Loan Dataset Case-study Report

- Rahul B
- Alhad Parashtekar

Aim

- ***The aim of the study is to analyze the risks associated with the Bank’s decision for loan approval when it receives a loan application using borrower information. The analysis will help us understand the underlying pattern connecting ‘loan_status’ column in the database and rest of the columns.***

- ***The aim of exploratory analysis is to figure out columns which are most influential or highly correlated with the ‘loan_status’ and gain additional insights.***

Each slide of the presentation would contain a figure showing impact of each variable present in the dataset on the loan status.

Columns from the database considered for exploratory analysis

- *The database contains 111 columns, out of which 54 columns contains only NaN values, and 2 columns contain only '0' as values. These 56 columns are not considered for the analysis (EDA).*

- *The 'member_id' and 'id' are not necessary to understand the correlation, hence not considered in EDA.*

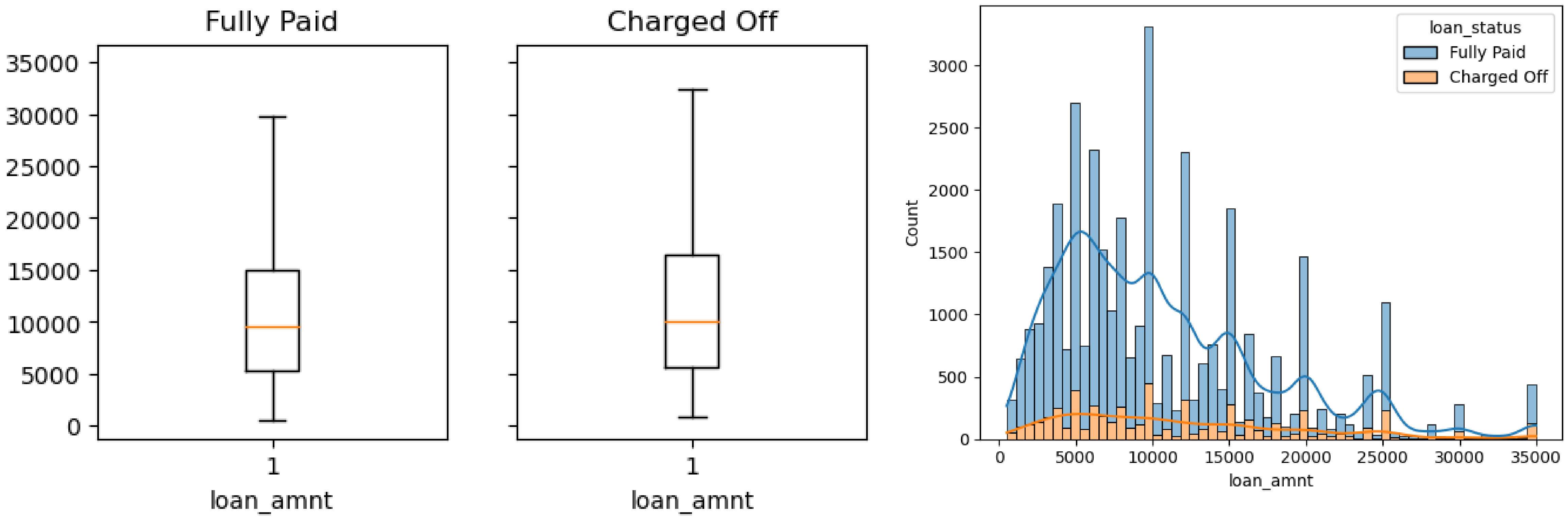
- Rest of the columns are selected for analysis:

  'loan_amnt', 'funded_amnt', 'int_rate', 'installment', 'grade', 'sub_grade', 'addr_state', 'all_util', 'annual_inc', 'chargeoff_within_12_mths', 'delinq_2yrs', 'funded_amnt_inv', 'loan_status', 'mort_acc', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_il_tl', 'num_sats', 'pymnt_plan', 'revol_bal', 'revol_util', 'tax_liens', 'tot_coll_amt', 'zip_code', 'term', 'purpose'

Univariate Analysis

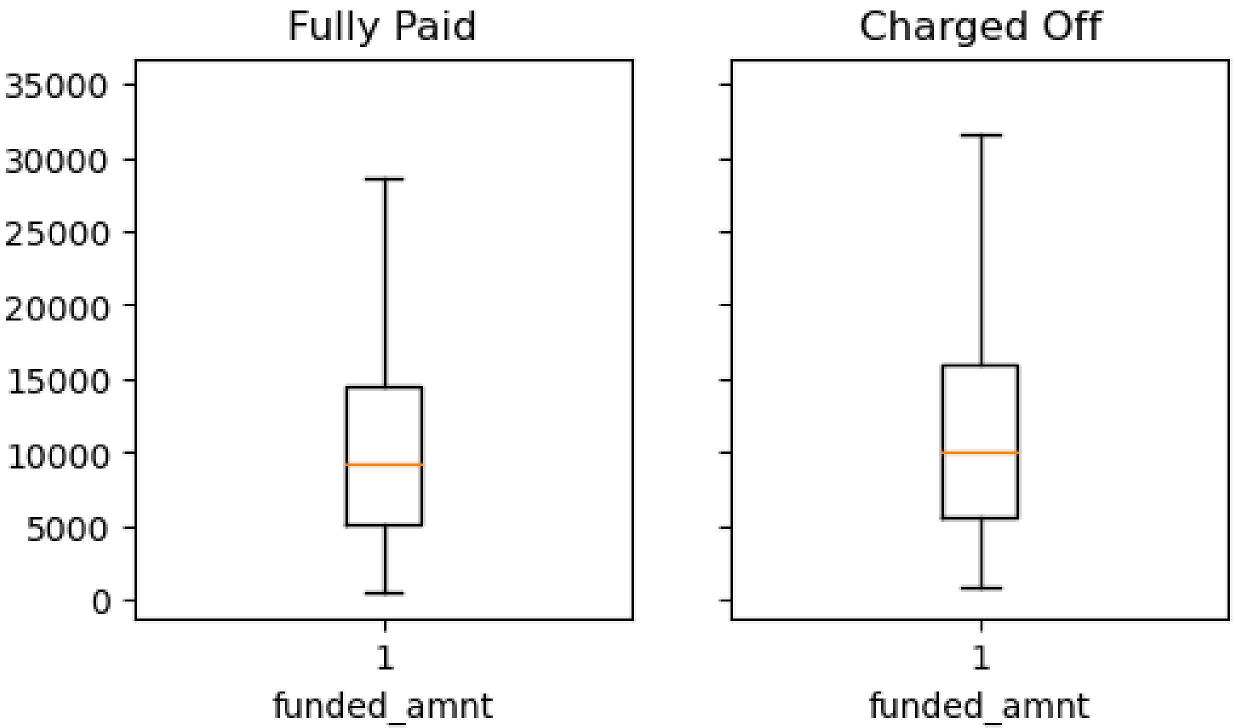
*Impact of **Loan Amount***

Though the median and other quartiles are higher in charged-off loans, no significant difference in distribution can be observed



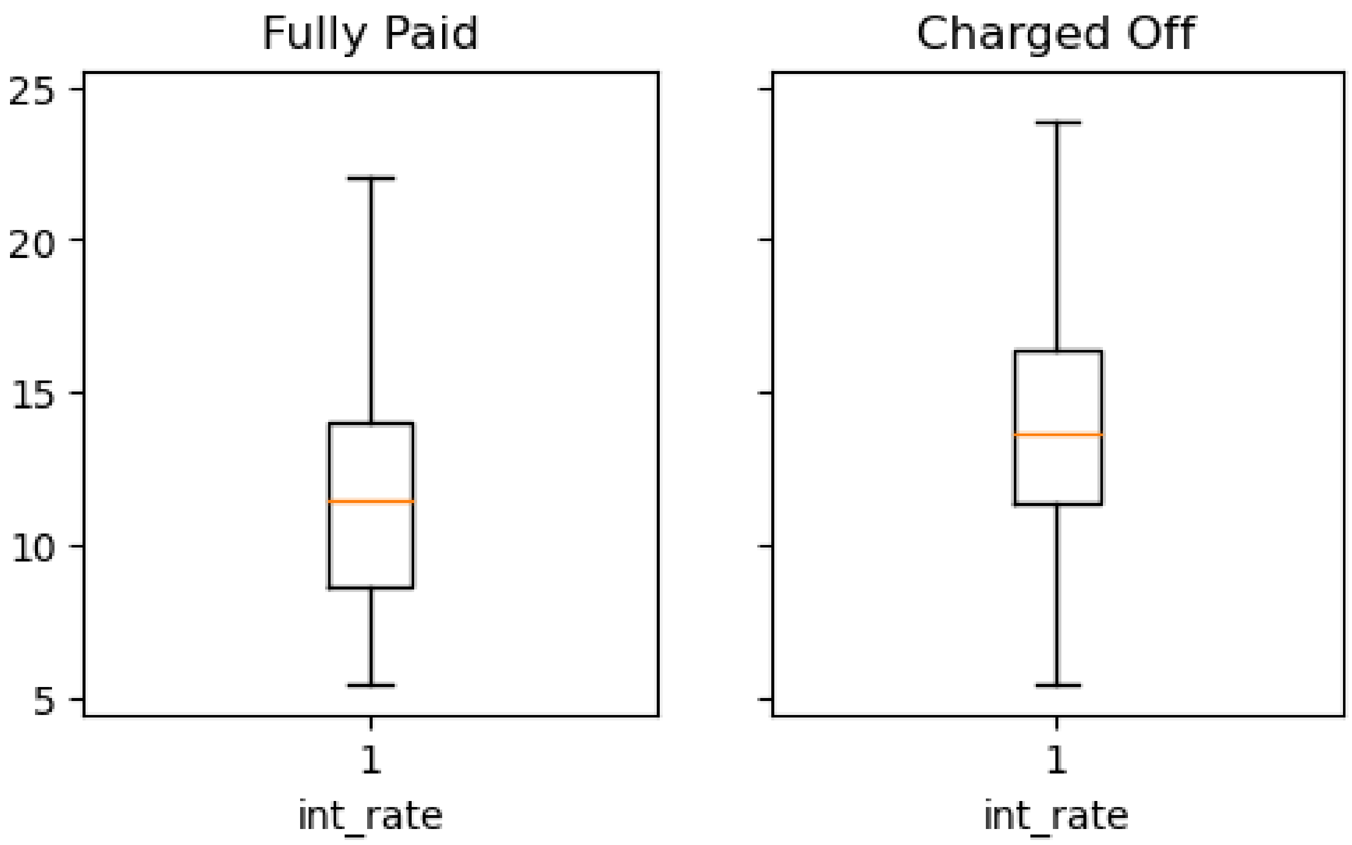
*Impact of **Funded Amount***

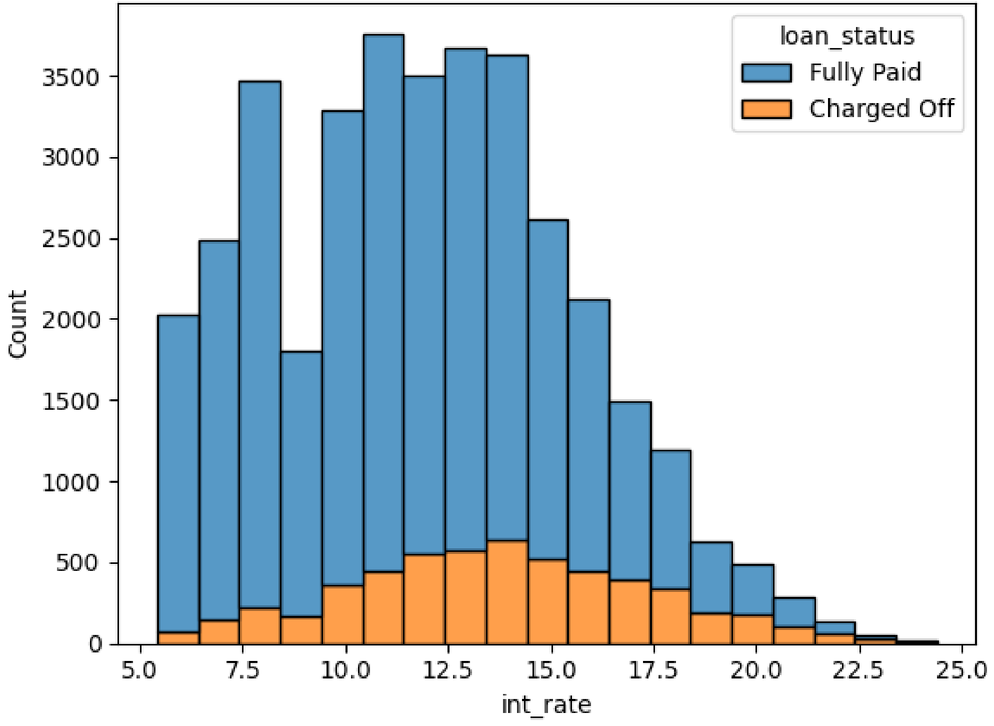
Though the spread of ‘funded_amnt’ is slightly larger for ‘charged_off’, no significant difference in distribution can be observed.



*Impact of **Interest Rate***

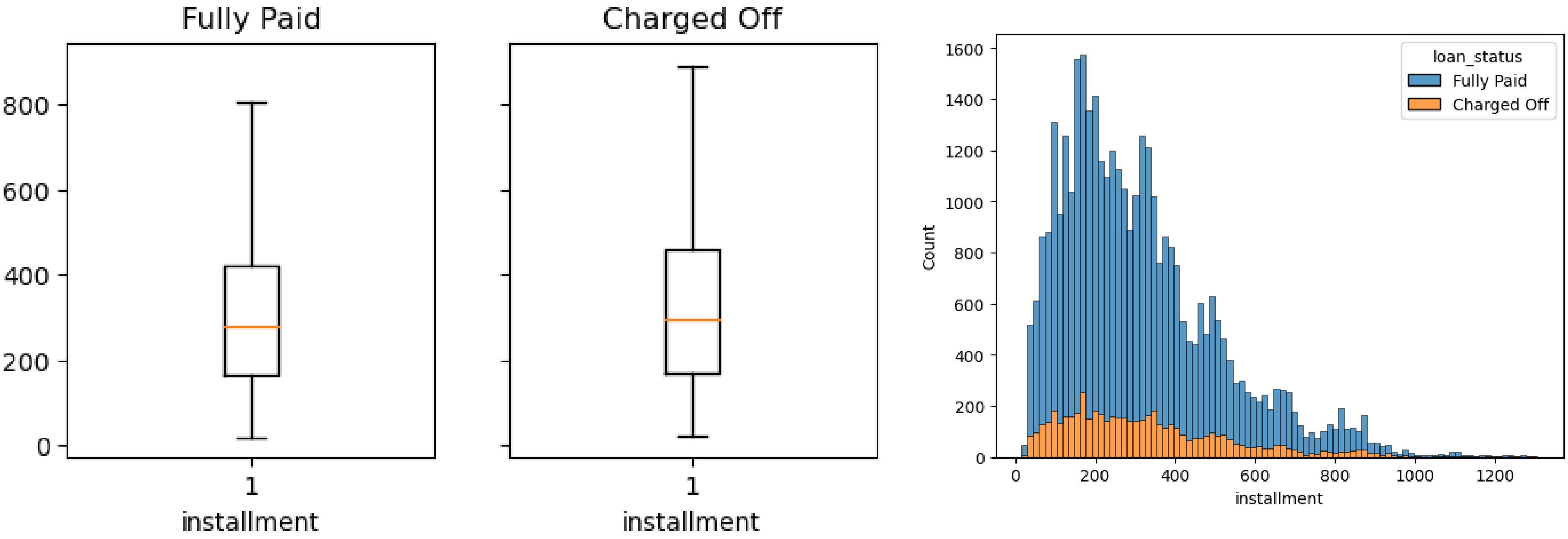
The median and other quartiles (25 and 75) are considerably higher in case of ‘charged_off’ loans than ‘fully_paid’ loans. This variable seems to influence the ‘loan_status’.





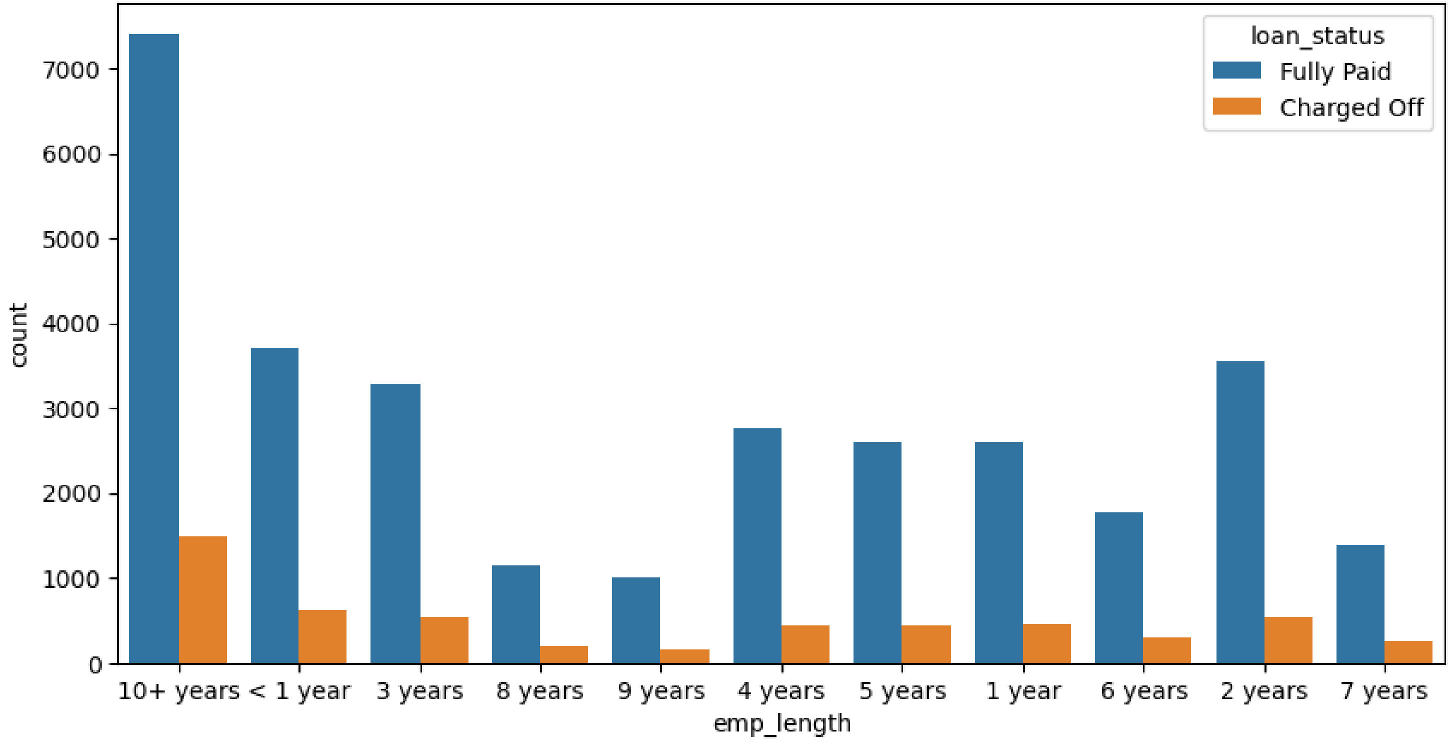
*Impact of **Interest Rate***

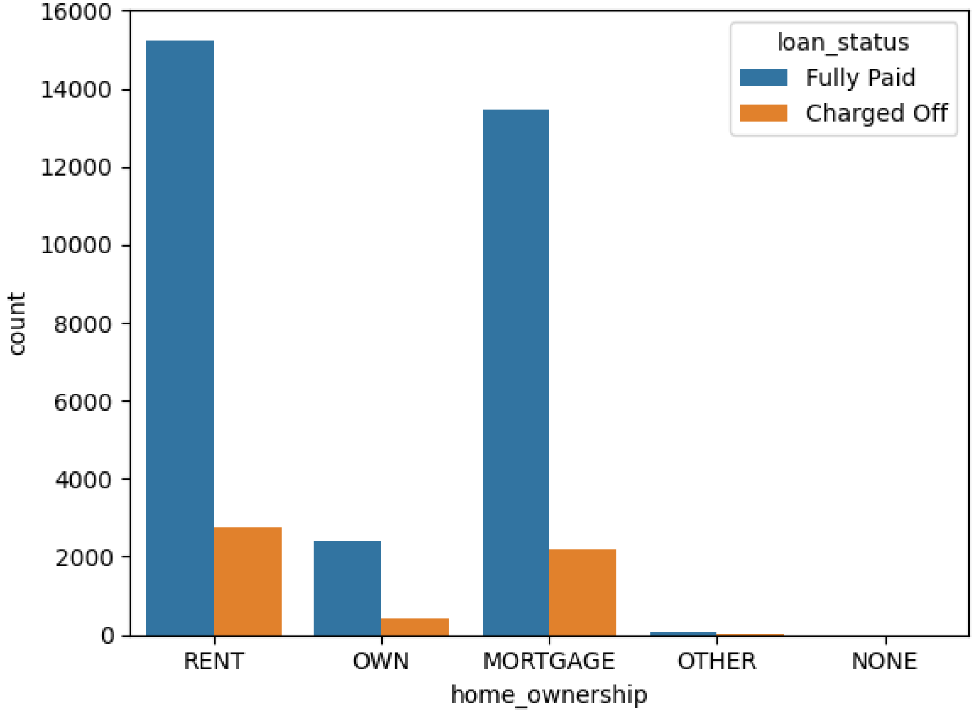
No significant difference can be observed here in distribution for ‘loan_amnt’.



*Impact of **Employment Length***

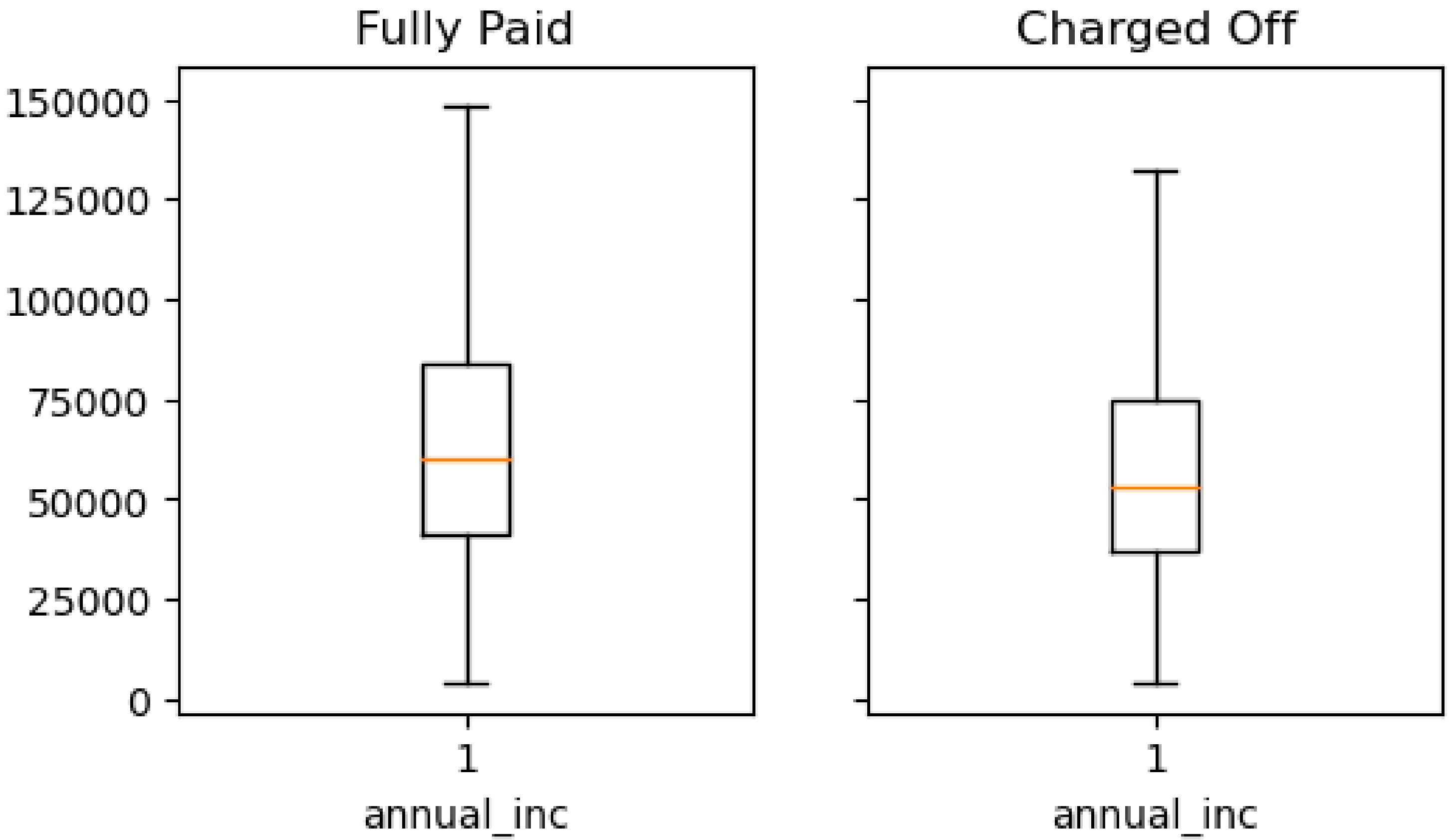
No significant difference can be observed here in distribution for ‘emp_length’ and ‘home_ownership’.





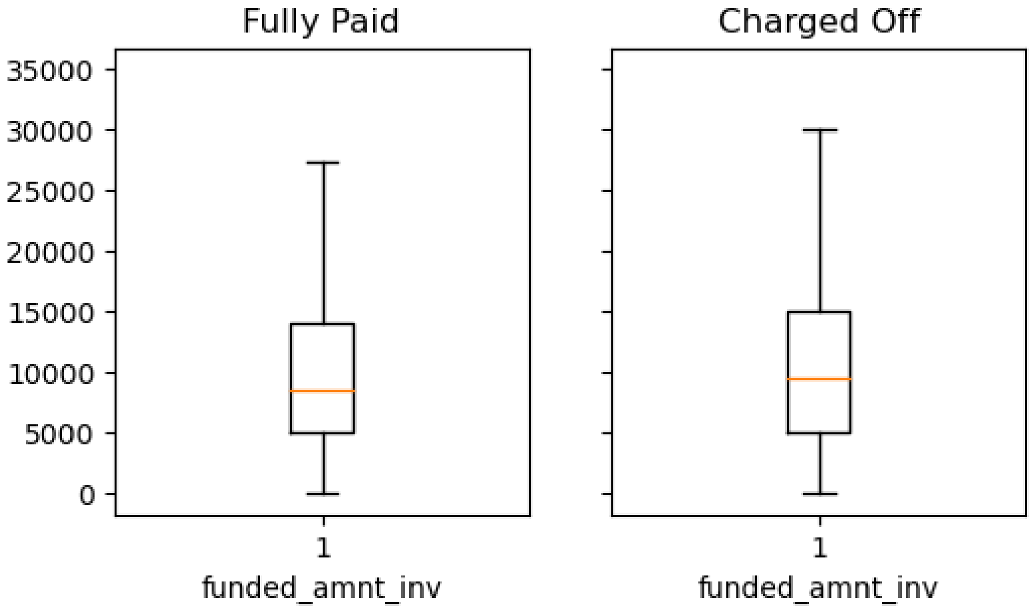
*Impact of **Annual Income***

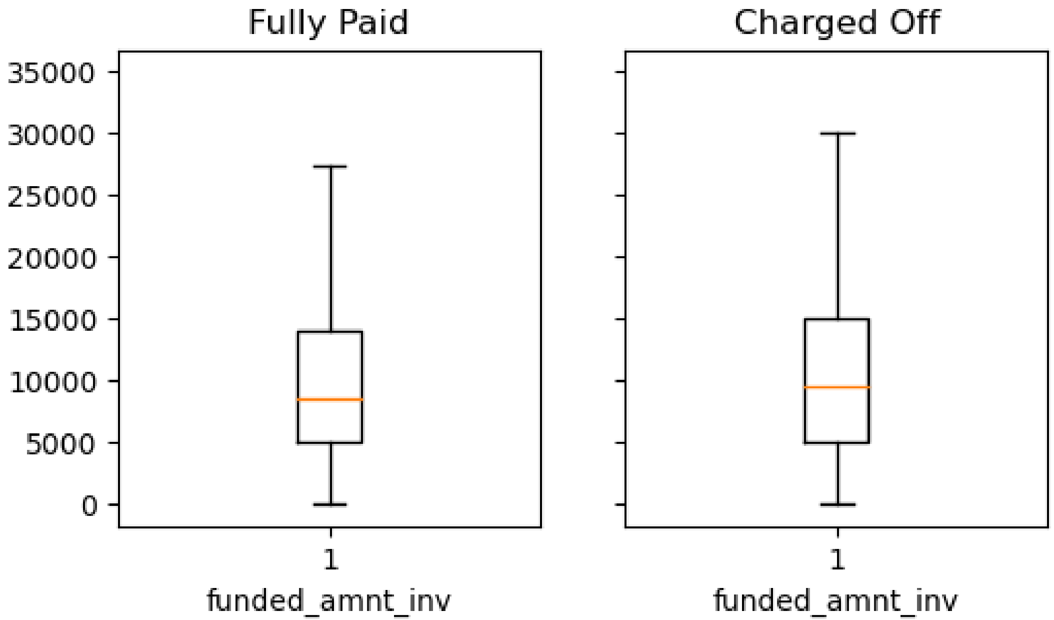
The fully paid loans have higher median and other quartiles for 'annual_income'. The variable seems to have some impact on the loan status.



*Impact of **Interest Rate***

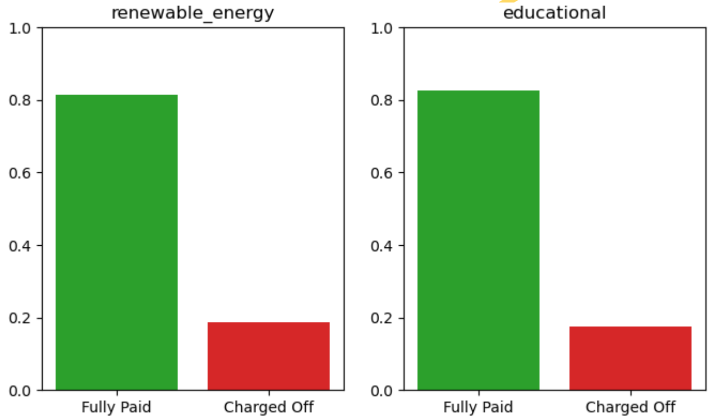
No significant difference can be observed here in distribution for ‘funded_amnt_inv’ and ‘revol_bal’.

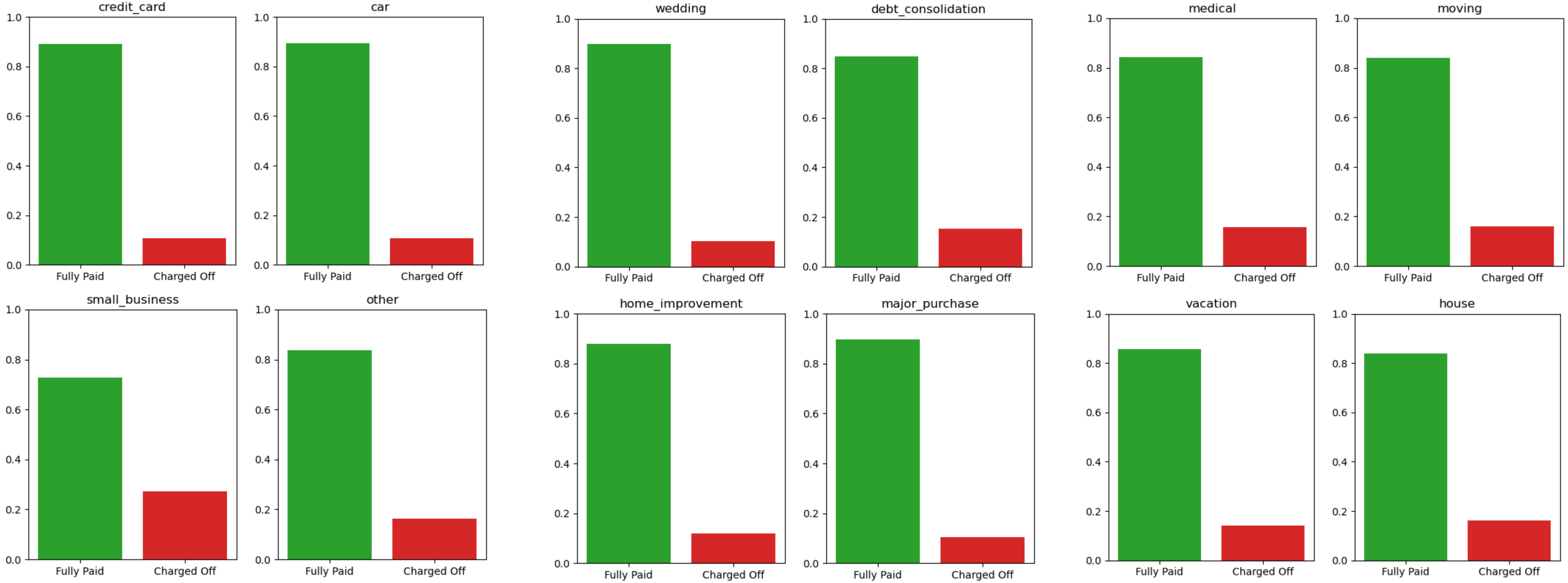




*Impact of **Purpose***

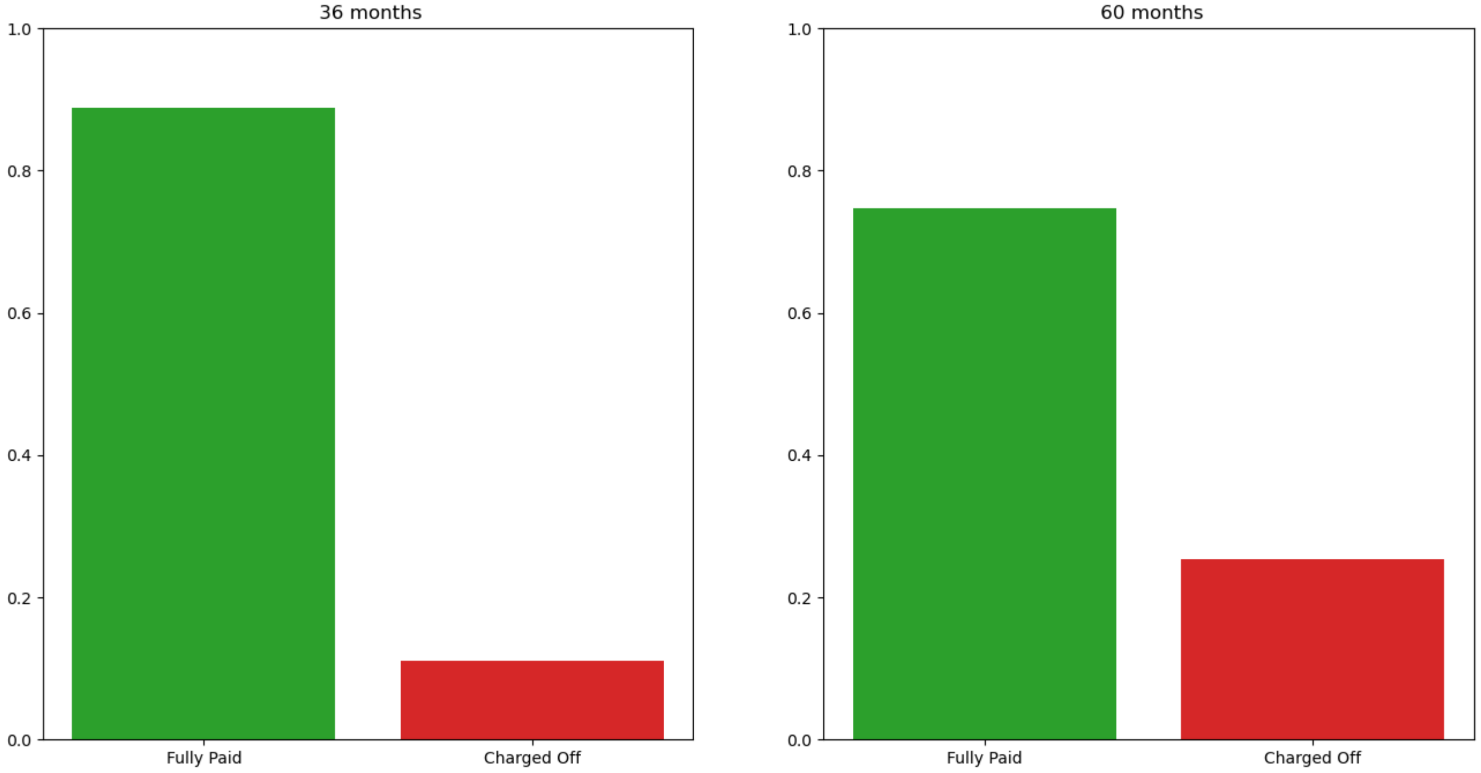
The purpose of the loan seems to impact the 'loan_status' variables. Some purpose values have higher percentage of the charged-off loans compared to others.





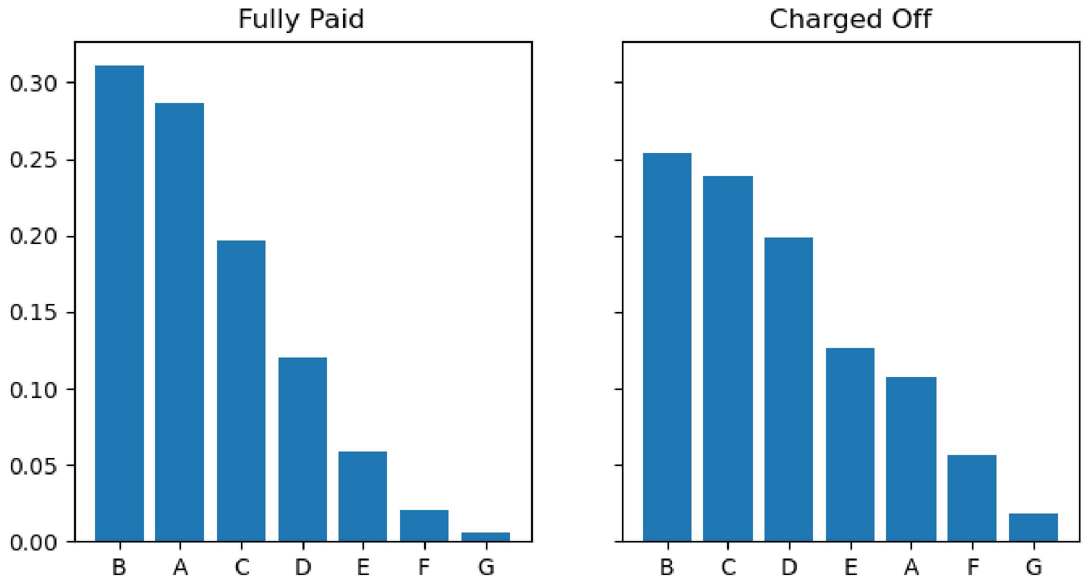
*Impact of **Term***

Peercentage of ‘charged_off’ loans is almost double in case of higher terms, i.e., ‘60 months’, when compared to shorted months ’36 months’. This variable seems to have larger impact on the loan status.

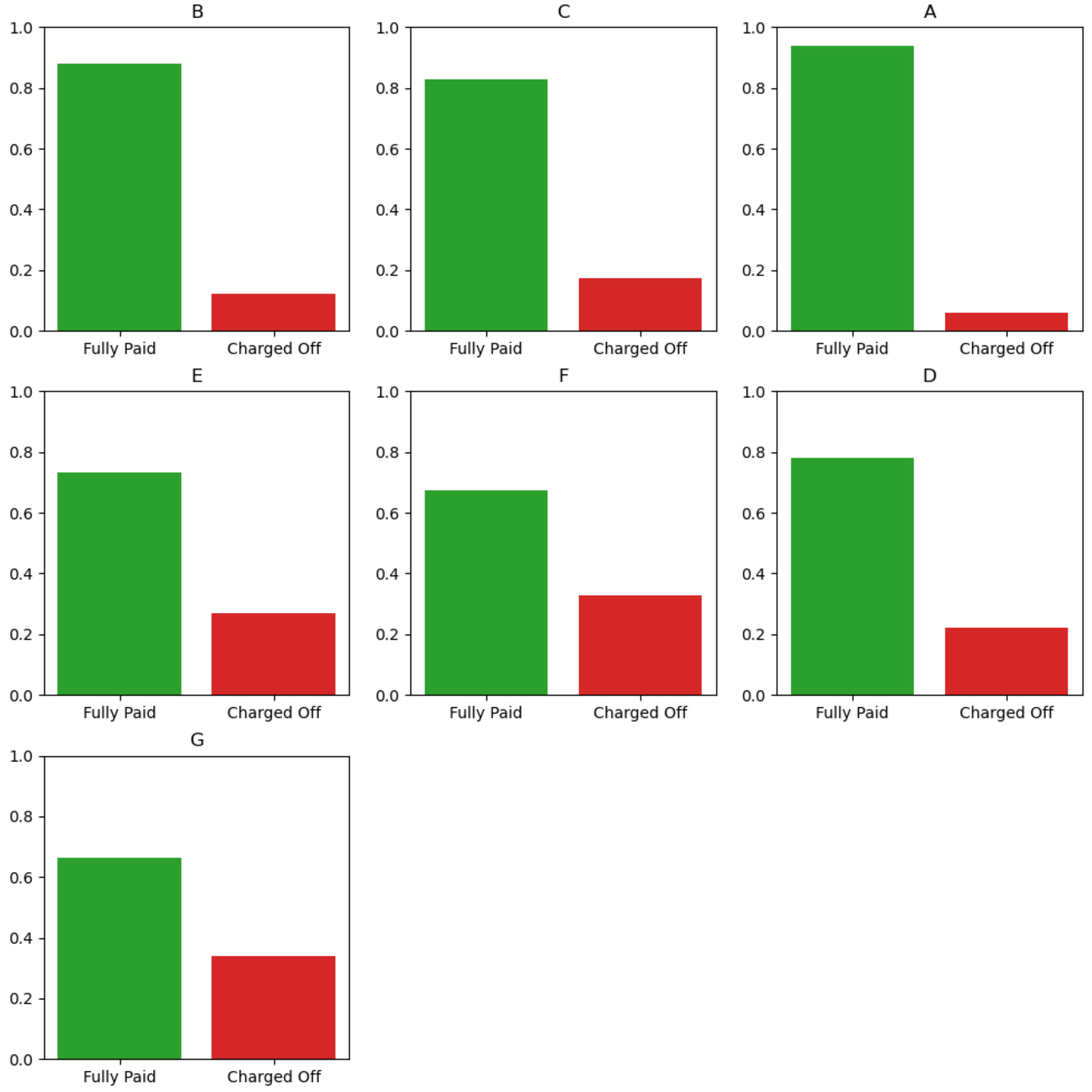


*Impact of **Grade***

As the grade quality is declining (A to G), we see higher percentage of the 'charged_off' loans. Hence, 'grade' is an influential variable.



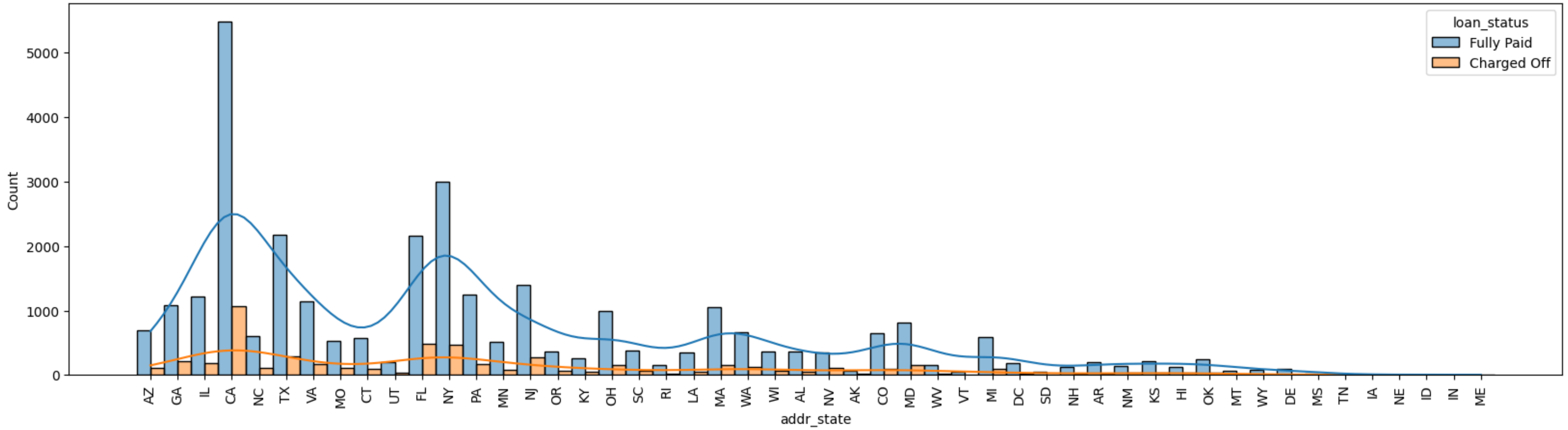
Percentage of each grade in the 'fully_paid' and 'charged_off' loans

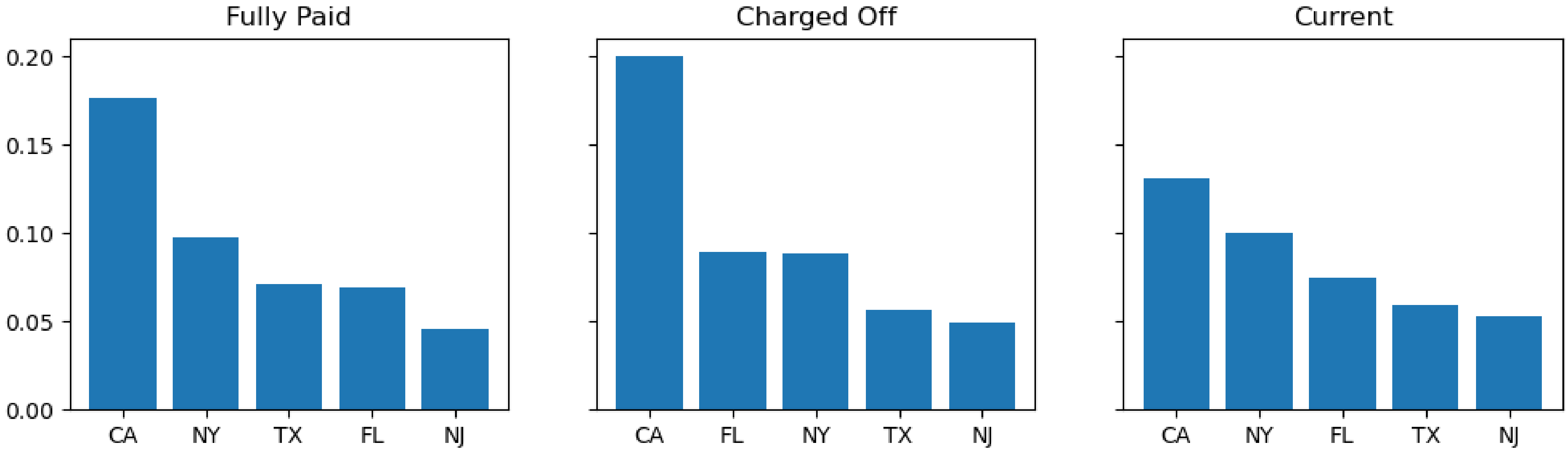


Percentage of 'fully_paid' and 'charged_off' loans in each grade.

*Impact of **City***

No particular impact of the city, i.e., ‘addr_state’ can be observed.



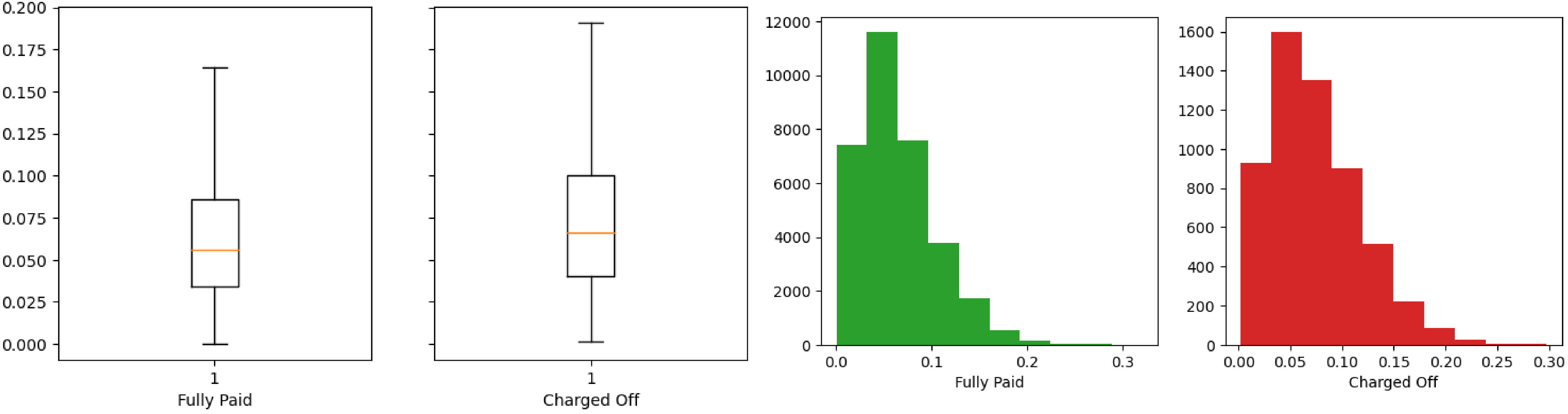


*Impact of **Installment to Income ratio***

We created a new column with following formula

Ratio = installment*12/annual_inc

The charged-off loans have slightly higher quartiles for the ‘ratio’. However, no significant difference can be observed here in distribution for ‘ratio’.



Bi-variate Analysis

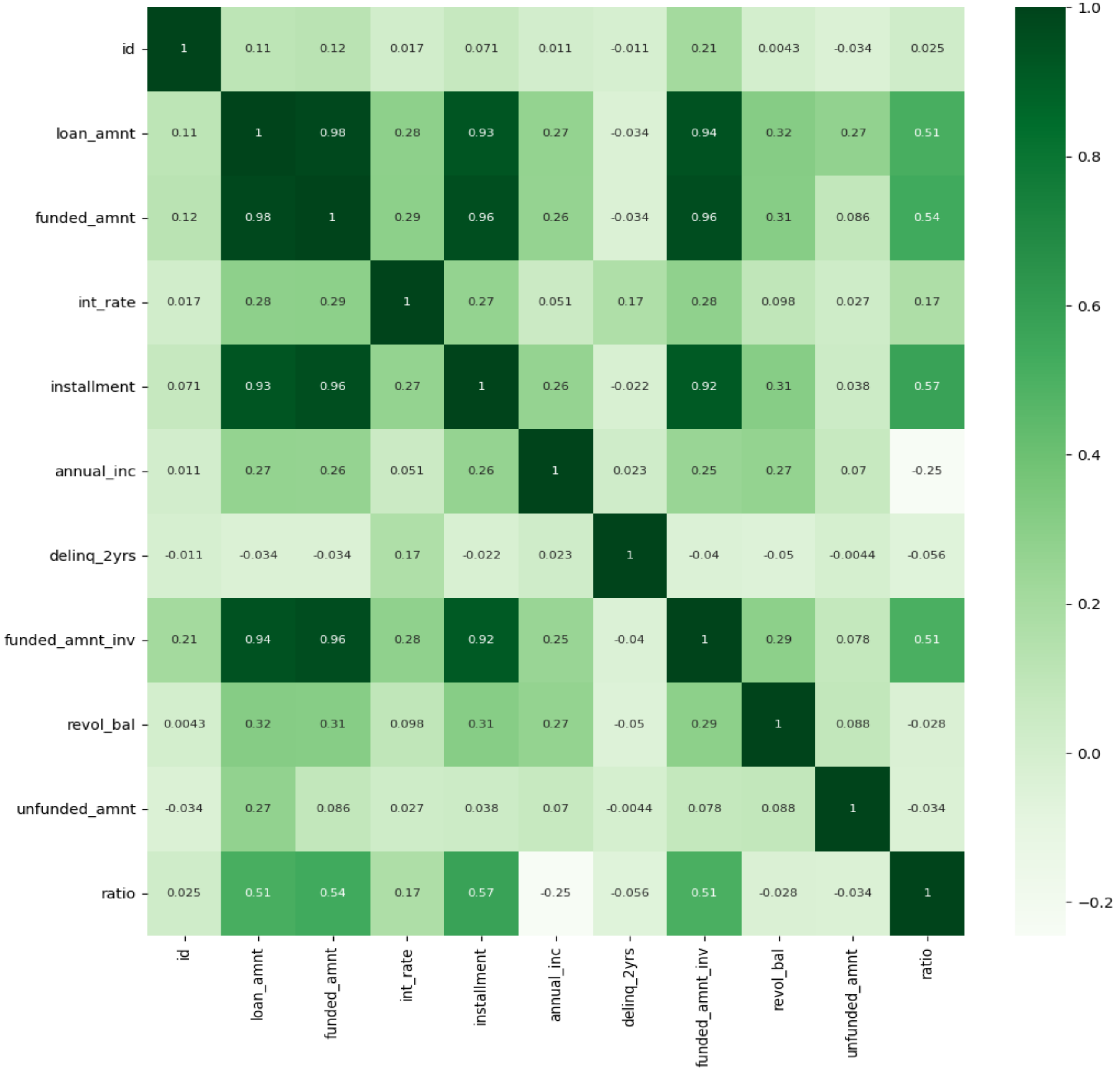
Correlation Analysis

Some variables show very high co-relation.

However, the correlation is very-much expected.

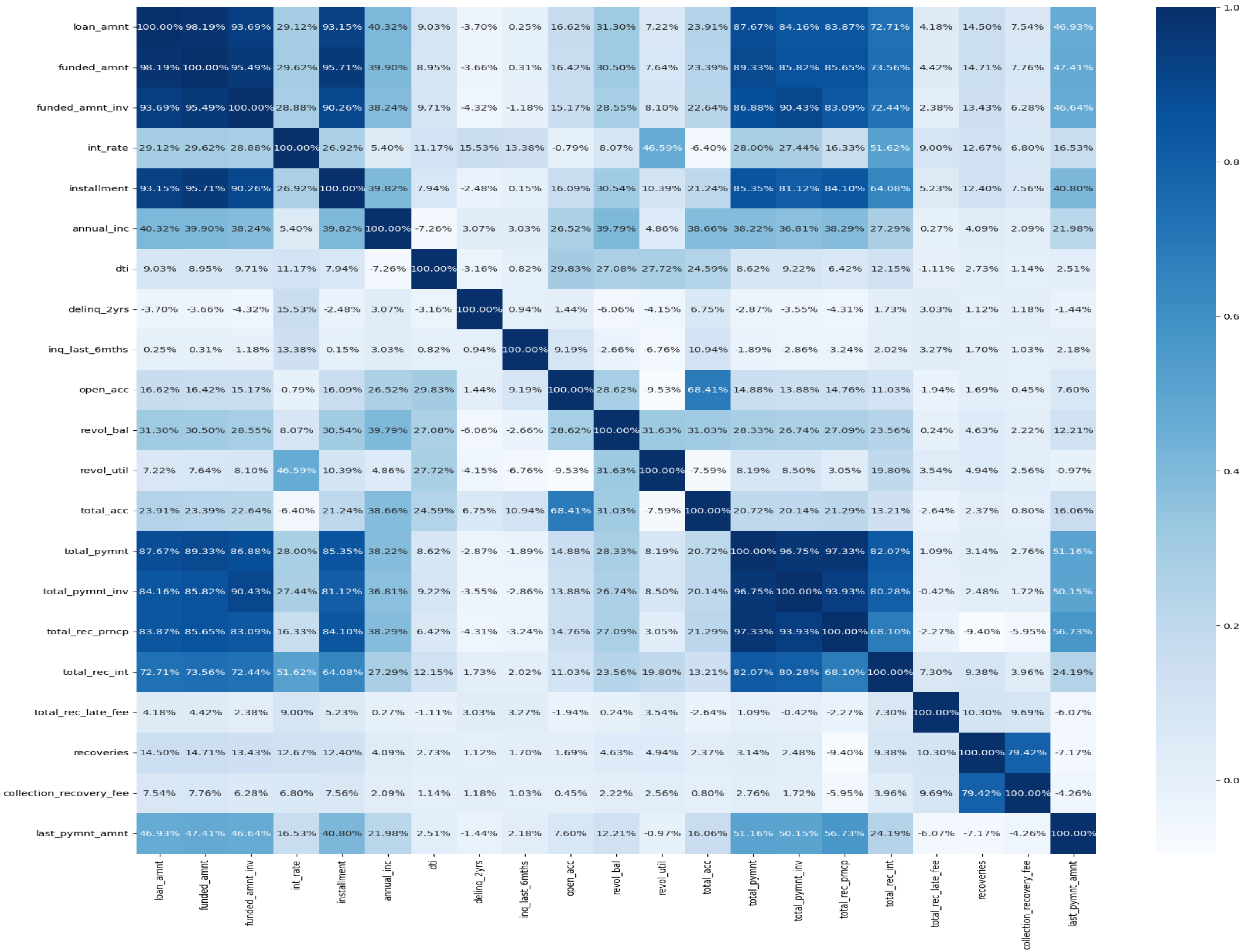
For example, it is obvious that the installment will be higher in the fraction of the loan_amnt.

From the highly co-related variables, i.e., 'loan_amnt', 'funded_amnt', 'installment', 'funded_amnt_inv', only one variable can be considered for model as the univariate analysis already showed that these variables do not have any significant impact.



	id	loan_amnt	funded_amnt	int_rate	installment	annual_inc	delinq_2yrs	funded_amnt_inv	revol_bal	unfunded_amnt	ratio
id	1	0.11	0.12	0.017	0.071	0.011	-0.011	0.21	0.0043	-0.034	0.025
loan_amnt	0.11	1	0.98	0.28	0.93	0.27	-0.034	0.94	0.32	0.27	0.51
funded_amnt	0.12	0.98	1	0.29	0.96	0.26	-0.034	0.96	0.31	0.086	0.54
int_rate	0.017	0.28	0.29	1	0.27	0.051	0.17	0.28	0.098	0.027	0.17
installment	0.071	0.93	0.96	0.27	1	0.26	-0.022	0.92	0.31	0.038	0.57
annual_inc	0.011	0.27	0.26	0.051	0.26	1	0.023	0.25	0.27	0.07	-0.25
delinq_2yrs	-0.011	-0.034	-0.034	0.17	-0.022	0.023	1	-0.04	-0.05	-0.0044	-0.056
funded_amnt_inv	0.21	0.94	0.96	0.28	0.92	0.25	-0.04	1	0.29	0.078	0.51
revol_bal	0.0043	0.32	0.31	0.098	0.31	0.27	-0.05	0.29	1	0.088	-0.028
unfunded_amnt	-0.034	0.27	0.086	0.027	0.038	0.07	-0.0044	0.078	0.088	1	-0.034
ratio	0.025	0.51	0.54	0.17	0.57	-0.25	-0.056	0.51	-0.028	-0.034	1

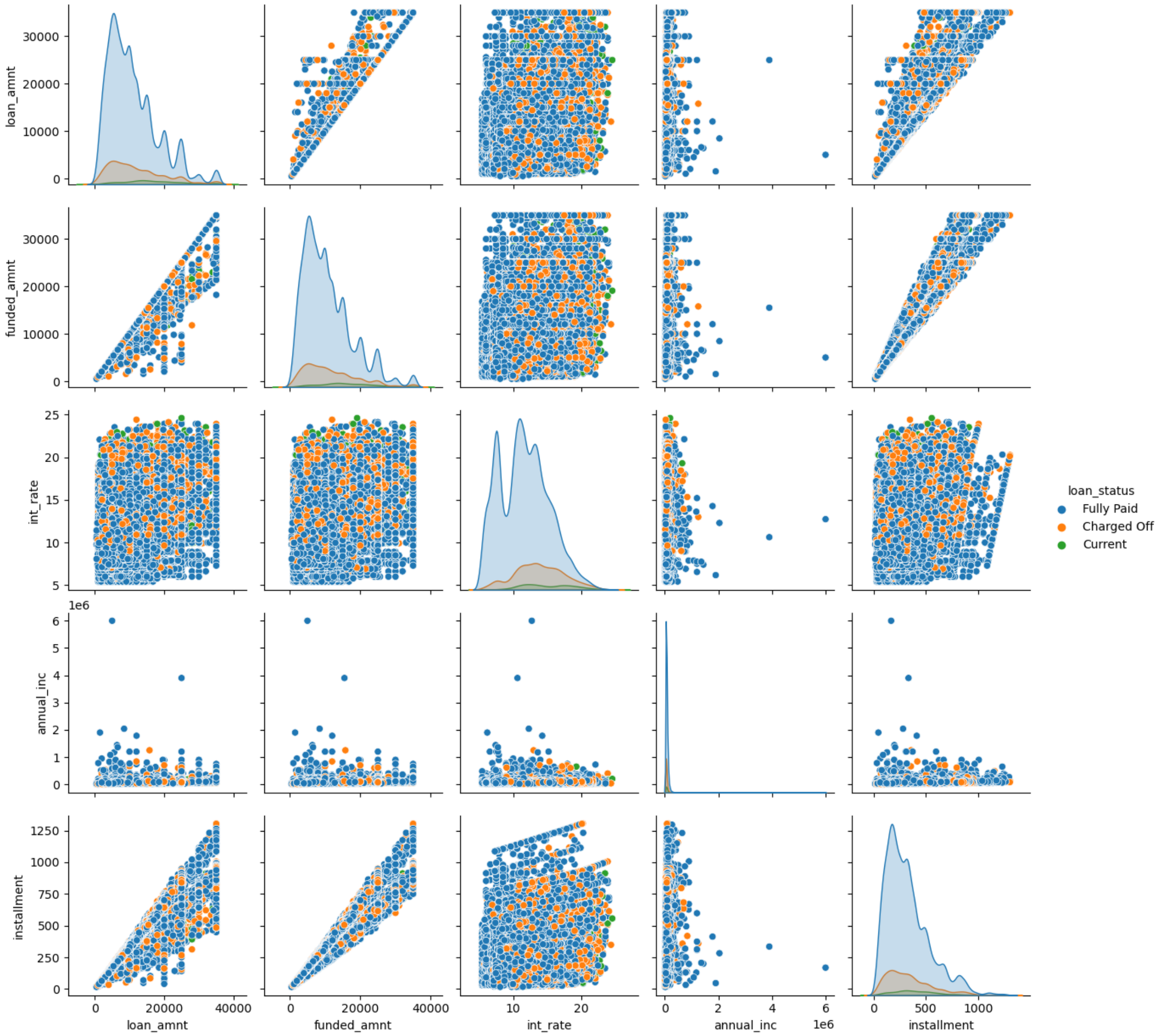
Correlation Analysis



	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	revol_bal	revol_util	total_acc	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	collection_recovery_fee	last_pymnt_amnt
loan_amnt	100.00%	98.19%	93.69%	29.12%	93.15%	40.32%	9.03%	-3.70%	0.25%	16.62%	31.30%	7.22%	23.91%	87.67%	84.16%	83.87%	72.71%	4.18%	14.50%	7.54%	46.93%
funded_amnt	98.19%	100.00%	95.49%	29.62%	95.71%	39.90%	8.95%	-3.66%	0.31%	16.42%	30.50%	7.64%	23.39%	89.33%	85.82%	85.65%	73.56%	4.42%	14.71%	7.76%	47.41%
funded_amnt_inv	93.69%	95.49%	100.00%	28.88%	90.26%	38.24%	9.71%	-4.32%	-1.18%	15.17%	28.55%	8.10%	22.64%	86.88%	90.43%	83.09%	72.44%	2.38%	13.43%	6.28%	46.64%
int_rate	29.12%	29.62%	28.88%	100.00%	26.92%	5.40%	11.17%	15.53%	13.38%	-0.79%	8.07%	46.59%	-6.40%	28.00%	27.44%	16.33%	51.62%	9.00%	12.67%	6.80%	16.53%
installment	93.15%	95.71%	90.26%	26.92%	100.00%	39.82%	7.94%	-2.48%	0.15%	16.09%	30.54%	10.39%	21.24%	85.35%	81.12%	84.10%	64.08%	5.23%	12.40%	7.56%	40.80%
annual_inc	40.32%	39.90%	38.24%	5.40%	39.82%	100.00%	-7.26%	3.07%	3.03%	26.52%	39.79%	4.86%	38.66%	38.22%	36.81%	38.29%	27.29%	0.27%	4.09%	2.09%	21.98%
dti	9.03%	8.95%	9.71%	11.17%	7.94%	-7.26%	100.00%	-3.16%	0.82%	29.83%	27.08%	27.72%	24.59%	8.62%	9.22%	6.42%	12.15%	-1.11%	2.73%	1.14%	2.51%
delinq_2yrs	-3.70%	-3.66%	-4.32%	15.53%	-2.48%	3.07%	-3.16%	100.00%	0.94%	1.44%	-6.06%	-4.15%	6.75%	-2.87%	-3.55%	-4.31%	1.73%	3.03%	1.12%	1.18%	-1.44%
inq_last_6mths	0.25%	0.31%	-1.18%	13.38%	0.15%	3.03%	0.82%	0.94%	100.00%	9.19%	-2.66%	-6.76%	10.94%	-1.89%	-2.86%	-3.24%	2.02%	3.27%	1.70%	1.03%	2.18%
open_acc	16.62%	16.42%	15.17%	-0.79%	16.09%	26.52%	29.83%	1.44%	9.19%	100.00%	28.62%	-9.53%	68.41%	14.88%	13.88%	14.76%	11.03%	-1.94%	1.69%	0.45%	7.60%
revol_bal	31.30%	30.50%	28.55%	8.07%	30.54%	39.79%	27.08%	-6.06%	-2.66%	28.62%	100.00%	31.63%	31.03%	28.33%	26.74%	27.09%	23.56%	0.24%	4.63%	2.22%	12.21%
revol_util	7.22%	7.64%	8.10%	46.59%	10.39%	4.86%	27.72%	-4.15%	-6.76%	-9.53%	31.63%	100.00%	-7.59%	8.19%	8.50%	3.05%	19.80%	3.54%	4.94%	2.56%	-0.97%
total_acc	23.91%	23.39%	22.64%	-6.40%	21.24%	38.66%	24.59%	6.75%	10.94%	68.41%	31.03%	-7.59%	100.00%	20.72%	20.14%	21.29%	13.21%	-2.64%	2.37%	0.80%	16.06%
total_pymnt	87.67%	89.33%	86.88%	28.00%	85.35%	38.22%	8.62%	-2.87%	-1.89%	14.88%	28.33%	8.19%	20.72%	100.00%	96.75%	97.33%	82.07%	1.09%	3.14%	2.76%	51.16%
total_pymnt_inv	84.16%	85.82%	90.43%	27.44%	81.12%	36.81%	9.22%	-3.55%	-2.86%	13.88%	26.74%	8.50%	20.14%	96.75%	100.00%	93.93%	80.28%	-0.42%	2.48%	1.72%	50.15%
total_rec_prncp	83.87%	85.65%	83.09%	16.33%	84.10%	38.29%	6.42%	-4.31%	-3.24%	14.76%	27.09%	3.05%	21.29%	97.33%	93.93%	100.00%	68.10%	-2.27%	-9.40%	-5.95%	56.73%
total_rec_int	72.71%	73.56%	72.44%	51.62%	64.08%	27.29%	12.15%	1.73%	2.02%	11.03%	23.56%	19.80%	13.21%	82.07%	80.28%	68.10%	100.00%	7.30%	9.38%	3.96%	24.19%
total_rec_late_fee	4.18%	4.42%	2.38%	9.00%	5.23%	0.27%	-1.11%	3.03%	3.27%	-1.94%	0.24%	3.54%	-2.64%	1.09%	-0.42%	-2.27%	7.30%	100.00%	10.30%	9.69%	-6.07%
recoveries	14.50%	14.71%	13.43%	12.67%	12.40%	4.09%	2.73%	1.12%	1.70%	1.69%	4.63%	4.94%	2.37%	3.14%	2.48%	-9.40%	9.38%	10.30%	100.00%	79.42%	-7.17%
collection_recovery_fee	7.54%	7.76%	6.28%	6.80%	7.56%	2.09%	1.14%	1.18%	1.03%	0.45%	2.22%	2.56%	0.80%	2.76%	1.72%	-5.95%	3.96%	9.69%	79.42%	100.00%	-4.26%
last_pymnt_amnt	46.93%	47.41%	46.64%	16.53%	40.80%	21.98%	2.51%	-1.44%	2.18%	7.60%	12.21%	-0.97%	16.06%	51.16%	50.15%	56.73%	24.19%	-6.07%	-7.17%	-4.26%	100.00%

Pair Plots

Pair plots provide same information as that we find in the univariate analysis.



Summary and Observations

- ***The variables ‘loan_amnt’, ‘funded_amnt’, ‘installment’, and ‘funded_amnt_inv’, are highly correlated. Only one variable can be considered for model as the univariate analysis already showed that these variables do not have any significant impact.***

- ***The variables ‘grade’, ‘term’, ‘purpose’, and ‘int rate’ seems to have larger influence on the ‘loan_status’.***

- ***Rest of the variables have middling to low impact on the ‘loan_status’.***