Loan Default Prediction: A Data-Driven Approach

Customer behavior analysis for loan granting

Prepared by: Saurabh Pandey
Shankar Leela

Objective

01

Study customer behavior for loan granting

02

Predict likelihood of default

03

Support decision-making: loan denial, amount adjustment, or higher interest rates for risky applicants

Case Study Understanding

This case study aims to provide insights into solving business problem through exploratory data analysis (EDA), focusing on risk analytics in banking and financial services. It highlights how data is used to minimize financial risk when lending to customers.

Business Understanding

- You are working for a consumer finance company that lends various types of loans to urban customers. The company must decide on loan approvals based on applicants' profiles, facing two main risks:
- Not approving a loan may result in lost business if the applicant can repay. Approving a loan may lead to financial loss if the applicant is likely to default. The dataset includes information on past loan applicants, indicating whether they defaulted. The goal is to identify patterns that predict loan defaults, which can guide decisions like denying loans, reducing loan amounts, or adjusting interest rates for risky applicants.
- Decision Scenarios: When a loan is applied for, the company can either:
- Approve the Loan: Fully paid: The applicant repays the loan completely. Current: The applicant is still paying installments. Charged-off: The applicant defaults on the loan. Reject the Loan: No data is available for these applicants since they did not proceed with the loan.

Business Objectives

- As the largest online loan marketplace, the company seeks to minimize credit loss, which occurs when borrowers default on loans. The aim is to identify risky applicants labeled as 'charged-off' to reduce potential losses.
- By understanding the key variables influencing loan defaults, the company can enhance its risk assessment and portfolio management. Additionally, conducting independent research on risk analytics is recommended to deepen your understanding of the relevant variables and their significance.

Data Preparation - Steps

- **Handling Missing Values**

Step 1: Identifying Missing Values

- Analyze and drop columns with excessive null values
- Ensure no null columns are left after processing

Step 2: Removing Duplicates

- Check for duplicates in the dataset and remove if found

- **Dropping Unnecessary Columns**

Step 1: Remove Unique/Descriptive Columns

- Columns with unique or text-based descriptions contribute little to analysis
- Focus on relevant columns for better accuracy

Step 2: Limit Analysis to Loan Grade Level

- Subgrade level detail removed

- **Correct Data Types & Derived Columns**

- Convert columns to appropriate data types
- Generate derived columns for enhanced insights

Univariate Analysis

- **Purpose: Understand individual features of the data**

- Graphical overview of all columns
- Handle outliers and create appropriate buckets for continuous variables

Observations

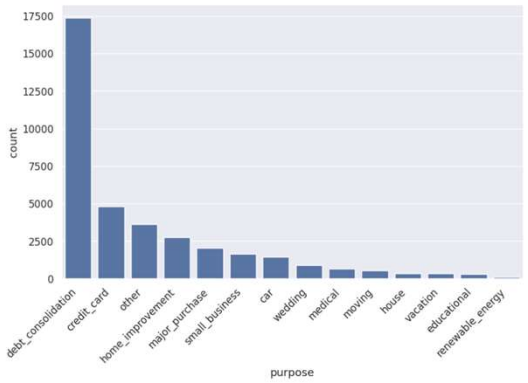
- **Loan Amount:**
  - Majority in range: 5.5k - 15k
  - Max: ~35k
- **Funded Amount:**
  - Most in range: 5.4k - 15k
  - Max funded amount: ~35k
- **Interest Rate:**
  - Average: 11.8%
  - Min: 5.4%, Max: 24.4%
  - Common range: 8.9% - 14.52%

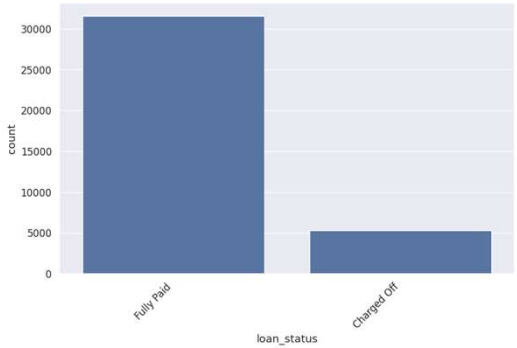
Categorical Variable Analysis

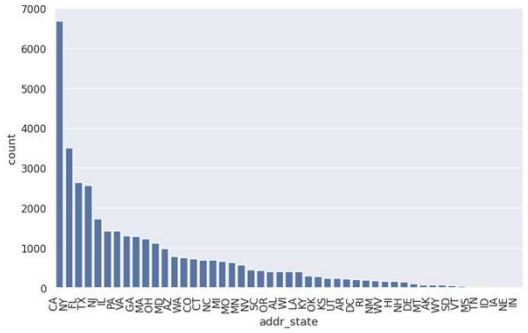
Observations

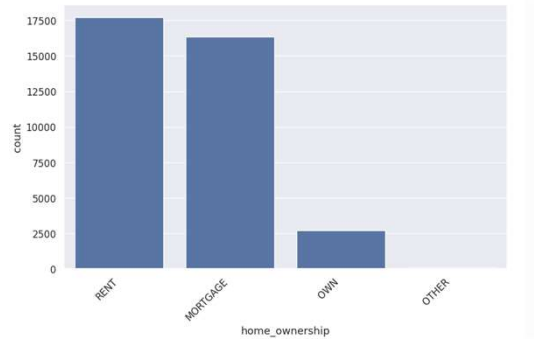
- Experience: Majority of applicants have 10+ years of experience
- Living Situation: Most applicants are either renting or have a mortgage
- Loan Purpose: Debt consolidation is the primary reason for applying
- Location: Most applicants are from California (CA)
- Loan Grade: Majority of loans are of Grade B
- Verification Status: 'Verified' applicants have a higher likelihood of default

Categorical Variable Analysis





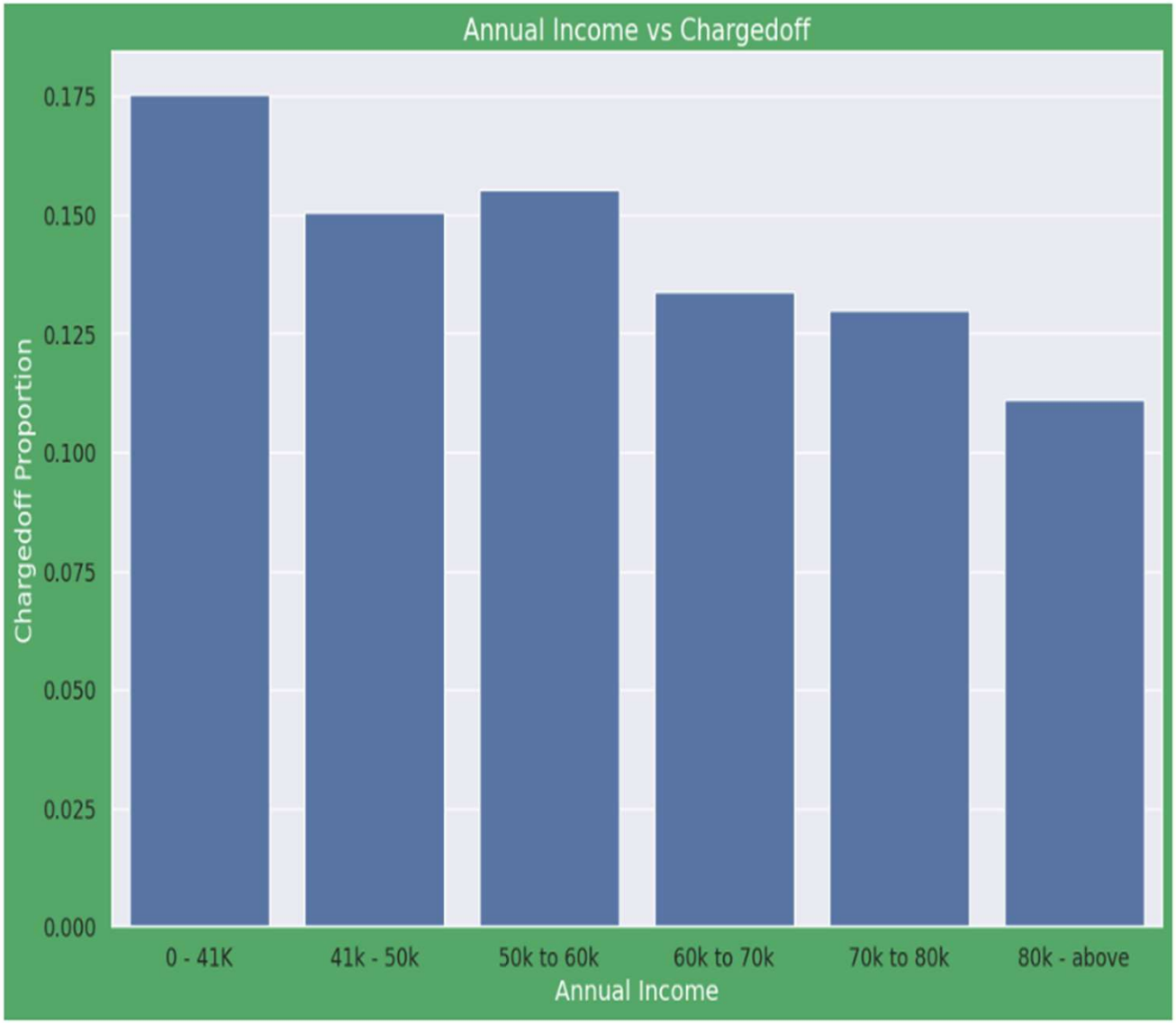




Annual Income Vs Charged off

Observations:

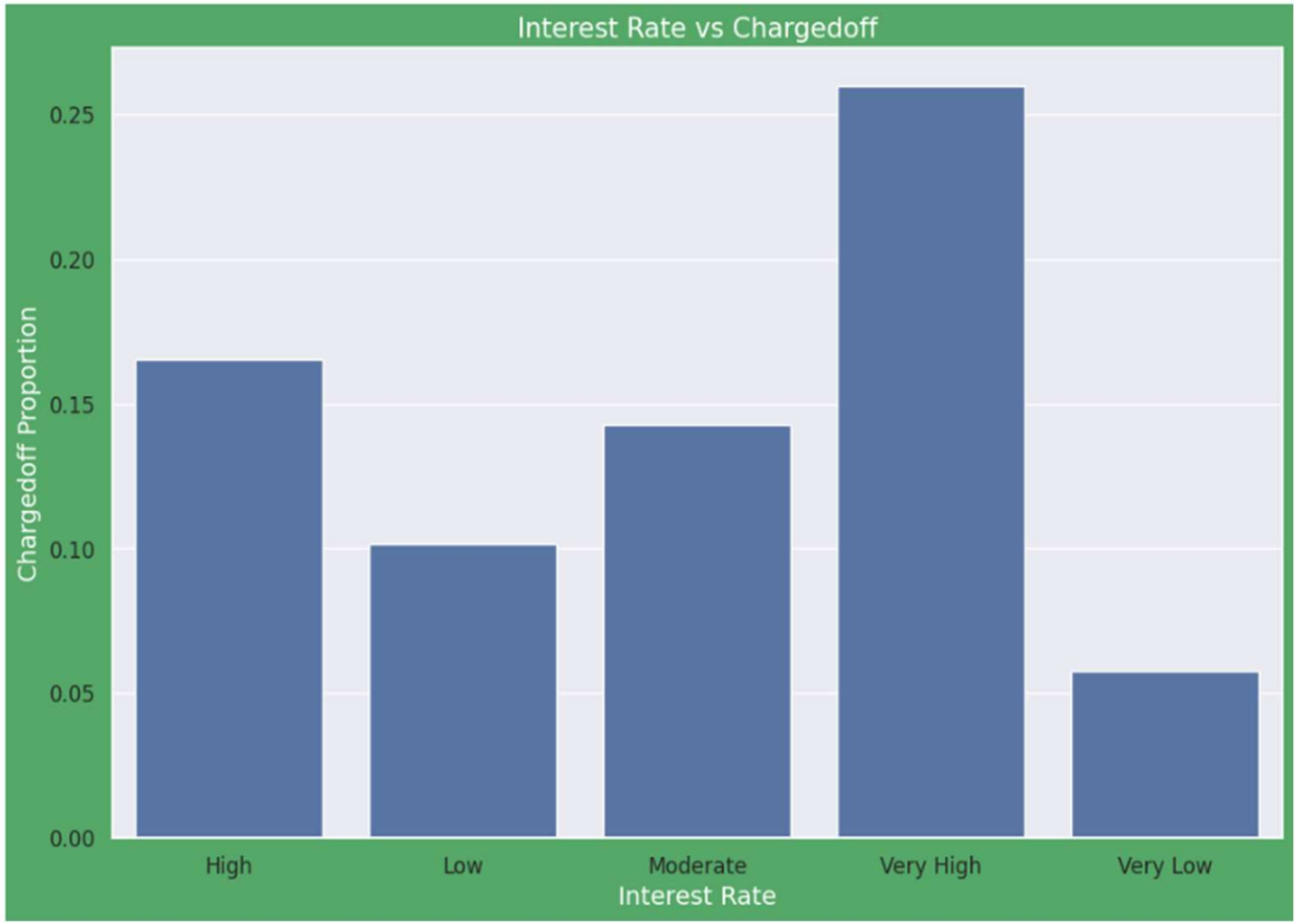
1. Income range 80000+ has less chances of charged off.
2. Income range 0-80000 has high chances of charged off.
3. Increase in annual income decreases charged off proportion.



Interest Rate Vs Charged off

Observations:

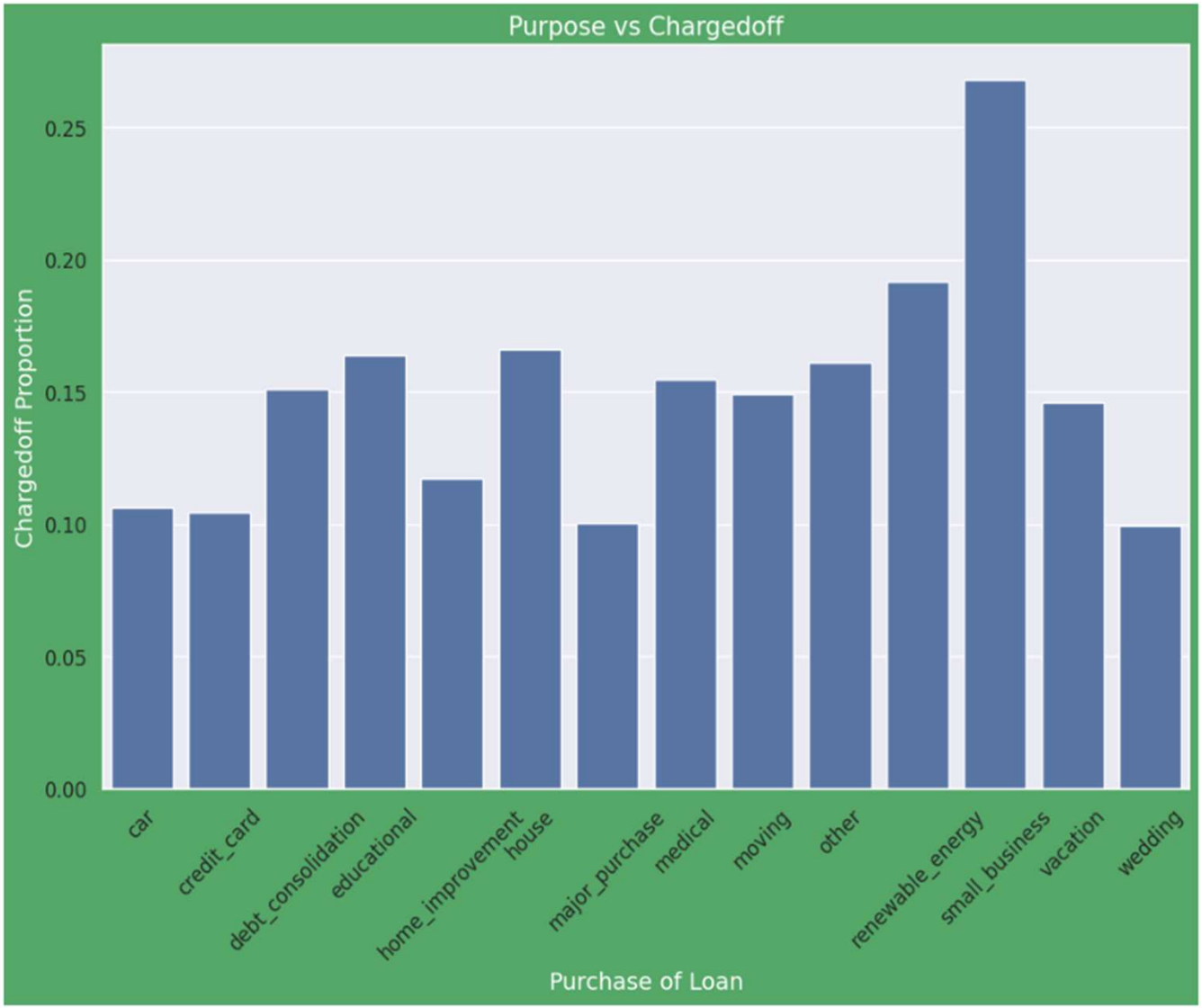
1. Interest rate less than 10% or very low has very less chances of charged off. Interest rates are starting from minimum 5 %.
2. Interest rate more than 15% or very high has good chances of charged off as compared to other category interest rates.
3. Reason could be they might be failing to pay the installments with higher interest rate.
4. Charged off proportion is increasing with higher interest rates.



Purpose Vs Charged off

Observations:

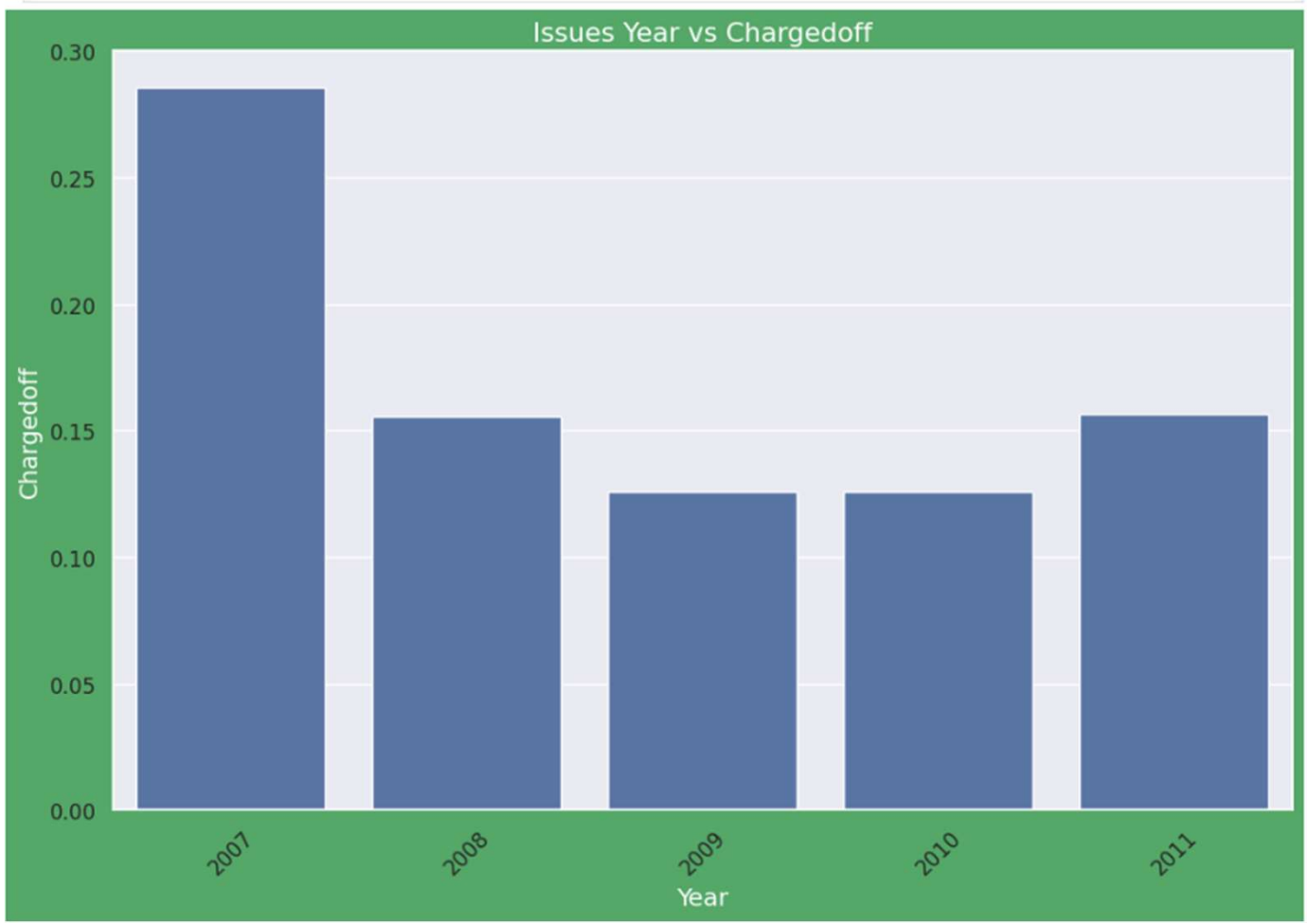
1. Those applicants who is having home loan is having low chances of loan defaults.
2. Those applicants having loan for small business is having high chances for loan defaults.



Issue Year Vs Charged off

Observations:

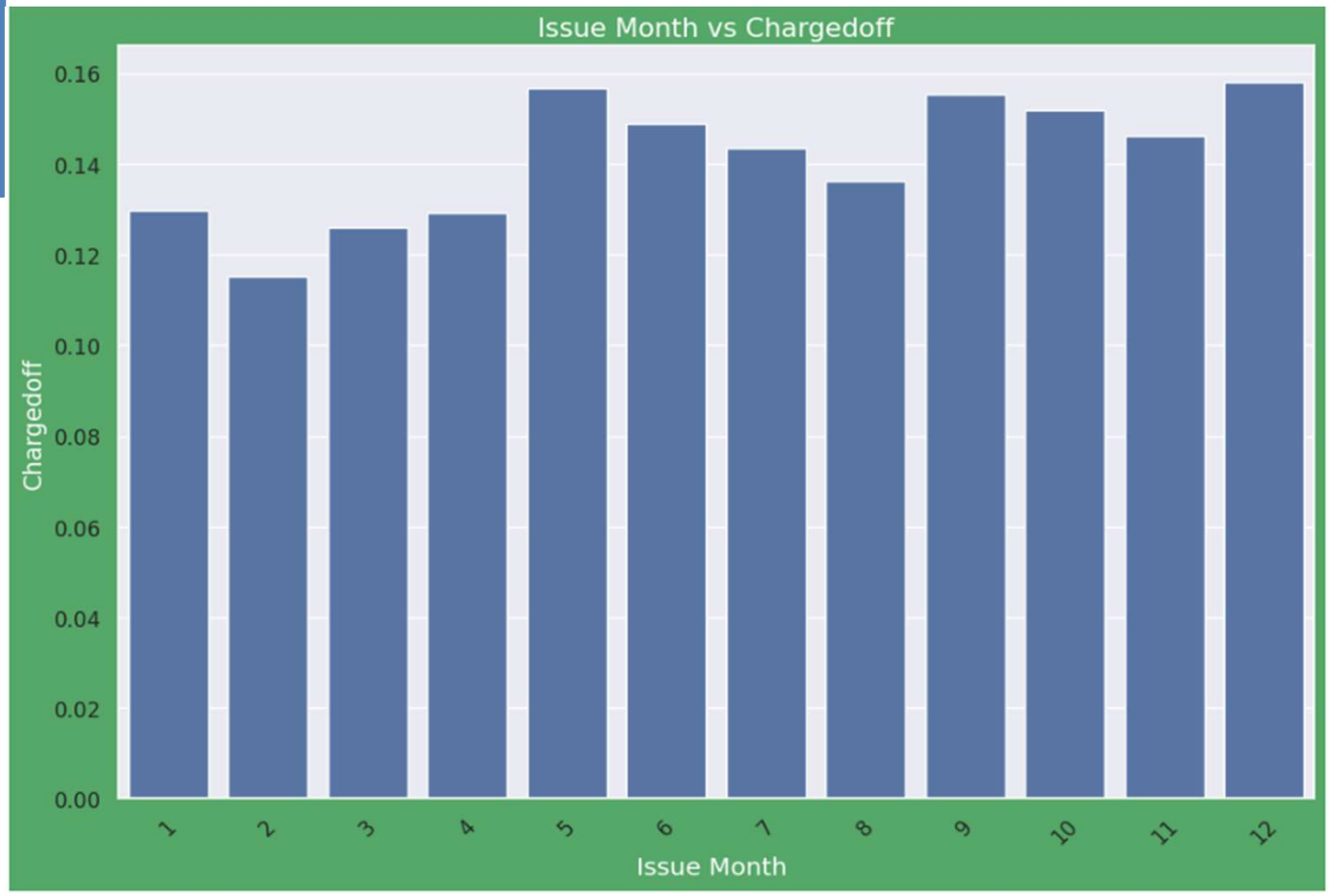
1. Year 2007 is having highest loan defaults.
2. Year 2009 is having lowest loan defaults.



Issue Month Vs Charged off

Observations:

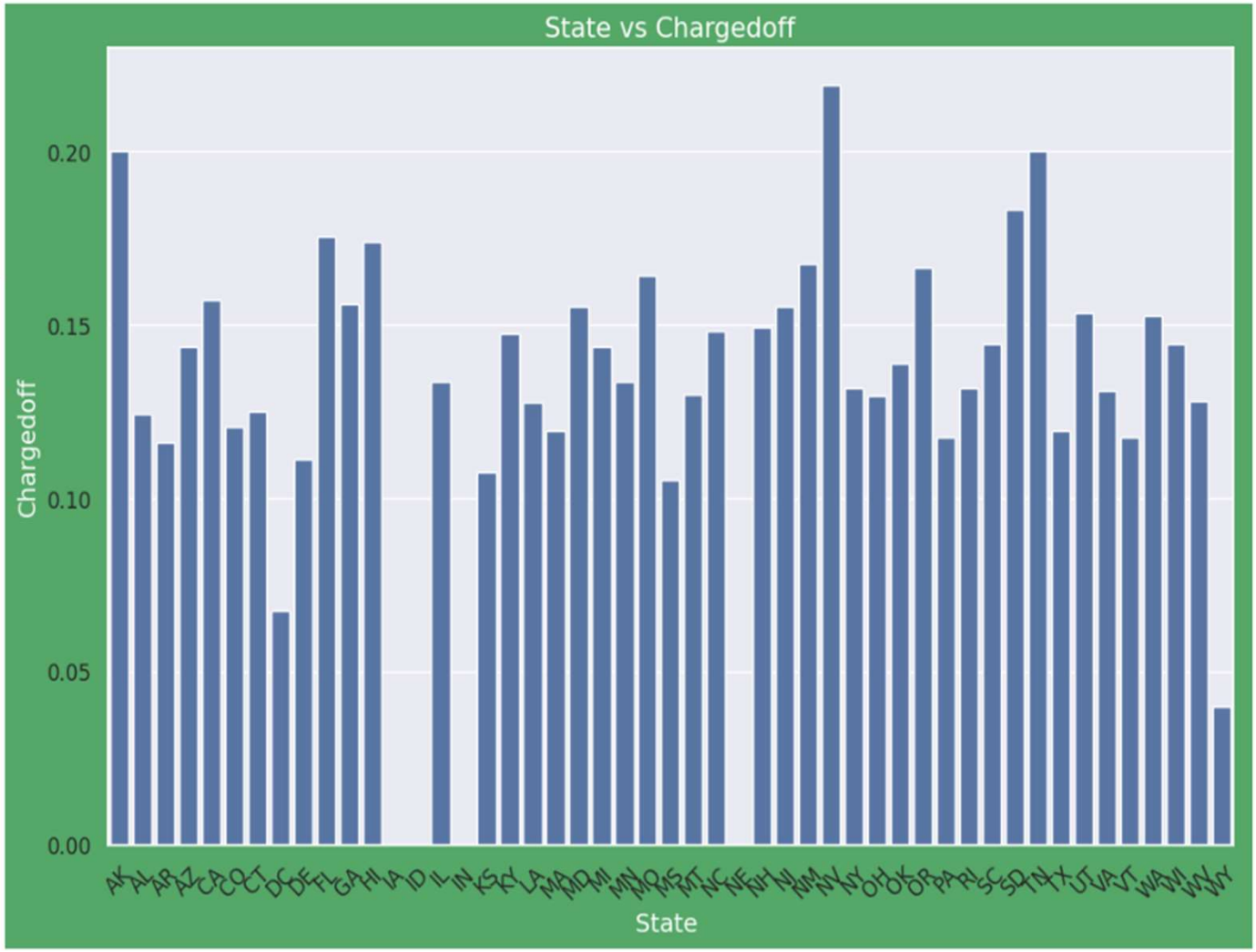
1. Those loan has been issued in May, September and December is having high number of loan defaults.
2. Those loan has been issued in month of February is having high number of loan defaults.
3. Majority of loan defaults coming from applicants whose loan has been approved from September to December.



State Vs Charged off

Observations:

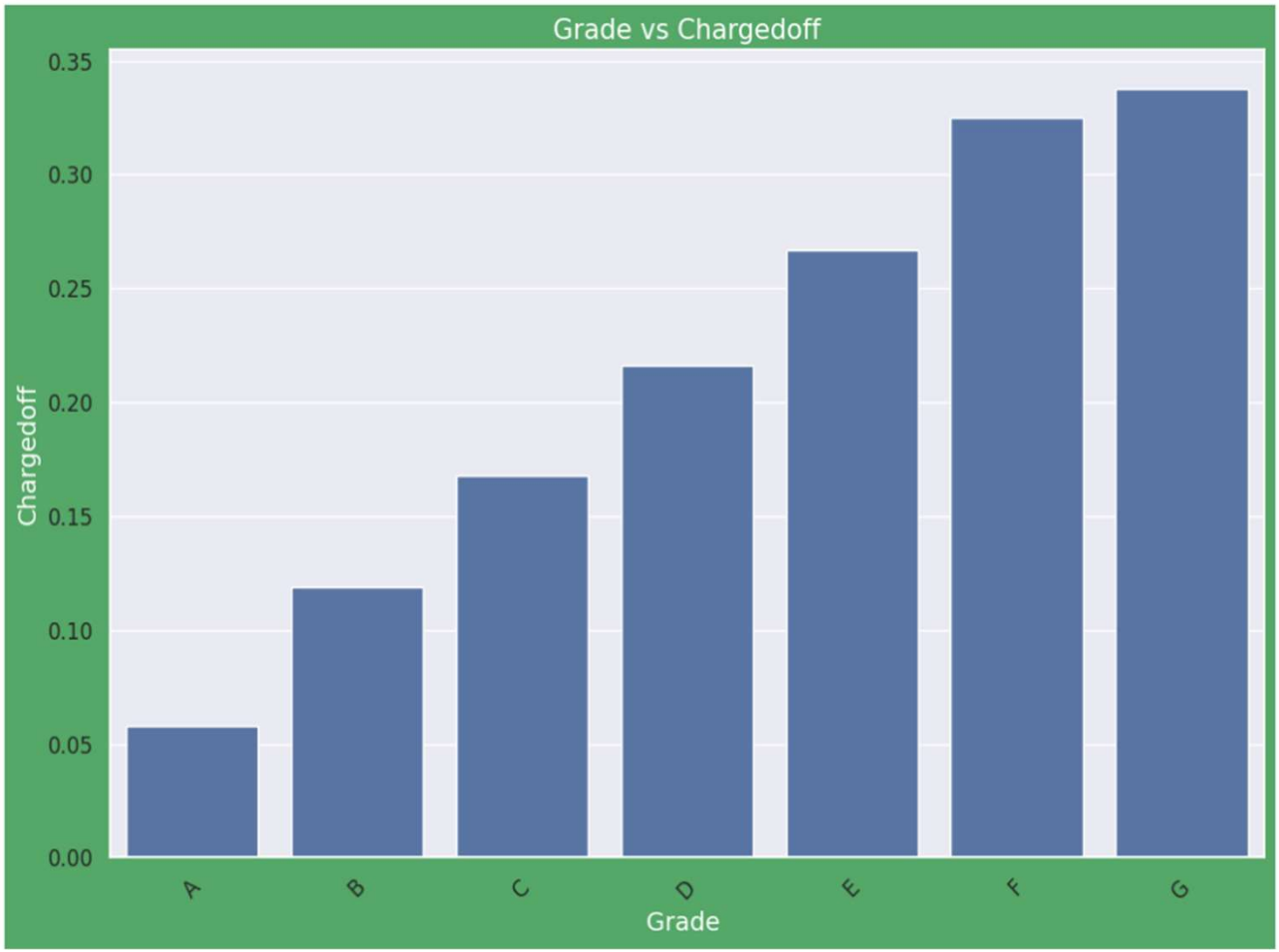
1. NV States is holding highest number of loan defaults.
2. WY is having low number of loan defaults



Grade Vs Charged off

Observations :

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



Positive Correlation:

1.Term has a positive correlation with loan amount.

2.Term has a positive correlation with interest rate.

3.Annual income has a positive correlation with loan_amount.

4. loan_amnt has strong positive correlation with Installment, funded_amnt, funded_amnt_inv.

Negative Correlation:

1.loan_amnt has negative correlation with pub_rec_bankrupticies.

2.annual income has a negative correlation with dti.

3.loan_amnt has negative correlation with delinq_2yrs