Lending Club Case Study

Exploratory Data Analysis

PROBLEM STATEMENT

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. Two **types of risks** are associated with the bank's decision:

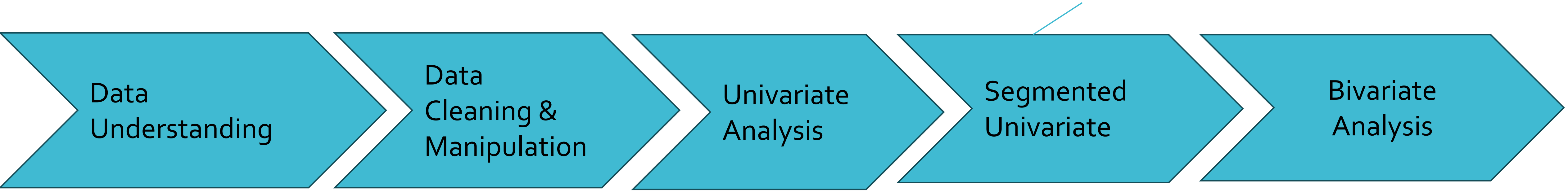
•If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company

•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

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, 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 utilize this knowledge for its portfolio and risk assessment.

EDA FLOW



Tools/Libraries used

Inputs

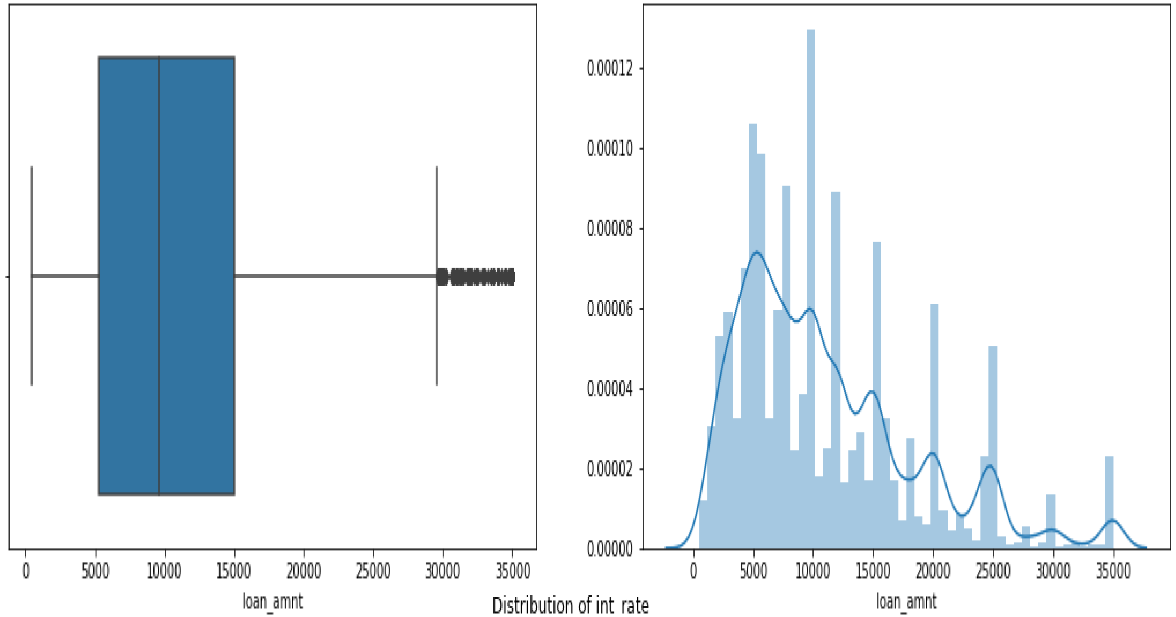
- Loan Dataset provided by Upgrad

Basic Flow

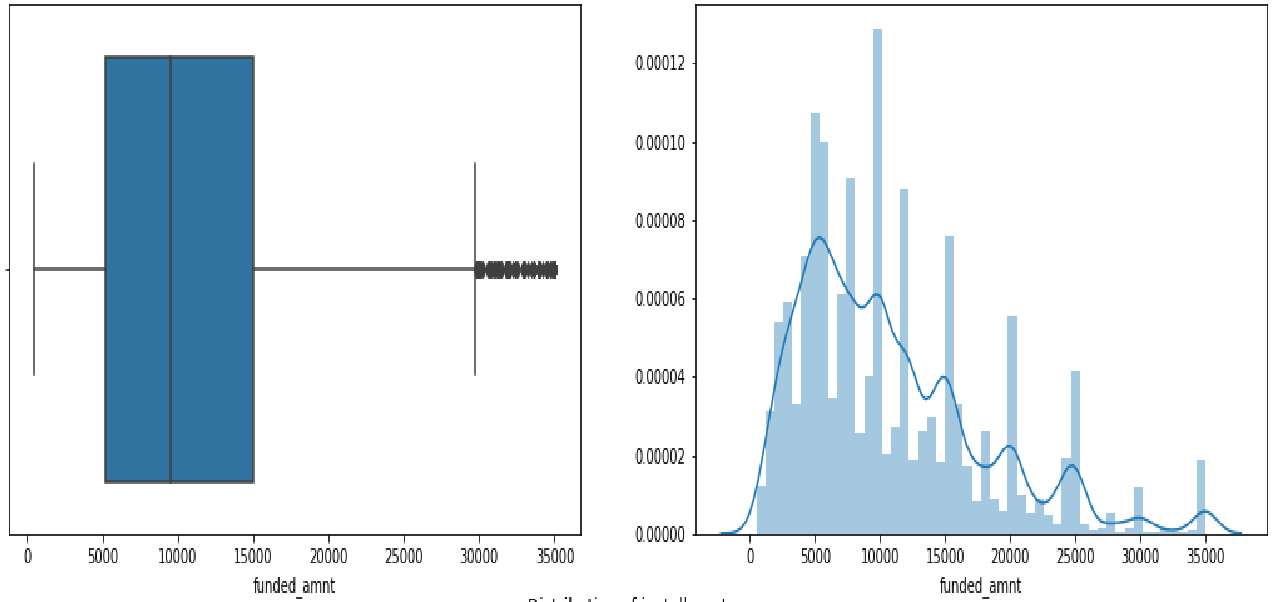
- Python in Jupyter notebook used for EDA
- NumPy, pandas, matplotlib, seaborn used for EDA.

Univariate Analysis - I

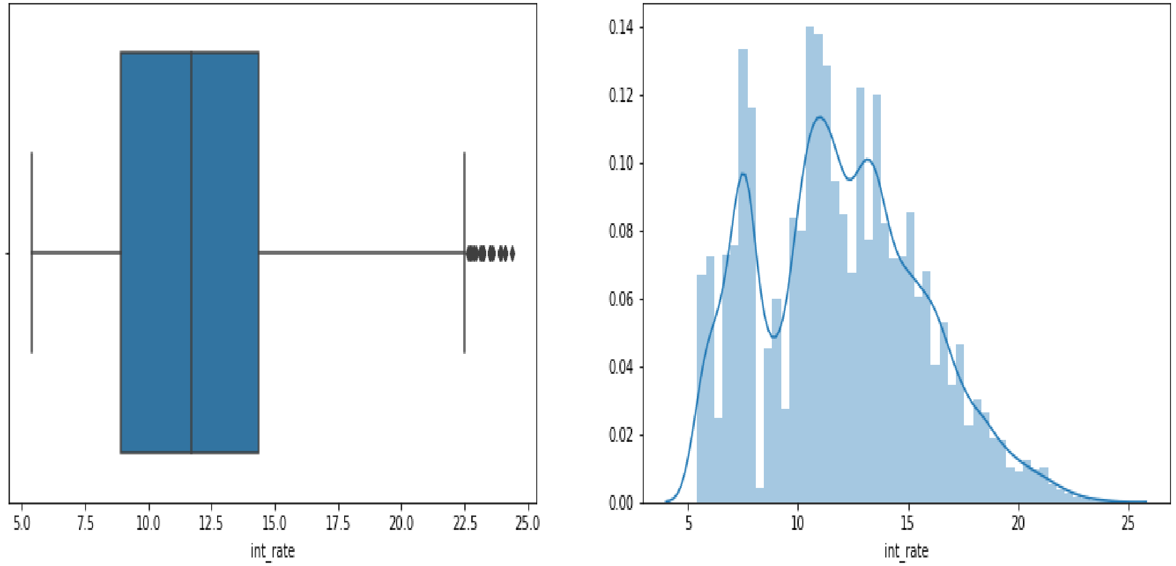
Distribution of loan_amnt



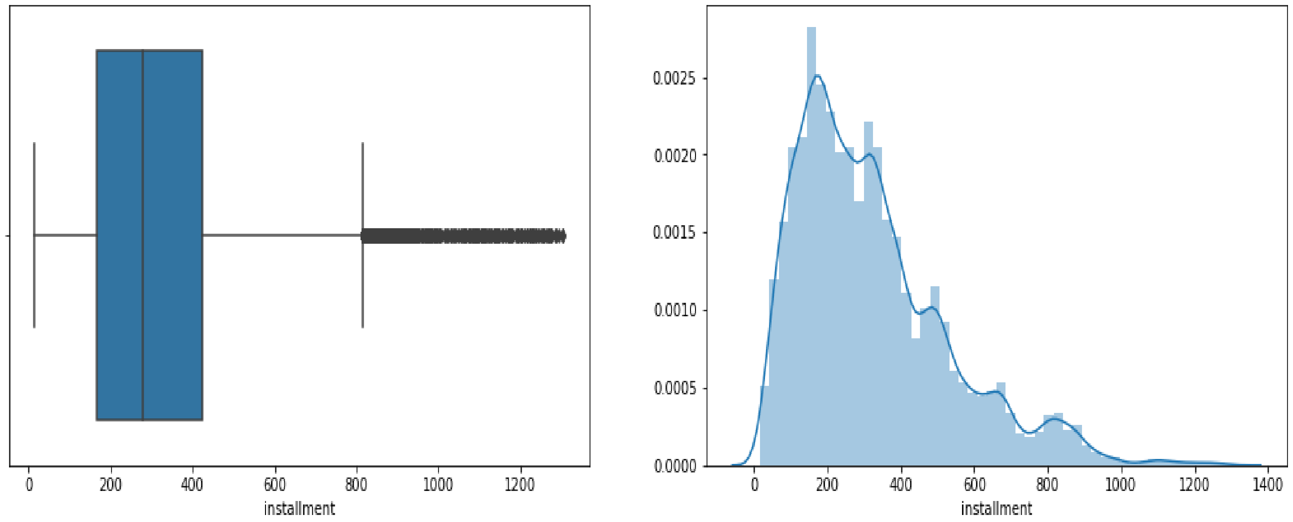
Distribution of funded_amnt



Distribution of int_rate



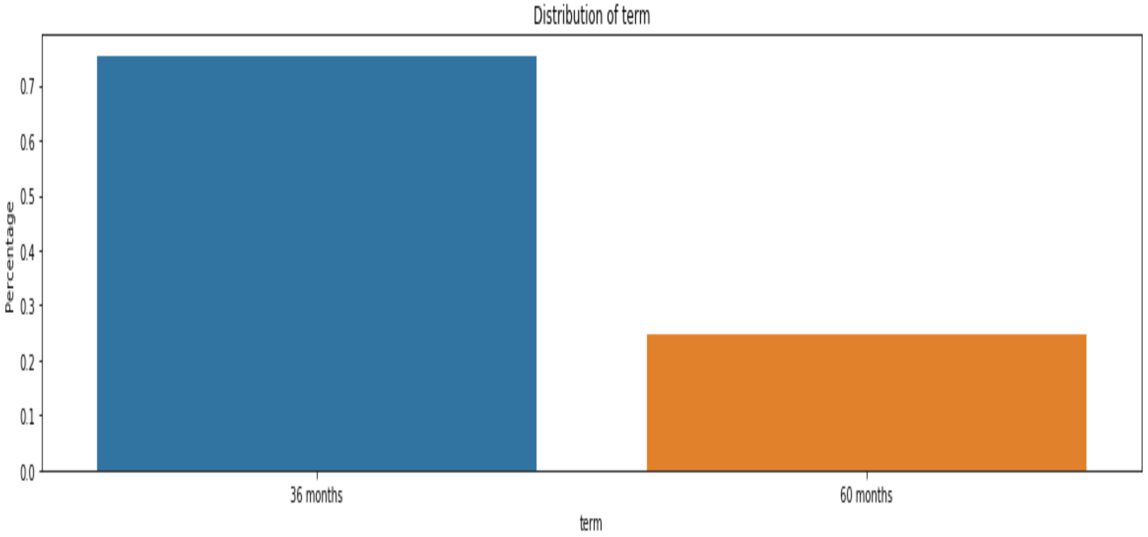
Distribution of installment

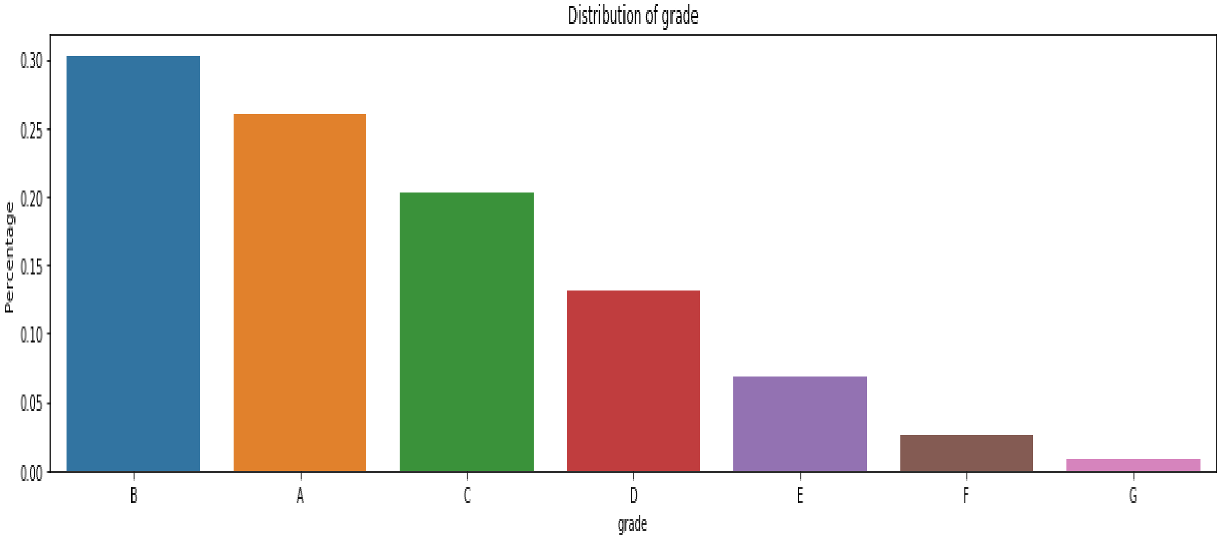


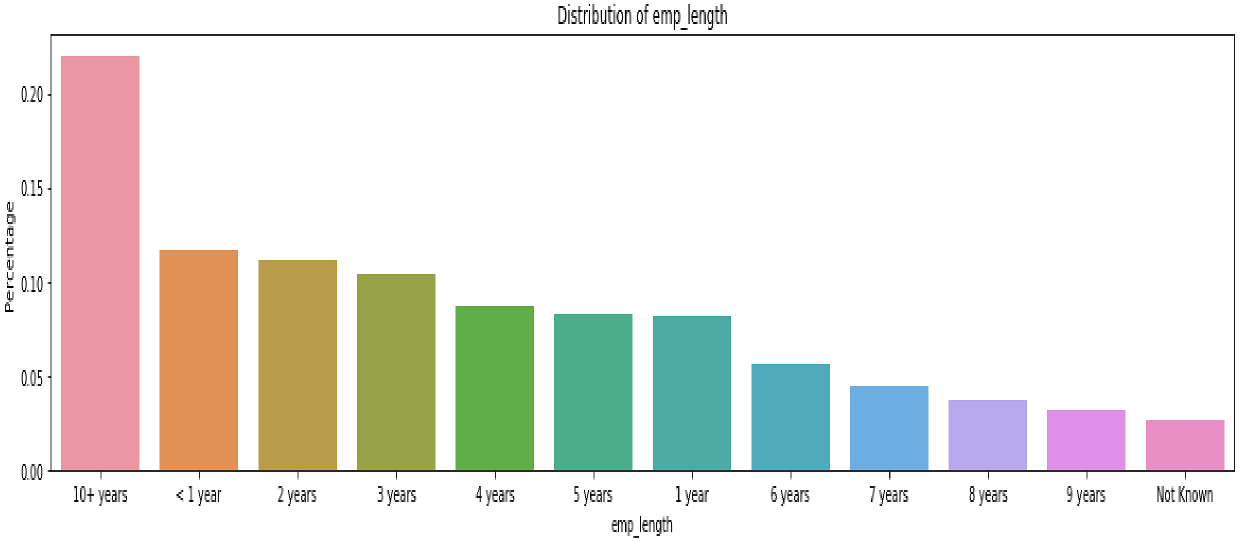
Univariate Analysis - II

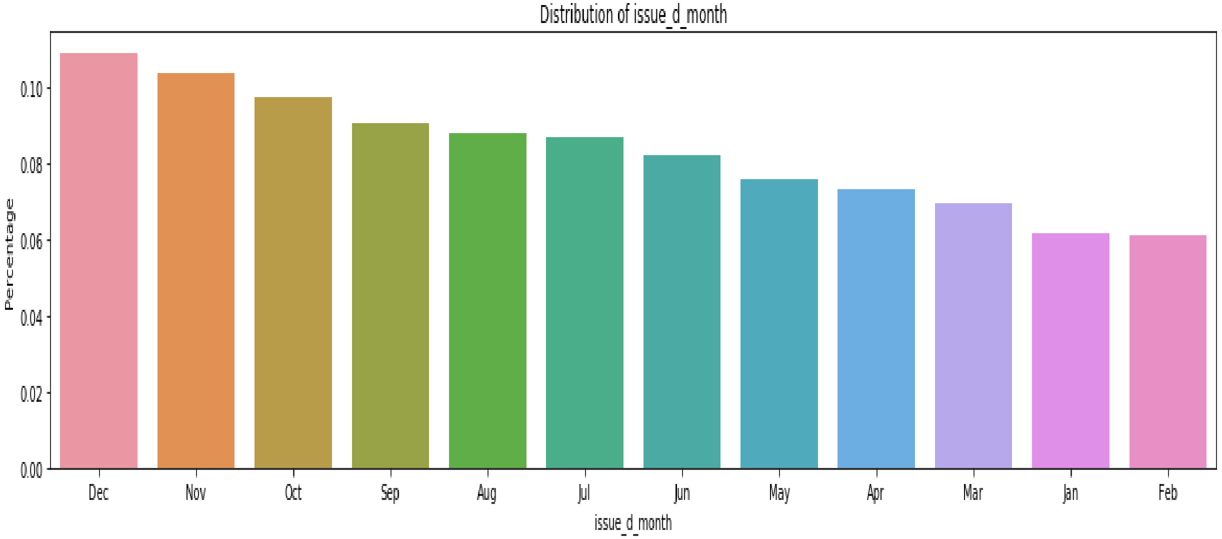
- 'id'
- This is a unique key which can be used a primary key.
- 'loan_amnt'
- From the above plots, the median is 10000. And very few people have taken a loan of 30000 or more. The IQR lies between 5000-15000.
- 'funded_amnt'
- Behaves similar to 'loan_amnt'. Which means most of the loan applications have been approved.
- 'funded_amnt_inv'
- Behave similar to 'loan_amnt'. This means most of the loan applications have been invested upon.
- 'int_rate'
- Interest rate for majority of the loans fall between 8 to 15%.
- 'installments'
- Number of installments have an IQR around 200-400.
- 'annual_inc'
- Annual Income of the applicants have an IQR between 40000 to 82000. Although there are quite a few high incomes which skew the data.
- 'dti'
- Looks like there are no outliers and the distribution is very much similar to normal distribution.
- This is good sign that all the loans are given to barrower's who have Debt to Income ration less than 30.
- 'open_acc'
- The numer of open credit lines can be a major indicating variable. It has an IQR between 5-12
- 'revol_bal'
- IQR between 25-72. Follows a poisson distribution
- 'revol_util'
- Follows a normal distribution. Indicates the percentage of credit used out of the available limit.
- 'total_acc'
- Follows a basic normal distribution. Total number of current credit lines.

Univariate Analysis - III





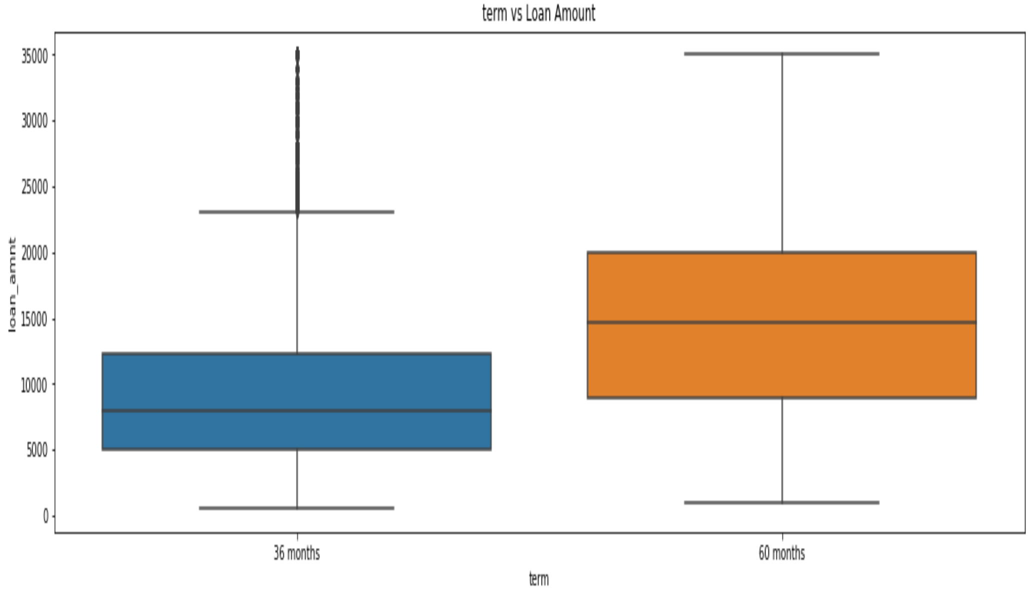


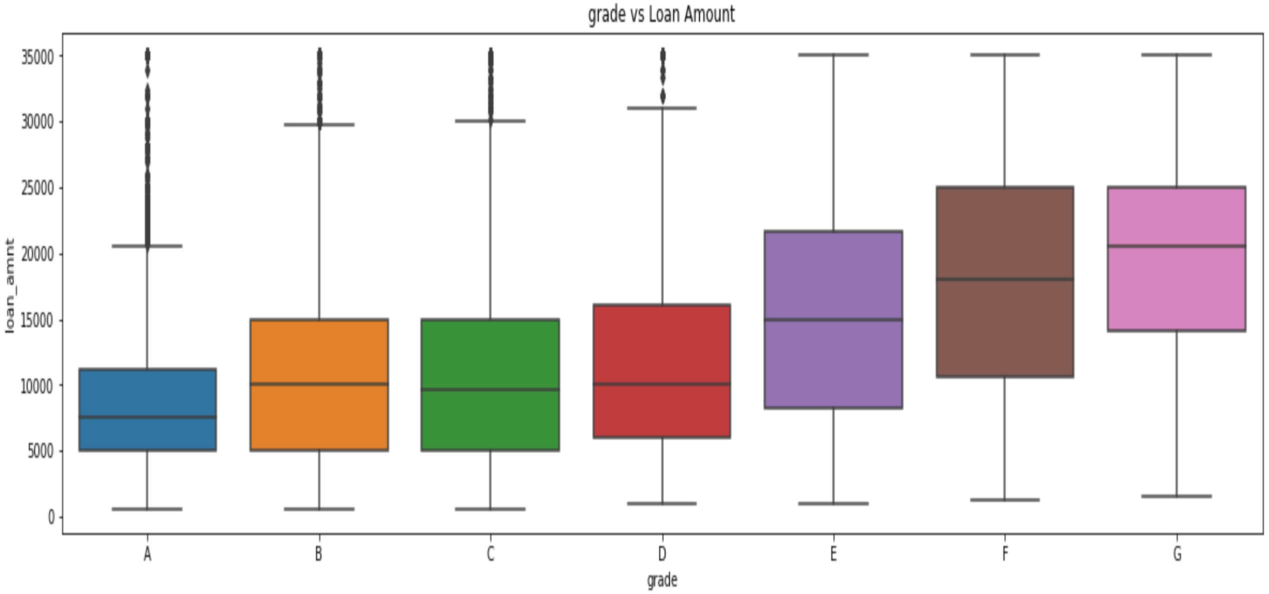


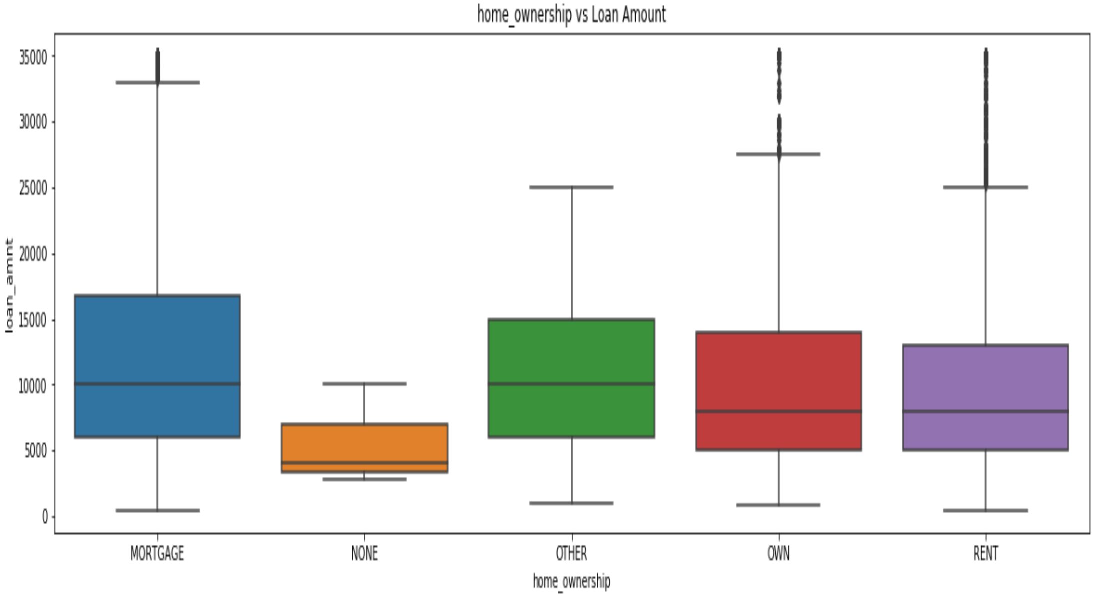
Univariate Analysis - IV

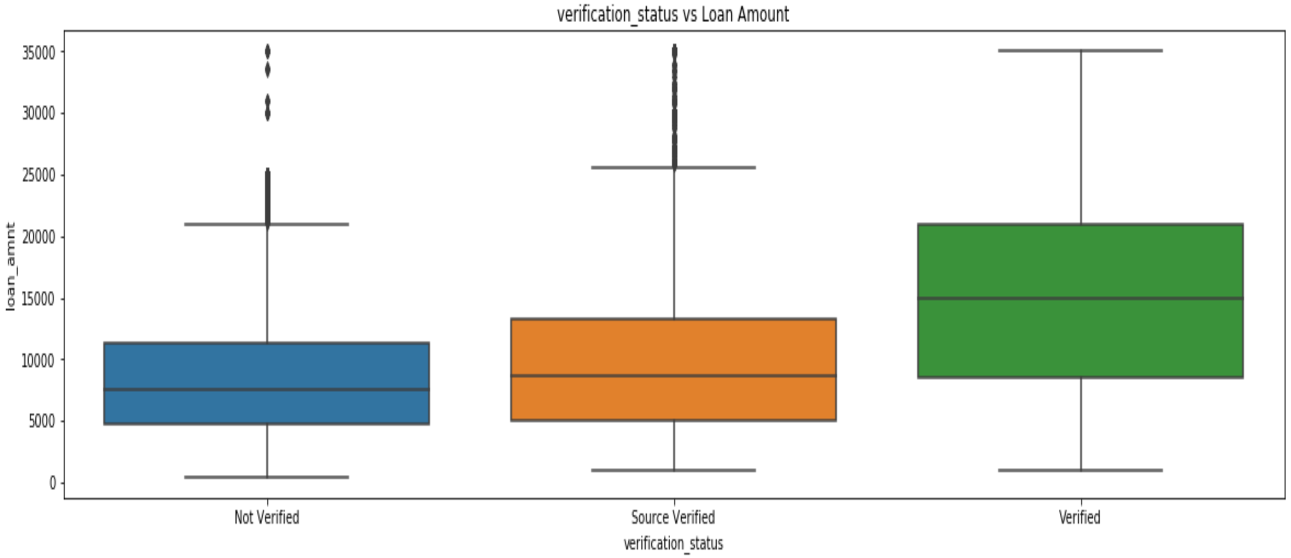
- 'pub_rec'
- Around 90% borrower's are having no public derogatory records.
- 'pub_rec_bankruptcies'
- There is no evidence of bankrupcies in 99% cases.
- 'term'
- Tenure of '36' months is more frequent.
- 'grade'
- 'A' and 'B' account for most of the grades among the appliers.
- 'sub_grade'
- 'A4', 'B3', 'A5' are the most frequent sub grades.
- 'emp_length'
- Applicants having more than 10 years of experience are the majority.
- 'home_ownership'
- Applicants mostly have a rented accomodation or have mortgage. Very few people have their own houses.
- 'verification_status'
- 'Not verified' applicants are majority.
- 'loan_status'
- More than 80% loans are paid off.
- 'purpose'
- 'Debt consolidation' and 'credit_card' account for the major reasons behind the application.
- 'addr_state'
- Most of the applicants are from 'CA' and 'NY'

Segmented Univariate - I









Segmented Univariate - II

Observations on 'int_rate'

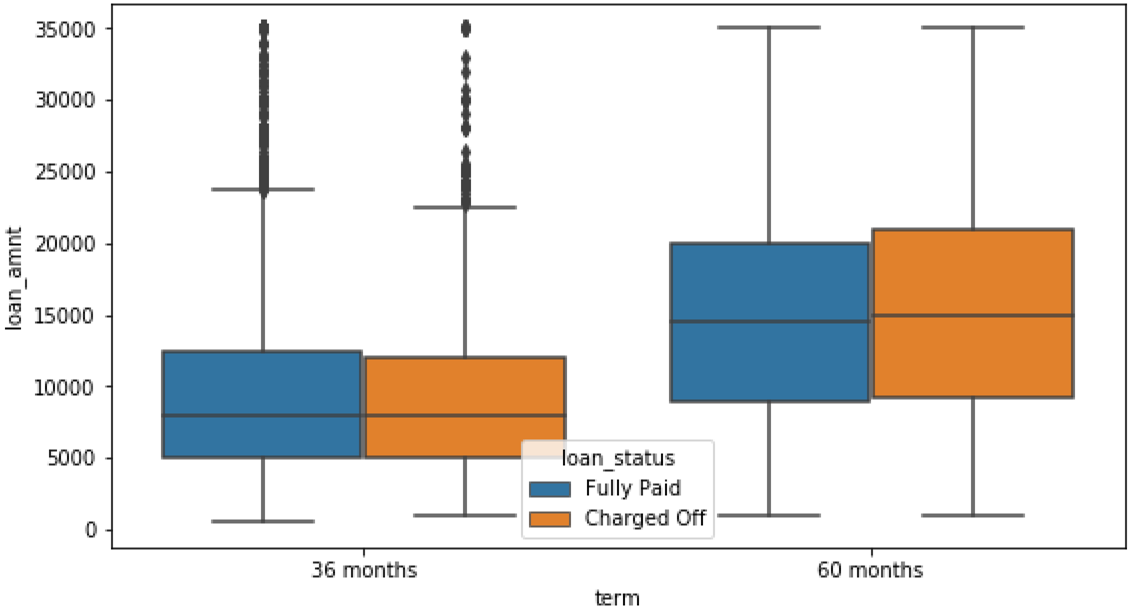
- Higher interest rates demanded for higher loan tenures(60 months).
- As the grades go from 'A' to 'G', interest rates increase.
- Applicants having mortgage are given loans with lower interest rates, because of the security.
- The verified applicants are charged higher interest rates.
- Higher interest rated loans default more compared to lower ones.
- 'Small Business' tops the chart in higher interest rates.
- Employment length isn't a factor towards interest rate.
- Interest rate has diversified over the years. Nothing stands out.

Observations on 'loan_amnt'

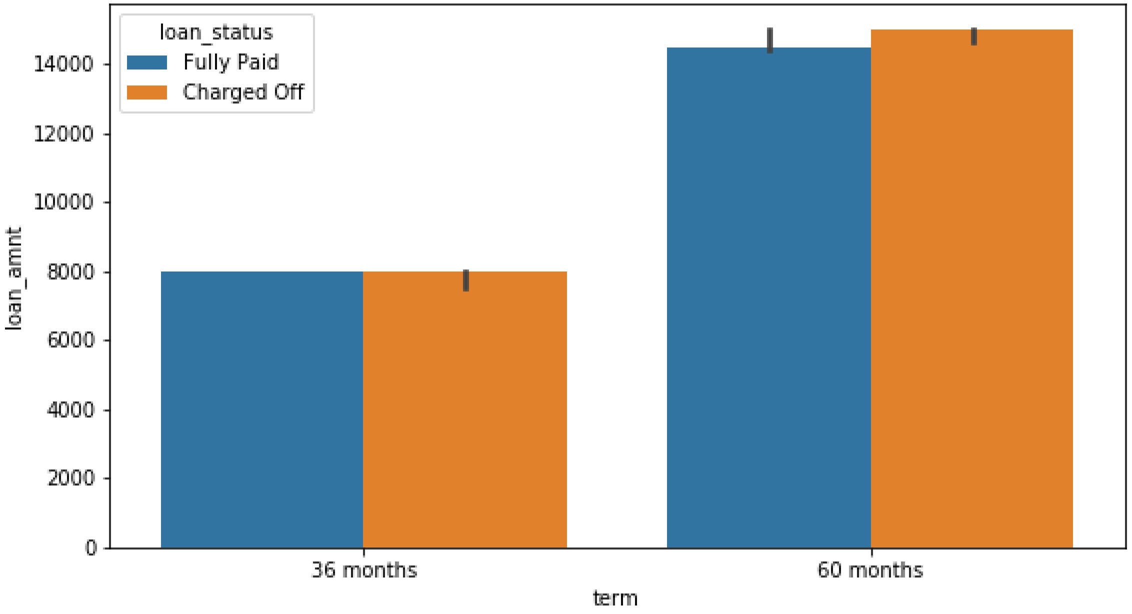
- High loan amount have higher loan tenures(60 months).
- Grade 'F' and 'G' have higher loan amounts. It is safe to say the loan amount is positively corelated with Grade.
- Applicants having mortgage are given higher amount of loans.
- 5000-10000 amounted loans are provided across all segments. There is higher inclination towards 'Verified' applicants once the loan amount crosses 10000.
- Defaulted loans have a slight inclination towards higher loan amount. But, not so significant.
- 'Small Business' tops the chart in retrieving high amounted loans.
- 10+ years experienced employees have higher amounted loans.
- There was a dip in loan amount in 2008, which makes sense as it was the yearof recession. Apart from that more or less the spread across years is even. December tends to be a month with high loans provided.

Bivariate Analysis - I

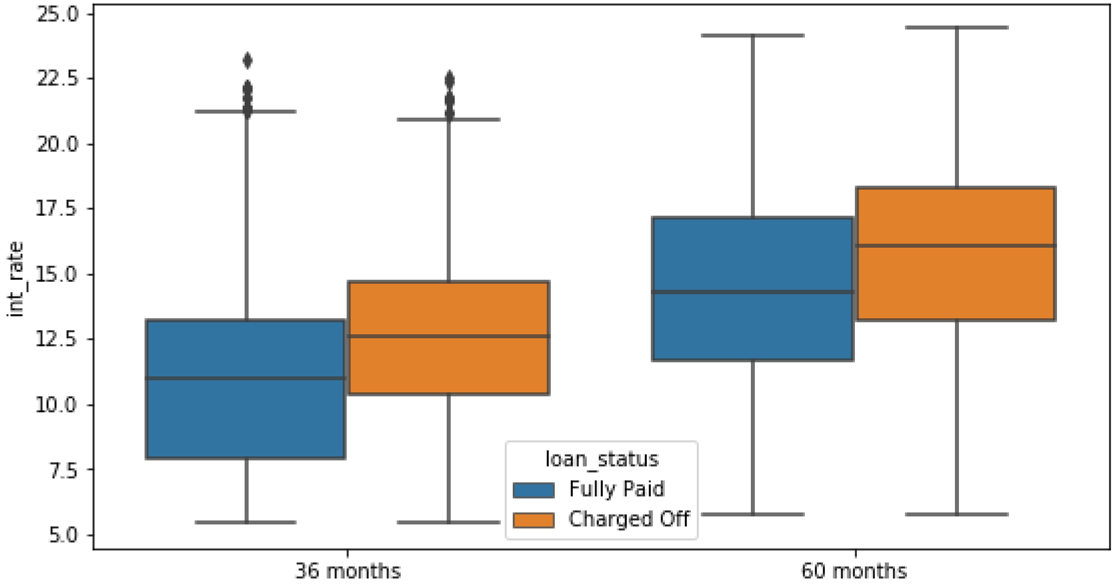
Term vs loan amount



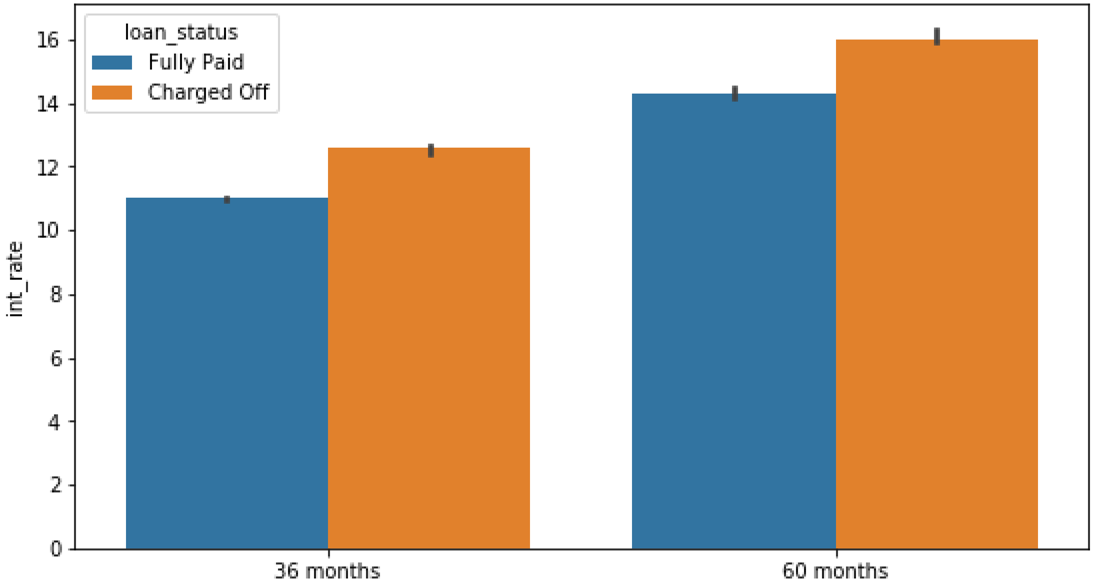
Term vs loan amount



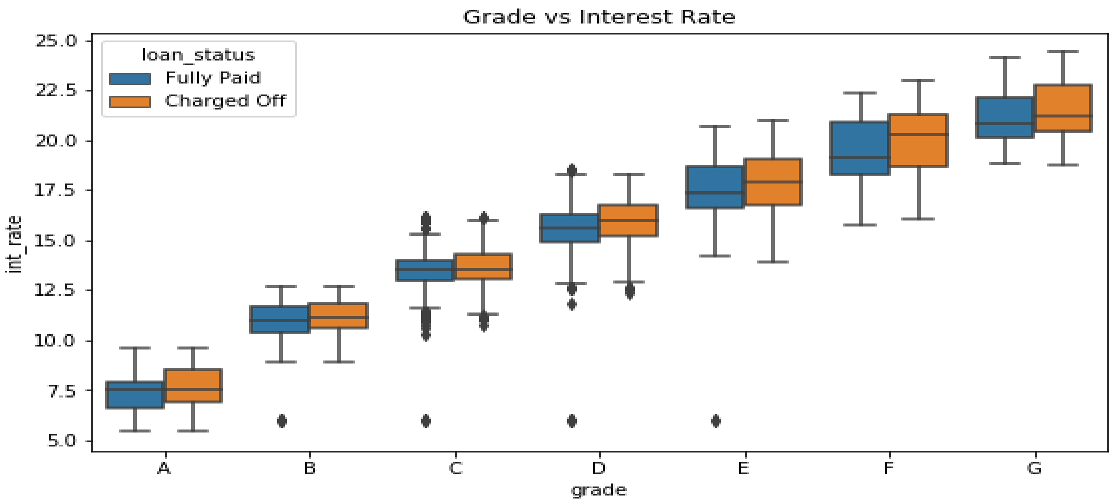
Term vs Interest rate



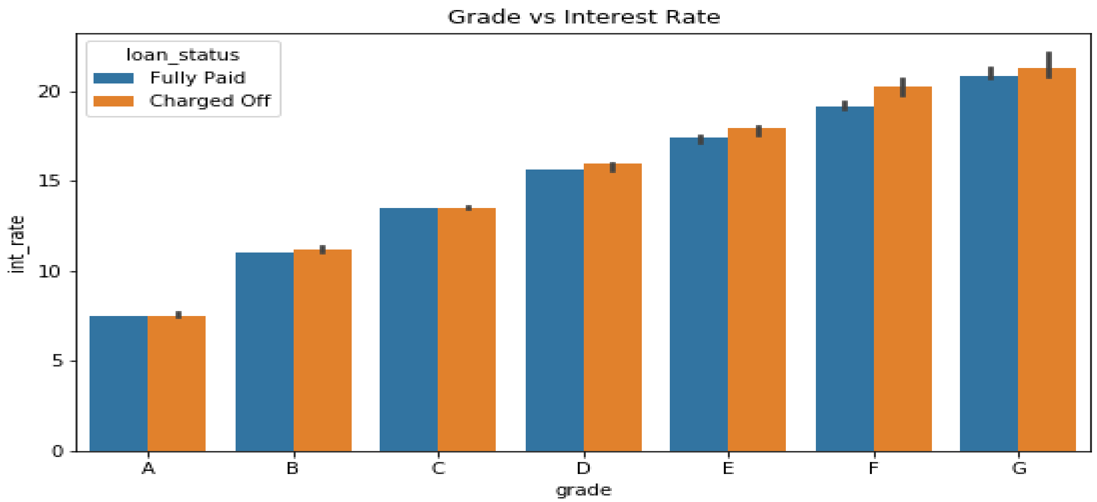
Term vs Interest rate



Bivariate Analysis - II

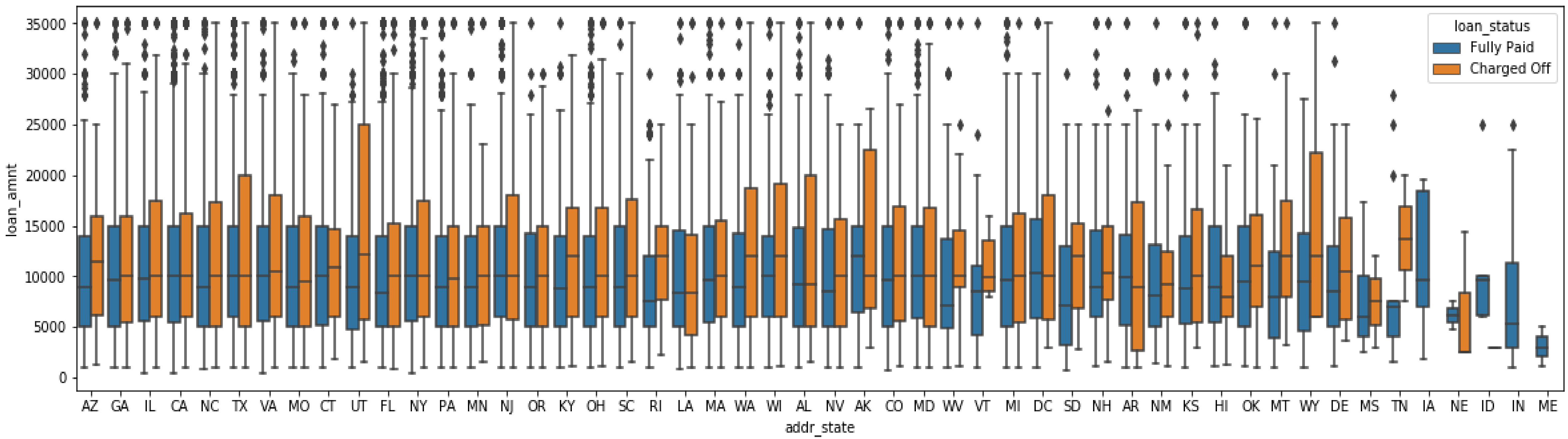


Grade vs Interest Rate



Grade vs Interest Rate

addr_state vs Loan amount



Bivariate Analysis - III

'term'

- Higher tenured loans on a high level have a higher chance to default. However, 'loan_amnt' is not a factor behind the defaulters in both tenure categories.
- Default rate is high when interest rate is high.
- Defaulters are higher for both tenures if the DTI is high. This should be factor while approving a loan.

'grade'

- There is a steady increase in defaulters when the grade changes from A-G, in that order. For lower grades 'F' and 'G' there are more difference between charged-off and fully paid, which means the lower grade applicants have taken higher amount of loans and also they are more prone to default the loan.
- There is a steady increase in interest rate and defaulters as we go from A-G
- DTI is not a major factor when it comes to grade and loan status together.

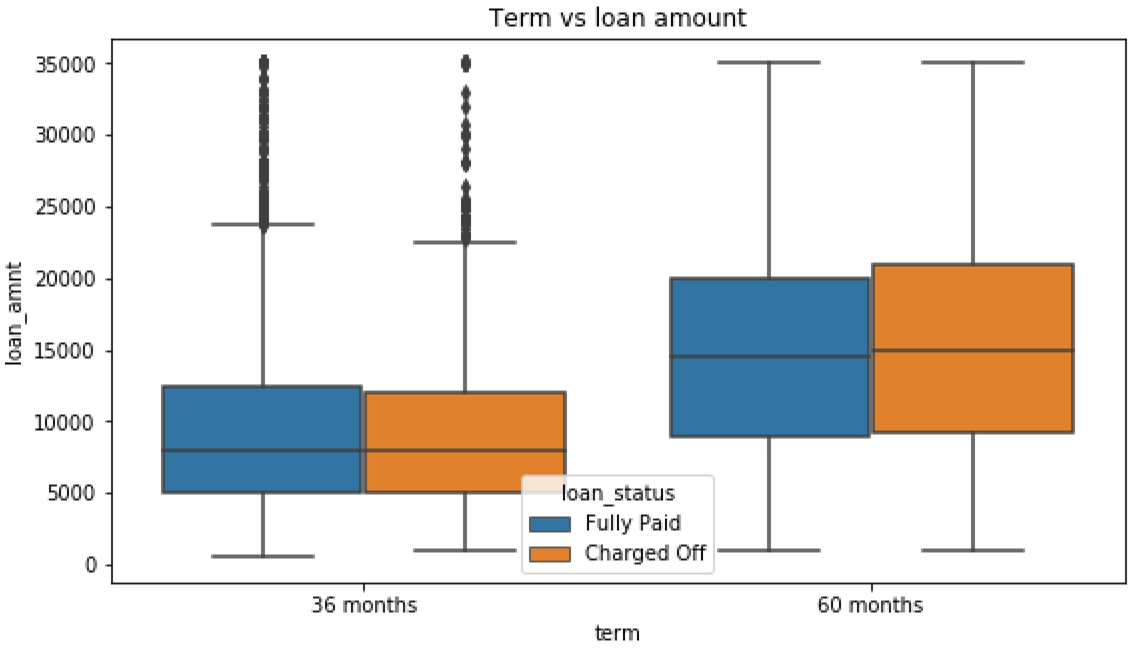
'employment length'

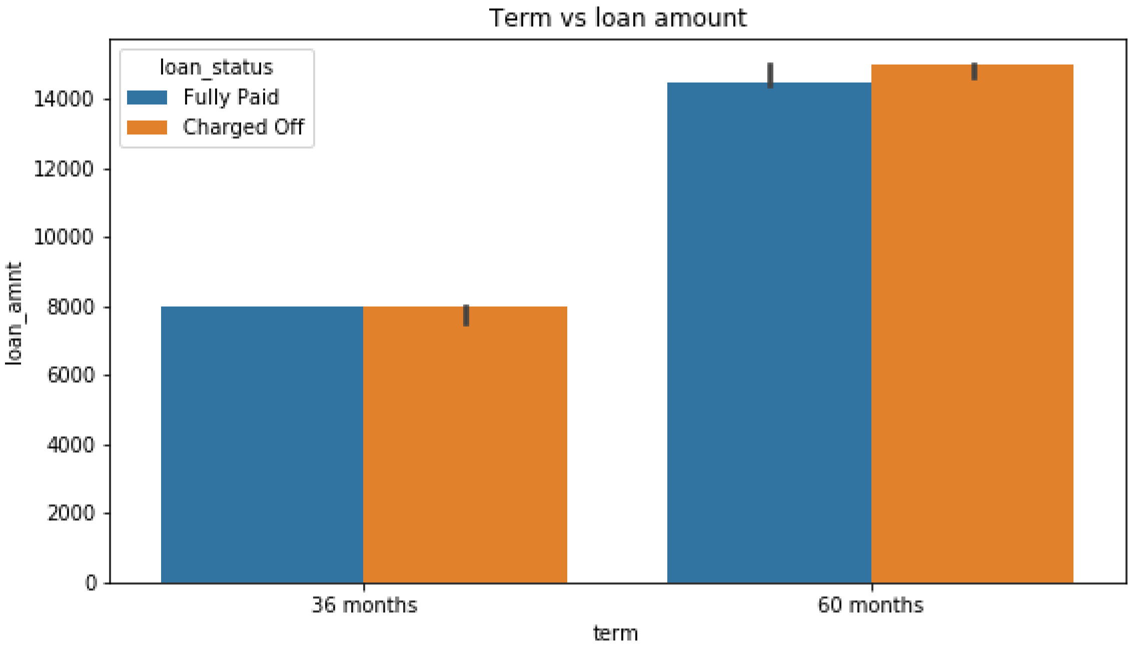
- The applicants having 10+ years employment experience have taken high amount of loan and have defaulted more.
- Loans with higher interest rates have defaulted more across employment lengths.

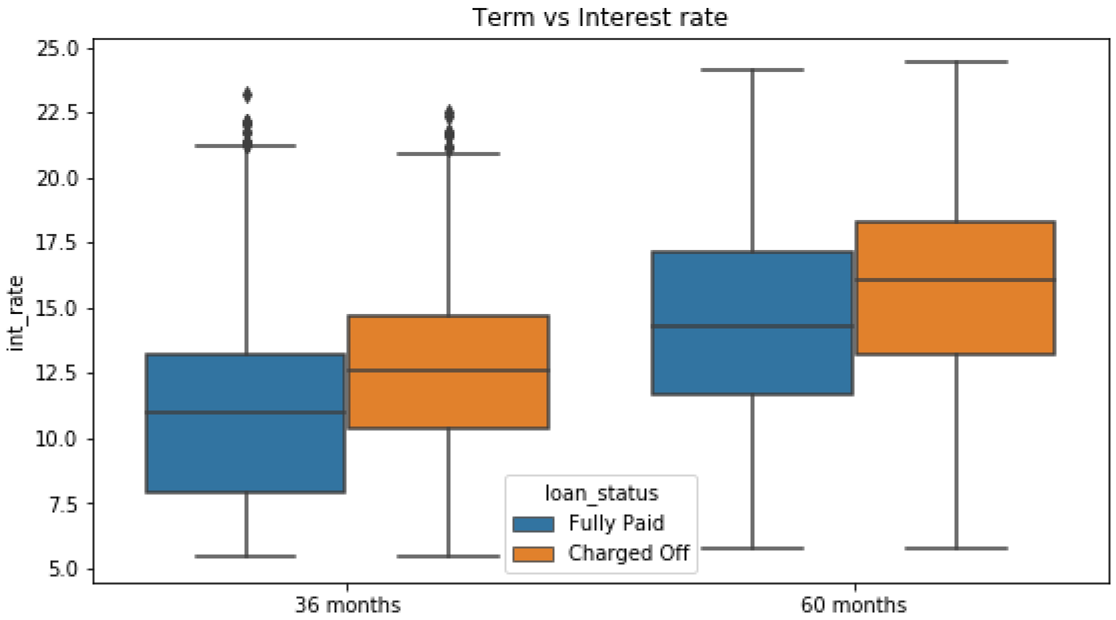
'home_ownership'

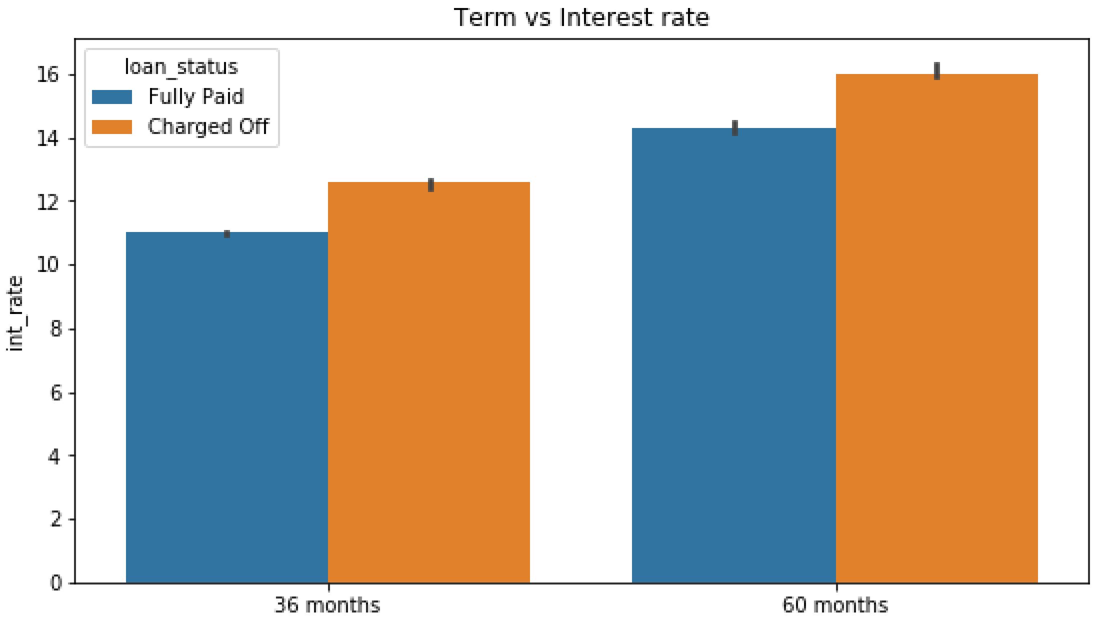
- There is a slight higher default rate in the 'OTHER' category.
- Loans with higher interest rates have defaulted more across home ownerships.
- The more the loan amount the more defaulters irrespective of home ownerships.
- There is equal possibility of home owners defaulting for all the home ownerships.

Bivariate Analysis - IV


Term vs loan amount


Term vs loan amount


Term vs Interest rate


Term vs Interest rate

Bivariate Analysis - V

'verification_status'

- Verified loans are given more loan amounts compared to others.
- HIgher loan amounts have defaulted for 'verified' category more.
- Loans with higher interest rates have defaulted more across verification status.
- 'Verified' applicants having higher DTI default more.

'purpose'

- 'Small Business' have higher defaulters.
- 'House' has higher defaults compared to others when the interest rate is high.

'addr_state'

- Huge numer of defaults in the state 'NE'.
- 'UT','WY','AK' have high defaulters with high loan amounts.
- 'UT','WY','AK' have high defaulters with high interest rates.

EDA Summary

● Lending club should reduce the high interest loans for 60 months tenure, they are prone to loan default. Maybe they can another tenure category which may cater to intermediatory customers.

● Grades are good metric for detecting defaulters. Lending club should examine more information from borrowers before issuing loans to Low grade (G to A).

● ‘UT’, ‘WY’, ‘AK’ have high defaulters. Further analysis should be done when an application request comes from those states.

● Small business loans are defaulted more. A proper plan should be put in place to stop/reduce the loan amount.

● Borrowers with mortgage home ownership are taking higher loans and defaulting the approved loans. Lending club should stop giving loans to this category when loan amount requested is more than 12000.

● People with more number of public derogatory records are having more chance of filing a bankruptcy. Lending club should make sure there are no public derogatory records for borrower.

- Applicants with lower incomes should not be provided high amount of loans.
- Applicants with lower incomes should be provided loans with less interest rates.