EDA : Lending Club Case Study

By: Poornima Ashok Nikam

ML C38 batch

Problem Statement:

Lending Club : Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

- A consumer finance company which specializes in lending various types of loans to urban customers, when 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 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 utilise this knowledge for its portfolio and risk assessment.
- 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.

Dataset:

- We are having the data of past customers who got approval for loan and along with their loan status , either they paid the full loan amount with interest, currently paying or they didn’t pay the loan and charged off.
- Our dataset have around 39717 customer’s past data with 111 different information of each for analysis.
- Dataset is available in “.csv” format.

Data Cleaning:

- Firstly we have filter the important data from the dataset which will be useful for case study.
- This includes:
  - We have drop the missing data
  - Kept the relevant data (with some assumption).
  - Identified the features that will contribute for our aim while dropping off the ones which doesn’t.
  - Converted the data info required data type by performing operations.
  - Derived metrics like issued month and year.
  - Filled the missing values.

Assumptions:

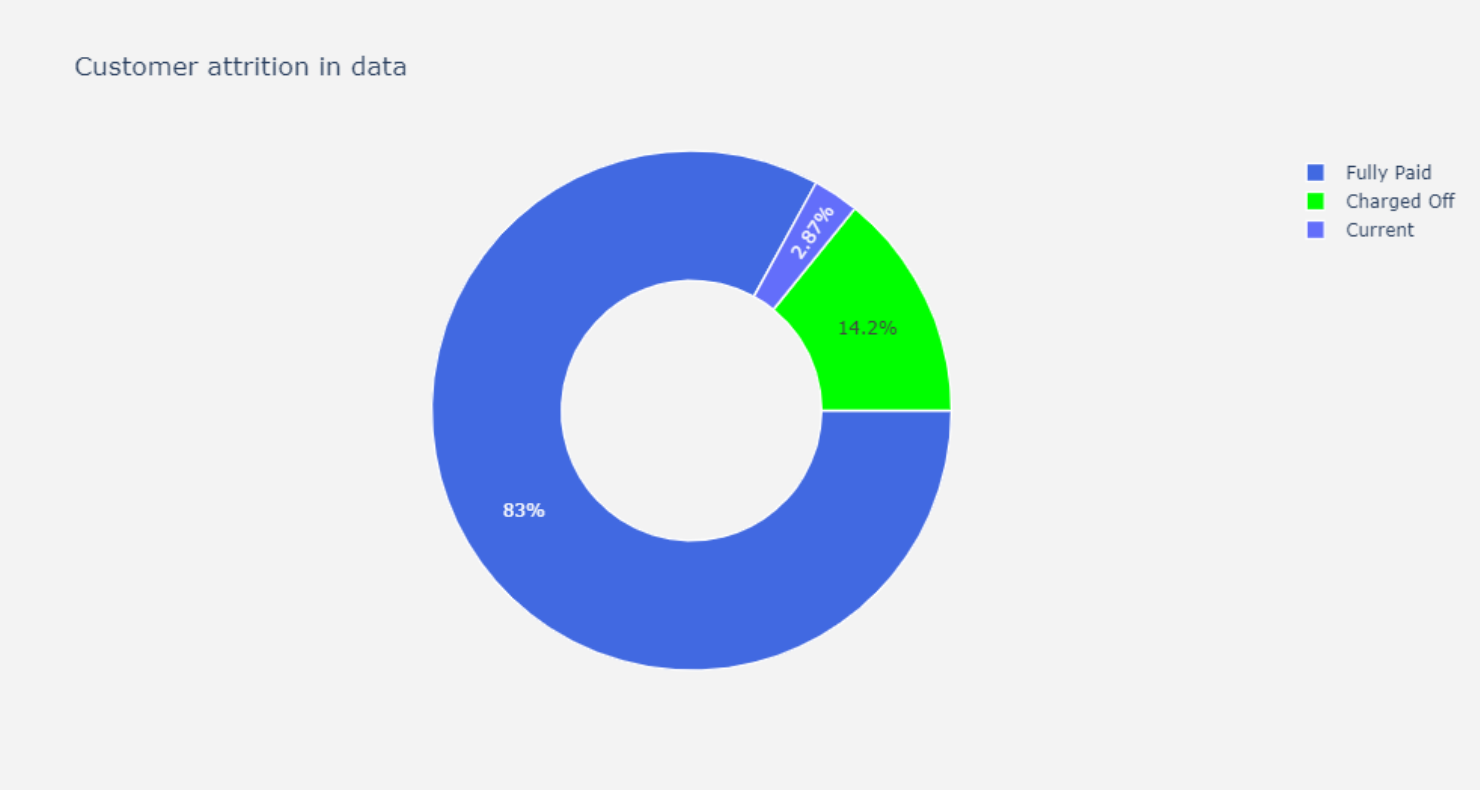
1. By going through the data dictionary, we can know there are some variables those are generated after the loan is approved such as:

   - delinq_2yrs, earliest_cr_line , inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, application_type.

- Since these variables are not available at the time of loan application, and hence **assuming that** these variables cannot be used as predictors for loan approval.

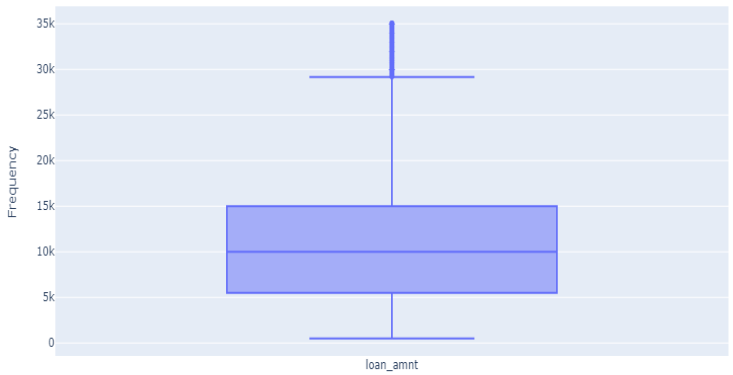
2. Source_verified and verified are **assume to** be same in verification_status column.

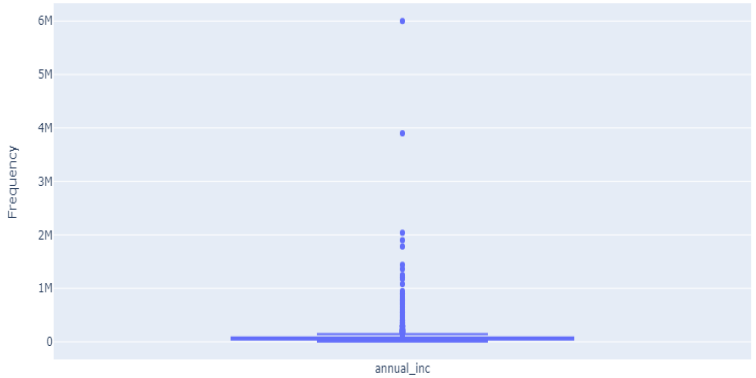
Target Variable:

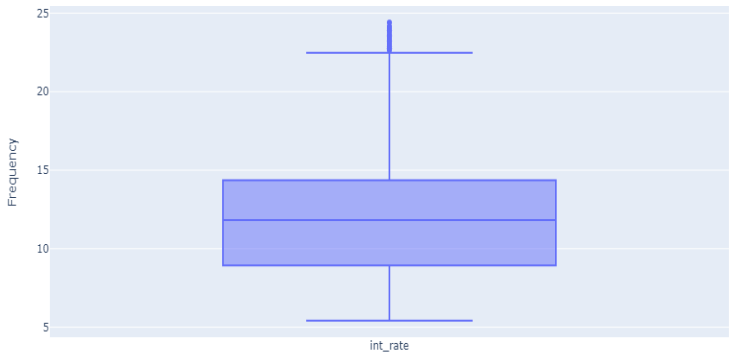


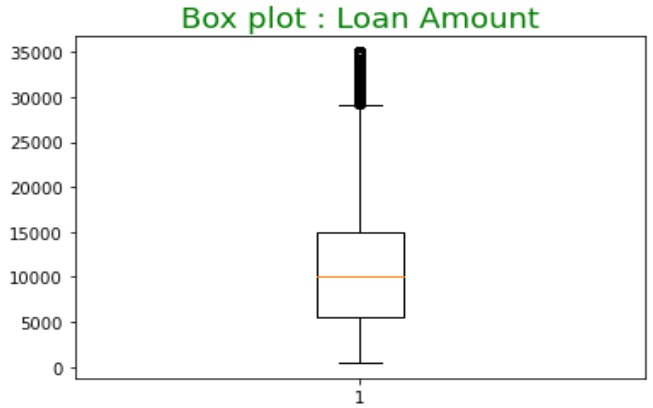
'Current' means the applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are neither labelled as 'defaulted' nor as 'fully paid'
Hence, we can drop off the records whose loan status is "Current"

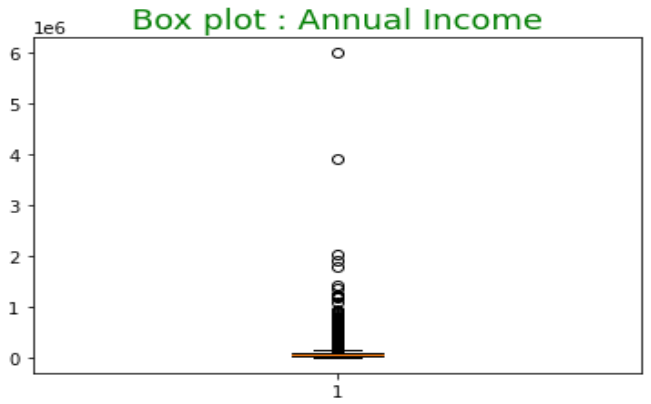
Univariate Analysis: Continuous Variables

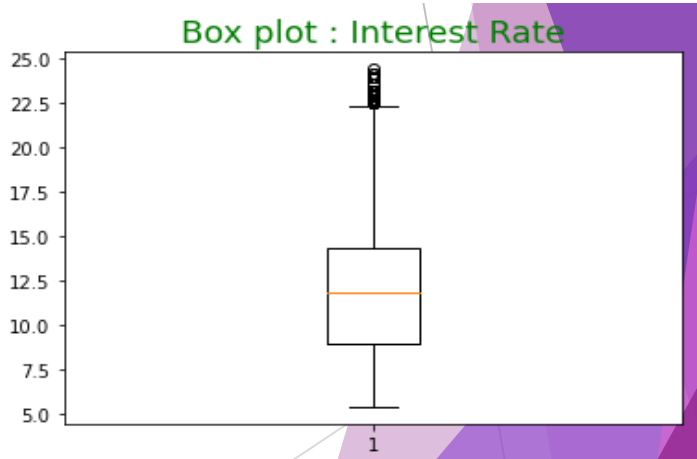






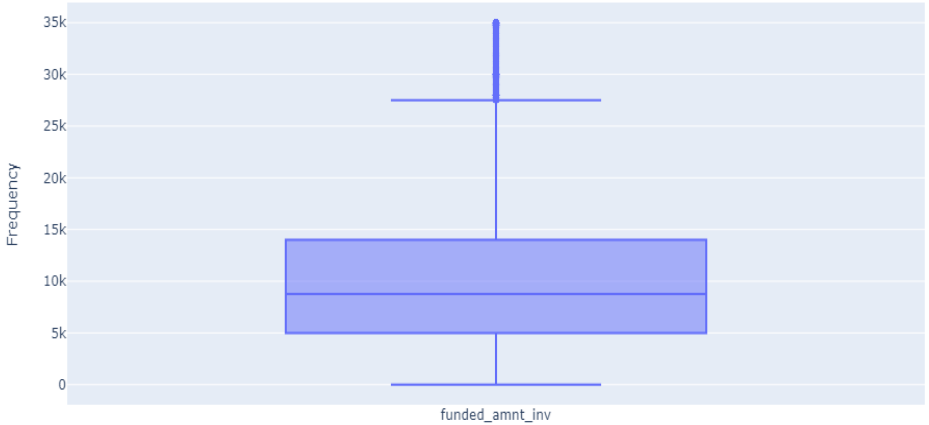


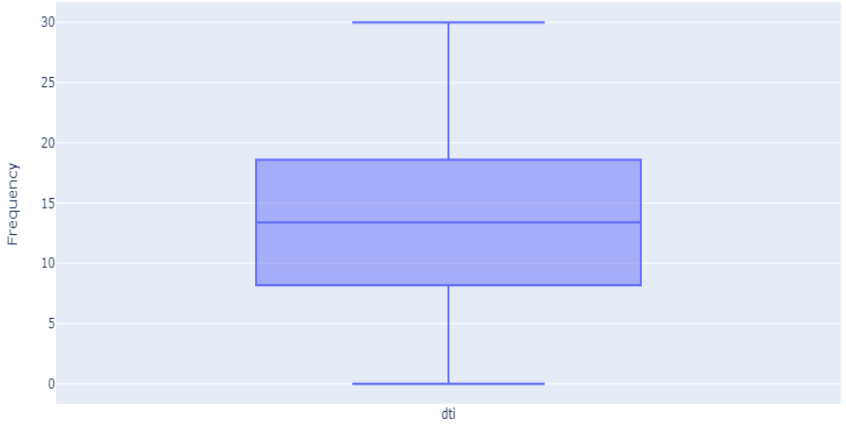


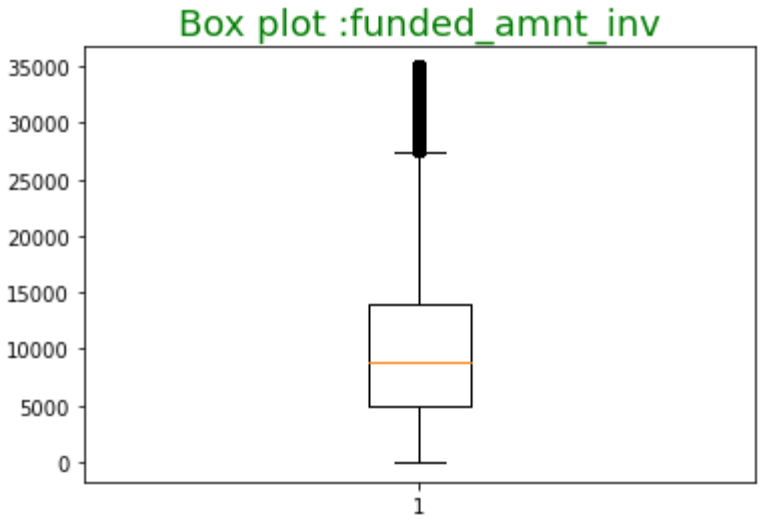


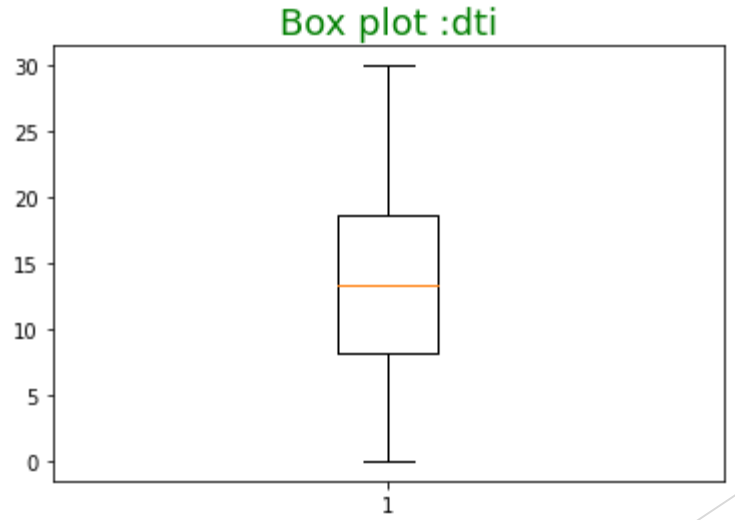
Loan Amount :Majority of loan amount is distributed in the range of 5500 and 15,000.
Annual Income : Here we can see so many outliers.
Interest Rate : Majority of interest rate is between 8.94 to 14.3575

Univariate Analysis: Continuous Variables



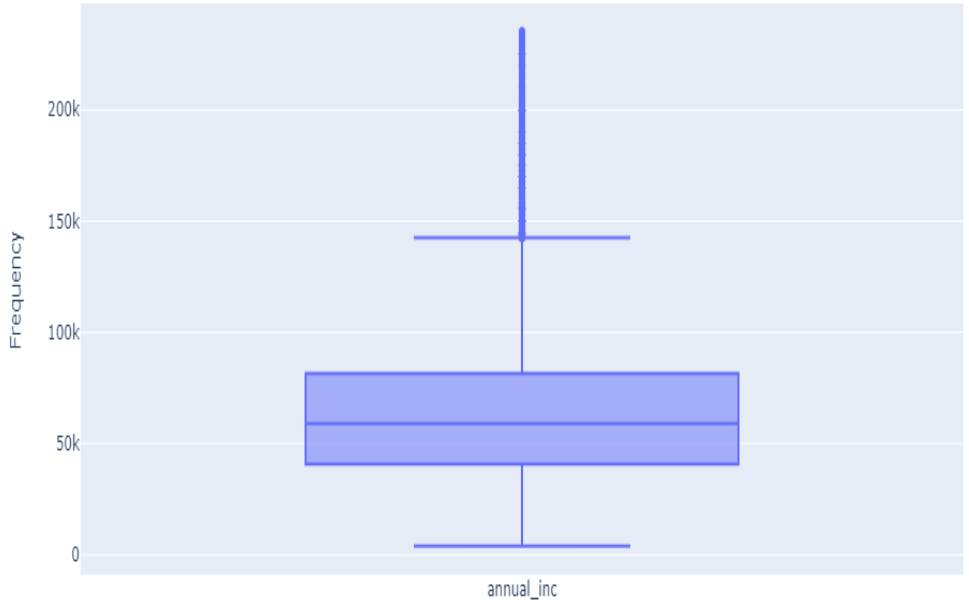


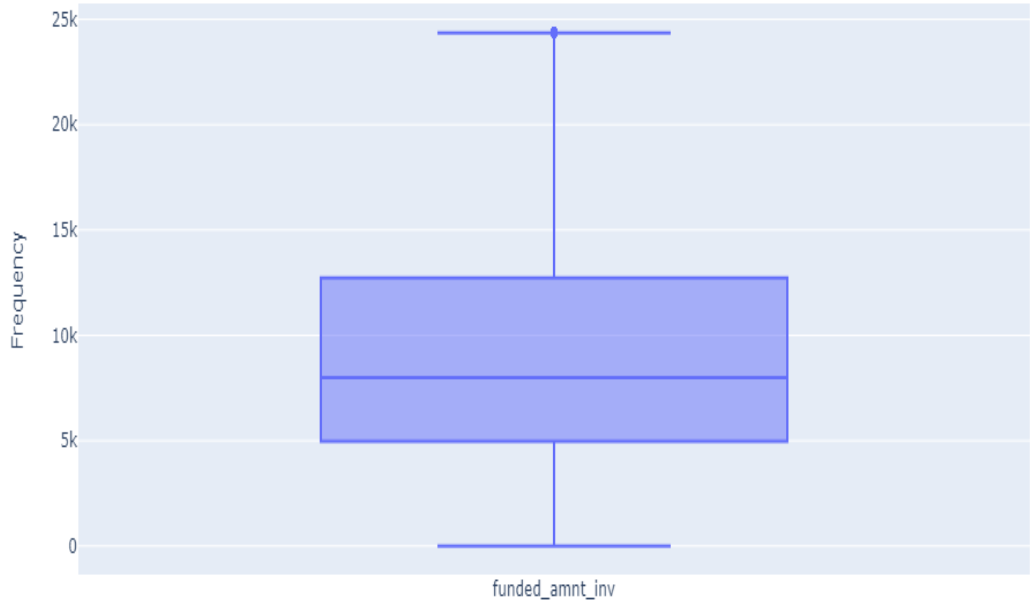




Funded_amt_inv : Here we can see lot number of outliers can be seen in "funded_amnt_inv" column.
DTI : Most of the "Debt to Income" ratio is distributed in the range of 8.24 to 18.6

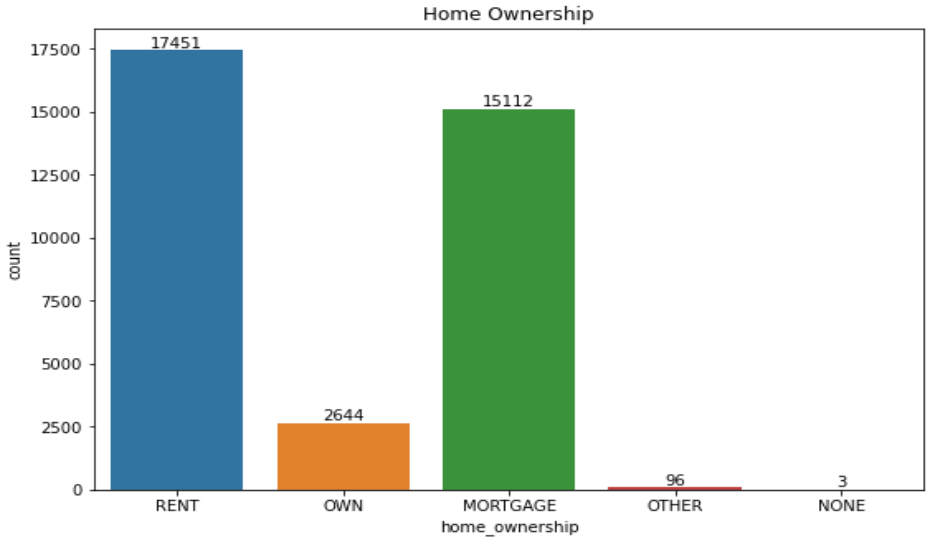
After Outlier Treatment:

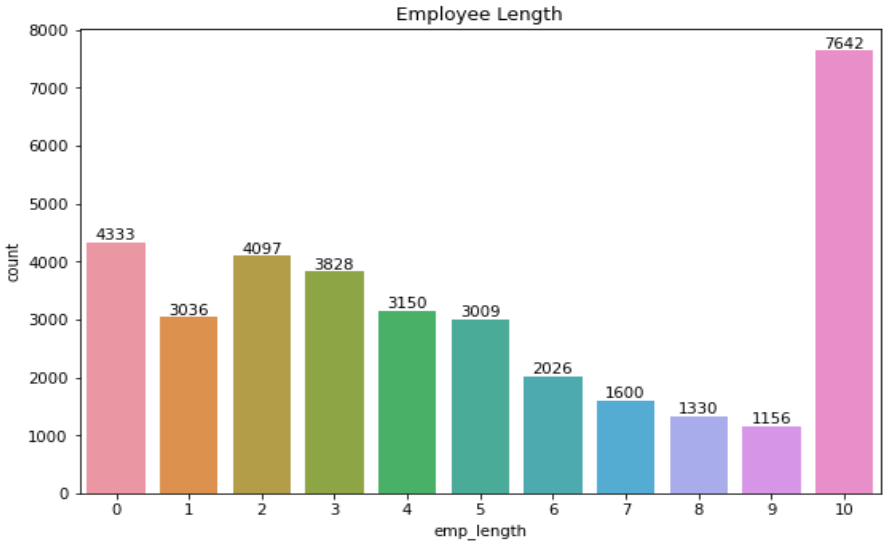


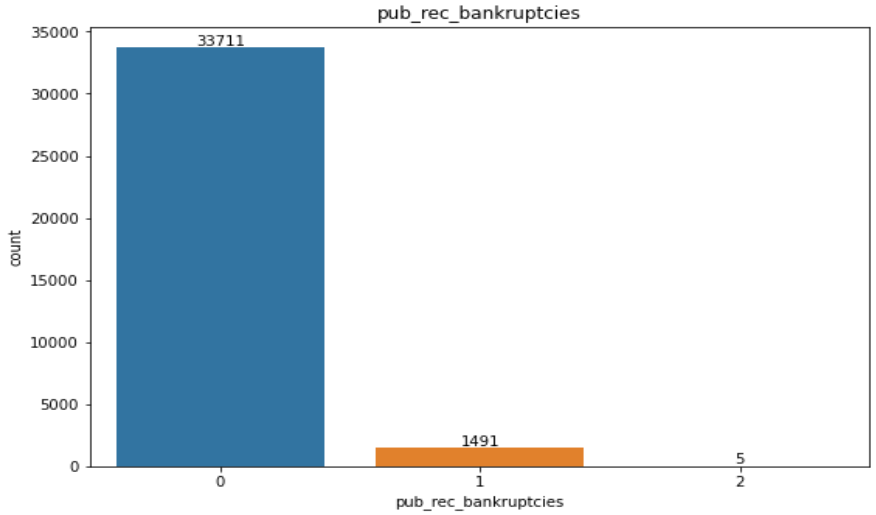


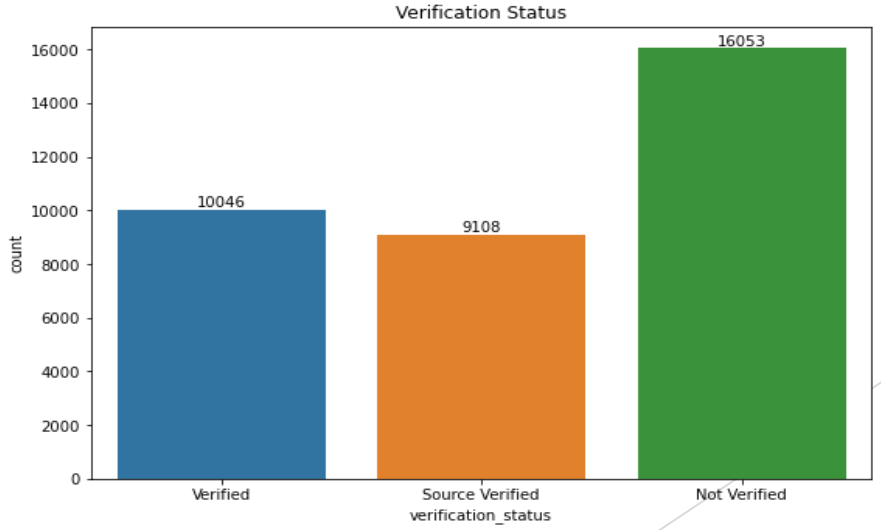
Annual Income now varies between 40.9k to 81.6k.
Majority of outliers are removed from "funded_amnt_inv" !!

Univariate Analysis: Categorical Variables



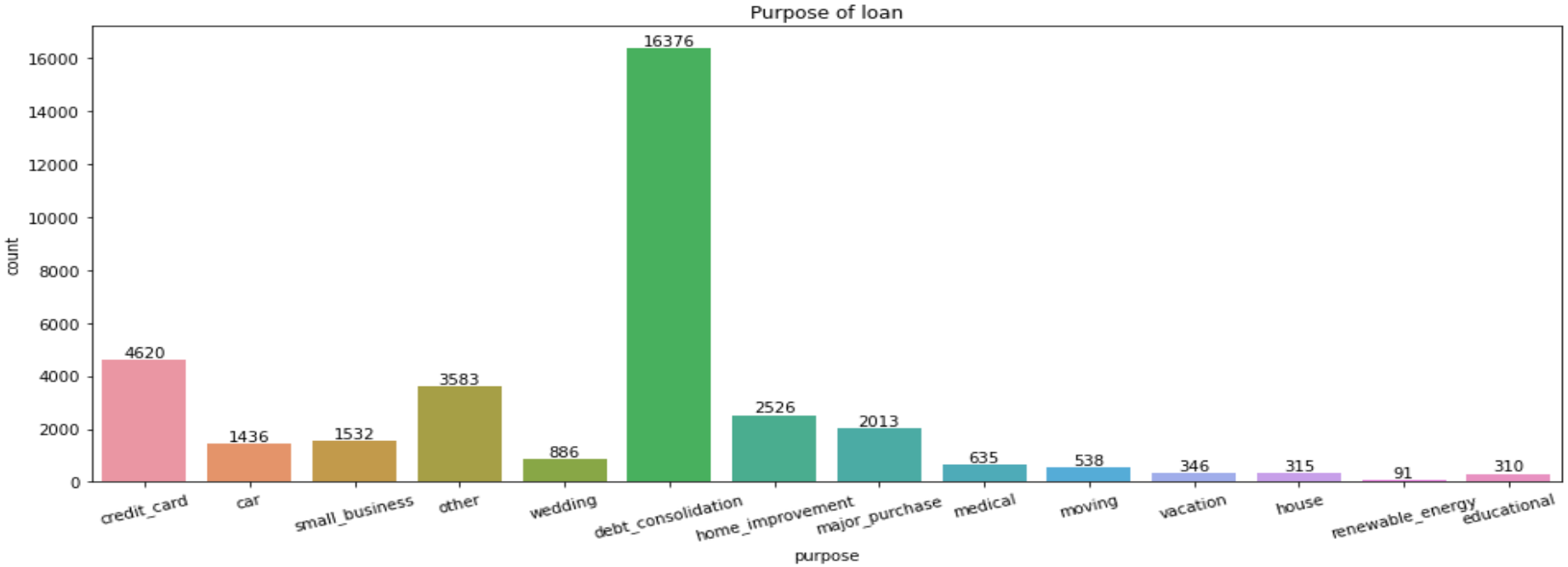


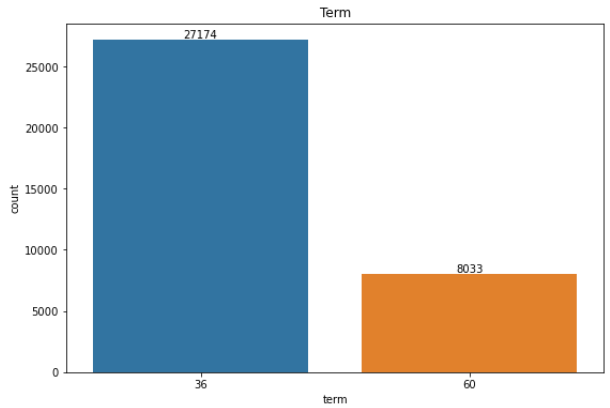


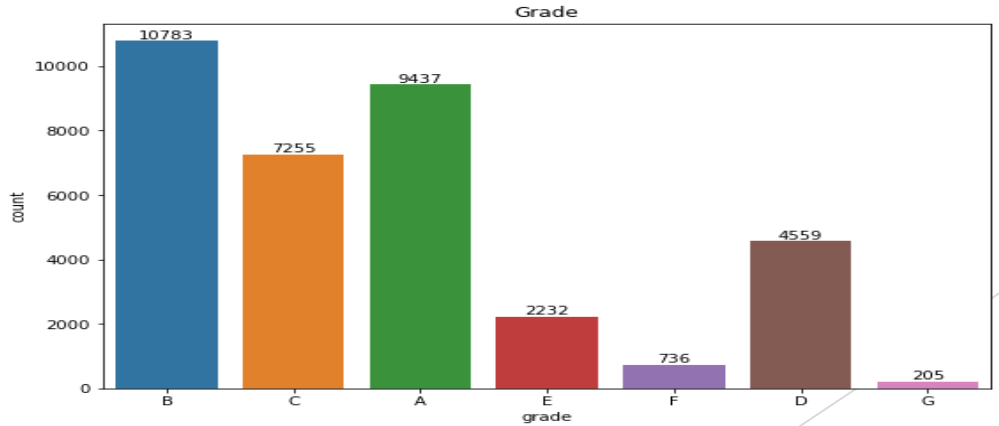


Here we can see the distribution of each categorical variables.

Univariate Analysis: Categorical Variables

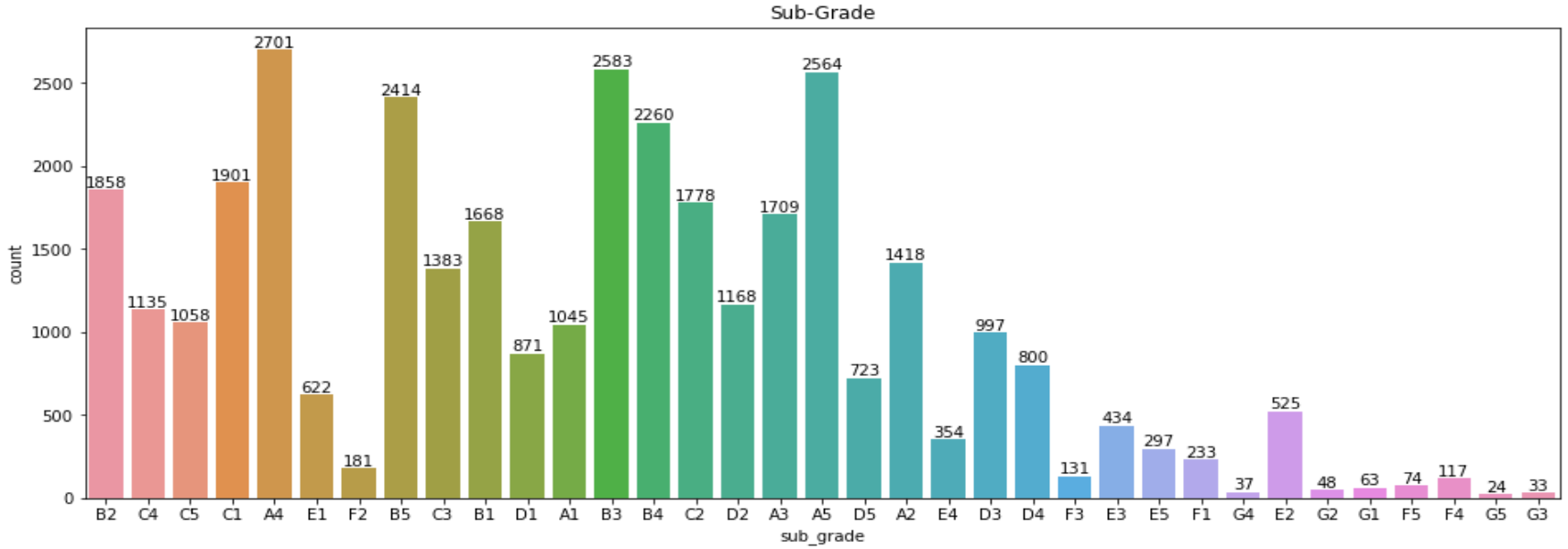


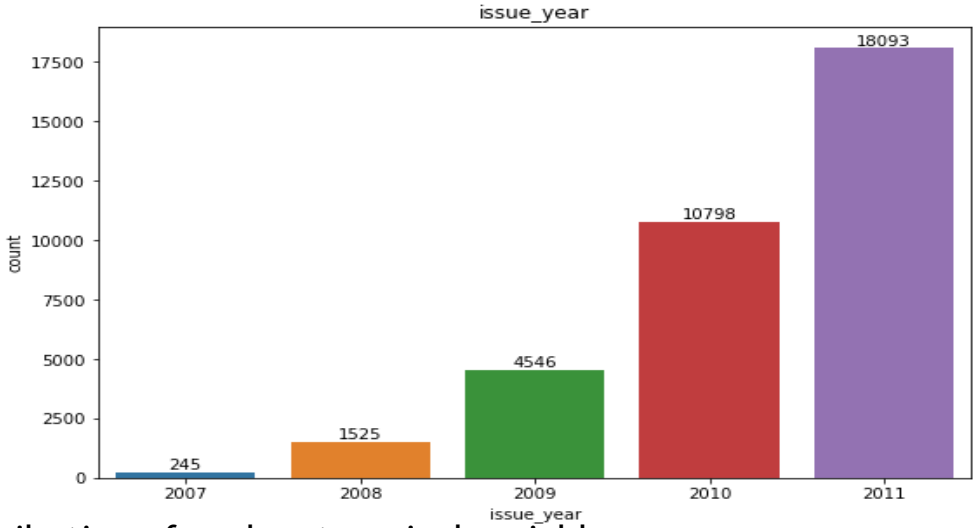




Here we can see the distribution of each categorical variables.

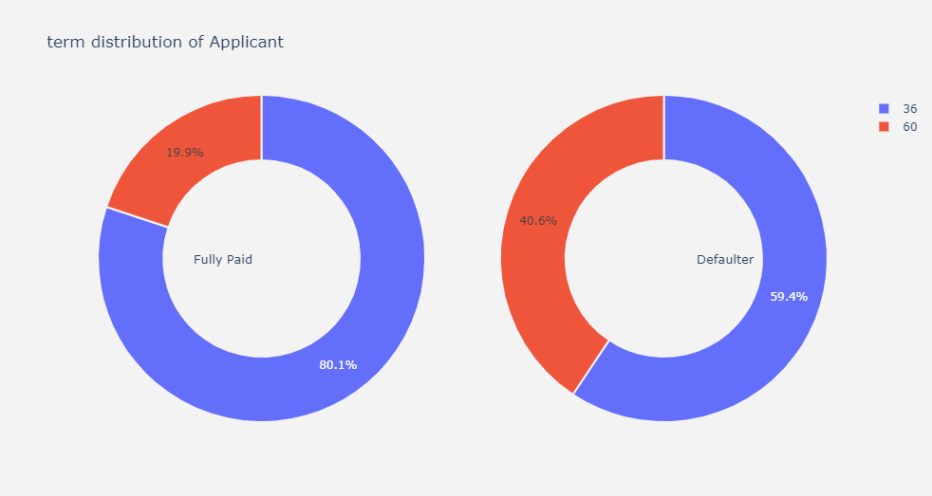
Univariate Analysis: Categorical Variables

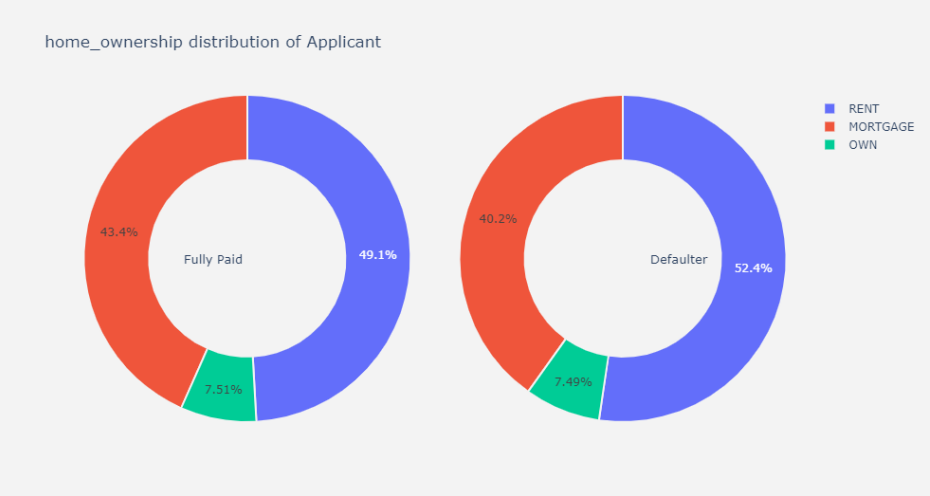


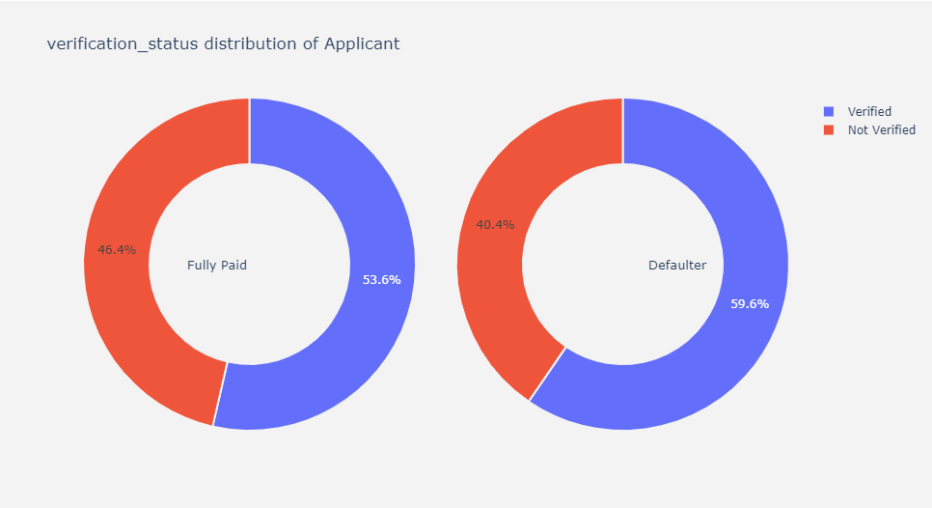


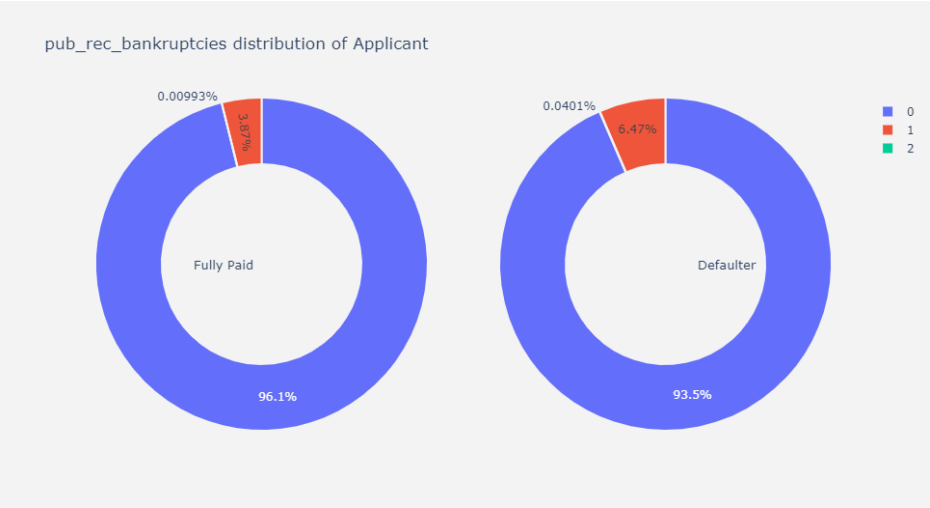
Here we can see the distribution of each categorical variables.

Bivariate Analysis: Categorical Variables



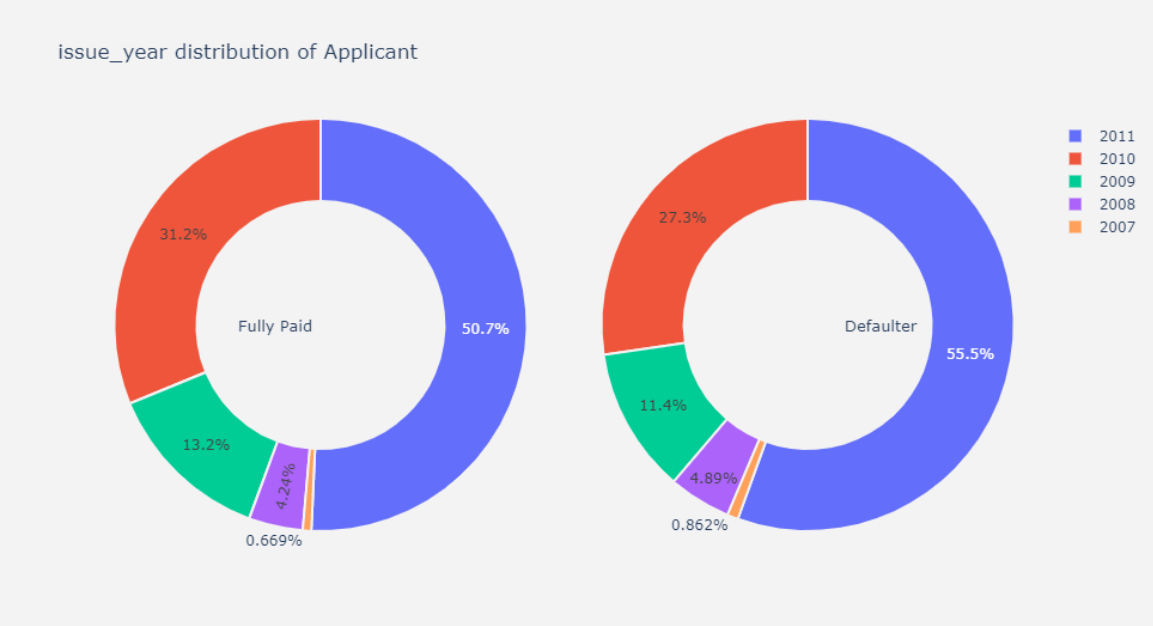






Here we can see the distribution of each categorical variables w.r.t target variable

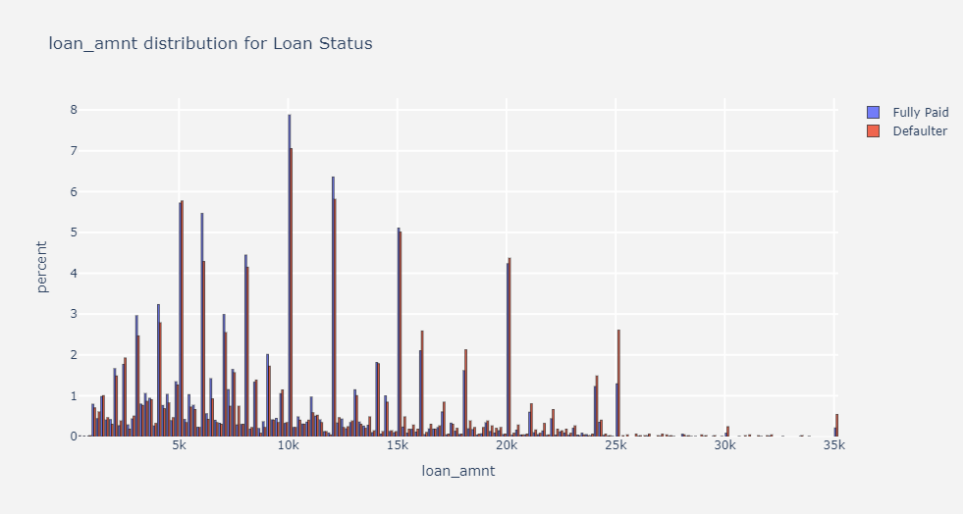
Bivariate Analysis: Categorical Variables

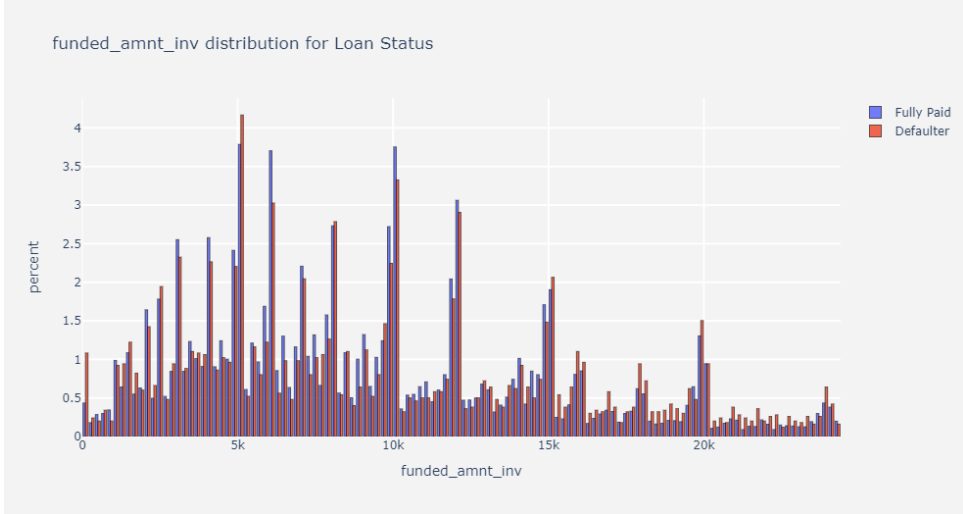


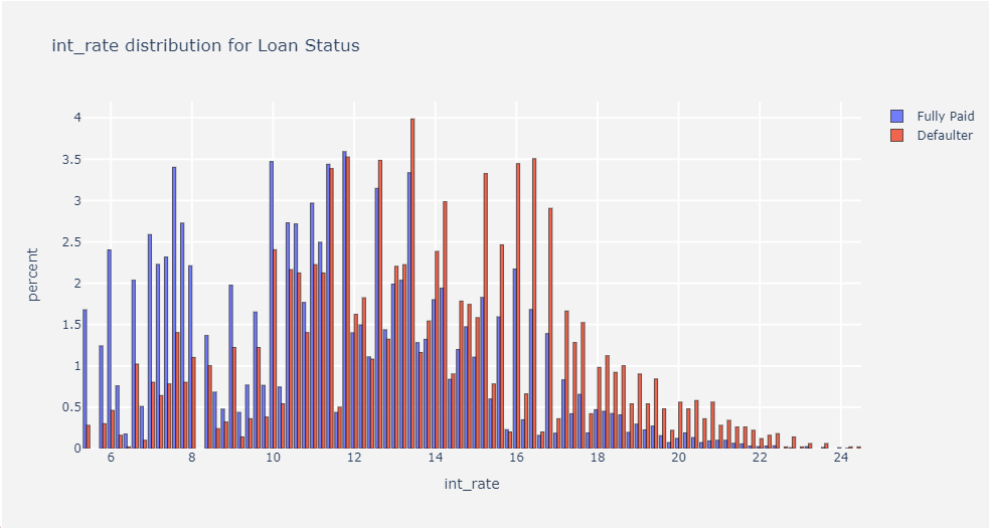
Observation of Bivariate Analysis (Categorical Variable) :

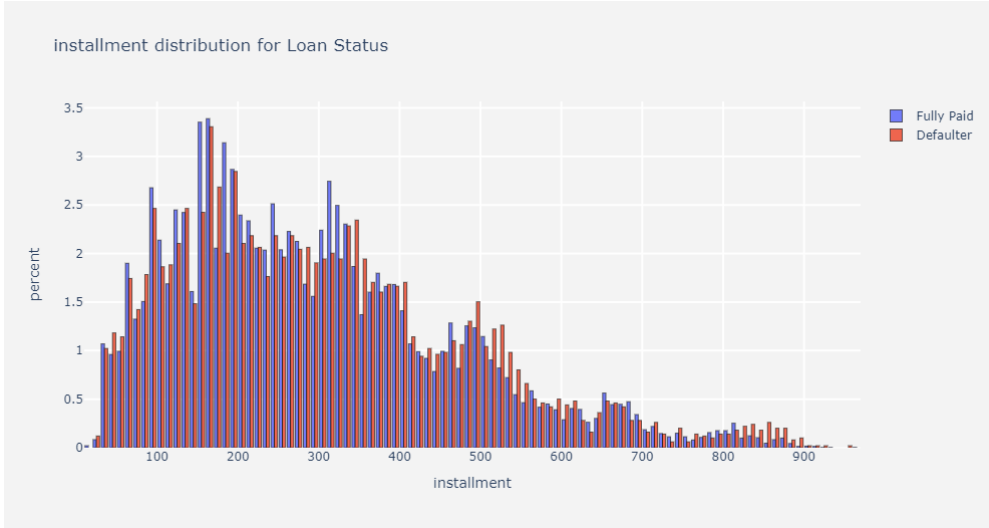
1. "Term" : Chance of defaulter is more for 60 than 30.
2. "Home Ownership" : Applicants living in Rent are more likely to default.
3. "Verification Status" : Verified applicants are more likely to default

Bivariate Analysis: Continuous Variables



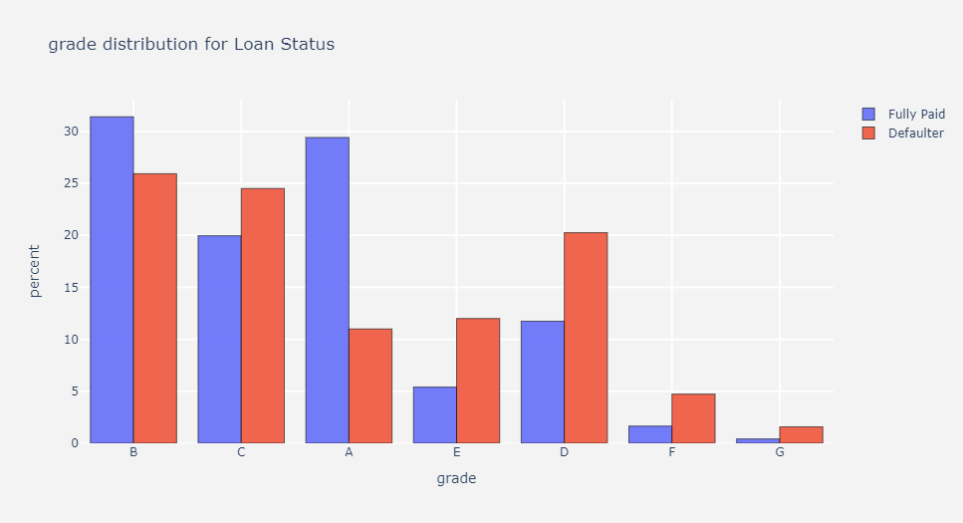


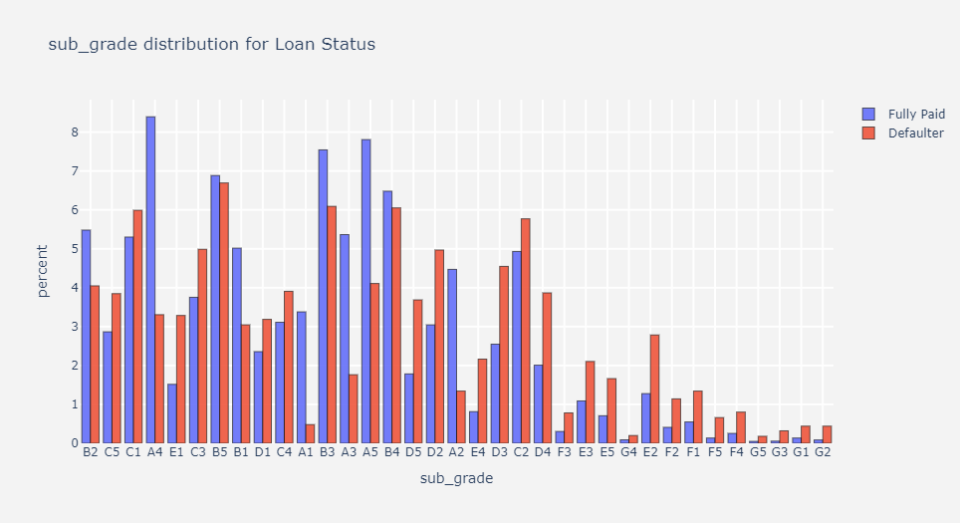


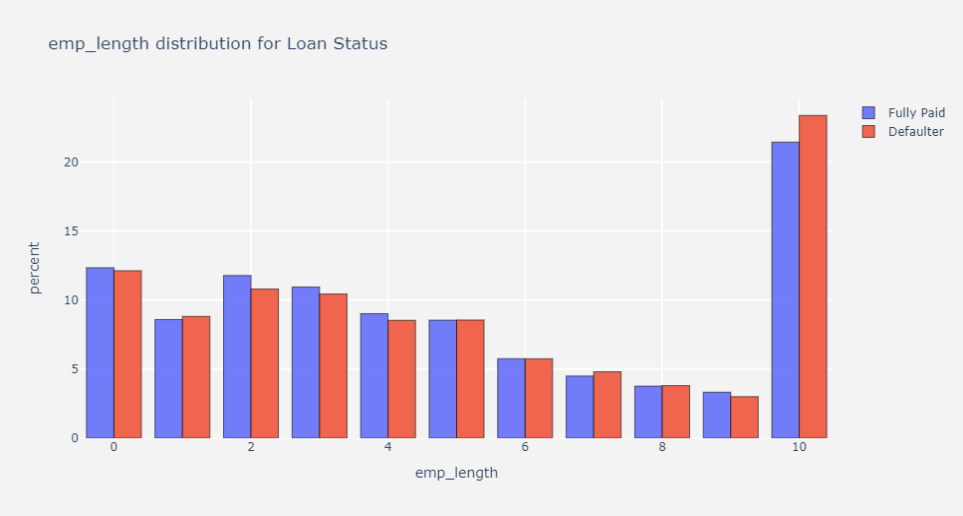


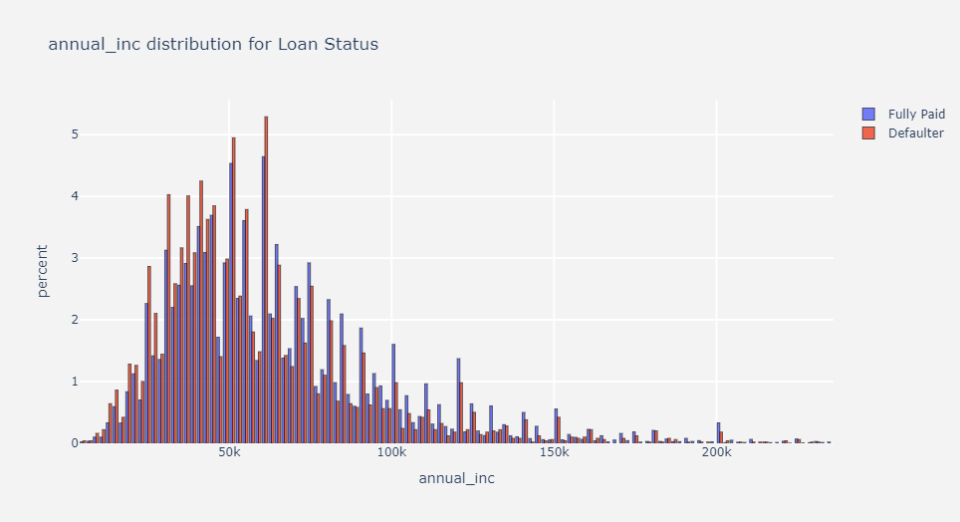
Here we can see the distribution of each continuous variables w.r.t target variable

Bivariate Analysis: Continuous Variables



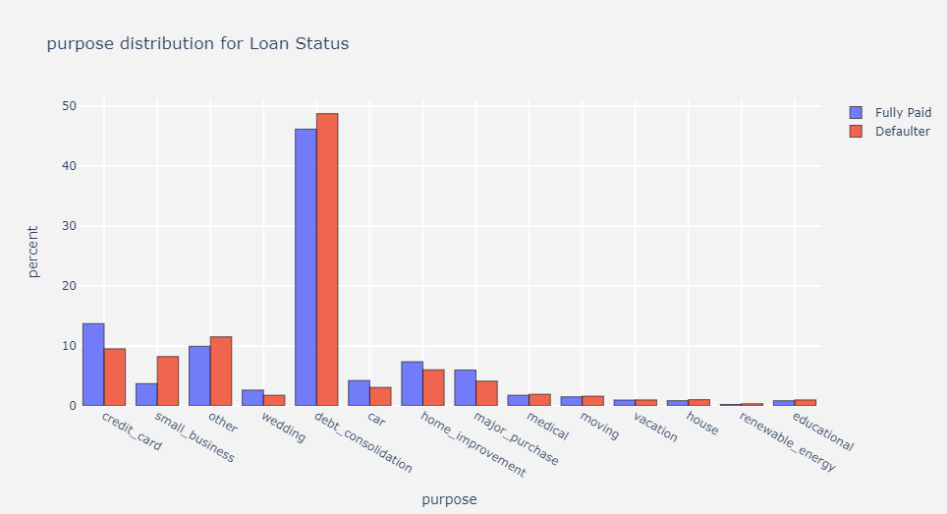


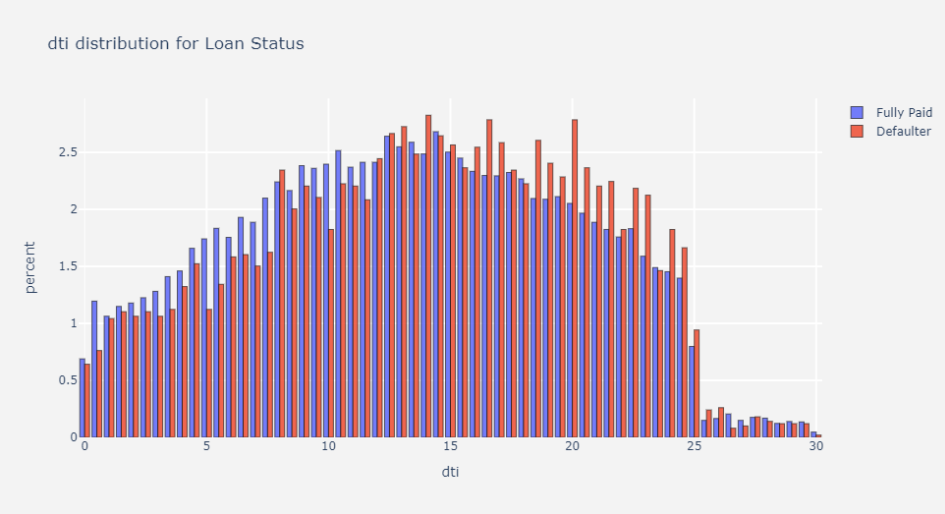




Here we can see the distribution of each continuous variables w.r.t target variable

Bivariate Analysis: Continuous Variables

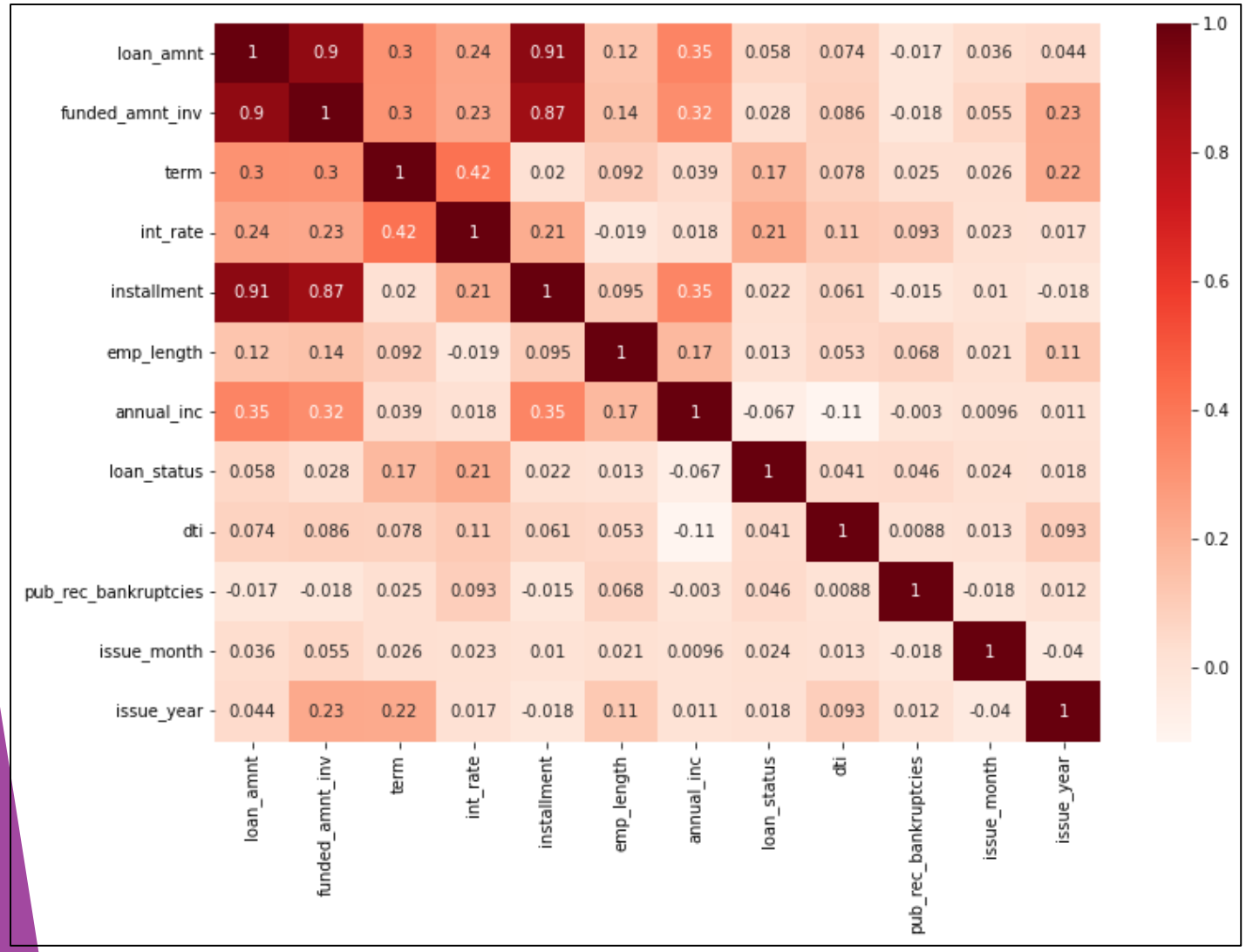




Observation from Bivariate analysis (continuous Variable) :

1. "Grade" : For grades C, D, E, F, G defaulter rate is more than paid ones.
2. "loan amt" : As loan amt increases, nnumber of defaulter increases (directly proportional).
3. "funded_amt_inv" : No such correlation can be concluded.
4. "int_rate" : Interest rate is directly proportional to defaulters(as int_rate increases, defaulters also increases)
5. "installment" : As installment increases, defaulters increases after certain limit.
6. "Emp lenghth" : Charged off percentage is almost constant.
7. "Annual income" : Annual income is indirectly proportional to defaulters (Annual income less==>more defaulters)
8. "Purpose:" : small business have more default ("debt con", "other" also defaults but small business % is more than double)
9. "DTI" : As dti increases , defaulters also increases.

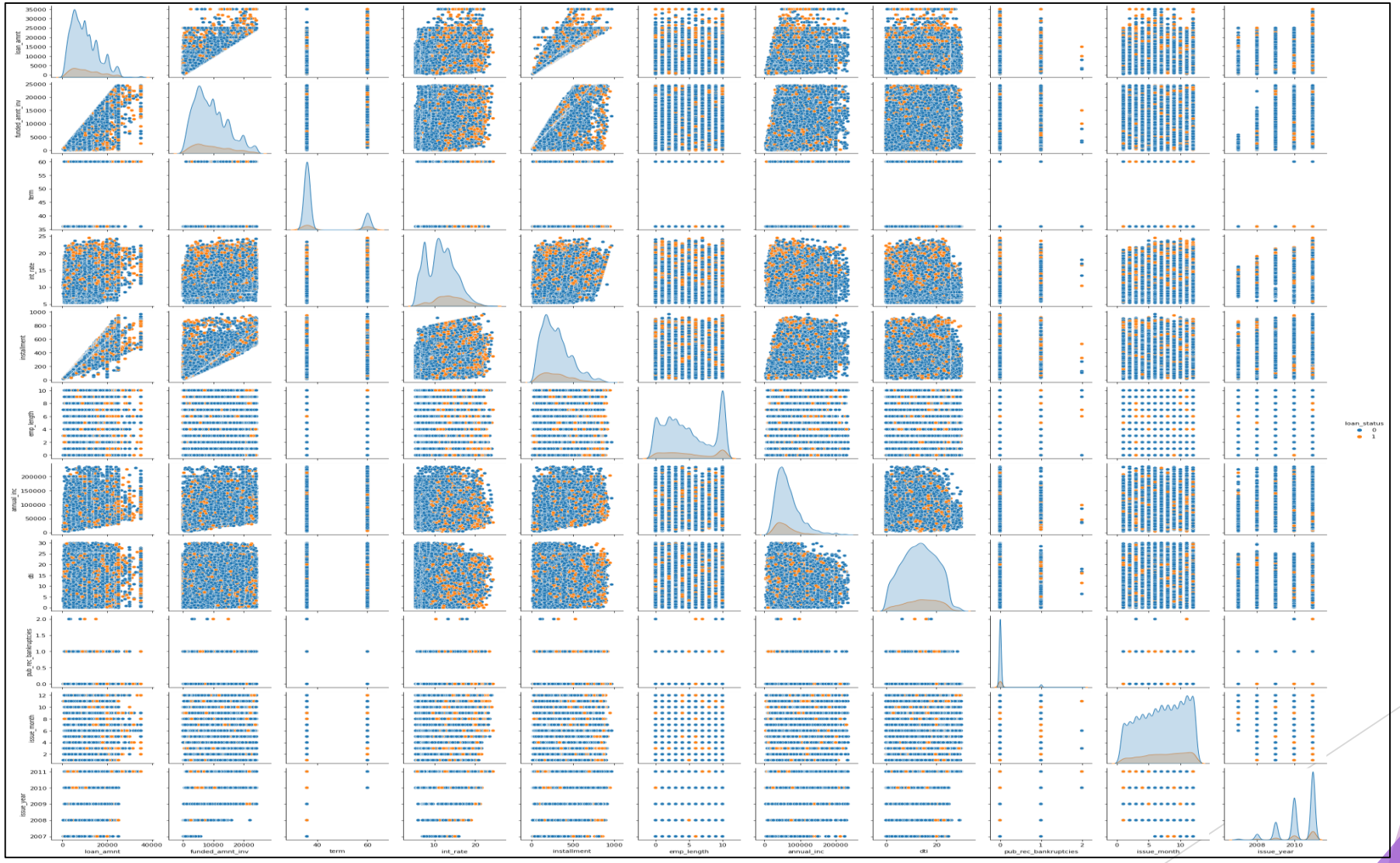
Multivariate Analysis: Heatmap and Correlation



Observation :

- "loan_amnt" have good positive correlation with "installment" and "funded_amnt_inv" variable i.e., "loan amount" will increase with increase in "installment" and "funded_amnt_inv".
- "funded_amnt_inv"have positive correlation with "installment".
- Other have very poor correlation so there is not much linear relationship between them

Multivariate Analysis: Pairplot



Conclusion

- At the end of Exploratory Data Analysis (EDA) part, we have identified the top most important features from the dataset (out of 111 features) given below:
  1. Annual income
  2. Loan Amount
  3. Purpose
  4. Term
  5. Grade
  6. Interest Rate
  7. Installment
  8. Debt-to-income
- These top 8 features will help the business to minimize the defaulters rate and this **will increases the business financial growth by approving loan of good customers who repays the loan within time.**

Thank You !!!