Table of Contents

Executive Summary

This report presented an in-depth analysis aimed at enhancing loan portfolio management through Cluster Analysis (CA) on 500 observations. The objective was to segment borrowers into distinct groups with similar traits in risks and economic benefits for Lending Club (LC), allowing tailored product offering, marketing effort optimisation, and customer service improvement. The methodology involved a rigorous process of variable selection, addressing outliers and multicollinearity to refine the dataset for analysis. Factor Analysis (FA) was applied to address multicollinearity and ensure a focus on the most informative attributes. Employing both hierarchical and K-means clustering techniques on the standardised dataset integrating both FA results and original variables, the study successfully identified four unique borrower clusters, differentiated by their risk and potential return levels. The validation demonstrated the high stability of our cluster solution, confirming its applicability in general portfolio management practices. The segmentation revealed that a significant portion of the existing customer base belonged to the unfavourable group, characterised by high risk but low returns, prompting a strategic imperative for LC to address this segment proactively. The desirable customer segments with substantial financial rewards were also identified, highlighting an attractive target for LC's marketing endeavours together with proper risk management strategies. Another emerging customer pool with young demographics was also pinpointed for LC to grasp the potential growth in their future loan demand.

Introduction

This report aims to adopt CA to significantly enhance loan portfolio management by organising borrowers according to their risk levels and potential returns. Our main hypothesis suggests that CA will reveal distinct borrower segments to target with tailored services, leading to better efficiency and customer satisfaction.

Data Preparation

Variable Selection

The project commenced with a dataset of 56 variables and narrowed down to 15 key variables, discarding 41 that did not meet our criteria for quality and relevance. Variables with more than 20% missing data or text-based were excluded due to their challenging quantification for clustering purposes. We also removed non-informative identifiers such as ID and member ID. Nominal categorical variables could potentially provide insightful information for cluster interpretation but are excluded during cluster construction. Additionally, variables exhibiting minimal variation, with more than 90% of the data falling into a single category, were considered not sufficiently distinctive, thus, eliminated.

We selected most relevant and insightful variables for the project objective of clustering customers by their risk and return profiles. Loan performance indicators such as loan status (encoded from best to worst based on risk level), principal, interest, along with the total payment received, paint a clear picture of the loan's health and the borrower's reliability. Borrower credit behaviours is represented through the debt-to-income ratio and encoded subgrades variable highlighting their credit management and potential risk factors. The earliest credit line opening date offers additional context on the borrower's credit history length. Together, these meticulously chosen variables provide a robust foundation for our CA, aiming to uncover meaningful patterns and insights into quality and potential returns of our loan portfolio. More detailed explanation of variables selection is shown in Appendix 1.

Outlier identification

The pre-analysis phase for outlier identification encompasses both univariate and multivariate methods. Outlier detection is critical due to CA’s sensitivity to outliers, which can skew the analysis. Utilising z-scores, we identified and removed 14 outliers exceeding a z-score of 4, specifically identified in installment, annual_inc, revol_bal, total_rec and earliest_cr_line_months_before_mar2024. This approach, informed by statistical rigor and visualised with box plots (Appendix 2), ensured the dataset's integrity.

Mahalanobis distance identified 27 significant outliers by evaluating multivariate relationships within the dataset. These outliers, determined through p-values less than 0.01 against a Chi-Square distribution, were examined if they represented a minority cluster. As no pattern was observed, those records were removed from further analysis.

Correlation and Multicollinearity Check

Our focus on correlation and multicollinearity through specific tests ensured the dataset's suitability for CA. The pairwise correlation matrix (Figure 1) highlighted significant correlations amongst loan_amnt, total_pymnt, total_rec_prncp, total_rec_int and installment (>0.70 coefficient for each pair), and between int_rate and sub_grade (0.99), suggesting it is crucial to treat correlation by FA prior to conduct CA. The Kaiser-Meyer-Olkin (KMO) test presents an overall score of 0.63, suggesting moderate suitability for uncovering underlying factors. Bartlett's test indicates a p-value significantly below 0.01, affirming the presence of notable correlations among variables.

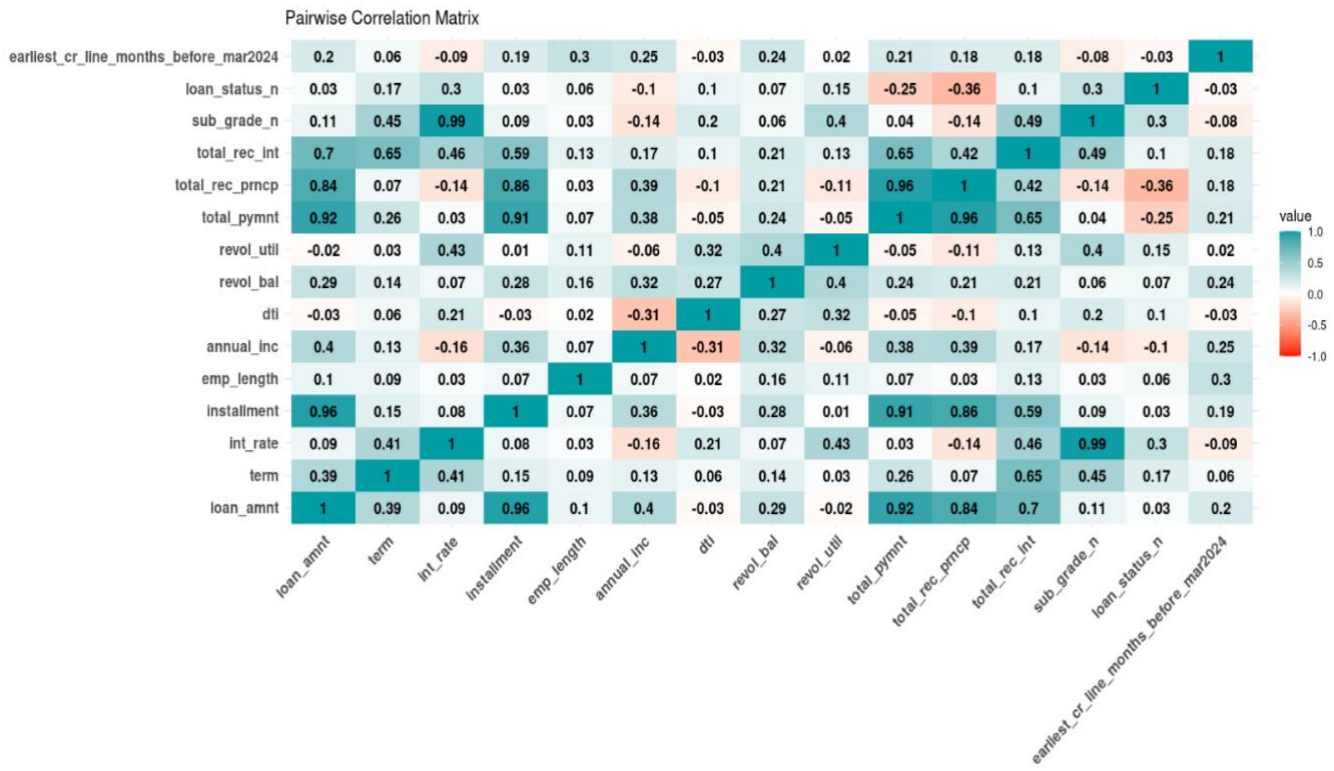


Figure 1 - Inter-variable Pairwise Correlation Matrix

Factor Analysis

FA was utilised to address the multicollinearity issues and reduce dimensionality. Whereas, both FA and Principal Component Analysis (PCA) are dimensionality reduction techniques, FA was selected to accompany CA due to its higher interpretability.

The two common factor extraction methods are PC and Maximum Likelihood (ML). The models exhibited similar patterns in terms of SS loadings and explained variance. Additionally, various factor numbers were examined with the optimal solution observed with 10 or 11 factors. Most variables had high loadings (>0.9) on a single factor, indicating strong associations. Overall, PC extraction with oblique rotation was selected as the final model due to its superlative performance with no cross-loading (Figure 2). Meanwhile, other models exhibited cross-loading and high inter-factor correlation, obscuring factor interpretation and distinctiveness. For example, the ML with oblique rotation model, highlighted correlation greater than 0.6 for some factors, suggesting potential redundancy.

Consequently, Loan Quantum (TC1), representing total_rec_prncp, installment, total_pymnt and loan_amnt and Risk Identifiers (TC2), comprising int_rate and sub_grade was utilised in the later analysis.

	item	TC1	TC2	TC6	TC10	TC3	TC8	TC9	TC5	TC4	TC7	TC11	h2
	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<dbl>
total_rec_prncp	11	1.00											0.9839172
installment	4	0.94											0.9830116
loan_amnt	1	0.86											0.9853822
total_pymnt	10	0.86											0.9889935
int_rate	3		1.00										0.9948820
sub_grade_n	13		0.99										0.9948210
loan_status_n	14			0.98									0.9933945
annual_inc	6				1.00								0.9999752
revol_bal	8					1.00							0.9999903
earliest_cr_line_months_before_mar2024	15						1.00						0.9999951
revol_util	9							1.00					0.9999953
emp_length	5								1.00				0.9999991
dti	7									1.00			0.9999992
term	2										0.97		0.9997992
total_rec_int	12											0.94	0.9990752

Figure 2 - PC Extraction with Oblique Rotation

Cluster Analysis

As factor scores TC1 and TC2 are used in our clustering, original variables are standardised prior to combining with factor scores for analysis.

Hierarchical Clustering

Linkage and distance selection

As there are 4 linkage methods, “average”, “single”, “complete” and “ward”, we calculated the agglomerative coefficient to select the best. Ward’s linkage method produced the highest figure (Table 1), thus, adopted in our final hierarchical cluster.

Table 1 - Agglomerative Coefficient for Linkage Methods

Linkage Method	Average	Single	Complete	Ward
Agglomerative Coefficient	0.745	0.587	0.824	0.96

We tested 3 methods of similarity measurement: Euclidean, Chebychev(Maximum) and Manhattan distance and compared the cluster outcomes to select the most appropriate model. Amongst various approaches deployed, Chebyshev distance and Ward linkage generated the most balanced cluster allocation (Figure 3). Refer to Appendix.3 for dendrograms of other methods.

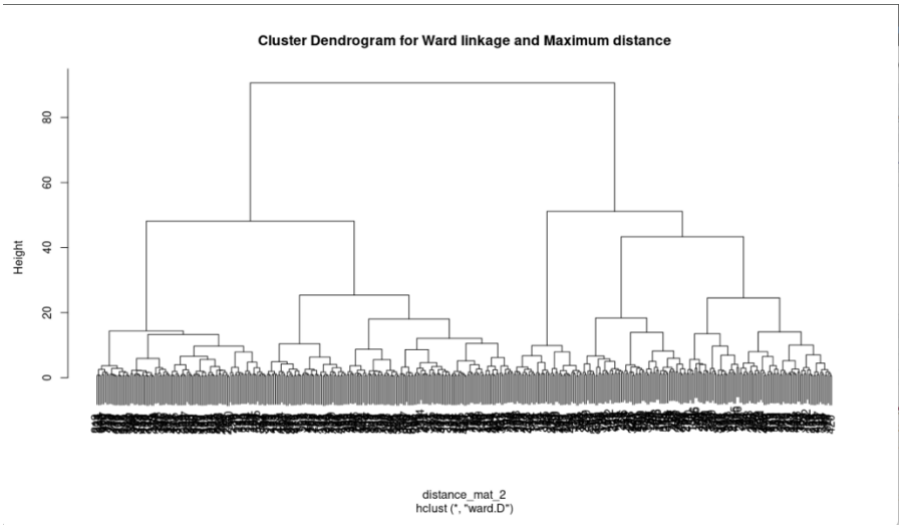


Figure 3 - Dendrogram of Ward Linkage and Chebychev(Maximum) Distance

Optimal number of clusters

We used the gap statistic to determine how many clusters the observations should be grouped into. As the gap statistic increased considerably from 3 to 4 (Figure 4), we categorised our observations into 4 clusters.

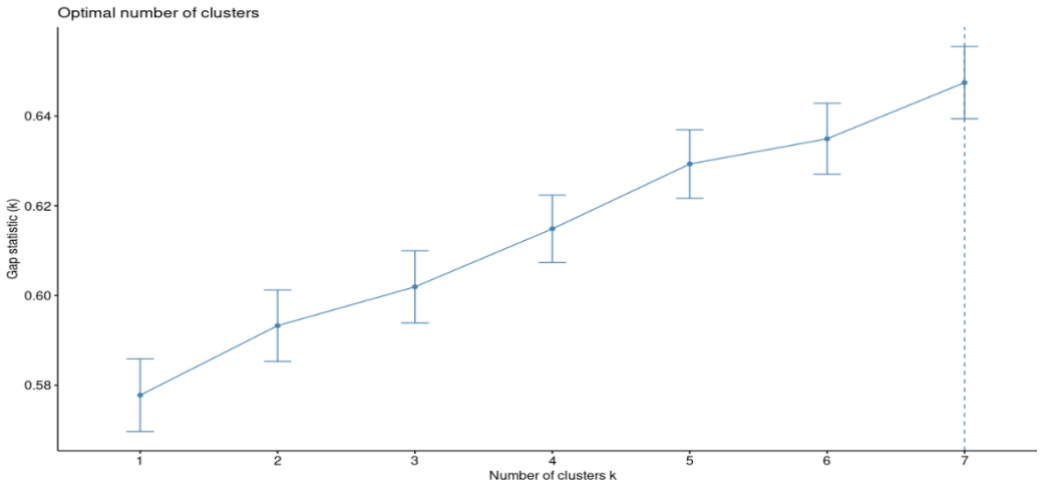


Figure 4 - Gap Statistic for Hierarchical Clustering

The final model for hierarchical method utilised Chebyshev distance and Ward linkage dividing observations into 4 clusters with the sizes of 154, 106, 43 and 156.

Non-Hierarchical Clustering (K-Means)

We utilised the optimal number of clusters from hierarchical clustering (4) during the operation of K-Means to allow observations to move freely among clusters and avoid early combination. K-Means clustering (Figure 5) was our final cluster solution, with cluster sizes of 117, 194, 61 and 87.

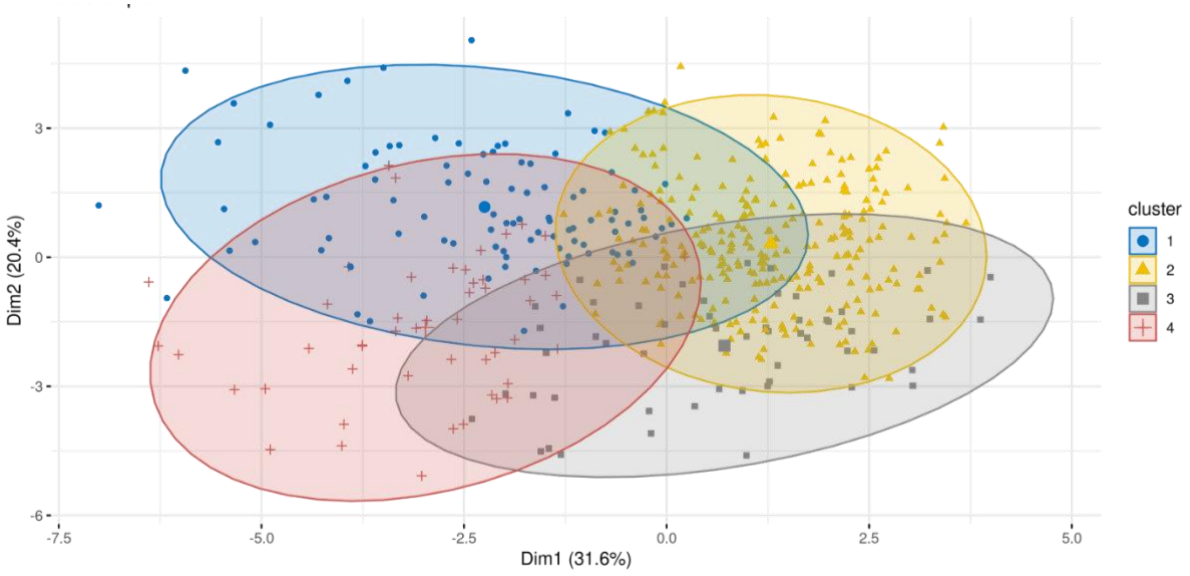


Figure 5 - Cluster plot from K-Means

Cluster Interpretation

The cluster centroids (Table 2) classified customers into four groups, delineating their economic benefits and risks to LC. Variables such as loan quantum (TC1), total interests collected, customers' annual income, employment length, credit revolving balances, revolving line utilisation rate, and customer tenure emerge as pivotal indicators of the economic gains derived from customers. Notably, clusters 3 and 4, exhibiting the highest means in these variables, are classified as profitable customers. Conversely, TC2, a risk identification, along with loan status and debt to income ratio (DTI), can measure customers' default risk levels, with higher values indicating greater risk. Hence, cluster 2 and 3, characterised by the highest central points in these indicators, are classified as high-risk customers, whilst cluster 1 and 4 are categorised as low-risk groups.

Table 2 - Cluster centroids

Cluster	term	emp_ length	annual_ inc	dti	revol_bal	revol_util	total_ rec_int	loan_ status_n	customer tenor	TC1	TC2	Profile
1	-0.39	-0.19	0.18	-0.79	-0.68	-1.02	-0.50	-0.36	0.11	-0.24	-0.85	Low Risk Low Return
2	-0.38	-0.12	-0.60	0.40	-0.23	0.40	-0.33	0.18	-0.39	-0.41	0.28	High Risk Low Return
3	2.44	0.25	0.32	0.22	0.32	0.14	1.75	0.46	0.17	0.54	1.11	High Risk High Return
4	-0.34	0.34	0.87	0.01	1.21	0.38	0.19	-0.24	0.61	0.85	-0.26	Low Risk High Return

Value colour coded from lowest to highest

These four clusters can then be plotted based on their risk and return spectrum (Figure 6).

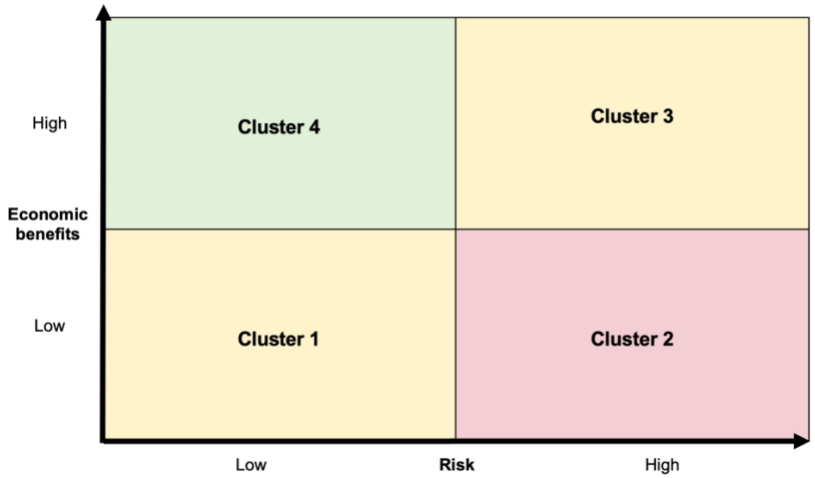


Figure 6 - Cluster labels

Validation

To assess CA's generalisability, we conducted internal validation, which found a high degree of stability, with only a 3.5% discrepancy observed between the training and validation datasets. Specifically, we randomly selected 200 out of the 500 observations, applying the same FA and CA techniques employed during the training phase. FA grouped total_rec_prncp, installment, total_pymnt and loan_amnt into TC1 while int_rate and sub_grade was categorised into TC2. Subsequently, K-means clustering method with four clusters were performed on the data set with TC1, TC2 and 9 other standardised variables. The cluster labels were then assigned to each observation based on their risk and return spectrum using centroid values (Table 3):

Table 3 – Cluster centroids of Validation set

Cluster	term	emp_ length	annual_ inc	dti	revol_bal	revol_util	total_ rec_int	loan_ status_n	customer tenor	TC1	TC2	Profile
4	-0.31	-0.20	0.36	-0.93	-0.69	-0.96	-0.43	-0.36	0.20	-0.14	-0.85	Low Risk Low Return
2	-0.31	0.03	-0.63	0.57	-0.13	0.36	-0.22	0.11	-0.30	-0.29	0.22	High Risk Low Return
3	3.17	0.24	0.17	0.02	0.27	0.06	1.53	0.80	0.01	0.37	1.38	High Risk High Return
1	-0.31	0.11	1.10	-0.15	1.31	0.46	0.48	-0.17	0.50	0.83	-0.01	Low Risk High Return

Value colour coded from lowest to highest

The validation cluster labels reported a 3.5% variance as compared to the original clusters (Table 4). This minimal level of discrepancy indicates our cluster solution can be representative of the general population and maintain stability across different datasets.

Table 3 - Validation results

Original labels	Cluster size	# different cases	*% deviation*
Low risk low return	54	3	*5.6%*
High risk low return	93	2	*2.2%*
High risk high return	17	0	*0.0%*
Low risk high return	36	2	*5.6%*

Cluster Profile

Our solution separated the customer base into four distinctive groups to assist LC in prioritising more profitable customers while managing their risk levels.

The most desirable customer group (Cluster 4), accounting for only 19% of the total sample size, should be of LC's strategic focus to expand. These customers paid the highest loan principals and interest amounts at relatively low default possibilities. This can be explained by their highest annual income and longest employment length (Figure 7). Therefore, they are more likely to have higher savings or stable income to pay for loan interests, leading to their default rate being one of the lowest among all clusters (Figure 8). Furthermore, 62.1% of the group had mortgages leading to their high loan demand (Figure 9).

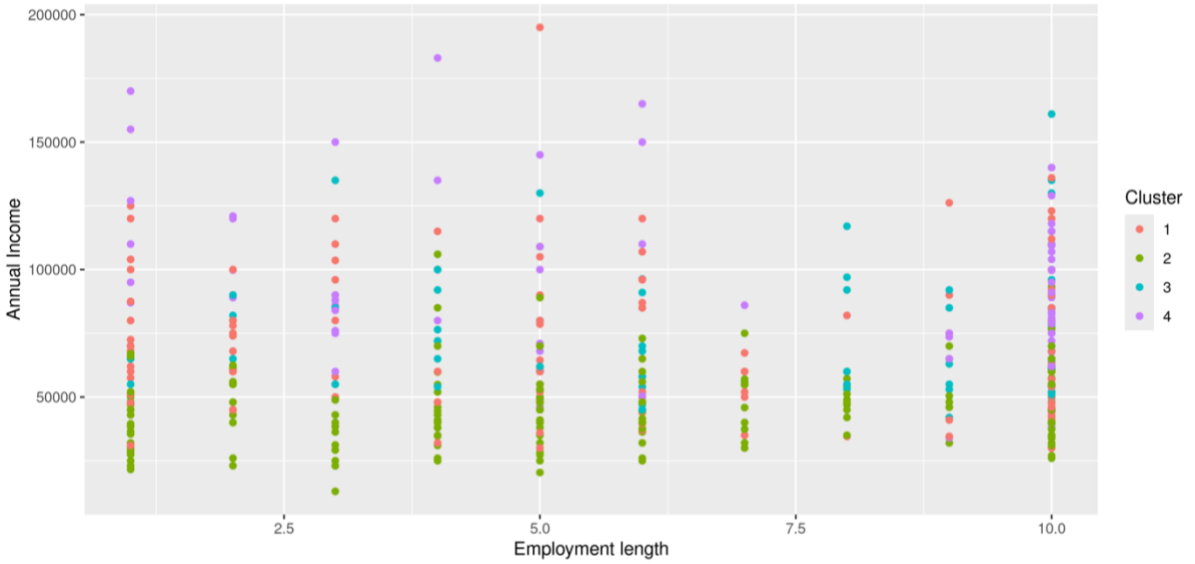


Figure 7 - Customers' annual income by employment length

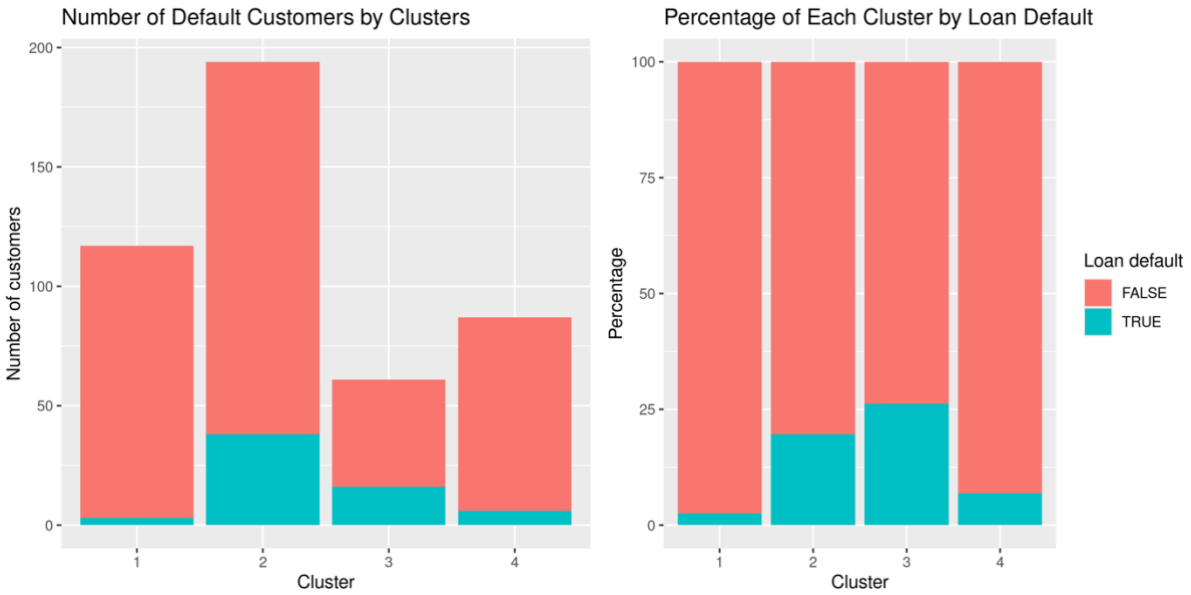


Figure 8 - Number of default customers by clusters

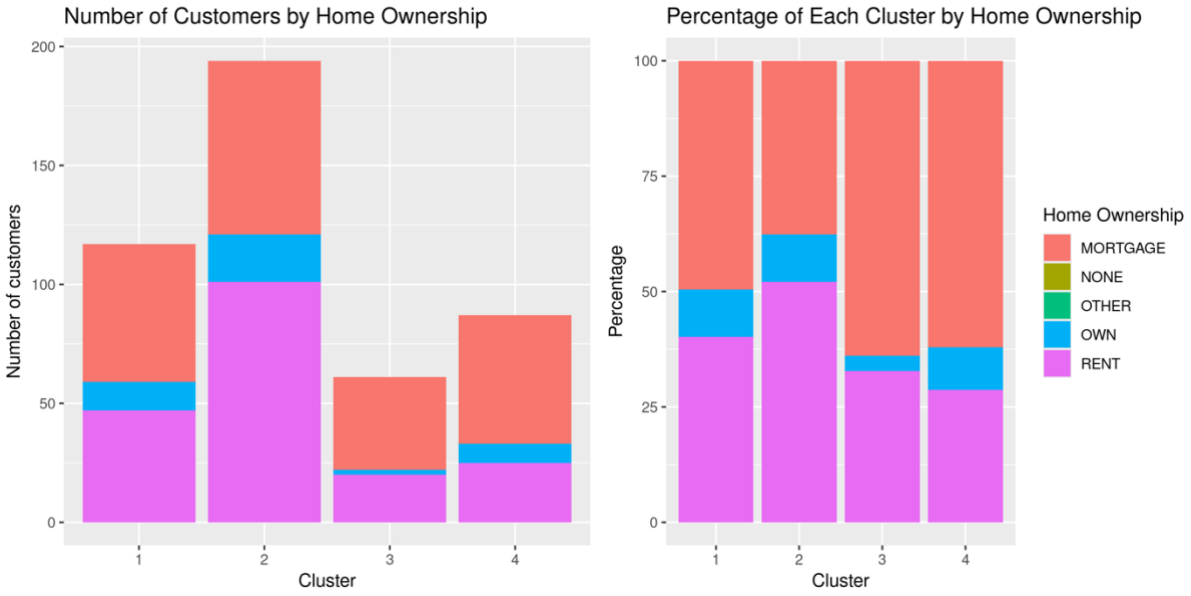


Figure 9 - Number of customers by home ownership

Another lucrative segment LC should concentrate on to increase the profitability is Cluster 3, comprising customers with both elevated risk and returns. Though having a higher default rate and worse loan status than Cluster 4, this group yield superior profits for LC due to their higher interest rates (Figure 10). Additionally, their substantial mortgage rate (63%), surpassing all other clusters, justifies their considerable loan demand (Figure 9). Therefore, it is advisable for LC to consciously grow this risky yet prosperous clientele.

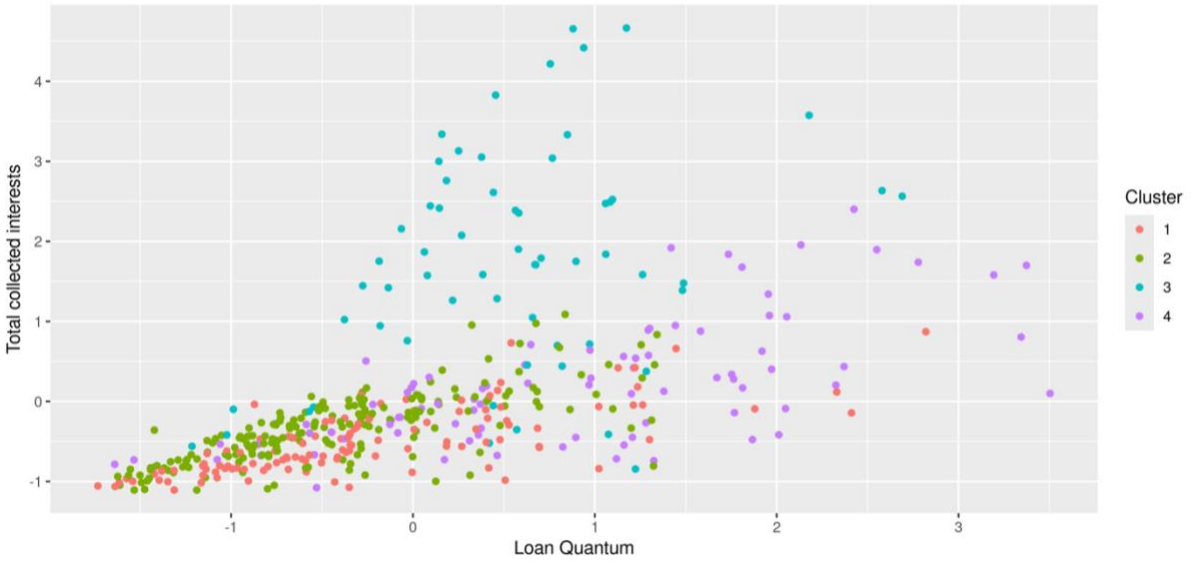


Figure 10 - Loan quantum by collected interests

Whilst not currently yielding high revenue, Cluster 1 is the least risky group with default risk of 2.6% (Figure 8). These are emerging customers, likely to be young demographics with the shortest credit history, shortest employment length yet not substantially lower annual income. They currently borrow least with lowest Loan Quantum and DTI. However, as they mature, both their incomes and credit demands will likely improve. Targeting this cluster with appropriate strategies will enable LC to nurture the relationship and capture their future loan demand growth, adjusting them to the high-value customer groups.

The most troublesome customer group, Cluster 2, with high risk and low returns, accounted for the majority of customer base (42%). This group paid the lowest interest amount to investors, given their smallest loan principals (Figure 11) while having the highest number of defaulters (Figure 8). This can be explained by their lowest annual income and shortest employment length among all groups (Figure 7). Therefore, LC is recommended to thoroughly manage this group for proper alleviation measures.

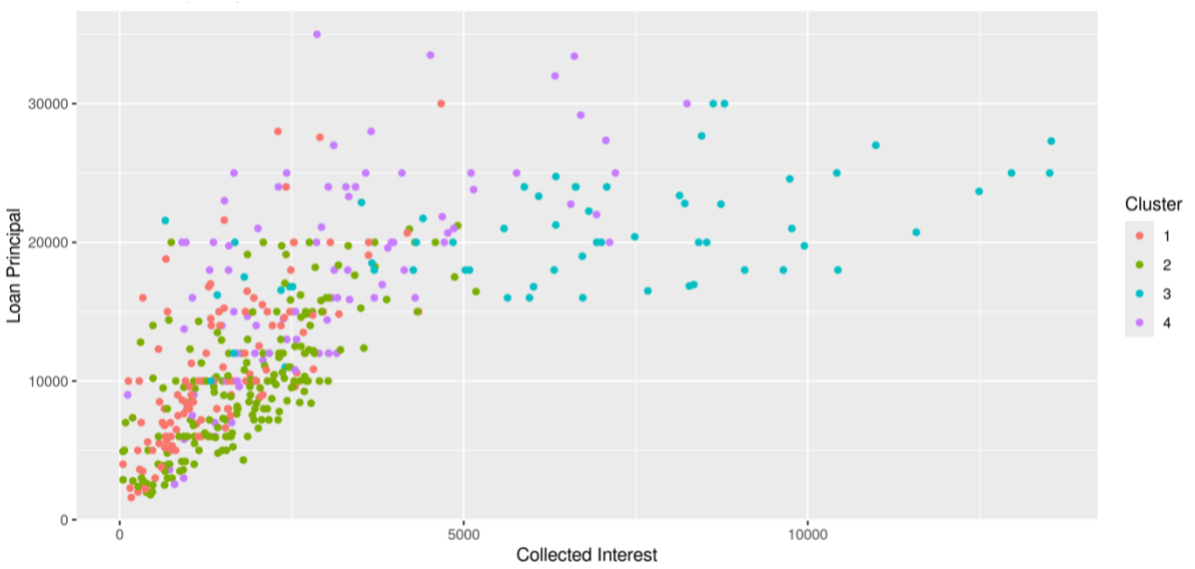


Figure 11 - Loan Principal by Collected Interest

Conclusion

Through a meticulous process of variable selection, data preparation, and application of both hierarchical and K-Means clustering methods, four distinct customer clusters were identified based on their risk levels and potential profitability. Validation demonstrates high stability of the clustering solution, with a low deviation between training and test datasets. This report highlights the importance of focusing on expanding the most desirable customer group (Cluster 4), characterised by high loan repayments and lower default risks, and managing the elevated risk yet potentially profitable customers in Cluster 3. Additionally, it advises close monitoring of Cluster 1 for future growth opportunities and recommends strategic management of Cluster 2, which poses substantial risk and low returns.

Appendix 1 - Variable table

Variable	Freqs (% of Valid)	Missing	Selection	Reason
id [numeric]	500 distinct values	0 (0.0%)	Remove	Non-informative
member_id [numeric]	500 distinct values	0 (0.0%)	Remove	Non-informative
loan_amnt [numeric]	196 distinct values	0 (0.0%)	Keep	Total loan amount the borrower can get
funded_amnt [numeric]	195 distinct values	0 (0.0%)	Remove	Similar to loan amount
funded_amnt_inv [numeric]	232 distinct values	0 (0.0%)	Remove	Similar to loan amount
term [numeric]	36: 40641 (81.3%) 60: 9359 (18.7%)	0 (0.0%)	Keep	Number of payments on the loan

int_rate [numeric]	45 distinct values	0 (0.0%)	Keep	Interest rate is an important indicator
installment [numeric]	430 distinct values	0 (0.0%)	Keep	Tells how much payment made per month by borrower
grade [character]	83 (16.6) % 18 (37.4) 7 % 11 (23.6) 8 % 70 (14.0) % 25 (5.0%) 15 (3.0%) 2 (0.4%)	0 (0.0%)	Remove	Sub grade gives a more detailed grade
sub_grade [factor]	12 (2.4%) 9 (1.8%) 16 (3.2%) 21 (4.2%) 25 (5.0%) 25 (5.0%) 38 (7.6%) 53 (10.6) % 38 (7.6%) 33 (6.6%) 23 (46.0) 0 %	0 (0.0%)	Keep	Tells the grade of the loan in detail
emp_title [character]	2 (0.4%) 2 (0.4%) 2 (0.4%) 1 (0.2%) 1 (0.2%) 1 (0.2%) 1 (0.2%) 1 (0.2%) 1 (0.2%) 1 (0.2%) 48 (97.4) 1 %	6 (1.2%)	Remove	Free text nominal categorical variable

emp_length [numeric]	1 : 5 (11.9 8%) 2 : 4 (8.00 %) 3 : 3 (6.84 %) 4 : 4 (8.21 %) 5 : 6 (12.00 0%) 6 : 4 (9.89 %) 7 : 1 (3.68 %) 8 : 2 (4.00 %) 9 : 2 (4.42 %) 10 : 15 (31.457 4%)	0 (0.0%)	Keep	Length of employment in years
home_ownership [factor]	25 (50.84 %) 0 (0.0%) 0 (0.0%) 45 (9.0%) 20 (40.21 %)	0 (0.0%)	Remove	Shows asset strength
annual_inc [numeric]	200 distinct values	0 (0.0%)	Keep	Annual income of the borrower shows asset strength
verification_status [factor]	20 (41.47 %) 84 (16.8%) 20 (41.89 %)	0 (0.0%)	Remove	Categorical variable

issue_d [POSIXct, POSIXt]	2 (4.4%) 2 3 (6.8%) 4 4 (8.0%) 0 6 (12.2) 1 % 5 (10.4) 2 % 6 (13.2) 6 % 6 (12.6) 3 % 6 (13.0) 5 % 6 (13.6) 8 % 2 (5.8%) 9	0 (0.0%)	Remove	Date format
loan_status [factor]	64 (12.8) % 70 (14.0) % 0 (0.0%) 35 (71.2) 6 % 3 (0.6%) 0 (0.0%) 7 (1.4%)	0 (0.0%)	Keep	An important status of current loan payment
pymnt_plan [character]	50 (100.0) 0 %	0 (0.0%)	Remove	>95% one class imbalance
desc [character]	1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%)	198 (39.6%)	Remove	> 20% missing value

addr_state [factor]	5 (1.0%) 4 (0.8%) 3 (0.6%) 11 (2.2%) 90 (18.0%) 15 (3.0%) 7 (1.4%) 0 (0.0%) 1 (0.2%) 37 (7.4%) 327 (65.4%)	0 (0.0%)	Remove	Non-informative
dti [numeric]	453 distinct values	0 (0.0%)	Keep	Shows debt compared to income of a borrower
delinq_2yrs [numeric]	0 : 424 (84.8%) 1 : 57 (11.4%) 2 : 12 (2.4%) 3 : 3 (0.6%) 4 : 1 (0.2%) 5 : 2 (0.4%) 10 : 1 (0.2%)	0 (0.0%)	Keep	An important figure to see how many times the borrower past-due
earliest_cr_line [POSIXct, POSIXt]	232 distinct values	0 (0.0%)	Keep	This show credit history often an important factor in banking credit

inq last 6 mths [numeric]	0 : 25 (50.3 6%) 1 : 12 (25.5 0%) 2 : 81 (16.2%) 3 : 38 (7.6 %) 4 : 3 (0.6 %)	0 (0.0%)	Remove	Non-informative
mths_since last delinq [numeric]	75 distinct values	287 (57.4%)	Remove	>20% missing variable
mths_since_last_record [numeric]	15 distinct values	480 (96.0%)	Remove	>20% missing variable
open_acc [numeric]	25 distinct values	0 (0.0%)	Remove	Non-informative
pub_rec [numeric]	0 : 48 (96.0 0%) 1 : 20 (4.0 %)	0 (0.0%)	Remove	95% one class imbalance
revol_bal [numeric]	495 distinct values	0 (0.0%)	Keep	Total credit revolving balance

revol_util [numeric]	371 distinct values	0 (0.0%)	Keep	The amount of credit the borrower is using relative to all available revolving credit.
total_acc [numeric]	48 distinct values	0 (0.0%)	Remove	Total number of credit line in the borrower's credit file
total_pymnt [numeric]	500 distinct values	0 (0.0%)	Keep	Shows how much the borrower has paid to date
total_pymnt_inv [numeric]	499 distinct values	0 (0.0%)	Remove	Similar to total payment
total_rec_prncp [numeric]	298 distinct values	0 (0.0%)	Keep	Total principal received to date
total_rec_int [numeric]	499 distinct values	0 (0.0%)	Keep	Total interest received to date
total_rec_late_fee [numeric]	16 distinct values	0 (0.0%)	Keep	Total late fee received to date

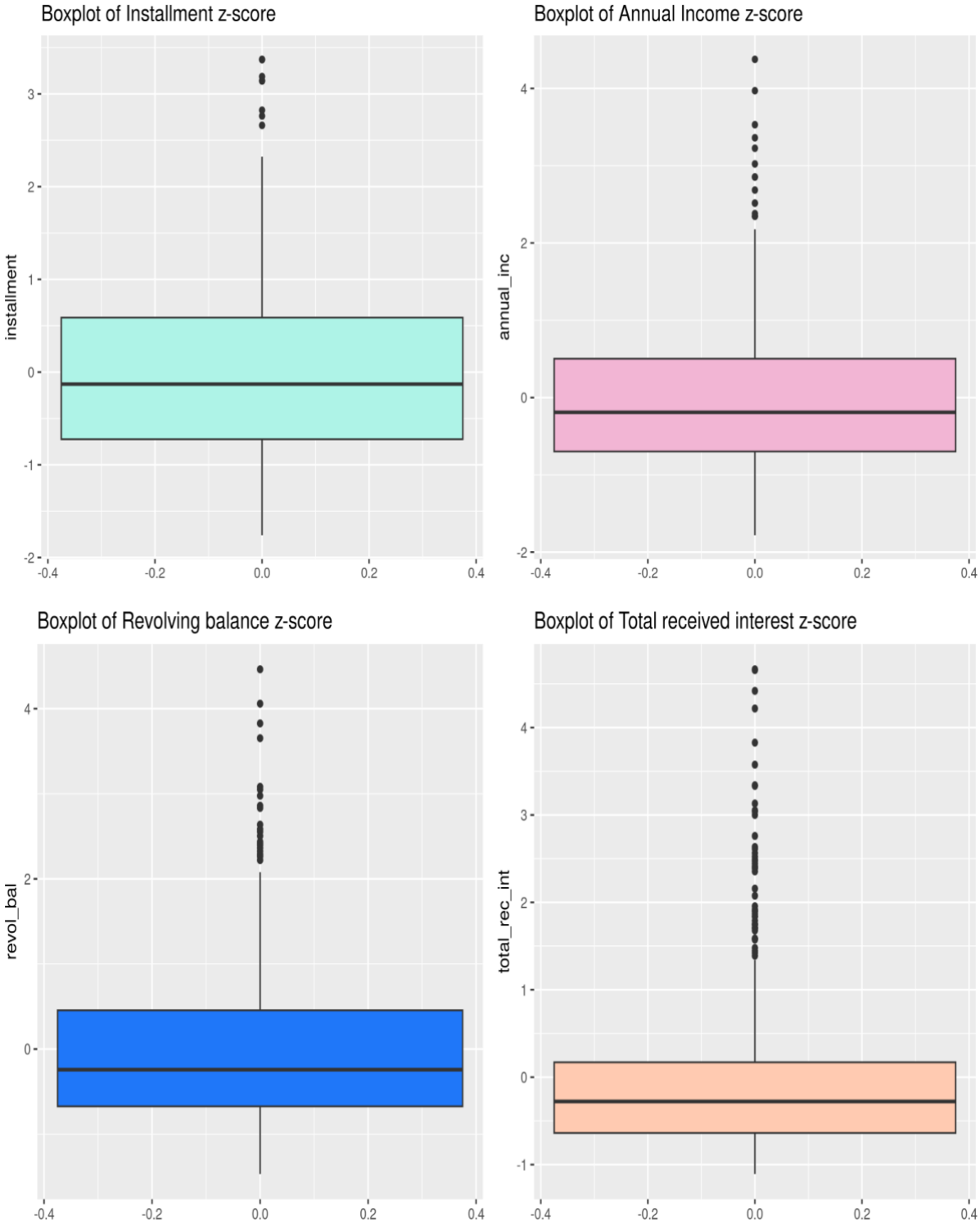
recoveries [numeric]	49 distinct values	0 (0.0%)	Remove	Only charged off loan has recovery figure
collection_recovery_fee [numeric]	46 distinct values	0 (0.0%)	Remove	Only charged off loan has recovery figure
last_pymnt_d [POSIXct, POSIXt]	41 distinct values	0 (0.0%)	Remove	Date format
last_pymnt_amnt [numeric]	493 distinct values	0 (0.0%)	Remove	Total payment is more important
next_pymnt_d [POSIXct, POSIXt]	6 (81.2%) 1 (18.8%)	420 (84.0%)	Remove	>20% missing variables
last_credit_pull_d [POSIXct, POSIXt]	38 distinct values	0 (0.0%)	Remove	Date format

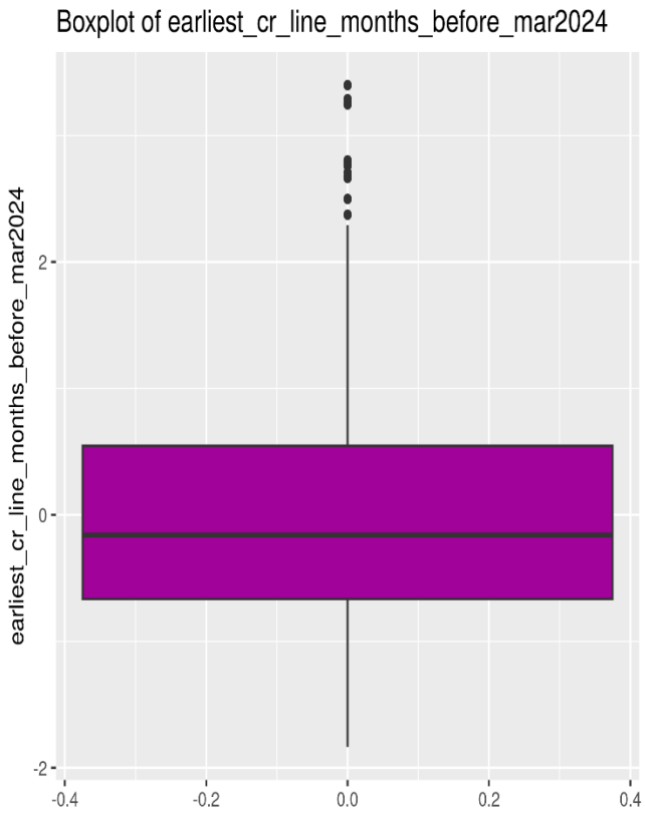
collections_12_mths_ex_med [numeric]	0: 5 (100.0%)	0 (0.0%)	Remove	>95% one class imbalance
mths_since_last_major_derog [numeric]	45 distinct values	438 (87.6%)	Remove	>20% missing variables
policy_code [numeric]	1: 5 (100.0%)	0 (0.0%)	Remove	>95% one class imbalance
acc_now_delinq [numeric]	0: 5 (100.0%)	0 (0.0%)	Remove	>95% one class imbalance
tot_coll_amt [numeric]	20 distinct values	150 (30.0%)	Remove	>20% missing variables
tot_cur_bal [numeric]	349 distinct values	150 (30.0%)	Remove	>20% missing variables

total_credit_rv [numeric]	261 distinct values	150 (30.0%)	Remove	>20% missing variables
loan is bad [logical]	426 (85.2%) 74 (14.8%)	0 (0.0%)	Remove	Indicate if a loan is default or not

Appendix 2 – Box plots for variables with potential outliers

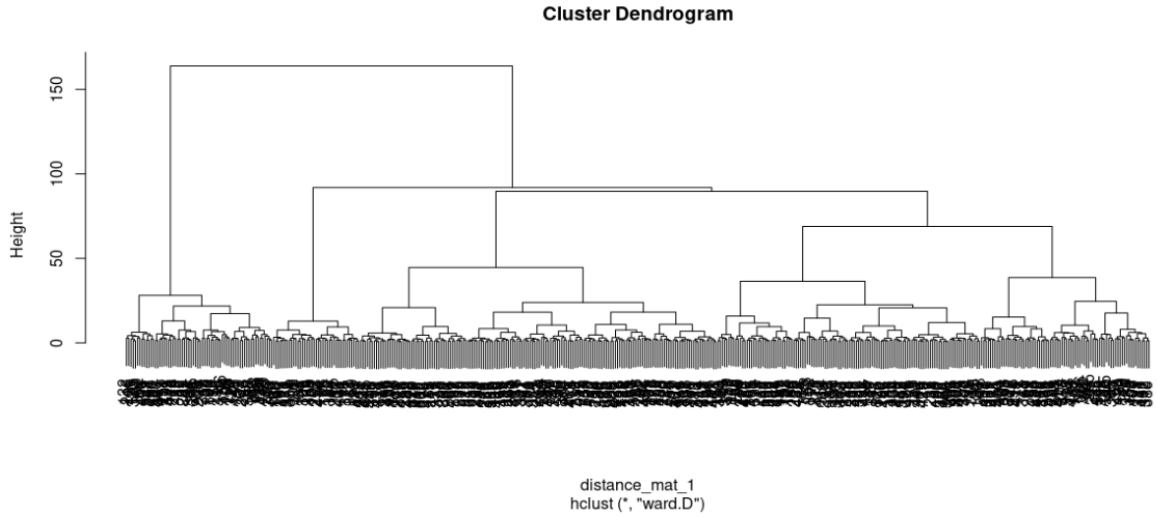
Appendix 2 – Z-Scores for different variables



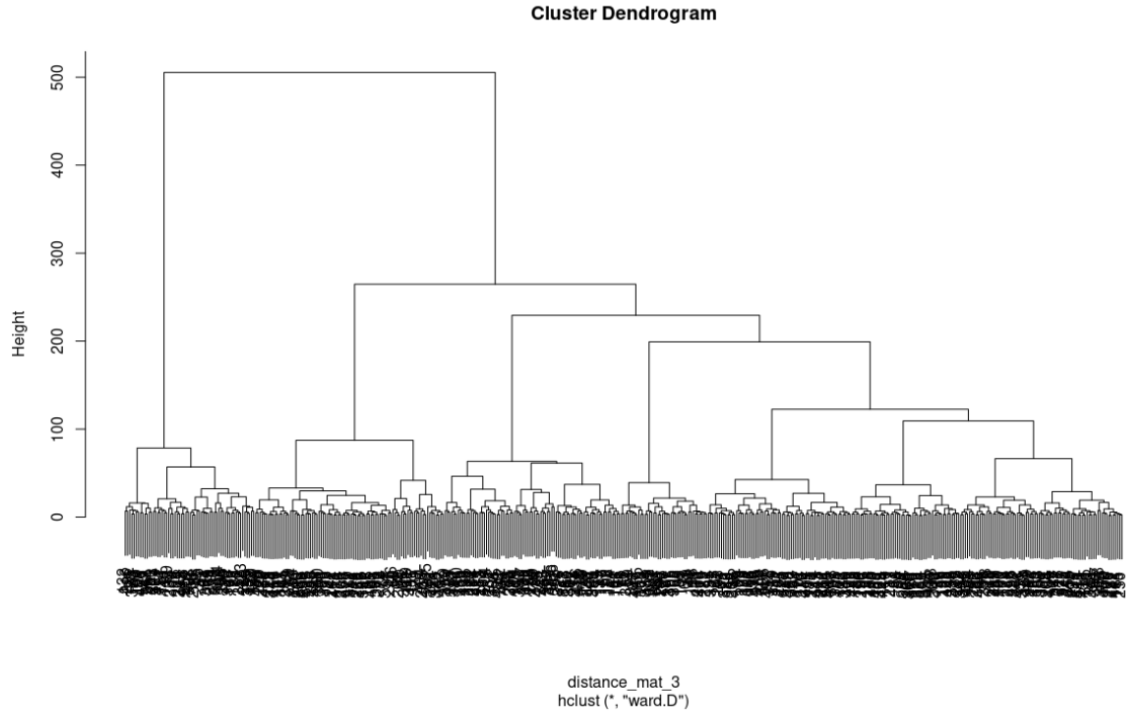


– Appendix 2a - Outlier Detection in Banking Data: Box Plot Analysis

Appendix 3 – Dendrograms for different distances



Appendix 3a - Dendrogram of Clusters, Ward Linkage, and Euclidean Distance



Appendix 3b - Dendrogram of Clusters, Ward Linkage, and Manhattan Distance