Lending Club Case Study

Exploratory Data Analysis

- Problem Statement
- Data Summary
- Data Cleaning
- Data conversions and Derived Columns
- Dropping/Imputing the Rows
- Outliers
- Univariate Analysis
- Bivariate Analysis
- Correlations
- Conclusion

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 not available with the company.

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’ does n’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 Conversion and 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

Dropping/Imputing the Rows

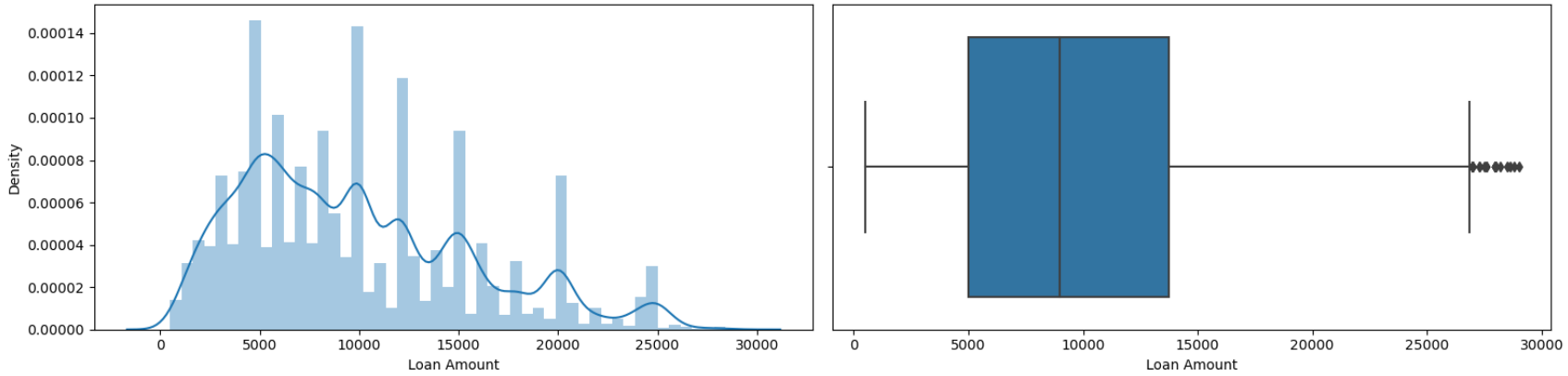
- ‘emp_lenght’ and pub_rec_bankruptcies contains 2.67% and 1.80% of rows as null, which is very small percetnage of data which we can drop it.
- Total % of rows deleted: 4.48%
- Outliers exits for numeric data 'loan_amnt', 'funded_amnt', ‘funded_amnt_inv','int_rate', 'installment', ‘annual_inc’.
- Outliers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

Univariate Analysis

Loan Amount

Observations:

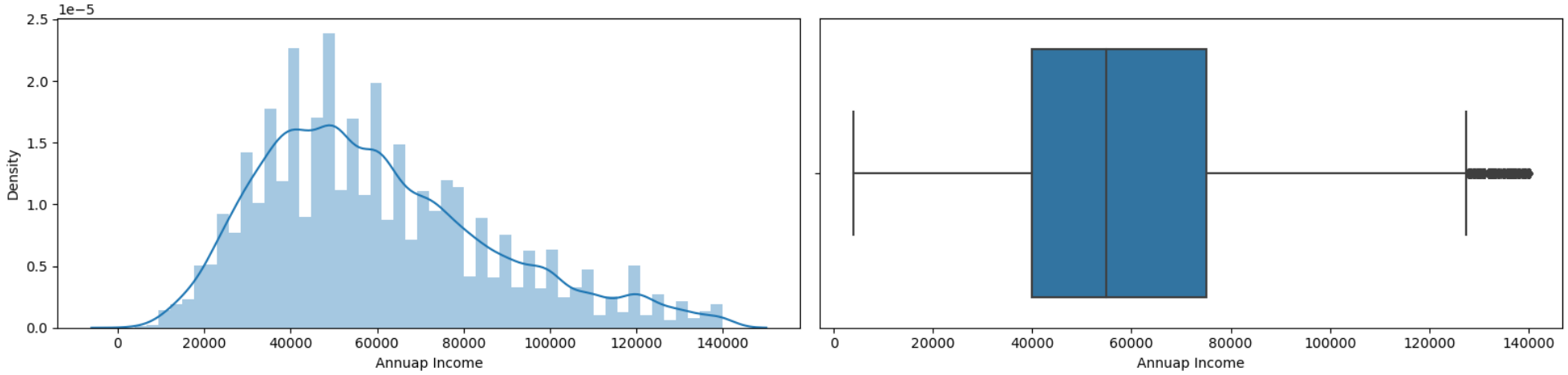
- Most of the loan amount applied was in the range of 5k-14k. •
- Max Loan amount applied was ~27k



Annual Income

Observations:

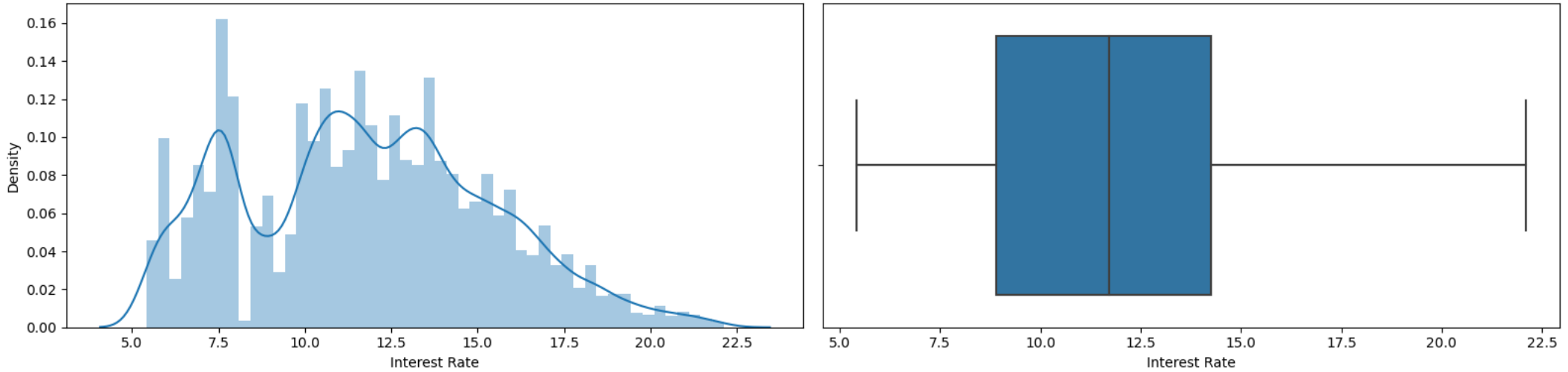
- The Annual income of most if applicants lies between 40k-75k.
- Average Annual Income is : 59883.0



Interest Rate

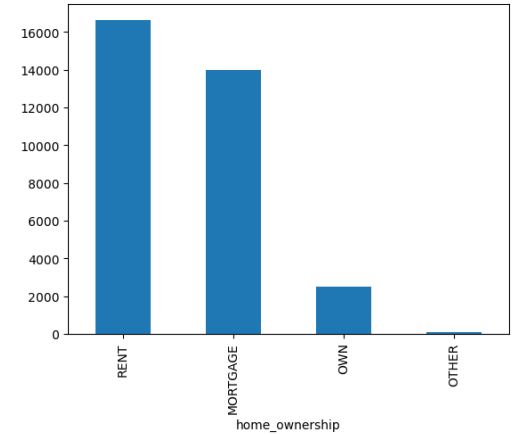
Observations:

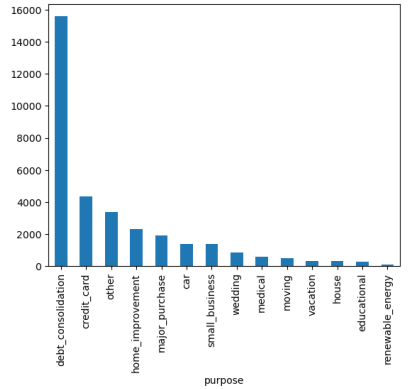
- Most of the applicant's rate of interest is between in the range of 8%-14%.
- Average Rate of interest of rate is 11.7%



Univariate Analysis

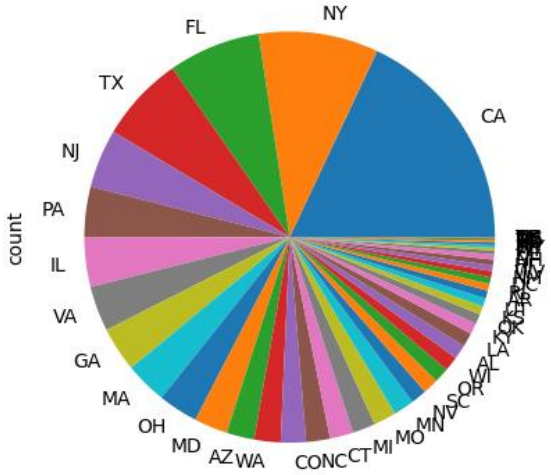
Unordered & Ordered Categorical Variable Analysis

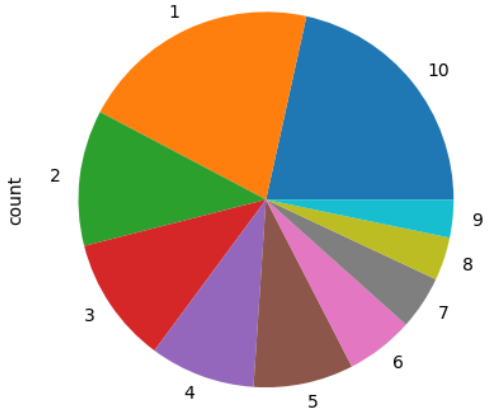




Observations:

- 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.





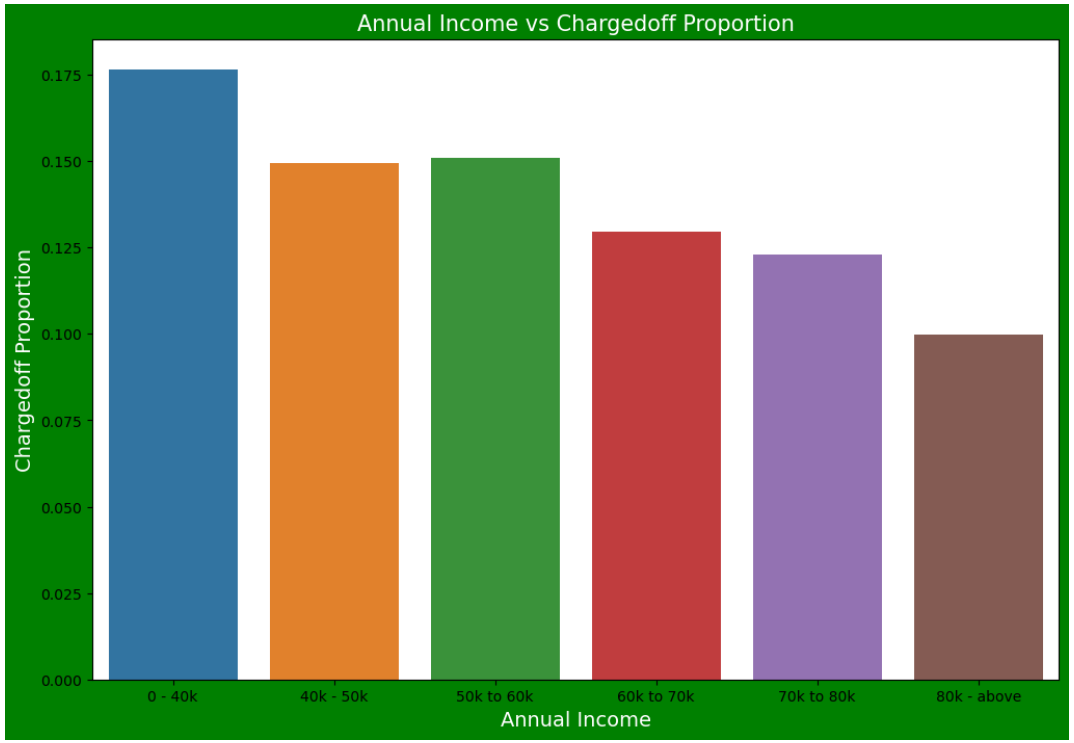
Bivariate Analysis

Annual Income vs Charged-off

loan_status	annual_inc_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
0	0 - 40k	1570	7326	8896	0.176484
2	50k to 60k	788	4435	5223	0.150871
1	40k - 50k	807	4593	5400	0.149444
3	60k to 70k	486	3261	3747	0.129704
4	70k to 80k	385	2749	3134	0.122846
5	80k - above	678	6113	6791	0.099838

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 decrease

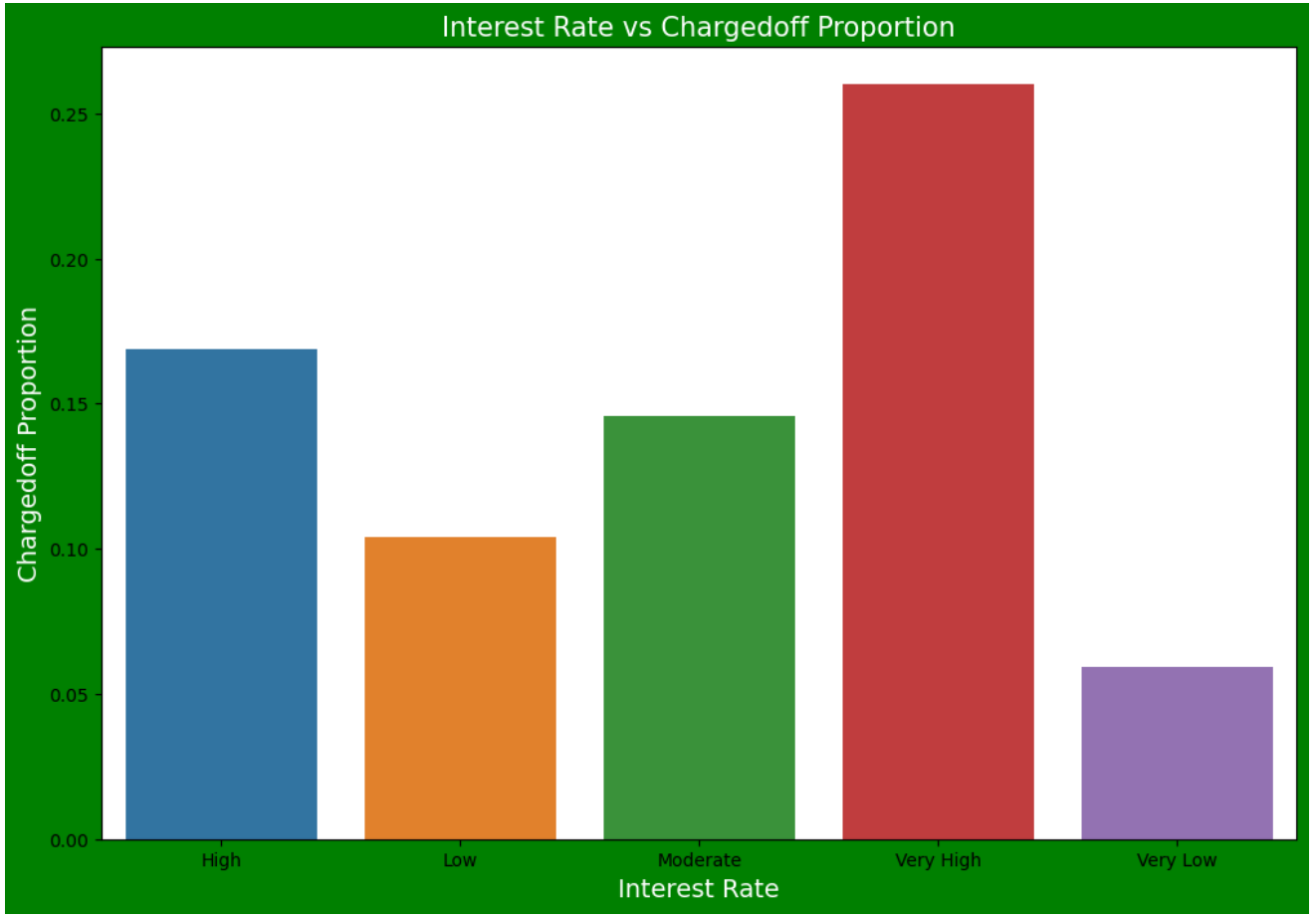


Interest rate vs Charged-off

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

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

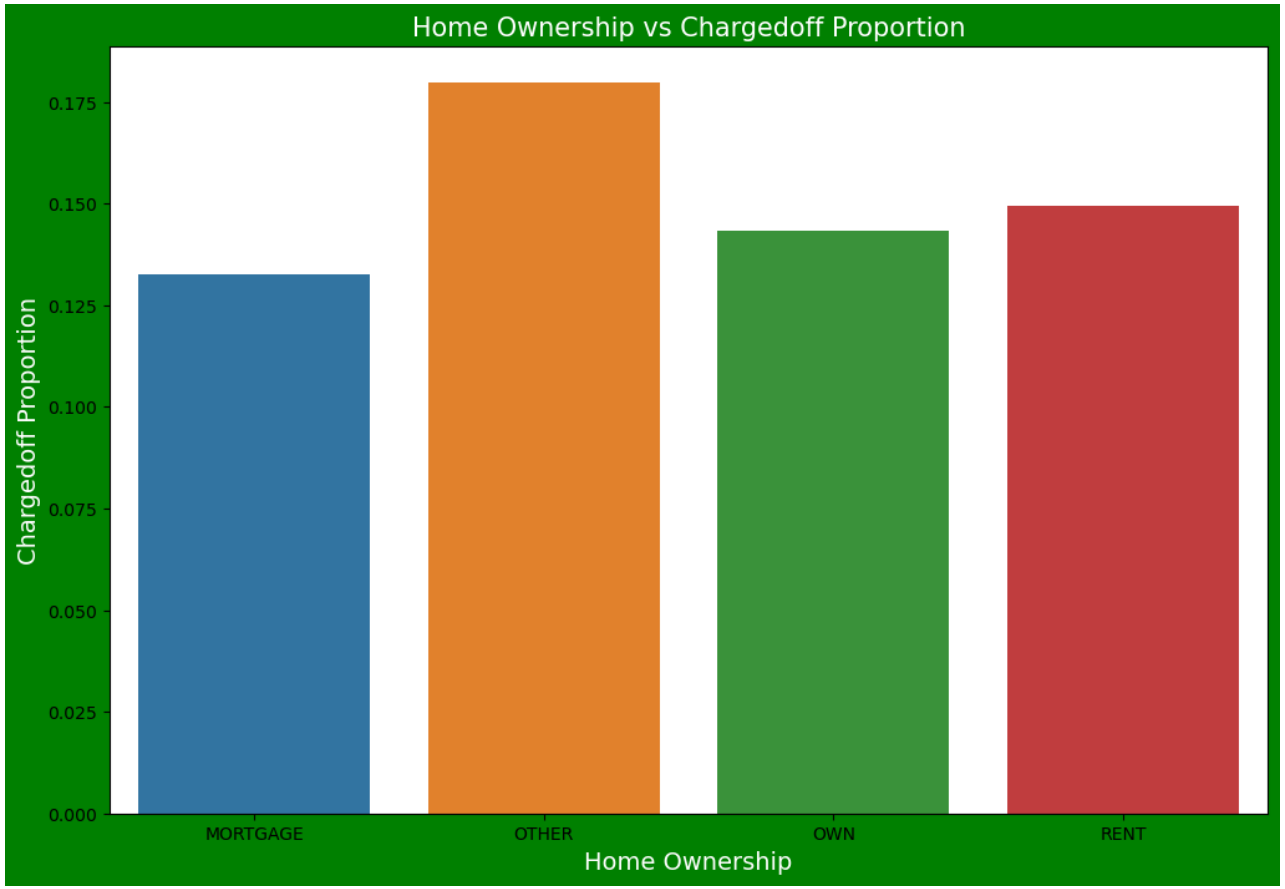


Home Ownership vs Charged-off

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

Observations:

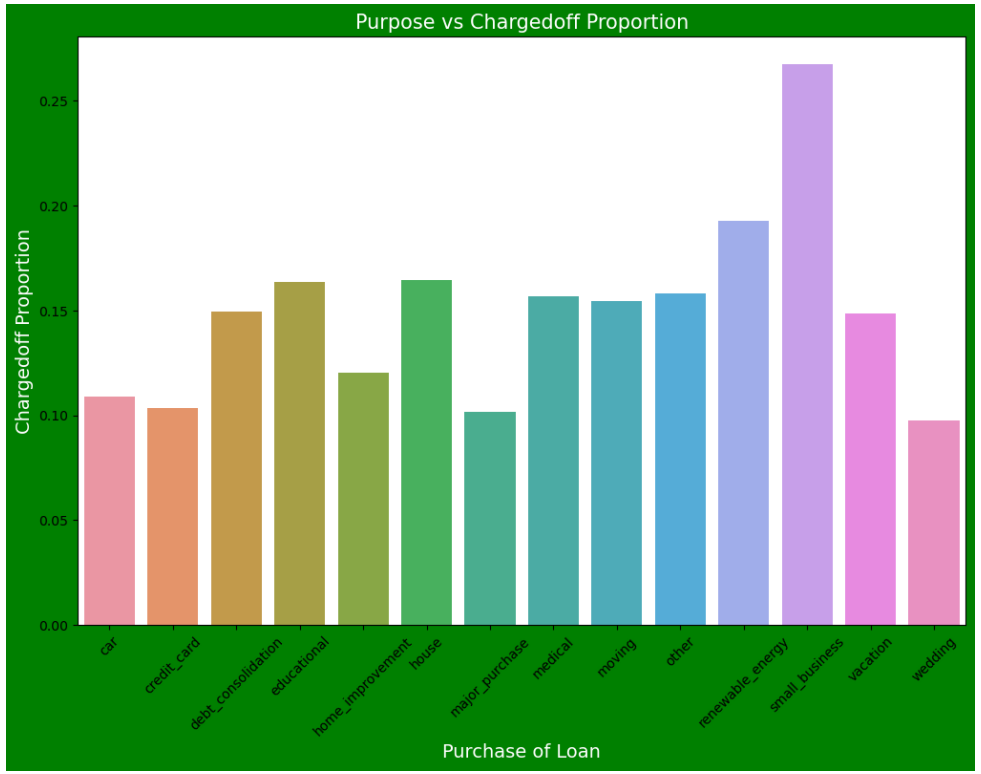
- Those who are not owning the home is having high chances of loan defaulter.
- From the graph even shows high chances of charged off. Proportions, but data available is very limited compared to other points



Purpose vs charged-of

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.



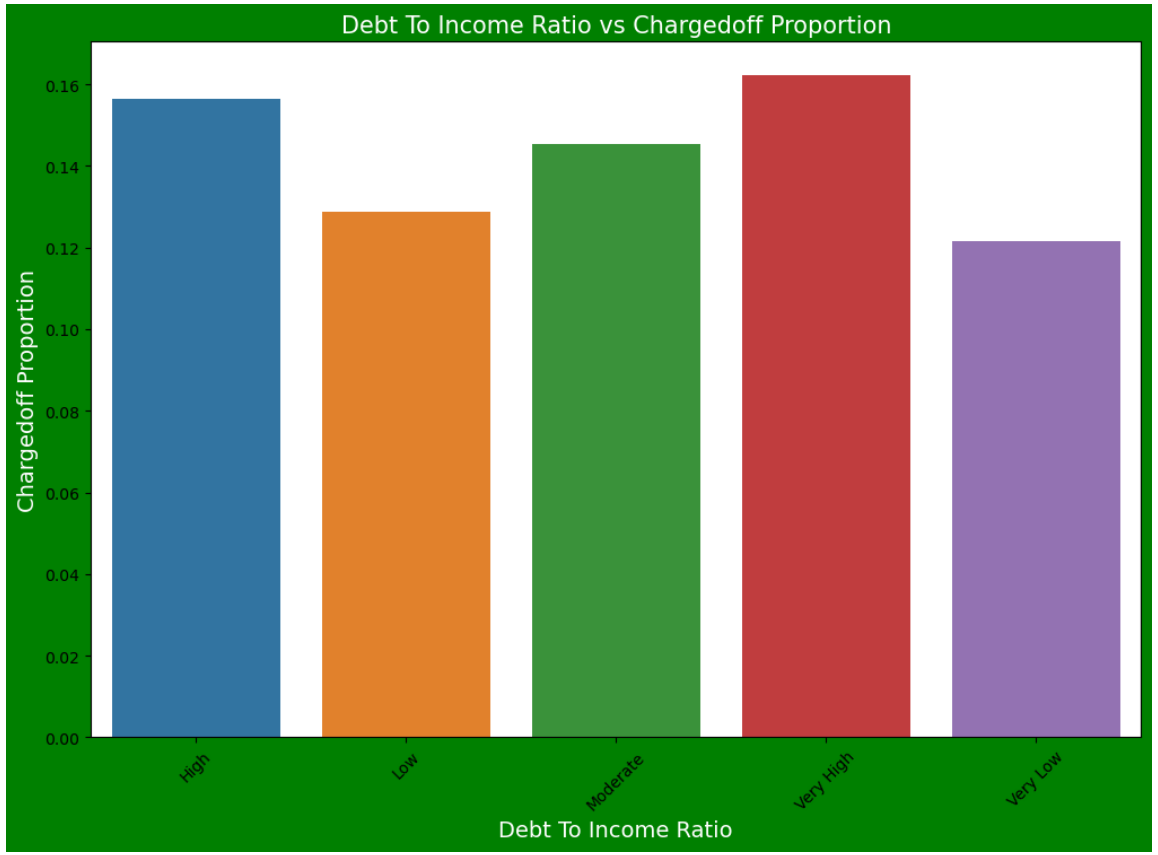
loan_status	purpose	Charged Off	Fully Paid	Total	Chargedoff_Proportion
11	small_business	366	1003	1369	0.267348
10	renewable_energy	16	67	83	0.192771
5	house	49	249	298	0.164430
3	educational	46	235	281	0.163701
9	other	531	2823	3354	0.158318
7	medical	95	510	605	0.157025
8	moving	79	433	512	0.154297
2	debt_consolidation	2329	13253	15582	0.149467
12	vacation	49	281	330	0.148485
4	home_improvement	277	2026	2303	0.120278
0	car	150	1224	1374	0.109170
1	credit_card	450	3894	4344	0.103591
6	major_purchase	195	1719	1914	0.101881
13	wedding	82	760	842	0.097387

DTI vs Charged-off

Observation:

- High DTI value having high risk of defaults
- Lower the DTO having low chances loan defaults

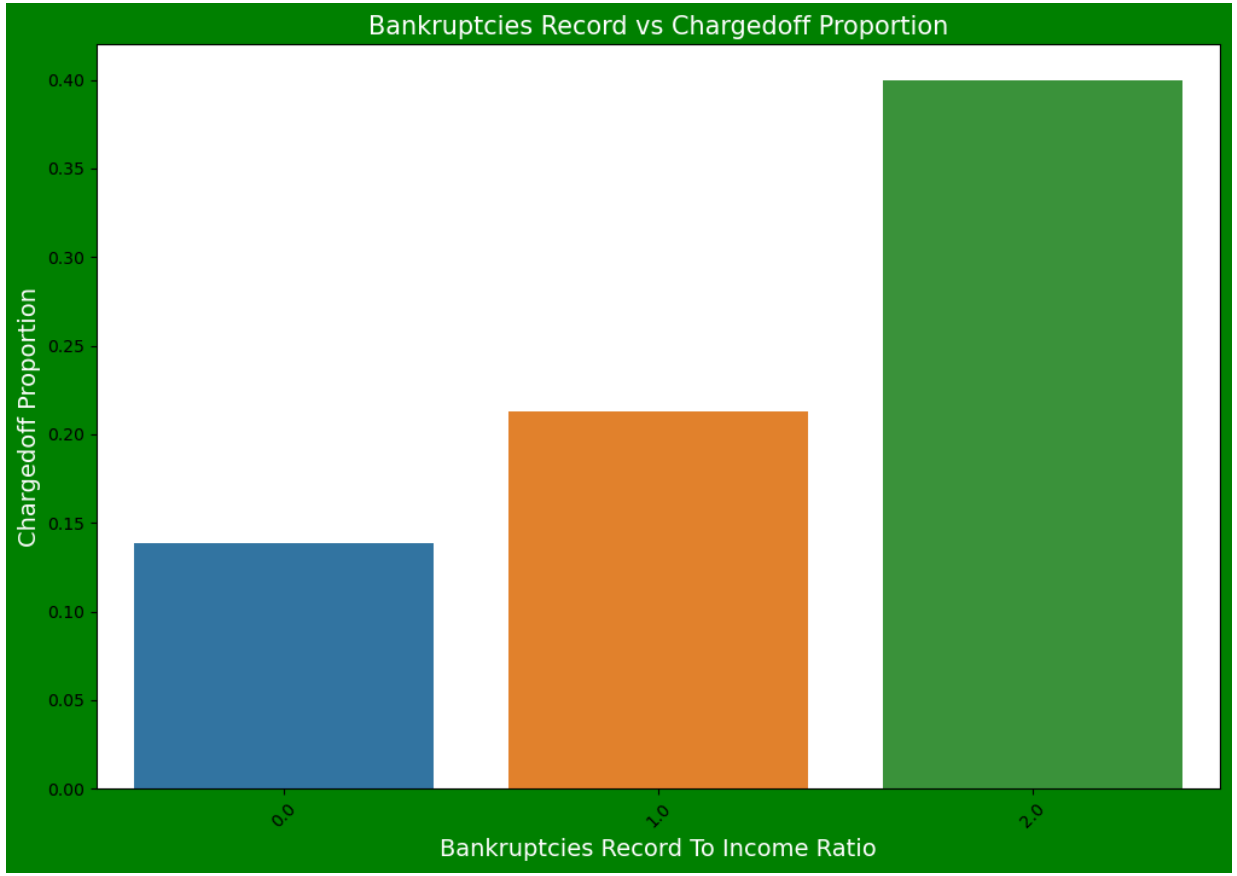
loan_status	dti_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1044	5387	6431	0.162339
0	High	948	5111	6059	0.156461
2	Moderate	985	5785	6770	0.145495
1	Low	789	5339	6128	0.128753
4	Very Low	948	6855	7803	0.121492



Bankruptcies Record vs Charged-off

Observations:

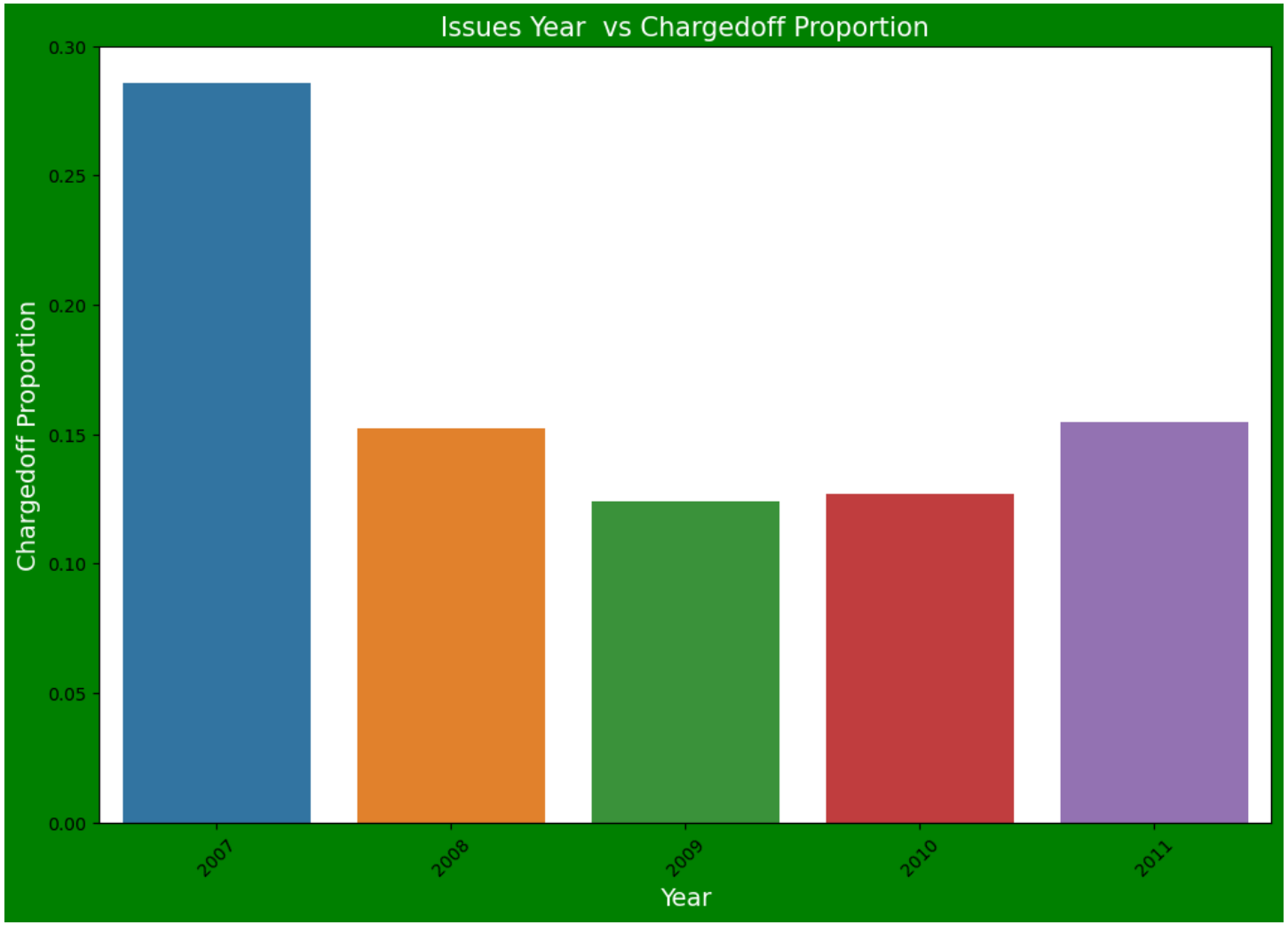
- Bankruptcies Record with 2 is having high impact on loan defaults
- Bankruptcies Record with 0 is low impact on loan defaults
- Lower the Bankruptcies lower the risk



Issue Year vs Charged-off

Observations:

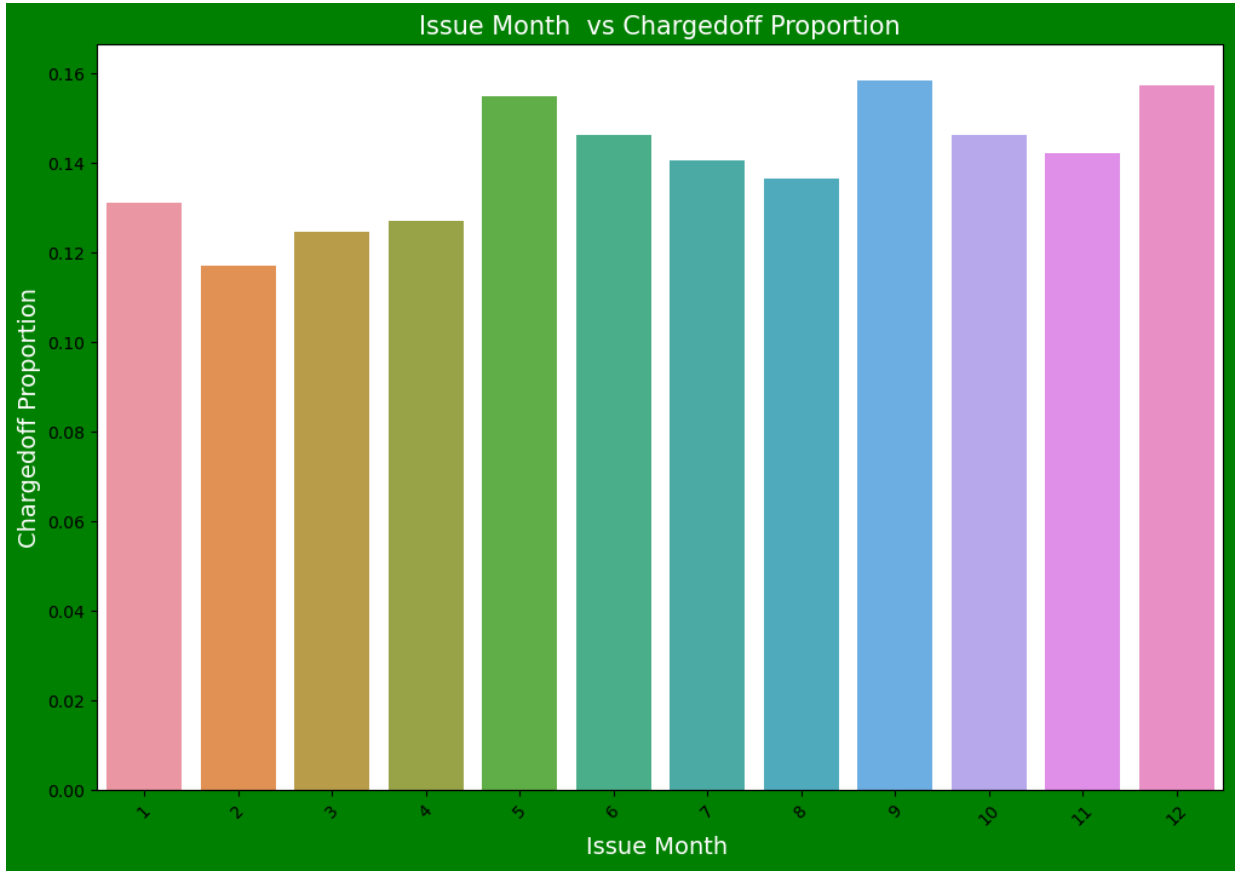
- Year 2007 is highest loan defaults.
- 2009 is having lowest loan defaults



Issue month vs Charged-off

Observations:

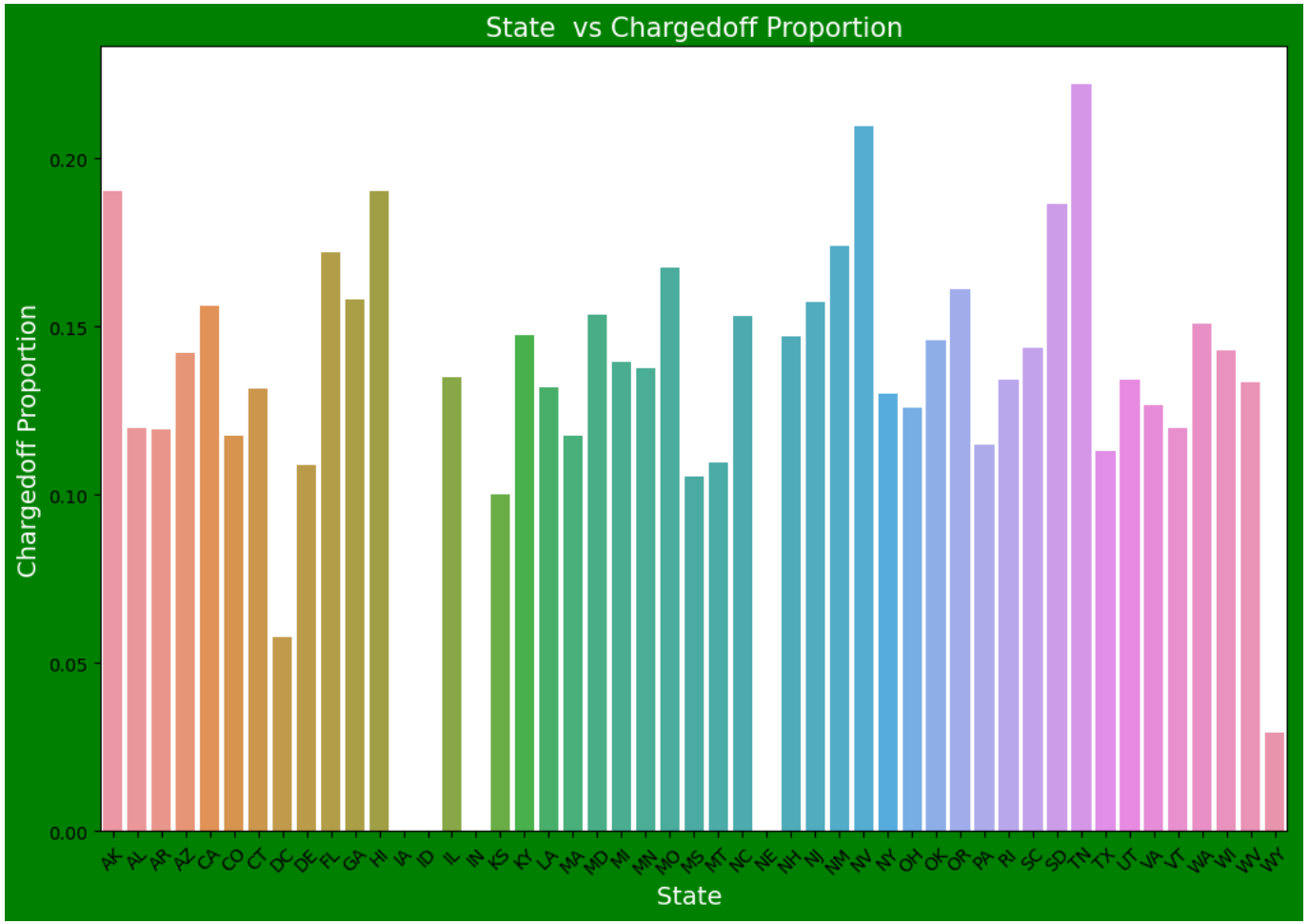
- Those loan has been issued in May, September and December is having high number of loan defaults
- Those loan has been issued in month of February is having high number of loan defaults
- Majority of loan defaults coming from applicants whose loan has been approved from September-to December



State vs Charged-off

Observations:

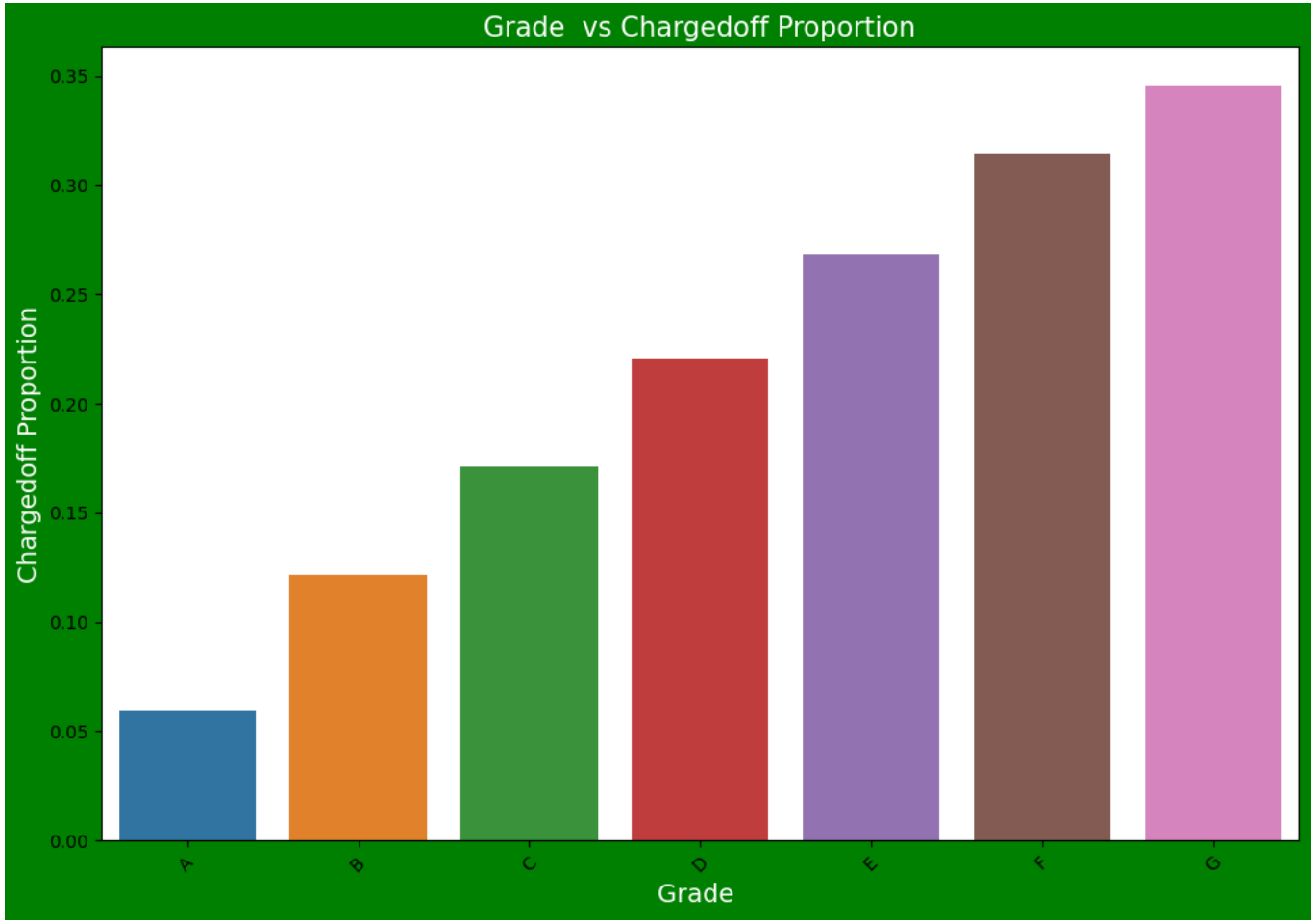
- DE States have highest number of loan defaults.
- CA have low number of loan default



Grade vs Charged-off

Observations:

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



Correlation

- **Negative Correlation:**

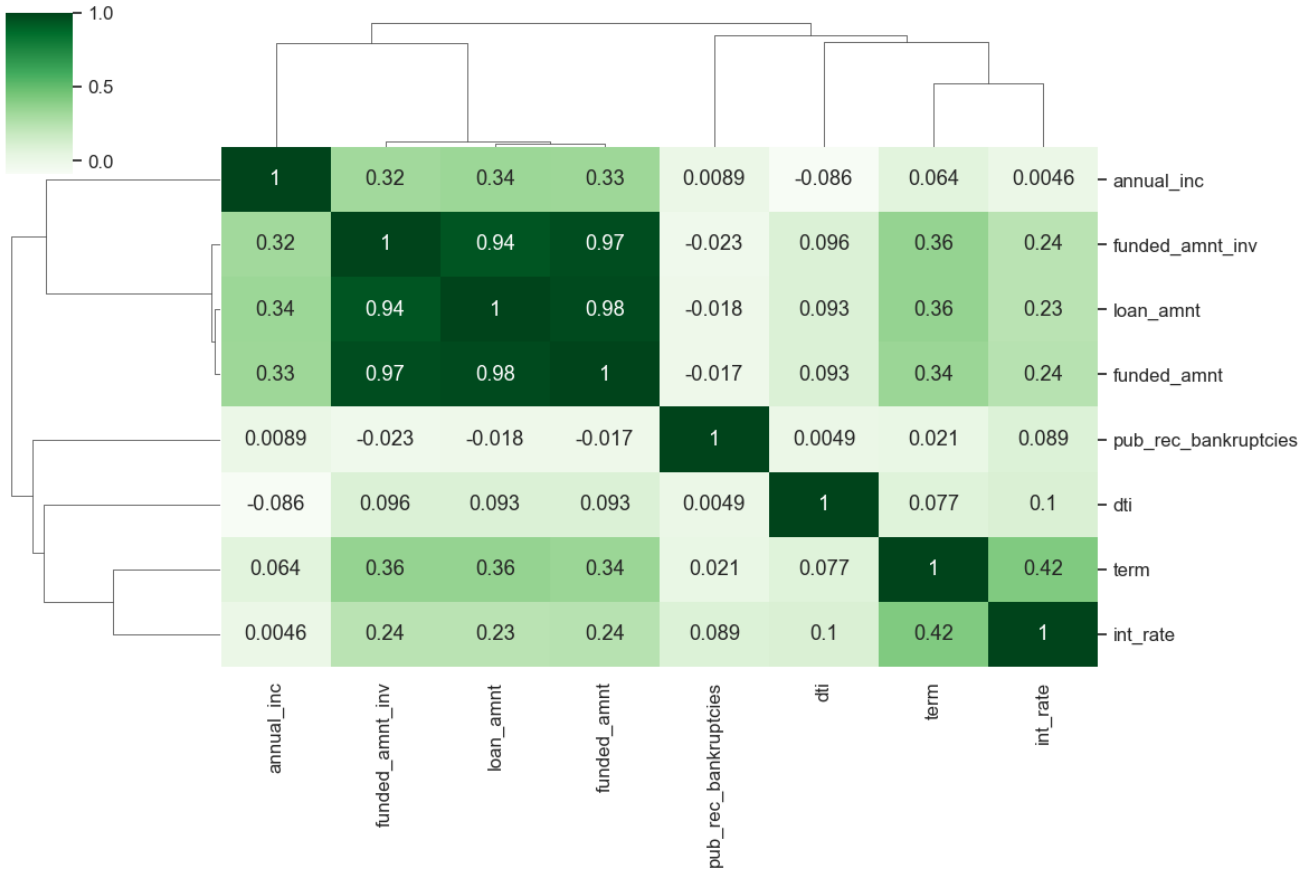
1. loan_amnt has negative correlation with pub_rec_bankrupticies

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