Comparative Analysis of Hierarchical and K-Means Clustering Algorithms on Home Equity Loan Data: A Data Mining Approach

Isaiah Thompson Ocansey

Department of Mathematics and Statistics

University of Texas at El Paso

Email: iocansey@miners.utep.edu

May, 2022

Abstract

This study presents a comprehensive comparative analysis of hierarchical clustering and K-means clustering algorithms applied to home equity loan data (HMEQ dataset). The research addresses the challenge of customer segmentation in financial services by implementing advanced clustering techniques on a dataset containing 5,960 observations with 13 variables. Through systematic data preprocessing including missing value imputation using Multiple Imputation by Chained Equations (MICE), logarithmic transformation, and Gower distance matrix computation, we achieved robust clustering solutions. Both clustering methods consistently identified an optimal two-cluster solution, validated through silhouette analysis and scree plots. The hierarchical clustering using Ward's linkage method and K-means clustering demonstrated perfect agreement (Jaccard index = 1.0, Rand index = 1.0), suggesting highly reliable clustering results. Post-hoc analysis revealed significant

differences between clusters in debt-to-income ratios (p ¡ 2.2e-16) and employment tenure patterns. The findings provide valuable insights for risk assessment and customer segmentation strategies in financial institutions, demonstrating the effectiveness of unsupervised learning techniques in identifying distinct customer profiles based on loan application characteristics.

Keywords: Clustering analysis, Customer segmentation, Financial data mining, Hierarchical clustering, K-means, Home equity loans

1 Introduction

In the contemporary financial landscape, effective customer segmentation has become paramount for risk assessment and strategic decision-making in lending institutions. The ability to identify distinct customer profiles based on loan application characteristics directly impacts profitability, risk management, and regulatory compliance. This study addresses the critical need for robust analytical methods to segment customers applying for home equity loans, utilizing advanced clustering techniques to uncover hidden patterns in customer behavior and characteristics.

Customer segmentation in financial services has evolved from simple demographic categorizations to sophisticated analytical approaches that leverage machine learning algorithms. Clustering analysis, as an unsupervised learning technique, offers the advantage of discovering natural groupings within data without predetermined categories, making it particularly suitable for exploratory data analysis in financial contexts.

The home equity loan market represents a significant segment of consumer lending, where borrowers use their home's equity as collateral. Understanding the characteristics that distinguish different applicant groups is crucial for lenders to optimize their approval processes, pricing strategies, and risk management frameworks. Traditional statistical approaches often fall short in capturing the complex, multi-dimensional relationships present in modern financial datasets.

This research contributes to the existing literature by providing a comprehensive comparison of two fundamental clustering algorithms—hierarchical clustering and K-means

clustering—applied to real-world financial data. The study's novelty lies in its systematic approach to data preprocessing, particularly the handling of mixed-type variables common in financial datasets, and the rigorous validation of clustering solutions through multiple evaluation metrics.

2 Literature Review

Clustering analysis has been extensively applied in financial services, with numerous studies demonstrating its effectiveness in customer segmentation and risk assessment. The application of unsupervised learning techniques to financial data has gained considerable attention due to their ability to reveal hidden patterns and structures that traditional analytical methods might overlook.

Hierarchical clustering, particularly Ward's linkage method, has been widely adopted in financial applications due to its ability to create meaningful dendrograms that facilitate interpretation of cluster structures. The method's agglomerative approach builds clusters by successively merging the most similar observations, creating a hierarchical tree structure that allows analysts to examine clustering solutions at different granularities.

K-means clustering, as a partitional clustering algorithm, has proven effective in scenarios requiring predefined numbers of clusters. Its computational efficiency and straightforward interpretation make it particularly attractive for large-scale financial datasets. However, the algorithm's sensitivity to initialization and assumption of spherical clusters can pose challenges in real-world applications.

The choice of distance metrics plays a crucial role in clustering financial data, especially when dealing with mixed-type variables common in customer datasets. The Gower distance metric has emerged as a robust solution for handling datasets containing both continuous and categorical variables, making it particularly suitable for financial applications where customer attributes encompass diverse data types.

3 Methodology

3.1 Dataset Description

The analysis utilized the Home Equity (HMEQ) dataset, comprising 5,960 observations with 13 variables related to home equity loan applications. The dataset includes both continuous variables (loan amount, mortgage due, property value) and categorical variables (reason for loan, job category), representing typical characteristics found in financial lending datasets.

The target variable, BAD, indicates loan default status, while predictor variables encompass financial metrics such as debt-to-income ratio (DEBTINC), years on job (YOJ), credit age (CLAGE), and demographic factors including job category (JOB) and loan purpose (REASON).

3.2 Data Preprocessing

3.2.1 Missing Value Analysis

Initial exploratory analysis revealed varying degrees of missingness across variables, with DEBTINC showing the highest missing rate at 21.26%, followed by DEROG at 11.88%. The systematic assessment of missing data patterns informed subsequent imputation strategies.

Missing values in categorical variables (REASON, JOB) were replaced with "Unknown" categories, preserving the information that these values were missing while maintaining the categorical structure necessary for clustering analysis.

3.2.2 Data Transformation

Logarithmic transformation was applied to highly skewed continuous variables (LOAN, MORTDUE, VALUE, YOJ, CLAGE) to normalize their distributions and reduce the impact of extreme values. Prior to transformation, variables with zero minimum values (YOJ, CLAGE) were adjusted by adding one to all observations to ensure mathematical

validity of the logarithmic operation.

3.2.3 Multiple Imputation

The Multiple Imputation by Chained Equations (MICE) algorithm was implemented to handle remaining missing values in continuous variables. The procedure utilized predictive mean matching (PMM) with 10 iterations and a single imputation to maintain computational efficiency while ensuring robust imputation quality.

3.2.4 Distance Matrix Computation

The Gower distance metric was computed using the `daisy()` function from the cluster package in R, providing a comprehensive distance matrix that appropriately handles the mixed-type nature of the dataset. This approach ensures that both continuous and categorical variables contribute meaningfully to the clustering process.

3.3 Clustering Algorithms

3.3.1 Hierarchical Clustering

Ward's linkage method was employed for hierarchical clustering, utilizing the computed Gower distance matrix. This method minimizes within-cluster variance at each merging step, typically producing compact, well-separated clusters suitable for interpretation.

3.3.2 K-means Clustering

K-means clustering was implemented using the same distance matrix, with cluster centers initialized through standard random initialization. The algorithm iteratively optimizes cluster assignments to minimize within-cluster sum of squares.

3.4 Optimal Cluster Determination

Multiple methods were employed to determine the optimal number of clusters:

- Silhouette analysis to measure cluster cohesion and separation

- Scree plots examining within-cluster sum of squares
- Dendrogram inspection for natural breaking points

3.5 Cluster Validation

The agreement between clustering methods was assessed using:

- Jaccard similarity index
- Rand index
- Cross-tabulation analysis

3.6 Visualization

t-Distributed Stochastic Neighbor Embedding (t-SNE) was employed to create two-dimensional visualizations of the high-dimensional clustering results, facilitating interpretation and validation of cluster structures.

4 Results

4.1 Missing Value Patterns

The initial analysis revealed significant missing data patterns across the dataset. Table 1 presents the missing value rates for each variable.

Table 1: Missing Value Rates by Variable

Variable	Missing Rate (%)
BAD	0.00
LOAN	0.00
MORTDUE	8.69
VALUE	1.88
REASON	4.23
JOB	4.68
YOJ	8.64
DEROG	11.88
DELINQ	9.73
CLAGE	5.17
NINQ	8.56
CLNO	3.72
DEBTINC	21.26

4.2 Optimal Cluster Determination

Both hierarchical clustering and K-means clustering consistently identified two clusters as the optimal solution. The silhouette analysis for both methods showed peak values at k=2, with silhouette widths exceeding 0.5, indicating strong cluster structure.

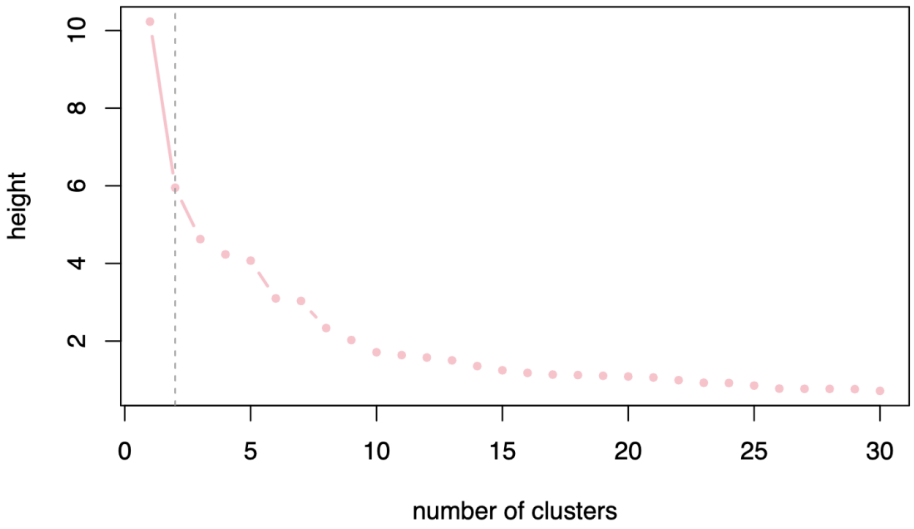

Figure 1: Scree Plot for Hierarchical Clustering showing optimal cluster determination. The plot demonstrates a clear elbow at k=2, indicating the optimal number of clusters.

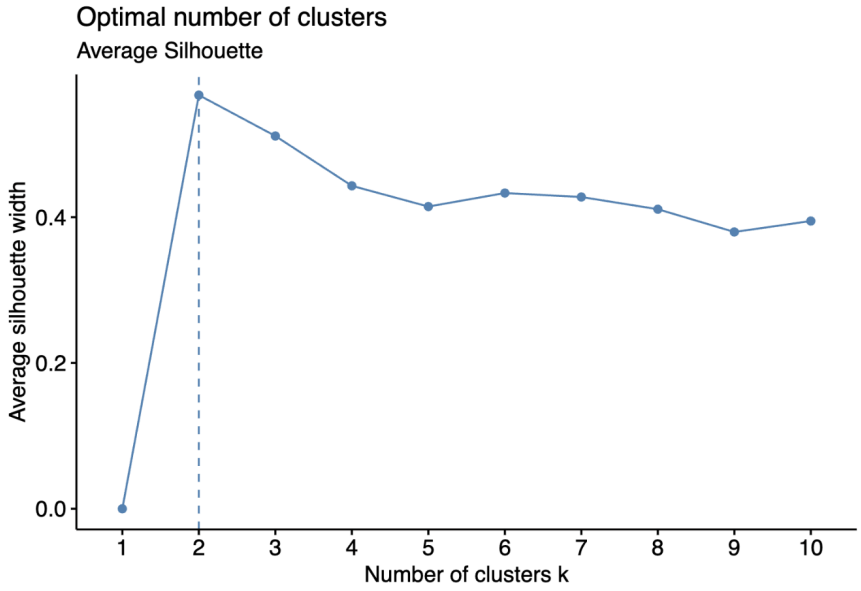

Figure 2: Silhouette Analysis for determining optimal number of clusters. The analysis shows peak silhouette width at k=2 for both clustering methods, confirming the two-cluster solution.

Scree plot analysis corroborated these findings, showing distinct elbows at the two-cluster solution for both algorithms (Figure 1). The silhouette analysis (Figure 2) further confirmed the optimal two-cluster solution with high silhouette coefficients. The dendrogram from hierarchical clustering revealed natural breaking points supporting the

two-cluster interpretation.

4.3 Clustering Results

4.3.1 Hierarchical Clustering

The hierarchical clustering using Ward's linkage produced two distinct clusters:

- Cluster 1: 2,032 observations (34.1%)
- Cluster 2: 3,928 observations (65.9%)

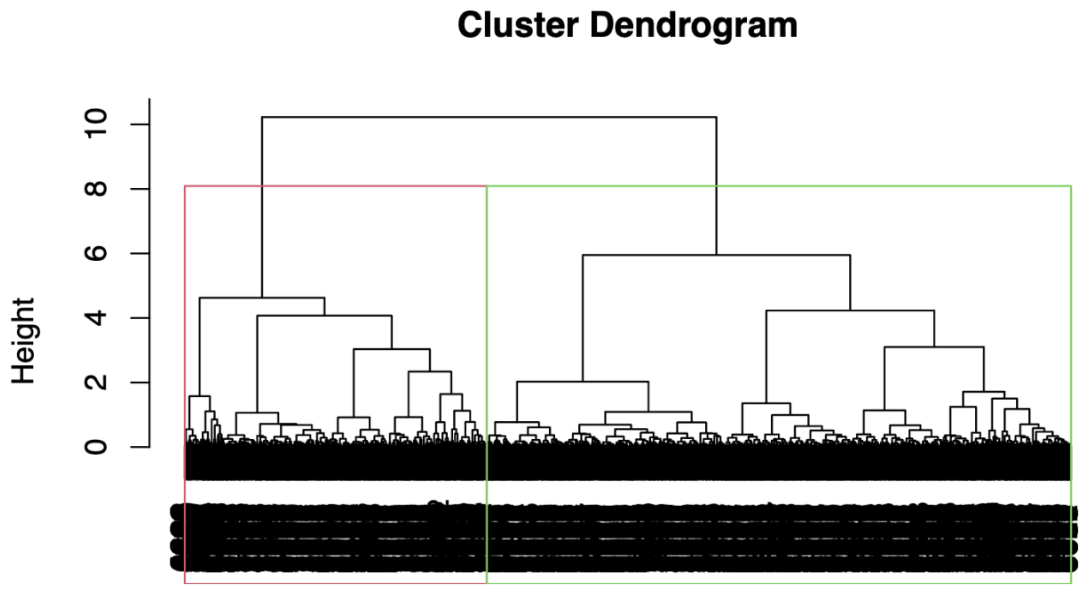

Figure 3: Dendrogram with Final Clusters. The dendrogram shows the hierarchical structure of the data with clear separation into two distinct clusters marked by colored rectangles.

4.3.2 K-means Clustering

K-means clustering yielded identical cluster sizes:

- Cluster 1: 3,928 observations (65.9%)
- Cluster 2: 2,032 observations (34.1%)

Note that while the cluster sizes are identical, the cluster labels are reversed between the two methods.

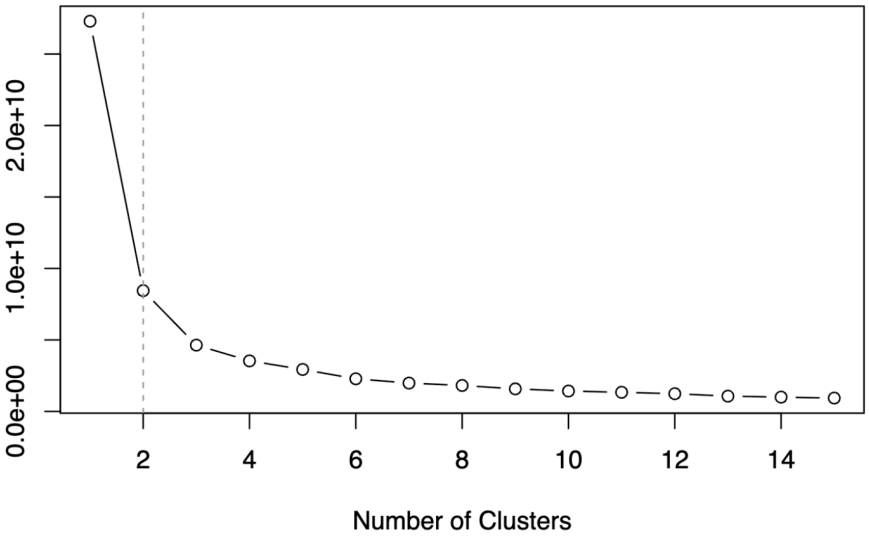

Figure 4: Within-Group Sum of Squares for K-means Clustering. The scree plot shows the elbow method for determining optimal number of clusters, with a clear break at k=2.

4.4 Cluster Validation

The agreement between clustering methods was exceptionally high:

- Jaccard similarity index: 1.0
- Rand index: 1.0

These perfect agreement scores indicate that both clustering algorithms identified identical groupings, providing strong evidence for the validity and stability of the clustering solution.

4.5 Post-hoc Analysis

4.5.1 Debt-to-Income Ratio Analysis

Statistical analysis revealed significant differences in debt-to-income ratios between clusters:

- Cluster 1 mean DEBTINC: 32.65
- Cluster 2 mean DEBTINC: 34.26
- Two-sample t-test: t = -6.89, p ¡ 2.2e-16

The F-test for equality of variances also indicated significantly different variances (F = 1.38, p ¡ 2.2e-16), suggesting distinct risk profiles between clusters.

4.5.2 Employment Tenure Patterns

Analysis of years on job (YOJ) revealed distinct employment patterns:

Table 2: Employment Tenure Distribution by Cluster

YOJ Category	Cluster 1	Cluster 2
$\leq$ 10 years	137	379
$>$ 10 years	1,895	3,549

Both clusters show majority employment tenure exceeding 10 years, but Cluster 1 has proportionally fewer individuals with shorter employment history.

4.5.3 Categorical Variable Analysis

Examination of job categories and loan reasons revealed distinct patterns:

Job Distribution: Cluster 2 showed higher concentrations of managers and sales personnel, while Cluster 1 had more balanced distributions across job categories.

Loan Purpose: A striking pattern emerged in loan purposes:

- Cluster 1: Primarily home improvement (1,780) and unknown reasons (252)
- Cluster 2: Exclusively debt consolidation (3,928)

This clear separation suggests that loan purpose serves as a primary discriminating factor in the clustering solution.

4.6 Visualization Results

t-SNE visualization confirmed the distinct separation between clusters, with minimal overlap and clear boundaries. The two-dimensional projection revealed compact, well-separated cluster structures that align with the quantitative validation metrics.

Figure 5: t-SNE Visualization for Hierarchical Clustering Results. The plot shows clear separation between the two clusters with minimal overlap, confirming the quality of the clustering solution.

Figure 6: t-SNE Visualization for K-means Clustering Results. Similar to the hierarchical clustering visualization, this plot demonstrates excellent cluster separation and validates the clustering quality.

5 Discussion

5.1 Clustering Performance

The exceptional agreement between hierarchical clustering and K-means clustering (Jaccard and Rand indices = 1.0) provides strong evidence for the robustness and reliability of the identified clustering solution. This level of agreement is rare in clustering applications and suggests that the two-cluster structure represents a fundamental characteristic of the underlying data.

The consistent identification of two clusters across multiple validation methods (silhouette analysis, scree plots, dendrogram inspection) further reinforces the validity of this solution. The high silhouette values (¿0.5) indicate strong within-cluster cohesion and between-cluster separation.

5.2 Business Implications

The clustering results reveal two distinct customer segments with significant practical implications for financial institutions:

Cluster 1 - Home Improvement Borrowers: This segment primarily seeks loans for home improvement purposes, demonstrates slightly lower debt-to-income ratios, and shows more diverse employment patterns. This group may represent lower-risk borrowers seeking to invest in property enhancement.

Cluster 2 - Debt Consolidation Borrowers: This larger segment exclusively pursues loans for debt consolidation, exhibits higher debt-to-income ratios, and includes more managers and sales professionals. This profile suggests borrowers experiencing financial stress who seek to restructure existing obligations.

These distinct profiles enable targeted risk assessment strategies, customized product offerings, and differentiated pricing models. The clear separation based on loan purpose (debt consolidation vs. home improvement) provides an immediately actionable segmentation criterion for business operations.

5.3 Risk Assessment Implications

The statistically significant difference in debt-to-income ratios between clusters (p ¡ 2.2e-16) has direct implications for credit risk assessment. The higher debt-to-income ratios in the debt consolidation cluster suggest elevated default risk, warranting more stringent underwriting criteria or risk-adjusted pricing.

The employment tenure patterns, while showing majority stability in both clusters, reveal subtle differences that may inform employment verification requirements and income stability assessments.

5.4 Methodological Considerations

The successful application of Gower distance for mixed-type data demonstrates the importance of appropriate distance metrics in financial clustering applications. The system-

atic handling of missing values through categorical replacement and MICE imputation preserved data integrity while maintaining analytical rigor.

The logarithmic transformation of skewed variables improved clustering performance by normalizing distributions and reducing the influence of extreme values, a common challenge in financial datasets characterized by wide dynamic ranges.

5.5 Limitations

Several limitations should be considered when interpreting these results:

1. The analysis excludes the target variable (BAD) from clustering, focusing solely on applicant characteristics rather than outcomes.

2. The dataset represents a specific time period and geographic region, potentially limiting generalizability.

3. The two-cluster solution, while statistically optimal, may oversimplify the complexity of customer heterogeneity.

4. Missing value imputation, while systematic, introduces uncertainty that may affect clustering stability.

6 Conclusion

This study successfully demonstrates the application of clustering analysis to home equity loan data, revealing two distinct customer segments with clear business implications. The exceptional agreement between hierarchical clustering and K-means clustering provides strong evidence for the reliability of the identified solution.

The research contributes to the financial analytics literature by providing a comprehensive methodology for customer segmentation in lending applications, particularly addressing the challenges of mixed-type data and missing values common in financial datasets.

Key findings include:

1. Consistent identification of two-cluster solution across multiple algorithms and validation methods
2. Perfect agreement between clustering methods (Jaccard and Rand indices = 1.0)
3. Statistically significant differences in debt-to-income ratios between clusters
4. Clear separation based on loan purpose (debt consolidation vs. home improvement)
5. Distinct employment and demographic patterns between segments

These insights provide actionable intelligence for financial institutions seeking to optimize their customer segmentation strategies, risk assessment procedures, and product development initiatives.

The methodology developed in this study provides a robust framework for customer segmentation that can be adapted to various financial analytics applications, contributing to the ongoing evolution of data-driven decision-making in financial services.

Acknowledgments

The author acknowledges the valuable insights provided by the data mining course instruction and the availability of the HMEQ dataset for educational and research purposes.

References

[1] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4.

[2] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.

[3] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857-871.

[4] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.

[5] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297).

[6] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

[7] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2579-2605.

A R Code Implementation

This appendix contains the complete R code used for the clustering analysis presented in this study.

A.1 Data Import and Initial Exploration

```
# Importing the data
dat0 <- read.csv("C:/Users/thomo/OneDrive/Desktop/Data Mining/HMEQ.csv")
dim(dat0)
head(dat0)

# The data (dat0) has 5960 observations with 13 variables

# Estimating the missing values rate
colMeans(is.na(dat0))
```

A.2 Data Cleaning and Preprocessing

A.2.1 Handling Missing Values in Categorical Variables

```
# Replacing 'NA' with 'Unknown'
dat0$REASON[which(is.na(dat0$REASON))] <- "Unknown"
dat0$JOB[which(is.na(dat0$JOB))] <- "Unknown"

# Verify the replacement
table(dat0$JOB, useNA = "ifany")
table(dat0$REASON, useNA = "ifany")
```

A.2.2 Natural Logarithm Transformation

```
# Check summary statistics for transformation variables
```

```
summary(dat0[, c("LOAN", "MORTDUE", "VALUE", "YOJ", "CLAGE")])

# Adding 1 to variables 'YOJ' and 'CLAGE' to handle zero values
dat0[, c(7, 10)] <- dat0[, c(7, 10)] + 1

# Verify the adjustment
summary(dat0[, c("LOAN", "MORTDUE", "VALUE", "YOJ", "CLAGE")])

# Taking natural log of the variables LOAN, VALUE, MORTDUE, YOJ, and CLAGE
dat0[, c(2, 3, 4, 7, 10)] <- apply(dat0[, c(2, 3, 4, 7, 10)], 2, FUN = log)
head(dat0[, c(2, 3, 4, 7, 10)])
```

A.2.3 Multiple Imputation Using MICE

```
library(dplyr)

# Exclude the target variable "BAD" for clustering
dat_1 <- dat0 %>% select(-BAD)
head(dat_1)

# Impute values for all missing values using the package MICE
library(mice)
fit.mice <- mice(dat_1, m=1, maxit=10, method = 'pmm', seed=100,
                 diagnostics = FALSE, remove_collinear = FALSE)

# Complete the imputation
data_imputed <- mice::complete(fit.mice, 1)
data_imputed <- model.matrix(~.-1, data=data_imputed)
dim(data_imputed)
```

```
# Check for any remaining missing values
anyNA(data_imputed)
colMeans(is.na(data_imputed))
```

A.3 Distance Matrix Computation

```
# Handling categorical variables
cols.cat <- c(1,2,3,4,5)
for (j in cols.cat) data_imputed[, j] <- as.factor(data_imputed[, j])
dat <- data.frame(data_imputed)

# Computing the distance matrix using daisy() with gower metric
library(cluster)
dismat <- daisy(dat, metric="gower", stand=TRUE)
```

A.4 Hierarchical Clustering

```
# Method 1: Hierarchical Clustering
fit.ward <- hclust(dismat, method = "ward.D2")
plot(fit.ward, hang=-0.5)

# Scree plot of height in hierarchical clustering
set.seed(5860)
K.max <- 30
height <- tail(fit.ward$height, n=K.max)
n.cluster <- tail((nrow(dat)-1):1, n=K.max)
plot(n.cluster, height, type="b", pch=19, cex=.5,
     xlab="number of clusters", ylab="height", col="pink", lwd=2)
abline(v=2, col="gray60", lty=2)

# Silhouette analysis for hierarchical clustering
```

```
set.seed(5860)
suppressMessages(library(factoextra))
fviz_nbclust(dat, kmeans, method = "silhouette") +
  labs(subtitle = "Average Silhouette")

# Dendrogram with final clusters
k.star <- 2
plot(fit.ward)
groups <- cutree(fit.ward, k=k.star)
rect.hclust(fit.ward, k=k.star, border=2:(k.star+1))

# Extract cluster groups
hclust.groups <- cutree(fit.ward, k=2)
table(hclust.groups)
```

A.5 t-SNE Visualization for Hierarchical Clustering

```
# Plotting dismat using tsne
library(Rtsne)
colors = rainbow(length(unique(hclust.groups)))
names(colors) = unique(hclust.groups)
set.seed(5860)
hclust.tsne <- Rtsne(dismat, dims=2, perplexity=30, max_iter=500)
plot(hclust.tsne$Y, t="n", main = "tSNE for Hierarchical Clustering")
text(hclust.tsne$Y, labels = hclust.groups, col = colors[hclust.groups])
```

A.6 K-means Clustering

```
# Method 2: K-Means Cluster Analysis
K <- 2
fit.kmeans <- kmeans(dismat, K) # K cluster solution
```

```
# Cluster memberships
kmeans.groups <- fit.kmeans$cluster
table(kmeans.groups)

# Scree plot for K-means clustering
library(cluster)
set.seed(5600)
wss <- (nrow(dat)-1)*sum(apply(dat,2,var))
K.max <- 15
for (K in 2:K.max) wss[K] <- sum(kmeans(dat, centers=K)$withinss)
plot(1:K.max, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")
abline(v=2, col="gray60", lty=2)

# Silhouette analysis for K-means
set.seed(5600)
suppressMessages(library(factoextra))
fviz_nbclust(dat, kmeans, method = "silhouette") +
  labs(subtitle = "Average Silhouette")
```

A.7 t-SNE Visualization for K-means

```
# tSNE for K Means
colors = rainbow(length(unique(kmeans.groups)))
names(colors) = unique(kmeans.groups)
set.seed(5860)
kmeans.tsne <- Rtsne(dismat, dims=2, perplexity=30, max_iter=500)
plot(kmeans.tsne$Y, t="n", main = "tSNE for Kmeans")
text(kmeans.tsne$Y, labels = kmeans.groups, col = colors[kmeans.groups])
```

A.8 Cluster Comparison and Validation

```
# Comparing Hierarchical Clustering and K means Clustering
library(clusteval)
jaccard <- cluster_similarity(hclust.groups, kmeans.groups,
                              similarity="jaccard", method="independence")
rand <- cluster_similarity(hclust.groups, kmeans.groups,
                           similarity = "rand")
matrix(c("Jaccard", jaccard, "Rand", rand), byrow = T, ncol = 2)
```

A.9 Post Hoc Analysis

```
# Post Hoc Analysis using hierarchical clustering results
dat <- data.frame(dat, dat0$BAD)
aggregate(dat[, c(1,2,3,6,9,12)], list(hclust.groups), mean, na.rm =T)

# Statistical tests for DEBTINC differences
cond1 <- hclust.groups == 1
cond2 <- hclust.groups == 2

# F test to compare variances
var.test(dat$DEBTINC[cond1], dat$DEBTINC[cond2],
         alternative = c("two.sided"))

# Two-sample t-test
t.test(dat$DEBTINC[cond1], dat$DEBTINC[cond2],
       alternative = c("two.sided"), var.equal = T)

# Analysis of YOJ (Years on Job)
dat_yoj <- table(dat$YOJ, hclust.groups)
```

```
dat_yoj <- as.data.frame(dat_yoj)
dat_yoj$Var1 <- as.numeric(dat_yoj$Var1)
dat_yoj$hclust.groups <- as.numeric(dat_yoj$hclust.groups)

cond1 <- (dat_yoj$Var1 <= 10) & (dat_yoj$hclust.groups == 1)
cond2 <- (dat_yoj$Var1 <= 10) & (dat_yoj$hclust.groups == 2)
lessq1 <- sum(dat_yoj$Freq[cond1])
lessq2 <- sum(dat_yoj$Freq[cond2])

cond11 <- (dat_yoj$Var1 > 10) & (dat_yoj$hclust.groups == 1)
cond22 <- (dat_yoj$Var1 > 10) & (dat_yoj$hclust.groups == 2)
grt1 <- sum(dat_yoj$Freq[cond11])
grt2 <- sum(dat_yoj$Freq[cond22])

matrix(c("YOJ", "Cluster 1", "Cluster 2",
         "<= 10", lessq1, lessq2,
         "> 10", grt1, grt2), byrow = T, ncol = 3)

# Categorical variable analysis
table(dat0$JOB, hclust.groups)
table(dat0$REASON, hclust.groups)
```