This research problem was to be able to identify if a loan will be given or not to an application based on data collected previously by the bank. However, this question will not be answered as the data is not labeled and consequently a classifier cannot be trained. The other objective is to identify and interpret two or three dimensional variables which contain as much information as the original variables in the dataset. The main issue from the original dataset is that there were too many points that have a strong correlation and therefore, the researchers would like to avoid multicollinearity and misinterpreting results.

The most appropriate method we recommend is Principles Components Analysis (PCA). This method uses eigenvalues which are inherently orthogonal and using only the first few components ultimately reduces the dimensionality of the data. The two main assumptions that underlie PCA for this case are: First there should be correlations among the variables. And second, the relationship between the variables is assumed to be linear.

We have attempted to find two or three principle components and eventually decided to use three. With three principal components, we were able to explain 84% of the variation in the data. And we used a cutoff of 0.4 for each principle component to list its corresponding variables. The first principle component includes information about the applicant such as the applicants age, length of residence in the area, length of time that applicant has been employed with their most current company, and finally the amount of savings they have in dollars. This principle component can be labeled as “Time Stability.” The second principle component comprises of debt and credit cards the applicant has. This principal component can be labeled as “Debt Management.” The third and final principal component has the applicant’s income in dollars and education in years. This principle component can be labeled as “Income Stability.”



