Contents

Problem Statement

The Lending Club wants to understand the driving variables which are strong indicators of loan default & loan repayment by customers by better risk assessment. This is to avoid financial loss & on the other hand, avoid business opportunity loss.

Definitions:

Default: Applicant has not paid the instalments in due time for a long period of time.
Charge-off: Applicant has not paid the instalments in due time for a long period of time.
Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
Current: Applicant is in the process of paying the instalments.

Approach considered

Exploratory Data Analysis to identify those factors and patterns in data.

Data Understanding

Data Source file: loan.csv , loan history of 39717 customers and 111 factors or variables.

Key Observation in the data:

- **Current** loan status can not to be included for analysis as these customers are still repaying loan & is not yet considered as defaulted or fully paid.
- Multiple missing values for various variables are observed in the data.
- Data for customers whose loans are rejected is not available due to relationship not continuing.
- Credit Score is not mentioned in the data, only Grade is assigned by the company to assess the risk.
- **Age** information of customers is not available
- **Interest rate** is provided as string data type instead of number.
- **Issue date** is the issue month instead of actual loan issue date.
- The **funded_amnt _inv <= loan_amnt.** Loan amount is amount applied for by customer & funded_amnt_inv field is the actual loan amounts funded by investors.

Data Cleaning

- Current loan status is excluded , 38577 rows are considered for the EDA.
- Dropped variables with missing values >85%.(dropped 58 col; missing value imputation is not considered for this analysis).
- Drop columns with same values in all rows ,unique value in each rows as these are not significant (dropped 14 col)
- Remaining useful data is 38,577 rows 39 columns.
- Month is derived from issue date.
- Interest Rate is string, converted into float data type to do numerical analysis.
- Out of 39 behavioral factors are removed, as these were not present at the time of application.

Data Cleaning – Manipulation

The columns of interest 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)
- Issue_Date (Month loan issued)
- home_ownership (Home type – rented, own, mortgage etc)
- verification_status (Whether details were verified)
- Installment (Monthly installment)

- Bins are created for numerical variable to understand how each buckets is related with charged Off/Fully Paid.
- Outliers are detected for all the numerical variables except DTI , but no treatment is considered for this study.

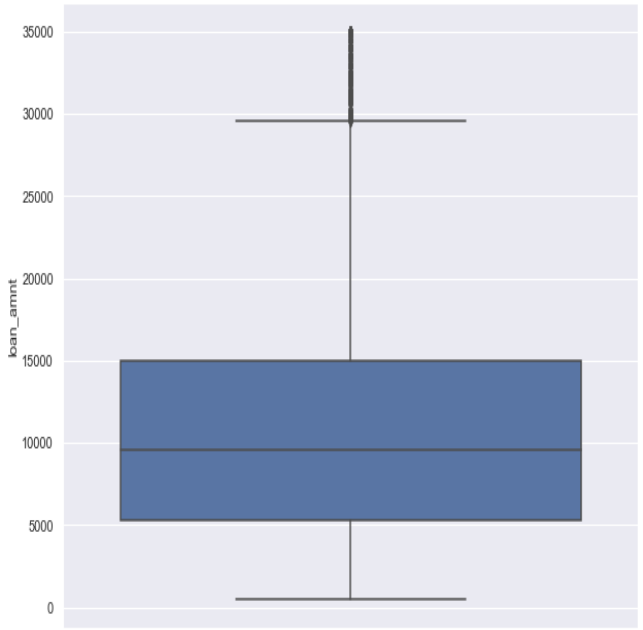
Summary Stats: Numeric Variables

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

Median is 50 percentile, since for most the variables median, mean and maximum values are highly different from each other so there is high chance of outliers in the respective variables.

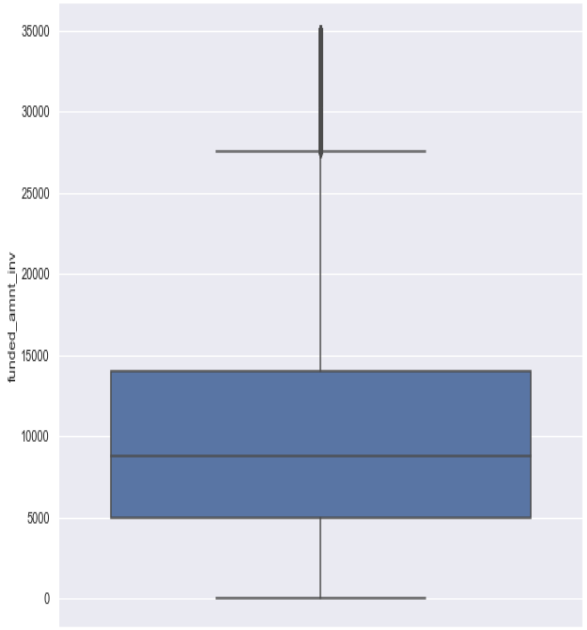
Univariate & Segmented Univariate Analysis

Loan Amt

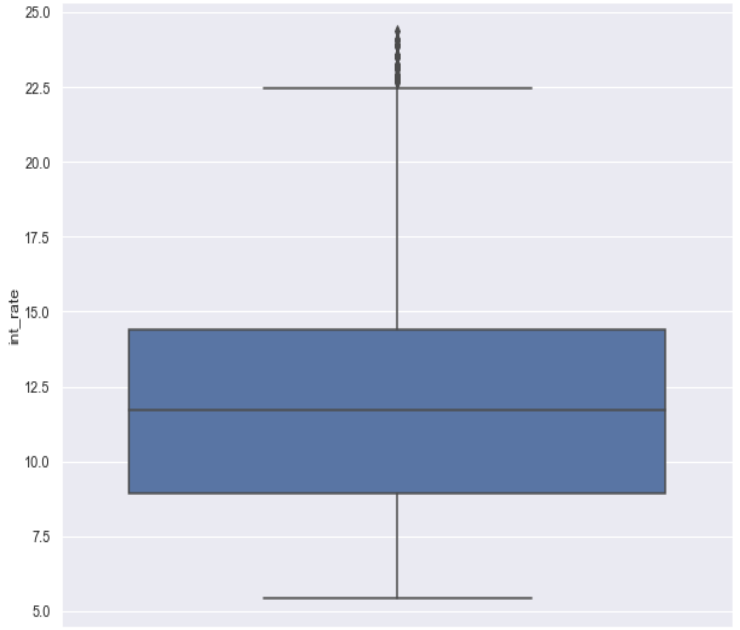


- Outliers are present in both Loan Amount & Funded Amount.
- Average loan amount applied 9.6K Vs Funded amount 8.7K.

Funded Amt

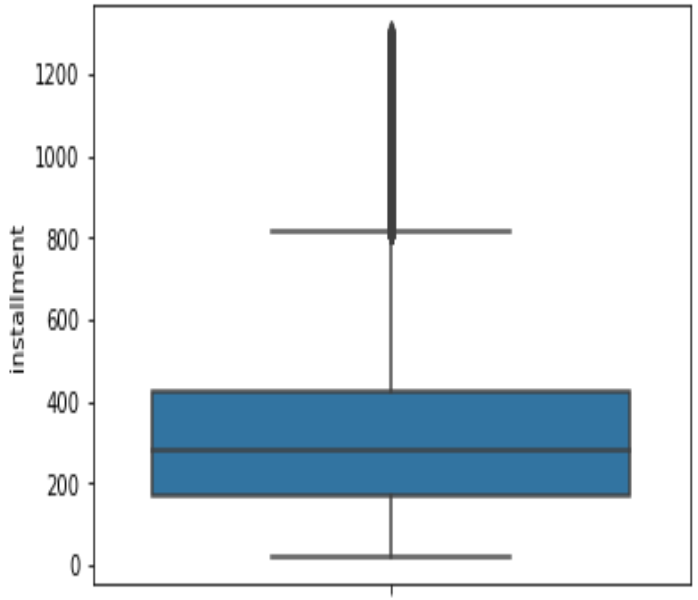


Interest Rate



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

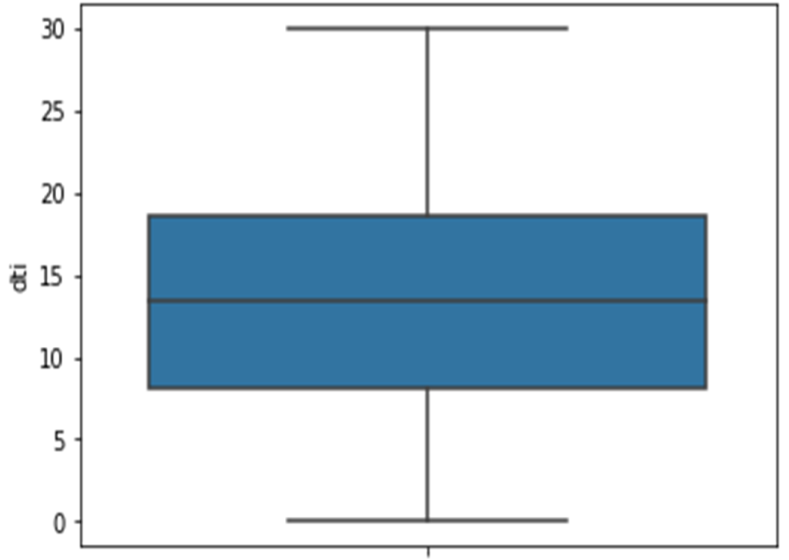
Installment



Monthly Installment Amount:

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

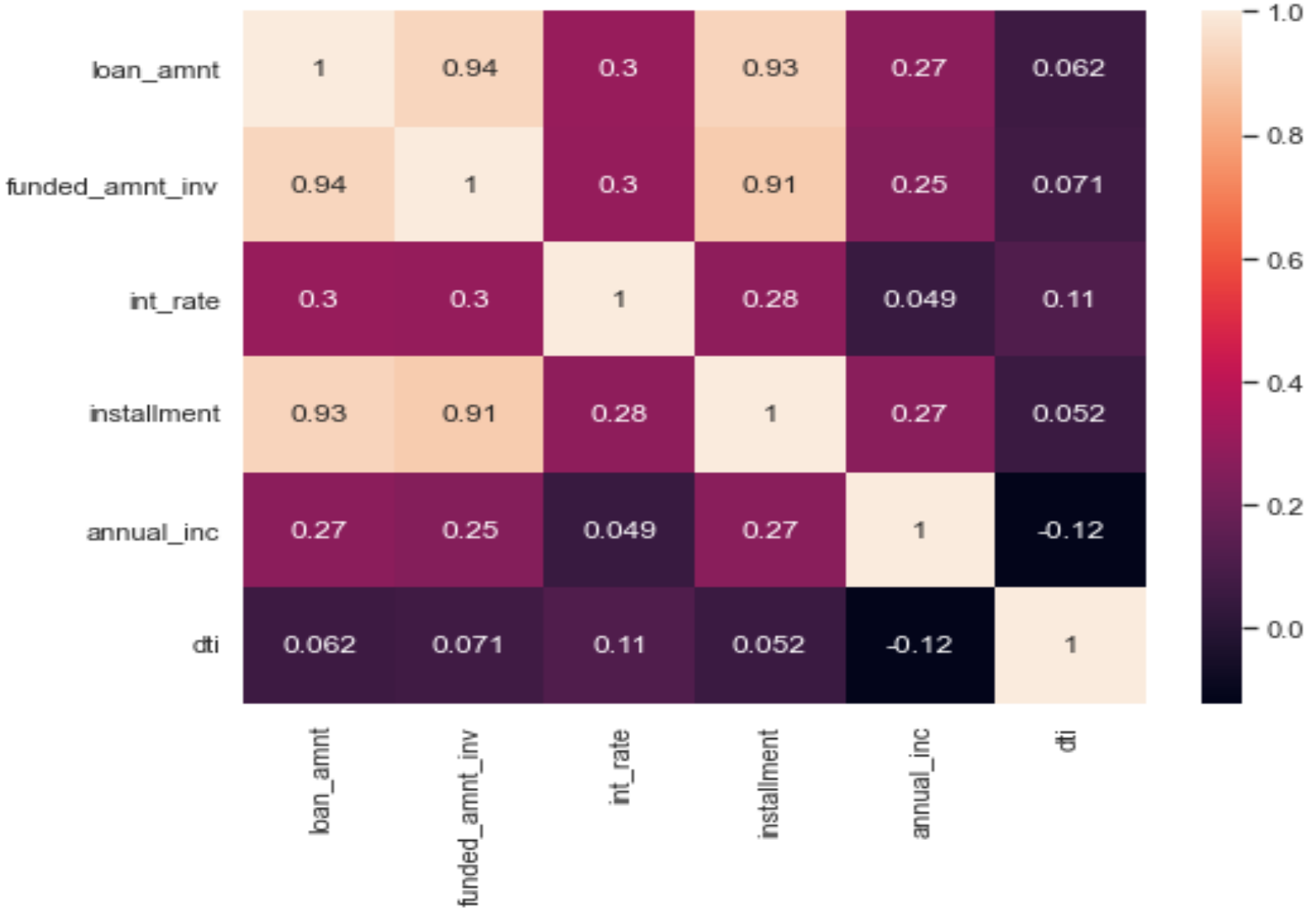
DTI(Debt to Income Ratio)



Debt to Income ratio:

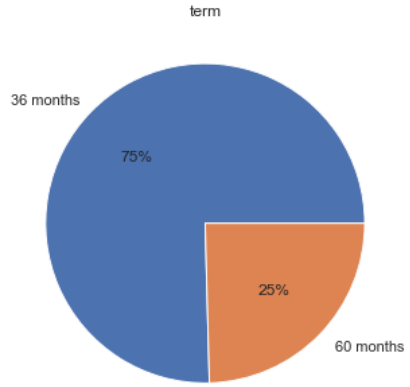
- Most of the dti are between 5 to 25 percent, median is 13.37
- No outliers is observed in DTI

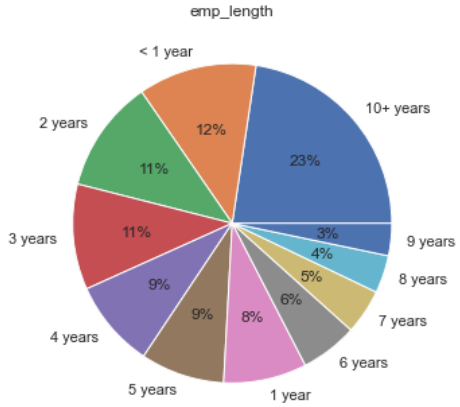
Correlation Matrix (Variables Of Interest)

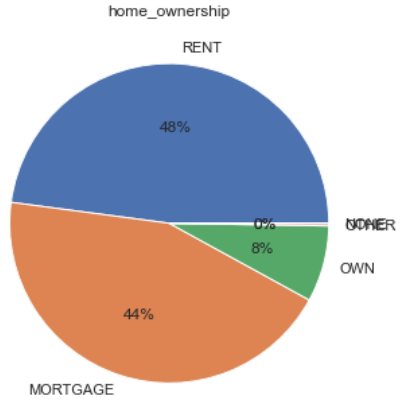


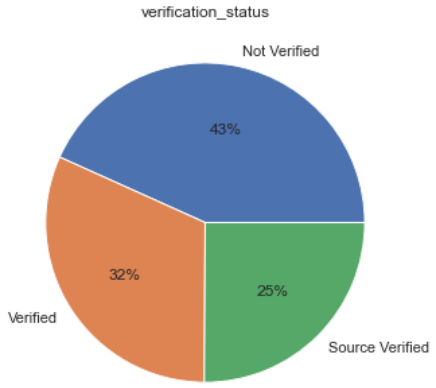
- Funded amount & Installment amount are highly correlated(r=.94)
- Higher installment to cover higher loan amount in limited period.
- dti is negatively least correlated with annual income. higher the income lower is the debt to income ratio.

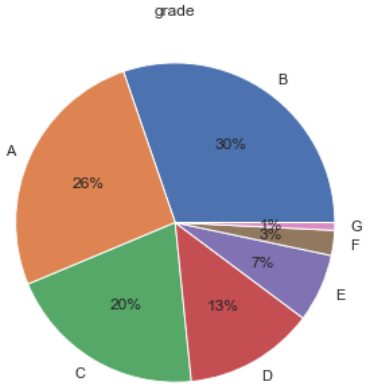
Univariate Analysis (Categorical Variables)





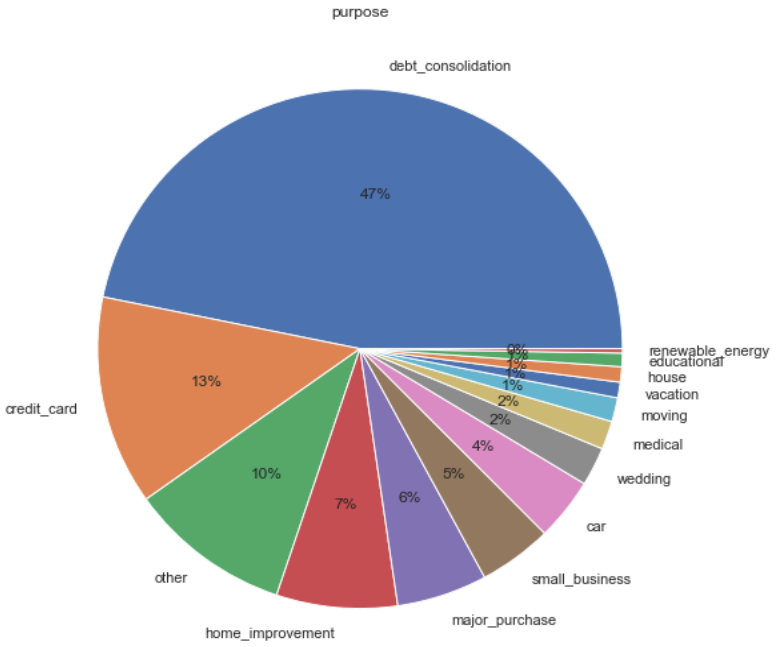


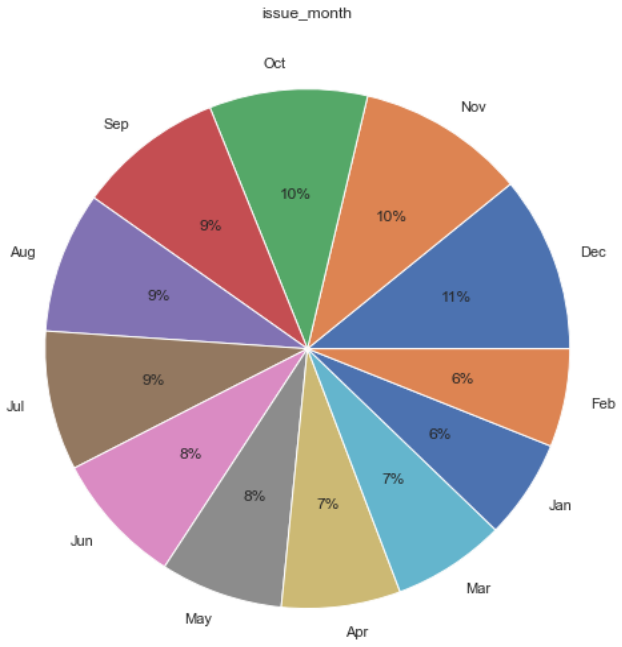


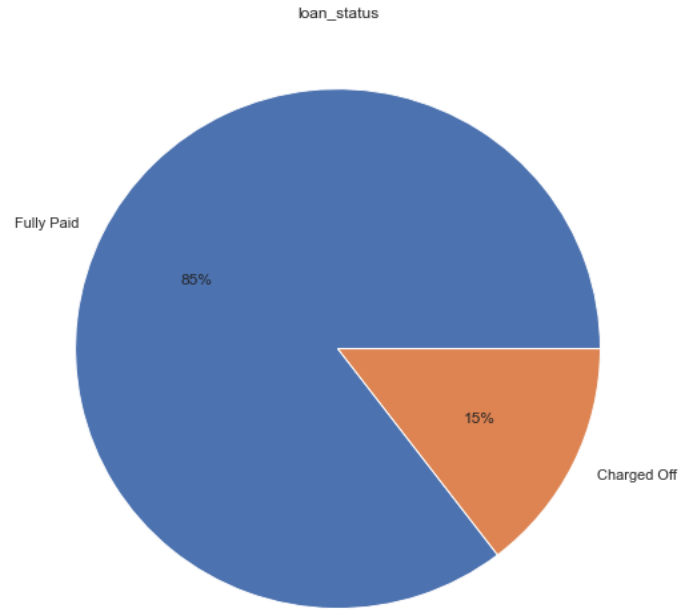


Major populations under each segments are:
Term **: 36 months**, Employment length :**10+ years & <=3 years**,
Home Ownership: Rental & Mortgage, Verification: **“Not verified”** .
Grades: **B, A & C**

Univariate Analysis (Categorical variables)





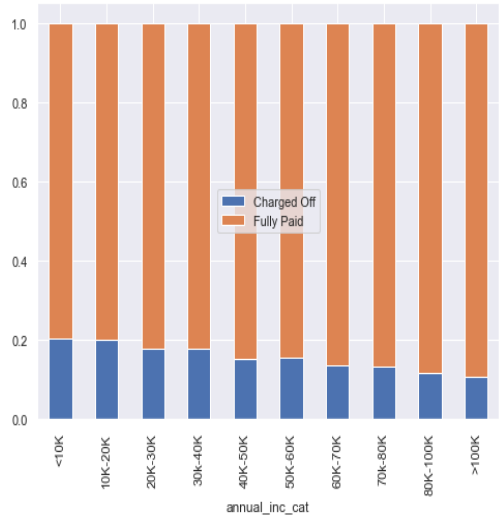


- **Purpose**: Major loans are taken for **debt consolidation**, **Credit Card**.
- **Loan Status**: 85% of approved loans are Fully paid & 15% are defaulted.
- **Months**: Sept to Dec loans approval are higher as compare other months

Bivariate Analysis

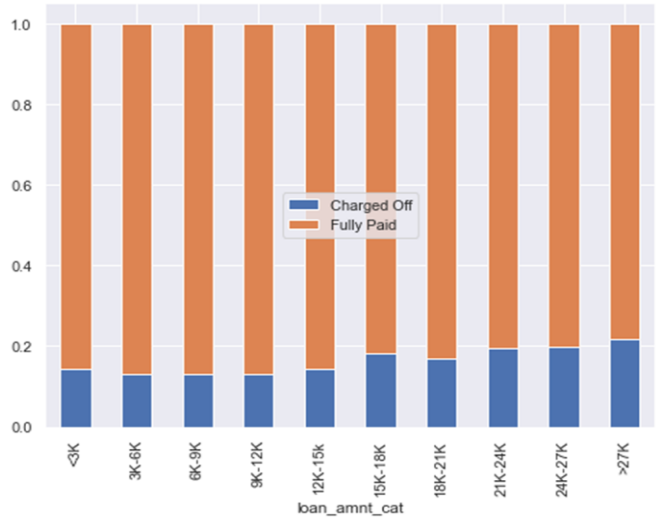
All Bivariate Analysis is between Loan Status Fully Paid/Charged Off Vs Variables

Annual Income



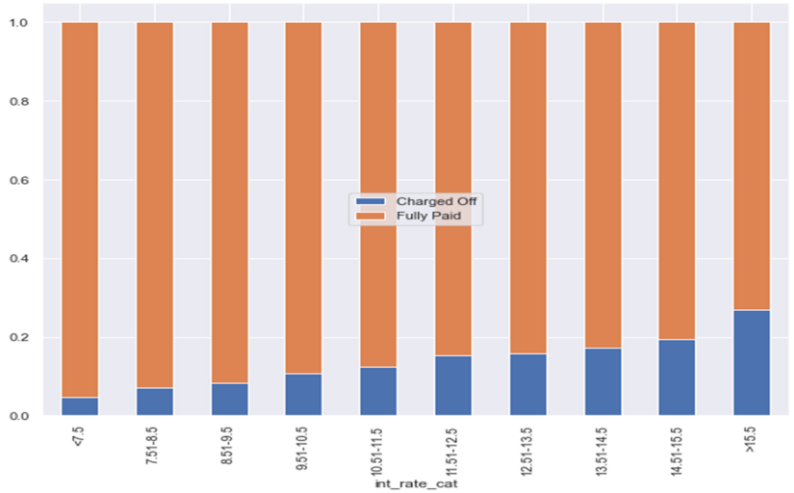
Annual Income	ChargedOff	FullyPaid
<10K	20.4%	79.6%
10K-20K	20.1%	79.9%
20K-30K	17.9%	82.1%
30k-40K	17.7%	82.3%
50K-60K	15.6%	84.4%
40K-50K	15.1%	84.9%
60K-70K	13.6%	86.4%
70k-80K	13.2%	86.8%
80K-100K	11.8%	88.2%
>100K	10.8%	89.2%

Loan Amount



Loan Amount	ChargedOff	FullyPaid
>27K	21.7%	78.3%
24K-27K	19.8%	80.2%
21K-24K	19.4%	80.6%
15K-18K	18.1%	81.9%
18K-21K	16.8%	83.2%
<3K	14.3%	85.7%
12K-15k	14.2%	85.8%
9K-12K	13.2%	86.8%
6K-9K	13.1%	86.9%
3K-6K	13.0%	87.0%

Interest Rates

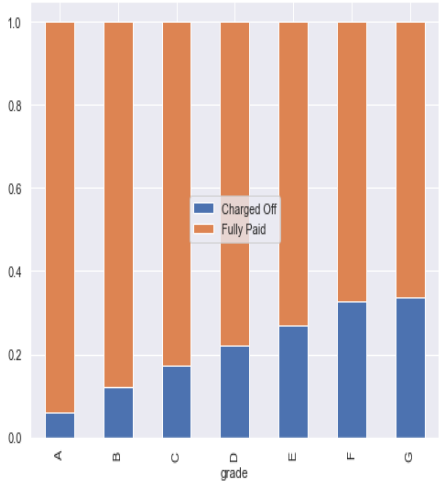


Interst Rate(in %)	ChargeOff	Paid
>15.5	26.8%	73.2%
14.51-15.5	19.4%	80.6%
13.51-14.5	17.3%	82.7%
12.51-13.5	15.8%	84.2%
11.51-12.5	15.2%	84.8%
10.51-11.5	12.3%	87.7%
9.51-10.5	10.7%	89.3%
8.51-9.5	8.2%	91.8%
7.51-8.5	7.1%	92.9%
<7.5	4.8%	95.2%

	High Risk
	Medium Risk
	Low Risk

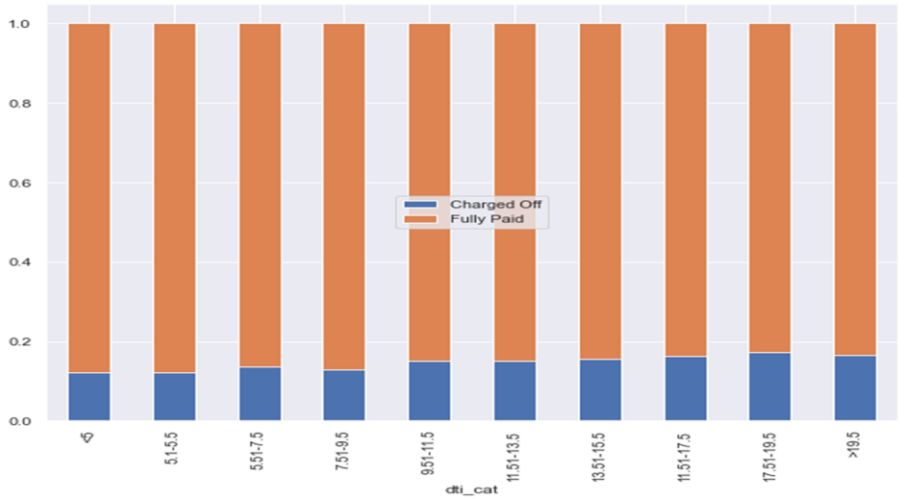
- Higher annual income, lower loan amount, low interest rates = low risk customers.
- Loan Amount and Funded Amount are highly correlated(r=.94), so can keep any one of them to avoid instability in the factors and redundancy. We preferred Loan Amount as customer requested amount (need) is more significant for this analysis.

Grade



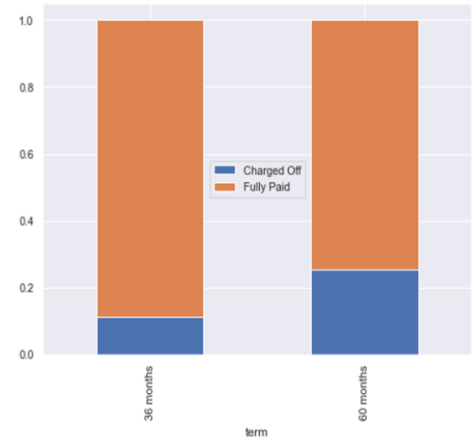
Grade	ChargedOff	Fully Paid
G	33.8%	66.2%
F	32.7%	67.3%
E	26.8%	73.2%
D	22.0%	78.0%
C	17.2%	82.8%
B	12.2%	87.8%
A	6.0%	94.0%

DTI



DTI(Debt /Income)	Charged Off	Fully Paid
19.51-21.5	17.2%	82.8%
>21.5	16.6%	83.4%
17.51-19.5	16.3%	83.7%
15.51-17.5	15.5%	84.5%
13.51-15.5	15.0%	85.0%
11.51-13.5	15.0%	85.0%
7.51-9.5	13.7%	86.3%
9.51-11.5	12.9%	87.1%
<5.5	12.1%	87.9%
5.51-7.5	12.1%	87.9%

Term

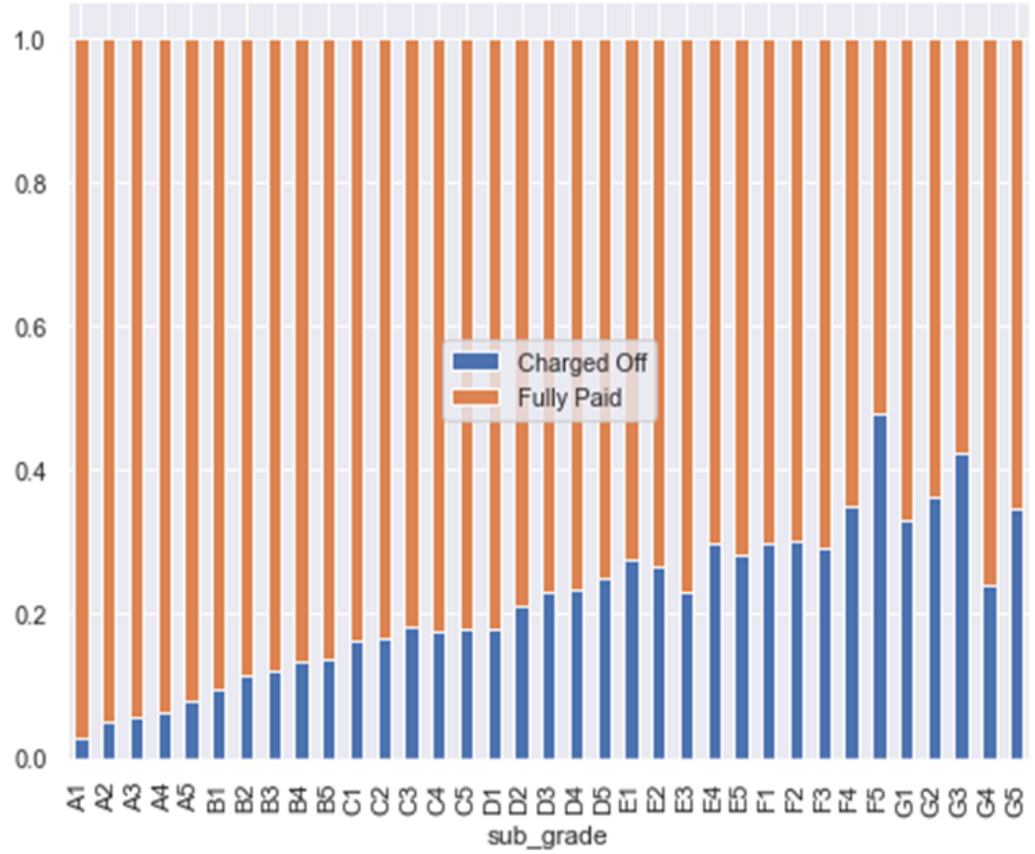


Term	Charged Off	Fully Paid
60 months	25.3%	74.7%
36 months	11.1%	88.9%

- **Grade** : A is lowest risk& G is highest risk
- **Term**: 60 months are more than twice risky than 36 months
- Lower DTI = Lower risk.

- **DTI, Grade & loan term are strong driving factors as these variables are showing significant pattern and rank order.**
- 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 default.

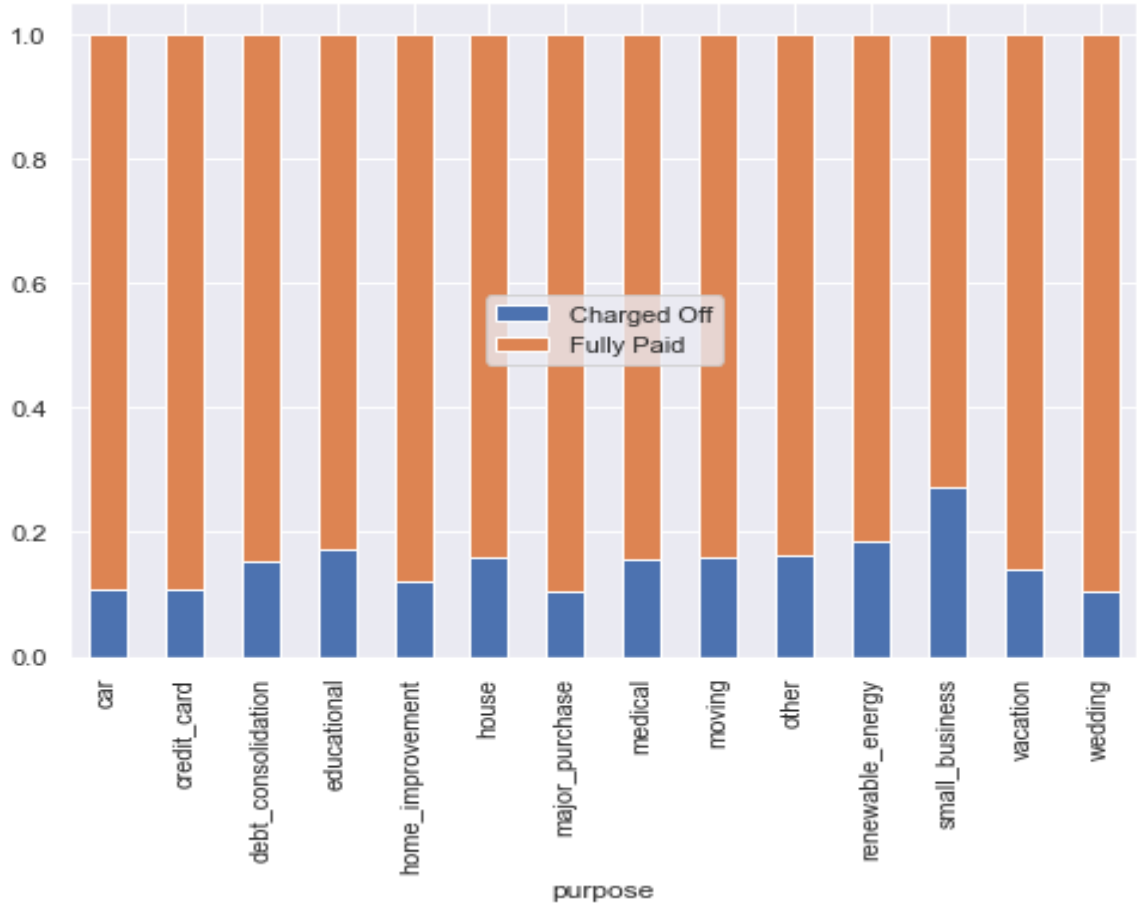
Sub_garde



Subgrade below B5 are risky, as grade below C (refer previous slide) are also at moderate risk so it is advisable not to approve loan below grade C .

SubGrade	ChargedOff	Fully Paid
F5	47.8%	52.2%
G3	42.2%	57.8%
G2	36.4%	63.6%
F4	35.1%	64.9%
G5	34.5%	65.5%
G1	33.0%	67.0%
F2	30.0%	70.0%
F1	29.8%	70.2%
E4	29.7%	70.3%
F3	29.3%	70.7%
E5	28.2%	71.8%
E1	27.4%	72.6%
E2	26.5%	73.5%
D5	25.1%	74.9%
G4	24.1%	75.9%
D4	23.4%	76.6%
E3	23.1%	76.9%
D3	22.9%	77.1%
D2	21.1%	78.9%
C3	18.1%	81.9%
C5	18.0%	82.0%
D1	17.9%	82.1%
C4	17.6%	82.4%
C2	16.6%	83.4%
C1	16.4%	83.6%
B5	13.6%	86.4%
B4	13.5%	86.5%
B3	12.1%	87.9%
B2	11.4%	88.6%
B1	9.5%	90.5%
A5	8.0%	92.0%
A4	6.2%	93.8%
A3	5.7%	94.3%
A2	4.9%	95.1%
A1	2.6%	97.4%

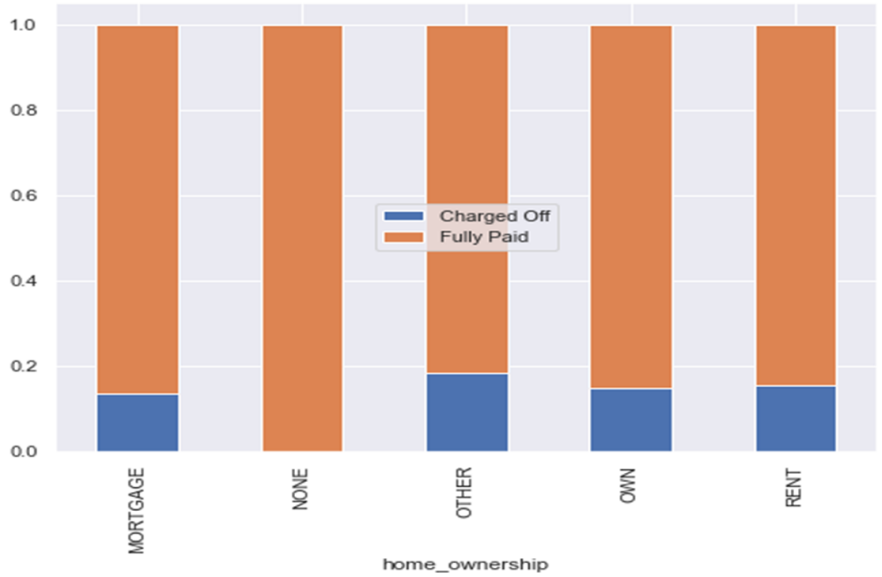
Purpose



Purpose	Charged Off	Fully Paid
small_business	27.1%	72.9%
renewable_energy	18.6%	81.4%
educational	17.2%	82.8%
other	16.4%	83.6%
house	16.1%	83.9%
moving	16.0%	84.0%
medical	15.6%	84.4%
debt_consolidation	15.3%	84.7%
vacation	14.1%	85.9%
home_improvement	12.1%	87.9%
credit_card	10.8%	89.2%
car	10.7%	89.3%
wedding	10.4%	89.6%
major_purchase	10.3%	89.7%

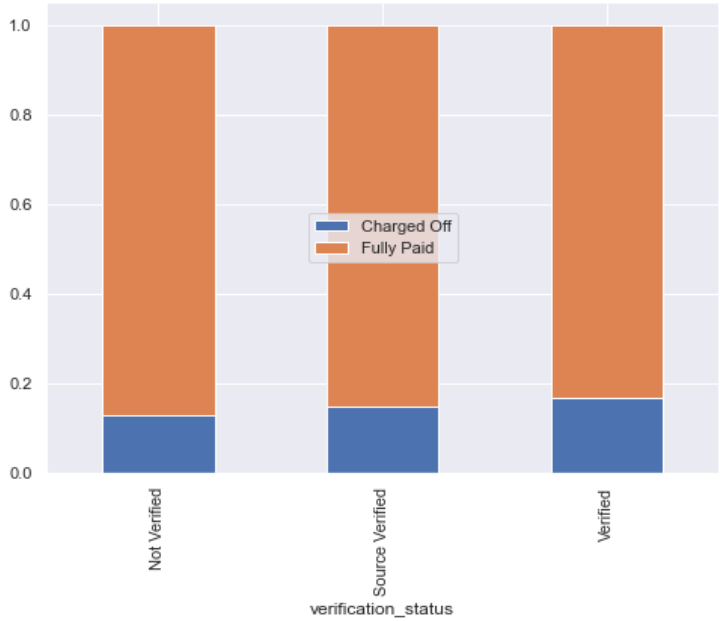
- Customers with purpose of loan as debt consolidation (earlier multiple debts), medical, education, renewable energy & small business are more likely to default
- Credit card, car purchase, wedding & major purchase are low risk categories for default

Home Ownership



Home Ownership	Charged Off	Fully Paid
OTHER	18.4%	81.6%
RENT	15.4%	84.6%
OWN	14.9%	85.1%
MORTGAGE	13.7%	86.3%
NONE	0.0%	100.0%

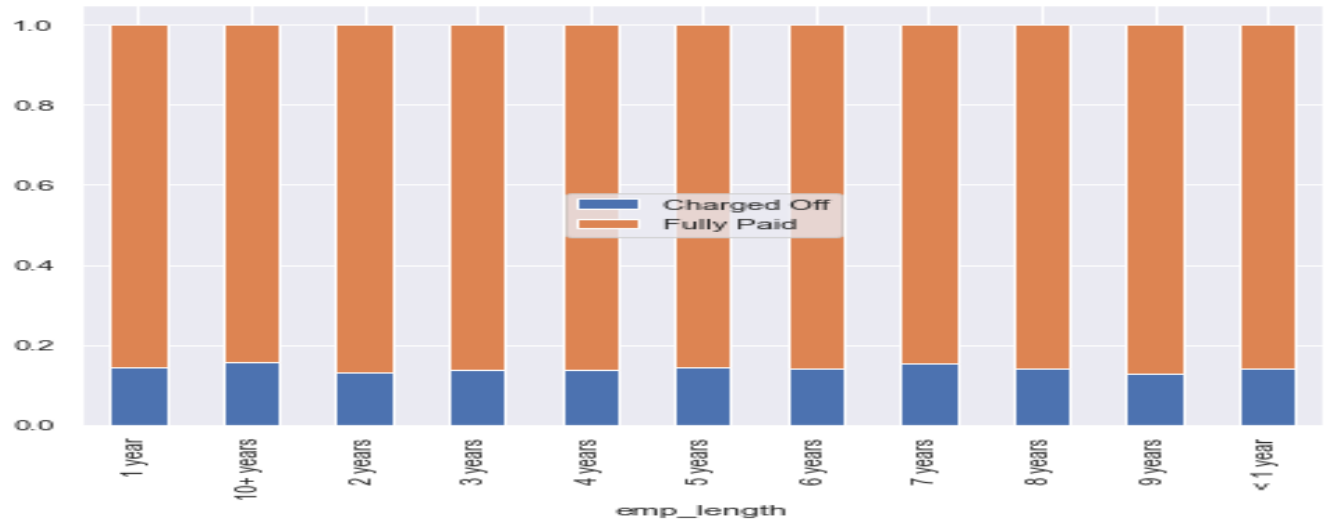
Verification Status



Verification Status	ChargedOff	Fully Paid
Verified	16.8%	83.2%
Source Verified	14.8%	85.2%
Not Verified	12.8%	87.2%

- Even though Verified customers are at slightly higher risk for default, it does not make business sense to avoid verification.
- Home ownership : Mortgage has lower charge off risk compared to “own”, “Others” and “rent” which are having more charged off potential.
- This also deviates from general assumption that House owner will have least charge off risk.
- Some Home Ownership types (Other) are not known, which should be captured for all customers to do better risk assessment.

Emp_length



Emp Length	Charged Off	Fully Paid
1 0+ years	15.7%	84.3%
1 year	14.4%	85.6%
< 1 year	14.2%	85.8%
7 years	15.4%	84.6%
5 years	14.3%	85.7%
6 years	14.2%	85.8%
8 years	14.1%	85.9%
3 years	13.8%	86.2%
4 years	13.8%	86.2%
2 years	13.2%	86.8%
9 years	12.9%	87.1%

- **Emp_length: Out of total charged off population, ~44.2% belong to emp_length 10+ years and <=1 year.**
- **Emp_length is not highly significant as the charge offs are almost uniformly distributed across the categories.**

Summary

The below 5 variables shows a clear and significant trend to identify the Fully Paid customers.

Driving Factors	Risk Appetite		
	Low	Medium	High
Loan Amount	<=15K	15K - <= 21K	>21K
Interest Rate	<=10.5%	>10.5 - <=13.5 %	>13.5%
DTI	<=9.5	>9.5 -<=17.5	>17.5
Term	36 months	60 months	60 months
grade	A	B,C	D,E,F,G

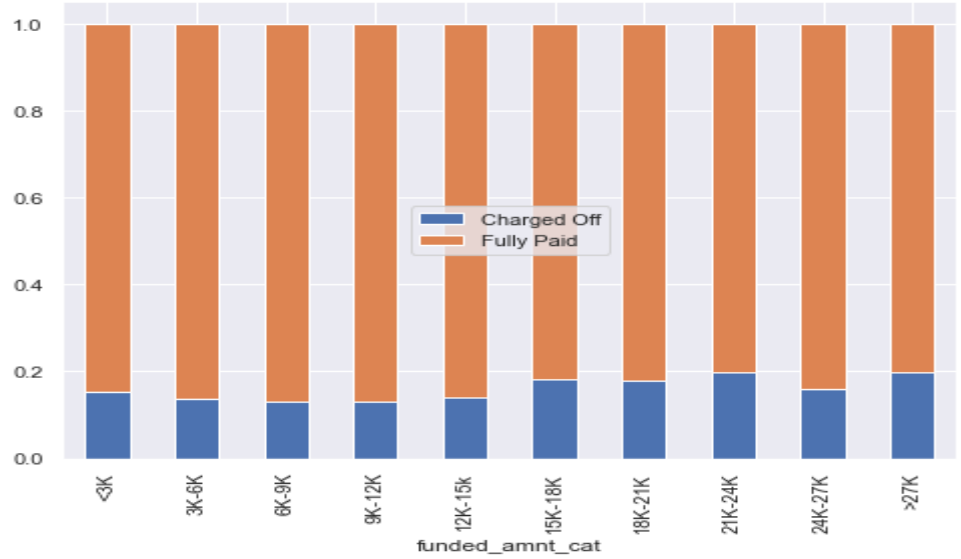
- Low Loan Amount is less likely to be charged off.
- Low interest rate is less likely to be charged off.
- Low Debt to Income Ratio is less likely to be charged off.
- Lesser term is less likely to be charged off
- Higher grade is less likely to charged off.

Reason for not considering the Annual Income:
High income does not always mean high ATP(Ability to Pay), rather low DTI(debt to income ratio) shows better ATP.

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

Appendix :

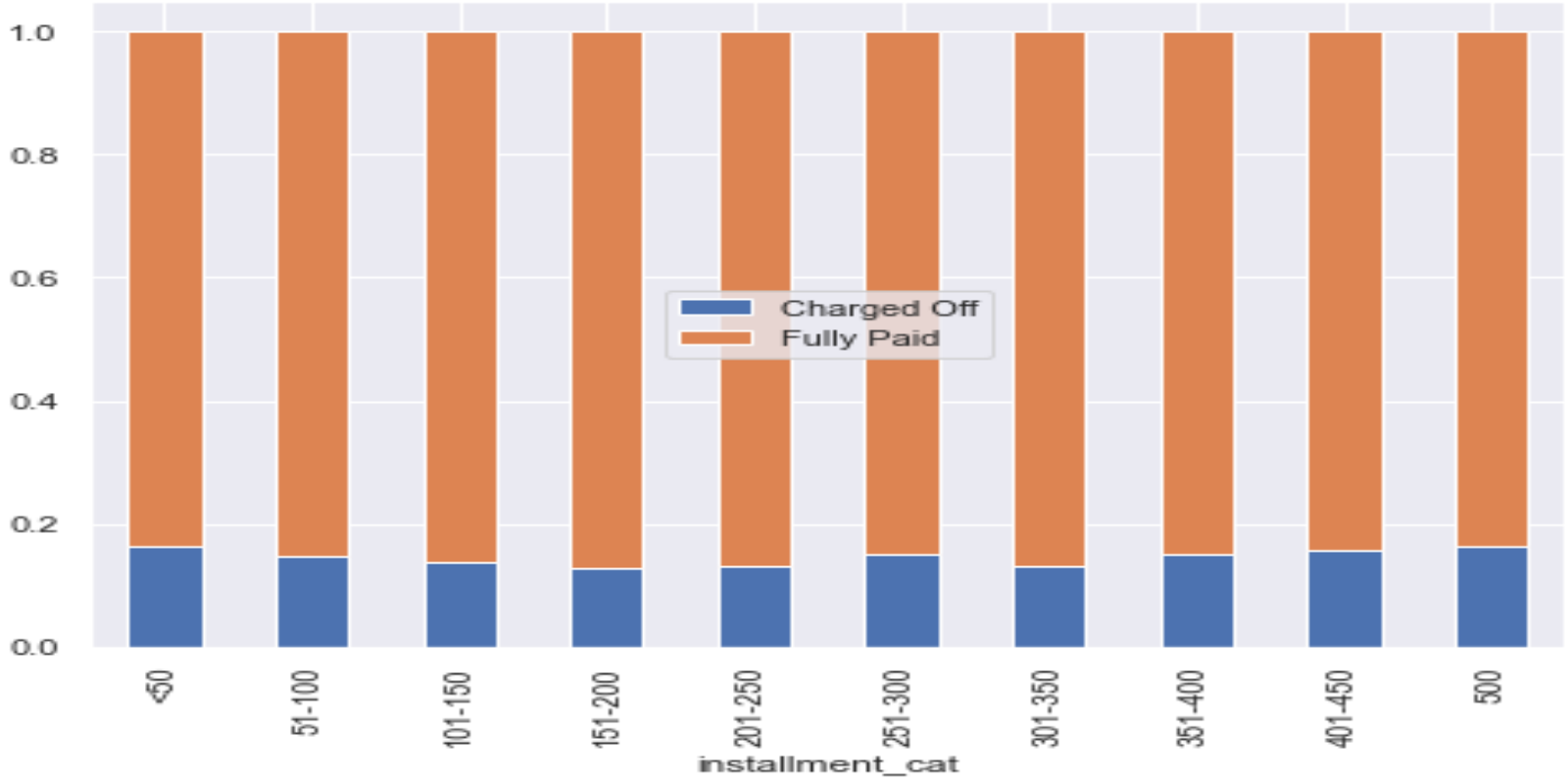
Funded Amount Inv



Funded_Amnt_Inv	ChargedOff	Fully Paid
21K-24K	19.9%	80.1%
>27K	19.9%	80.1%
15K-18K	18.2%	81.8%
18K-21K	17.9%	82.1%
24K-27K	16.0%	84.0%
<3K	15.3%	84.7%
12K-15k	14.1%	85.9%
3K-6K	13.6%	86.4%
9K-12K	13.1%	86.9%
6K-9K	13.0%	87.0%

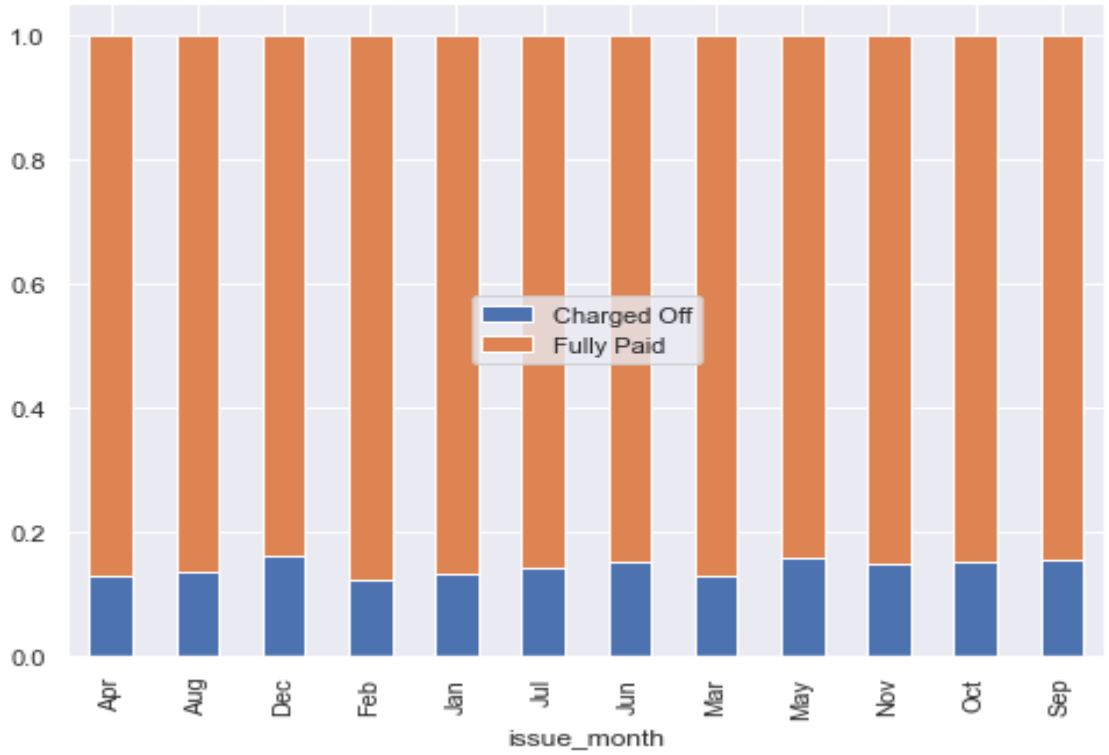
- Funded Amount >15K (actual approved loan Amount) higher is chance of default. Except funded amt <3K the trend is consistent.
- Loan Amount and Funded Amount are highly correlated(r=.94), so can keep any one of them to avoid instability in the factors and redundancy.
- We preferred Loan Amount as customer requested amount (need) is more significant for this analysis.

Installment Amount



- Installment is applicable only when the loan are approved. We considered it to check how the charged off are related with installment amount for the existing customer.
- It shows no significance.

Issue month



Lon_Issue_Month	Charged Off	Fully Paid
Dec	16.1%	83.9%
May	16.0%	84.0%
Sep	15.6%	84.4%
Oct	15.4%	84.6%
Jun	15.2%	84.8%
Nov	14.9%	85.1%
Jul	14.3%	85.7%
Aug	13.8%	86.2%
Jan	13.5%	86.5%
Apr	13.1%	86.9%
Mar	12.9%	87.1%
Feb	12.3%	87.7%

Months: sept,oct,dec,may are higher charged off as compare to other months, the difference is not significant enough

The End

Thank You