PROBLEM STATEMENT :

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

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 utilise this knowledge for its portfolio and risk assessment.

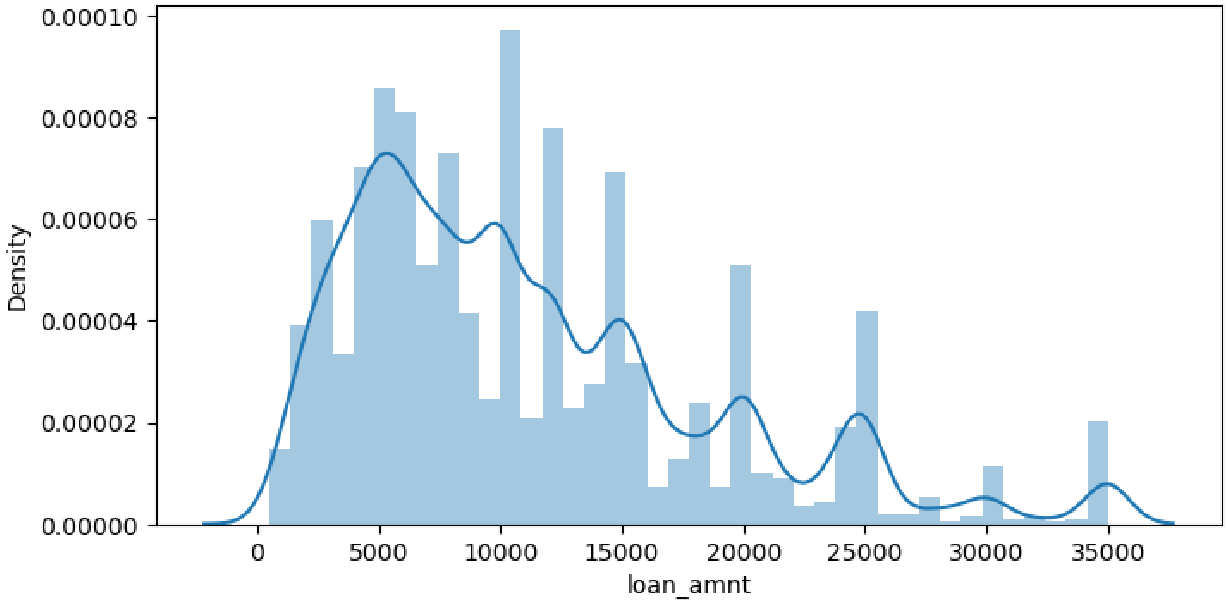
GENERAL INFORMATION

The aim of this project was to analyze and gain insights into the loan portfolio of a financial institution. The dataset consisted of various loan-related attributes such as loan amount, loan status, borrower information, and other relevant factors. Through extensive data exploration and visualization, we examined the distribution of loan amounts, loan status, and borrower characteristics. Additionally, we conducted bivariate analyses to understand the relationships between different variables. By leveraging data analysis techniques, we identified trends, patterns, and risk factors associated with credit risk. Overall, this project highlights the value of data analysis in mitigating credit risk, enhancing loan portfolio quality, and ensuring the financial stability of the institution.

An Approach To Find A Solution To Problem Statement:

Loan Amount:

Checking the distribution of Loan Amount Column

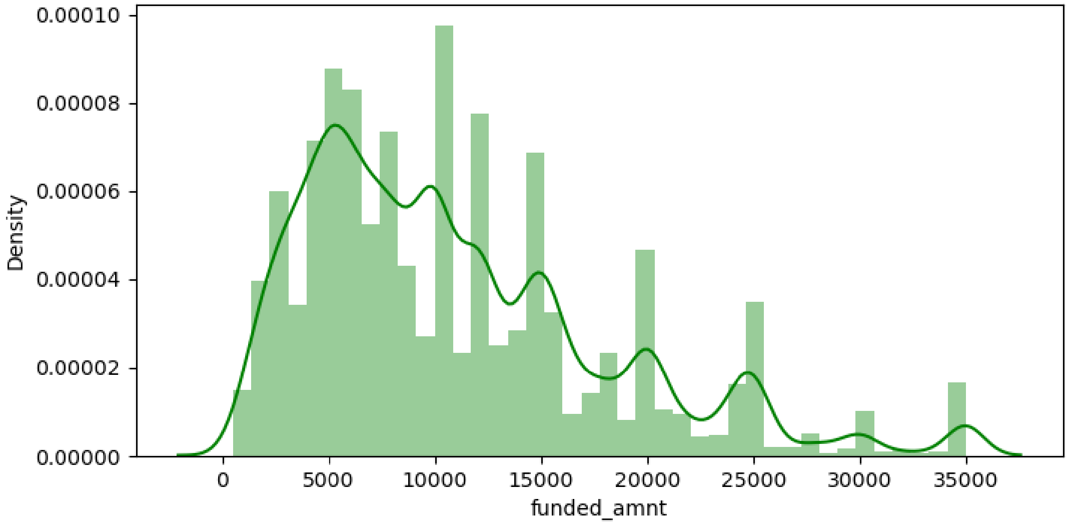


Observation:

The distribution of the loan amount column appear to be right skewed and most of the loans those were advanced lie between 1 to 15000 dollars (assuming the currency is dollars).

Funded Amount:

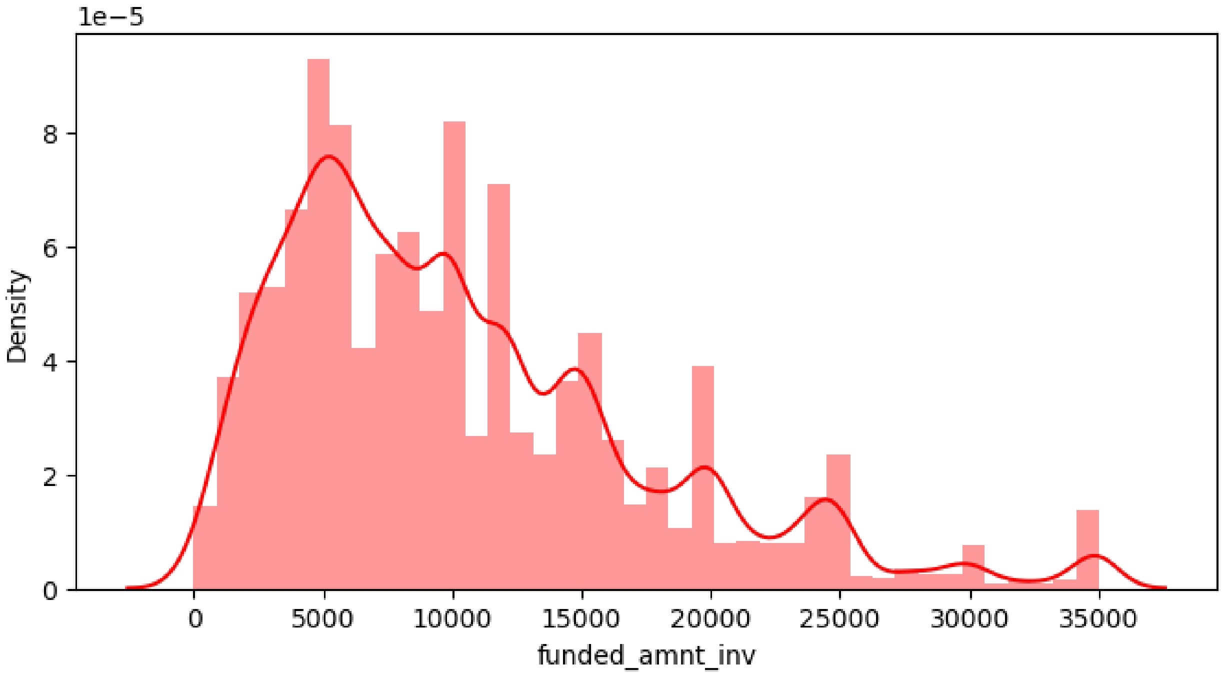
The total amount committed to that loan at that point in time.

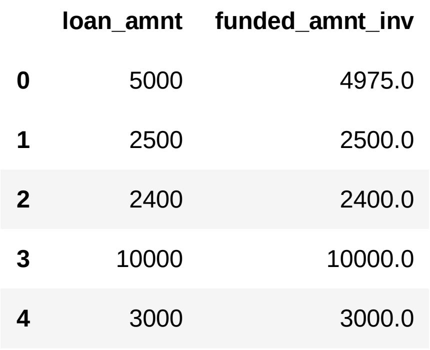


Observation:

From the description of loan amount column and the funded amount column and comparing the values in these two columns it appears that both the column contains same information. Hence, to avoid data duplicity and redundancy its better to drop one of these two columns.

Funded_Amount_Inv





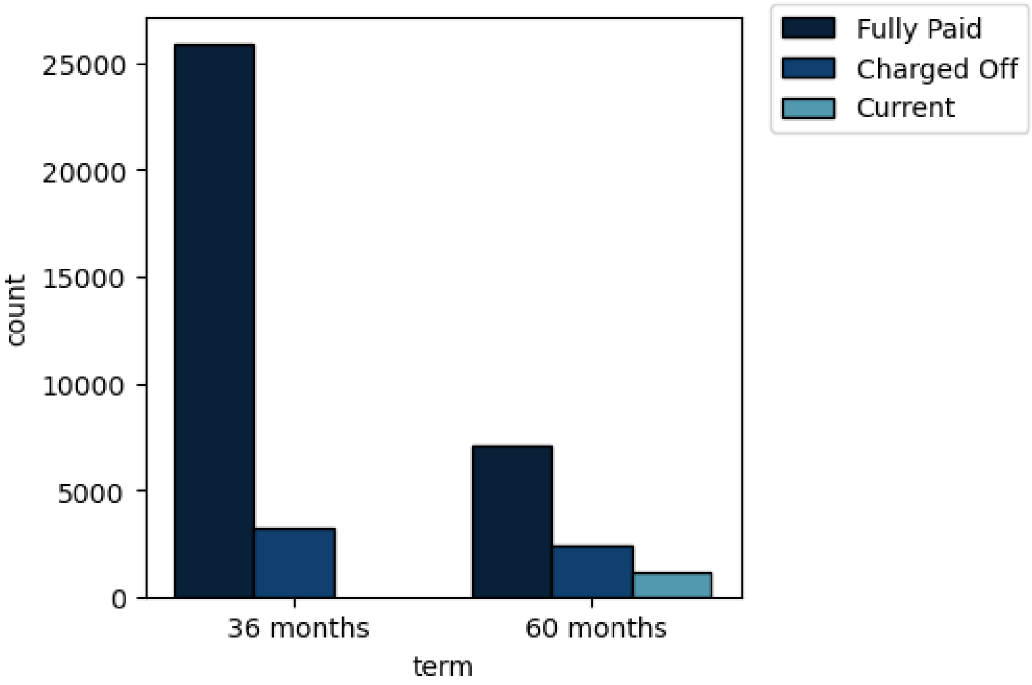
	loan_amnt	funded_amnt_inv
0	5000	4975.0
1	2500	2500.0
2	2400	2400.0
3	10000	10000.0
4	3000	3000.0

Observation:

Out of total of 39717 loans those were sanctioned there are 20189 borrowers who did not avail the entire amount that was sanctioned to them and instead committed to the lesser loan amount.

Term:

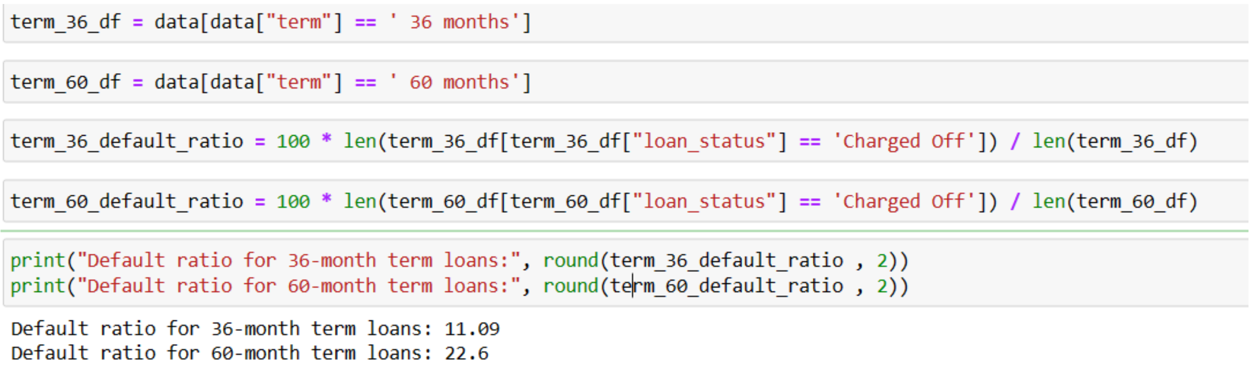
The number of payments on the loan. Values are in months and can be either 36 or 60.



Observation :

Most Loans were given for 36 Months term

Checking Default Ratio wrt Term:



```
term_36_df = data[data["term"] == ' 36 months']
```

```
term_60_df = data[data["term"] == ' 60 months']
```

```
term_36_default_ratio = 100 * len(term_36_df[term_36_df["loan_status"] == 'Charged Off']) / len(term_36_df)
```

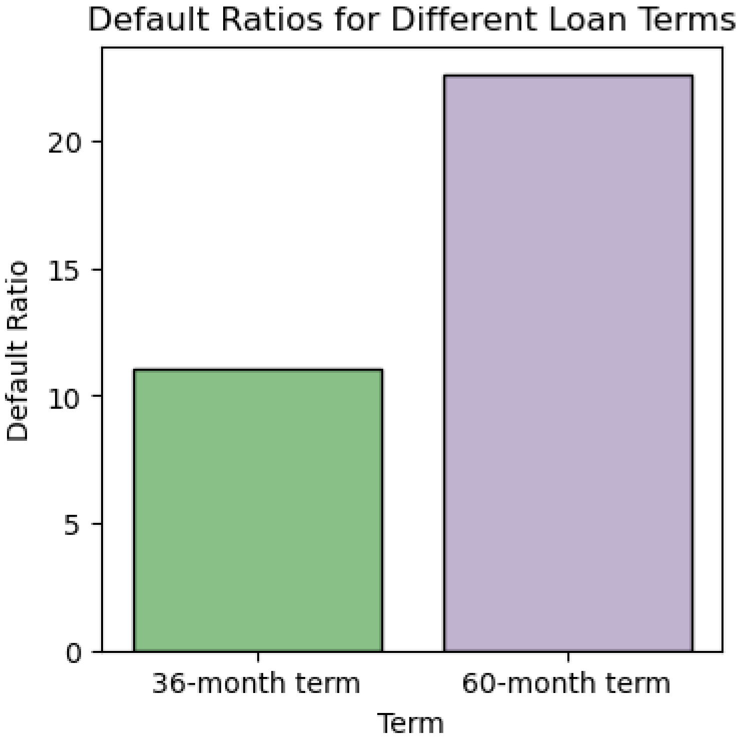
```
term_60_default_ratio = 100 * len(term_60_df[term_60_df["loan_status"] == 'Charged Off']) / len(term_60_df)
```

```
print("Default ratio for 36-month term loans:", round(term_36_default_ratio , 2))
print("Default ratio for 60-month term loans:", round(term_60_default_ratio , 2))
```

```
Default ratio for 36-month term loans: 11.09
Default ratio for 60-month term loans: 22.6
```

Observation:¶

Clearly from the above plot is appears that if the term of the loan is 60 months the likelihood of default is more.

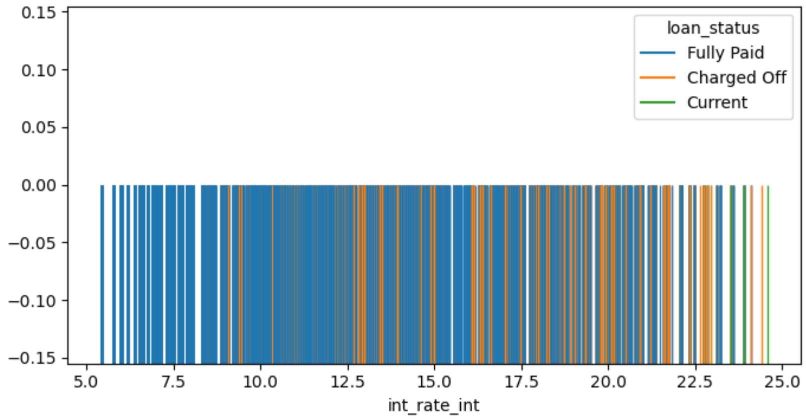


Observation:¶

Clearly from the above plot is appears that if the term of the loan is 60 months the likelihood of default is more.

Interest Rate

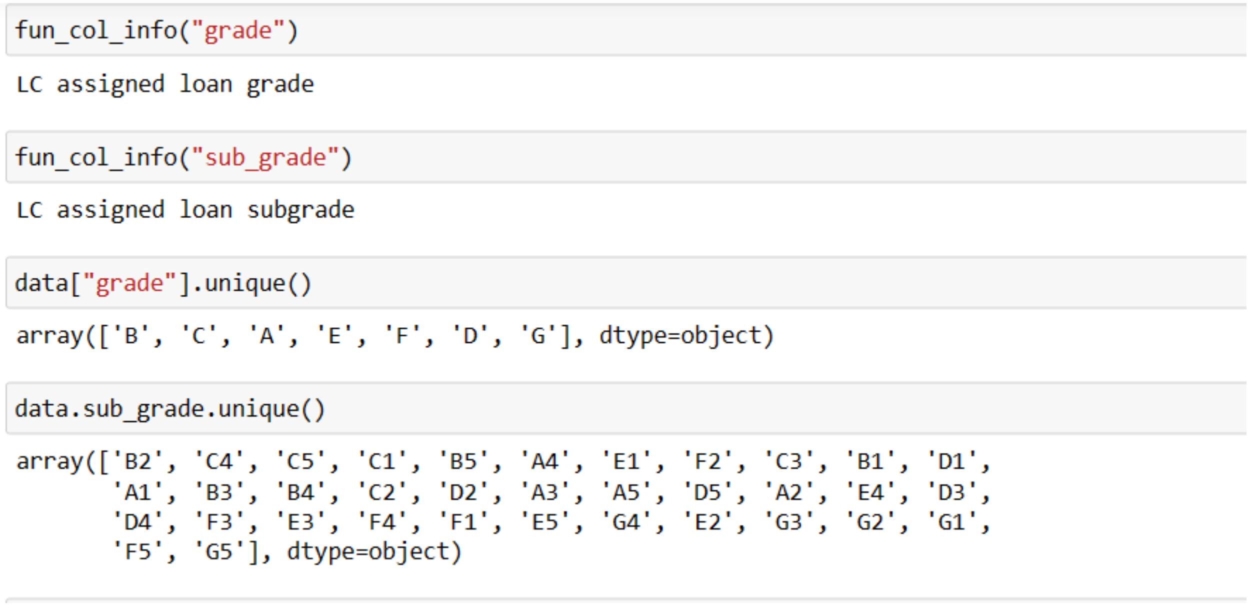
Interest Rate on the loan



Observation:

Clearly form the above rugplot it appears that as the rate of interest on loan increases there is a likelihood that the person with interest rate above 12.5% might default.

Grade and Subgrade



```
fun_col_info("grade")
```

```
LC assigned loan grade
```

```
fun_col_info("sub_grade")
```

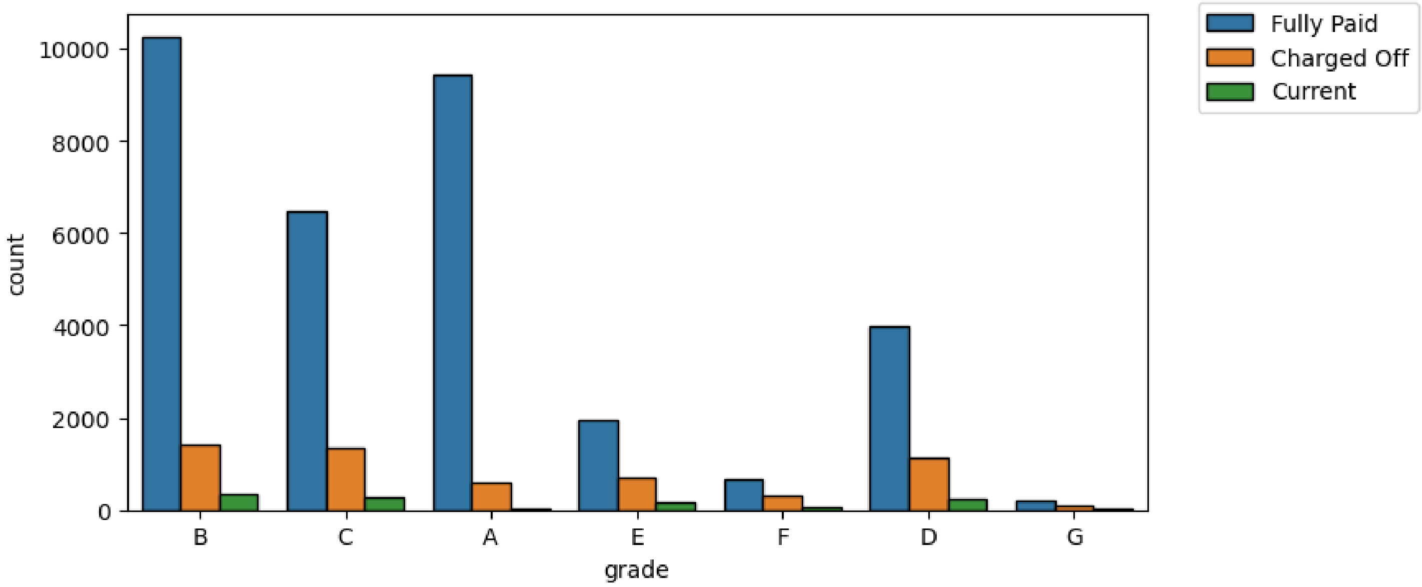
```
LC assigned loan subgrade
```

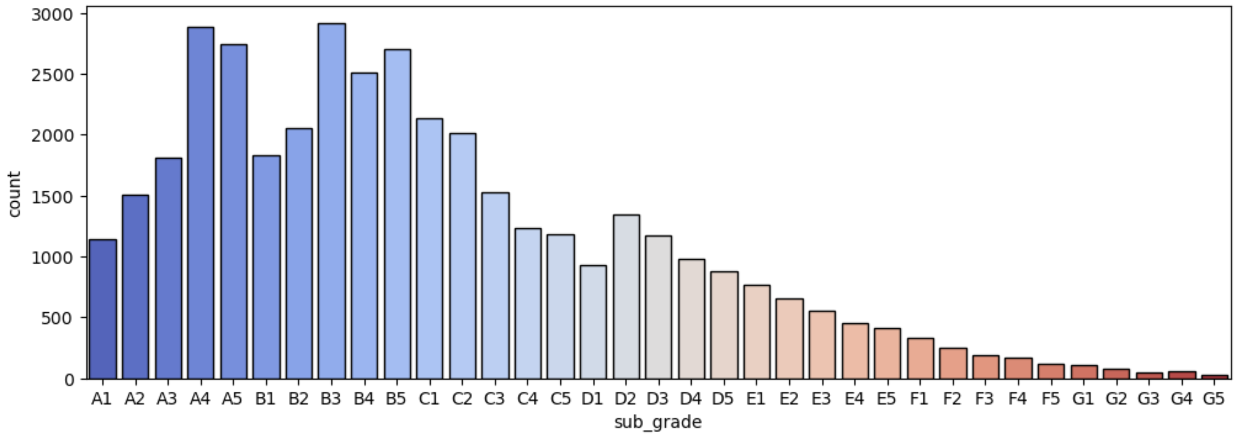
```
data["grade"].unique()
```

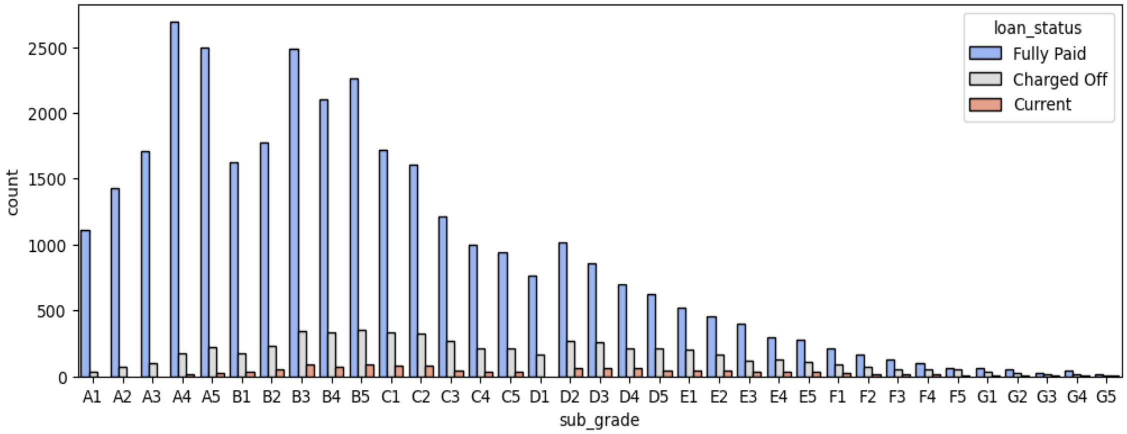
```
array(['B', 'C', 'A', 'E', 'F', 'D', 'G'], dtype=object)
```

```
data.sub_grade.unique()
```

```
array(['B2', 'C4', 'C5', 'C1', 'B5', 'A4', 'E1', 'F2', 'C3', 'B1', 'D1',
       'A1', 'B3', 'B4', 'C2', 'D2', 'A3', 'A5', 'D5', 'A2', 'E4', 'D3',
       'D4', 'F3', 'E3', 'F4', 'F1', 'E5', 'G4', 'E2', 'G3', 'G2', 'G1',
       'F5', 'G5'], dtype=object)
```



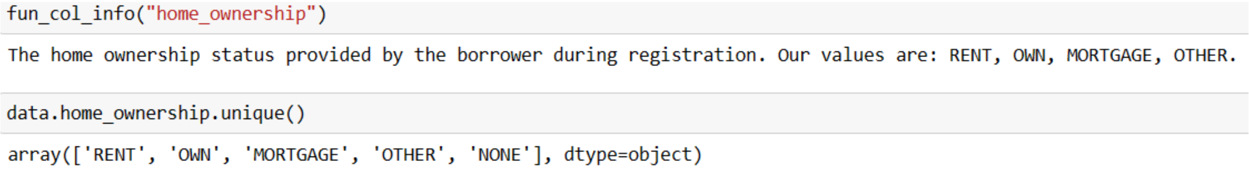




Observation:

We can see that the loan status for the F and G grade is almost the same. Which means there is a high chance that if the lending is done to someone with F and G grade then the person might default the loan.

Home Ownership

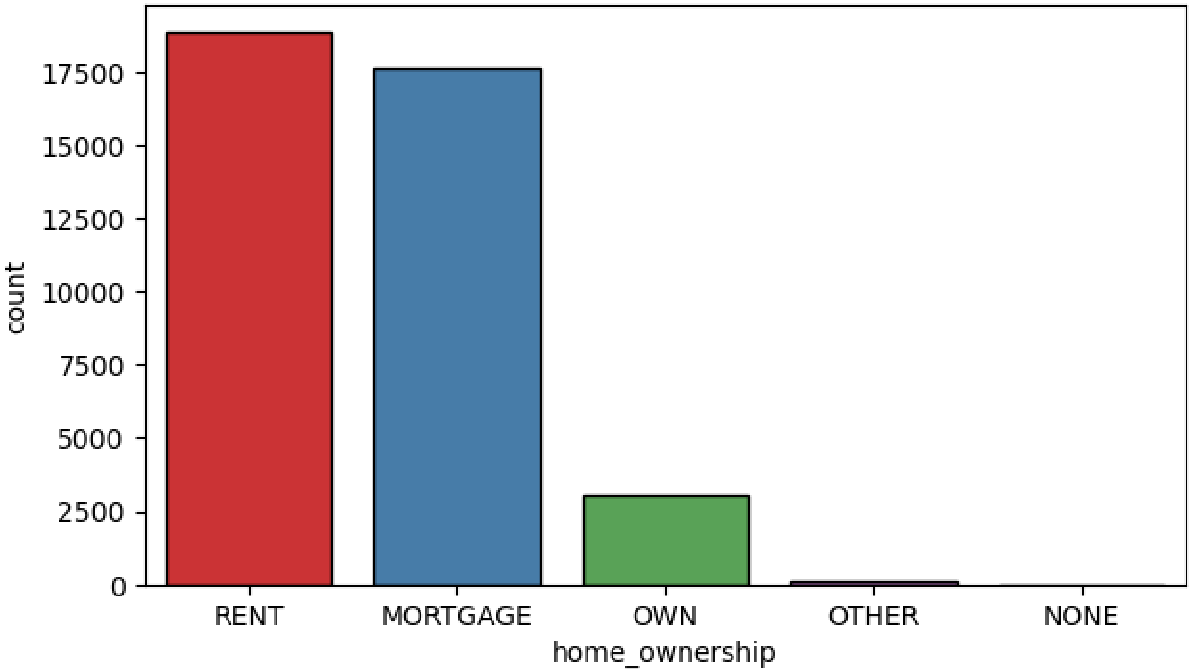


```
fun_col_info("home_ownership")
```

```
The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
```

```
data.home_ownership.unique()
```

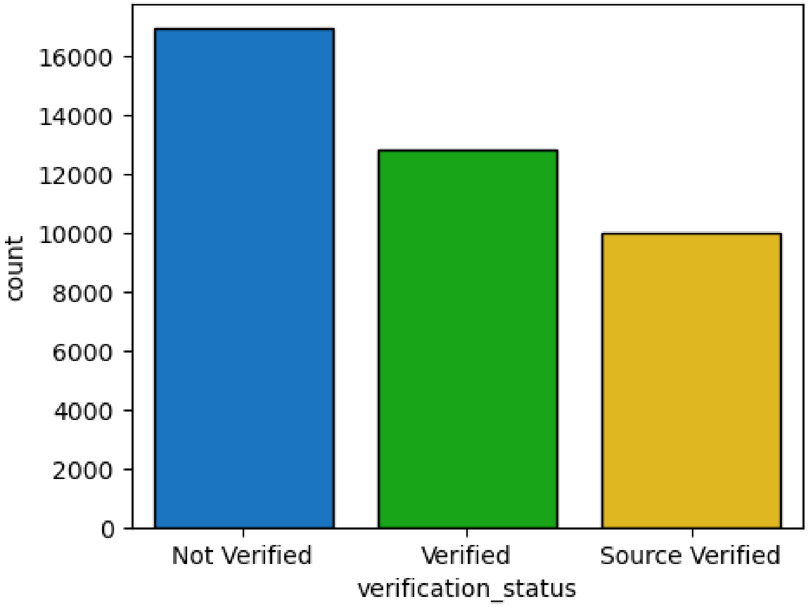
```
array(['RENT', 'OWN', 'MORTGAGE', 'OTHER', 'NONE'], dtype=object)
```



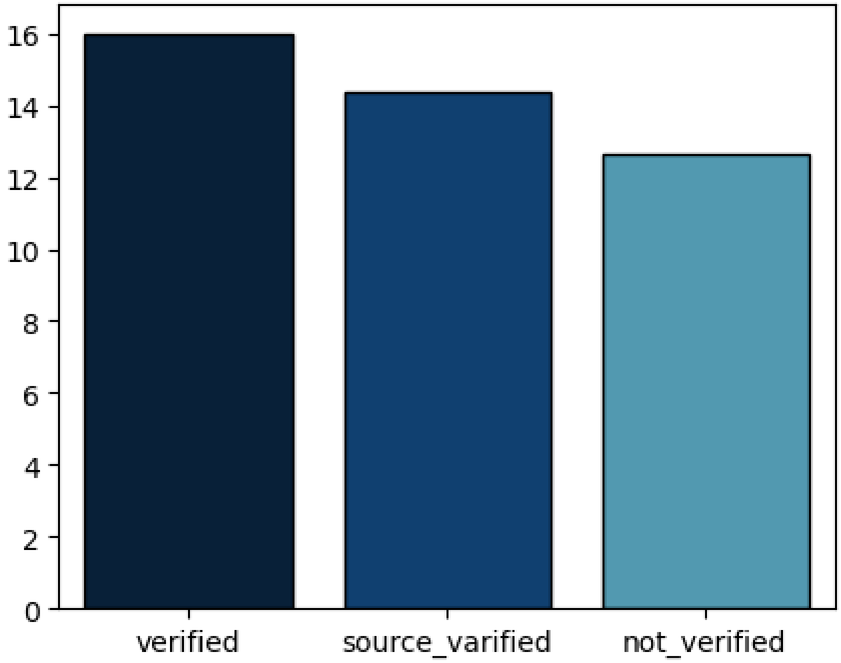
Observation:

It appears that in the given dataset there are a very few people who own house. Majority of the people have either rented the home or have mortgaged.

Verification Status



Checking Default ratio wrt Verification Status

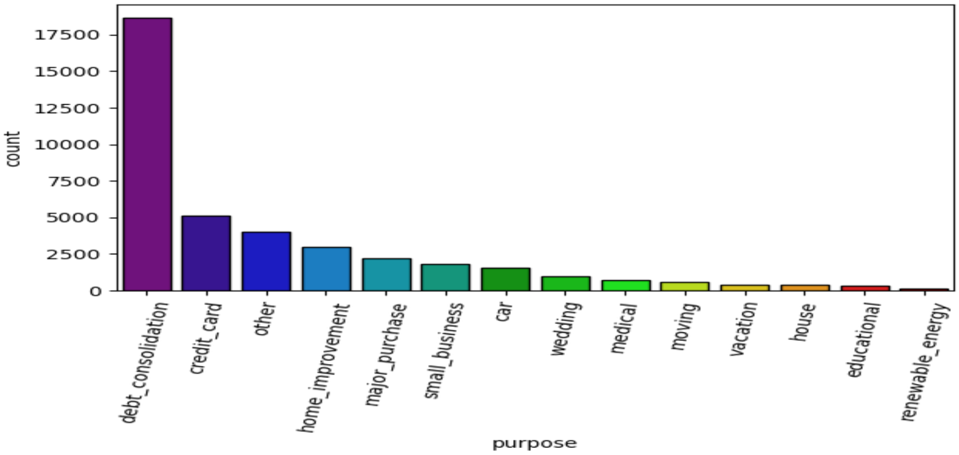


Observation:

It is strange that the people whose income was verified by either Lending Club staff or Source Vrified seems to have high default ratio than those whose income was not verified.

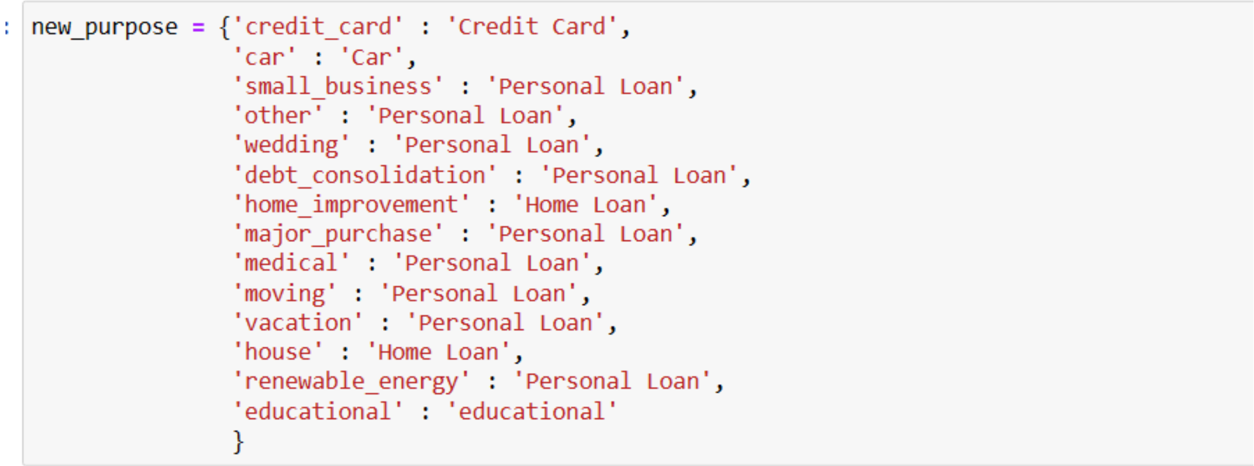
Purpose

A category provided by the borrower for the loan request.

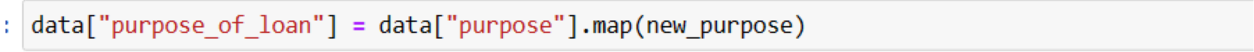


Observation:

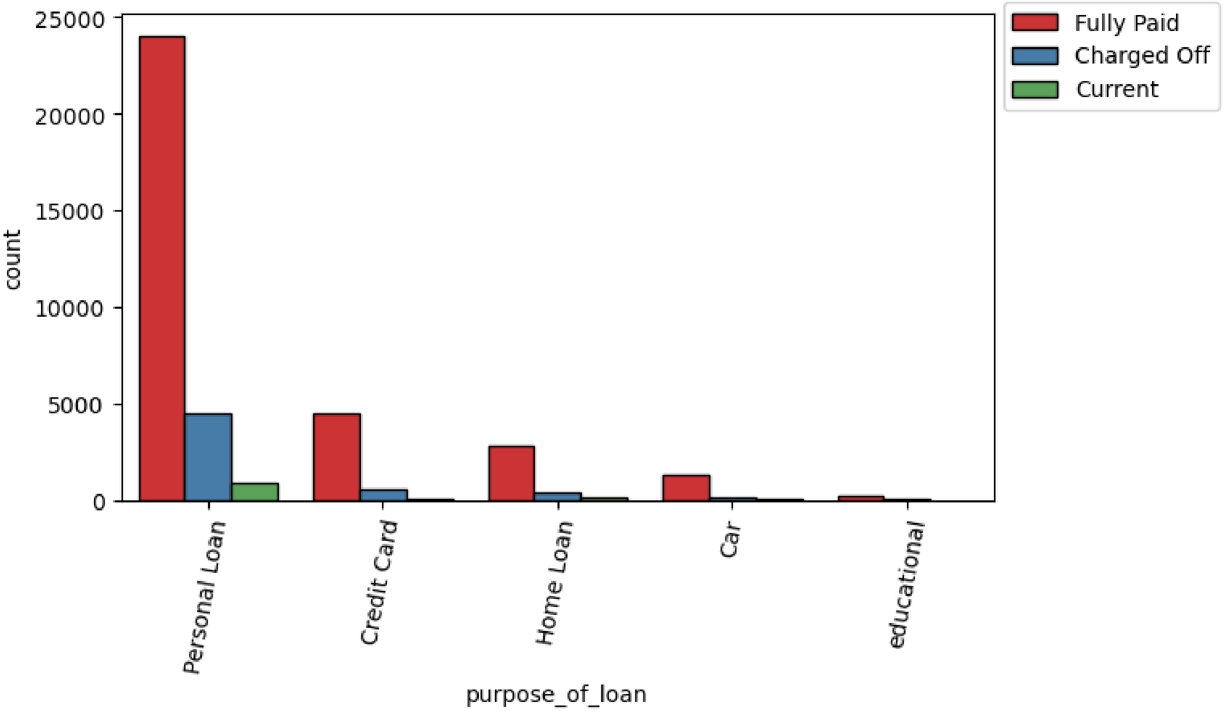
There are 14 categories under purpose of loan. But if we study each category we understand that there is no need to have these many categories because purposes like wedding, moving, medical, vacation etc can be clubbed under Personal Loan category. So keeping only 5 main categories as purpose as below.



```
: new_purpose = {'credit_card' : 'Credit Card',
                 'car' : 'Car',
                 'small_business' : 'Personal Loan',
                 'other' : 'Personal Loan',
                 'wedding' : 'Personal Loan',
                 'debt_consolidation' : 'Personal Loan',
                 'home_improvement' : 'Home Loan',
                 'major_purchase' : 'Personal Loan',
                 'medical' : 'Personal Loan',
                 'moving' : 'Personal Loan',
                 'vacation' : 'Personal Loan',
                 'house' : 'Home Loan',
                 'renewable_energy' : 'Personal Loan',
                 'educational' : 'educational'
                 }
```



```
: data["purpose_of_loan"] = data["purpose"].map(new_purpose)
```

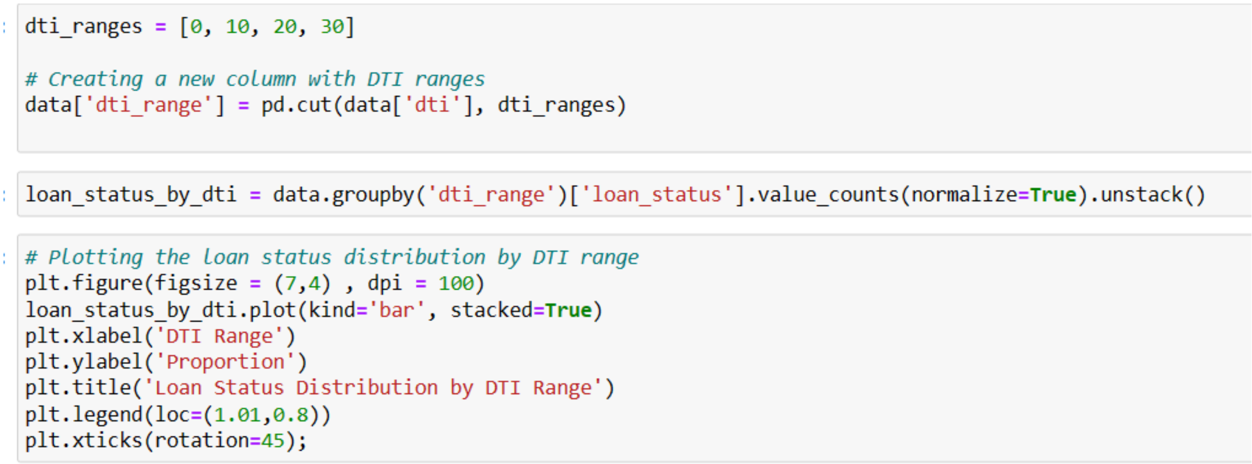


Observation:

The Lending Club appear to advance more advance more personal loans especially for Debt consolidation. Debt consolidation is a liability created to clear other liabilities. If the person is creating liabilities to clear of other previous liabilities then it can be inferred that the person is spending way more than his income. And hence at some point of time the persons' Tangible Net Worth to Outside Liability ratio might increase and this may lead to person committing default. And this clearly appears from the above plot. Hence, Lending Club should focus on advancing more secured loans like housing loans where the house is mortaged or the vehicle loan where vehicle/car is hypothecated.

Debt To Income Ratio (dti)

A ratio calculated using the borrower’s total monthly debt payments on the tota l debt obligations, excluding mortgage and the requested LC loan, divided by th e borrower’s self-reported monthly income.

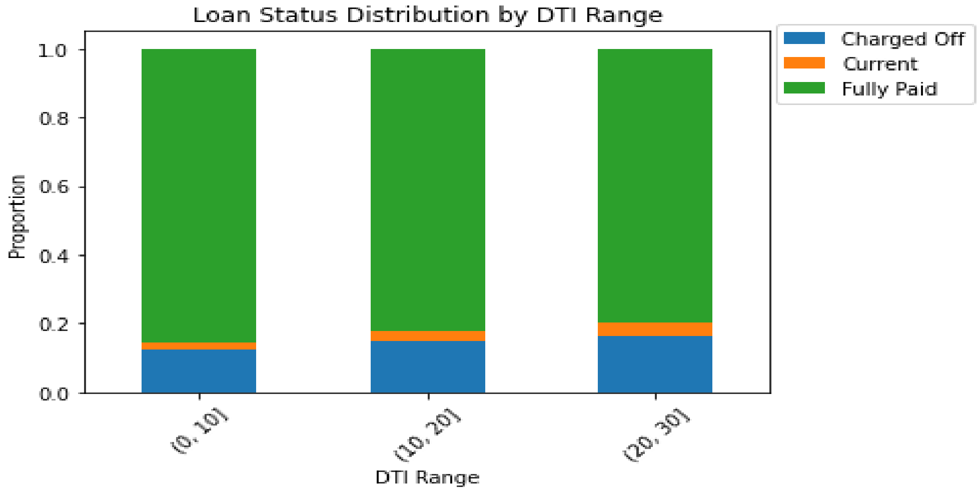


```
dti_ranges = [0, 10, 20, 30]

# Creating a new column with DTI ranges
data['dti_range'] = pd.cut(data['dti'], dti_ranges)
```

```
loan_status_by_dti = data.groupby('dti_range')['loan_status'].value_counts(normalize=True).unstack()
```

```
# Plotting the loan status distribution by DTI range
plt.figure(figsize = (7,4) , dpi = 100)
loan_status_by_dti.plot(kind='bar', stacked=True)
plt.xlabel('DTI Range')
plt.ylabel('Proportion')
plt.title('Loan Status Distribution by DTI Range')
plt.legend(loc=(1.01,0.8))
plt.xticks(rotation=45);
```

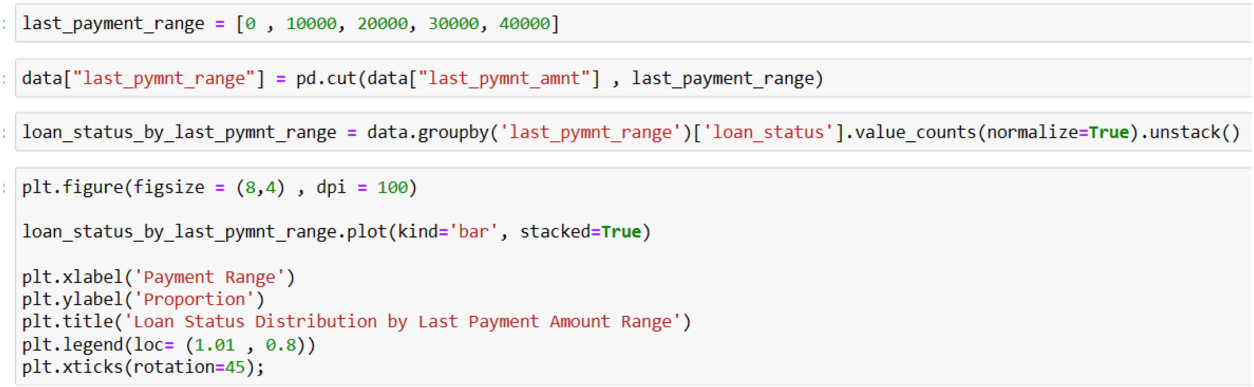


Observation:

It is clearly evident from the abobe plot that as the debt to income ratio increases the rate of loan default increases. We can corelate this with people borrowing the loan for debt consolidation as commented above.

Last Payment Amount

```
Last total payment amount received
```



```
last_payment_range = [0 , 10000, 20000, 30000, 40000]
```

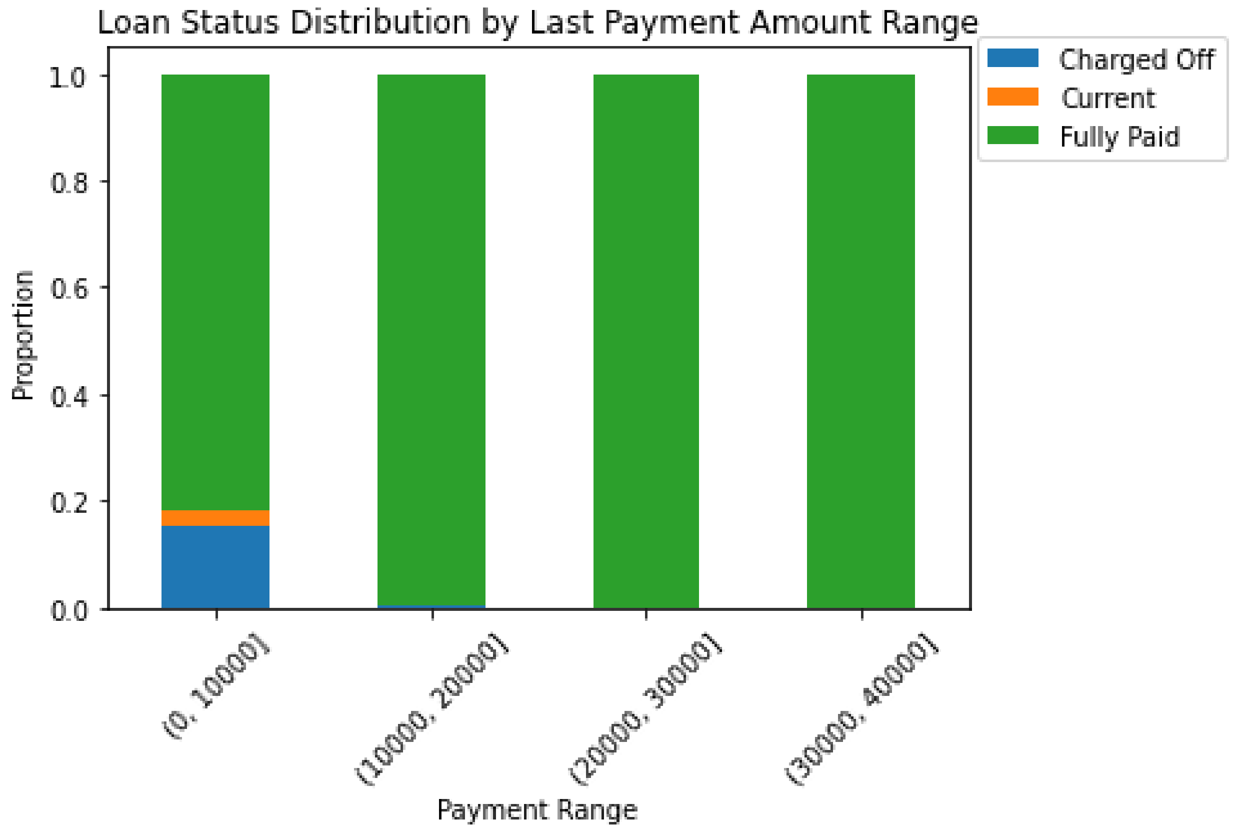
```
data["last_pymnt_range"] = pd.cut(data["last_pymnt_amnt"] , last_payment_range)
```

```
loan_status_by_last_pymnt_range = data.groupby('last_pymnt_range')['loan_status'].value_counts(normalize=True).unstack()
```

```
plt.figure(figsize = (8,4) , dpi = 100)

loan_status_by_last_pymnt_range.plot(kind='bar', stacked=True)

plt.xlabel('Payment Range')
plt.ylabel('Proportion')
plt.title('Loan Status Distribution by Last Payment Amount Range')
plt.legend(loc= (1.01 , 0.8))
plt.xticks(rotation=45);
```

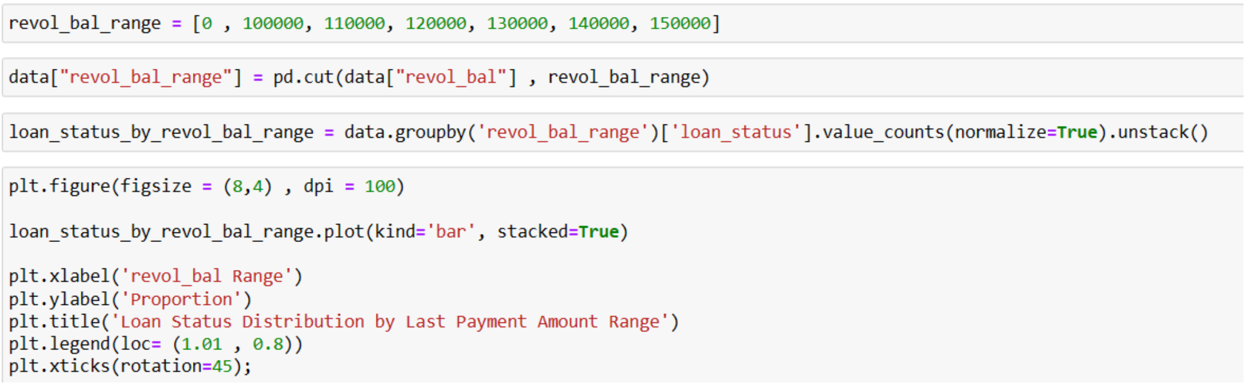


Observation:

Clearly from the above figure it is evident that if the last payment made by the borrower is more than 10,000 dollars then likelihood of default is negligible.

Revolving Balance

Total credit revolving balance



```
revol_bal_range = [0 , 100000, 110000, 120000, 130000, 140000, 150000]
```

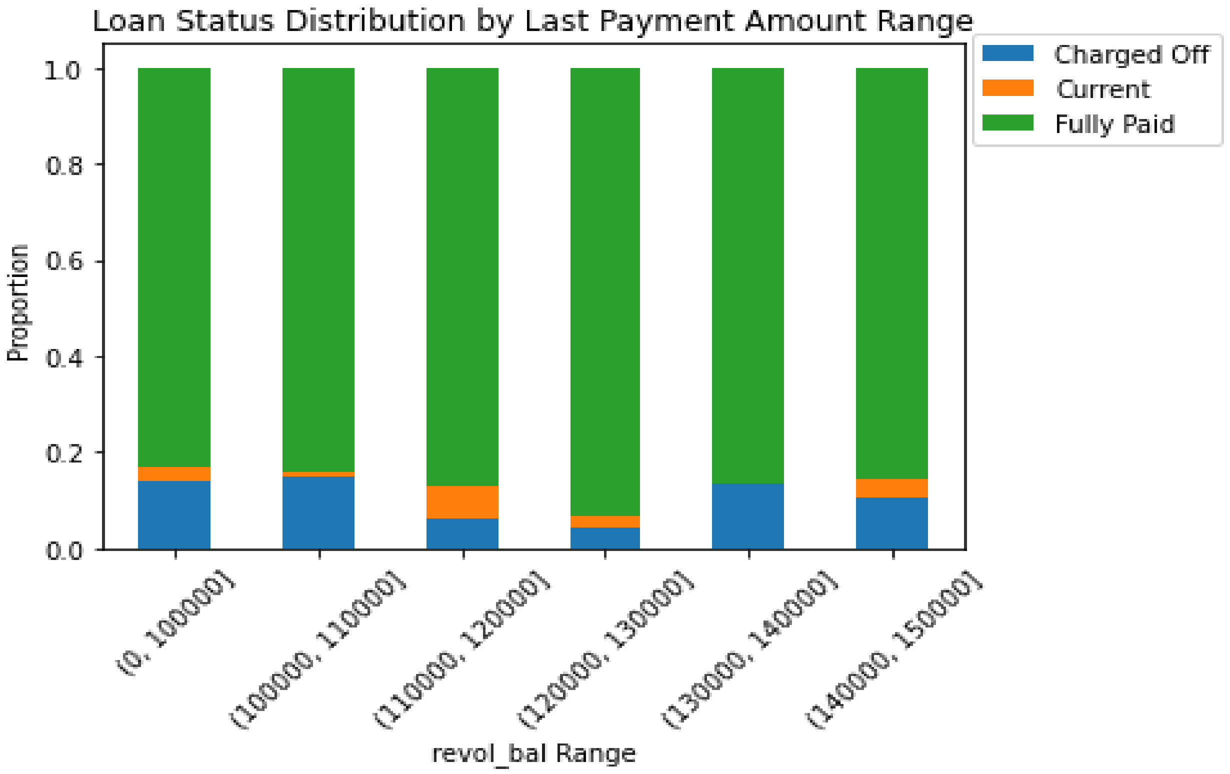
```
data["revol_bal_range"] = pd.cut(data["revol_bal"] , revol_bal_range)
```

```
loan_status_by_revol_bal_range = data.groupby('revol_bal_range')['loan_status'].value_counts(normalize=True).unstack()
```

```
plt.figure(figsize = (8,4) , dpi = 100)

loan_status_by_revol_bal_range.plot(kind='bar', stacked=True)

plt.xlabel('revol_bal Range')
plt.ylabel('Proportion')
plt.title('Loan Status Distribution by Last Payment Amount Range')
plt.legend(loc= (1.01 , 0.8))
plt.xticks(rotation=45);
```

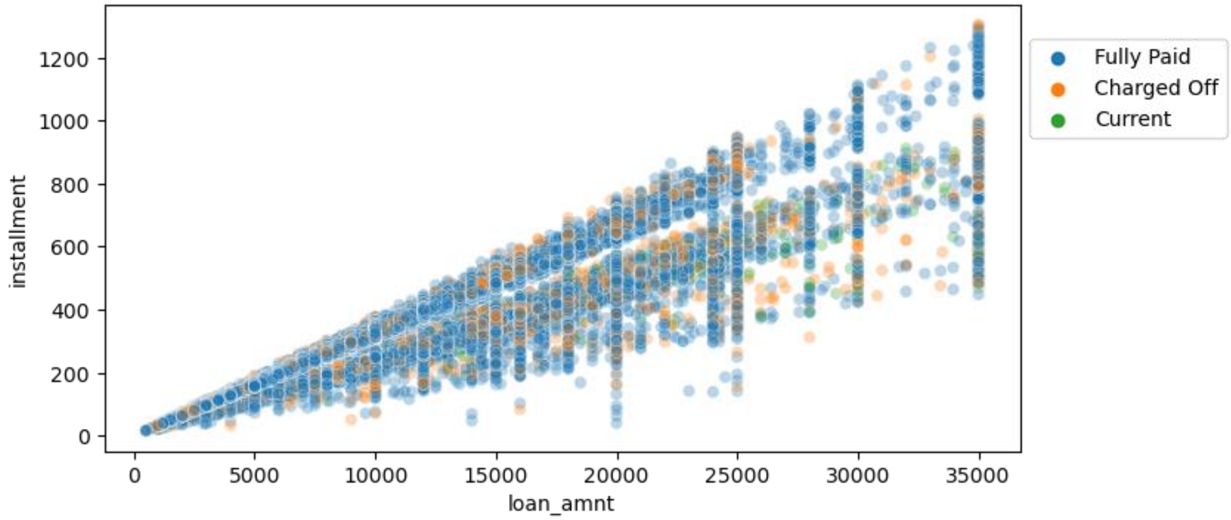


Observation:

If the revolving balance is in the range of 110000 to 130000 dollars then the likelyhood of default is less.

Bivariate Analysis

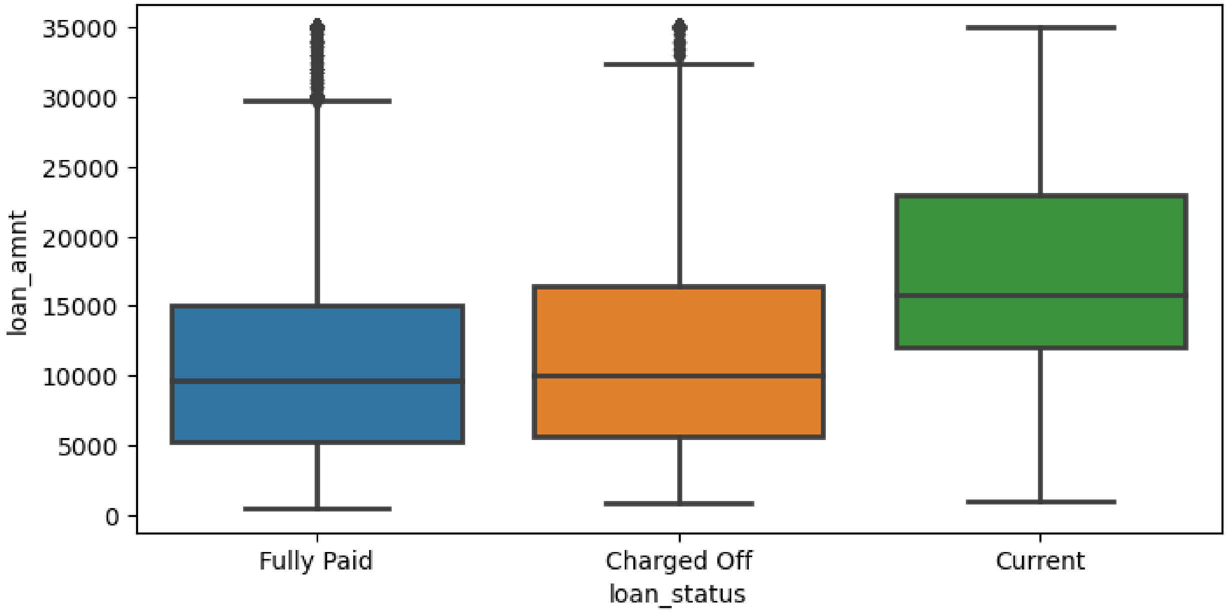
Loan Amount vs Installment



Observation:

The loan installment is proportional to loan amount. Further, the likelihood of default seem to increase when the loan amount tis greater then 15000 dollars.

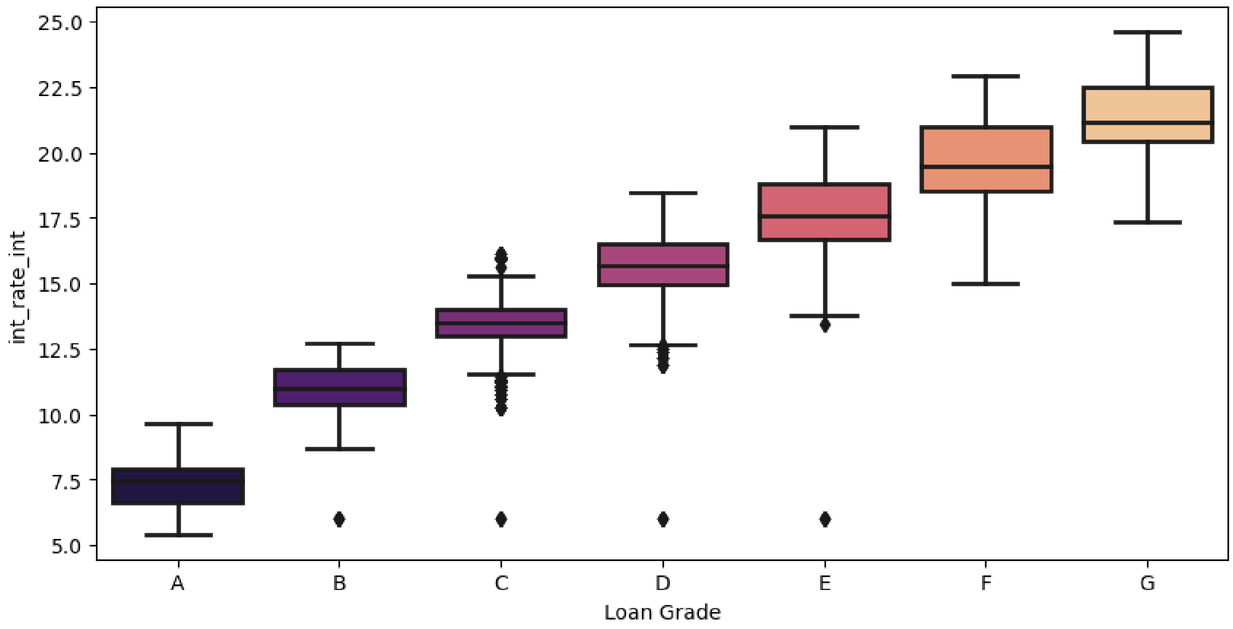
Loan Status vs Loan Amount



Observation:

It can be inferred from the boxplot that mean of loan amount of the people who didnt repay the loan is higher.

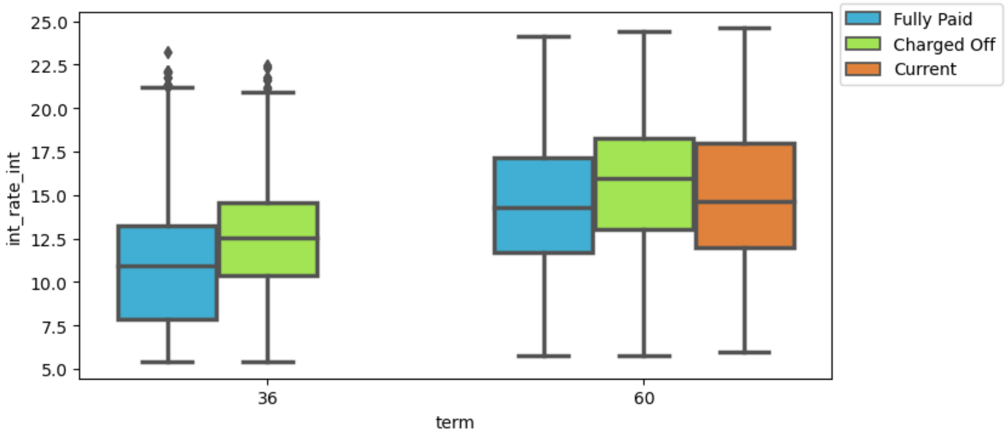
Grades vs Loan Interest



Observation:

Clearly, as the grades start to worsen the interest rate goes up. It is a good technique to manage the credit risk

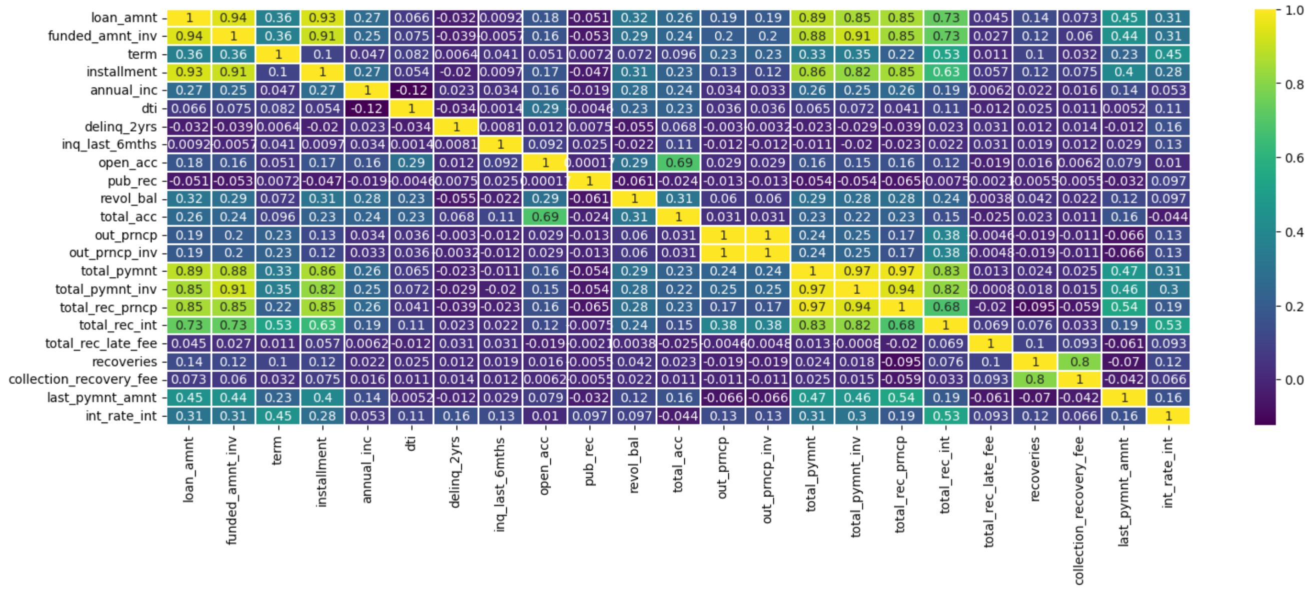
Term vs Interest Rate



Observation:

The loans with 60 Months seem to have higher interest values and within that those loans which were charged off seem to have higher mean interest values.

Correlation of between variables



	loan_amnt	funded_amnt_inv	term	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	total_acc	out_prncp	out_prncp_inv	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	collection_recovery_fee	last_pymnt_amnt	int_rate_int
loan_amnt	1	0.94	0.36	0.93	0.27	0.066	-0.032	0.0092	0.18	-0.051	0.32	0.26	0.19	0.19	0.89	0.85	0.85	0.73	0.045	0.14	0.073	0.45	0.31
funded_amnt_inv	0.94	1	0.36	0.91	0.25	0.075	-0.039	-0.0057	0.16	-0.053	0.29	0.24	0.2	0.2	0.88	0.91	0.85	0.73	0.027	0.12	0.06	0.44	0.31
term	0.36	0.36	1	0.1	0.047	0.082	0.0064	0.041	0.051	0.0072	0.072	0.096	0.23	0.23	0.33	0.35	0.22	0.53	0.011	0.1	0.032	0.23	0.45
installment	0.93	0.91	0.1	1	0.27	0.054	-0.02	0.0097	0.17	-0.047	0.31	0.23	0.13	0.12	0.86	0.82	0.85	0.63	0.057	0.12	0.075	0.4	0.28
annual_inc	0.27	0.25	0.047	0.27	1	-0.12	0.023	0.034	0.16	-0.019	0.28	0.24	0.034	0.033	0.26	0.25	0.26	0.19	0.0062	0.022	0.016	0.14	0.053
dti	0.066	0.075	0.082	0.054	-0.12	1	-0.034	0.0014	0.29	-0.0046	0.23	0.23	0.036	0.036	0.065	0.072	0.041	0.11	-0.012	0.025	0.011	0.0052	0.11
delinq_2yrs	-0.032	-0.039	0.0064	-0.02	0.023	-0.034	1	0.0081	0.012	0.0075	-0.055	0.068	-0.003	-0.0032	-0.023	-0.029	-0.039	0.023	0.031	0.012	0.014	-0.012	0.16
inq_last_6mths	0.0092	-0.0057	0.041	0.0097	0.034	0.0014	0.0081	1	0.092	0.025	-0.022	0.11	-0.012	-0.012	-0.011	-0.02	-0.023	0.022	0.031	0.019	0.012	0.029	0.13
open_acc	0.18	0.16	0.051	0.17	0.16	0.29	0.012	0.092	1	0.00017	0.29	0.69	0.029	0.029	0.16	0.15	0.16	0.12	-0.019	0.016	0.0062	0.079	0.01
pub_rec	-0.051	-0.053	0.0072	-0.047	-0.019	-0.0046	0.0075	0.025	0.00017	1	-0.061	-0.024	-0.013	-0.013	-0.054	-0.054	-0.065	-0.0075	-0.0021	-0.0055	-0.0055	-0.032	0.097
revol_bal	0.32	0.29	0.072	0.31	0.28	0.23	-0.055	-0.022	0.29	-0.061	1	0.31	0.06	0.06	0.29	0.28	0.28	0.24	0.0038	0.042	0.022	0.12	0.097
total_acc	0.26	0.24	0.096	0.23	0.24	0.23	0.068	0.11	0.69	-0.024	0.31	1	0.031	0.031	0.23	0.22	0.23	0.15	-0.025	0.023	0.011	0.16	-0.044
out_prncp	0.19	0.2	0.23	0.13	0.034	0.036	-0.003	-0.012	0.029	-0.013	0.06	0.031	1	1	0.24	0.25	0.17	0.38	-0.0046	-0.019	-0.011	-0.066	0.13
out_prncp_inv	0.19	0.2	0.23	0.12	0.033	0.036	-0.0032	-0.012	0.029	-0.013	0.06	0.031	1	1	0.24	0.25	0.17	0.38	-0.0048	-0.019	-0.011	-0.066	0.13
total_pymnt	0.89	0.88	0.33	0.86	0.26	0.065	-0.023	-0.011	0.16	-0.054	0.29	0.23	0.24	0.24	1	0.97	0.97	0.83	0.013	0.024	0.025	0.47	0.31
total_pymnt_inv	0.85	0.91	0.35	0.82	0.25	0.072	-0.029	-0.02	0.15	-0.054	0.28	0.22	0.25	0.25	0.97	1	0.94	0.82	-0.0008	0.018	0.015	0.46	0.3
total_rec_prncp	0.85	0.85	0.22	0.85	0.26	0.041	-0.039	-0.023	0.16	-0.065	0.28	0.23	0.17	0.17	0.97	0.94	1	0.68	-0.02	-0.095	-0.059	0.54	0.19
total_rec_int	0.73	0.73	0.53	0.63	0.19	0.11	0.023	0.022	0.12	-0.0075	0.24	0.15	0.38	0.38	0.83	0.82	0.68	1	0.069	0.076	0.033	0.19	0.53
total_rec_late_fee	0.045	0.027	0.011	0.057	0.0062	-0.012	0.031	0.031	-0.019	-0.0021	0.0038	-0.025	-0.0046	0.0048	0.013	-0.0008	-0.02	0.069	1	0.1	0.093	-0.061	0.093
recoveries	0.14	0.12	0.1	0.12	0.022	0.025	0.012	0.019	0.016	-0.0055	0.042	0.023	-0.019	-0.019	0.024	0.018	-0.095	0.076	0.1	1	0.8	-0.07	0.12
collection_recovery_fee	0.073	0.06	0.032	0.075	0.016	0.011	0.014	0.012	0.0062	-0.0055	0.022	0.011	-0.011	-0.011	0.025	0.015	-0.059	0.033	0.093	0.8	1	-0.042	0.066
last_pymnt_amnt	0.45	0.44	0.23	0.4	0.14	0.0052	-0.012	0.029	0.079	-0.032	0.12	0.16	-0.066	-0.066	0.47	0.46	0.54	0.19	-0.061	-0.07	-0.042	1	0.16
int_rate_int	0.31	0.31	0.45	0.28	0.053	0.11	0.16	0.13	0.01	0.097	0.097	-0.044	0.13	0.13	0.31	0.3	0.19	0.53	0.093	0.12	0.066	0.16	1

Observations from heatmap:

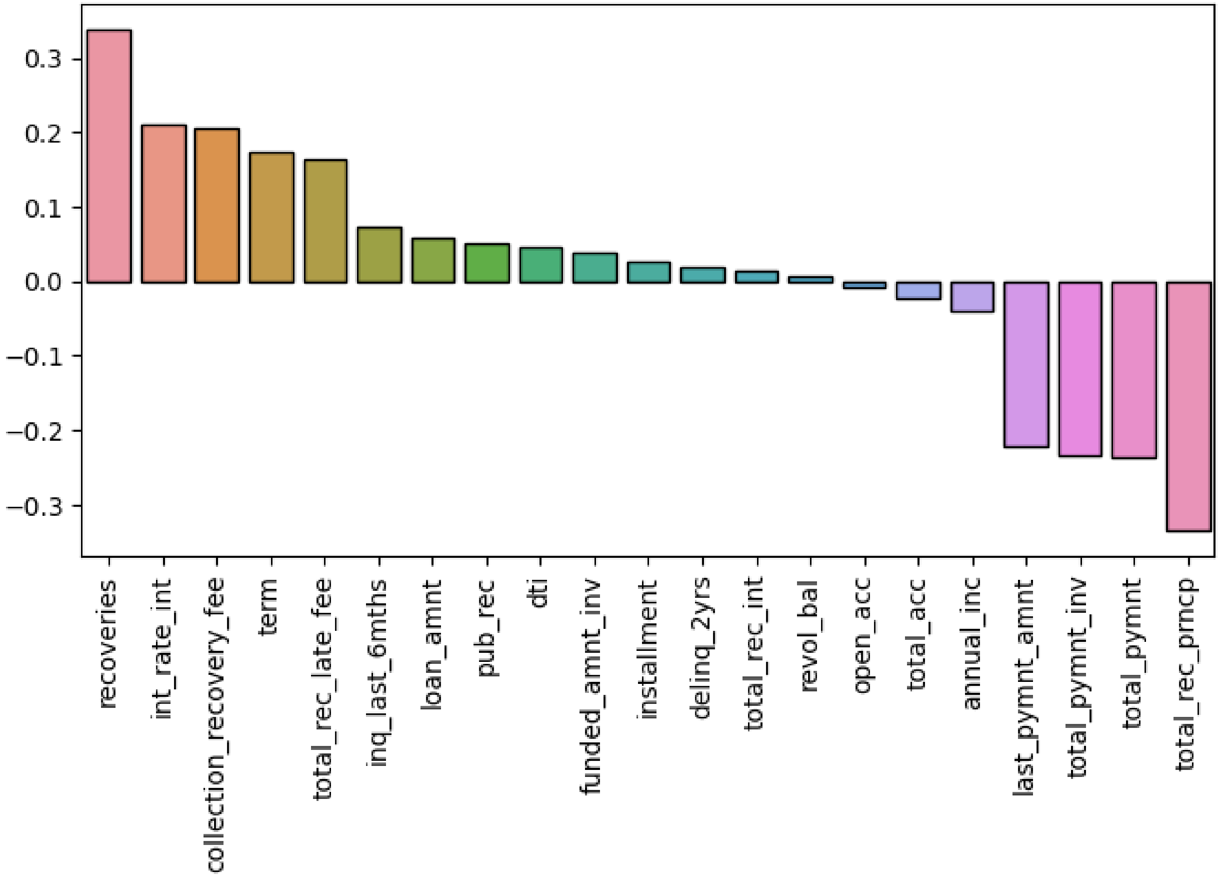
1) There is almost perfect relationship between loan amount, funded amount, installment and total payment.

2) Total payment has almost perfect relationship between total recovered principle.

3) There is perfect relationship between outstanding principle and remaining outstanding principal for portion of total amount funded by investors.

All the above mentioned columns may cause multicollinearity issue while training the machine learning model.

Finding The Correlation Of Independent Variables With Target Variable (Loan_Status_Encoded)



Observation:

1) The columns like recoveries, interest rate, recovery fee, term etc have strong positive correlation with loan status column which is the target column.

2) The columns like total recovered principle, total payment etc have negative correlation with our target variable.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have supported and contributed to the successful completion of this data analysis project. Firstly, I would like to thank the financial institution for providing access to the dataset and resources necessary for conducting this analysis. I am grateful for the guidance and mentorship provided by Upgrad Mentors. Their expertise and insights were invaluable in shaping the direction of the analysis. Additionally, I would like to thank my fellow colleagues and classmates who provided assistance

and constructive feedback during various stages of the project. Finally, I extend my appreciation to the developers and contributors of the open-source tools and libraries that were utilized in this project. Their efforts have made data analysis more accessible and efficient.

This project would not have been possible without the collective support and collaboration of all these individuals and organizations.

Contact

Created by:

Swapnil Meshram

https://github.com/swapnilmeshram1404