Lending club case study

Vivek Prahlada, Soham Halbandge
10/08/2022

Problem statement

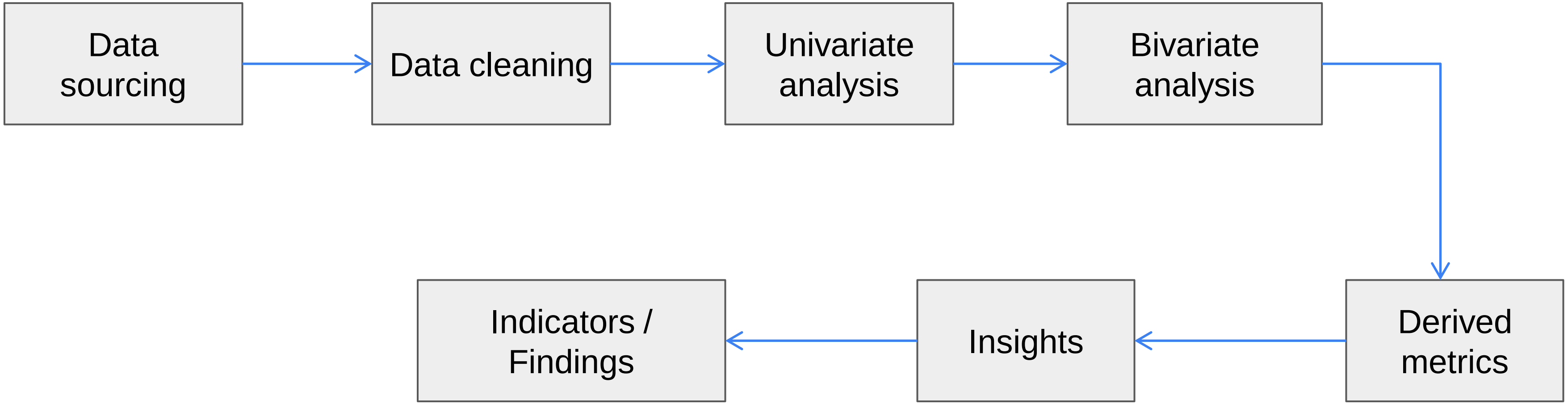
- Financial institution(Lending club) wishes to reduce the credit loss. Data from the previous granted/approved applicants is provided with may dimensions and values
- Objective: Identify the indicators of potential defaulters from the behavior of the defaulters in the data set
- Findings or indicators may be then used by institution to reduce the credit loss

Data description

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 111 columns):
```

- It is a data set with 111 dimensions and 39717 values provided in .csv format
- Headers are clear and no additional or extra headers, page breakers are observed

EDA steps



Data sourcing
Data cleaning
Univariate analysis
Bivariate analysis
Derived metrics
Insights
Indicators / Findings

Data cleaning : Fixing rows and columns

```
mths_since_last_major_derog
annual_inc_joint
dti_joint
verification_status_joint
tot_coll_amt
tot_cur_bal
open_acc_6m
open_il_6m
open_il_12m
open_il_24m
mths_since_rcnt_il
total_bal_il
il_util
open_rv_12m
open_rv_24m
max_bal_bc
all_util
total_rev_hi_lim
inq_fi
total_cu_tl
inq_last_12m
acc_open_past_24mths
avg_cur_bal
bc_open_to_buy
bc_util
mo_sin_old_il_acct
mo_sin_old_rev_tl_op
mo_sin_rcnt_rev_tl_op
mo_sin_rcnt_tl
mort_acc
mths_since_recent_bc
mths_since_recent_bc_dlq
mths_since_recent_inq
mths_since_recent_revol_delinq
num_accts_ever_120_pd
num_actv_bc_tl
num_actv_rev_tl
num_bc_sats
num_bc_tl
num_il_tl
```

- 54 columns (mentioned to the left) contains only missing values. All of them shall be dropped
- 3 columns (mentioned to the right) have more than 60% missing values. Cannot be imputed. All of them shall be dropped
- One row where the values are shifted. This shall be deleted
- Missing values in other columns are simply ignored as the python and libraries handles the operations with missing values in the series

```
mths_since_last_delinq
mths_since_last_record
next_pymnt_d
```

Data cleaning: Filtering of unwanted dimensions

Below variables are assumed to be very less or no use for the stated objective and hence are not considered for analysis further. They are not deleted to keep the option open if needed during bivariate analysis

delinq_amnt --> Number of delinquency makes more sense than amount of delinquency

earliest_cr_line --> Earliest credit line doesnt provide the information about defaulters or in worst case mislead the analysis. If customer uses a credit card which is quite normal and this can lead in opening of credit line

funded_amnt, funded_amnt_inv --> Since the entire data is of accepted application, sum of funded amount and funded amount investors will be reflected in total loan amount (requested loan amount). This information will not be available at the time of processing loan application

initial_list_status --> This provides information of wether or not completely funded by investors or by lending clud. Therefore it doesnt help us in finding indicators of potential defaulters. This information will not be available at the time of processing loan application

installment --> Monthly installment is function of interest and total loan amount. Any impact of this installment is reflected categories of total loan amount and interest rate

last_credit_pull_d --> Month when the credit history was pulled. This information is assumed to be used to avoid pulling more frequent credtis. Therefore, it doesnt help the objective

last_pymnt_amnt, last_pymnt_d --> Will not help objective. This information will not be available at the time of processing loan application

open_acc --> Open account depends on different credit options the applicant has. This is usually open. Hence will not help the objective

out_prncp --> This is non zero only for current and charged off applicants. Since current is not important and charged off is identifiable by loan_status variable. This doesnt help. Also, This information will not be available at the time of processing loan application

out_prncp_inv --> same as out_prncp

recoveries --> Doesn't help. This is non zero for only charged off account. No objective finding behaviors of charged off applicant

revol_bal, revol_util --> Doesnt help the objective. Not very clear about the meaning of these dimensions

title --> This is provided by borrower. Non standard values. Difficult to analyse. Purpose variable better serves the need

total_acc --> Same as open account

total_pymnt, total_pymnt_inv, total_rec_int, total_rec_prncp --> same as funded amount

total_rec_late_fee --> This doesnt give pattern for a new applicant. This information will not be available at the time of processing loan application

zip_code --> Additional information addr_state

issue_d --> It is assumed that this variable doesnt provide the insight towards objective

Univariate analysis

- Values of the below dimensions are not record or has only one unique value. Hence deleted
    - Collections_12_mths_ex_med
    - Tax_liens
    - pymnt_plan

- Desc, url columns doesnt provide information needed to achieve objective hence deleted

- Collection_recovery_fee is applicable for the loans that are charged off and doesnt provide any information about the defaulter behavior. Thus removed

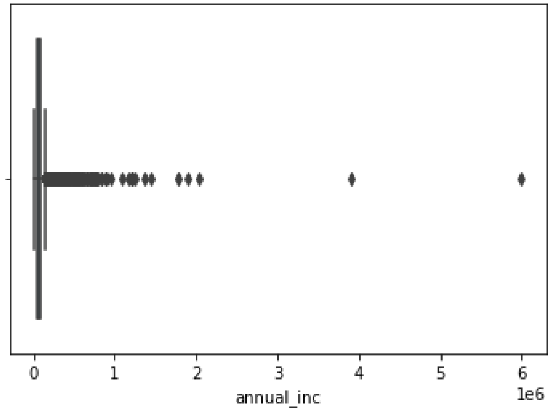
- Id, member_id are unique values and used to find the duplicate entries. None found. Post this these variables doesnt provide any information about he behavior of the defaulter. Hence deleted

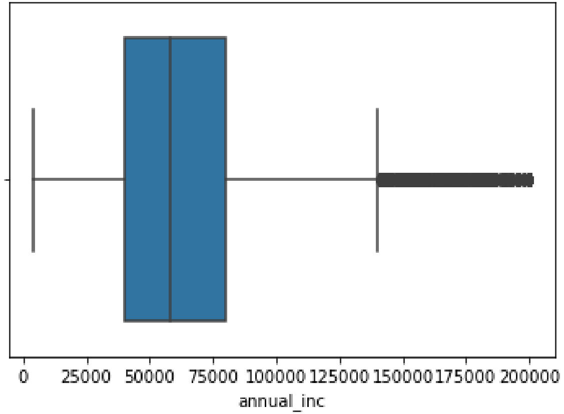
- Emp_title : Employee title is an nominal categorical variable with lot of categories. Also this needs lot of cleaning as same category value is expressed in different ways. This will not help in analysis and hence not considered

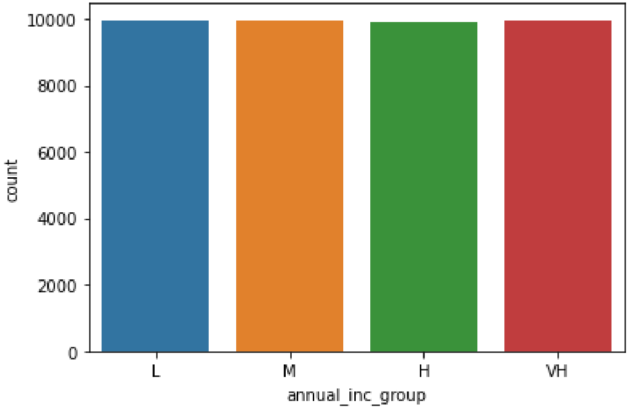
- Statistically(box plots) outliers are observed but they seem not outliers from business context and hence not treated

Univariate analysis/derived metrics

Annual income:



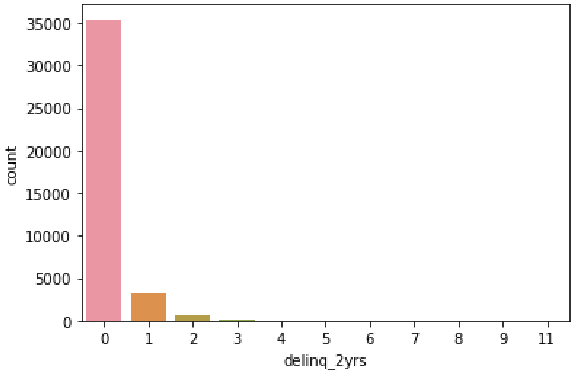




- Four groups per quartile of the annual income is created to support categorical analysis
- These groups are further used in bi-variate analysis
- No insight towards objective can be directly understood from this univariate analysis

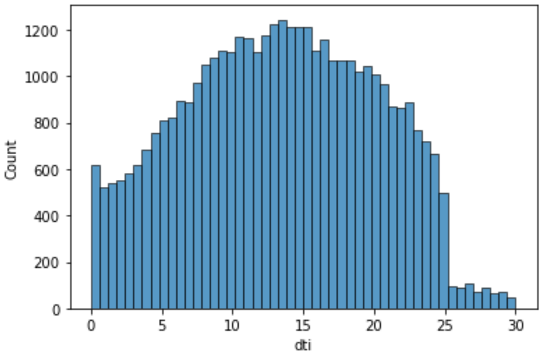
Univariate analysis/derived metrics

delinq_2yrs



- Data spread is highly imbalanced. Hence derived 4 groups of no delinq, 1 delinq, 2 delinq and >2 delinq
- No insight towards the objective can be derived

dti



- Cannot be directly compared with categorical variables. Hence 5 equal groups of VL, L, M, H and VH equal buckets are formed
- No insight towards the objective can be derived

Univariate analysis/derived metrics

inq_last_6mths

```
0    19300
1    10970
2     5812
3     3048
4      326
5      146
6       64
7       35
8       15
```

- Data spread is highly imbalanced. Hence derived 4 groups of no inquiry, 1 inquiry, 2 inquiry and more than 2 inquiry
- No insight towards the objective can be derived

loan_amnt

```
count      586.000000
mean     19370.776451
std       9683.917646
min       1000.000000
25%      12000.000000
50%      20000.000000
75%      25000.000000
max      35000.000000
Name: loan_amnt, dtype: float64
```

- Cannot be directly compared with categorical variables. Hence 4 equal groups of L, M, H and VH equal buckets are formed
- No insight towards the objective can be derived

Univariate analysis/derived metrics

pub_rec

count	39716.000000
mean	0.055066
std	0.237203
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	4.000000

0	37600
1	2056
2	51
3	7
4	2

- Data spread is highly imbalanced. Hence derived 2 groups of no public record and public record are created
- No insight towards the objective can be derived

int_rate

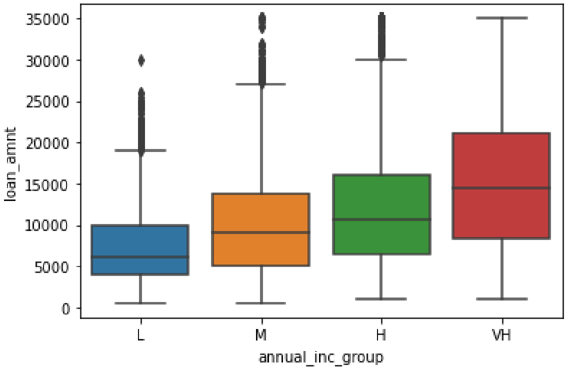
count	39716.000000
mean	12.021108
std	3.724847
min	5.420000
25%	9.250000
50%	11.860000
75%	14.590000
max	24.590000

Name: int_rate_conv, dtype: float64

- Cannot be directly compared with categorical variables. Hence 5 equal groups of VL, L, M, H and VH equal buckets are formed
- No insight towards the objective can be derived

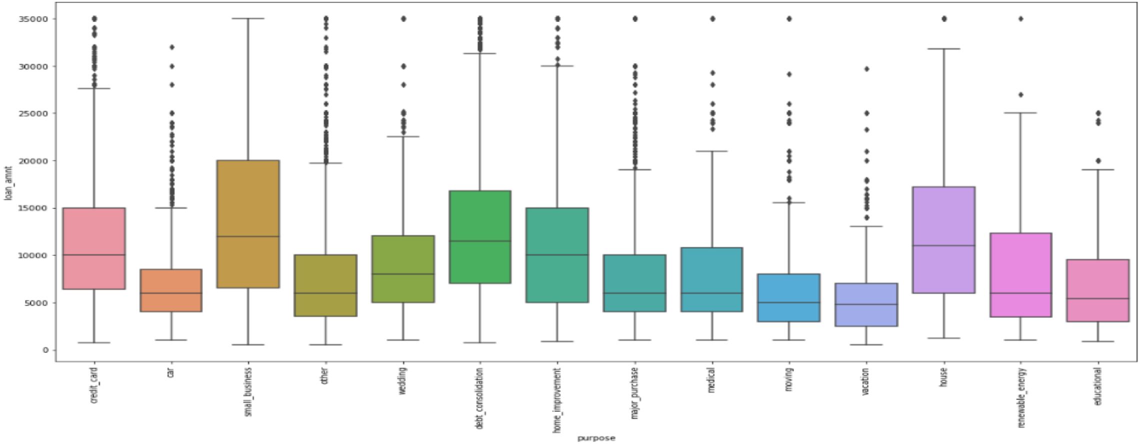
Segmented univariate

Annual income



- Lower the annual income lower the loan amount granted which is plausible

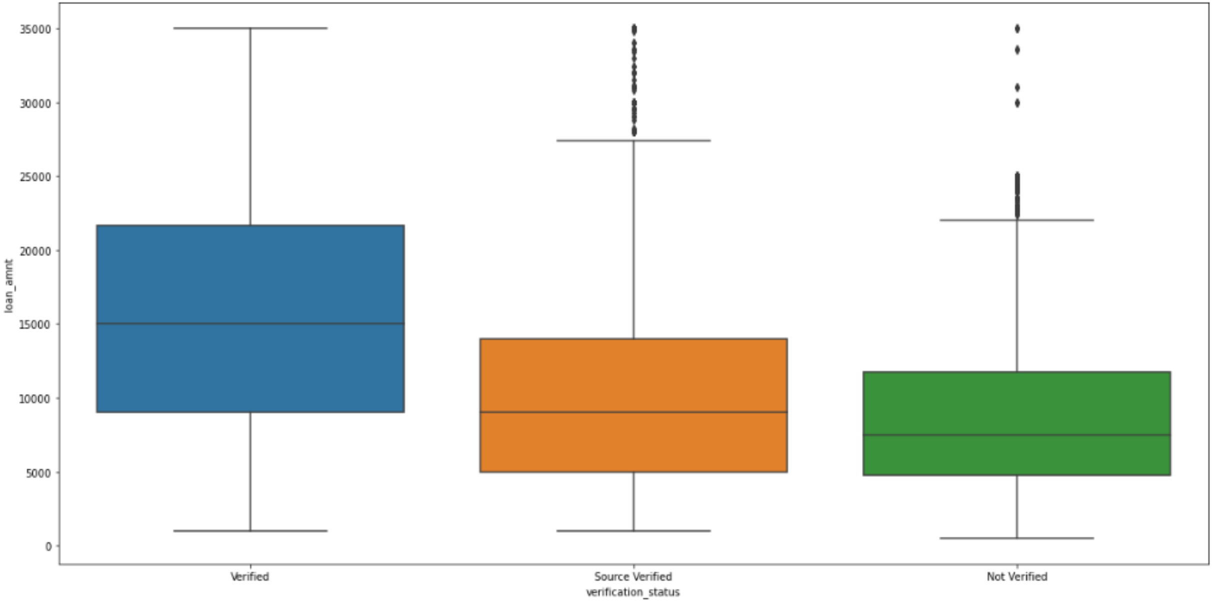
Loan amount



- Small business were granted higher loan amounts together with debt and housing purpose
- Impact of this can be understood in bivariate analysis

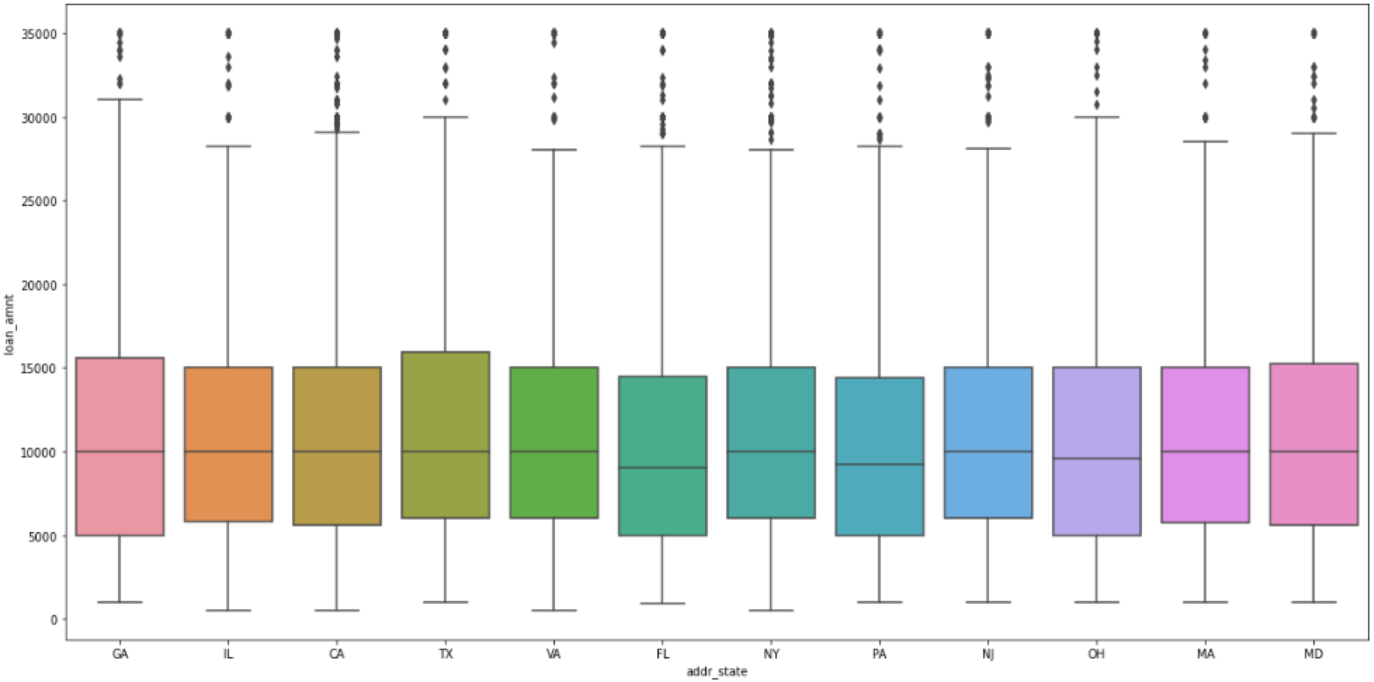
Segmented univariate

Verification status



- Higher loan amount are given to verified applicants compared to not verified. This makes sense
- Impact of this can be analyzed in bivariate analysis

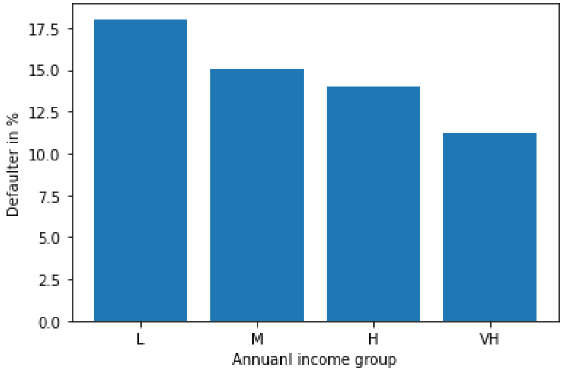
Segmented univariate

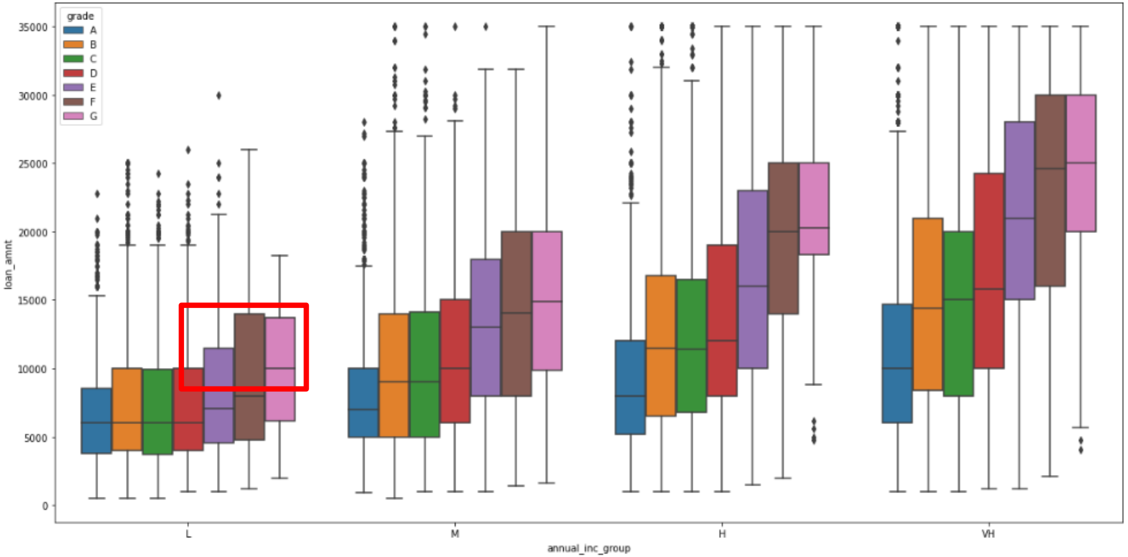


There are almost no difference in the loan amount provided at the different states

Note: states with > 1000 data points are only considered

Bi-variate analysis : Annual income

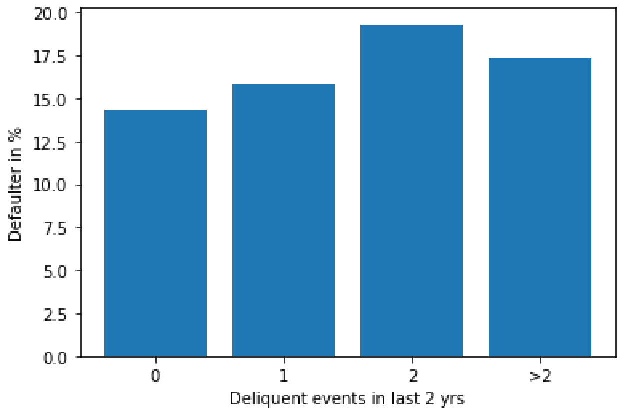




- Lower income group is having higher defaulter percentage
- Hypothesis:

within lower income groups, higher loan amount is provided to risker applicants (grade >D). Even though the pattern is same across the income groups, riskier lower income groups applicants are more vulnerable than riskier higher income group.

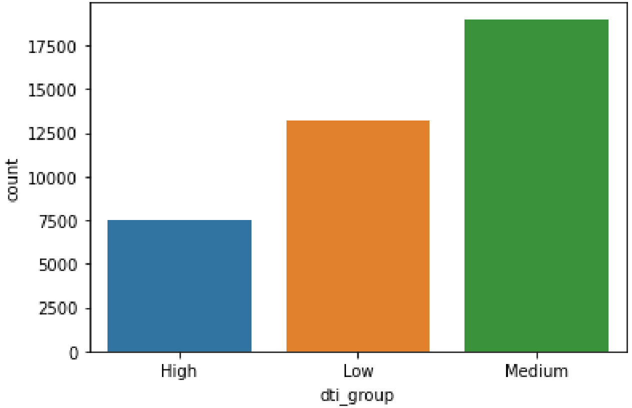
Bi-variate analysis : Delinquency

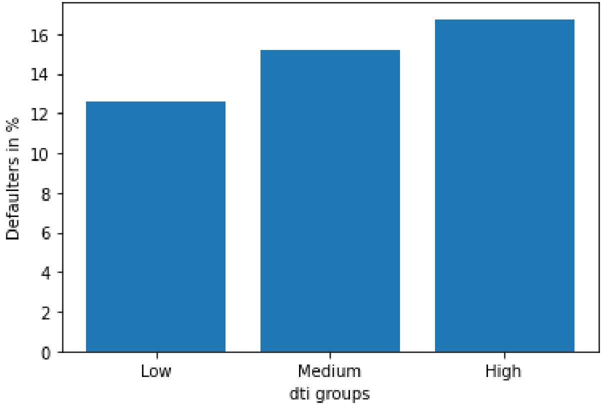


- Higher the delinquency higher is the risk. This is natural as one with the behaviour of late payments is more vulnerable than others.

Note:
>2 migh show less % of defaulters than 2 entries that probably because the number of data points are less understanding that delinq_2yrs data will be available at the time of processing loan application

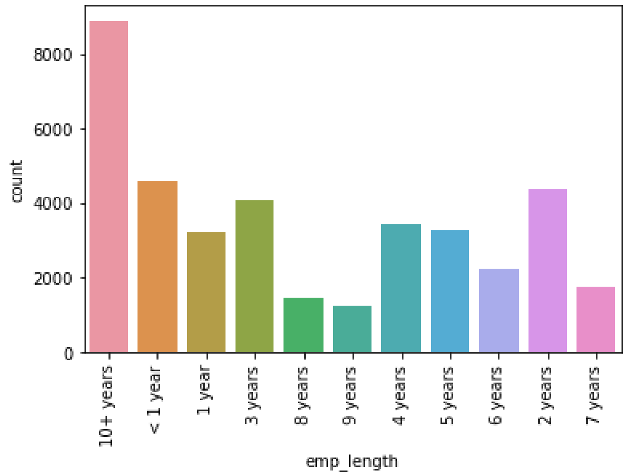
Bi-variate analysis : dti

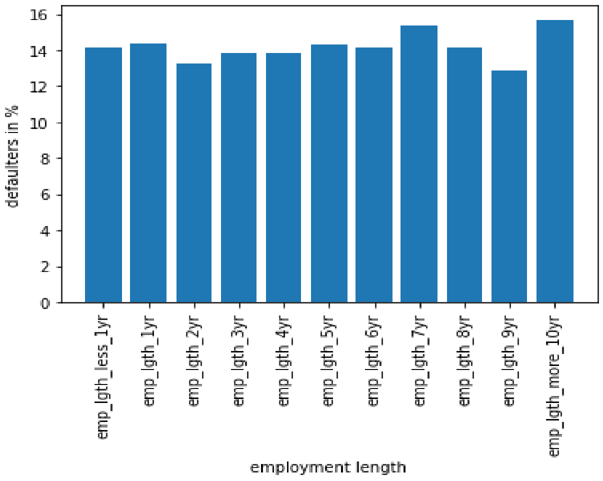




- From the derived grouping of dti, it is clear than higher the dti, higher is the risk of becoming default
- This is sensible that if income is lower than the due payment, they tend to default or divert money to different needs/priorities

Bi-variate analysis : employment length





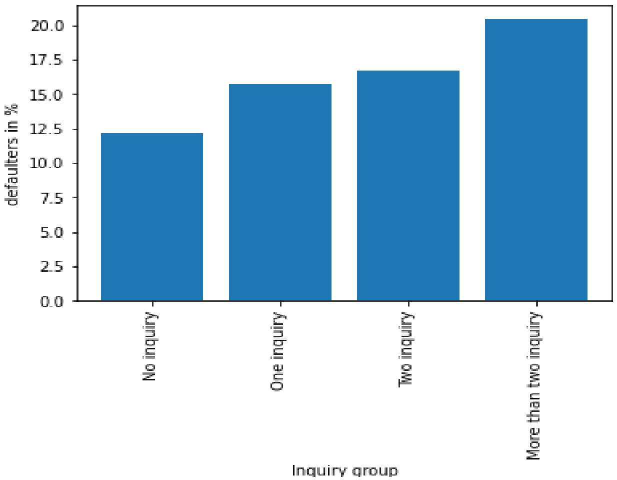
- Data distribution inside the employment length group is uneven
- Looks like employment duration doesn't impact the behaviour of defaulters

Bi-variate analysis: Home ownership

- House ownership is not a strong indicator of charged Off. If the applicant select the house owning status as others, they tend to have higher charged off tendency
- It is better to be careful when the home status is unknown or others

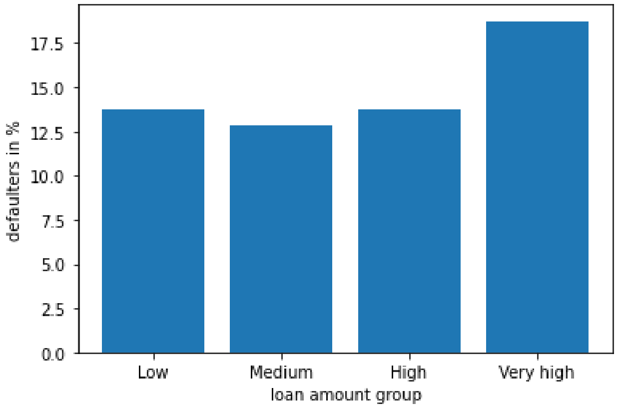
Note:
Other category have lesser data points

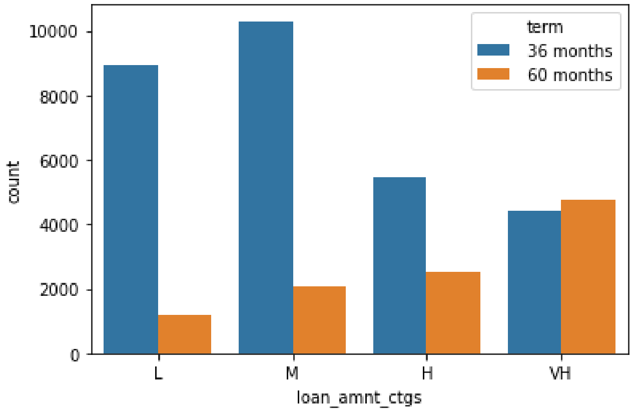
Bi-variate analysis : Inquiries in last 6 months



- It is a clear pattern that higher the number of inquiries in 6 months more is the risk of getting default applicant. It should be noted that the more than two inquiry means that he could also be rejected by other lending firms

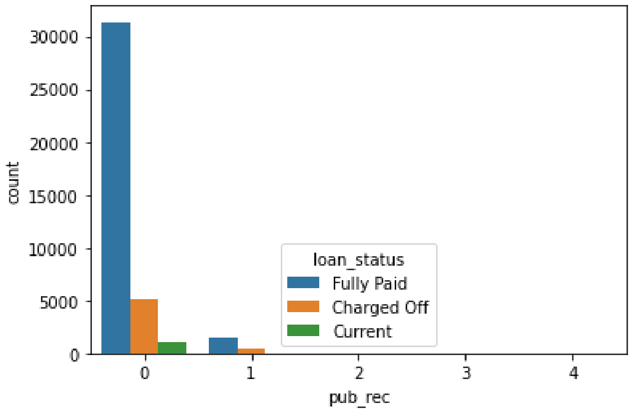
Bi-variate analysis : Loan amount

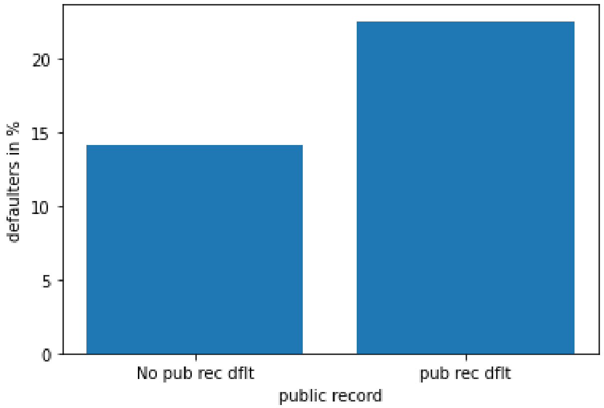




- Defaulters are highest in the very high loan amount category (75 percentile to max value). No reason can be directly understood except for the reason of higher monthly payment
- Also, higher the loan amount longer term starts to become prominent

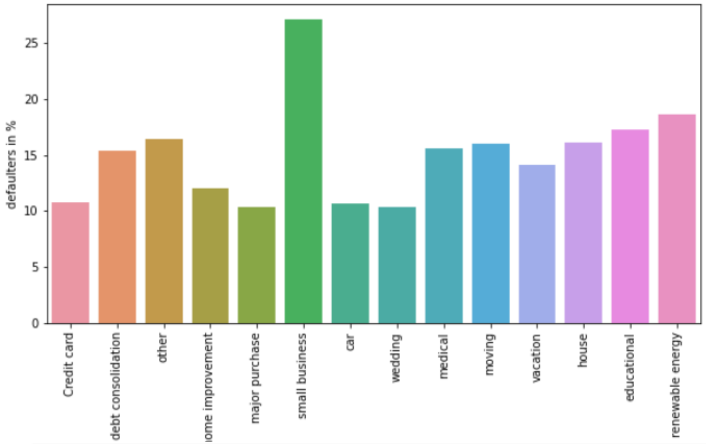
Bi-variate analysis : Public record

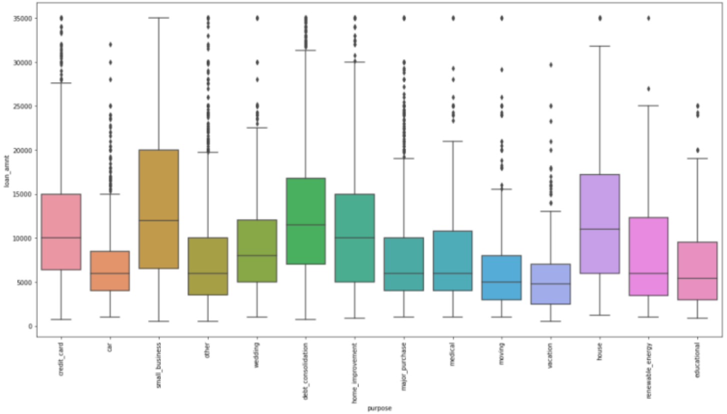


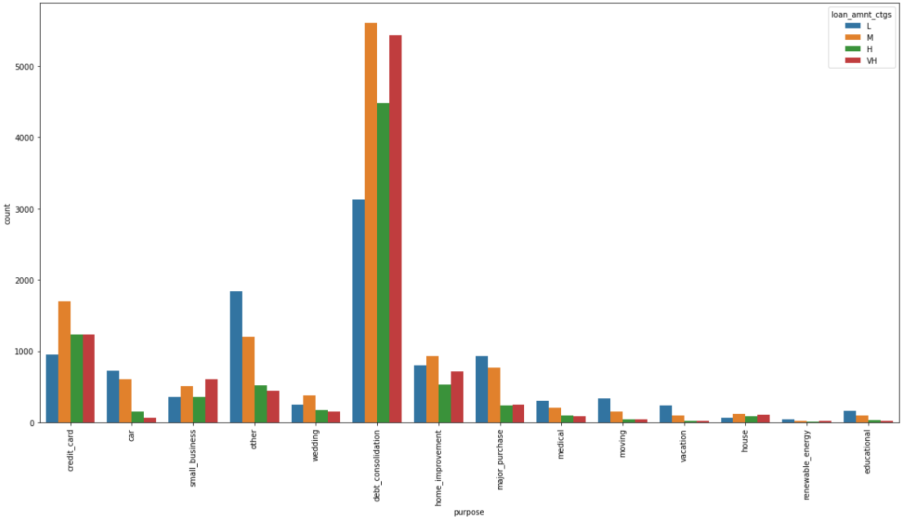


- Defaulter are higher with public derogatory records. This goes with understanding that they are probably less serious or in more difficulty compared to no public record applicants

Bi-variate analysis : Purpose



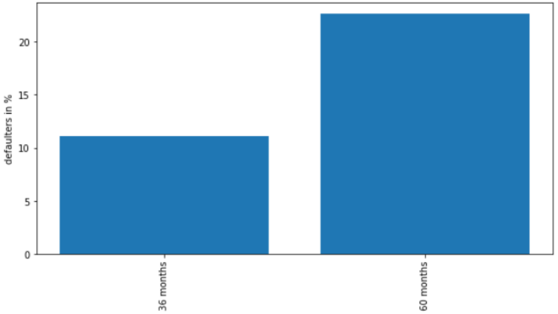


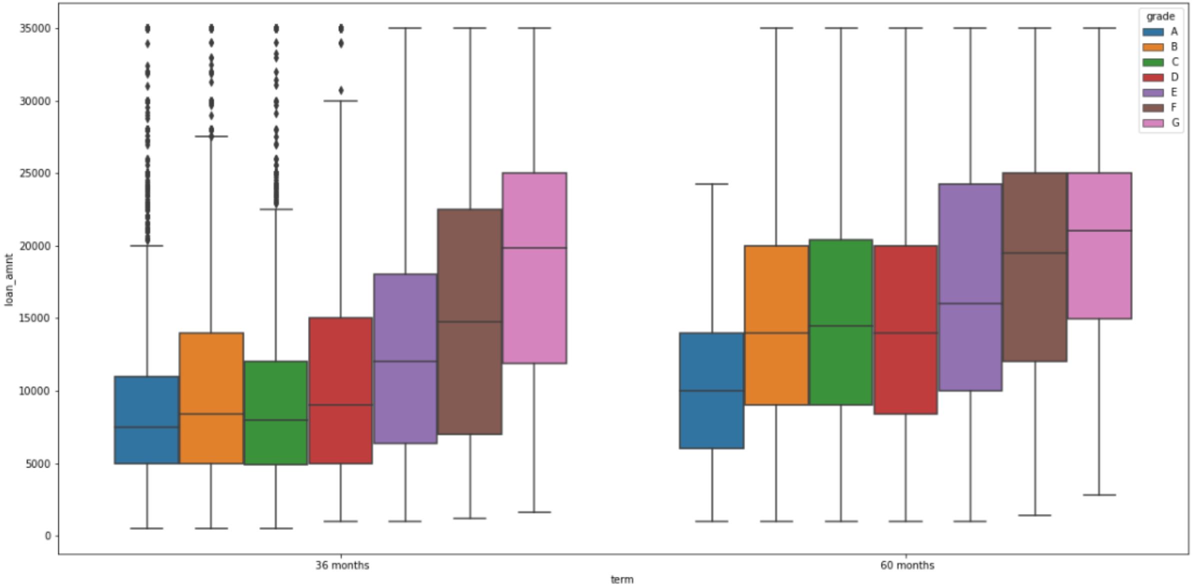


- Small business, house, debt consolidation purpose loans are leading to higher defaulters. Higher loan amounts are sanctioned.

- This may also be reason why higher loan amounts have higher defaulters

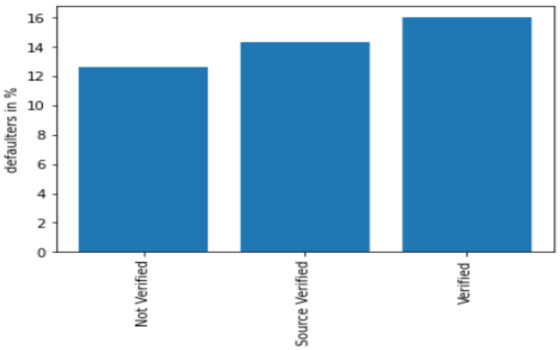
Bi-variate analysis : Term

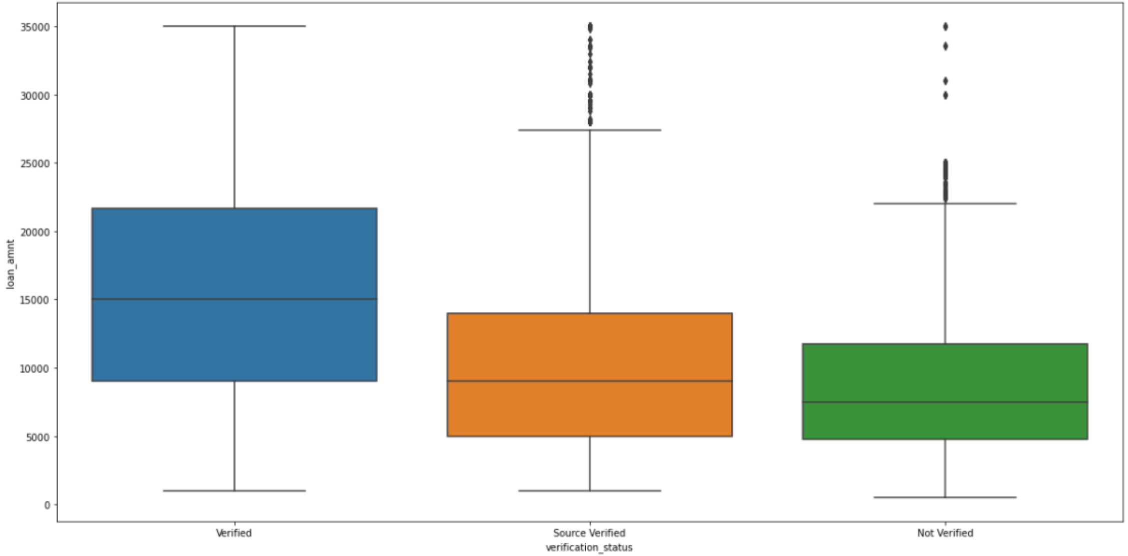




- Defaulters are almost double in the higher term.
- Higher loan amounts are provided B,C,D,E grade applicants in longer term loans there by increasing the risk compared to lower term
- B,C,D,E applicants are risker in long term with higher loan amount. Probably because of the higher interest

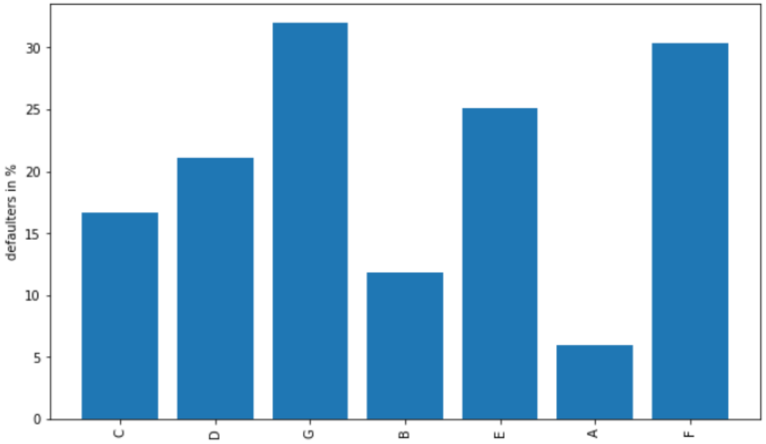
Bi-variate analysis : Verification status

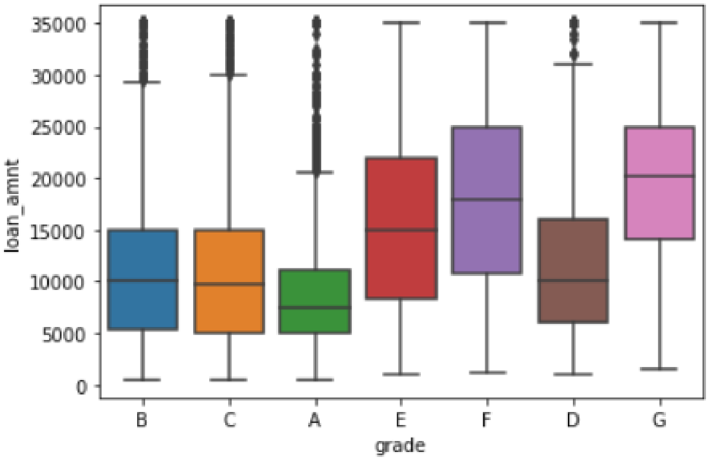




- Surprisingly, verified applicants have higher default percentage
- When analysed with loan amount given, verified applicants are given higher loan amounts. Therefore all the analogy of higher loan amount is applicable here

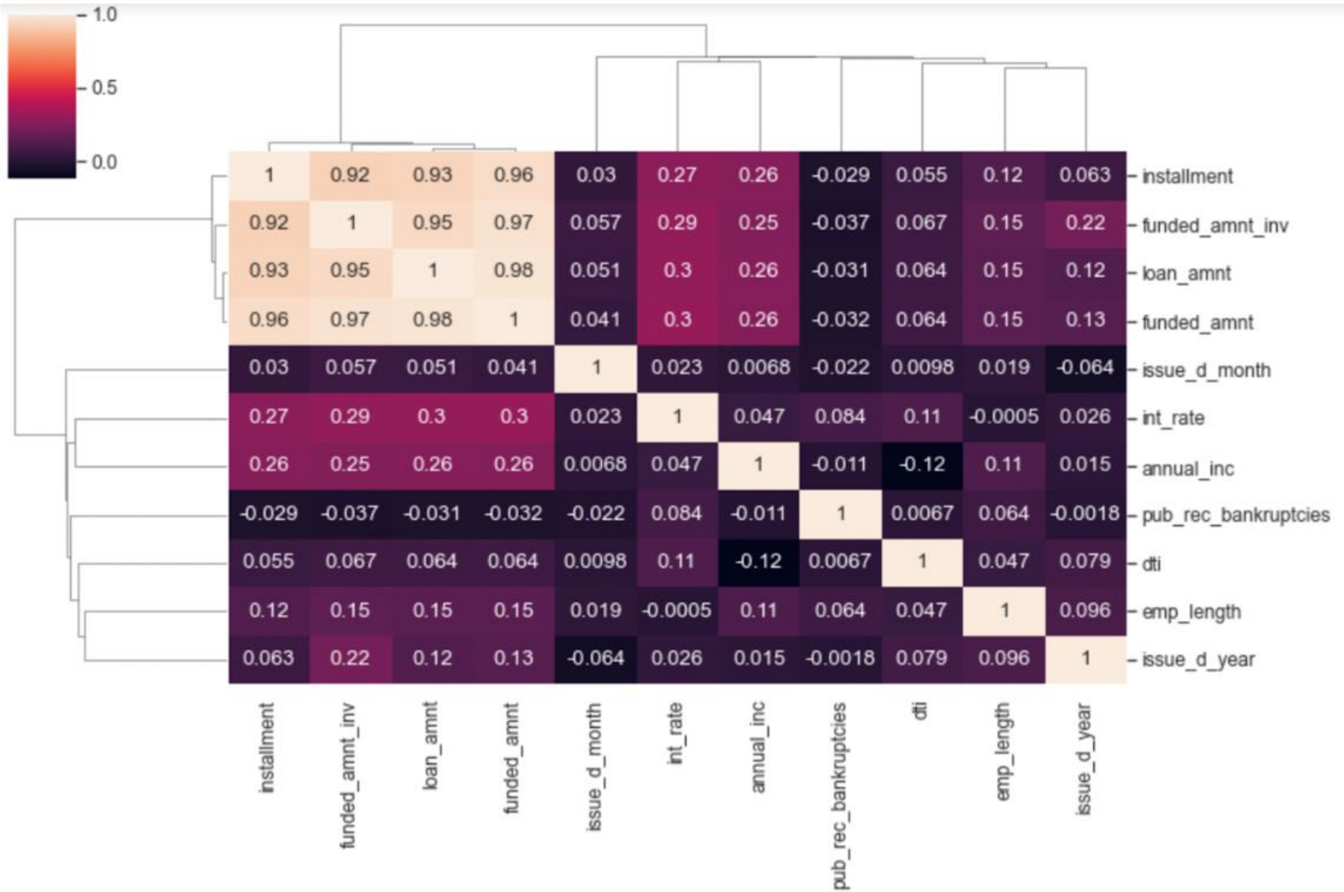
Bi-variate analysis : Grade and interest rate





- Higher the grade higher the risk. This is true and hence the interest rate are also higher for higher grade
- Higher loan amounts are provided for the higher grades making the whole business riskier
- Interest rate is not important for analysis as this is the resultant of the grade, other indicators and probably loan amount. This will not be known at the time of application directly. Hence not so relevant for the study

Correlation check



- Not lot of insights towards objective is available from the insights

Insights

- Small Business, house, debit consolidation and credit card purpose are given higher loan amount
- Loan amount distribution is almost same across the states with more than 1000 data points. This can potentially nullify the impact of any state based frauds
- Employment length doesn't seem to have any impact on defaulting behavior
- Higher the loan amount, longer term becomes prominent compared to lower loan amount
- Higher loan amounts have higher defaulters
- B,C,D,E applicants are risker in long term with higher loan amount. Probably because of the higher interest
- When analysed with loan amount given, verified applicants are given higher loan amounts. Therefore all the analogy of higher loan amount is applicable here
- Higher loan amounts are provided for the higher grades making the whole business riskier as the interest rate are relatively high for higher grades

Findings/Indicators

Findings/indicators	Dimension of interest
Applicant with public derogatory record tends to default than with no record	pub_rec (slide#22)
Small business, renewable energy, educational, house and debt consolidation loans have higher defaulters compared to others and therefor may be riskier	purpose (slide#23)
B,C,D,E grade applicants are provided higher loan amounts for longer duration making the vulnerable as they will have higher interest rate for longer duration and tend to become defaulters. Specially important among low income applicants	annual_inc (slide#15)
Verified customers should not be granted the higher loan amounts straight away. Other indicators shall be considered for the calculation of loan amount	verification (slide#25)
Applicant with higher inquiries within last 6 months tend to be defaulters compared to lesser or no inquiries	inq_last_6mths (slide#20)
Unknown/other house ownership tends highly to default to having correct house ownership status	home_ownership (slide#19)
Higher delinquency events tends to have higher defaulters compared to lower or no delinquency event applicants	delinq_2yrs (slide#16)
Higher the DTI risker is the applicant	dti (slide#17)

Back - up

Findings/Indicators

- Applicant with public derogatory record tends to default than with no record
- Small business, renewable energy, educational, house and debt consolidation loans have higher defaulters compared to others and therefor may be riskier
- B,C,D,E grade applicants are provided higher loan amounts for longer duration making the vulnerable as they will have higher interest rate for longer duration and tend to become defaulters. Specially important among low income applicants
- Verified customers should not be granted the higher loan amounts straight away. Other indicators shall be considered for the calculation of loan amount.
- Applicant with higher inquiries within last 6 months tend to be defaulters compared to lesser or no inquiries
- Unknown/other house ownership tends highly to default to having correct house ownership status
- Higher the DTI risker is the applicant
- Higher delinquency events tends to have higher defaulters compared to lower or no delinquency event applicants