Lending Club Case Study

UpGrad Assignment : Using EDA

Content

- Problem Statement
- Data Cleaning
- Data Conversion
- Data Analysis
- Summary

Problem Statement

1. You work for 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.
2. When a person applies for a loan, there are two types of decisions that could be taken by the company:
   - **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:
     - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
     - Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
     - Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted other loan.
   - **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.).Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset).

Data Loading

- Data Loading and cleaning is the process of importing or reading data from various sources such as files (e.g., CSV, Excel). Loading data is the first step in any data analysis.
- Below steps were performed to load the data
- Different python libraries were imported

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        import datetime as dt
```

- The Data file named ‘loan.csv’ provided was loaded to Pandas data frame

  df_loan=pd.read_csv(loan.csv)

Data Cleaning

- Data cleaning, is the process of identifying and correcting errors, inconsistencies, and missing values in the dataset. This may include tasks such as removing duplicates, handling missing data, correcting data types, and standardizing formats to ensure the data is accurate and ready for analysis
- Total 39717 rows and 111 columns were present in the given data set.
- Data Cleaning included below steps :
  - Basic checks of header and footer
  - The columns that were not required for the current analysis were dropped.
    - Drop columns with unique value
    - Drop columns with null values in all rows and in more than 50% rows
    - Drop columns with unique single value in all rows
    - Drop columns that will not contribute for analysis
  - The rows that were not required for the current analysis were dropped.
    - Drop rows of Loan status as ‘Current’ as this attribute will not contribute for this analysis
    - Drop duplicate rows
    - Drop/Impute rows that has less than 5% data we can drop, otherwise we need to impute.

After following the data cleaning process final data set has 36847 rows and 21 columns for analysis.

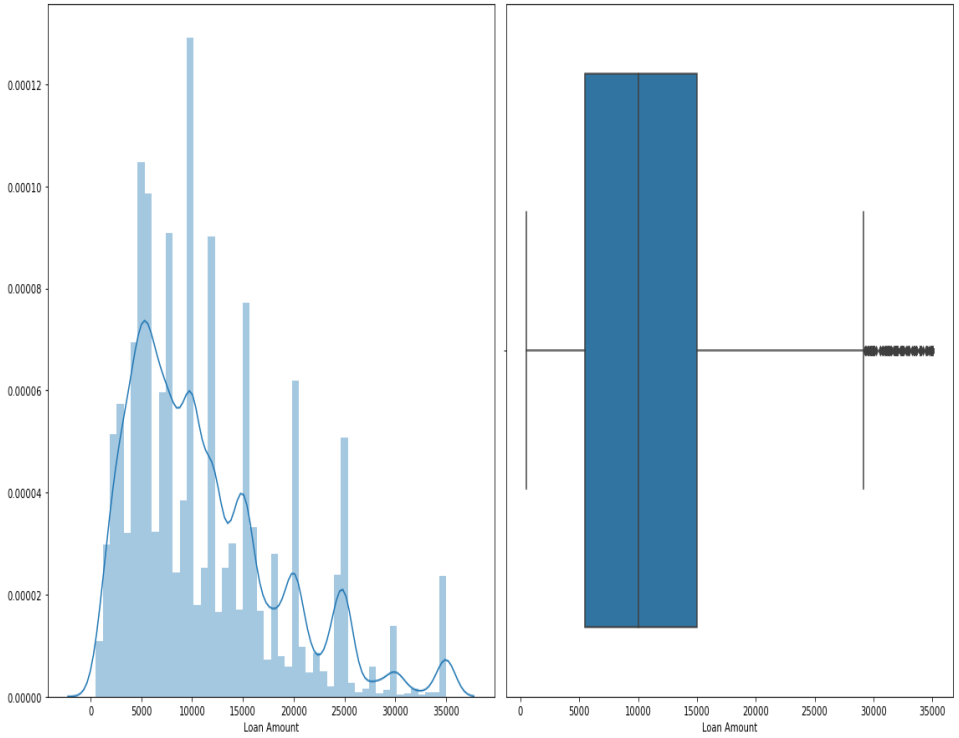
Data Conversion

- Data conversion refers to the process of converting data from one format, structure, or type to another. Data conversion is necessary to ensure compatibility between different systems or to prepare data for specific analysis.
- In this step data that are converted as:
  - Variable ‘term’ from obj type to int type
  - Variable issue_d from obj type to datetime type to year-month-date format where we can derive issue year and issue month for further analysis.
  - Converted int_rate from string type to int.
  - Converted loan_amnt and funded_amnt from int to float type
  - Converted emp_length from string to int type.

Data Analysis

- Data analysis is the process of interpreting data to uncover insights, patterns, trends, and relationships.
- Below are the analysis performed:
  - Univariate Analysis
    - Numerical Analysis
    - Categorical Analysis
  - Segmented Univariate Analysis
  - Bivariate Analysis

Data Analysis - Univariate Analysis with outliers

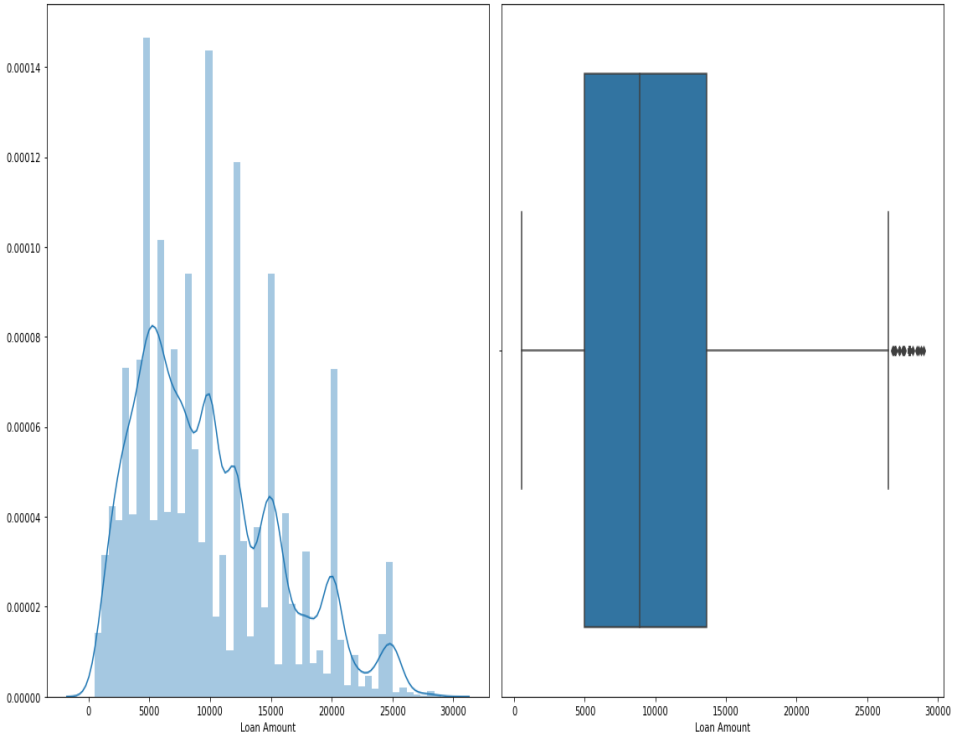



```
count 36847.000000
mean 11141.327652
std 7369.988994
min 500.000000
25% 5500.000000
50% 10000.000000
75% 15000.000000
max 35000.000000
```

From above univariate numeric variable analysis we can find out from description result as well as from plot that

- majority of loan amount applied was for range 5k to 15k(between range 25% to 75%)
- Min and max loan amount applied 500 and 35k respectively.
- From Box plot we can see the mean is around 1000 and there are some outliers as well.

Data Analysis - Univariate Analysis after removing outliers

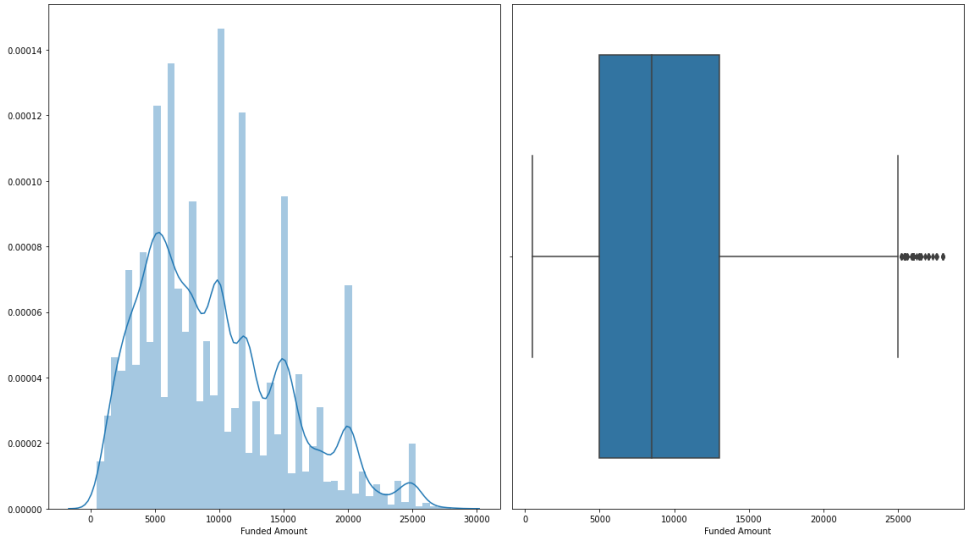



```
count 33123.000000
mean 9793.820759
std 5784.600495
min 500.000000
25% 5000.000000
50% 8875.000000
75% 13637.500000
max 29000.000000
```

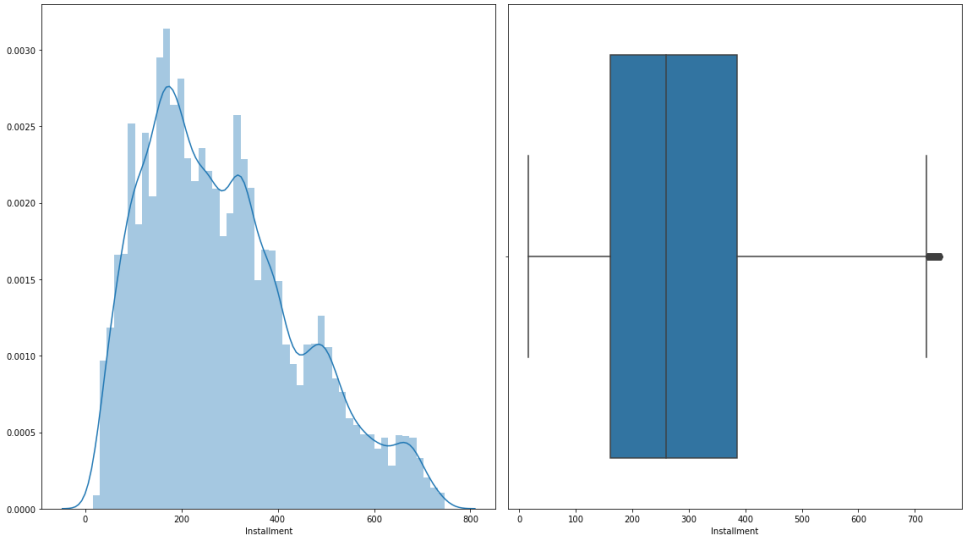
From above univariate numeric variable analysis we can find out from description result as well as from plot that

- majority of loan amount applied was for range 5k to 14k(between range 25% to 75%)
- Min and max loan amount were 500 and 29k respectively.

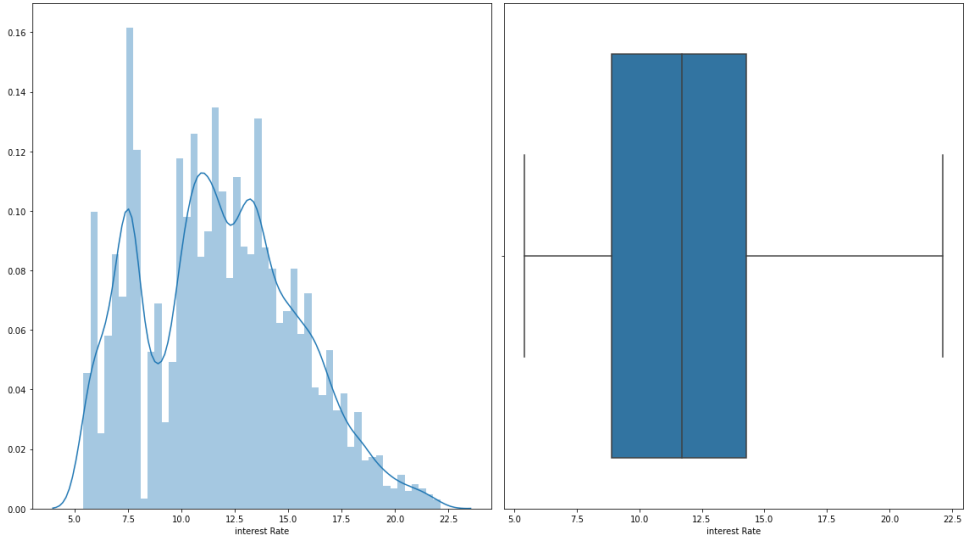
Data Analysis - Univariate Analysis of some more variable after removing outliers



Observation : Funded amount range is 5k – 13k (25% - 75%)

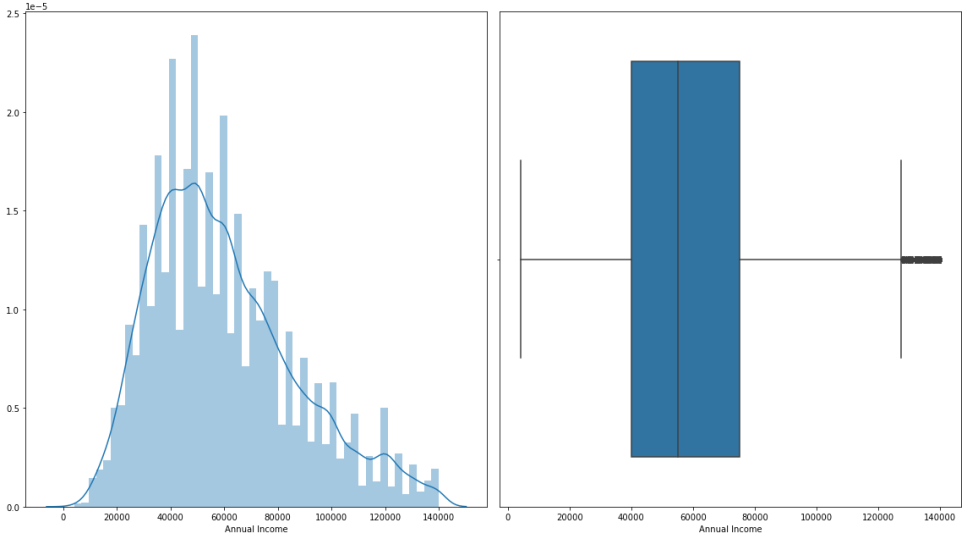


Observation : Common Installment range is from 160-385



Observation :

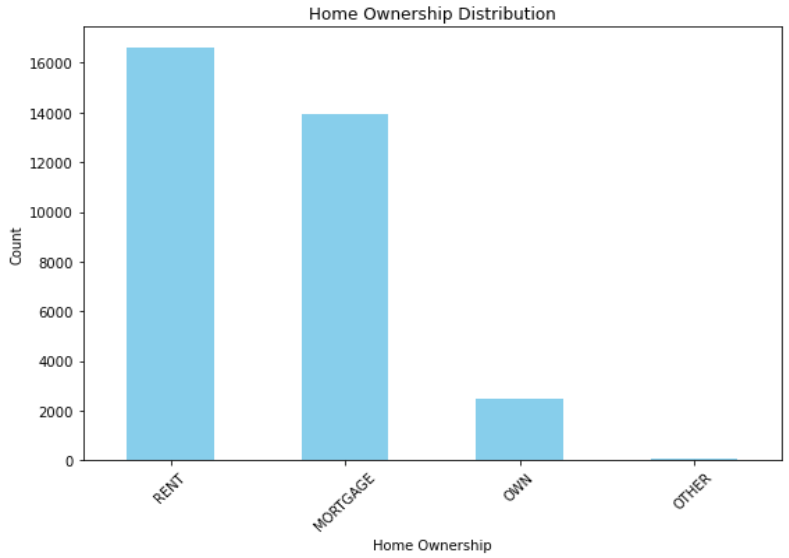
- Majority applicant's interest rate range is 8.9% to 14%
- Average interest rate is 11.8%



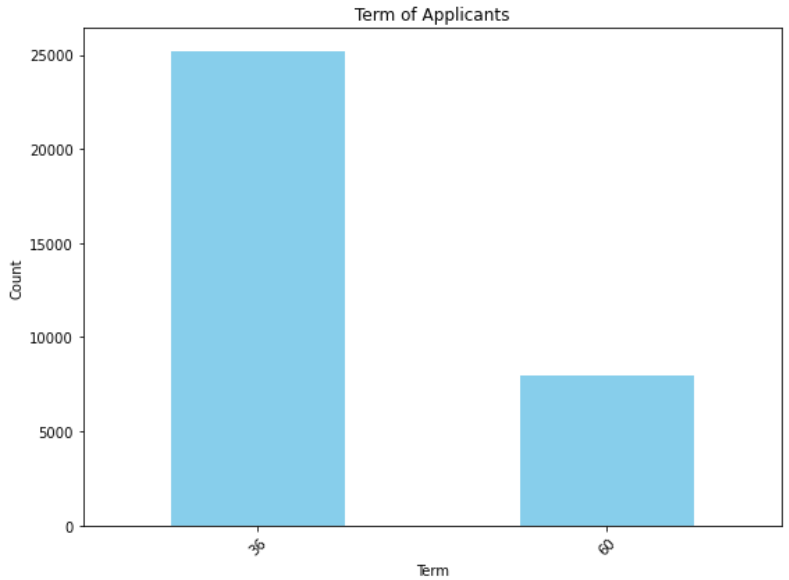
Observation :

- Applicants' Annual income is in normal distribution range of 40K-75K

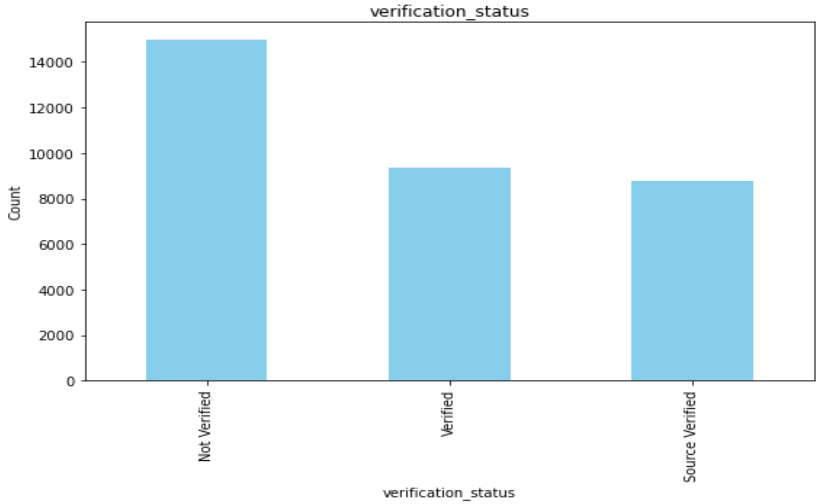
Categorical Univariate Analysis of some more variable after removing outliers



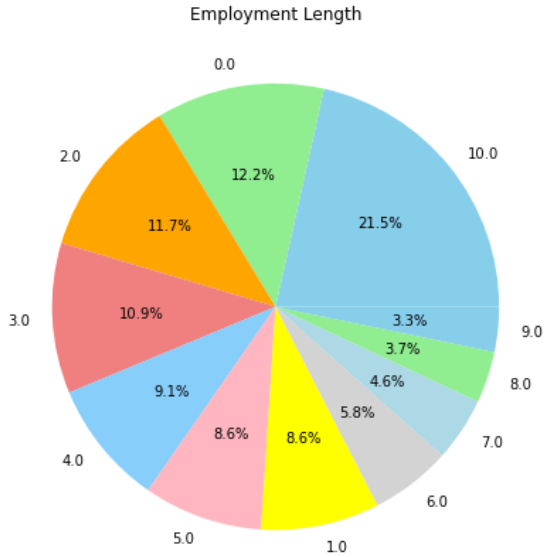
Observation : Maximum loans were applied by Rented applicants and Mortgage compared to who already owns home.



Observation : Maximum number of applicants have term 36 months.



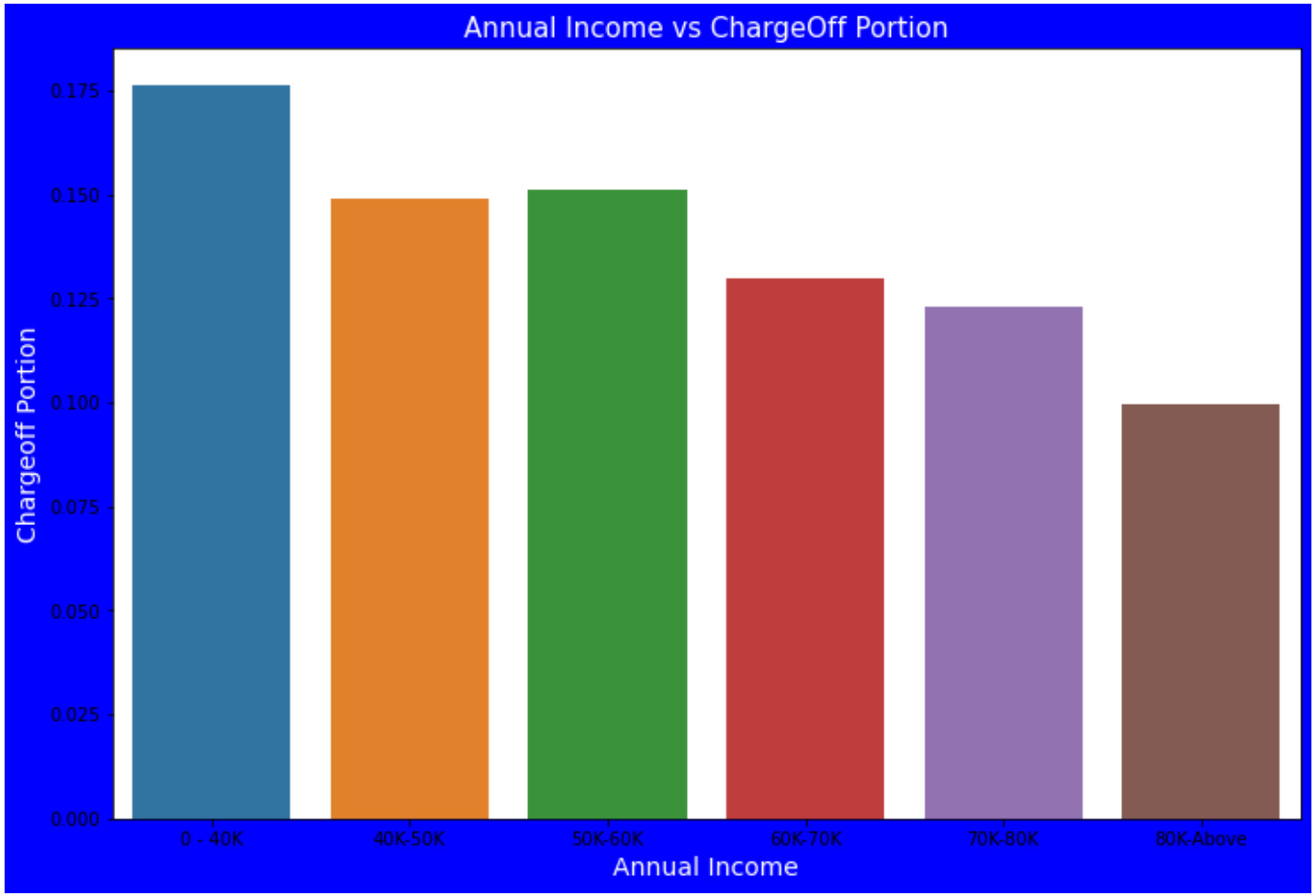
Observation : Maximum applicants loan approved are not verified and minimum is source verified



Observation : Majority applicants are 10+ employment experience.

Bivariate Analysis

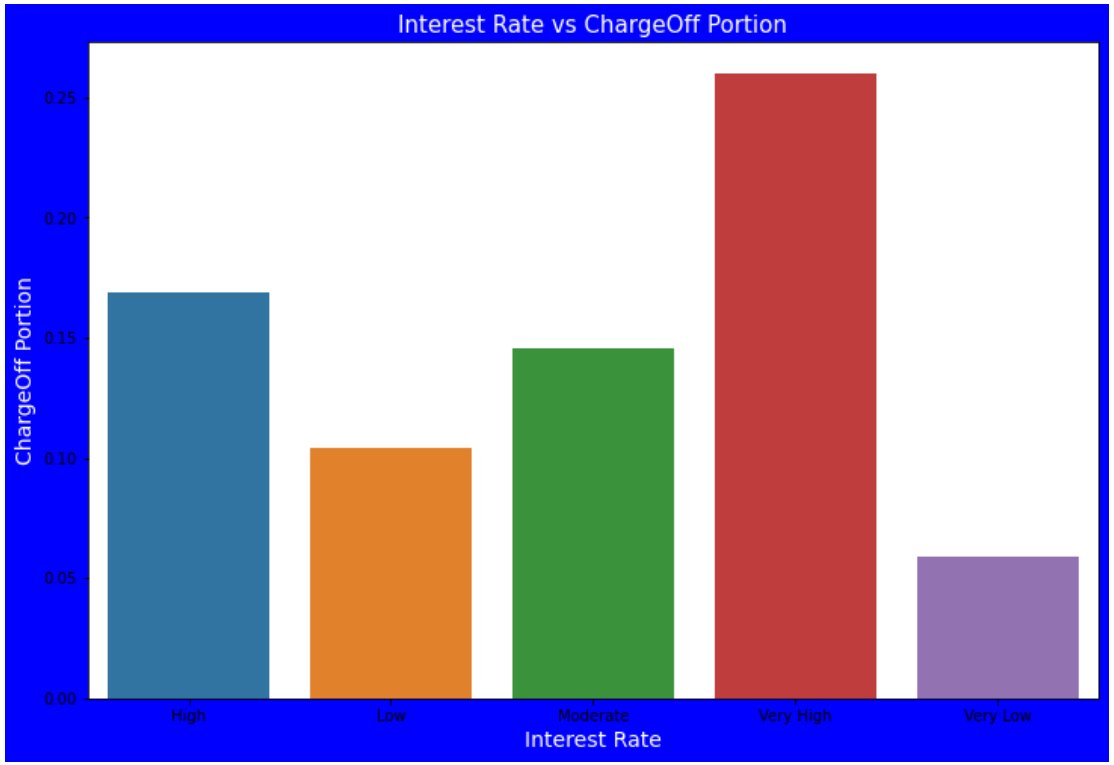
Using bivariate analysis we can compare , correlate with multiple variables. We can use derived variables or categorize and bucket into ranges. For this analysis I have bucketed few variables **loan_amnt_b, annual_inc_b, int_rate_b .**

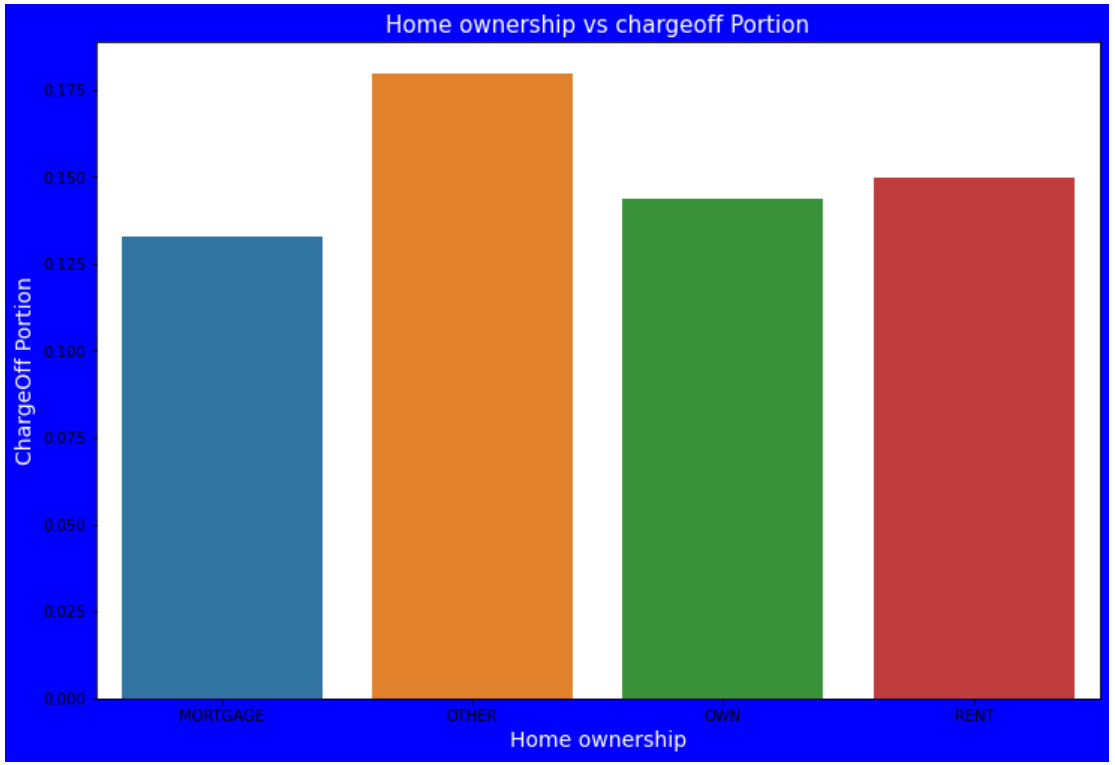


Observation on Analysis of annual income against chargeoff:

- Income range 80K and above has less chances to get charged off .
- Income range 0-40K has more chances to get charged off.
- So, with increase in income charge off portion decreases.

Bivariate Analysis





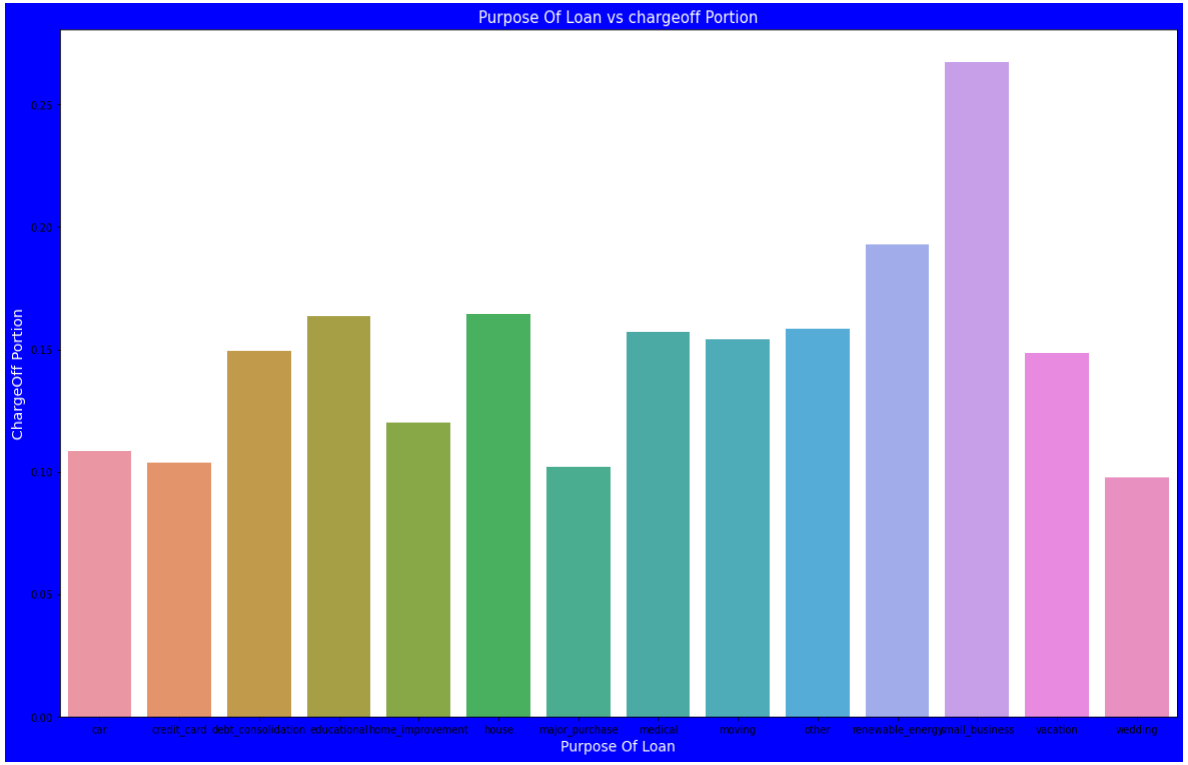
Observations:

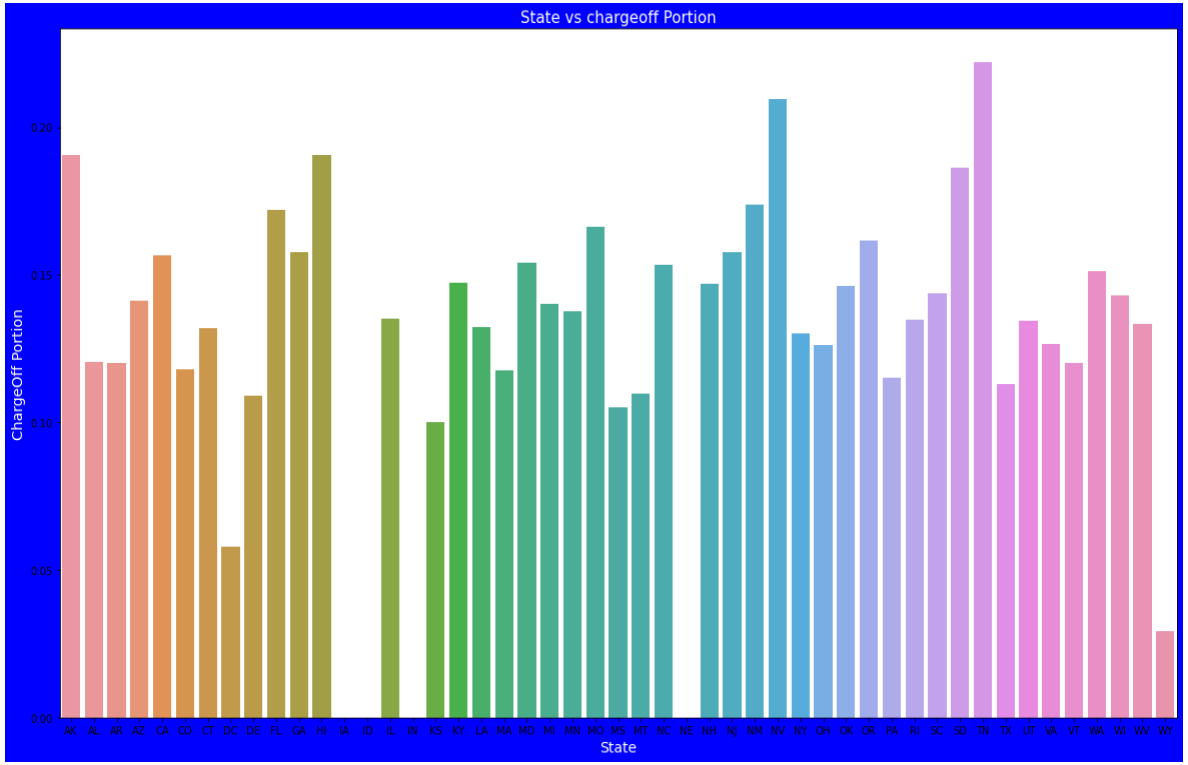
- Interest rate less than 10% or very low has very less chance of charged off.
- Interest rate more than 16% or very high has very high chance of charged off.
- With increase in interest rate charge off portion increases.

Observation :

- Applicants who owns house or on mortgage are having low chances of loan defaults.
- Applicants who doesn't own are on high chances of loan defaults.

Bivariate Analysis





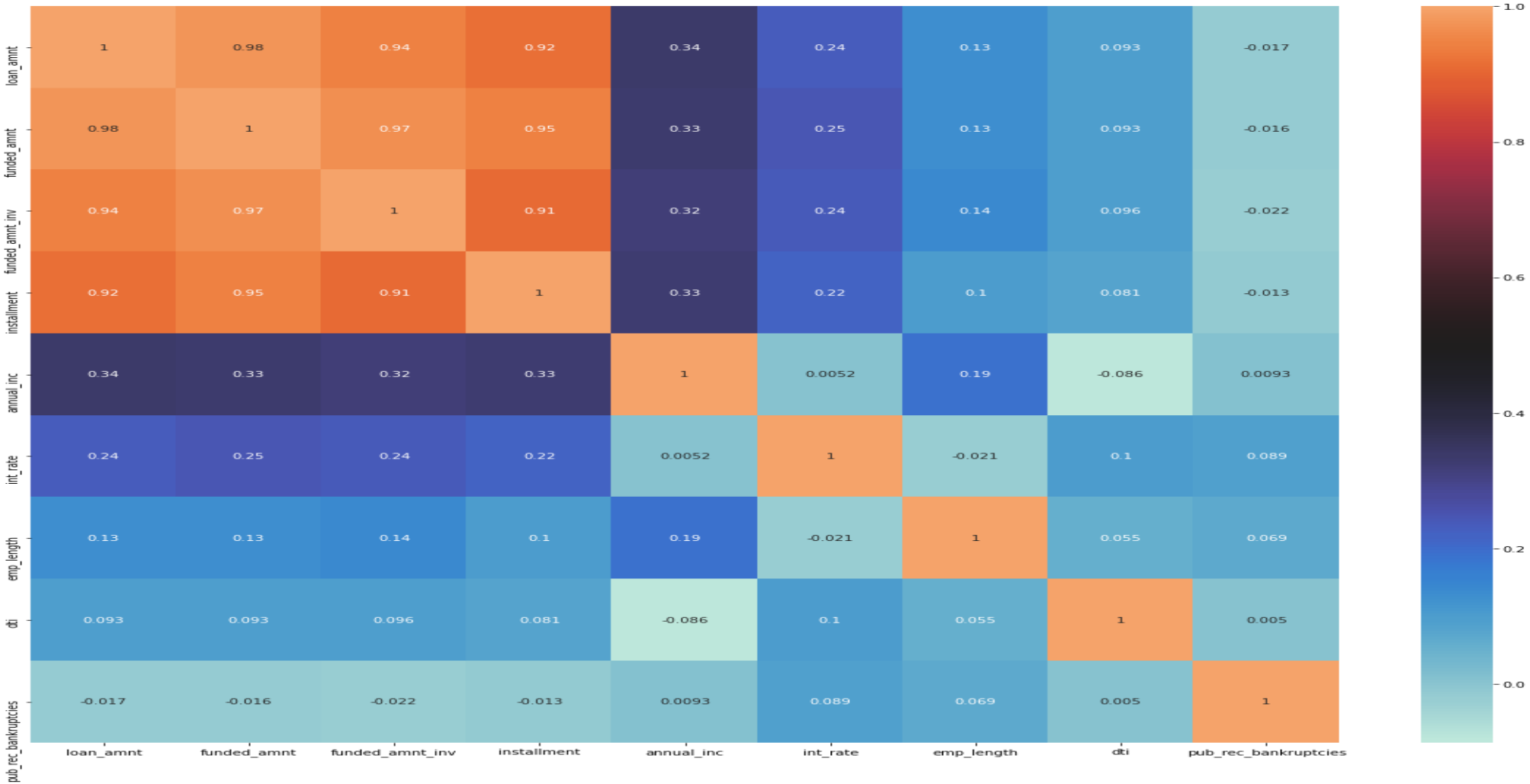
Observations:

- loan applied for purpose of Small business are highly charged off or in loan defaults.
- Wedding, credit card loan purposes are in low risk of loan defaults.

Observation :

- Maximum chargedoff were in state Tennessee followed by Nevada in US.
- Whereas the states NE,IA,ID,IN are with no loan defaults.

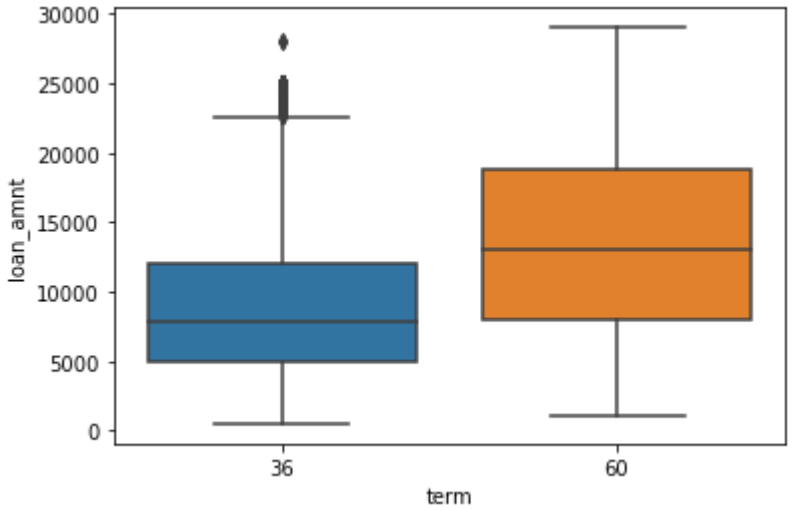
Bivariate Analysis



Observation :

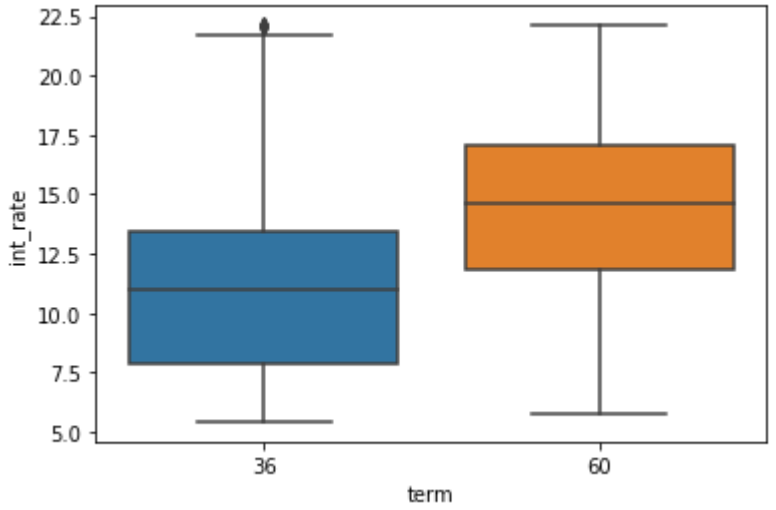
- From above correlation matrix using the variables loan amount, interest rate, installment, income, funded amount, dti and bankruptcy record, we can clearly see that
- Negative Correlation : with a bankruptcy record it is negatively or very poorly correlated to any other variable like loan amount, installment, etc.variable Dti is -vely correlated to other variables.
- Strong Correlation:variable 'funded_amount' is higly correlated to loan_amount by factor .98 or most likely to get the loan approved .Infact funded_amount shares high correlation factor .95 with funded_amnt_investors and with installments.

Bivariate Analysis



Observations:

- This above boxplot correlation shows if term is increasing the loan amount is increasing.



Observation :

- Same with term and interest rate correlation with high term interest rate is increasing and vice versa

Summary of the analysis

- The findings from the analysis are that
  - Risky Applicants
    - Applicants with high bankruptcy records
    - Applicants with Small Business
    - Applicants with less employment length
    - Applicants with Annual income less than 40K.
  - Favorable Loan Conditions
    - Loans with less interest rate
    - Loans with low term (Time period)
    - Loans with high Grade.

Thank You