Analysis of Cardiovascular Diseases

Ridhiwan Ramadhan, R.R.M., MSEYA

Middle East Technical University Department of Chemistry, ridhiwanmseya@gmail.com

İlay, İ.B., BAYBAĞAN

Middle East Technical University Department of Geodesy and Geography Information Technologies,
ilaybaybagan@gmail.com

In this paper, the dataset about cardiovascular disease will be analyzed. The dataset has several attributes such as age, gender, height, weight, blood pressure types, cholesterol, glucose, smoking, alcohol, physical activity and finally cardiovascular disease. In this study, effects of smoking and alcohol consumption on cardiovascular disease with the hypothesis about the age attribute. For this study, some data statistical methods will be applied. Firstly, projection methods which are Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), Isomaps, Local Linear Embedding will be applied on dataset and after that HierarchicalClustering, K-Means Clustering and PAM Clustering will be performed and finally, Internal and External Validation will be carried out.

Keywords: Data Analysis**,** Cardiovascular Disase, Projection, PCA, MDS, nMDS, Hierarchical Clustering, K-means Clustering, PAM Clustering, Validation

1 INTRODUCTION

Cardiovascular disease is a type of disease s which involve blood or heart vessels. Cardiovascular disease covers coronary artery diseases like heart attack. There are many different mechanisms which trigger cardiovascular disease. In this study, some of these mechanisms will be examined and their effects on cardiovascular disease will be revealed. These mechanisms can be high blood pressure, lack of exercise, obesity, smoking, high blood cholesterol, overabundant alcohol consumption, poor diet, etc. For looking the relationship with these mechanisms and cardiovascular disease, a dataset is chosen about cardiovascular disease.

1.1 Dataset Details

The chosen dataset is "Cardiovascular Disease Dataset (clean)". This dataset was used in previous assignments and for this paper, it will examined again with different perspective. The dataset is clean and it gave good results previous studies. Therefore, for this study, same dataset is chosen. The used dataset which is clean version [1] and its parent dataset [2] are available at Kaggle website. This dataset reveals the relationship between cardiovascular disease and some mechanisms by associating it with different parameters such as age, height and weight. Therefore, it will be analyzed again for this study from different perspectives.

The dataset consists of 68783 samples and 12 attributes. The attributes are:

- AGE: integer (years of age)
- GENDER: categorical (0: female, 1: male)
- HEIGHT: integer (cm)
- WEIGHT: integer (kg)
- AP_HIGH: systolic blood pressure, integer
- AP_LOW: diastolic blood pressure, integer
- CHOLESTEROL: categorical (1: normal, 2: above normal, 3: well above normal)
- GLUCOSE: categorical (1: normal, 2: above normal, 3: well above normal)
- SMOKE: categorical (0: no, 1: yes)
- ALCOHOL: categorical (0: no, 1: yes)
- PHYSICAL_ACTIVITY: categorical (0: no, 1: yes)
- CARDIO_DISEASE: categorical (0: no, 1: yes)

The last attribute is the "target" column that indicates whether the individual has a cardiovascular disease or not using the numbers 1 and 0. Moreover, the characteristics of attributes are examined and given in the Table 1.

Table 1 Characteristics of attributes

Name of Attribute	Value	Type of Attribute	
Age	30-65	Continuous Variable	Ratio
Gender	0: Female, 1: Male	Binary Variable	Nominal
Height	55-250	Continuous Variable	Ratio
Weight	11-200	Continuous Variable	Ratio
Ap_High	60-240	Continuous Variable	Ratio
Ap_Low	40-190	Continuous Variable	Ratio
Cholesterol	1: Normal, 2: Above Normal, 3: Well Above Normal	Class Variable	Ordinal
Glucose	1: Normal, 2: Above Normal, 3: Well Above Normal	Class Variable	Ordinal
Smoke	0: No, 1: Yes	Binary Variable	Nominal
Alcohol	0: No, 1: Yes	Binary Variable	Nominal
Physical_Activity	0: No, 1: Yes	Binary Variable	Nominal
Cardio_Disease	0: No, 1: Yes	Binary Variable	Nominal
Age	30-65	Continuous Variable	Ratio

1.2 Question and Hypothesis

In order to look at relationship with some attributes and cardiovascular diseases, the mechanisms to be examined should be specified. For this reason, firstly a question and a hypothesis are determined and then some analyzes will be made on them. For this study, the question is "What is the relation between age and cardiovascular diseases and how can external factors like smoking and alcohol consumption affect this relation?" and the hypothesis will be "Heart diseases mostly affect people above 50 years of age but alcohol and smoking can exurcibate the effects regardless of the gender."

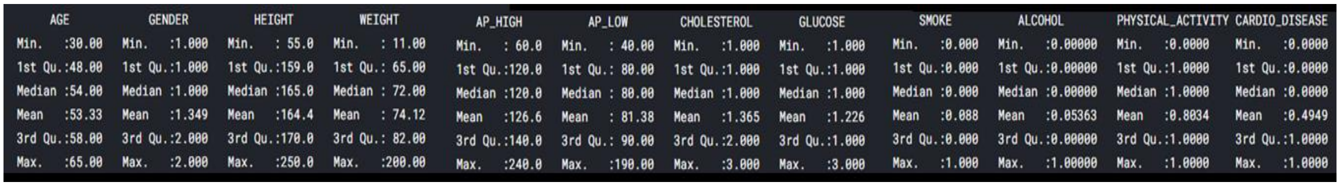
2 PRELIMINARY ANALYSIS

Preliminary analysis will consist of inspection of the dataset to see how clean it is and to have a clear a view of its structure. Therefore, after seeing the structure of the dataset, some projections will be tried to have an idea of what we can expect as we work on the data further down the line.

2.1 Structure of Dataset

➢ *Summary of Dataset*

Firstly, the summary of the dataset was created ant the results are given in Figure 1 that is the output of the code which is available in Appendix A.



```
      AGE            GENDER          HEIGHT          WEIGHT           AP_HIGH          AP_LOW         CHOLESTEROL      GLUCOSE          SMOKE           ALCOHOL       PHYSICAL_ACTIVITY CARDIO_DISEASE
Min.   :30.00   Min.   :1.000   Min.   : 55.0   Min.   : 11.00   Min.   : 60.0   Min.   : 40.00   Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:48.00   1st Qu.:1.000   1st Qu.:159.0   1st Qu.: 65.00   1st Qu.:120.0   1st Qu.: 80.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000
Median :54.00   Median :1.000   Median :165.0   Median : 72.00   Median :120.0   Median : 80.00   Median :1.000   Median :1.000   Median :0.000   Median :0.00000   Median :1.0000   Median :0.0000
Mean   :53.33   Mean   :1.349   Mean   :164.4   Mean   : 74.12   Mean   :126.6   Mean   : 81.38   Mean   :1.365   Mean   :1.226   Mean   :0.088   Mean   :0.05363   Mean   :0.8034   Mean   :0.4949
3rd Qu.:58.00   3rd Qu.:2.000   3rd Qu.:170.0   3rd Qu.: 82.00   3rd Qu.:140.0   3rd Qu.: 90.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :65.00   Max.   :2.000   Max.   :250.0   Max.   :200.00   Max.   :240.0   Max.   :190.00   Max.   :3.000   Max.   :3.000   Max.   :1.000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

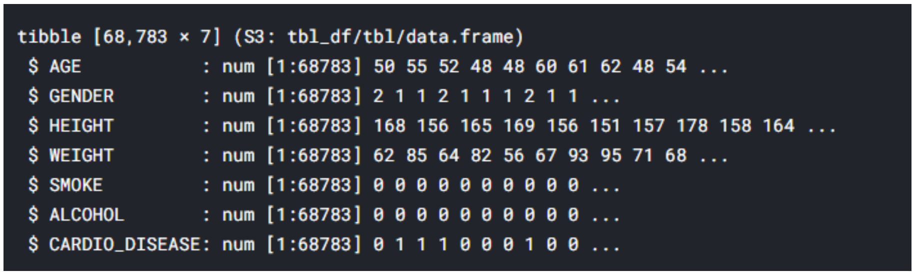
Figure 1 Summary of the dataset

➢ *Missing Data*

After that, the dataset was checked for missing data and there was it was found that there is no missing data since the dataset is clean version of its parent dataset.

- *Removing Columns*

According to the question will be asked to the dataset we will mostly need only the “Age”, “Gender”, “Smoke”, “Alcohol”, “Height”, “Weight” and "Target" columns. Therefore, we can further reduce the dataset to include only these columns. The output for this process is given in Figure 2.



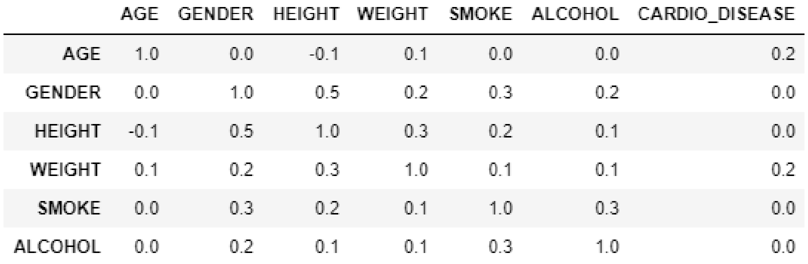
```
tibble [68,783 × 7] (S3: tbl_df/tbl/data.frame)
 $ AGE          : num [1:68783] 50 55 52 48 48 60 61 62 48 54 ...
 $ GENDER       : num [1:68783] 2 1 1 2 1 1 1 2 1 1 ...
 $ HEIGHT       : num [1:68783] 168 156 165 169 156 151 157 178 158 164 ...
 $ WEIGHT       : num [1:68783] 62 85 64 82 56 67 93 95 71 68 ...
 $ SMOKE        : num [1:68783] 0 0 0 0 0 0 0 0 0 0 ...
 $ ALCOHOL      : num [1:68783] 0 0 0 0 0 0 0 0 0 0 ...
 $ CARDIO_DISEASE: num [1:68783] 0 1 1 1 0 0 0 1 0 0 ...
```

Figure 2 Dataset with columns extracted

- *Correlation*

A correlation matrix will be plotted to determine how attributes' independence from each other. The results are given in Table 2 and Figure 3.

Table 2 A matrix: 6 × 7 of type dbl



	AGE	GENDER	HEIGHT	WEIGHT	SMOKE	ALCOHOL	CARDIO_DISEASE
AGE	1.0	0.0	-0.1	0.1	0.0	0.0	0.2
GENDER	0.0	1.0	0.5	0.2	0.3	0.2	0.0
HEIGHT	-0.1	0.5	1.0	0.3	0.2	0.1	0.0
WEIGHT	0.1	0.2	0.3	1.0	0.1	0.1	0.2
SMOKE	0.0	0.3	0.2	0.1	1.0	0.3	0.0
ALCOHOL	0.0	0.2	0.1	0.1	0.3	1.0	0.0

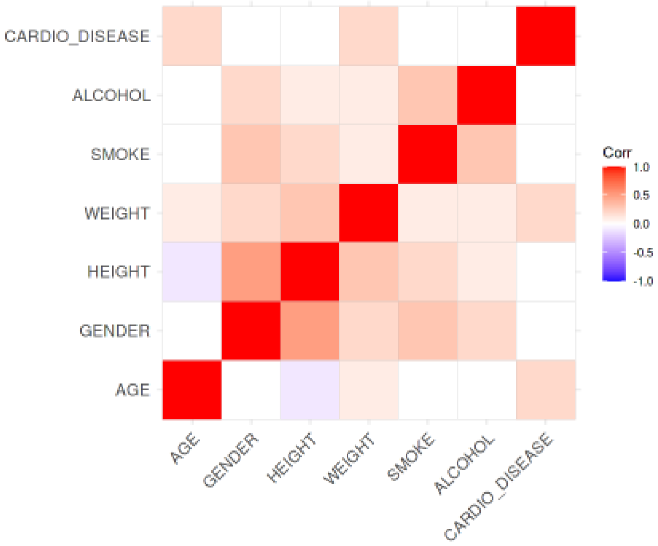
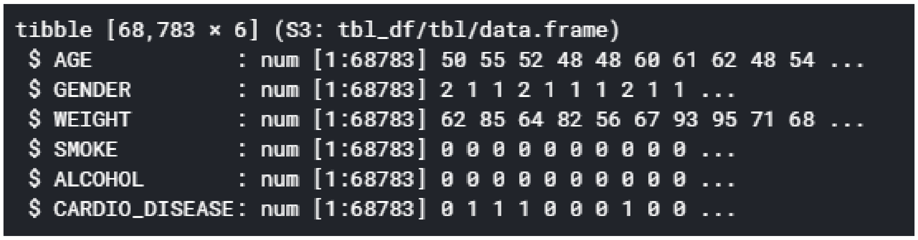


Figure 3 Correlation of attributes

From the above plot age and height have a negative correlation while the rest of the attributes are either not correlated or have some positive correlation. We will remove the height column to avoid any introduction of unwanted artifacts to our analysis. The final form is given in Figure 4.

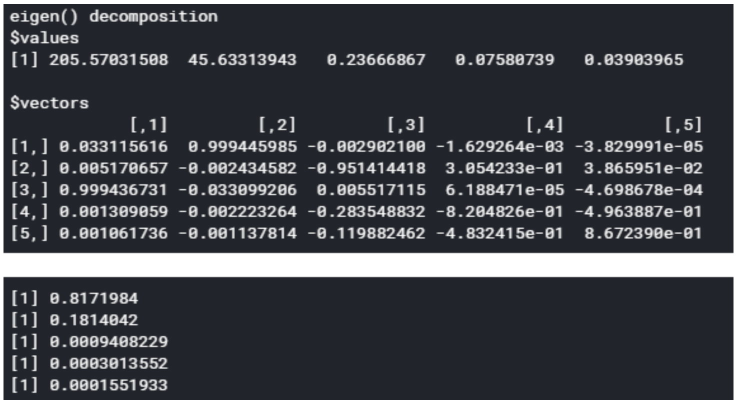


```
tibble [68,783 × 6] (S3: tbl_df/tbl/data.frame)
 $ AGE           : num [1:68783] 50 55 52 48 48 60 61 62 48 54 ...
 $ GENDER        : num [1:68783] 2 1 1 2 1 1 1 2 1 1 ...
 $ WEIGHT        : num [1:68783] 62 85 64 82 56 67 93 95 71 68 ...
 $ SMOKE         : num [1:68783] 0 0 0 0 0 0 0 0 0 0 ...
 $ ALCOHOL       : num [1:68783] 0 0 0 0 0 0 0 0 0 0 ...
 $ CARDIO_DISEASE: num [1:68783] 0 1 1 1 0 0 0 1 0 0 ...
```

Figure 4 Final dataset with columns extracted

2.2 Principle Components Analysis

As a preliminary analysis, Principle Component Analysis (PCA) Method will be applied to understand the dataset and determine the expectations for further analysis. Therefore, linear projection method PCA is applied to have a first glimpse of the structure and influences associated with the attributes of the dataset. The eigenvalues of dataset is in Figure 5 and Figure 6.



```
eigen() decomposition
$values
[1] 205.57031508  45.63313943   0.23666867   0.07580739   0.03903965

$vectors
            [,1]         [,2]         [,3]          [,4]          [,5]
[1,] 0.033115616  0.999445985 -0.002902100 -1.629264e-03 -3.829991e-05
[2,] 0.005170657 -0.002434582 -0.951414418  3.054233e-01  3.865951e-02
[3,] 0.999436731 -0.033099206  0.005517115  6.188471e-05 -4.698678e-04
[4,] 0.001309059 -0.002223264 -0.283548832 -8.204826e-01 -4.963887e-01
[5,] 0.001061736 -0.001137814 -0.119882462 -4.832415e-01  8.672390e-01
```

```
[1] 0.8171984
[1] 0.1814042
[1] 0.0009408229
[1] 0.0003013552
[1] 0.0001551933
```

Figure 5 Eigenvalues

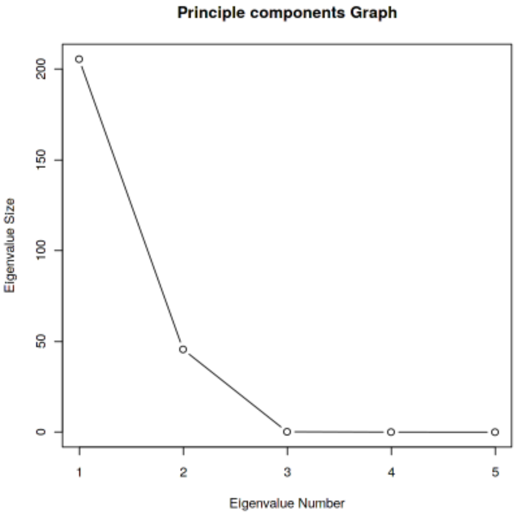
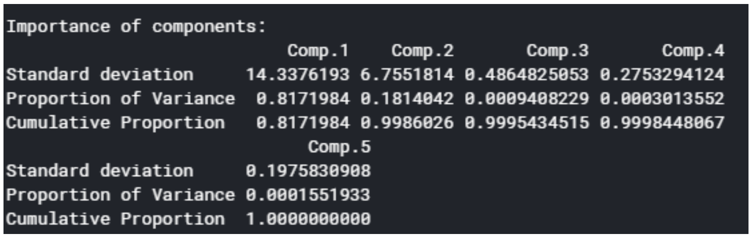


Figure 6 PCA graph for eigenvalues

After finding eigenvalues of the dataset, visualization of PCA will be done. The summary of PCA and the plots are given in Figure 7, Figure 8 and Figure 9.



```
Importance of components:
                          Comp.1    Comp.2       Comp.3       Comp.4
Standard deviation     14.3376193 6.7551814 0.4864825053 0.2753294124
Proportion of Variance  0.8171984 0.1814042 0.0009408229 0.0003013552
Cumulative Proportion   0.8171984 0.9986026 0.9995434515 0.9998448067
                             Comp.5
Standard deviation     0.1975830908
Proportion of Variance 0.0001551933
Cumulative Proportion  1.0000000000
```

Figure 7 Importance of components

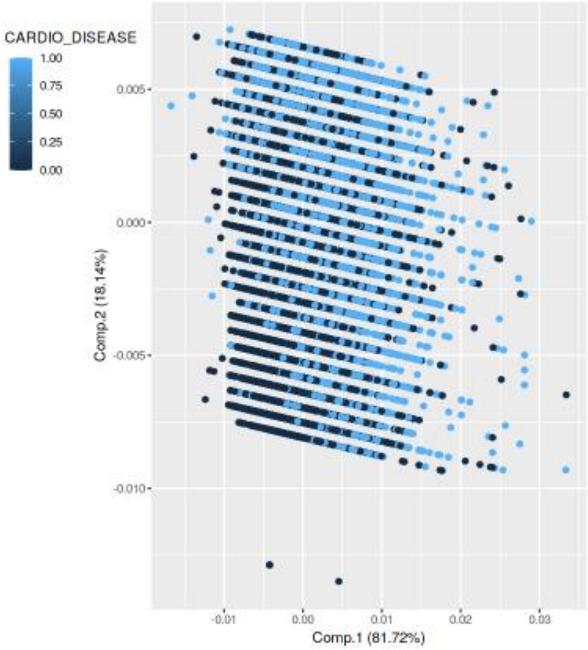


Figure 8 Graph for PC1 and PC2

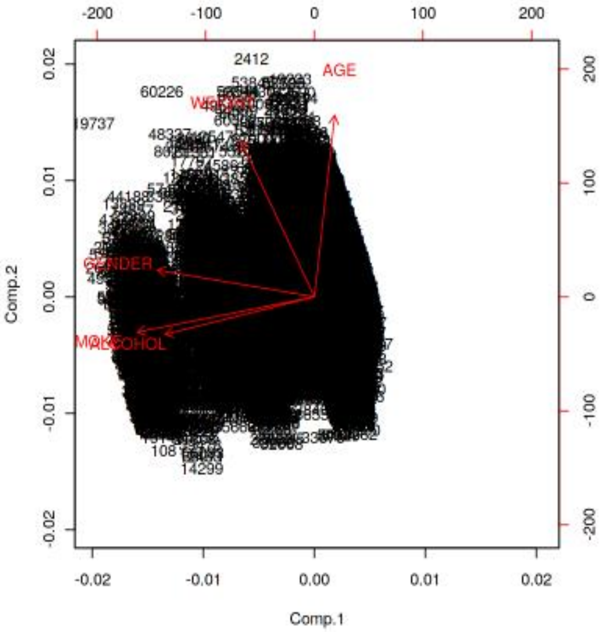


Figure 9 Directions of all PCs on the PC1 and PC2

The first two principle components separate the data well enough showing that it will be good for clustering analysis and the PCA with correlation gives us an idea of the attributes and their direction of influence i.e Alcohol and Smoking both affect the data in the same direction and manner.

3 DATA PROJECTION

The preview of the dataset has shown that our data is clean and well-structured for further analysis including projections and clustering. We will continue with linear projections using Multi-dimensional Analysis and non-linear projection using Isomaps and Local Linear Embedding. The goal is to try and observe any kind of clustering that might present itself in these projections. The projections will provide a guide on how to continue towards the last step of data clustering and analysis.

3.1 Linear Projection Methods

The purpose of this section of the report is applying the linear projection methods to the dataset. As a preliminary analysis, Principle Component Analysis (PCA) Method was applied and the second selected method for linear projection will be Classical Multi-Dimensional Scaling.

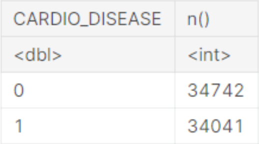
3.1.1 Classical Multi-Dimensional Scaling (MDS)

Initially, using a distance matrix with euclidean distances will be tried even though we have some binary values in our rows. The projection maybe of some significance but the differences in distance will not be representative of all the data under analysis. The dataset that we have is very large and very tasking for the computer to manage easily especially when making matrices. Thus, we will take a sample of our data randomly instead of using the entire dataset. We will try to ensure the ratio is identical to that of the original dataset.

- *Data Sampling*

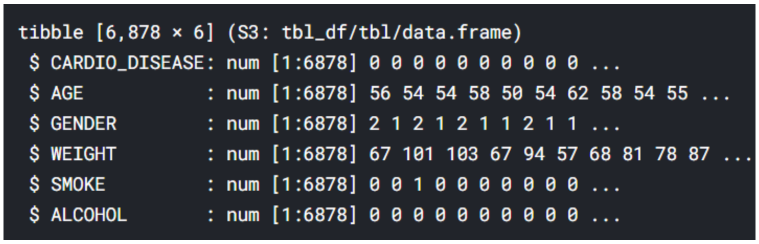
The ratio of those who have a disease and those who don't from the original dataset will be determined.

Table 3 The ratio graph



CARDIO_DISEASE	n()
<dbl>	<int>
0	34742
1	34041

From the above result we will take a sample of a tenth of the original dataset for those that have the cardiovascular disease and those that do not. We will use this sample for the rest of the analysis of this report and it is given in Figure 10.



```
tibble [6,878 × 6] (S3: tbl_df/tbl/data.frame)
 $ CARDIO_DISEASE: num [1:6878] 0 0 0 0 0 0 0 0 0 0 ...
 $ AGE           : num [1:6878] 56 54 54 58 50 54 62 58 54 55 ...
 $ GENDER        : num [1:6878] 2 1 2 1 2 1 1 2 1 1 ...
 $ WEIGHT        : num [1:6878] 67 101 103 67 94 57 68 81 78 87 ...
 $ SMOKE         : num [1:6878] 0 0 1 0 0 0 0 0 0 0 ...
 $ ALCOHOL       : num [1:6878] 0 0 0 0 0 0 0 0 0 0 ...
```

Figure 10 Final sample dataset

- *Visualization*

After sample data was created, data will be visualized.

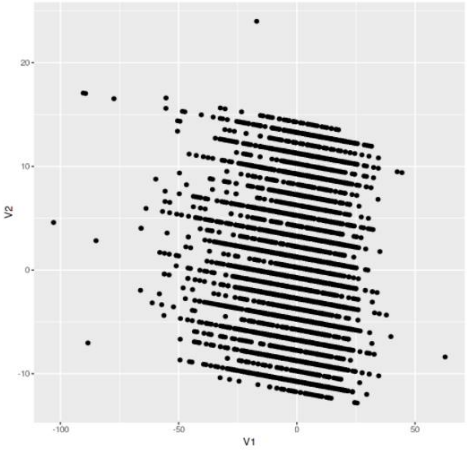


Figure 11 Classical MDS with distance matrix

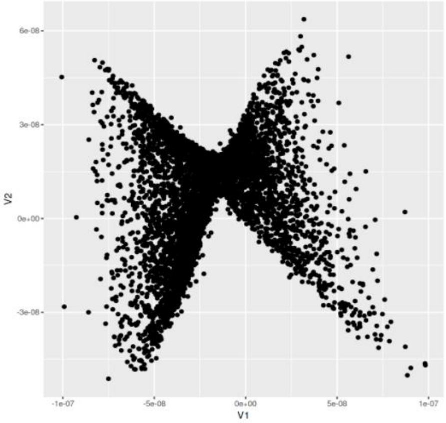


Figure 12 Classical MDS correlation matrix

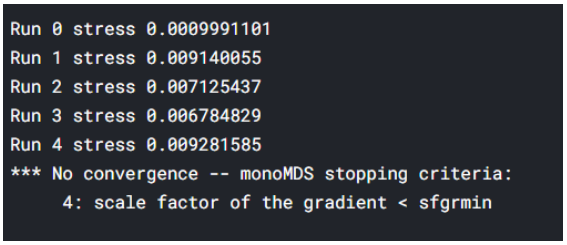
From the above results the distance matrix doesn't perform very well on The multi-dimensional analysis as expected, because of the mixture of types of variables in the sample dataset i.e continuous and nominal. On the other hand the correlation matrix of our data when projected on a 2 dimensional multi-dimensional analysis shows a separation of some kind of two joined clusters of our data. This indicates a possibility of clustering our data down the line.

3.2 Non-Linear Projection Methods

The purpose of this section of the report is applying the non-linear projection methods to the dataset. Non Metric Multi-Dimensional Scaling and Isomaps will be applied to the dataset.

3.2.1 Non Metric Multi-Dimensional Scaling

NMDS uses metaMDS function from the vegan library to call monoMDS to get a stable solution by starting from a different number of tries dictated by the "try" and "trymax" options. The scaling in the result is standardized and easier to interpret. Depending on the number of tries the calculation may or may not reach convergence but we can still have an interpretation of the data.



```
Run 0 stress 0.0009991101
Run 1 stress 0.009140055
Run 2 stress 0.007125437
Run 3 stress 0.006784829
Run 4 stress 0.009281585
*** No convergence -- monoMDS stopping criteria:
     4: scale factor of the gradient < sfgrmin
```

Figure 13 The output values

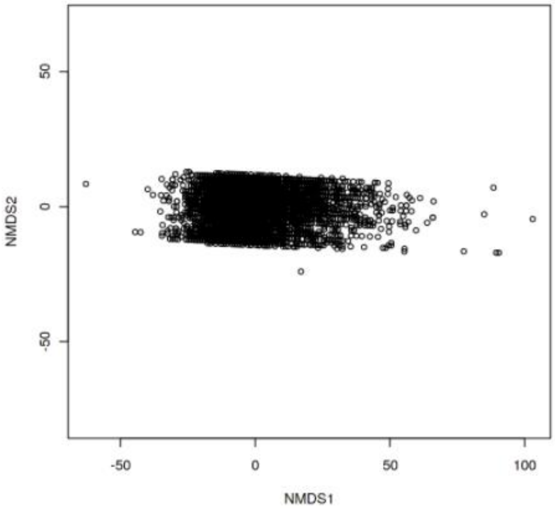


Figure 14 The plot for nMDS

3.2.2 Isomaps

Isomapping is a non-linear projection technique that uses a dissimilarity matrix based on geodesic distances. The isomap function will give us a different view of our data as we might have clearer dissimilarities that are non linear compared to linear similarities in the sample data being analyzed.

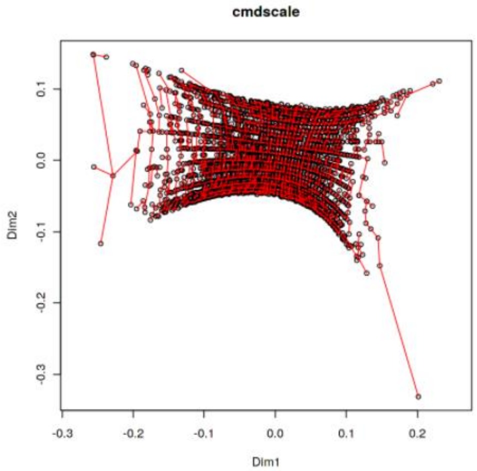


Figure 15 The plot for isomap

We were not able to plot isomaps using the isomap function because the calculations ran for so long regardless of the chosen epsilon. The above isomap uses a spantree function to visualize the map and it shows that our data in 2 dimensions looks like one cluster with very few outliers compared to the two clusters from MDS analysis observed before.

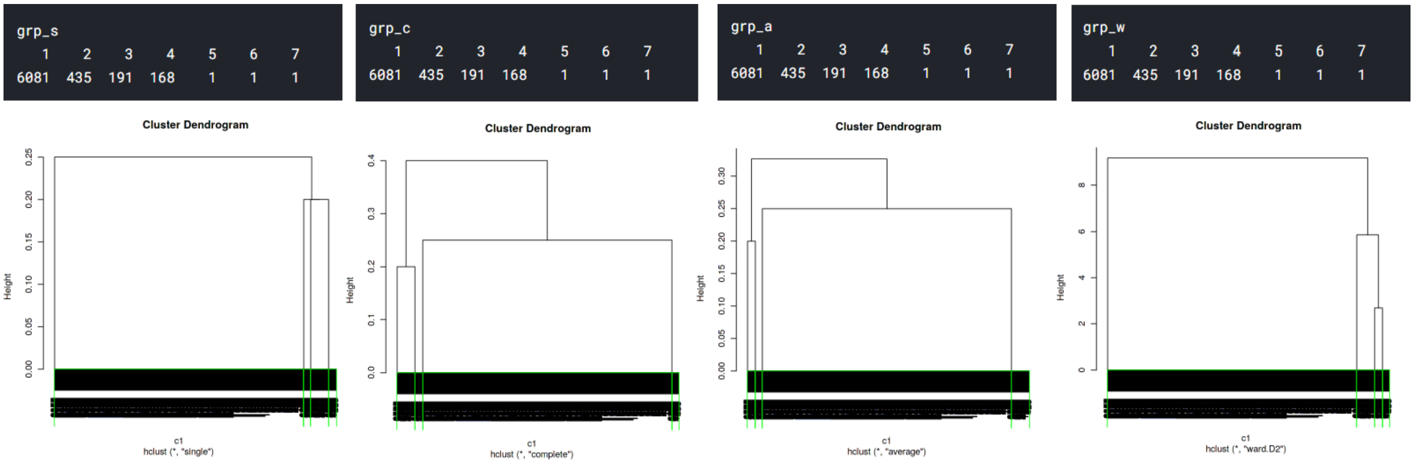
From the above projections our data has a linear correlation that can be clustered into two major clusters. These clusters could be in accordance with our hypothesis where by age is the differentiating factor on who has the highest risk to get a cardiovascular disease.

4 CLUSTERING

In this part of the report, clustering analysis will be applied on the dataset.

4.1 Hierarchical Clustering

The best way to cluster is by distance and not correlation but from previous MDS plots there is only separation when we plot with the similarity matrix by correlation and not distance. Euclidean distances on Nominal data is not the best practice thus, binary method will be used. Single, complete, average and ward hierarchical clustering is applied and their dendograms are given in Figure 16.



```
grp_s
   1    2    3    4    5    6    7
6081  435  191  168    1    1    1
```

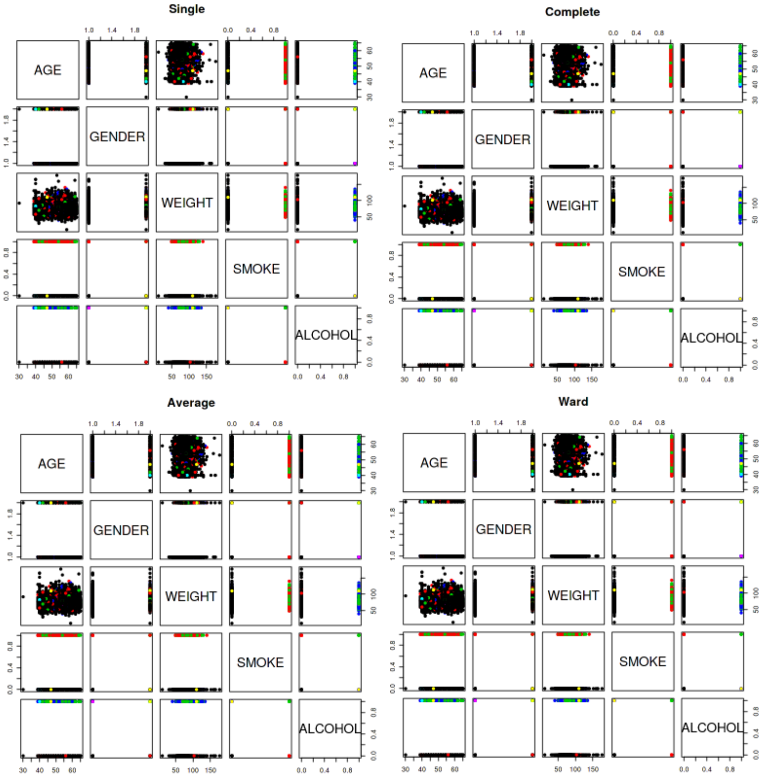
```
grp_c
   1    2    3    4    5    6    7
6081  435  191  168    1    1    1
```

```
grp_a
   1    2    3    4    5    6    7
6081  435  191  168    1    1    1
```

```
grp_w
   1    2    3    4    5    6    7
6081  435  191  168    1    1    1
```

Figure 16 Dendograms

- *Visualization*


Single

Complete

Average

Ward

Figure 17 Visualizationof hierarchical clustering

Hierarchical clustering has showed that our data has mainly 4 major clusters regardless of the method being used to cluster. The visualization of all the attributes involved in clustering shows that age and weight are the most differentiating factors in our clustering.

4.2 K-Means Clustering

The partition clustering method of k-means can be better at clustering complex data compared hierarchical clustering. It will be employed in this clustering analysis to project the previously predicted 4 main clusters from hierarchical clustering. This will be done by first predicting the probable number of clusters.

- *Number of clusters*

The graph for optimal number of clusters is given in Figure 18.

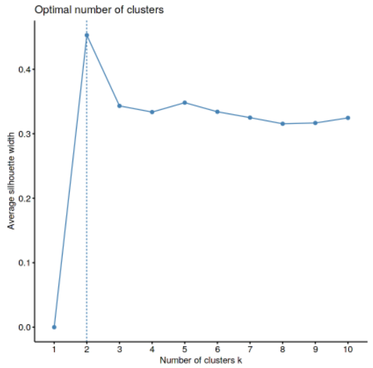

Optimal number of clusters

Figure 18 Optimal number of clusters

- Clusters

The optimal number of clusters is predicted to be 2 clusters from Figure 18 but we will work on the possibility of 4 clusters and visualize our data accordingly with k-means clustering.

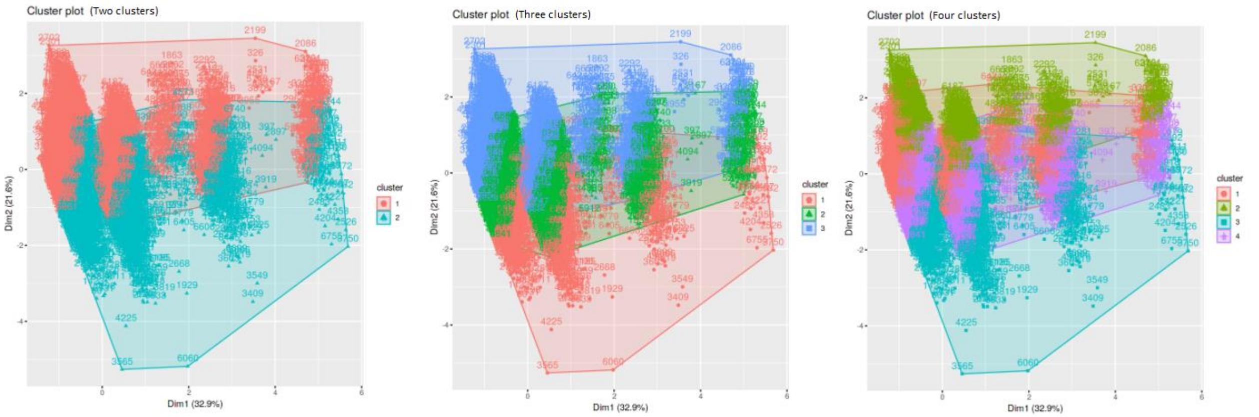


Figure 19 K-Means cluster plots

From k-means the clusters formed do not have a clear separation regardless of th number of clusters chosen or the number of initial configurations. The third and fourth clusters are formed within the first two major clusters, showing that we might to main clusters in our data rather than four. Since not all our attributes are continuous and some are nominal, we will apply kmedoids to try and achieve better clustering.

4.3 Partitioning Around Medoids(PAM)

For this method, two, three and four clusters is created and they are given in Figure 20.

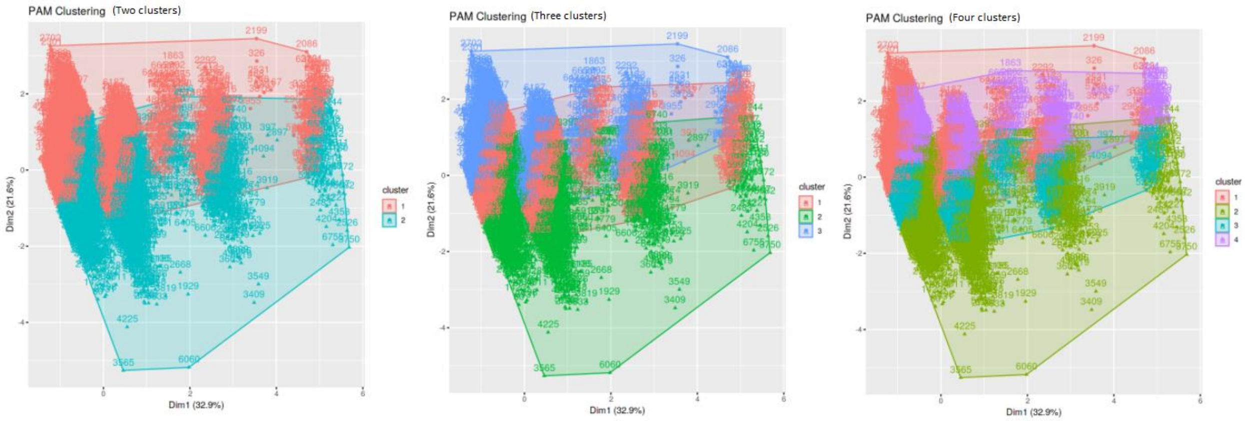


Figure 20 PAM cluster plots

All clustering methods have indicated that our data might have between 2 and 4 clusters but the most probable number of clusters seem to be two but this will only be certain after we have performed a clustering validation analysis. The two main clusters are potentialy between those who have the cardiovascular disease and those who do not have it.

5 CLUSTERING VALIDATION ANALYSIS

5.1 Internal Validation

5.1.1 Hierarchical Clustering Validation

Complete & Ward

Firstly, cluster plot is created for both complete and ward hierarchical clusters and it is given in Figure 21.

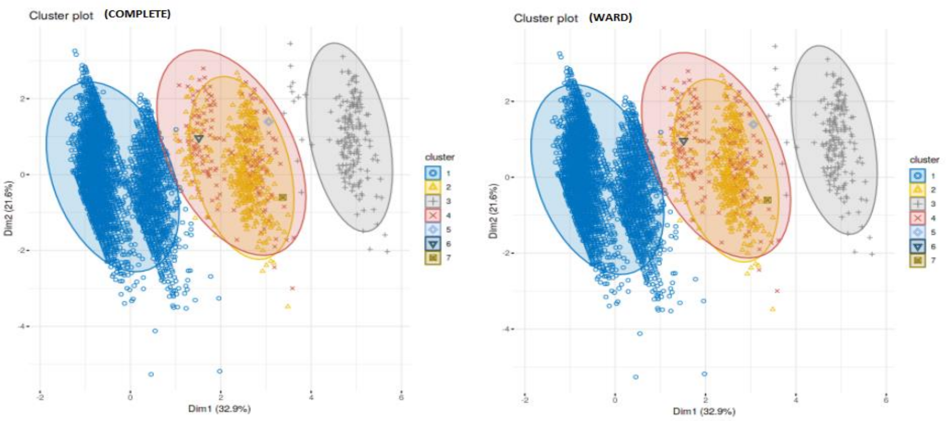
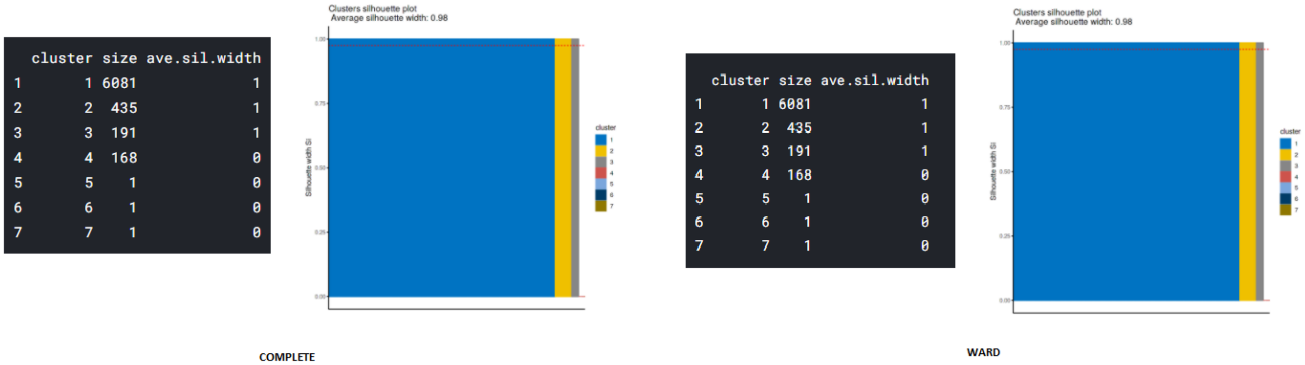


Figure 21 Cluster plots

The visualization shows three major clusters from hierarchical clustering the remaining 4 clusters are within the three major clusters and do not count. Now we will perform the silhouette and dunn indices.

Silhouette Index



```
  cluster size ave.sil.width
1       1 6081             1
2       2  435             1
3       3  191             1
4       4  168             0
5       5    1             0
6       6    1             0
7       7    1             0
```

```
  cluster size ave.sil.width
1       1 6081             1
2       2  435             1
3       3  191             1
4       4  168             0
5       5    1             0
6       6    1             0
7       7    1             0
```

Figure 22 Silhouette plot

According to the silhouette index of 0.98 the clusters are fairly well separated with hierarchical clustering using the complete method.

Dunn Index

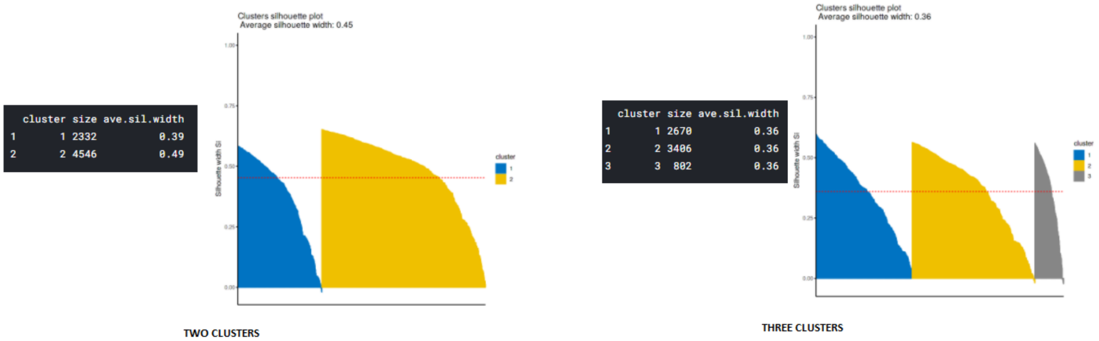
The Dunn Index is found as “0.00602016520523227” for complete one and “0.00602016520523227” for ward one. The dunn index is very low but this is due to the presence of some clusters within other clusters.

From hierarchical clustering internal validation it is shown that they are three clear clusters well separated and among the three the first cluster is dominant with the most population compared to the remaining two clusters.

5.1.2 K-Means Clustering Validation

The internal validation for k-means clustering is done for two and three clusters.

Silhouette Index



```
  cluster size ave.sil.width
1       1 2332          0.39
2       2 4546          0.49
```

TWO CLUSTERS

```
  cluster size ave.sil.width
1       1 2670          0.36
2       2 3406          0.36
3       3  802          0.36
```

THREE CLUSTERS

Figure 23 Silhouette plot

Dunn Index

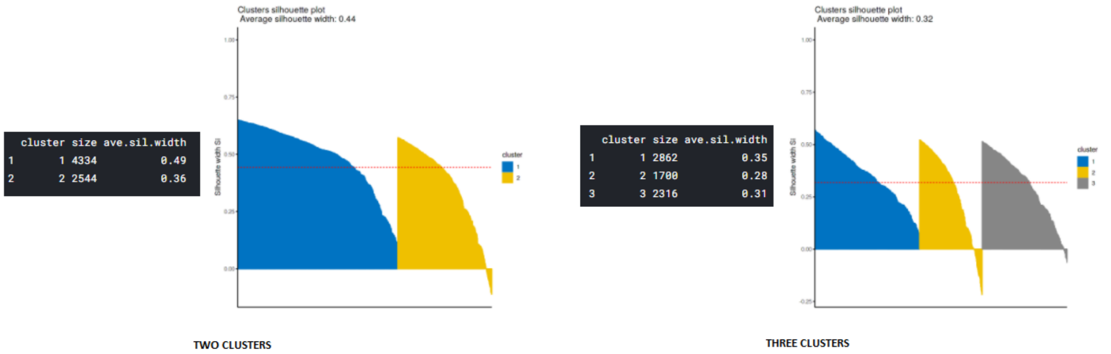
The Dunn Index is found as “0.0100382179809853” for two clusters and “0.0100382179809853” for three clusters.

The k-means clustering has done much better than the hierarchical clustering on the dunn index but it is still low and the average silhouette index from hierarchical clustering was much better.

5.1.3 PAM Clustering Validation

The internal validation for PAM clustering is done for two and three clusters.

Silhouette Index



```
  cluster size ave.sil.width
1       1 4334          0.49
2       2 2544          0.36
```

TWO CLUSTERS

```
  cluster size ave.sil.width
1       1 2862          0.35
2       2 1700          0.28
3       3 2316          0.31
```

THREE CLUSTERS

Figure 24 Silhouette plot

There is a few negative silhouette widths on the second and third clusters potentially indicating misplaced memebers of the respective clusters. Thus the clustering was not very good and could use some improvement.

Dunn Index

The Dunn Index is found as “0.00973630911761081” for two clusters and” 0.0104427608461279” for three clusters.

From the internal methods of validation, the best clusters were acquired using hierarchical clustering while kmeans was good enough and PAM had the worst performance. This shows that it isnt necessary to have a more complex method of clustering depending on how the data is structured.

5.2 External Validation

In external validation we will need synthetic true labels for our clusters and since we do not have any real labels we will use the supposedly best clustering from hierarchical clustering using the ward method.

5.2.1 Hierarchical Clustering Validation

For external validation, adjusted rand and Meila's VI indices are found as 1 and -1.11022302462516e-16, respectively for complete hierarchical clustering.

5.2.2 K-Means Clustering Validation

For two clusters, adjusted rand and Meila's VI indices are found as 0.0138223944741246 and 1.13329532105085, respectively. And, for three clusters, adjusted rand is found as 0.0125478099852573 and Meila's VI indices is found as 1.43852765469504.

5.2.3 PAM Clustering Validation

For two clusters, adjusted rand and Meila's VI indices are found as 0.0138223944741246 and 1.11442546132157, respectively. And, for three clusters, adjusted rand is found as 0.00189198692372031 and Meila's VI indices is found as 1.5479576580552.

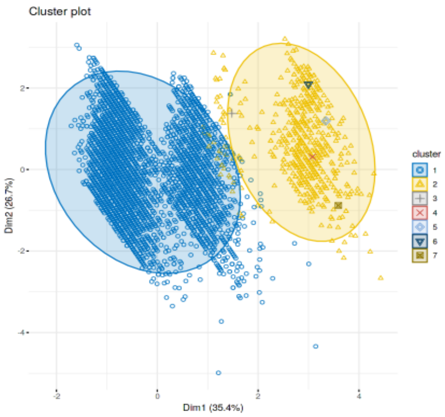
With the adjusted rand index the closest non-hierarchical clustering to the external clustering is the kmeans clustering with two clusters. The furthest is the PAM with three clusters clustering. For our dataset apart from hierarchical clustering the next best choice is kmeans and not PAM.

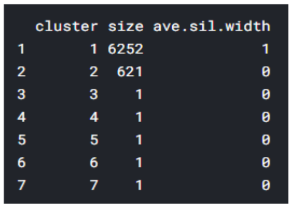
6 CONCLUSION

In this report, the relationship between cardiovascular disease and the external effects such as smoking or alcohol was examined. The preview of the data showed that the data could be representated by the first two principal components and could be clustered. Decreasing the dimensionality of the data was not a big challenge most of the linear and nonlinear methods applied accomplished the task quite well. The clustering methods also performed adequetly but the best clustering was from hierarchical clustering with three clusters. The first cluster is much bigger than the following two clusters. We predict that this cluster is the cluster of individuals with the cardiovascular disease combined with those who dont have it but consume alcohol or smoke. These two attributes also contribute in the formation of the remaining two clusters which contain the individuals in the data that do not have the disease but also do not drink alcohol or smoke.

The consequent elimination of these attributes will result in the formation of a single cluster failing to clearly distinguish between those who have the disease or are at high risk of getting the disease with those that are healthy and are not atany risk of developing cardiovascular diseases.The inclusion of more attributes could help differentiate the clusters more clearly in the advanced methods of clustering like k-means and PAM.We conclude that the data in this report was succesfully clustered and the hypothesis posed at the beginning of the report is sound. That is, regardless of the gender the factors that contribute to the increase in likelihood of developing cardiovascular diseasesis the consumption of alcohol and smoking.

- *Removing alcohol attribute*




```
  cluster size ave.sil.width
1       1 6252             1
2       2  621             0
3       3    1             0
4       4    1             0
5       5    1             0
6       6    1             0
7       7    1             0
```

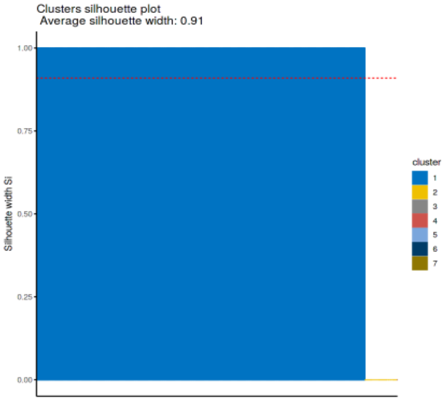
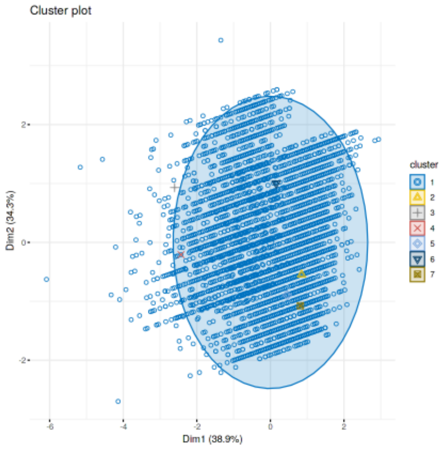
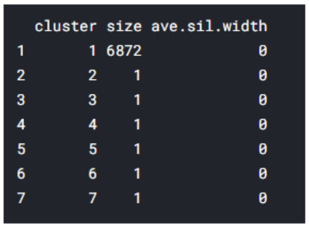


Figure 25 Removing alcohol attribute

- Removing both alcohol and smoking attributes




```
  cluster size ave.sil.width
1       1 6872             0
2       2    1             0
3       3    1             0
4       4    1             0
5       5    1             0
6       6    1             0
7       7    1             0
```

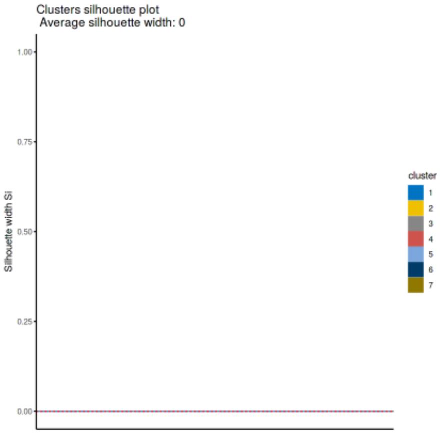


Figure 26 Removing both alcohol and smoking attributes

7 REFERENCES

[1] Kaggle: Cardiovascular diseases dataset (clean). (2020). Retrieved 10 January 2021, from https://www.kaggle.com/aiaiaidavid/cardio-data-dv13032020

[2] Kaggle: Cardiovascular disease dataset. (2020). Retrieved 10 January 2021, from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

APPENDIX A

Codes for Preliminary Analysis

Structure of Dataset

```
library(tidyverse)
card <- read.csv("../input/cardio-data-dv13032020/cardiovascular_diseases_dv3.csv")
library(rio)
export(card,"card.csv")

library(rio)
convert("card.csv", "card.xlsx")

library(readxl)
card <- read_excel("card.xlsx")
names(card) <-
c("AGE","GENDER","HEIGHT","WEIGHT","AP_HIGH","AP_LOW","CHOLESTEROL","GLUCOSE","SMOKE","ALCOHOL","PHYSICAL_ACTIVITY","CARDIO_DISEASE")
summary(card)
#list.files(path = "../input/cardio-data-dv13032020")

sum(is.na(card))

card_new =
subset(card,select=c(1,2,3,4,9,10,12))
str(card_new)

library(ggcorrplot)
card_corr <- round(cor(card_new), 1)
head(card_corr)
ggcorrplot(card_corr)

card_final =
subset(card,select=c(1,2,4,9,10,12))
str(card_final)
```

PCA Method

```
card_reorder <- card_final[,c(6,1,2,3,4,5)]
card_S <- cov(card_reorder[2:6])
sum(diag(card_S)) # Total variance
card_S.eigen <- eigen(card_S) #Check the eigen values
card_S.eigen
for (card_S in card_S.eigen$values) {
  print(card_S / sum(card_S.eigen$values))
}

plot(card_S.eigen$values, xlab = 'Eigenvalue Number', ylab = 'Eigenvalue Size', type = 'b', main = 'Principle components Graph')

card.pca <- princomp(card_reorder[,2:6])
card.pca1 <- princomp(card_reorder[,2:6], cor=TRUE)
summary(card.pca)

library(ggfortify)
pca.plot <- autoplot(card.pca, data = card_reorder, colour = 'CARDIO_DISEASE')
pca.plot
biplot(card.pca1)
```

Codes for Data Projection

Classical MDS

```
#ratio of those who have a disease and those who dont from the
#original data set

library(tidyverse)
card_reorder %>%
 group_by(CARDIO_DISEASE) %>%
 summarize(n())

# Make two datasets, with and without the disease.
card_reorder_0 <- card_reorder %>%
  filter(CARDIO_DISEASE == 0)
card_reorder_1 <- card_reorder %>%
  filter(CARDIO_DISEASE == 1)

# Randomize the selection of the sample data
set.seed(312)
index_0 <- sample(1:nrow(card_reorder_0),3474 )
card_0 <- card_reorder_0[index_0, ]

index_1 <- sample(1:nrow(card_reorder_1),3404 )
card_1 <- card_reorder_1[index_1, ]

# Merge the two sample datasets to form a final sample dataset
card_sample <- rbind(card_0,card_1)
str(card_sample)

c <- dist(card_sample[2:6])
fit <- cmdscale(c,eig=TRUE, k=2)

library(apcluster)
c1 <- corSimMat(card_sample[2:6], sel=NA, r=1, signed=TRUE, method="pearson")
fit1 <- cmdscale(c1,eig=TRUE, k=2)

library(ggfortify)
autoplot(fit, title="Classical MDS with distance matrix")
autoplot(fit1, title="Classical MDS correlation matrix")
```

Non-Metric MDS

```
library(vegan)
c_mt <- metaMDS(comm=c, try=4, trymax=4)
plot(c_mt,type="p")
```

Isomap

```
library(vegan)
d <- vegdist(card_sample[2:6])
t <- spantree(d)
plot1 <- ordiplot(cmdscale(d), main="cmdscale")
lines(t, plot1, col="red")
```

Codes for Clustering

Hierarchical Clustering

```
c1 <- dist(card_sample[2:6],method="binary")
h_s <- hclust(c1,method="single")
plot(h_s)
#abline(h=110,col="green")
grp_s <- cutree(h_s,k=7)
table(grp_s)
rect.hclust(h_s,k=5,border="green")
h_c <- hclust(c1,method="complete")
plot(h_c)
grp_c <- cutree(h_c,k=7)
table(grp_c)
rect.hclust(h_c,k=5,border="green")
h_a <- hclust(c1,method="average")
plot(h_a)
grp_a <- cutree(h_a,k=7)
table(grp_a)
rect.hclust(h_a,k=5,border="green")
h_w <- hclust(c1,method="ward.D2")
plot(h_w)
grp_w <- cutree(h_w,k=7)
table(grp_w)
rect.hclust(h_w,k=5,border="green")
plot(card_sample[2:6],
pch=20,             col=grp_s,
main="Single")
plot(card_sample[2:6],
pch=20,             col=grp_c,
main="Complete")
plot(card_sample[2:6],
pch=20,             col=grp_a,
main="Average")
plot(card_sample[2:6],
pch=20, col=grp_w,
main="Ward")
```

K-Means Clustering

```
library(cluster)
library(factoextra)
fviz_nbclust(card_sample[2:6], kmeans, method =
"silhouette")
k_2 <- kmeans(card_sample[2:6], centers = 2,
nstart = 25)
fviz_cluster(k_2, data = card_sample[2:6])
k_3 <- kmeans(card_sample[2:6], centers = 3, nstart =
50)
fviz_cluster(k_3, data = card_sample[2:6])
k_4 <- kmeans(card_sample[2:6], centers = 4,
nstart = 100)
fviz_cluster(k_4, data = card_sample[2:6])
```

PAM Clustering

```
p_2 <- eclust(card_sample[2:6],FUNcluster="pam",
k=2,hc_metric = "binary")
p_3 <- eclust(card_sample[2:6],FUNcluster="pam",
k=3,hc_metric = "binary")
p_4 <- eclust(card_sample[2:6],FUNcluster="pam",
k=4,hc_metric = "binary")
```

Codes for Validation

Internal Validation – Complete Hierarchical Clust. Val.

```
set.seed(321)
h_c1 <- eclust(card_sample[2:6], "hclust", k = 7,
hc_metric = "binary",hc_method = "complete", graph =
FALSE)
fviz_cluster(h_c1, geom = "point", ellipse.type =
"norm",palette = "jco", ggtheme = theme_minimal())
fviz_silhouette(h_c1, palette = "jco",ggtheme =
theme_classic())
silinfo <- h_c1$silinfo
library(fpc)
library(NbClust)
hc_stats <- cluster.stats(c, h_c1$cluster)
hc_stats$dunn
```

Internal Validation – Ward Hierarchical Clust. Val.

```
set.seed(321)
h_c1 <- eclust(card_sample[2:6], "hclust", k = 7,
hc_metric = "binary",hc_method = "ward.D2", graph =
FALSE)
fviz_cluster(h_c1, geom = "point", ellipse.type =
"norm",palette = "jco", ggtheme = theme_minimal())
fviz_silhouette(h_c1, palette = "jco",ggtheme =
theme_classic())
silinfo <- h_c1$silinfo
hc_stats <- cluster.stats(c, h_c1$cluster)
hc_stats$dunn
```

Internal Validation – K-Means Clust. Val.

```
k_2 <- eclust(card_sample[2:6], "kmeans", k = 2,
nstart = 25, graph = FALSE)
fviz_silhouette(k_2, palette = "jco",ggtheme =
theme_classic())
silinfo <- k_2$silinfo
km_stats <- cluster.stats(c, k_2$cluster)
km_stats$dunn


k_3 <- eclust(card_sample[2:6], "kmeans", k = 3,
nstart = 25, graph = FALSE)
```

```
fviz_silhouette(k_3, palette = "jco",ggtheme = 
theme_classic())
silinfo <- k_3$silinfo
km_stats <- cluster.stats(c, k_2$cluster)
km_stats$dunn
```

Internal Validation – PAM Clust. Val.

```
fviz_silhouette(p_2, palette = "jco",ggtheme = 
theme_classic())
silinfo <- p_2$silinfo
pm_stats <- cluster.stats(c, p_2$cluster)
pm_stats$dunn

fviz_silhouette(p_3, palette = "jco",ggtheme = 
theme_classic())
silinfo <- p_3$silinfo
pm_stats <- cluster.stats(c, p_3$cluster)
pm_stats$dunn
```

External Validation – Complete Hierarchical Clust. Val.

```
cluster.stats(c,clustering=h_c1$cluster,alt.clu
stering=h_w1$cluster,silhouette=FALSE,wgap=FALS
E,sepindex = FALSE,sepwithnoise = 
FALSE,compareonly = TRUE)
```

External Validation – K-Means Clust. Val.

```
cluster.stats(c,clustering=k_2$cluster,alt.clus
tering=h_w1$cluster,silhouette=FALSE,wgap=FALSE
,sepindex = FALSE,sepwithnoise = 
FALSE,compareonly = TRUE)

cluster.stats(c,clustering=k_3$cluster,alt.clus
tering=h_w1$cluster,silhouette=FALSE,wgap=FALSE
,sepindex = FALSE,sepwithnoise = 
FALSE,compareonly = TRUE)
```

External Validation – PAM Clust. Val.

```
cluster.stats(c,clustering=p_2$cluster,alt.clus
tering=h_w1$cluster,silhouette=FALSE,wgap=FALSE
,sepindex = FALSE,sepwithnoise = 
FALSE,compareonly = TRUE)

cluster.stats(c,clustering=p_3$cluster,alt.clus
tering=h_w1$cluster,silhouette=FALSE,wgap=FALSE
,sepindex = FALSE,sepwithnoise = 
FALSE,compareonly = TRUE)
```

Codes for Conclusion Part

Removing Alcohol Attribute

```
h_w2 <- eclust(card_sample[2:5], "hclust", k = 
7, hc_metric = "binary",hc_method = "ward.D2", 
graph = FALSE)
fviz_cluster(h_w2, geom = "point", ellipse.type 
=   "norm",palette   =   "jco",   ggtheme   = 
theme_minimal())
fviz_silhouette(h_w2, palette = "jco",ggtheme = 
theme_classic())
silinfo <- h_w2$silinfo
```

Removing Alcohol and Smoking Attributes

```
h_w3  <-  eclust(card_sample[2:4],  "hclust",  k  =  7, 
hc_metric  =  "binary",hc_method  =  "ward.D2",  graph  = 
FALSE)

fviz_cluster(h_w3, geom = "point", ellipse.type = 
"norm",palette = "jco", ggtheme = theme_minimal())
fviz_silhouette(h_w3,   palette   =   "jco",ggtheme   = 
theme_classic())
silinfo <- h_w3$silinfo
```