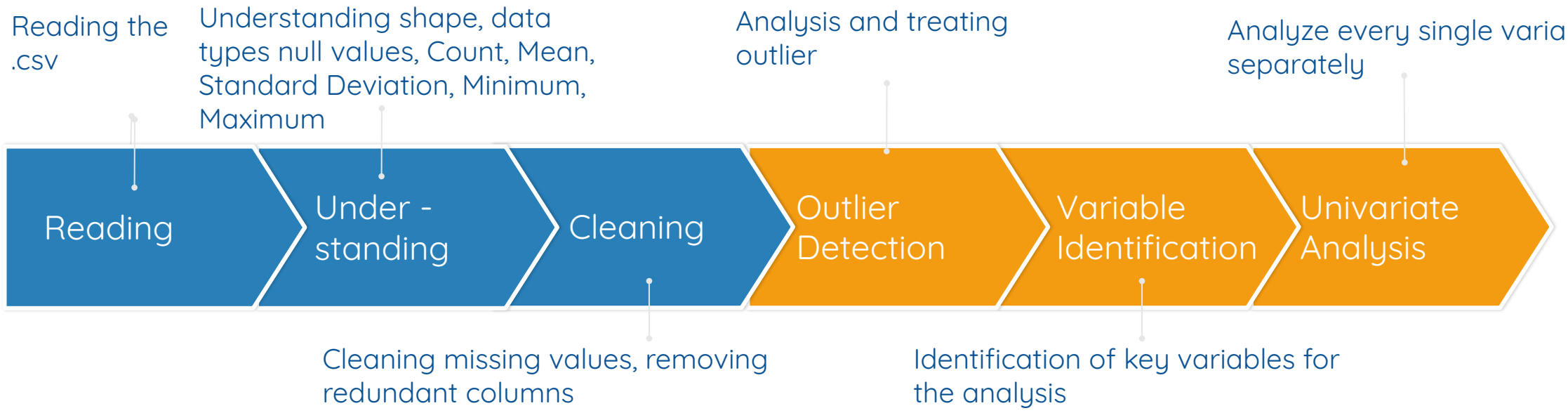
LENDING CLUB CASE STUDY

ANKIT SHARMA
ARPIT SAXENA

Exploratory Data Analysis (EDA)

Lets perform EDA to gain an in-depth understanding of the data.



Reading the Data! Data! Data!

Importing the required Libraries

```
#Importing libraries
import numpy as np # linear algebra
import pandas as pd # data processing,
CSV file I/O (e.g. pd.read_csv)
# Input data files are available in the
"../input/" directory.
import os
import matplotlib.pyplot as
plt#visualization
from PIL import  Image
%matplotlib inline
import pandas as pd
import seaborn as sns#visualization
import itertools
import warnings
warnings.filterwarnings("ignore")
import io
import plotly.offline as py#visualization
py.init_notebook_mode(connected=True)
#visualization
import plotly.graph_objs as
go#visualization
import plotly.tools as tls#visualization
import plotly.figure_factory as
ff#visualization

loan = pd.read_csv("loan.csv", sep=",")
```

id	member_id	loan_amnt	funded_amn	funded_amn	term	int_rate	installment	grade	sub_grade	emp_title	em
1077501	1296599	5000	5000	4975	36 months	10.65%	162.87	B	B2		10+
1077430	1314167	2500	2500	2500	60 months	15.27%	59.83	C	C4	Ryder	< 1
1077175	1313524	2400	2400	2400	36 months	15.96%	84.33	C	C5		10+
1076863	1277178	10000	10000	10000	36 months	13.49%	339.31	C	C1	AIR RESOUR	10+
1075358	1311748	3000	3000	3000	60 months	12.69%	67.79	B	B5	University M	1 ye
1075269	1311441	5000	5000	5000	36 months	7.90%	156.46	A	A4	Veolia Trans	3 ye
1069639	1304742	7000	7000	7000	60 months	15.96%	170.08	C	C5	Southern Sta	8 ye
1072053	1288686	3000	3000	3000	36 months	18.64%	109.43	E	E1	MKC Account	9 ye
1071795	1306957	5600	5600	5600	60 months	21.28%	152.39	F	F2		4 ye
1071570	1306721	5375	5375	5350	60 months	12.69%	121.45	B	B5	Starbucks	< 1
1070078	1305201	6500	6500	6500	60 months	14.65%	153.45	C	C3	Southwest R	5 ye
1069908	1305008	12000	12000	12000	36 months	12.69%	402.54	B	B5	UCLA	10+
1064687	1298717	9000	9000	9000	36 months	13.49%	305.38	C	C1	Va. Dept of C	< 1
1069866	1304956	3000	3000	3000	36 months	9.91%	96.68	B	B1	Target	3 ye
1069057	1303503	10000	10000	10000	36 months	10.65%	325.74	B	B2	SFMTA	3 ye
1069759	1304871	1000	1000	1000	36 months	16.29%	35.31	D	D1	Internal reve	< 1
1065775	1299699	10000	10000	10000	36 months	15.27%	347.98	C	C4	Chin's Restau	4 ye
1069971	1304884	3600	3600	3600	36 months	6.03%	109.57	A	A1	Duracell	10+
1062474	1294539	6000	6000	6000	36 months	11.71%	198.46	B	B3	Connection In	1 ye
1069742	1304855	9200	9200	9200	36 months	6.03%	280.01	A	A1	Network Inte	6 ye
1069740	1284848	20250	20250	19142.1611	60 months	15.27%	484.63	C	C4	Archdiocese	3 ye
1039153	1269083	21000	21000	21000	36 months	12.42%	701.73	B	B4	Osram Sylva	10+
1069710	1304821	10000	10000	10000	36 months	11.71%	330.76	B	B3	Value Air	10+
1069700	1304810	10000	10000	10000	36 months	11.71%	330.76	B	B3	Wells Fargo	5 ye
1069559	1304634	6000	6000	6000	36 months	11.71%	198.46	B	B3	bmg-educati	1 ye
1069697	1273773	15000	15000	15000	36 months	9.91%	483.38	B	B1	Winfield Pat	2 ye
1069800	1304679	15000	15000	8725	36 months	14.27%	514.64	C	C2	nyc transit	9 ye
1069657	1304764	5000	5000	5000	60 months	16.77%	123.65	D	D2	Frito Lay	2 ye
1069799	1304678	4000	4000	4000	36 months	11.71%	132.31	B	B3	Shands Hosp	10+
1047704	1278806	8500	8500	8500	36 months	11.71%	281.15	B	B3	Oakridge hor	< 1
1032111	1261745	4375	4375	4375	36 months	7.51%	136.11	A	A3		7 ye

Understanding the data

Checking Shape, Head, Describe and Info

Shape

Rows	Columns
39717	111

Data has 39717 rows and 111 Columns.

Info

Info is not showing Non-Null Count and Dtype because data is not clean.

So, lets clean the data to understand it better!!

Head

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	...
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000	39717.000000	3.971700e+04	39717.000000	39717.000000	39717.000000	...
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868	324.561922	6.896893e+04	13.315130	0.146512	0.869200	...
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439	208.874874	6.379377e+04	6.678594	0.491812	1.070219	...
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	0.000000	...
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000	167.020000	4.040400e+04	8.170000	0.000000	0.000000	...
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000	280.220000	5.900000e+04	13.400000	0.000000	1.000000	...
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000	430.780000	8.230000e+04	18.600000	0.000000	1.000000	...
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	11.000000	8.000000	...

8 rows × 87 columns

Cleaning the data

Missing Values

56 Columns has 90% or more missing values.
Dropping them for better readability and analysis,

Dropping desc mths_since_last_delinq with 32.6% and 64.7% missing values respectively.

Dropping 9 columns where all values are zero.

No row has more than 5 missing values.

This reduces the shape of data to (39717, 44).

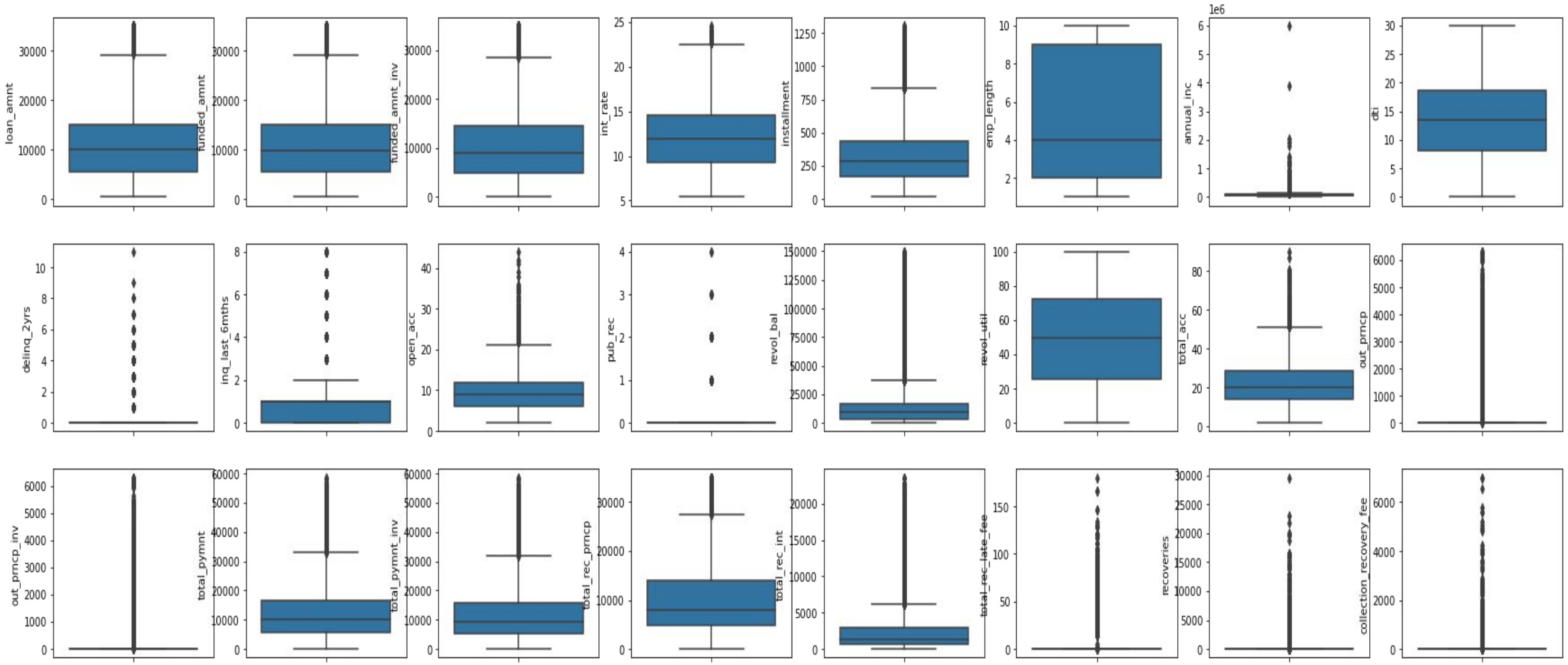
Incorrect Data Types

Two columns, int_rate and revol_util has dtype = 'object' because of % at the end.
Lets convert them to integer.

Five date columns are stored as object.
Lets convert them to date.

Outlier Detection

Checking outliers on the lower and upper side in all the continuous columns
And replacing them with the lower and upper bound of inter-quartile range respectively.



Variable Identification

Identification of key variables for the analysis

The goal is to identify predictors of default so that we can use those variables for approval/rejection of the loan at the time of loan application.

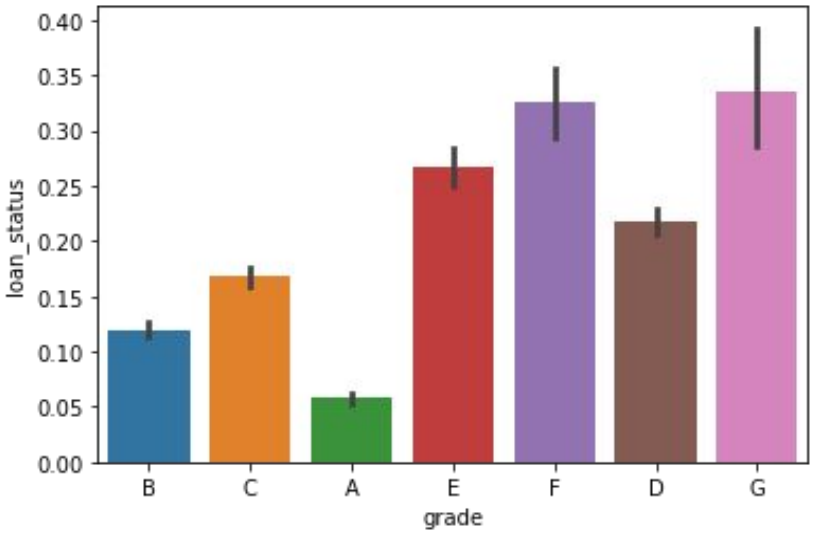
Three types of variables -
1. Those related to the applicant (demographic variables such as age, occupation, employment details etc.)
2. loan characteristics (loan amount, interest rate, the purpose of loan etc.)
3. Customer behaviour variables (those generated after the loan was approved, such as delinquent two years, revolving balance, next payment date etc.).

Now, the customer behaviour parameters are not there at the time of loan approval, and thus they cannot be used as parameters for credit approval.

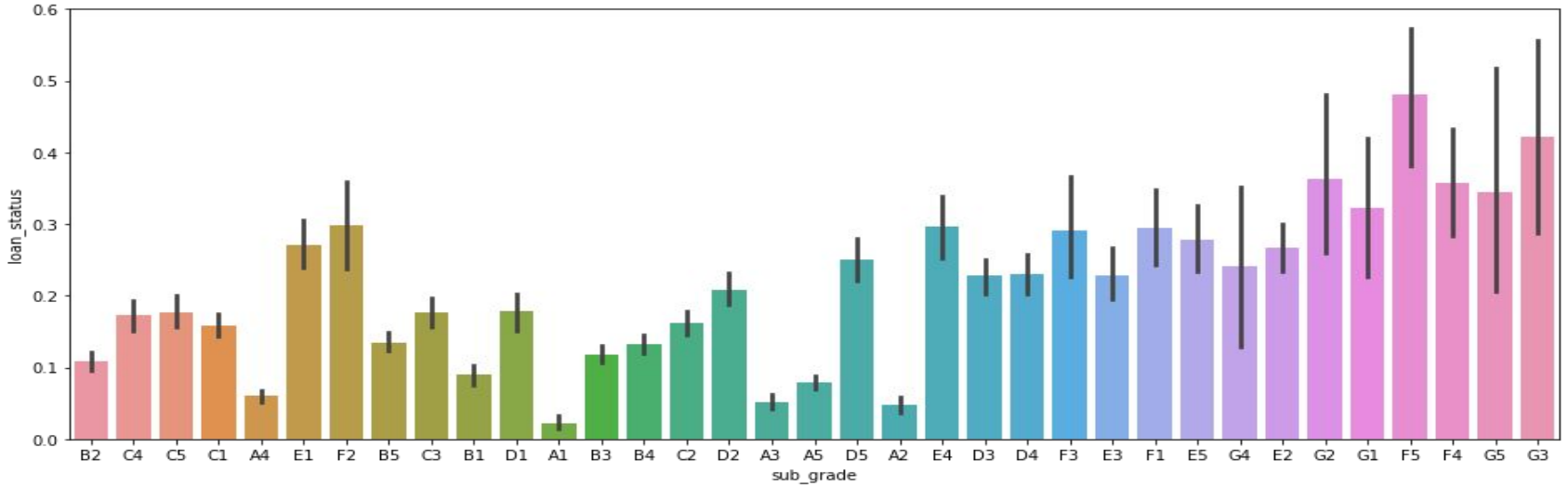
Thus, as we advance, we will use only the other two types of variables.

Creating dummy variables for loan_status
'Current' are neither fully paid not defaulted, so let's get rid of the current loans
Marking ‘Fully Paid’ as 0
And ‘Charged Off’ as 1

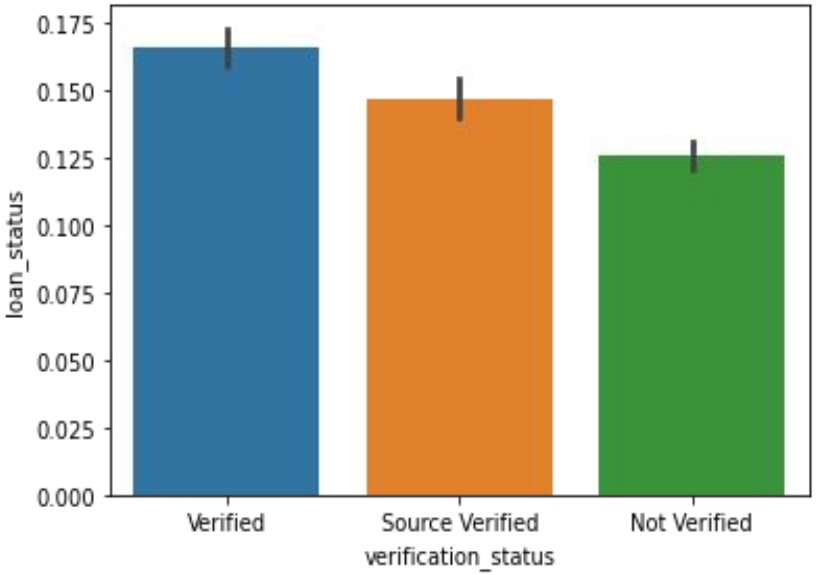
Univariate Analysis of Categorical Variables



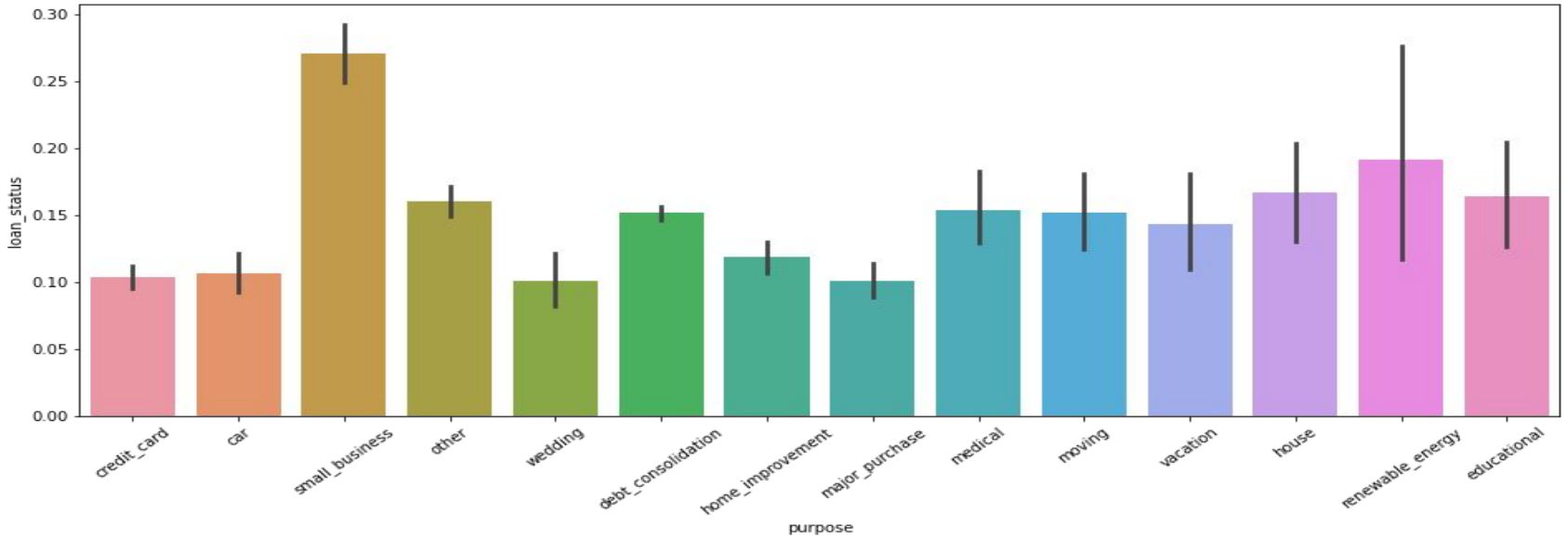
Default Rate increases from grade A to G.



Default Rate increases with sub grade rank across all grades.

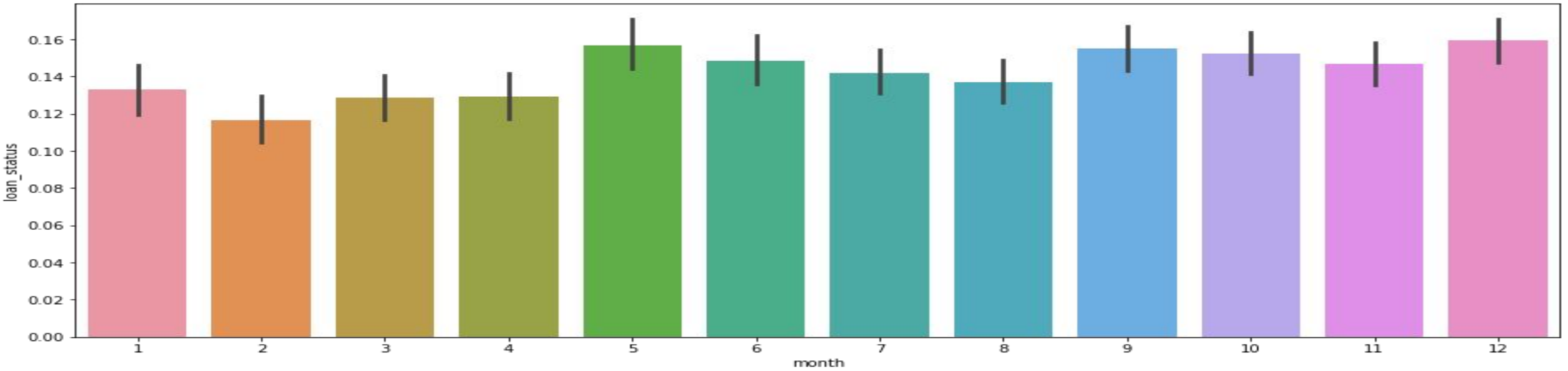


Default Rate is high in verified loans.

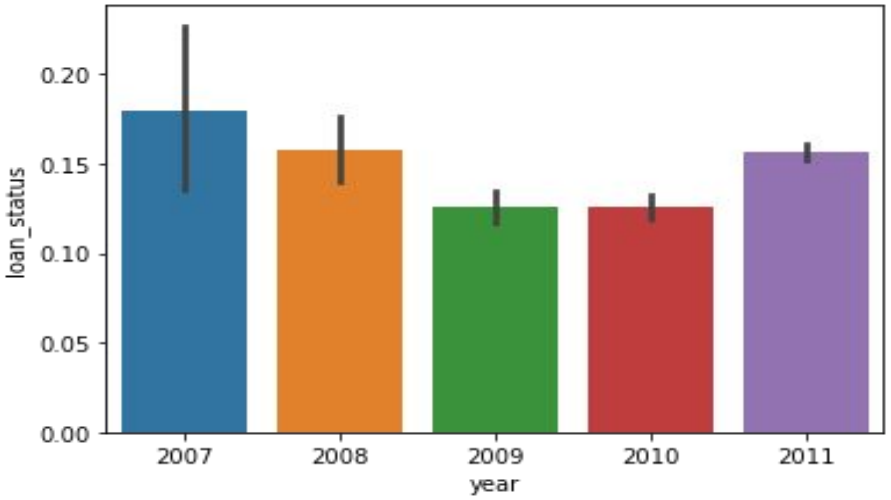


Default Rate is max in small businesses.

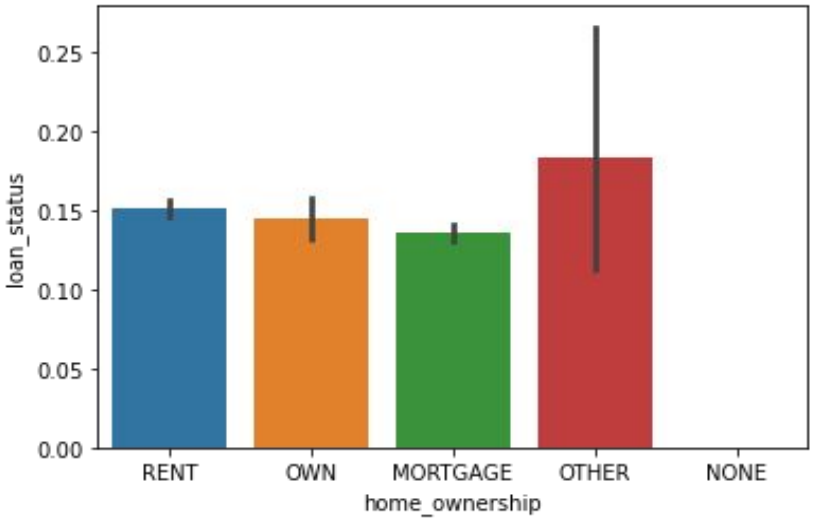
Univariate Analysis of Categorical Variables



Maximum Loan approvals happens in December.



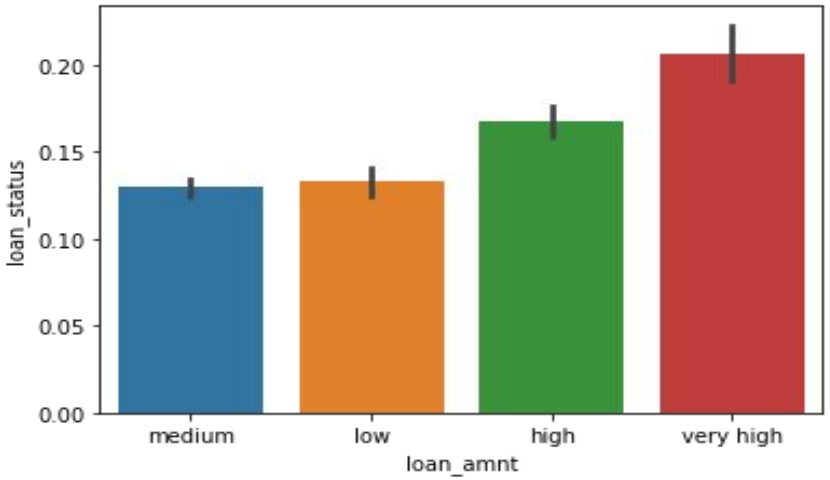
2007 & 2009 have max and min loan_status.



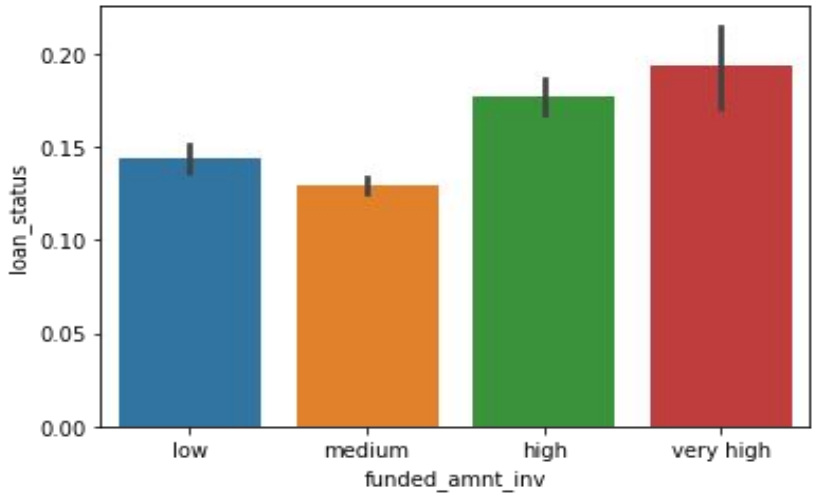
Default Rate is highest in small businesses.

Univariate Analysis of Continuous Variables

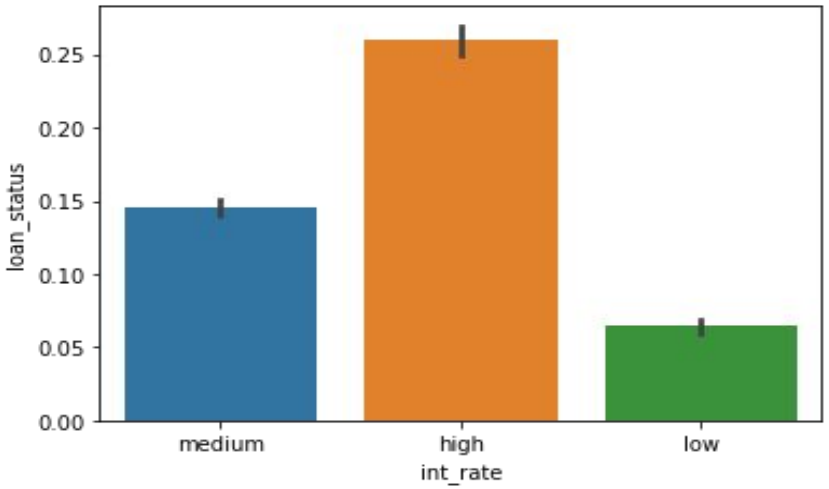
Making buckets of continuous variables to analyze in a more meaningful manner.



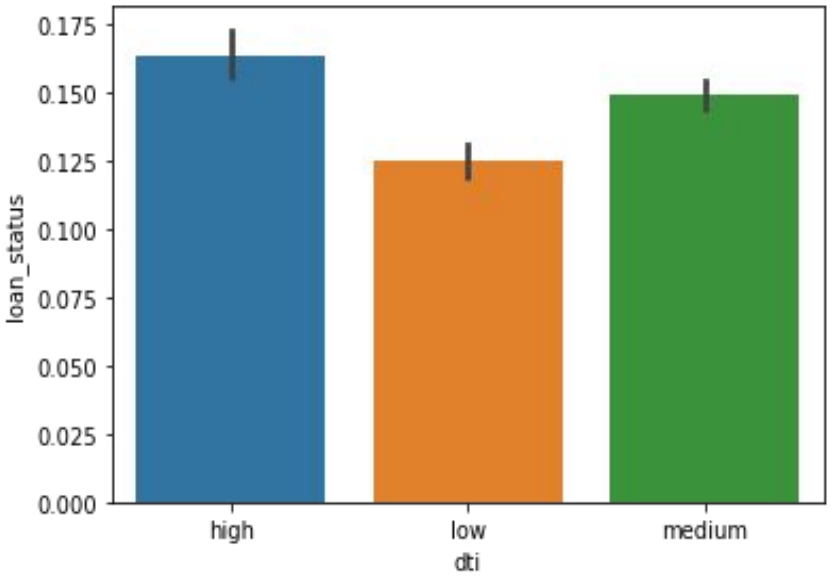
Default Rate is high for “very high” loan amount.



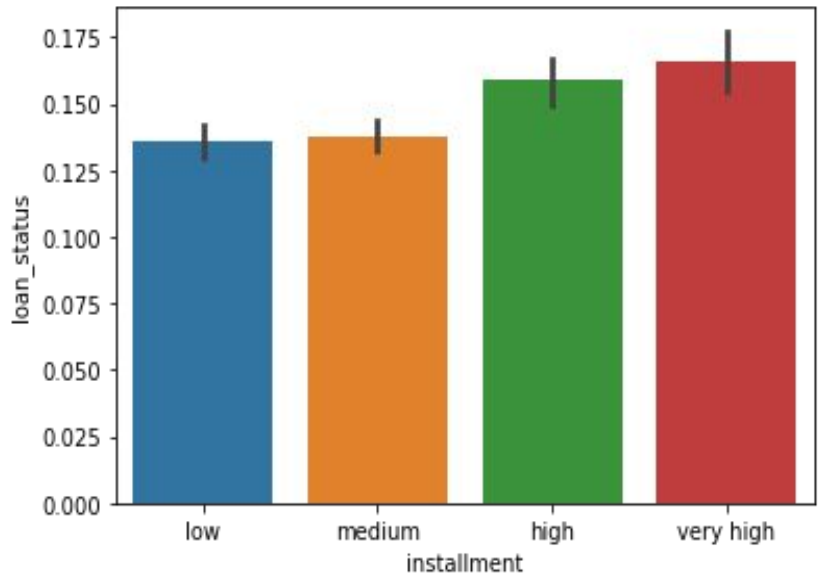
Default Rate is high for “very high” funded amount inv.



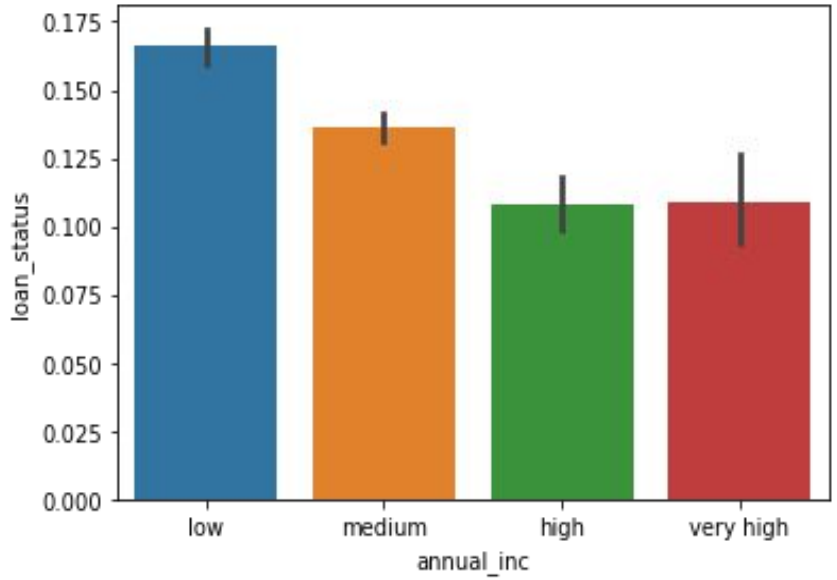
Default Rate is high for “high” interest rate.



Default Rate is high for “high” debt to income ratios.

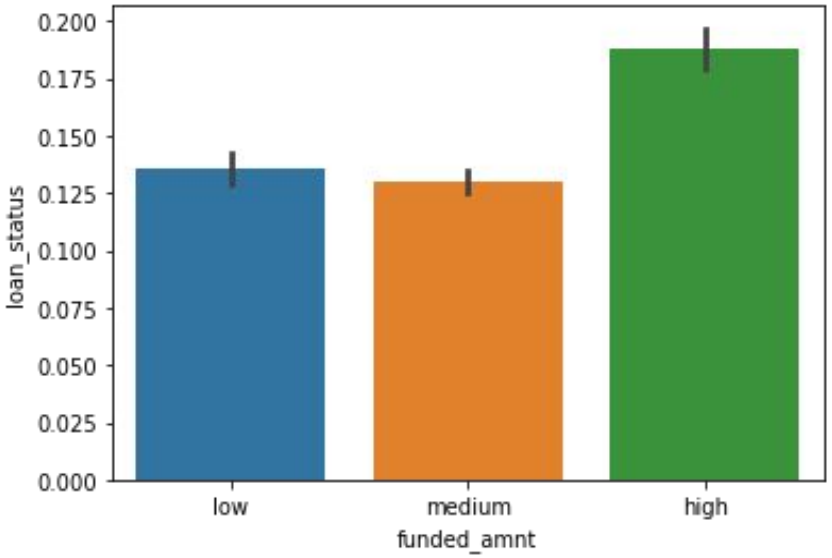


Default Rate is high for “very high” count of installment.

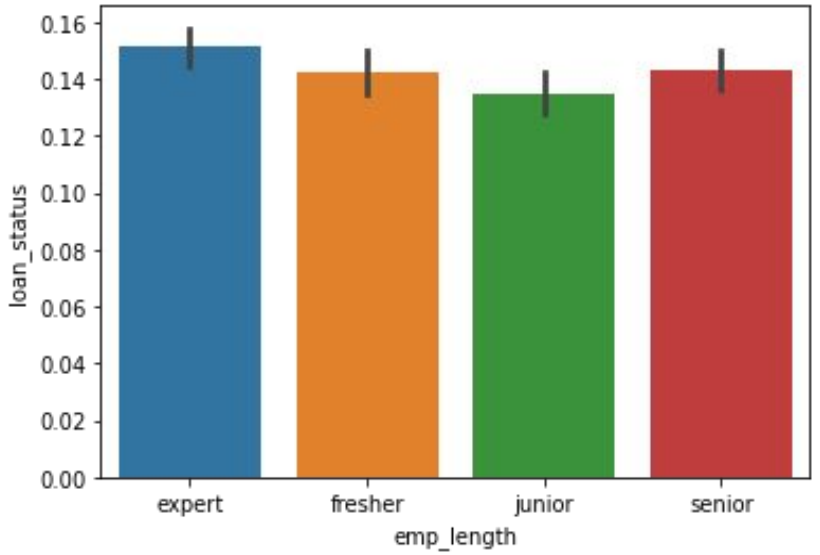


Default Rate is high for “low” annual income.

Univariate Analysis of Continuous Variables

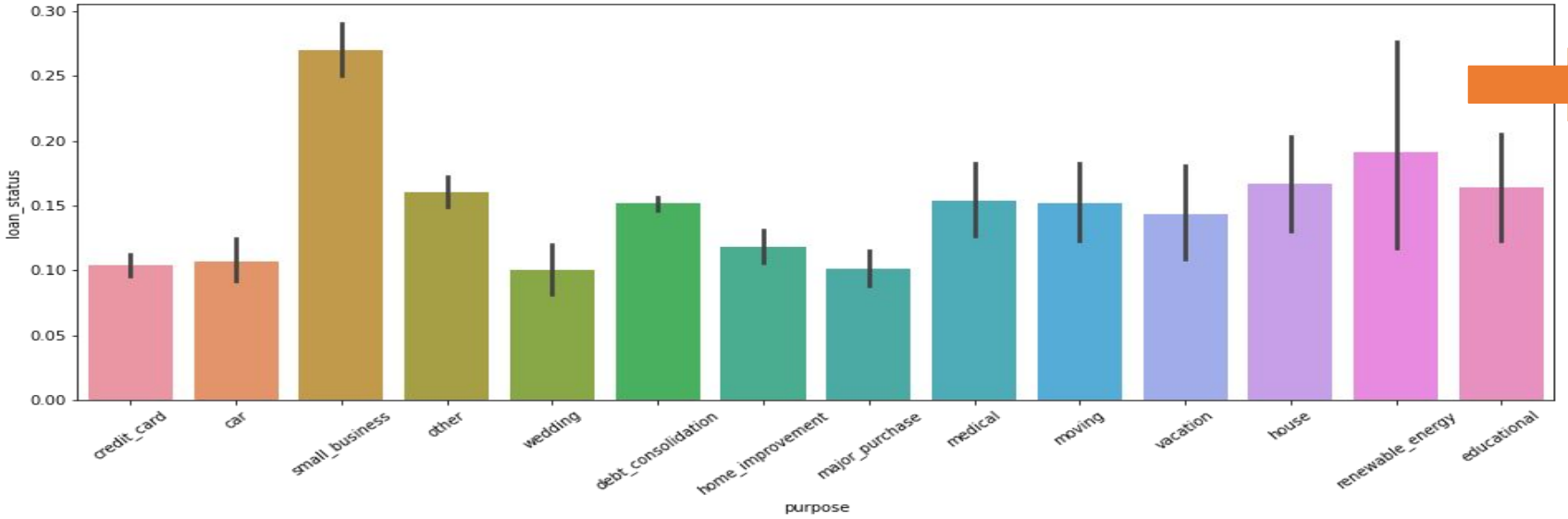


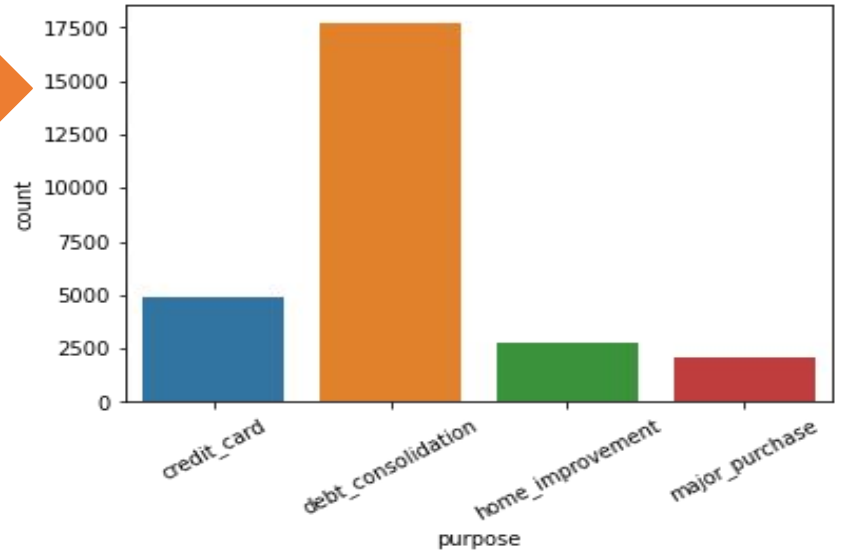
Default Rate is high for “high” funded amount.



Default Rate is high for experienced.

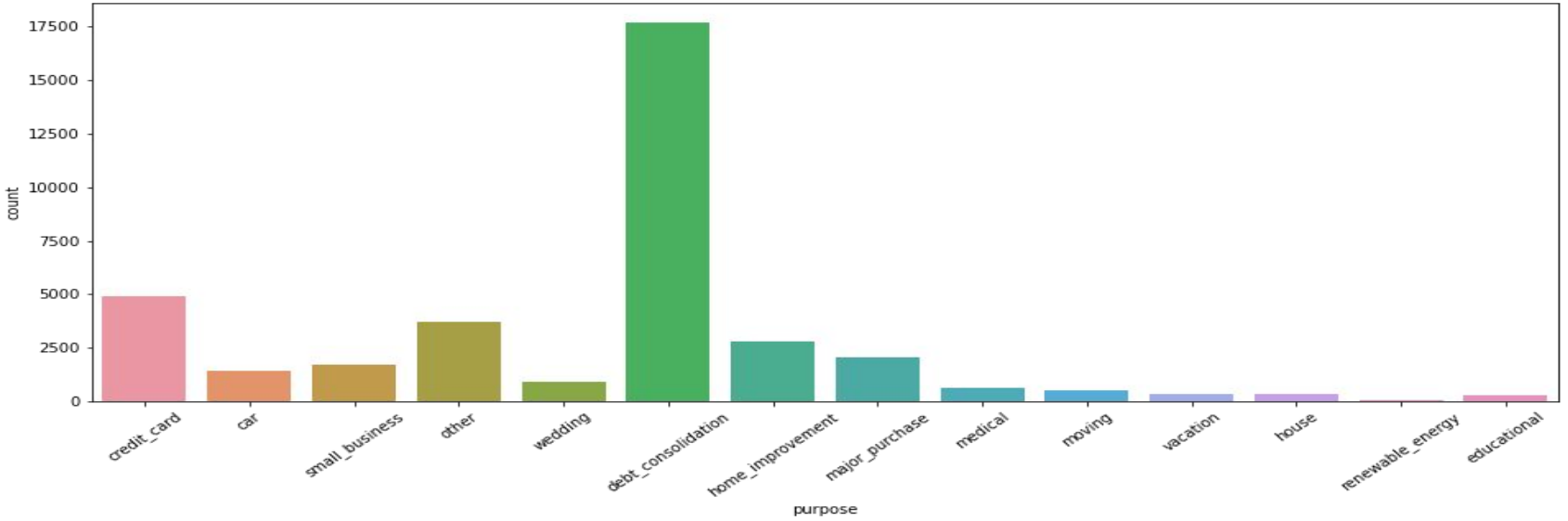
Segmented Univariate Analysis

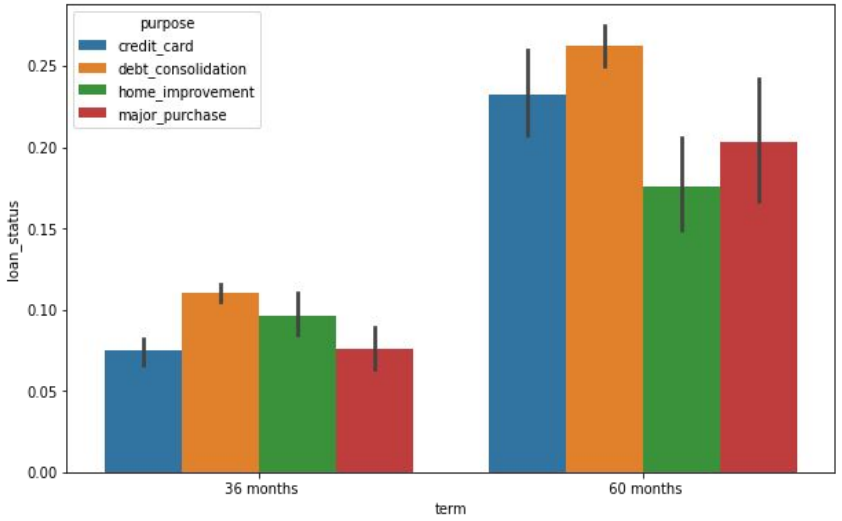




Default Rate is max in small businesses.

4 major categories are mentioned above.
Count of loan disbursed is max for debt consolidation.

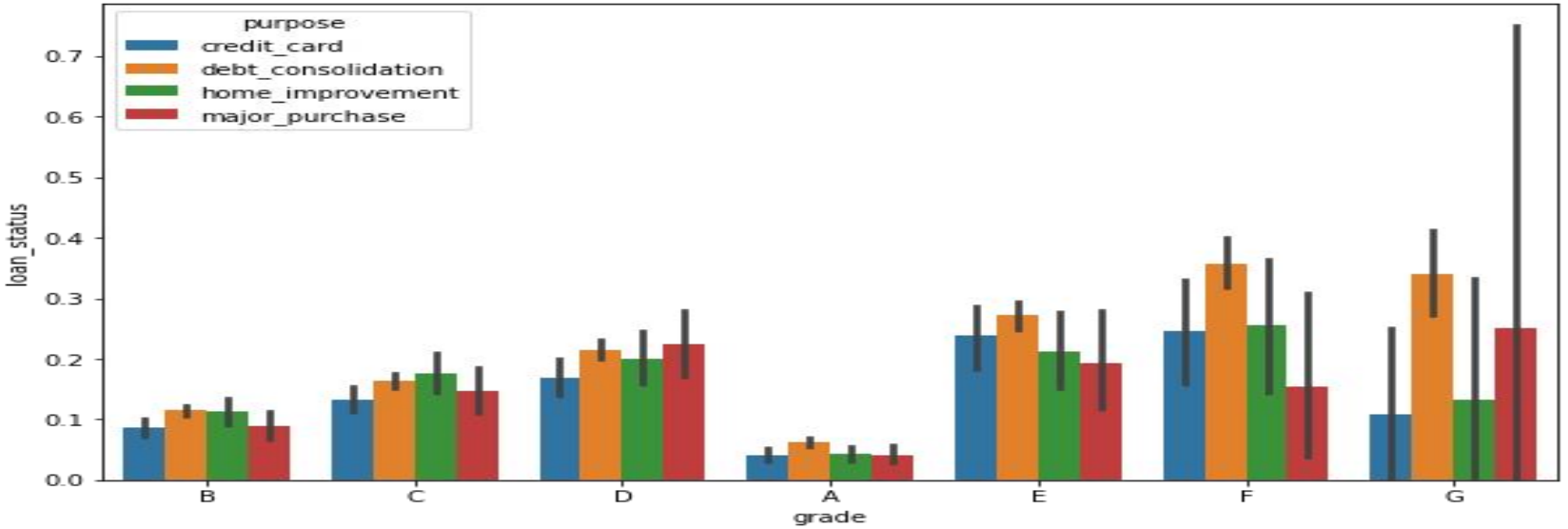




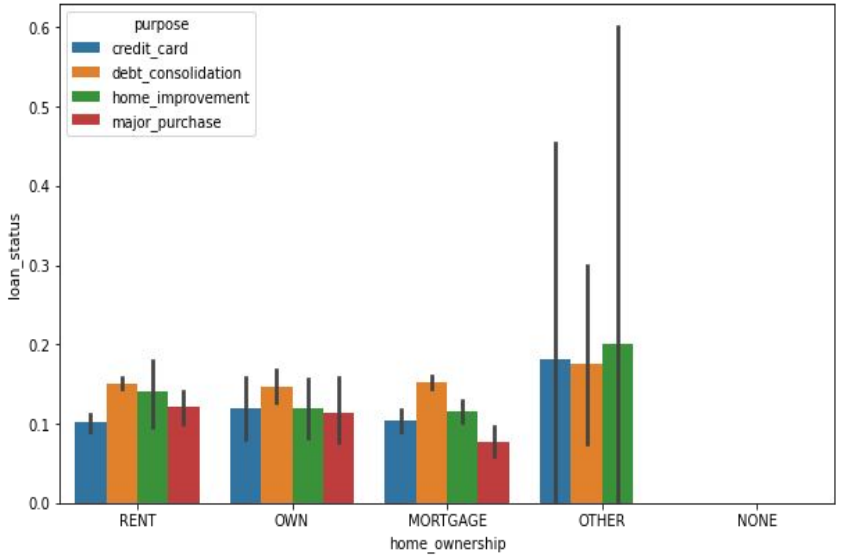
Most loans are debt consolidation (to repay other debts), then credit card, major purchase etc.

Debt consolidation is still highest either in 36 or 60 months term.

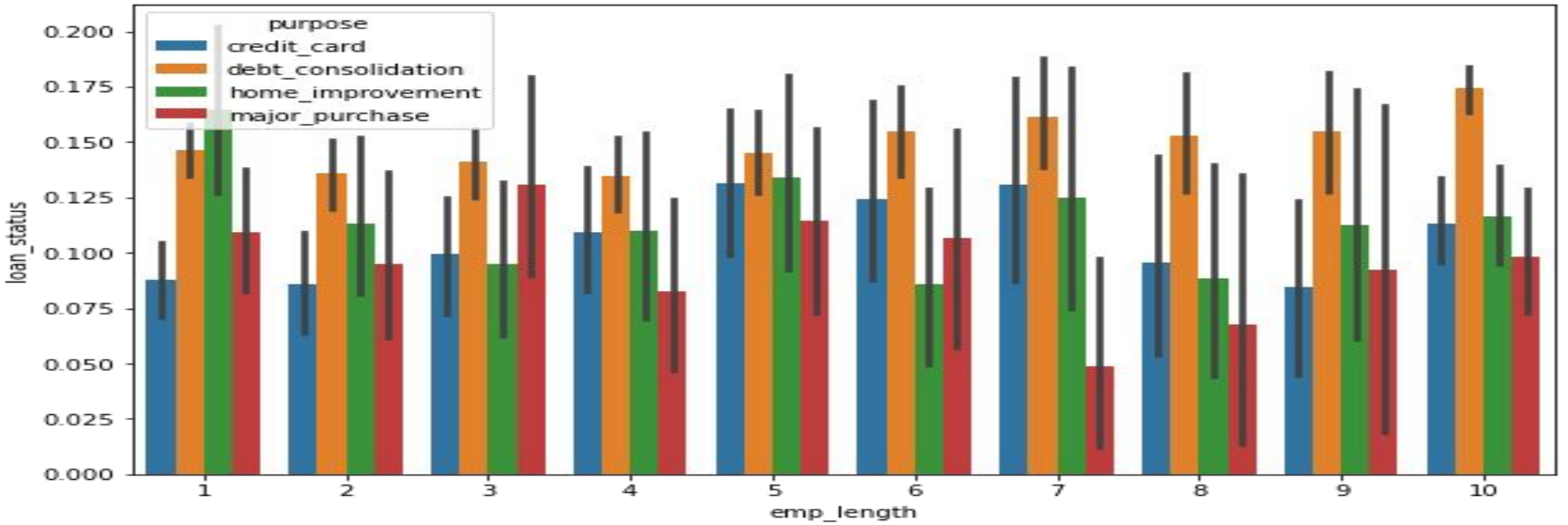
Segmented Univariate Analysis



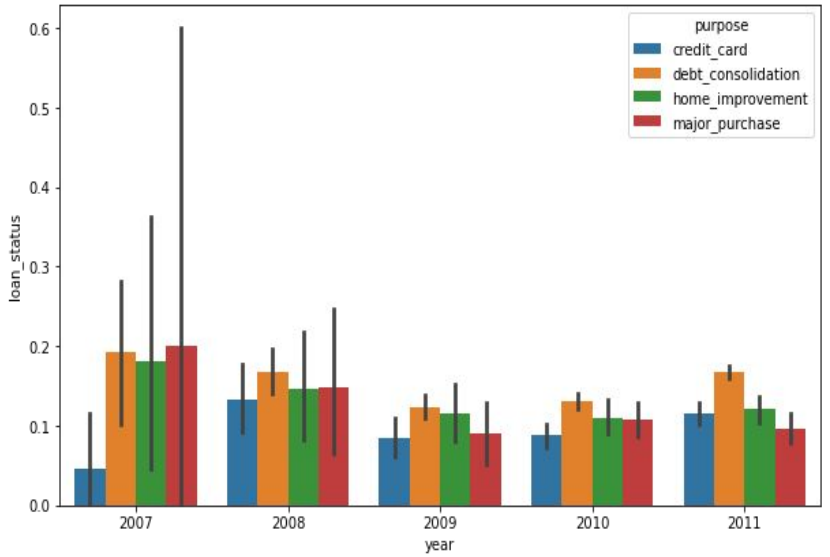
Default Rate is highest for debt consolidation is highest among all grades.



Debt consolidation is highest among all home_owneship except others.

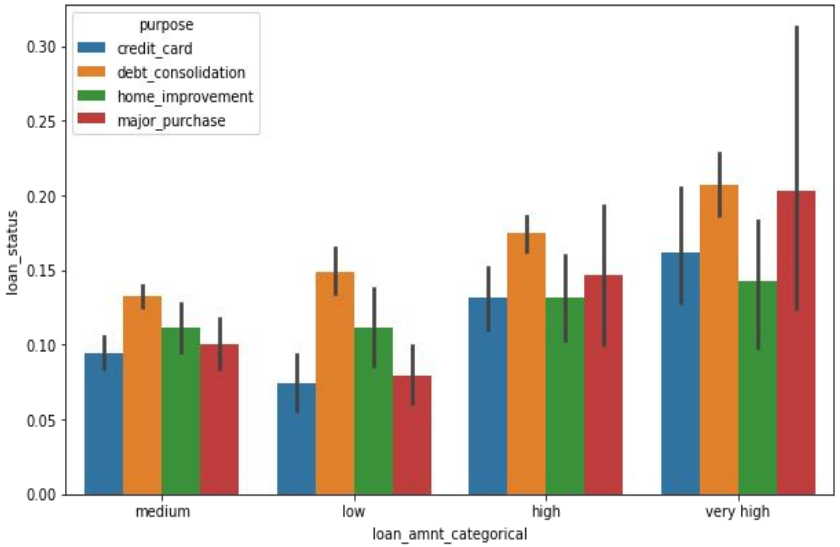


Debt consolidation is highest among employees experience of all years except 1 year.

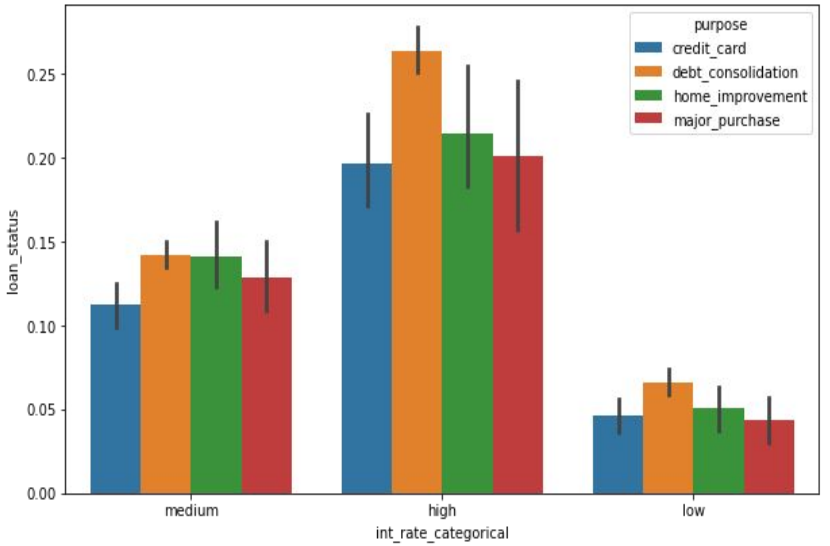


Debt consolidation is highest among all years except 2007.

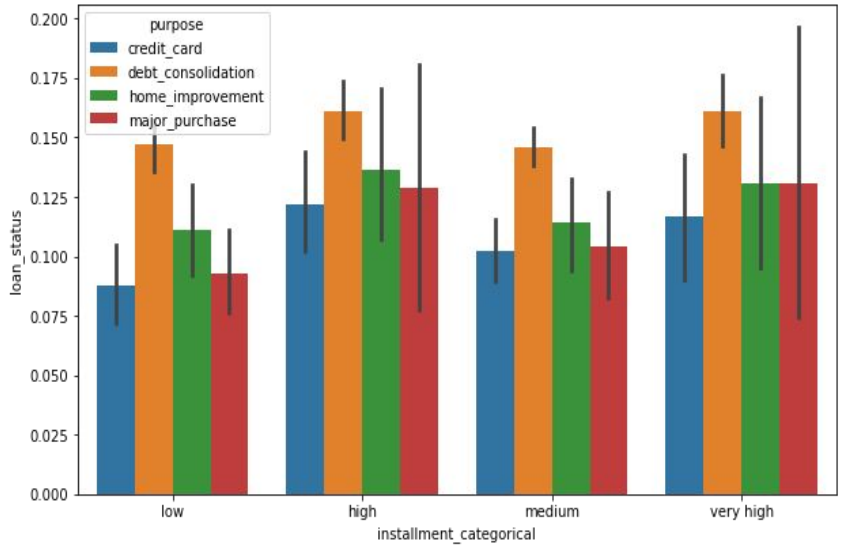
Segmented Univariate Analysis



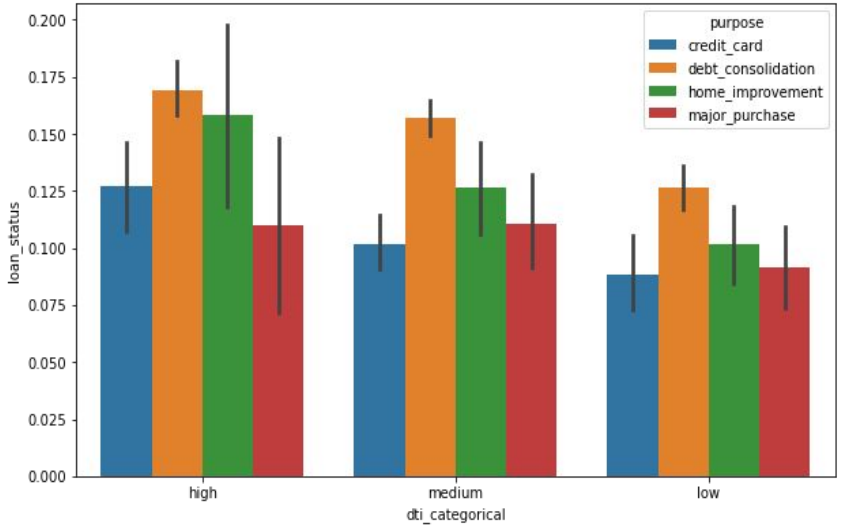
All 4 purpose are are high in “very high” loan amount.



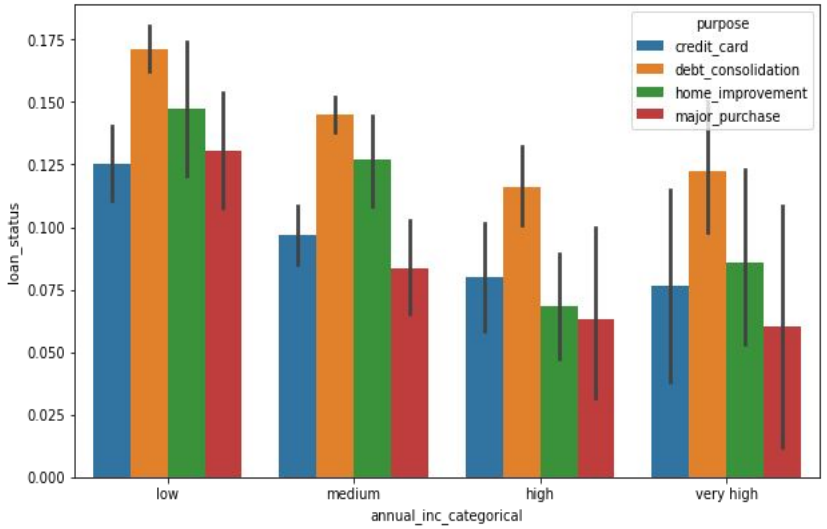
All 4 purpose are are high in high” interest rate..



Installment also has debt consolidation as the highest factor.

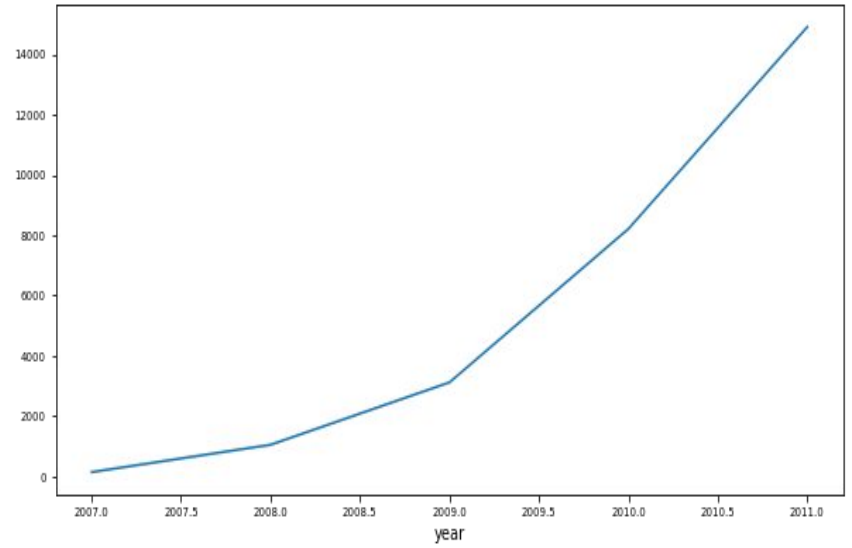


All 4 purpose are are high in “high” debt to income ratio.

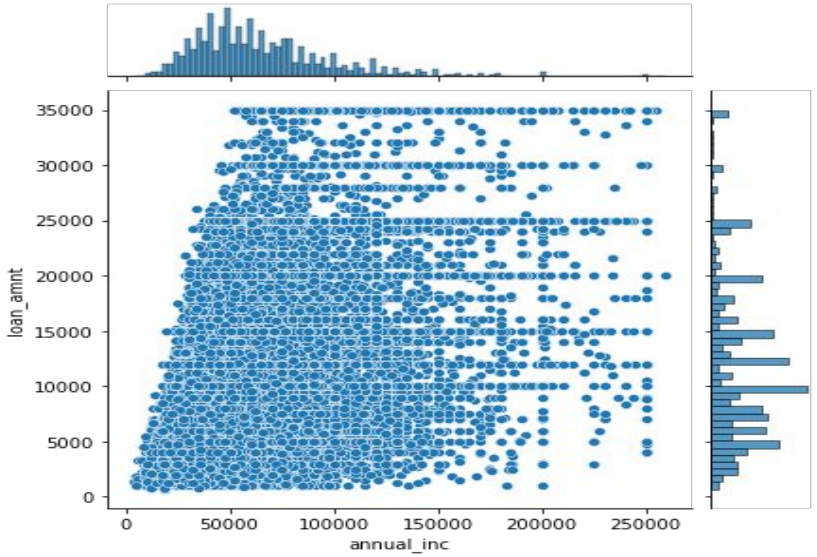


All 4 purpose are are high in “low” annual income.

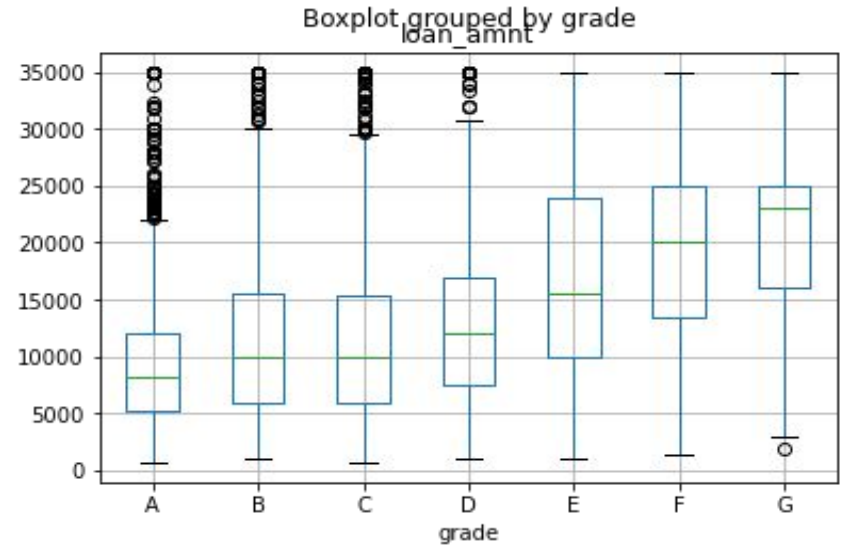
Bivariate Analysis



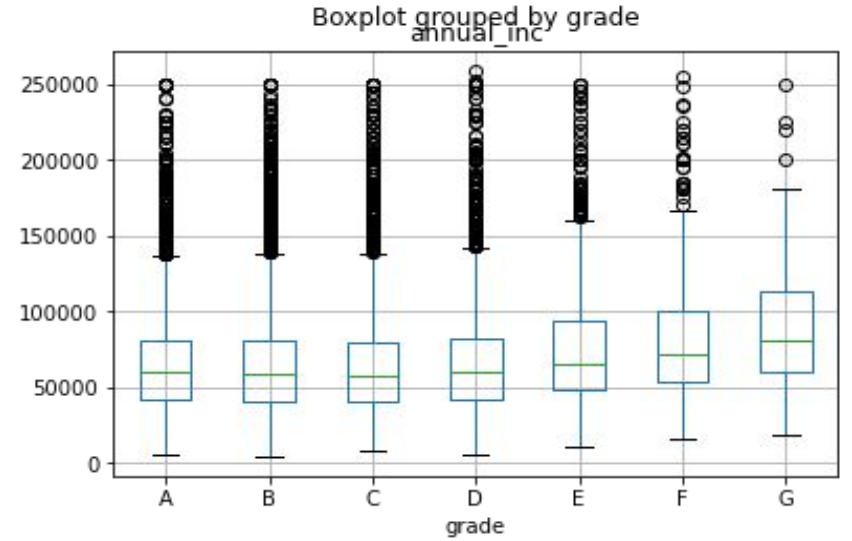
Average loan amount dropped sharply when subprime mortgage crisis hit



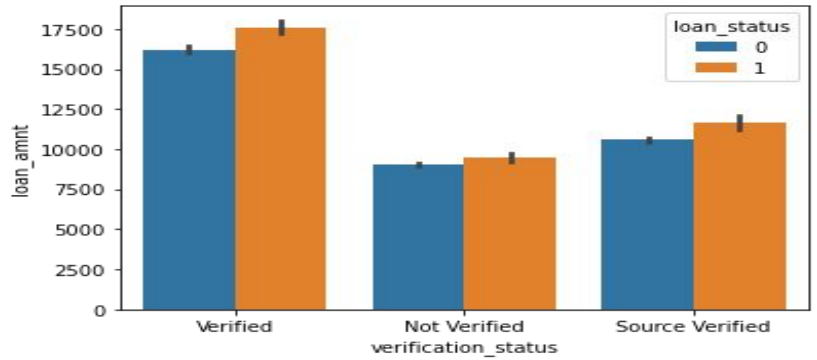
There are people with average income lower than 50000 taking loans of 25000 or higher. These would be risky loans.



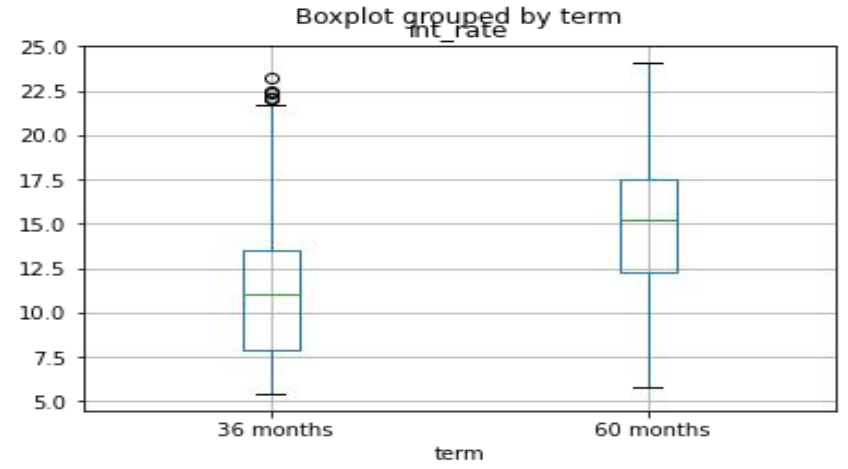
Larger loans generally appear to be given a lower grade, with the median loan amount for a grade G loan being almost 10000 higher than that of a grade A, B, or C loan.



Median grows from grade A to G gradually.

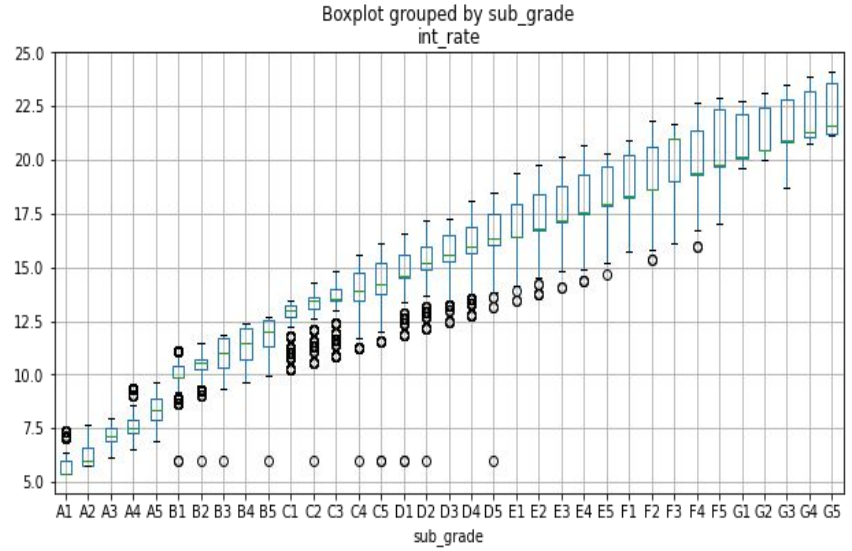


Higher loan amounts are Verified more often.
We already know that larger loans are less in number, but see a higher charge off rate.
This, combined with previous observation, explains why verified loans see a higher rate of default. It's not the verified status per se, it's the fact that higher loan amounts are riskier and are also verified more often by Lending Club.

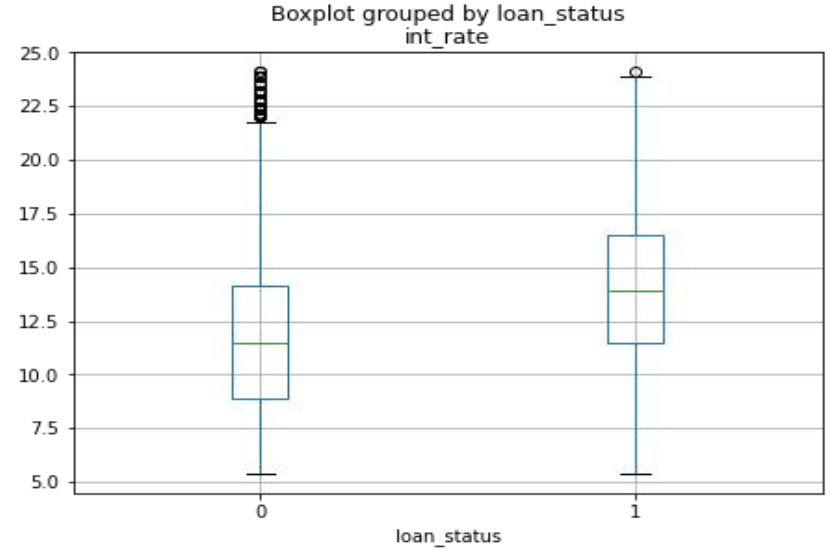


Interest rates are biased on term. Larger amounts were seen to be given for higher term. The rate of interest associated with them is also high.

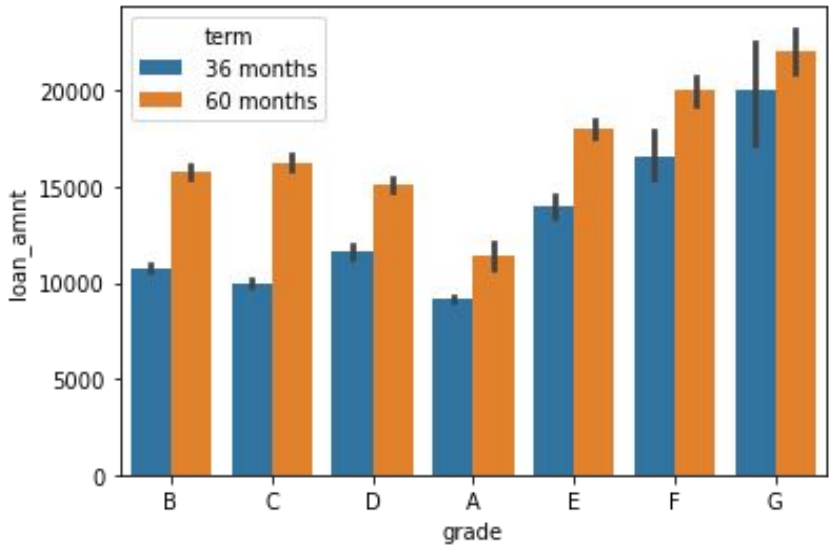
Bivariate Analysis



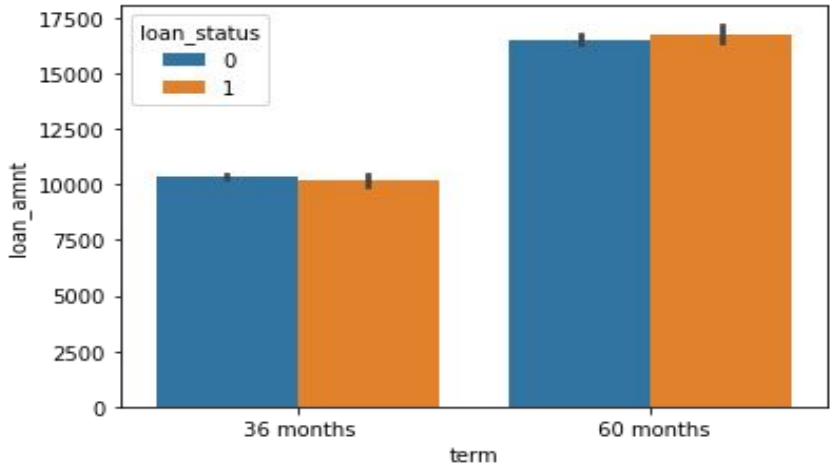
Interest rates varies directly with the subgrade. Larger or worst the sub grade, higher are the rate of interest for the loan.



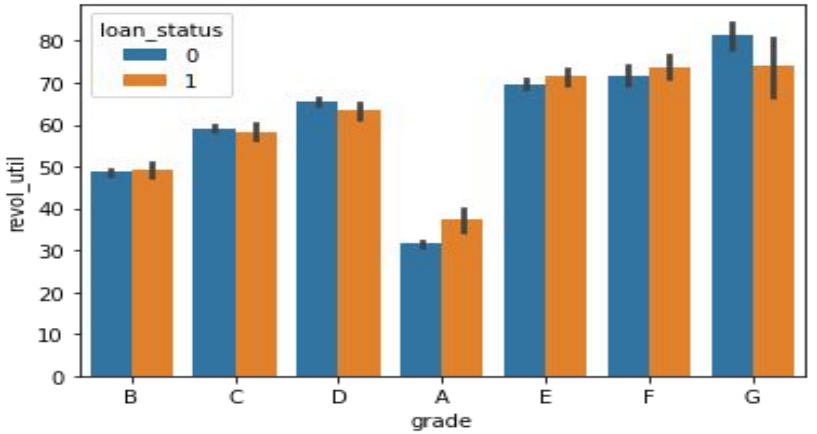
Loans at a higher interest rate are more likely to be Charged Off.



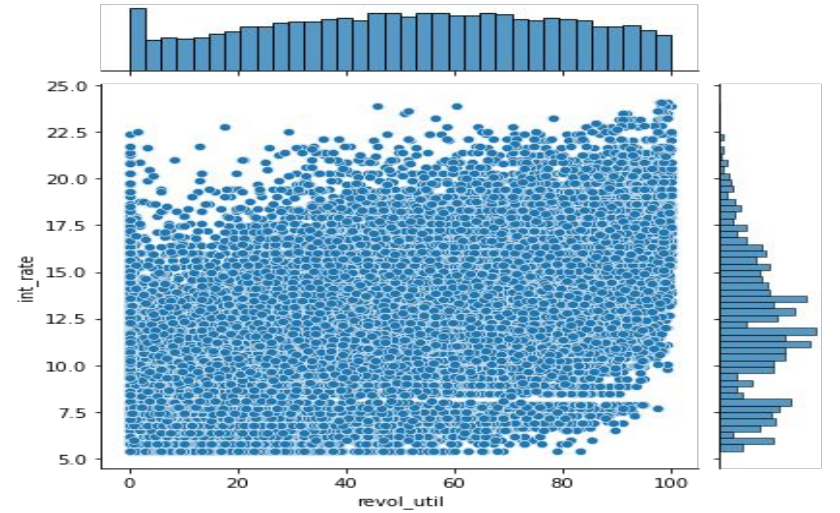
Our assumption made during univariate analysis is more evident with this plot. Higher loan amount are associated with lower grade for longer terms.



Our assumption made during univariate analysis is more evident with this plot. Higher loan amount are associated with longer terms and see higher Charge Offs.

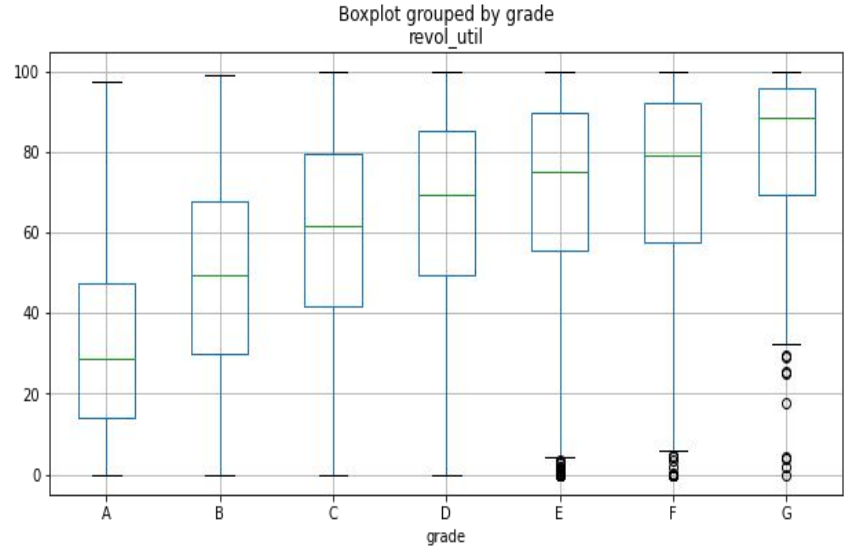


revol_util and grade(and therefore int_rate) are correlated in some way. The revol_util is positively correlated to the grade. As the grade goes from A to E the revol_util also increases. This may be because higher loan amounts are associated with higher grades.

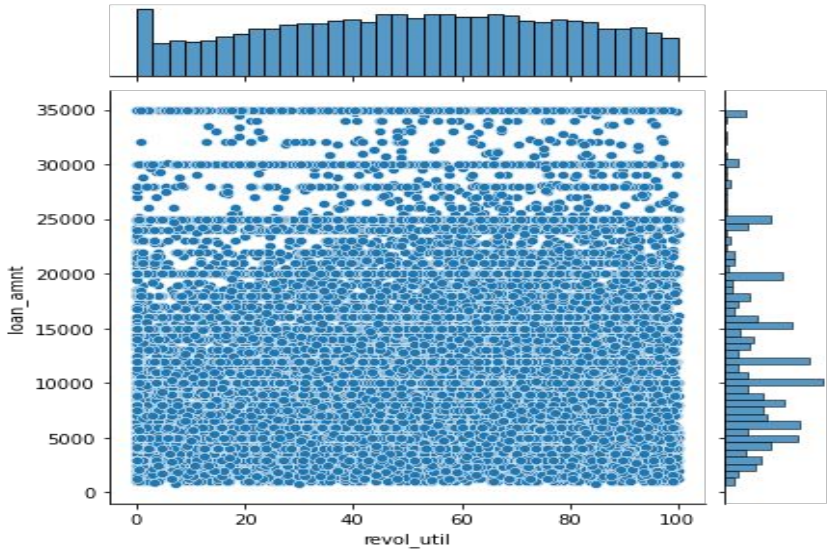


Utilization rate for loan amount is maximum for medium interest rates.

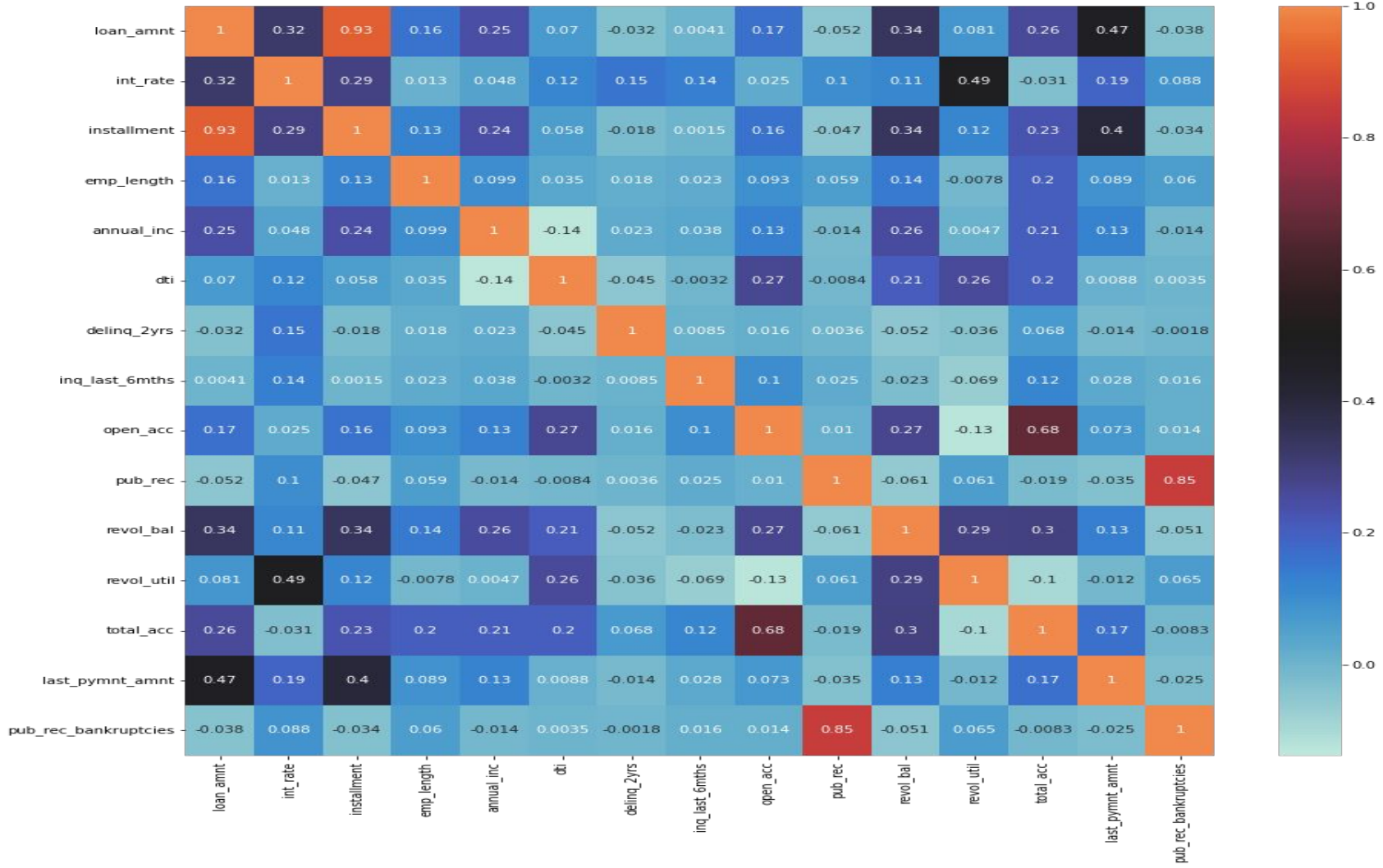
Bivariate Analysis



Median for revol_util increases from grade A to G.

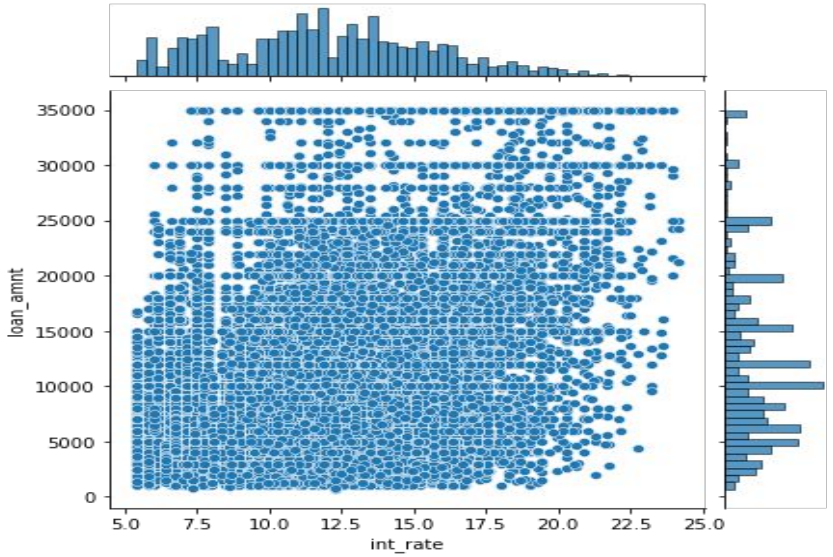


Utilization rate for loan amount is higher for medium loan amount.

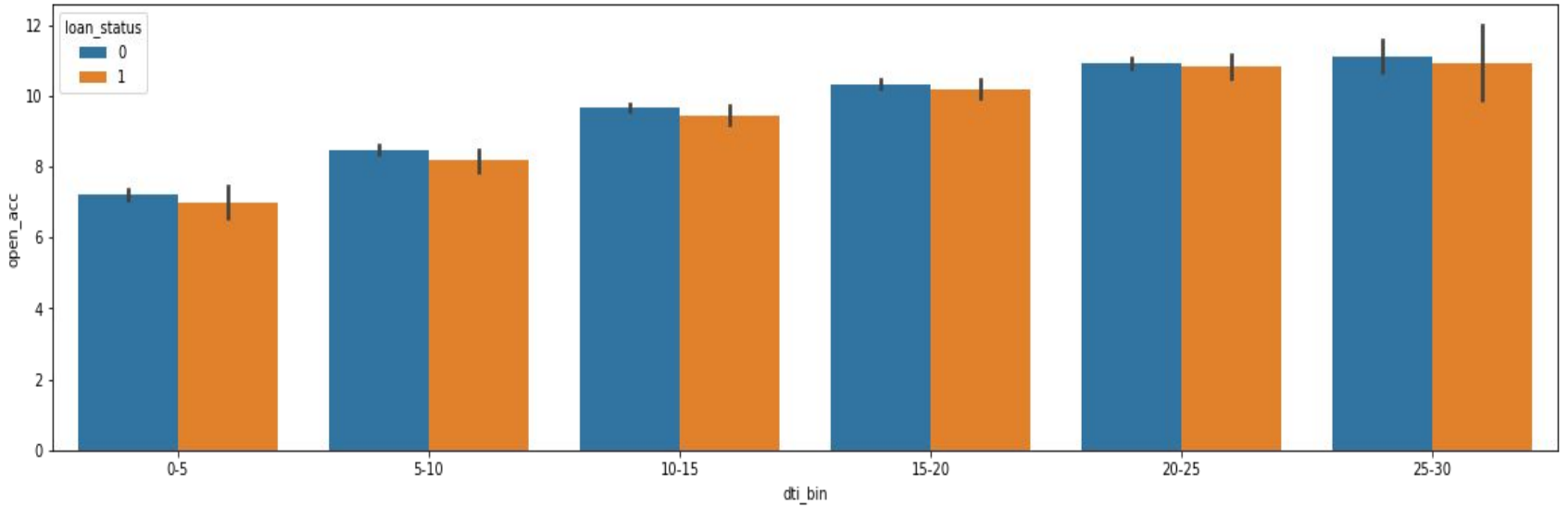


loan_amnt is correlated to last_payment_amount with r factor.44, as expected
int_rate is correlated to revol_util with r factor of .47 - This is good, as company is charging higher interest from riskier loan.
loan_amnt revol_bal are correlated with r factor .35 - This is not good as it suggests that higher loan amount is being approved to riskier borrowers.
delinq_2yrs is totally un-correlated with public record of bankruptcy. Therefore they represent distinct features with individual predictive value.

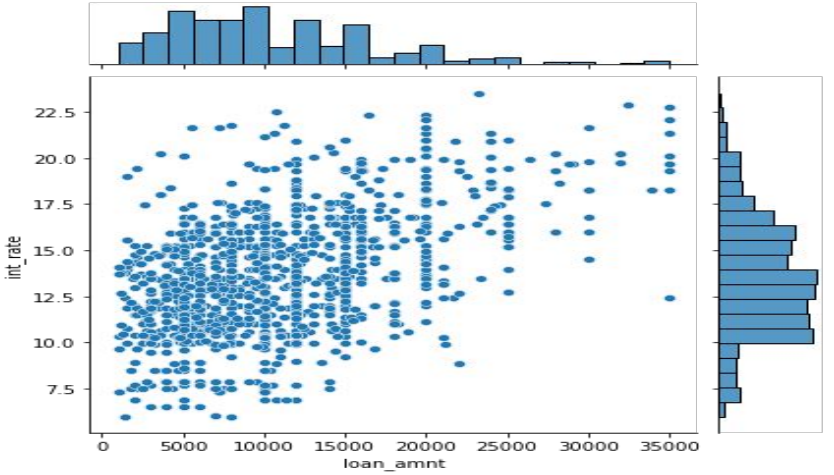
Bivariate Analysis



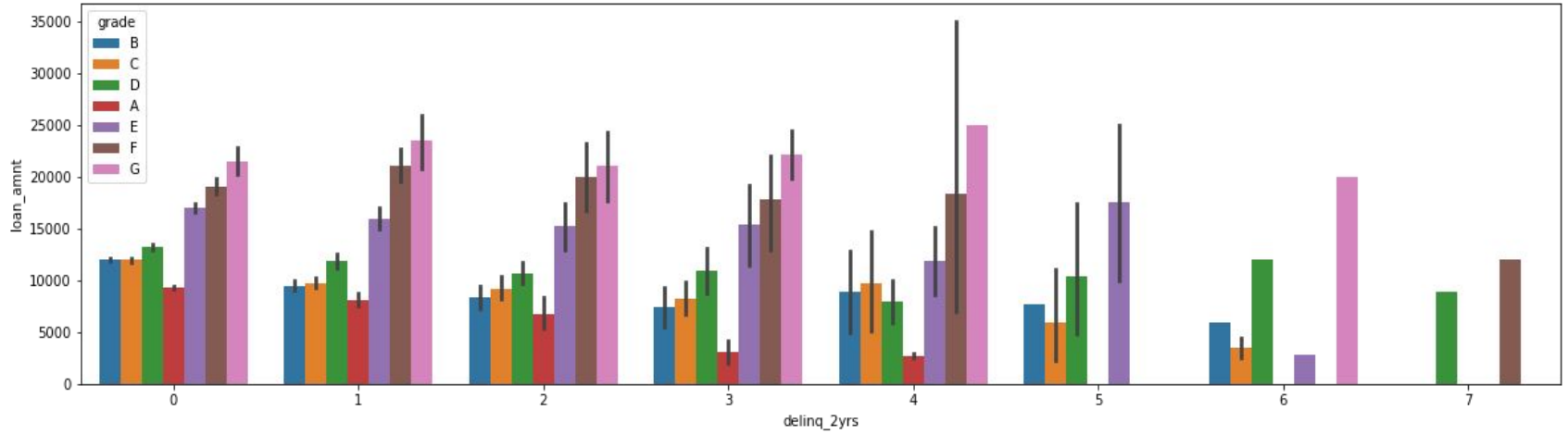
Interest rates varies directly with the subgrade. Larger or worst the sub grade, higher are the rate of interest for the loan.



Fully paid has slightly more open account than charged off across dti_bin.

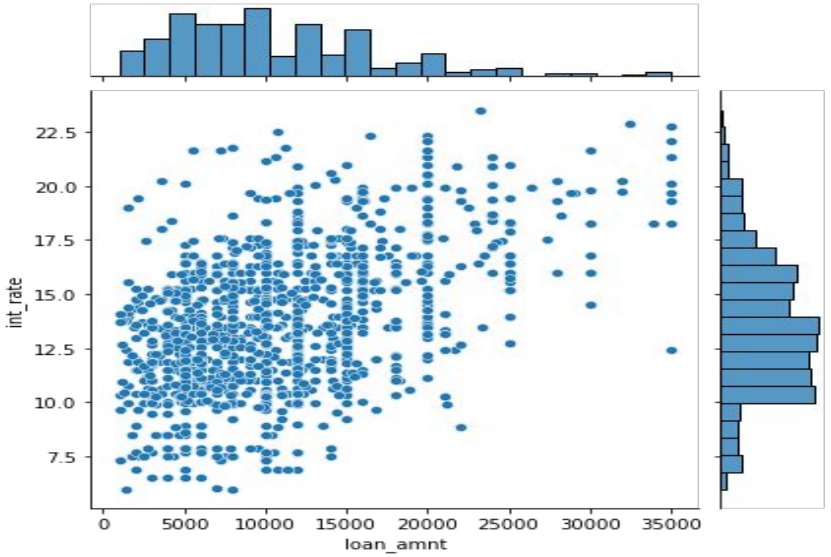


Interest Rate increases proportionally to Loan Amount for pub_rec_bankruptcies.

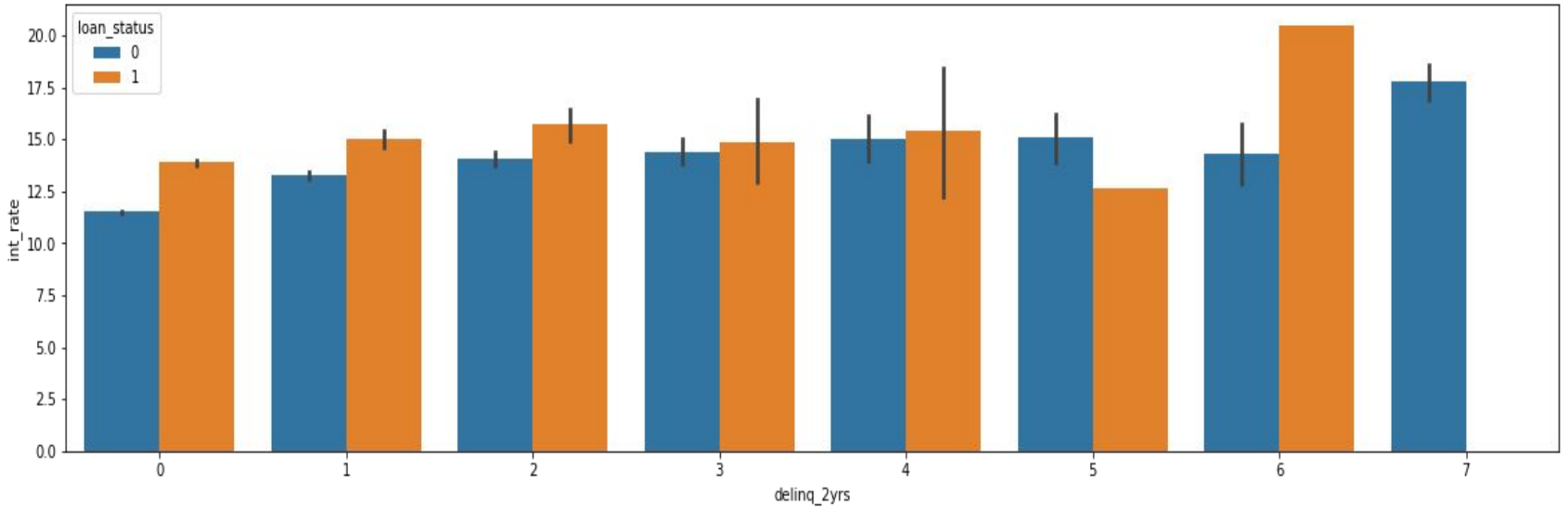


Grade E, F, G has relatively higher loan amount than A. B, C, D across delinq_2yrs.

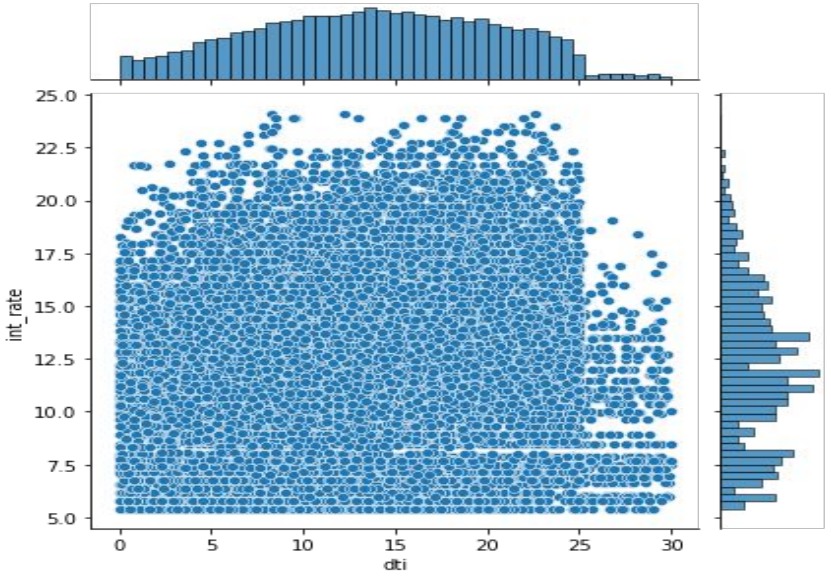
Bivariate Analysis



Interest Rate increases proportionally to Loan Amount for pub_rec.

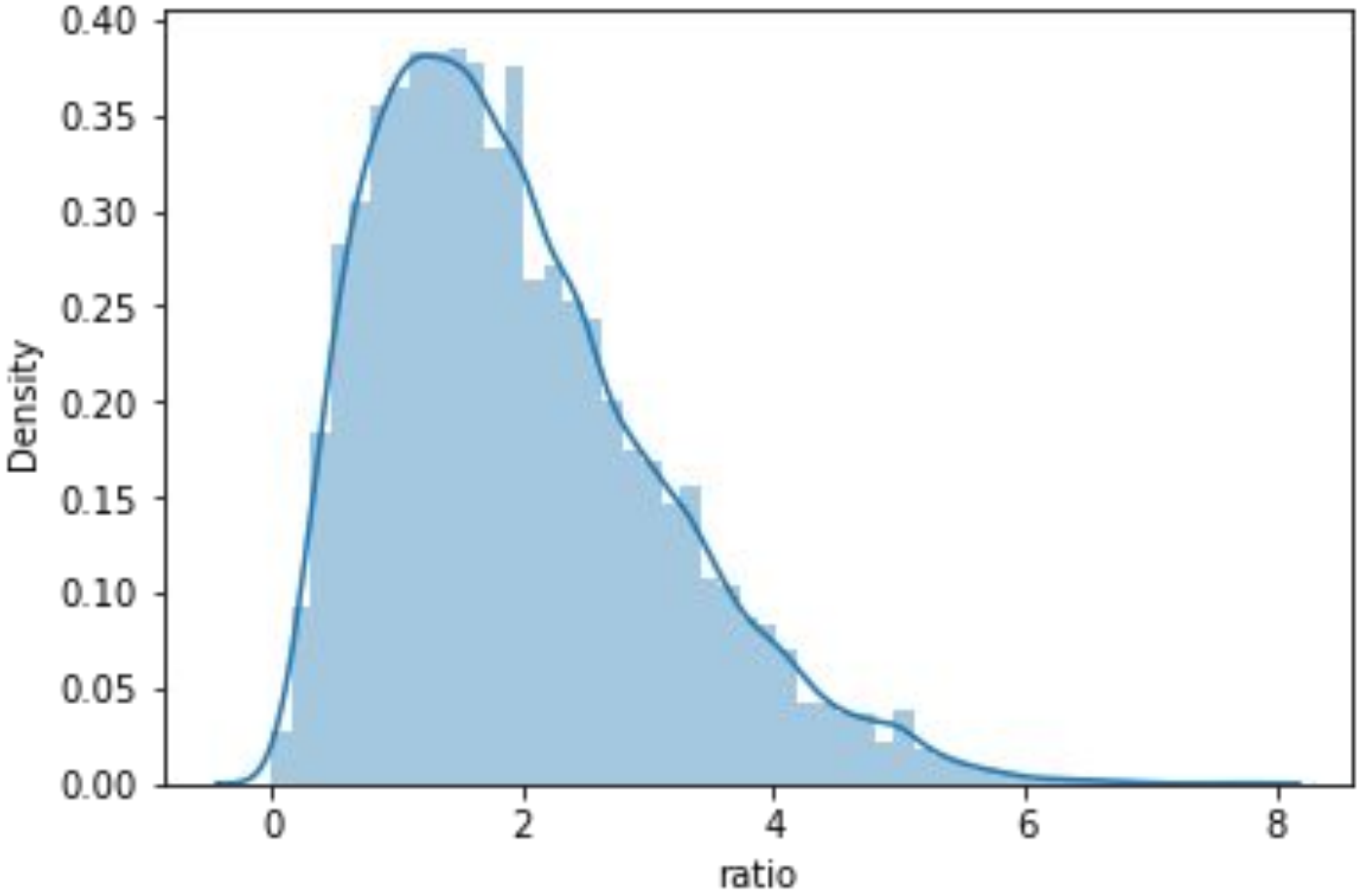


Interest Rate is higher for Charged Off than Fully Paid across delinq_2yrs except for 5.



Sharp drop in density of interest rate post 25 debt to interest ratio

Derived Metrics



Ratio of loan amount to annual income.
Maximum occurs where loan amount is equal or twice of annual income
than it declines sharply.

Results Summary

Univariate Analysis of Categorical Variables

1. Default Rate increases from grade A to G.
2. Default Rate increases with sub grade rank across all grades.
3. Default Rate is high in verified loans.
4. Default Rate is max in small businesses.
5. Maximum Loan approvals happens in December.
6. 2007 & 2009 have max and min loan_status.
7. Default Rate is highest in small businesses.

Univariate Analysis of Continuous Variables

1. Default Rate is high for “very high” loan amount.
2. Default Rate is high for “very high” funded amount inv.
3. Default Rate is high for “high” interest rate.
4. Default Rate is high for “high” debt to income ratios..
5. Default Rate is high for “very high” count of installment.
6. Default Rate is high for “low” annual income.
7. Default Rate is high for “high” funded amount.
8. Default Rate is high for experienced.

Segmented Univariate Analysis

1. Default Rate is max in small businesses.
2. 4 major categories are mentioned above.
   Count of loan disbursed is max for debt consolidation.
3. Debt consolidation is highest among employees experience of all years except 1 year.
4. Debt consolidation is still highest either in 36 or 60 months term.
5. Default Rate is highest for debt consolidation is highest among all grades.
6. Debt consolidation is highest among all home_owneship except others.
7. Debt consolidation is highest among employees experience of all years except 1 year.
8. Debt consolidation is highest among all years except 2007.
9. All 4 purpose are are high in “very high” loan amount.
10. All 4 purpose are are high in high” interest rate..
11. Installment also has debt consolidation as the highest factor.
12. All 4 purpose are are high in “high” debt to income ratio.
13. All 4 purpose are are high in “low” annual income.

Results Summary

Bivariate Analysis

1. Average loan amount dropped sharply when subprime mortgage crisis hit
2. There are people with average income lower than 50000 taking loans of 25000 or higher. These would be risky loans.
3. Larger loans generally appear to be given a lower grade, with the median loan amount for a grade G loan being almost 10000 higher than that of a grade A, B, or C loan.
4. Median grows from grade A to G gradually.
5. Higher loan amounts are Verified more often.
   We already know that larger loans are less in number, but see a higher charge off rate.
   This, combined with previous observation, explains why verified loans see a higher rate of default. It's not the verified status per se, it's the fact that higher loan amounts are riskier and are also verified more often by Lending Club.
6. Interest rates are biased on term. Larger amounts were seen to be given for higher term. The rate of interest associated with them is also high.
7. Interest rates varies directly with the subgrade. Larger or worst the sub grade, higher are the rate of interest for the loan.
8. Loans at a higher interest rate are more likely to be Charged Off.
9. Our assumption made during univariate analysis is more evident with this plot. Higher loan amount are associated with lower grade for longer terms.
10. Our assumption made during univariate analysis is more evident with this plot. Higher loan amount are associated with longer terms and see higher Charge Offs.
11. revol_util and grade(and therefore int_rate) are correlated in some way. The revol_util is positively correlated to the grade. As the grade goes from A to E the revol_util also increases. This may be because higher loan amounts are associated with higher grades.

Results Summary

Bivariate Analysis continued…

12. Utilization rate for loan amount is maximum for medium interest rates.
13. Median for revol_util increases from grade A to G.
14. Utilization rate for loan amount is higher for medium loan amount.
15. loan_amnt is correlated to last_payment_amount with r factor.44, as expected
    int_rate is correlated to revol_util with r factor of .47 - This is good, as company is charging higher interest from riskier loan.
    loan_amnt revol_bal are correlated with r factor .35 - This is not good as it suggests that higher loan amount is being approved to riskier borrowers.
    delinq_2yrs is totally un-correlated with public record of bankruptcy. Therefore they represent distinct features with individual predictive value.
16. Interest rates varies directly with the subgrade. Larger or worst the sub grade, higher are the rate of interest for the loan.
17. Fully paid has slightly more open account than charged off across dti_bin.
18. Interest Rate increases proportionally to Loan Amount for pub_rec_bankruptcies.
19. Grade E, F, G has relatively higher loan amount than A. B, C, D across delinq_2yrs.
20. Interest Rate increases proportionally to Loan Amount for pub_rec.
21. Interest Rate is higher for Charged Off than Fully Paid across delinq_2yrs except for 5.
22. Sharp drop in density of interest rate post 25 debt to interest ratio.

Recommendations

Highest default rate is for debt consolidation which in turn increase the overall debt exposure of individual.
Hence, suggesting to settle an amount and charge off the loan in order to not increase the amount further.

1. Loan approval can be reduced for small business purposes.

2. Loan approval can be reduced for people with prior bad record or at least stopped for high-value loans.

3. Loan approval can be reduced where amount/income is higher than 30%.

4. High Value Loan approval can be reduced when revolving line utilization rate greater than 75%.

5. Start–charging higher interest rates for loans with dti greater than 20.
   Higher interest rate will further led to increase in debt.
   Hence, highly correlated to default rates.

THANK YOU!!