Lending Club Case Study

Exploratory Data Analysis

By:

Tejas & Deenaz

Problem Statement:

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:

- 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

Data Understanding:

Thoroughly validating the entire dataset involved a comprehensive examination of each column to understand the nature of the values they represent. During this analysis, various data quality issues were identified, including numerous N/A values, incomplete information, date format discrepancies, and data type inconsistencies. These issues will be systematically addressed during the forthcoming Data Cleaning process to ensure the dataset's integrity and reliability. Additionally, consulting the provided Data Dictionary enhanced our understanding of the business domain, facilitating more informed and contextually aware analyses in the future.

No duplicates found in the loan data

Loan data has 39717 rows and 111 Columns

https://www.lendingclub.com/ was referenced for further understanding

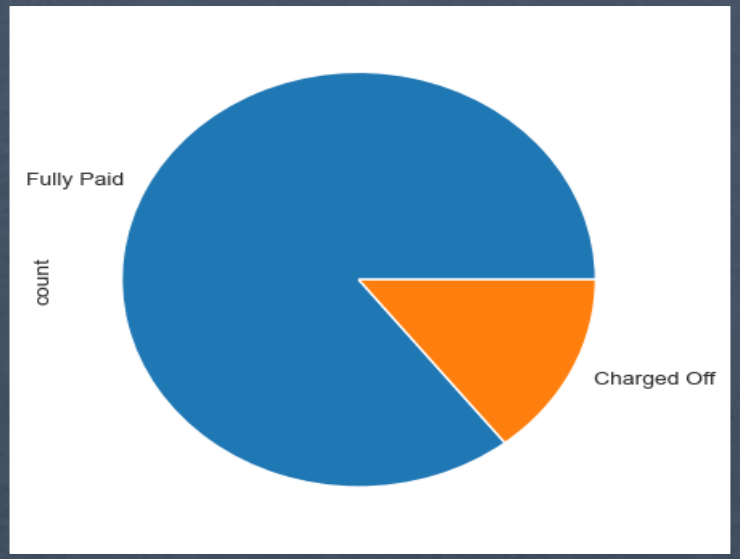
Data Cleaning and Preprocessing:

- No duplicates rows found.
- We excluded 1,160 rows with the loan status labeled as '**current**' from the dataset as this particular status does not contribute to the analysis.
- Exclude columns that consist entirely of null values
- Exclude rows that have null values across all columns.
- Excluded all those columns which had single value constantly.
- Fixed Column Datatypes to ensure accurate analysis and efficient use of memory.
- Performed outlier checking on the 'funded_amnt_inv' and 'annual_inc' column and removed outliers.

Univariate Analysis

Unordered Categorical Analysis

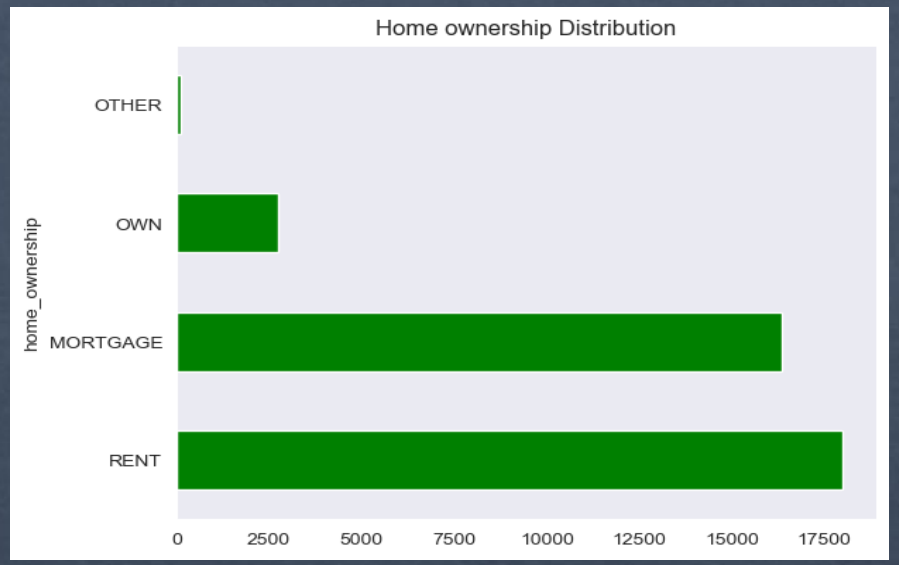
Distribution of Loan



Loan_Status:
Fully Paid - 31813
Charged Off - 5360

- Predominantly, fully paid loans make up the majority of the records.

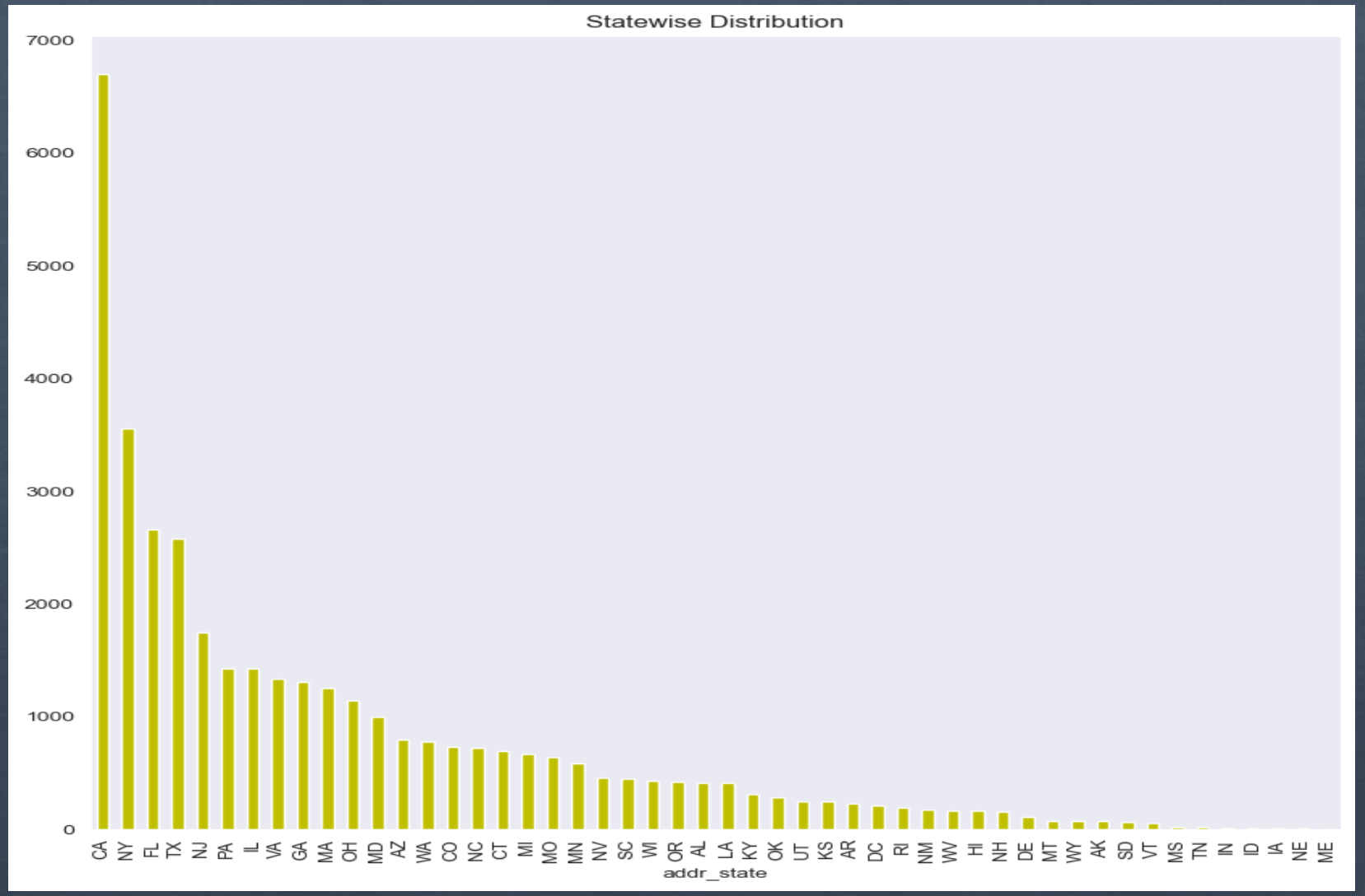
Home Ownership Distribution



RENT - 17999
MORTGAGE - 16331
OWN - 2744
OTHER - 99

- The majority of applicants indicate home ownership through either a mortgage or rental arrangement.

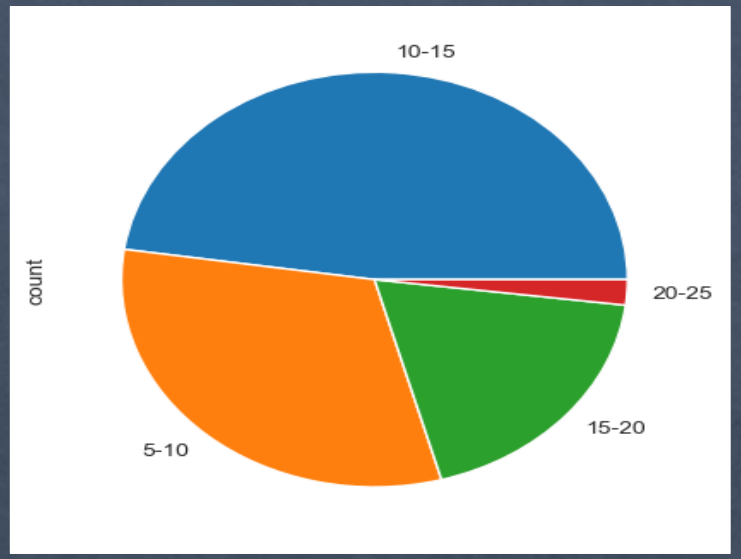
Loan Application Distribution



- Indicates that California (CA) has the highest number of loan applicants among all states.

Ordered Categorical Analysis

Interest Rate

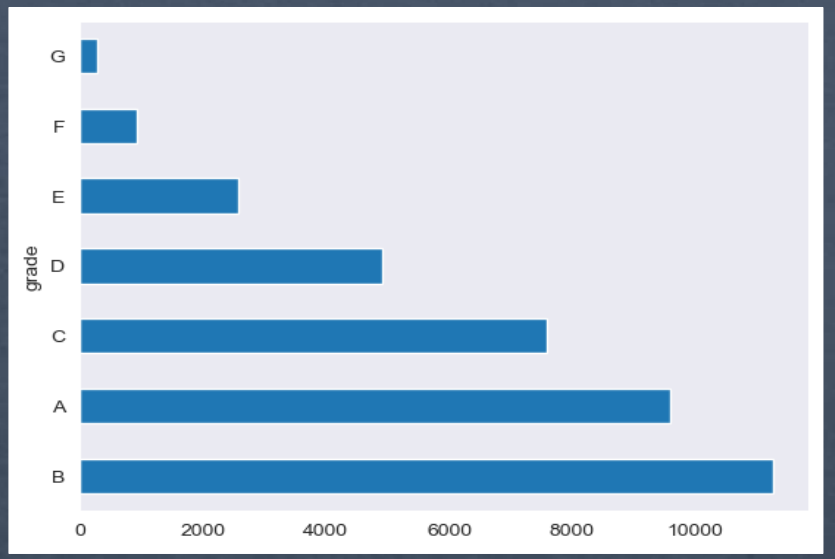


Interest Groups:

5-10	11746
10-15	17720
15-20	6975
20-25	732

- Indicates that a significant portion of loans falls within the interest rate range of '10-15'.

Grade wise Loan Applicants



Grades:

B	11268
A	9593
C	7593
D	4924
E	2569
F	938
G	288

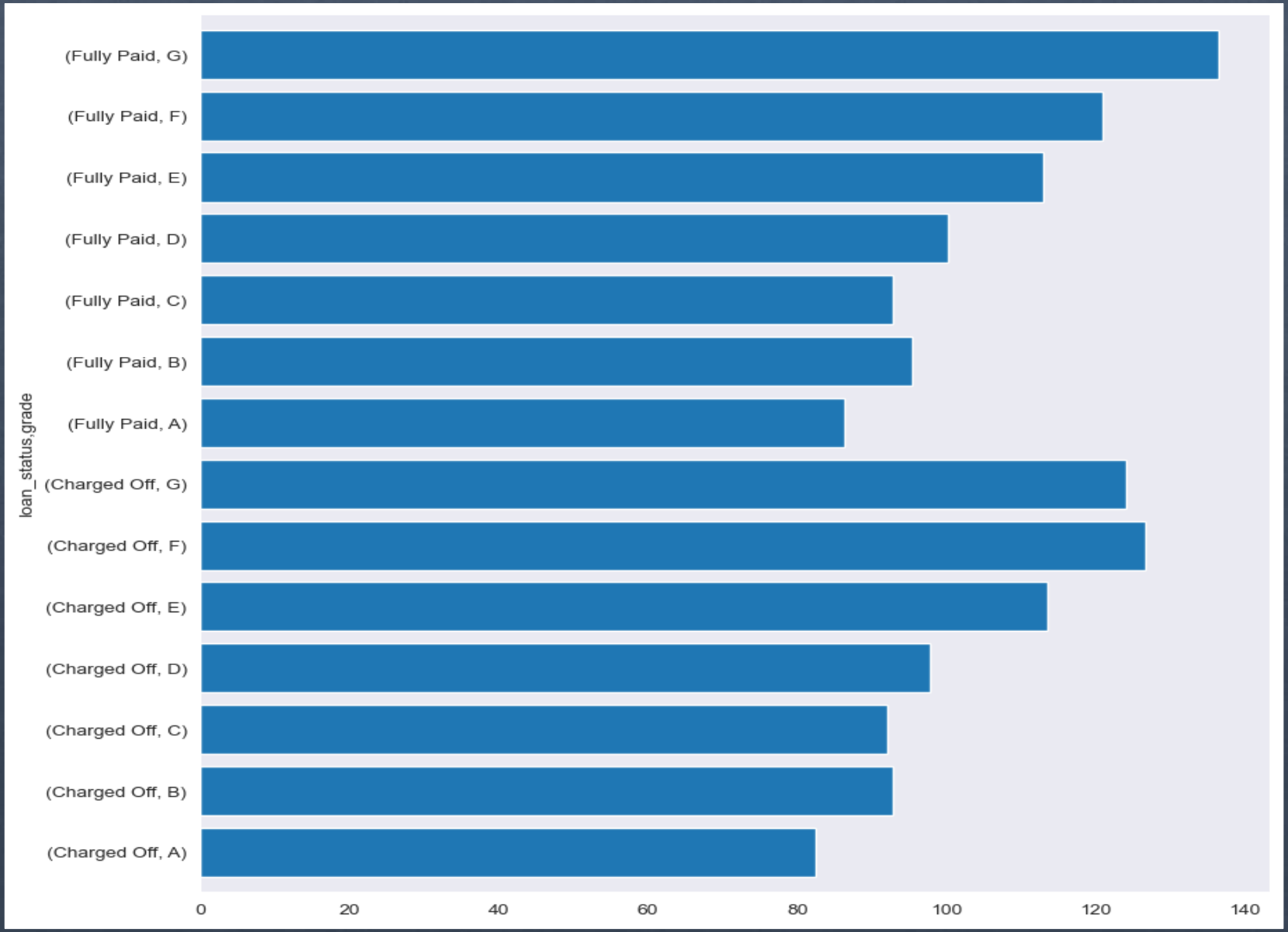
- Indicates that the majority of loans are categorized as Grade B, with Grade G having a lower representation.

Loan Term Plan Distribution

- Indicates that significant number of loans have a 3-year term.

Segmented Univariate

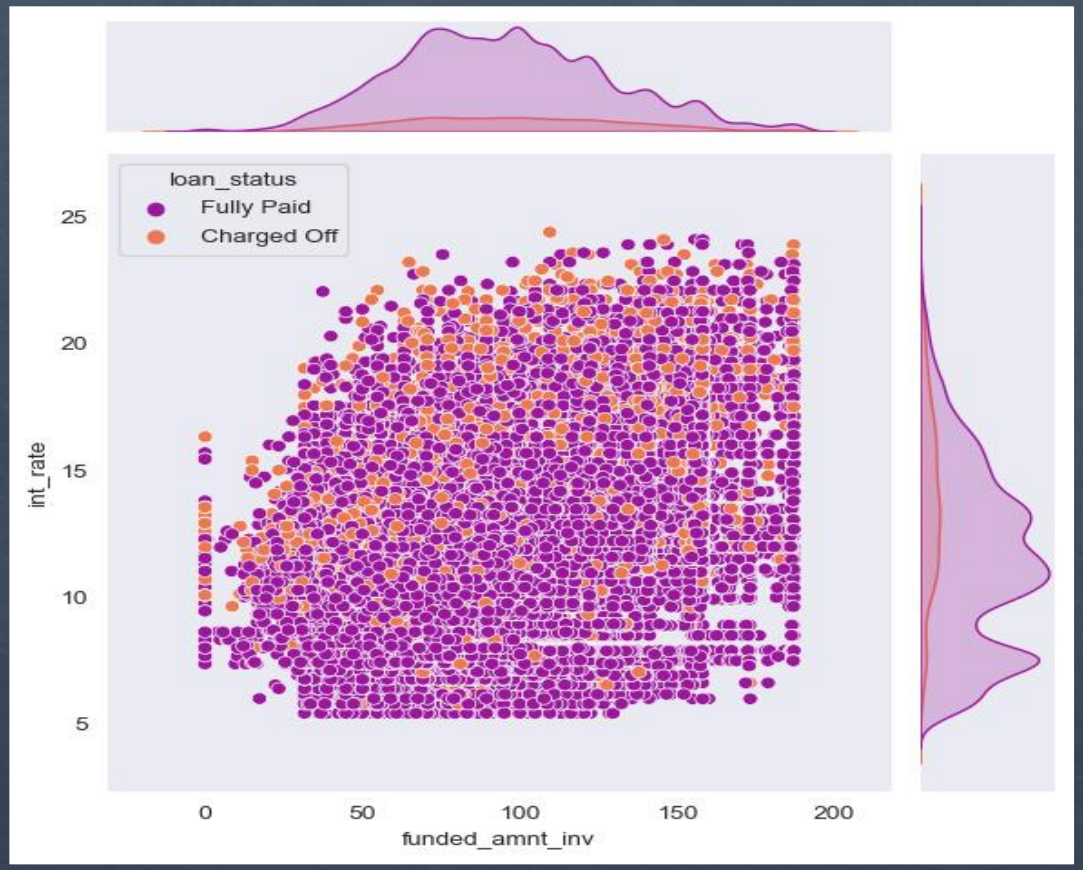
Segmented Representation of Grade wise Funded Amount



- Indicates that fully paid Grade G loans exhibit the highest average loan amount provided by Lending Club.
- Signifies that the majority of charged-off loans belong to grades E, F, and G, with an average loan amount equal to or exceeding 130,000.

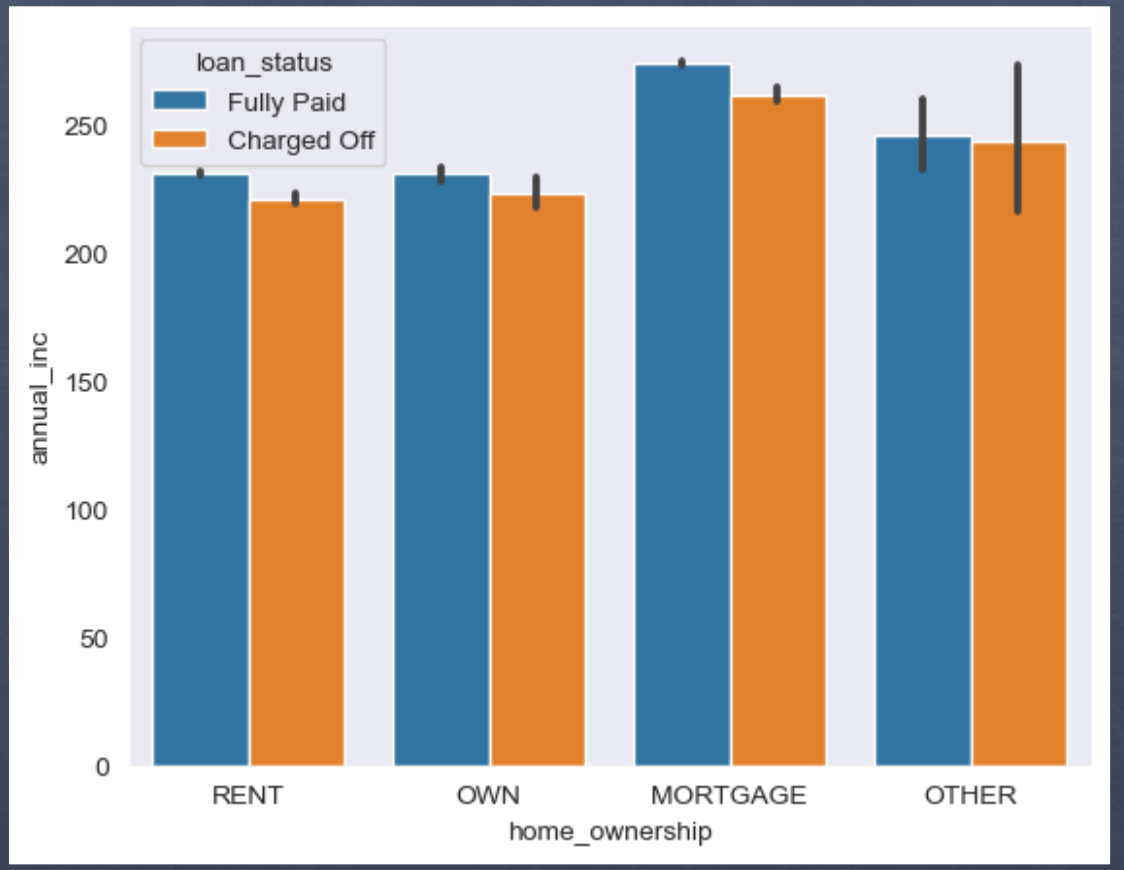
Bivariate Analysis

Joint Plot Representation of Interest_Rate vs Funded_Amount



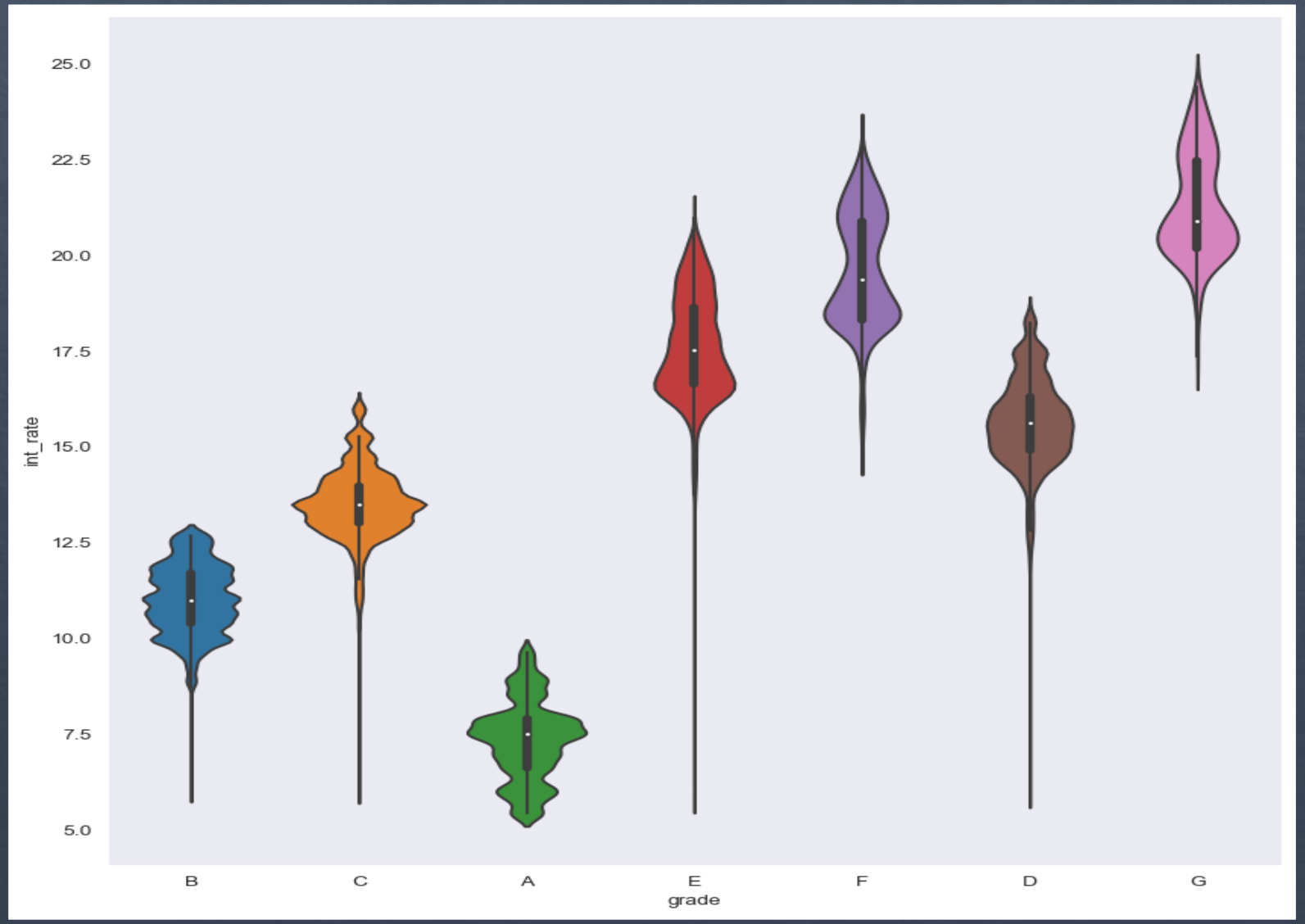
- Indicates that a predominant proportion of charged-off loans exhibit interest rates exceeding 12%.

Bar Plot Representation of Annual_Income vs Home_Ownership



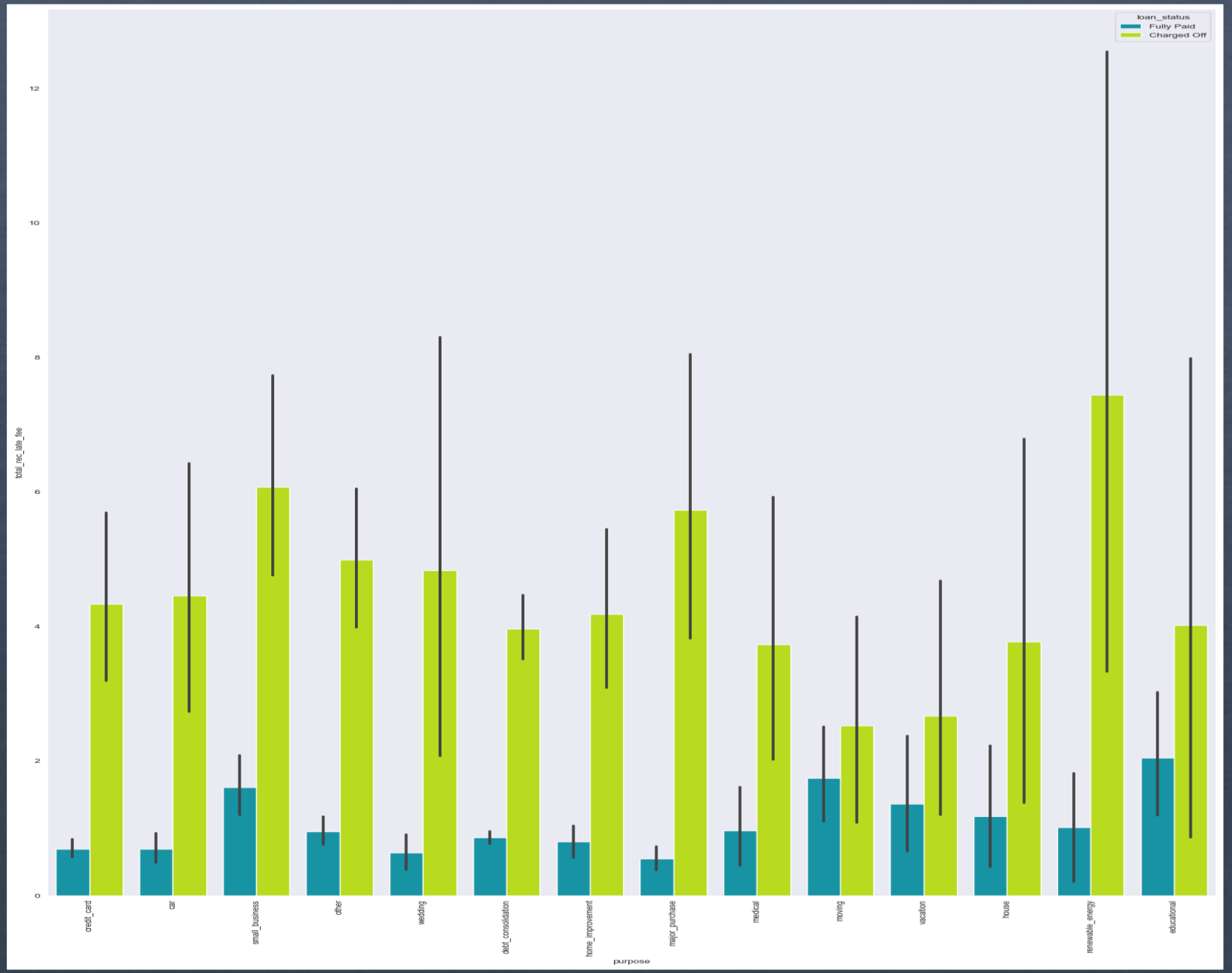
- Applicants with a mortgage home ownership status tend to have the highest average annual income, observed across both fully paid and charged-off loans.

Violinplot Plot Representing Distribution of Interest Rates Across Different Loan Grades



- The interquartile range of interest rates is more significant for Grade G loans compared to other grades.
- Grade A loans tend to have the least interest rates.

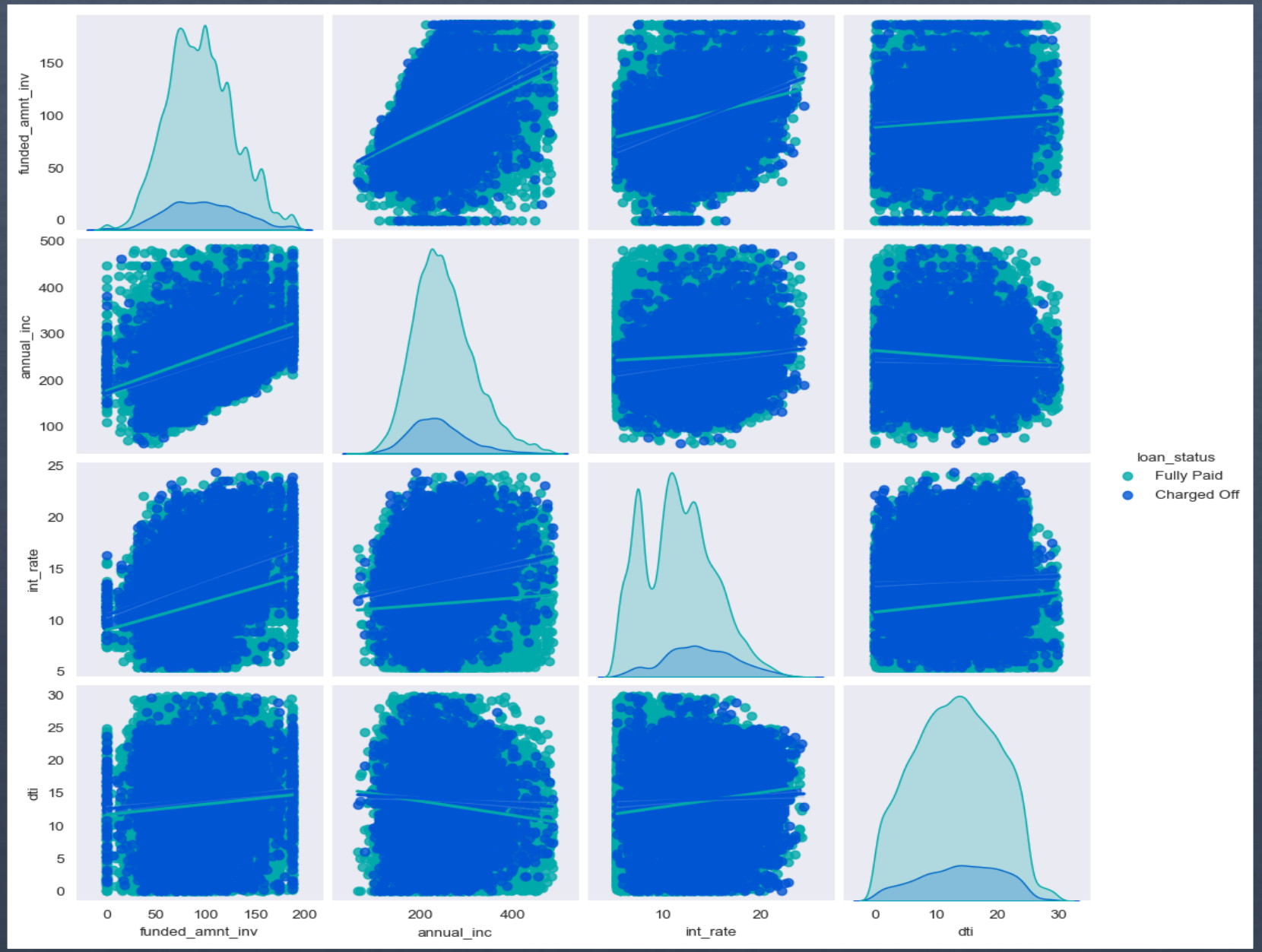
Bar Plot Representing Total Late Fees Received for Different Loan Purposes, Categorized by Loan Status



- Loans with the purpose of Renewable Energy and Small Business tend to have the highest late fees among Charged Off loans.

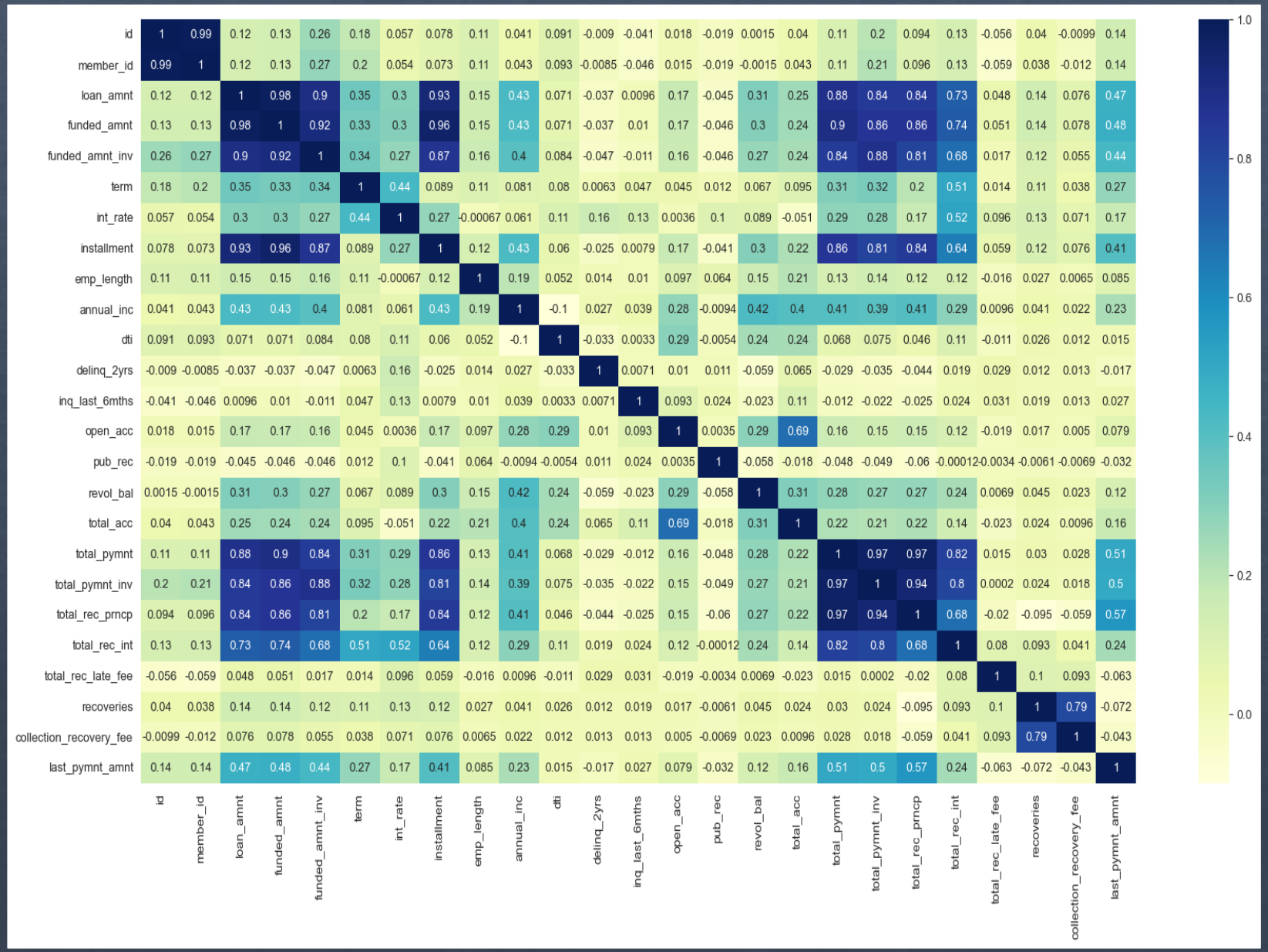
- **Recommendation:** Considering this, the lending company can reassess its lending policies and risk evaluation criteria for loans associated with these particular purposes, implementing measures to mitigate the risk of defaults and enhance overall portfolio performance.

Pair plot Visualization of key variables such as funded amount invested, annual income, debt-to-income ratio, interest rate, and loan status



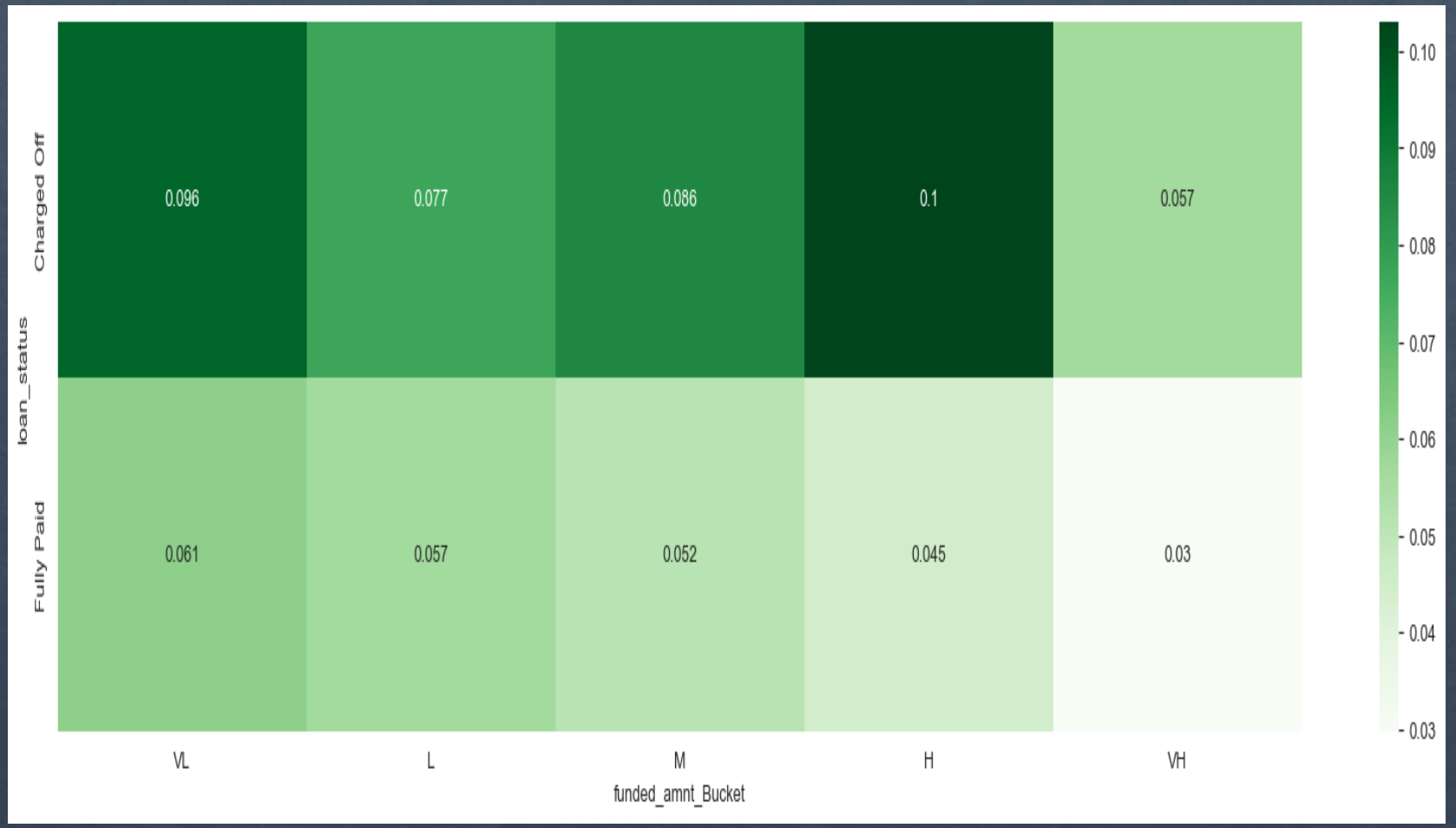
- plot suggests an inverse relationship between dti and annual income, indicating that higher debt-to-income ratios may be associated with lower annual incomes.
- There appears to be a direct proportionality between annual income and funded amount invested, suggesting that as annual income increases, the loan amount provided also tends to increase.
- The plot indicates that interest rates tend to increase with higher debt-to-income ratios.

Correlation Heatmap Providing a Comprehensive Overview of the Relationships Between Numeric Variables



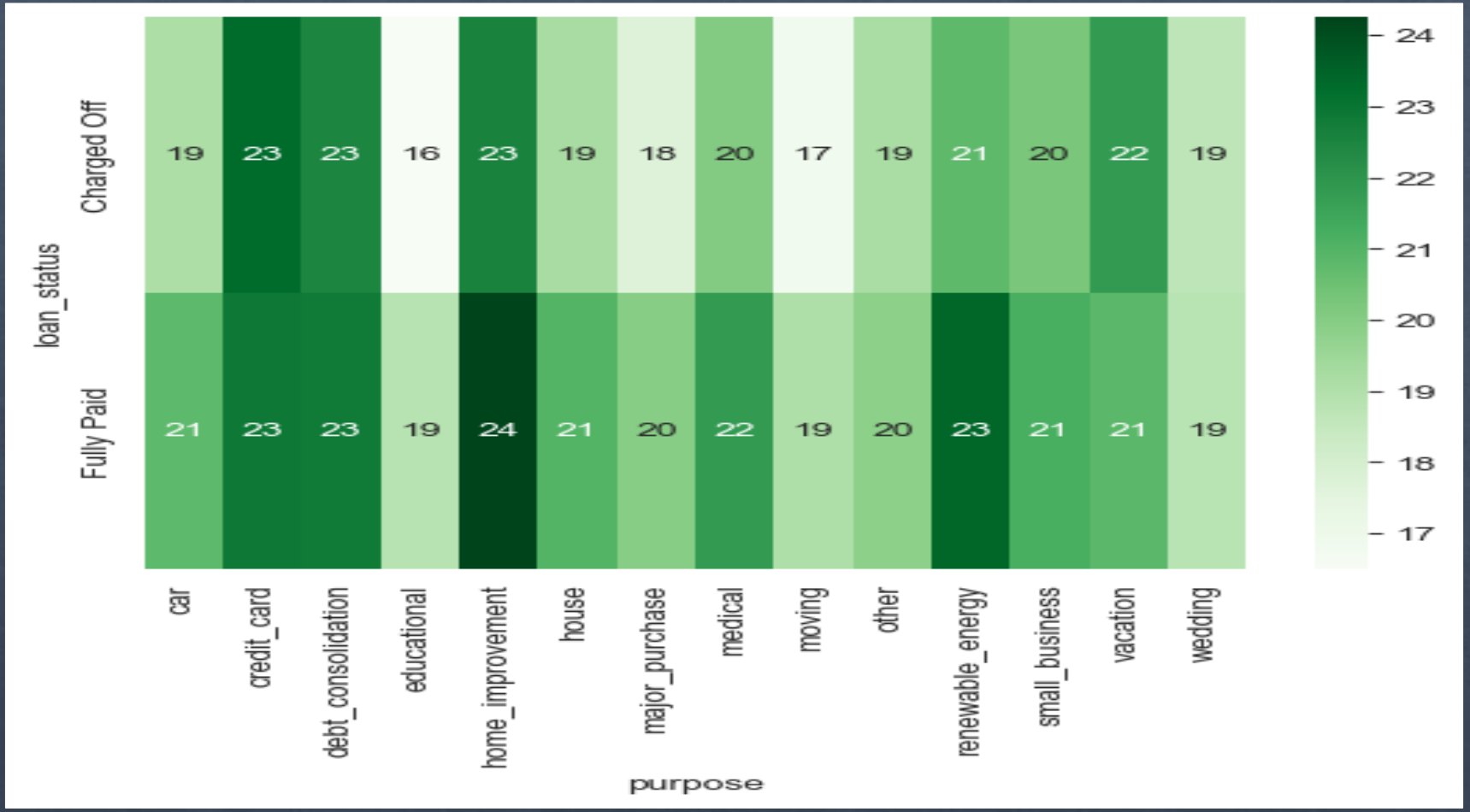
- Notable observations include a negative correlation between the debt-to-income ratio (dti) and the number of derogatory public records (pub_rec), suggesting that higher debt-to-income ratios may be associated with a lower number of derogatory public records
- Total payment amount invested (total_pymnt_inv) has a zero correlation with the total late fee, indicating that the amount paid till date does not have a significant impact on the total late fees incurred.
- Loan_amnt,Funded_amount,funded_amt_inv have a strong Correlation.

The Heatmap Visualization of the Pivot table, Focusing on Loan Status and funded amount inv Bucket with no of public derogatory values



- Shows that charged off loans have a highest number of derogatory records fall in High range bucket of loans and pose credit risk to lenders.
- This finding highlights the potential influence of loan purpose on the derogatory of borrowers with negative credits and emphasizes the importance of considering purpose-specific characteristics in risk assessment and credit analysis.

The Heatmap Visualization of the Pivot table, Focusing on Loan Status and Purpose with total account values

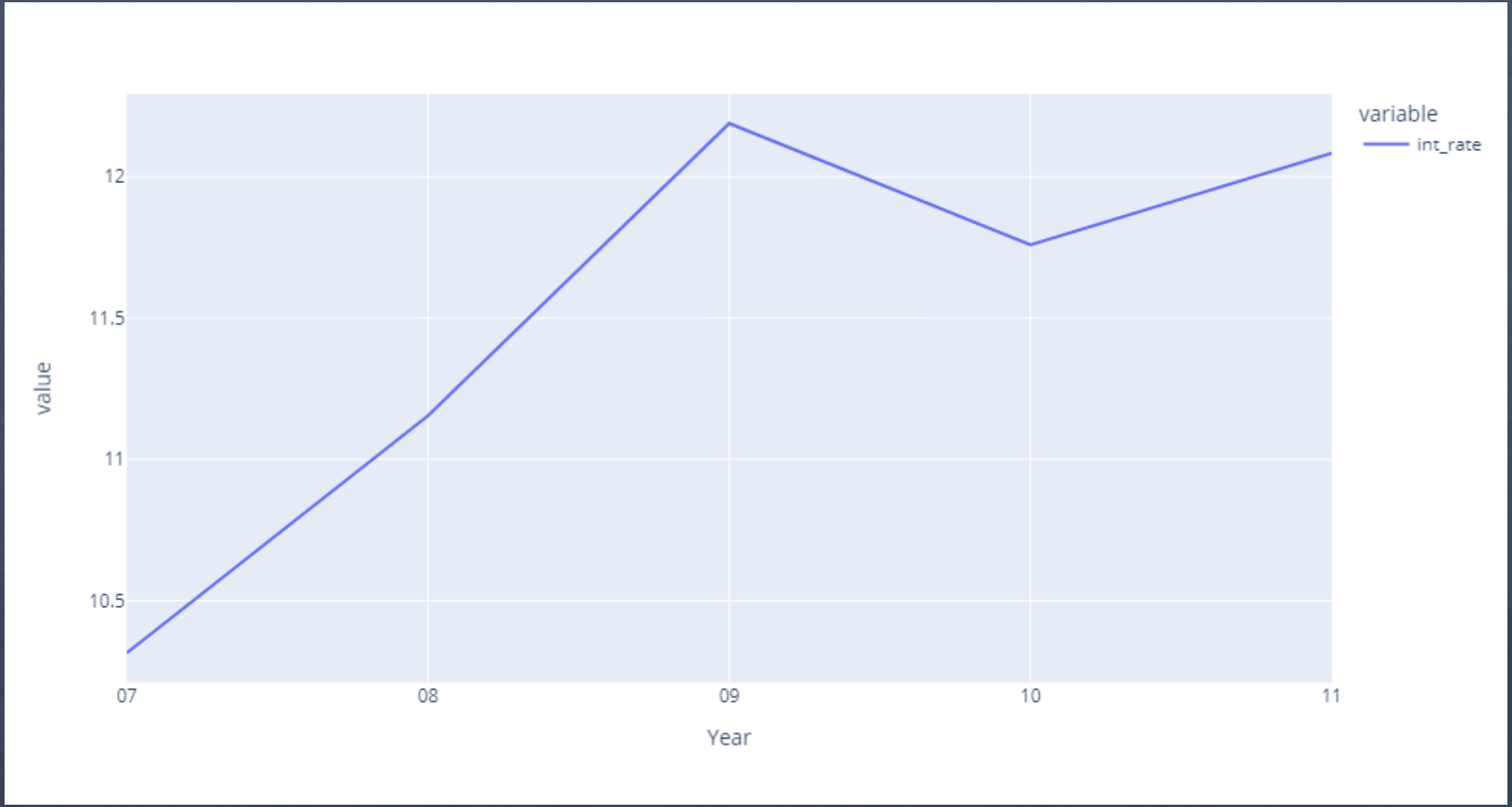


- Shows that charged-off loans associated with the purposes of home improvement and debt consolidation tend to have a higher count of opened credit lines (total_acc).

- This finding highlights the potential influence of loan purpose on the credit behavior of borrowers and emphasizes the importance of considering purpose-specific characteristics in risk assessment and credit analysis.

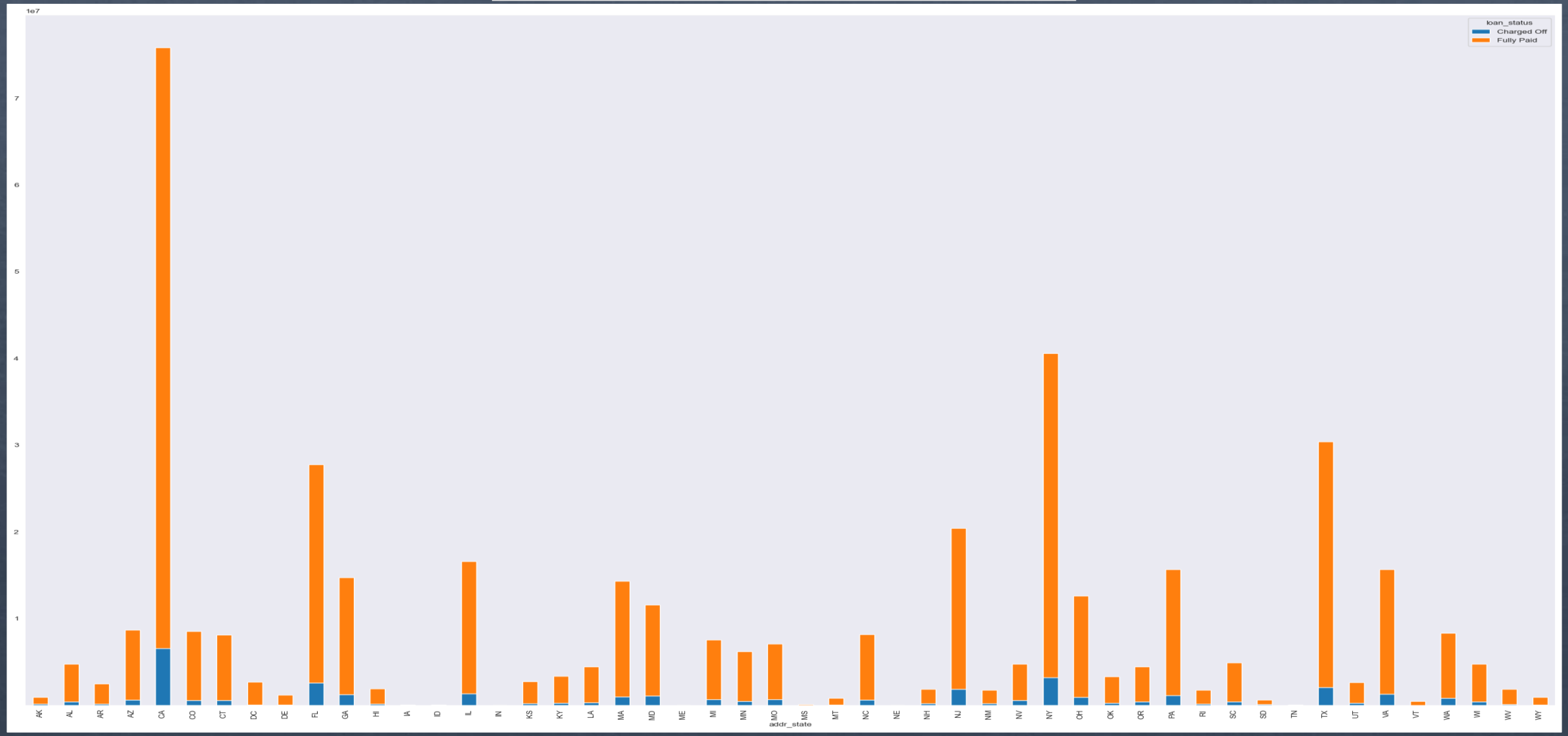
Derived Metric Analysis

Line Plot Visualization for Interest Rate Across Different Years



- The observed spike in interest rates in the years 2009 and 2011, as depicted in the line plot, signifies a notable fluctuation in lending rates during those specific periods.

Stacked Chart Representation of Loan Status and ratio of Total Payment Invoice Amounts to Loan amount Across Different States Values



- States such as CA, NY, and TX exhibit a significant accumulation of both fully paid and charged-off loans, emphasizing the importance of considering geographic factors in analyzing loan performance.

Driving Factors for default Loans

Home Ownership

Purpose

DTI

Annual income

Pub_rec

Grade

Summary and Key Take Aways

- Loans with an interest rate exceeding 12% are more likely to be charged off compared to lower interest rate categories.
- Individuals without home ownership are at a higher risk of loan default.
- Applicants seeking loans for Renewable Energy and Small Business purposes have an elevated likelihood of defaulting.
- Elevated Debt-to-Income (DTI) ratios pose a greater risk of defaults, as they may be linked to lower annual incomes.
- A higher count of bankruptcies is associated with an increased probability of loan defaults.
- Loans classified under Grade G have the highest likelihood of default among different loan grades.