Lending Club Case Study

Exploratory Data Analysis

Group Members:
Dhananjay &
Krithigha

- Problem Statement
- Data Summary
- Data Cleaning
- Data conversions vs Derived Columns
- Dropping/Imputing the Rows
- Outliers
- Univariate Analysis
- BivariateAnalysis
- Correlations
- Conclusions

Problem Statement

Problem:

- 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.

Objective:

  - Use EDA to understand how consumer attributes and loan attributes influence the tendency of default

Constraints:

- When a person applies for a loan, there are two types of decisions that could be taken by the company:
  - **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:
    - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    - **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'.
    - **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
  - **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 notavailable with the company (and thus in this dataset)

Data Summary

- Loan.csv file contains 39717 rows and 111 columns.
- There are two types of attributes Loan Attribute and Customer attributes.

Data Cleaning

- There were no header, footers, summary or Total rows found.
- There were no duplicates rows found.
- There were 1140 rows present of loan_status=‘current’ which has been deleted as loan_status =‘current’ doesn’t participate in analysis.
- There were 55 columns which is having all the rows values as null/blank and doesn’t participate in analyse 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, behavioural data is captured and hence will not available during the loan approval and doesn’t participate in analysis. 21 behavioural 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.
- After all the Data cleaning process we are left with 38577 rows and 20 columns.

Data Conversions vs Derived Columns

- Rounding off long float decimal columns to 2 - total_pymnt,total_rec_late_fee,collection_recovery_fee
- Identifying and converting Date time fields - last_pymnt_d , last_credit_pull_d ,earliest_cr_line ,issue_d
- Dropping Duplicates from data
- Converting column - int_rate having % and as object data type to integer data type
- Converting column - revol_util having % and as object data type to integer data type
- Making the Emp length column as a range of years values
- Removing months word from the Term column , Sanity Check of data and changes made

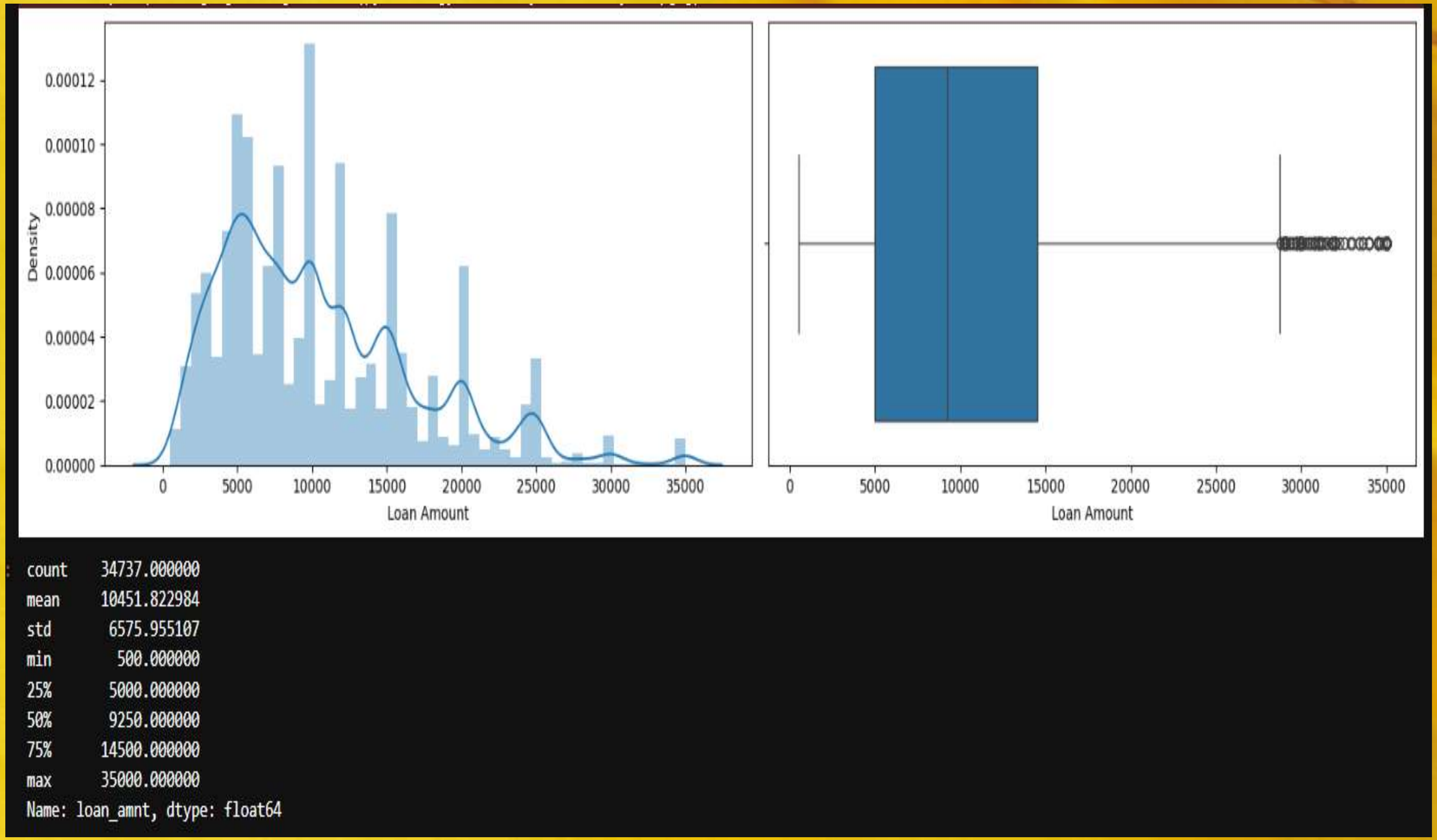
Dropping/Inputing the rows

- ‘emp_lenght’ and pub_rec_bankruptcies contains 2.67% and 1.80% of rows as null, which is very small percetnage of data which we can drop it.
- Total % of rows deleted: 4.48%,
- Outliers exits for numeric data 'loan_amnt', 'funded_amnt', ‘funded_amnt_inv','int_rate', 'installment', ‘annual_inc'.
- Outliers treatment has been done for above fields using quantile mechanism.

Univariate Analysis

Loan Amount

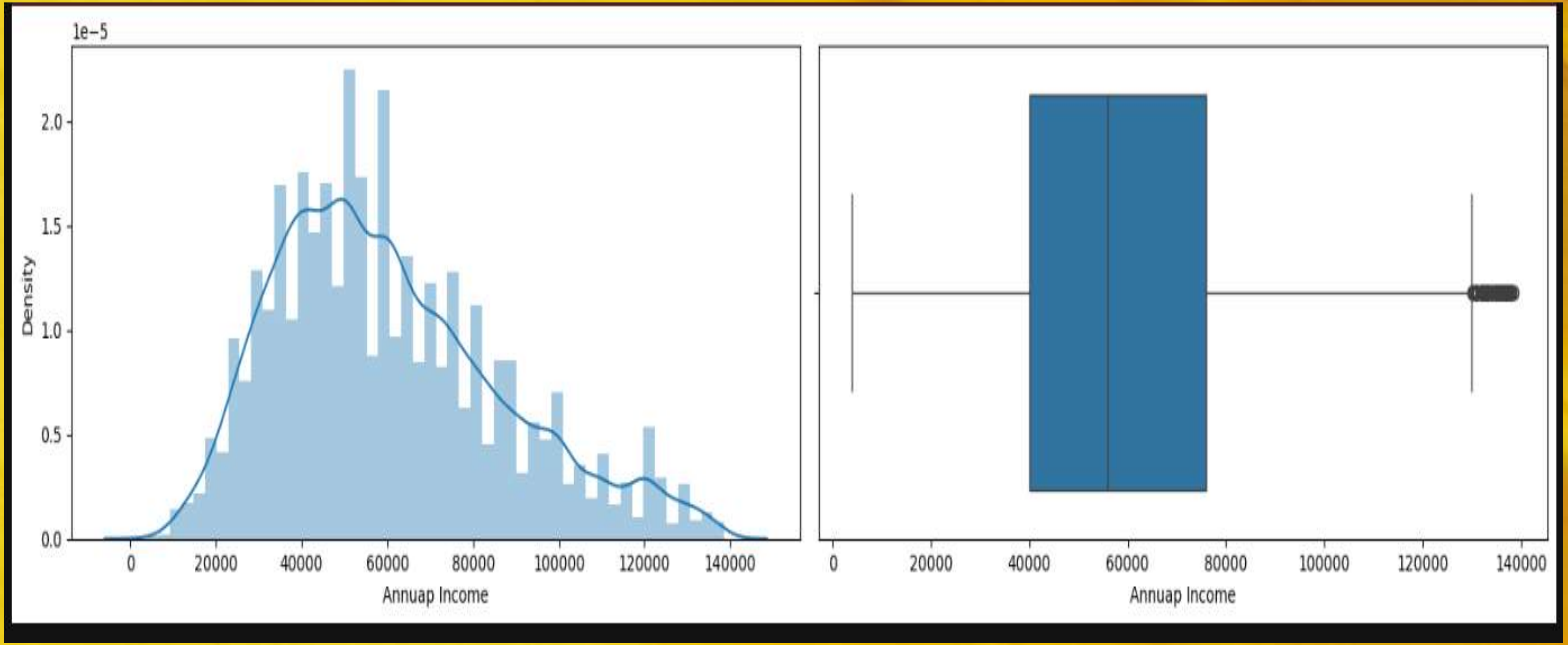
- **Observations:**
  - Most of the loan amount applied was in the range of 5k-14k.
  - Max Loan amount applied was ~27k.



60433.0

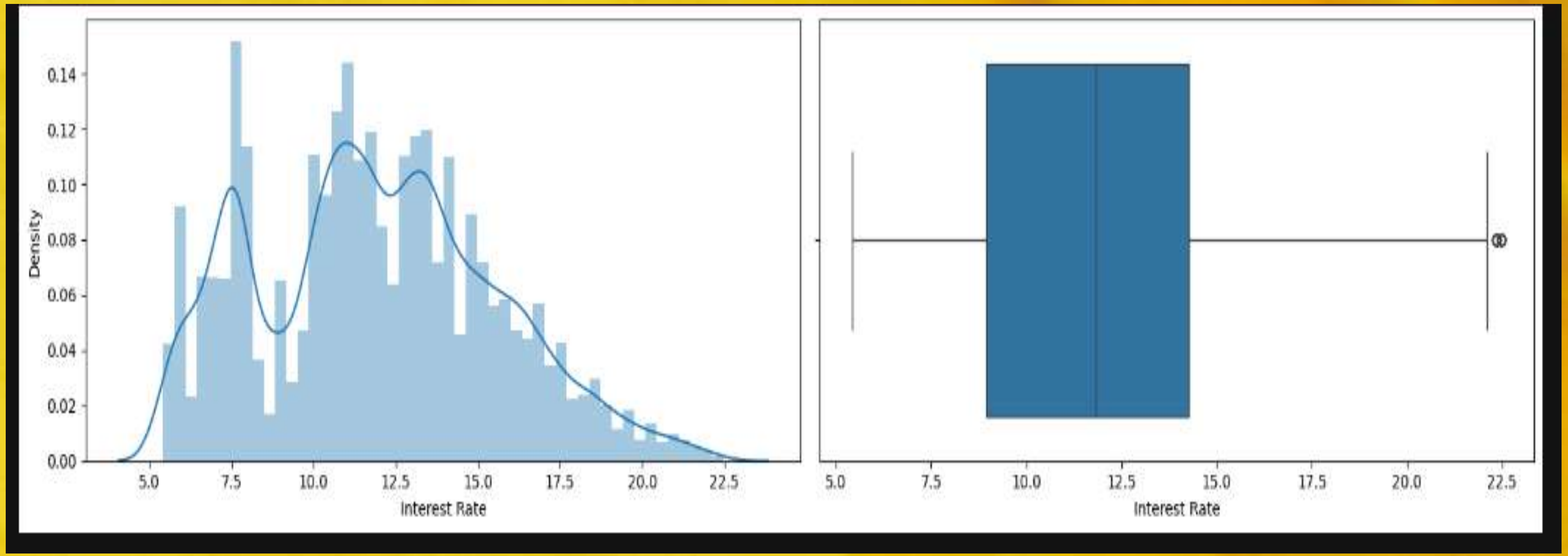
Annual Income

- **Observations:**
  - The Annual income of most if applicants lies between 40k-75k.
  - Average Annual Income is :
  - 60433.0



Interest Rate

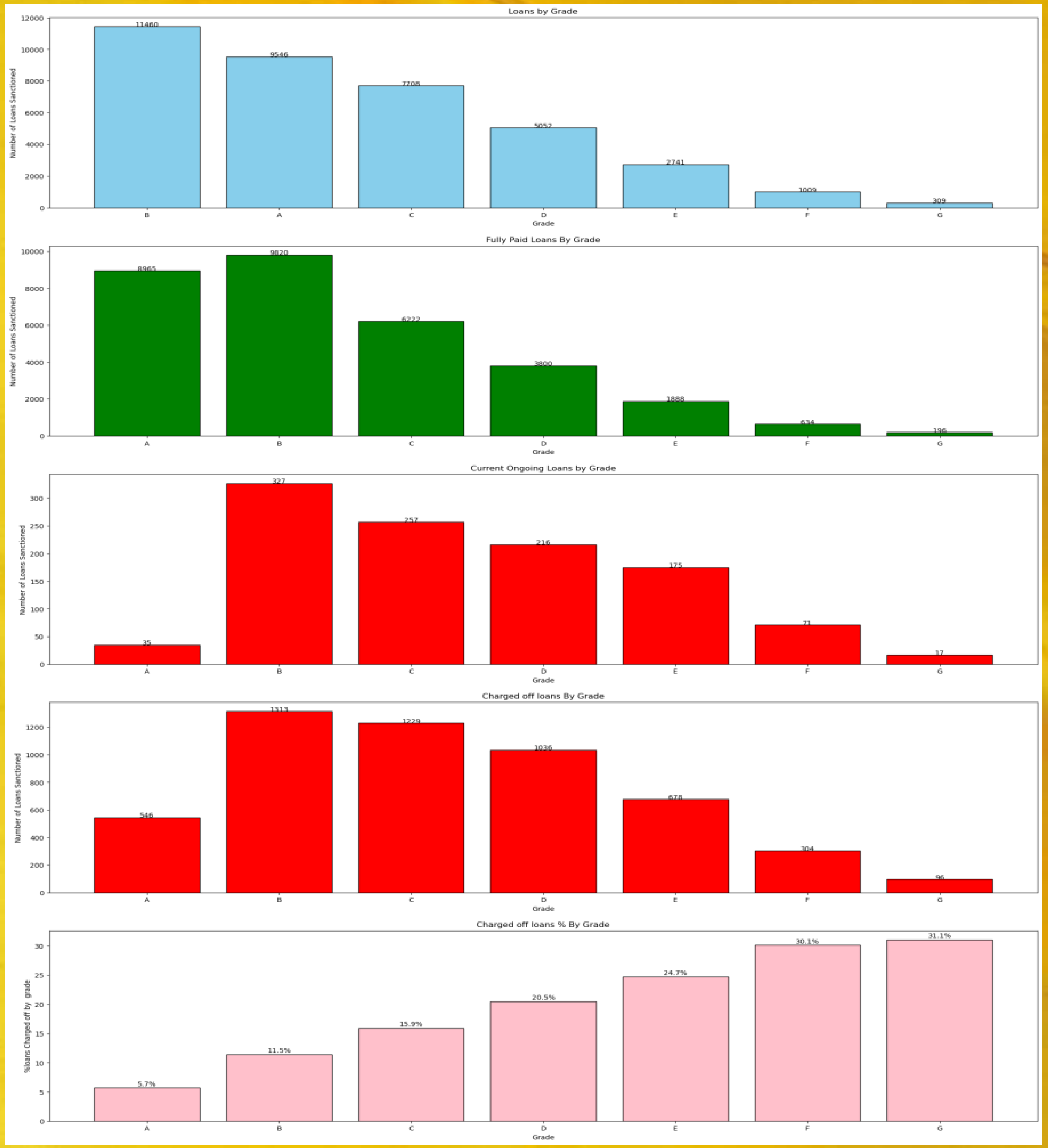
- **Observations:**
  - Most of the applicant's rate of interest is between in the range of 8%-14%.
  - Average Rate of interest of rate is 11.7 %



Analysis on Charged Off, Current and Fully Paid loan distribution across grades

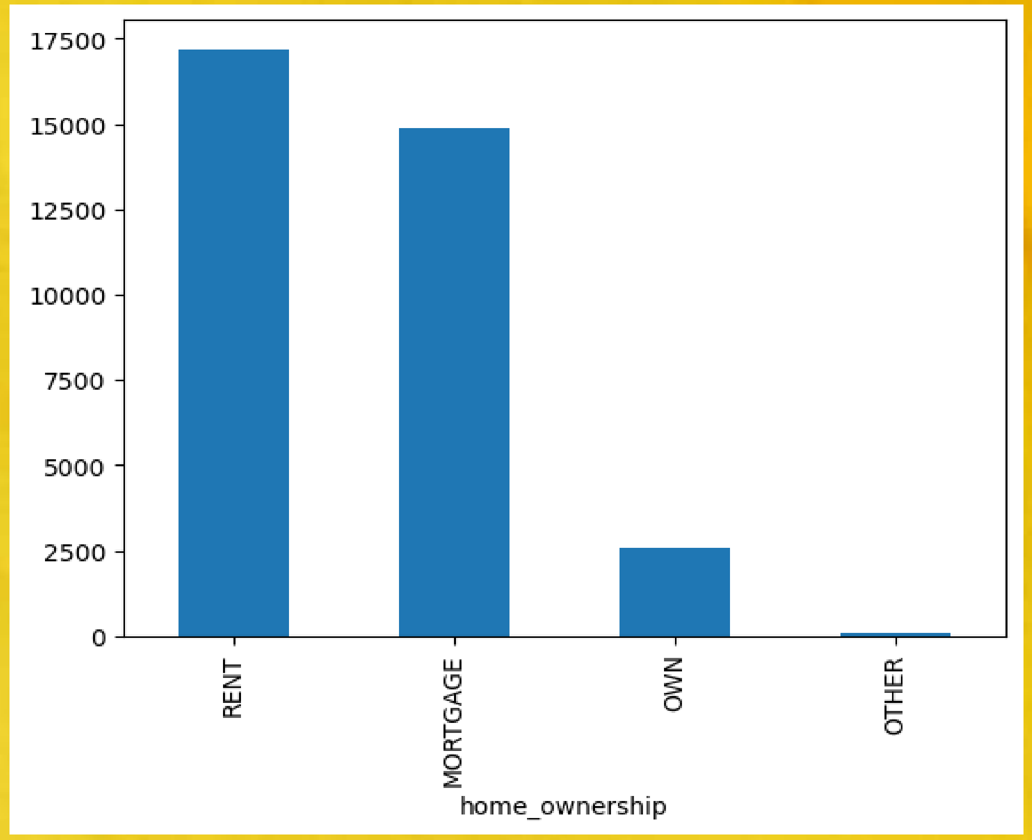
Observation:

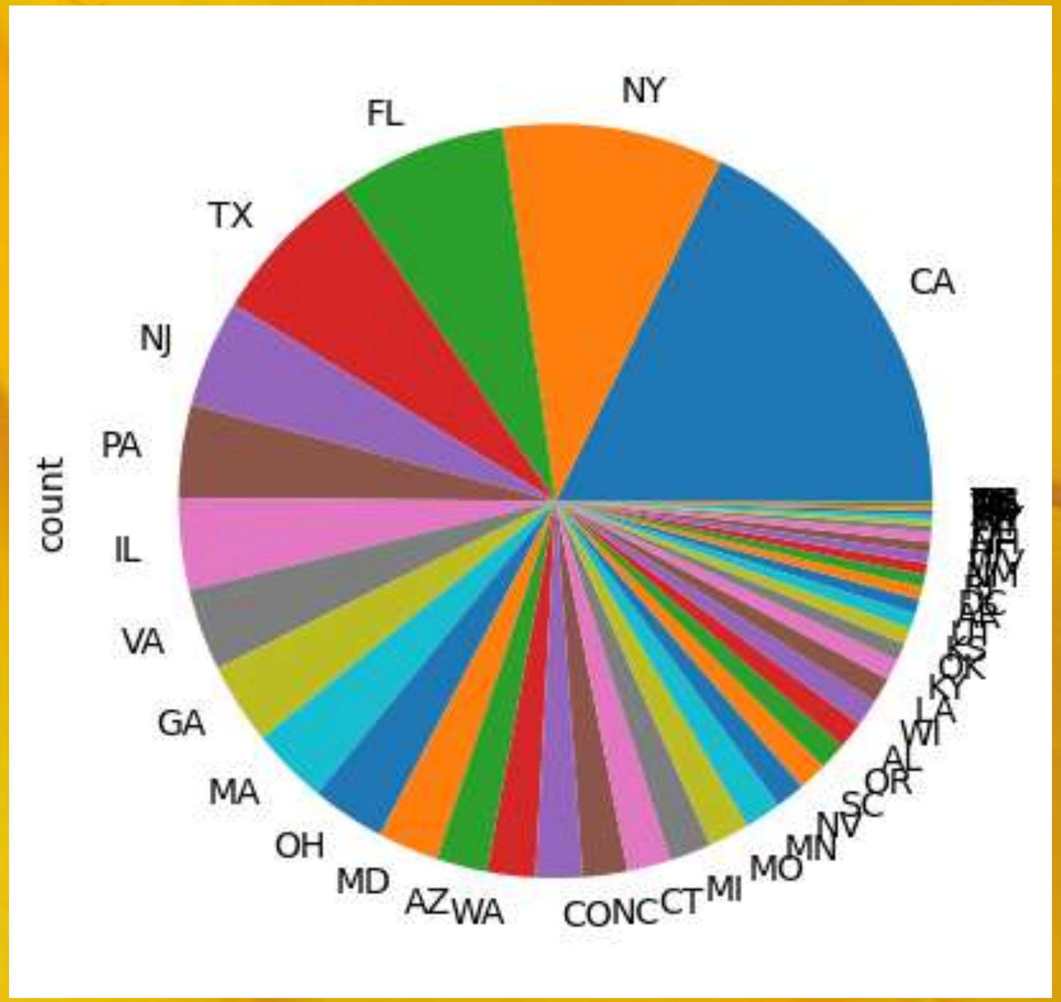
- The analysis reveals that B grade customers have the highest number of loan sanctions.
- The analysis reveals that B grade customers have the highest number of fully paid loans.
- The analysis reveals that B grade customers have the highest number of ongoing loans.
- The analysis reveals that B grade customers have the highest number of loans charged off, with the number being 1313.
- The analysis reveals that G grade customers have the highest percentage of loans charged off, at 31%.
- The analysis reveals that A grade customers have the lowest percentage of loans charged off, at 5.7%.



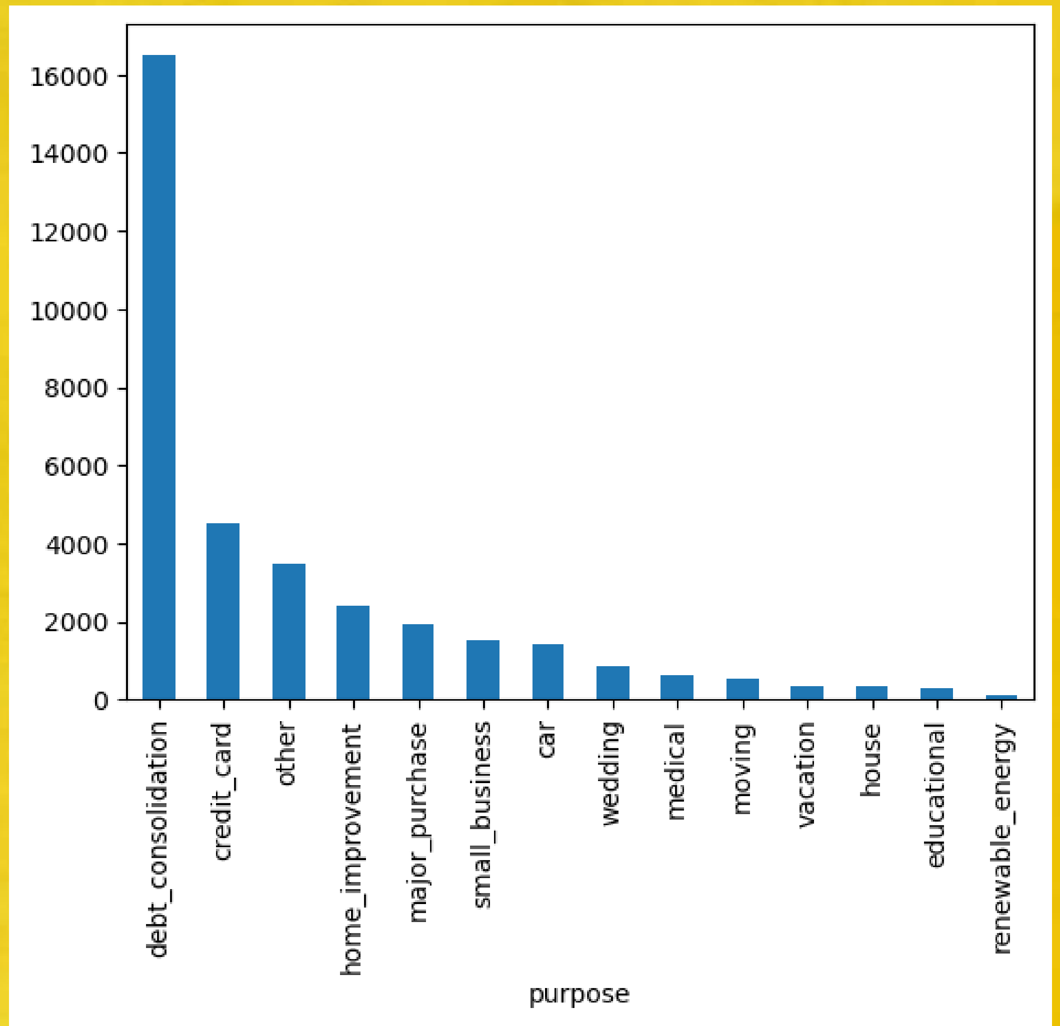
Univarients Analysis

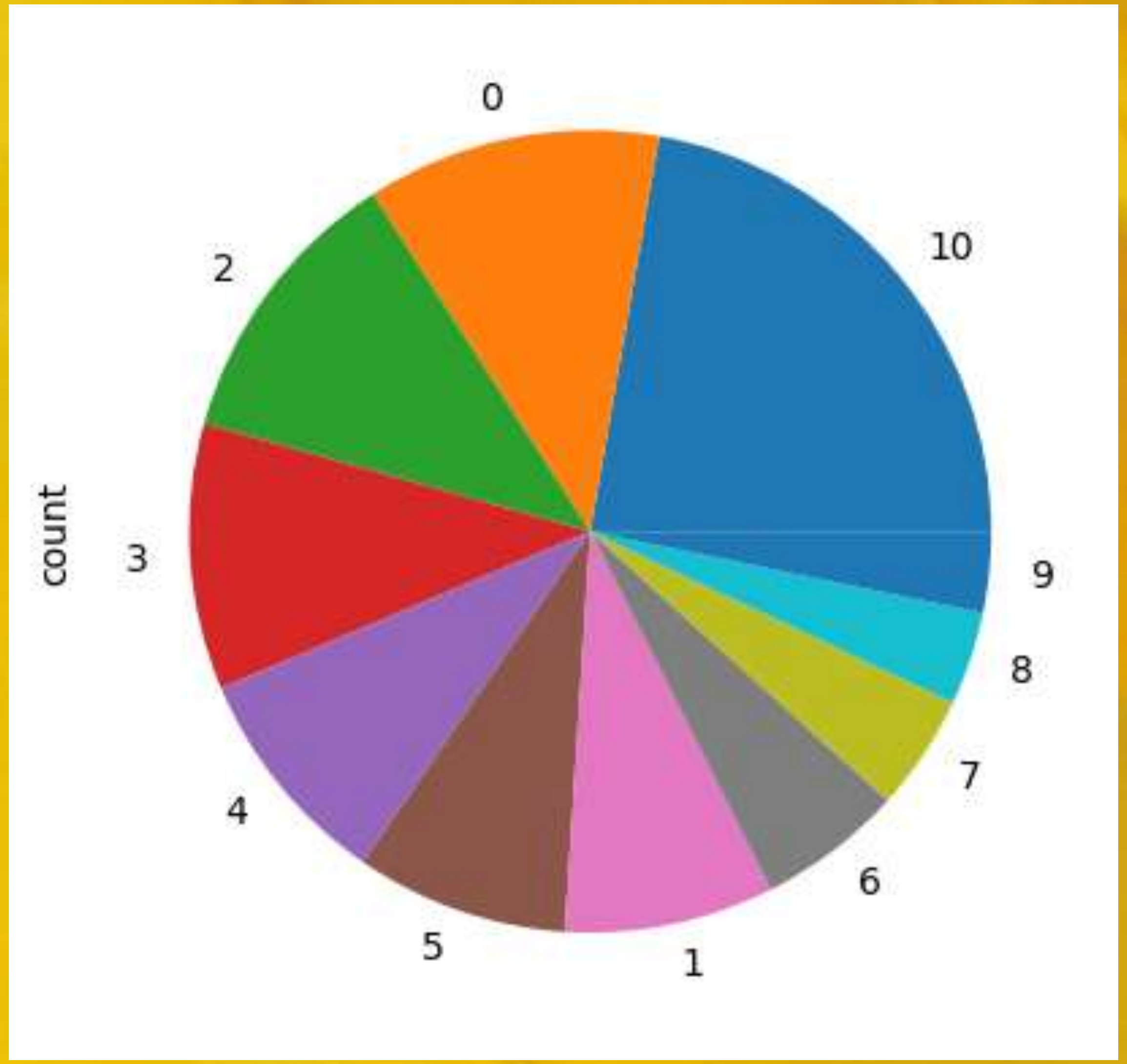
Unordered & Ordered Categorical Variable Analysis





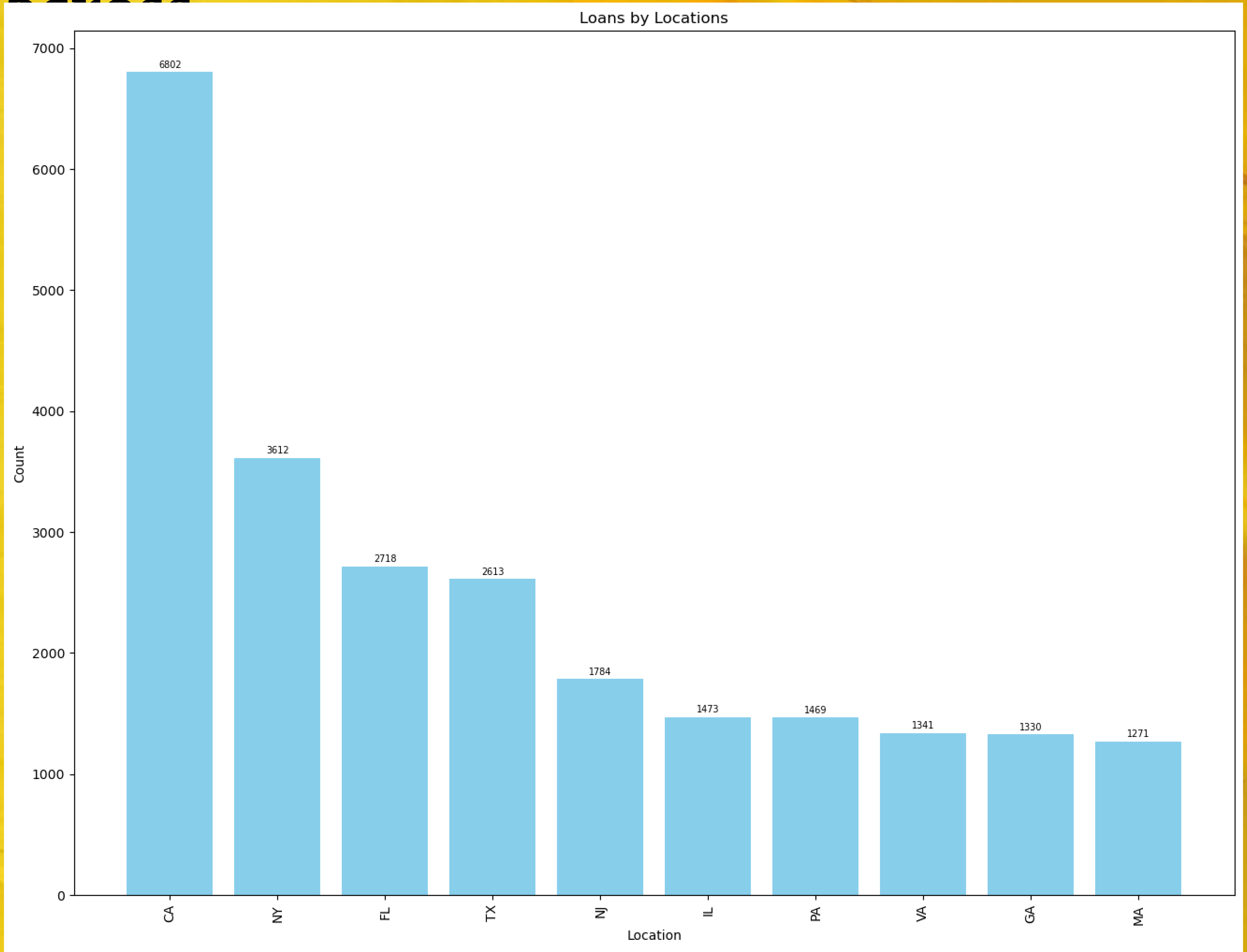
- **Observations**:
  - Majority of loan applicants are either living on Rent or on Mortgage
  - Most of the loan applicants are for debt_consolidations
  - Most of the Loan applicants are from CA(State).
  - Most of the applications are having 10+ yrs of Exp.





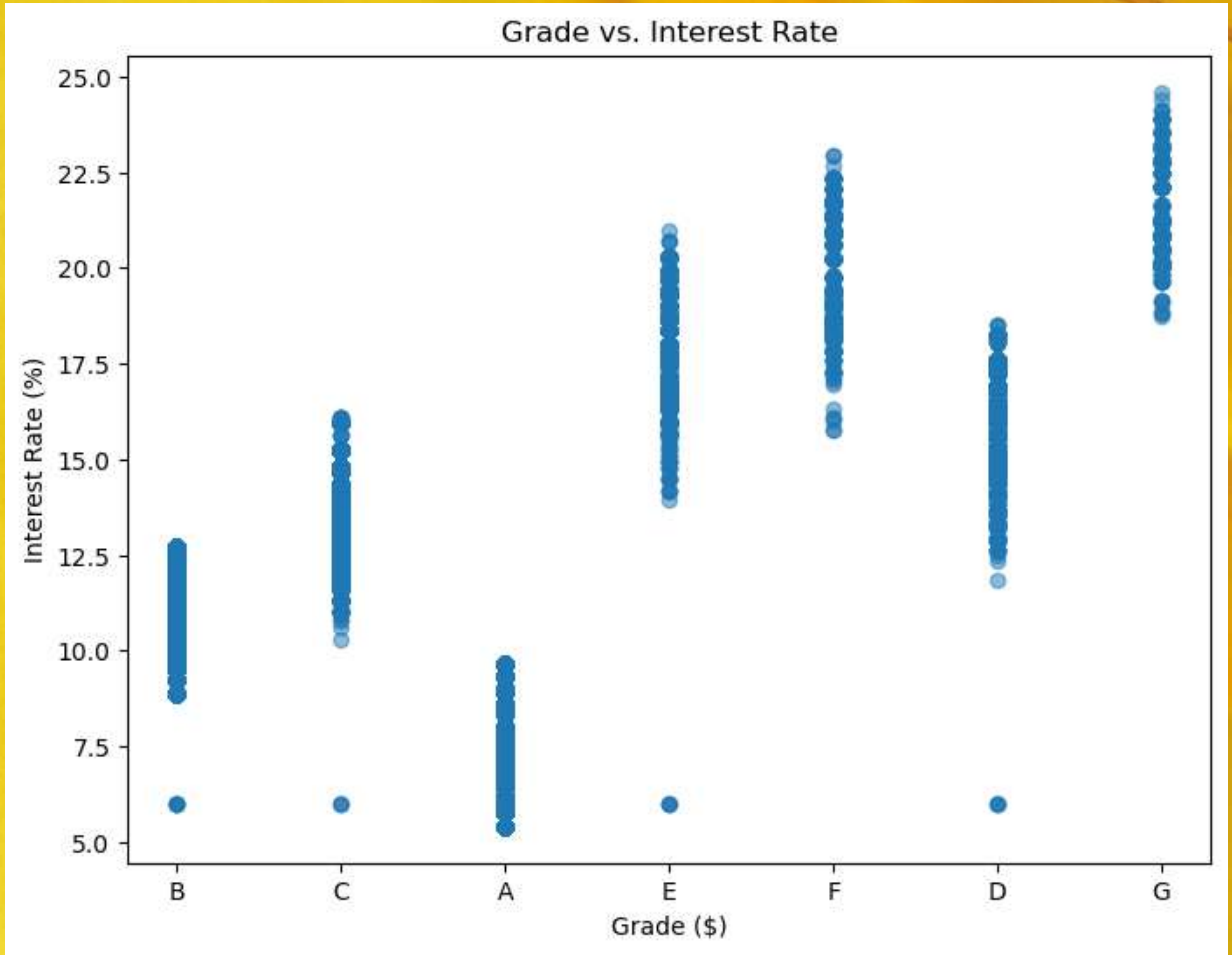
Loan Distribution across Top Ten States

1. CA state has the highest number of loan count
2. MA has the lowest number of loan count



Analysis on Intrest Percentage across Grade

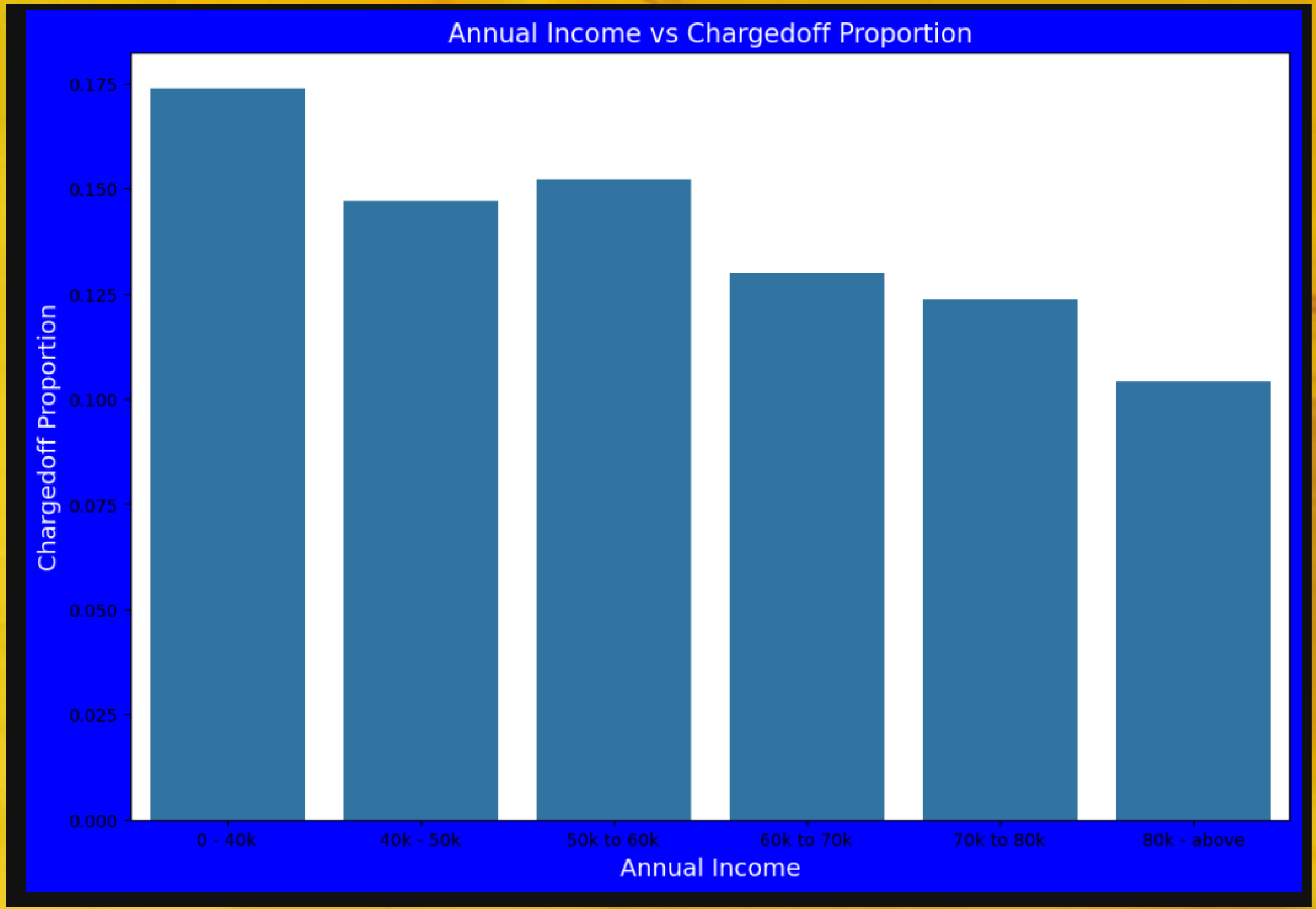
1. Above concludes Lowest Interest Percentage for loan is for A grade
2. Above concludes Highest Interest Percentage for loan is for G grade closer to 25%



Bivariate Analysis

Annual income vs Charged Off

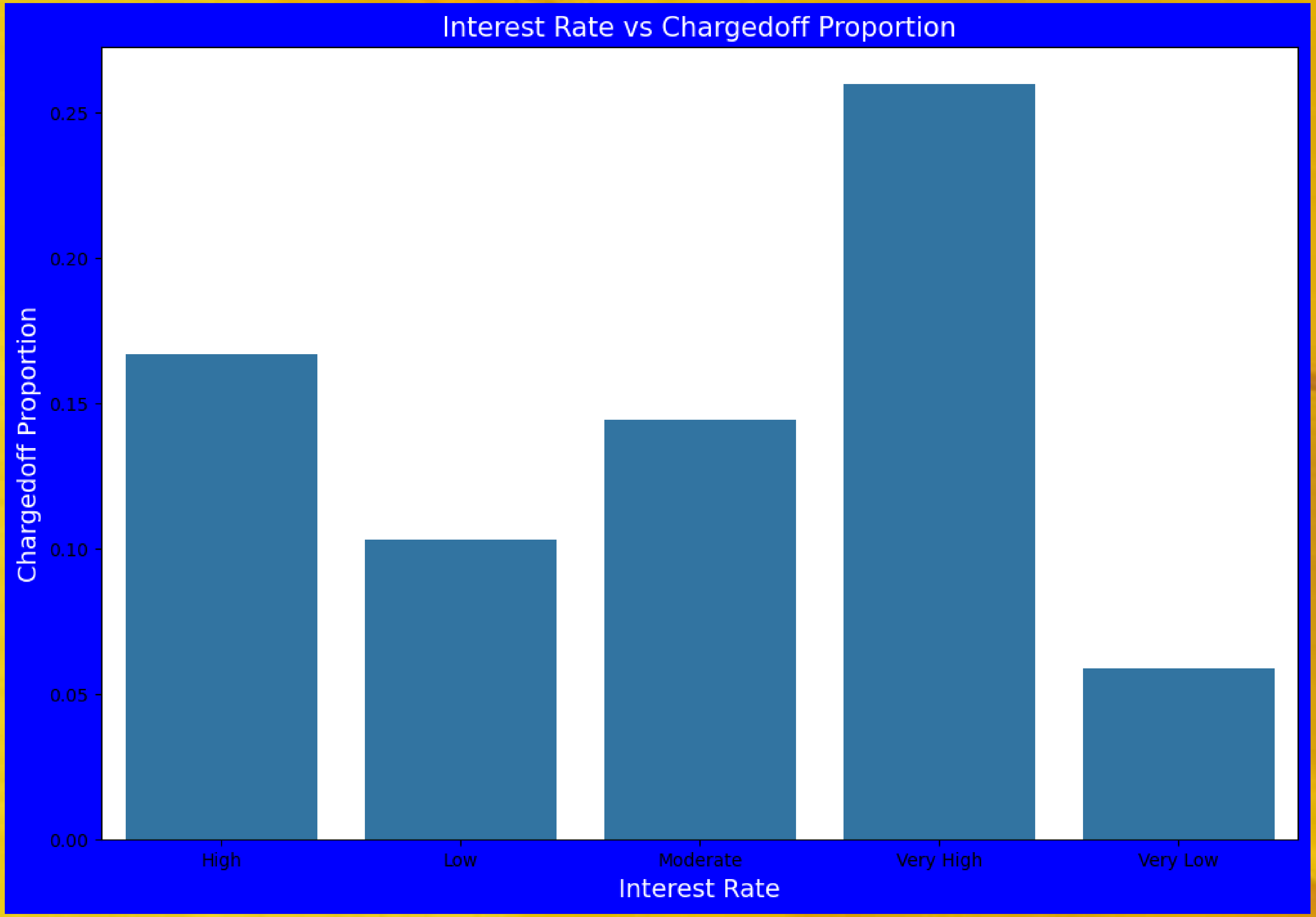
- **Observations:**
  - Income range 80000+ has less chances of charged off.
  - Income range 0-20000 has high chances of charged off.
  - Notice that with increase in annual income charged off proportion got decreased.



oan_status	int_rate_b	Charged Off	Current	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1719	431	4905	6624	0.259511
0	High	983	164	4922	5905	0.166469
2	Moderate	978	223	5814	6792	0.143993
1	Low	588	115	5121	5709	0.102995
4	Very Low	514	36	8224	8738	0.058824

Interest Rate vs Charged off

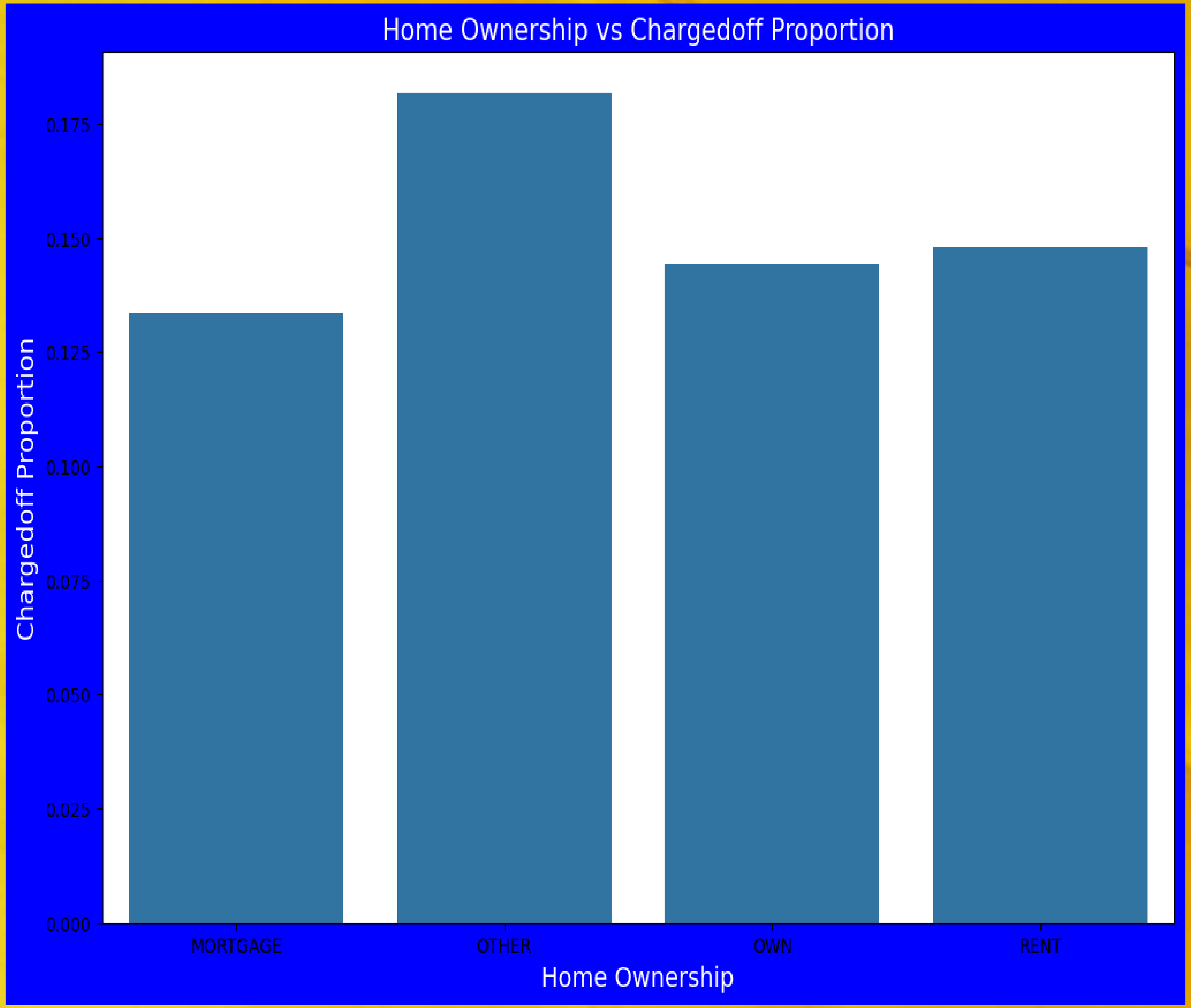
- **Observations:**
  - Interest rate less than 10% or very low has very less chances of charged off. Interest rates are starting from minimum 5 %.
  - Interest rate more than 16% or very high has good chances of charged off as compared to other category interest rates.
  - Charged off proportion is increasing with higher interest rates.



loan_status	home_ownership	Charged Off	Current	Fully Paid	Total	Chargedoff_Proportion
1	OTHER	16.0	0.0	72.0	88.0	0.181818
3	RENT	2486.0	384.0	14312.0	16798.0	0.147994
2	OWN	361.0	63.0	2142.0	2503.0	0.144227
0	MORTGAGE	1919.0	522.0	12460.0	14379.0	0.133459

Home Ownership vs Charged off

- **Observations:**
  - Those who are not owning the home is having high chances of loan defaulter.
  - From the graph even shows high chances of charged off. Proportions, but data available is very limited compared to other points

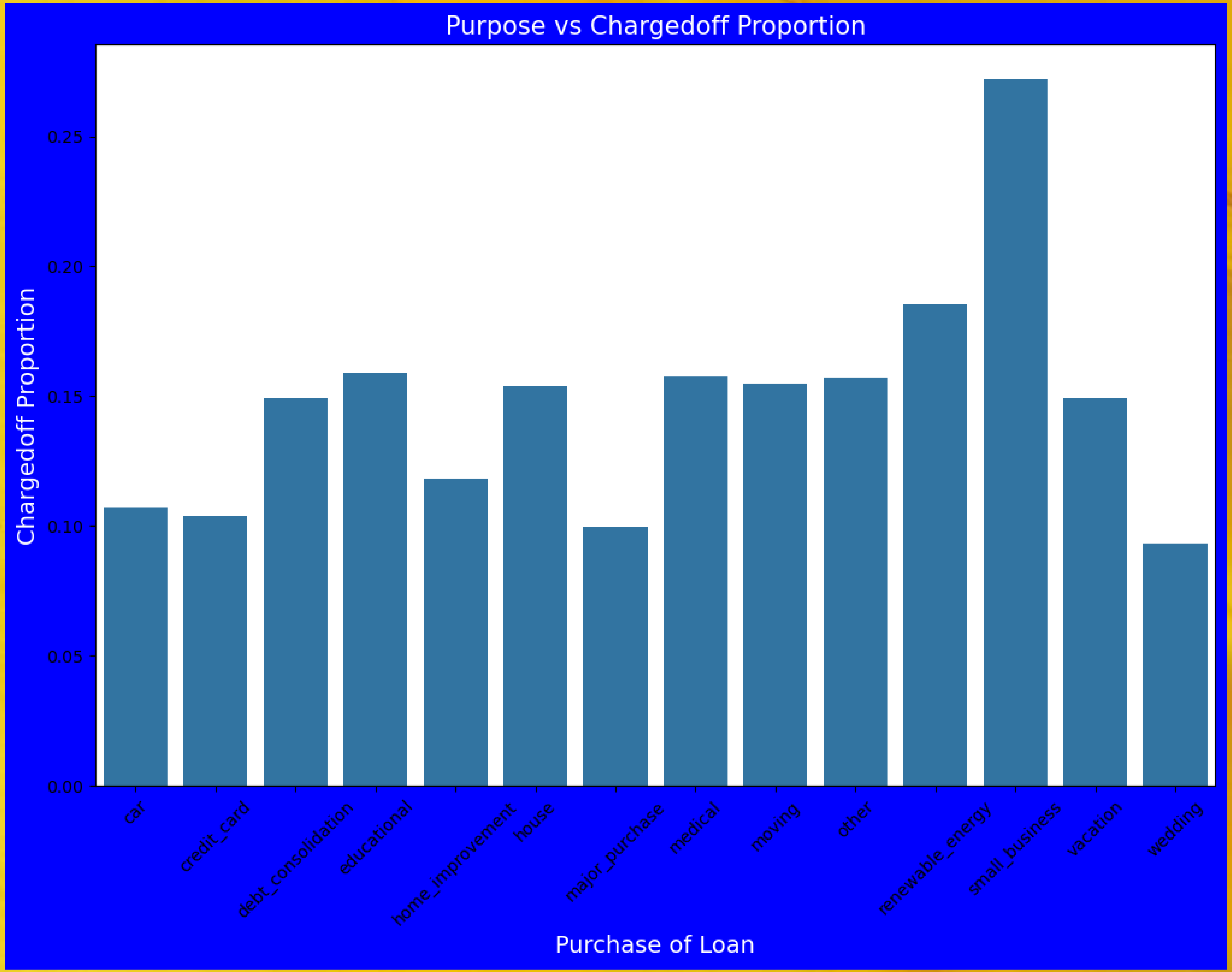


loan_status	home_ownership	Charged Off	Current	Fully Paid	Total	Chargedoff_Proportion
1	OTHER	16.0	0.0	72.0	88.0	0.181818
3	RENT	2486.0	384.0	14312.0	16798.0	0.147994
2	OWN	361.0	63.0	2142.0	2503.0	0.144227
0	MORTGAGE	1919.0	522.0	12460.0	14379.0	0.133459

Purpose vs Charged Off

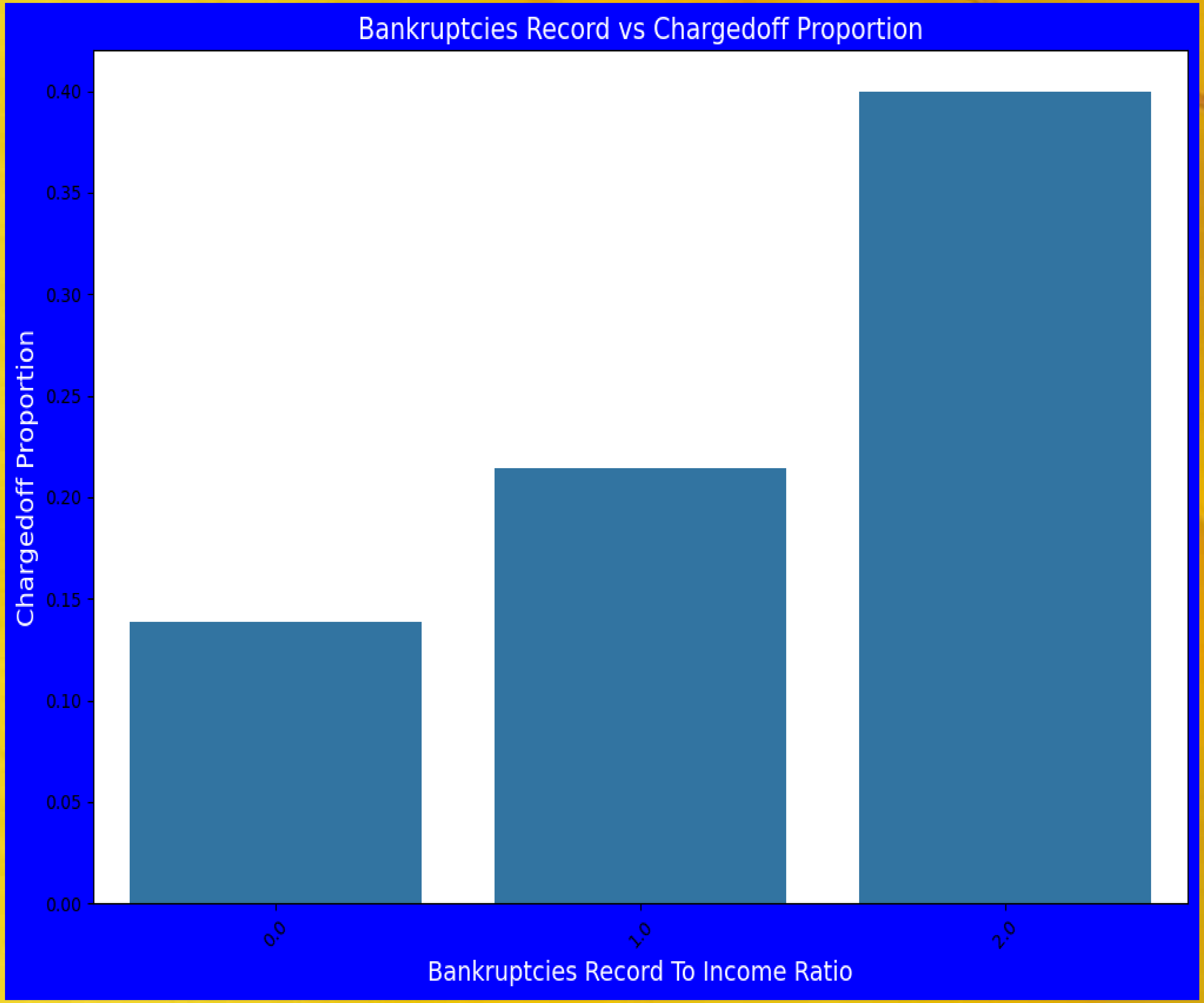
- **Observations:**
  - Those applicants who is having home loan is having low chances of loan defaults.
  - Those applicants having loan for small business is having high chances for loan defaults.

loan_status	purpose	Charged Off	Current	Fully Paid	Total	Chargedoff_Proportion
11	small_business	389.0	68.0	1042.0	1431.0	0.271838
10	renewable_energy	15.0	1.0	66.0	81.0	0.185185
3	educational	44.0	0.0	233.0	277.0	0.158845
7	medical	95.0	7.0	508.0	603.0	0.157546
9	other	528.0	107.0	2837.0	3365.0	0.156909
8	moving	79.0	7.0	431.0	510.0	0.154902
5	house	47.0	11.0	258.0	305.0	0.154098
12	vacation	49.0	4.0	279.0	328.0	0.149390
2	debt_consolidation	2385.0	502.0	13608.0	15993.0	0.149128
4	home_improvement	277.0	73.0	2068.0	2345.0	0.118124
0	car	147.0	48.0	1223.0	1370.0	0.107299
1	credit_card	458.0	89.0	3950.0	4408.0	0.103902
6	major_purchase	191.0	33.0	1723.0	1914.0	0.099791
13	wedding	78.0	19.0	760.0	838.0	0.093079



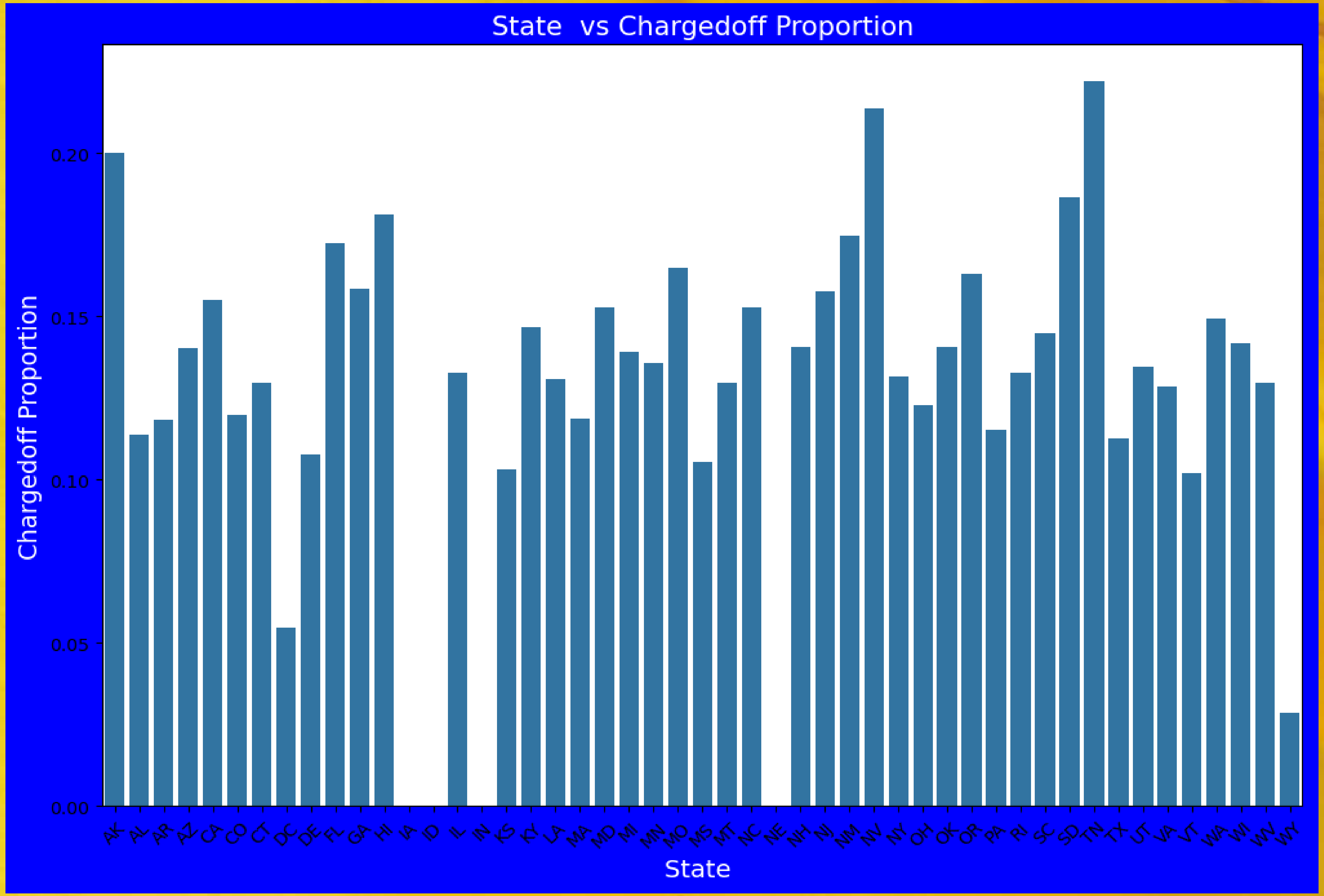
Bankruptcies Record vs Charged off

- **Observations:**
  - 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.



State vs Charged off

- **Observations:**
  - DE States is holding highest number of loan defaults.
  - CA is having low number of loan defaults



Correlation

Correlations

- Negative Correlation:

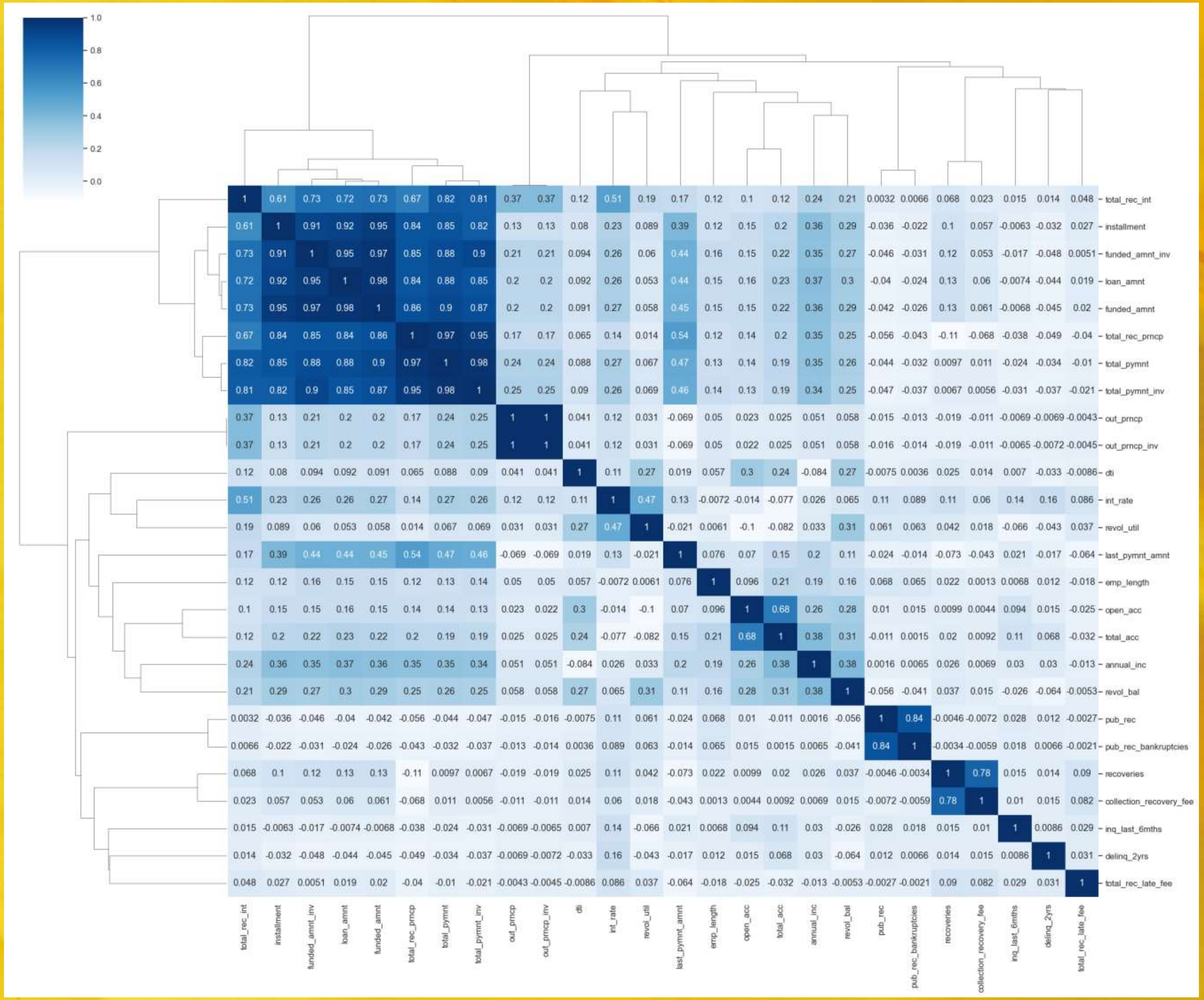
1. loan_amnt has negative correlation with pub_rec_bankrupticies
2. annual income has a negative correlation with dti

- Strong Correlation:

  1.term has a strong correlation with loan amount

  2. term has a strong correlation with interest rate

  3. annual income has a strong correlation with loan_amount



Conclusions

- Income range between 0-20000 has high chances of charged off.
- Interest rate more than 16% has good chances of charged off as compared to other category interest rates.
- Those who are not owning the home is having high chances of loan defaulter.
- Those applicants having loan for small business is having high chances for loan defaults.
- High DTI value having high risk of defaults.
- Higher the Bankruptcies record higher the chance of loan defaults.
- DE States is holding highest number of loan defaults.
- The Loan applicants with loan Grade G is having highest Loan Defaults.