Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	5521960, 5539729, 5504970, 5586034, 5576453, 5577172
Module Code:	IB98D0
Module Title:	Advanced Data Analysis
Submission Deadline:	18 March 2024
Date Submitted:	17 March 2024
Word Count:	1820
Number of Pages:	14
Question Attempted: *(question number/title, or description of assignment)*	Question 1: Cluster Analysis
Have you used Artificial Intelligence (AI) in any part of this assignment?	No

Executive Summary

This report gives insights from cluster analysis aimed at improving operational efficiency and customer satisfaction. By considering hierarchical and non-hierarchical methods, three distinct clusters were classified, varying in financial status and loan portfolio. We concluded that tailoring loan products and marketing strategies are paramount for fulfilling different customer's needs. Additionally, we recommended that providing flexibility on loan offerings, streamlining application procedures and customising customer services are conducive to effective operations and customer satisfaction.

Table of Contents

1 Introduction

Recognising the challenges in loan approval efficiency, risk assessments and customer satisfaction, the implementation of applying cluster analysis to our existing customers could be potentially beneficial. The objective is to understand how customers behave when they use our loan services. By grouping customers with similar profiles, we aim to gain valuable insights into their portfolio, such as employment length and current balance, thereby tailoring marketing strategies, loan products and customer services.

2 Data Preparation

The dataset contains loan records issued between 2012 and 2013, with 53 variables and 50,000 observations.

Variable	Reason
annual_Inc	Indicate financial stability, use in conjunction with other variables to determine borrowers' credit risk
delinq_2yrs	High number of delinquencies can indicate financial irresponsibility
dti	High debt-to-income ratio borrowers can be categorized as high-risk customers
emp_length	Borrowers with longer employment histories may have greater loan repayment capacity, thereby being perceived as lower-risk borrower
funded_amnt	Higher funded amount may indicate greater financial need
inq_last_6mths	A higher number of inquiries in the past 6 months may indicate active credit-seeking behaviour
pub_rec	High number of derogatory public record indicate high-risk borrowers
revol_util	Provides insight into how borrowers manage their credit
tot_cur_bal	Reflect overall debt burden
total_credit_rv	Indicate borrower's credit capacity which is reflective of borrower's creditworthiness
total_acc	Borrowers with a high number of credit lines may have a history of actively seeking credit

Table 1. Reasons for including selected variables

After selecting the variables, we performed data checking and identified variables with missing values, as shown in Table 2. Upon examining the dataset, we concluded that these values were missing at random. For instance, some observations lacked data for "tot_cur_bal" and "total_credit_rv", but other variables indicated active credit, suggesting the absence wasn't indicative of no credit. We decided to eliminate records with any missing values, leaving us with 34,052 observations. Despite the potential for information loss, these remaining observations were deemed sufficient for our analysis. We then selected 500 samples from this dataset through random sampling for further cluster analysis.

Variable	Number of NA values	% of NA values from 50K observations
emp_length	1802	3.60%
revol_util	31	0.06%
total_cur_bal	14,618	29.24%
total_credit_rv	14,618	29.24%

Table 2. NA values of selected variables

3 Methodology

3.1 Outliers

The Mahalanobis distance method was employed to address the potential impact of extreme values on classification performance in identifying meaningful clusters. This method calculates the distance between each observation and the mean centre. Observations with a p-value less than 0.001 can be classified as outliers. Table 3 presents the result, showing that 3.8% of the sample, or 19 out of 500 observations, had a p-value of less than 0.001.

Figure 1 shows the distribution of each variable from the sample. By comparing with the 19 outliers shown in Table 3, it was noted that the highest value of “annual_inc”, “total_cur_bal” and “total_credit_rv” were included among the observations. Furthermore, unusual patterns were observed among the observations. For instance, the 6th observation in Table 3 exhibited a relatively high annual income with the highest number of delinquencies. Such cases may lead to poor cluster results as they may not align well with the general patterns of the sampled data. Consequently, these outliers were removed from our sample.

```
> summary(loandata_sample)
   annual_inc       delinq_2yrs         dti          emp_length      funded_amnt     inq_last_6mths
 Min.   : 15000   Min.   :0.000   Min.   : 0.48   Min.   : 1.000   Min.   : 1200   Min.   :0.000
 1st Qu.: 47000   1st Qu.:0.000   1st Qu.:11.25   1st Qu.: 2.000   1st Qu.: 8400   1st Qu.:0.000
 Median : 64000   Median :0.000   Median :17.09   Median : 6.000   Median :12212   Median :1.000
 Mean   : 72320   Mean   :0.258   Mean   :17.33   Mean   : 6.046   Mean   :14268   Mean   :0.988
 3rd Qu.: 90000   3rd Qu.:0.000   3rd Qu.:23.13   3rd Qu.:10.000   3rd Qu.:19812   3rd Qu.:2.000
 Max.   :357000   Max.   :7.000   Max.   :34.99   Max.   :10.000   Max.   :35000   Max.   :5.000
    pub_rec        revol_util       tot_cur_bal     total_credit_rv     total_acc
 Min.   :0.000   Min.   :0.0010   Min.   :  1518   Min.   :  1000   Min.   : 3.00
 1st Qu.:0.000   1st Qu.:0.4522   1st Qu.: 24757   1st Qu.: 13500   1st Qu.:16.75
 Median :0.000   Median :0.6190   Median : 59510   Median : 22656   Median :23.00
 Mean   :0.066   Mean   :0.5983   Mean   :128307   Mean   : 27909   Mean   :24.64
 3rd Qu.:0.000   3rd Qu.:0.7768   3rd Qu.:200403   3rd Qu.: 36350   3rd Qu.:32.00
 Max.   :2.000   Max.   :0.9760   Max.   :940724   Max.   :122700   Max.   :63.00
```

Figure 1. Summary of 11 selected variables from sampled data

> LoanMaha %>% filter(MahaPvalue<0.001)

	annual_inc	delinq_2yrs	dti	emp_length	funded_amnt	inq_last_6mths	pub_rec	revol_util	tot_cur_bal	total_credit_rv	total_acc	maha	MahaPvalue
1	64000	0	3.34	10	6000	2	2	0.366	4244	11600	25	61.44419	1.929718e-09
2	357000	0	22.15	4	30000	1	0	0.694	459994	113700	34	69.19455	6.341324e-11
3	50000	4	23.21	10	7200	2	0	0.749	273982	7600	29	29.80953	9.202863e-04
4	220000	0	30.28	3	27575	0	0	0.412	495409	115000	50	30.36884	7.453498e-04
5	150000	1	4.79	8	13225	0	0	0.544	805972	57348	26	31.07904	5.693464e-04
6	110000	7	13.75	7	5000	0	0	0.959	199999	29600	31	91.44314	2.769875e-15
7	98000	4	10.71	10	5000	2	0	0.771	63544	18000	33	32.48646	3.321251e-04
8	140000	0	26.71	10	14000	0	0	0.770	294326	117400	24	31.29684	5.240174e-04
9	48000	5	14.00	8	10000	1	0	0.881	11966	5300	14	47.81803	6.699975e-07
10	36400	5	22.48	5	14500	0	0	0.585	9217	13400	46	46.47995	1.173787e-06
11	300000	0	7.44	2	5000	2	0	0.878	940724	45100	32	67.71014	1.224636e-10
12	85000	1	8.47	7	9000	1	0	0.953	687569	43800	11	32.14811	3.783020e-04
13	85000	0	18.31	3	24000	2	0	0.957	405510	115900	19	38.37546	3.264379e-05
14	32500	5	33.90	8	10850	0	0	0.939	21869	10200	29	46.65975	1.088783e-06
15	225000	0	11.80	2	35000	0	0	0.706	84926	107400	34	37.66486	4.340201e-05
16	350000	0	5.96	6	20000	1	0	0.365	240116	17000	18	82.97720	1.306457e-13
17	31000	0	17.69	3	10000	0	0	0.839	16685	98000	11	34.20038	1.708028e-04
18	160000	0	10.79	2	18000	1	0	0.064	562646	122700	51	31.11222	5.622025e-04
19	140000	0	6.50	1	4375	2	1	0.334	56008	7500	47	32.46058	3.354532e-04

Table 3. Outliers identified by Mahalanobis distance

3.2 Multicollinearity

As highly correlated variables can overpower other variables and influence the clustering process, multicollinearity check has been conducted to eliminate redundant variables. Typically, a correlation value greater than the absolute value of 0.8 indicates a strong relationship between variables. According to the correlation results shown in Table 4, all variables exhibit correlation between -0.5 and 0.5, indicating suitability for cluster analysis without the need for PCA or FA prior to clustering.

> lowerCor(LoanMaha_new)

	annl_	dln_2	dti	emp_l	fndd_	in__6	pb_rc	rvl_t	tt_c_	ttl__	ttl_c
annual_inc	1.00										
delinq_2yrs	0.02	1.00									
dti	-0.21	0.08	1.00								
emp_length	0.08	0.09	0.12	1.00							
funded_amnt	0.48	-0.02	0.04	0.11	1.00						
inq_last_6mths	0.03	-0.01	0.00	-0.13	0.01	1.00					
pub_rec	-0.02	-0.08	-0.11	0.00	-0.03	-0.03	1.00				
revol_util	0.04	-0.07	0.28	0.14	0.10	-0.11	-0.04	1.00			
tot_cur_bal	0.49	0.09	0.05	0.11	0.32	0.01	0.02	0.08	1.00		
total_credit_rv	0.37	-0.01	0.08	0.10	0.36	0.06	-0.06	-0.27	0.33	1.00	
total_acc	0.35	0.21	0.24	0.19	0.31	0.07	0.05	-0.01	0.37	0.39	1.00

Table 4. Correlations between each variable

3.3 Standardisation

According to Table 5, the means and standard deviations across the 11 variables vary significantly due to differences in scale, especially “annual_inc”, “funded_amnt”, total_cur_bal” and “total_credit_rv”, highlighting the necessity for standardization in the pre-processing step to minimise bias cluster formation. Z-score standardisation was applied, ensuring each variable has a mean of zero and a standard deviation of one, thereby guaranteeing equal weights to generate meaningful clusters.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
annual_inc	1	500	72319.56	39788.99	64000.00	66868.35	29088.61	15000.00	357000.00	342000.00	2.60
delinq_2yrs	2	500	0.26	0.75	0.00	0.08	0.00	0.00	7.00	7.00	4.46
dti	3	500	17.33	8.05	17.09	17.14	8.85	0.48	34.99	34.51	0.19
emp_length	4	500	6.05	3.49	6.00	6.18	5.93	1.00	10.00	9.00	-0.15
funded_amnt	5	500	14268.00	7988.05	12212.50	13583.69	7746.58	1200.00	35000.00	33800.00	0.68
inq_last_6mths	6	500	0.99	1.11	1.00	0.84	1.48	0.00	5.00	5.00	0.89
pub_rec	7	500	0.07	0.26	0.00	0.00	0.00	0.00	2.00	2.00	3.84
revol_util	8	500	0.60	0.23	0.62	0.61	0.24	0.00	0.98	0.98	-0.52
tot_cur_bal	9	500	128306.67	141518.34	59510.50	103486.42	69952.03	1518.00	940724.00	939206.00	1.73
total_credit_rv	10	500	27909.16	20693.51	22655.50	24732.88	15789.69	1000.00	122700.00	121700.00	1.71
total_acc	11	500	24.64	11.13	23.00	23.95	10.38	3.00	63.00	60.00	0.63

Table 5. Summary statistics of each variable

3.4 Perform Cluster Analysis - Hierarchical and Non-hierarchical

We explored two clustering algorithms: hierarchical and a combination of hierarchical and non-hierarchical methods. While the results from both methods are similar, we found that the latter yielded slightly better outcomes in terms of achieving a more balanced number of observations across clusters and providing a sensible interpretation of cluster profiles.

The different distance and linkage measures used in hierarchical method can be found in Appendix 4, and cluster result in Appendix 5.

We utilised the gap statistic method to identify the optimal number of clusters. Figure 2 suggested for one cluster according to the "1 standard deviation rule" (Tibshirani et al., 2001). However, to tailor loan products and business services across different customer segments effectively, we sought multiple clusters. We experimented with both three and four clusters using K-means clustering and found that three clusters were easier to interpret and profile.

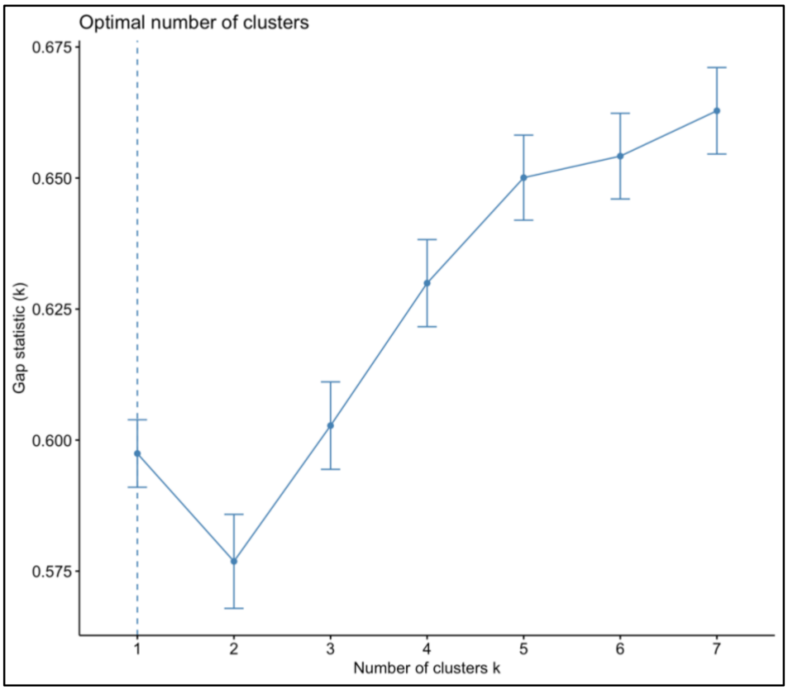


Figure 2. Gap statistic plot

4 Results

4.1 Cluster Analysis

The borrower's profiles are grouped into three clusters and the results can be seen in Table 6. The number of observations is 30, 282, and 169 in each cluster respectively.

Variables	Cluster 1	Cluster 2	Cluster 3
annual_inc	-0.08	-0.457	0.776
delinq_2yrs	-0.319	-0.127	0.268
dti	-0.425	0.001	0.074
emp_length	-0.017	-0.133	0.225
funded_amnt	-0.129	-0.433	0.746
inq_last_6mths	-0.112	-0.012	0.039
pub_rec	3.873	-0.258	-0.258
revol_util	-0.162	0.007	0.017
tot_cur_bal	0.086	-0.452	0.739
total_credit_rv	-0.227	-0.415	0.733
total_acc	0.184	-0.449	0.716

Table 6. Cluster results from a combination of hierarchical and non-hierarchical method

Cluster 1: Average income, average employment length, and good credit management but with presence of derogatory public records

This cluster represents borrowers with annual incomes and employment similar to the sample mean. They exhibit responsible credit management, with low debt-to-income ratio and a good payment history over the past two years. A lower number of inquiries in the past 6 months also indicates minimal recent credit-seeking behaviour. However, a notable feature is the higher number of derogatory public records, suggesting past financial challenges or legal issues.

Cluster 2: Low income, limited employment history, and conservative borrowing behaviour

Borrowers in this cluster have lower incomes and shorter employment histories, suggesting younger individuals who have recently entered the workforce or those who have faced limited job opportunities. A significant portion of borrowers in this group rents their homes. They demonstrate conservative borrowing behaviour and relatively clean credit histories, with lower loan amounts and fewer outstanding balances. Additionally, they have fewer number of credit lines and access to less credit compared to other clusters.

Cluster 3: High income, stable employment history, credit-active borrowers, and significant credit usage

In this cluster, borrowers have higher incomes and longer employment histories, indicating financial stability. However, they are also identified as high-risk borrowers due to a history of delinquency over the past two years and high debt-to-income ratios. Most borrowers in this cluster own their homes through mortgage arrangements. Moreover, they tend to take out high loan amounts and exhibit active credit-seeking behaviour, with numerous credit lines and inquiries in the past six months. Additionally, they heavily rely on credit cards or other revolving credit accounts to meet their financial obligations, resulting in high revolving line utilisation rates and significant total current balances across all accounts.

4.2 External Validation

An additional random sample of 500 observations has been generated to validate the cluster results obtained in Table 6.

Following the same data preparation and methodology as described in Section 2 and 3, the findings from the gap statistic, as illustrated in Figure 3, align with our previous tests, indicating that the optimal number of clusters is three clusters or more.

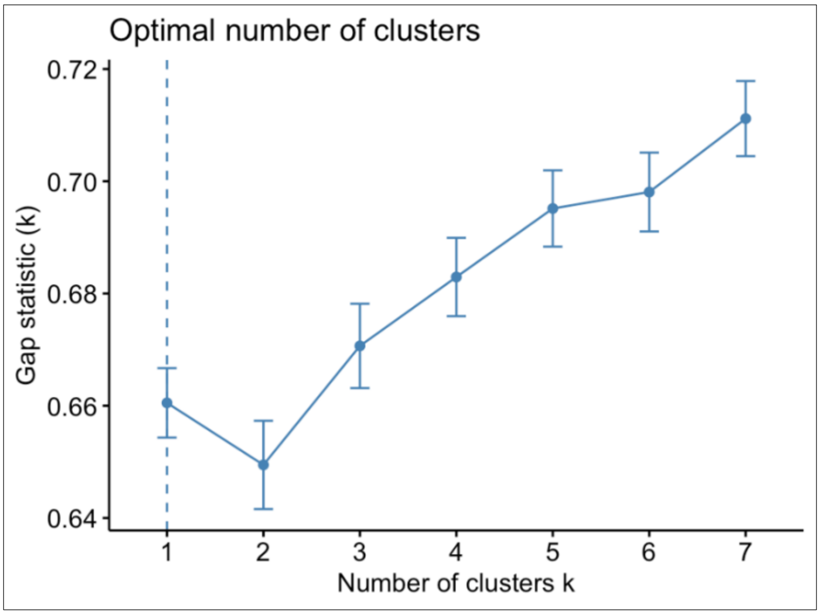


Figure 3. Gap statistic of external validation

Table 7 displays the cluster results obtained from the combination of hierarchical and non-hierarchical method. The number of observations in each cluster is 23, 198, and 258 respectively, totalling approximately 12% cases assigned to different clusters compared to previous cluster results. Furthermore, in contrast to our previous cluster result, we observed a shift in one variable, “revol_util”, to a different cluster. Initially, “revol_util” was predominantly linked with the cluster of high-income earners. However, the validation results indicate that it aligns more closely with the average-income cluster. Nevertheless, this change does not affect the overall profiling of each cluster, the cluster solution is stable and should represent the population.

Variable	Cluster 1	Cluster 2	Cluster 3
annual_inc	0.119	0.628	-0.493
delinq_2yrs	-0.181	0.061	-0.031
dti	-0.266	0.086	-0.042
emp_length	-0.099	0.314	-0.232
funded_amnt	-0.099	0.681	-0.514
inq_last_6mths	-0.233	0.174	-0.112
pub_rec	4.448	-0.224	-0.224
revol_util	0.053	-0.060	0.041
tot_cur_bal	-0.077	0.666	-0.504
total_credit_rv	-0.391	0.596	-0.422
total_acc	-0.018	0.626	-0.479

Table 7. Cluster results of external validation dataset

5 Recommendations

Cluster 1: Average income borrowers with presence of derogatory public records

Customers from Cluster 1 are deemed to have a decent chance to further improve their creditability due to their average income and employment length. To help improve their public records, we could offer customised loan products such that credit scores can be improved if they demonstrate consistency in making on-time payments.

To further support average income borrowers improving their credit scores, educational contents such as tips and successful cases could be shared through social media platforms and advertisements. Customers engagement can also be achieved by frequently replying to messages from them.

Knowing that customers may have poor loan historical records, we may establish a dedicated team consists of pre-eminent customer relationship managers, specialising on providing professional advice to overcome financial challenges.

Cluster 2: Low-income borrowers with conservative borrowing behaviour

Cluster 2 are customers with a relatively low income and shorter employment period, suggesting that they are mostly belonged to a younger age group. To cater their financial needs, we could provide loan products with longer repayment periods and lower interest rates, thus encouraging them to apply for loan services continuously while facilitating a positive credit score.

In terms of marketing strategy, special offers such as referral rewards and cash rebates could be attractive to them, which can be attained by referring friends and colleagues. While referrers could gain benefits from discounts, we could also be benefited by acquiring new customers, facilitating customer satisfaction and organisation's reputation.

In view of their conservative borrowing behaviour, a lenient and automated loan approval process could be proposed. This could be achieved by setting a lower minimum salary requirement and implementing a digital platform for online applications, hence reducing manually approval process and maintaining high operational efficiency.

Cluster 3: High-income borrowers with significant credit usage

For Cluster 3 customers who are financially stable but have a riskier credit profile due to past delinquency records, we may offer a higher loan amount and a flexible monthly interest rate. A lower rate can be rewarded if customers pay back on time. However, due to poor repayment history, it is suggested that a shorter repayment period should be enforced by customising loan terms, hence mitigating the risk of financial loss.

Cluster 3 has the largest customer-based implies that they are the major source of company's profit. In view of this, utilising a multitude of communicating channels plays a vital role in maintaining connections with them. We could promote new products and services through social media platforms and send personalised emails to notify their credit usage. By leveraging digital channels effectively, we could increase chances to retain these valuable customers.

It is noticeable that number of enquiries for Cluster 3 is higher than the average. For this reason, understanding their potential issues and pain points is of utmost importance to alleviate their confusions and frustrations. For instance, they might raise concerns on changing repayment schedule and reduction on credit scores due to late payment. By providing pragmatic and actionable solutions, maintaining customer loyalty, and fostering long-lasting relationship could be achieved.

6 Conclusions

In conclusion, implementing cluster analysis on our customers is highly suggested and could be beneficial to overcoming operational efficiency and customer satisfaction. By classifying into 3 different clusters, we can identify and generalise customers' financial status and loan portfolio, enabling us to determine business strategies and directions.

To maintain reputation and competitiveness in lending business, we aim to utilise cluster results by tailoring loan products, marketing strategies and customer services, thereby developing rapport with customers and facilitating seamless operational process.

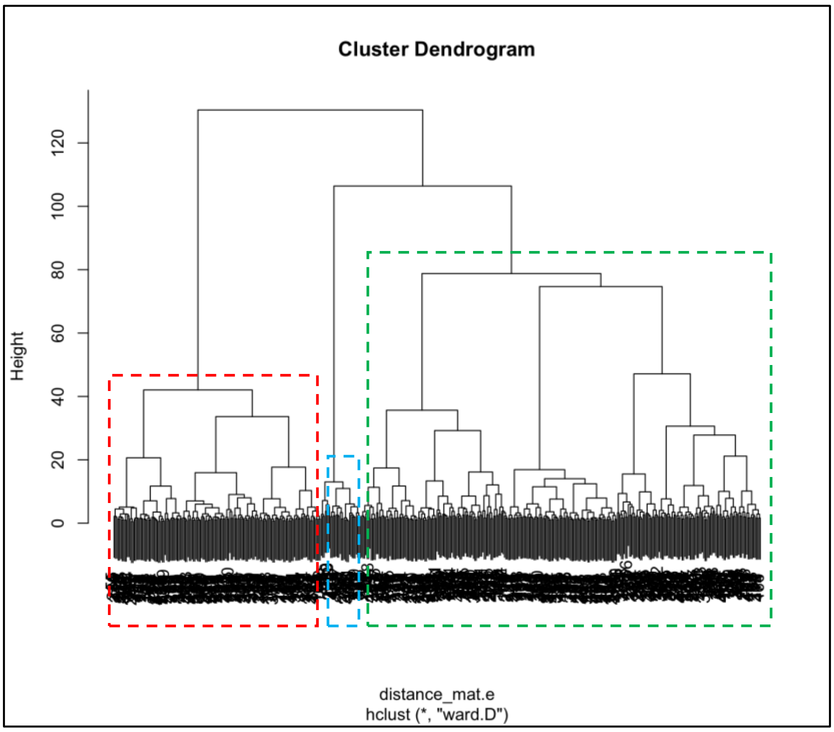
7 References

Tibshirani, R., Walther, G. & Hastie, T. 2001, "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society. Series B, Statistical methodology,* vol. 63, no. 2, pp. 411-423.

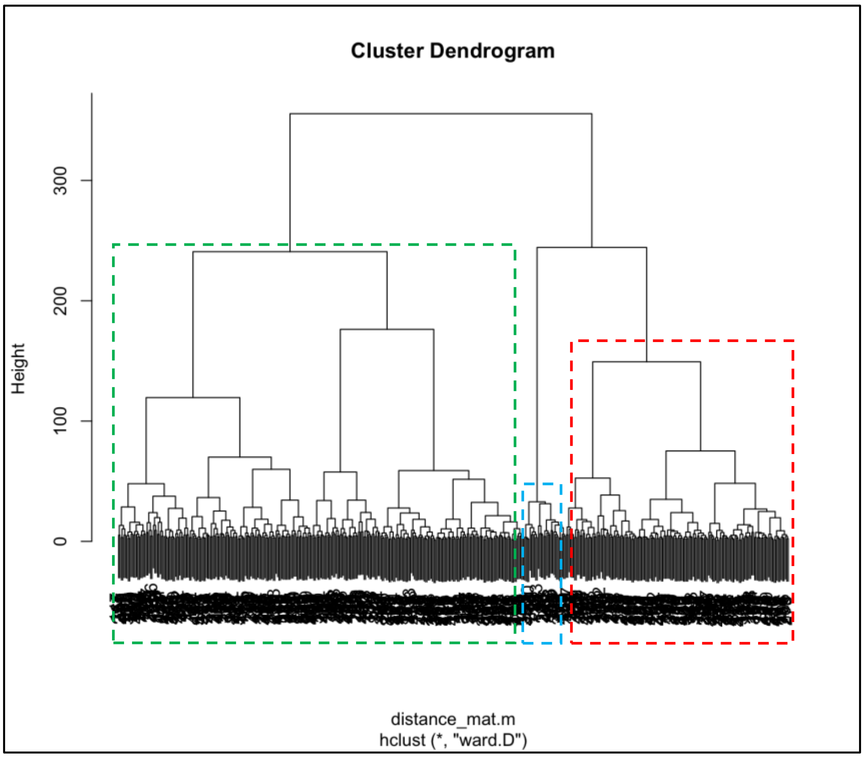
8 Appendices

Variable	Description
annual_inc	The self-reported annual income provided by the borrower during registration.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
funded_amnt	The total amount committed to that loan at that point in time.
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
pub_rec	Number of derogatory public records
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
tot_cur_bal	Total current balance of all accounts
total_credit_rv	Total revolving credit
total_acc	The total number of credit lines currently in the borrower's credit file

Appendix 1. Data dictionary of selected variables



Appendix 2. Dendrogram of Ward's method with Euclidean distance



Appendix 3. Dendrogram of Ward's method with Manhattan distance

Distance Measure	Linkage Method	Clusters		
		1	2	3
Euclidean	Ward	297	154	30
Manhattan	Ward	291	160	30

Appendix 4. Distribution of observations with Ward's method and Manhattan distance

	Cluster 1	Cluster 2	Cluster 3
annual_inc	0.259	-0.456	-0.080
delinq_2yrs	0.185	-0.276	-0.319
dti	0.263	-0.399	-0.424
emp_length	0.325	-0.587	-0.017
funded_amnt	0.316	-0.551	-0.129
inq_last_6mths	0.069	-0.104	-0.112
pub_rec	-0.258	-0.258	3.873
revol_util	0.195	-0.324	-0.162
tot_cur_bal	0.261	-0.492	0.086
total_credit_rv	0.154	-0.237	-0.227
total_acc	0.351	-0.673	0.184

Appendix 5. Cluster result from hierarchical method

````
---
title: "Untitled"
output: html_document
date: "2024-03-07"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(psych)
library(psychTools)
library(readxl)
library(GPArotation)
library(factoextra)
library(cluster)
library(dplyr)
```

#Load data
```{r}
loandata <- read_excel("loan_data_ADA_assignment.xlsx")
summary(loandata)
```

#Data preparation
```{r}
#Select targeted variables
loandata_variables <-
select(loandata,"annual_inc","delinq_2yrs","dti","emp_length","funded_amnt","inq_last_6mths","pub_rec","revol_util","tot_cur_bal","total_credit_rv","total_acc")

#View data summary
summary(loandata_variables)

#Remove NA values
loandata_variables_clean <- na.omit(loandata_variables)
summary(loandata_variables_clean)

#Create random sample of 500 observations
set.seed(4)
loandata_sample <- slice_sample(loandata_variables_clean, n=500)
summary(loandata_sample)
```

#Data check-Mahalanobis distance
```{r}
#Describe data
#Huge variation of mean and standard deviation-need to standardize the data
describe(loandata_sample)

#Calculate mahalanobis distance
maha <- mahalanobis(loandata_sample,colMeans(loandata_sample),cov(loandata_sample))
print(maha)

#Check p-value of the mahalanobis distance
MahaPvalue <-pchisq(maha,df=10,lower.tail = FALSE)
print(MahaPvalue)

#filter p-value<0.001 (check for outliers)
#19 observations (3.8% of total sample)
print(sum(MahaPvalue<0.001))

#Add the mahalanobis distance and its p-value to the data
LoanMaha<-cbind(loandata_sample, maha, MahaPvalue)
LoanMaha %>% filter(MahaPvalue<0.001)
````
````

```
#Remove outliers
LoanMaha_new <- filter(LoanMaha, MahaPvalue>=0.001)


```

#Data check-Correlation
```{r}
#Remove maha and MahaPvalue columns from dataset
LoanMaha_new<-select(LoanMaha_new,-maha,-MahaPvalue)

#Check correlation between variables
#None >0.8
LoandataMatrix <- cor(LoanMaha_new)
round(LoandataMatrix,2)

```

#Data preparation-Standardize data
```{r}
LoanMaha_scaled <-scale(LoanMaha_new)
```

#Cluster analysis-gap statistic
```{r}
#Determine optimal number of clusters
#Calculate gap statistic
gap_stat_h <- clusGap(LoanMaha_scaled, FUN = hcut, nstart = 25, K.max = 7, B = 50)

#Produce the plot
#Optimal 3+ clusters--choose 3 clusters
fviz_gap_stat(gap_stat_h)

```
#Cluster analysis-distance measure
```{r}
#Try several distance measures
#Squared Euclidean distance, chebychev not exist in the dist function
distance_mat.e <- dist(LoanMaha_scaled, method = 'euclidean')
distance_mat.m <- dist(LoanMaha_scaled, method = 'manhattan')
```

#Cluster analysis-linkage method
```{r}
#Finding best linkage method
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

#Compute agglomerative coefficient
ac <- function(x) {
agnes(LoanMaha_scaled, method = x)$ac
}

#Calculate agglomerative coefficient for each linkage method
#Ward (0.95) closest to 1 therefore the best method
sapply(m, ac)
```

#Cluster analysis-clustering
```{r}
#Hierarchical procedures-agglomerative methods

#Euclidean-ward
#set.seed(240)
```

```
#Hierar_cl <- hclust(distance_mat.e, method = "ward.D")
#Hierar_cl
#Manhattan-ward
set.seed(240)
Hierar_cl.m <- hclust(distance_mat.m, method = "ward.D")
Hierar_cl.m
#Manhattan-complete
#set.seed(240)
#Hierar_cl.e <- hclust(distance_mat.m, method = "complete")
#Hierar_cl.e
#Euclidean-complete
#set.seed(240)
#Hierar_cl.c <- hclust(distance_mat.e, method = "complete")
#Hierar_cl.c

#Plotting dendogram
plot(Hierar_cl.m)

#Choosing the number of cluster
#fit <- cutree(Hierar_cl, k = 3 )
#fit
fit.m <- cutree(Hierar_cl.m, k = 3 )
fit.m
#fit.a <- cutree(Hierar_cl.e, k = 3 )
#fit.a
#fit.c <- cutree(Hierar_cl.c, k = 3 )
#fit.c

#Find number of observations in each cluster
#Euclidean-ward: 297 154 30
#table(fit)
#Manhattan-ward: 291 160 30--*best result*
table(fit.m)
#Manhattan-complete: 334 30 117
#table(fit.a)
#Euclidean-complete: 99 358 24
#table(fit.c)

#Append cluster labels to original data
final_data<-cbind(LoanMaha_scaled, cluster=fit.m)

#Mean values for each cluster
hcentres<-aggregate(x=final_data, by=list(cluster=fit.m), FUN="mean")
print(hcentres)
``` 

#Cluster analysis-Non-Hierarchical
```{r}
#K-means clustering
#30 282 169 *better result than hierarchical method*--BUT only small difference--still choose hierarchical method
set.seed(240)
k_cl <- kmeans(LoanMaha_scaled,3,nstart=25)
k_cl

clustering_vector <- k_cl$cluster
print(clustering_vector)
``` 

#External validation-Data Sample
```{r}
#Take 500 sample
set.seed(5)
loandata_sample_validate <- slice_sample(loandata_variables_clean, n=500)
summary(loandata_sample_validate)
``` 
```
```

```
#External validation-Mahalanobis distance
```{r}
#Describe data
#Huge variation of mean and standard deviation-need to standardize the data
describe(loandata_sample_validate)

#Calculate mahalanobis distance
maha_validate <-
mahalanobis(loandata_sample_validate,colMeans(loandata_sample_validate),cov(loandata_sample_validate))
print(maha_validate)

#Check p-value of the mahalanobis distance
MahaPvalue.v <-pchisq(maha_validate,df=10,lower.tail = FALSE)
print(MahaPvalue.v)

#filter p-value<0.001 (check for outliers)
#21 observations (4.2% of total sample)
print(sum(MahaPvalue.v<0.001))

#Add the mahalanobis distance and its p-value to the data
LoanMaha.v<-cbind(loandata_sample_validate, maha_validate, MahaPvalue.v)
LoanMaha.v %>% filter(MahaPvalue.v<0.001)

#Remove outliers
LoanMaha_validate <- filter(LoanMaha.v, MahaPvalue.v>=0.001)
```

#External validation-Correlation test
```{r}
#Remove maha and MahaPvalue columns from dataset
LoanMaha_new_validate<-select(LoanMaha_validate,-maha_validate,-MahaPvalue.v)

#Check correlation between variables
#None >0.8
LoandataMatrix.v <- cor(LoanMaha_new_validate)
round(LoandataMatrix.v,2)
```

#External validation-Standardize data
```{r}
LoanMaha_scaled_validate <-scale(LoanMaha_new_validate)
```

#External validation-Perform cluster analysis
```{r}
#Determine optimal number of clusters
#Calculate gap statistic
gap_stat_h_v <- clusGap(LoanMaha_scaled_validate, FUN = hcut, nstart = 25, K.max = 7, B = 50)

#Produce the plot
#Optimal 3+ --choose 3 clusters
#Validation result: same no. of optimal clusters!
fviz_gap_stat(gap_stat_h_v)

#Non-hierarchical clustering
#23 198 258
set.seed(240)
k_cl.v <- kmeans(LoanMaha_scaled_validate,3,nstart=25)
k_cl.v

clustering_vector.v <- k_cl.v$cluster
print(clustering_vector.v)
```
```

Appendix 6. R codes