- **Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation and the benefits that your analysis provides to the business or stakeholders of this data.**

The dataset is about the loan status of students studying in the college. Based on the time, principal, loan amount and age, we want to predict weather the student id going to pay time in due time or not

- **Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.**

Field: - Description

Loan_status: - Whether a loan is paid off or not

Principal: - Basic principle loan amount at the

Terms: - Origination terms which can be weekly (7 days), biweekly, and monthly payoff schedule

Effective_date: - When the loan got originated and took effects

Due_date: - Since it's one-time payoff schedule, each loan has one single due date

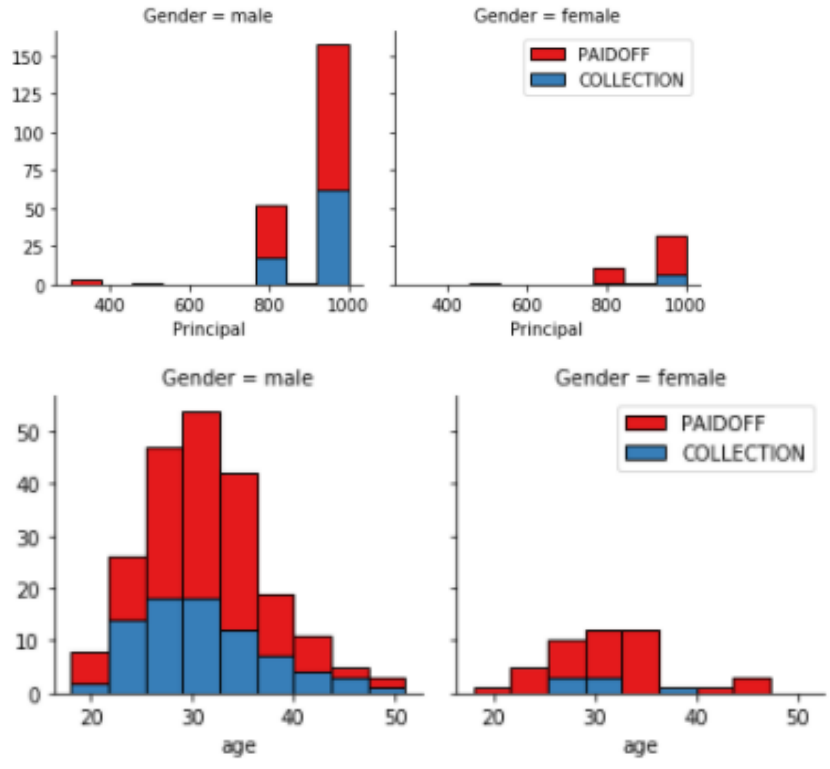
Age: - Age of applicant

Education: - Education of applicant

Gender: - The gender of applicant

- **Brief summary of data exploration, actions taken for data cleaning and feature engineering**.

The date columns are object data typed. So, first the date columns need to be converted to datetime and the loan status, gender, education need to be encoded since that is a categorical feature.



From both of these visualizations, I came to know that: -

1) Males are more in number than females
2) People who are older paid off their loans whereas younger people still have not. (it can be said since, older people have savings)

- **Summary of training at least three different classifier models, preferably of different nature in explainability and predictability. For example, you can start with a simple logistic regression as a baseline, adding other models or ensemble models. Preferably, all your models use the same training and test splits, or the same cross-validation method.**

```
KNN - The best accuracy was 0.7766540244416351 with k= 7
DECISION TREES - The best accuracy was 0.7064793130366899 with max_depth = 1
SVM - The best accuracy was 0.7714285714285715 with kernel = rbf
LOGISTOC REGRESSION - The best accuracy was 0.7572463768115942 with kernel = rbf
```

- **A paragraph explaining which of your classifier models you recommend as a final model that best fits your needs in terms of accuracy and explainability.**

As observed in the above predicts, I would definitely use KNN as my model. Since, it is performing well on the test set.

- **Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your classifier model.**

As my model is more biased to males, this can work well for male students but not much good for female students. Not considering the category, older students are more likely to pay their student loans. So, age is the key factor that is driving our model

- **Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model after adding specific data features that may help you achieve a better explanation or a better prediction.**

The dataset that I chose is good but it can be improved by adding more female students. So that the predictions can be more accurate while coming to female students

BY,

SAI GOWTHAM BABU AMBURI.