GRAMENER CASE STUDY – Lending Club Dataset

1. *Sudar Abisheck Saravanan*
2. *Ashish Gupta*
3. *Bibhuti Bhusan Sahu*
4. *Sunil Yousef*

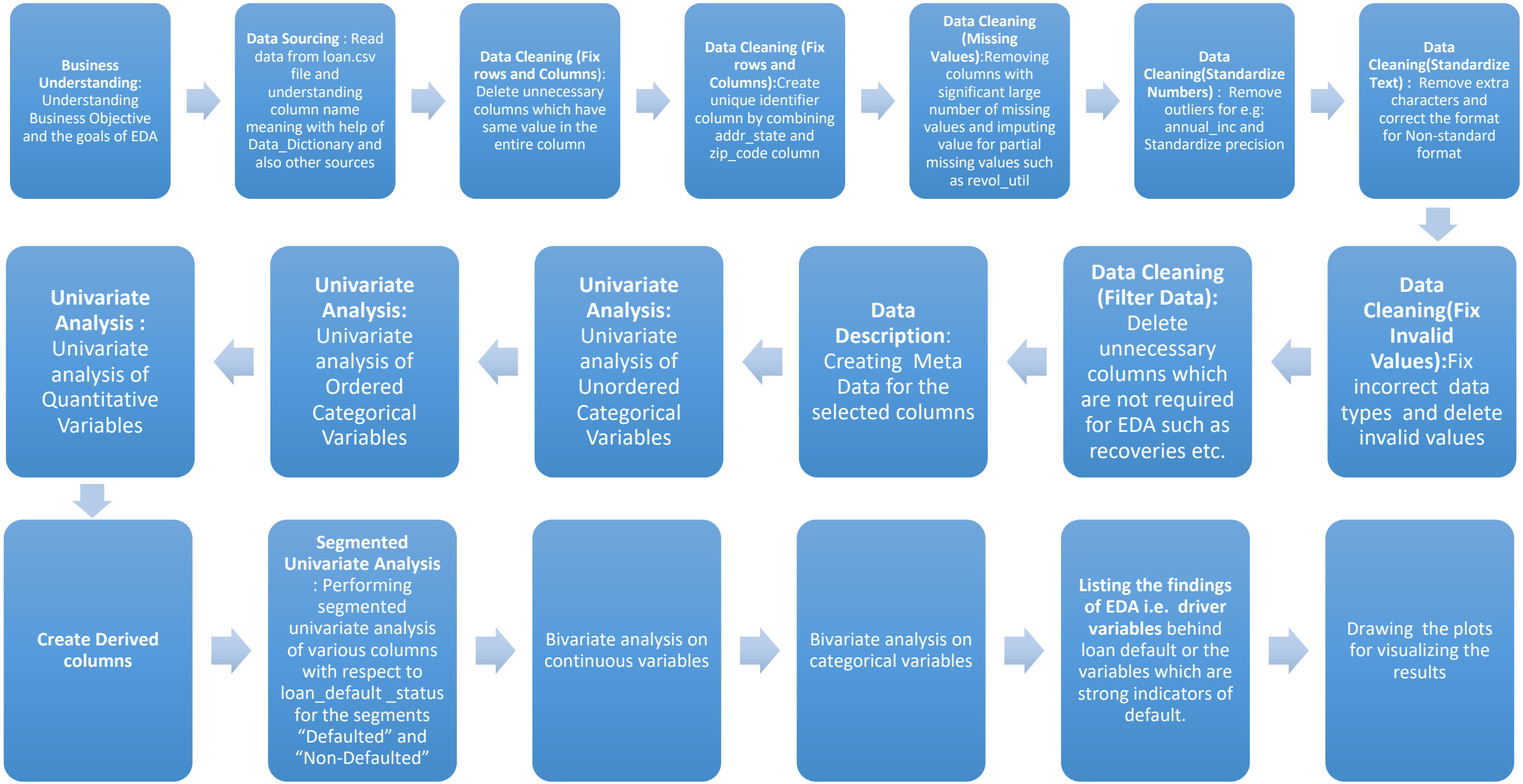
Objective: *To understand the driving factors behind loan default.*

- ✓ To understand key *driving factors (or driver variables)* behind loan default,
  - ✓ This can be used by Lending Club to identify the risky loan applicants who are likely to default.
- ✓ Lending Club can cut down the credit loss by reducing the amount of risky loans.
- ✓ The knowledge gathered by this activity can be used by Lending Club for portfolio and risk assessments to,
  - ✓ Reduce the financial loss for the company if the applicant is not likely to repay the loan.
  - ✓ Reduce the loss of business to the company if the applicant is likely to repay the loan.
- ✓ Use EDA to identify the driving factors of a risky loan applicant.

Assumptions:

- The variables like *recoveries, total_pymnt, total_pymnt_inv, total_rec_prncp* ***etc.***. which normally get captured only after a loan is accepted, will not be available at the time of a new loan application. So these type of variables can be removed from the dataset.
- Since bankruptcy filings, tax liens and judgments are the three kinds of public records that appears on a credit report, this information should already be captured in column *pub_rec* which contains derogatory public records.
  - Number of values in *pub_rec_bankruptcies* is greater then *pub_rec*, Hence we can drop column *pub_rec_bankruptcies*.
- *purpose* and *title* have redundant information.
- *emp_title* column has so much discrepancies in its values(e.g. The same employer name is mentioned in various formats) .Also as it has many unique values it would not give any useful insights about the pattern for loan defaulting.

Problem Solving Approach



Data Sourcing:

- *Get the Loan data set from here*
- *Get the Data Dictionary from here*

Data Cleaning:

- Remove Following Columns which do not provide any meaning to EDA analysis.
  - *id, member_id, Grade, url, funded_amnt, emp_title, earliest_cr_line, pymnt_plan, desc, 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, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, initial_list_status, next_pymnt_d, collections_12_mths_ex_med, policy_code, collection_recovery_fee, inq_last_6mths, out_prncp, out_prncp_inv, tax_liens, delinq_amnt, chargeoff_within_12_mths, acc_now_delinq.*
- Application_type has same value this shall be removed. i.e if any column has only one value then drop it.
- '*purpose*' and *'title*' has redundant information, here 'Purpose' can be used as a categorical variable, so remove 'title' column.
- Remove all columns which has only '**NA**' and/or **0's**.
- Remove all columns with **90% or more zeros and NA's**.
- Convert dates to **python datetime** object.
- '*funded_amnt_inv*', '*installment*', Round to 2 decimal places.
- **Strip Months** from '*term*' Column.

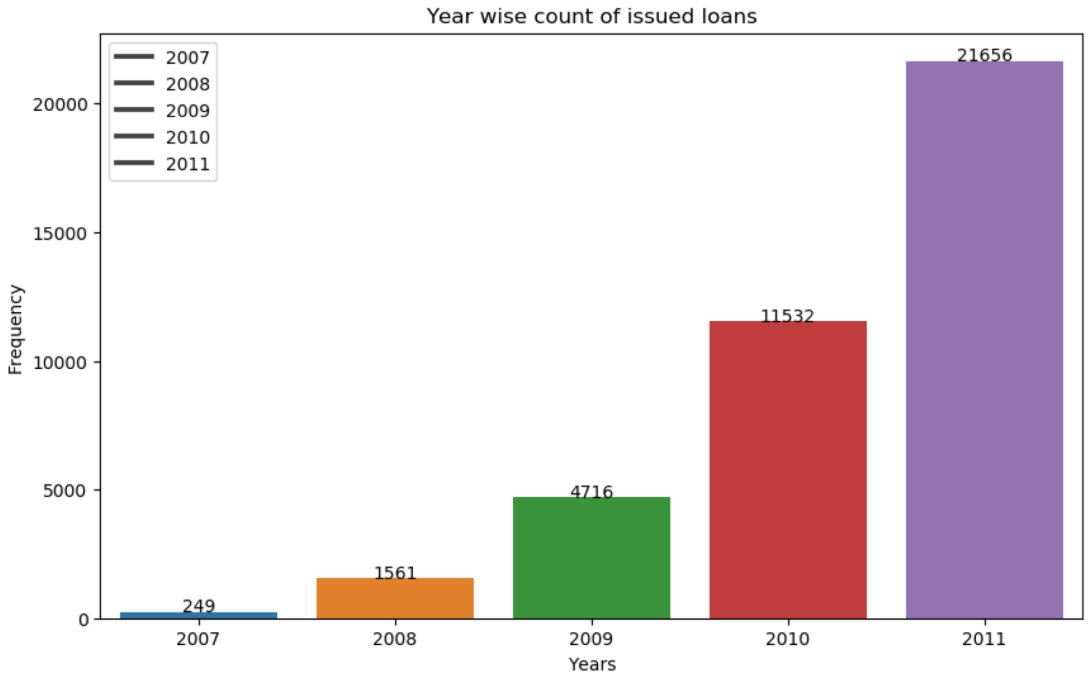
Data Cleaning: *Continued*

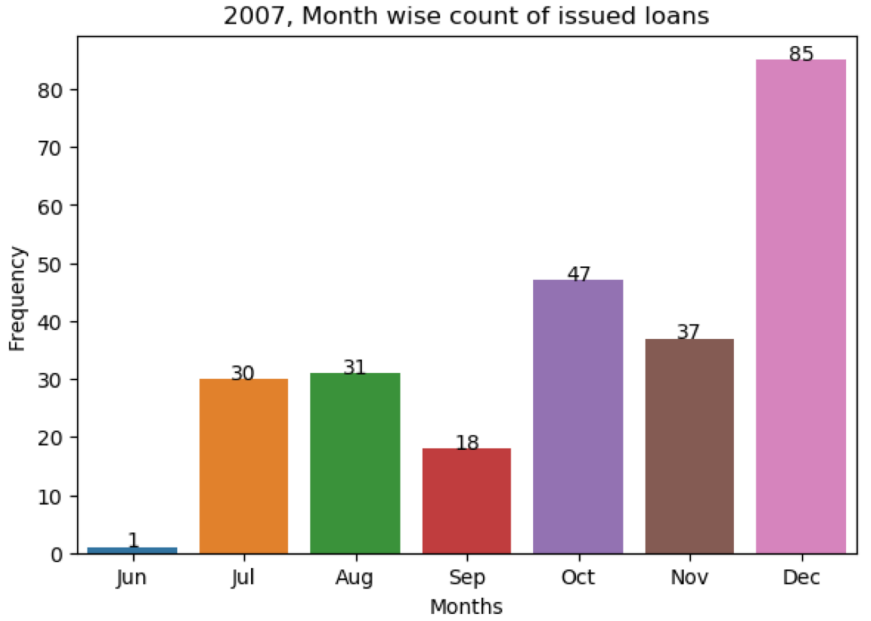
- Strip '%' from '*int_rate*' column.
- Column *emp_length* has substring '+', 'years' strip this. and change '< 1' to 0, keep 10+ as 10.
- '*mths_since_last_delinq*' : Since 65% of data is NaN, which is high, so dropping this column.
- Remove 3 rows which has "*NONE*" as the value in "*home_ownership*" column.
- Remove '*xx*' from '*zip_code*'
- Combine '*zip_code*' and '*addr_state'* to one column as '*address*' Also column like '*total_**" and '*recoveries*' should be removed as per the assumptions.
- Impute the *2.71 % of rows* of '*emp_length'* column, which contain NA values, with the mean value of the column.

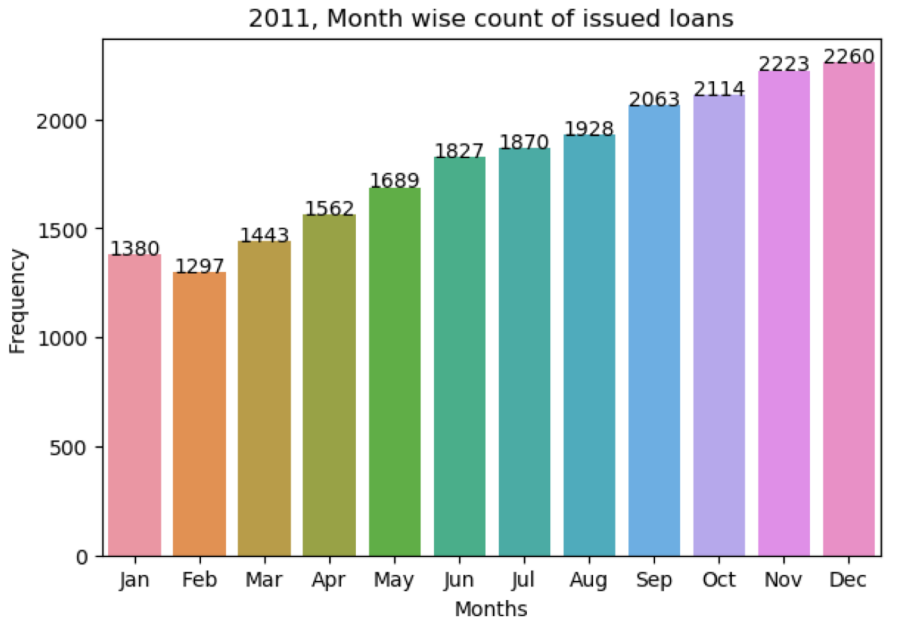
Univariate Analysis: *Metadata*

Column Name	Type	Missing	Uniques	Top	Min	Mean	Median	Impute With	Type of Variables (Ordered/Unorderd Catagorical, Quantitative)
loan_amnt	int64	0	885	35000	500	1.12E+04	10000		Quantitative
funded_amnt_inv	float64	0	7939	35000	0	1.04E+04	8975		Quantitative
term	int64	0	2	60	36	4.24E+01	36		Ordered Categorical
int_rate	float64	0	371	24.59	5.42	1.20E+01	11.86		
installment	float64	0	15381	1305.19	15.69	3.25E+02	280.23		Quantitative
sub_grade	object	0	35	G5	A1	NaN	NaN	NA	Ordered Categorical
emp_length	float64	1075	11	10	0	4.97E+00	4	Mean	Ordered Categorical
home_ownership	object	0	4	RENT	MORTGAGE	NaN	NaN	Rows having "NONE" values are removed	Unordered Categorical
annual_inc	float64	0	5318	6.00E+06	4000	6.90E+04	59000	NA	Quantitative
verification_status	object	0	3	Verified	Not Verified	NaN	NaN	NA	Unordered Categorical
issue_d	datetime64 [ns	0	55	2011-12-01 0:00:00	2007-06-01 0:00:00	NaN	NaN	NA	Ordered Categorical
loan_status	object	0	3	NaN	NaN	NaN	NaN	NA	Unordered Categorical
purpose	object	0	14	NaN	NaN	NaN	NaN	NA	Unordered Categorical
zip_code	object	0	823	NaN	NaN	NaN	NaN	NA	Unordered Categorical
addr_state	object	0	50	NaN	NaN	NaN	NaN	NA	Unordered Categorical
dti	float64	0	2868	29.99	0	1.33E+01	13.4		Quantitative
delinq_2yrs	int64	0	11	11	0	1.47E-01	0		Ordered Categorical
mths_since_last_delinq	float64	0	96	120	-1	1.20E+01	-1		Ordered Categorical
open_acc	int64	0	40	44	2	9.29E+00	9		Ordered Categorical
pub_rec	int64	0	5	4	0	5.51E-02	0		Ordered Categorical
revol_util	object	50	1089	NaN	NaN	48.8	49.3	Median	Quantitative
pub_rec_bankruptcies	float64	694	3	2	0	4.33E-02	0		Ordered Categorical

Univariate Analysis: *Loans issued*



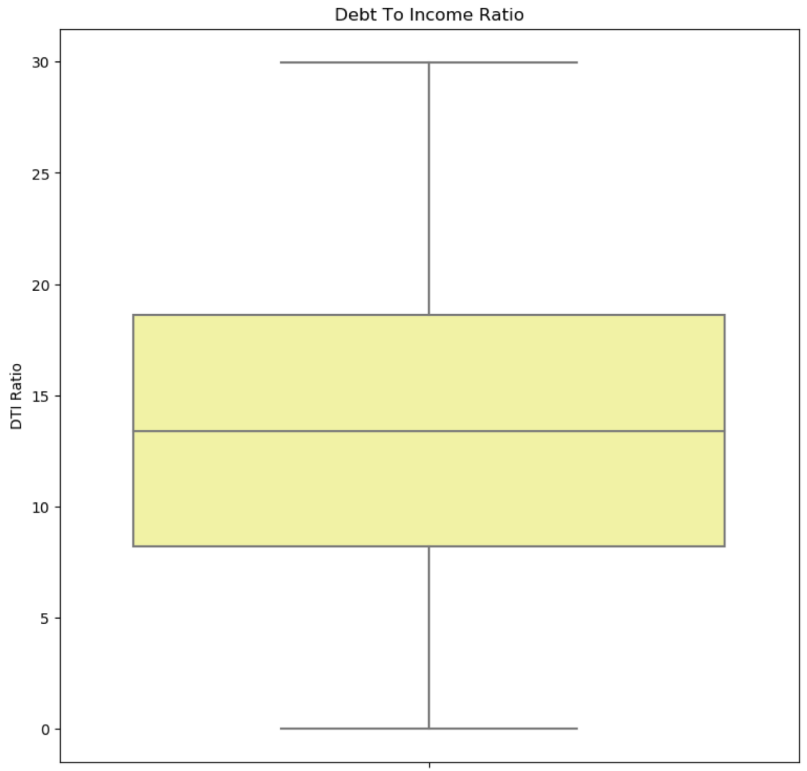


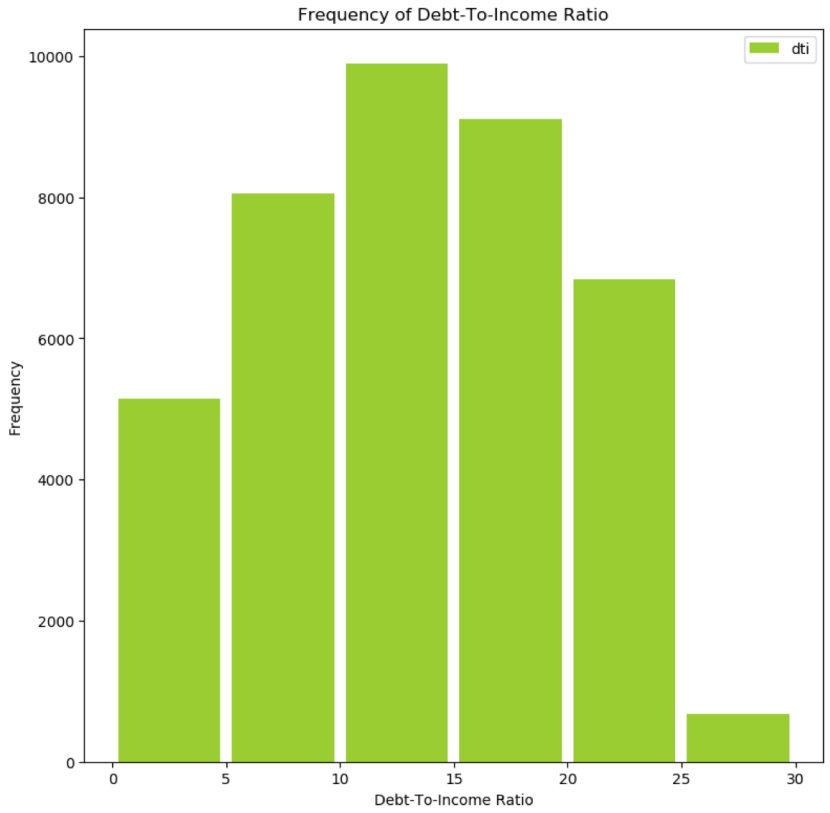


Inference: *The Number of issued loans increases year by year.*

Percentage increase for year 2007 = 0 %
Percentage increase for year 2008 = 527 %
Percentage increase for year 2009 = 202 %
Percentage increase for year 2010 = 145 %
Percentage increase for year 2011 = 88 %

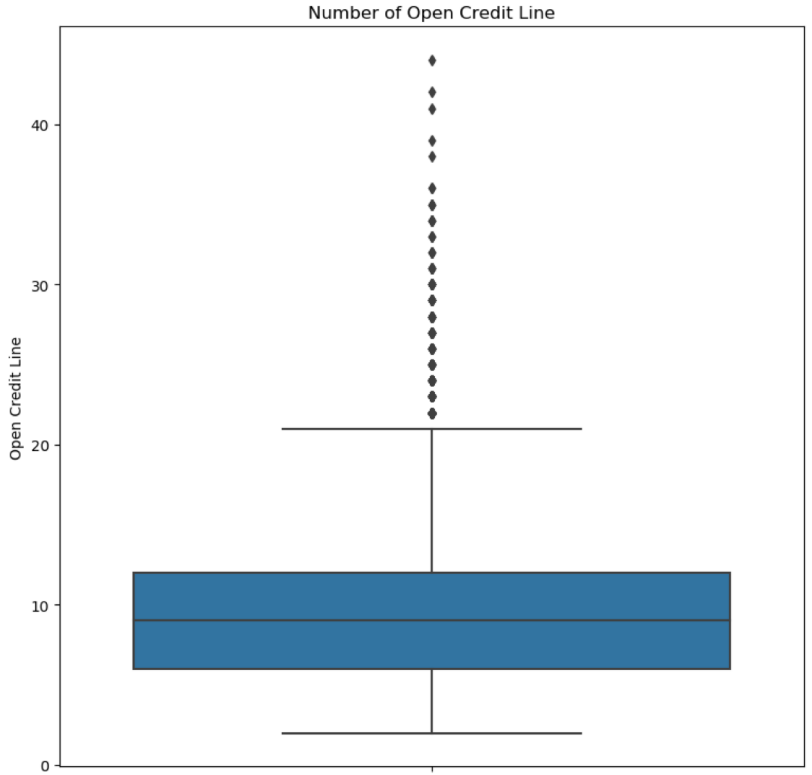
Univariate Analysis: *Debt-to-Income Ratio*

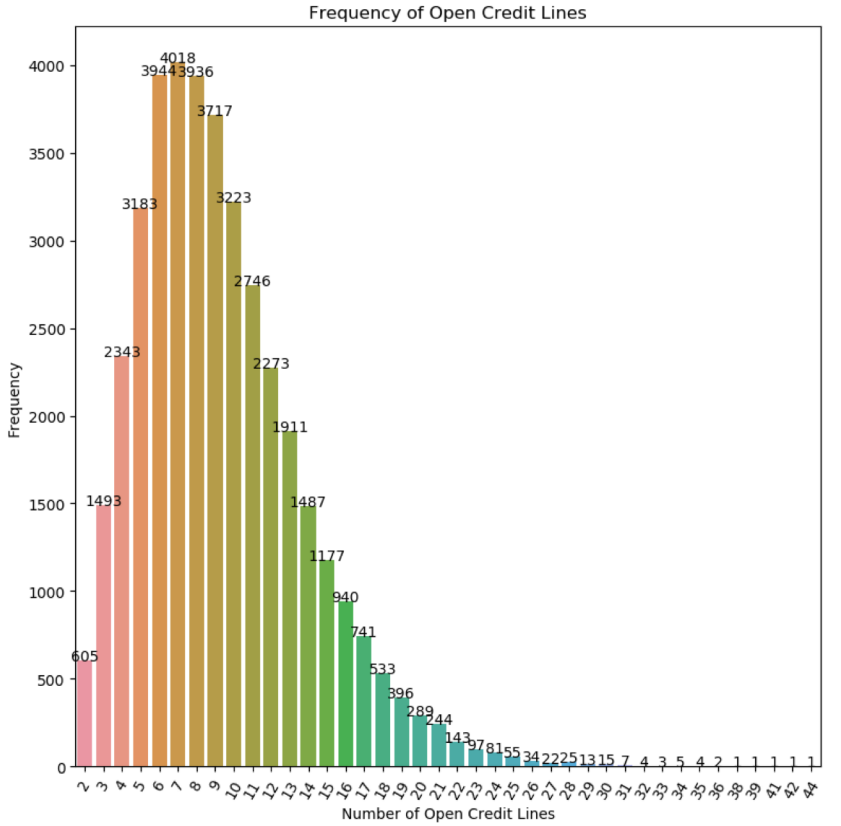




Inference: *dti values are equally distributed along the median between 25-75 percentile with max value, Approximately at 30, minimum at 0 and maximum borrowers given loan having DTI value between 10-15%*

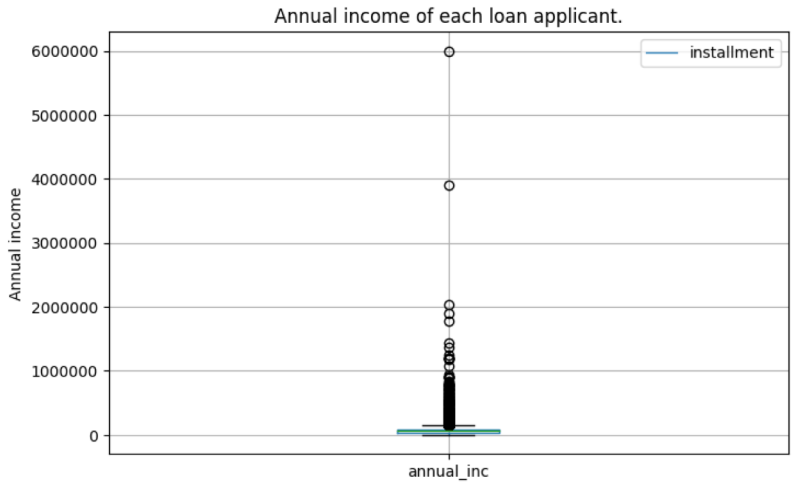
Univariate Analysis: *Open Credit Lines*



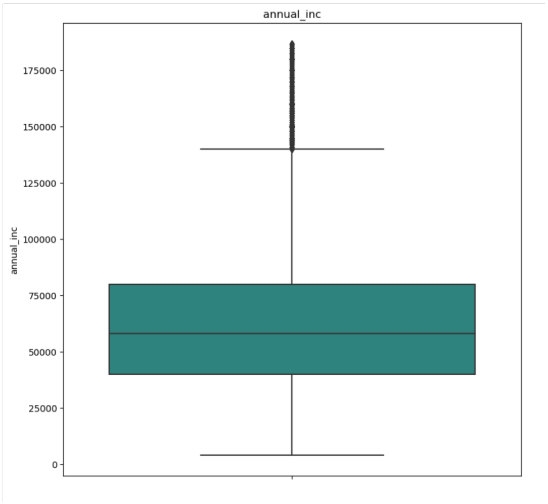


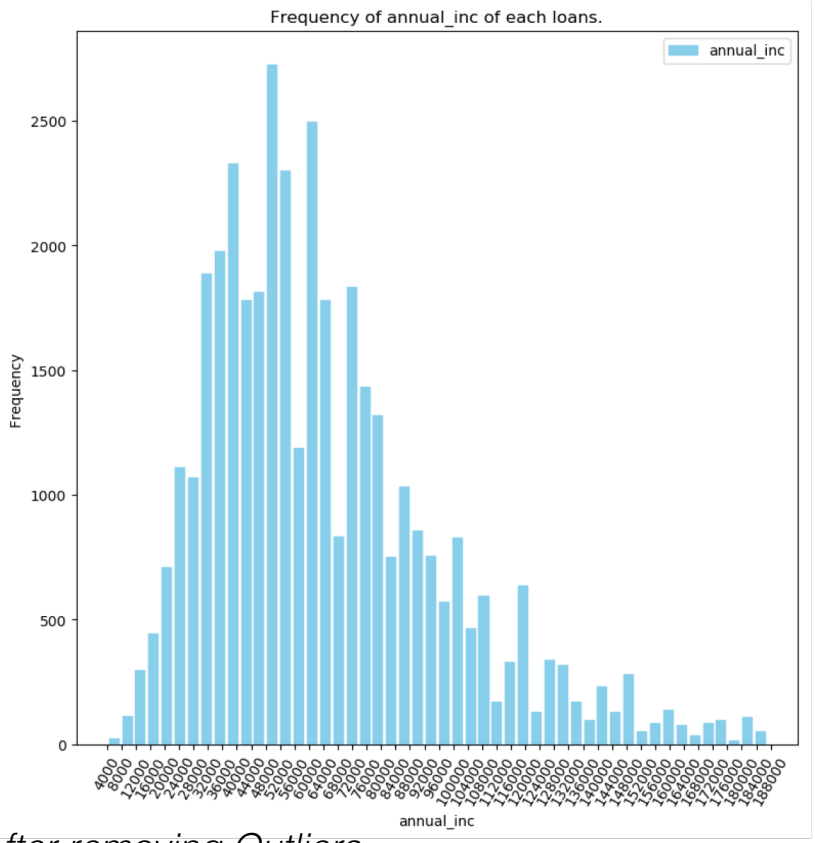
Inference: *open_acc values seems equally distributed along the median(9) for 25-75 percentile with some outliers.*

Univariate Analysis: *Annual Income*



Annual Income – With Outliers



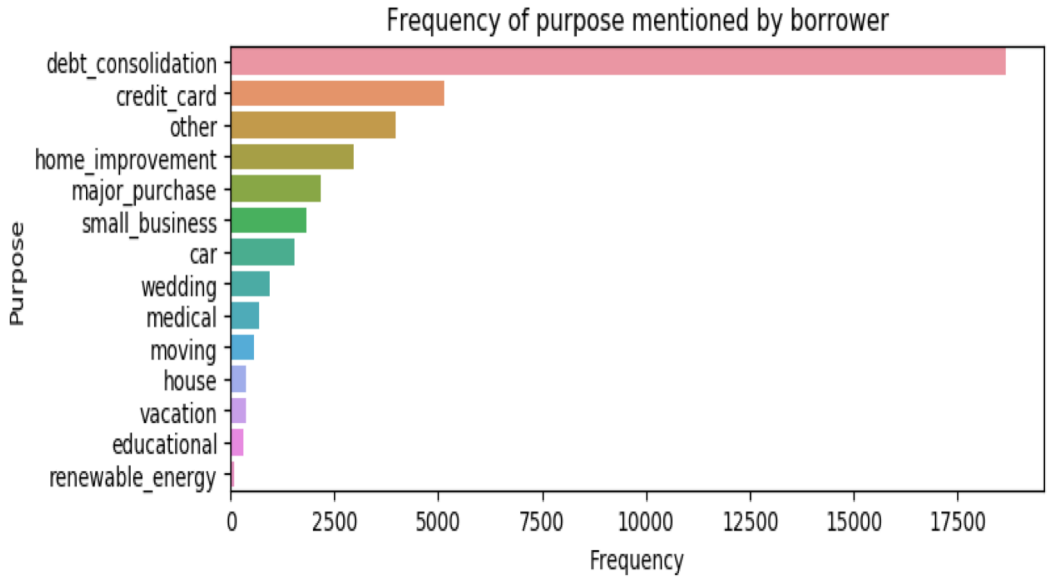


Annual Income – After removing Outliers

Inference: From the 'Annual income' column, the mode of Annual income is 60000.0.
Since there are clearly some outliers above 98 percentile, this has to be removed.

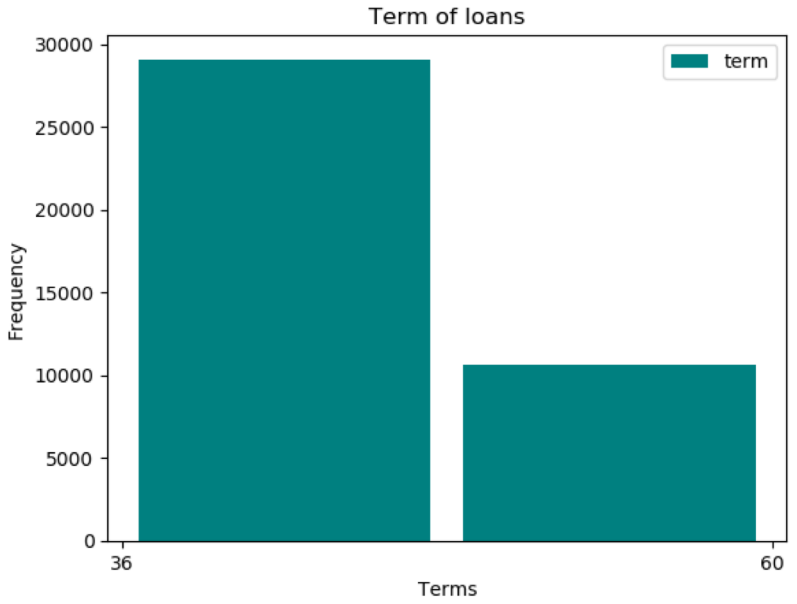
Inference: *After removing the outliers in the 'annual_inc' column, most frequent income range is between $46000 to $52000.*

Univariate Analysis: *continued …*



Purpose of Loans – Frequency

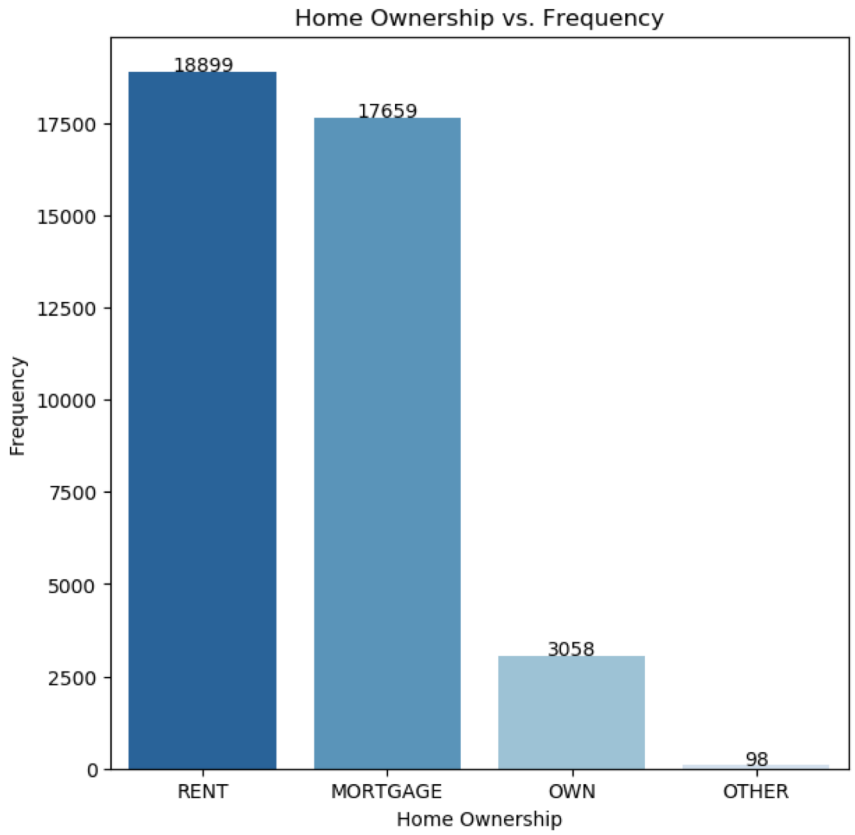
Inference: *Most of the loans are taken for debt Consolidation.*



Term of Loans - Frequency

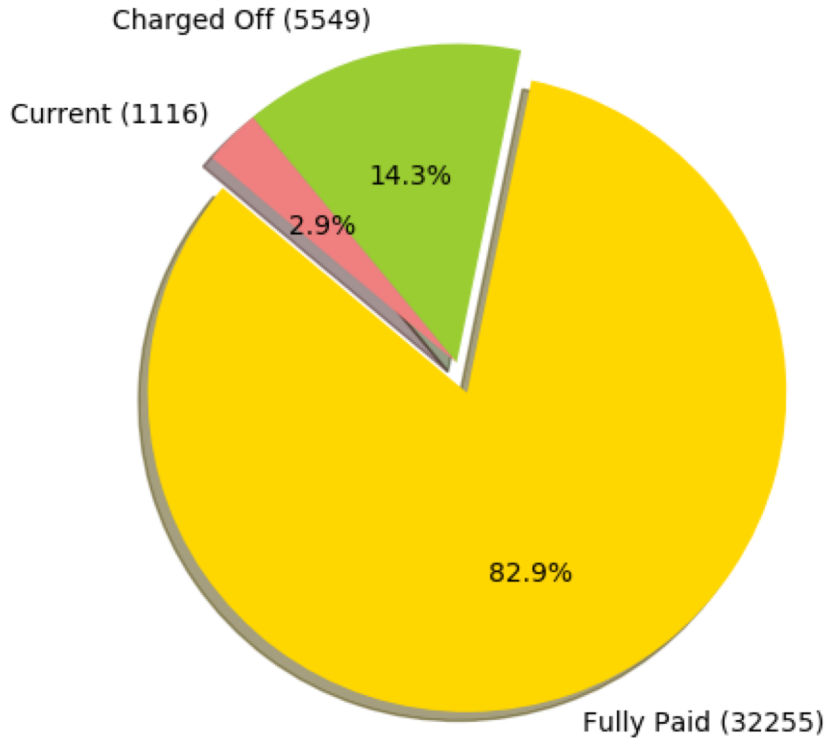
Inference: *The most frequent term for loans is 36 months.*

Univariate Analysis: *continued …*



Home Ownership

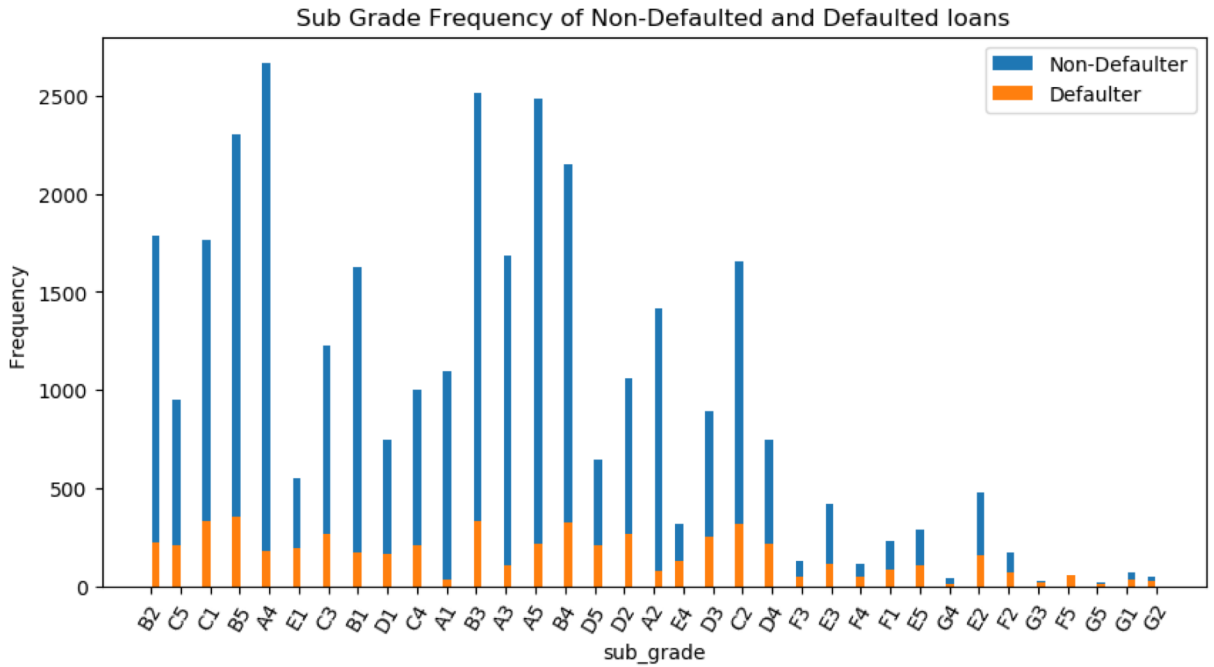
Inference: *People who are staying on rented house are taking most of the loans, followed by people staying in a mortgaged home.*



Loan Status

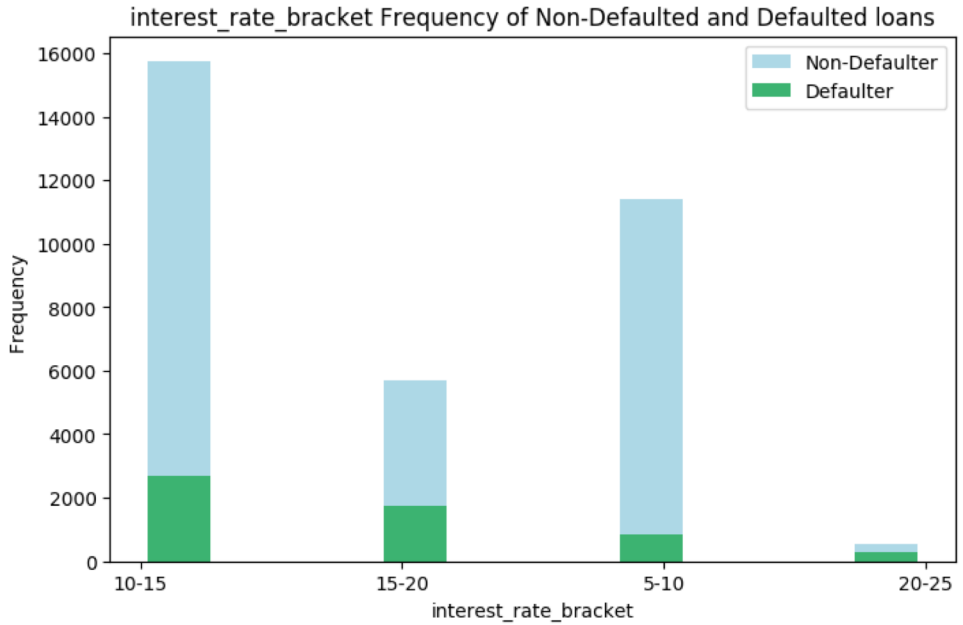
Inference: *Most of the loans in the dataset are fully paid and only a small percentage of loans are current loans.*

Segmented Univariate Analysis: *Defaulters vs. Non-Defaulters*



Sub Grade – Frequency

Inference: *Lower the sub-grade higher is the chances of defaulting.*

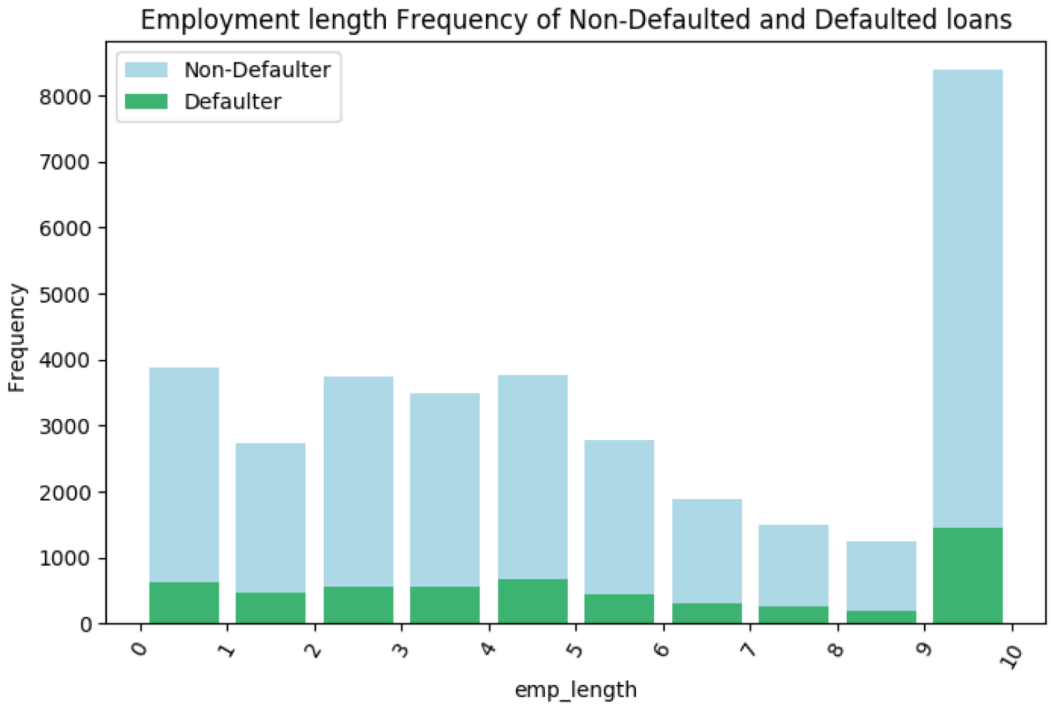


Interest Rate Bracket – Frequency

loan_default_status	Defaulter	Non-Defaulter
interest_rate_bracket		
10-15	14.54	85.46
15-20	23.63	76.37
20-25	34.59	65.41
5-10	6.73	93.27

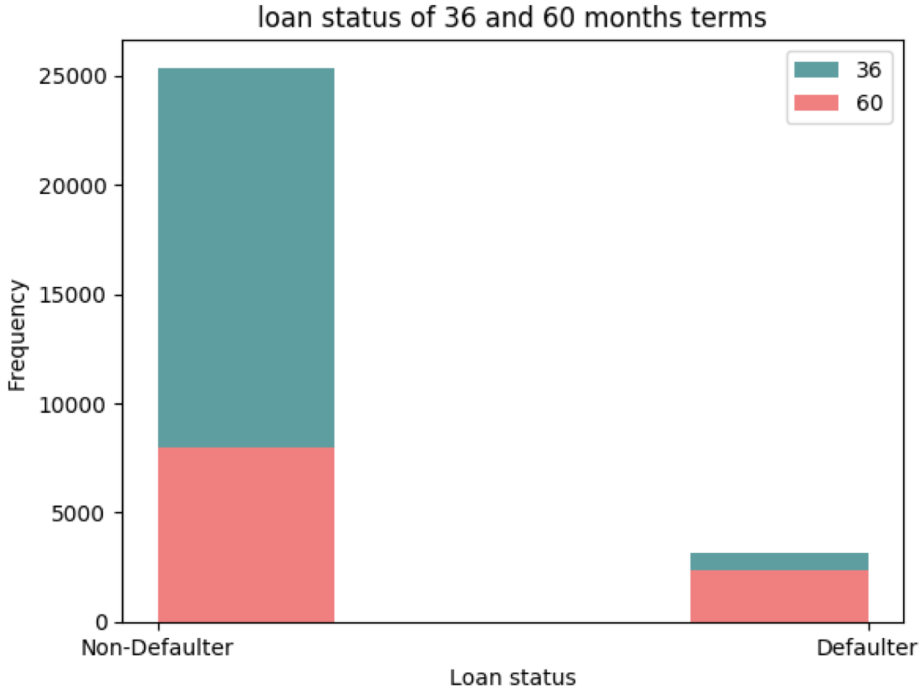
Inference: *Nearly 35% of loans in the interest bracket 20-25% have been defaulted followed by 15-20% interest bracket which has defaulted loan of 24%*

Segmented Univariate Analysis: *Defaulters vs. Non-Defaulters*



Employment Length – Frequency

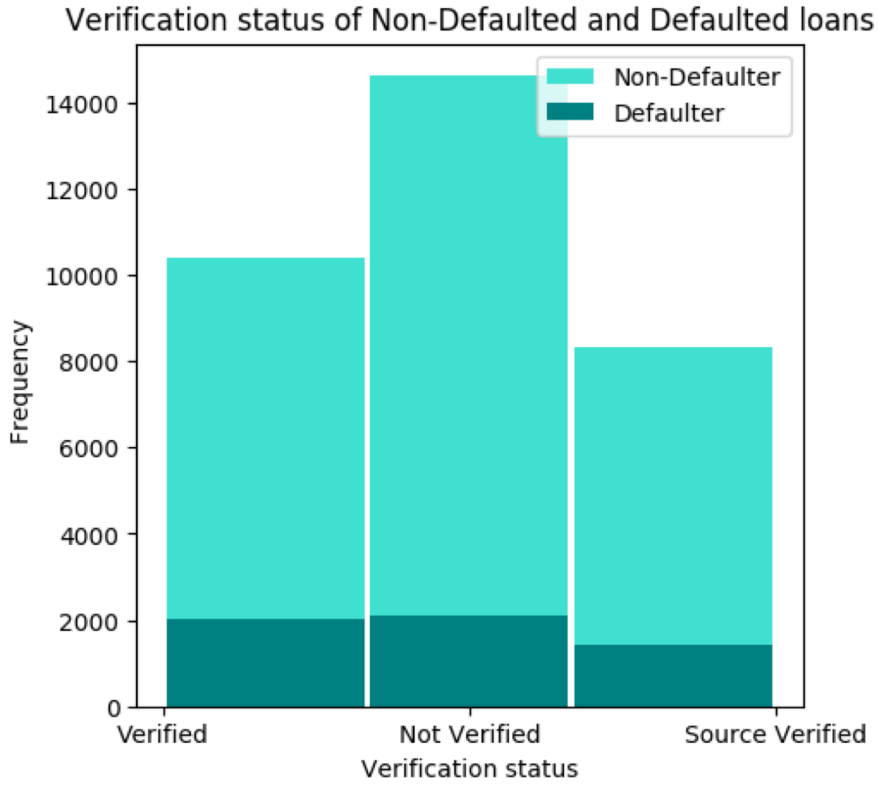
Inference: *Lower experience employees tend to default more on loan payment than high experience employees..*



Loan Status - Freq

Inference: *The defaulted loans has mostly 60 months term.*

Segmented Univariate Analysis: *Defaulters vs. Non-Defaulters*



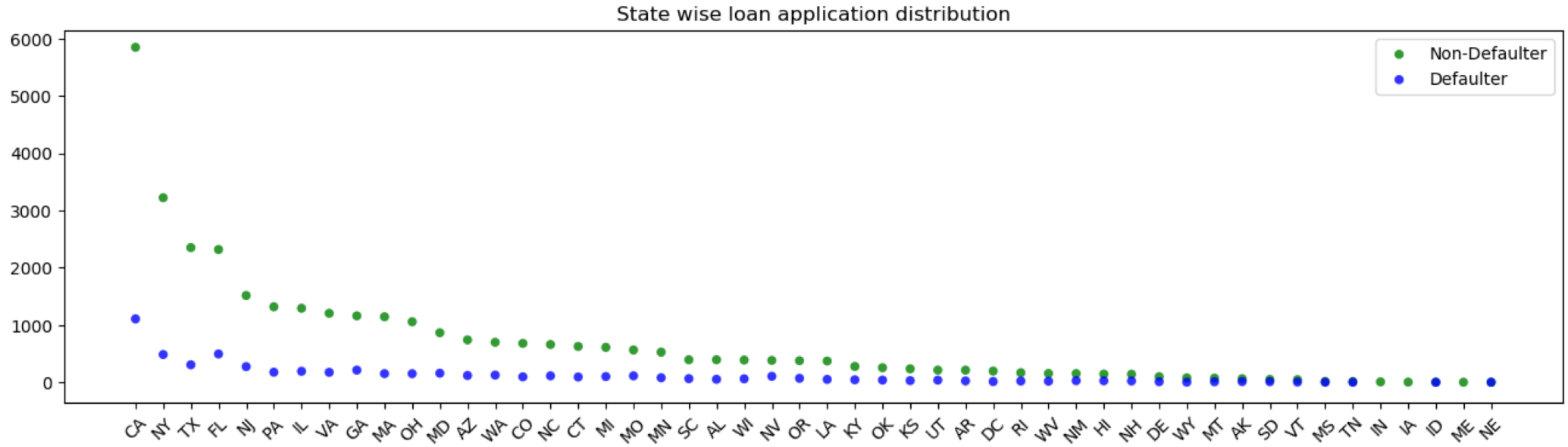
Verification Status – Freq.

loan_default_status	Defaulter	Non-Defaulter
verification_status		
Not Verified	12.66	87.34
Source Verified	14.50	85.50
Verified	16.23	83.77

Inference:

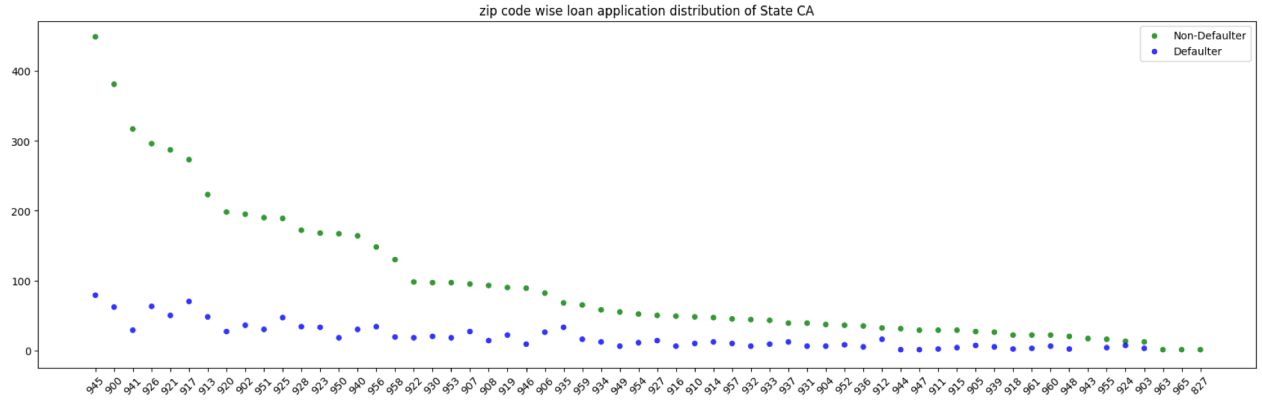
Maximum percentage (16%) of loans sanctioned in "Verified" category have been defaulted in loan while Not Verified have the least(12%)

Segmented Univariate Analysis: *State-wise Loan Application Dist.*



Inference:

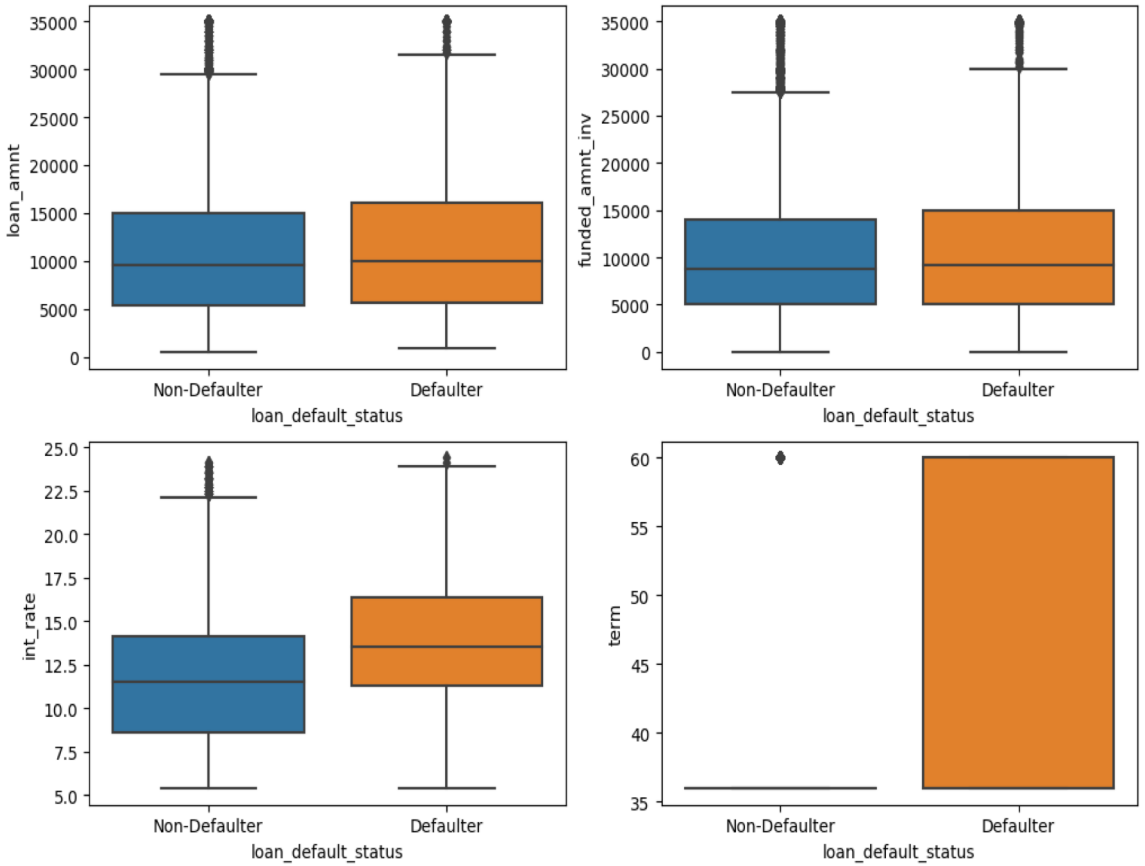
The state CA has most of the loans...
The state CA has most of the 'Non-Defaulted' loans...
The state CA has most of the 'Defaulted' loans...



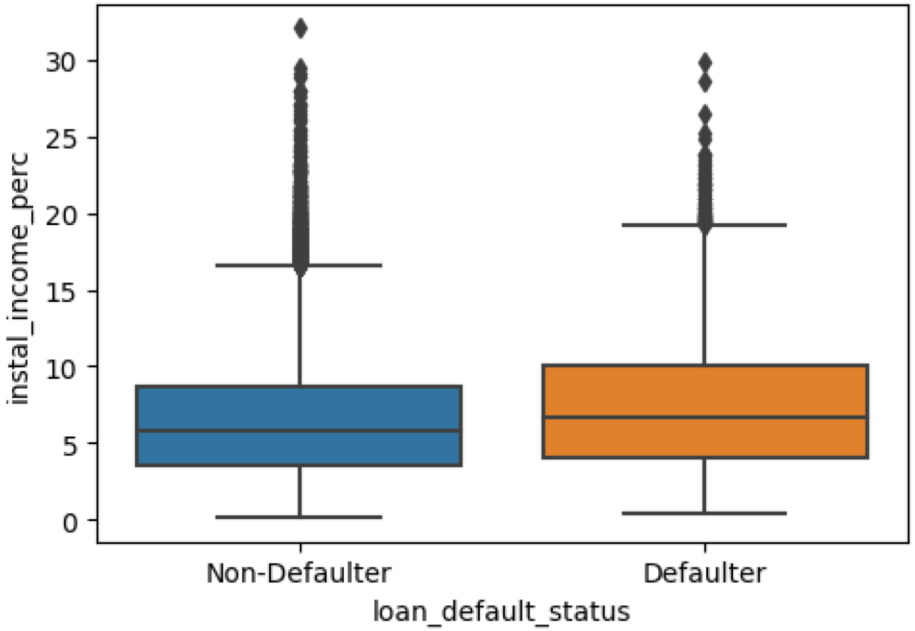
Conclusion:

State CA and ZipCode '945' tops in
both Non-Defaulter and Defaulter categories of loans.

Segmented Analysis: *loan status for loan amount, funded amount & interest rate*

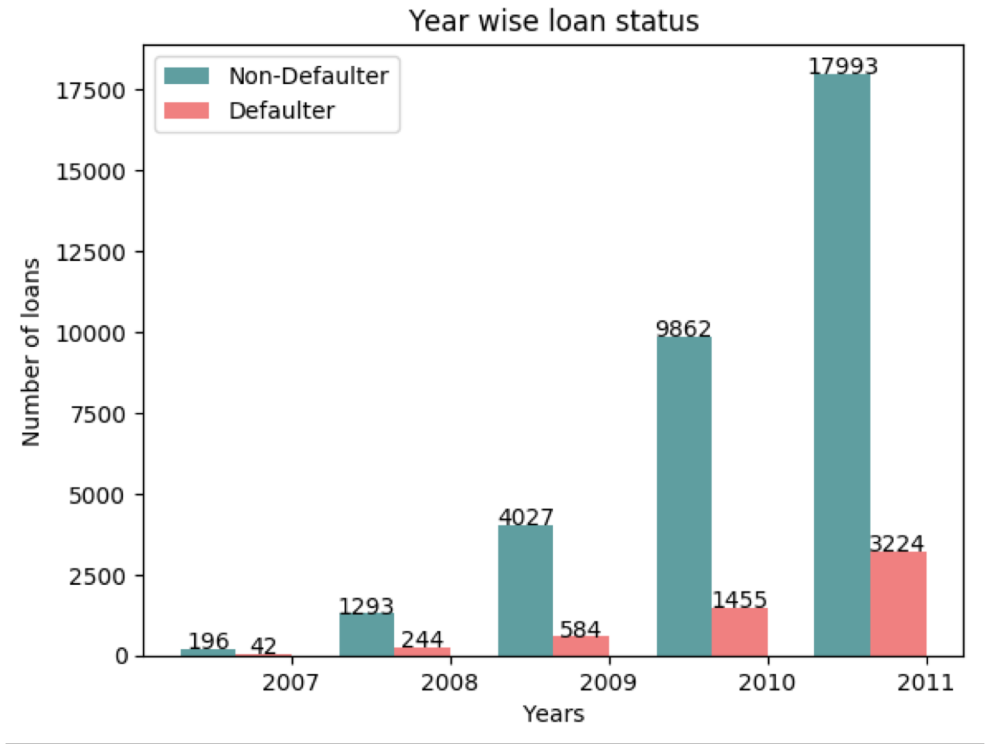


Inference:
The loan_amnt and funded_amnt_inv has almost equal spread and median for defaulted and non-defaulted loans.
The interest rate bracket for defaulted loans is clearly higher than non-defaulted loans.
Almost All the loans Defaulted are given for 60 months interval.



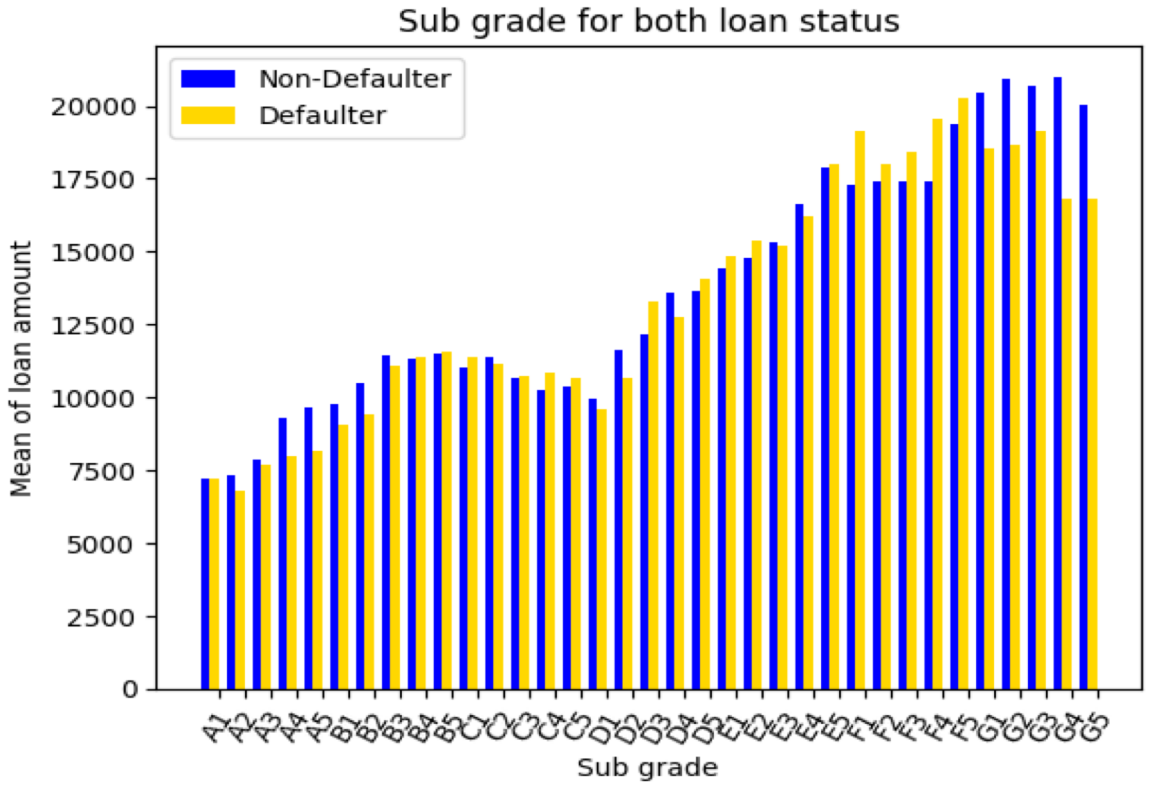
Inference: *The instalment to income perc is slightly higher for defaulted loans than non-defaulted loans.*

Bivariate Analysis: *Defaulters vs. Non-Defaulters*



Year wise loan status

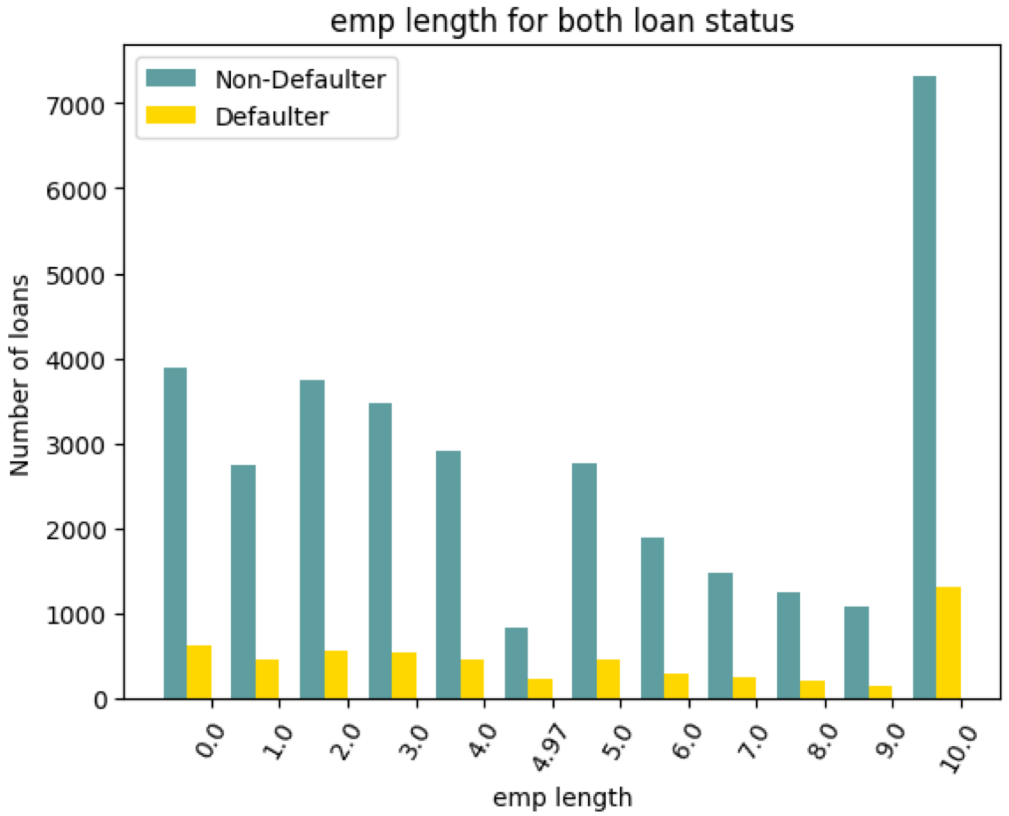
Inference: *Year 2011 has most loans on both categories.*



Sub Grade for both loan status

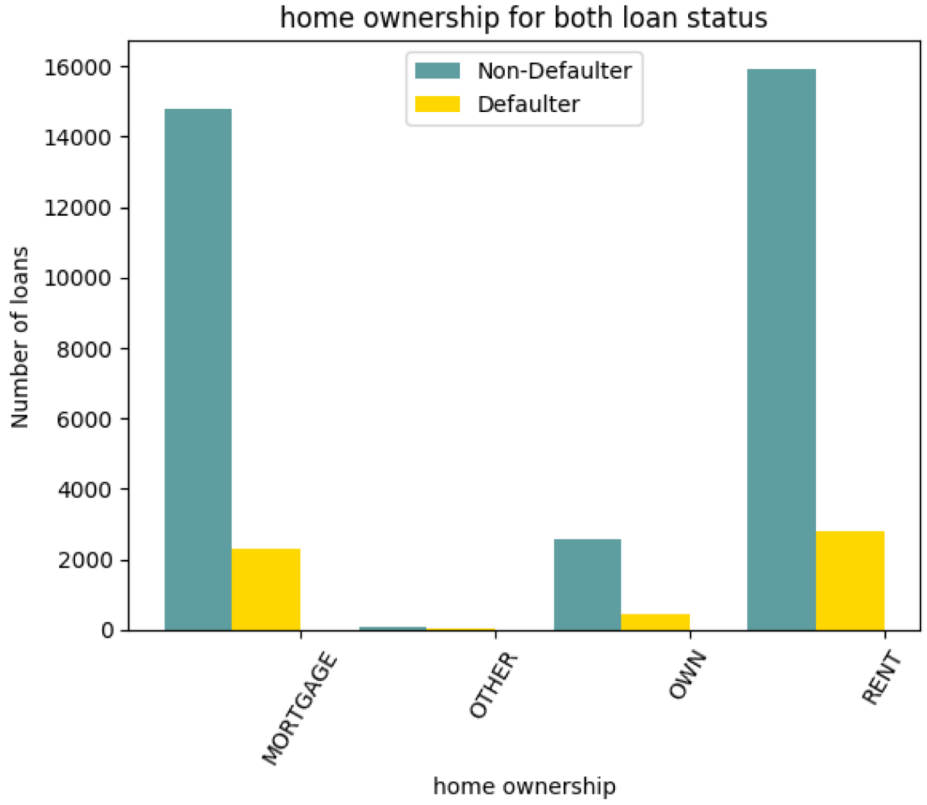
Inference: *Lower the grade higher is the average loan amount and risk for defaulting except for few exceptions*

Bivariate Analysis: *Defaulters vs. Non-Defaulters*



Employment length for both loan status

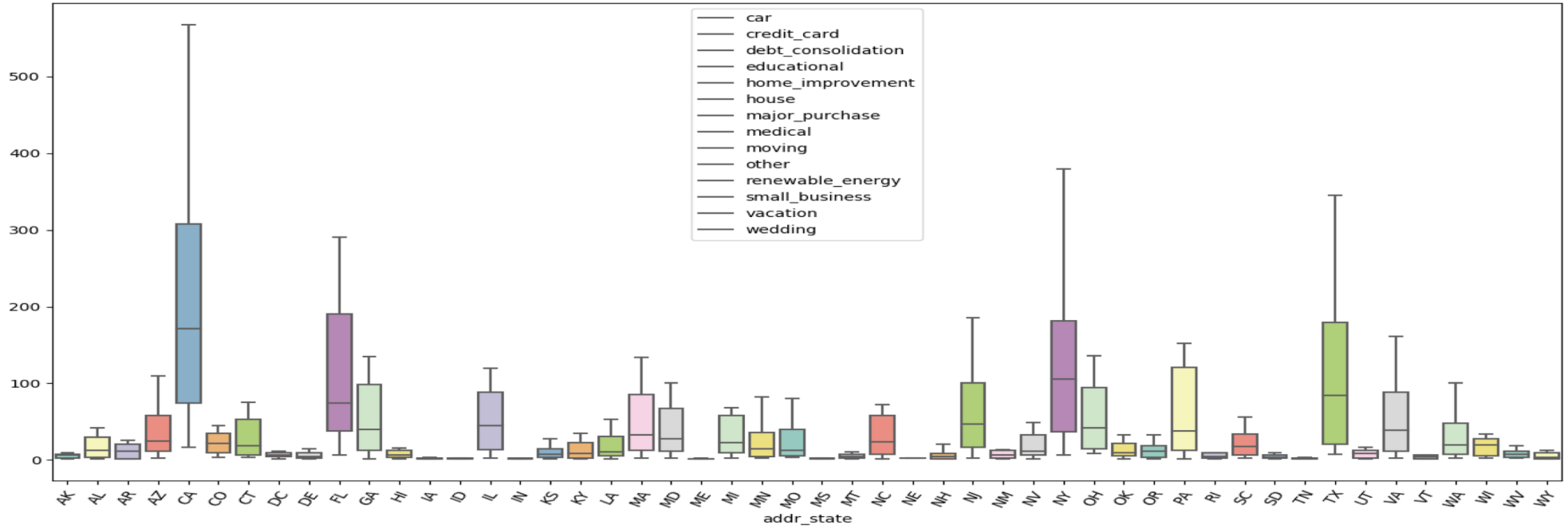
Inference:
Employees with longer duration are processing more loans in both categories.
There is no much deviation in defaulting rate based on employees' experience.



Home Ownership for both loan status

Inference:
Applicant who are in rental homes are more likely to default followed by mortgaged home owners.

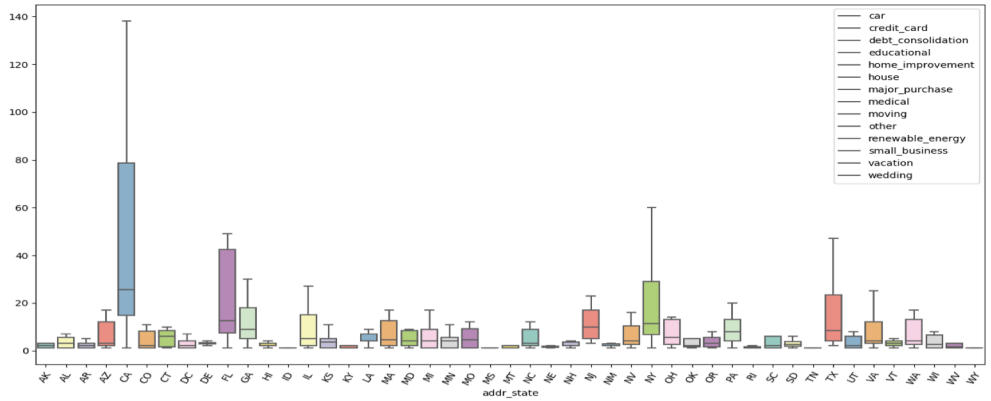
Bivariate Analysis: *state, purpose and loan status*



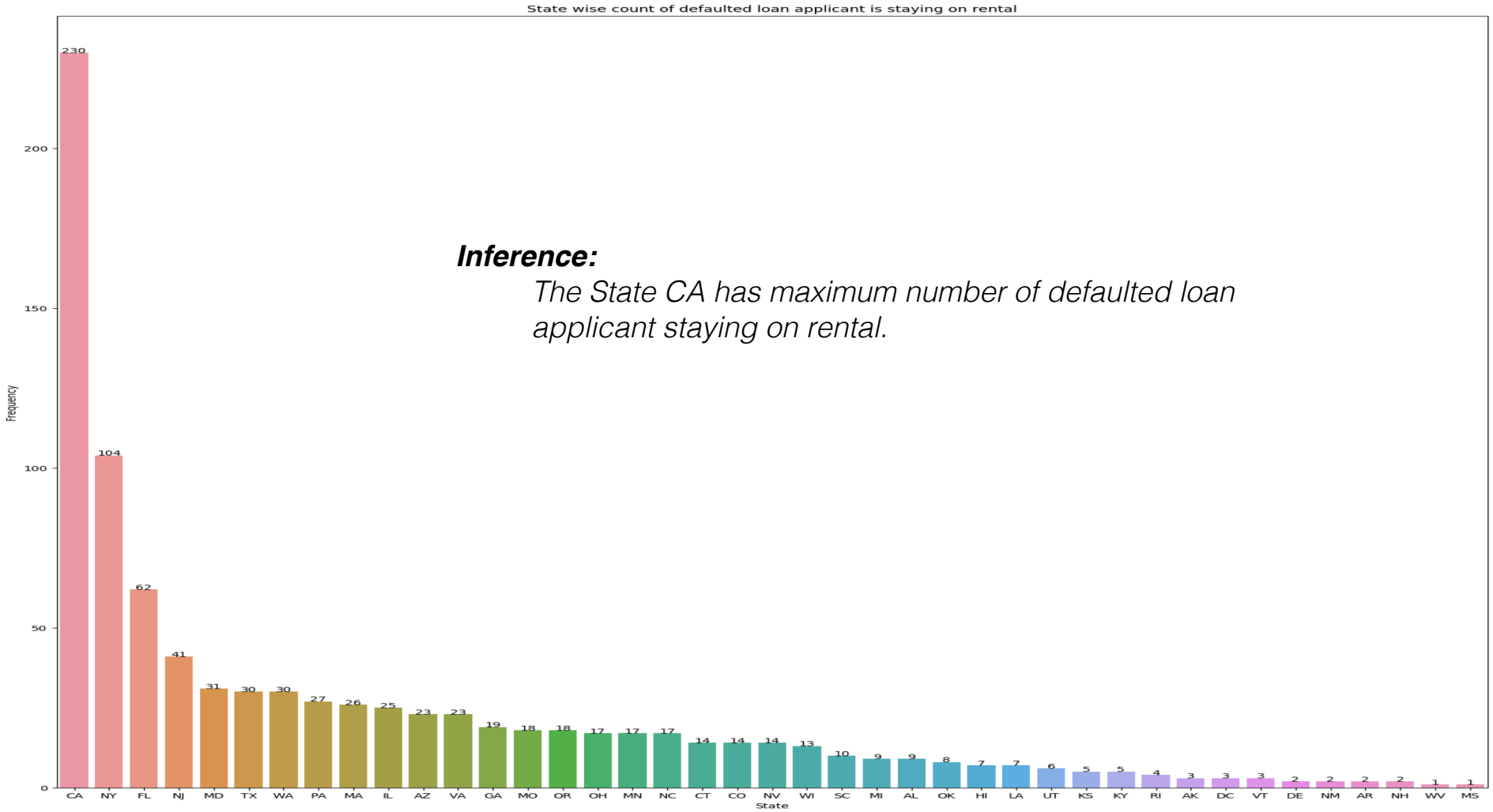
Inference:

The loan taken for debt consolidation has highest chances of defaulting

The state CA has highest percentage of defaulters



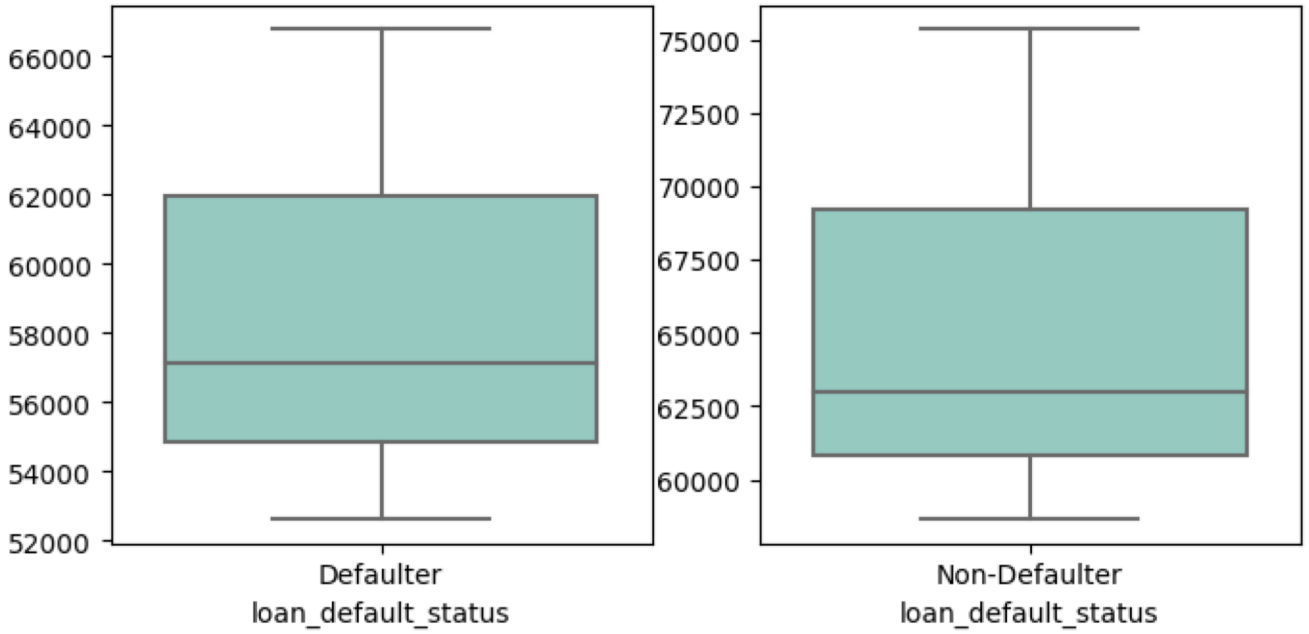
Bivariate Analysis: *address, verification status, home ownership and loan status*



Inference:

The State CA has maximum number of defaulted loan applicant staying on rental.

Bivariate Analysis: *verification status and loan status based on annual income*

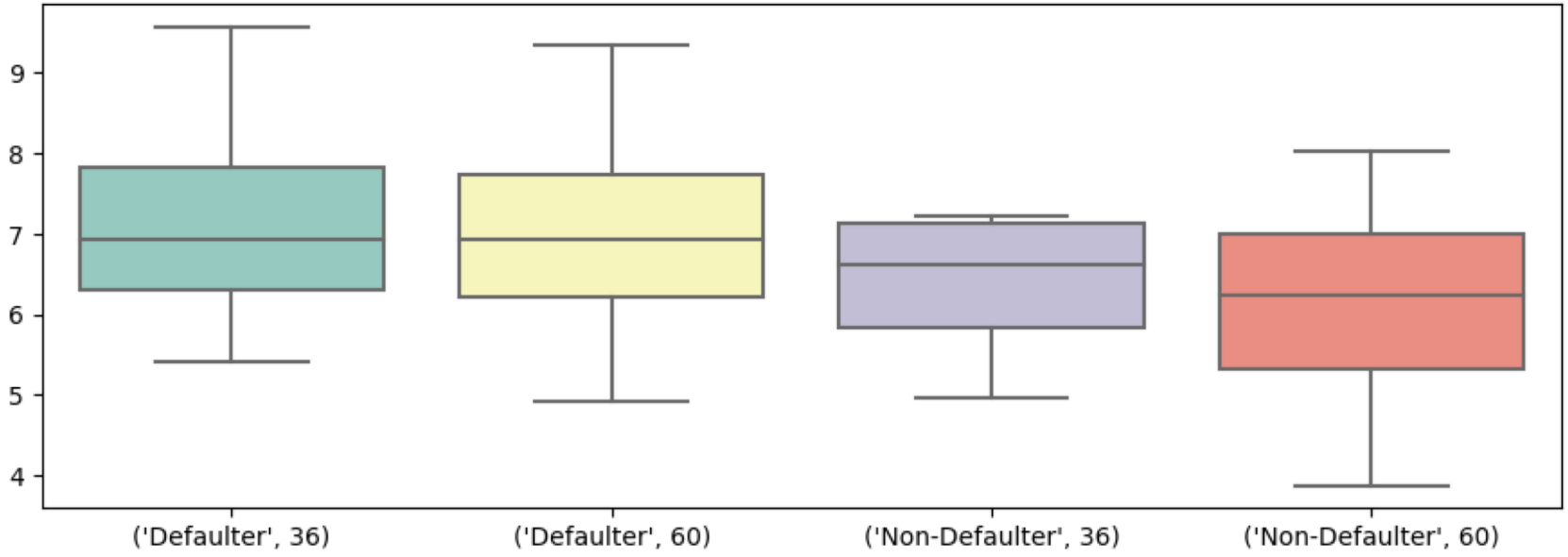


loan_default_status	Defaulter
verification_status	
Not Verified	52635.181697
Source Verified	57112.503800
Verified	66787.825280

loan_default_status	Non-Defaulter
verification_status	
Not Verified	58669.808540
Source Verified	63014.910050
Verified	75343.250518

Inference: *In each of the 'verification_status' group (Not verified, source verified, verified), Fully paid always has higher average 'annual_inc' than charged off*

Bivariate Analysis: *interest rate based on instalment to income ratio and term*



loan_default_status	Defaulter		Non-Defaulter	
term	36	60	36	60
interest_rate_bracket				
10-15	6.594200	6.645357	6.132338	5.800984
15-20	7.258667	7.203643	7.100320	6.657143
20-25	9.570312	9.347769	7.218121	8.037000
5-10	5.415128	4.921428	4.951257	3.875707

Inference from the table:
For each term term loan the instalment to income ratio of for each of the interest slabs are as below. This influences whether a loan is likely to be defaulted or not.

Conclusion & Recommendations

Key Driving Factors						
1	2	3	4	5	6	7
home-ownership	interest_rate	term	instalment to income ratio	loan_purpose	address	loan_amnt

Based on the following inferences

- *Loan taken for 60 month interval has higher chances of defaulting.*
- *Nearly 35% of loans in the interest bracket 20-25% have been defaulted followed by 15-20% interest bracket which has defaulted loan of 24%. So if the loan has higher interest rate (about 15%), there is higher chance of defaulting.*
- *Instalment to income ratio is driving factor. Having high value means high chances of default.*
  - *For 60 months term, if the instalment to income ratio is greater than 9%, there is maximum probability of defaulting*
- *Purpose of loan: Loan taken for "Debt consolation" has high chances of defaulting.*
- *Loan applications from State CA and Zipcode '945'*
- *Applicant who are in rental homes are more likely to default followed by mortgaged home owners*

Thank You