GRAMENER CASE STUDY

SUBMISSION

Group Name:

1. Richa Dhir
2. Imran Khan
3. Lakshmi Bhargavi P.

Abstract

- The study was performed by a consumer finance company to understand the trends among defaulters, i.e., customers who borrow money and refuse to pay or runs away with the money owed.

- The study focuses on two objectives:
  1. To understand how consumer attributes influence the tendency of default.
  2. To understand how loan attributes influence the tendency of default.

- The overall business strategy is to 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, thereby avoiding potential customers who are likely to default on a loan.

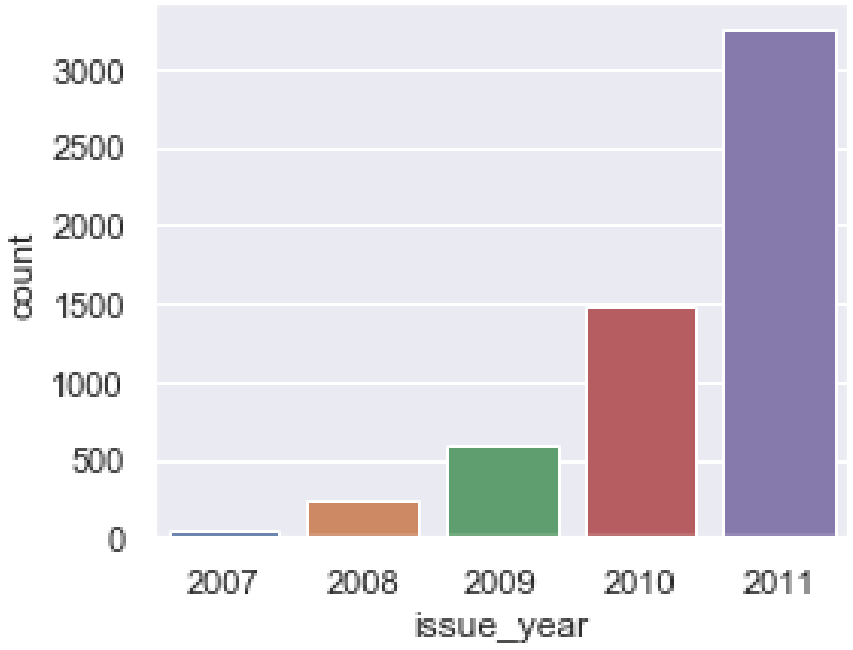
Problem solving methodology

- The data analysis strategy is to showcase the various factors that are strong indicators of loan default.
- The first step is to understand the business domain and variables the variables given in the dataset.
  - This is accomplished by performing industry research and referring to the data dictionary.
- Then, perform data cleaning after importing the data from the CSV file.
  - This involves dropping columns (i.e., those containing no values, those that have high percentage of missing values, those that are not required for the analysis), imputing missing values for certain columns, standardizing the column names of all variables, standardizing the casing of values in a column, ensuring that variable are stored in the proper data type (e.g., date stored as an object data type instead of datetime data type).
- Next, determine the derived metrics (business-driven metrics, data-driven metrics, and type-driven metrics) that would help us draw inferences from the dataset.
  - This includes issue year (i.e., the year in which loan was approved; it's a type-driven metric), income cat (i.e., the category of income ; it's a data-driven metric), and employment years (i.e., the number of years since the applicant has been employed; it's a data-driven metric).
- Plot various graphs of the relationship between the variables in the dataset to determine if they are the driving factors of loan default.
  - This includes box plots of numerical variables, numerical variable with hue = categorical variables.
- Finally, discover the variables that are strong indicators of default.

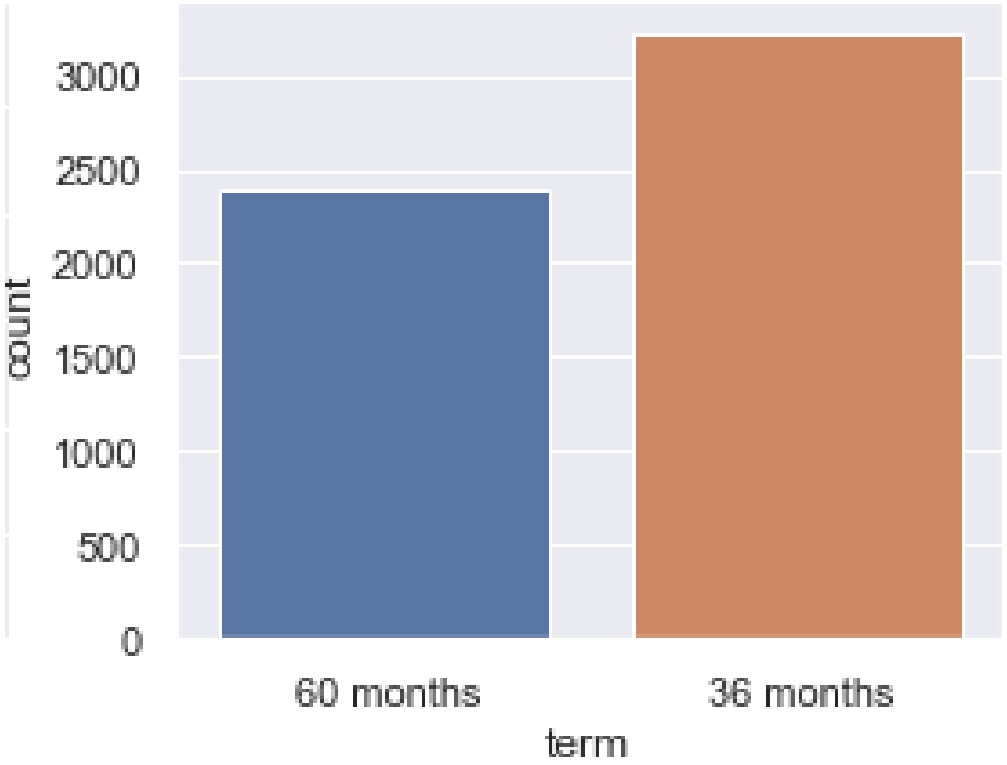
Analysis

- The following types of data analysis were performed.
  - Univariate Analysis
  - Bivariate Analysis
  - Segmented Analysis

Univariate Analysis

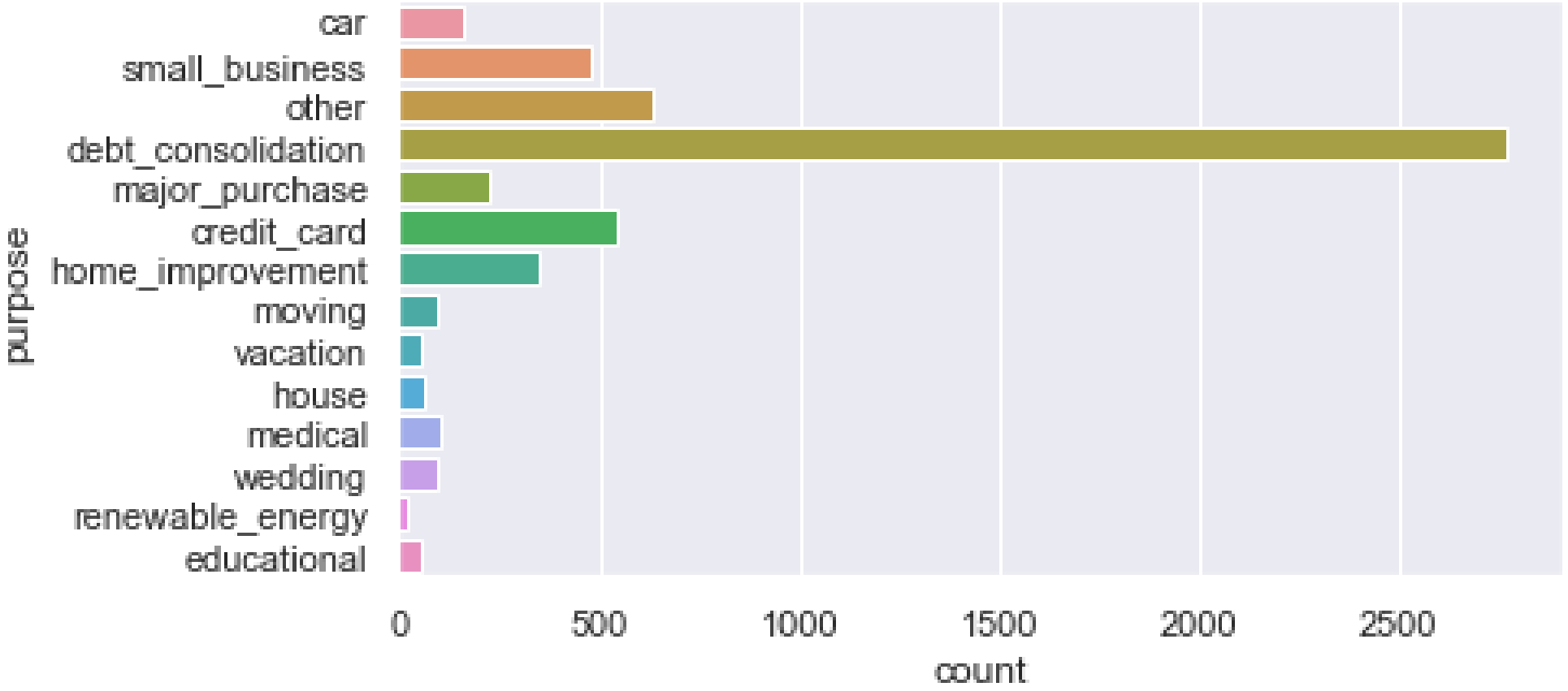


This shows that most loans that defaulted were **issued** in the year **2011**.



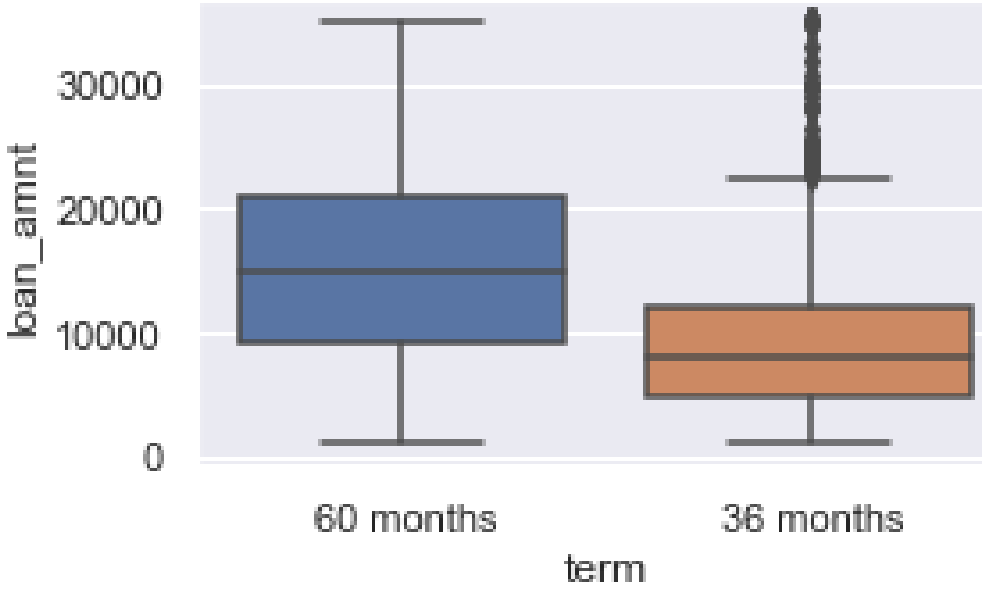
This shows that the **term** of most loans that defaulted was **36 months**.

Bivariate Analysis

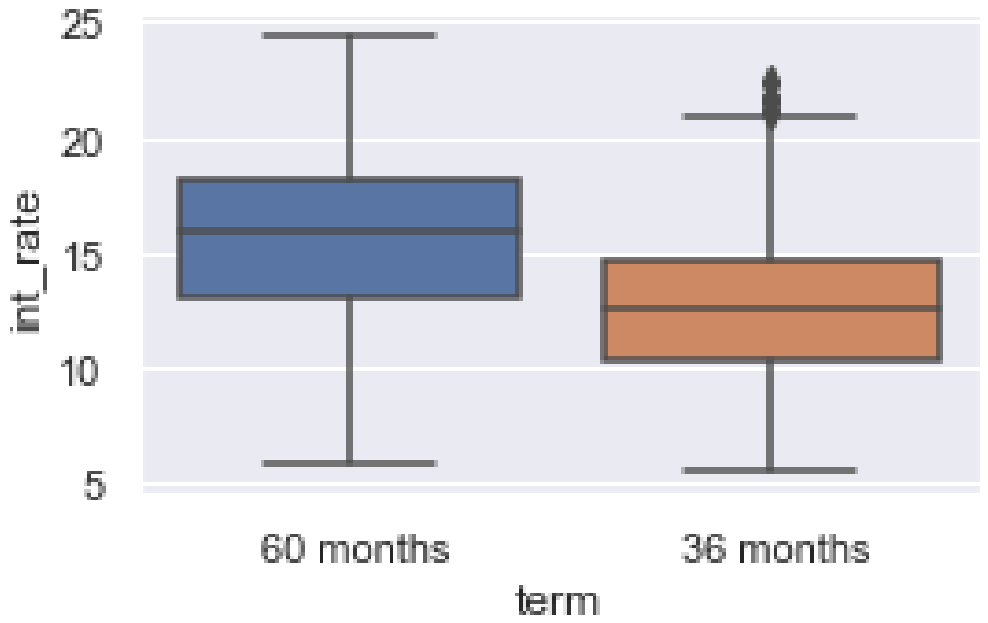


This shows that most loans that defaulted were issued for the **purpose** of **debt consolidation**.

Bivariate Analysis

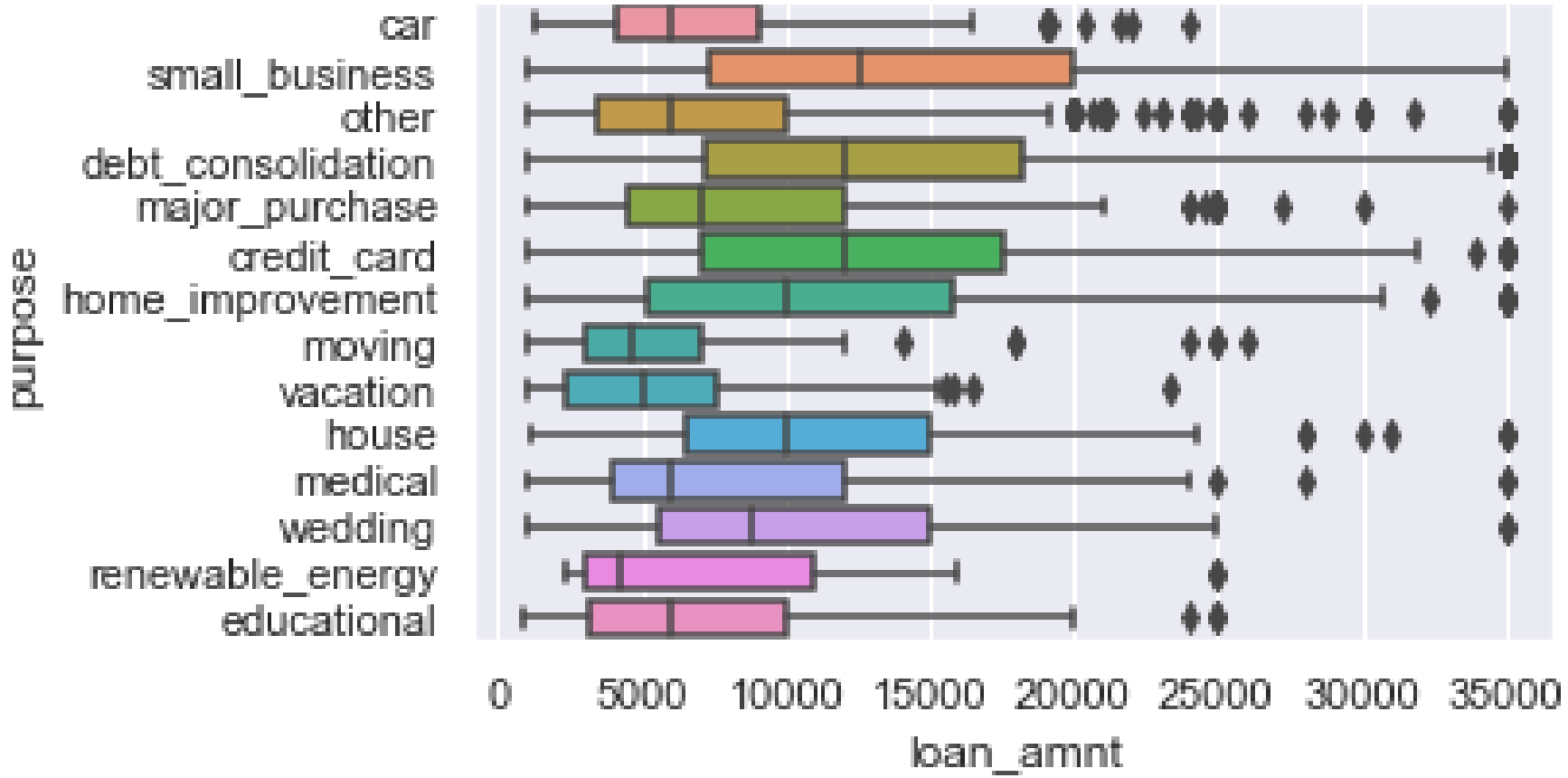


This box plot shows that the **maximum loan amount** was issued when the **term** is **60 months**.



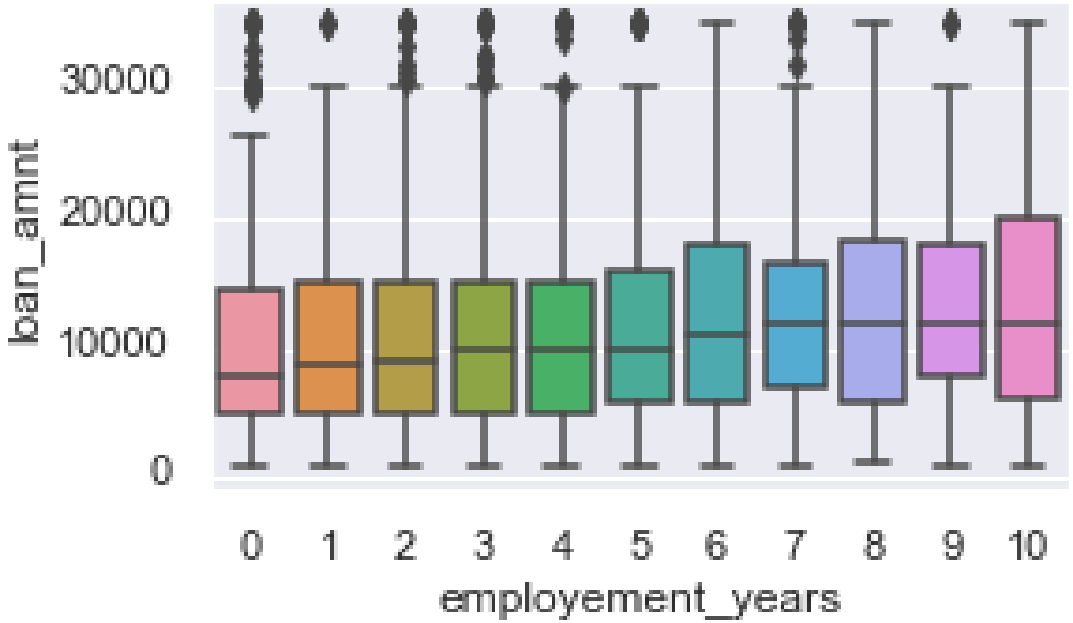
The box plot shows *a proportional relationship* between the loan's interest rate and its term.

Bivariate Analysis

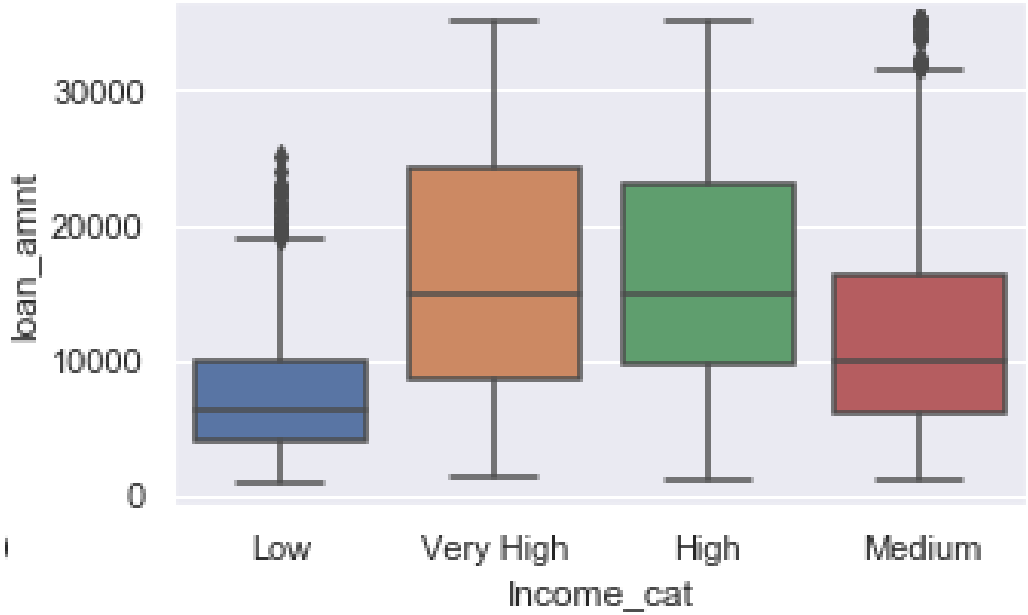


This box plot shows that the **maximum loan amount defaulted** was issued for the **purpose** of **small businesses**.

Bivariate Analysis

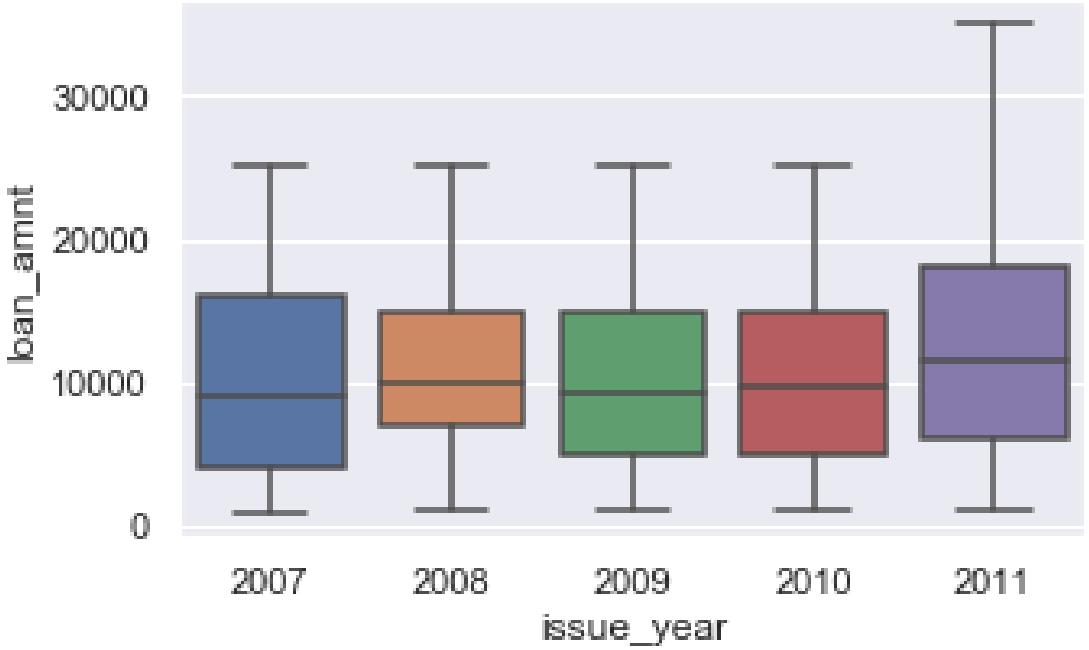


This box plot shows *a proportional relationship* between the **loan amount issued** and the **applicants' years of employment**, i.e., Year 0 has the lowest loan amount issued and Year 10 has the highest loan amount issued.

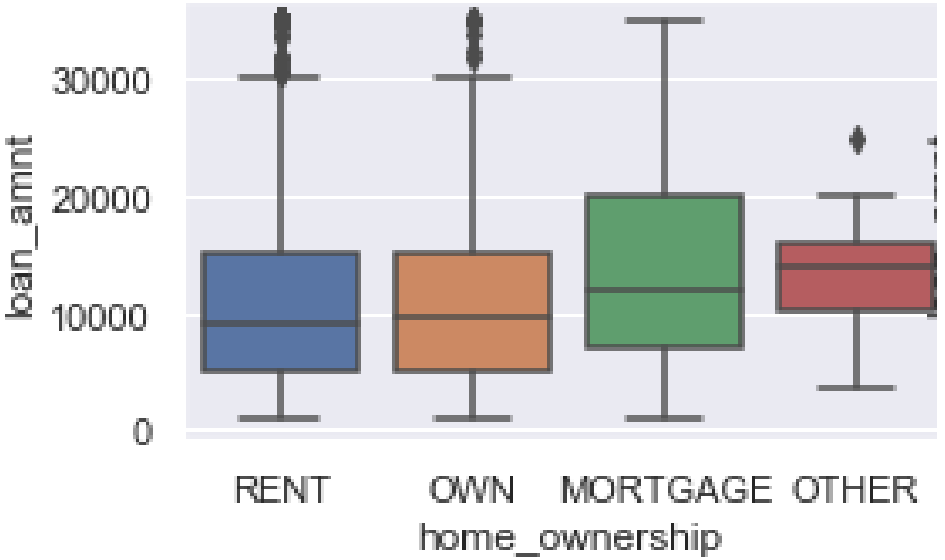


This box plot shows *a proportional relationship* between the **loan amount issued** and the **applicants' income category**, i.e., low-income customers are issued the lowest loan amount and very high-income customers are issued the highest loan amount.

Bivariate Analysis

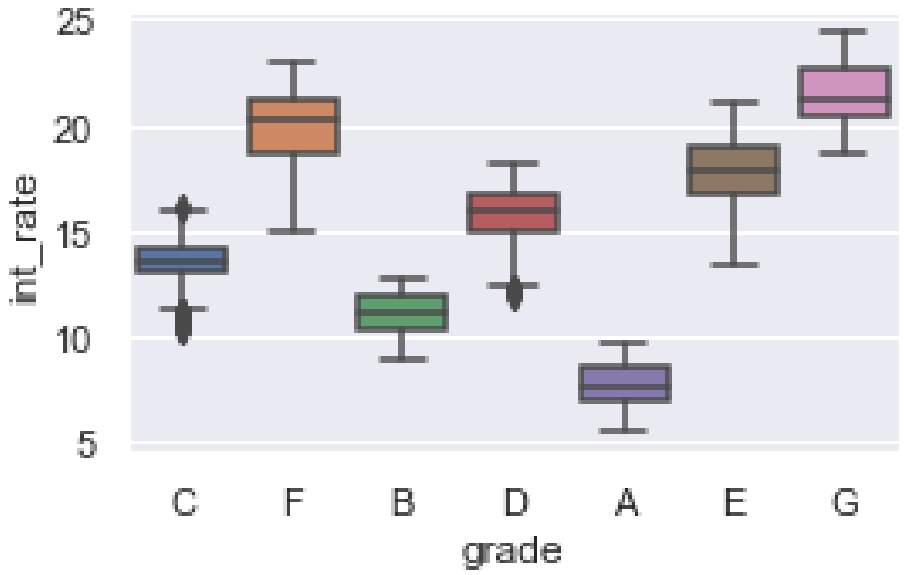


This box plot shows that maximum loan amount that defaulted was **issued** in the year **2011**.

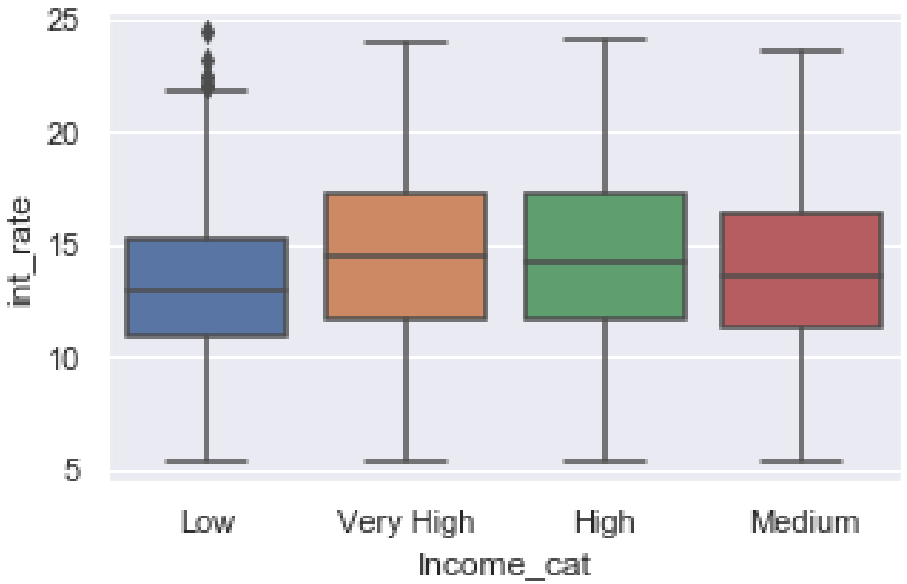


This box plot shows that maximum loan amount that defaulted was **issued** to people who paid mortgage on their home.

Bivariate Analysis



This box plot shows *a proportional relationship* between the loan's **grade** and **interest rate**, i.e., Grade A has the lowest interest rate and Grade G has the highest interest rate.



The box plot shows *a proportional relationship* between the loan's interest rate and the applicant's income category.

Conclusions and Inferences

- The driving factors of loan default are as follows:
  - **Applicant's income category:** There is a proportional relationship between the applicant's income category and the amount of defaulted loans.
  - **Home ownership:** Consumers who are paying mortgage are more likely to default on loans than rent, own, and others.
  - **Purpose:** Customers who have taken loans for the purpose of small business are more likely to default on the loans.
  - **Term:** Customers who have applied for a term of 60 months are more likely to default the loans.
  - **Grade:** There is a proportional relationship between the loan's grade and the probability of the defaulted loans.