SCMA 648 Business Data Analytics

Assignment 3 - Cluster Analysis and Principal Component Analysis

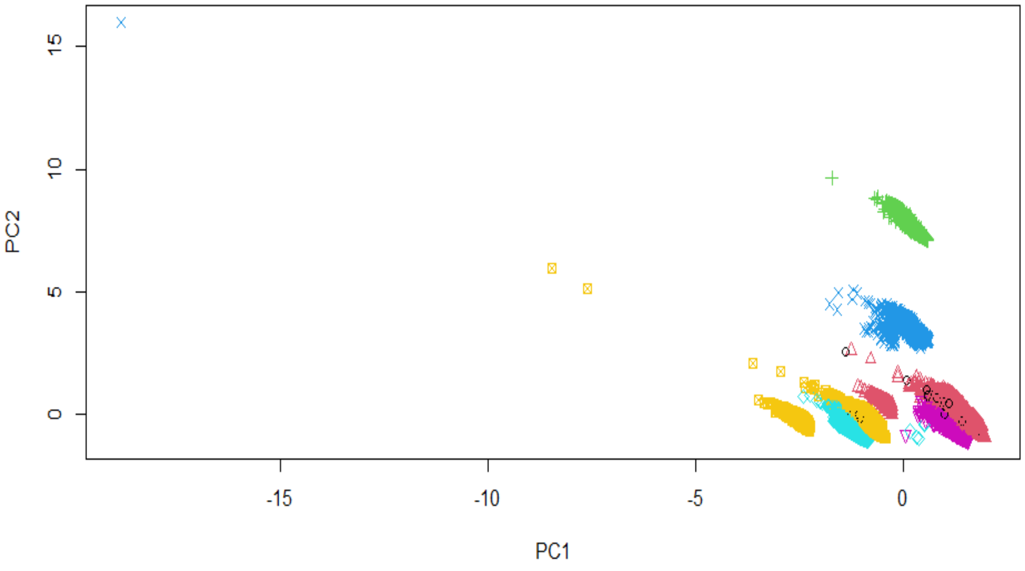
The working directory was set according to the source file location. The saved rda file that was created for quarter 3 was loaded to the R environment along with the required packages.

Pre-processing steps:

- The random number generator seed has been used to get the same results each time.
- The rda file was loaded which contained the data frame called df.
- To perform the cluster analysis, the columns income, loan amount, employment length, home ownership status, and debt-to-income ratio have been selected.
- There were null values for the column dti (debt-to-income ratio). These were imputed with the median value. After imputation, there were no null values observed in the summary of the data frame.
- From the columns considered, there are three numeric and two factor columns. The variables emp_length and home_ownership have several categories. Dummy variables have been created for both these variables. One column each for each category present in the variables has been created. The original column has been dropped.
- For the feature emp_length, there were “n/a” some values present as a category. These have been retained assuming that “not applicable” can be a category in that column.
- The variables have been encoded as numeric. Following this, the data has been scaled and centered as the features have been measured on different scales.

After the pre-processing, the data was clustered into seven groups and the principal component scores were found. However, there were outliers in the data. Therefore, the outliers were removed based on the values obtained from summary of the scores. The analysis was conducted again with the outliers removed.

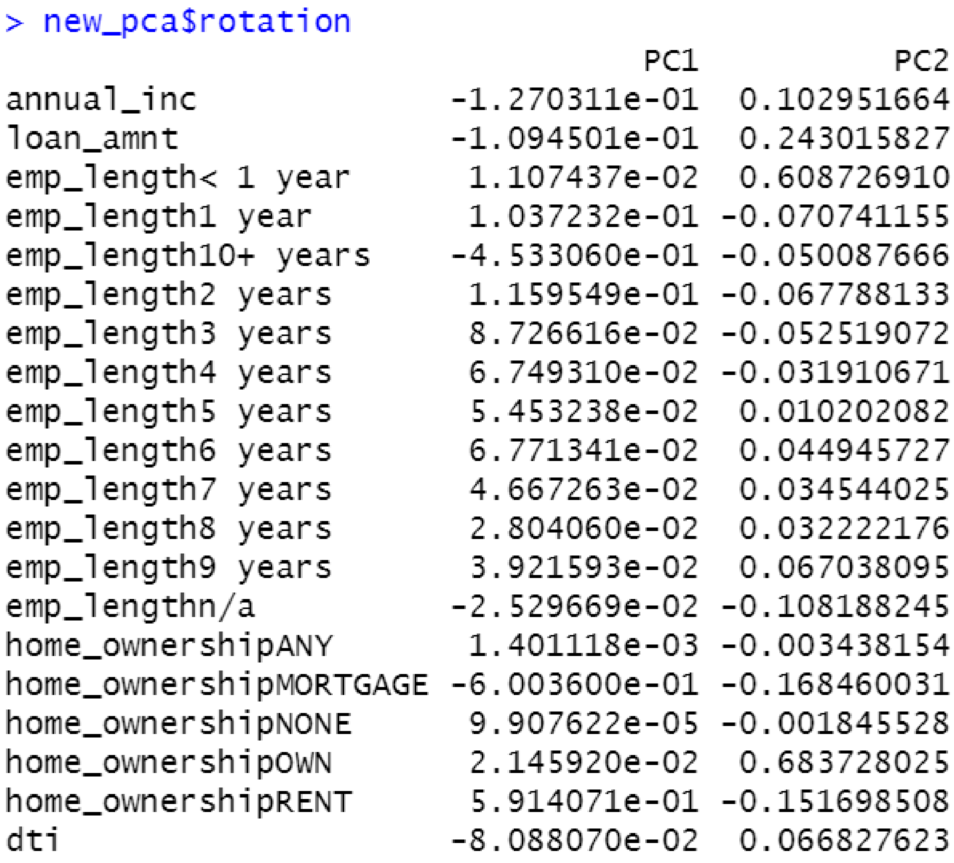
1. Cluster the borrowers into seven groups using k-means clustering.



There are a lot of observations. Therefore, instead of hierarchical clustering, k-means clustering has been used.

2. Use principal component analysis to identify characteristics of each cluster.

The lending data was clustered into seven groups. The clusters were plotted on the first two principal component scores. The outlier has been assigned to some other point this time. The blue, green and yellow clusters are mostly well-separated. The purple, cyan and red clusters seem to overlap each other. Some points in cyan cluster are not well clustered. There is a variation in data because of which we cannot see these clusters clearly.

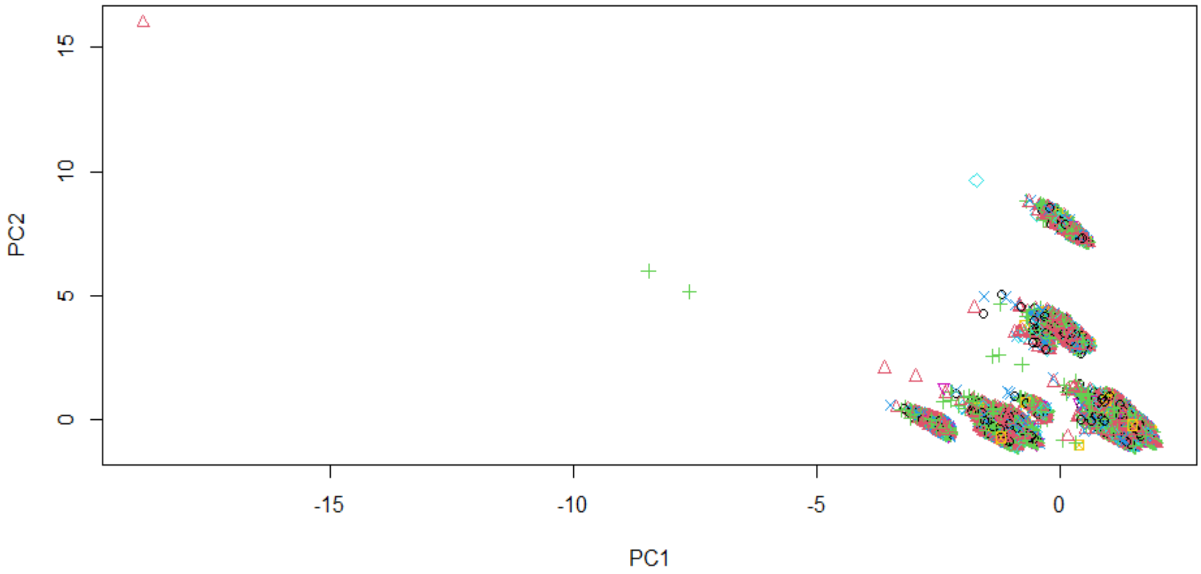


```
> new_pca$rotation
                                 PC1          PC2
annual_inc             -1.270311e-01  0.102951664
loan_amnt              -1.094501e-01  0.243015827
emp_length< 1 year      1.107437e-02  0.608726910
emp_length1 year        1.037232e-01 -0.070741155
emp_length10+ years    -4.533060e-01 -0.050087666
emp_length2 years       1.159549e-01 -0.067788133
emp_length3 years       8.726616e-02 -0.052519072
emp_length4 years       6.749310e-02 -0.031910671
emp_length5 years       5.453238e-02  0.010202082
emp_length6 years       6.771341e-02  0.044945727
emp_length7 years       4.667263e-02  0.034544025
emp_length8 years       2.804060e-02  0.032222176
emp_length9 years       3.921593e-02  0.067038095
emp_lengthn/a          -2.529669e-02 -0.108188245
home_ownershipANY       1.401118e-03 -0.003438154
home_ownershipMORTGAGE -6.003600e-01 -0.168460031
home_ownershipNONE      9.907622e-05 -0.001845528
home_ownershipOWN       2.145920e-02  0.683728025
home_ownershipRENT      5.914071e-01 -0.151698508
dti                    -8.088070e-02  0.066827623
```

The variation in the first component is mostly due to the employment_length and home_ownership status. Borrowers with larger values on the first component will have larger values for these variables. The variation in the second component is due to employment_length less than one year and home_ownership status as their own.

The green and blue clusters seem to have positive values for PC1 and PC2. The borrowers will have higher values for all the corresponding variables. Cyan and yellow clusters have negative value for PC1 but positive values for PC2. The clusters purple, red and black seem to have a positive value for PC1 but a negative value for PC2. This indicates that here are a variety of values for the corresponding variables.

3. Evaluate how your clusters compare to assigning applicants to clusters by loan grade (you do not need to run k-means again for this step - simply change the visualization based on loan grade). Support your comparison with visualizations.



For the clusters assigned to applicants by loan grade, all the clusters seem to overlap each other. None of the clusters are clearly separable. This means that there is a lot of variation in the data. In the previous graph, some clusters were clearly separable.