## 1. Executive summary

This project aims to enhance loan portfolio management by understanding customer behaviour and identifying distinct borrower segments through cluster analysis. Five main steps were employed: data preparation, Principal Component Analysis (PCA), Factor Analysis (FA), Cluster Analysis, and internal validation. Our analysis began with the preparation of loan data from 2012-2013. We randomly sampled 500 cases before narrowing down key variables, removing outliers, and standardising data. PCA and FA can eliminate multicollinearity and cross-loading issues, leading to optimal models. Our findings indicate that the optimal model involves 9-PC extraction with oblique rotation for cluster analysis. Additionally, clustering revealed three distinct borrower segments. Importantly, high-accuracy internal validation results confirmed our model's suitability for improving loan portfolio management.

## 2. Introduction

Cluster analysis is a way to combine groups of things so that things in the same group are more alike than those in other groups. (Rodriguez, Comin et al. 2019). In the report, we use cluster analysis on many loan data to learn more about how customers behave, find different groups of borrowers, and improve loan offerings. The main objective is to enhance loan portfolio management by putting customers into groups with similar traits. This makes it possible to create personalised loan products, more effective marketing plans, and better ways to help customers, all of which meet each group's specific wants.

## 3. Data Preparation

The dataset, consisting of loans issued during 2012-2013, illustrates a comprehensive view of various loan characteristics, such as current loan status and loan amount. The selection of variables for our analysis is based on their relevance to the loan performance and reliability. Hence, we strategically selected 12 critical variables from the initial 53 variables according to common sense, as detailed in Table 3.1. This selection aims to capture representative variables to offer deep insights into borrowers' creditworthiness and loan performance, allowing for a targeted and informed analysis. After selecting variables, we converted them into a suitable type for the following analysis.

**Table 3.1 Data Dictionary**

| | |
|---|---|
| **loan_amnt** | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| **sub_grade** | LC assigned loan subgrade |
| **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. |
| **annual_inc** | The self-reported annual income provided by the borrower during registration. |
| **loan_status** | Current status of the loan |
| **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. |
| **delinq_2yrs** | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| **open_acc** | The number of open credit lines in the borrower's credit file. |
| **revol_util** | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| **total_pymnt** | Payments received to date for total amount funded |
| **tot_coll_amt** | Total collection amounts ever owed |
| **tot_cur_bal** | Total current balance of all accounts |

3.1 Removing Missing Value

In our case, removing missing values (NAs) rather than imputing them is made considering the sufficient sample size and the aim to work with real, unaltered data. This approach is appropriate as we are only supposed to focus on 500 samples from the raw data. Even if we remove all missing values, we still have more than 500 observations.

3.2 Checking Distribution and Outliers Removal

Histogram analysis examines variable distributions after sampling the data. The visualisations showed a range of distribution patterns, from normal to skew. Variables with non-continuous values are categorical. This initial examination determines the suitability of statistical

methods and data transformation. They also suggest outliers in the data, which may affect cluster analysis (Appendix: Table A1).

We removed these outliers using univariate and multivariate approaches to ensure the quality of our analysis. To find outliers, the univariate technique calculates Z-scores for each observation. As we have a big sample dataset, the z-score threshold is 4 to minimise the chance of mistaking a natural deviation for an anomaly (Nwodo Benita Chikodili et al., 2021). Based on z-score, 18 observations were identified as outliers. Meanwhile, the Mahalanobis distance approach detects multivariate outliers. This technique accounts for variables' covariance, allowing for advanced detection that considers multidimensional data. Li et al. (2019) define outliers as distances greater than four times the degree of freedom and p-values less than 0.001. Our distance cut-off and p-value identified 7 and 16 significant multivariate outliers that could bias the cluster analysis, respectively. Table A2 in appendix shows the z-score and Mahalanobis value. After identification, we removed these outliers.

Further, scaling with standardisation prevented more extensive variables from dominating grouping. It provides all variables with a mean of zero and a standard deviation of one, allowing comparisons and equal contribution to the study.

## 4. Conducting Principal Component Analysis and Factor Analysis

### 4.1 Data Verification

After data preparation, we proceeded further to analyse its multicollinearity. We examined pairwise correlation, Kaiser-Meyer-Olkin (KMO), and Bartlett's test to decide whether to use PCA. Firstly, generating the correlation matrix shows multiple pairwise correlations above 0.8, indicating strong correlations among variables. The number of correlations exceeded 0.3, indicating significant variable relationships. Subsequently, the KMO measure evaluates all variable correlations. With an overall KMO value of 0.51, exceeding 0.5, our dataset is suitable for PCA. However, some variables have KMO values below 0.5; therefore, they may not correlate with other variables. This will be addressed in PCA.

Additionally, Bartlett's test determines if the correlation matrix has strong variable correlations (Hair et al., 2019). With a p-value $< 0.05$, PCA may be effective in treating multicollinearity. Thus, all three methods show that the dataset has multicollinearity, requiring management to plan for cluster analysis.

4.2 Principal Component Analysis

PCA, an interdependent multivariate statistical method, creates new variables. We consider SS-loadings (eigenvalues), a scree plot, and cumulative variance to determine the correct number of PCs. However, we want to reduce multicollinearity without information loss. For this, we seek to maintain most PCs with a cumulative variance greater than 0.9, which suggests 9 PCs. Therefore, it can provide most of the information for further research.

Initially, PCA without rotation was performed. In Table 4.3, 9 cross-loading issues were found, with factor loadings exceeding the minimum permissible limit of ±0.40. Higher numbers might suggest that the variable represents many factors. Thus, factor analysis with rotation was needed to solve this problem and simplify interpretation. Table A4 in appendix shows that variables with KMO values less than 0.5 correlate; hence, no variables should be removed.

**Table 4.3 The Results of FA Implementation**

| Models | No. of factor | No. of cross-loading | Cross-loading problem* |
|---|---|---|---|
| PCA without rotation | 9 | 9 | Most variables are explained by many PCs, except for total_pymnt, loan_amnt, and revol_util. |
| PC extraction with Oblique rotation | 9 | 1 | Factor 2 and 9 captured revol_util |
| PC extraction with Orthogonal rotation | 9 | 2 | Factor 1, and 4 captured annual_income.<br>Factor 2, and 9 captured revol_util. |

* Hair et al. (2019) mentioned that "Factor loadings in the range of ±0.30 to ±0.40 are considered to meet the minimal level for interpretation of the structure."

4.3 Factor Analysis – PC extraction with rotation

As each variable should have at most one factor, FA aims to minimise cross-loading and simplify interpretation. For this, we tried oblique and orthogonal rotation to analyse PC extraction. Table 4.3 shows rotation results. With only one cross-loading, PC extraction with oblique rotation, which is more flexible than orthogonal rotation, yields the best results. Hence, this model is suitable for further investigation.

## 5. Performing Cluster Analysis

Cluster analysis was used to group clients by credit risk, borrowing behaviours, and financial condition, including loan amount, annual income, loan status, etc.

Basic clustering in financial data analysis often uses hierarchical and K-means methods (Ezugwu et al., 2022). We also used both methods to obtain ideal cluster values, where 'Manhattan' distance was used for our project in hierarchical clustering. Distance is used to gravitate each observation to the nearest mean.

Different linkage approaches were tested to determine cluster strength. The results were derived using an agglomerative coefficient, with a good fit assumed to be closer to 1. After analysis, the ward's technique yielded the greatest value (Appendix: Table A7), which will be used for clustering. Meanwhile, the gap statistic, which compares the total intra-cluster variation for different values of k with their expected values for distribution without clustering, gave the optimal number of clusters as 3 using k-means (Appendix: Figure A1). Then, we proceed to hierarchical clustering, as combining both methods allows items to move between clusters and yields better results.

Following the hierarchical grouping, the data was effectively divided into three distinct categories. The 375, 50, and 52 cluster sizes account for 78.62%, 10.48%, and 10.90% of 477 data points, respectively. This hierarchical clustering with 3 clusters proved to be the most effective for this dataset, with well-separated and connected clusters, ensuring the best grouping possible.

Figure 5.1 below shows that Cluster 1 has a low value on TC8 and TC5, which affects loan status and delinquency. A low loan status standardised value indicates full payment. Cluster 1 has highly committed loanees with minimal risk due to a fully paid loan and few delinquencies. Cluster 2 has low TC8, high TC7 and TC5 values for loan status, employment term, and delinquency. The borrower has consistent employment and currently pays the loan; however, they have historical past due payments in the past two years, indicating moderate risk. Cluster 3 has high TC8 value, suggesting the loanee is charged off or lost; low TC1 and high TC9 scores imply low-grade loanees with low total payments, representing high-risk borrowers. The variables related to each TC can be found in Table A6 in appendix.

| cluster | TC1 | TC2 | TC9 | TC4 | TC3 | TC8 | TC7 | TC5 | TC6 |
|---|---|---|---|---|---|---|---|---|---|
| <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 0.0244516 | 0.033942940 | -0.08075009 | -0.05989392 | -0.003856792 | -0.3302450 | -0.05866702 | -0.31997640 | 0.03177184 |
| 2 | 0.2511834 | -0.006948996 | 0.19622873 | 0.22200010 | 0.131433514 | -0.3948257 | 0.31459250 | 2.44494417 | -0.11958986 |
| 3 | -0.4178561 | -0.238099092 | 0.39365094 | 0.21846568 | -0.098564974 | 2.7612149 | 0.12058666 | -0.04338576 | -0.11413362 |

Figure 5.1 The Result of Cluster Analysis.

## 6. Validation

We performed repeated clustering analysis on a lower random dataset as our first internal validation to verify model accuracy. This can be seen in the figure 6.1 below.

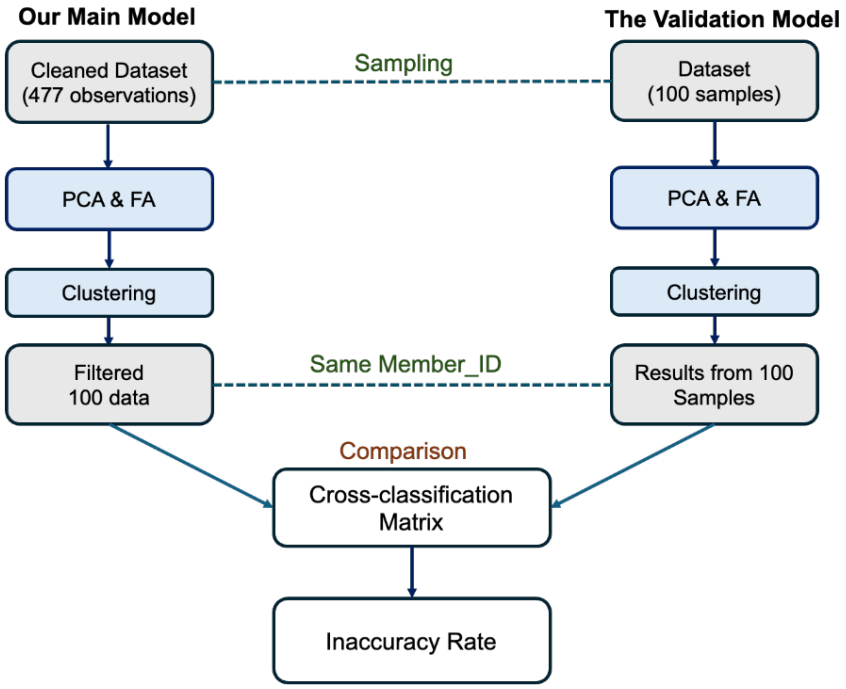



Figure 6.1 Internal Validation Process

We assigned an index (Member_ID) to clean the data after eliminating the outliers so we knew who was selected after sampling 100 observations and used it for further analysis. PCA, FA, and clustering were conducted to validate clusters following similar processes. It gives the same analysis as the main model, 9 PCs with oblique rotation for the best PC Extraction, Ward's technique, and the number of clusters is 3 from k-means gap statistics, as seen in Figure A2 in appendix.

Figure 6.2 shows the results of internal validation for cluster analysis. Firstly, cluster 1 has low TC1, TC2, TC5, TC4, TC8, TC6, and TC3 values with high-grade loanees and fully paid loans, indicating low credit risk. Secondly, cluster 2 has a high value on TC2 and TC3, indicating long-term employment and a moderate loan level, and a high value on TC8, indicating a charged-off loan status and high risk. Finally, cluster 3 has high scores on TC1, TC5, TC4, TC6, and TC3,

reflecting the borrower has high annual income, extended employment, and consistent credit activity. However, they have been delinquent in the prior two years, suggesting moderate risk.

| cluster | TC1 | TC2 | TC5 | TC4 | TC8 | TC6 | TC7 | TC3 | TC9 |
|---|---|---|---|---|---|---|---|---|---|
| <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | -0.4152204 | -0.3530810 | -0.3289576 | -0.3252452 | -0.4055501 | -0.43744148 | 0.1590465 | -0.4617293 | -0.02646895 |
| 2 | -0.2674507 | 0.3625376 | -0.1481554 | 0.1352274 | 2.2445265 | -0.07797605 | -0.1352291 | 0.4715041 | -0.26901348 |
| 3 | 0.5964018 | 0.2953662 | 0.4496201 | 0.3434190 | -0.3169876 | 0.55516446 | -0.1431276 | 0.3871851 | 0.12846742 |

Figure 6.2 The Result of Cluster Analysis for Validation.

Afterwards, we compared the primary and validation models using two methods. The first step is determining whether the observation is still in the same cluster. We compared both models using the same index before generating a cross-classification matrix, as in Table 6.1, to determine which cluster corresponds to which validation model. We also checked the characteristics of the clusters in the validation model, and they match the initial model; details are shown in Table 6.2. The matrix validation result shows an inaccuracy rate of approximately 29%, indicating a nearly 'fairly stable solution' (Hair, Black, and Babin, 2019).

| Cluster Number from Model | Cluster Number from Validation Model | | | Total |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1 | **45** | 0 | 27 | 72 |
| 2 | 0 | 0 | **12** | 12 |
| 3 | 2 | **14** | 0 | 16 |
| **Total** | 47 | 14 | 39 | 100 |

Table 6.1 Cross-Classification to Assess Cluster Stability

| Model | Low-risk Cluster | Moderate-risk Cluster | High risk Cluster |
|---|---|---|---|
| Main Model | 1 | 2 | 3 |
| Internal Validation Model | 1 | 3 | 2 |

Table 6.2 Cluster Identification from Initial Model to Validation Model

We proposed the second internal validation way as connectivity, the Dunn Index, and Silhouette Width are considered important measures (Brock et al., 2008) that use properties like clusters' connectedness, separation between them, and compactness of clusters to validate the analysis. We used the ‘clValid’ package (Brock et al., 2008) to perform these tests, and hierarchical clustering outperformed all tests. Hierarchical clustering with 3 clusters worked best for this dataset; refer to Figure A3 in appendix. It had ideal clustering—separated, compact, and well-connected.

## 7. Recommendations

We advised Lending Club to use the cluster analysis results to guide credit activity based on the low-risk (Cluster 1), moderate-risk (Cluster 2), and high-risk loanees (Cluster 3). A system of fixing the maximum borrowing limit for each category, along with continuous monitoring of the customers, is much needed to avoid default.

Lending Club can provide financial education and counselling to help high-risk borrowers improve their financial literacy and stability to compensate for the risk of high-risk clusters. Additionally, Lending Club might offer flexible loan products and terms to moderate-risk borrowers to meet their needs. Finally, Lending Club might offer low-risk borrowers competitive interest rates and loan terms to keep them.

## 8. Conclusion

Cluster analysis was performed on the data to determine if Lending Club loanees might be categorised to understand them better. Clustering using PC, 9-PC extraction, and oblique rotation found three groups: low-risk, moderate-risk, and high-risk. In addition, internal validation results with a 29% error rate indicate a 'fairly stable solution' and support our model's loan proposals. Lending Club should use these results for data-driven decision-making, such as offering financial education for high-risk clusters, customised terms for moderate-risk clusters, and competitive interest rates for low-risk borrowers.

**9. References**

Brock, G., Pihur, V., Datta, S. and Datta, S. (2008). *clValid , an R package for cluster validation*. [online] Available at: https://cran.r-project.org/web/packages/clValid/vignettes/clValid.pdf [Accessed 16 Mar. 2024].

Dalmaijer, E.S., Nord, C.L., and Astle, D.E. (2022). *Statistical power for cluster analysis*. BMC bioinformatics, 23(1), pp.205.

Davidson, I., (2002). *Understanding K-means non-hierarchical clustering.* Computer Science Department of State University of New York (SUNY), Albany.

Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I. and Akinyelu, A.A. (2022). *A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects.* Engineering Applications of Artificial Intelligence, 110, p.104-743. doi:https://doi.org/10.1016/j.engappai.2022.104743.

Hair et al. (2019). *Multivariate data analysis*. 8th edn. Hamshire: Pearson Education, Inc [England].

Li, X., Deng, S., Li, L., and Jiang, Y. (2019). *Outlier Detection Based on Robust Mahalanobis Distance and Its Application.* Open Journal of Statistics, 09(01), pp.15–26. doi:https://doi.org/10.4236/ojs.2019.91002.

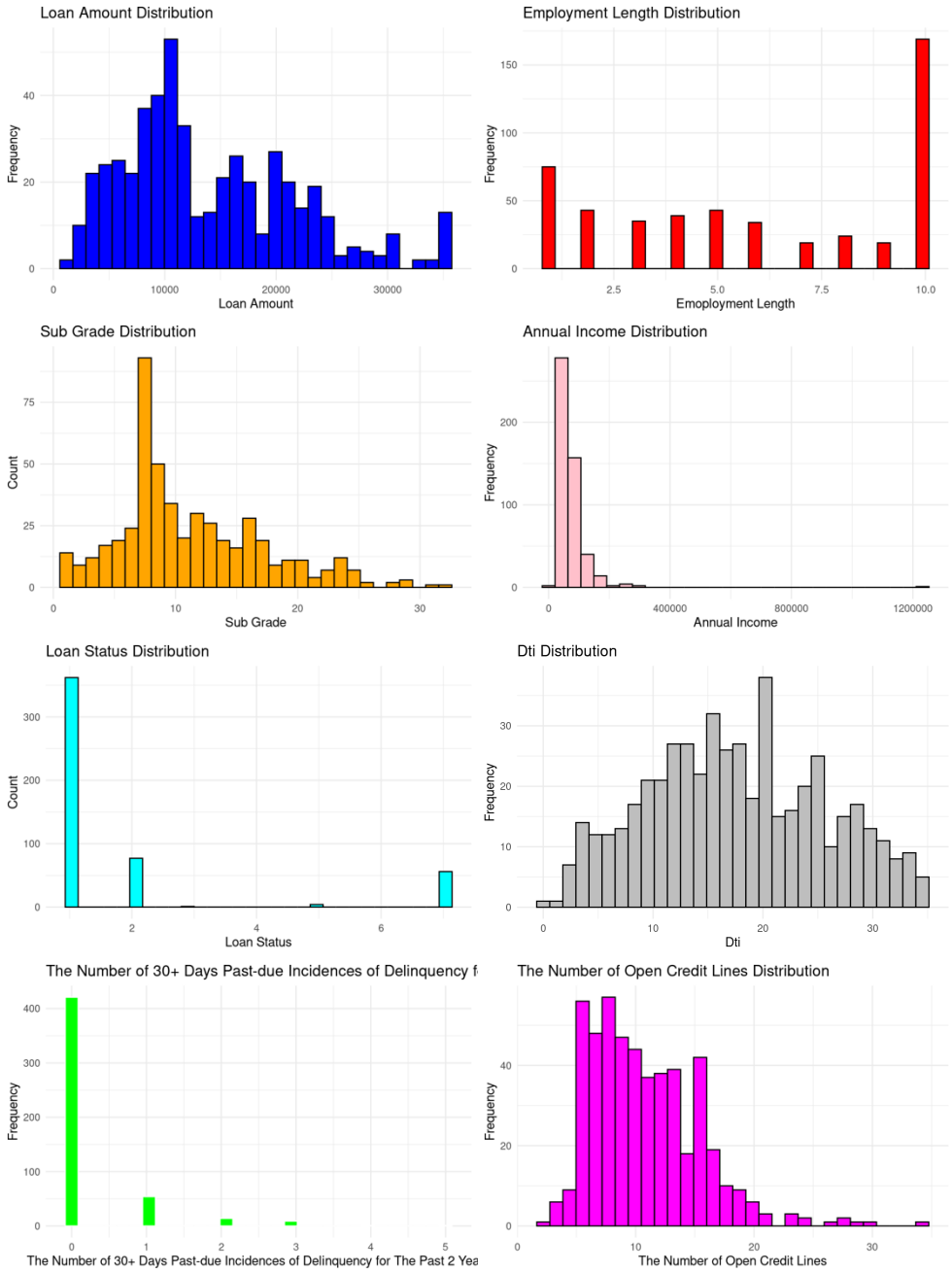
Murtagh, F., and Contreras, P. (2011). *Algorithms for hierarchical clustering: an overview.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp.86–97. doi:https://doi.org/10.1002/widm.53

Chikodili et al. (2021). *Outlier Detection in Multivariate Time Series Data Using a Fusion of K-Medoid, Standardized Euclidean Distance and Z-Score*. Communications in computer and information science, pp.259–271. doi:https://doi.org/10.1007/978-3-030-69143-1_21.

Rodriguez, M. Z., et al. (2019*). "Clustering algorithms: A comparative approach."* PloS one 14(1): e0210236.

## 10. Appendix

Table A1: Visualisations of the distribution



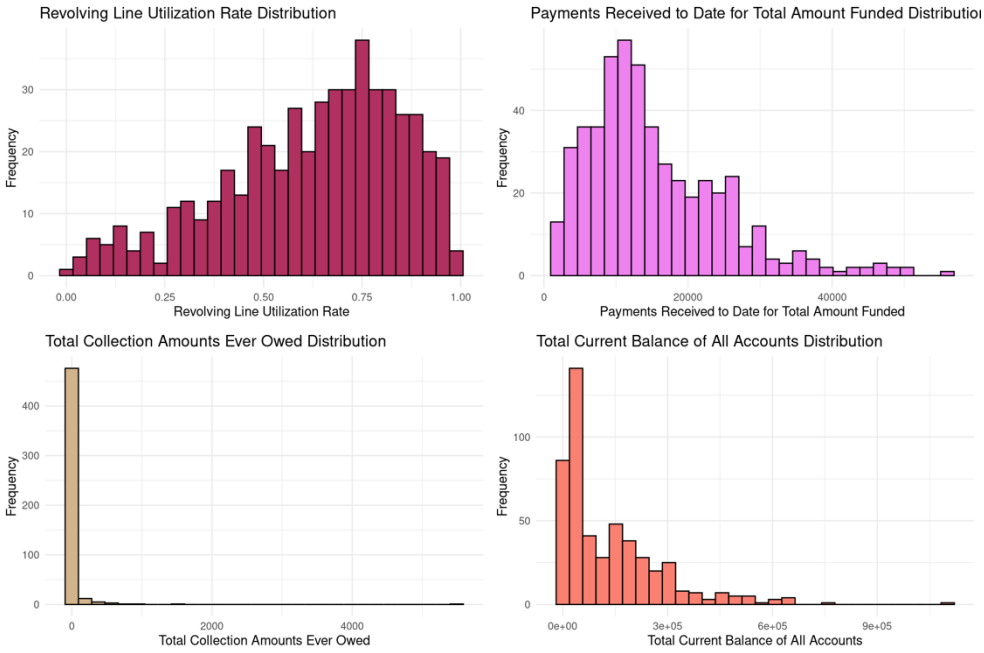



Table A2: Outliers

| loan_amnt | sub_grade | emp_length | annual_inc | loan_status | dti | delinq_2yrs | open_acc | revol_util | total_pymnt | tot_coll_amt | tot_cur_bal | loan_amnt_Z | sub_grade_Z | emp_length_Z | annual_inc_Z | loan_status_Z | dti_Z_score | delinq_2yrs_Z | open_acc_Z_s | revol_util_Z_sc | total_pymnt_Z | tot_coll_amt_Z | tot_cur_bal_Z | MahaDistanc | MahaPvalue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8500 | 14 | 5 | 63000 | 1 | 23.14 | 3 | 6 | 0.581 | 10157.27 | 0 | 27272 | -0.70 | 0.47 | -0.30 | -0.15 | -0.46 | 0.70 | 4.32 | -1.06 | -0.21 | -0.54 | -0.11 | -0.73 | 25.71 | 0.007 |
| 8000 | 31 | 6 | 1.00E+05 | 2 | 20.71 | 0 | 30 | 0.822 | 9103.37 | 0 | 97918 | -0.76 | 3.33 | -0.02 | 0.41 | 0.07 | 0.39 | -0.37 | 4.15 | 0.86 | -0.65 | -0.11 | -0.25 | 35.15 | 0.000 |
| 10000 | 16 | 10 | 87500 | 1 | 14.44 | 3 | 10 | 0.781 | 12829.42 | 0 | 254240 | -0.51 | 0.80 | 1.13 | 0.22 | -0.46 | -0.39 | 4.32 | -0.19 | 0.68 | -0.27 | -0.11 | 0.81 | 20.82 | 0.035 |
| 11000 | 17 | 10 | 53620 | 7 | 28.67 | 5 | 8 | 0.854 | 8882.63 | 0 | 251472 | -0.38 | 0.97 | 1.13 | -0.29 | 2.71 | 1.39 | 7.45 | -0.62 | 1.00 | -0.68 | -0.11 | 0.80 | 64.37 | 0.000 |
| 35000 | 23 | 10 | 99000 | 7 | 31.44 | 0 | 11 | 0.818 | 7652.32 | 0 | 107796 | 2.68 | 1.98 | 1.13 | 0.40 | 2.71 | 1.73 | -0.37 | 0.03 | 0.84 | -0.81 | -0.11 | -0.18 | 83.22 | 0.000 |
| 12000 | 8 | 10 | 69000 | 1 | 12.7 | 4 | 8 | 0.902 | 12578.92 | 0 | 376727 | -0.25 | -0.54 | 1.13 | -0.06 | -0.46 | -0.60 | 5.88 | -0.62 | 1.21 | -0.29 | -0.11 | 1.65 | 40.28 | 0.000 |
| 8000 | 14 | 10 | 33900 | 1 | 28.39 | 0 | 15 | 0.864 | 10143.96 | 5491 | 48787 | -0.76 | 0.47 | 1.13 | -0.59 | -0.46 | 1.35 | -0.37 | 0.90 | 1.05 | -0.55 | 20.58 | -0.58 | 424.55 | 0.000 |
| 13200 | 10 | 1 | 44000 | 7 | 18.57 | 3 | 17 | 0.59 | 14455.36 | 0 | 27347 | -0.10 | -0.20 | -1.45 | -0.43 | 2.71 | 0.13 | 4.32 | 1.33 | -0.17 | -0.10 | -0.11 | -0.73 | 38.64 | 0.000 |
| 9000 | 14 | 10 | 75000 | 1 | 27.86 | 3 | 13 | 0.73 | 10872.38 | 0 | 133342 | -0.63 | 0.47 | 1.13 | 0.03 | -0.46 | 1.28 | 4.32 | 0.46 | 0.45 | -0.47 | -0.11 | -0.01 | 22.73 | 0.019 |
| 10050 | 9 | 3 | 88000 | 7 | 29.59 | 0 | 15 | 0.925 | 9381.13 | 0 | 642370 | -0.50 | -0.37 | -0.88 | 0.23 | 2.71 | 1.50 | -0.37 | 0.90 | 1.32 | -0.63 | -0.11 | 3.45 | 32.74 | 0.001 |
| 6250 | 14 | 5 | 41000 | 7 | 15.37 | 0 | 28 | 0.124 | 7291.95 | 0 | 186228 | -0.98 | 0.47 | -0.30 | -0.48 | 2.71 | -0.27 | -0.37 | 3.71 | -2.23 | -0.84 | -0.11 | 0.35 | 36.04 | 0.000 |
| 9000 | 18 | 10 | 185000 | 1 | 15.28 | 3 | 11 | 0.898 | 10956.24 | 0 | 315190 | -0.63 | 1.14 | 1.13 | 1.70 | -0.46 | -0.28 | 4.32 | 0.03 | 1.20 | -0.46 | -0.11 | 1.23 | 24.73 | 0.010 |
| 20000 | 18 | 10 | 1.00E+05 | 1 | 8.46 | 0 | 7 | 0.511 | 25996.95 | 1504 | 16808 | 0.77 | 1.14 | 1.13 | 0.41 | -0.46 | -1.13 | -0.37 | -0.84 | -0.52 | 1.11 | 5.56 | -0.80 | 39.08 | 0.000 |
| 20000 | 5 | 3 | 1250000 | 1 | 2.01 | 0 | 19 | 0.317 | 22862.28 | 0 | 575475 | 0.77 | -1.05 | -0.88 | 17.79 | -0.46 | -1.93 | -0.37 | 1.76 | -1.38 | 0.78 | -0.11 | 3.00 | 376.69 | 0.000 |
| 14000 | 11 | 10 | 82000 | 1 | 9.32 | 3 | 11 | 0.698 | 16091.22 | 0 | 242795 | 0.01 | -0.04 | 1.13 | 0.14 | -0.46 | -1.02 | 4.32 | 0.03 | 0.31 | 0.07 | -0.11 | 0.74 | 21.76 | 0.026 |
| 16800 | 32 | 4 | 110000 | 1 | 18.5 | 0 | 29 | 0.505 | 17524.17 | 0 | 371937 | 0.36 | 3.49 | -0.59 | 0.56 | -0.46 | 0.12 | -0.37 | 3.93 | -0.54 | 0.22 | -0.11 | 1.61 | 36.00 | 0.000 |
| 35000 | 14 | 9 | 185800 | 7 | 9.27 | 0 | 12 | 0.604 | 7014.7 | 0 | 611937 | 2.68 | 0.47 | 0.84 | 1.71 | 2.71 | -1.03 | -0.37 | 0.24 | -0.11 | -0.87 | -0.11 | 3.25 | 80.80 | 0.000 |
| 12000 | 7 | 10 | 175000 | 1 | 20.05 | 1 | 11 | 0.575 | 13990.72 | 0 | 1103872 | -0.25 | -0.71 | 1.13 | 1.55 | -0.46 | 0.31 | 1.19 | 0.03 | -0.23 | -0.15 | -0.11 | 6.59 | 54.14 | 0.000 |
| 12000 | 8 | 5 | 80000 | 1 | 7.49 | 3 | 9 | 0.598 | 14184.49 | 0 | 172096 | -0.25 | -0.54 | -0.30 | 0.11 | -0.46 | -1.25 | 4.32 | -0.41 | -0.13 | -0.12 | -0.11 | 0.26 | 22.36 | 0.022 |
| 6250 | 17 | 1 | 38356 | 2 | 15.56 | 3 | 15 | 0.896 | 7735 | 0 | 60520 | -0.98 | 0.97 | -1.45 | -0.52 | 0.07 | -0.25 | 4.32 | 0.90 | 1.19 | -0.80 | -0.11 | -0.50 | 27.93 | 0.003 |
| 35000 | 28 | 10 | 120000 | 1 | 22.45 | 0 | 19 | 0.58 | 55145.01 | 0 | 85444 | 2.68 | 2.82 | 1.13 | 0.71 | -0.46 | 0.61 | -0.37 | 1.76 | -0.21 | 4.14 | -0.11 | -0.33 | 36.77 | 0.000 |
| 4800 | 21 | 1 | 51000 | 1 | 11.43 | 1 | 34 | 0.558 | 5337.32 | 0 | 186267 | -1.17 | 1.64 | -1.45 | -0.33 | -0.46 | -0.76 | 1.19 | 5.01 | -0.31 | -1.05 | -0.11 | 0.35 | 44.00 | 0.000 |
| 6000 | 13 | 3 | 250000 | 7 | 14.36 | 3 | 16 | 0.791 | 3696.08 | 0 | 752481 | -1.02 | 0.30 | -0.88 | 2.68 | 2.71 | -0.40 | 4.32 | 1.11 | 0.72 | -1.22 | -0.11 | 4.20 | 48.14 | 0.000 |

Table A3. The result of Kaiser-Meyer-Olkin factor adequacy on the main model

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = data_no_outliers_std)
Overall MSA =  0.51
MSA for each item =
   loan_amnt     sub_grade    emp_length    annual_inc   loan_status           dti
        0.50          0.64          0.76          0.65          0.13          0.44
 delinq_2yrs     open_acc    revol_util   total_pymnt  tot_coll_amt  tot_cur_bal
        0.75          0.57          0.53          0.49          0.63          0.66
```

Table A4. The result of PCA in 12 PCs without rotation on the main model

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **loan_amnt** | 0.84 | | -0.35 | | | | | | | | | |
| **total_pymnt** | 0.83 | | -0.38 | | | | | | | | | |
| **annual_inc** | 0.67 | -0.44 | | 0.33 | | | | | | | 0.39 | |
| **tot_cur_bal** | 0.58 | | 0.43 | 0.32 | | | | | | -0.39 | | |
| **dti** | | 0.6 | | -0.49 | | | | | | | | |
| **sub_grade** | 0.43 | 0.59 | | | | | | | -0.49 | | | |
| **revol_util** | 0.34 | 0.56 | | 0.39 | | | -0.36 | | | 0.33 | | |
| **open_acc** | 0.42 | | 0.56 | -0.43 | | | | | -0.36 | | | |
| **delinq_2yrs** | | | 0.3 | 0.32 | -0.64 | | | 0.49 | | | | |
| **loan_status** | | 0.54 | | | 0.54 | | 0.35 | 0.39 | | | | |
| **tot_coll_amt** | | | 0.44 | | | 0.84 | | | | | | |
| **emp_length** | | | 0.35 | | | | 0.68 | -0.44 | | | | |

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SS loadings | 2.86 | 1.63 | 1.36 | 1.08 | 0.97 | 0.96 | 0.93 | 0.81 | 0.59 | 0.45 | 0.31 | 0.04 |
| Proportion Var | 0.24 | 0.14 | 0.11 | 0.09 | 0.08 | 0.08 | 0.08 | 0.07 | 0.05 | 0.04 | 0.03 | 0.00 |
| Cumulative Var | 0.24 | 0.37 | 0.49 | 0.58 | 0.66 | 0.74 | 0.82 | 0.88 | 0.93 | 0.97 | 1.00 | 1.00 |
| Proportion Explained | 0.24 | 0.14 | 0.11 | 0.09 | 0.08 | 0.08 | 0.08 | 0.07 | 0.05 | 0.04 | 0.03 | 0.00 |
| Cumulative Proportion | 0.24 | 0.37 | 0.49 | 0.58 | 0.66 | 0.74 | 0.82 | 0.88 | 0.93 | 0.97 | 1.00 | 1.00 |

Table A5. The result of PCA in 9 PCs without rotation on the main model

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| **loan_amnt** | 0.84 | | | | | | | | |
| **sub_grade** | 0.43 | 0.59 | | | | | | | -0.49 |
| **emp_length** | | | | | | | 0.68 | -0.44 | |
| **annual_inc** | 0.67 | -0.44 | | | | | | | |
| **loan_status** | | 0.54 | | | 0.54 | | | | |
| **dti** | | 0.6 | | -0.49 | | | | | |
| **delinq_2yrs** | | | | | -0.64 | | | 0.49 | |
| **open_acc** | 0.42 | | 0.56 | -0.43 | | | | | |
| **revol_util** | | 0.56 | | | | | | | |
| **total_pymnt** | 0.83 | | | | | | | | |
| **tot_coll_amt** | | | 0.44 | | | 0.84 | | | |
| **tot_cur_bal** | 0.58 | | 0.43 | | | | | | |

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| SS loadings | 2.86 | 1.63 | 1.36 | 1.08 | 0.97 | 0.96 | 0.93 | 0.81 | 0.59 |
| Proportion Var | 0.24 | 0.14 | 0.11 | 0.09 | 0.08 | 0.08 | 0.08 | 0.07 | 0.05 |
| Cumulative Var | 0.24 | 0.37 | 0.49 | 0.58 | 0.66 | 0.74 | 0.82 | 0.88 | 0.93 |
| Proportion Explained | 0.26 | 0.15 | 0.12 | 0.10 | 0.09 | 0.09 | 0.08 | 0.07 | 0.05 |
| Cumulative Proportion | 0.26 | 0.40 | 0.52 | 0.62 | 0.71 | 0.79 | 0.87 | 0.95 | 1.00 |

Table A6. The result of PC Extraction, 9 factors, oblique rotation on the main model

| | TC1 | TC2 | TC9 | TC4 | TC3 | TC8 | TC7 | TC5 | TC6 |
|---|---|---|---|---|---|---|---|---|---|
| loan_amnt | 0.99 | | | | | | | | |
| total_pymnt | 0.95 | | | | | | | | |
| tot_cur_bal | | 0.84 | | | | | | | |
| annual_inc | | 0.69 | | -0.36 | | | | | |
| sub_grade | | | 0.95 | | | | | | |
| revol_util | | 0.45 | 0.51 | 0.37 | -0.38 | | | | |
| dti | | | | 0.93 | | | | | |
| open_acc | | | | | 0.92 | | | | |
| loan_status | | | | | | 0.99 | | | |
| emp_length | | | | | | | 1 | | |
| delinq_2yrs | | | | | | | | 1 | |
| tot_coll_amt | | | | | | | | | 1 |

| | TC1 | TC2 | TC9 | TC4 | TC3 | TC8 | TC7 | TC5 | TC6 |
|---|---|---|---|---|---|---|---|---|---|
| SS loadings | 2.05 | 1.50 | 1.26 | 1.20 | 1.11 | 1.04 | 1.02 | 1.02 | 1.01 |
| Proportion Var | 0.17 | 0.12 | 0.10 | 0.10 | 0.09 | 0.09 | 0.08 | 0.08 | 0.08 |
| Cumulative Var | 0.17 | 0.30 | 0.40 | 0.50 | 0.59 | 0.68 | 0.76 | 0.85 | 0.93 |
| Proportion Explained | 0.18 | 0.13 | 0.11 | 0.11 | 0.10 | 0.09 | 0.09 | 0.09 | 0.09 |
| Cumulative Proportion | 0.18 | 0.32 | 0.43 | 0.54 | 0.64 | 0.73 | 0.82 | 0.91 | 1.00 |

Table A7. Linkage methods used for cluster analysis on the main model

| average | single | complete | ward |
|---|---|---|---|
| 0.8252392 | 0.7366227 | 0.8935709 | 0.9554026 |

Figure A1. ‘Gap statistics’ graphs from K-means clustering on the main model

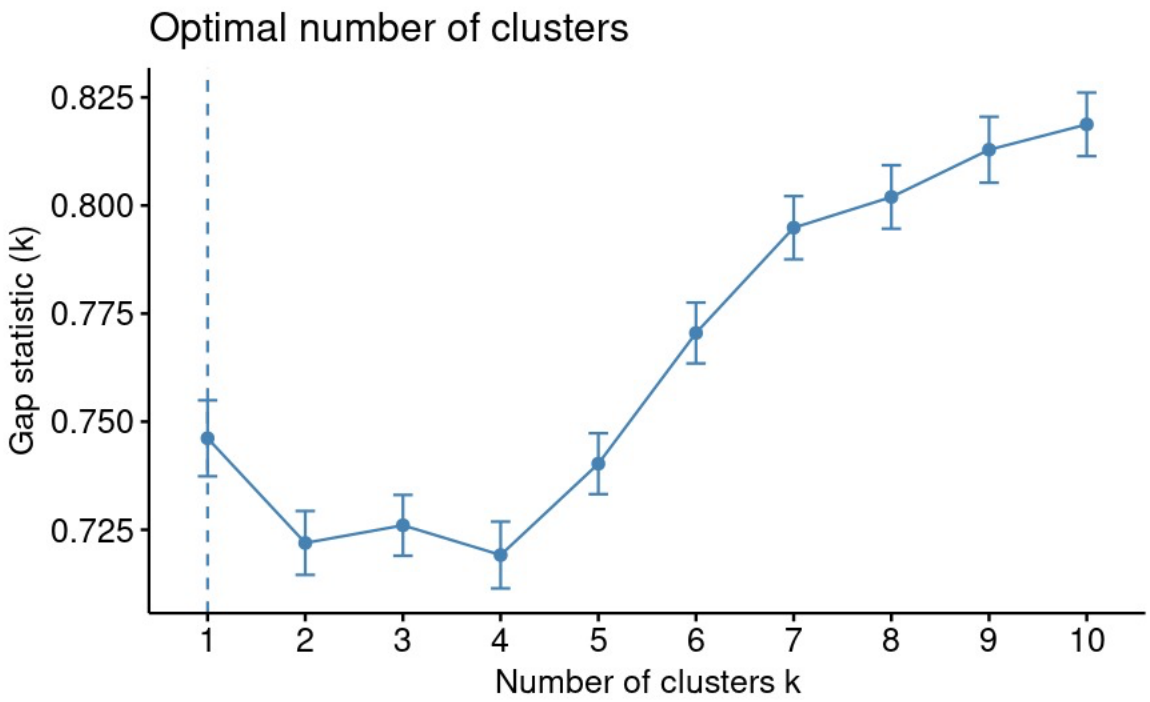


Table A8. The result of Kaiser-Meyer-Olkin factor adequacy on the internal validation model.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = data_int_val_std)
Overall MSA =  0.56
MSA for each item =
   loan_amnt    sub_grade   emp_length   annual_inc  loan_status          dti  delinq_2yrs     open_acc
        0.55         0.62         0.62         0.61         0.21         0.63         0.79         0.65
  revol_util  total_pymnt tot_coll_amt  tot_cur_bal
        0.61         0.52         0.64         0.62
```

Table A9. The result of PCA in 12 PCs without rotation on the internal validation model

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| loan_amnt | 0.83 | | | | | | | | | | | |
| total_pymnt | 0.81 | | -0.34 | | | | | | | | | |
| tot_cur_bal | 0.6 | -0.33 | 0.34 | 0.32 | -0.33 | | | | | | -0.33 | |
| sub_grade | 0.56 | 0.55 | | | | | | 0.31 | 0.32 | | | |
| annual_inc | 0.52 | -0.48 | 0.34 | | | | | | | | | |
| revol_util | 0.4 | 0.58 | | | -0.46 | | | | | 0.31 | | |
| dti | 0.5 | 0.55 | | | | | | -0.32 | | -0.31 | | |
| emp_length | | | 0.63 | | 0.37 | | 0.35 | -0.47 | | | | |
| loan_status | | 0.54 | 0.57 | | | | | 0.4 | -0.31 | | | |
| open_acc | 0.46 | | | 0.61 | 0.31 | -0.32 | | | | | | |
| delinq_2yrs | 0.46 | | | | -0.41 | -0.65 | 0.4 | | | | | |
| tot_coll_amt | | | -0.44 | 0.55 | | 0.35 | 0.59 | | | | | |

Table A10. The result of PCA in 9 PCs without rotation on the internal validation model

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| loan_amnt | 0.83 | | | | | | | | |
| sub_grade | 0.56 | 0.55 | | | | | | | |
| emp_length | | | 0.63 | | | | | -0.47 | |
| annual_inc | 0.52 | -0.48 | | | | | | | |
| loan_status | | 0.54 | 0.57 | | | | | 0.4 | |
| dti | 0.5 | 0.55 | | | | | | | |
| delinq_2yrs | 0.46 | | | | -0.41 | -0.65 | | | |
| open_acc | 0.46 | | | 0.61 | | | | | |
| revol_util | 0.4 | 0.58 | | | -0.46 | | | | |
| total_pymnt | 0.81 | | | | | | | | |
| tot_coll_amt | | | -0.44 | 0.55 | | | 0.59 | | |
| tot_cur_bal | 0.6 | | | | | | | | |

Table A11. The result of PC Extraction, 9 factors, oblique rotation on internal validation model

| | TC1 | TC2 | TC5 | TC4 | TC8 | TC6 | TC7 | TC3 | TC9 |
|---|---|---|---|---|---|---|---|---|---|
| loan_amnt | 1 | | | | | | | | |
| total_pymnt | 0.94 | | | | | | | | |
| revol_util | | 0.99 | | | | | | | |
| dti | | 0.47 | -0.41 | 0.42 | | | | | |
| annual_inc | | | 0.9 | | | | | | |
| open_acc | | | | 1 | | | | | |
| loan_status | | | | | 0.99 | | | | |
| delinq_2yrs | | | | | | 1.01 | | | |
| tot_coll_amt | | | | | | | 1 | | |
| emp_length | | | | | | | | 1 | |
| sub_grade | | 0.44 | | | | | | | -0.61 |
| tot_cur_bal | | | 0.53 | | | | | | 0.57 |

Table A12. Linkage methods used for cluster analysis on the internal validation model

| average | single | complete | ward |
|---|---|---|---|
| 0.8128179 | 0.7462503 | 0.8283755 | 0.8600994 |

Figure A2. 'Gap statistics' graphs from K-means clustering on the internal validation model

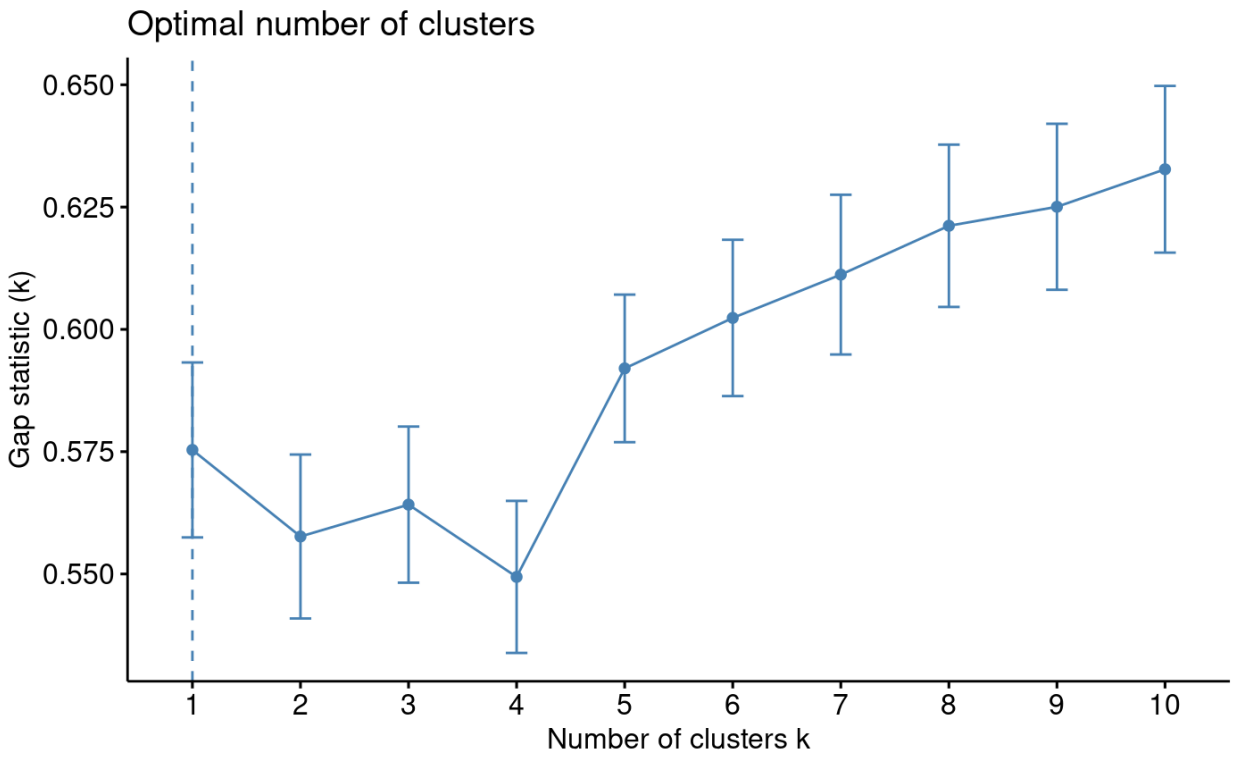



Figure A3. Internal Validation Using 'clValid' Package

```
##
## Clustering Methods:
##  hierarchical kmeans
##
## Cluster sizes:
##  3 4 5 6 7 8
##
## Validation Measures:
##                                  3        4        5        6        7        8
##
## hierarchical Connectivity   12.9925  49.3694  52.0984  53.7734  53.9734  54.8845
##              Dunn            0.3496   0.2192   0.2192   0.2192   0.2192   0.2192
##              Silhouette      0.4170   0.3232   0.2669   0.2532   0.2426   0.2434
## kmeans       Connectivity  159.4353 170.5238 162.9921 217.9540 281.6782 279.4948
##              Dunn            0.0986   0.0986   0.1053   0.1078   0.0911   0.0911
##              Silhouette      0.2074   0.2043   0.1879   0.1709   0.1279   0.1305
##
## Optimal Scores:
##
##              Score   Method       Clusters
## Connectivity 12.9925 hierarchical 3
## Dunn          0.3496 hierarchical 3
## Silhouette    0.4170 hierarchical 3
```