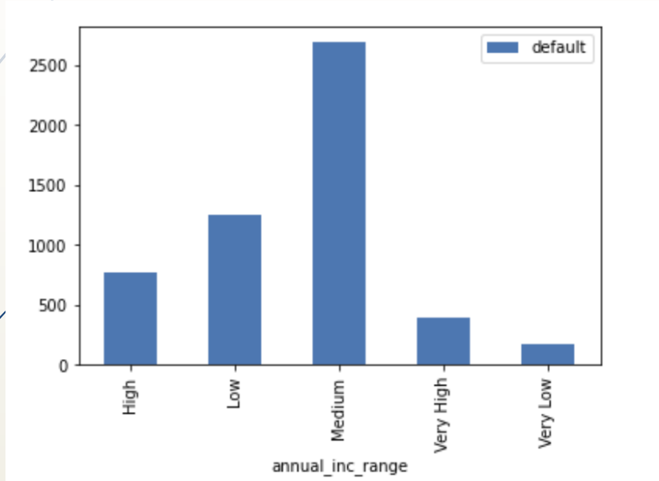
#Majority of the employment length of the customers are 10+ years and then in the range of 0-2 years# Majority of loan application are in the category of debt_consolidation

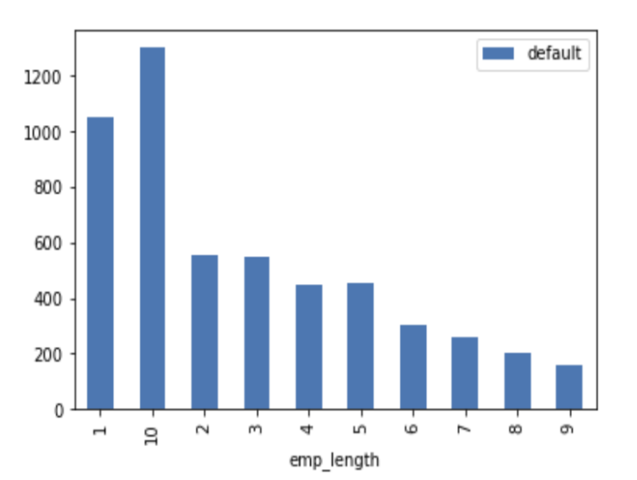
# Majority of loan application counts fall under the catogory of **Grade B**

# Majority of the loan applications counts are in the term of 36 months

# Majority of the loan applicants are in the category of not having an public record of bankruptcies

Univariate Segmented Analysis



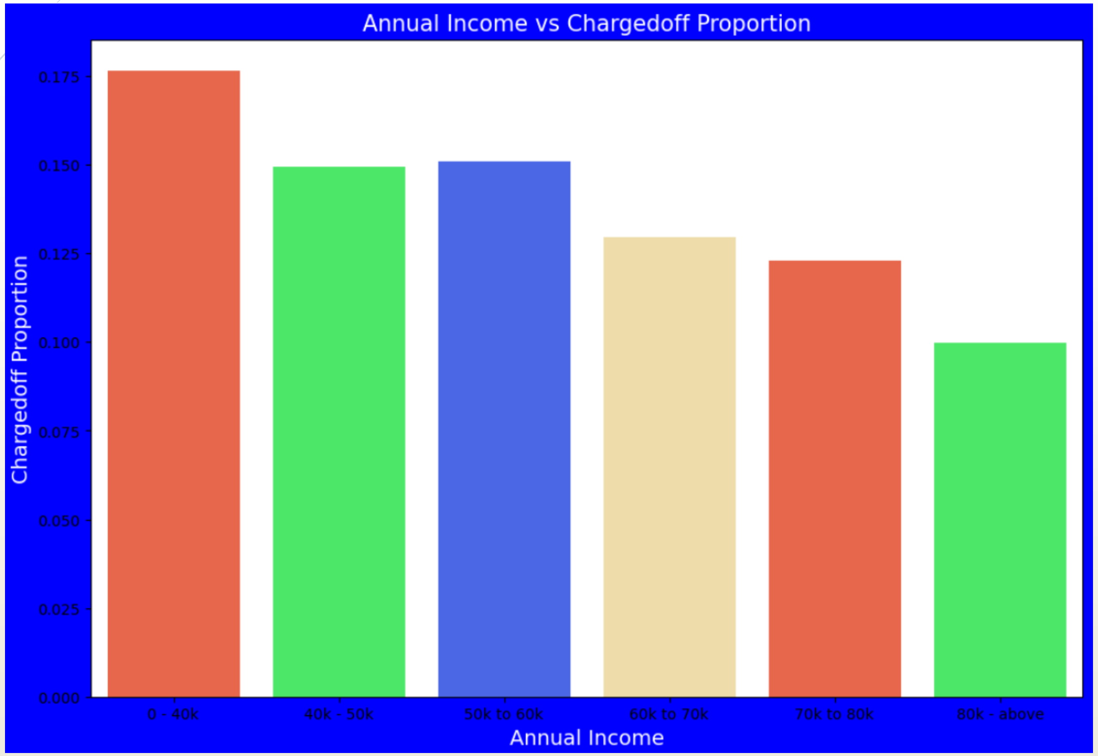


Observations :

- Annual income range segment Medium (80k-1.2L) has highest number of defaulters
- Employees having experience of >10 and 1 are highest defaulters

Bivariate Analysis

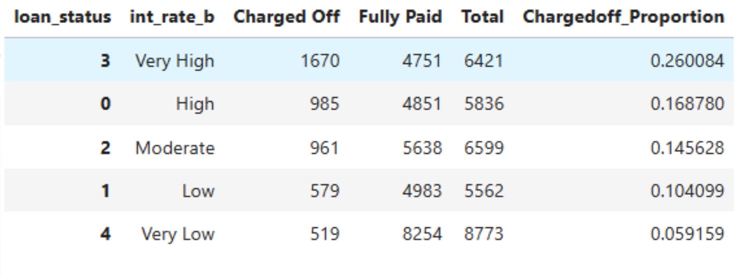
- Annual income vs Charged Off



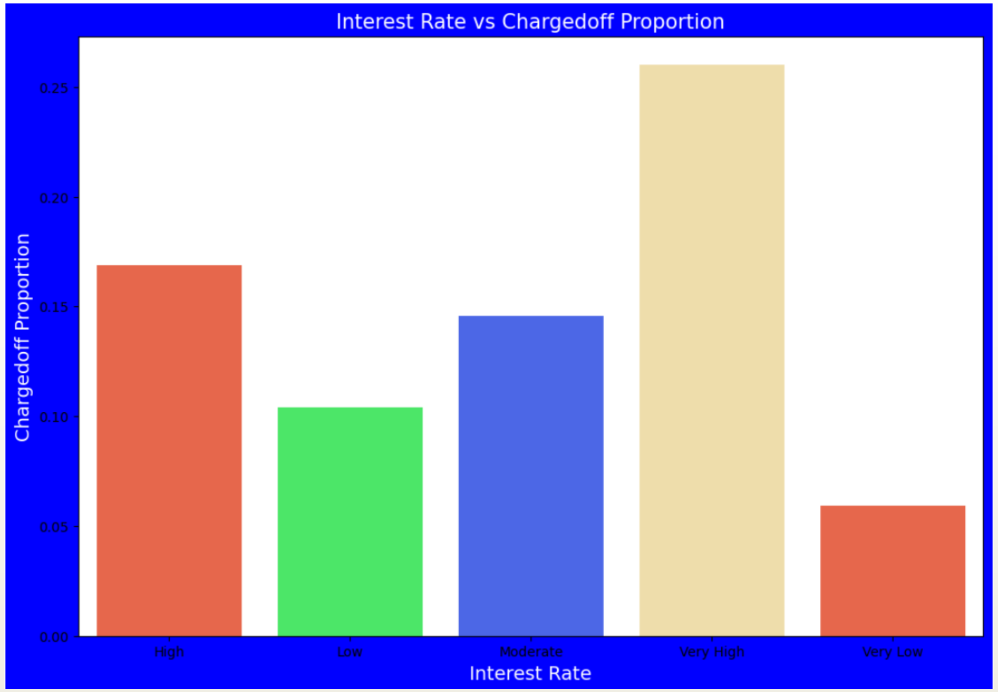
Observation:

# Income range 80000+ has less chances of charged off
# Income range 0-40000 has high chances of charged off
# Notice that with increase in annual income charged off proportion got decreased.

Intrest rate vs Charged Off



loan_status	int_rate_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1670	4751	6421	0.260084
0	High	985	4851	5836	0.168780
2	Moderate	961	5638	6599	0.145628
1	Low	579	4983	5562	0.104099
4	Very Low	519	8254	8773	0.059159


Interest Rate vs Chargedoff Proportion

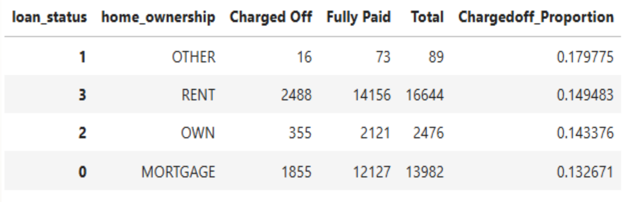
Observation:

# interest rate less than 10% or very low has very less chances of charged off. Intrest rates are starting from min 5 %.

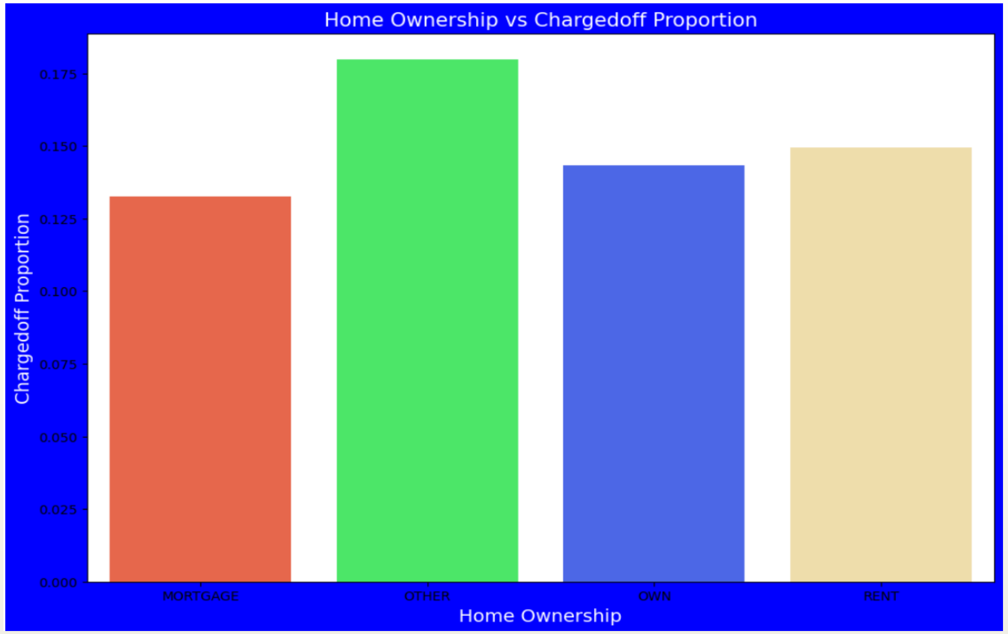
# interest rate more than 16% or very high has good chances of charged off as compared to other category intrest rates.

# Charged off proportion is increasing with higher intrest rates.

Home Ownership vs Charged Off



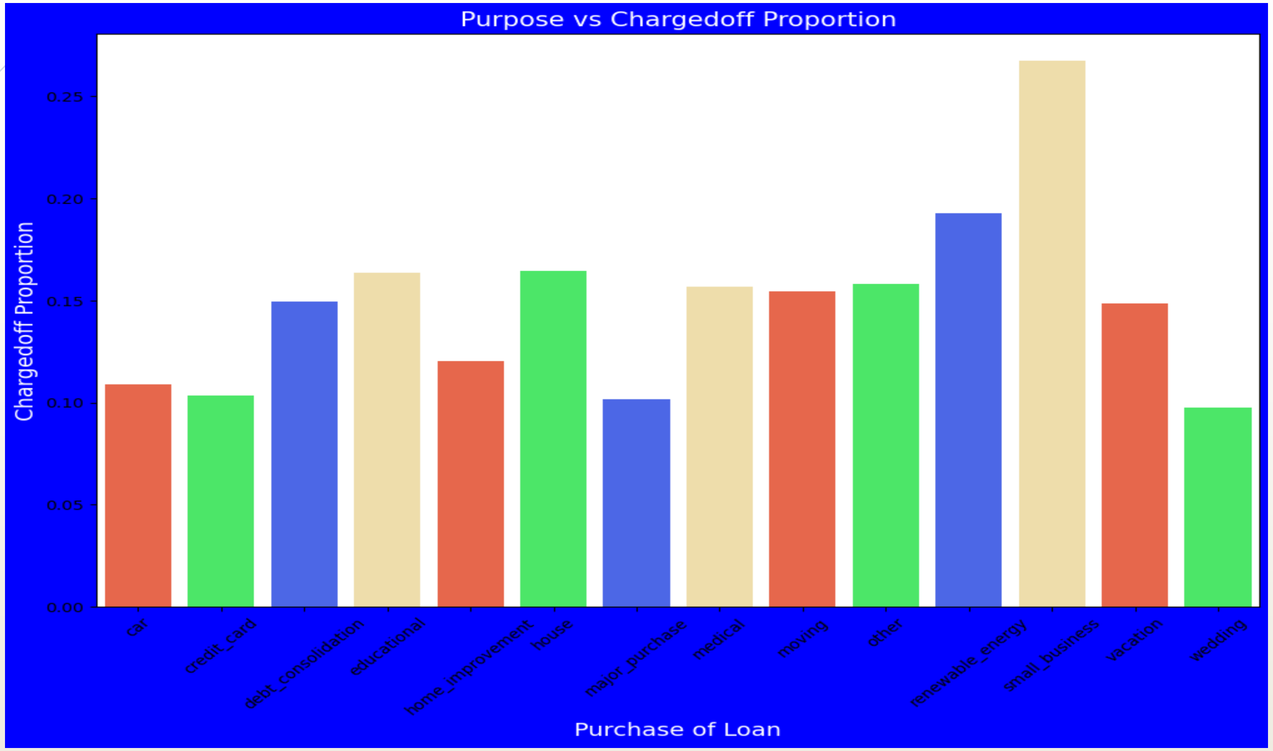
loan_status	home_ownership	Charged Off	Fully Paid	Total	Chargedoff_Proportion
1	OTHER	16	73	89	0.179775
3	RENT	2488	14156	16644	0.149483
2	OWN	355	2121	2476	0.143376
0	MORTGAGE	1855	12127	13982	0.132671



Observation:

# Those who are not owning the home is having high chances of loan defaults.

Purpose vs Charged Off

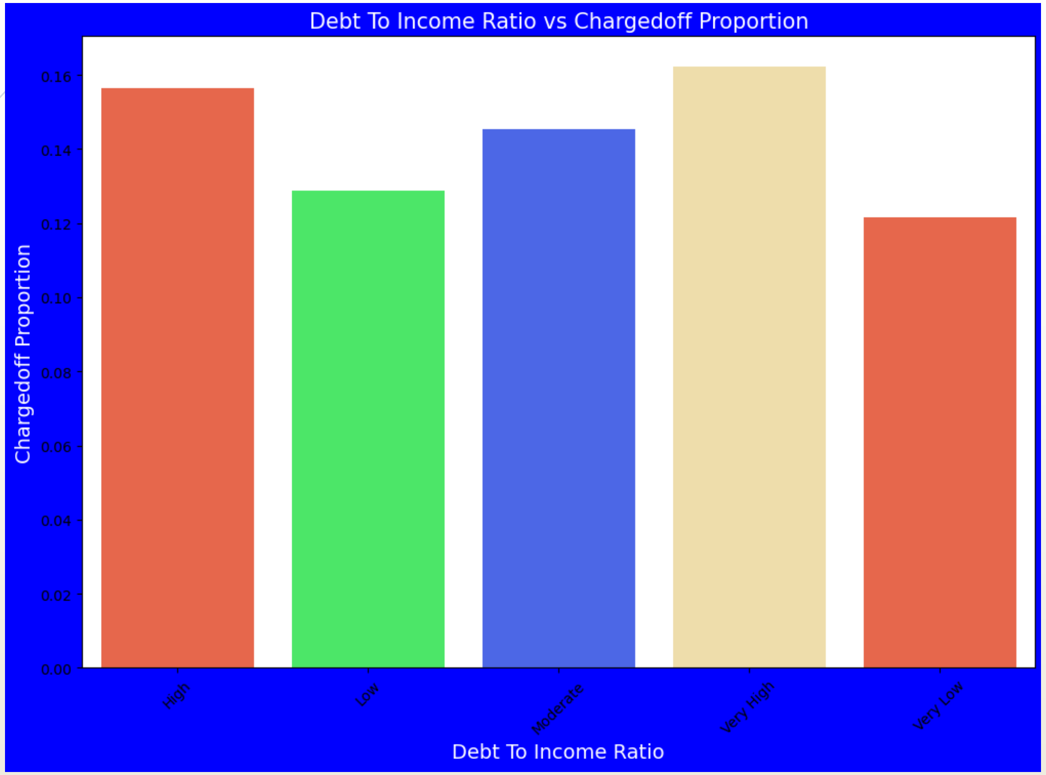


Observation:

# Those applicants who is having home loan or car loan is having low chances of loan defaults.

# Those applicants having loan for small business is having high chances for loan defaults.

- Debt to Income (Dti) vs Charged Off

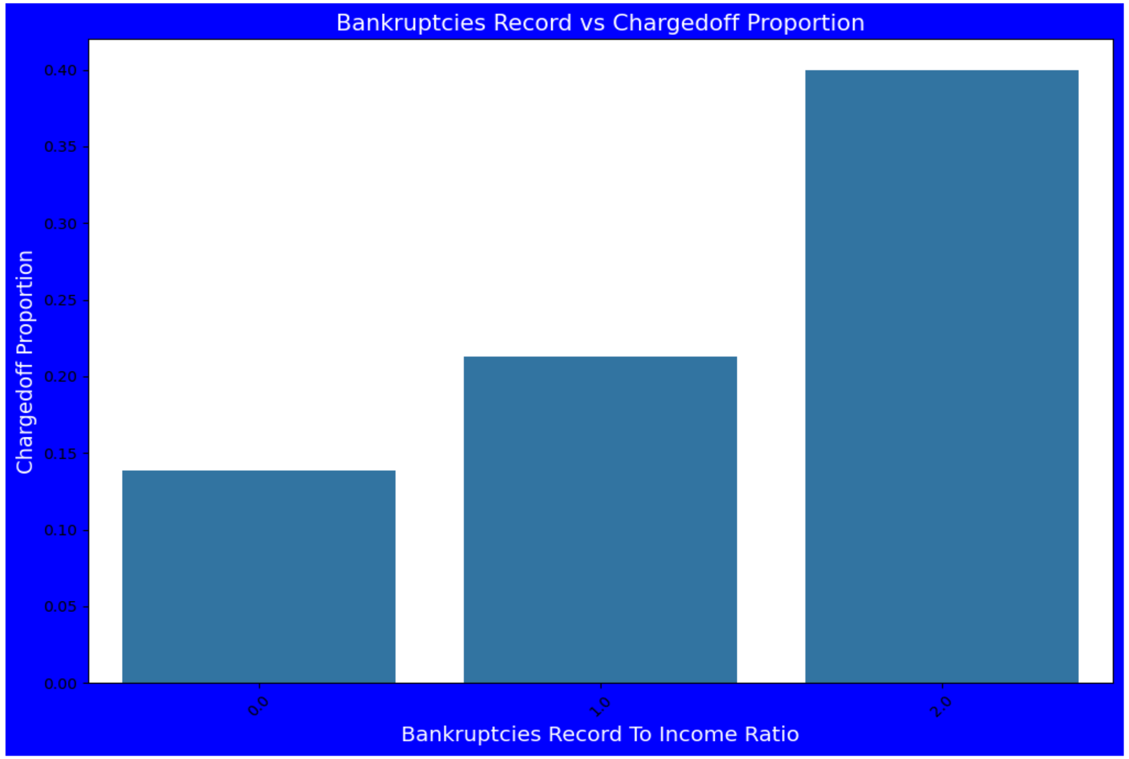


Observation:

# High DTI value is having high risk of defaults

# Lower the DTI having low chances loan defaults.

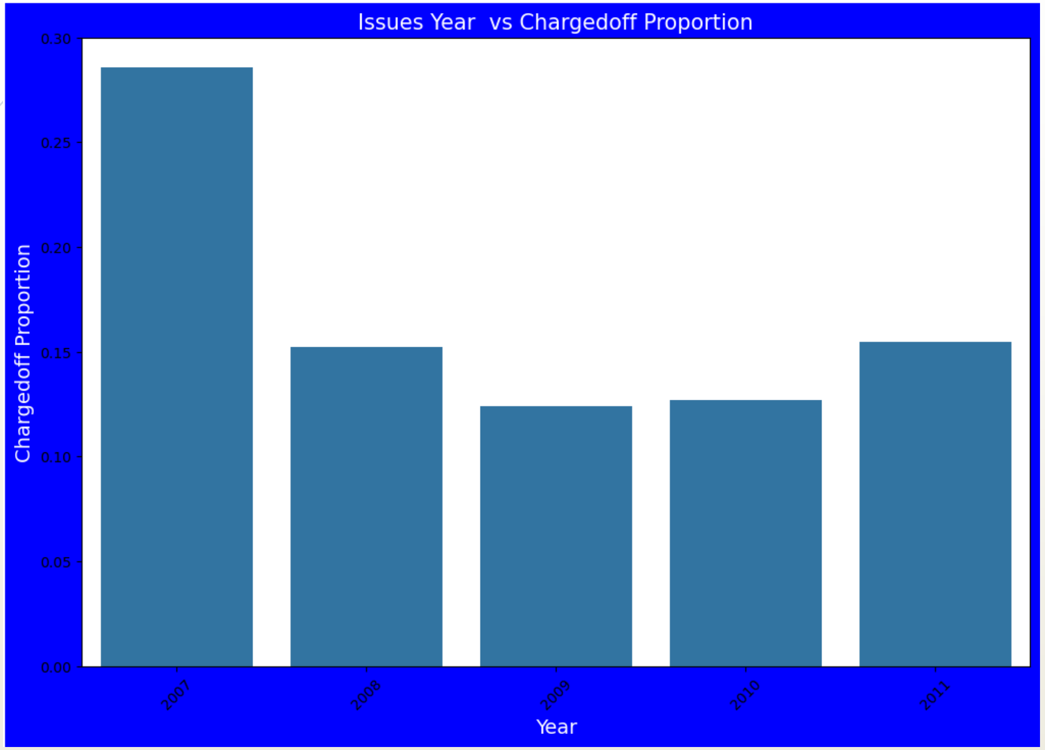
## Bankruptcies Record vs Charged Off



Observation:

# Bankruptcies Record with 2 is having high impact on loan defaults
# Bankruptcies Record with 0 is low impact on loan defaults
# Lower the Bankruptcies lower the risk

- Issue Year vs Charged Off

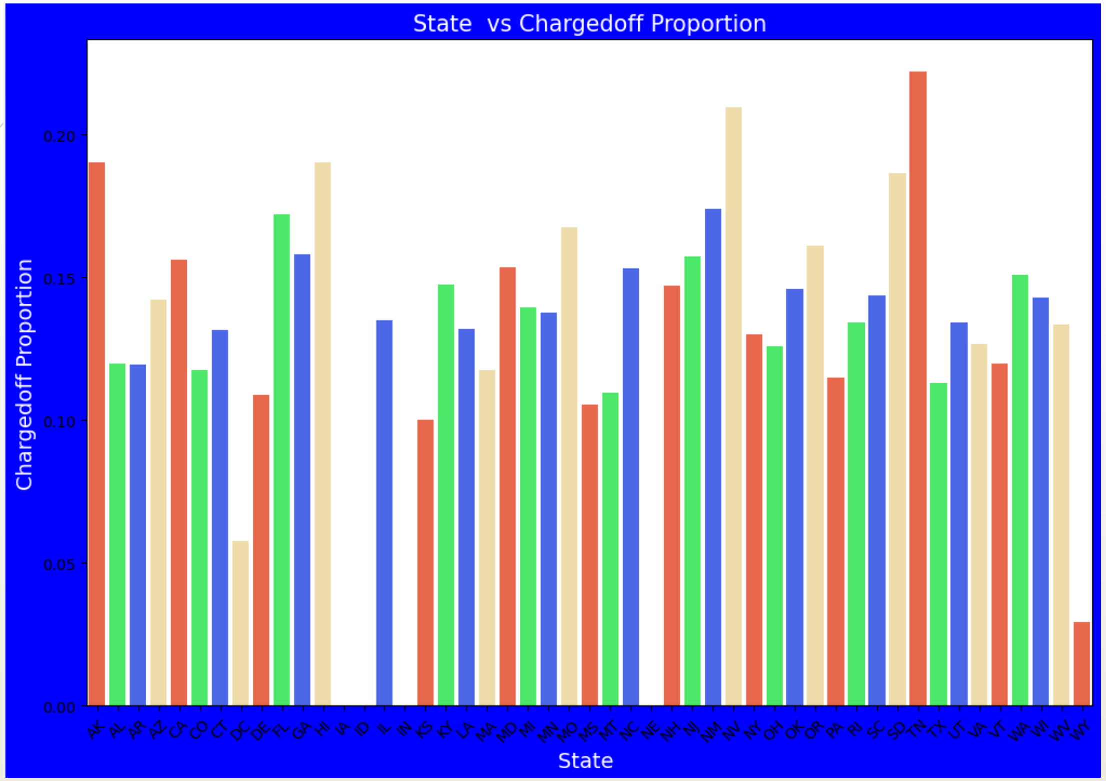


Observation:

# Year 2007 is highest loan defaults

# 2009 is having lowest loan defaults.

State vs Charged Off

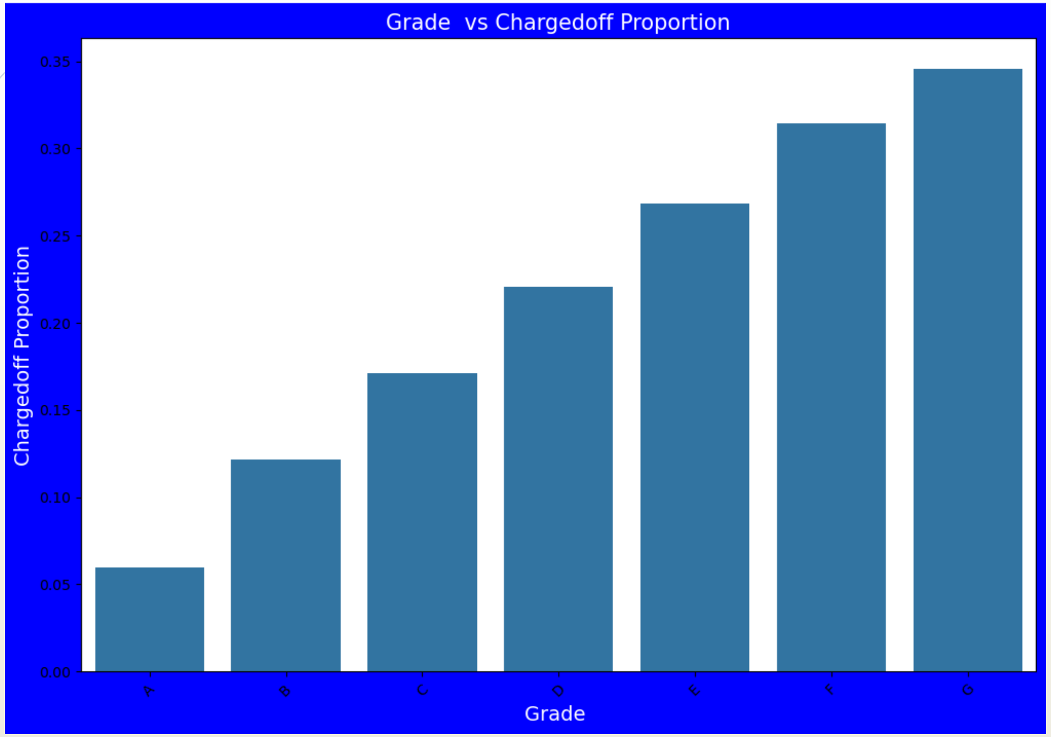


Observation:

# TN States is holding highest number of loan defaults

# WY is having low number of loan defaults

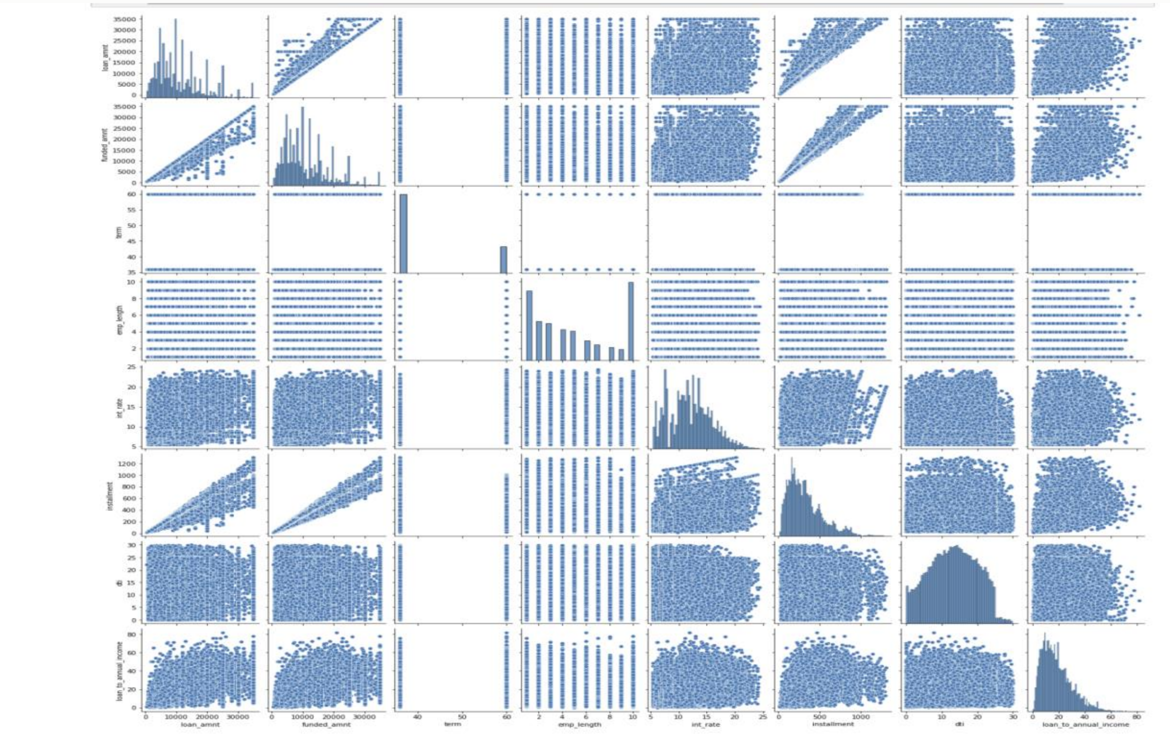
- Grade vs Charged Off



Observation:

# The Loan applicants with loan Grade G is having highest Loan Defaults
# The Loan applicants with loan A is having lowest Loan Defaults

Correlation



Observations :

- Funded amount increased with Loan amount in almost linear fashion.
- Employees of different experience years have no strong relation with the loan amount, interest rate and term

Conclusions/Recommendations

- Most of the loan amount applied was in the range of 5k-14k
  Most of the applicant's rate of interest is between in the range of 8%-14%
- CA state has the maximum amount of loan applications
- Income range 0-40000 has high chances of charged off
- interest rate more than 16% or very high has good chances of charged off as compared to other category intrest rates. Charged off proportion is increasing with higher intrest rates.
- Those who are not owning the home is having high chances of loan defaults.
- Those applicants who is having home loan or car loan is having low chances of loan defaults. Those applicants having loan for small business is having high chances for loan defaults.
- High DTI value is having high risk of defaults
- Higher the Bankruptcies record higher the chance of loan defaults
- TN States is holding highest number of loan defaults
- The Loan applicants with loan Grade G is having highest Loan Defaults. The Loan applicants with loan A is having lowest Loan Defaults
- Year 2007 is highest loan defaults. 2009 is having lowest loan defaults
- Purpose : Debt Consolidation has high chances of loan defaults
- Employee experience : 10 years has high chances of loan default
- Annual income range : Medium(40K-80K) segment has high chances of loan default
- Home ownership and Verification status : Rent (Verified and Not Verified) and Mortgage (Verified) has high chances of loan default