UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Statistical Sciences

Natalie Bianca Alexander

The University of Cape Town

The Department of Statistical Sciences

STA5077Z Unsupervised Learning

Assignment 1: Question 1

Due Date: 6 October 2023

Table of Contents

Question 1

Foetal Health Data

Abstract

The aim of this project is to cluster observations of the foetal health data set into its relevant classes of foetal health, namely: “healthy”, “suspect” and “pathological” based on the available features in the foetal health dataset. The objectives include: (1) Clustering the data into three clusters, which performs well across various metrics, and (2) to investigate if three foetal health classes are appropriate. As a result, I find that the K-means clustering algorithm outperforms all other clustering algorithms based on performance metrics such as the average silhouette width, average within cluster Jaccard similarity, average cluster instability, average between cluster dissimilarity and average within cluster dissimilarity. In addition, analysis of the optimal number of clusters confirms that k = 3 clusters is the most representative clustering structure for the foetal health data set.

Introduction

Child mortality is a key indicator of human progress [1]. With that being said, the United Nations' (UN) Sustainable Development Goals emphasize the need to reduce child mortality. The UN forecasts that all countries should be able to end preventable deaths of newborns and children under 5 years of age by the year 2030. To achieve this goal, all countries are encouraged to reduce under-5 mortality to at least 25 per 1,000 live births.

However, one cannot ignore maternal mortality, while investigating child mortality. Maternal mortality accounts for 295 000 deaths during and after pregnancy and childbirth (as of 2017). A large proportion of these deaths (94%) have occurred in low-resource settings, most of which were preventable.

Cardiotocograms (CTGs) have been proposed as a simple and cost-effective solution to assess foetal health so that health professionals (in formal and informal settings) may take the necessary preventative measures to reduce child and maternal mortality. CTGs operate by sending ultrasound pulses and reading its response. As a result, foetal heart rate (FHR), foetal movements, and uterine contractions may be monitored, among other factors. Obstetricians perform these cardiotocogram exams, and the results may assist in the early diagnosis of foetal anomalies. Based on the CTG results, foetal health can be classified into three main classes, namely: “healthy”, “suspect” and “pathological”.

This project aims to assess data obtained by cardiotocogram exams, cluster foetal health into its relevant classes and investigate these clusters. Insight into these clusters may assist in the reduction of foetal and maternal mortality.

Data

The foetal health data set was obtained in CSV format on the VULA site. The data set may also be obtained from the Kaggle database [1].

Exploratory Data Analysis (EDA)

A. Checking the data

The dimensions of the data was assessed using the dim() function, which suggested that this is a 2126 row by 21 column dataset. The ncol() and nrow() functions confirmed that there are 2126 rows and 21 columns in the dataset. I then checked for missing data, and found that there were no rows with missing data.

Table 1 in **Appendix A** shows the column names of the independent variables, which was assessed us ing the colnames() function. The description of each independent variable is also provided [1]. Thereafter, the column names were formatted by capitalizing column names and removing underscores.

Table 2-4 in the Appendix A shows the first 6 rows across all available features in the foetal health data set. **Table 5-7 in the Appendix A** shows the last 6 rows across all available features in the foetal health data set. It is evident that the foetal health data set is a high dimensional dataset, where the features have different ranges, scales, and units.

Thereafter, it was necessary to assess the data types of each column (also known as the independent variables). All independent variables were numeric. Based on the description of the independent variables in **Table 1** in **Appendix A**, there were no categorical or binary independent variables present in the data set that needed to be converted to factors.

B. Summary statistics

Table 8 below shows the summary statistics for each independent variable, where summary statistics such as the minimum value, first quartile, median, mean, third quartile and maximum value are provided. Based on these summary statistics, it is evident that the minimum and maximum values differ across the independent variables, which suggest variation in the range across these independent variables. Differences in the range among the independent variables support the concept of dissimilar distributions among these independent variables. For example, the “Baseline” variable has a narrow range, where the minimum value = 106 and the maximum value = 160. On the other hand, the “Histogram width” variable, has a wider range, where the minimum value = 3 and the maximum value = 180.

It is also apparent that the means and medians differ among the independent variables, which further suggest differences in the underlying distributions for each independent variable. For example, the “Abnormal short-term variability” variable has a mean value of 46.990 which is smaller than its median value of 49, suggesting a left-skewed distribution. On the other hand, the “Percentage of time with abnormal long-term variability” variable has a mean value of 9.847, which is larger than the median value of 0, suggesting a right-skewed distribution.

Finally, there are also differences among the quartile 1 and quartile 3 values for each of the independent variables, suggesting differences in the interquartile range (IQR = Q3 – Q1) among the independent variables. Certain independent variables have a small interquartile range (i.e., a small distance between quartile 1 and quartile 3), which suggests a narrower spread of observations. In contrast, certain independent variables have a large interquartile range (i.e., a large distance between quartile 1 and quartile 3), which suggests a greater spread of observations. For example, the “Uterine contractions” variable has a Q1-value of 0.002 which is smaller than its Q3-value of 0.007, suggesting an IQR = 0.005. On the other hand, the “Severe decelerations” variable has its Q1 = Q3 = 0, and so the IQR = 0.

Table 8: Summary statistics of the foetal health dataset, showing the minimum value, first quartile, median, mean, third quartile and maximum value.

Independent variable	Minimum value	Quartile 1	Median	Mean	Quartile 3	Maximum value
Baseline value	106	126	133	133.304	140	160
Accelerations	0	0	0.002	0.003	0.006	0.019
Foetal movement	0	0	0	0.009	0.003	0.481
Uterine contractions	0	0.002	0.004	0.004	0.007	0.015
Light decelerations	0	0	0	0.002	0.003	0.015
Severe decelerations	0	0	0	0	0	0.001
Prolongued decelerations	0	0	0	0	0	0.005
Abnormal short-term variability	12	32	49	46.990	61	87
Mean value of short-term variability	0.200	0.700	1.200	1.333	1.700	7
Percentage of time with abnormal long-term variability	0	0	0	9.847	11	91
Mean value of long-term variability	0	4.600	7.400	8.188	10.800	50.700
Histogram width	3	37	67.500	70.446	100	180
Histogram min	50	67	93	93.579	120	159
Histogram max	122	152	162	164.025	174	238
Histogram number of peaks	0	2	3	4.068	6	18
Histogram number of zeroes	0.0	0	0	0.324	0	10
Histogram mode	60	129	139	137.452	148	187
Histogram mean	73	125	136	134.611	145	182
Histogram median	77	129	139	138.090	148	186
Histogram variance	0	2	7	18.808	24	269
Histogram tendency	-1	0	0	0.320	1	1

For ease of formatting, **Table 9** shows the variance and standard deviation for each independent variable. Dark green cells indicate independent variables with relatively large variance and standard deviation values. Lighter green cells indicate independent variables with moderately large variance and standard deviation values. Dark orange cells indicate independent variables with relatively small variance and standard deviation values. Lighter orange cells indicate independent variables with moderately small variance and standard deviation values.

The independent variables with relatively large variance suggest that on average, the observations are far from the mean for this independent variable. In contrast, the independent variables with small variance suggest that on average, the observations are close to the mean of this independent variable.

I also assessed the standard deviation for a more representative and comparable measure of the average distance of observations to the mean. The standard deviation unit of measurement is the same as that of the independent variable in question. The trend shown in **Table 9** for the standard deviation is the same as that of the variance values.

Certain independent variables such as "Histogram width" value has a large variance of 1517.546 and a large standard deviation of 38.956. On the other hand, certain independent variables such as "Mean value of long-term variability" has a moderately large variance of 31.677 and a moderately large standard deviation of 5.628. In contrast, certain independent variables such as "Accelerations" has a relatively small variance of 0 and standard deviation of 0.004, while other independent variables have a moderately small variance and standard deviation such as "Mean value of short-term variability" with variance = 0.78 and a standard deviation = 0.883.

Table 9: Variance and standard deviation of the foetal health dataset.

Independent variable	Variance*	Standard deviation*
Baseline value	96.842	9.841
Accelerations	0	0.004
Foetal movement	0.002	0.047
Uterine contractions	0	0.003
Light decelerations	0	0.003
Severe decelerations	0	0
Prolongued decelerations	0	0.001
Abnormal short-term variability	295.593	17.193
Mean value of short-term variability	0.78	0.883
Percentage of time with abnormal long-term variability	338.445	18.397
Mean value of long-term variability	31.677	5.628
Histogram width	1517.546	38.956
Histogram min	873.806	29.56
Histogram max	321.994	17.944
Histogram number of peaks	8.699	2.949
Histogram number of zeroes	0.499	0.706
Histogram mode	268.347	16.381
Histogram mean	243.16	15.594
Histogram median	209.282	14.467
Histogram variance	839.703	28.978
Histogram tendency	0.37	0.611

*Dark green cells suggest that the independent variable has a relatively large variance and standard deviation value. Lighter green cells indicate independent variables with moderately large variance and standard deviation values. Dark orange cells indicate independent variables with relatively small variance and standard deviation values. Lighter orange cells indicate independent variables with moderately small variance and standard deviation values.

C. *Visualizing the distribution of the non-histogram independent variables*

Figure 1 below shows the box plots for each non-histogram independent variable. **Figure 2** shows the density plots for each non-histogram independent variable. Below, I discuss in detail, the distribution of each non-histogram independent variable.

In **Figure 1,** the Abnormal short-term variability box plot has a left-skewed distribution, which is supported by **Table 8**, where the mean of 46.990 is less than the median of 49. The density plot in **Figure 2** further suggests a bimodal distribution.

In **Figure 1,** the Accelerations boxplot has a right skew, which is supported by **Table 8**, where the mean (0.003) is greater than the median (0.002). The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1**, the Baseline value boxplot has an approximate symmetric shape, which is supported by **Table 8**, where the mean (133.304) is similar to that of the median (133). The density plot in **Figure 2** further suggests a symmetric, unimodal distribution, where most values are centred at the mean.

In **Figure 1,** the Foetal movement boxplot has a right skew, which is supported by **Table 8**, where the mean (0.009) is greater than the median (0). The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Light decelerations boxplot has a right skew, which is supported by **Table 8**, where the mean (0.002) is greater than the median (0). The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Mean value of long-term variability boxplot has a right skew, which is supported by **Table 8**, where the mean (8.188) is greater than the median (7.400). The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Mean value of short-term variability boxplot has a right skew, which is supported by **Table 8**, where the mean (1.333) is greater than the median (1.200). The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Percentage of time with abnormal long-term variability boxplot has a right skew, which is supported by **Table 8**, where the mean (9.847) is greater than the median (0). The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Prolonged decelerations boxplot has a right skew due to the presence of outliers However, Table **8**, suggests that the mean and median are both approximately 0 values. The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Severe decelerations boxplot has a right skew due to the presence of outliers. However, **Table 8**, suggests that the mean and median are both approximately 0 values. The density plot in **Figure 2** further suggests a right-skewed distribution.

In **Figure 1,** the Uterine contractions has a symmetric shape which is supported by **Table 8** where the mean = median = 0.004. The density plot in **Figure 2** suggests a bimodal distribution.

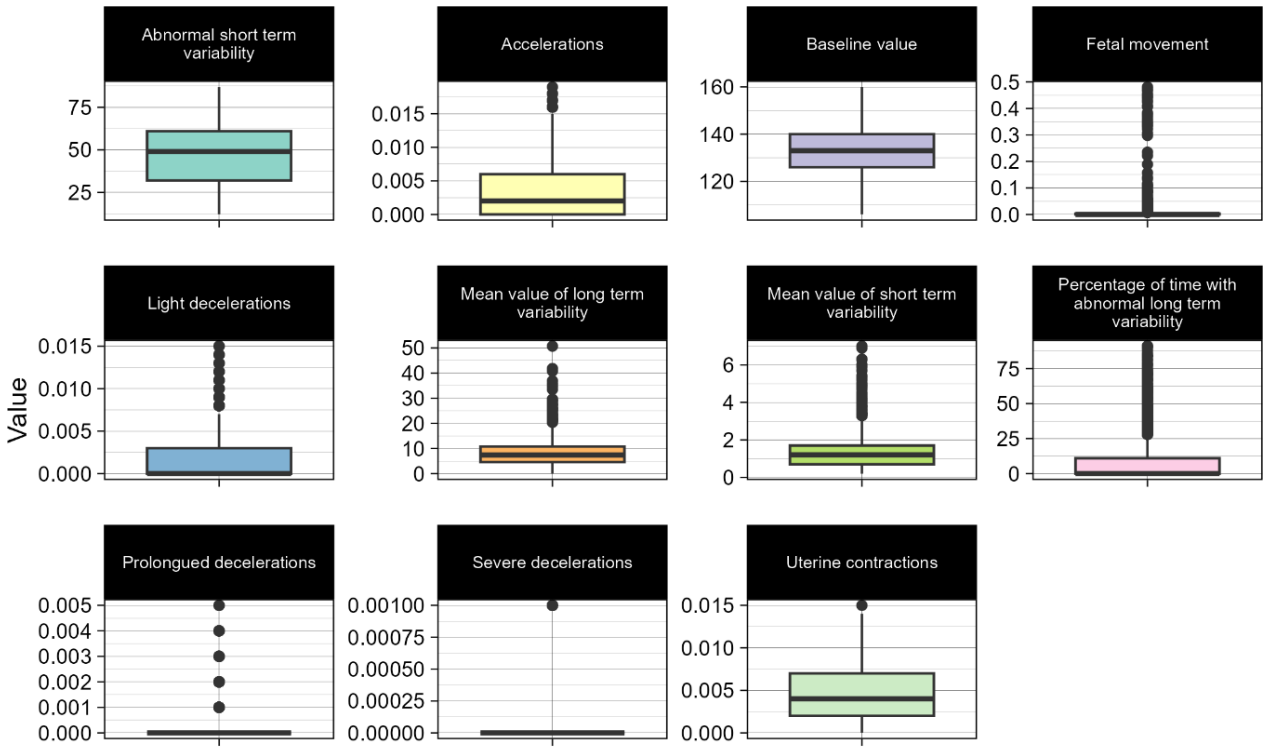

Figure 1: Boxplots showing distribution of non-histogram independent variables.

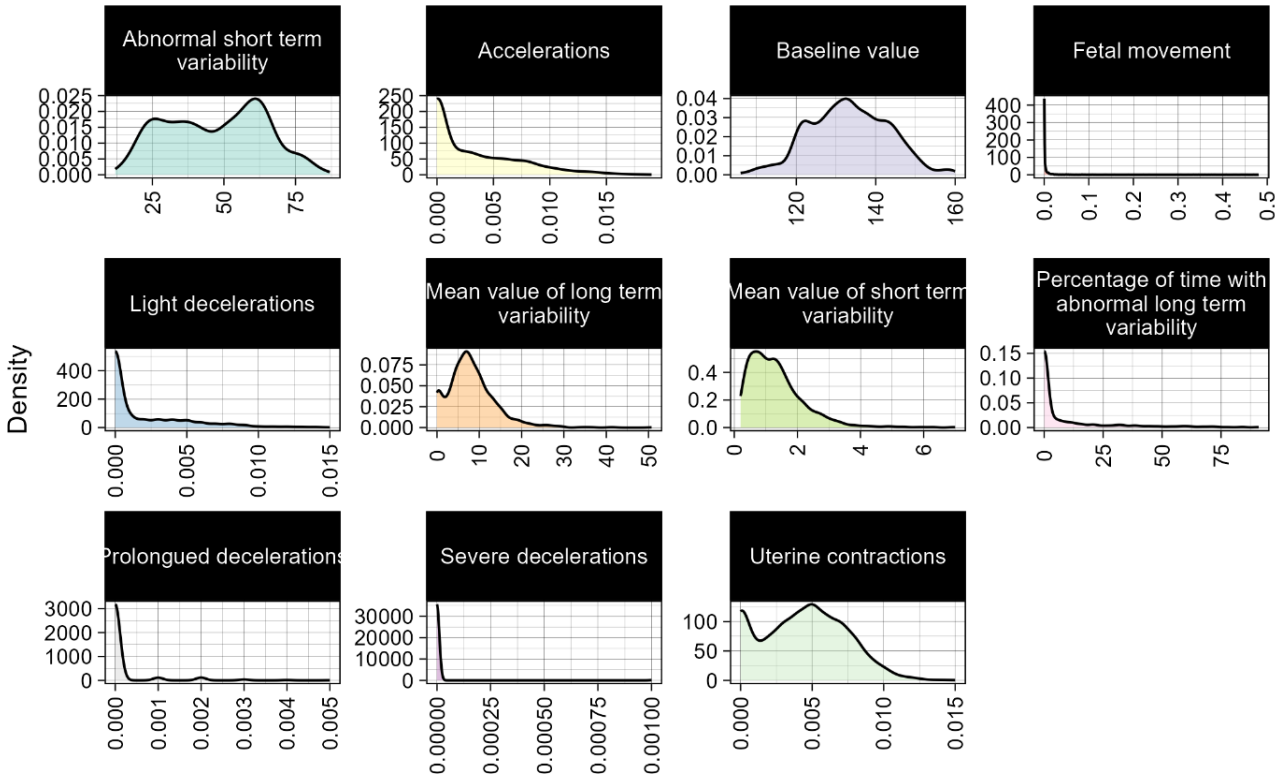

Figure 2: Density plots showing distribution of non-histogram independent variables.

D. Visualizing the distribution of the histogram independent variables

Figure 3 below shows the box plots for each histogram independent variable. **Figure 4** shows the density plots for each histogram independent variable. Below, I discuss in detail, the distribution of each histogram independent variable.

In **Figure 3**, we see the Histogram max box plot has a right skew, which is supported by Table 8, where the mean (164.025) is greater than the median (162). The density plot in **Figure 4** further suggests a right-skewed distribution.

In **Figure 3,** we see the Histogram mean boxplot has a left skew, which is supported by **Table 8**, where the mean (134.611) is less than the median (136). The density plot in **Figure 4** further suggests a left-skewed distribution.

In **Figure 3,** we see the Histogram median boxplot has a left skew, which is supported by **Table 8**, where the mean (138.090) is less than the median (139). The density plot in **Figure 4** further suggests a left-skewed distribution.

In **Figure 3,** we see the Histogram min boxplot has an approximate symmetric shape, which is supported by **Table 8**, where the mean (93.579) is approximately equal to the median (93). The density plot in **Figure 4** suggests a bimodal distribution.

In **Figure 3,** we see the Histogram mode boxplot has a left skew, which is supported by **Table 8**, where the mean (137.452) is less than the median (139). The density plot in **Figure 4** further suggests a left-skewed distribution.

In **Figure 3,** we see the Histogram number of peaks boxplot has a right skew, which is supported by **Table 8**, where the mean (4.068) is greater than the median (3). The density plot in **Figure 4** further suggests a right-skewed distribution.

In **Figure 3,** we see the Histogram number of zeroes boxplot has a right skew, which is supported by **Table 8**, where the mean (0.324) is greater than the median (0). The density plot in **Figure 4** further suggests a right-skewed distribution.

In **Figure 3,** we see the Histogram tendency boxplot has a right skew, which is supported by **Table 8**, where the mean (0.320) is greater than the median (0). The density plot in **Figure 4** suggests a multi-modal distribution.

In **Figure 3,** we see the Histogram variance boxplot has a right skew, which is supported by **Table 8**, where the mean (18.808) is greater than the median (7). The density plot in **Figure 4** further suggests a right-skewed distribution.

In **Figure 3,** we see the Histogram width boxplot has a right skew, which is supported by **Table 8**, where the mean (70.446) is greater than the median (67.500). The density plot in **Figure 4** suggests a bimodal distribution.

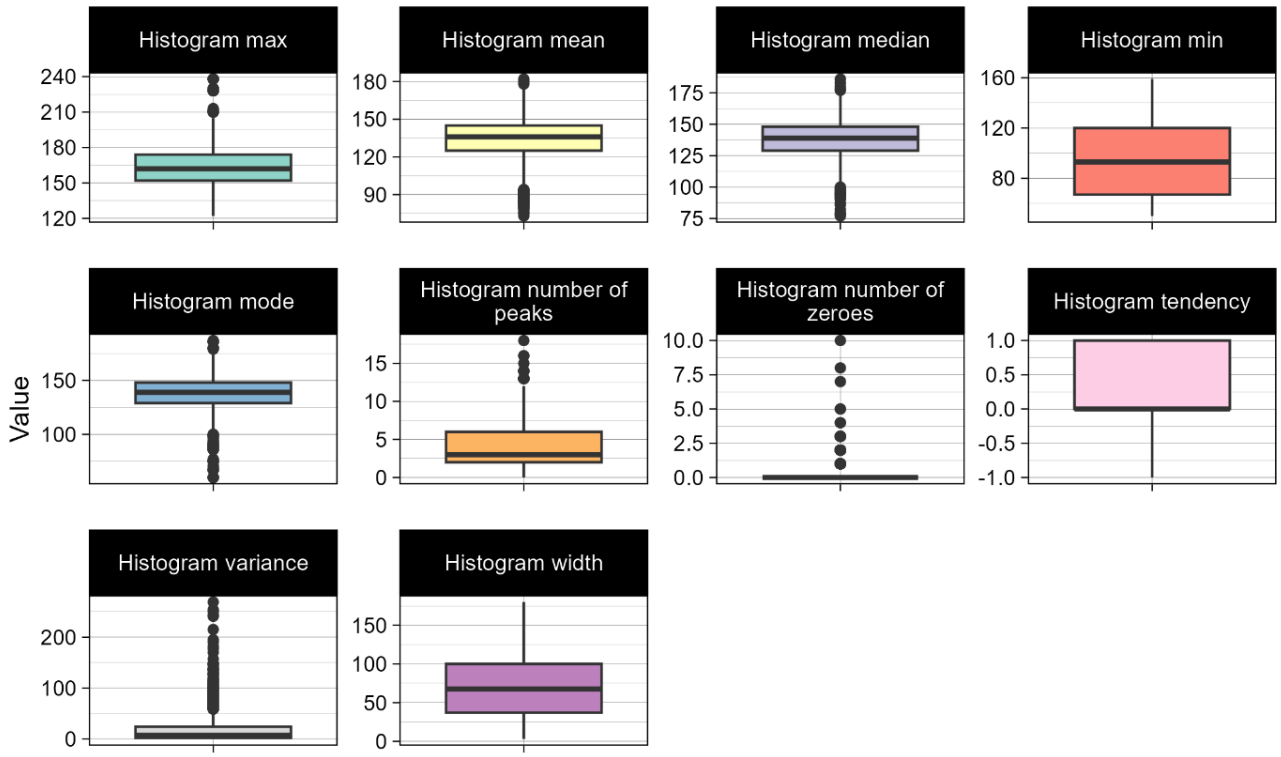

Figure 3: Boxplots showing the distribution of the histogram-related independent variables.

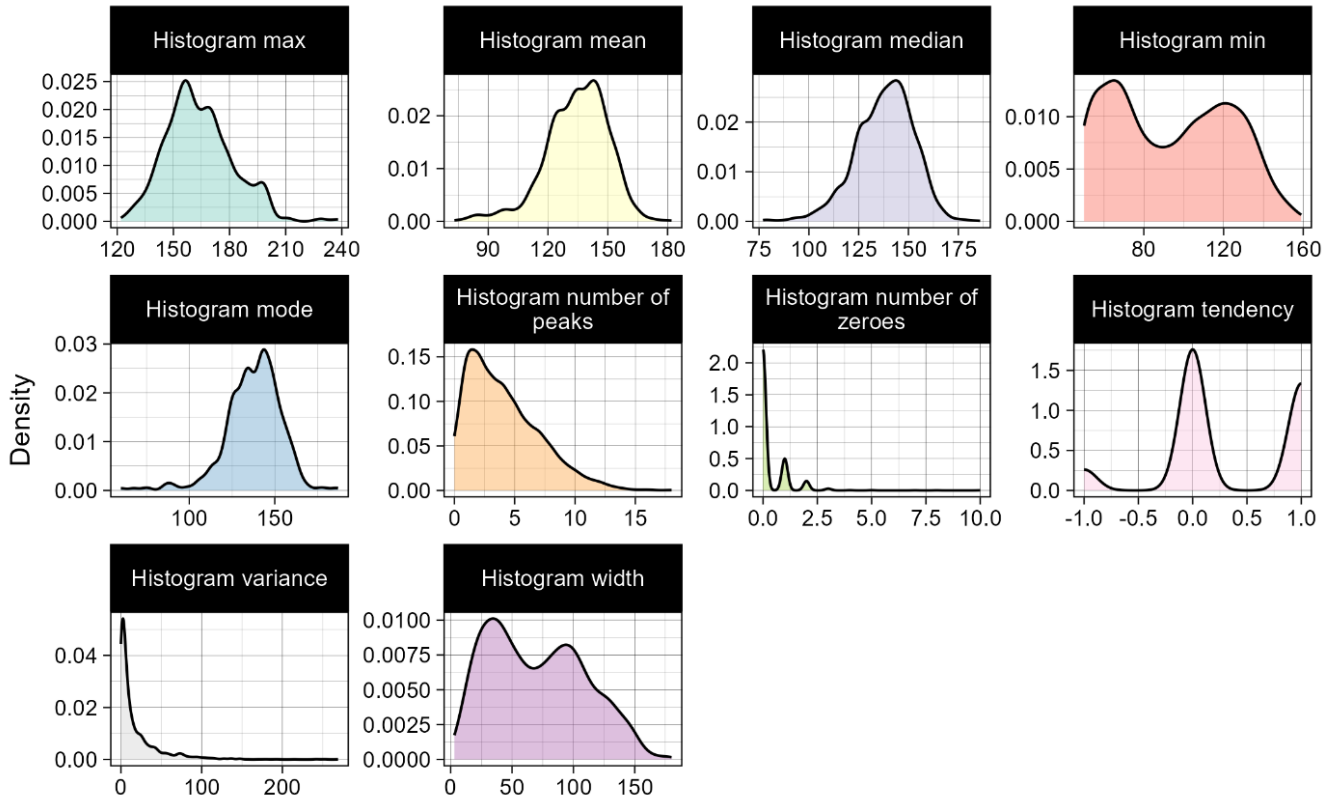

Figure 4: Density plots showing the distribution of the histogram-related independent variables.

E. Correlation

Figure 5 shows the pairwise correlation between the independent variables. The bottom triangle of the correlation matrix is assessed since the correlation matrix is symmetric in shape. As expected, the diagonal has a Pearson correlation coefficient = 1 (shown in dark blue), suggesting perfect relationship between an independent variable and itself. In addition, we see a strong, positive Pearson-correlation coefficient between the following pairs of independent variables:

- Histogram mode and Baseline value.
- Histogram mean and Baseline value.
- Histogram median and Baseline value.
- Histogram mode and Histogram mean.
- Histogram mode and Histogram median.
- Histogram median and Histogram mean.
- Histogram width and Histogram max.
- Histogram width and Histogram number of peaks.
- Histogram width and Histogram variance.

We also see a strong, negative Pearson-correlation coefficient between the following pairs of independent variables:

- Histogram width and Histogram min
- Histogram number of peaks and Histogram min
- Mean value of short-term variability and Histogram min
- Light decelerations and Histogram min

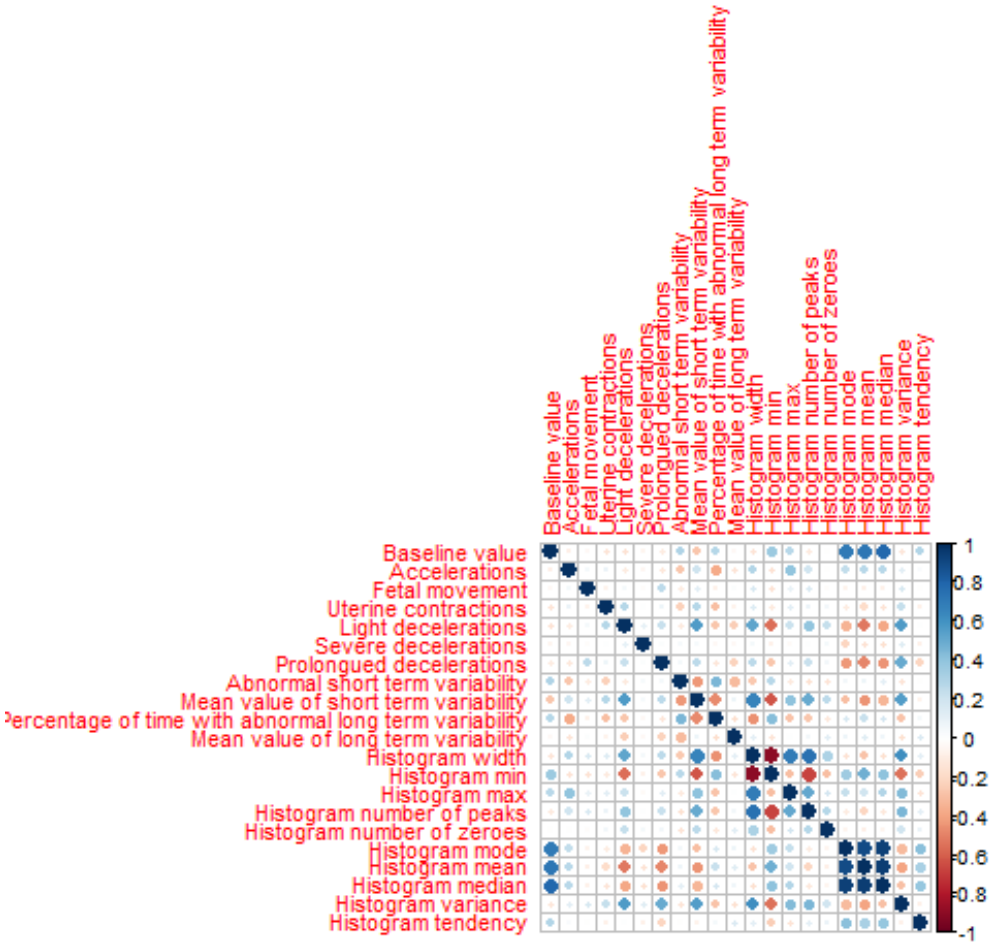

Figure 5: Correlation plot showing pairwise Pearson-correlation between independent variables. The Pearson correlation metric is shown along the right, where values close to -1 are shown in dark red and values close to +1 are shown in dark blue.

Methods

A. *Removing correlated independent variables.*

Based on the correlation plot in **Figure 5** above, it is recommended that one of the correlated variables within a pairwise correlation is removed. In this way, redundancy is removed from the dataset. As a result, the following variables were removed:

- "Histogram mode", "Histogram median", and "Histogram mean" variables were removed since these variables are highly correlated with each other and with the "Baseline value" variable. The "Baseline value" variable remained in the data set, since it provides non-redundant information to the cluster analysis.
- "Histogram width" was also removed since it is highly correlated with "Histogram max", "Histogram number of peaks", "Histogram variance" and "Histogram min".
- "Histogram tendency" was also removed since it provides redundant information about the histogram, which can be inferred by other variables.

As a result, 16 independent variables remained, after removing highly correlated variables.

B. *Data standardization*

The boxplots in **Figure 1** and **Figure 3** and the density plots in **Figure 2** and **Figure 4** show evidence of varying distributions across the independent variables. In parallel, **Table 8** indicates the independent variables have different mean and median values. **Table 8** also suggests that the independent variables differ in terms of the spread of the data, where the minimum and maximum values vary considerably between the independent variables (i.e., differences in the range). **Table 9** supports this idea, where the variance and standard deviation differ among the independent variables. Furthermore, the descriptions of the independent variables in **Table 1: Appendix A** suggest that the unit of measurement differs among the independent variables. For this reason, standardization of the data is necessary.

The data are standardized using the Z-score normalization, so that observations range between 0 and 1. Normalization is necessary to prevent bias in the cluster assignment, where bias exists for extreme values. In addition to this, normalization balances the weight of impact of the observations on the distance calculations.

In order to standardize the data, the scale() function was used in r, where both center = True and scale = True, to ensure that the data are both centred and scaled.

C. *Principal component analysis*

The data set still suffers from high dimensionality, albeit removing highly correlated variables. I am left with 16 independent variables after removing highly correlated variables. Thus, it was necessary to perform dimensionality reduction [2,3].

Principal component analysis (PCA) was performed on the correlation matrix of the scaled and centred data using the princomp() function. The correlation matrix here, represents the correlation between the observations and not between the independent variables. The loadings of the PCA score matrix was then determined and the eigenvalues obtained by squaring the standard deviation of each component. The eigenvalues are important because they represent the proportion of variance in the data that each component explains.

Next, it was necessary to compute the PCA scores as recommended by E. Kaloyanova [2].

To determine the optimal number of principal components, a scree plot was generated, representing the number of components along the x-axis and the eigenvalues along the y-axis. The elbow point was chosen as the optimal number of principal components. In addition, I computed the cumulative proportion of variance explained by the principal components. Initially, I chose the number of

principal components that explained 80% of cumulative variance in the data. However, this method produced suboptimal results, and so the elbow-method was preferred going forward.

The results of the principal component analysis can be seen under the ***Results*** section: ***A. Principal component* results**.

D. Proximity measures

The pairwise dissimilarity between the observations were determined by using the dist() function with the following inputs: (1) the PCA scores obtained by the optimal number of principal components and (2) the distance measure, namely: Euclidean distance, Manhattan distance, and the Chebyshev Maximum distance measure.

E. Agglomerative hierarchical clustering algorithms

Hierarchical clustering was performed using the hclust() function, taking the following parameters as input: (1) the pairwise dissimilarity matrix of the PCA scores for the first few principal components obtained above in ***D. Proximity measures*** and (2) the linkage method, namely: complete linkage, single linkage, average linkage, centroid linkage, median linkage, Ward D and Ward D2 algorithms. Models 1 to 21 were constructed taking in different combinations of dissimilarity matrices and linkage methods.

F. Partitioning clustering algorithms

The K-means clustering algorithm was implemented by using the kmeans() function with the following parameters as input: (1) centers = 3 cluster centres, (2) nstart = 100 random initial starting centres to be chosen and (3) iter.max = a maximum of 1000 iterations.

The K-Medoids clustering algorithm was also investigated. The PAM clustering algorithm was implemented by using pam() with k = 3 clusters for the Euclidean and Manhattan distance matrices, respectively. The CLARA clustering algorithm was also implemented using clara() with k = 3 clusters for the Euclidean and Manhattan distance matrices, respectively.

G. Objective 1: Model comparison and choice of best clustering solution

Models were assessed based on the clustering performance and the visualization of reliable clusters. The performance metrics computed and compared are the cophenetic correlation coefficient (for agglomerative hierarchical clustering models), the average silhouette width across all observations, the gap-statistic, the average within cluster dissimilarity and the average between cluster dissimilarity. In addition, I computed the average within cluster Jaccard similarity for 1000 bootstrap samples and the average instability for 1000 bootstrap samples. By bootstrapping with replacement, I can test cluster adequacy to determine if the clusters identified are valid representations of the true clusters within this data set.

Results of the models can be viewed in the ***Results*** section under ***B. Comparison of model performance.*** Below each of the performance metrics are discussed in detail, their relevance and how they were used as a criterion for assessing the clustering models.

i. Cophenetic corelation coefficient

The cophenetic correlation coefficient for the agglomerative hierarchical models was considered as a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data [4]. A cophenetic correlation measure close to 1 suggests that the dendrogram does a good job at preserving the pairwise distances between the original observations, while a cophenetic correlation coefficient value close to 0 infers that the dendrogram does not do a good job at preserving the pairwise distances between the original observations.

The hierarchical clustering models were filtered by discarding models with a cophenetic correlation coefficient < 0.6, suggesting that these models do a relatively poor job at preserving the pairwise distances between the original observations. A threshold of 0.75 and above is preferred, however due to the range of values obtained for these models, the threshold was relaxed to 0.6.

ii. Average silhouette width

The remaining models which passed the first filter were assessed by means of their average silhouette width to test the goodness of fit for each model [5]. The average silhouette width ranges from -1 to 1, that is from poor fit to good fit. An average silhouette width close to -1 suggests that the observations have incorrect cluster membership, while an average silhouette width close to 0 suggest that the observations are on the boundary between two clusters and an average silhouette width of 1 suggests correct cluster membership. Clustering models with an average silhouette width < 0.400 were discarded, since the best models had an average silhouette width around 0.400.

iii. Gap statistic

The gap statistic for the remaining models were then assessed. The gap statistic estimates the optimal number of clusters (k) by comparing the total within cluster variation for different values of k to the expected values for a distribution with no apparent clustering [6]. The optimal number of clusters k is the point that maximizes the gap statistic. Thus, by comparing the remaining models, I accept models having a relatively large gap statistic at k = 3 clusters.

iv. Average within cluster Jaccard similarity and average cluster instability

Next the average within cluster Jaccard similarity and the average instability for each model was assessed to determine whether the groupings found are a valid representation of cohesive clusters in the data set. For this reason, bootstrapping was performed, where 1000 samples were randomly resampled with replacement and the clustering algorithm is executed at each iteration.
The average Jaccard similarity reflects the frequency with which the data points in each cluster co-exist at each iteration. A large average Jaccard similarity suggests that the cluster membership is the same most of the time.
On the other hand, the average instability of the clusters measures how stable clusters are across all bootstrap samples [7]. A small average instability value is ideal and suggests that the clusters are highly stable, and observations are less likely to change cluster membership across the bootstrap samples. Models with an average within cluster Jaccard similarity < 0.6 and an average instability > 0.5 were removed since these clusters are highly unstable [7].

v. Average within cluster dissimilarity vs average between cluster dissimilarity

The models were then assessed by means of the cluster.stats() function to confirm that the average between cluster dissimilarity is greater than the average within cluster dissimilarity. This suggests that there is a greater dissimilarity between observations in different clusters compared to observations that co-exist within the same cluster. Ideally, we want to maximize the between cluster dissimilarity indicating large cluster separation. In addition, we want to minimize the within cluster dissimilarity indicating a large degree of similarity within clusters. If a model had an average between cluster dissimilarity smaller than the average within cluster dissimilarity, the model was discarded since the model does not produce reliable clusters.

vi. Visualization of clusters

Finally, the clusters for each remaining model were assessed by means of dendrograms and cluster plots for the agglomerative hierarchical clustering models and cluster plots for the partitioning clustering methods. If the model shows clear separation between the three clusters, the model is a reliable representation of the clustering structure withing the dataset. If the clusters overlap to a large extend, the model is discarded.

H. Objective 2: Investigate if three foetal health classes are appropriate.

i. Using NbClust to determine the optimal number of clusters.

The NbClust() function was used to determine the optimal number of clusters for this dataset by performing a majority vote across 30 indices. The optimal clustering method and distance metric was used as input for NbClust()

ii. Within cluster sum of squares as a criterion for choosing the number of clusters

Ideally, each cluster should have minimum within cluster variance and within cluster sum of squares. As the number of clusters (k) increase, the total within cluster variance should decrease, before reaching 0 at k = n observations. In this investigation, I used the elbow method [8] to assess the total within cluster sum of squares across k = 2 to k = 20 clusters. The K-means clustering method was implemented in this investigation, which takes 500 random initial starting points run over 1000 iterations. I obtained the optimal number of clusters (k), where the plot elbows off, suggesting that the total within cluster variance is at its global minimum, with very little decrease in the within cluster variance with an additional cluster k. Beyond this point where within cluster variance = 0 and within sum of squares = 0, we see that k = n observations, which is not what we are looking for.

iii. Average silhouette width vs number of clusters for the K-means method

The average silhouette width was computed for each cluster k =2 to k = 20 using the K-means clustering method. The K-means clustering algorithm was implemented using 500 random initial starting points run at 1000 iterations. The k that generates the largest average silhouette width was chosen as the optimal number of (k) clusters. By choosing the value of k with the largest average silhouette width, I am guaranteed the best model fit.

iv. Average silhouette width vs number of clusters for the hierarchical clustering method

The average silhouette width was computed for each cluster k =2 to k = 20, using an agglomerative hierarchical clustering method, namely complete linkage. I used the Manhattan distance as input since the Manhattan distance measure is robust against the effects of outliers when computing dissimilarity. The k that generates the largest average silhouette width was chosen as the optimal number of (k) clusters. By choosing the value of k with the largest average silhouette width, I am guaranteed the best model fit.

v. Gap statistic vs number of clusters

The optimal number of clusters (k) under this criterion is chosen by maximizing the gap statistic at k. Based on **Equation 1** below, k is chosen at the point where the gap statistic at k is greater than or equal to the gap statistic at k+1 minus the standard error over all simulations.

$$Gap_n(k) \geq Gap_n(k+1) - s(k+1) \ldots \textbf{Equation 1}$$

We want to maximize the gap statistic since the gap statistic indicates that the total within cluster variation for k, $W(k)$ differs from the expected value under the null reference distribution of the data where there is no apparent clustering.

Below I explain the logic for choosing the best number of clusters k by using this criterion.

- **K-means clustering algorithm:**
  The K-means clustering algorithm was used to compute the gap statistic at k = 2 to k = 20 clusters. This implementation of K-means randomly finds 25 initial starting points and runs

the algorithm for 100 iterations. This process is repeated by means of Monte Carlo bootstrapping, where B = 50 samples.

- **Hieratical clustering method:**

The agglomerative hierarchical clustering algorithm was used to compute the gap statistic at k = 2 to k = 20 clusters. This implementation uses complete linkage and takes in the Manhattan distance matrix as input. This process is repeated by means of Monte Carlo bootstrapping, where B = 50 samples.

Results

A. *Principal component results*

Figure 6 below shows the scree plot of the principal components plotted along the x-axis vs the eigenvalues for each principal component plotted along the y-axis. The eigenvalues represent the variance explained by each principal component. We see that the scree plot elbows off at 2 principal components. **Figure 7** below shows the cumulative proportion of variance that the principal components explain. We see that 8 principal components cumulatively explain greater than 80% of the variance observed in the data (cumulative sum = 0.805). However, choosing 8 principal components produced suboptimal clustering results and reduced performance. As a result, 2 dimensions were chosen going forward, in accordance with the scree plot in **Figure 6**. Thus, I extracted 2 principal components from the PCA score matrix.

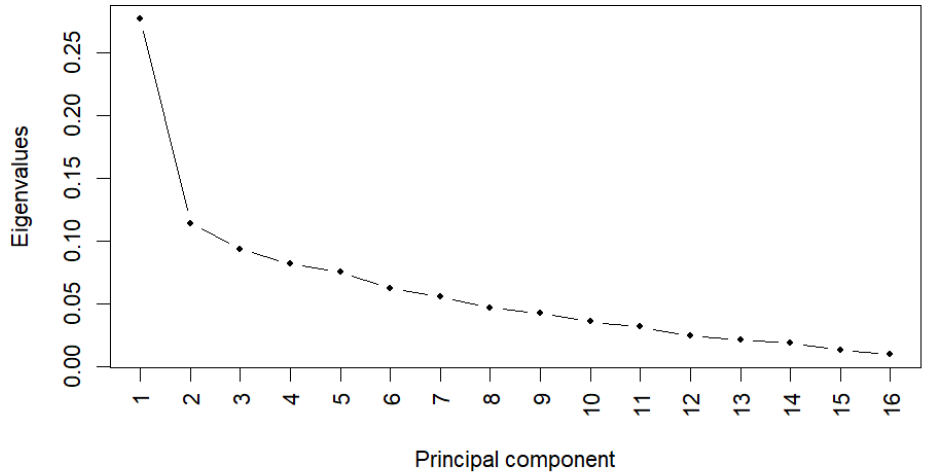

Figure 6: Scree plot showing the proportion of variance explained by each principal component. The elbow point of the scree plot is chosen as the reliable number of principal components for dimensionality reduction.

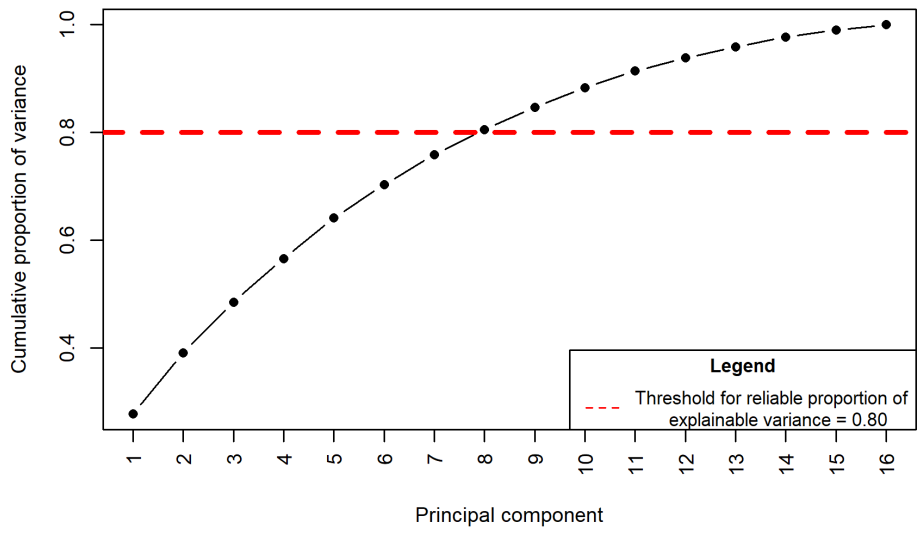

Figure 7: Line plot showing cumulative proportion of variance explained by the principal components. Ideally, the principal components should cumulatively explain at least 80% of variance observed in the data.

B. Comparison of model performance

Table 10 below shows the cophenetic correlation coefficient for each agglomerative hierarchical model, the average silhouette width and the gap statistic for all models. **Table 11** shows the per model average within cluster Jaccard similarity-index, average cluster instability, average within cluster dissimilarity and average between cluster dissimilarity.

In **Table 10**, we see that the following agglomerative hierarchical models (shown in green cells) have a cophenetic correlation coefficient ≥ 0.6, suggesting that these models do a relatively good job at preserving the pairwise distances between the original observations: model 1, model 2, model 7, model 8, model 9, model 11, model 12, model 13, model 14, model 15 and model 17. These agglomerative hierarchical models passed the first filter, in addition to the partitioning clustering models 22-26. All other models shown in orange cells did not pass the filter (cophenetic correlation coefficient < 0.6) and were discarded as a result.

In **Table 10**, we also see that the following agglomerative hierarchical models have an average silhouette width ≥ 0.400: model 2, model 7, model 9, model 14, model 15 and model 17. In addition, the partitioning clustering models 22-26 also have an average silhouette width ≥ 0.400. These models have the best model fit among all other models and were chosen as candidate models going forward. All other models did not pass the filter (average silhouette width < 0.400), and were discarded as a result.

In **Table 10**, we also see that model 7, model 9, model 14 and model 15 have a gap statistic of 0.349, 0.391, 0.213, and 0.384, respectively. Each of these models have a gap statistic value that is relatively smaller than the rest of the models. Since we want to maximize the gap statistic, model 7, model 9, model 14 and model 15 were discarded. The following models had a relatively large gap statistic, namely: model 2, model 17, model 22, model 23, model 24, model 25 and model 26 remained.

In **Table 11** below, we see that model 2 has a relatively small average within cluster Jaccard similarity-index (0.555) and a relatively large average cluster instability (0.430), and so model 2 was discarded. Model 17 and models 22-26 remained, since these models had the largest average within cluster Jaccard similarity indices and relatively small values for the average cluster instability.

In **Table 11,** we see also see that the average between cluster dissimilarity > average within cluster dissimilarity for all remaining models 17 and 22-26. Thus, it was necessary to compute the proportion of between cluster dissimilarity in relation to the total dissimilarity for each model. We find that all models had a relatively large proportion of between cluster dissimilarity, where:

- Model 17 had a 69.27% between cluster dissimilarity.
- **Model 22 had a 69.34% between cluster dissimilarity.**
- Model 23 had a 69.00% between cluster dissimilarity.
- Model 24 had a 68.98% between cluster dissimilarity.
- Model 25 had a 68.70% between cluster dissimilarity.
- Model 26 had a 68.95% between cluster dissimilarity.

Model 22 had the largest proportion of between cluster dissimilarity, suggesting that these clusters are well-separated and observations within a cluster are most similar. As a result, model 22 was chosen as the optimal model for clustering the foetal health dataset.

Table 10: Assessing clustering model performance using the cophenetic correlation coefficient, average silhouette width and the gap statistic.

Model ID	Model type	Distance measure	Linkage	Model params	*Cophenetic correlation	*Average silhouette width	*Gap statistic
1	hclust	Euclidean	Complete	k=3	0.686	0.369	0.501
2	hclust	Manhattan	Complete	k=3	0.651	0.401	0.674
3	hclust	Maximum	Complete	k=3	0.488	0.285	0.496
4	hclust	Euclidean	Single	k=3	0.534	0.614	0.678
5	hclust	Manhattan	Single	k=3	0.523	0.615	0.678
6	hclust	Maximum	Single	k=3	0.539	0.615	0.678
7	hclust	Euclidean	Average	k=3	0.729	0.407	0.349
8	hclust	Manhattan	Average	k=3	0.722	0.381	0.514
9	hclust	Maximum	Average	k=3	0.722	0.437	0.391
10	hclust	Euclidean	Median	k=3	0.561	0.309	0.612
11	hclust	Manhattan	Median	k=3	0.615	0.273	0.573
12	hclust	Maximum	Median	k=3	0.63	0.223	0.515
13	hclust	Euclidean	Centroid	k=3	0.724	0.341	0.219
14	hclust	Manhattan	Centroid	k=3	0.721	0.442	0.213
15	hclust	Maximum	Centroid	k=3	0.696	0.437	0.384
16	hclust	Euclidean	Ward D	k=3	0.598	0.382	0.682
17	hclust	Manhattan	Ward D	k=3	0.612	0.435	0.759
18	hclust	Maximum	Ward D	k=3	0.503	0.304	0.696
19	hclust	Euclidean	Ward D2	k=3	0.582	0.355	0.719
20	hclust	Manhattan	Ward D2	k=3	0.598	0.409	0.682
21	hclust	Maximum	Ward D2	k=3	0.548	0.337	0.701
22	K-means	Euclidean (to calculate the silhouette width)	-	centers=3, nstart=100, iter.max=1000		0.428	0.726
23	PAM	Euclidean	-	k=3		0.418	0.736
24	PAM	Manhattan	-	k=3		0.429	0.736
25	CLARA	Euclidean	-	k=3		0.412	0.742
26	CLARA	Manhattan	-	k=3		0.433	0.742

***Green cells indicate models which have passed filters for that column criteria and orange cells indicate models which did not pass the filter for that column criteria and have been discarded as a result. White or grey cells indicate models which have been discarded in previous column filters.**

Table 11: Assessing clustering model performance using the average within cluster Jaccard similarity-index, average cluster instability, average within cluster dissimilarity and average between cluster dissimilarity.

Model ID	Average Jaccard similarity	Average instability	Average within cluster dissimilarity	Average between cluster dissimilarity
1	0.555	0.430	2.215	4.435
2*	0.555	0.430	2.418	5.221
3	0.555	0.430	1.997	3.288
4	0.800	0.237	3.003	9.066
5	0.800	0.237	3.859	11.767
6	0.800	0.237	2.673	8.222
7	0.514	0.498	2.583	5.231
8	0.514	0.498	2.875	5.375
9	0.514	0.498	2.196	4.397
10	0.420	0.592	2.147	3.760
11	0.420	0.592	2.897	4.787
12	0.420	0.592	2.092	4.008
13	0.579	0.454	3.017	5.828
14	0.579	0.454	3.856	10.626
15	0.579	0.454	2.324	4.673
16	0.762	0.063	1.842	3.992
17*	0.762	0.063	2.198	4.954
18	0.762	0.063	1.591	3.269
19	0.650	0.246	1.778	3.647
20	0.650	0.246	2.379	5.173
21	0.650	0.246	1.604	3.243
22*	0.650	0.246	1.693	3.829
23*	0.650	0.246	1.690	3.763
24*	0.650	0.246	2.182	4.852
25*	0.650	0.246	1.706	3.745
26*	0.650	0.246	2.218	4.925

***Green cells in the "Model ID" column indicate models that have passed filters in Table 10. Green cells in all other columns indicate models that have passed filters for that column criterion, while orange cells indicate models which did not pass the filter for that column criterion and were discarded as a result.**

C. *Objective 1: Optimal clustering model*

Model 22, which uses the K-means clustering method was chosen as the best clustering model based on the results of the decision criteria explained above in ***B. Comparison of model Performance.***

D. *Dendrogram of the hierarchical clustering models*

The dendrograms for the discarded models – the agglomerative hierarchical models can be seen in **Figures 8-14** in **Addendum A.**

Figure 9 in particular, which represents the single linkage models display chaining which is undesirable. **Figure 13** and **Figure 14** which show Ward D and Ward D2 linkage show the most reliable clustering among the agglomerative hierarchical models. However, these clusters will not be considered since these models have been discarded based on the results of the decision criteria explained above in ***B. Comparison of model Performance.***

E. *Cluster plot of the optimal clustering model*

Figure 15 and **Figure 16** below shows the clustering plots for model 22, which uses the K-means partitioning clustering method. In **Figure 15** we see that the foetal health data can be clustered into 3 clusters shown in red, blue and green for cluster 1, 2 and 3 respectively. The clusters are plotted against component 1 along the x-axis and component 2 along the y-axis. It is clear that cluster 1 (shown in red) is the largest cluster, followed by cluster 3 (shown in green) and cluster 2 (shown in blue). The cluster plot in **Figure 16** shows a similar clustering structure, where the clusters are also plotted against component 1 along the x-axis and component 2 along the y-axis. Additionally, we see that these two clusters explain 100% of the point variability.

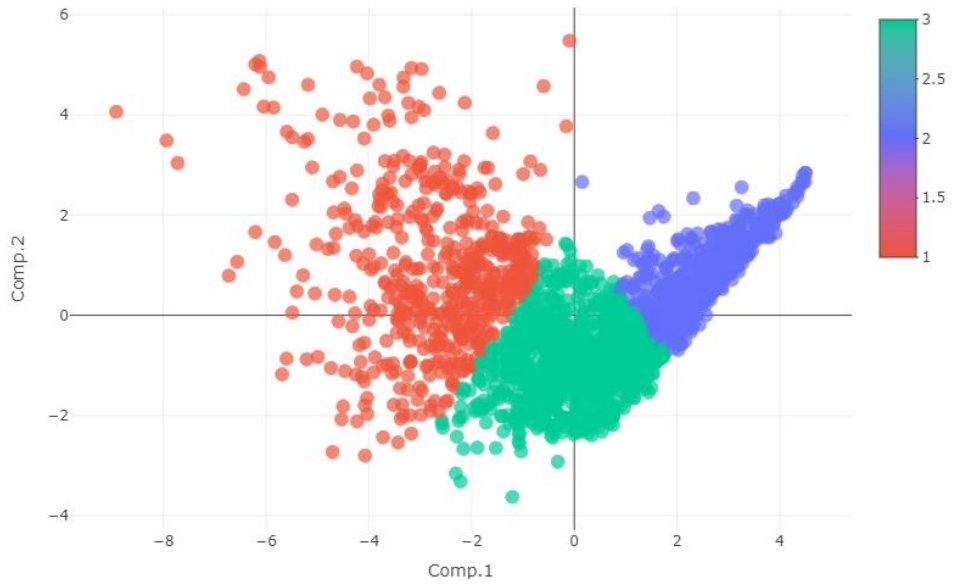

Figure 15: Cluster plot showing model 22, a K-means clustering model that clusters the foetal health data into 3 distinct clusters plotted against component 1 along the x-axis and component 2 along the y-axis.

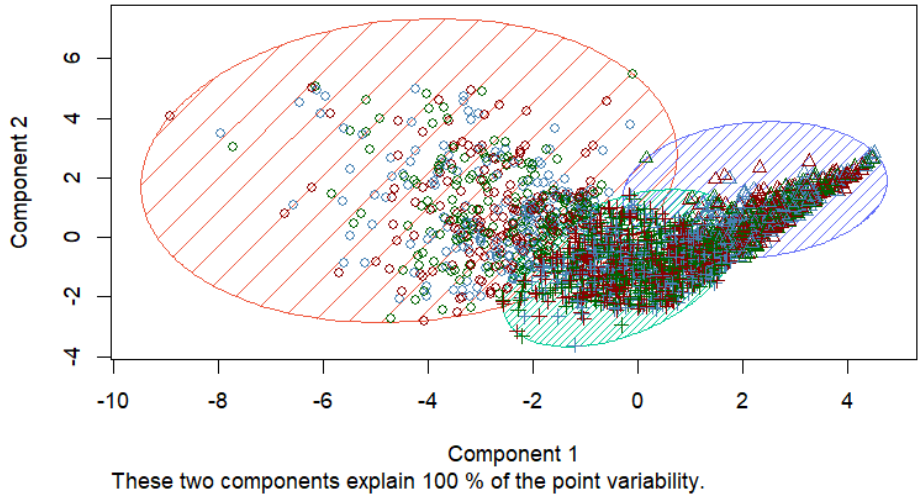

Figure 16: Cluster plot showing model 22, a K-means clustering model that clusters the foetal health data into 3 distinct clusters plotted against component 1 along the x-axis and component 2 along the y-axis.

F. Silhouette plot of the optimal clustering model

In **Figure 17** below, we see the silhouette width of the best performing model 22 based on the results of the filtering criteria in **Table 10** and **Table 11.**

Shown in **Figure 17**, cluster 2 (shown in blue) has a total of 543 observations, cluster 3 (shown in green) has a total of 552 observations and cluster 1 (shown in red) has a total of 1031 observations.

Most of the observations that are found in cluster 2 (shown by the blue horizontal bars) have an average silhouette width at approximately 0.5, suggesting that some observations in cluster 2 are most likely correctly assigned to cluster 2. However, some observations in cluster 2 have a negative silhouette width suggesting incorrect cluster assignment. Most of the observations that are found in cluster 3 (shown by the green horizontal bars) have an average silhouette width ≥ 0.5, suggesting that these observations are likely correctly assigned to cluster 3. Most of the observations in cluster 1 (shown by the red horizontal bars) also have an average silhouette width $\geq$ 0.5, suggesting that these observations are most likely correctly assigned to cluster 1.

Cluster 2 (shown in blue) has an overall average silhouette width of 0.30 suggesting that some observations are correctly assigned to cluster 2. As noted before, some observations in cluster 2 have a negative silhouette width indicating incorrect cluster assignment and so these incorrect cluster assignments, bring down the average silhouette width for cluster 2. Cluster 3 (shown in green) has an average silhouette width of 0.55 suggesting that, on average, most observations are correctly assigned to cluster 3. Cluster 1 (shown in red) has an average silhouette width of 0.43 suggesting that on average, most observations are incorrectly assigned to cluster 1.

Overall, the average silhouette width = 0.43 suggesting that the structure is weak and possibly artificial. However, considering the average silhouette width across all 26 models in **Table 10**, this average silhouette width is relatively good.

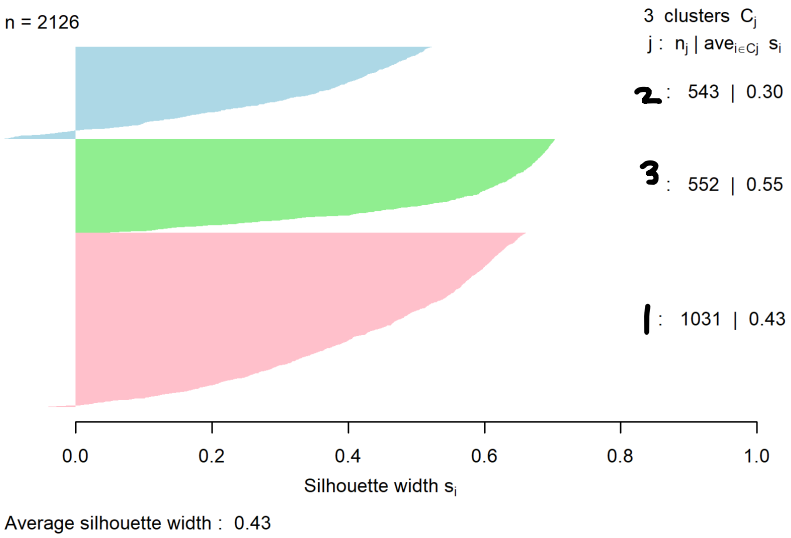

Figure 17: Average silhouette width showing model 22, which uses the K-means clustering algorithm to cluster the foetal health dataset into 3 distinct clusters. The Euclidean distance measure was used to compute the silhouette width.

G. *Objective 2: Optimal number of clusters*

i. Using NbClust to determine the optimal number of clusters

The NbClust package in r suggests that there are indeed 3 clusters. The histogram below in **Figure 18 (a)** shows that the majority vote across all metrics, is at k = 3 clusters, when using the K-means clustering algorithm. In total, 7 votes (out of 23) voted for k = 3 as the optimal number of clusters.

In **Figure 18 (b),** we see the line plot showing the Hubert index. At k = 3, we see a significant knee that corresponds to a significant increase of the value of this measure i.e., a significant peak in Hubert index second differences plot **Figure 18 (c).**

In **Figure 18 (d),** we see the line plot showing the D-index. At k = 3, we see a significant knee that corresponds to a significant increase of the value of this measure i.e., a significant peak in the D-index second differences plot **Figure 18 (e).**

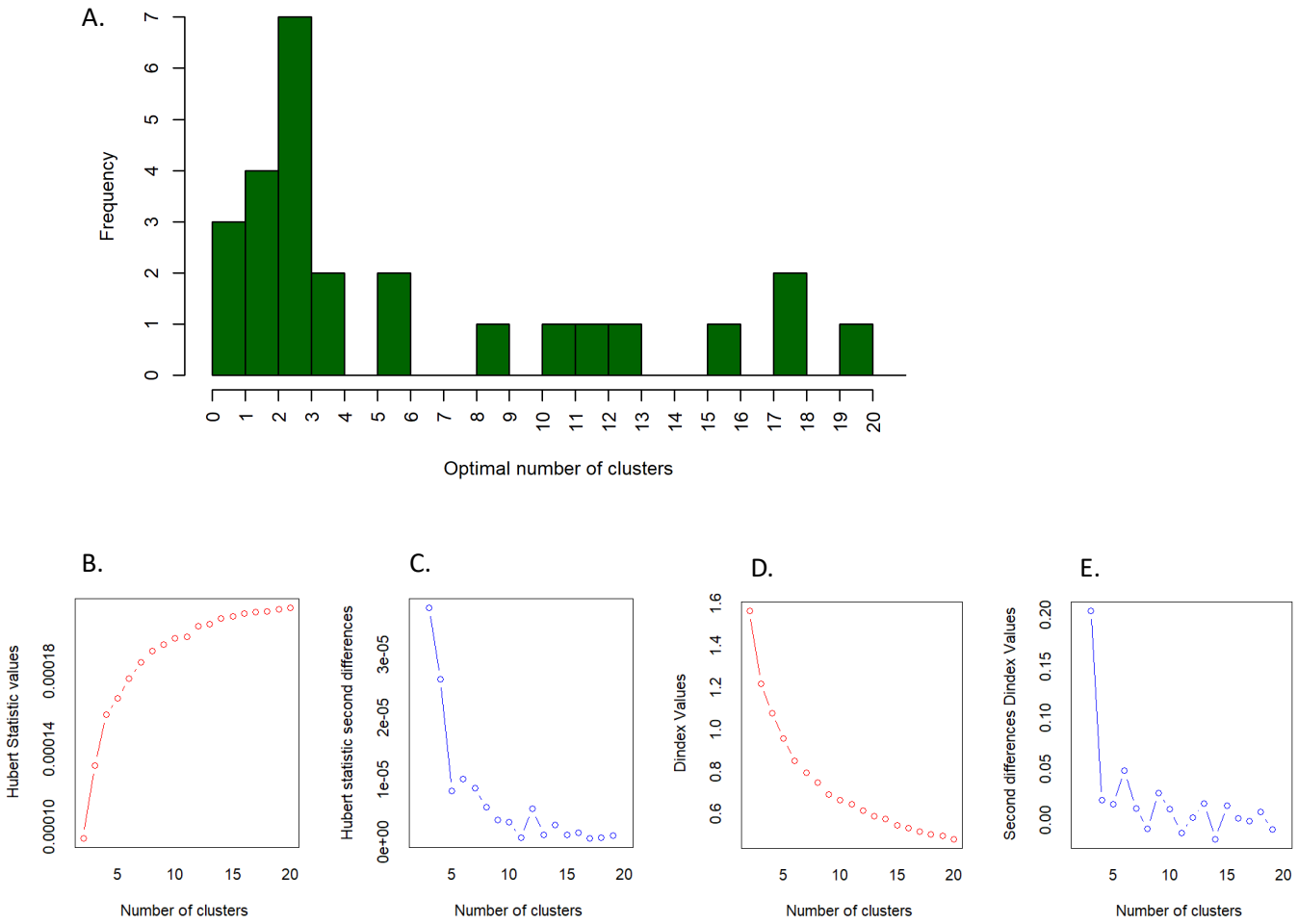

Figure 18: Determining the optimal number of clusters k using Nbclust, Hubert statistics and D-index values.
(a) Histogram showing the Nbclust majority vote for the optimal number of clusters k across 30 indices.
(b) Number of clusters vs Hubert statistic values (c) Number of clusters vs Hubert statistic second differences
(d) Number of clusters vs D-index values (c) Number of clusters vs Second differences D-index values

ii. Within cluster sum of squares as a criterion for choosing the number of clusters

The line plot below in **Figure 19** shows that the optimal number of clusters are somewhere between k = 3 to k = 6 clusters, where the plot elbows off. The investigation that follows clarifies which value of k should be chosen as the optimal number of clusters.

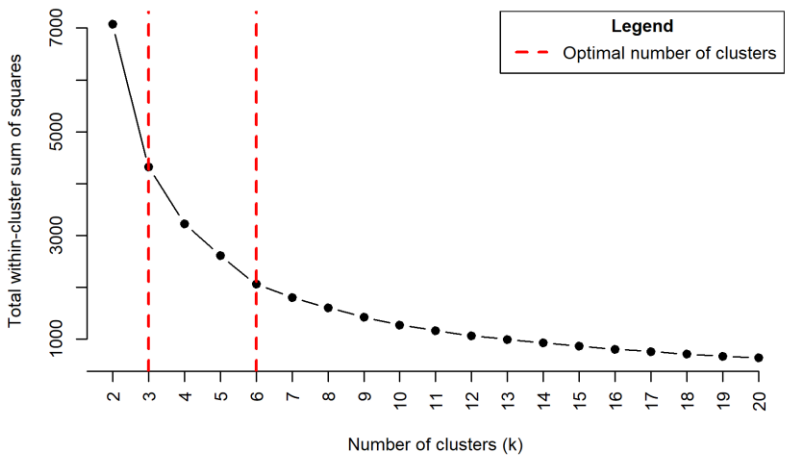

Figure 19: Line plot showing number of clusters (k) vs total within cluster sum of squares computed using the K-means clustering algorithm.

iii. Average silhouette width vs number of clusters for the K-means clustering method

Figure 20 below, suggests that k = 3 is the optimal number of clusters for the K-means model, with an average silhouette width of 0.437 being the largest among all other values of k. This suggests that the K-means model at k = 3 clusters has the best fit among all other values of k. In parallel to the methods above, k = 3 is chosen as the optimal number of clusters in this investigation.

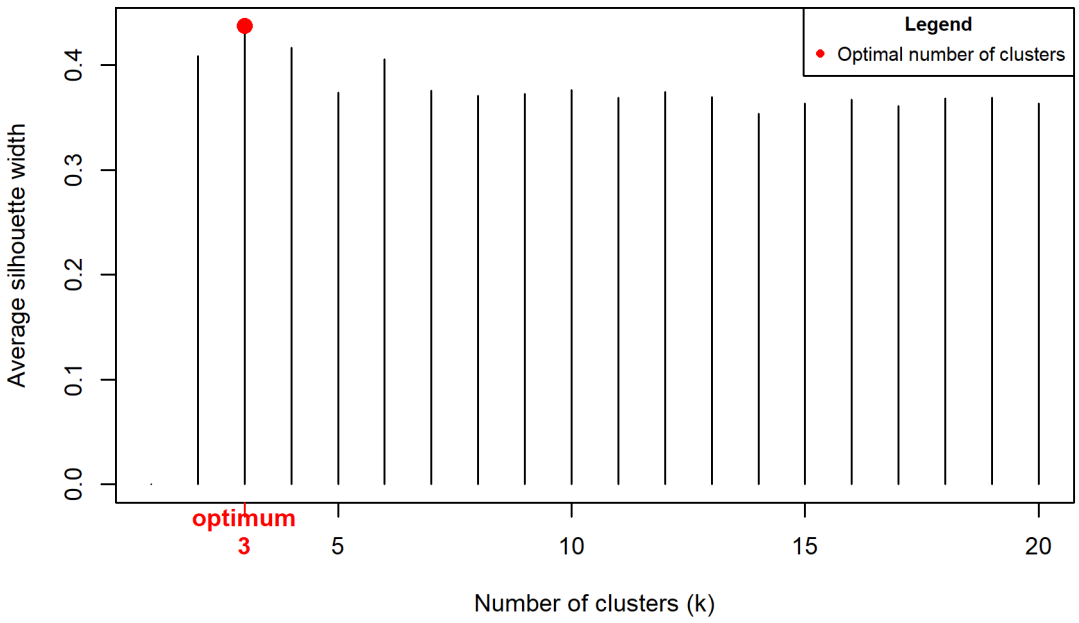

Figure 20: Line plot showing number of clusters (k) vs total average silhouette width computed using the K-means clustering algorithm.

iv. Average silhouette width vs number of clusters for the hierarchical clustering method

. This implementation of hierarchical clustering uses complete linkage and Manhattan distance. **Figure 21** below, suggests that k = 2 and k = 3 may be the optimal number of clusters for the hierarchical clustering model, with an average silhouette width of 0.401 being the largest among all other values of k. However, since my previous methods suggest that k = 3 is the optimal number of clusters, k = 3 is also chosen as the optimal number of clusters for this investigation.

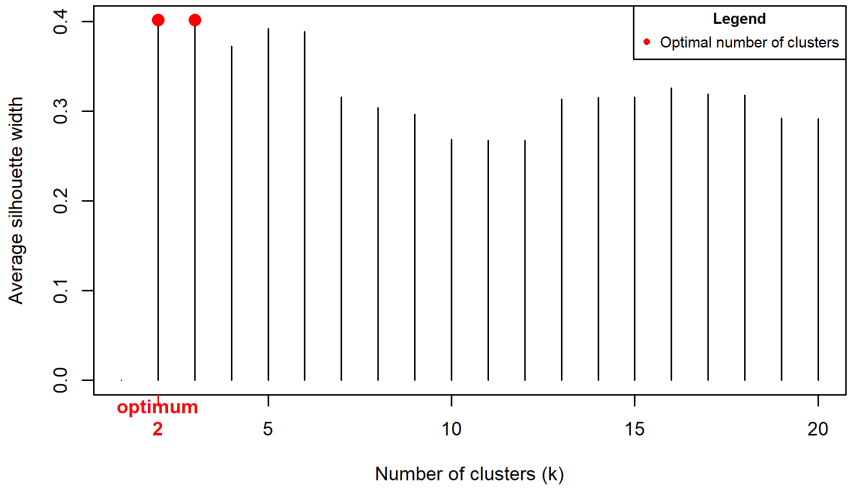

Figure 21: Line plot showing number of clusters k vs average silhouette width computed using the complete linkage and the Manhattan distance measure.

v. Gap statistic vs number of clusters

In agreement with the methods above, both the K-means and agglomerative hierarchical clustering methods below in **Figure 22** and **Figure 23**, respectively suggest that k = 3 is the optimal number of clusters. For the K-means method in **Figure 22**, the largest Gap statistic = 0.727 at k = 3. Likewise, for the agglomerative hierarchical method in **Figure 23**, the largest Gap statistic = 0.674 at k = 3. In other words, the criterion is satisfied so that $Gap_n(k) \geq Gap_n(k+1) - s(k+1)$, for k = 3. This implies that by increasing the number of clusters of k beyond k = 3, there will be no further improvement in the gap statistic at k+1, where the gap statistic compares the total within cluster variation with their expected values under the null distribution of the data. It is evident that at k = 3, we have maximized the Gap statistic.

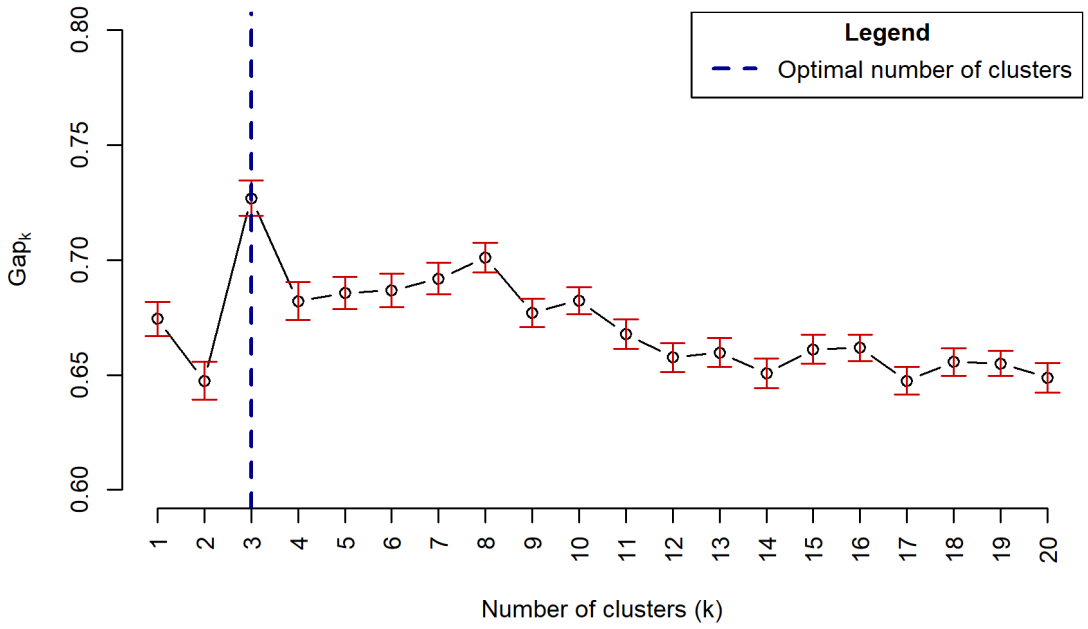

Figure 22: Line plot showing the number of clusters (k) vs the gap statistic at k for the K-means clustering method, where k ranges from 1-20 clusters.

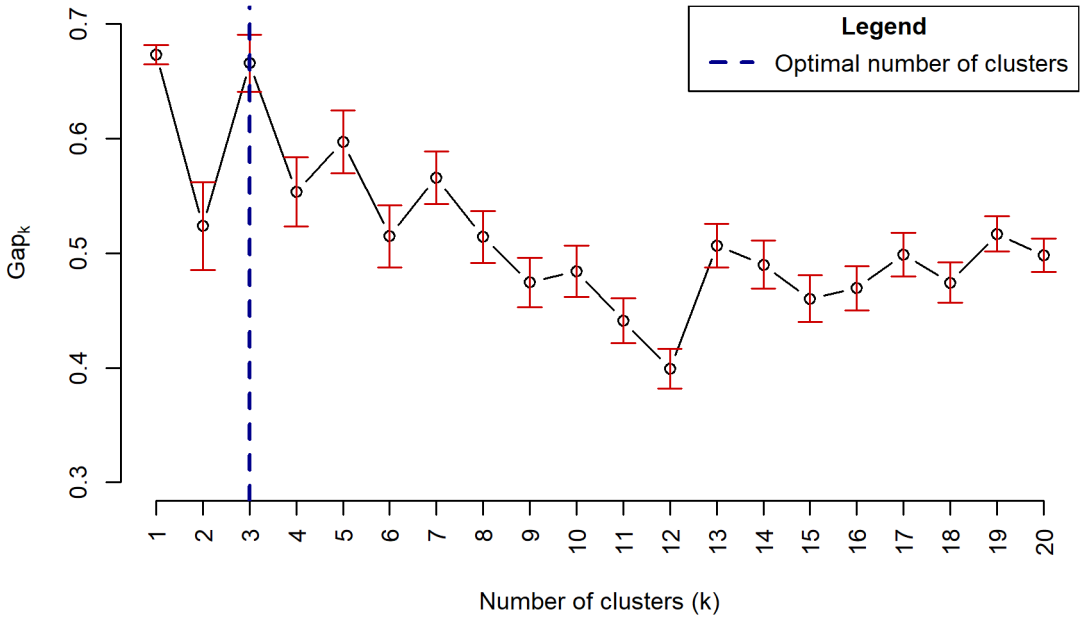

Figure 23: Line plot showing the number of clusters (k) vs the gap statistic at k for the complete linkage clustering method, where the values of k range from 1 to 20 clusters. This implementation of complete linkage takes the Manhattan distance matrix as input.

vi. Optimal number of clusters

Based on the investigations above, I can conclude that k = 3 is the optimal number of clusters for the foetal health dataset.

Discussion

Model 22, which uses the K-means partitioning clustering algorithm, outperformed all other models based on all filtering criteria mentioned in the ***Results*** section under ***B. Comparison of model performance.***

In this exercise, a partitioning clustering algorithm was the best choice since we prespecify the number of clusters k. Whereas, hierarchical clustering algorithms determine the best clustering structure without prespecifying the number of clusters beforehand. For these models, one is required to "cut" the dendrogram at the desirable value of k.

In addition, using the K-means clustering algorithm makes intuitive sense for the foetal health dataset which has three distinct classes (healthy, suspect, and pathological). The K-means algorithm finds mutually exclusive clusters with very little overlap between clusters, compared to the hierarchical clustering algorithms which allows clusters to overlap. In this dataset, an observation cannot be classified as both pathological and suspect, pathological and healthy or suspect and healthy; the observation is either one or the other.

An added advantage of the K-means clustering method is that convergence is guaranteed, and the algorithm generalizes well to clusters of different shapes and sizes such as elliptical clusters which is apparent in the cluster plot in **Figure 16**, which can be viewed under the ***Results*** section [3]

Some notable shortfalls of the K-means clustering algorithm is that it has difficulty in clustering data of different sizes and density when generalization is not implemented. However, in this scenario, clusters of various size and density were identified (shown in the cluster plot in **Figure 16** under the ***Results*** section). In addition, the K-means algorithm is sensitive to outliers and outliers may drag centroids from its optimal position, distorting clusters as a result. Interestingly, I observed similar average silhouette width results when using either Euclidean or Manhattan distance measures, where the Manhattan distance reduces the effect of outliers. Another disadvantage of the K-means clustering method is that it does not do well with data having a large number of dimensions. However, this shortfall was circumvented by reducing the dimensionality of the data set, by means of PCA.

The resultant model 22 produced 3 clusters, where cluster 1 has more observations (n = 1031) than cluster 3 (n = 552) which in turn has more observations than cluster 2 (n = 543). This suggests that the 3 classes of foetal health: "healthy", "suspect" and "pathological", have varied distributions. Kaggle suggests that there are 1655 healthy cases, 295 suspect cases and 176 pathological cases [1]. The order of magnitude for each category in the clustered data set is consistent with the classes in the labelled dataset on Kaggle. Cluster 1 may correspond to the "healthy" class, cluster 2 may correspond to the "pathological" class and cluster 3 may correspond to the "suspect class".

Notably, the "suspect" class falls between the "healthy" and pathological classes. This makes intuitive sense, since cluster 3 (shown in green in **Figure 15**) falls between the largest cluster, (cluster 1 shown in red in **Figure 15**) and the smallest cluster (cluster 2 shown in blue in **Figure 15).** Based on this trend, one may speculate further that the largest cluster (cluster 1 shown in red in **Figure 15**) represents the "healthy" category, the second largest cluster (cluster 3 shown in green in **Figure 15**) may represent the "suspect" category and the smallest cluster (cluster 2 shown in blue in **Figure 15**) may represent the "pathological" category. However, further investigation is required to confirm this trend.

Conclusion

The K-means partitioning clustering algorithm (model 22) outperformed all other hierarchical and partitioning clustering models. In addition, the K-means clustering method was able to faithfully represent the clustering structure of the foetal health data set by partitioning the dataset into k = 3 clusters. The clustering structure produced is consistent with the structure of the labelled data set, found on Kaggle. Further Analysis also suggests that k = 3 is indeed the optimal number of clusters for the foetal health dataset. This suggests that the 3 clusters accurately represent the 3 classes of foetal health: “healthy”, “suspect”, and “pathological”.

References

1. Larxel (2017). *Fetal Health Classification*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification [Accessed 23 Sep. 2023].
2. 365 Data Science. (2020). *How to Combine PCA and K-means Clustering in Python? | 365 Data Science*. [online] Available at: https://365datascience.com/tutorials/python-tutorials/pca-k-means/ [Accessed 28 Sep. 2023].
3. Google for Developers. (2022). *k-Means Advantages and Disadvantages*. [online] Available at: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages [Accessed 6 Oct. 2023].
4. ResearchGate. (2020). *What is cophenetic correlation? | ResearchGate*. [online] Available at: https://www.researchgate.net/post/What-is-cophenetic-correlation [Accessed 6 Oct. 2023].
5. Bhardwaj, A. (2020). *Silhouette Coefficient - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c [Accessed 6 Oct. 2023].
6. Nyirenda, J.C. Cluster Analysis [Slide show].
7. Rpubs.com. (2020). *RPubs - An introduction to Clustering Methods in R*. [online] Available at: https://rpubs.com/pjmurphy/599072 [Accessed 6 Oct. 2023].
8. Rpubs.com. (2018). *RPubs - k-means clustering and dendrogram analysis*. [online] Available at: https://www.rpubs.com/dvallslanaquera/clustering [Accessed 6 Oct. 2023].

Appendix A

Table 1: Description of independent variables in the Foetal Health dataset.

Independent variable name	Description
1. **Baseline value**	Foetal heart rate (FHR) baseline (beats per minute)
2. **Accelerations**	Number of accelerations per second
3. **Foetal movement**	Number of foetal movements per second
4. **Uterine contractions**	Number of uterine contractions per second
5. **Light decelerations**	Number of light decelerations per second
6. **Severe decelerations**	Number of severe decelerations per second
7. **Prolonged decelerations**	Number of prolonged decelerations per second
8. **Abnormal short-term variability**	Percentage of time with abnormal short-term variability
9. **Mean value of short-term variability**	Mean value of short-term variability
10. **Percentage of time with abnormal long-term variability**	Percentage of time with abnormal long-term variability
11. **Mean value of long-term variability**	Mean value of long-term variability
12. **Histogram width**	Width of FHR histogram
13. **Histogram min**	Minimum (low frequency) of FHR histogram
14. **Histogram max**	Maximum (high frequency) of FHR histogram
15. **Histogram number of peaks**	# of histogram peaks
16. **Histogram number of zeroes**	# of histogram zeroes
17. **Histogram mode**	Histogram mode
18. **Histogram mean**	Histogram mean
19. **Histogram median**	Histogram median
20. **Histogram variance**	Histogram variance
21. **Histogram tendency**	Histogram tendency

Table 2: Head of the foetal health data set, showing the first 6 rows for columns 1 to 7.

Baseline value	Accelerations	Foetal movement	Uterine contractions	Light decelerations	Severe decelerations	Prolonged decelerations
120	0.000	0	0.000	0.000	0	0.000
132	0.006	0	0.006	0.003	0	0.000
133	0.003	0	0.008	0.003	0	0.000
134	0.003	0	0.008	0.003	0	0.000
132	0.007	0	0.008	0.000	0	0.000
134	0.001	0	0.010	0.009	0	0.002

Table 3: Head of the foetal health data set, showing the first 6 rows for columns 8 to 14.

Abnormal short-term variability	Mean value of short-term variability	Percentage of time with abnormal long-term variability	Mean value of long-term variability	Histogram width	Histogram min	Histogram max
73	0.5	43	2.4	64	62	126
17	2.1	0	10.4	130	68	198
16	2.1	0	13.4	130	68	198
16	2.4	0	23.0	117	53	170
16	2.4	0	19.9	117	53	170
26	5.9	0	0.0	150	50	200

Table 4: Head of the foetal health data set, showing the first 6 rows for columns 15 to 21.

Histogram number of peaks	Histogram number of zeroes	Histogram mode	Histogram mean	Histogram median	Histogram variance	Histogram tendency
2	0	120	137	121	73	1
6	1	141	136	140	12	0
5	1	141	135	138	13	0
11	0	137	134	137	13	1
9	0	137	136	138	11	1
5	3	76	107	107	170	0

Table 5: Tail of the foetal health data set, showing the last 6 rows for columns 1 to 7.

Baseline value	Accelerations	Foetal movement	Uterine contractions	Light decelerations	Severe decelerations	Prolonged decelerations
140	0.000	0.000	0.005	0.001	0	0
140	0.000	0.000	0.007	0.000	0	0
140	0.001	0.000	0.007	0.000	0	0
140	0.001	0.000	0.007	0.000	0	0
140	0.001	0.000	0.006	0.000	0	0
142	0.002	0.002	0.008	0.000	0	0

Table 6: Tail of the foetal health data set, showing the last 6 rows for columns 8 to 14.

Abnormal short-term variability	Mean value of short-term variability	Percentage of time with abnormal long-term variability	Mean value of long-term variability	Histogram width	Histogram min	Histogram max
77	0.7	17	6.0	31	124	155
79	0.2	25	7.2	40	137	177
78	0.4	22	7.1	66	103	169
79	0.4	20	6.1	67	103	170
78	0.4	27	7.0	66	103	169
74	0.4	36	5.0	42	117	159

Table 7: Tail of the foetal health data set, showing the last 6 rows for columns 15 to 21.

Histogram number of peaks	Histogram number of zeroes	Histogram mode	Histogram mean	Histogram median	Histogram variance	Histogram tendency
2	0	145	143	145	2	0
4	0	153	150	152	2	0
6	0	152	148	151	3	1
5	0	153	148	152	4	1
6	0	152	147	151	4	1
2	1	145	143	145	1	0

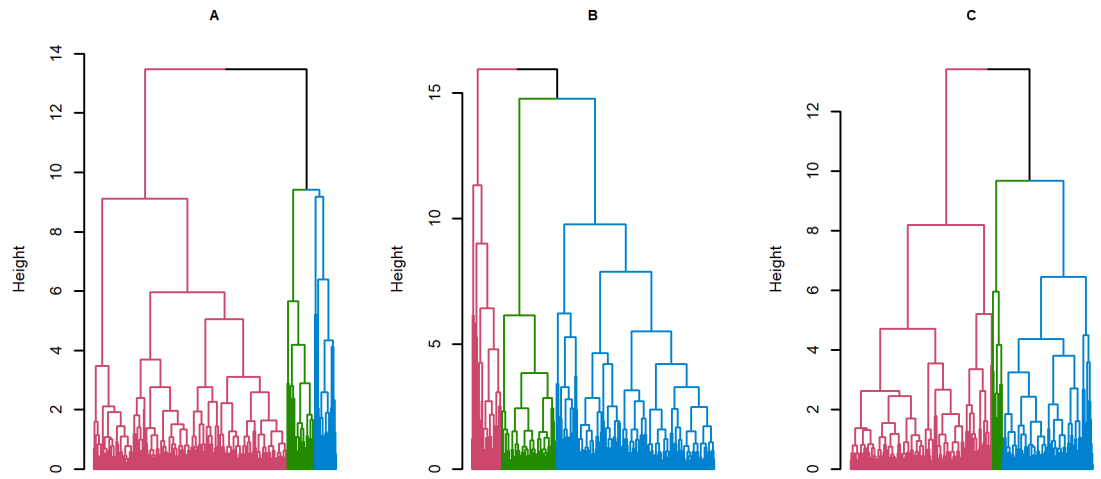

Figure 8: Dendrograms of the agglomerative, hierarchical models which use complete linkage. Frames a to c represent models 1-3, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distances.

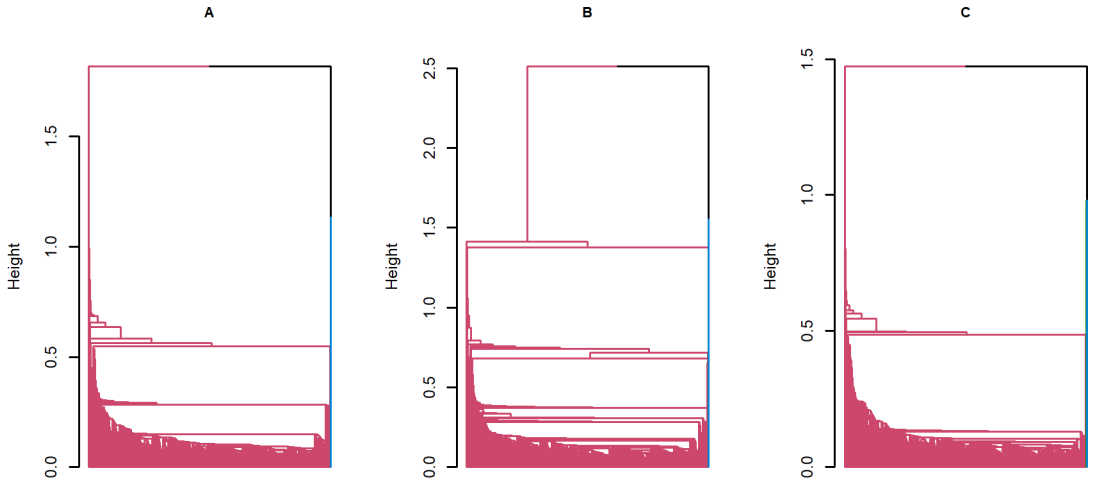

Figure 9: Dendrograms of the agglomerative, hierarchical models which use single linkage. Frames a to c represent models 4-6, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distance.

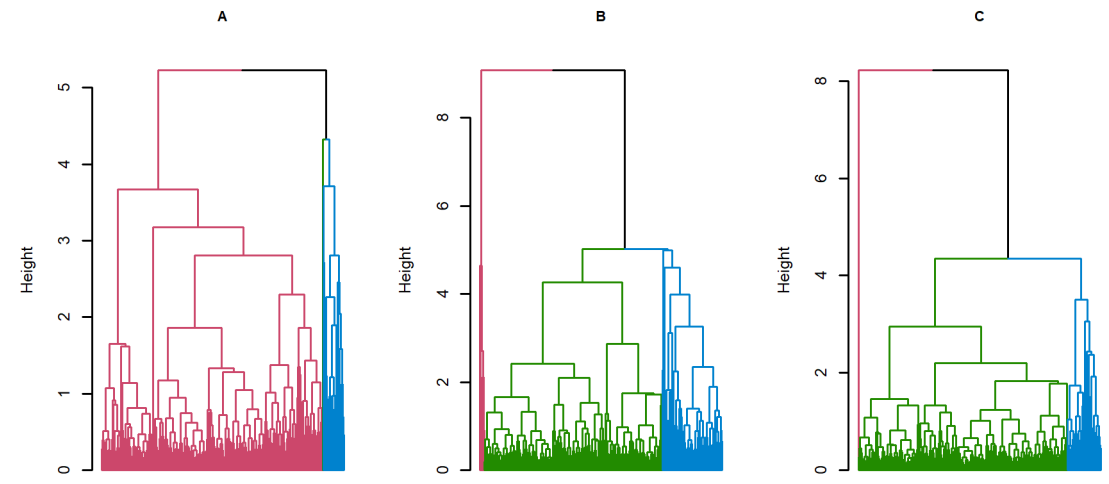

Figure 10: Dendrogram of the agglomerative, hierarchical models which use average linkage. Frames a to c represent models 7-9, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distance.

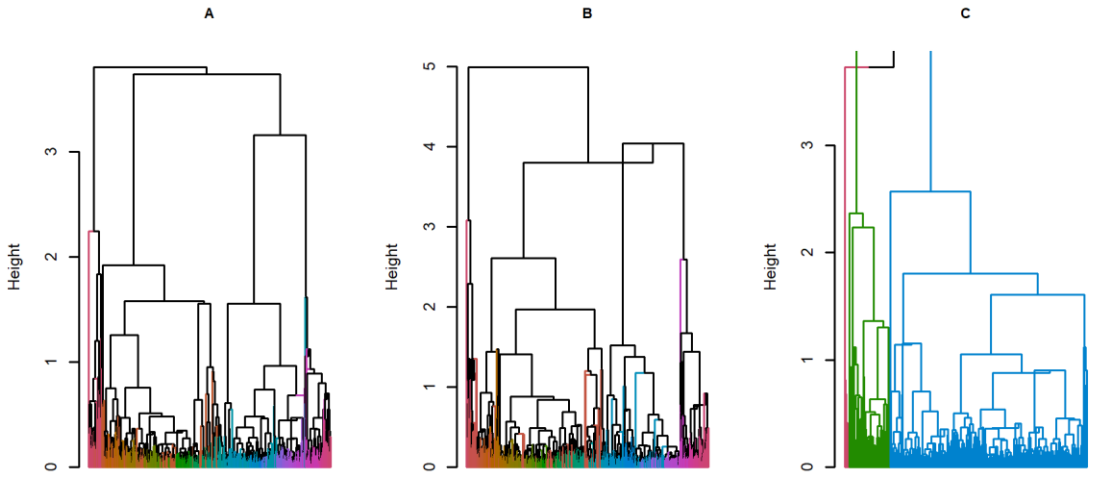

Figure 11: Dendrogram of the agglomerative, hierarchical models which use median linkage. Frames a to c represent models 10-12, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distance.

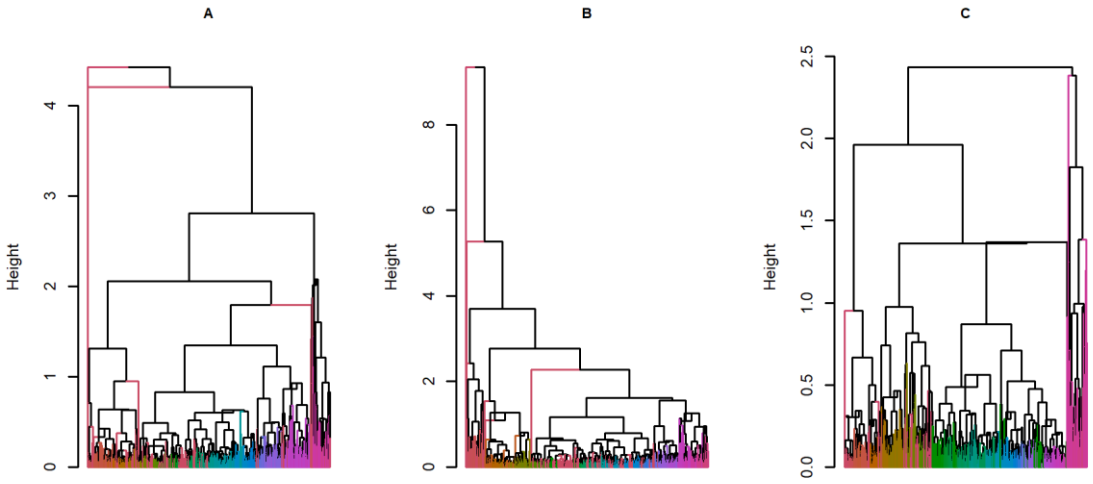

Figure 12: Dendrogram of the agglomerative, hierarchical models which use centroid linkage. Frames a to c represent models 13-15, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distance.

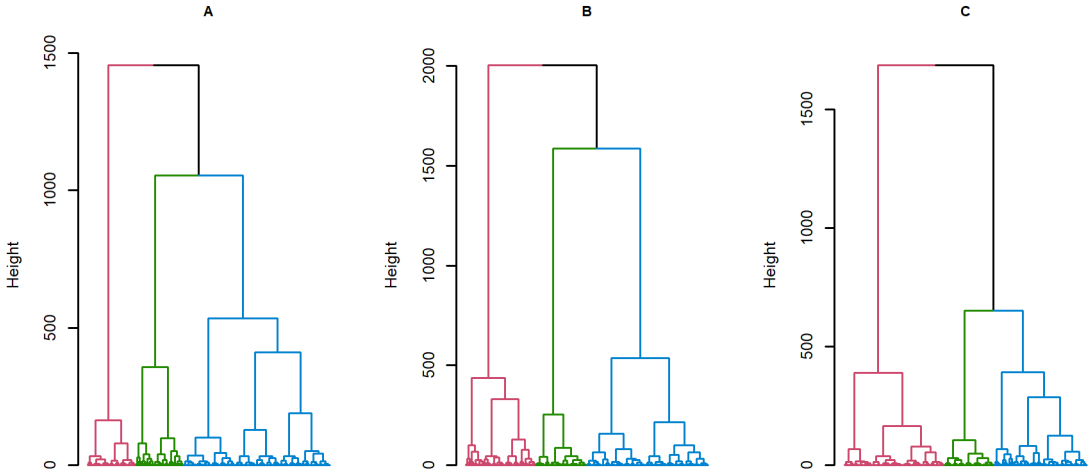

Figure 13: Dendrogram of the agglomerative, hierarchical models which use ward D linkage. Frames a to c represent models 16-18, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distance.

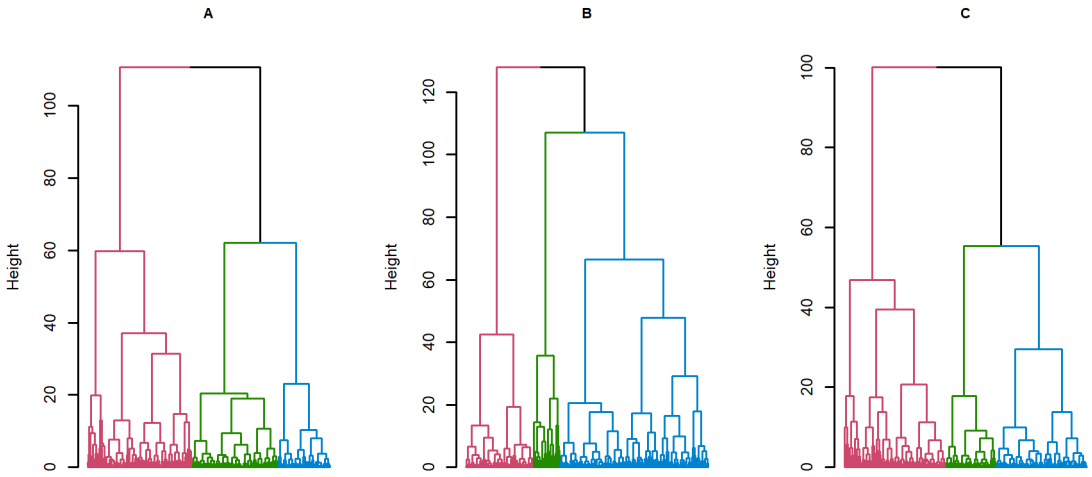

Figure 14: Dendrogram of the agglomerative, hierarchical models which use ward D2 linkage. Frames a to c represent models 19-21, which use the following dissimilarity matrix: (a) Euclidean, (b) Manhattan, and (c) Chebyshev Maximum distance.