Lending Case Study

Authors: Vinnakota Santosh Aravind
Subhash Chandra Bose Kondapalli

Problem Statement:

- We work for a **consumer finance company** which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:
- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company
- Find out how **consumer attributes** and **loan attributes** influence the tendency of default?

Cleaning the Data

- We received a data set containing 111 columns with 39717 records
- Out of 111 columns we have 57 columns with 40% to 100% of data as Nan values. So lets remove them:
  - 'mths_since_last_delinq', 'mths_since_last_record', 'next_pymnt_d', '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', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit’

- 'id','member_id','url‘ are just identifiers to distinguish the users => not useful for the analysis, So let's remove them

Cleaning the Data

- Remove the Customer Behavior 20 variables which are not available at the time of loan application
  - 'delinq_2yrs','earliest_cr_line','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_d','last_credit_pull_d',

    'last_pymnt_amnt', 'pub_rec_bankruptcies'
- Removing the columns which have only 1 value (and NaN values) across the users
  - 'pymnt_plan', 'initial_list_status', 'policy_code', 'application_type', 'acc_now_delinq', 'delinq_amnt', "collections_12_mths_ex_med", "chargeoff_within_12_mths","tax_liens"
- categorical variables whose distribution is widespread i.e., lot of unique entries can be removed
  - 'emp_title','desc','title','zip_code'

Cleaning the Data

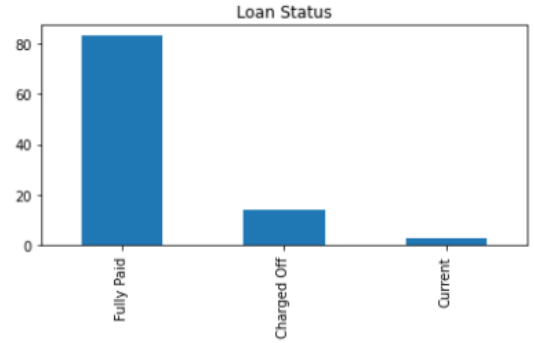
- After removing all the unwanted and non meaning full columns we are left with 18 columns and null values are less than 3 %.

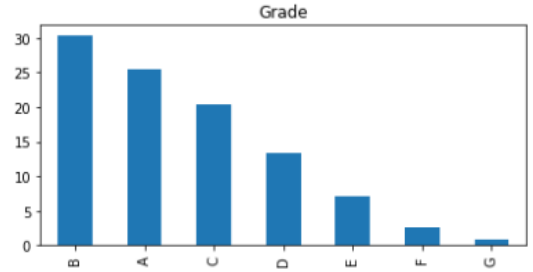
- Term column represents loan tenure in months. So removing the “months” value and keeping only the number in term column
- int_rate column represents the percentage of loan interest. Hence, removing the “%” symbol
- revol_util is having 50 null values and hence imputing null values with mode value “0%”
- revol_util contains entire data in “%” format. Removing the “%” from the column and keeping only numerical values

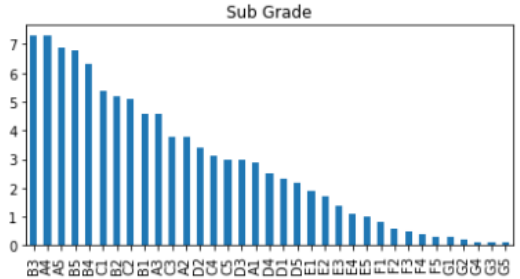
Univariate Analysis

purpose of this is to understand the distribution of data

- After cleaning the data, we are now left with 17 variables out of which 9 categorical variables and 8 numerical variables.
- Plotting the bar graph for Loan status, grade and sub grade







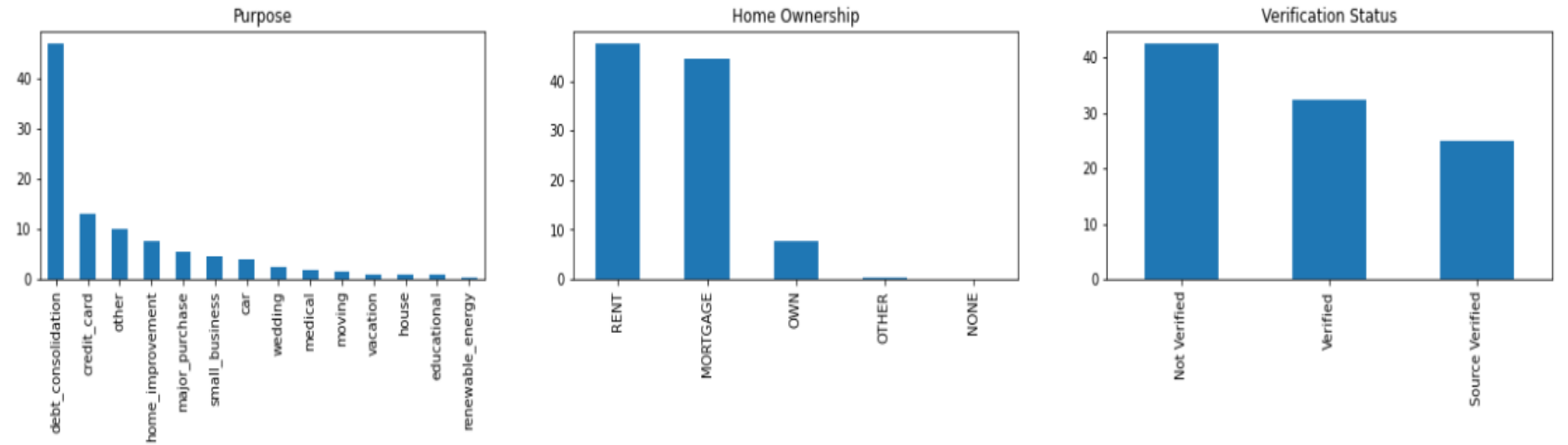
1. Although most of the customers have fully paid their respective loans, there is a significant percentage of people(20%) who Charged off their loan

2. Customers with top grade ratings take significant number of loans.

Univariate Analysis

purpose of this is to understand the distribution and variance of the data

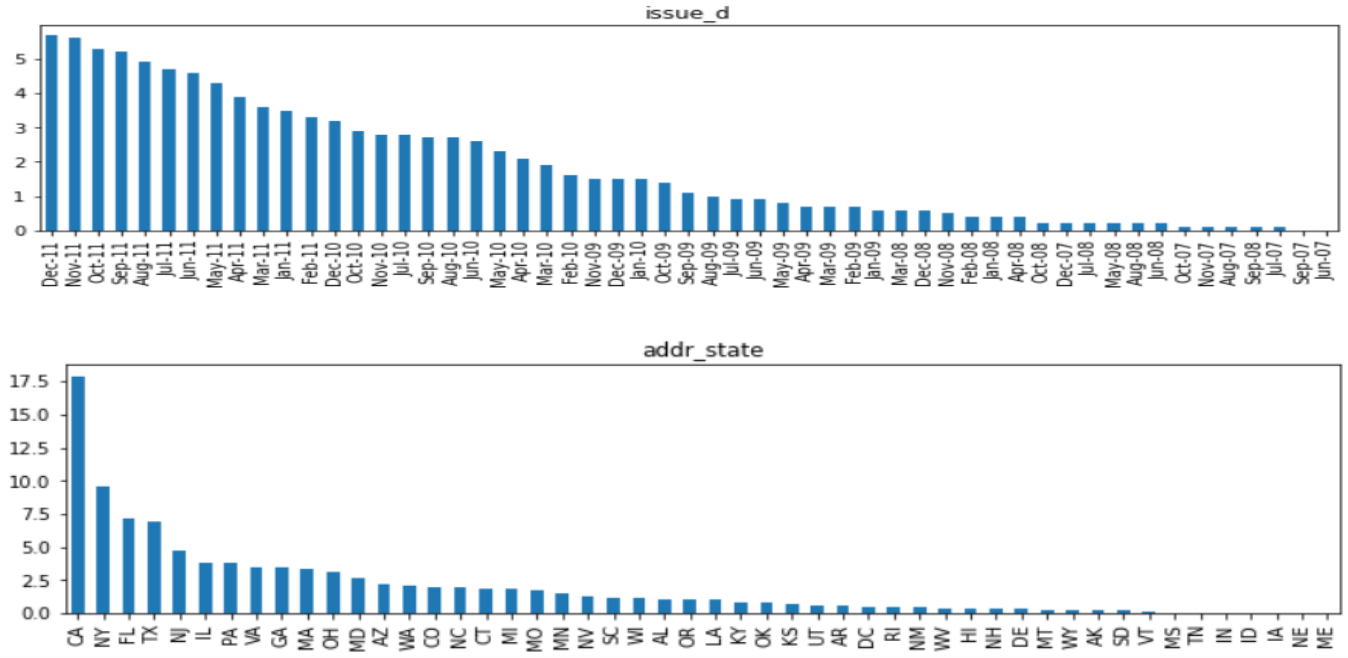
- Plotting the bar chart for Purpose, Home Ownership and verification status



- 3. Customers with Home Ownership as "Rent" and "Mortgage" are the potential loan Customers and majority of the customers took loan for their "Debt Consolidation”

Univariate Analysis

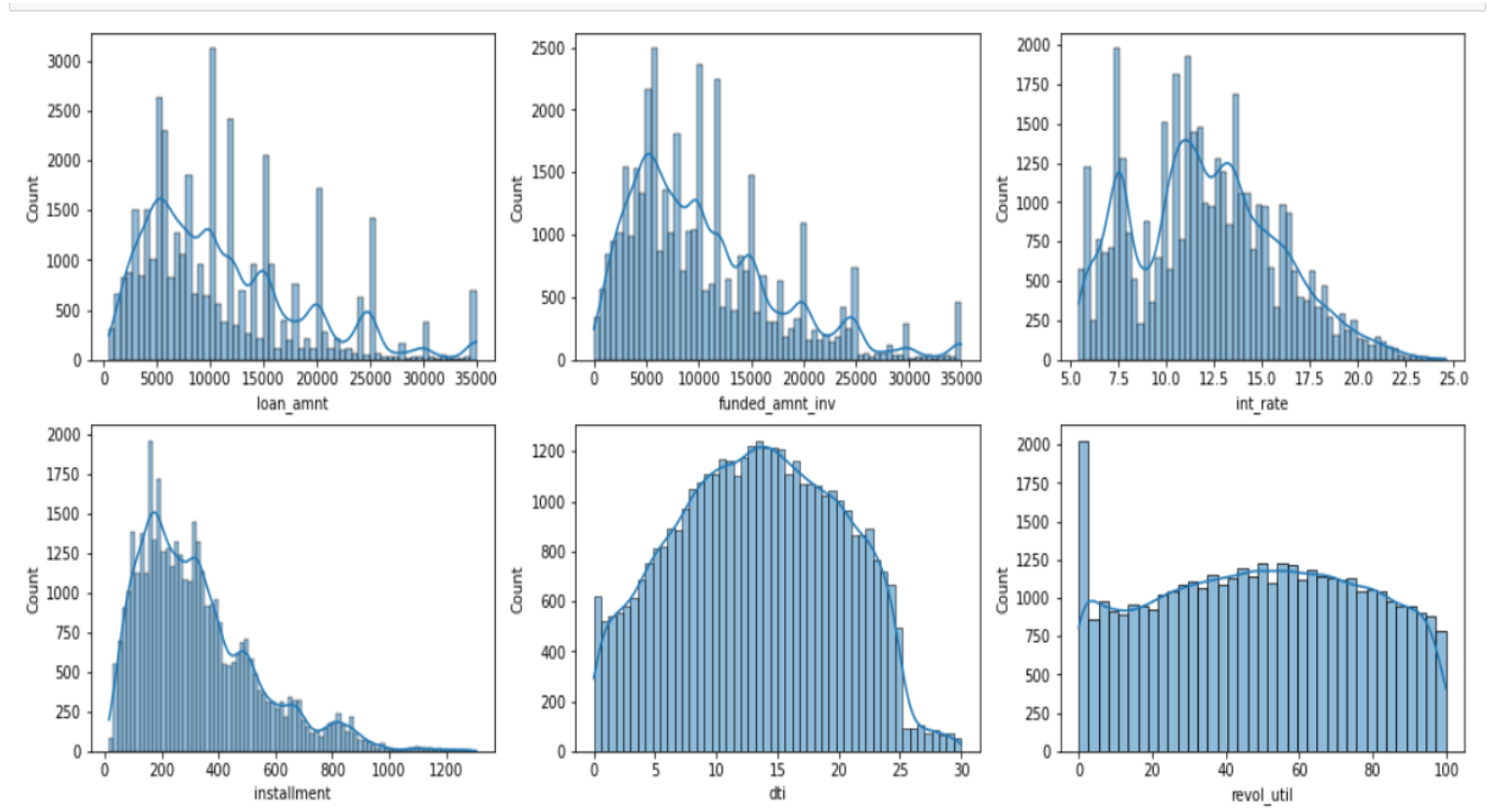
- Plotting the graph for issue_d and addr_state



- 4. Over the years the number of customers taking loans is increasing which indicates an increasing Customer Base
- 5. A significant Customer Base is from states like CA, NY, FL and TX. Implies, greater scope for the company expansion in states where we have less defaulters.

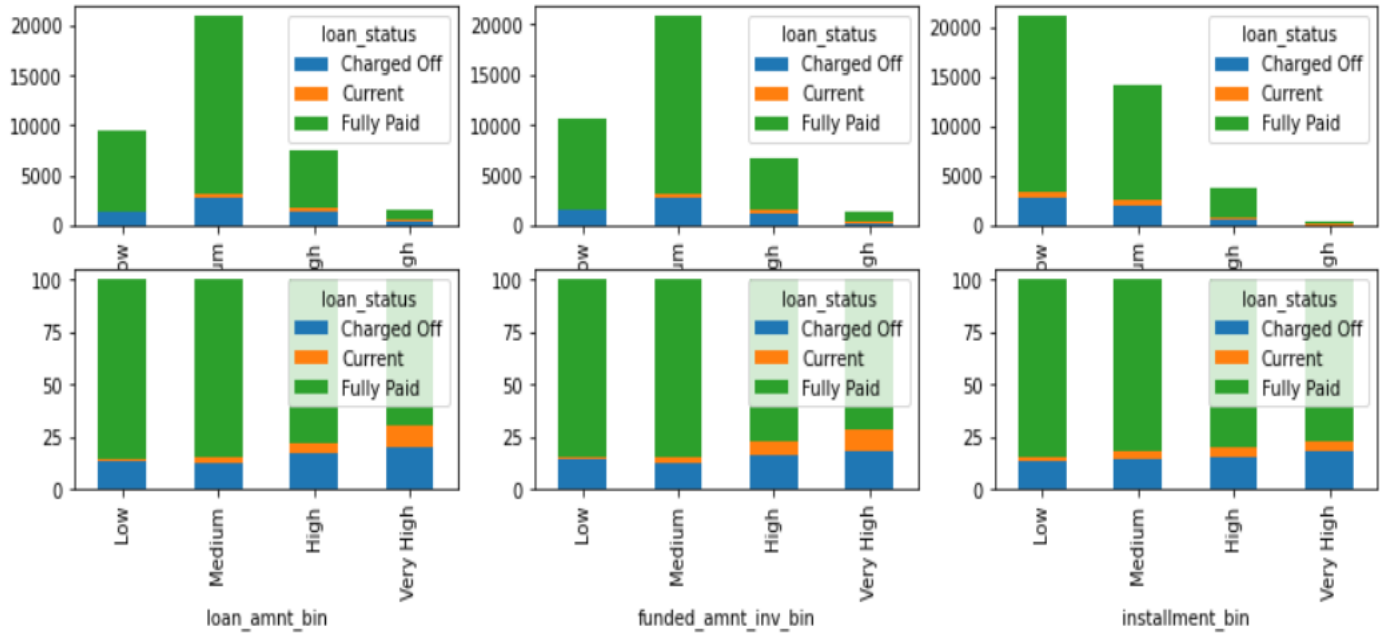
Univariate Analysis

- Plotting the graphs to view the dispersion of the data for loan_amnt, funded_amnt_inv, int_rate, installment, dti, revol_util



Bivariate Analysis

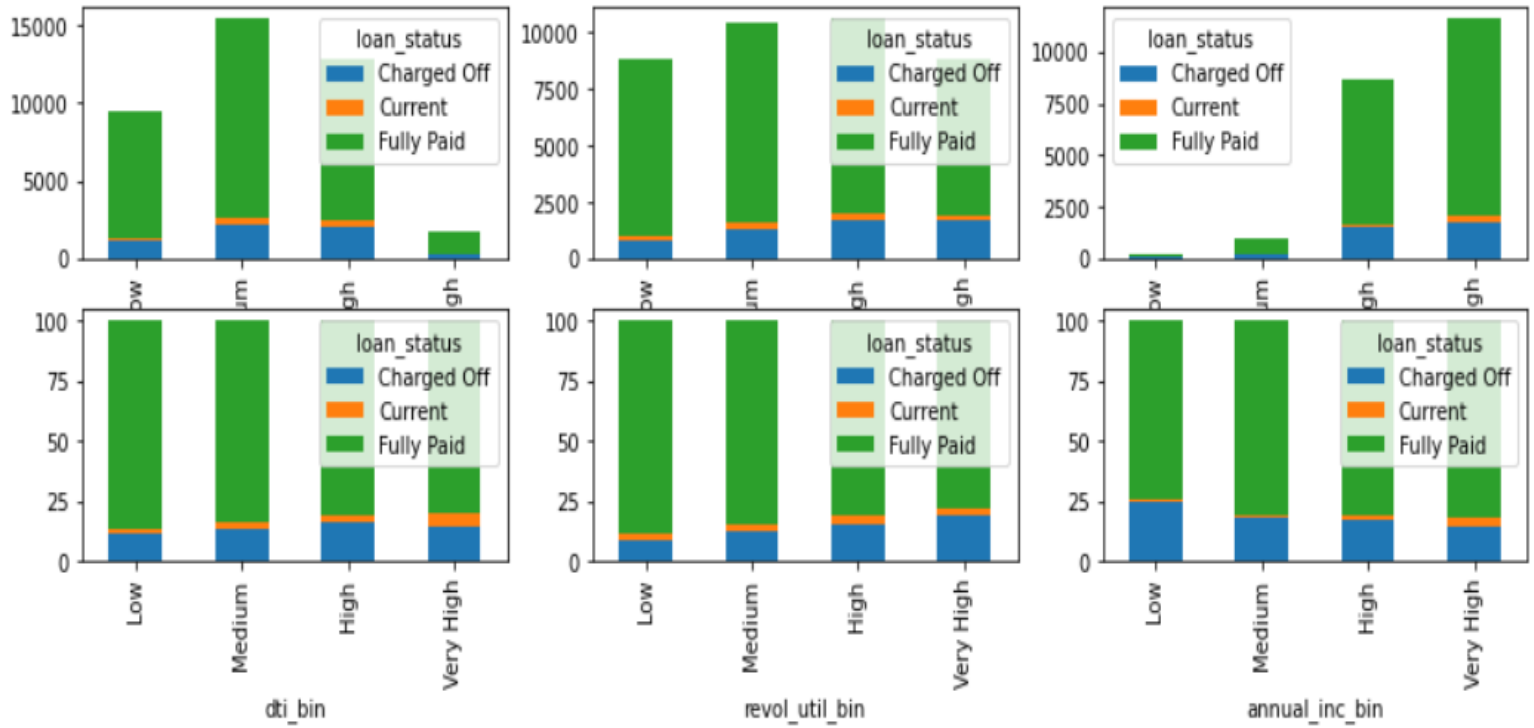
- We are plotting graphs according to their Counts(in 1st row) and Percentages(in 2nd row) to get more information



- 6. Customers who are receiving loans with higher loan amount or funded amount($25000-35000) have a higher default rate

Bivariate Analysis

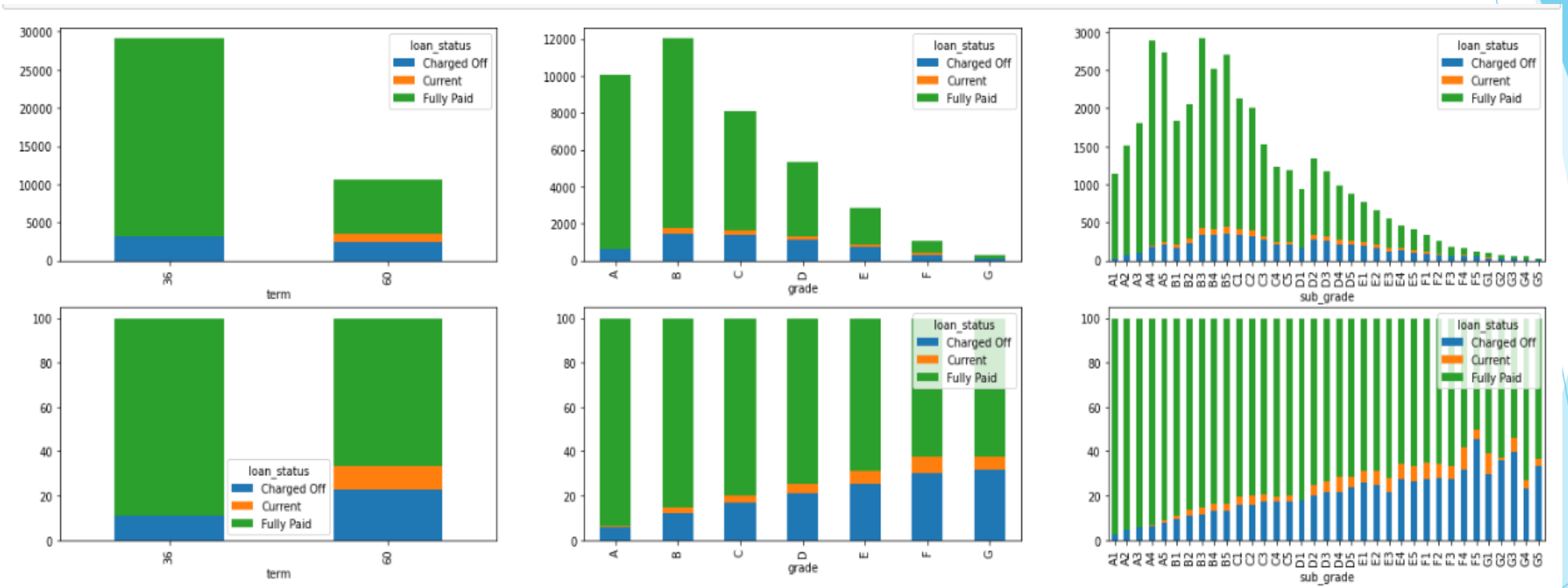
- Plotting the graphs among dti, revol_util, annual_inc



- 7. Customers with low and medium annual incomes(< $12000), have a high default rate.

Bivariate Analysis

- Plot the graph among term, grade, sub_grade against loan_status

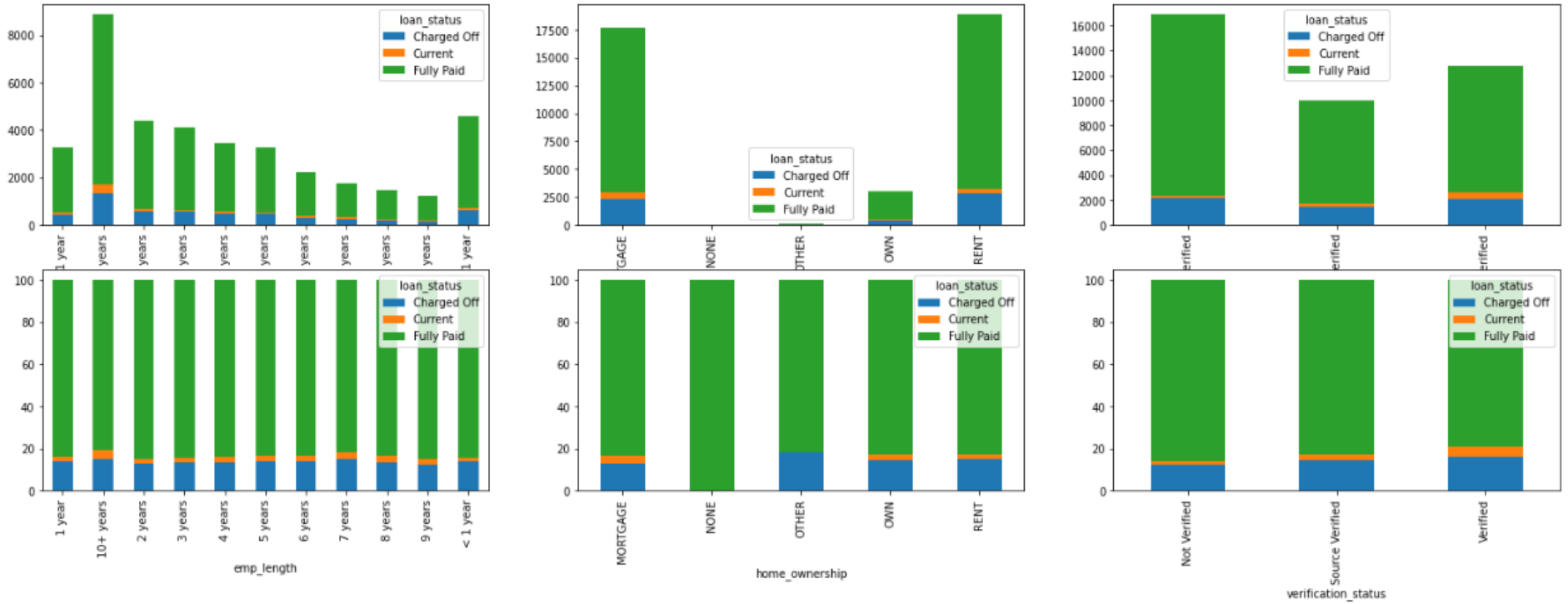


8. Customers taking loans for 60 months tend to default more than the lower term period one's.

9. we need to be more cautious in lending the loan for the customers with lower subgrade(from B2-G5) ratings and a loan term of 60months

Bivariate Analysis

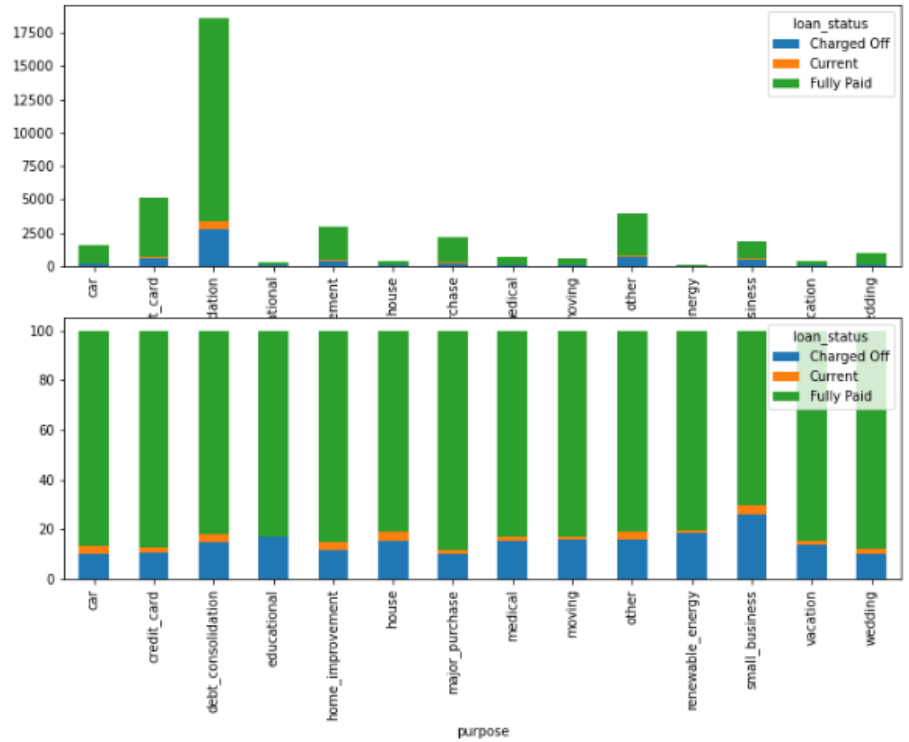
- Plot the graph among emp_length, home_ownership, verification_status against loan_status

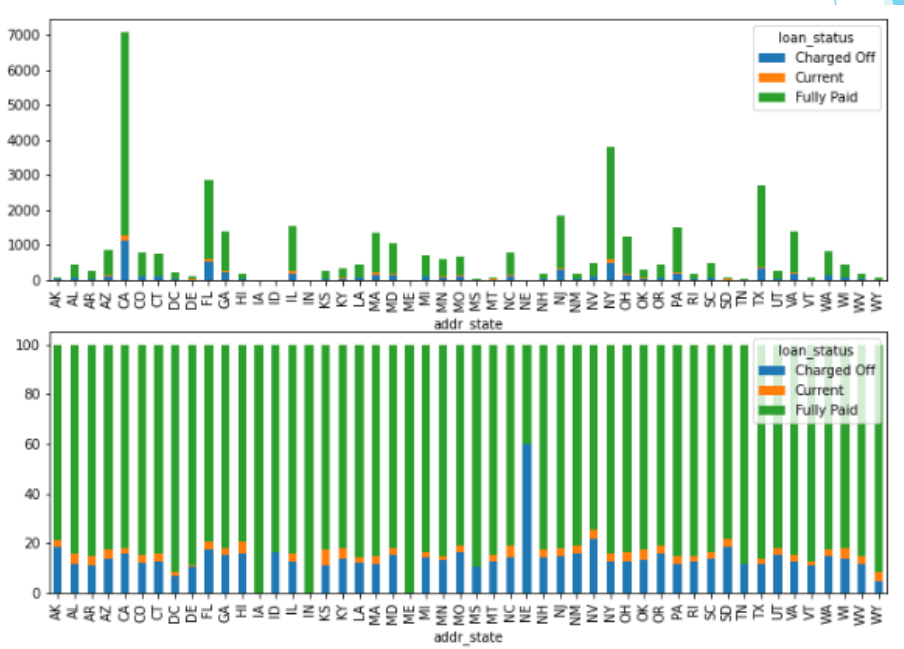


- 10. We have observed that Customers with Home Ownership as Rent and Mortgage are the potential loan applicants, but however, there is not significant rise in default rate for these customers, so it is safe to target these customers to increase the customer base of the company

Bivariate Analysis

- Plot the graph among purpose and addr_state against loan_status

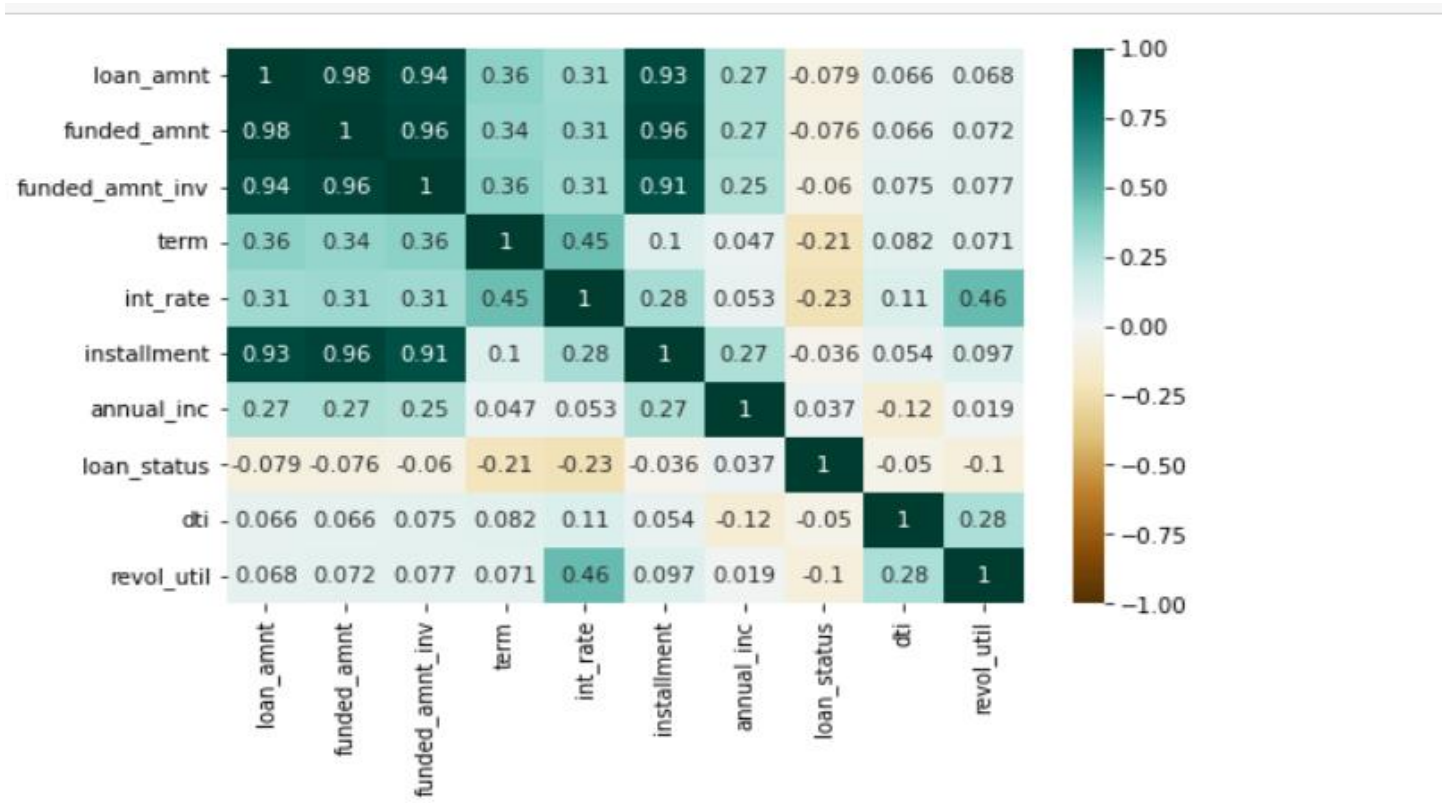




- 11. Customers taking loan for “small business” purpose have a high default rate. Let’s analyze more and consider necessary actions to mitigate this.
- 12. It is suggestible to stop lending loans(if not, be very cautious before lending) to the customers from the state of "NE" as 65% of these customers have defaulted.

Multivariate Analysis

- Plot the correlation among the variables to verify if there is any high correlation



- Not much information can be extracted from the correlation matrix w.r.to direct impact on loan status

Segmented Analysis

- Lets conclude our observations by segmenting the customers based on the above insights(11,12 points)

loan_status	purpose	annual_inc	loan_amnt	int_rate	revol_util
Charged Off	debt_consolidation	36120.0	14400.0	12.800	93.50
	home_improvement	53500.0	2500.0	11.220	19.45
Fully Paid	debt_consolidation	52500.0	6150.0	11.915	75.10

loan_status	annual_inc	loan_amnt	int_rate	revol_util
Charged Off	10800.0	2250.0	12.990	35.5
Fully Paid	10800.0	2400.0	12.455	31.0

loan_status	home_ownership	annual_inc	loan_amnt	int_rate	revol_util
Charged Off	MORTGAGE	66000.0	15000.0	13.785	43.05
	OTHER	78500.0	13500.0	15.825	34.50
	OWN	56000.0	10500.0	14.110	49.50
	RENT	50000.0	10000.0	14.265	49.40
Fully Paid	MORTGAGE	76584.0	12437.5	12.150	34.90
	OTHER	83500.0	13350.0	12.050	25.40
	OWN	53400.0	10000.0	11.110	25.80
	RENT	54000.0	10000.0	12.840	32.90

loan_status	loan_amnt_bin	annual_inc	loan_amnt	int_rate	revol_util
Charged Off	Low	54000.0	4000.0	7.51	14.00
	Medium	60000.0	10000.0	7.66	11.70
	High	64000.0	18000.0	7.51	15.90
	Very High	121500.0	30000.0	7.50	8.00
Fully Paid	Low	60000.0	4000.0	6.99	9.60
	Medium	68000.0	9600.0	7.29	10.90
	High	83500.0	18500.0	7.49	12.90
	Very High	120000.0	30000.0	7.90	10.05

Conclusions

Over the years, although the customer base of loan applicants is increasing(which is a positive sign), the default rate(Charged off Customers) is at ~20% which is a significant number, because the total amount lended to these customers is Company's loss

To avoid Loss

1. The most immediate step we can take is to avoid lending loans to customers from "NE" state whose annual income is comparable to the loan amount and revol_util is high(>80%). The default rate for these customers is >65% i.e., 65 out of 100 customers are Charged off Customers.

2. Customers who took loans for "Small Business" purpose tend to Charge off their laon if they have high revol_util value(>35%). We can experiment these customers by marginally reducing the interest rates to the high Grade, low revol_util customers so that chances of Paying off loans increases

3. Customers with low annual incomes(<12000 dollars) and low grading are to be offered less loan amount with slightly higher interest rates as these customers have high chances of defaulting the money.

To increase Profit Margins

1. For the 'A' Grade customers with high Annual income(>40000 dollars) and low revol_util(<25%), we can either offer them marginally more loan amount as per their requirement or slightly increase the interest rates(less suggestible)

2. For the long-term growth, We can expand our services to states like "IA", "ME" where we have very less customer base and almost 0% default rate and target the customers who have Home Ownership as "Rent" or "Mortgage" as these are the potential customer base we observed from our previous records