UpGrad

iiit-b

LendingClub

Case Study
by Vimala Subramanian

Problem Statement

A consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Solution Approach

Data Loading and Cleaning

- Import Libraries.
- Read CSV file in Dataframe.
- Drop/Impute Columns

Fix Data Type & Categorize

- Fix data types.
- Add derived columns.
- Categorize columns for analysis.

Univariate Analysis

- Numerical
- Categorical
  - Orderded
  - Unordered

Bivariate Analysis

- Numerical vs Numerical
- Numerical vs Categorical
- Categorical vs Categorical

Multivariate Analysis

- Correlation Analysis using heatmap

Data Cleaning

Initial Observation about the dataframe

1. There are lot of columns with all null values, which can be dropped
2. Few columns has very less null values
3. Some of the fields are of datatype = object which needs to be converted to int or float after data cleaning and imputing
4. Columns with missing values needs to be imputed either by removing or replacing it with a standard value
5. Date fields like payment_d are with dtype as object - this might need to be convered to an data/time field
6. Some columns might not be of meaningful value for further analysis like url which can be dropped off. But let us analyze before dropping off.

Steps taken for data cleaning

1. Removing All null value Columns [Removed from 111 to 57 columns]
2. Removing columns which doesn't add meaning full value for analysis
   a. For ex: Columns having same values for all rows or having more zero values
   b. Fields like desc, url, emp_title etc. doesn't support in further analysis
3. Removing All rows with Nan or null values
4. Data Imputing - Fixing missing values either by using majority of values used
   a. For ex: mths_since_last_record = filling blank or Nan rows with 0
5. Adding Derived columns for year and month for issue_d columns

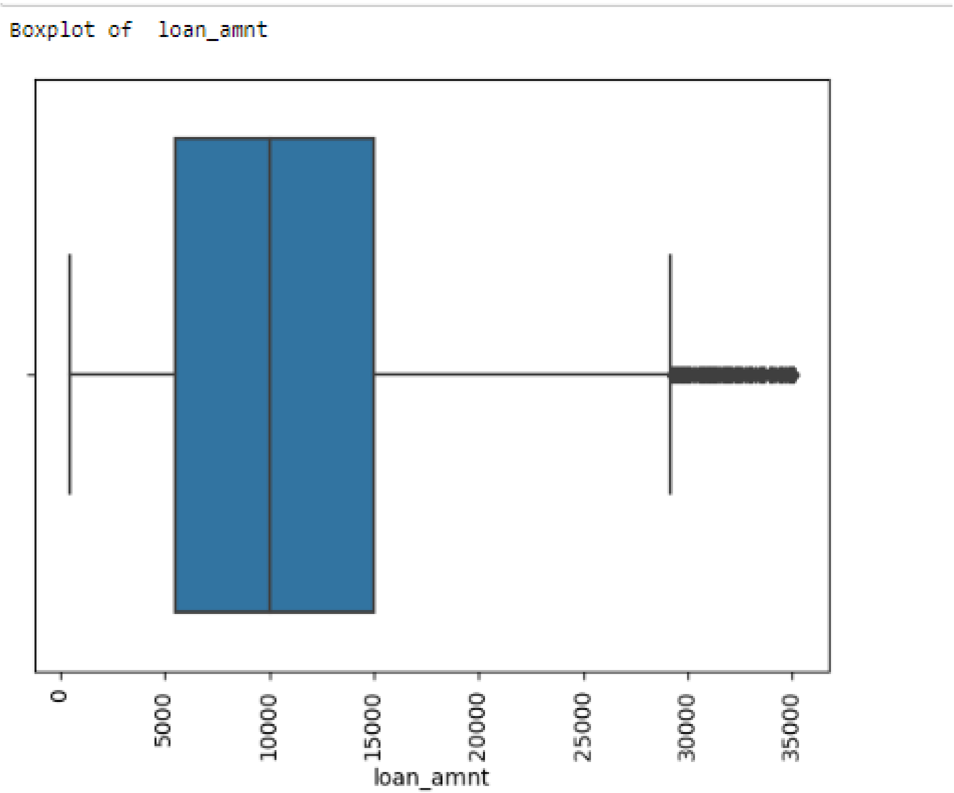
Univariate

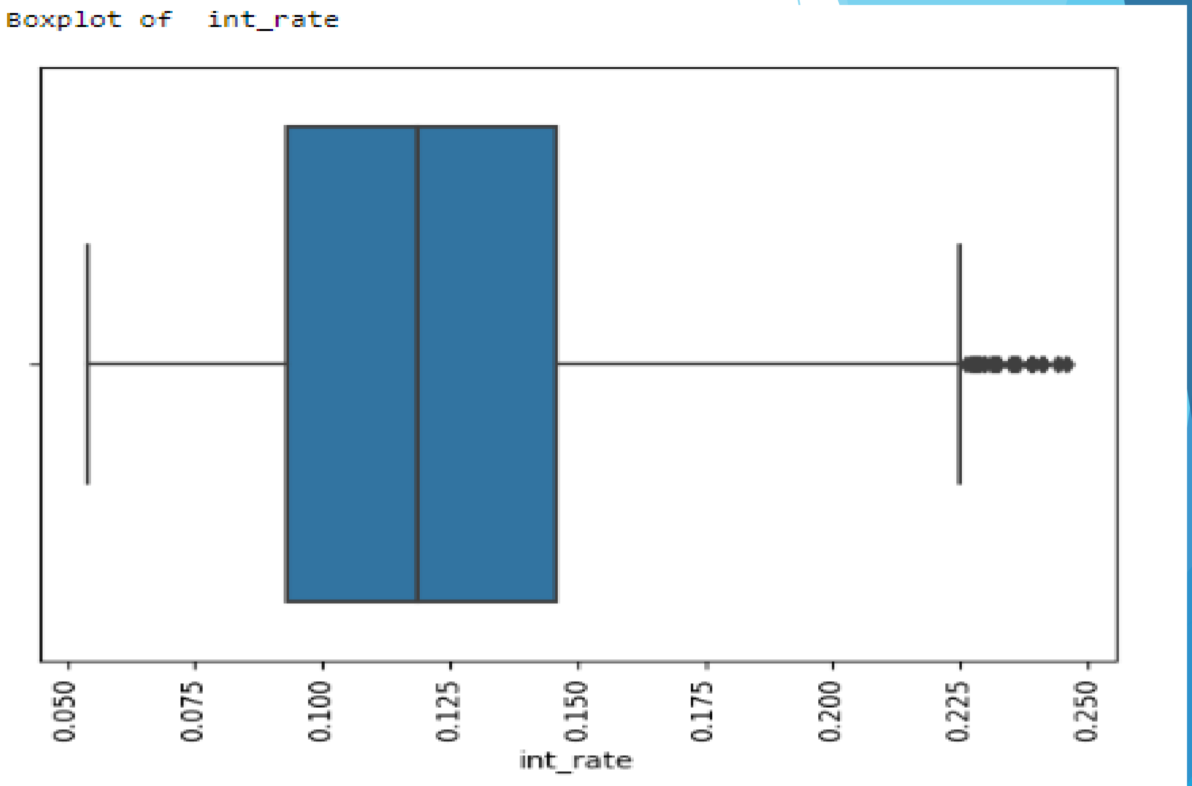
(Numerical)

Bivariate

Multivariate

Analysis





Inference

1.Box Plot for loan_amnt shows, outliers >30000 which means the rows greater than 30000 can be dropped of.

2.Box Plot for funded_amnt shows, same as loan_amnt which means the column can be dropped of.

3.Box Plot for int_rate shows, outliers > 22.5% interested rate which means the rows greater than can be dropped of.

4.Box Plot for emp_length shows, the median is 4.

5.Box Plot for total_pymnt shows, outliers >30000

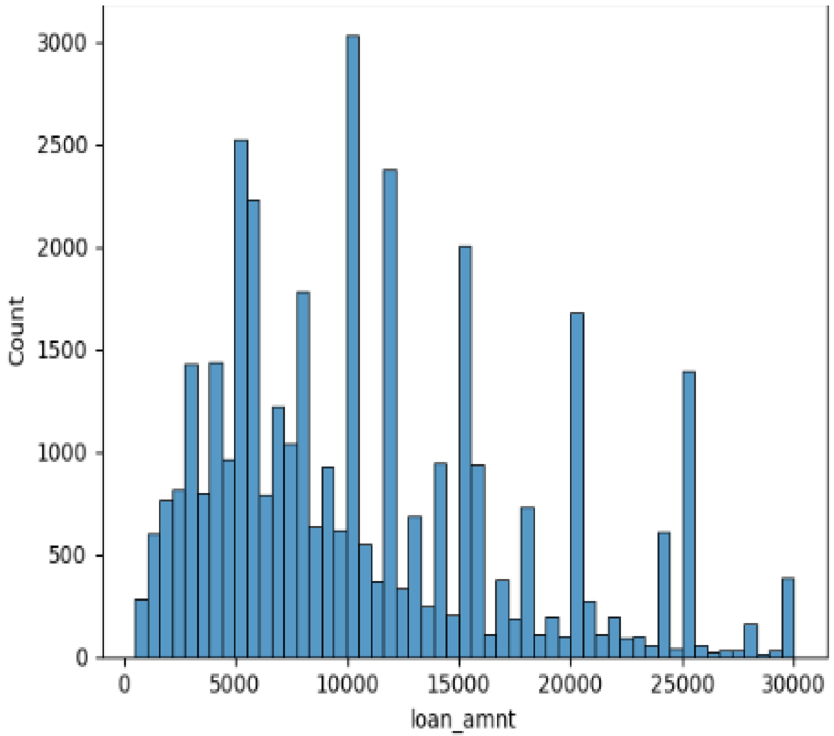
Univariate

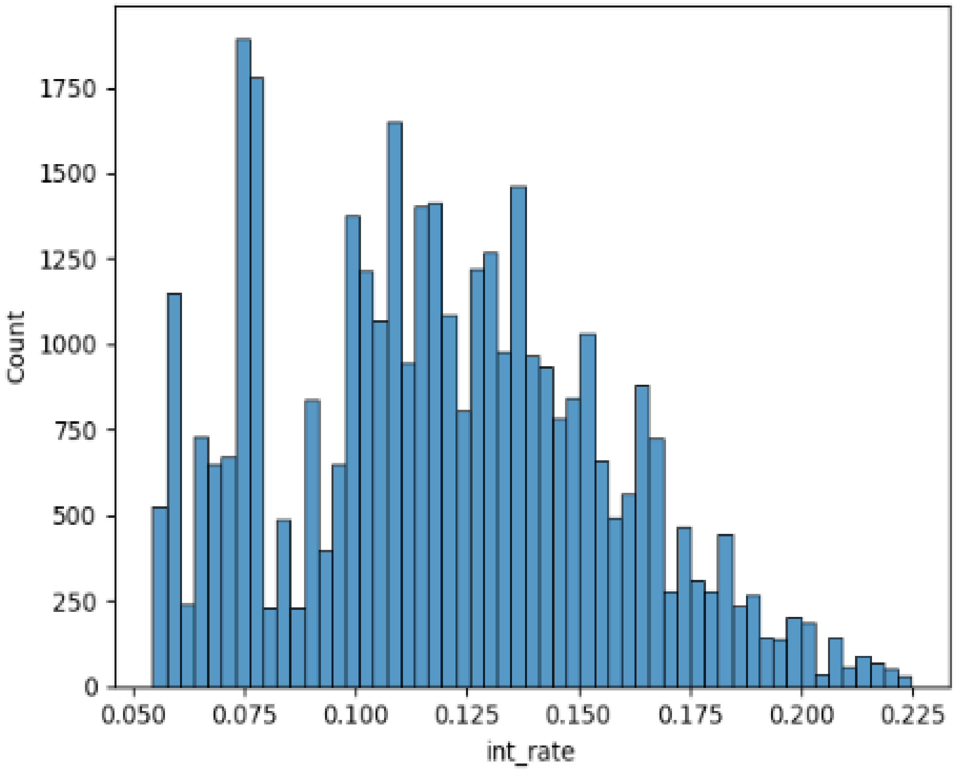
(Numerical) after data clean up

Bivariate

Multivariate

Analysis





Inference

1.Box Plot for loan_amnt and int_rate after removing the outliers

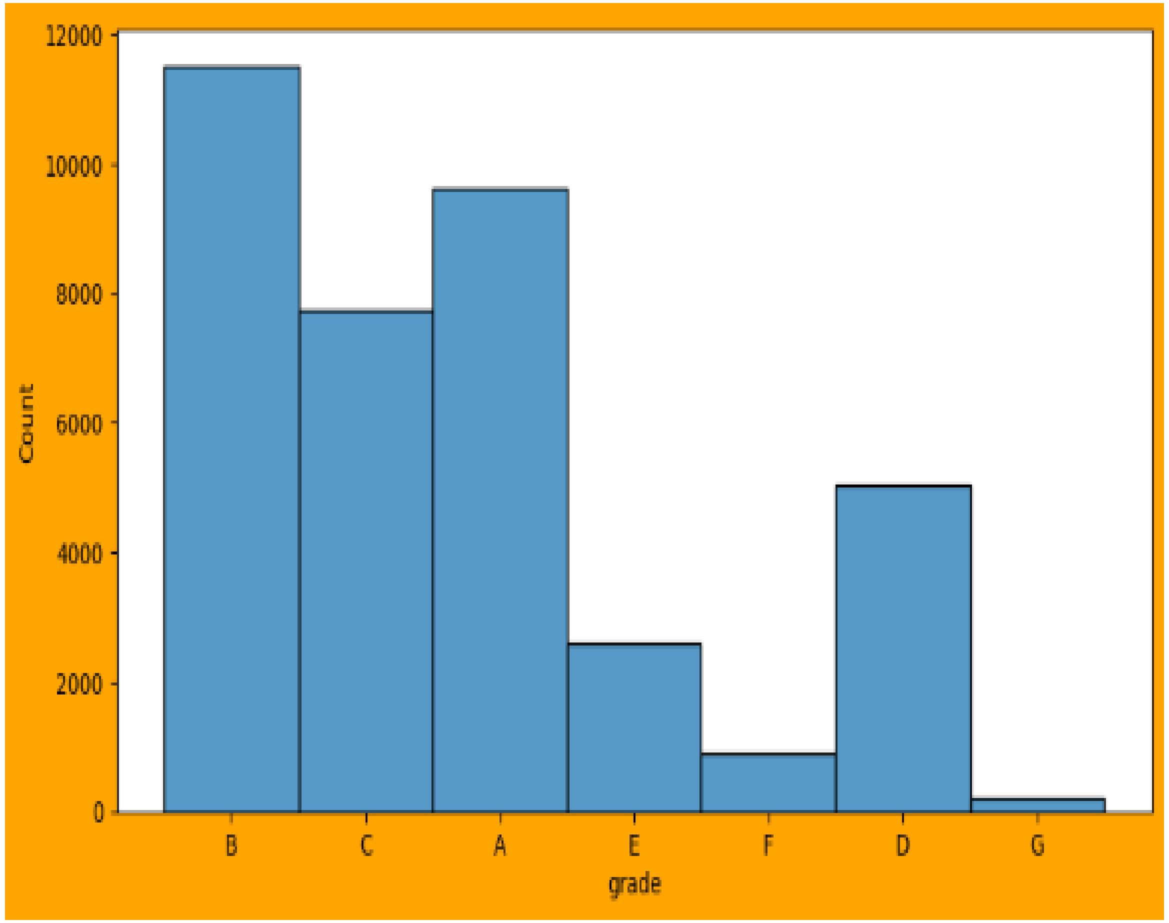
Analysis

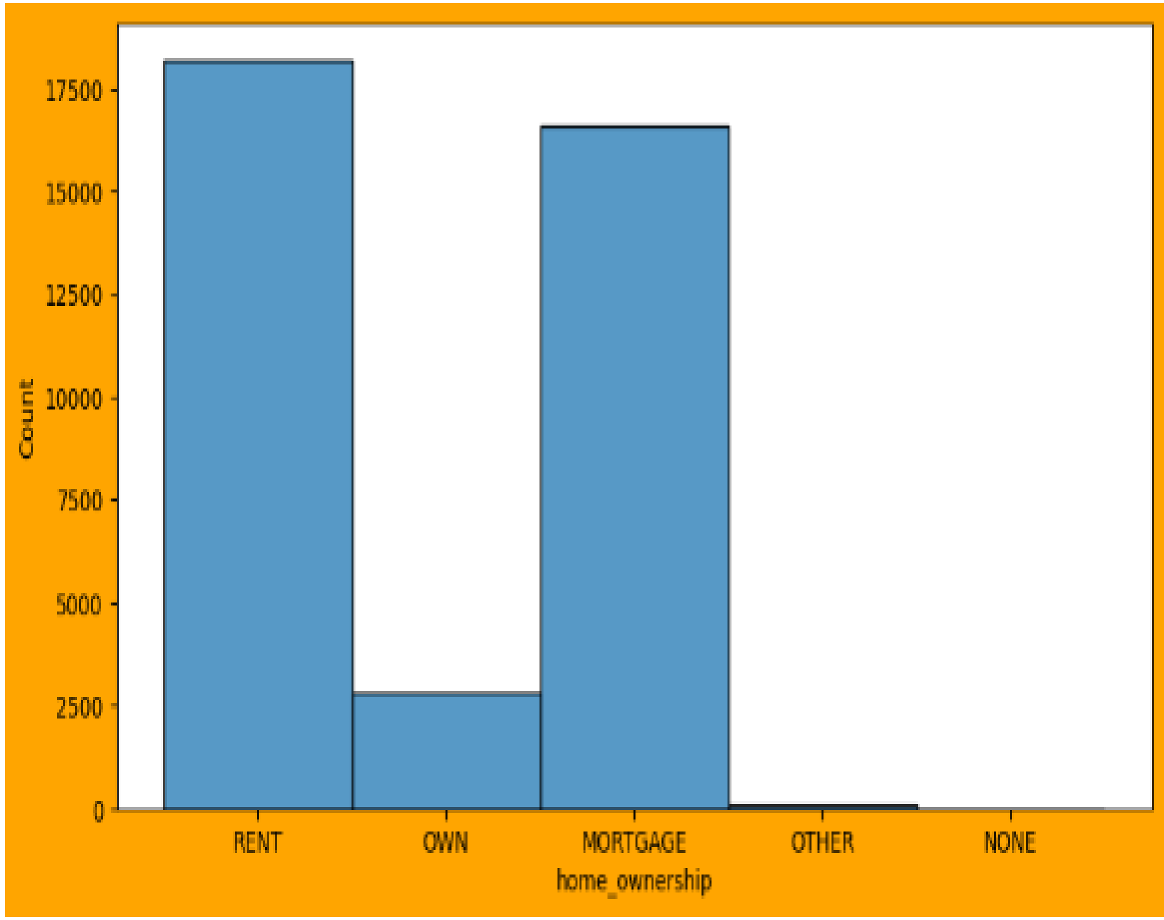
Univariate

(Categorical – unordered)

Bivariate

Multivariate





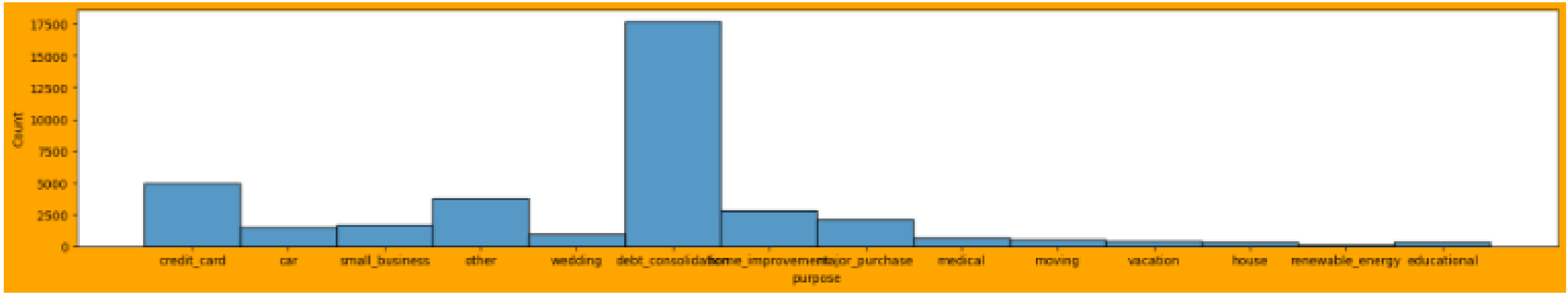
Analysis

Univariate

(Categorical – unordered)

Bivariate

Multivariate



Inference

1. **For higher loan amount the term is 36 months and loan amount <= 10000 the term is 60 months which is odd**

2. Employees with Grade B & A are highest loan seekers and there are more outliers in Grade A

3. People with Subgrades A4, B3 ,B4, B5 have taken loans

4. **The main purpose is for paying Rent or Mortgage which could be cause for becoming defaulters**

5. **The major purpose for loan request is debt_consolidation - which means this loan is taken to repay another loan**

6. **More number of people from CA has taken followed by NY. This is a key parameter to analyse and who has not repayed**

7. Employees with more experience have taken long term loans

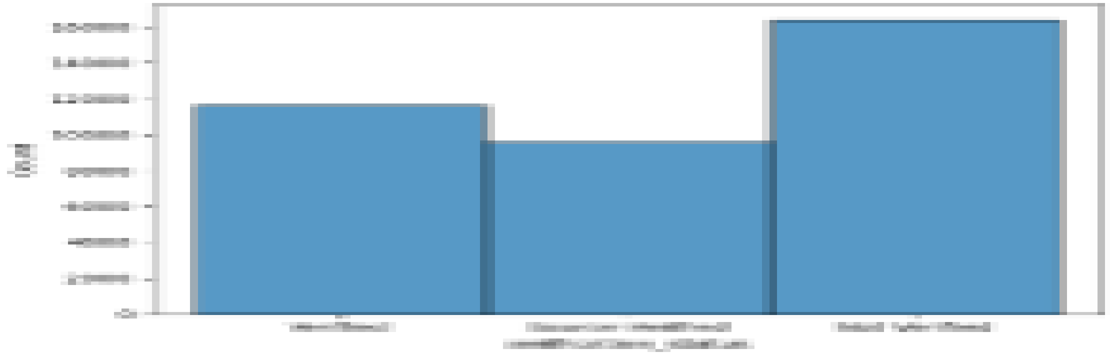
Univariate

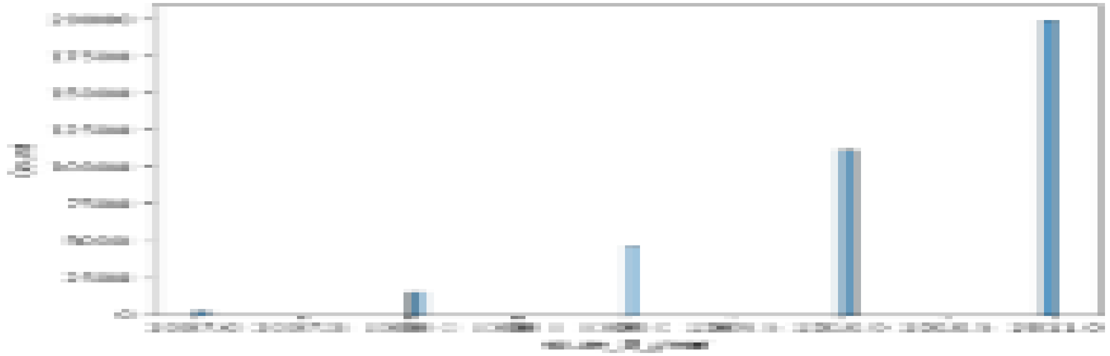
(Categorical – ordered)

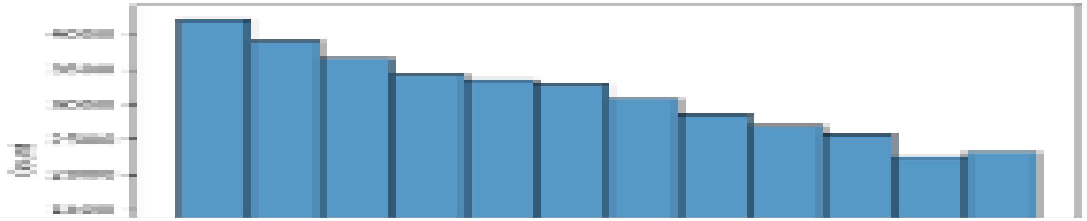
Bivariate

Multivariate

Analysis







Inference

1. There is around 50% of cases where income is not Verified, This is something the bank has to consider
2. For the Year 2011 the number of loans given is high and is increasing from 2007
3. Loans given during Dec month is high

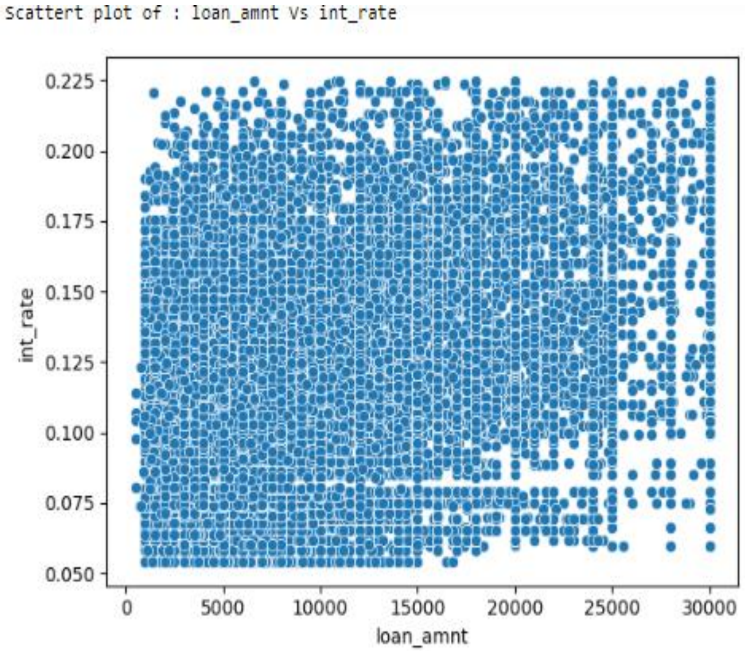
Univariate

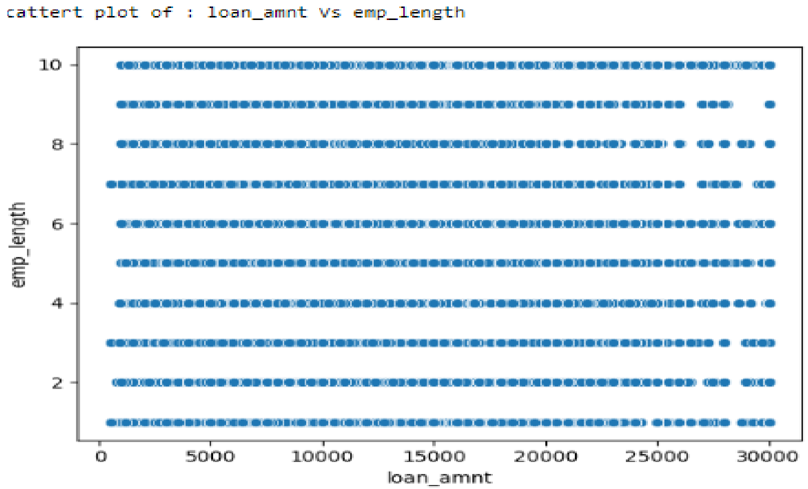
Bivariate

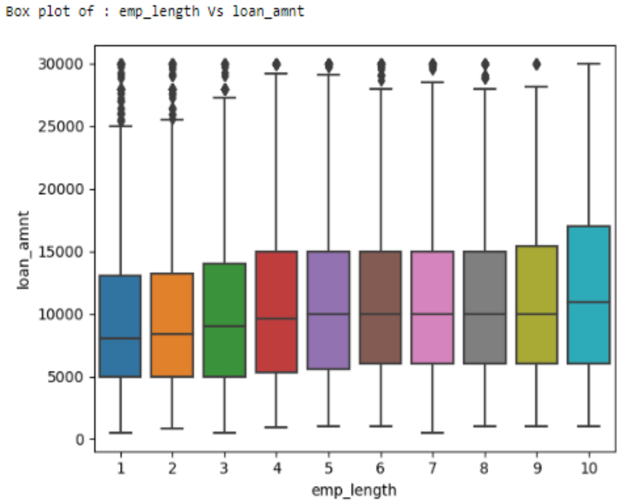
Numerical vs Numerical

Multivariate

Analysis







Inference

1. The higher the loan amount the interest is higher
2. Employees with less employee length like 1, 2, 3 years have taken loans and there are outliers with high loan_amnt

Analysis

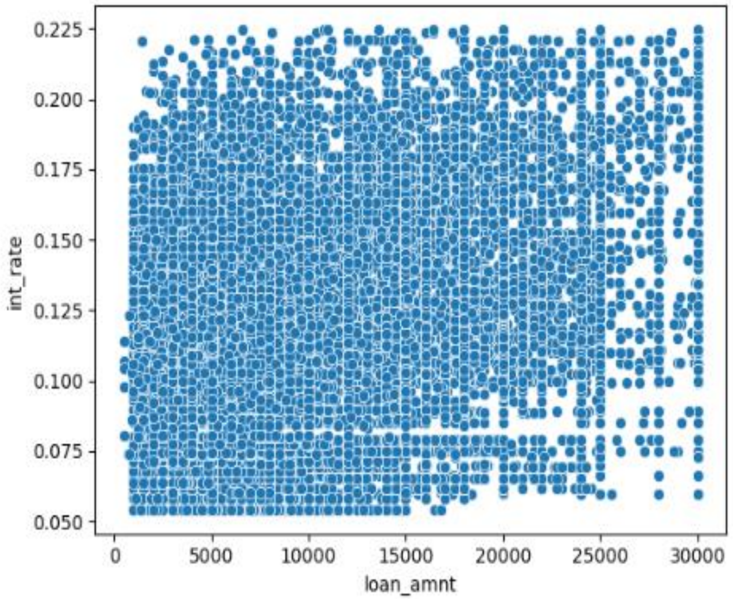
Univariate

Bivariate

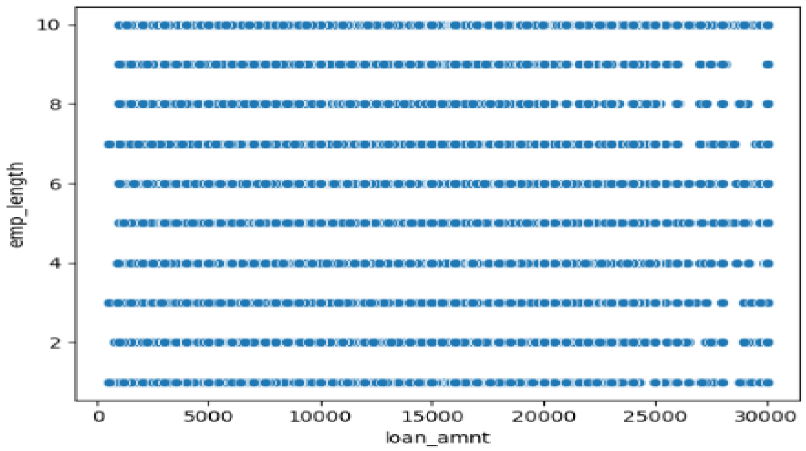
Numerical vs Numerical

Multivariate

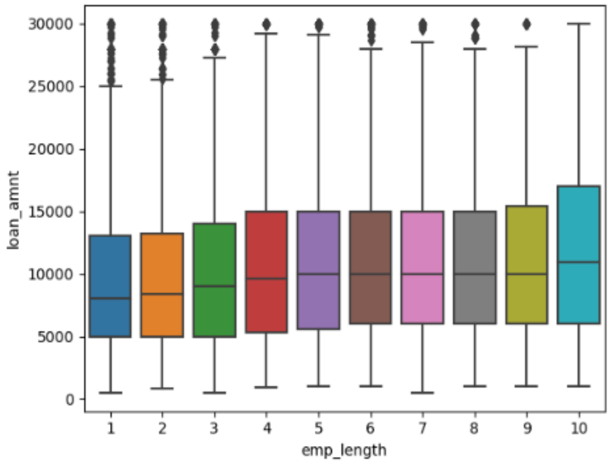
Scattert plot of : loan_amnt Vs int_rate



cattert plot of : loan_amnt vs emp_length



Box plot of : emp_length Vs loan_amnt



Inference

1. The higher the loan amount the interest is higher
2. Employees with less employee length like 1, 2, 3 years have taken loans and there are outliers with high loan_amnt – **There is high probability of being defaulters**

Univariate

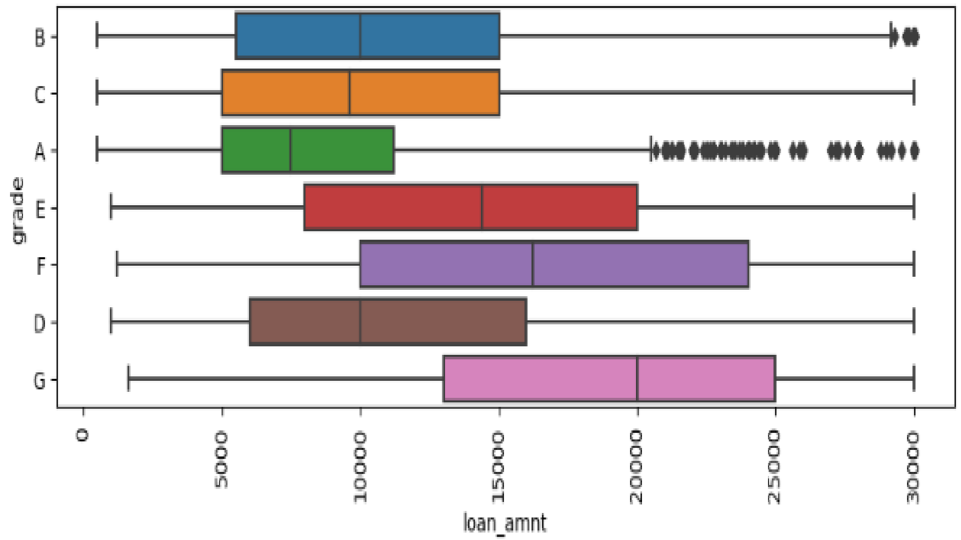
Bivariate

Numerical vs Categorical

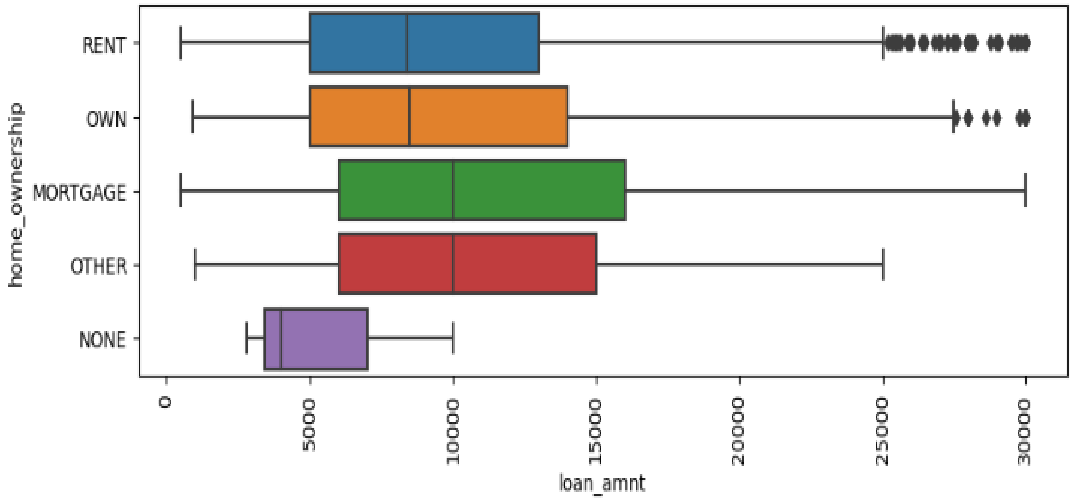
Multivariate

Analysis

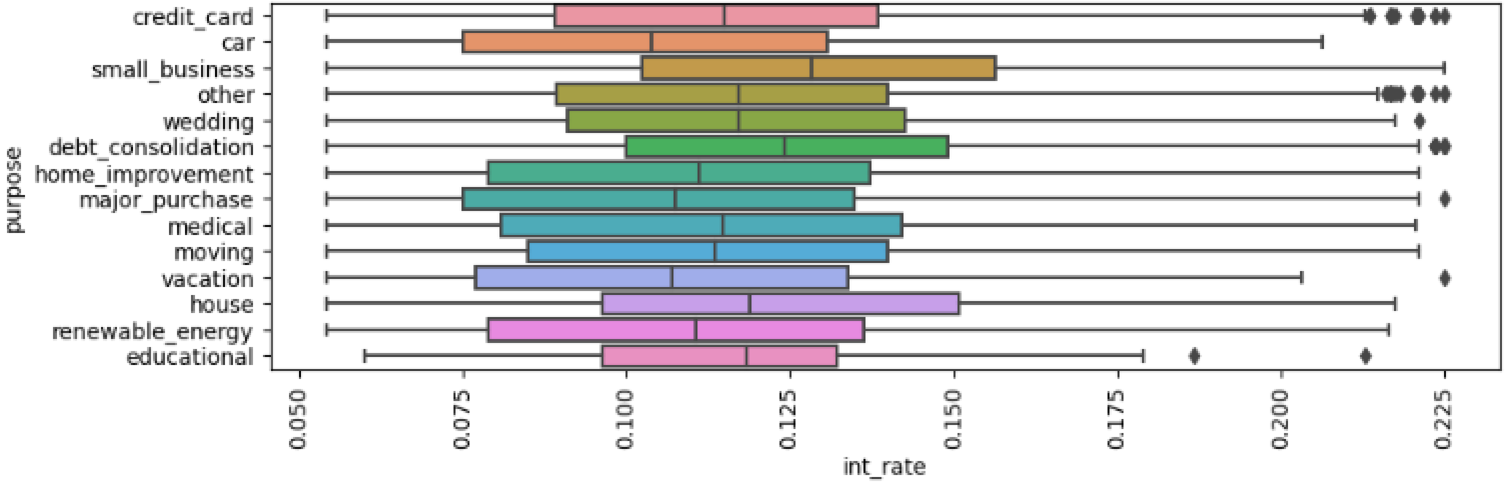
Box Plot of : loan_amnt Vs grade



Box Plot of : loan_amnt Vs home_ownership



Box Plot of : int_rate Vs purpose



Analysis

Univariate

Bivariate

Numerical vs Categorical

Multivariate

Observations

1.The higher the loan amount the term is also increased to 60 months
2.The majority of the loan lies around 10000
3.For loan_status of Fully Paid and Charged Off the loan_amnt median is close to 10000 where as in Current the median is 15000
4.Majority of the loans taken are with the purpose debt_consolidation -which means this loan is taken to repay another loan and small_business and second top reason is paying credit_card and house
5.The higher the interest_rate there a quite a few outliers with term of 36 months
6.Employees with Grade B & A are highest loan seekers and there are more outliers in Grade A
7.Employees with Grade F & G have higher int_rate
8.People with Subgrades A4, B3 ,B4, B5 have taken loans
9.Foe home_ownership of RENT, there outliers and the int_rate for all types is close to .125
10.For Source Not Verified the loan amount median is around 7000 and interest rate also seems to be less compared to other sources
11.More number of people from CA has taken followed by NY. This is a key parameter to analyze and who has not repayed
12.Employees with more experience have taken long term loans

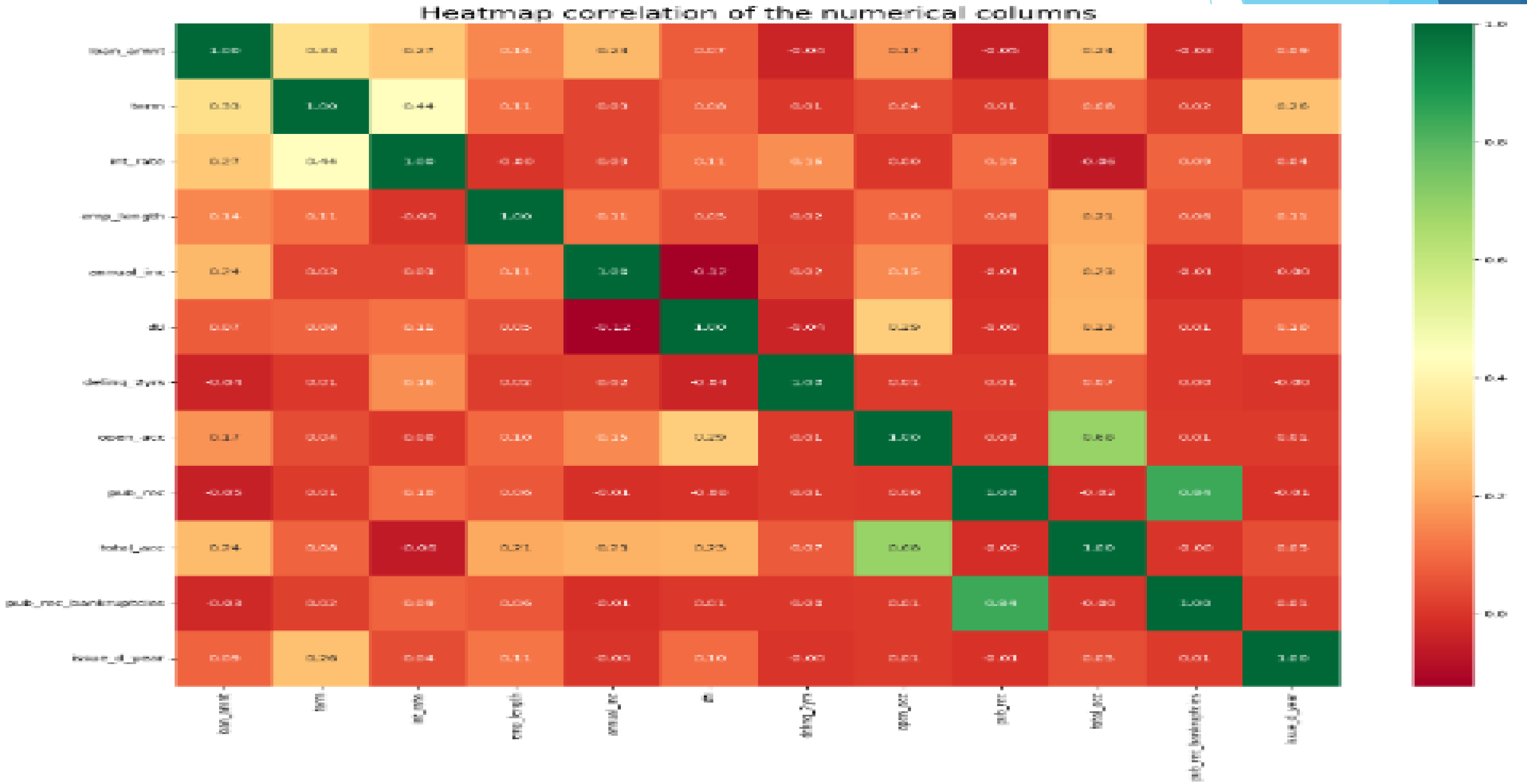
Univariate

Bivariate

Multivariate

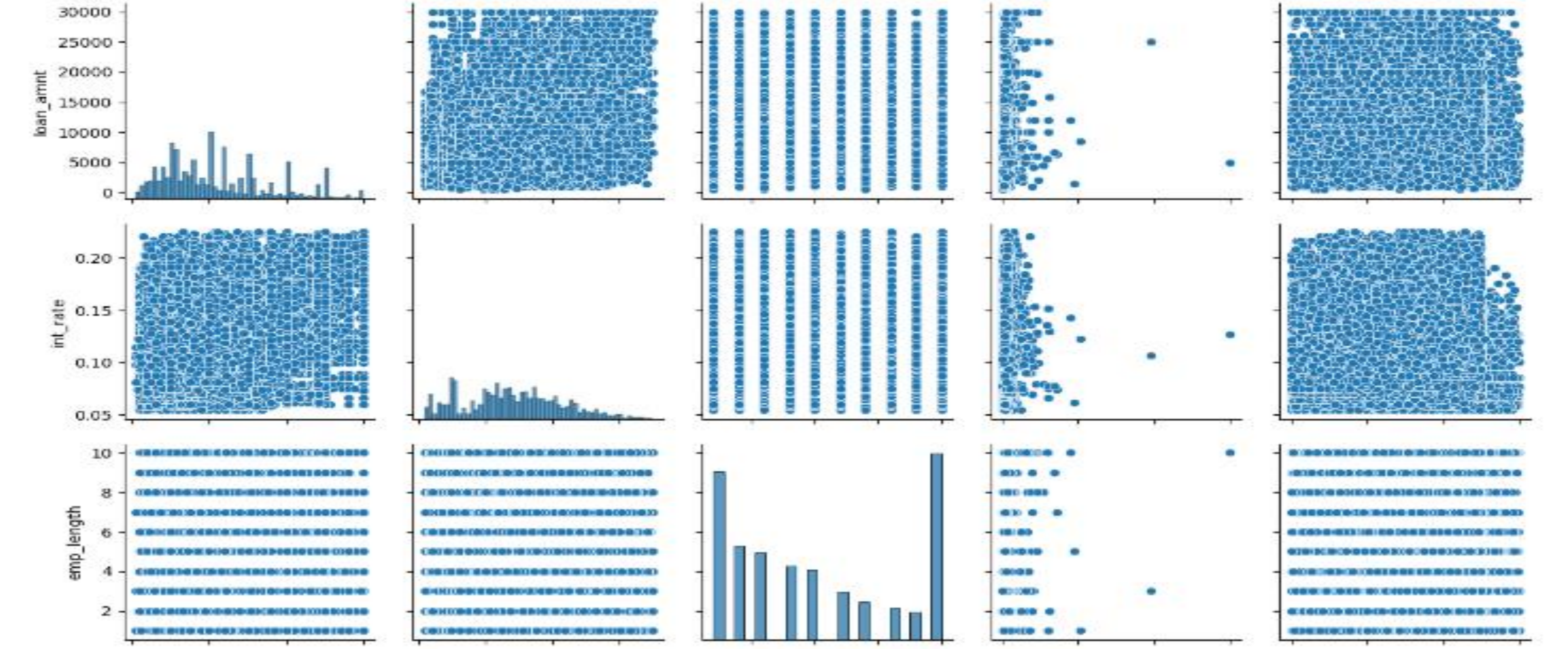
Correlation map

Analysis



Analysis

Univariate → Bivariate → Multivariate



1. Public record and public record bankruptcies are highly correlated
2. Annual income and debt to income ratio has negative correlation, meaning if one increases other decreases
3. Annual income have low correlation to the customers public record bankruptcies.
4. Based on the employment length, interest rates are not offered to the employees.
5. Customers with high interest rate have strong correlation with the loan term.

Final Prediction

The following parameters that could enable bank to look to avoid the risk of having more deaulters

1. **For higher loan amount the term is 36 months and loan amount <= 10000 the term is 60 months which is odd. Also employees with less emp_length have taken higher loan amount which they might end up In becoming a defaulter**
2. Employees with Grade B & A are highest loan seekers and there are more outliers in Grade A
3. People with Subgrades A4, B3 ,B4, B5 have taken loans
4. **The main purpose is for paying Rent or Mortgage which could be cause for becoming defaulters**
5. **The major purpose for loan request is debt_consolidation - which means this loan is taken to repay another loan. This parameter should be considered by bank before providing loan**
6. **More number of people from CA has taken followed by NY. This is a key parameter to analyse and who has not repayed**
7. **For Source Not Verified the loan amount median is around 7000 and interest rate also seems to be less compared to other sources**

THANK YOU