Lending Club Case Study

Agrim Koundal
Ajith Shenoy

Objective:

- The goal is to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA using the given dataset, is the aim of this case study.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

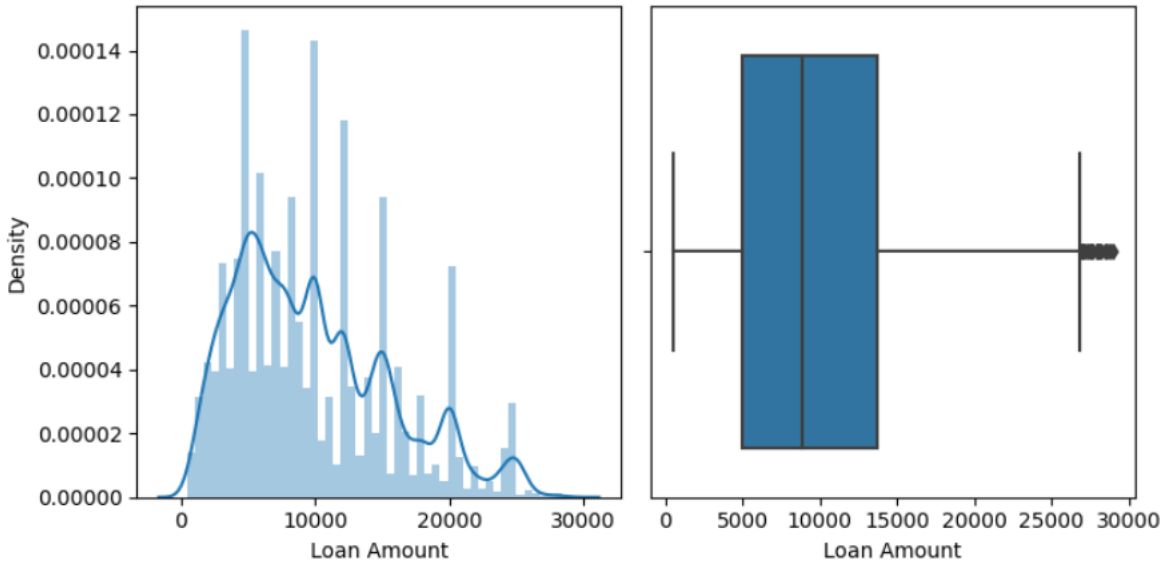
Steps followed:

- Data Cleaning and Manipulation
  - Dropping unnecessary Rows and Columns
  - Data Type Conversion for columns
  - Handle null values and Duplicate Data
  - Create Derived Columns
  - Filter necessary data and remove outliers
- Univariate Analysis
- Bivariate Analysis
- Derived Column Analysis
- Correlation Analysis

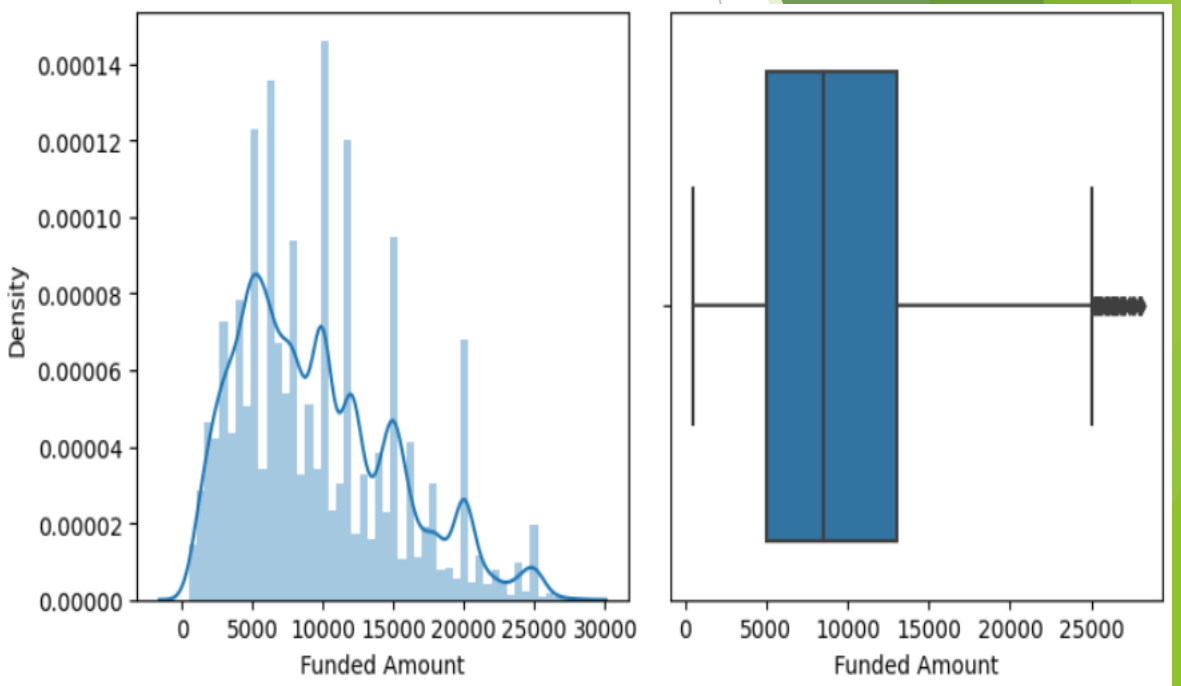
Note:

- **The analysis presented in this file is very limited and only important graphs and conclusions are presented. For detailed analysis, please refer to the jupyter notebook**
- **Missing Data Rules**
  - Columns with high percentage of missing values will be dropped (65% above for this case study)
  - Columns with less percentage of missing value will be imputed
  - Rows with high percentage of missing values will be removed (65% above for this case study)
- **Derived columns**
  - verification_status_n added. Considering domain knowledge of lending = Verified > Source Verified > Not Verified. verification_status_n correspond to {Verified: 3, Source Verified: 2. Not Verified: 1} for better analysis
  - issue_y is year extracted from issue_d
  - issue_m is month extracted from issue_d

Univariate Analysis:

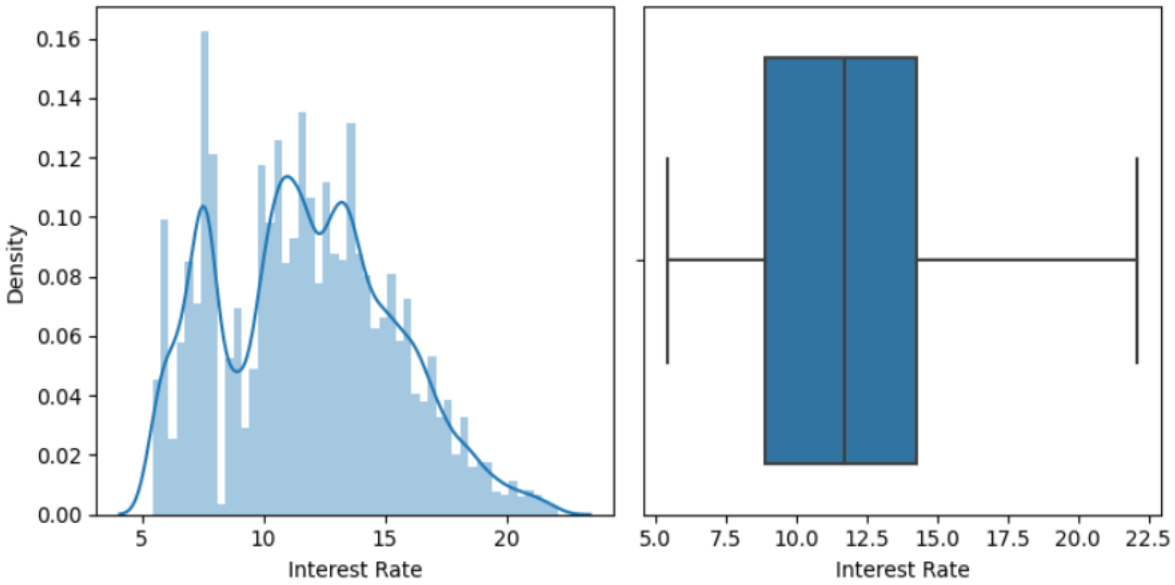


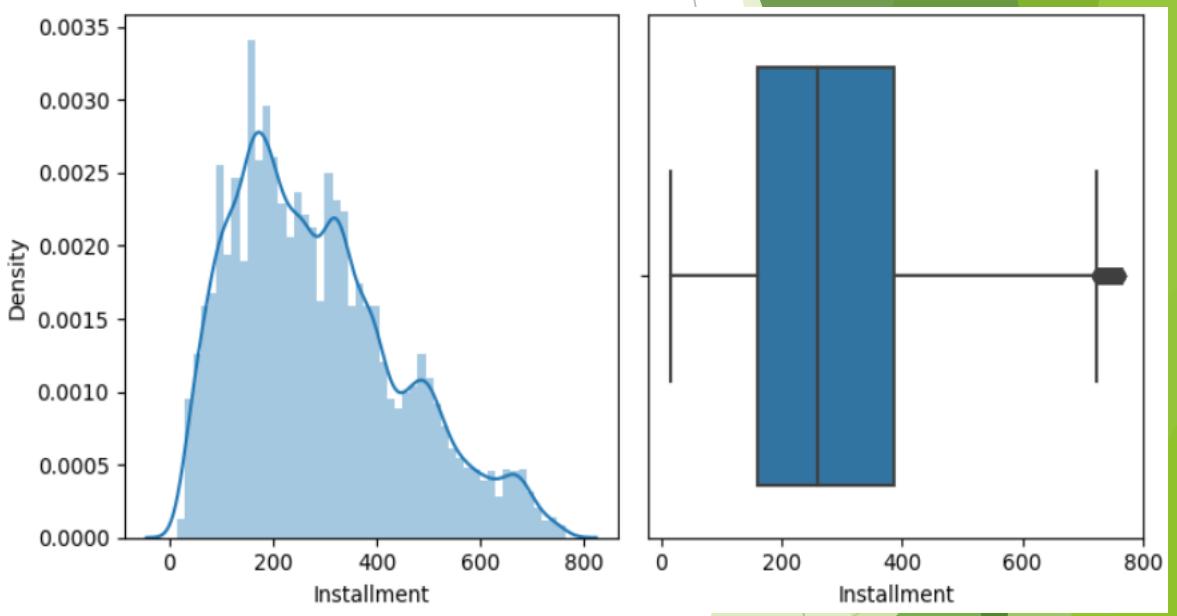
Majority of the loan_amount is in the range of 5K to 14K



Majority of the funded_amnt is in the range of 5K to 13K

Univariate Analysis:

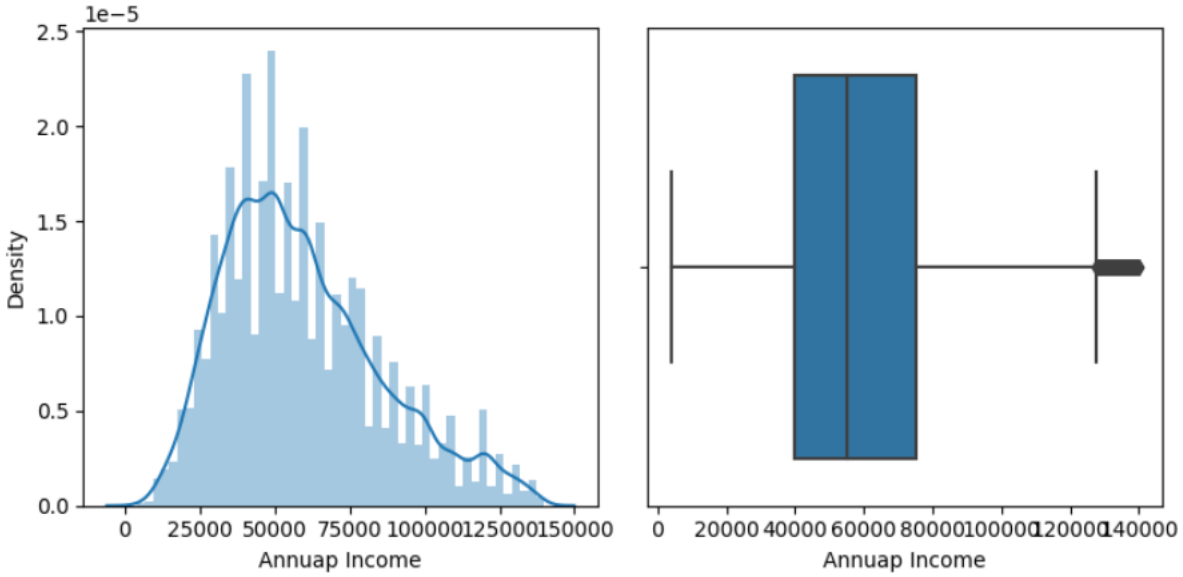


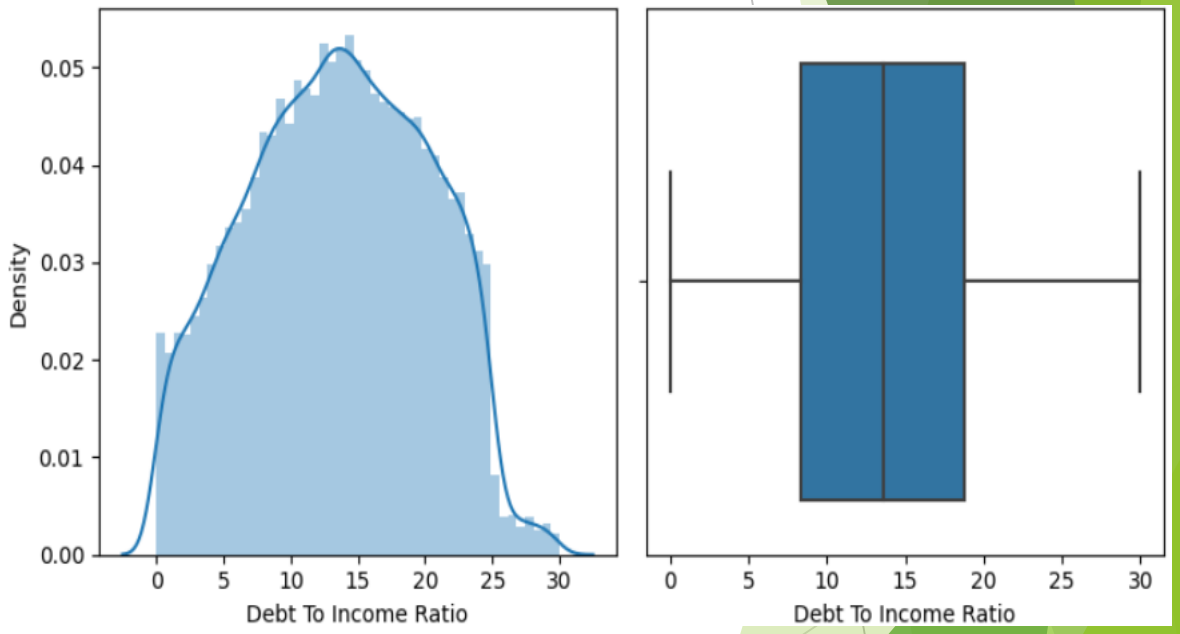


Majority of the interest rate is in the range of 5% to 16% going at the max to 22%

Majority of the installment is in the range of 20 to 400 going at the max to 700

Univariate Analysis:

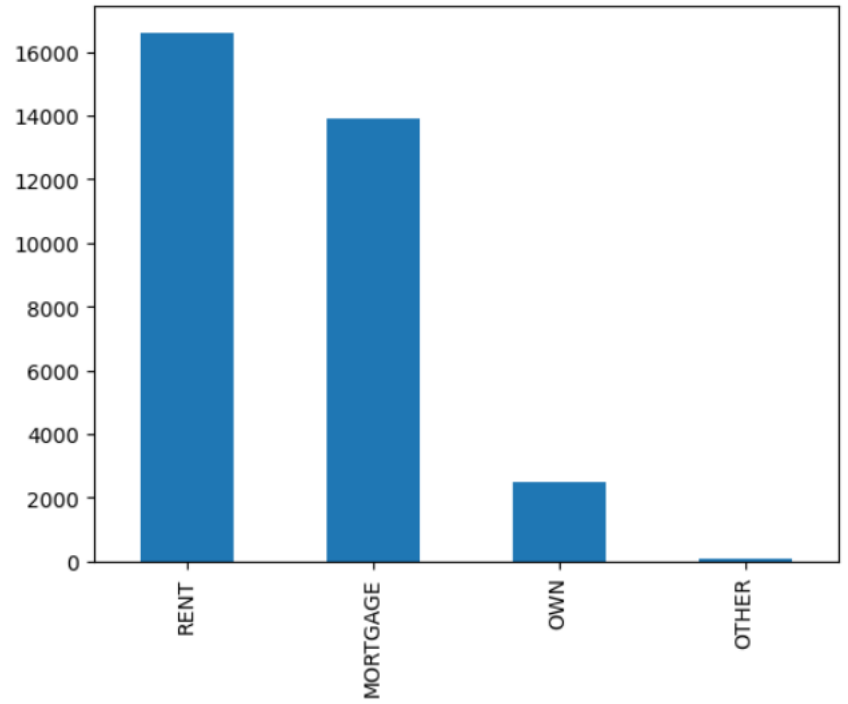




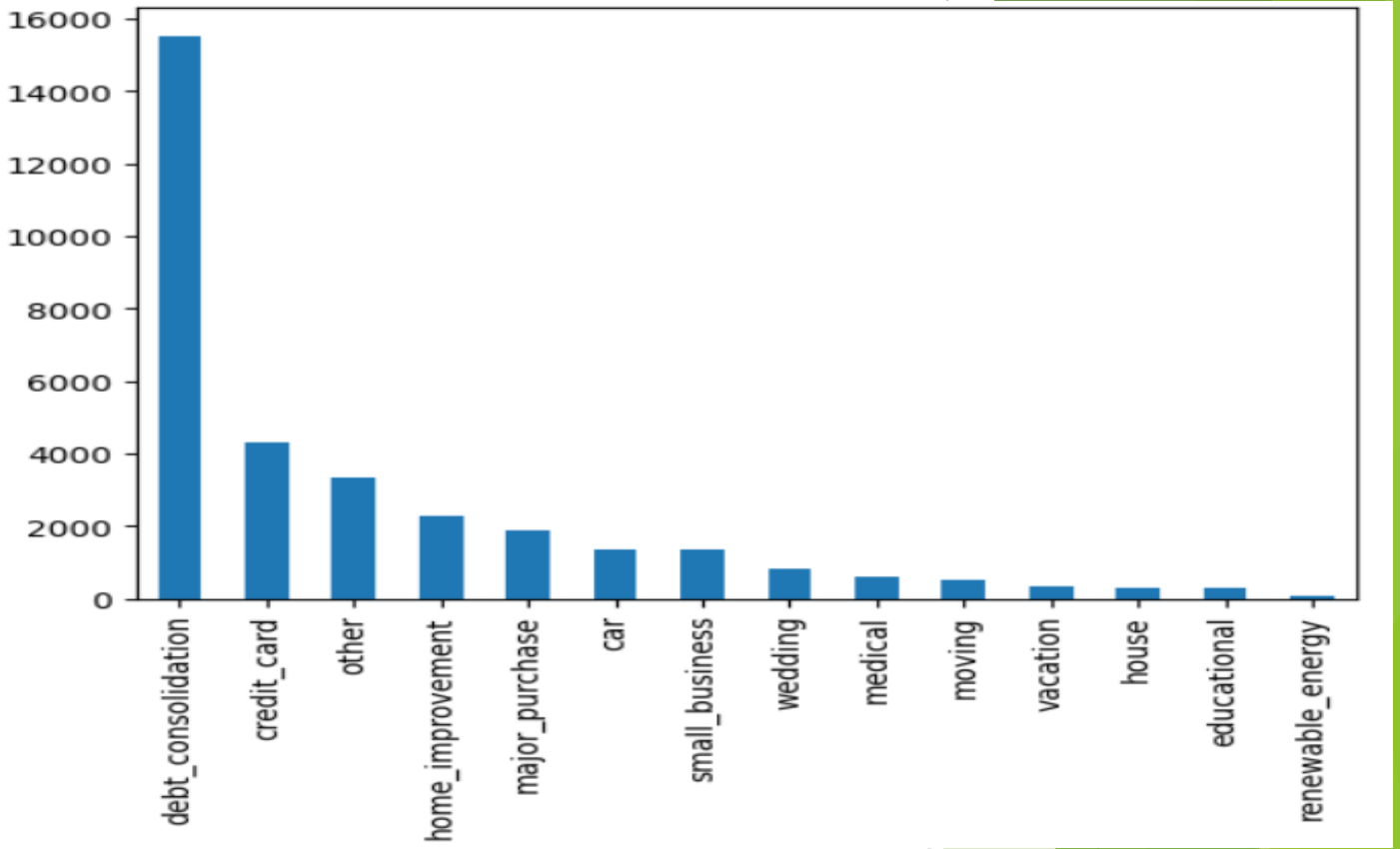
Majority of the annual income is in the range of 4k to 40k going at the max to 120k.

Majority of the debt to income ratio is in the range of 0 to 20 going at the max to 30

Univariate Analysis:



Majority of the home owner status are in status of RENT and MORTGAGE



Majority of loan application are in the category of debt_consolidation

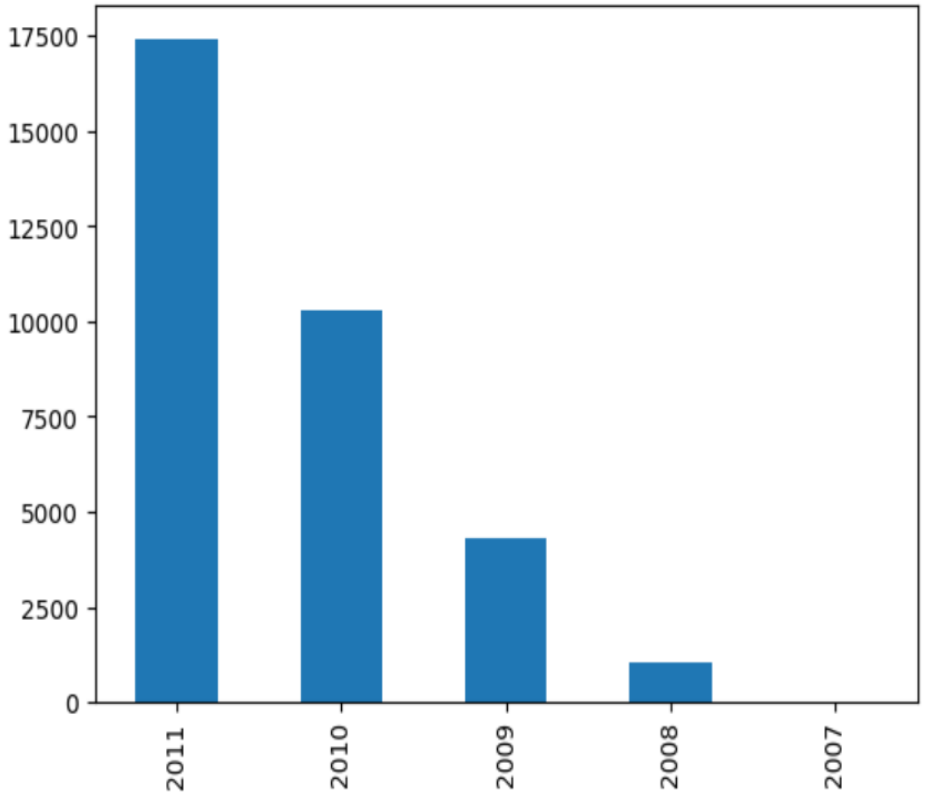
Univariate Variable Summary:

Customer Demographics
- Majority of the loan applicants are in the range of 0 - 40K annual income
- Majority of the debt to income ratio is in the range of 0 to 20 going at the max to 30
- Majority of the home owner status are in status of RENT and MORTGAGE
- Highest loan applications are in the category of debt consolidation
- CA (California) state has the maximum amount of loan applications
- Majority of the loan applicants are in the category of not having an public record of bankruptcies
- Majority of the employment length of the customers are 10+ years and then in the range of 0-2 years

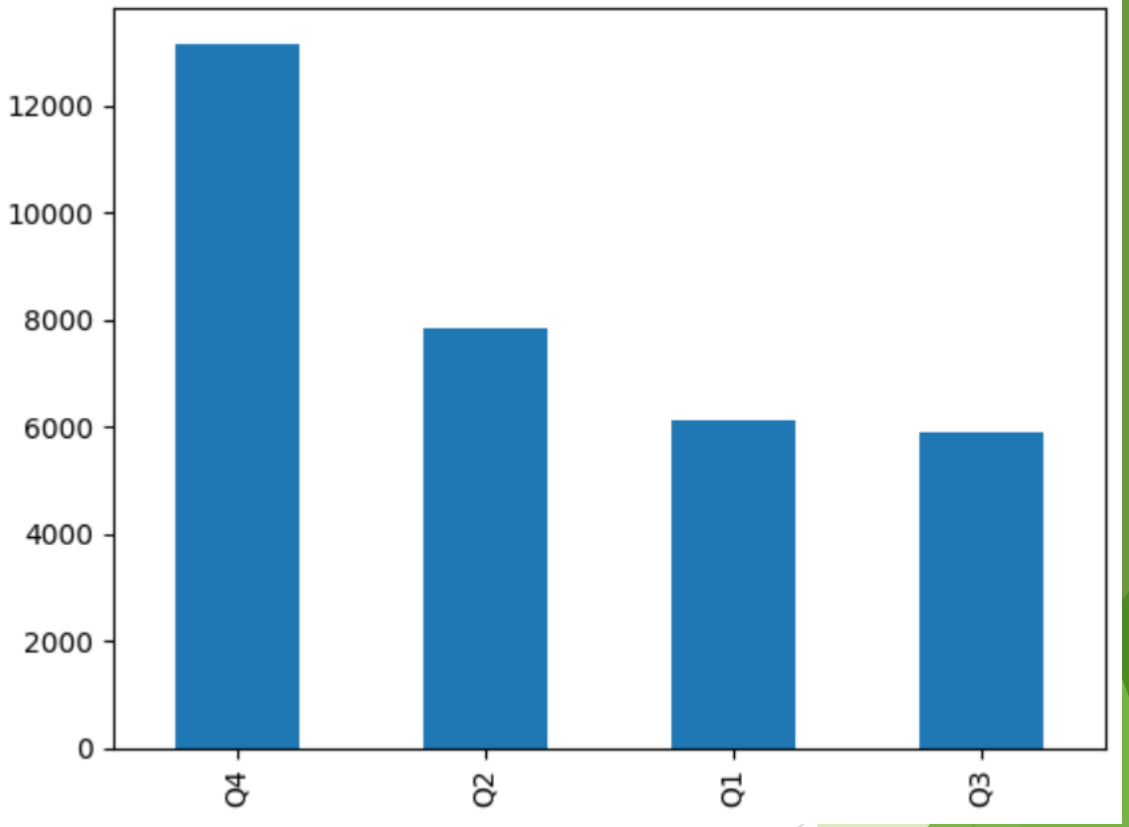
Loan Demographics
- Highest loan amount applications fall in the range of 5k to 10k
- Majority of the interest rate is in the range of 5% to 16% going at the max to 22%
- Majority of the loan applications counts are in the term of 36 months
- Majority of loan application counts fall under the category of Grade B

Time Based Analysis
- Loan application counts are increasing year over year
- Highest loan application volume in Quarter 4 of every year
- Lowest loan applications are in Q1
    - Possibly because by year ends people face the financial challenges
    - Possibly because of festive seasons
    - Possibly because they are consolidating debt by year end

Derived Variable Analysis:



Loan application counts are increasing year over year. Maybe the risk exposure is increasing over the year

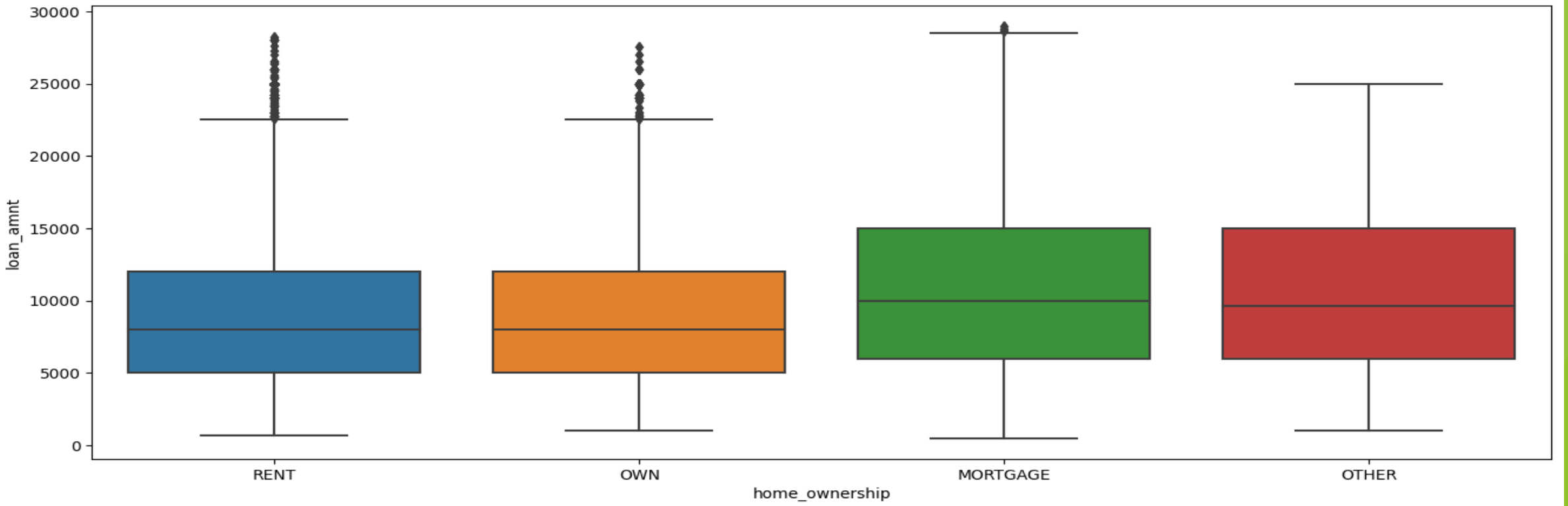


Highest loan application volume in Quarter 4 of a year

Derived Variable Summary:

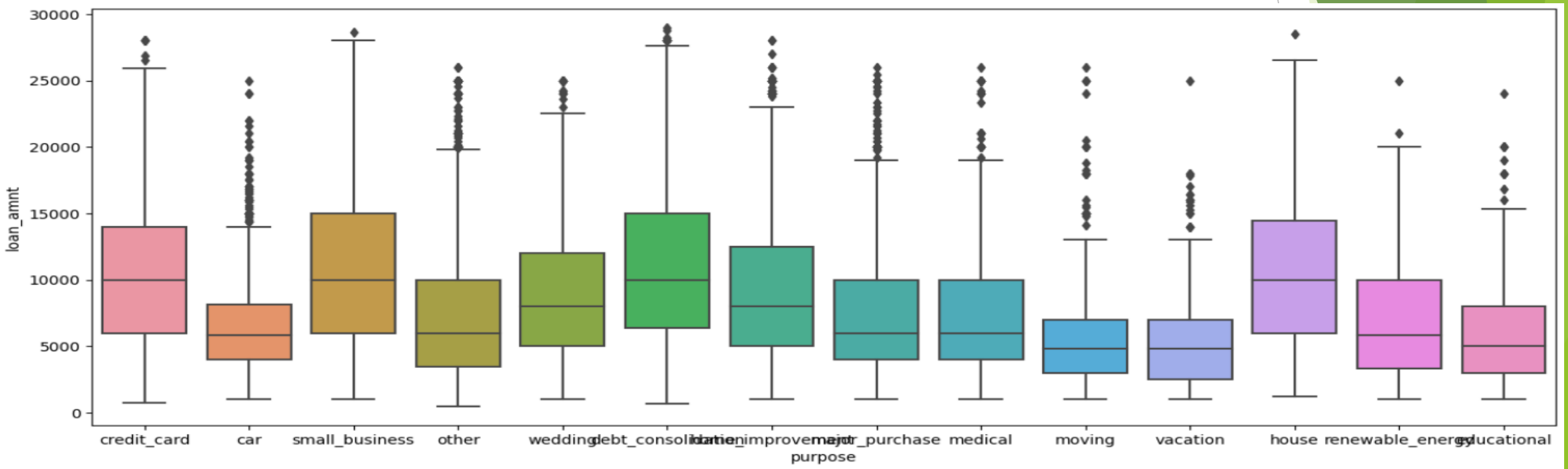
- Highest loan application volume in Quarter 4 of a year
- Loan application counts are increasing year over year. Maybe the risk exposure is increasing over the year
- The lowest loans application count are in the month of Jan/Feb/March and highest counts are in Nov/Dec.
- Highest loan amount applications fall in the range of 5k to 10k
- Majority of the loan applications are in the category of Very Low interest rates
- Majority of the loan applicants are in the range of 0 - 40K annual income
- Majority of the loan applications are in Moderate debt to income ratio ratio

Bivariate Analysis and Summary:



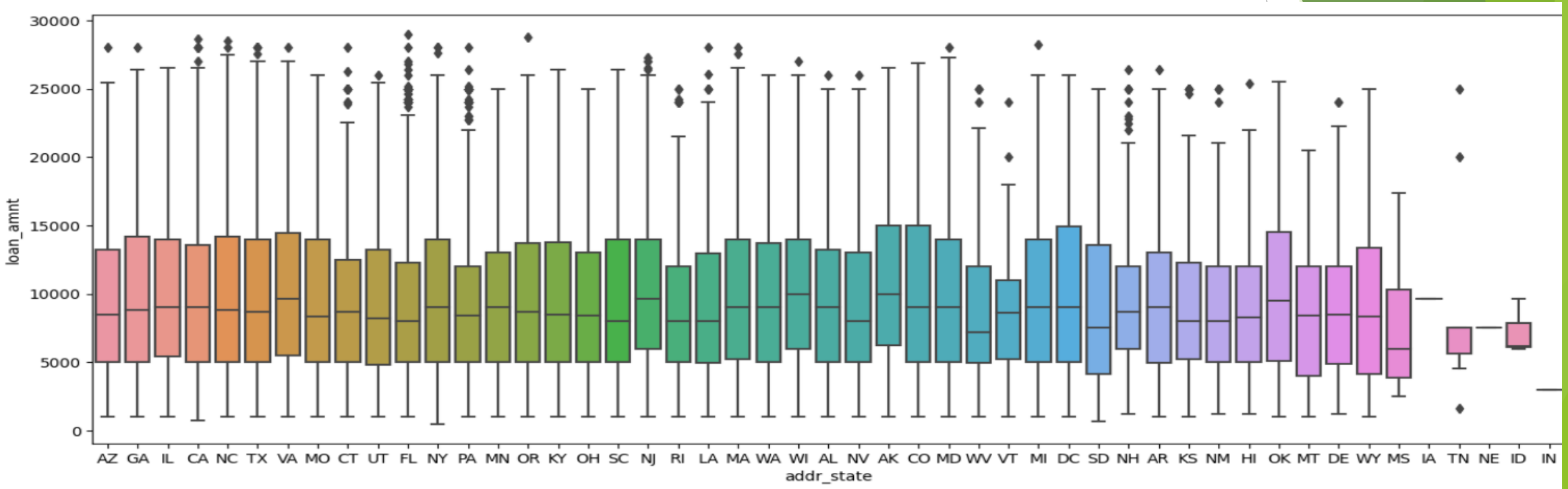
- Overall highest Charge Off numbers are in the category of RENT and MORTGAGE
- Within each home_ownership category, the ratio of Charge Off's for others is higher

Bivariate Analysis and Summary:



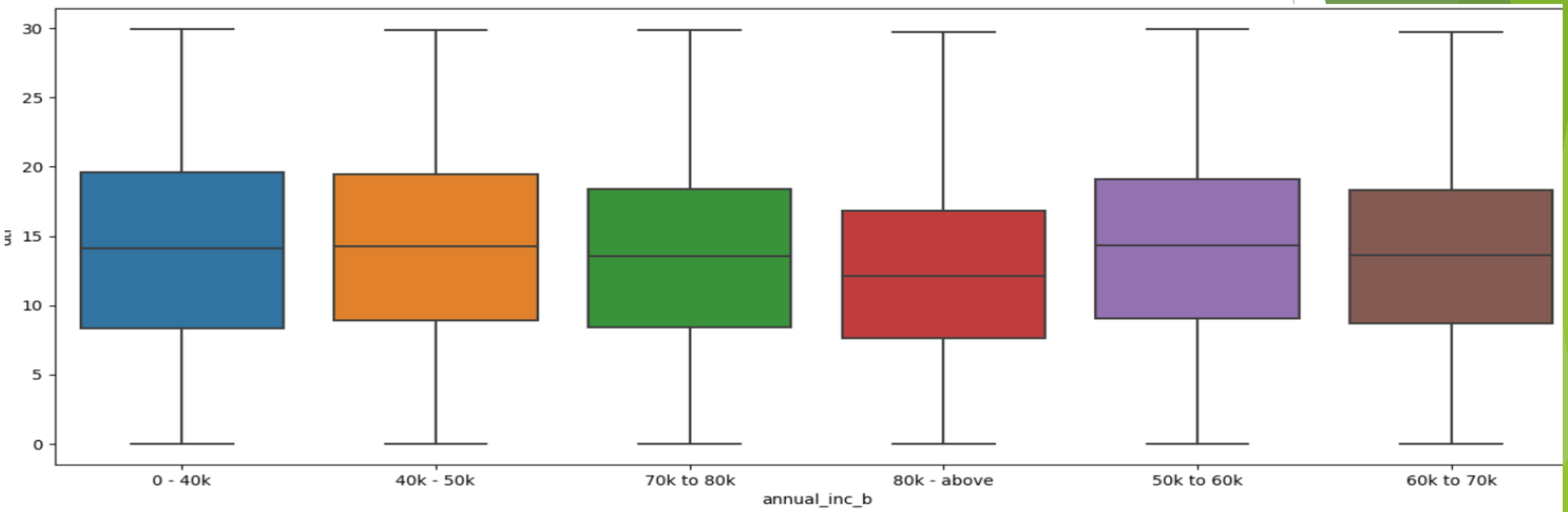
- The highest risk of Charge Offs are the category of debt_consolidation
- The highest probability of Charge Offs within a category is small_business but the volume is extremely low
- The highest loan amount ranges are in small business, debt consolidation and house

Bivariate Analysis and Summary:



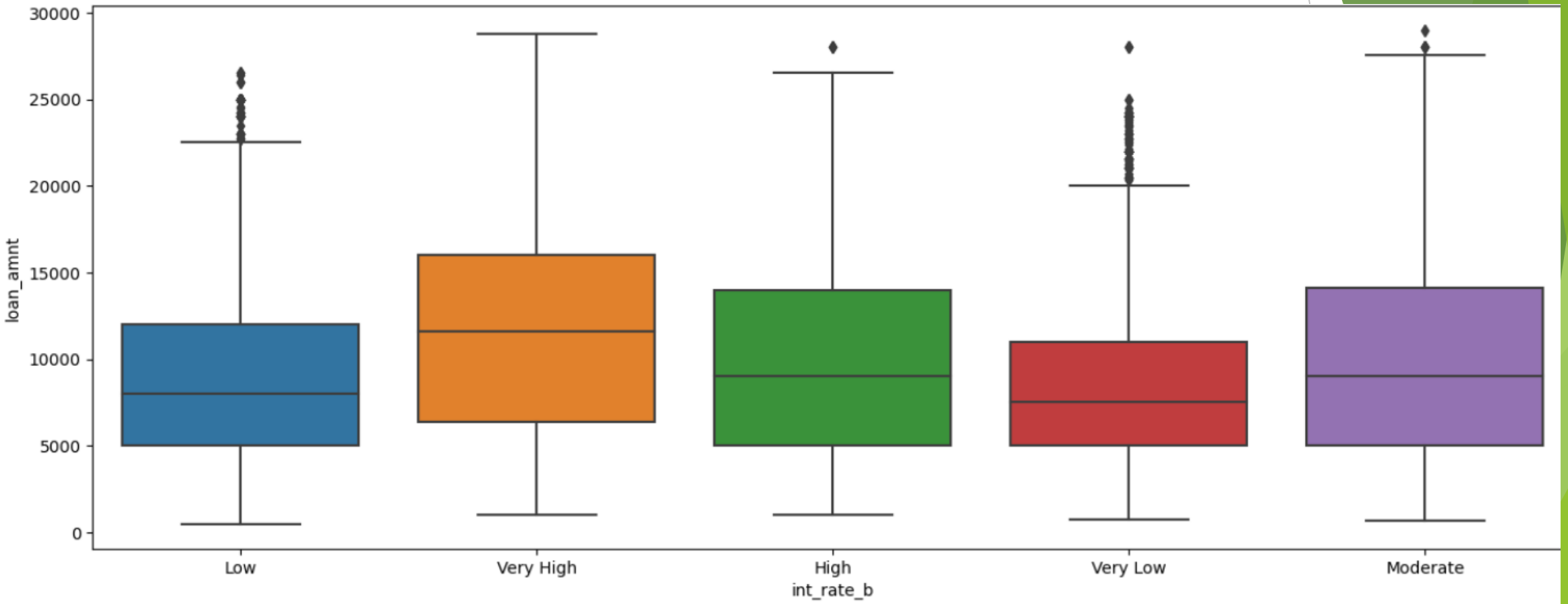
- The highest volume of loans is from CA and purely based on volumes the highest Charge Off's are from CA
- Within each state, NE and NV have the highest Charge Offs
- NE has a very low volume this cannot be considered
- Loan applications from NV will have a high risk

Bivariate Analysis and Summary:



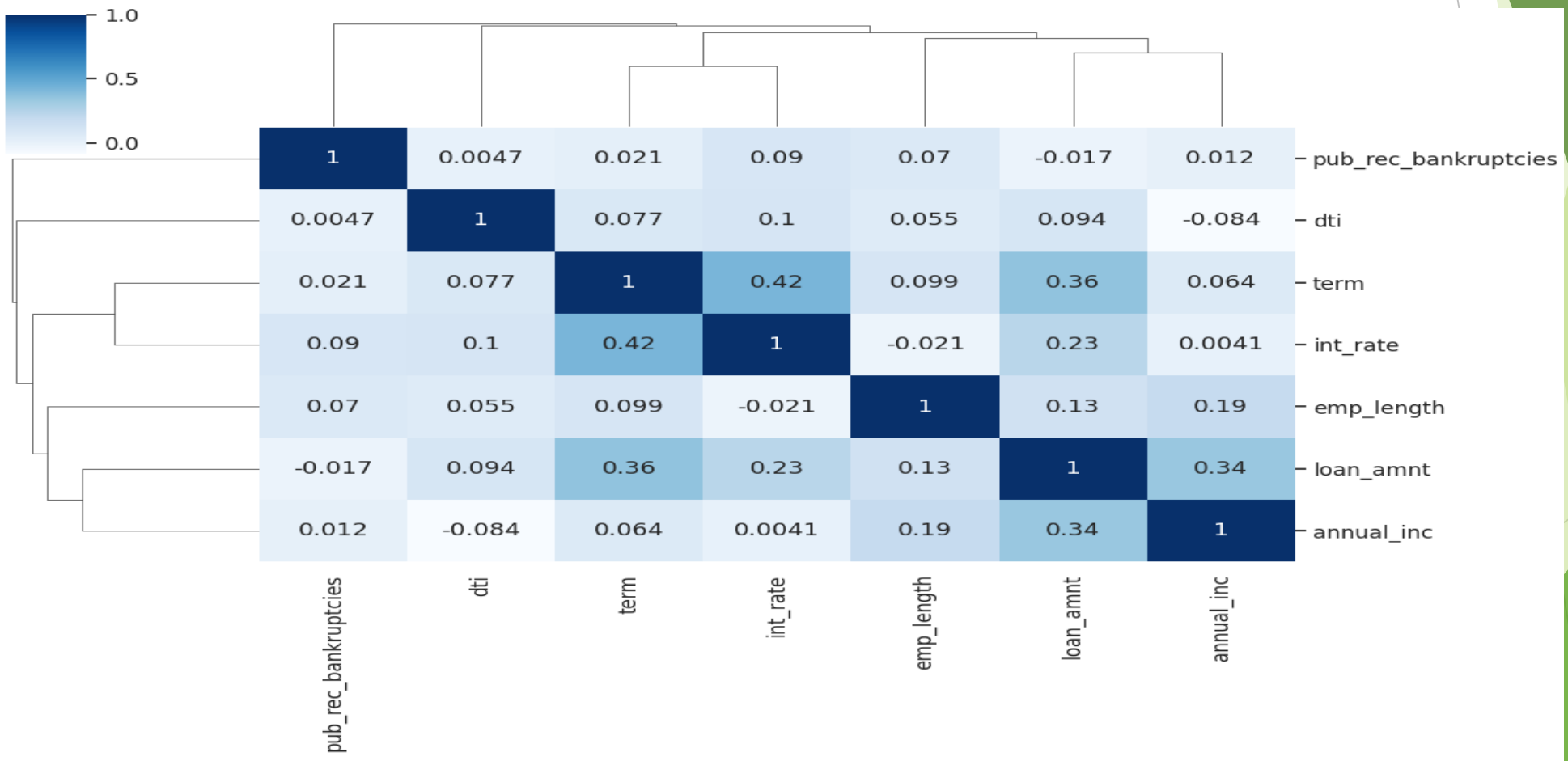
- Annual income range of 0-40K has the highest charge offs
- Charge-off ratio within the bucket of 0-40K have the highest Charge Offs

Bivariate Analysis and Summary:



- Based on volume and based on the Charge Off ratio within the category, the Very High-interest rates are at risk of Charge Off
- Very High-interest rate is 15% and above

Correlation Analysis:



Correlation Summary:

- **Negative Correlation**
  - loan_amnt has a negative correlation with pub_rec_bankrupticies
  - annual income has a negative correlation with dti (Debt to income ratio)
- **Strong Correlation**
  - term has a strong correlation with loan amount
  - term has a strong correlation with interest rate
  - annual income has a strong correlation with loan_amount
- **Weak Correlation**
  - pub_rec_bankruptcies has a weak correlation with most of the fields

Conclusion:

- **Major Driving factors which can be used to predict the chance of defaulting and avoiding Credit Loss:**
  - DTI (Debt to income ratio)
  - Grades
  - Verification Status
  - Annual income
  - Pub_rec_bankruptcies

- **Other reasons for defaults :**
  - Borrowers not from large urban cities like California, new york, texas, florida etc.
  - Borrowers having annual income in the range 50000-100000.
  - Borrowers having Public Recorded Bankruptcy.
  - Borrowers with least grades like E,F,G which indicates high risk.
  - Borrowers with very high debt to income ratio value.
  - Borrowers with working experience 10+ years.