Lending Club Case Study

BY: TUSHAR SOOD

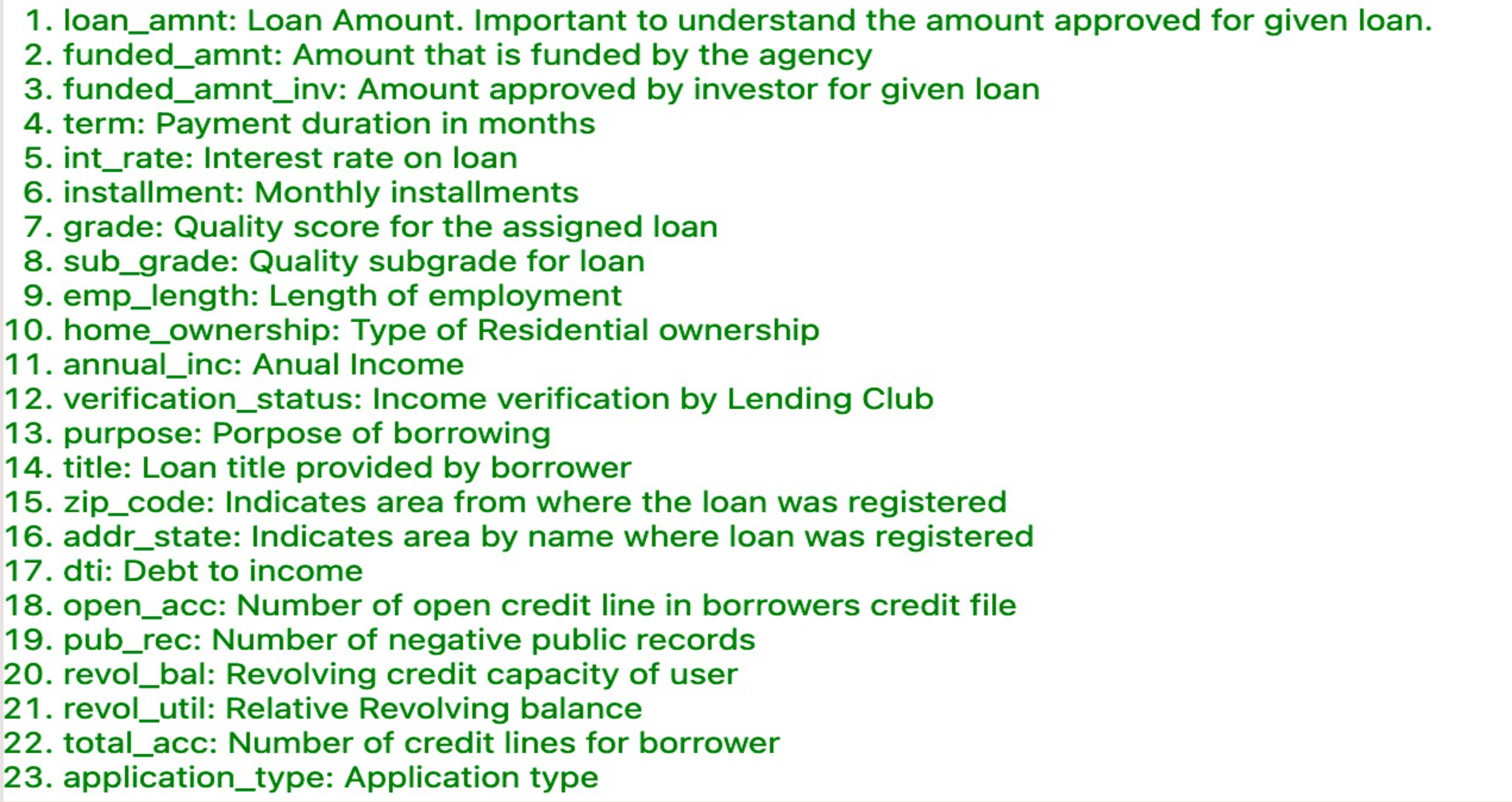
Problem Statement

- A Lending firm needs to reduce credit loss, by smartly identifying risks identified by lending loans
- If repayment does not happen it is a credit loss for the lending company
- If repayment is done it leads to good credit and hence profits for the lending company

Requirement

- Data set with 39717 rows and 111 columns are given along with metadata defining each column is given
- As part of the data exploration activity, I need to identify the pattern which will be useful for identifying the possibility of good and bad credit in future

What Interesting Columns are there?



1. loan_amnt: Loan Amount. Important to understand the amount approved for given loan.
2. funded_amnt: Amount that is funded by the agency
3. funded_amnt_inv: Amount approved by investor for given loan
4. term: Payment duration in months
5. int_rate: Interest rate on loan
6. installment: Monthly installments
7. grade: Quality score for the assigned loan
8. sub_grade: Quality subgrade for loan
9. emp_length: Length of employment
10. home_ownership: Type of Residential ownership
11. annual_inc: Anual Income
12. verification_status: Income verification by Lending Club
13. purpose: Porpose of borrowing
14. title: Loan title provided by borrower
15. zip_code: Indicates area from where the loan was registered
16. addr_state: Indicates area by name where loan was registered
17. dti: Debt to income
18. open_acc: Number of open credit line in borrowers credit file
19. pub_rec: Number of negative public records
20. revol_bal: Revolving credit capacity of user
21. revol_util: Relative Revolving balance
22. total_acc: Number of credit lines for borrower
23. application_type: Application type

Case Study Understanding…

- There are 3 categories of data given

  Data about ongoing or current loans

  Data about fully paid off loans

  Data about charged-off or defaulters

- If I can find a pattern in the charged-off category, then it will be easy to recommend credit loss risk to the lending company
- There is no meaning in exploring fully paid or ongoing customers are these will not lead to credit loss

Data Cleaning Activity

- In slide 4, I indicated 23 important columns, hence I focused on finding patterns among these
- Step 1: Transform columns from strings to floats

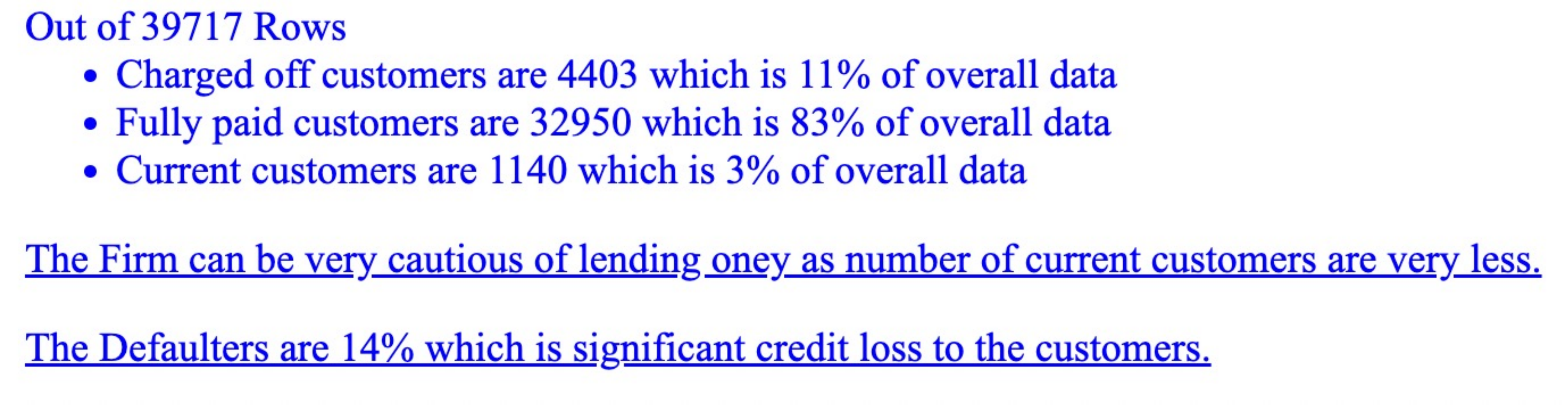
  The term is given as <NUMBER><SPACE><“MONTH”> -> <NUMBER>

  Remove ‘%’ symbol from rates like int_rate, revol_util
- Step 2: Remove columns having Nan values (Missing Values)
- Step 3: Remove outliers by quantile with high, low values as (0.01, 0.90)

  Fields like ‘annual_income’, ‘loan_amnt’ and ‘funded_amnt_inv’
- Step 4: Convert types to float

  Fields like ‘int_rate’ and ‘dti’ need to be float instead of string

Truth from cleaned Data!



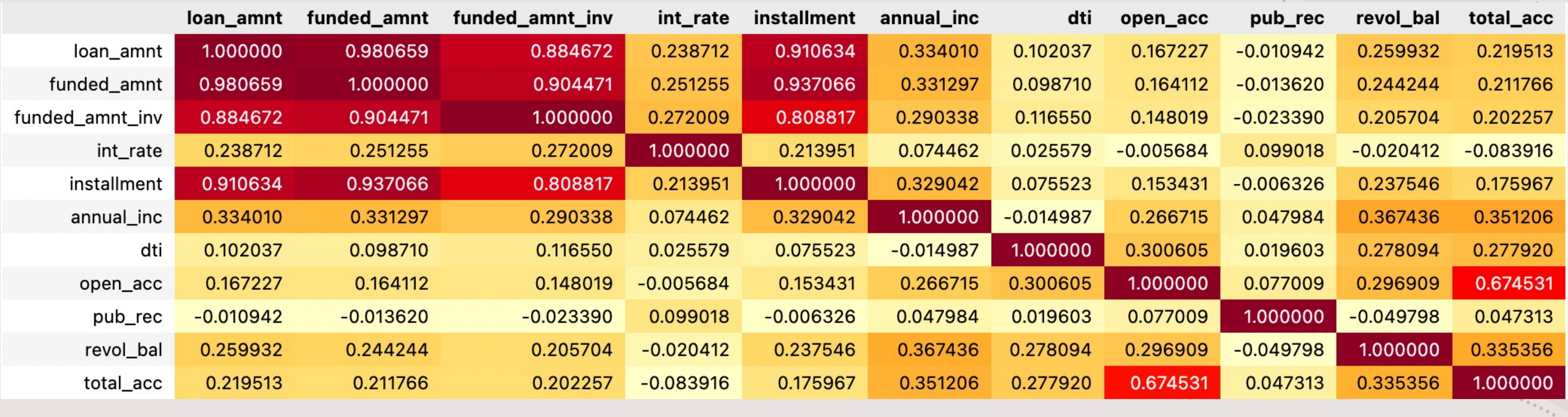
Out of 39717 Rows

- Charged off customers are 4403 which is 11% of overall data
- Fully paid customers are 32950 which is 83% of overall data
- Current customers are 1140 which is 3% of overall data

The Firm can be very cautious of lending oney as number of current customers are very less.

The Defaulters are 14% which is significant credit loss to the customers.

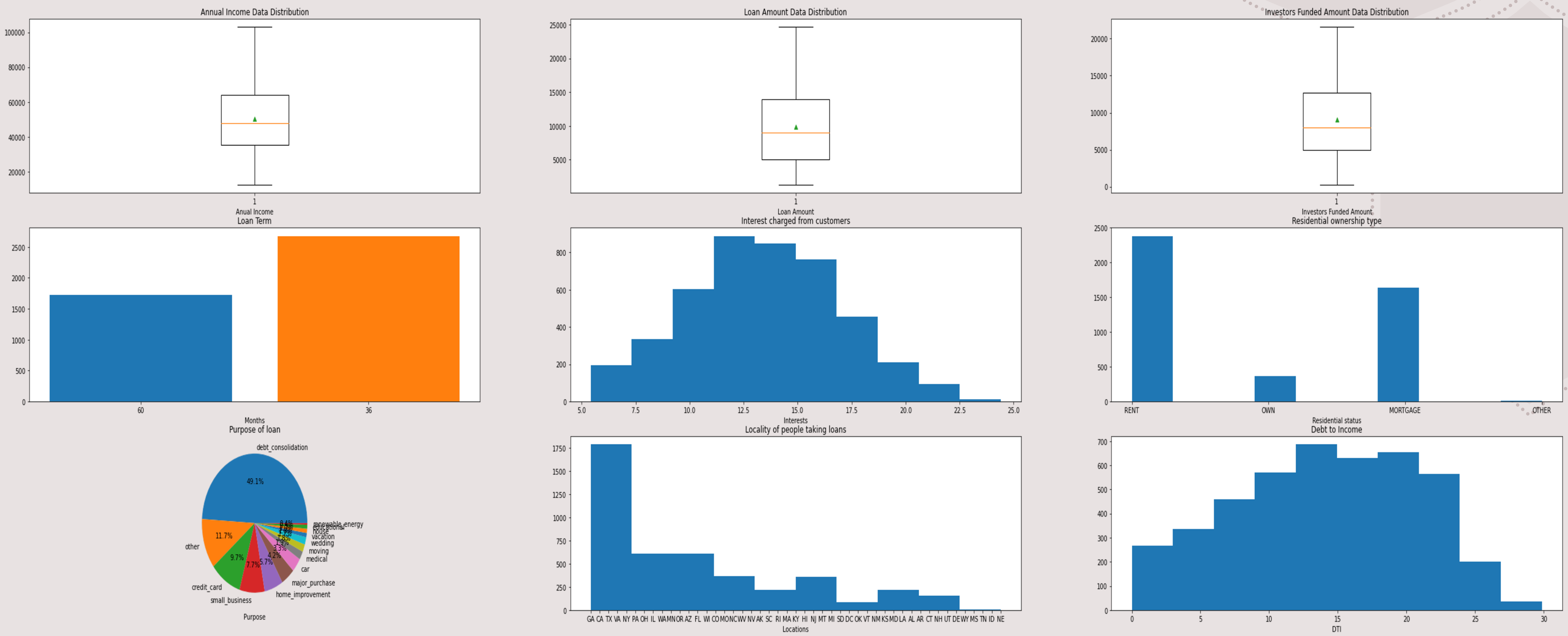
Correlation on cleaned data!



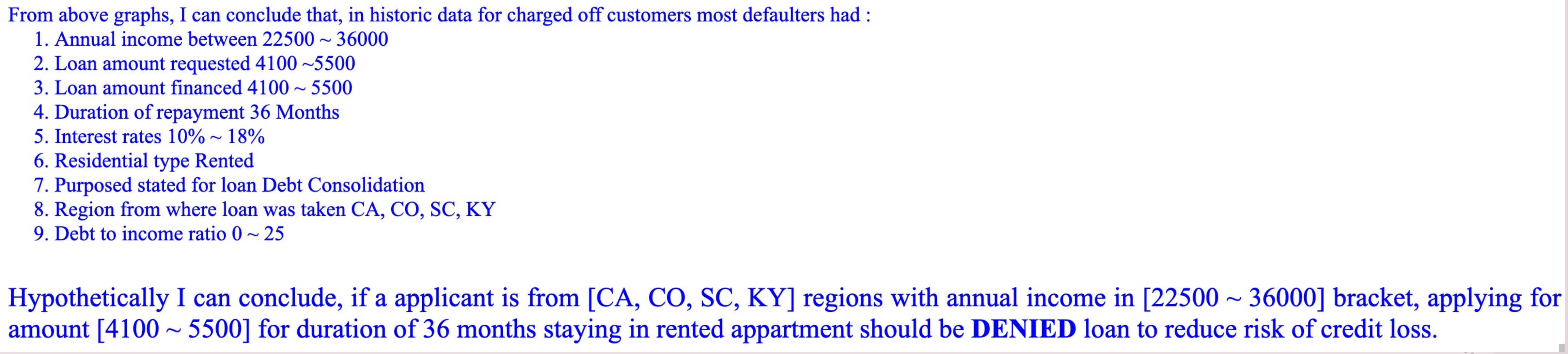
	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	open_acc	pub_rec	revol_bal	total_acc
loan_amnt	1.000000	0.980659	0.884672	0.238712	0.910634	0.334010	0.102037	0.167227	-0.010942	0.259932	0.219513
funded_amnt	0.980659	1.000000	0.904471	0.251255	0.937066	0.331297	0.098710	0.164112	-0.013620	0.244244	0.211766
funded_amnt_inv	0.884672	0.904471	1.000000	0.272009	0.808817	0.290338	0.116550	0.148019	-0.023390	0.205704	0.202257
int_rate	0.238712	0.251255	0.272009	1.000000	0.213951	0.074462	0.025579	-0.005684	0.099018	-0.020412	-0.083916
installment	0.910634	0.937066	0.808817	0.213951	1.000000	0.329042	0.075523	0.153431	-0.006326	0.237546	0.175967
annual_inc	0.334010	0.331297	0.290338	0.074462	0.329042	1.000000	-0.014987	0.266715	0.047984	0.367436	0.351206
dti	0.102037	0.098710	0.116550	0.025579	0.075523	-0.014987	1.000000	0.300605	0.019603	0.278094	0.277920
open_acc	0.167227	0.164112	0.148019	-0.005684	0.153431	0.266715	0.300605	1.000000	0.077009	0.296909	0.674531
pub_rec	-0.010942	-0.013620	-0.023390	0.099018	-0.006326	0.047984	0.019603	0.077009	1.000000	-0.049798	0.047313
revol_bal	0.259932	0.244244	0.205704	-0.020412	0.237546	0.367436	0.278094	0.296909	-0.049798	1.000000	0.335356
total_acc	0.219513	0.211766	0.202257	-0.083916	0.175967	0.351206	0.277920	0.674531	0.047313	0.335356	1.000000

- Funded_amount are strongly related to instalments
- Annual_inc has a decent correlation with a funded loan amount

Plotting to conclude hypothesis



Conclusion from plotting's!

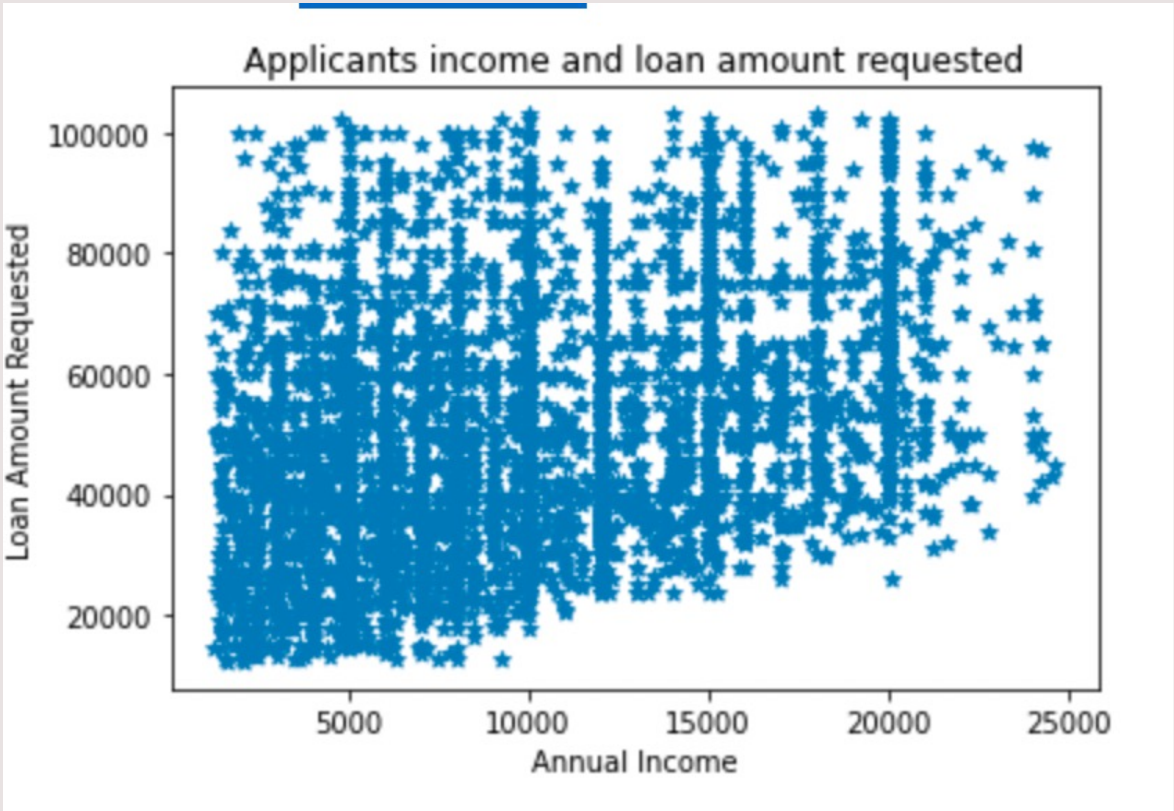


From above graphs, I can conclude that, in historic data for charged off customers most defaulters had :

1. Annual income between 22500 ~ 36000
2. Loan amount requested 4100 ~5500
3. Loan amount financed 4100 ~ 5500
4. Duration of repayment 36 Months
5. Interest rates 10% ~ 18%
6. Residential type Rented
7. Purposed stated for loan Debt Consolidation
8. Region from where loan was taken CA, CO, SC, KY
9. Debt to income ratio 0 ~ 25

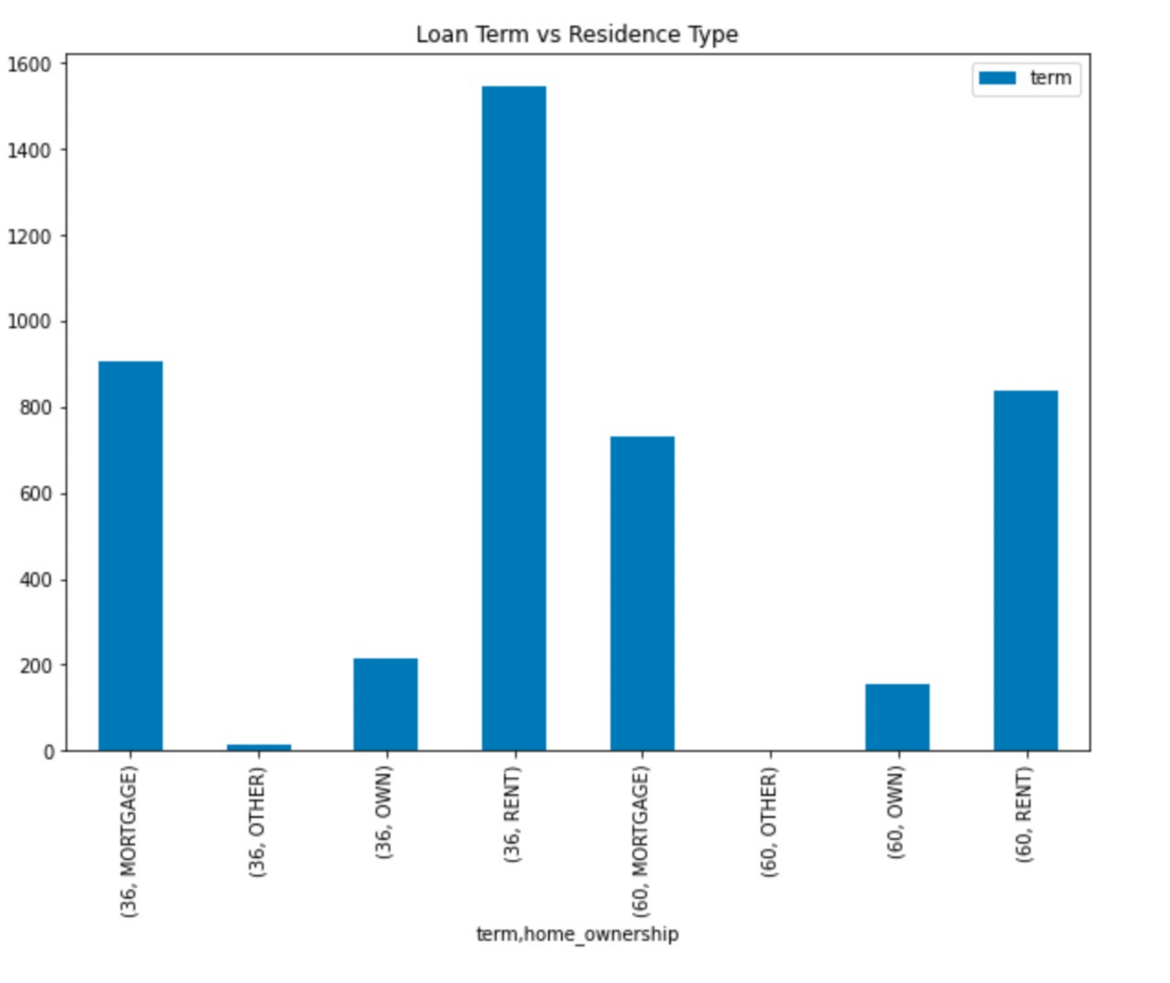
Hypothetically I can conclude, if a applicant is from [CA, CO, SC, KY] regions with annual income in [22500 ~ 36000] bracket, applying for amount [4100 ~ 5500] for duration of 36 months staying in rented appartment should be **DENIED** loan to reduce risk of credit loss.

Finding deep patterns #1



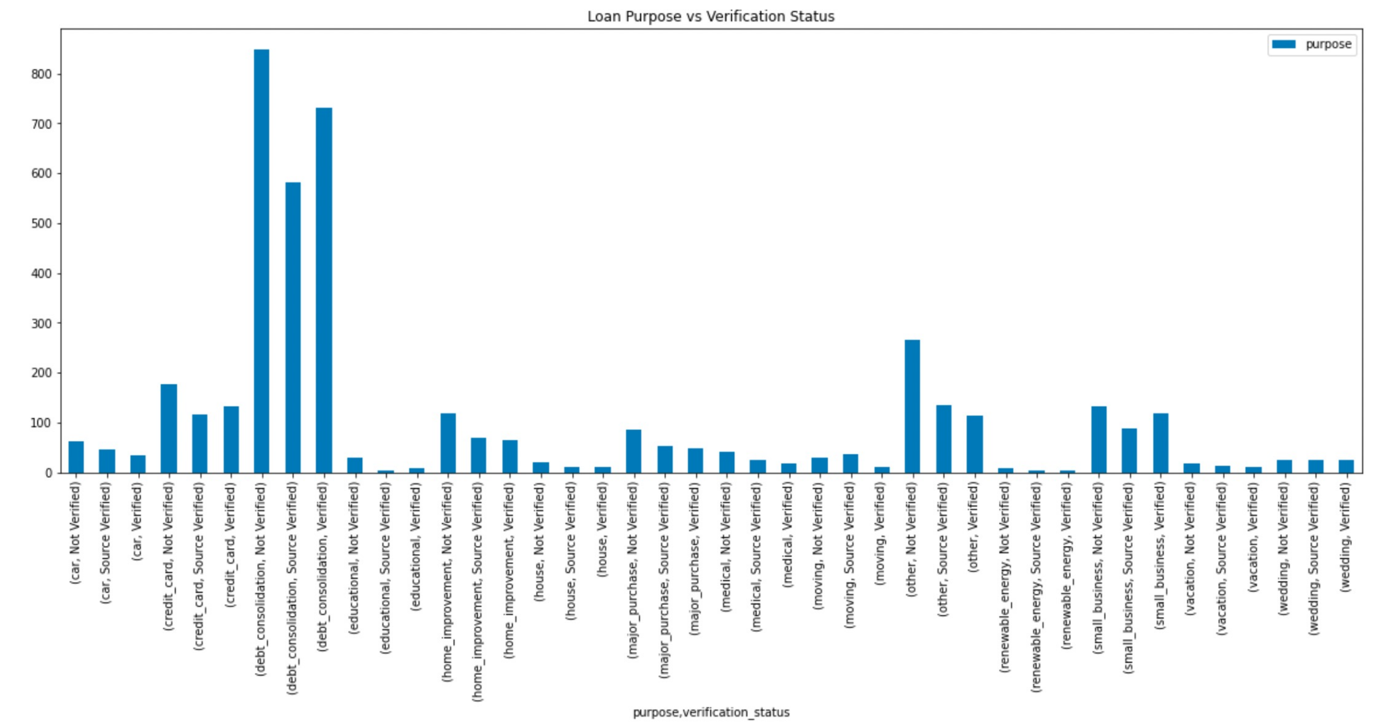
Among charged off applicants, Dense cluster is seen with annual salary above 1000 and below 10000. These applicants were applying loan in between 10000 to 55000.

Finding deep patterns #2



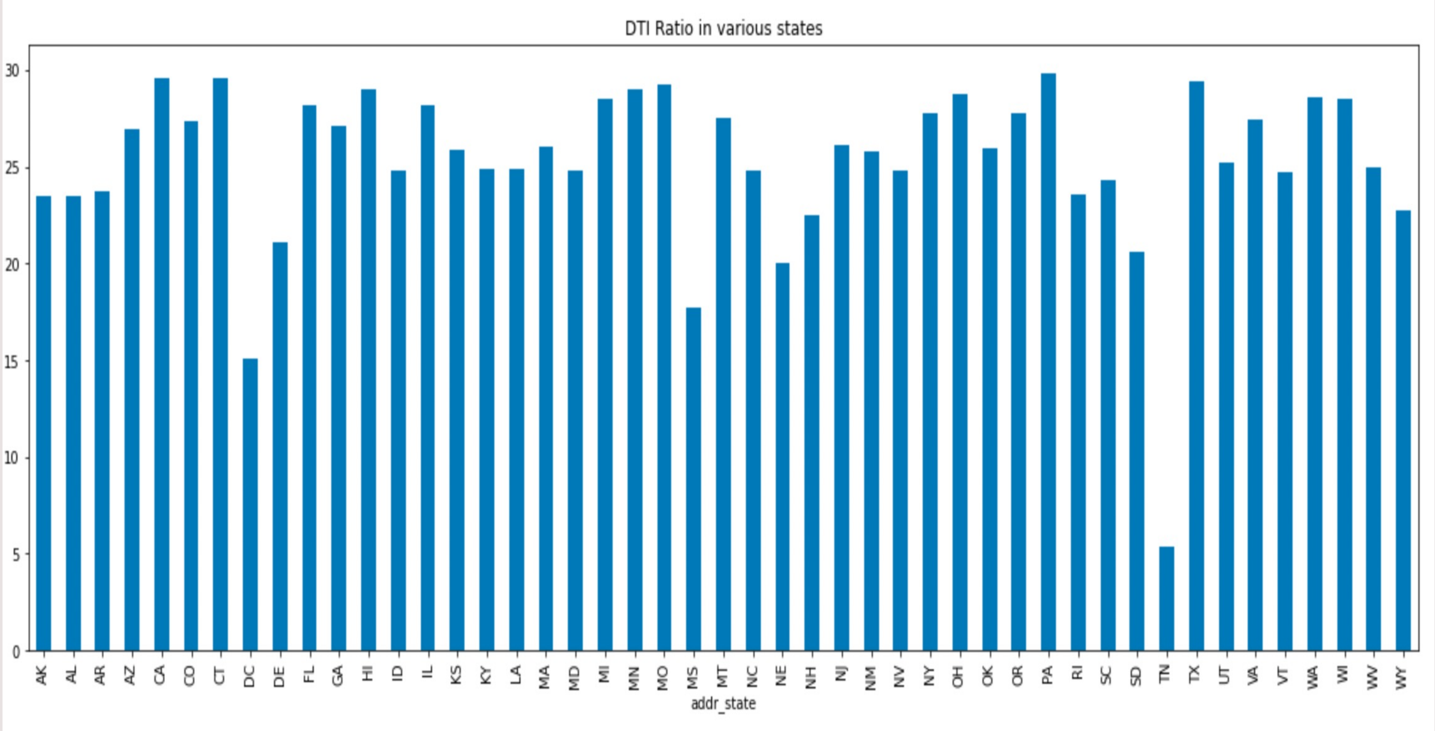
Among charged off applicants, Defaulters mostly have rented accomodation and has selected term to be 36 months for repayment.

Finding deep patterns #3



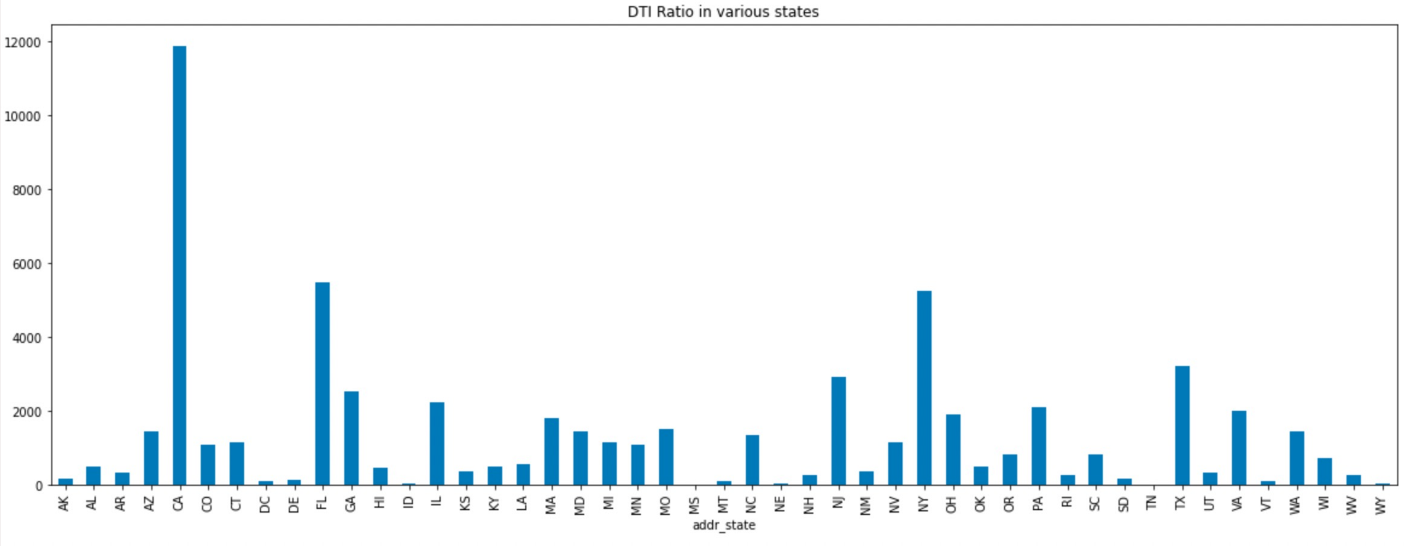
Many People have taken loan for debt_consolidation who have defaulted on loan. more then 800 of these were Not Verified, source verified and only 750 were verified.

Finding deep patterns #4



People from DC, MS, TN have lowest DTI ratio, hence I can conclude that defaulters belong to these states may be paying less debts from their income still they have defaulted the loan

Finding deep patterns #5



People from CA outnumbers the DTI ratio being highest, this indicates that people from CA are more in debt then any other states.

Thank You!