Lending Club Case Study

For ,

Institute of Information Technology, Bangalore
Upgrad Education

Prepared by,

Naveen Cheruku
Shashank Mishra

Agenda

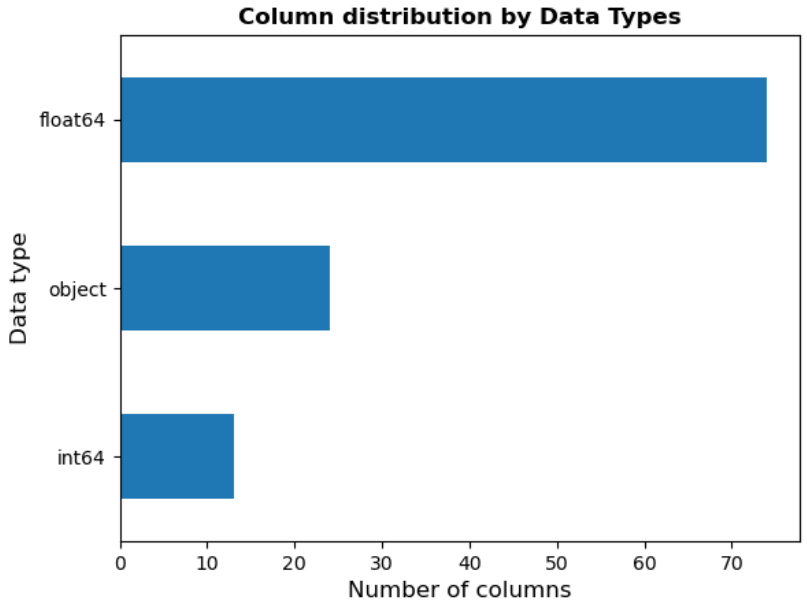
- Objective
- Data Source
- Approach
- Data Cleaning and Manipulation
- Data shortlisting required from Business Perspective
- Data Conversion
- Outlier analysis and Handling
- Univariate Analysis and Outlier Handling
- Segmented Univariate Analysis
- Derived Metrics
- Data points of Interest
- Bivariate Analysis

Objective

To assist Lending Club in mitigating credit losses by identifying high-risk loan applicants through Exploratory Data Analysis. The aim is to understand key factors leading to loan default, thereby enhancing portfolio and risk assessment.

Data Source

- Data provided in CSV file
  - Associated data dictionary explains meaning of data columns



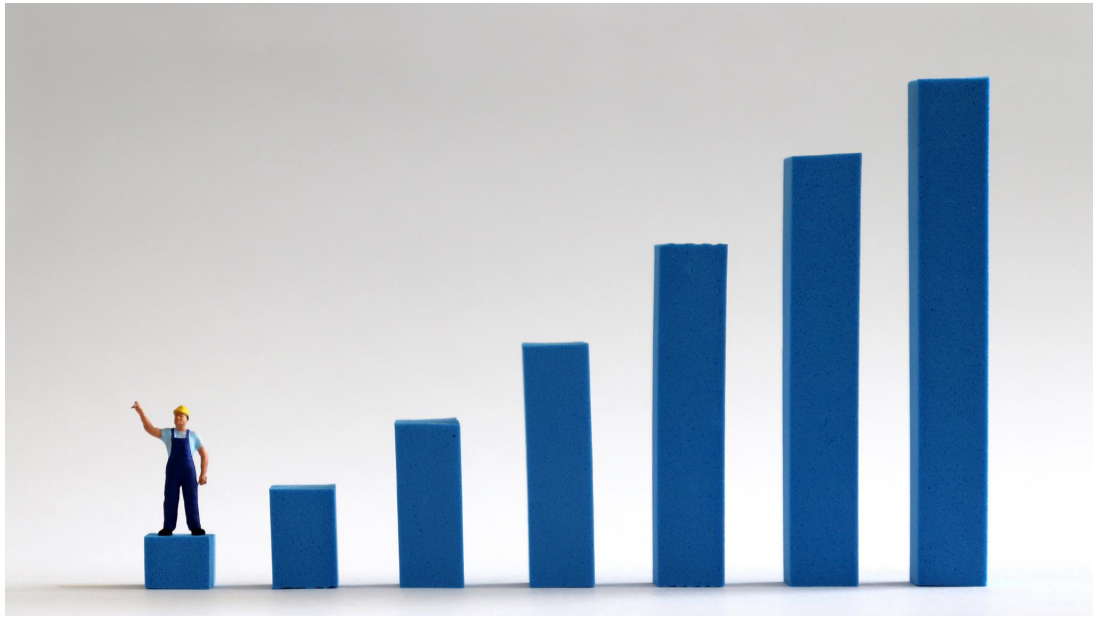
Approach

- Data Cleaning
  - Remove irrelevant or incorrect data
- Univariate Analysis
  - Analyze individual variables
- Segmented Variate Analysis
  - Analyze variables by segment
- Bivariate Analysis
  - Analyze relationship between two variables
- Derived Metrics
  - Create new metrics from existing data
- Collate and Conclude Findings
  - Summarize findings and provide suggestions

Data Cleaning and Manipulation: Identification and Removal of Irrelevant Columns

- Columns with no unique values
  - tax_liens, delinq_amnt, chargeoff_within_12_mths, acc_now_delinq, application_type, policy_code, collections_12_mths_ex_med, initial_list_status, pymnt_plan
- Columns with all unique values
  - id, url and member_id are all unique. We can use one of them as identifier and drop other two
- Columns which are not useful
  - emp_title and title columns seem to be just designations. Hence they can be removed

Data Cleaning and Manipulation: Missing values in rows



- Check maximum number of missing values in a row
  - Ignore if less than 5
- Number of null values in irrelevant columns
  - Leave for now, drop in final evaluation

Data Cleaning and Manipulation: Handling 'null' values

- Evaluate Imputation of Null Values
  - Check if null values can be imputed
  - If not, remove them

Data Cleaning and Manipulation: Other Issues

Remove padded spaces from 'Term' column

Data shortlisting required from Business Perspective

- Relevant attributes help make decision before loan sanction
  - Irrelevant attributes are removed
- Loans with status Current are removed
  - Only Fully paid or Charged Off loans are required
- Columns representing data gathered after loan sanction are dropped
  - Examples: delinq_2yrs, earliest_cr_line, inq_last_6mths
- Zip_code is a masked column and is dropped
  - Funded_amnt_inv is internal data and not required

Data Conversion

- Term column
  - Convert to 'int' after removing string literal 'months'
- Grade column
  - Convert to 'category' type
- Sub_grade column
- Emp_length column
- Home_ownership column
- Verification_status column
- Purpose column
- Addr_state column
- Issue_d
- Int_rate column

Outlier analysis and Handling

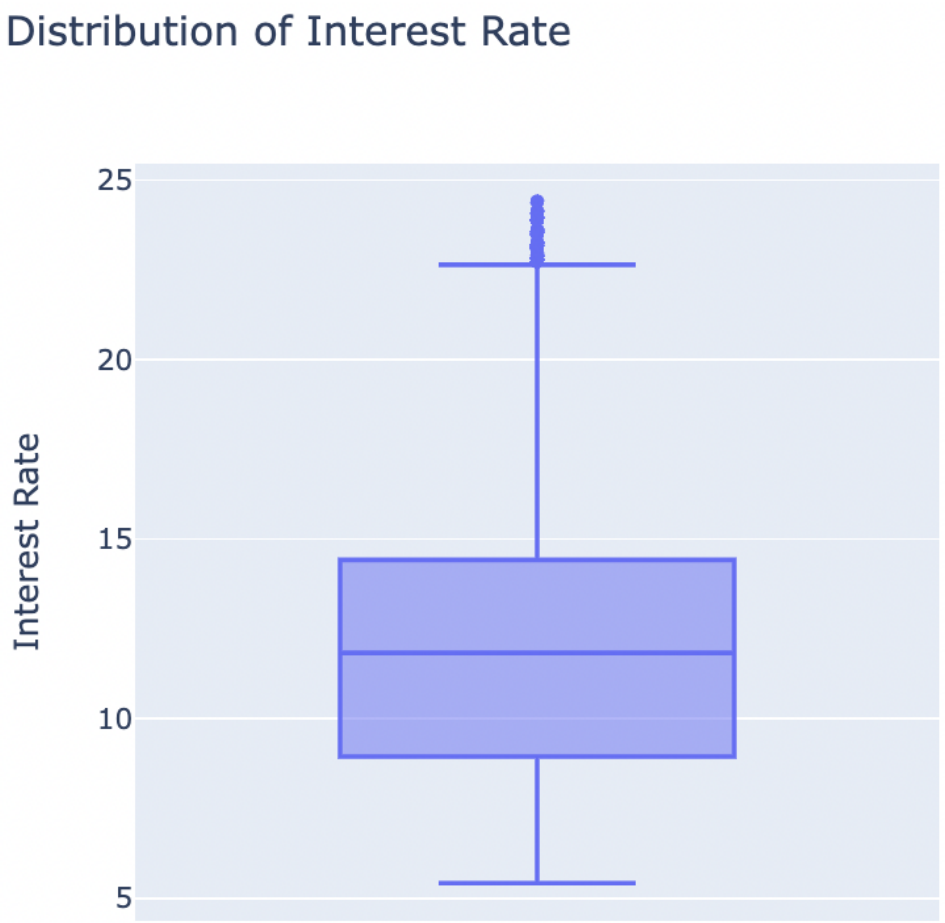
- Outlier Check Performed on Continuous Variables
  - Loan Amount
  - Interest Rate
  - Annual Income
  - DTI - Debt to Income Ratio
- Outliers Handled for Annual Income
  - Data Made Uniformly Distributed

Univariate Analysis

Univariate Analysis and Outlier Handling: DTI - Debt to Income ratio

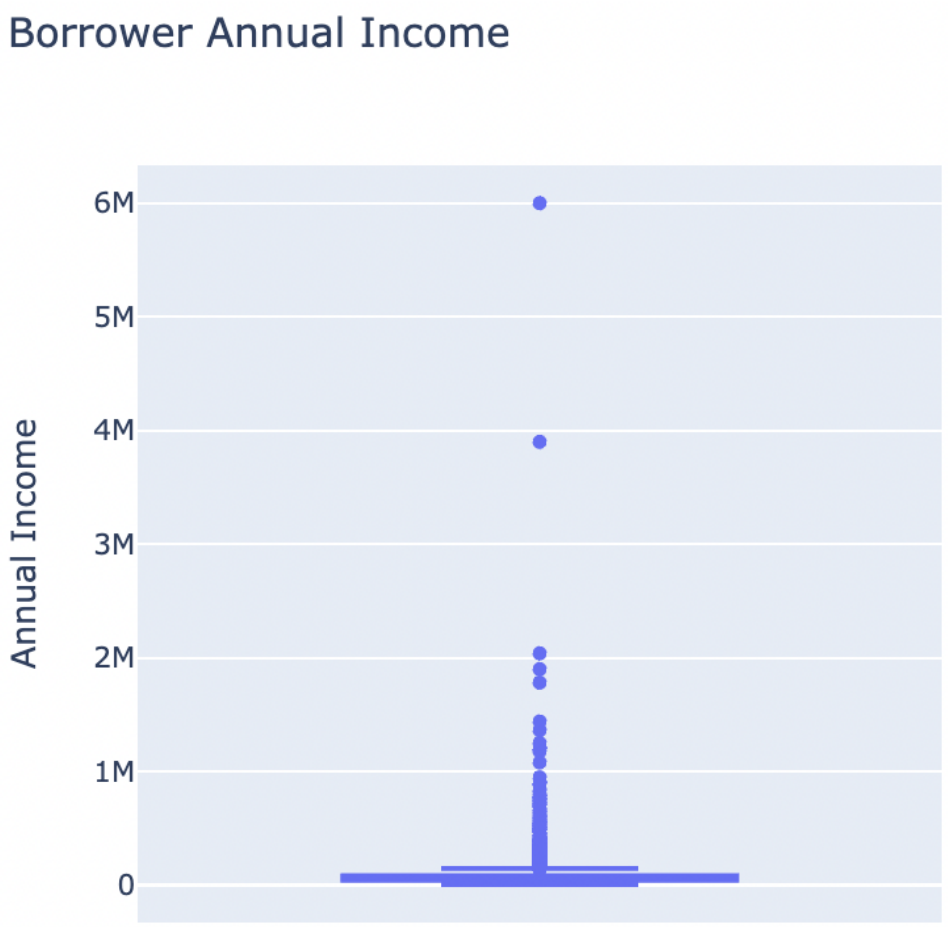
DTI - Debt to Income ratio

Univariate Analysis and Outlier Handling: Interest Rate



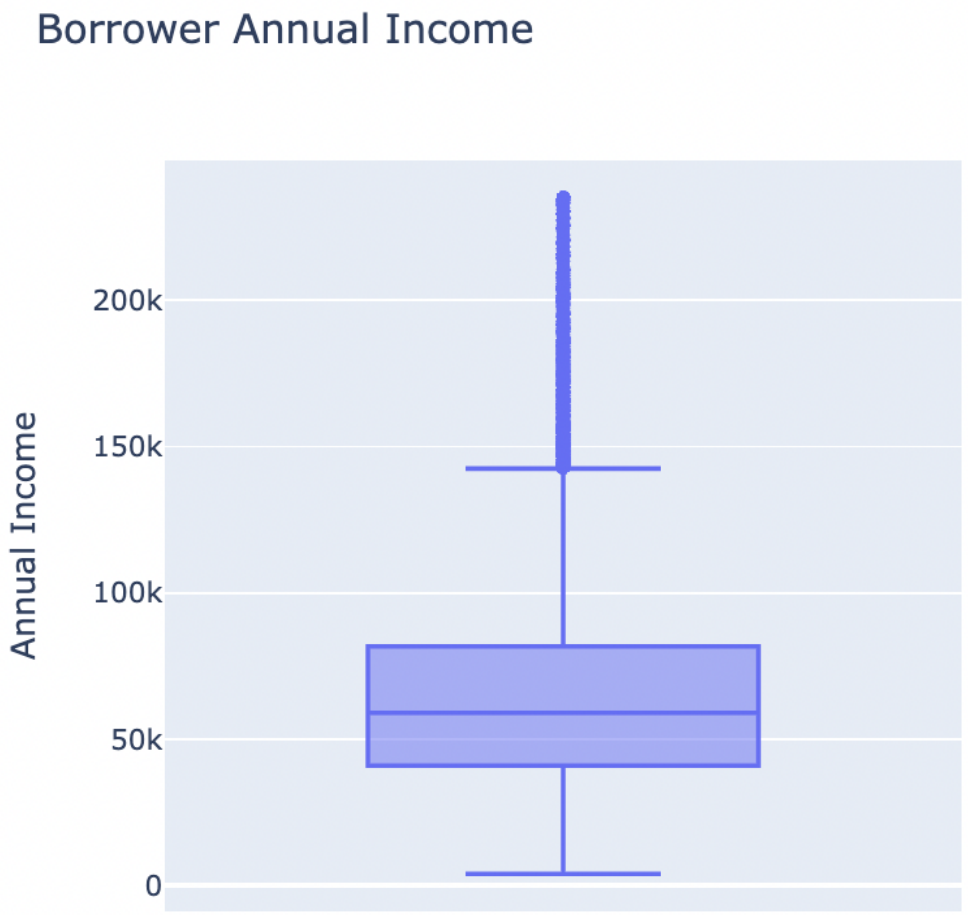
Univariate Analysis and Outlier Handling: Annual Income with Outliers in raw data

3. Annual Income with Outliers in raw data



Univariate Analysis and Outlier Handling: Annual Income with 99th percentile outlier removed

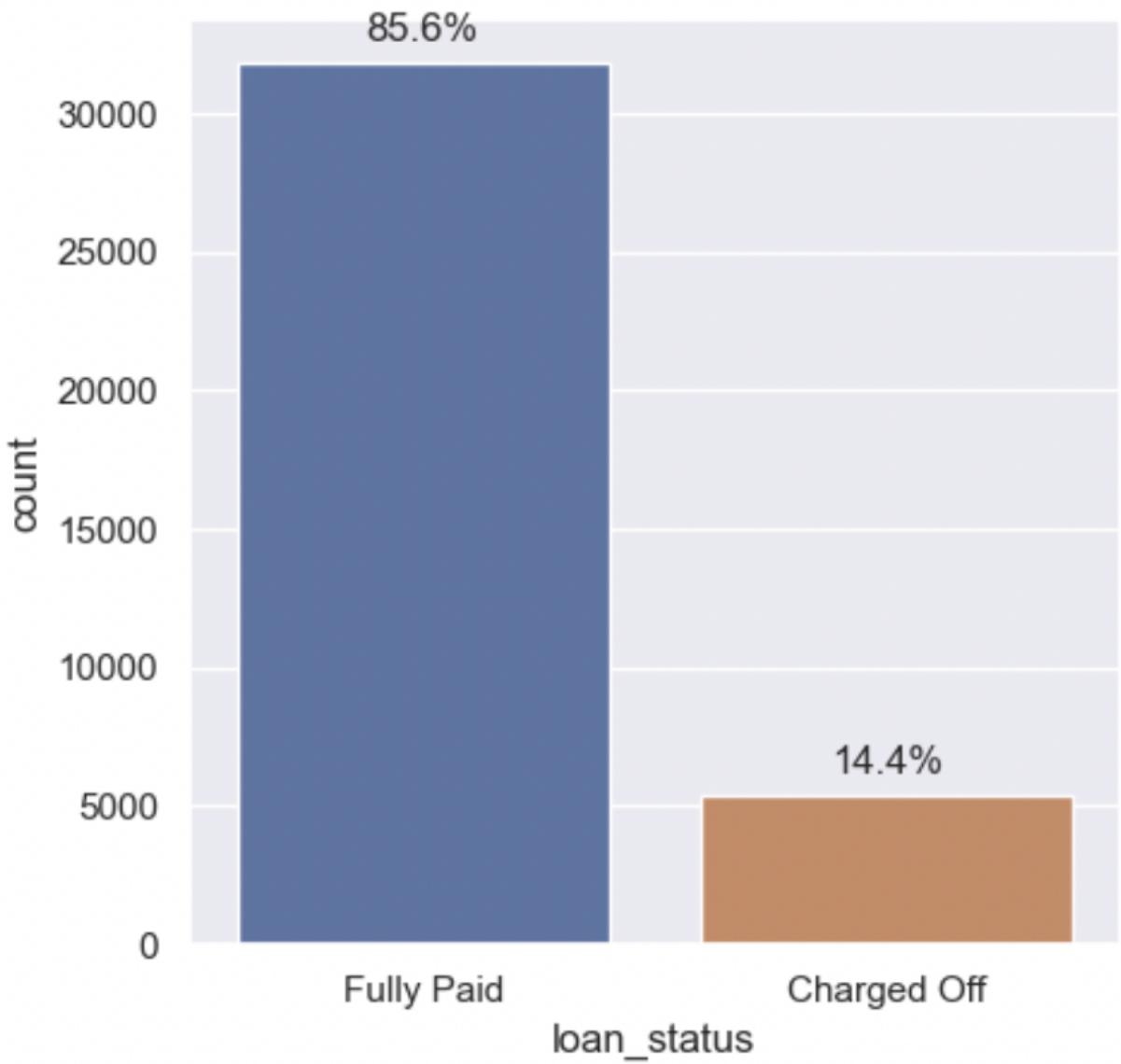
4. Annual Income with 99th percentile outlier removed



Univariate Analysis and Outlier Handling: DTI - Debt to Income Ratio

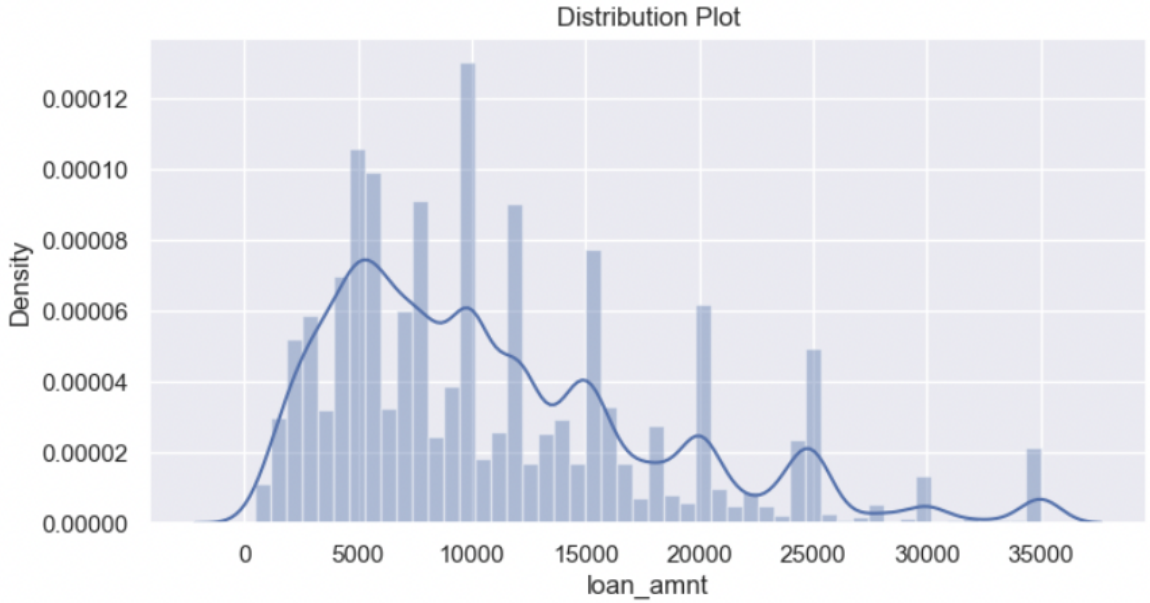
Univariate Analysis and Outlier Handling: Loan_status

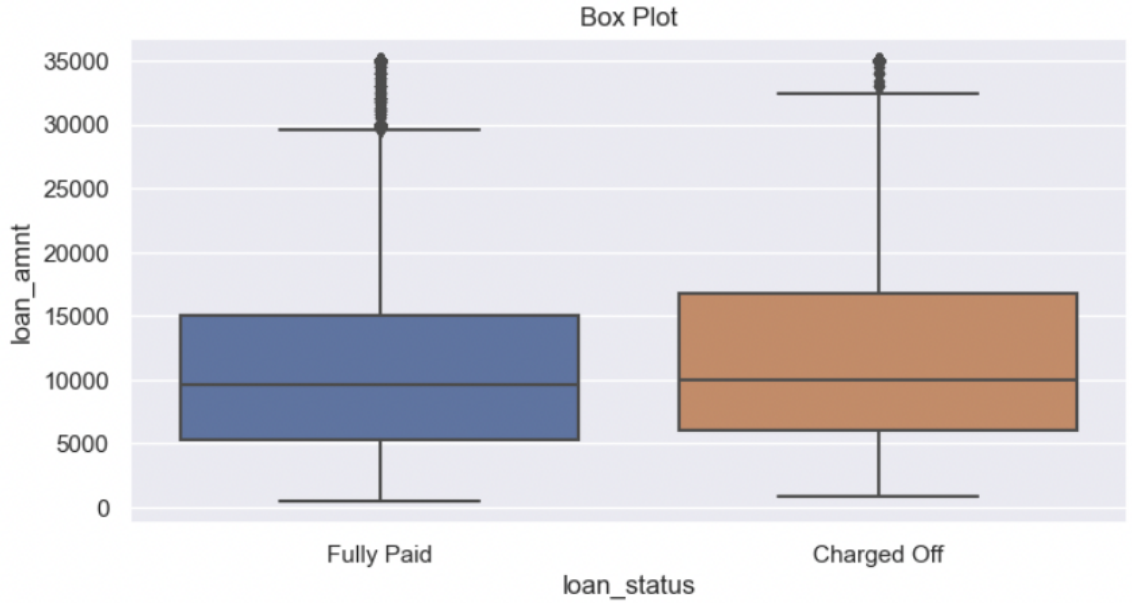
- Loan_status is the target column for analysis
  - Comparison with other columns
  - Analysis of Customer and Loan attributes



Univariate Analysis and Outlier Handling: Loan amount

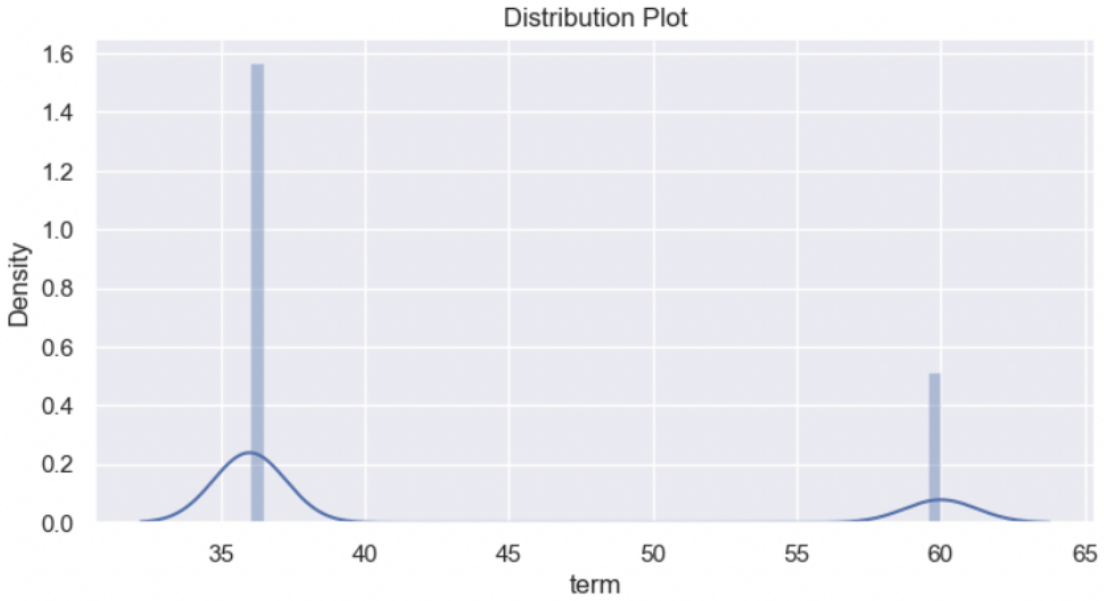
7. loan amount

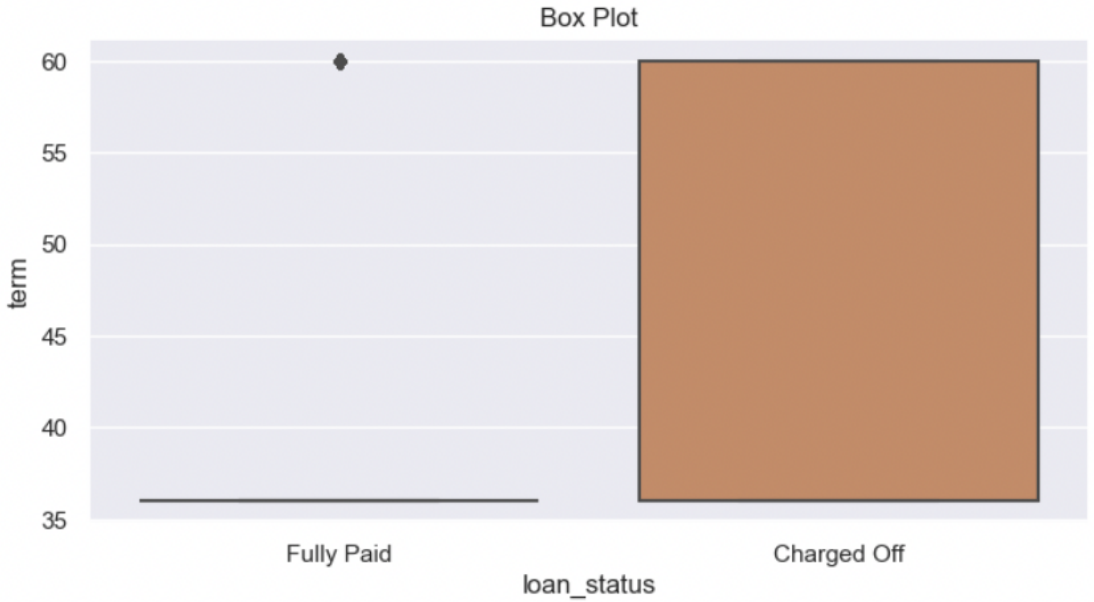




Univariate Analysis and Outlier Handling: Term

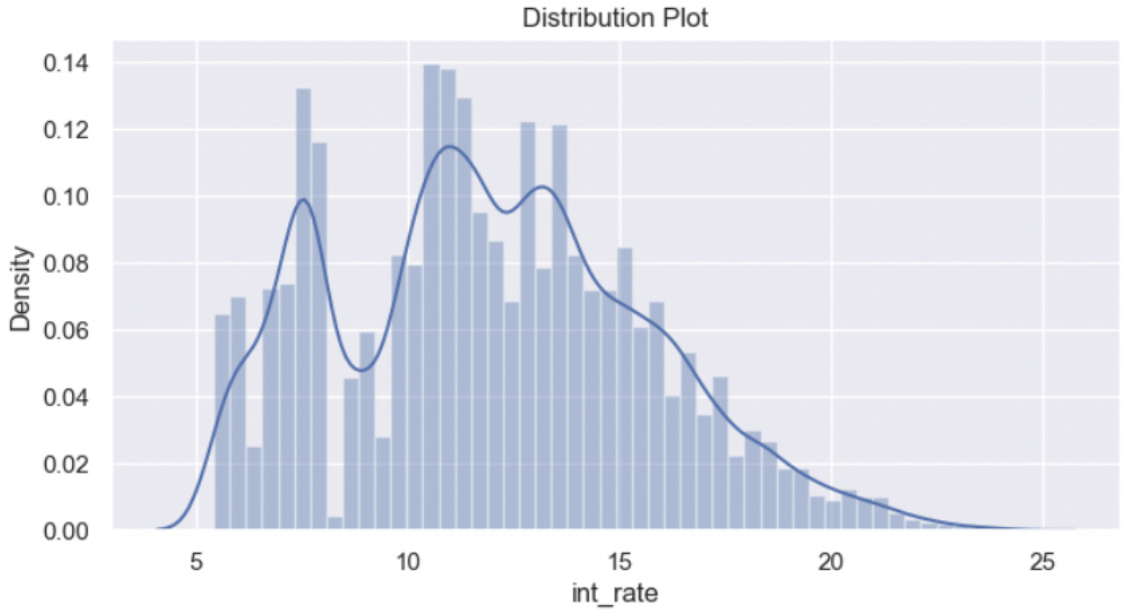
8. term

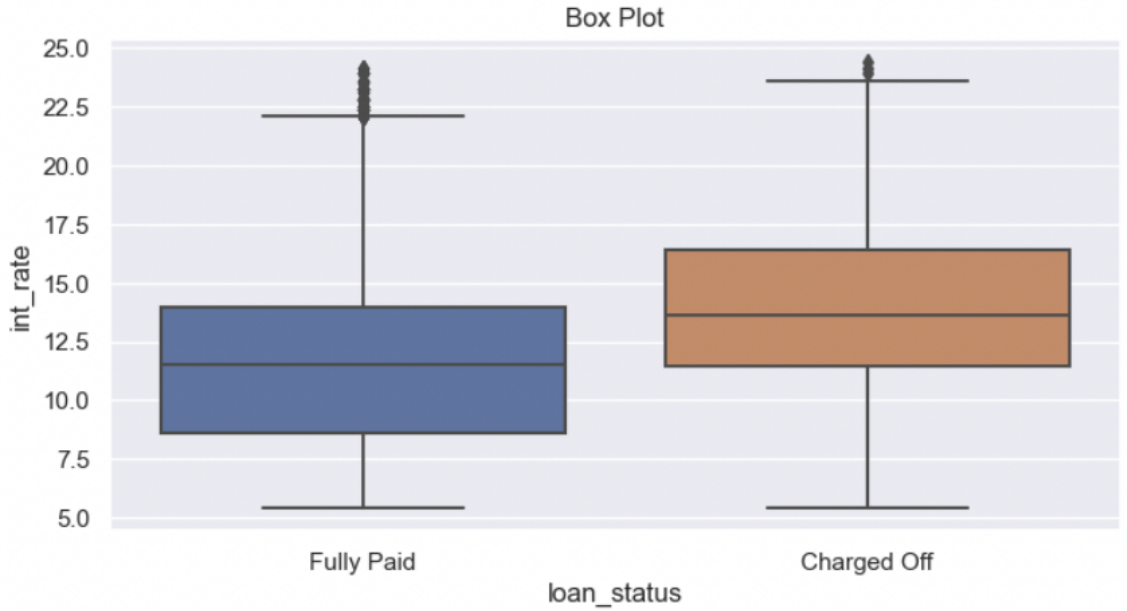




Univariate Analysis and Outlier Handling: Interest rate

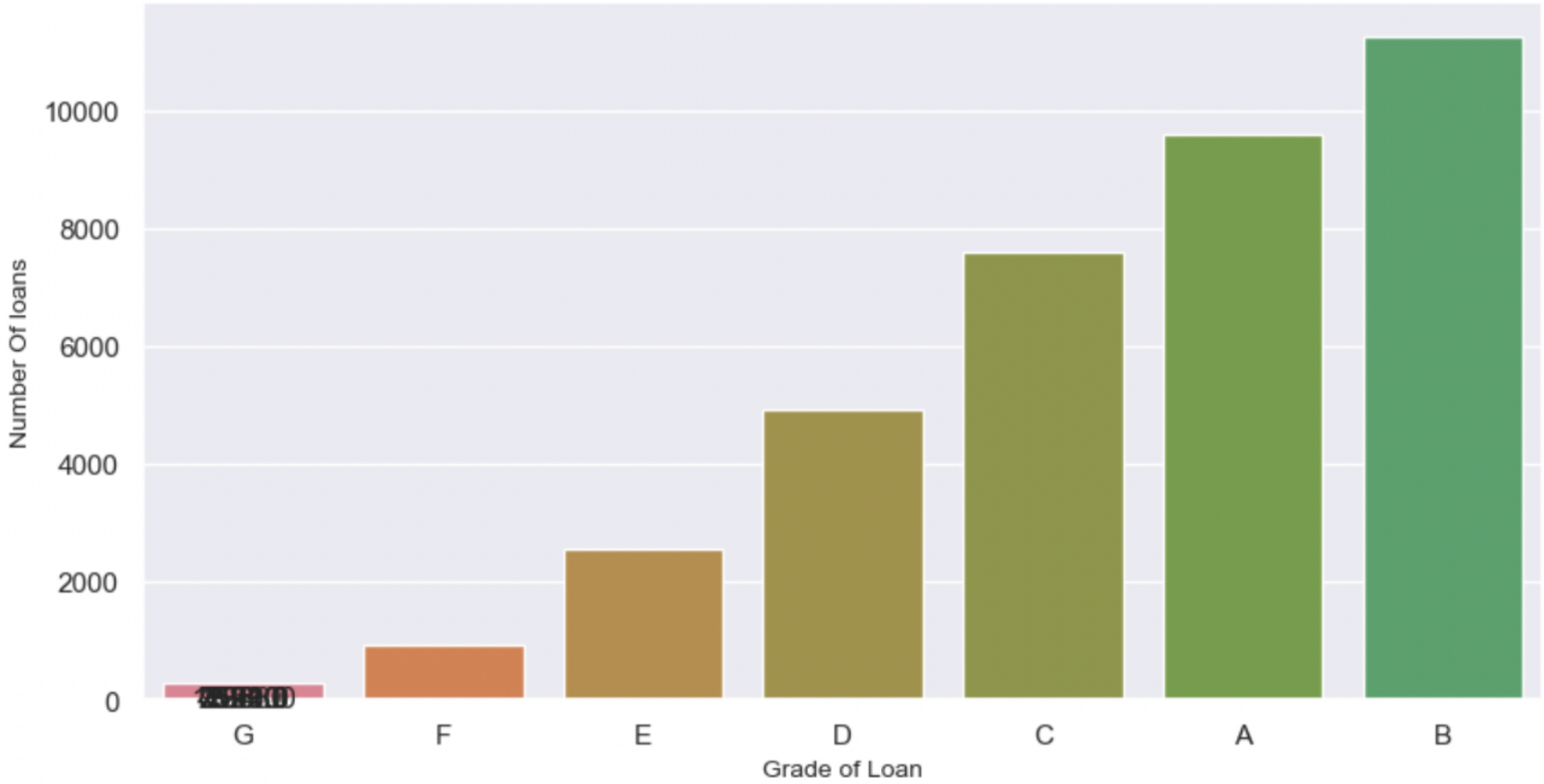
9. Interest rate





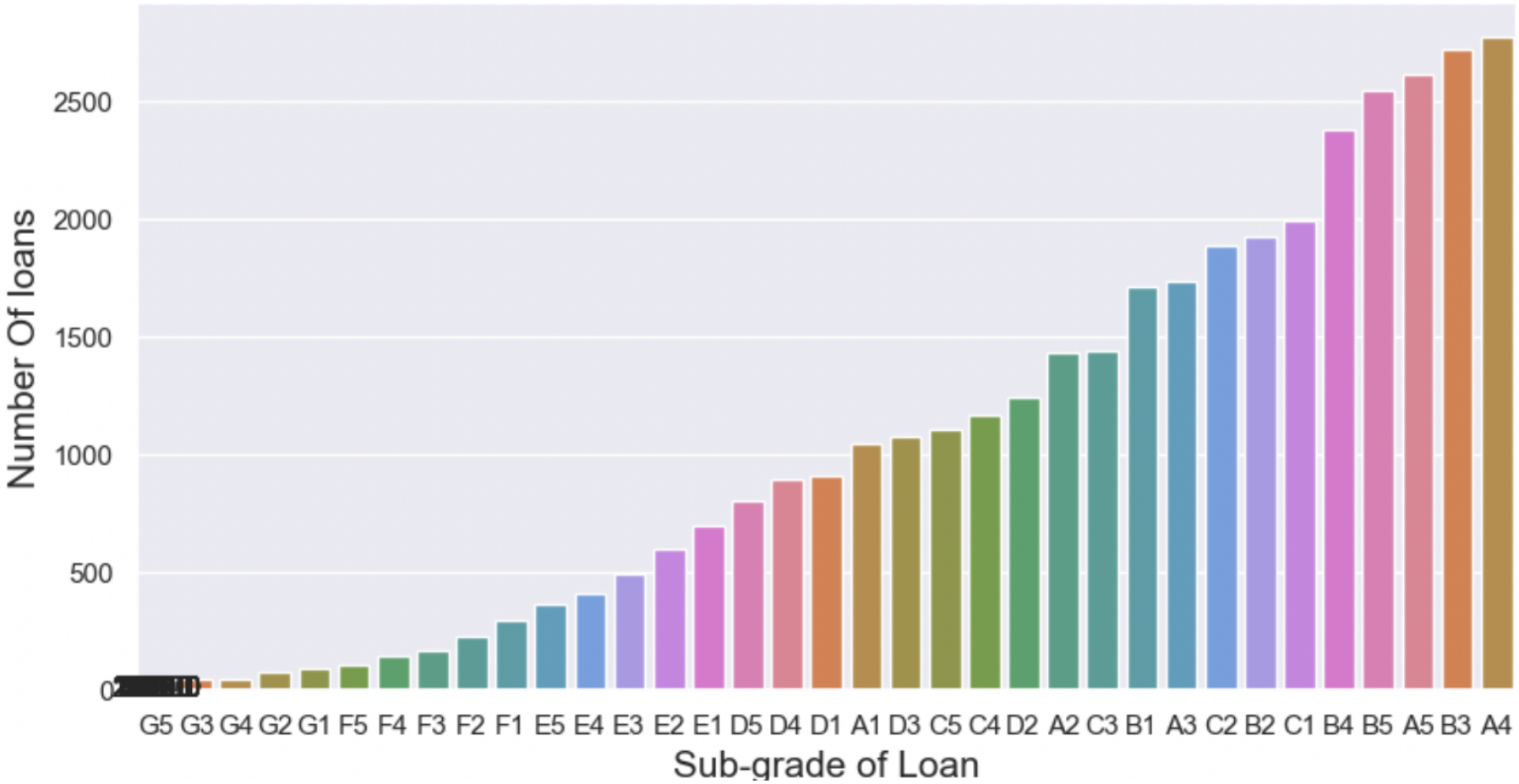
Univariate Analysis and Outlier Handling: Grade

10. Grade



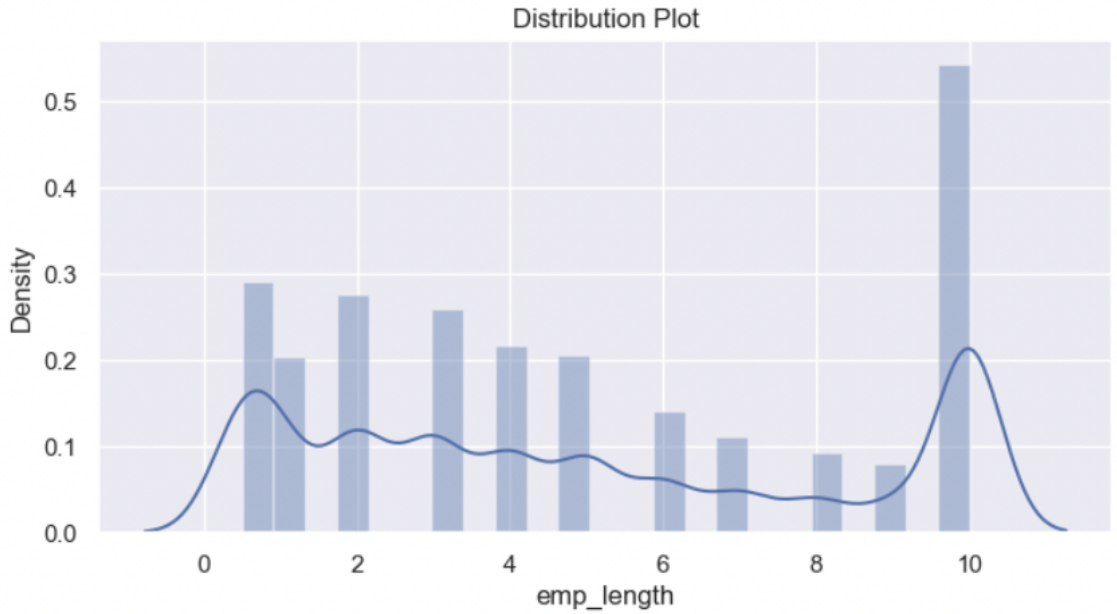
Univariate Analysis and Outlier Handling: Sub-grade

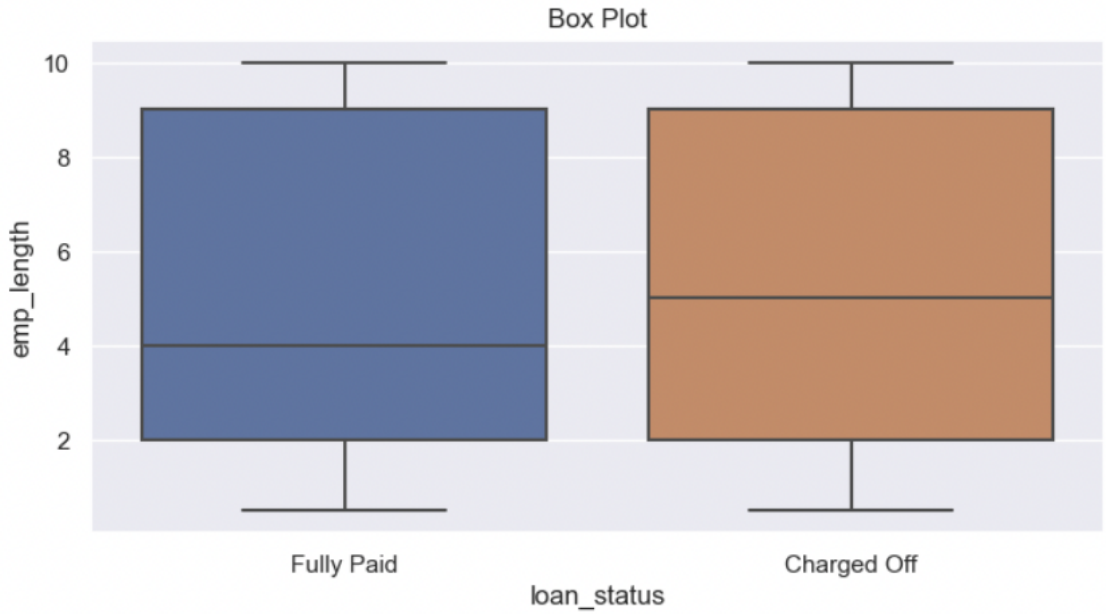
11. Sub-grade



Univariate Analysis and Outlier Handling: Employment Length

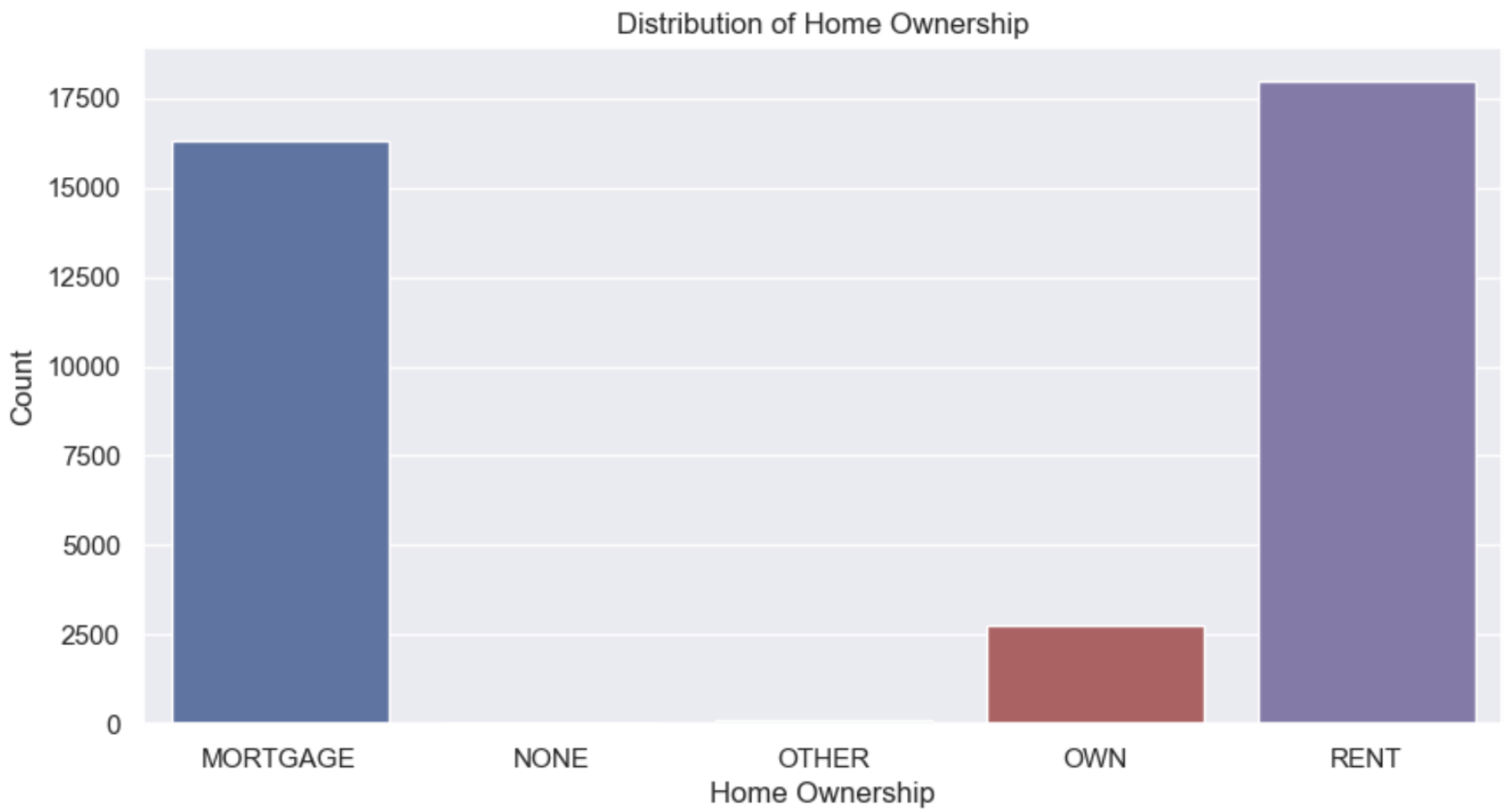
12. Employment Length





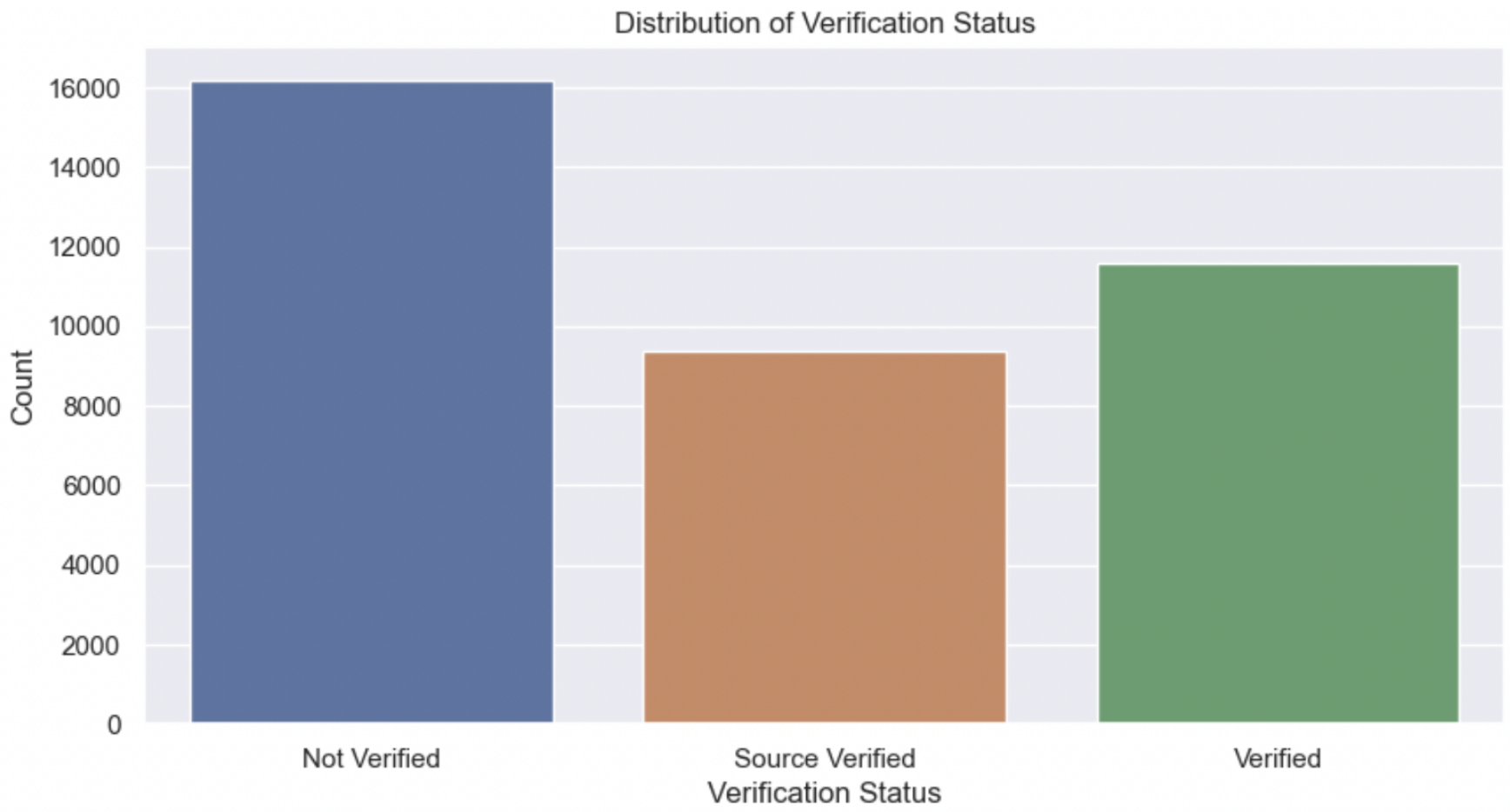
Univariate Analysis and Outlier Handling: Home ownership

13. Home ownership



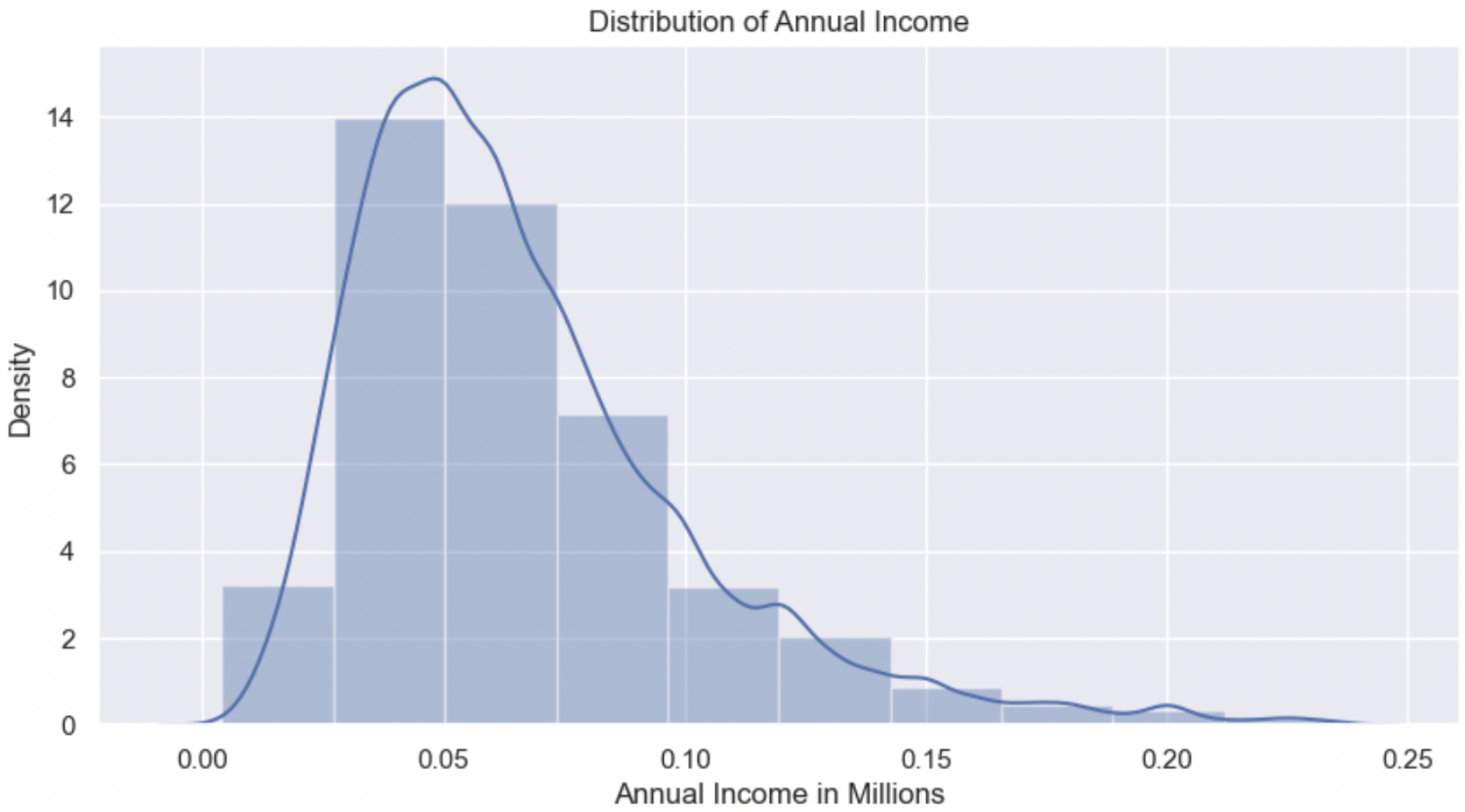
Univariate Analysis and Outlier Handling: Verification status

14. Verification status



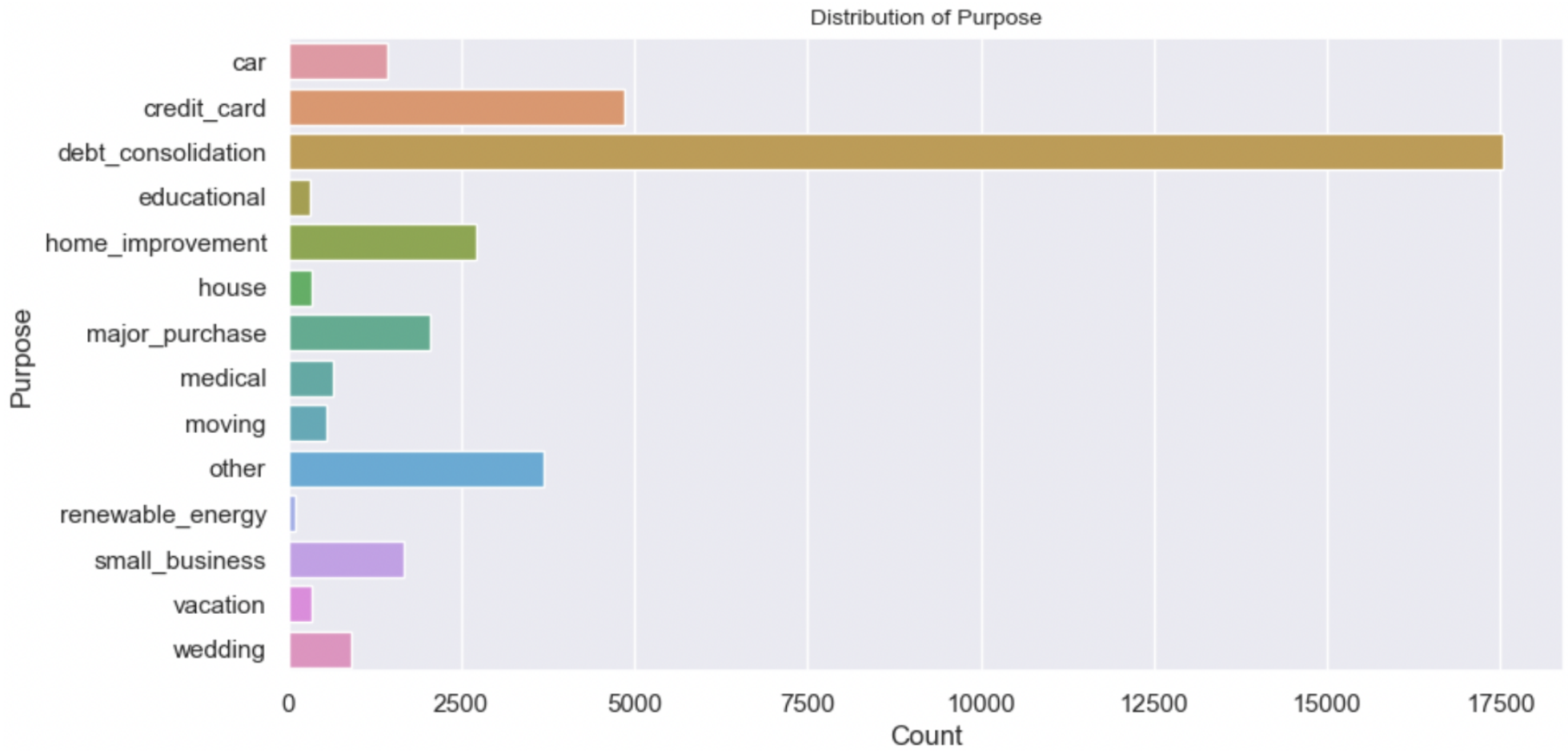
Univariate Analysis and Outlier Handling: Annual Income

15. Annual Income



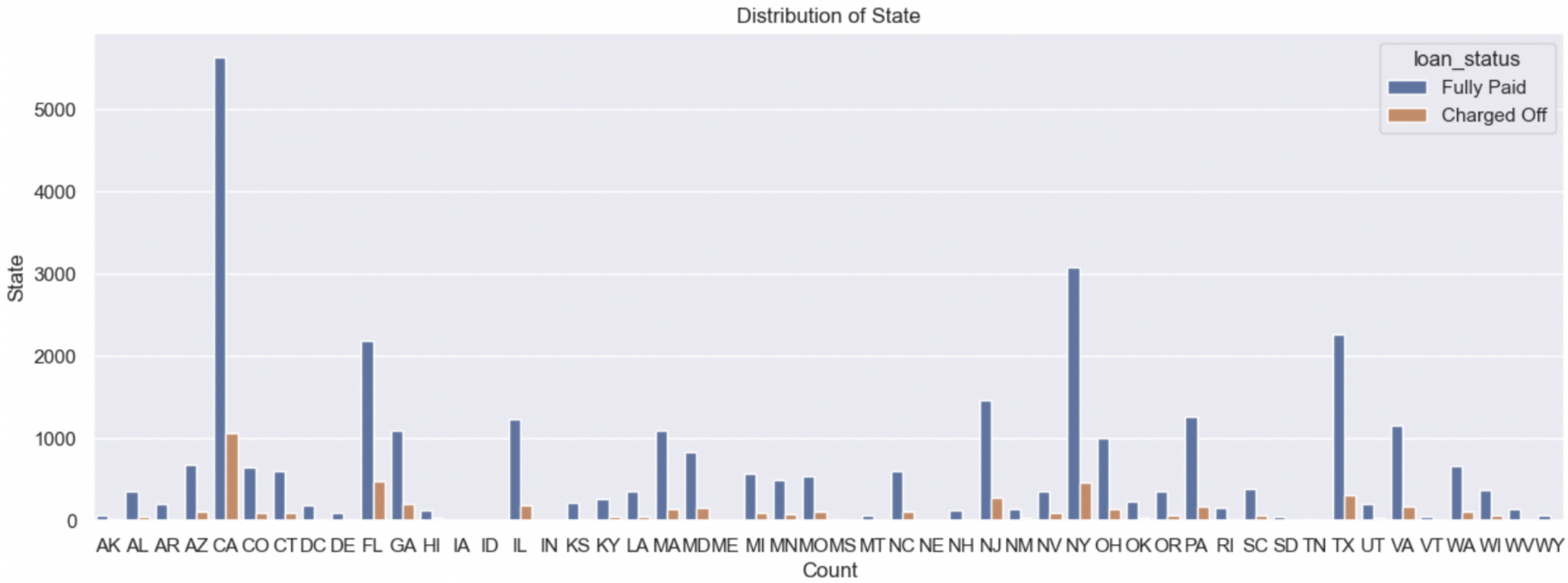
Univariate Analysis and Outlier Handling: Purpose

16. Purpose



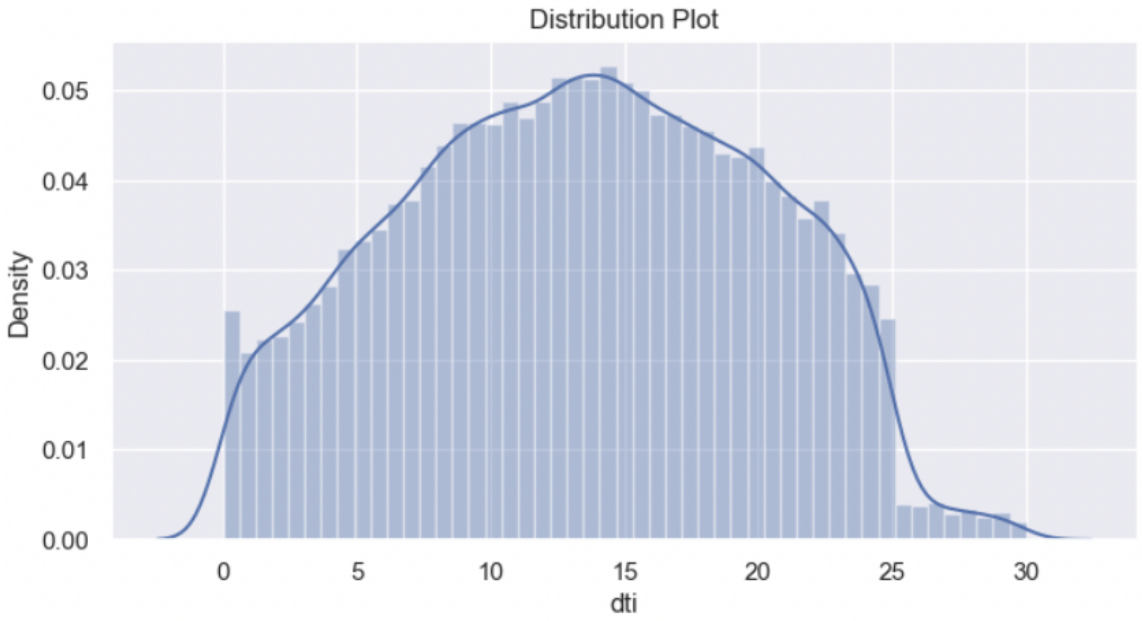
Univariate Analysis and Outlier Handling: Address State

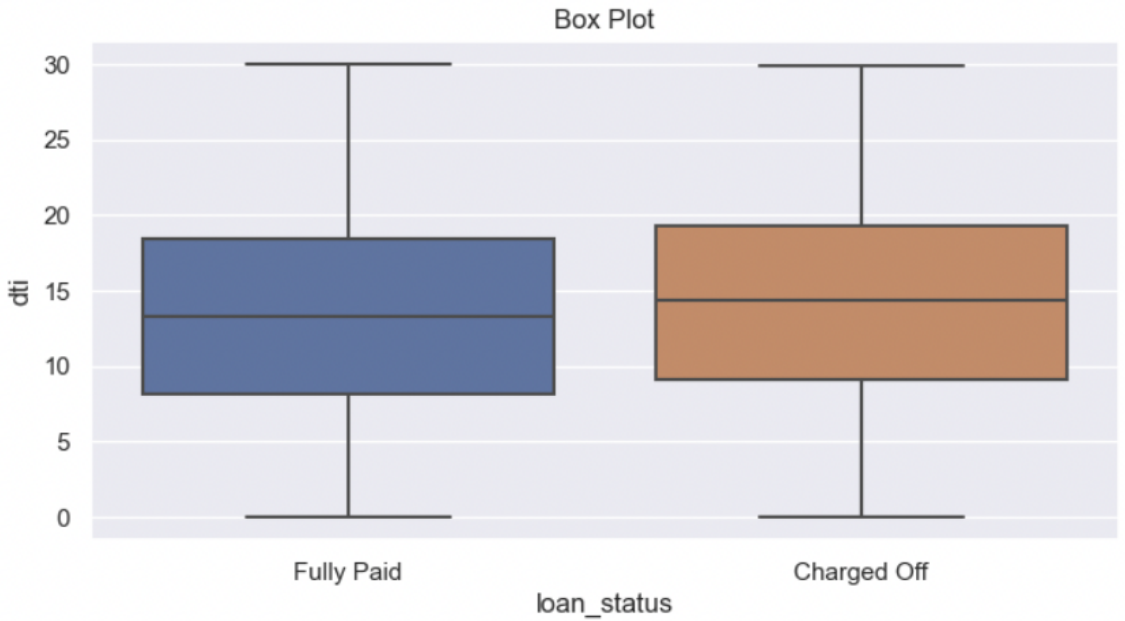
17. Address State



Univariate Analysis and Outlier Handling: DTI - Debt to Income ratio

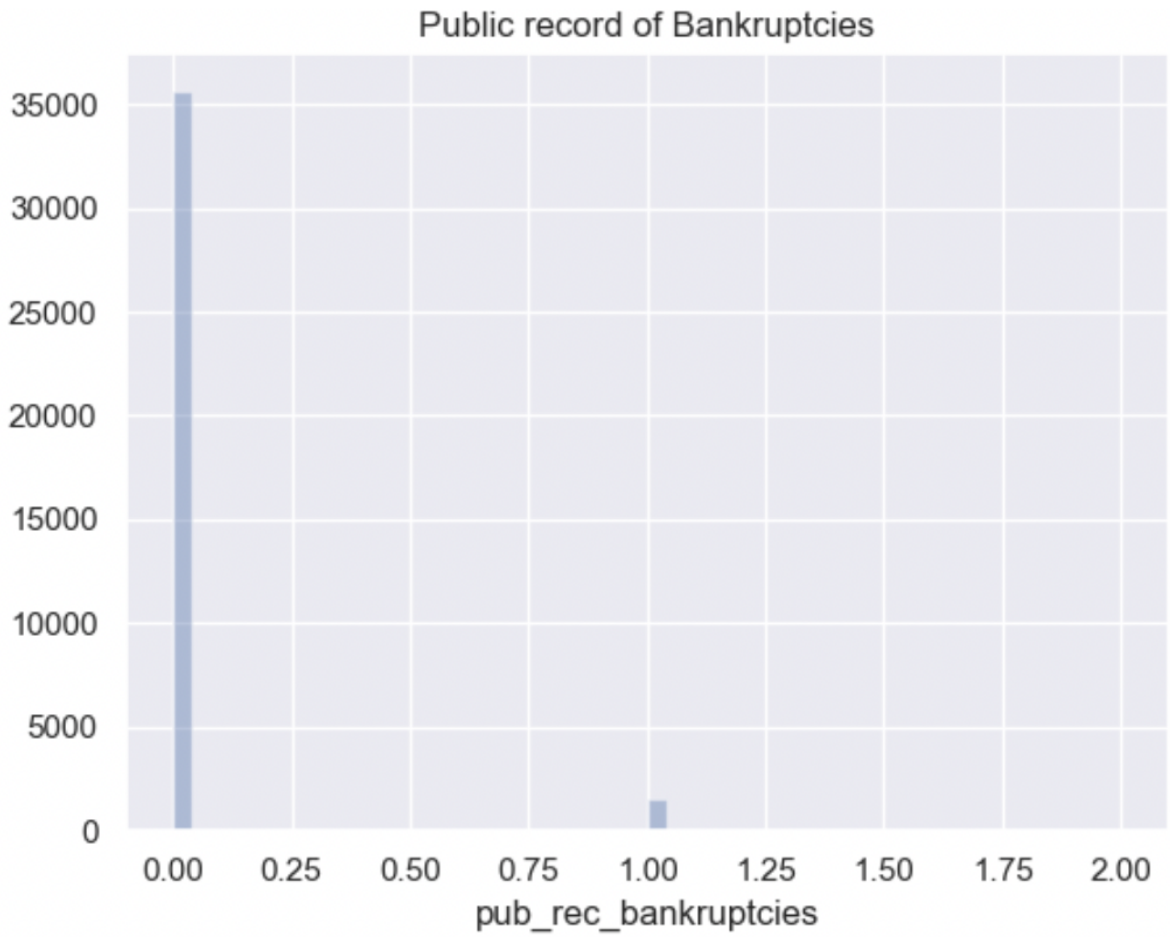
18. DTI - Debt to Income ratio





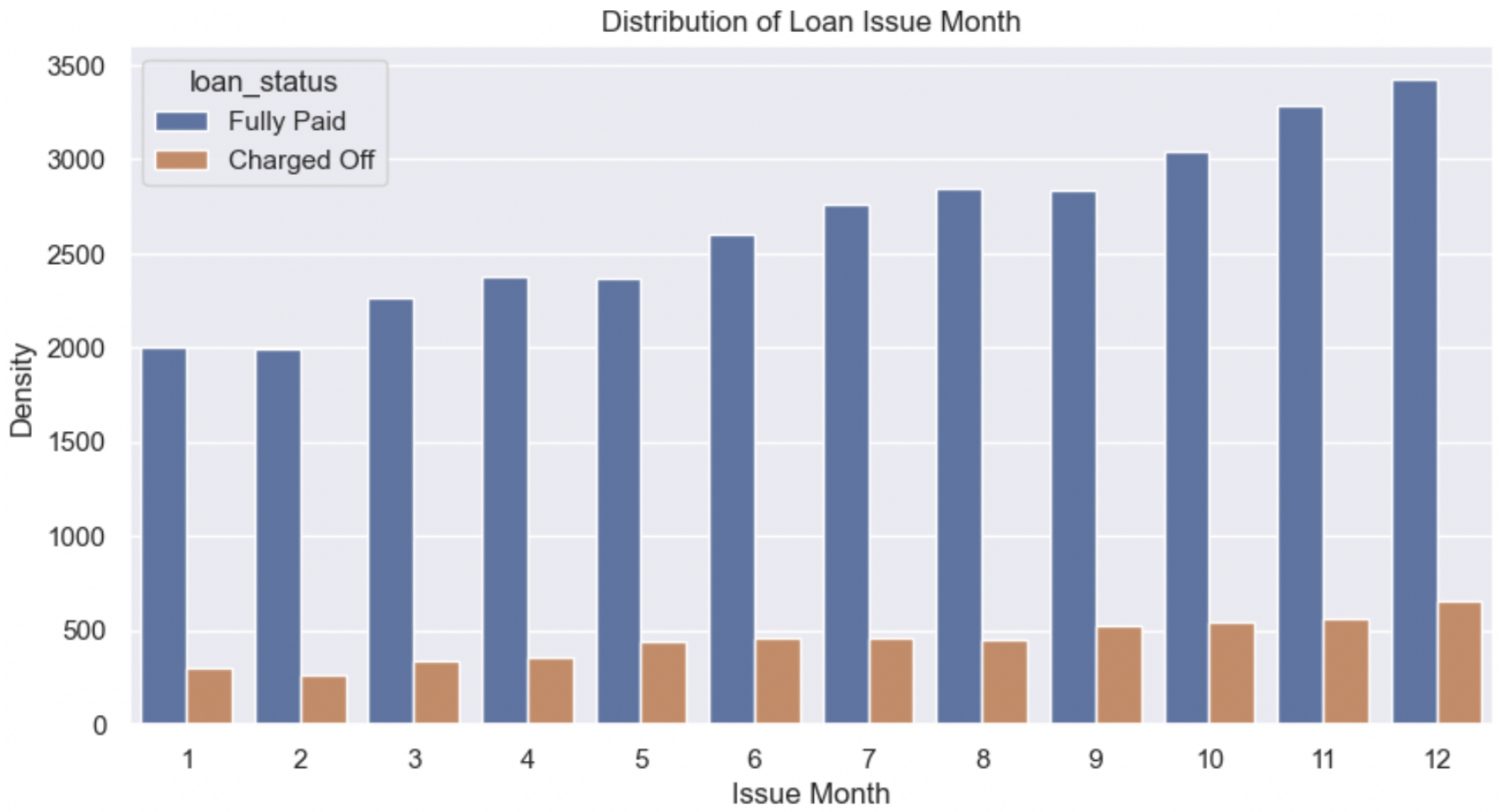
Univariate Analysis and Outlier Handling: Public record of Bankruptcy

19. Public record of Bankruptcy



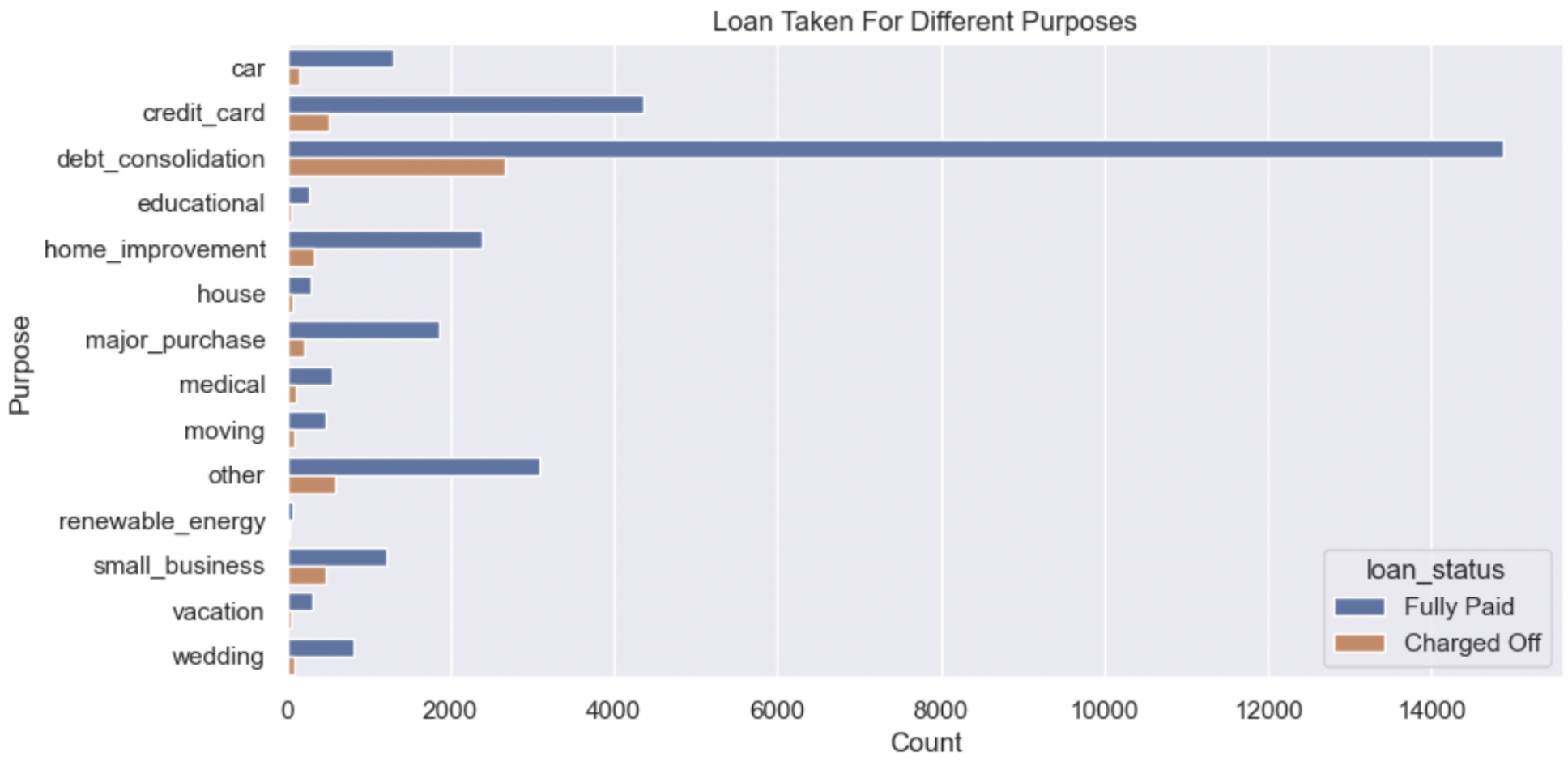
Univariate Analysis and Outlier Handling: Derived Metric - Loan Issue Month

20. Derived Metric - Loan Issue Month

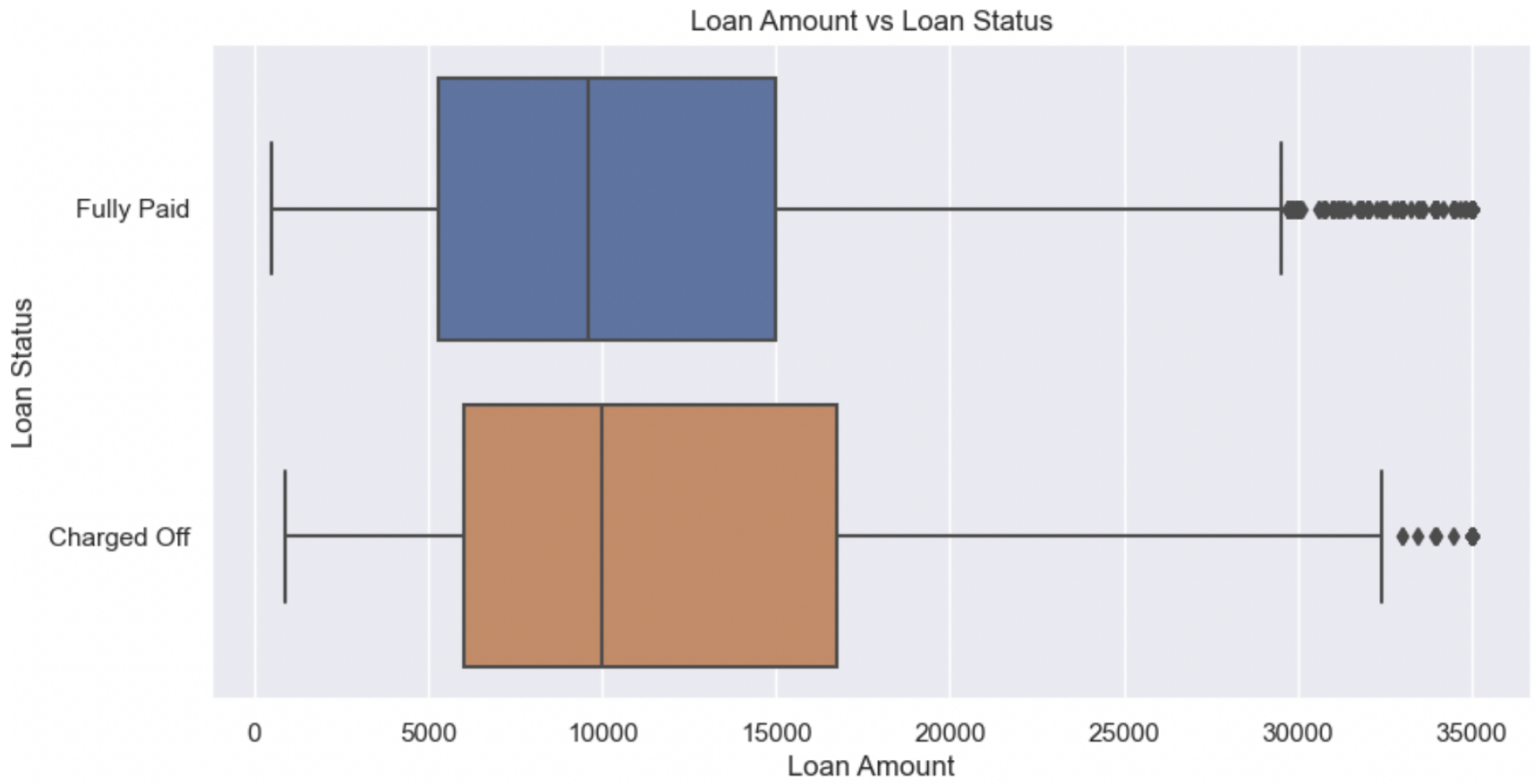


Segmented Univariate Analysis: Purpose

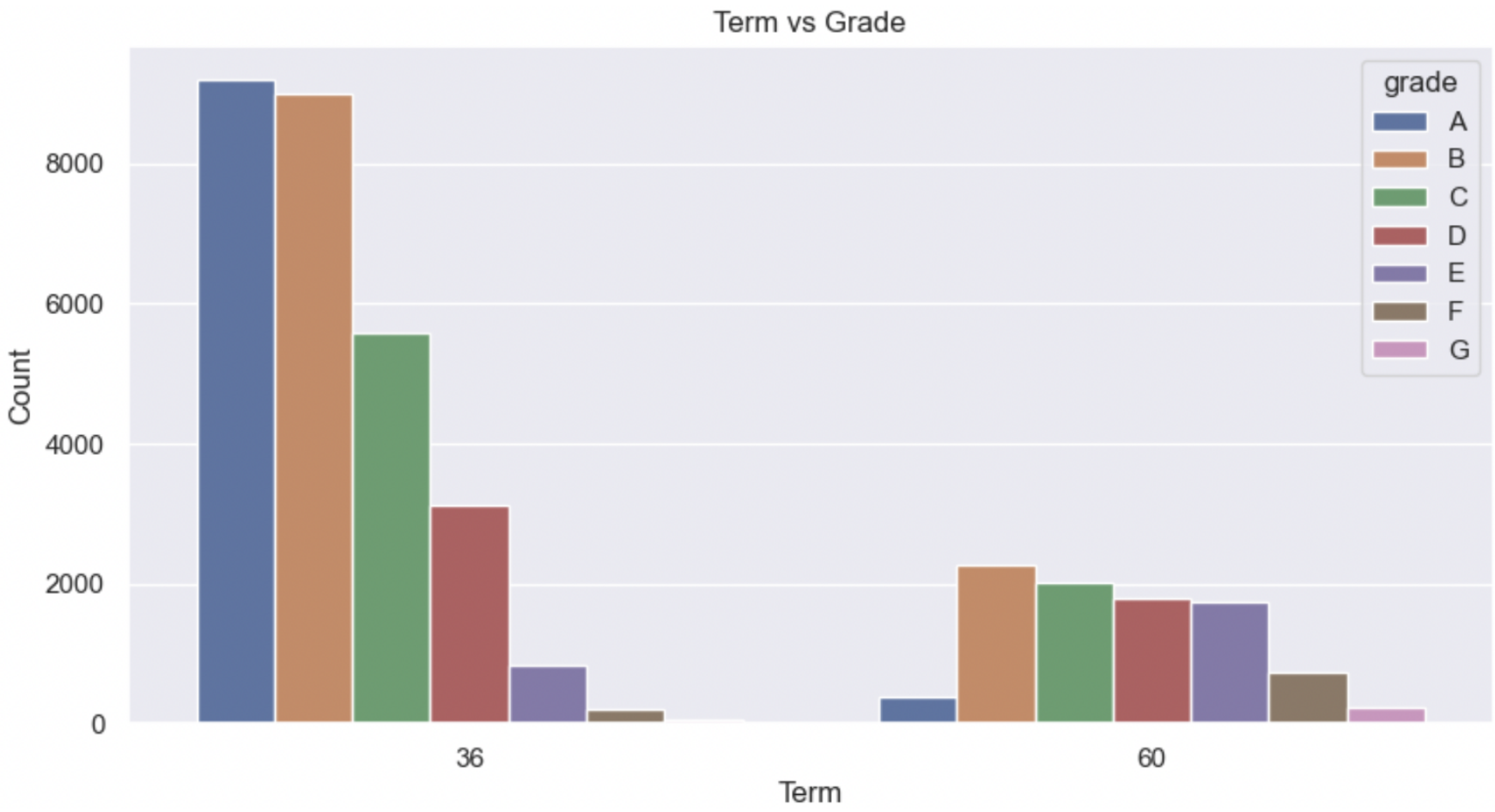
Purpose



Segmented Univariate Analysis: Loan Amount vs Loan Status

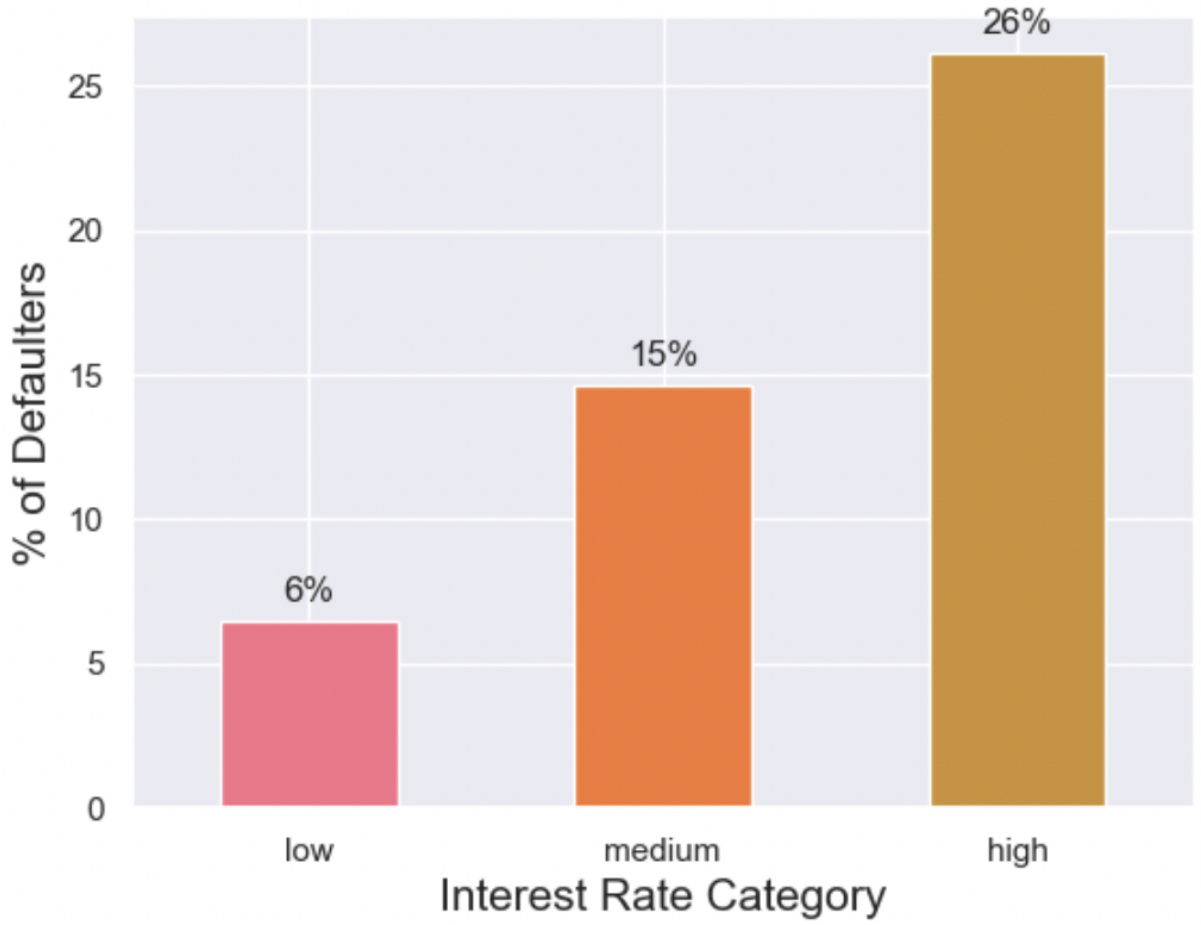


Segmented Univariate Analysis: Loan Term vs Grade



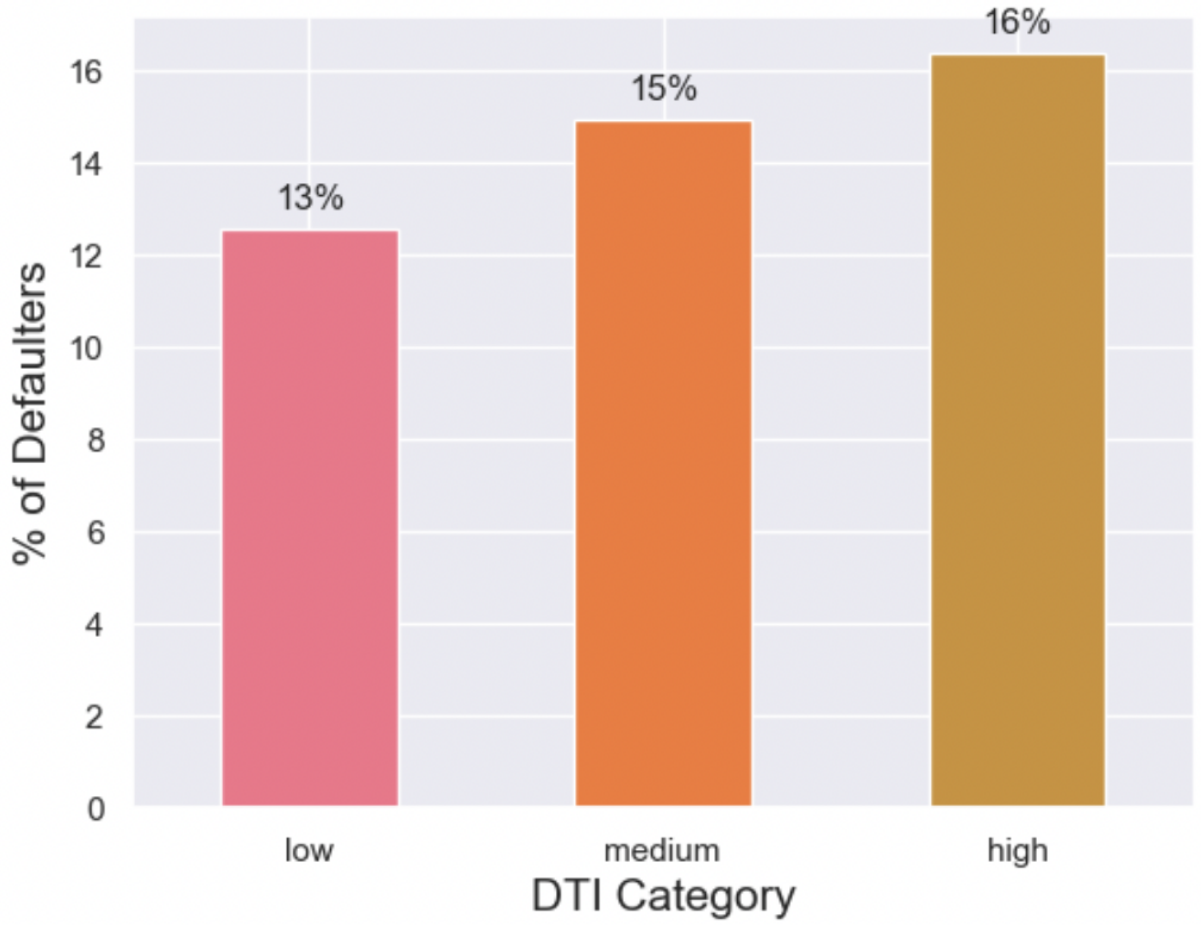
Type Driven Metrics : int_rate_bin

Type Driven Metrics :
int_rate_bin



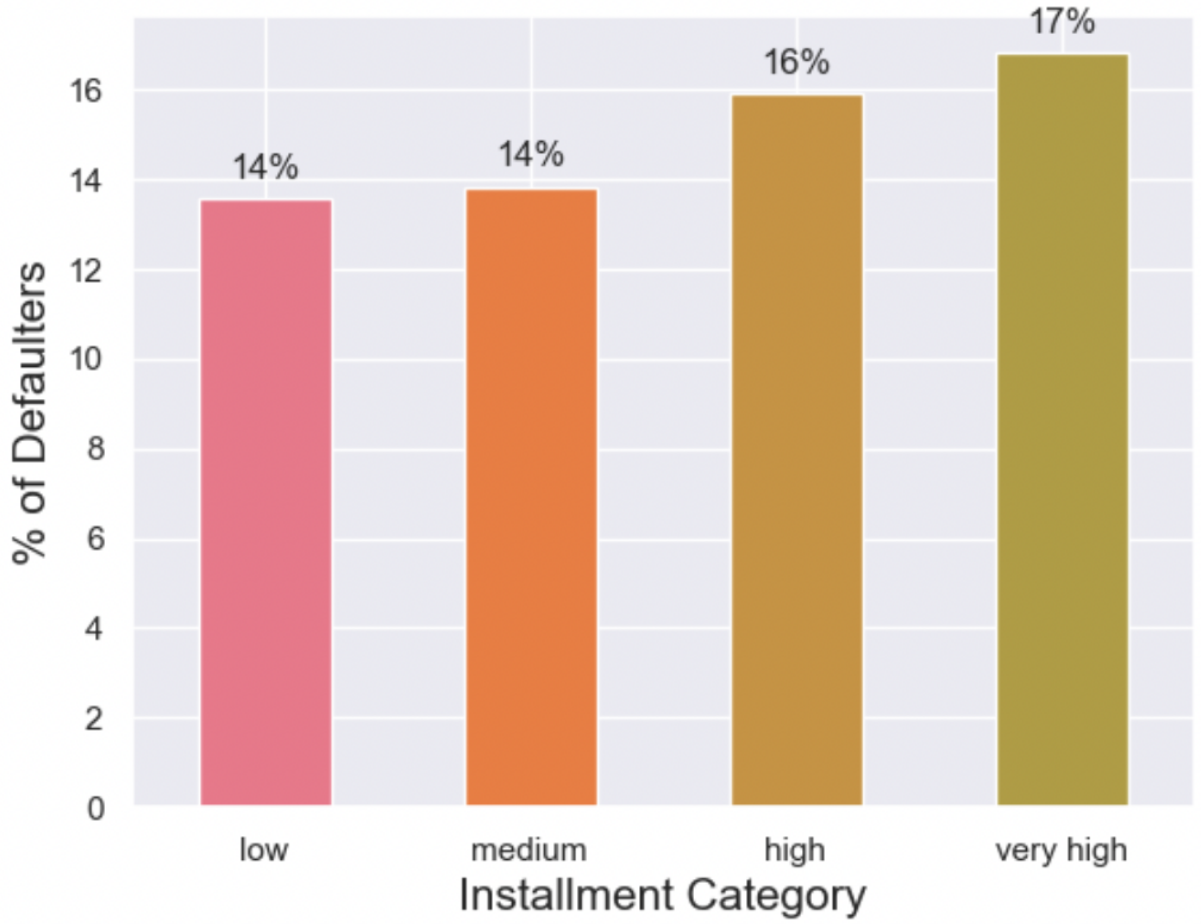
Type Driven Metrics : dti_bin

Type Driven Metrics : dti_bin



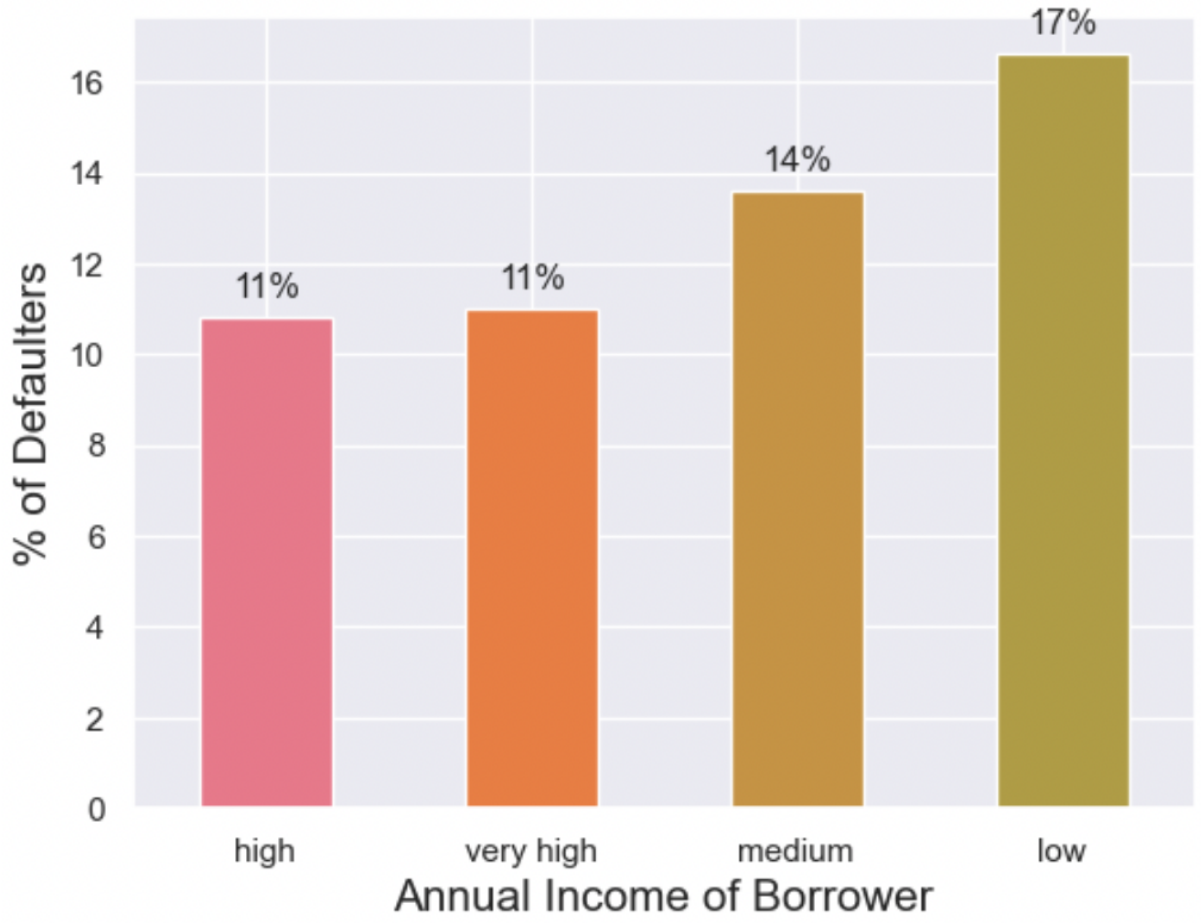
Type Driven Metrics : installment

Type Driven Metrics : installment



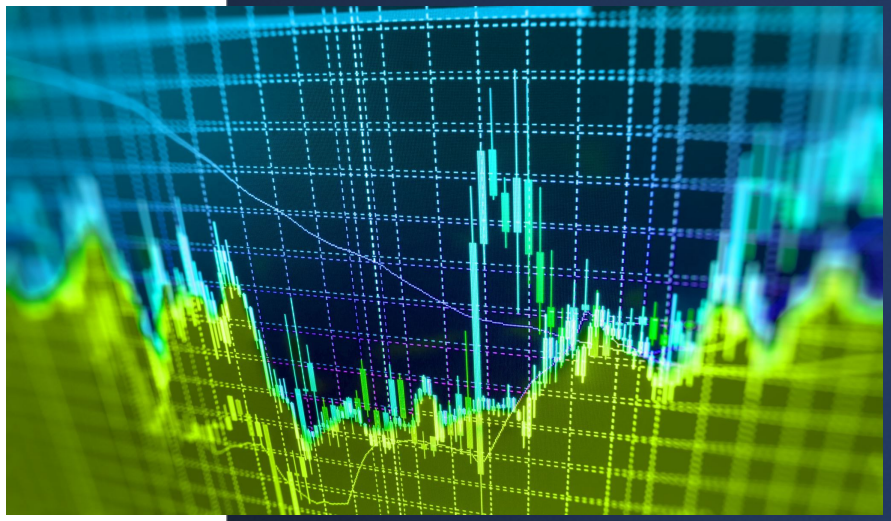
Type Driven Metrics : annual_inc

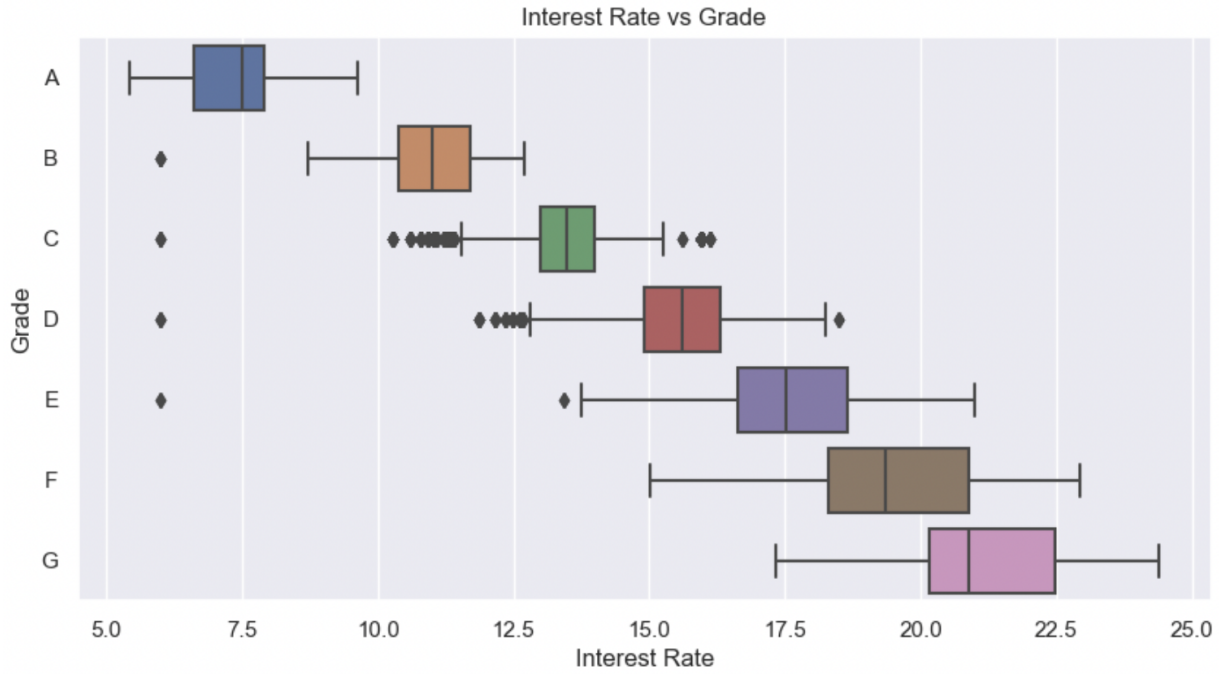
Type Driven Metrics : annual_inc



Data points of Interest

- Loan Amount (loan_amnt)
- Purpose (purpose)
- Home Ownership (home_ownership)
- Issue Date (issue_d)
- Sub-Grade (sub_grade)
- Term (term)
- Annual Income (annual_inc)
- DTI (dti)
- Public Record Bankruptcies (pub_rec_bankruptcies)
- Employment Length (emp_length)



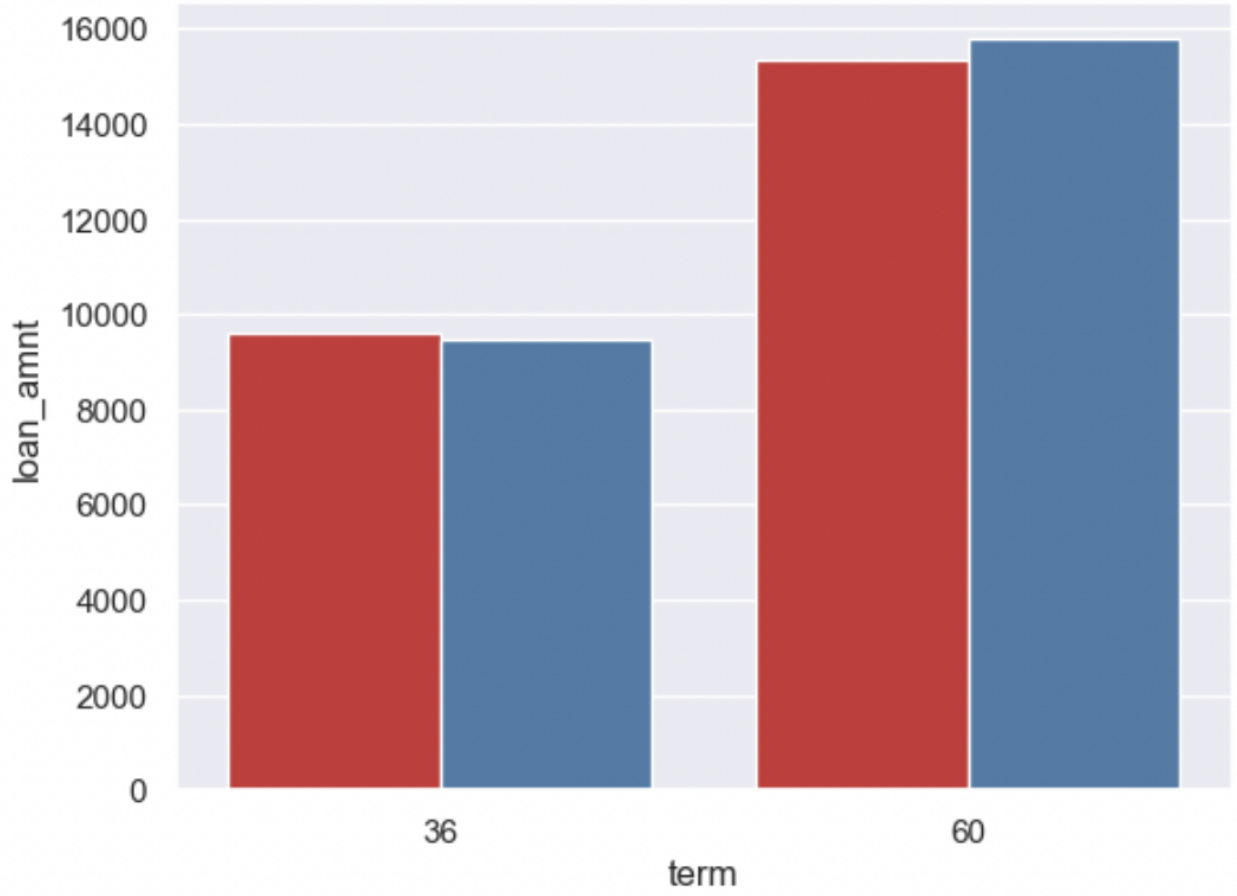


- Lower Grade = Higher Risk
  - Lower grade with high interest rate increases the risk

Bivariate Analysis

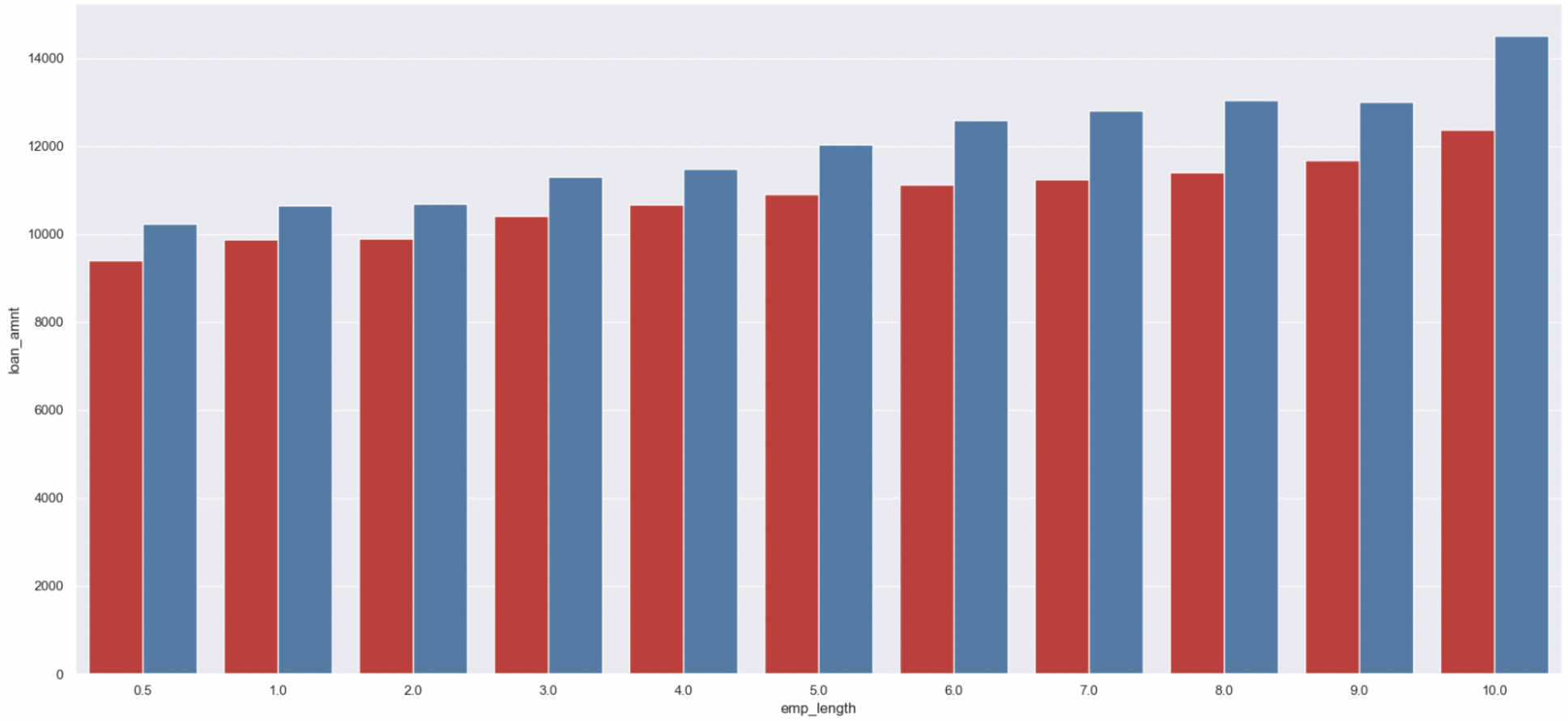
Multi-Variate Analysis: Loan Amount vs Term vs loan_status

Loan Amount vs Term vs loan_status



Multi-Variate Analysis: Loan Amount vs Employment Length vs loan_status

Loan Amount vs Employment Length vs loan_status



	loan_amnt	funded_amnt	term	int_rate	installment	grade	emp_length	annual_inc	loan_status	dti	pub_rec_bankruptc
loan_amnt	1.000000	0.981574	0.347447	0.296455	0.931098	0.280973	0.149245	0.412178	0.064965	0.070918	-0.028
funded_amnt	0.981574	1.000000	0.325587	0.300729	0.957259	0.282535	0.148890	0.407337	0.061781	0.070580	-0.029
term	0.347447	0.325587	1.000000	0.439440	0.088707	0.426705	0.105716	0.072589	0.176019	0.079515	0.019
int_rate	0.296455	0.300729	0.439440	1.000000	0.272861	0.947899	-0.000144	0.063253	0.214626	0.114593	0.085
installment	0.931098	0.957259	0.088707	0.272861	1.000000	0.258188	0.121445	0.408688	0.031707	0.060419	-0.026
grade	0.280973	0.282535	0.426705	0.947899	0.258188	1.000000	-0.000507	0.066128	0.204583	0.100309	0.078
emp_length	0.149245	0.148890	0.105716	-0.000144	0.121445	-0.000507	1.000000	0.172843	0.016942	0.052581	0.063
annual_inc	0.412178	0.407337	0.072589	0.063253	0.408688	0.066128	0.172843	1.000000	-0.059898	-0.112099	-0.010
loan_status	0.064965	0.061781	0.176019	0.214626	0.031707	0.204583	0.016942	-0.059898	1.000000	0.041832	0.044
dti	0.070918	0.070580	0.079515	0.114593	0.060419	0.100309	0.052581	-0.112099	0.041832	1.000000	0.007
pub_rec_bankruptcies	-0.028905	-0.029832	0.019322	0.085265	-0.026723	0.078934	0.063764	-0.010598	0.044683	0.007245	1.000

- High correlation is observed between:
  - funded_amnt and loan_amnt
  - Loan_amnt and installment
  - Grade and interest rate

Correlation HeatMap

Analysis Results

Analysis Results: Univariate Analysis

1. Loan amounts range from 500 to 35000, with an average of 9800.
2. 75% of loans have a term of 36 months.
3. Most loans have an interest rate of 5-10 and 10-15.
4. Grade B loans are the most common.
5. Borrowers with 10+ years of employment are most prevalent.
6. Most borrowers live in rented or mortgaged accommodations.
7. Borrowers' income sources are mostly verified.
8. Most borrowers have low annual income.
9. Debt consolidation is the most common reason for borrowing.
10. The majority of borrowers are from California, New York, Texas, and Florida.
11. Most borrowers have a high debt-to-income ratio.
12. 85.5% of borrowers have no public bankruptcy records.
13. The last quarter of the year sees the highest loan disbursement.
14. Debt consolidation and credit card loans have the highest fully paid and defaulted rates.
15. Loan amounts between 10K and 17K have a higher default rate.
16. Shorter term loans are more common for higher grade loans

Analysis Results: Segmented Univariate Analysis

- Higher interest rates lead to more defaults.
- Borrowers with high debt-to-income ratio (>20) are more likely to default.
- High installment payments are associated with higher default rates.
- Borrowers with annual income less than 50K are more likely to default.
- Lower grade loans with high interest rates are riskier.
- Higher loan amounts and longer terms have more defaults and fully paid loans.
- Borrowers who own their property default less than those on rent or mortgage.
- Borrowers with low annual income are more likely to default.
- Larger loan amounts have a higher chance of defaulting.
- Debt consolidation is the most common purpose and has the highest number of defaults.
- Fully paid loans are increasing exponentially compared to defaulted loans.
- Default loan amount increases with interest rate and declines after 17.5%.
- Borrowers with more than 10 years of employment take larger loans.
- Most defaulters with large loans have more than 10 years of employment.

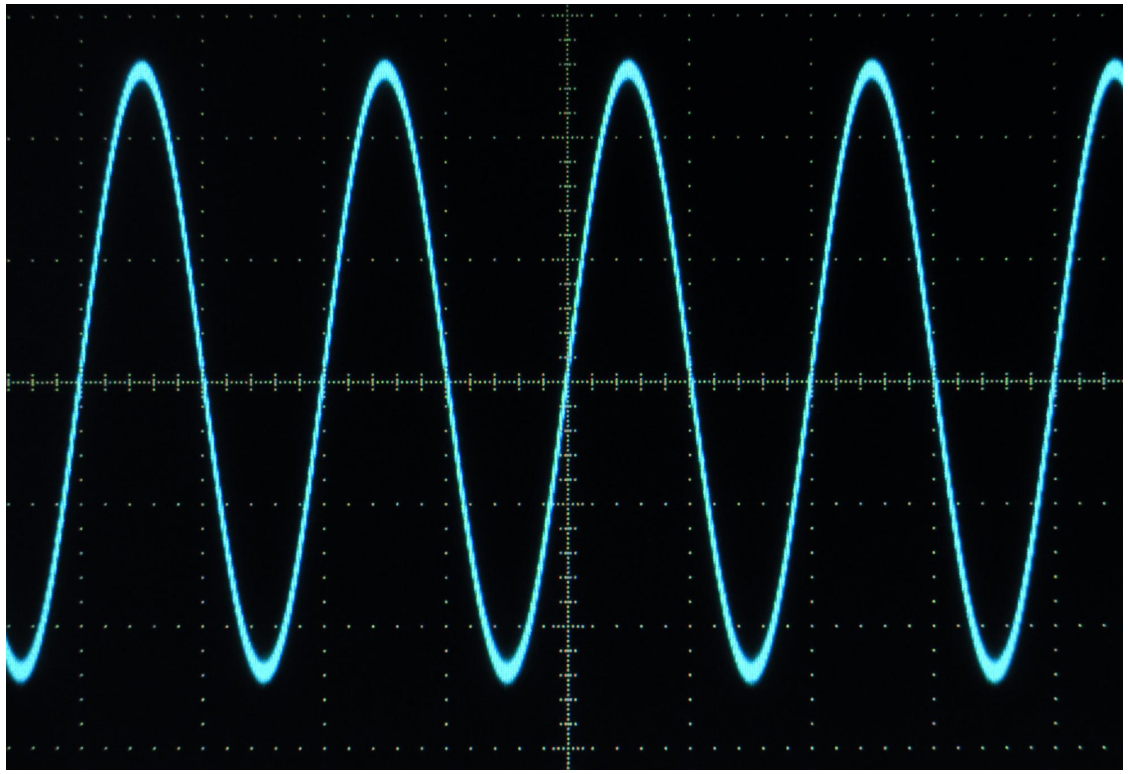
Analysis Results: Bivariate Analysis

- The Grade represents the risk factor. Lower grade high risk of default.
- The borrowers with no record of Public Recorded Bankruptcy can be a safe choice for lending.

Correlation

High correlation is observed between

- funded_amnt and loan_amnt
- Loan_amnt and installment
- Grade and interest rate



Suggestions and Recommendations

Suggestions & Recommendations

Data points that can be used to predict whether there will be a default and avoiding Credit Loss:

- DTI
- Grades
- Verification Status
- Annual income
- Pub_rec_bankruptcies

With respect to borrowers following considerations can be considered as risk for 'defaults' :

- Should not be from large urban cities like California, New York, Texas, Florida etc.
- Not having annual income in the range 50000-100000.
- Not having Public Recorded Bankruptcy.
- Have least grades like E,F,G which indicates high risk.
- Have very high Debt to Income value.
- Have working experience of 10+ years.