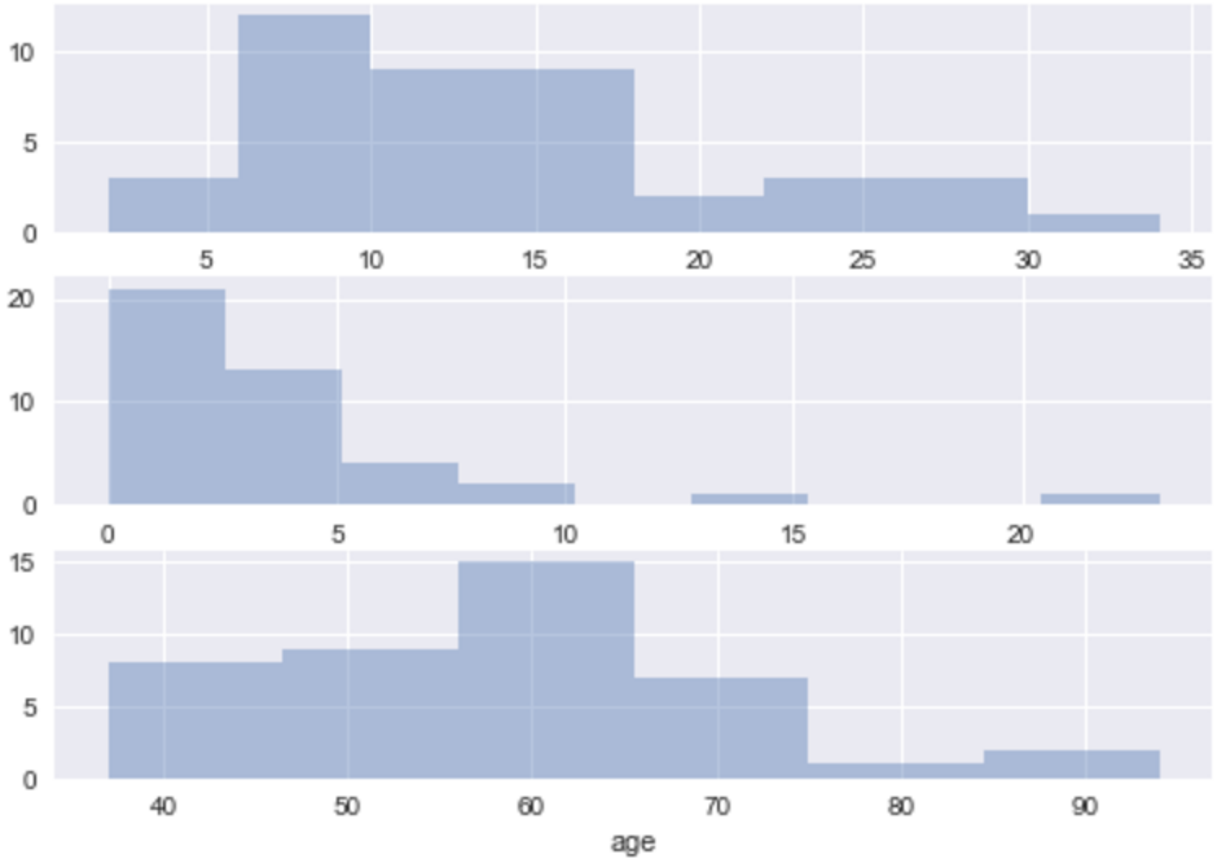
Homework 2 Report

Summary of Methods
In this assignment, I predicted the instance of serious delinquency in two years using a credit dataset. In exploring the data, it appeared that there were missing values for monthly income and the number of dependents a creditor had. I decided to fill missing income values with the average of the training and testing datasets, respectively. This ensures that we can utilize the 3000 rows without monthly income while retaining the integrity of the monthly income feature. Exploratory analysis also uncovered that *debt_ratio* and *monthly_income* features had extreme values, far above the normal rate. As an attempt to utilize these values, rather than discard them, I used a scaling function that accounts for outliers. To complete the feature engineering process, I created bins for the following variables: income, age and number of dependents.

Analysis of High Debt Ratio
Persons with high debt ratios had significantly more real estate loans compared to someone who did not (1 loan versus 3 loans on average).

Distribution Graphs for persons with High Debt Ratios
(top to bottom: number of credit lines, number of real estate loans, age)



Findings and Efficacy of the Model

Using a random forest classifier, I discovered that the *numberoftime30-59dayspastduenotworse* *revolvingutilizationofunsecuredlines* and *numberoftimes90dayslate* were the most important features for classification, in addition to the *scaled debt ratio*. Unfortunately, the model also detected *person_id* as a significant feature, which is completely erroneous. Further, the Logistic Regression model I trained predicted the correct outcome with 93% accuracy, but this percentage is also the frequency of the delinquency variable in the dataset. This let me to the conclusion that the model does not outperform a simple decision rule that were to score an incoming person as a 0 (or rather, no serious delinquency in two years) 93% of the time. Likewise, I evaluated a strategy of predicting zero every time, which had an accuracy score of 99%. More work must be done to engineer features before a better model can be developed.

Please refer to the HW2 jupyter notebook for code and additional analysis.