upGrad

#LifeKoKaroLift

Lending Club Case Study: Pre-Assignment Session

13-07-2022

upGrad

Course : Machine Learning

Lecture On : Lending Club Case Study

Instructor : Siddhesh Gunjal

Data understanding

upGrad

- The dataset had close to 110+ columns. By going through the descriptions of each column, it looked like we needed to choose selected columns for the analysis. It also had 39717 rows.
  ('Shape :', (39717, 111)) ('#Rows :', 39717) ('#Columns :', 111).
- Some of the column variables were dealing with the client data which had to be recorded once the loan application is accepted and processed. Eg: *total_rec_int,total_rec_late_fee,etc*
- Since our use-case is meant to focus on the variables which helps us to make a decision for loan approval based on the applicant's profile, we dropped the following columns :
  delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, 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_pymnt_amnt, last_credit_pull_d, application_type
- Also some of the loan variables directly dealt with loan variables like: *amount of loan, interest rate,etc* . This can be directly used for analysis.
- Some of the variables were not directly related to loan charatericts like: *employment details , grade,etc.* This can be coupled with direct loan variable for analysis .

Data cleaning

upGrad

- Looking at the data and we see there are many columns which may not be usefull and similary there will be many rows which may be having invalid data.
  Now lets do some data cleaning and try to keep only data which will help us.
- Any column where the %of missing data was more than 80 had been dropped. The columns were- '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'
- *Desc , emp_title and mths_since_last_delinq* columns were again dropped since it dindt provide us much information.

Data cleaning

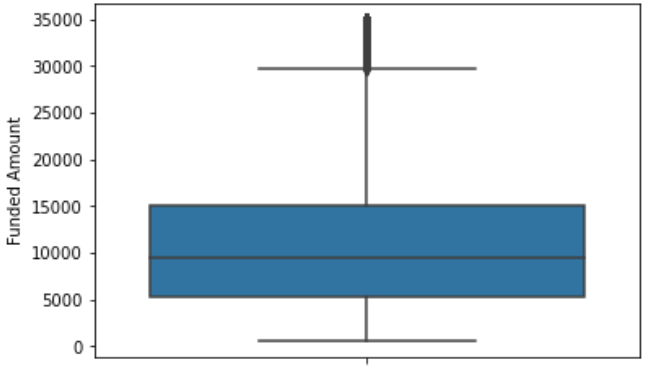
upGrad

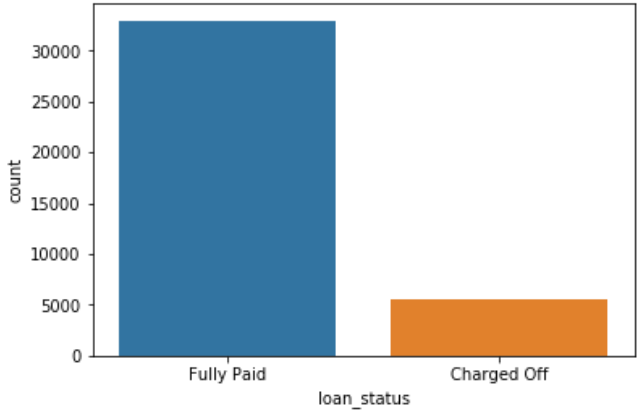
- Lets see fully paid vs ChargedOff row count value*:*
  Fully Paid 32950
  Charged Off 5627
  Current 1140
- We do not need Current (ongoing loan) so removed them from the dataset
- Create a new column *term_val* from term which just has number values of the tenure(36/60)

Data Analysis – Univariate Analysis

upGrad

1. Box plot of *funded_amnt . We can see that 75% of the loan amount funded lies below 30k.*



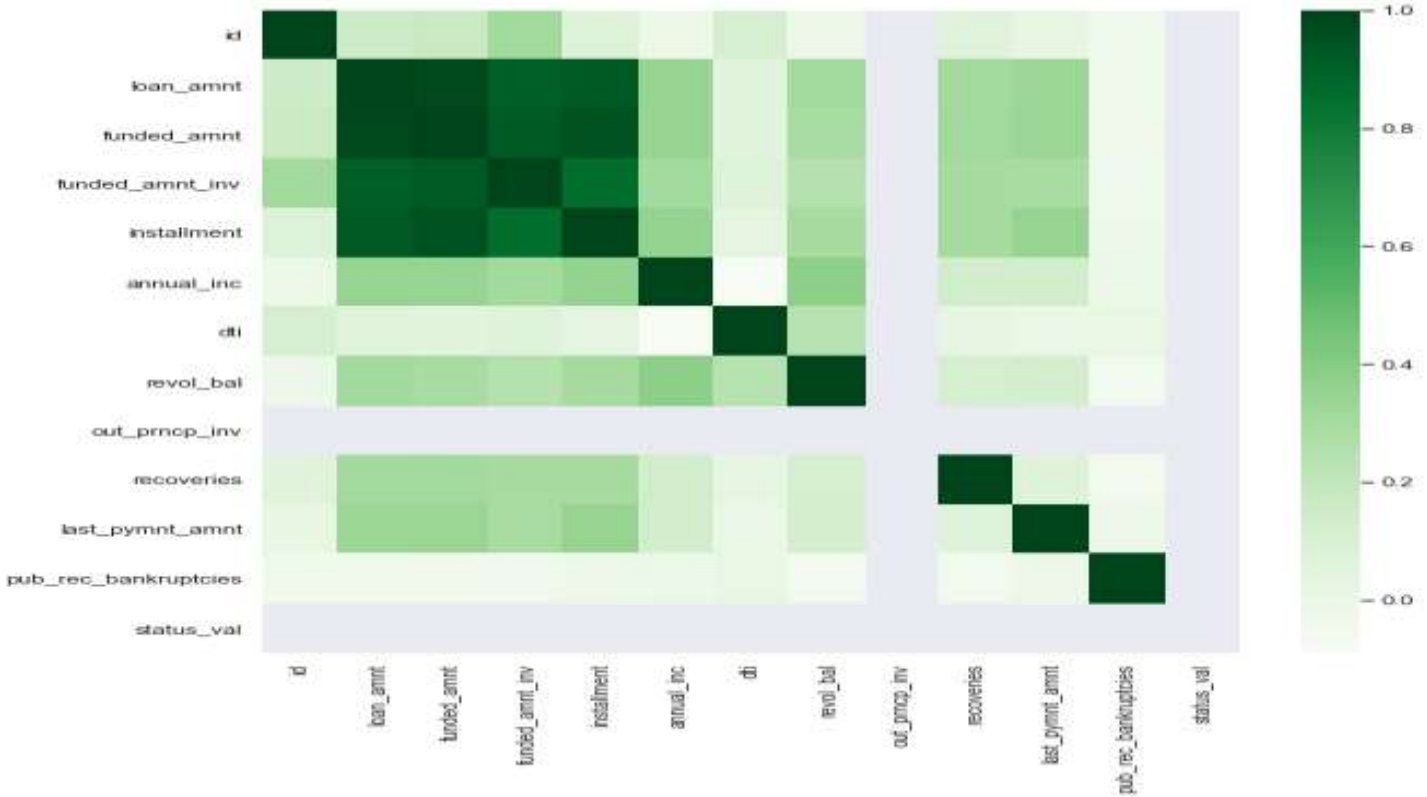


2. Count plot of *loan_status*. We can see the 14% of the total number of loans are defaulted.

Data Analysis – Univariate Analysis

upGrad

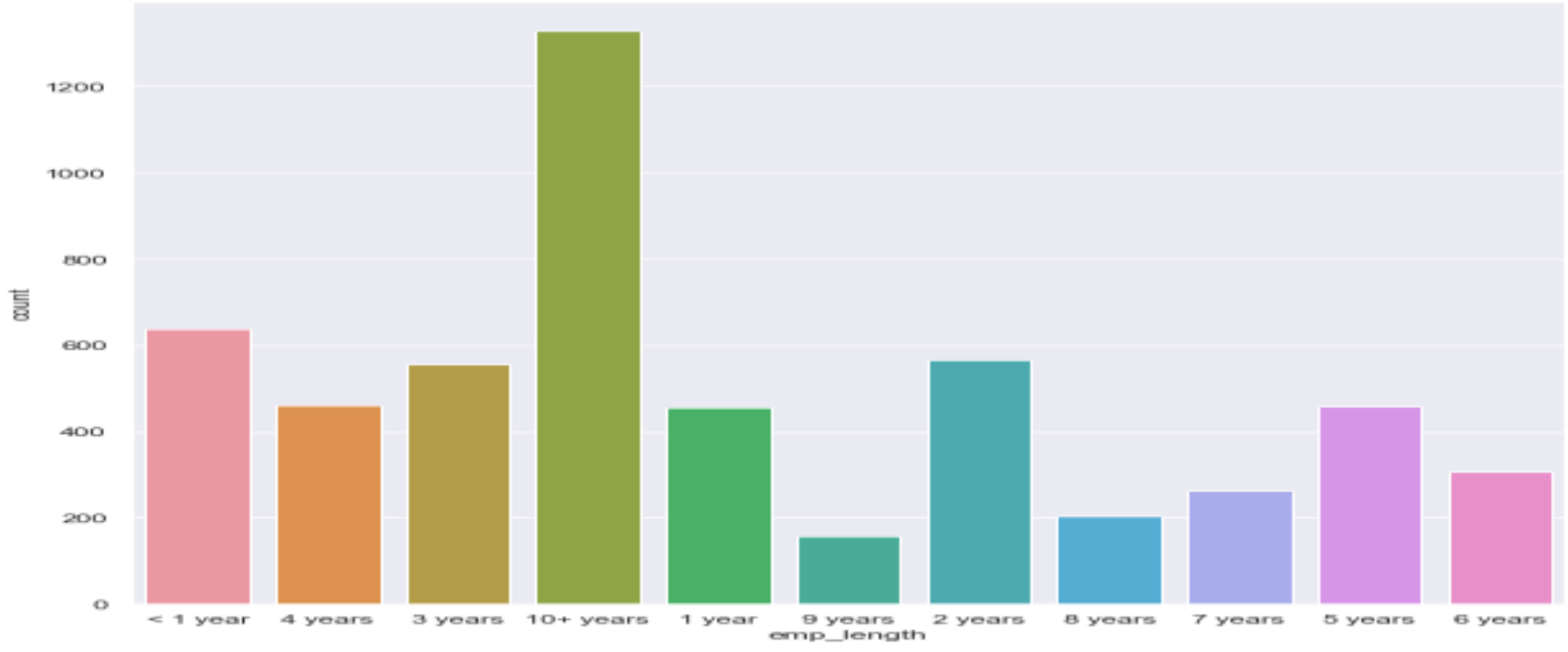
3. Heatmap across various variables. we can see dark areas have high correlation.



Data Analysis – Bivariate Analysis

upGrad

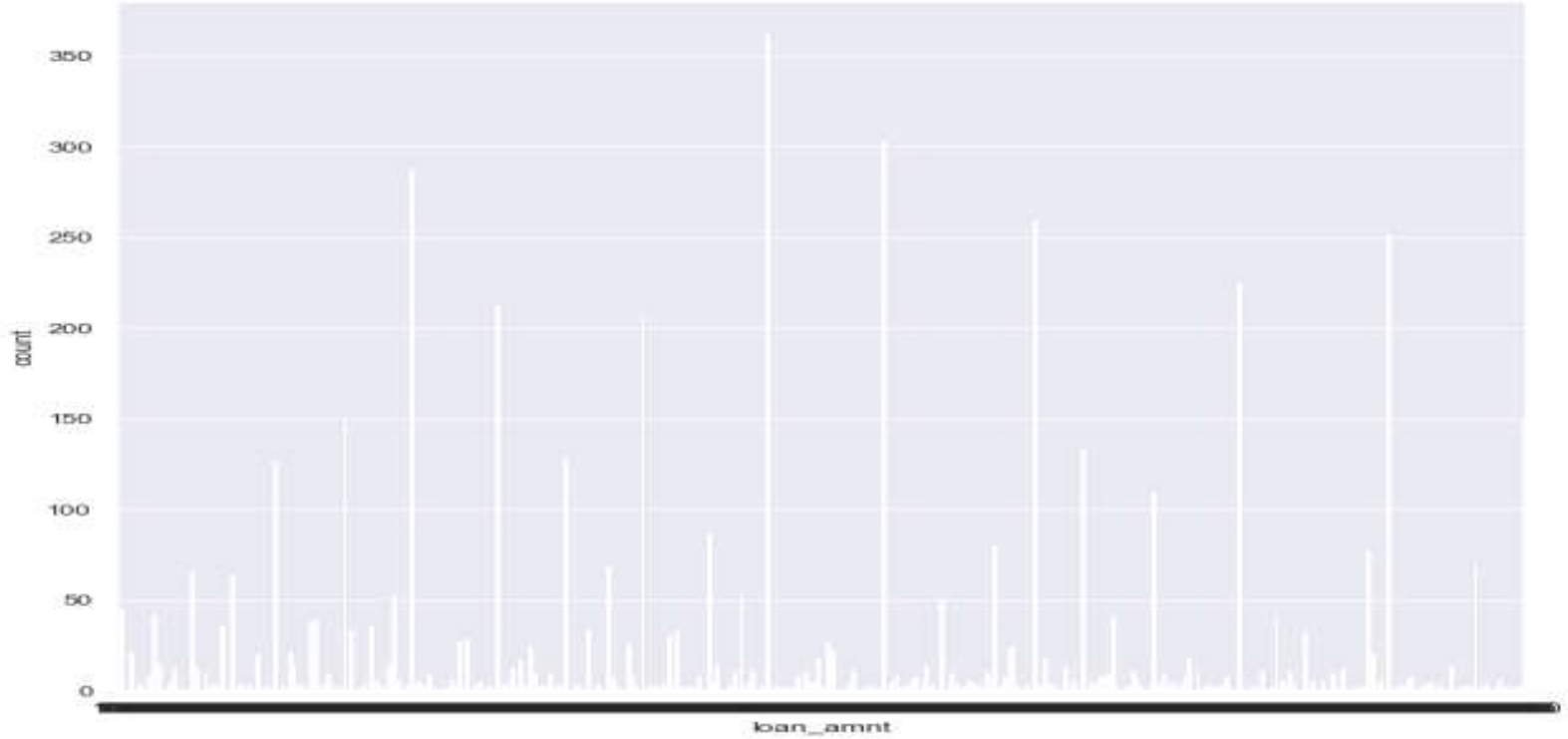
4. Count plot of emp_length among the charged off employees. We can see 10+ years are maximum defaulted.



Data Analysis – Bivariate Analysis

upGrad

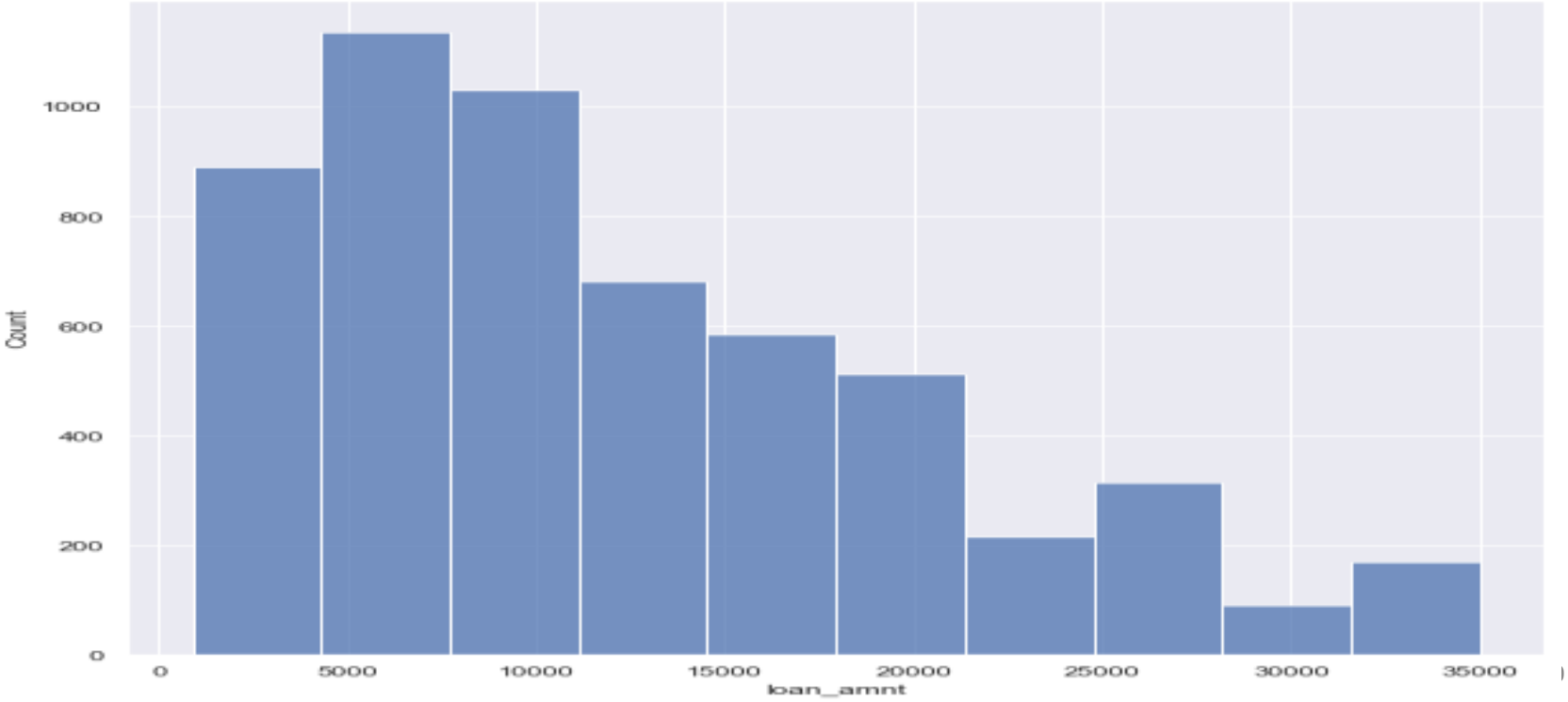
5. Count plot of loan_amount among the charged off employees.



Data Analysis – Segmented Analysis

upGrad

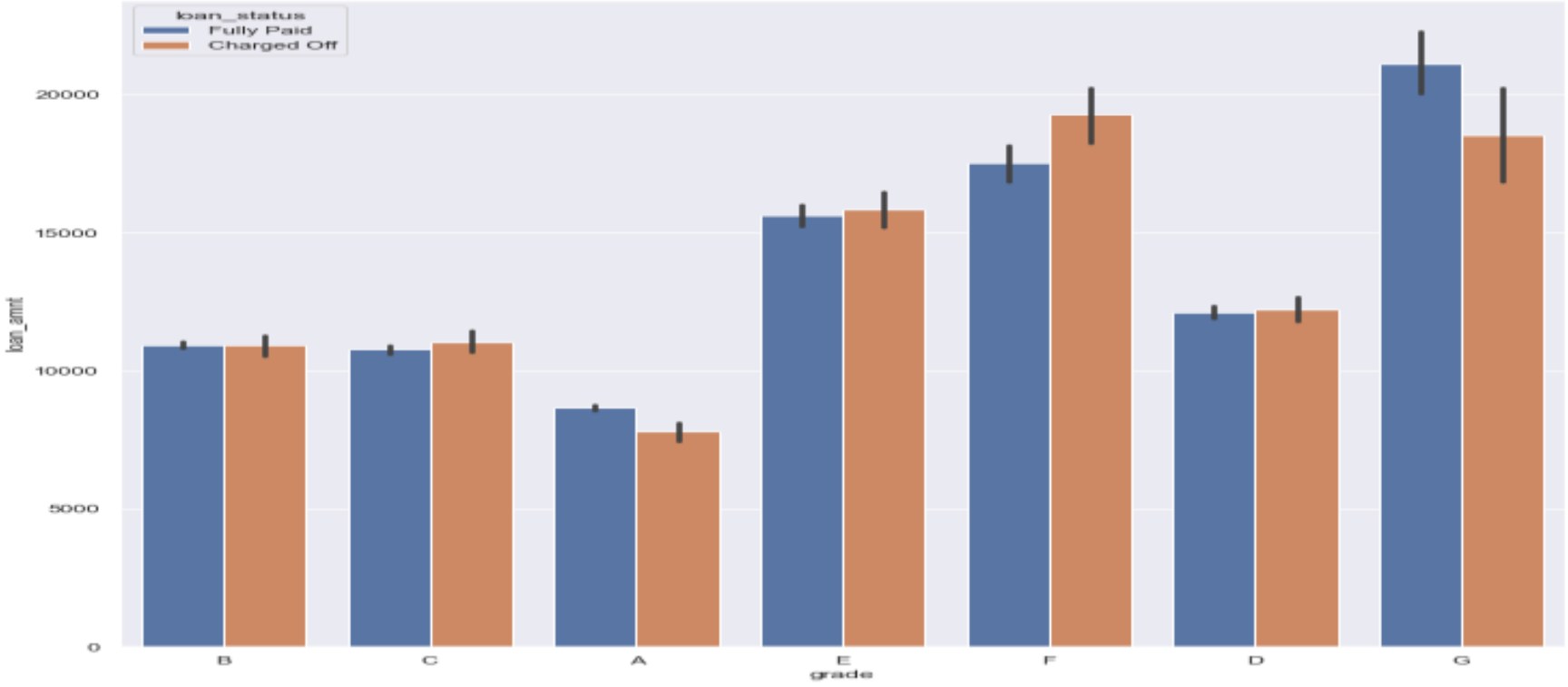
The above plot was not very continuous and meaningful. So after binning them we can see that the employees with lesser loan_amount have higher defaults.



Data Analysis – Bivariate Analysis

upGrad

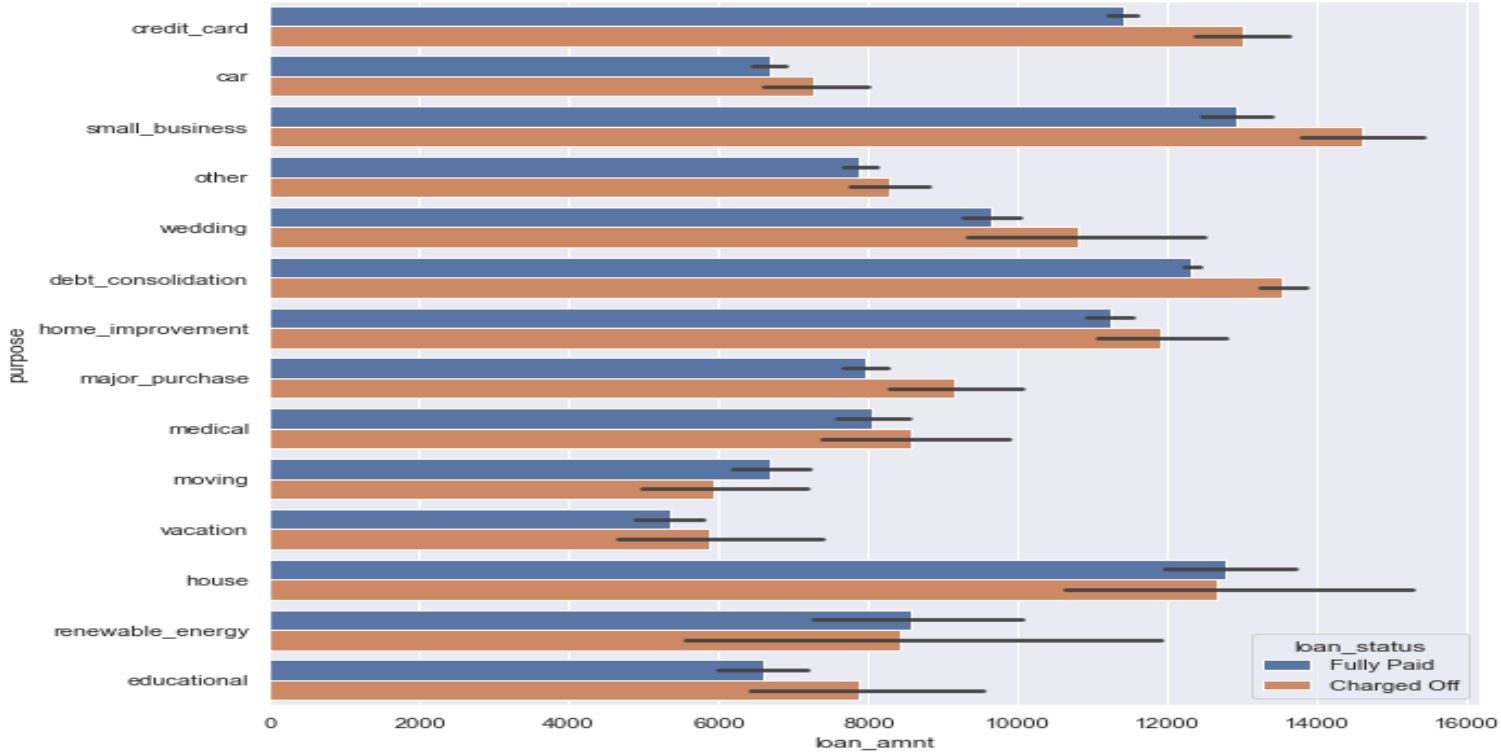
6. BarPlot of grade with loan status. As we can see employees with F grade are defaulting more and G are defaulting less.



Data Analysis – Bivariate Analysis

upGrad

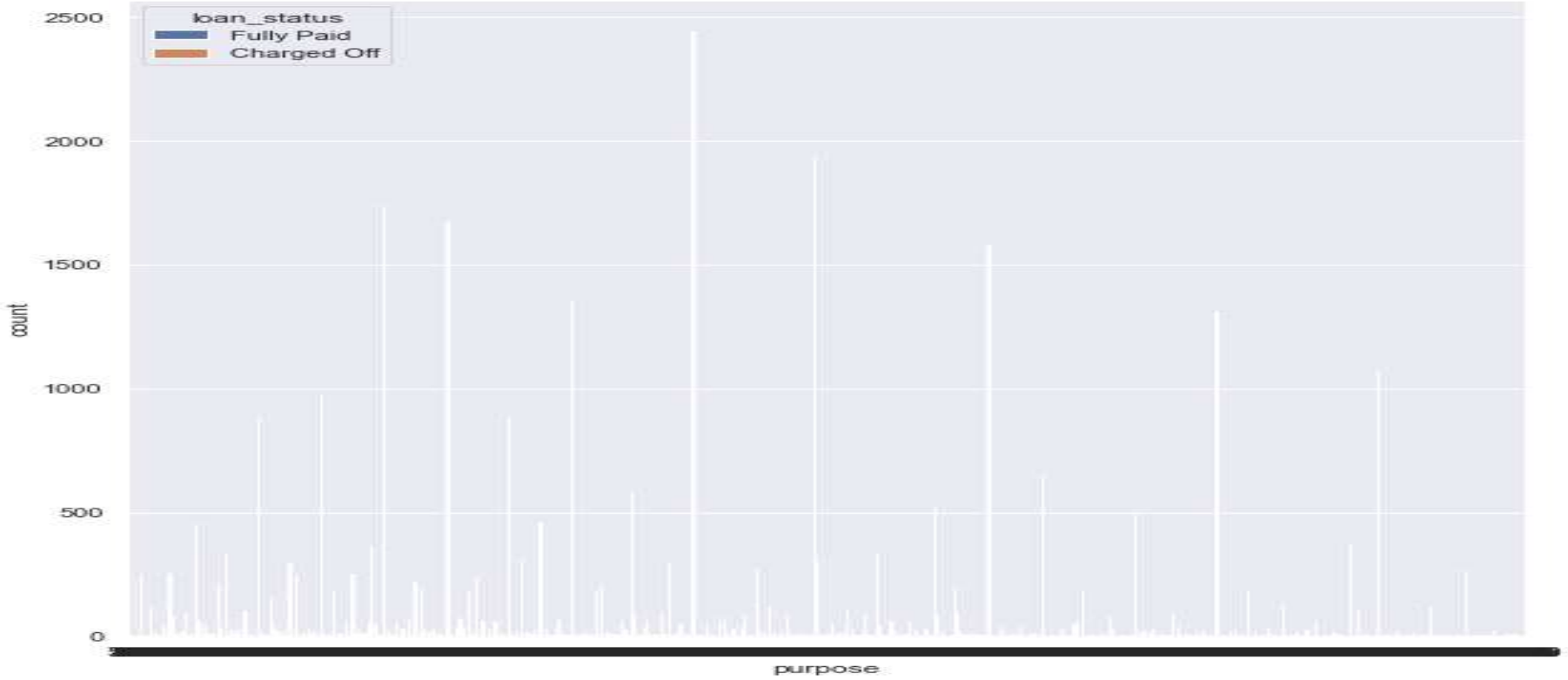
7. BarPlot of loan purpose with loan status. Small Business group has defaulted more.



Data Analysis – Bivariate Analysis

upGrad

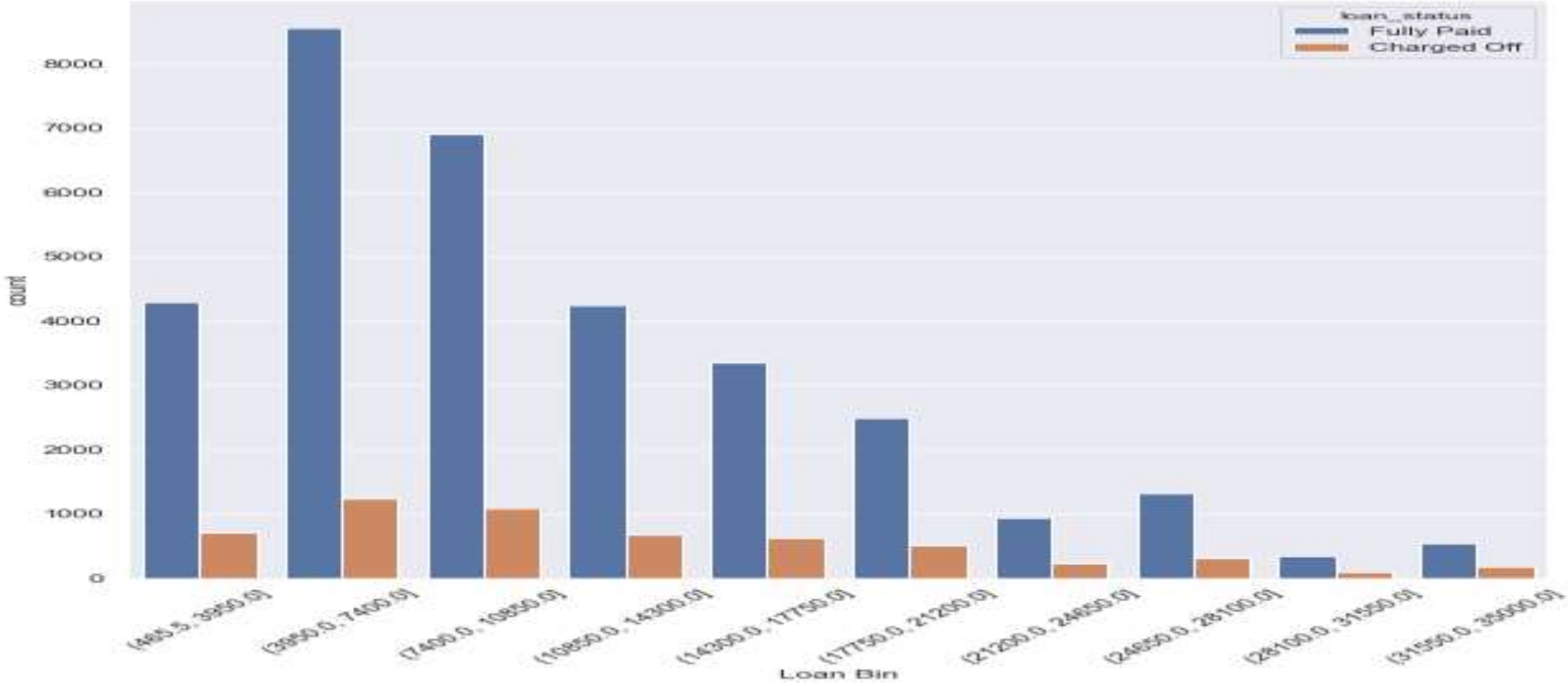
8. CountPlot of loan amount against loan status.



Data Analysis – Segmented Analysis

upGrad

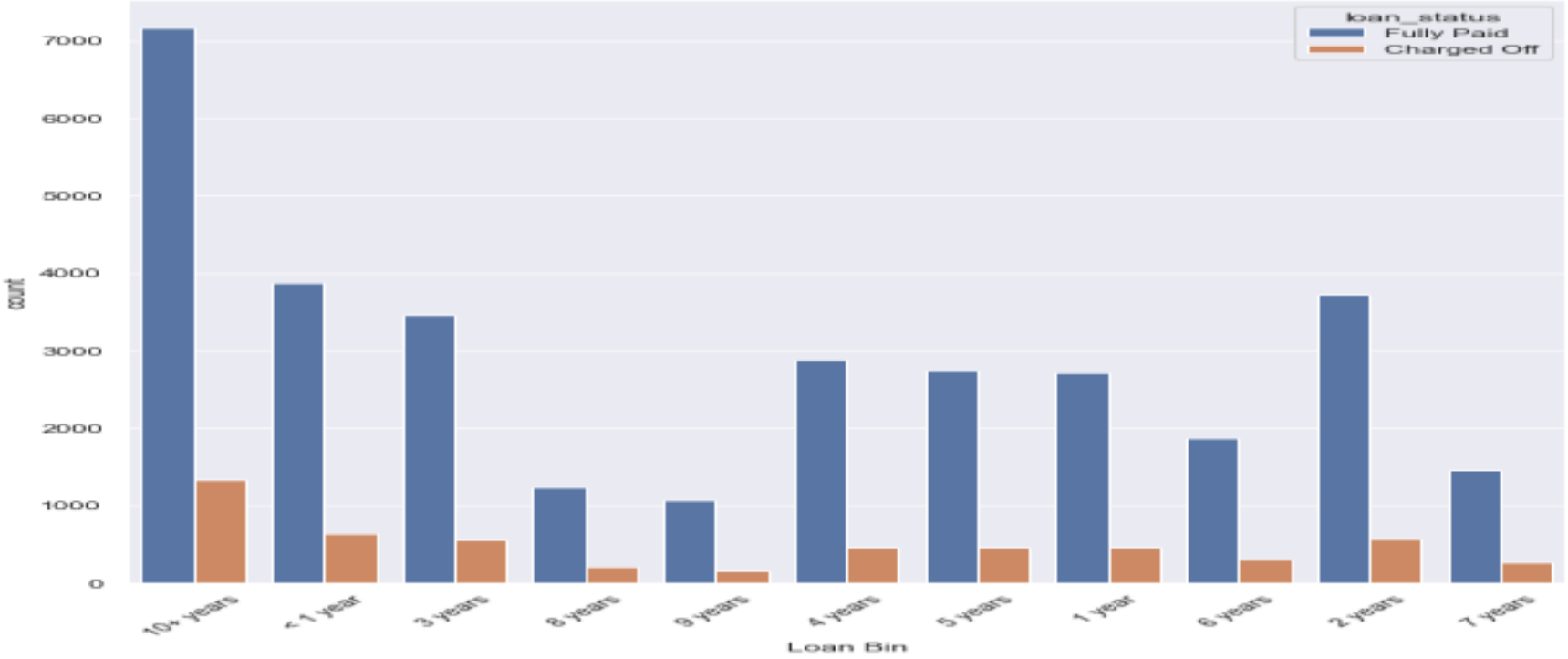
The previous plot was not continuous. After binning, we can see maximum % defaults are in the range 31000 – 35000 where maximum defaults are in the range 3950-7400.



Data Analysis – Bivariate Analysis

upGrad

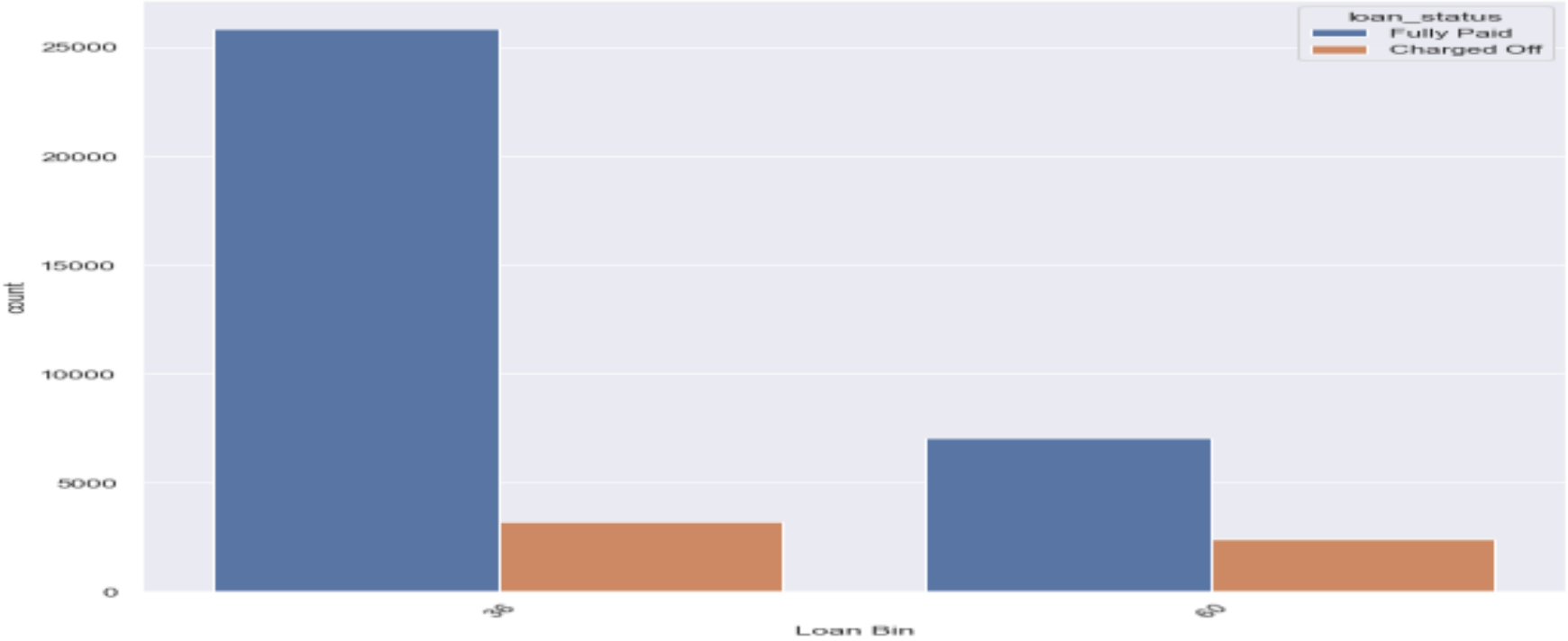
9. Count plot of employment length against loan status. We can see the default % is less in 10+ years.



Data Analysis – Bivariate Analysis

upGrad

10. Count plot term_val vs loan_status. Its clearly visible that loan with 60 months term has higher default %.



Lending Club: Recommendations

upGrad

- We can see the atleast 15% of the total number of loans are defaulted.
- Employment term with 10+ years are maximum defaulted. So its better to reduce the number of loan approvals within such group.
- Employees with lesser loan_amount have higher defaults. So lesser loan amount needs to be restricted further.
- Employee with grade as C,D,E,F tend to default more . So its better to reduce the number of loan approvals within such group.
- Loan purpose like small business, credit card and debt_consolidation have defaulted more, so reduce the number of approvals there.
- Maximum % defaults are in the range 31000 – 35000 loan amount, so we need to be care -ful while approving loans there.
- Loan with 60 months term has higher default %. So its better to enforce more number of loans which has 36 months term.

upGrad

#LifeKoKaroLift

Thank You!