Applied Multivariate Analysis

Project #3 - Clustering

Omar Moustafa

900222400

Nour Kahky

900221042

MACT 4233

Spring 2025

Introduction

Complete Source of the Data:

The dataset to be used and analyzed is obtained from Kaggle and originates from the Child Health and Development Studies (CHDS), a long-term research initiative based in California, USA. It contains birth and maternal health information collected as part of a larger initiative to study the vital factors in prenatal and postnatal development. This is achieved by including key variables that describe the health conditions of mothers and their newborns, along with lifestyle factors specifically related to the mothers. Therefore, this dataset provides valuable insights into the relationships between maternal health, lifestyle decisions, and newborn well-being.

What Are the Main Objectives?

The primary objective of this study is to apply several different clustering techniques to a multivariate dataset to reveal the natural patterns and groupings within the data. Specifically, the aim is to implement a comprehensive set of clustering combinations by pairing five different distance measures between the observations and then another five between the clusters. These 25 distance-based variations are then applied to Hierarchical and K-Means Clustering, resulting in over 50 unique clustering results. By implementing this thorough method, the investigation will evaluate the structure and performance of each clustering result to establish the combinations that best capture the existing patterns of the dataset. Overall, the analysis will comprehensively describe the strengths and limitations of each method and offer justifiable recommendations when choosing the most appropriate clustering method based on the characteristics of the data.

Description of the Data

What Are the Observations?

The dataset includes 1236 observations, each representing a single recorded birth in the study. Every observation corresponds to the birth of one baby and encompasses attributes related to the baby's birth, such as birth weight and length of gestation, along with characteristics of the mother, including her age, height, weight, smoking status, and pregnancy history.

Full Definition of the Variables:

Variable Name:	**Variable Description:**	**Variable Type:**	**Unit of Measurement:**
case	ID Number	Unique Identifier	N/A
bwt	Birth weight of the newborn baby	Quantitative	Ounces (oz)
gestation	Length of gestation	Quantitative	Days
parity	Indicator for first pregnancy	Binary	0 = First Pregnancy, 1 = Not First Pregnancy
age	Age of the mother	Quantitative	Years
height	Height of the mother	Quantitative	Inches (in)
weight	Weight of the mother	Quantitative	Pounds (lbs)
smoke	Indicate whether the mother smokes or not	Binary	0 = Non-smoker, 1 = Smoker

Table 1

Data Analysis

Outlier Detection:

Technically speaking, the first step of this study would normally be to detect the existing outliers, where a robust multivariate distance-calculating method, such as Hadi's Distance, is used to fulfill the necessity of considering both the leverage and influence within the data. While it is a powerful methodology for identifying influential observations, implementing it has been relieved of the tasks in this particular context.

Instead, computing five standard pairwise distances between observations—Euclidean, Manhattan, Canberra, Minkowski, and Hamming—each combined with five different linkage methods—Single, Complete, Average, Centroid, and Ward's Method—for Hierarchical Clustering, followed by the K-Means Clustering method, was thoroughly explored.

Hierarchical Clustering:

To identify natural groupings within the dataset, Hierarchical Clustering was applied using 25 combinations of observation-level and cluster-level distance methods.

Initially, five distance measures between the observations were calculated. These five were split into the four that computed observation-level distances for the standardized numeric variables, which were Euclidean, Manhattan, Canberra, and Minkowski, and Hamming's distance did the same but for the categorical variables. Each captures a different aspect of similarity or difference between observations, providing flexibility in the formation of the clustering structures.

Following this, each of these five observation-level distance measures was paired with five different linkage methods, which are Single, Complete, Average, Centroid, and Ward's method. Therefore, this results in 25 unique configurations of Hierarchical Clustering. These 25 configurations were split into 5 groups of 5 dendrograms, allowing for visualizations of how clusters form and where natural breaks take place.

For consistency and easing the upcoming comparisons, each dendrogram was cut into 3 clusters using the **rect.hclust()** function. All of the visual outputs and their corresponding analyses are as follows:

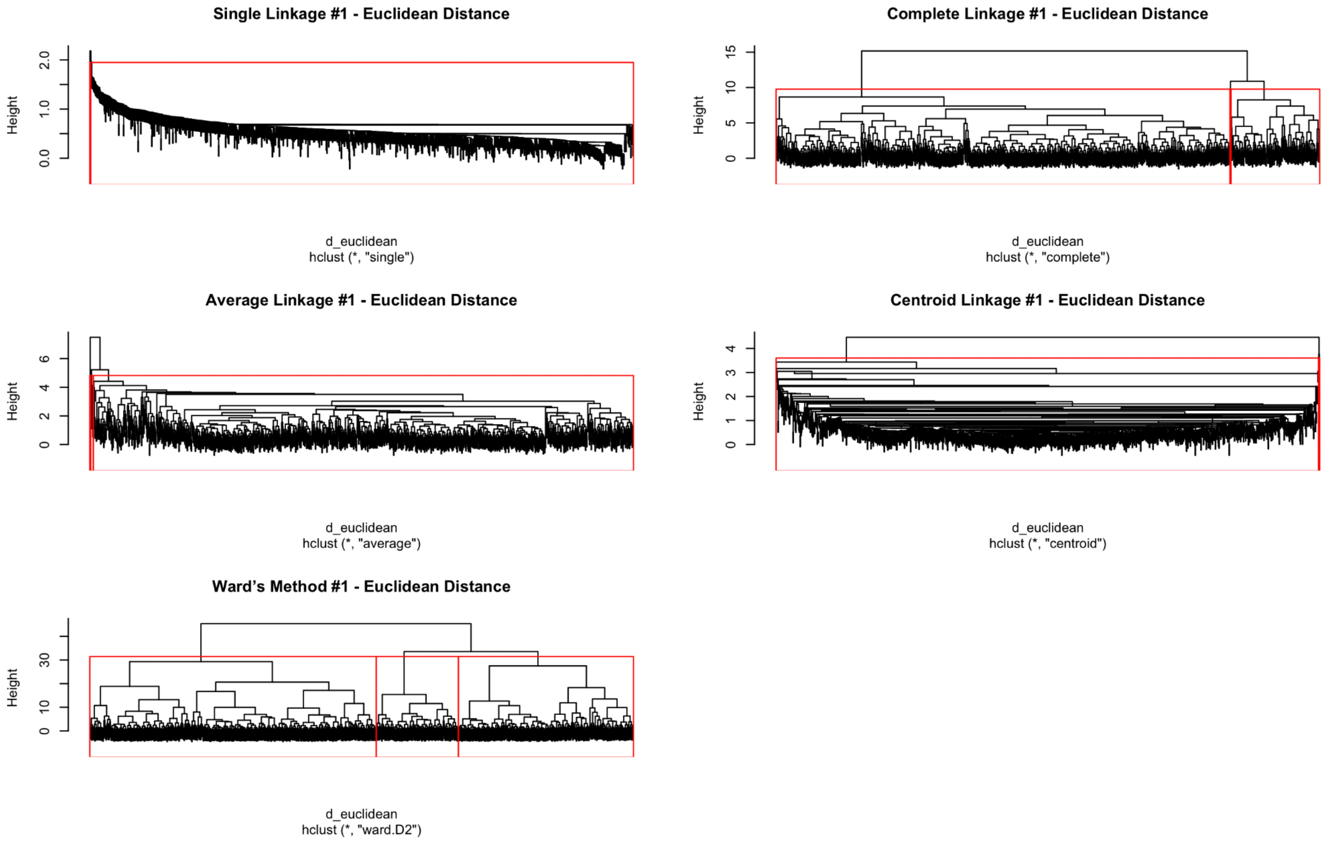


Figure 1

Figure 1 Analysis:

As shown above in Figure 1, which is based on Euclidean distances, well-structured clustering is the case across most linkage methods. Particularly well-structured clustering was accomplished with Complete Linkage and Ward's Method, as they separate the groups with noticeable vertical spacing before the final merges. Not as well-structured, but still decently formed clusters can be seen with Average and Centroid linkages. As for Single Linkage, it shows a gradually changing effect with numerous small-height merges, which is normal for this type of linkage.

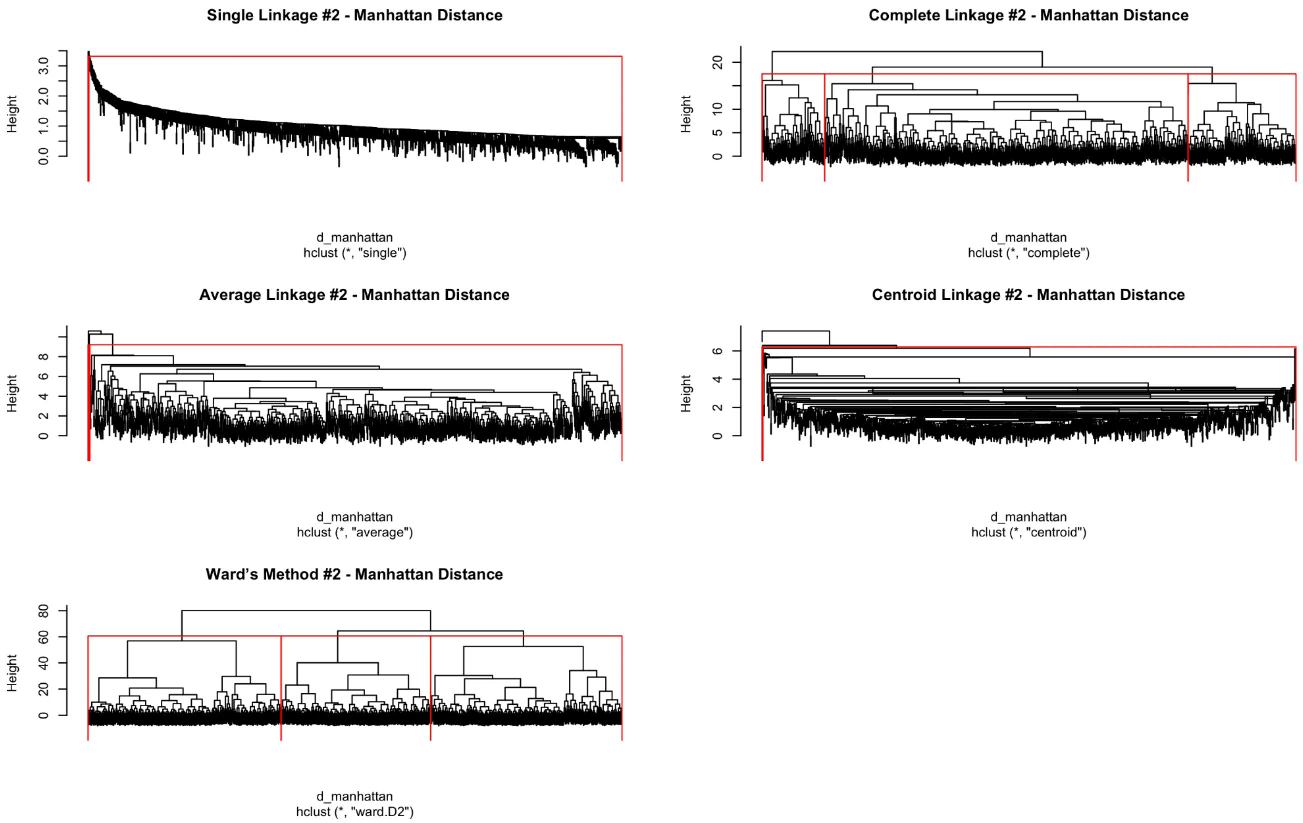
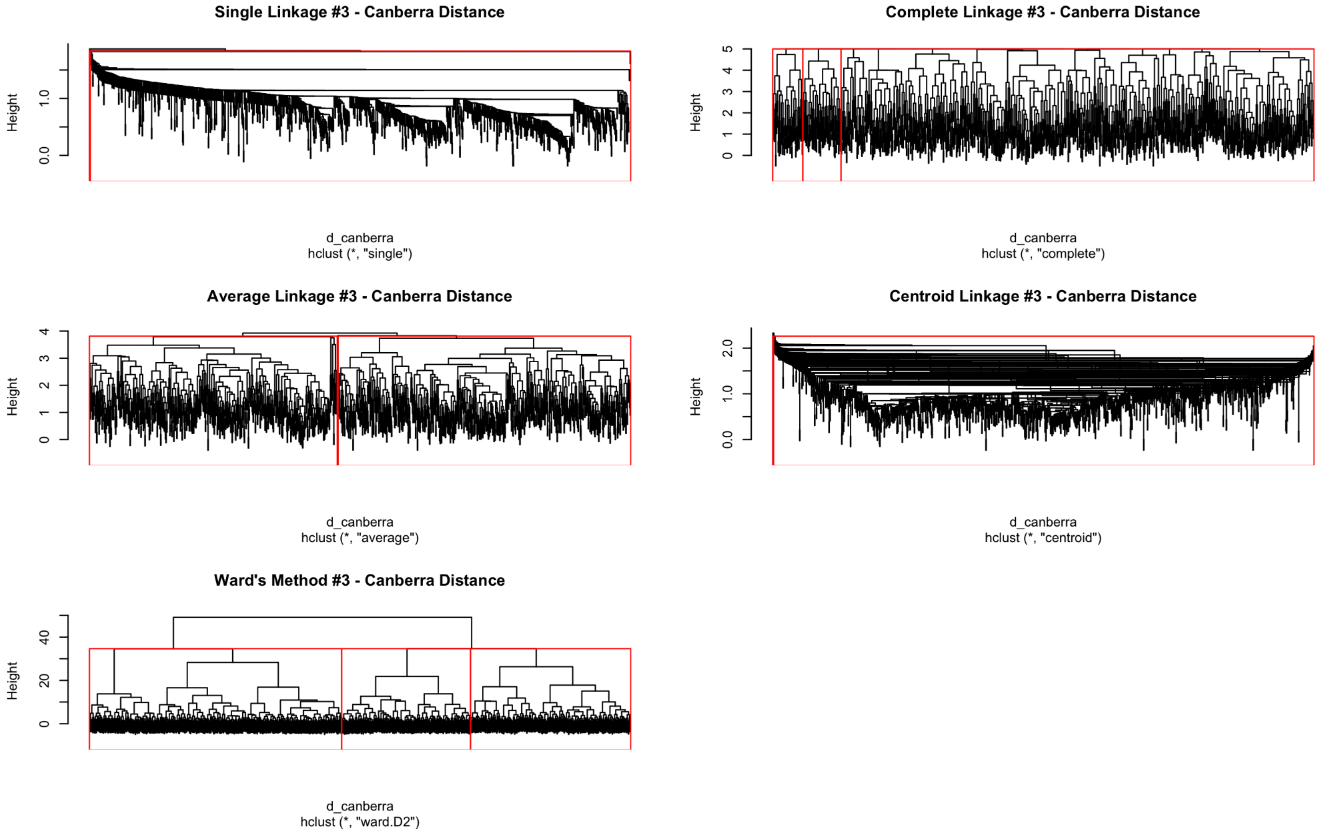


Figure 2

<u>Figure 2 Analysis:</u>

The Manhattan-based dendrograms show similar clustering behavior and characteristics to the Euclidean-based dendrograms. Specifically, Ward's Method again shows highly distinct clusters with precise vertical jumps, indicating an even stronger structure. Complete and Average linkages also look to have performed well, and Single Linkage sticks to its typical behavior of producing a more elongated dendrogram.


Single Linkage #3 - Canberra Distance

Complete Linkage #3 - Canberra Distance

Average Linkage #3 - Canberra Distance

Centroid Linkage #3 - Canberra Distance

Ward's Method #3 - Canberra Distance

Figure 3

<u>Figure 3 Analysis:</u>

The dendrograms above from Canberra distance, displayed above in Figure 3, show much more noise than their two predecessors, and it is especially noticeable with the Single and Centroid linkages. Due to Canberra's distance, stressing smaller values, and being more susceptible to variation within the observations, this outcome was quite expected. Similar to the Euclidean and Manhattan-based dendrograms, Ward's Method continues to produce clean groupings, albeit here in Canberra, there is slightly less vertical separation than in the two previous Ward's Method plots, but still much better than the four other methods when paired with Canberra.

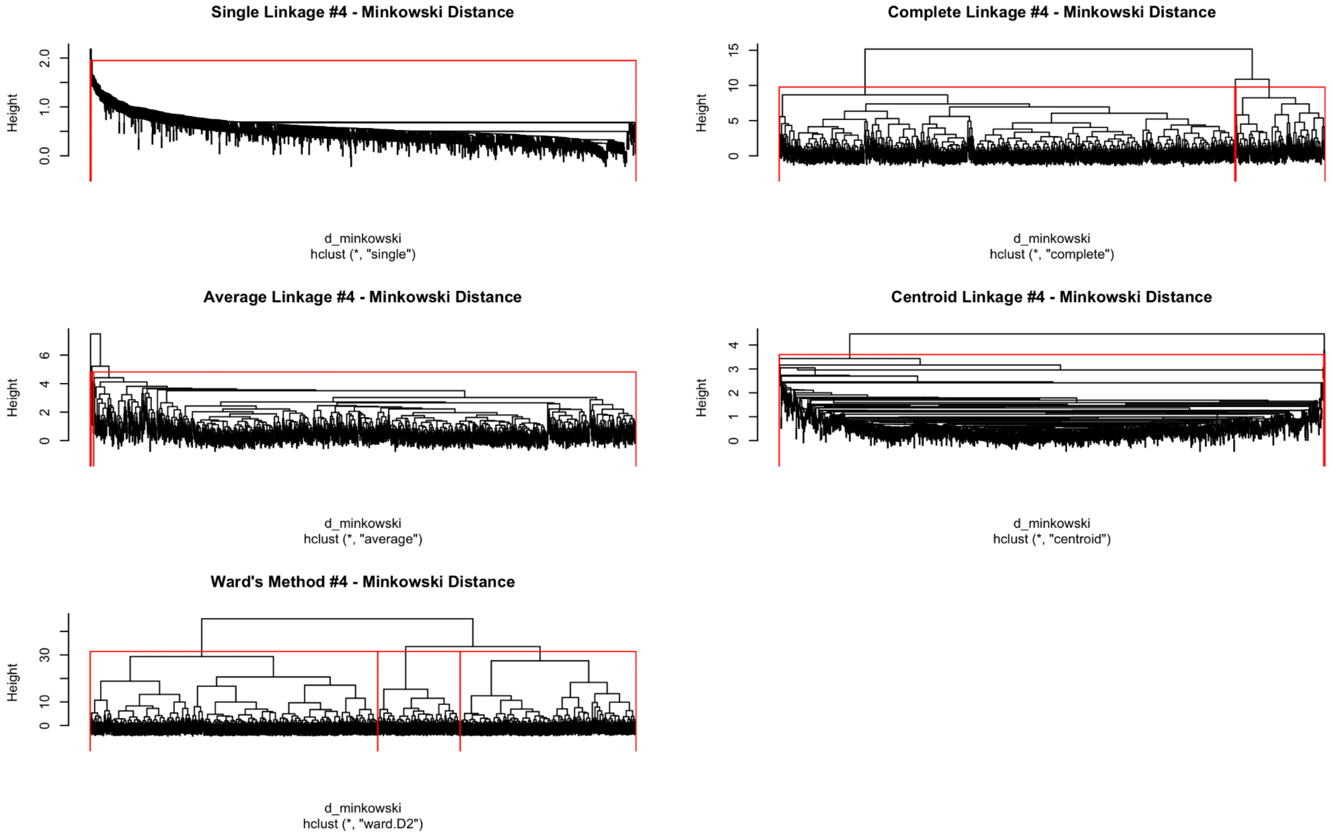


Figure 4

Figure 4 Analysis:

The overall structure and results of the Minkowski distance appear to combine traits from Euclidean and Manhattan distances. This is because Ward's Method, combined with Complete Linkage, once again provides the most accurate and balanced clusters with clear visual distinctions between said clusters. Single Linkage still forms a long and gradual chain, and Centroid Linkage here does show tighter merges. Also, this marks the fourth consecutive time where, no matter the observational-level distance measure, the Single Linkage demonstrates a gradually descending effect with many small-height merges, further supporting the notion that it is the typical behavior of this type of Linkage.

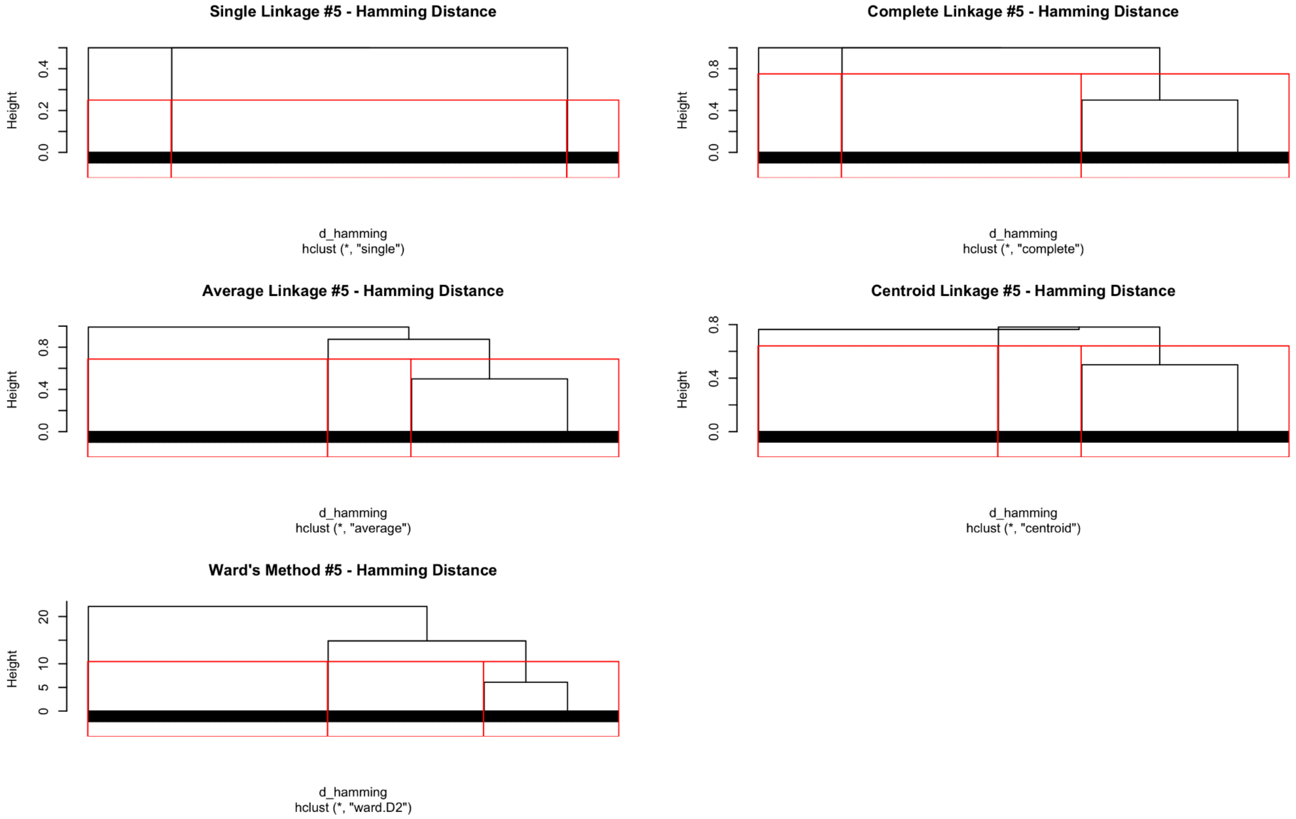


Figure 5

<u>Figure 5 Analysis:</u>

Finally, Figure 5, placed right above, shows that the detriment caused by the limited number of categorical or binary variables within the dataset is that the five Hamming-based dendrograms turn out completely flat and much less defined. Due to this dataset only containing two binary variables, these plots show that there are fewer jumps in height, and several merges take place at similar heights and levels. It can be said that Complete Linkage and Ward's Method did perform slightly better; however, the separations between clusters are significantly weaker than those of the four numeric distances.

Overarching Comparisons Between Figures 1-5:

Altogether, the five Hierarchical Clustering methods reveal quite consistent results and patterns in performance across the different distance and linkage combinations. Euclidean and Manhattan distances stood out as the two strongest performers, producing the most distinct and balanced clusters, especially when combined with Ward's Method and sometimes Complete Linkage. Clear vertical jumps were observed and evidenced in these dendrograms, suggesting significant cluster separation.

Canberra and Minkowski distances introduced more variability in their plots and results, and still performed well when paired with Ward's Method; however, their traits of being sensitive to small values or high variance restrained their clarity in certain linkage combinations. Finally, Hamming's distance, using only two binary categorical variables, provided the least distinct clustering, emphasizing how important dimensionality is when dealing with categorical variables and implementing their specific methods.

Also, it is important to note that Single Linkage constantly produced chaining structures, as in long, gradual merges with minimal vertical changes in height, highlighting that Single Linkage may not be adequate or ideal for this dataset.

K-Means Clustering and the Elbow Method:

Following Hierarchical Clustering, K-Means Clustering was implemented to utilize multiple clustering techniques and provide a direct comparison with the obtained results. Unlike Hierarchical Clustering, which works to construct dendrograms based on linkage trees, K-Means Clustering partitions the data by maximizing the Between-Cluster Sum of Squares (BSS) and minimizing the Within-Cluster Sum of Squares (WSS) through an iterative reassignment of the observations. The implementation and analysis below explore the performance of the K-Means methodology using multiple k-values and evaluating the strength of clustering solutions using the balance of the cluster sizes, WSS (error rate), and R^2 (goodness-of-fit), along with the elbow method.

On the standardized quantitative data, K-Means Clustering was first performed using k = 2 clusters, then k = 3, up to k = 6, allowing for a solid range to evaluate performance. The cluster sizes, WSS, and R^2 values were recorded in the following table:

k	**Cluster Sizes (n)**	**WSS**	**R^2**
2	592 592	4658.97	0.2123
3	442 491 251	4052.16	0.3149
4	268 365 395 156	3581.61	0.3945
5	369 157 300 211 147	3250.60	0.4503
6	195 114 214 271 294 96	3034.71	0.4869

Table 2

Analyzing Table 2, it turns out that, as expected, the R^2 values increased with higher k values, starting from 0.2123 for k = 2 and reaching 0.4869 at k = 6. This increase represents the improved ability of the model when it comes to explaining the variability in the data as more clusters are added. Although the R^2 values were increasing, it has to be said that these increases were steady, as the gains started to diminish after k = 4, indicating that it had reached a point of declining returns. Additionally, it is evident that as k increased, the size of the clusters became more and more uneven, with k = 6 yielding a cluster with as few as 96 observations.

Then, looking at the WSS values, a steady decrease from 4658.97 at k = 2 down to 3034.71 at k = 6 is observed. This downward trend indicates that there is an improvement in the compactness within the clusters; however, this specific improvement also seems to flatten after k = 4. This flattening behavior heavily supports the observation regarding R², further confirming that k = 3 or 4 offers the best trade-off between analysis and execution for this dataset.

Following this, the traditional Elbow Method was implemented to plot the WSS for k values ranging from 2 to 10 to confirm the findings from Table 2, specifically the diminishing returns in both R² and WSS. This visualization, known as the "L-Curve," helps identify the point where additional clusters result in the least amount of error reduction. This is represented by the following plot:

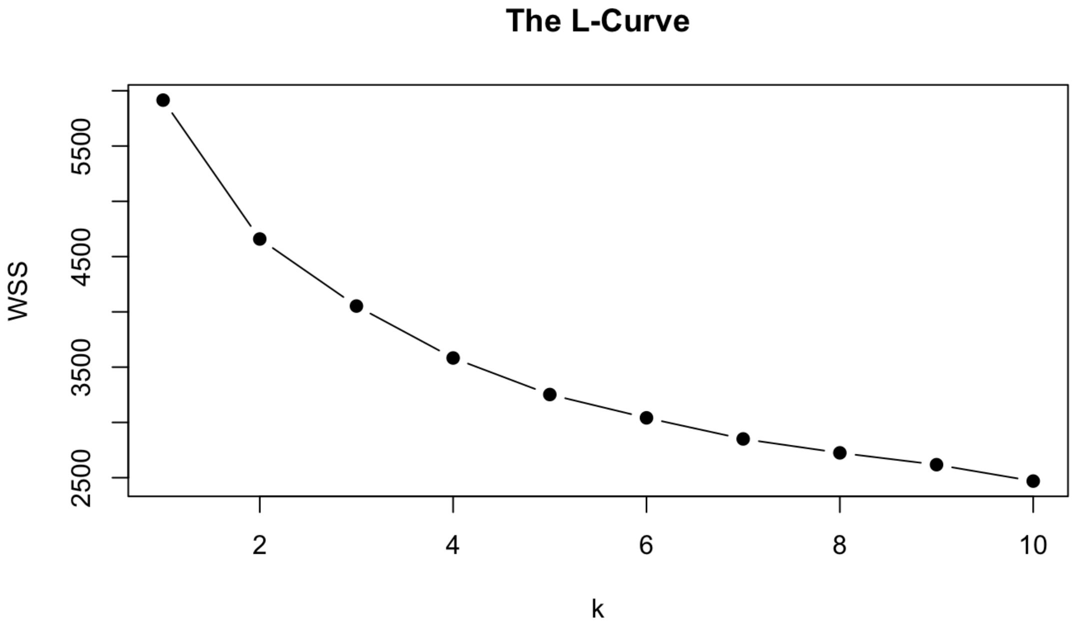


Figure 6

<u>Figure 6 Analysis:</u>

The resulting L-Curve shows a very steep drop in WSS from k = 2 to k = 4, after which the curve begins to flatten. The shape of the plot above indicates that the majority of the clustering benefit is achieved by the time k reaches the value of 4, with very small returns afterwards. Therefore, this heavily reinforces the finding drawn from Table 2, suggesting that k being equal to either 3 or 4 leads to the convenience of having a simple model whose structure within the data is well explained.

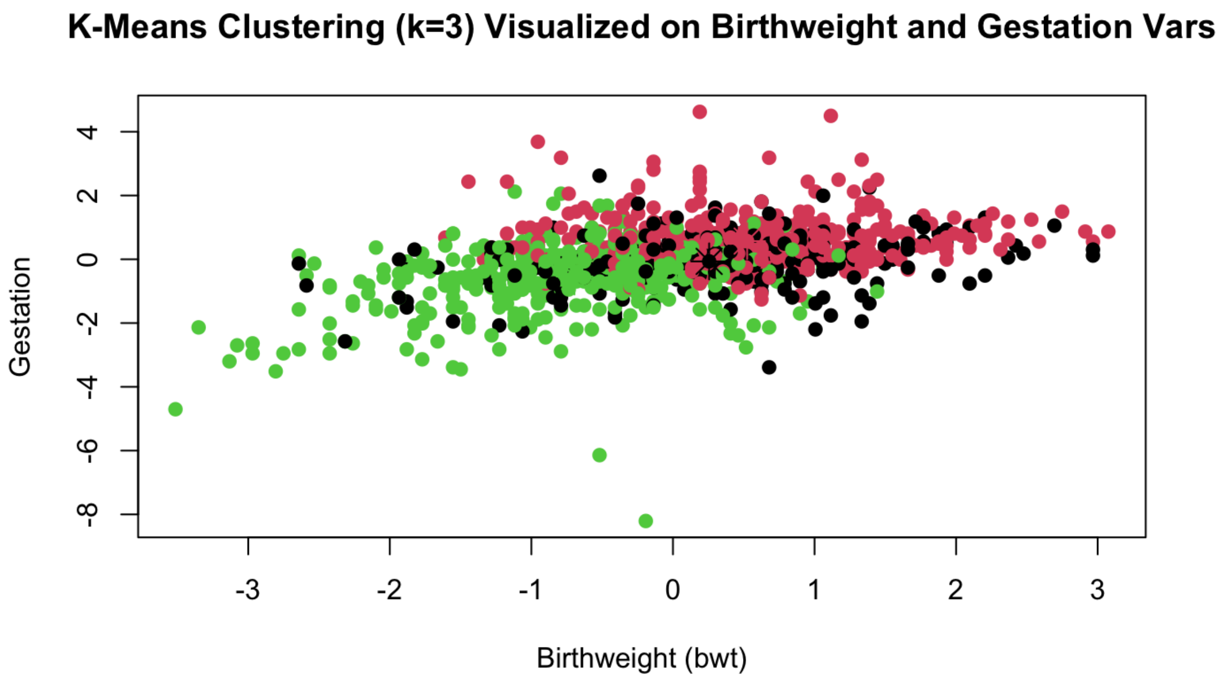


Figure 7

Figure 7 Analysis:

The K-Means result with k = 3 is displayed above as an additional visualization, and it shows the projection onto two standardized quantitative variables: "gestation" and "birthweight (bwt)." This 2D projection provides a simplified view of how the K-Means method works to divide the data into multiple clusters, despite there being five quantitative variables.

The cluster allocations from the K-Means output are represented by the three color-coded groups in this figure. Even though there is substantial overlap, particularly in the middle, there are some patterns that do still stand out. For instance, the red cluster records slightly longer gestation periods, while the green cluster groups observations with lighter initial birth weights and shorter gestation periods.

Additionally, the black cluster seems to contain observations with average values for both birth weights and gestation periods. This suggests that the black cluster is serving as more of a transition zone between the red and green clusters, hence why it is overlapping between the red and green clusters and mainly occupying the center of the plot. Furthermore, the black cluster, horizontally, looks much more spread out, but vertically, it looks relatively constant, indicating more variability in the birth weights but tighter grouping in the gestation periods.

Hierarchical vs. K-Means Clustering:

After performing 25 Hierarchical Clustering combinations and several K-means runs with varying values of k, a meaningful and worthy comparison can be made between the two methods. Hierarchical clustering—especially Ward's Method with Euclidean distance—produced the most visually interpretable and distinct dendrograms with well-separated clusters and consistent performance across different distance measures. It also led to a clearer and more comprehensible understanding of the number of clusters because of the structure and formation of the dendrogram.

K-Means, on the other hand, offered flexibility by allowing multiple runs with varying k-values. Both quantitative evaluations and visualizations were conducted using R^2 scores and the Elbow Method (WSS), respectively, which showed steady increases up until $k = 4$. However, after that point, the R^2 increased very minimally, and the results revealed even more unbalanced cluster sizes.

While the K-Means Clustering method carried out slightly better results, with higher R^2 values, especially at $k = 4$, Ward's Method produced more understandable results and visually appealing clustering without requiring a fixed number of clusters. Therefore, Ward's Method is considered the most suitable choice for this dataset, highlighting its consistent separation and structural balance. Ward’s Method consistently delivered adequate results regardless of the observational-level distance measure it was paired with, which is a trait that cannot be overlooked or undermined.

Conclusion

In conclusion, this study successfully conducted a thorough clustering analysis utilizing both Hierarchical and K-Means methods. A total of 25 hierarchical combinations and various K-Means runs were performed, resulting in over 50 unique clustering outcomes. The analysis revealed that Euclidean and Manhattan distance measures, when combined with Ward’s Method, yield the most well-structured and interpretable results.

The K-Means method further confirmed the strength of these clusters, particularly when k was equal to 3 and 4, shown by the table of R^2 scores, WSS values, and cluster sizes (*Table 2*) as well as the L-Curve plot (*Figure 6*). Although both clustering approaches are useful and have their benefits, Ward’s Method, paired with Euclidean distance, provided the most consistent results with its informative and clear visualizations, along with an admirable level of robustness, making it the most suitable of the bunch.

Overall, this investigation truly emphasizes the importance of implementing various clustering methods and strategies, evaluating both quantitative outputs and visual plots, and using data-driven research methods and reasoning. These steps all play key roles in leading to eventual success and informed decision-making when choosing the most effective clustering approach to understand the natural groupings within multivariate data.

References

Ferretti, Jacopo. *Child Weight at Birth and Gestation Details*. Kaggle. April 20, 2025. https://www.kaggle.com/datasets/jacopoferretti/child-weight-at-birth-and-gestation-details.