Dated: 03rd Sep 2023

Lending Club Case Study

1

Name: DEBASISH DEATY

Objectives

The Objective of this case study is to implement the EDA technique on a real-world problem understand the insights and present in a business-first manner via presentation.

Benefits of the case study:

- Gives an idea about how EDA is used in real-life business problems.
- It also develops a basic understanding of risk analytics in banking and financial services.
- How the data is used to minimize loss of money while lending it to clients.
- It improves our understating of visualization and what charts to use for real life data.

Context:

Lending Club wants to understand the **driving factors** behind loan default or non-default which are strong parameters of default. The company can utilize this knowledge for its portfolio and risk assessment in terms of loan issues.

Problem Statement:

When the company receives a loan application, the company must decide 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.

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

Business Understanding

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)

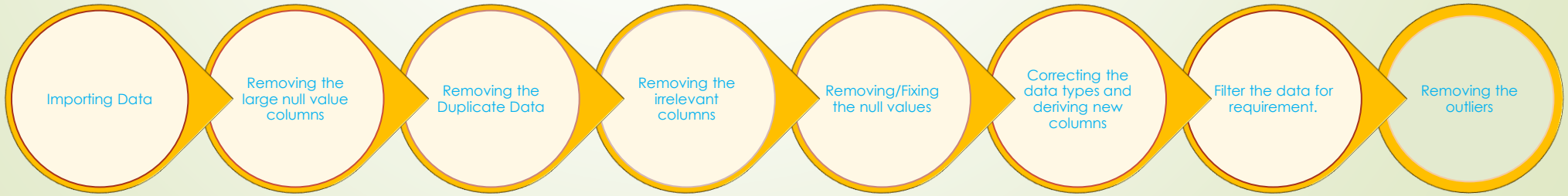
Current: Applicant is in the process of paying the installments, i.e., the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.

Charged-off: Applicant has not paid the installments in due time for a long period of time, i.e., he/she has defaulted on the loan.

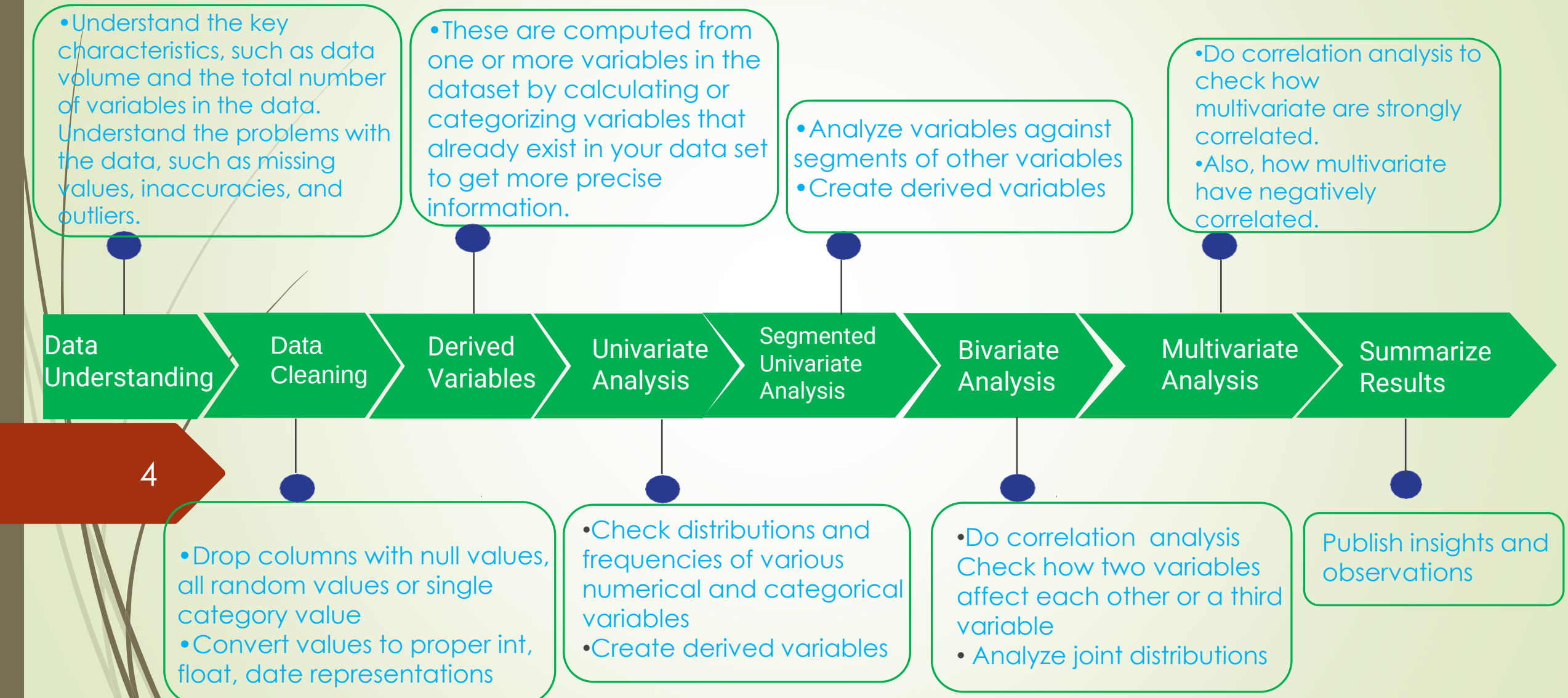
Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labeled as 'charged-off' are the 'defaulters'.

The company wants to understand the driving factors behind loan default, i.e., the variables that are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Data Clean-up and preparation process



Analysis Approach & Problem-Solving Methodology



Data Understanding

- The dataset has 39717 rows and 111 columns.
- There are many columns that consist of null values and NAN values.
- The dataset has a mixture of categorical and continuous data.
- Some of the columns only consist of NAN values.
- There are some columns that have string values also.
- The columns in the dataset have duplicate values.
- Our dataset needs a good amount of pre-processing.

Data Cleaning

- **Data Cleaning:**

  Removing null valued columns, single unique columns, unnecessary columns then manipulation of data such as conversion of data types, removing outliers, deriving new variables, and many more.

  - In the first step we have dropped all the duplicate values in the dataset.
  - We have removed the columns that have the columns with single occurrence values.
  - Then we drop the columns where all values are NULL.
  - Then we have dropped the unnecessary columns like “id” and “URL” etc.
  - Also removing columns that have more than 10000 null values in them.
  - Storing only the columns that have a unique value equal to 1 in single_unique
  - Dropping all the columns that have unique values equals to 1
  - After analysis we can see the null value and unique value of **emp_title** is more, so we can drop this column as it will disturb our analysis.
  - There are some **customer behavior variables** that are not available at the time of loan application, and thus they cannot be used as a prediction of credit approval therefore we can remove those variables from our dataset.
  - For loan status the "**Current**" value doesn't give any information for approving or rejecting loan applications. So, we can drop the rows having value current.
  - Analysis that there are empty space at the start of the **term** values, so remove it now.
  - Now our dataset is ready for data analysis.

Data Analysis

➢ **Data Analysis:**

The right problem is solved which is coherent with the needs of the business. The analysis has a clear structure, and the flow is easy to understand.

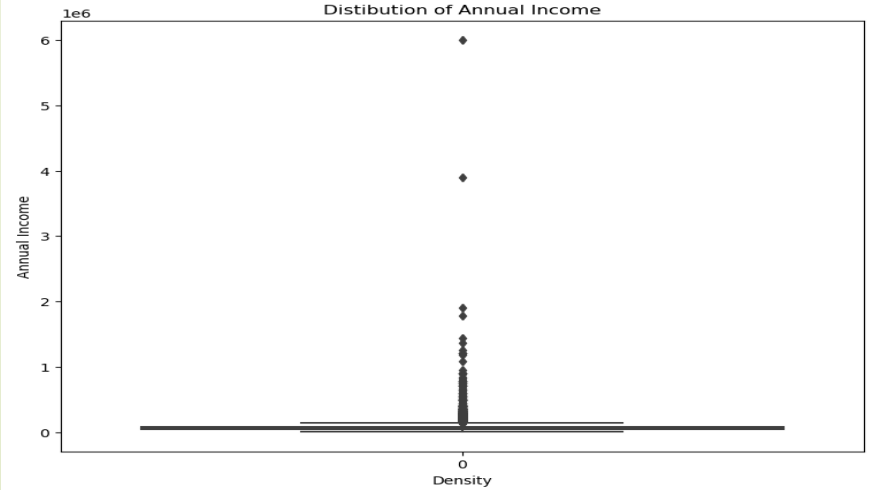
- For further analysis we will do the **Univariate and segmented univariate analysis**. Once it is done correctly and appropriate realistic assumptions are made wherever required. Also, will analyze to identify some of the important driver variables (i.e., variables that are strong indicators of default).
- Will create Business-driven, type-driven, and data-driven metrics for the important variables and utilize them for analysis. For further analysis, the explanation for creating the **derived metrics** is mentioned.
- Further the **Bivariate analysis** is performed to identify the important combinations of driver variables so that they make business or analytical sense.
- Further will perform the **Multivariate Analysis**, Through this we can analyze more than two variables and determine the correlation matrix between them.
- 'll make sure that the most useful insights are explained correctly in the comments.
- Will ensure that the appropriate plots are created to present the results of the analysis. The choice of plots for respective cases is correct. The plots should clearly present the relevant insights and should be easy to read. The axes and important data points are labeled correctly.
- Now our dataset is ready for further data analysis.

Data Manipulation

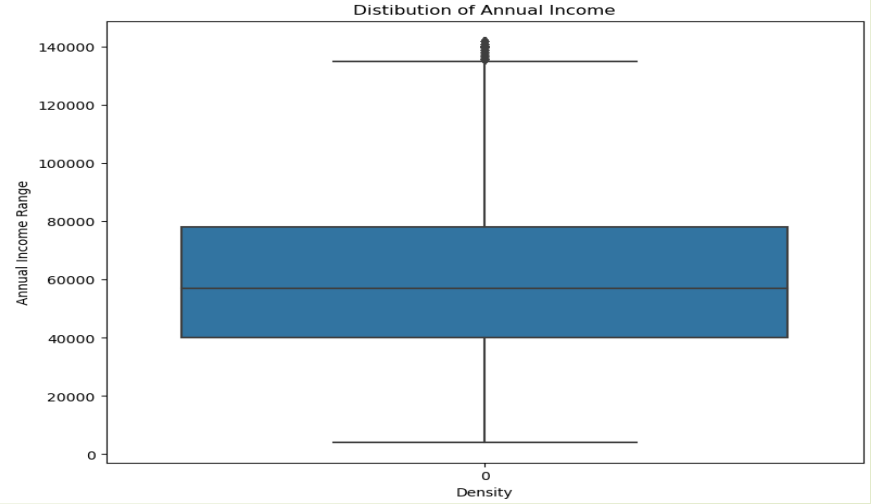
- **Data Manipulation:**
  Understanding the process of organizing or arranging data in order to make it easier to interpret and make reading or interpreting the insights from the data more structured and comprises of having better design.

  - Observe that the column 'int_rate' have "%" because of which it is showing data type - Object. So we can remove "%" and convert it into float data type. We have removed the columns that have the columns with single occurrence values.
  - Analyze outliers in annual_inc. So let us remove the values after the 95 percentile.

First analyze outliers in annual_inc



Then analyze after Removing the values after 95 percentile



Derived Variables

- **Derived Variables:**

  Derived variables are variables that are computed from one or more variables in the dataset. They are created by calculating or categorizing variables that already exist in your data set.

  - Creating new columns 'month' and 'year' from 'issue_d' column.
  - Approved the loan amount percentage with the help of 'funded_amnt_inv' and 'loan_amnt' by creating new column 'approved_loan_amt_percent'.
  - Binning the data to analyze more efficiently.
  - Changing the 'loan_status' to a numeric variable 'loan_status_count', assign 1 for defaulted loans and 0 for paid-off ones.
  - As we were done with cleaning and manipulating the data, derived new variables for our analysis and removed the outliers. Now, we can start analyzing the data.

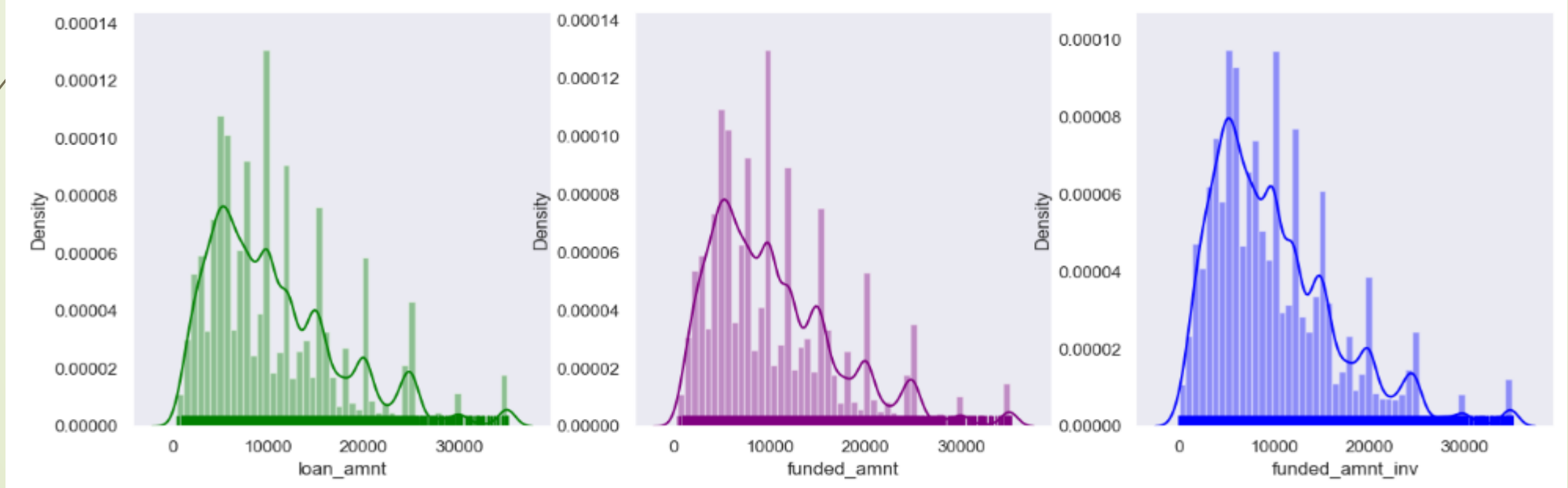
Univariate Analysis

- **Univariate Analysis:**

  Analyzing each column and plotting the distribution of each to get more information.

  Quantitative Variables:

  - Let's see the distribution of loan amount, funded amount, and funded amount by the investor using a distribution plot.

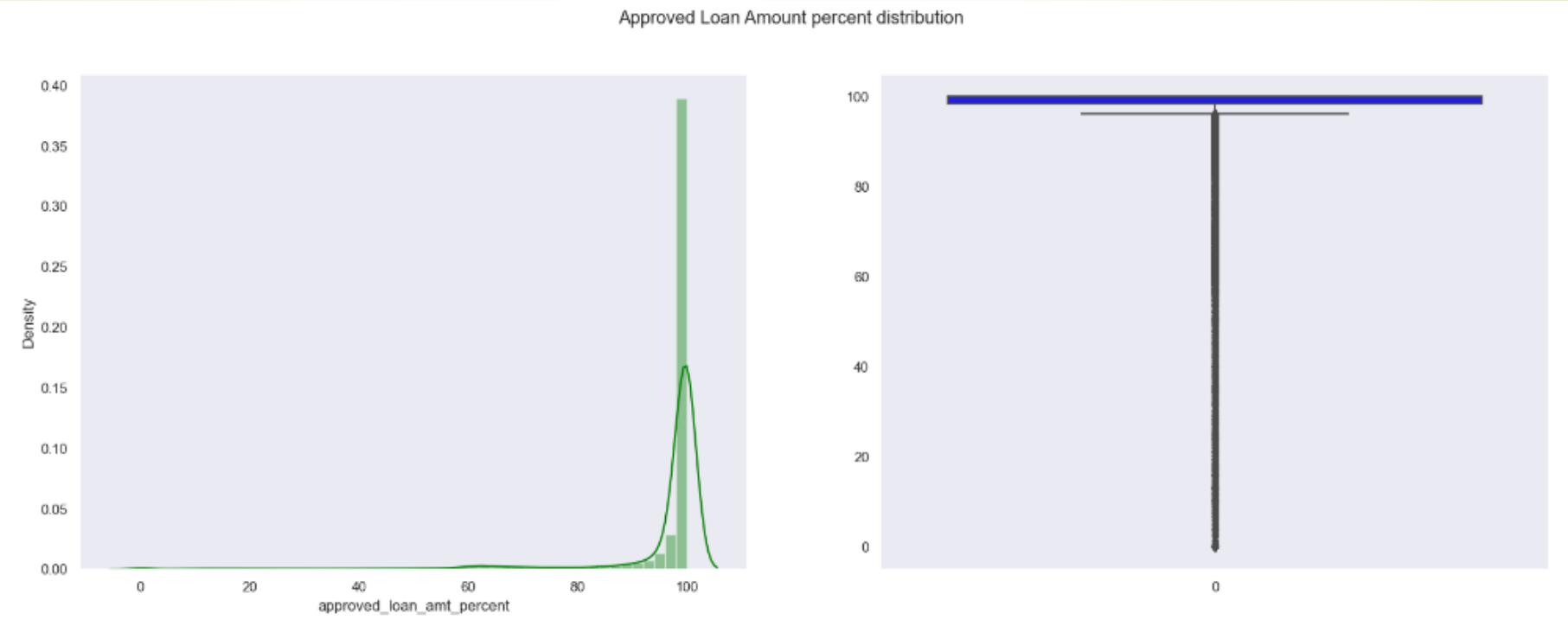


Observations:

Above, we can say that the amount distribution looks very similar. So, we can use loan_amnt for our further analysis.

Univariate Analysis

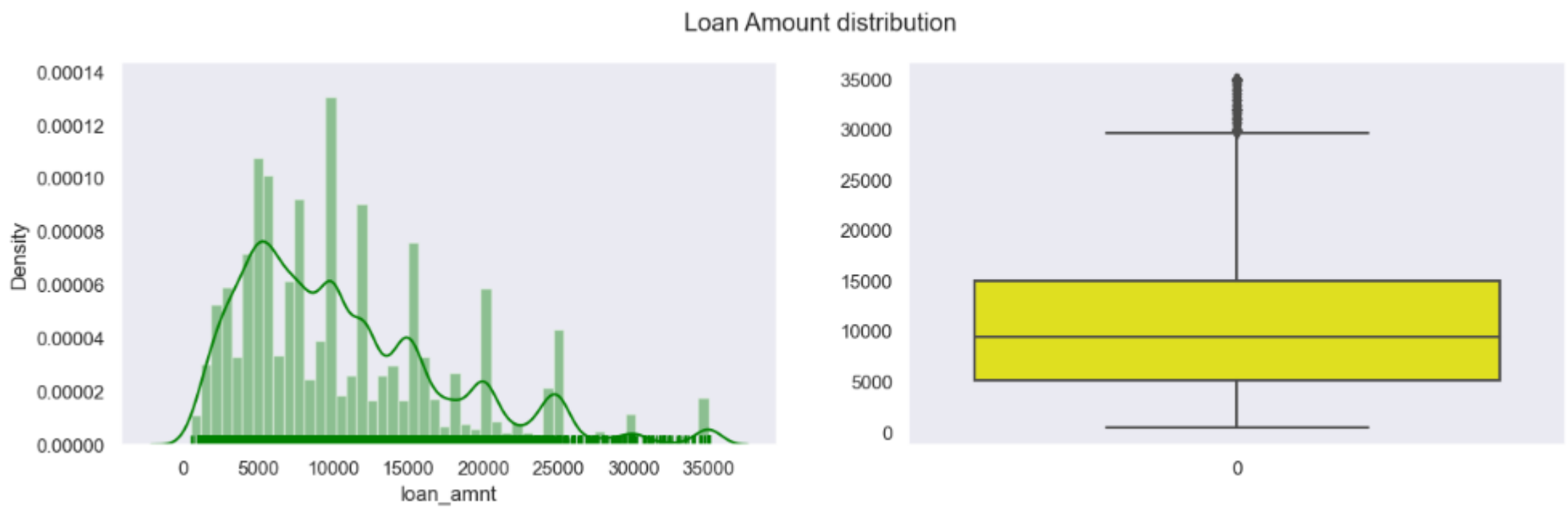
- **Analyzing Approved Loan Amount percentage.**



Observations:
80% of Borrowers got 100% loan amount from investors.

Univariate Analysis

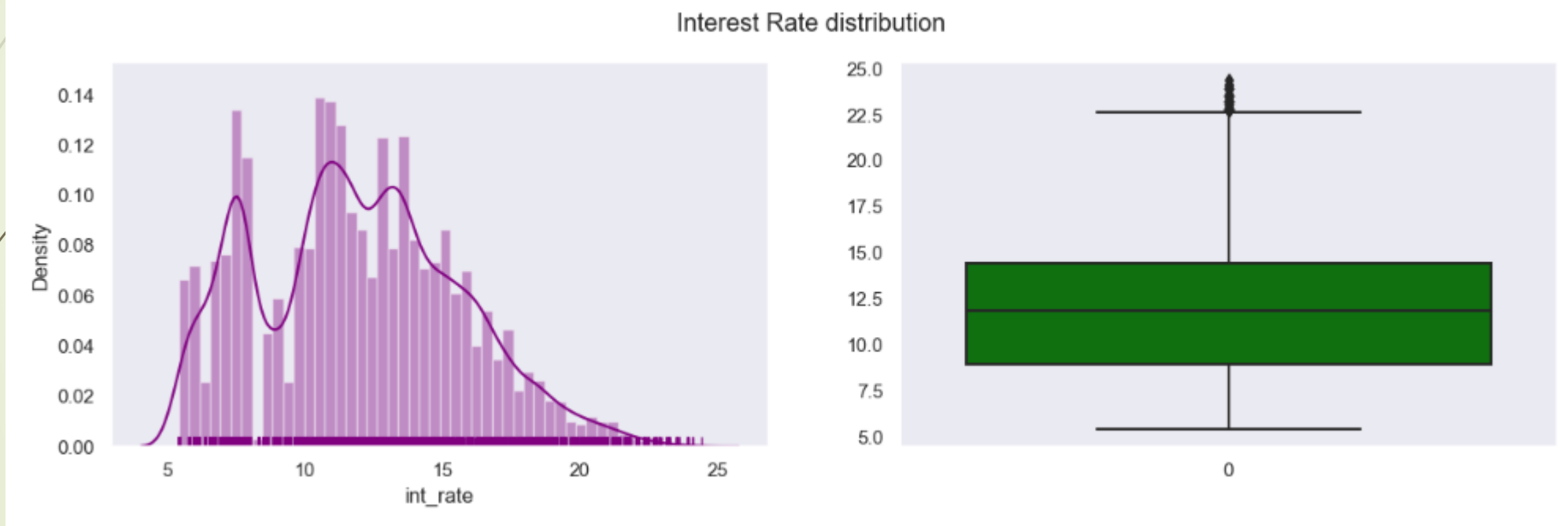
- **Analyzing Loan amount:**



Observations:
From the above loan amount data, we can say that most of them have taken their loan between 5000 and 15000

Univariate Analysis

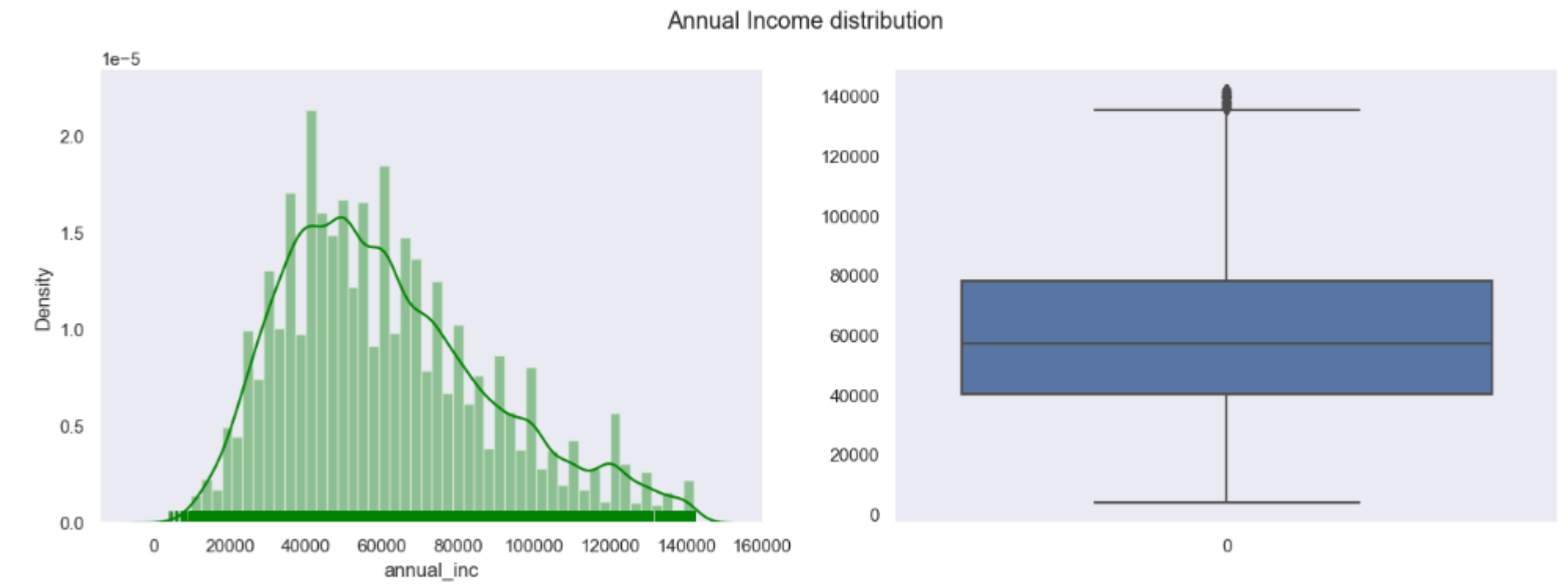
- **Analyzing Interest Rate:**



Observations:
From the above interest rate data, we can say that most of the interest rate lies between 9% to 14.5%.

Univariate Analysis

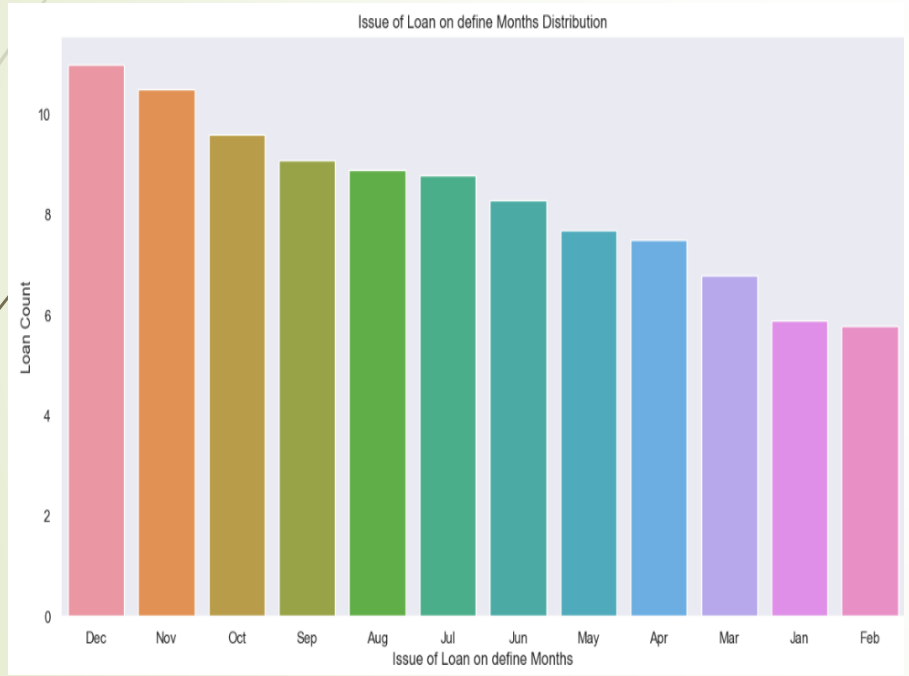
- **Analyzing Annual Income:**

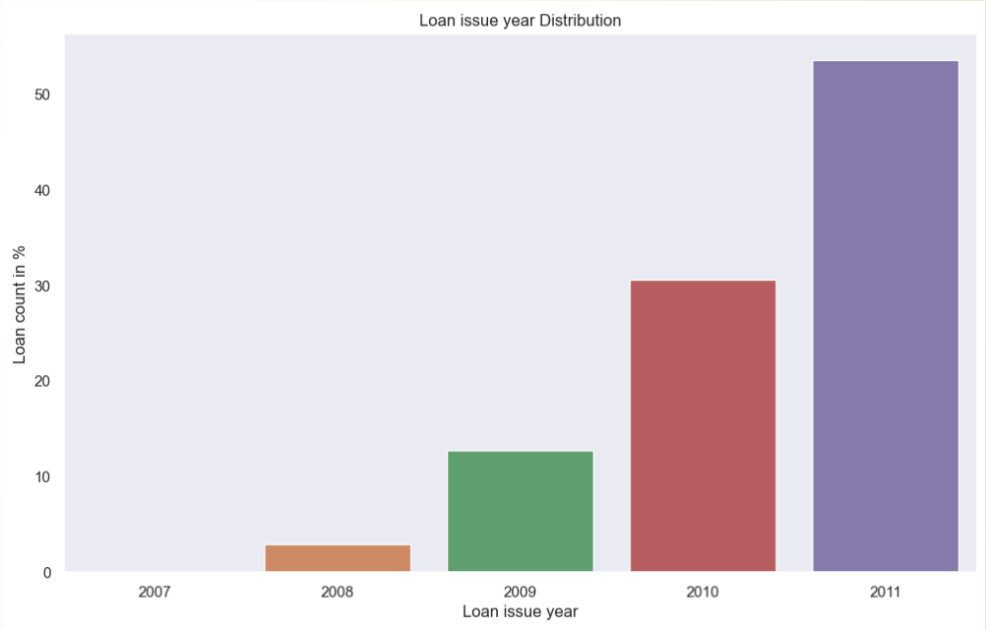


Observations:
From the above annual income data, we can say that most of the borrower's annual income is in the range of 40k to 80k.

Univariate Analysis

- **Converting the 'value_count' into percentage and analyzing on “Issue of Loan on define Months Distribution” & “Loan issue year Distribution” :**



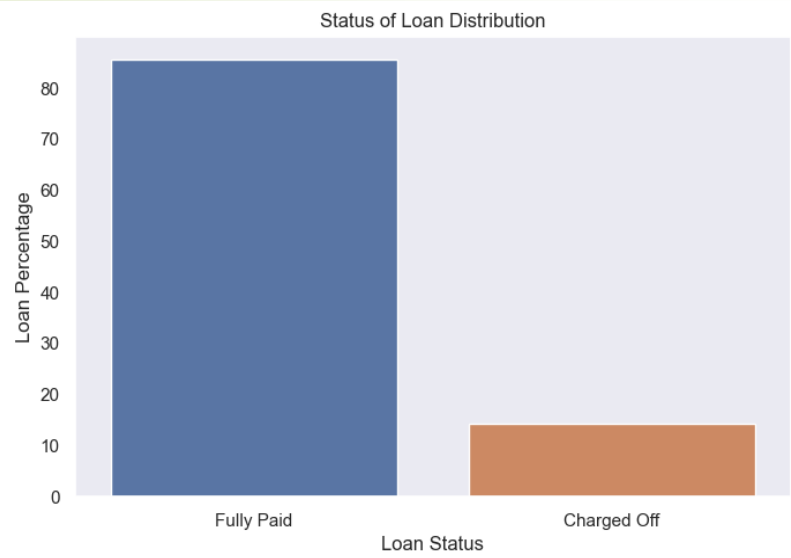


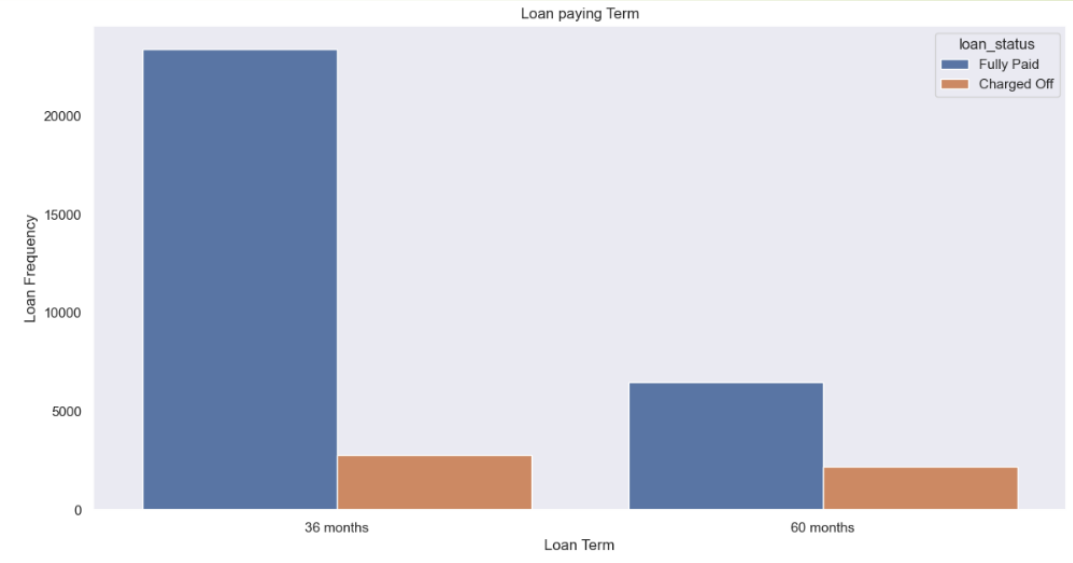
Observations:
From the above **“Issue of Loan on defined Months Distribution”** data, we can say that the issue of loans is increasing every month from Jan to Dec, and in the final quarter of the year there are more loans issued due to vacation.

Observations:
From the above “**Loan Issue Year Distribution”** data, we can say that the lending club has really expanded year by year, every year the number of loans doubled.

Univariate Analysis

- **Unordered Categorical Variables:** Converting THE "**value_count**" into a percentage and analyzing the 'Status of Loan Distribution'
- **Ordered Categorical Variables:** Converting THE "**value_count**" into a percentage and analyzing the 'Loan Category / Purpose'





Observations

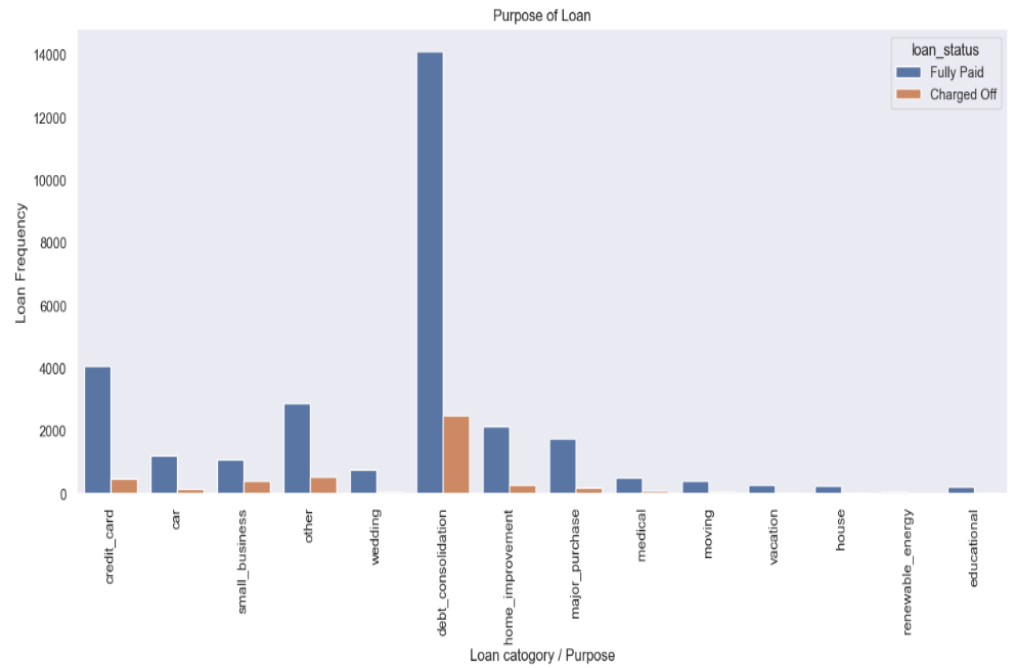
•From the above **Status of Loan Distribution** data, we can say that 85.7% have fully paid whereas 14.3% are charged off.

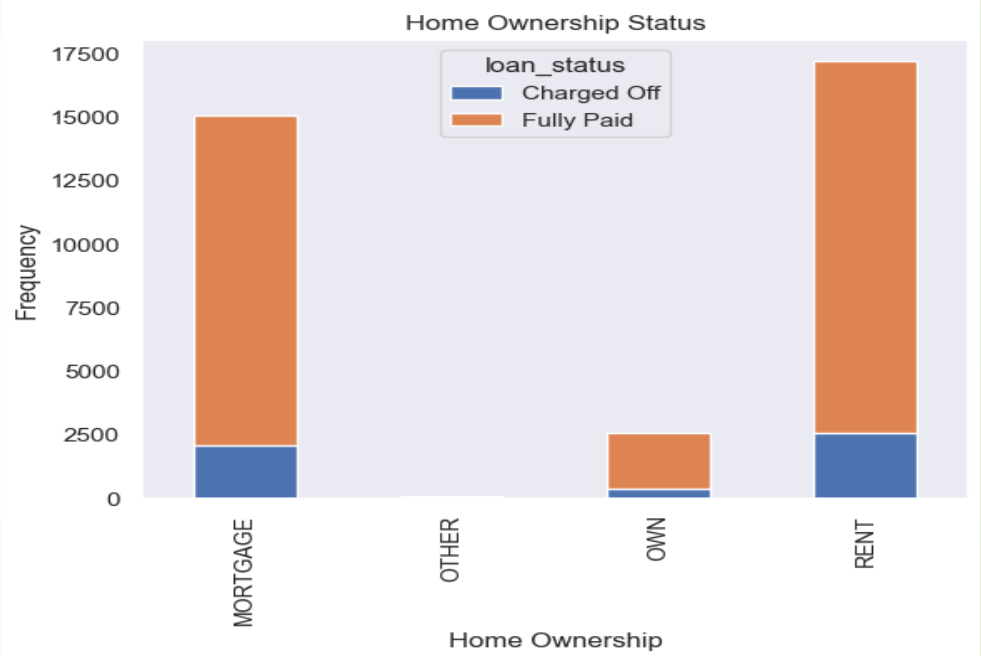
Observations:

•From the above **Loan Paying Term** data, we can say that there are 2 loan terms and most of the borrowers took 36 months tenure. But the ratio of charged off is high in 60 months tenure.

Univariate Analysis

- **Analyzing the "Purpose of Loan" & "Home Ownership":**





Observations

•From purpose data, we can say that most of them have taken loans for **Debt Consolidation** and paying credit card bills.

•Charged Off loan status is also high for debt consolidation.

Observations:

•From the above **Home ownership status** data, we can say that most of them have taken loans who are in **rent** or **mortgage** their home.

•Charged Off loan status is also high for these above two home ownership (**rent** or **mortgage**).

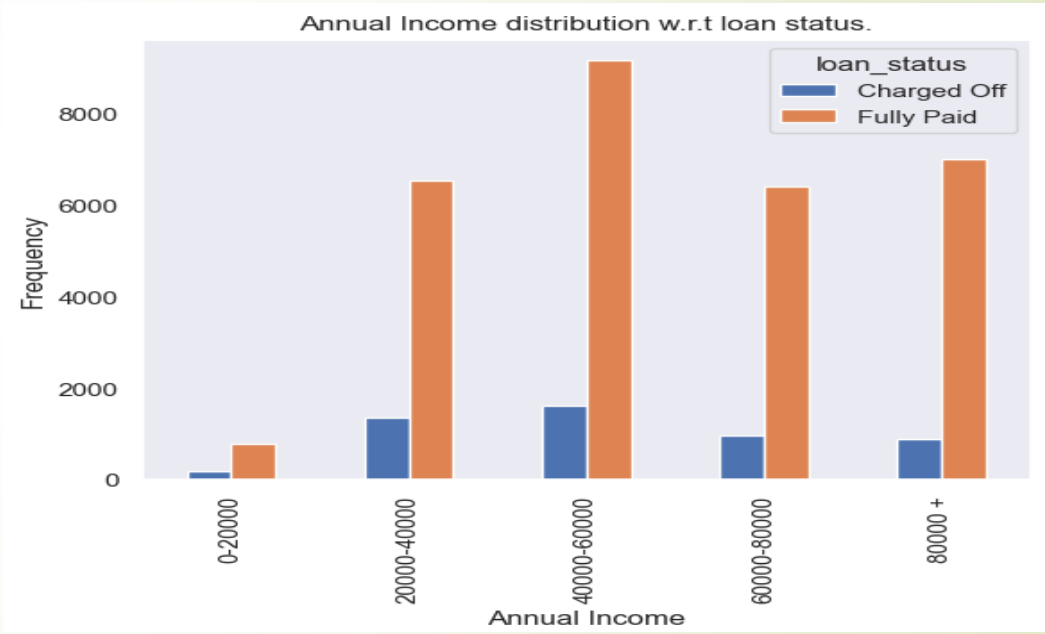
Segmented Univariate Analysis

- **Analyzing the "Loan Amount distribution with respect to loan status":**

Observations

•From above we can say that under 14k most of the borrowers take the loan amount, and charged off status is also high for those amounts.

- **Analyzing the "Annual Income distribution with respect to loan status":**

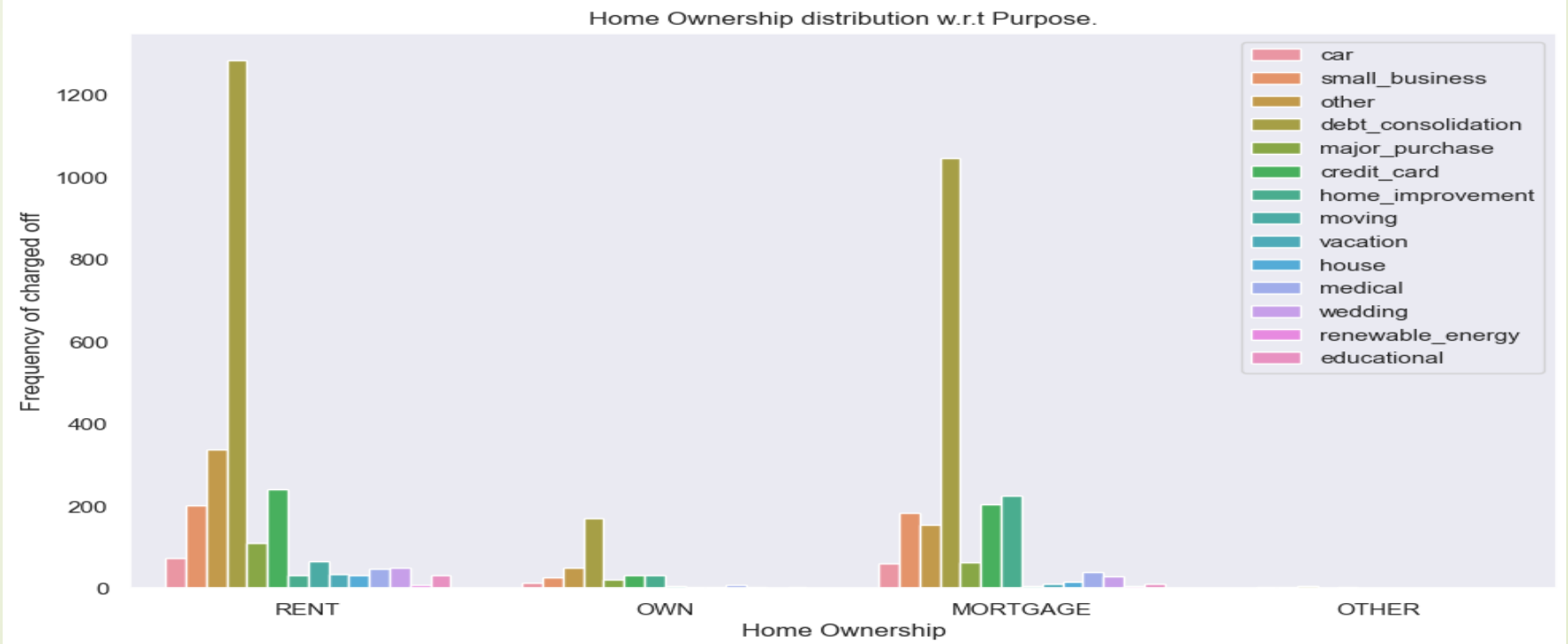


Observations:

•From the above chart we can say that most of the borrower's annual income is in the range of 40k to 60k.

Segmented Univariate Analysis

- **Analyzing the "Home Ownership distribution w.r.t Purpose in comparison with Frequency of charged off**

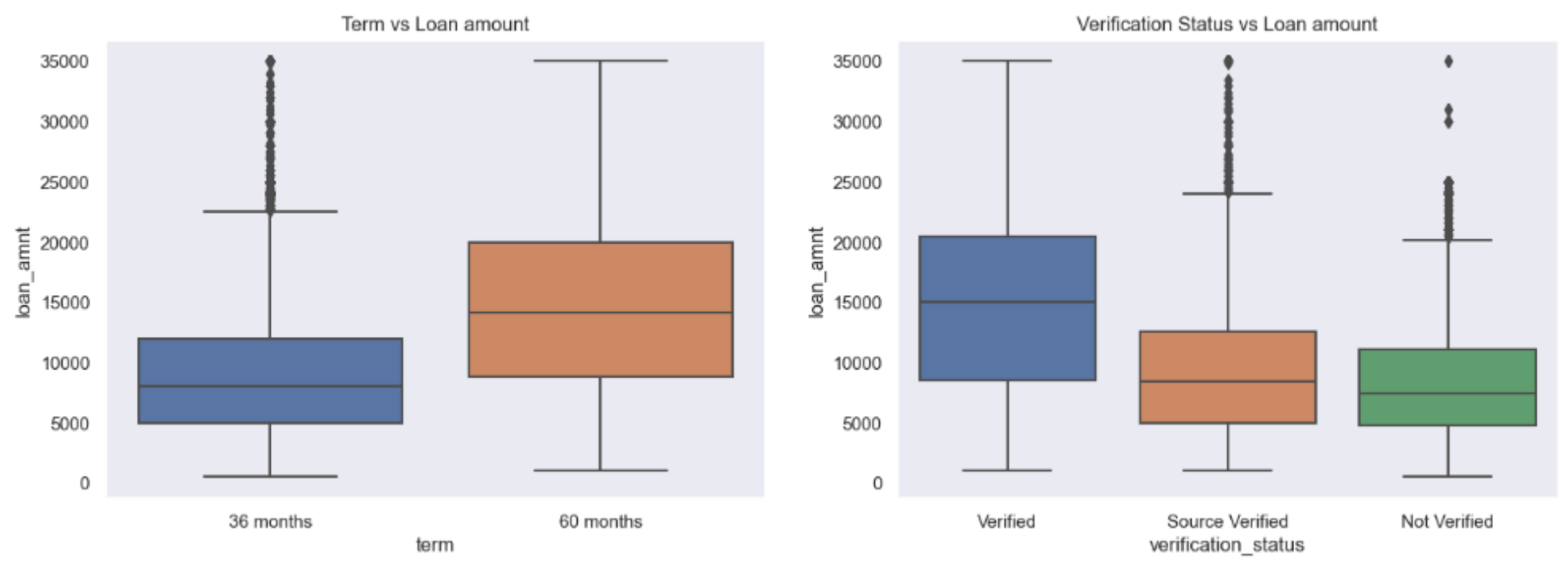


Observations

•From the above chart we can say that most of the borrowers who took loan for the purpose of debt consolidation has the highest number of Charged-off status and those who are in rent as the most.

Bivariate Analysis

- **Analyzing the Loan amount with "Loan Term vs Loan amount" and "Verify Status vs Loan amount"**



Observations

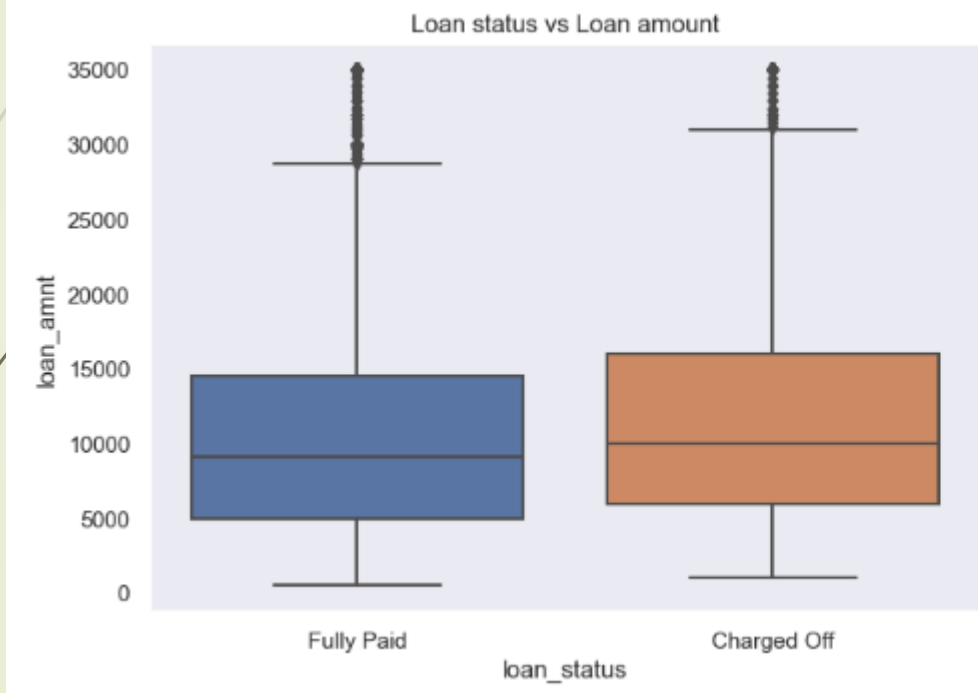
•The borrower who takes a higher loan amount tends to choose the loan term of 60 months.

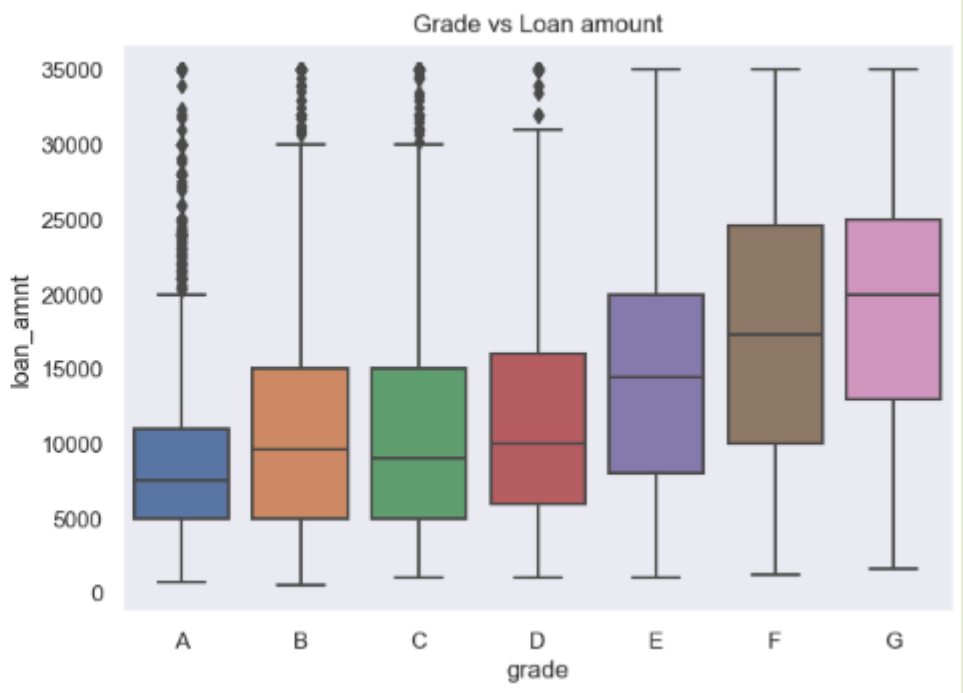
Observations

•Mostly verified borrowers are getting higher loan amounts due to security reasons.

Bivariate Analysis

- **Analyzing the Loan amount with "Loan status vs Loan amount" and "Grade vs Loan amount"**





Observations

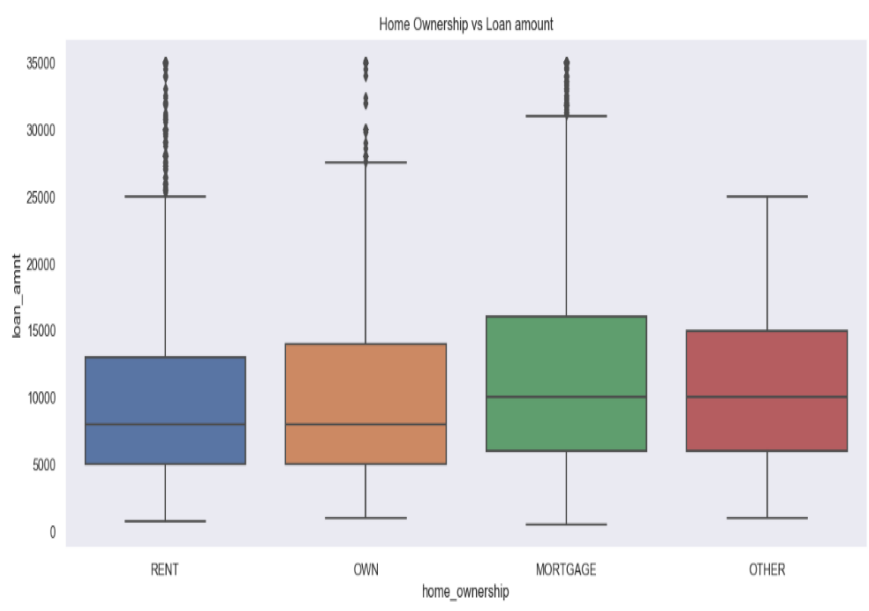
•In the **loan amount vs. Loan status** variable we can say that charged-off have higher loan amounts than fully paid.

•In **Grade vs. loan amount** we can say that grade F & G have having max amount of loan. As the grade decreases amount of the loan is increasing.

•From this we can say that the higher the grade more is the risk of default.

Bivariate Analysis

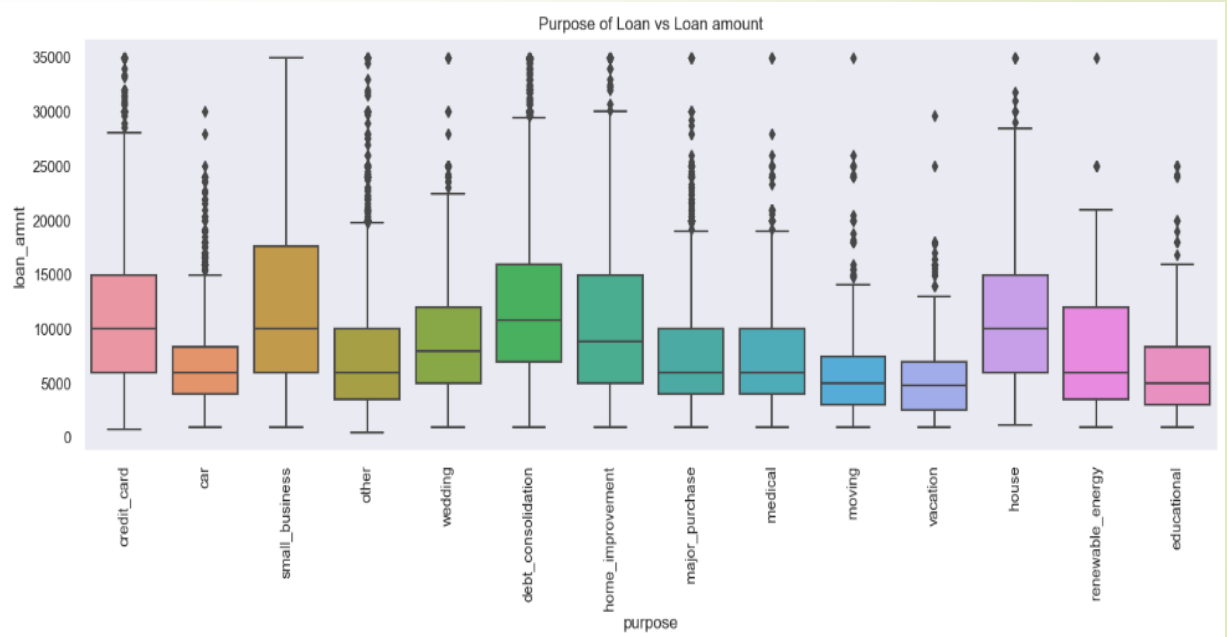
Analyzing the Loan amount with "Home Ownership vs Loan amount"



Observations

•In loan **amount vs home ownership** variable, we can say that the borrower who are from Mortgage have taken higher amount of loan than the others

Analyzing the Loan amount with "Purpose of Loan vs Loan amount"



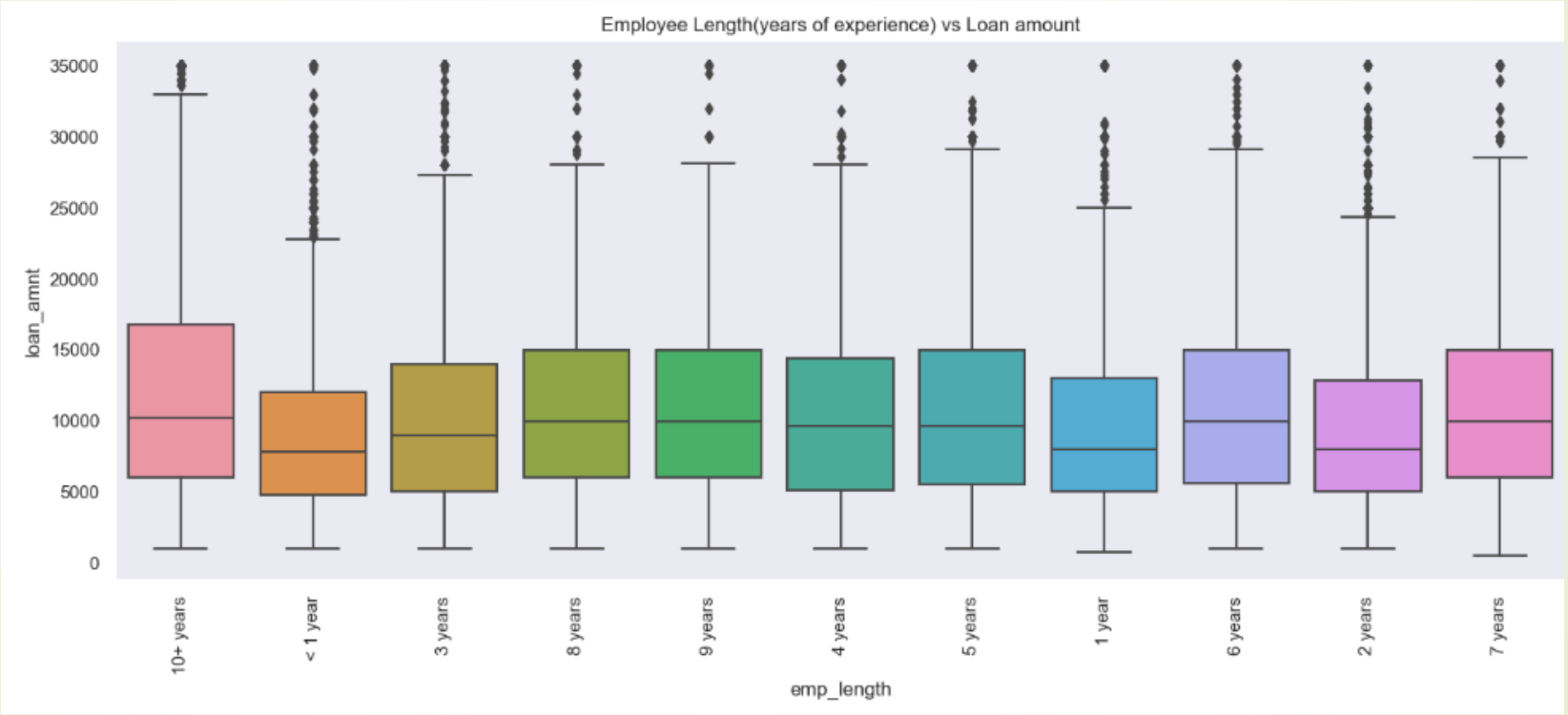
Observations

•In the **loan amount vs. purpose** variable, we can say that startups with **small business** borrowers are taking higher loan than others. Then comes **debt_consolidation** is second and **Credit card** comes 3rd.

•Median,95th percentile, and 75th percentile of loan amount is highest for loans taken for small business purposes among all purposes

Bivariate Analysis

Analyzing the Loan amount with "Employee Length(years of experience) vs Loan amount"

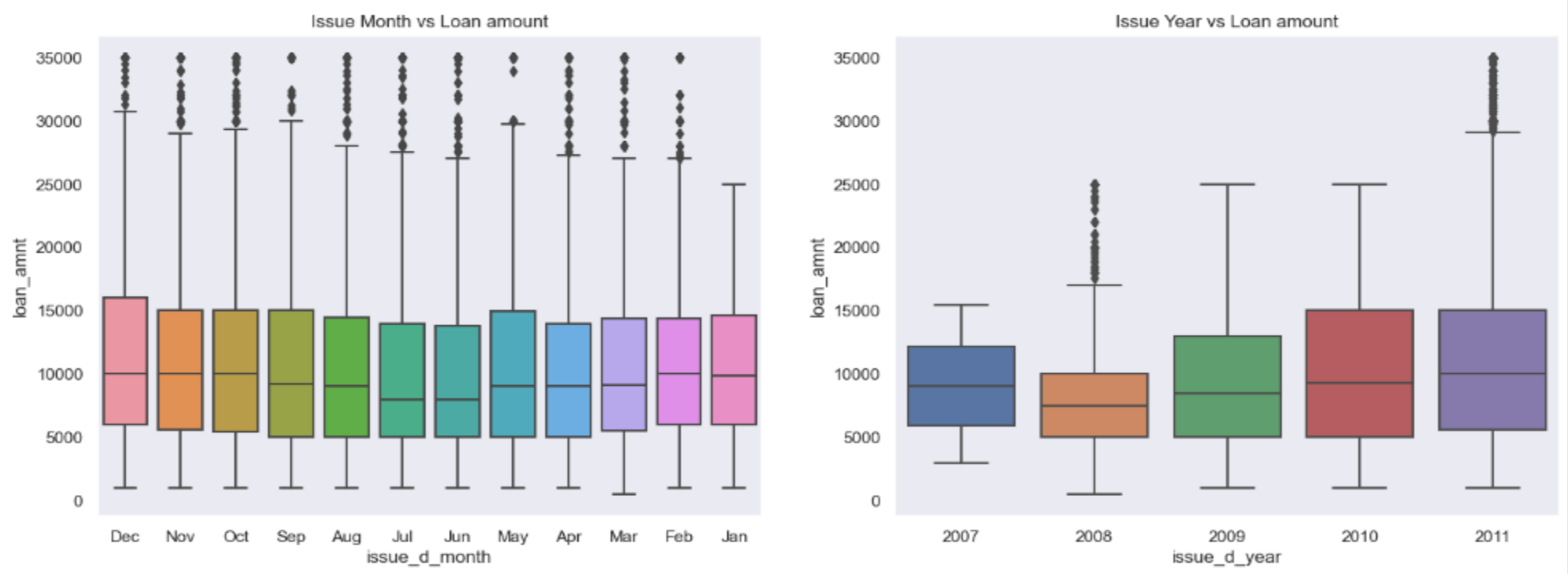


Observations

•In **loan amount vs employee length** variable, we can say that the borrower who has 10+ years of experience is taking a higher amount of loan than others, and borrowers with less than 1 year of experience are taking a lesser amount of loan compared to others.

Bivariate Analysis

Analyzing the Loan amount with "Issue Month vs Loan amount" with "Issue Year vs Loan amount"



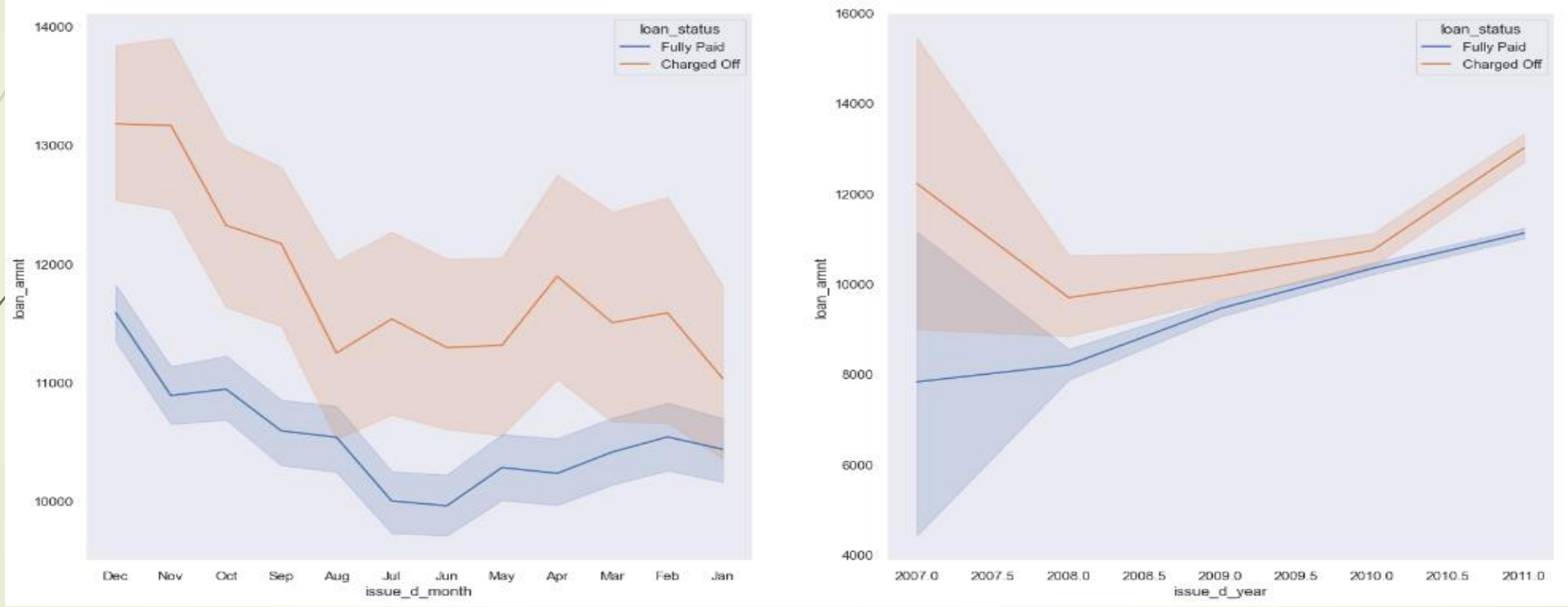
Observations

•In **loan amount vs issue month** variable, we can say that the highest loan amount is taken in the months of Dec and May whereas the median value doesn't vary too much.

•In **Issue year vs. loan amount,** we can say that the highest loan amount is taken in the years 2008 and 2011 as we can see in outliers by some borrowers. We can say that the median value doesn't vary too much but as the year increases high amount of loans are taken.

Bivariate Analysis

Analyzing the Loan amount with Loan Amount vs. Issue month vs. Issue year to verify the frequency



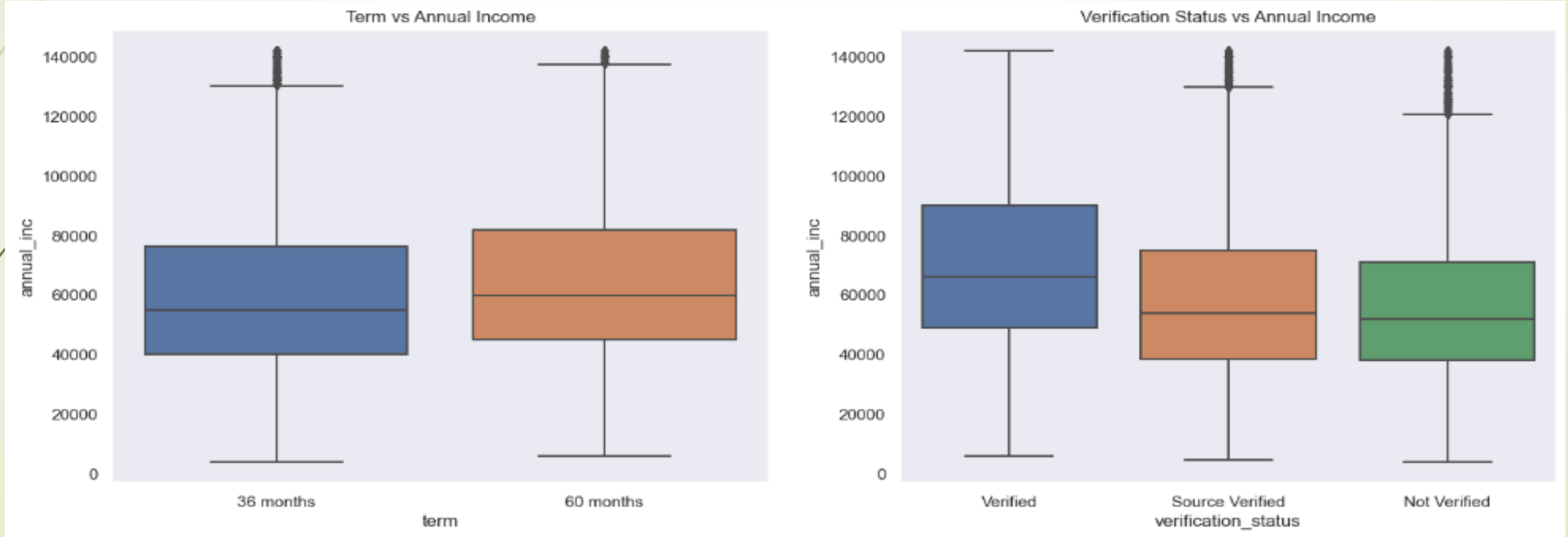
Observations

•From the above **Lineplot** we can say that the higher the loan amount, the more the charged-off frequency.

•And the borrower who took the loan in the months of November and December has the highest charged-off ratio with the highest loan amount.

Bivariate Analysis

Analyzing the Annual Income with "Term vs Annual Income" and "Verification Status vs Annual Income"



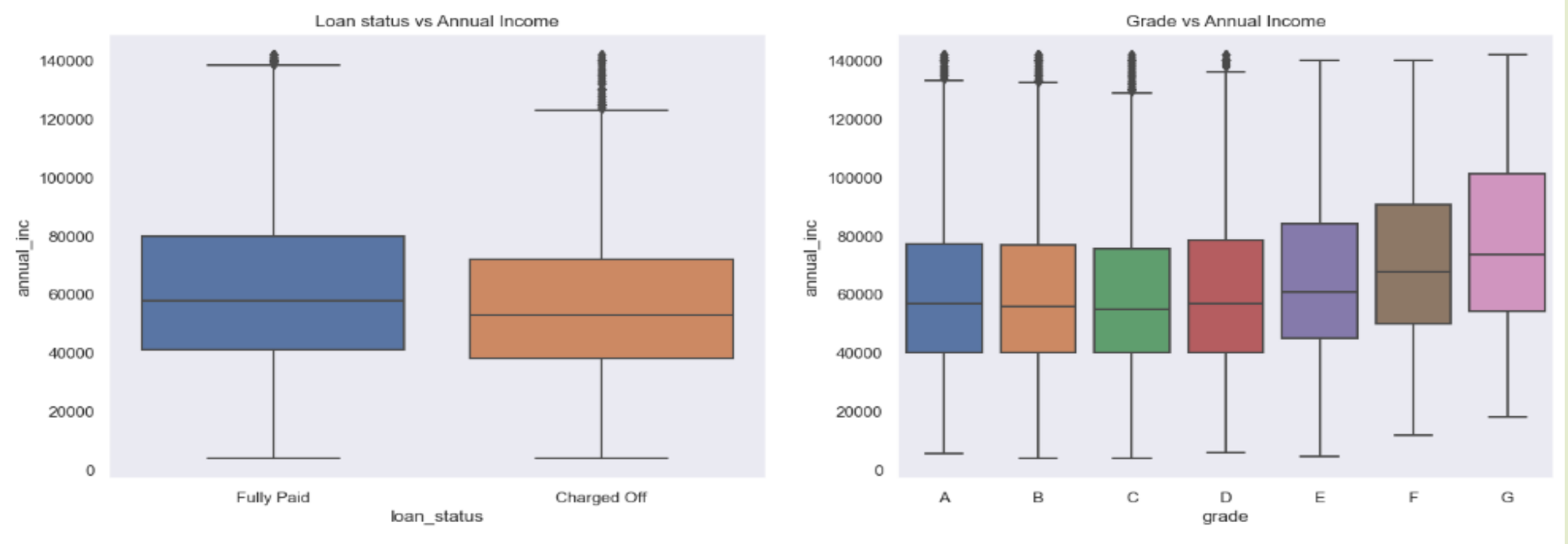
Observations

•In the Term vs. Annual Income variable, we can say that the borrowers who have high annual income are taking loans for 60 months tenure as compared to 36 months.

•In Verification status vs. annual income, we can say that, mostly the verified borrowers are having high annual income than others.

Bivariate Analysis

Analyzing the Annual Income with "Loan status vs Annual Income" and "Grade vs Annual Income"

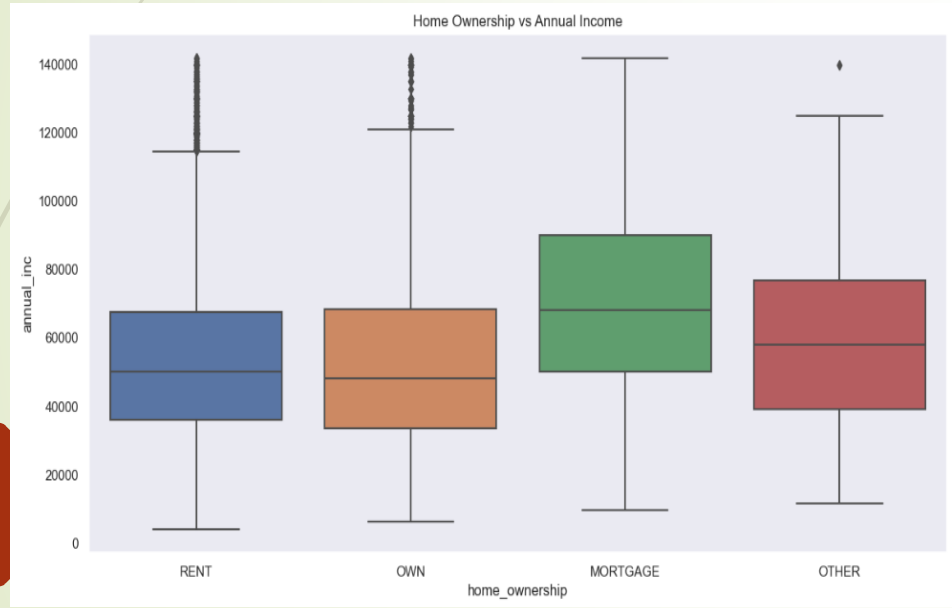


Observations

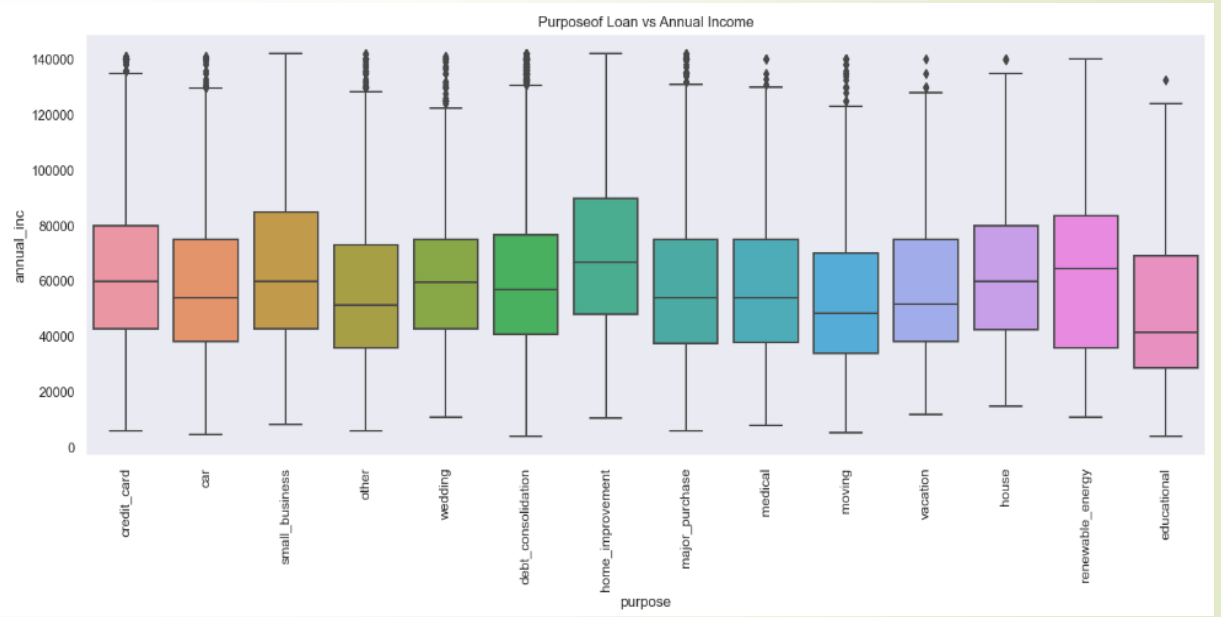
- In **the loan amount vs. term variable**, we can say that the loan amount higher is the tenure i.e. 60 months, Its median is only 15k whereas the median of 36 months is 8k.
- In **Verification status vs loan amount** we can say that the verified borrower gets more loan amount than Non-verified and Source Verified i.e. above 10k loan amount everyone is verified.

Bivariate Analysis

Analyzing the Annual Income with "Home Ownership vs Annual Income"



Analyzing the Annual Income with "Purpose of loan vs Annual Income"



Observations

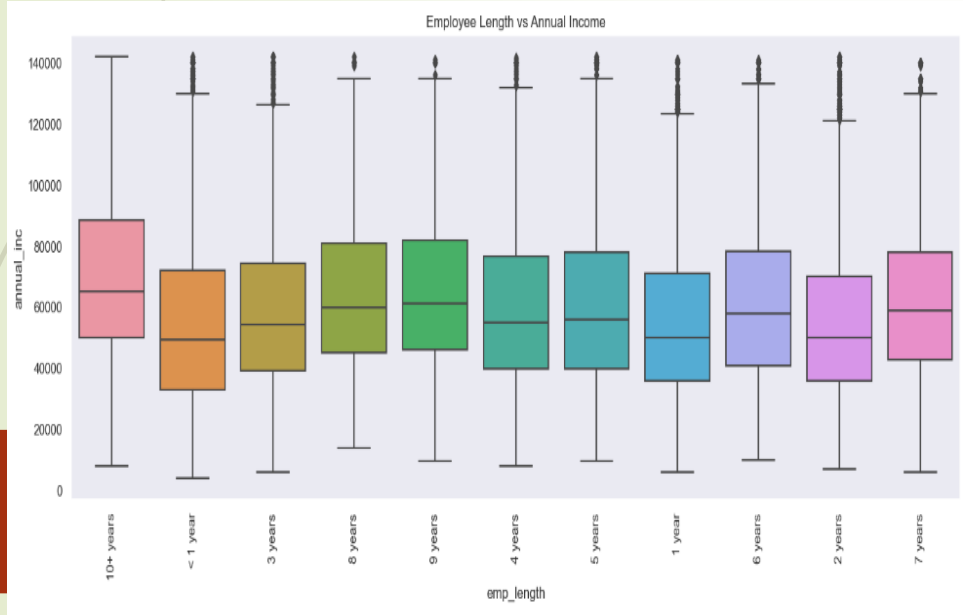
•In the **Annual Income vs Home Ownership** variable, we can say that the borrowers who have the status as Mortgage have higher annual income than others.

Observations

•In the Annual Income vs. Purpose variable, we can say that the borrowers who have high annual income are taking loans mostly for home improvement and small business.

Bivariate Analysis

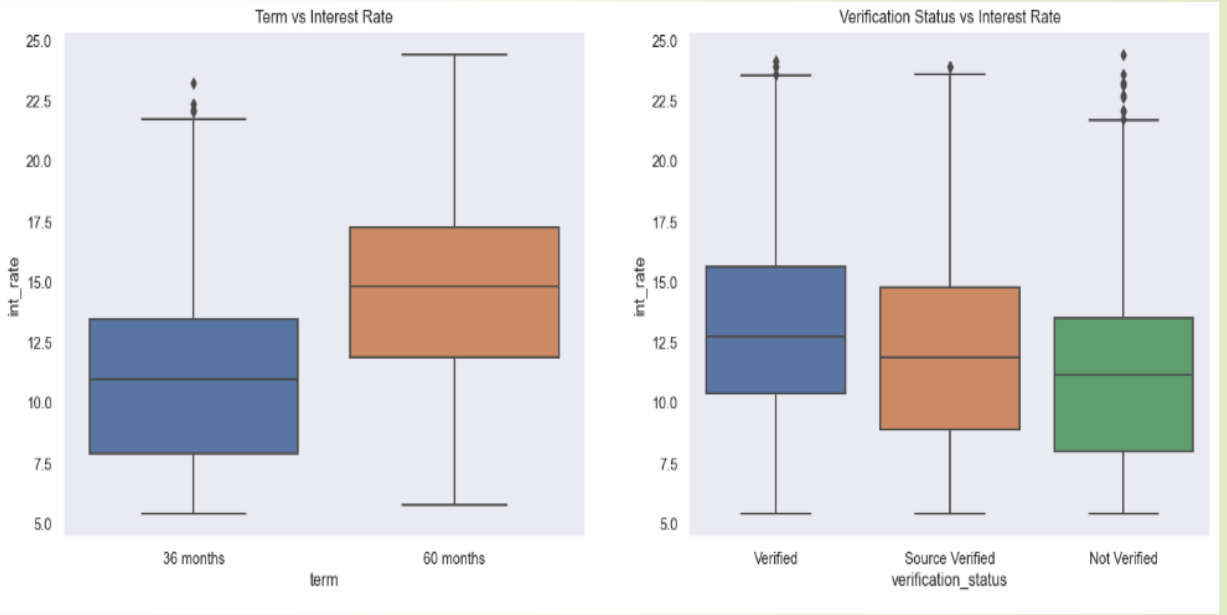
Analyzing the Annual Income with "Employee Length vs Annual Income"



Observations

•In the **Annual income vs. employee length** variable, we can say that, as we expected the borrowers who have 10+ experience have having highest annual income than others and >1 has having least.

Analyzing the Segment Interest Rate with "Term vs Interest Rate" and "Verification Status vs Interest Rate"



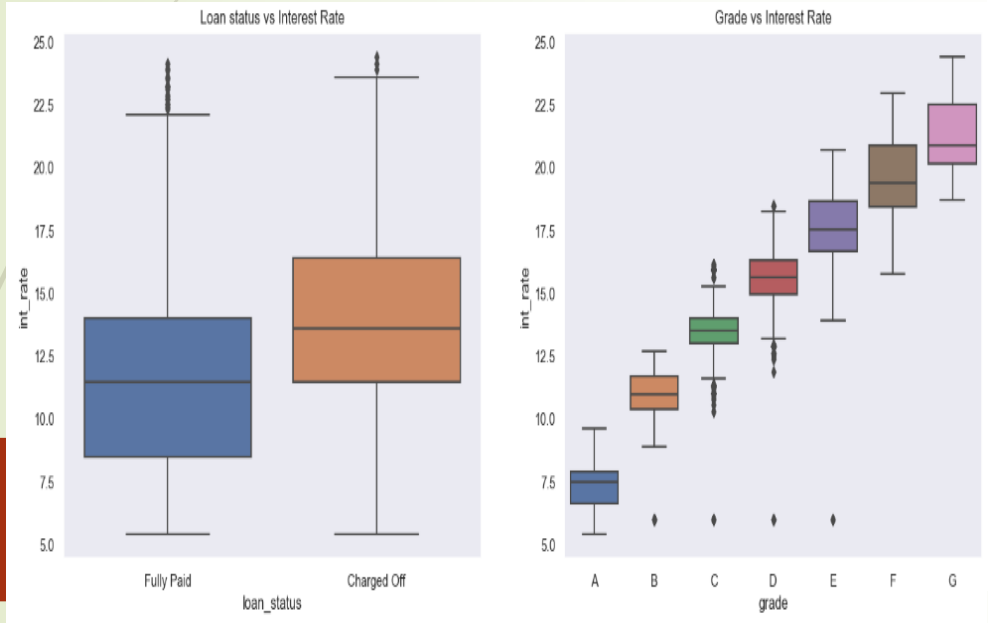
Observations

•In the **term vs. interest rate** variable, we can say that the interest rate is less for those who take the loan for 36 months and high for 60 months.

•In the **Verification Status vs. interest rate** variable, we can say that a verified borrower gets more loans at higher interest rates compared to others

Bivariate Analysis

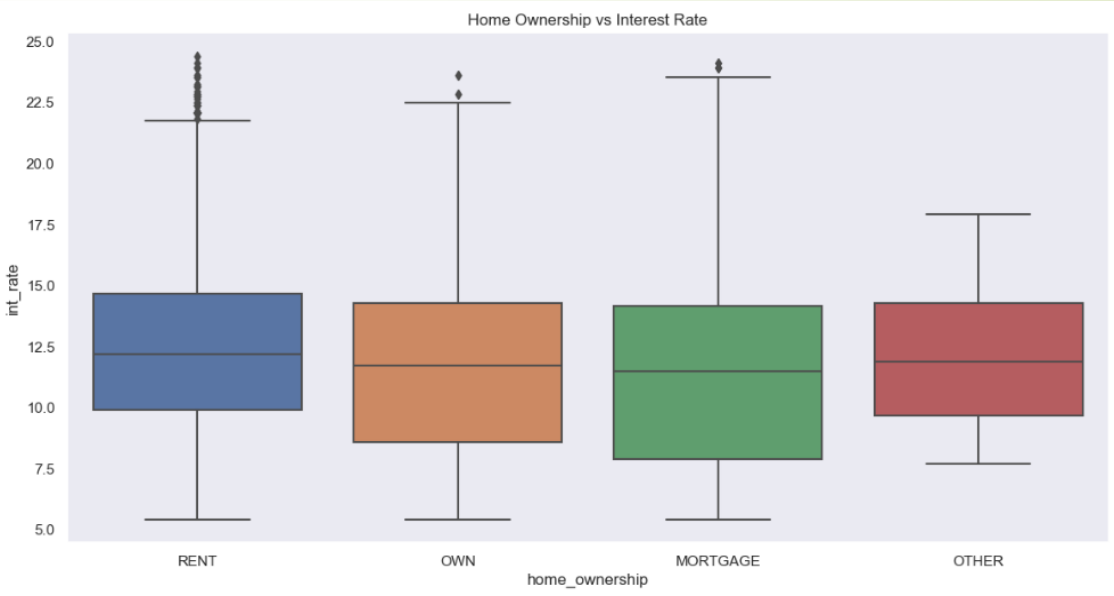
Analyzing the Segment Interest Rate with "Loan status vs Interest Rate" and "Grade vs Interest Rate"



Observations

- In the **loan status vs. interest rate** variable, we can say that the borrowers who have high interest rates are mostly defaulters.
- In **grade vs. interest rate** we can say that, as the grade decreases, interest rate is also increasing.

Analyzing the Segment Interest Rate with "Home Ownership vs Interest Rate"

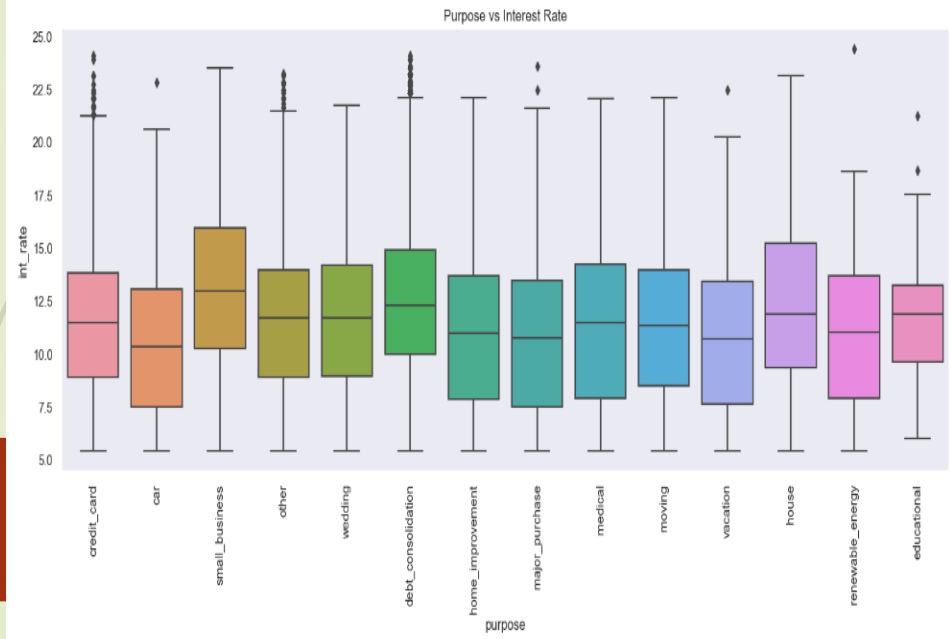


Observations

- In the **home ownership vs interest rate** variable, we can say that the borrowers who are in rent are getting high interest rates. The borrowers with their own mortgages are getting loans with lower interest rates also due to security purposes.

Bivariate Analysis

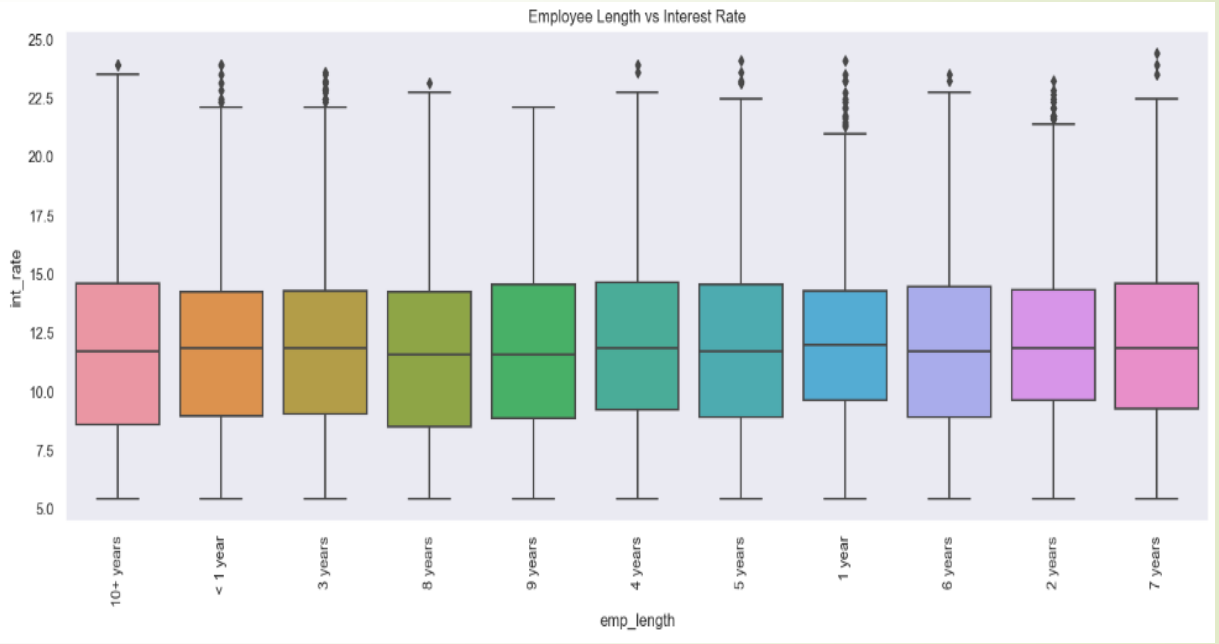
Analyzing the Segment Interest Rate with "Purpose vs Interest Rate"



Observations

•In the **purpose vs. interest rate** variable, we can say that small business, debt consolidation, and house loans are getting loan at higher interest rate than others

Analyzing the Segment Interest Rate with "Employee Length vs Interest Rate"

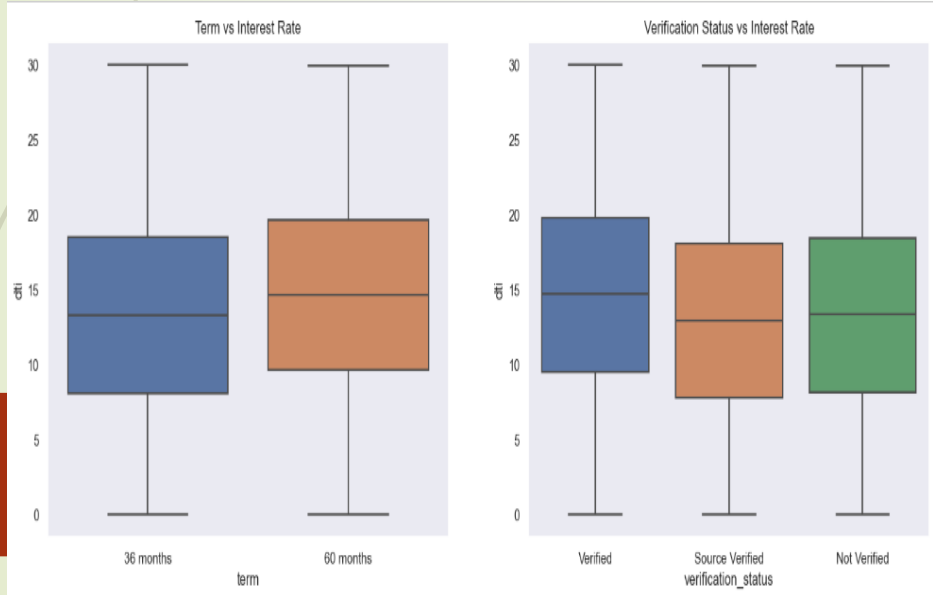


Observations

•There is no much relation between Employment length and interest rate.

Bivariate Analysis

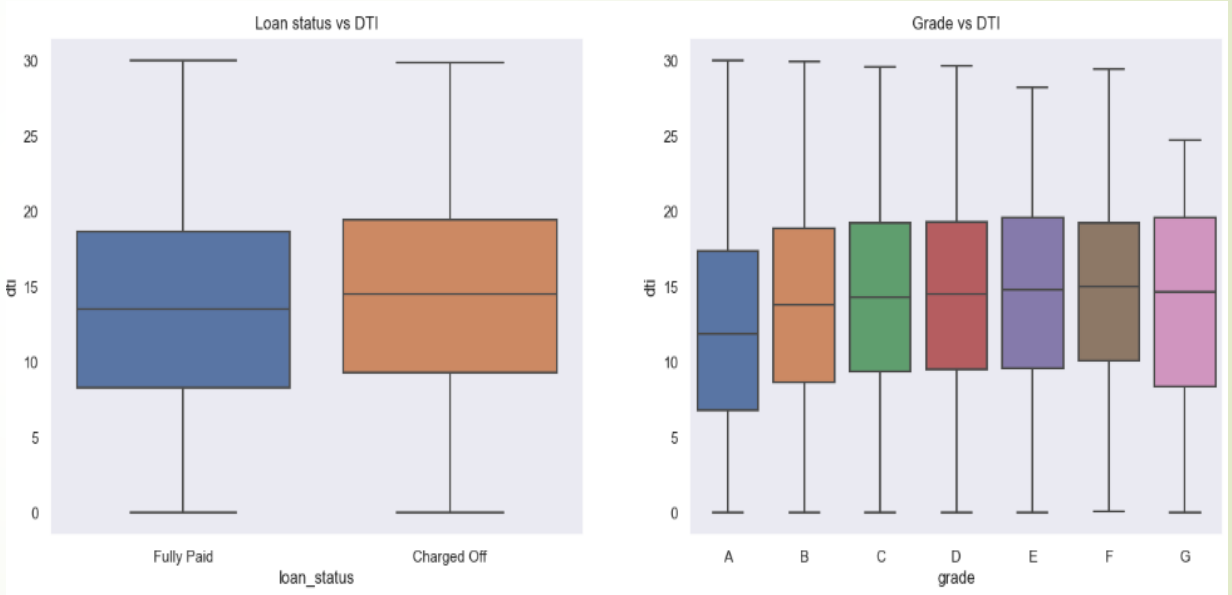
Analyzing the In Debt to income ratio compare "Term vs Interest Rate" and "Verification Status vs Interest Rate"



Observations

- In the **interest rate vs term** variable, we can say that dti ratio is high for the borrower who has 60 months tenure.
- In **Verification status vs dti** we can say that, verified borrowers are having high dti ratio.
-

Analyzing the In Debt to income ratio compare "Loan status vs DTI" and "Grade vs DTI"

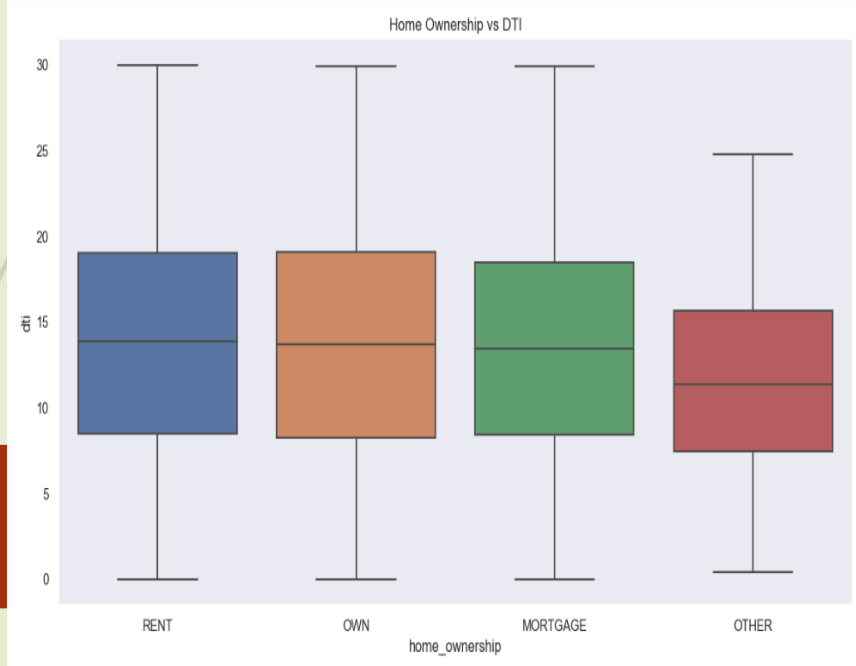


Observations

- Most of the charged-off borrowers have had a high dti ratio.
- As grade decreases dti ratio is increasing.

Bivariate Analysis

Analyzing the In Debt to income ratio compare "Home Ownership vs DTI"

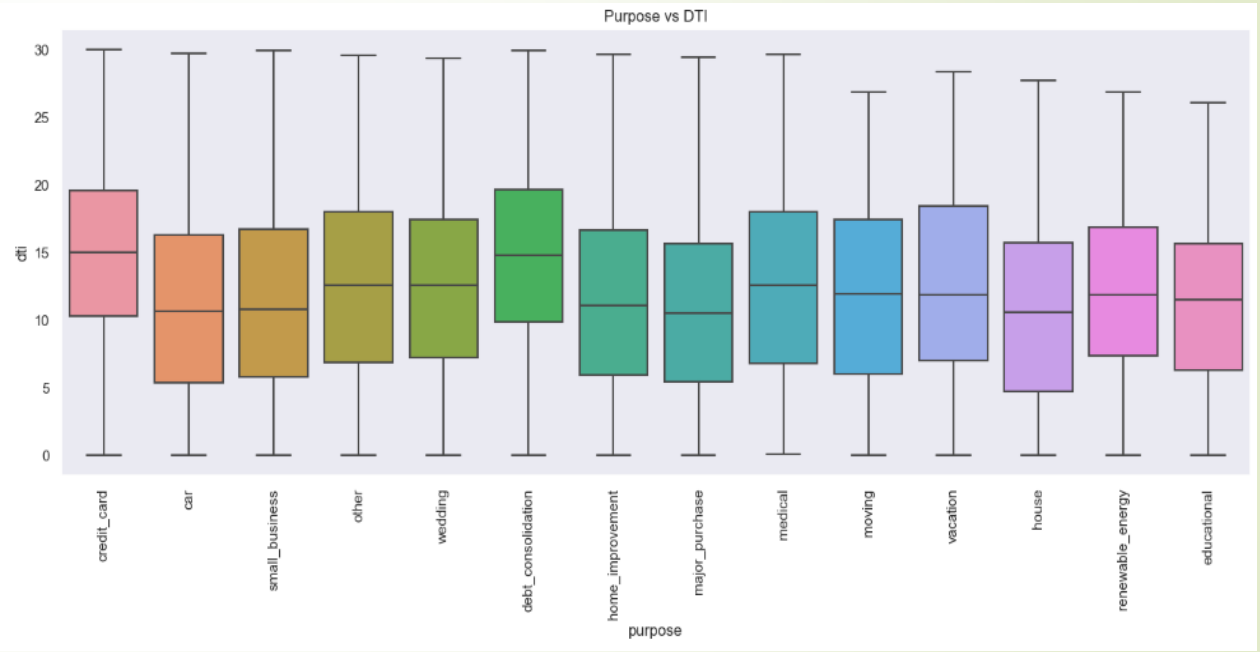


Observations

•In the **dti vs. home ownership** variable we can say that the borrowers with own and rent home status are having high dti ratio than others.

•

Analyzing the In Debt to income ratio compares "Purpose of loan vs DTI"

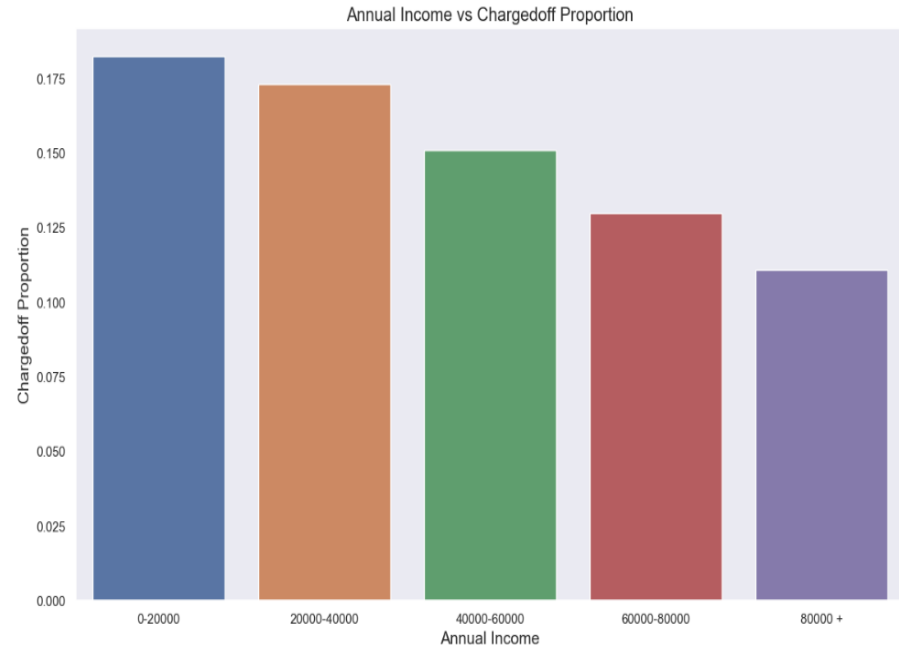


Observations

•In the **purpose vs dti** variable we can say that the borrower who takes loan for the purpose of debt consolidation and credit card are having high dti ratio than others.

Bivariate Analysis

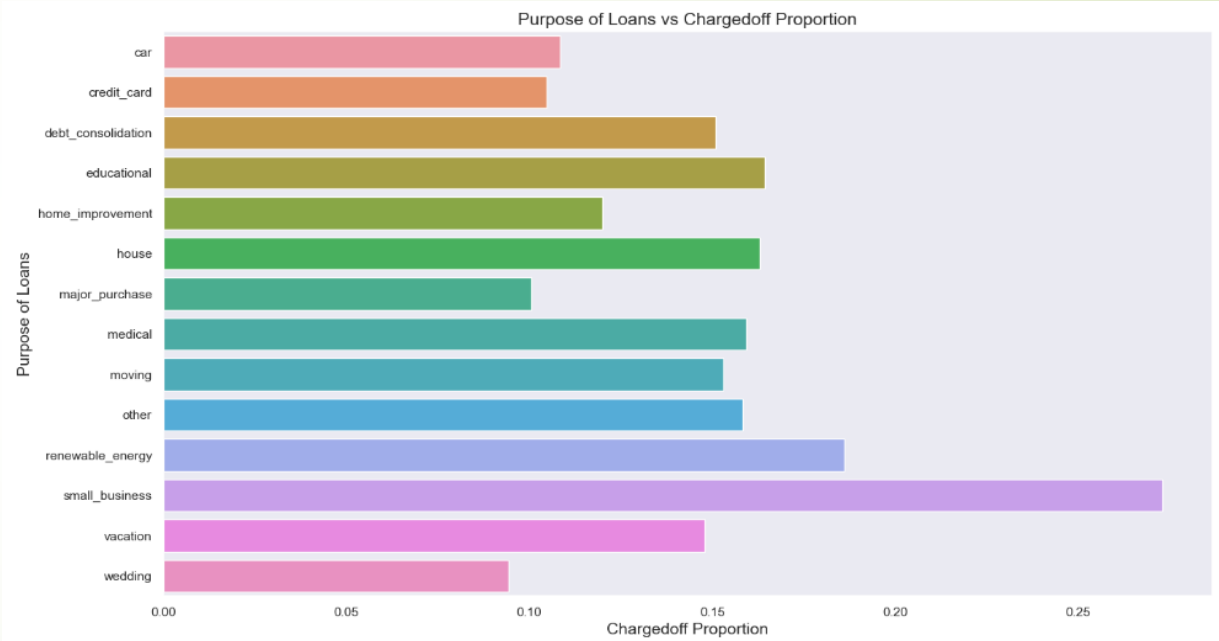
Analyzing the Charged Off Proportion with 'Annual Income vs Charged off Proportion



Observations

- From the above chart we can say that, as annual income is increasing charged-off proportion is decreasing.
- So, the highest charged-off proportion are in the range of 0 to 20k annual income.
-

Analyzing the Charged Off Proportion with 'Purpose of Loans vs Charged off Proportion'

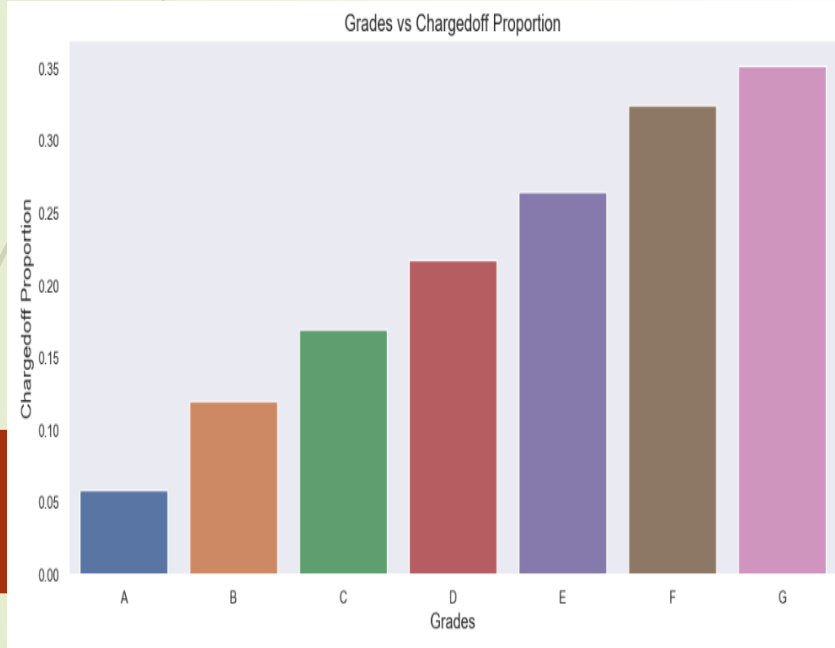


Observations

- From above we can say that the borrower who takes a loan for the purpose of **small business** has the maximum charged-off proportion.

Bivariate Analysis

Analyzing the Charged-Off Proportion with 'Grades vs Charged-off Proportion'

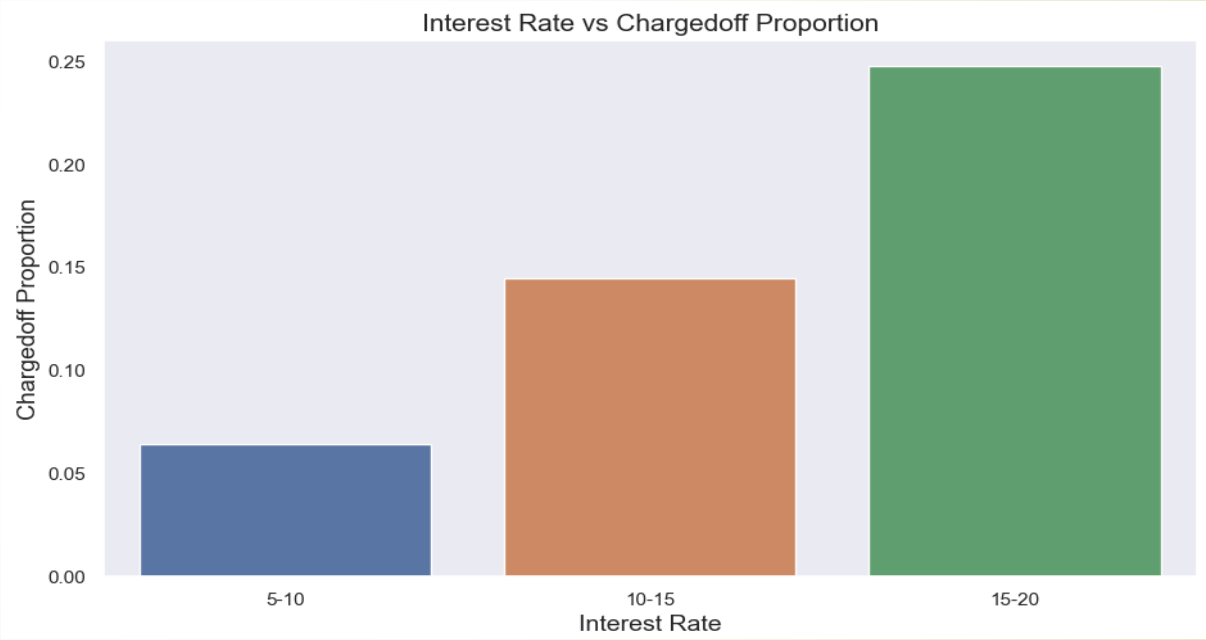


Observations

•From the above chart we can say that, as grades are decreasing charged-off proportion is increasing.

•

Analyzing the Charged-Off Proportion with 'Interest Rate vs Charged off Proportion'



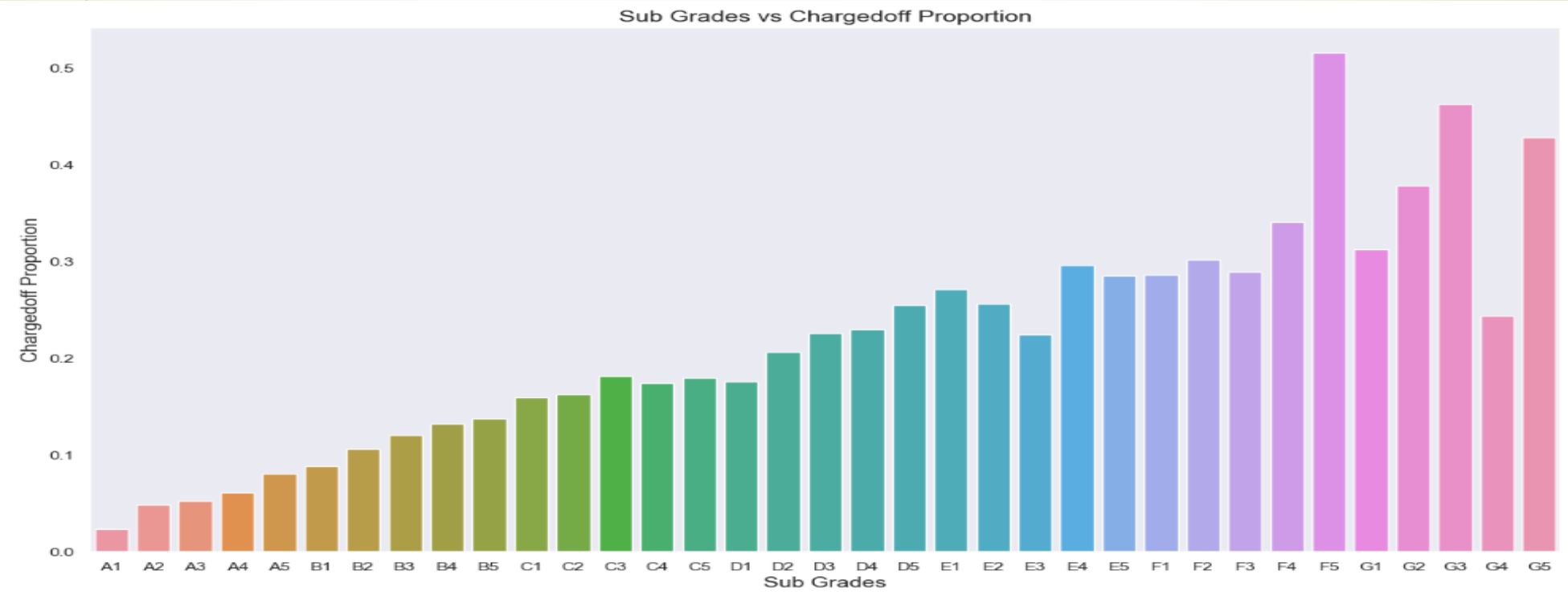
Observations

•As interest rates are increasing, the charged-off proportion is also increasing.

•The borrower who takes a loan at the interest rate of 15 to 20 are getting mostly charged-off.

Bivariate Analysis

Analyzing the Charged-Off Proportion with 'Sub Grades vs Charged off Proportion'

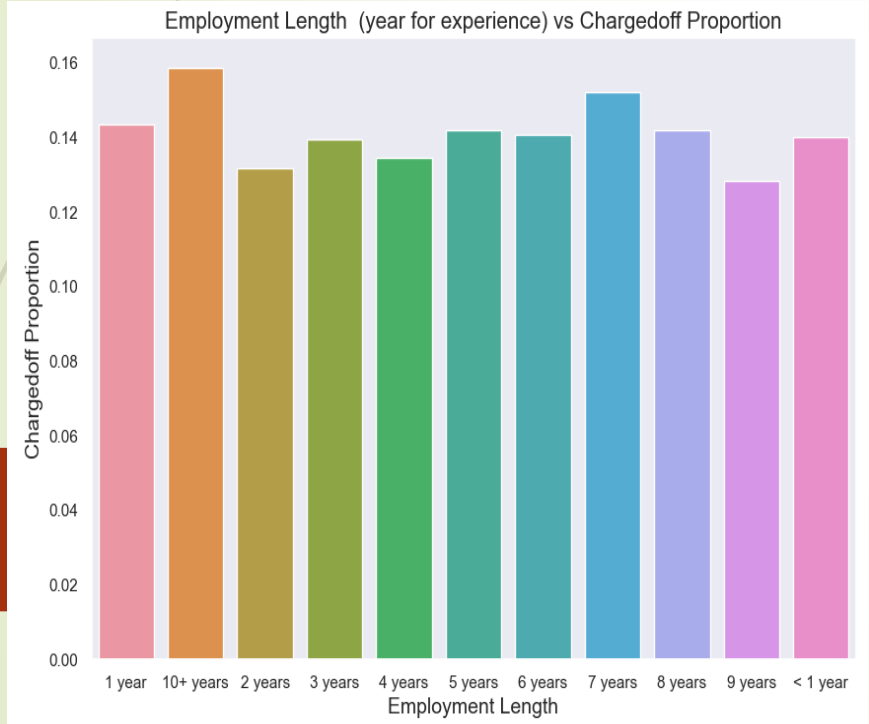


Observations

•Subgrade F5, G3, and G5 have maximum charged-off proportion.

Bivariate Analysis

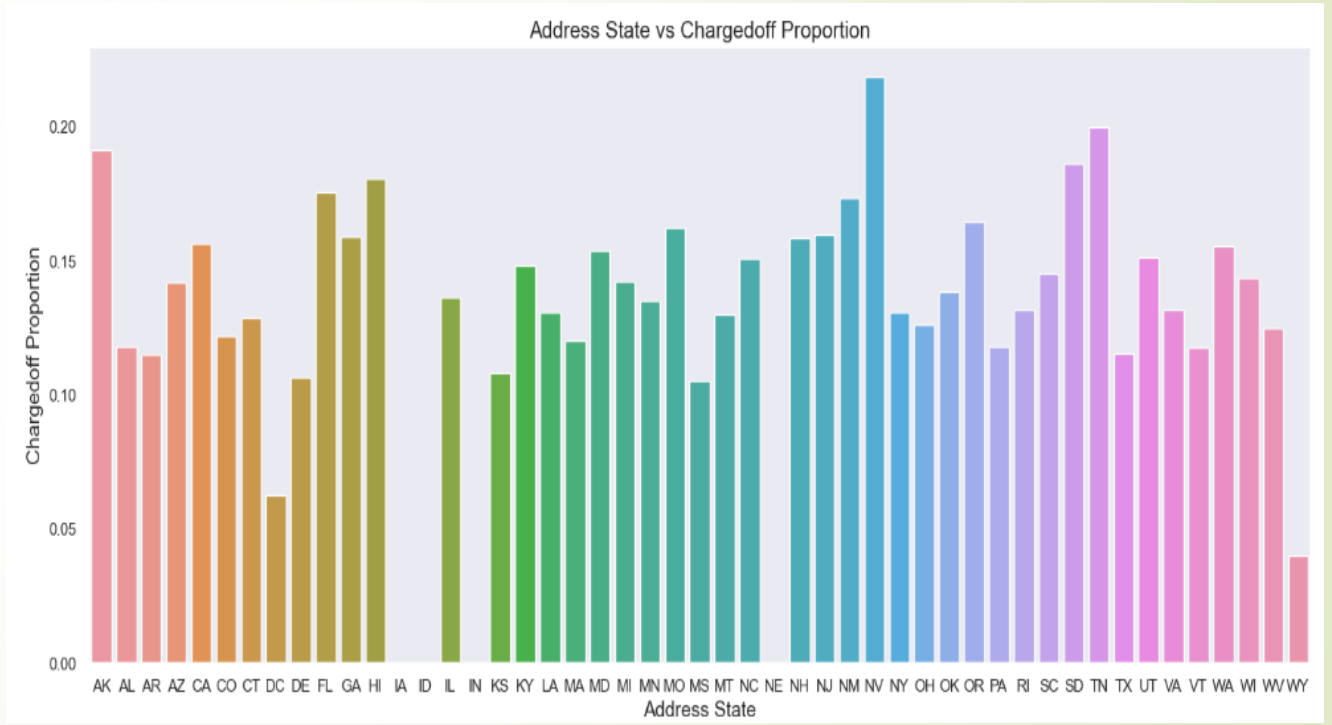
Analyzing the Charged Off Proportion with 'Employment Length (year for experience) vs Charged off Proportion'



Observations

•The borrowers whose employee length is less than 1 year, 1 year, and 10+ years are mostly getting charged off.

Analyzing the Charged-Off Proportion with 'Address State vs Charged off Proportion'

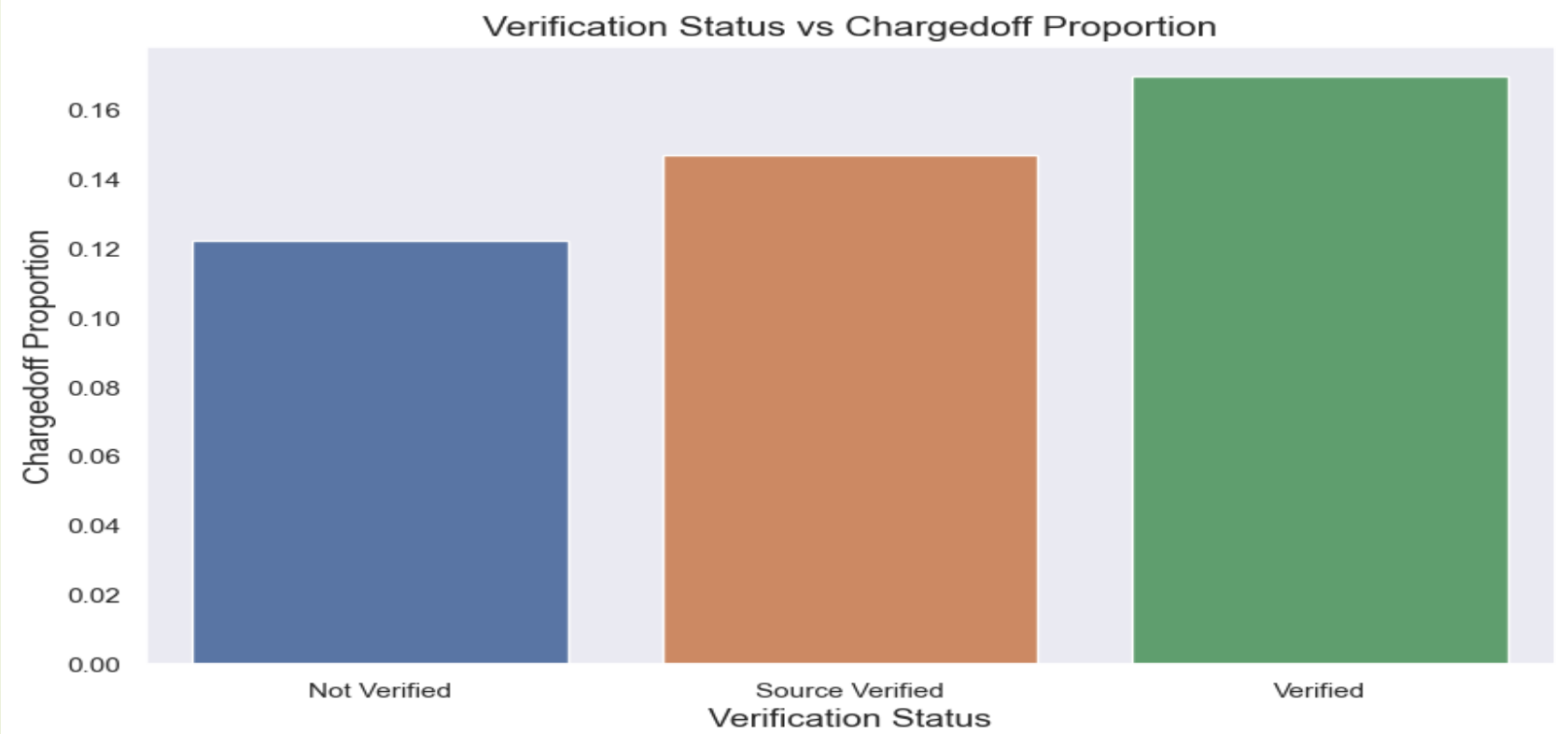


Observations

•The borrowers who are from NV, AK, TN state are mostly getting charged-off.

Bivariate Analysis

Analyzing the Charged-Off Proportion with 'Verification Status vs Charged off Proportion'

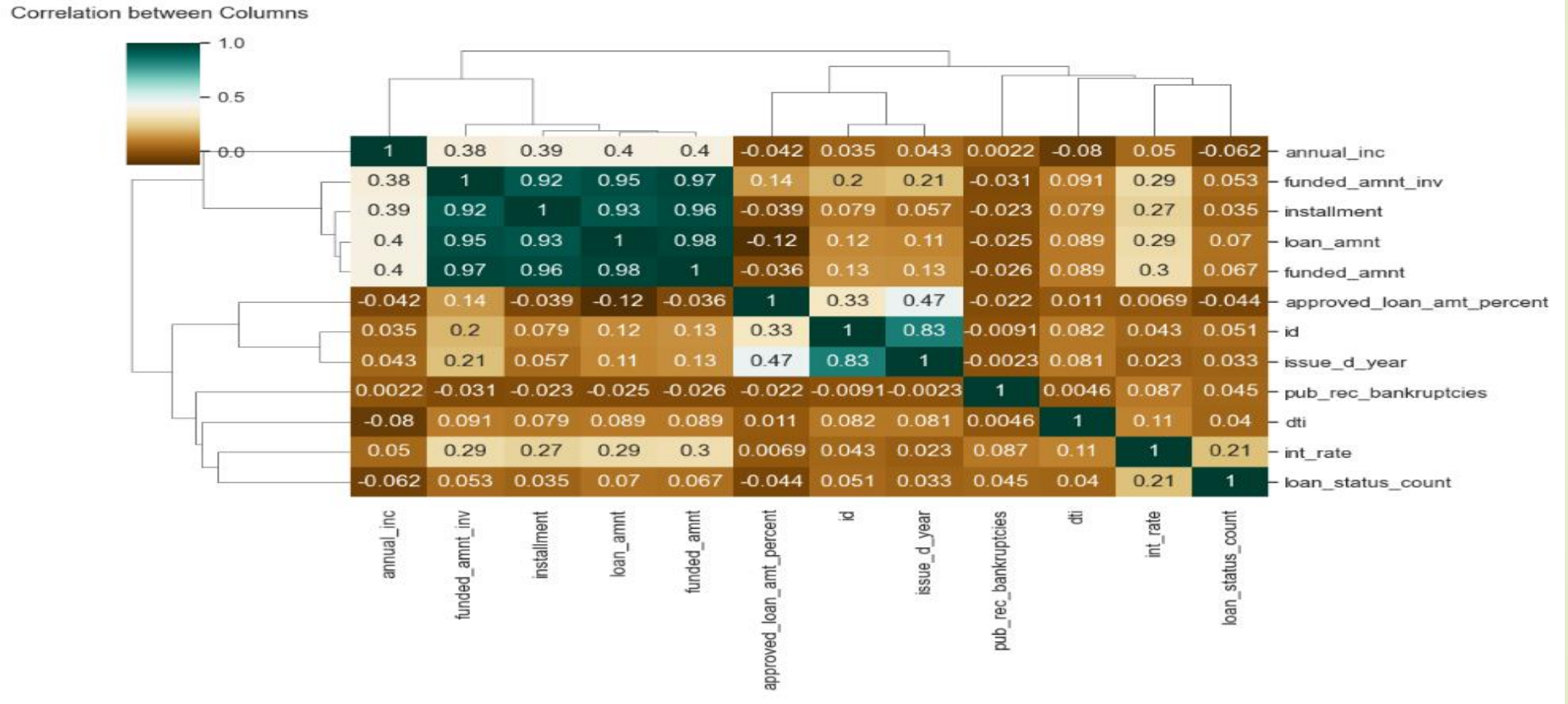


Observations

•Most of the verified borrowers are getting charged off than others.

Multivariate Analysis

With multivariate Analysis finding the correlation matrix on Cluster-map

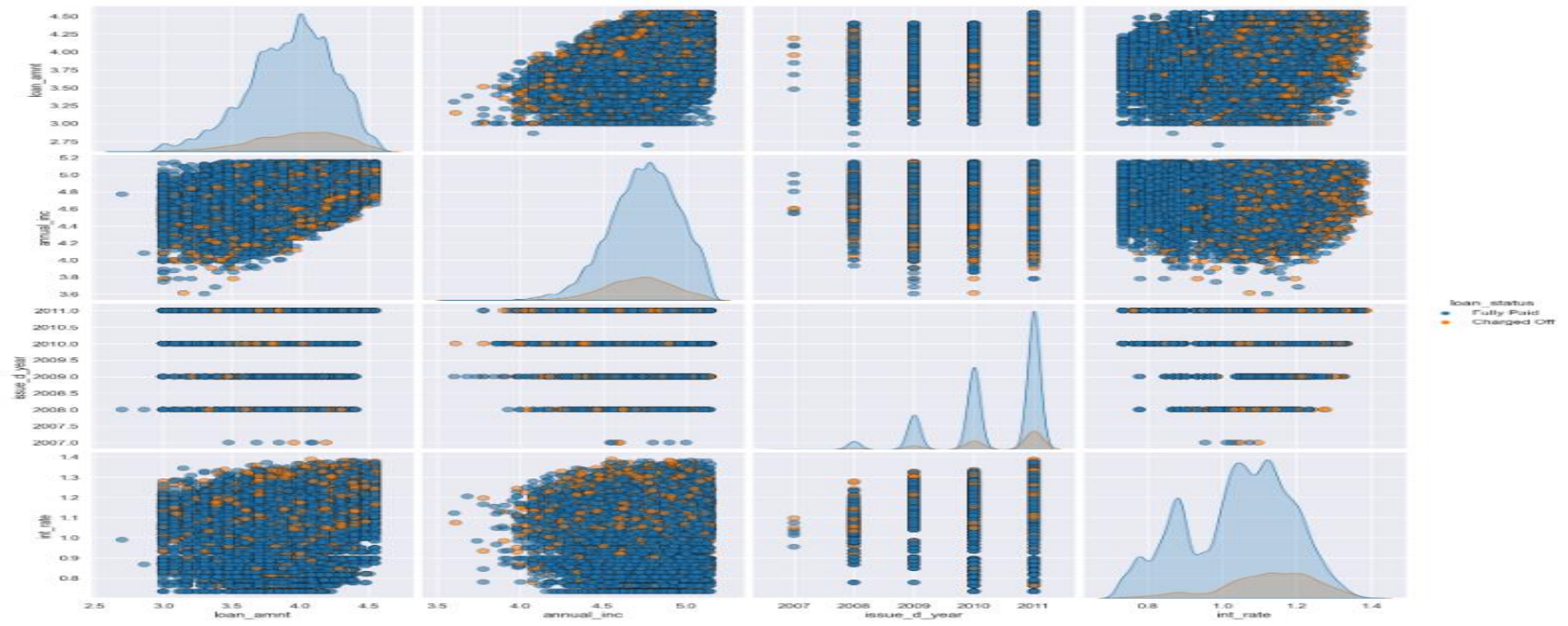


Observations:

- **Loan_amnt , funded_amnt , funded_amnt_inv** , installment are strongly correlated.
- **Annual_inc and dti** is negatively correlated.
- Debt income ratio is the percentage of a borrower's monthly gross income that goes toward paying debts.
- Which means when **Annual_inc** is low, debt is high and vice versa.

Multivariate Analysis

With multivariate Analysis finding the correlation matrix on pair plot by loan status with a density plot of the diagonal and format the scatter plots.



Observations:

- The higher the annual income, the higher the loan amount slightly.
- Interest rate is increasing with loan amount increase.
- Increase in the number of charged-off with an increase in year.
- The higher the interest rate, the higher the charged-off ratio.

Recommendations / Summarize Results

From all the above analysis, we can say that there is a higher probability of defaulting when:

- Borrowers are taking loans for a term of 60 months.
- Borrowers whose loan status is 'Verified' as they took a high loan amount with 60 months tenure.
- Borrowers who have home ownership as 'Rent' and take loans for the purpose of debt consolidation.
- Borrowers whose annual income is low i.e. (0-20000).
- Borrowers who take loan amounts in the range of 0 to 14000.
- Borrowers who receive interest at the rate of 15-20%.
- Borrower who takes a loan for the purpose of small business.
- Borrowers with lower Grades i.e., F<G.
- Borrower's whose subgrade is F5, G3, G5.

Thank you.