Lending Club: EDA Case Study

By:

Debasish Mondal
Sharath Chandra

Problem Statement

The primary objective of this case study is to identify the key driver variables that influence loan default, enabling the consumer finance company to make informed decisions on loan approvals and risk assessment. By conducting an Exploratory Data Analysis (EDA), we aim to discover patterns and relationships in the data that can help the company minimize credit loss, manage a healthy loan portfolio, and optimize their lending strategies.

Specifically, the case study will focus on:

1. Understanding how consumer attributes and loan attributes influence the likelihood of default.
2. Identifying the variables that are strong indicators of loan default.
3. Providing actionable insights and recommendations for the company's portfolio and risk assessment processes.

Solution approach summary

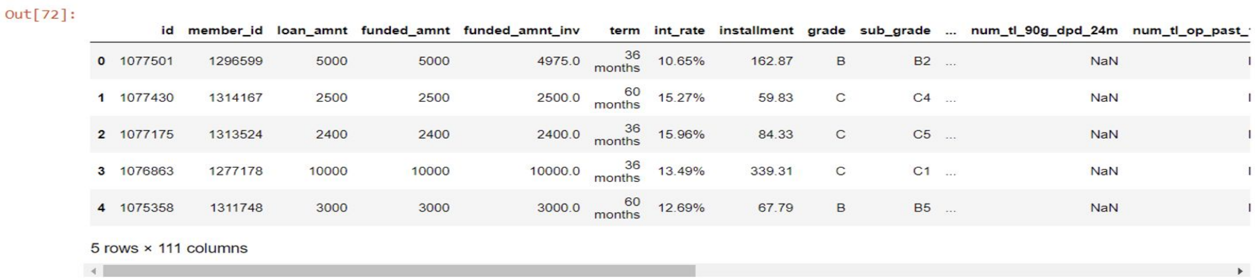
- **Data Cleaning:**
  - Import the dataset and perform preliminary data cleaning, including handling missing values, dropping irrelevant columns, and converting data types as needed.
- **Univariate Analysis:**
  - Analyze the distribution of each variable in the dataset using appropriate visualization techniques.
- **Bivariate Analysis:**
  - Examine relationships between pairs of variables to understand their associations with loan default.
- **Feature Engineering:**
  - Create new variables or transform existing ones to better represent the information in the dataset.

Solution approach summary

- **Multivariate Analysis:**
    - Investigate the relationships among multiple variables simultaneously using advanced techniques such as PCA, clustering, or regression analysis.
- **Identifying Key Driver Variables:**
    - Determine the variables that have the strongest relationship with loan default using correlation coefficients, statistical tests, or machine learning algorithms.
- **Interpretation and Recommendations:**
    - Summarize the findings from the analysis and provide actionable insights to help the company make informed decisions on loan approvals and risk assessment.

Data summary

- Data contains 39717 rows and 111 columns where each row contains data related to user’s each loan request
- Following are the important variables which we selected for our analysis and ignored others which may not provide info to the analysis
  Loan_amnt, term, interest_rate, grade, subgrade, annual_income, loan_purpose, debt_to_income, emp_length, loan_date, home_ownership, verification_status



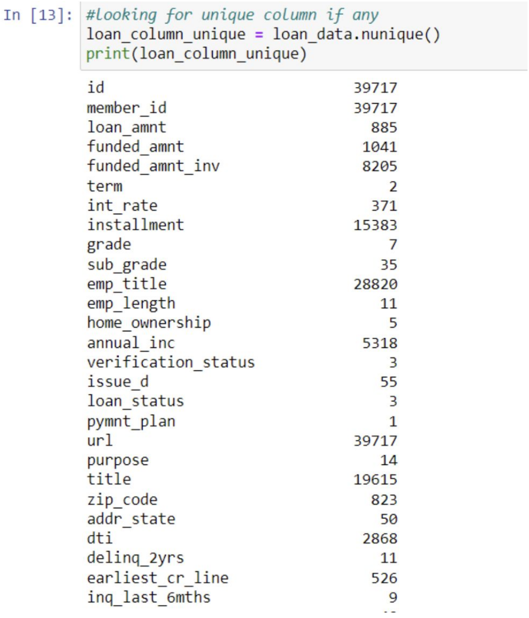
Out[72]:

	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
3. Dropped column of unique value ;



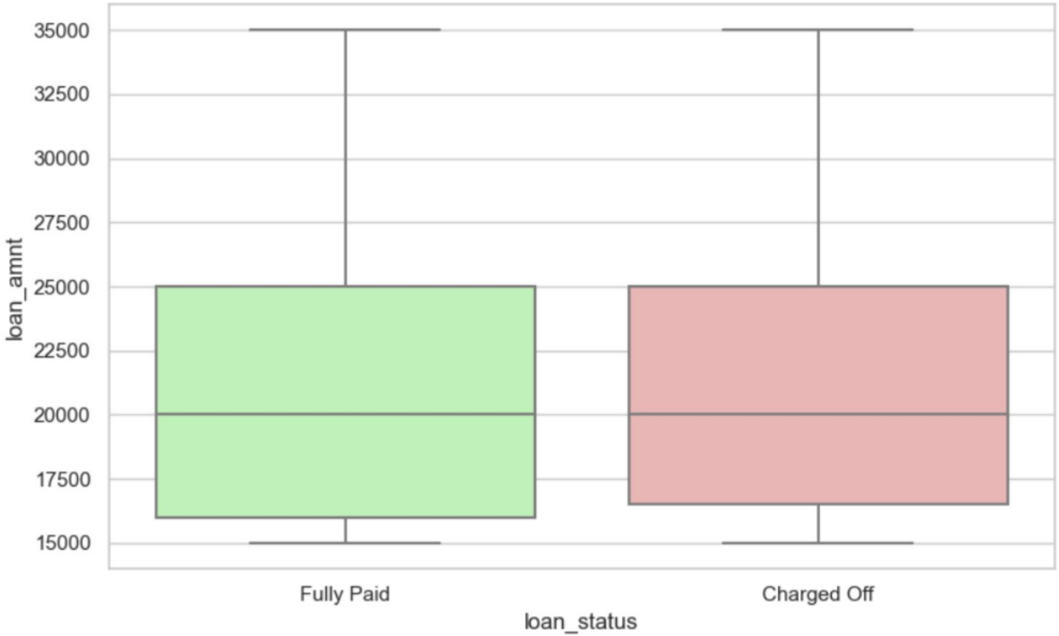
```
In [13]: #looking for unique column if any
         loan_column_unique = loan_data.nunique()
         print(loan_column_unique)

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
```

Univariate Analysis

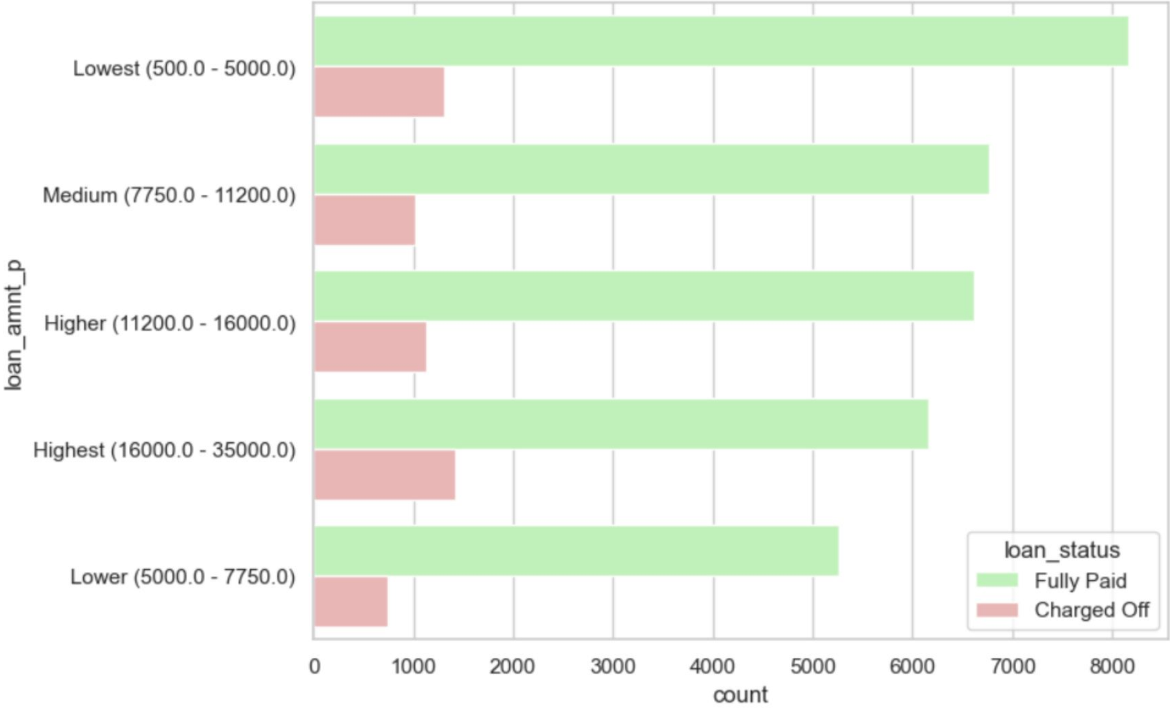
What is/are the main feature(s) of interest in your dataset?

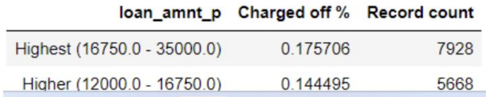
- A number of columns with Factor type and a the rest are continous or discrete form of numerical values. Among the numeric fields, the Loan Amount, Annual Income, Interest Rate are of particular interest.Of all the categorical fields (Factors), Home ownership, Loan Status, Loan Grade, Term, are interesting.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

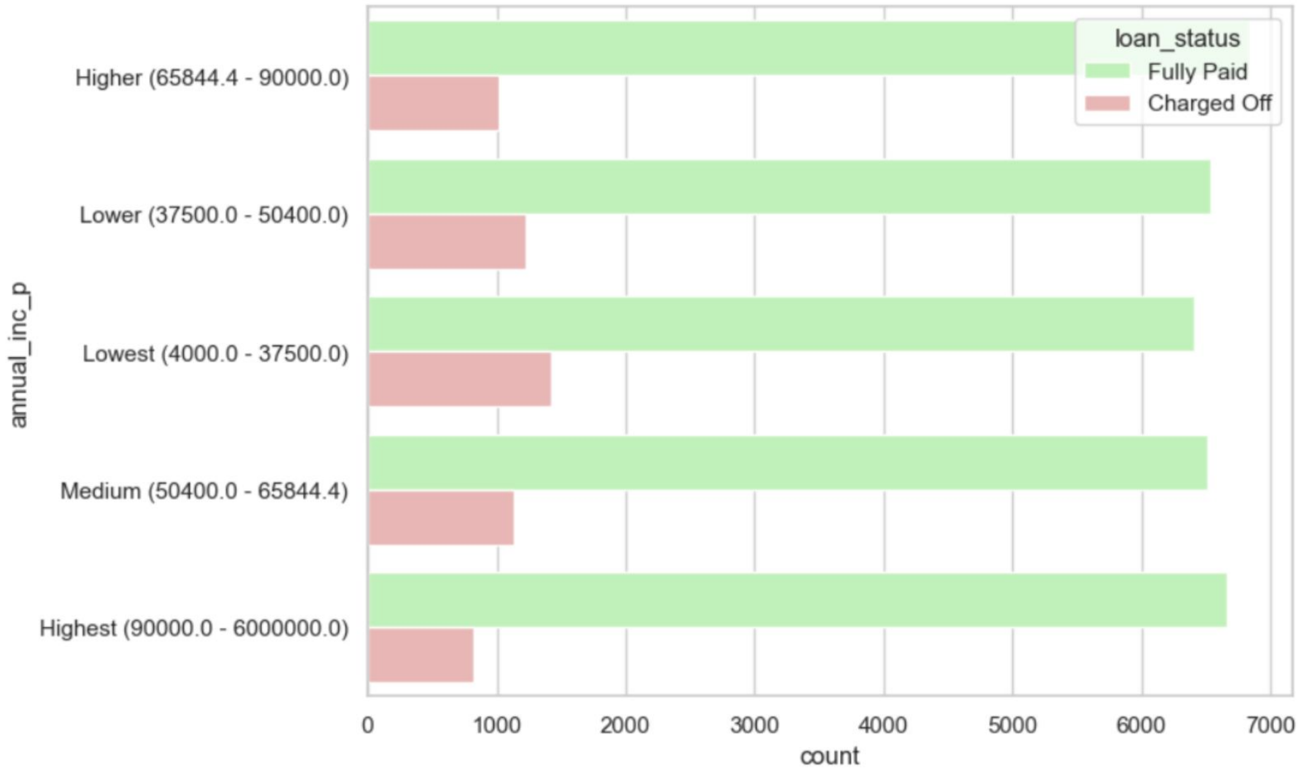
- From below plot, we can conclude that Higher the loan amount, greater the chance of the loan getting default.

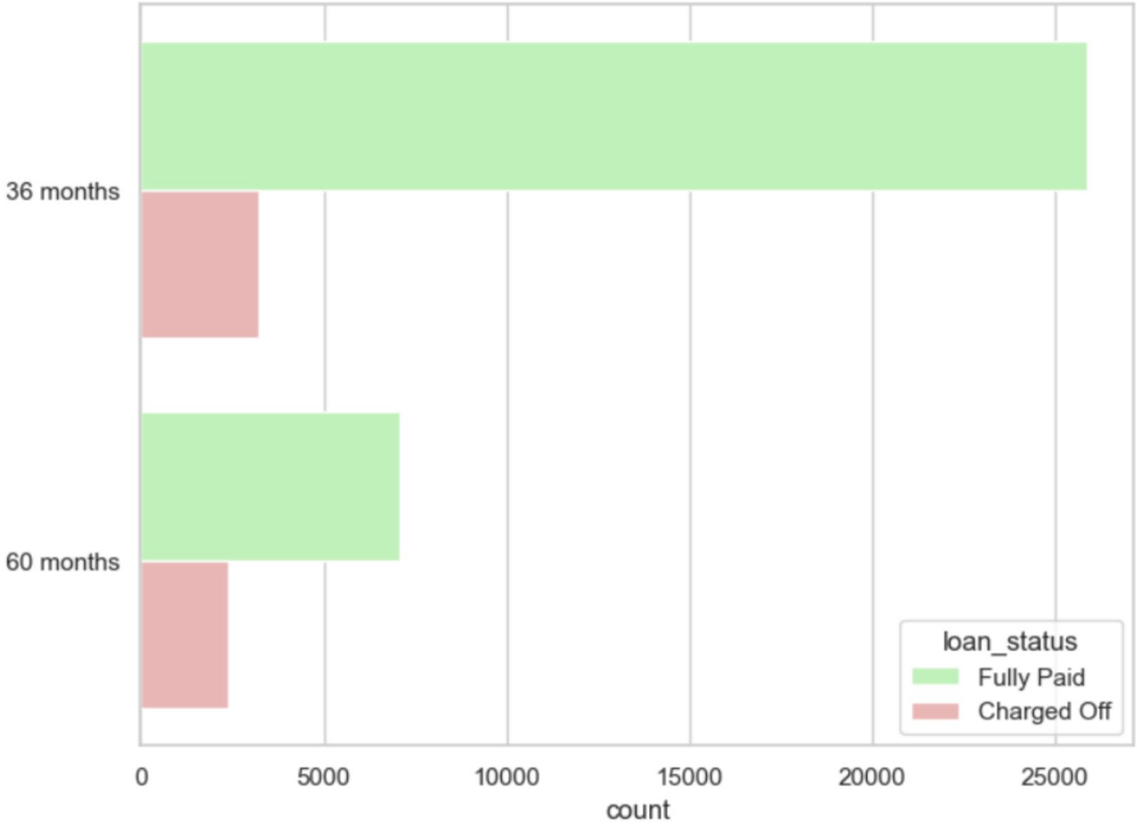




loan_amnt_p	Charged off %	Record count
Highest (16750.0 - 35000.0)	0.175706	7928
Higher (12000.0 - 16750.0)	0.144495	5668

- From below plot, we can conclude that Higher the income higher the repayment percentage





- 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

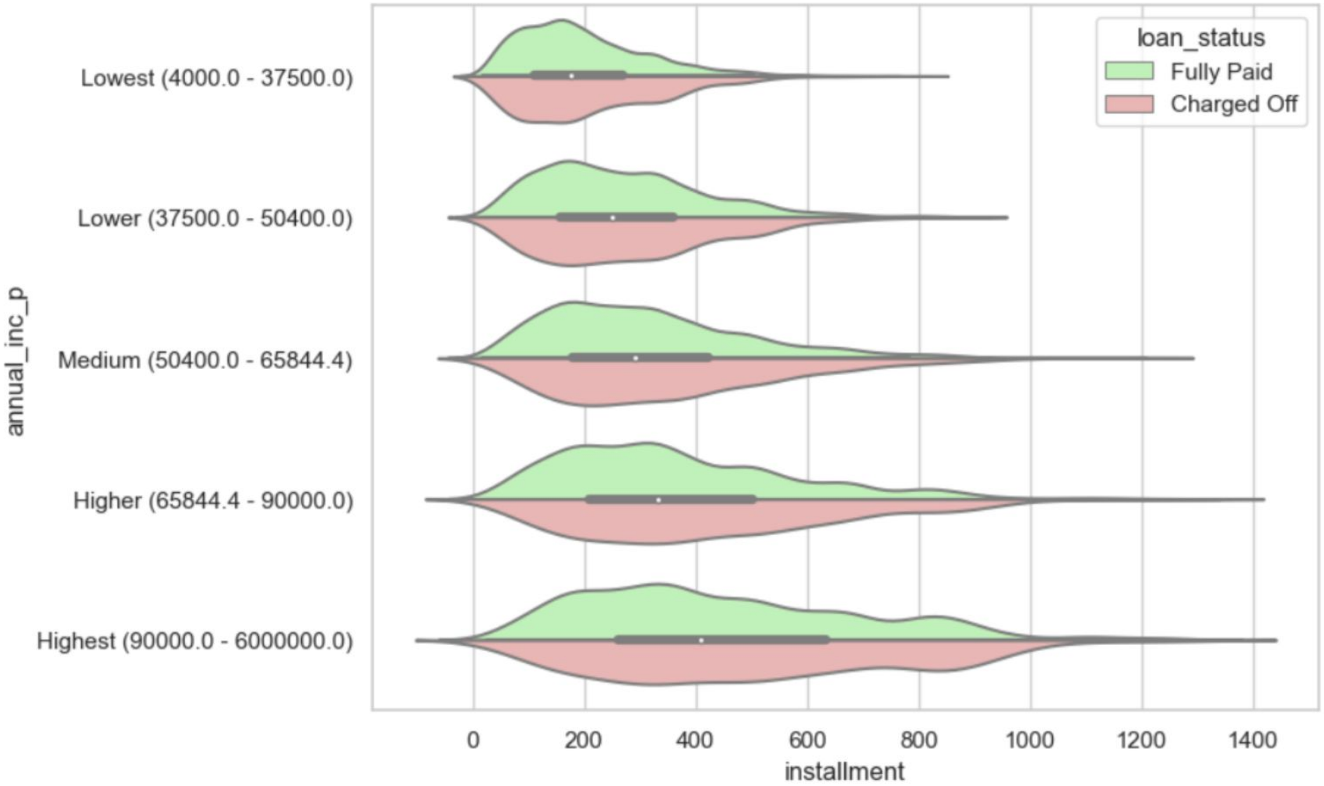
Based on what you saw in the univariate plots, what relationships between variables might be interesting to look at in this section?

I was particularly interested in the relationship between loan amount and the following fields:

- Home Ownership
- Loan Grade
- Loan Status
- Interest Rate
- Annual 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



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

Conclusion

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)