CREDIT RISK ANALYSIS

EDA CASE STUDY

BY –RAJ KUMAR SARKAR

CONTENTS

CHAPTER 1

INTRODUCTION

INTRODUCTION

PROBLEM STATEMENT

A consumer finance company specializes in providing various types of loans to urban customers. When the company receives a loan application, it must decide whether to approve or reject it based on the applicant's profile. Two types of risks are associated with the bank's decision:

Risk of Loss due to Non-Approval:

If an applicant is likely to repay the loan, not approving the loan results in a loss of business for the company.

Risk of Default:

If an applicant is not likely to repay the loan (i.e., they are likely to default), approving the loan may lead to financial loss for the company.

OBJECTIVE:

The goal of this EDA project is to identify patterns and insights that indicate whether a person is likely to default on a loan. These findings can be used to make informed decisions, such as denying the loan, reducing the loan amount, or lending to risky applicants cautiously.

CHAPTER 2

DATA OVERVIEW

DATA OVERVIEW

We performed an extensive analysis using historical loan data to uncover essential patterns and risk factors related to loan defaults. Our objective is to offer actionable insights that can improve decision-making in loan approvals. Below, you'll find the flowchart outlining our approach.

DATA LOADING

- LIBRARIES USED
  - Pandas – Used for Data Manipulation and Analysis
  - Numpy – Used for Numerical Computational
  - Matplotlib and Seaborn - Used for Data Visualization
- USING THE 'loan.csv' FILE
  - The 'loan.csv' file contains vital information about loan applications. Efficiently loading this dataset is the first crucial step in data analysis.
    The example code to load the dataset using the Pandas library is provided:

$$loan\_df = pd.read\_csv('loan.csv')$$

DATA CLEANING

- HANDLING MISSING VALUES
  - Detect and understand the extent of missing data in the dataset.
- DELETING COLUMNS

Removing all colums having only null values

```
In [6]:  empty_cols=loan_df.columns[loan_df.isnull().all()]

In [7]:  loan_df.drop(columns=empty_cols,inplace=True)
```

- There are 53 columns which are 100 % empty And 3 columns which have more than 60 % empty are dropped
- There are 17 columns which are important for our analysis so all other are deleted to increase efficiency.

DATA CLEANING

DELETING ROWS

- For analysis we have deleted all the rows which have value as 'current' for the loan status column.

```
loan_df=loan_df[~(loan_df['loan_status']=='Current')]
```

MODIFYING THE COLUMNS

- There are two columns 'term' and 'int_rate' whose data types are corrected to 'int' and 'float' respectively.

```
loan_df.term=loan_df.term.apply(lambda x:x[:3]).astype('int8')
```

```
loan_df.int_rate=loan_df.int_rate.str.strip("%").astype('float32')
```

DATA CLEANING

CREATION OF COLUMNS

- Three new columns are created

```
# Extracting Year and month from Issue_Date
loan_df['issue_year'] = pd.to_datetime(loan_df['issue_d'], format='%b-%y').dt.year
loan_df['issue_month'] = pd.to_datetime(loan_df['issue_d'], format='%b-%y').dt.month

```

```
# Created a new column for loan to income ratio
loan_df['loan_to_income_ratio']=round(loan_df['installment']/(loan_df['annual_inc']/12) *100,2)
```

- 5 new columns are created to group our existing values into categories for better analysis following some assumptions .

```
# Creating bins for intrest rate column
b0 = [0, 9, 11, 13, float('inf')]
l0 = ['Low', 'Medium', 'High', 'Very High']
loan_df['rate_category'] = pd.cut(loan_df['int_rate'], bins=b0, labels=l0, right=False)
```

DATA CLEANING

```python
# Creating bins for loan amount column
b11 = [0, 5400, 9599, 14999, float('inf')]
l11 = ['Low', 'Medium', 'High', 'Very High']
loan_df['loan_amnt_category'] = pd.cut(loan_df['loan_amnt'], bins=b11, labels=l11, right=False)
```

```python
# Creating bins for dti column
b = [0, 8, 13, 18, float('inf')]
l = ['Low', 'Medium', 'High', 'Very High']
loan_df['dti_category'] = pd.cut(loan_df['dti'], bins=b, labels=l, right=False)
```

```python
# Loan to Income Ratio Categories:
b3 = [0, 10, 17, 25, float('inf')]
l3 = ['Low', 'Medium', 'High', 'Very High']
loan_df['lti_category'] = pd.cut(loan_df['loan_to_income_ratio'], bins=b3, labels=l3, right=False)
```

```python
#Emp Length Categories
category_mapping = {'< 1 year': 0, '1 year': 1, '2 years': 2, '3 years': 3,'4 years':4,'5 years':5,
                    '6 years': 6,'7 years':7,'8 years':8,'9 years':9,'10+ years':10}
#ln['emp_category'] = ln['emp_length'].map(category_mapping)
b5 = [0, 2, 4, 9, 111]
l5 = ['Entry Level', 'Junior Level', 'Middle Level', 'Senior Level']
loan_df['emp_len_category'] = pd.cut(loan_df['emp_length'].map(category_mapping), bins=b5, labels=l5, right=False)
```

DATA CLEANING

CHECKING NULL IN COLUMNS

- After all the process there is one column 'emp_length' which has empty rows . For our analysis we can leave them as it is . Deleting or Filling in data may lead to some anomaly.

FINAL INFO

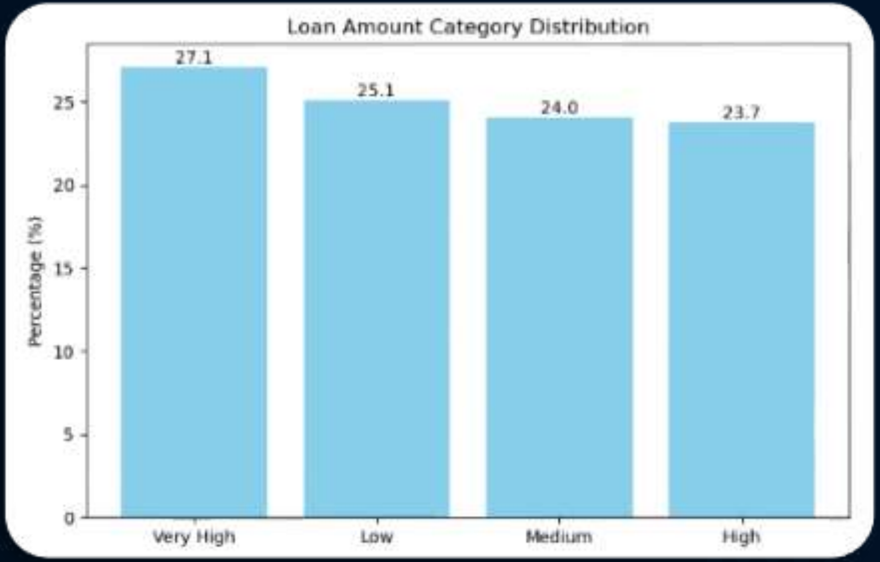
- There are 38577 rows and 26 columns in the dataset after the data cleaning process.

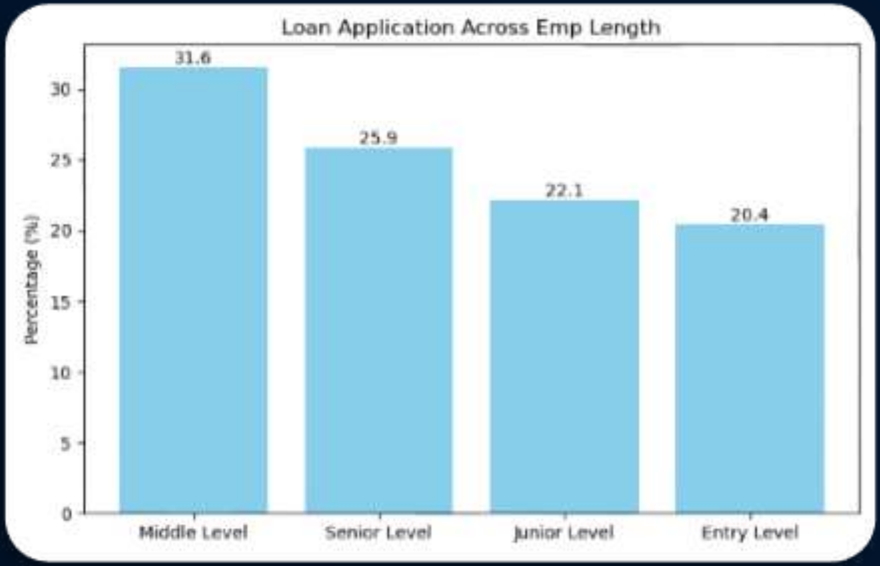
```
loan_df.shape
(38577, 26)
```

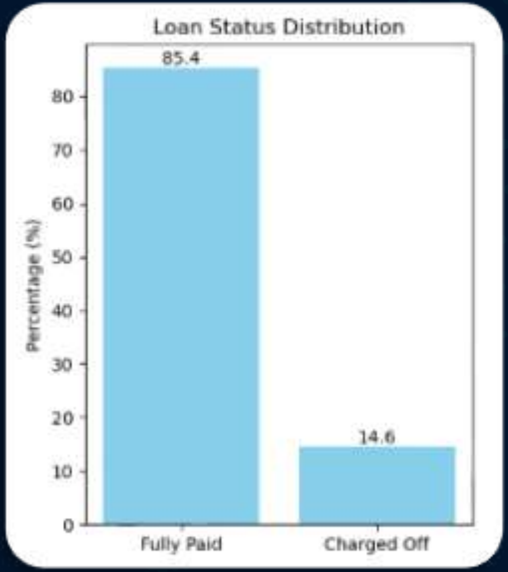
CHAPTER 3

UNIVARIATE ANALYSIS

ANALYSIS OF LOAN AMOUNT,EMPLOYMENT LENGTH,LOAN STATUS





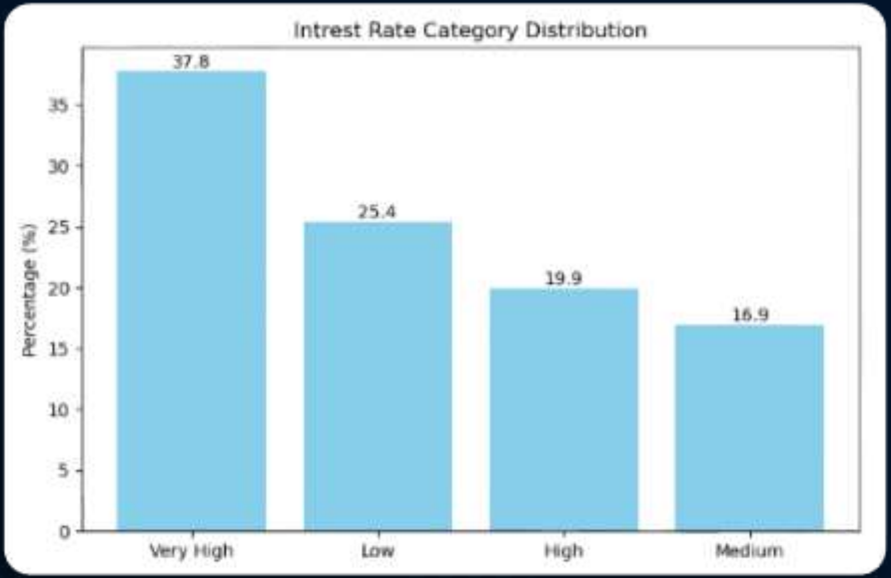


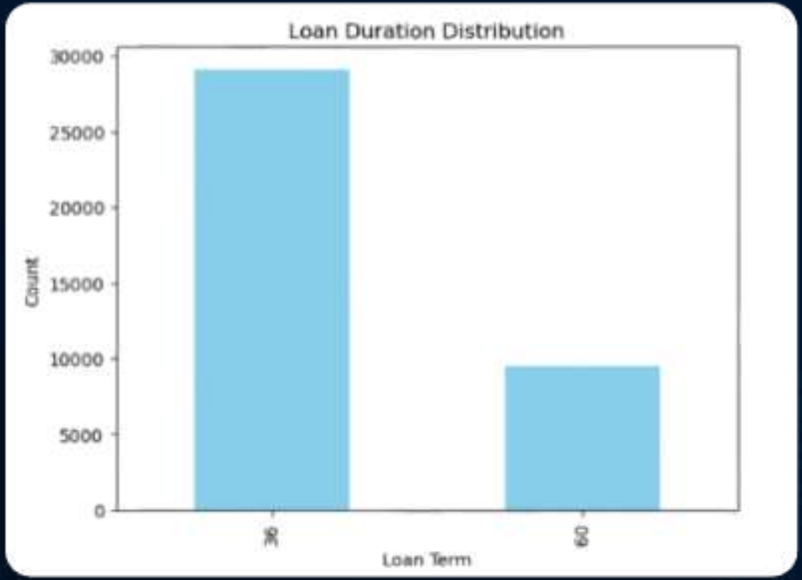
Loan Amount Distribution: Very High and low categories dominate the loan amount distribution, suggesting a preference for moderate loan amounts among borrowers.

Loan Application by Term: Senior Level loans have the highest application rate, indicating a possible trend of experienced individuals seeking loans more frequently.

Loan Status: A significant majority of loans are fully paid, which could reflect a healthy financial status among borrowers or effective loan management by the lender.

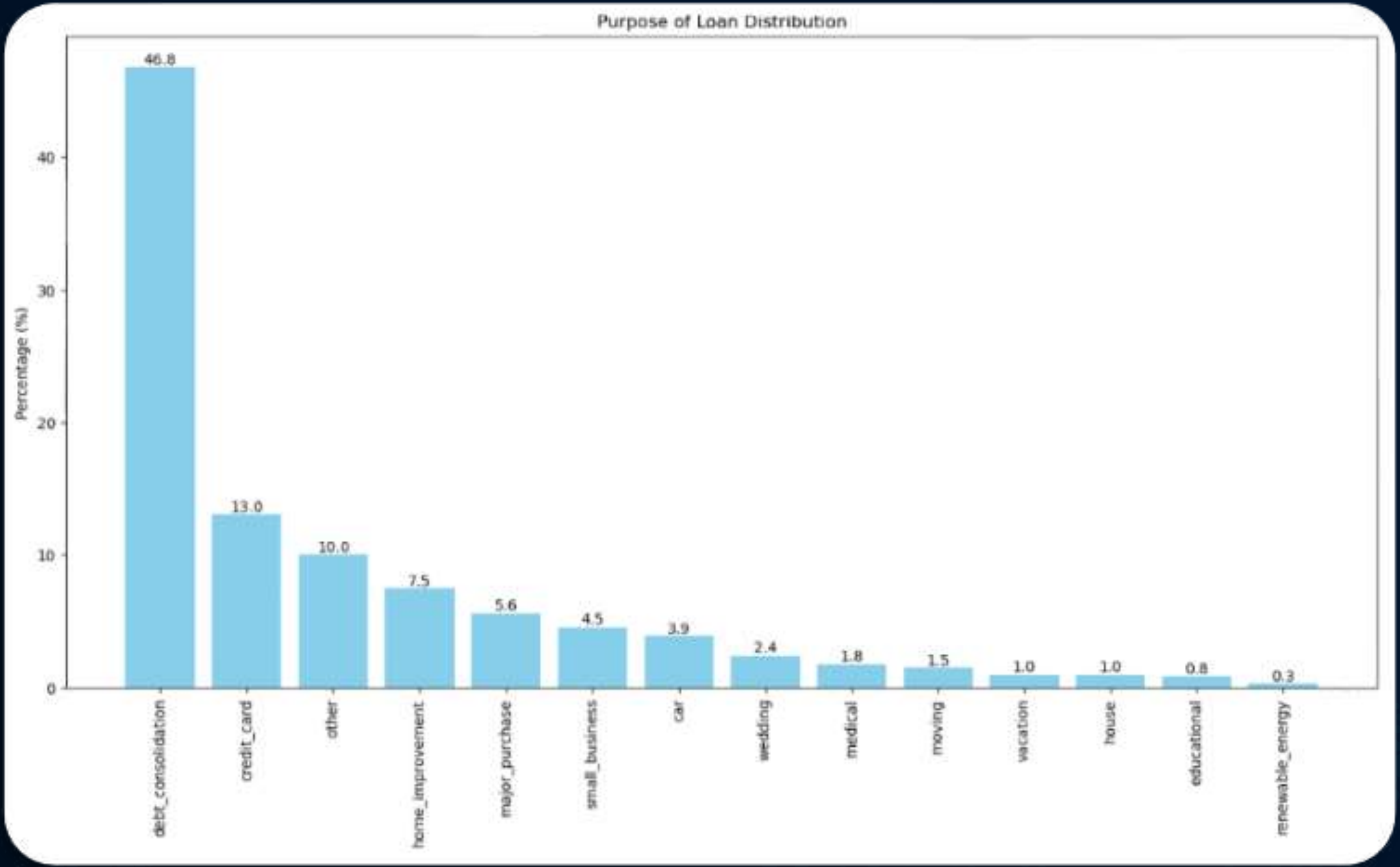
ANALYSIS OF INTREST RATE AND LOAN DURATION





- The two graphs provide the following insights:
- Interest Rate Distribution:
  - The "Very High" interest rate category has the largest share at approximately 31.8%.
  - The "Low" category follows with around 25.4%.
- Loan Duration:
  - The graph shows a single category "<=60 months" with a high count just above 30,000.
  - This indicates that the majority of loans have a term of 36 months .

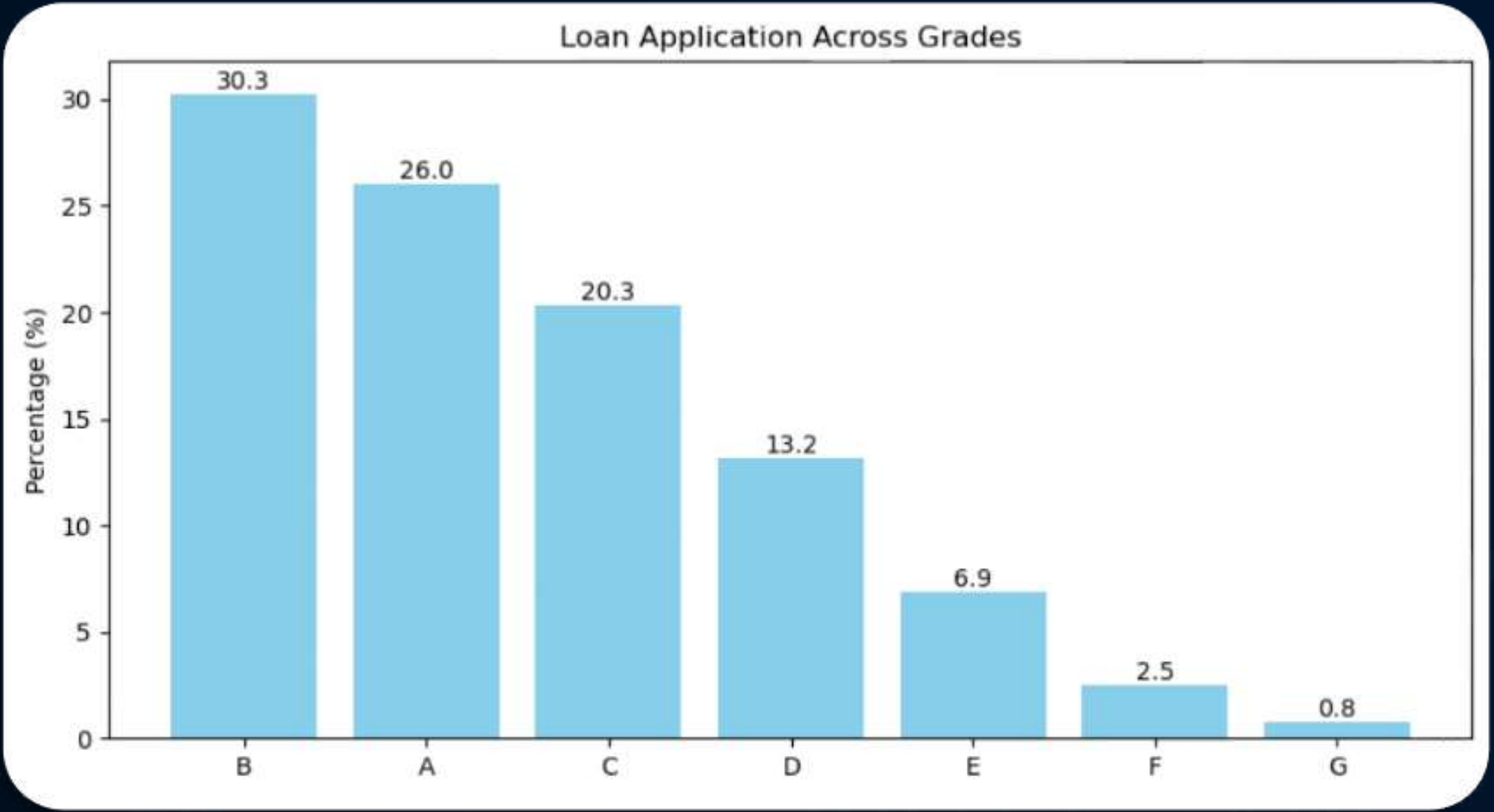
ANALYSIS OF PURPOSE OF LOAN



The bar graph titled "Purpose of Loan Distribution" provides the following insights:

1. Debt Consolidation: The most common purpose for loans, occupying a significant 46.1% of the total.
2. Credit Card: The second most common reason, accounting for 13.0% of loans.
3. Home Improvement: Represents 10.0%, indicating a notable number of borrowers investing in their homes.

ANALYSIS OF LOAN ACROSS GRADES



The bar graph provides insights into the distribution of loan applications by credit grade:

1 .Grade B: The most applications at approximately 30.3%.

2. Grade A: The second-highest with about 26.0%.

3. Grade C: Follows with roughly 20.3%.

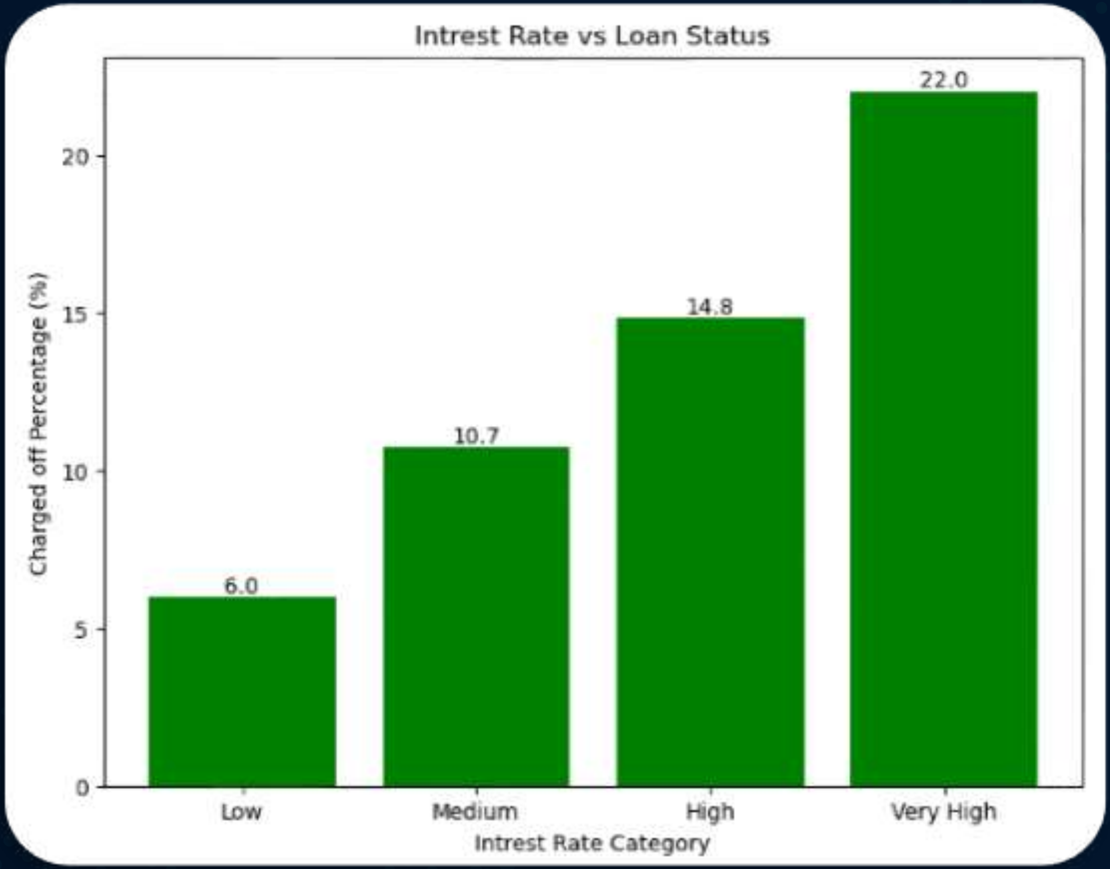
CHAPTER 4

BIVARIATE ANALYSIS

INTREST RATE VS LOAN STATUS

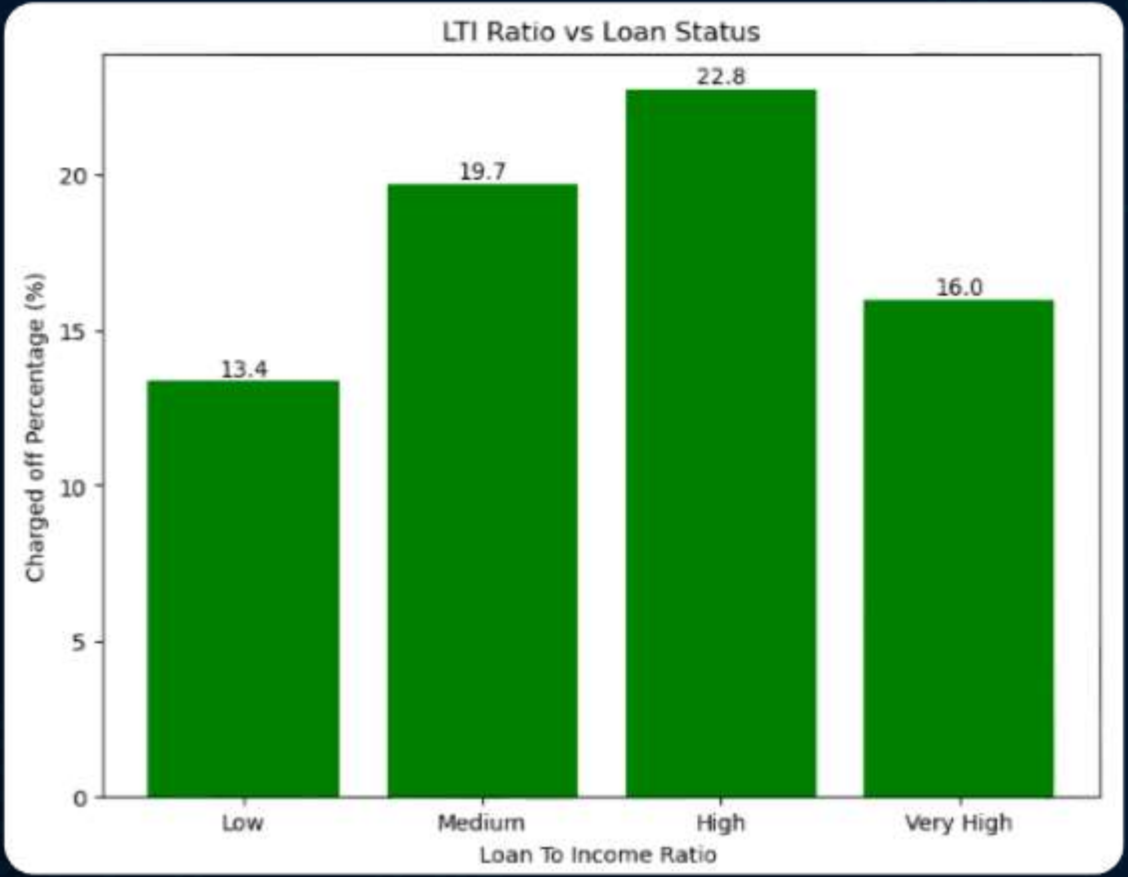
- The bar graph shows a clear trend where higher interest rates are associated with a greater percentage of charged-off loans. Here are the key insights:
- Interest Rate Categories:
    - Low: 6.0% charged-off loans
    - Medium: 10.7% charged-off loans
    - High: 14.8% charged-off loans
    - Very High: 22.0% charged-off loans

This pattern suggests that loans with higher interest rates carry a higher risk of not being repaid, which is an important consideration for credit risk management.



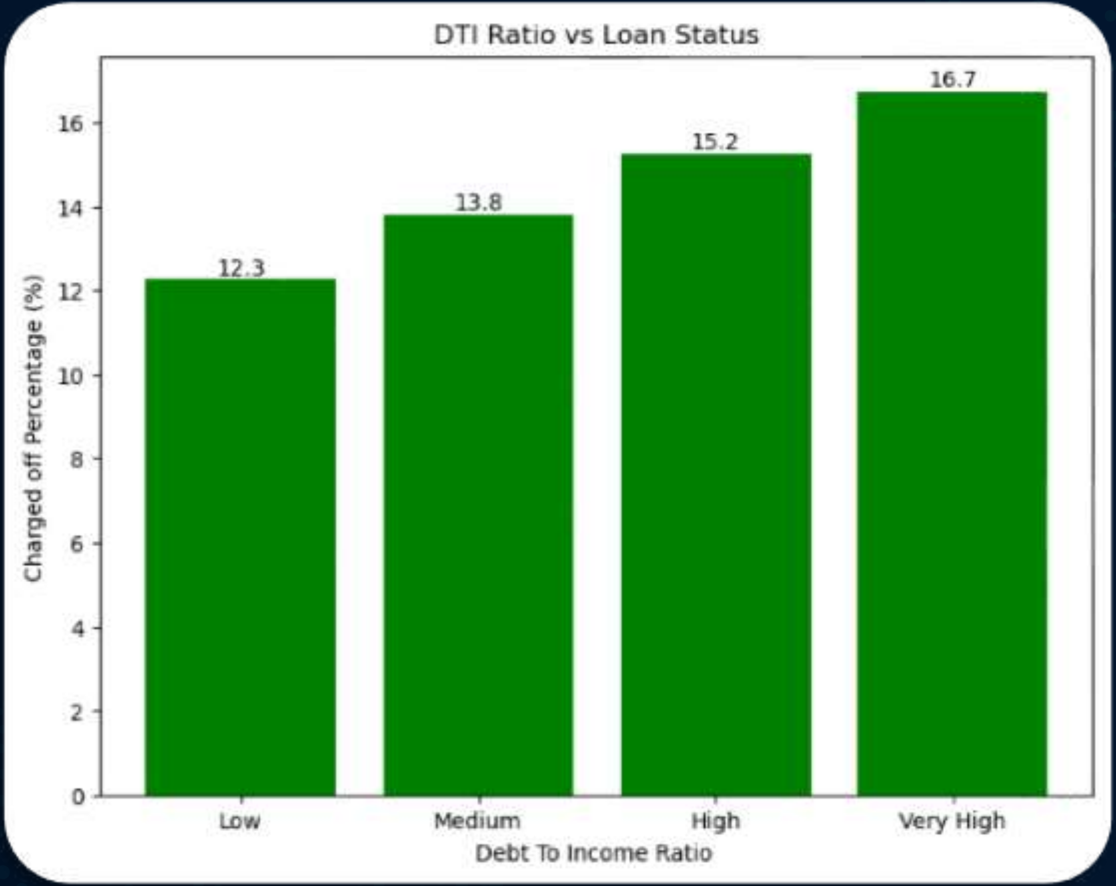
LTI RATIO VS LOAN STATUS

- The bar chart reveals the relationship between loan-to-income (LTI) ratios and loan charge-off percentages:
  1. Low LTI Ratio: 13.4% of loans are charged off.
  2. Medium LTI Ratio: 19.7% of loans are charged off.
  3. High LTI Ratio: 22.8% of loans are charged off.
  4. Very High LTI Ratio: 16.0% of loans are charged off.
- The data suggests that loans with high and very high LTI ratios tend to have higher charge-off rates, indicating a potential risk factor for lenders.

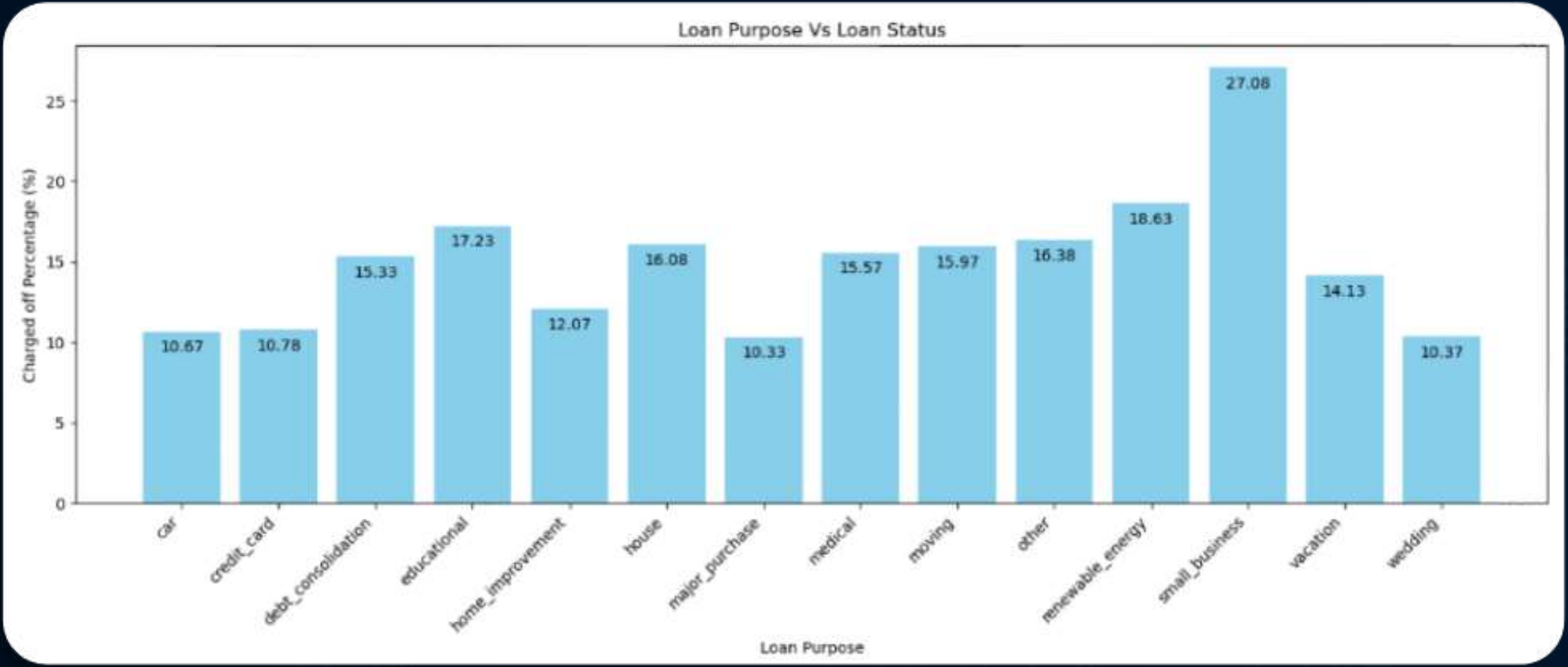


DTI RATIO VS LOAN STATUS

- The bar chart reveals the relationship between debt-to-income (DTI) ratios and loan charge-off percentages:
  1. Low LTI Ratio: 12.3% of loans are charged off.
  2. Medium LTI Ratio: 13.8% of loans are charged off.
  3. High LTI Ratio: 15.2% of loans are charged off.
  4. Very High LTI Ratio: 16.7% of loans are charged off.
- The data suggests that loans with high and very high DTI ratios tend to have higher charge-off rates, indicating a potential risk factor for lenders.



LOAN PURPOSE VS LOAN STATUS



The bar graph shows the charge-off rates for various loan purposes:

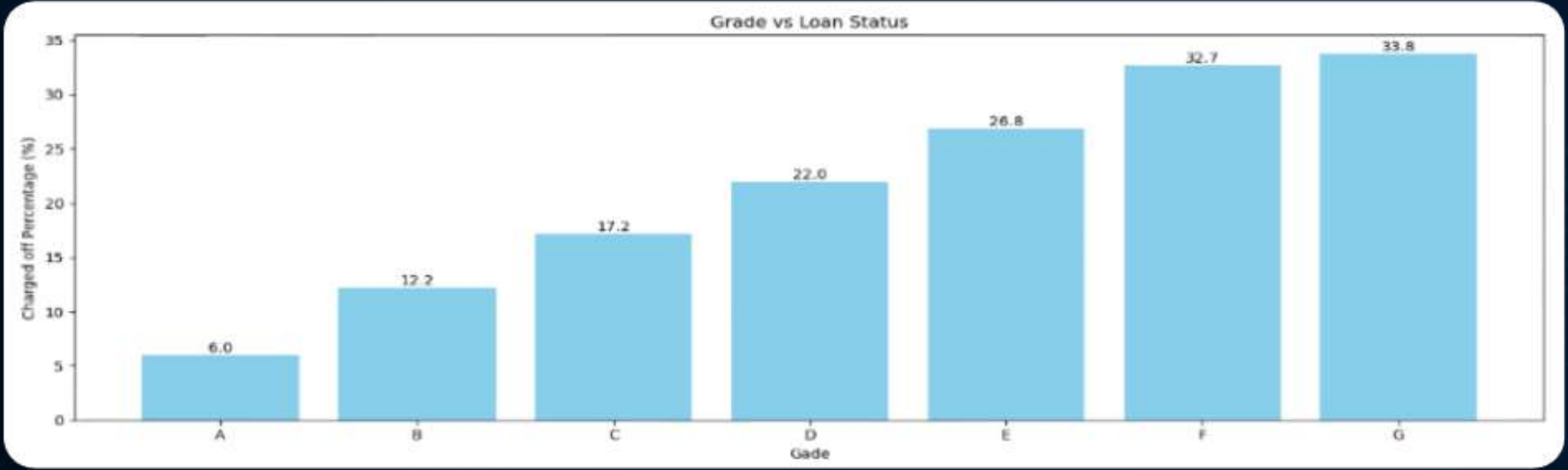
1 .Small Business: Highest risk with a 27.08% charge-off rate.

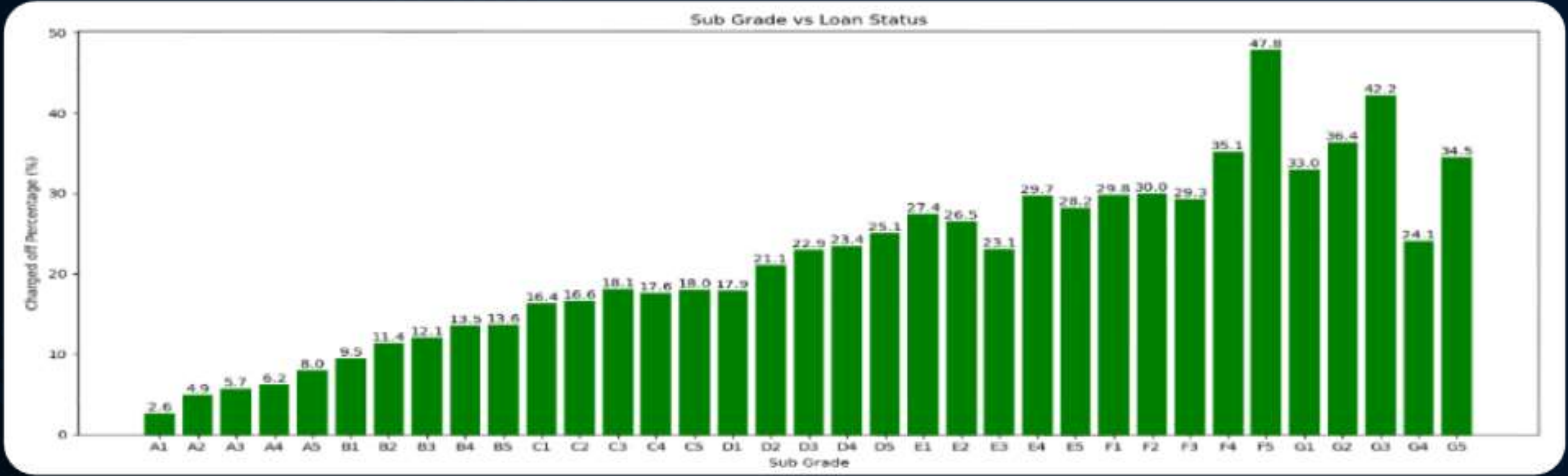
2. Renewable Energy: Next highest risk at 18.63%.

3. Educational: Charge-off rate of 17.23%.

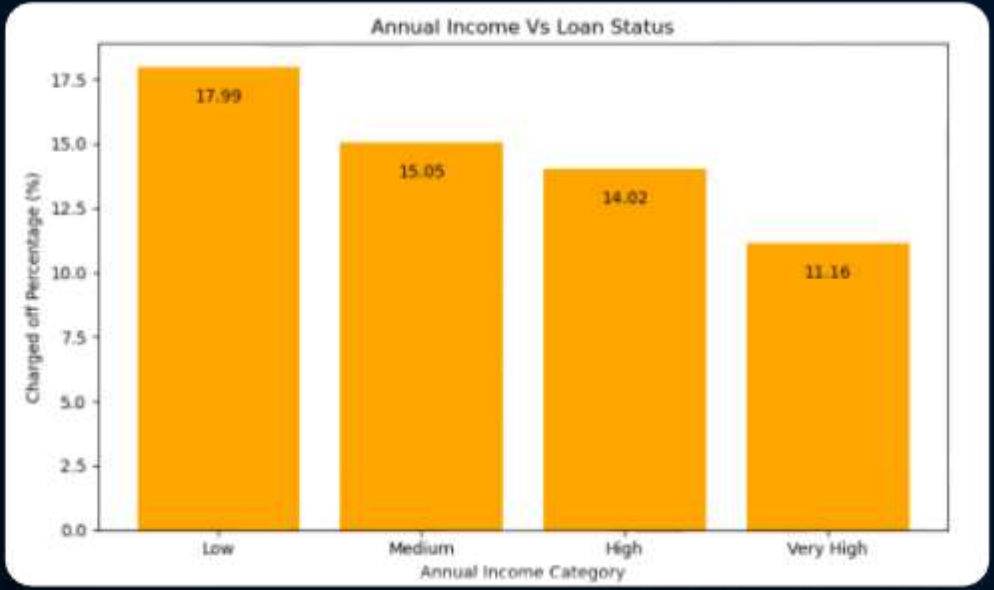
Loans for small businesses and Renewable Energy and Educational loans carry higher risks.

GRADE .SUB GRADE VS LOAN STATUS

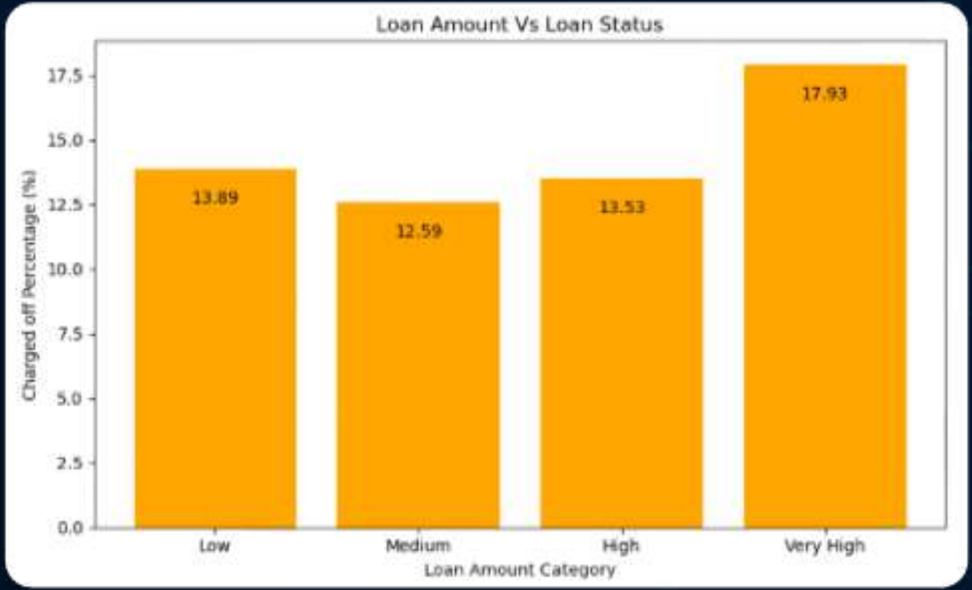




ANNUAL INCOME, LOAN AMOUNT VS LOAN STATUS

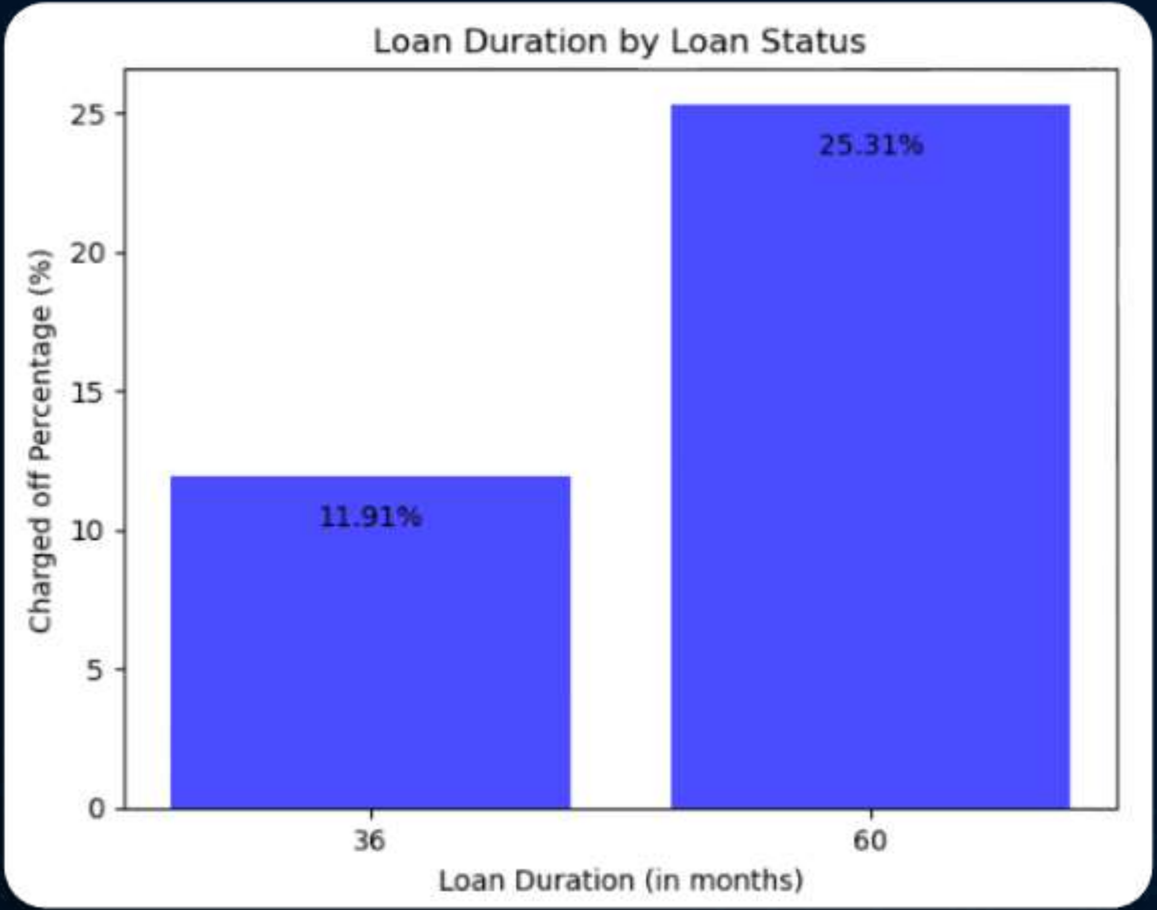


Income Correlation with Loan Charge-Offs: The graph suggests a trend where lower annual income categories have higher charged-off percentages. Specifically, the “Low” income category has the highest charged-off percentage at approximately 17.99%,



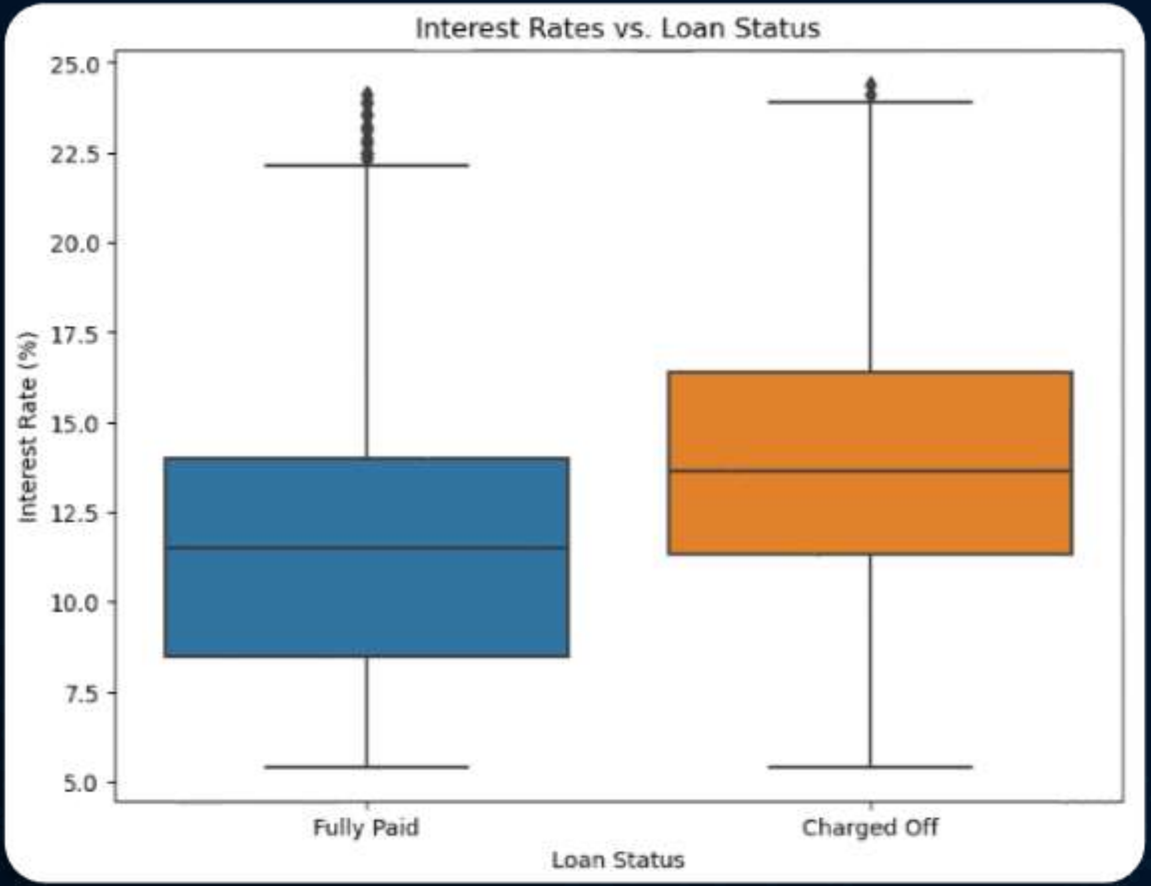
Higher Risk with Larger Loans: The graph suggests that loans categorized as “Very High” have the highest charged-off percentage at 17.93%, indicating a greater risk of default compared to other categories. In contrast, “Medium” loans have the lowest charged-off percentage at 12.59%. This implies that as the loan amount increases, so does the likelihood of charge-offs.

LOAN DURATION VS LOAN STATUS



- The graph indicates that loans with a term of 60 months have a higher charged-off percentage compared to those with a term of 36 months. This suggests that longer-term loans may carry a higher risk of charge-offs.

INTREST RATE VS LOAN STATUS



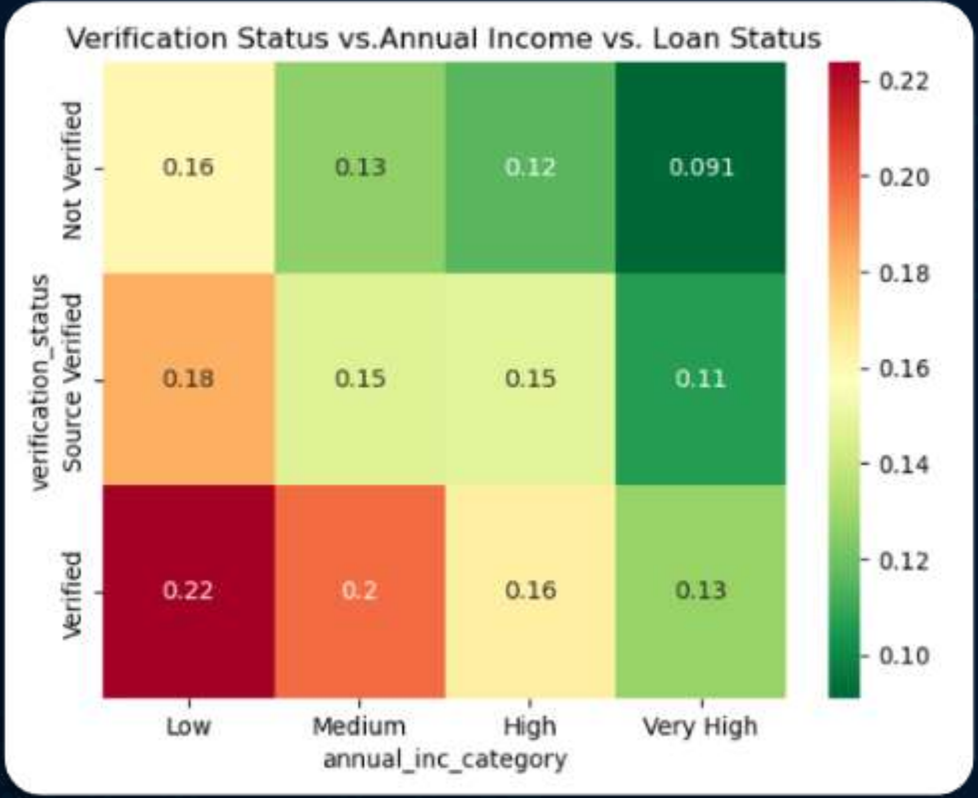
- Charged off loans tend to have higher interest rates.
- This suggests a correlation between higher rates and loan defaults.
- Fully paid loans generally cluster around lower interest rates.

CHAPTER 5

MULTIVARIATE ANALYSIS

VERIFICATION STATUS VS.ANNUAL INCOME VS. LOAN STATUS

- The heatmap analysis indicates the following insights, where red represents a higher probability of bad loans and green represents a higher probability of good loans:



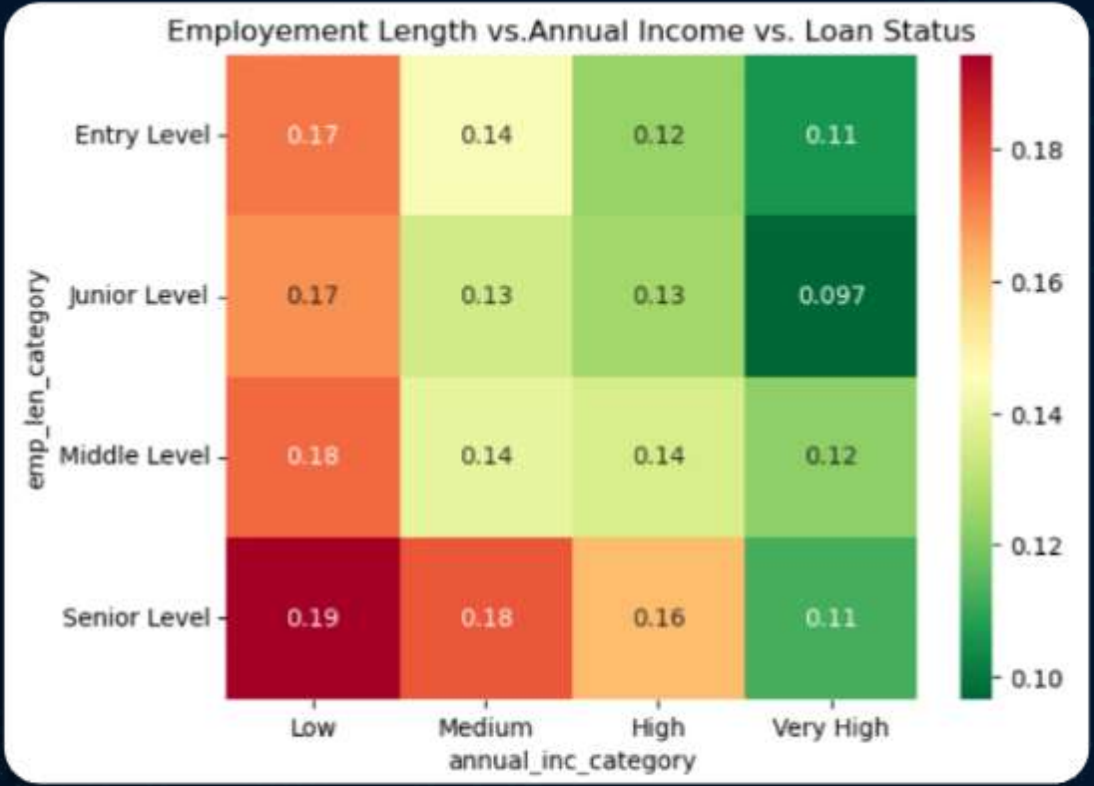
- Verification Status:
  1. Not Verified: Higher proportion of bad loans across all income levels.
  2. Source Verified: Moderate risk with a mix of good and bad loans.
  3 Verified: Lower proportion of bad loans, especially in higher income categories.

- Annual Income:
  1. Low: Higher risk of bad loans regardless of verification status.
  2. Medium to Very High: As income increases, the likelihood of good loans increases, particularly when verified.

EMPLOYEMENT LENGTH VS.ANNUAL INCOME VS. LOAN STATUS

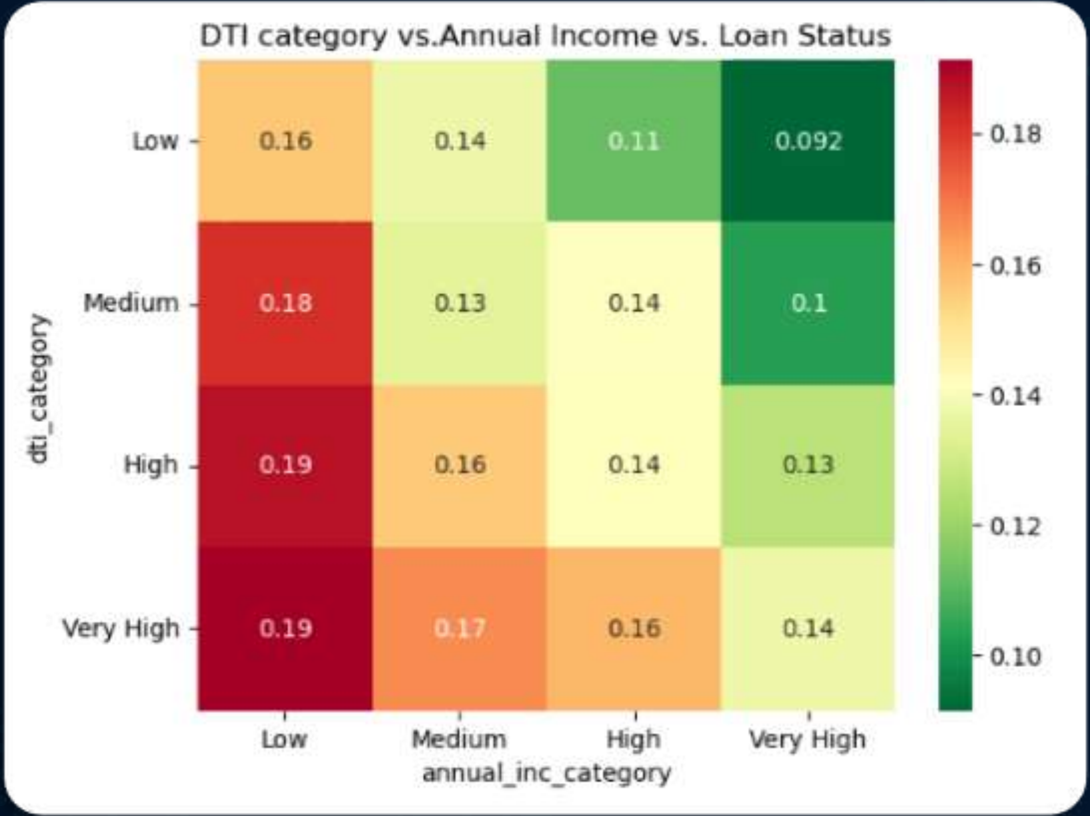
- Overall, the trend suggests that individuals with higher annual incomes tend to have lower charge-off percentages across all employment levels, indicating a potential correlation between income level and loan repayment reliability.

- Additionally, Senior Level employees with Very High annual incomes have the lowest charge-off percentage, suggesting that experience and higher income may contribute to financial stability and loan repayment success.

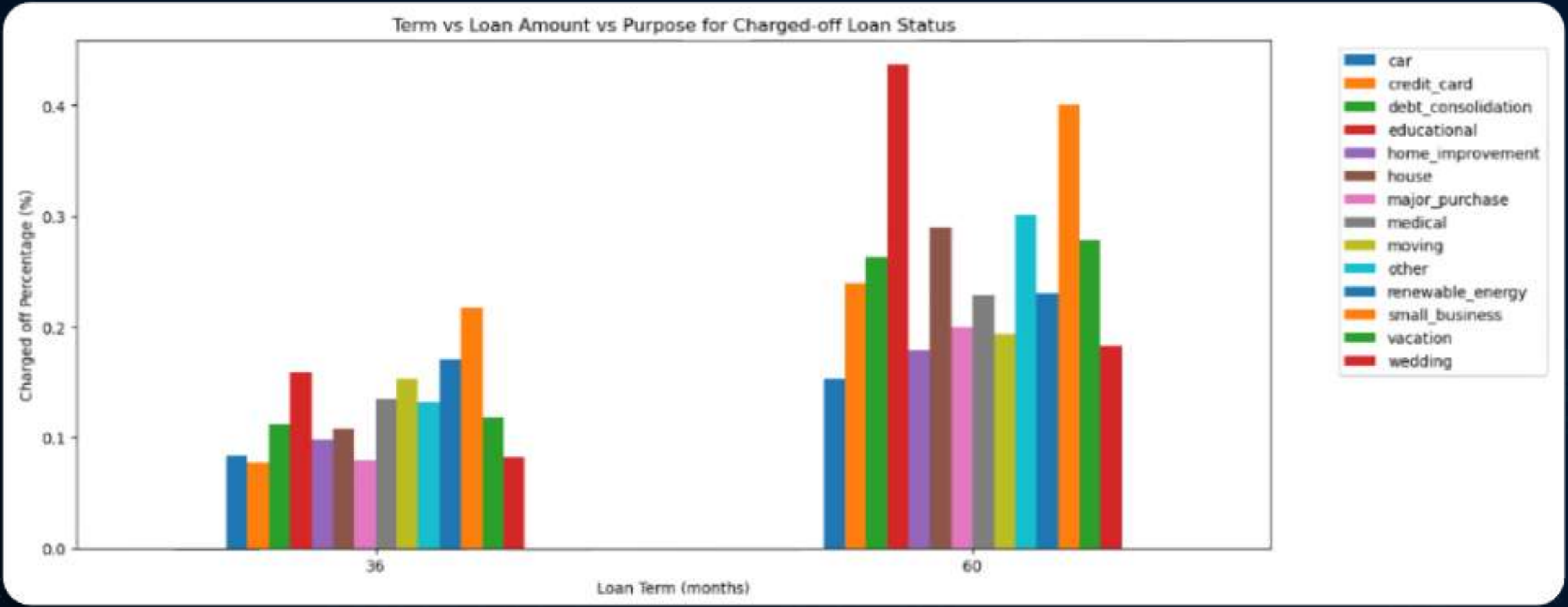


DTI CATEGORY VS.ANNUAL INCOME VS. LOAN STATUS

- The data suggests that individuals with very high incomes have the lowest charge-off percentages at low DTI levels, indicating better financial stability. However, as DTI increases, even those with very high incomes show an increased risk of loan charge-offs.

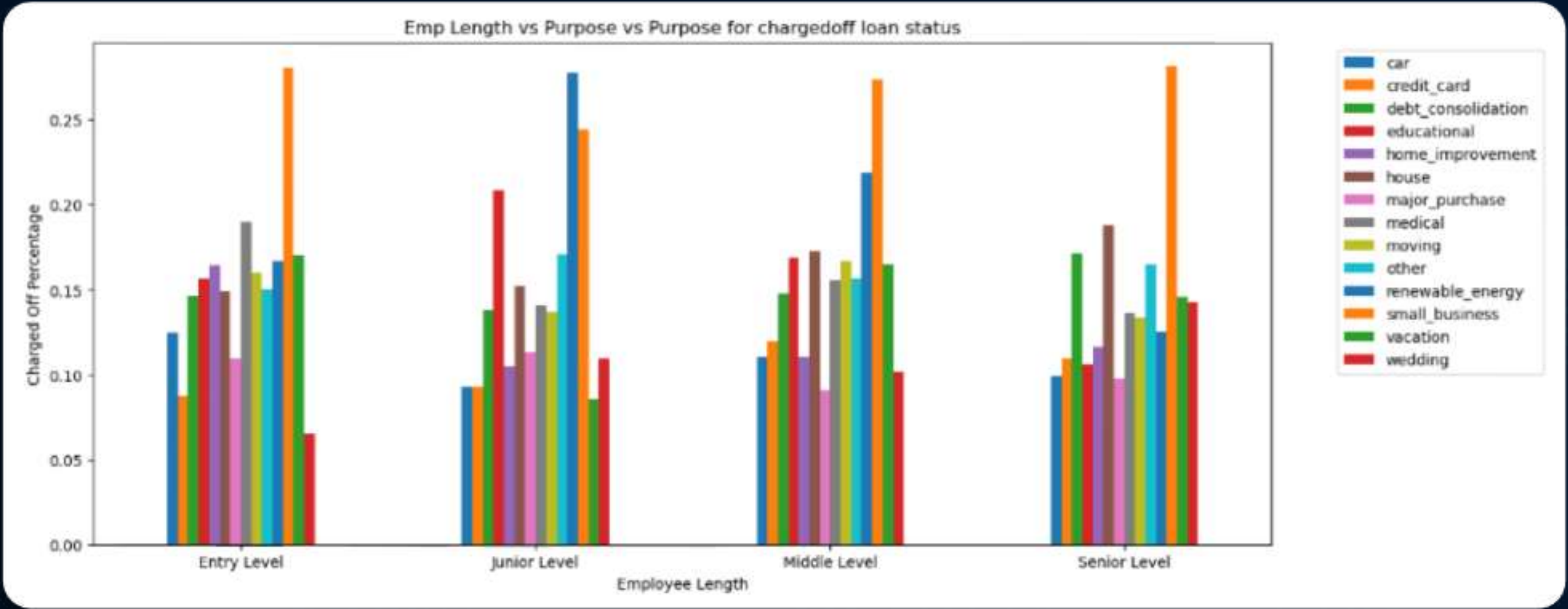


TERM VS LOAN AMOUNT VS PURPOSE FOR CHARGED-OFF LOAN STATUS

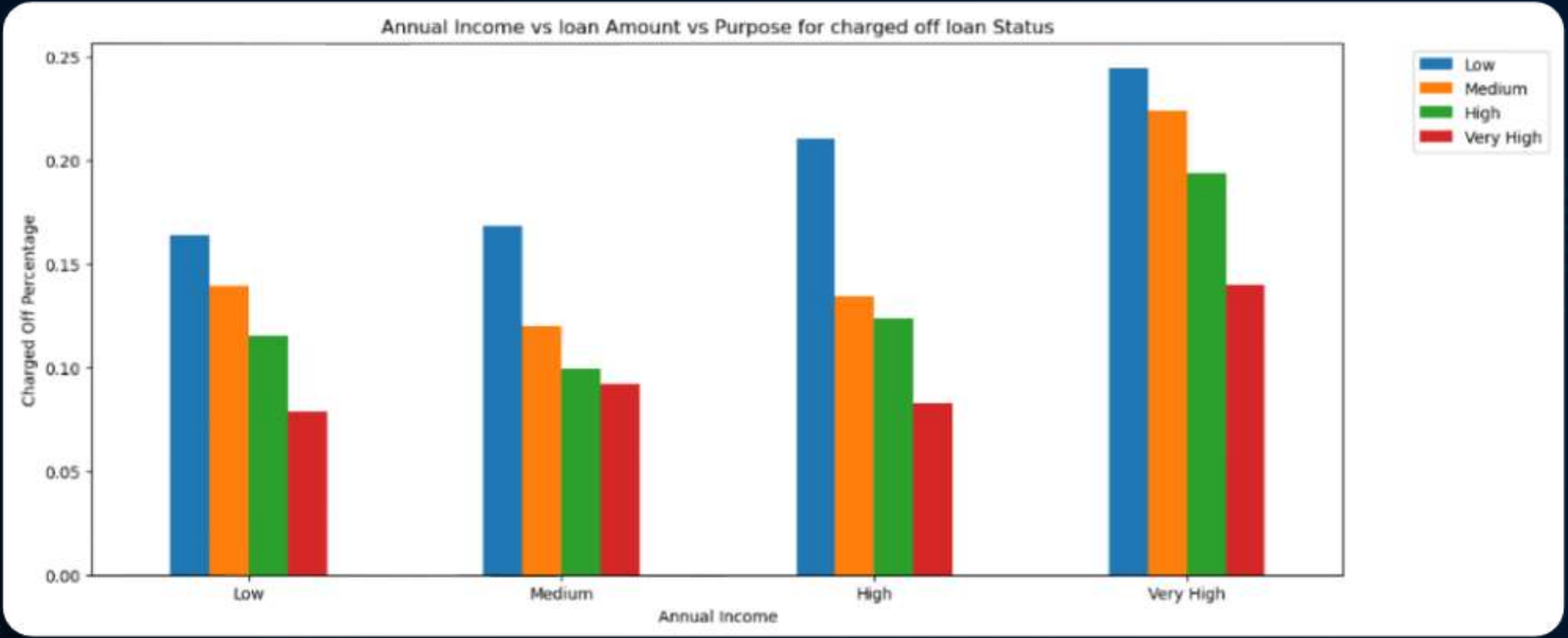


- **Loan Term:** Typically, longer loan terms might show a higher charge-off rate due to the increased risk over time.
- **Loan Amount:** Higher loan amounts could correlate with a higher charge-off rate as they represent a larger financial burden on the borrower.
- **Loan Purpose:** Certain loan purposes, such as small business or debt consolidation, might exhibit higher charge-off rates compared to personal or educational loans.

EMP LENGTH VS PURPOSE VS PURPOSE FOR CHARGED OFF LOAN STATUS



EMP LENGTH VS PURPOSE VS PURPOSE FOR CHARGED OFF LOAN STATUS



CHAPTER 6

CONCLUSIONS

CONCLUSION

Based on the analysis, the driving factors (or strong indicators) behind loan defaults can be summarized as follows:

- Higher Loan Amounts:
  - There is a correlation between higher loan amounts and an increased risk of default.
  - Charged-off loans tend to have higher average loan amounts compared to fully paid loans.

- Higher Interest Rates:
  - Loans with higher interest rates are more likely to end up as charged off.
  - Interest rates play a significant role in determining the risk of default.

- Lower Credit Grades:
  - Lower credit grades (e.g., grades D, E, F, G) are associated with a higher proportion of charged-off loans.
  - Applicants with lower credit grades pose a higher risk of default.

CONCLUSION

- Higher Debt-to-Income (DTI) Ratios:
  - Loans with higher DTI ratios are more likely to default.
  - Charged-off loans tend to have higher DTI ratios compared to fully paid loans.

- Employment Length:
  - Although not explicitly mentioned in the summary, employment length can also be a driving factor.
  - Applicants with shorter employment histories or unstable employment records might be at a higher risk of default.

- Annual Income:
  - Although not strongly highlighted in the provided insights, annual income often correlates with the ability to repay loans.
  - Higher income levels generally indicate better repayment capability.