greatlearning

Guidelines for PGPDSE FT Capstone Project – Interim Report

Project Group Info:

BATCH DETAILS	DSE online January 2022
TEAM MEMBERS	• Sree Ganesha Chellappa • Faishal Sharif • Hasitha vaddi • Nikhil Nivalagi • Ashish Shukla.
DOMAIN OF PROJECT	Predictive Analysis in Finance
PROJECT TITLE	Loan Default Prediction
GROUP NUMBER	GROUP – 4
TEAM LEADER	Hasitha vaddi
MENTOR NAME	Mrs Vidhya Kannaiah

Abstract:

The bank Indessa has not been performing well for the last three quarters and would like to improve their bank's performance by reducing their NPA's Since the data which is collected is very messy identification of defaulters who are the major cause which increases the NPA's has become difficult. so we are devising a model which can predict whether the person who is a loan applicant or so already has a loan will be a potential defaulter in the future

Objectives:

Bank Indessa face challenges in identifying the loan applicants of those who can be potential defaulter. Any discrepancies in data faced by the employees against a set of standards can indicate that the loan the applicant can be a potential defaulter and can mark the applicants for any differences against some set standards if they have crossed the set number of marking (parameters which have been found) they can be identified as a potential defaulter and the approval of the loan can be rejected.

The objective of this Capstone Project is to predict the optimal parameters for the identification of potential defaulters of a loan to meet the stockholder confidence in their company and to carry out any future increase of NPA's (non-performing assets) and increase the company's share in the market.

Industry Review:

Commercial lending options provide flexible long-term lending, which is a major commercial lending market driver.

In addition, payment collection collaborations between digital lending organizations and FinTech companies are expected to grow in the market

The commercial lending market size was valued at USD 8,823.53 Billion in 2020 and is projected to reach USD 29,379.83 Billion by 2030, growing at a CAGR of 13.1% from 2021 to 2030.

Many business owners had to take out commercial loans to keep their businesses afloat as COVID-19 cases continued to rise and more restrictions were

imposed during the pandemic. This resulted in a surge in commercial lending market growth.

Commercial lending offers the lowest interest rates on all loan options, allowing business owners to get needed funds while keeping overhead costs low. Borrowers who choose fixed monthly repayments can use them accurately in their business planning and forecasting, allowing them to structure their business finance with a bit more certainty.

Furthermore, commercial lending payment plans are typically for several years, allowing a company to focus on other important business matters such as sales, overhead management, and employee training. As a result, this is a significant driving force in the commercial lending market.

Problem statement

Bank Indessa has not done well in the last 3 quarters. Their NPAs (Non-Performing Assets) has reached an all-time high. It is starting to lose the confidence of its investors. As a result, its stock has fallen by 20% in the previous quarter alone.

After careful analysis, it was found that the majority of NPA was contributed by loan defaulters. With the messy data collected over the years, this bank has decided to use machine learning to figure out a way to find these defaulters and devise a plan to reduce them.

This bank uses a pool of investors to sanction their loans.

We will help this bank by predicting the probability that a member will default.

Project Outcome:

By implementing the resultant models built using the above methods, we can suggest the ideal standards or parameters required on which the loan defaulters can be identified in a short duration to reduce the increase in NPAs and improve efficiency.

Dataset and Domain:

Data Dictionary:

Real-time bank dataset obtained from an organization.

The indessa.xlsx This data set comprises information captured in December 2016.

The dataset has 5,32,428 records and 45 attributes

The attribute/feature/column names are given below:

Index(['member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'batch_enrolled', 'int_rate', 'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'pymnt_plan', 'desc', 'purpose', 'title', 'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'mths_since_last_delinq', 'mths_since_last_record', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'initial_list_status', 'total_rec_int','total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'collections_12_mths_ex_med', 'mths_since_last_major_derog',

'application_type', 'verification_status_joint', 'last_week_pay', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim', 'loan_status'])

There are 27 numerical and 18 object dtypes

dtype - numerical - ['member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'mths_since_last_delinq', 'mths_since_last_record', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'collections_12_mths_ex_med', 'mths_since_last_major_derog', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim', 'loan_status']

dtype - object -['term', 'batch_enrolled', 'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'verification_status', 'pymnt_plan', 'desc', 'purpose', 'title', 'zip_code', 'addr_state', 'initial_list_status', 'application_type', 'verification_status_joint', 'last_week_pay']

Variable	Description	Dtype
member_id	unique ID assigned to each member	int64
loan_amnt	loan amount ($) applied by the member	int64
funded_amnt	loan amount ($) sanctioned by the bank	int64
funded_amnt_inv	loan amount ($) sanctioned by the investors	float64
term	term of loan (in months)	object
batch_enrolled	batch numbers allotted to members	object

Variable	Description	Dtype
int_rate	interest rate (%) on loan	float64
grade	grade assigned by the bank	object
sub_grade	grade assigned by the bank	object
emp_title	job / Employer title of member	object
emp_length	employment length, where 0 means less than one year and 10 means ten or more years	object
home_ownership	status of home ownership	object
annual_inc	annual income ($) reported by the member	float64
verification_status	status of income verified by the bank	object
pymnt_plan	indicates if any payment plan has started against loan	object
desc	loan description provided by member	object
purpose	purpose of loan	object

Variable	Description	Dtype
title	loan title provided by member	object
zip_code	first three digits of area zip code of member	object
addr_state	living state of member	object
dti	ratio of member's total monthly debt repayment excluding mortgage divided by self reported monthly income	float64
delinq_2yrs	number of 30+ days delinquency in past 2 years	float64
inq_last_6mths	number of inquiries in last 6 months	float64
mths_since_last_delinq	number of months since last delinq	float64
mths_since_last_record	number of months since last public record	float64
open_acc	number of open credit line in member's credit line	float64
pub_rec	number of derogatory public records	float64

Variable	Description	Dtype
revol_bal	total credit revolving balance	float64
revol_util	amount of credit a member is using relative to revol_bal	float64
total_acc	total number of credit lines available in members credit line	float64
initial_list_status	unique listing status of the loan - W(Waiting), F(Forwarded)	object
total_rec_int	interest received till date	float64
total_rec_late_fee	Late fee received till date	float64
recoveries	post charge off gross recovery	float64
collection_recovery_fee	post charge off collection fee	float64
collections_12_mths_ex_med	number of collections in last 12 months excluding medical collections	float64
mths_since_last_major_derog	months since most recent 90 day or worse rating	float64

Variable	Description	Dtype
application_type	indicates when the member is an individual or joint	object
verification_status_joint	indicates if the joint members income was verified by the bank	object
last_week_pay	indicates how long (in weeks) a member has paid EMI after batch enrolled	object
acc_now_delinq	number of accounts on which the member is delinquent	float64
tot_coll_amt	total collection amount ever owed	float64
tot_cur_bal	total current balance of all accounts	float64
total_rev_hi_lim	total revolving credit limit	float64
loan_status	status of loan amount, 1 = Defaulter, 0 = Non Defaulters	int64

Nulls

BEFORE

member_id	0
loan_amnt	0
funded_amnt	0
funded_amnt_inv	0

term	0
batch_enrolled	85149
int_rate	0
grade	0
sub_grade	0
emp_title	30833
emp_length	26891
home_ownership	0
annual_inc	3
verification_status	0
pymnt_plan	0
desc	456829
purpose	0
title	90
zip_code	0
addr_state	0
dti	0
delinq_2yrs	16
inq_last_6mths	16
mths_since_last_delinq	272554
mths_since_last_record	450305
open_acc	16
pub_rec	16
revol_bal	0
revol_util	287
total_acc	16
initial_list_status	0
total_rec_int	0
total_rec_late_fee	0
recoveries	0
collection_recovery_fee	0
collections_12_mths_ex_med	95
mths_since_last_major_derog	399448
application_type	0
verification_status_joint	532123
last_week_pay	0
acc_now_delinq	16
tot_coll_amt	42004
tot_cur_bal	42004
total_rev_hi_lim	42004
loan_status	0

AFTER

loan_amnt	0
funded_amnt	0
funded_amnt_inv	0
term	0
int_rate	0
emp_length	0
annual_inc	0
verification_status	0
dti	0
delinq_2yrs	0
inq_last_6mths	0
mths_since_last_delinq	0
open_acc	0
pub_rec	0
revol_bal	0
revol_util	0
total_acc	0
initial_list_status	0
total_rec_int	0
total_rec_late_fee	0
recoveries	0
collection_recovery_fee	0
collections_12_mths_ex_med	0
last_week_pay	0
acc_now_delinq	0
tot_coll_amt	0
tot_cur_bal	0
total_rev_hi_lim	0
loan_status	0
grade_num	0
addr_state_NE	0
addr_state_S	0
addr_state_W	0

Univariate analysis:

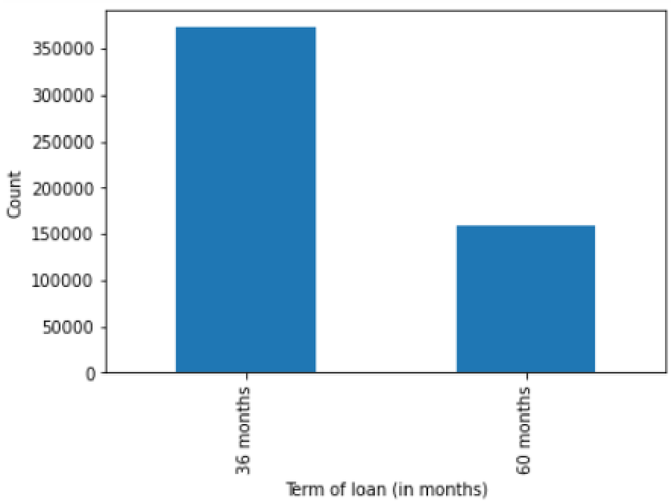
Univariate Analysis:

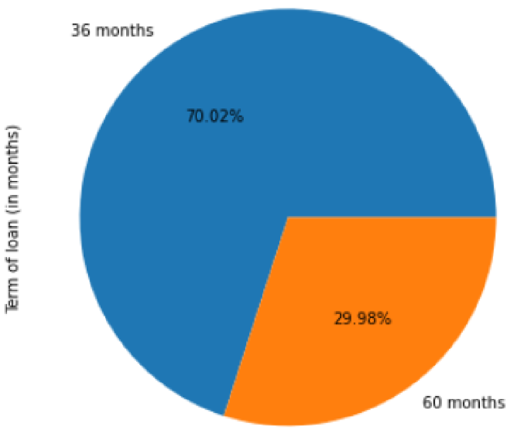
Analysis of Categorical Variables:

From the below plots, each of the categorical variables is analysed individually.

Term of Loan Attribute:

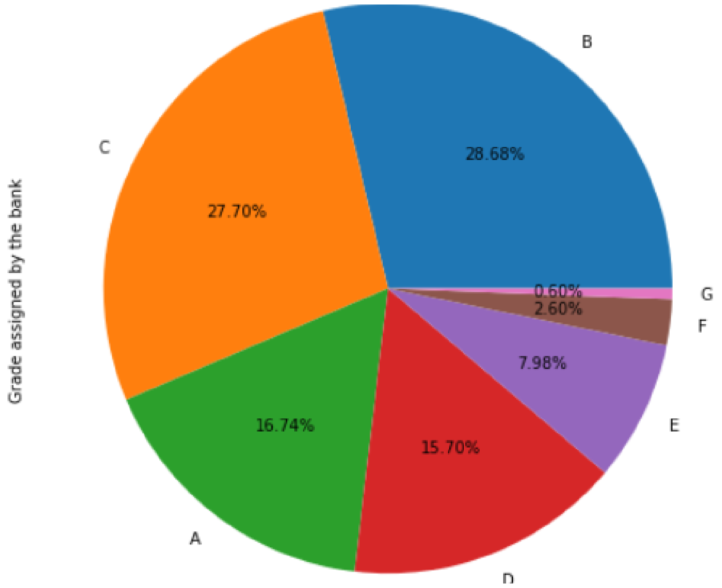
From the above plot, we can observe that 70% of members need the loan for the duration of 36 months and the remaining 30% of members for 60 months.





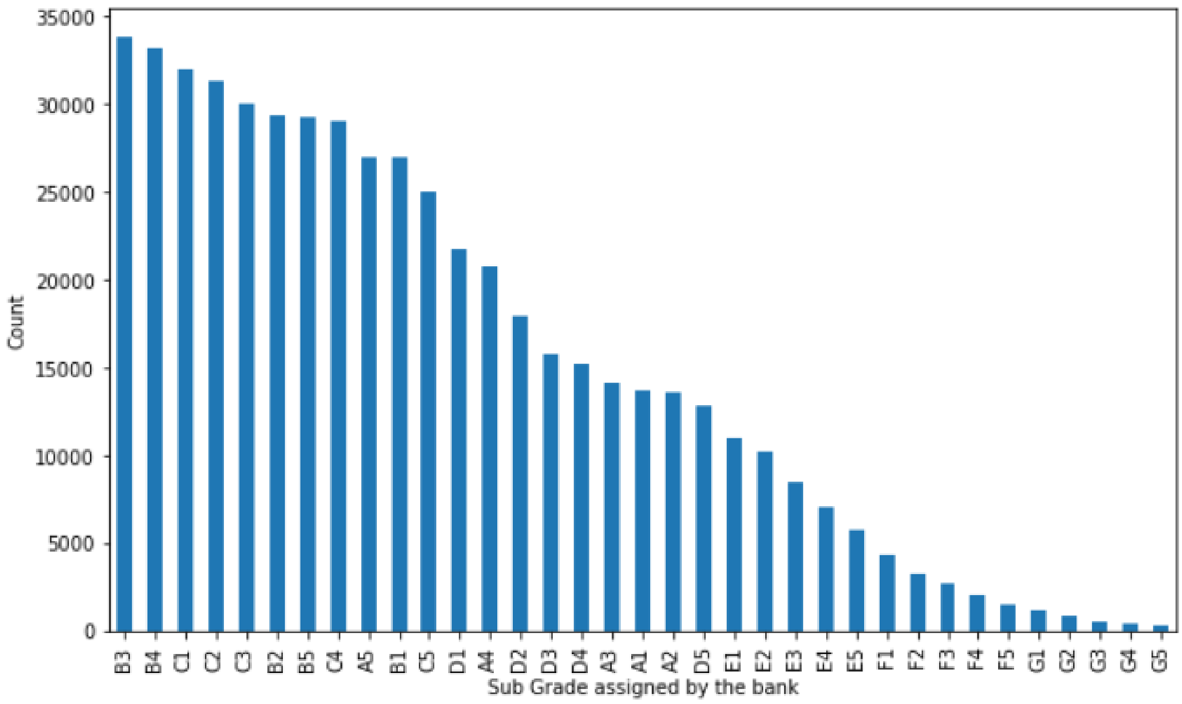
Grade Attribute:

From the above plot, we can observe that the highest no. of members belongs to Grade B and the lowest no. of members from Grade G. Also the no. of members from Grade A and D is approximately the same.

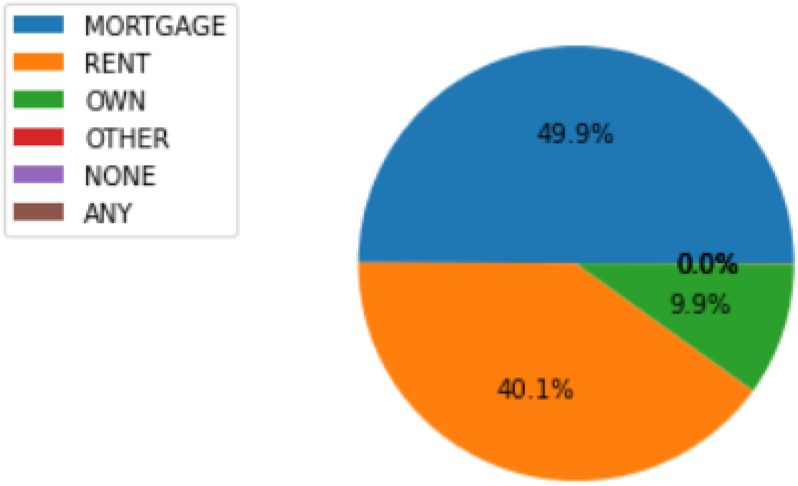


Sub Grade Attribute:

From the above plot, we can observe that the maximum number of members are assigned in Sub Grade B3 and the lowest members are assigned in Sub Grade G5. More than 30000 members are assigned in these 5 Subgrades B3, B4, C1, C2, and C3.

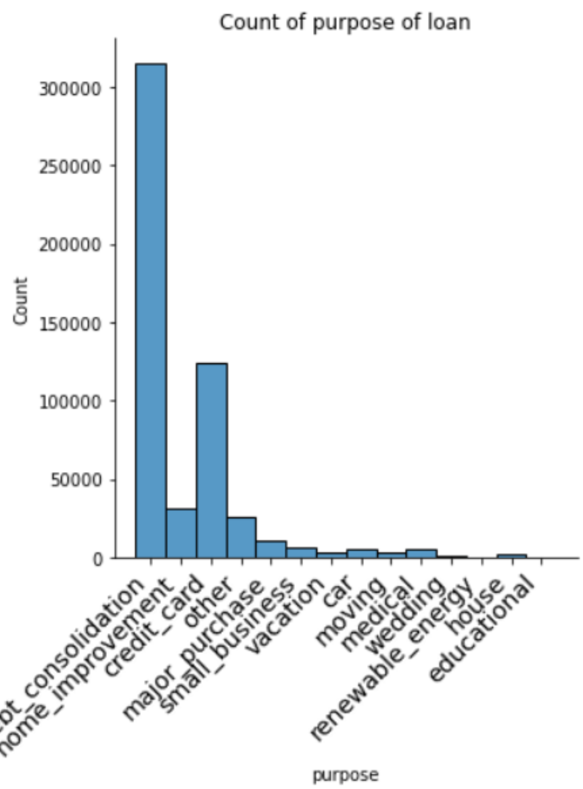


Home Ownership Attribute:



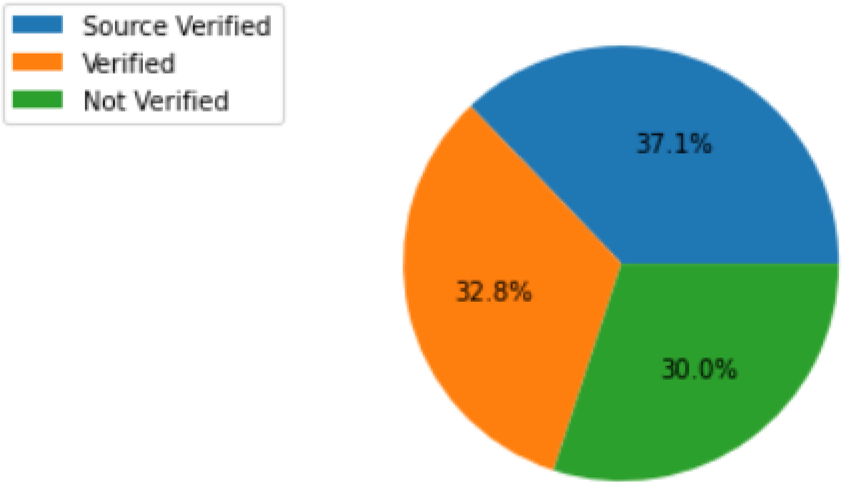
From the above plot, we can observe that almost 50% of home ownership status is of the ‘Mortgage’ type. However, mortgage, rent, and own totally comprise the 99.99% home ownership status of the total members.

Purpose Attribute:



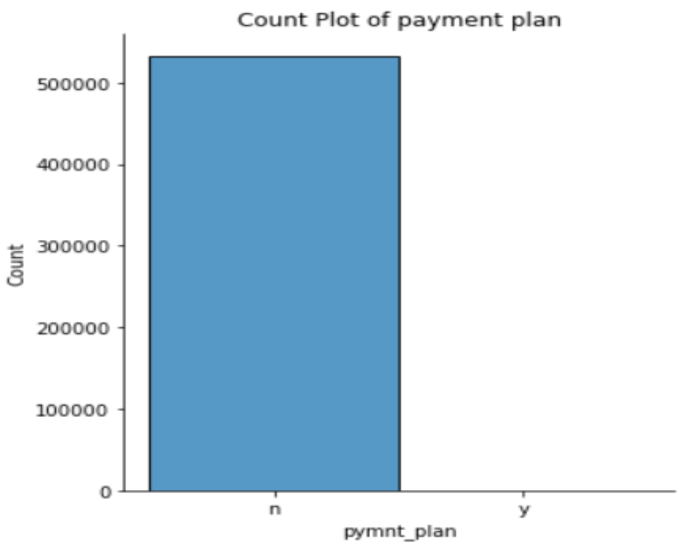
- From the above count plot, it can be observed that the maximum number of members (60%) have the purpose of the loan as 'debt_consolidation'.
- More than 1lakh employees have put their purpose of the loan as 'credit card'.

Verification Status Attribute:



- From the above pie plot, it can be observed that the maximum status of income verified by the bank is ‘source verified’.However not verified members are also in significant numbers. All together all the three categories contribute significantly as all three have members present in good numbers.

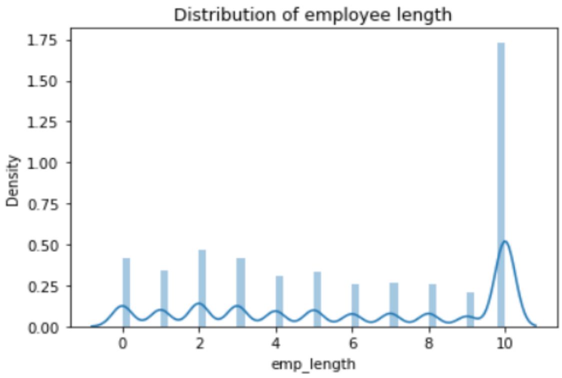
Payment Plan attribute:

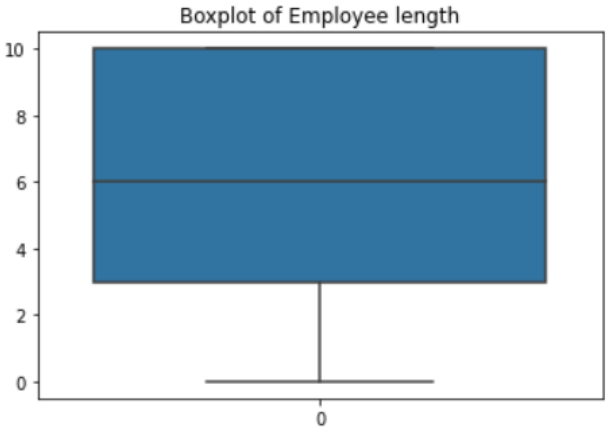


From the above count plot, we can observe that almost all the members don't have any payment plan started against the loan.

last_week_pay: too many categories to use

Employee Length Attribute:

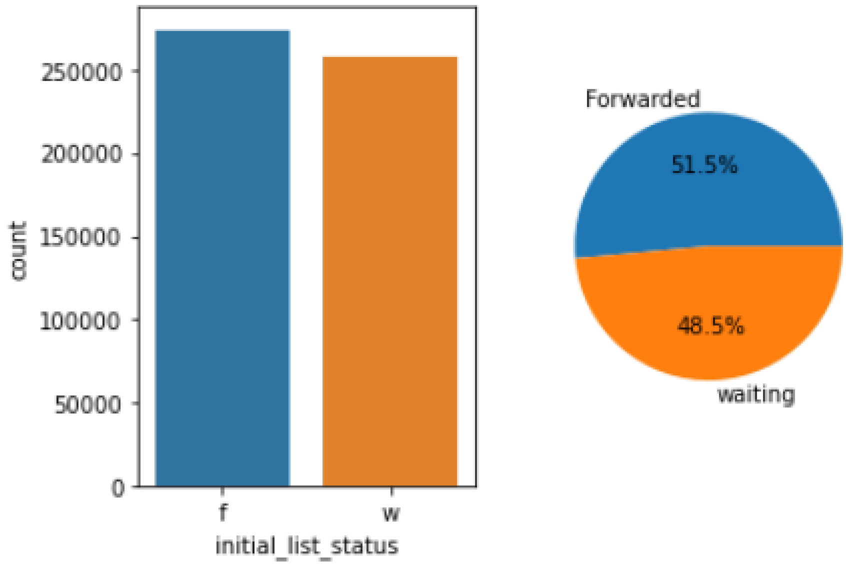




- From the above distribution plot we can assume that the distribution is highest at 10 years. That means the max members have an employment length of 10 years.
- Similarly, the above boxplot tells us that the mean employment length of 6 years approximately.
- Majority of the members have employment length in the range of 3 years to 10 years.

initial_list_status:

unique listing status of the loan - W(Waiting), F(Forwarded)



#forwarded loan status() occupies 51.5 per cent of the total data count

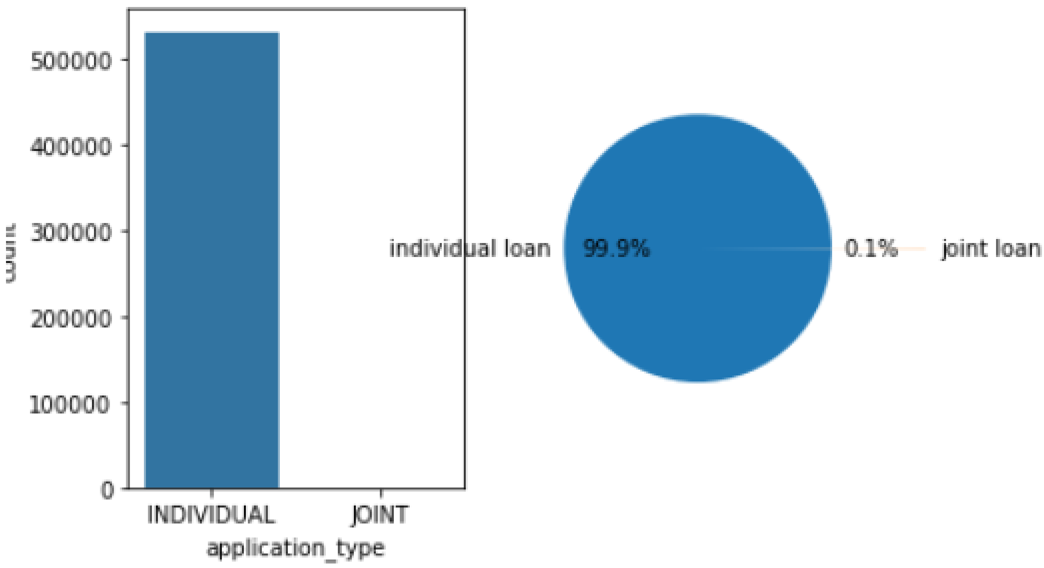
#waiting loan status() occupies 48.5 per cent of the total data count

#unique listing status of the loan - W(Waiting), F(Forwarded)

forwarded loan status() (274018)is slightly greater than waiting for status ()

application_type:

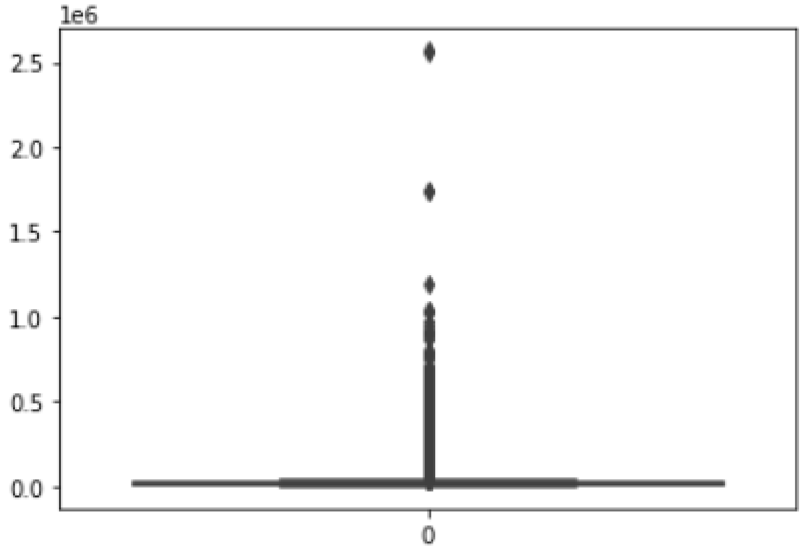
indicates when the member is an individual or joint

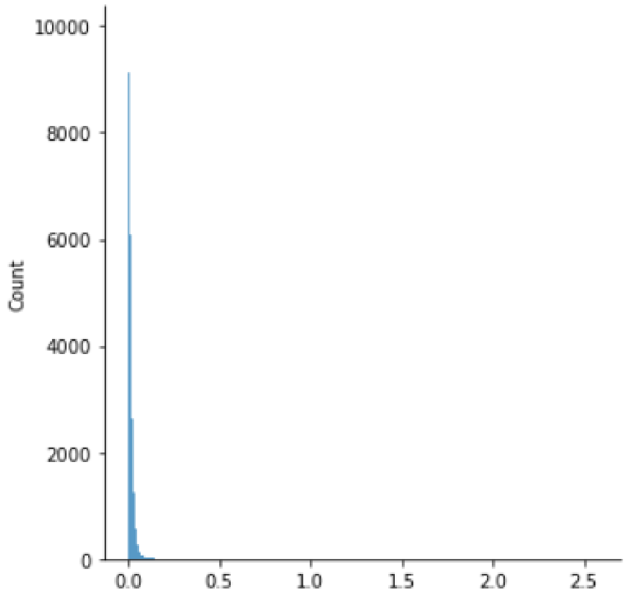


#these joint loan takers hold a minority of 0.1 per cent and the rest is 99.9 per cent

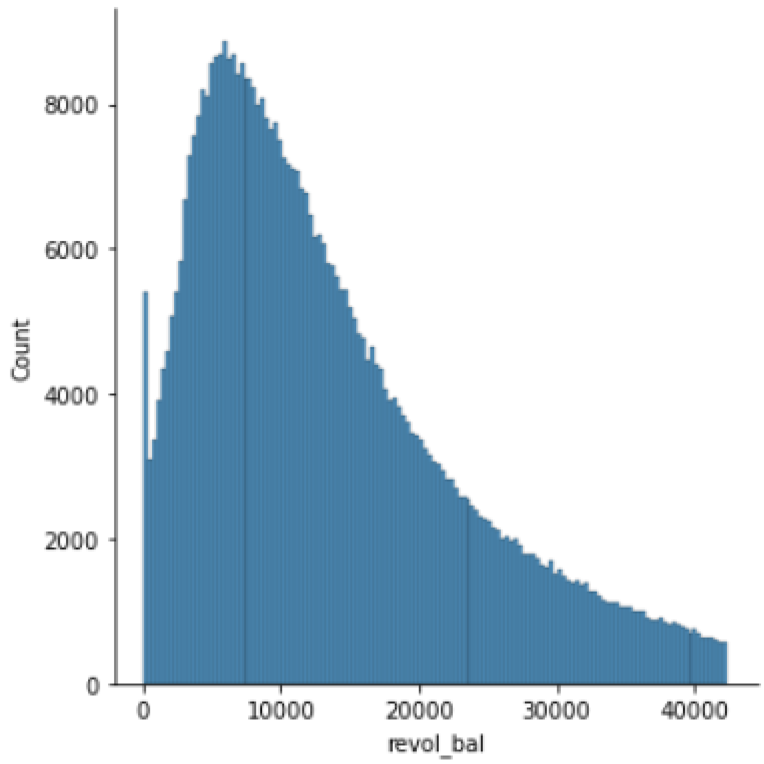
<u># Numerical data</u>

revol_bal: If you don't pay the balance on your revolving credit account in full every month, the unpaid portion carries over to the next month.





before



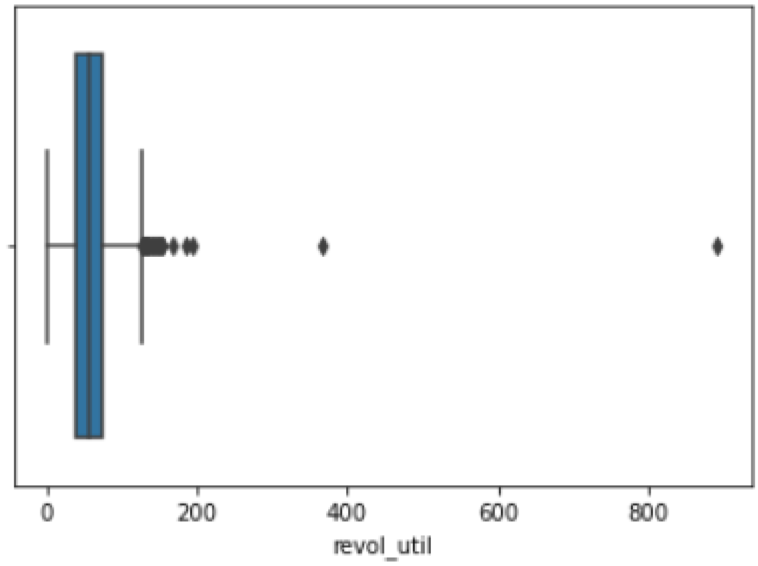
The data has an extreme outlier of value of 2.5*10^5 which affect the mean, and other outliers

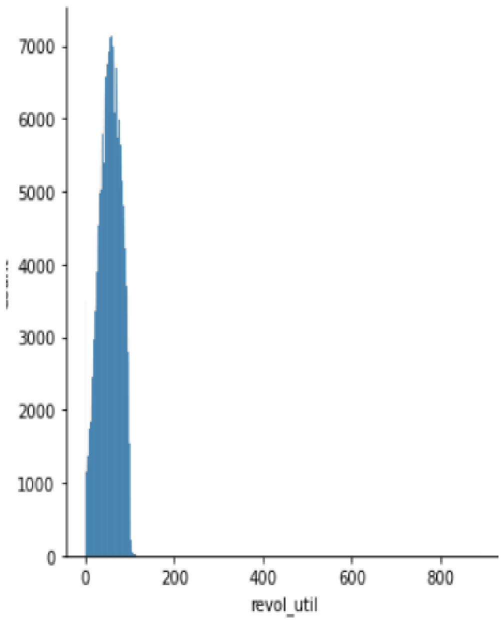
but the majority of the data has a total credit revolving balance in the range of 0 to 40000

the data is skewed towards the right

revol_util:

the amount of credit a member is using relative to revol_bal



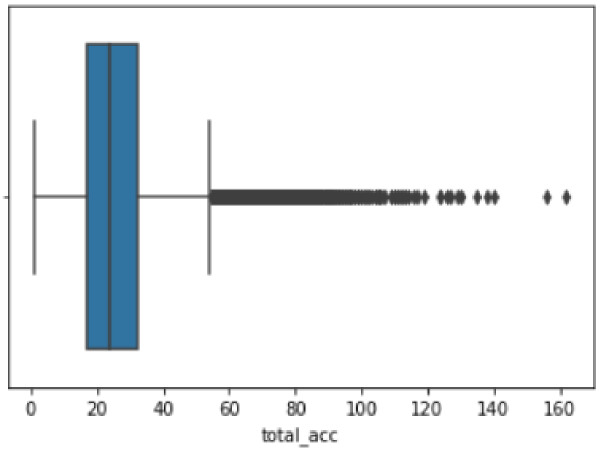


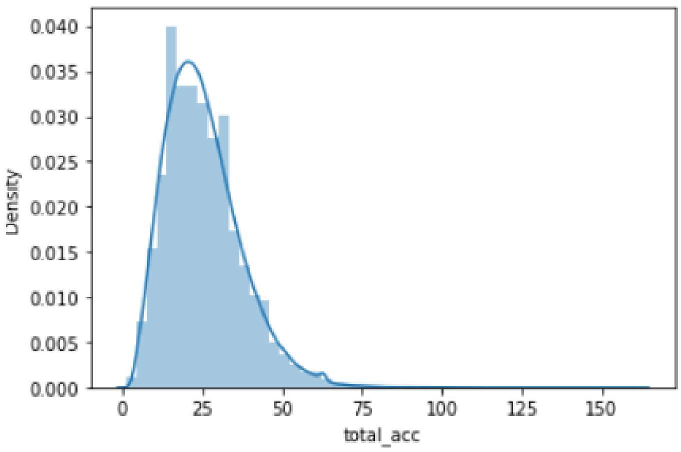
After removal of the outliers the data is slightly skewed, but majority of the credit usages lies between 0 to 100k where the

Mean credit usage is 65k

total_acc:

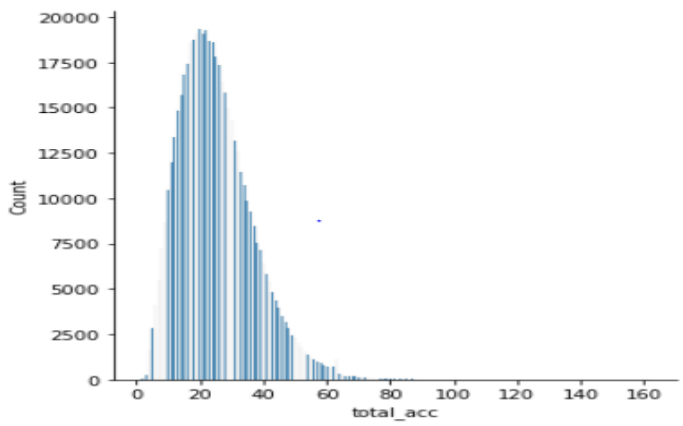
total number of credit lines available in members credit line

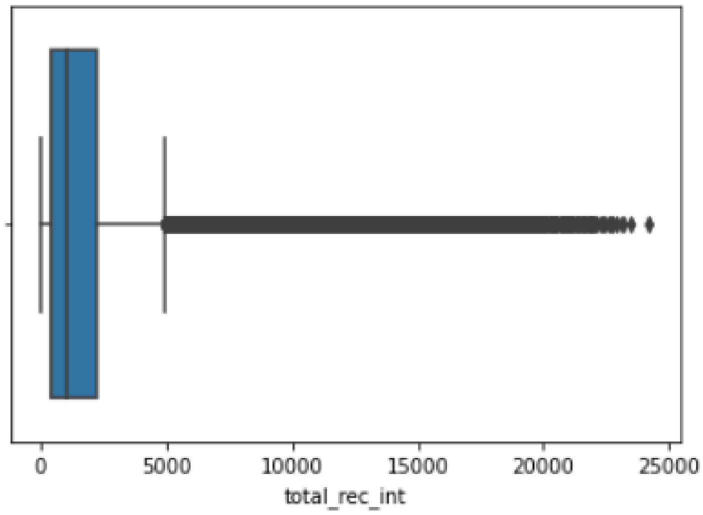




Since the data indicates the number of total credit lines available for in members credit line the range of credit lines available lie between 1 to 162 but the maximum number lies between 1-60 where the median is around 24, the data lies majorly between 0 to 50(IQR) the data also appears to be skewed to the right

total_rec_int: Interest received till date





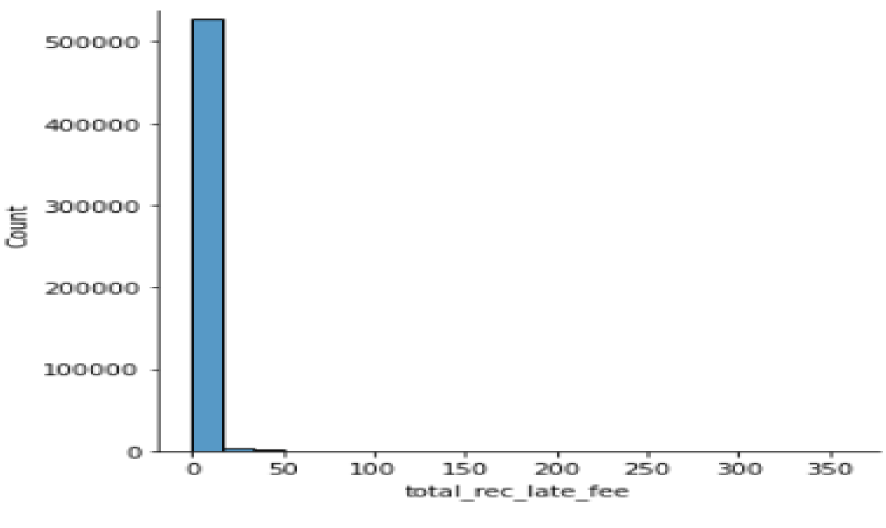
The data also appears to be highly skewed to the right

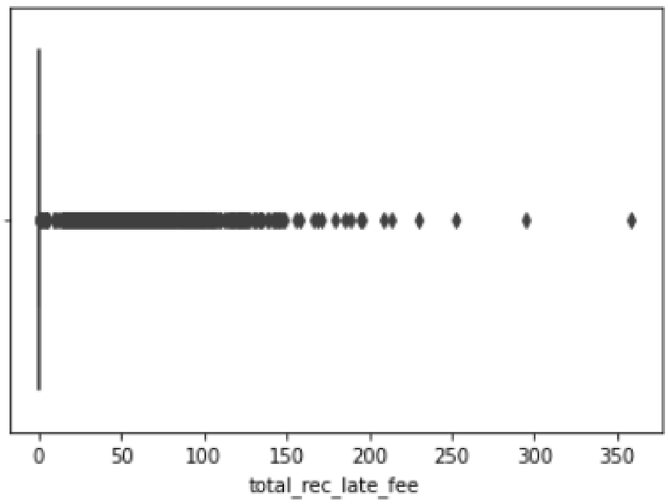
the interest received by the bank ranges between 0 to a maximum of 24205

majorities of the data lies between 200-1900 after outlier removal

the highest number of interest received lies between 0 to 2500

total_rec_late_fee: Late fee received till date

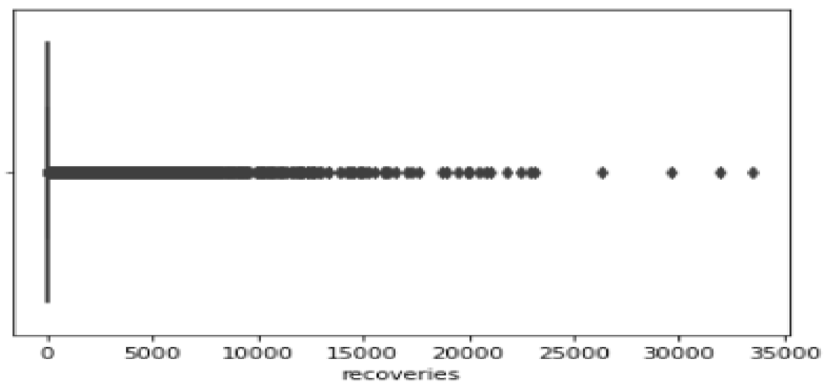


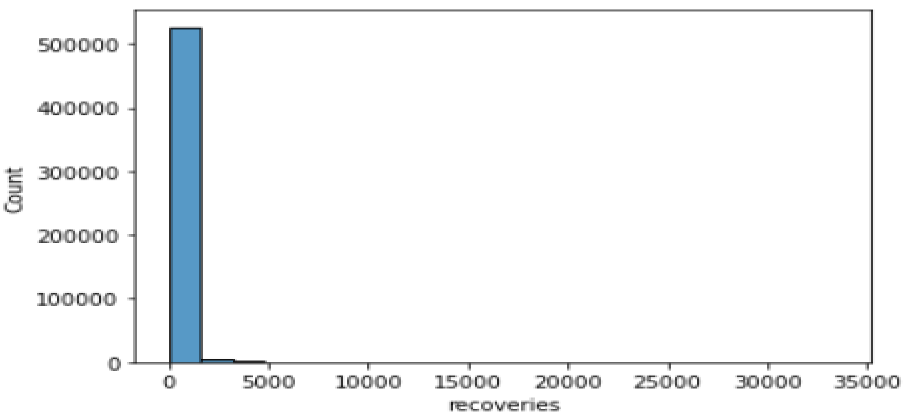


The range of late fee lies between 0 - 358, but the majority of the late fee is 0. One can conclude that most individuals pay on time. The outlier after zero to 350

recoveries:

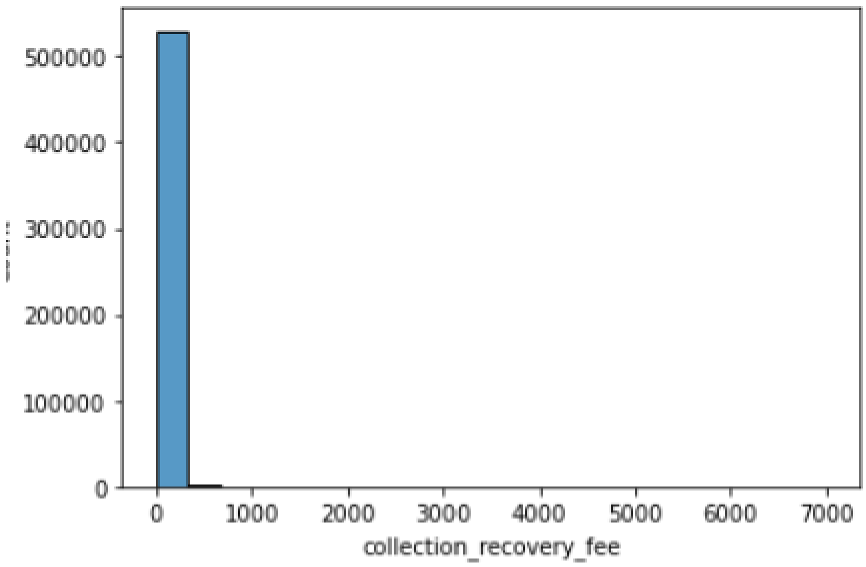
post charge off gross recovery

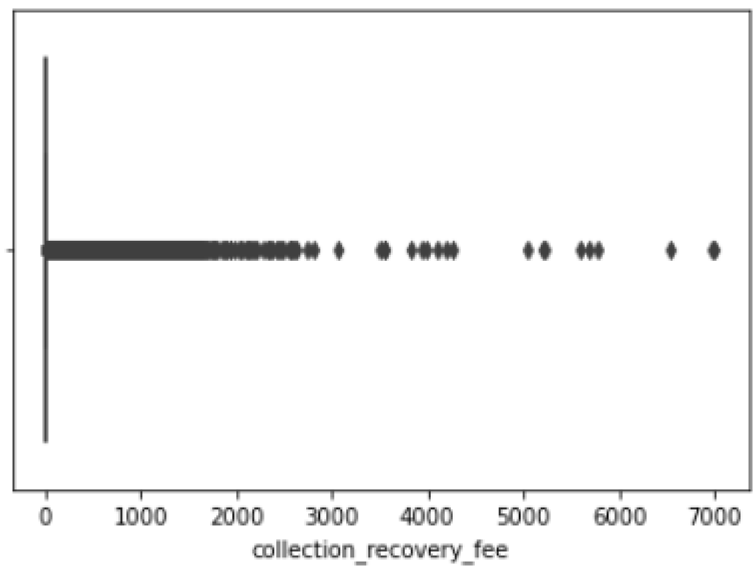




The range of late fees lies between 0 - 35000, but the majority of the late fee is 0 #one can conclude that most of the individuals pay on time and the recovery cost would not exist

collection_recovery_fee: post charge off collection fee

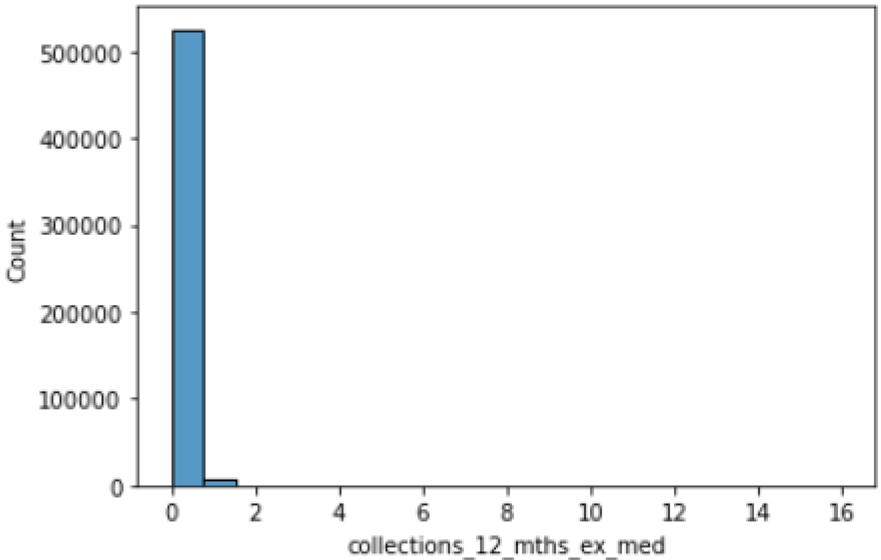


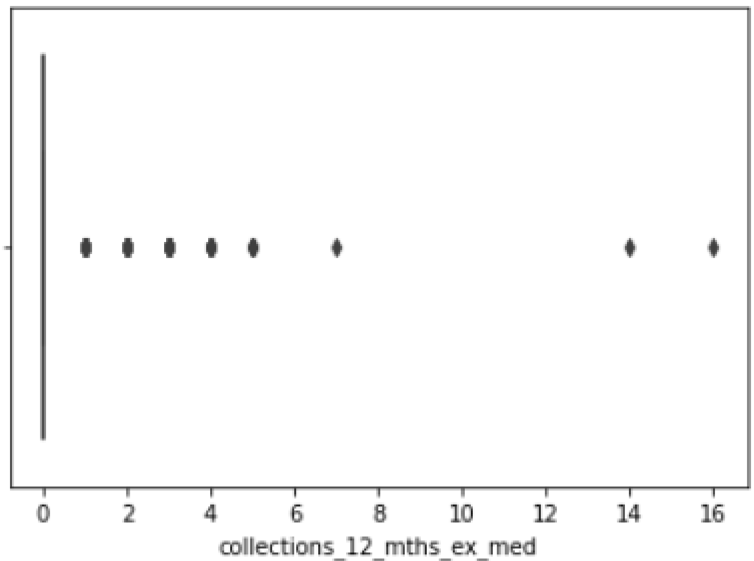


The data has a lot of outliers but avg collection_recovery fee -0.4 to 0.4 after the removal of the outliers. There is a lot of multi correlation between collection recovery fee, recoveries, total_rec_late_fee

collections_12_mths_ex_med:

Number of collections in last 12 months excluding medical collections



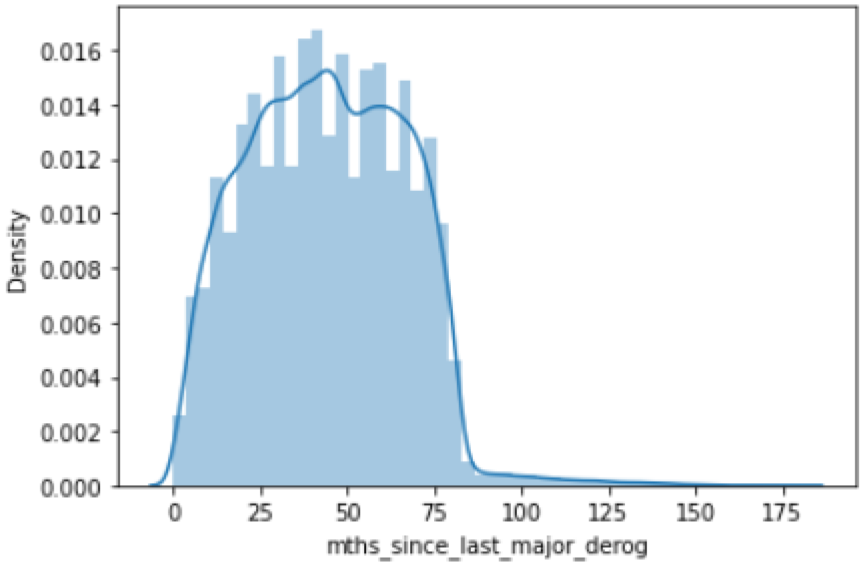


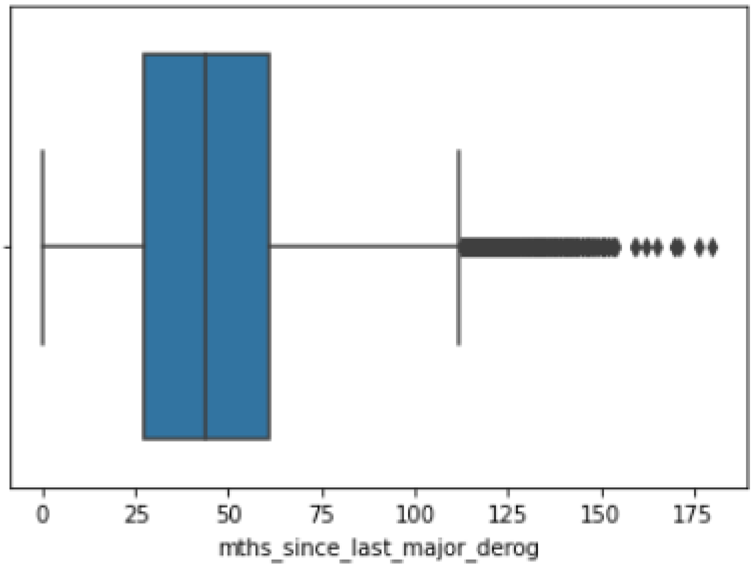
All the 4 data columns have right skew ,the collections_12_mths_ex_med multicollinear.

Data majorly lies around 0

mths_since_last_major_derog:

months since the most recent 90-day or worse rating

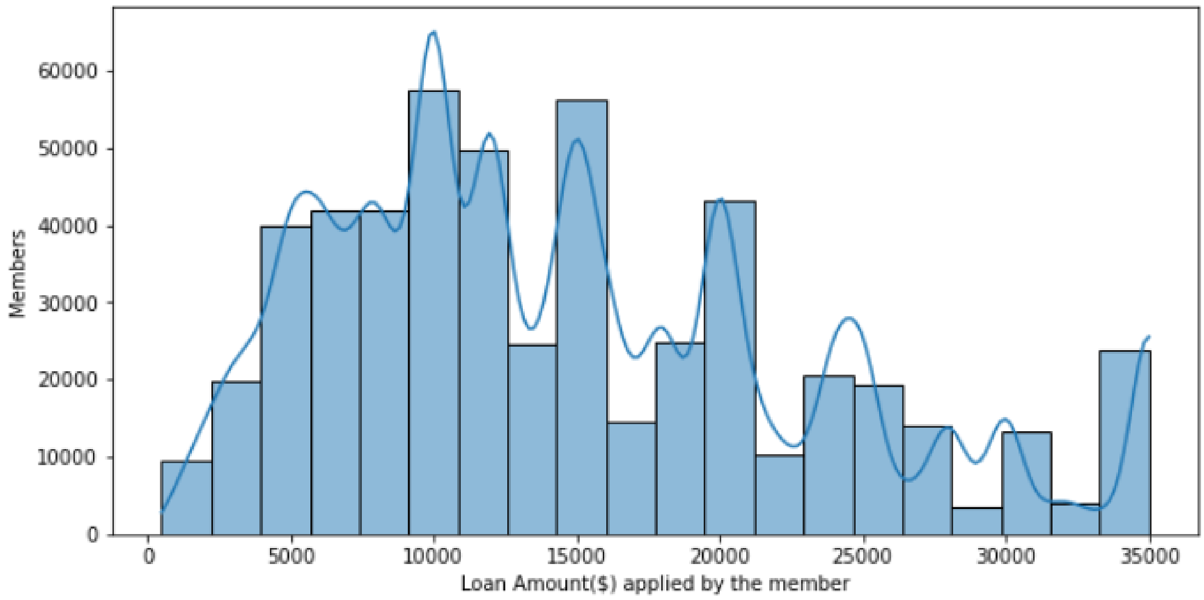




The iqr is between 25 -65, the mean appears to be 44, And has a slightly positive skew, outliers exist.

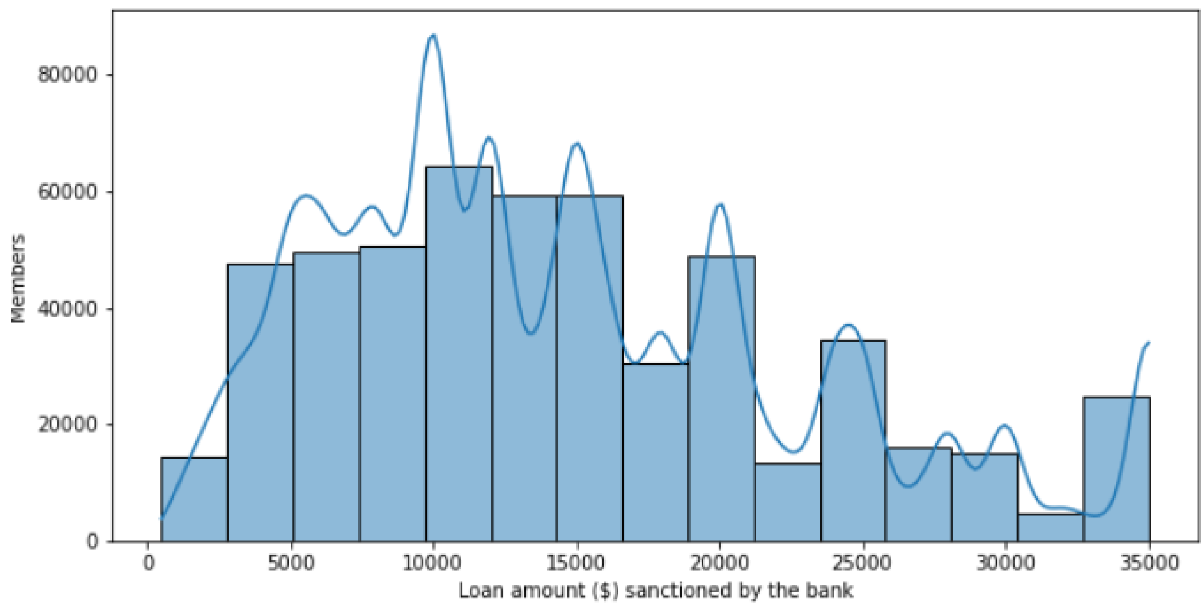
Loan Amount Attribute:

From the above plot, we can observe that the maximum loan amount applied by the member is 35000$ and approximately 58000 members applied the loan amount of 10000$ and similarly 56000 members applied the loan amount of 15000$.



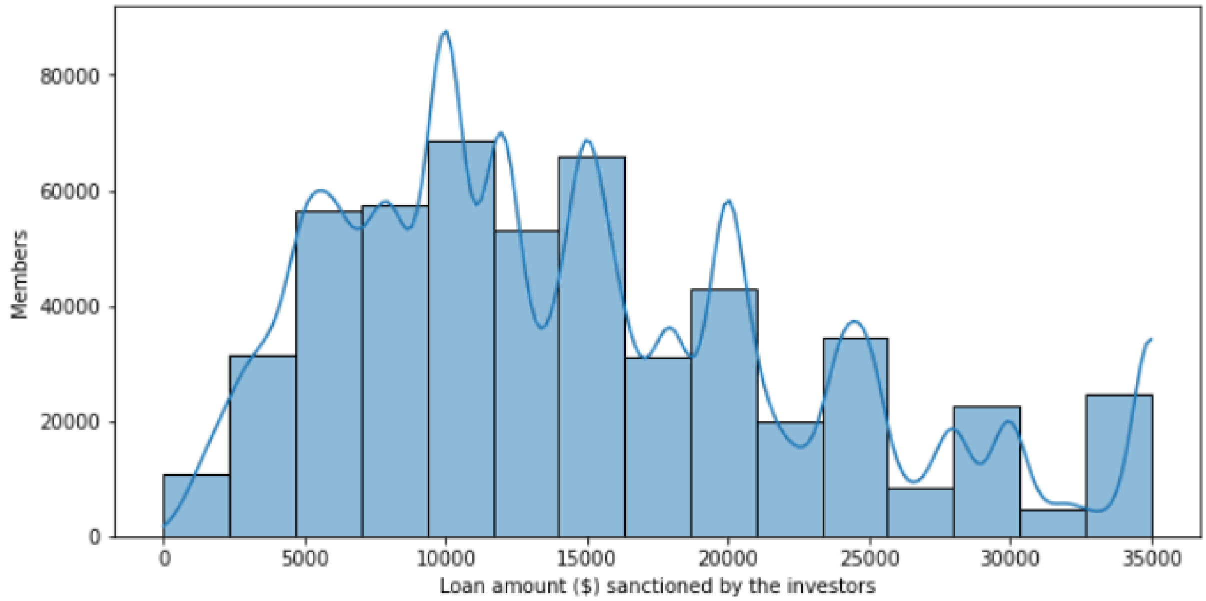
Funded Amount Bank Attribute:

From the above plot, we can observe that the maximum loan amount sanctioned by the bank is 35000$ and banks have also sanctioned the loan amount of 10000$ for more than 60000 members.



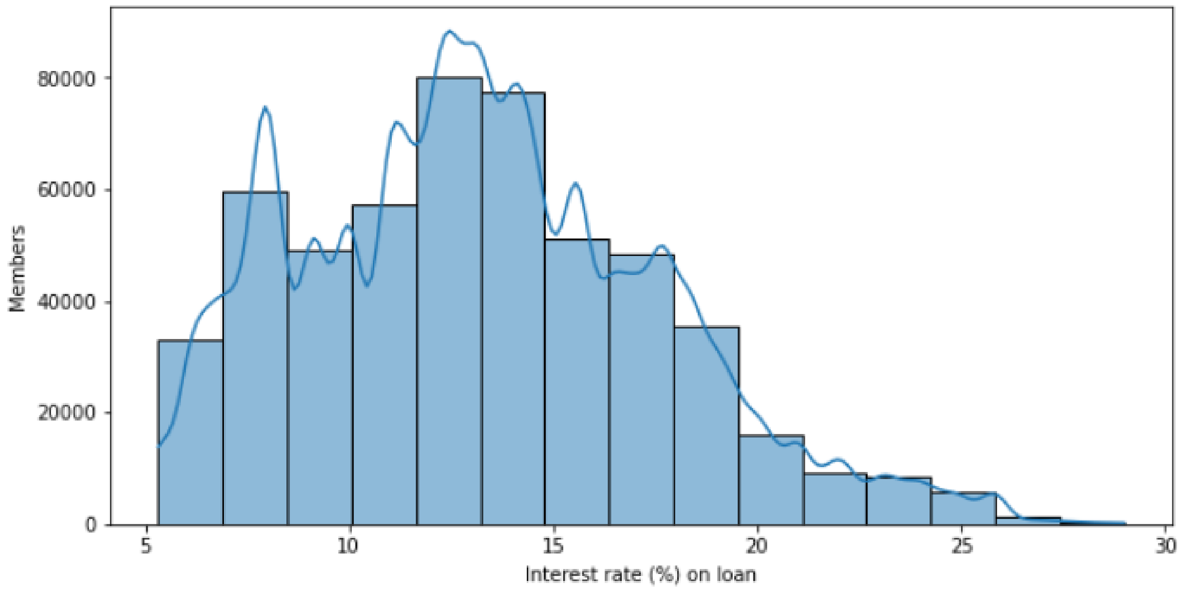
Funded Amount Investors Attribute:

From the above plot, we can observe that the maximum loan amount sanctioned by the investors is 35000$ and investors have also sanctioned the loan amount of 10000$ more than 68000 members

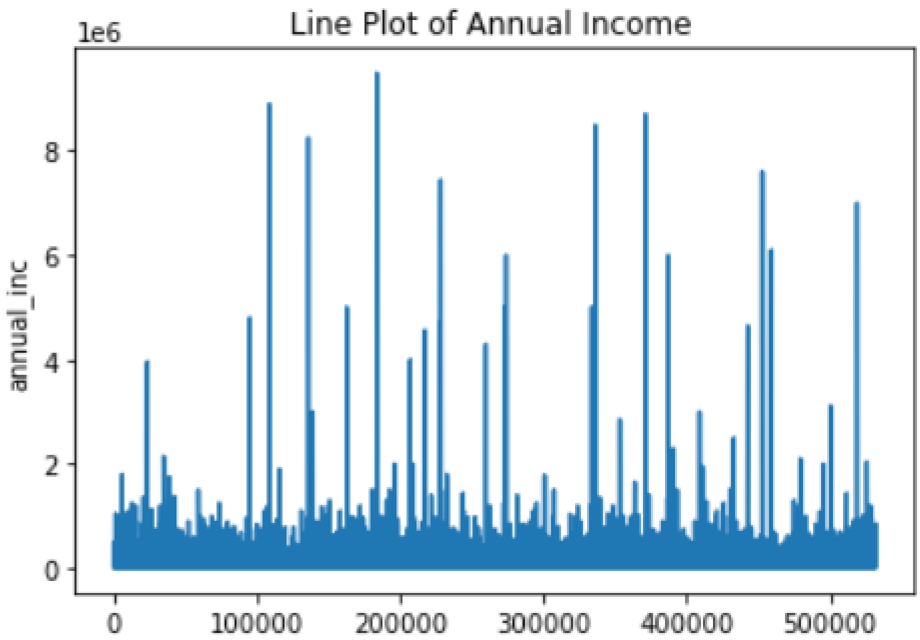


Interest Rate Attribute:

From the above plot, we can observe that the maximum interest rate given on a loan is around 26% and the minimum interest rate given on a loan is 6%. Approximately 80,000 members are paying the interest of around 13%.



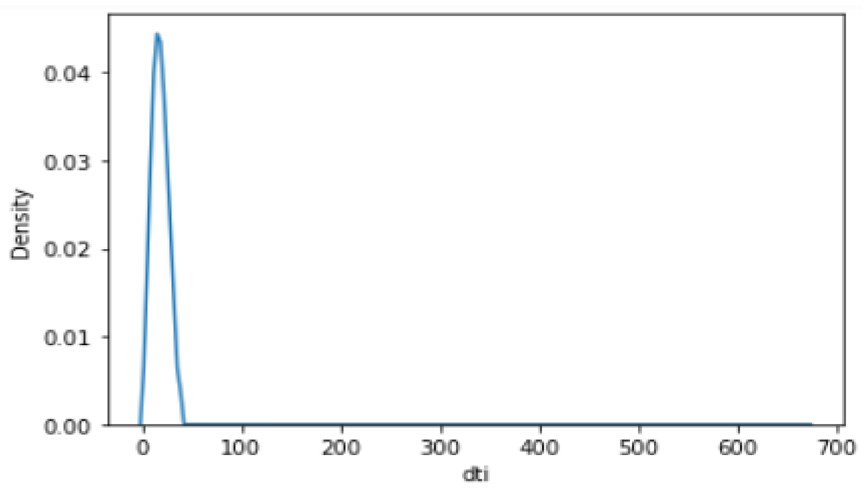
Annual Income Attribute:



Mean annual income reported by the members is approximately 75000$.Max annual income is 9500000$. The range of the annual incomes reported by members lies between 45000 $to 90000$.

DTI :

ratio of member's total monthly debt repayment excluding mortgage divided by self reported monthly income

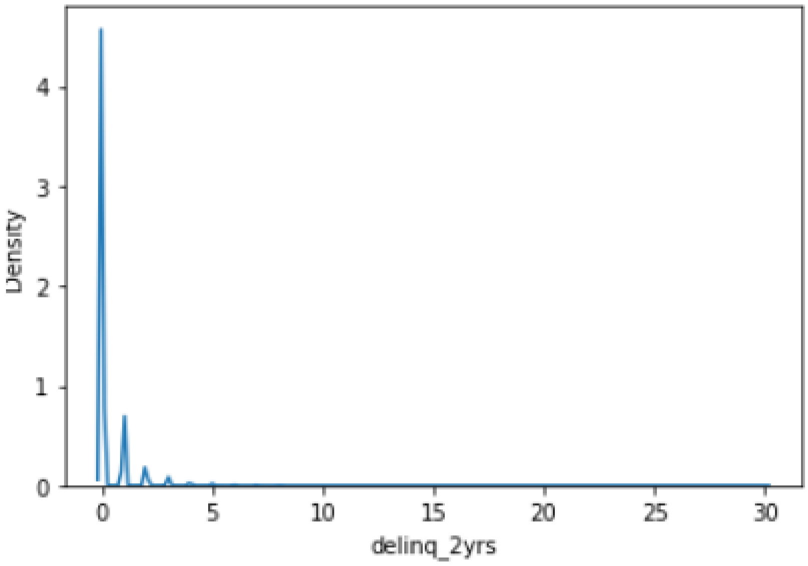


Lenders typically say the ideal front-end ratio should be no more than 28 per cent, and the back-end ratio, including all expenses, should be 36 per cent or lower. Lenders prefer borrowers with a lower DTI because that indicates less risk that you'll default on your loan. The majority of the data lies in between the range of 0 to 40.

delinq_2yrs: number of 30+ days delinquency in past 2 years

delinquency:- minor crime

If the payments are missed on payment in a single month across various credit accounts.

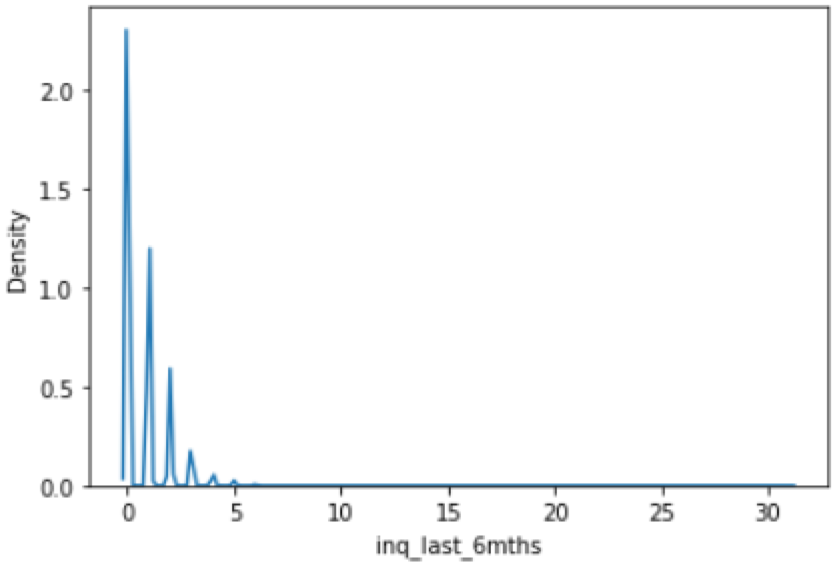


Most of the values lie at zero

inq_last_6mths: number of inquiries in last 6 months

How many inquiries is too many in 6 months?

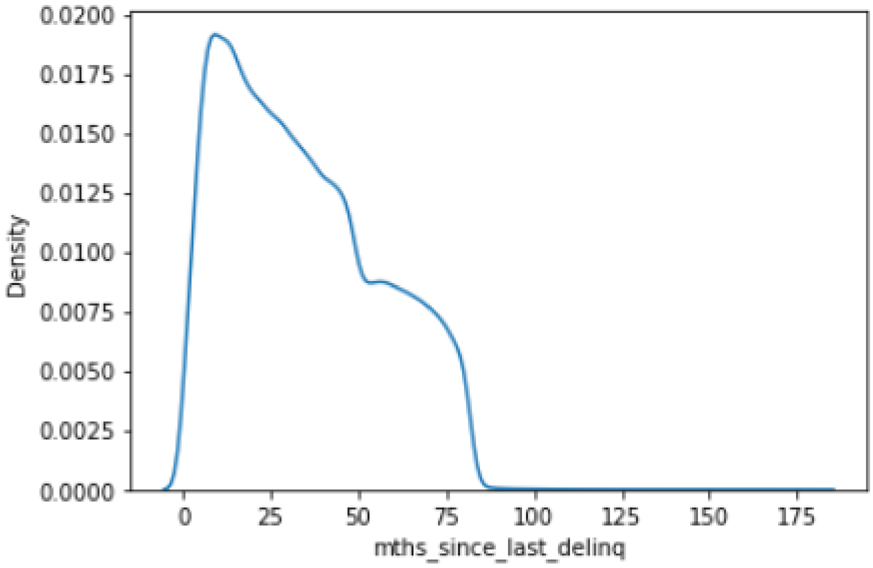
For many lenders, six inquiries are too many to be approved for a loan or bank card. Even if you have multiple hard inquiries on your report in a short period of time, you may be spared negative consequences if you are shopping for a specific type of loan.



Most of the values lie at zero. There are outliers and there are a few null values, most of them are below 6 in our graph data. Increased the multiplying factor to 5 and reduced the outliers.

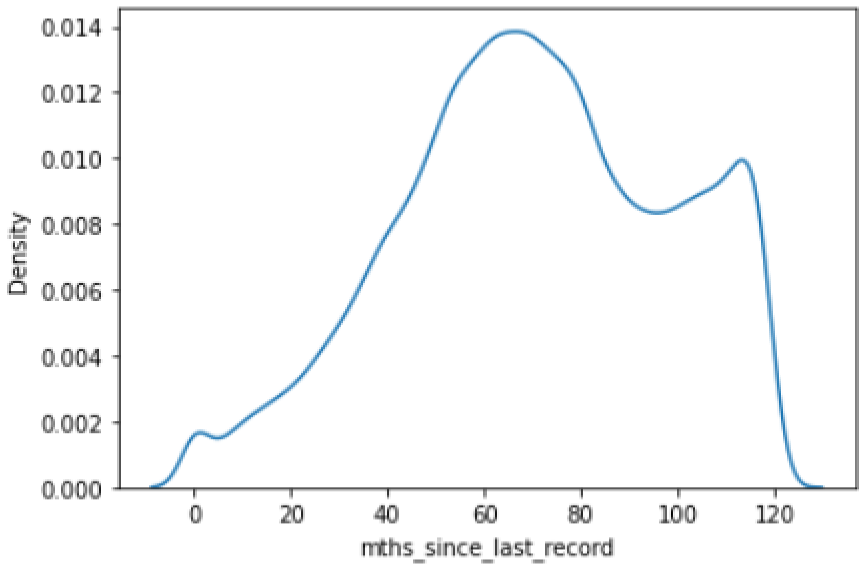
mths_since_last_delinq

Number of months since last delinq , few outliers and lots of null values



mths_since_last_record: Number of months since last public record

Public records may indicate you stopped paying your bills, No outliers and its two peaked.

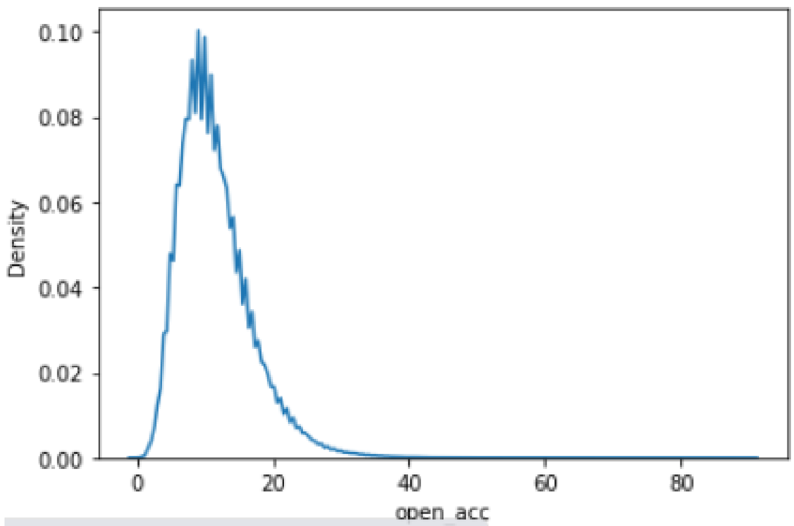


In the above graph, we could see that the record for defaulters has a high density towards the right side compared to the non-defaulters

open_acc:

- number of open credit lines in members' credit line
- Open credit is a pre-approved loan between a lender and a borrower.
- It allows the borrower to make repeated withdrawals up to a certain limit and then
- Make subsequent repayments before the payments become due

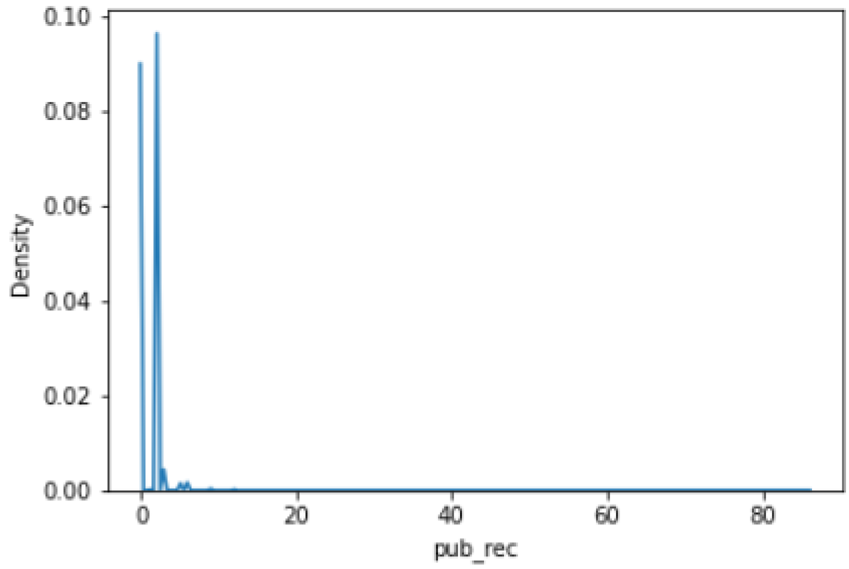
A new line of credit may improve your credit score. However, you should never take out an additional line of credit unless necessary. Applying for multiple lines of credit in a short period is not advised, and having too many lines of credit makes you look risky to lenders.



There is no evidence that the more the number of credit lines more is the chance of getting approved for a loan

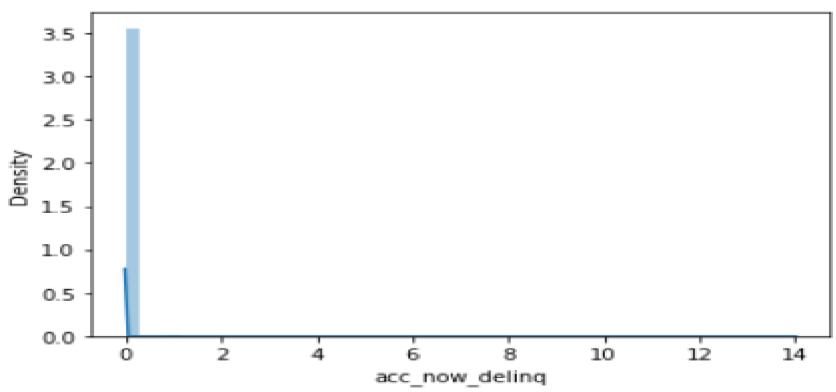
pub_rec: a number of derogatory public records

A derogatory account is one that is seriously past due. Most commonly, the term derogatory refers to accounts that are 60 or 90 days past due or more. It also includes collection accounts, charge-offs, repossessions and foreclosures The only type of public record information that would appear on your credit report is a bankruptcy filing.



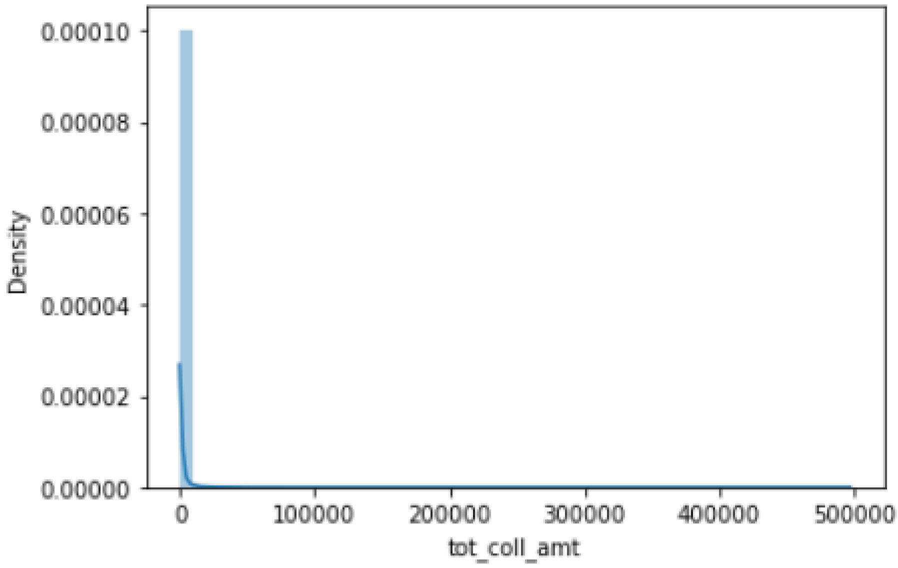
acc_now_delinq: number of accounts on which the member is delinquent

Most of the members pay on time and no default exists although there are a few outliers indicating that some people have a delay in their payment and their accounts have been marked



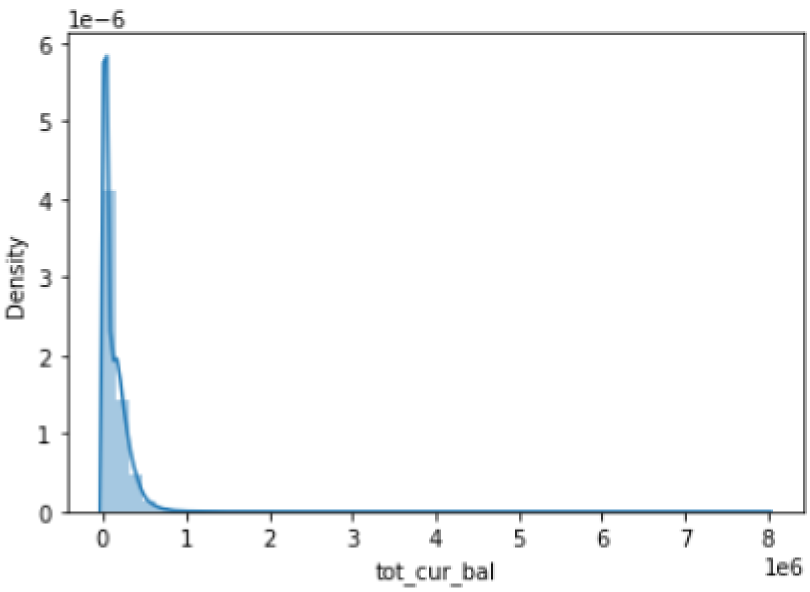
tot_coll_amt:

most of the data lies in zero which can be understood that people have yet to start to pay.



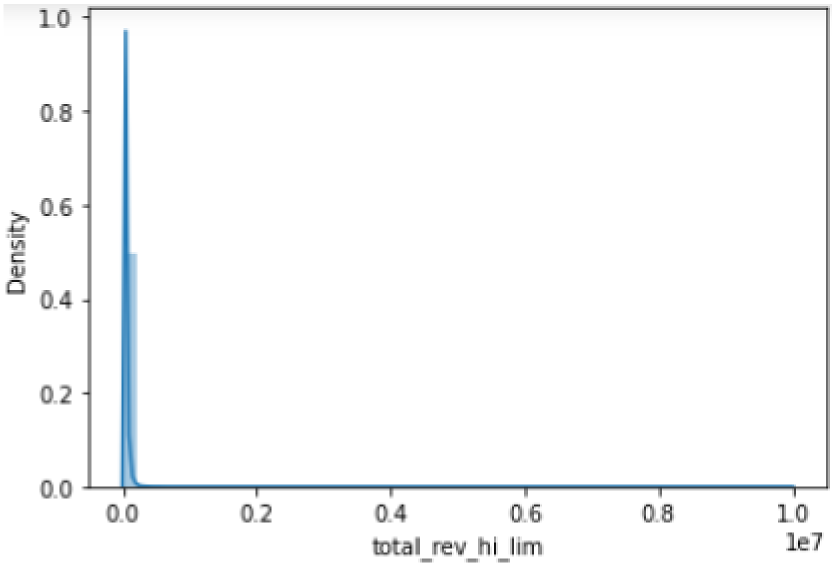
tot_cur_bal:

most of the data lie between 0 and 1, that people have yet to start to pay



total_rev_hi_lim:

most of the data lie between 0 and 0.2, that people have low credit revolving limit



Employee Title:

With more than 1lakh unique job/employer title of member, ‘Teacher’ title is the most number of titles given by an employer, almost 8000 members are with the Teacher employee title.

Desc:

This Attribute talks about the loan description given by the members. With more than 85% null values present in the said attribute. It is very insignificant . Hence can be dropped from the dataset.

Title:

This attribute is about the loan title given by the members. With almost 40000 unique records. ’Debt Consolidation’ is the loan title provided by maximum members. Almost 2.48lakh members opted for Debt Consolidation as their loan title.

Batch Enrolled Attribute:

From the data, we can observe that Batch numbers allotted to members are mostly different.

Member ID Attribute:

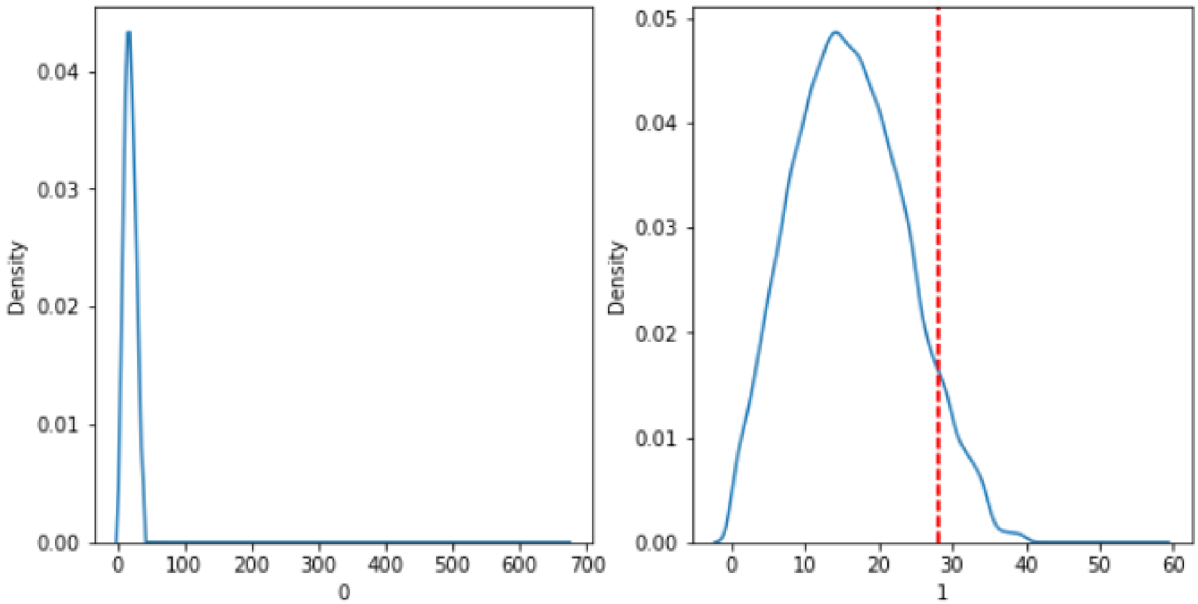
From the data, we can observe that member id of each member is unique.

Outlier detection:

we have observed the presence of outliers in every numerical data type column which is mentioned during the univariate analysis

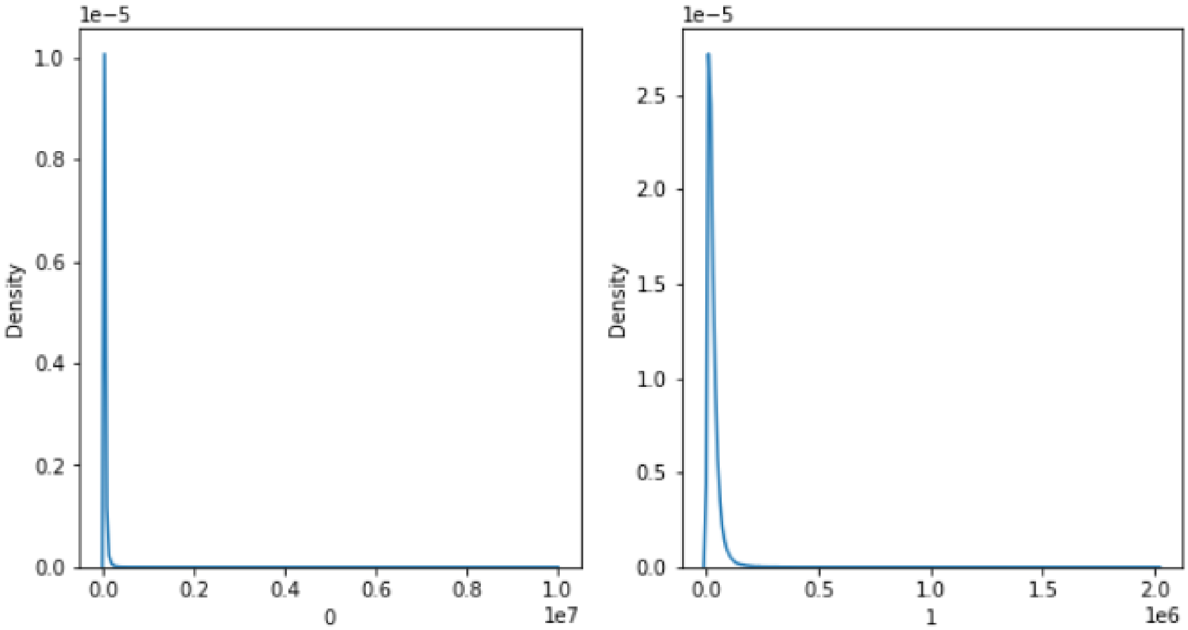
Bivariate data:

dti vs loan_status¶



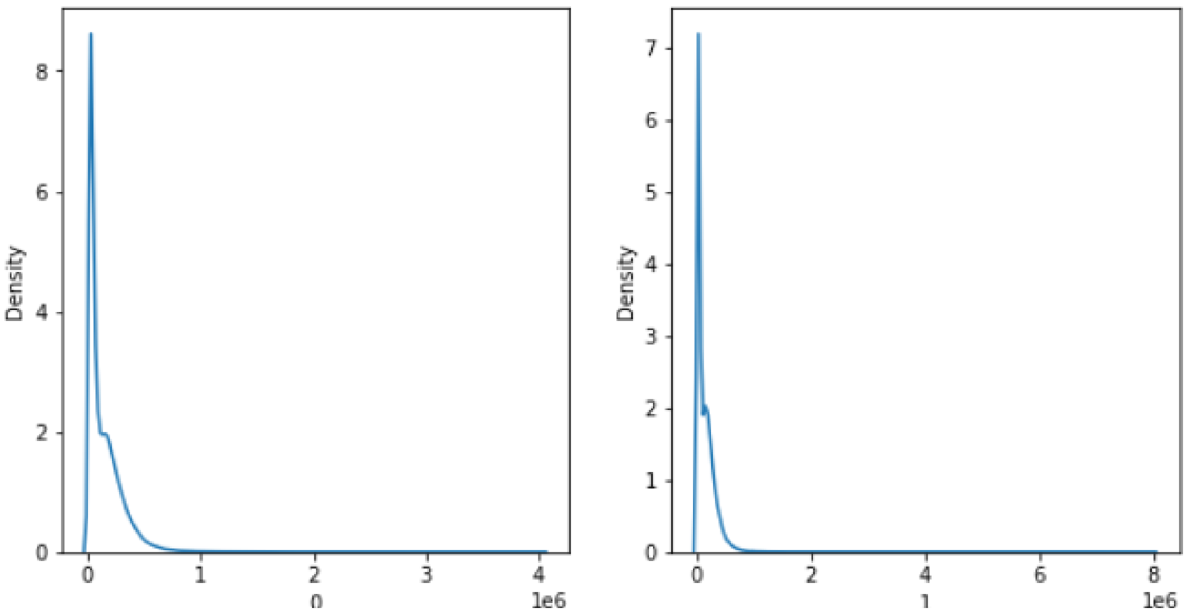
- Below graph shows two distributions of a non-defaulter and a defaulter respectively
- In both cases we could see the majority of the data lies in between the range of 0 to 40
- In default graph we could see a lot of the member have ratios greater than 28 which could lead to rejection in loan

loan_status vs 'total_rev_hi_lim':



- Below graph shows two distribution of a non-defaulter and a defaulter respectively
- in non-default we see a most of the density of the data is situated in the zero value
- In default the area of density is more and has values greater than zero too.

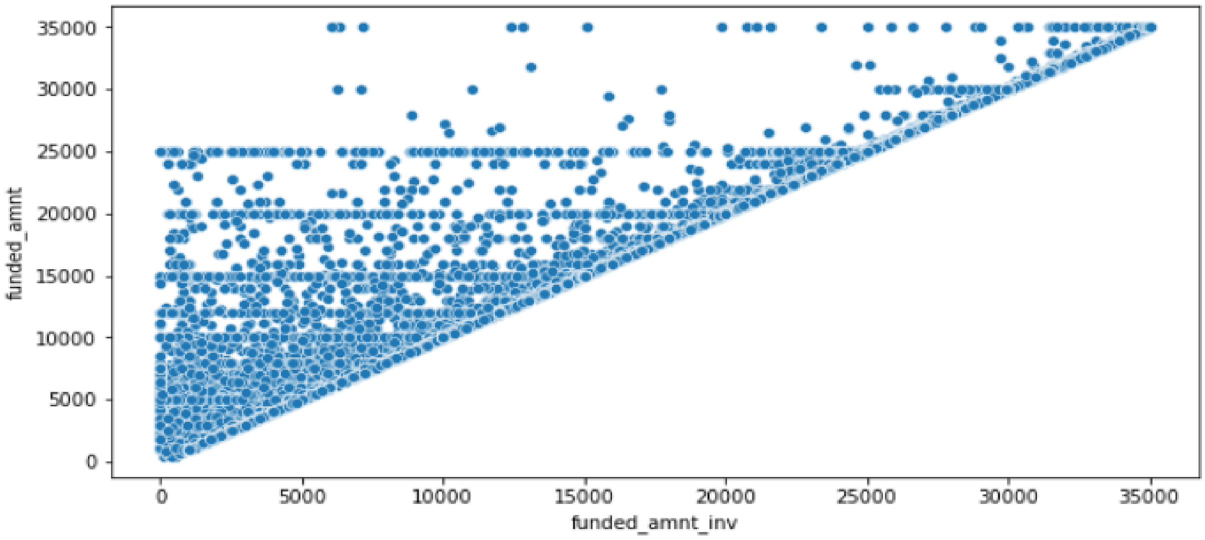
loan_status vs tot_cur_bal



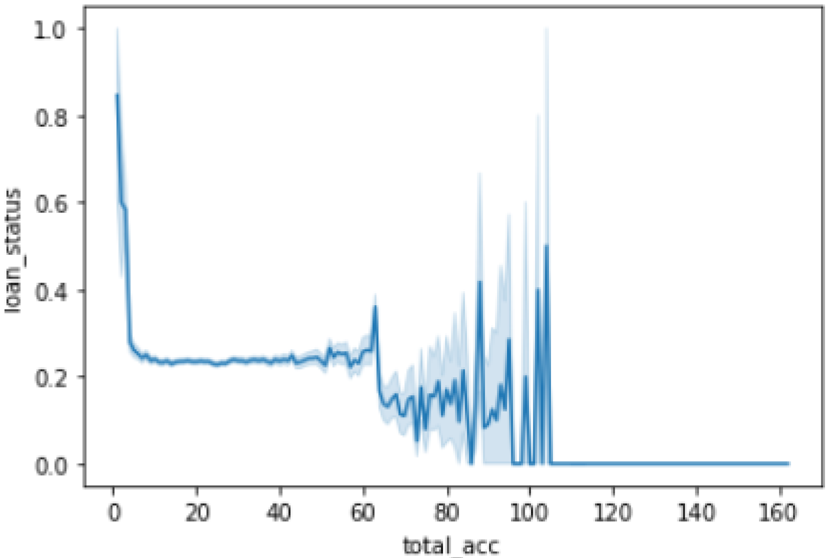
- Below graph shows two distributions of a non-defaulter and a defaulter respectively
- In both the cases we can see the majority of the data lies in the non-default data

- In the default graph we can see a lot of the members have zero values

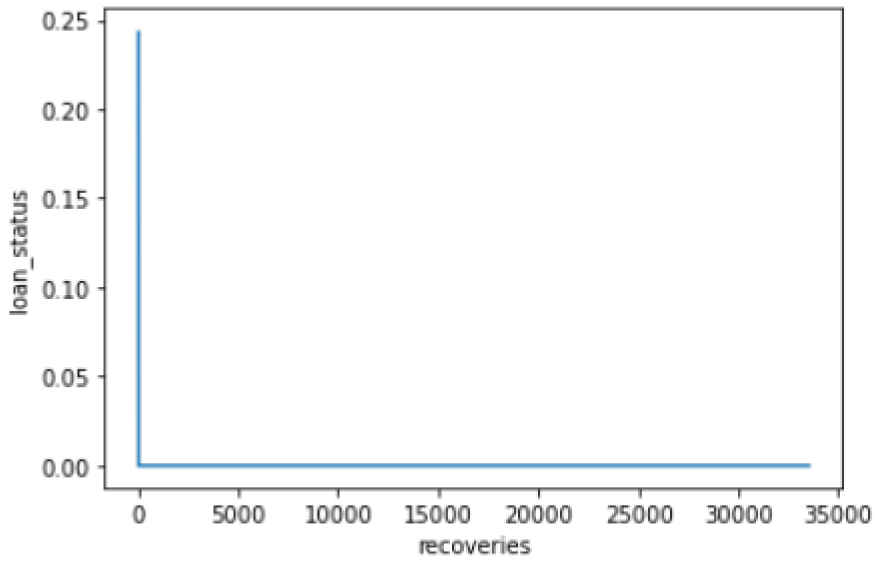
funded_amnt_inv vs funded_amnt: we can observe that Loan amount sanctioned by the bank is less or equal to the amount sanctioned by the investors.



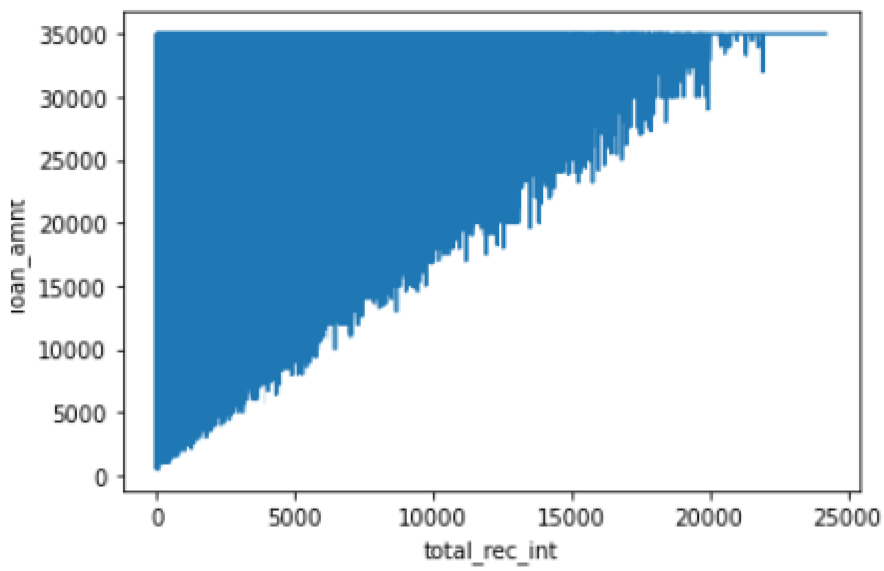
total_acc vs loan_status: there is a sudden drop which settles down and then we can observe and a smaller drop in loan status as total_acc increases

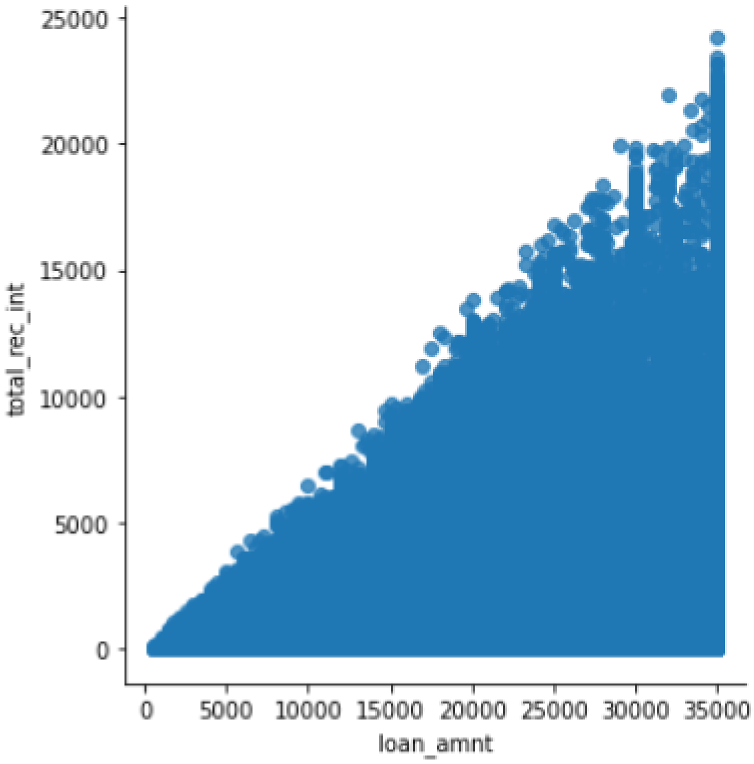


recoveries vs loan_status: L graph this indicates that most of the values lie on zero and there exists outliers.

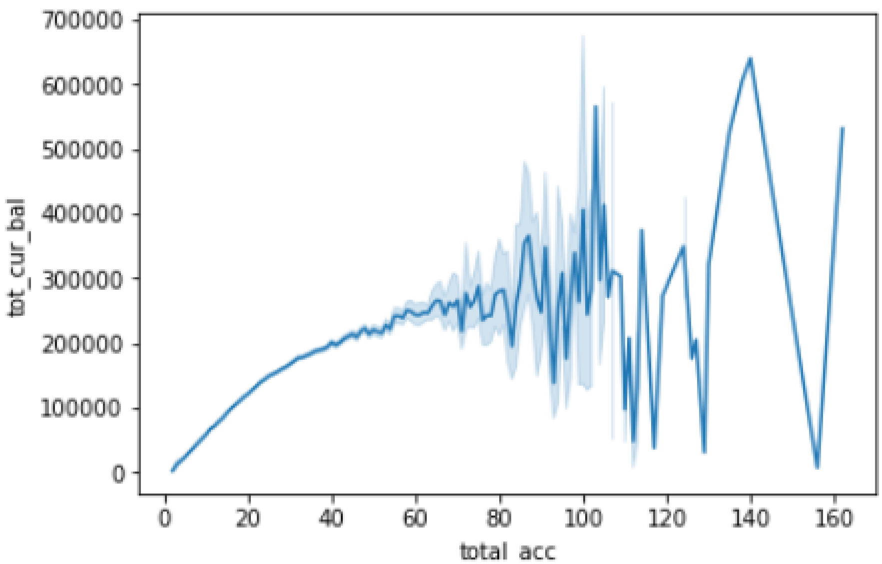


total_rec_int vs loan_amnt: Triangle-based filled graph indicating with total_rec_int increase there is a decrease in loan amount

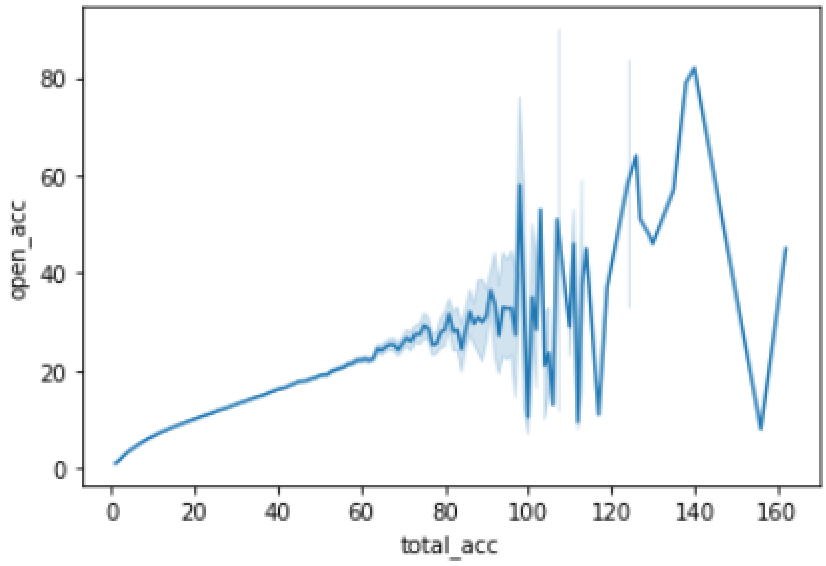




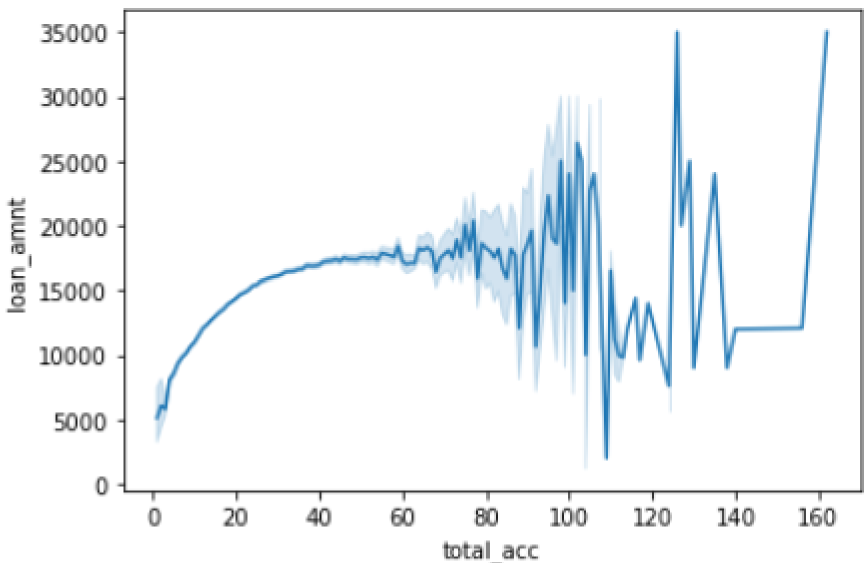
total_acc vs tot_cur_bal: With the increase in the total number of credit lines the loan status decreases



total_acc vs open_acc: The open_acc increases with an increase in total_acc and total_rec_int

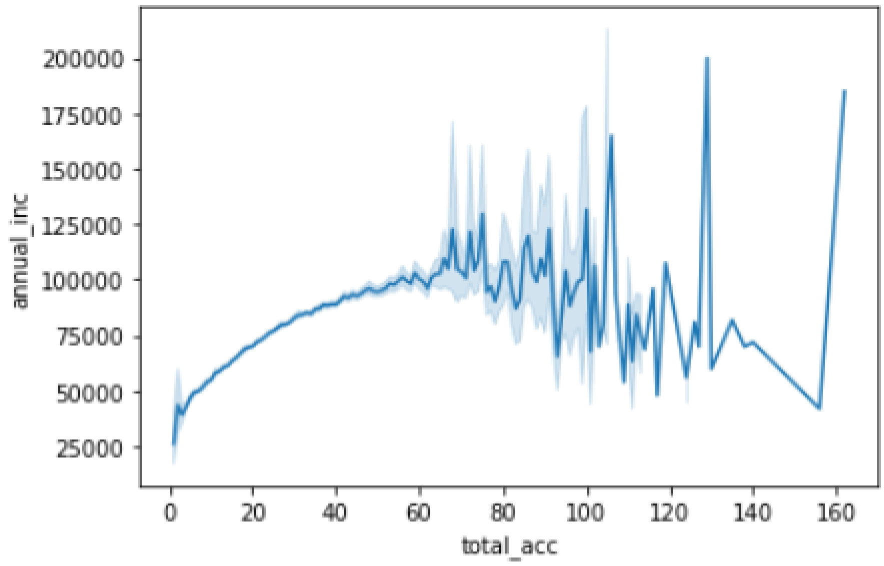


total_acc vs loan_amnt: Initially there is an exponential change and then it hits a plateau



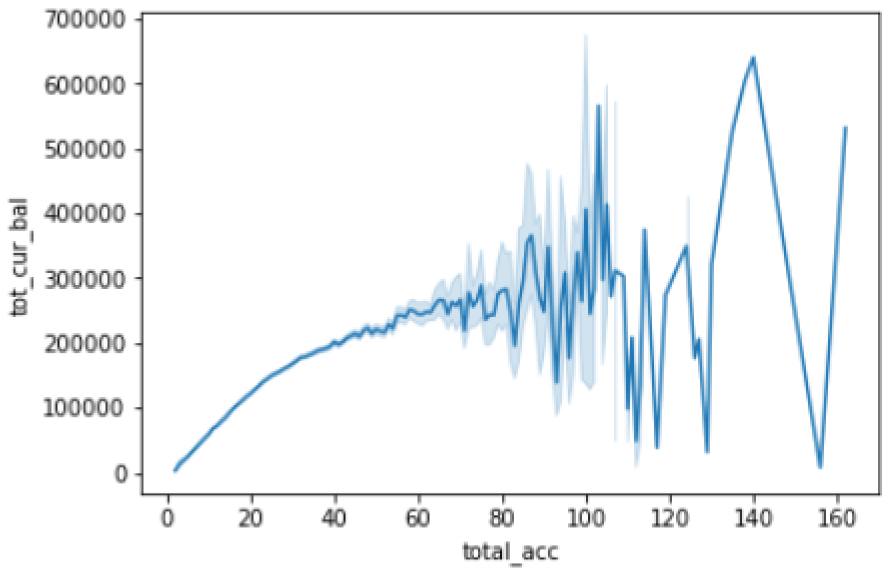
total_acc vs annual_inc:

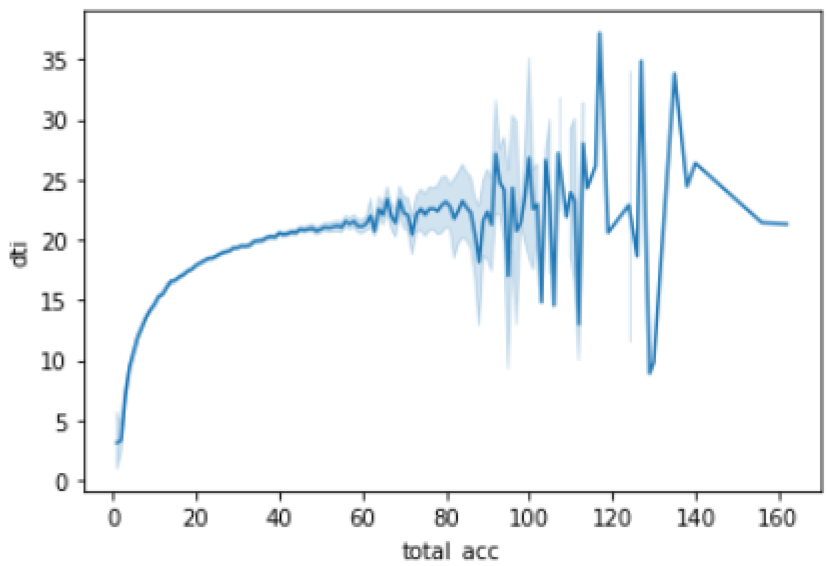
annual_inc increases with increase in total_acc once total_acc reaches 80 it begins to drop



total_acc vs dti & ['tot_cur_bal']:

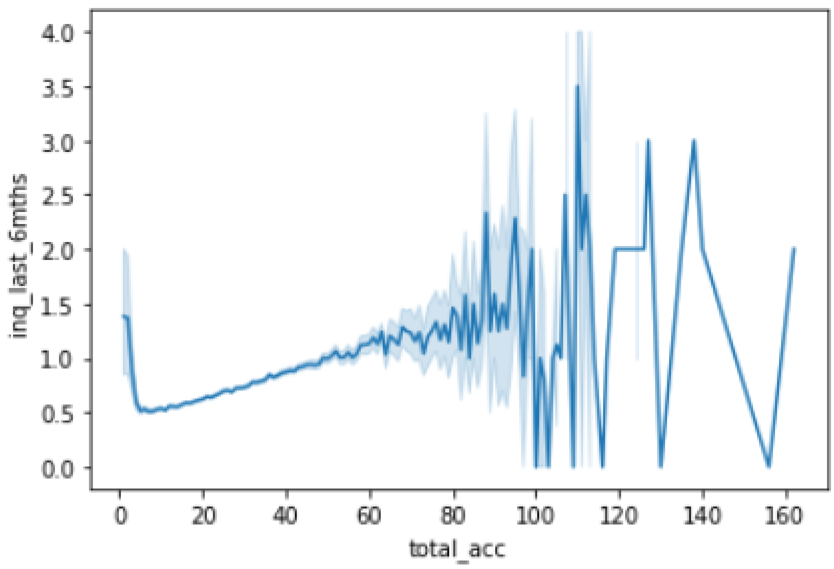
- initially there is a exponential change and then it hits a plateau at 40





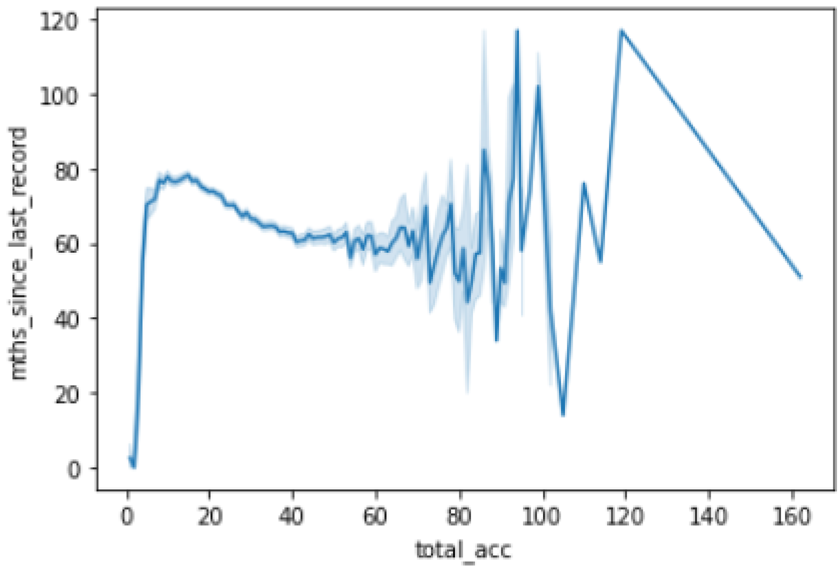
total_acc vs inq_last_6mths:

There was sudden drop and there was slow growth



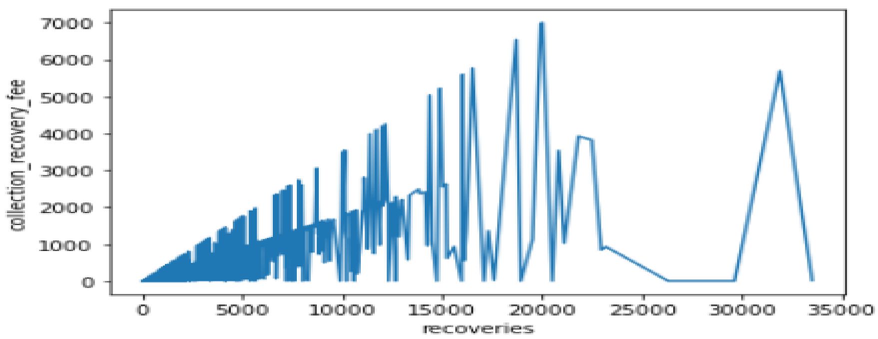
Total_acc vs mths_since_last_delinq and mths_since_last_record:

increases and then hits a plateau



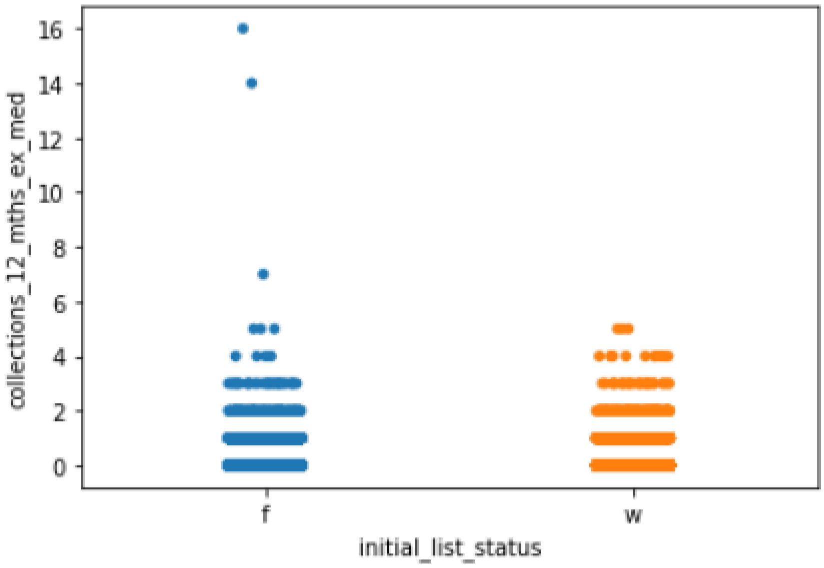
recoveries vs collection_recovery_fee:

linear vague line



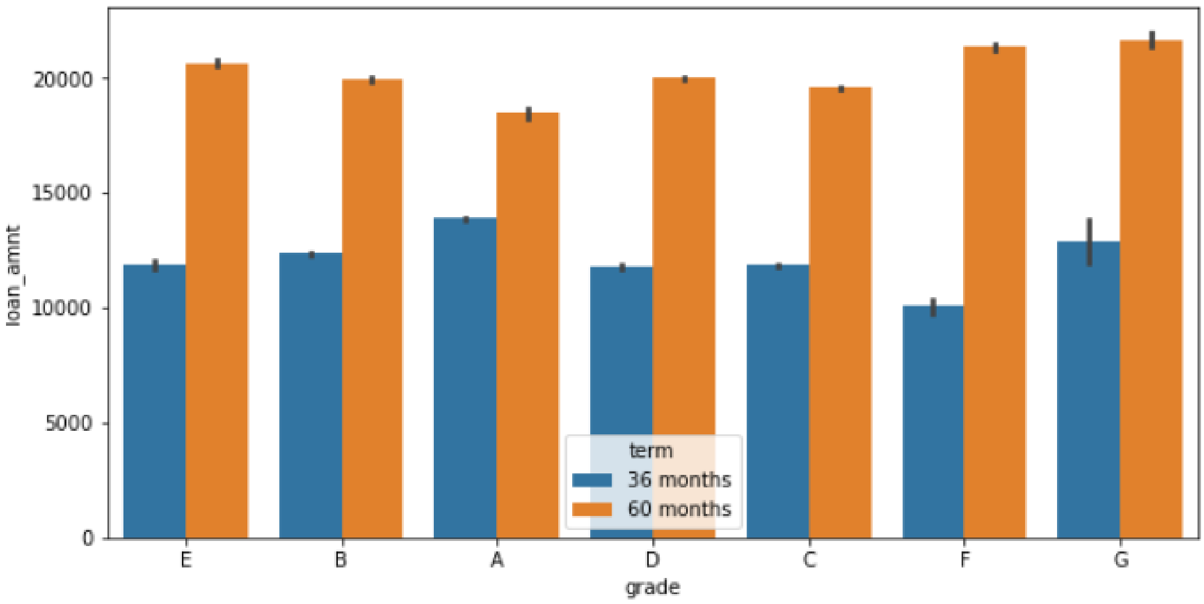
initial_list_status vs collections_12_mths_ex_med:

most of the values concentrate around 0-5



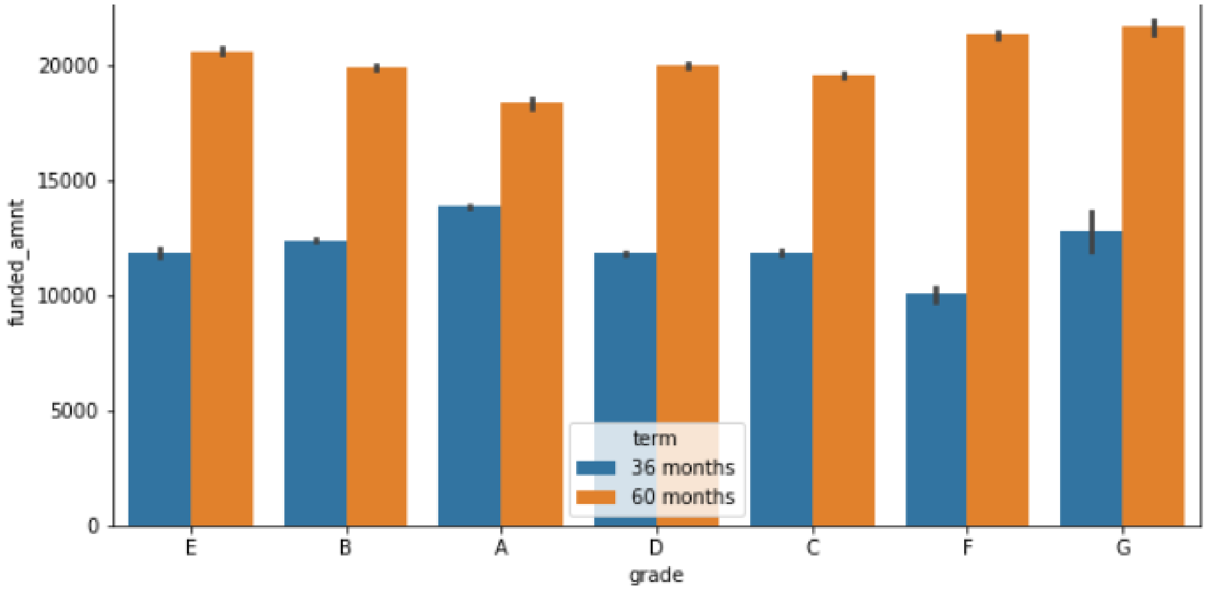
MULTIVARIATE

Grade vs Loan_amnt (Hue: Term)



- We can observe that the maximum loan amount applied to the bank is around 60 months and this is assigned under Grade G.

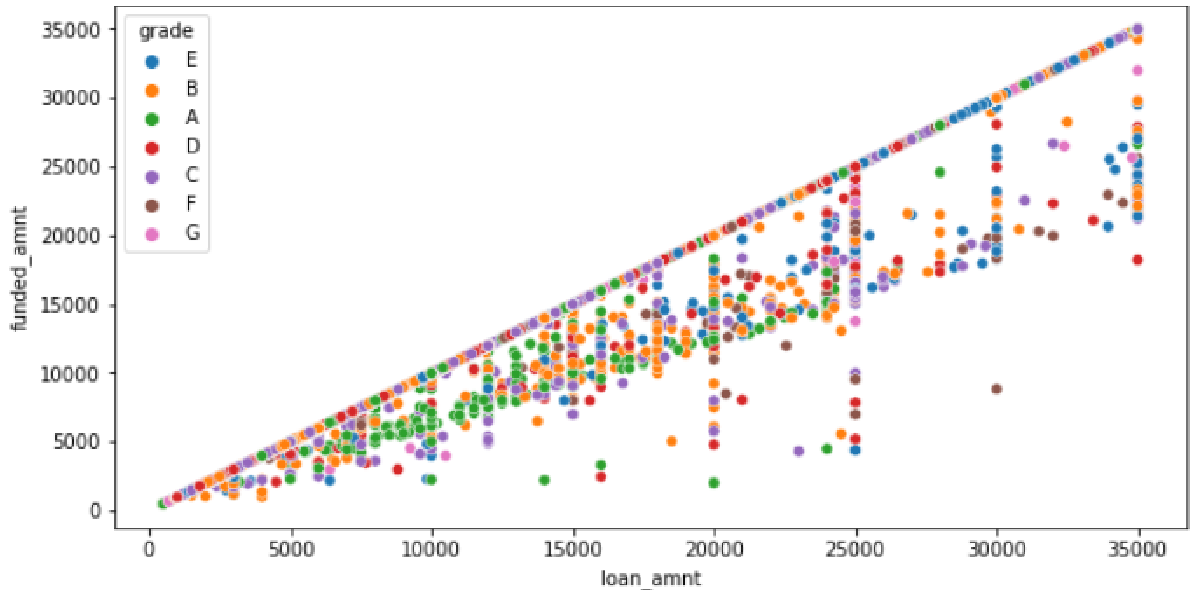
Grade vs funded_amnt (Hue: Term)



- we can observe that the maximum funded amount offered by the bank for the 60 months and it also comes under Grade G

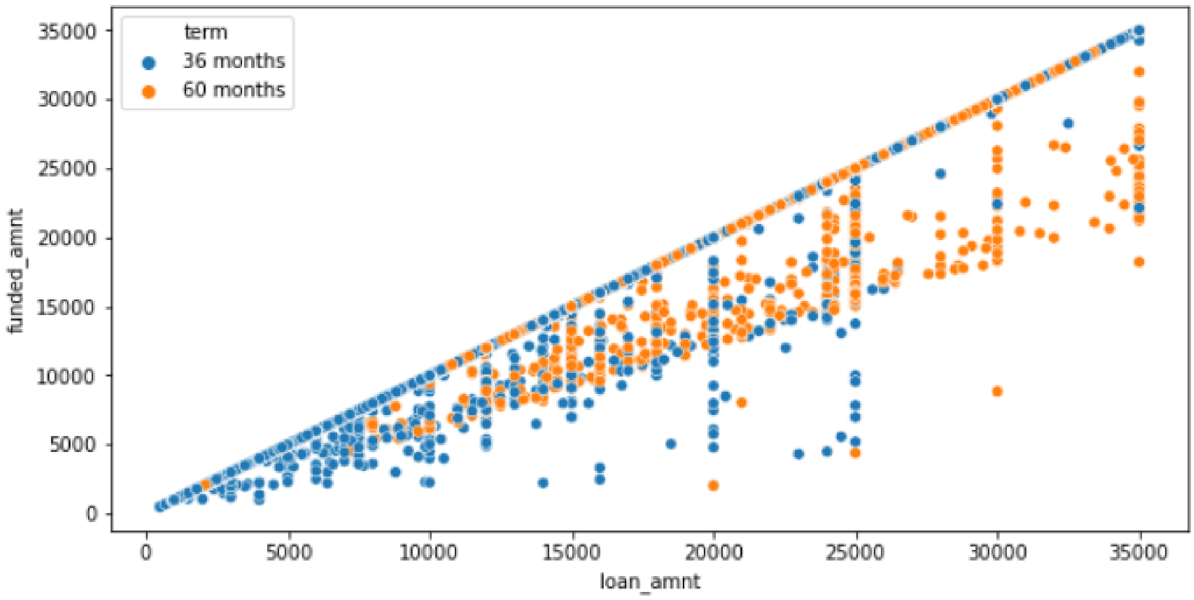
loan_amnt vs funded_amnt (Hue: grade)

- From the below plot, we can observe that as the loan amount is increasing the funded amount is also increasing.
- This is the directly proportional relationship between them.



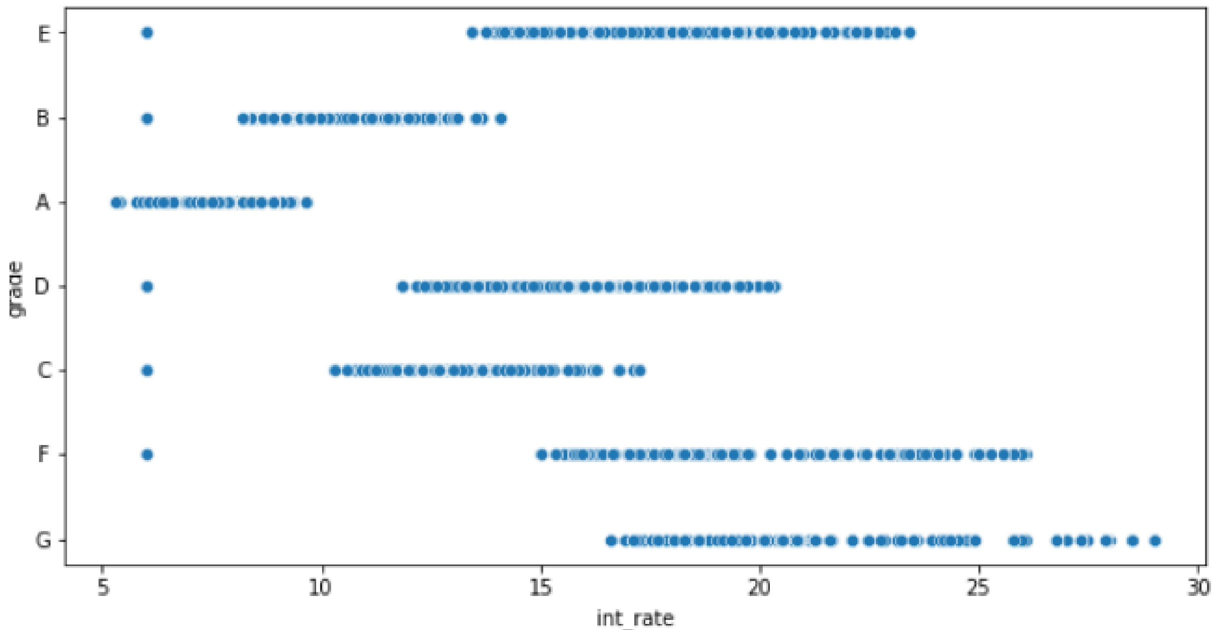
loan_amnt vs funded_amnt (Hue: term)

- From the below plot, we can observe that as the loan amount is increasing the funded amount is also increasing.
- This is the directly proportional relationship between them.
- when the loan_amount increases the number of default members also increases



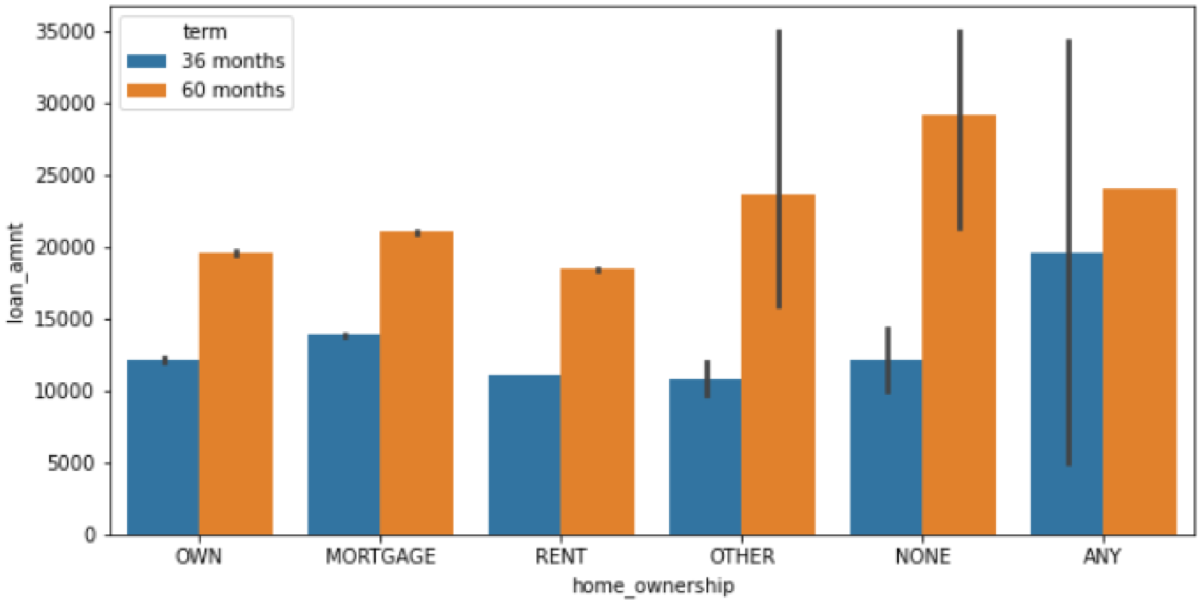
int_rate vs grade:

- From the below plot, we can observe that the maximum interest rate is applicable for the Grade G.



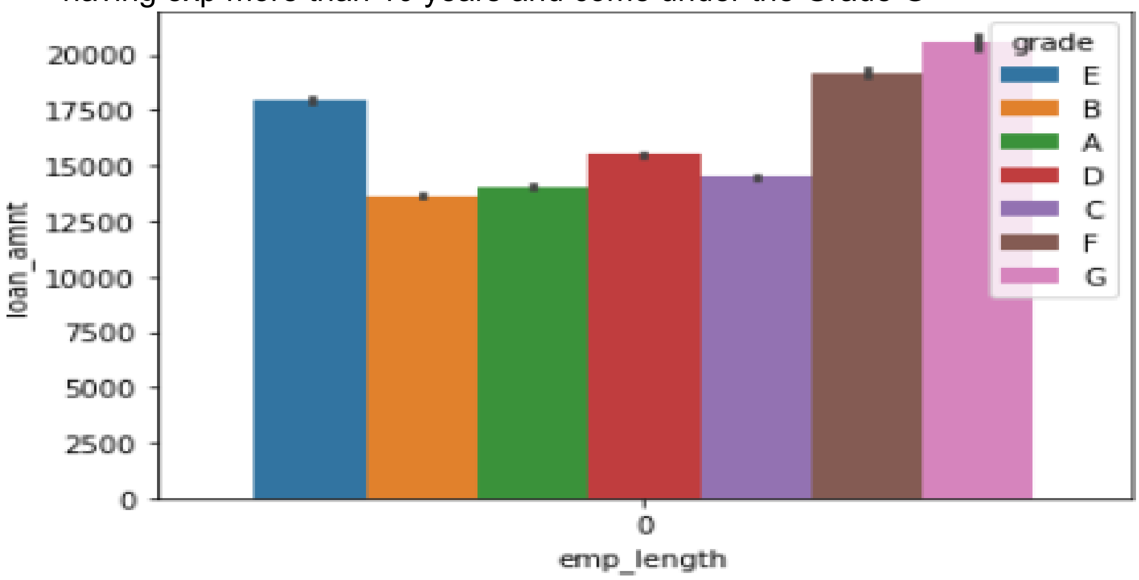
home_ownership vs loan_amnt (Hue: term)

- we can observe that maximum loan amount (30000$) applied by the members whose ownership is categorised by 'None'.
- From the hist plot, the 'None' category home is very less.
- Maximum members (around 50%) come under the "Mortgage" category who applied the loan amount around 21000$



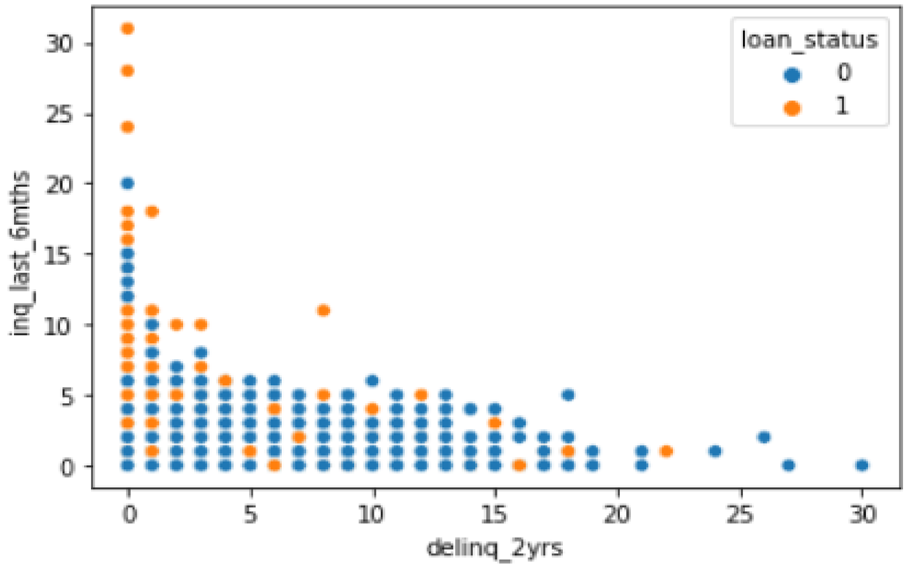
emp_length vs loan_amnt (Hue: grade)

- we can observe that maximum loan amount offered by the member having exp more than 10 years and come under the Grade G



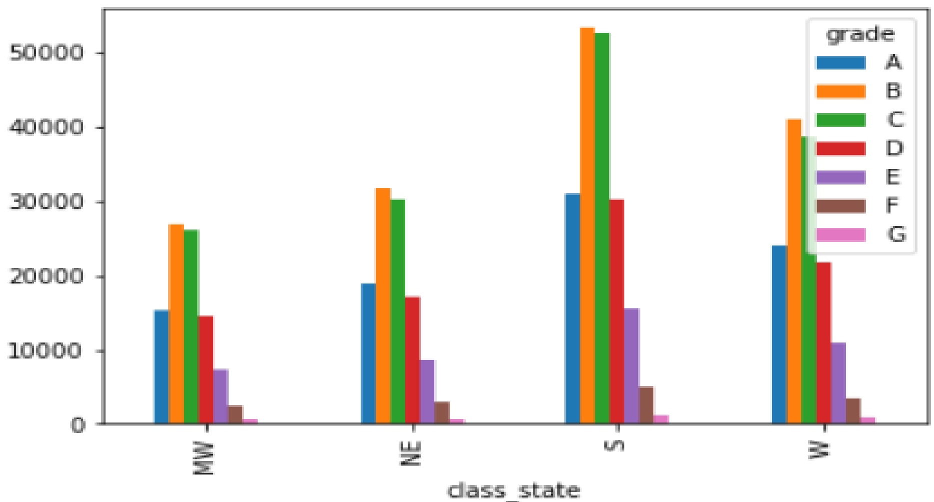
delinq_2yrs vs inq_last_6mths (hue: loan_status)

- From the plot we see that when the delinquency is zero the number of inquiries is more
- Although the delinquency is zero due to the high number in inquiries we see a lot of default loan status



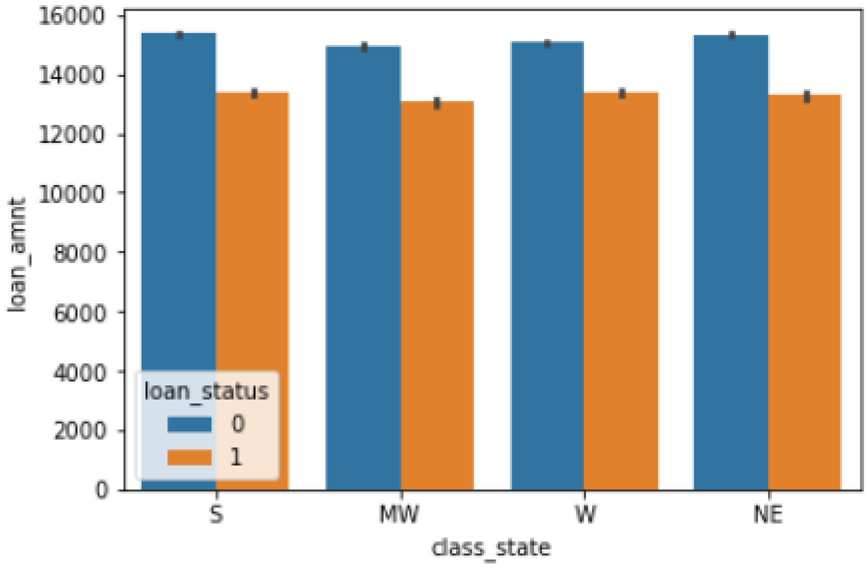
class_state vs grade

- The ratio of all the grade in every state is the same



class_state vs loan_amnt (hue: loan_status)

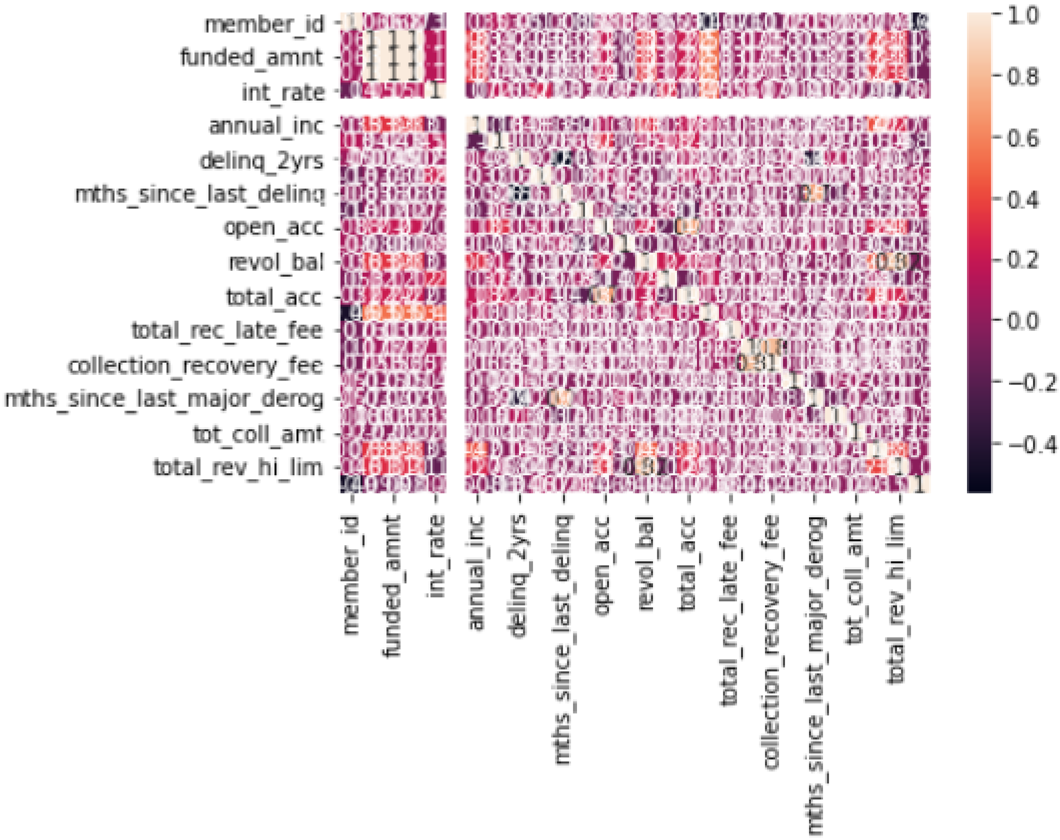
- Every state has a equal average and a equal non-default to default ratio



CORRELATION PLOT:

Inferences

- Member_id has a very weak negative correlation with loan_status.
- funded_amnt,funded_amnt_inv are perfectly positively correlated with loan_amnt.
- similarly funded_amnt,funded_amnt_inv,loan_amnt are perfectly positively correlated to each other.
- last_week_pay has a significant negative correlation with member_id.
- total_acc has a significant positive correlation with open_acc.
- total_rev_hi_lim has a significant positive correlation with revol_bal.
- collection_recovery_fee has a significant positive correlation with recoveries.
- total_rec_int has a weak positive correlation with loan_amnt,funded_amnt,funded_amnt_inv,last_week_pay columns.
- mths_since_last_dealing has a weak negative correlation with delinq_2yrs.
- member_id has a weak negative correlation with total_rec_int columns.
- mths_since_last_major_derog has weak positive relation with mths_since_last_delinq.
- tot_cur_bal has a fragile positive relation with annual_inc.



- due to numerous columns in the data frame we have refrained from displaying it but it can be verified in the file attached
- before scaling the data it was observed that the data had a positive skew
- Also in the univariate analysis we noticed the presence of outliers and have chosen not the remove them since they can help in the identification of the possible defaulters
-

Model building:

Assumptions of logistic regression:

1. The logistic regression assumes that it is minimal or no multicollinearity among the independent variables.

2. The Logistic regression assumes that the independent variables are linearly related to the log of odds.

3. Logistic regression usually requires a large sample size to predict properly.

4. The Logistic regression which has two classes assumes that the dependent variable is binary and ordered logistic regression requires the dependent variable to be ordered. (here we have taken the loan status as our binary target column)

5. The Logistic regression assumes the observations to be independent of each other.

we are using logistic regression: the reason being the target column is a categorical dtype

Train and Test data:

```
print('X_train', X_train.shape)
print('y_train', y_train.shape)

# print dimension of test set
print('X_test', X_test.shape)
print('y_test', y_test.shape)
```

```
X_train (425929, 33)
y_train (425929, 1)
X_test (106483, 32)
y_test (106483,)
```

the model score is currently: 0.780

Logistic Regression Summary:

Logit Regression Results

Dep. Variable:	loan_status	No. Observations:	425929
Model:	Logit	Df Residuals:	425896
Method:	MLE	Df Model:	32
Date:	Fri, 22 Jul 2022	Pseudo R-squ.:	0.1820
Time:	13:01:14	Log-Likelihood:	-1.9051e+05
converged:	False	LL-Null:	-2.3290e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-16.6707	0.147	-113.609	0.000	-16.958	-16.383
loan_amnt	0.0003	1.76e-05	19.453	0.000	0.000	0.000
funded_amnt	0.0003	2.58e-05	10.043	0.000	0.000	0.000
funded_amnt_inv	-0.0006	1.82e-05	-32.410	0.000	-0.001	-0.001
term	-0.4677	0.012	-39.576	0.000	-0.491	-0.444
int_rate	0.5464	0.005	117.194	0.000	0.537	0.556
emp_length	0.0008	0.001	0.729	0.466	-0.001	0.003
annual_inc	-1.065e-07	7.76e-08	-1.372	0.170	-2.59e-07	4.56e-08
verification_status	-0.1506	0.009	-16.320	0.000	-0.169	-0.133
dti	-0.0341	0.001	-58.722	0.000	-0.035	-0.033
delinq_2yrs	-0.1909	0.006	-32.992	0.000	-0.202	-0.180
inq_last_6mths	0.1356	0.004	32.636	0.000	0.127	0.144
mths_since_last_delinq	-0.0037	0.000	-20.036	0.000	-0.004	-0.003
open_acc	-0.0439	0.001	-37.725	0.000	-0.046	-0.042
pub_rec	-0.3223	0.009	-34.555	0.000	-0.341	-0.304
revol_bal	4.176e-06	4.11e-07	10.161	0.000	3.37e-06	4.98e-06
revol_util	-0.0075	0.000	-34.288	0.000	-0.008	-0.007
total_acc	0.0274	0.000	55.965	0.000	0.026	0.028
initial_list_status	-0.7697	0.009	-87.336	0.000	-0.787	-0.752
total_rec_int	-0.0002	3.91e-06	-48.012	0.000	-0.000	-0.000
total_rec_late_fee	-0.0087	0.001	-7.450	0.000	-0.011	-0.006
recoveries	-1.9192	115.970	-0.017	0.987	-229.216	225.378
collection_recovery_fee	-231.3463	4330.291	-0.053	0.957	-8718.560	8255.868
collections_12_mths_ex_med	-0.6024	0.043	-14.022	0.000	-0.687	-0.518
last_week_pay	0.0092	0.000	70.476	0.000	0.009	0.009
acc_now_delinq	-0.2480	0.064	-3.892	0.000	-0.373	-0.123
tot_coll_amt	-5.73e-05	4.05e-06	-14.157	0.000	-6.52e-05	-4.94e-05
tot_cur_bal	1.171e-07	3.33e-08	3.511	0.000	5.17e-08	1.82e-07
total_rev_hi_lim	-3.494e-06	3.06e-07	-11.436	0.000	-4.09e-06	-2.9e-06
grade_num	0.1660	0.002	105.933	0.000	0.163	0.169
addr_state_NE	0.0880	0.013	6.531	0.000	0.062	0.114
addr_state_S	0.1132	0.012	9.311	0.000	0.089	0.137
addr_state_W	0.2736	0.013	21.708	0.000	0.249	0.298