Lending Club

Upgrad -ACP AI/NLP- Case Study Submission

Sindhu N Kurup

Submitted by deadline-17/05/2023

Background

- This assignment is done as part of ACP AI/ML & NLP to implement the understanding of EDA (exploratory data analytics)

Problem Statement

- This case study Lending club refers to customer risk profiling challenge faced by a consumer finance company which specializes in lending various types of loans to urban customers.
- The data that has been provided as past loan applicants and whether they ‘defaulted’ or not.
- The aim is to identify patterns which indicate if a person is likely to default.
- This may be used by the company for taking actions such as
  - denying the loan,
  - reducing the amount of loan,
  - lending (to risky applicants) at a higher interest rate, etc.
- The data that has been provided for the period 2007 to 2011.
- The data contains , 3 scenarios , customers 1) Fully paid 2) Current (payment in progress) 3) Defaulted/Charged off

Domain Understanding

- Loan defaulting happens for 3 main reasons, and below can be broadly used for profiling the defaulters ;

1) Personal Attitude
   a) Multiple follow-ups ,though account balance seems fine

      Columns mapping - avg_cur_bal Vs no of inquiries

   b) Too many loans & commitment , in proper planning

      Columns mapping - FICO score , No of personal finance inquiries , Revolving credit balance , No of finance trades

2) Financial downgrade – job loss , property loss , financial loss , Accidents , natural calamities , health issues , theft ,delayed salary :Columns mapping
   a) No of mortgage accounts
   b) Account Balance at the time of opening vs now
   c) Income verification status
   d) Total number of credit lines

3) Fraud
   a) Non existent or relocation without intimation , absconding
      i. Permanent vs temporary address
      ii. Backup address availability
      iii. Home ownership
      iv. Income verification status
      v. Balance on installment accounts
      vi. Total num of credit lines

Data Analysis Approach

Data Cleansing
- Removed columns with no values – 54-111 columns
- Removed rows 'current' as it is not helping in default analysis
- After initial comparison of data , selected columns as in next section
- Removed outliers for annual income , loan amount etc
- Interest rate – removed % symbol , corrected type
- Term was analysed
- Many numerical values were null – had to drop off as mentioned in next section
- No of revolving accounts

Opted for Segmented univariate analysis
- That is ,Took only rows where loan status is "charged off"
- Analysed the effect of loan grade, subgrade columns on the default status
- Loan status vs interest rates impact

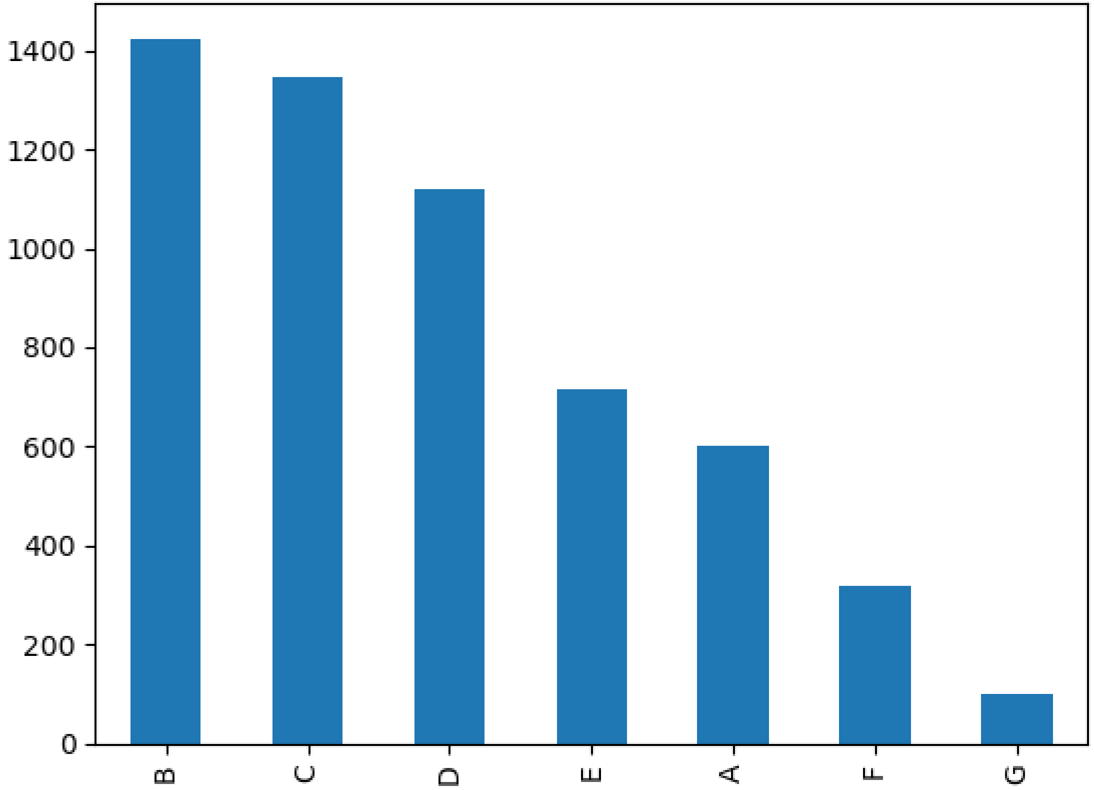
Type driven metrics
- Grade /Subgrade
- Term
- Emp_length
- verification_status
- Home_ownership

Business driven
- annual income Vs interest rate
- revol_util vs loan amount

Data driven metric
- Binning interest rate against no of defaulters

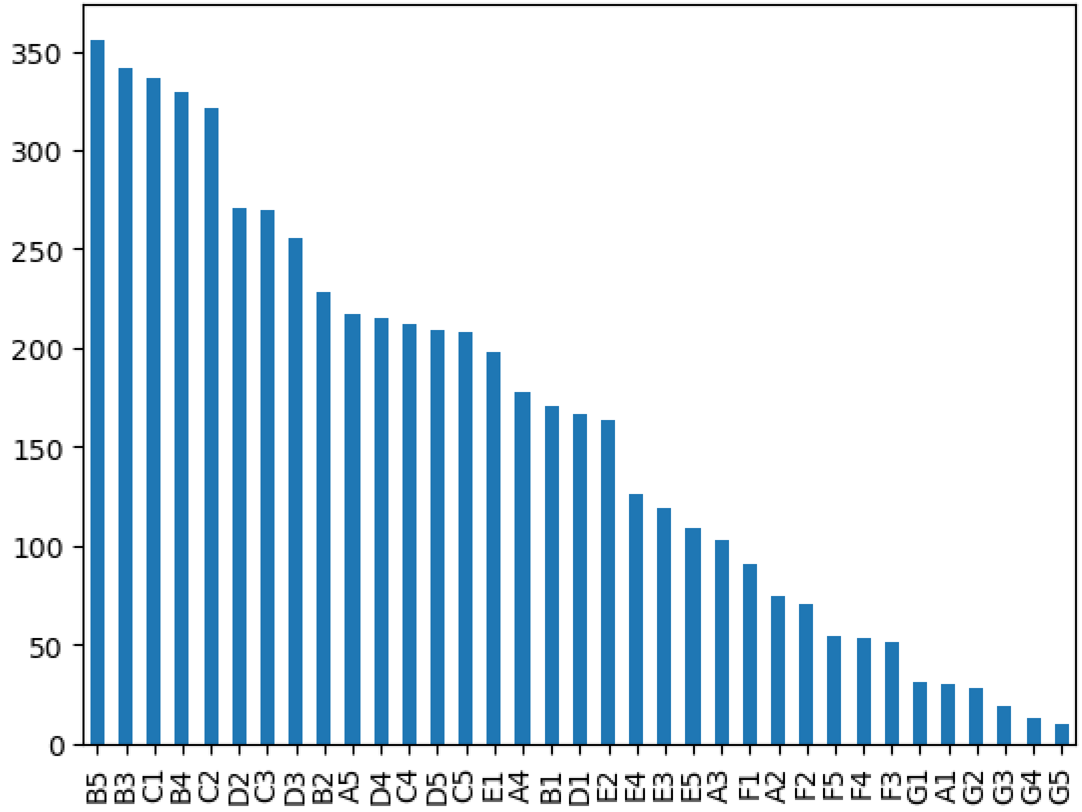
Inference

Plotting hist of grade and sub grade



Inference : There seems an in increasing num of defaulters where the loan grade is B.

Plotting hist of grade and sub grade



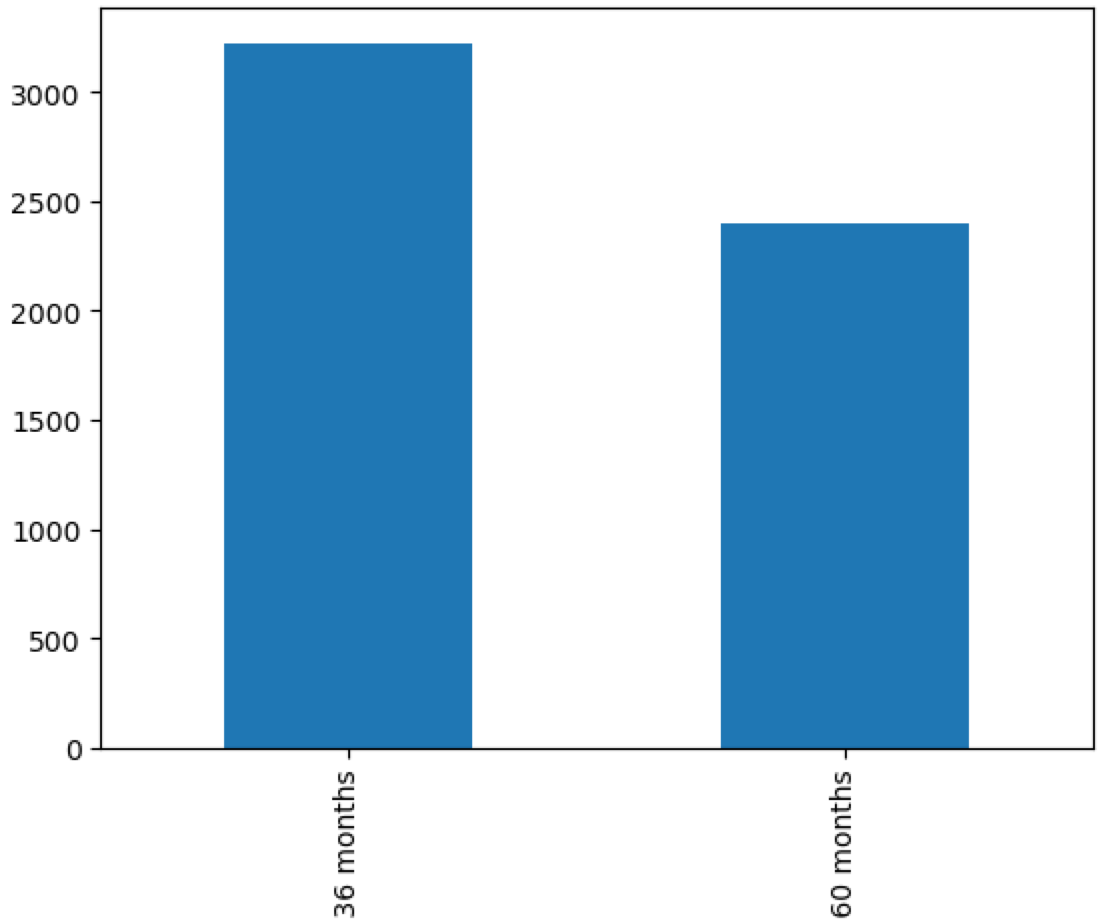
Inference : Grade B5, B3, C1, B4,C2 seems to be having highest number of defaulters

percent_bc_gt_75 , out_prncp_inv

The below were identified as potential columns for analysis to check bank limit and outstanding amount had any co-relation , but both had null values and couldnt use.
percent_bc_gt_75 -Percentage of all bankcard accounts > 75% of limit - had null values - hence couldnt infer much out_prncp_inv- Remaining outstanding principal for portion of total amount funded by investors tot_cur_bal = Total current balance of all accounts

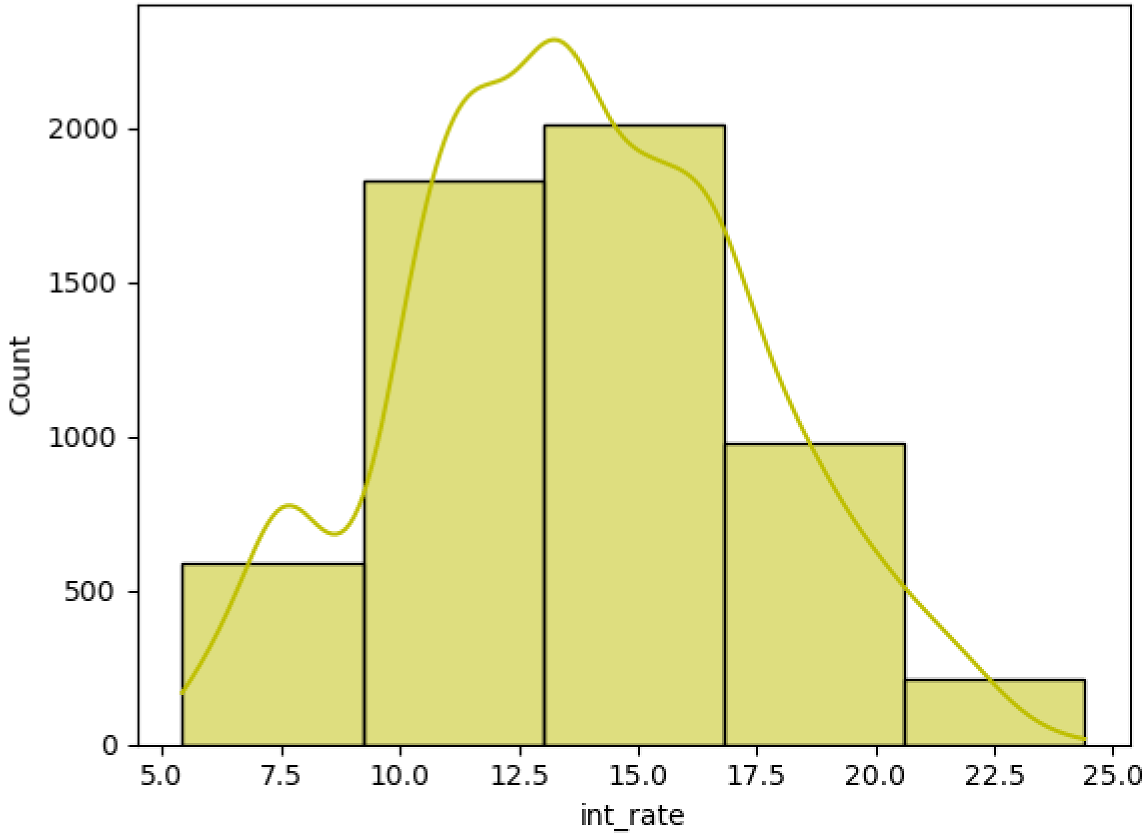
Inference : No conclusions

loan term



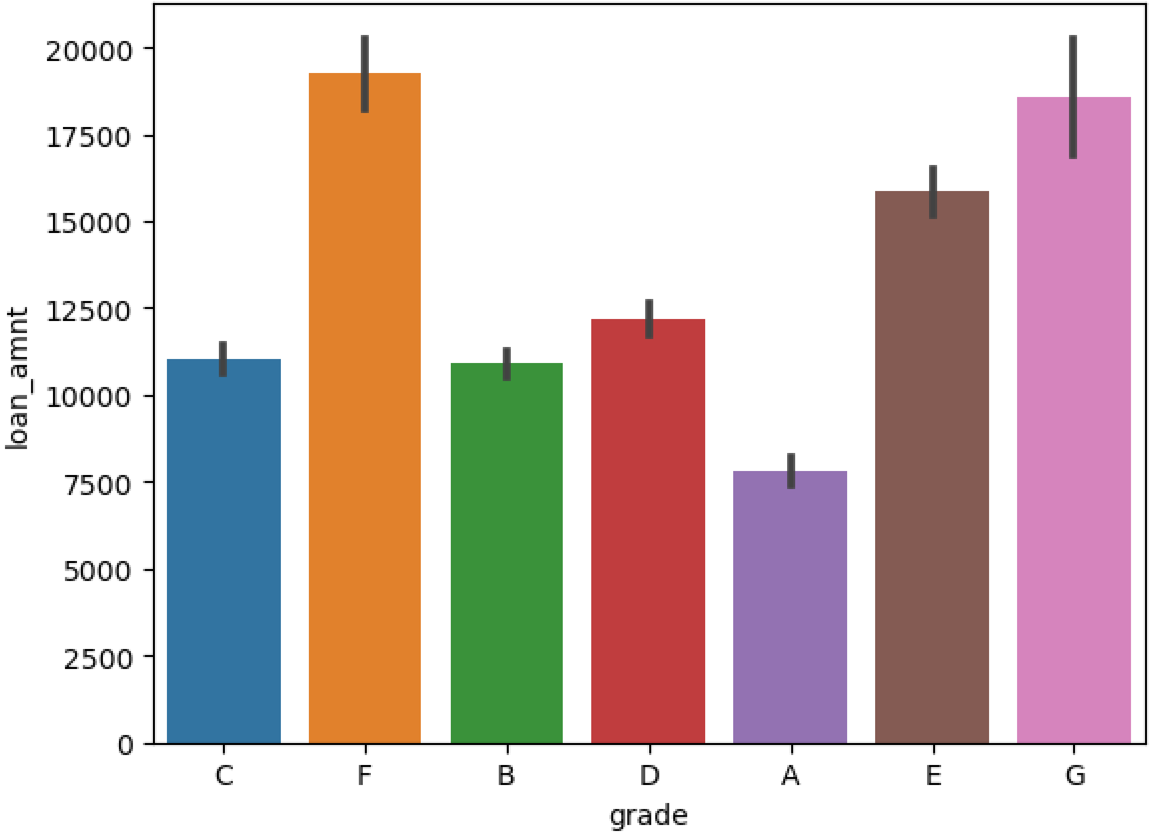
Inference : Shorter the term more defaulters

interest rate



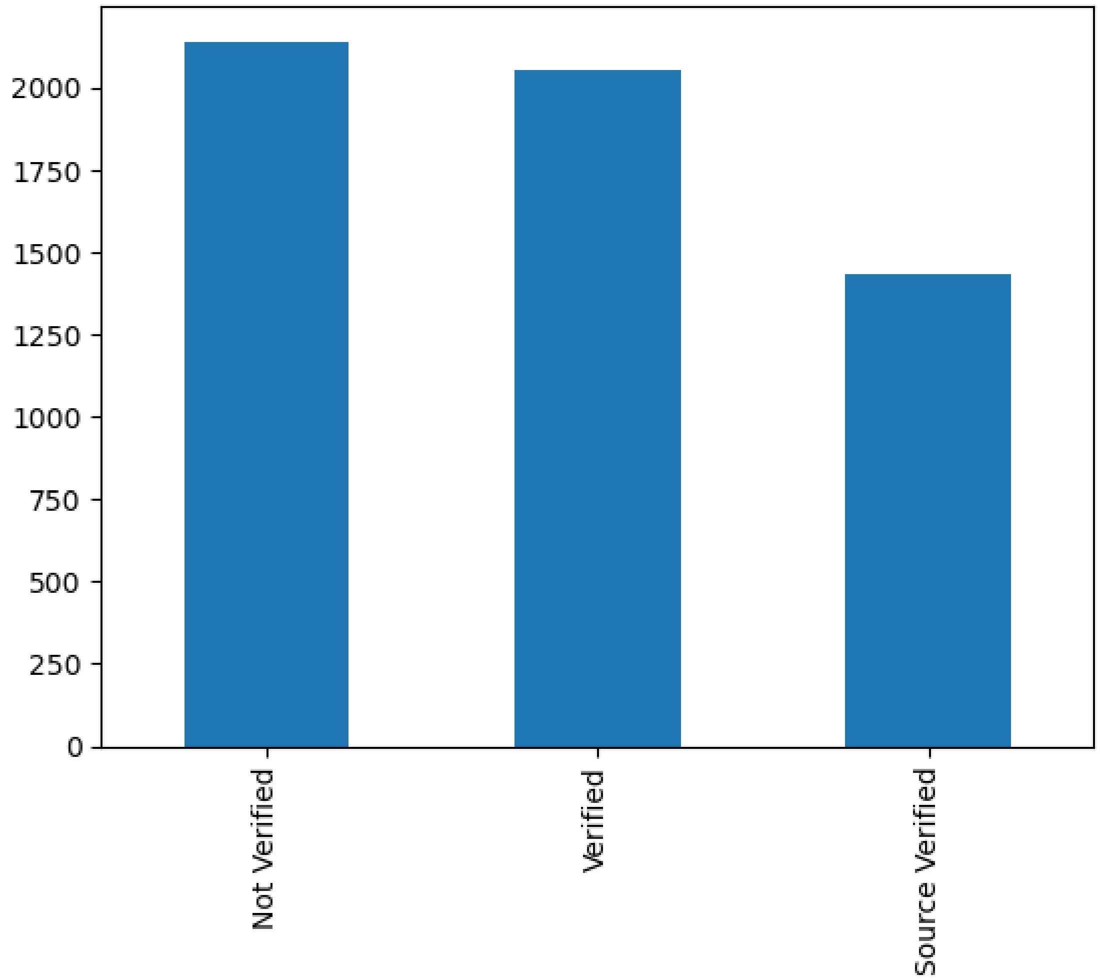
Inference : Most of the defaulters are in the interest rate 15 ..

Loan amount vs grade



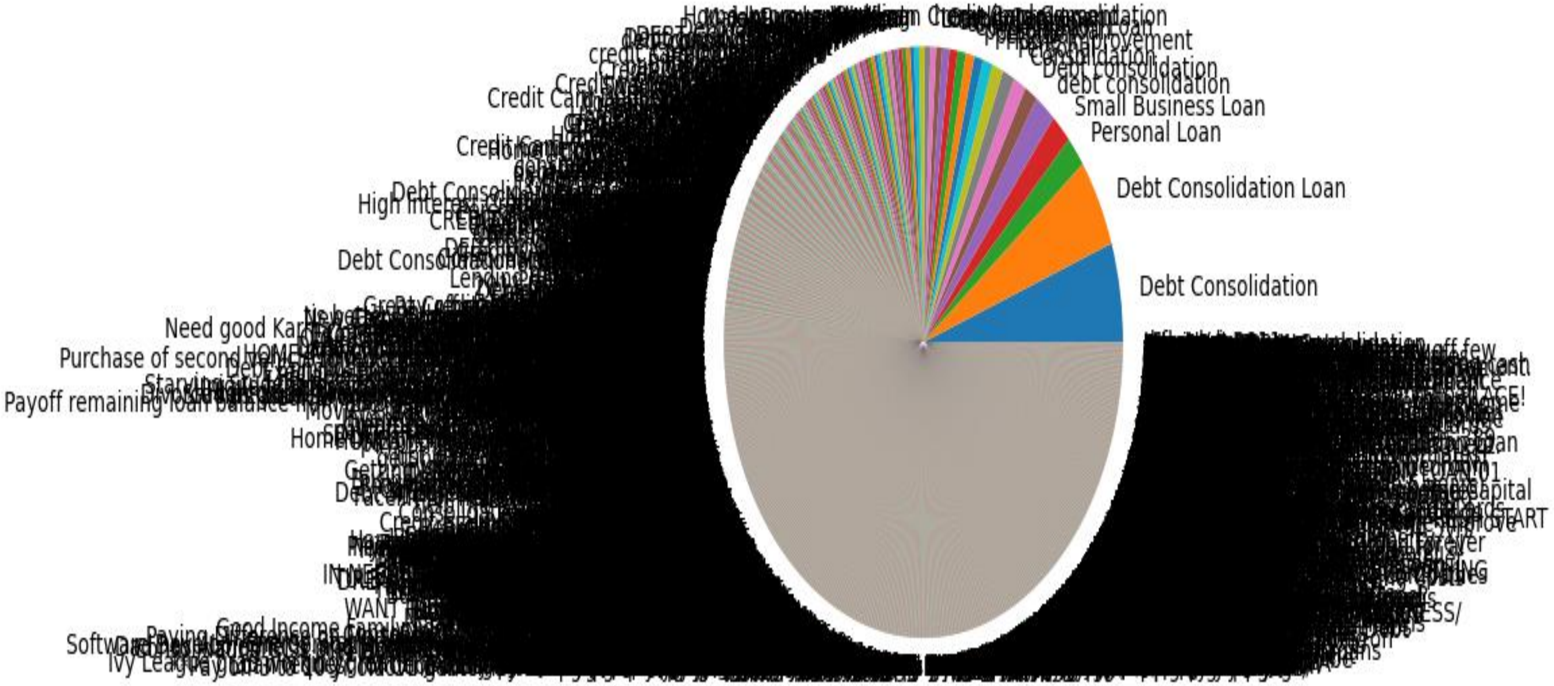
Inference: loan_amt for grade F and Grade G seems to have some co-relation , which can studied further

Income verification status and defaulter spread



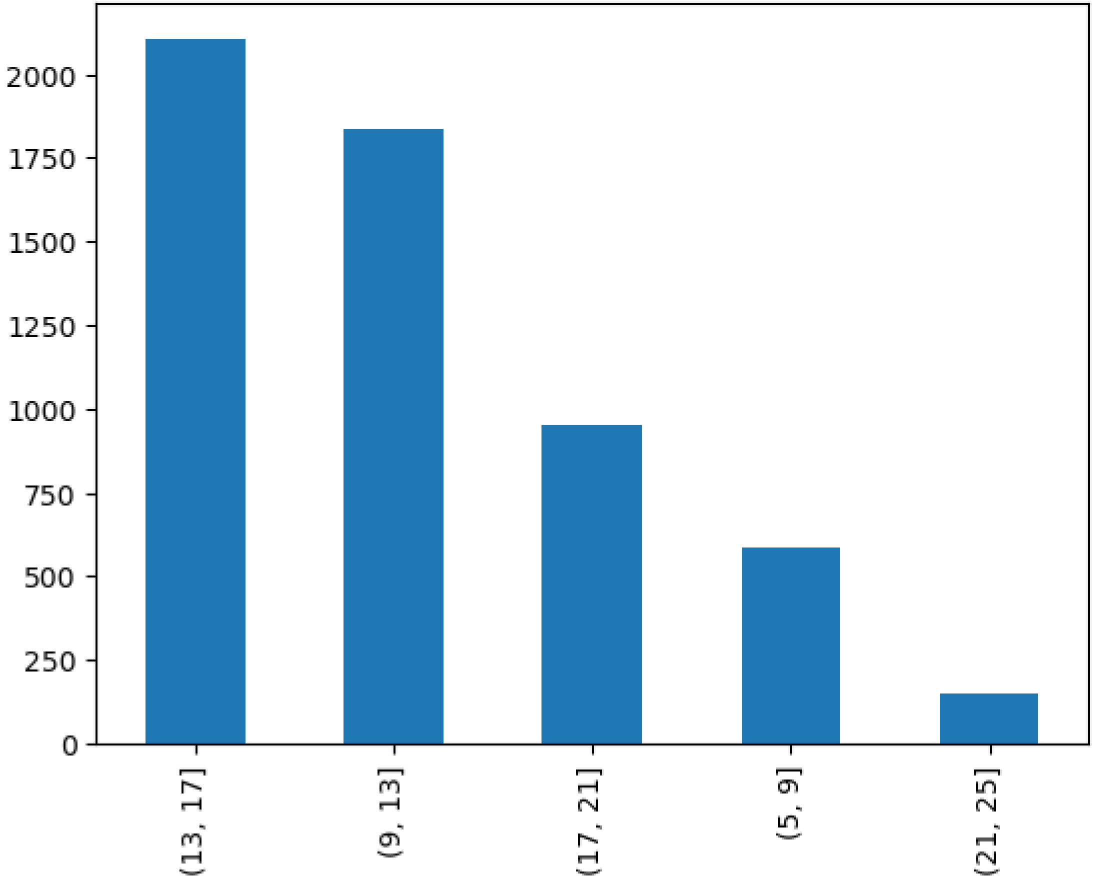
Inference :Most of the defaulters have in the status Not verified , which can be studied further

Title vs defaulters



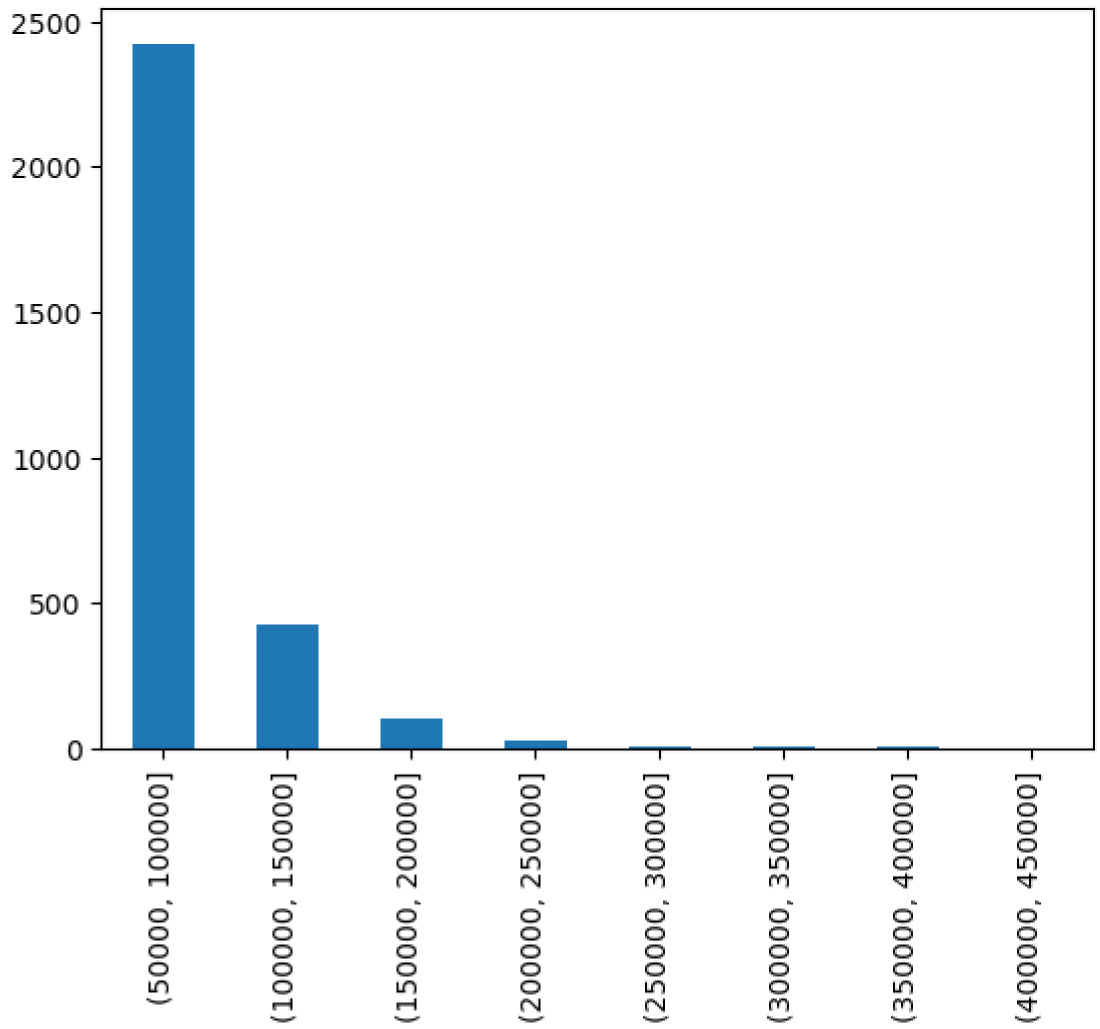
Inference : Top 3 titles are small Business loan , Debt consolidateion , Debt Consolidation loan , these categories might need further investigation.

interest rate - binning



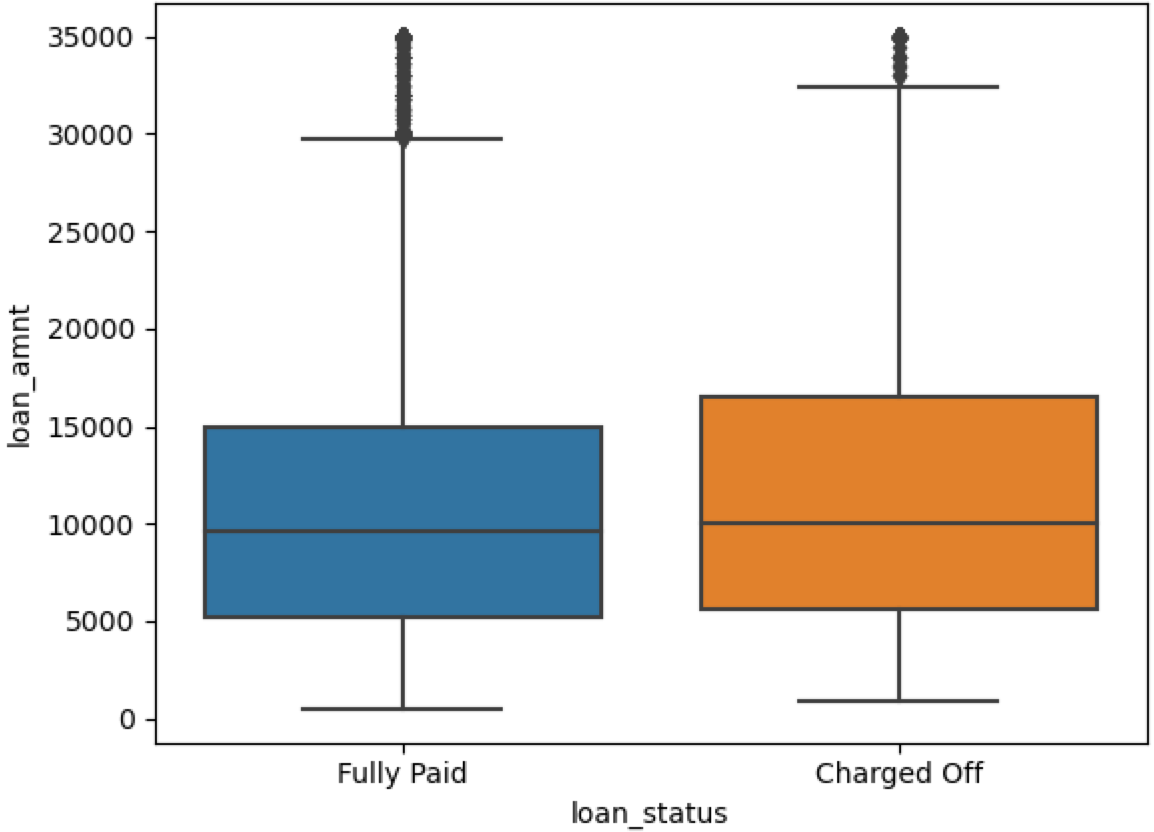
Inference : Most falling under 13-17% category

annual income - binning



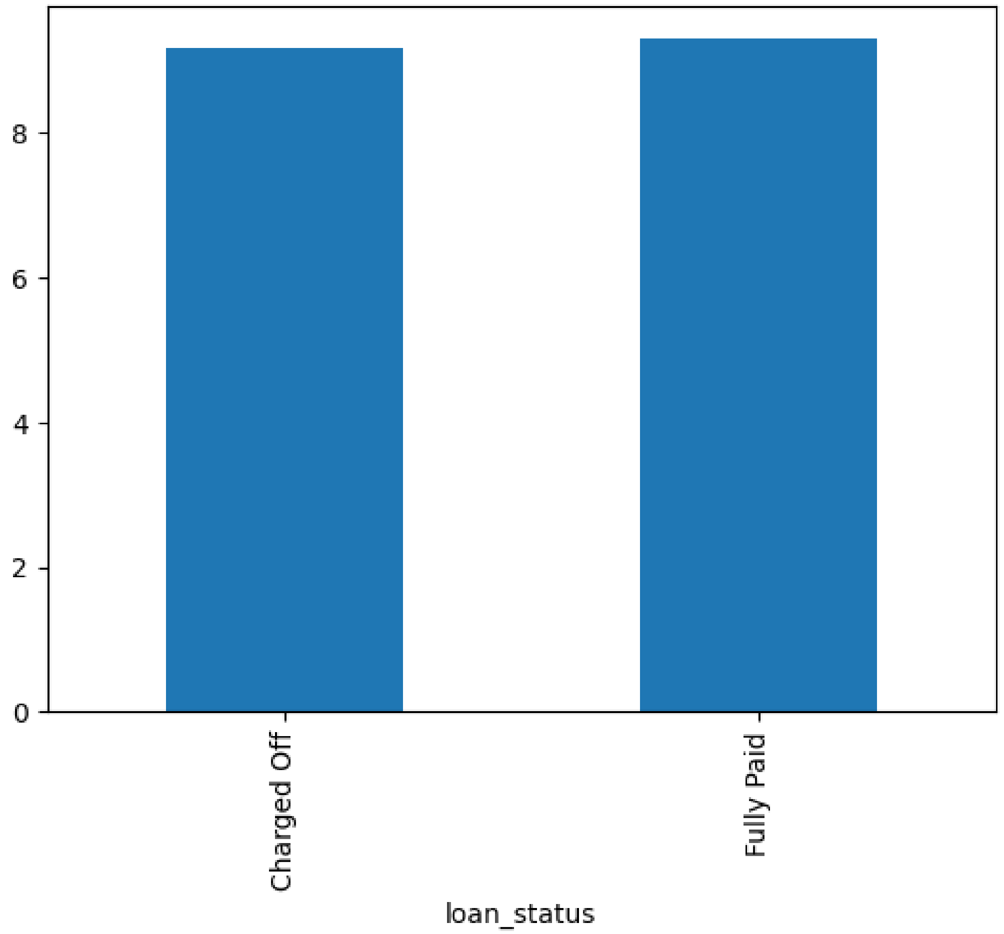
Inference : Most of the defaulters are in the lowest income range , which needs further study

Loan amount vs Loan status



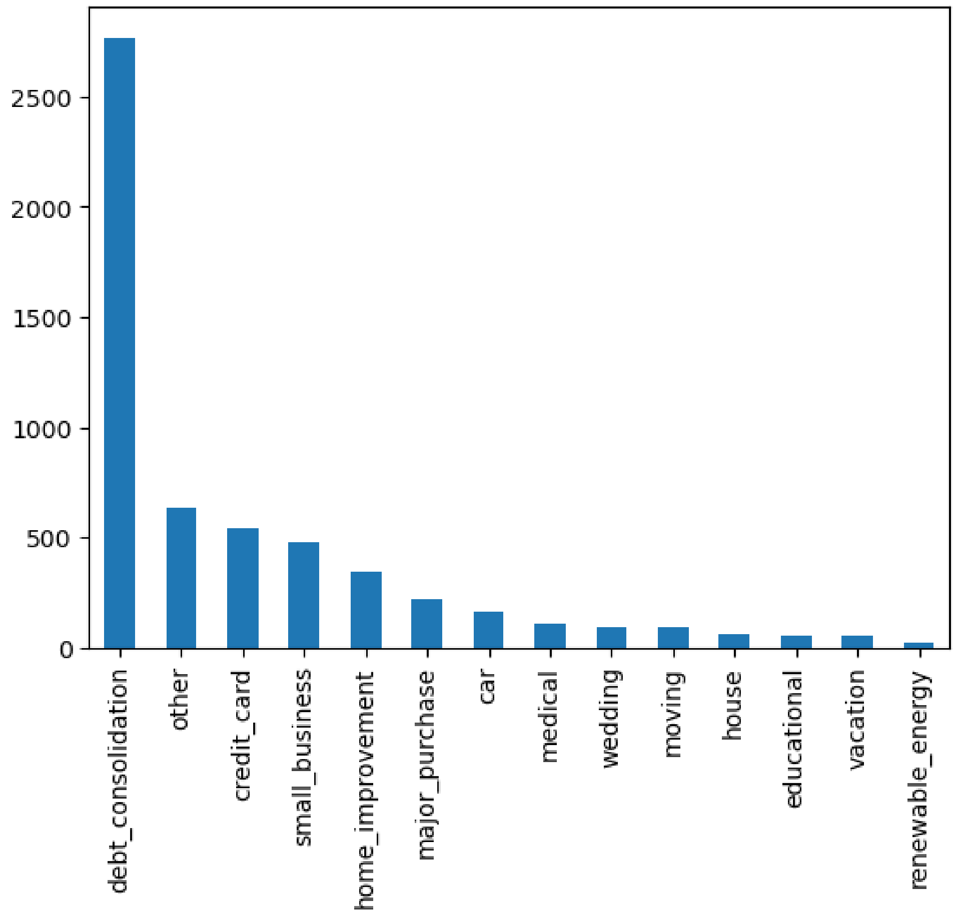
Inference : 75% of loan amount taken by Charged Off status is higher than the 75% of loan amount taken by fully paid , is there any rules to be applied on loan amount. It seems like people with higher amount has defaulted

checking the defaulters against no of credit line.



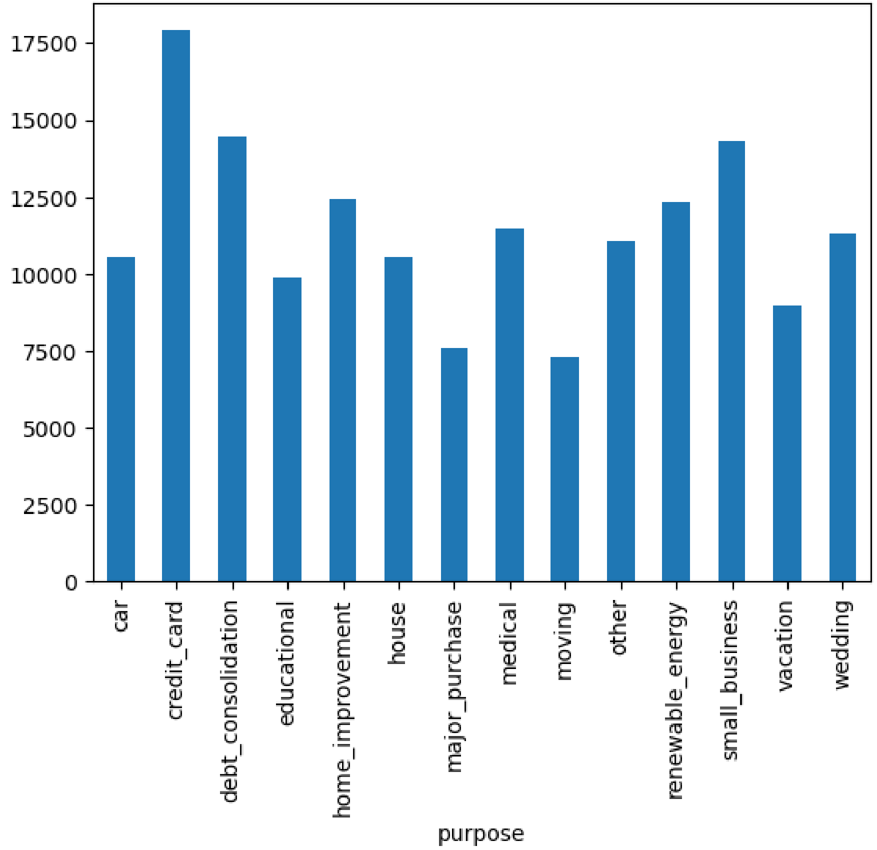
Inference : The number of credit lines doesnt seem to be having any co-relation with defaulters ; fully paid members seems to have an average 9. something which almost same as those of Charged Off members

Purpose



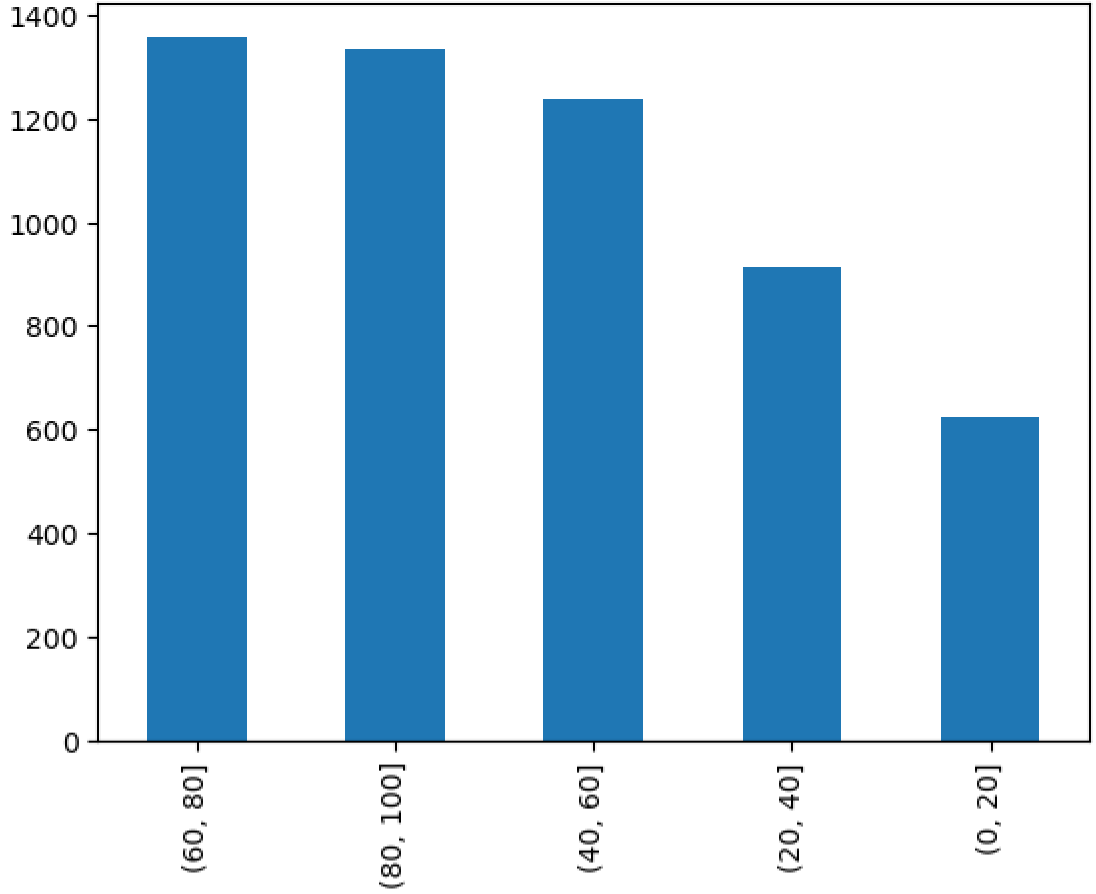
Inference - Most of defaulters seems to be falling under purpose debt consolidation , may this category needs bit more scrutiny

#checking wehther purpose has anything to be income bins for defaulters



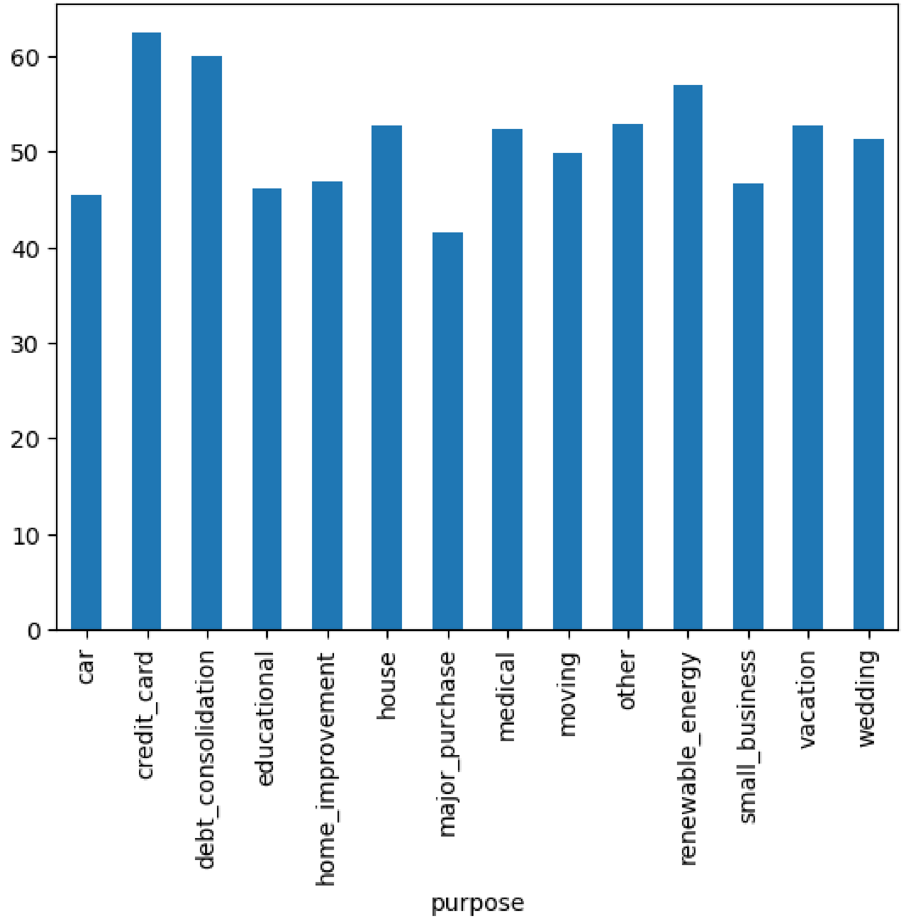
Inference : The number of credit lines doesnt seem to be having any co-relation with defaulters ; fully paid members seems to have an average 9. something which almost same as those of Charged Off members

#checking revol util and defaulters



Inference : Most of defaulters are in the revol_util (Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.) of 60-80%

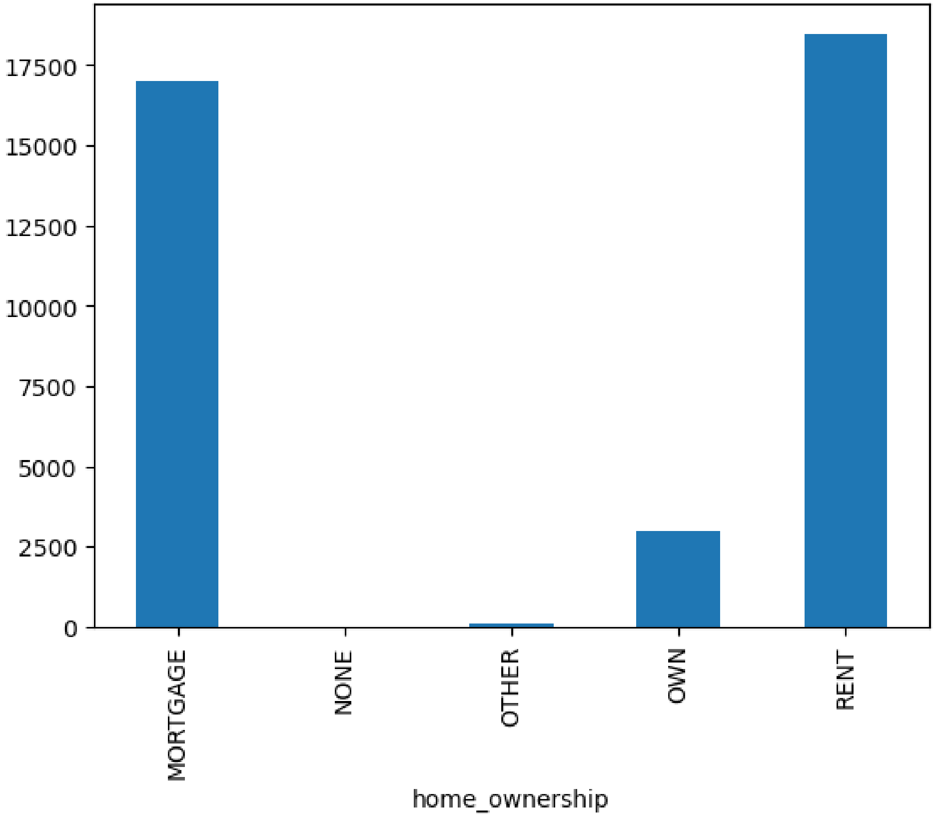
#checking whether purpose and revol util has any co-relation



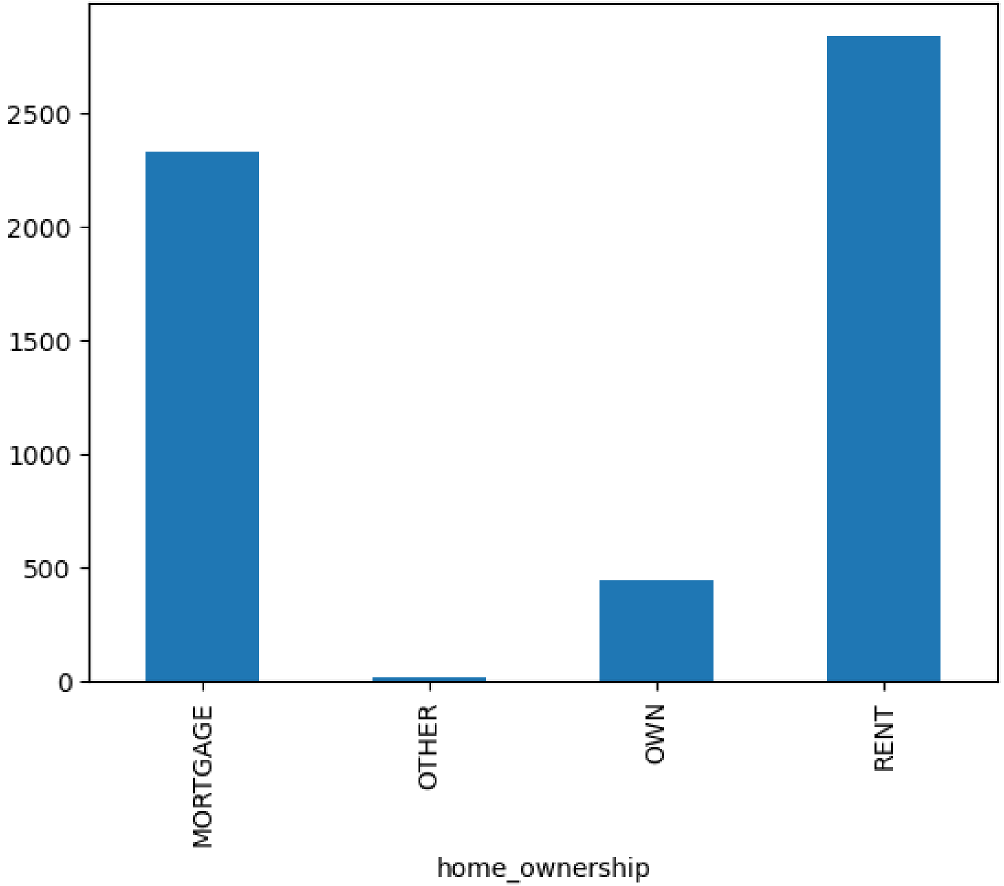
Inference : Those borrowing money for credit card payment (?) and debt consolidation seems to be having 60% revol util - which seems to be where defaulter count is concentrated

#home ownership vs loan status

Fully Charged + Charged Off



Charged Off



Inference : Couldn’t find much co-relation of whether defaulters are those who stay in rented house in isolation

Risk Profiling Vs Results (1 of 3)

Profile Category	Analysis	What	Columns used	Approach	Conclusion
Attitude /Habitual	Analysis 1- Univariate	Whether loan grade has an impact	sub_grade, grade	Barplot for grade and sub_grade to check the spread	Grade B has most defaulters Grade B5, B3, C1, B4,C2 seems to be having highest number of defaulters
	Analysis 2- Univariate	Whether the term has an impact	term	Bar plot	Term seems to be having an impact As shorter the term- more defaulters
	Analysis 3- Bivariate	No of revolving accounts vs delinq amount	num_op_rev_tl Vs delinq_amnt	These figures not available for analysis – hence couldn't infer anything	None
	Analysis 4- bivariate	Total balance vs outstanding principal	out_prncp vs tot_cur_bal	These figures not available for analysis – hence couldn't infer anything	None
	Analysis 5 – Univariate	Analysis bank accounts with >75 % limit	percent_bc_gt_75	These figures not available for analysis – hence couldn't infer anything	None
	Analysis 7- Univariate	Income	annual_inc	Checking the income range	Yes, lowest income members seems to be defaulters mostly

Risk Profiling Vs Results (2 of 3)

Profile Category	Analysis	What ?	Columns used	Approach	Conclusion
Financial	Analysis 1(Univariate)	Whether income source was verified	verification_status	Checking which verification status contributes to most defaulters	Falls under 'Not Verified ' status
	Analysis 2(univariate)	Living in own house or rented house	Analyse the column home_ownership and annual income	Finding any relation with income and home ownership	Couldn't conlude
	Analysis 3(bivariate)	Income range	Analyse total income with outstanding amount	Null column for outstanding amount	No inference
	Analysis 4(bivariate)	Employement status	emp_length vs out_prncp	Null column for outstanding principal	No inference
	Analysis5(bivariate)	Trade delinquent % vs outstanding amount	pct_tl_nvr_dlq	This column is NA	Couldn't proceed with this check

Risk Profiling Vs Results (3 of 3)

Profile Category	Analysis	What	Columns used	Approach	Conclusion
Fraud	Analysis 1- univariate	Whetther bankruptsy ahs cause person to absond	pub_rec_bankruptcies	NA	This data was NA
	Analysis 2- univariate	Any tax default actions	tax_liens	NA	This data was NA
	Analysis 3- univariate	Title	title	Checking whether employment length has any co-relation	Those employed for more than 10 years seems to be the most defaulters
	Analysis 4 – Univarite	Purpose of loan	Purpose	Checking whether purpose has any co-relation	Yes, purpose could be attributed to loan status