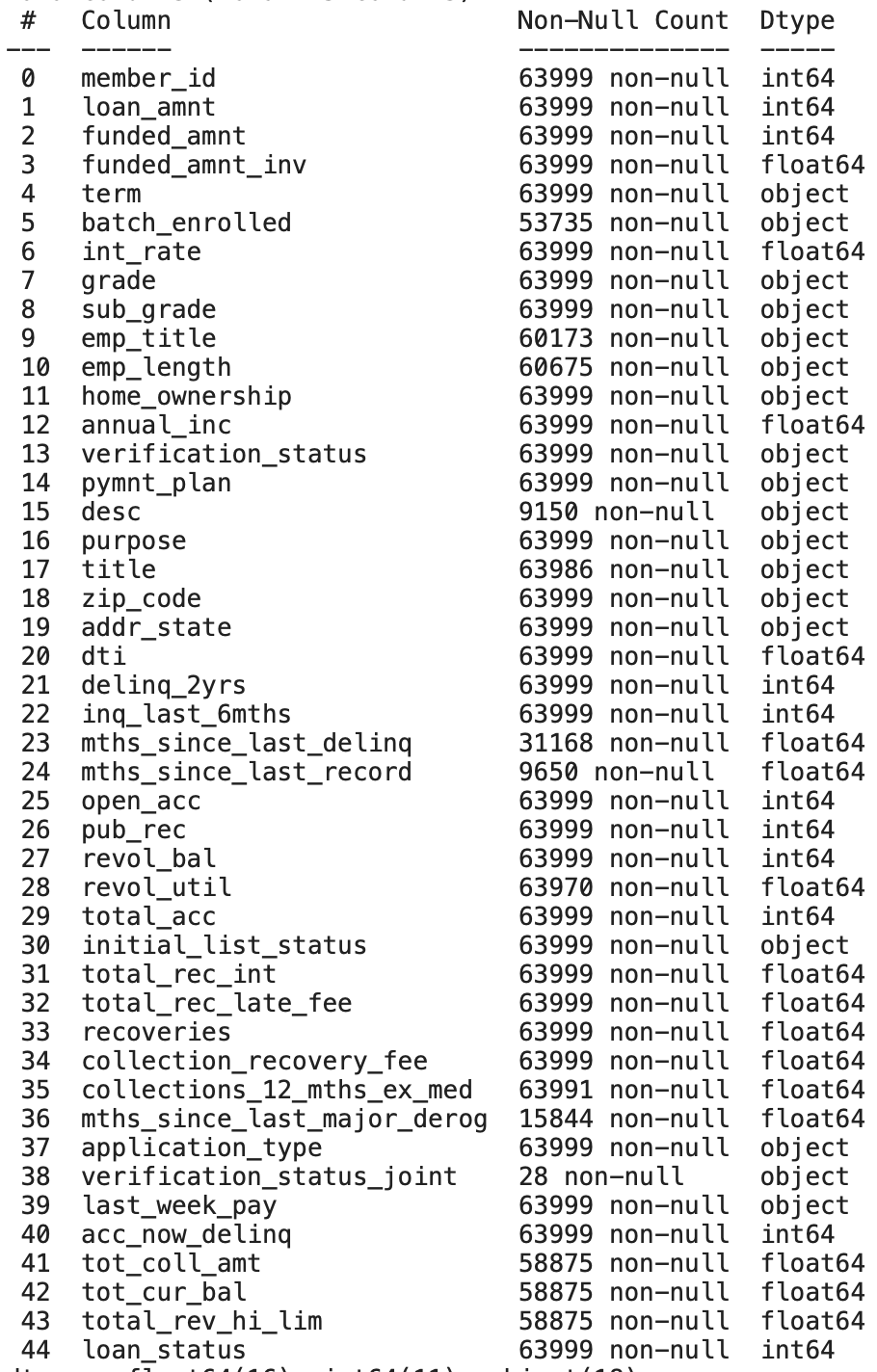
Model to Predict Loan Eligibility Status Through Machine Learning

# Solution Approach Document

Link to Dataset: <https://www.kaggle.com/shadabhussain/credit-risk-loan-eliginility>

**About the Dataset:**

Dataset has 63,999 rows and 45 columns. The following are the attributes of the dataset:

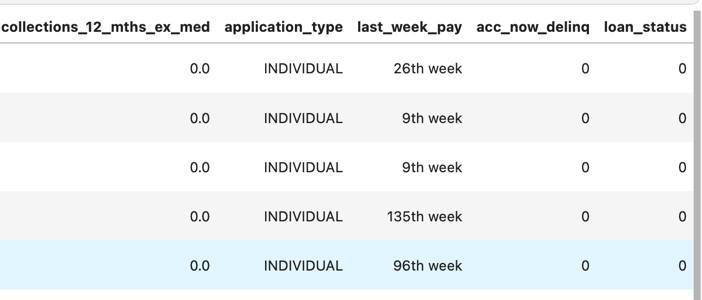
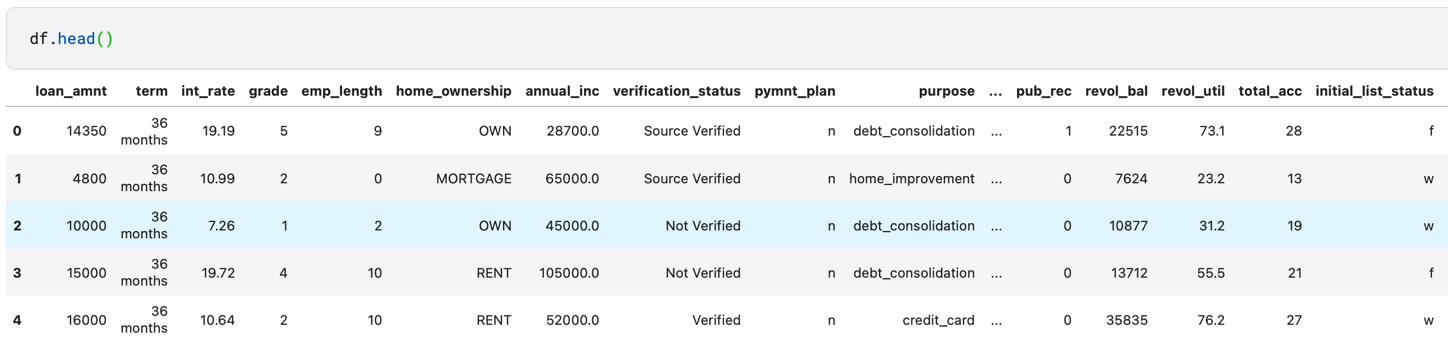


**Note:**

* After taking a look at the dataset and going through data preprocessing and exploratory data analysis, it was determined that not all of these columns would be significant in determining the loan eligibility status either due to insignificance or too many null values.
* All the code to all the analyses will be provided in the final solution rather than this solution approach document

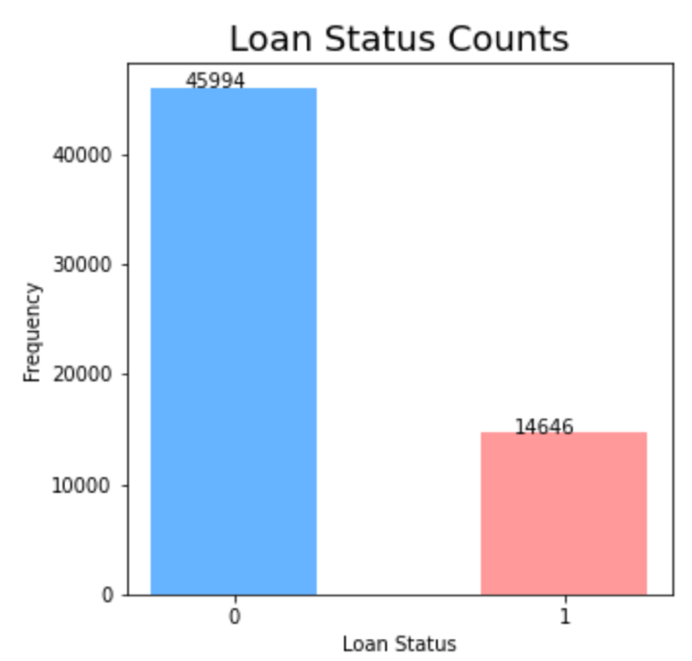
**Exploratory Data Analysis:**

Here is a preview of the dataset after dropping columns:



**Bar Plot of Loan Status Counts:**

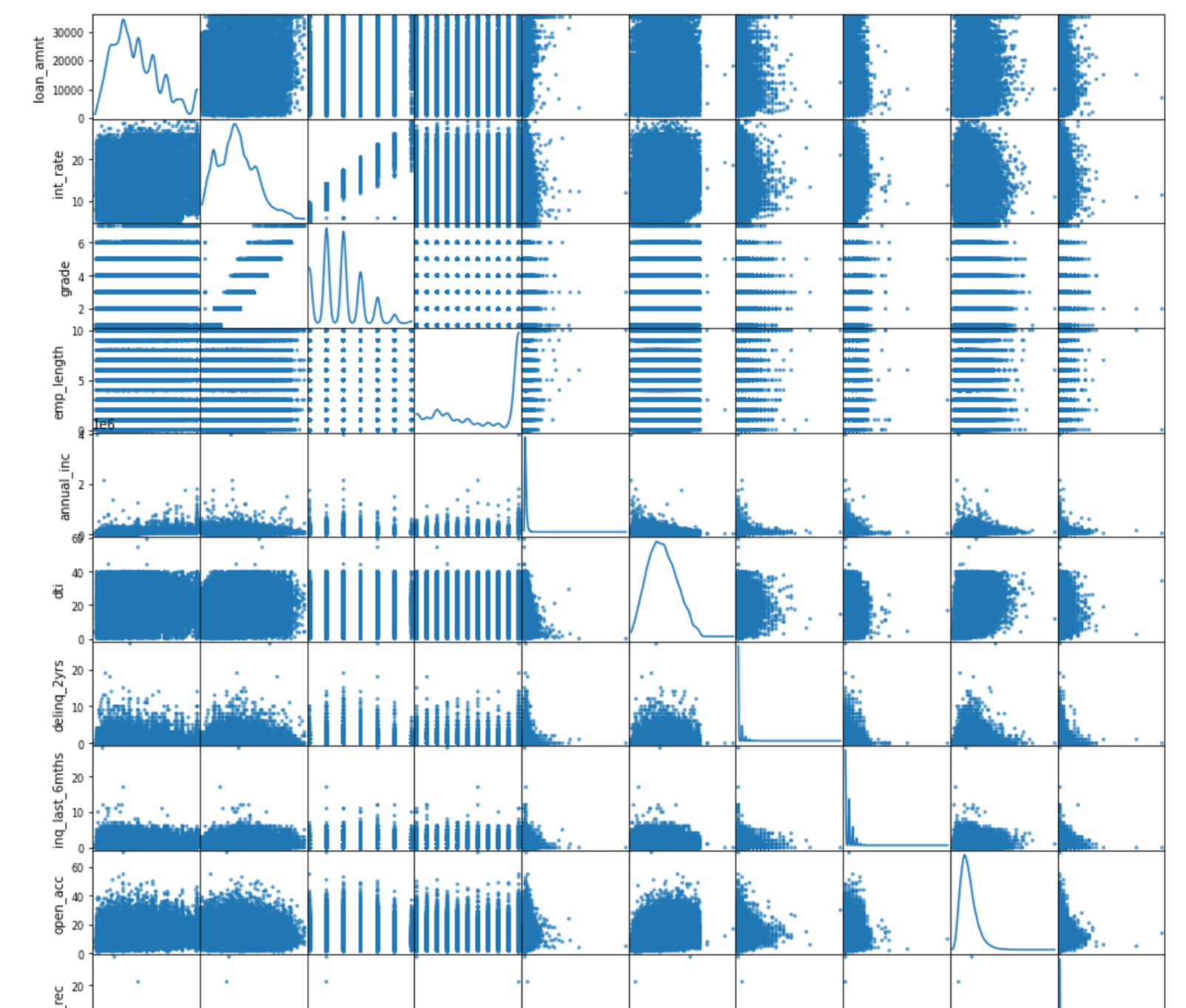
This bar plot is charted in terms of 1s and 0s, just as the dataset, and serves to provide an idea of how many loans were considered to be approved or not based on the conditions provided.



As seen in the bar chart, most of the Loans received a loan status of ‘0’, specifically 76.25%, is more dominant in that data column.

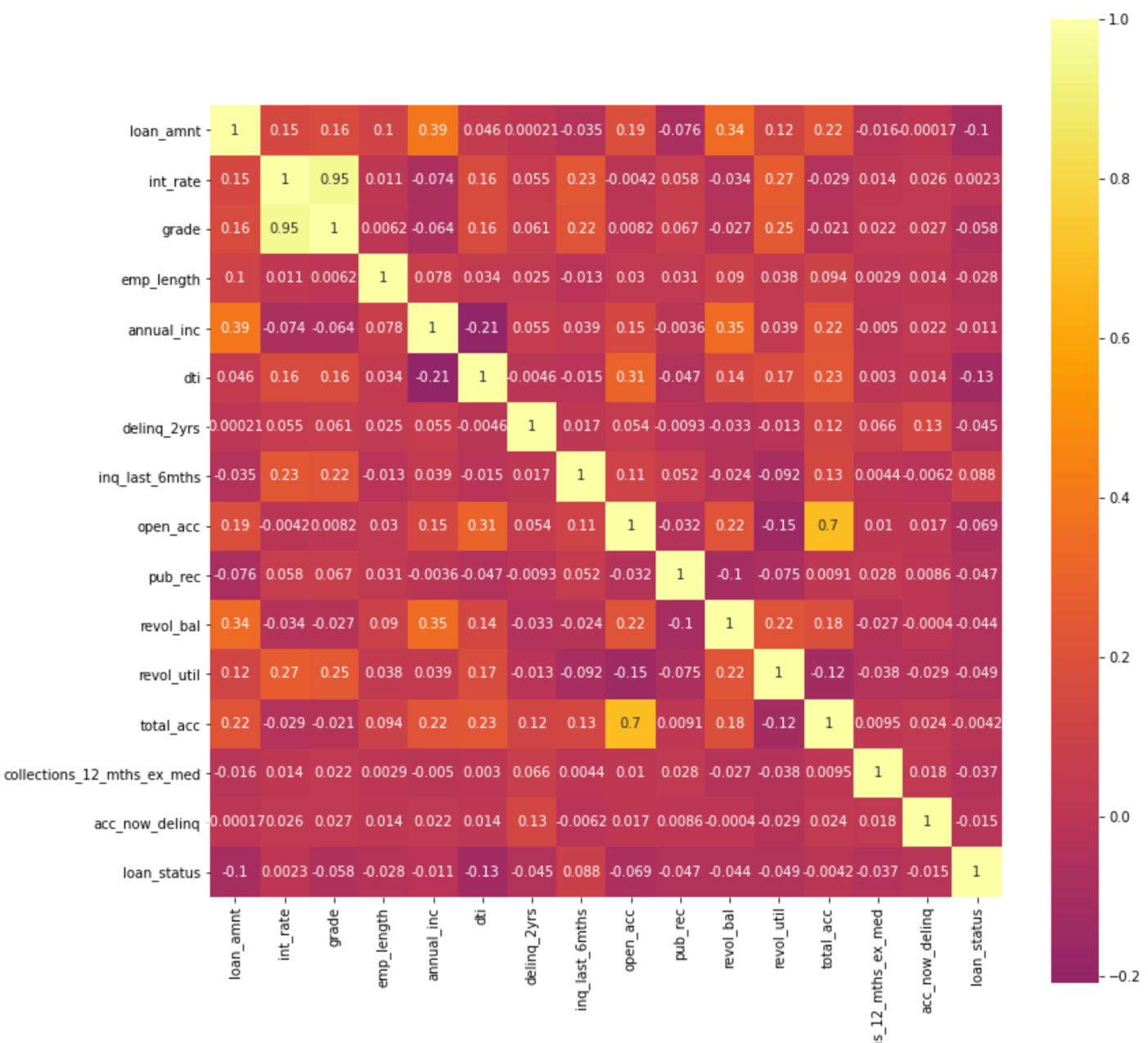
**Pairplot:**

The following shows the pairplot of this dataset which allows us to see various correlations between different variables in a compact manner.



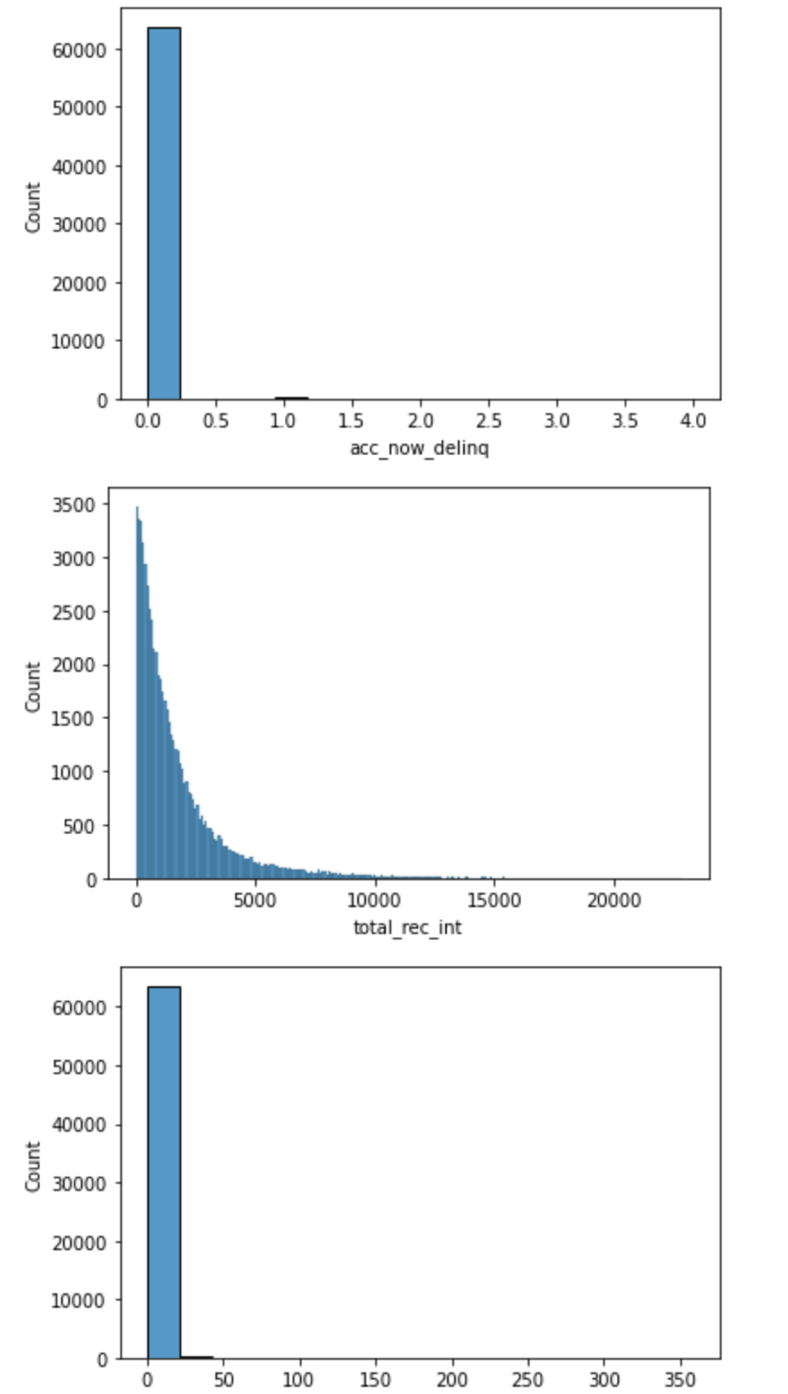
**Correlation Heatmap:**

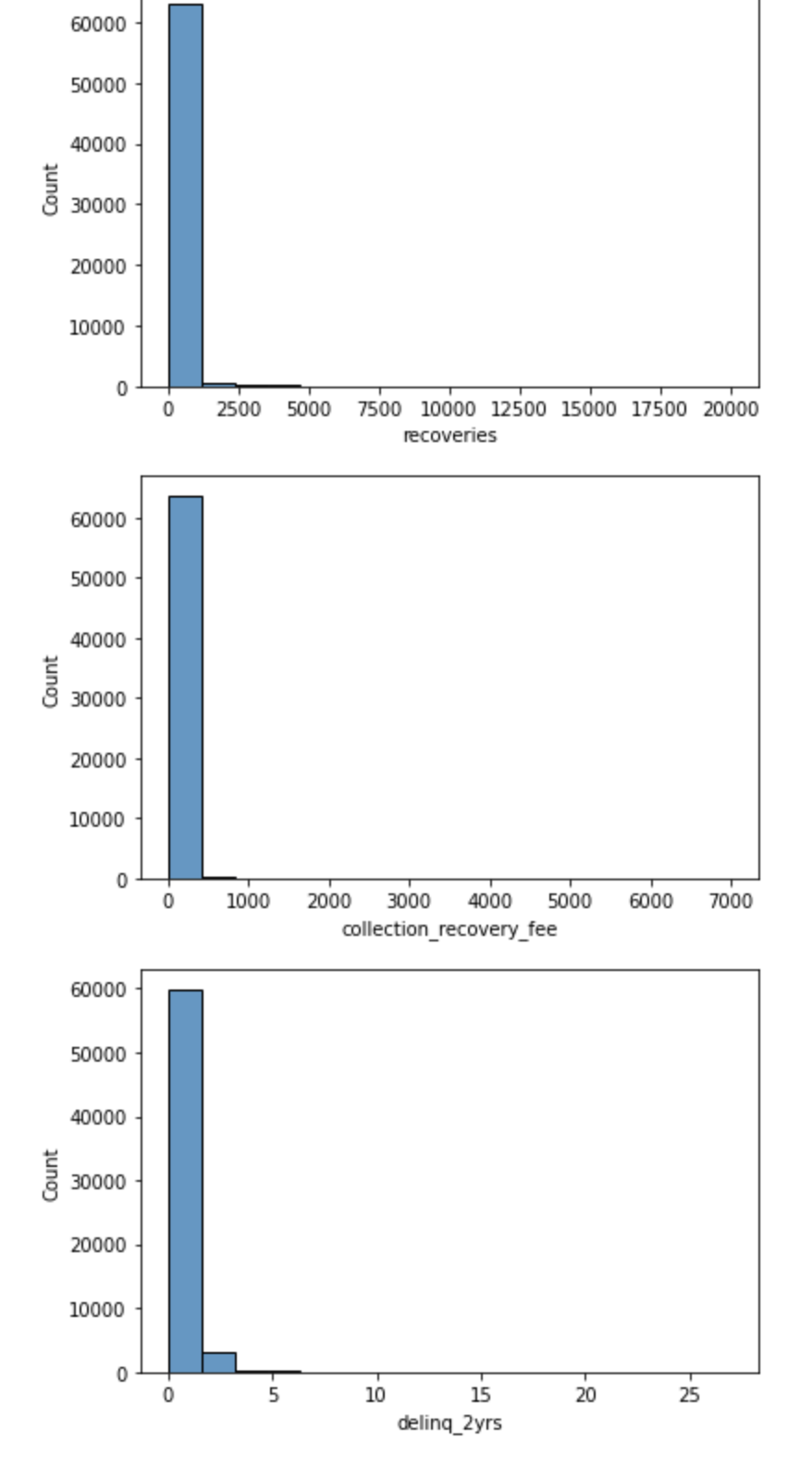
This heatmap displays the correlation between different features of the dataset in order to help us understand what the dataset and its attributes are like.

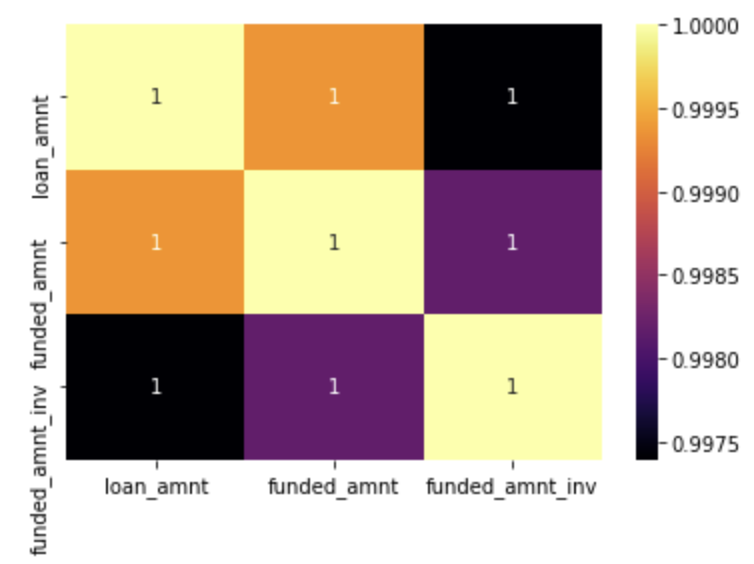


**Data Values Check:**

By taking a look at the following histograms, we can get an idea of the distribution of the values in each data column and see how biased the data is. After looking at all these plots, it can be seen that all of them are biased and the data is lobsided. Hence, these all of these were dropped and grouped as not being useful in building the model.







The above heatmap displays the correlation between the three columns “loan\_amnt”, “funded\_amnt”, and “funded\_amnt\_inv.” Since the correlation between these three columns is 1 in all of the squares in the heatmap, two of these three columns can be dropped.

**Building the model:**

After EDA and Data preprocessing on the dataset, the edited dataset would be exported from jupyter notebooks and then inputted into Vertex AI which would be used to build a model to predict the “loan status” or the loan eligibility based on certain features. After the model is built and validated, it can be exported from Vertex AI and deployed on the internet for usage.