EDA on Lending Club's Loan Data

Analysis Done By

Name: **Arunkumar Gunasekaran**

Batch: **ML C57** Assignment: **Lending Club Case Study**

Email: manigunasekaran30@gmail.com

Problem Statement

- The Lending Club is a consumer finance company which provides loan to urban consumer. The company receives the application from the consumer, Company has to make decision whether to **Approve/Reject** the Request.

There are two types of Risk

1. If the applicant is **likely to repay** the loan, then **not approving** the loan results in **a loss of business** to the company
2. 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

- As an ML Engineer need to **perform EDA** on customer loan data to understand how **consumer attributes and loan attributes** influence the tendency of default.So the company can mitigate two type of risk mentioned above.

Business Objective

- 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
- 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.

Analysis Design

Analysis will be carried out three stages

1. Exploratory analysis to understand the data set, clean the data set, and identify key column for the analysis
2. Perform univariate, bivariate, and multivariate analysis for target categorical and numerical variables
3. Identifying the variables which can help in predicting customer loan default.

Data Preparation

- **Private data** as been collected from the Lending Club Customer in form of **csv** containing the complete loan data for all loans issued through the time period 2007 t0 2011
- Loan Status on the dataset indicated whether customer defaulted or not (Charged off – defaulted,

Successfully paid- Not defaulted)

Data Exploration Actions:

1. Loading the data
2. Understanding data structure
2. Data Quality checks

Data Cleansing Actions

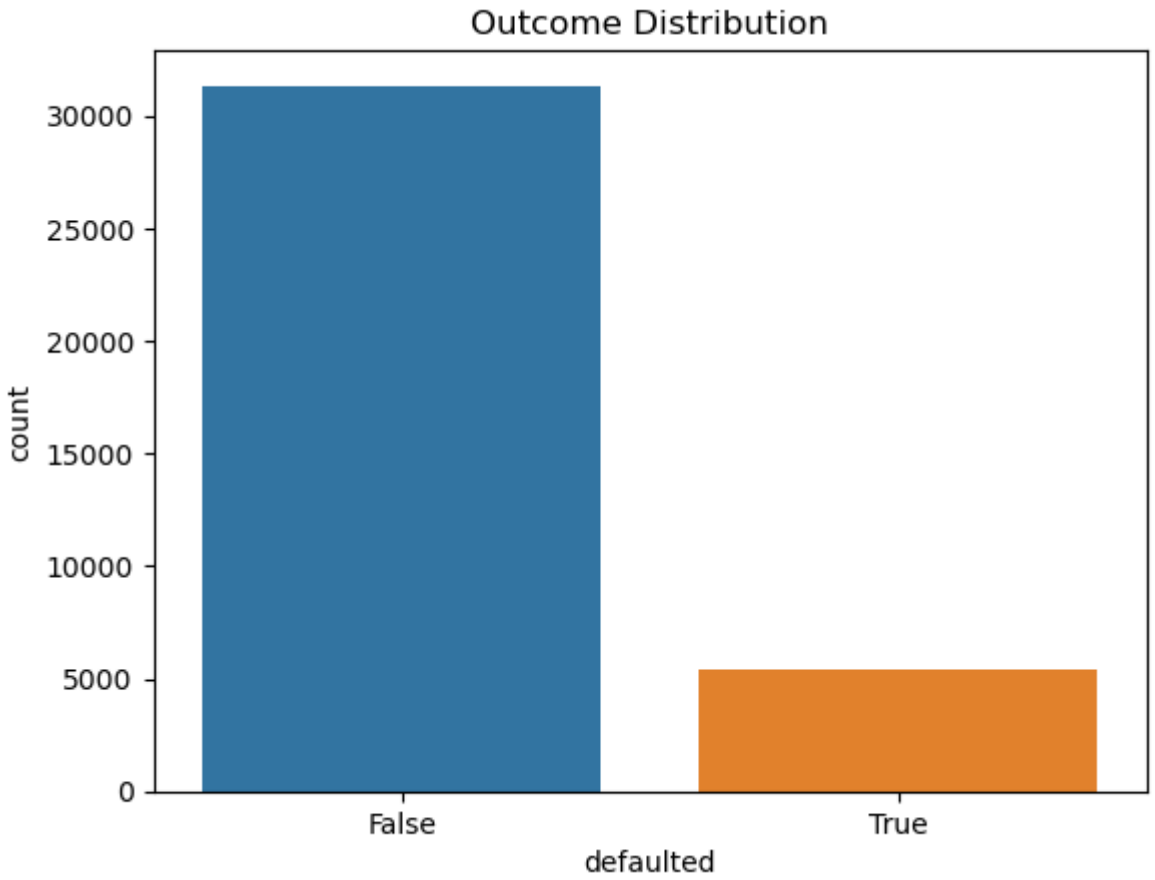
1. Removed current loan data as for analyze we need only Charged off and Successfully paid
2. Data quality checks done on columns and rows
3. Columns with more than 50% missing values are removed
4. Outlier are checked and removed on columns
5. Dropped columns not relevant to Analysis (constant columns, unique columns, data available post loan processing)
6. Fixing data types of the columns
6. Derived additional columns required for analysis

Univariate Analysis

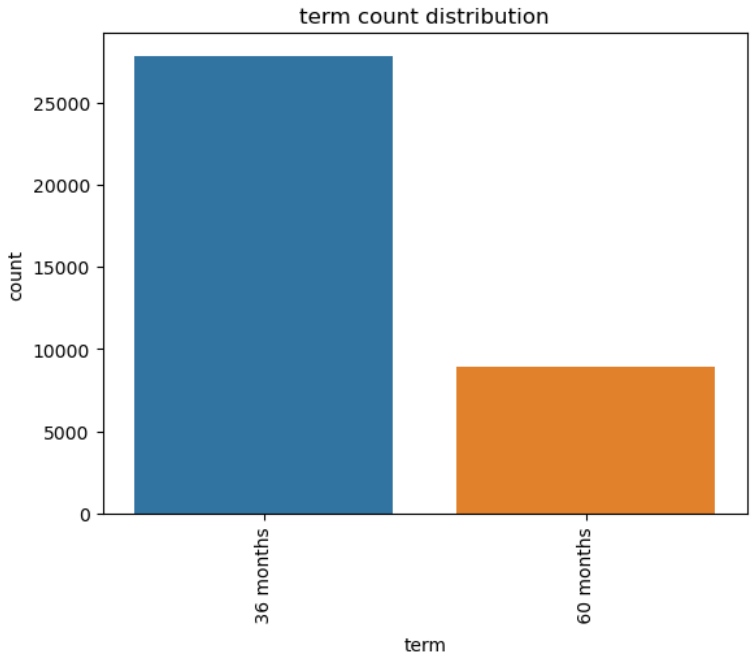
Univariate Analysis on outcome variables

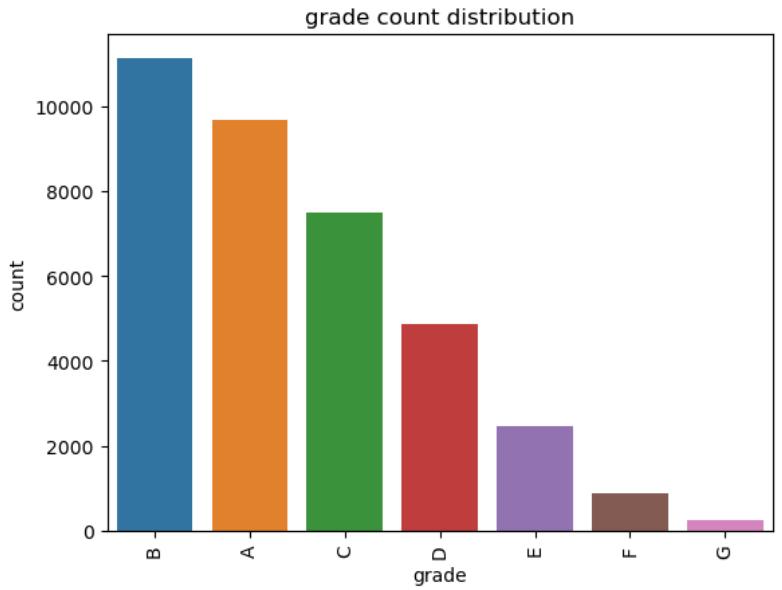
Bar chart on left shows outcome distribution

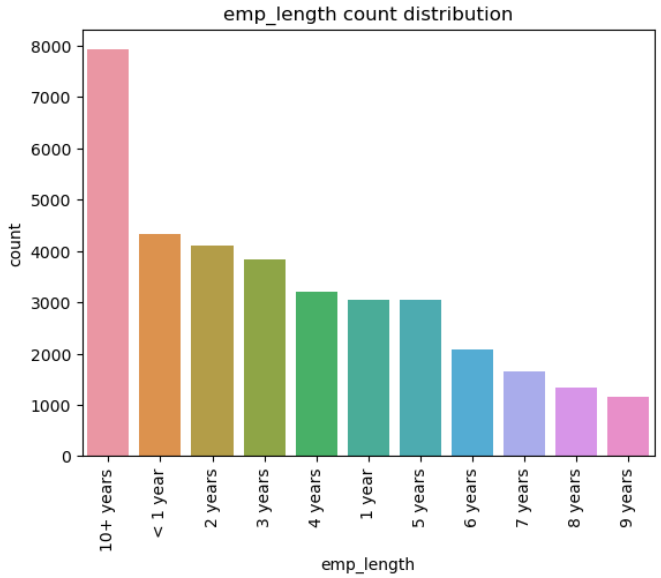
1. 14.77 % of the loan application in the data set are defaulted
2. We can also observe imbalance in the data



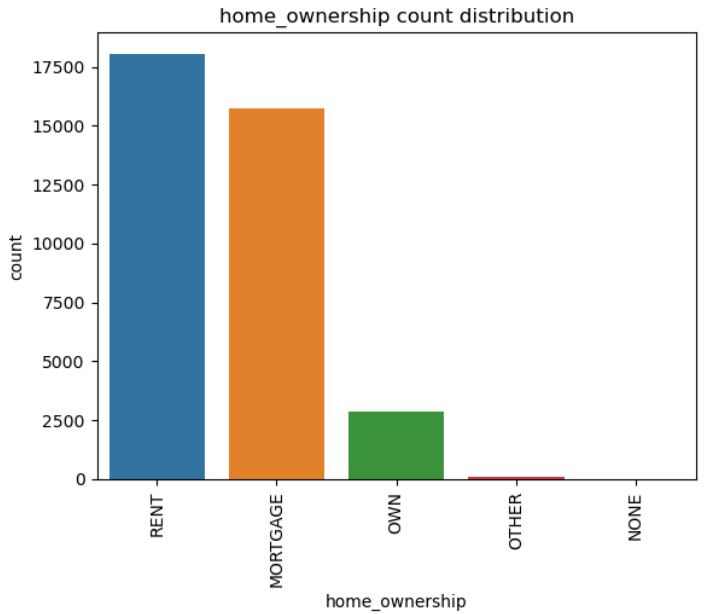
Univariate Analysis on Categorical Variables



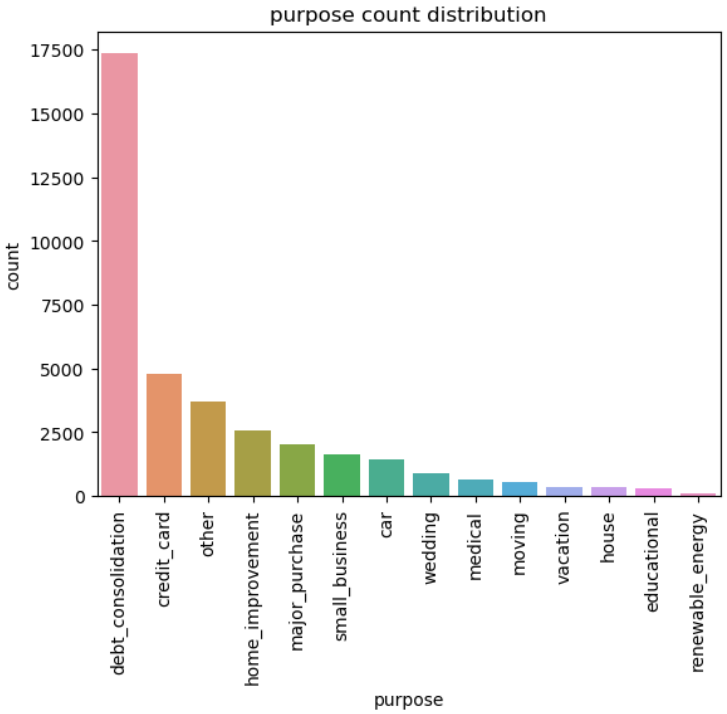


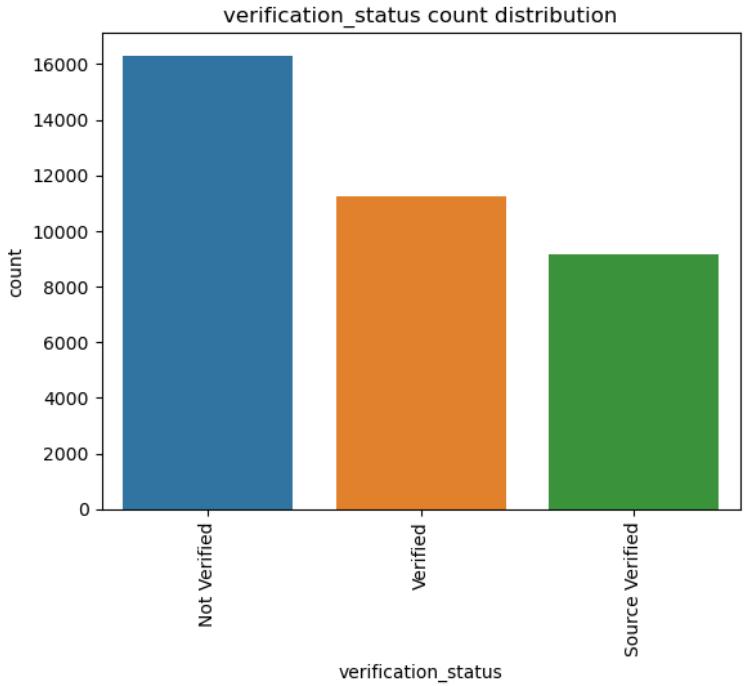


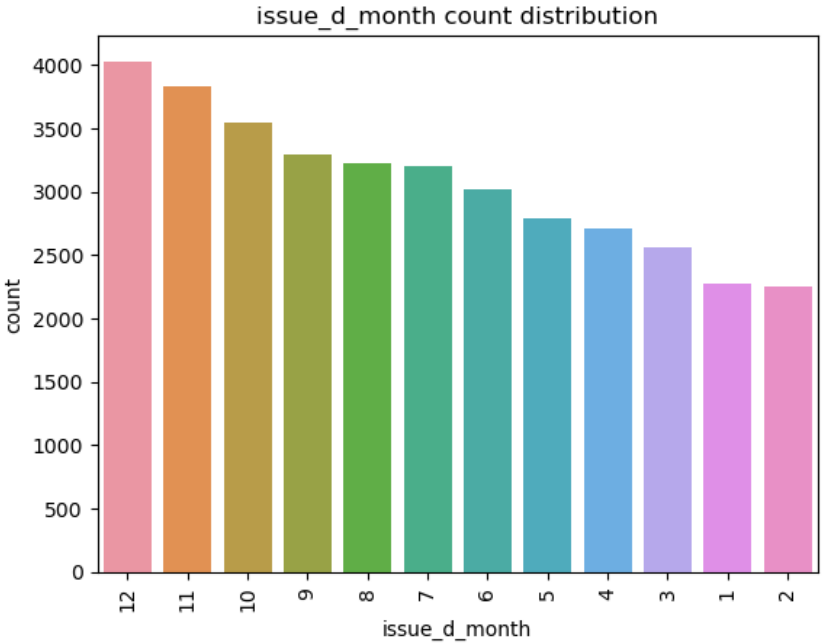
1. More loans are taken for 36 months term
2. More Loans given with Loan grades B, A, C
3. More people applying for loan with 10+ years and 0 to 3 years experience
4. People own the house the doesn't apply for loan much compared to rented and mortgage people

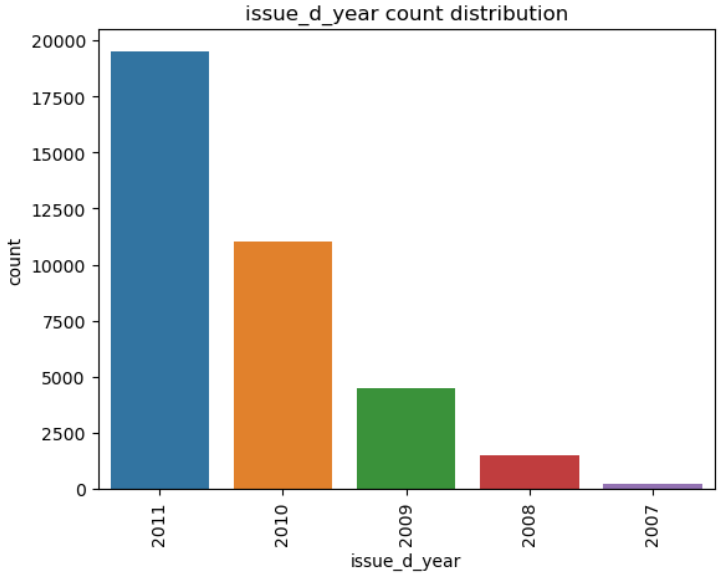


Univariate Analysis on Categorical Variables



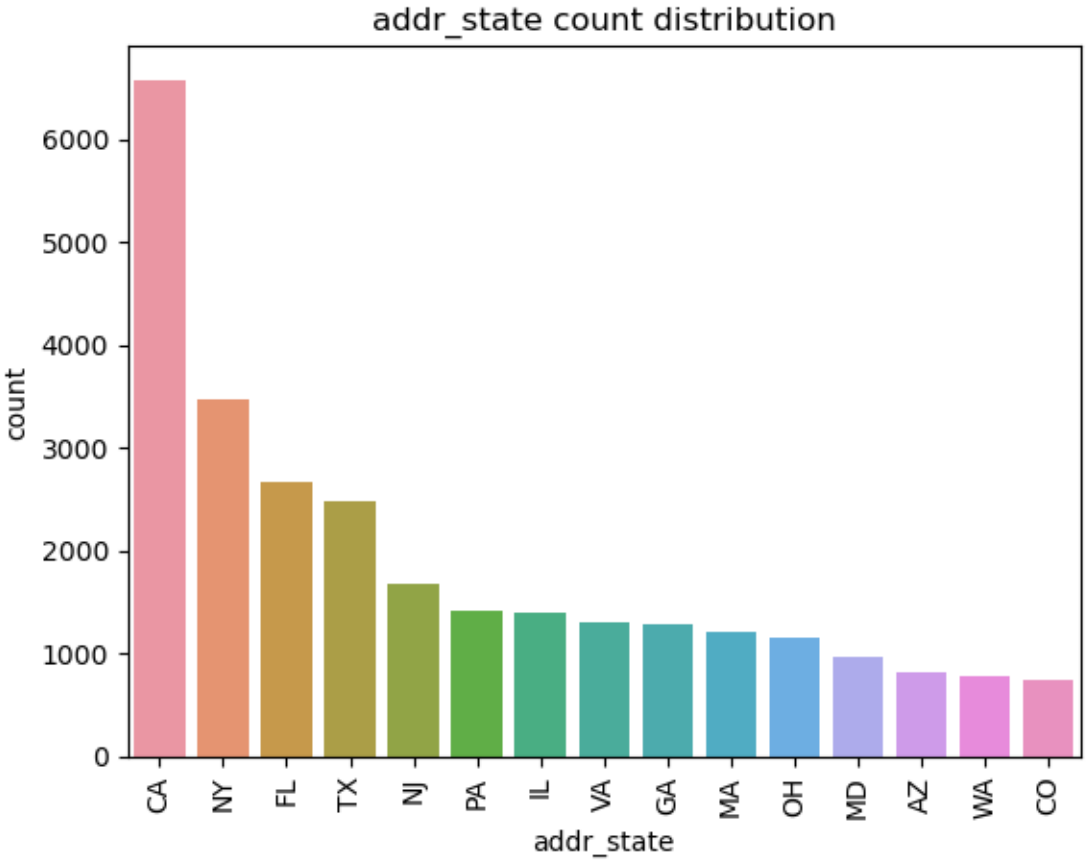


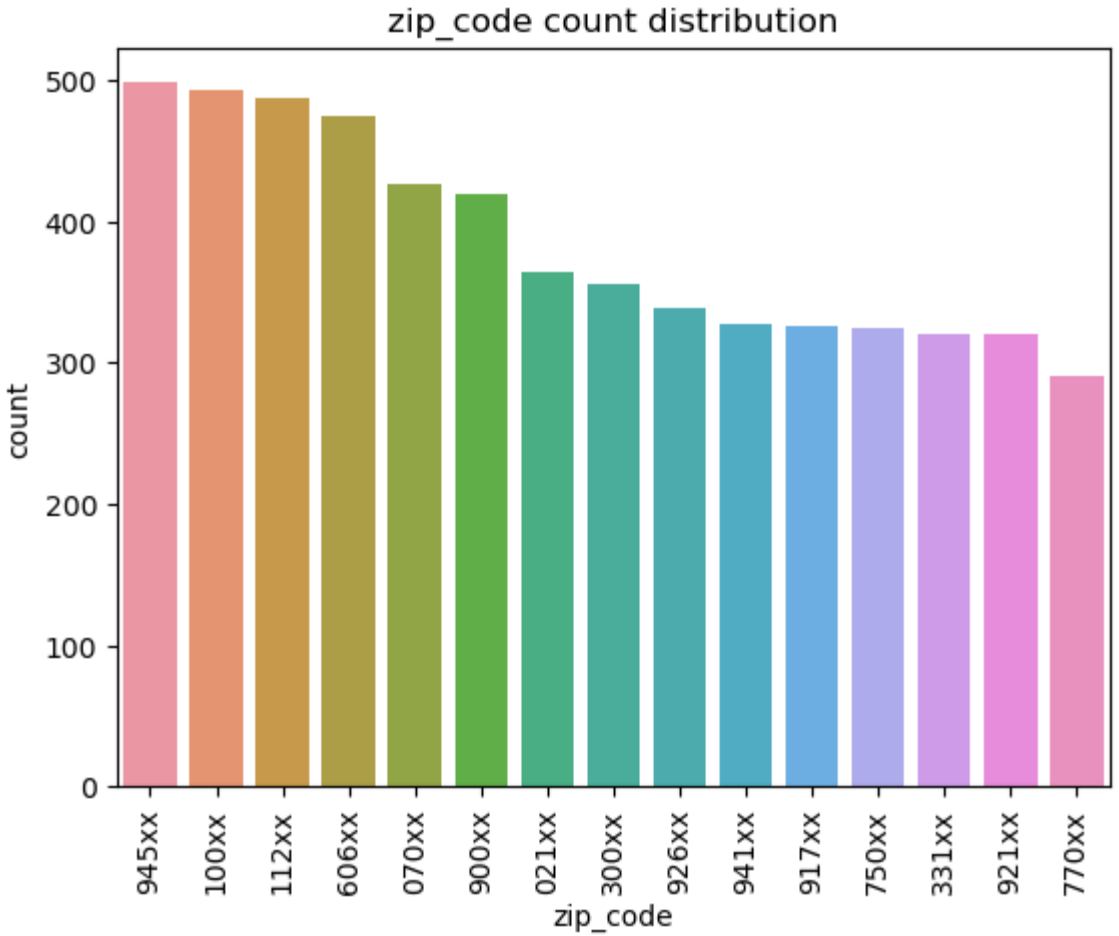




1. Most loans are taken for the purpose of debt_consolidation, credit_card, home_improvement
2. Most of the loan application falls in not “Not verified” category
3. Most Loans taken at the end of the year at Dec, Nov, Oct
4. Number of loans taken steadily increasing over the years, This might be due to the company getting popularity and growth

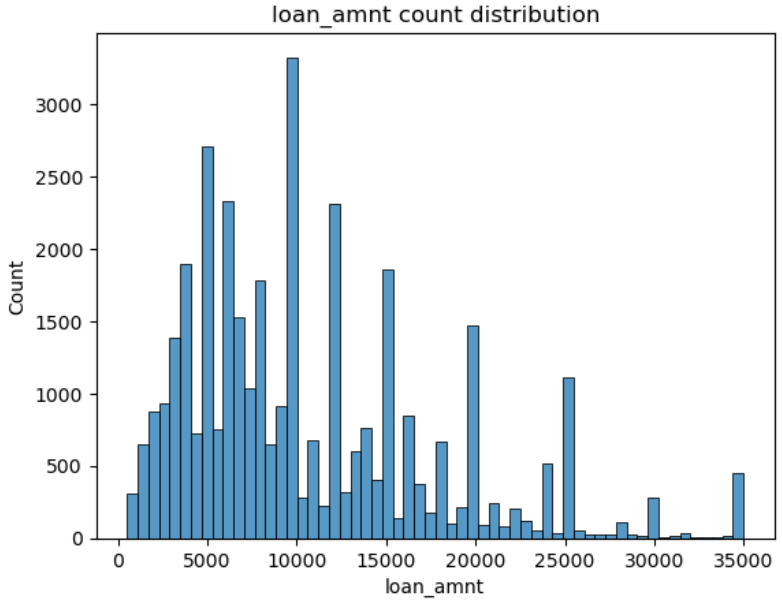
Univariate Analysis on Categorical Variables

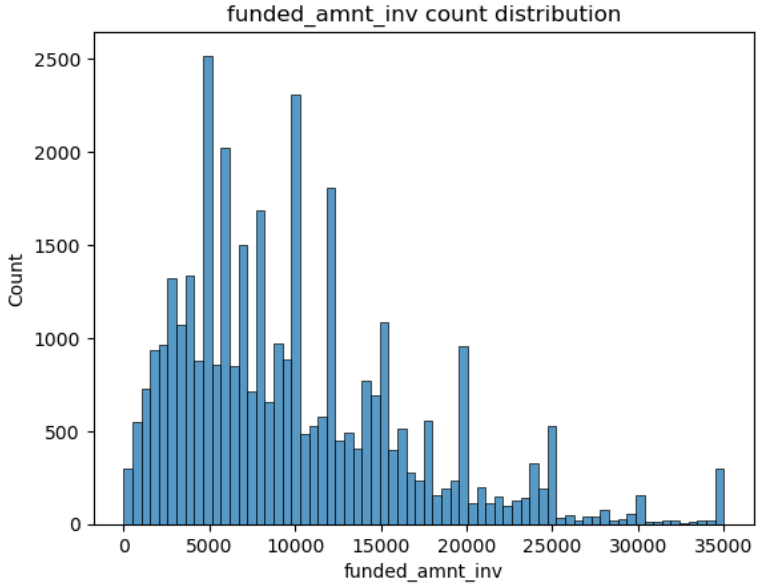


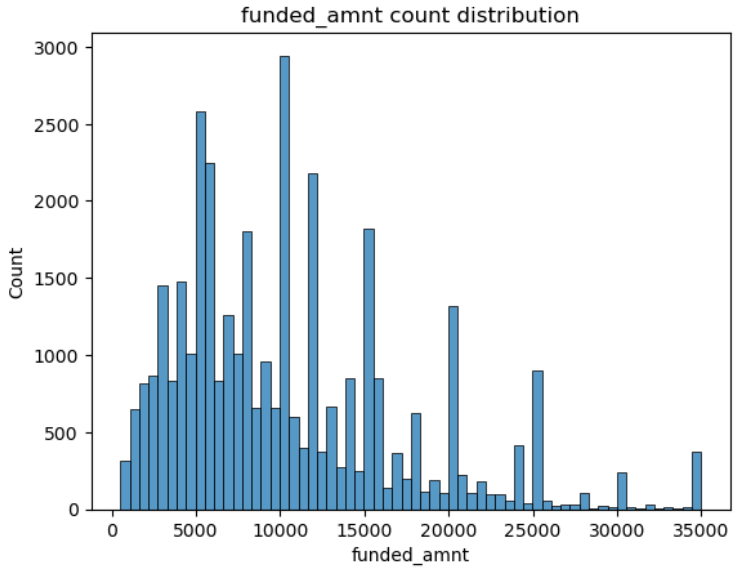


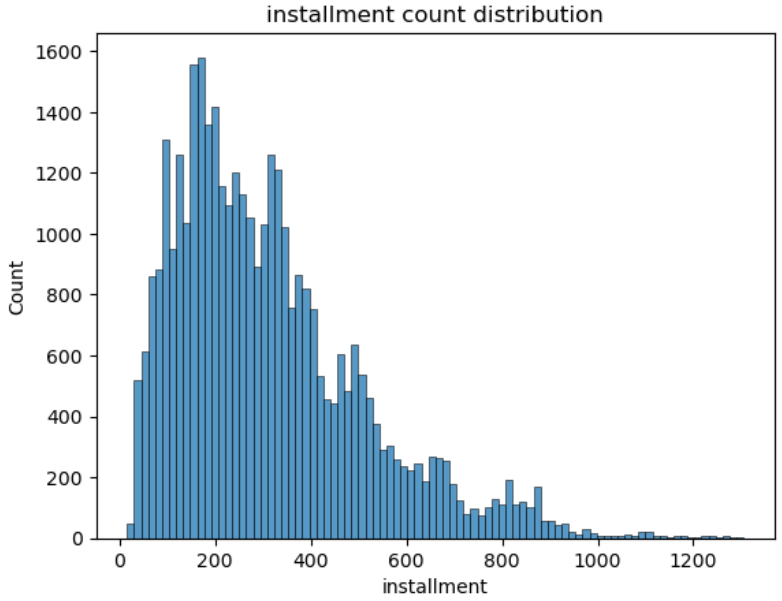
Most loans taken at states California, New York, Florida, Texas

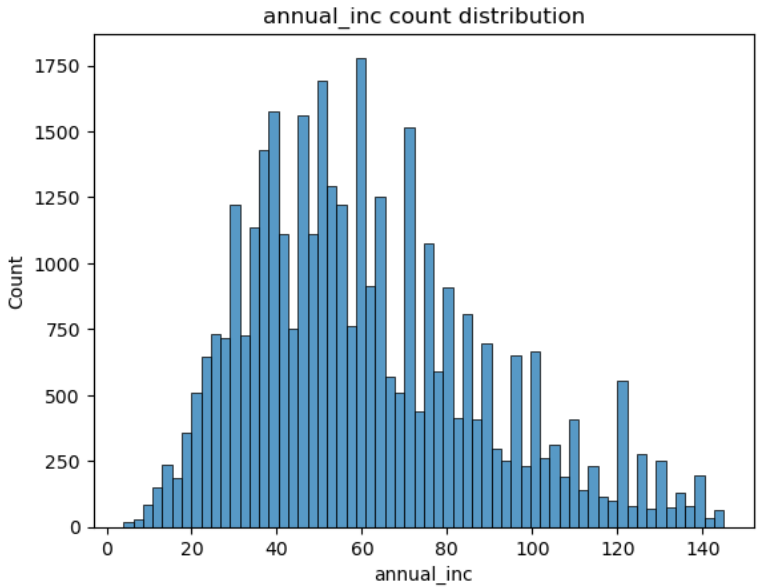
Univariate Analysis on Continuous Variables

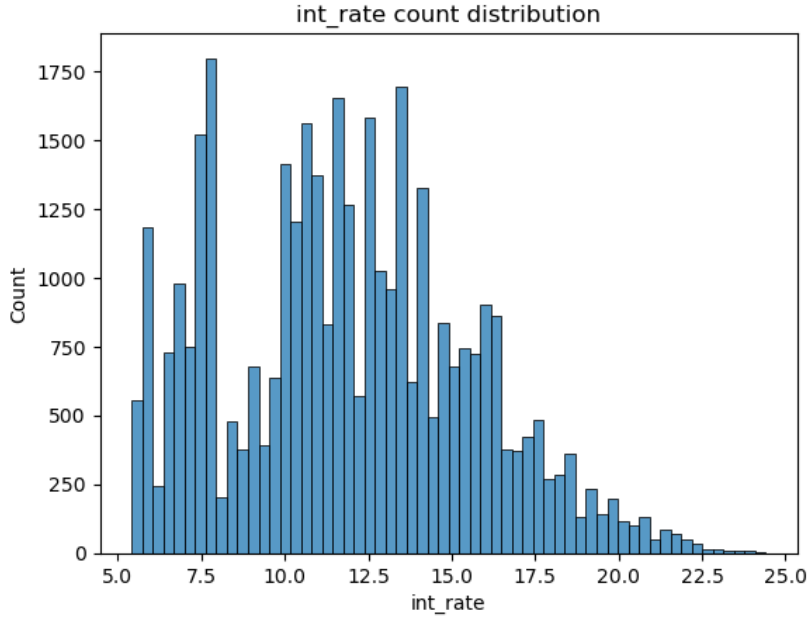




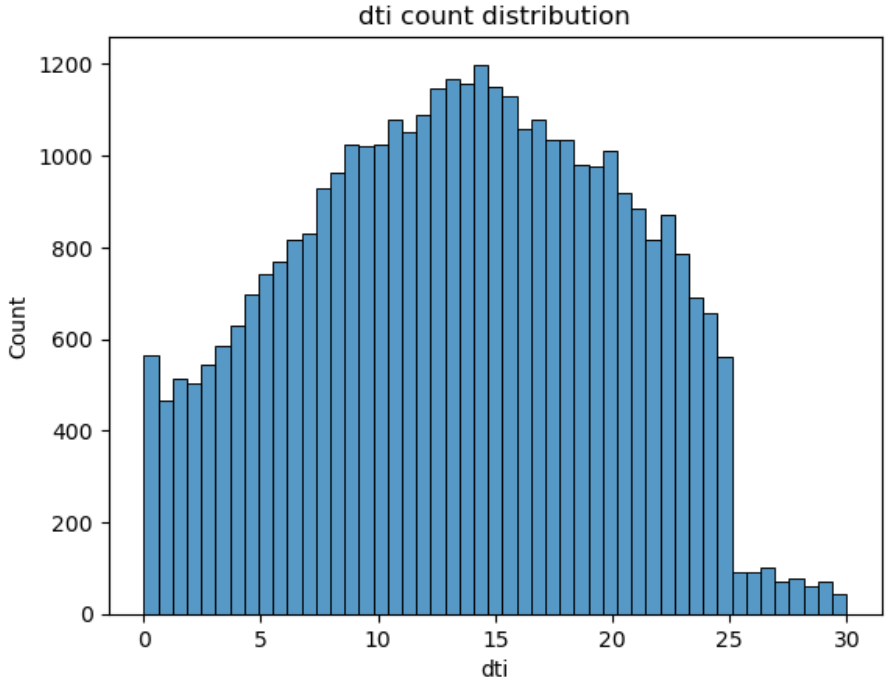


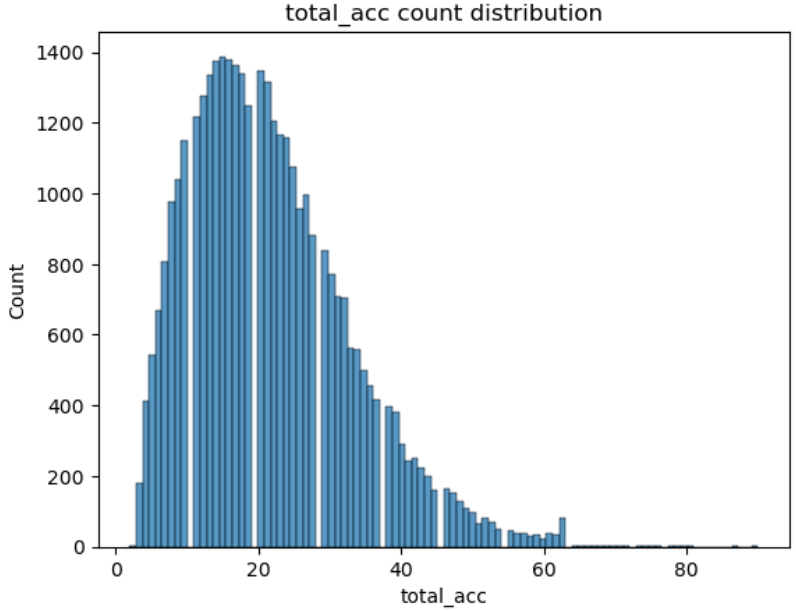




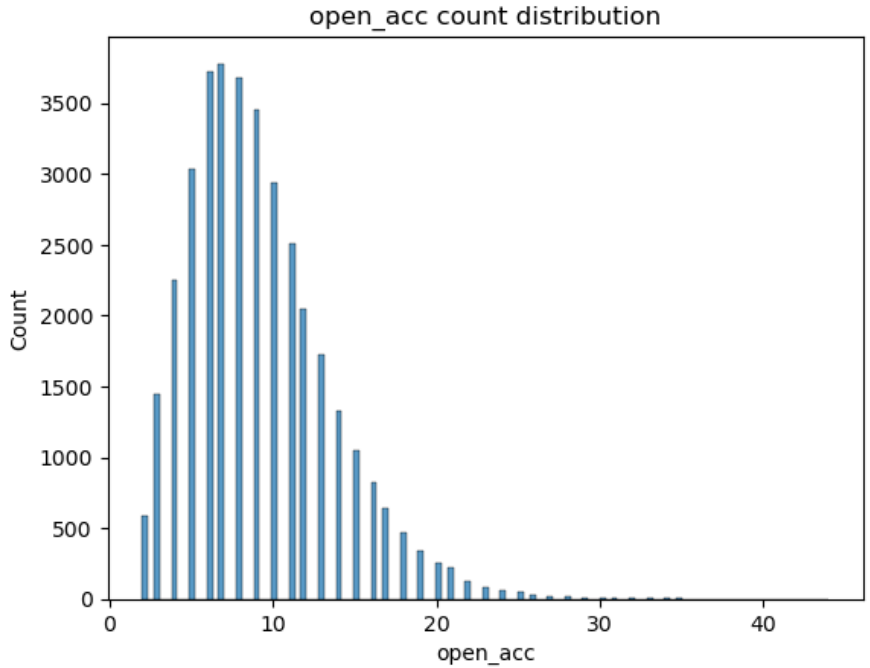


Univariate Analysis on Continuous Variables



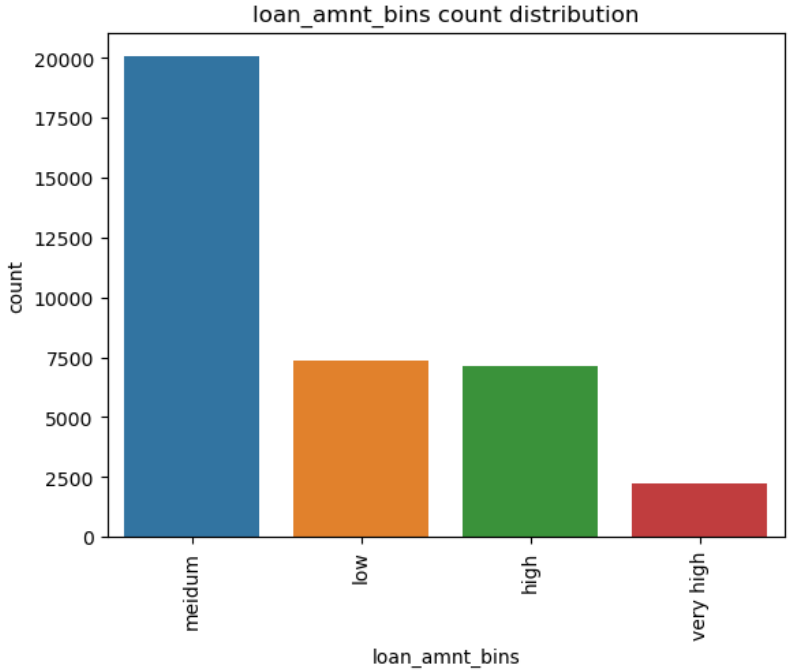


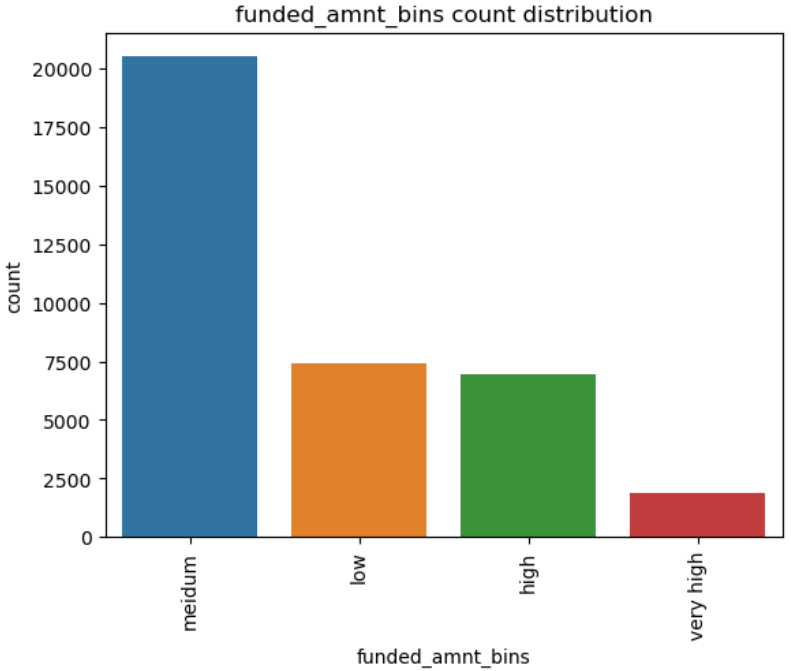
1. Loan amount, funded amount, investor funded amount more loan present in the range 5000 to 15000, Also Spikes explains that most of the loans are taken as multiple of 5000
2. More loan having interest rate between 10 to 15 %
3. More people taking loans at the income range of 40 to 80 K
4. Loans are concentrated with 100 to 400 range installment amount
5. Most Loans having dti 10 to 20
6. Total accounts gradually increase till 20 then drops
7. Open accounts gradually increase till 10 then drops

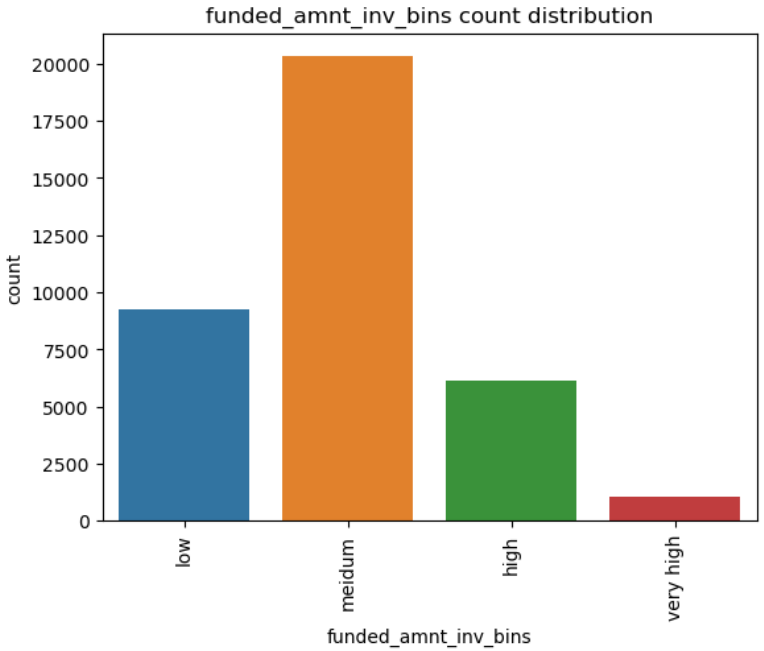


Univariate Analysis on Binned Columns

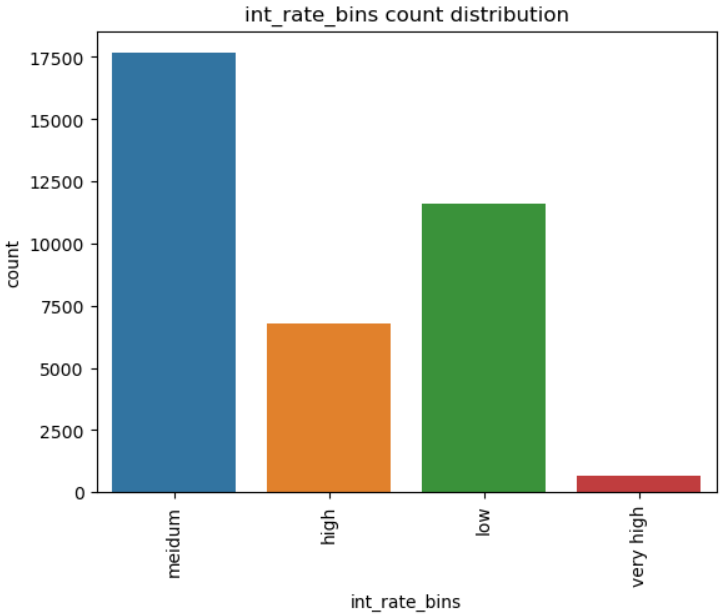
1. Loan amount, funded amount, funded amount investor – More loans taken for medium value of amount
2. funded amount investor – investors fund medium loans for decent return and low for safe return compared to high and low



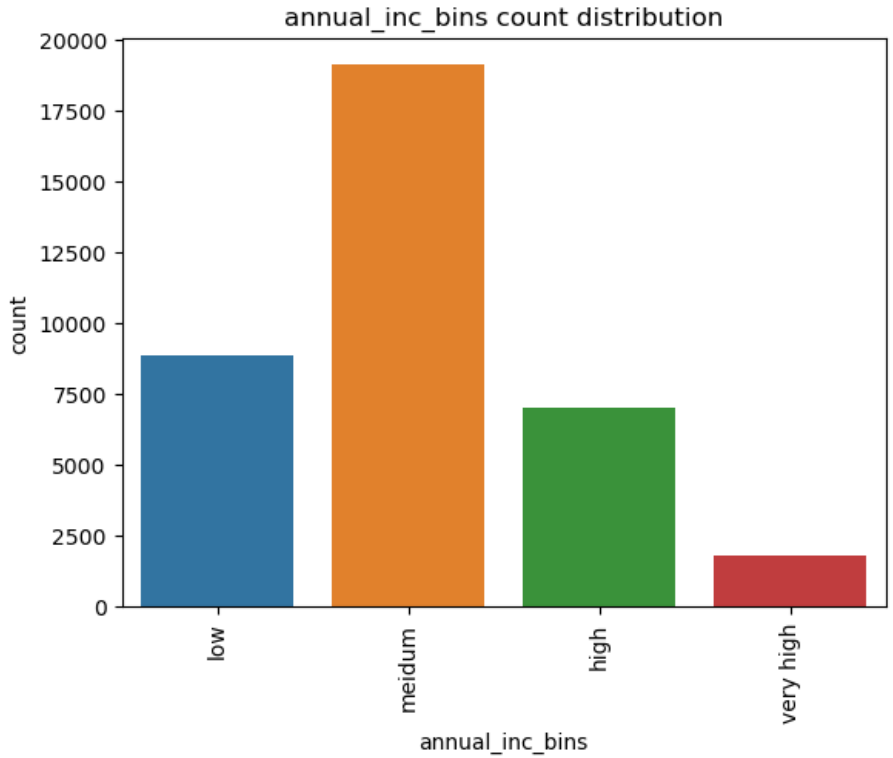


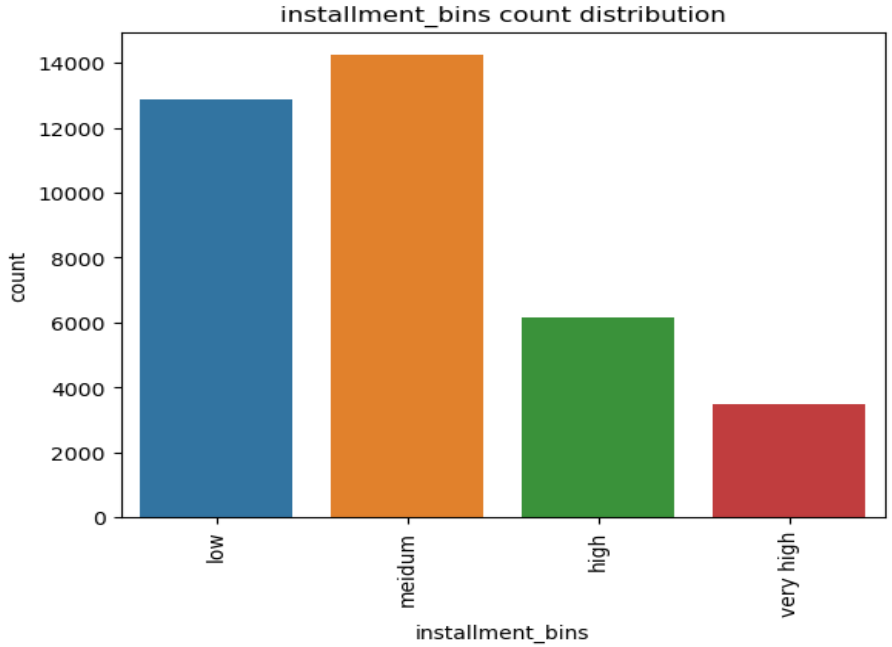


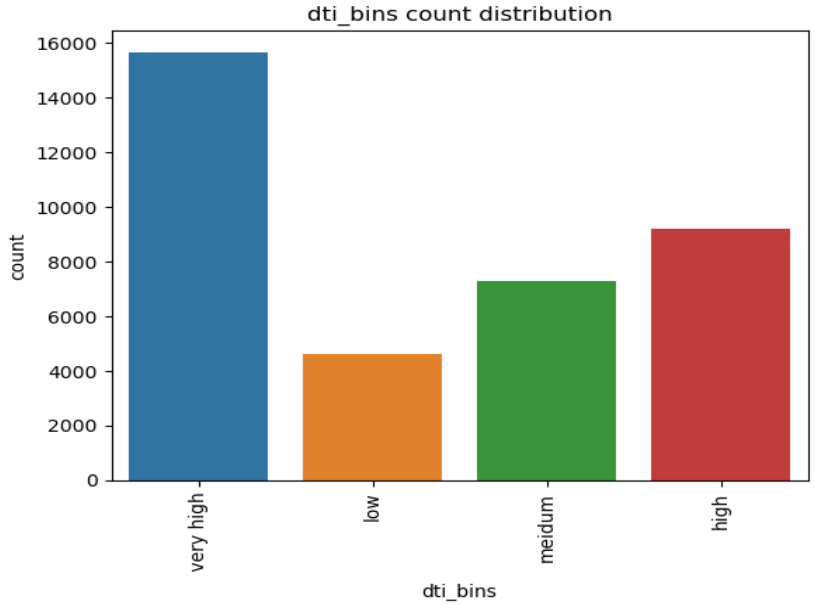
1. More Loans given at medium interest rate



Univariate Analysis on Binned Columns

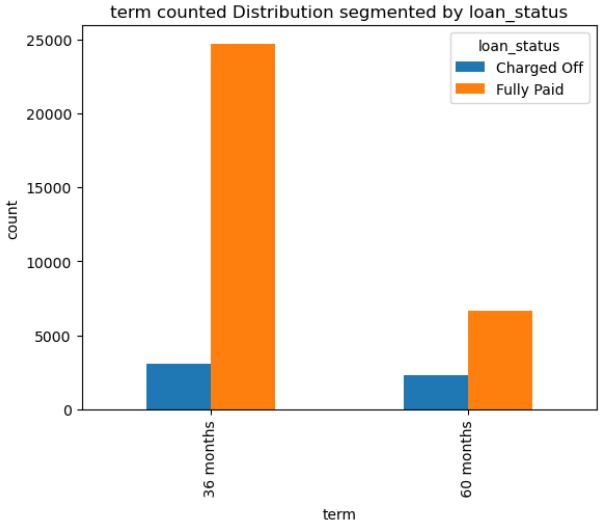


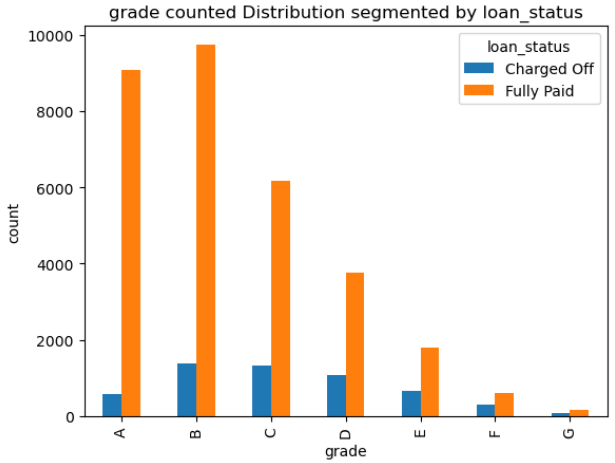


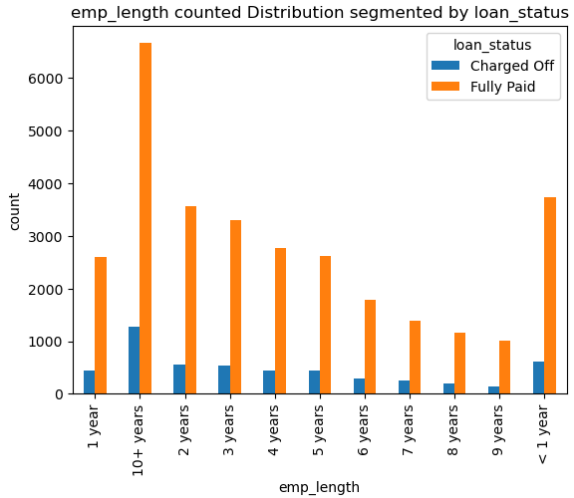


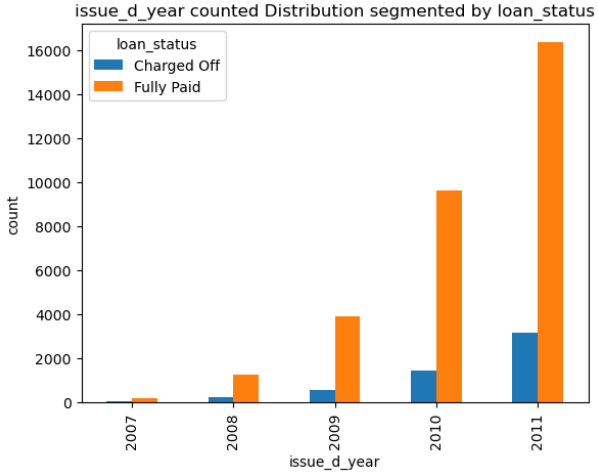
1. Most people takes loan are at medium annual income category
2. Most loans are given at medium and low installment amount
3. Most loan taken have vey high dti

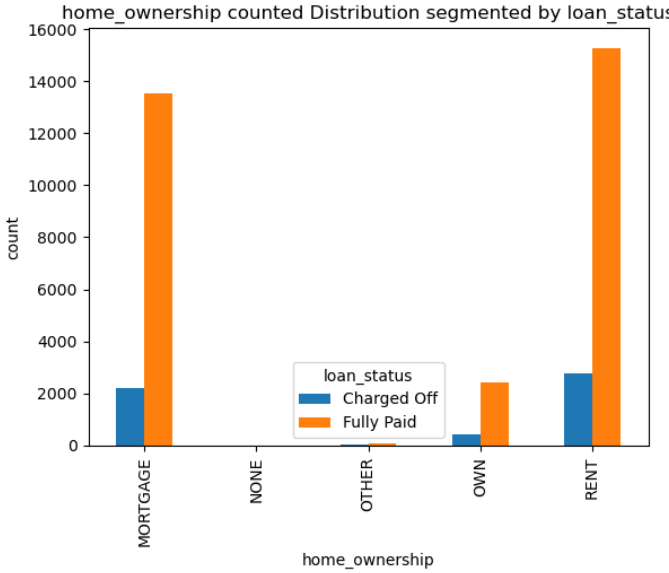
Segmented Univariate Analysis on Categorical Variables

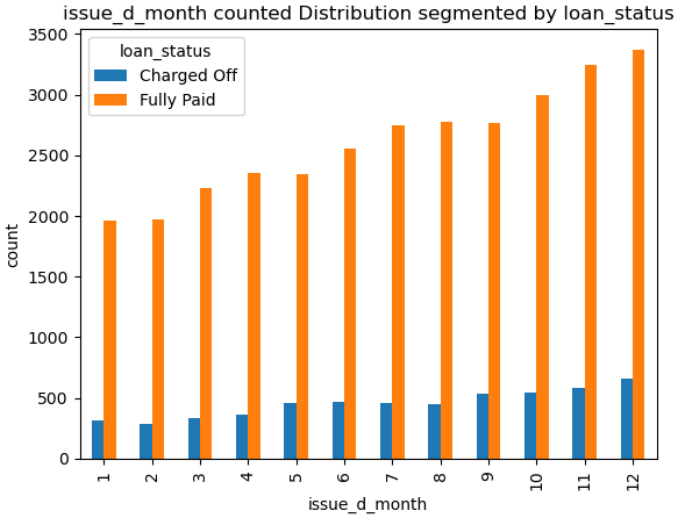


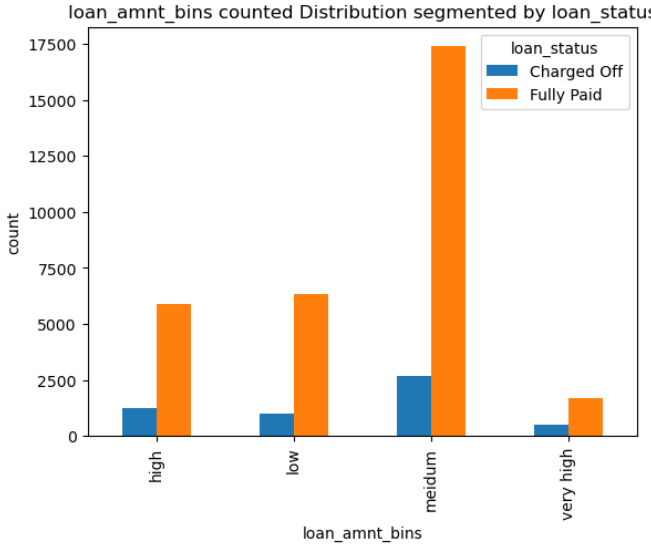




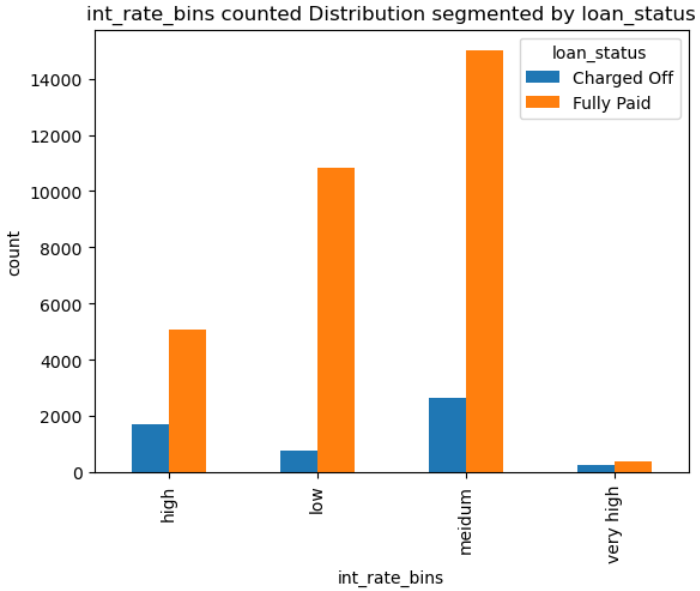


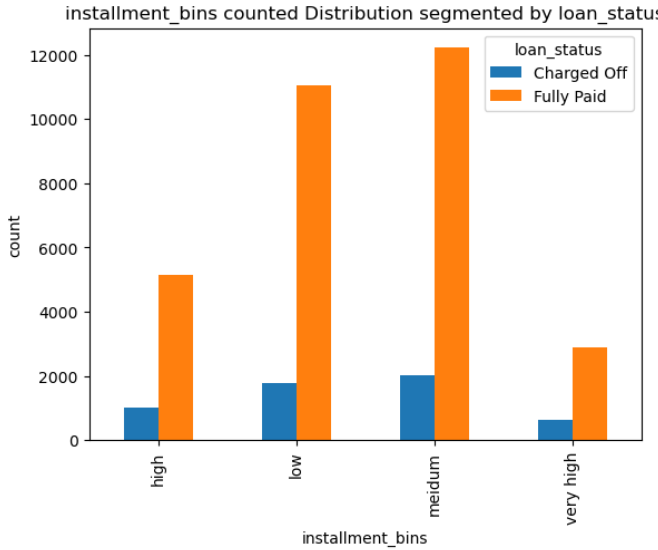


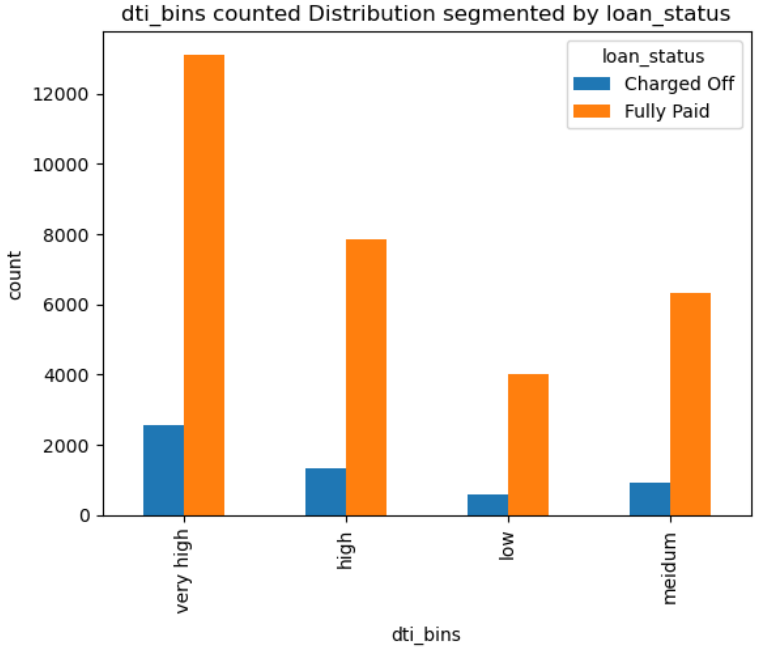




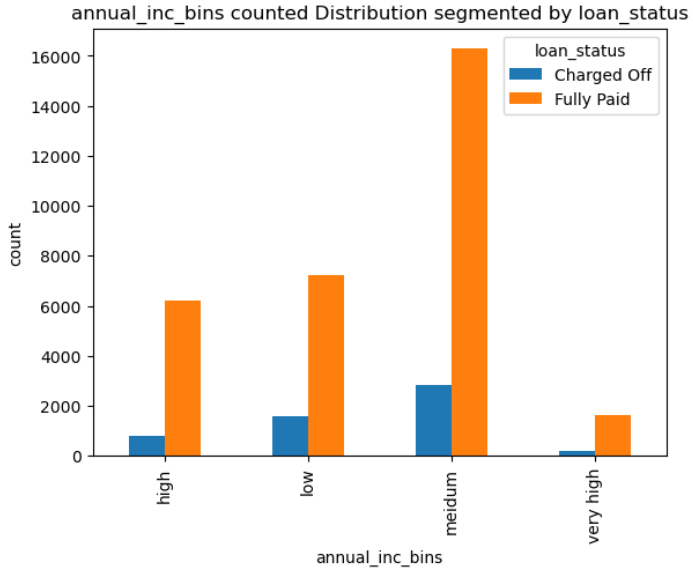
Segmented Univariate Analysis on Categorical Variables



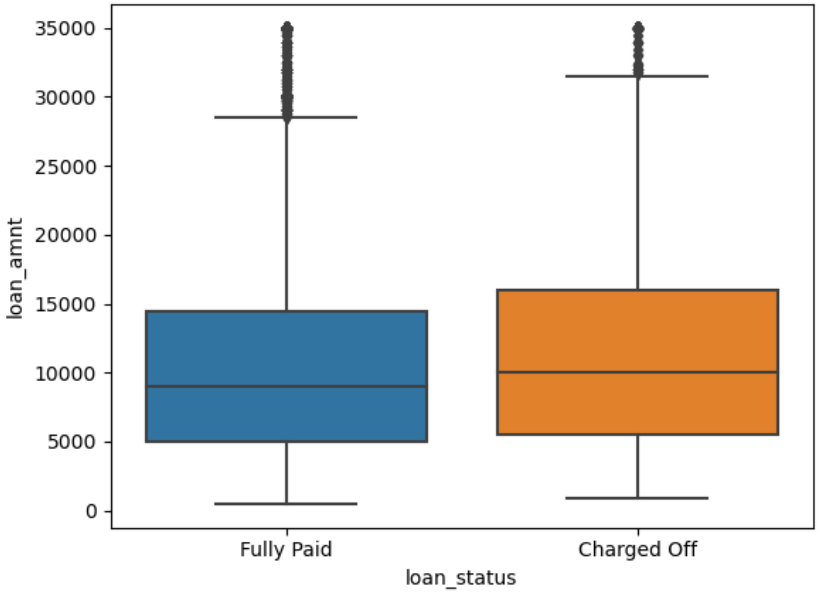


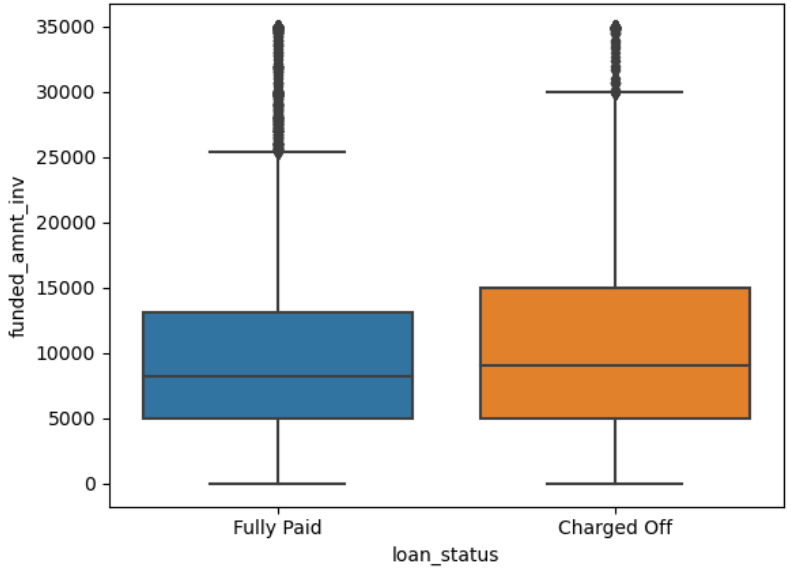


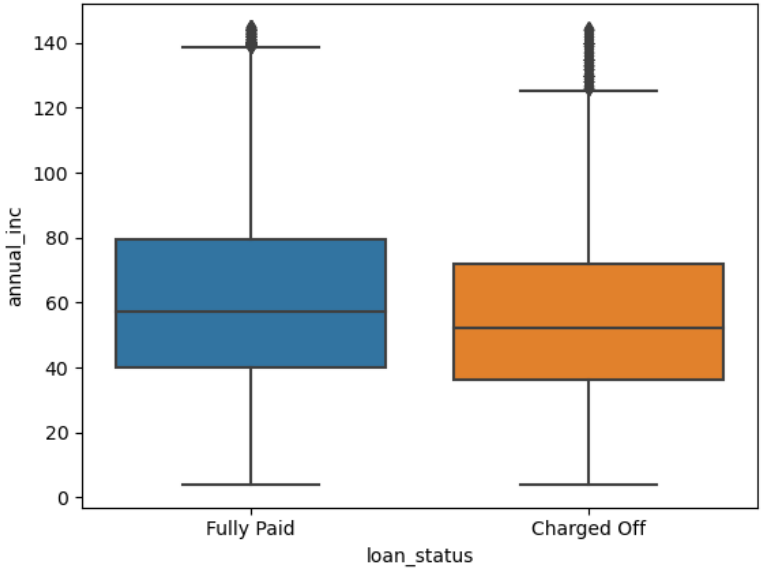
1. Counts plot provides same insights as univariate analysis
2. Successfully paid insights looks as a scaled version of charged off
3. We can derive more insights by comparing against rate of default



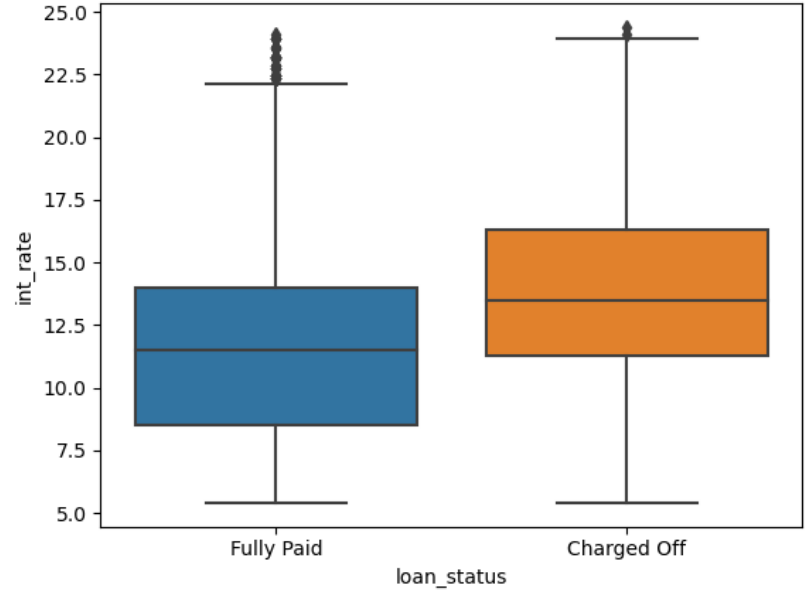
Segmented Univariate Analysis on Continuous Variables





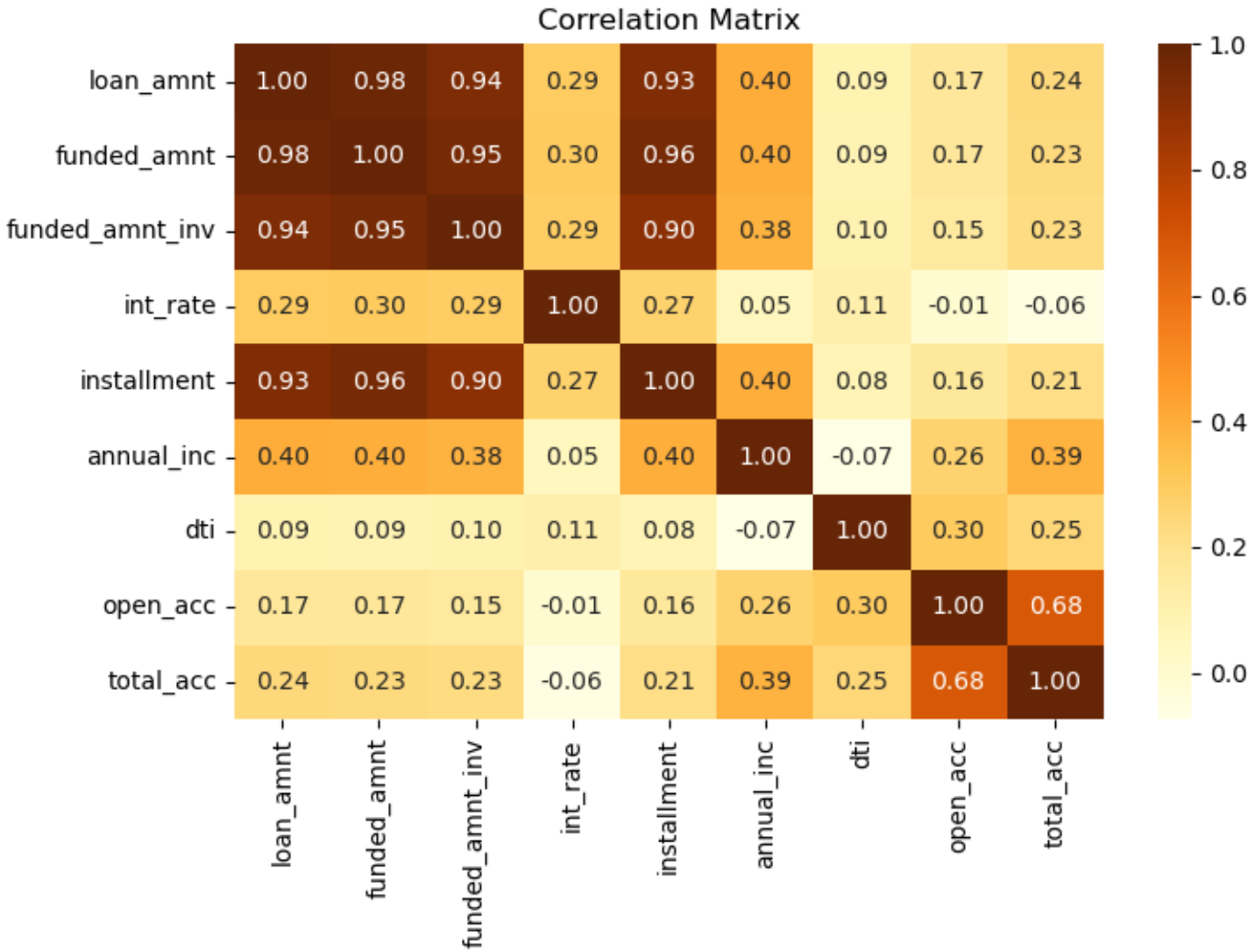


1. More Loans are getting charged off at High and Very High loan amount category
2. More Loans are getting charged off at High and Very High interest rate category
3. People at low income category defaulting more loans

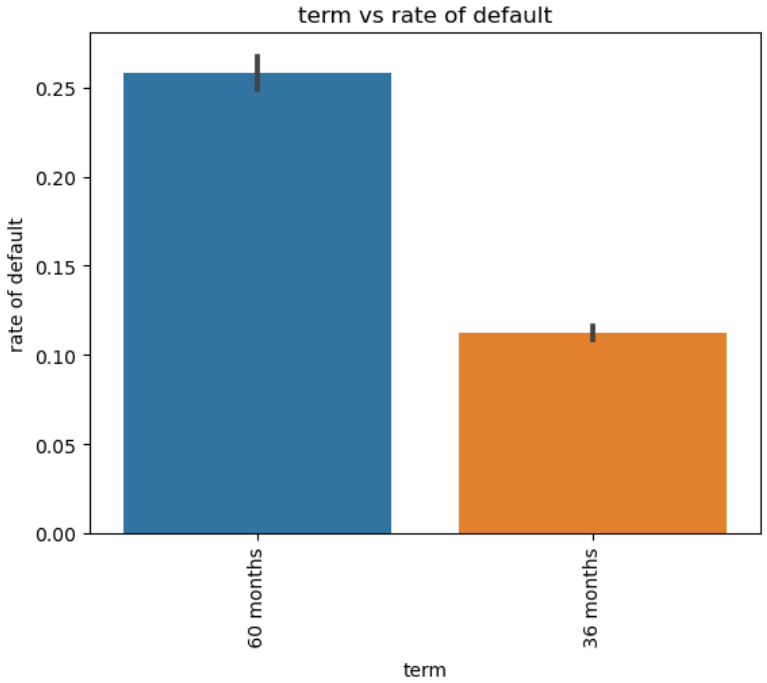


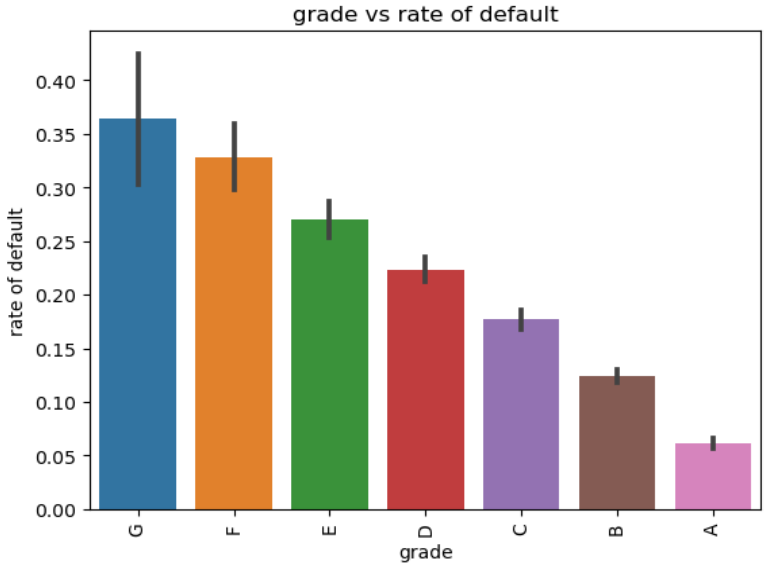
Bivariate Analysis on Continuous Variables

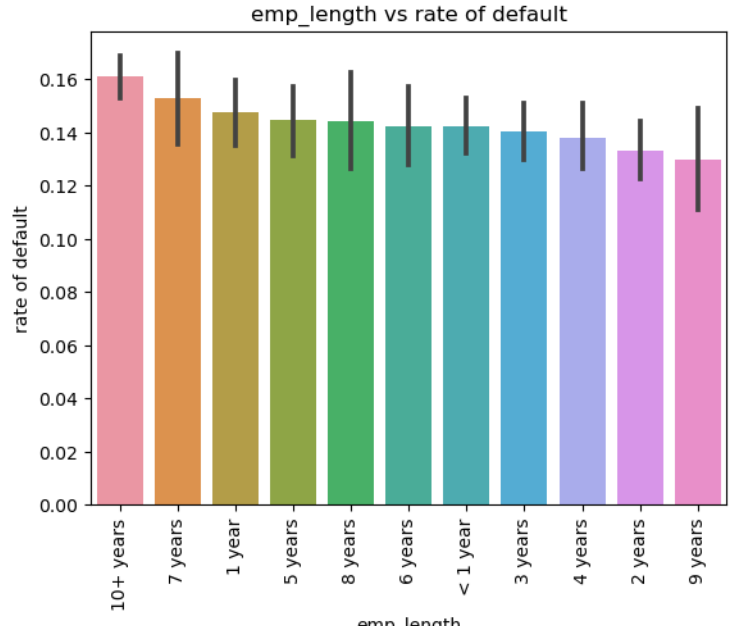
1. **loan_amt,funded_amt,funded_amnt_inv Installment** are highly positive corelated with each other (>0.9)
2. **Annual_income** having decent positive correlation with **loan_amt,funded_amt,funded_amnt_inv, Installment** (~ 0.4)
3. **int_rate** having decent positive correlation with **loan_amt,funded_amt,funded_amnt_inv, Installment** (~ 0.3)
4. total_acc decent positive correlation with annual income (~0.4)
5. dti not having much correlation with other variables



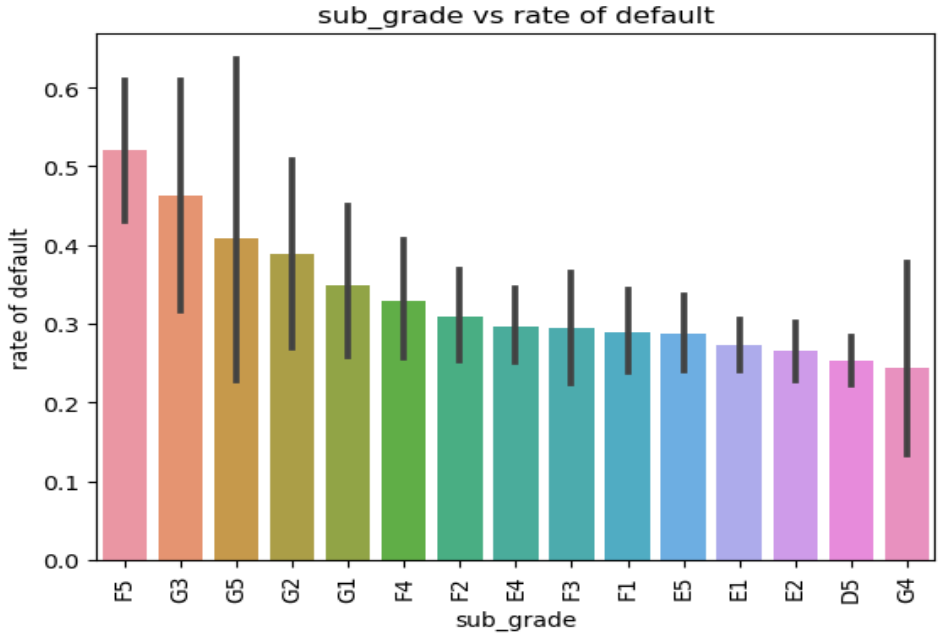
Bivariate Analysis (Rate of Default vs Categorical Variables)



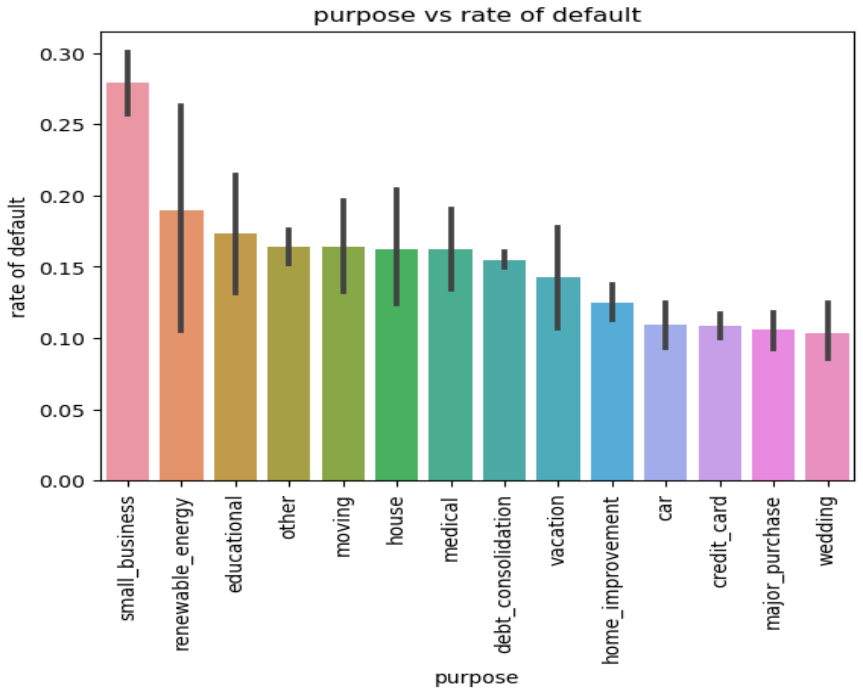


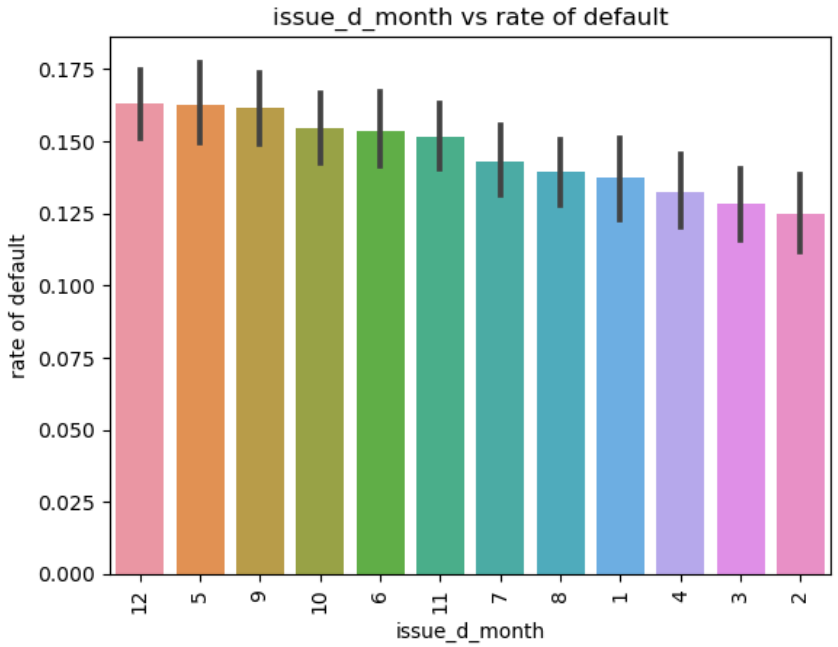


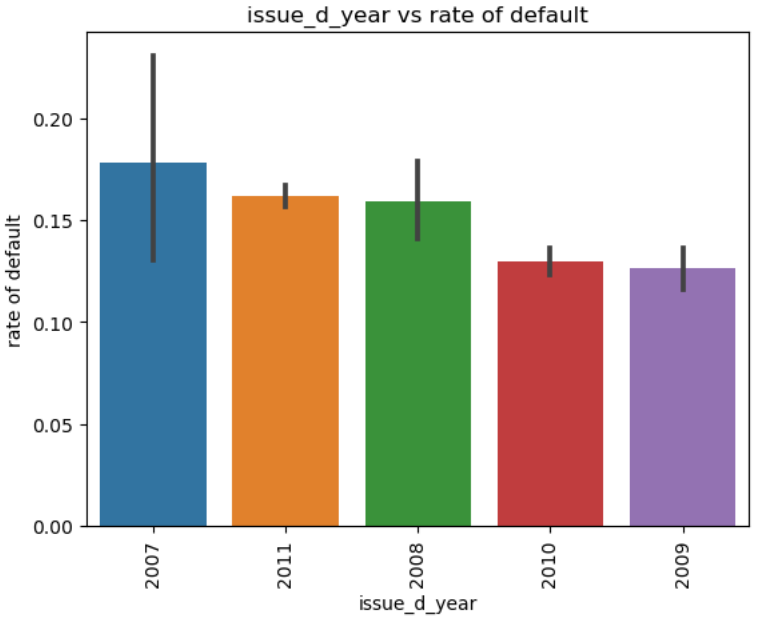
1. **Term** - High Default rate on 60 months term ~ 26%
2. **Grade** - Default rate increase from loan grade A to G, risk is increasing from A to G (6 to 36%)
3. **Sub Grade –** F5(52%),G3(46%),G5(40%),G2(38%) are more risk loans
4. **Employment Length** - Higher default rate occurs when emp_length is 10+ Years. Only 2% diff between lowest to highest default rate



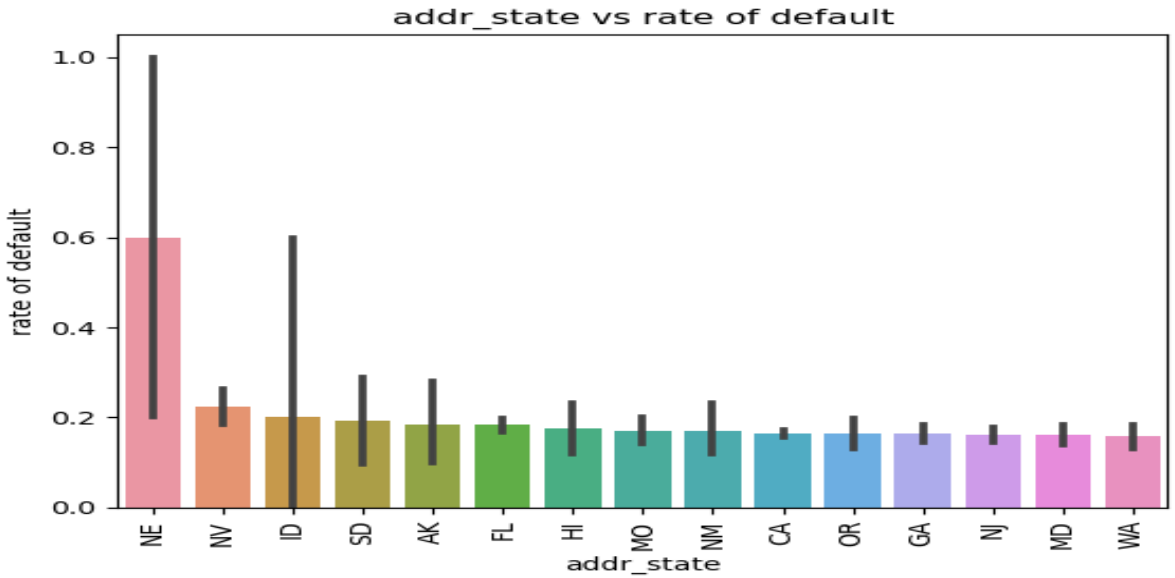
Bivariate Analysis (Rate of Default vs Categorical Variables)



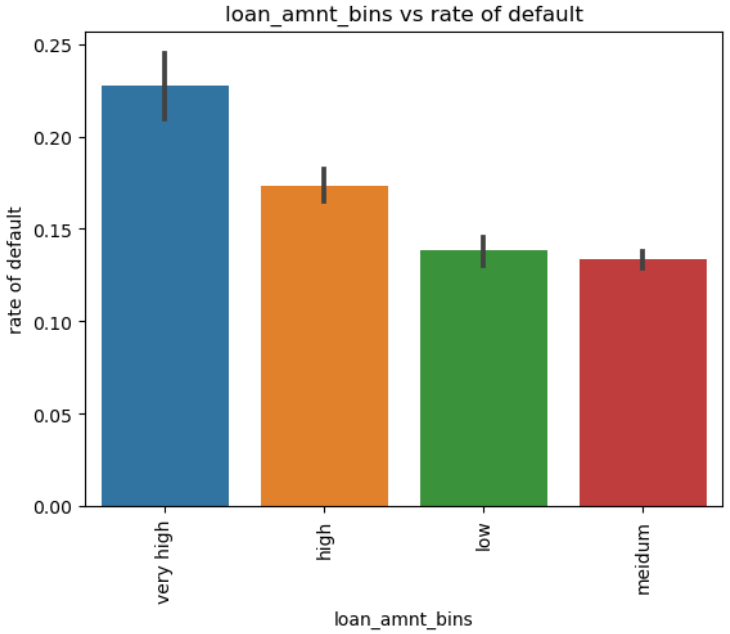


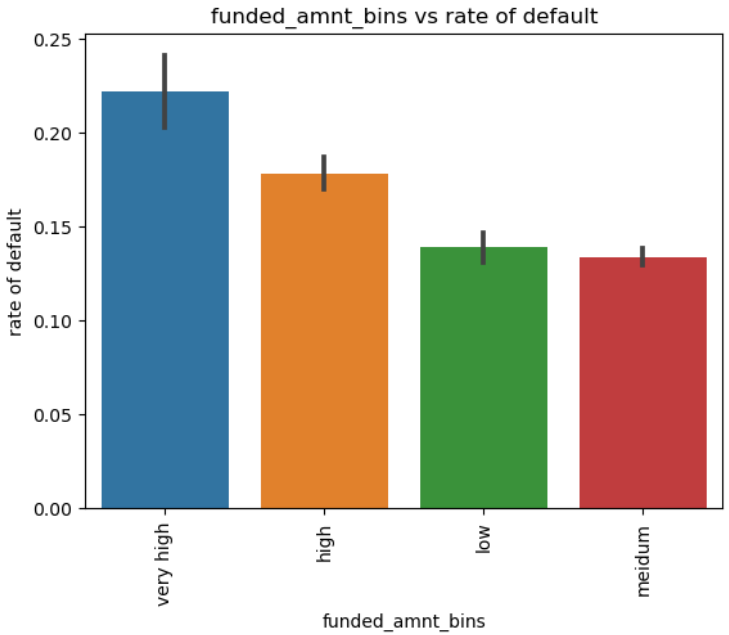


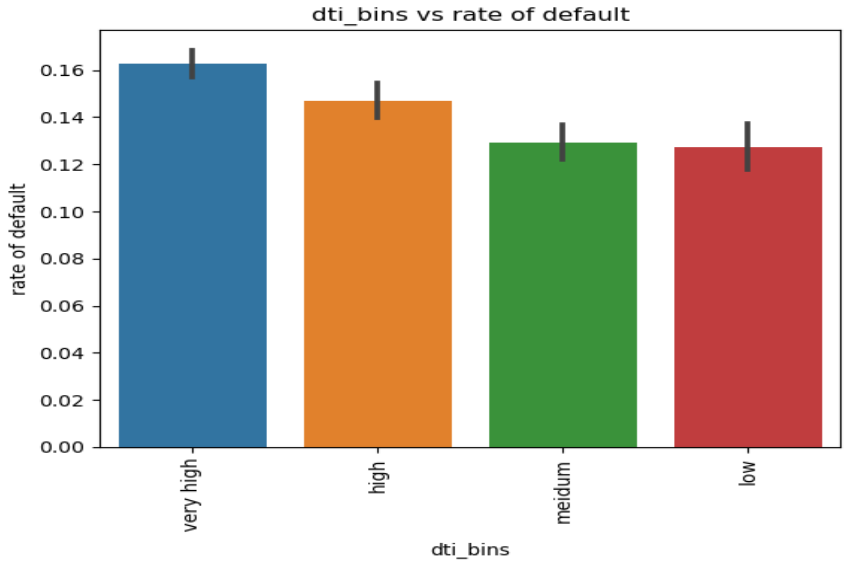
1. **Purpose** – People taking loans for the purpose small business(27%), renewable_energy(19%), educational(17%) tends to default more
2. **Issue Date** – 2007 had default rate of 18% compared to low on 2009(12%). 2007 might be impacted by economic crisis
3. **Address State** - NE, NV,ID, SD, AK, FL ,HI has higher default rate



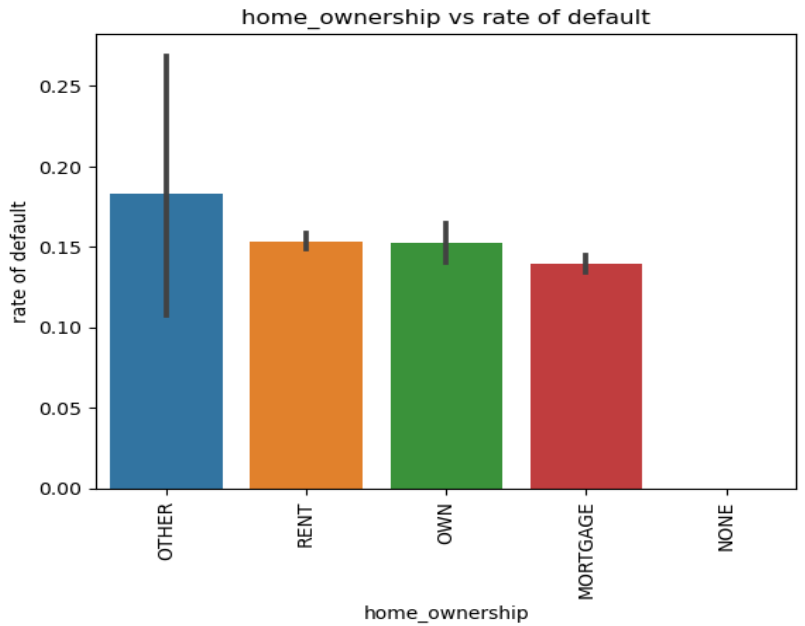
Bivariate Analysis (Rate of Default vs Categorical Variables)



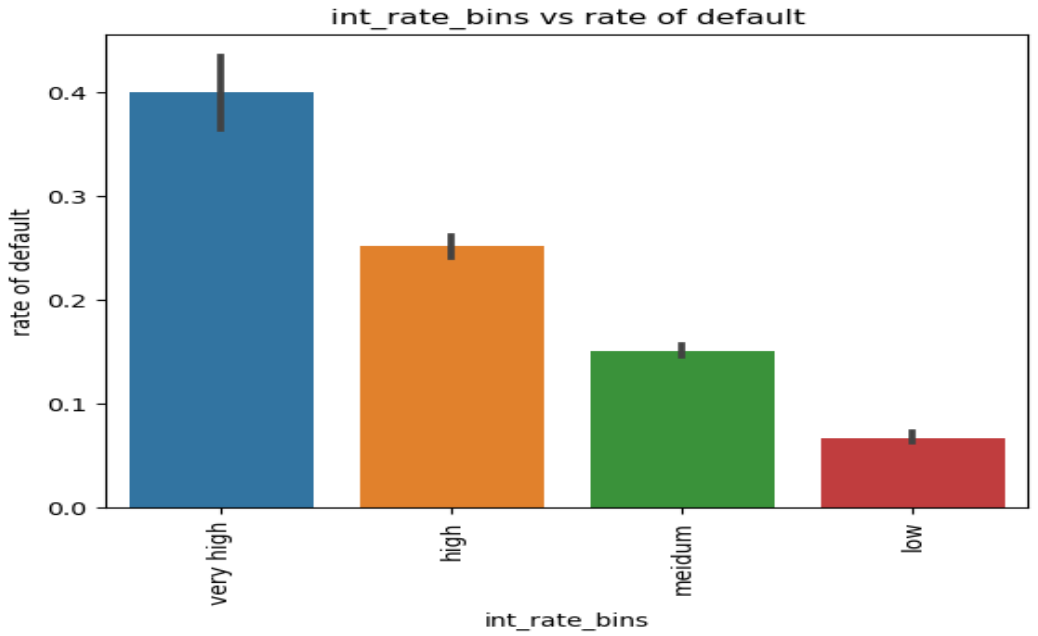


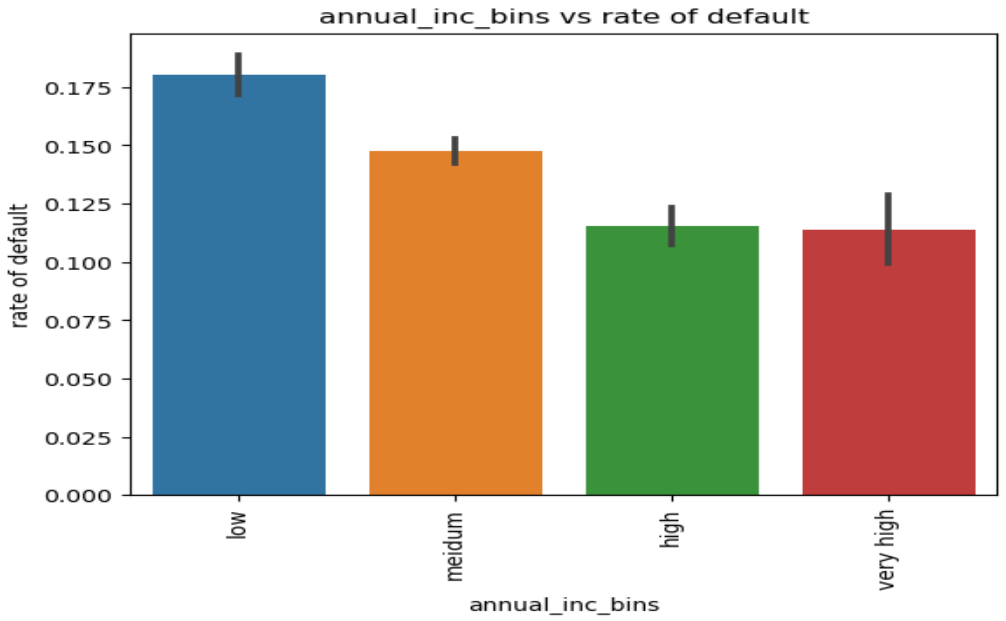


1. **loan_amnt, funded_amnt_bins , funded_amnt_inv_bins** having same patterns. Very High loan Amounts having default rate 22% compared to medium and low 13%
2. **dti** - Default rate increase from low to very high (12 to 16%)
3. **Home Ownership** - people with mortgage home_ownership tends to default lesser ~14% compared to highest Other ~18%

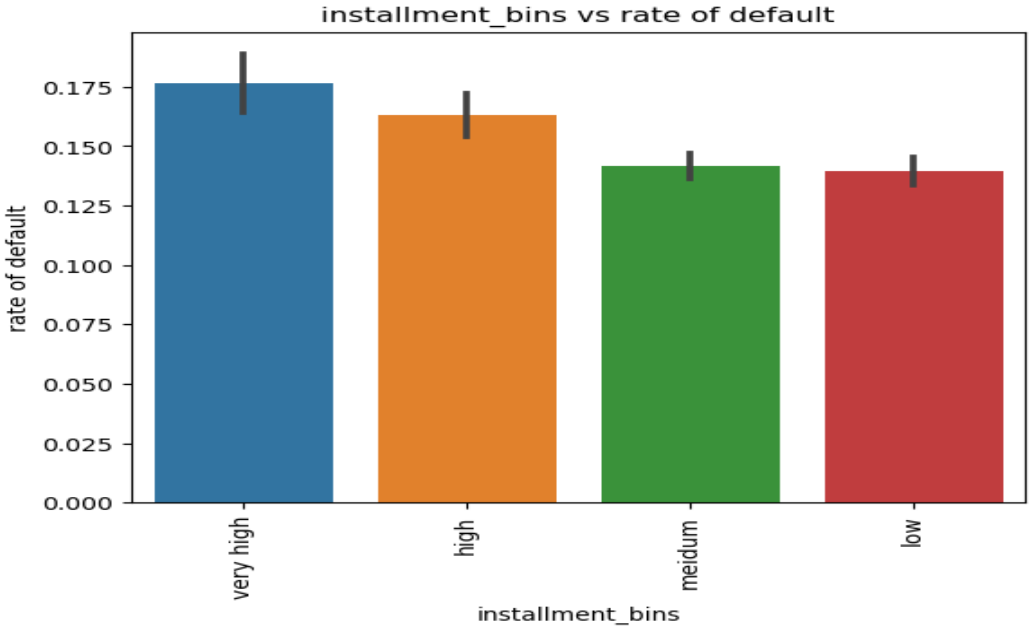


Bivariate Analysis (Rate of Default vs Categorical Variables)

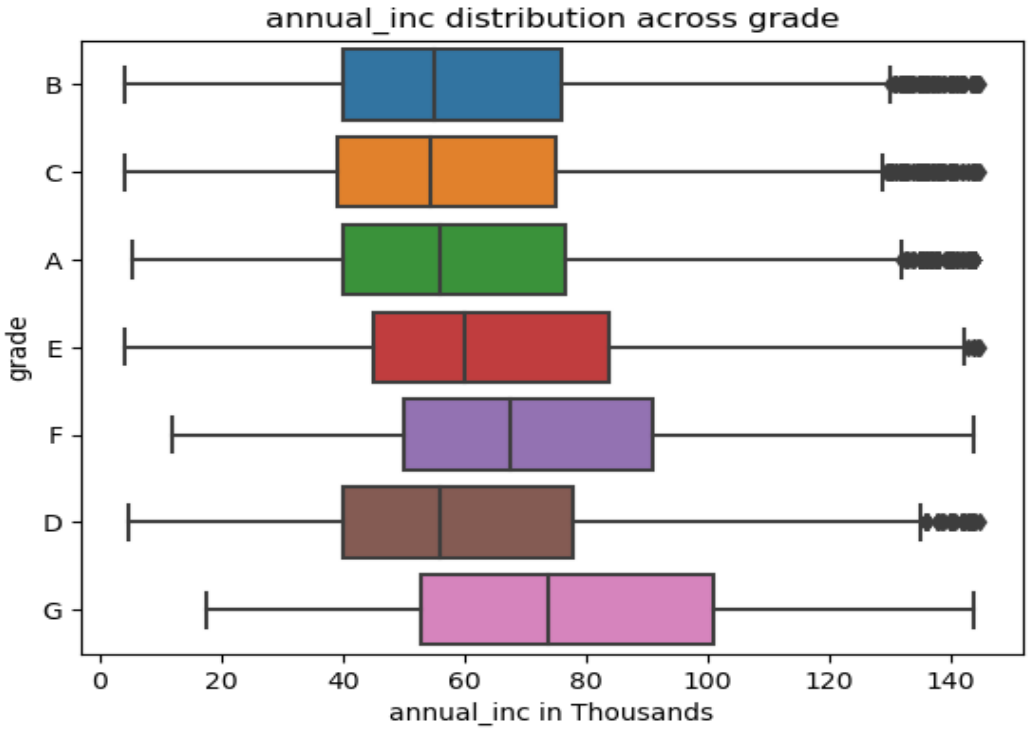


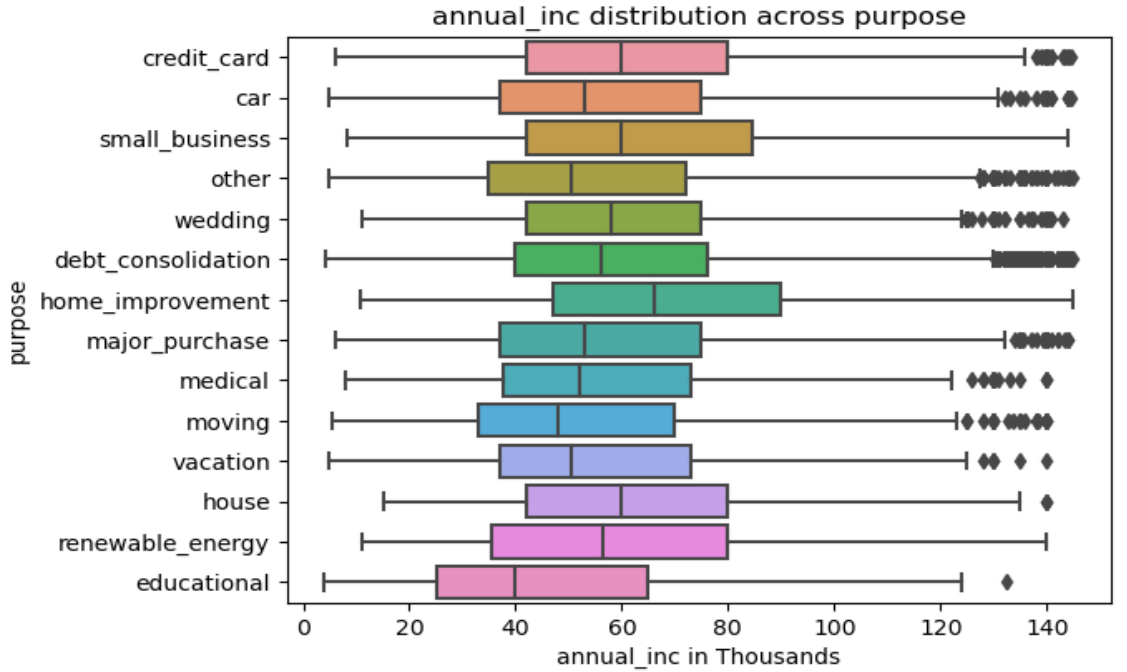


1. **int_rate** - Default rate increase from low to very High (6 to 40%)
2. **annual_inc** - Default rate **decrease** from low to very high (18 to 11%)
3. **installment** - Default rate increase from low to very high (14 to 18%)

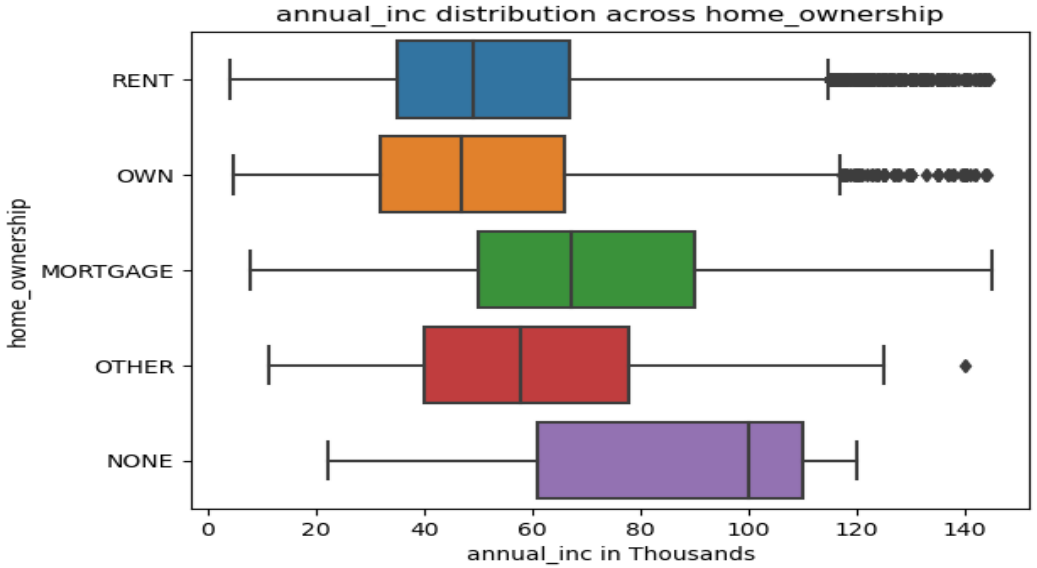


Bivariate Analysis (Annual Income vs Categorical Variables)

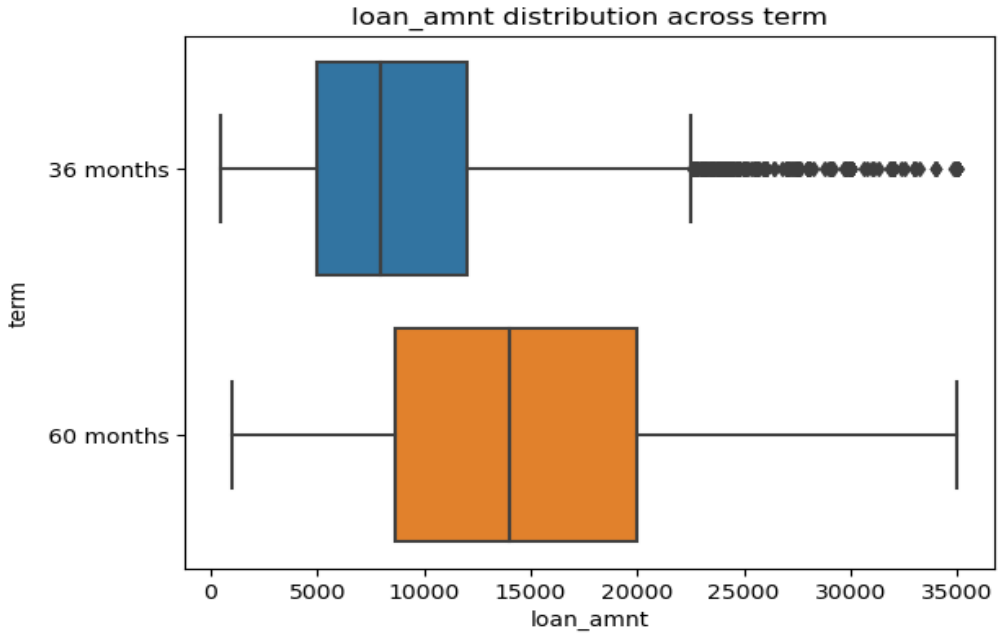


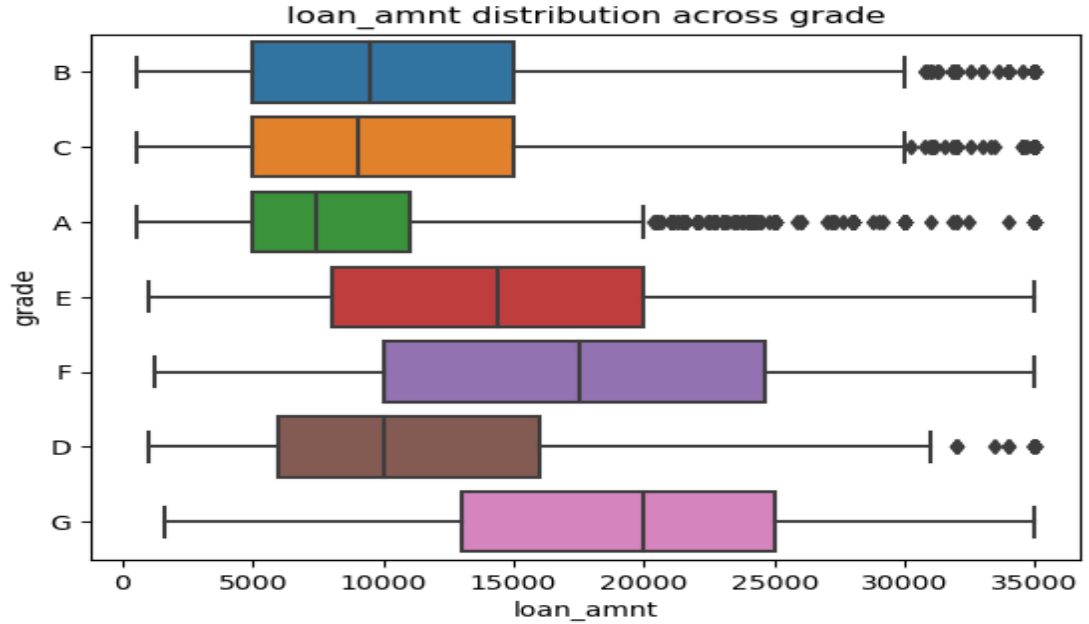


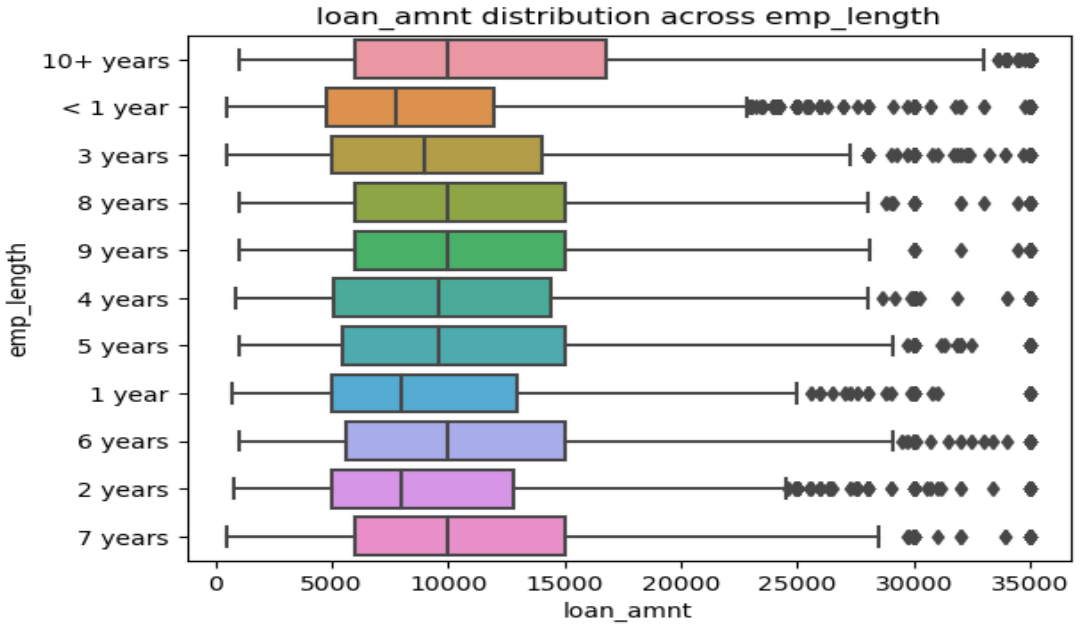
1. People with having income range 50 to 100K tends to take more G grade loans
2. People with having income range 40 to 90K tends to take more loans for home improvement and small business



Bivariate Analysis (Loan Amount vs Categorical Variables)

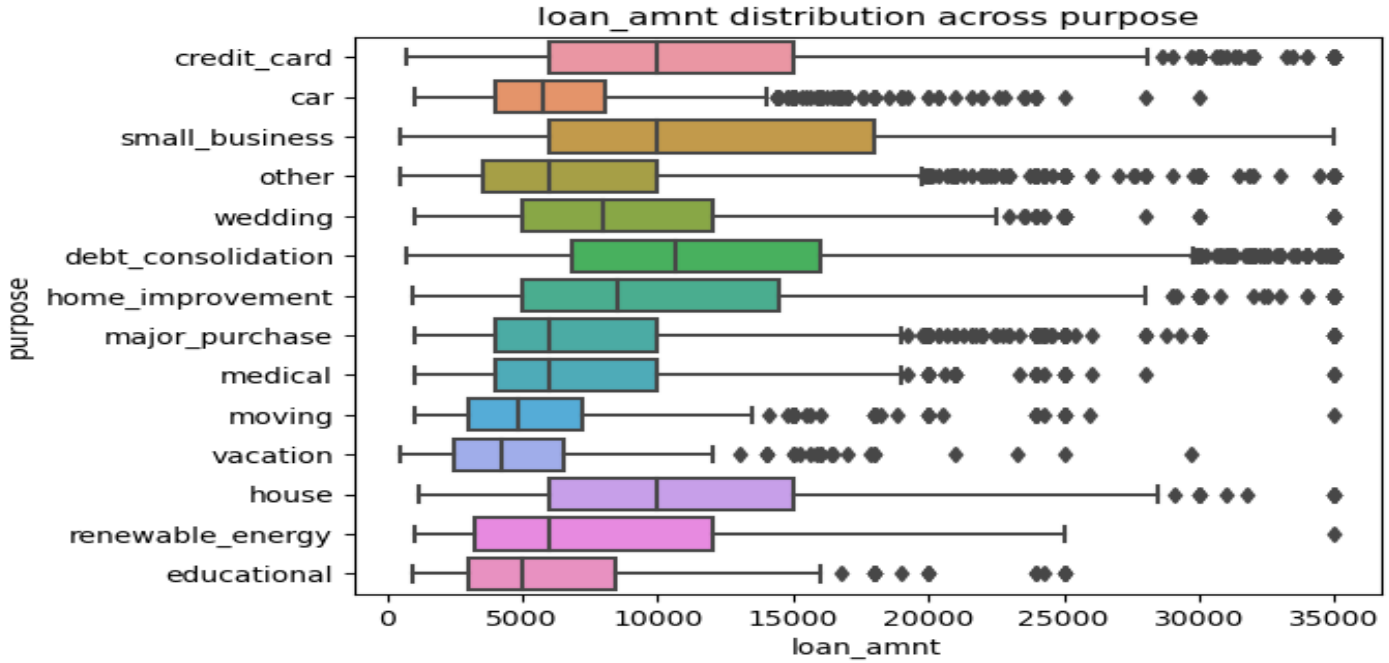


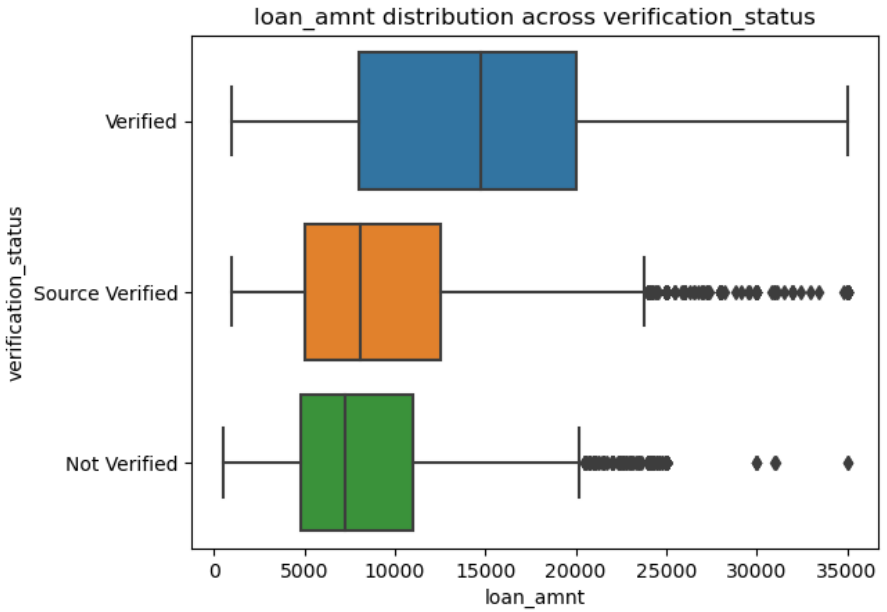




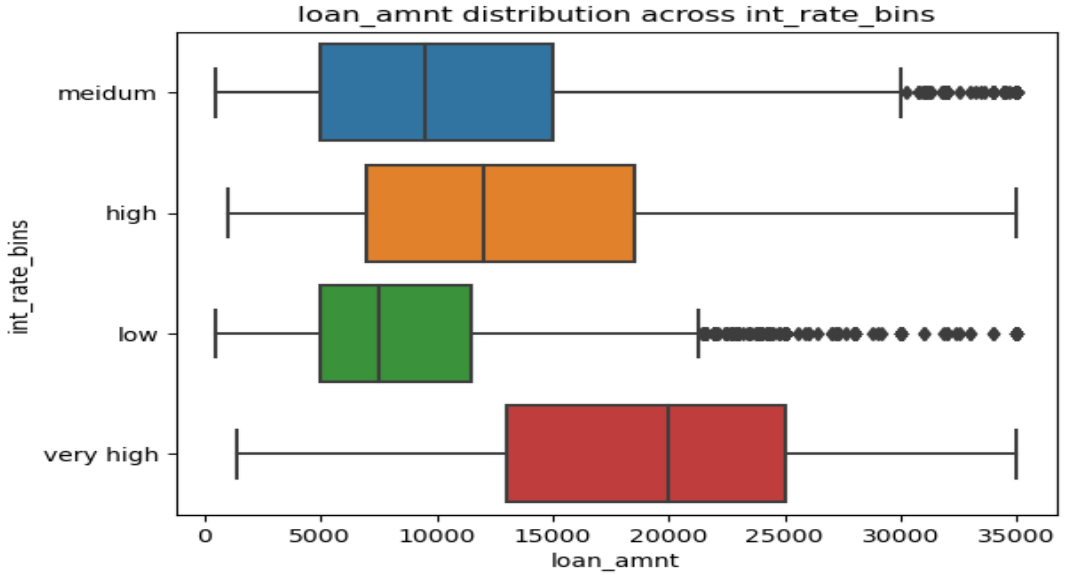
1. 60 months term loan mostly give for High and Very High loan amount category (15 to 35 K)
2. A grade loan given at 5 to 10K loan amount which also why very less default rate on A grade loans
3. People with 10+ tends to takes more loan with high and very loan amount category. Which also explains why more defaulter. High loan amounts tends to default

Bivariate Analysis (Loan Amount vs Categorical Variables)

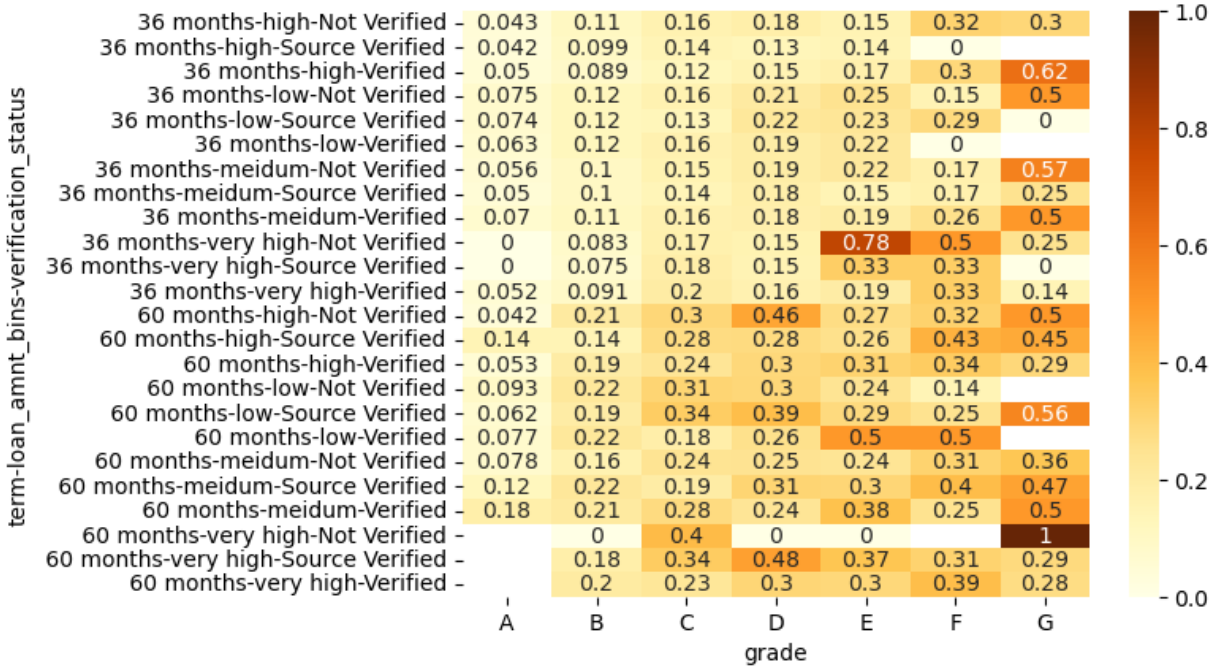


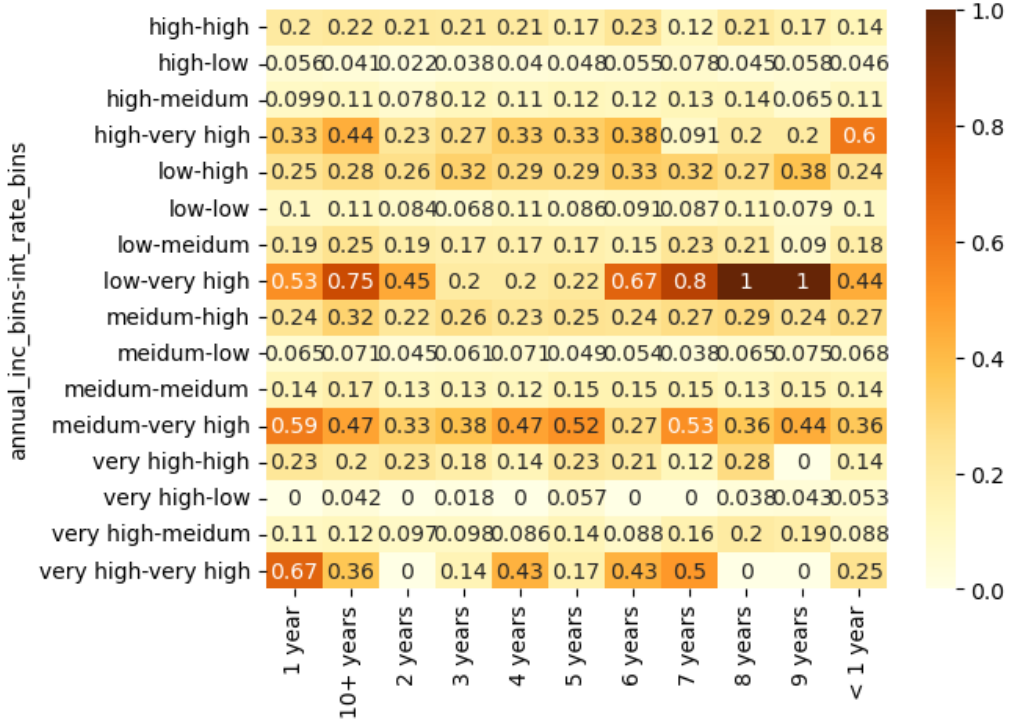


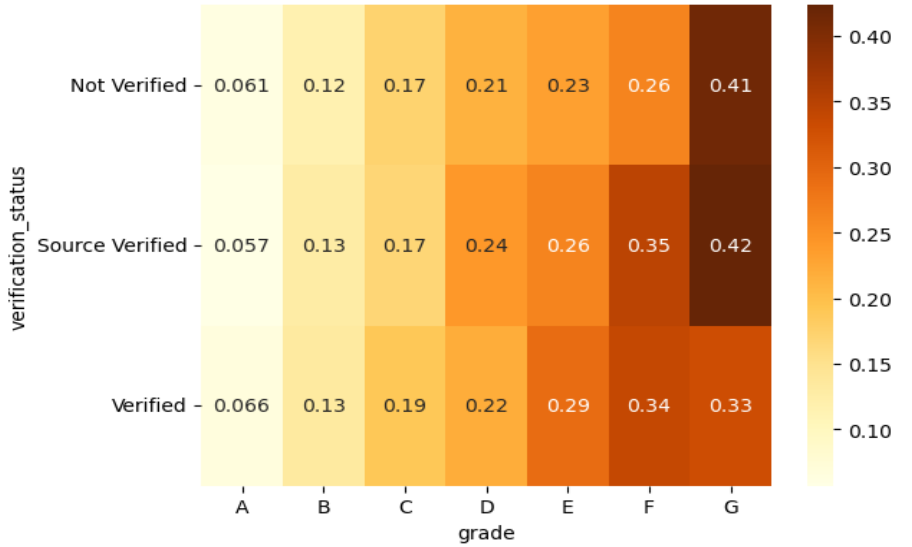
1. High and Very High Loan amounts taken for the purpose small business, debt consolidation, house, credit cards
2. High and Very High loan amount given with higher interest rates



Multivariate Analysis

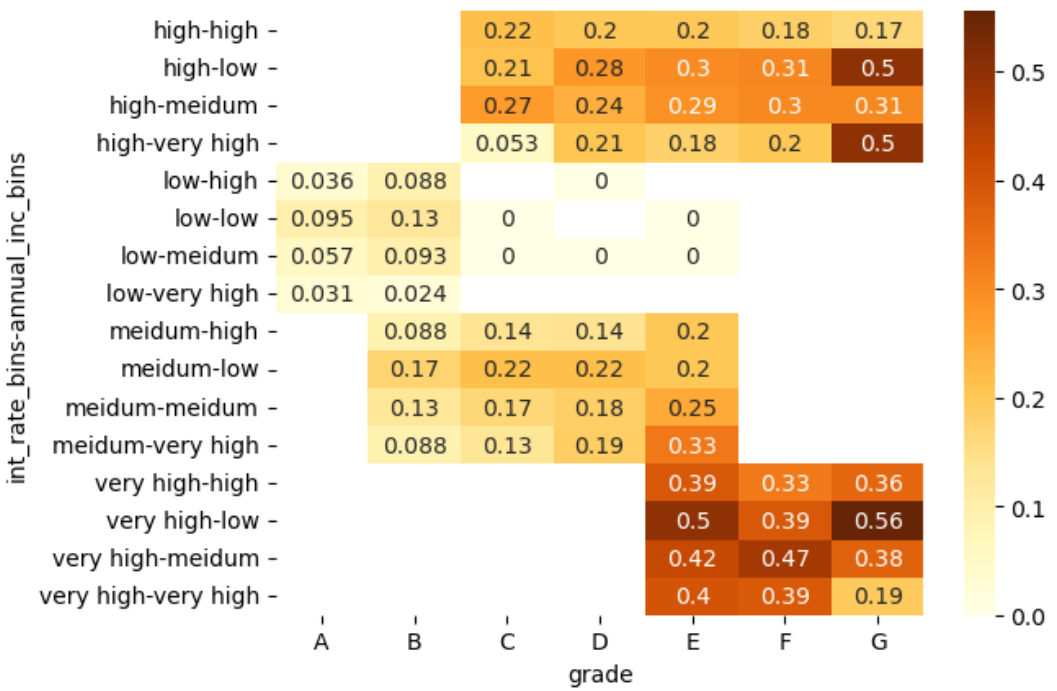


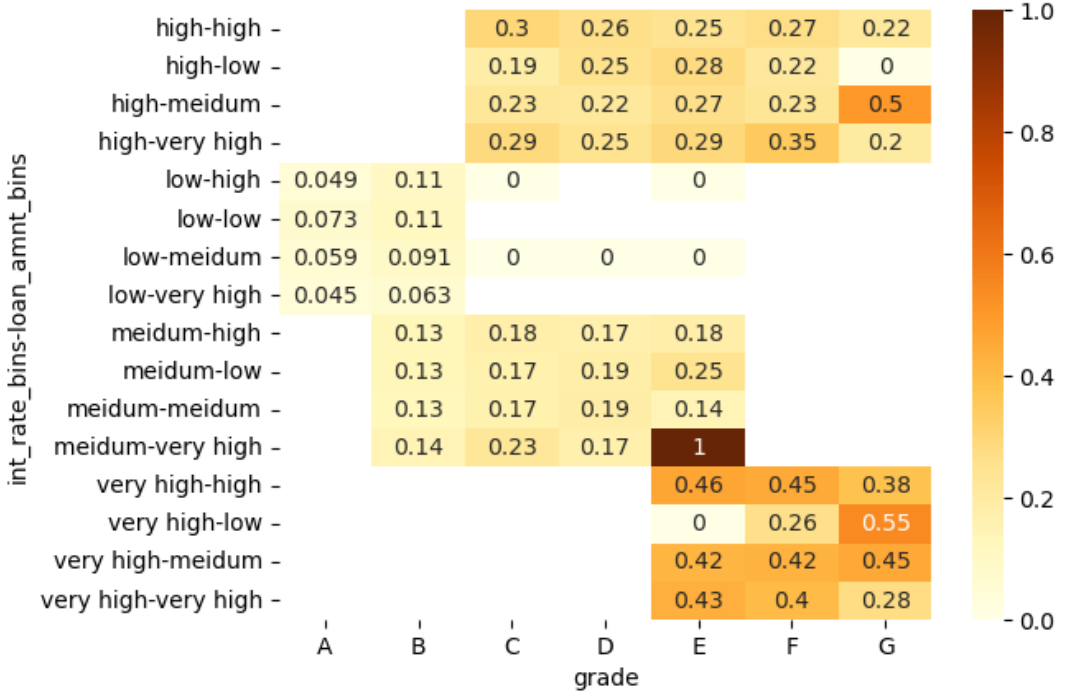




1. 60 months term loans are defaulted most on the lower grade (E,F,G)
2. 30 months term, high loan amt and not verified E grade defaulted more
3. 10+ years experience taking very high interest rate loans tends to default
4. low annual income and very high interest rate tends when experience is <=2(40 to 55%) or >=6(70 to 80%)
5. people with 1 year experience taking vey high interest loan tends to default across all income range
6. E, F, G loan grades not verified having more probability to default

Multivariate Analysis





1. Very High interest rate are given E, F, G category tendency to default irrespective annual_inc and loan_amnt category

Conclusion

Following variables identified as influencing loan default

Column Name	Column label	Type	Note
grade	grade	Categorical	
sub_grade	Sub grade	Categorical	
Int_rate	Interest Rate	Continuous	
loan_amnt	Loan amount	Continuous	funded_amnt_bins, funded_amnt_inv_bins follows same pattern
purpose	Purpose	Categorical	
term	term	Categorical	
annual_inc	Annual income	Continuous	
installment	Installment Amount	Continuous	
verification_status	Verification status	Categorical	
home_ownership	Home ownership	Categorical	
dti	Debt to income ratio	Continuous	
emp_length	Employment length	Categorical	

Trends and Patterns:

1. 14.77 % of the loan application in the data set are defaulted
2. Most loans are taken for the purpose of debt_consolidation, credit_card, home_improvement
3. Most Loans taken at the end of the year at Dec, Nov, Oct
4. Number of loans taken steadily increasing over the years
5. Most loans taken at states California, New York, Florida, Texas
6. More loan taken at 5000 to 15000 loan amount range
7. More loan having interest rate between 10 to 15 %
8. More people taking loans at the income range of 40 to 80 K
9. Loans are concentrated with 100 to 400 range installment amount
10. Most Loans having dti 10 to 20
11. People with having income range 50 to 100K tends to take more G grade loans
12. People with having income range 40 to 90K tends to take more loans for home improvement and small business
13. 60 months term loan mostly given for High and Very High loan amount category (15 to 35 K)
14. A grade loan given at 5 to 10K loan amount which also why very less default rate on A grade loans
15. People with 10+ tends to takes more loan with high and very loan amount category. Which also explains why more defaulter. High loan amounts tends to default
16. High and Very High Loan amounts taken for the purpose small business, debt consolidation, house, credit cards
17. High and Very High loan amount given with higher interest rates

Correlations

1. loan_amt,funded_amt,funded_amnt_inv, Installment are highly positive corelated with each other (>0.9)
2. Annual_income having decent positive correlation with loan_amt,funded_amt,funded_amnt_inv, Installment (~ 0.4)
3. int_rate having decent positive correlation with loan_amt,funded_amt,funded_amnt_inv, Installment (~ 0.3)

Default Patterns

1. Grade - Default rate increase from loan grade A to G, risk is increasing from A to G (6 to 36%)
2. int_rate - Default rate increase from low to very High (6 to 40%)
3. annual_inc - Default rate decrease from low to very high (18 to 11%)
4. installment - Default rate increase from low to very high (14 to 18%)
5. Issue Date – 2007 had default rate of 18% compared to low on 2009(12%). 2007 might be impacted by economic crisis
6. Employment Length - Higher default rate occurs when emp_length is 10+ Years. Only 2% diff between lowest to highest default rate
7. dti - Default rate increase from low to very high (12 to 16%)
8 . 60 months term loans are defaulted most on the lower grade (E,F,G)
9. 30 months term, high loan amt and not verified E grade defaulted more
10. E, F, G loan grades not verified having more probability to default

When did it happen most

1. Grade - G(36%), F(32%), E(26%)
2. Sub Grade – F5(52%),G3(46%),G5(40%),G2(38%)
3. Interest rate – Very High –(39%) , High (25%)
4. Purpose – small business(27%), renewable_energy(19%), educational(17%)
5. Term - 60 months ~ 26%
6. Address State - NE, NV,ID, SD, AK, FL ,HI
7. loan_amnt, funded_amnt_bins , funded_amnt_inv_bins - Very High loan Amounts ~22%
8. Home Ownership - Other ~18%
9. installment - very High ~18 %
10. dti - very high ~18%

When did it happen least

1. Grade - A grade 6%
2. Interest rate - low - 6%
3. Purpose – wedding, credit card ,car - 10%
4. Term - 36 months ~ 11%
5. annual income - very high - 11%
6. dti - low 12%