Cluster Analysis

Stage 1- Loading Libraries and Reading CSV Files

```
# Load the necessary libraries

library(tidyverse)
library(dendextend)
library(factoextra)
library(cluster)


# read the data file
student_data <- read.csv("Student_Data.csv")
```

4 packages were loaded:

tidyverse: This package is a collection of R packages for data science. It includes tools for data manipulation (dplyr), visualization (ggplot2), and data importing (readr), among others. It simplifies working with data frames and helps streamline data analysis workflows.

dendextend: This package is designed to extend and customize dendrograms, which are visual representations of hierarchical clustering. It allows users to adjust, compare, and visualize dendrogram objects for better insights into clustering results.

factoextra: This package provides tools for data visualization, especially for multivariate data analysis and clustering results. It helps create easy-to-interpret graphs for Principal Component Analysis (PCA), hierarchical clustering, and k-means clustering.

cluster: This package contains methods for cluster analysis, including hierarchical and k-means clustering. It provides functions like daisy, agnes, diana, pam, and clara, which are used to explore the structure of data and identify natural groupings.

Afterwards, the dataset “student_data” was loaded into R studio using the “read.csv” function and stores in a variable appropriately named “student_data”.

Stage 2 – Variable Exploration

Group 1: Academic performance. Grouping of G1-3 columns for both subjects.

There is clear evidence of a relationship between school academic performance and student health/well-being. Low grades may depress a student who hopes to get into top universities, while high grades may elate a student as they feel they are one step closer to their dream university. Even for students with average grades, they may still feel badly, hoping instead for perfect or near-perfect grades. This is especially prevalent in certain cultures such as Asian ones where top grades are celebrated, while anything below perfect is deemed not good enough (Naumann et al., 2012). While certain subjects may be given more importance than others, overall academic performance is still highly valued and, thus, has a high effect on stress/happiness of students (Nordlander et al., 2014).

Group 2: Parent's Education. Grouping of Medu and Fedu (both numeric-interval)

Grouping Medu (mother's education) and Fedu (father's education) into a combined Parent's Education variable is essential to capture the socio-economic background and its influence on student well-being. Research has shown that students who have parents with higher educational attainment perform better academically (Idris et al., 2020). However, these students might also face higher academic pressure due to the elevated expectations set by their parents (who have achieved highly academically, themselves), leading to potential stress. Thus, both the advantages and pressures associated with higher parental education need to be considered.

Conversely, students whose parents have lower educational attainment may face fewer academic expectations but experience more financial challenges and less access to academic resources, which can increase stress and negatively affect their well-being. The combination of Medu and Fedu into a single variable enables a comprehensive understanding of how parental education shapes a student's academic journey, emotional support, and overall happiness. Grouping these variables together provides a clear picture of the socio-economic pressures and support systems at home that affect student stress and happiness.

Group 3: Family Background. Grouping of famsize (binary) Pstatus (binary)

The grouping of famsize (family size) and Pstatus (parental status, whether parents live together) provides insight into the family environment's impact on a student's stress and happiness. Family structure and dynamics, such as whether parents are living together and the size of the family, play a significant role in shaping a student's emotional and psychological

well-being. Studies indicate that children from smaller families tend to receive more individualized attention from their parents, which can lead to better emotional support and reduced stress. This increase in individualized attention can also have cognitive and behavorial advantages (Blake, 2023). On the other hand, larger families might have more limited resources and parental attention, potentially leading to higher stress and less academic or emotional support.

Additionally, Pstatus, which reflects whether parents are living together or separated, directly affects a student's sense of stability. Research shows that students whose parents are not living together may face increased emotional stress and, thus, academic stress, due to family conflict or instability at home (Cassum, 2018). Conversely, students in intact families (where both parents are living together) often benefit from a more stable home environment, which can contribute to better mental health and academic success. By grouping famsize and Pstatus, we can better understand the family dynamics and their influence on a student's stress levels and overall well-being.

Group 4: Academic support. Grouping of schoolsup, famsup, and paid (all binary).

There is substantial evidence showing that academic support systems, both from schools and families, have a profound impact on students' well-being and success. For instance, research suggests that school-based support, such as tutoring or extracurricular help, can alleviate educational stress and improve academic outcomes. This leads to greater feelings of accomplishment and happiness among students. Parental support for education similarly plays a critical role in shaping students' academic emotions. Positive parental involvement, such as encouraging learning and helping with homework, has been linked to lower stress and greater school satisfaction (Shanlax International Journal of Education, 2022)

Additionally, financial support, such as paid tutoring or access to paid educational resources, can further alleviate educational burden (Bowman-Perrott, 2023). Students who receive such paid support are often better equipped to handle academic challenges, reducing stress levels and improving their overall academic experience. Grouping schoolsup, famsup, and paid together is justified, as all three variables work in tandem to provide a supportive framework that can influence student stress and happiness by addressing different aspects of academic support.

Group 5: Academic Struggles. Grouping of failures (non-oridinal numeric), studytime (ordinal) and absences (non-ordinal numeric)

Research shows that failures are a significant predictor of both academic and emotional struggles. Students who experience repeated academic failures often face

decreased motivation and increased stress, which can lead to a vicious cycle of continued academic underperformance and emotional distress (Cabrera et al., 2001). Studytime, or the amount of time a student dedicates to studying, plays an essential role in academic success; research suggests that inadequate study time is linked to poor performance, while structured study schedules can mitigate the effects of academic pressure (Briones et al., 2019).

Similarly, absences are a well-known indicator of academic struggles. Chronic absenteeism has been consistently linked to lower academic performance and increased dropout rates, as missed instruction and learning opportunities severely affect students' ability to keep up with coursework (Center for Research in Education and Social Policy, 2018). The grouping of these variables together effectively captures a student's academic effort, their experience of failure, and the extent to which absenteeism affects their learning, all of which are key indicators of academic struggles.

Group 6: Health. Grouping of health, dalc, and walc (all ordinal)

The World Health Organization (WHO) has determined that **any** amount of alcohol consumption is detrimental to one's health (Monteiro, 2020). There is a strong association between alcohol consumption and lower overall health. Another analysis found that harmful alcohol use among university students significantly contributes to non-communicable diseases and various risk behaviors, which negatively impact health (Htet et al., 2020).

Grouping health, daily alcohol consumption (dalc), and weekend alcohol consumption (walc) together is crucial for understanding how these behaviors collectively impact students' well-being. This clustering captures the daily habits and weekend social behaviors associated with alcohol use, providing a comprehensive view of their relationship with overall health.

Group 7: Social Perspective. Grouping of goout, famrel, and freetime (all numeric - interval)

Research shows that spending time socializing with friends (goout) positively correlates with emotional health, offering a buffer against academic stress by providing social support and relaxation. Strong family relationships (famrel) are equally important, as they provide emotional stability and act as a protective factor against mental health difficulties. A strong family bond is associated with lower levels of emotional stress and anxiety in students (Aloia, 2020).

Moreover, the amount of freetime a student has can also reflect their ability to balance social and academic life. Properly structured leisure time, including social and physical activities, allows students to decompress, fostering well-being (Doerksen et al., 2014).

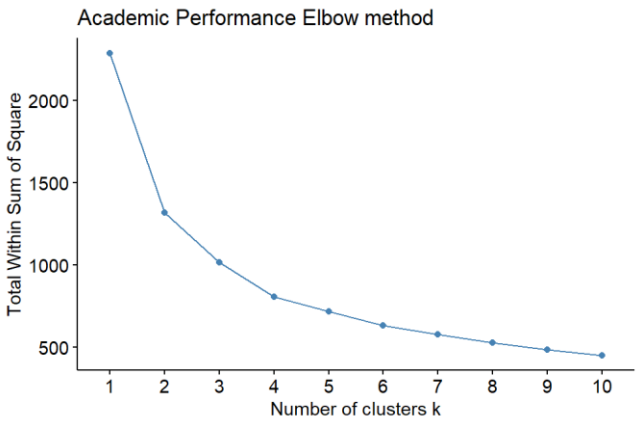
By grouping goout, famrel, and freetime, this cluster captures the intricate relationship between social interactions, family dynamics, and personal time, all of which contribute to a student's overall well-being.

Stage 3 – Variable Construction Through Hierarchical Clustering

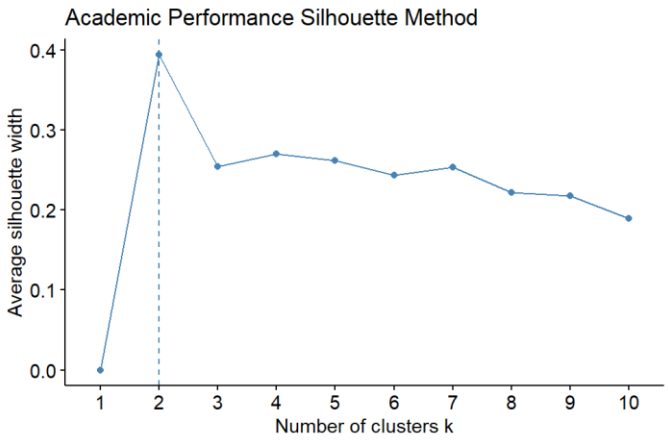
Academic Performance

Checking Ideal Number of Clusters

Elbow Method



Silhouette Method



Codes generated for the above (and all subsequent) elbow and silhouette methods are in the appendix.

From the elbow method, k=2 and k=4 seem ideal, while from the silhouette method k=2 seems ideal. So k=2 and k=4 will be tried.

The Euclidean distancing method was then used on the scaled numeric data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible, while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each desired k. The three methods attempted were complete, ward, and centroid linkage methods. The results of each method are as follows:

Complete Method

Sample sizes for k=2

	cluster1_complete_k2	n
1	1	231
2	2	151

Sample sizes for k=4

	cluster1_complete_k4	n
1	1	99
2	2	105
3	3	151
4	4	27

Sample sizes seem relatively okay, although cluster 4 for k=4 seems a bit small.

Ward's Method

Sample sizes for k=2

	cluster1_ward_k2	n
1	1	228
2	2	154

Sample sizes for k=4

	cluster1_ward_k4	n
1	1	159
2	2	59
3	3	95
4	4	69

The results using ward's method seem slightly better than using the complete method. Especially for the clusters in k=4, where the sample sizes are more balanced.

Centroid Method

Sample sizes for k=2

	cluster1_centroid_k2	n
1	1	381
2	2	1

Sample sizes for k=4

	cluster1_centroid_k4	n
1	1	375
2	2	1
3	3	2
4	4	4

Codes used to generate the above and all subsequent sample size tables are in the appendix.

Sample sizes for certain clusters in both k=2 and k=4 are unacceptable. Some clusters only have a sample size of 1 or 2.

Therefore, ward's method will be chosen.

Interpretation for academic performance

Summary for k=2

	cluster1_ward_k2	`math first period`	`math second period`	`math final period`	`language first period`	`language second period`	`language final period`
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	12.5	12.7	12.9	13.5	13.6	14.1
2	2	8.4	7.7	6.6	10.1	10.2	10.2

Both groups present scores that do not stray too far from the average.

Cluster 1: High Achievers. Students in this cluster are high academic performers. They maintain consistently strong scores in both math and language across all periods. These students are likely motivated, academically focused, and perform well in a variety of subjects.

They may represent students who are engaged in their studies and consistently put in the effort to excel.

Cluster 2: Struggling Learners. This cluster represents students who may be struggling academically. Their math scores are notably lower compared to Cluster 1, and their language scores, while better than math, are still lower overall. These students might be experiencing difficulties in grasping certain concepts or lack access to resources or support, impacting their performance.

Summary for k=4

	cluster1_ward_k4	`math first period`	`math second period`	`math final period`	`language first period`	`language second period`	`language final period`
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	11.1	11.5	11.6	12.7	12.8	13.3
2	2	7.4	5.6	2.6	11.1	11.1	11.6
3	3	9	9	9.1	9.4	9.6	9.4
4	4	15.8	15.7	16.1	15.3	15.5	15.9

Both initial clusters have split into 2 new clusters.

Cluster 1: Consistent learners. This cluster represents consistent and moderate performers. Their math and language scores are steady and above average across all periods. These students are likely those who maintain consistent effort throughout their academic career, achieving stable and good performance without major fluctuations.

Cluster 2: Language-driven learners. This group of students performs better in language than in math. Their language scores remain above average while their math scores are very weak, indicating that they might prefer or excel more in verbal or linguistic skills. These students could represent individuals who are more focused on humanities or social sciences.

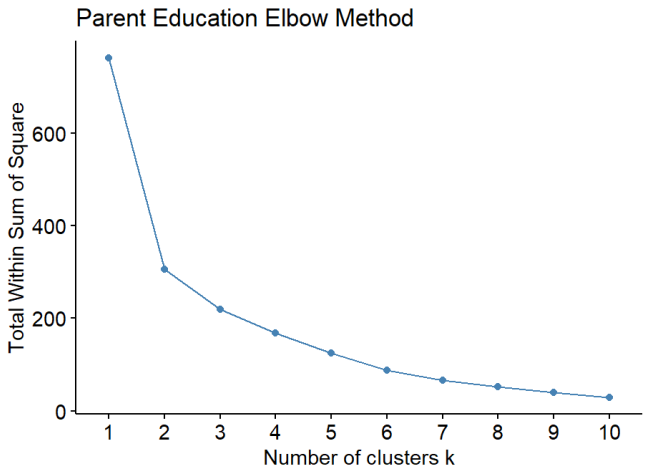
Cluster 3: Struggling learners. This group represents students who are struggling academically, particularly in language subjects, where they have the lowest scores across all periods. Their math performance, while slightly better than their language scores, is still relatively low compared to other clusters. This group could consist of students who face difficulties in both subjects and might need additional support or resources to improve their academic performance. They may lack confidence in these subjects or have gaps in foundational knowledge.

Cluster 4: Grade A students: This group represents top performers in both math and language. They maintain the highest scores in both subjects across all periods, suggesting they are academically gifted or highly motivated students who perform consistently at a high level. These students might represent the top academic tier, excelling across all areas.

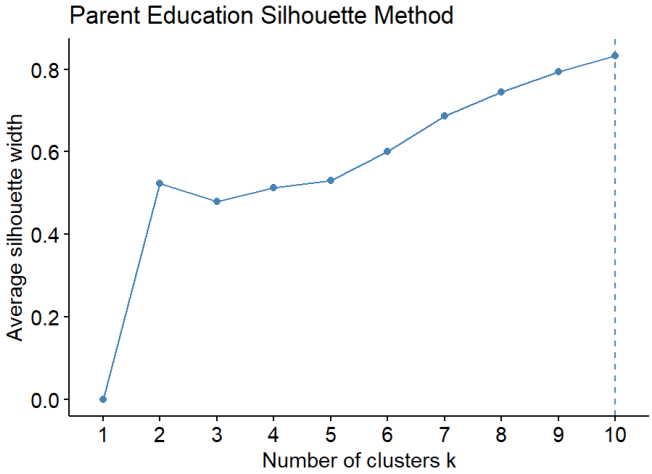
Parent's Education

Checking ideal number of clusters using elbow and silhouette method

Elbow



Silhouette



The elbow method shows k=2 to be ideal. The silhouette method shows an ideal number of clusters at k > 10, which is unreasonable and too complicated. Silhouette shows a slight peak at k=2. So based off of both methods, k=2 seems to be the ideal number of clusters.

The Euclidean distancing method was then used on the scaled numeric data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible, while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each desired k. The three methods attempted were ward, complete, and centroid linkage methods. The results of each method are as follows:

Ward's Method

Sample sizes for k=2

	cluster2_ward_k2	n
1	1	130
2	2	252

Complete Method

Sample sizes for k=2

	cluster2_complete_k2	n
1	1	204
2	2	178

Average Method

Sample sizes for k=2

	cluster2_average_k2	n
1	1	381
2	2	1

Based on the tables above, the complete linkage method has produced the most balanced sample sizes for its clusters. Therefore, complete linkage will be chosen.

Interpretation for Parent's Education

Summary

	cluster2_complete_k2	Medu	Fedu
	<int>	<dbl>	<dbl>
1	1	2.1	1.7
2	2	3.6	3.6

Cluster 1: Students With Lowly-educated parents. This cluster may represent students whose parents have attained lower levels of formal education. Students in this group are likely to come from families with limited academic resources and a lower socioeconomic status. Due to their parents' lower educational attainment, these students may experience fewer academic expectations at home. The parents may face difficulties in guiding their children academically, either because of a lack of formal education or the resources required to provide adequate support. Consequently, students from these families might experience higher academic stress and less support in navigating educational challenges. This aligns with research suggesting that parental educational attainment is strongly linked to children's academic outcomes and overall well-being.

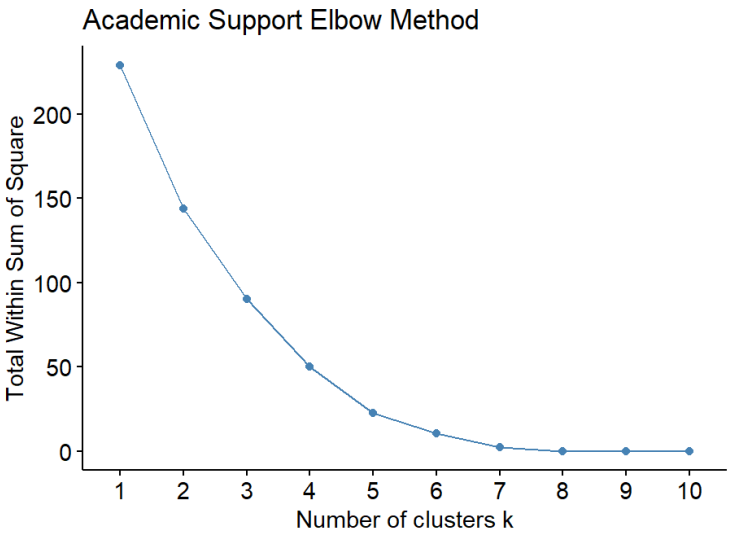
Cluster 2: Students With Highly-educated parents. This cluster represents students where both parents have higher levels of educational attainment. These students are likely to benefit from the academic guidance and expectations that come with having more educated parents. Families in this group may place a higher value on education, creating an environment conducive to academic success. Students in this group might also have access to better social and economic resources (Sirin, 2005), contributing to lower academic stress and more academic achievement. Studies indicate that parents with higher educational levels are better able to support their children's academic development, providing both emotional and material resources for their success.

Academic Support

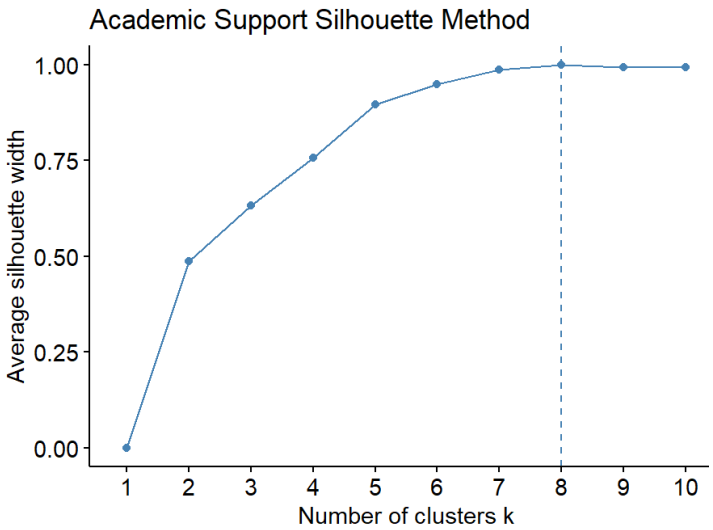
The relevant variables were chosen as determined in stage 2 and formed into a new data frame (See Appendix for example codes)

Checking ideal number of clusters using elbow and silhouette method

Elbow



Silhouette



Based on the elbow method, k=2, k=5, and k=7 all seem ideal. Based on the silhouette method, k=8 seems ideal. Therefore k=2, k=5, k=7, and k=8 will be tried. Codes for this and all subsequent elbows and silhouettes are available in the appendix.

The binary distancing method was then used on the binary data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each desired k. The three methods attempted were complete, ward, and centroid linkage methods. The results of each method are as follows:

Complete Linkage Method

Sample sizes for k=2

	cluster3_complete_k2	n
1	1	343
2	2	39

Sample sizes for k=5

	cluster3_complete_k5	n
1	1	37
2	2	201
3	3	14
4	4	91
5	5	39

Sample sizes for k=7

	cluster3_complete_k7	n
1	1	37
2	2	116
3	3	85
4	4	11
5	5	91
6	6	39
7	7	3

Sample sizes for k=8

	cluster3_complete_k8	n
1	1	19
2	2	18
3	3	116
4	4	85
5	5	11
6	6	91
7	7	39
8	8	3

We can see that the complete method has produced off-balanced sample sizes for all k. Sample sizes for clusters using k=2 and k=5 may not have been strictly too small (none of n > k), however, they were still off-balanced with certain clusters having a significantly bigger sample size than others. Clusters with a sample size too small may produce unreliable results, so it is best to get clusters with as sufficient of sample sizes as possible.

For k=7 and k=8, some clusters produced completely unacceptable sample sizes (cluster 7 for k=7 and cluster 8 for k=8) where n < k. These results may indicate that the complete linkage method, even for k=2 and k=5, may not be the most suitable

Ward's Minimum Variance Linkage Method

Sample Sizes for k=2

	cluster3_ward_k2	n
1	1	291
2	2	91

Sample Sizes for k=5

	cluster3_ward_k5	n
1	1	51
2	2	116
3	3	85
4	4	91
5	5	39

Sample Sizes for k=7

	cluster3_ward_k7	n
1	1	22
2	2	18
3	3	116
4	4	85
5	5	11
6	6	91
7	7	39

Sample Sizes for k=8

	cluster3_ward_k8	n
1	1	19
2	2	18
3	3	116
4	4	85
5	5	11
6	6	91
7	7	39
8	8	3

We can see that ward's linkage method produced better results with more balanced clusters. However, for k=8 cluster 8 has too small of a sample size (n < k).

Centroid Linkage Method

Sample Sizes for k=2

	cluster3_centroid_k2	n
1	1	291
2	2	91

Sample Sizes for k=5

	cluster3_centroid_k5	n
1	1	238
2	2	11
3	3	91
4	4	39
5	5	3

Sample Sizes for k=7

	cluster3_centroid_k7	n
1	1	37
2	2	116
3	3	85
4	4	11
5	5	91
6	6	39
7	7	3

Sample Sizes for k=8

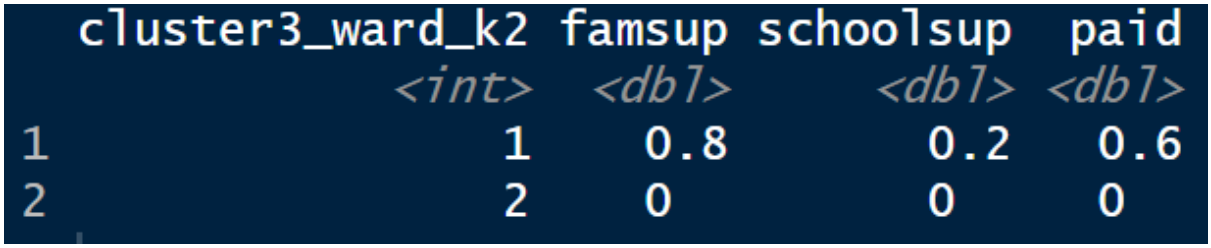
	cluster3_centroid_k8	n
1	1	19
2	2	18
3	3	116
4	4	85
5	5	11
6	6	91
7	7	39
8	8	3

The centroid method produced less adequate results. For k=5, k=7, and k=8, the final cluster had too small of a sample size (n < k).

Based on these results, ward's linkage method was deemed the most suitable and was used. However, k=8 was disregarded as the final cluster had too small of a sample size (n < k)

Interpretation for Academic Support

Summary for K=2



	cluster3_ward_k2	famsup	schoolsup	paid
	<int>	<dbl>	<dbl>	<dbl>
1	1	0.8	0.2	0.6
2	2	0	0	0

Famsup: 1 represents existence of family education support, 0 represents familial educational support.

Schoolsup: 1 represents extra educational support, 0 represents no extra educational support

Paid: 1 represents extra paid classes (within course subject), 0 represents no extra paid classes

Cluster 1: Well-supported Students. These students receive a high level of support in all aspects. With an 80% rate of family support (famsup), they benefit from the emotional and possibly financial resources provided by their families. Additionally, while their school support (schoolsup) is lower at 20%, this could imply that family support plays a more prominent role in their academic assistance. On top of this, 60% of students in this group are receiving extra paid classes (paid), indicating that they are likely investing more into their education through external resources, which can further help alleviate academic pressure.

Cluster 2: Self-reliant Students. In stark contrast to Cluster 1, these students receive no significant support in any form. With 0% family support (famsup), school support (schoolsup), and no access to paid extra classes (paid), these students may be facing challenges in terms of resources and guidance. The lack of external support could increase stress and make it harder for these students to cope with academic demands, potentially affecting their overall well-being and academic outcomes (Gammon & Morgan-Samuel, 2005). This lack of support may stem from various socioeconomic or institutional barriers, leaving them at a disadvantage compared to their peers in Cluster 1.

Summary for K=5

	cluster3_ward_k5	famsup	schoolsup	paid
	<int>	<dbl>	<dbl>	<dbl>
1	1	0.7	1	0.4
2	2	1	0	1
3	3	1	0	0
4	4	0	0	0
5	5	0	0	1

We can see that cluster 2 from the previous example remained the same here (represented by cluster 4). Therefore, this cluster did not break up into other clusters. It's interpretation remains the same.

Cluster 1: School-focused achievers. These students are well-supported by both their families and schools, with 70% receiving family support and 100% benefiting from school support. However, only 40% of them rely on paid external tutoring. This group likely represents students who receive adequate support from their schools and families, preferring to focus on institutional resources rather than seeking private tuition.

Cluster 2: Family-Driven Scholars. Students in this cluster receive strong family support (100%) but rely more on external, paid tutoring (100%) rather than school support (0%). This suggests that their academic progress is largely driven by family investment in private education, perhaps supplementing any gaps in school support. These might be students from well-off families that prioritize external tutoring.

Cluster 3: Holistically-Supported Students. Students in this cluster benefit from full support across the board, receiving both family (100%) and school support (100%), but without participating in paid tutoring. These students likely have access to a well-rounded support system, making additional paid tutoring unnecessary. This group may represent students who thrive under institutional and family guidance.

Cluster 5: Privately-Tutored Independents. These students rely solely on private tutoring, with no family or school support. All of them (100%) participate in paid tutoring, indicating that private education is their main source of academic assistance. This group may include students whose families have chosen to invest in private tutoring rather than depending on school or family support.

Summary for K=7

	cluster3_ward_k7	famsup	schoolsup	paid
	<int>	<dbl>	<dbl>	<dbl>
1	1	0.9	1	1
2	2	1	1	0
3	3	1	0	1
4	4	1	0	0
5	5	0	1	0
6	6	0	0	0
7	7	0	0	1

Clusters 2, 3, 4, 5 from k=5 have remained the same. Cluster 1, however, has broken up into 3 clusters represented by Cluster 1, Cluster2, and Cluster 5.

Cluster 1: Overly-Supported Students. Cluster 1 students are almost fully supported by both family (famsup = 0.9) and school (schoolsup = 1), and they additionally participate in paid extracurricular support (paid = 1). This cluster represents students who benefit from extreme levels of academic and emotional support across multiple avenues. The intensity of their support network suggests that their families and schools place a significant emphasis on ensuring their success, which may be linked to high expectations and aspirations for these students. The fact that families in this cluster can afford all the extra support shows that they may be richer or come from a higher socioeconomic background.

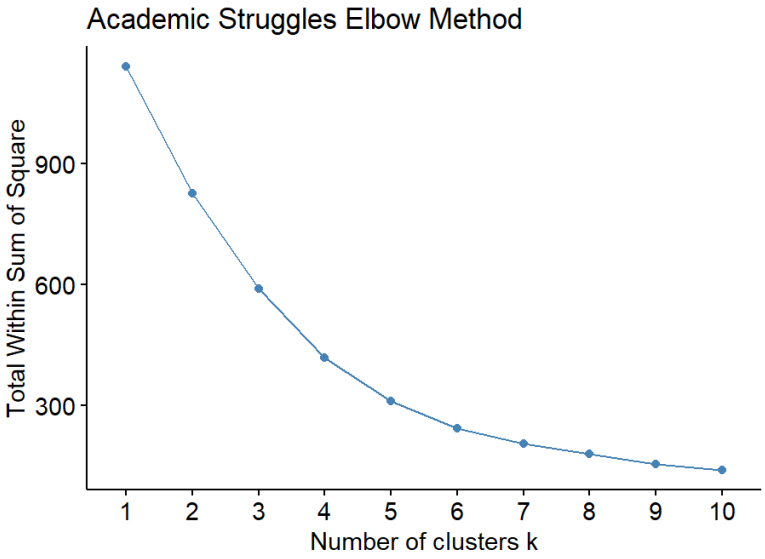
Cluster 2: Family-Backed Scholars. In this cluster, students receive full support from family (famsup = 1) and schools (schoolsup = 1), but do not engage in paid tutoring or extracurricular academic support (paid = 0). This group relies heavily on internal family and school resources to succeed. They might come from households that prioritize the importance of direct family involvement in education but either cannot afford or do not see the need for paid academic assistance. The strong family and school support suggest they are still well positioned for academic success, but the absence of paid support might indicate a more traditional or financially constrained background.

Cluster 5: School-Supported Independents. Students in this cluster do not receive any family support (famsup = 0), but they do benefit from school-based academic support (schoolsup = 1), with no reliance on paid tutoring or resources (paid = 0). This group may represent students who rely primarily on school-provided resources to succeed, perhaps due to a lack of support at home. They may be attending schools that have robust support systems in place, compensating for the absence of family and financial resources. These students could be considered more independent in their learning, leaning on their school for guidance while managing without external or family help.

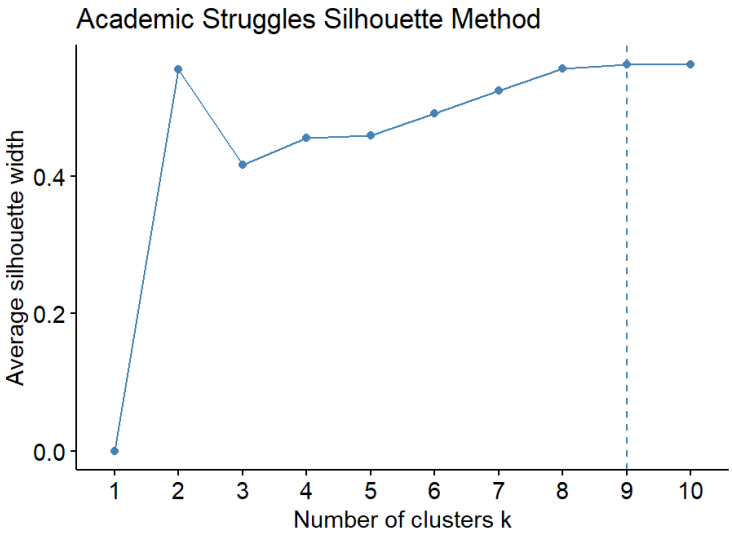
Academic Struggles

Checking ideal number of clusters using elbow and silhouette method

Elbow



Silhouette



Elbow method showed no clear results. Silhouette method showed k=9 to be ideal. It also showed a significant peak at k=2. Therefore, k=2 and k=9 will be tried.

The Euclidean distancing method was then used on the scaled numeric data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible, while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each of the desired k. The three methods attempted were complete, ward, and single linkage methods. The results of each method are as follows:

Complete Method

Sample Sizes for k=2

	cluster4_complete_k2	n
1	1	379
2	2	3

Sample Sizes for k=9

	cluster4_complete_k9	n
1	1	6
2	2	26
3	3	225
4	4	81
5	5	19
6	6	10
7	7	3
8	8	6
9	9	6

It was already clear from k=2 that the complete linkage method would not be suitable for this grouping. K=9 further proved this as there are many clusters with too small of sample sizes.

Ward's Method

Sample Sizes for k=2

	cluster4_ward_k2	n
1	1	300
2	2	82

Sample Sizes for k=9

	cluster4_ward_k9	n
1	1	29
2	2	64
3	3	56
4	4	118
5	5	26
6	6	18
7	7	49
8	8	19
9	9	3

K=2 for ward's method showed okay sample sizes, while k=9 produced a sample size too small for cluster 9

Single Linkage Method

Sample sizes for k=2

	cluster4_single_k2	n
1	1	379
2	2	3

Sample sizes for k=9

	cluster4_single_k9	n
1	1	37
2	2	9
3	3	313
4	4	2
5	5	15
6	6	2
7	7	1
8	8	1
9	9	2

Similar to the complete method, k=2 produced results already showing that the method may not be suitable. K=9 justified this by producing many clusters with too small of sample sizes.

Given that the complete and single methods were not suitable, ward's method was chosen by process of elimination. K=9 is not suitable so only k=2 will be interpreted

Interpretation For Academic Struggles

Summary for k=2

	cluster4_ward_k2	failures	studytime	absences
	<int>	<dbl>	<dbl>	<dbl>
1	1	0.4	1.7	5.8
2	2	0	3.3	3.7

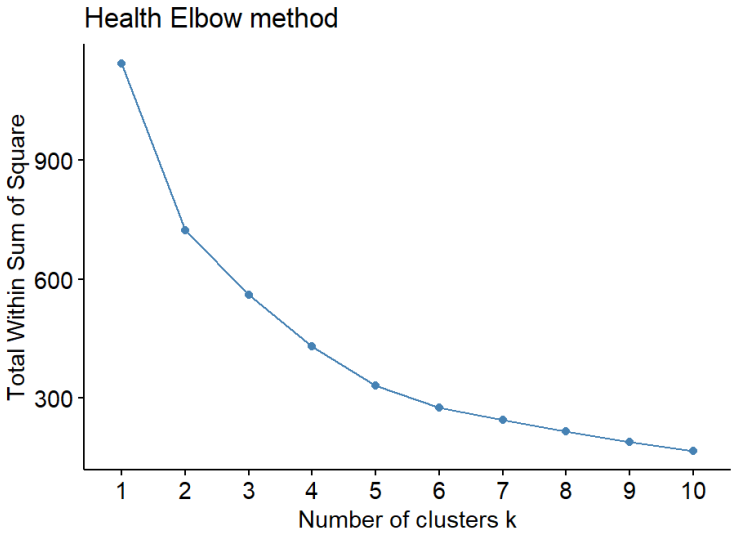
Cluster 1: Struggling Students. This group represents students who are experiencing academic difficulties. They report low study time, high absences, and a moderate number of failures. These students might be struggling with time management or face challenges outside the classroom that lead to inconsistent attendance and reduced study hours. As a result, their performance is affected by their inability to keep up with the workload, leading to higher failure rates and lower overall academic performance.

Cluster 2: Committed and successful students. On the other hand, students in this group tend to dedicate more time to their studies, with study times almost double that of Cluster 1. They also exhibit better attendance and no failures, indicating a stronger commitment to their academic responsibilities. These students may have a structured approach to studying and are likely to benefit from better academic outcomes due to their diligence and consistent efforts. This group may represent students who have better study habits, access to academic support, or face fewer distractions that might impede their learning.

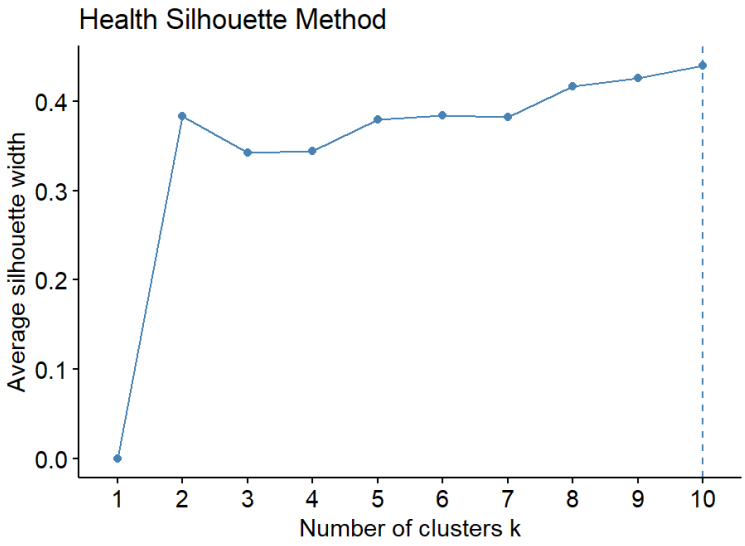
Health

Checking ideal number of clusters using elbow and silhouette method

Elbow



Silhouette



Based on the elbow method, k=2, k=5, and, very subtly, k=6 seem ideal. Based on the silhouette method, a k > 10 seems ideal, however, that is unreasonable and too complicated. Silhouette shows a peak at k=2. Based on both methods, k=2, k=5, and k=6 will be tried.

The Euclidean distancing method was then used on the scaled numeric data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible, while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each of the desired k. The three methods attempted were complete, ward, and centroid linkage methods. The results of each method are as follows:

Complete Method

Sample sizes for k=2

	cluster5_complete_k2	n
1	1	341
2	2	41

Sample sizes for k=5

	cluster5_complete_k5	n
1	1	108
2	2	200
3	3	33
4	4	16
5	5	25

Sample sizes for k=6

	cluster5_complete_k6	n
1	1	108
2	2	72
3	3	128
4	4	33
5	5	16
6	6	25

The complete method produced barely acceptable sample sizes for all desired k. Other methods may produce better results.

Ward's Method

Sample sizes for k=2

	cluster5_ward_k2	n
1	1	222
2	2	160

Sample sizes for k=5

	cluster5_ward_k5	n
1	1	59
2	2	114
3	3	72
4	4	91
5	5	46

Sample sizes for k=6

	cluster5_ward_k6	n
1	1	59
2	2	68
3	3	72
4	4	46
5	5	91
6	6	46

The sample sizes look very balanced! An interesting observation is that the overall balance improved from k=5 to k=6, which seems counterintuitive as increasing k usually means it is harder to get enough sample sizes in all clusters. However, that doesn't seem to be the case here.

Centroid Method

Sample sizes for k=2

	cluster5_centroid_k2	n
1	1	368
2	2	14

Sample sizes for k=5

	cluster5_centroid_k5	n
1	1	368
2	2	2
3	3	10
4	4	1
5	5	1

Sample sizes for k=6

	cluster5_centroid_k6	n
1	1	368
2	2	1
3	3	10
4	4	1
5	5	1
6	6	1

The centroid method is clearly very unsuitable here. For k=2 the sample size for cluster 2 was quite small. For k=5 and k=6 there were multiple clusters with a sample size of only 1.

The final chosen method is ward's linkage method, which produced very balanced cluster sizes.

Interpretation for health

Summary for k=2

	cluster5_ward_k2	health	Dalc	Walc
	<int>	<dbl>	<dbl>	<dbl>
1	1	3.4	1	1.4
2	2	3.8	2.1	3.5

At k=2, both groups have relatively similar "health" levels. Both groups have increased weekend alcohol consumption, which is logical.

Cluster 1: Infrequent Drinkers. This group exhibits relatively moderate levels of health (3.4), coupled with low alcohol consumption, both on weekdays (Dalc = 1) and weekends (Walc = 1.4). The lower levels of alcohol consumption suggest that this group may prioritize other aspects of well-being and social activities that don't revolve around alcohol. Their weekend consumption is slightly higher, which aligns with typical social drinking patterns, but it's still significantly lower than that of Cluster 2. Despite being labeled "Infrequent Drinkers," their health score is just slightly lower than Cluster 2. This suggests that while they consume less alcohol, they may still face other health-related factors that prevent them from achieving a higher health rating.

Cluster 2: Frequent Drinkers. This group shows slightly better overall health (3.8), despite having a much higher alcohol consumption rate both on weekdays (Dalc = 2.1) and weekends (Walc = 3.5). The higher weekend alcohol consumption is particularly notable, indicating that this group likely engages in more social activities or environments where drinking is a norm. The fact that their health score is still relatively high despite their alcohol consumption could be due to other factors such as physical activity, diet, or genetic resilience that buffer the negative effects of alcohol. However, in the long run, this pattern of alcohol consumption could become detrimental to their overall health.

Summary for k=5

	cluster5_ward_k5	health	Dalc	Walc
	<int>	<dbl>	<dbl>	<dbl>
1	1	1.4	1	1.6
2	2	4.4	1.7	3.4
3	3	5	1	1.3
4	4	3.4	1	1.3
5	5	2.4	3.1	3.8

From k=2 to k=5 we can see that both initial clusters have broken up into 5 clusters in total.

Cluster 1: Chronic Health Sufferers. This group could represent individuals who are suffering from chronic health conditions, such as long-term illnesses or mental health challenges. They abstain from heavy drinking, potentially because of their health issues, which are not related to alcohol. These individuals might represent those dealing with physical limitations or significant stressors that impact their health.

Cluster 2: Resilient Social Drinkers. This group may include socially active individuals who can balance moderate weekday drinking and heavy weekend drinking with relatively good health. These people could represent young professionals or active social individuals who

indulge on weekends but take care of themselves during the week. Their resilience to alcohol's effects could reflect good genetics or an active lifestyle.

Cluster 3: Abstainers in Peak Health. This group may represent health-conscious individuals or those who abstain from alcohol for personal, cultural, or religious reasons. They maintain the best health among all groups and drink very little, if at all. These individuals may prioritize physical fitness, mental well-being, and are less likely to engage in risky behaviors.

Cluster 4: Non-drinkers with minor health issues. Similar to Cluster 3, this group has low alcohol consumption, but they experience slightly lower health. These individuals might be experiencing minor health issues unrelated to alcohol. This group could represent people who avoid drinking for various reasons but still face other challenges impacting their health, such as stress or environmental factors.

Cluster 5: Frequent drinkers Facing Consequences. This group represents individuals who are likely experiencing the negative consequences of heavy alcohol consumption. They drink heavily on both workdays and weekends, and their health reflects the toll this behavior is taking. This group could include individuals dealing with stress or high-pressure lifestyles, leading to consistent and excessive drinking as a form of coping, which impacts their overall well-being.

Summary for K=6

	cluster5_ward_k6	health	Dalc	Walc
	<int>	<dbl>	<dbl>	<dbl>
1	1	1.4	1	1.6
2	2	4.6	2.2	3.3
3	3	5	1	1.3
4	4	4.1	1	3.5
5	5	3.4	1	1.3
6	6	2.4	3.1	3.8

We can see that, from k=5, cluster 2 has split into 2 new clusters, cluster 2 and cluster 4. The remaining clusters remain the same.

In k=5, cluster 2 was interpreted as healthy moderate drinkers. In k=6, Cluster 2 has a higher health score than cluster 4 and a higher alcohol consumption during workdays. Weekend alcohol consumption for both groups are similar.

The new cluster 2 could be interpreted as balanced hedonists. This group may represent professionals or individuals who manage to strike a balance between their drinking habits and their health. They are likely to have the ability to maintain a relatively good health

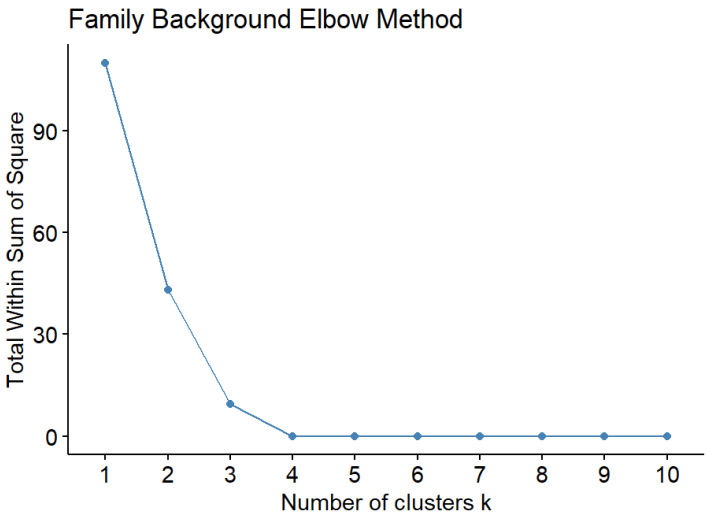
score despite drinking moderately during workdays and indulging more heavily on weekends. These individuals could be part of a group that enjoys socializing after work and maintains a generally balanced lifestyle but leans into alcohol consumption for social or stress-relief purposes. This group could include young professionals, corporate workers, or active socialites who juggle work, social engagements, and leisure activities, and who view alcohol as part of their weekly relaxation or networking routine.

The new cluster 4 could be interpreted as weekend escapists. This group might consist of people who view weekend drinking as an escape or reward from their otherwise disciplined workweek. They consume little or no alcohol during the workdays but indulge more heavily on the weekends. The slightly lower health score may indicate that their weekend drinking habits take a toll on their overall well-being, perhaps leading to fatigue, stress, or other health issues. These individuals might engage in weekend binge drinking as a means to escape from the stress of their work or personal lives. This group could represent blue-collar workers, students, or those in high-stress jobs who engage in heavy drinking on the weekends to "escape" from their responsibilities or stresses of daily life. They might refrain from drinking during the week but use the weekends as an opportunity to let loose.

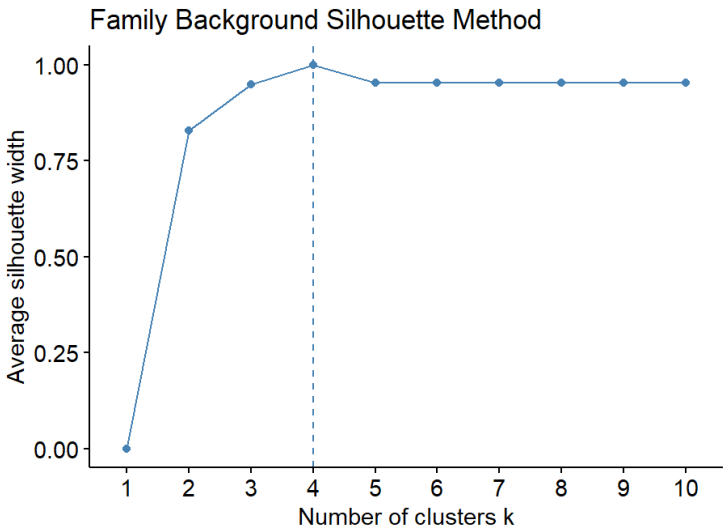
Family Background

Checking ideal number of clusters using elbow and silhouette method

Elbow



Silhouette



Based on the elbow method, k=2, k=3, and k=4 seem suitable. Based on the silhouette method, k=4 seems suitable. Therefore, k=2, k=3, and k=4 will be tried.

The binary distancing method was then used on the binary data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible, while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each desired k. The three methods attempted were complete, ward, and centroid linkage methods. The results of each method are as follows:

Complete Method

Sample sizes for k=2

	cluster6_complete_k2	n
1	1	364
2	2	18

Sample sizes for k=3

	cluster6_complete_k3	n
1	1	344
2	2	20
3	3	18

Sample sizes for k=4

	cluster6_complete_k4	n
1	1	258
2	2	86
3	3	20
4	4	18

The sample sizes for the clusters for all k are quite off-balanced. Complete linkage may not be the best method.

Ward's Method

Sample sizes for k=2

	cluster6_ward_k2	n
1	1	278
2	2	104

Sample sizes for k=3

	cluster6_ward_k3	n
1	1	278
2	2	86
3	3	18

Sample sizes for k=4

	cluster6_ward_k4	n
1	1	258
2	2	86
3	3	20
4	4	18

For the first time, ward's method doesn't seem to be producing the best results. While none of the sample sizes are completely unacceptable, they are still very off-balance, with some clusters having less-than-ideal sample sizes.

Centroid Method

Sample sizes for k=2

	cluster6_centroid_k2	n
1	1	364
2	2	18

Sample sizes for k=3

	cluster6_centroid_k3	n
1	1	344
2	2	20
3	3	18

Sample sizes for k=4

	cluster6_centroid_k4	n
1	1	258
2	2	86
3	3	20
4	4	18

All linkage methods have produced the same results for k=4. None of the linkage methods produced ideal cluster sizes, however, ward seems to be the best and, thus, will be chosen.

Interpretation for Family Background

Summary for k=2

	cluster6_ward_k2 <int>	famsize <dbl>	living_together <dbl>
1	1	1	0.9
2	2	0	0.8

Famsize: 1 represents families >=3, 0 represents families < 3

Living together: 1 represents parent's living together, 0 represents living separately

Cluster 1: Large families with cohabiting parents. Cluster 1 represents large families (3 or more members) where both parents are living together. The famsize value of 1 indicates larger family units, but the key factor here is that the parents are cohabiting, as indicated by the living together value of 0.9. This suggests that the family structure is traditional in the sense

that both parents share the same household. The large family size could imply that these are multigenerational households or households with several children. Even though the living together variable only refers to the parents, the larger family size could mean that other family members also cohabit, reinforcing a sense of family unity and interdependence.

Cluster 2: Small families with cohabiting parents. Cluster 2 consists of small families (fewer than 3 members) where both parents are living together. The living together value of 1 signifies that all parents in this group cohabit, just like in Cluster 1, but the famsize value of 0 indicates a smaller family size. These families could represent nuclear families with fewer children or households where only the parents and one child live together. The emphasis here is on the parents' cohabitation, suggesting a stable family environment, though the overall household is smaller. This could reflect more modern nuclear family dynamics where smaller family units still value parental cohabitation as a key part of family stability.

Summary for k=3

	cluster6_ward_k3	famsize	living_together
	<int>	<dbl>	<dbl>
1	1	1	0.9
2	2	0	1
3	3	0	0

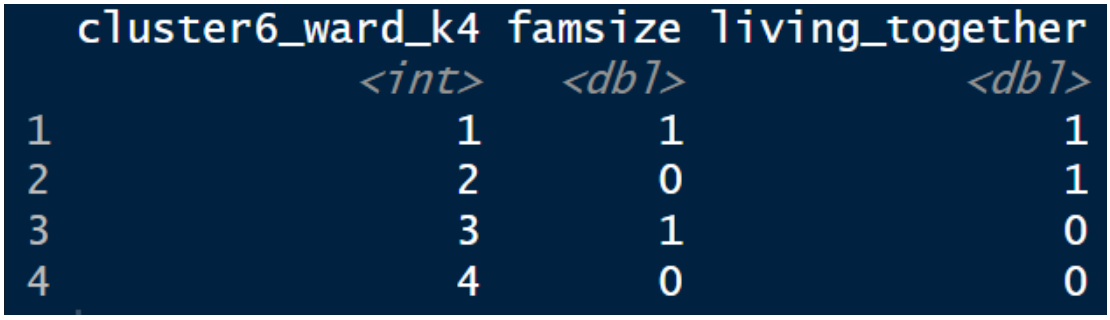
We can see that cluster 2 has broken up into 2 new clusters. Cluster 1 remains the same. In the previous example the 2nd cluster was interpreted as small cohabiting families where most families are living together. These have now been separated into 2 clusters, where one cluster represents small families with cohabiting parents (cluster 2 in the above image) and where the other cluster represents small families with separated parents (cluster 3 in the above image).

Cluster 2: Small families with cohabiting parents. Cluster 2 represents small families where both parents live together, but there are no children in the household. The famsize value of 0 indicates that these are small households with fewer than 3 members, which in this case likely means couples without children. The living together value of 1 indicates that the couples in this group cohabit, but the absence of children distinguishes them from Cluster 1. These families may represent younger couples who have not yet had children, or older couples whose children no longer live with them. In either case, the emphasis is on cohabitation between the parents, but in a childless household.

Cluster 3: Small single-parent households. Cluster 3 consists of small families where no parents are living together, as indicated by the living together value of 0. The famsize value of 0 suggests that these families have fewer than 3 members, and given that the parents are

not cohabiting, this cluster likely represents single-parent households. These households may include one parent raising a single child, where the other parent lives separately or is absent. This cluster reflects a non-traditional family structure, where the focus is on a single parent managing the household without the support of a cohabiting partner. This may occur due to divorce, separation, or work-related circumstances.

Summary for k=4



	cluster6_ward_k4	famsize	living_together
	<int>	<dbl>	<dbl>
1	1	1	1
2	2	0	1
3	3	1	0
4	4	0	0

Now cluster 1 from k=2 has broken up into 2 clusters, represented by cluster 1 and cluster 3 in the above image.

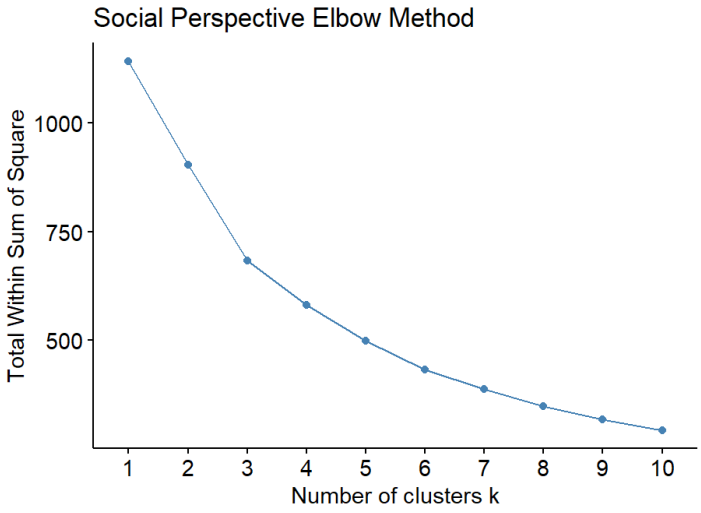
Cluster 1: Large cohabiting families with parents together. Cluster 1 consists of larger families where both parents are cohabiting. The famsize value of 1 indicates that these are families with 3 or more members, and the living together value of 1 confirms that both parents are living together. This suggests a traditional family structure, where cohabitation is a key element. These families likely emphasize family unity and stability, and could involve both parents, children, and possibly extended family members living under the same roof.

Cluster 3: Large families with separated parents. Cluster 3 represents larger families with 3 or more members, but the parents are not living together. The living together value of 0 implies that these are families where the parents are separated, possibly due to divorce, separation, or other circumstances. A large family size may imply children living with one parent while the other lives separately, resulting in a divided household dynamic.

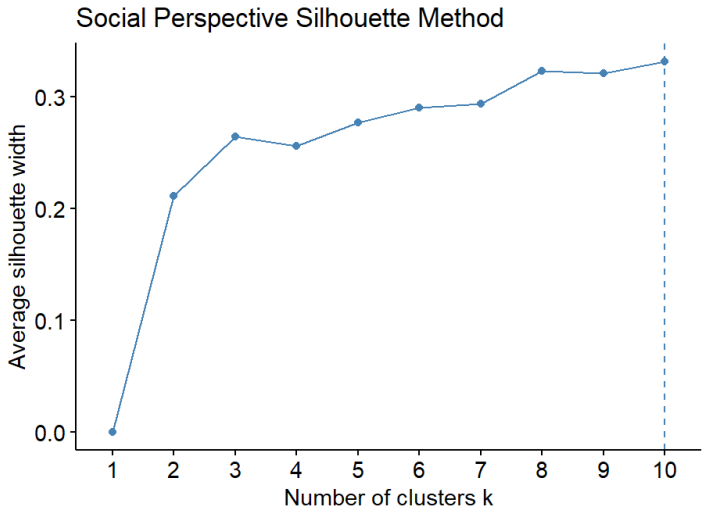
Social Perspective

Checking ideal number of clusters using elbow and silhouette method

Elbow



Silhouette



Based on the elbow method, k=3 and, very subtly, k=6 seem ideal. Silhouette method shows an ideal cluster size at a k > 10 which is, once again, unreasonable and too complicated. K=3 and k=6 will be tried.

Euclidean distancing method was then used on the scaled numeric data to measure how different the data points are from each other, making sure that points within the same cluster are as close as possible, while clusters themselves are as far apart as possible.

Afterward, three different linkage methods were tried for each desired k. The three methods attempted were complete, ward, and average linkage methods. The results of each method are as follows:

Complete Method

Sample sizes for k=3

	cluster7_complete_k3	n
1	1	224
2	2	142
3	3	16

Sample sizes for k=6

	cluster7_complete_k6	n
1	1	191
2	2	33
3	3	107
4	4	20
5	5	16
6	6	15

Some clusters produced using this method seem a little bit too small. Complete method may not be the most suitable.

Ward's Method

Sample sizes for k=3

	cluster7_ward_k3	n
1	1	87
2	2	185
3	3	110

Sample sizes for k=6

	cluster7_ward_k6	n
1	1	87
2	2	51
3	3	49
4	4	37
5	5	110
6	6	48

Ward's method seems (once again) to produce very good results. The sample sizes for the clusters in both k=3 and k=6 seem adequate

Average Method

Sample sizes for k=3

	cluster7_average_k3	n
1	1	379
2	2	1
3	3	2

Sample sizes for k=6

	cluster7_average_k6	n
1	1	24
2	2	44
3	3	299
4	4	1
5	5	12
6	6	2

The average linkage method produced unacceptable sample sizes even at k=3. Some clusters only have a sample size of 1 or 2.

Therefore the most suitable linkage method is ward. K=3 and k=6 are both suitable for interpretation (enough sample size for all the clusters).

Interpretation for Social Perspective

Summary for k=3

	cluster7_ward_k3	goout	famrel	freetime
	<int>	<dbl>	<dbl>	<dbl>
1	1	3.1	2.6	2.8
2	2	2.5	4.3	3
3	3	4.2	4.4	3.9

Cluster 1: Independently social students. This cluster might represent students who are somewhat independent and prefer to balance socializing with their responsibilities, but may not rely on family for socialization. They might be experiencing family challenges or simply don't prioritize family relationships, focusing more on their friendships and personal activities. These students could be from disconnected family backgrounds or situations where the family environment is not strong, driving them to develop independence from a younger age.

Cluster 2: Family-oriented introverts. This group likely includes students who prefer family interactions over social outings. They may be introverted or simply more family-oriented, enjoying a strong emotional bond with their family. These students may come from households where family time is cherished, potentially close-knit families with values that emphasize emotional support at home over external social engagements.

Cluster 3: Outgoing, well-balanced students. This cluster could represent well-balanced students who are both socially active and maintain strong family relationships. They may come from supportive families that encourage socializing and provide a good environment for personal development. These students might be involved in a variety of activities, excelling at balancing social life, academics, and family time.

Summary for k = 6

	cluster7_ward_k6	goout	famrel	freetime
	<int>	<dbl>	<dbl>	<dbl>
1	1	3.1	2.6	2.8
2	2	1.8	3.9	3.7
3	3	2.7	4.3	1.8
4	4	3	4	3
5	5	4.2	4.4	3.9
6	6	2.6	5	3.4

Cluster 2 from the previous k=3 interpretation has now split into four distinct clusters: Cluster 2, Cluster 3, Cluster 4, and Cluster 6. Clusters 1 and 5 have remained unchanged from the previous analysis.

Cluster 2: Introverted, balanced students. This cluster represents introverted students who prefer to spend time with their families rather than engage in social activities with friends. They maintain stronger family connections and have a healthy amount of free time, likely focusing on personal or family-related activities. These students may come from tight-knit families where social interactions outside the family are not a priority.

Cluster 3: Busy, family-oriented students. These students are family-oriented but have limited free time, likely due to academic or extracurricular commitments. They still maintain strong family bonds and socialize from time to time outside of family.

Cluster 4: Well-rounded students. This group represents students who are balanced in both social and family aspects of their lives. They maintain average family relationships and engage in a fair number of social activities. Their lives appear to be well-rounded and without any extremes in social or family dynamics.

Cluster 6: Very family-oriented students with a balanced schedule. This cluster is similar to cluster 3 except they have a much stronger family relationship level (the strongest) and have more free time. This group has excellent time management skills, better than cluster 3. They manage to keep very good relations with family, while also socializing occasionally and having a good amount of free time to spare.

Stage 4: Final Clustering

After combining all our variables, we now do one final clustering on our freshly defined variables: “Academic Performance,” “Parent’s Education,” “Academic Support,” “Academic Struggles,” “Health,” “Family Background,” and “Social Perspective”.

```
final <- data.frame(academic_data_ward_k4$cluster1_ward_k4,
                    parent_educ_data_complete_k2$cluster2_complete_k2,
                    support_data_ward_k7$cluster3_ward_k7,
                    struggles_data_ward_k2$cluster4_ward_k2,
                    health_data_ward_k6$cluster5_ward_k6,
                    family_info_data_ward_k4$cluster6_ward_k4,
                    social_data_ward_k6$cluster7_ward_k6)
```

```
> str(final)
'data.frame':   382 obs. of  7 variables:
 $ academic_data_ward_k4.cluster1_ward_k4           : int  1 2 1 3 1 1 2 3 4 3 ...
 $ parent_educ_data_complete_k2.cluster2_complete_k2: int  1 1 1 2 2 2 2 1 1 2 ...
 $ support_data_ward_k7.cluster3_ward_k7            : int  1 2 1 1 3 1 4 1 4 5 ...
 $ struggles_data_ward_k2.cluster4_ward_k2          : int  1 1 1 2 1 2 1 1 2 2 ...
 $ health_data_ward_k6.cluster5_ward_k6             : int  1 2 1 3 2 3 2 4 2 5 ...
 $ family_info_data_ward_k4.cluster6_ward_k4        : int  1 1 1 1 1 1 1 2 2 3 ...
 $ social_data_ward_k6.cluster7_ward_k6             : int  1 1 2 2 3 2 3 3 2 4 ...
```

By examining the structure of the final dataset, we observe that all variables are of the "int" data type. However, these integers are not continuous, scalar values; rather, they represent the respective cluster that each observation belongs to. These clusters were defined and named in previous stages of the analysis. Hence, despite being technically classified as "int" data types, these variables should be treated as categorical, since the integers merely refer to the cluster categories that specific observations are assigned to.

```
# Convert columns to factor to ensure they are treated as categorical variables
# This is necessary for calculating the distance matrix with Gower distance
final_factor <- final %>% mutate(across(everything(), as.factor))
```

All the columns in the final dataset are transformed to factors in R studio. This is to reflect the categorical nature of these variables, and to ensure accurate distance matrix calculation.

```
# Distance matrix using daisy (Gower distance) to handle categorical data
dist_final <- daisy(final_factor, metric = "gower")
```

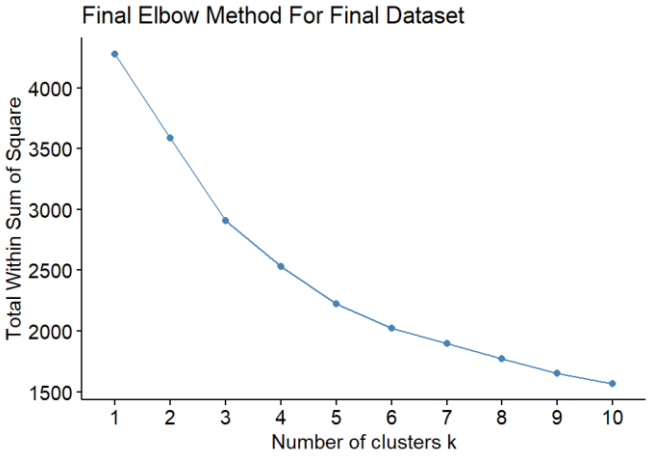
As the dataset consists of categorical variables, where the integers indicate different clusters rather than numerical quantities or binary variables, using the standard dist() function to calculate a distance matrix would not be appropriate. Instead, the daisy() function with the "gower" method is more suitable, as it is designed to handle mixed data types—including categorical data. In this case, even though the dataset only consists of categorical variables

(cluster assignments), daisy() with "gower" remains effective for calculating distances between observations in a way that properly respects the categorical nature of the data.

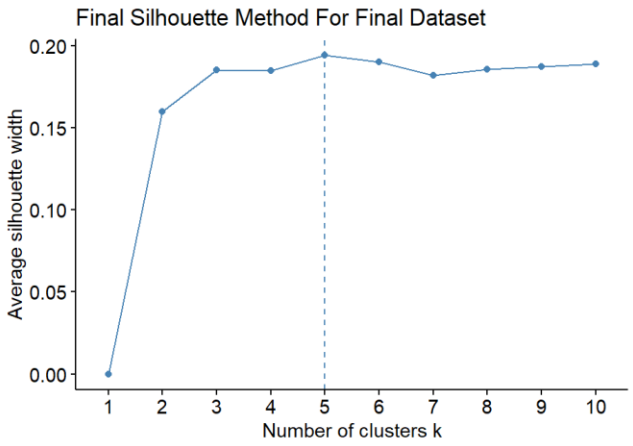
Stage 5 – Hierarchical Clustering for Happiness/Stress Level

Checking Ideal Number of Clusters Using Elbow and Silhouette Method

Elbow



Silhouette



Based on the elbow method, there seems to be 2 bends at k=3 and k=6. For the silhouette method, k=5 is shown to have the highest peak, and, thus, is the optimal number of clusters. Therefore, k=3, k=5, and k=6 will be tried for the final clustering.

Afterward, three different linkage methods were tried for each desired k. The three methods attempted were complete, ward, and average linkage methods. The results of each method are as follows:

Complete Method for Final Clustering

Sample sizes for k=3

	clusterf_complete_k3	n
1	1	350
2	2	21
3	3	11

Sample sizes for k=5

	clusterf_complete_k5	n
1	1	296
2	2	24
3	3	30
4	4	21
5	5	11

Sample sizes for k=6

	clusterf_complete_k6	n
1	1	254
2	2	42
3	3	24
4	4	30
5	5	21
6	6	11

The cluster sizes using the complete linkage method seem off-balanced. In both all 3 cases, cluster 1 has a significantly larger sample size compared to the other clusters.

Ward Method for Final Clustering

Sample sizes for k=3

	clusterf_ward_k3	n
1	1	167
2	2	72
3	3	143

Sample sizes for k=5

	clusterf_ward_k5	n
1	1	28
2	2	79
3	3	72
4	4	143
5	5	60

Sample sizes for k=6

	clusterf_ward_k6	n
1	1	28
2	2	79
3	3	72
4	4	113
5	5	60
6	6	30

The sample sizes using ward's method for the clusters seem more balanced.

Average Method for Final Clustering

Sample sizes for k=3

	clusterf_average_k3	n
1	1	302
2	2	74
3	3	6

Sample sizes for k=5

	clusterf_average_k5	n
1	1	167
2	2	74
3	3	135
4	4	2
5	5	4

Sample sizes for k=6

	clusterf_average_k6	n
1	1	167
2	2	65
3	3	135
4	4	2
5	5	4
6	6	9

The cluster sizes using the average linkage methods produce clusters with too small of sample sizes.

Based on the results, ward's method will be used for the final clustering.

Interpretation of Final Clusters

Summary for K=3

clusterf_ward_k3	mode_Acad Perform	mode_Parent Educ	mode_Acad Supp	mode_Acad Strugg	mode_Health	mode_Fam Bckgrnd	mode_Soc Perspective
1	1	1	6	1	5	1	5
2	1	1	3	2	5	1	1
3	1	2	3	1	2	1	5

Given the commonalities across academic performance and family background for all 3 clusters, it seems these factors don't matter much when segmenting students for stress/happiness analysis. Research suggests a very weak link between academic performance and student happiness/stress. (Wenli Chen et al., 2023) suggests that the grades of Chinese high school students are not correlated with their daily happiness levels, but makes a mention of those students experiencing stress and anxiety attacks. Based on this, it may be concluded that academic performance doesn't affect happiness, but affects stress levels. At k=3, academic performance is seen not to be important in segmenting clusters for happiness/stress analysis. However, at k=5 (as will be seen next), clusters form into 2 different categories for academic performance and at k=6, clusters form into 3 different categories

based on academic performance. Based on the literature and the final clustering data, it may be said that the impact of academic performance on the combination of stress and happiness of students is mixed. It also seems that, at k=3, family background is not important in segmenting the clusters.

The three clusters can be interpreted as follows:

Cluster 1: This cluster has students with the following differentiating characteristics:

- Students with lowly-educated parents (defined under “parent’s education”).
- Self-reliant students (defined under “academic support”).
- Struggling students (defined under “academic struggles”)
- Non-drinkers with minor health issues (defined under “health”)
- Outgoing, well-balanced students (defined under “social perspective”)

Based on the above characteristics, this cluster of students can be named “independent, resilient students”. This cluster of students is characterized by resilience and self-reliance in the face of multiple challenges. Coming from families with low parental education, they lack the academic support that many others may receive at home. Despite these obstacles, they demonstrate a strong sense of independence, managing their studies without relying heavily on external academic assistance. However, this independence is coupled with academic struggles. Nevertheless, they persevere and continue pushing through these challenges, showcasing their inner strength and determination.

In addition to their academic efforts, these students maintain a well-balanced social life and exhibit responsible health behaviors. They are outgoing and socially well-adjusted, which likely serves as a coping mechanism to balance the stress of their academic and personal challenges. They refrain from engaging in risky behaviors, such as drinking, and while they may have minor health issues, they make responsible choices to maintain their well-being. Overall, these students demonstrate resilience, maturity, and a strong capacity for self-reliance, navigating their difficulties while maintaining a balanced lifestyle.

Cluster 2: This cluster has students with the following differentiating characteristics:

- Students with lowly-educated parents (defined under “parent’s education”).
- Family-driven Scholars (defined under “academic support”).
- Committed and successful students (defined under “academic struggles”)
- Non-drinkers with minor health issues (defined under “health”)
- Independently social students (defined under “social perspective”)

Based on the above characteristics, this cluster of students can be named “School-focused independent students. These students are similar to the first cluster in that they come from homes with lowly-educated parents. However, despite their parent’s low education levels, the parents recognize the importance of education for their children and support them in every way they can, including extra paid classes for specific subjects. These parents may recognise their own shortcomings due to not having a high education and, thus, emphasise its importance to their children. This support pays off as these students have very few academic struggles, being committed and successful instead.

Despite the ample family support, these students maintain a somewhat weak relationship with their family, instead preferring to have a more independent and balanced approach to socializing, where they go out from time to time and leave a good amount of free time for themselves. It is clear that the priority for this group of students is academics with socializing coming afterwards. In addition to prioritizing academics, these students also prioritize their health, presenting as non-drinkers with minor health issues. It is clear that students in this cluster are responsible human beings. The parents heavy support and involvement in their children’s academics may be the key to them being disciplined and responsible, allowing for their social-emotional development (Martinez-Yarza, 2024).

Cluster 3: This cluster has students with the following differentiating characteristics:

- Students with highly-educated parents (defined under “parent’s education”).
- Family-driven Scholars (defined under “academic support”).
- Struggling students (defined under “academic struggles”)
- Balanced hedonists (defined under “health”)
- Outgoing, well-balanced students (defined under “social perspective”)

This cluster can be named “privileged, social, irresponsible students”. This cluster represents a group of students who come from highly-educated families and receive considerable academic support from their parents, yet face significant academic struggles. Despite the strong foundation provided by their parents, these students may not be fully capitalizing on their educational opportunities. The fact that they struggle academically

suggests that, despite having the resources and guidance, they may prioritize other aspects of their lives over their studies.

These students exhibit characteristics of balanced hedonists, which implies that they engage in regular social activities, including alcohol consumption. Unlike students who abstain from drinking or maintain a strict focus on their academics, this group manages a lifestyle that balances their health and social indulgences. They may consume alcohol moderately during weekdays while indulging more on weekends. Their outgoing and well-balanced social lives further indicate that social activities and relaxation are significant priorities. This is not necessarily negative, as they seem to maintain good health despite their drinking habits. However, it suggests that their focus is on immediate social enjoyment rather than long-term academic success.

While these students maintain a strong social presence and appear well-adjusted in terms of health and social interactions, their academic struggles may point to a lack of discipline or motivation to focus on their studies. The support from their family may be present, but they appear to invest their energy into social activities and indulgences rather than academics. This could indicate a mismatch between their priorities and the resources available to them. Thus, the group might be seen as individuals who enjoy the privileges of family support and a well-balanced social life, yet lack the discipline or drive to channel these advantages into academic success.

Summary for K=5

clusterf_ward_k5	mode_Acad Perform	mode_Parent Educ	mode_Acad Supp	mode_Acad Strugg	mode_Health	mode_Fam Bckgrnd	mode_Soc Perspective
1	1	1	3	1	1	1	2
2	3	1	6	1	5	1	5
3	1	1	3	2	5	1	1
4	1	2	3	1	2	1	5
5	3	1	6	1	2	2	5

From k=3 to k=5, the cluster “school-focused independent students” has remained the same. It is now presented as the 3rd cluster in the above table for k=5.

The cluster “privileged, social, irresponsible students” has also remained the same. It is now presented as the 4th cluster in the above table for k=5.

Therefore, the first initial cluster for k=5, “independent, resilient students”, has broken up into 3 new clusters. These new clusters are represented by the 1st, 2nd, and 5th clusters in the above table for k=5.

Therefore, “independent, resilient students” can be further clustered into:

New cluster 1:

- Consistent learners (defined under “academic performance”)
- Students with lowly-educated parents (defined under “parent’s education”)
- Family-driven scholars (defined under “academic support”)
- Struggling students (defined under “academic struggles”)
- Chronic health sufferers (defined under “health”)
- Large cohabiting families with parents together (defined under “family background”)
- Introverted, balanced students (defined under “social perspective”)

Based on the above characteristics, the first subgroup of resilient, independent students could be named as “resilient scholars with chronic struggles”. This group is defined by its persistent academic engagement despite multiple challenges. While these students are consistent learners, maintaining steady progress academically, their lives are notably impacted by chronic health issues, which distinguish them from other groups. The ongoing health difficulties may act as a barrier to reaching their full potential, yet they continue to push through these challenges with determination. Their lowly-educated parents provide family support, and they are family-driven scholars, emphasizing the strong role their families play in motivating and assisting them academically. However, despite this support, they still face academic struggles, perhaps due to the compounding effects of poor health and limited academic resources at home.

In addition, this group comes from large, cohabiting families, offering a strong sense of family structure and emotional backing. They are introverted, balanced students, preferring smaller social circles, which could be influenced by both their health struggles and the need for stability in their personal lives. Their resilience is admirable as they navigate life’s difficulties while balancing academic pressures and family responsibilities.

New cluster 2:

- Struggling learners (defined under “academic performance”)
- Students with lowly-educated parents (defined under “parent’s education”)
- Self-reliant students (defined under “academic support”)
- Struggling students (defined under “academic struggles”)
- Non-drinkers with minor health issues (defined under “health”)
- Large cohabiting families with parents together (defined under “family background”)
- Outgoing, well-balanced students (defined under “social perspective”).

The second subcluster of “independent, resilient students”, based on the above characteristics, can be named “self-reliant strugglers with social balance”. These students are defined primarily by their academic struggles, both in terms of performance and effort. Despite these struggles, they exhibit remarkable self-reliance, managing their academic challenges without relying heavily on family or external support. The students come from lowly-educated parents, which could contribute to their academic difficulties, as their families may lack the resources or knowledge to provide academic assistance. However, these students maintain a strong sense of independence, possibly stemming from the necessity to navigate educational hurdles on their own.

Although they are struggling learners, they balance these academic challenges with a vibrant social life. They are outgoing and well-balanced students, suggesting that they manage to maintain their social well-being despite the academic pressures. This social engagement could be a key factor in helping them cope with stress. Additionally, their large cohabiting families may offer emotional support, even if they cannot assist academically. The students are non-drinkers with minor health issues, indicating responsible lifestyle choices, which further reflect their capacity for balance and self-management in the face of both personal and academic challenges.

In essence, these self-reliant strugglers with social balance are students who face considerable academic difficulties but exhibit personal strength and independence. They prioritize maintaining a healthy social life and take care of their physical well-being, allowing them to maintain a semblance of balance despite their educational struggles.

New Cluster 5:

- Struggling learners (defined under “academic performance”)
- Students with lowly-educated parents (defined under “parent’s education”)
- Self-reliant students (defined under “academic support”)
- Struggling students (defined under “academic struggles”)
- Balanced hedonists (defined under “health”)
- Small families with cohabiting parents (defined under “family background”)
- Outgoing, well-balanced students (defined under “social perspective”)

This subcluster can be named “self-reliant hedonists with a social balance”. Students in this cluster face significant challenges in their academic performance, and despite their efforts, they continue to struggle academically. These struggling learners come from lowly-educated parents, which may limit the academic guidance and support available at home. However, they exhibit self-reliance in managing their studies, showing independence and

perseverance despite these hardships. Their self-reliance could be a result of necessity, as they must navigate their academic difficulties without strong external support.

In terms of health and lifestyle, these students are characterized as balanced hedonists. While they are not neglecting their health entirely, they engage in moderate drinking, often balancing this with an otherwise healthy lifestyle. Their social lives reflect this balance: they are outgoing and well-balanced, suggesting that they prioritize social interactions and may use these engagements as a way to cope with academic stress. These students are also from small families with cohabiting parents, which likely provides a stable family environment, although the smaller family size may reduce the amount of support they receive compared to larger families.

These Self-Reliant Hedonists with Social Balance are a group of students who balance personal independence, social engagement, and hedonistic tendencies with the challenges they face in both academics and health. They manage to maintain a healthy social life and avoid more serious negative health outcomes, but their academic struggles remain a key hurdle.

An interesting observation across all 3 subclusters is the commonality across parent's education and academic struggles. All 3 subclusters have parents that are lowly educated and all face academic struggles. It seems these 2 factors are not important when segmenting this cluster into subclusters. However, this may provide insight into the role of parental education and the academic struggles that their resilient, independent children face, as a result. Yet despite these academic struggles, all 3 subgroups are filled with resilient students, demonstrating that a child's socioeconomic background (as determined by their parents through their level of education, albeit not a perfect proxy for socioeconomic status) is not associated with their inner drive, passion, and motivation (Takashiro & Clarke, 2020).

Summary for K=6

clusterf_ward_k6	mode_Acad Perform	mode_Parent Educ	mode_Acad Supp	mode_Acad Strugg	mode_Health	mode_Fam Bckgrnd	mode_Soc Perspective
1	1	1	3	1	1	1	2
2	3	1	6	1	5	1	5
3	1	1	3	2	5	1	1
4	1	2	4	1	2	1	5
5	3	1	6	1	2	2	5
6	4	2	3	1	1	2	6

From k=5 to k=6, cluster 4 has broken up into 2 new clusters, presented by cluster 4 and cluster 6 in the above picture for k=6.

In k=5, the naming and interpretation of cluster 4 was carried down from k=3. In k=3, cluster 4 (which is in cluster 3 there) is defined as “privileged, social, irresponsible students”. This has now been broken up into 2 subclusters:

New cluster 4:

- Consistent learners (defined under “academic performance”)
- Students with highly-educated parents (defined under “parent’s education”)
- Holistically-supported students (defined under “academic support”)
- Struggling students (defined under “academic struggles”)
- Balanced hedonists (defined under “health”)
- Small families with cohabiting parents (defined under “family background”)
- Outgoing, well-balanced students (defined under “social perspective”)

This subcluster inherits all the characteristics of its main cluster, besides academic support. While the “privileged, social, irresponsible students” cluster are family-driven scholars, this subcluster is holistically-supported. They are supported through both their family and their school, but do not receive extra paid classes. In terms of privilege, this subcluster is very similar to its original cluster. Being so similar, I will name this subcluster the same as its main cluster, “privileged, social, irresponsible students”

New Cluster 6:

- Grade A students (defined under “academic performance”)
- Students with highly-educated parents (defined under “parent’s education”)
- Family-driven scholars (defined under “academic support”)
- Struggling students (defined under “academic struggles”)
- Chronic health sufferers (defined under “health”)
- Large cohabiting families with parents together (defined under “family background”)

- Very family-oriented students with a balanced schedule (defined under “social perspective”)

This subcluster can be defined as “intelligent, self-destructive family lovers”. These students have a very solid family relationship. In addition to having a large family, they maintain very good relations with them. The family, in return, support the students academically. This support pays off, as this subcluster is the highest-achieving out of any subcluster or cluster. However, this high-achievement comes with a cost; these students have academic struggles. Given that these are very high-achieving grade A students, it is unlikely that their academic struggles stem from failure. Instead, it is likely that the academic struggles are related to study time and absences. Given that these students achieve so highly despite their lack of study time and/or absences, it can be deemed that these students are quite intelligent. For these students, a lack of studying and/or school absences may not even be a “struggle” as they achieve high grades nonetheless.

However, these may be a sign of self-destructive behavior that these students engage in. These students face chronic health issues, likely due to their very high alcohol consumption throughout the week. This consumption may be a way to cope with the stress that they (or their highly-educated parents) put onto them to achieve very high grades in school. As we know, alcohol is a self-destructive coping mechanism (Cargiulo, 2007) that creates more problems and solves none. Hence, students in this cluster can be characterized as self-destructive.

Stage 6 – Conclusion

In conclusion, this analysis provides valuable insights into the relationship between various factors such as academic performance, parental education, health, family background, and social perspective, and their impact on student well-being. Through hierarchical clustering, several distinct clusters of students emerged, each demonstrating unique characteristics and patterns. Ward's minimum variance linkage method consistently emerged as the preferred linkage type, highlighting its efficacy at handling different kinds of datasets.

One significant insight gained is the role that parental education and academic support play in shaping student outcomes. Students from highly-educated families tend to perform better academically, benefiting from family-driven support, whereas those with lowly-educated parents often face greater academic struggles, highlighting the importance of parental influence. However, it is important to note that not all students with lowly-educated parents struggle academically, as some demonstrate remarkable resilience and independence, navigating challenges with minimal external support.

Health also surfaced as a key differentiator, particularly in distinguishing clusters of students who engage in varying levels of alcohol consumption. The "balanced hedonists" group, for example, showed that students who maintain a moderate drinking lifestyle can still balance their health and social life, although in some cases, this lifestyle could negatively impact their academic performance and well-being.

The analysis of social perspective revealed that family relationships and social interactions play a crucial role in student well-being. Students with strong family bonds tend to have better emotional stability, whereas more independent and outgoing students might rely on their social life to balance the pressures of academic life.

Finally, this study underscores that despite differences in background, many students share common struggles, particularly around academic performance and health. Resilient students emerge across multiple clusters, demonstrating that regardless of their family background or academic struggles, they can maintain balance through strong social or family support. A student's inner passion, drive, and grit is not determined or foreseeable based on parental education levels/socioeconomic status.

These insights suggest the importance of tailored interventions that consider not just academic factors but also the socio-emotional and health-related needs of students to support their overall success and well-being.

Appendix

Example codes

Code for creating and scaling datasets and calculating distance for created groupings (numeric only)

```
academic_performance <- data.frame(student_data$G1.x,
                                   student_data$G2.x,
                                   student_data$G3.x,
                                   student_data$G1.y,
                                   student_data$G2.y,
                                   student_data$G3.y
)


# Standardize

academic_performance_scaled <- scale(academic_performance)

str(academic_performance_scaled) # not data frame

# Convert to data frame

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

The data.frame function is used to create a new data frame for all variables in a certain cluster. In the above example, the variables chosen for academic performance are selected from the main student_data dataset using the $ symbol.

After creating the data frame, the “scale” function is then applied to standardize variables for unbiasedness.

After scaling the data frame, and using the “str” function to check the structure of the resulting data, it was seen that the resulting data type was not a data frame. Thus, the “as.data.frame” function was applied to the scaled data to convert it back to a data frame.

```
dist_struggles <- dist(academic_struggles_scaled)
```

For numeric data frames, calculating the distance can be done by simply applying the “dist” function. The method used to calculate the distance, by default, is the Euclidean distance, which is the type applicable to numeric data. Thus, there was no need to specify a calculation type.

Code for creating datasets and calculating distance for groupings (binary and categorical only)

```r
# Convert categorical "yes"/"no" to logical TRUE/FALSE

academic_support <- data.frame(
  famsup = student_data$famsup == "yes",
  schoolsup = student_data$schoolsup == "yes",
  paid = student_data$paid == "yes"
)

# Calculate binary distance
dist_support <- dist(academic_support, method = "binary")
```

Famsup, schoolsup, and paid were all variables that took either “yes” or “no” as values. For simplicity, I converted them to binary variables, with 1 representing “yes” and 0 representing “no”. Afterwards I was able to apply the “dist” function using the “binary” method to calculate the distance between these now-binary variables.

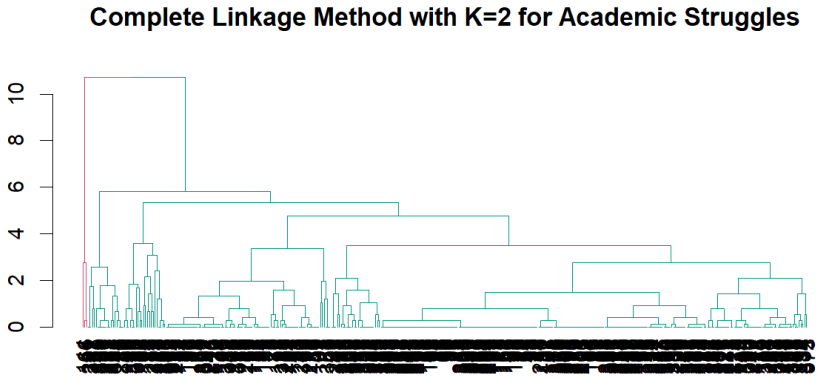
Dendrogram Coding and Output

Code

```
# Plot the dendrogram with k = 2 clusters
plot(color_branches(as.dendrogram(struggles_hclust_complete), k=2),
     main = "Complete Linkage Method with K=2 for Academic Struggles")
```

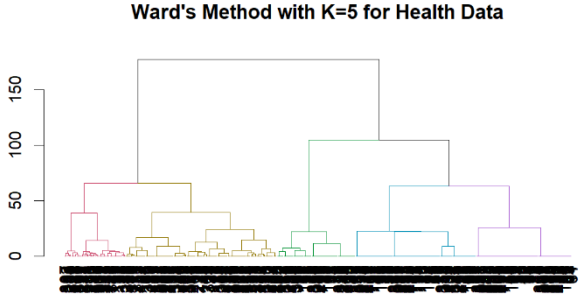
"plot" was used to plot a visual graph. "color_branches" is used to specify that we want a colored dendrogram, with different colors representing different clusters. "as.dendogram" was used to specify that the visual is a dendrogram, and the "k" argument specifies the number of clusters in the dendrogram, determined previously using the elbow and silhouette methods. "main" specifies the title of the dendrogram.

Off-balanced dendrogram



The above dendrogram is an example of a very off-balanced dendrogram. As you can see there are 2 clusters represented by red and green. The green cluster takes up significantly more space and has significantly more branches. These dendrograms help give a clue to the sample size of each cluster. For example for the above, it was predicted that the complete linkage method for academic struggles with k=2 would generate very off-balanced sample sizes and it did. Cluster 2 had a very small sample size compared to cluster 1

Balanced Dendrogram



Example Codes for Clustering Algorithm

Codes for Checking Optimal Number of Clusters Using Elbow and Silhouette Methods

```
health_elbow <- fviz_nbclust(health_scaled, FUNcluster = hcut, method = "wss",
                                          k.max = 10) +
  labs(title="Health Elbow method")

health_elbow

# There seems to be 3 bends, k=2, 5, and 6
health_silhouette <- fviz_nbclust(health_scaled, FUNcluster = hcut, method = "silhouette",
                                            k.max = 10) +
  labs(title="Health Silhouette Method")

health_silhouette
```

Checking appropriate number of clusters using elbow method (wss) or silhouette. fviz_nbclust from the factoextra package is the function used to do this. The function is used to visualize and determine the optimal number of clusters in a dataset. FUNcluster is used to determine the type of clustering to be performed. In this case, “hcut” was always used because we were doing hierarchical clustering.

Codes for Performing Clustering

```
# Perform hierarchical clustering using Complete Linkage method
health_hclust_complete <- hclust(dist_health, method = "complete")

# Cut the dendrogram to form 2 clusters
health_cluster_complete_k2 <- cutree(health_hclust_complete, k=2)

# Add cluster labels to the data
health_data_complete_k2 <- health %>% mutate(cluster5_complete_k2 = health_cluster_complete_k2)

# Count the number of items in each cluster
health_data_complete_k2 %>% count(cluster5_complete_k2)
```

hclust: Performs hierarchical clustering using a distance matrix (dist_health) and the "complete" linkage method, which merges clusters based on the maximum distance between points in different clusters.

cutree: Cuts the hierarchical tree (dendrogram) into 2 clusters (k=2), assigning each observation to a cluster based on the dendrogram's structure.

mutate: Adds the cluster assignments (health_cluster_complete_k2) as a new column (cluster5_complete_k2) to the original health dataset, enabling the tracking of which observations belong to which cluster.

count: Counts the number of observations in each cluster, summarizing how many data points belong to each of the two clusters.

Codes For Generating Summary Table

```r
names(health_data_ward_k2)[c(1,2,3)] <- c("health", "Dalc", "Walc")

health_data_ward_k2 %>% group_by(cluster5_ward_k2) %>%
  summarise(across(names(health_data_ward_k2)[c(1,2,3)], ~round(mean(.x, na.rm = TRUE), 1)))
```

The code first renames the specified columns in the dataset, groups the data by cluster, and then calculates and rounds the mean values of the selected columns in each cluster. This allows for easy comparison of averages across the clusters.

Code For Generating Summary Table of Final Clustering

```r
# Function to calculate mode for summarizing clusters
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))
}
```

The get_mode function in R calculates the mode (the most frequent value) of a numeric vector. It first extracts the unique values from the vector using unique(), then creates a frequency table using table() to count the occurrences of each value. The function identifies the mode by finding the maximum count in the frequency table and extracting the values associated with that count. If there are multiple values with the same maximum frequency, all such values are considered modes. Finally, the function returns the mode(s) as numeric values using as.numeric(). This function handles both single and multiple modes effectively.

References

Aloia, L., & Strutzenberg, C. (2020). The influence of family cohesion and relational maintenance strategies on stress in first semester college students. *Communication Quarterly, 68*, 457–471. https://doi.org/10.1080/01463373.2020.1821735.

Blake, J. (2023). Family size and achievement. https://doi.org/10.5860/choice.27-0595.

Bowman-Perrott, L., Davis, H., Vannest, K., Williams, L., Greenwood, C., & Parker, R. (2013). Academic benefits of peer tutoring: A meta-analytic review of single-case research. *School Psychology Review, 42*, 39–55. https://doi.org/10.1080/02796015.2013.12087490.

Briones, V. A. C., Dela Vega, M. J. M., Laureta, M. J. C., Nave, W. A. L., Valdez, M. A. A., & Perfecto, M. G. A. (2019). Impact of poor time management in the academic performance of Grade 12 ABM students in Bestlink College of the Philippines S.Y. 2018-2019. *Ascendens Asia Singapore – Bestlink College of the Philippines Journal of Multidisciplinary Research, 1*(1). Retrieved from https://ojs.aaresearchindex.com/index.php/aasgbcpjmra/article/view/1426.

Cabrera, A. F., & La Nasa, S. M. (2001). On the path to college: Three critical tasks facing America's disadvantaged. *Research in Higher Education*, 119–149.

Cargiulo, T. (2007). Understanding the health impact of alcohol dependence. *American Journal of Health-System Pharmacy, 64*(5 Suppl 3), S5–11. https://doi.org/10.2146/AJHP060647.

Cassum, L. (2018). Impact of divorce on students' life. *Journal of Educational Psychology, 11*, 54–58. https://doi.org/10.26634/jpsy.11.4.14222.

Center for Research in Education and Social Policy. (2018). Chronic absenteeism and its impact on achievement (P18-002.5). Newark, DE: University of Delaware.

Chen, W. (2023). Happy but not happy: The correlation between Chinese high school students' grades and their level of happiness. *Lecture Notes in Education Psychology and Public Media*. https://doi.org/10.54254/2753-7048/6/20220366.

Doerksen, S. E., Elavsky, S., Rebar, A. L., & Conroy, D. E. (2014). Weekly fluctuations in college student leisure activities and well-being. *Leisure Sciences, 36*(1), 14–34. https://doi.org/10.1080/01490400.2014.860778.

Gammon, J., & Morgan-Samuel, H. (2005). A study to ascertain the effect of structured student tutorial support on student stress, self-esteem and coping. *Nurse Education in Practice, 5*(3), 161–171. https://doi.org/10.1016/j.nepr.2004.09.003.

Htet, H., Saw, Y. M., Saw, T. N., Htun, N. M. M., Lay Mon, K., Cho, S. M., et al. (2020). Prevalence of alcohol consumption and its risk factors among university students: A cross-sectional study across six universities in Myanmar. *PLoS ONE, 15*(2), e0229329. https://doi.org/10.1371/journal.pone.0229329.

Idris, D., Hussain, D., & Ahmad, N. (2020). Relationship between parents' education and their children's academic achievement. https://doi.org/10.46662/jass-vol7-iss2-2020(82-92).

Martinez-Yarza, N., Solabarrieta-Eizaguirre, J., & Santibáñez-Gruber, R. (2024). The impact of family involvement on students' social-emotional development: The mediational role of school engagement. *European Journal of Psychology of Education.* https://doi.org/10.1007/s10212-024-00862-1.

Monteiro, M. (2020). The World Health Organization's SAFER initiative and the challenges in Brazil to reducing harmful consumption of alcoholic beverages. *Epidemiologia e Servicos de Saude, 29*(1), e2020000. https://doi.org/10.5123/S1679-49742020000100011.

Naumann, L., Guillaume, E., & Funder, D. (2012). The correlates of high parental academic expectations. *Journal of Cross-Cultural Psychology, 43*, 515–520. https://doi.org/10.1177/0022022112438398.

Nordlander, E., & Stensöta, H. (2014). Grades – for better or worse? The interplay of school performance and subjective well-being among boys and girls. *Child Indicators Research, 7*, 861–879. https://doi.org/10.1007/s12187-014-9233-y.

Shanlax International Journal of Education. (2022). How can we make students happier at school? Parental pressure or support for academic success, educational stress, and school happiness of secondary school students. ERIC. https://eric.ed.gov/?id=EJ1328706.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. https://doi.org/10.3102/00346543075003417.

Takashiro, N., & Clarke, C. H. (2020). Low-socioeconomic status students turn their academic failure to success: A synthesis of qualitative research. In Vanderheiden, E., & Mayer, C. H. (Eds.), *Mistakes, errors and failures across cultures.* Springer. https://doi.org/10.1007/978-3-030-35574-6_19.