MONASH University

Individual Assignment 2:
Hierarchical Clustering for Student Stress

ETW2500 : Unsupervised Learning for Business

By Ng Chen Ting [31861148]

Table of Contents

Stage 1: Loading Libraries and Reading the CSV files

```r
# Load the necessary libraries
library(pacman)
p_load(tidyverse, factoextra, cluster, dendextend)
```

Firstly, the necessary libraries are loaded into R. The pacman package contains the function called p_load() which will be used to load needed missing packages. It is used to load tidyverse for data manipulation, factoextra for clustering visualization, cluster for cluster analysis and dendextend for dendrogram visualizations.

```r
# Load the csv file
student_data <- read.csv("Student_Data.csv")
student_data <- as.data.frame(student_data)
```

Once the libraries are ready, we proceed to load the student dataset containing information on students, stored in a CSV file. This dataset includes various demographic, academic, and extracurricular variables such as school, gender, age, family size, parental education, and more.

Stage 2: Variable Exploration

```r
str(student_data)
```

```
'data.frame':   382 obs. of  36 variables:
 $ school    : chr  "GP" "GP" "GP" "GP" ...
 $ sex       : chr  "F" "F" "F" "F" ...
 $ age       : int  15 15 15 15 15 15 15 15 15 15 ...
 $ address   : chr  "R" "R" "R" "R" ...
 $ famsize   : chr  "GT3" "GT3" "GT3" "GT3" ...
 $ Pstatus   : chr  "T" "T" "T" "T" ...
 $ Medu      : int  1 1 2 2 3 3 3 2 3 3 ...
 $ Fedu      : int  1 1 2 4 3 4 4 2 1 3 ...
 $ Mjob      : chr  "at_home" "other" "at_home" "services" ...
 $ Fjob      : chr  "other" "other" "other" "health" ...
 $ reason    : chr  "home" "reputation" "reputation" "course" ...
 $ nursery   : chr  "yes" "no" "yes" "yes" ...
 $ internet  : chr  "yes" "yes" "no" "yes" ...
 $ guardian  : chr  "mother" "mother" "mother" "mother" ...
 $ traveltime: int  2 1 1 1 2 1 2 2 2 1 ...
 $ studytime : int  4 2 1 3 3 3 3 2 4 4 ...
 $ failures  : int  1 2 0 0 2 0 2 0 0 0 ...
 $ schoolsup : chr  "yes" "yes" "yes" "yes" ...
 $ famsup    : chr  "yes" "yes" "yes" "yes" ...
 $ paid      : chr  "yes" "no" "yes" "yes" ...
 $ activities: chr  "yes" "no" "yes" "yes" ...
 $ higher    : chr  "yes" "yes" "yes" "yes" ...
 $ romantic  : chr  "no" "yes" "no" "no" ...
 $ famrel    : int  3 3 4 4 4 4 4 4 4 4 ...
 $ freetime  : int  1 3 3 3 2 3 2 1 4 3 ...
 $ goout     : int  2 4 1 2 1 2 2 3 2 3 ...
 $ Dalc      : int  1 2 1 1 2 1 2 1 2 1 ...
 $ Walc      : int  1 4 1 1 3 1 2 3 3 1 ...
 $ health    : int  1 5 2 5 3 5 5 4 3 4 ...
 $ absences  : int  2 2 8 2 8 2 0 2 12 10 ...
 $ G1.x      : int  7 8 14 10 10 12 12 8 16 10 ...
 $ G2.x      : int  10 6 13 9 10 12 0 9 16 11 ...
 $ G3.x      : int  10 5 13 8 10 11 0 8 16 11 ...
 $ G1.y      : int  13 13 14 10 13 11 10 11 15 10 ...
 $ G2.y      : int  13 11 13 11 13 12 11 10 15 10 ...
 $ G3.y      : int  13 11 12 10 13 12 12 11 15 10 ...
```

Figure 1.1: Data types of original dataset

This data has 36 variables, 19 of them are numeric and 17 of them are categorical. These variables can be grouped to construct new variables.

To cluster a student's current stress level, new variables like social perspective, well being and family background can be created as they can play a huge role in providing insights on a students' stress levels.

New Variable 1: Social Perspective

This is a measurement of a student's engagement with their social environment. Certain social factors including social exclusion have 3.87 times the odds of experiencing moderate to high stress compared to those with lower stress levels (Othman et al., 2019). Moreover, positive social interactions can alleviate stress as well. This highlights the impact that social interactions can have on stress levels.

Table 2.1: Variables used to construct Social Perspective

Variable Name	Variable Type	How it's used
goout	Ordinal	Used in **[Social Perspective]** where a high frequency of interaction with friends means a high level of social engagement.
activities	Binary	Used in **[Social Perspective]** where participation in extra-curricular activities means a high level of social engagement.

New Variable 2: Well-Being

This is a measurement of a students' lifestyle choices and overall mental and physical health which directly influences stress levels. There is a negative correlation between stress and self-rated health meaning when a student feels unhealthy, their stress levels are likely to rise. (Othman et al., 2019).

Table 2.2: Variables used to construct Well Being

Variable Name	Variable Type	How it's used
Walc	Ordinal	Used in **[Well Being]** where high alcohol consumption may correlate with certain lifestyle choices or stress management practices that can impact overall well-being.

Health	Ordinal	Used in **[Well Being]** which can reflect both physical and mental well-being.

New Variable 3: Family Background

A measurement of a student's family environment and background could also impact stress levels. Family dynamics and family support play a significant role in a student's emotional well-being and have been shown to affect stress levels. For instance, strong parental support and positive relationships helps students cope better with academic and social pressures.

Table 2.3: Variables used to construct Family Background

Variable Name	Variable Type	How it's used
FamSup	Binary	Used in **[Family Background]** where the presence of family support in education is linked to lower stress.
FamRel	Ordinal	Used in **[Family Background]** to assess family relationships quality. A higher rating indicates stronger family bonds which positively influence stress levels.

```
# Extract variables of interest to create Master Dataset
master_data <- data.frame(
                          # Social Perspective
                          student_data$goout,
                          student_data$activities,

                          # Well-being
                          student_data$Walc,
                          student_data$health,

                          # Family Background
                          student_data$famsup,
                          student_data$famrel)
```

All the variables of interest listed above are then extracted to create a Master Dataset.

```
> # Check for NA values
> any(is.na(master_data))
[1] FALSE
```

There are also no missing values found in the Master Dataset.

```r
# Randomization process for sample selection
set.seed(31861148)
master_sample <- master_data[sample(1:nrow(student_data), 200), ]
```

To ensure that the natural variability in the data is captured, a random process is employed for sample selection but by setting a seed, the random selection process becomes reproducible,

```r
# Change to boolean for Activities and Family Support
master_sample$student_data.activities <- ifelse(master_sample$student_data.activities == "yes"
,TRUE, FALSE)
master_sample$student_data.famsup <- ifelse(master_sample$student_data.famsup == "yes",TRUE,
FALSE)
```

Lastly, Activities and Family Support in the student_data dataset which were originally formatted as "yes" or "no" are being converted into logical binary variables (TRUE/FALSE). This conversion simplifies the data for later clustering analysis. "Yes" values are converted to TRUE, and "No" values are converted to FALSE.

Cluster Focus Variable: Stress

The stress variable integrates social perspective, well being and family background to provide a thorough assessment of student stress.

Stage 3: Variable Construction through Hierarchical

[Variable 1: Social Perspective]

```
# Select variables for Social Perspective
set.seed(31861148)
social_perspective <- data.frame(
  master_sample$student_data.goout,
  master_sample$student_data.activities
)

# Scaled and ensure it is in numerical format for "goout"
social_perspective$master_sample.student_data.goout <- as.numeric(scale
(social_perspective$master_sample.student_data.goout))

# Ensure it's in a dataframe format
social_perspective <- as.data.frame(social_perspective)
```

The variables of interest to Social Perspectives are first extracted from the Master Sample Dataset to create a new data frame. Variable “Go Out” is then scaled to prevent variables with larger ranges from dominating the results in clustering.

Elbow Method

```
# Elbow Method : Social Perspective
set.seed(31861148)
social_perspective_elbow <- fviz_nbclust(social_perspective, FUNcluster = hcut, method = "wss",
                            k.max = 6) +
  labs(title="A : Elbow Method [Social Perspective]")
social_perspective_elbow
```

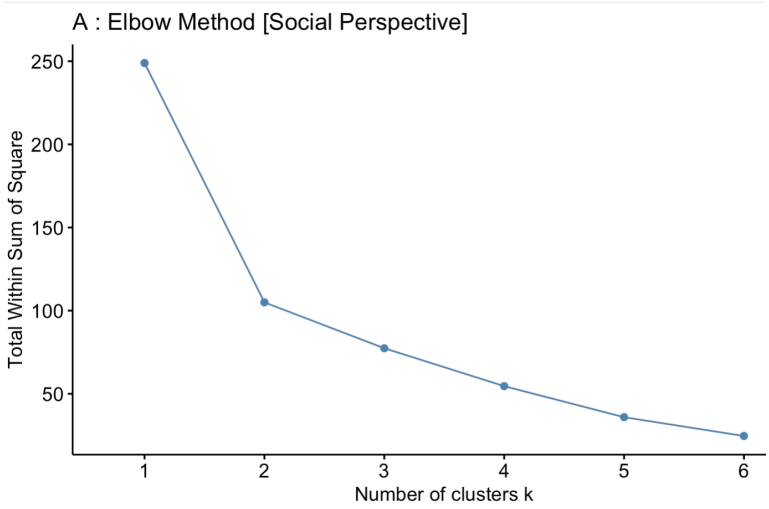

Figure 3.1: Elbow Method of Social Perspective

The elbow plots' maximum number of clusters is limited to 6 as more clusters would be impractical and difficult to interpret. There is a steep gradient between k = 2 and k = 3 and between k = 5 and k = 6, indicating that adding a third and sixth cluster offers diminishing returns.

Silhouette Method

```
# Silhouette Method : Social Perspective
set.seed(31861148)
social_perspective_silhouette <- fviz_nbclust(social_perspective, FUNcluster = hcut, method =
"silhouette",
                                k.max = 6) +
  labs(title="A : Silhouette Method [Social Perspective]")
social_perspective_silhouette
```

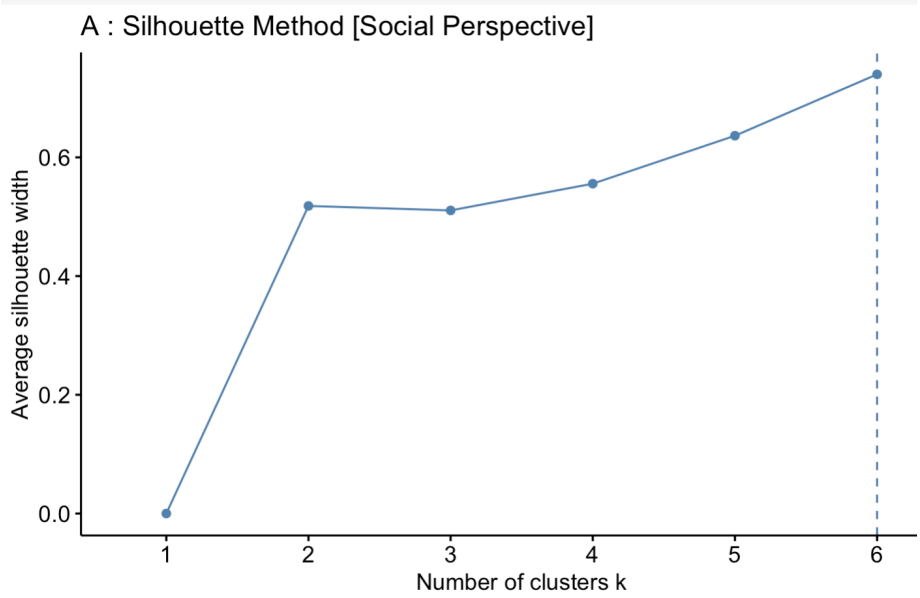

Figure 3.2: Silhouette Method of Social Perspective

From the silhouette method, it was observed that both k = 5 and k = 6 yielded close and good results in terms of silhouette values. Given this similarity in both plots and to maintain a simpler clustering structure, k = 5 was selected as the optimal number of clusters.

Clustering using different Hierarchical Methods

```
# Set binary variables as factors
social_perspective$master_sample.student_data.activities <- as.factor(social_perspective$master_sample.student_data.activities)

# Compute Gower distance
gower_dist <- daisy(social_perspective, metric = "gower")

# Perform hierarchical clustering - Complete
social_perspective_complete <- hclust(gower_dist, method = "complete")
plot(color_branches(as.dendrogram(social_perspective_complete), k=5), main = "Complete Method with K=5")

# Perform hierarchical clustering - Average
social_perspective_average <- hclust(gower_dist, method = "average")
plot(color_branches(as.dendrogram(social_perspective_average), k=5), main = "Average Method with K=5")

# Perform hierarchical clustering - Ward
social_perspective_ward <- hclust(gower_dist, method = "ward.D2")
plot(color_branches(as.dendrogram(social_perspective_ward), k=5), main = "Ward Method with K=5")
```

The dendrograms illustrating the hierarchical clustering can be referred to in the appendix which provides insights into the clustering process using Ward Method, Average Method and Complete Method.

Check Count

```
# Check Count - Complete
set.seed(31861148)
social_perspective_cluster_complete_k5 <- cutree(social_perspective_complete, k=5)
social_perspective_complete_k5 <- master_sample %>% mutate(cluster1_ck5 = social_perspective_cluster_complete_k5)

# Check Count - Average
set.seed(31861148)
social_perspective_cluster_average_k5 <- cutree(social_perspective_average, k=5)
social_perspective_average_k5 <- master_sample %>% mutate(cluster1_ak5 = social_perspective_cluster_average_k5)

# Check Count - Ward
set.seed(31861148)
social_perspective_cluster_ward_k5 <- cutree(social_perspective_ward, k=5)
social_perspective_ward_k5 <- master_sample %>% mutate(cluster1_wk5 = social_perspective_cluster_ward_k5)
```

Table 3.1: Cluster sizes of the three hierarchical clustering methods

```
> social_perspective_complete_k5 %>%  count(cluster1_ck5)
  cluster1_ck5  n
1            1 31
2            2 65
3            3 13
4            4 57
5            5 34
> social_perspective_average_k5 %>%  count(cluster1_ak5)
  cluster1_ak5  n
1            1 31
2            2 16
3            3 49
4            4 70
5            5 34
> social_perspective_ward_k5 %>%  count(cluster1_wk5)
  cluster1_wk5  n
1            1 61
2            2 35
3            3 39
4            4 31
5            5 34
```

To better measure the size of each cluster, the dendrograms were cut at k = 5 to create five distinct clusters. It was observed that the Ward method produces the most even split across the clusters, distributing the observations more uniformly.

Whereas in contrast, the Complete and Average method creates imbalanced clusters, with some clusters having significantly more observations than others. This suggests that the Ward method is the most desirable as balanced group sizes.

Check Mean

Table 3.2: Cluster values of each variable in Social Perspective

```
# Count the mean of each variable in each cluster
social_perspective_ward_k5 %>%
  group_by(cluster1_wk5) %>%
  summarise(across(c( "student_data.goout", "student_data.activities"),
                   ~round(mean(as.numeric(.x), na.rm = TRUE),0)))
```

A tibble: 5 × 3

cluster1_wk5 <int>	student_data.goout <dbl>	student_data.acti... <dbl>
1	2	0
2	4	0
3	4	1
4	3	1
5	2	1

Interpretation:

Cluster 1 - Homebody : These students rarely go out with friends and do not participate in extracurricular activities. They indicate a preference for solitude over social engagement, which may limit their opportunities for building social networks. These students likely have limited social interaction, which could result in a more isolated social perspective.

Cluster 2 - Social butterfly : These are students who go out with friends frequently but do not engage in extracurricular activities. This suggests a highly social group in casual, peer-based environments. They hold a higher social perspective, especially if their frequent socializing outside of school allows them to engage with a larger, more diverse community.

Cluster 3 - Socially Active Leaders : These students actively participate in extracurriculars and are extremely popular among their peers. These students hold the strongest social position, as they balance formal and informal social interactions. They demonstrate a strong connection to their school community and most likely hold leadership roles, which can inspire and influence their peers positively leading to an excellent social perspective.

Cluster 4 - Active Moderates: This balanced group likely has a well-rounded social life, engaging in both formal activities and casual social outings. Their social perspective is likely more diverse with moderate peer-to-peer interaction.

Cluster 5 - Reserved Participants : Although these students are involved in structured activities, their limited informal socializing may result in a more reserved or selective social perspective. They likely develop strong bonds within formal group settings but have fewer spontaneous social interactions with peers showing a preference for quieter social interactions.

[Variable 2: Well Being]

```
# Select variables for Well Being
set.seed(31861148)
well_being <- data.frame(
  master_sample$student_data.health,
  master_sample$student_data.Walc
)

# Scaling
well_being <- scale(well_being)
well_being <- as.data.frame(well_being)
```

Similarly, the two variables are extracted from the Master Dataset and scaled to create the “Well Being” dataset.

Elbow Method

```
# Elbow Method : Well Being
set.seed(31861148)
well_being_elbow <- fviz_nbclust(well_being, FUNcluster = hcut, method = "wss",
                                 k.max = 6) +
  labs(title="A : Elbow Method [Well Being]")
well_being_elbow
```

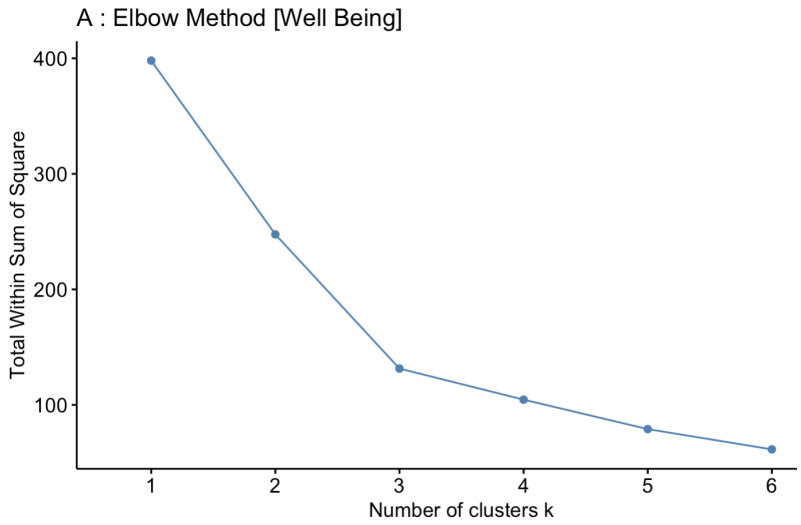

Figure 3.3: Elbow Method of Well Being

The elbow plot shows a steep gradient between k = 3 and k = 4, indicating that adding a fourth cluster offers diminishing returns.

Silhouette Method

```
# Silhouette Method : Well Being
set.seed(31861148)
well_being_silhouette <- fviz_nbclust(well_being, FUNcluster = hcut, method = "silhouette",
                                      k.max = 6) +
  labs(title="A : Silhouette Method [Well Being]")
well_being_silhouette
```

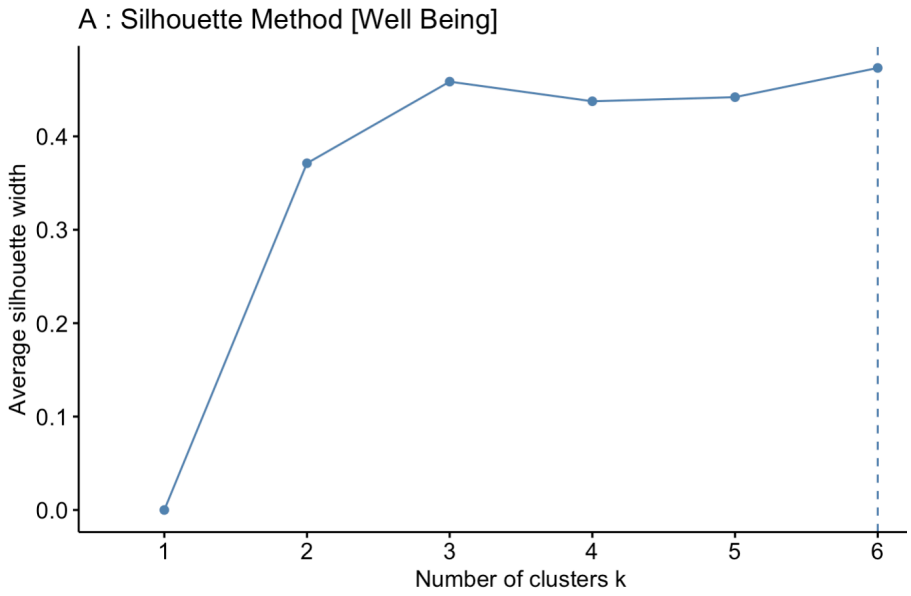

Figure 3.4: Silhouette Method of Well Being

From the silhouette method, it can also be observed that k = 3 is the second highest peak making it the optimal number of clusters.

Clustering using different Hierarchical Methods

```
# Perform hierarchical clustering - Complete
set.seed(31861148)
well_being_complete <- hclust(dist(well_being), method = "complete")
plot(color_branches(as.dendrogram(social_perspective_complete), k=3), main = "Complete Method with K=3")

# Perform hierarchical clustering - Average
set.seed(31861148)
well_being_average <- hclust(dist(well_being), method = "average")
plot(color_branches(as.dendrogram(social_perspective_average), k=3), main = "Average Method with K=3")

# Perform hierarchical clustering - Ward
set.seed(31861148)
well_being_ward <- hclust(dist(well_being), method = "ward.D")
plot(color_branches(as.dendrogram(social_perspective_ward), k=3), main = "Ward Method with K=3")
```

The dendrograms can be observed in the appendix.

When compared to the sizes of other clusters, having 3 clusters provides the best split. This size effectively distinguishes a specific subset of students who prioritize their well-being, allowing for a clearer understanding of their characteristics.

Check Count

```
# Check count - Complete
set.seed(31861148)
well_being_cluster_complete_k3 <- cutree(well_being_complete, k=3)
well_being_complete_k3 <- master_sample %>% mutate(cluster2_ck3 = well_being_cluster_complete_k3)

# Check count - Average
set.seed(31861148)
well_being_cluster_average_k3 <- cutree(well_being_average, k=3)
well_being_average_k3 <- master_sample %>% mutate(cluster2_ak3 = well_being_cluster_average_k3)

# Check count - Ward
set.seed(31861148)
well_being_cluster_ward_k3 <- cutree(well_being_ward, k=3)
well_being_ward_k3 <- master_sample %>% mutate(cluster2_wk3 = well_being_cluster_ward_k3)
```

Table 3.3: Cluster sizes of the three hierarchical clustering methods in Well Being

```
> well_being_complete_k3 %>%  count(cluster2_ck3)
  cluster2_ck3   n
1            1  37
2            2  51
3            3 112
> well_being_average_k3 %>%  count(cluster2_ak3)
  cluster2_ak3  n
1            1 88
2            2 31
3            3 81
> well_being_ward_k3 %>%  count(cluster2_wk3)
  cluster2_wk3  n
1            1 84
2            2 31
3            3 85
```

While both the Ward and Average methods produced similar cluster ranges, Ward is slightly better in terms of cluster sizes making it the preferred choice. On the other hand, the Complete method performed poorly, resulting in highly uneven cluster sizes and less distinguishable groupings. Therefore, the Ward method was chosen for the final clustering solution.

Check Mean

Table 3.4: Cluster values of each variable in Well Being

```
# Count the mean of each variable in each cluster
well_being_ward_k3 %>%
  group_by(cluster2_wk3) %>%
  summarise(across(c("student_data.health","student_data.Walc"),
                   ~round(mean(as.numeric(.x), na.rm = TRUE), 0)))
```

A tibble: 3 × 3

cluster2_wk3 <int>	student_data.hea... <dbl>	student_data.Walc <dbl>
1	2	2
2	5	4
3	4	2

Interpretation:

Cluster 1: Stressed Drinkers - Students in this cluster displayed poor self-rated health while maintaining low levels of alcohol consumption. This combination may suggest that despite their minimal alcohol intake, underlying factors such as unhealthy lifestyle choices, or other stressors are contributing to their health issues.

Cluster 2: Partying Health Enthusiasts - These students feel very healthy but indulge in high levels of weekend alcohol consumption. They balance a high-energy lifestyle, excelling in health. While their physical health may seem favorable, heavy drinking can lead to negative impacts on mental well-being.

Cluster 3 Extremely Healthy Individuals - This group reports high health and keeps their alcohol intake low. They prioritize their well-being by focusing on fitness or wellness, and tend to limit alcohol consumption as part of a healthier lifestyle.

[Variable 3: Family Background]

```
# Select variables for Family Background
set.seed(31861148)
family_background <- data.frame(
  master_sample$student_data.famrel,
  master_sample$student_data.famsup
)

family_background$master_sample.student_data.famrel <- as.numeric(scale
(family_background$master_sample.student_data.famrel))

family_background <- as.data.frame(family_background)
```

Similarly, the variables are extracted from the Master Dataset and scaled for “Family Background” dataset.

Elbow Method

```
# Elbow Method : Family Background
set.seed(31861148)
family_background_elbow <- fviz_nbclust(family_background, FUNcluster = hcut, method = "wss",
                           k.max = 6) +
  labs(title="A : Elbow Method [Family Background]")
family_background_elbow
```

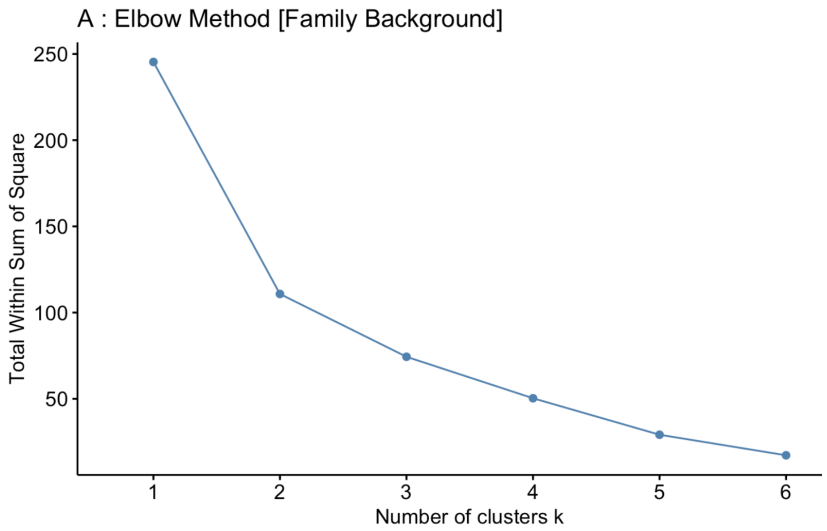

Figure 3.5: Elbow Method of Family Background

The resulting elbow plot shows two elbow bends between k = 2 and k = 3 and between k = 5 and k = 6, indicating that selecting the third and sixth cluster is not a great clustering choice.

Silhouette Method

```
# Silhouette Method : Family Background
set.seed(31861148)
family_background_silhouette <- fviz_nbclust(family_background, FUNcluster = hcut, method =
"silhouette",
                                k.max = 6) +
  labs(title="A : Silhouette Method [Family Background]")
family_background_silhouette
```

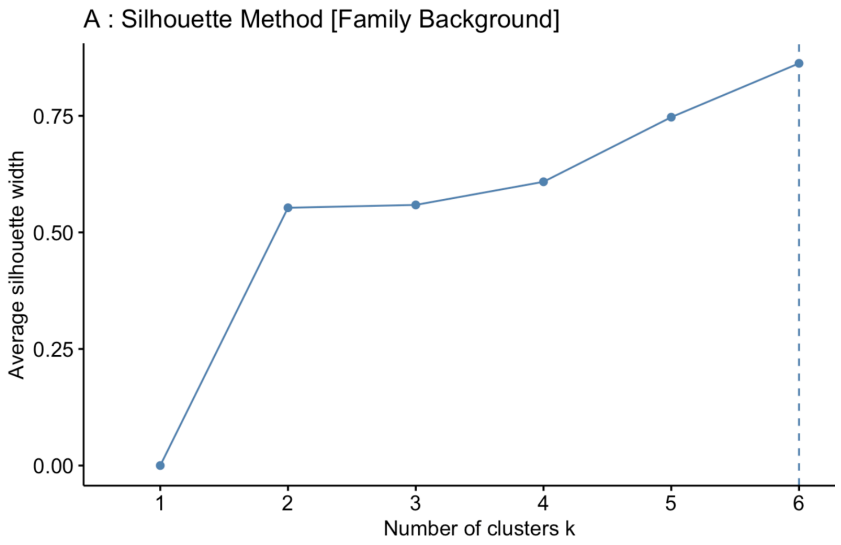

Figure 3.6: Silhouette Method of Family Background

In the silhouette method, the average silhouette width values continuously increase, making k = 5 the second highest and together with its elbow bend reasoning, it is the best choice to avoid creating too many clusters.

Clustering using different Hierarchical Methods

```
# Set binary variables as factors
family_background$master_sample.student_data.famsup <- as.factor(family_background$master_sample.student_data.famsup)

# Compute Gower distance
gower_dist <- daisy(family_background, metric = "gower")

# Perform hierarchical clustering - Complete
family_background_complete <- hclust(gower_dist, method = "complete")
plot(color_branches(as.dendrogram(family_background_complete), k=5), main = "Complete Method with K=5")

# Perform hierarchical clustering - Average
family_background_average <- hclust(gower_dist, method = "average")
plot(color_branches(as.dendrogram(family_background_average), k=5), main = "Average Method with K=5")

# Perform hierarchical clustering - Ward
family_background_ward <- hclust(gower_dist, method = "ward.D2")
plot(color_branches(as.dendrogram(family_background_ward), k=5), main = "Ward Method with K=5")
```

Dendrograms can be referred to in the appendix.

Check Count

```
# Check count - Complete
set.seed(31861148)
family_background_cluster_complete_k5 <- cutree(family_background_complete, k=5)
family_background_complete_k5 <- master_sample %>% mutate(cluster3_ck5 = family_background_cluster_complete_k5)

# Check count - Average
set.seed(31861148)
family_background_cluster_average_k5 <- cutree(family_background_average, k=5)
family_background_average_k5 <- master_sample %>% mutate(cluster3_ak5 = family_background_cluster_average_k5)

# Check count - Ward
set.seed(31861148)
family_background_cluster_ward_k5 <- cutree(family_background_ward, k=5)
family_background_ward_k5 <- master_sample %>% mutate(cluster3_wk5 = family_background_cluster_ward_k5)
```

Table 3.5: Cluster sizes of the three hierarchical clustering methods in Family Background

```
> family_background_complete_k5 %>%  count(cluster3_ck5)
  cluster3_ck5  n
1            1 91
2            2 21
3            3 52
4            4 33
5            5  3
> family_background_average_k5 %>%  count(cluster3_ak5)
  cluster3_ak5  n
1            1 91
2            2 70
3            3 33
4            4  3
5            5  3
> family_background_ward_k5 %>%  count(cluster3_wk5)
  cluster3_wk5  n
1            1 59
2            2 21
3            3 32
4            4 52
5            5 36
```

Using the Ward method k = 5 provides the best split when compared to the cluster sizes obtained through the complete and average method as it balances the distinctiveness of the clusters.

Complete method shows an obvious minority within the clusters, with sizes ranging dramatically from 3 to 91. This imbalance highlights the limitations of the complete method, as it may lead to clusters that are not as meaningful or easy to interpret.

Additionally, the Average method performs even worse, yielding highly uneven and poorly defined clusters, which makes it unsuitable for analysis. Therefore, the Ward method is chosen for its ability to produce more evenly distributed clusters.

Check Mean

Table 3.6: Cluster values of each variable in Family Background

```
family_background_ward_k5 %>%
  group_by(cluster3_wk5) %>%
  summarise(across(c("student_data.famrel","student_data.famsup"),
                   ~round(mean(as.numeric(.x), na.rm = TRUE), 0)))
```

A tibble: 5 × 3

cluster3_wk5 <int>	student_data.fam... <dbl>	student_data.fam... <dbl>
1	4	1
2	3	0
3	5	1
4	4	0
5	3	1

Interpretation:

Cluster 1: Supported Harmonizers - These students have strong family relationships and receive family support in their education. With a good family background, they likely benefit from a balanced, encouraging home environment.

Cluster 2: Struggling Independents - With lower family relationships and no family educational support, these students come from a distant family background. The lack of presence and support may contribute to feelings of isolation and stress, making them more reliant on themselves.

Cluster 3: Ideal Support System - Students in this cluster have perfect family relationships and strong educational support from their families. They have a nurturing and academically supportive home environment.

Cluster 4: Self-Starters - These students have decent family relationships but no family educational support. While they have good personal connections at home, managing academic responsibilities independently may bring stress.

Cluster 5: Supported but Strained - Their family background creates a dynamic where academic encouragement exists alongside emotional challenges with strained family relationships.

Stage 4: Final Dataset

Table 4.1: Display of the Final Dataset first 4 rows

```
> final <- data.frame(social_perspective_ward_k5$cluster1_wk5, well_being_ward_k3$cluster2_wk3, family_background_ward_k5$cluster3_wk5)
> final
  social_perspective_ward_k5.cluster1_wk5 well_being_ward_k3.cluster2_wk3 family_background_ward_k5.cluster3_wk5
1                                       1                               1                                     1
2                                       2                               2                                     2
3                                       1                               1                                     3
4                                       2                               3                                     4
```

In this step, we are creating the final dataset for stress clustering by combining the clusters derived from three different aspects: social perspective, well-being, and family background. The resulting data frame, final, includes the classifications from the social perspective cluster (k = 5), the well-being cluster (k = 3), and the family background cluster (k = 5). This will enable us to analyze their impact on student stress levels.

Stage 5: Hierarchical Clustering for Stress Level

Elbow Method

```
set.seed(31861148)
final_elbow <- fviz_nbclust(final, FUNcluster = hcut, method = "wss",
                    k.max = 6) +
  labs(title="Elbow method for Stress Level")
final_elbow
```

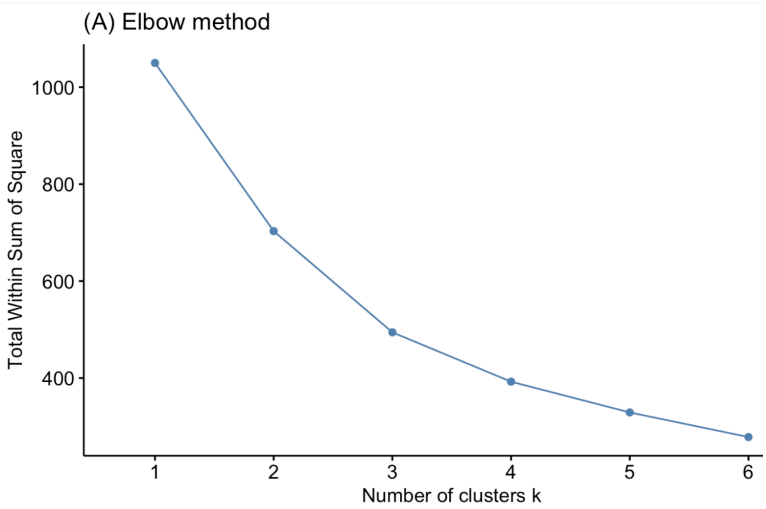

Figure 5.1: Elbow Method of Stress Levels

There is a noticeable bend between k = 3 and k = 4. This suggests that a three-cluster solution offers the most meaningful distinctions within the data.

Silhouette Method

```
set.seed(31861148)
final_silhouette <- fviz_nbclust(final, FUNcluster = hcut, method = "silhouette",
                    k.max = 6) +
  labs(title="Silhouette method for Stress Level")
final_silhouette
```

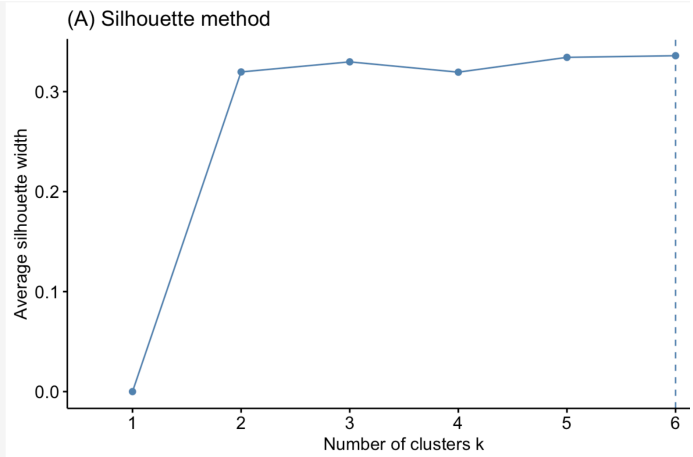

Figure 5.2: Silhouette Method of Stress Levels

Additionally, the silhouette analysis clearly indicates that k = 3 represents the initial peak for cluster differentiation further supporting the decision to select k = 3.

Clustering using different Hierarchical Methods

```
# Perform hierarchical clustering - Complete
set.seed(31861148)
final_hclus_complete <- hclust(daisy(final,metric="gower"), method = "complete")
plot(color_branches(as.dendrogram(final_hclus_complete), k=3), main = "Complete Method with K=3")

# Perform hierarchical clustering - Average
set.seed(31861148)
final_hclus_average <- hclust(daisy(final,metric="gower"), method = "average")
plot(color_branches(as.dendrogram(final_hclus_average), k=3), main = "Average Method with K=3")

# Perform hierarchical clustering - Ward
set.seed(31861148)
final_hclus_ward <- hclust(daisy(final,metric="gower"), method = "ward.D")
plot(color_branches(as.dendrogram(final_hclus_ward), k=3), main = "Ward Method with K=3")
```

Check Count

```
# Check count - Complete
set.seed(31861148)
final_cluster_comp_k3 <- cutree(final_hclus_complete, k=3)
final_data_comp_k3 <- final %>% mutate(clusterf_ck3 = final_cluster_comp_k3)

# Check count - Average
final_cluster_average_k3 <- cutree(final_hclus_average, k=3)
final_data_average_k3 <- final %>% mutate(clusterf_ak3 = final_cluster_ward_k3)

# Check count - Ward
final_cluster_ward_k3 <- cutree(final_hclus_ward, k=3)
final_data_ward_k3 <- final %>% mutate(clusterf_wk3 = final_cluster_ward_k3)
```

Table 5.1: Cluster sizes of the three hierarchical clustering methods in Stress Levels

```
> final_data_comp_k3 %>%  count(clusterf_ck3)
  clusterf_ck3  n
1            1 84
2            2 30
3            3 86
> final_data_average_k3 %>%  count(clusterf_ak3)
  clusterf_ak3  n
1            1 84
2            2 79
3            3 37
> final_data_ward_k3 %>%  count(clusterf_wk3)
  clusterf_wk3  n
1            1 84
2            2 79
3            3 37
```

Although the values of the clusters are similar for both methods, the complete method provides a more evenly spread distribution of cluster sizes. Therefore, the complete method is selected for the final clustering.

Additionally, it is more appropriate to use the complete linkage method for clustering for categorical data. As it is designed to handle categorical variables by focusing on maximizing the distance between clusters this helps ensure that distinct groups are formed based on categorical characteristics.

Check Mean

Table 5.2: Cluster values of each variable in Stress Levels

```
get_mode <- function(x) {
  uniq_x <- unique(x)
  freq_table <- table(x)
  modes <- names(freq_table[freq_table == max(freq_table)])
  return(as.numeric(modes))
}

# Reframe to get mode for each variable
summary_data <- final_data_comp_k3 %>%
  group_by(clusterf_ck3) %>%
  reframe(across(everything(), ~ get_mode(.x), .names = "mode_{.col}"))

# Rename columns
colnames(summary_data) <- c(
  "Cluster",
  "Social_Perspective",
  "Well_Being",
  "Family_Background")
summary_data
```

A tibble: 3 × 4

Cluster <int>	Social_Perspective <dbl>	Well_Being <dbl>	Family_Background <dbl>
1	1	1	1
2	3	2	4
3	1	3	1

Interpretation:

Cluster 1: Most Stressed - This cluster reflects students who exhibit low social perspective and poor well-being, relying on their strong family relationships for support. When stress levels are excessively high, this can have negative academic and emotional repercussions, which may lead to the adoption of multiple unhealthy behaviors (Pitzer and Skinner, 2016). Despite their familial backing, their limited social interactions and health challenges may contribute to heightened stress levels.

Cluster 2: Least Stressed - This cluster is characterized by students with a high social perspective and an active partying lifestyle, all while maintaining strong family relations. These students experience the lowest levels of stress among the clusters, as their social engagement and independence serve as effective buffers against stressors. Although independent in educational support, the combination of a supportive family environment and an active social life fosters resilience, enabling these students to navigate challenges more effectively.

Cluster 3: Moderately Stressed - Students in this cluster prioritize their health and enjoy strong family support. The dynamics of the parent-child relationship and parental aspirations for achievement are key factors that can influence academic stress in these students (Pui & Chen, 2022). Their focus on well-being helps mitigate some stress, but their more reserved social interactions may lead to occasional feelings of isolation.

Stage 6: Conclusion

In analyzing the stress levels of students based on their social behavior, extracurricular involvement, and family dynamics, three clusters are observed.

The "Most Stressed" cluster, characterized by low social interaction and poor health, highlights the need for healthier behaviors. In contrast, the "Least Stressed" cluster benefits from high social engagement and strong family relationships, as these factors serve as effective buffers against stress. Lastly, the "Moderately Stressed" group reflects the connection between health consciousness and family dynamics, showing the importance of fostering social connections to alleviate feelings of isolation.

References

Othman, N., Ahmad, F., Christo El Morr, & Ritvo, P. (2019). Perceived impact of contextual determinants on depression, anxiety and stress: a survey with university students. *International Journal of Mental Health Systems*, *13*(1). https://doi.org/10.1186/s13033-019-0275-x

Pui, E., & Chen, J.-K. (2022). The Correlates of Academic Stress in Hong Kong. International Journal of Environmental Research and Public Health, 19(7), 4009–4009. https://doi.org/10.3390/ijerph19074009

Pitzer, J., and E. Skinner. 2016. “Predictors of Changes in Students’ Motivational Resilience over the School Year: The Roles of Teacher Support, Self-Appraisals, and Emotional Reactivity.” International Journal of Behavioral Development 1–5. doi:10.1177/0165025416642051.

Appendix

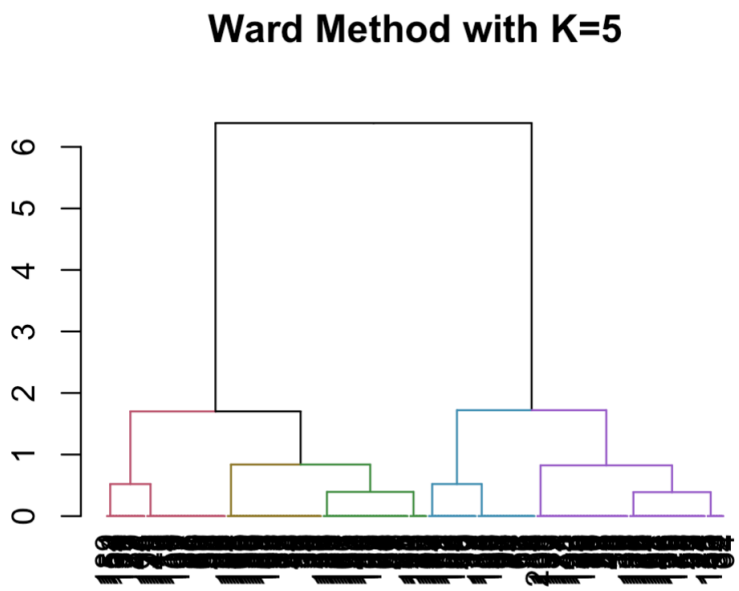

Figure 1: Social Perspective - Dendrogram for Ward Method

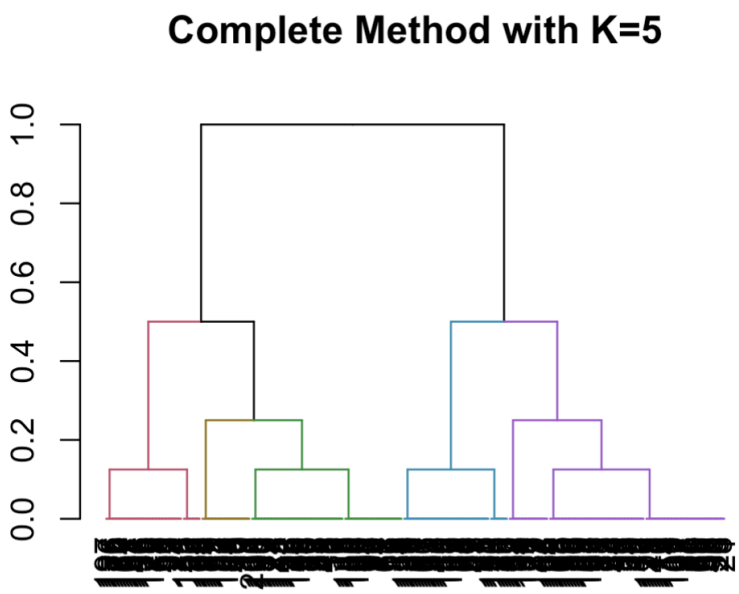

Figure 2: Social Perspective - Dendrogram for Complete Method

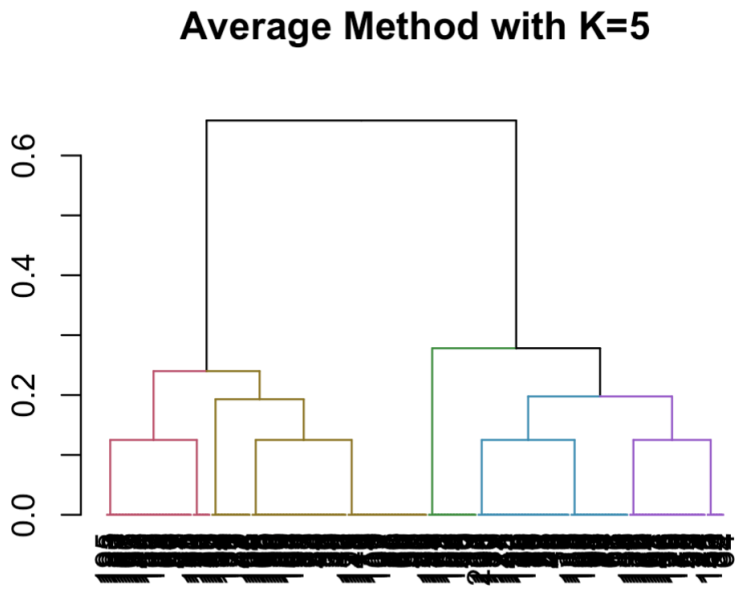

Figure 3: Social Perspective - Dendrogram for Average Method

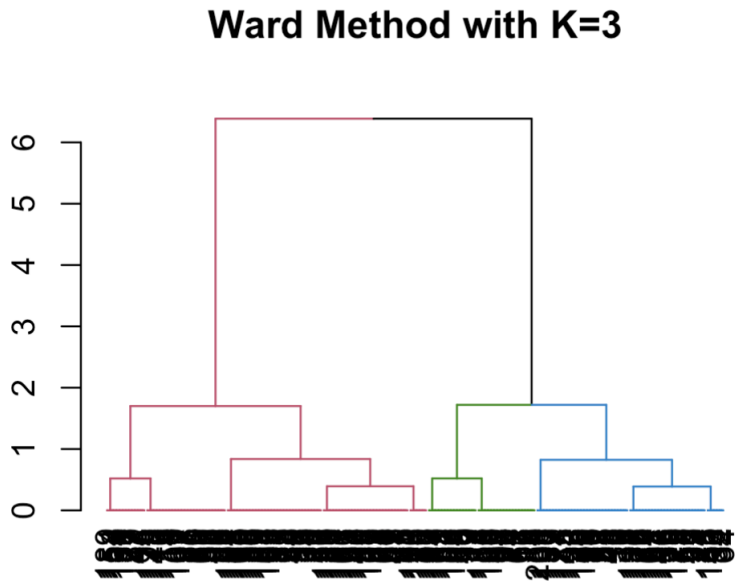

Figure 4: Well Being - Dendrogram for Ward Method

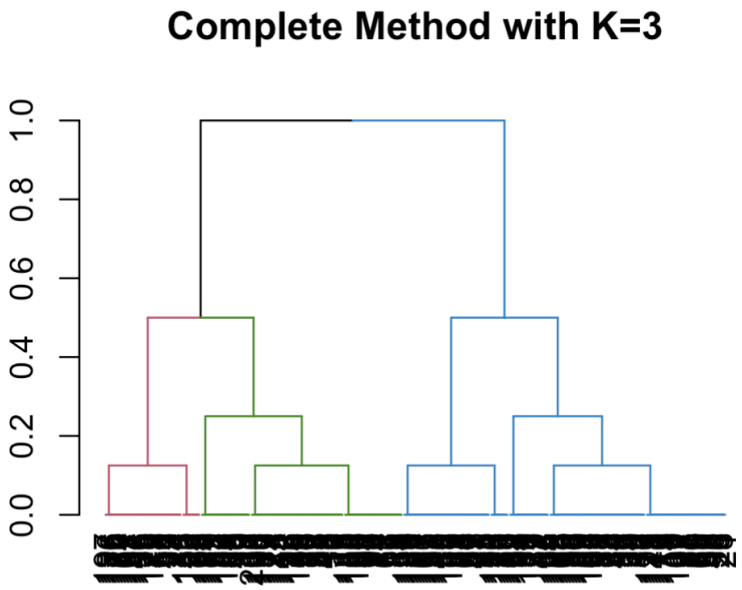

Figure 5: Well Being - Dendrogram for Complete Method

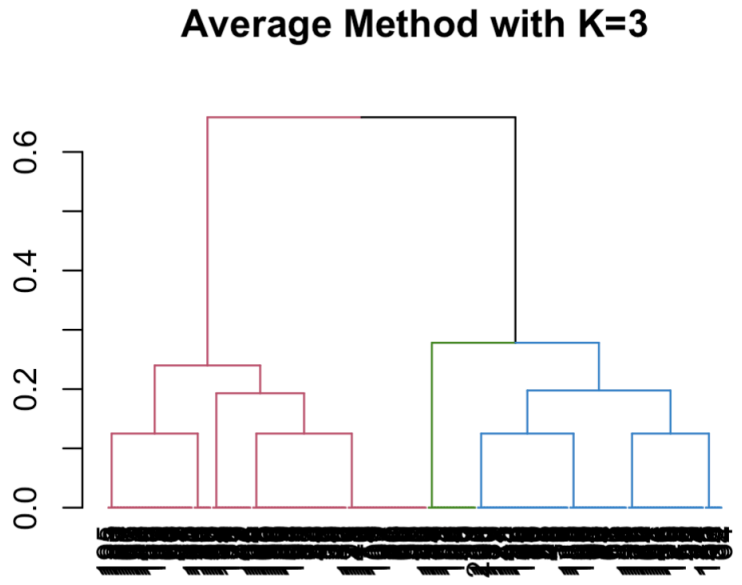

Figure 6: Well Being - Dendrogram for Average Method

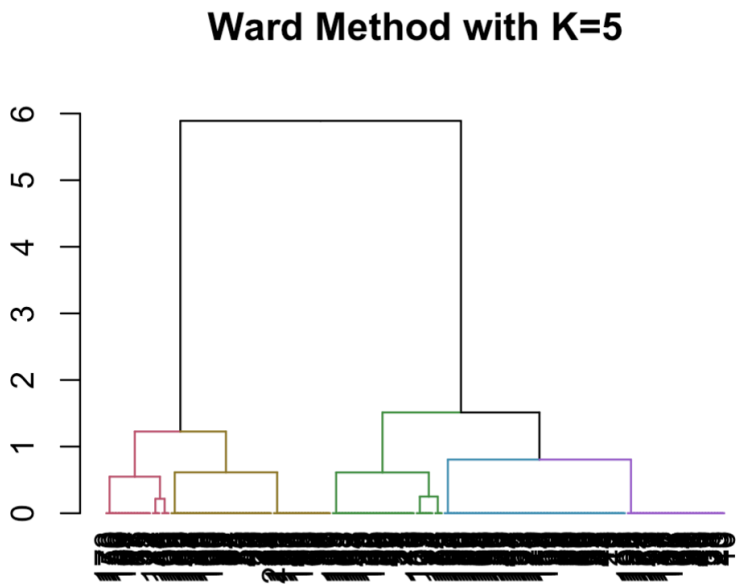

Figure 7: Family Background - Dendrogram for Ward Method

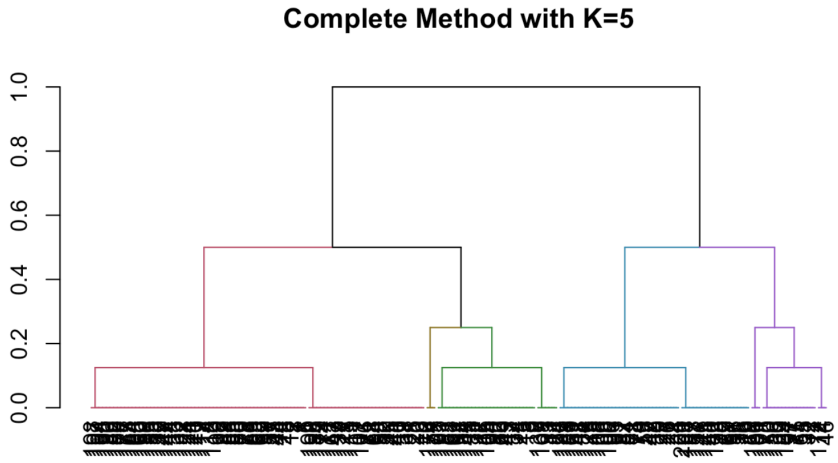

Figure 8: Family Background - Dendrogram for Complete Method

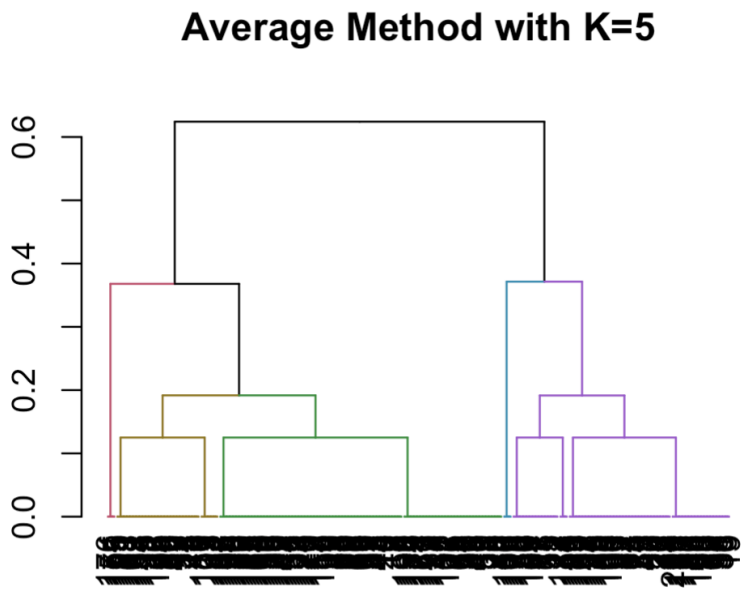

Figure 9: Family Background - Dendrogram for Average Method

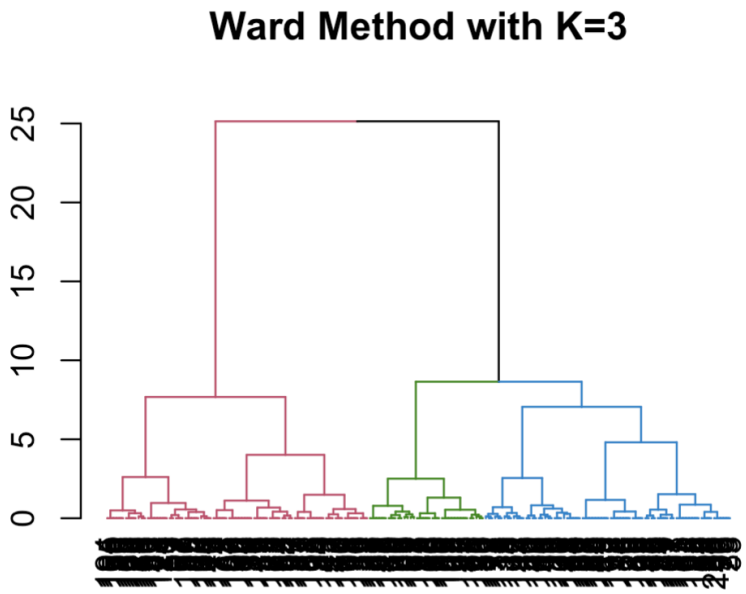

Figure 10: Stress Level - Dendrogram for Ward Method

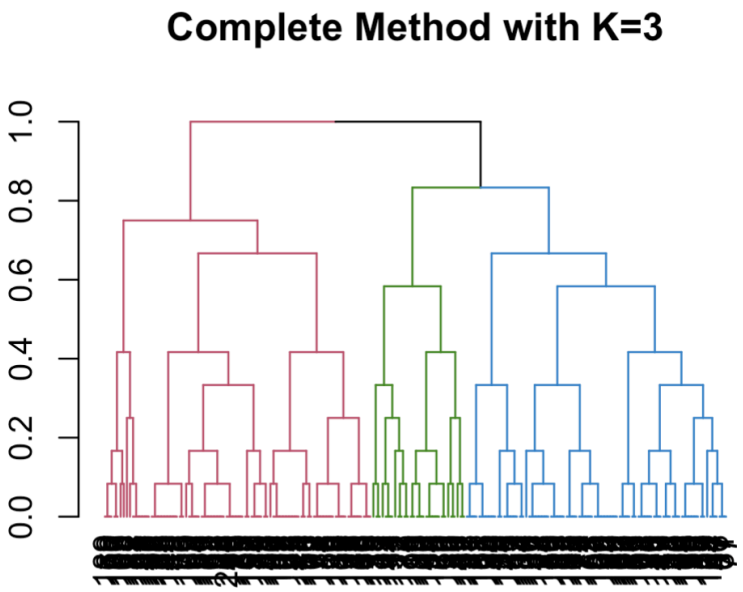

Figure 11: Stress Level - Dendrogram for Complete Method

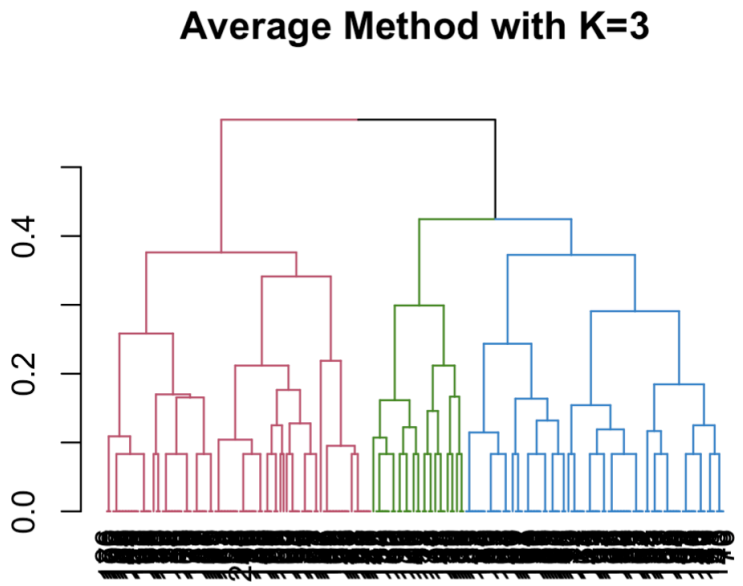

Figure 12: Stress Level - Dendrogram for Average Method