PROJECT ON LENDING CLUB CASE STUDY

SUBMITTED BY : SMIT SAGAR DASH

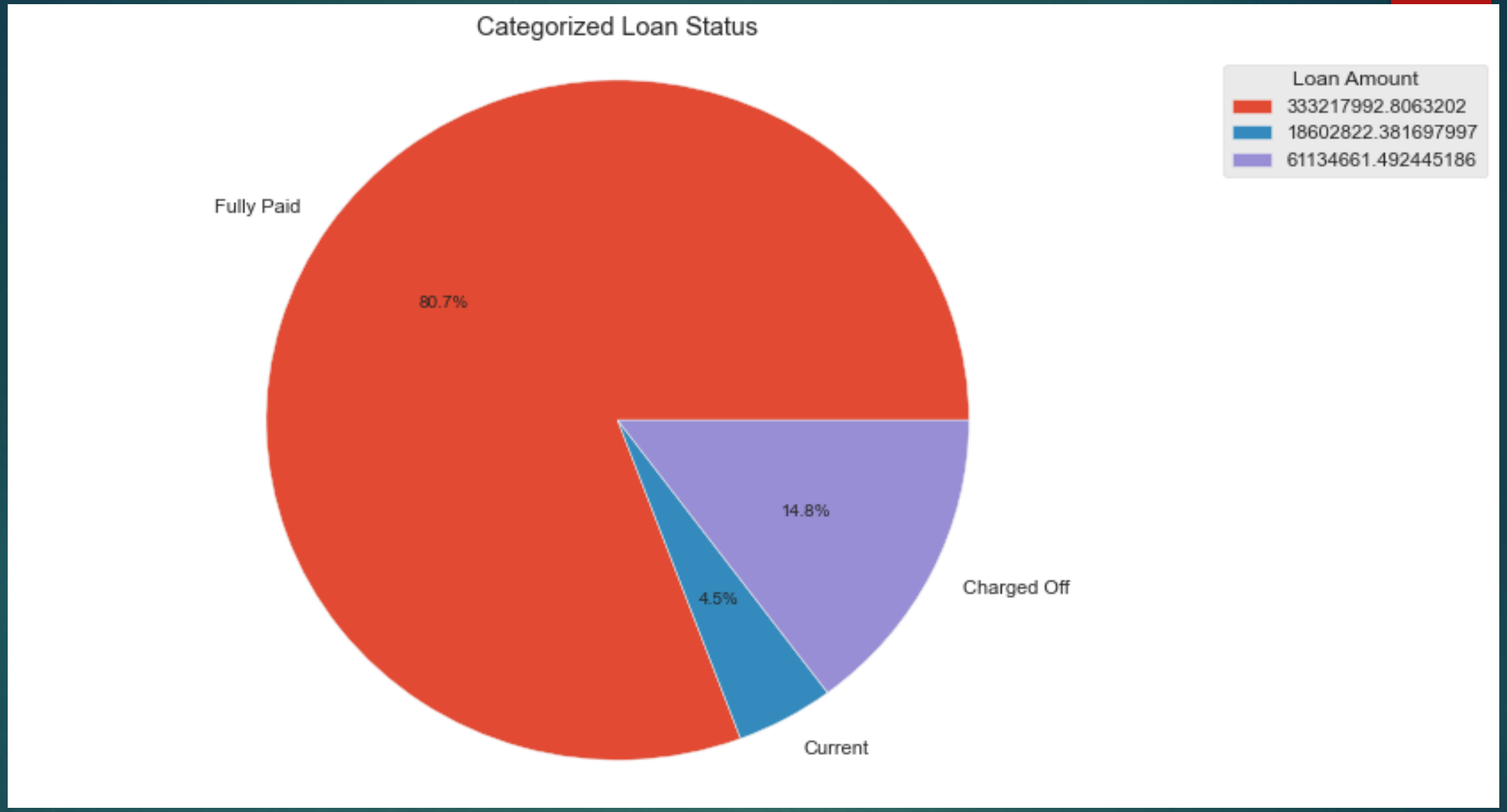
Description

In this presentation, various charts and plot observation is shown :

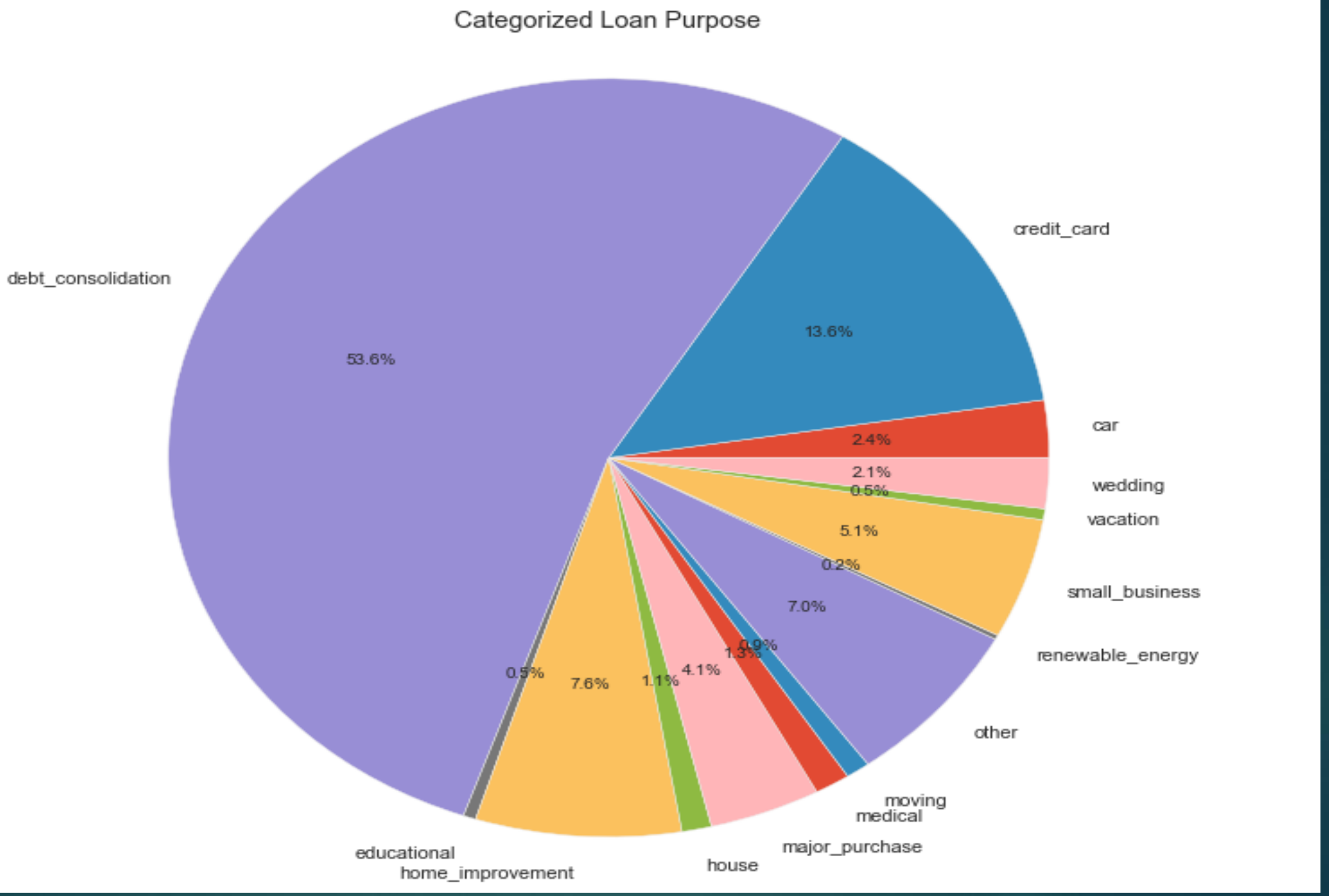
- The pie chart observation of loan_status data.
- Univariate, bivariate and multi-variate analysis on the basis of defaulters in loan_status.
- All analysis is done after cleaning unnecessary data from given dataset

Observations

Pie Charts



loan_status is categorised on the basis of Fully Paid, Current and Charged Off category with the help of pie-chart.

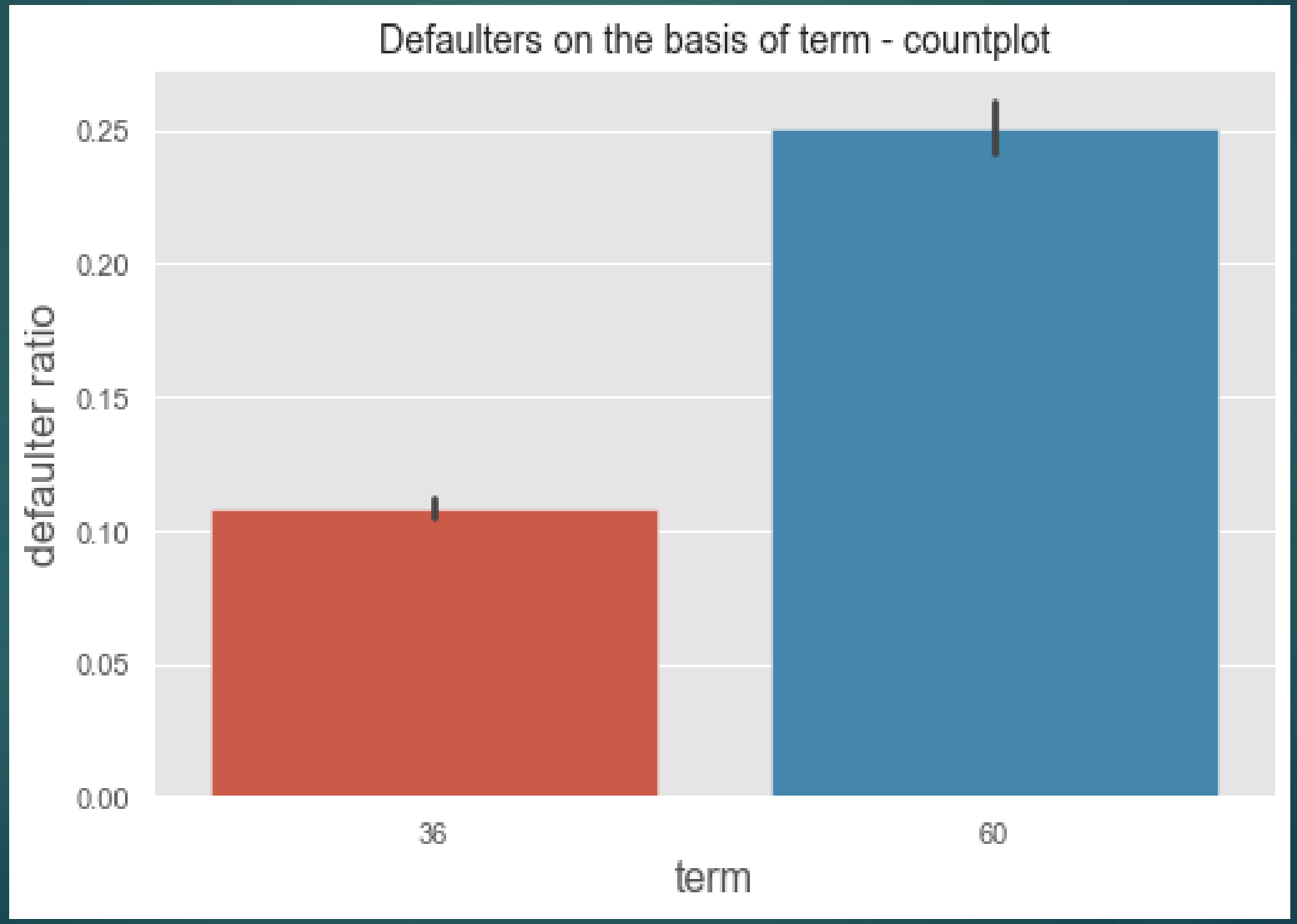


In this pie-chart, we can see various types purpose from the dataset has been categorised.

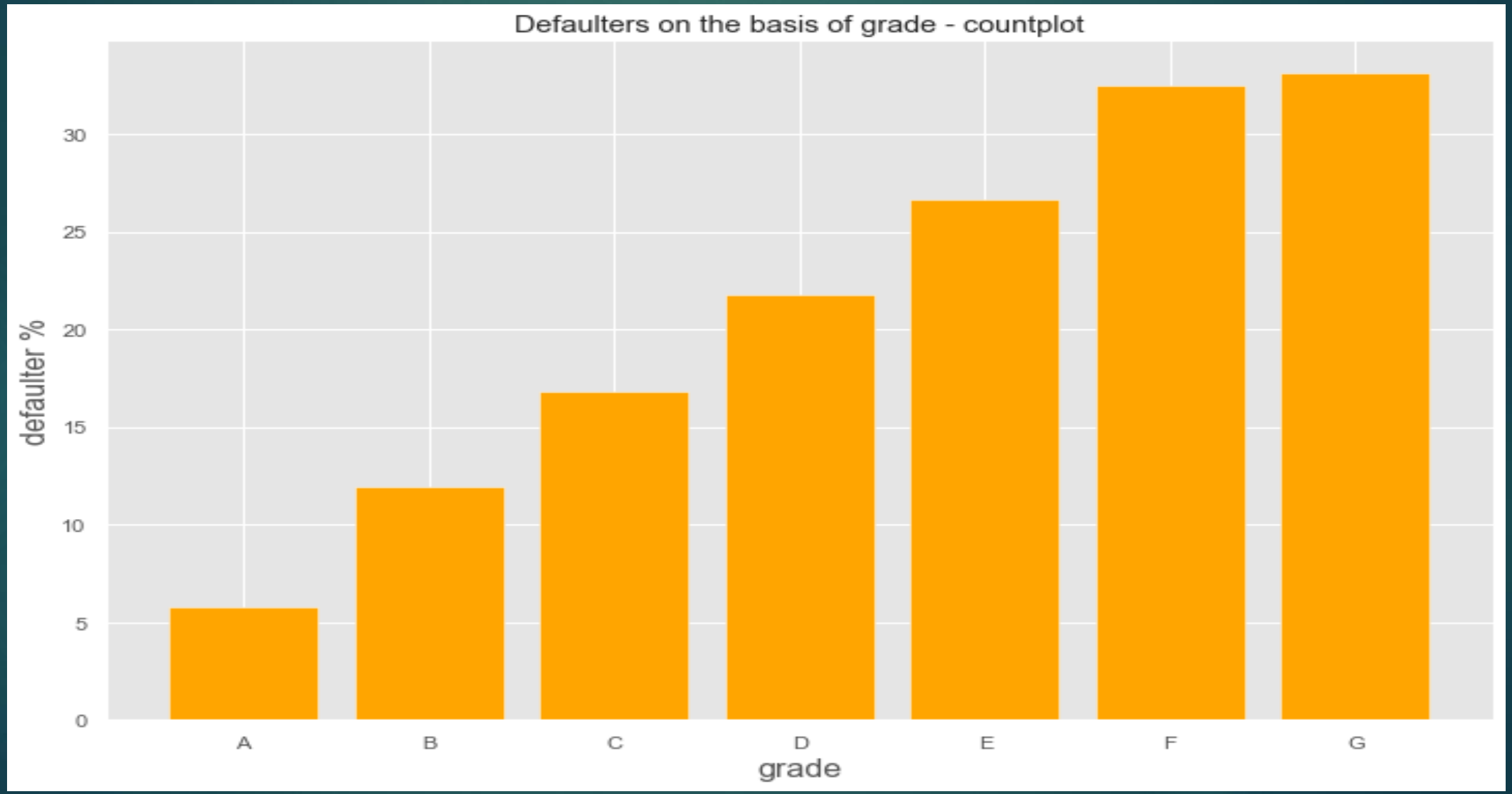
Uni-variate Analysis

(*using count-plot*)

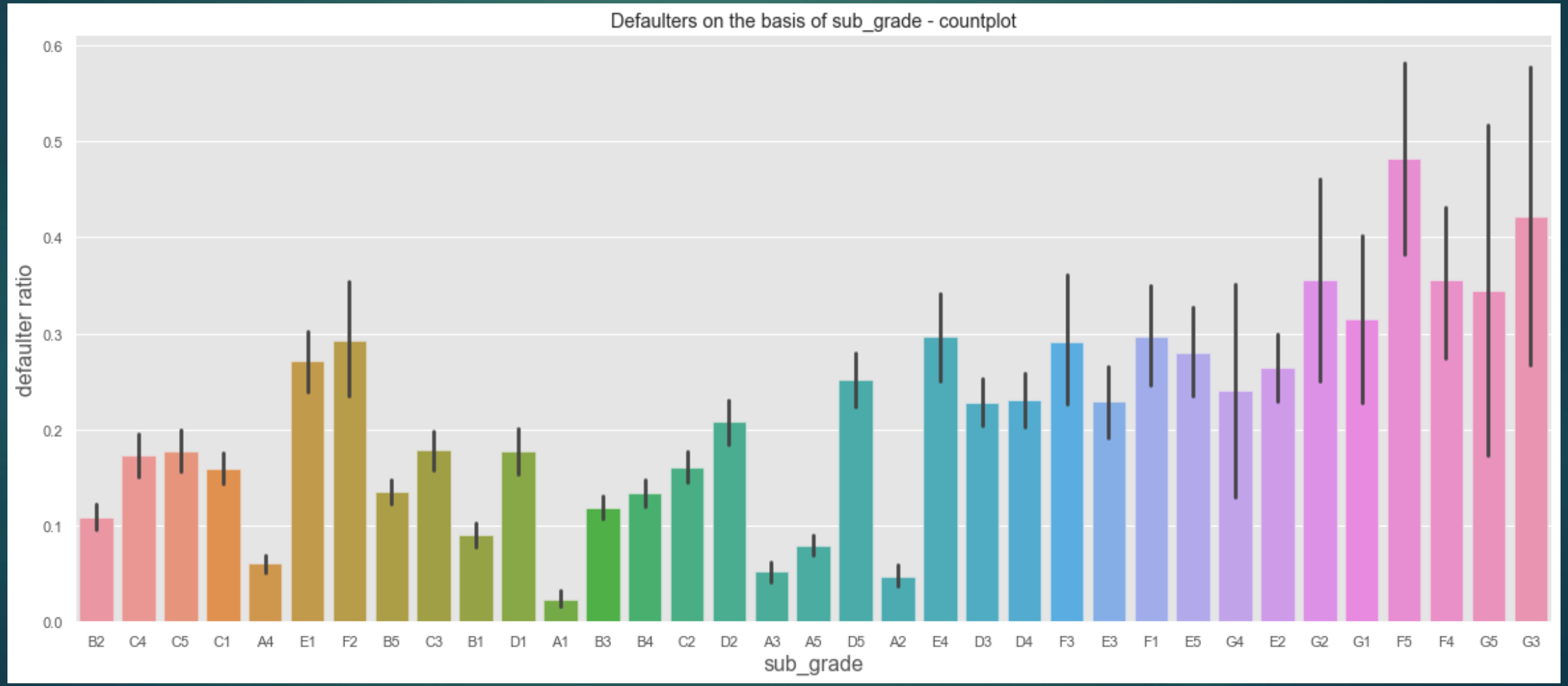
Term count-plot



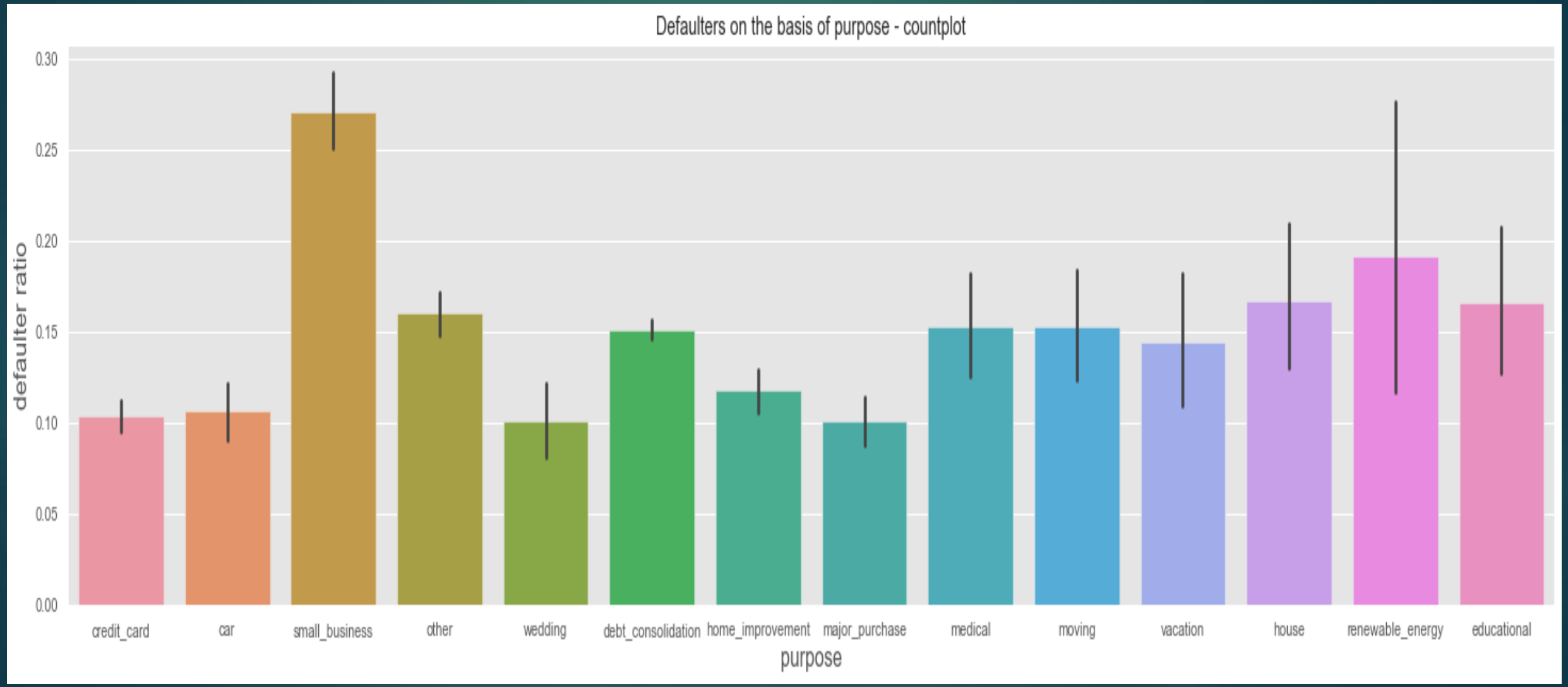
Grade count-plot



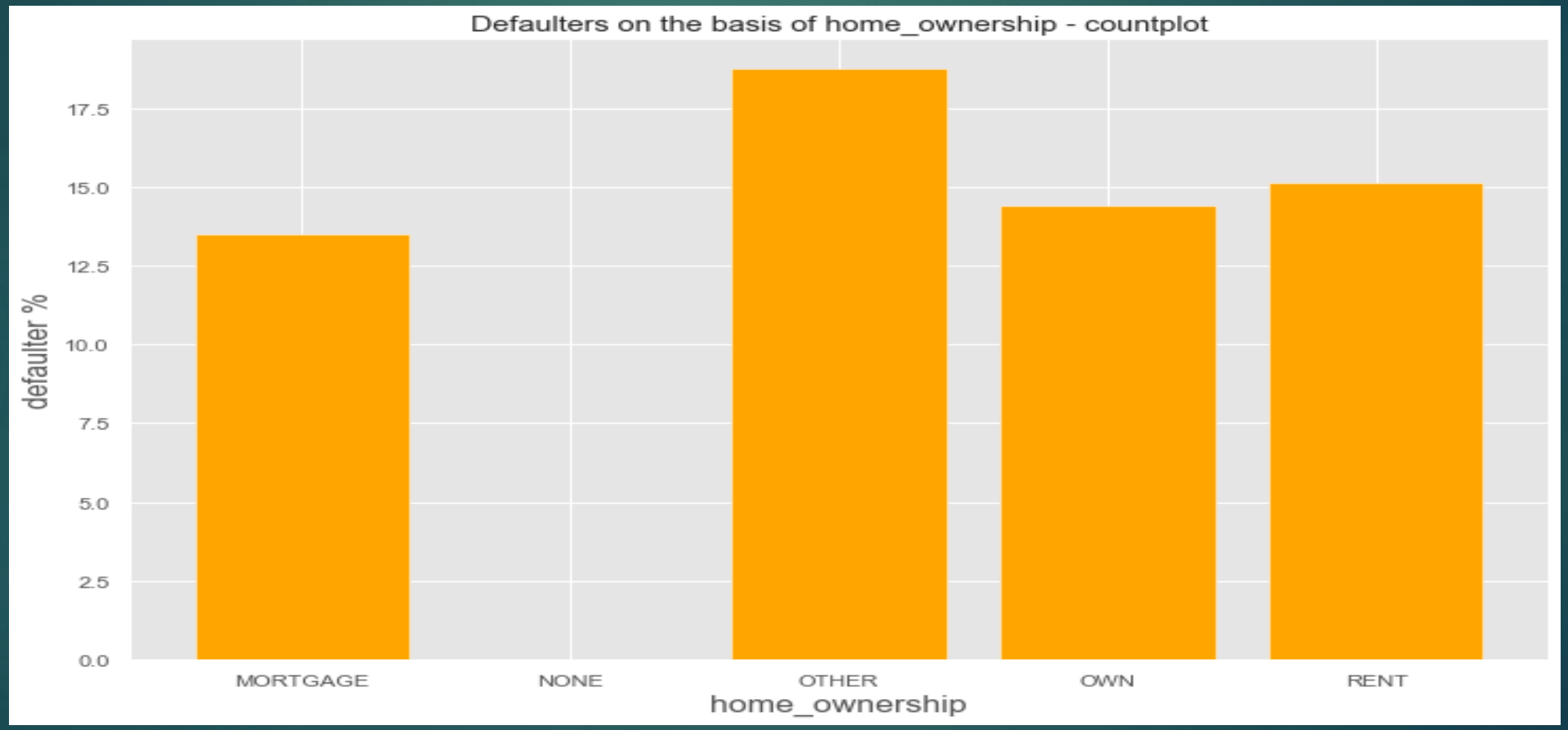
Sub-grade count-plot



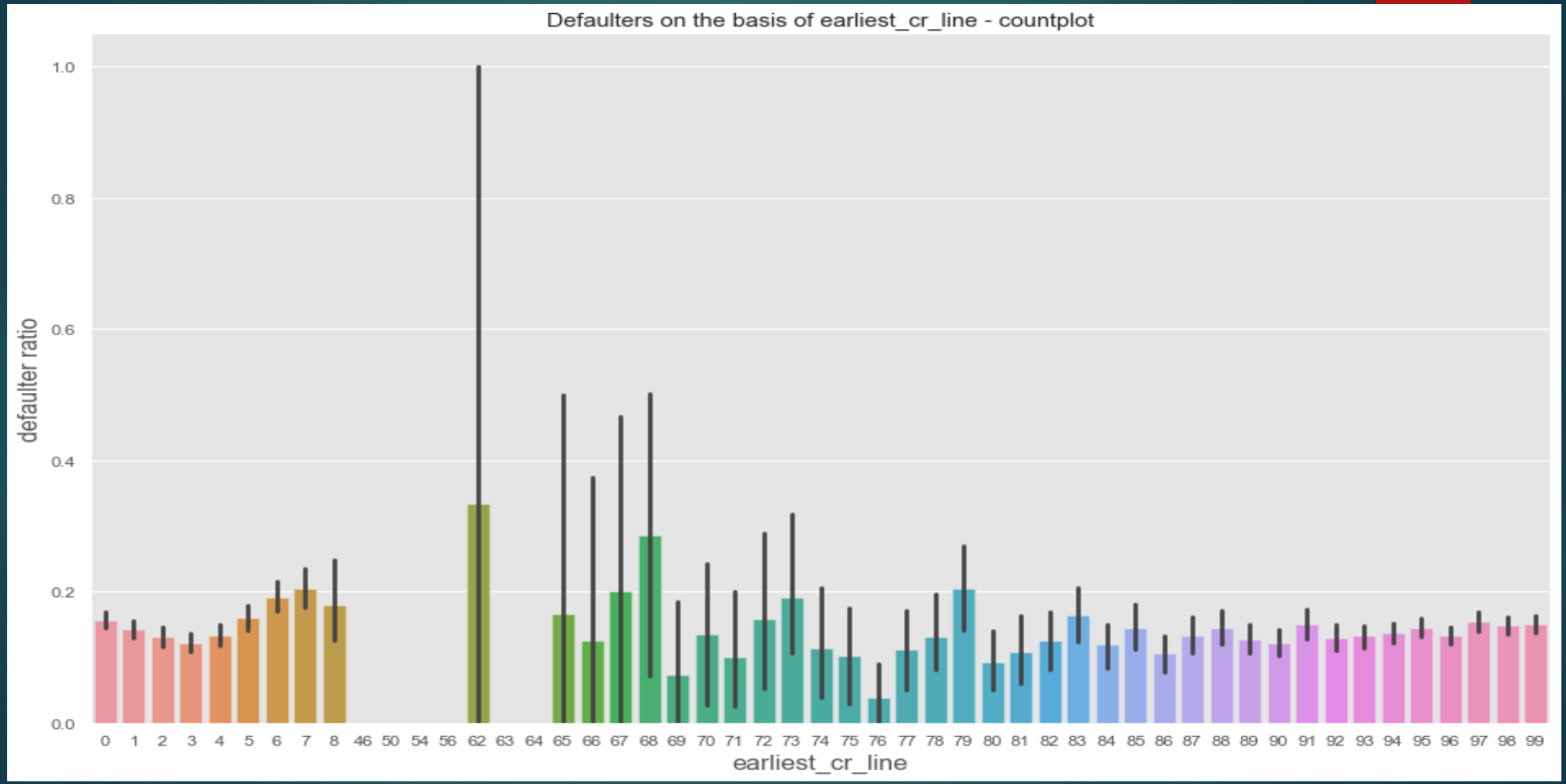
Purpose count-plot



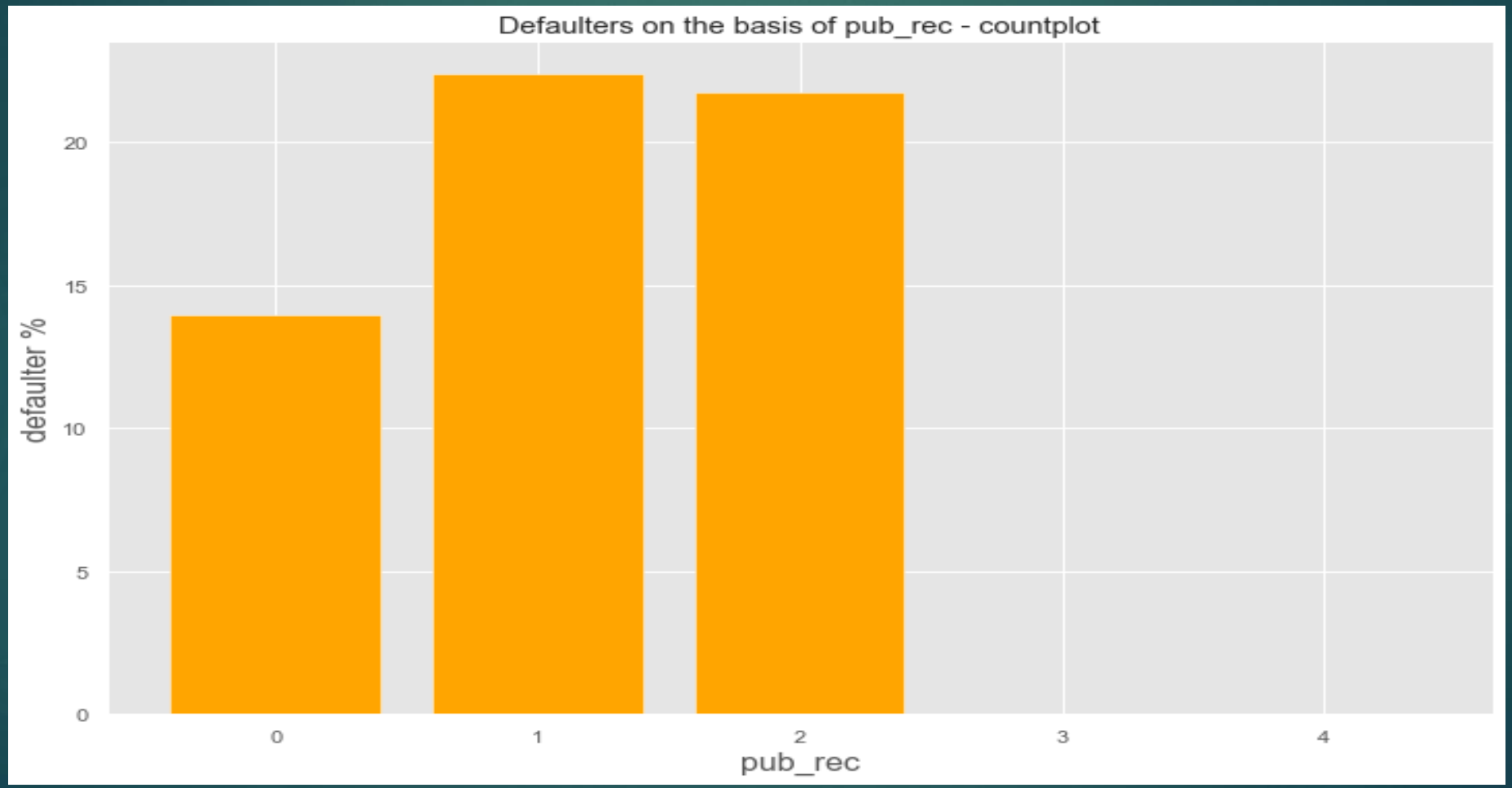
Home ownership count-plot



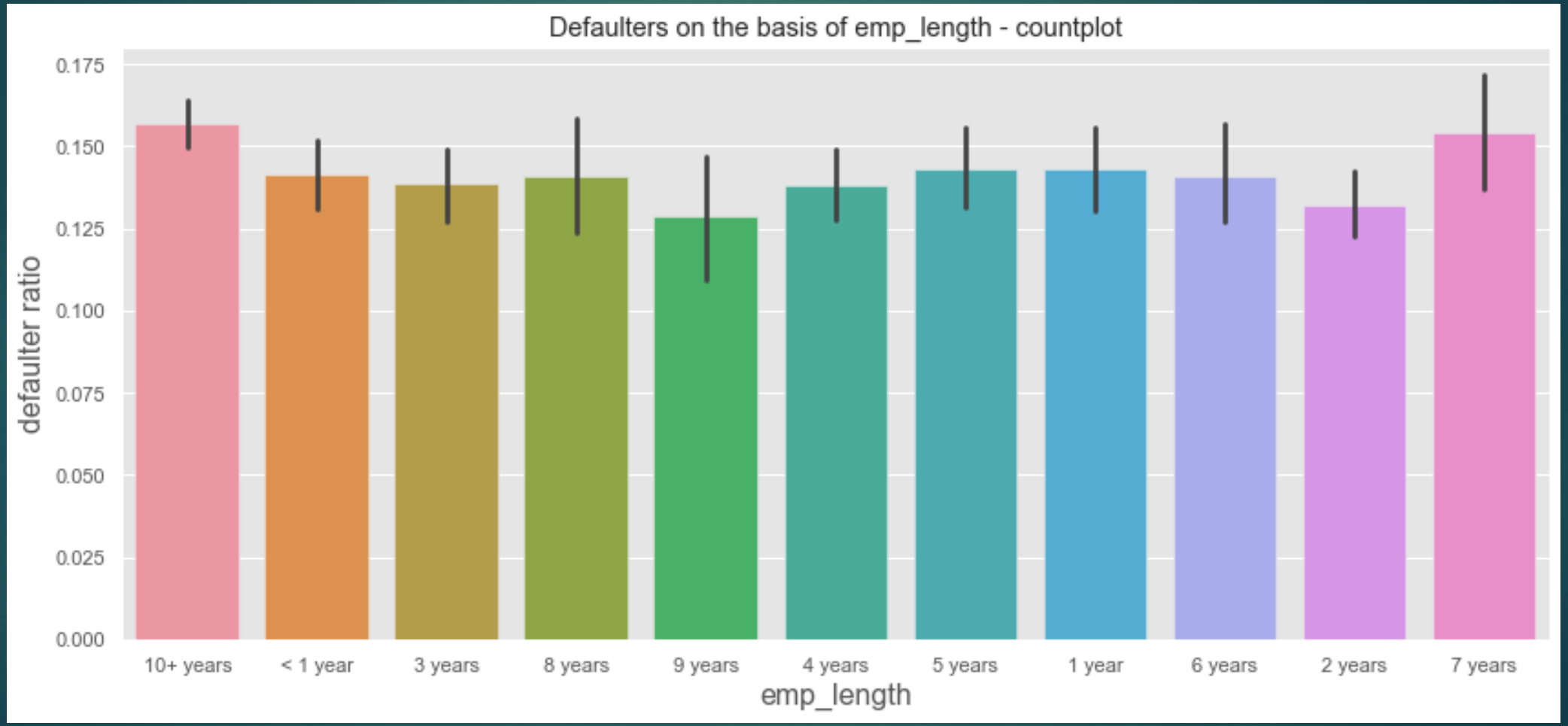
Earliest credit line count-plot



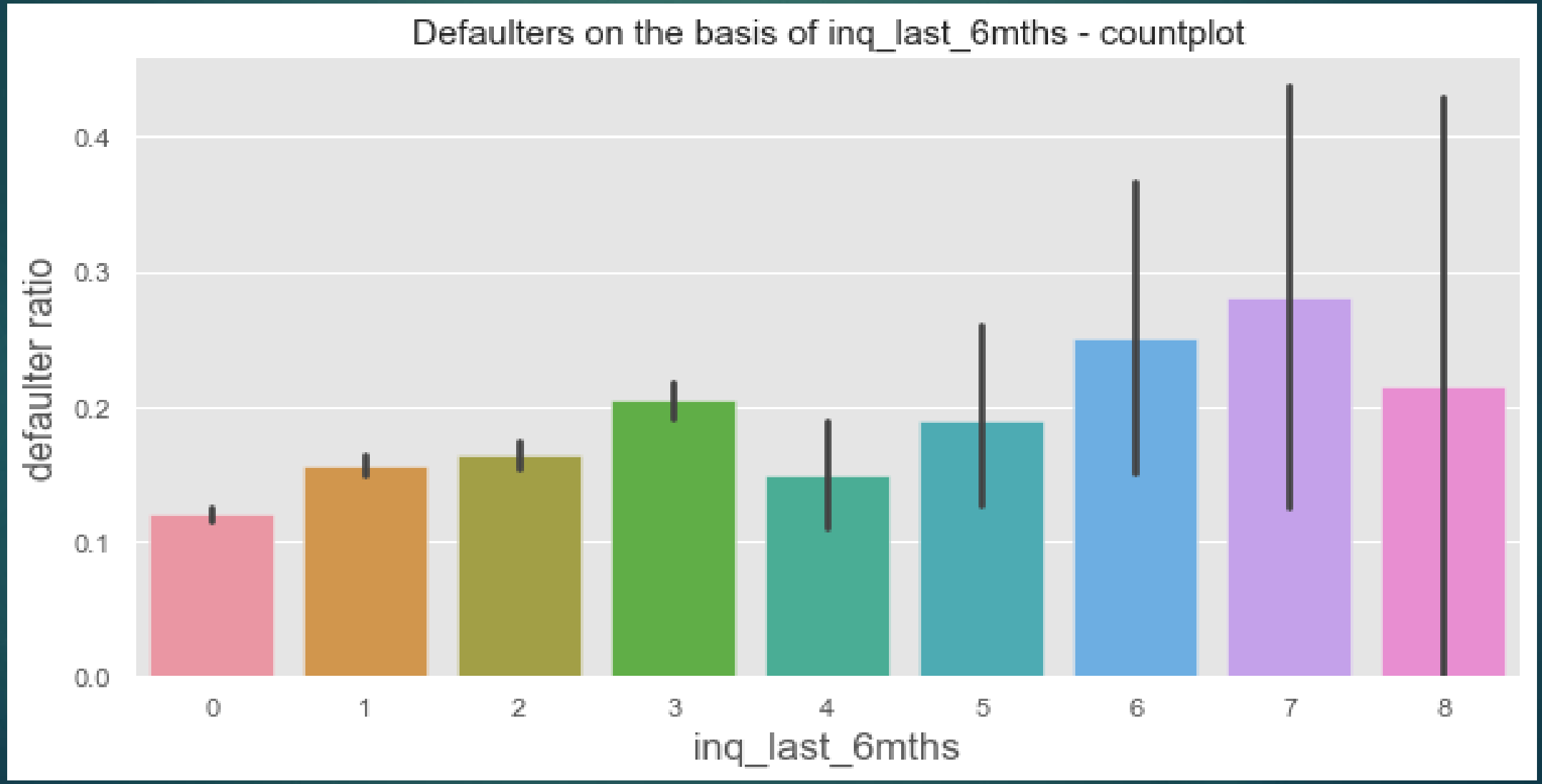
Public records count-plot



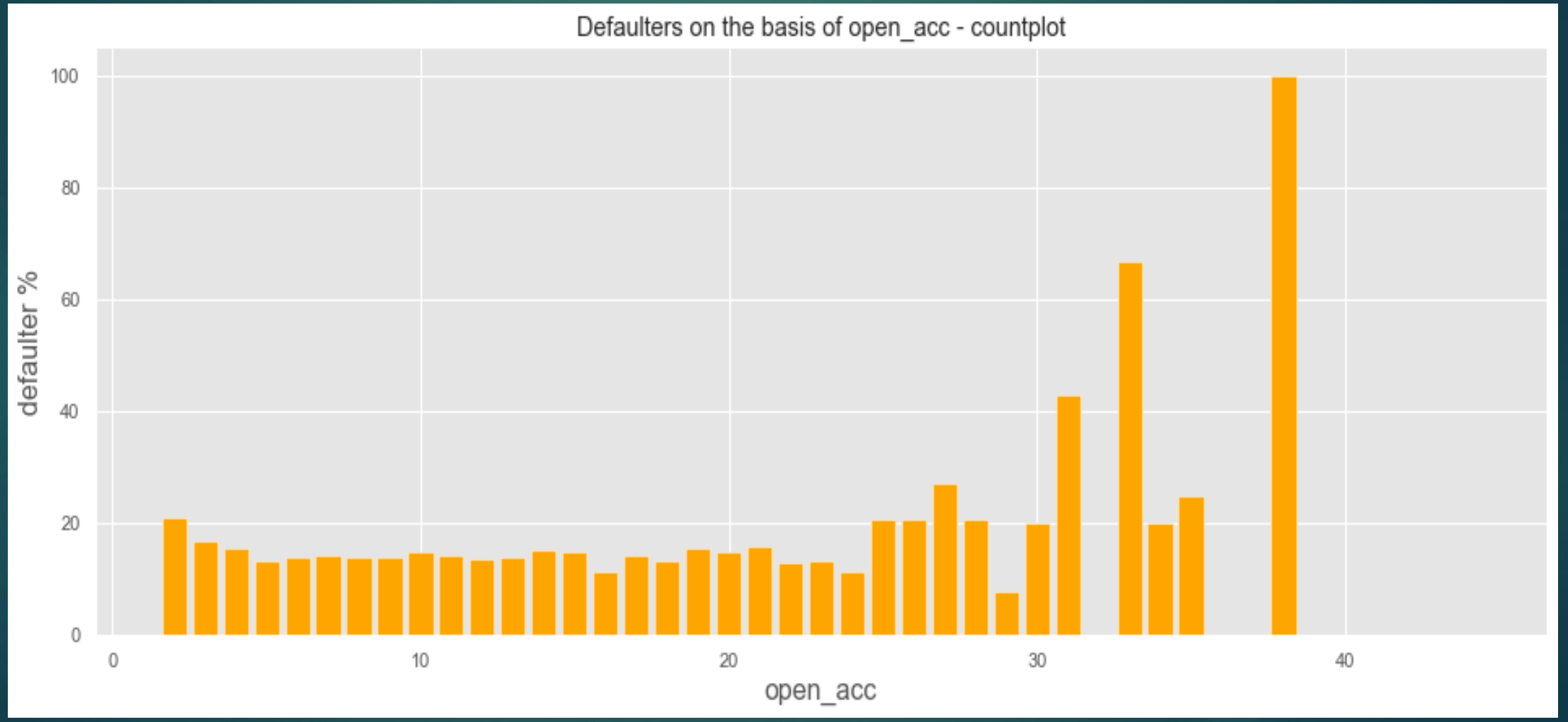
Employment length count-plot



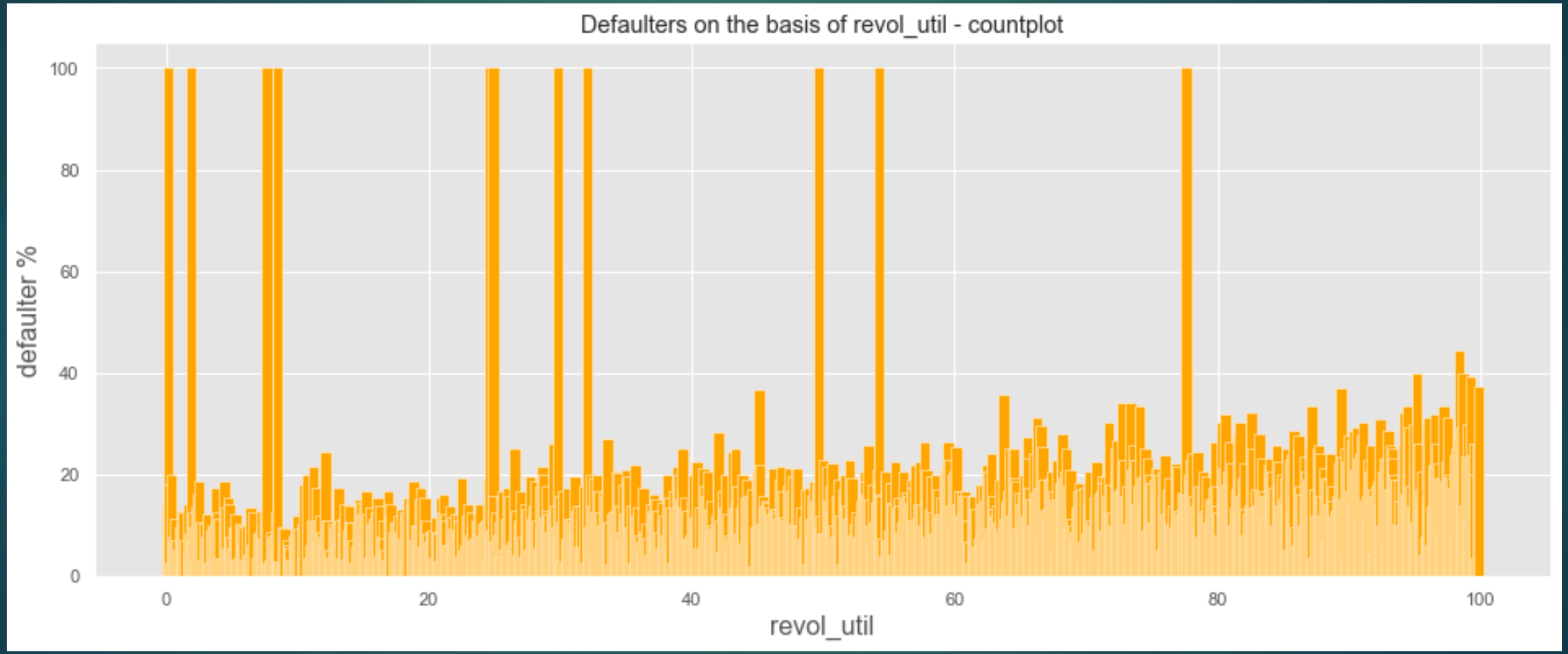
Count-plot of Inquires in last 6 months



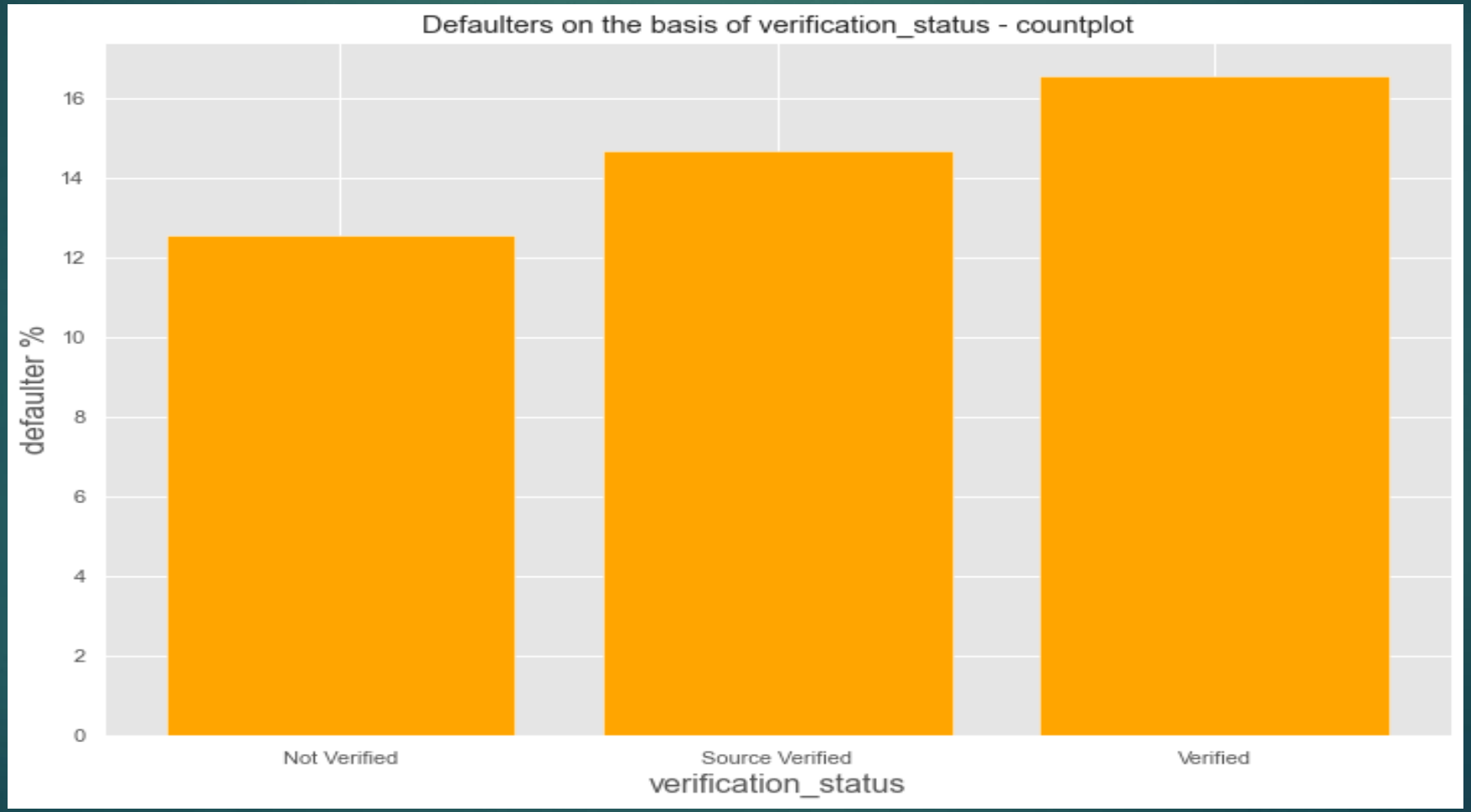
Open account count-plot



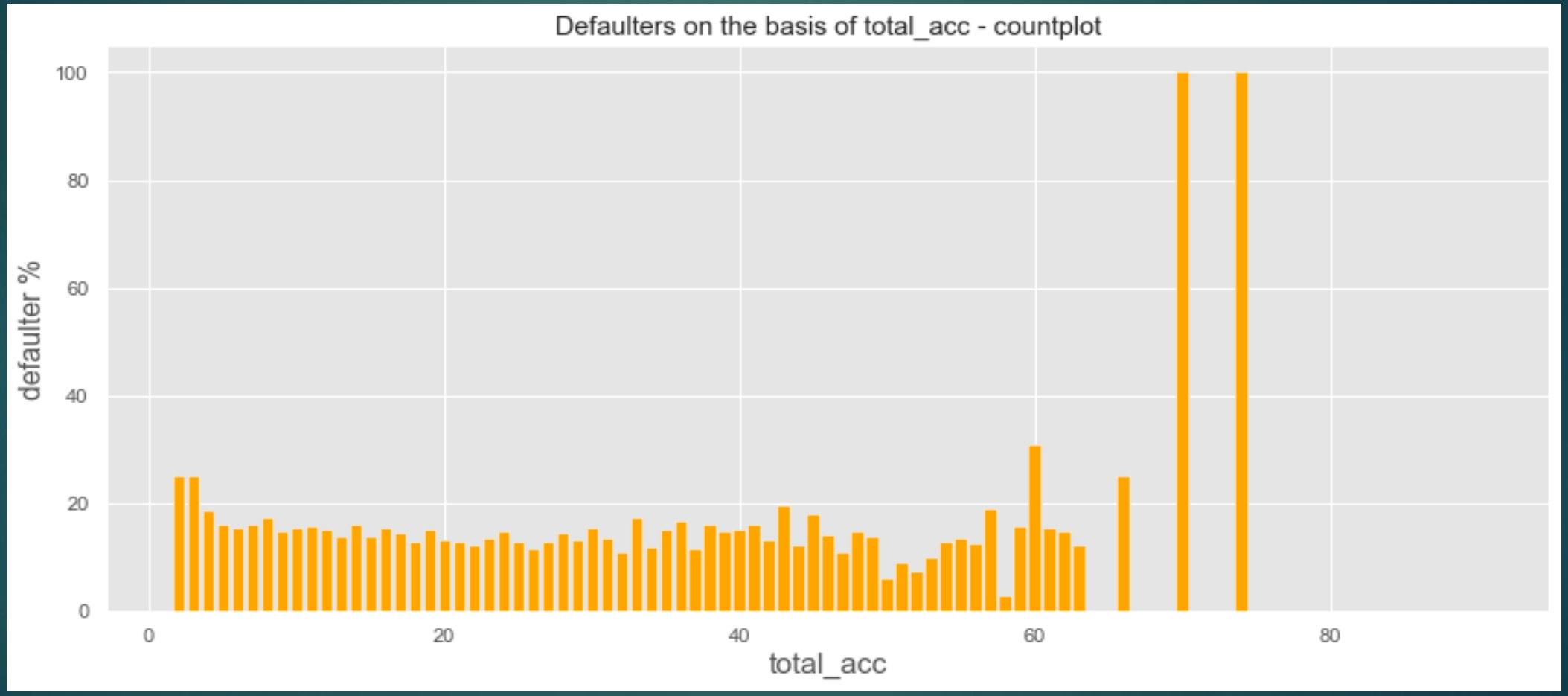
Count-plot of revolving line utilization



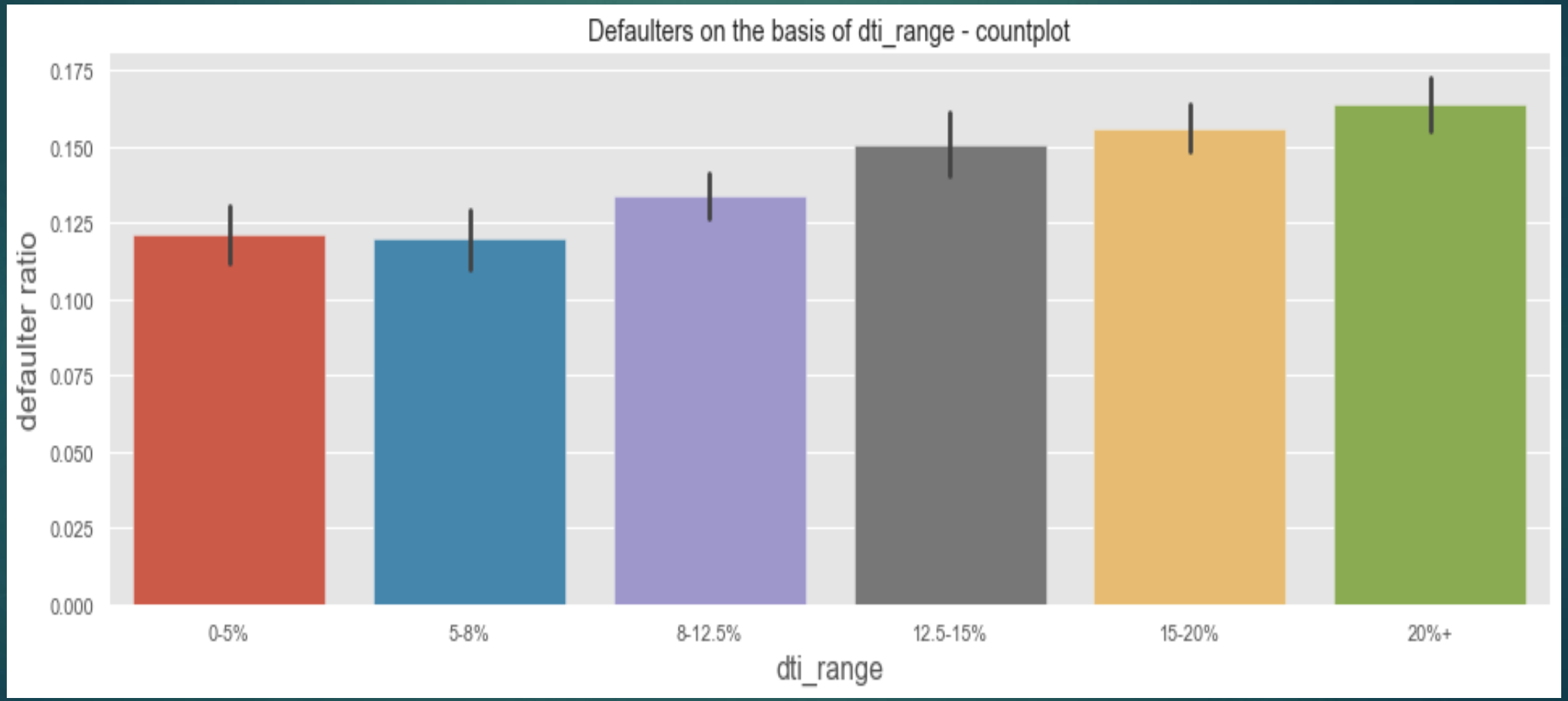
Verification status count-plot



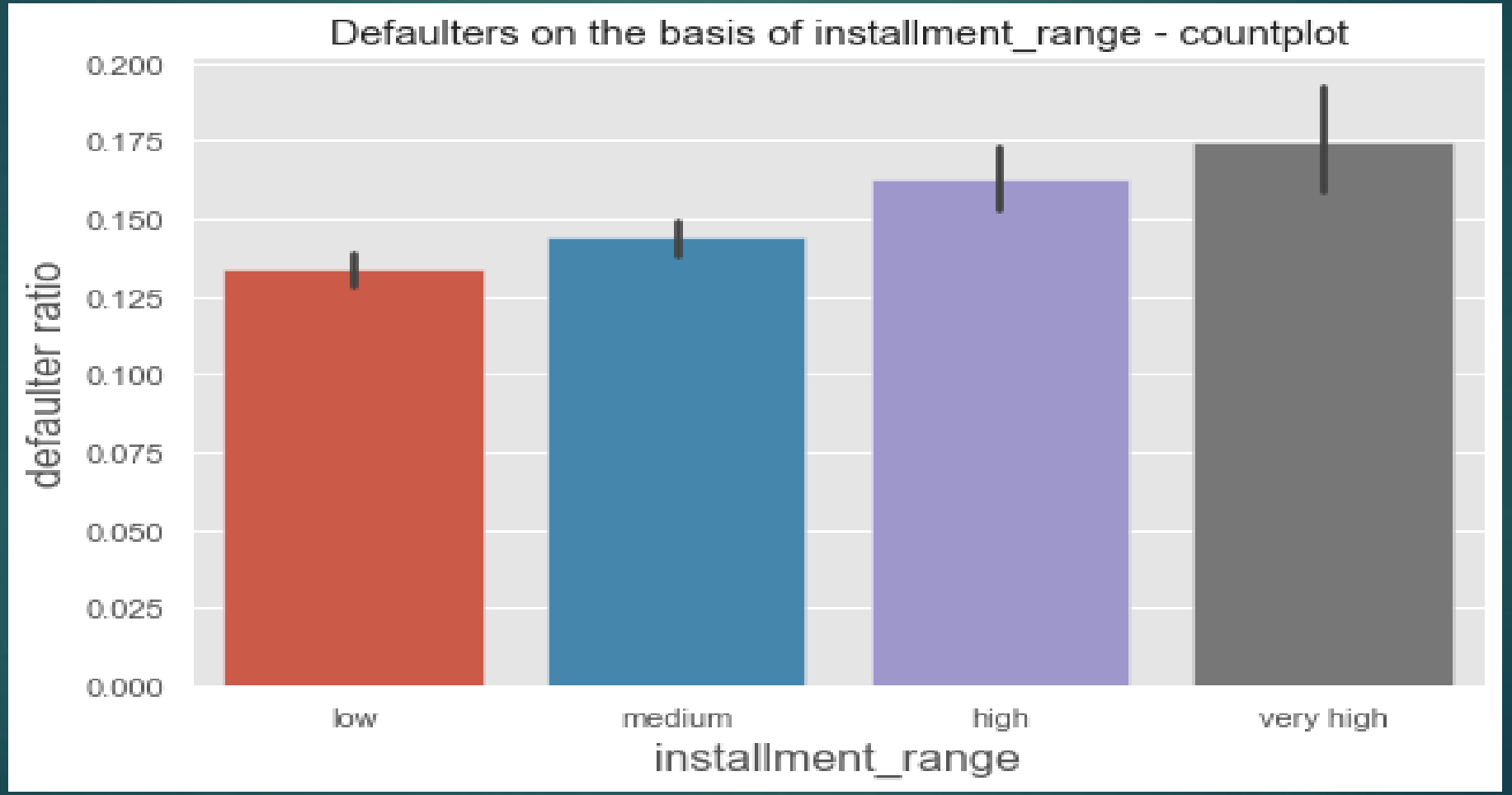
Count-plot for total credit lines



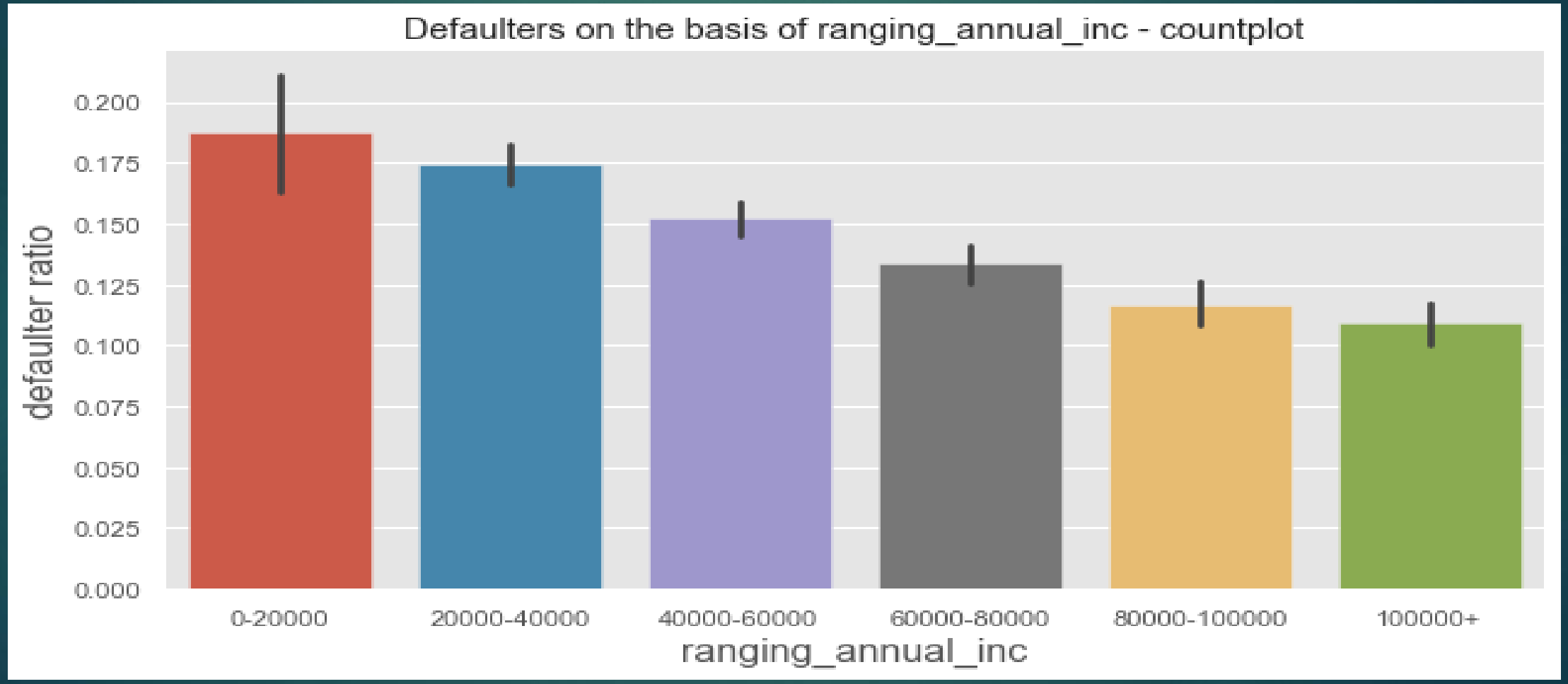
Count-plot for *dti* in dataset



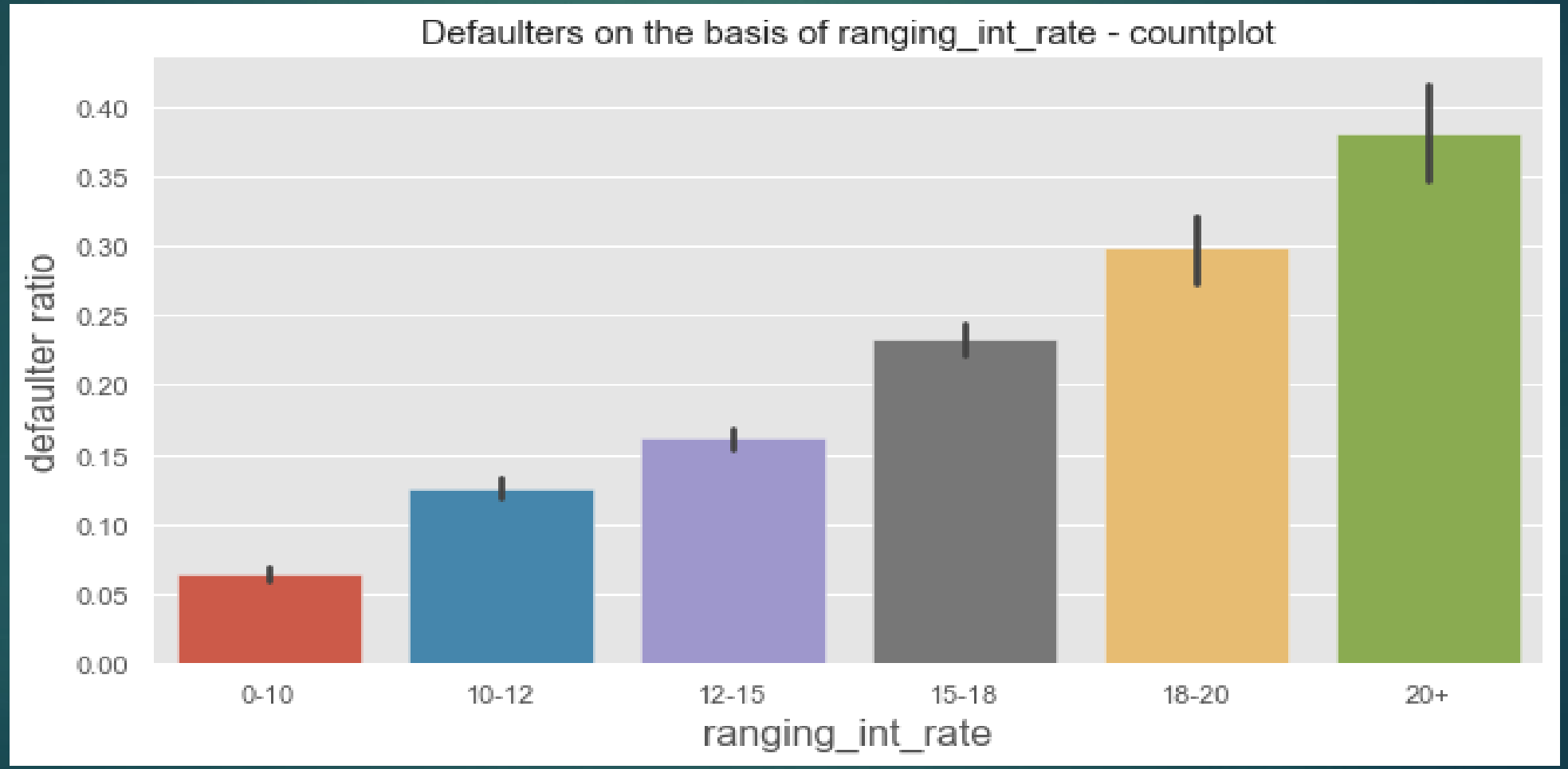
Installment count-plot



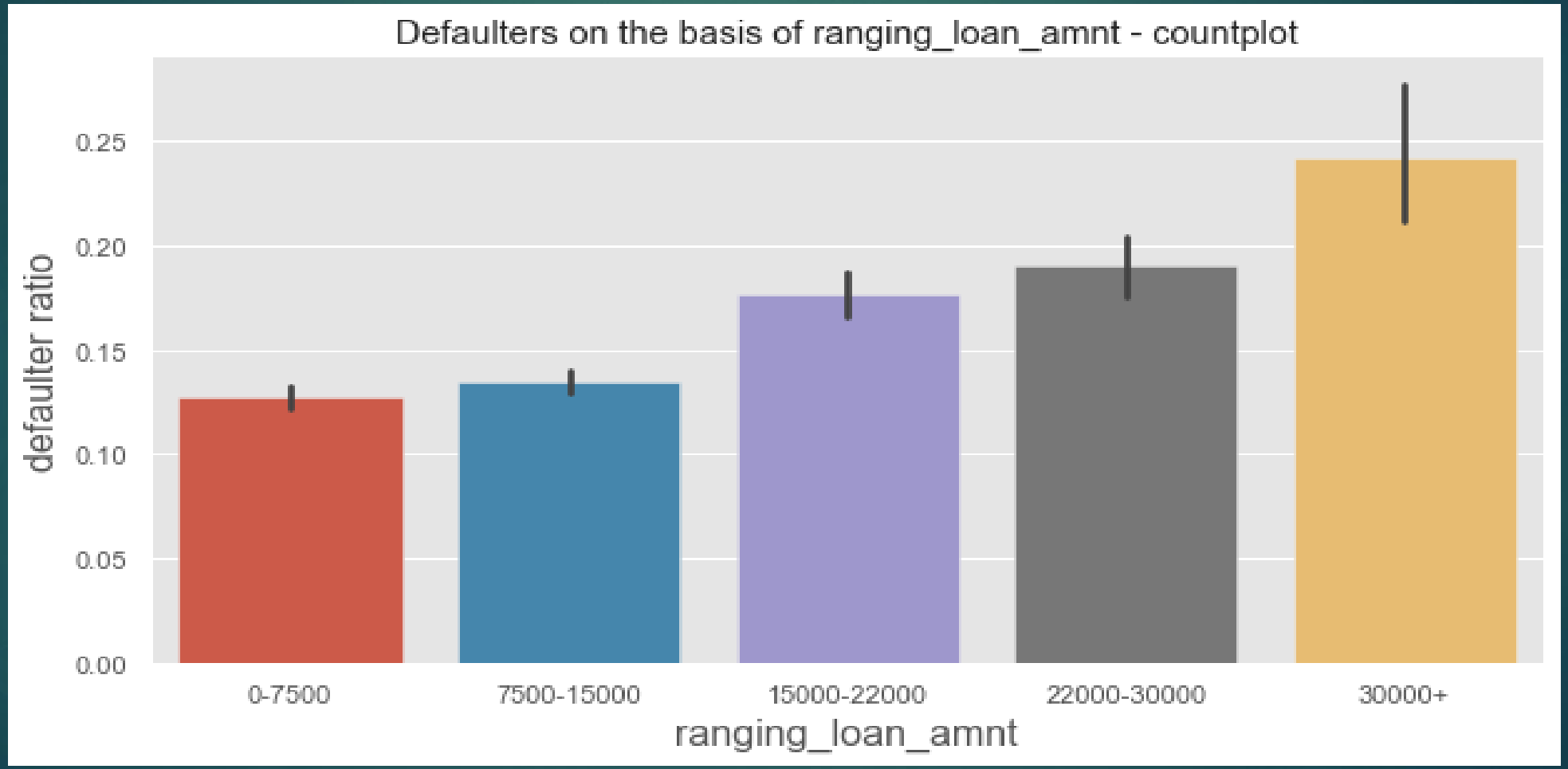
Count-plot for annual income



Interest rate count-plot



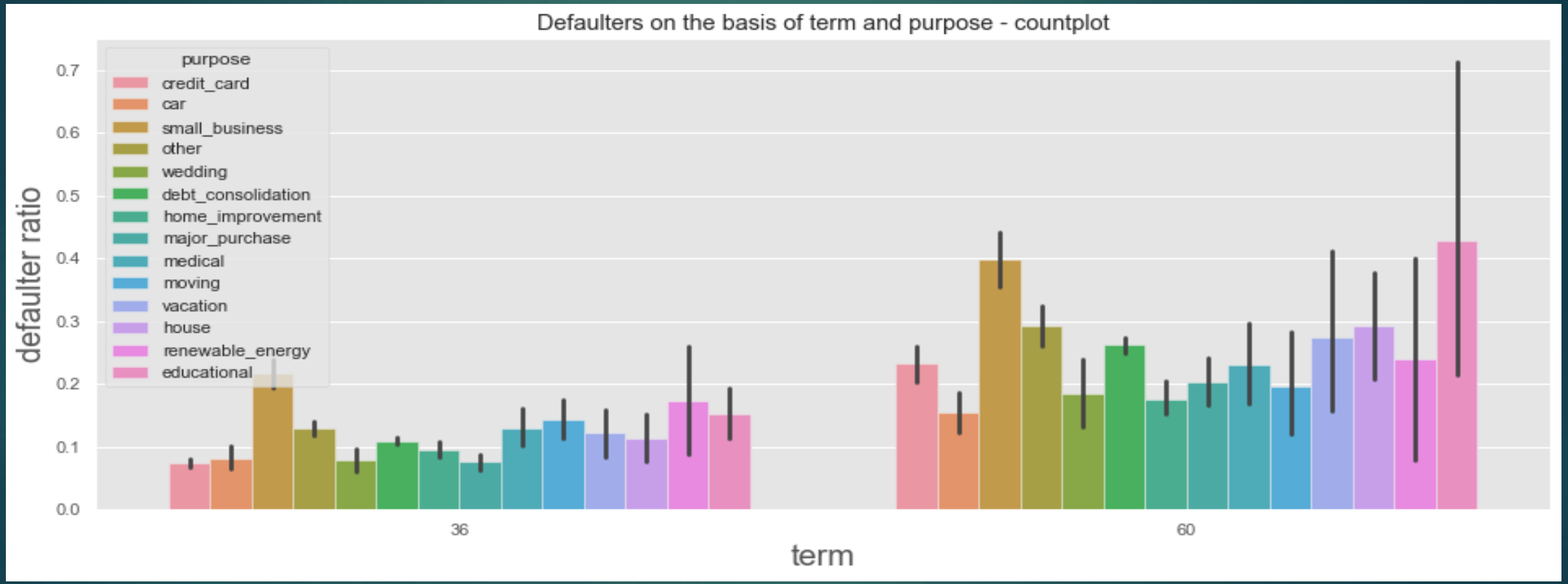
Count-plot for loan amount



Bi-variate Analysis

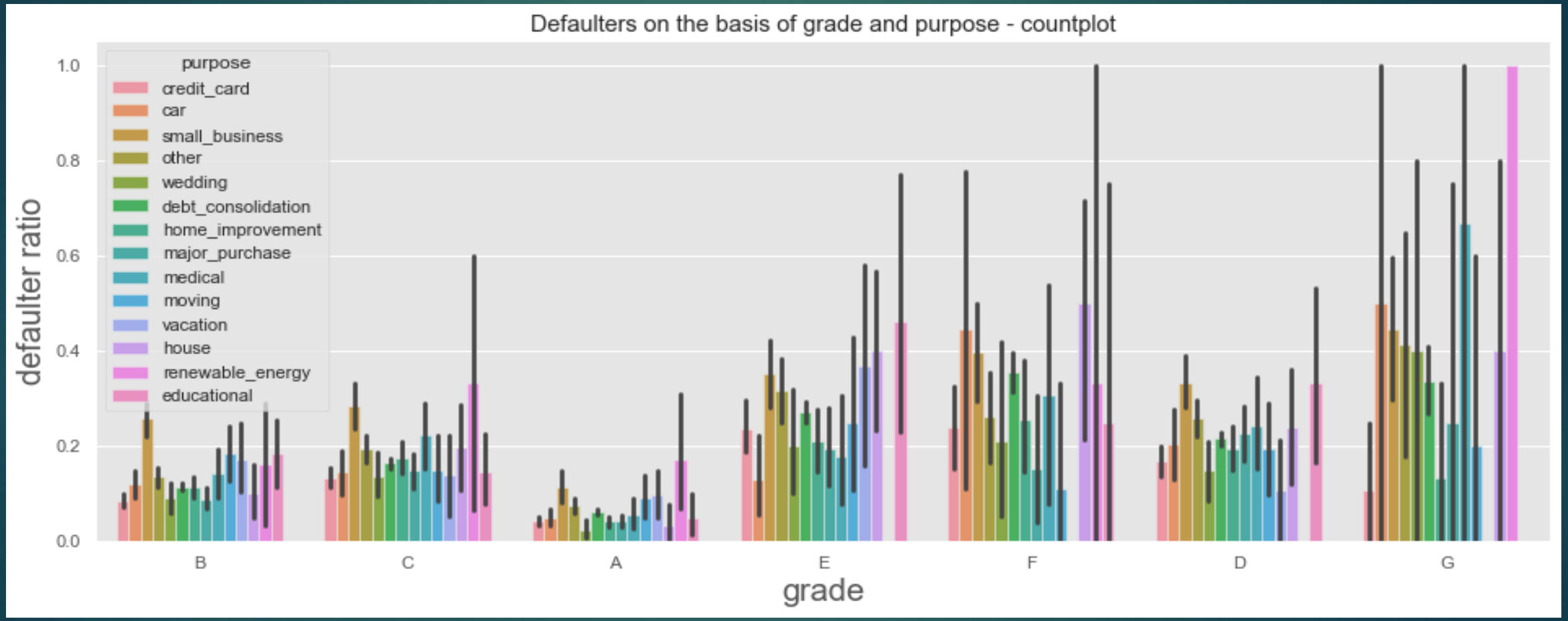
Bivariate Bar Plot

Comparing dataset of *purpose* on basis of *term*



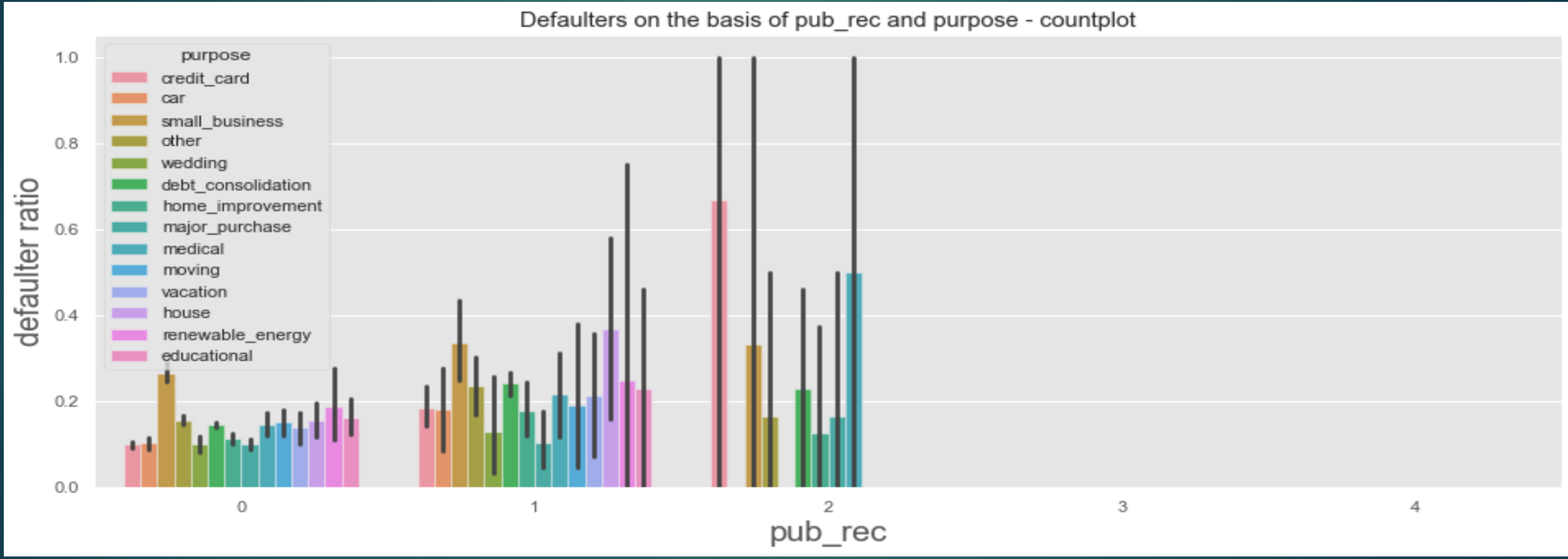
There is co-relation between purpose and term, i.e., on basis of term of defaulters purpose varies.

Comparing dataset of *purpose* on basis of *grade*



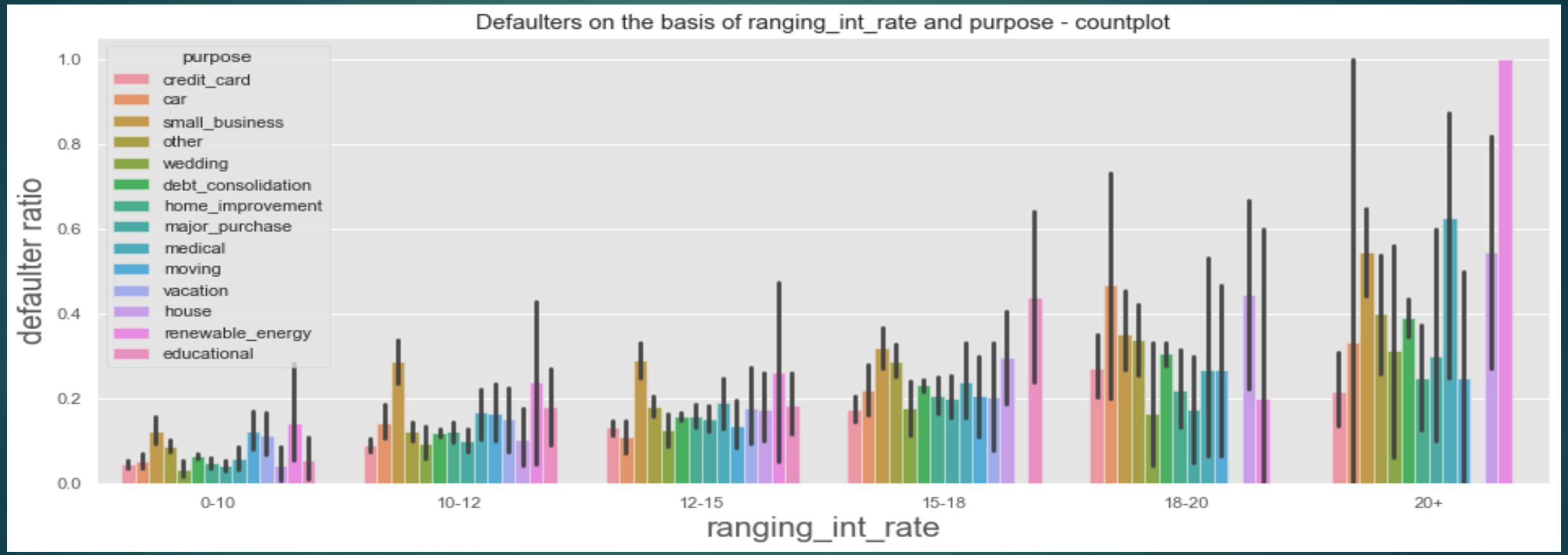
Purpose increases as related to grade of defaulters.

Comparing dataset of *purpose* on basis of *pub_rec*



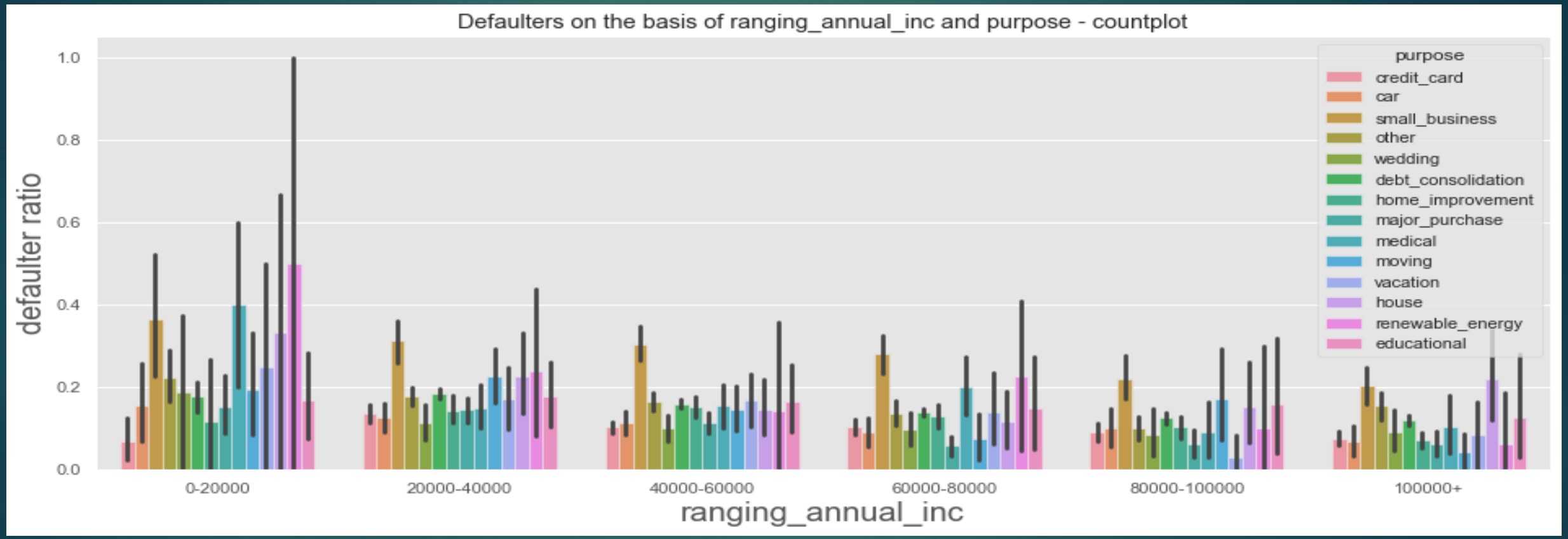
There is co-relation seen in this comparison. Hence, this data is not so useful.

Comparing dataset of *purpose* on basis of *int_rate*



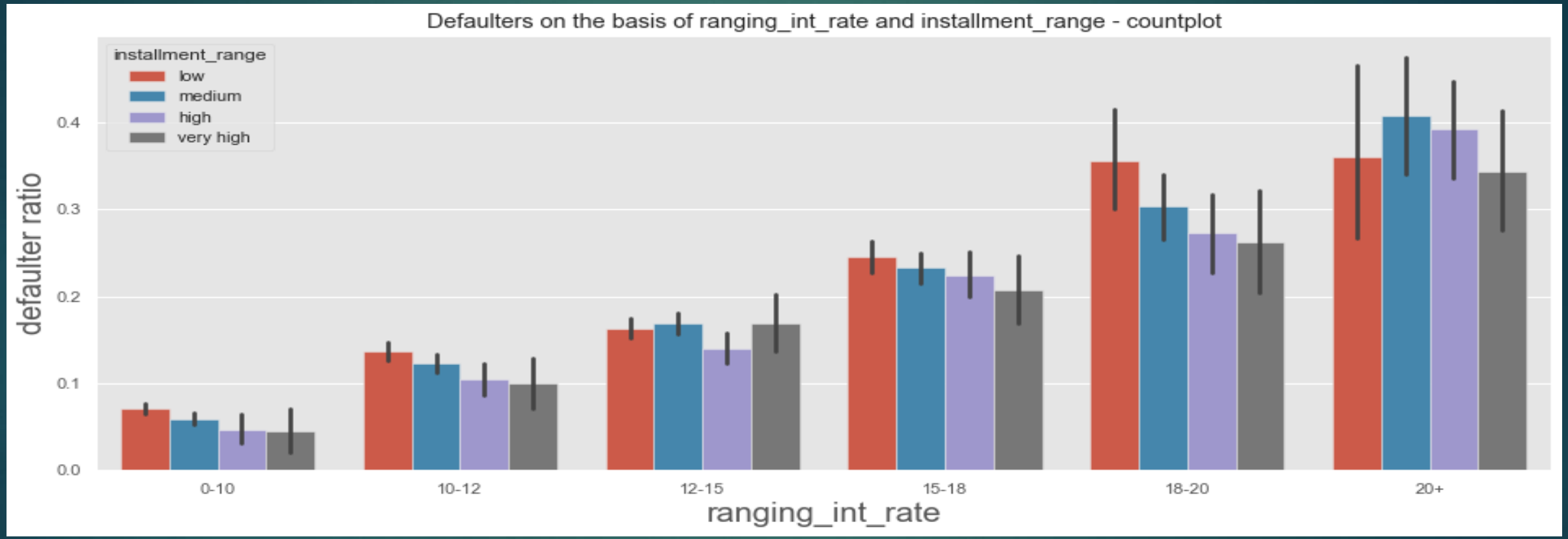
Purpose increases as per the rate of interest.

Comparing dataset of *purpose* on basis of *annual_inc*



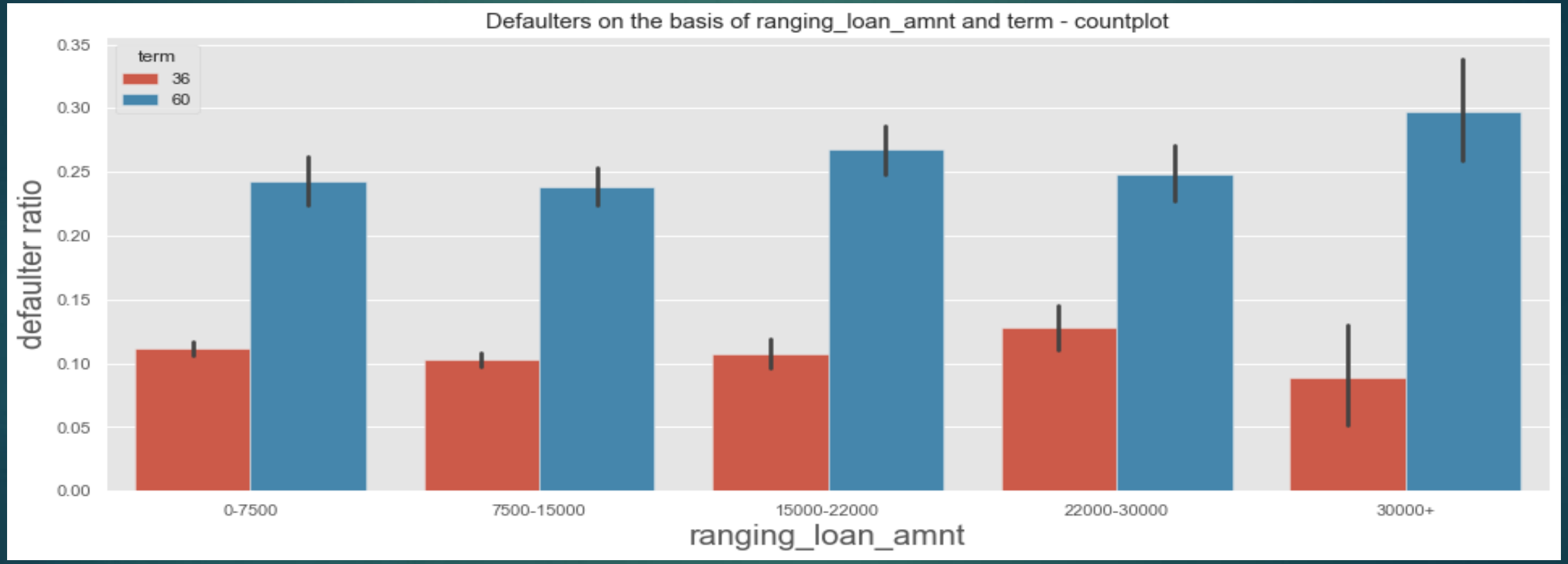
Purpose decreases as per the increase in annual_inc of the dataset.

Comparing dataset of *installment* on basis of *int_rate*



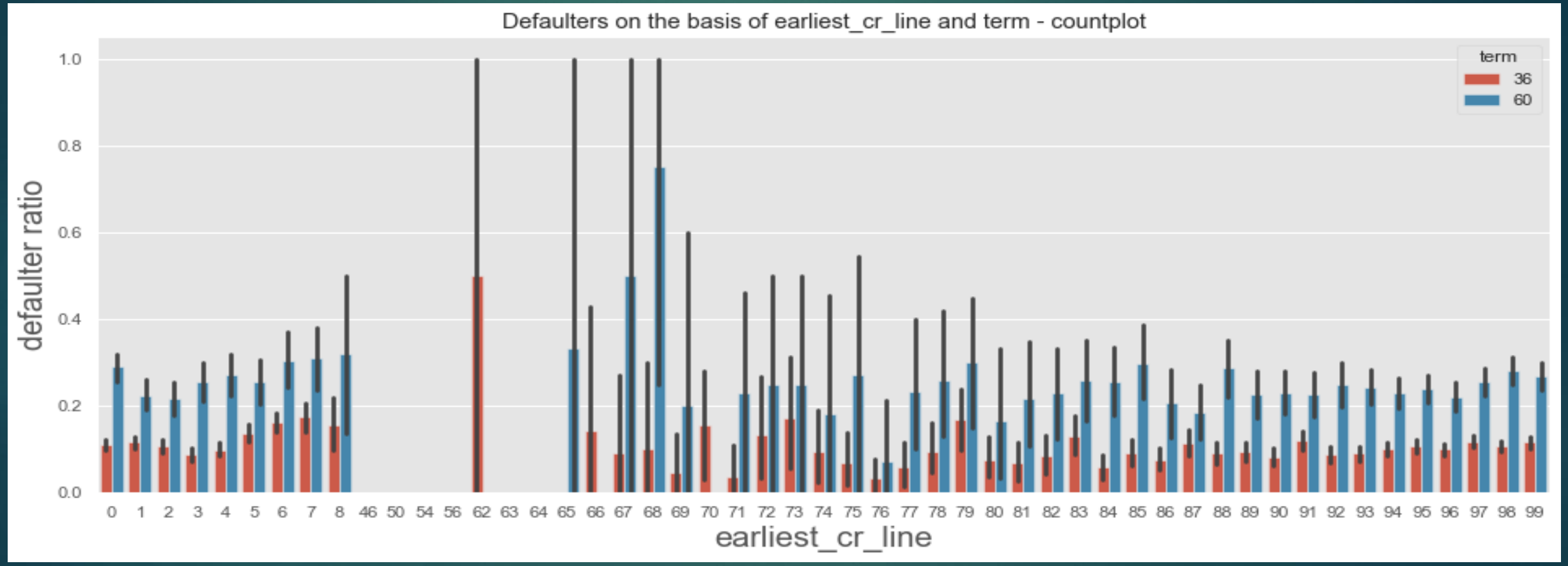
Installment increases with increase of int_rate. Hence, this data is useful.

Comparing dataset of *term* on basis of *loan_amnt*



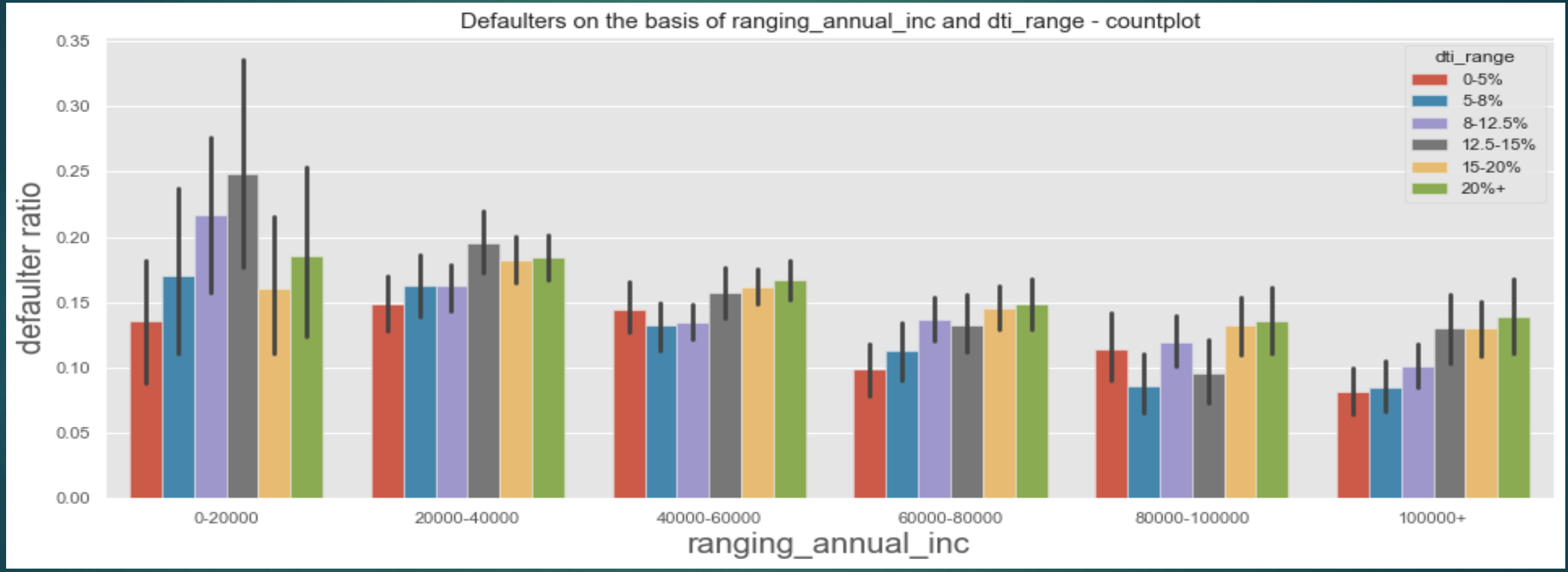
Defaulters take more amount of loan as compare to fully paid clients. But this data is not important.

Comparing dataset of *term* on basis of *earliest_cr_line*



Defaulters are constantly high with earliest_cr_line wrt term of the dataset.

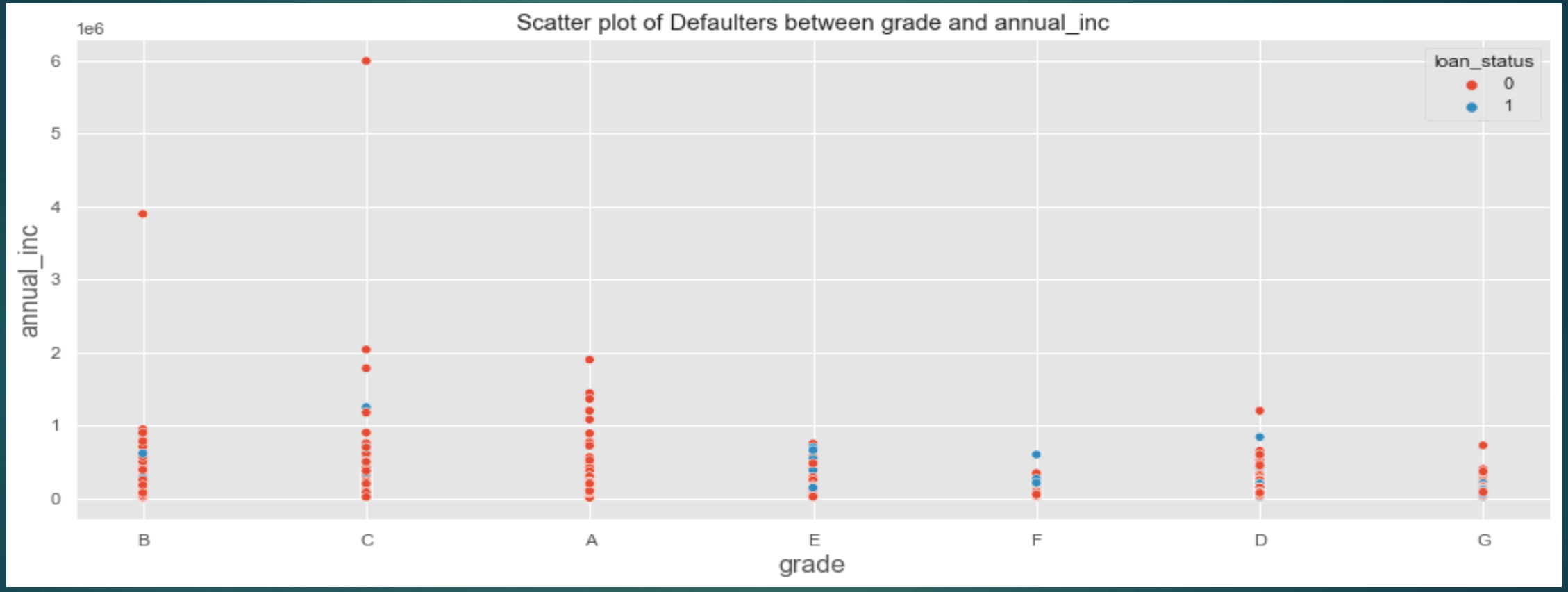
Comparing dataset of *dti* on basis of *annual_inc*



dti decreases with increase of annual_inc of defaulters in dataset.

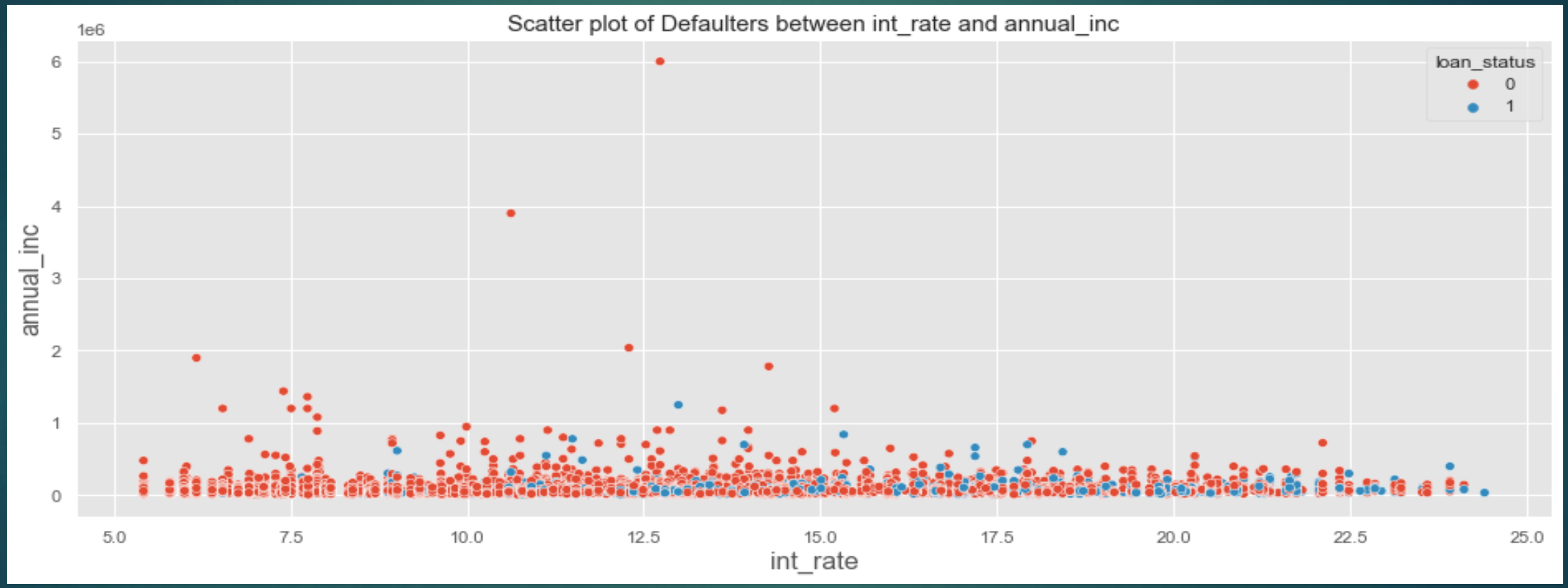
Scatter Plot

Comparing datasets of *annual_inc* and *grade*



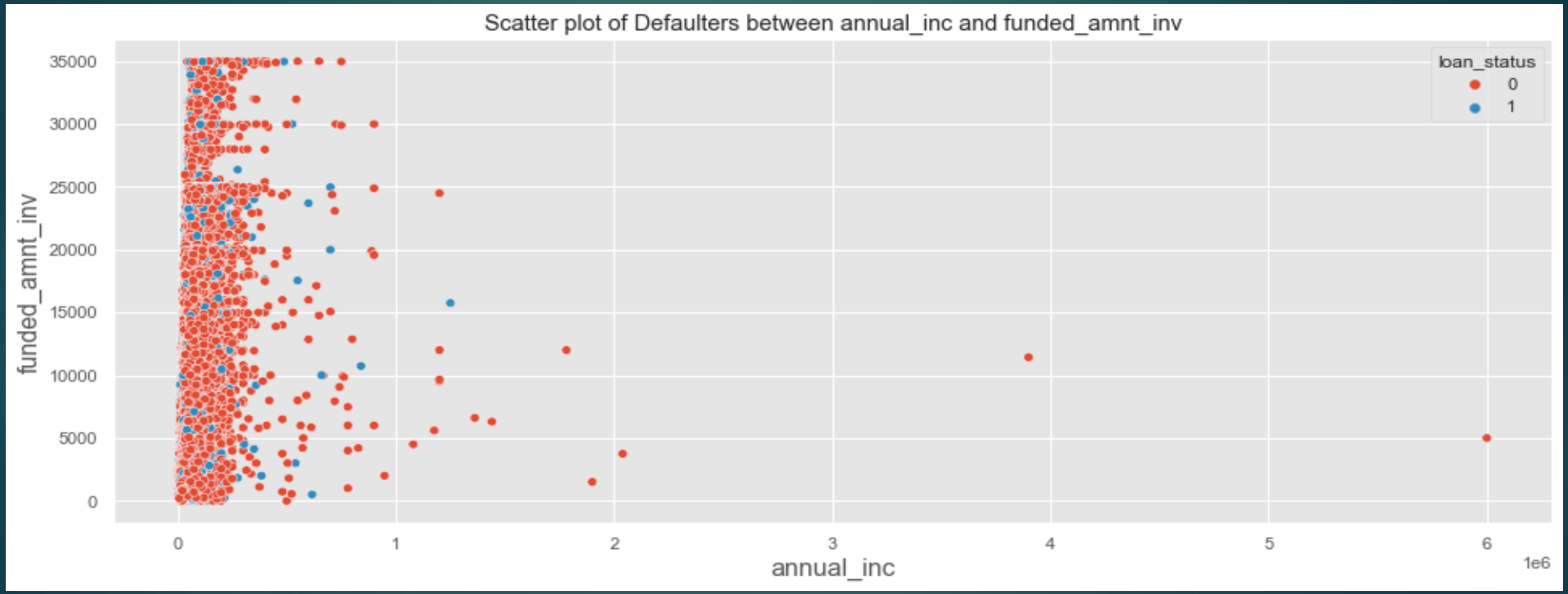
There is co-relation seen in between annual_inc and grade in the dataset.

Comparing datasets of *annual_inc* and *int_rate*



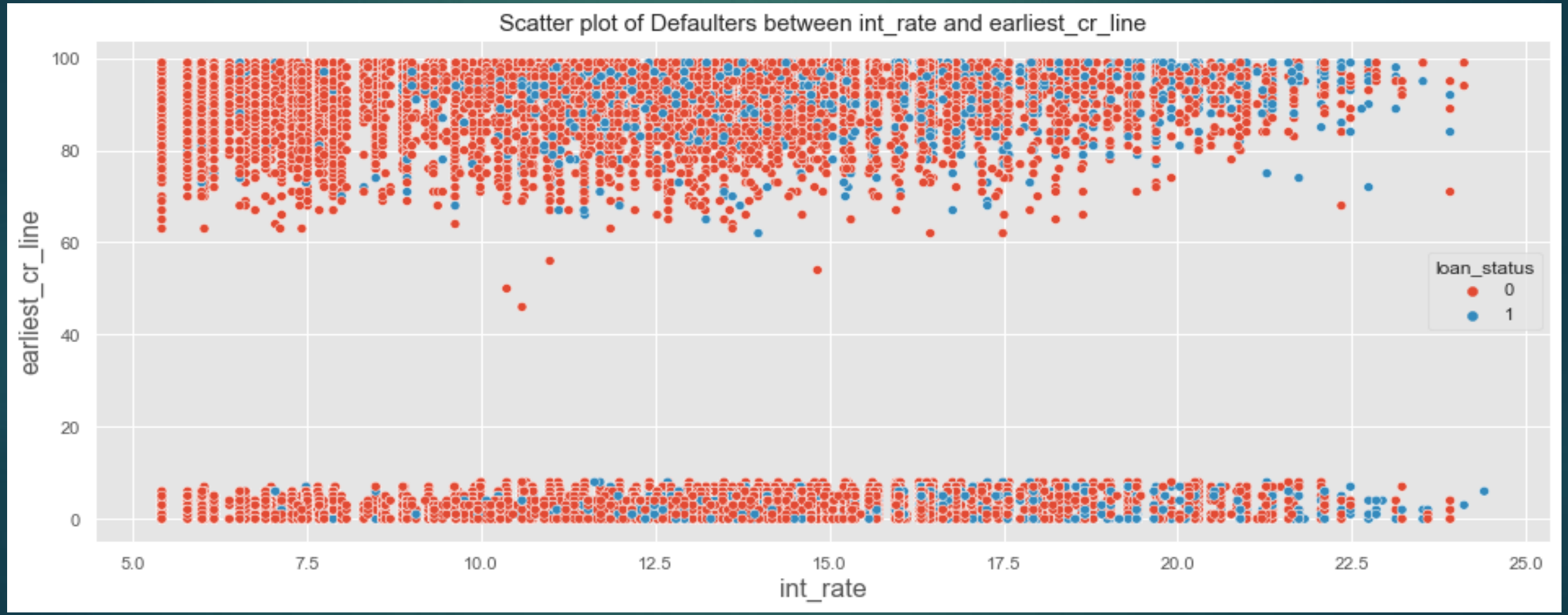
Increase of annual_inc also increases the int_rate of the defaulters.

Comparing datasets of *funded_amnt_inv* and *annual_inc*



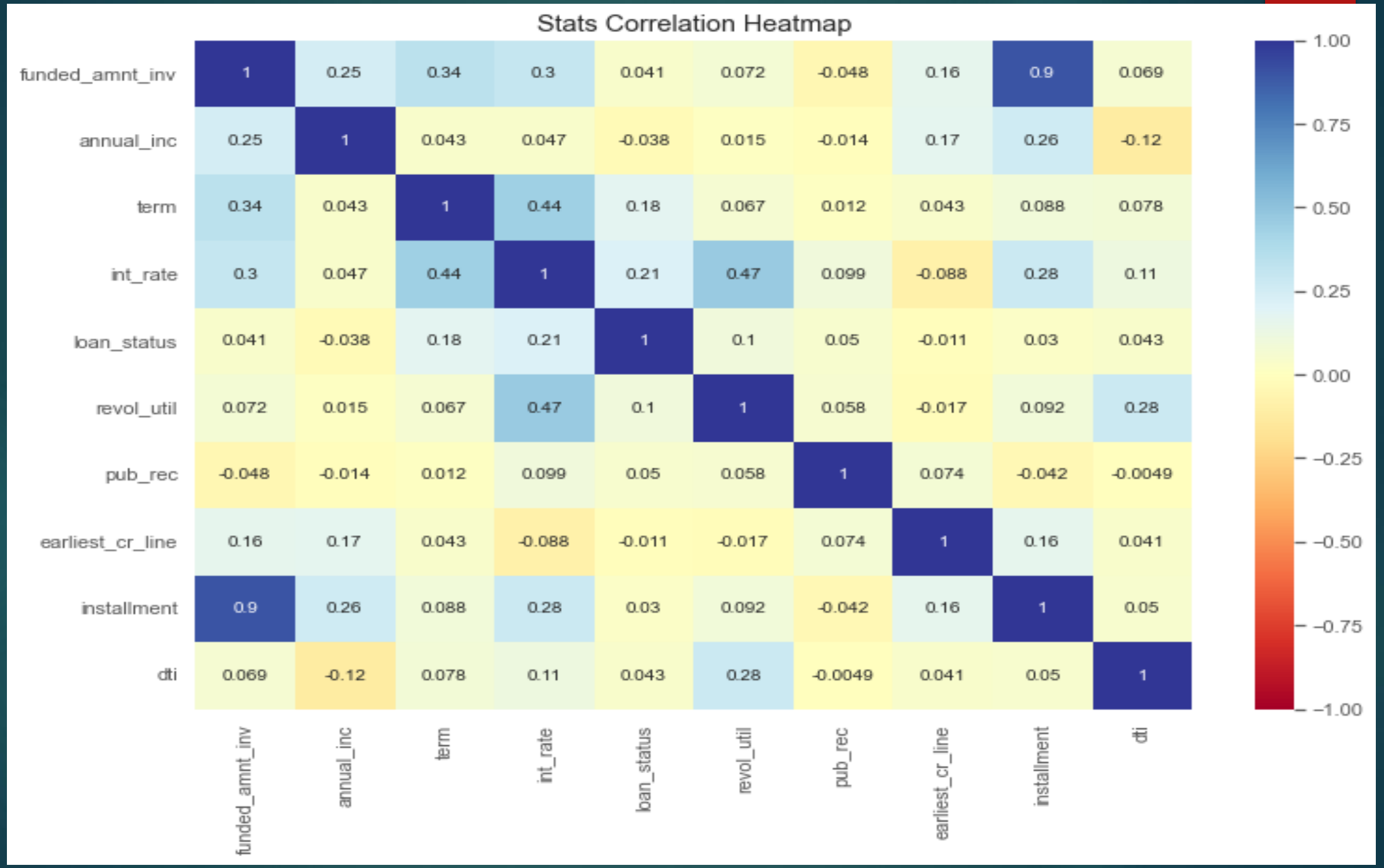
There is no co-relation between funded_amnt_inv and annual_inc.

Comparing datasets of *earliest_cr_line* and *int_rate*



Defaulter are seen more in number with earliest_cr_line as int_rate increases.

Multi-variate Analysis


Stats Correlation Heatmap
funded_amnt_inv 1 0.25 0.34 0.3 0.041 0.072 -0.048 0.16 0.9 0.069
annual_inc 0.25 1 0.043 0.047 -0.038 0.015 -0.014 0.17 0.26 -0.12
term 0.34 0.043 1 0.44 0.18 0.067 0.012 0.043 0.088 0.078
int_rate 0.3 0.047 0.44 1 0.21 0.47 0.099 -0.088 0.28 0.11
loan_status 0.041 -0.038 0.18 0.21 1 0.1 0.05 -0.011 0.03 0.043
revol_util 0.072 0.015 0.067 0.47 0.1 1 0.058 -0.017 0.092 0.28
pub_rec -0.048 -0.014 0.012 0.099 0.05 0.058 1 0.074 -0.042 -0.0049
earliest_cr_line 0.16 0.17 0.043 -0.088 -0.011 -0.017 0.074 1 0.16 0.041
installment 0.9 0.26 0.088 0.28 0.03 0.092 -0.042 0.16 1 0.05
dti 0.069 -0.12 0.078 0.11 0.043 0.28 -0.0049 0.041 0.05 1
funded_amnt_inv
annual_inc
term
int_rate
loan_status
revol_util
pub_rec
earliest_cr_line
installment
dti
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Conclusion

- On observation of univariate analysis, the important data from dataset are *term, grade, purpose, int_rate, annual_inc, funded_amnt_inv,earliest_cr_line*.
- On observation of bivariate analysis, the important data from dataset are *term, grade, sub_grade, purpose, emp_length, earliest_cr_line, revol_util, loan_amnt, annual_inc*.
- On observation of correlation heatmap of multi-variate analysis, the important data from dataset are *term, funded_amnt_inc, int_rate, earliest_cr_line, annual_inc, revol_util*.

After observing all the analysis, the data that effect the most in the dataset are *term, grade, purpose, int_rate, annual_inc, revol_util*.