I first cleaned the data like finding the null values and changed the null values in education to 0 since very likely NaN in education means the person is not educated and their was no 0 value beforehand. Next converted the string data of proof submitted to numeric by creating a function which allots a separate integer from 0 to 4 to each of the type of proof submitted and adding the column to the dataframe. Plotted the correlation heatmap to see the variation of properties with the required variable, since our data has 0,1 dependent variable the dependency coefficient are around 0.1 for each of the independent variable. Now plotted the boxplot to check for the outliers and also see the statistical data for maximum and minimum to find out more info. Like for no of loans and current loans the maximum data is quite high like 109 and the data for high values are mostly below 10 cases. So dropped the rows with less than 10 cases for no of loans and 15 for curr loans. Now in order to find a threshold to eliminate the fliers for loan amount and asset cost ,I used scatter plot to know the data. And similarly filtered the data for these to features. Now in order to make model more efficient I scaled the data for loan amount and asset cost using standard scaler. Now I used KNN classifier to make the model to predict the loan defaults.