PREDICTING BANK LOAN APPROVAL BASED ON RISK

DSCI 550 TEAM 6

RAVISH KAMATH | EDDIE ZHAO | JACKSON CRAWFORD | CAMERON ELA

BUSINESS PROBLEM

- We are working in a modern commercial bank as a Data Scientist where we built regression models.
- We build regression models for predicting the probability that those loans would be defaulted on. However, we realize that these models do not exactly work.
- TASK:
  - Build a more efficient default probability model that can be used in production for deciding whether a customer should receive a loan or not.

DATA SET AND DESCRIPTION

	BrowseNotesFile	Description
0	loanAmnt	The listed amount of the loan applied for by t...
1	annualInc	The self-reported annual income provided by th...
2	application_type	Indicates whether the loan is an individual ap...
3	avg_cur_bal	Average current balance of all accounts
4	chargeoff_within_12_mths	Number of charge-offs within 12 months
5	delinq2Yrs	The number of 30+ days past-due incidences of ...
6	dti	A ratio calculated using the borrower's total ...
7	emp_length	Employment length in years. Possible values ar...
8	grade	LC assigned loan grade
9	homeOwnership	The home ownership status provided by the borr...
10	inq_last_12m	Number of credit inquiries in past 12 months
11	installment	The monthly payment owed by the borrower if th...
12	job	Job Description
13	loanAmnt	The listed amount of the loan applied for by t...
14	loanDefault	0: Loan was uptimated paid in full. 1: A defau...
15	mortAcc	Number of mortgage accounts.
16	num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in...
17	pub_rec_bankruptcies	Number of public record bankruptcies
18	purpose	A category provided by the borrower for the lo...
19	term	The number of payments on the loan. Values are...
20	Year	Year of Issue of the loan

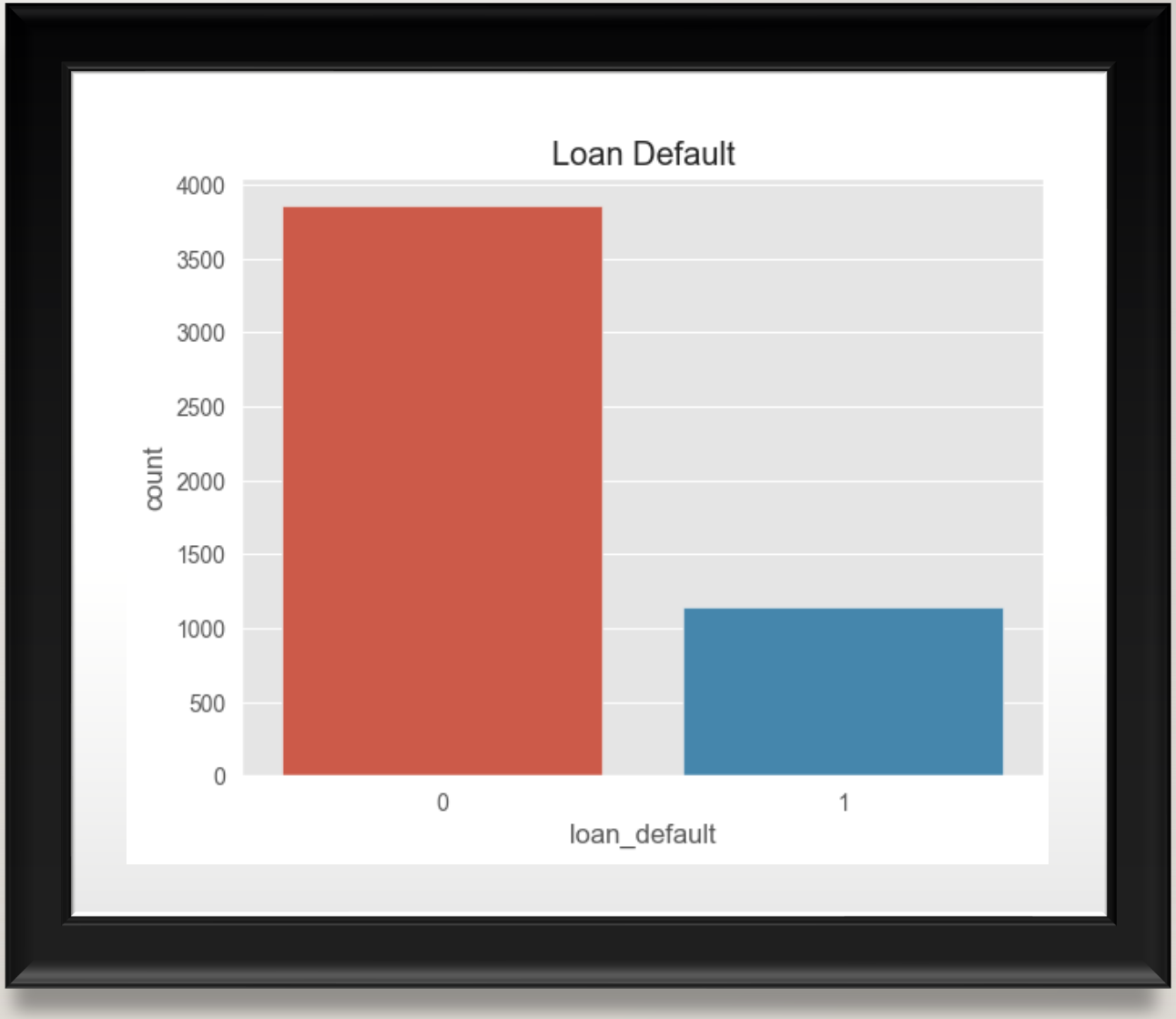
ATTRIBUTES:

- Loan Default **(VARIABLE OF INTEREST)**
- Loan Amount
- Annual Application Type
- Average Current Balance
- Charge-offs within 12 months
- Delinq 2Yrs
- DTI
- Employment Length
- Grade (Loan)
- Home Ownership
- Number of Credit Inquires
- Installment
- Job
- # of Mortgage Accounts
- # of Accounts 90 or more days due
- # of public record bankruptcies
- Purpose
- Term
- Year

EXPLORING THE DATA

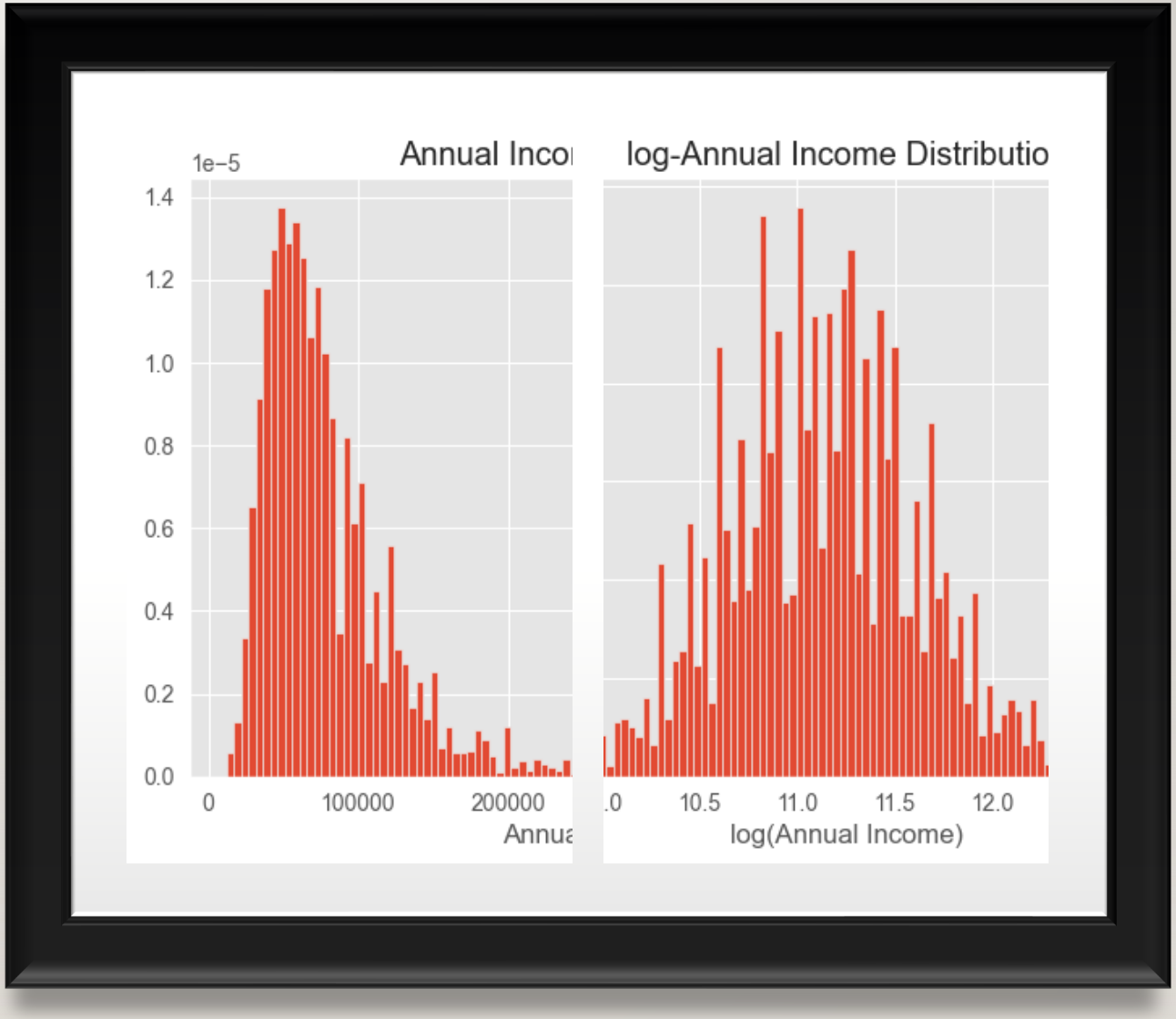
TARGET VARIABLE

- We can see that around 20-25 percent of all loans in the data set are defaults
- Due to the biasness towards non-defaults, we need to consider resampling our dataset to create equal amount of each category.
- Using F1 Score for deciding the success of a model would make sense.



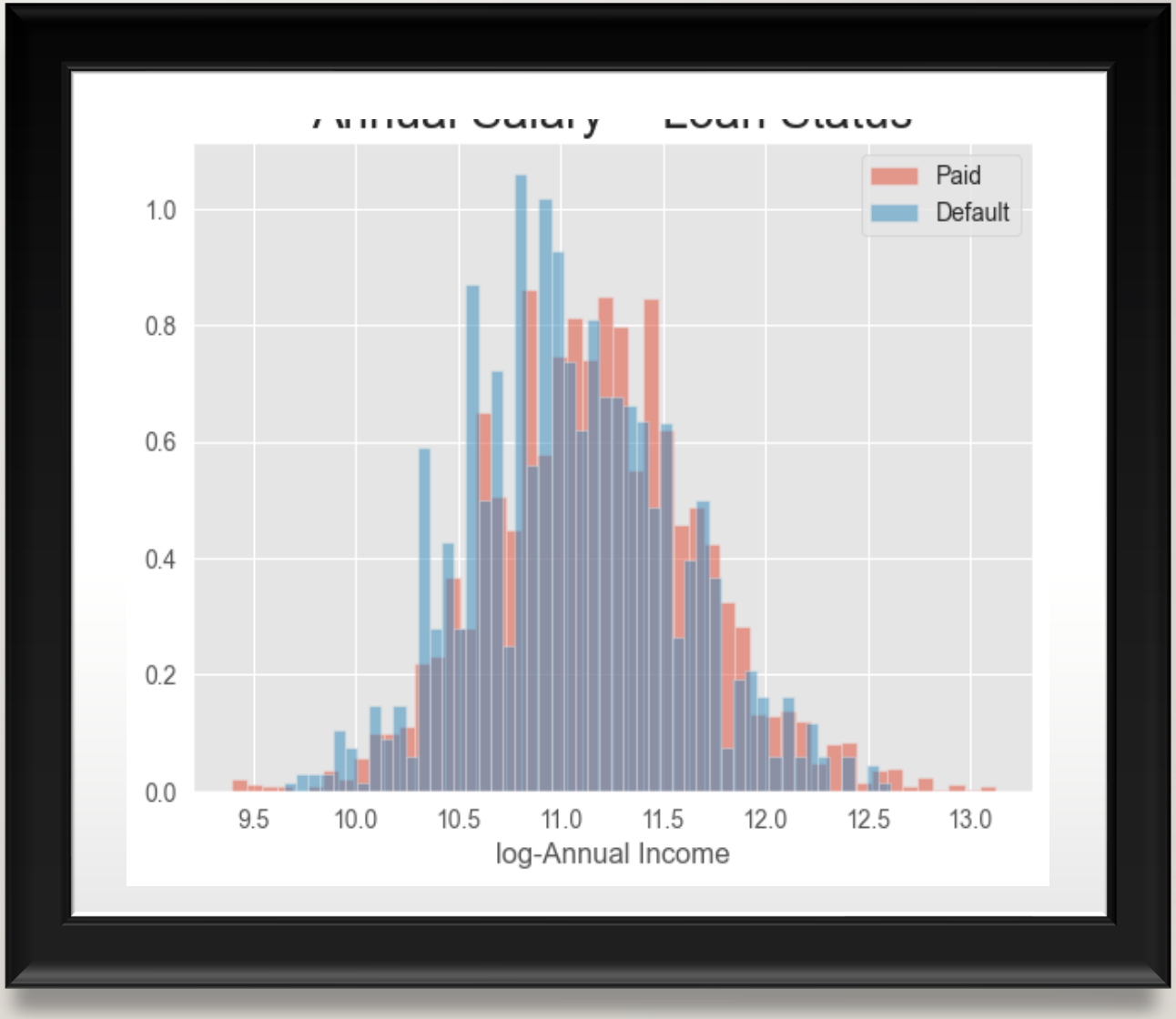
ANNUAL INCOME

- We can see in the left graph that annual income is skewed left.
- We can proceed to do a log transformation, to create a more normally distributed data for annual income.



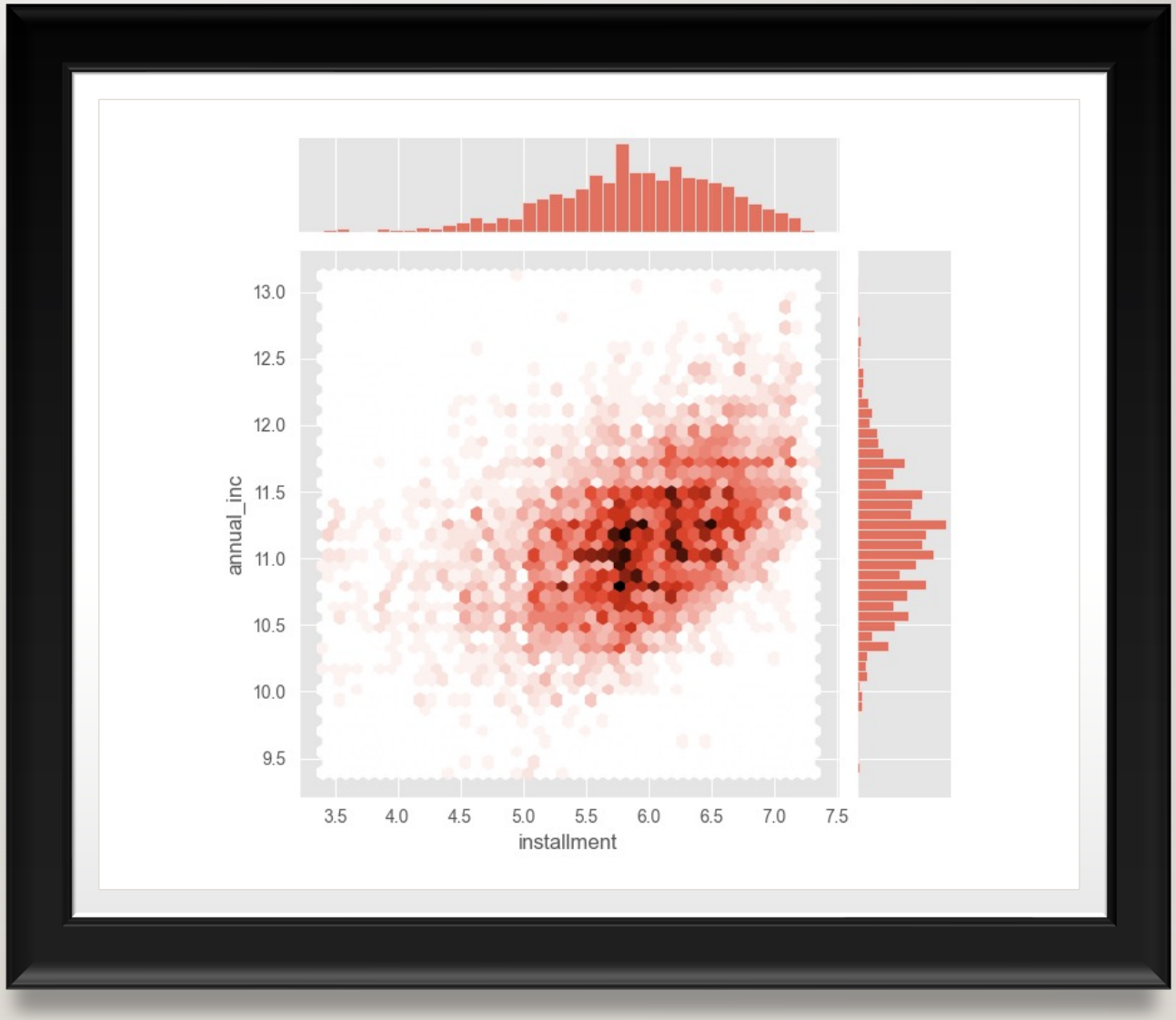
ANNUAL INCOME ON DEFAULT STAT

- We can see that the distribution for both the groups in terms of their annual income are similarly distributed.
- This can indicate that income is not likely to explain the difference in loan status.



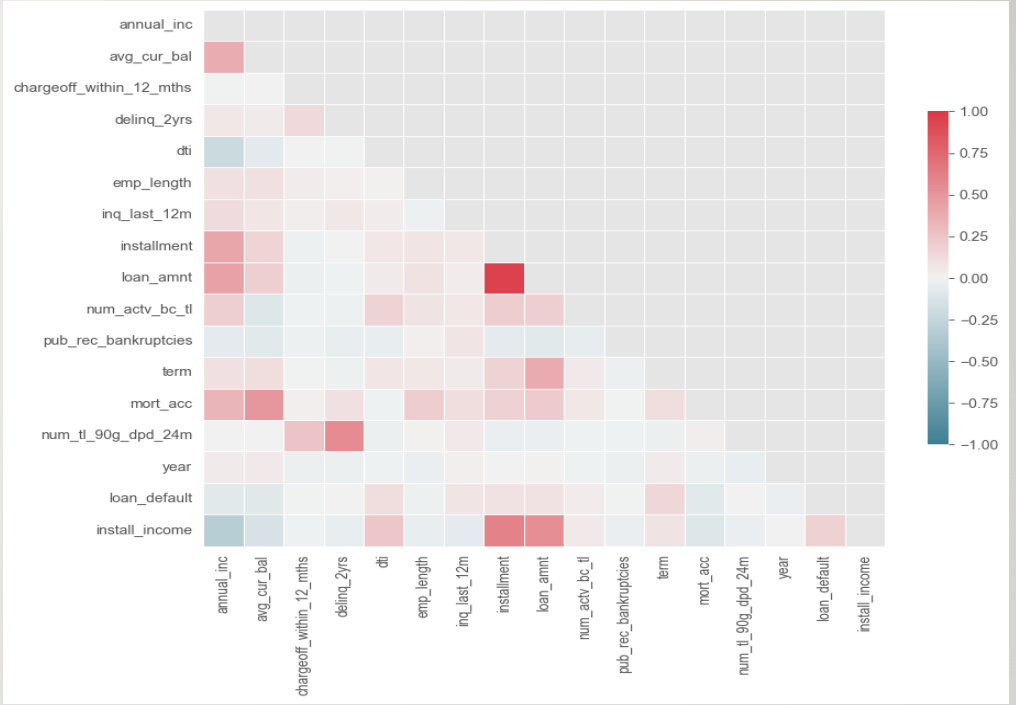
ASSOCIATION: ANNUAL INCOME & MONTHLY INSTALLMENTS

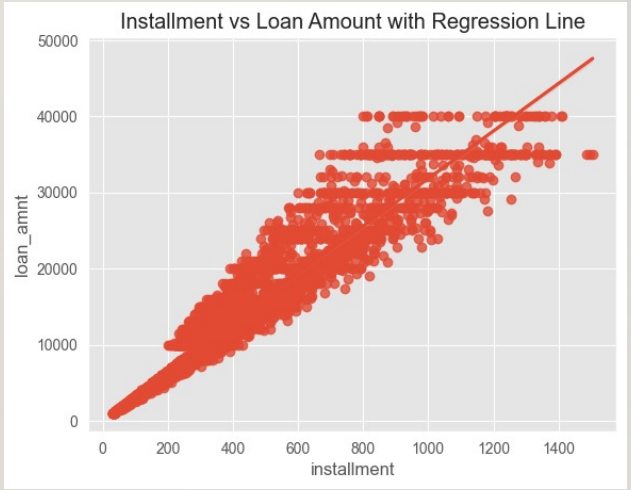
- This plot looks at the relationship between annual income and monthly installments.
- We can see that there is a more linear relationship between these 2 variables, with a few outliers on the far left.



CORRELATION PLOT

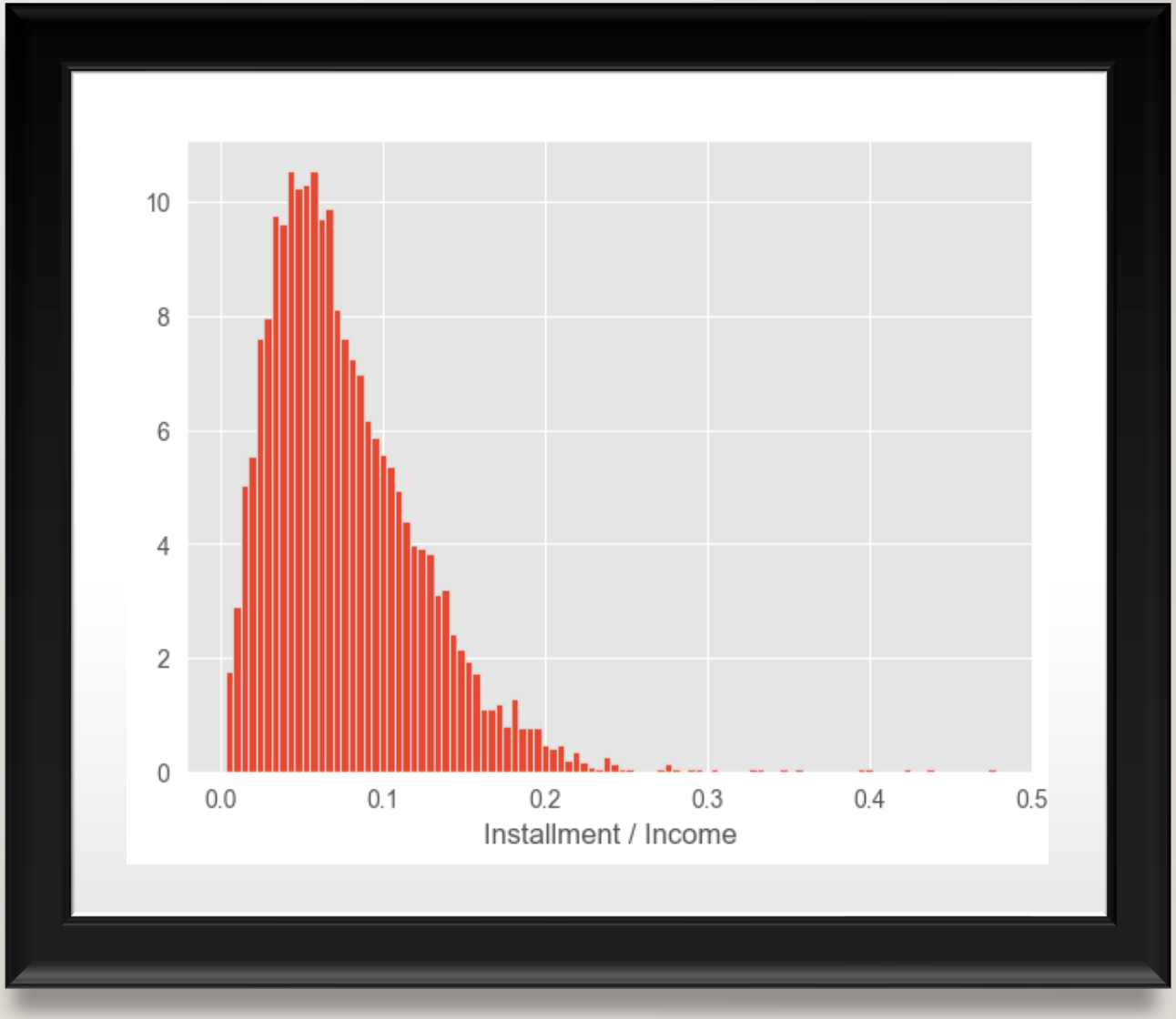
- Created a correlation plot to see which variables relate towards our target variable.
- We notice a high correlation between loan amount and installment. This could lead to multicollinearity problems. Based to drop or transform these variables.





NEW VARIABLE

- The yearly payment owed by the borrower, as a fraction of annual income, is a standard metric used in evaluating whether a loan should be issued.
- We create a new variable that calculates this using the installment variable and annual income.
- Install_income = 12 * Installment/ (annual Income)



END OF PRESENTATION