Lending Club Case Study

BISWARANJAN AICH
ML C69 SEPTEMBER 2024 BATCH
18-11-2024

Table Of Content:

Business Problem Understanding

Context:

The company facilitates loans to urban customers through an online platform.
Loan approval decisions impact business profitability and risk exposure.

Key Risks:

Loss of business if loans to creditworthy applicants are rejected.
Financial losses if loans to risky applicants are approved and they default.

Dataset Description:

- Contains information on past applicants, loan statuses, and default history etc.

Key categories:

- **Fully Paid:** No issues, loan closed.
- **Current:** Still paying, no defaults.
- **Charged-off:** Defaulted, causing credit loss.

Business Objectives

Primary Objective:

- Minimize credit loss by identifying **risky loan applicants** before approval.

Expected Outcomes:

- Develop patterns to assess applicant risk.
- Provide actionable insights to:
- Deny risky loans.
- Adjust loan amounts.

Purpose of Analysis:

- Identify driving factors behind loan defaults (e.g., applicant and loan attributes).
- Enable better risk and portfolio management.

Data Understanding

Load Data & Initial Inspection:

- Import the dataset, check the data types, dimensions, null values, and summary statistics. Ensure no overlooked data quality issues by reviewing the distribution and types of values.
- Go through the given Data dictionary to understand the meaning of each and every columns.

Interpretation of Variables:

- Study the context of each variable and annotate the dataset in the notebook with clear descriptions of each, emphasizing their relevance to default prediction. This helps when selecting driver variables.

Identify and Document Data Quality Issues:

- Clearly state all observed data quality issues, such as missing values, outliers, duplicate entries, or inconsistencies.

Data Cleaning And Manipulation

Address Missing Values:

- Impute or remove missing values based on the business significance of each variable.
- Drop columns with excessive missing values (>80% missing)
- For 'desc': drop if irrelevant, else impute with 'No description'

Outlier Detection and Treatment:

- Use visualization techniques (box plots, histograms) to identify and cap or transform outliers.

Feature Engineering:

- Derive new variables that could better represent default risk factors (e.g., debt-to-income ratio).

Data Type and Format Adjustments

- Ensure dates and strings are cleaned and converted into appropriate formats for ease of analysis.

Data Visualization

Effective Plotting:

- Appropriate plots being used based on variable types (e.g., scatter plots for continuous data relationships, bar plots for categorical variables), making sure visual support the narratives.

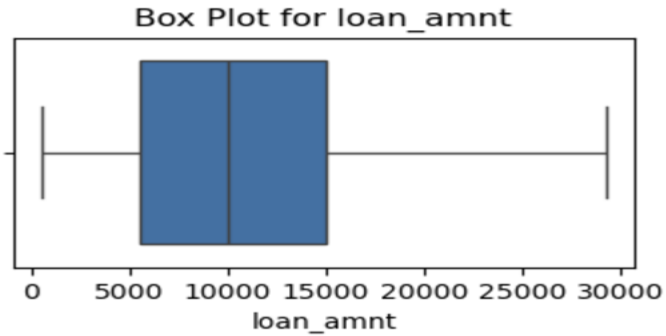
- Understanding distribution of individual variables, relationship between two variables and complex relationships

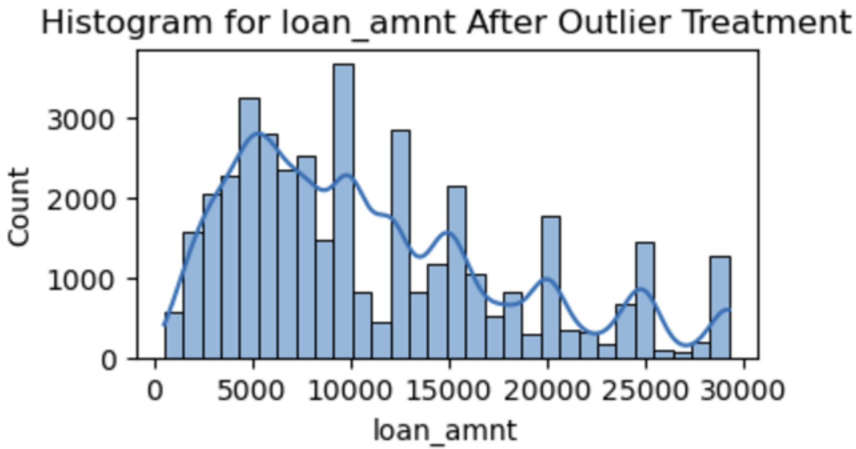
Visualization of loan_amnt after Outlier treatment

upper bound: 29250.0

Example

- Visualize Outliers of loan_amnt Using Box Plots
- Identify Outliers Using the IQR Method
- Then Cap Outliers for loan_amnt with upper bound 29250.0



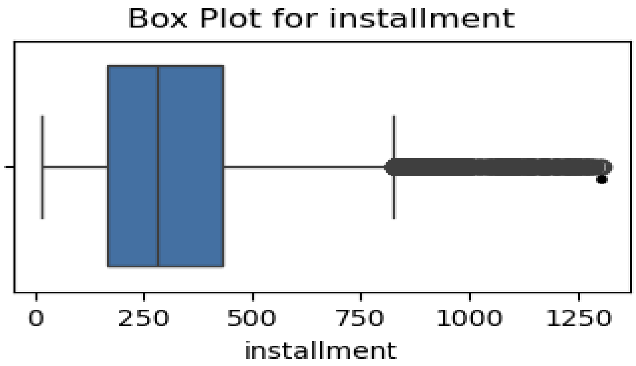


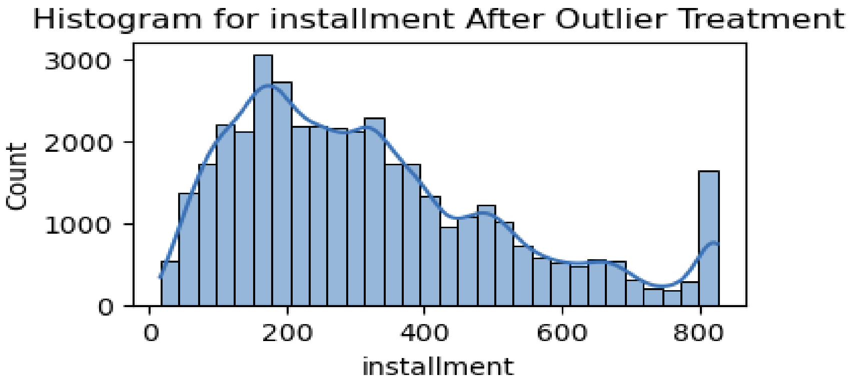
Visualization of installment column after Outlier treatment

upper bound:826.42

Example2

- Visualize detected outliers of installment using Box Plots
- Identify Outliers Using the IQR Method
- Then Cap Outliers for installment with upper bound: 826.42





Data Analysis

Univariate Analysis

Segmented Univariate Analysis

Bivariate Analysis

Univariate Analysis:

- Systematically explore each variable to understand its distribution and relevance. Use visuals like histograms and bar charts, especially for continuous variables like loan amount, income, and credit history.

Segmented Univariate Analysis

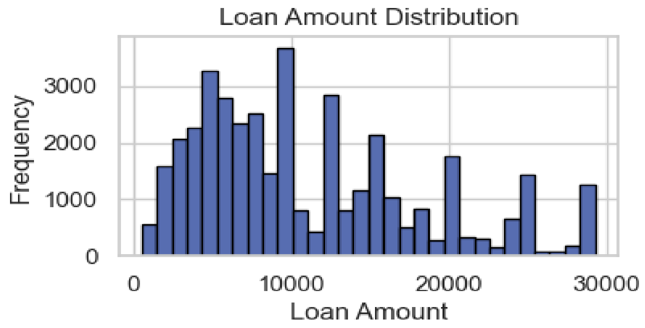
- Perform segmented analysis by default status (charged-off, fully paid) for each variable. Identify differences that can indicate risky patterns.

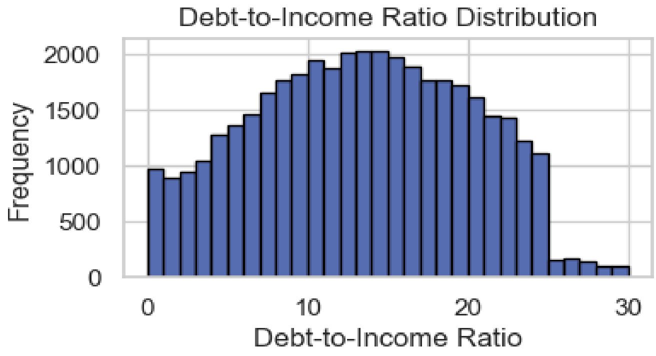
Bivariate Analysis

- Relationships between pairs of key variables. Look for patterns where certain variable combinations indicate higher default risk.
- Use scatter plots to reveal insights.
- Identify and summarize the key relationships, emphasizing those that can help differentiate defaulters from non-defaulters.

Univariate Analysis

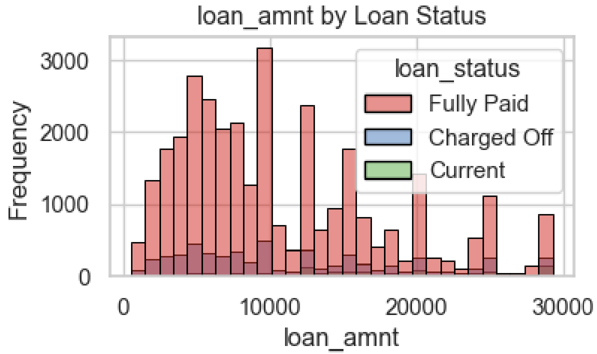
- In loan amount distribution, within range(0-10000) more loan get distributed more than 3k
- DTI(Debt-to-income-ratio) distribution, within the range(10-20) get high frequency equivalent to 2k

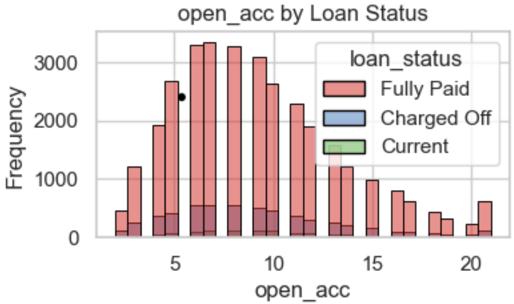




Segmented Univariate Analysis

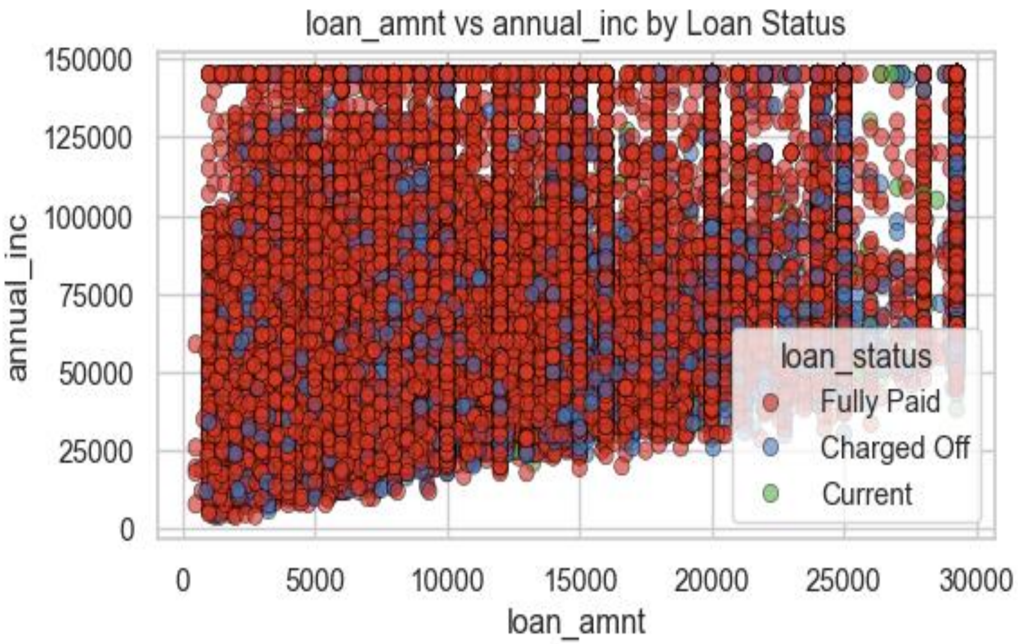
- Within the range (0-10k) having maximum (>3k) loans having Fully paid status.
- Within range(5-10) having maximum open account with more than (>3k) loans having status Fully paid.





Bivariate Analysis

- Plot a scatter plot to visualize relationships between two key variables(loan_amnt & annual_inc), segmented by loan status (default vs non-default).



Observations & Insights

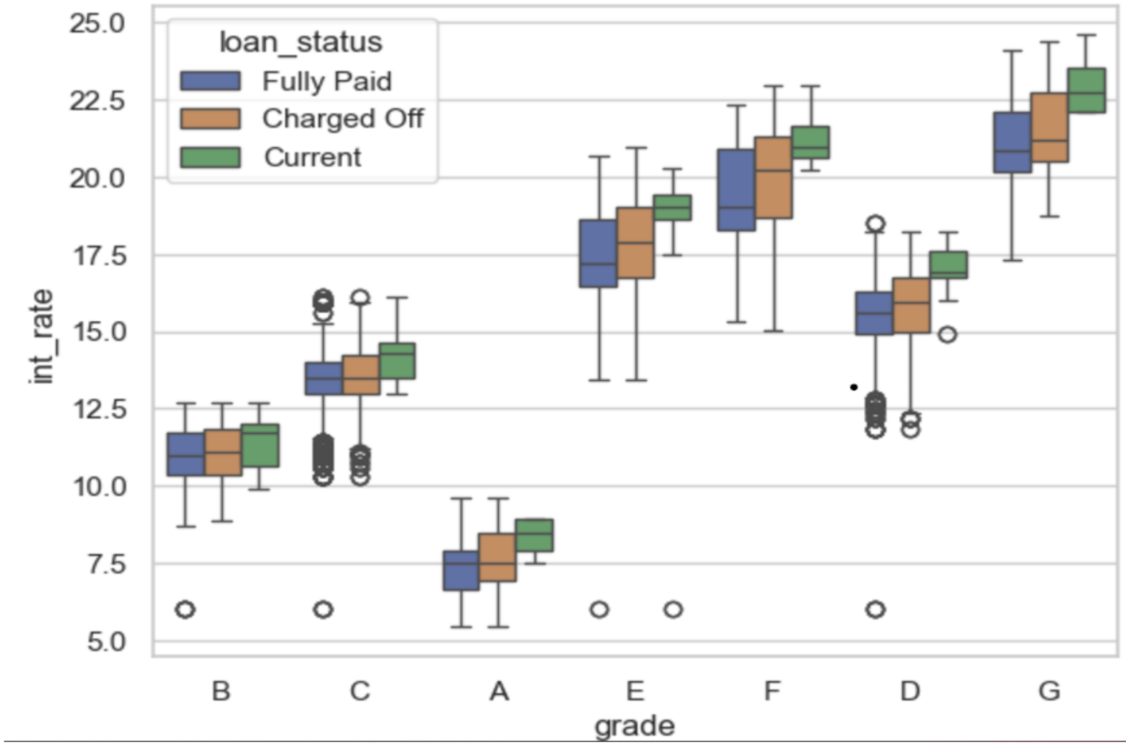
int_rate vs Grade

loan_status

Charged Off 13.82
Current 15.03
Fully Paid 11.61

Observation:

1. Higher interest rates are linked to default loans (Charged Off) and lower grades.

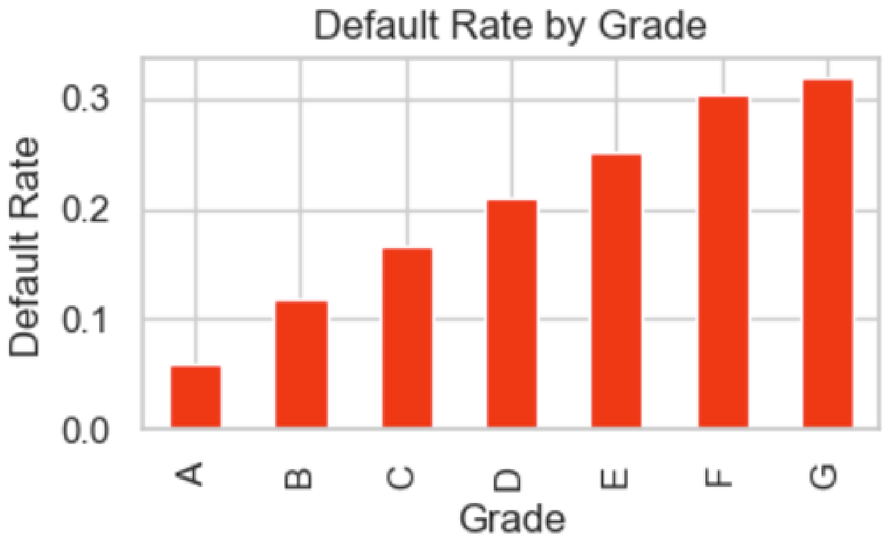


Observations after Data Analysis

Default Rate VS Grade

Observation:

2. Default rates increase as grades worsen (A to G)



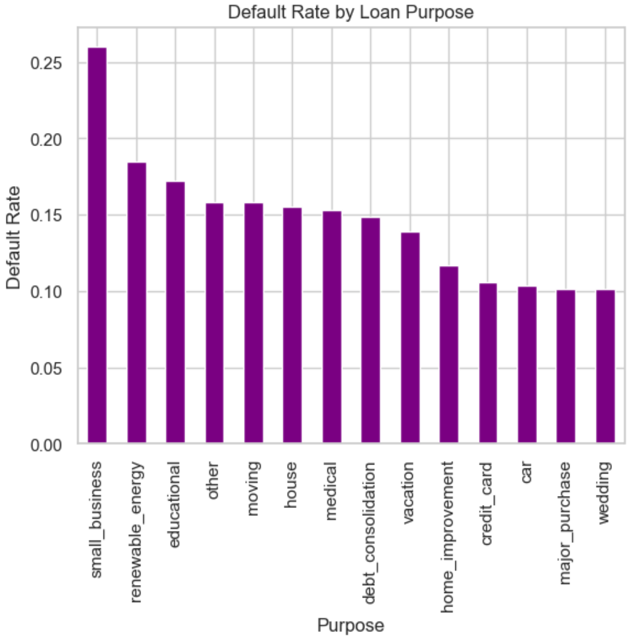
Observations after Data Analysis

Default Rate VS Grade

Observation:

3. Loans for small_business and debt_consolidation show higher defaults.

```
default_rate_by_purpose = df[df['loan_status'] == 'Charged Off'].groupby('purpose').size() / df.groupby('purpose').size()
default_rate_by_purpose.sort_values(ascending=False).plot(kind='bar', color='purple')
plt.title("Default Rate by Loan Purpose")
plt.ylabel("Default Rate")
plt.xlabel("Purpose")
plt.show()
```



Observations after Data Analysis

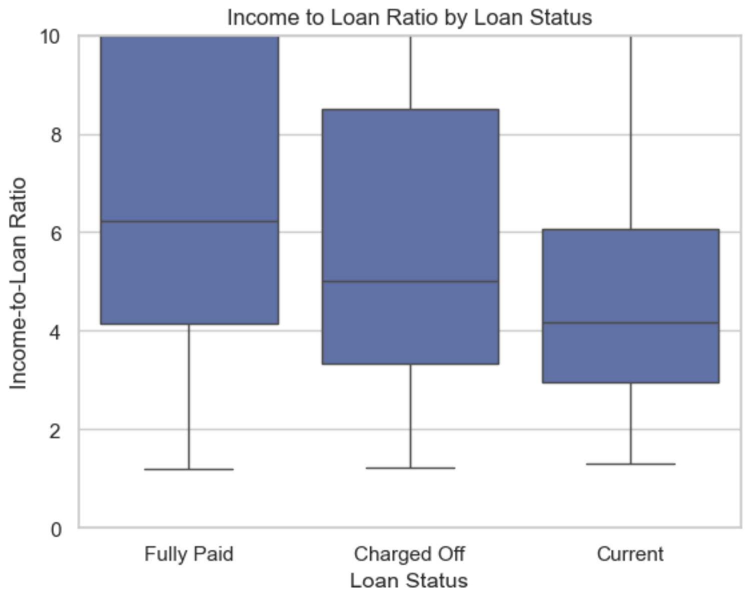
loan_status
Charged Off 5.00
Current 4.17
Fully Paid 6.24

Observation:

4. Lower ratios (<2x income vs. loan amount) are associated with higher defaults.

```
df['income_to_loan_ratio'] = df['annual_inc'] / df['loan_amnt']
sns.boxplot(data=df, x='loan_status', y='income_to_loan_ratio')
plt.title("Income to Loan Ratio by Loan Status")
plt.ylabel("Income-to-Loan Ratio")
plt.xlabel("Loan Status")
plt.ylim(0, 10)  # Focus on the key range
plt.show()

# Summary statistics
ratio_summary = df.groupby('loan_status')['income_to_loan_ratio'].median()
print(ratio_summary)
```



Observations after Data Analysis

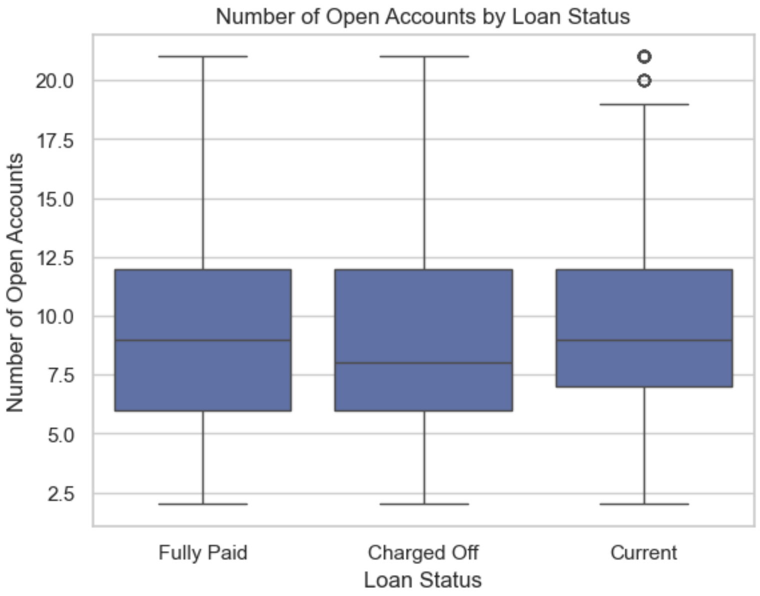
Observation:

5. Loans with shorter terms (36 months) tend to have a higher default rate compared to longer terms (60 months).

Observations after Data Analysis

Observation:

6. Open account does not impact on loan defaulter.



Thank You