ANALYSIS OF THE BANK LOAN DATA REPORT

GROUP 1

Isini Asalya	s15023
Aravinda Bandara	s15024
Vidura Chathuranga	s15030
Lakni Kodithuwakku	s15057

Contents

1. Objectives of the Analysis

The objectives of this data analysis are three-fold. First, to identify the variables or factors that are associated with customers who have defaulted on their payments in the past. Second, to segment customers into homogeneous groups based on their characteristics and identify the group of customers who are more likely to default on their payments in the future, along with their common characteristics. Finally, to compare the characteristics of this high-risk group with those of customers who have already defaulted in the past. By accomplishing these objectives, it may be possible to gain a better understanding of the factors that contribute to defaults and implement targeted strategies to reduce the risk of default. Furthermore, comparing the characteristics of high-risk customers with those of customers who have already defaulted may help to identify additional factors that should be considered when predicting future defaults, and enable the development of more accurate predictive models.

2. Description of data

Name of the dataset: **bankloan.csv**   No of Records- 1500   No of variables- 12

Variable No	Variable Name	Variable descriptions	Description of categories
1	branch	Branch	
2	ncust	Number of customers	
3	customer	Customer ID	
4	age	Age in years	
5	ed	Level of education	1 – Did not complete high school 2 – High school degree 3 – Some college 4 – College degree 5 – Post/Under-graduate degree
6	employ	Years with current employer	
7	address	Years at current address	
8	income	Household income in thousands	
9	debtinc	Debt to income ratio (x100)	

10	creddebt	Credit card debt in thousands	
11	othdebt	Other debt in thousands	
12	default	Previously defaulted	0 – No 1-yes

Table 1 : Description of data

3. Data pre-processing and Feature Engineering

Data set does not include any missing values, so no need to remove or impute missing values.A new variable named “default_name” was created by recoding the values of 0 and 1 in Default variable in to “Yes” and “No” respectively. The “branch”, “ncust” and “customer” variables have been omitted since those variables do not provide any meaningful information towards this study. Also, the rest of all predictor variables was standardized for further analysis.

4. Descriptive Data Analysis

According to the Figure 1, It can be observed that there is a significant difference in the distribution of age between the two groups. It seems that Group “Not default” has a slightly higher median compared to Group “Default”. Moreover, Group “Default” has fewer outliers than Group “Not Default”, which suggests a more homogeneous distribution of income in that group.

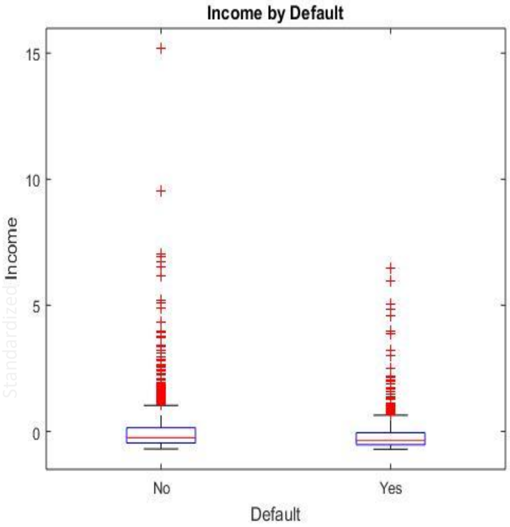


Figure 1

Based on the analysis of Figure 2, it can be observed that the normality assumption is violated for the income variable. Therefore, a non-parametric statistical test, namely Wilcoxon’s rank sum test, was conducted to investigate whether there is a significant difference between the median values of the two distributions.

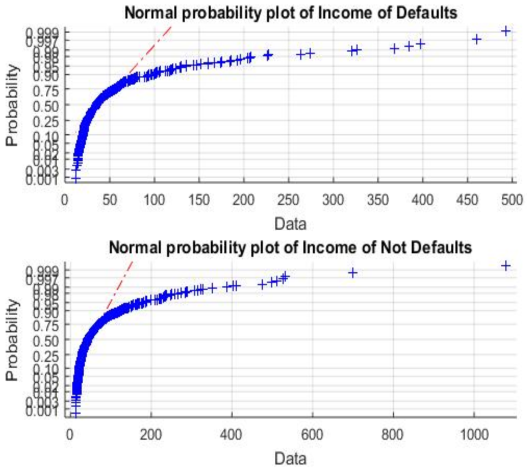


Figure 2

The results of the test indicate that the p-value is 6.4153e-09, which suggests that the medians of the two groups are significantly different at a 95% level of confidence. These findings suggest that there appears to be an association between the age variable and the default variable.

Based on the observation of Figure 3, it can be inferred that the distribution of Age within the "Not default" group is wider compared to that of the "Default" group. Furthermore, a significant majority of the observations belong to the "Default" group.

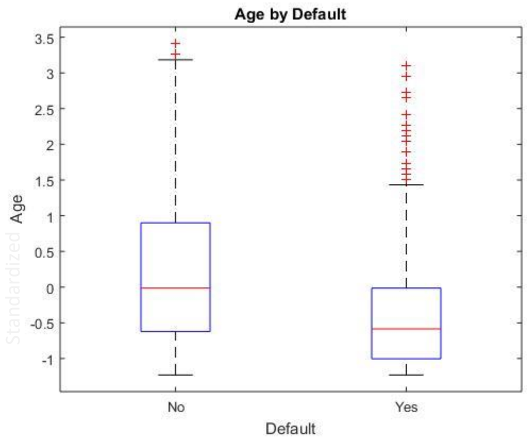


Figure 3

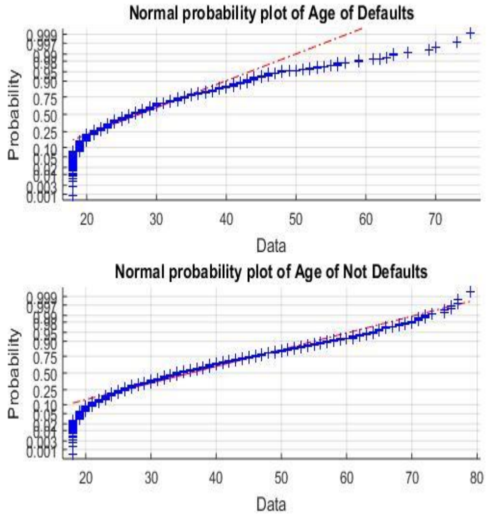


Figure 4

Figure 4 reveals that since the normality assumption is violated, we have conducted the Wilcoxon's rank sum test and the p value obtained is 2.0446e-30. Since the p value is lower than the significance level we considered, this suggest that there is a significant difference between median values of two distributions and there may be an association between the Age variable and Default variable.

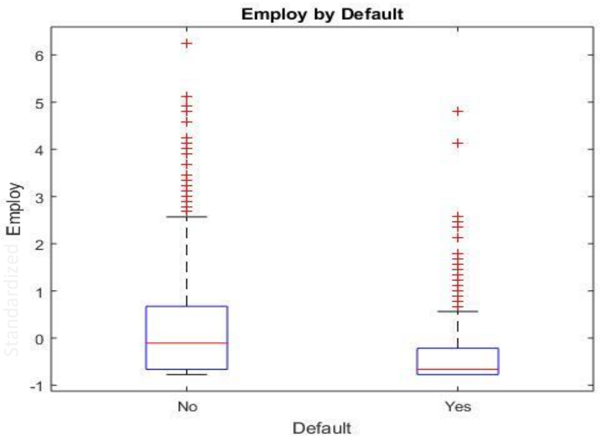


Figure 6

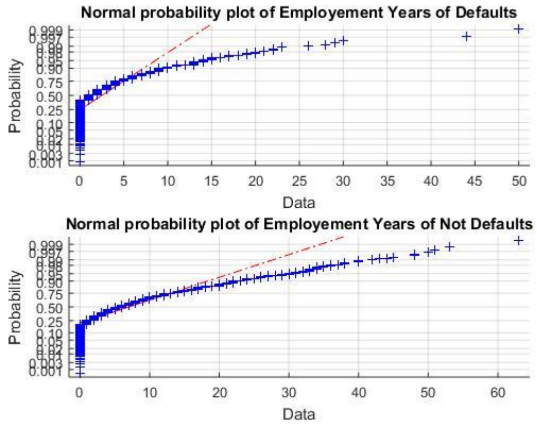


Figure 5

According to the Figure 5 there is a wide distribution in "Not Default" group than "Default" group in Years with current employer. We also can see that median no of years of the "Not Default" is higher than "Default". But, here also we have conducted the Wilcoxon's test since the normality assumption is violated and obtained the p - value as 2.1604e-37. So, it implies that there is a significant difference in medians between "Default" and "Not default" groups which also suggests an association between the Employ variable and Default variable.

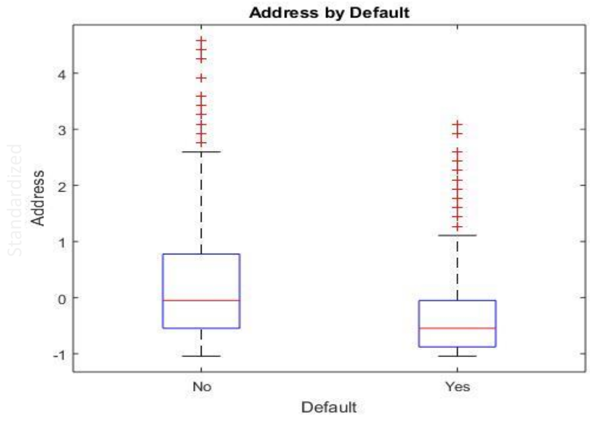


Figure 8

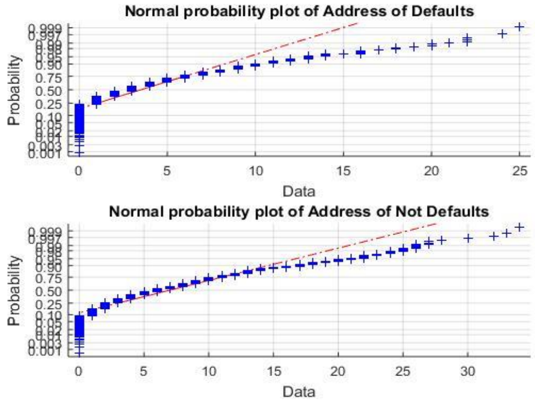


Figure 7

In Figure 7 also, there is a wider distribution in “Not Default” group than “Default” group in Years at current address. We also can see that median number of years at current address in “Not Default” is higher than “Default”. We have conducted the Wilcoxon’s test and obtained the p- value as 3.7661e-29. So, it implies that there is a significant difference in medians between “Default” and “Not default” groups which again suggests an association between the Employ variable and Default variable.

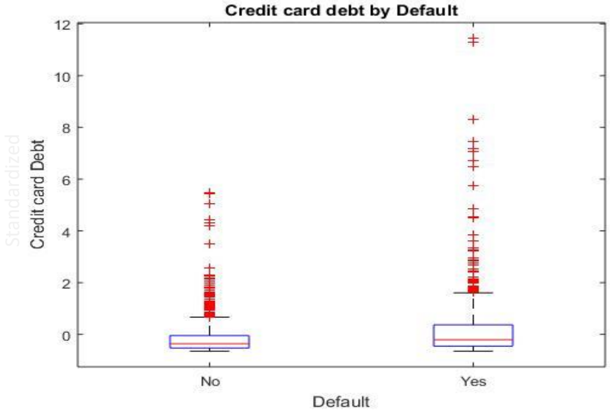


Figure 9

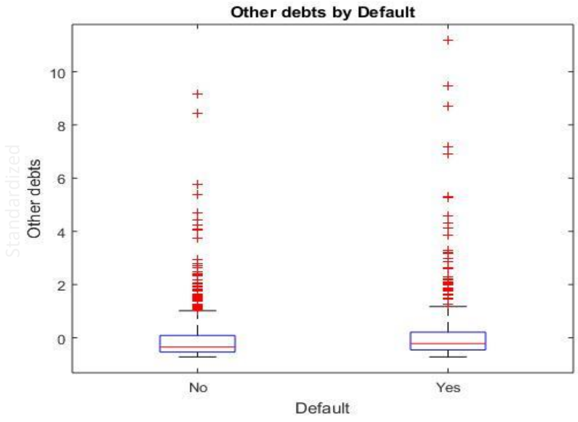


Figure 10

The analysis was extended since because of the need of identifying the association between Default variable and Debt amount of a person. For that we have considered the distribution of credit card debts and other debts of customers according to Default variable. Figure 9 and 10 illustrates the how these debts have been distributed among “Default” and “Not Default” groups. In both figures since it was hard to identify whether there is a significant difference between medians of “Default” and “Not Defaults”. So, we again conducted Wilcoxon’s sign rank test and obtained these P – values 2.8817e-06 and 2.1484e-12 respectively. This suggests that there seems to have an association of Credit card debt and other debts with the Default variable.

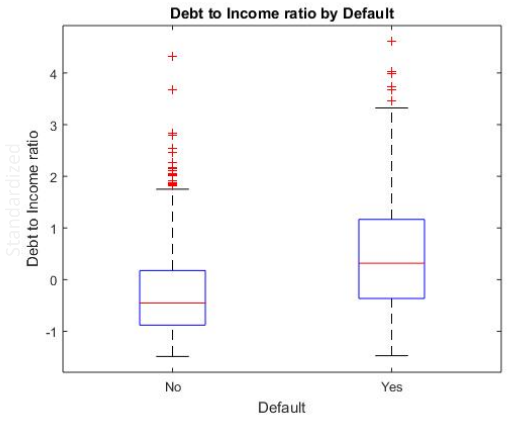


Figure 11

Figure 11 examines the association between Debt-to-income ratio and the Default variable. It can be observed that the distribution of the "Default" group is wider than that of the "Not default" group. Further analysis through a Wilcoxon's rank sum test yielded a p-value of 1.8471e-42, indicating a significant difference between medians of two groups. This result suggests that there may be an association between Debt-to-income ratio and the Default variable.

Since there is an ordinal variable named "Education" which provides information about the education level of the customer, the association between Education variable and Default variable also needs to be considered. Figure 12 illustrates how the number of customers who are Default and Not Default is varying according to each education level.

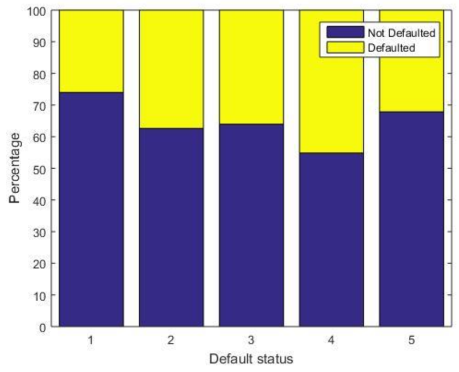


Figure 12

A Pearson's Chi-squared test was conducted to investigate the presence of an association between Education level and the Default variable. The obtained p-value for the test was 0.000153, which suggests a significant association between the two variables. This result indicates that Education level may be a relevant factor in predicting the likelihood of Default.

5. Advanced Data Analysis

The most crucial variables were chosen using a decision tree approach. The classification tree algorithm was first fitted using the default arguments to the training dataset. The best pruning level for the fitted tree was then determined using the "cvLoss" function in the Matlab software. It was determined that

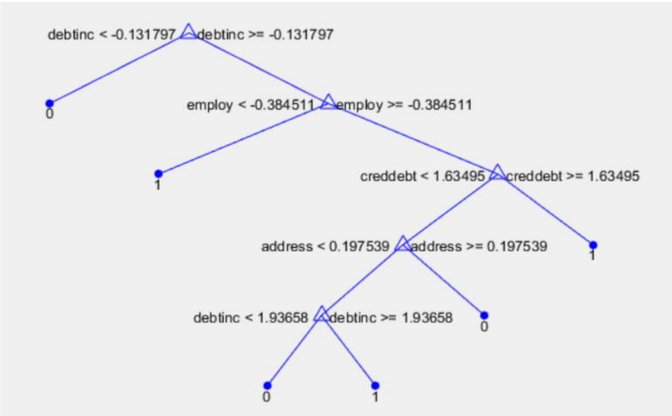


Figure 13

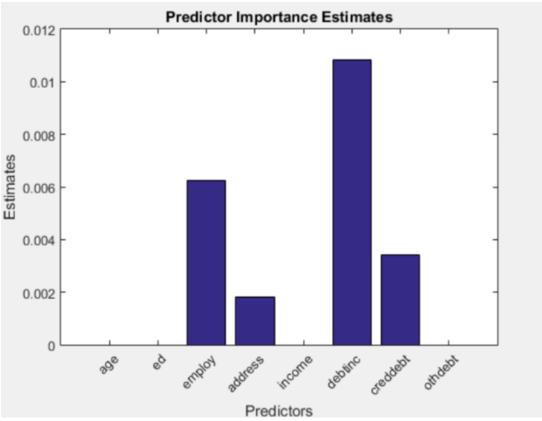


Figure 14

level 21 was the best pruning level out of 24 levels by using the smallest tree whose cost is within one standard error of the minimal cost.

The fitting tree was post-pruned using the best pruning level, as shown by figure 13-the decision tree above (21). Then, important predictors were chosen using the variable importance plot in figure 14. The most crucial factors in this dataset are employ, address,debtinc and creddebt.

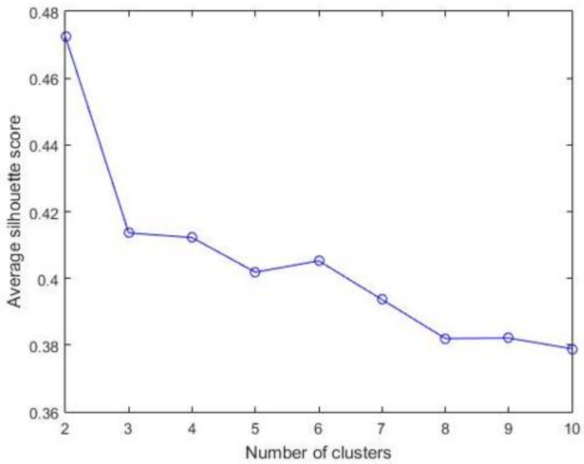


Figure 15

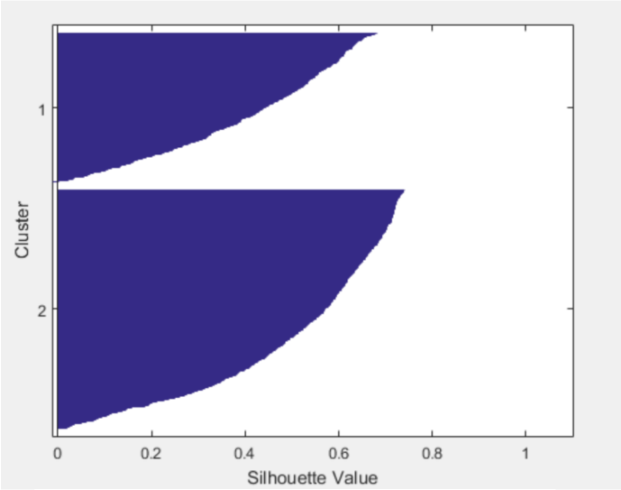
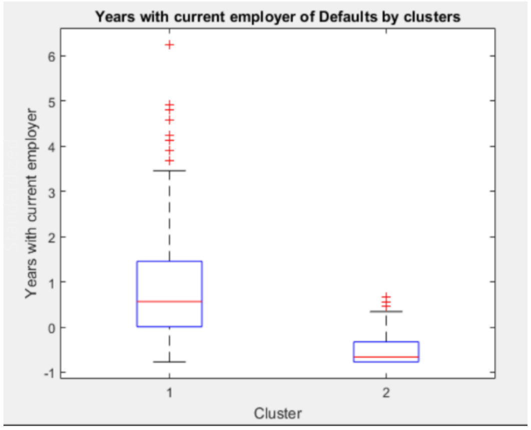
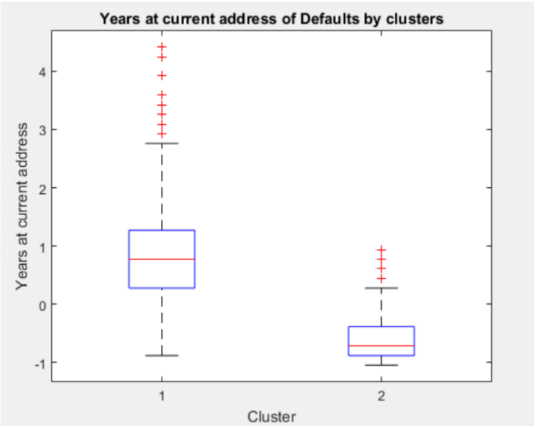


Figure 16

Dividing customers into homogeneous groups involves clustering customers into groups considering their similarities on the predictor variables. The process of dividing customers into similar groups based on their similarities is done using kmeans clustering. For the identification of the best number of clusters, the kmeans clustering algorithm was used with the distance measure, cosine. As shown in figure 15 it ended up with 2 clusters whose average silhouette value is maximum which is 0.4724. The clusters have been separated as the following figure 16 of silhouette plot which shows how perfectly the clusters have been divided. By the clustering, from the test data set 461 observations were allocated to cluster 1 and 739 observations were allocated to cluster 2. Through the analysis it was identified that the cluster which has the highest chance of defaulting in the future is cluster 2.

Refer below graphs to consider the common characteristics of most important variables of cluster two.





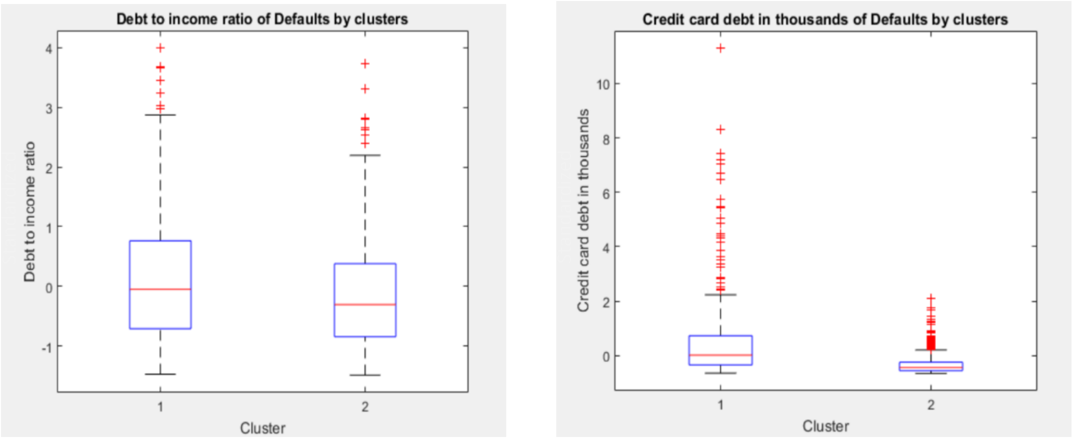


Figure 17

When considering the employ and address variables, it is clear that in cluster two, which is the group of customers who have higher chance to default in the future have the lower median than cluster one. Indebtinc and creddebt variables' medians are also slightly lower in cluster two. When we consider the spread of those variables, the spread of cluster two seems to be lower than that of cluster one. By taking the standard deviation of these variables in each cluster, it seems that there is low standard deviation in cluster two in each important variable.

Medians

	employ	address	debtinc	creddebt
Cluster1	0.56228	0.77626	0.04936	0.02286
Cluster2	0.66298	0.71189	0.30416	-0.43444

Table 2

Standard Deviation

	employ	address	debtinc	creddebt
Cluster1	1.11676	0.94146	1.08031	1.40121
Cluster2	0.30500	0.40731	0.88606	0.34511

Table 2

6. Concluding Remarks

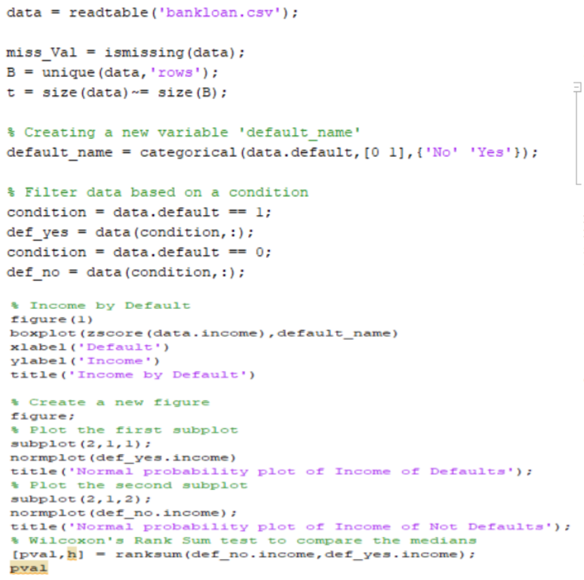
The results of comparison of the common characteristic of most important variable identified in the 2nd cluster with the common characteristics of customers who previously defaulted will be discussed here.

Based on the results of the descriptive analysis, it was observed that the median Years with Current Employer (employ) and median Years at Current Address for customers who had defaulted in the past were lower compared to those who had not. Similarly, these findings were consistent with the results of the clustering, as cluster 2 which was identified as having a higher likelihood of default had a higher median for these two variables compared to the other clusters.

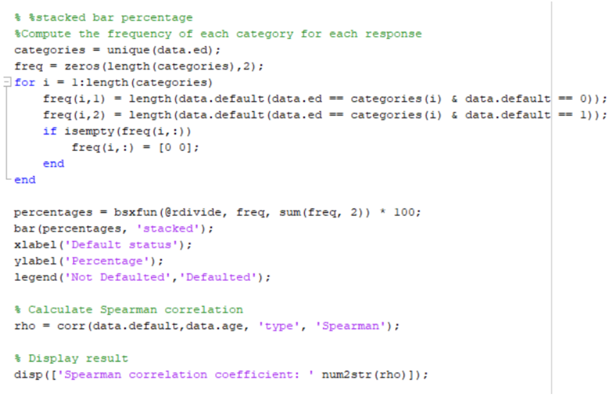
Previous descriptive analysis has indicated that the median debt-to-income ratio credit card debts of individuals who have defaulted in the past is higher than that of those who have not defaulted. However, on the contrary clustering has revealed that individuals with a high likelihood of defaulting have a lower median value compared to others. This is due to cluster 2 having a lower median value for both variables.

Appendices

1) Descriptive Analysis

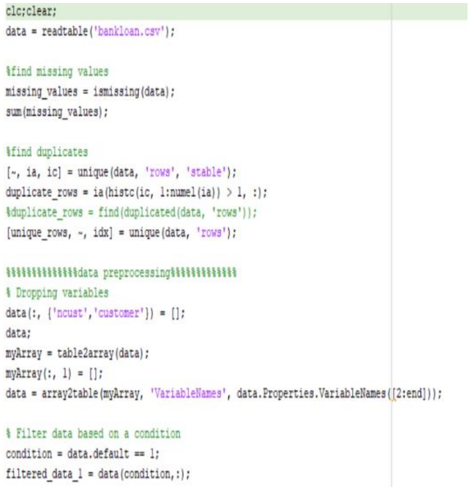


```
data = readtable('bankloan.csv');

miss_Val = ismissing(data);
B = unique(data,'rows');
t = size(data)~= size(B);

% Creating a new variable 'default_name'
default_name = categorical(data.default,[0 1],{'No' 'Yes'});

% Filter data based on a condition
condition = data.default == 1;
def_yes = data(condition,:);
condition = data.default == 0;
def_no = data(condition,:);

% Income by Default
figure(1)
boxplot(zscore(data.income),default_name)
xlabel('Default')
ylabel('Income')
title('Income by Default')

% Create a new figure
figure;
% Plot the first subplot
subplot(2,1,1);
normplot(def_yes.income)
title('Normal probability plot of Income of Defaults');
% Plot the second subplot
subplot(2,1,2);
normplot(def_no.income);
title('Normal probability plot of Income of Not Defaults');
% Wilcoxon's Rank Sum test to compare the medians
[pval,h] = ranksum(def_no.income,def_yes.income);
pval
```

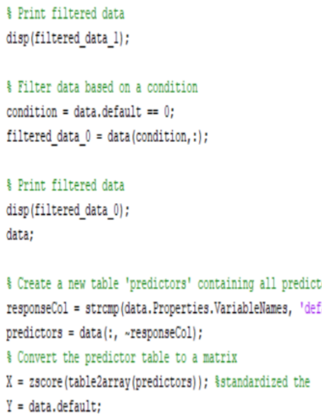


```
% %stacked bar percentage
%Compute the frequency of each category for each response
categories = unique(data.ed);
freq = zeros(length(categories),2);
for i = 1:length(categories)
    freq(i,1) = length(data.default(data.ed == categories(i) & data.default == 0));
    freq(i,2) = length(data.default(data.ed == categories(i) & data.default == 1));
    if isempty(freq(i,:))
        freq(i,:) = [0 0];
    end
end

percentages = bsxfun(@rdivide, freq, sum(freq, 2)) * 100;
bar(percentages, 'stacked');
xlabel('Default status');
ylabel('Percentage');
legend('Not Defaulted','Defaulted');

% Calculate Spearman correlation
rho = corr(data.default,data.age, 'type', 'Spearman');

% Display result
disp(['Spearman correlation coefficient: ' num2str(rho)]);
```

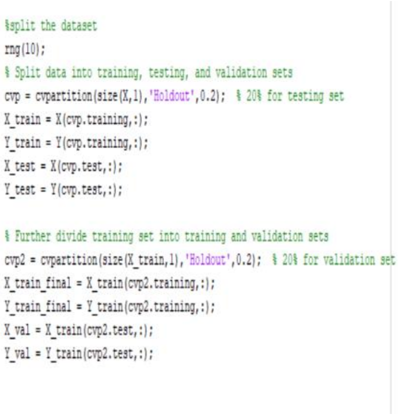
2) Classification



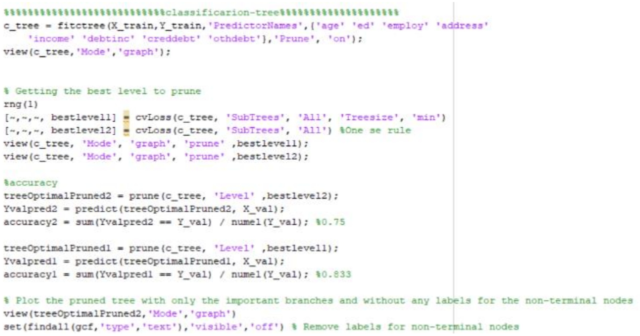
```
clc;clear;
data = readtable('bankloan.csv');

%find missing values
missing_values = ismissing(data);
sum(missing_values);

%find duplicates
[~, ia, ic] = unique(data, 'rows', 'stable');
duplicate_rows = ia(histc(ic, 1:numel(ia)) > 1, :);
%duplicate_rows = find(duplicated(data, 'rows'));
[unique_rows, ~, idx] = unique(data, 'rows');

%%%%%%%%%%%%%%data preprocessing%%%%%%%%%%%%%
% Dropping variables
data(:, {'ncust','customer'}) = [];
data;
myArray = table2array(data);
myArray(:, 1) = [];
data = array2table(myArray, 'VariableNames', data.Properties.VariableNames([2:end]));

% Filter data based on a condition
condition = data.default == 1;
filtered_data_1 = data(condition,:);
```



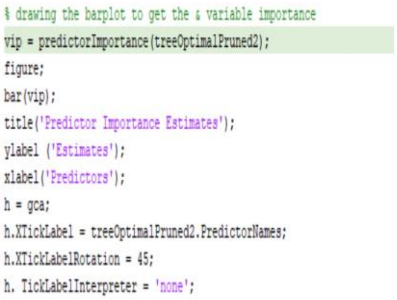
```
% Print filtered data
disp(filtered_data_1);

% Filter data based on a condition
condition = data.default == 0;
filtered_data_0 = data(condition,:);

% Print filtered data
disp(filtered_data_0);
data;

% Create a new table 'predictors' containing all predictor variables
responseCol = strcmp(data.Properties.VariableNames, 'default');
predictors = data(:, ~responseCol);
% Convert the predictor table to a matrix
X = zscore(table2array(predictors)); %standardized the
Y = data.default;
```



```
%split the dataset
rng(10);
% Split data into training, testing, and validation sets
cvp = cvpartition(size(X,1),'Holdout',0.2);  % 20% for testing set
X_train = X(cvp.training,:);
Y_train = Y(cvp.training,:);
X_test = X(cvp.test,:);
Y_test = Y(cvp.test,:);

% Further divide training set into training and validation sets
cvp2 = cvpartition(size(X_train,1),'Holdout',0.2);  % 20% for validation set
X_train_final = X_train(cvp2.training,:);
Y_train_final = Y_train(cvp2.training,:);
X_val = X_train(cvp2.test,:);
Y_val = Y_train(cvp2.test,:);
```

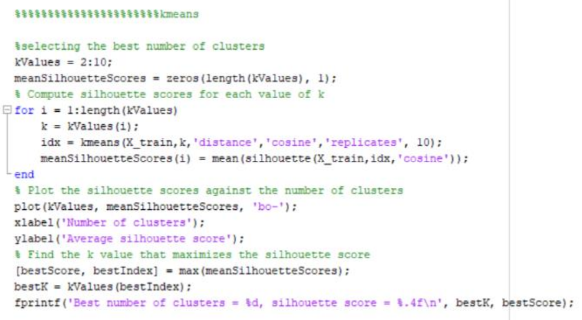


```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%classificarion-tree%%%%%%%%%%%%%%%%%%%%
c_tree = fitctree(X_train,Y_train,'PredictorNames',{'age' 'ed' 'employ' 'address'
    'income' 'debtinc' 'creddebt' 'othdebt'},'Prune', 'on');
view(c_tree,'Mode','graph');


% Getting the best level to prune
rng(1)
[~,~,~, bestlevel1] = cvLoss(c_tree, 'SubTrees', 'All', 'Treesize', 'min')
[~,~,~, bestlevel2] = cvLoss(c_tree, 'SubTrees', 'All') %One se rule
view(c_tree, 'Mode', 'graph', 'prune' ,bestlevel1);
view(c_tree, 'Mode', 'graph', 'prune' ,bestlevel2);

%accuracy
treeOptimalPruned2 = prune(c_tree, 'Level' ,bestlevel2);
Yvalpred2 = predict(treeOptimalPruned2, X_val);
accuracy2 = sum(Yvalpred2 == Y_val) / numel(Y_val); %0.75

treeOptimalPruned1 = prune(c_tree, 'Level' ,bestlevel1);
Yvalpred1 = predict(treeOptimalPruned1, X_val);
accuracy1 = sum(Yvalpred1 == Y_val) / numel(Y_val); %0.833

% Plot the pruned tree with only the important branches and without any labels for the non-terminal nodes
view(treeOptimalPruned2,'Mode','graph')
set(findall(gcf,'type','text'),'visible','off') % Remove labels for non-terminal nodes
```

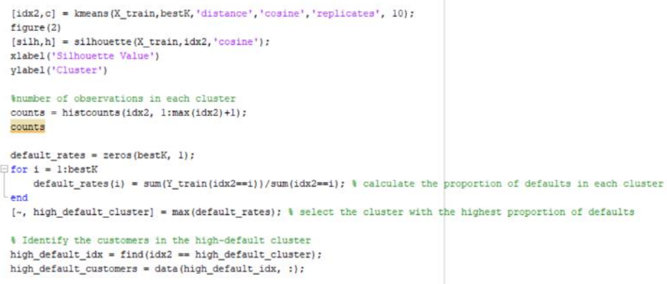


```
% drawing the barplot to get the & variable importance
vip = predictorImportance(treeOptimalPruned2);
figure;
bar(vip);
title('Predictor Importance Estimates');
ylabel ('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = treeOptimalPruned2.PredictorNames;
h.XTickLabelRotation = 45;
h. TickLabelInterpreter = 'none';
```

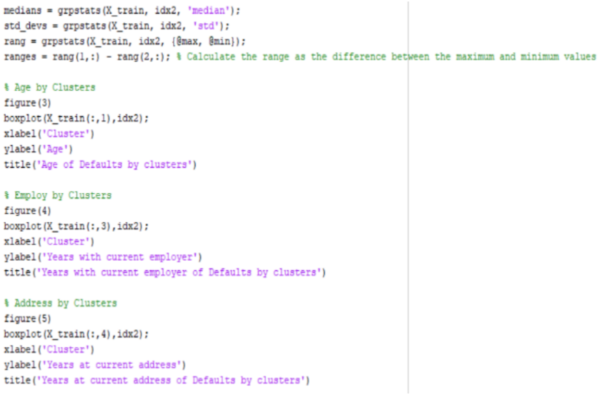
3) Clustering



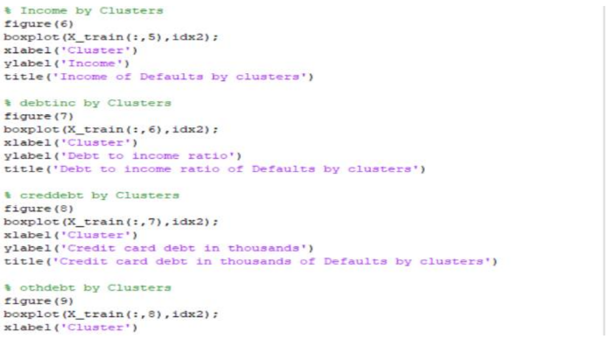
```
%%%%%%%%%%%%%%%%%%%%%%kmeans

%selecting the best number of clusters
kValues = 2:10;
meanSilhouetteScores = zeros(length(kValues), 1);
% Compute silhouette scores for each value of k
for i = 1:length(kValues)
    k = kValues(i);
    idx = kmeans(X_train,k,'distance','cosine','replicates', 10);
    meanSilhouetteScores(i) = mean(silhouette(X_train,idx,'cosine'));
end
% Plot the silhouette scores against the number of clusters
plot(kValues, meanSilhouetteScores, 'bo-');
xlabel('Number of clusters');
ylabel('Average silhouette score');
% Find the k value that maximizes the silhouette score
[bestScore, bestIndex] = max(meanSilhouetteScores);
bestK = kValues(bestIndex);
fprintf('Best number of clusters = %d, silhouette score = %.4f\n', bestK, bestScore);
```



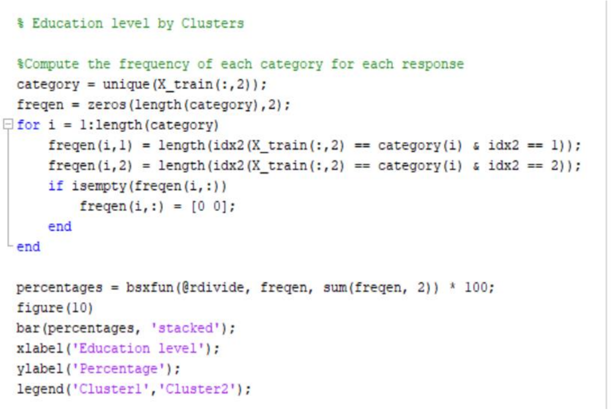
```
[idx2,c] = kmeans(X_train,bestK,'distance','cosine','replicates', 10);
figure(2)
[silh,h] = silhouette(X_train,idx2,'cosine');
xlabel('Silhouette Value')
ylabel('Cluster')

%number of observations in each cluster
counts = histcounts(idx2, 1:max(idx2)+1);
counts

default_rates = zeros(bestK, 1);
for i = 1:bestK
    default_rates(i) = sum(Y_train(idx2==i))/sum(idx2==i); % calculate the proportion of defaults in each cluster
end
[~, high_default_cluster] = max(default_rates); % select the cluster with the highest proportion of defaults

% Identify the customers in the high-default cluster
high_default_idx = find(idx2 == high_default_cluster);
high_default_customers = data(high_default_idx, :);
```



```
medians = grpstats(X_train, idx2, 'median');
std_devs = grpstats(X_train, idx2, 'std');
rang = grpstats(X_train, idx2, {@max, @min});
ranges = rang(1,:) - rang(2,:); % Calculate the range as the difference between the maximum and minimum values

% Age by Clusters
figure(3)
boxplot(X_train(:,1),idx2);
xlabel('Cluster')
ylabel('Age')
title('Age of Defaults by clusters')

% Employ by Clusters
figure(4)
boxplot(X_train(:,3),idx2);
xlabel('Cluster')
ylabel('Years with current employer')
title('Years with current employer of Defaults by clusters')

% Address by Clusters
figure(5)
boxplot(X_train(:,4),idx2);
xlabel('Cluster')
ylabel('Years at current address')
title('Years at current address of Defaults by clusters')
```



```
% Income by Clusters
figure(6)
boxplot(X_train(:,5),idx2);
xlabel('Cluster')
ylabel('Income')
title('Income of Defaults by clusters')

% debtinc by Clusters
figure(7)
boxplot(X_train(:,6),idx2);
xlabel('Cluster')
ylabel('Debt to income ratio')
title('Debt to income ratio of Defaults by clusters')

% creddebt by Clusters
figure(8)
boxplot(X_train(:,7),idx2);
xlabel('Cluster')
ylabel('Credit card debt in thousands')
title('Credit card debt in thousands of Defaults by clusters')

% othdebt by Clusters
figure(9)
boxplot(X_train(:,8),idx2);
xlabel('Cluster')
```



```
% Education level by Clusters

%Compute the frequency of each category for each response
category = unique(X_train(:,2));
freqen = zeros(length(category),2);
for i = 1:length(category)
    freqen(i,1) = length(idx2(X_train(:,2) == category(i) & idx2 == 1));
    freqen(i,2) = length(idx2(X_train(:,2) == category(i) & idx2 == 2));
    if isempty(freqen(i,:))
        freqen(i,:) = [0 0];
    end
end

percentages = bsxfun(@rdivide, freqen, sum(freqen, 2)) * 100;
figure(10)
bar(percentages, 'stacked');
xlabel('Education level');
ylabel('Percentage');
legend('Cluster1','Cluster2');
```