Assignment : Lending Club Case Study

Team : Debasish Mondal & Sharath

Date : 05 APR 2023

Business Understanding

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return. 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

.

Objective

The main objective is to be able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Company wants to understand the factor behind loan default (loan_status = 'Charged Off')

There are four major parts that are needed to be done for this case study:

1. Data understanding

2. Data cleaning (cleaning missing values, removing redundant columns etc.)

3. Data Analysis

4. Recommendations

Out[72]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_past_
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	I
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	I
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	I
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	I
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	I

5 rows × 111 columns

Data cleaning

1. cleanup of NULLs from rows and columns of the loan dataframe
2. cleanup of NA record

In [13]:
```
#looking for unique column if any
loan_column_unique = loan_data.nunique()
print(loan_column_unique)
```

```
id                          39717
member_id                   39717
loan_amnt                     885
funded_amnt                  1041
funded_amnt_inv              8205
term                            2
int_rate                      371
installment                 15383
grade                           7
sub_grade                      35
emp_title                   28820
emp_length                     11
home_ownership                  5
annual_inc                   5318
verification_status             3
issue_d                        55
loan_status                     3
pymnt_plan                      1
url                         39717
purpose                        14
title                       19615
zip_code                      823
addr_state                     50
dti                          2868
delinq_2yrs                    11
earliest_cr_line              526
inq_last_6mths                  9
open_acc                       40
```

3.

Dropped column of unique value ;

Data preparation

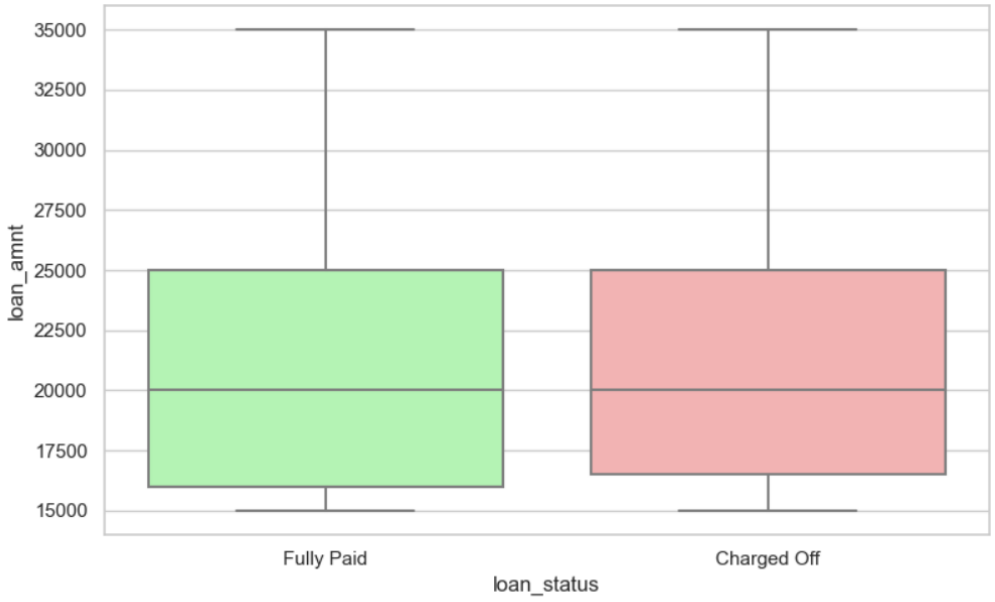
Univariate Analysis

- **What is/are the main feature(s) of interest in your dataset?**

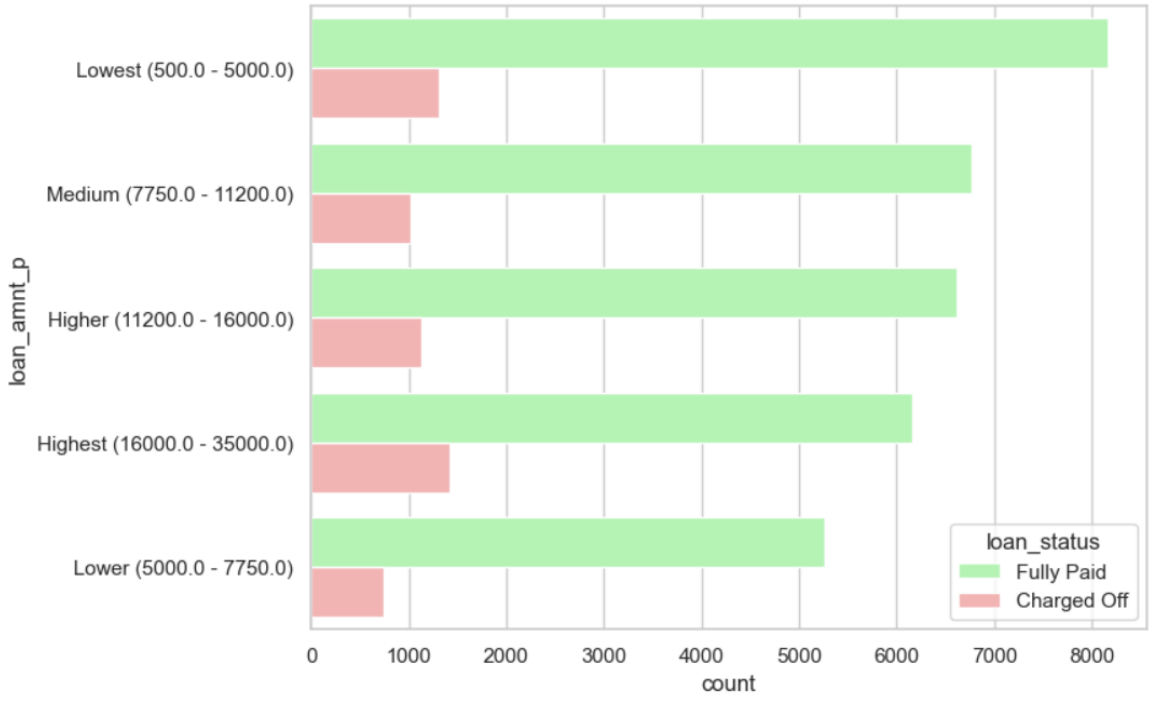
A number of columns with Factor type and a the rest are continous or discrete form of numerical values. Among the numeric fields, the Loan Amount, Annual Income, Interest Rate are of particular interest.Of all the categorical fields (Factors), Home ownership, Loan Status, Loan Grade, Term, are interesting.

understanding the correlation between the different numeric fields and see if they are related (high correlation values)

Loan status vs Numerical continuous variables: compare the loan_status fields with all the numerical variable.

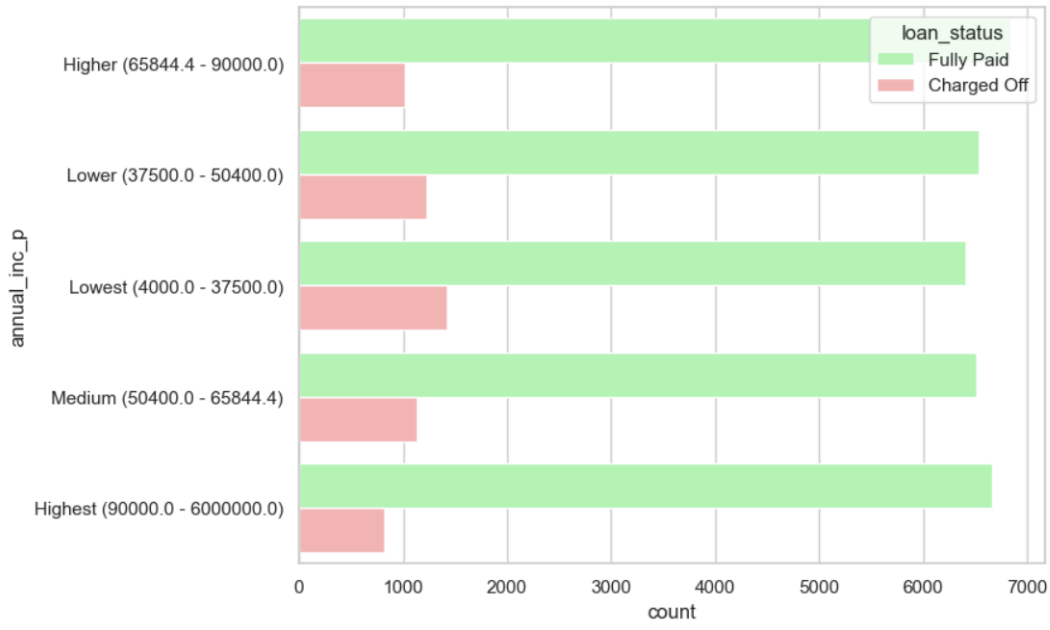


from above box plot we can conclude that "heigher the amount" will tend to "Write off"

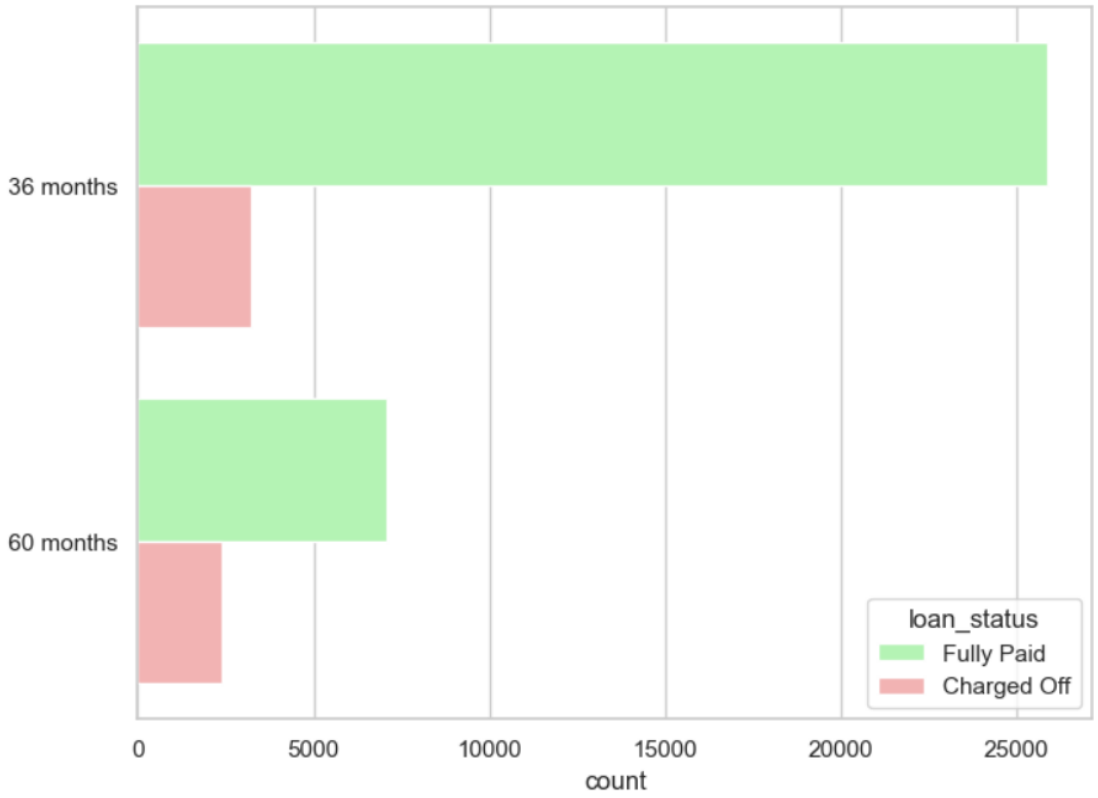


loan_amnt_p	Charged off %	Record count
Highest (16750.0 - 35000.0)	0.175706	7928
Higher (12000.0 - 16750.0)	0.144495	5668

conclusion Higher the loan amount, greater the chance of the loan getting default.



conclution Higher the income hiegher the repayment %



year repayment term, the default percent is 25%. And for 3 year loan repayment term, the default is only for 11% of the cases

Bivariate and Multivariate Plots Section

- **Based on what you saw in the univariate plots, what relationships between variables might be interesting to look at in this section?**
- I was particularly interested in the relationship between `loan amount` and the following fields:
- Home Ownership
- Loan Grade
- Loan Status
- Interest Rate
- Annula Income
- Income-to-Loan-ratio

1.Applicants with high incomes should have more chances of loan approval.

2.Applicants who have repaid their previous debts should have higher chances of loan approval.

3.Loan approval should also depend on the loan amount. If the loan amount is less, the chances of loan approval should be high.

4.Lesser the amount to be paid monthly to repay the loan, the higher the chances of loan approval.

#Let’s try to test the above-mentioned hypotheses using bivariate analysis

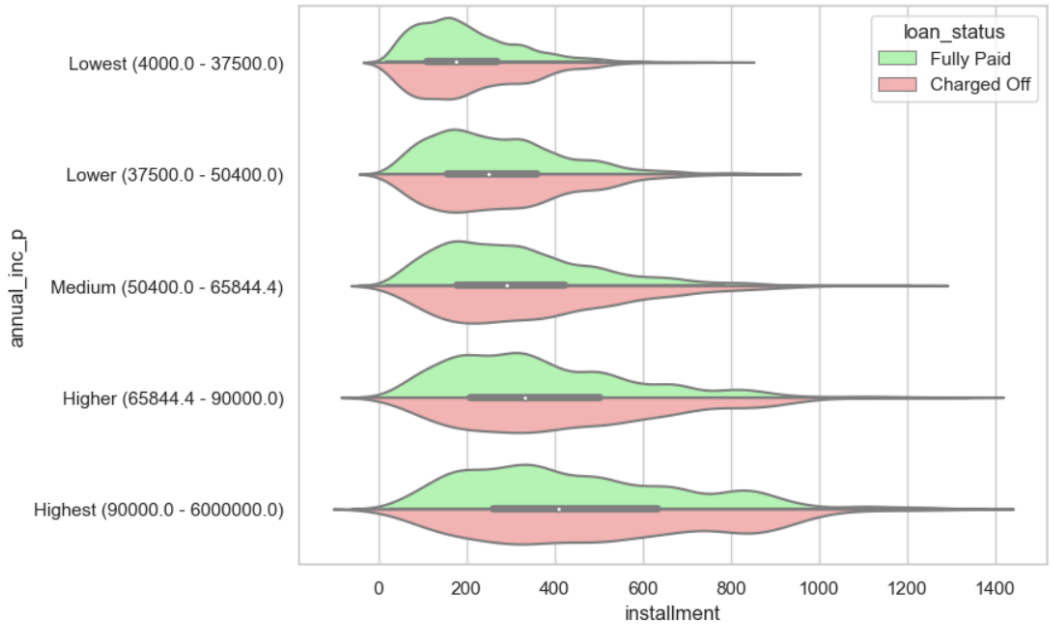


figure shows that for higher installments for any income group have more number of defaults.

Hereby we come to an end of the EDA of the loan data set and finding some of the drivers for loan default

.Higher loan amount (above 16K

•Higher installment amount)

- Lower annual income
- Higher interest rate (above 13%)
- Repayment term (5 years)
- Loan grade & sub-grade
- Missing employment record
- Loan purpose (small business, renewable energy, educational)