LENDING CLUB CASE STUDY

EXPLORATORY DATA ANALYSIS

BY ANISH DHONDI

CONTENTS

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

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:

- **Risk 1:** If the applicant is likely to repay the loan, then not approving the loan results in a **loss of business** to the company.
- **Risk 2:** If the applicant is not likely to repay the loan (i.e., likely to default), then approving the loan may lead to a **financial loss** for the company.

Objective:

- Use **Exploratory Data Analysis (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 made by the company:

- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios:
    - **Fully paid:** Applicant has fully paid the loan (the principal and the interest rate).
    - **Current:** Applicant is in the process of paying the installments; the loan tenure has not yet been completed. These candidates are not labeled as **'defaulted'**.
    - **Charged-off:** Applicant has not paid the installments on time for a long period of time and has defaulted on the loan.
- **Loan rejected:** The company rejected the loan because the candidate does not meet their requirements. Since the loan was rejected, there is **no transactional history** of those applicants with the company, and thus, this data is **not available** in this dataset.

DATA SUMMARY

- **Dataset**: Loan.csv
- **Total Rows**: 39,717
- **Total Columns**: 111
- **Types of Attributes**:

Loan Attributes: Information related to the loan itself (e.g., loan amount, interest rate, term, etc.)

Customer Attributes: Information about the customer (e.g., annual income, credit score, home ownership status, etc.)

DATA CLEANING

Initial Dataset:

Rows: 39,717

Columns: 111

Steps Taken:

1. Rows with loan_status = 'current':

- 1,140 rows removed as they don't participate in the analysis.

2. Columns with All Null Values:

- 55 columns removed that had only null/blank values.

3. Unique Columns Removed:

- **'url'** and **'member_id'** were unique in nature and removed.
- **'desc'** and **'title'** were text/description fields and irrelevant to analysis.

4. Behavioral Data Columns:

- 21 columns related to behavioral data removed based on domain knowledge, as they are unavailable during loan approval.

5. Columns with Constant Values:

- 8 columns where the value was always '1', indicating uniqueness, were dropped.

6. High Percentage of Missing Data:

- 2 columns with more than 50% missing data were removed.

Final Dataset:

Remaining Rows: 38,577

Remaining Columns: 20

DATA CONVERSIONS VS DERIVED COLUMNS

Data Conversions:

- **‘term’**:
  - Trimmed additional string values and converted to **int** data type.
- **‘int_rate’**:
  - Converted from **string** to **int** by trimming the '%' symbol.
- **‘loan_funded_amnt’** and **‘funded_amnt’**:
  - Converted to **float** data type.
- **‘loan_amnt’, ‘funded_amnt’, ‘funded_amnt_inv’, ‘int_rate’, ‘dti’**:
  - Rounded to **two decimal points** for consistency.
- **‘issue_d’**:
  - Converted to a **datetime** data type for better analysis.

Derived Columns:

- **‘issue_year’ and ‘issue_month’**:
  - Extracted from **‘issue_d’** for further analysis.
- **‘loan_amnt_b’, ‘annual_inc_b’, ‘int_rate_b’, ‘dti_b’**:
  - Created **bucketed columns** from continuous data for more effective analysis.

DROPPING/INPUTING THE ROWS

Dropping Rows:

1. Loan Status = ‘Current’:

- Removed **1140 rows** where loan_status = 'current', as they do not participate in the analysis.

2. Columns with Null/Blank Values:

- Removed **55 columns** where all rows contained **null or blank values**, as they didn’t contribute to the analysis.

3. Unique Columns (‘url’ and ‘member_id’):

- Removed **‘url’** and **‘member_id’** since they are unique in nature and don’t contribute to the analysis.

4. Text/Description Columns (‘desc’ and ‘title’):

- Dropped **‘desc’** and **‘title’** as they contain text descriptions and do not participate in the analysis.

5. Behavioral Data Columns:

- Removed **21 behavioral data columns** as they were not relevant to loan approval or analysis.

6. Columns with Constant Values:

- Dropped **8 columns** that had the same value (1) for all rows, indicating uniqueness and lack of variance.

7. Columns with >50% Missing Data:

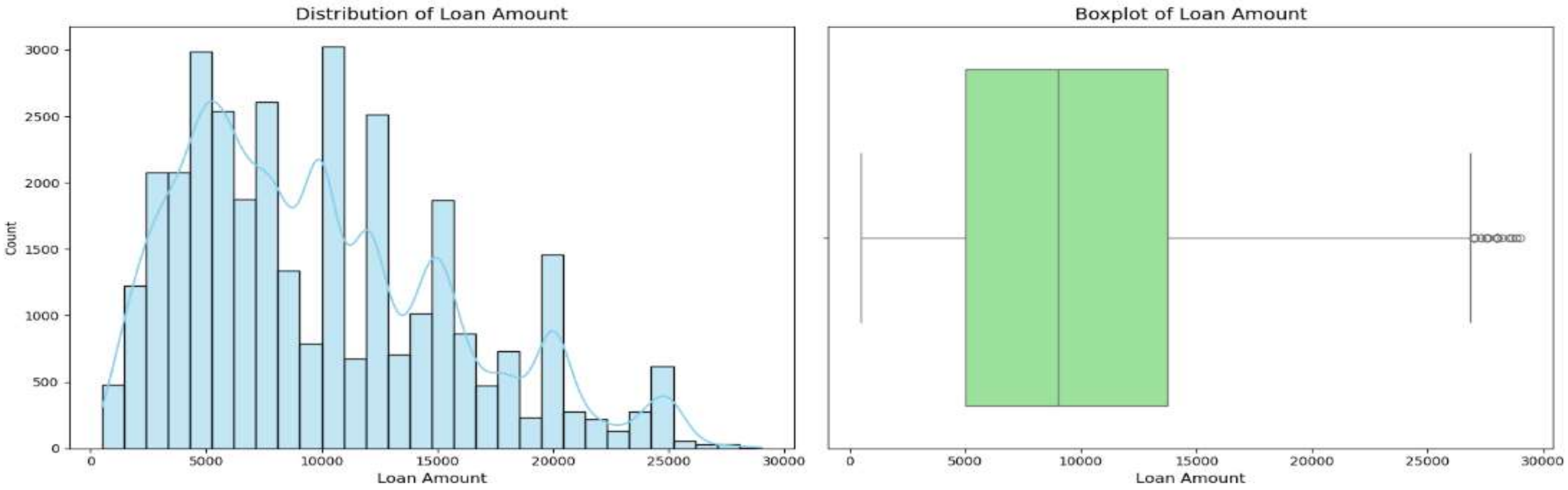
- Removed **2 columns** with over **50% missing values** as they would not provide meaningful insights.

Inputing Rows:

- No row imputations were required as all columns were either dropped or kept after handling null values through the cleaning process.

UNIVARIATE ANALYSIS

LOAN AMOUNT

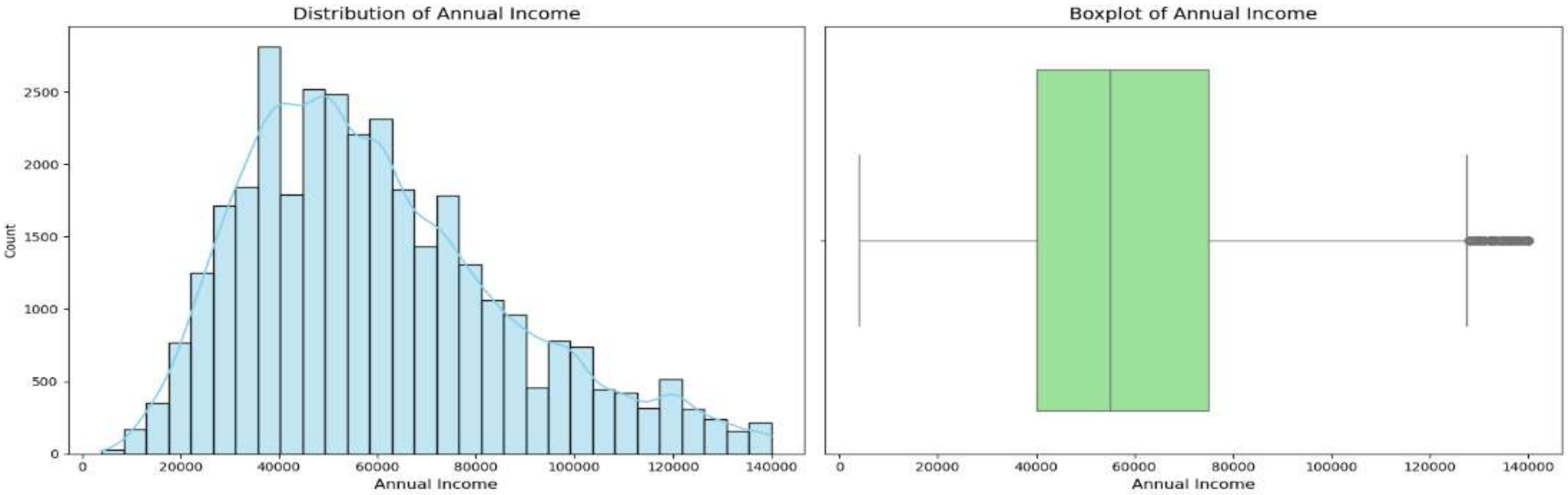


Observation:

1. Most loan applications have loan amounts in the range of 5k to 14k, which indicates that the majority of borrowers prefer smaller loan amounts within this bracket.

2. The maximum loan amount applied for is around 29k, which stands out as an outlier compared to the general distribution of loan amounts.

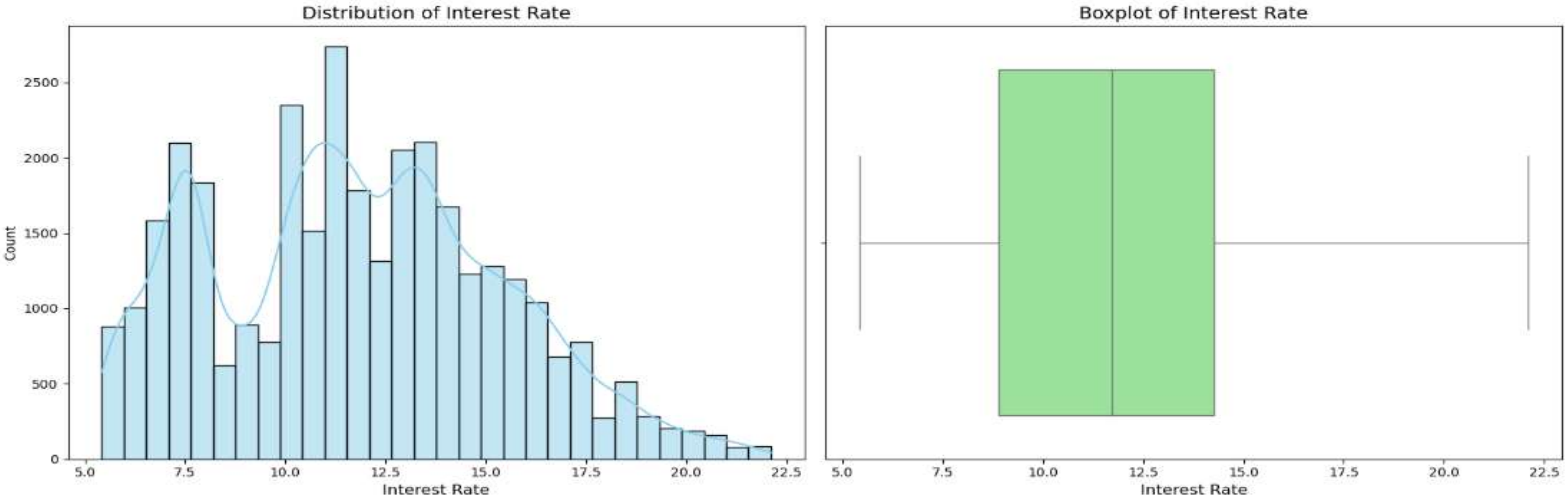
ANNUAL INCOME



Observation:

1. The majority of applicants' annual incomes fall between 40k and 75k, as seen from the distribution.
2. The average annual income across all applicants is approximately $59,883, indicating a central tendency towards a middle-income range.
3. A larger proportion of applicants have incomes close to or slightly below the average, with a few outliers on both lower and higher ends of the spectrum.

INTEREST RATE

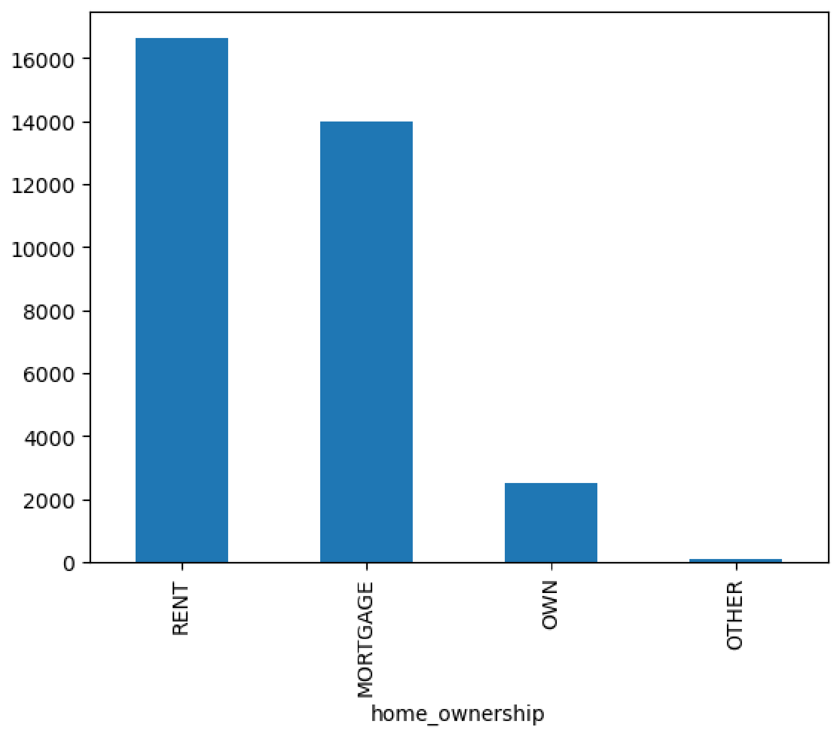


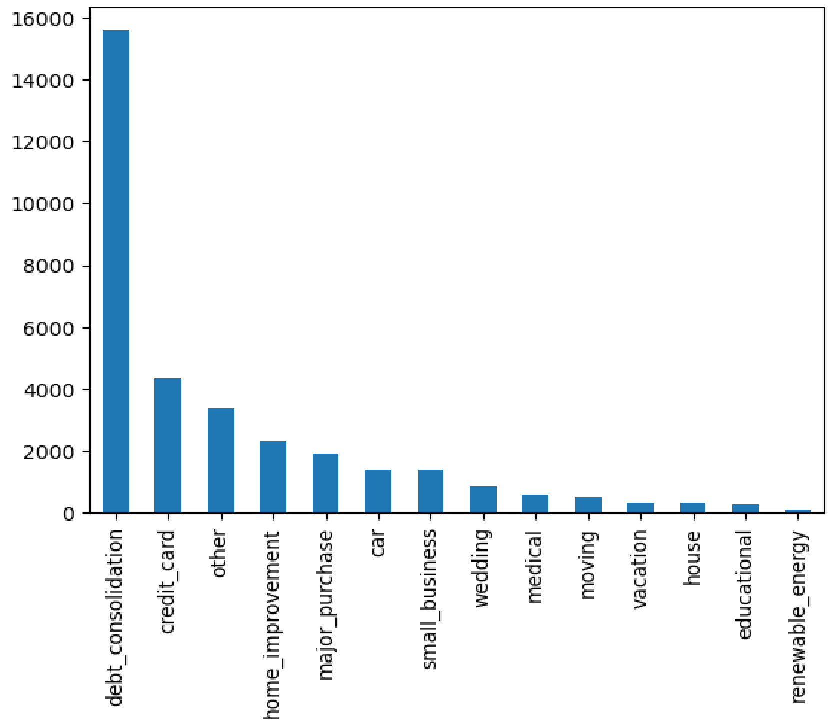
Observation:

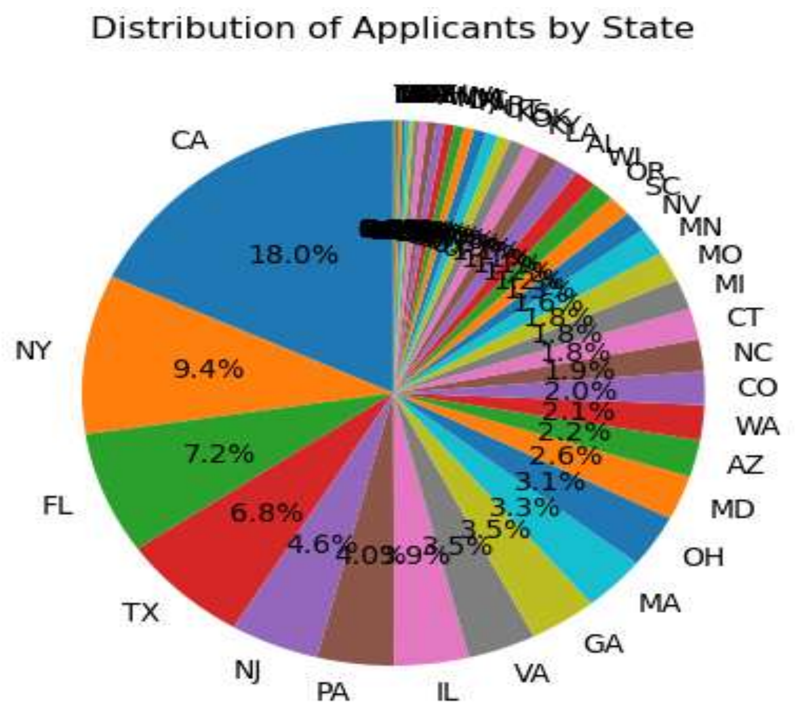
1. The majority of interest rates fall between 8.9% and 14.26%, with a mean interest rate of approximately 11.78%.
2. The minimum interest rate is 5.42%, while the highest interest rate recorded is 22.11%.

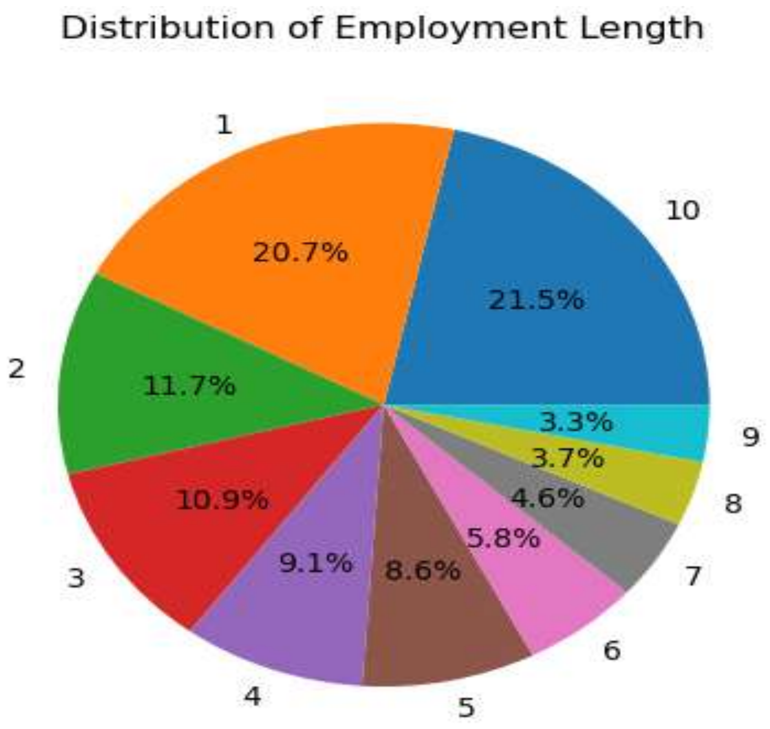
UNIVARIENTS ANALYSIS

Unordered & Ordered Categorical Variable Analysis









Observations:

Housing Status:

- Majority of loan applicants are either living **on Rent** or **on Mortgage**.

Loan Purpose:

- A significant portion of loan applicants are applying for **debt consolidation**.

Geographic Distribution:

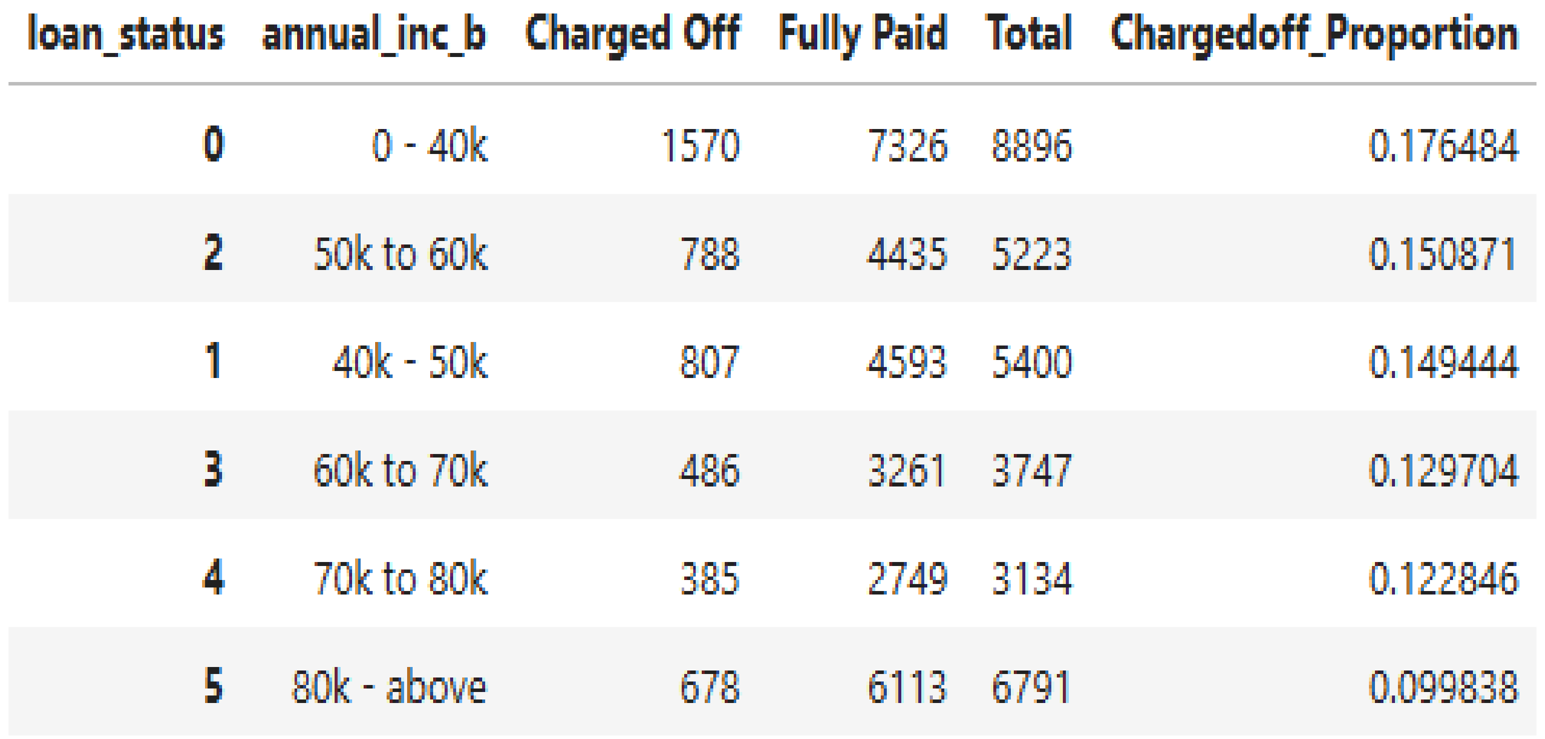
- The majority of loan applicants are from the state of **California (CA)**.

Experience Level:

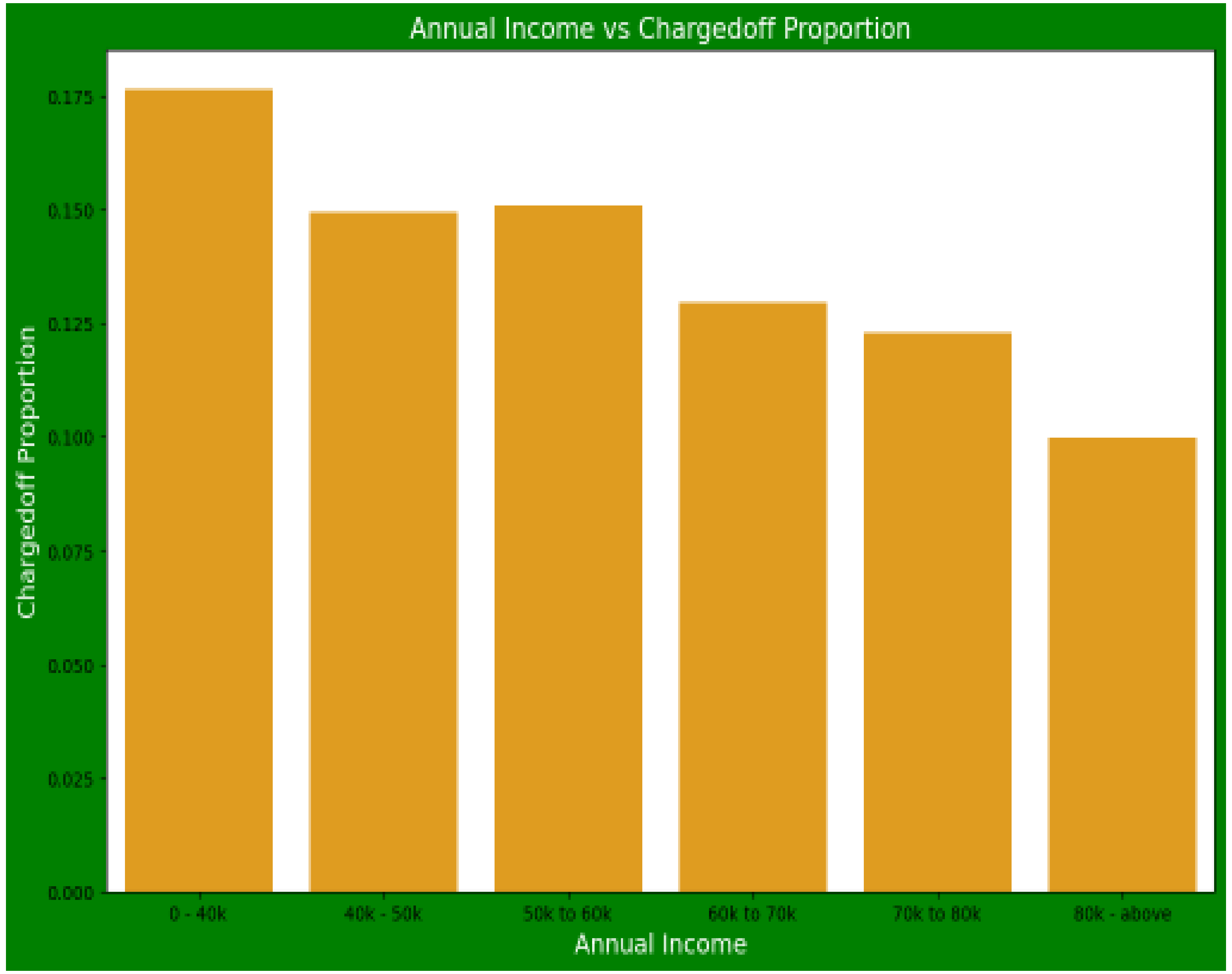
- A large number of loan applicants have **10+ years of experience**.

BIVARIATE ANALYSIS

ANNUAL INCOME VS CHARGED OFF



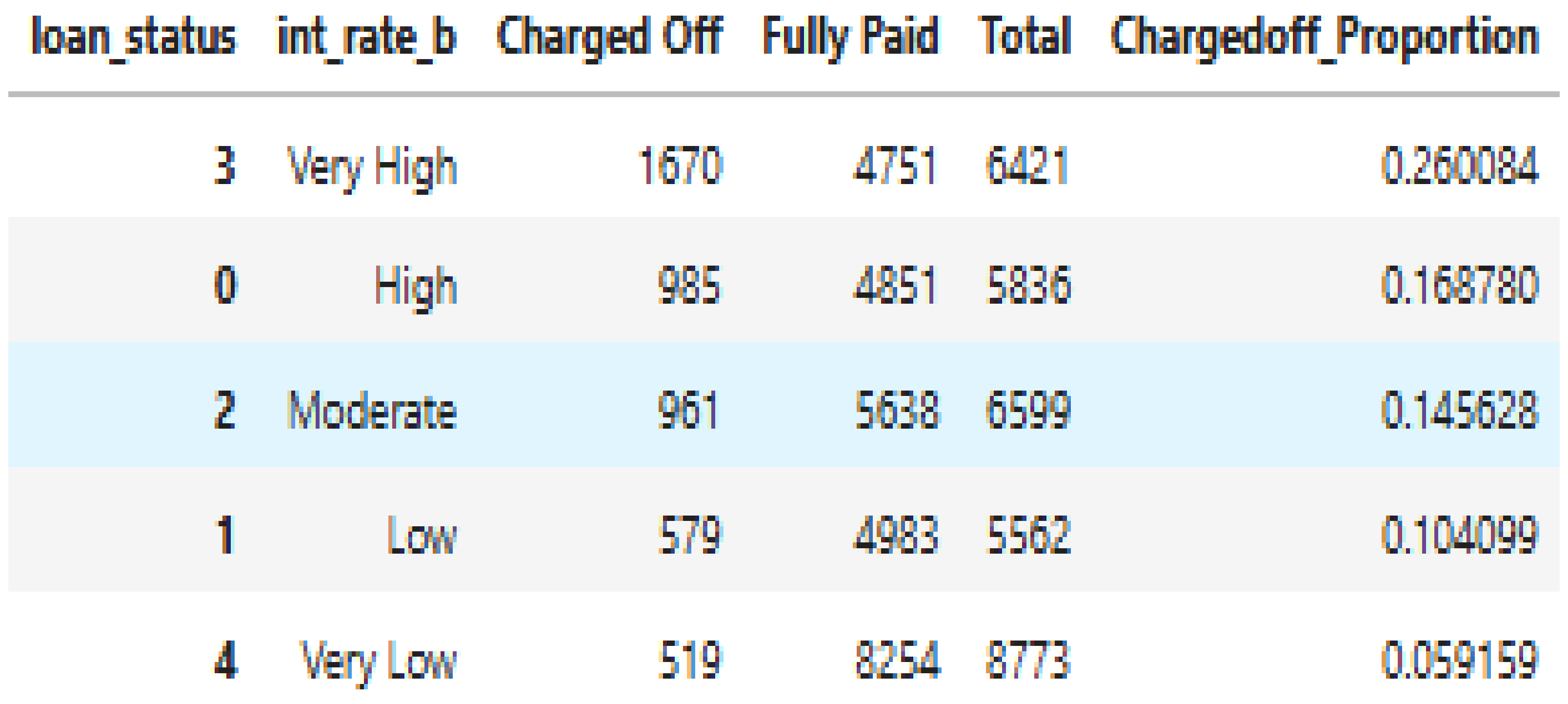
loan_status	annual_inc_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
0	0 - 40k	1570	7326	8896	0.176484
2	50k to 60k	788	4435	5223	0.150871
1	40k - 50k	807	4593	5400	0.149444
3	60k to 70k	486	3261	3747	0.129704
4	70k to 80k	385	2749	3134	0.122846
5	80k - above	678	6113	6791	0.099838



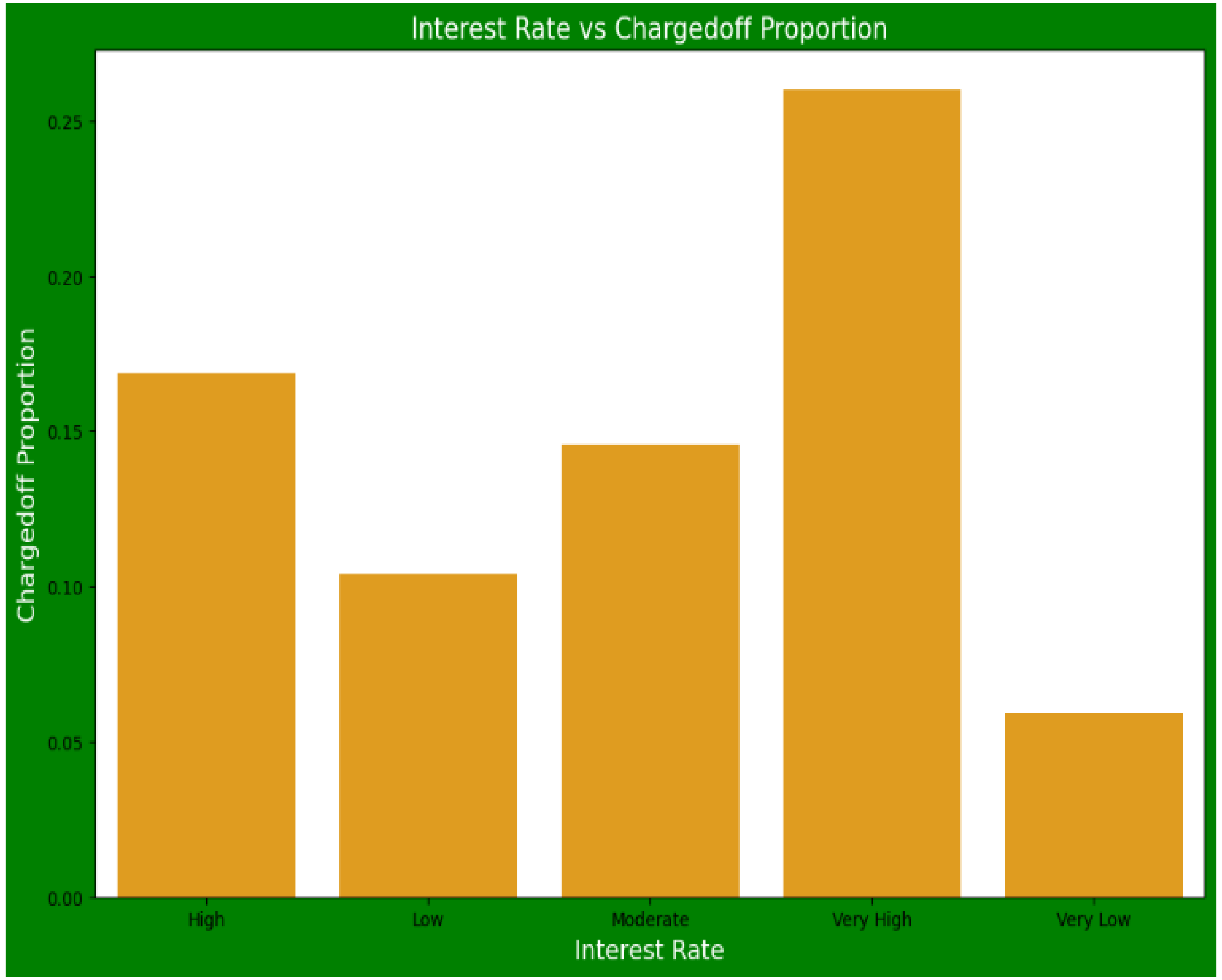
Observations:

1. The income range of 80k+ has lower chances of charge-off, indicating that higher-income individuals are less likely to default on their loans.
2. The income range of 0-40k has higher chances of charge-off, suggesting that individuals in this income bracket are more prone to default on their loans.
3. Overall, as the annual income increases, the charge-off proportion decreases, showing a negative correlation between income level and loan defaults.

INTEREST RATE VS CHARGED OFF



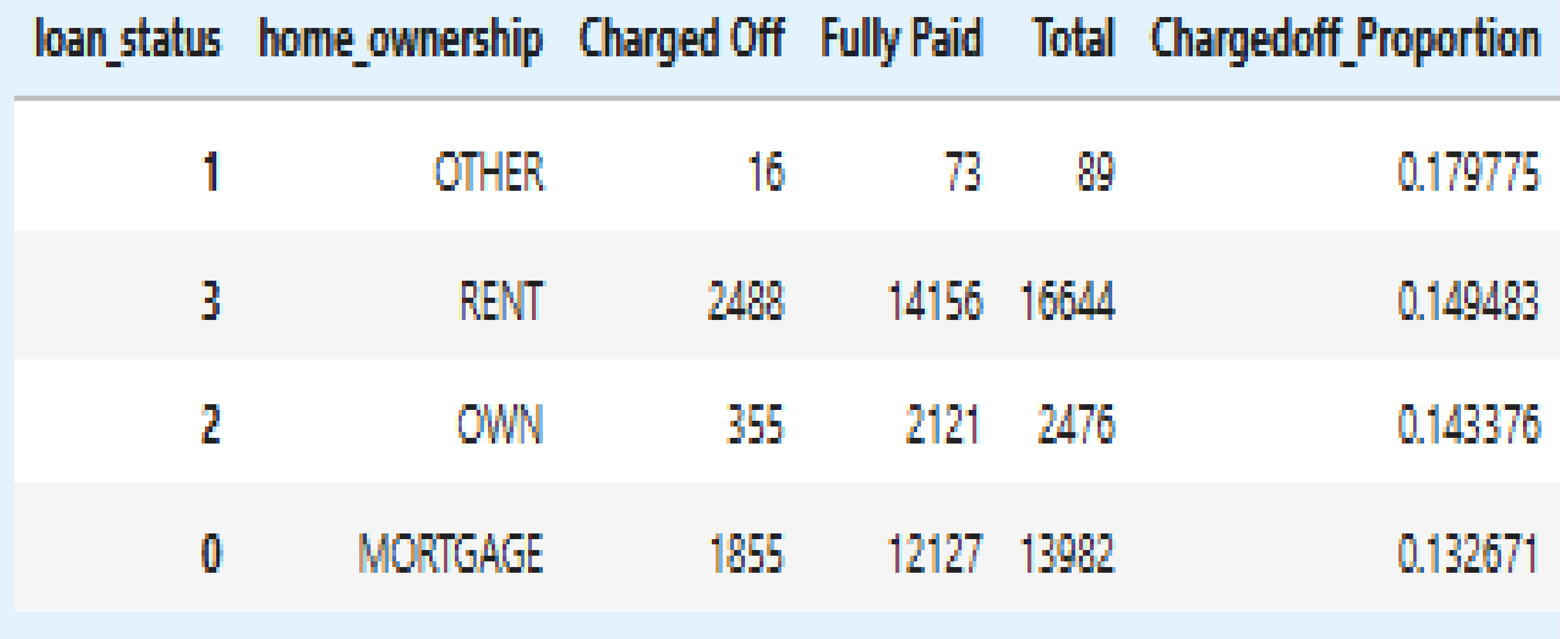
loan_status	int_rate_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1670	4751	6421	0.260084
0	High	985	4851	5836	0.168780
2	Moderate	961	5638	6599	0.145628
1	Low	579	4983	5562	0.104099
4	Very Low	519	8254	8773	0.059159



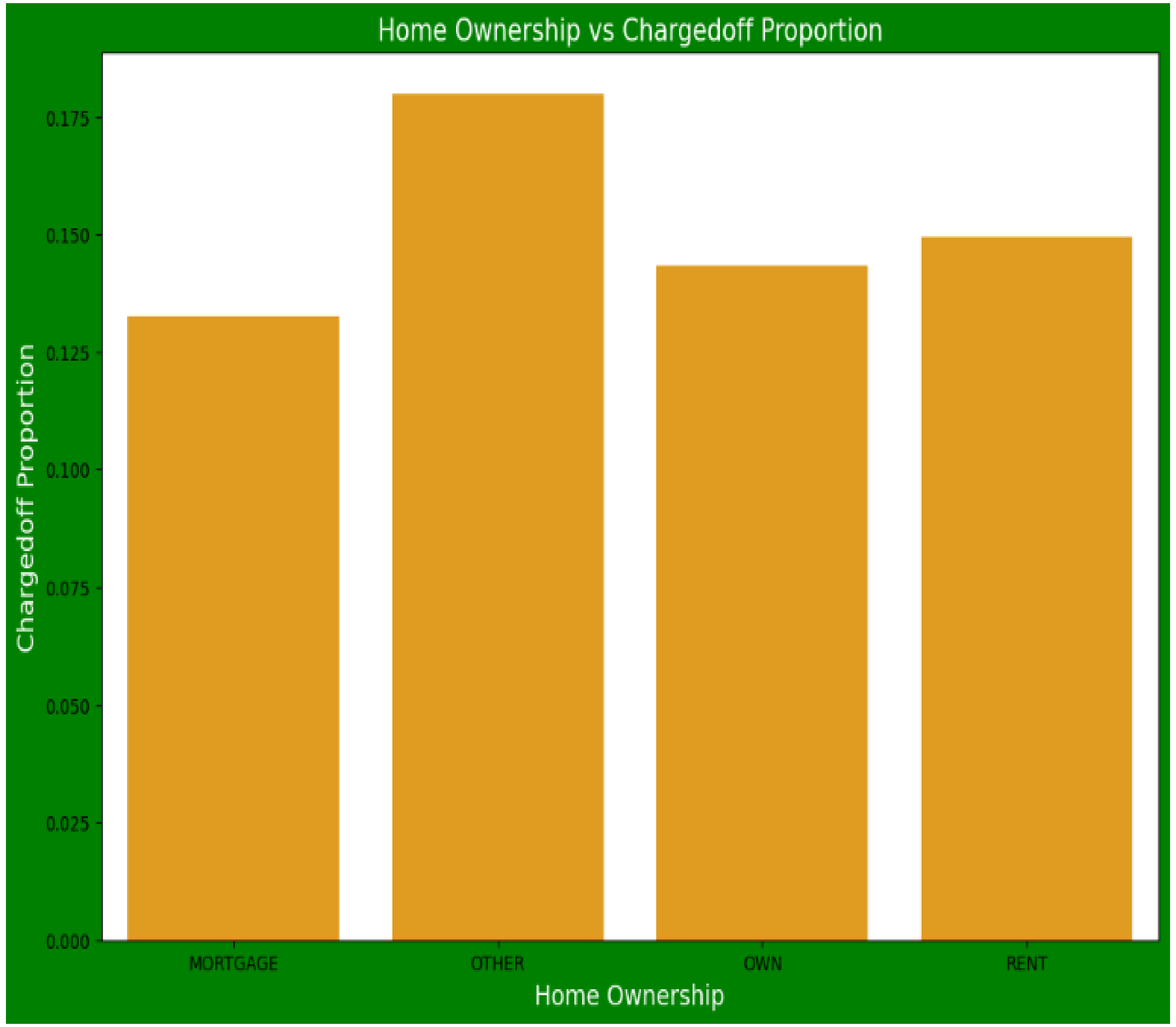
Observation:

1. Loans with an interest rate of less than 10% (very low) have a significantly lower chance of being charged off, with the interest rates starting from a minimum of 5%.
2. Loans with interest rates higher than 16% (very high) show a higher proportion of charged-off loans compared to the other categories.
3. The charged-off proportion tends to increase as the interest rate rises, indicating a correlation between higher interest rates and higher loan defaults.

HOME OWNERSHIP VS CHARGED OFF



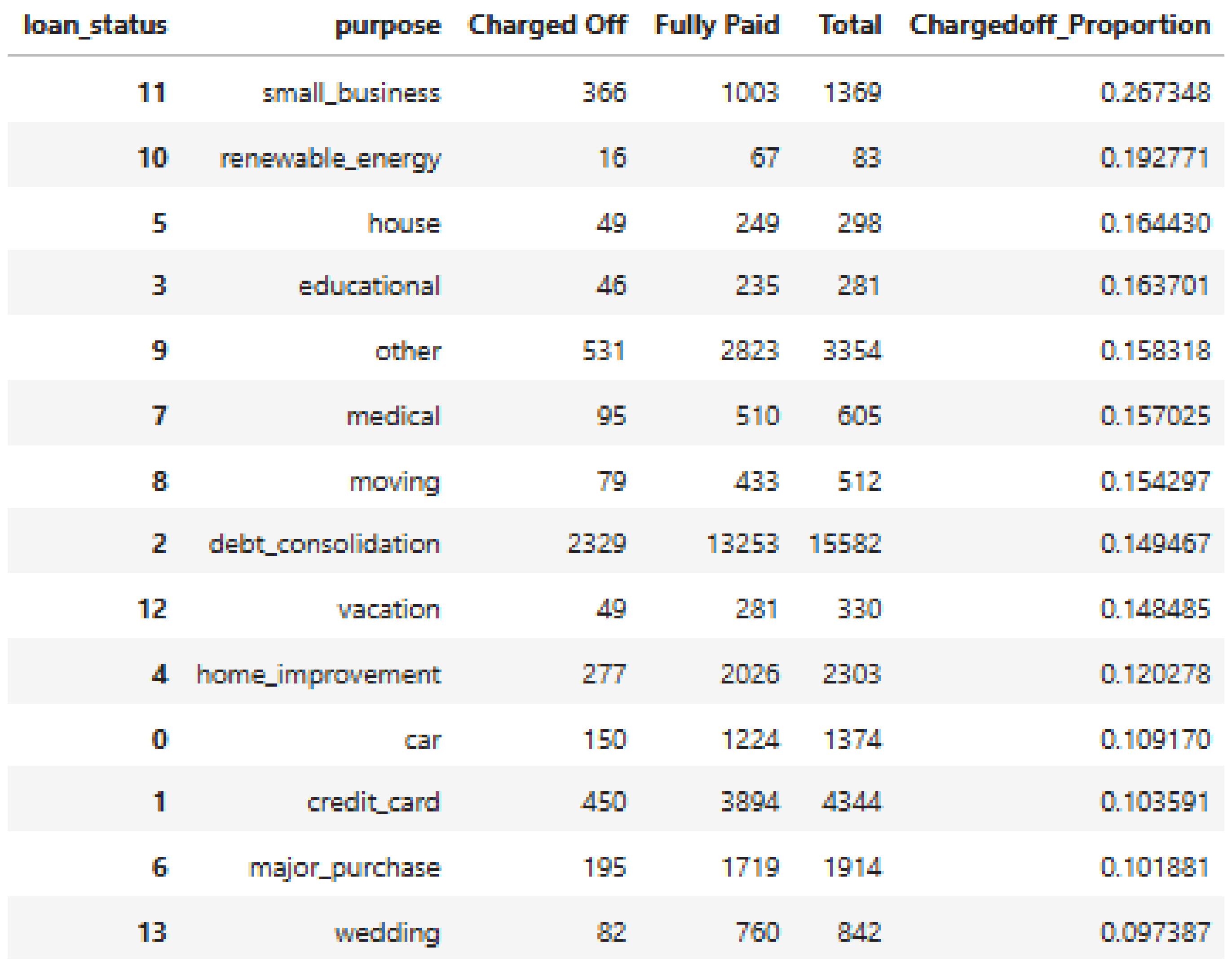
loan_status	home_ownership	Charged Off	Fully Paid	Total	Chargedoff_Proportion
1	OTHER	16	73	89	0.179775
3	RENT	2488	14156	16644	0.149483
2	OWN	355	2121	2476	0.143376
0	MORTGAGE	1855	12127	13982	0.132671



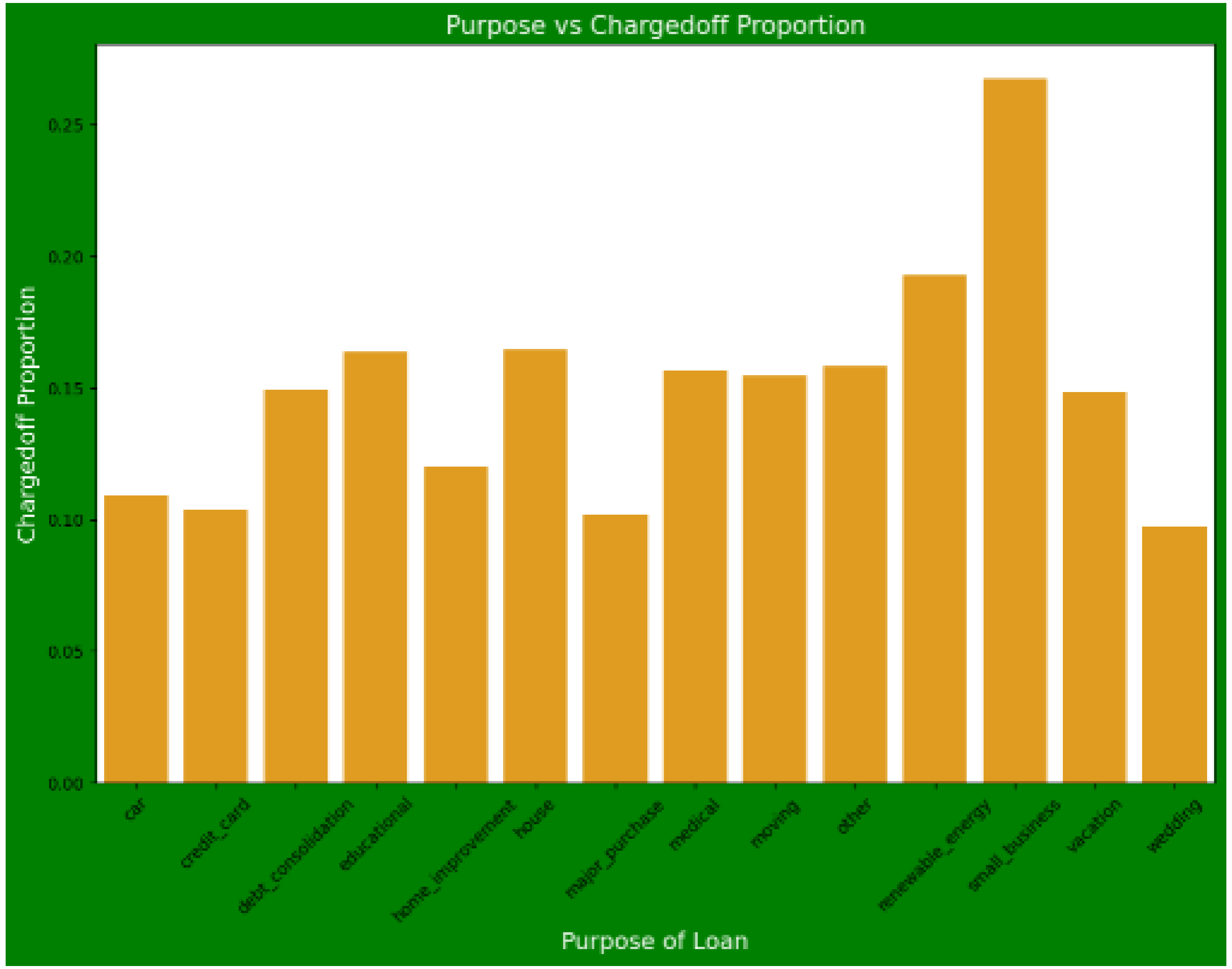
Observation:

1. Individuals who do not own a home have a higher likelihood of loan defaults compared to those with home ownership.

PURPOSE VS CHARGED OFF



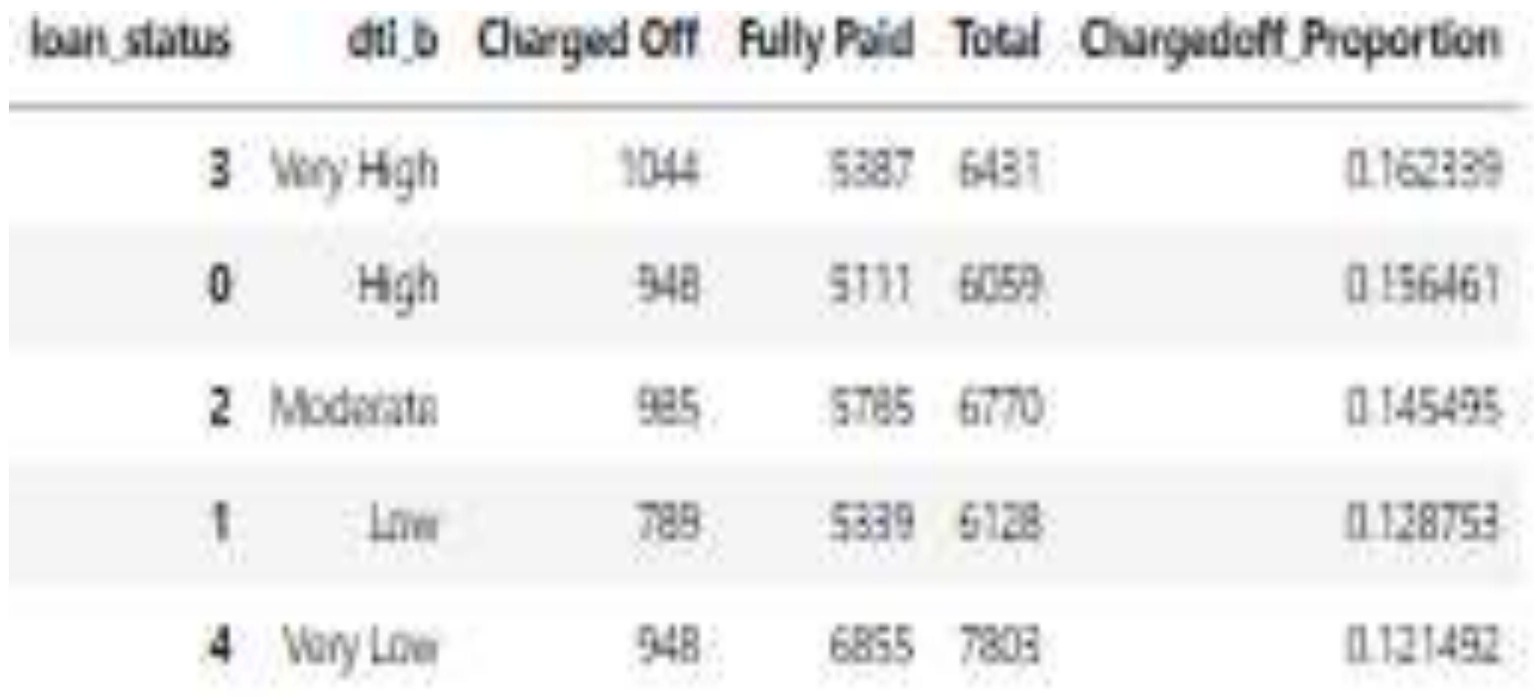
loan_status	purpose	Charged Off	Fully Paid	Total	Chargedoff_Proportion
11	small_business	366	1003	1369	0.267348
10	renewable_energy	16	67	83	0.192771
5	house	49	249	298	0.164430
3	educational	46	235	281	0.163701
9	other	531	2823	3354	0.158318
7	medical	95	510	605	0.157025
8	moving	79	433	512	0.154297
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



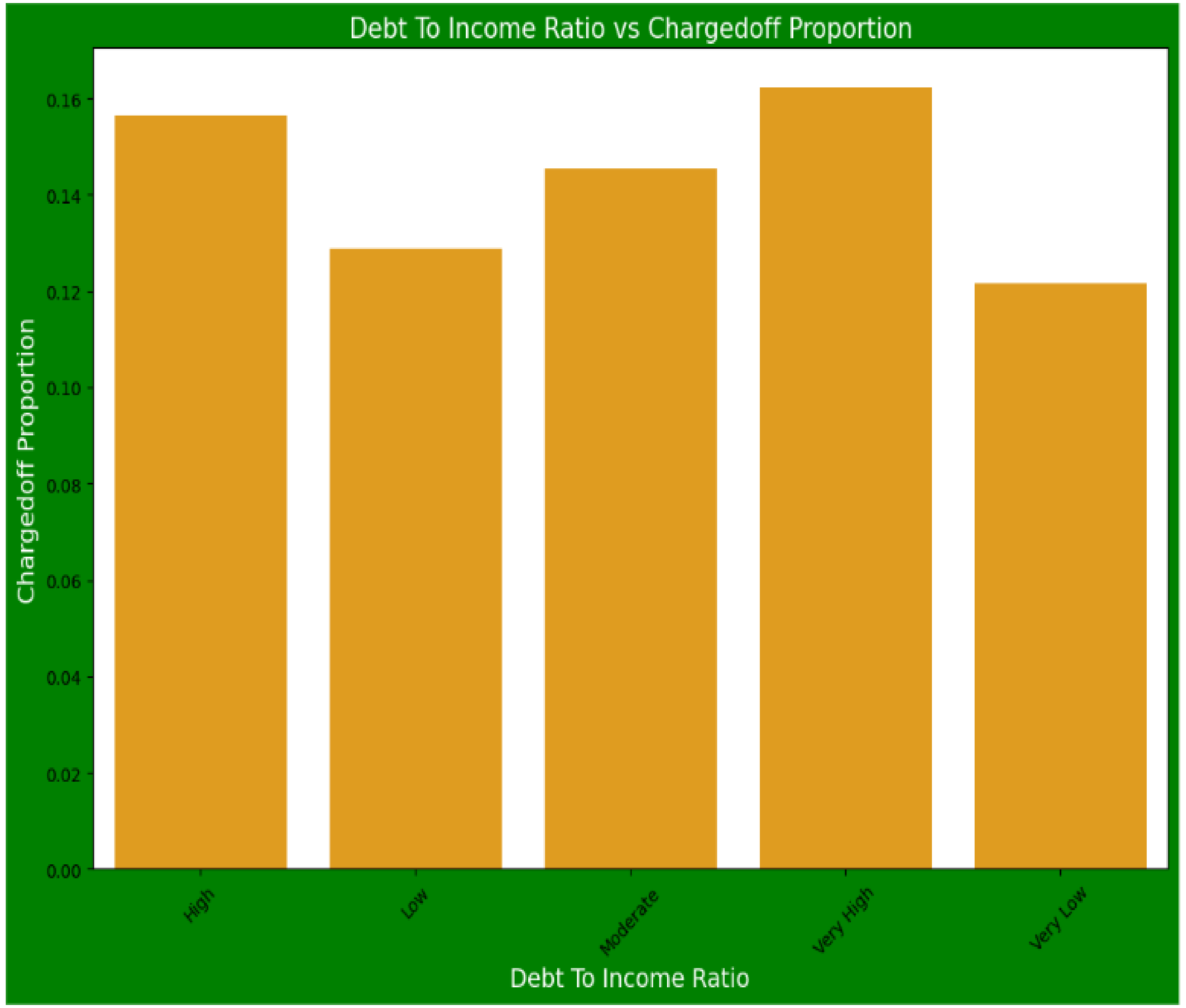
Observations:

1. Applicants who took loans for wedding purposes tend to have lower chances of loan defaults.
2. Applicants who borrowed for small business purposes exhibit higher chances of loan defaults.

DTI VS CHARGED OFF



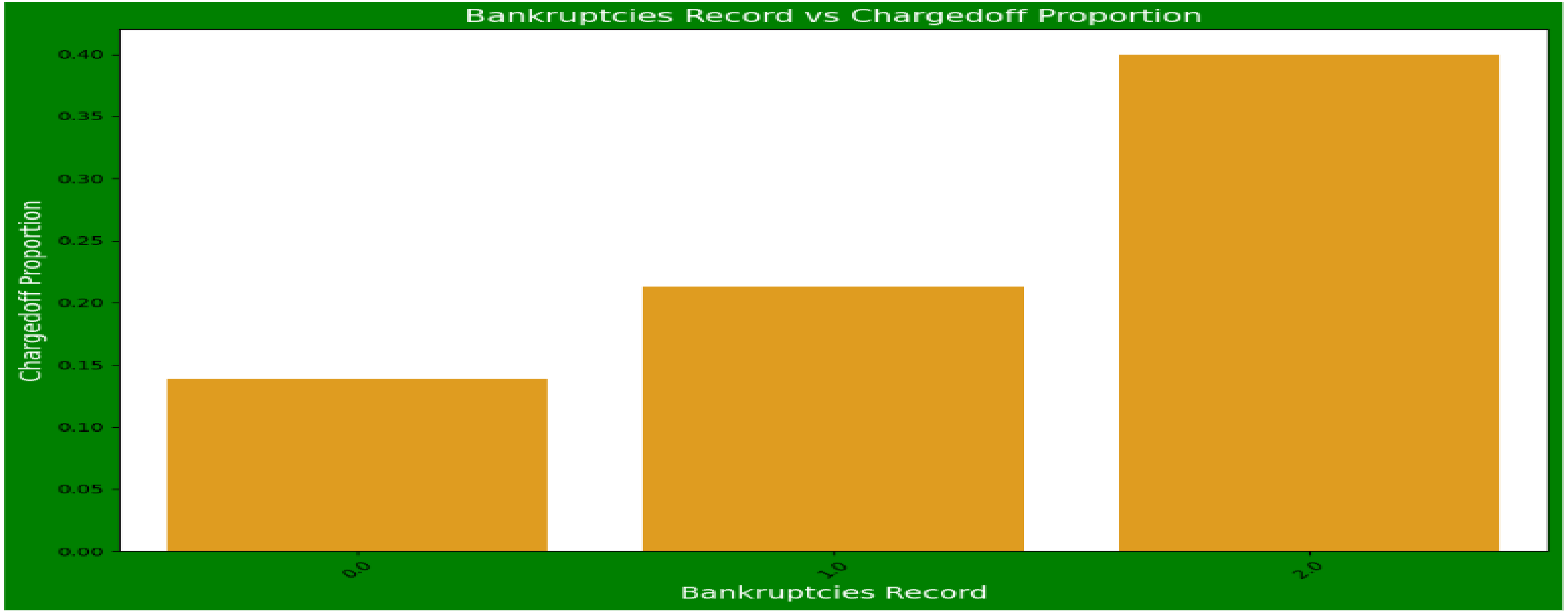
loan_status	dti_b	Charged Off	Fully Paid	Total	Chargedoff_Proportion
3	Very High	1044	5387	6431	0.162339
0	High	948	5111	6059	0.156461
2	Moderate	985	5785	6770	0.145495
1	Low	789	5339	6128	0.128753
4	Very Low	948	6855	7803	0.121492



Observations:

1. Higher Debt-to-Income (DTI) values are associated with a higher risk of loan defaults.
2. Lower DTI values indicate a lower probability of loan defaults.

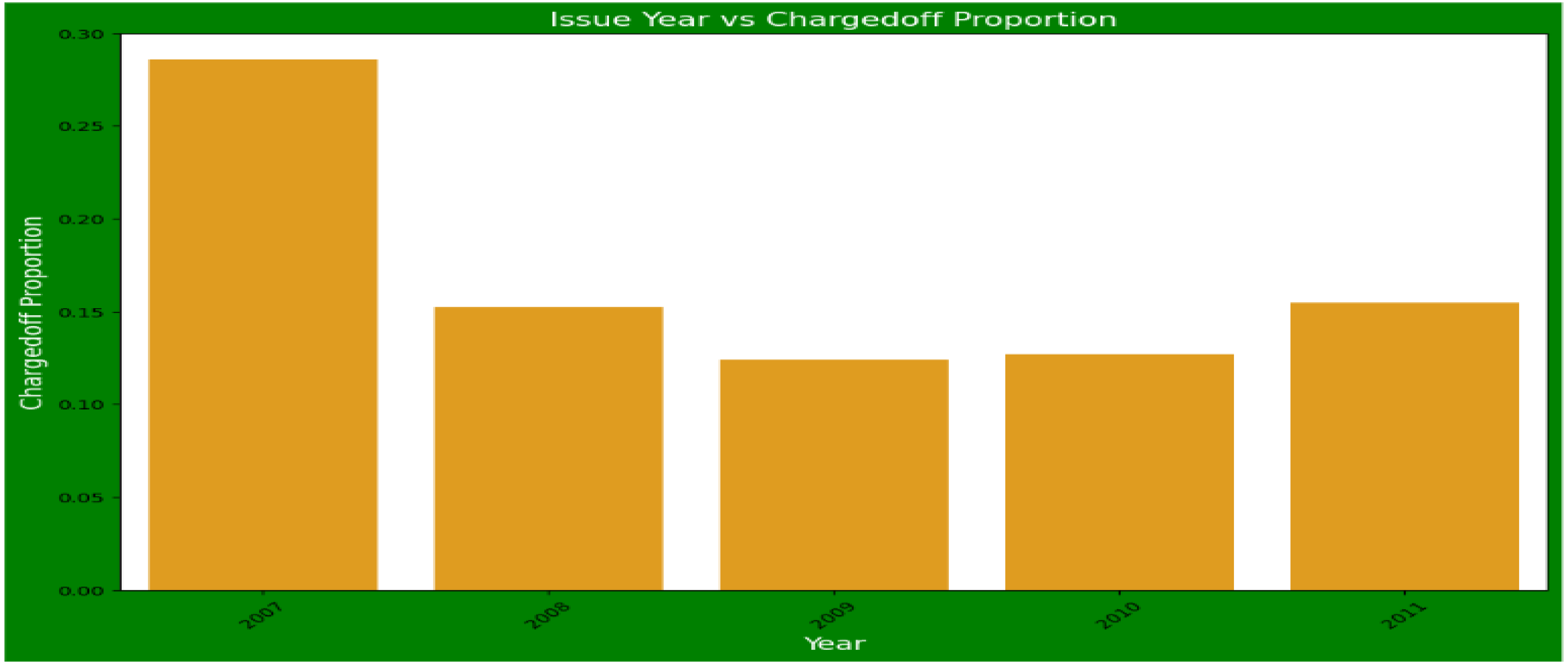
BANKRUPTCIES RECORD VS CHARGED OFF



Observations:

1. Loan defaults are significantly higher among applicants with 2 bankruptcy records.
2. Applicants with no bankruptcy records (0) have a much lower risk of defaulting on their loans.
3. Generally, the fewer bankruptcy records an applicant has, the lower the risk of default.

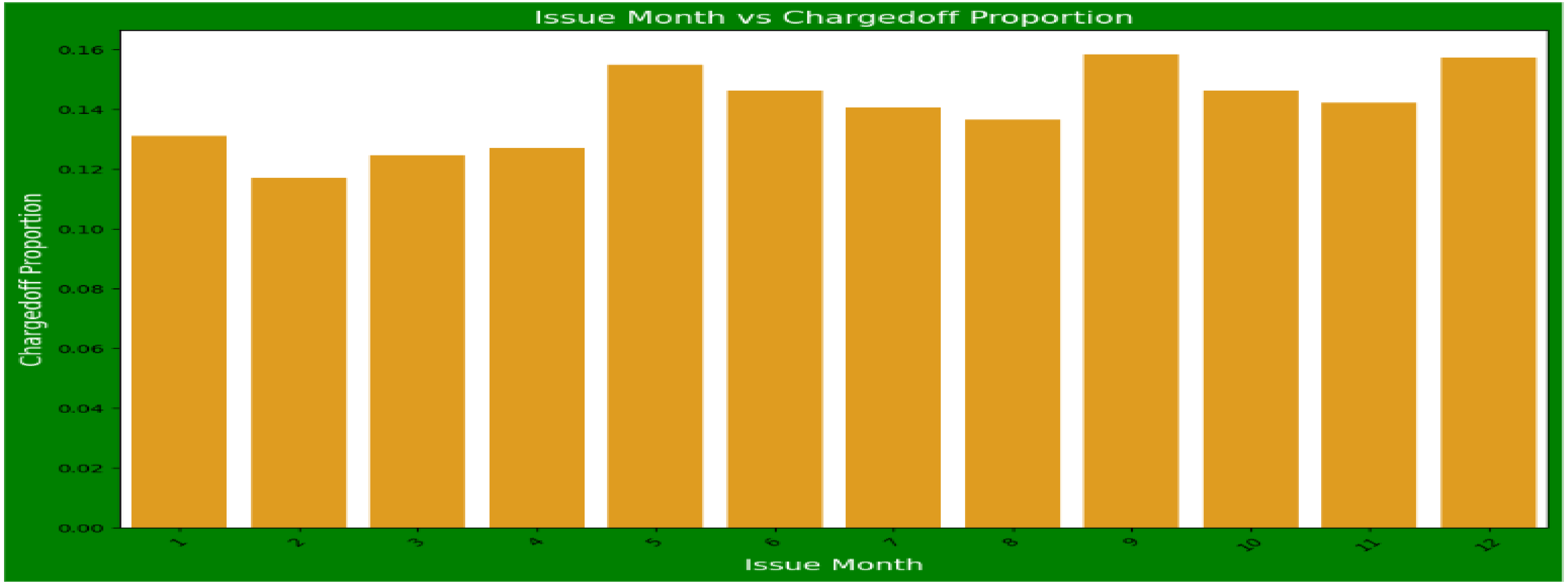
ISSUE YEAR VS CHARGED OFF



Observations:

1. Year 2007 has the highest proportion of loan defaults, indicating a higher risk of charge-offs in that year.
2. Year 2009 shows the lowest proportion of loan defaults, indicating a relatively better repayment behavior for loans issued that year.

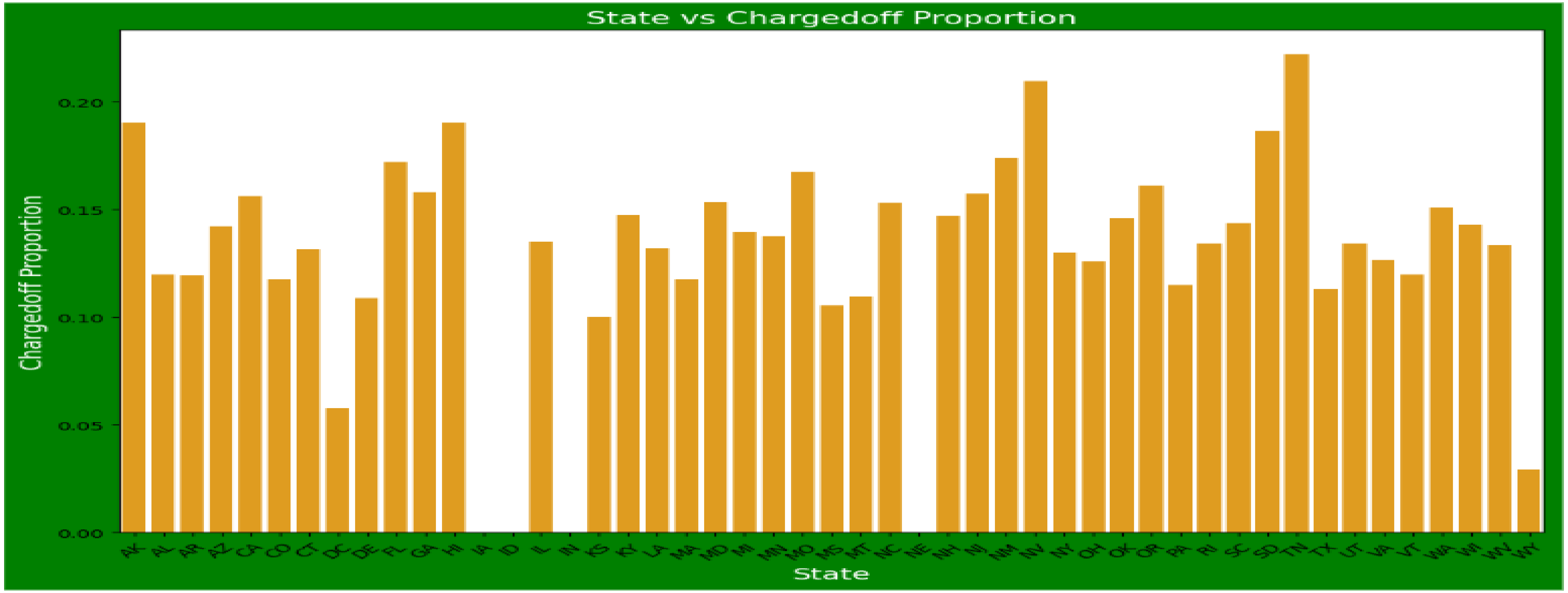
ISSUE MONTH VS CHARGED OFF



Observations:

1. Loans issued in May, September, and December exhibit a higher number of loan defaults.
2. February is another month with a significant number of loan defaults.
3. A substantial portion of loan defaults originates from loans approved between September and December.

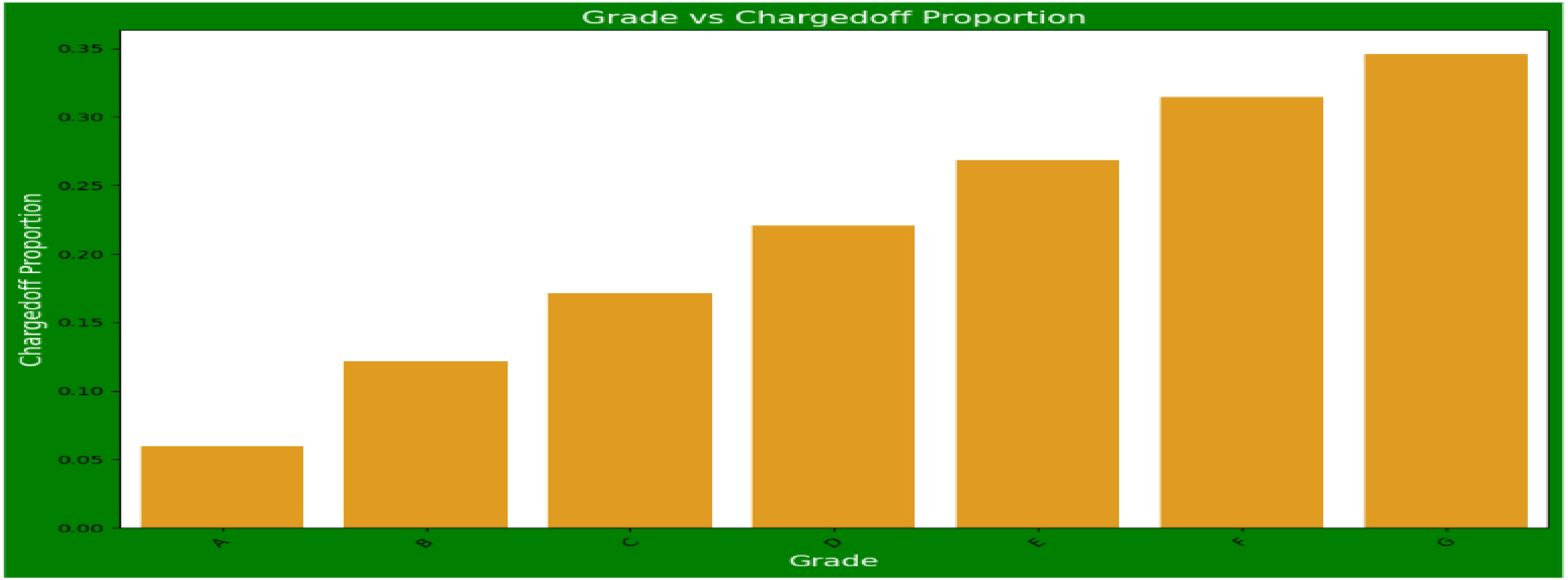
STATE VS CHARGED OFF



Observations:

1. DE States have the highest number of loan defaults.
2. CA (California) has the lowest number of loan defaults.

GRADE VS CHARGEDOFF



Observations:

1. Loan applicants with Grade G have the highest proportion of loan defaults (charged-off loans).
2. Loan applicants with Grade A exhibit the lowest proportion of loan defaults, indicating better loan repayment rates in this group.

CORRELATION

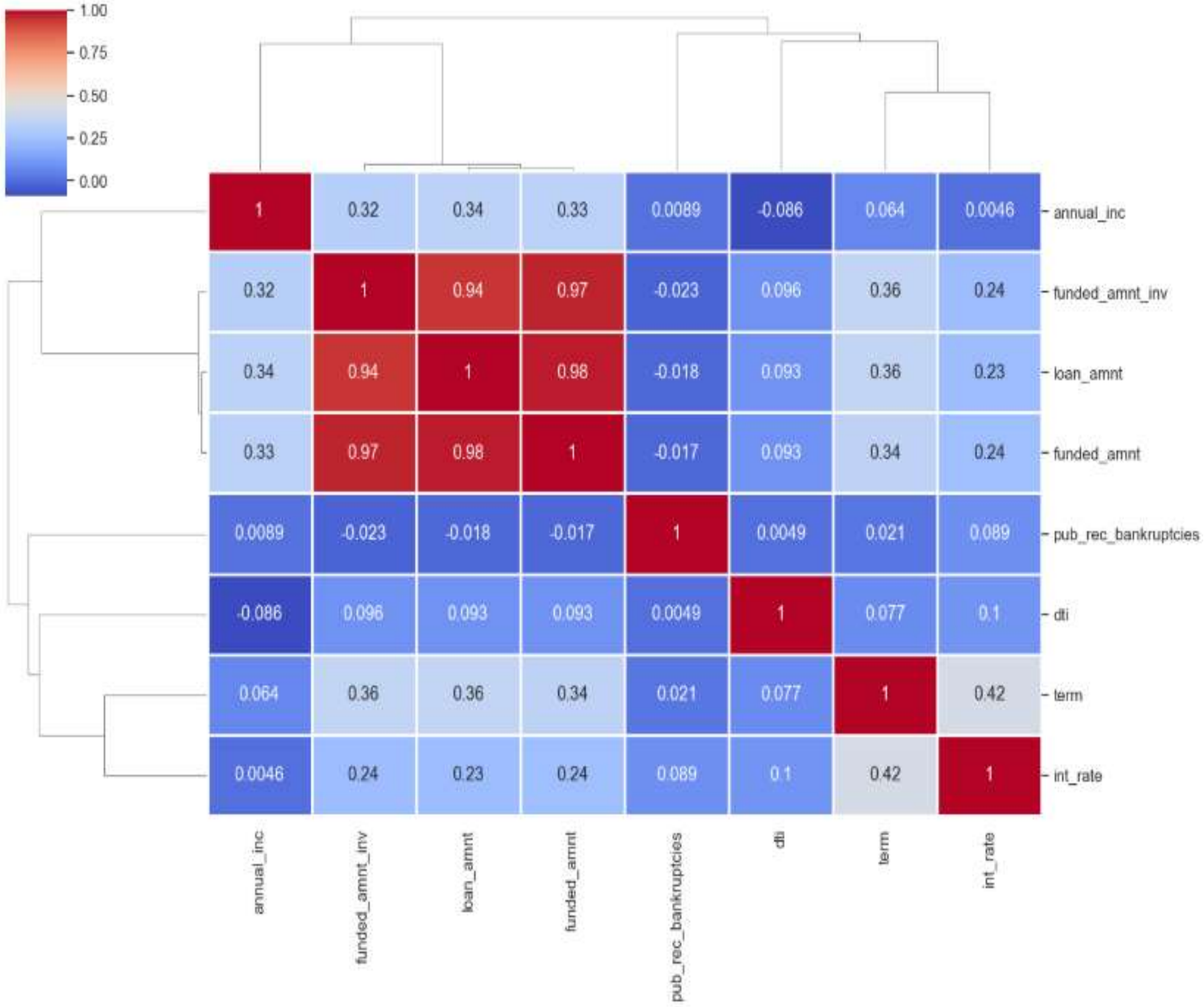
CORRELATIONS

Negative Correlations:

1. Loan Amount and Bankruptcies: Higher bankruptcies are associated with lower loan amounts.
2. Annual Income and Debt-to-Income Ratio: Higher income leads to a lower debt-to-income ratio.

Strong Positive Correlations:

1. Loan Term and Loan Amount: Longer loan terms are linked to higher loan amounts.
2. Loan Term and Interest Rate: Longer terms are associated with higher interest rates.
3. Annual Income and Loan Amount: Higher income corresponds to larger loan amounts.



	annual_inc	funded_amnt_inv	loan_amnt	funded_amnt	pub_rec_bankruptcies	dti	term	int_rate
annual_inc	1	0.32	0.34	0.33	0.0089	-0.086	0.064	0.0046
funded_amnt_inv	0.32	1	0.94	0.97	-0.023	0.096	0.36	0.24
loan_amnt	0.34	0.94	1	0.98	-0.018	0.093	0.36	0.23
funded_amnt	0.33	0.97	0.98	1	-0.017	0.093	0.34	0.24
pub_rec_bankruptcies	0.0089	-0.023	-0.018	-0.017	1	0.0049	0.021	0.089
dti	-0.086	0.096	0.093	0.093	0.0049	1	0.077	0.1
term	0.064	0.36	0.36	0.34	0.021	0.077	1	0.42
int_rate	0.0046	0.24	0.23	0.24	0.089	0.1	0.42	1

CONCLUSIONS

1. **Income**:
   Applicants with an income range of **0-20,000** are at the highest risk of **loan defaults**.

2. **Interest Rates**:
   Loans with an **interest rate above 16%** have significantly higher chances of being **charged off**.

3. **Home Ownership**:
   Non-homeowners are more likely to **default on loans** compared to those who own homes.

4. **Loan Purpose**:
   Loans taken for **small businesses** exhibit the highest default rates compared to other loan purposes.

5. **Debt-to-Income (DTI)**:
   High DTI values correlate with a greater risk of loan defaults.

6. **Credit History**:
   Applicants with a history of **bankruptcies** have a significantly higher probability of defaulting.

7. **State-Level Defaults**:
   The state of **Delaware (DE)** records the highest proportion of **loan defaults**.

8. **Loan Grades**:
   Applicants with loans in **Grade G** are the most likely to default, while **Grade A** has the least defaults.

THANK YOU