LENDING ASSIGNMENT

SUBMISSION

Name:

1. Tina Shrivastava – Group Facilitator
2. Ravi Kumar

Overall Methodology

UpGrad

Key Objectives	Business Understanding	Approach	Key Takeaways	Deliverables
• To understand the **driving factors** behind loan defaults. • Utilize this knowledge for its portfolio and risk assessment to **minimize credit loss and business loss.**	• Consumer finance company tending to urban customers. • **Risks** 1. Loss of business when consumer likely to repay and loan not approved. 2. Loss to business if consumer not likely to repay and loan approved.	1. Data Loading 2. Data cleaning 3. Data Modification 4. Univariate analysis 5. Bivariate Analysis 6. Segment Analysis 7. Derived Metrics Analysis 8. Correlation Analysis	• Understand the consumer and loan attributes which influence the tendency of default. • Recommend important driver variables	One zip file containing: 1. Jupyter Notebook 2. Presentation in the PDF format.

Program Solving Approach

Data Cleansing

- ✓ Handling NaN
- ✓ Handling zero's (irrelevant to analysis)
- ✓ Handling outliers

Data Modification

- ✓ Removing symbols from numeric values
- ✓ Converting date string to proper date format.
- ✓ Extracting year from relevant columns

Data Analysis

Exploratory data analysis is related with gaining insights from the data presented.

- ✓ Univariate analysis of both categorical and continuous variables
- ✓ Bivariate Analysis of relevant variables to gain insight about the defaulting loan

Plotting

Using matplotlib and seaborn to plot the graph related with the data.

This helps in

- ✓ Finding and fixing data related issues.
- ✓ Univariate and Bivariate analysis

It further helps in communicating the interferences with the bigger audience.

Continuous Variables - Summary

	count	mean	std	min	10%	25%	50%	75%	90%	max
loan_amnt	34400.00	11068.73	7068.25	500.00	3500.00	5800.00	10000.00	15000.00	20675.00	35000.00
funded_amnt	34400.00	10809.26	6818.27	500.00	3500.00	5600.00	9800.00	15000.00	20000.00	35000.00
funded_amnt_inv	34400.00	10253.96	6760.96	0.00	3000.00	5000.00	9000.00	14000.00	19925.00	35000.00
int_rate	34400.00	11.90	3.69	5.42	6.99	8.94	11.71	14.35	16.77	24.40
installment	34400.00	322.55	199.73	16.08	104.46	171.43	283.20	423.11	603.14	1305.19
emp_length	34400.00	5.02	3.44	0.00	1.00	2.00	4.00	9.00	10.00	10.00
annual_inc	34400.00	63561.94	26508.72	24044.00	33600.00	42500.00	59278.00	80000.00	102000.00	141996.00
dti	34400.00	13.52	6.63	0.00	4.28	8.47	13.68	18.75	22.40	29.99
delinq_2yrs	34400.00	0.15	0.49	0.00	0.00	0.00	0.00	0.00	1.00	11.00
inq_last_6mths	34400.00	0.87	1.07	0.00	0.00	0.00	1.00	1.00	2.00	8.00
open_acc	34400.00	9.35	4.34	2.00	4.00	6.00	9.00	12.00	15.00	44.00
revol_bal	34400.00	12855.33	14031.09	0.00	1241.90	3930.00	9028.50	16811.50	27751.00	149000.00
revol_util	34373.00	48.97	28.17	0.00	8.80	25.80	49.50	72.30	87.70	99.90
total_acc	34400.00	22.25	11.15	2.00	9.00	14.00	21.00	29.00	37.00	90.00
total_pymnt	34400.00	11868.57	8466.41	0.00	3122.18	5681.73	9933.14	16031.05	23680.40	58563.68
total_pymnt_inv	34400.00	11280.82	8370.49	0.00	2604.58	5261.06	9316.13	15302.52	22833.33	58563.68

- ✓ After removing all the irrelevant data, 37 columns were left for analysis.
- ✓ Also, 2 columns (addr_state, zip_code) could be used for demographic analysis.
- ✓ To identify the outliers, boxplot was used. After identifying those, proper percentile was used to remove them.

Continuous Variables - Summary

	count	mean	std	min	10%	25%	50%	75%	90%	max
total_rec_prncp	34400.00	9653.36	6738.92	0.00	2400.00	4800.00	8000.00	13000.00	20000.00	35000.02
total_rec_int	34400.00	2116.37	2317.37	0.00	327.76	669.76	1335.66	2676.35	4837.67	23563.68
recoveries	34400.00	97.56	696.06	0.00	0.00	0.00	0.00	0.00	14.49	29623.35
collection_recovery_fee	34400.00	12.35	148.05	0.00	0.00	0.00	0.00	0.00	0.00	7002.19
last_pymnt_amnt	34400.00	2745.23	4402.65	0.00	100.06	227.61	579.94	3545.99	8438.15	36115.20
pub_rec_bankruptcies	33808.00	0.04	0.21	0.00	0.00	0.00	0.00	0.00	0.00	2.00
issue_year	34400.00	2010.32	0.88	2007.00	2009.00	2010.00	2011.00	2011.00	2011.00	2011.00
loan_income_ratio	34400.00	0.19	0.11	0.01	0.06	0.10	0.17	0.25	0.35	0.82

Derived Fields:

- ✓ issue_year

  Derived by extracting the year from the issue_d

- ✓ loan_income_ratio

  Derived by dividing the loan (loan_amnt) requested by the customer by total annual income (annual_inc)

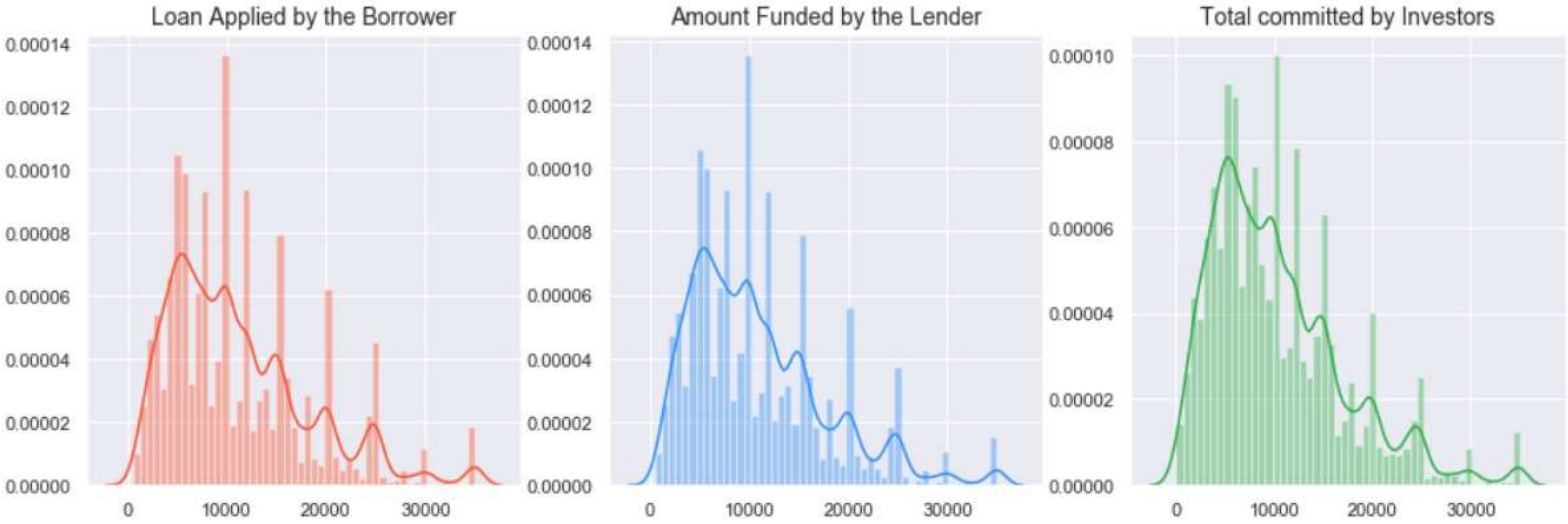
Important fields: The following variables are identified as important and for which univariate analysis was carried out:

1. loan_amnt
2. int_rate
3. annual_inc
4. dti
5. issue_d
6. emp_length
7. loan_income_ratio
8. issue_year

Continuous Variables – Key Graphs

UpGrad

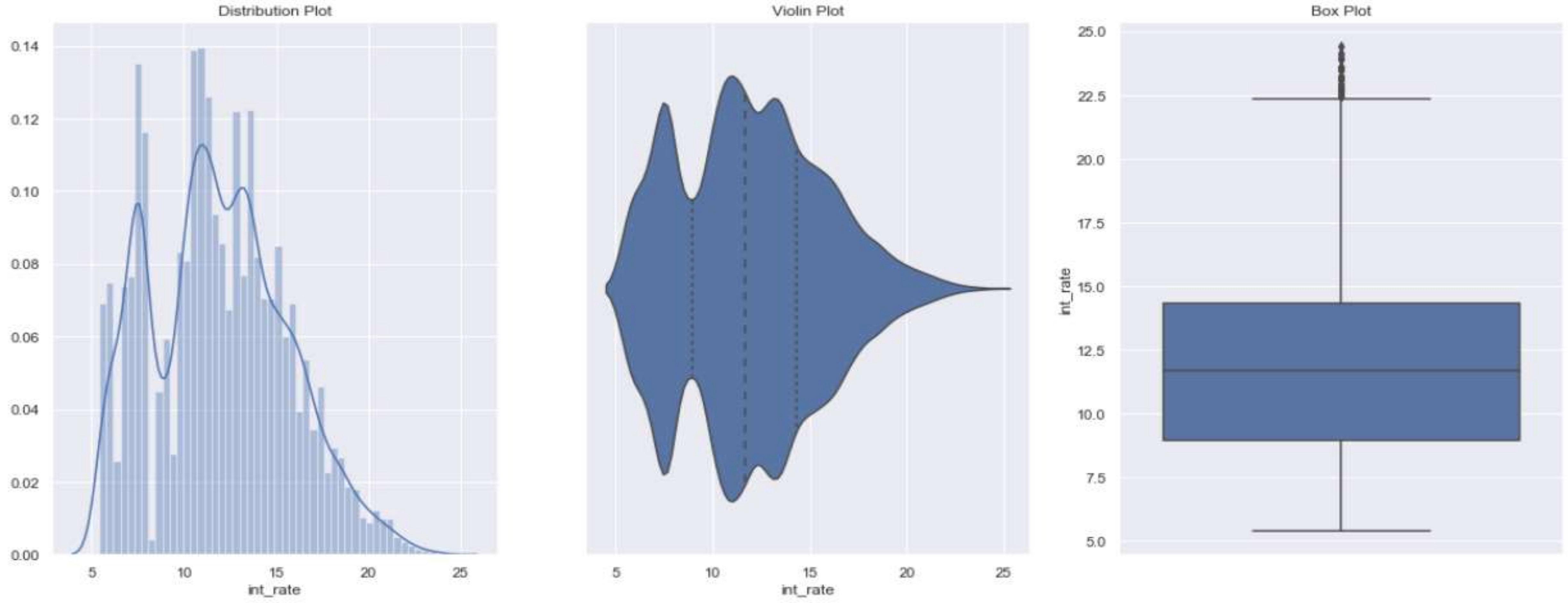
loan_amnt, funded_amnt, funded_amnt_inv



Observation : loan amount and funded amoount are same. In this case, funded amount can be removed from the data set

Continuous Variables – Key Graphs

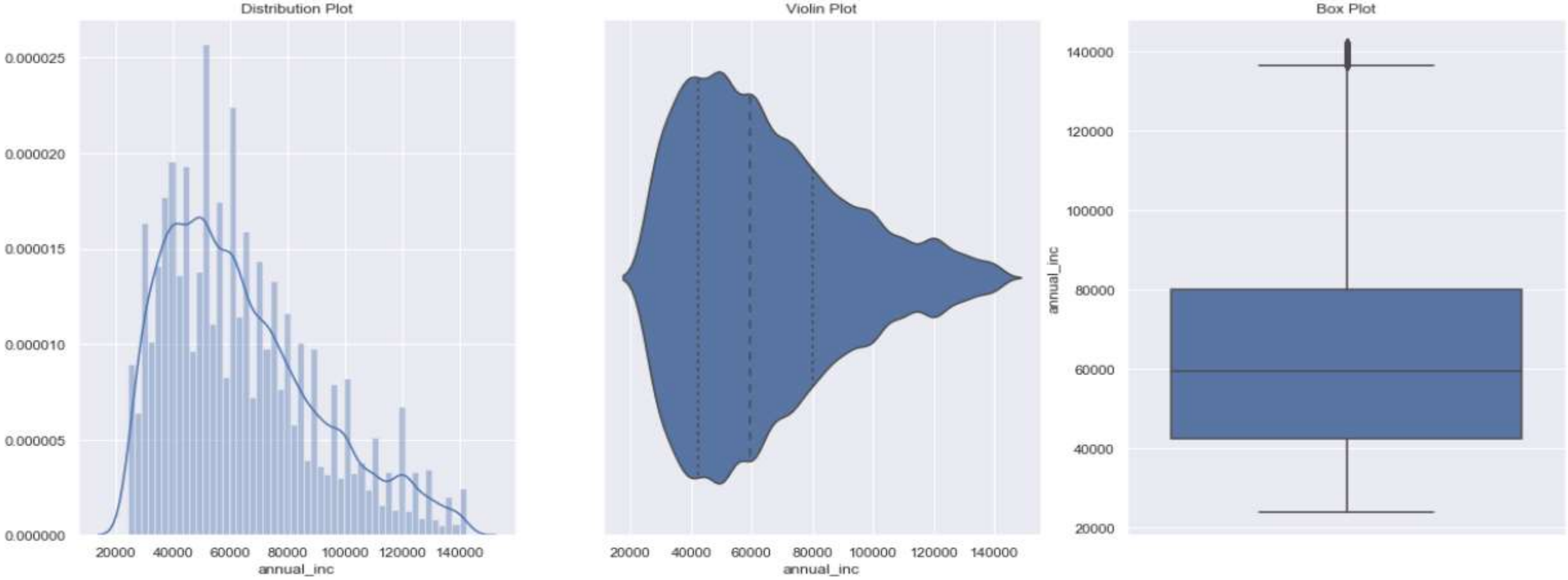
int_rate



Observation : Most of the interest rates are between 5 and 10% with a maximum interest rate of 24% approx.

Continuous Variables – Key Graphs

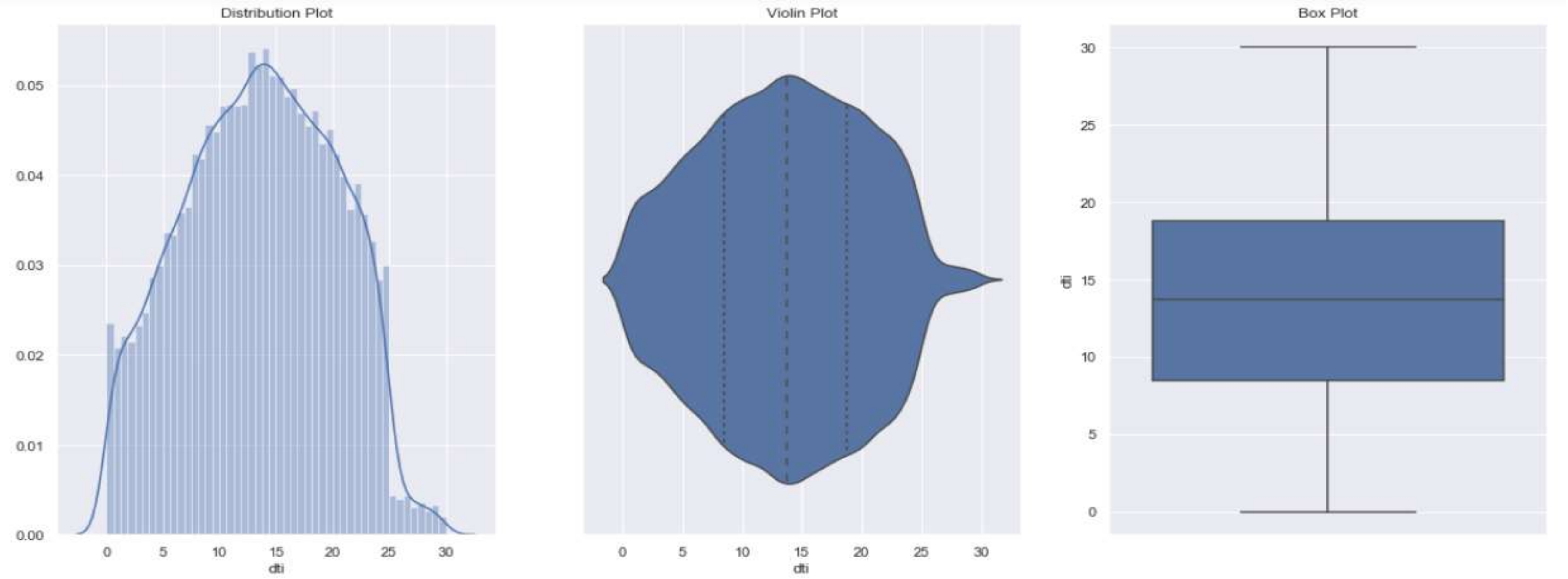
annual_inc



Observation : In 75% cases, the annual income of the client is less than 80000. There were extremely large annual income in the dataset which were more than 150 times than the mean annual income. These has been removed from the dataset

Continuous Variables – Key Graphs

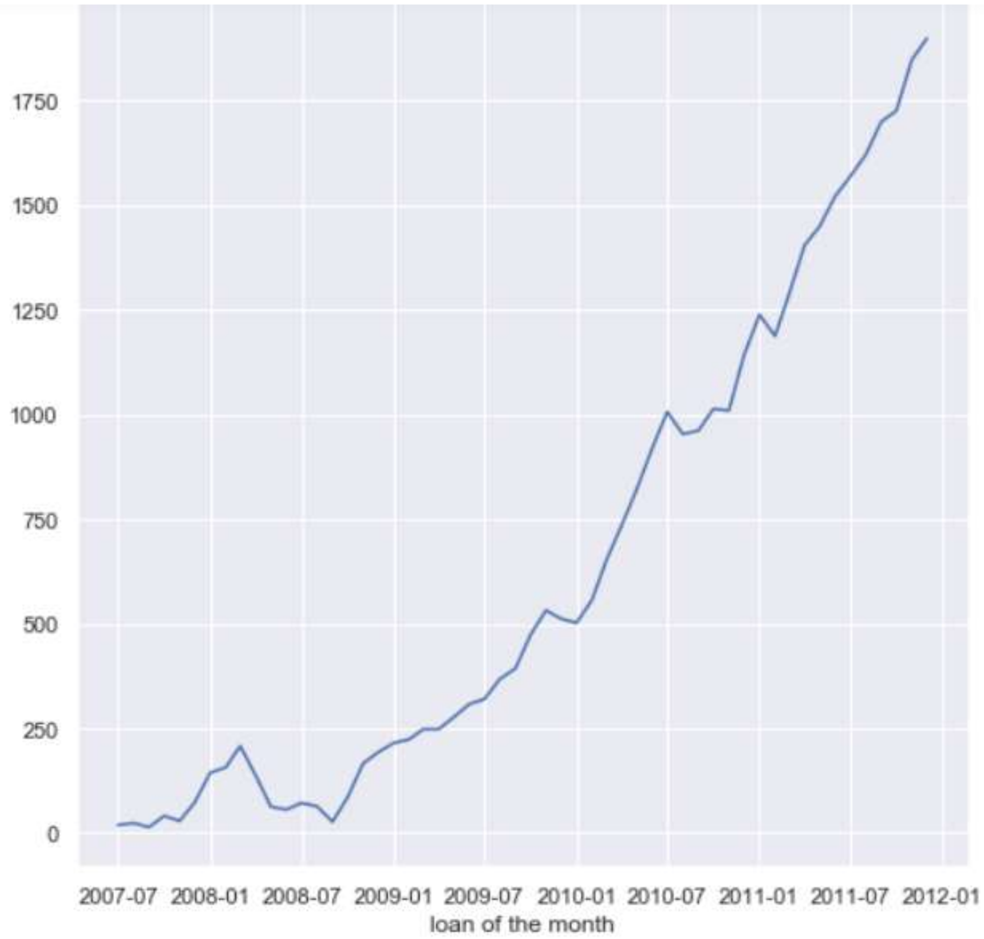
dti



Observation : dti is a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income

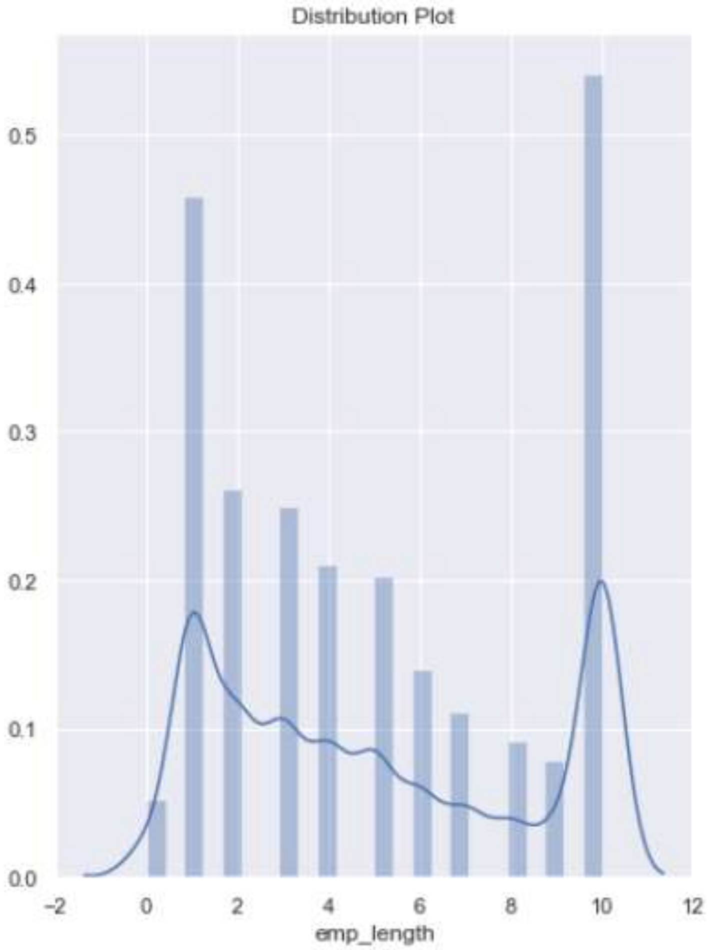
Continuous Variables – Key Graphs

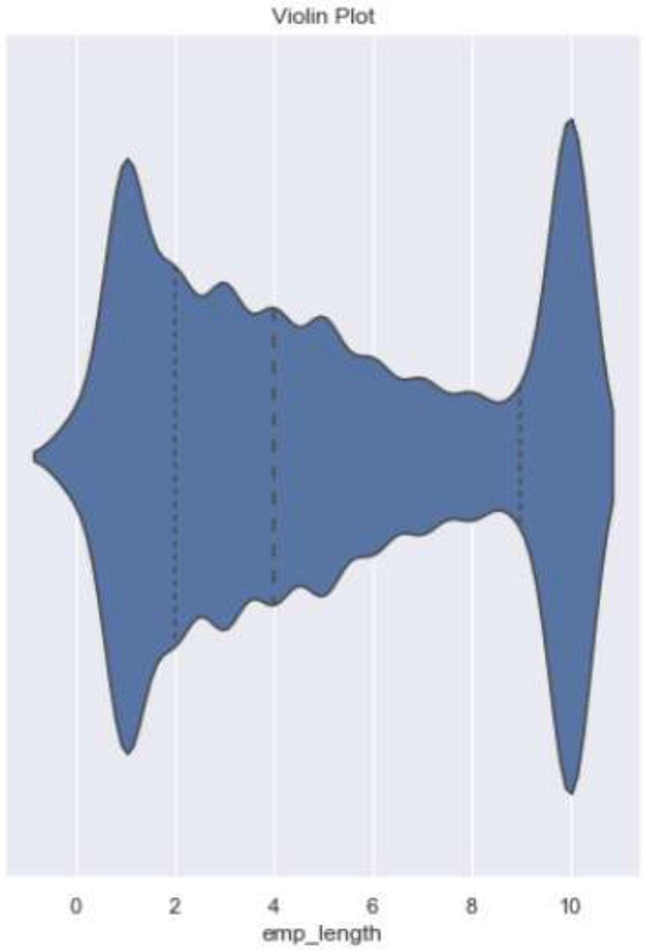
issue_d

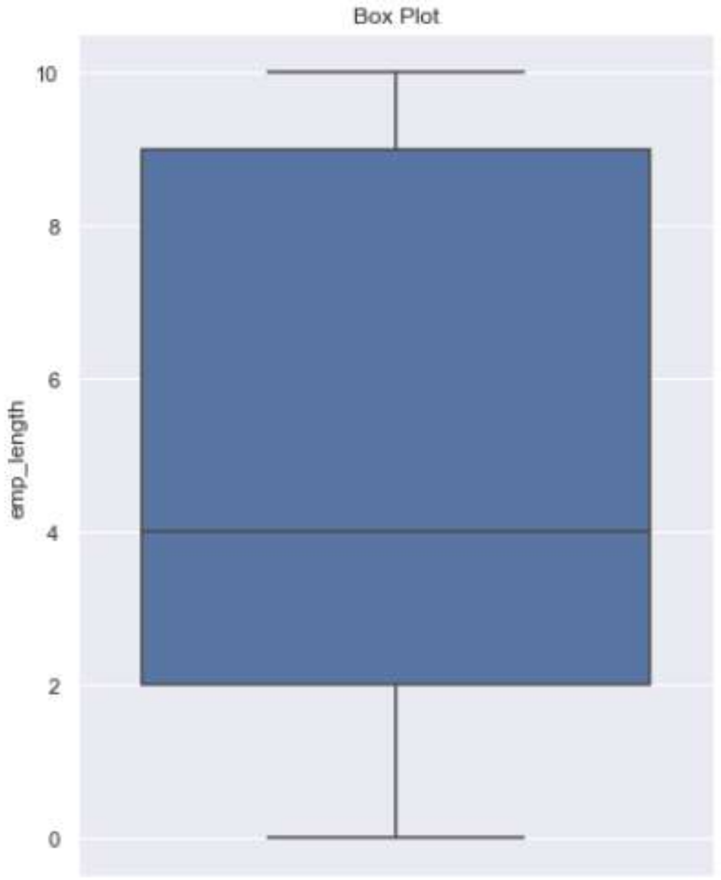


Observation : The highest number of loans has been taken in the month of January,2012. It seems the spending capacity of the people increased which in turn indicates economic growth

emp_length





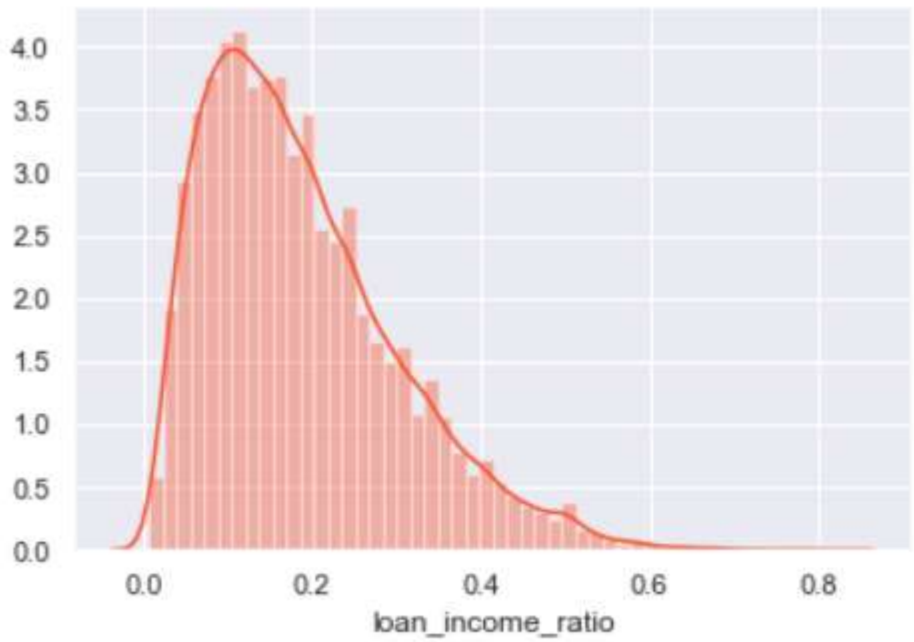


Observation : People tend to take more loans in their 4th year of employment whereas after a long employment tenure (10 years), the number of loans taken by them reduces

Continuous Variables – Key Graphs

UpGrad

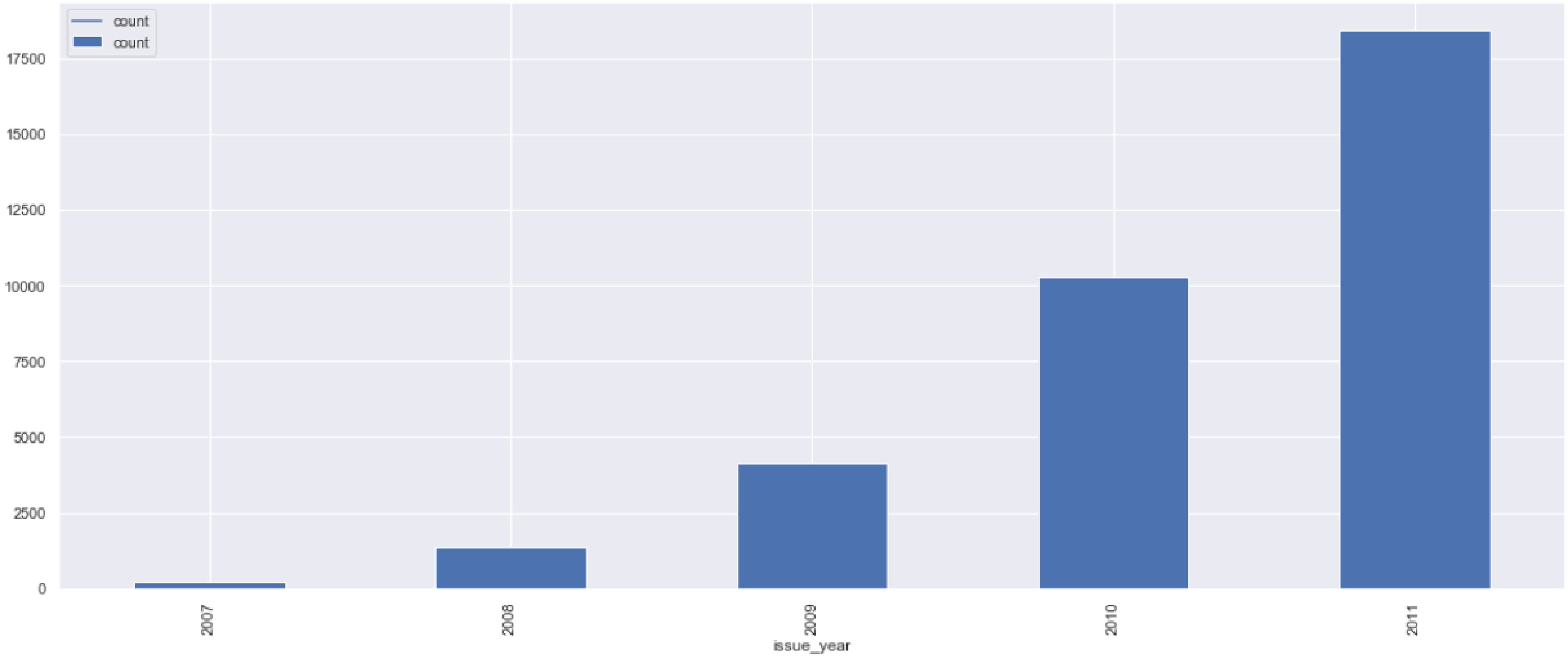
loan_income_ratio



Observation : On analyzing the loans, it seems that people tends to take around 1\5th of their annual income as loan. There can be a correlation between the loan income ratio and the loan status.

Continuous Variables – Key Graphs

issue_year



Observation : There is a gradual increase in the number of people taking loans from 2007 to 20011 with maximum number of people taking loans in the year 2011 for the duration of 2007-2011

Categorical Variables – Summary

Important fields: The following variables are identified as important categorical variables and for which univariate analysis was carried out:

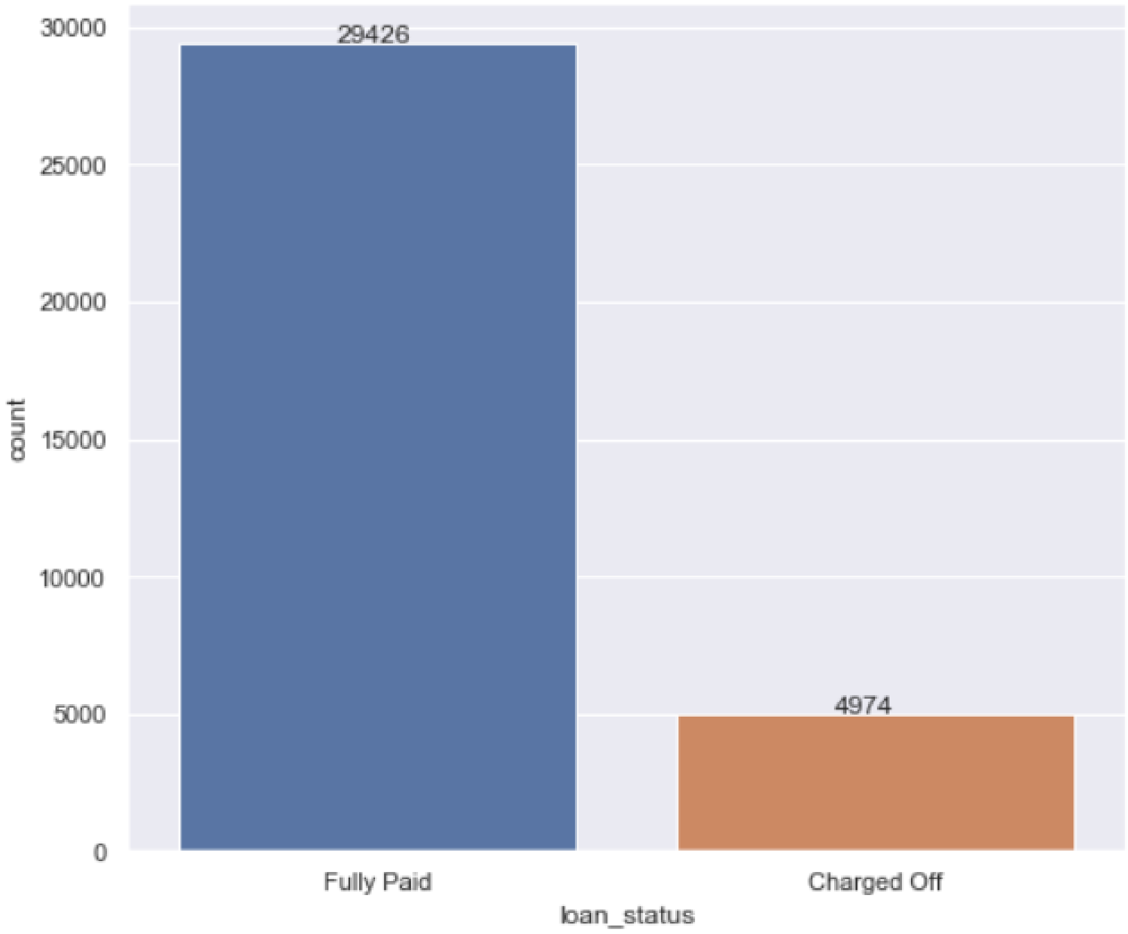
1. grade
2. sub_grade
3. home_ownership
4. verification_status
5. loan_status
6. purpose

	count	unique	top	freq
grade	34400	7	B	10385
sub_grade	34400	35	A4	2582
home_ownership	34400	3	RENT	16441
verification_status	34400	3	Not Verified	15100
loan_status	34400	2	Fully Paid	29426
purpose	34400	14	debt_consolidation	16446

Categorical Variables – Key Graph

UpGrad

loan_status

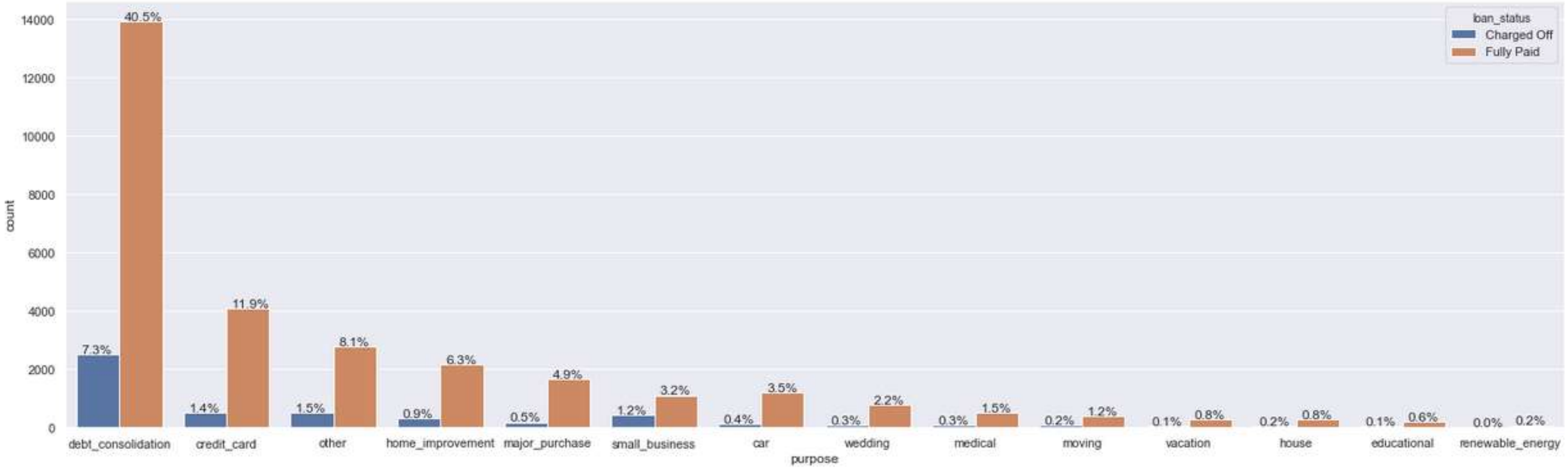


Observation : Around 14% of the loans have the status as charged off. This means that these loans have defaulted.

Categorical Variables – Key Graph

UpGrad

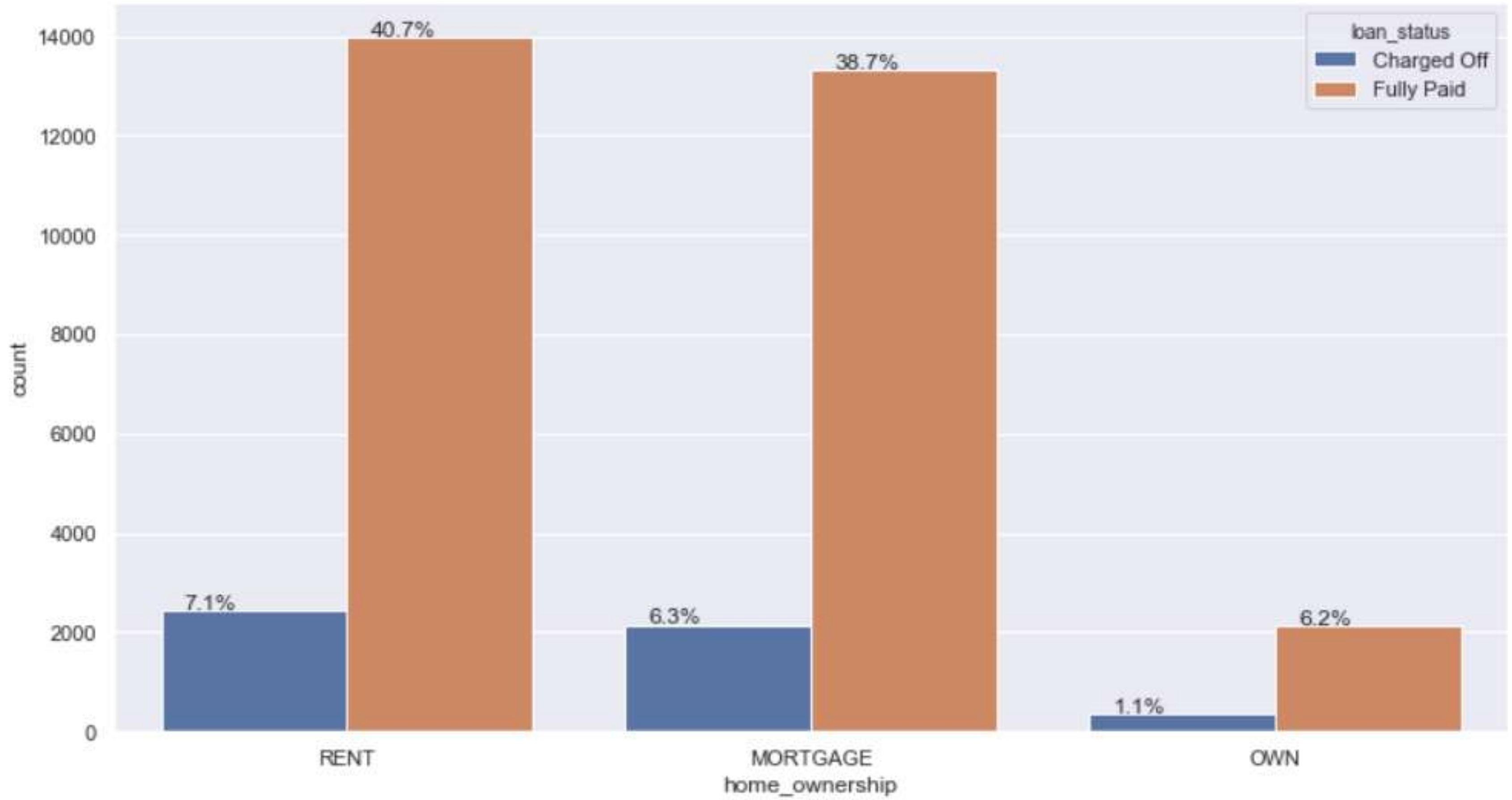
purpose



Observation : Approximately 40% of people took loan to settle thier other loans. And out of these people 7.3% defaulted their loan which is the highest among all other people who took for another reason. This means that there is a more tendency that around 7% people can default their loan if they have taken for repaying other loans

Categorical Variables – Key Graph

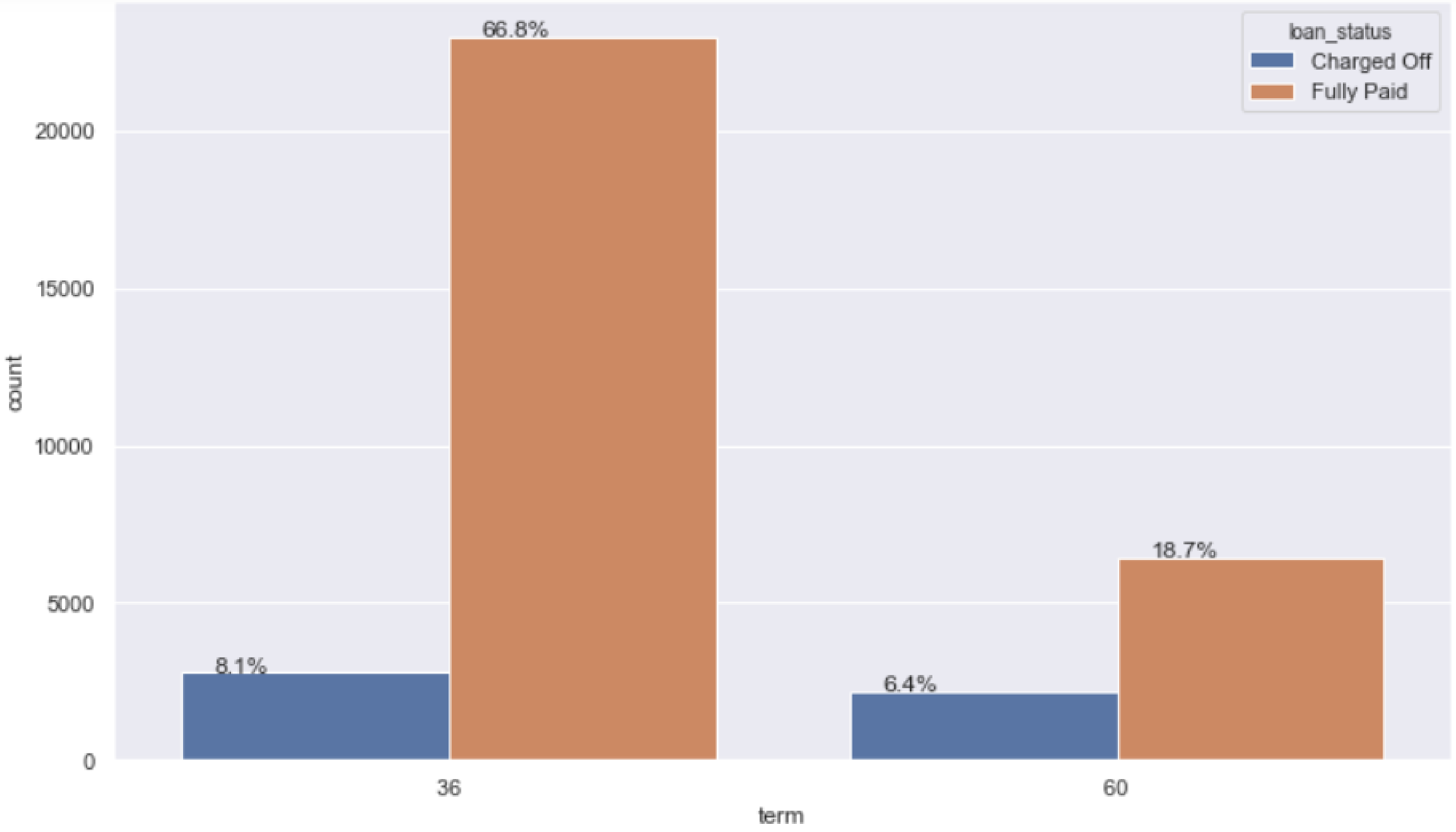
home_ownership



Observation : From the above chart it is clear that people with their own homes have defaulted significantly less as compared to people living on rent and morgage.

Categorical Variables – Key Graph

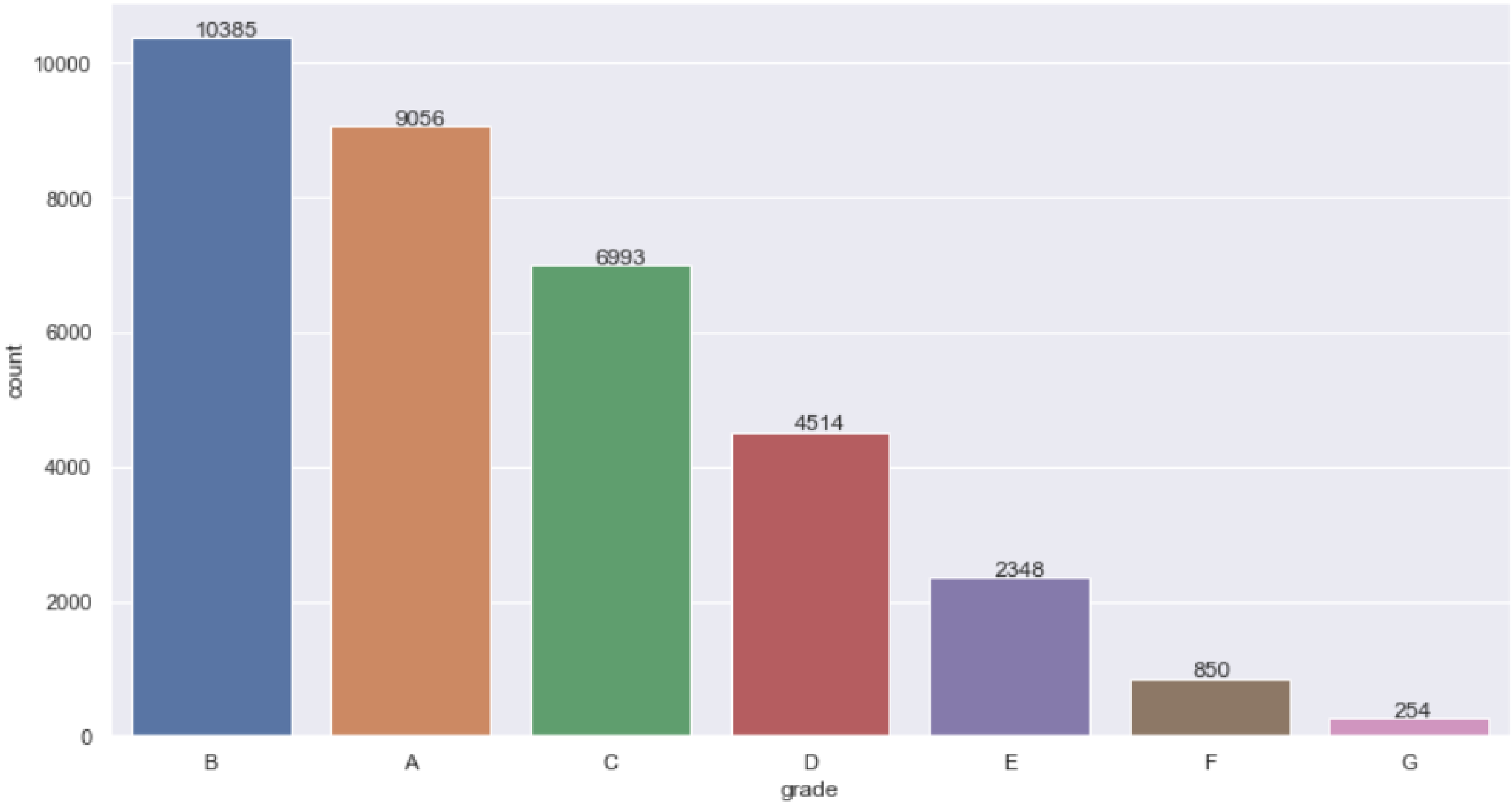
term



Observation : More people are inclined to take loans of shorter duration as compared to longer duration of 60 months. But there is not a significant difference in the number of people defaulting loan based on the tenure of the loans

Categorical Variables – Key Graph

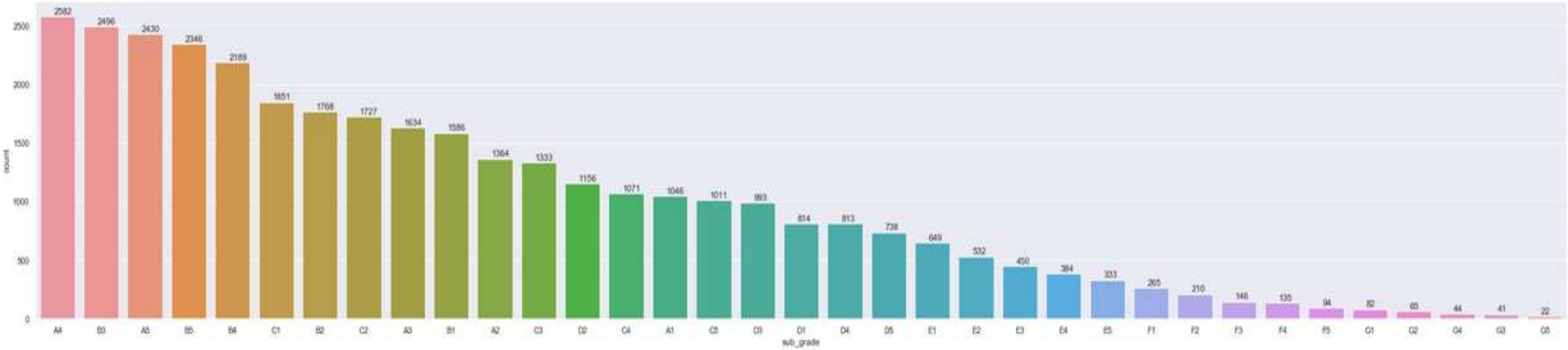
grade



Observation : As compared to the people which are rated higher (A Grade), people with low rating (B Grade) tend to take more loans

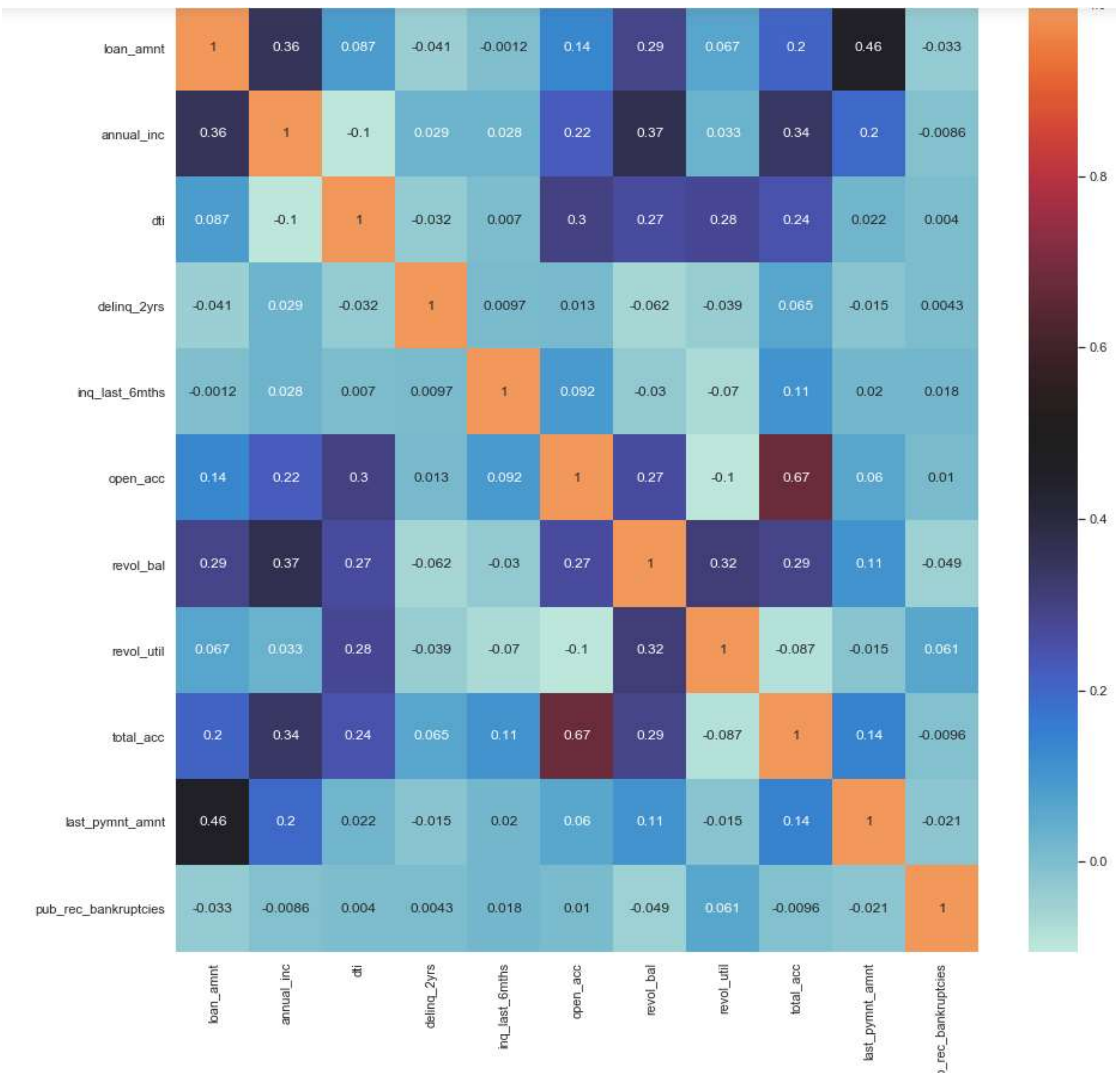
Categorical Variables – Key Graph

sub_grade



Observation : But on a deeper lookout, people with grade A4 tends to take more loans

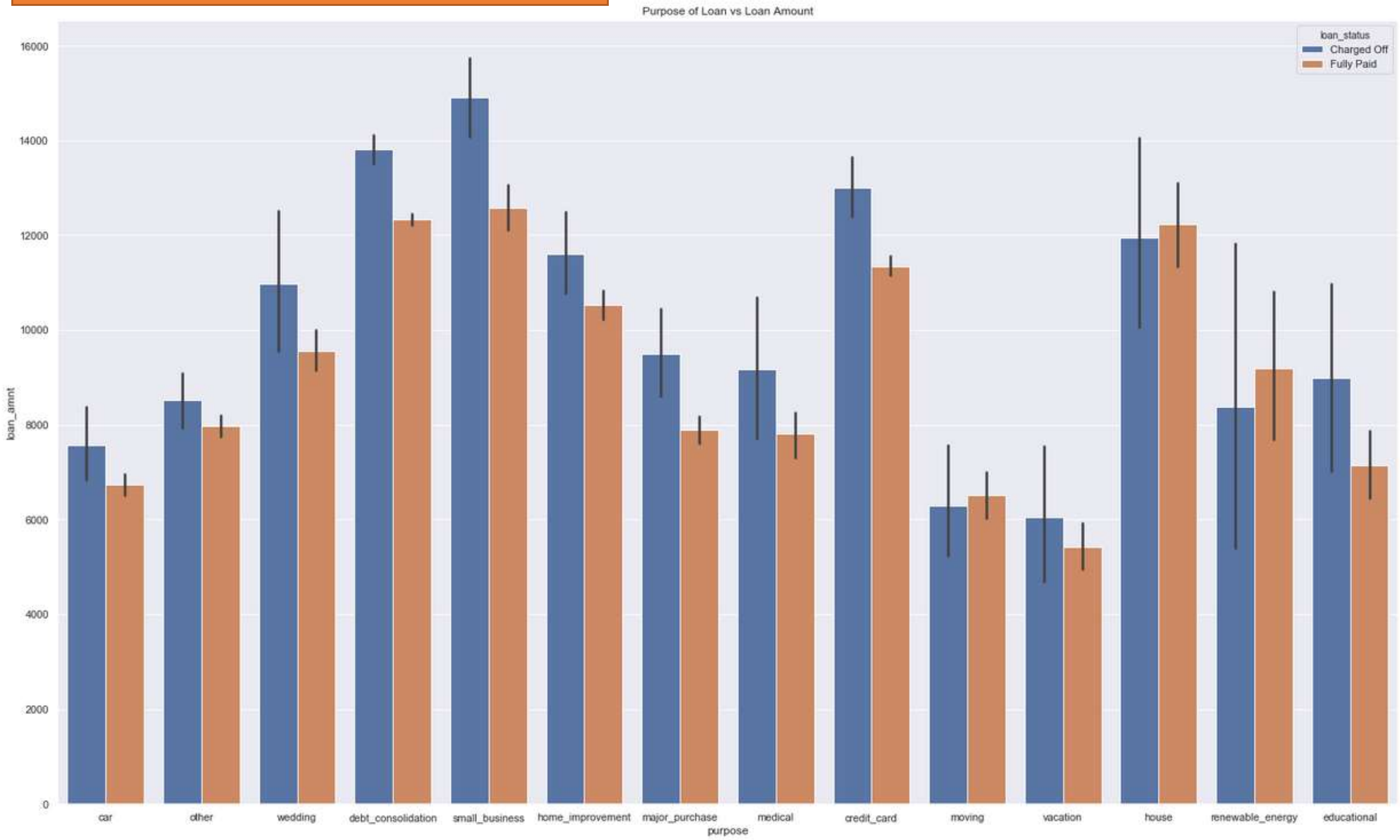
Correlation plot



Bivariate Analysis

UpGrad

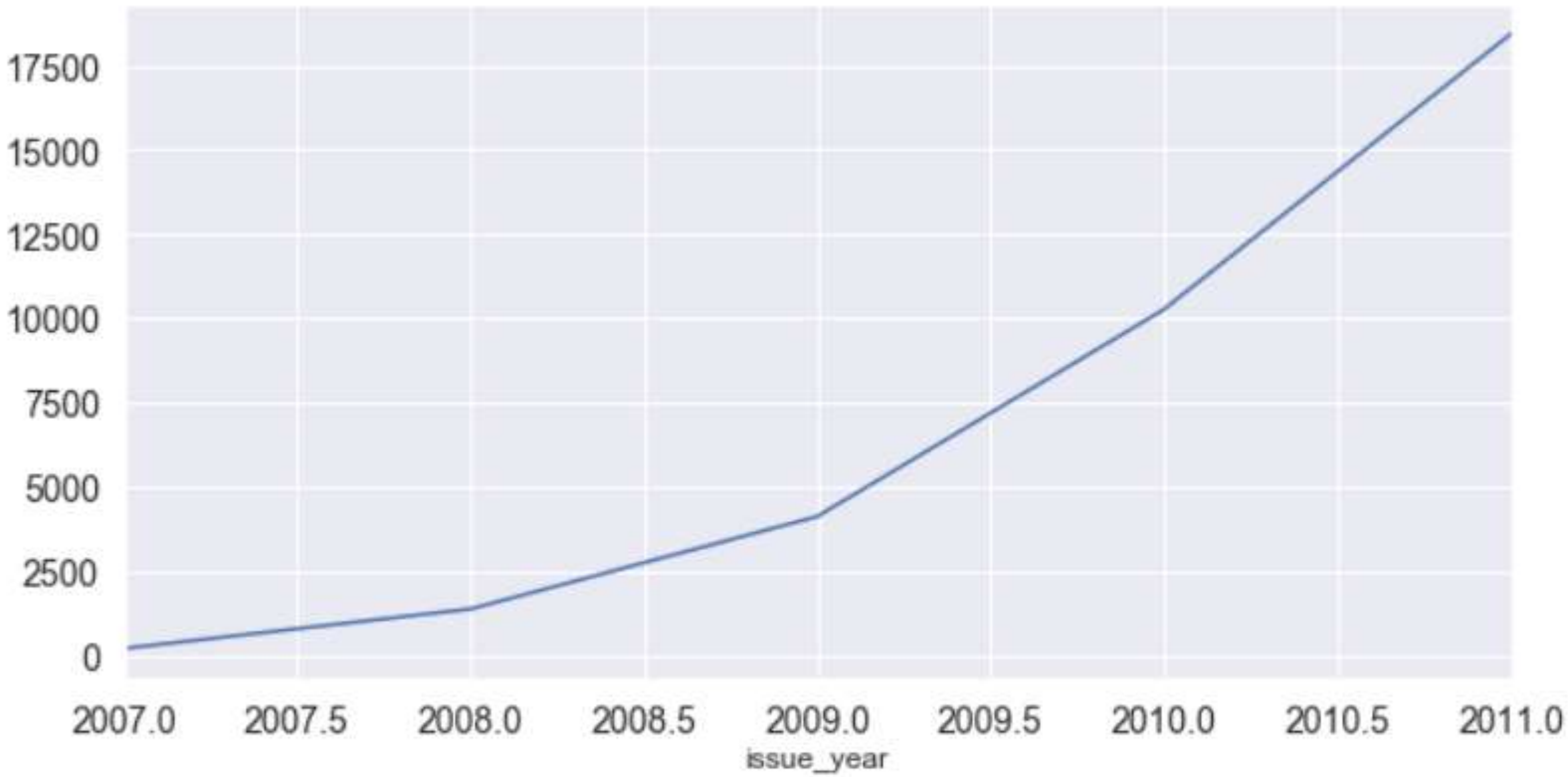
Purpose of Loan vs loan amount



Observation : People with small business tends to take more loans as compared to another purpose and they are also the ones which default the most

Bivariate Analysis

Loan amount vs Time

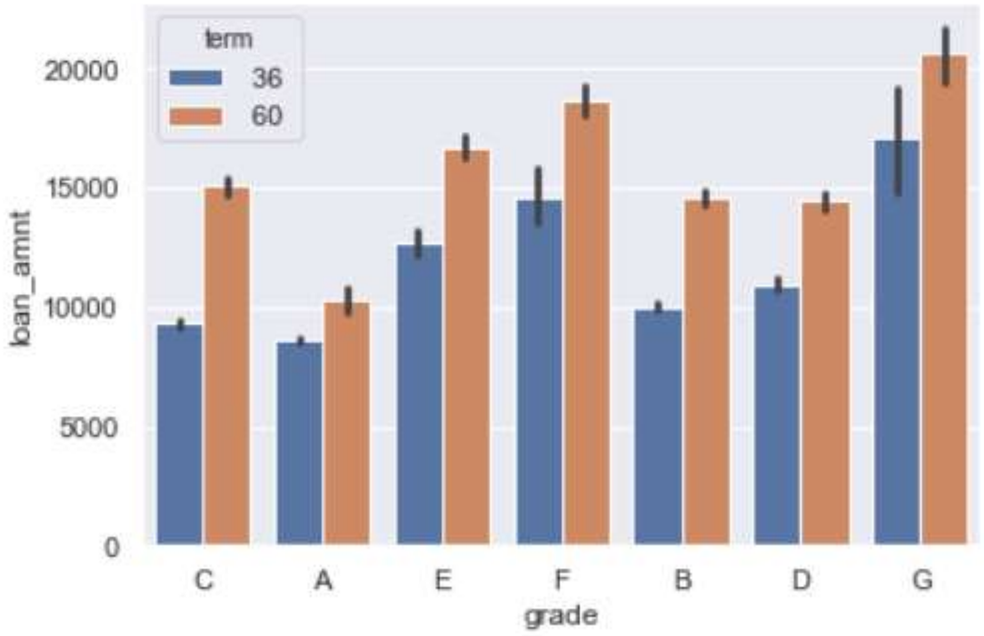


Observation : From 2007 onwards, people are taking more loan with maximum loan taken during the yar of 2011. This is another indication of economic growth

Bivariate Analysis

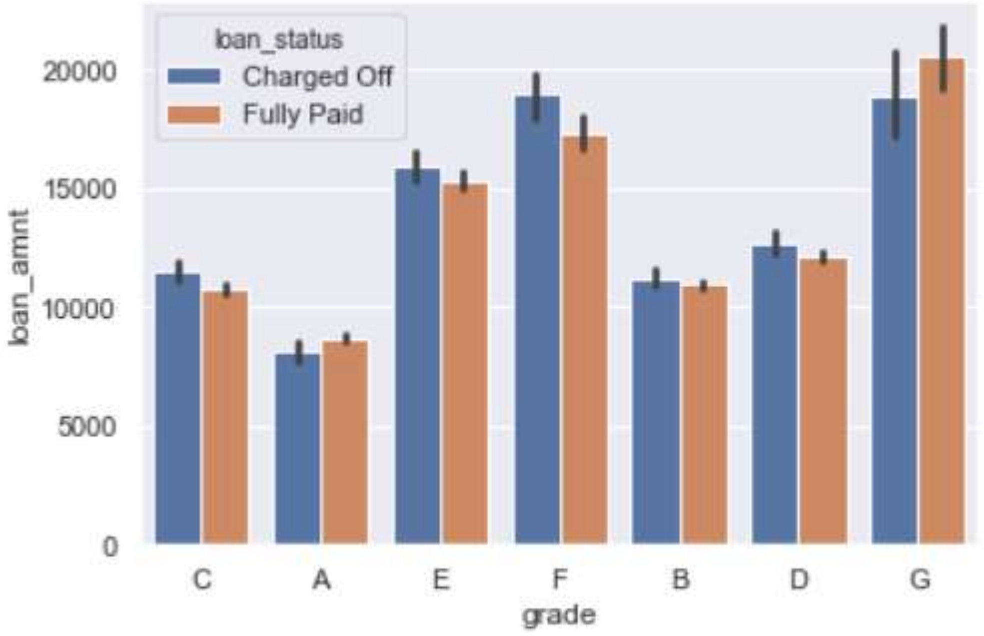
UpGrad

Loan amount vs Loan Amount vs Grade vs Term



Observation : People belonging to all the grades have taken loans for 36 months as well as 60 months

Loan amount vs Grade vs Loan Status

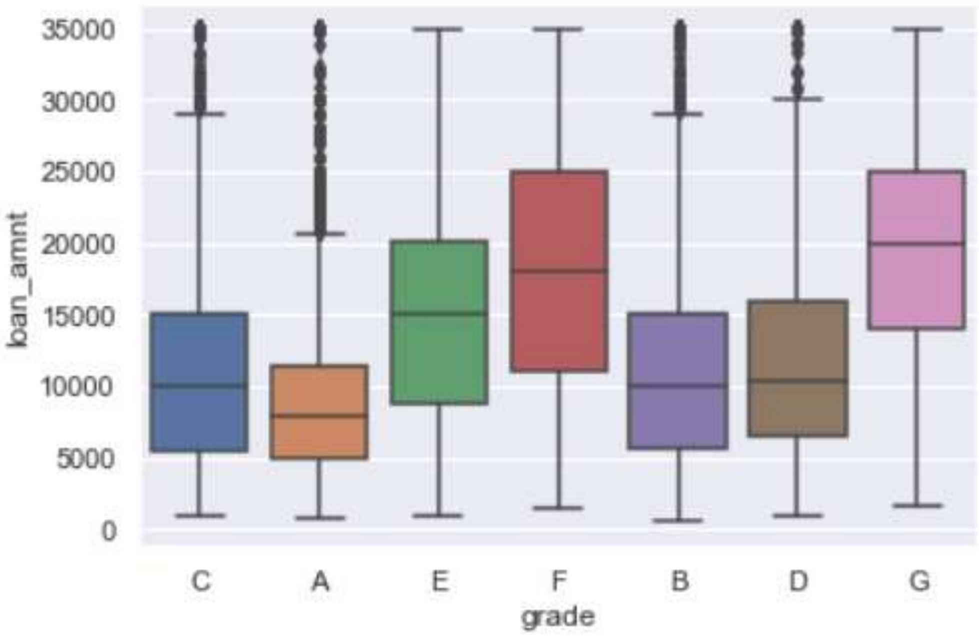


Observation : From the above graph, it is clear that the people with grade F, are defaulting more as compared to others.

Bivariate Analysis

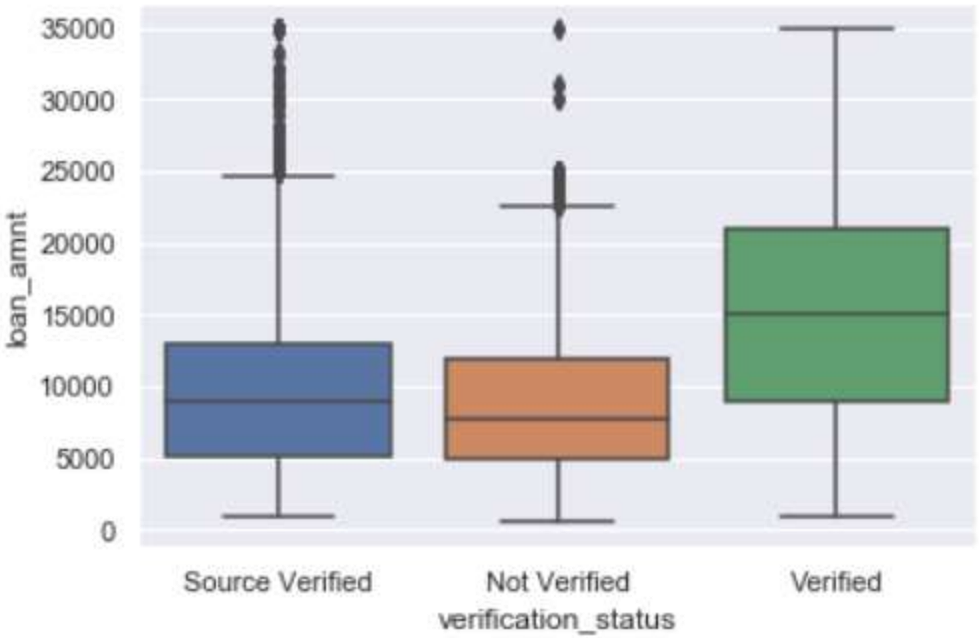
UpGrad

Loan amount vs Grade



Observation : There is a great variation in the distribution of loans amount the grades

Loan amount Vs Verification Status

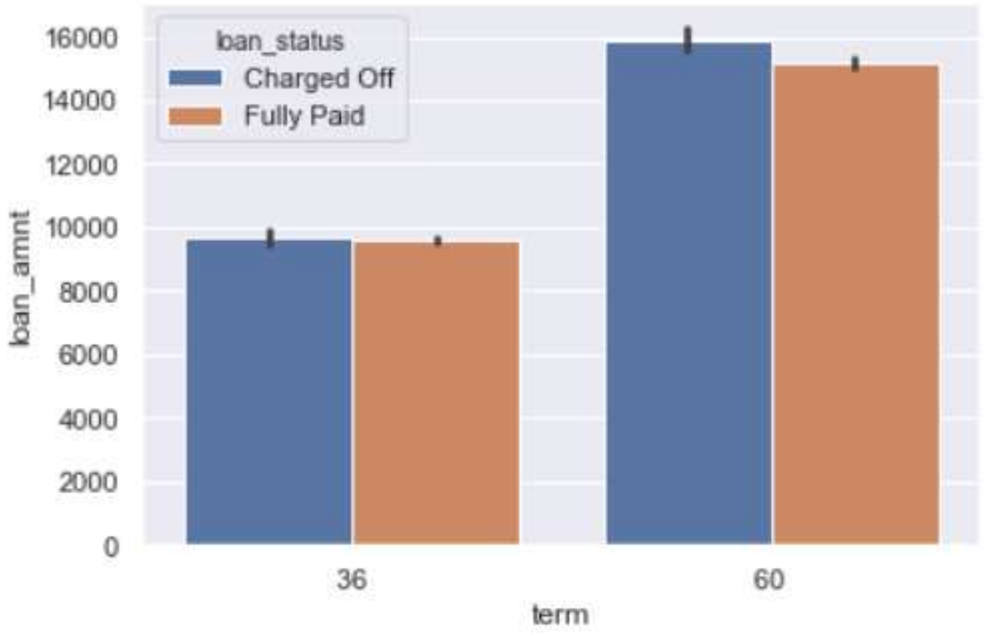


Observation : Verified people are granted more loan as compared to other verification status

Bivariate Analysis

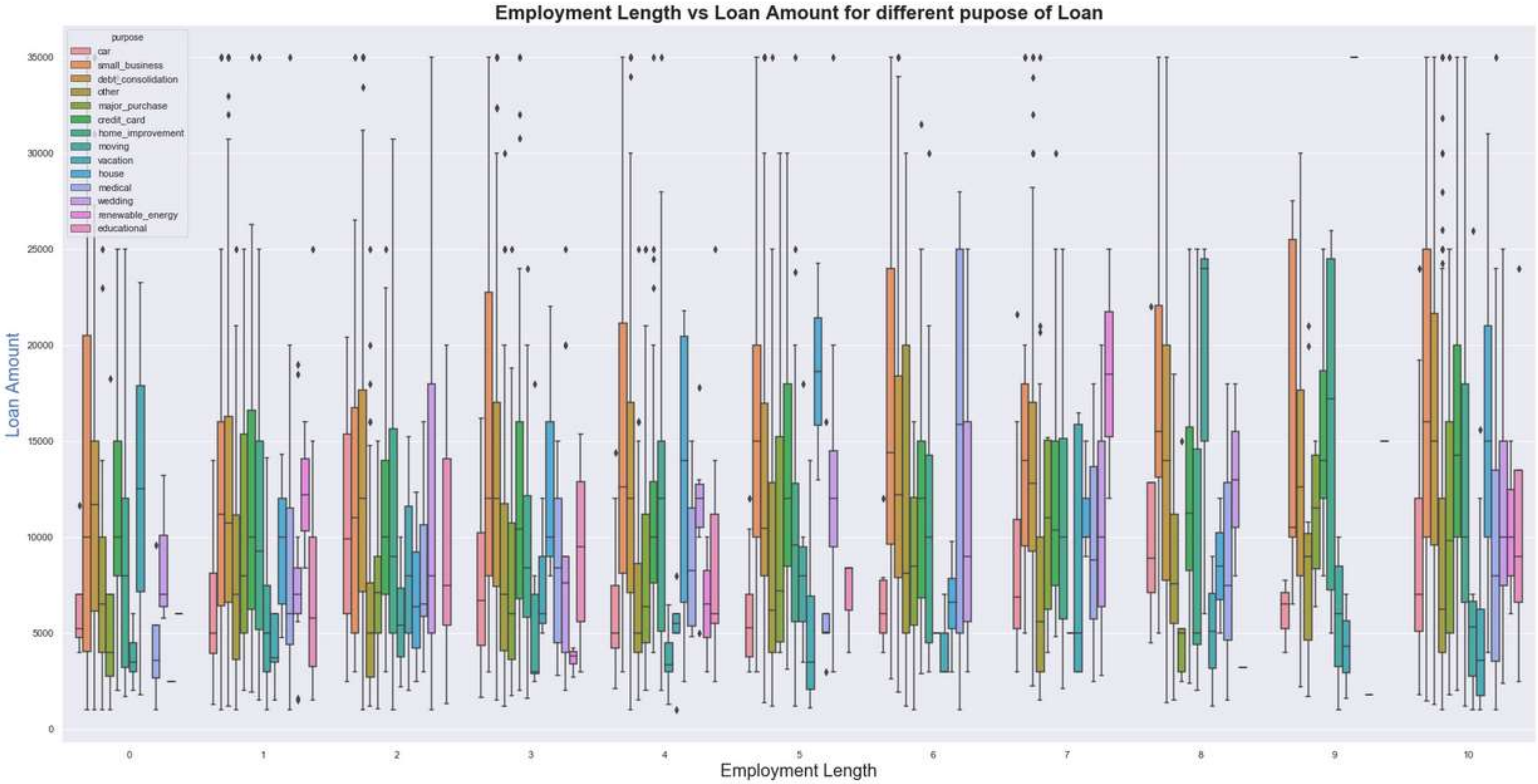
UpGrad

Loan amount vs term vs loan status

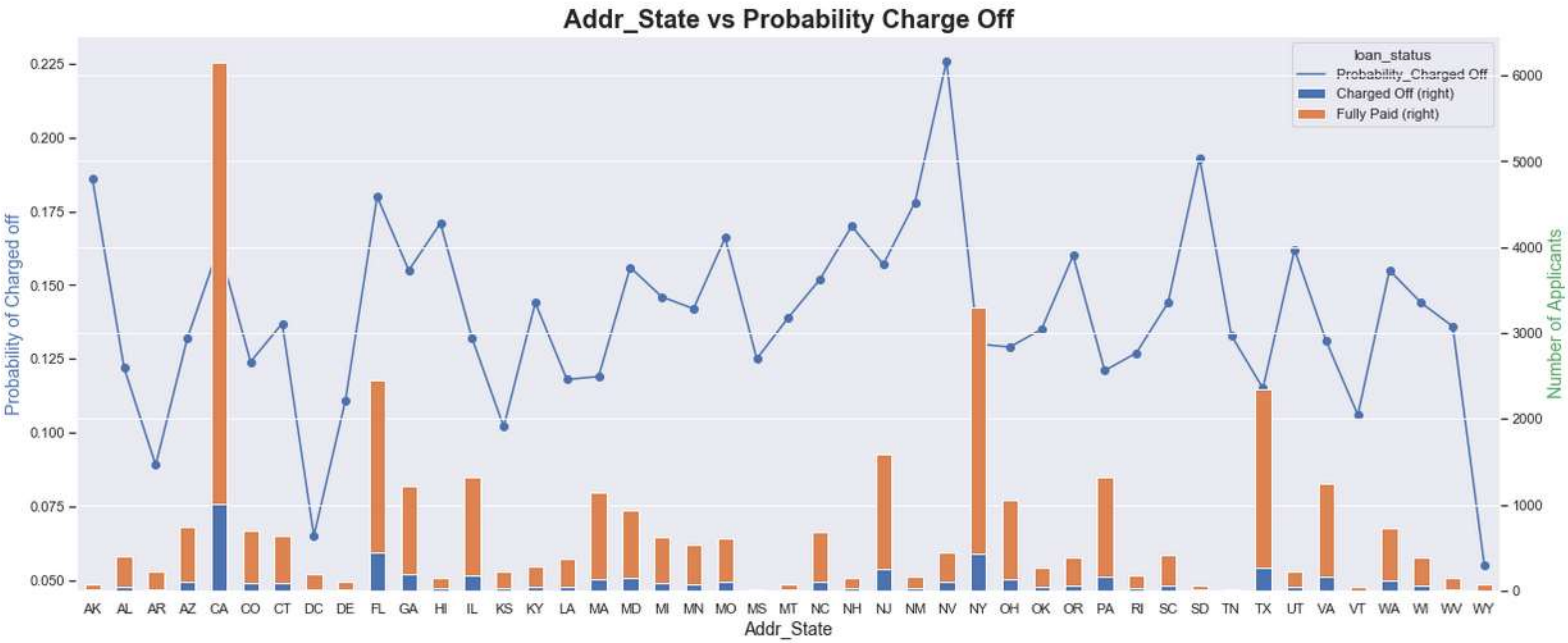


Observation : Higher loan amount are associated with longer terms and higher charges off.

Probability plot

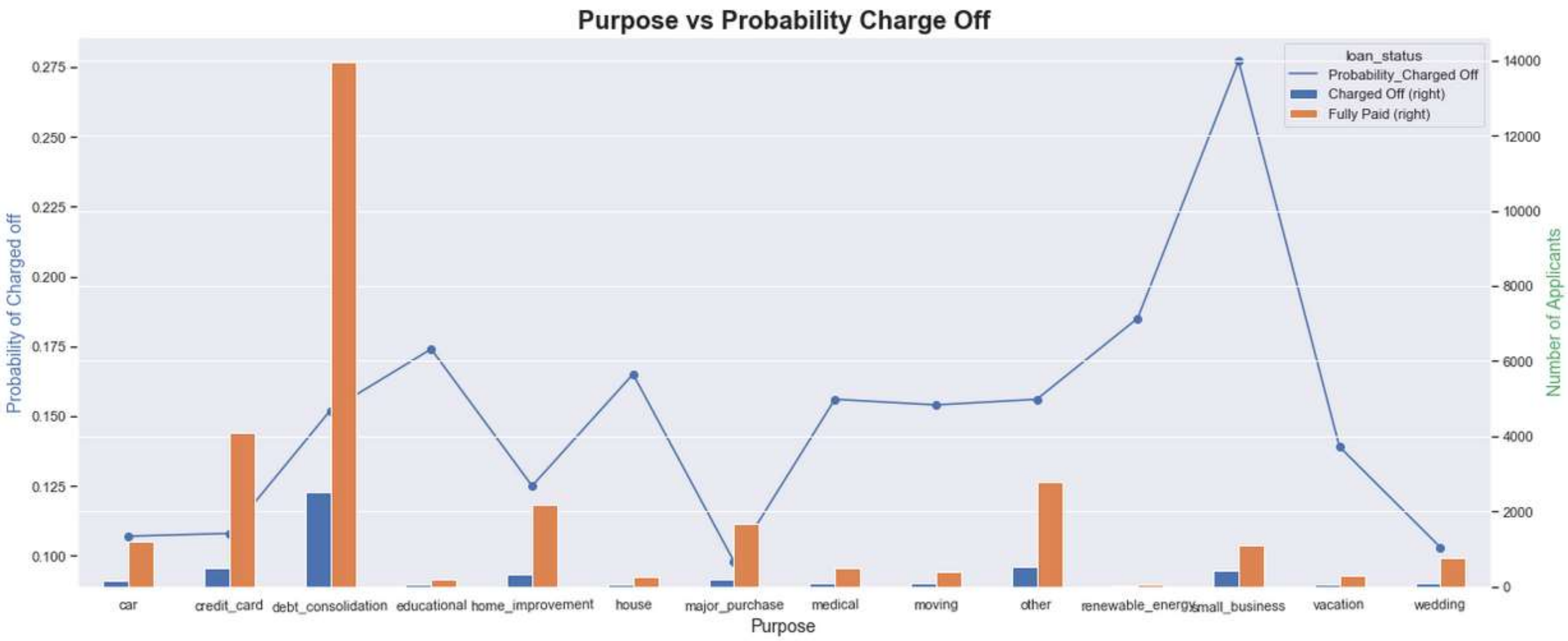


Probability plot



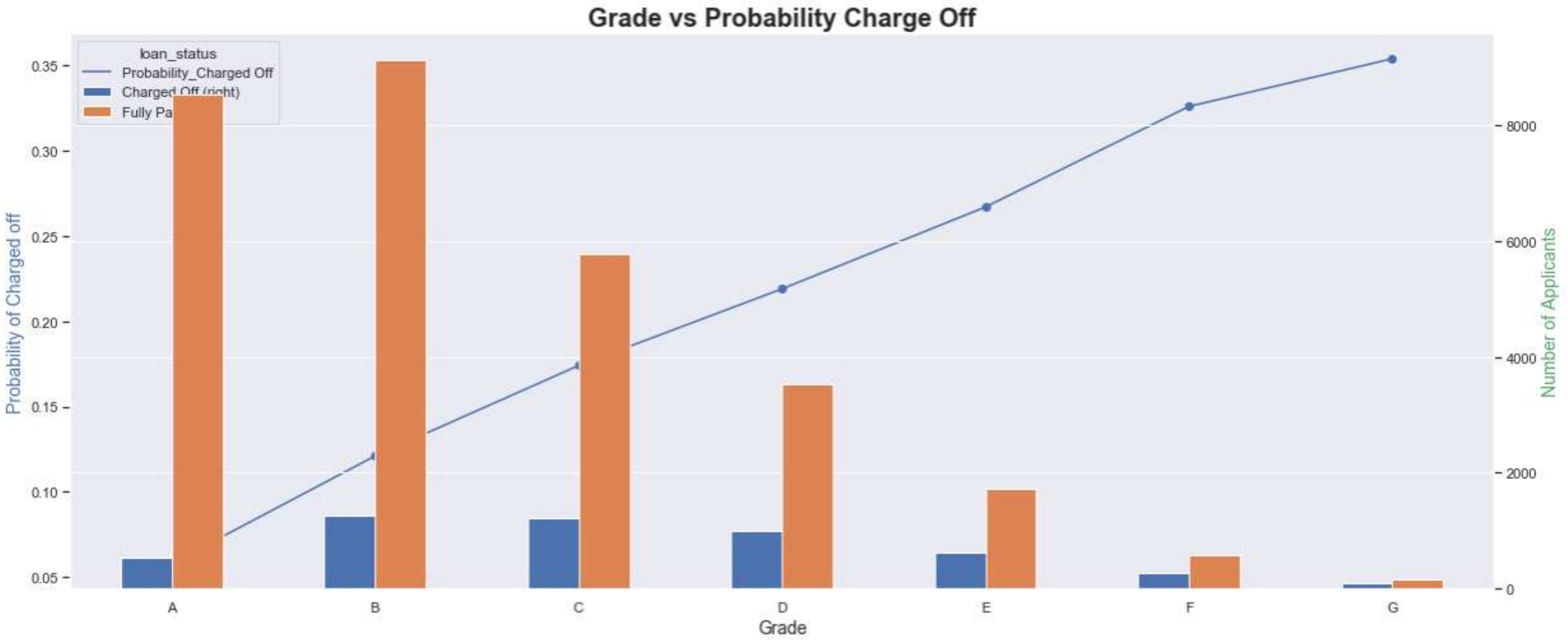
Observation : There are multiple States/Provinces with high probability where people tends to default more as compared to other states

Probability plot



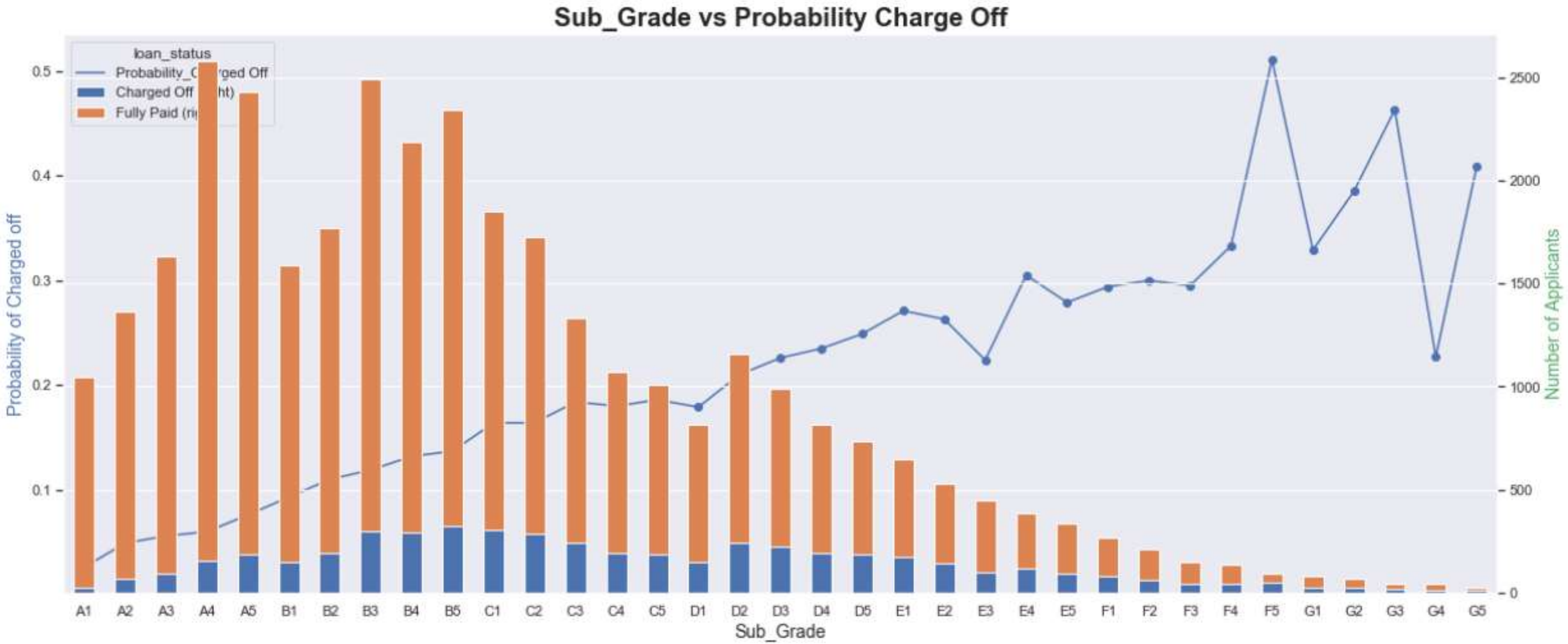
Observation : Applicants who has taken the Loan for 'small business' has the highest probabilty of charge off of 26%. So bank should take extra caution like take some asset or guarentee while approving the loan for purpose of 'small business'

Probability plot



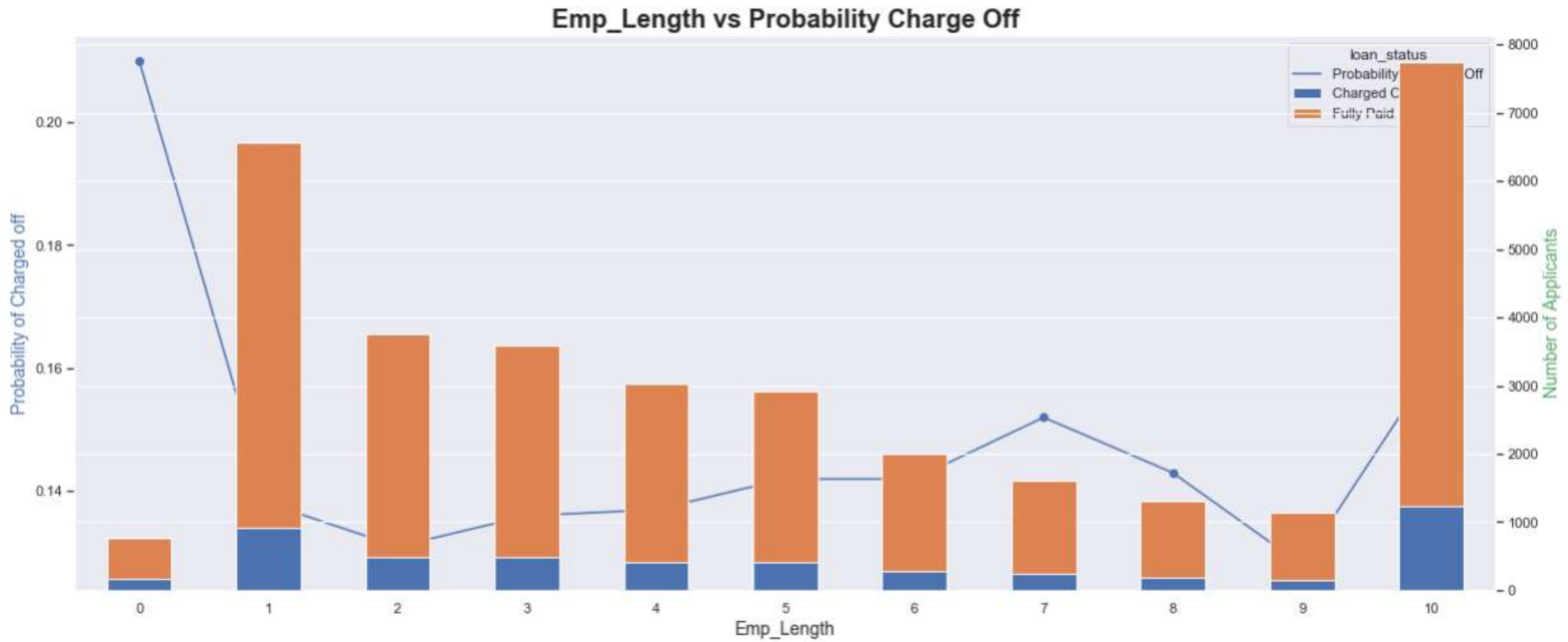
Observation: This graph show the result of probability charged off as per the number of applicants grade.

Probability plot



Observation : As we move from Grade A to G, probability that person will default on their loan is gradually increasing.

Probability plot



Observation : As the annual income is decreasing the probability that person will default is increasing with highest of 16% at (0 to 25000) salary bracket.

Important driver variables

Purpose of the loan

Employment Length

Grade

Interest Rate

Term

THANK YOU