Background

A company provides various types of loans to customers

Company receives a loan application with customer details

Company needs to make one of the following decision based on the data

a. To approve the loan application

b. To reject the loan application

Decision is based on likelihood of whether

- Customer is likely to repay the loan
- Customer is likely to default on loan

Risks involved

- Approving application of customer who is likely to default → Financial loss
- Rejecting application of customer who is likely to repay → Business opportunity loss

Problem Statement

Company wants to understand the likelihood of customer repaying or defaulting the loan and the driving factors.

-Data for past loan applicants is available with company

- The intention of this study is using EDA to identify patterns seen in data
- Using the patterns, provide inferences
- Suggest actionable points to company which are useful to minimize both types of risk (Business Opportunity Loss & Financial Loss)
- This can be achieved by using the analytical insights in the process of loan approval

Analysis Approach

Exploratory Data Analysis of past applicants' data

Steps:

- Observation of data
- Data cleaning
- Treatment of missing values & redundancy handling
- Univariate analysis to see existing trends of variable factors
- Bivariate analysis to specifically find effect of multiple variables factor on outcomes of company's interest (Loan repaid or loan defaulted)

Primary Data Observations

- Initial data file received in .csv format
- This is internally collected data by company (Data for customers with loan approved)
- 39717 records seen with 111 columns
- Data available for 3 types of customers:
  - Current customer (Currently repaying loan)
  - Customers who repaid loan amount
  - Customers who defaulted

Primary Data Observations (Continued..)

Notes on limitations of data:

- Data for customers with loan approval rejected is not available with company,.
- The relationship with these customers did not continue after loan rejection, hence this data is not available for analysis
- Except for a “grade” assigned by company at initial assessment, for the purpose of risk based interest rest increment, any External data (credit score, other ongoing loans with other companies etc.) is not included in given data)
- Data related to age is not provided
- Multiple empty values observed which need to be dropped by applying certain threshold

Data Cleaning & Basic Operations

- Data for current customers is dropped for this analysis, as the outcome related to default is not yet final
- Dropped the columns with missing values (Threshold defined at >85% missing values)
- Dropped the columns with same values in all rows or exclusively unique value in each row, as these are not of significance
- Remaining useful data has 38,577 rows & 39 columns
- Month is extracted from Issue date.
- Interest rate is provided as string data type. Converted to numeric.
- Bins are created for numeric variables to view how variables under each bucket is related to charged off/fully paid.
- Outliers are not excluded in this study
- Imputation is not done for missing values in order to avoid distortion. Dropped with threshold instead

Columns of Interest

Columns identified as significant for this analysis are

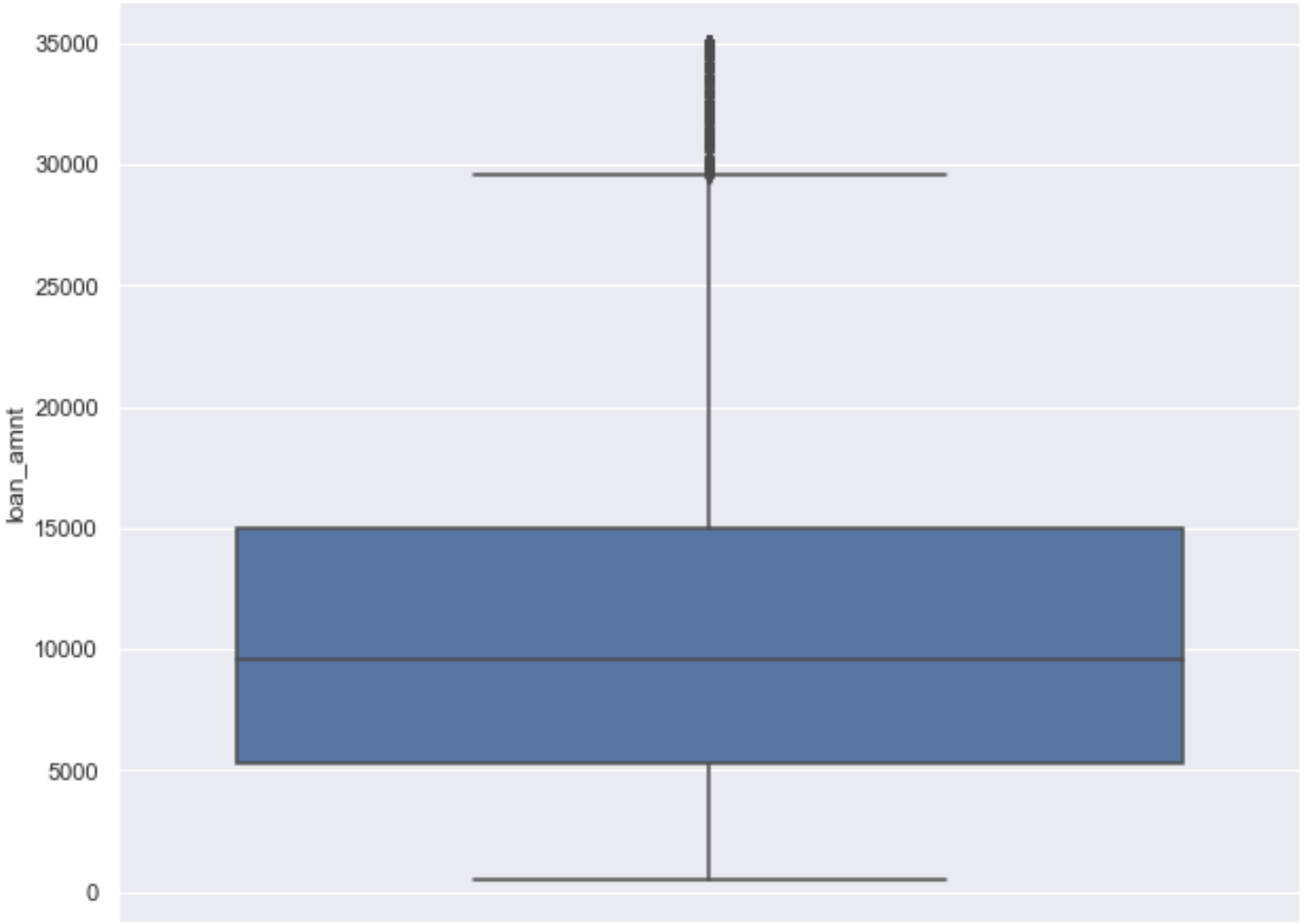
- loan_amnt (Loan amount as applied by borrower)
- funded_amnt_inv (Amount actually approved)
- term (Term of loan in months)
- int_rate (Interest rate in %)
- grade (Risk grade assigned by the company)
- sub_grade (Risk sub grade assigned by company)
- annual_inc (Annual income)

- purpose (Purpose of loan)
- dti (Debt to Income ratio)
- emp_lenght (Years of employment)
- Loan_Date (Month)
- home_ownership (Home type – rented, own, mortgage etc)
- verification_status (Whether details were verified)
- Installment (Monthly installment)

Summary Stats of Numeric variable

Attributes	count	mean	std	min	25%	50%	75%	max
loan_amnt	38577	11047.0254	7348.44165	500	5300	9600	15000	35000
funded_amnt_inv	38577	10222.4811	7022.72064	0	5000	8733.44	14000	35000
int_rate	38577	11.932219	3.691327	5.42	8.94	11.71	14.38	24.4
installment	38577	322.466318	208.639215	15.69	165.74	277.86	425.55	1305.19
annual_inc	38577	68777.9737	64218.6818	4000	40000	58868	82000	6000000
dti	38577	13.272727	6.673044	0	8.13	13.37	18.56	29.99

Univariate Analysis (Loan Amount Requested)

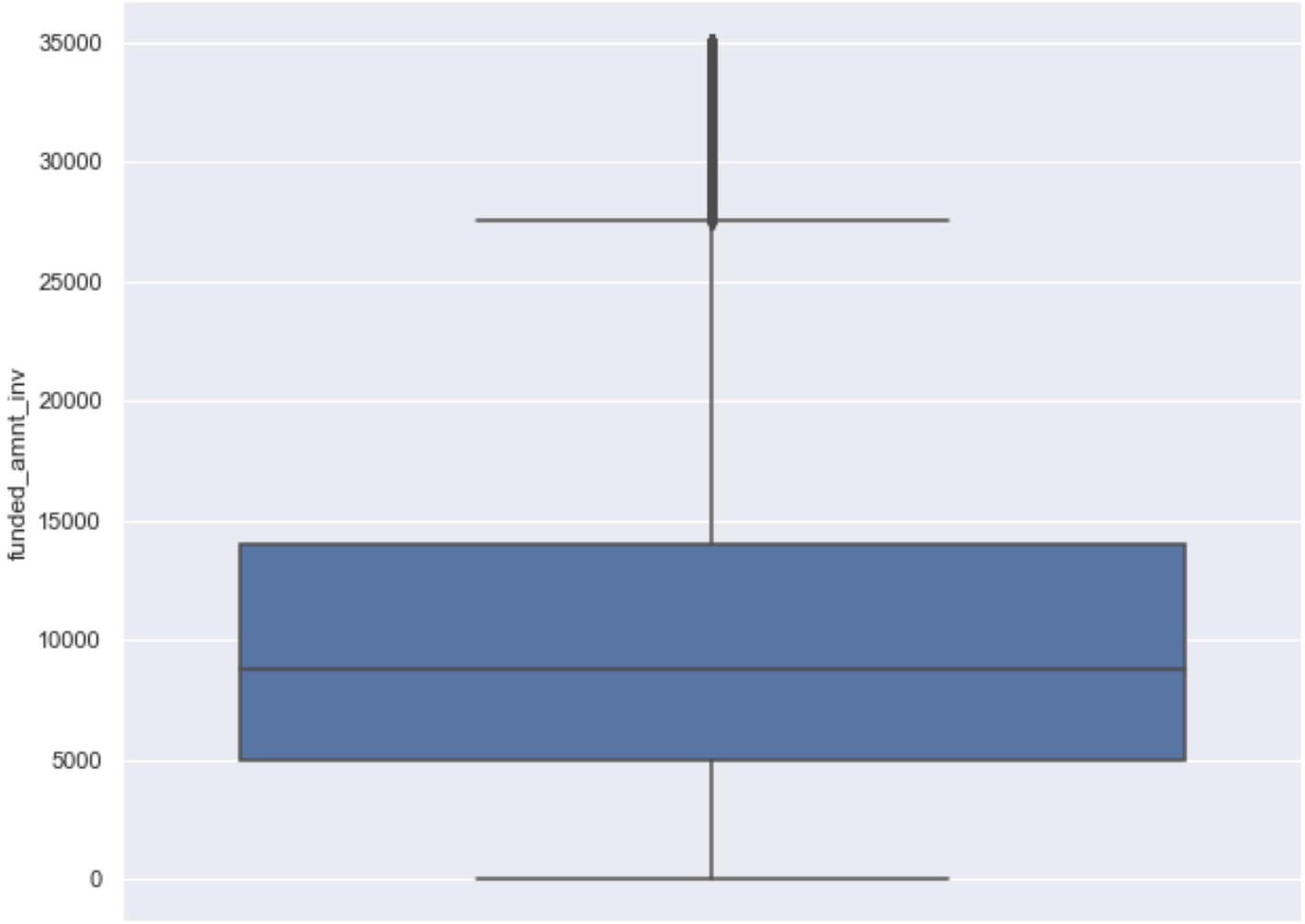


Loan amount applied by customer:

Observation:

-Median loan amount demand is 9600.
-50% of demand amounts are between amount of 5300 to 15000

Univariate Analysis (Loan Amount Funded)

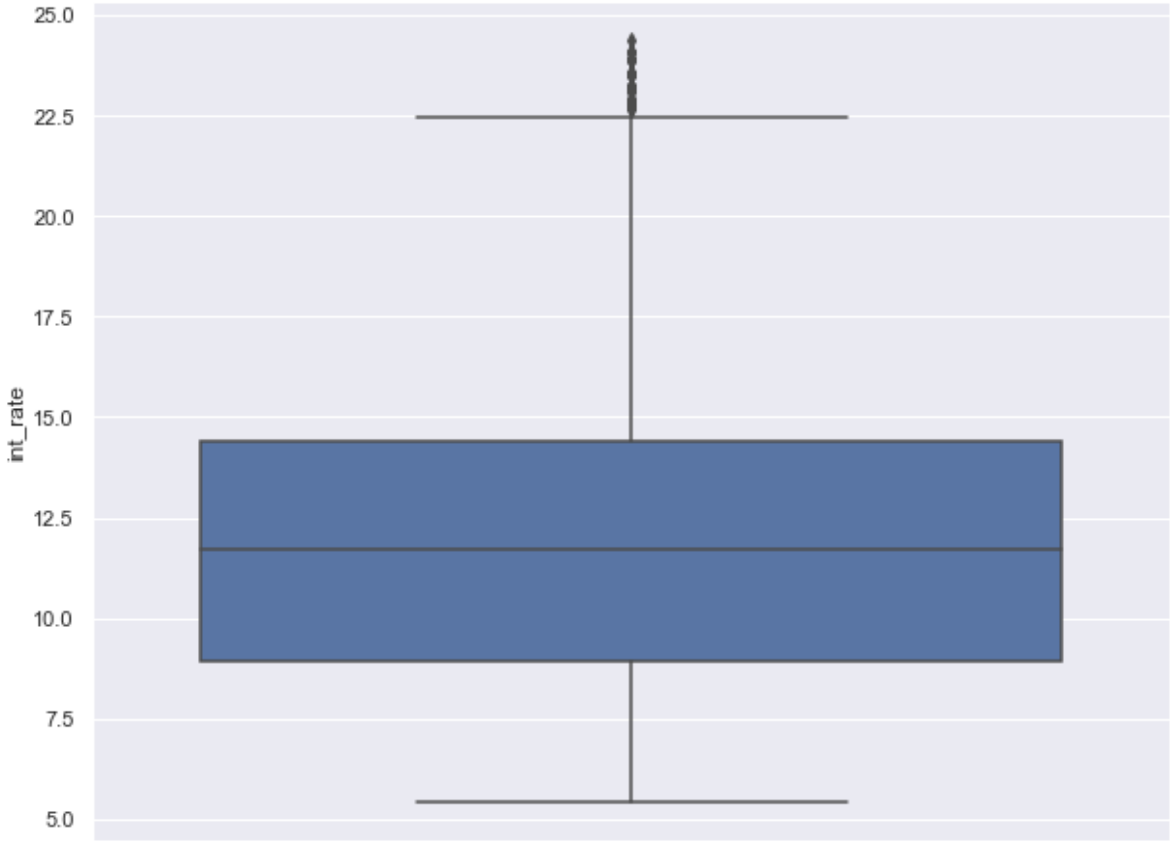


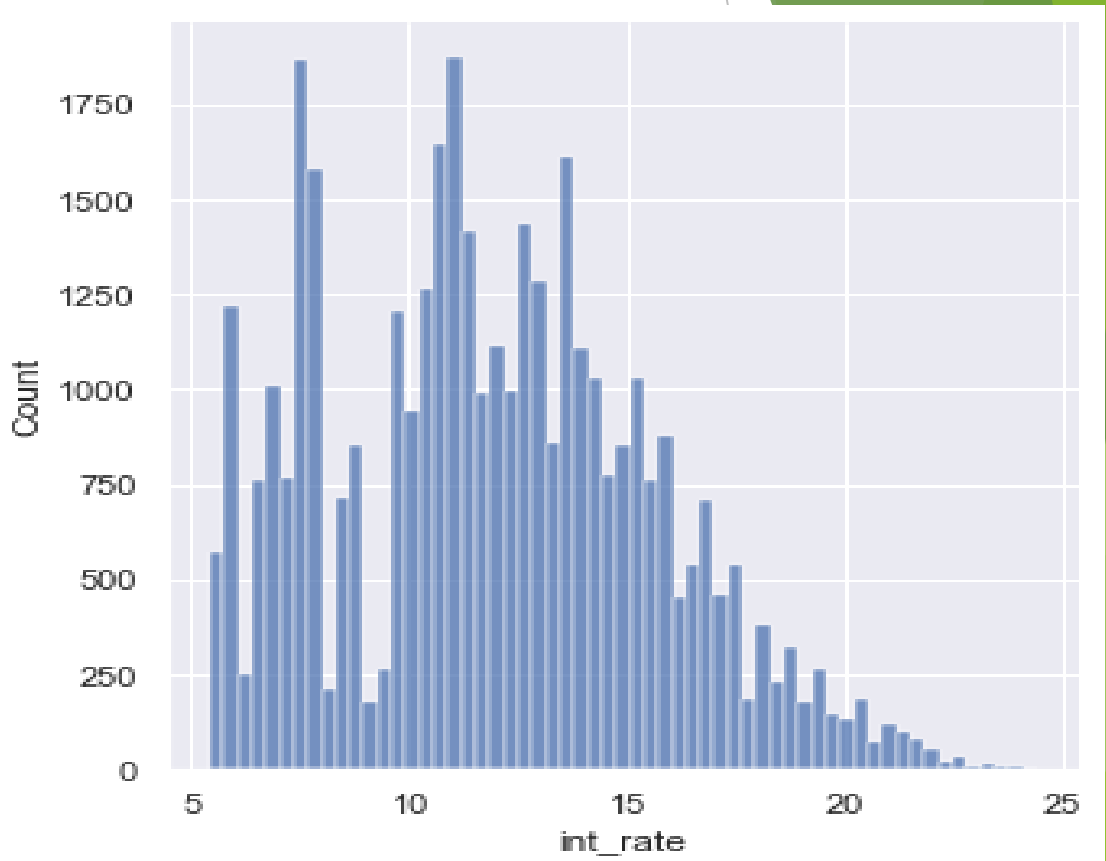
Loan amount finally disbursed:

Observation:

-Median loan amount demand is 8733.
-50% of demand amounts are between amount of 5000 to 14000

Univariate Analysis (Interest Rate)



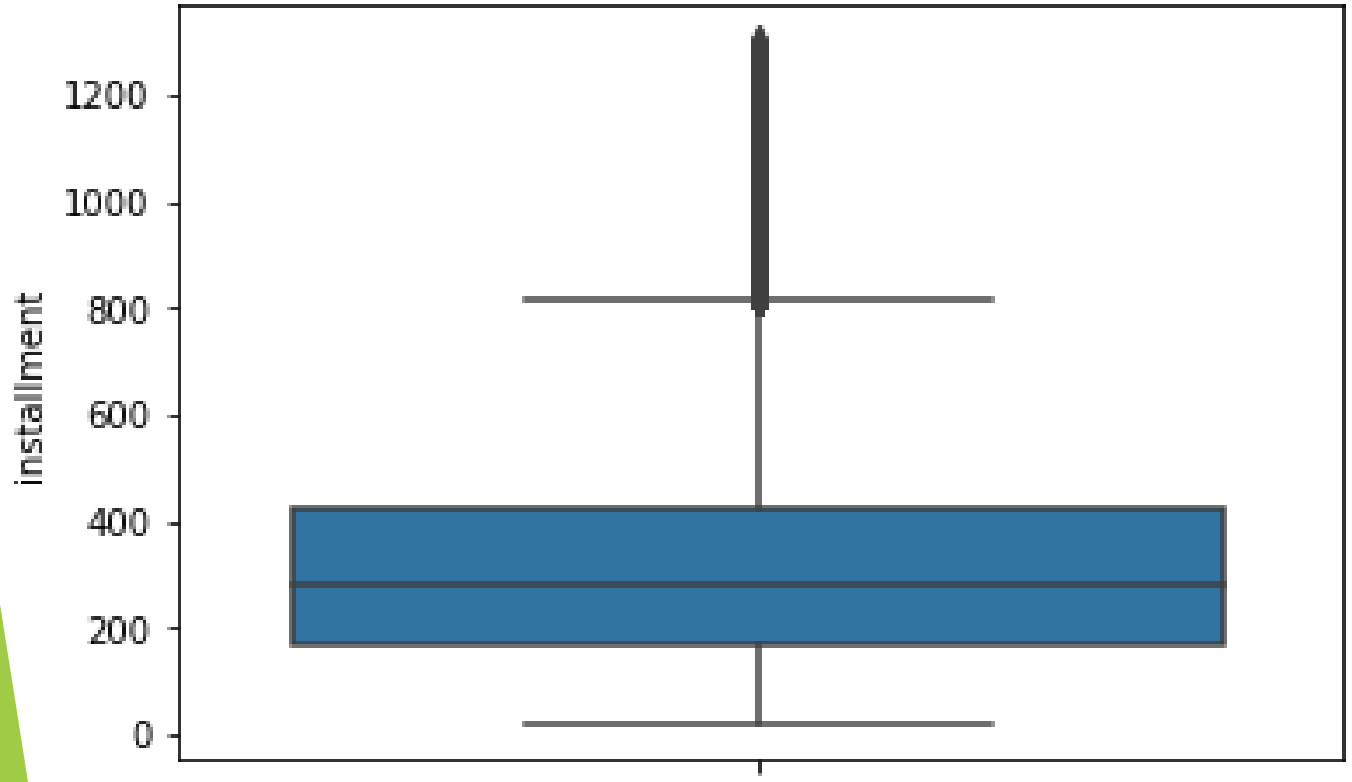


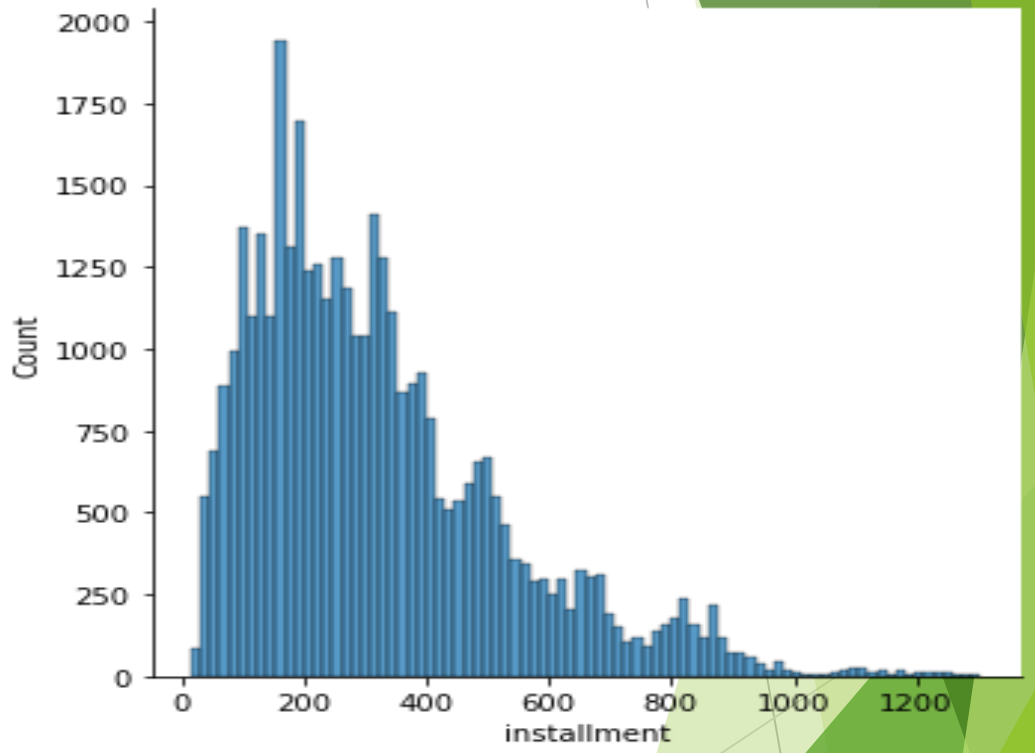
Interest rate at which loan was funded:

Observation:

- Most frequent interest rates are between 10 to 17 percent
- Outliers with interest rates higher than 22.5% are present

Univariate Analysis (Installment Amount)



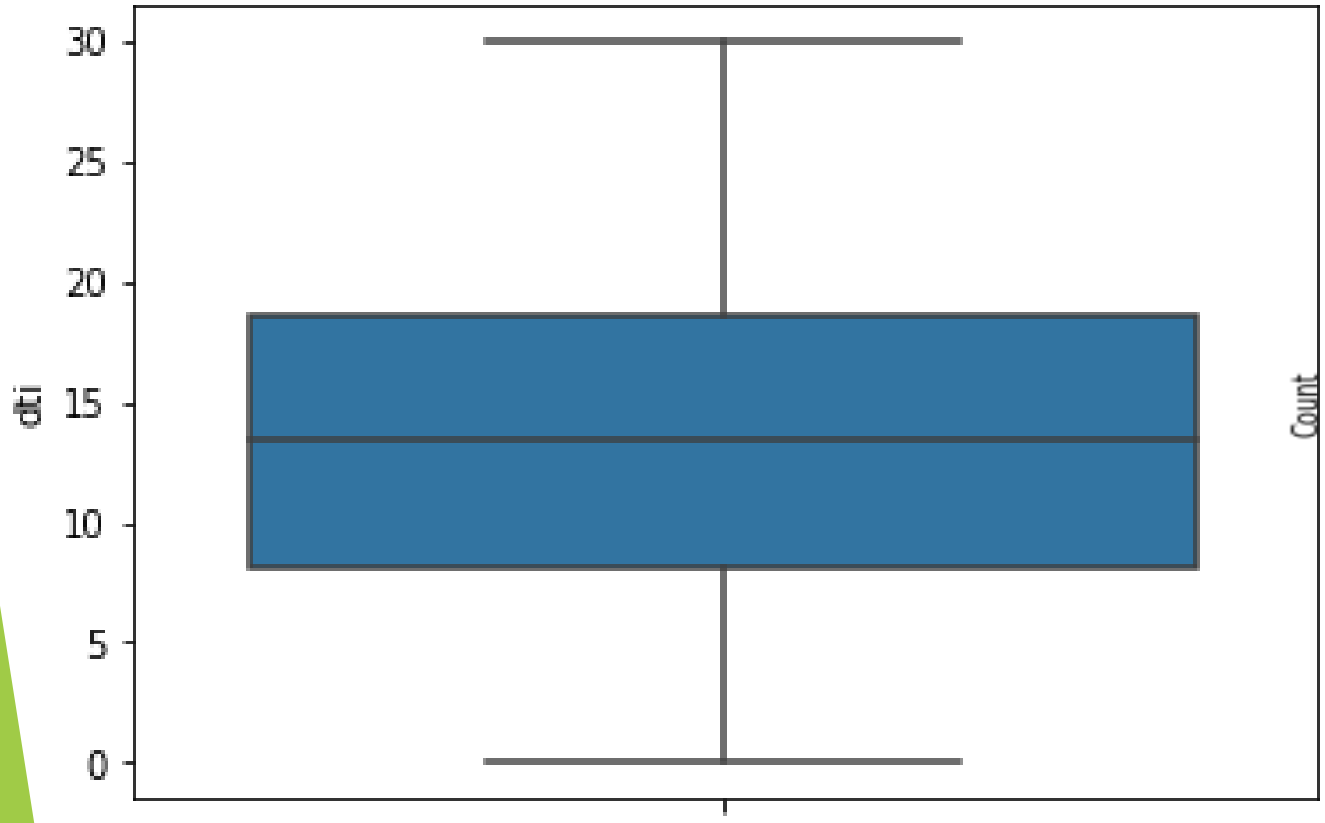


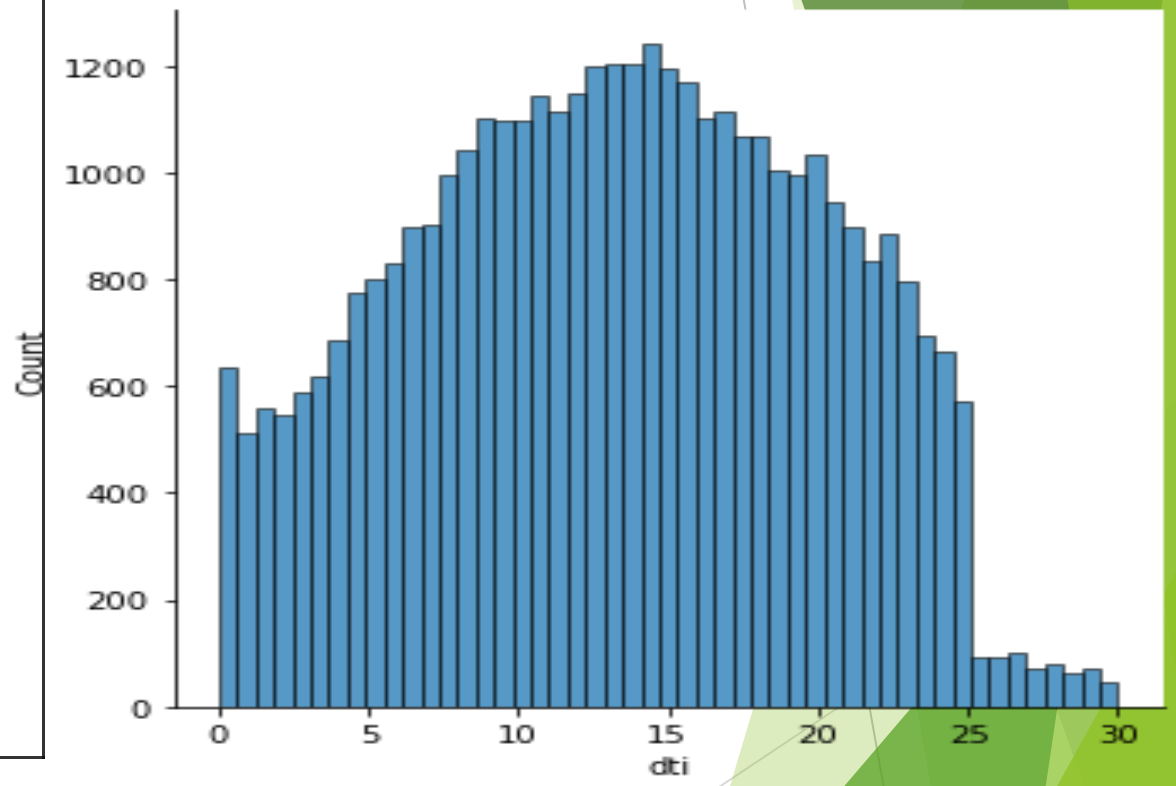
Monthly Installment Amount:

Observation:

- Maximum prevalent installment amount is about 200
- Outliers are present in installment amount

Univariate Analysis (DTI)



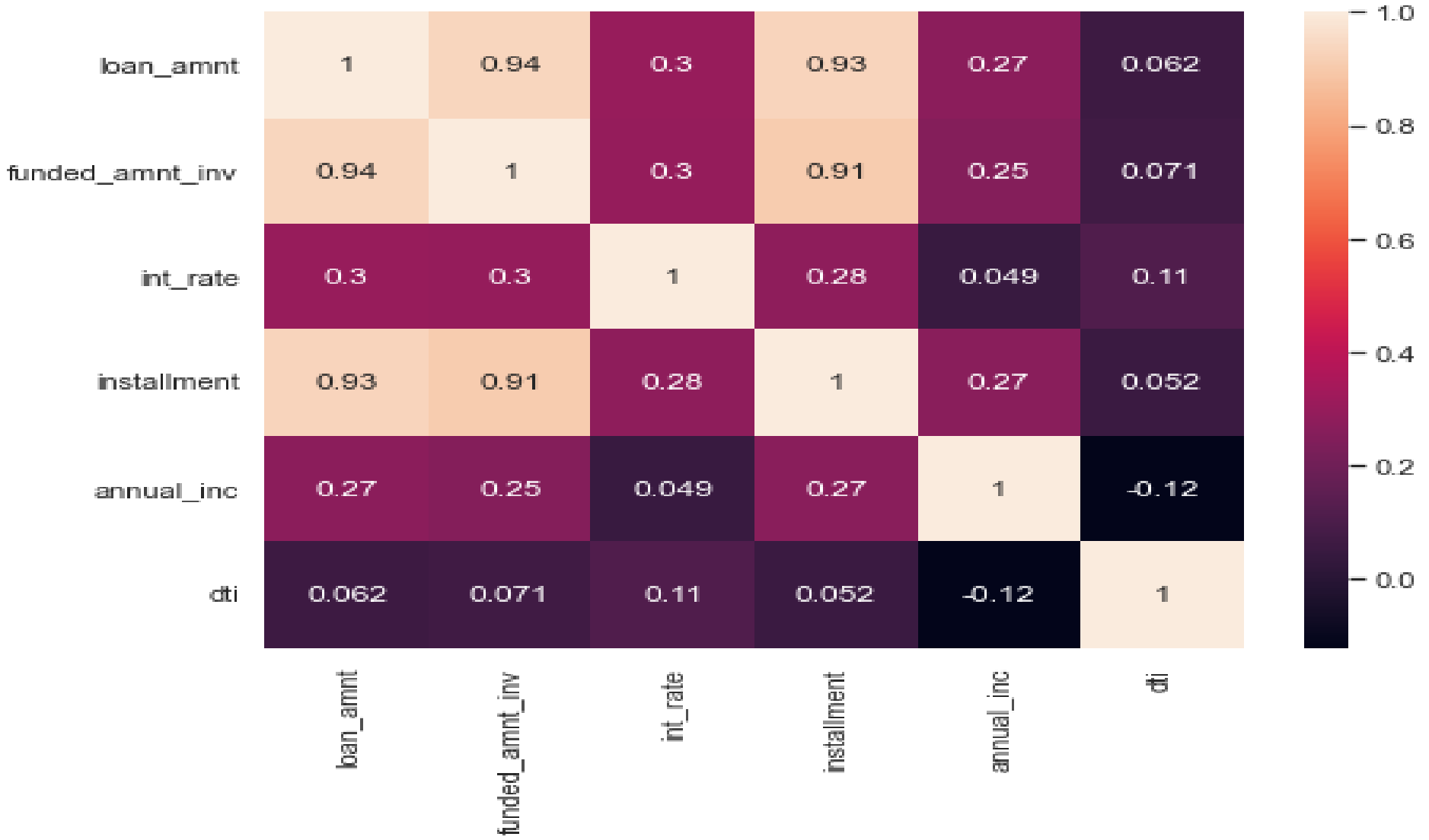


Debt to Income ratio

Observation:

- Most of the dti are between 5 to 25 percent, median is 13.37
-No outliers observed in DTI

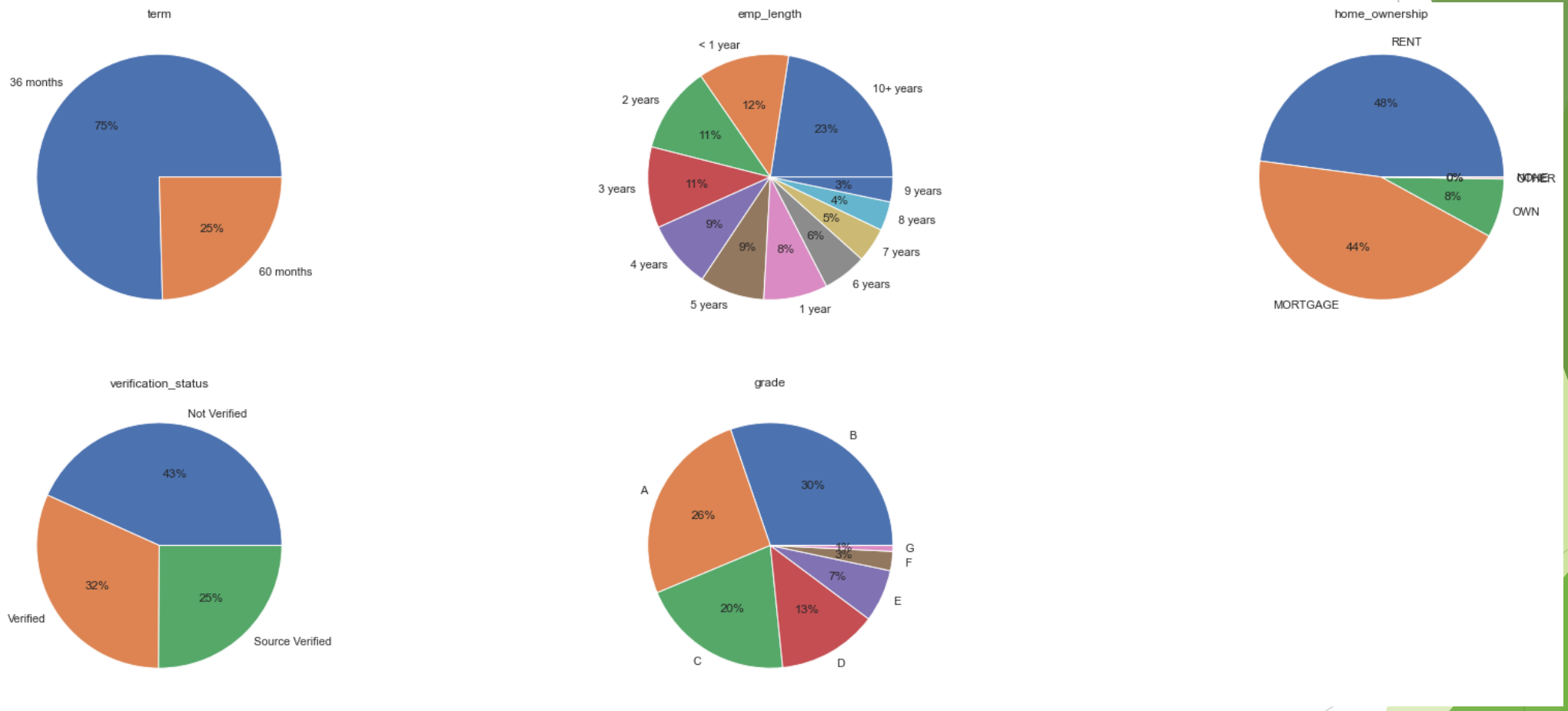
Correlation Matrix (Variables Of Interest)



Funded amount & Installment amount are highly correlated

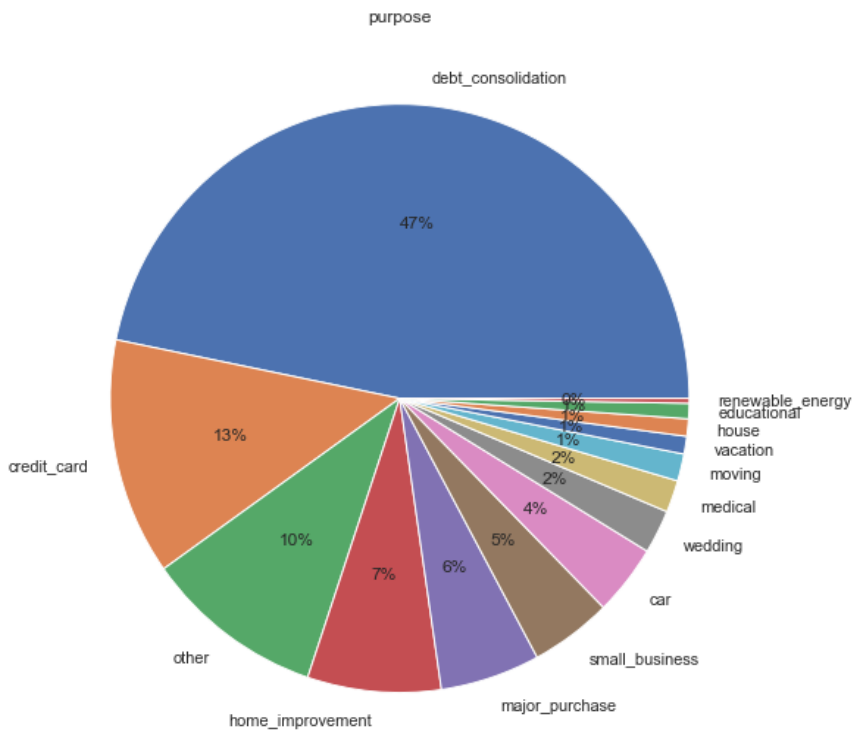
Higher installment to cover higher loan amount in limited period

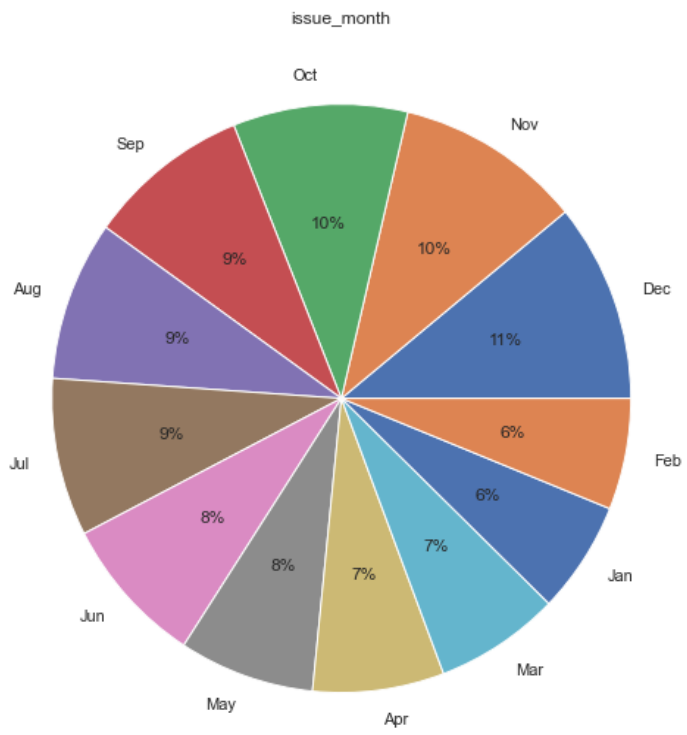
Univariate Analysis (Categorical Variables)

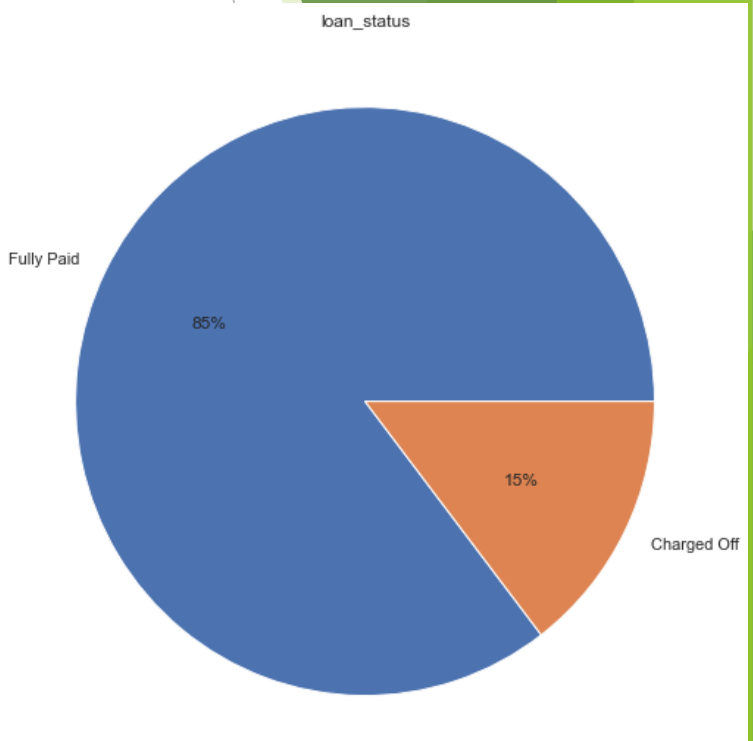


Major populations under each category are: **Term : 36 months**, Employment length :10+ years & <=3 years, Home Ownership: Rental & Mortgage, Verrification: “Not verified” category, Grades: B, followed by A & C

Univariate Analysis (Share of Loan Purpose, Issue Month & Loan Status)

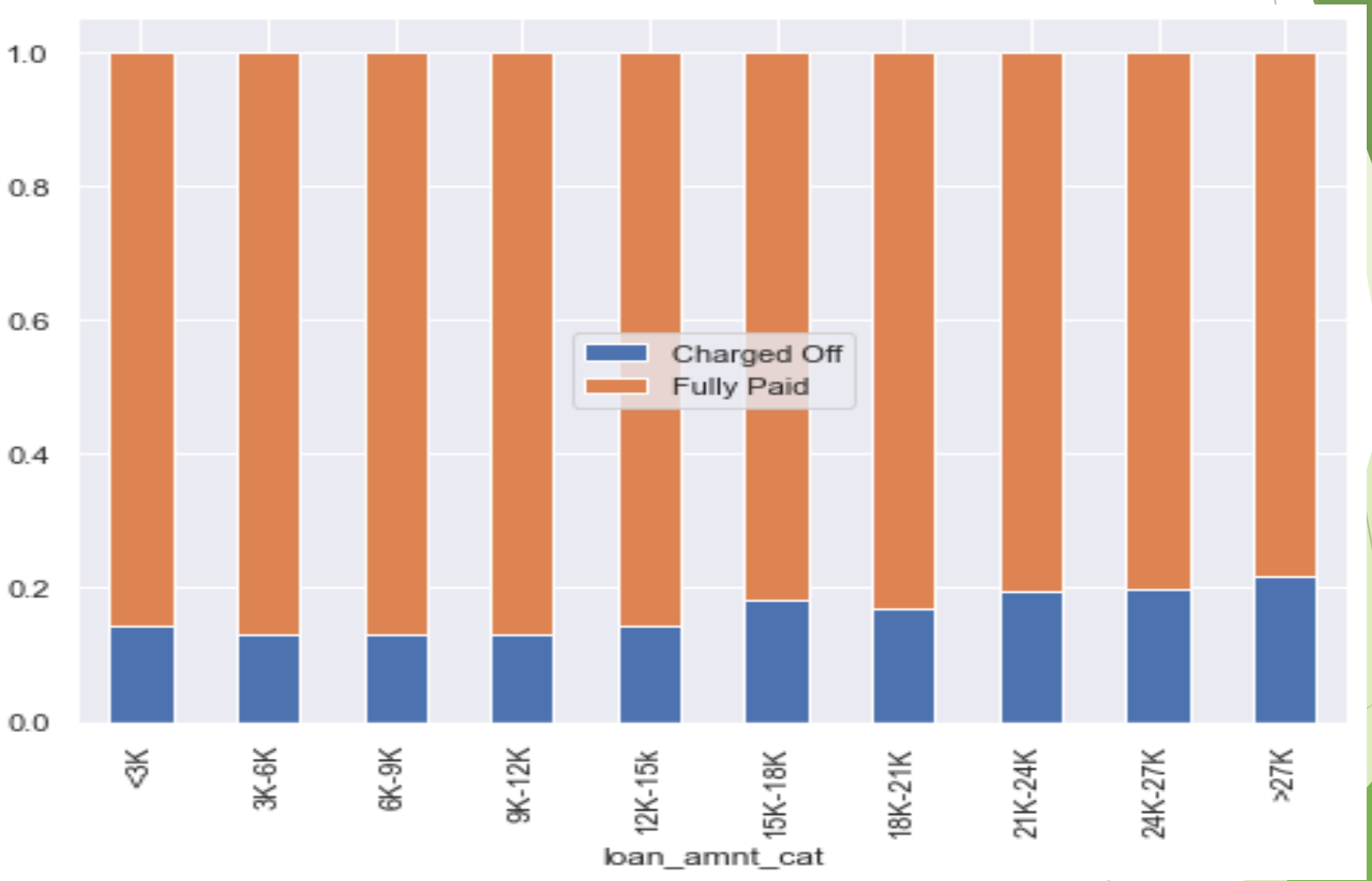






Maximum loans are taken for debt consolidation
85% of approved loans are paid off & 15% are defaulted

Bivariate Analysis (Applied Loan Amount Vs Paid Status: Fully Paid/Charged Off)



We considered loan amount applied instead of funded. Customers demanded amount (need) is more significant for this analysis

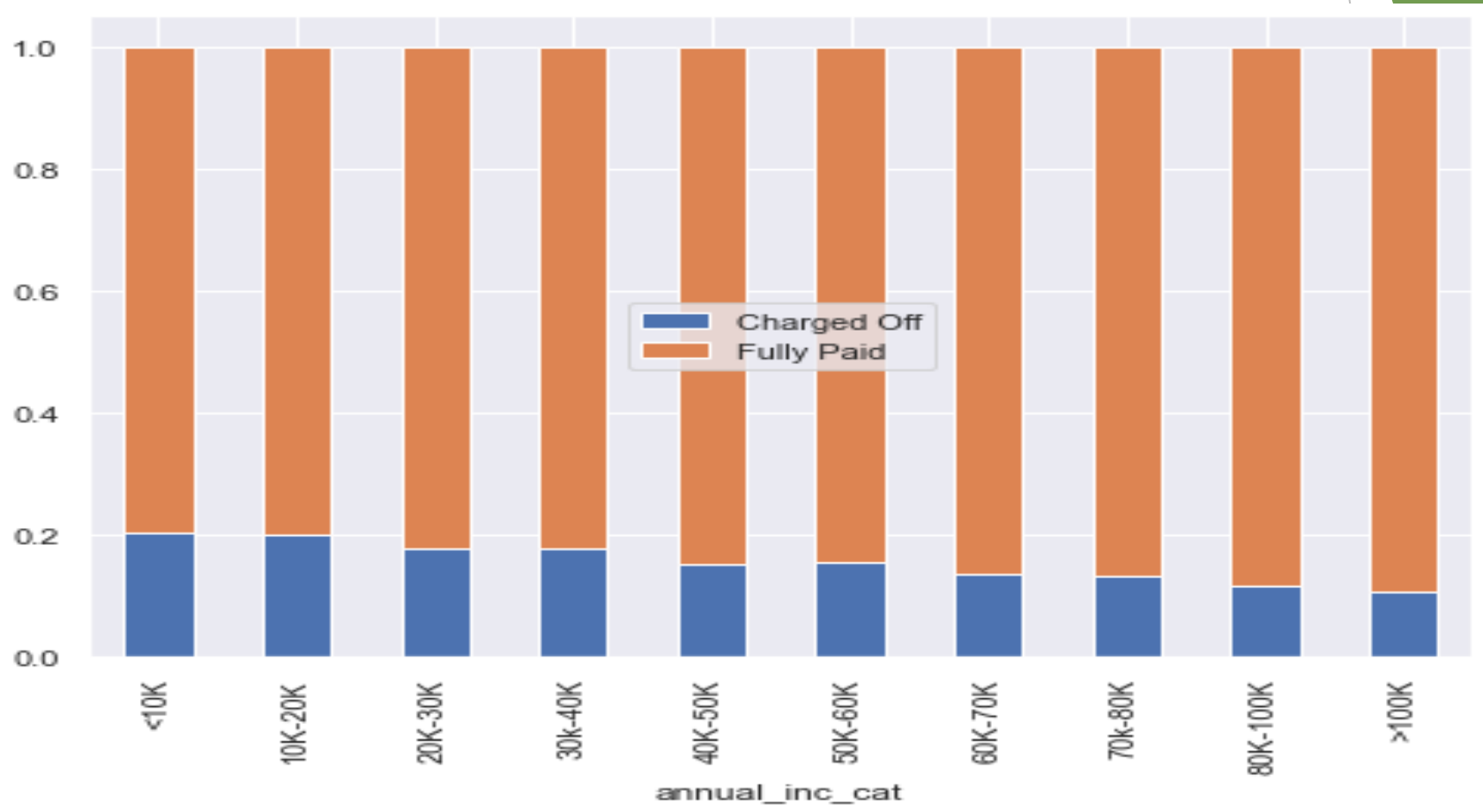
Charge offs are significantly higher for loan amounts higher than 15 K

Trend of higher loan amount causing higher defaults is seen

Bivariate Analysis (Funded Amount Vs Paid Status: Fully Paid/Charged Off)

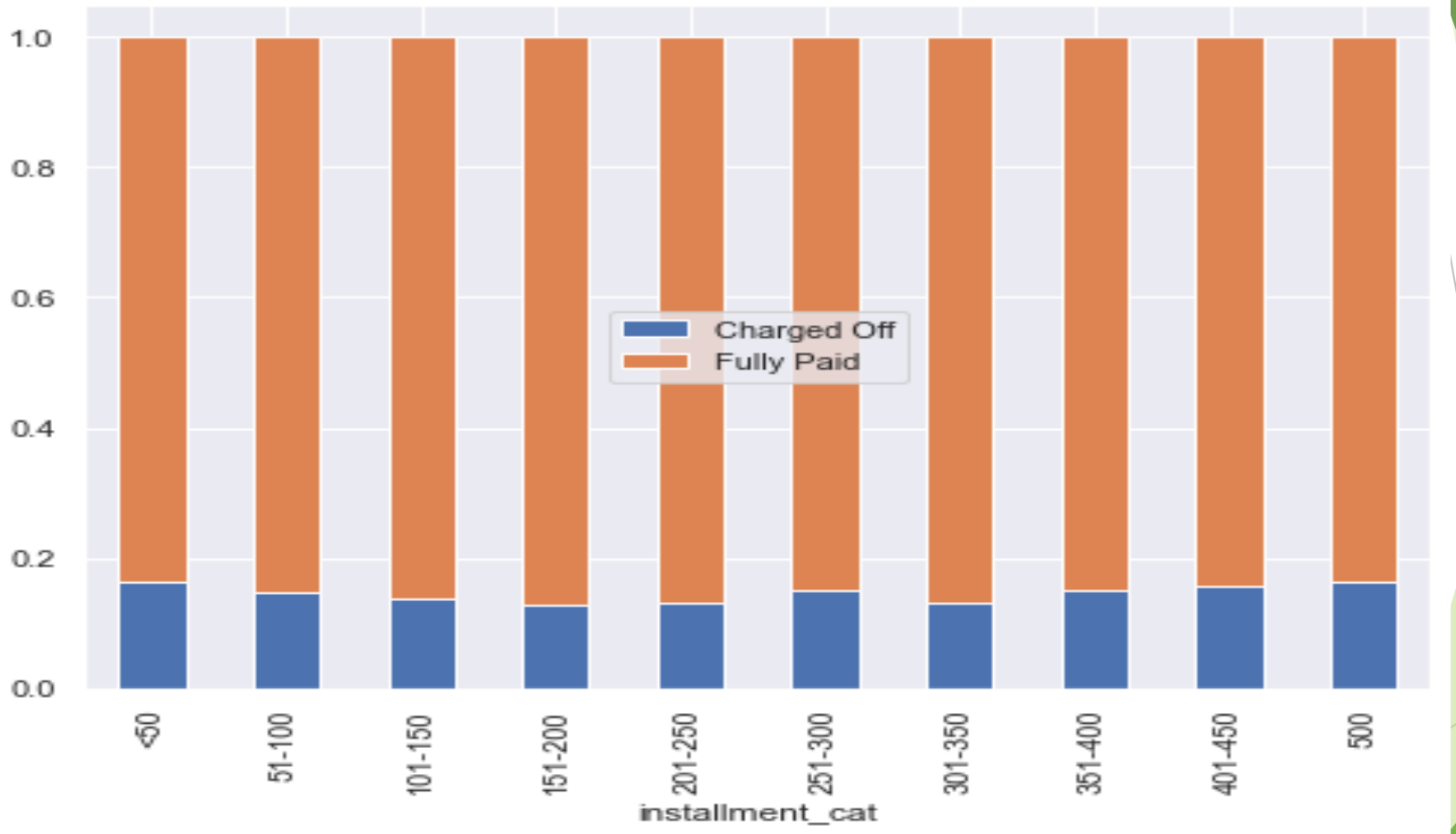
Trend similar to applied loan amount

Bivariate Analysis (Annual Income Vs Paid Status: Fully Paid/Charged Off)



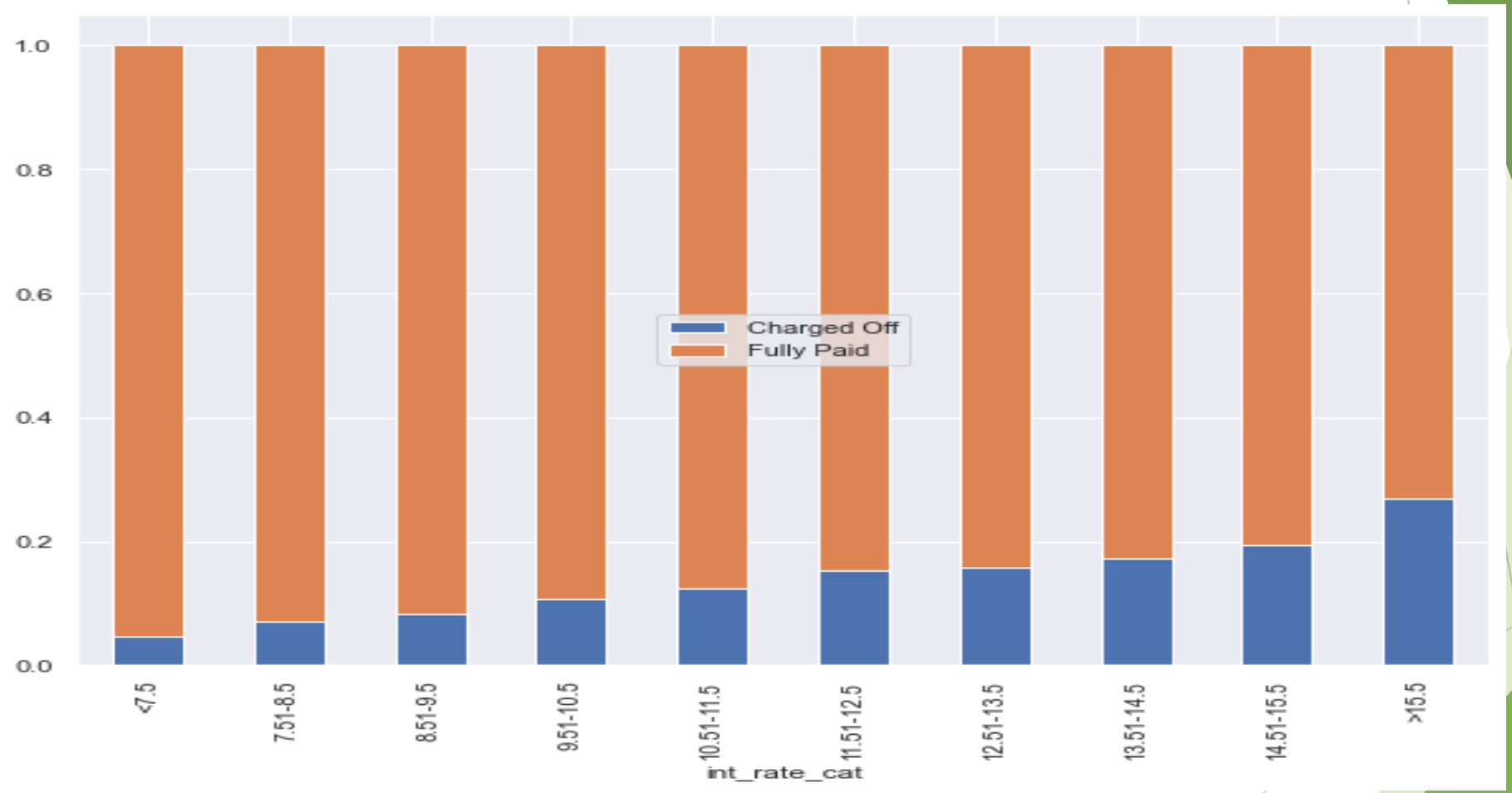
Higher annual income strongly related to lesser defaults

Bivariate Analysis (Installment Amount Vs Paid Status: Fully Paid/Charged Off)



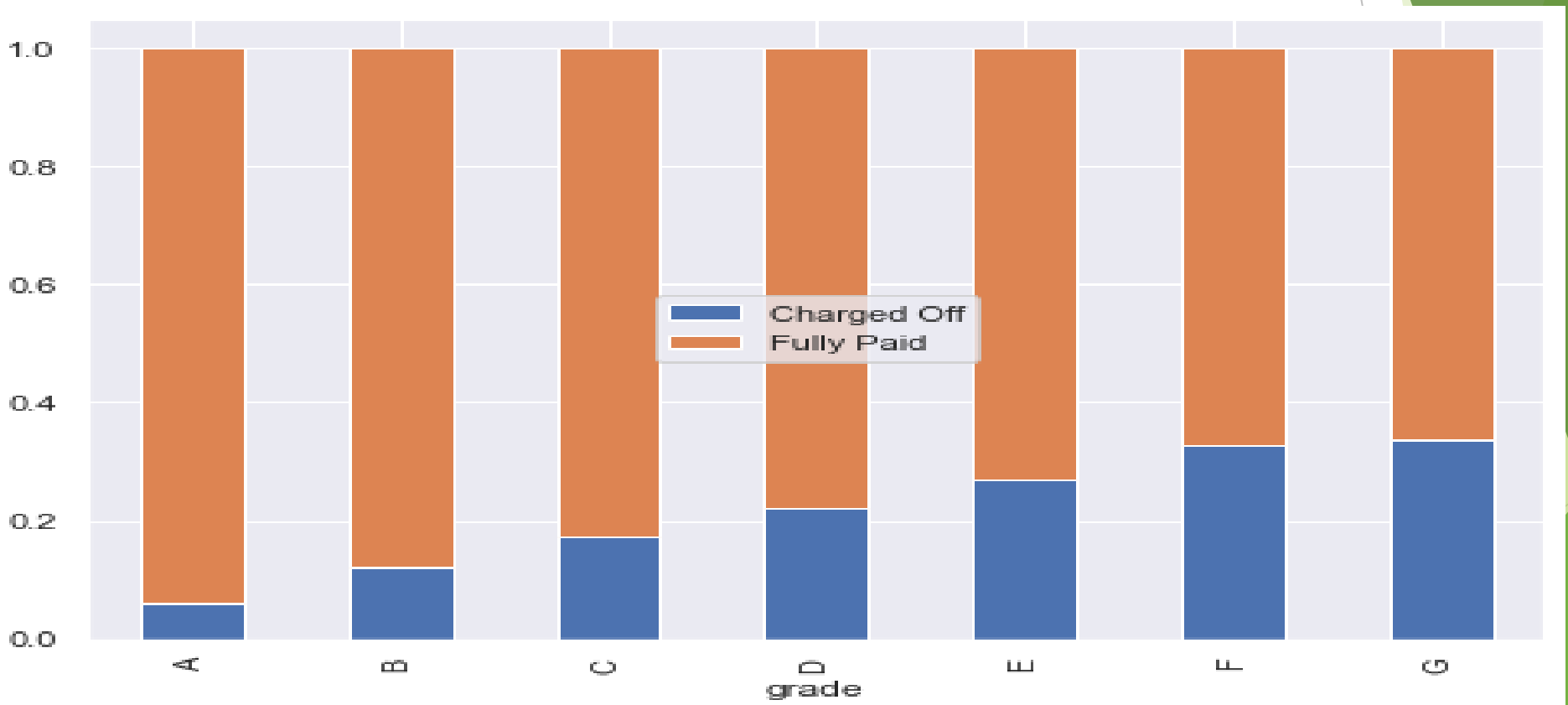
Installment amount has no significant influence on paid status

Bivariate Analysis (Interest Rate Vs Paid Status: Fully Paid/Charged Off)



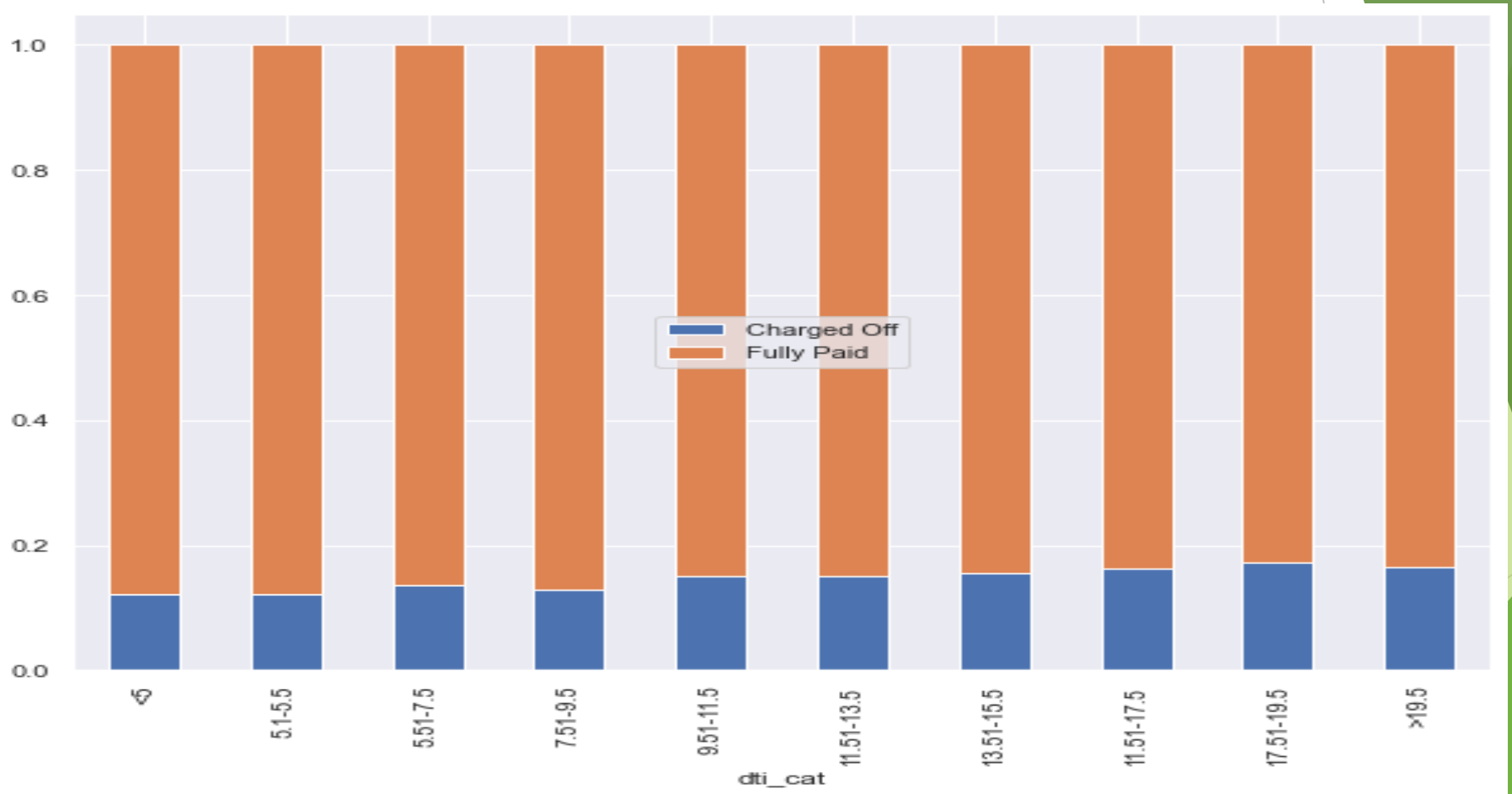
Higher interest rate causes significantly higher number of defaults
Interest rate is a function of risk grade assigned to customer at the time of application
Thus the correlation is also with grade

Bivariate Analysis (LC assigned Grade Vs Paid Status: Fully Paid/Charged Off)



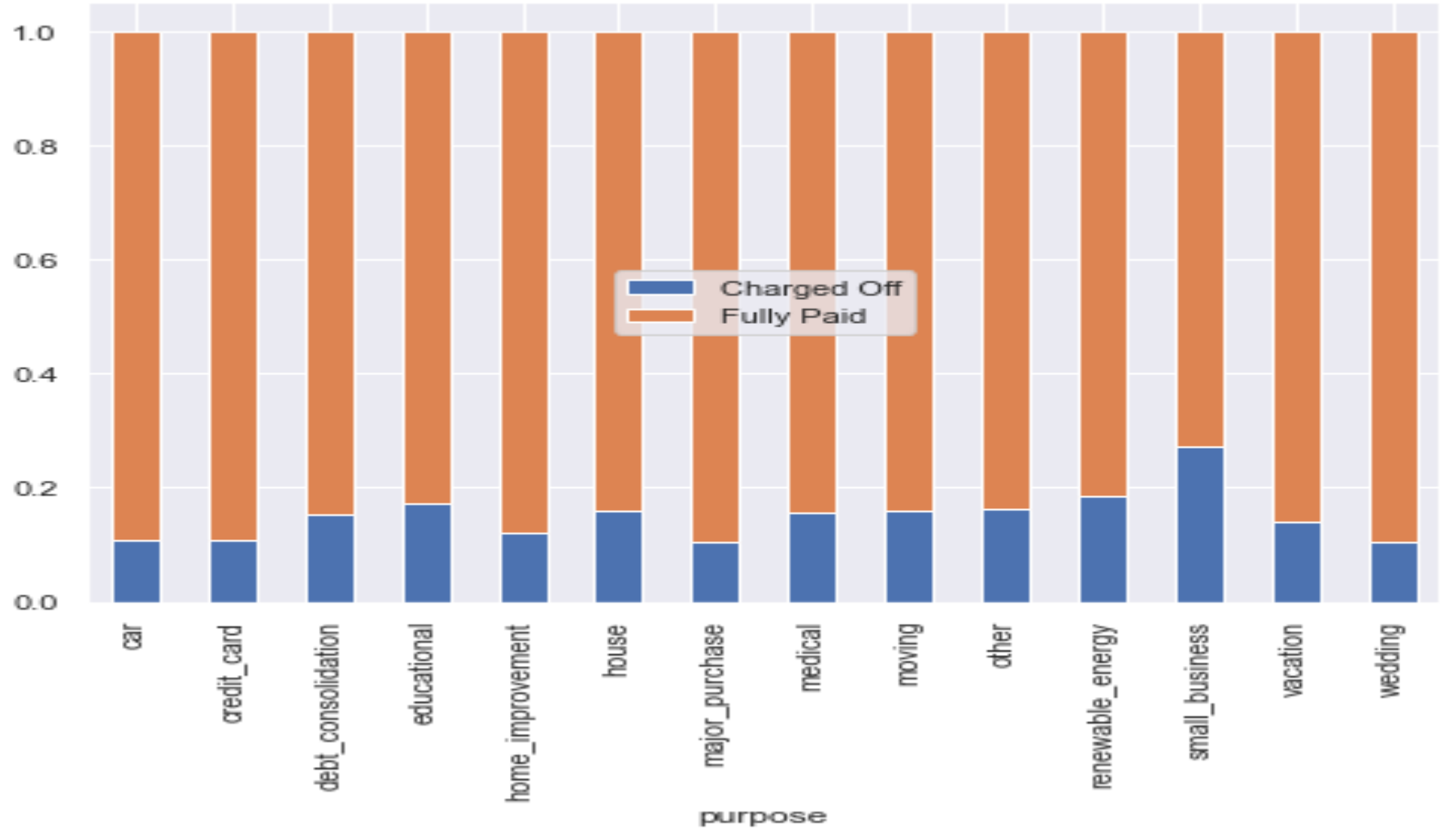
Grades are provided as 'A' with lowest risk & 'G' with highest
Lower the grade, higher the default rate

Bivariate Analysis (DIT Vs Paid Status: Fully Paid/Charged Off)



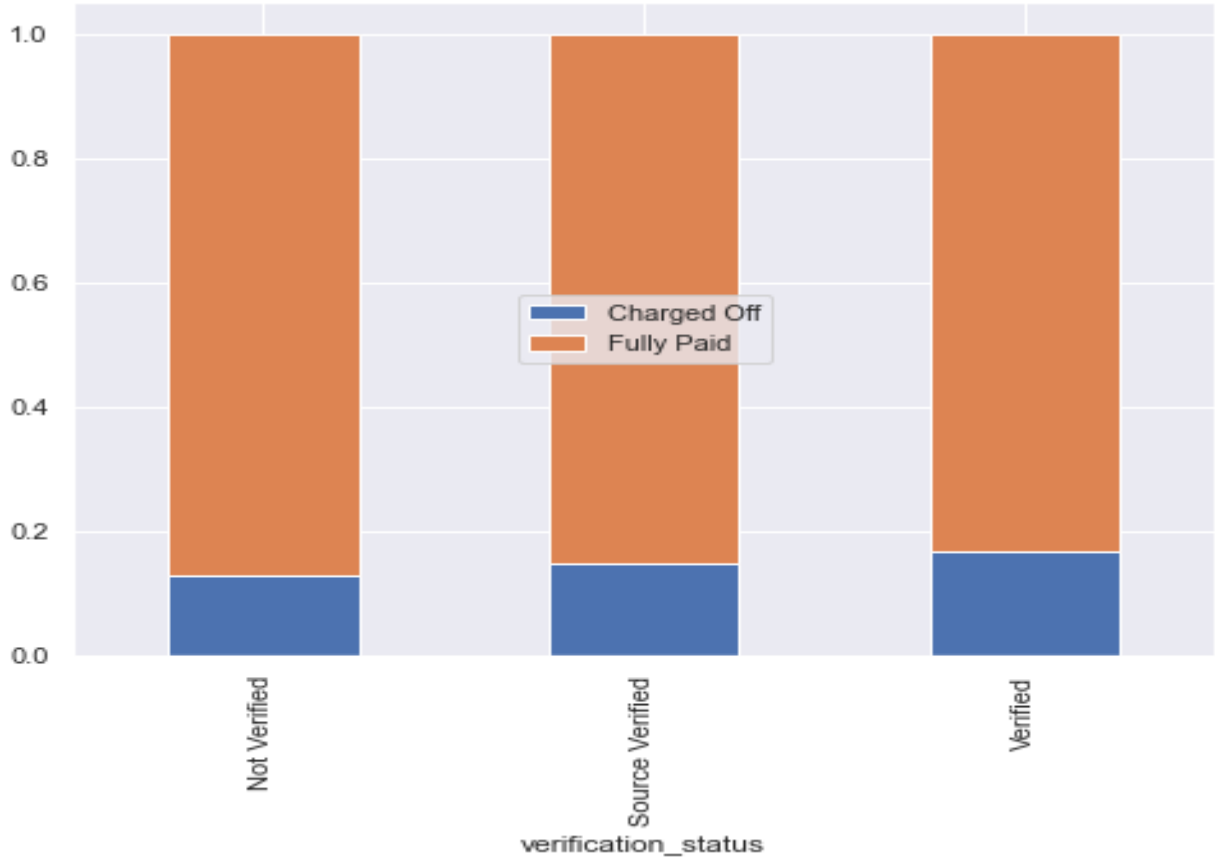
Debt to income ratio >9.5 are more likely to default

Bivariate Analysis (Purpose Vs Paid Status: Fully Paid/Charged Off)



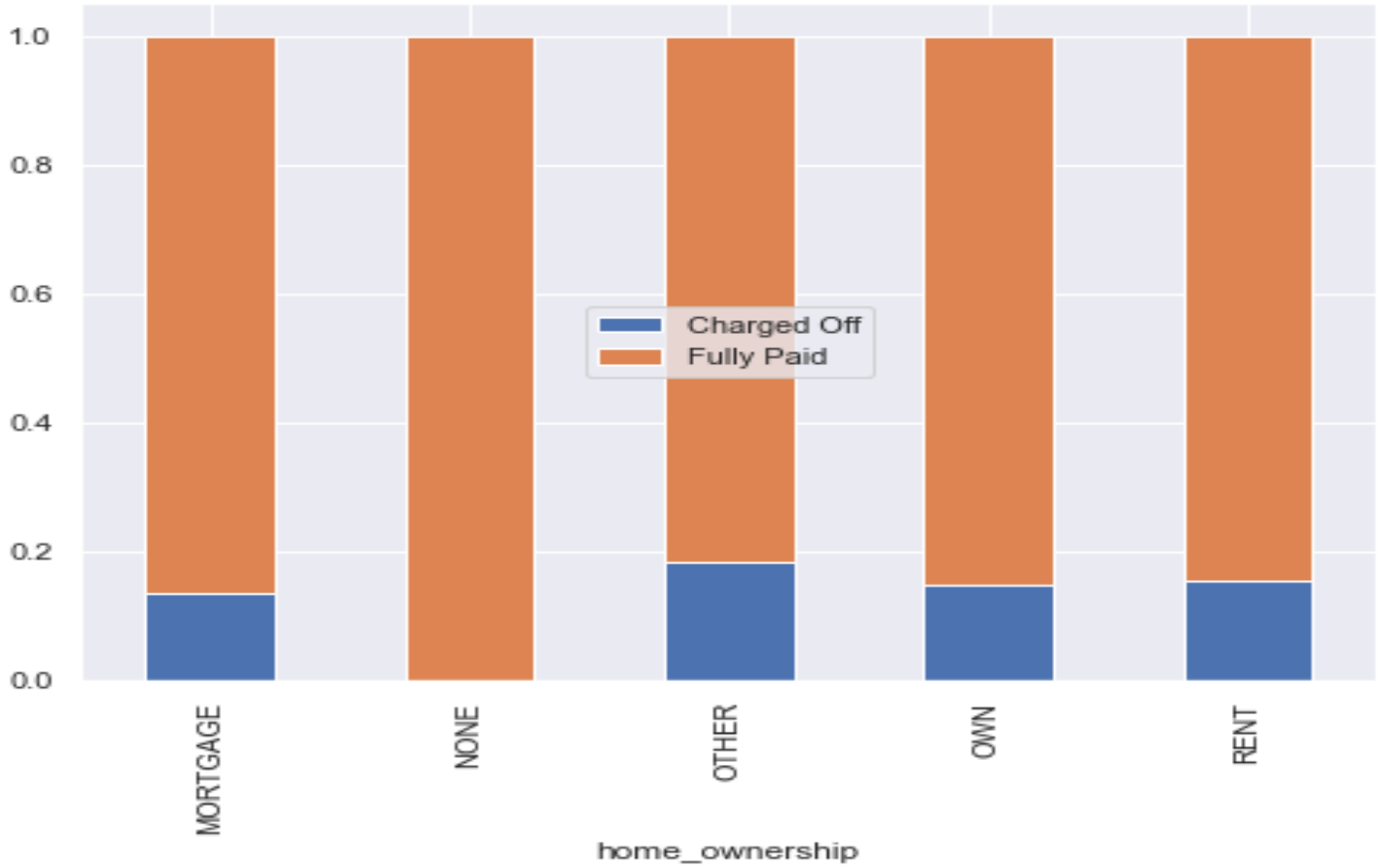
Defaults are significantly higher in loans taken for small business & debt consolidation
Customers with earlier multiple debts are more likely to default
Car purchase, wedding & major purchase are low risk categories for default

Bivariate Analysis (Verification Status Vs Paid Status: Fully Paid/Charged Off)



Verified customers are at slightly lower risk for default

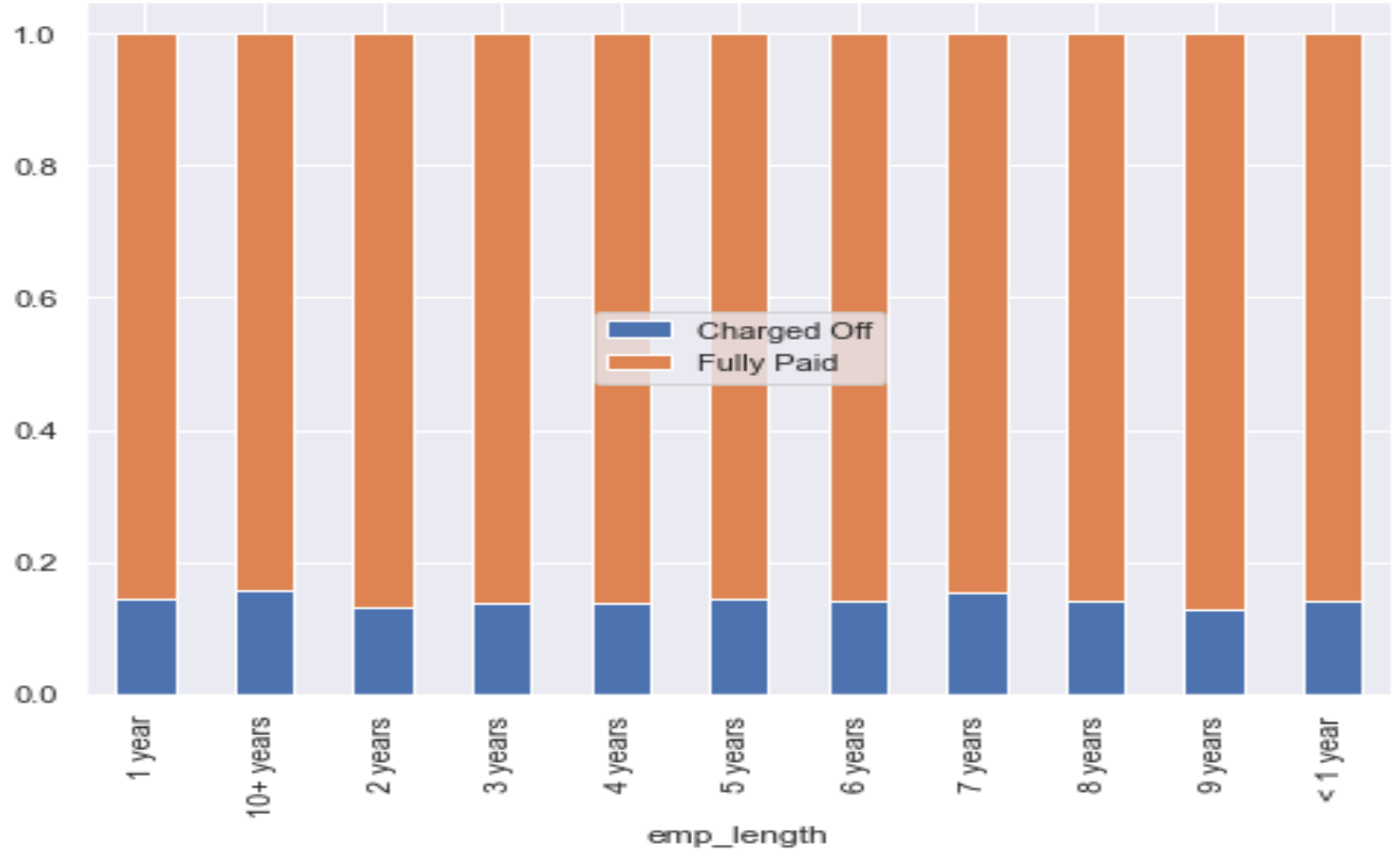
Bivariate Analysis (Home ownership Status Vs Paid Status: Fully Paid/Charged Off)



No significant variation seen between home ownership types

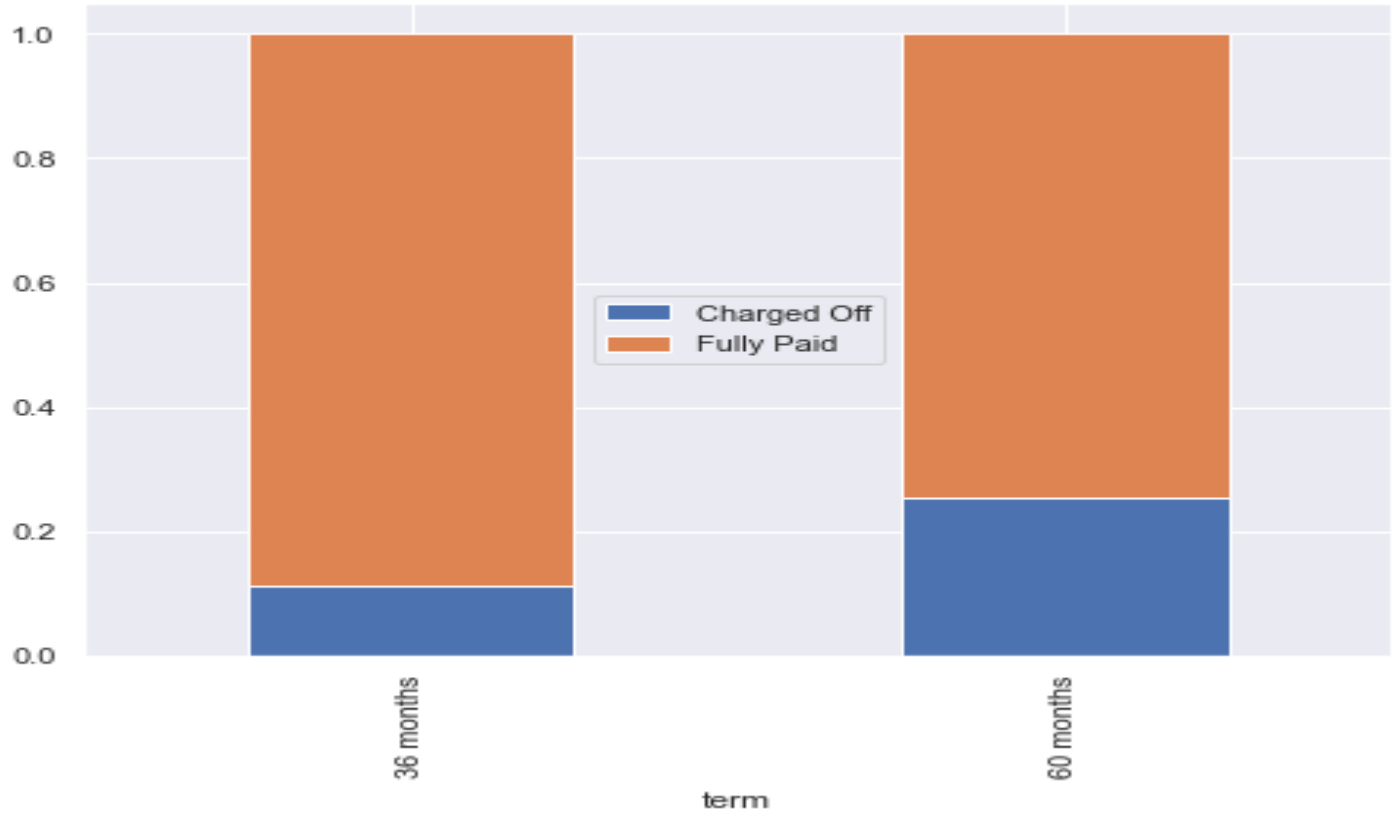
Some values are not known, which should be captured for all customers as a process

Bivariate Analysis (Employment Duration Vs Paid Status: Fully Paid/Charged Off)



Employment length < 1 years & >10+ years are at slightly higher risk of default

Bivariate Analysis (Loan Term Vs Paid Status: Fully Paid/Charged Off)



Customers with longer loan term are significantly higher at default

Inference Summary

High risk for default (To be considered during loan approval):
-High loan Amt (>15K)
-Low income
- Longer term
-High Debt to Income ratio
-Purpose: Small business
-Purpose: Education
-Purpose: Renewable energy
-Purpose: Debt consolidation (accumulated debt, probable multiple habitual debts)
-Very less (<1 yr) or very high (>10 yrs) employment duration
    -Probably due to lesser stability (<1 yr group) & accumulated debt lines (>10 yrs) respectively

Low risk for default:
-Verified source
- Shorter term
-Higher grade at application time
-Purpose: Car purchase
-Purpose: Wedding

Note: We recommend not to increase interest rate or term of loan to compensate for higher default risk, as both these factors are strongly related to further default probability