WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	5568440, 2027360, 5554542, 5591743, 55865 5503558
Module Code:	IB98D0
Module Title:	Advanced Data Analytics
Submission Deadline:	Monday 18th March at 12:00:00
Date Submitted:	Monday 18th March 2024
Word Count:	1994
Number of Pages:	26
Question Attempted: *(question number/title, or description of assignment)*	1
Have you used Artificial Intelligence (AI) in any part of this assignment?	No

Academic Integrity Declaration

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that this work is being submitted on behalf of my group and is all our own, , except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

Table of Content

Executive Summary

This report focuses on differentiating between different borrower segments, analysing customer behaviour, and optimising strategies and products. The methodology focuses on the application of cluster analysis to critically analyse the provided dataset of 50,000 observations and 53 variables. The data were pre-processed and screened for 10 relevant variables and randomly sampled 500 observations for the study. As there is no multicollinearity, the factor analysis was conducted. Six factors are identified, and the eight most representative variables are filtered out. Then, this report utilises hierarchical clustering to obtain potential optimal number clusters. Based on that, the results were generated by K-means clustering. External validation confirms the effectiveness of the identified clusters, ensuring the robustness of the analysis and recommendations.

Introduction

Recently, lending companies have encountered challenges in efficiently approving loans, assessing risks associated with borrowers, and ensuring customer satisfaction. To overcome these hurdles, the application of principal component analysis (PCA), factor analysis (FA), and cluster analysis (CA) present an opportunity to gain deeper insights into customer behaviours, segmentation, and risk identification analysis.

This research aims to categorize customers into distinct clusters based on their behaviours and risk profiles. Therefore, lending companies can develop personalized loan products and tailored marketing strategies, enhancing customer engagement and satisfaction. Additionally, this analysis aids in refining decision-making processes and risk management strategies, proactively mitigating potential credit defaults to optimize lending practices.

Data Preparation and Variables Selection

To ensure the data quality and suitability of the dataset for the cluster analysis, the process involved several crucial steps including removing missing values, normalising data, and investigating multicollinearity.

Initially, seven variables with many missing values were identified and removed from the dataset. This step was taken to avoid potential biases in the CA that could arise from utilizing variables with substantial missing data.

Considering the logical relationship between the variables and our targets, we select the top 10 variables to do analysis (see Appendix A and G).

The next step is transforming three ordinal categorical variables - including 'grade', 'home_ownership', and 'loan_status' - to numeric data. There are 7 categories for ‘grade’, 3 categories for 'home_ownership', and 4 categories for 'loan_status'. All ‘1’ represents the best from perspective of lending loan (see Appendix A).

After converting categorical data, multicollinearity was assessed using correlation analysis. However, through investigation, the selected variables show no significant multicollinearity (see Appendix B and H). Therefore, we decided not to perform PCA and jump right to FA.

Factor Analysis

To understand the idea of factor analysis, first check the correlations between observed variables in terms of factors which helps in to remove variable home_ownership, delinq_2yrs for the data redundancy.

To initiate Factor Analysis, we opted a sample of 500 observation and scaled the data for standardization. Further, we removed 13 outliers based on Mahalanobis distance. This helps us to calculate for the correlation matrix to identify variable relationships.

Before Factor Analysis, there must be a strong foundation supporting the assumption of underlying structure by checking multicollinearity. There was no multicollinearity. (Refer Appendix C and I)

We executed the factor analysis to find the suitable factors for your cluster analysis using different methods of analysis.

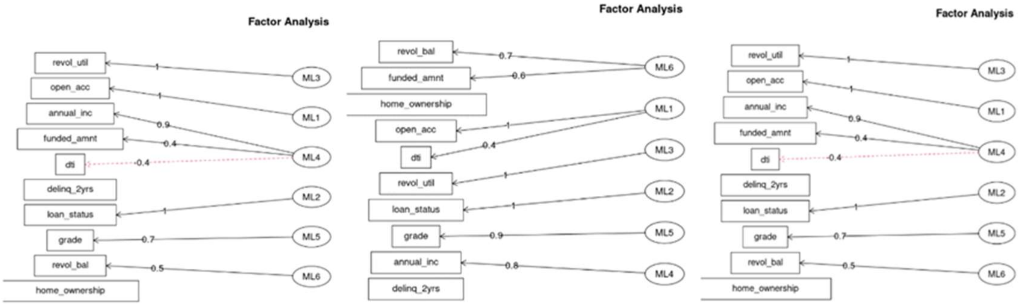


Fig 1: From left, Factor analysis with no rotation reveals negative correlations, varimax rotation demonstrates no correlations, while oblimin rotation also indicates negative correlation.

This result shows that Varimax captures maximum variables and shows less cross-loading compared to the ML extraction and Oblimin rotation. (Refer Appendix J)

Clustering Analysis

The main assumption of clustering analysis is to divide groups into different risk levels when improving loan portfolio management, risk strategy, and marketing strategy. Then the company can provide customized loan products and differentiated customer services based on the different needs.

We compared CA1 with factors and CA2 using eight variables directly with the different optimal number of clusters. Janrao, Mishra, and Bharadi (2019) indicates the value of between_SS/total_SS can be used to preferentially evaluate the quality of clustering, as it indicates better explanatory power and a more pronounced separation of clusters. According to Table 1, the CA1's ratio signifies a lower proportion of variance explained, implying potentially weaker clustering performance. In contrast, the CA2's ratio indicates that nearly half of the total variability in the data is explained by clustering, suggesting a substantial impact on the data's structure and a clearer separation between clusters. Therefore, the following focus is a detailed and critical analysis of the CA2.

Table 1 The comparison of the performance of two different clustering analysis

Type	Linkage method	The optimal number of clusters	Between_SS / Total_SS
CA1	ward	5	41.2 %
CA2	ward	6	45.8 %

Create Clusters using hierarchical and K-Means

Instead of applying just one technique, we innovatively utilize hierarchical clustering to get the results of distance matrix between borrowers, the potential optimal number of clusters and corresponding attributes of each cluster, then we generate the results by K-means clustering, to adopt the most advantages of the two techniques.

Firstly, comparing with other agglomerative methods, Ward's minimum variance method can make the highest agglomerative coefficient, which can contribute most for the following analysis (see Table 2).

Table 2 The agglomerative coefficient of the different agglomerative methods

Agglomerative methods	average	single	complete	ward
Coefficient	0.7899684	0.6231242	0.8617598	0.9616273

Determining the optimal cluster number, shorter lines in Figure 3 suggest greater similarity among clusters. While fewer clusters simplify interpretation and Figure 2 favors six for higher gap statistics, Figure 3's line 1 points to five for better performance. However, considering fit values and the between_SS/total_SS ratio, six clusters provide a more balanced Cluster Analysis with broader coverage (refer to Appendices C, D and K). Thus, six clusters were chosen.

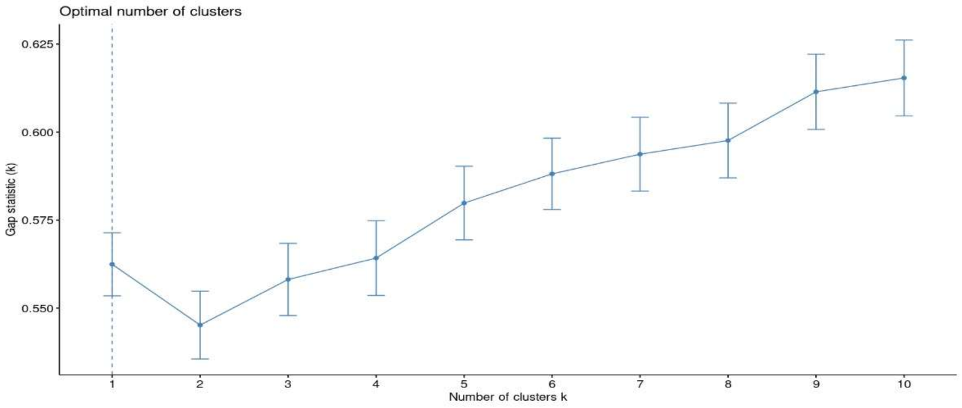


Figure 2 The optimal number of clusters with gap statistics method

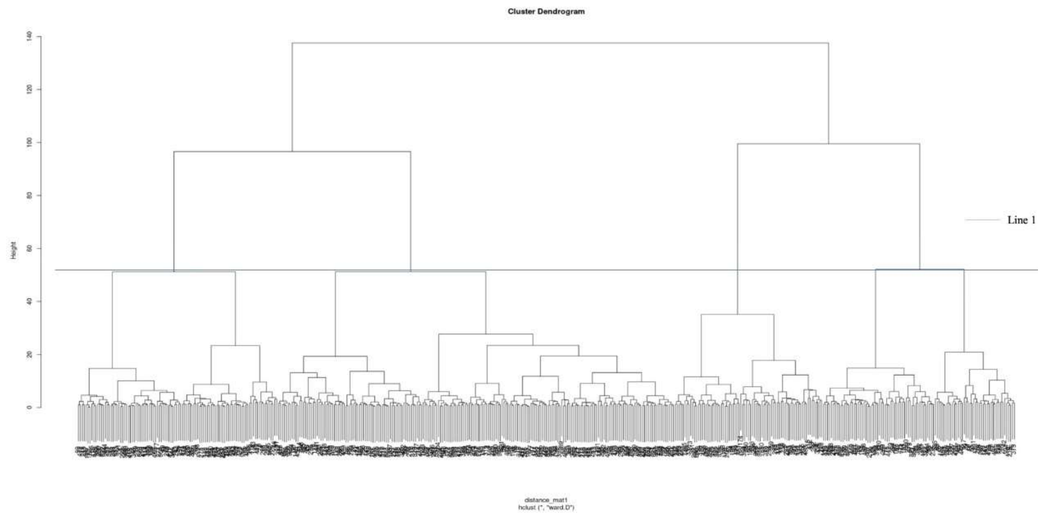
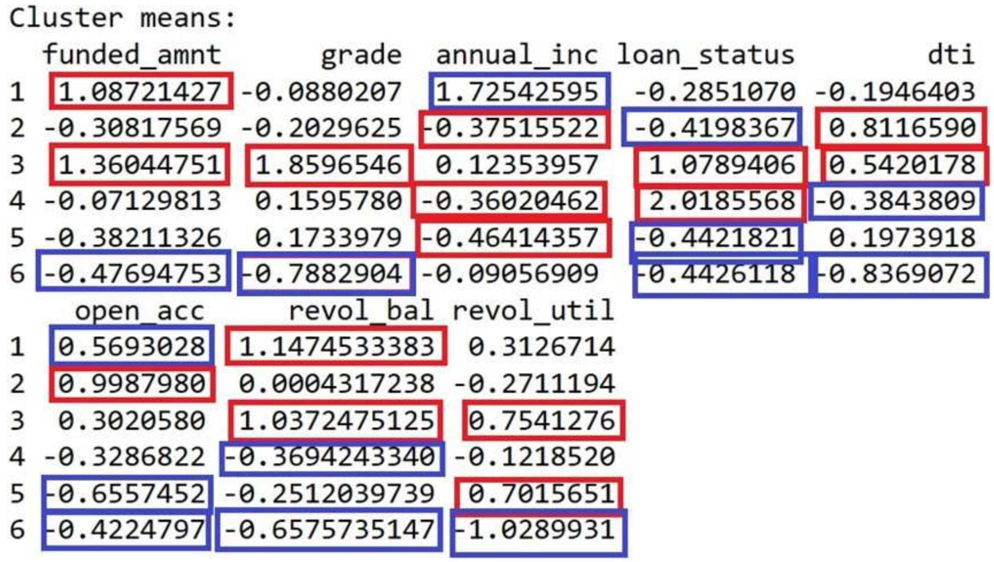


Figure 3 Cluster Dendrogram Visualization Using Hierarchical Clustering

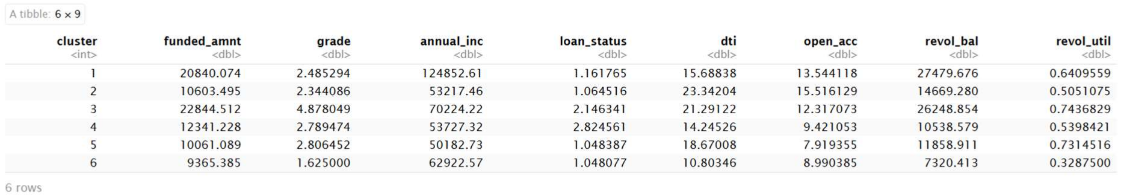
Interpretation and Recommendation

In analysing the clusters, we focused on variables showing notable deviations from the mean (scaled to '0') positive or negative in Figure 4. Figure 5 serves as a complementary which explains the pre-scaling mean data for each cluster.



```
Cluster means:
  funded_amnt      grade  annual_inc loan_status        dti
1  1.08721427 -0.0880207  1.72542595  -0.2851070 -0.1946403
2 -0.30817569 -0.2029625 -0.37515522  -0.4198367  0.8116590
3  1.36044751  1.8596546  0.12353957   1.0789406  0.5420178
4 -0.07129813  0.1595780 -0.36020462   2.0185568 -0.3843809
5 -0.38211326  0.1733979 -0.46414357  -0.4421821  0.1973918
6 -0.47694753 -0.7882904 -0.09056909  -0.4426118 -0.8369072
   open_acc     revol_bal revol_util
1  0.5693028  1.1474533383  0.3126714
2  0.9987980  0.0004317238 -0.2711194
3  0.3020580  1.0372475125  0.7541276
4 -0.3286822 -0.3694243340 -0.1218520
5 -0.6557452 -0.2512039739  0.7015651
6 -0.4224797 -0.6575735147 -1.0289931
```

Figure 4 Results of K-Means



A tibble: 6 × 9

cluster	funded_amnt	grade	annual_inc	loan_status	dti	open_acc	revol_bal	revol_util
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	20840.074	2.485294	124852.61	1.161765	15.68838	13.544118	27479.676	0.6409559
2	10603.495	2.344086	53217.46	1.064516	23.34204	15.516129	14669.280	0.5051075
3	22844.512	4.878049	70224.22	2.146341	21.29122	12.317073	26248.854	0.7436829
4	12341.228	2.789474	53727.32	2.824561	14.24526	9.421053	10538.579	0.5398421
5	10061.089	2.806452	50182.73	1.048387	18.67008	7.919355	11858.911	0.7314516
6	9365.385	1.625000	62922.57	1.048077	10.80346	8.990385	7320.413	0.3287500

6 rows

Figure 5 Denormalized Cluster Mean value for selected variables

Cluster 1 Medium Risk and High Income Borrowers

They are individuals with higher-than-average incomes who exhibit certain characteristics indicating medium risk can be categorised as "Affluent Borrowers". They demand larger loan amounts and maintain numerous open credit lines. Despite their high credit balances, their mean annual income significantly outweighs their total revolving balance. While a higher income may enable them to qualify for larger loans, it also implies a higher debt capacity. However, larger loan amounts also entail higher repayment obligations, increasing the risk of default if individuals experience financial setbacks or unexpected expenses as they already have a lot of open credit lines with high balances due.

Recommendation

1. **Personalised Instalments –** Offers personalised loan packages with higher loan limits but suggest custom repayment options or interest align with their higher income and utility.
2. **Relationship Marketing** – Foster long term relationship as they are high income individuals and can bring high returns to lenders by assigning dedicated relationship manager, seminars and providing VIP perks *(Raghunathan, 2023)***.**

Cluster 2: Medium Risk and Low Income Borrowers

These customers are potentially at a medium risk level. They have the highest average dti among all clusters of customers, suggesting a high level of debt relative to their income. Also, they have more opening lines, indicating more usage of credits, which is aligned with the findings from high average dti. However, the loan_status values are more negative, suggesting most of their loans are on the side of paid off. Despite their generally consistent loan repayment history, the relatively lower income levels could suggest a higher risk due to their substantial debt.

Recommendations:

1. **Consistent Monitoring:** To identify customers within this cluster who are moving towards higher risk profiles, it is recommended to employ consistent monitoring strategies, including regular reviews of credit limits and financial health checks, alongside dynamic risk assessments.

2. **"Pay Later" with low limits:** "Pay Later" can be appealing to them as they might need to make a purchase but do not have the immediate funds available *(Kempson & Philps ,2022)*.

Cluster 3 High Risk and Non- Performing Borrowers

This comprises of individuals exhibiting high-risk characteristics, particularly concerning loan repayment history, dti, and credit utilization, and can be categorised as "Risky Borrowers". These borrowers typically request higher loan amounts, primarily for high-risk grade loans, but demonstrate poor repayment behaviour. Moreover, their high dti indicates that they spent a large portion of their monthly income on repaying their debts, which would negatively impact their ability to qualify for credit. Additonally, they carries larger revolving balances and use a substantial amount of their available credit, which implies they heavily depend on credit and may be experiencing financial stress, hence they should be avoided.

Recommendation

1. **Stricter Approval Criteria** – When lending we should perform credit checks, assess repayment history, and scrutinise dti based on which imposing higher credit score thresholds to mitigate risk.
2. **Target Risk mitigation Products** – Market products guarantor loans which provide additional security for lenders while enabling borrowers to access financing options despite their credit challenges *(Kempson & Philps ,2022)*.

Cluster 4 High-to-Medium Risk and Low-Income Borrowers

They consist of individuals with comparatively low annual incomes, and history of non-repayment indicated by instances late payments, charge-offs, or non-payment. Despite their low income, they have a favourable dti ratio suggesting monthly debt payment compared to monthly income categorising them as "Low-Earning Debt Managers". This suggests an enhanced ability to repay their debts over time. Additionally, they maintain a low revolving balance, indicating a conservative approach to credit usage and a reduced reliance on credit lines.

Recommendations

1. **Asset Purchase Loans –** Suggest an asset purchase loan to this group. If this type of people fails to repay the loan, the lender has the right to seize the asset to recover their funds *(Kempson & Philps ,2022).*
2. **Small Instalment credit cards**– Offer opportunities like secured credit cards, small instalment loans to improve their credit history and gradually enhance their creditworthiness.

Cluster 5: Medium-to-Low Risk and Low-Income Borrowers

Customers with notably lowest income levels and a moderate to low risk profile. Their positive but relatively low dti implies that their debt may be a larger proportion of their income than the lowest risk groups. The relatively high credit utilization further underscores a substantial dependence on credit resources, which might suggest behaviors such as relying on credit for day-to-day expenses due to limited liquidity. However, their negative average loan_status implies good payback history and responsible profiles.

Recommendation

1. **"Pay Later" with higher limits:** short-term products like "Pay Later" with higher limits or Bridging Loans, due to good dti, can be introduced to them more, which meets their financial needs *(Kempson & Philps ,2022).*
2. **Supportive Resources:** Resources and tools for budget management and financial planning can be offered to assist customers from this cluster to enhance their financial standing and responsible profiles over time.

Cluster 6: Low Risk, Conservative Borrowers

Customers who exhibit financial stability and conservative borrowing behaviour. They have lower average dti, suggesting a manageable level of debt relative to their income, which signals a lower credit risk. Also, they tend to take smaller and less risky loans, as indicated by negative funded_amnt and grade. Moreover, they are more conservative with their credit, as indicated by fewer open credit, carry less debt, and smaller percentage of usage of their available credit limit. Additionally, the loan_status values in this cluster are more negative, indicating a high number of loans being current or fully paid off.

Recommendations:

1. **Better Loan Terms:** From a lending perspective, these customers can be considered low risk and reliable borrowers, which could qualify them for more favourable loan terms.
2. **Engagement Marketing:** The company should actively engage with them through emails and calls, delve into their borrowing motivations and preference, and ultimately provide good loan offers tailored to their needs and profiles *(Aite, 2022)*.

External Validation

In this part, we did a validation method for ensuring the robustness the quality of the clustering. This is important because the role of validation method is measuring the quality of the clusters formed by a cluster algorithm. Internal validation will be preferable when we have not yet decided the optimal clustering. However, external validation is better to fully capture the full complexity of patterns that might be present in the data (*Zerabi & Meshoul, 2017).*

Table 3 The comparison of the customer profiles resulted from final clustering analysis and external validation.

Profiling	Medium Risk, High Income	Medium Risk, Low Income	High Risk, Non-performing	High-to-Medium Risk, Low-Income	Medium-to-Low Risk, Low-Income	Low Risk, Conservative
Final Cluster Analysis	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
% Cluster	14%	19%	8%	12%	26%	21%
External validation	Cluster 5	Cluster 3	Cluster 6	Cluster 1	Cluster 2	Cluster 4
% Cluster	13%	22%	12%	13%	23%	17%
Diff	1%	(3%)	(4%)	(1%)	3%	4%

Our external validation results show very similar clusters can be generated from the external data using the same clustering algorithm. (Refer Appendix L) Firstly, the SS ratio, indicating total variability captured by the clustering algorithm, of the external validation (45.50%) closely aligns with that of the final clustering analysis (45.80%). Secondly, the results shown in Table 3 illustrate that we can create similar profiling clusters with a comparable proportion of customers categorised into each cluster. For example, Cluster 1 from the final cluster analysis—comprising customers with medium risk and high income—shows a similar profile to Cluster 5 from external validation, with only a 1% difference in the proportion of customers classified in this cluster, 14% versus 13%.

Conclusion

In conclusion, this report applied hierarchical and non-hierarchical cluster analysis on loan data to identify 6 distinguish segments, where the corresponding recommendations are provided to lending company, which will contribute to improving the risk strategy, customized products and operating strategy.

References

Janrao, P., Mishra, D. and Bharadi, V. (2019) “Performance evaluation of principal component analysis for clustering on sugarcane dataset,” in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*. IEEE.

S. Zerabi and S. Meshoul   (2017), "External clustering validation in big data context," in   3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat, Morocco, 2017, pp. 1-6, doi: 10.1109/CloudTech.2017.8284735.

Joel Kempson and Rhiannon Philps (2022) [online] “19 Types of Loans: Which Should You Choose?” in https://www.nerdwallet.com/uk/loans/personal-loans/types-of-loan-faq/, Secured Loan, Guarantor Loan, Pay later loan, Bidging Loan

Cassy Aite (2022), [online] “10 Effective engagement marketing tactics to boost customer engagement” in https://www.hoppier.com/blog/engagement-marketing, What is engagement marketing

Sunitha Raghunathan (2023), [online] “What is relationship marketing: examples and strategies” in https://birdeye.com/blog/relationship-marketing/, Different types of relationship marketing

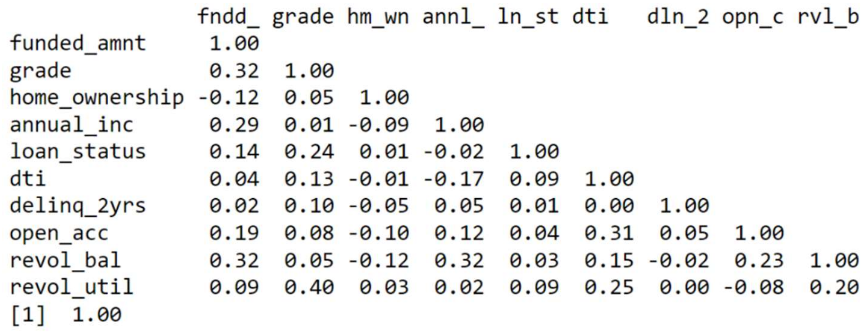
Appendix

Appendix A: Description of Variables Selected for Culstering Analysis

Variable	Level	Description
funded_amnt	num	The amount of money the bank commits to lend to a lender at a given point in time, which is related with loan amount.
grade	chr	LC assigned loan grade. We encoded A as 1, which represents the least risky loan, G represents the riskiest (A-1, B-2, C-3, D-4, E-5, F-6, G-7).
home_ownership	chr	Borrower's home ownership at the time of registration, which is related with repayment ability and loan purpose. We encoded based on lenders view stating the 'own' status as 1, 'mortgage' as 2, 'rent' as 3 (other and none category are encoded as 1 because of clarity and less number of observation).
annual_inc	num	A record of the borrower's combined annual income, which is related with repayment ability and loan amount.
loan_status	chr	The loan's repayment situation, which is related with the borrower's credit rating. We encoded 'fully paid' status to 1, 'current' to 2, 'charged of' to 3, 'default' or 'In Grace Period', 'Late (16-30 days)' , 'Late (31-120 days)' to 4.
dti	num	A ratio about the borrower's total monthly debt payments divided by the total self monthly income (excluding mortgages and required letter of credit loans) divided by the borrower's self-reported monthly income, which is related with repayment type and repayment ability.
delinq_2yrs	num	The number of times a borrower's credit file is more than 30 days past due, within the last 2 years, which is related with

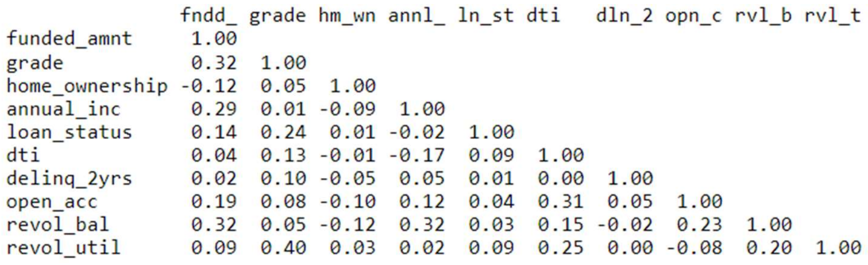
		repayment ability and credit rating.
open_acc_6m	num	Number of open trades in the last 6 months, which is related with repayment ability.
revol_bal	num	The sum of the borrower's revolving credit balance, which is related with loan amount.
revol_util	num	A ratio regarding the total current balance with a high credit/credit limit in all revolving accounts.

Appendix B: Correlation Table for the Chosen Variables



```
               fndd_ grade hm_wn annl_ ln_st dti   dln_2 opn_c rvl_b
funded_amnt     1.00
grade           0.32  1.00
home_ownership -0.12  0.05  1.00
annual_inc      0.29  0.01 -0.09  1.00
loan_status     0.14  0.24  0.01 -0.02  1.00
dti             0.04  0.13 -0.01 -0.17  0.09  1.00
delinq_2yrs     0.02  0.10 -0.05  0.05  0.01  0.00  1.00
open_acc        0.19  0.08 -0.10  0.12  0.04  0.31  0.05  1.00
revol_bal       0.32  0.05 -0.12  0.32  0.03  0.15 -0.02  0.23  1.00
revol_util      0.09  0.40  0.03  0.02  0.09  0.25  0.00 -0.08  0.20
[1]  1.00
```

Appendix C: Test for Multicollinearity before Factor Analysis



```
               fndd_ grade hm_wn annl_ ln_st dti   dln_2 opn_c rvl_b rvl_t
funded_amnt     1.00
grade           0.32  1.00
home_ownership -0.12  0.05  1.00
annual_inc      0.29  0.01 -0.09  1.00
loan_status     0.14  0.24  0.01 -0.02  1.00
dti             0.04  0.13 -0.01 -0.17  0.09  1.00
delinq_2yrs     0.02  0.10 -0.05  0.05  0.01  0.00  1.00
open_acc        0.19  0.08 -0.10  0.12  0.04  0.31  0.05  1.00
revol_bal       0.32  0.05 -0.12  0.32  0.03  0.15 -0.02  0.23  1.00
revol_util      0.09  0.40  0.03  0.02  0.09  0.25  0.00 -0.08  0.20  1.00
```

Fig: This shows that there is no correlation between all the variables

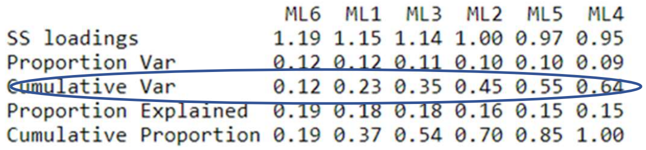
We also performed the Kaiser-Meyer-Olkin (MSA) > 0.5 and Barlett's test (p-value < 0.05) before factor analysis.

Kaiser-Meyer-Olkin Overall MSA	0.54

Bartlett's Test of Sphericity	Approx. Chi-Square	857.5433
P value		1.058988e-150
df		45

Table: Shows overall MSA is more than 0.50 and p value <0.05 which shows that data is suitable for the factor analysis.

For Factor analysis, it must find the suitable number of factors to perform the analysis to increase the stability of the model.



```
                        ML6  ML1  ML3  ML2  ML5  ML4
SS loadings            1.19 1.15 1.14 1.00 0.97 0.95
Proportion Var         0.12 0.12 0.11 0.10 0.10 0.09
Cumulative Var         0.12 0.23 0.35 0.45 0.55 0.64
Proportion Explained   0.19 0.18 0.18 0.16 0.15 0.15
Cumulative Proportion  0.19 0.37 0.54 0.70 0.85 1.00
```

Fig: Enough factors to meet a specified percentage of variance explained, usually 60% or higher.

Appendix D: The comparison of Fit Value 1 with Six Clusters and Fit Value 2 with Five Clusters

Cluster	1	2	3	4	5	6
Fit value 1	104	58	131	77	42	75
Fit value 2	104	58	206	77	42	

Appendix E: The comparison of the Ratio of Between_SS / Total_SS with Five Clusters and with Six Clusters

Type	Linkage method	The optimal number of clusters	Between_SS / Total_SS
CA2	ward	5	41.8 %
CA2	ward	6	45.8 %

Appendix F: Coding Setup

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

~~~
library(dplyr)
library(car)
```

# Master data
```{r}
loan_data <-read_excel("loan_data_ADA_assignment.xlsx")
```
~~~

Appendix G: Data Exploration and Data

~~~
## Looking at the Data, and Basic Analysis
```{r}
describe(loan_data)
```

desc, mths_since_last_delinq and mths _since_last_record, next_payment_d, mths_since_last_derog and totak_coll_amount has more than 10% values missing.

generally, loan offered is for 14% interest rate.
Approx whole amount lent is funded by investors.
grade to be analysed further..
generally people lent loan have mean employment length of 6 years
generally people lent loan have mean annual income of 71,316
generally people lent loan have mean revol utiliy of 59%

```{r}
summary(loan_data)
str(loan_data)
```

```{r}
headTail(loan_data)
```
# Data cleaning
## removing variable with large NA
```{r}
# Removing Column
columns_to_remove <- c("desc", "mths_since_last_delinq", "mths_since_last_record",
"next_pymnt_d", "mths_since_last_major_derog", "tot_coll_amt","tot_cur_bal")

loan_data <- loan_data[, !names(loan_data) %in% columns_to_remove]

# Print the modified dataframe
describe(loan_data)
~~~

~~~
```
~~~

Appendix H: Variable Selection

~~~
# Variable selection
```{r}
#Select Desired Variables
loan_data_revised <- loan_data %>%
  select("funded_amnt", "grade", "home_ownership", "annual_inc", "loan_status","dti",
"delinq_2yrs", "open_acc", "revol_bal","revol_util")

describe(loan_data_revised)

loan_data_revised<- na.omit(loan_data_revised)

str(loan_data_revised)

#Visualize 3 major variables
#DTI (debt toi income ratio)
ggplot(loan_data_revised, aes(x = dti)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Debt-to-Income Ratio (DTI)",
       x = "Debt-to-Income Ratio (DTI)",
       y = "Frequency") +
  theme_minimal()

#Annual income
ggplot(loan_data_revised, aes(x = annual_inc)) +
  geom_histogram(fill = "lightgreen", color = "black") +
  scale_x_log10() +
  labs(title = "Distribution of Annual Income",
       x = "Annual Income (Log Scale)",
       y = "Frequency") +
  theme_minimal()
#Grade
plot_grade <- ggplot(loan_data_revised, aes(x = grade)) +
  geom_bar(fill = "orange", color = "black") +
  labs(title = "Distribution of Loan Grades",
       x = "Loan Grade",
       y = "Count") +
  theme_minimal()

```
~~~

~~~
# Normalizing data (Convering chr variable into numerical by grading)
```{r}

loan_data_revised$grade <- as.numeric(factor(loan_data_revised$grade, levels = c("A", "B",
"C", "D", "E", "F", "G")))

str(loan_data_revised)

loan_data_revised$loan_status <- as.numeric(factor(loan_data_revised$loan_status, levels =
c("Fully Paid", "Current", "Charged Off", "Default", "In Grace Period", "Late (16-30 days)",
"Late (31-120 days)")))

loan_data_revised$loan_status <- ifelse(loan_data_revised$loan_status %in% 4:7, 4,
loan_data_revised$loan_status)

loan_data_revised$home_ownership <-
as.numeric(factor(loan_data_revised$home_ownership, levels = c("OWN", "MORTGAGE",
"RENT", "NONE", "OTHER")))

loan_data_revised$home_ownership <- ifelse(loan_data_revised$home_ownership %in% 4:5,
1, loan_data_revised$home_ownership)


str(loan_data_revised)
summary(loan_data_revised)

```
~~~

Appendix I: Multicollinearity Checking and Sampling

~~~
# Multicollinearity
We check the data if there are any multicollinearity in our dataset
## Check multicollinearity by correlation
```{r}

datacorr <- cor(loan_data_revised)

round(datacorr, 2)

lowerCor(loan_data_revised)

```


## sampling preparation before FA and clustering
```{r}
~~~

~~~~
#Sampling
set.seed(100)
representative_sample <- loan_data_revised[sample(nrow(loan_data_revised), 500, replace = FALSE),]

```


## Check Mahalanobis distance

```{r}
#FA Prep
Maha <- mahalanobis(representative_sample, colMeans(representative_sample, na.rm = TRUE), cov(representative_sample, use = "complete.obs"))
# Calculate p-values
MahaPvalue <- pchisq(Maha, df=10, lower.tail = FALSE)
# Identify outliers
outliers <- which(MahaPvalue < 0.001)
# Remove outliers
cleaned_sample <- representative_sample[-outliers, ]
```
~~~~

Based on the mahalanobis test we found that there are 13 outlier data that we removed in order to create more comperhensive analysis

Appendix J: Factor Analysis

~~~~
# Factor Analysis Preparation
## Standardise each variable with mean of 0 and sd of 1
```{r}

scaled_cleaned_sample <- scale(cleaned_sample)

headTail(scaled_cleaned_sample)

summary(scaled_cleaned_sample)
lowerCor(scaled_cleaned_sample)

```


## KMO anddd barlett Test
```{r}
# KMO test might require a matrix without NA values or non-numeric columns
KMO(scaled_cleaned_sample)
~~~~

~~~~
# Bartlett's test
cortest.bartlett(cleaned_sample, n=487)
```
for KMO, the overall MSA is 0.54 which is suitable for doing Factor Analysis. however, there
are several variables that have a lower score (dti,delinq_2yrs,open_acc,revol_until). in the
barlett test, the p value <0.05 which is also suitable for the factor analysis.


# Factor Analysis
## Factor Analysis Using ml method
```{r}
factor_analysis1 <- fa(scaled_cleaned_sample,6,n.obs = 487, fm = "ml")
print(factor_analysis1, cut = 0.3, sort = "TRUE")

scores_FA1 <- factor_analysis1$scores

```
```{r}
fa.diagram(factor_analysis1)
```
## Factor Analysis using Varimax rotation
```{r}
factor_analysis2 <- fa(scaled_cleaned_sample,6,n.obs = 487,rotate = "varimax", fm = "ml")
print(factor_analysis2, cut = 0.3, sort = "TRUE")
scores_FA2 <- factor_analysis2$scores

```



```{r}
fa.diagram(factor_analysis2)
```
## FA Analysis using Oblimin Rotation
```{r}
factor_analysis3 <- fa(scaled_cleaned_sample,6,n.obs = 500,rotate = "oblimin", fm = "ml")
print(factor_analysis3, cut = 0.3, sort = "TRUE")
scores_FA3 <- factor_analysis2$scores
```


```{r}
fa.diagram(factor_analysis3)
```
Based on the 3 FA maethod that we use, we decided to use Varimax rotation because this
method captures more variables than the other
~~~~

~~~~
# Finalize data for further analysis
we delete home ownership and Delinq_2yrs in ourr observation data because we found that those two variables are not corelated with other variables
## Delete "Home Ownership" and "Delinq_2yrs"
```{r}

final_sampling <- subset(scaled_cleaned_sample, select = -c(home_ownership, delinq_2yrs))

summary(final_sampling)
factor_analysis_final <- fa(final_sampling,6,n.obs = 487,rotate = "varimax", fm = "ml")
print(factor_analysis_final, cut = 0.3, sort = "TRUE")
scores_FA <- factor_analysis_final$scores


fa.diagram(factor_analysis_final)


```
## Fscores Comparison
```{r}
fscores <- factor_analysis_final$scores #varimax
describe(fscores)
headtail(fscores)
```

check assumptions to see whether the data are suitable for Cluster Analysis:

```{r}

FscoresMatrix<-cor(fscores) #varimax
print(FscoresMatrix)
```
```{r}
round(FscoresMatrix, 2)
```

```{r}
lowerCor(fscores)
```
~~~~

Appendix K: Cluster Analysis

~~~~
# Cluster Analysis
~~~~

we decided to do 2 cluster analysis, the first one is using FA and the second one is only using 8 variables that we already chose before to see which one create a better results

```{r}
#Final data without home ownershipn and delinq2years
# Define linkage methods
m <- c("average", "single", "complete", "ward")
names(m) <- c("average", "single", "complete", "ward")
```

## Clustering after FA
```{r}
ac <- function(x) {
  agnes(fscores, method = x)$ac
}
sapply(m, ac)
```

Ward's minimum variance method produces the highest agglomerative coefficient, thus we'll use that as the method for our final hierarchical clustering:

```{r}
gap_stat1 <- clusGap(fscores, FUN = hcut, nstart = 25, K.max = 10, B = 50)
fviz_gap_stat(gap_stat1)
```

### Finding distance matrix and performing hierarchical clustering
```{r}
distance_mat1 <- dist(fscores, method = 'euclidean')
Hierar_cl1 <- hclust(distance_mat1, method = "ward")
plot(Hierar_cl1)
```
we found that based on the dendogram and gap stat, we found that 5 cluster is the best one
### Choosing number of clusters
```{r}

fit1 <- cutree(Hierar_cl1, k = 5)
table(fit1)
```

### Appending cluster labels to the factor scores
```{r}
```

```
final_data1 <- cbind(fscores, cluster = fit1)
head(final_data1)
```

### Find mean values for each cluster
```{r}
hcentres1 <- aggregate(x = final_data1, by = list(cluster = fit1), FUN = "mean")
print(hcentres1)
```

### Kmeans clustering
```{r}
set.seed(100)
k_cl1 <- kmeans(fscores, 5, nstart = 25)
k_cl1
fviz_cluster(list(data = fscores, cluster = k_cl1$cluster), geom = "point", ellipse = TRUE, ellipse.type = "norm")
```

"Within cluster sum of squares by cluster" refers to the sum of squares of the distances between each point in a cluster and the centroid of that cluster. Essentially, it's a measure of how tightly grouped the points within each cluster are.

The numbers (408.6356, 481.8271, 485.6417) are the within-cluster sum of squares for each cluster. Lower values generally indicate better clustering because it suggests that the points within each cluster are closer to each other, which implies more compact and distinct clusters.

The value "(between_SS / total_SS = 41.2%)" is the ratio of between-cluster sum of squares to the total sum of squares. This ratio indicates the proportion of variance in the data that is explained by the clustering. A higher percentage generally indicates better separation between clusters.

## Cluster Analysis with 8 variables
```{r}
ac <- function(x) {
  agnes(final_sampling, method = x)$ac
}
sapply(m, ac)
```

Ward's minimum variance method produces the highest agglomerative coefficient, thus we'll use that as the method for our final hierarchical clustering:

```{r}
```

~~~
gap_stat2 <- clusGap(final_sampling, FUN = hcut, nstart = 25, K.max = 10, B = 50)
fviz_gap_stat(gap_stat2)
```
based on this score, we choose 6 clustering since it has the highest gap statistics

### Finding distance matrix and performing hierarchical clustering
```{r}
distance_mat2 <- dist(final_sampling, method = 'euclidean')
Hierar_cl2 <- hclust(distance_mat2, method = "ward")
plot(Hierar_cl2)
```
based on the dendogram and gap statistic we choose 6 as our cluster number
### Choosing number of clusters
```{r}

fit2 <- cutree(Hierar_cl2, k = 6)
table(fit2)
```
### Appending cluster labels to the factor scores
```{r}

final_data2 <- cbind(final_sampling, cluster = fit2)
head(final_data2)
```
### Find mean values for each cluster
```{r}
hcentres2 <- aggregate(x = final_data1, by = list(cluster = fit2), FUN = "mean")
print(hcentres2)
```
### Kmeans clustering
```{r}
set.seed(100)
k_cl2 <- kmeans(final_sampling, 6, nstart = 25)
k_cl2




fviz_cluster(list(data = final_sampling, cluster = k_cl2$cluster), geom = "point", ellipse =
TRUE, ellipse.type = "norm")
```
~~~

Appendix L: Validation

~~~
#Validation
~~~

~~~
##External Validation
```{r}
validation_sampling <- subset(loan_data_revised, select = -c(home_ownership, delinq_2yrs))
```

```{r}
set.seed(50)
ex_validation_sampling <- validation_sampling[sample(nrow(validation_sampling), 500, replace = FALSE),]
```

```{r}
ex_validation_sampling <- scale(ex_validation_sampling)
```



```{r}
Maha_validation <- mahalanobis(ex_validation_sampling, colMeans(ex_validation_sampling, na.rm = TRUE), cov(ex_validation_sampling, use = "complete.obs"))

# Calculate p-values
MahaPvalue_validation_ex <- pchisq(Maha_validation, df=10, lower.tail = FALSE)
# Identify outliers
outliers_validation_ex <- which(MahaPvalue_validation_ex < 0.001)
# Remove outliers
ex_validation_sampling <- ex_validation_sampling[-outliers_validation_ex, ]
str(ex_validation_sampling)
```

### Finding distance matrix and performing hierarchical clustering
```{r}
dist_matrix_ex <- dist(ex_validation_sampling)  # This computes the Euclidean distance matrix by default

# Perform hierarchical clustering with the updated method
Hierar_cl_ex <- hclust(dist_matrix_ex, method = "ward.D")
```
based on the dendogram and gap statistic we choose 6 as our cluster number
### Choosing number of clusters
```{r}

fit_ex <- cutree(Hierar_cl_ex, k = 6)
table(fit_ex)
```
~~~

~~~
### Cluster Analysis
#### Kmeans clustering
```{r}
library(factoextra)
set.seed(123)
k_cl_ex <- kmeans(ex_validation_sampling, 6, nstart = 25)
k_cl_ex

```
~~~

Appendix M: Minutes of Meeting

All Team members participated in all meeting listed below:

Student ID - 5568440, 2027360, 5554542, 5591743, 5586595, 5503558

Meeting 1 - 03-03-24, 11 to 1 pm

1. Reading through the question and summarizing the objectives and work to be done for the assignment and approach we want to go forward with which is "Clustering Analysis".
2. We read through the data file and data dictionary for understanding of variables.
3. We highlighted some important variables and divided the team with some people doing the research of methodology and some for coding part.

Meeting 2 - 06-03-24, 3 to 6 pm

1. We discussed the final 10 variables selection and the reasoning for the same for noting in report.
2. We discussed the coding part needed for data preparation.
3. We discussed the structure of the report.

Meeting 3 - 13-03-24, 3 to 6 pm

1. We discussed the clustering analysis and using factor analysis or original variables.
2. We discussed the best method for clustering.

Meeting 4 - 15-03-24, 5 to 7 pm

1. We read through the part 1 of report other than interpretation.
2. We discussed the interpretation and how to profile and recommendation for each profile.
3. We discussed the validation process we followed external validation.

Meeting 5 - 17-03-24, 11 to 4 pm

1. Final Draft read of the report.
2. Making changes and adding few details.

Appendix L: Members' Contribution

2027360: Took part in interpreting cluster analysis result, wrote for half of the interpretation and recommendation, wrote for introduction, and contributed to the review of the report and codes and report refinement.

5568440: Took part in interpreting cluster analysis result, wrote for half of the interpretation and recommendation, Chunks of coding related to missing values, encoding and clsutering, and contributed to the review of the report and codes and report refinement.

5503558: Took part in writing report on Introduction, Data preparation, External validation. Coding contributed to checking multicollinearity and outlier checking and contributed to the review of the report and codes and report refinement.

5591743: Took part in taking charge of executive summary, variable selection, clustering analysis and conclusion part of report and contributed to the review of the report and codes and report refinement.

5554542: Took part in taking charge of factor analysis critical evaluation, discussion of methods and contributed to the review of the report and codes and report refinement.

5586595: Took part in coding and refinement of the part after data cleaning from subset creation till validation and contributed to the review of the report and codes and report refinement.