University of Glasgow | School of Mathematics & Statistics

Cluster Analysis of Acute Respiratory Distress Syndrome - PreECMO data

Hsuan Meng

Degree of Master of Science
in
Data Analytics
School of Mathematics and Statistics
University of Glasgow
September 2023

Abstract

Cluster analysis has been thoroughly researched and used in the medical field. The goal of this dissertation is to use two different clustering methods, hierarchical clustering and K-means, to find the clusters in 450 patients' biosignatures data before medical treatment and apply Fisher's exact test to investigate the association between found clusters and survival.

Compared to hierarchical clustering, K-means performed better. It separates the cluster fairly. Conversely, hierarchical clustering partitions the data into two clusters but is poorly categorised. Consequently, only the results from K-means are taken to apply further analysis. This study finds that two clusters are detected in this data and are associated with the outcome variable.

TABLE OF CONTENTS

1. Introduction

The utilisation of cluster analysis has been extensively studied in a wide range of fields. For example, medical. Traditional biomarker analysis is getting more complex with the development of "high throughput" biological marker computations. Henk-Jan van den Ham et al. (2009) demonstrated the specific disease patterns using hierarchical clustering algorithms on biomarkers. Wierenga et al. (2019) applied the K-means clustering algorithm, and the results were used to find the relation between tumour features and the clinical data.

This type of analysis investigates processes for categorising variables based on calculating the important properties (Jain, 2008). Chiu, Douglas and Li (2009) also concluded that cluster analysis based on both K-means and hierarchical agglomerative can be used to group participants that share similar capabilities after summarising the data.

In the medical field, there is a term called Acute Respiratory Distress Syndrome (ARDS), a kind of acute hypoxic respiratory failure in which the PaO2 to FiO2 ratio is less than 300 mmHg. It is the most common complication for severely ill patients, which could increase their incidence mortality rate (Bos et al., 2017). Throughout decades of study, no medical therapies have been shown to be valuable in treating ARDS, and the primary reason for this is the biological heterogeneity (Sinha et al., 2018). However, a short-term life supporter, Extracorporeal Membrane Oxygenation (ECMO), could allow patients to get additional time for medical assistance (Rajsic et al., 2022).

Biomarkers, which have been repeatedly mentioned above, are a key factor in this study. It could be used to assess how effectively the body reacts to therapy. Moreover, it is a biomolecule that exists in blood or other fluids from the body that shows one's physical condition (National Cancer Institute, n.d.).

This project will focus on studying cluster analysis of ARDS and applying two cluster methods, hierarchical clustering and K-means to 450 patients' multiple biomarkers before the ECMO treatment and find out if the spotted clusters relate to the disease survival outcomes.

2. Methodology

The two majority clustering approaches being used in this project are both unsupervised machine learning. Unlabelled datasets are analysed and grouped by utilising machine learning techniques (IBM, 2021).

2.1. Hierarchical Clustering

Hierarchical clustering is an approach that constructs a step-by-step arrangement of clusters by combining or separating them according to their similarities. This results in a dendrogram, a tree-like structure illustrating the links and levels of hierarchy between clusters at multiple stages. This method helps imagines connections among clusters, detecting patterns and unusual data points (Sharma, 2019).

This clustering is broken up into two types, agglomerative hierarchical clustering and divisive hierarchical clustering.

- Agglomerative Hierarchical Clustering

  Agglomerative clustering is an iterative procedure that starts with treating each observation as a separate cluster. As the algorithm progresses, it gradually merges the nearest clusters until all observations are included in a single cluster. It is a bottom-up strategy in which clusters are merged one by one based on a chosen distance or similarity metric.

- Divisive Hierarchical Clustering

  Divisive clustering, also referred as top-down clustering, works oppositely with agglomerative clustering. Starting with every data point from a single cluster, the algorithm repeatedly splits the clusters into two based on a dissimilarity criterion during each stage. This method stops until each data point has its own cluster.

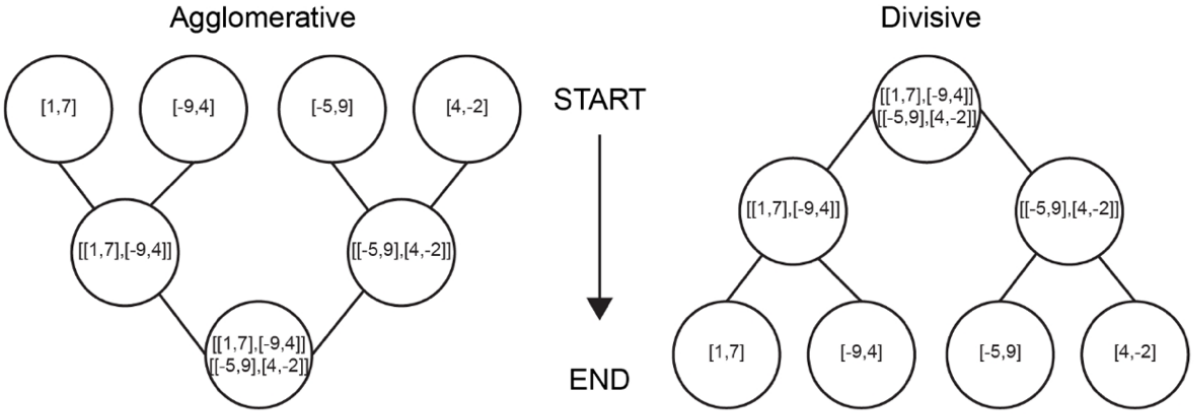


Figure 1: Difference between agglomerative and divisive hierarchical clustering (Johnston, Jones and Kruger, 2019)

2.1.1. Linkage Measures in Hierarchical Clustering

Euclidean distance is a method for calculating the distance between two points in a multi-dimensional space. However, it could only handle distances between pairs of points rather than three or more data points simultaneously. The aim of clustering analysis is to compute the scores that differ between two groups while no single value is associated with each variable (Yim and Ramdeen, 2015). Various measurements are used to determine the distance between two clusters in the linkage approach of clustering. These include single linkage, complete linkage, and average linkage, which respectively define the distance between any pair of points from two clusters as the minimum, maximum, and average distance.

- Single linkage

  This method is also known as the minimum method. With single linkage clustering, the distance between the nearest points of clusters X and Y is calculated. However, it is important to note that single linkage is sensitive to outliers, which might reduce its efficiency.

  The expression for the single linkage function is:

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y)$$

where X and Y represent any two sets of clusters, $D(X,Y)$ is the distance between two clusters and $d(x,y)$ is the distance between elements $x \in X, y \in Y$ (Wikipedia, 2023).

- Complete linkage
  It is also known as the maximum method. With complete linkage, the distance between the furthest points of two clusters is computed. Compared to single linkage, it is less vulnerable to outliers.
  The expression for the complete linkage function is:

  $$D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$$

  where X and Y are any two sets of clusters, $D(X, Y)$ is the distance between two clusters and $d(x, y)$ denotes the distance between two points in each cluster X and Y (Wikipedia, 2023).

- Average linkage
  The average distance between all pairs of data points belonging to two clusters is computed using average linkage. This method strikes a balance between single and complete linkage, resulting in relatively compact clusters.
  The expression for the average linkage function is:

  $$D(X, Y) = T_{XY}/(N_X \times N_Y)$$

  where T_{XY} denotes the sum of all pairwise distances between clusters X and Y, N_X and N_Y are the sizes of the clusters X and Y, respectively (Frontline Solvers, n.d.).

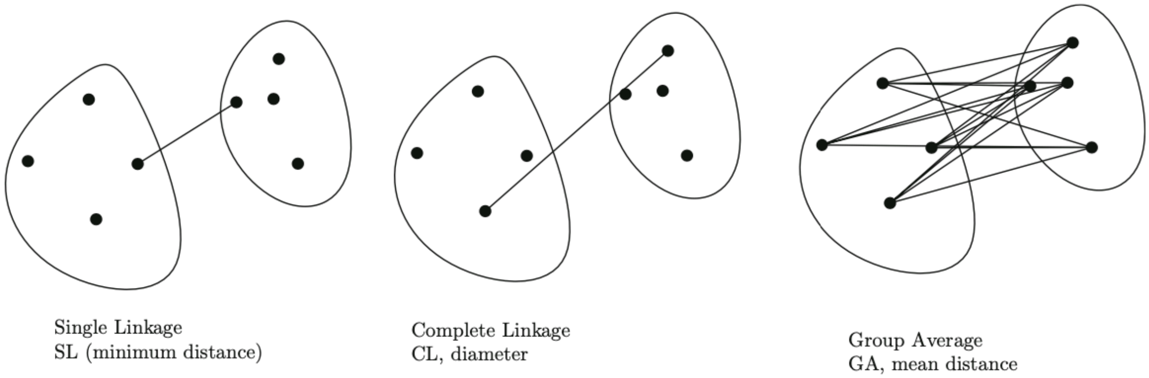


Figure 2: Difference between three different linkages (Nielsen, 2016)

2.2. Partitional Clustering

Partitional clustering, also known as partitioning clustering, is another commonly used strategy in cluster analysis. It aims to separate a dataset into distinct clusters by grouping the data points that share a high degree of similarity together (Data Novia, n.d.).

K-means is one of the well-known approaches for partitioning cluster analysis. It takes the mean value within a cluster as the centroid. By evaluating the squared distance between each data point and the cluster centre, this approach effectively quantifies cluster compactness. Its purpose is to minimise the within-cluster sum of squares.

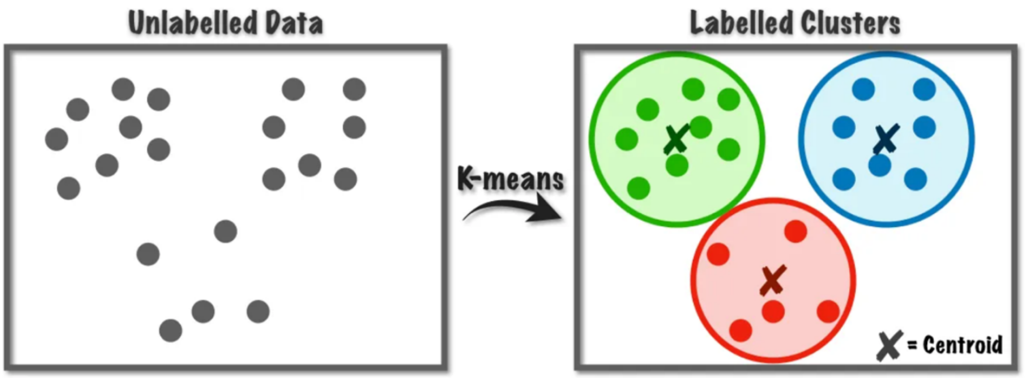


Figure 3: K-means diagram (Jeffares, 2019)

The K-means clustering approach usually takes the Euclidean distance as the metric for evaluating the similarities between data points (Singh, 2013).

The expression for the Euclidean distance function is:

$$D(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

2.3. Dimension reduction

Dimension reduction would be helpful when dealing with high-dimensional data (El Bouchefry and de Souza, 2020). This strategy attempts to keep the most critical details while effectively reducing the total number of dimensions.

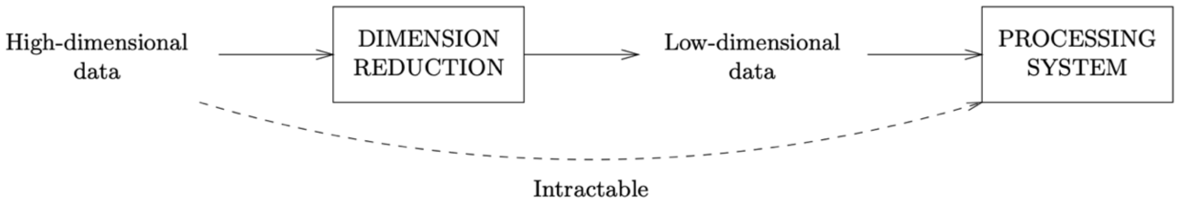


Figure 4: Dimension reduction process (Carreira-Perpiñán, 1997)

One of the dimension reduction methodologies for continuous data is Principal Component Analysis (PCA). It is accomplished by generating principal components by combining *n* different linear combinations of the *n* initial variables and capturing the highest amount of variation in the dataset. Each group of data points may be expressed as a few figures rather than numerous variables with only specific components. This enables samples to visually analyse similarities and differences and determine if they can be categorised (Ringnér, 2008).

2.3.1. Proportion of Explained Variance

While reducing the number of dimensions, keeping information as much as possible is equally important. The proportion of explained variance (PEV) is one of the strategies.

Explained variance is a statistical process that calculates the total amount of variation in the data explained by a set of principal components (Kumar, 2023), and the proportion of explained variance estimates how much information is kept when a certain number of components are included.

Suppose that there are *n* principal components. The proportion of explained variance of the *i*th principal component can be calculated as:

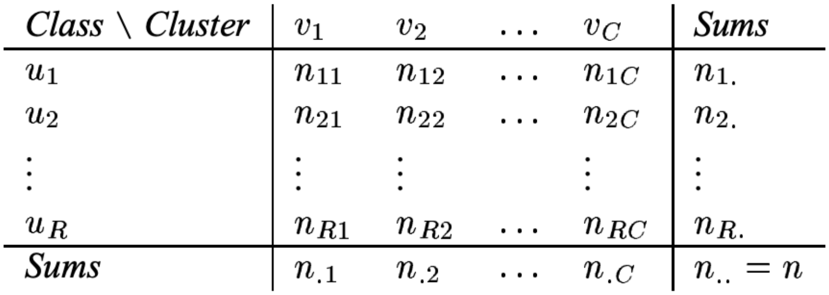
$$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

where λ_i is the *i*th explained variance in the *i*th principal component.

2.4. Adjusted Rand Index

Rand Index (RI) is a method of examining how similar the two data grouping are (Wikipedia, 2023). The RI ranges from 0 to 1, with 0 showing that there is no agreement between the two clustering and 1 indicating that the two data clustering are precisely the same.

Adjusted Rand Index (ARI) is an extension of RI. Assume *u* and *v* are two random partitions containing several clusters with a fixed number of objects. The contingency table can be demonstrated in Figure 5.



Class \ Cluster	v_1	v_2	$\ldots$	v_C	*Sums*
u_1	n_{11}	n_{12}	$\ldots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.C}$	$n_{..} = n$

Figure 5: Adjusted Rand Index Contingency Table (Yeung and Ruzzo, 2001)

where n_{RC} be the number of objects that are in both clusters u_R and v_C, and $n_{R\cdot}$ and $n_{\cdot C}$ denote the number of objects in each cluster separately.

The formula of ARI can be shown as follows:

$$\mathrm{ARI} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}$$

$$= \frac{\sum_{RC}\binom{n_{RC}}{2} - [\sum_{R}\binom{n_{R\cdot}}{2}\sum_{C}\binom{n_{\cdot C}}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_{R}\binom{n_{R\cdot}}{2} + \sum_{C}\binom{n_{\cdot C}}{2}] - [\sum_{R}\binom{n_{R\cdot}}{2}\sum_{C}\binom{n_{\cdot C}}{2}]/\binom{n}{2}}$$

The ARI value can vary between -1 and 1. When the index is the same as the expected value, the ARI will take the value 0.

2.5. Fisher's Exact Test

Fisher's exact test is a statistical test used to assess whether there is a significant relationship between categorical variables (Yu, 2014). It is preferable to the chi-squares test when the sample size is small or at least one expected cell frequency is less than five (Leon, 1998). In other words, the *a*, *b*, *c* and *d* in the contingency table shown in Table1 should be smaller than five.

		Category1		
		Group1	Group2	Total
Category2	Group1	*a*	*b*	*a*+*b*
	Group2	*c*	*d*	*c*+*d*
	Total	*a*+*c*	*b*+*d*	*n*

Table1: Contingency table of the Fisher's exact test

By setting the significant level α, the p-value in Fisher's exact test could indicate whether there is an association between categorical variables.

The p-value could be calculated as below:

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

3. Data Analysis

3.1. Explanatory Data Analysis

Before classification analysis, an explanatory data analysis should be performed. It can provide information on how data is distributed and whether it contains unusual values, such as outliers and missing data.

This data contains numerous biomarkers from 450 patients. There are 65 variables, 62 numeric variables and three-character variables. Among these variables, there are 49 continuous variables containing missing values. To develop the analysis, data preprocessing should be done initially.

3.1.1. Data Preprocessing

Preprocessing data is an essential step in data analysis and machine learning. It entails transforming and processing raw data to make it more suitable for further analysis or modelling.

Plenty of unnecessary variables and couples of possible outliers shows in Figure 6, are contained in the original data. Furthermore, one of the biomarkers-related variables (Albumin) contains 209 missing values which is more than 50% of its data. Although Gupta and Lis (2010) have shown that albumin is associates with the survival, these variables will still need to be removed. The median of the available data for that variable will be replaced by other continuous variables with less than 50% of missing data. In addition, categorical variables will be encoded into a numerical representation. Since the variables are measured differently, scaling the features is also important. It ensures the data are on a similar scale, which can help the algorithms perform better in future tasks.

All the variables that need to be removed, encoded, replaced or scaled will be done in this stage. Finally, the filtered data contains 28 numerical variables and 409 observations.

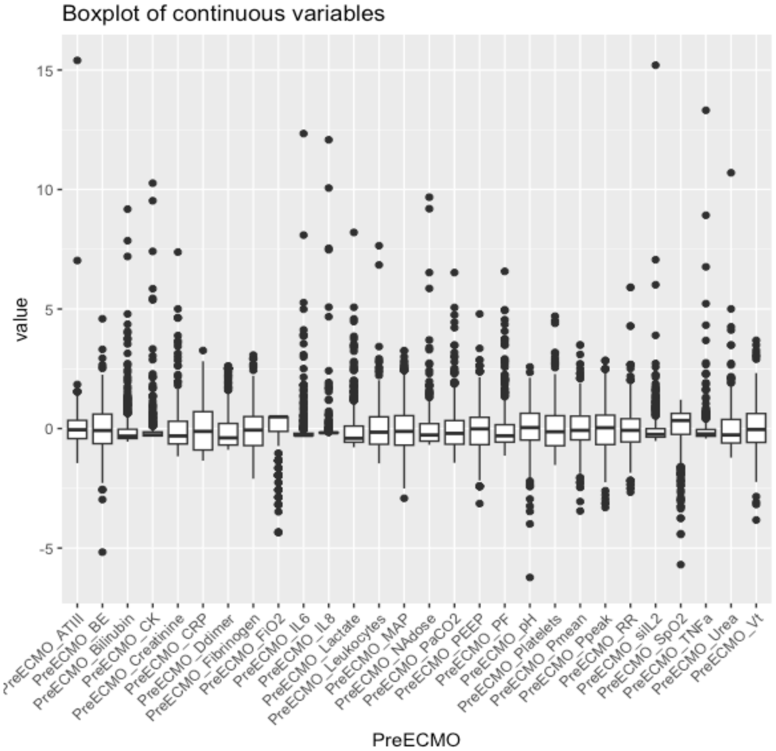


Figure 6: Boxplot of continuous variables

In addition, the skewness of the data will have to be checked as well. The data is normally distributed when the skewness value is close to 0. Figure 7 indicates the skewness value for each variable. It can be said that those variables are highly skewed. Other than that, some negative values are contained. Therefore, offset and log transform should be applied to reduce skewness.

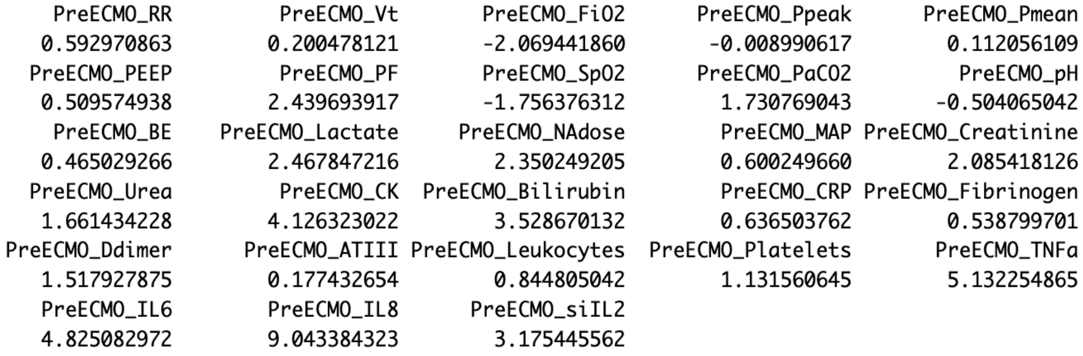

```
      PreECMO_RR         PreECMO_Vt       PreECMO_FiO2      PreECMO_Ppeak      PreECMO_Pmean
     0.592970863        0.200478121       -2.069441860       -0.008990617        0.112056109
    PreECMO_PEEP         PreECMO_PF       PreECMO_SpO2      PreECMO_PaCO2         PreECMO_pH
     0.509574938        2.439693917       -1.756376312        1.730769043       -0.504065042
      PreECMO_BE    PreECMO_Lactate     PreECMO_NAdose        PreECMO_MAP PreECMO_Creatinine
     0.465029266        2.467847216        2.350249205        0.600249660        2.085418126
    PreECMO_Urea         PreECMO_CK  PreECMO_Bilirubin        PreECMO_CRP PreECMO_Fibrinogen
     1.661434228        4.126323022        3.528670132        0.636503762        0.538799701
  PreECMO_Ddimer      PreECMO_ATIII PreECMO_Leukocytes  PreECMO_Platelets       PreECMO_TNFa
     1.517927875        0.177432654        0.844805042        1.131560645        5.132254865
     PreECMO_IL6        PreECMO_IL8      PreECMO_siIL2
     4.825082972        9.043384323        3.175445562
```

Figure 7: Skewness of each variable

Figure 8 concisely presents the pairwise scatterplot between the first ten variables. It shows that the data contains some potential outliers. However, several outliers have already been dropped in the previous section, these outliers will be kept at this moment.

There are significant linear relationships between some pairs of variables. Take two pairs of variables for example, PreECMO_Ppeak and PreECMO_Pmeans, PreECMO_PaCO2 and PreECMO_pH.

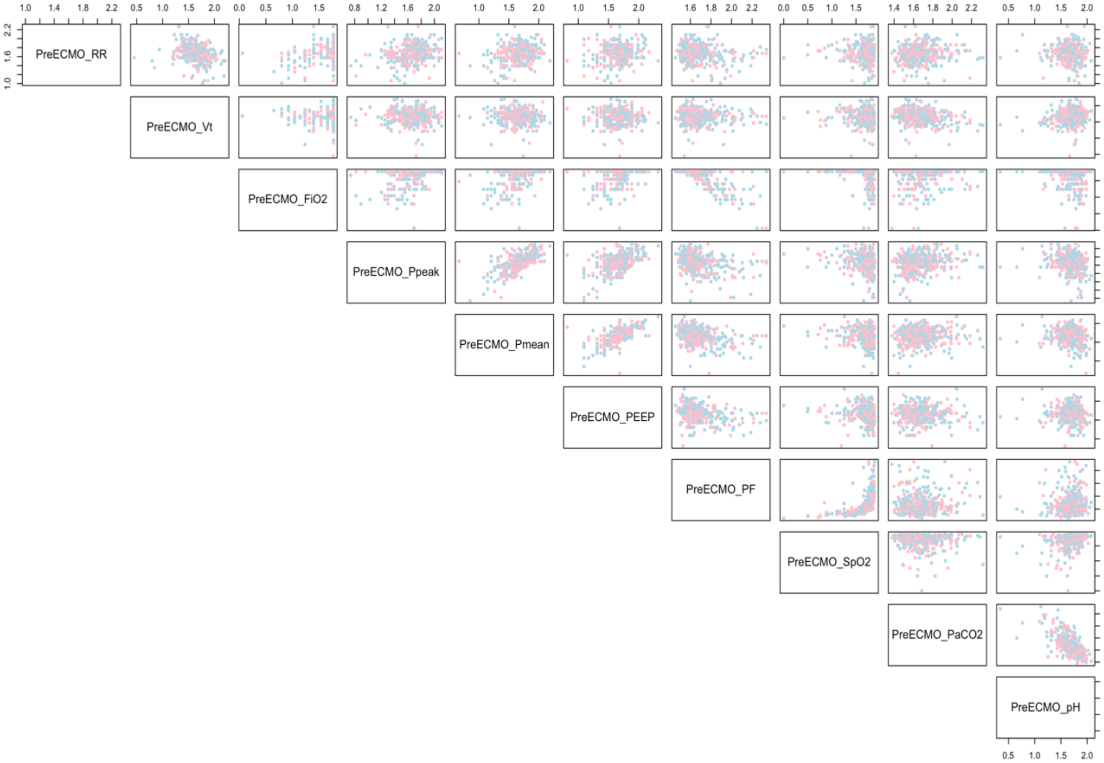


Figure 8: Pair plot

3.1.2. Correlation

For the reason that this data contains numerous observations and variables, dimension reduction is taken into consideration. Principal Component Analysis (PCA) will be used to reduce the dimension of this dataset. Before applying the PCA, the correlation between variables is checked. Figure 9 presents the correlation plot of the PreECMO variables. From the details given in the correlation matrix, it can be noticed that the correlation ranges from relatively low (0.76) to fairly high (0.7) and other correlations in the intermediate range suggest that dimension reduction may be possible.

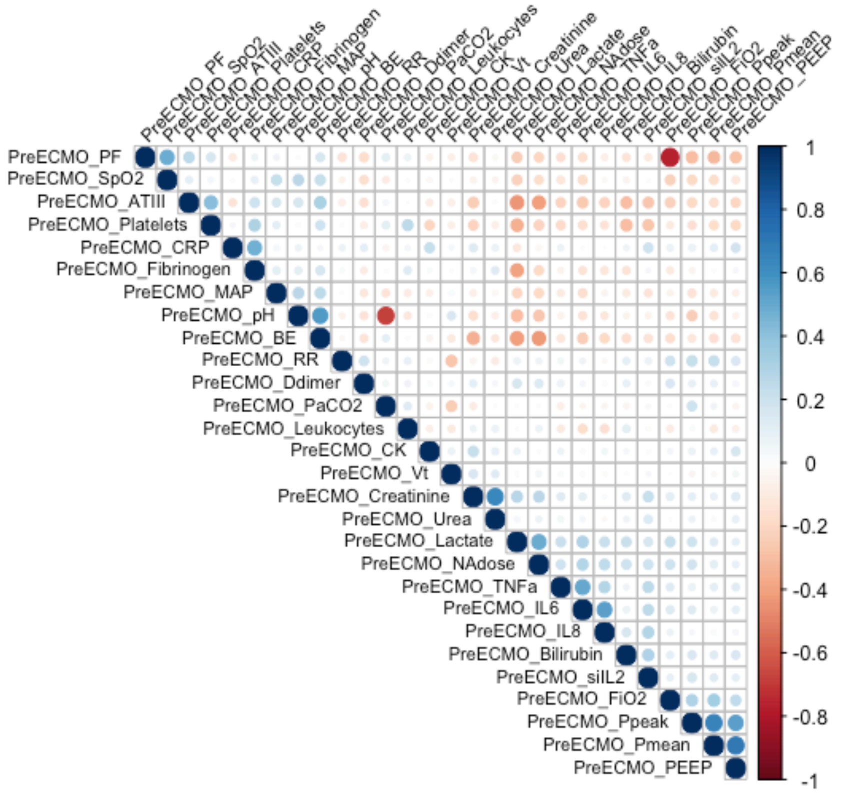
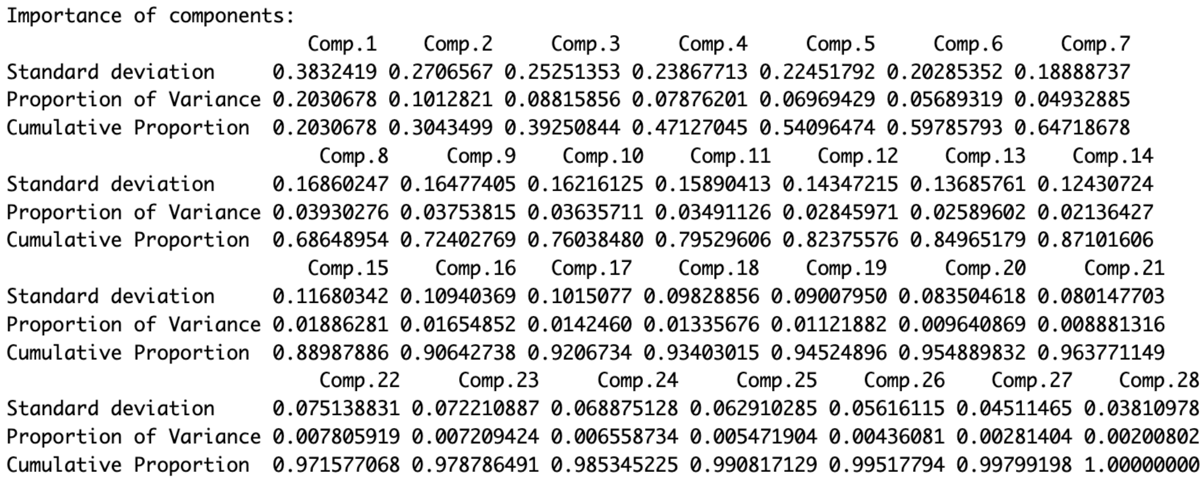


Figure 9: Correlation plot

3.1.3. Dimension Reduction

PCA can be impacted by the scale of the variables (Borgognone, Bussi and Hough, 2001). In the data preprocessing step, the scaling method has priorly standardised the variables and removed the differences in scale. Therefore, when performing the PCA, it is preferred to use the covariance matrix rather than the correlation matrix. The `prcomp()` function in R uses `cor=FALSE`, which means it conducts PCA based on the covariance matrix. It is noticeable that plenty of noisy data remains in the filtered dataset, which might influence the result of PCA. This issue will be left to discuss later.

Figure 10 shows the summary of PCA. This project will investigate at least 80% of the explained variability. By examining the cumulative proportion, the minimum number of the components based on the proportion of explained variance strategy will be 12.

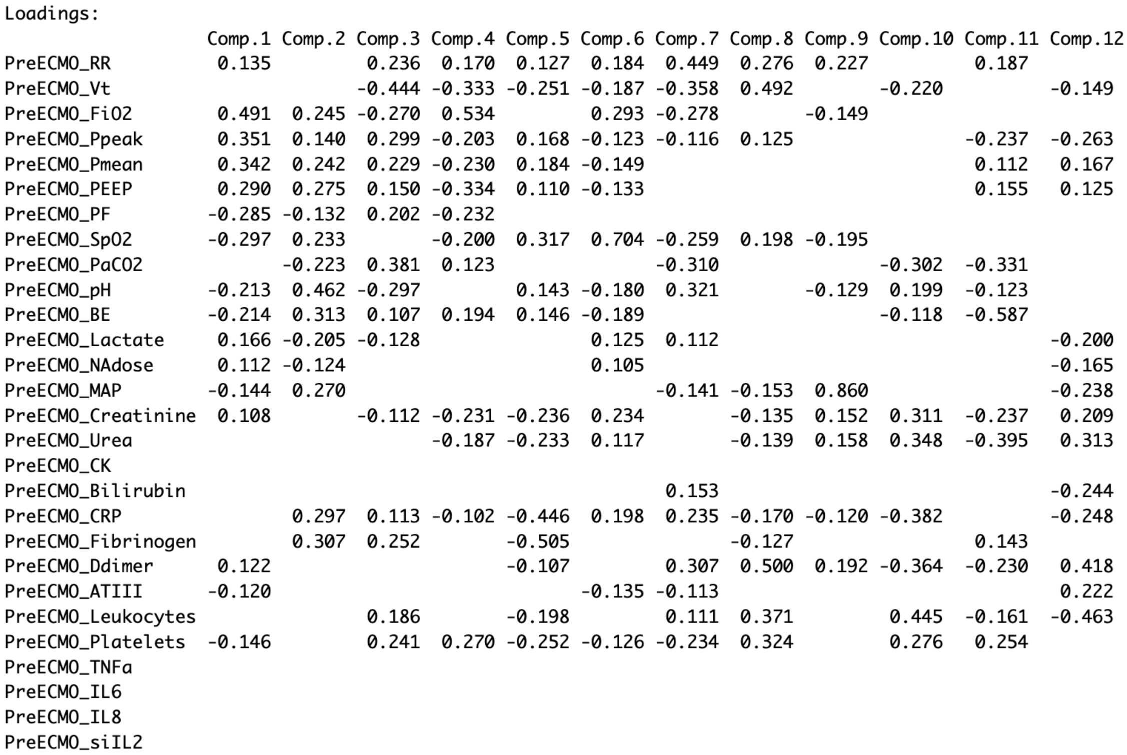


```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
Standard deviation     0.3832419 0.2706567 0.25251353 0.23867713 0.22451792 0.20285352 0.18888737
Proportion of Variance 0.2030678 0.1012821 0.08815856 0.07876201 0.06969429 0.05689319 0.04932885
Cumulative Proportion  0.2030678 0.3043499 0.39250844 0.47127045 0.54096474 0.59785793 0.64718678
                           Comp.8     Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14
Standard deviation     0.16860247 0.16477405 0.16216125 0.15890413 0.14347215 0.13685761 0.12430724
Proportion of Variance 0.03930276 0.03753815 0.03635711 0.03491126 0.02845971 0.02589602 0.02136427
Cumulative Proportion  0.68648954 0.72402769 0.76038480 0.79529606 0.82375576 0.84965179 0.87101606
                          Comp.15    Comp.16   Comp.17    Comp.18    Comp.19     Comp.20     Comp.21
Standard deviation     0.11680342 0.10940369 0.1015077 0.09828856 0.09007950 0.083504618 0.080147703
Proportion of Variance 0.01886281 0.01654852 0.0142460 0.01335676 0.01121882 0.009640869 0.008881316
Cumulative Proportion  0.88987886 0.90642738 0.9206734 0.93403015 0.94524896 0.954889832 0.963771149
                           Comp.22     Comp.23     Comp.24     Comp.25    Comp.26    Comp.27    Comp.28
Standard deviation     0.075138831 0.072210887 0.068875128 0.062910285 0.05616115 0.04511465 0.03810978
Proportion of Variance 0.007805919 0.007209424 0.006558734 0.005471904 0.00436081 0.00281404 0.00200802
Cumulative Proportion  0.971577068 0.978786491 0.985345225 0.990817129 0.99517794 0.99799198 1.00000000
```

Figure 10: Principal Component Analysis Summary

Loading for those PCs is displayed in Figure 11. Loading entry in PCA is the coefficients that specify how much each variable contributes to the principal components. Variables with positive loadings have a positive relationship with the principal component. On the other hand, variables with negative loadings have a negative relationship with the principal component (Holland, 2019).

Figure 11 shows the loadings with the 12 chosen PCs. It could be seen that there are five variables (PreECMO_CK, PreECMO_TNFa, PreECMO_IL6, PreECMO_IL8 and PreECMO_siIL2) that do not contribute to the data.



```
Loadings:
                   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12
PreECMO_RR          0.135         0.236  0.170  0.127  0.184  0.449  0.276  0.227           0.187
PreECMO_Vt                       -0.444 -0.333 -0.251 -0.187 -0.358  0.492         -0.220          -0.149
PreECMO_FiO2        0.491  0.245 -0.270  0.534         0.293 -0.278        -0.149
PreECMO_Ppeak       0.351  0.140  0.299 -0.203  0.168 -0.123 -0.116  0.125                 -0.237  -0.263
PreECMO_Pmean       0.342  0.242  0.229 -0.230  0.184 -0.149                                 0.112   0.167
PreECMO_PEEP        0.290  0.275  0.150 -0.334  0.110 -0.133                                 0.155   0.125
PreECMO_PF         -0.285 -0.132  0.202 -0.232
PreECMO_SpO2       -0.297  0.233        -0.200  0.317  0.704 -0.259  0.198 -0.195
PreECMO_PaCO2             -0.223  0.381  0.123               -0.310                -0.302  -0.331
PreECMO_pH         -0.213  0.462 -0.297         0.143 -0.180  0.321        -0.129   0.199  -0.123
PreECMO_BE         -0.214  0.313  0.107  0.194  0.146 -0.189                       -0.118  -0.587
PreECMO_Lactate     0.166 -0.205 -0.128                0.125  0.112                                 -0.200
PreECMO_NAdose      0.112 -0.124                       0.105                                        -0.165
PreECMO_MAP        -0.144  0.270                             -0.141 -0.153  0.860                   -0.238
PreECMO_Creatinine  0.108        -0.112 -0.231 -0.236  0.234        -0.135  0.152   0.311  -0.237   0.209
PreECMO_Urea                            -0.187 -0.233  0.117        -0.139  0.158   0.348  -0.395   0.313
PreECMO_CK
PreECMO_Bilirubin                                             0.153                                 -0.244
PreECMO_CRP                0.297  0.113 -0.102 -0.446  0.198  0.235 -0.170 -0.120  -0.382          -0.248
PreECMO_Fibrinogen         0.307  0.252        -0.505               -0.127                   0.143
PreECMO_Ddimer      0.122                      -0.107         0.307  0.500  0.192  -0.364  -0.230   0.418
PreECMO_ATIII      -0.120                             -0.135 -0.113                                  0.222
PreECMO_Leukocytes                0.186        -0.198         0.111  0.371          0.445  -0.161  -0.463
PreECMO_Platelets  -0.146         0.241  0.270 -0.252 -0.126 -0.234  0.324          0.276   0.254
PreECMO_TNFa
PreECMO_IL6
PreECMO_IL8
PreECMO_siIL2
```

Figure 11: Loading for chosen PCs

3.2. Classification Analysis

This section will go through two clustering methods: hierarchical cluster analysis and K-means. Determining the optimal number of clusters *k* is crucial in these two analyses. To split the data into *k* clusters, it may be calculated using several techniques, including the elbow method, average silhouette method and gap statistic method. This project will rely on the average silhouette method.

Briefly, the average silhouette method calculates the average silhouette of observations for various *k* values (Et-Taleby, Boussetta and Benslimane, 2020). The one that maximises the average silhouette across the range of *k* value is the optimal number of clusters. A large average silhouette width suggests that the clustering is effective.

3.2.1. Hierarchical Cluster Analysis

Firstly, a line plot using silhouette analysis in Figure 12 represents the ideal number of clusters. It shows that the optimal is when the number of clusters is 2. By looking further at Figure 13, the silhouette plot implies that the average linkage has the highest average silhouette width, making the clustering the most functional while grouping the data. It is noticeable that there are few observations with negative average silhouette width in the blue cluster, suggesting that these observations are poorly assigned to their cluster or misclassified. Under the circumstance that the influence on the clustering is fragile, those observations will not be removed.

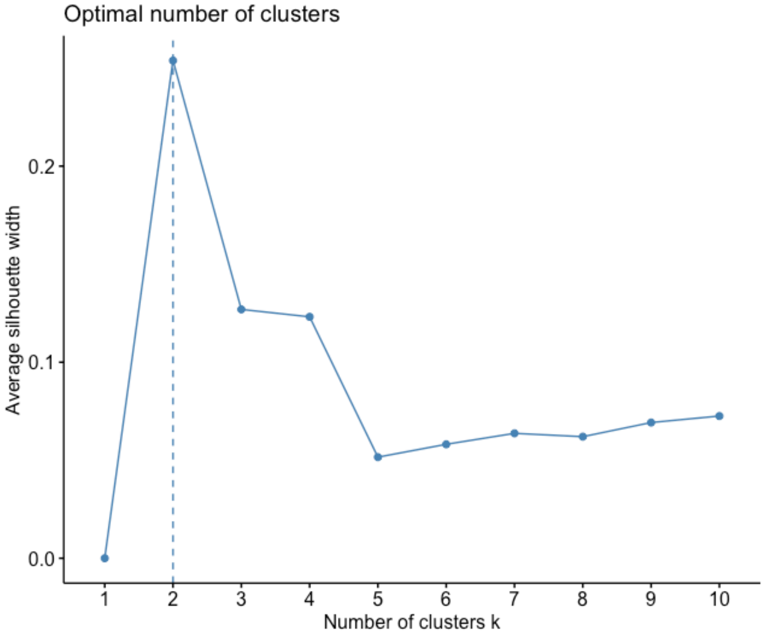


Figure 12: Optimal number of clusters for hierarchical clustering

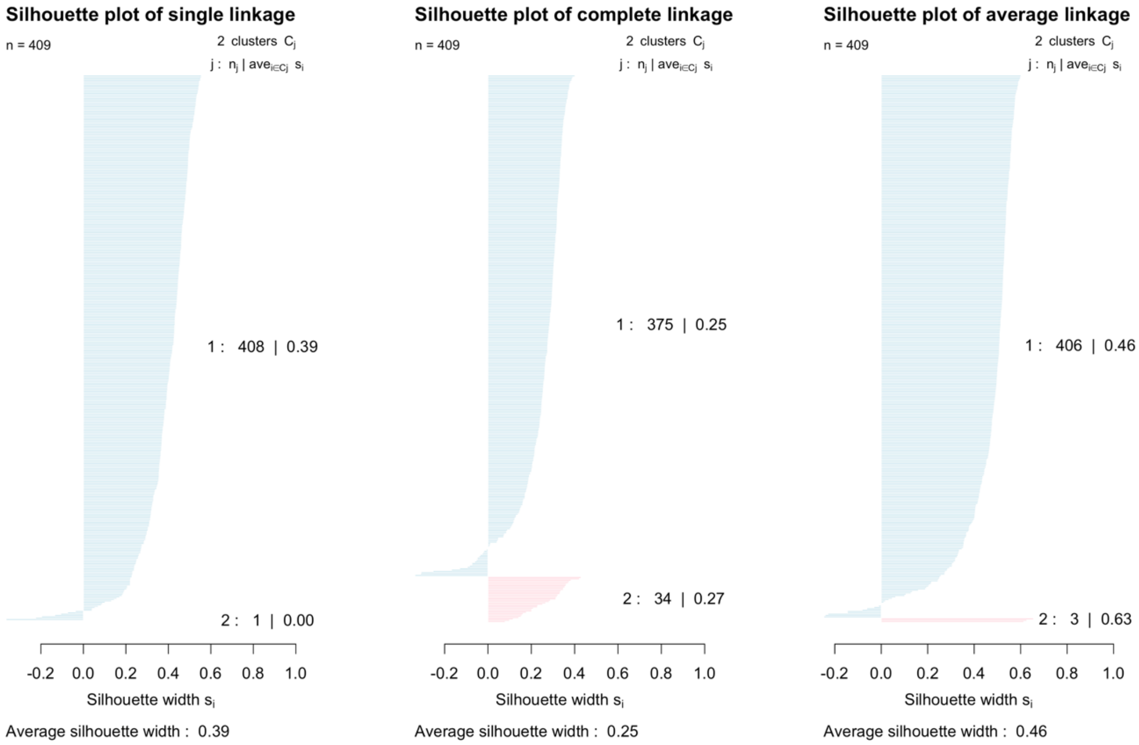


Figure 13: Silhouette plot of single, complete and average linkage

The dendrogram can be plotted after confirming the best number of clusters and linkage method. Clustering tree in Figure 14 separates the tree into two partitions and displays each cluster with a different colour. There are three observations in the blue cluster, and the rest are in another cluster.

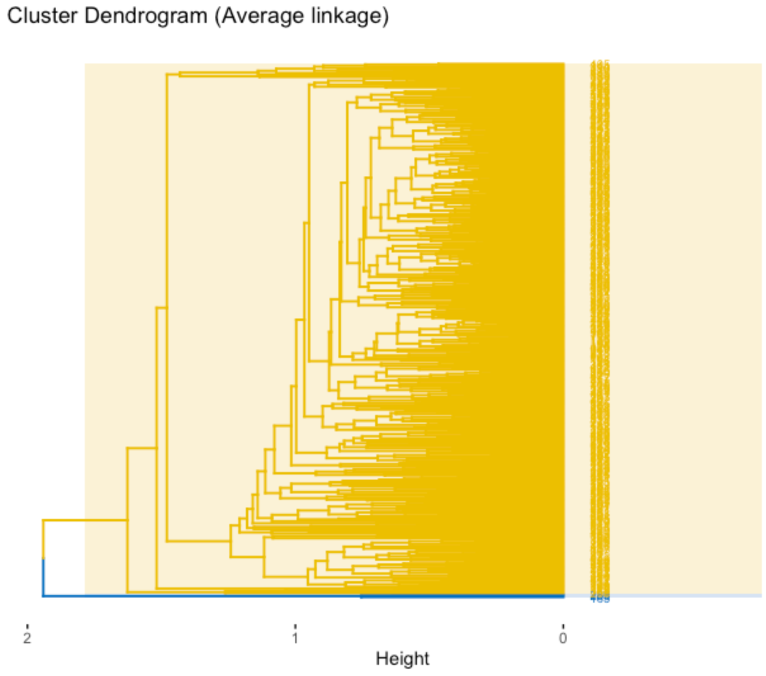


Figure 14: Cluster Dendrogram (Average linkage)

However, this result doesn't go well. The result will not be indicative when only a few observations are in one cluster and the remaining in another. Some possible reasons will be discussed further in the next chapter.

3.2.2. K-means

Using the same measurement as previous to determine the optimal number of clusters for K-means, the ideal number of clusters is found to be at 2.

Figure 15 briefly shows the K-means cluster plots between the first five chosen PCs. Comparing all the cluster plots, those that are associated with component 1 have a clear separation. This implies that the variables in Component 1 are representative of cluster results, and the chosen components effectively capture the difference between clusters.

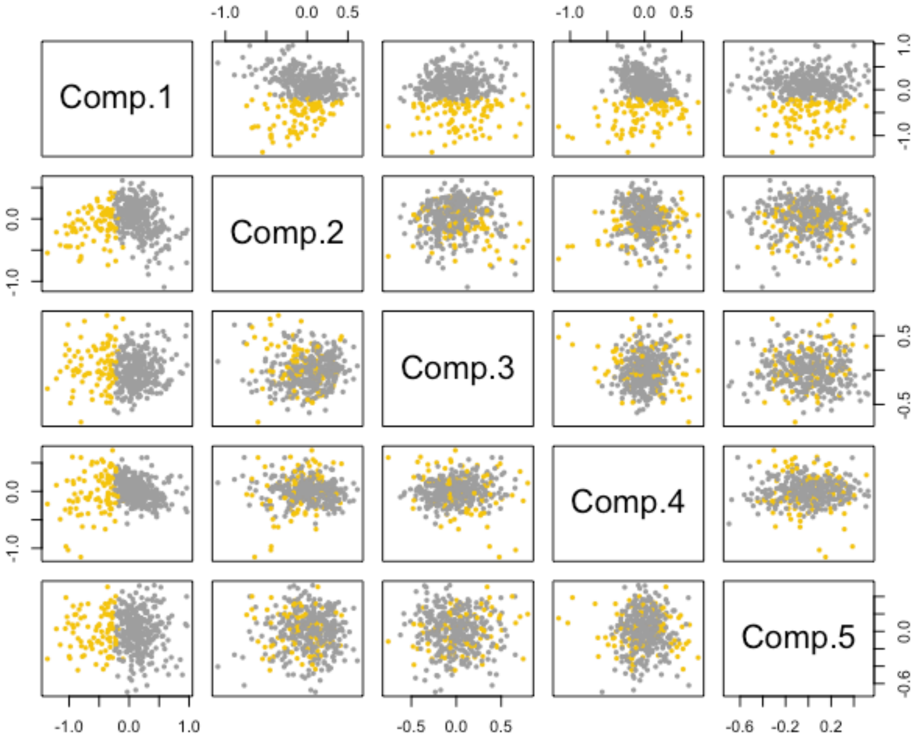


Figure 15: K-means cluster plot between the first five PCs

As a consequence of the first component is more representative while splitting data, in this section, other components will be compared with it in order to take a further look at the classification. In this case, clusters with substantial levels of similarity and clusters with considerable degrees of dissimilarity will be picked. The adjusted rand index (ARI) will be utilised.

- Two clusters with high similarity

  Table 2 suggests that the strongest similarity between the two clusters was identified at the maximum ARI value (0.034), which is constructed by component 1 and component 2, and the cluster plot is shown in Figure 16.

		Component 1		
		Group1	Group2	Total
Component 2	Group1	56	53	109
	Group2	196	104	300
	Total	252	157	409

Table 2: Contingency table (Component 1 by Component 2)

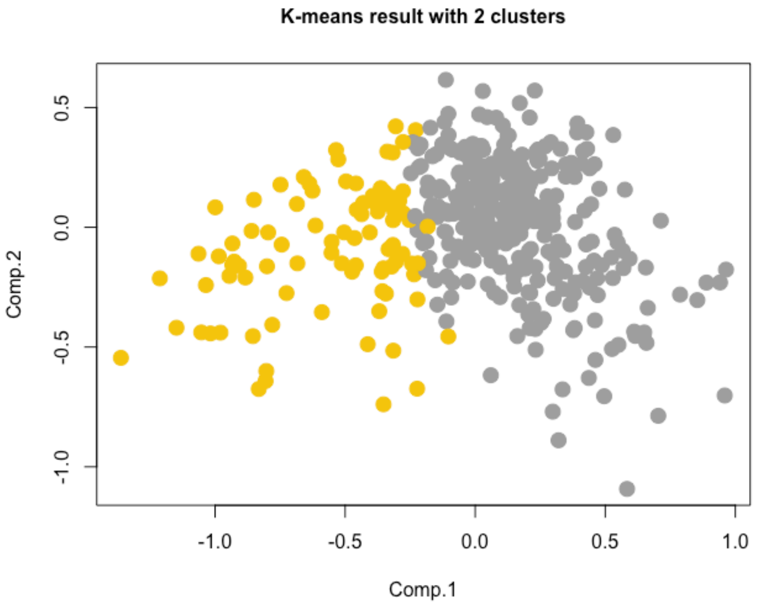


Figure 16: K-means cluster plot (Component 1 by Component 2)

- Two clusters with high dissimilarity
  The lowest ARI value indicates that those two clusters being compared have the biggest dissimilarity. The ARI value could be calculated through Table 3, which the value in this data is approximately -0.003, and the cluster plot formed by components 1 and 9 in Figure 17 pictures the clusters with the largest dissimilarity.

		Component 1		
		Group1	Group2	Total
Component 9	Group1	61	48	109
	Group2	163	137	300
	Total	224	185	409

Table 3: Contingency table (Component 1 by Component 9)

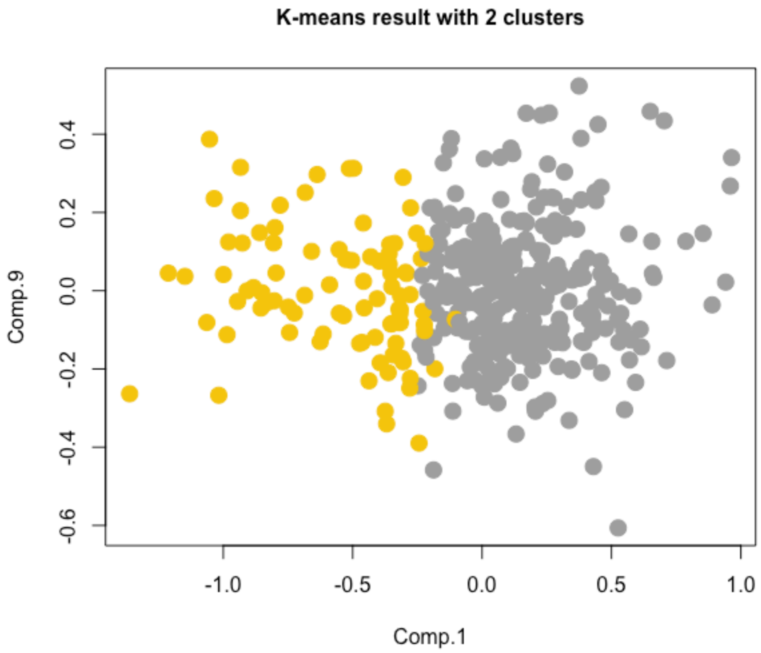


Figure 17: K-means cluster plot (Component 1 by Component 9)

The comparison of the two distinct plots above demonstrates that while comparing with the first component, the clusters are clearly classified with either high or low values of ARI.

3.3. Fisher's exact test

The Fisher's exact test is used here to test whether there is a significant association between cluster results from the above clustering analysis and outcome variables for Hospital_Survival at 5% level of significance.

Considering that the result given in hierarchical clustering did not work well, this section will only focus on testing K-means results.

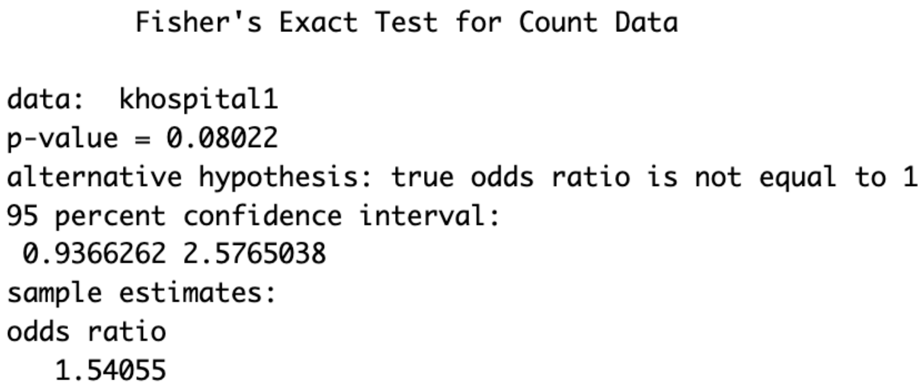
The null hypothesis and alternative hypothesis in this test will be set as below:

$$\begin{cases} H_0: \text{There is no association between clustering result and survival} \\ H_1: \text{There is significant association between clustering result and survival} \end{cases}$$

- First Component

  By reason of the first component is the most representative, it will used to examine whether the grouping results from is corresponds to the survival.

  Figure 18 indicates Fisher's exact test. The null hypothesis is not rejected with a p-value of 0.08>0.05. Suggesting that the clustering result and the outcome variable for survival are unrelated.



```
         Fisher's Exact Test for Count Data

data:  khospital1
p-value = 0.08022
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9366262 2.5765038
sample estimates:
odds ratio
   1.54055
```

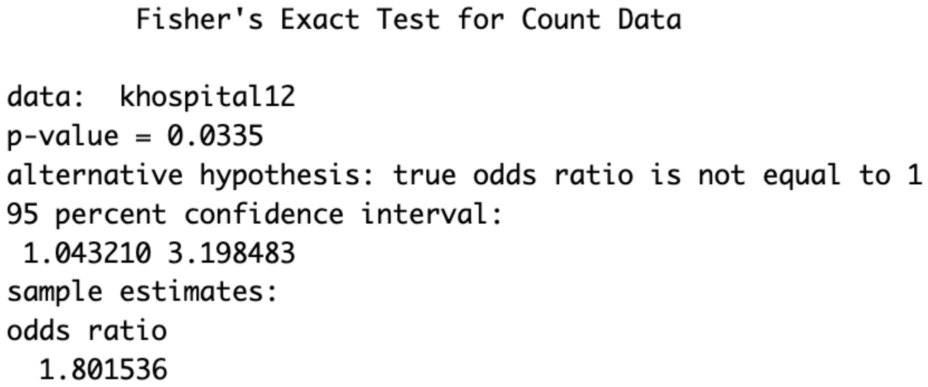
Figure 18: Fisher's Exact Test for the first component

From the result above, although component 1 is the most illustrative while grouping the data, there is no evidence showing that there is a significant relation between those variables and the outcome.

In order to check if selecting at least 80% of the explained variability (12 components) is sufficient to signify the entire dataset, the components containing at least 80% and 100% of the explained variability will be tested next.

- 12 Components

  Figure 19 presents the Fisher's exact test for the association between the first 12 components' clustering and survival. Given that the p-value in this test is 0.04<0.05, the null hypothesis is rejected, and the odds ratio in this test is greater than one (1.80), meaning that the association between the two groups is positive.



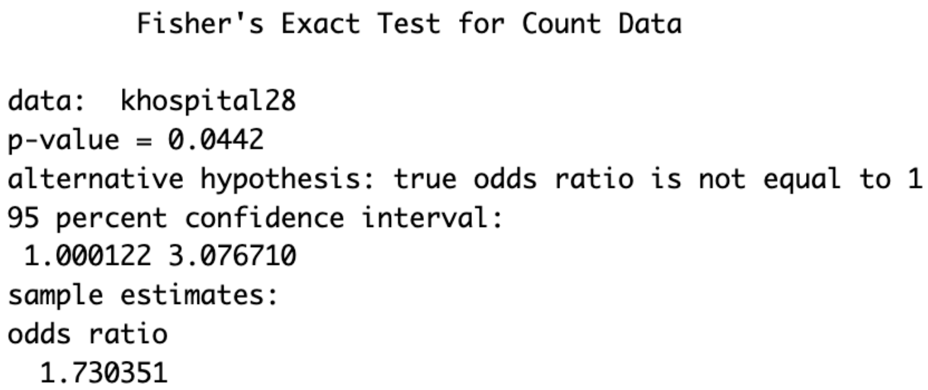
```
         Fisher's Exact Test for Count Data

data:  khospital12
p-value = 0.0335
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.043210 3.198483
sample estimates:
odds ratio
  1.801536
```

Figure 19: Fisher's Exact Test for 12 Components

- Entire Components

  The Fisher's exact test for the association between the clustering result with the 28 components grouping and survival is shown in Figure 20. With the p-value of 0.04<0.05, the null hypothesis is rejected. It appears that the clustering result and the outcome variable for survival are associated. Also, the odds ratio in this test is greater than one (1.73). That means the association between the groups which were examined is positive.



```
                Fisher's Exact Test for Count Data

data:  khospital28
p-value = 0.0442
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.000122 3.076710
sample estimates:
odds ratio
  1.730351
```

Figure 20: Fisher’s Exact Test for the entire components

4. Conclusion and Discussion

4.1. Conclusion

It is obvious that there is a great difference between the two clustering results. Despite the fact that both clustering approaches separated the data into two clusters, K-means performed better than hierarchical clustering. Under this circumstance, only the K-means clustering findings were examined to check if they are corresponded to the experimental variable for survival.

Based on Fisher’s exact for this data, although component 1 is the most representative while grouping the data, there is no evidence showing that there is a significant relation between those variables and the outcome. However, it is reasonable to conclude that either the components contain 80% of explained variability or the entire components are associated to the survival. In other words, it suggests that the collective information captured by those two groups of components provides more details into the connection with the outcome variable than the first component alone.

4.2. Discussion

There are several possible reasons why hierarchical clustering failed to work effectively in this project. Firstly, outliers. Outliers can influence the linkage. Even if multiple unusual data points were removed before formal data analysis, additional possible observations in the data might still be considered outliers while conducting hierarchical clustering. Secondly, the distance metrics. In this project, Euclidean distance was used to calculate the linkage. However, different distance metrics will also influence the result of hierarchical clustering. Last but not least, the linkage method. According to average silhouette width, average linkage is chosen in this project. However, Figure 12 shows that the clustering outcome for complete linkage separates data into two groups. One has 375 observations, whereas the other has 34 observations. This would perform better than the others when comparing the three linkage strategies.

Section 3.3 results in both the first 12 components and entire components having an association with the outcome, it implies that the principal components successfully preserved important patterns in the data. Moreover, even though there are several meaningless data being stored, those relevant data are powerful enough to be distinguished.

5. Future Extension

Given the time constraints and data limitations, other clustering methods could not be compared, and most importantly, better hierarchical clustering outcomes could not be tested. To make the result more reliable, more factors could be included in the analysis in the future. For example, gender, age, disease or medical condition and lifestyle. These may affect whether a patient survives. Furthermore, it might be more precise if alternative linkage methods could be applied in hierarchical clustering and additional cluster algorithms were computed.

References

Henk Jan van den Ham *et al.* (2009). 'Differential cytokine profiles in juvenile idiopathic arthritis subtypes revealed by cluster analysis', *Rheumatology*, 48(8), pp. 899–905.

Wierenga, A.P.A. *et al.* (2019). 'Aqueous Humor Biomarkers Identify Three Prognostic Groups in Uveal Melanoma', *Investigative Ophthalmology & Visual Science*, 60(14), pp. 4740–4747.

Jain, A.K. (2008). 'Data Clustering: 50 Years Beyond K-means', *Machine Learning and Knowledge Discovery in Databases*, pp. 3–4.

Chiu, C.Y., Douglas, J.A. and Li, X. (2009). 'Cluster analysis for cognitive diagnosis: Theory and applications', *Psychometrika*, 74(4), pp. 633–665.

Bos, L.D. *et al.* (2017). 'Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis', *Thorax*, 72(10), pp. 876–883.

Sinha, P. *et al.* (2018). 'Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study', *Intensive Care Medicine*, 44(11), pp. 1859–1869.

Rajsic, S. *et al.* (2022). 'ECMO in Cardiogenic Shock: Time Course of Blood Biomarkers and Associated Mortality', *Diagnostics*, 12(12), p. 2963.

National Cancer Institute (no date). 'Definition of biomarker' [online]. Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker (Accessed: 18 June 2023).

IBM (2021). 'What is Unsupervised Learning?' [online]. Available at: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/ (Accessed: 13 July 2023).

Sharma, P. (2019) 'What is Hierarchical Clustering in Python?' [online]. Available at: https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/#h-what-is-hierarchical-clustering (Accessed: 18 June 2023).

Johnston, B., Jones, A. and Kruger, C. (2019) 'Applied Unsupervised Learning with Python Table of Contents Preface' [online]. Available at: https://www.packtpub.com/product/applied-unsupervised-learning-with-python/9781789952292 (Accessed: 15 August 2023).

Yim, O. and Ramdeen, K.T. (2015) 'Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data', *The quantitative methods for psychology*, 11(1), pp. 8–21.

Wikipedia (2023). 'Single-linkage clustering' [online]. Available at: https://en.wikipedia.org/wiki/Single-linkage_clustering (Accessed: 21 June 2023).

Wikipedia (2023). 'Complete-linkage clustering' [online]. Available at: https://en.wikipedia.org/wiki/Complete-linkage_clustering (Accessed: 21 June 2023).

Frontline Solvers (no date). 'Hierarchical Clustering' [online]. Available at: https://www.solver.com/xlminer/help/hierarchical-clustering-intro (Accessed: 21 June 2023).

Nielsen, F. (2016). 'Hierarchical Clustering', *Introduction to HPC with MPI for Data Science*, pp.195–211.

Data Novia (no date). ‘Partitional Clustering in R: The Essentials’ [online]. Available at: https://www.datanovia.com/en/courses/partitional-clustering-in-r-the-essentials/ (Accessed: 28 June 2023).

Jeffares, A. (2019). ‘K-means: A Complete Introduction.’ [online]. Available at: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c (Accessed: 26 July 2023).

Singh, A. (2013). ‘K-means with Three different Distance Metrics’, *International Journal of Computer Applications*, 67(10).

El Bouchefry, K. and de Souza, R.S. (2020). ‘c’, *Knowledge Discovery in Big Data from Astronomy and Earth Observation: Astrogeoinformatics*, pp. 225–249.

Carreira-Perpiñán, M.A. (1997). ‘A Review of Dimension Reduction Techniques’, *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, *9,* pp. 1–69.

Ringnér, M. (2008). ‘What is principal component analysis?’, *Nature Biotechnology*, 26(3), pp. 303–304.

Kumar, A. (2023). ‘PCA Explained Variance Concepts with Python Example’ [online]. Available at: https://vitalflux.com/pca-explained-variance-concept-python-example/#What_is_Explained_Variance (Accessed: 26 June 2023).

Wikipedia (2023). ‘Rand index’ [online]. Available at: https://en.wikipedia.org/wiki/Rand_index#cite_note-rand71-1 (Accessed: 27 June 2023).

Yeung, K.Y. and Ruzzo, W.L. (2001). ‘Details of the Adjusted Rand index and Clustering algorithms, supplement to the paper an empirical study on Principal Component Analysis for clustering gene expression data’, *Bioinformatics*, 17(9), pp. 763–774.

Yu, Y. (2014). ‘Tests of Independence in a Single 2x2 Contingency Table with Random Margins’, *Worcester Polytechnic Institute.*

Leon, A.C. (1998). ‘Descriptive and Inferential Statistics’, *Comprehensive Clinical Psychology*, pp. 243–285.

Gupta, D. and Lis, C.G. (2010). ‘Pretreatment serum albumin as a predictor of cancer survival: A systematic review of the epidemiological literature’, *Nutrition Journal*, 9(1), pp. 1–16.

Borgognone, M.G., Bussi, J. and Hough, G. (2001) ‘Principal component analysis in sensory analysis: covariance or correlation matrix?’, *Food Quality and Preference*, 12(5–7), pp. 323–326.

Holland, S.M. (2008). ‘Principal components analysis (PCA)’, *Department of Geology, University of Georgia, Athens, GA, 30602, 2501.*

Et-Taleby, A., Boussetta, M. and Benslimane, M. (2020). ‘Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image’, *International Journal of Photoenergy*, pp. 1–7.