The initial question asked if we could ensure that customers would pay back their loans. After exploring and visualizing the data, there are some useful business insights to be gained. From the 30,000 customers offered loans, 6,636 defaulted on their loan. Given only 22% of the customers defaulted, we can conclude that the majority of the customers repaid back their loans, and the predictive model we plan to build will need to focus on predicting what customers will be the ones likely to default. Client ages ranged from 21 to 79, and 3 out of 4 customers were under the age of 41. However, the age of a customer did not provide any useful insight as to whether or not the client would default. The probability of default based on education showed that customers with graduate level education defaulted only 19% of the time. On the other hand, customers with high school level and university education defaulted 25% and 24% of the time, respectively. Divorced people were found to default 26% of the time. Married and single people were found to default 23% and 21% of the time respectively. The data provided by CreditOne contained features named ‘pay\_1’…’pay\_6’; these features described the repayment status of each customer. The higher the ‘pay\_n’ value, the further behind the customer was to paying back his or her loans. I added the values from the ‘pay\_n’ features to create a collective feature ‘pay’; this feature ranged from a value of -12 to 36. As the ‘pay’ value increased, so did the probability of default; this makes sense, because the higher the value is, the further behind the customer usually is to paying back their loans and the higher the probability of them defaulting. The ‘limit\_bal’ feature is also useful in predicting probability of default; this feature tells us the size of the loan the customer was approved for. Customers that have a history of paying back their loans will get pre-approved for larger loans. Therefore, the more money a person is approved for, the more responsible the person is rated to be and the lower their probability of defaulting will be. The ‘balance’ feature was created by subtracting the total ‘pay\_amt’ from the total ‘bill\_amt’. Then one could see how close a customer is to reaching the limit of their loan by dividing their ‘balance’ by their ‘limit\_bal’; this feature was created under the ‘limit\_balance\_ratio’ name. It proved to be useful as the higher the ‘limit\_balance\_ratio’ was, the higher the probability of the customer defaulting was. This makes sense, because the closer a person is to using all of their allowed balance, the harder it will be for them to pay it back.

Apart from these business insights, I learned how to use multiple libraries to visualize and understand relationships in the data. During the visualization part of the exploration I used the seaborn library for the first time to plot factor plots and regression plots of the data. These showed correlations between features and the probability of a client defaulting. Using the data wrangling pandas cheat sheet, I learned the various commands to separate the dataset into subsets that allowed for a more specific look into the different variables. Overall, the exploration of the data allowed me to get comfortable with the use of jupyter notebooks and other python commands useful in the manipulation of data.

My final recommendations are for the continuation of the predictive model creation. The predictive model will help CreditOne determine a probability of a customer defaulting based on their previous behavior and pertaining data. From there, a cutoff probability can be selected to determine what customers will be eligible for future loans and which ones will not. Future collection of the data could contain features not in the current dataset; for example, a more detailed collection of the marriage and education features of the customers could help us find useful information regarding the customers belonging to the ‘other’ categories.