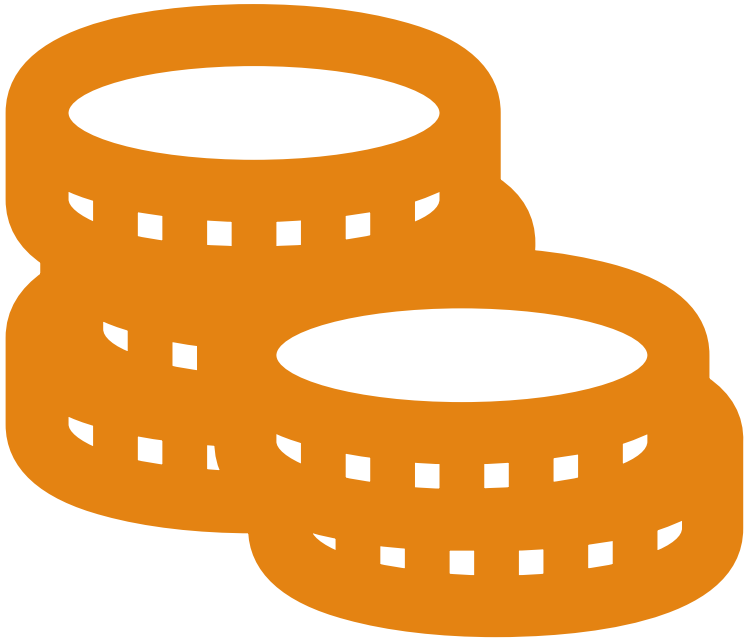


Lending Club Case Study

EXPLORATORY DATA ANALYSIS

SANGEETHA RAMMOHAN

SAMEEULLA

Content

Problem Statement

Problem:

A consumer finance company which specialises in lending various types of loans to urban customers, receives loan applications. 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 analyse the influence of consumer attributes and loan attributes on tendency of default

Constraint:

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)

Data Summary

The Data provided has information about 39717 applicants and 111 attributes.

The attributes are classified into Loan attributes and Customer attributes.

Data Cleaning and Manipulation

Data Cleaning

- There are no headers, footers, summary and total rows for removal
- There are no duplicate application data identified
- Identified and removed 54 attributes with 100% null data
- Identified and removed unique attributes like ‘url’ and ‘member_id’ as they don’t influence charge-off calculation, preserved ‘id’ for reference
- Identified and removed description attributes like ‘desc’ and ‘title’
- Limiting the analysis to group level and hence removed sub-group attribute
- Removed the data of ‘Current’ loan applicants as it does not participate in the analysis.
- With the support of domain knowledge, 21 behavioural attributes that would not be available during loan approval and hence doesn’t participate in analysis were removed
- Attributes with more than 50% null values are removed
- Post data cleaning we are left with 19 attributes of 38577 applications for analysis.

Data Cleaning and Manipulation

Data Manipulation

Identified below attributes that are in text format and converted them to numeric fields: ‘term’, ‘int_rate’, ‘loan_amt’, ‘funded_amt’

‘issue_d’ converted to date format

‘issue_year’ and ‘issue_month’ derived from the date ‘issue_d’

‘loan_amnt_b’, ‘annual_inc_b’, ‘int_rate_b’ and ‘dti_b’ derived into bins for better analysis

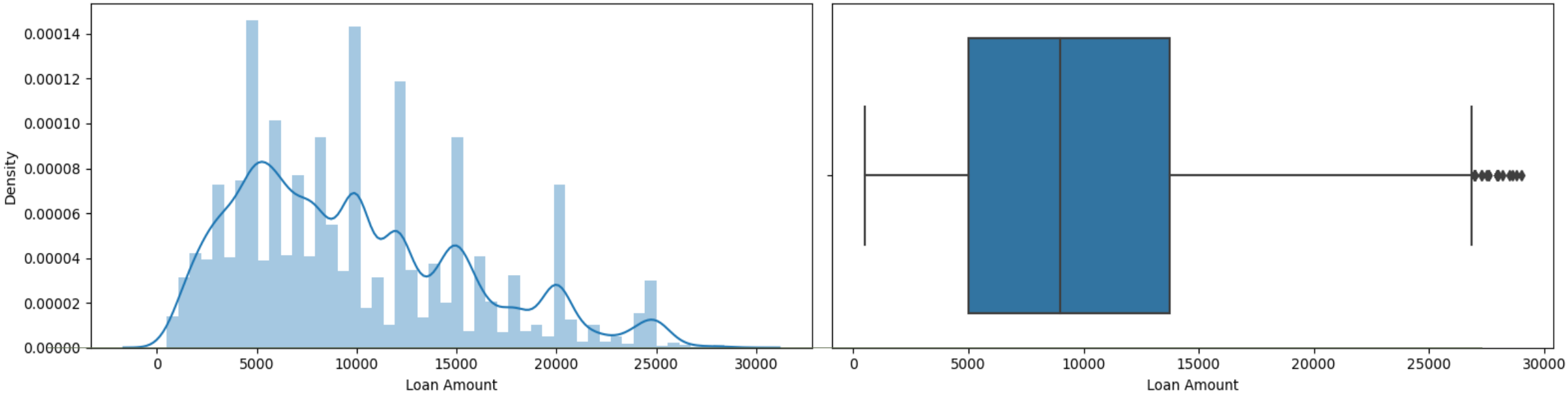
Removal of rows that have null values less than 5% (emp_length – 2.67%, pub_rec_bankruptcies – 1.8%)

Outliers identified and treated with quantile mechanism

Loan Amount

Observations:

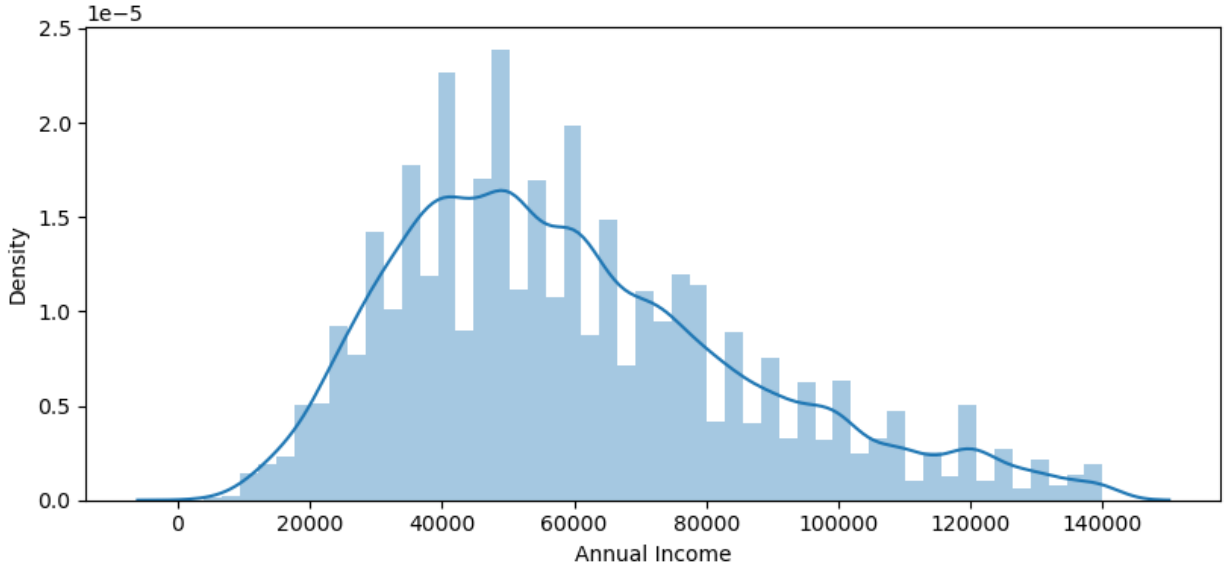
- Most of the loan amount applied is in the range of 5k to 14k
- Maximum loan amount applied for is 29k

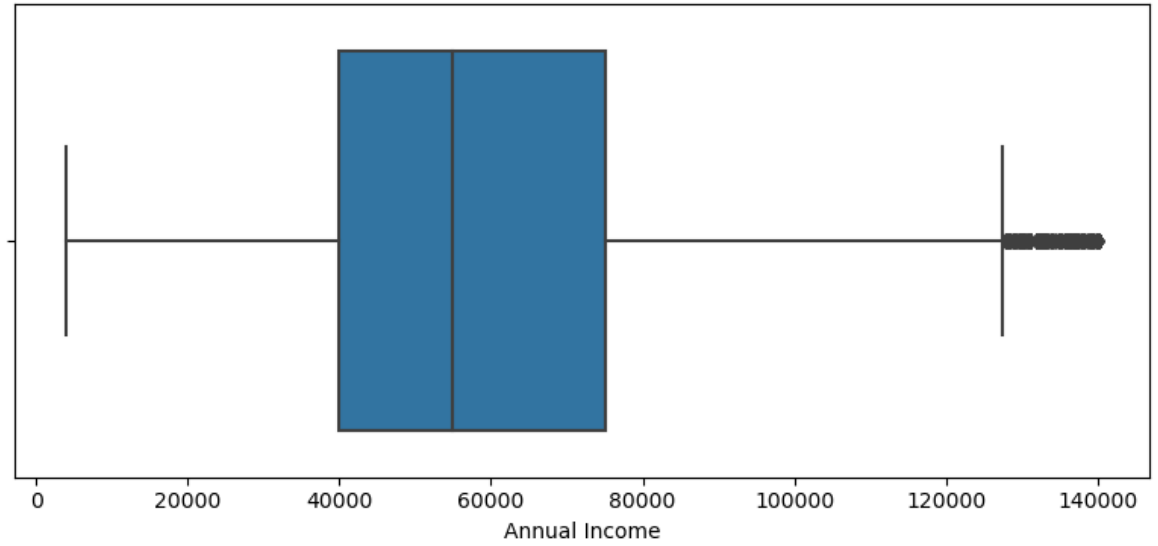


Annual Income

Observations:

- Most of the annual income lies in the range of 40K to 75K
- Average annual income is 59883.0

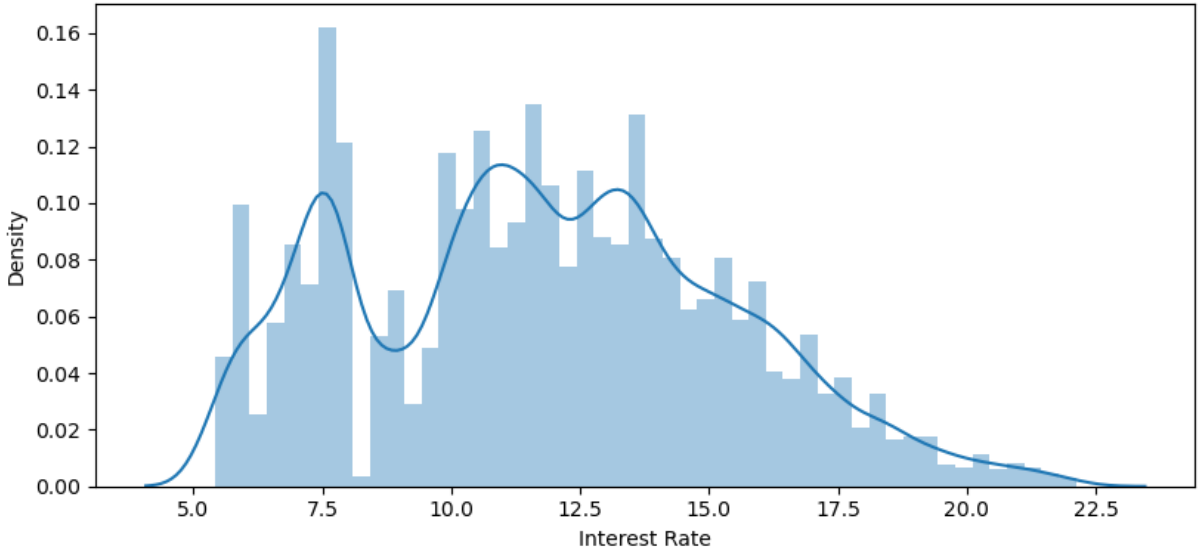


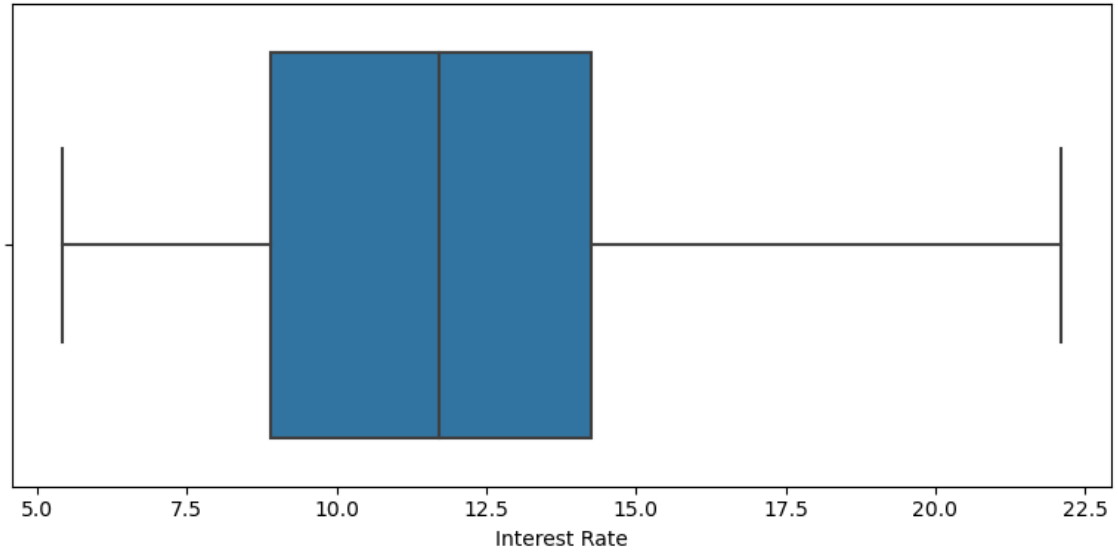


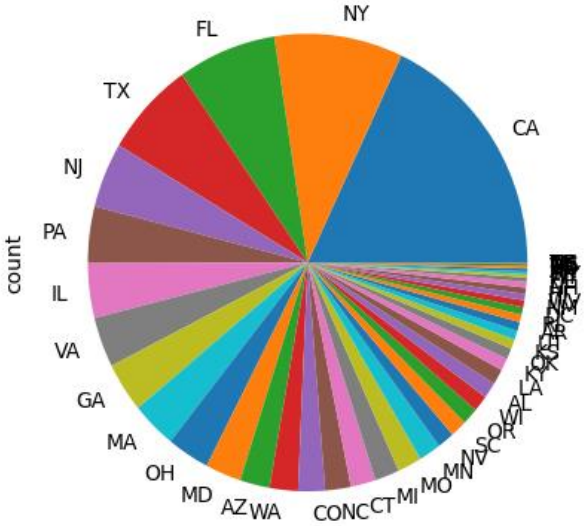
Interest Rate

Observations:

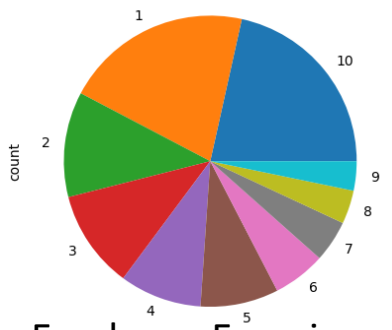
- Interest rates lie in the range of 8.9% to 14.26%
- Average interest rate is 11.78%
- Maximum interest rate is 22.11%



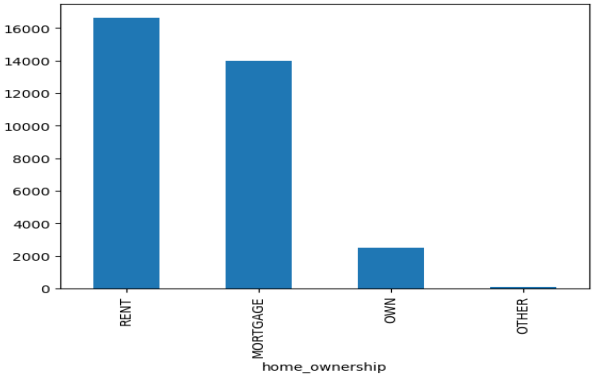


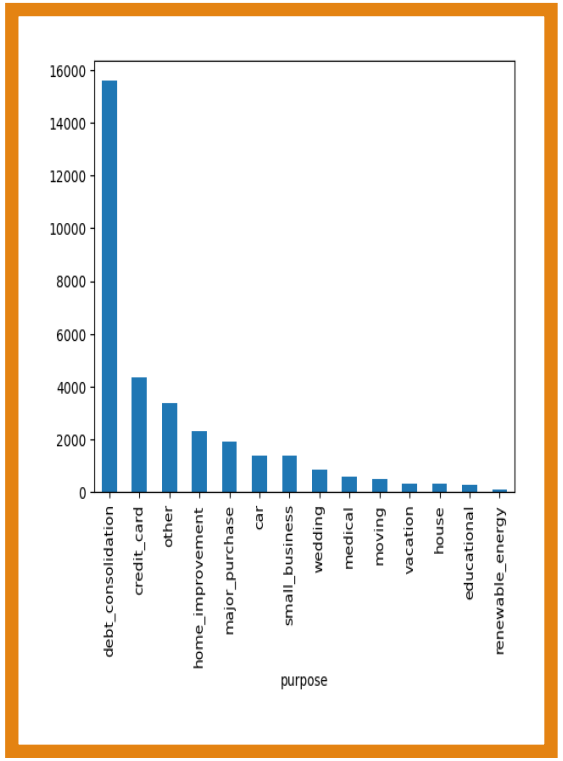


Region



Employee Experience





Univariate Analysis

Observations:

- Most of the loan applicants are from ‘CA’ region
- Most of the applicants have over 10 years of employment.
- Majority of applicants are living on rent or mortgage
- Most of the applicants apply for loan for debt consolidation

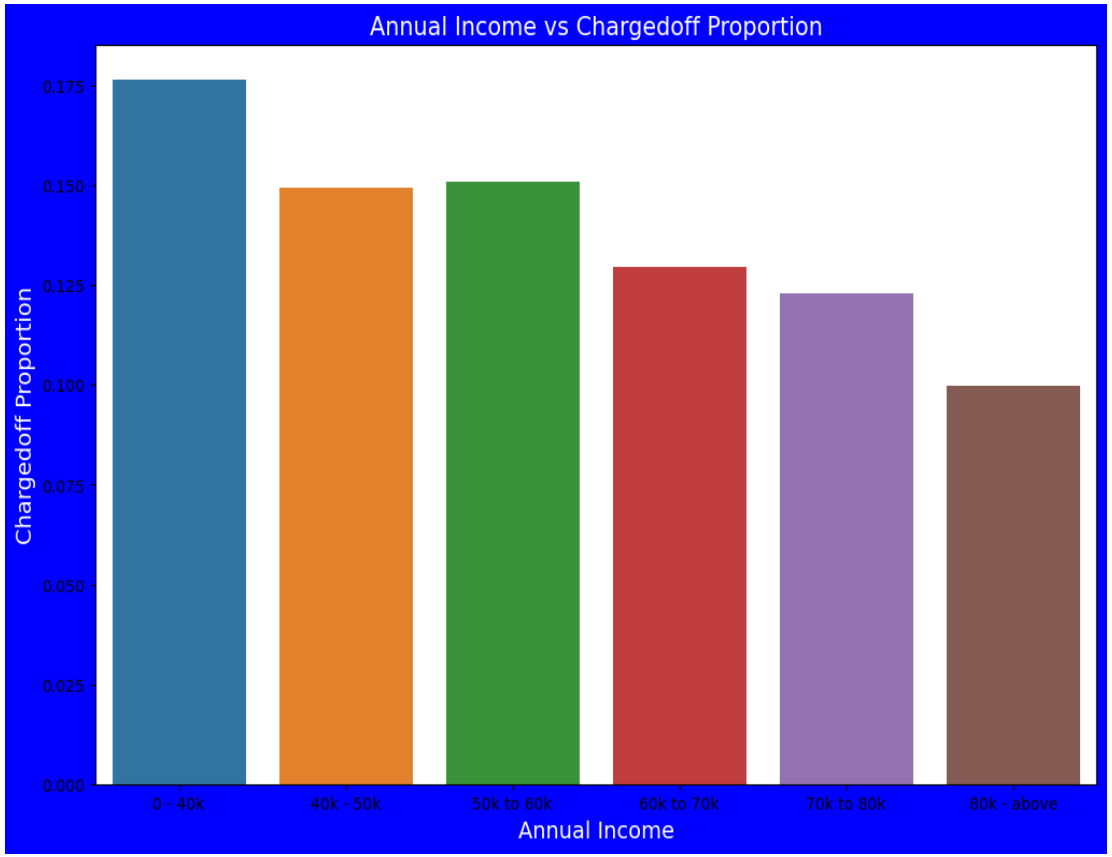
Bivariate analysis

Annual Income vs Charge Off

Influence of annual income on charge-off:

- Applicants with income range of over 80K are less likely to be loan defaulters
- Applicants in the income range of 0 to 40K have higher probability of loan defaults
- It is noticed that with increase in annual income, the charge off proportion reduces.

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

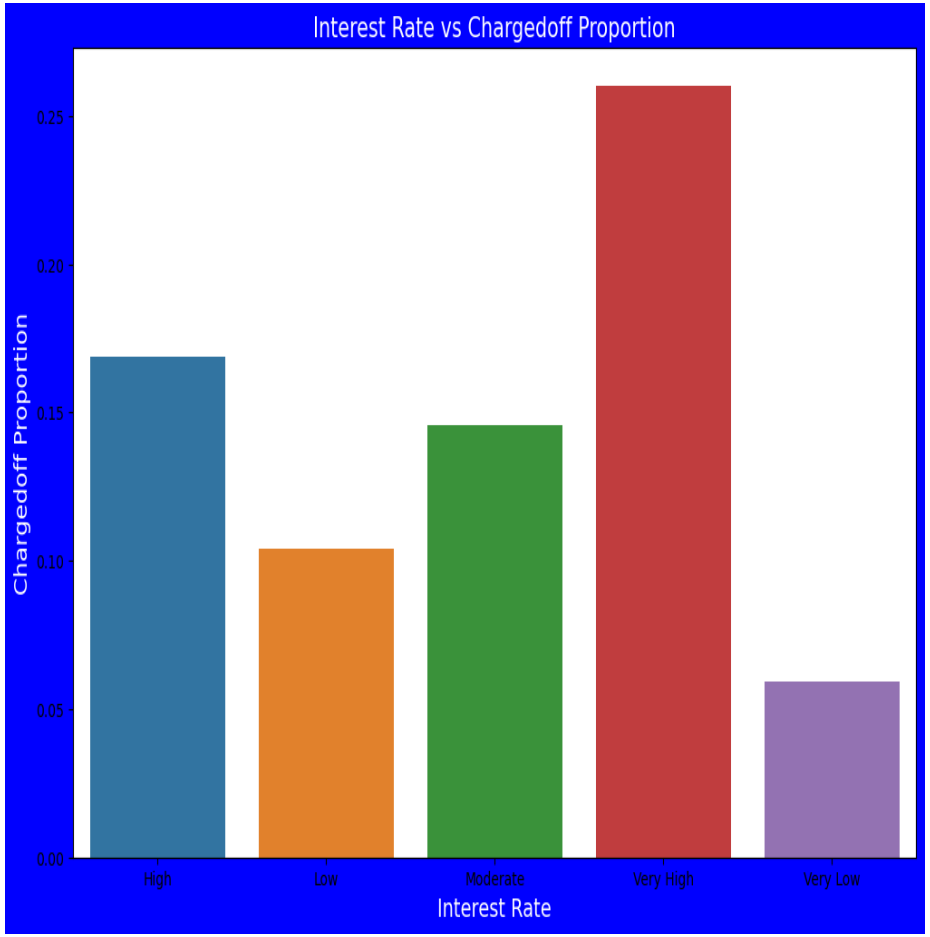


Interest Rate vs Charge Off

Influence of Interest-Rate on charge-off:

- Loans with very high interest rate of above 15% are most likely to be defaulted
- Charge-offs increase with increase in interest rate

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

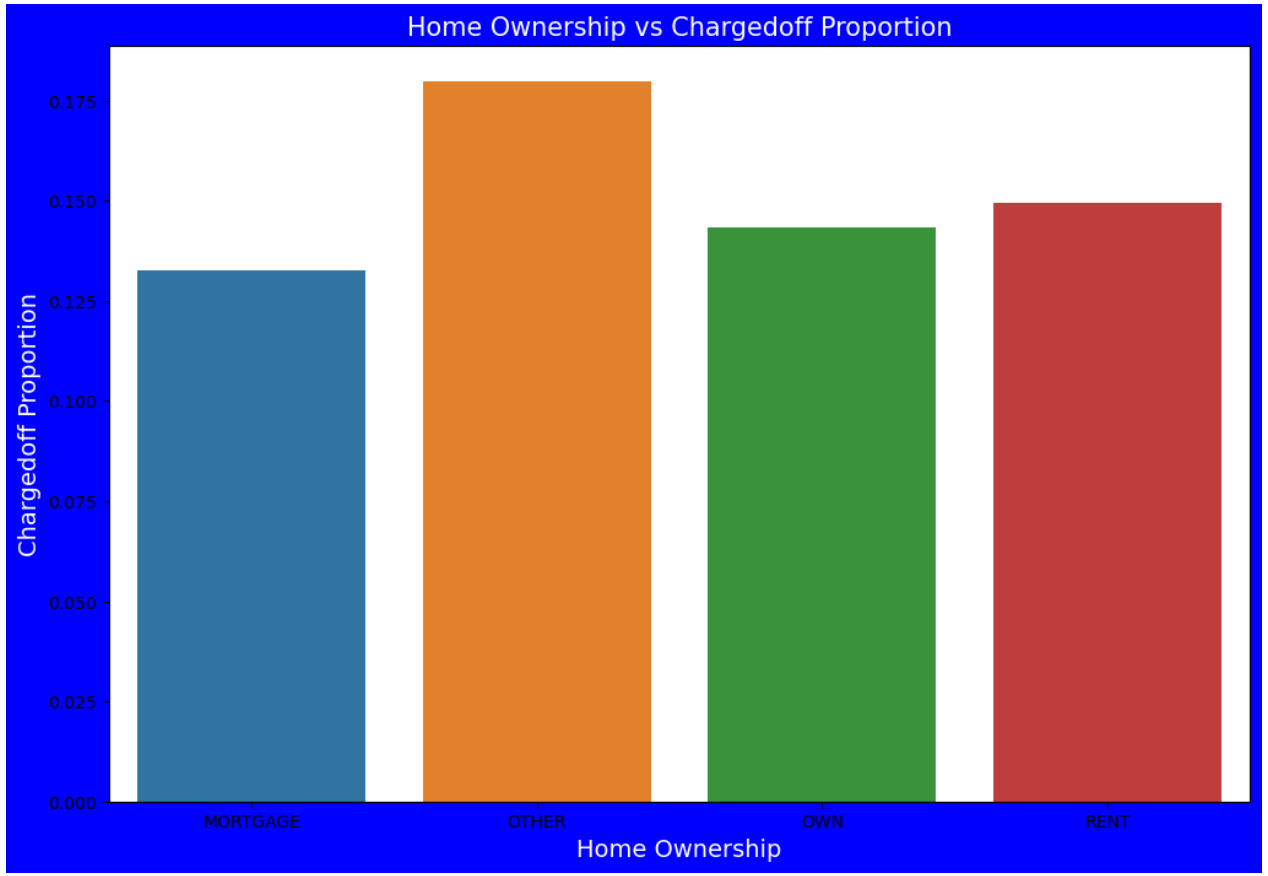


Home Ownership vs Charge off

Influence of Home Ownership on charge-off:

- With the limited data available, the graph shows loan applicants with mortgaged and own house are less likely to default loans

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

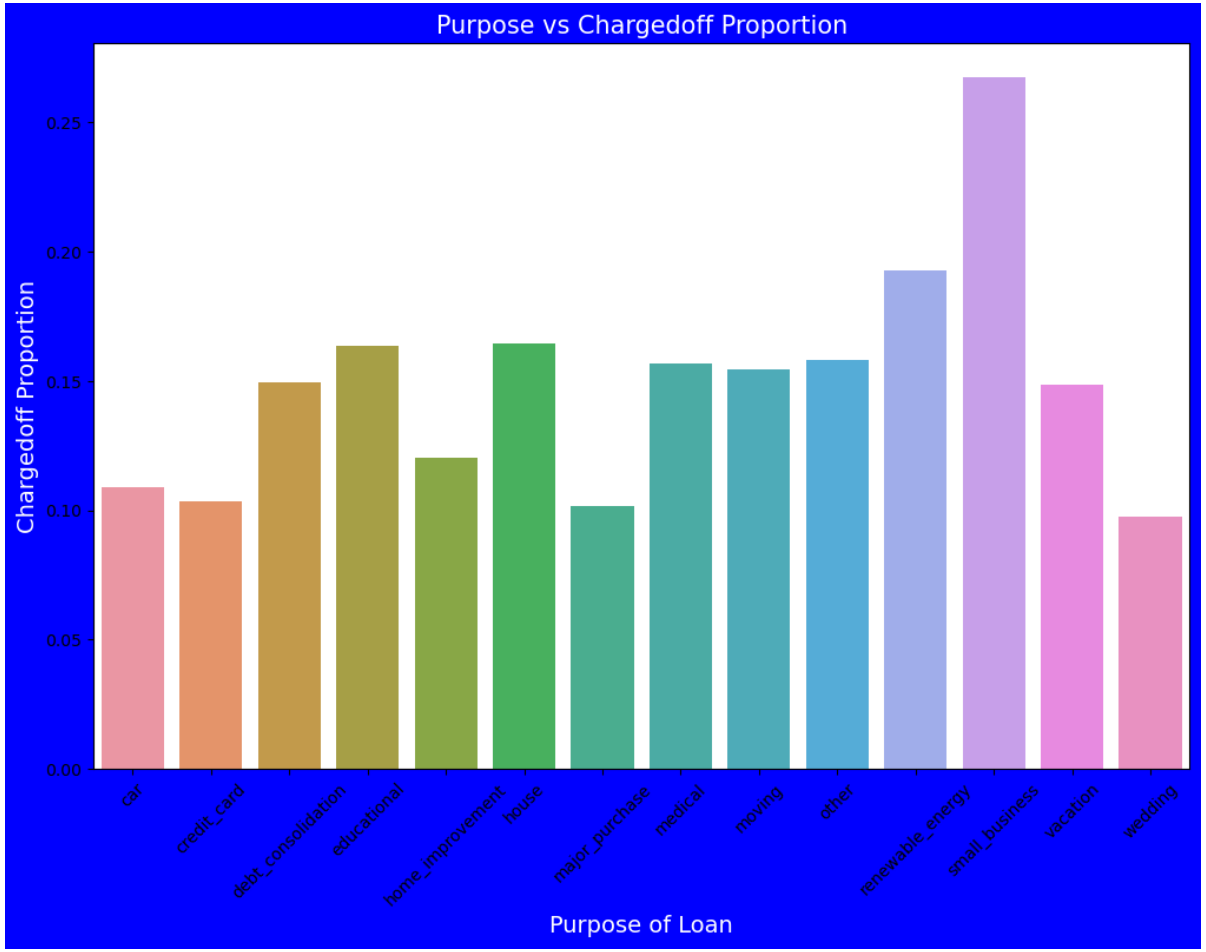


Purpose of Loan vs Charge Off

Influence of loan purpose on charge off

- Loans for Small business are most likely to be defaulted
- Loans for weddings and major purchases are least likely to be defaulted

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

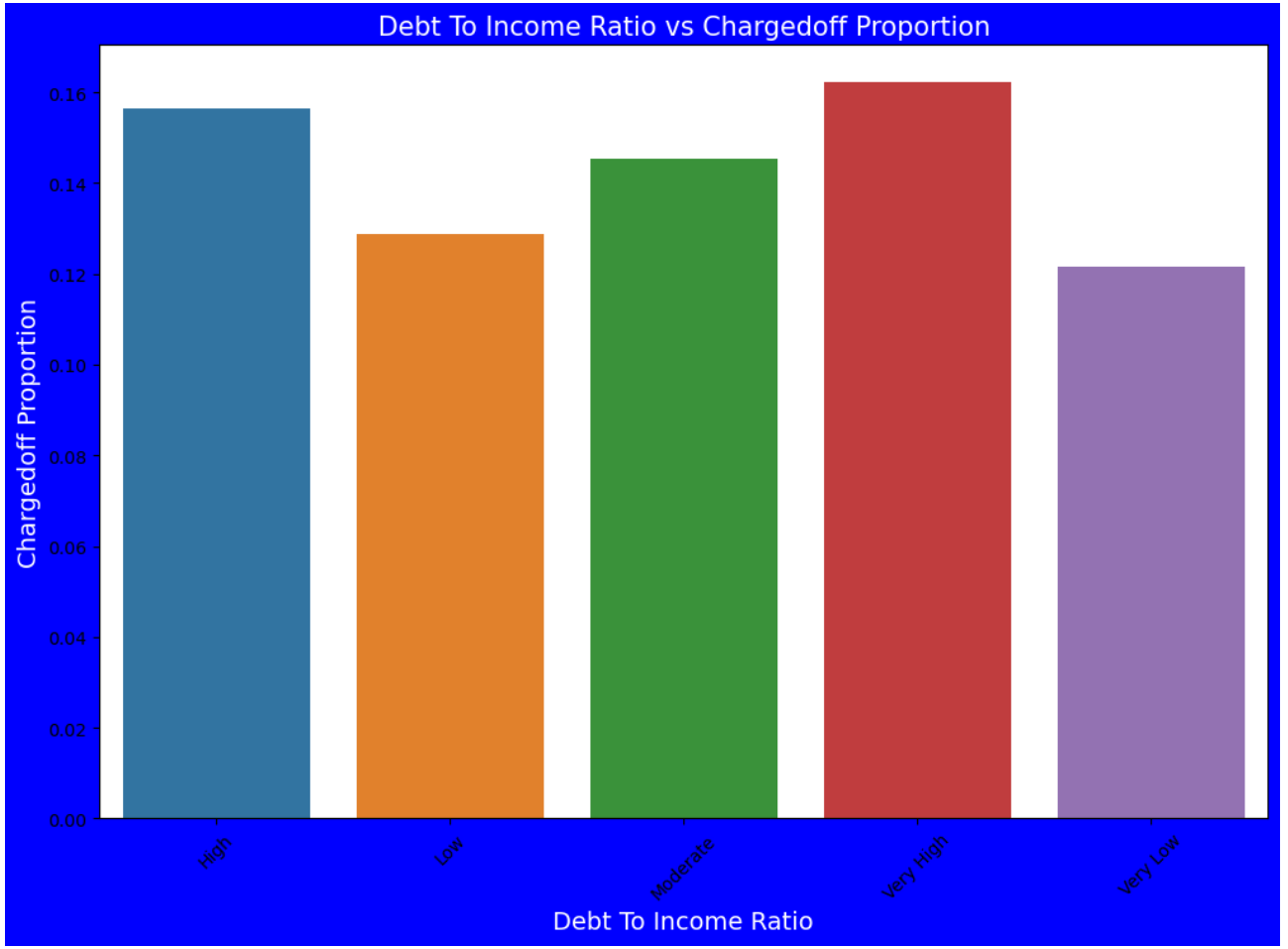


Debt to Income Ratio vs Charge Off

Influence of Debt-to-Income ratio on Charge off

- Applicants with very high DTI have higher chances of loan default
- As the DTI ratio decreases the chances of loan default also reduces

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

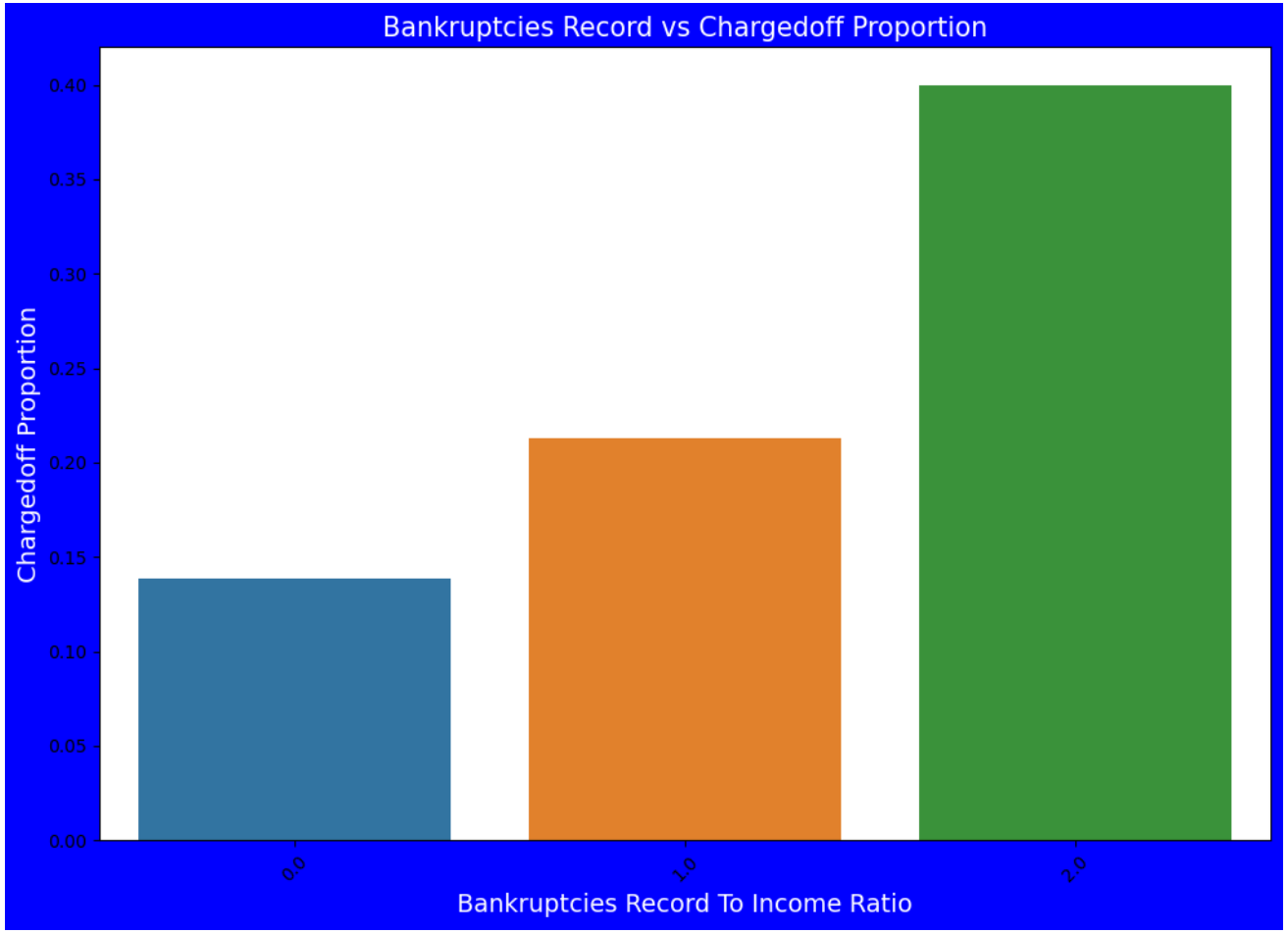


Bankruptcies record vs Charge Off

Influence of Bankruptcies record on Charge off

- Bankruptcies record with 0 have low impact on loan defaults
- Bankruptcies record with 2 have high impact on loan defaults
- Lower the bankruptcies lower the risk

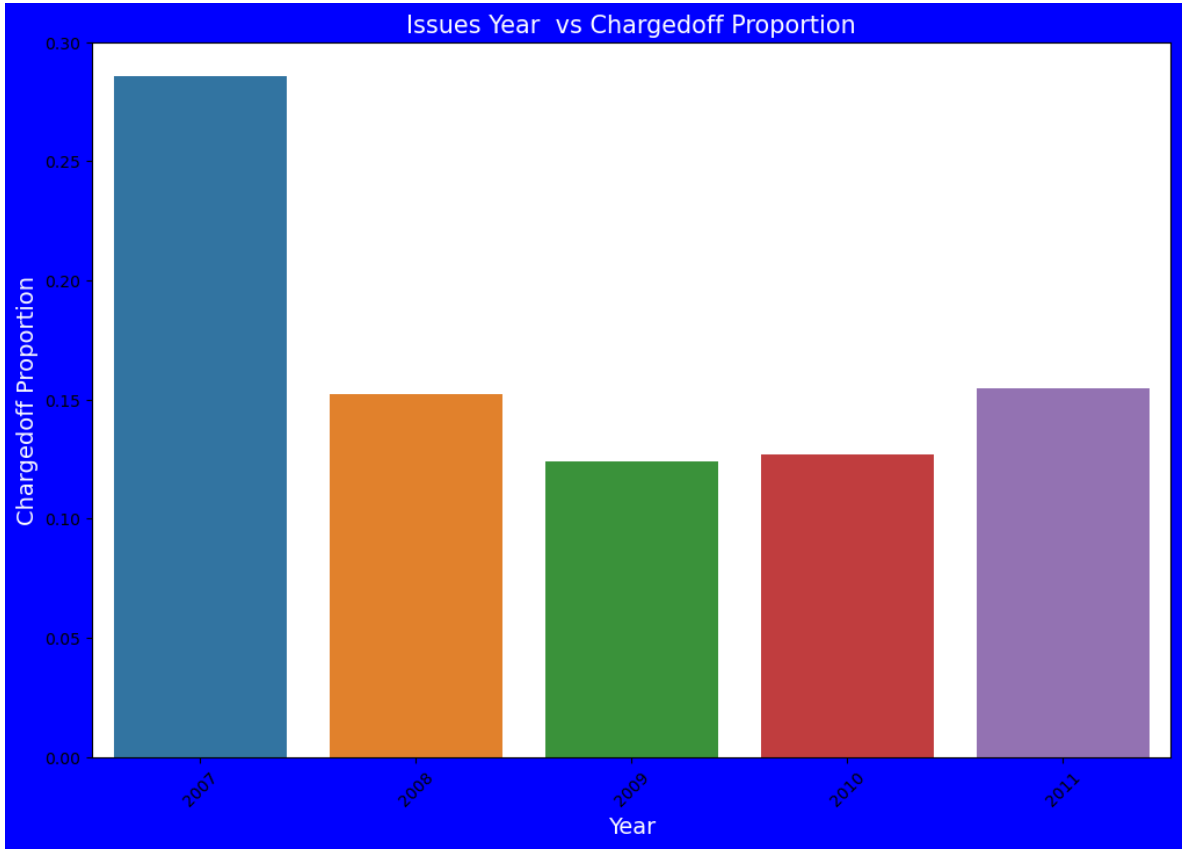
loan_status	pub_rec_bankruptcies	Charged Off	Fully Paid	Total	Chargedoff_Proportion
2	2.0	2	3	5	0.400000
1	1.0	308	1137	1445	0.213149
0	0.0	4404	27337	31741	0.138748



Issue year vs Charge Off

Influence of Issue Year on Charge off

- Majority of the loan defaults are for the loans issued in the year 2007
- Least number of loan defaults are from the loans issued in 2009

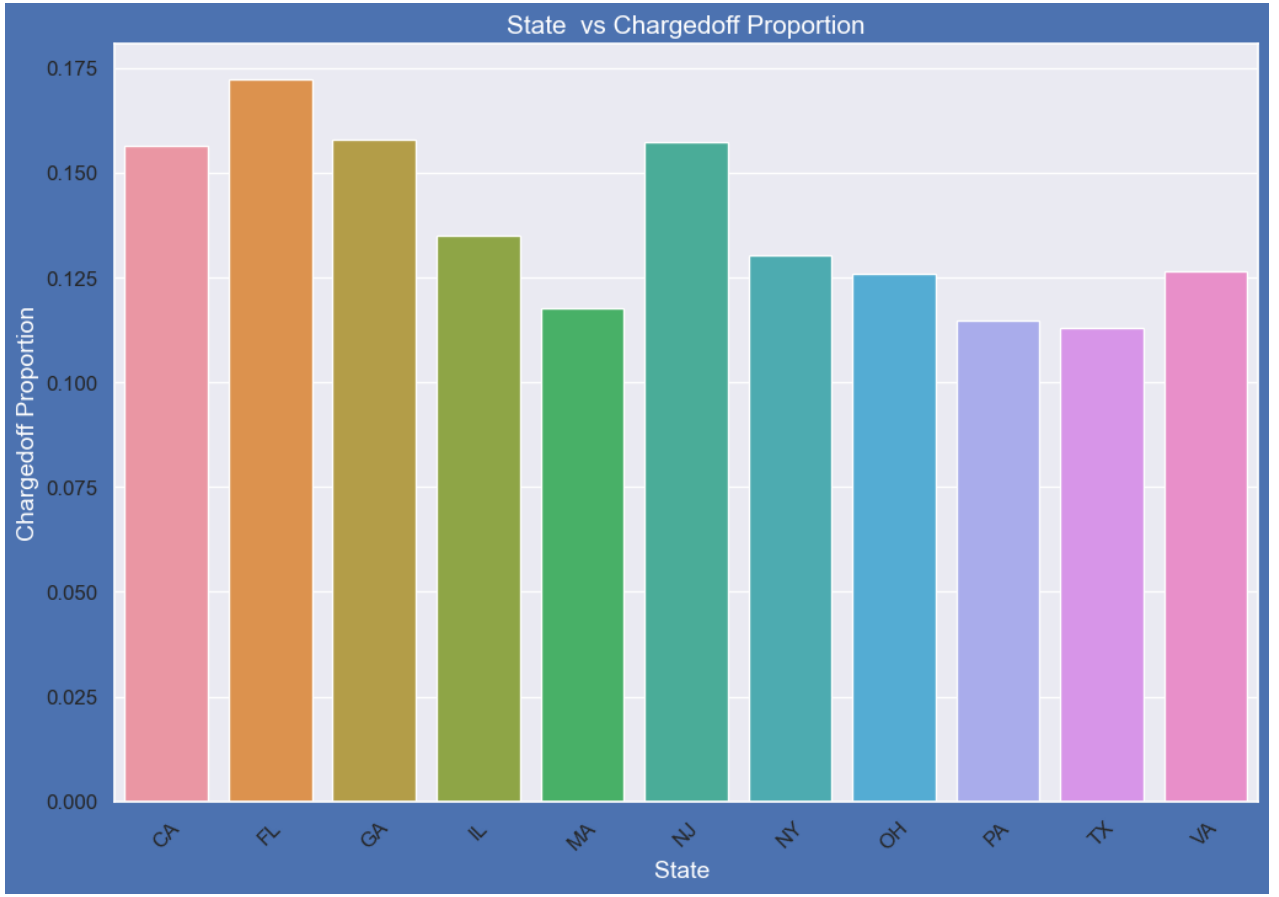


State vs Charge Off

Influence of State on Charge off

Among the states where there are more than 1000 loan applications,

- Highest loan defaults are from the state FL
- Lowest loan defaults are from the state TX

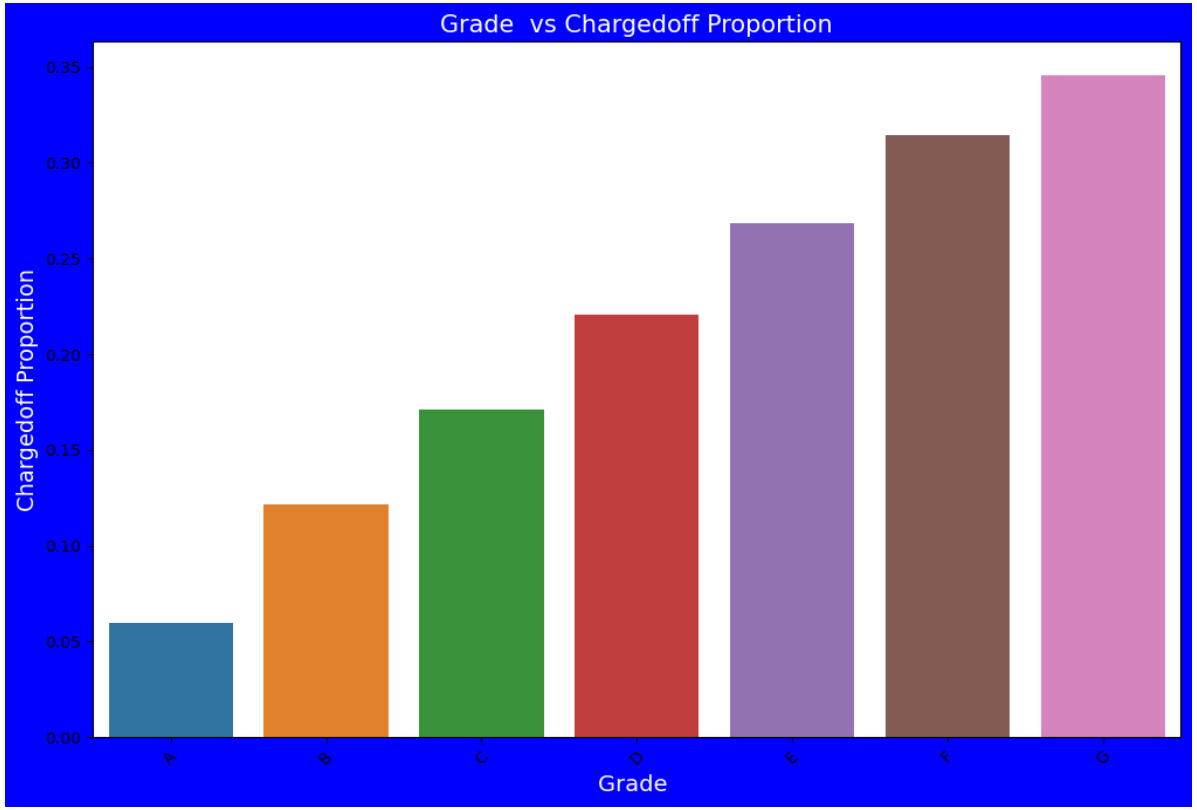


Grade vs Charge Off

Influence of State on Charge off

- Highest loan defaults are with the grade G
- Lowest loan defaults are with the grade A

loan_status	grade	Charged Off	Fully Paid	Total	Chargedoff_Proportion
6	G	55	104	159	0.345912
5	F	220	480	700	0.314286
4	E	557	1518	2075	0.268434
3	D	941	3329	4270	0.220375
2	C	1177	5702	6879	0.171100
1	B	1225	8857	10082	0.121504
0	A	539	8487	9026	0.059716



Correlation Analysis

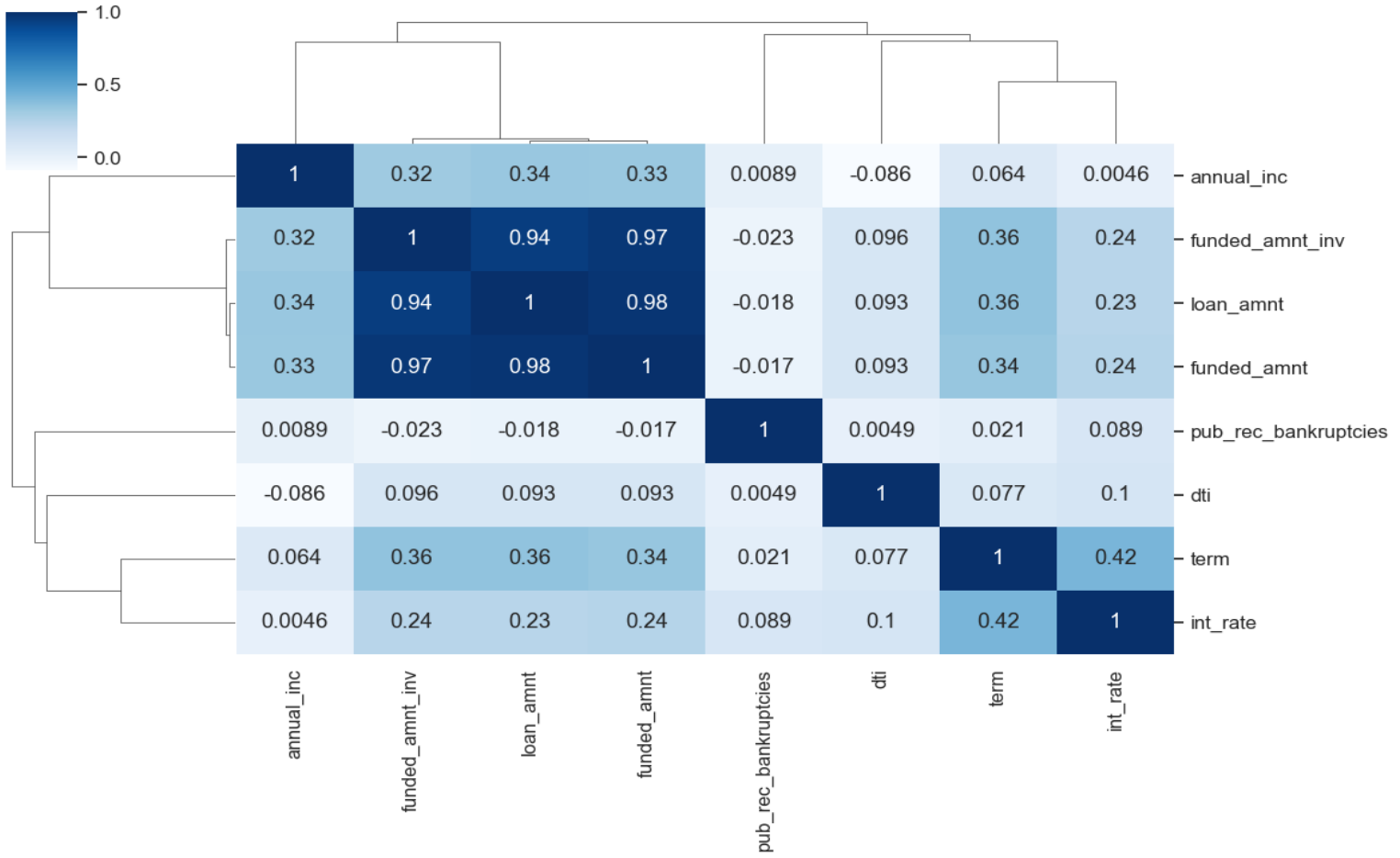
Correlations

Strong Correlation:

- Term has a strong correlation with interest rate and loan amount
- Annual income has strong correlation with funded amount

Negative Correlation:

- DTI have negative correlation with annual income
- Funded loan amount is negatively correlated with bankruptcies record published



Conclusion

- Applicants with annual income range 0 to 40K have high probability of charge off
- High Interest Rate of greater than 15% have high chances of charge off
- Applicants who do not own a home have high chances of defaulting the loan
- Applicants who apply loan for small business have high probability of charge off
- Applicants from the state FL are more likely to be loan defaulters
- Applicants with higher bankruptcies record have higher chance of charge off
- Applicants with higher DTI have higher chances of loan default
- Applicants with loan grade ‘G’ are more likely to be defaulted