EDA Lending Case Study

By Anshul Ahuja & Ankur Kumar

Agenda

- INTRODUCTION
- BUSINESS OBJECTIVE AND DATASET OVERVIEW
- DATA CLEANING AND MANIPULATION
- UNIVARIATE ANALYSIS
- BIVARIATE ANALYSIS
- SEGMENTED ANALYSIS
- MULTIVARIATE ANALYSIS
- RESULTS
- RECOMMENDATION

Introduction

In this case study, we will dive into extensive analysis of a dataset encompassing **loan details** issued from **2007 to 2011**. Our primary objective is to apply `Exploratory Data Analysis (EDA)` techniques to uncover key factors that contribute to **loan defaults**.

When the company receives a loan application, it must make a critical decision regarding loan approval based on the applicant's profile. This decision involves assessing two types of risks:

- **Opportunity Loss:** If the applicant is likely to repay the loan, then not approving the loan results in a loss of potential business for the company.

- **Financial Risk:** If the applicant is not likely to repay the loan, i.e., is likely to default, then approving the loan may lead to a financial loss for the company.

The dataset provided contains information about past loan applicants and whether they 'defaulted' or not. This information can be used to make more informed decisions regarding loan approval, such as:

Denying the Loan: For applicants likely to default.
Reducing the Loan Amount: To minimize risk exposure.
Adjusting Interest Rates: Charging a higher rate for riskier applicants.

Business Objective:

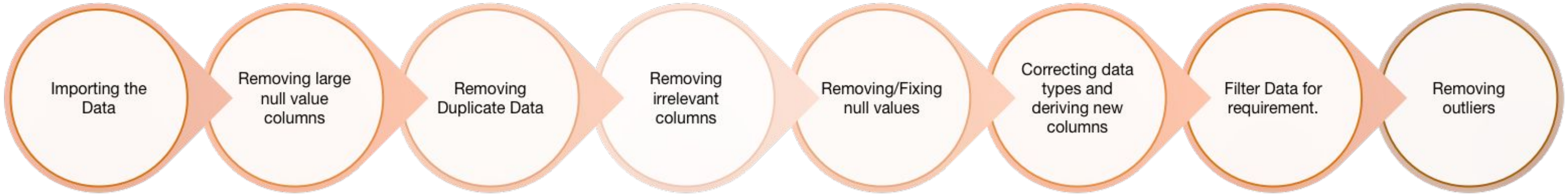
Identifying risky loan applicants is crucial for reducing credit loss. By leveraging `Exploratory Data Analysis (EDA)`, the company aims to uncover the driving factors behind loan defaults. Understanding these **driver variables**, i.e., the variables that are strong indicators of **default**, allows the company to improve its portfolio and risk assessment strategies.

Dataset Overview:

The dataset under examination includes comprehensive records of loans issued over a five-year period, with features ranging from borrower demographics to loan characteristics and payment history. Key attributes are:

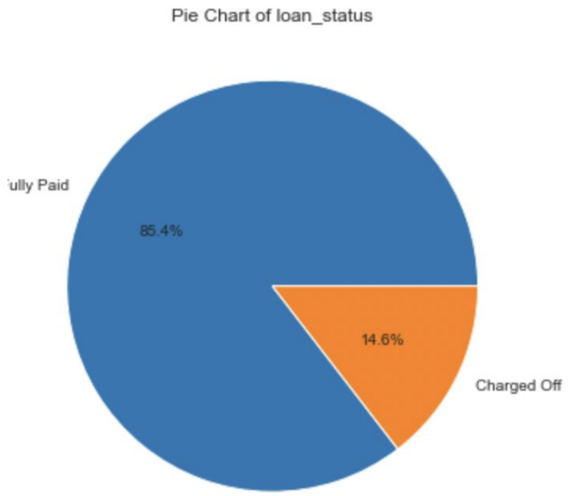
- **loan_status:** The status of the loan (e.g., Fully Paid, Charged Off, Current). This is the `target` variable.
- **annual_inc:** Annual income of the applicant.
- **loan_amnt:** Amount of the loan.
- **int_rate:** Interest rate of the loan.
- **purpose:** Purpose of the loan (e.g., credit_card, debt_consolidation).
- **dti:** Debt to income ratio of applicant.
- **grade:** Credit rating of the applicant (e.g., A,B..G).
- **term:** Duration for which loan has been taken (e.g., 36 Months, 60 Months)

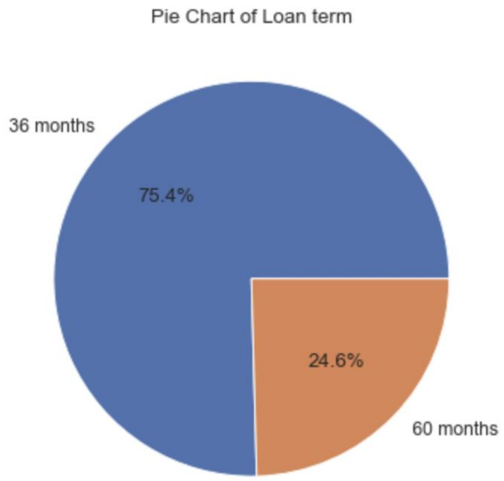
Data Cleaning and Preparation



1. **Handling Null Values:** Identified and removed columns with 54 features entirely containing null values.
2. **Duplicate Rows:** Ensured data integrity by verifying and confirming no duplicates based on the unique key 'id'.
3. **Outlier Management:** Detected and addressed significant outliers in `annual_inc`, `installment`, and `open_acc` to enhance analysis accuracy.
4. **Data Type Conversion:** Converted critical attributes like `int_rate` and `revol_util` from object type to `int64` for improved analysis capability.
5. **Handling Current Loans:** Excluded records where `loan_status` is 'Current' as they do not contribute to the current analysis goals.
6. **Temporal Attributes:** Derived new attributes 'Year' and 'Month' from `issued_dt` to facilitate temporal analysis.
7. **Creation of Categorical Variables:**Implemented binning techniques to create categorical variables like (`loan_amnt_categories`, `annual_inc_categories`, `interest_rate_categories`, for effective bivariate analysis.
8. **Imputation of Missing Values:** Imputed attributes like `emp_length, pub_rec_bankruptcies etc with proper values like median/mode`
9. **Column Pruning:**Removed columns such as `policy_code`, `acc_now_deliqn`, and others with only one unique value to streamline the dataset.

Univariate Analysis





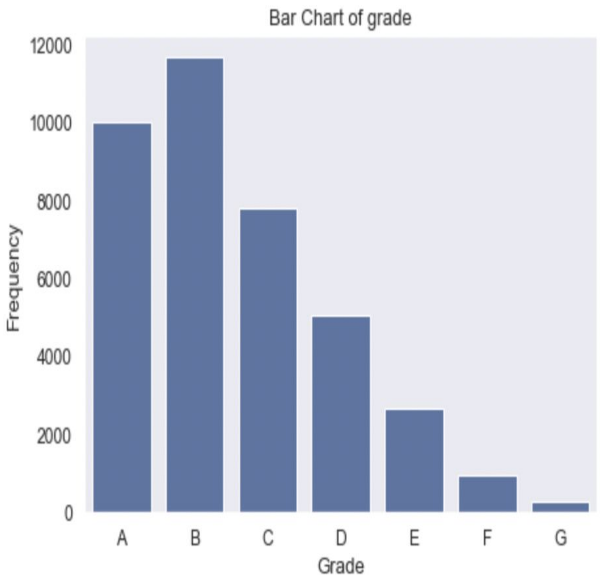
Observation:

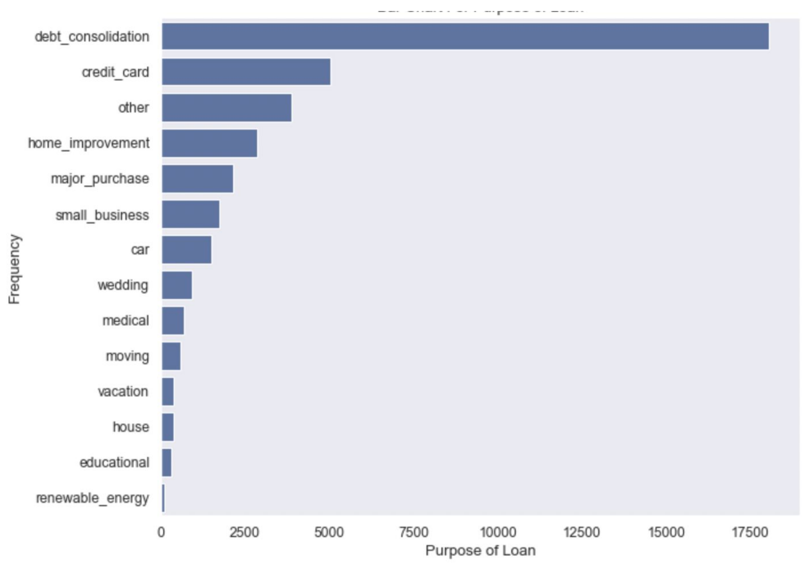
Only **14.6%** of the total loans were charged off.

Observation:

Above Pie chart shows that out of total loans issued ~76% have taken loan for 36 Months period

Grade and Loan Bar Chart





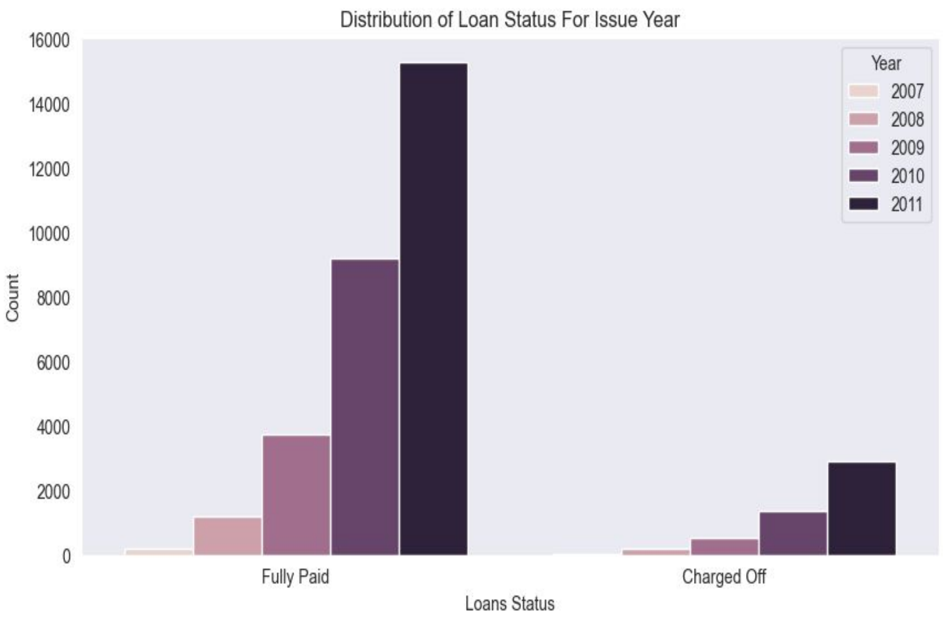
Observations:

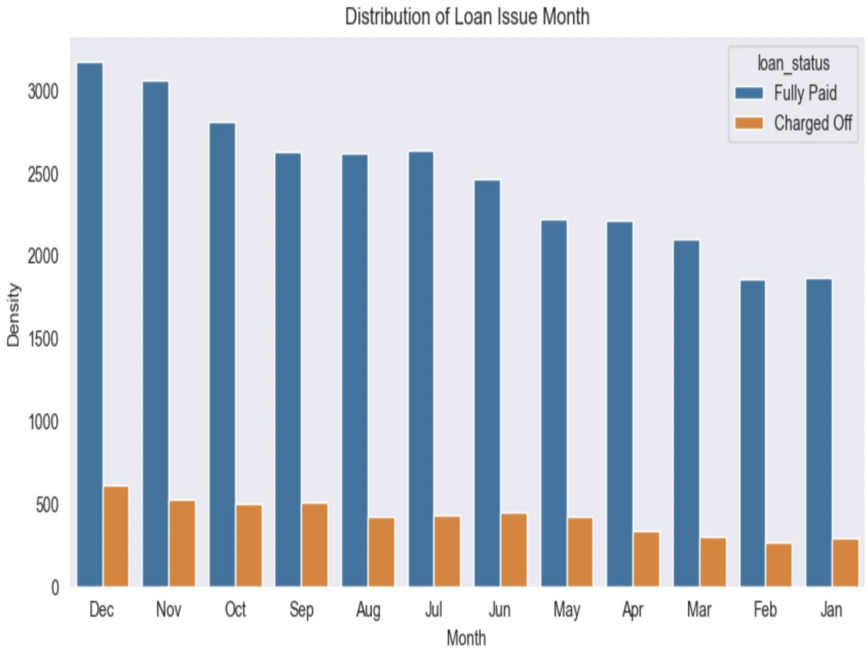
- We can see borrowers count decreases as we go from high credit rating (A) to low credit Rating (B)
- A reason for this could be, lending platform release loan easily to borrowers who has good credit rating

Observations:

- The bar plot shows that most loans were taken for debt_Consolidation.

Employment Length & Year Wise Loan Status





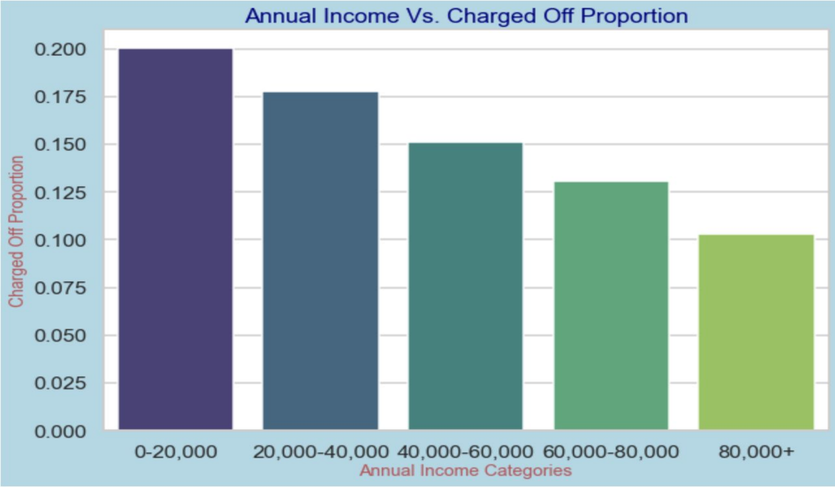
Observations:

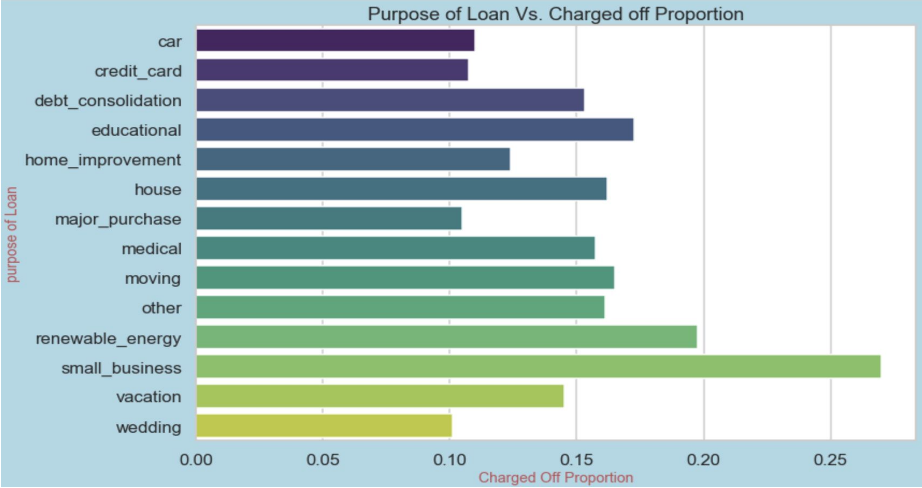
- With each passing year total number of fully paid loans and charged off count are increasing exponentially.

Observations:

- Majority of the loans are given in last quarter of the year.

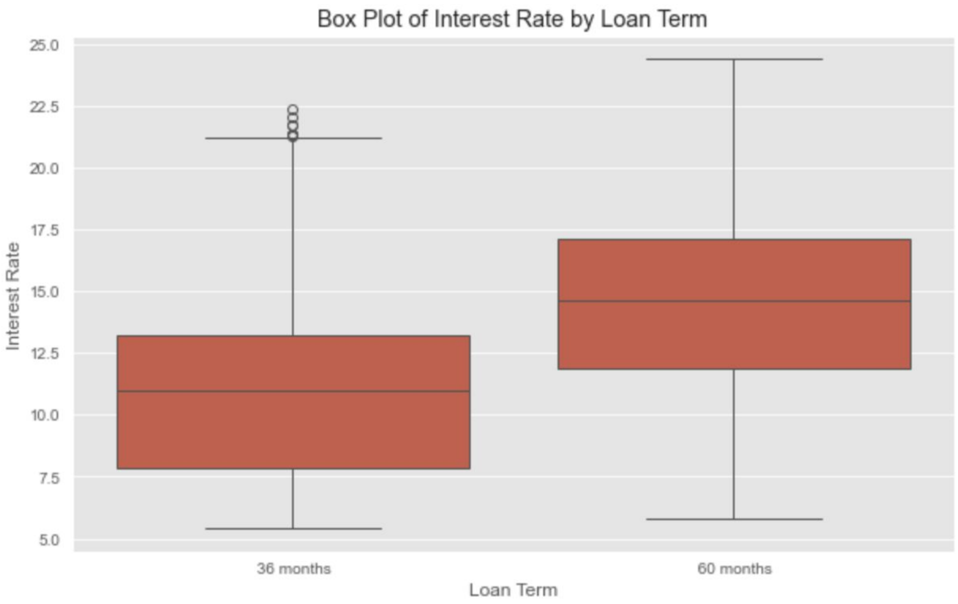
Bivariate Analysis





Observations :

1. **Income and Charged-Off Proportion:**
   - There is a clear inverse relationship between the annual income of borrowers and the proportion of charged-off loans. Lower annual incomes tend to correlate with higher charged-off proportions.
2. **Loan Purpose and Default Risk:**
   - Loans taken for small business purposes exhibit the highest charged-off proportion compared to other loan purposes. This suggests that small business loans carry a higher default risk.

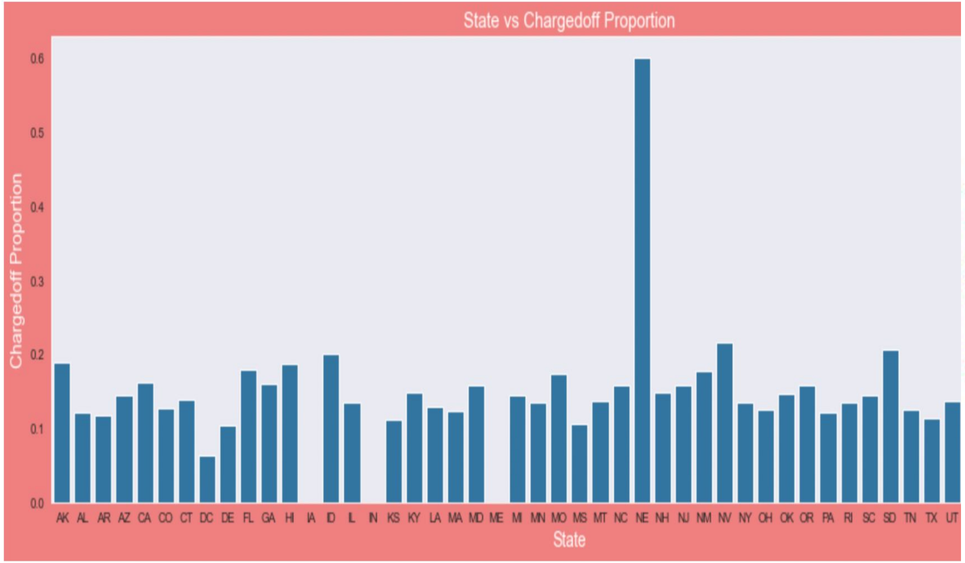


Observations:

- Longer loan terms are associated with higher interest rates. This indicates that borrowers opting for longer terms incur higher interest costs.

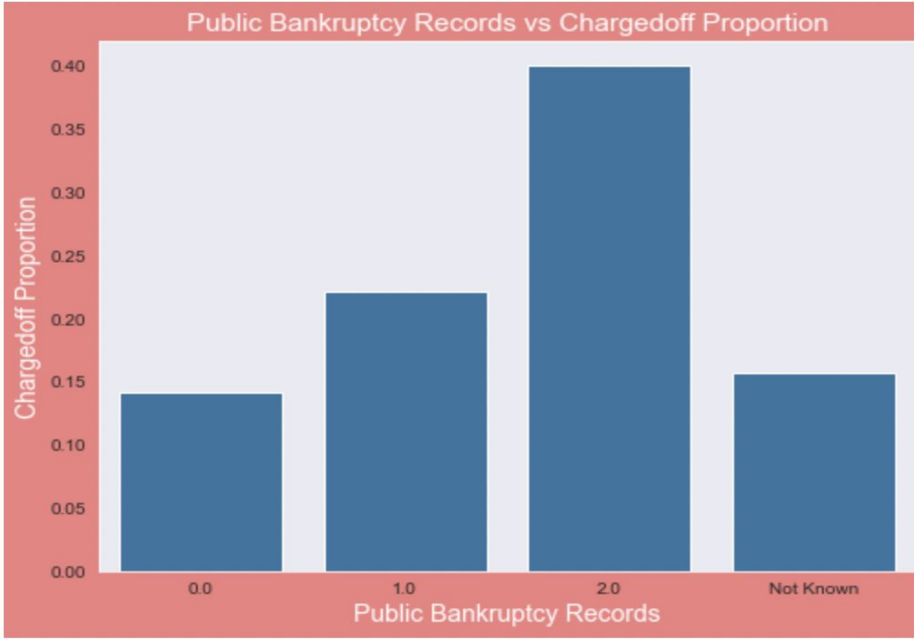
Observations:

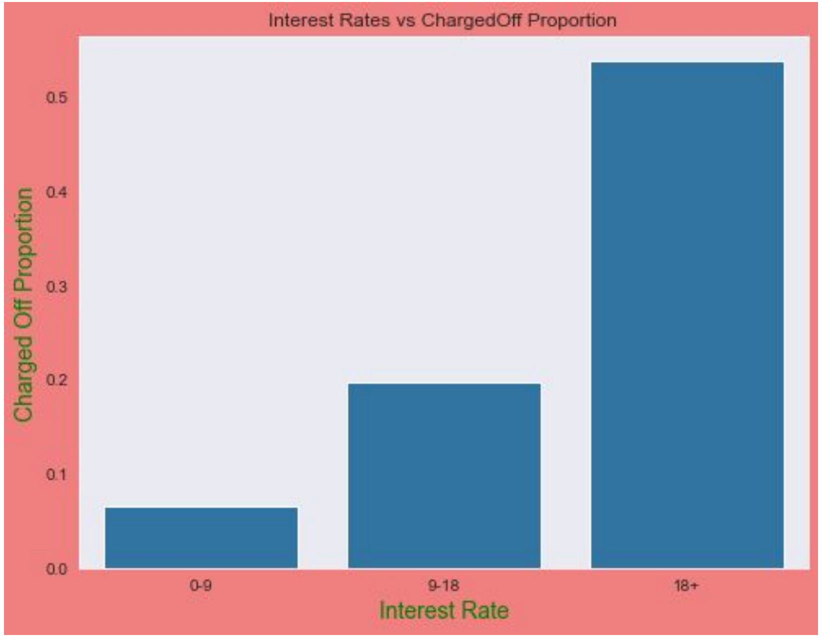
- There is a positive relationship between the debt-to-income ratio and interest rate. As the debt-to-income ratio increases, so does the interest rate.



Observations:

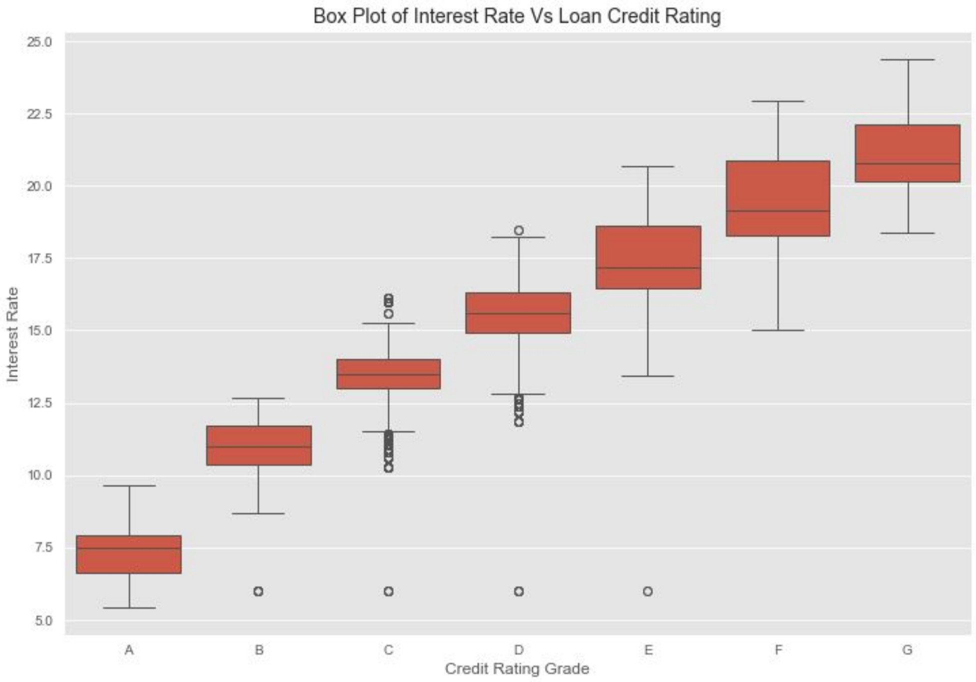
- From the above barplot we can say Borrowers who are from NE has high chance of being chargedOff, But number of applicants from NE is very less (5) and so don't have enough evidence to conclude this.
- As the LC assigned loan grade (`grade`) decreases from 'A' to 'G', the charged-off proportion increases. Grade 'A' represents the highest credit rating, while 'G' represents the lowest.

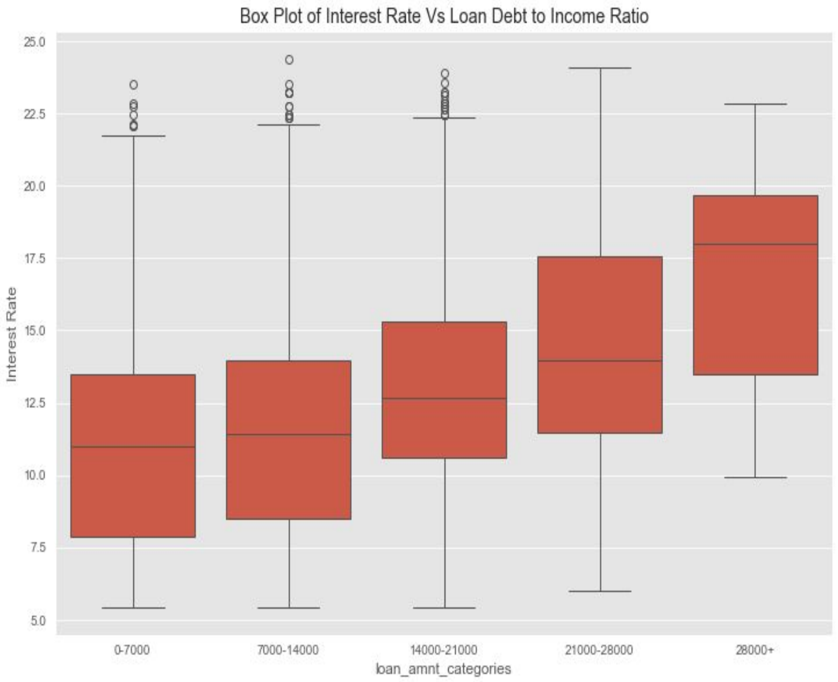




Observations:

- The charged-off proportion increases with an increase in the count of public bankruptcy record number for borrowers.
- There is a positive correlation between the interest rate (`int_rate`) and charged-off proportion. Higher interest rates are associated with higher default risk.

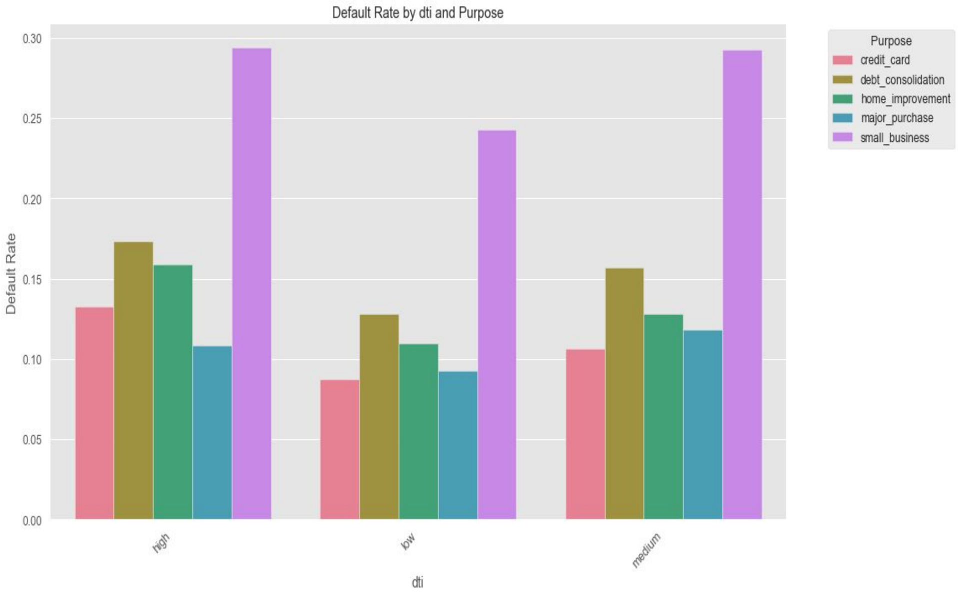


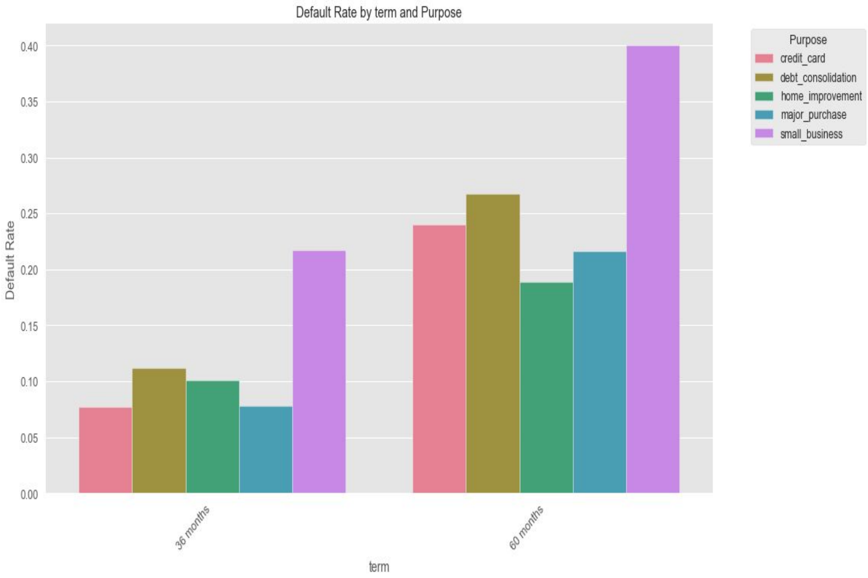


Observations:

- As we move from Grade 'A' to Grade 'G', assigned by the lender, the interest rate increases, indicating that higher credit ratings typically secure lower interest rates for borrowers.
- We observe that the interest rate increases with the loan amount, likely because larger loans are often taken for longer terms, which, as noted earlier, tend to have higher interest rates.

Segmented Analysis



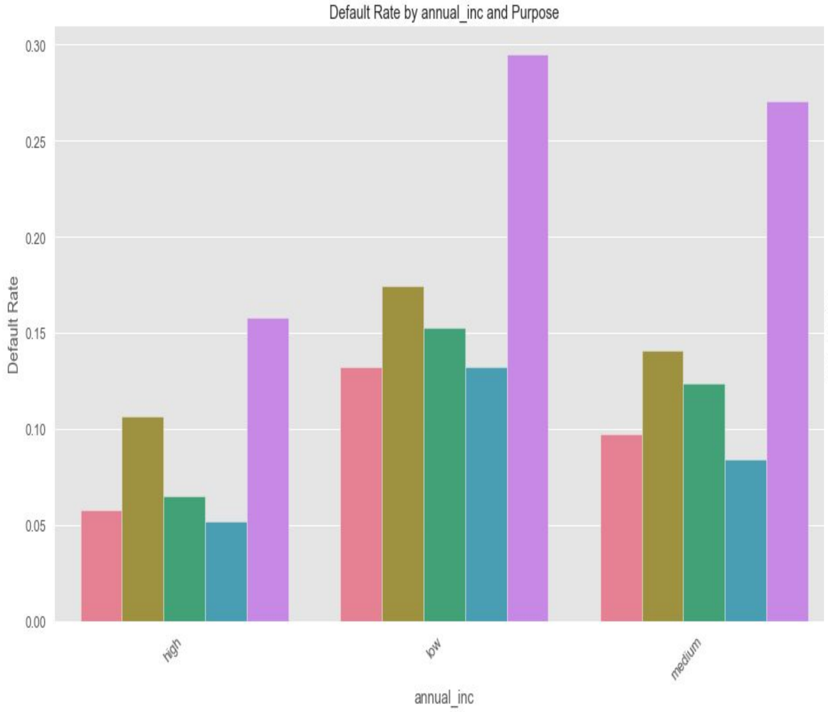


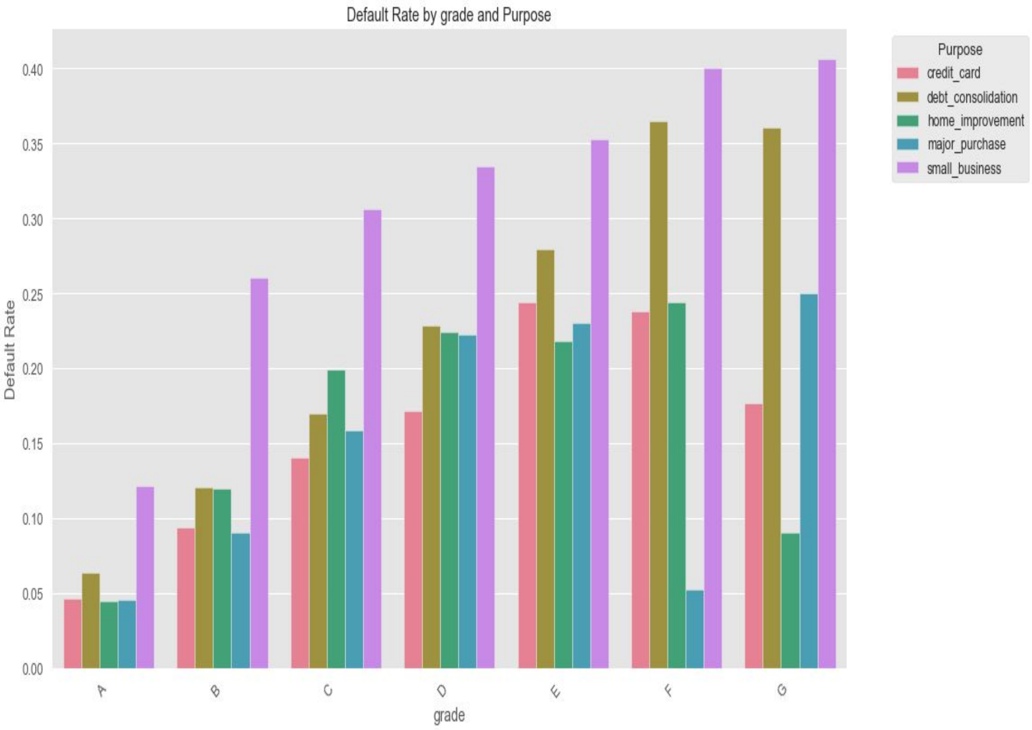
Observations:

- The default rate for small business loans is higher than for other purposes in both 36-month and 60-month term loans.
- Additionally, the default rate for all loan purposes is higher for 60-month loan terms compared to 36-month loan terms.

Observations:

- small_business purpose loan has high default rate across dti category.
- We have same trend across across purpose, nothing conclusive here.

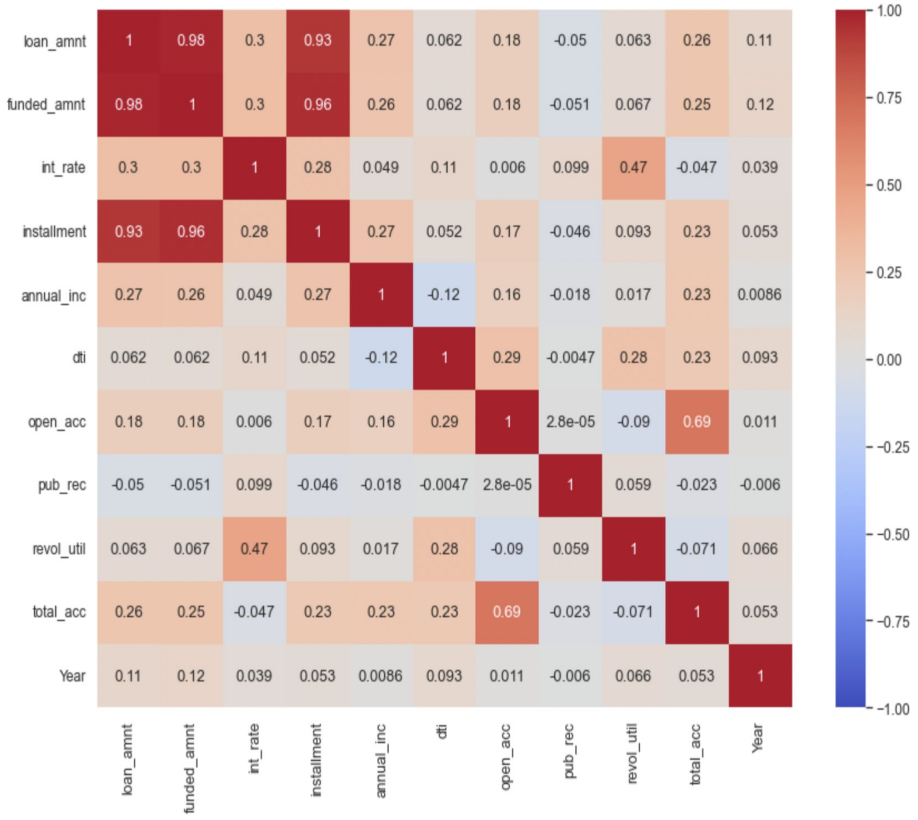




Observations:

- The default rate decreases as annual income increases, and this trend holds true for all loan purposes.
- Loans taken for small business purposes have the highest default rate across all grades from 'A' to 'G'.

Correlation between all Numerical variables

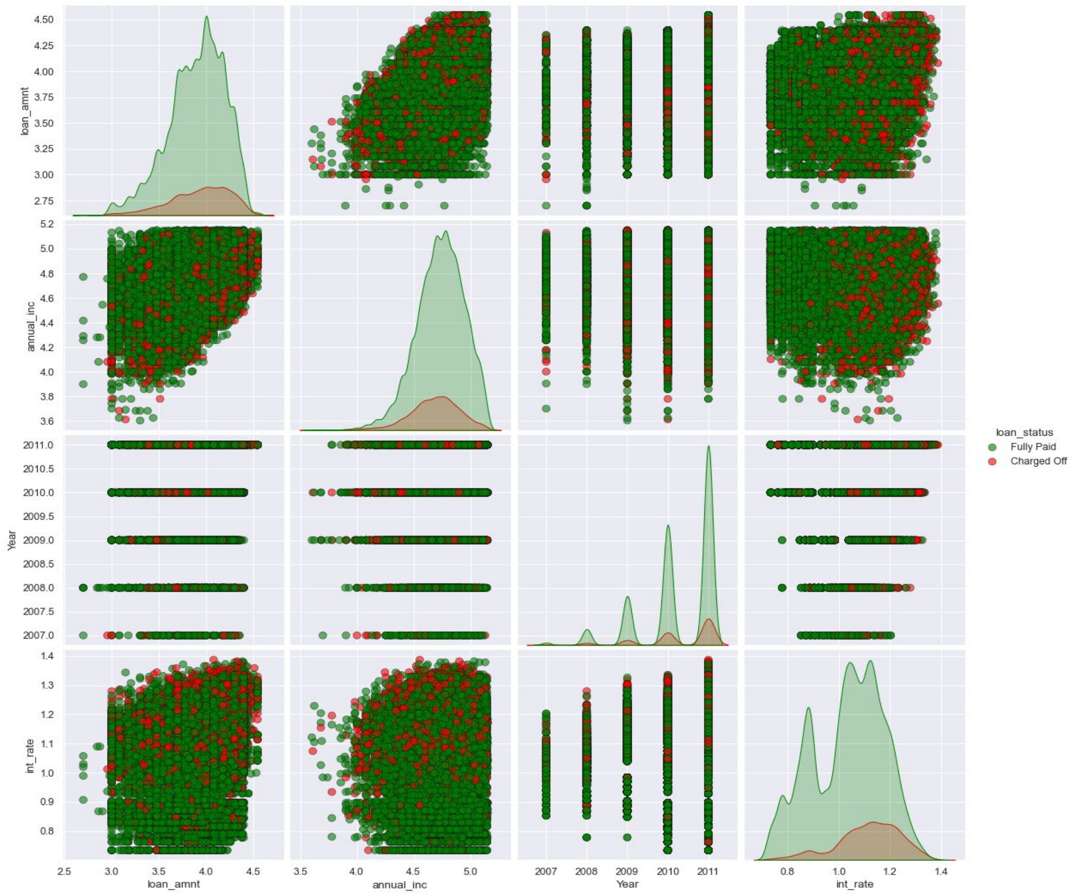


Observations:

- Positive correlations means directly proportional
- Negative correlations means inversely proportional

1. Loan_amount and Interest_Rate are positively correlated, which means higher the loan_amount higher the interest_rate
2. Loan_amount and annual_income are positively correlated, which means higher the annual_income higher the loan_amount
3. interest_rate and installment are positively correlated, higher the interest_rate higher the installment.
4. Annual_income and debt_to_income_ratio is negatively correlated, higher the annual_income lower the debt_to_income_ratio
5. Interest_rate and credit_utilization is positively correlated, higher the credit_utilization (revol_util) higher is the interest_rate

Multivariate Analysis



Observations:

1. Higher the loan amount, higher is the annual income.
2. Higher the annual income, high is the loan amount
3. Higher the interest rate, higher is the chargedOff count
4. Yearly, there is increase in total number of loans taken.

Results

- **Income and Charged-Off Proportion:**
  - There is a clear inverse relationship between the annual income of borrowers and the proportion of charged-off loans. Lower annual incomes tend to correlate with higher charged-off proportions.
- **Loan Purpose and Default Risk:**
  - Loans taken for small business purposes exhibit the highest charged-off proportion compared to other loan purposes. This suggests that small business loans carry a higher default risk.
- **Loan Grade and Default Risk:**
  - As the LC assigned loan grade (`grade`) decreases from 'A' to 'G', the charged-off proportion increases. Grade 'A' represents the highest credit rating, while 'G' represents the lowest.
- **Interest Rate and Default Risk:**
  - There is a positive correlation between the interest rate (`int_rate`) and charged-off proportion. Higher interest rates are associated with higher default risk.
- **Public Bankruptcy Records:**
  - The charged-off proportion increases with an increase in the count of public bankruptcy record number for borrowers.
- **Loan Term and Interest Rate:**
  - Longer loan terms are associated with higher interest rates. This indicates that borrowers opting for longer terms incur higher interest costs.
- **Debt-to-Income Ratio and Interest Rate:**
  - There is a positive relationship between the debt-to-income ratio and interest rate. As the debt-to-income ratio increases, so does the interest rate.
- **Loan Grade and Interest Rate:**
  - Similar to default risk, as the LC assigned loan grade (`grade`) decreases from 'A' to 'G', the interest rate increases. Lower-grade loans are associated with higher interest rates.

Results Cont..

- **Loan Amount and Interest Rate:**
  - Interest rates tend to increase with higher loan amounts. This suggests that larger loans are associated with higher interest costs.
- **Interest Rate by Purpose:**
  - Loans for small business purposes have the highest interest rates compared to other loan purposes.
- **Loan Amount and Annual Income:**
  - There is a positive correlation between the loan amount and annual income of borrowers. Higher annual incomes tend to correlate with higher loan amounts.
- **Annual Income and Debt-to-Income Ratio:**
  - There is a negative correlation between annual income and debt-to-income ratio. Higher incomes tend to be associated with lower debt-to-income ratios.
- **Small Business Loans and Default Rate Across Grades:**
  - Small business loans consistently exhibit the highest default rates across all loan grades ('A' to 'G'). Additionally, the default rate for small business loans increases as the credit rating of borrowers decreases.
- **Default Rates Across Loan Terms:**
  - Small business loans have higher default rates compared to other loan purposes, across both 36-month and 60-month term loans.
- **Small Business Loans Across Housing Types:**
  - Small business loans consistently exhibit the highest default rates across borrowers residing in rented houses, owning houses, or having mortgaged houses.
- **Interest Rate and Credit Utilization:**
  - There is a positive correlation between interest rate and credit utilization `revol_util`. Higher the credit utilization higher is the interest rate.

Conclusion & Recommendations

Based on the findings from the analysis, the following recommendations are suggested for the loan provider/bank:

1. **Risk Assessment and Loan Approval:**
   a. Implement a stricter assessment for borrowers with lower annual incomes, particularly for loans with small business purposes, given their higher default risk.
2. **Credit Rating and Interest Rates:**
   a. Adjust interest rates based on the LC assigned loan grade ('A' to 'G') to better reflect the risk associated with different credit ratings. Lower-grade loans should incur higher interest rates.
3. **Debt Management and Loan Terms:**
   a. Encourage borrowers to manage their debt-to-income ratios effectively, as higher ratios are linked to higher interest rates and potential default risks, especially for longer loan terms.
4. **Monitoring Small Business Loans:**
   a. Enhance monitoring and risk management strategies specifically for small business loans, which consistently exhibit higher default rates across different metrics.
5. **Housing Status and Loan Performance:**
   a. Consider the housing status of borrowers (rented, owned, mortgaged) when assessing default risks for small business loans, as default rates remain consistently high across different housing types.

Thank You