Lending Club Case Study
EDA and Risk Assessment

Prachi Chintan Shah
Pradeep Gadamsetty

Contents

Business Problem

- ❏ The Lending Club facilitates online loans for individuals.

- ❏ One major challenge the company faces is the risk of borrowers defaulting on their loans, leading to credit losses.

- ❏ Borrowers who fail to repay their loans cause financial losses, which the company aims to minimize. Identifying risky loan applicants is crucial for reducing the likelihood of defaults and protecting profits.

Business Analysis

Primary Goal:

- Use EDA to crack the patterns which helps to predict loan defaults.

Key Objectives:

- Load and Understand the data.
- Analyze customer and loan data to identify risky applicants.
- Determine which variables strongly indicate the risk of default.
- Add the insights for the company in minimizing financial losses by making more informed lending decisions.

Data Loading

The dataset consists of loan application data from past applicants. The attributes include personal information and loan details. The csv file contains 39717 rows and 111 columns.

Need to focus on key attributes which are:

- ❏ **Loan amount**
- ❏ **Interest rate**
- ❏ **Loan status**
- ❏ **Employment length**
- ❏ **Grade**
- ❏ **Term**

Data Understanding

- ❏ **Loan amount:** The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
- ❏ **Interest rate:** Interest rate on the loan.
- ❏ **Loan status:** Current status of the loan. It consists of three values: Fully paid, Charged off and Current. For this case study analysis, only first two values will be required in further analysis.
- ❏ **Grade:** LC assigned loan grade
- ❏ **Employment length:** Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- ❏ **Term:** The number of payments on the loan. Values are in months and can be either 36 or 60.

Data Cleaning

While data cleaning, we focused on handling missing/null values or incorrect values.

Data cleaning strategy:

❏ Dropped columns with more than 60% missing data to reduce columns that will be of no use in analysis.

❏ Converted string-based categorical values like percentages, months, years, special characters and employment length to numerical values for analysis.

❏ Reformatted dates to ensure consistency and accuracy in the dataset.

❏ Check for the duplicated rows and columns if any.

❏ Imputed the values.

❏ Remove the outliers.

After Data cleaning and deriving the columns we were left with **38577** rows and **25** columns.
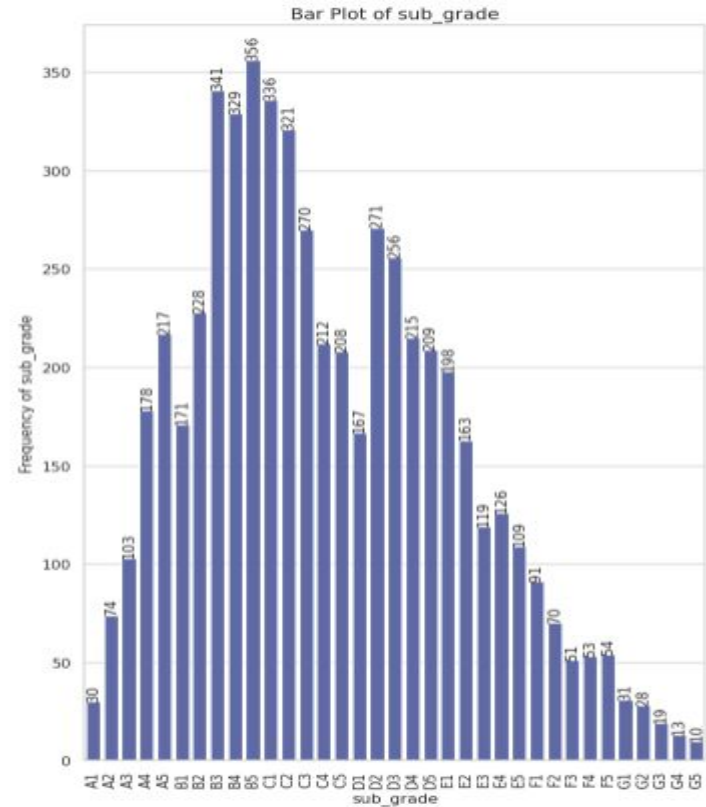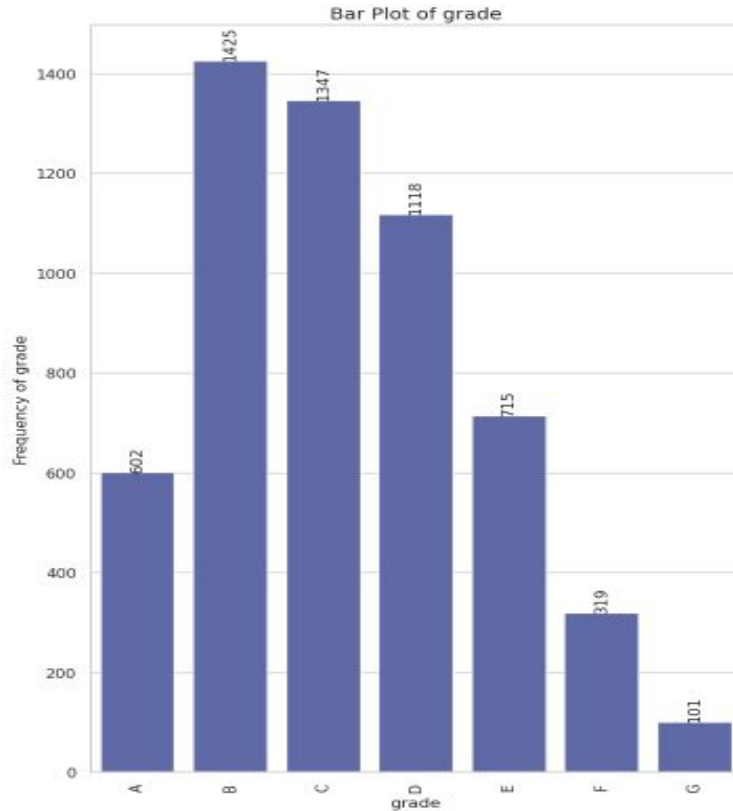
Univariate Analysis

Univariate analysis deals with analysing variables one at a time. We performed univariate analysis on individual attributes to understand their distribution and properties:
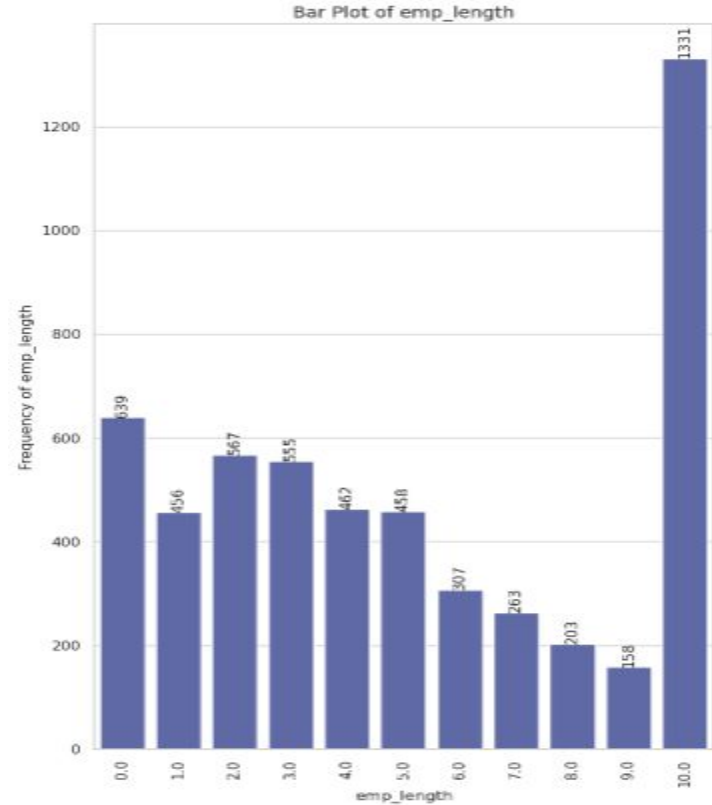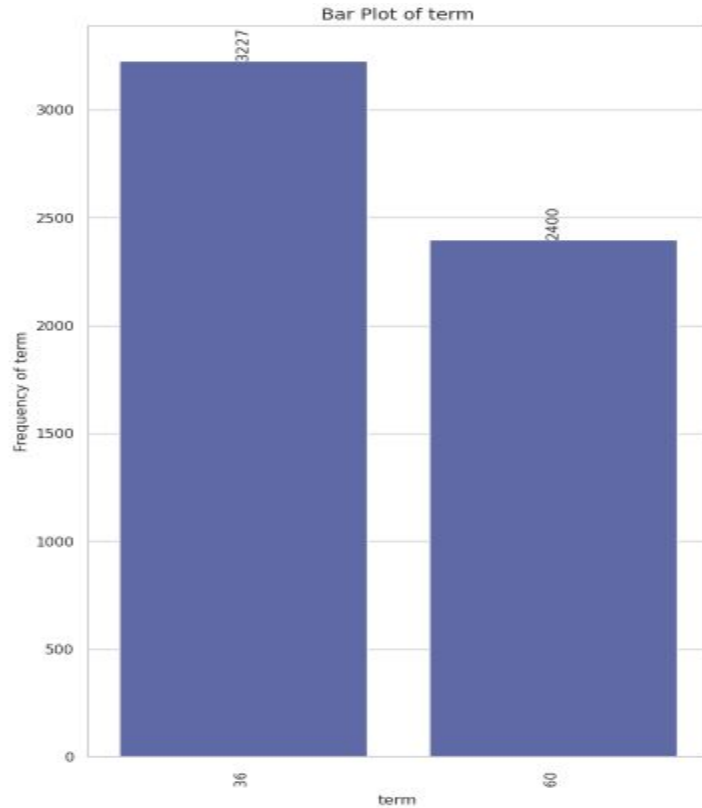
Categorical	
Ordered	**Unordered**
Grade	Addr state
Subgrade	Loan Purpose
Term	Home ownership
Emp length	Loan status
Issue year and month	

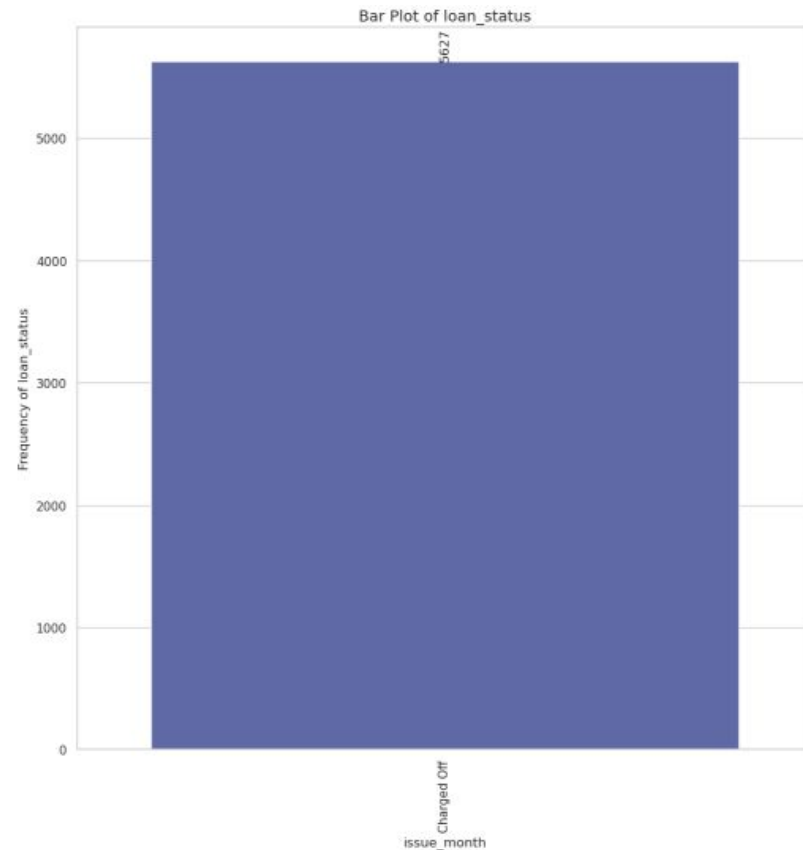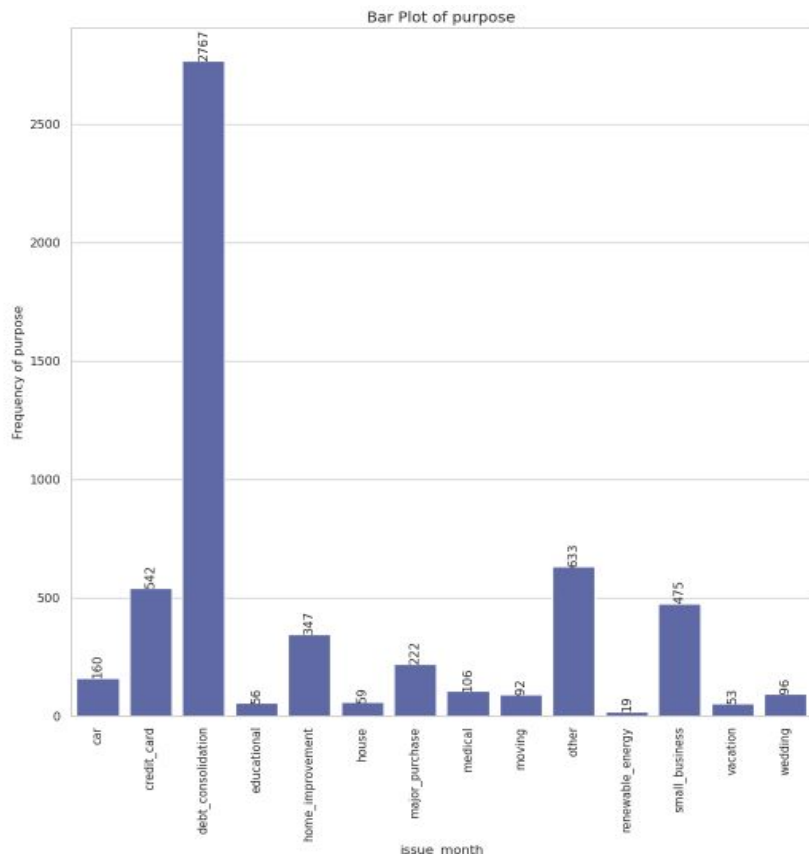| Quantitative |
|---|
| Int rate |
| Annual income |
| Loan Amount |
| Funded amount |
| DTI |
| Monthly installment |

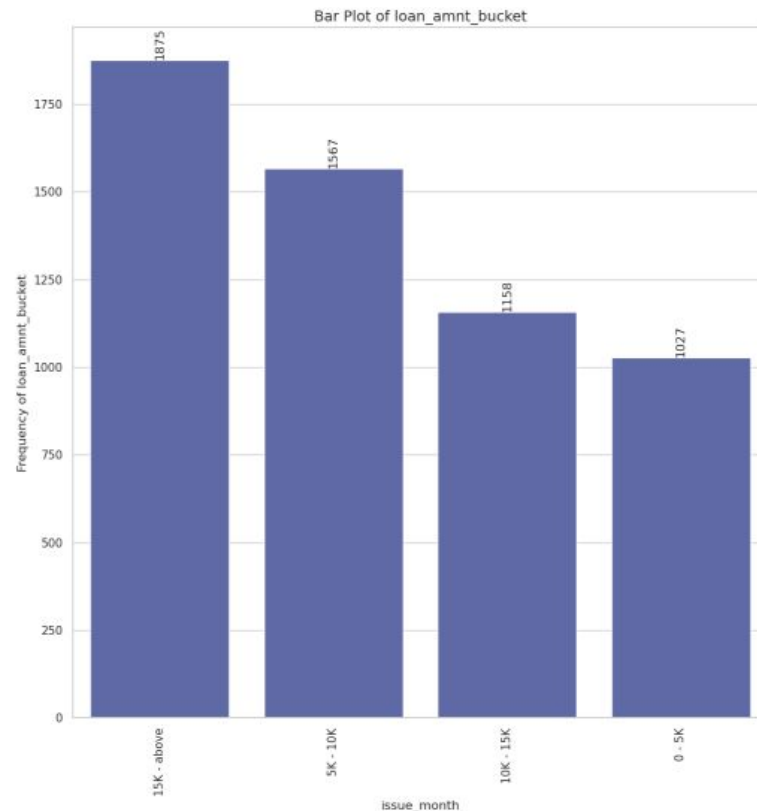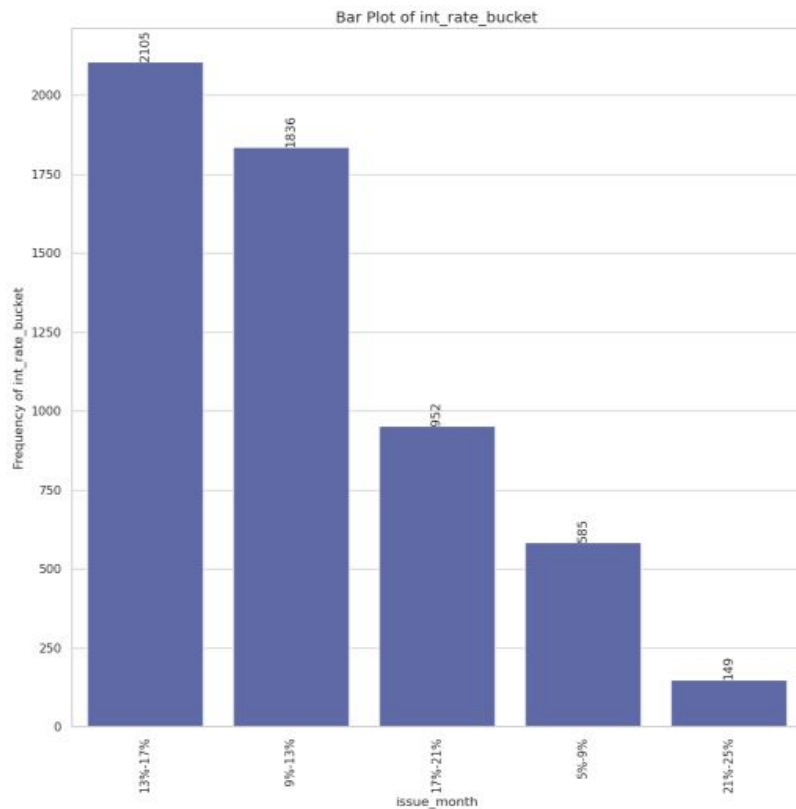# Bar plot of Grade and SubGrade -Ordered

Bar plot of term and emp_length - Ordered

Bar plot of loan purpose and status - Unordered

Bar plot of int_rate and loan_amt- Quantitative

Univariate Analysis Inferences

Ordered Categorical Variables

❏ Grade B had the highest number of "Charged Off" applicants, totaling 1,425, suggesting that borrowers in this grade experienced significant difficulty in repaying their loans.

❏ A majority of "Charged Off" applicants, around 3,227, had opted for 36-month loans, indicating a trend where those who defaulted tended to choose shorter loan terms.

❏ Borrowers with over 10 years of work experience accounted for the most "Charged Off" loans, with a total of 1,331. This demonstrates that extensive work experience alone didn't guarantee successful loan repayment.

❏ The year 2011 saw the highest number of "Charged Off" loans, with 3,256 instances. This could reflect broader economic challenges or financial instability during that period.

❏ Dec month saw the highest number of "Charged Off" loans with 678.

Univariate Analysis Inferences

Unordered Categorical Variables

❏ CA California led in terms of "Charged Off" applicants, with 1,125 borrowers defaulting on their loans. The company should consider stricter lending criteria for applicants from states with higher default rates like California.

❏ Debt consolidation was the most common loan purpose among "Charged Off" borrowers, with 2,767 applicants. Loans for this purpose need more careful evaluation, as it is a significant factor in loan defaults.

❏ A large portion of the "Charged Off" borrowers, totaling 2,839, lived in rented properties. The financial stability of applicants living in rented homes should be closely assessed, as they may be more vulnerable to economic fluctuations.

❏ Total of 5627 of loan status has Charged off.

Univariate Analysis Inferences

Quantitative variables

❏ A total of 1,595 "Charged Off" applicants had annual incomes below $40,000. For borrowers in this income bracket, the company should conduct stringent income verification and assess repayment capabilities to minimize the risk of default.

❏ The largest group of defaulting borrowers (2,105) fell into the interest rate range of 13%-17%. Offering more competitive, lower interest rates may help reduce the risk of defaults.

❏ A substantial number of defaulting applicants (1,875) had loan amounts of $15,000 or more. Higher loan amounts should be granted with caution, ensuring that borrowers have strong credit histories and proven repayment abilities.

❏ Among the "Charged Off" borrowers, 1,782 received funded amounts of $15,000 or above. The company must carefully assess whether the funded amounts align with the borrower's financial capacity to repay larger loans.

❏ High DTI ratios were common among defaulters, with 1,208 "Charged Off" borrowers having elevated DTI ratios. The company should enforce stricter DTI ratio thresholds to prevent lending to borrowers with excessive debt burdens.
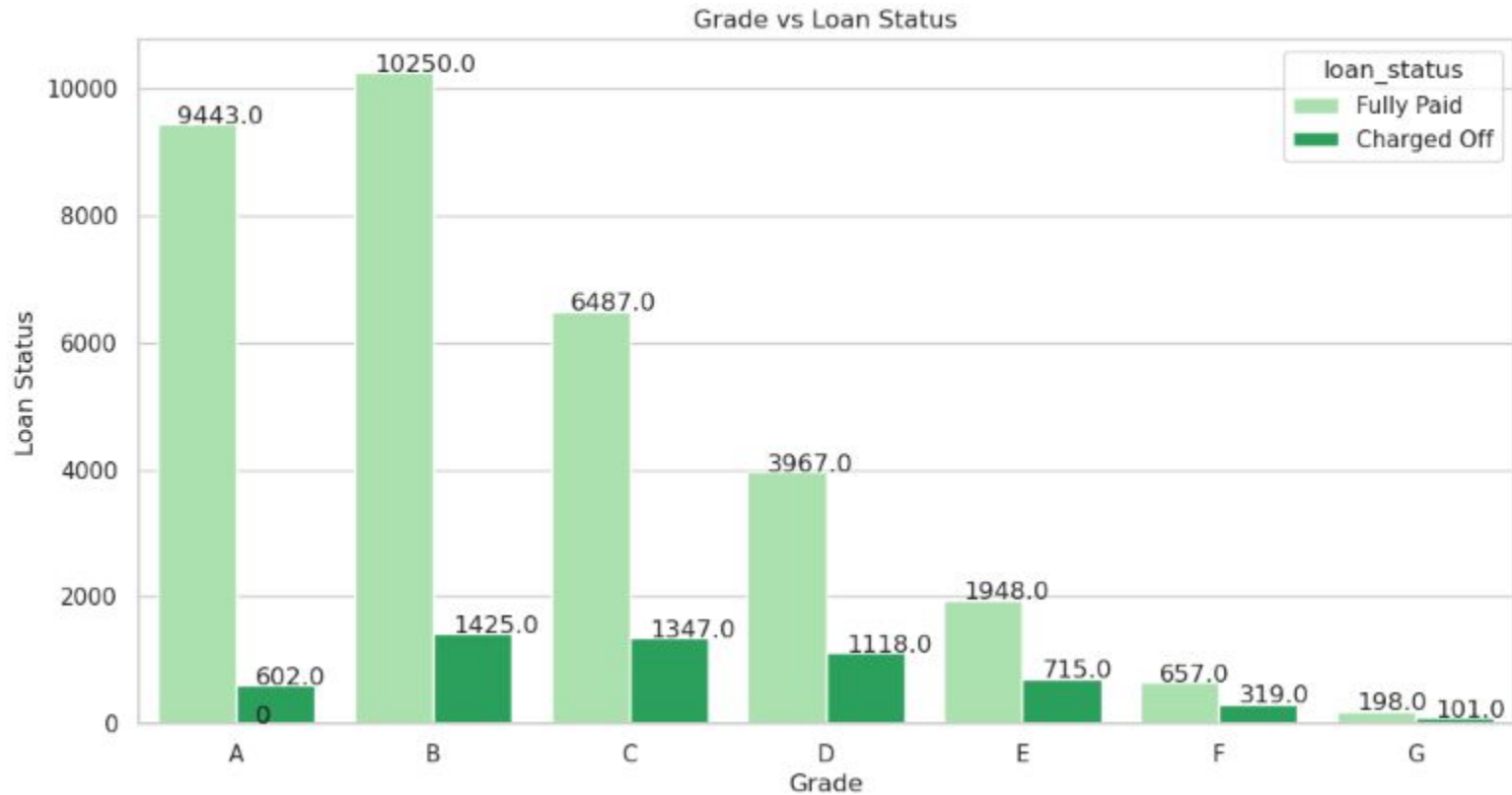
Bivariate Analysis

Bivariate analysis is to understand the relationship between two variables. We explored relationships between different features to identify correlations and patterns related to loan default:

| Categorical		Quantitative
Ordered	**Unordered**	Annual inc and loan status
Grade and loan status	Purpose and loan status	Funded amt and loan status
Subgrade and loan status	Home ownership and loan status	Int rate and loan status
Term and loan status	Verification and loan status	Dti and loan status
Emp length and laon status	Addr state and loan status	Loan amt and loan status
Loan year and loan status		
Loan month and loan status		

Grade VS Loan status - Ordered

Loan purpose VS Loan status - Unordered
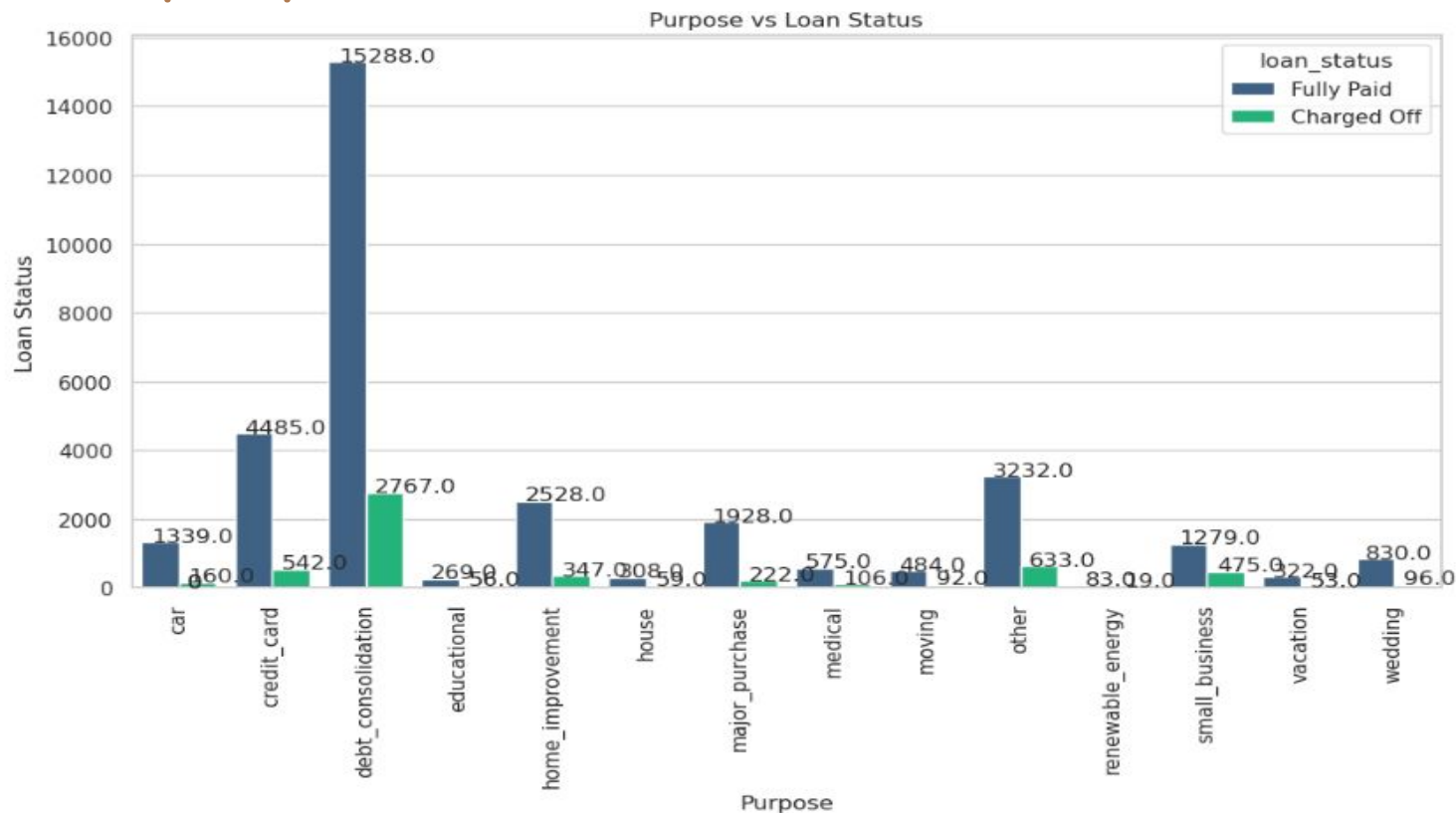
Int rate VS Loan status - Quantitative
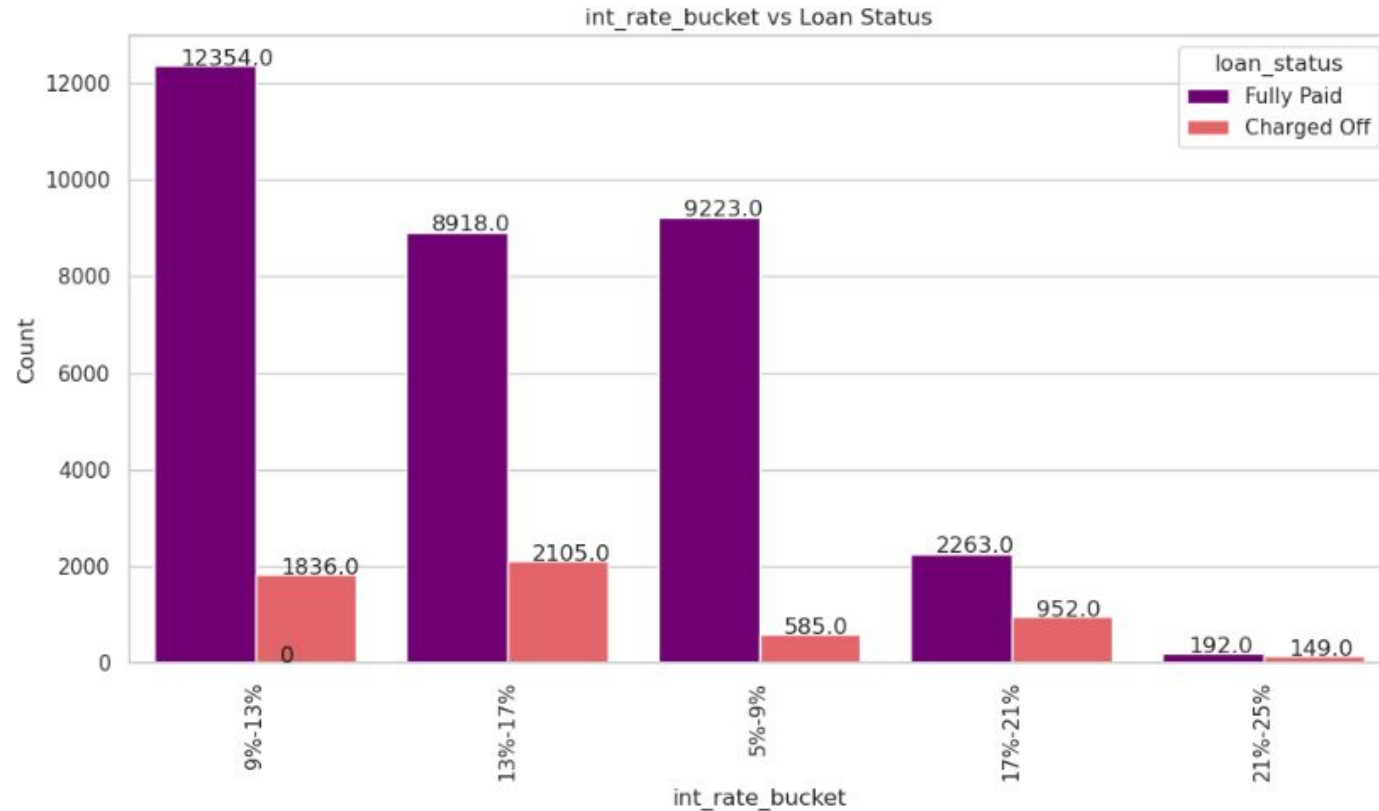
Bivariate Analysis Inferences

- ❏ **Risk Mitigation for Grades**: Since loan applicants from Grades B, C, and D contribute to most of the "Charged Off" loans, the company should tighten risk evaluation and apply stricter underwriting criteria for borrowers in these grades.
- ❏ **Focus on Subgarades**: Applicants in Subgrades B3, B4, B5, C1 and C2 show a higher likelihood of default. The company might consider applying additional risk controls or offering these applicants lower loan amounts to mitigate risk.
- ❏ **Experience and Default Probability**: Loan applicants with ten or more years of experience are more likely to default. This suggests that experience alone may not be a reliable indicator of creditworthiness. The company should use a more comprehensive credit scoring system that factors in other risk-related attributes.
- ❏ **Growth in Loan Applicants**: The steady rise in loan applications from 2009 to 2011 shows market growth. The company should capitalize on this opportunity while reinforcing risk management strategies to stay competitive and secure.
- ❏ **Seasonal Loan Demand**: The company sees an uptick in loan applications in December and Q4, possibly driven by holiday spending. It should prepare for this seasonal demand by optimizing loan processing during peak periods.

Bivariate Analysis Inferences

- ❏ **Debt Consolidation Risk**: Since debt consolidation is the category with the maximum number of loans and high default rates, the company should carefully evaluate applicants seeking debt consolidation loans and potentially adjust interest rates or offer financial counseling services.
- ❏ **Housing Status and Default Risk**: Applicants living in rented or mortgaged houses are more likely to default. This information can be considered in the underwriting process to assess housing stability and its impact on repayment ability.
- ❏ **Verification Process Review**: Verified loan applicants are defaulting more than those who are not verified. The company should review its verification process to ensure it effectively assesses applicant creditworthiness and consider improvements or adjustments.
- ❏ **High Loan Amount Risk**: Applicants requesting loans of 5k-10k or more have a greater default risk. The company should consider conducting more thorough assessments for higher loan amounts and potentially limiting loan size for high-risk borrowers.
- ❏ **Int Rates**: High DTI ratios and interest rates in the range of 9% to 13% are linked to increased default rates. The company should revise how it determines interest rates and align them more closely with borrowers' DTI ratios to improve repayment capabilities.
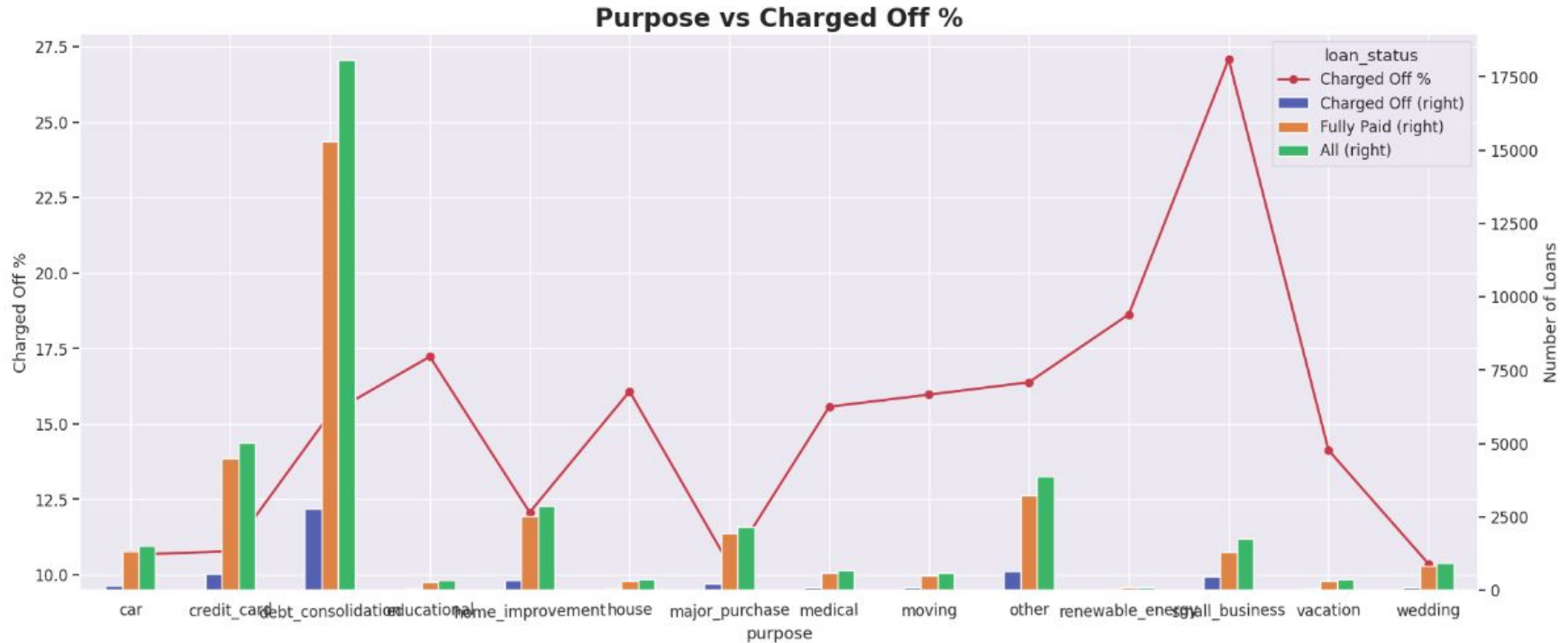
Multivariate Analysis

Multivariate against Charged Off
Grade
Sub Grade
Emp length
Addr state
Home ownership
Verification
Annual inc
DTI
Int rate

Grade and Charged Off
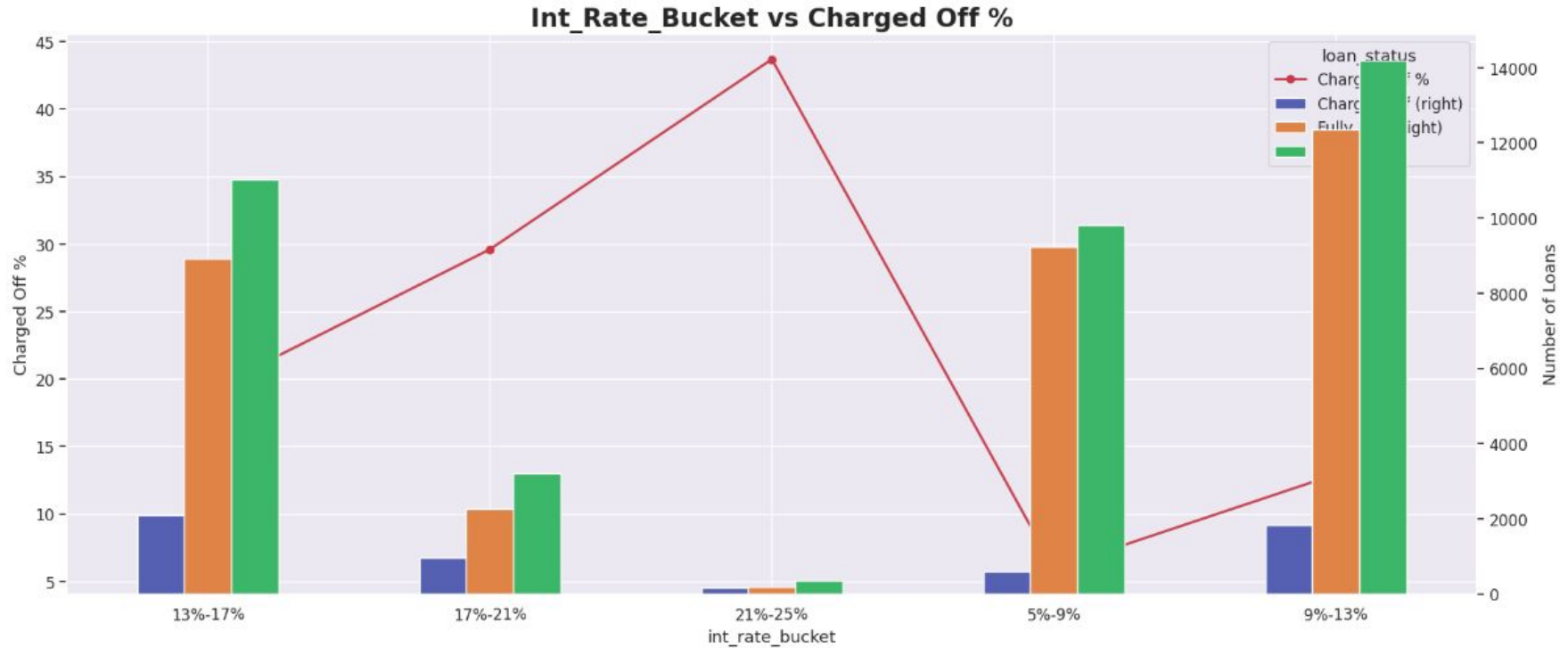

Grade vs Charged Off %
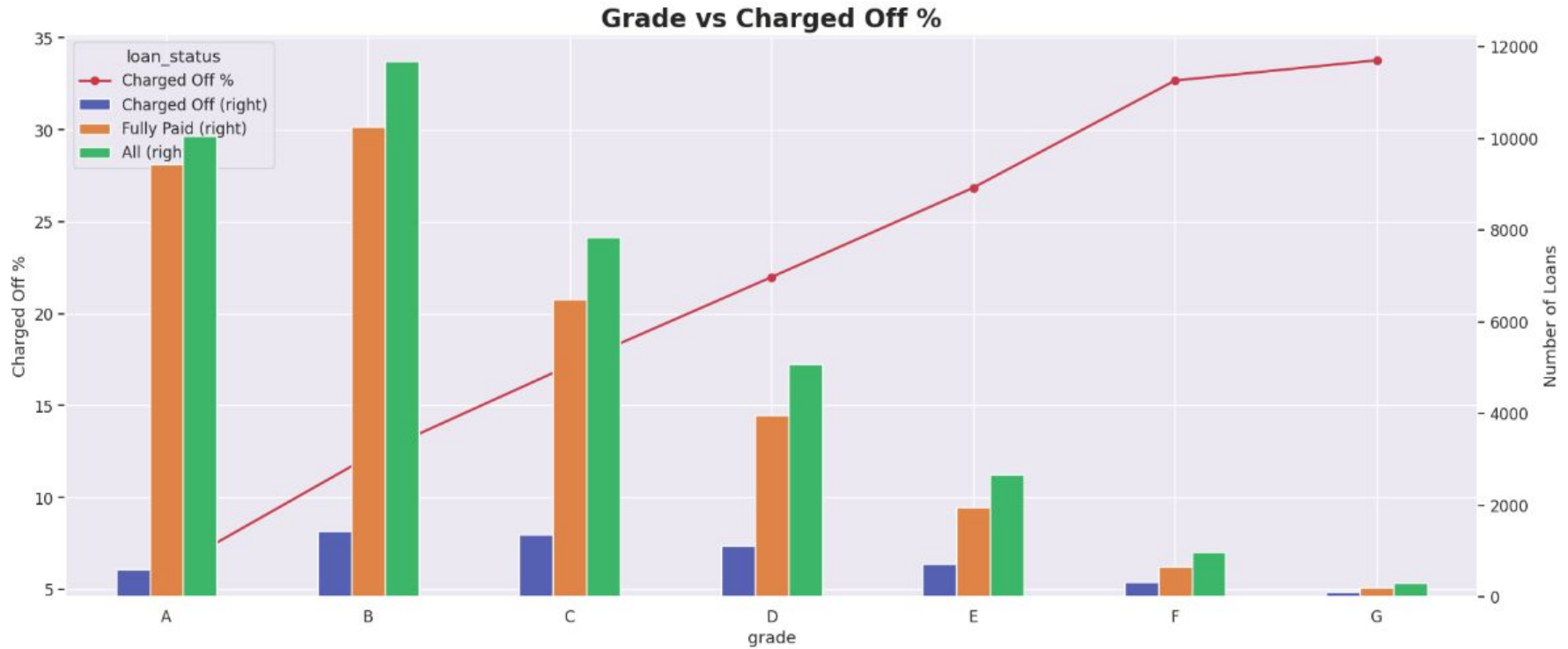
Purpose and Charged Off

Purpose vs Charged Off %

Int rate and Charged Off

Correlation

Strong Correlation

- ❏ installment shows a strong correlation with funded_amnt, loan_amnt, and funded_amnt_inv
- ❏ term shows a strong correlation with interest rate
- ❏ annual_inc shows a strong correlation with loan_amount

Weak Correlation

- ❏ dti shows weak correlation with many fields
- ❏ emp_length shows weak correlation with many fields

Negative Correlation

- ❏ pub_rec_bankrupticies shows a negative correlation with almost every field
- ❏ annual_inc shows a negative correlation with dti

Correlation