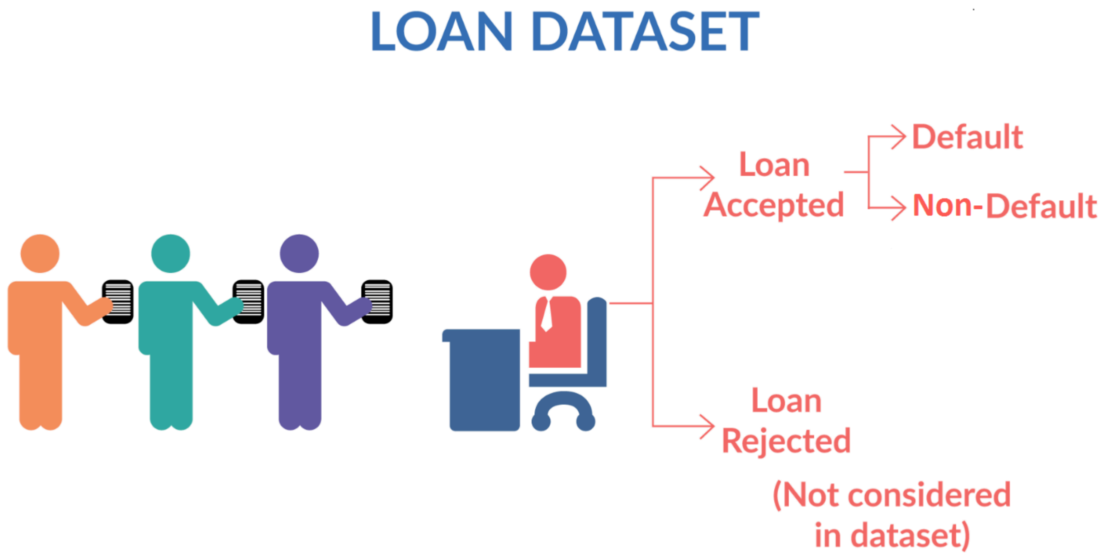
Lending Club Case Study

Submitted By —

Kunal Arneja
Pankaj Dixit

Introduction



When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company

- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

The data given below contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Exploratory Data Analysis

Exploratory data analysis is the first and foremost step to analyse any kind of data. Rather than a specific set of procedures, EDA is an approach, or a philosophy, which seeks to explore the most important and often hidden patterns in a data set. In EDA, we explore the data and try to come up with a hypothesis about it which we can later test using hypothesis testing. Statisticians use it to take a bird's eye view of the data and try to make some sense of it.

Steps Involved In EDA are:

1. Data sourcing

2. Data cleaning

3. Univariate analysis

4. Bivariate analysis

5. Derived metrics

Data Sourcing

To solve a business problem using analytics, you need to have historical data to come up with actionable insights. Data is the key — the better the data, the more insights you can get out of it.

Here, the dataset was shared with us in the assignment. So it was a **private source** of data.

Deleted Columns with no values(always null)

num_op_rev_tl	num_op_rev_tl	100.000000
num_rev_accts	num_rev_accts	100.000000
num_rev_tl_bal_gt_0	num_rev_tl_bal_gt_0	100.000000
num_sats	num_sats	100.000000
num_tl_120dpd_2m	num_tl_120dpd_2m	100.000000
num_tl_30dpd	num_tl_30dpd	100.000000
num_tl_90g_dpd_24m	num_tl_90g_dpd_24m	100.000000
pct_tl_nvr_dlq	pct_tl_nvr_dlq	100.000000
percent_bc_gt_75	percent_bc_gt_75	100.000000
tot_hi_cred_lim	tot_hi_cred_lim	100.000000
total_bal_ex_mort	total_bal_ex_mort	100.000000
num_il_tl	num_il_tl	100.000000
mo_sin_rcnt_rev_tl_op	mo_sin_rcnt_rev_tl_op	100.000000
verification_status_joint	verification_status_joint	100.000000
mo_sin_old_il_acct	mo_sin_old_il_acct	100.000000
next_pymnt_d	next_pymnt_d	100.000000
mths_since_last_major_derog	mths_since_last_major_derog	100.000000
annual_inc_joint	annual_inc_joint	100.000000
dti_joint	dti_joint	100.000000
total_bc_limit	total_bc_limit	100.000000
tot_coll_amt	tot_coll_amt	100.000000
tot_cur_bal	tot_cur_bal	100.000000
open_acc_6m	open_acc_6m	100.000000
open_il_6m	open_il_6m	100.000000
open_il_12m	open_il_12m	100.000000
open_il_24m	open_il_24m	100.000000
mths_since_rcnt_il	mths_since_rcnt_il	100.000000
mo_sin_old_rev_tl_op	mo_sin_old_rev_tl_op	100.000000
total_bal_il	total_bal_il	100.000000
open_rv_12m	open_rv_12m	100.000000
open_rv_24m	open_rv_24m	100.000000
max_bal_bc	max_bal_bc	100.000000
all_util	all_util	100.000000
total_rev_hi_lim	total_rev_hi_lim	100.000000
inq_fi	inq_fi	100.000000
total_cu_tl	total_cu_tl	100.000000
inq_last_12m	inq_last_12m	100.000000
acc_open_past_24mths	acc_open_past_24mths	100.000000
avg_cur_bal	avg_cur_bal	100.000000
bc_open_to_buy	bc_open_to_buy	100.000000
bc_util	bc_util	100.000000
il_util	il_util	100.000000
total_il_high_credit_limit	total_il_high_credit_limit	100.000000

Deleted Columns with all rows having same value

```
17                  pymnt_plan
33         initial_list_status
34                   out_prncp
35               out_prncp_inv
46  collections_12_mths_ex_med
47                 policy_code
48            application_type
49              acc_now_delinq
50     chargeoff_within_12_mths
51                 delinq_amnt
53                   tax_liens
Name: Variable, dtype: object
```

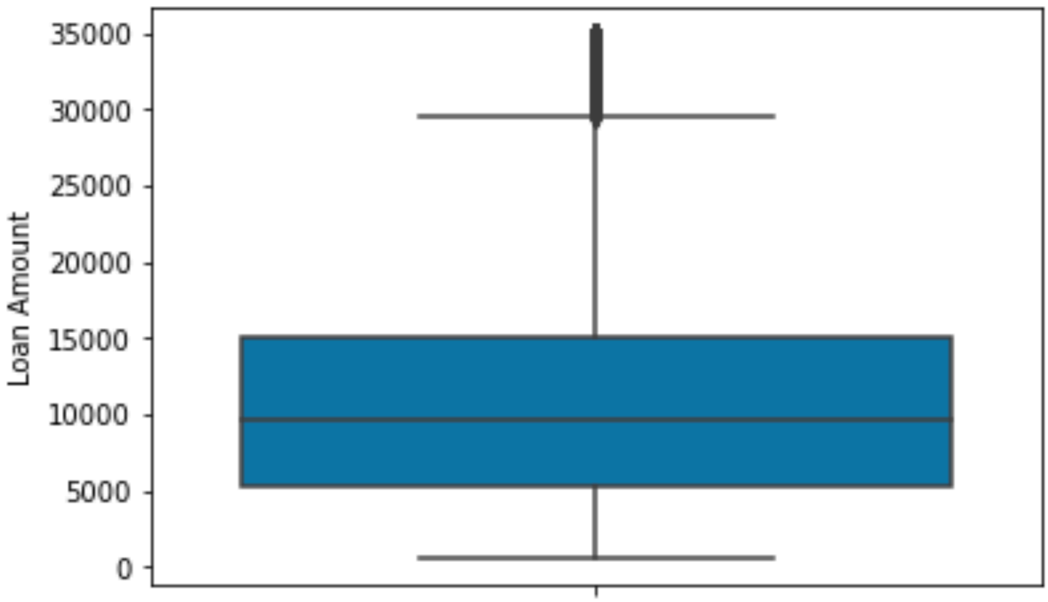
Insights from above

1. A lot of features are having a single value, hence can be dropped.

Deleted id related columns with all values as distinct value

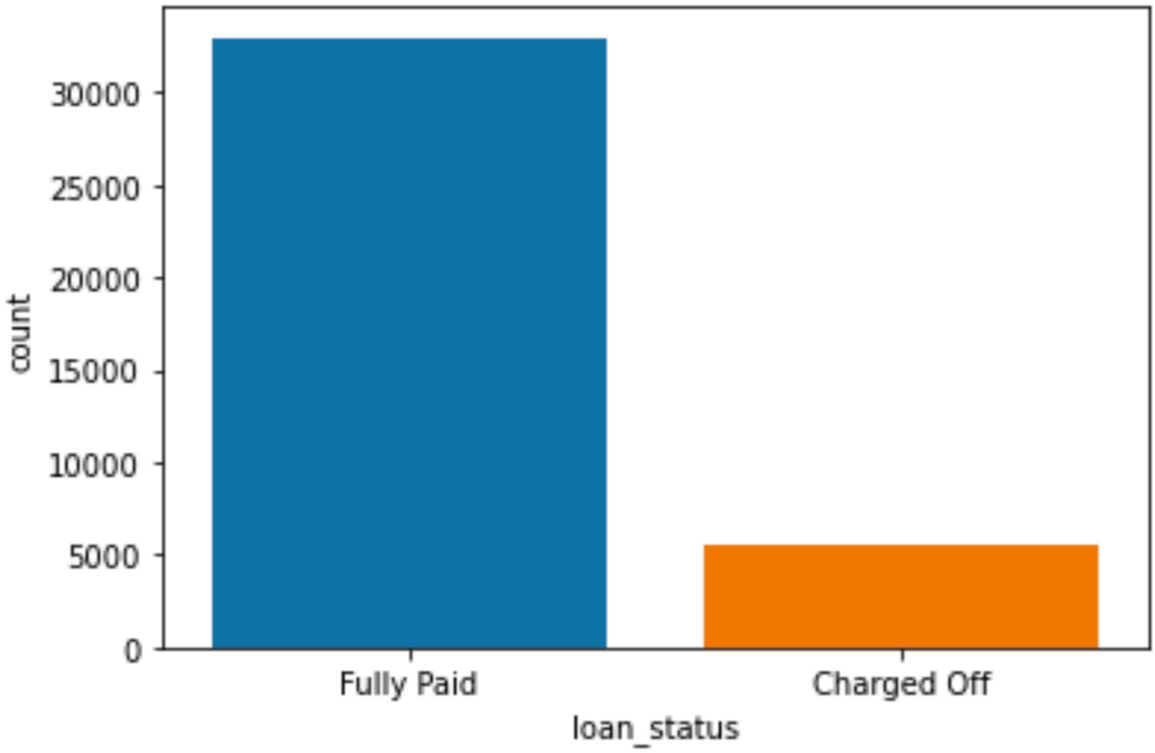
Univariate Analysis

Loan Amount



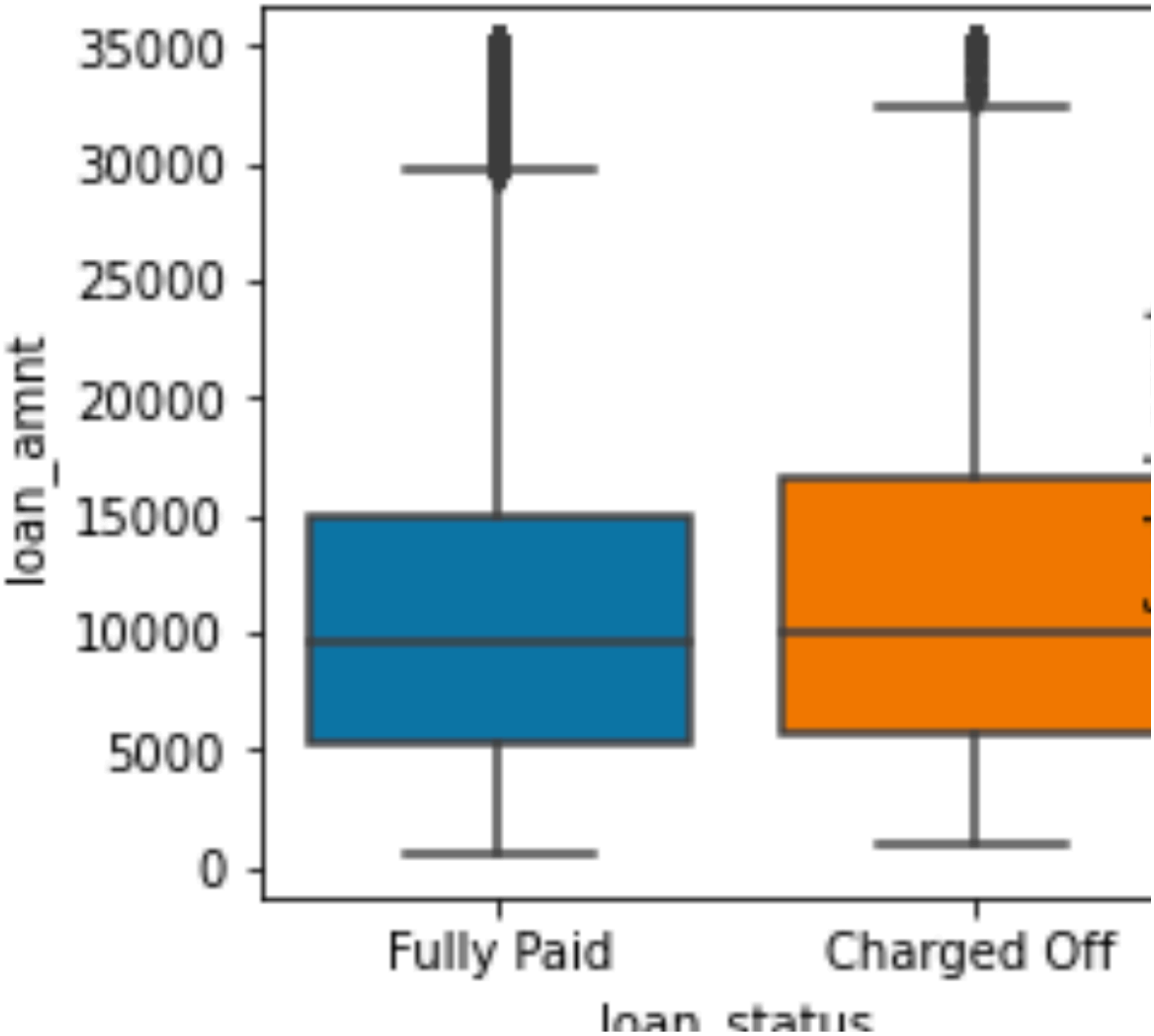
The loan amount varies from 0 to 35,000 having mean of 10,000

Loan Status



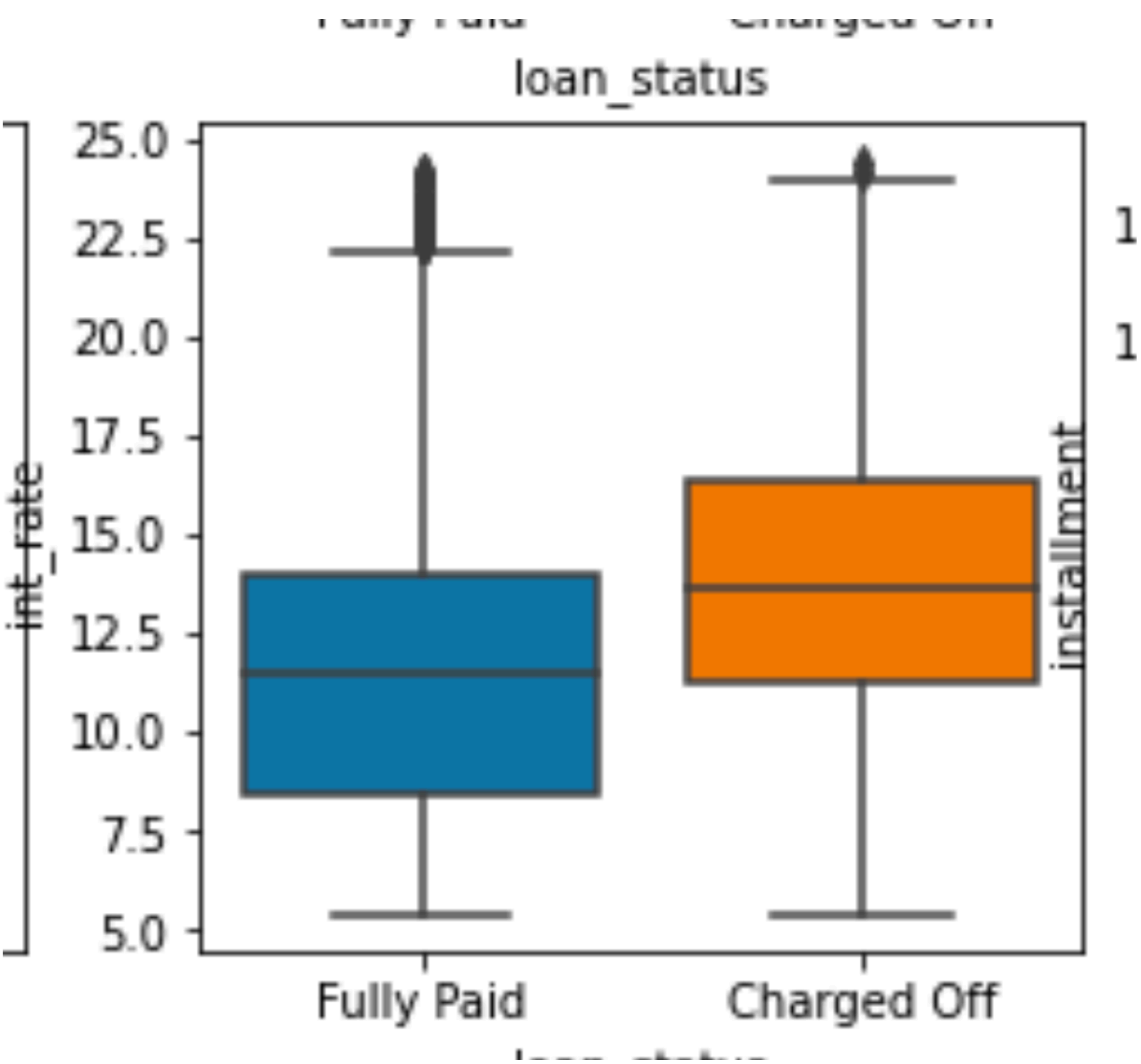
86% loans are fully paid while only 14% are defaulted.
There are class imbalances present.

Loan Amount vs Loan Status



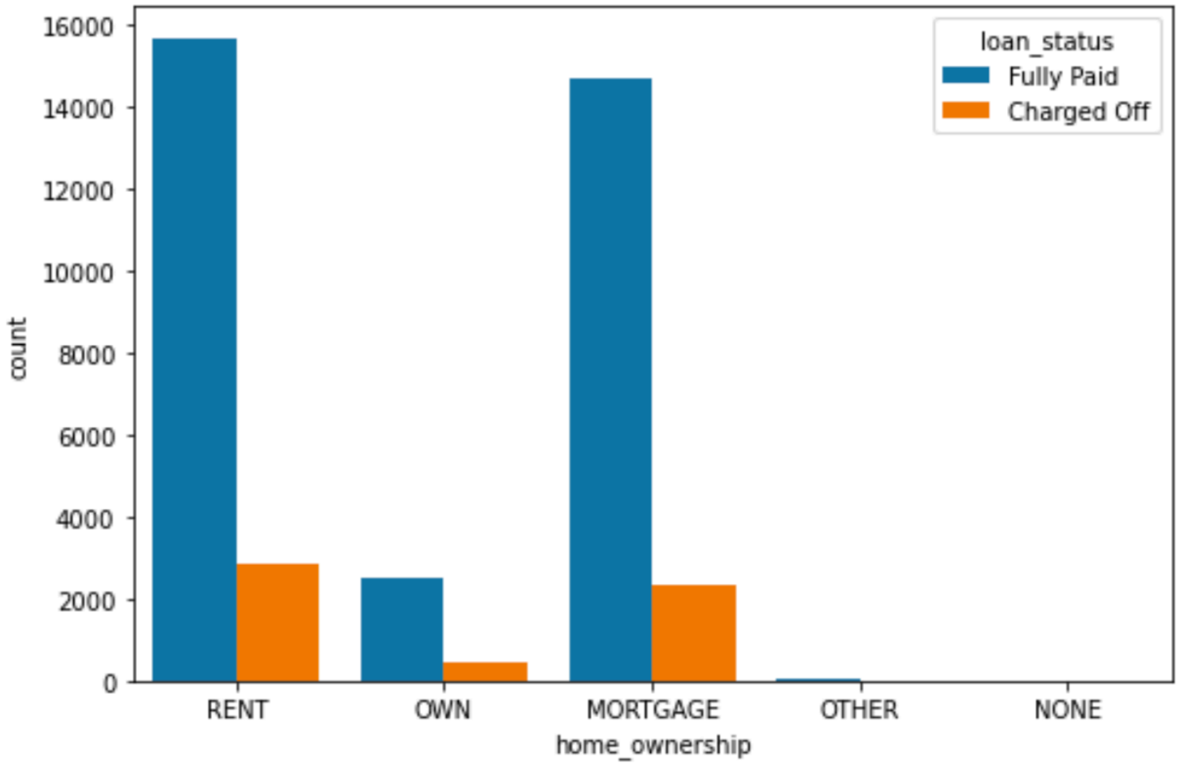
Charged Off Loans have slightly higher loan amounts.

Interest Rate vs Loan Status



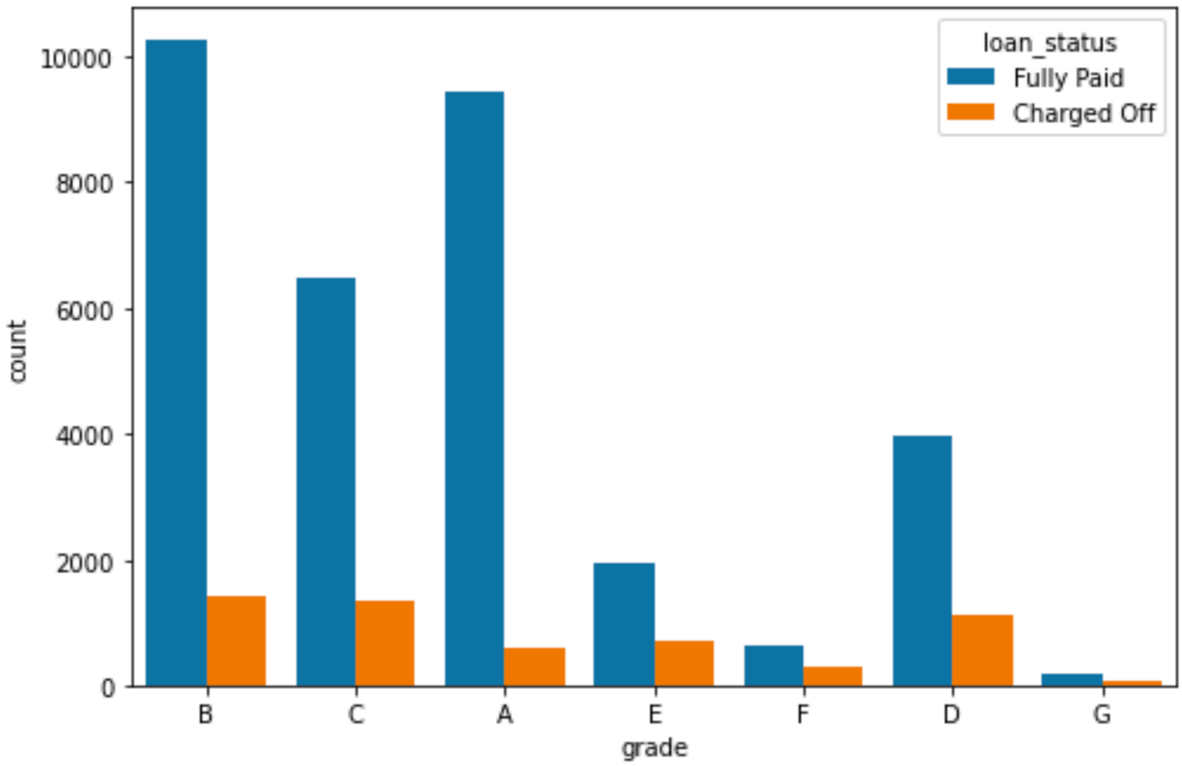
Higher Installments have significantly higher chances of being charged off.

Purpose vs Loan Status



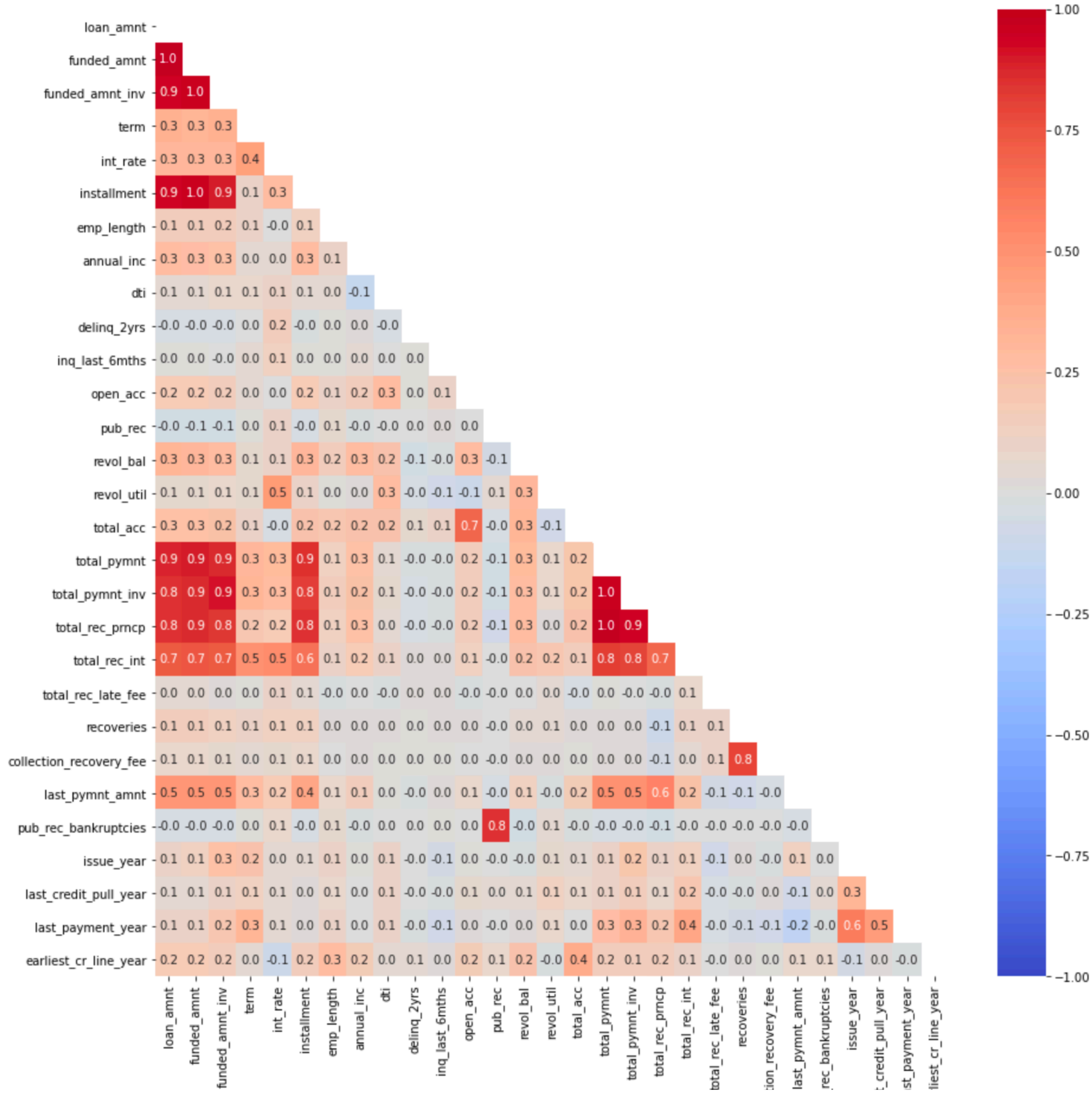
Most loand are for category RENT, OWN, MORTGAGE. RENT and MORTGAGE have higher charged off to paid ratio compared to OWN.

Grade vs Loan Status



Most loans are of class A,B and C. D has highest charged off to paid ratio followed by E, C, B and A.

Bivariate Analysis



1. Loan Amount, Funded Amount, Funded Amount invested and installment are highly correlated with each other.
2. Annual Income is negatively correlated with DTI.