Lending Club Case Study

Data Analysis

Sahadeb Patro

Sivaprasad M

Agenda

- Understand Problem Statement
- Data Sourcing & Cleaning
- Univariate Analysis
- Bivariate Analysis
- Derived Metrics
- Conclusion

Problem Statement:

Problem statement

- A consumer finance company into lending loans to individuals faces the **challenge of optimizing loan approvals while minimizing the risks of defaulters**. Task here to analyze the historical data and find the driving factors to optimize the loan defaulters.

Business Objectives:

- **Risk Assessment**: Utilize EDA on historical loan data (2007-2011) to evaluate the risk associated with loan applicants based on their likelihood of repayment.
- **Default Identification**: Identify patterns indicating the likelihood of loan default by analyzing **consumer and loan attributes**. Apply EDA techniques to analyze consumer and loan attributes.
- **Data Understanding**: Understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- **Mitigate credit loss** by identifying and optimizing loans to high-risk applicants.

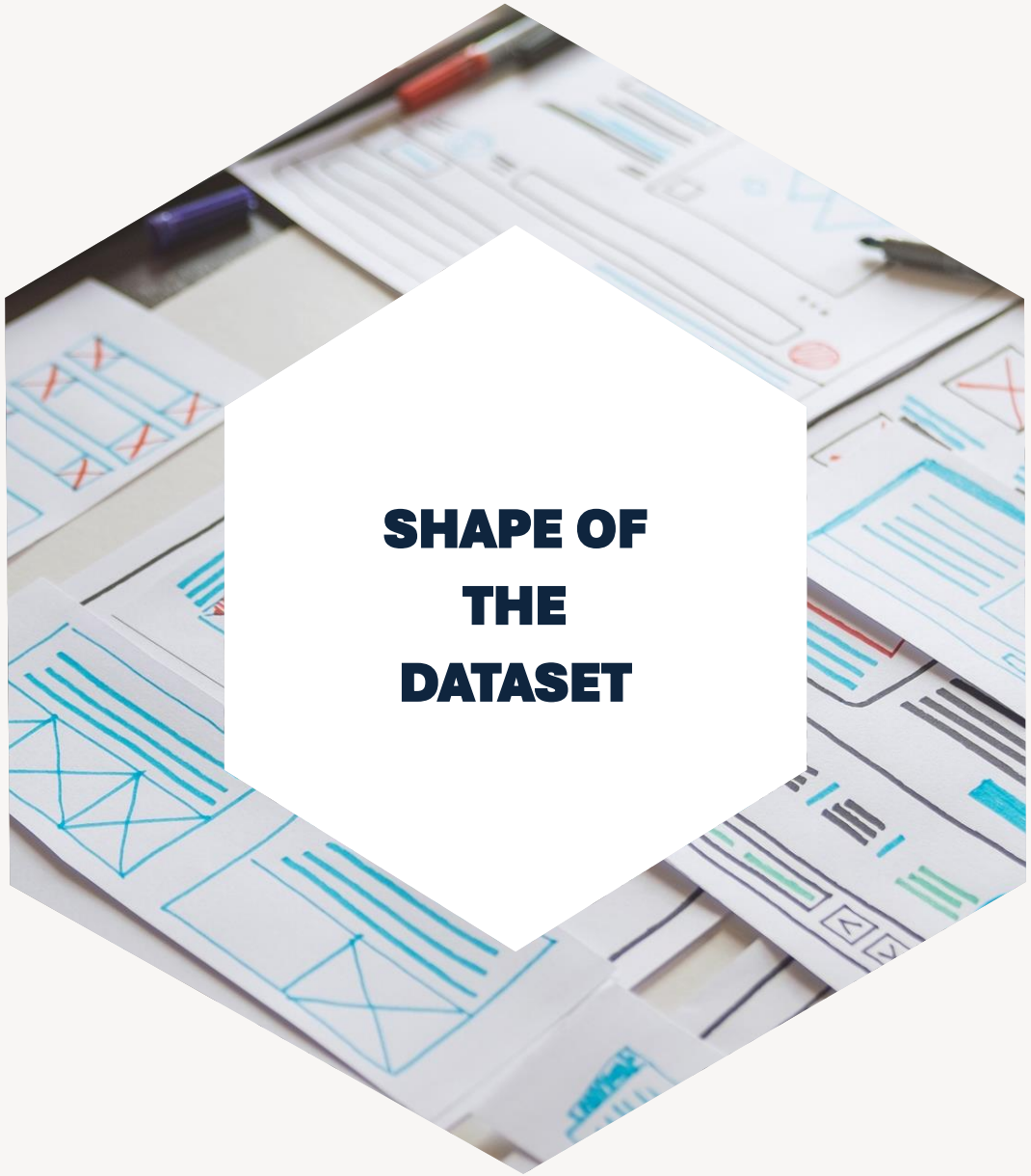
Loan Outcome Categories:

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:
    - **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.
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).

Strategy for Analysis

Data Sourcing	Data Cleaning	Data Conversion & Derived columns	Data Analysis	Conclusion
Import modules Import CSV Data Understand Data	Null/Missing values Invalid values Drop Columns Drop rows Data filtering Imputing rows Managing Outliers	Data Type conversion Formatting Data types Creating Derived columns	Univariate Analysis Bivariate Analysis Correlation	Summary of Data Analysis



SHAPE OF THE DATASET

Imported Dataset (loan.csv) :

111	39717
Columns	Rows

Data Cleaning

Columns:

- Dataset lacks headers and footers.
- Removed columns with 100% null values; a total of 54 columns.
- Dropped two descriptive columns ('desc' and 'title') since they do not impact the outcome.
- Eliminated 9 columns with a single value, making them non-quantifiable.
- Excluded 12 behavioral columns that were either unique or irrelevant for the outcome.
- 'member_id' column has been removed; 'id' is now the primary key for analysis.

Rows:

- No duplicate rows in the dataset.
- **Dropped 1140** rows with a loan status of **'Current**,' as they are still in progress and not fully paid nor defaulted.
- After data cleaning, the dataset now comprises **38577** rows and **34** columns.

Data Conversion, Formatting & Derived Columns

Conversion & Formatting

- Rounded all floating-point numbers in the dataset to two decimal points for consistency.
- Formatted the existing **'issue_d'** column to the date data type.
- Removed the percentage symbol from the **'int_rate'** and **'revol_util'** columns, converting them to the float data type for numerical analysis.

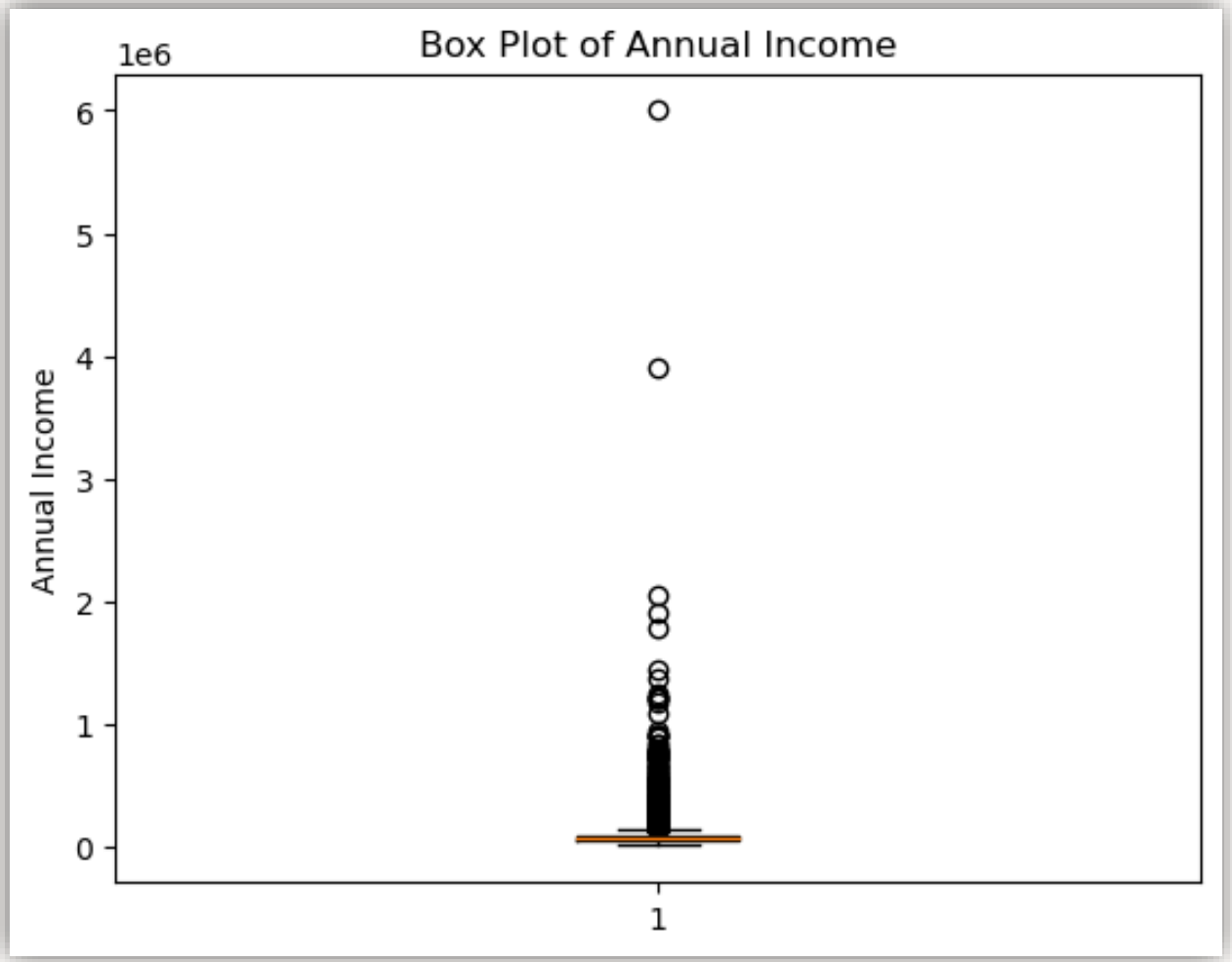
Derived Columns

- Extracted the **Year** and **Month** components from the existing **'issue_d'** date column to facilitate temporal analysis.
- Generated data **buckets** for essential numeric columns—specifically, **'annual_inc**,' **'loan_amnt**,' **'int_rate**,' and **'dti**.' These buckets will be utilized in Univariate analysis for a more granular examination of the distribution and patterns within these key numeric features.

Outliers

Outliers have been identified in the annual income field, with 171 borrowers reporting annual incomes exceeding 300k. Given the relatively low number of such instances, their presence could potentially distort the overall outcome. Consequently, these data points have been excluded from the analysis.

- annual_inc
- loan_amnt
- int_rate
- dti



Univariate Analysis

Univariate analysis can play a crucial role in predicting loan defaulters in the Lending Club case study by examining individual variables in isolation.

By systematically analyzing each variable independently, univariate analysis helps identify patterns, trends, and potential outliers that could be indicative of default risk.

Boxplot and **bar plot** can be jointly analyzed for Univariate analysis.

Loan Amount

Observation:

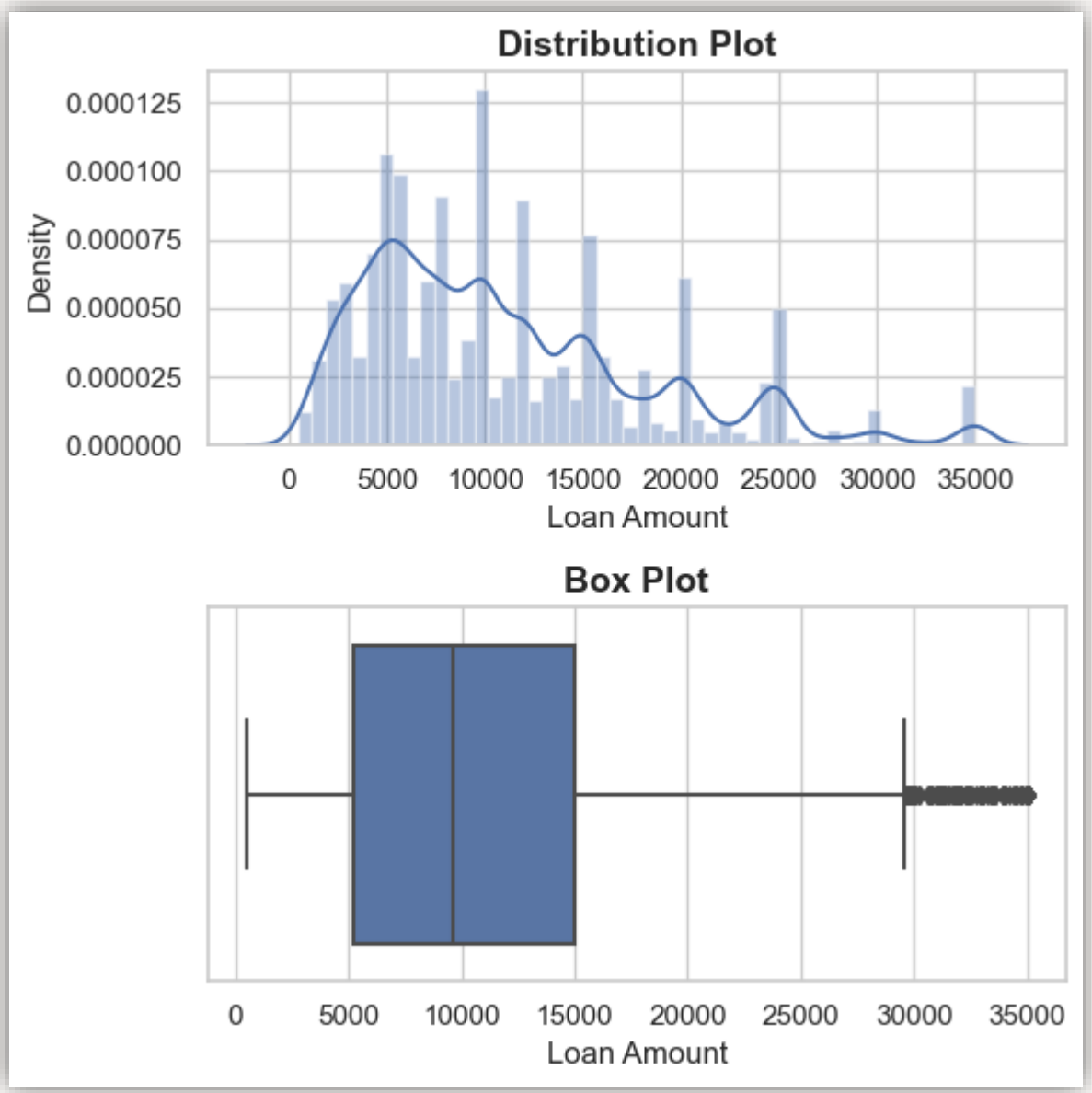
Min Loan amount: 500

Max Loan amount: 35,000

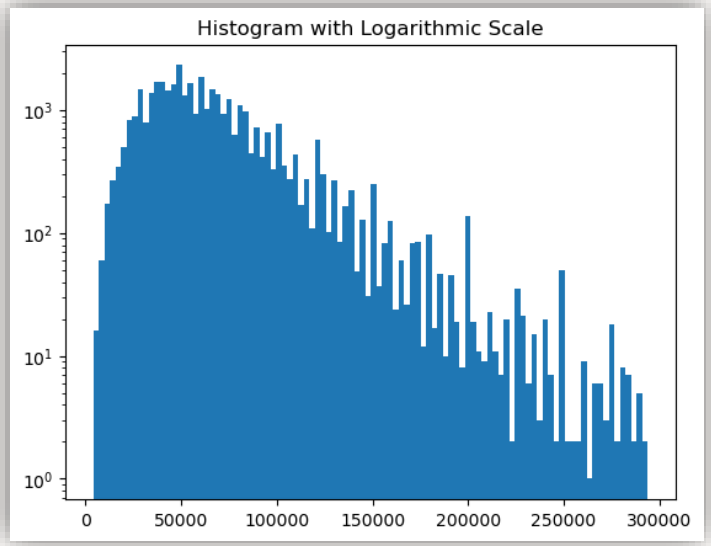
Average Loan amount: 11,000 (approx.)

Most of the applied loan amount seems to be in between 5000 to 15000 range.

People prefer to take the loan amounts which are multiple of 5k, ex: 5k, 10k, 15k, etc.



Annual Income



Observation:

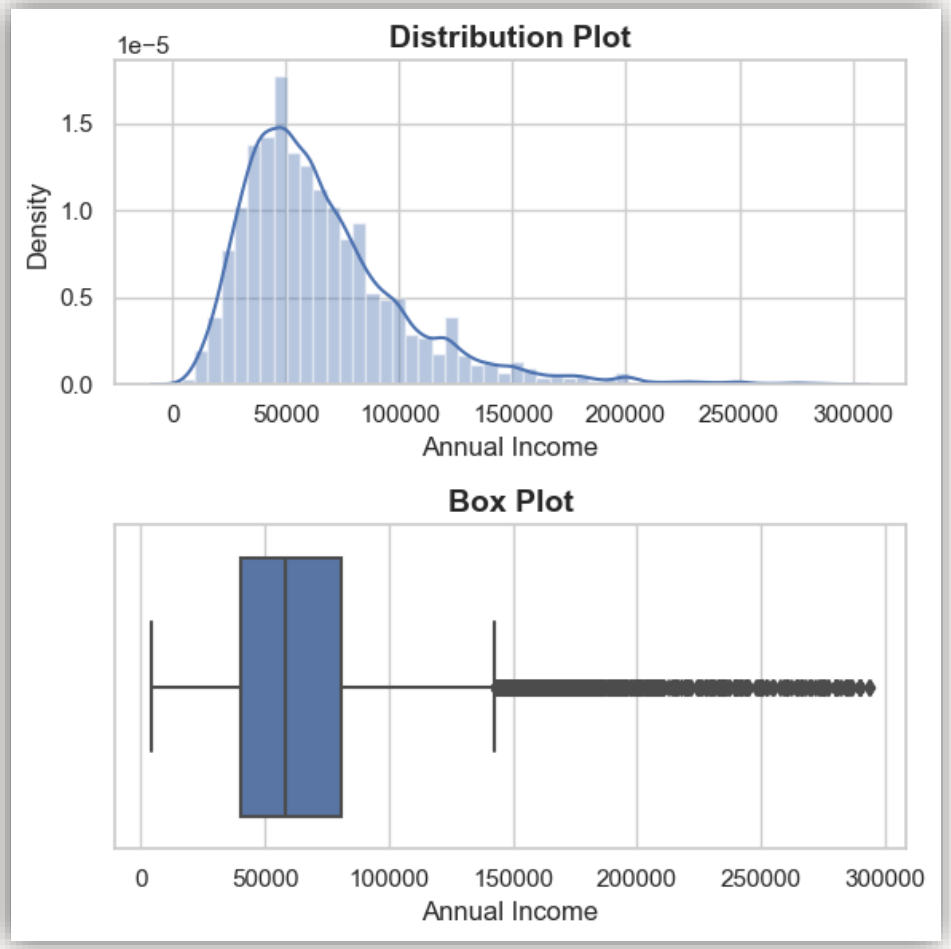
Min Income: 4000

Max Income: 2,94,000

Average Income: 66,000 (approx.)

Most of the borrower's income range is in between 40,000 to 81,000 range.

There are many borrower's whose annual income is above the annual income range.



Loan Status

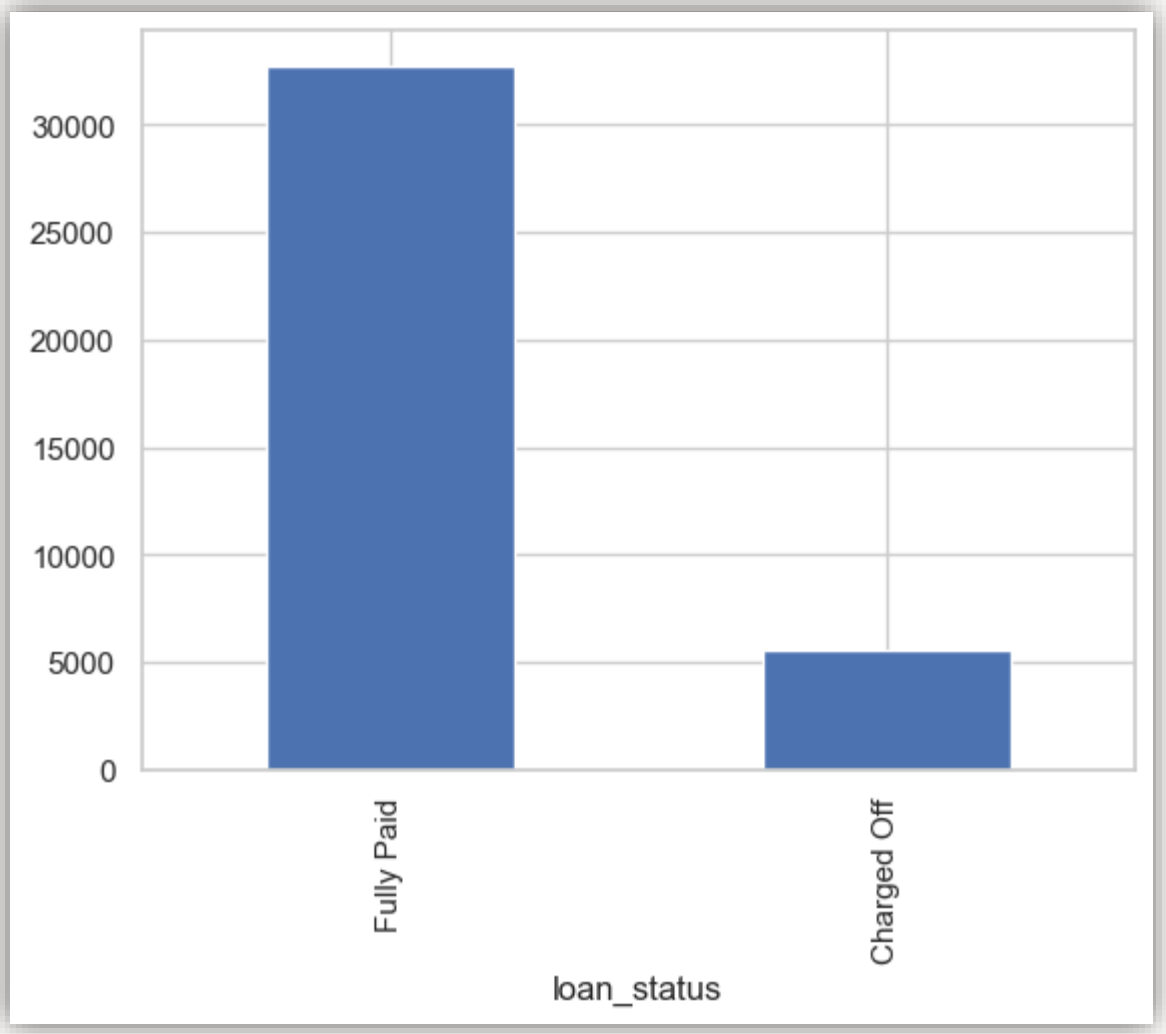
Observation:

Total Fully Paid : 32756

Total Charged Off : 5603

% of Defaulters on the existing data: 14.61 %

The defaulter's data can be analyzed to predict the pattern further.



Interest Rate

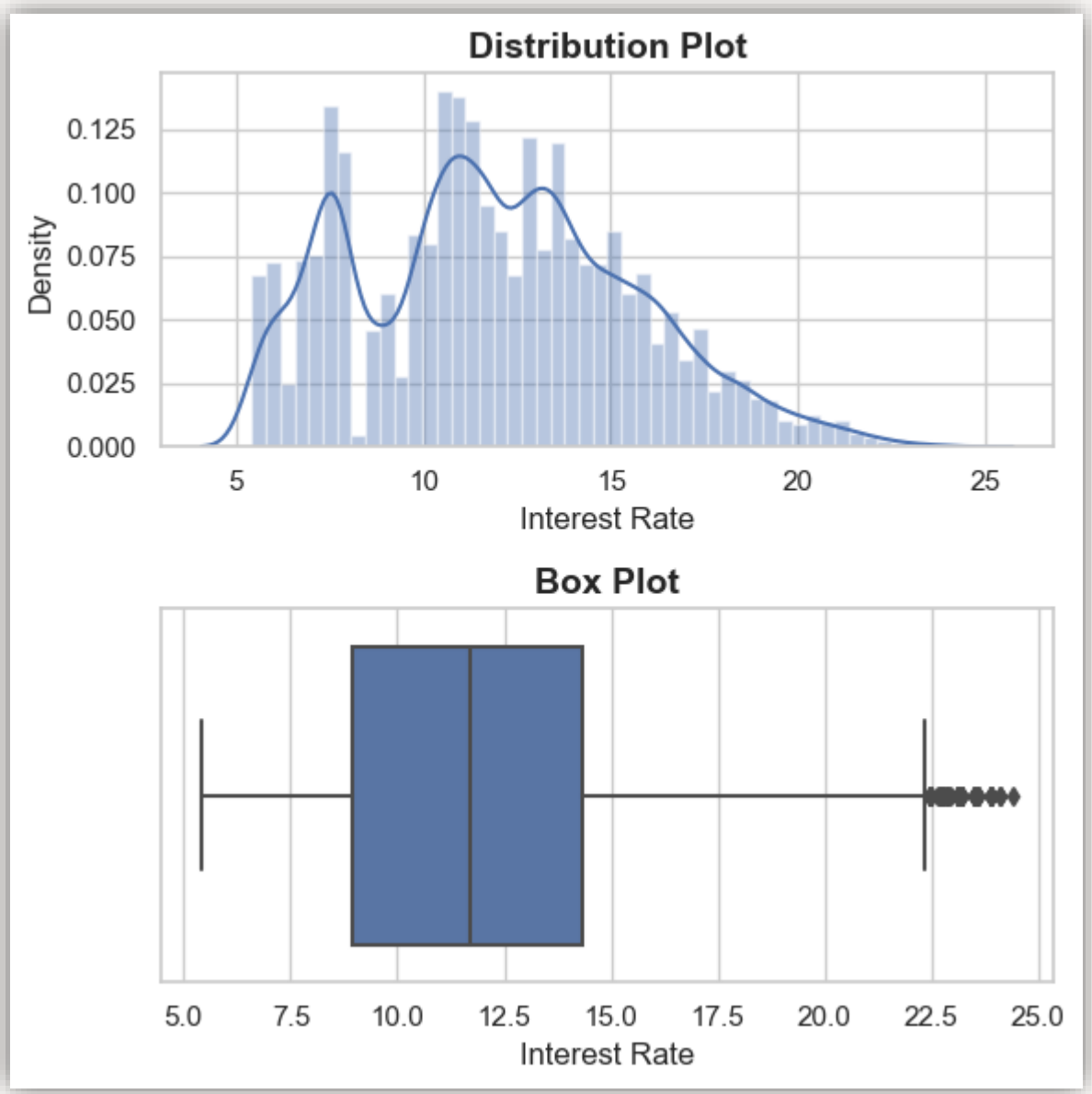
Observation:

Min Interest Rate: 5.2 %

Max Interest rate: 24.4%

Average Loan amount: 11.92 %

Maximum interest rates are in the bracket of 8.94% to 14.35%.



Categorical Data

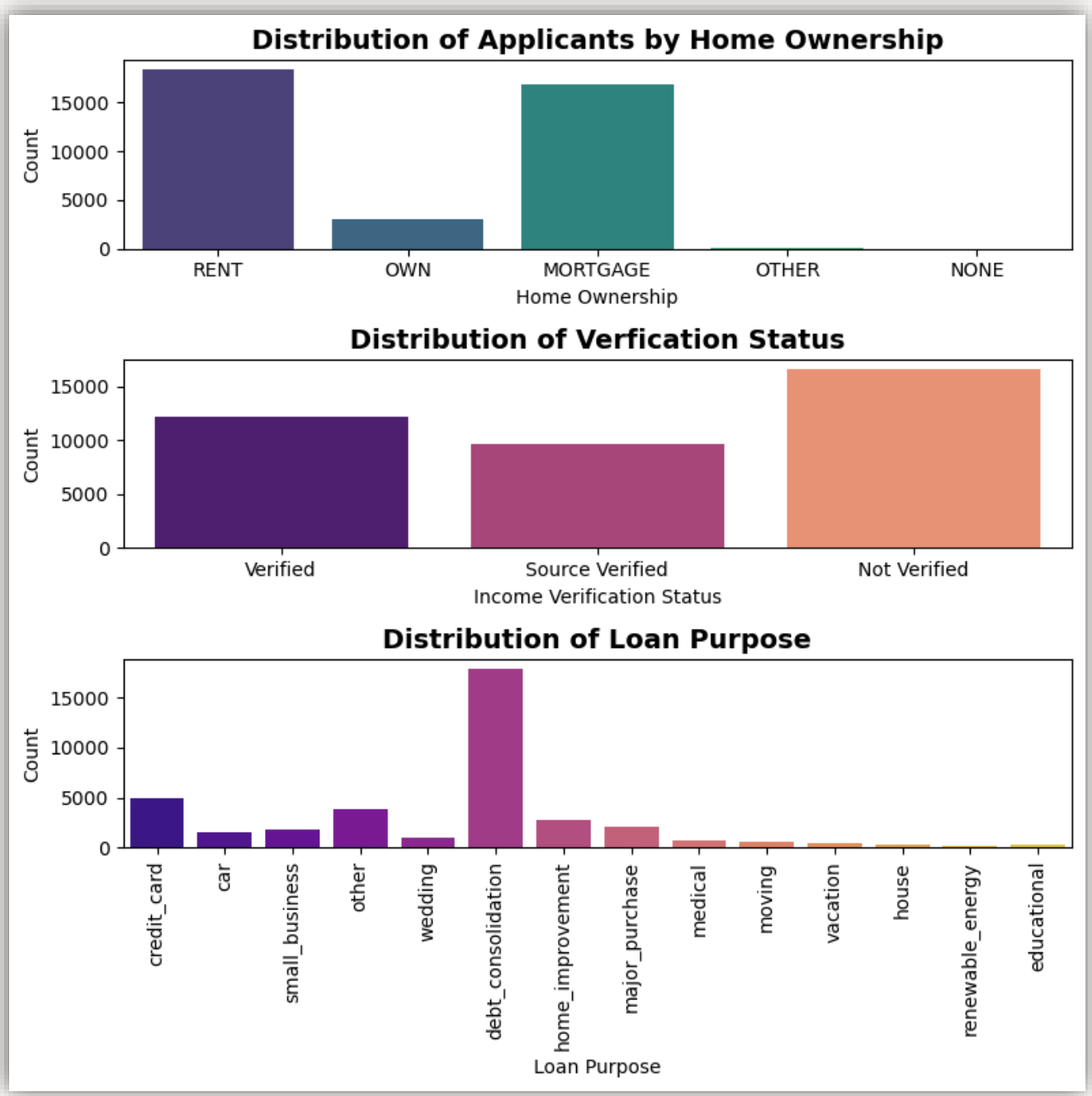
Observation:

- Income Verification Status
- Loan Purpose
- Home Ownership

Majority of loan applicants are either living on **Rent** or on **Mortgage**, followed by Own house. People owning a rented house takes more loan than those who own a house.

Most of the loan applications are for **debt_consolidations**.

Most of the applications are **Not verified** status, followed by the **Verified** status.



Categorical Data (cntd.)

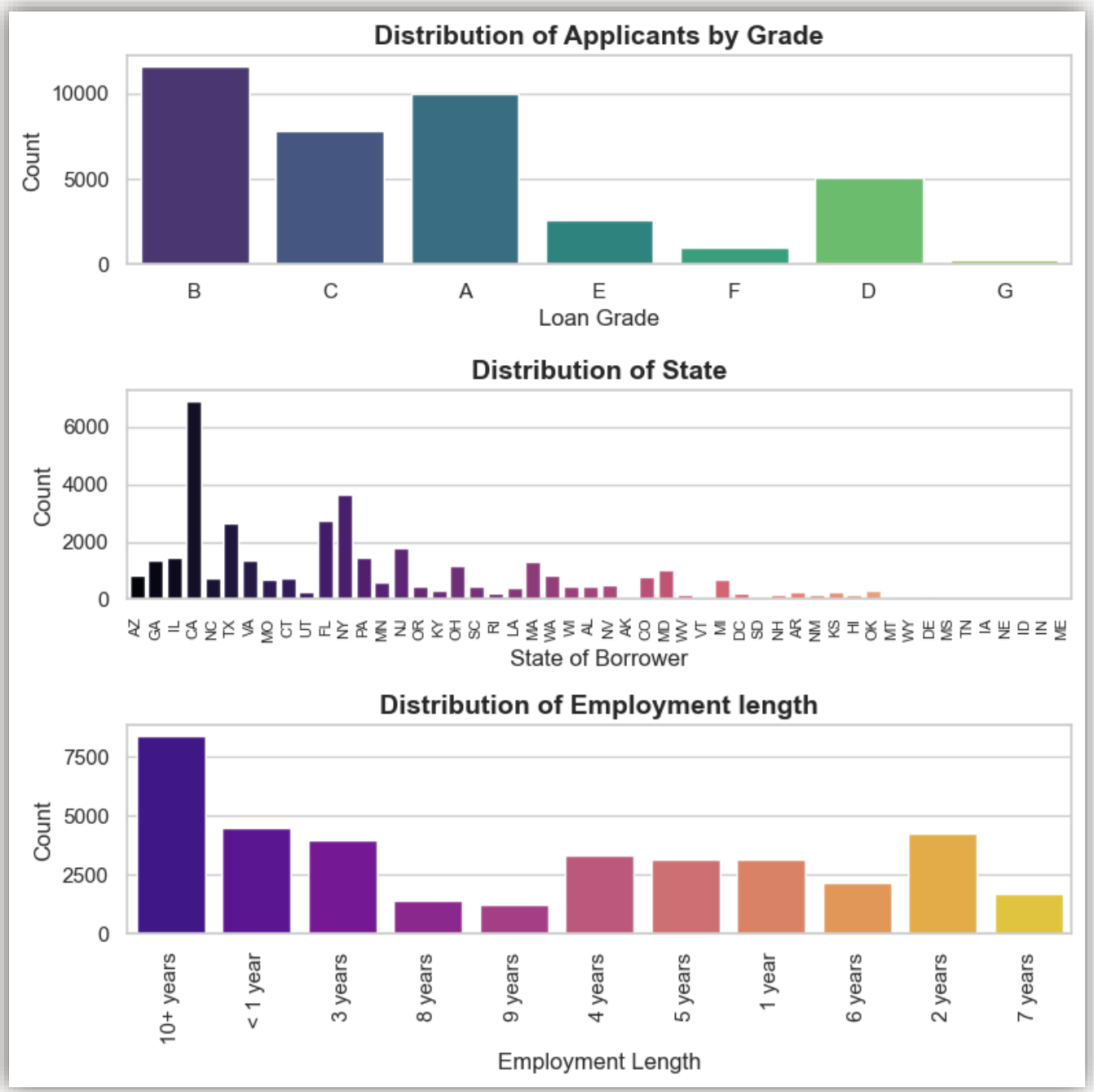
Observation:

- Loan Grade
- Loan State
- Employment Length

Most of the loan has grade B followed by A & C. Pint to notice here, there are borrowers in Grade D and could be the ones which eventually be defaulted mostly.

Maximum loan applied from state CA followed by NY, FL,TX & NJ. People from other states looks less interested in the loan.

Noticed that, **most of the borrowers are of above 10 years of employment experience**. Surprisingly, there is no pattern below 10 years of experience and number of applied loans do not have any relation with experience.



Categorical Data (cntd.)

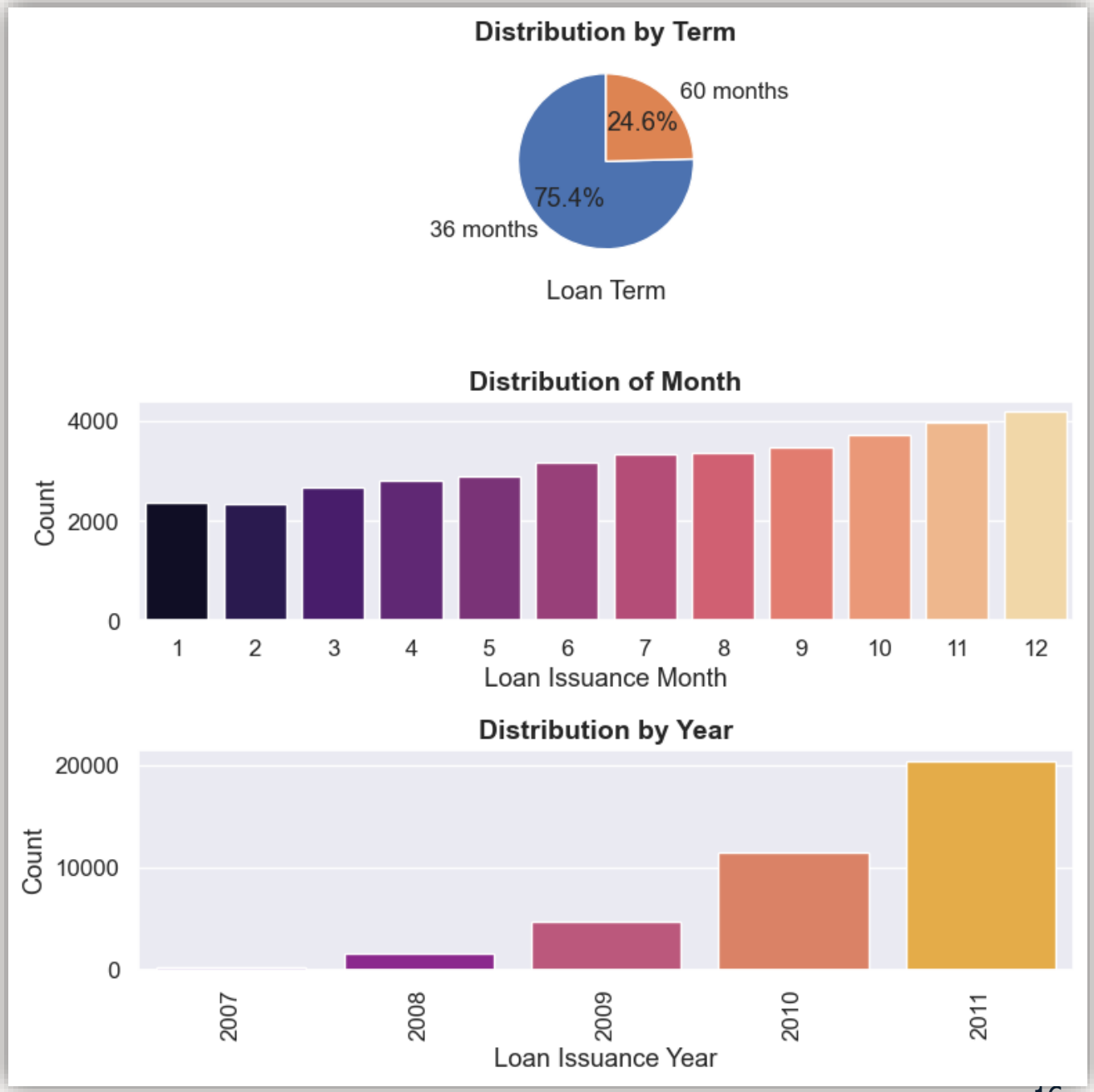
Observation:

- Loan Term
- Loan taken each Year
- Loan taken each Month

Borrowers are more interested in taking loan for 36-month term then 60 months. That means, they're more keen to close the loan early.

There is a steady increase of loan disbursements from month Jan to Dec. Which means, Jan being the slowest month in terms of providing loan or borrowers interested in taking loan, but high in Dec.

There has been a significant increase in loan applications year-on-year from 2007 to 2011. 2011 being the year with highest number of loan applications.



Bivariate Analysis

Bivariate analysis in the Lending Club case study helps in identifying relationships between key variables. By examining pairs such as interest rate and loan status, income and loan amount, debt-to-income ratio and loan status, and others, it provides insights into potential connections influencing loan default.

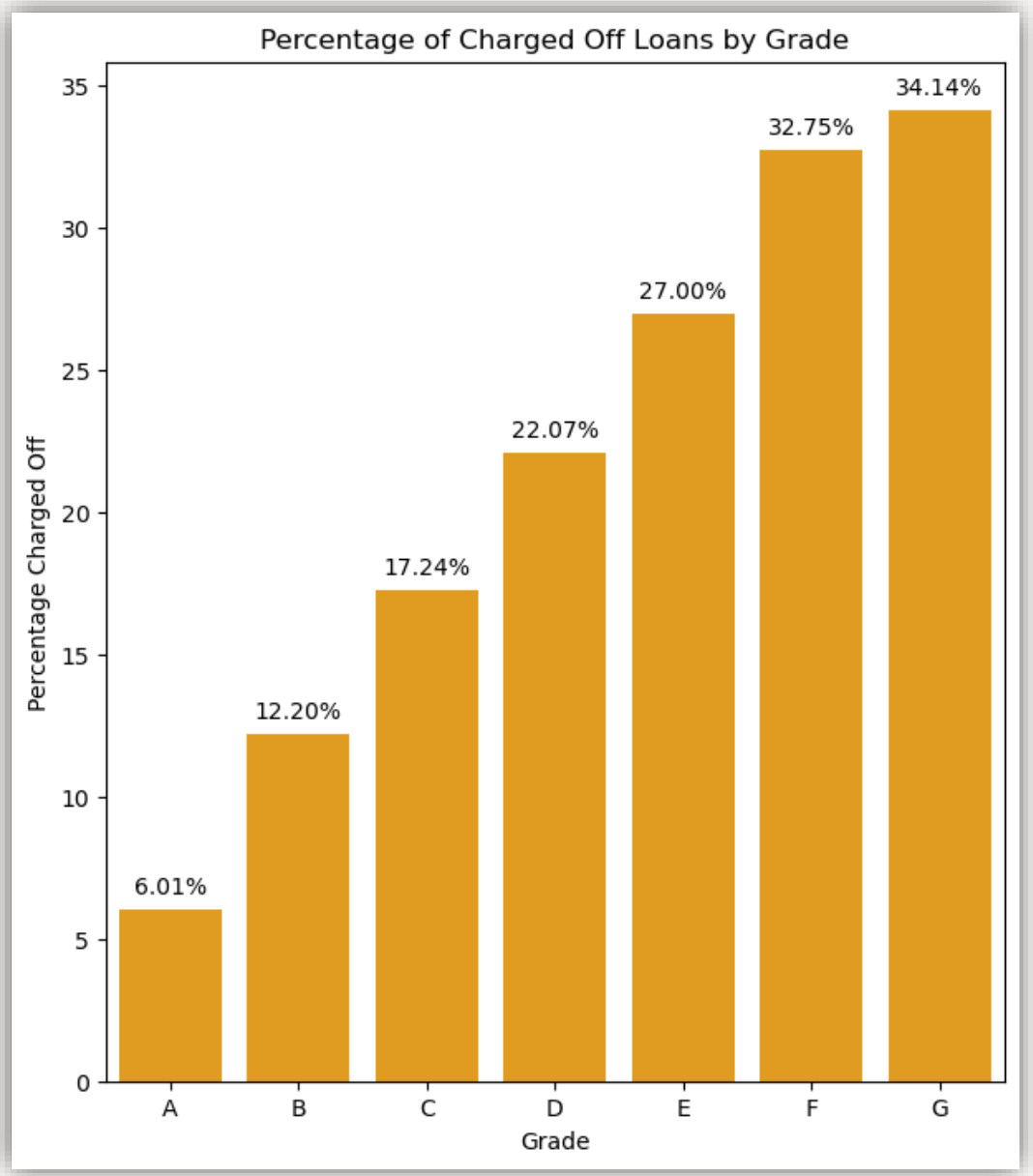
Grade Vs Charged Off

Observation:

Based on the data, it looks like Percentage of "charged off" rates increase from grades A to G.

Important:

Which means, the higher grades like D,E,G,F require more scrutiny on the loan applicants before approval, as they're the ones who more likely to be defaulted.



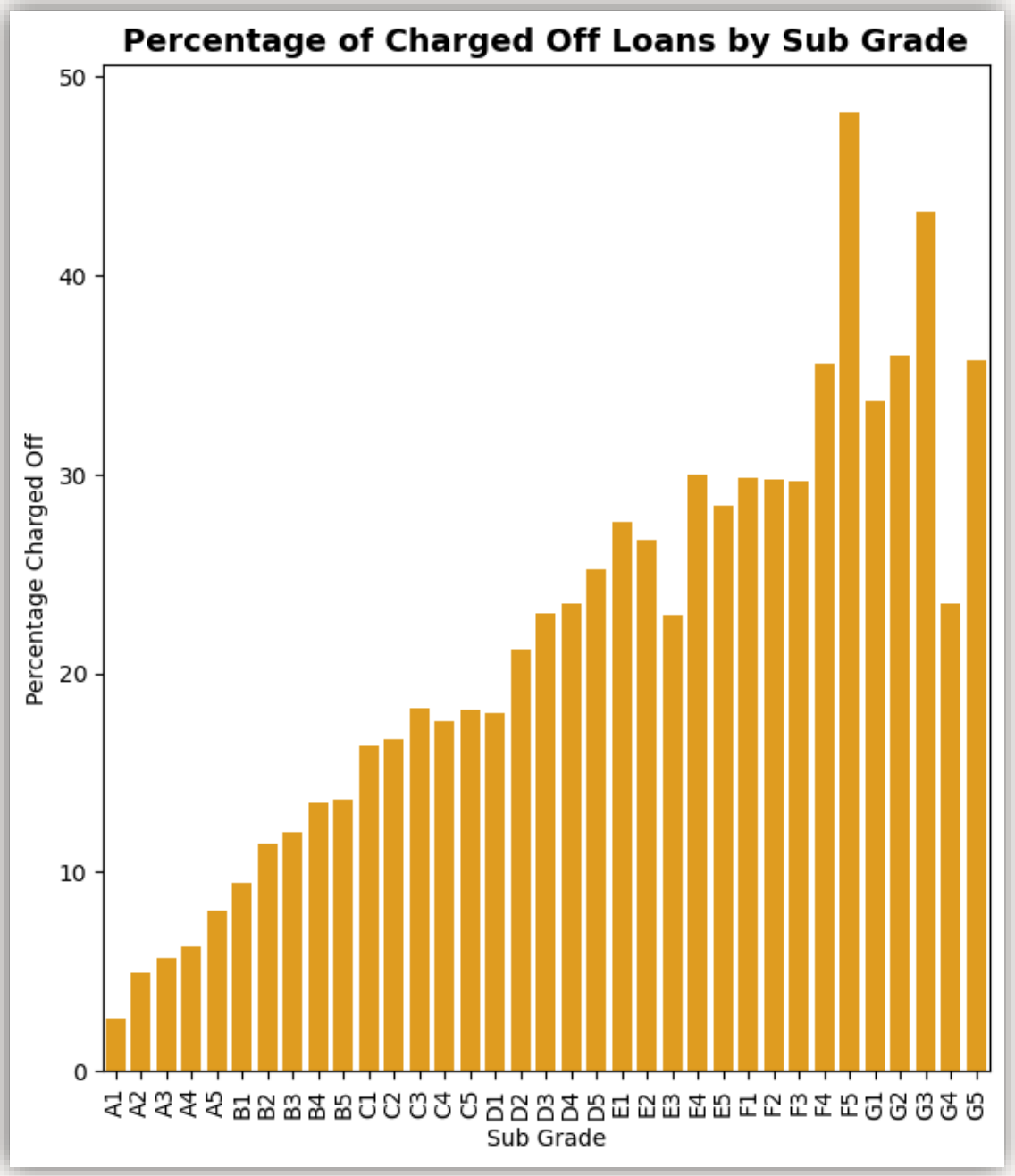
Sub Garde Vs Charged Off

Observation:

Based on the data, it looks like Percentage of "charged off" rates increase from sub grades A1 to G5.

Important:

Noticed, F5 and G3 are the sub grades which are more vulnerable for being defaulters. So, the lenders has to be more careful while lending loans in this category.



Interest Rate Vs Charged Off

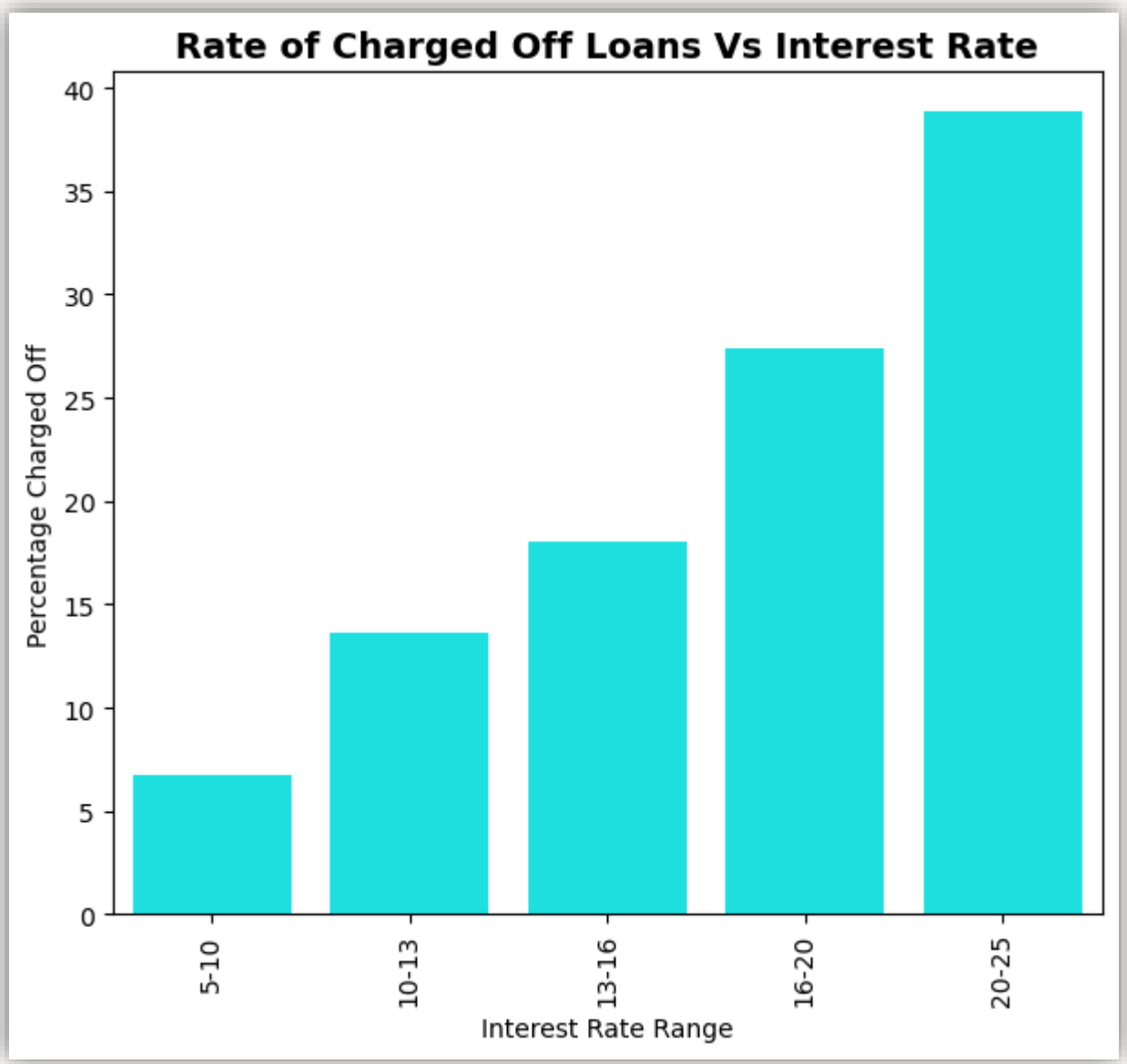
Observation:

Charged off proportion is increasing with higher interest rates.

Important:

There is a significant increase of charged off percentage for the interest rates from 16%-25%.

Important watch out for interest rate bucket 20%-25%.



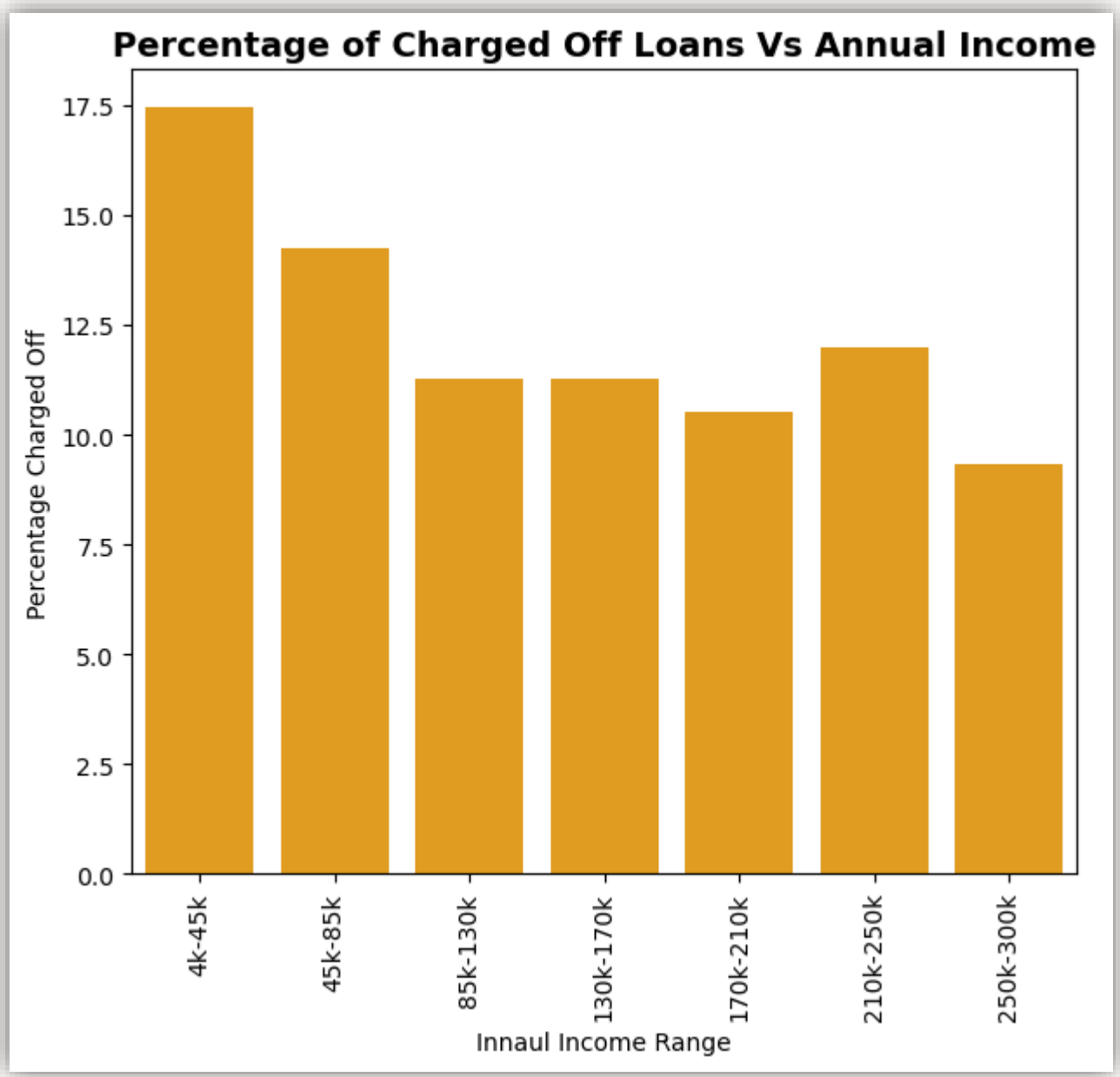
Annual Income Vs Charged Off

Observation:

Based on the data, it looks like the percentage of Charged off decreases (not for all cases with some exception like the income range 210k-250k) when annual income increases.

Important:

Noticed, the annual income range 4k-45k is more prone for defaulters. As the annual income is less, there could be challenges in loan repayment.



Purpose, Home ownership Vs Charged Off

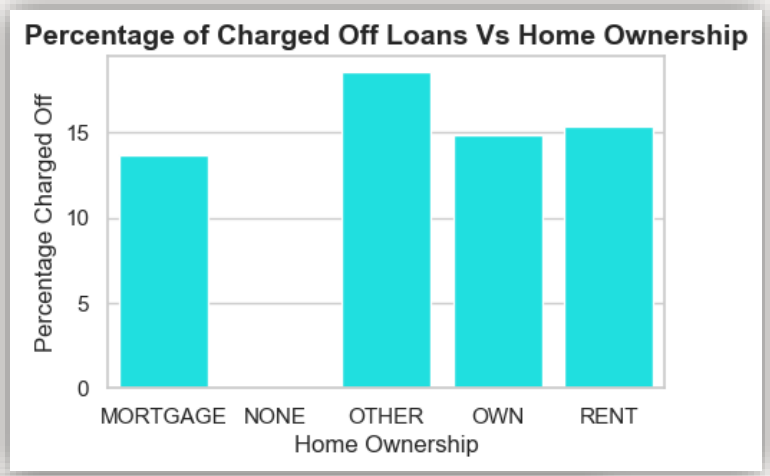
Observation:

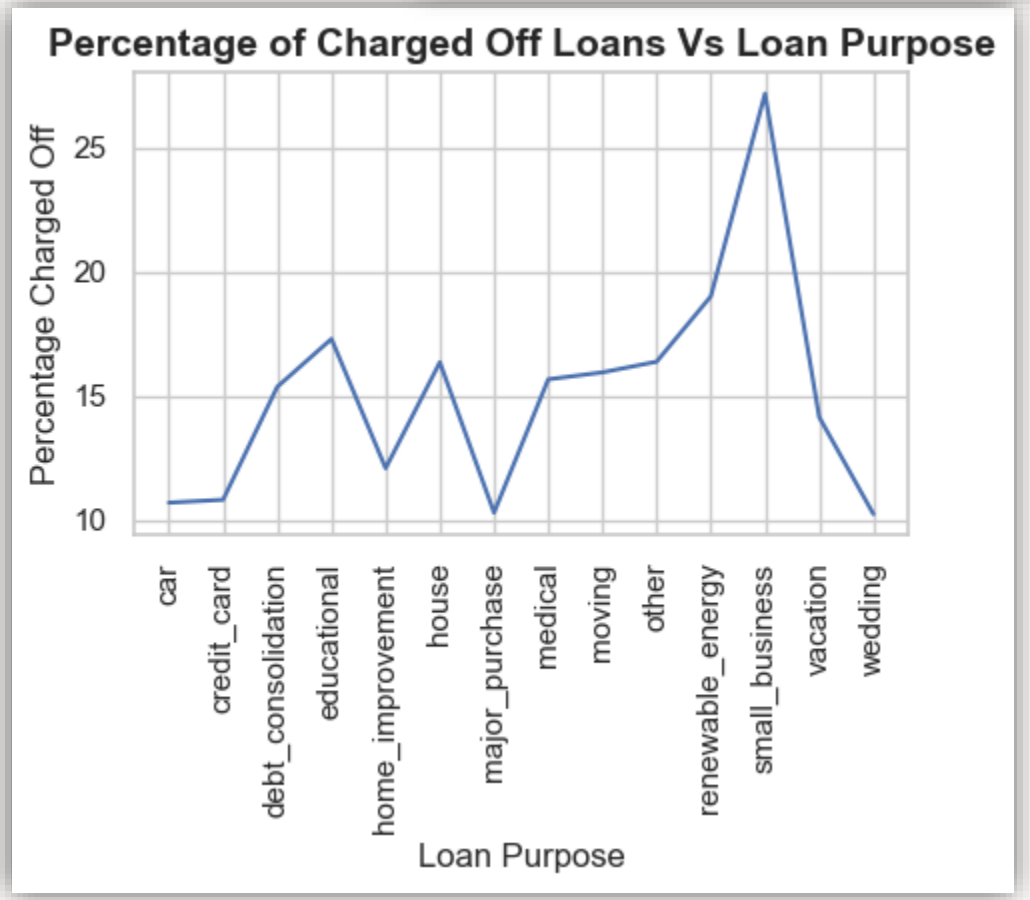
Loan taken towards **Small businesses** are more likely to be defaulted. This could be because of many risk of starting small business and establishing it. However, larger businesses could have all potentials/ assets for loan payment.

Renewable energy purpose for loan is also a critical point for loan defaulters.

So, lenders has to be additional careful in lending to these purposes.

Those who are not owning a house are more likely to become the defaulters.

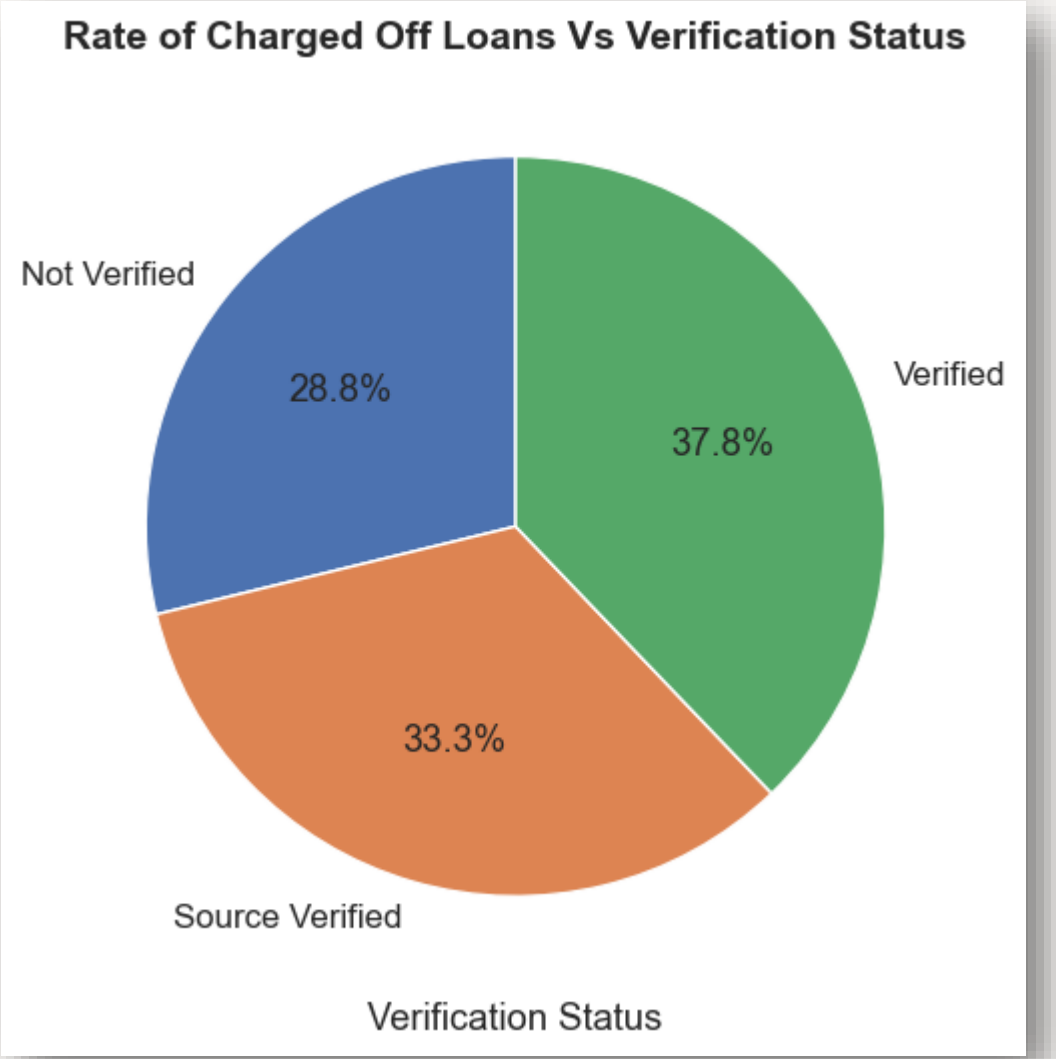




Verification Status Vs Charged Off

Observation:

Verification status alone may not be a clear-cut indicator of whether a loan will be charged off. Specifically, the percentages of charged-off loans are distributed somewhat evenly across different verification statuses.



DTI Vs Charged Off

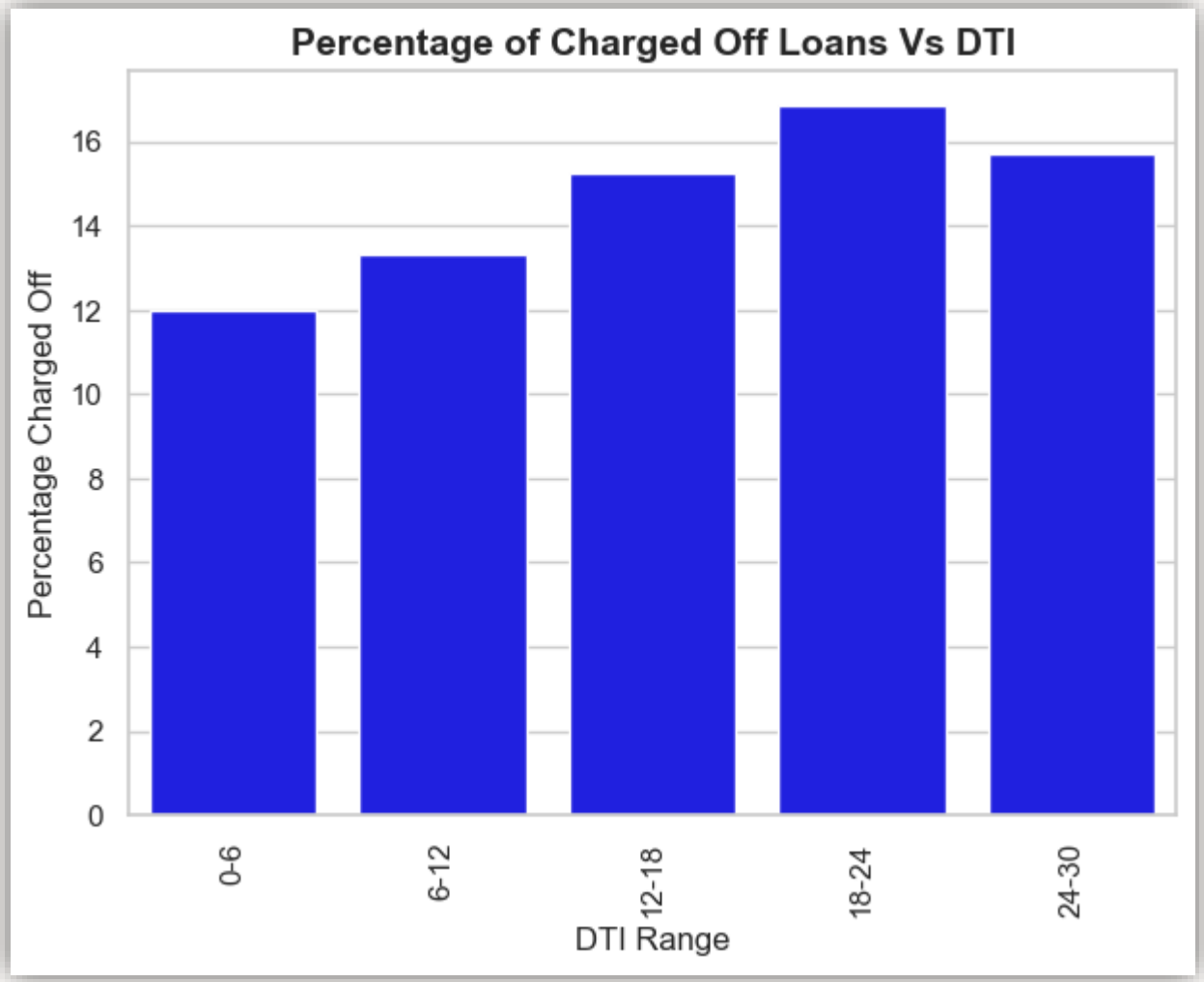
Observation:

Based on the data, it looks like higher the DTI larger the chances of charged off. However, there is an exception for the bracket 24-30 DTI.

Important:

High DTI has higher risk of defaults.

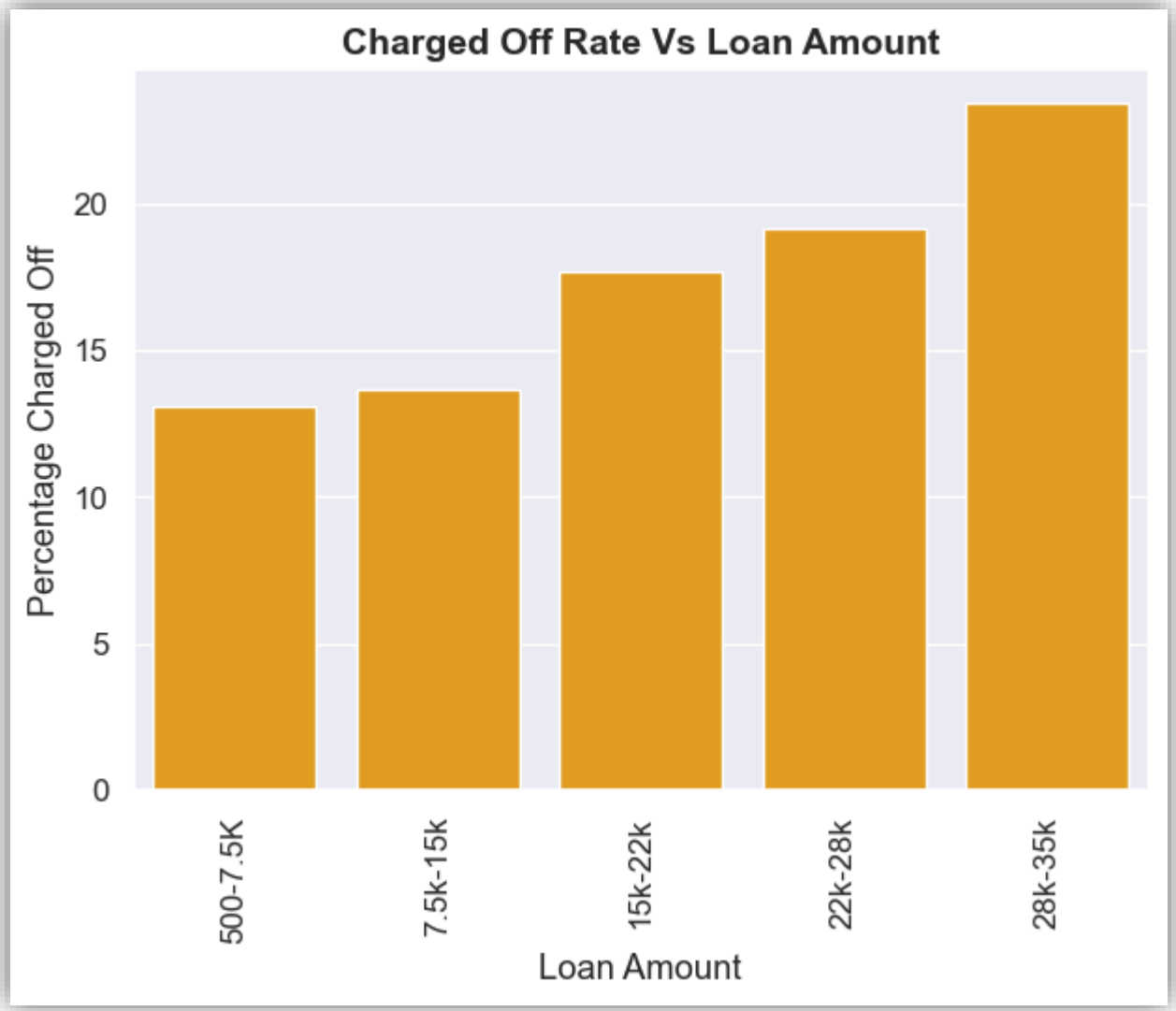
Noticed, for the DTI in range of 18-24 are higher the chances of charged off.



Total Loan Amount Vs Charged Off

Observation:

As per the report, higher the loan amount, more is the risk of defaults.

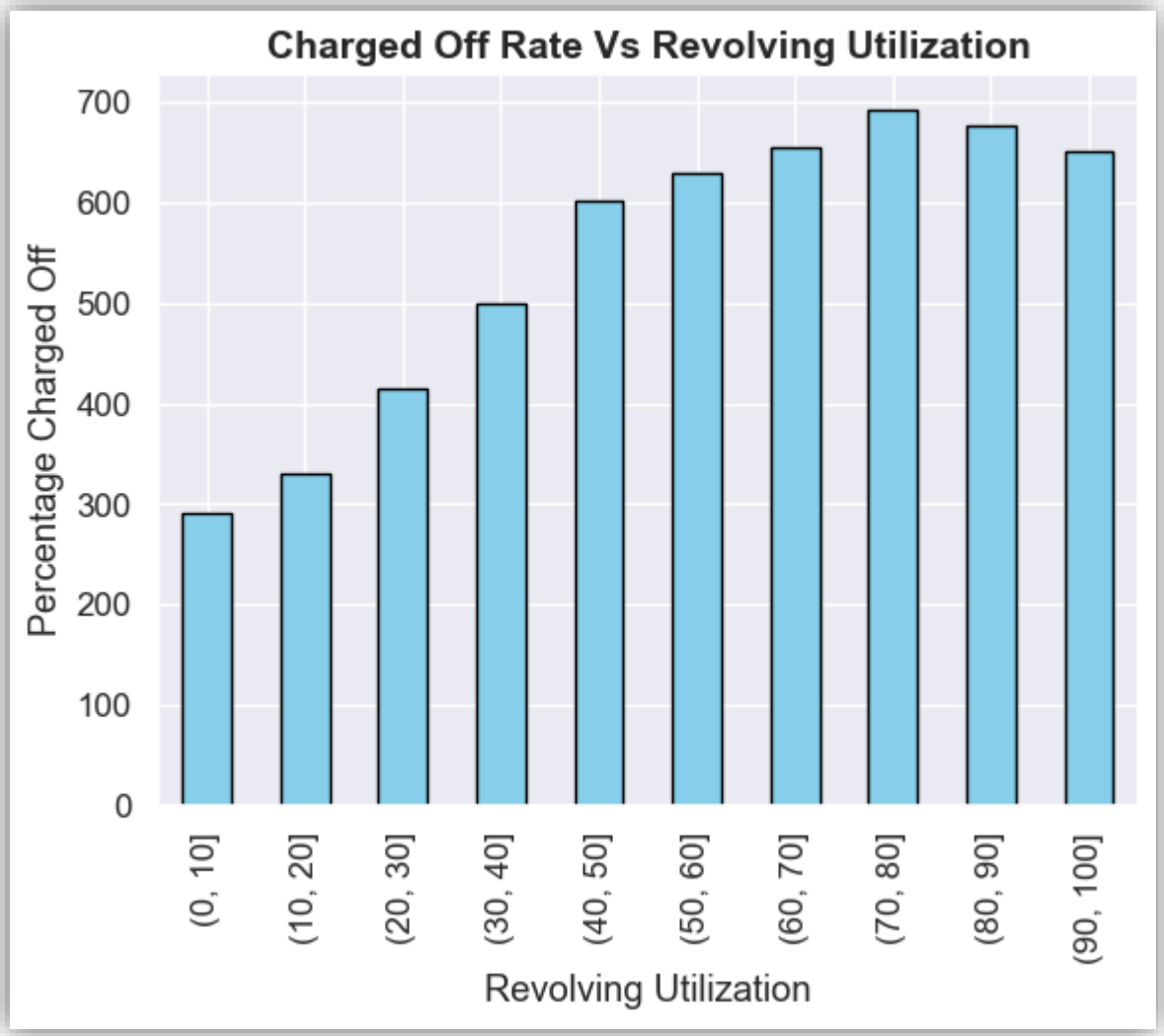


Revolving Utilization Vs Charged Off

Observation:

Having a high revolving utilization suggests a higher chance of default.

Interestingly, the risk plateaus after reaching 50%, indicating a critical threshold.

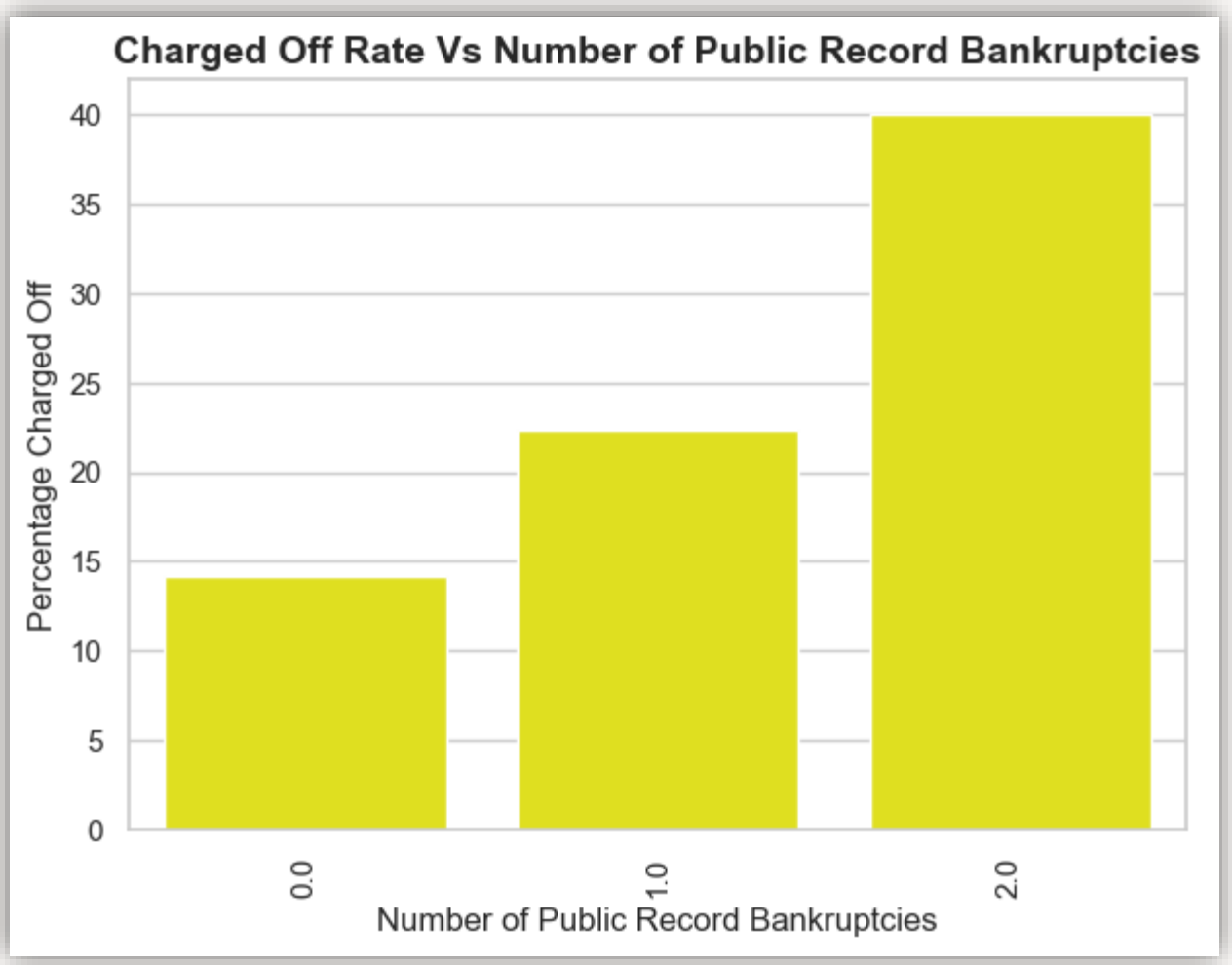


of Public Record Bankruptcies Vs Charged Off

Observation:

High the number of public bankruptcies has higher risk of defaults.

So, lenders has be careful providing loans to the applicates who has already declared bankruptcies in past.



Correlation Plot

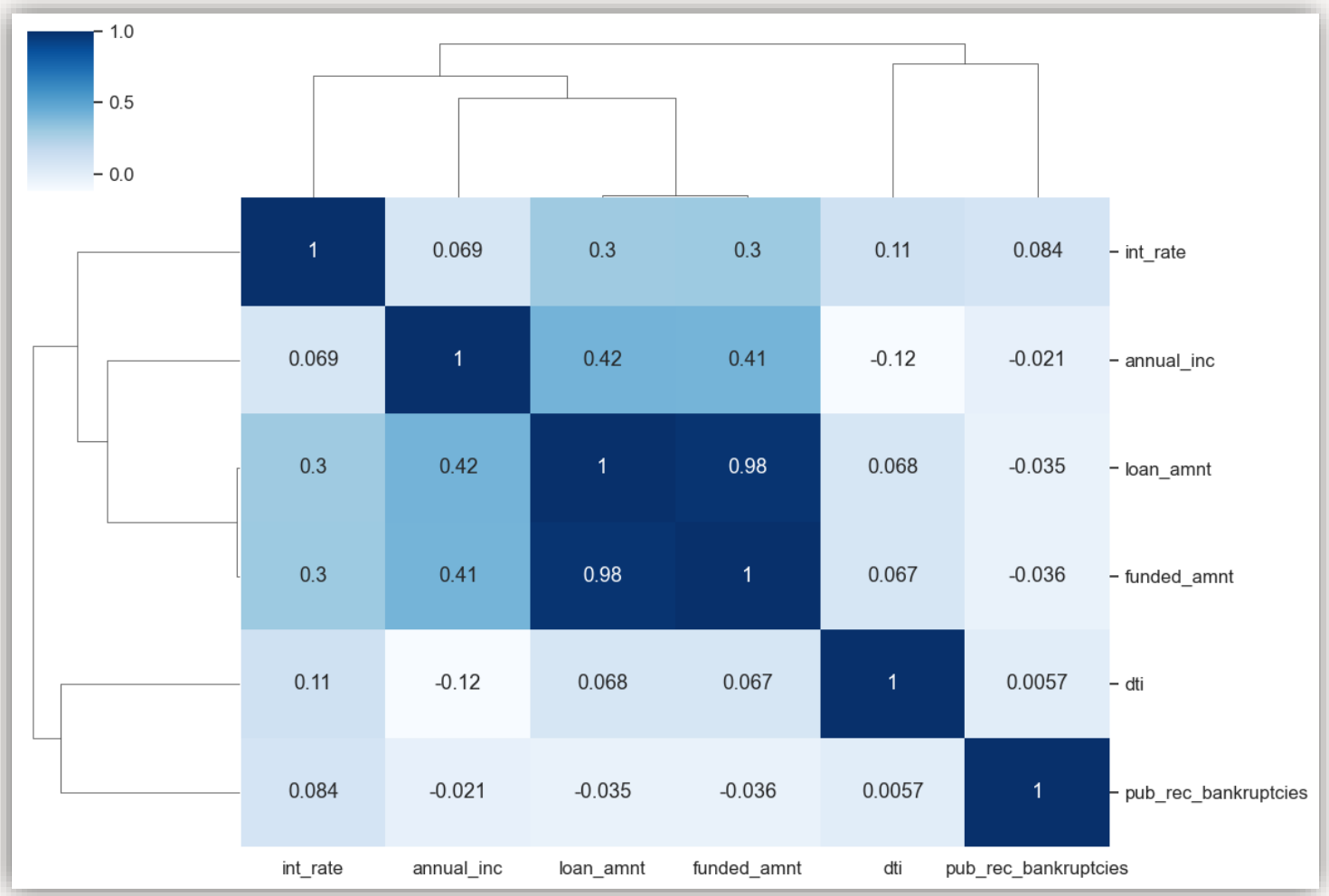
Observation:

Annual income has positive correlation with loan amount and funded amount. Applicants with higher incomes are likely to be approved for larger loans.

DTI has a weak negative correlation with annual income, implying a slight decrease in DTI as income rises.

Interest rates show a moderate positive correlation with loan amounts, indicating potential higher costs for larger loans.

Public bankruptcies have limited correlation with financial variables, suggesting a weaker impact.



Conclusion

- Higher loan amounts, particularly in the range of $15,000 to $35,000, correlate with an increased likelihood of default.
- Higher loan amount tend to increase the risk of defaulting.
- Annual incomes falling between $4,000 and $45,000 are associated with a higher risk of default, indicating a connection between lower income levels and repayment challenges.
- Loans for small businesses and renewable energy purposes demonstrate a higher tendency to default, possibly due to increased business risks.
- Applicants without property ownership, such as a house, exhibit a higher likelihood of becoming defaulters.
- Debt-to-Income Ratio (DTI), especially in range of 24-30, significantly raises the chances of loan default, emphasizing the importance of evaluating an applicant's debt burden.
- Credit grades F & G showcase higher default rates, warranting careful scrutiny and cautious approval for applicants in these categories. To be more specific, sub grade type F5 and G3 are very chance of being defaulters.
- A history of public bankruptcies increases the risk of default, emphasizing the need for thorough evaluation before approving loans for individuals with such backgrounds.

Thank you !