Lending Club

Case Study by

Rahul Nanwani and Nitin Katiyar

Agenda

- Introduction
- Problem Statement
- Approach
- Conclusion

Introduction

When a loan application is received, the company must make a loan approval decision based on the applicant's profile. The bank's decision is associated with two types of risks:

- If the applicant is likely to repay the loan, then the company loses business by not approving the loan.
- If the applicant is unlikely to repay the loan, i.e. if he or she will default, approving the loan may result in a financial loss for the company.

Introduction (contd.)

1. Loan accepted: If the loan is approved by the company, there are three possible outcomes as described below:
    a. Fully paid: The applicant has paid off the loan in full (the principal and the interest rate)
    b. Current: The applicant is in the process of paying the installments, so the loan's tenure has not yet been completed. These candidates are not marked as 'defaulted'.
    c. Charged-off: The applicant has not paid the installments on time for an extended period of time, indicating that he or she has defaulted on the loan.

Introduction (contd.)

2. Loan rejected: The loan had been rejected by the company (because the candidate does not meet their requirements etc.). Because the loan was rejected, the applicants have no transactional history with the company, and thus this data is not available to the company (and thus in this dataset)

Problem Statement

Assist the company to minimize the risks and losses by making use of exploratory data analysis and studying the dataset provided to identify which categories of people will not default and which should be avoided when making loan decisions.

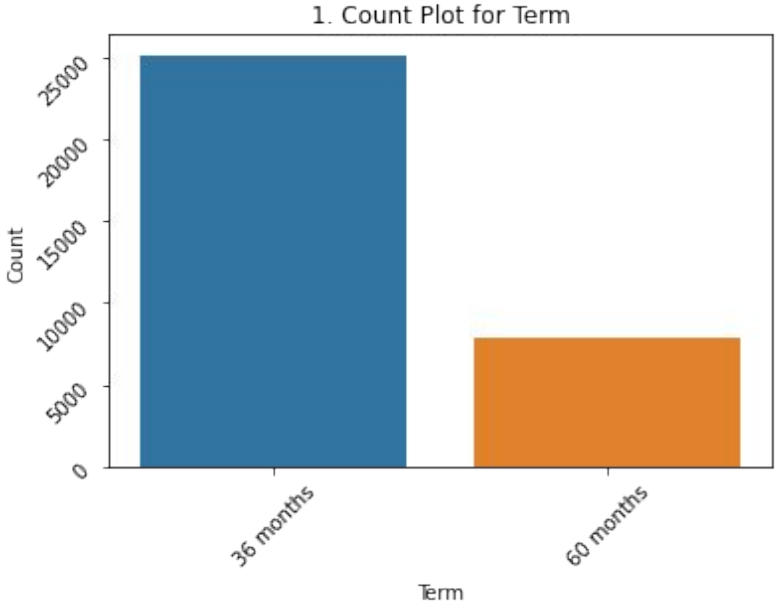
Approach

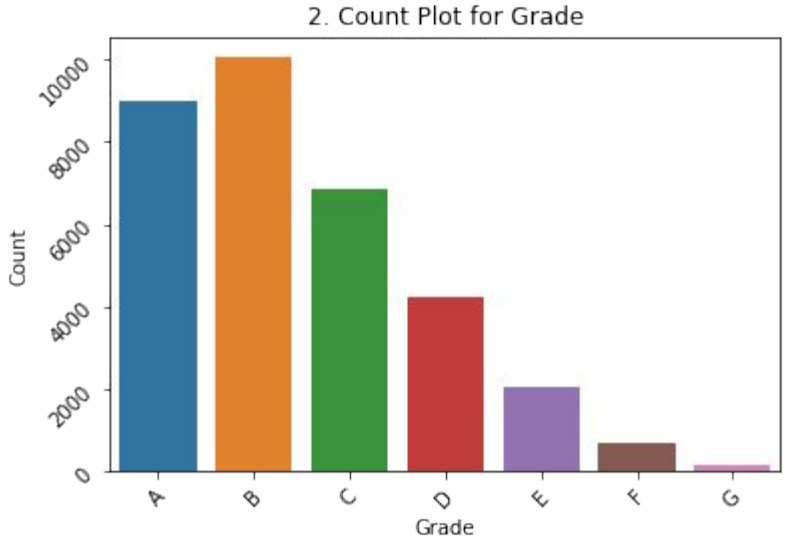
Data Sourcing	Data Cleaning	Data Analysis	Conclusion
We already have the dataset provided. We need to understand this dataset that has 39717 rows and 111 columns.	The dataset provided may have null values or shifted values or junk values. We need to deal with these values.	We need to conduct data analysis using univariate, bivariate analysis, deriving new metrics from the data, etc.	From the data we have analysed making decisions would be easier just by looking at the plots.

Data Cleaning

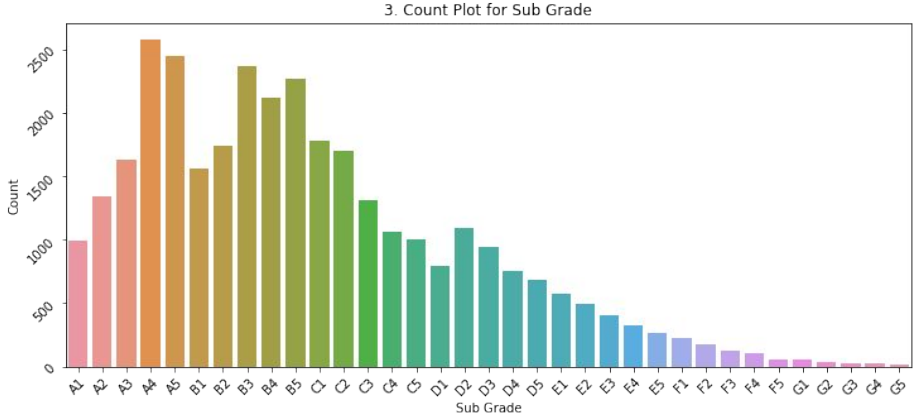
- We start by filtering, ie. we remove the irrelevant data.
- Next, we need to handle the missing values, since the proportion of the missing values is very low, it is good to drop these records.
- Now, we need to standardize and clean the data.
- Finally we need to handle the outliers, the best way to do so in our case would be to remove them directly.

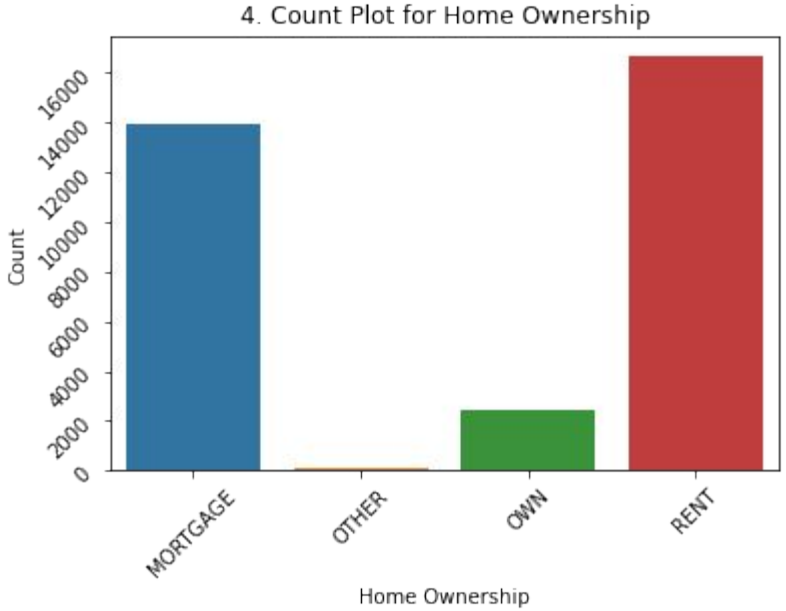
Data Analysis



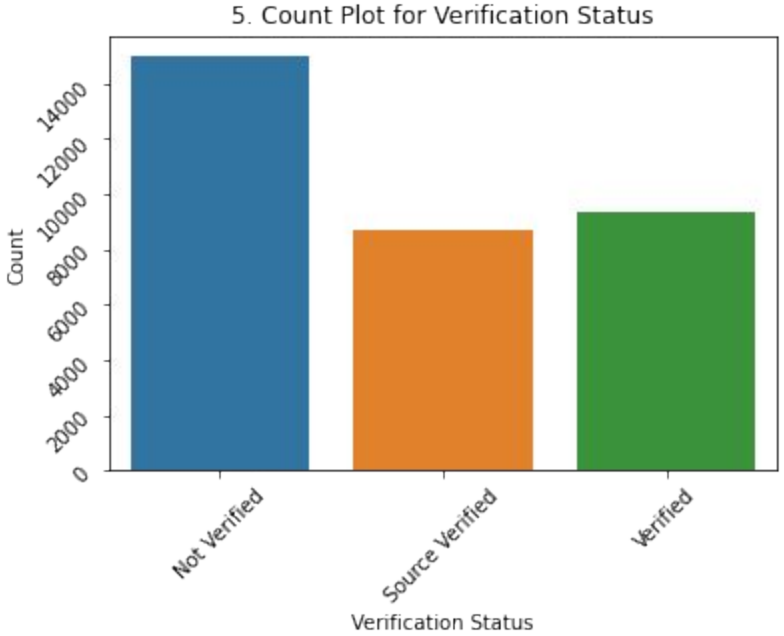


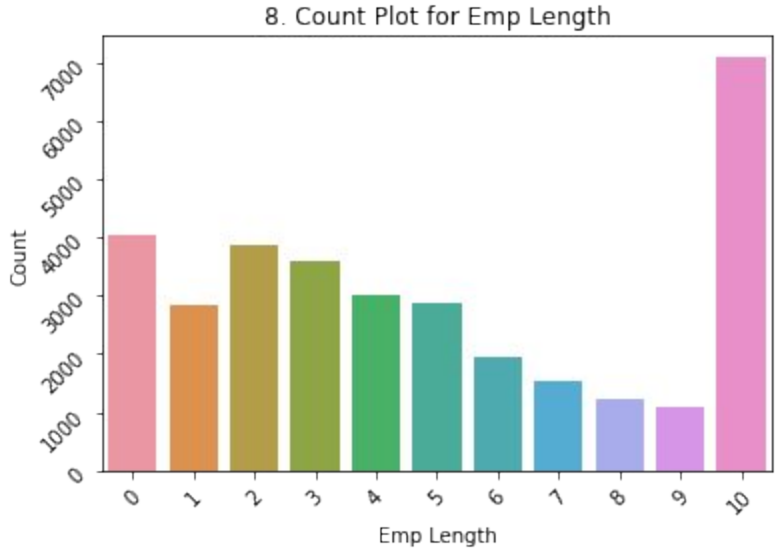
We can consider term and grade for our analysis, with a note that F and G grades has low data.



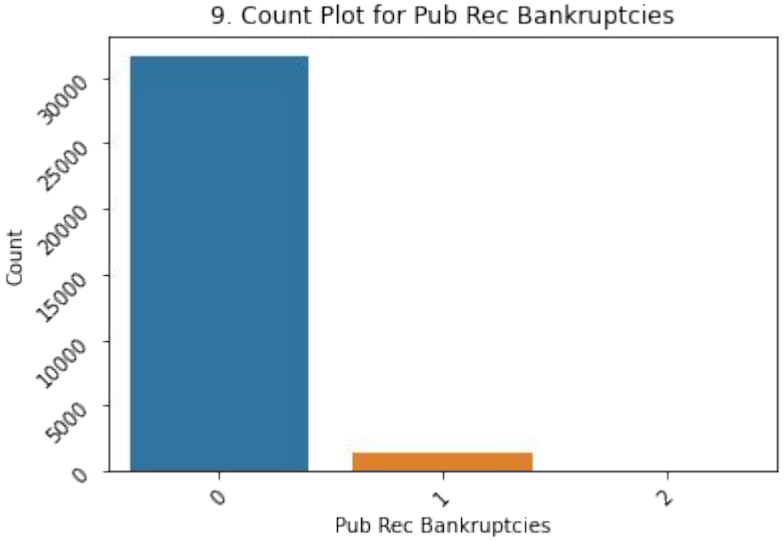


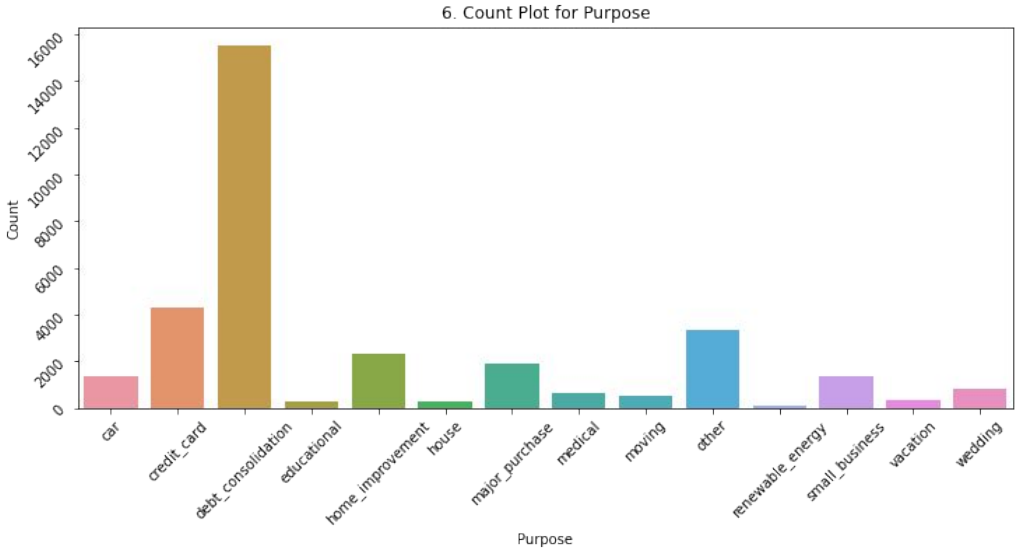
- We need to ignore sub-grade as many of the grades have less data.
- We can consider home ownership with a note that other should be excluded in our analysis.



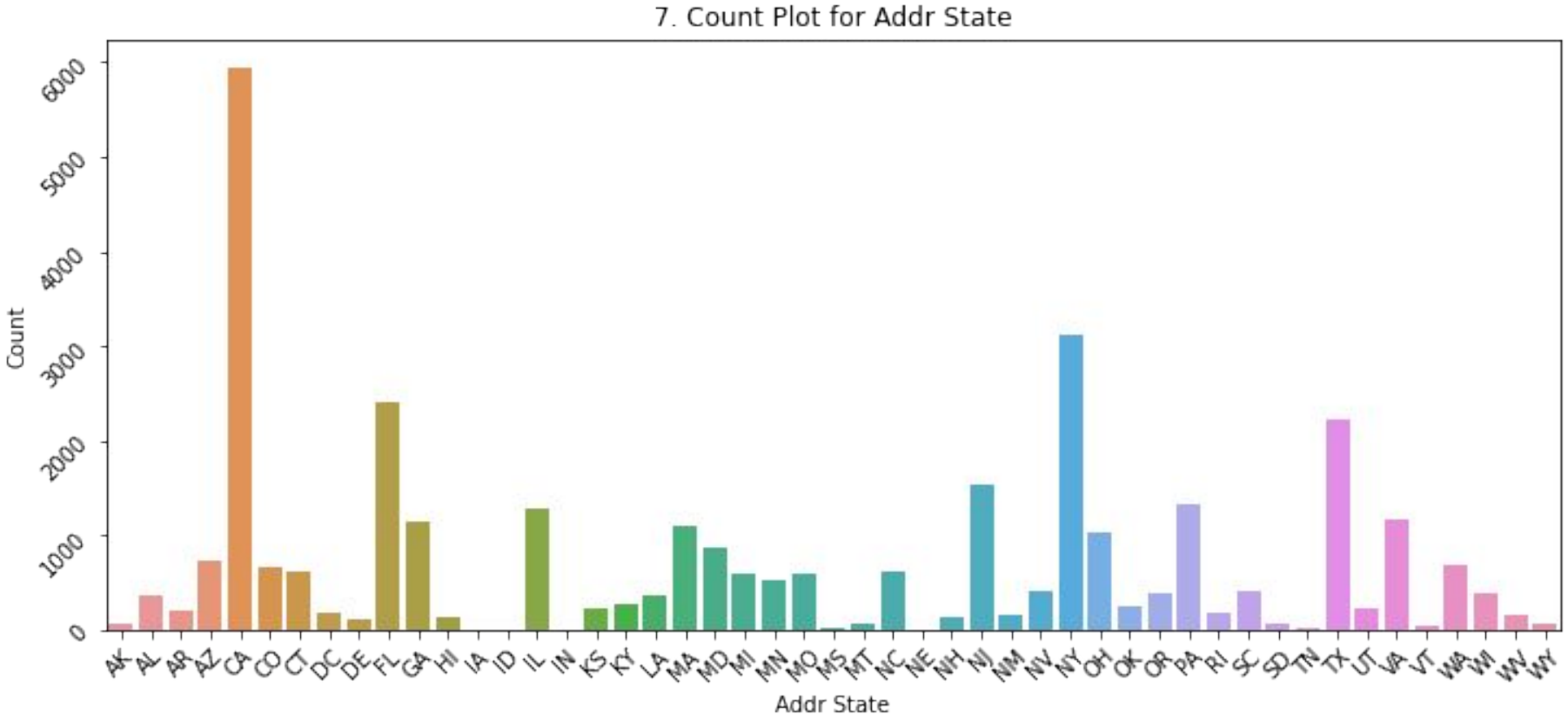


We can include verification status and employment length in our analysis.

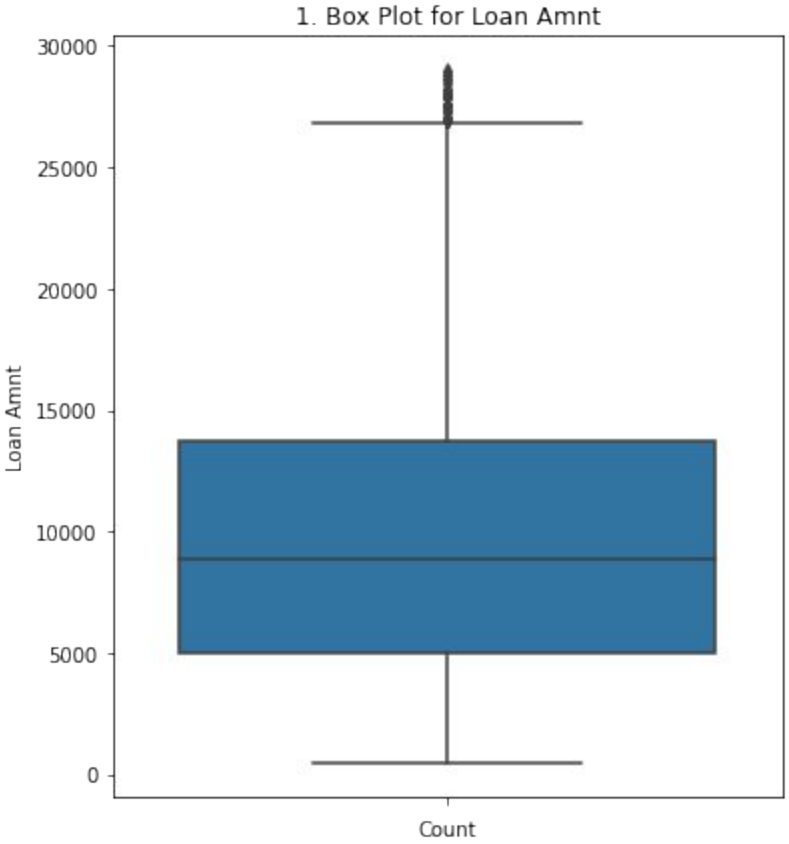


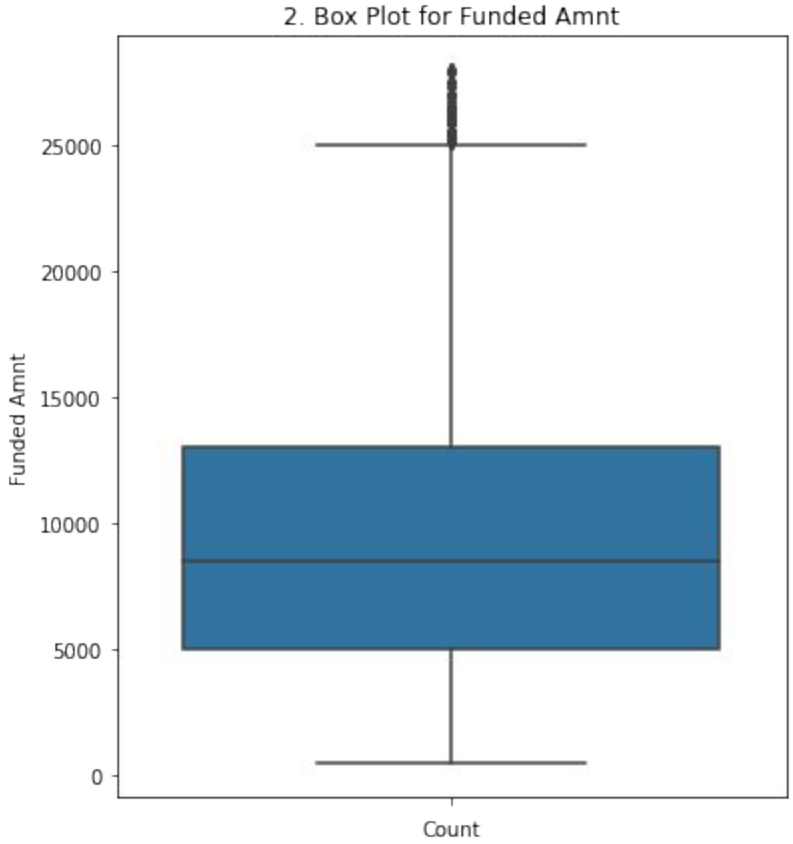


We need to ignore both of these variables as they have concentrated data for just one value, this may not give us correct analysis.

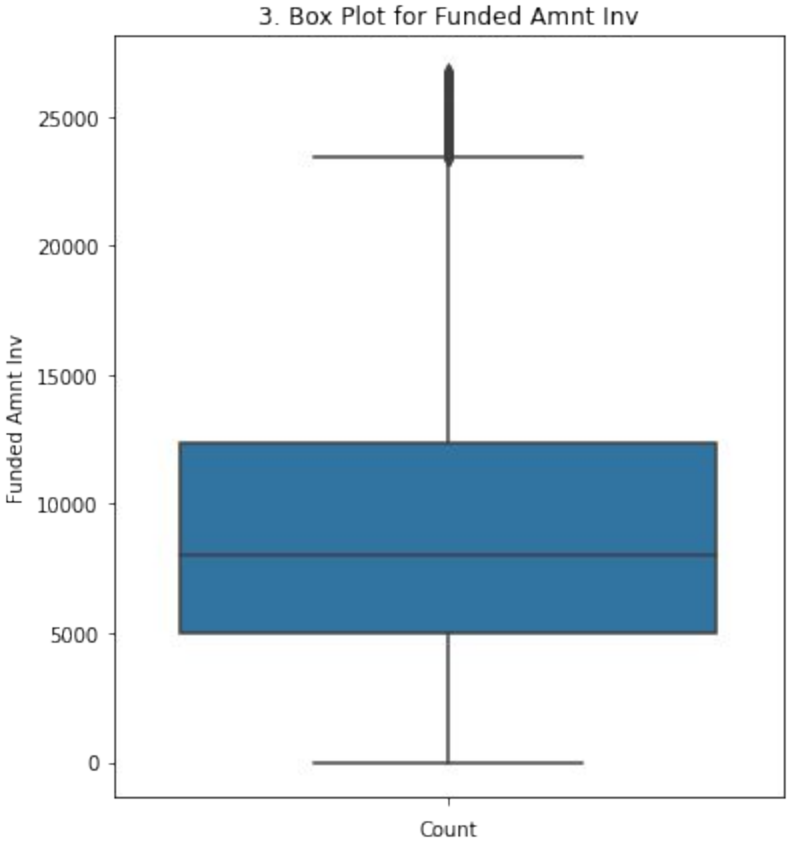


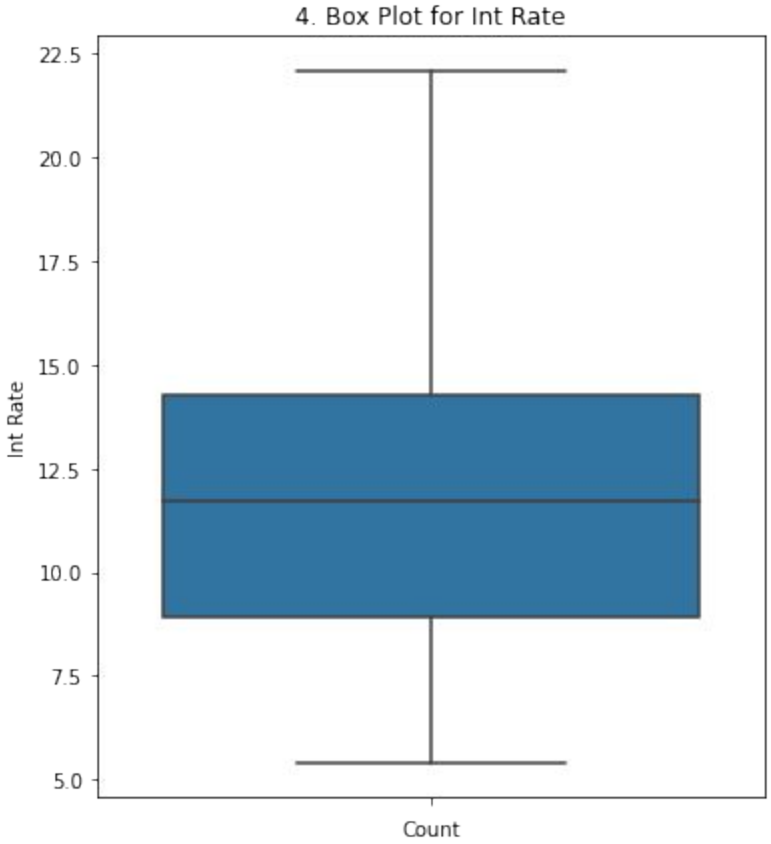
We need to ignore address state as it has concentrated data for just one value, this may not give us correct analysis.



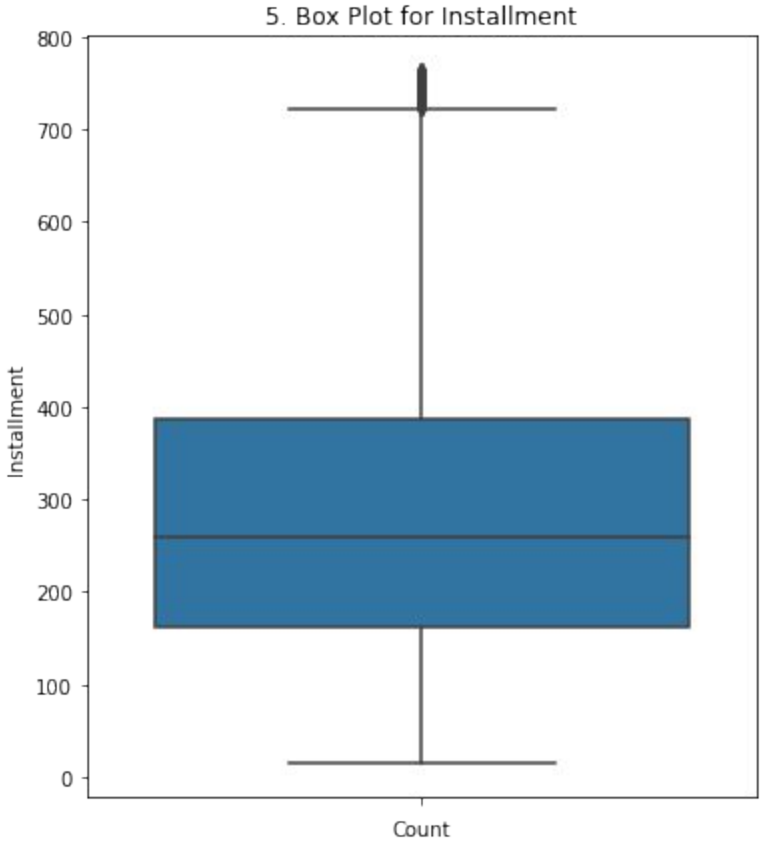


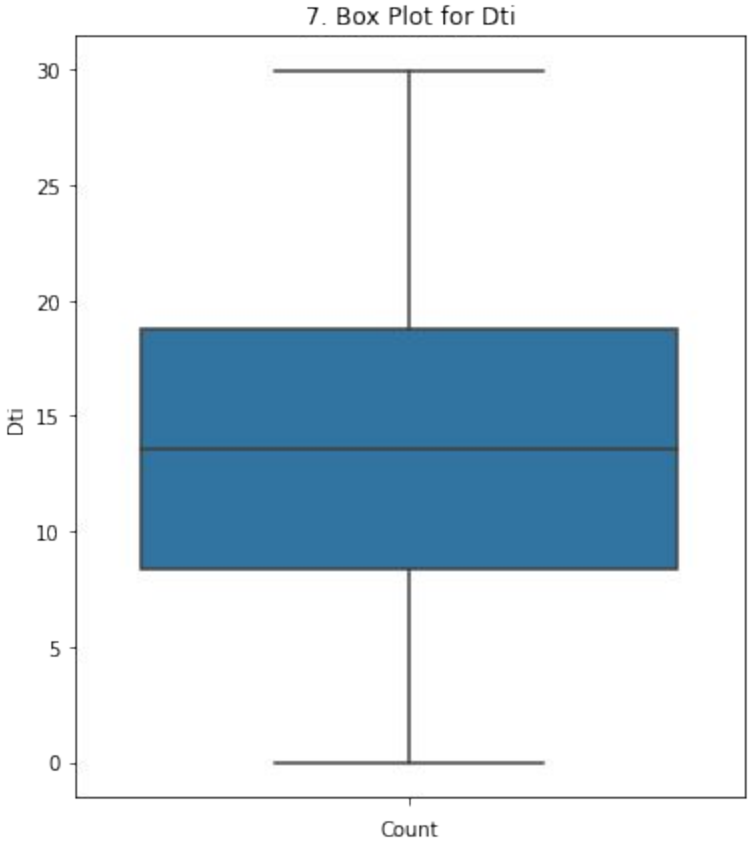
We cannot say much about these fields as it has many outliers.



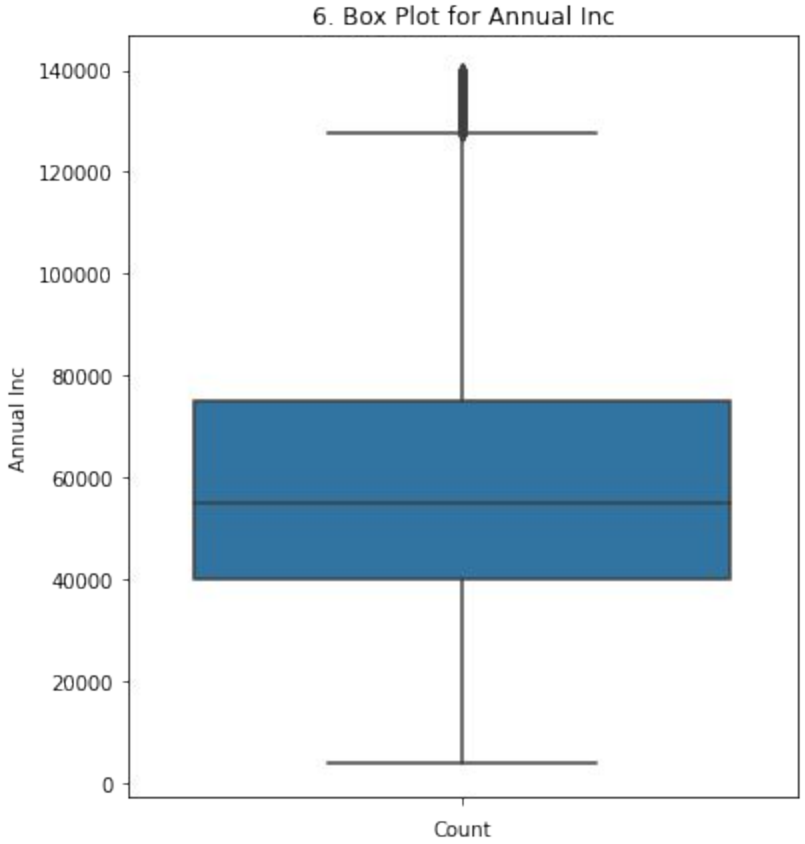


We can use interest rate for our analysis

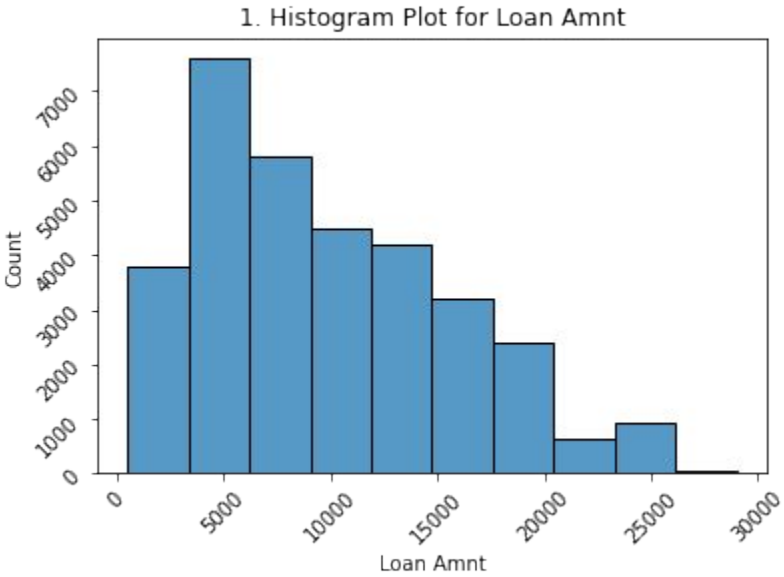


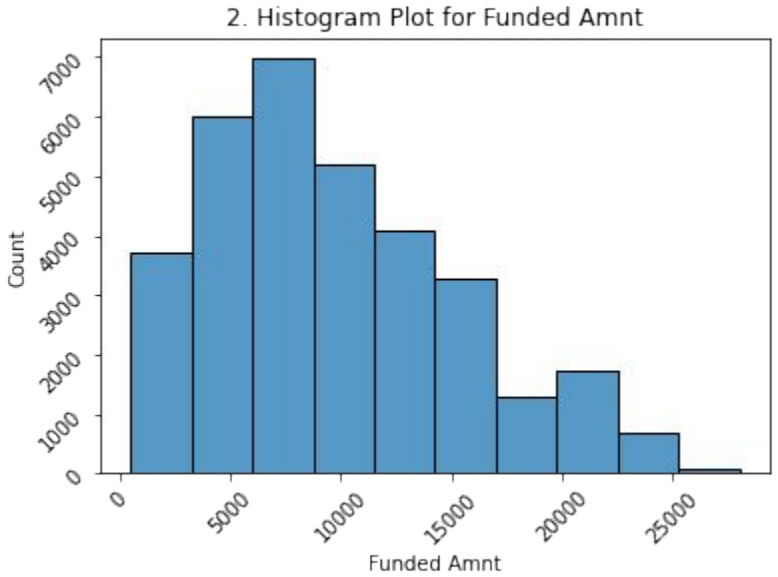


We can use DTI for our analysis.

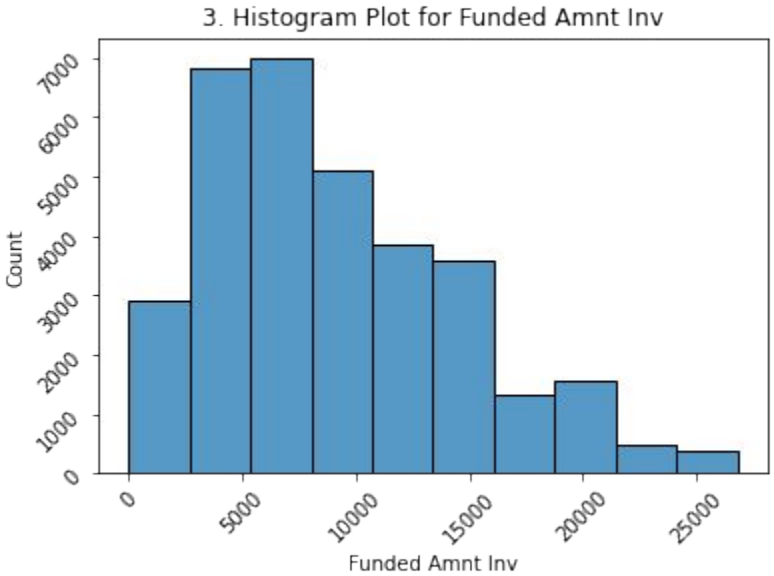


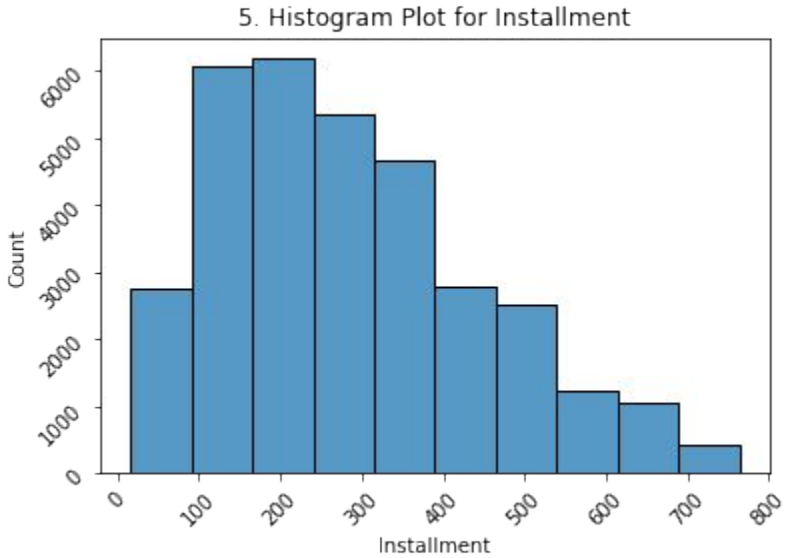
We cannot say much about this field as it has many outliers.
Since, many have outliers we should look at plotting a histogram.



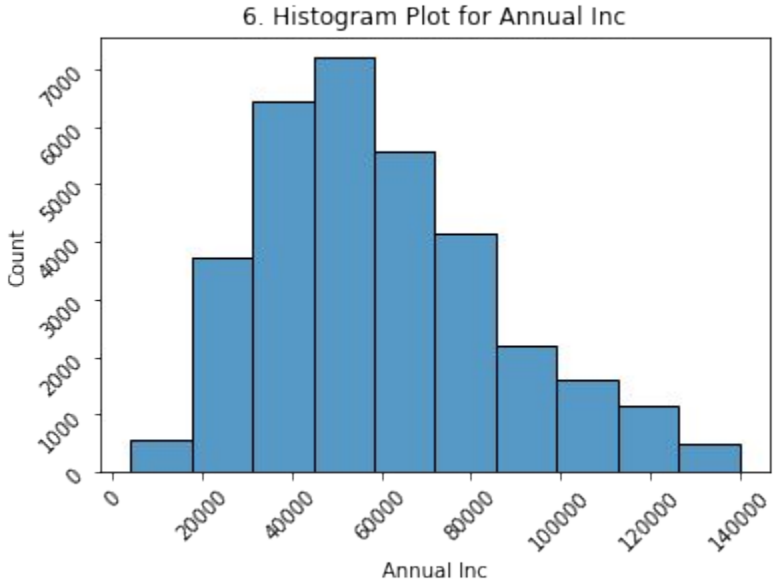


We can consider funded amount for our analysis.

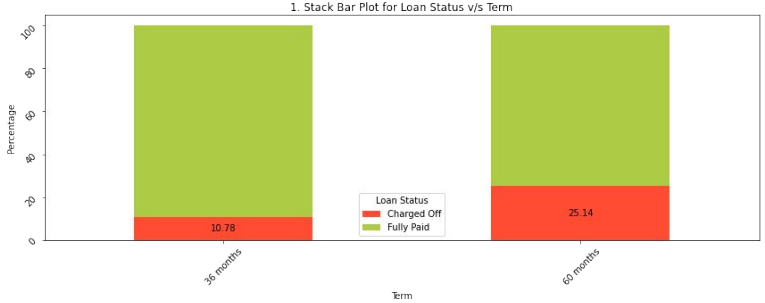


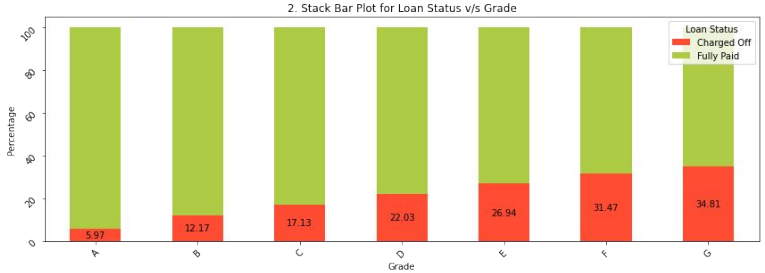


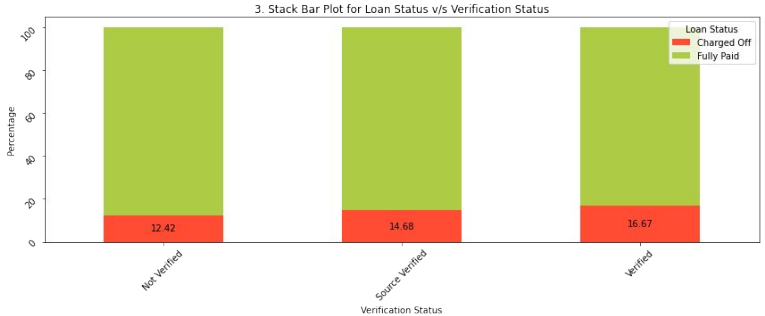
We need to avoid these as they are extremely right skewed.



We can consider annual income for our analysis.

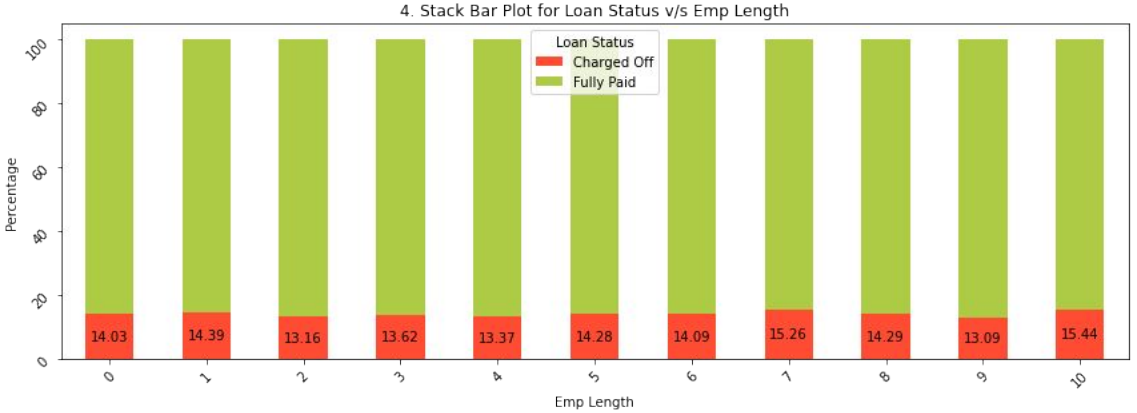


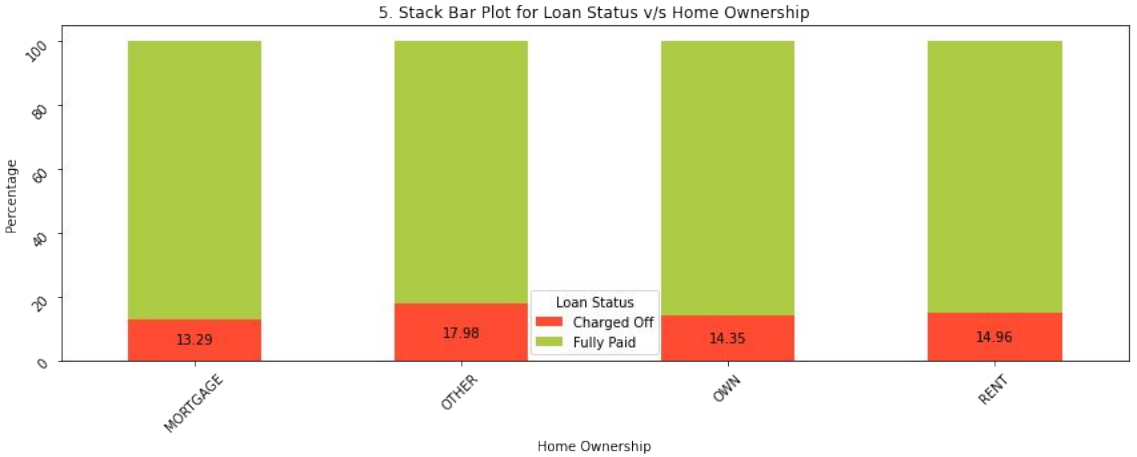


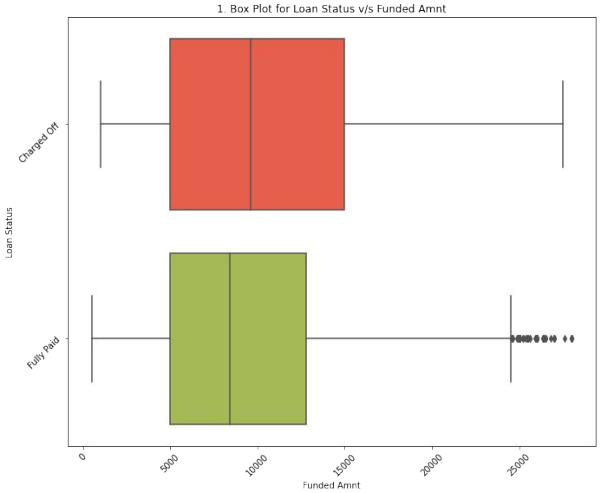


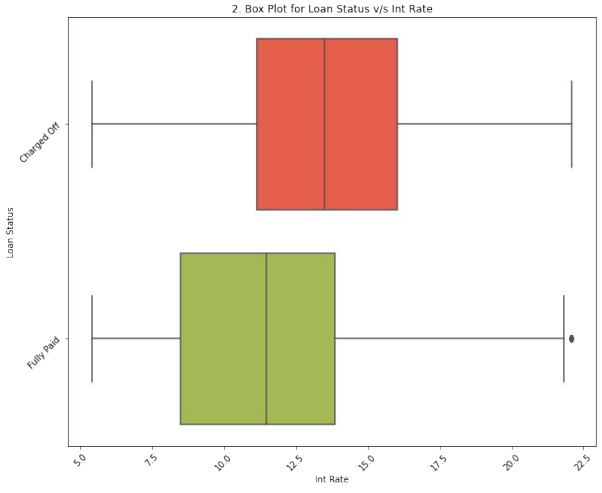
- Term: Loans taken for 60 months are almost 2.5x more likely to default as compared to loans taken for 36 months.
- Grade: This clearly states the risk factor increases as the grade increases from A through G. We do ignore F and G grades as we do not have enough data but we do see a pattern here.
- Verification Status: Status being source verified and verified slightly increases the default rate.

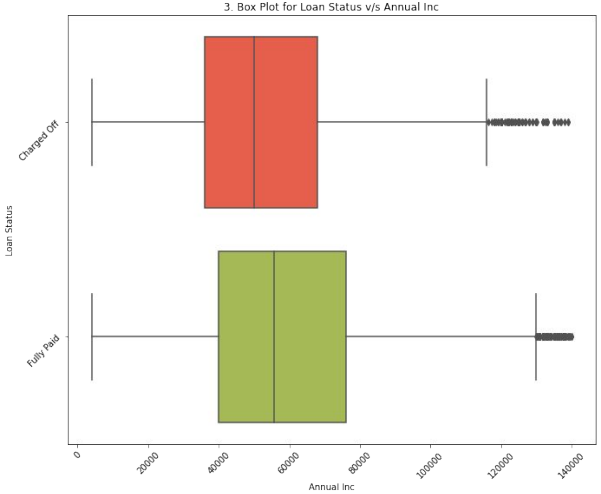
- Employment Length: We can see there is very slight deviation in the default rates in across the employment terms. Hence, it does not have much impact with the risk.
- Home Ownership: We need to ignore other category in our analysis, hence, we can see that owned or rented home ownerships are a bit more likely to default than the mortgaged. But this is very slight difference and could change with more data.

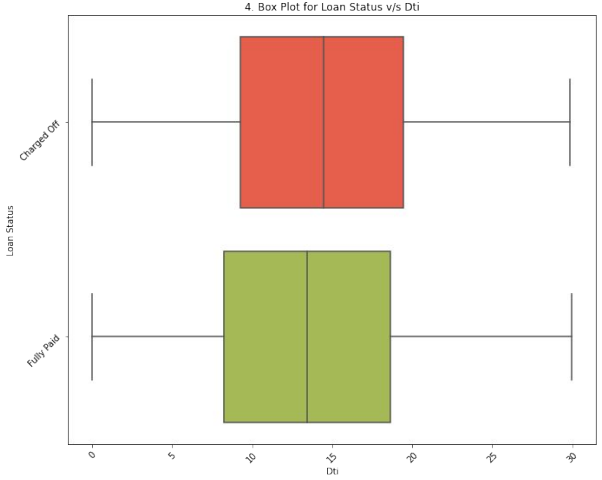








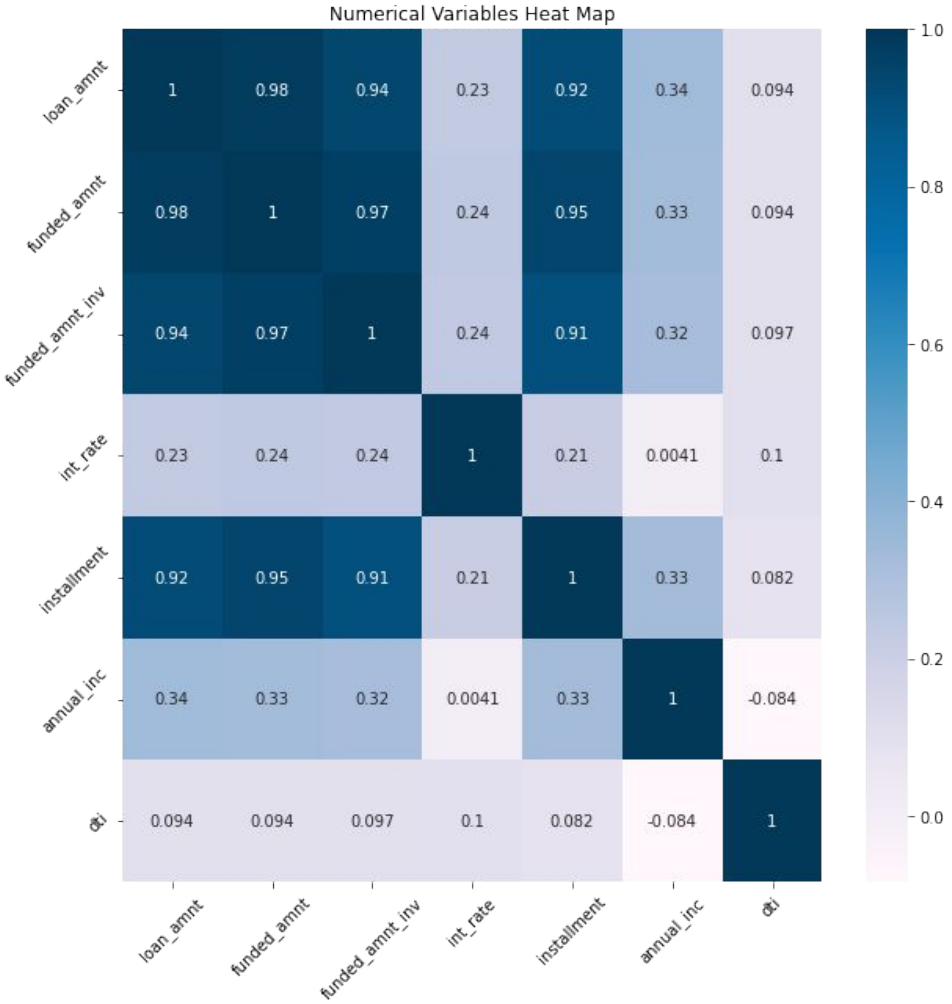




- Funded Amount: Plot for this is not helping us derive any conclusions.
- Interest Rate: We can clearly see that higher interest rates may increase the risk of default.
- Annual Income: This plot has many outliers, and it would not be safe to draw any strong conclusions but higher income may slightly reduce the risk of defaulting the loan.
- Debt-To-Income: As DTI increases, the risk of defaulting increases as well.

The two variables we can draw analysis from are int_rate and dti.

- int_rate does show a slightly positive correlation with loan_amnt, funded_amnt, funded_amnt_inv, and installment.
- dti does not show much correlation with any of the factors except for a very slight positive correlation with int_rate and even slighter negative correlation with annual_inc which is expected.



Conclusion

- Term has the most impact on default rate, loan term being 60 months are 2.5x more likely to default than 36 months.
- Loan Grades A through G shows a pattern of increased risk.
- Employment Length does not show much impact on default rates.
- Verification Status being source verified and verified slightly increases the default rate.
- Home ownerships being stated mortgaged, owned or rented, very slightly increases the default risk in the order. Mortgaged being the least.
- Higher the interest rate, higher is the risk of default.
- Higher annual income may very slightly reduce the risk of defaulting the loan.
- As DTI(Debt-To-Income) ratio increases, the risk of default increases as well.