CEDRIC-JIONGO NGNIMPIABA

DISTANCE MATRICES COMPUTATION

MASTER

2022

NEAR EAST UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

DEPARTMENT OF BIOSTATISTICS

Distance Matrix Computation Applied on HCV Dataset.

M.Sc. THESIS

CEDRIC JIONGO NGNIMPIABA

Nicosia

February 2022

NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF BIOSTATISTICS

Distance Matrix Computation Applied on HCV Dataset

M.Sc. THESIS

Cedric JIONGO NGNIMPIABA

Supervisor
Assist. Prof. Dr. Özgür TOSUN

Nicosia
February 2022

Approval

We certify that we have read the thesis submitted by Cedric Jiongo Ngnimpiaba titled **"Distance Matrix Computation Applied on HCV dataset"** and that in our combined opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Educational Sciences.

Examining Committee	Name-Surname	Signature
Head of the Committee:	Prof. Dr. İlker Etikan	…..…..……
Committee Member:	Assoc. Prof. Dr. Uğur Bilge	
Supervisor:	Assist. Prof. Dr. Özgür Tosun	…..…..……

Approved by the Head of the Department

..../..../2022

Prof. Dr. İlker Etikan

Title, Name-Surname

Head of Department

Approved by the Institute of Graduate Studies

..../…../2022

Prof. Dr. Kemal Hüsnü Can Başer

Head of the Institute

Declaration

I hereby state that the work contained in this manuscript is my own work and that I conducted it ethically, From the stage of planning until writing through analysis. All the information contained inside the thesis and which is not from me have their references cited in the reference section of this work.

Cedric JIONGO NGNIMPIABA

14/01/2022

Acknowledgment

The work we have done in this thesis wouldn't have been possible without some special people:

- ✓ God, the Almighty, for all the blessings in my life.
- ✓ Assist. Prof. Dr. Ozgur Tosun, my academic advisor in the Department of Biostatistics, Thank you for the guidance, patience, and support.
- ✓ Pr. Dr. ILKER ETIKAN, head of the Department of Biostatistics in Near East University, for his assistance and support
- ✓ Assist. Pr. Etienne GNIMPIEBA ZOHIM of the department of Biomedical Engineering in the University of South Dakota; for his unconditional love and support even in times of conflict.
- ✓ My parents, I will never be that much happy to see where we come from and what we have archived so far. The best is yet to come, your joy shall never miss
- ✓ My mates Vidal Ymelong, Serge Tonlio, Francxa Waguie, Alviane Kassep, Victorine Bekotto, and Rafna-Chloe Jiongo, for all the love and the joy you share daily.
- ✓ My family for all the ups and downs we have gone through and survived so far. I love you

Abstract

Distance matrix computation applied Hepatitis C Virus dataset

Jiongo Ngnimpiaba Cedric

Ass. Prof. Dr. Özgür Tosun

Master, Department of Biostatistics

14/01/2022 66 pages

With the advent of new technology, the profusion of data has created a new area for research. The data originating from various domains such as Biology, computer science, etc… has proven to be a serious problem when it comes to exploiting them since they come into Categorical, Continuous, and mixed types. However, exploiting those data requires a selection of good metrics depending on the type of work when want to process.

In this work, we explored various distances measures or metrics such as Euclidean distance, Manhattan distance, Canberra Distance, and many more. Besides, we collected the Hepatitis C virus (HCV) dataset on which we computed distance matrices using five well-known distances. In addition, we built a heatmap for those distances both in the variables and Individuals spaces;

After building the matrices, we compared them using the Mantel Correlation score and found that the Euclidean distance matrix (EDM) and Minkowski Distance matrix (MINDM) are 100% correlated and account for the highest correlation between two different distances measures used to compute distance matrices. Besides, the Canberra distance matrix (CANDM) shares the least similarity score of 59.59% with the maximum distance matrix (MAXDM). In addition, the couple (EDM and MAXDM) & (MAXDM and MINKDM) have the same mantel correlation score of 99.10%.

Key Words: distance measures, distance matrix, multivariate distance matrix regression

Table of Contents

List of Tables

Page

Table of figures

List of Abbreviations

TRNC:	Turkish Republic of North Cyprus
MNE:	Ministry of National Education
HCV:	Hepatitis C Virus Dataset
KNN:	K-Nearest Neighbors
DM:	Distance Matrix, Distance Measures
EDM:	Euclidean distance matrix
MANDM	Manhattan Distance Matrix
MAXDM	Maximum Distance Matrix
CANDM	Canberra Distance Matrix
MINKDM	Minkowski Distance Matrix

1 CHAPTER I

Introduction

In this chapter, the study context, the problem, the aims, the objectives, the concepts definition, and the significance of the study are evaluated.

1.1 Context of the Study

Contemporary research in many domains generates a massive amount of data. In biology, the data being produced from technologies such as genotyping platforms, imaging, gene expressions microarrays are intensive and complex. Thus, fueled researchers and scientists with data that requires improved and advanced techniques to be analyzed. Improvement of digital health technologies has resulted in a huge amount of both qualitative and quantitative data, which contain important information about user interactions and transactions that could benefit caregivers as well as patients [1]. Data have been considerably generated over the years and stored in large databases for commercial and research purposes. Now, analyzing this huge amount of data remains an important challenge. However, data mining provides software and methods which help automate, analyze and explore huge and intricate data sets. Research fields such as biostatistics, computer science, database management, medicine, and machine learning, just to name a few, are furthering works to extract scientific knowledge from the data that could benefit society. Data reductions technics, such as clustering, factors analysis have yielded important good information from datasets. However, hierarchical clustering and phylogenetic analysis require the computation of a distance matrix. But constructing a distance matrix implies the selection of an appropriate distance measure [2]. The distance used in a matrix could or could not be a function that defines a distance between each pair of points elements in a set. Albeit, building the given matrix has never been an easy-to-do task since the nature of the data in a dataset differs from being qualitative, quantitative, or mixed. Therefore, come the issues of which statistical distance should be used depending on the data type.

1.2 Statement of the problem

The flow in a profusion of data generated from every research field especially in biological research requires methods to find similarities between participants. Here come the notions of distance matrices and building such a matrix implies calculating

pairwise distances using statistical distances. However, which distances metric should be used for qualitative, quantitative, and mixed data types?

1.3 Research Questions

- Which distance measures might be better for building a matrix for quantitative data?
- Which distance measures might be better for building a matrix for qualitative data?
- Which distance measures might be better for building a matrix for a mixed data type?

1.4 Aims and Objectives

1.4.1 Aims

The purpose of this work is to investigate distance measures and build a distance matrix accordingly based on the type of data. Therefore, evaluate how statistical distances affect the properties of a distance's matrix given that the type of data used to build the matrix could be qualitative, quantitative, or Mixed (both Categorical and Continuous).

1.4.2 Objectives

At the end of the work, we should be able to:

a) Recommend which distances measures may be suitable for building a matrix of quantitative data.
b) Propose which statistical distances are appropriate for building a matrix of qualitative data.
c) Suggest which distances measure works well for building a matrix for the mixed data type.

1.5 Significance of the Study

The research is suitable to show how to build a distance matrix. Data are ubiquitous and available in different forms. This thus brings the concern of having an appropriate distance matrix for each type of data. which helps to find a relation of (dis)similarity between the individuals in the given data set.

1.6 Definition of Concepts

Statistical distance: it is a numerical measure that presents how distant two points or objects are apart.

Distance matrix: it is a two-dimensional matrix containing pairwise distance computed using a given distance measure.

Multivariate Distance Matrix Regression: it is a data analysis technic that takes advantage of data reduction methods and is engraved in traditional linear models.

Similarity: it shows how closed or distant two objects are.

1.7 Structure of the Work

To further this work, its second part will focus on former related work, the third part of it will be based on materials and methods used to proceed with the task. Besides, the fourth part will emphasize the results while the last part will conclude and recommend.

2 CHAPTER II

Literature Review

2.1 Concept of Distances Matrices

A distance Matrix (DM) is a two-dimensional array containing distances in pairs of a set of samples[3]. The usage of DM can be found in many domains such as Bioinformatics[3], Information Retrieval[4], Images Analysis[5], Data Clustering[6], and Pattern recognition[7]. Xing Hua et al. [8] tested several microbiome distance matrices that helped them develop a statistical method to identify host genetic variants linked to microbiome composition. The computation of a DM requires the selection of distance measures which is calculated either between variables in the columns or between individuals in rows, let us consider the sample data table below for a proper explanation.

Table 1

Example of Individuals by Variables Tables

		Variables	
		V1	V2
Individuals	I1	3	5
	I2	4	1

The distance between the variables V1(3,4) and V2(5,1) in the space of individuals is pictured in *figure 1;* while the distance between individuals I1(3,5) and I2(4,1) in the space of variables is given by *figure 2.*

Figure 1

Variables by Individuals' Space

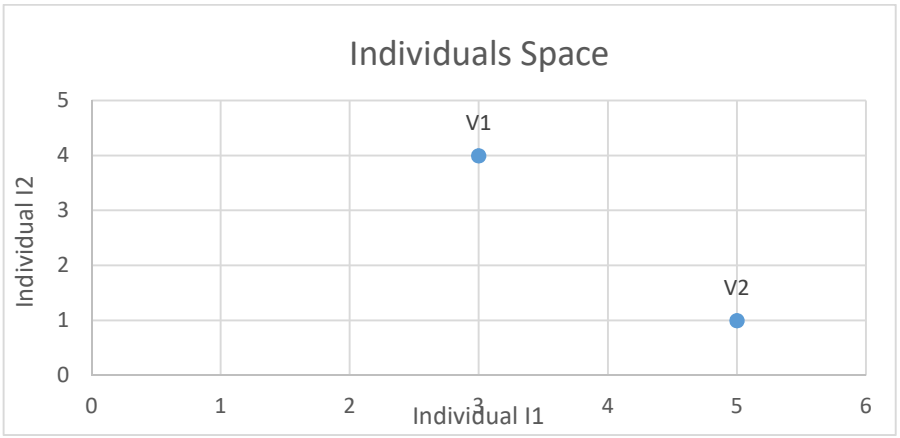
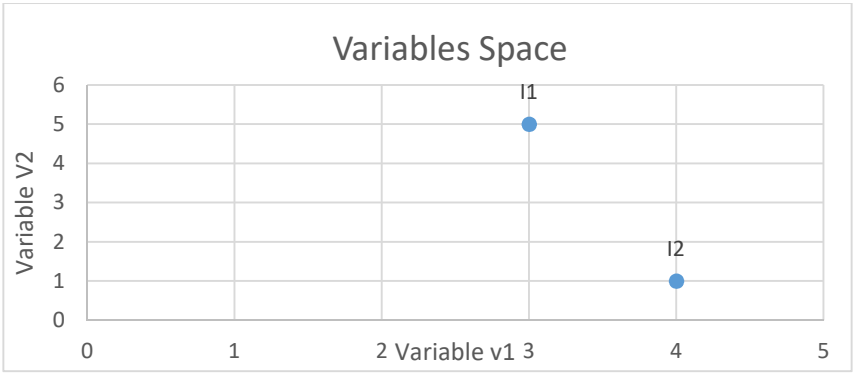


Figure 2

Individuals by Variables Space



The two figures above draw a better picture of the space to consider while calculating the distance measures.

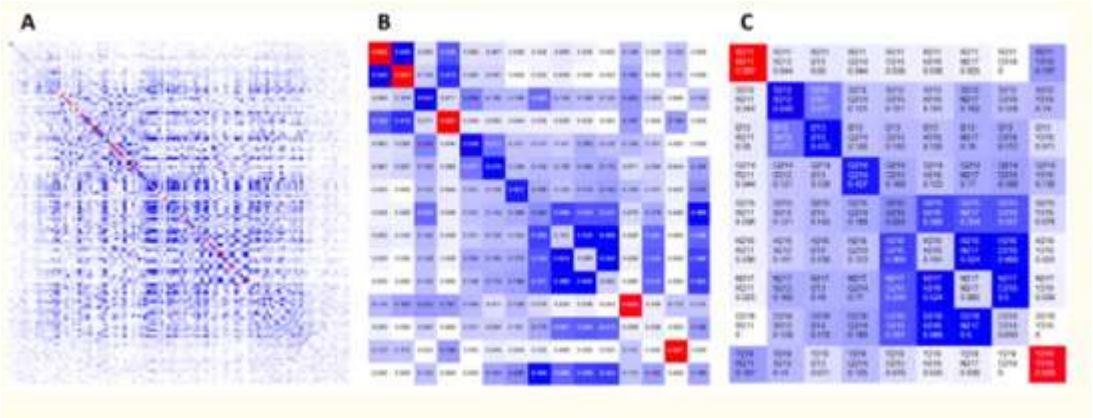
The Properties of a Distance Matrix could be affected by the choice of a statistical distance and therefore its analysis. After choosing a distance and building a matrix, the Multivariate Distance Matrix Regression (MDMR) analysis can be used to evaluate the properties of the given matrix. MDMR is a statistical technique that allows researchers to relate P variables to M factors collected on N individuals, where P >> N[2]. The MDMR evaluates the distances pairwise between individuals based on the selected features.

Shehzap et al.(2014)[9] developed a method entitled Connectome-Wide Association Studies (CWAS), that uses the MDMR analysis to evaluate the correlation between functional connectivity and phenotypes. They found that, compared to univariate methods that require serious improvement for multiple comparisons, the MDMR analysis method considerably reduced the number of comparisons between phenotypes and connectivity.

Baker & Porollo (2018)[10] stated that interpreting huge matrix especially those with thousands of features is a tricky task. They used the CoeViz tool to visualize matrices of protein, human estrogen receptor-alpha (ESR1) containing about six hundred amino acid residues. The heat map they obtained shows the covariance data based on the covariance metrics Chi-Square statistic, Pearson Correlation, and joint Shannon entropy (a conservative measure)[11]. The visual describing the covariance data is given in *figure 3* below where the gradient from the color that states no covariance to red meaning high covariance. The main diagonal contains frequencies of the given amino acids observed at the individual positions in a given multiple sequence alignment.

Figure 2:

Implementation of the Interactive Heatmap Visualization of Covariance Metrics in CoeViz



F. N. Baker and A. Porollo, "CoeViz: A web-based integrative platform for interactive visualization of large similarity and distance matrices," Data, vol. 3, no. 1, Mar. 2018, DOI: 10.3390/data3010004.

2.2 Concept of Distance measures

A distance measure between two instances calculates how far or close they are. The shorter the distance the closer they are and the longer the distance the farther they are. This best defines the concept of similarity and dissimilarity. They are known to play key roles in the fields of Machine Learning, Statistics, and other scientific-related domains[12]. In statistics, they are distances between two statistical objects.

Markatou et al.(2019) [13] explored statistical distances also known as divergences or metrics in the constructions of Model Adequacy – some of the distances investigated were the popular Euclidean Distance, Manhattan, or City-Block distance, Mahalanobis distance, and simple Matching Distance.

Mulak and Talhar (2013) [14] performed a kNN classification using Euclidean, Chebyshev, and Manhattan distance measures on the KDD data set which contains numeric data in 2 classes for 41 features. They estimated the sensitivity, accuracy, and specificity to evaluate the performance of the kNN algorithm for each distance measure. They concluded that the Manhattan distance performed well than the Chebyshev and Euclidian distance with 96.76% sensitivity, 98.35% specificity, and 97.80% accuracy.

Haneen Arafat et al. (2018) [15] investigated in a comprehensive review study the impact of 54 distinct distances measures on 28 different data sets collected from the UCI machine-learning repository. Their work showed that Hassanat distance received the best performance compared to other distances on the majority of data sets.

Todescheni et al. [16] examined the kNN classifier with 18 distinct distances measures, on 8 benchmark data sets. The distances used in their project include Manhattan, Soergel, Lance-Williams, Euclidean, Bhattacharyya, Lagrange, Mahalanobis, Cosine, Correlation, contracted Jaccard-Tanimoto, Clark, and 4 centered Mahalanobis distances. To estimate the efficiency of the measure, they used the average rank of each distance measure and the rate of non-errors. The results showed that contracted Jaccard-Tanimoto, Manhattan, Soergel, Euclidean, and Lance-William's distance measures proved higher accuracy.

Rezvan and Finn (2020) [17] evaluated the performance of the kNN algorithm on 4 different cancer data sets (*Brain, Breast, Lung, and Prostate*) using 12 distance measures (*Fisher, Sobolov, Clark, Bhattacharyya, Soergel, Hassanat, Euclidean, Chebyshev, Hamming, Canberra, and Bray-Curtis*). It resulted that, on the Brain cancer data set, for all the *k-values* tested, the Canberra distance has the best

PRECISION and F1 scores, the Manhattan and the hamming distances have the best *RECALL* score, and the Hassanat distance has the best *ACCURACY* score. On the Breast Cancer data set, Bray-Curtis and Clark both have the best *ACCURACY* score, Clark itself has the best *RECALL and F1 scores,* whereas Bray-Curtis has the best *PRECISION* score. In addition, only Canberra distance has the best score regarding the *PRECISION, the RECALL, the F1, and the ACCURACY* score on the *k*NN performance applied on the Prostate cancer data set. Besides, on the Lung cancer data set*, Fisher and Sobolev* have the best *F1* score, Fisher has the best *RECALL* score, and Sobolev has the best *PRECISION and ACCURACY score. In summary, the Hassanat ranked first; followed by Manhattan distance on the general evaluation of the k*NN classifier on all 4 data sets. That last one was the hamming distance.

Dixon et al. (2009)[18] analyzed metabolomics data without stating an appropriate way to choose a better distances measure, but proved that compare to the Euclidian distance measures, Canberra Distance and a New Precision-Weighted Manhattan Distance are most repeatable.

Hu et al. (2016)[19] on a kNN classification project using medical data sets evaluated 4 different distance measures; Euclidean, cosine, chi-square, and Minkowski. They emphasized 3 different types of data, consisting of numerical, categorical, and mixed data types. The data sets originated from the UCI machine learning repository of data sets, they used cross-validation (30% testing and 70% training) to measure the performances, with k-values between 1 and 15. The experiments portrayed the chi-square distance measure to be best for the 3 different data types, whereas the Euclidean, cosine, and Minkowski distances lead to the lowest accuracy score on the mixed-type data set.

Sung Hyuk Cha (2021)[20] partitioned distances measures according to how well is the correlation between each other and clustered them hierarchically. His work is a great source of understanding distances without reviewing how they are used. Besides, he provided a semantic and syntactic grouping of similarity and distance measures based on the application to probability distribution functions.

2.2.1 Metric Distances

A distance is said to be a metric if and only if it meets the four following criteria:

1- *Positivity: d(p, q)>=0*, for two observations p and q that are distinct.

2- *Symmetry*: *d(p, q)=d(q, p),* for any two distinct observations q and p.
3- *Triangle inequality*: *d(p, q)=<d(p, r) + d(r, q),* for all p,r,q
4- *d(p, q)=0* in case of p=q

Using the previously mentioned characteristics, the following distances are considered metrics.

- **Canberra Distance**

It was first introduced as a software metric by Lance & William in the years 60[21], and it is a weighted version of the Manhattan distance or L1 classical distance. This distance is equally useful in functional genomics when the comparison of the ranked lists is needed. In addition, it is used in Clustering, Classification, harm/spam detection, and computer security. It is computed as the total of absolute values of the differences between ranks divided by their sum and is given by the following formula:

$$Ca(x, y) = \sum_{j=1}^{n} \frac{|x_j - y_j|}{|x_j| + |y_j|}$$

Where x and y are two vectors of reals.

- **Manhattan Distance**

This distance is used to evaluate the absolute difference between the coordinates of pairs of objects[22]. Let n be the number of data, x_{ij} the data located at the center of the cluster to k, k the symbol of each data and d the distance between i Centre of the cluster) and j (the attribute data). The Manhattan distance is given by the formula below:

$$d(i, j) = \sum_{k=1}^{n} |x_{ij} - y_{jk}|$$

- **Mahalanobis Distance**

This distance is suitable for comparing groups of objects depicted by the same variables. Besides, it removes the differences in scale and the effect of correlation between variables. Its computation happens by using a covariance matrix from the

input matrix. This metric between two *d*-dimensional numerical vectors ***x*** *and* ***x'*** *is given by* $d^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')T\mathbf{M}(\mathbf{x} - \mathbf{x}')$, where **M** is a $d \times d$ dimension matrix.

- **χ2 distance**

This distance is approximative to the χ2 metric, except that, the weighted Euclidean distances are timed by the sum of all values in the raw data matrix. This turns the weights in the Euclidean distances to **probabilities** rather than to column totals. The metric is mainly used in the principal correspondence analysis and other related analyses.

- **χ2 metric**

This metric is asymmetric and its computation turns a matrix of quantitative values into a conditional probability matrix in which, based on the row values of the obtained matric, a weighted Euclidean distance is computed. Weights, considered as the reciprocal variables totals from the raw data help to reduce the influence of the highest measured values.

- **Chord Distance**

This given distance is also an asymmetric measure which is the Euclidean distance computed for a row standardized matrix. Instead of comparing absolute values, the metric compares objects based on the proportion of suggested values to the total of all variables values along the row belonging to that given object. The sites will be considered similar as long as the raw values are different in more variable and proportionately equal when standardized.

- **Hellinger distance**

The asymmetric distance is most like the χ2 metric. Although there are **no weights** applied, the **square roots of conditional probabilities** are used as variance and therefore stabilizing data transformations. The variable with small non-zero counts is provided with lower weights and the distance measure performs well in linear ordination.

- **Coefficient of racial Likeness**

The coefficient is suitable to compare groups of objects that are described by the same variable. In addition, it does not remove the correlation effects between variables. It

could be appropriate when the samples are really small to considerably eliminate the effect of the correlation.

- **Euclidian Distance**

In a two-dimensional space, it represents the length between two points. It helps to determine how similar or dissimilar the two points are. The formula below is used to evaluate that degree of similarity.

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{jk})^2}$$

Where n is the number of data, x_{ik} the data located at the center of the cluster to k, k the symbol of each data and d the distance i (center of the cluster) and j (the attribute data).

2.2.2 Semi-metric Distances

This type of metric does not always comply with the triangle inequality and cannot be always used to represent a dissimilarity measure in the Euclidean space unless proper transformation is done. They are mainly used in principal component analysis and non-metric dimensional scaling.

- **Bray-Curtis Distance**

It is an asymmetric measure that is often used for raw data count. In addition, it treats the differences between low and high variables values the same and is also mainly used in ecology.

- **Hamming Distance**

$$\boldsymbol{d}(i,j) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \mathbf{1}_{\boldsymbol{x_i \neq y_i}}$$

In comparing strings of the same length, the hamming distance counts the number of characters from which they differ. It is mainly used with categorical variables and helps to find the similarity or dissimilarity between non-continuous variables. For instance, let us consider two strings coded respectively with 11011001 and 10011101. The hamming distance between them is 2 since they are different in only two digits.

- **Minkowski Distance**

This Distance is a generalization of distance metric and is given by the formula below.

$$d(i,j) = \sqrt[p]{\sum_{k=1}^{n}(x_{ik} - y_{jk})^p}$$

Where, for p=1, it yields the Manhattan distance and for p=2, the distance represents the classical Euclidian distance. For p= +∞ the distance become the maximum distance.

- **Cosine Distance**

The cosine distance is a metric that is widely used in natural language processing (NLP), text mining, and information retrieval systems.

2.3 Clustering

Clustering is an unsupervised learning method used to classify data in such a way that similar data belong to the same group whereas dissimilar data are part of different groups.

Sami Nouali et al. (2019) [23]investigated over 30 categorical clustering algorithms and classified them into partition and hierarchical clustering algorithms. The most prominent ones were selected based on their accuracy, recall, and precision score. Among those developed algorithms, rough, fuzzy, and hard-set based methods were developed and the rough set-based clustering methods yielded efficient results compared to the other two methods.

Melodie Angeletti et al. [24] clustered the MRIF (Multi-Resolution Interest Fusion) data and the distance matrix used to cluster took 80% of the total computation time. They proposed three algorithms for computing the Euclidian Distance Matrix (EDM). The given parallel algorithms improved the computation time of the EDM on parallel computers.

3 Chapter III

Material and Methods

3.1 Data source description

The dataset used in this research originated from the online open-access UCI Machine Learning Repository. It is a web-based application containing datasets from various domains such as biology, sales, engineering, and more. The datasets are used to serve the machine learning community. The authors, David Aha and his fellow graduate students from the University of California Irvine, School of Information and Computer Science created the database in 1987. Since its inception, the dataset has been mainly used by researchers, students, educators, as the primary source for machine learning projects. It currently hosts 588 datasets ready by people in need.

The credentials through which a dataset can be selected from the UCI website, are listed below as:

- **Name**

This is the name of the dataset

- **Data type**

It varies in between Multivariate, Univariate, Sequential, Time-series, Text, Domain-theory and other data types.

- **Attribute Type**

It is amongst Categorical, Numerical, and Mixed attribute type

- **Area**

The domain area includes Life-Sciences, Physical Sciences, Computer Science & Engineering, Social Sciences, Business, Game, Other.

- **#Attibutes**

It categorizes the dataset into three groups. The first one with less than 10 attributes, the second with at least 10 and at most 100 attributes, and the last one with more than 100 attributes.

- **#Instances**

Based on the attributes of the number of instances, the categorization is also done in three groups. The first one with less than 100 instances, the second with more than 100 and less than 1000 instances, and the last one contains more than 1000 instances.

- **Format Type**

It has two types either matrix or non-matrix format.

3.2 Data acquisition

The HCV data is an open-access database and is free of collection and usage. Therefore, we selected the data published in 2020, with less than 1000 instances, the number of attributes between 10 and 100, and the data type multivariate.

3.3 Dataset description

The HVC_DATA dataset has 14 columns where each of the following listed represents a variable. They are X, Category, Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT.

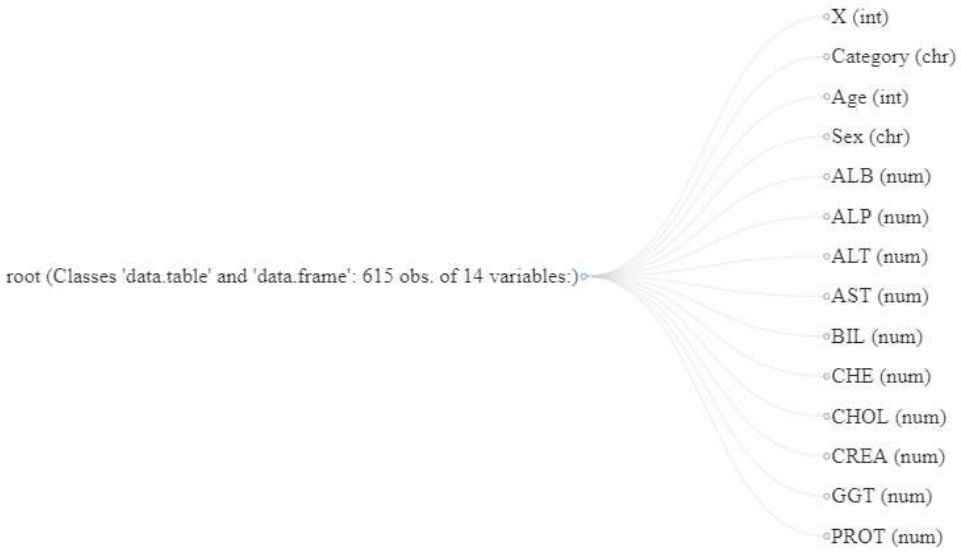
- X: is a numeric variable that represents a patient ID.
- Category: a categorical variable that divides groups into which a patient fit.
- Age: it is a numeric variable that holds the patient age
- Sex: is the patient gender
- ALB stands for the albumin level
- ALP stands for Alkaline phosphatase. Which is an enzyme found in the liver, bile ducts, and bone.
- ALT stands for Alanine Transaminase and represents an enzyme level found in the Liver that help0 the body metabolize protein. Its normal value is 7 to 55 units per liter.
- AST stands for Aspartate transaminase. This enzyme helps metabolize alanine- it is an amino acid.

- BIL stands for Bilirubin and is a yellow bile pigment that appears red blood cells break down.
- CHE stands for Cholinesterase and improves the health functionality of the nervous system.
- CHOL stands for Cholesterol that regulates the biological process of substrate presentation and the enzymes that use substrate presentation as a mechanism of their activation.
- CREAT stands for Creatinine and is used to serum creatine.
- GGT stands for Gamma-glutamyltransferase. Which is an enzyme found in the blood.
- PROT stands for a proteolytic enzyme that breaks long chainlike molecules of proteins into peptides and eventually into amino acids.

The figure below recalls all the variables and their types.

Figure 3

List of dataset variables



3.4 Data Analysis

The analysis of the data was done with the R language through R studio software on a 16go of ram computer. The steps followed after acquiring the data consisted of Data cleaning, Data description, and distance matrices computation. It resulted that the dataset has 31 missing values and we imputed them in the dataset using the *mice library*.

After clearing the missing values, we computed distance matrices for five distances measures that are: Euclidean, Manhattan, Canberra, maximum, and Minkowski on the first 20 patients of the dataset. This was done just to have a clear heatmap for each distance matrix. Because computing all the individuals yields an unreadable heatmap.

We used the whole dataset to compare the computed distance matrices using the Mantel correlation test and found that Euclidean and Minkowski distance matrices are 100% similar while Canberra and Maximum distance matrices are just 59.59% similar.

Besides, a binary distance matrix has not been computed for the reason that the dataset has no binary values. Since the data also contains a categorical value, the Gower distance was used to compute a distance matrix for mixed data (both categorical and continuous values).

Furthermore, for each of the distance matrices computed, a multivariate distance matrix regression (MDMR) was computed and the results showed that Minkowski and Euclidean distance matrices seem to be similar.

4 CHAPTER IV

Findings

In this chapter, we mainly discuss the results of our analysis.

4.1 Data exploration

4.1.1 Data Observation

Figure 4

Dataset Observation using the Glimpse Function

```
> glimpse(df)
Rows: 615
Columns: 14
$ X        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, ~
$ Category <chr> "0=Blood Donor", "0=Blood Donor", "0=Blood Donor", "0=Blood Donor", "0=Bloo~
$ Age      <int> 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33, 33, 33,~
$ Sex      <chr> "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", ~
$ ALB      <dbl> 38.5, 38.5, 46.9, 43.2, 39.2, 41.6, 46.3, 42.2, 50.9, 42.4, 44.3, 46.4, 36.~
$ ALP      <dbl> 52.5, 70.3, 74.7, 52.0, 74.1, 43.3, 41.3, 41.9, 65.5, 86.3, 52.3, 68.2, 78.~
$ ALT      <dbl> 7.7, 18.0, 36.2, 30.6, 32.6, 18.5, 17.5, 35.8, 23.2, 20.3, 21.7, 10.3, 23.6~
$ AST      <dbl> 22.1, 24.7, 52.6, 22.6, 24.8, 19.7, 17.8, 31.1, 21.2, 20.0, 22.4, 20.0, 22.~
$ BIL      <dbl> 7.5, 3.9, 6.1, 18.9, 9.6, 12.3, 8.5, 16.1, 6.9, 35.2, 17.2, 5.7, 7.0, 6.8, ~
$ CHE      <dbl> 6.93, 11.17, 8.84, 7.33, 9.15, 9.92, 7.01, 5.82, 8.69, 5.46, 4.15, 7.36, 8.~
$ CHOL     <dbl> 3.23, 4.80, 5.20, 4.74, 4.32, 6.05, 4.79, 4.60, 4.10, 4.45, 3.57, 4.30, 5.3~
$ CREA     <dbl> 106, 74, 86, 80, 76, 111, 70, 109, 83, 81, 78, 79, 78, 65, 63, 71, 90, 102,~
$ GGT      <dbl> 12.1, 15.6, 33.2, 33.8, 29.9, 91.0, 16.9, 21.5, 13.7, 15.9, 24.1, 18.7, 19.~
$ PROT     <dbl> 69.0, 76.5, 79.3, 75.7, 68.7, 74.0, 74.5, 67.1, 71.3, 69.9, 75.4, 68.6, 68.~
```

A rapid observation shows that the dataset has 615 rows (patients) and 14 columns (Variables). Besides, the function shows some values of each variable and we can see that two of them are categorical while the rest is continuous.

4.1.2 Data Status

Table 2

Dataset Status Before imputation

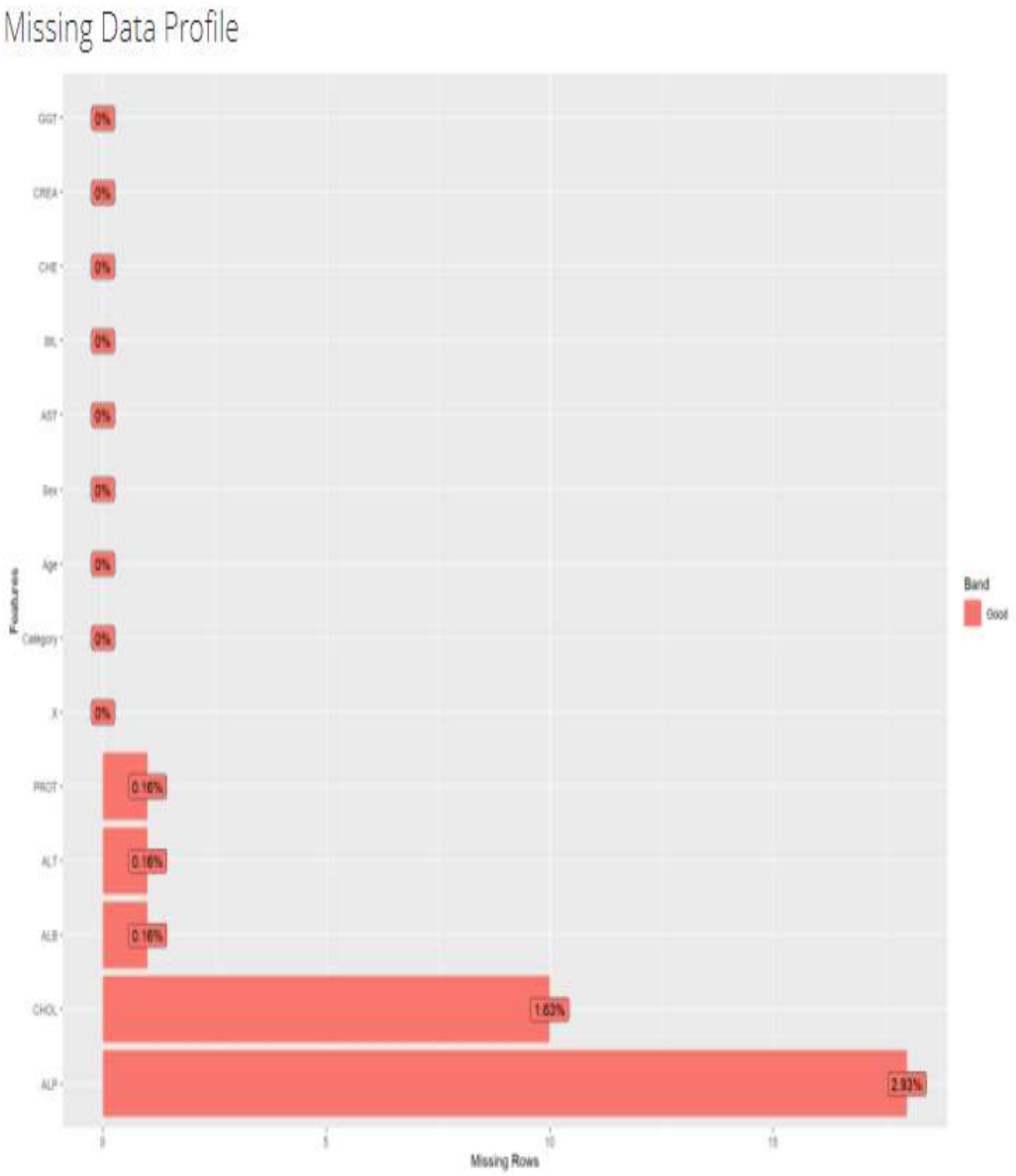
	q_zeros	p_zeros	q_na	p_na	q_inif	p_nif	type	unique
X	0	0	0	0	0	0	integer	615
Category	0	0	0	0	0	0	character	5
Age	0	0	0	0	0	0	integer	49
Sex	0	0	0	0	0	0	character	2
ALB	0	0	1	0.001	0	0	numeric	189
ALP	0	0	18	0.029	0	0	numeric	414
ALT	0	0	1	0.001	0	0	numeric	341
AST	0	0	0	0	0	0	numeric	297
BIL	0	0	0	0	0	0	numeric	188
CHE	0	0	0	0	0	0	numeric	407
CHOL	0	0	10	0.016	0	0	numeric	313
CREA	0	0	0	0	0	0	numeric	117
GGT	0	0	0	0	0	0	numeric	358
PROT	0	0	1	0.001	0	0	numeric	198

The ***q_na*** states the number of missing values from each variable. The variable ALP has the greatest number of missing values that equals 18 followed by the CHOL variable with 10 missing values and PROT, ALT, ALB have respectively one missing value.

4.1.3 Missing data Histogram by Percentage

Figure 5

Histogram of Missing Value by Percentages

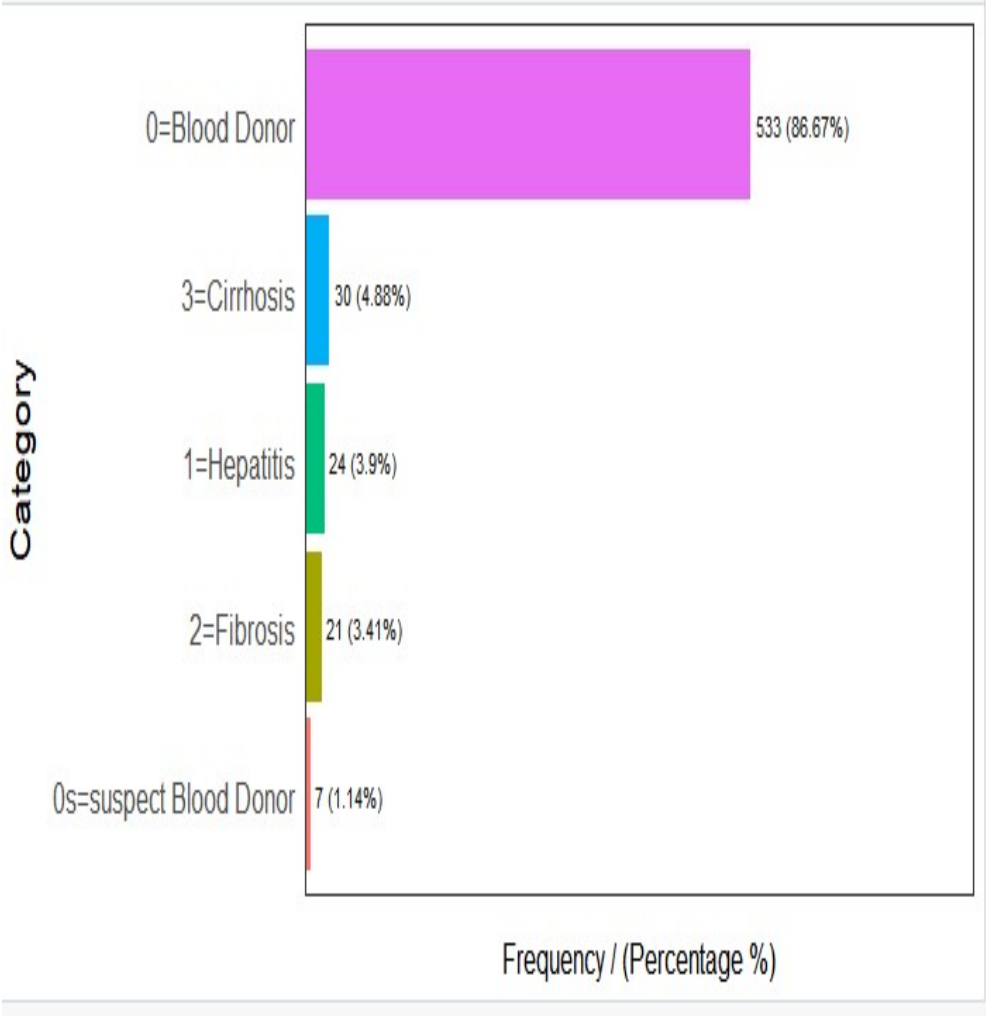


. The histogram shows that ALP has 2.92%, the highest percentage, and accounts for the highest number of missing values. The least percentage is for three variables and accounts for 0.16% of the missing information.

4.1.4 *Analysis of the Categorical Variable "Category"*

Figure 6

Frequency Distribution Histogram for the variable 'Category'

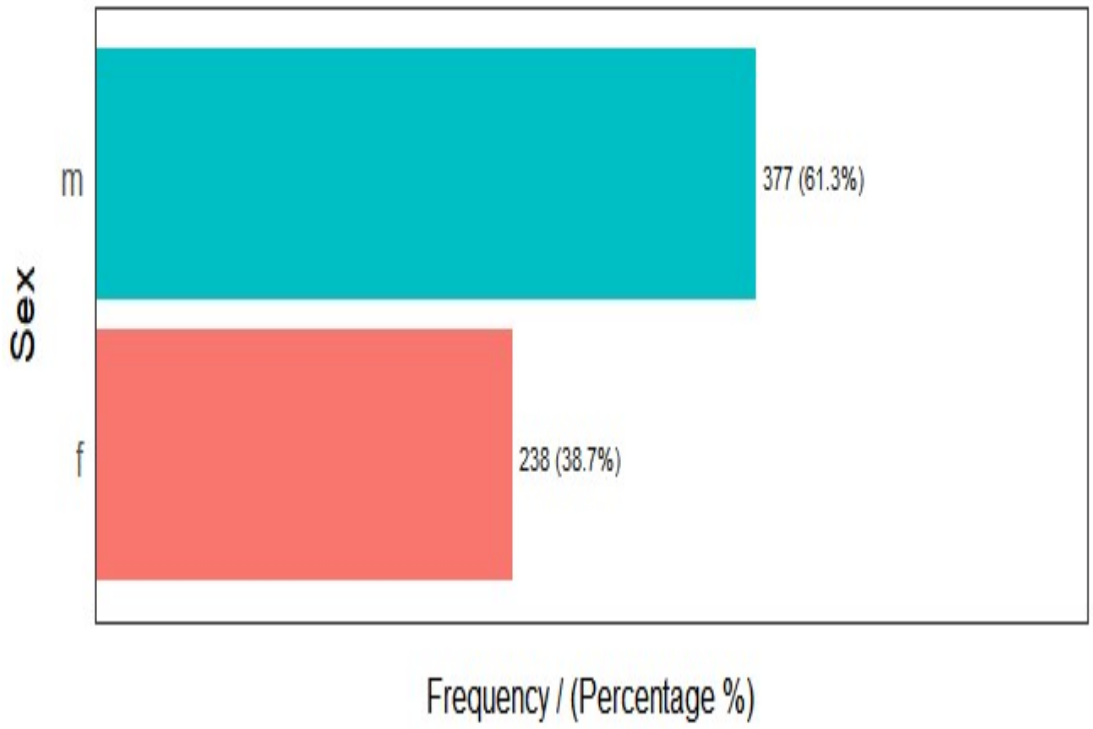


Interpretation:

The histogram shows that the variables have 5 groups known in ascending order as: "0s=suspect Blood Donor" (with 7 instances and accounts for 1.14%), "2=Fibrosis" (21 instances and accounts for 3.41%), "1=Hepatitis" (24 instances of 3.9%), "3=Cirrhosis"(30 instances of 4.88%), and "0=Blood Donnor"(533 instances for 86.67%). The latter accounts for the greatest amount of patients in the group.

Figure 7

Frequency Distribution Histogram for the variables 'SEX'



Interpretation:

The variable SEX has two attributes: "*f* for female" and "*m* for male". The former has 238 individuals and accounts for 38.7% while the latter has 377 individuals for 61.3%. Therefore, there are more male patients than female patients in this dataset.

4.1.5 Managing Missing Values

Table 3

Status of the Dataset After Imputing Missing Values

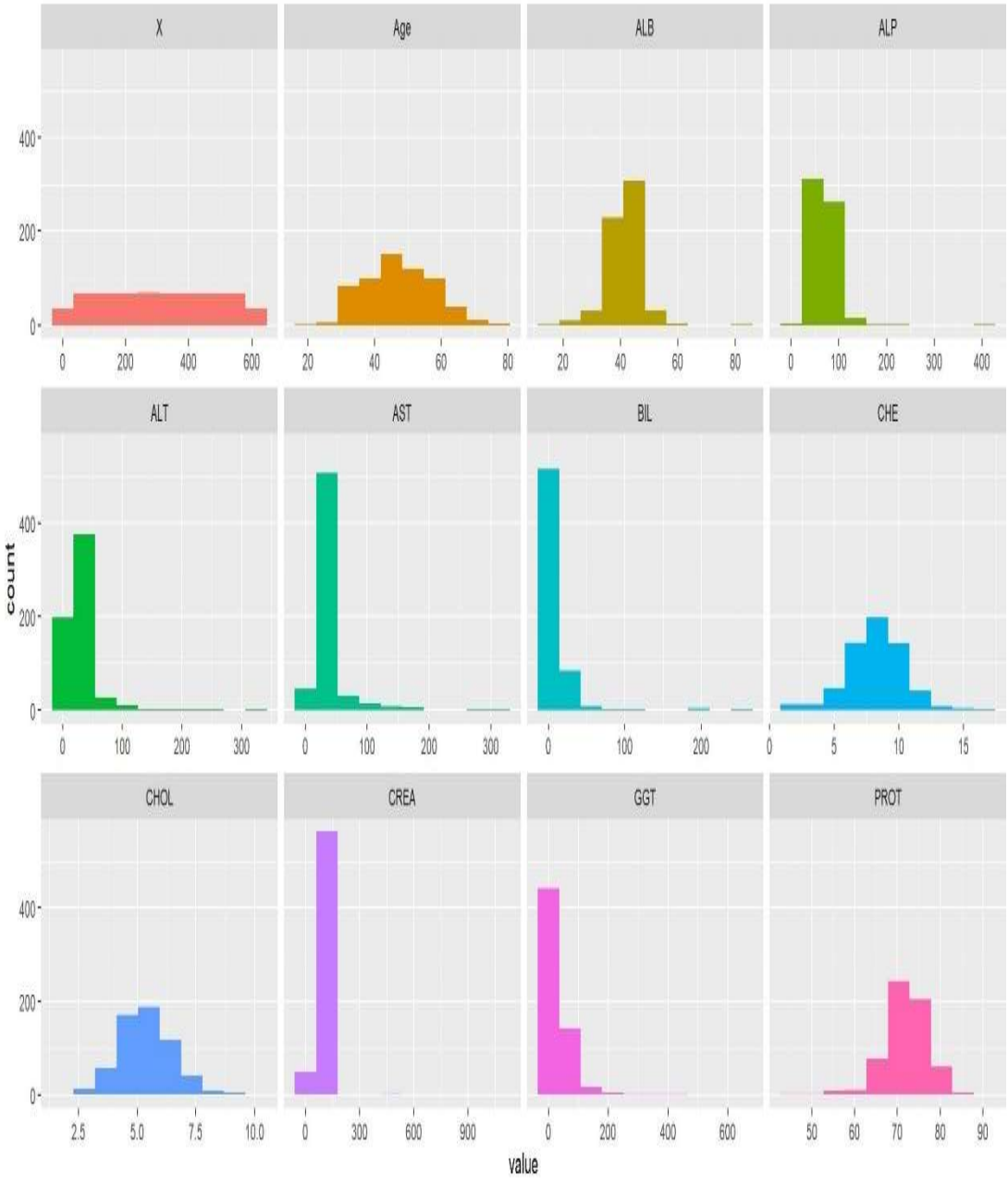
	q_zeros	p_zeros	q_na	p_na	q_inif	p_nif	type	unique
X	0	0	0	0	0	0	integer	615
Category	0	0	0	0	0	0	character	5
Age	0	0	0	0	0	0	integer	49
Sex	0	0	0	0	0	0	character	2
ALB	0	0	0	0	0	0	numeric	189
ALP	0	0	0	0	0	0	numeric	414
ALT	0	0	0	0	0	0	numeric	341
AST	0	0	0	0	0	0	numeric	297
BIL	0	0	0	0	0	0	numeric	188
CHE	0	0	0	0	0	0	numeric	407
CHOL	0	0	0	0	0	0	numeric	313
CREA	0	0	0	0	0	0	numeric	117
GGT	0	0	0	0	0	0	numeric	358
PROT	0	0	0	0	0	0	numeric	198

This data status table shows that there are no more missing values after their imputation.

4.1.6 Analysis of Continuous Variables

Figure 8

Distribution of Continuous Data

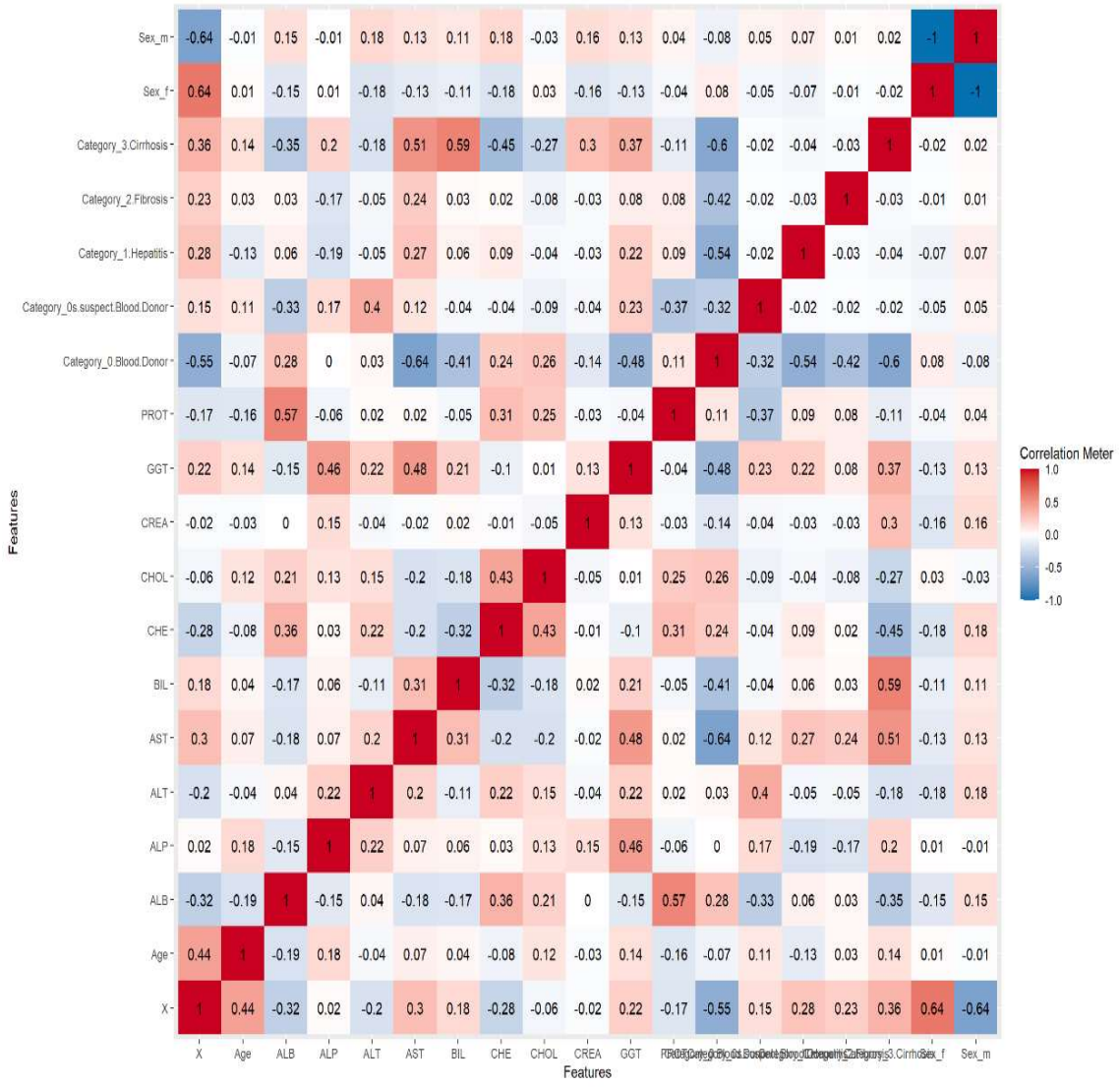


The figure above presents the normal distribution of all the continuous variables. The dataset has 12 variables among which, 5 variables (AST, ALT, BIL, CREAT, and GGT) exhibit outliers' values. In the meantime, variables, PROT, CHOL, and CHE tend to be normally distributed.

4.1.7 Correlation Matrix of Variables Before Imputing Missing Values

Figure 9

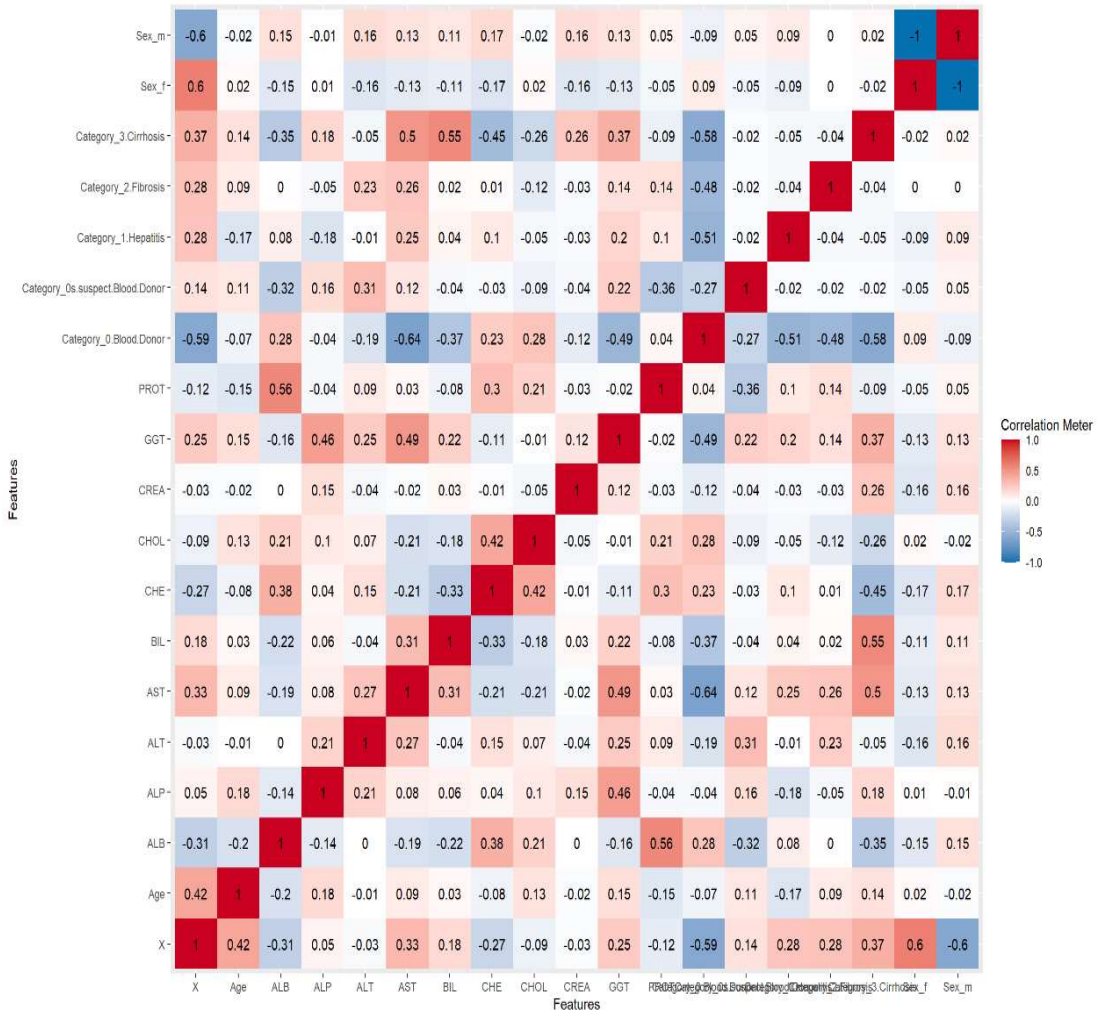
Correlation Matrix Before Imputation of Missing Values



4.1.8 Correlation Matrix of Variables after Imputing Missing Values

Figure 10

Correlation Matrix of Variables after Missing Values Imputation



F*igure 11* and *figure 12* above show the correlation matrices of variables *before* and *after* imputing missing values in the dataset. The Pearson correlation coefficient has been applied to the dataset variable. It results that the level of PROT and ALB are positive and moderately correlated with a correlation coefficient of 57% before imputation and a slight decrease to 56 % after the imputation of missing values. The graphs also depict the moderate and positive correlation between "category3=Cirrhosis" and BIL with a correlation coefficient of 59% before managing missing values and drops to 55% after the imputation.

In addition, the correlation between “category0=Blood Donor” and variable ALT remained negatively correlated at a constant of -64% both before and after the imputation of missing values.

4.2 Distance matrices computation

In this part of the work, we select five distances and compute their distance matrices on the dataset with imputed missing values. The distance matrices computed for individuals or patients account for only 20 rows for the sake of visibility.

Distances matrices were computed on both individuals and variables spaces using six-well known distances that are: Euclidean, Manhattan, Minkowski, Canberra, maximum, and Gower.

First of all, in the space of variables, distance matrices for individuals were computed. The heatmaps (see from Figure 14 to Figure 25) in this work show the similarities of the first 20 patients. This sample finds more understanding in the frame of clarity. Using all the patients yielded an unreadable heatmap. So for each of the distances measures or metric, the output heatmaps are colored from pure white to show the smallest distances—their similarities and pure red to exhibit the longest distance and therefore their dissimilarities. Amongst the first 20 patients selected in the dataset, the 6^{th} seems to be more distant with the rest from Euclidean, Manhattan, Minkowski, and Maximum distance matrices heatmaps (See picture).

Just to see how patients are closed to each other based on the variables, distance matrices were computed in the individuals' space and their heatmaps show us that using the Euclidean distance, the CREA, is dissimilar to ALT, AST, and BIL. In addition, BIL is very dissimilar to almost all the variables except for ALT and CREA. With Manhattan distance, CHE is distant to ALP, GGT, and CHOL, while ALB is very dissimilar to all, except GGT, PROT, and BIL respectively. For maximum distance, CHE is distant to all the variables but very distant to BIL at the same time the variables in the couples (AST, ALT) and (GGT, CHOL) seem to be similar. Moreover, GGT is very dissimilar to ALP, BIL, AGE, ALT, AST, and ALB. Coming up to Canberra distance, AST is closer to BIL, AGE, CREA, and ALT than ALP, CHOL, CHE, and GGT. Considering the Minkowski distance, the variables are almost similar or dissimilar as there are with the Euclidean distance.

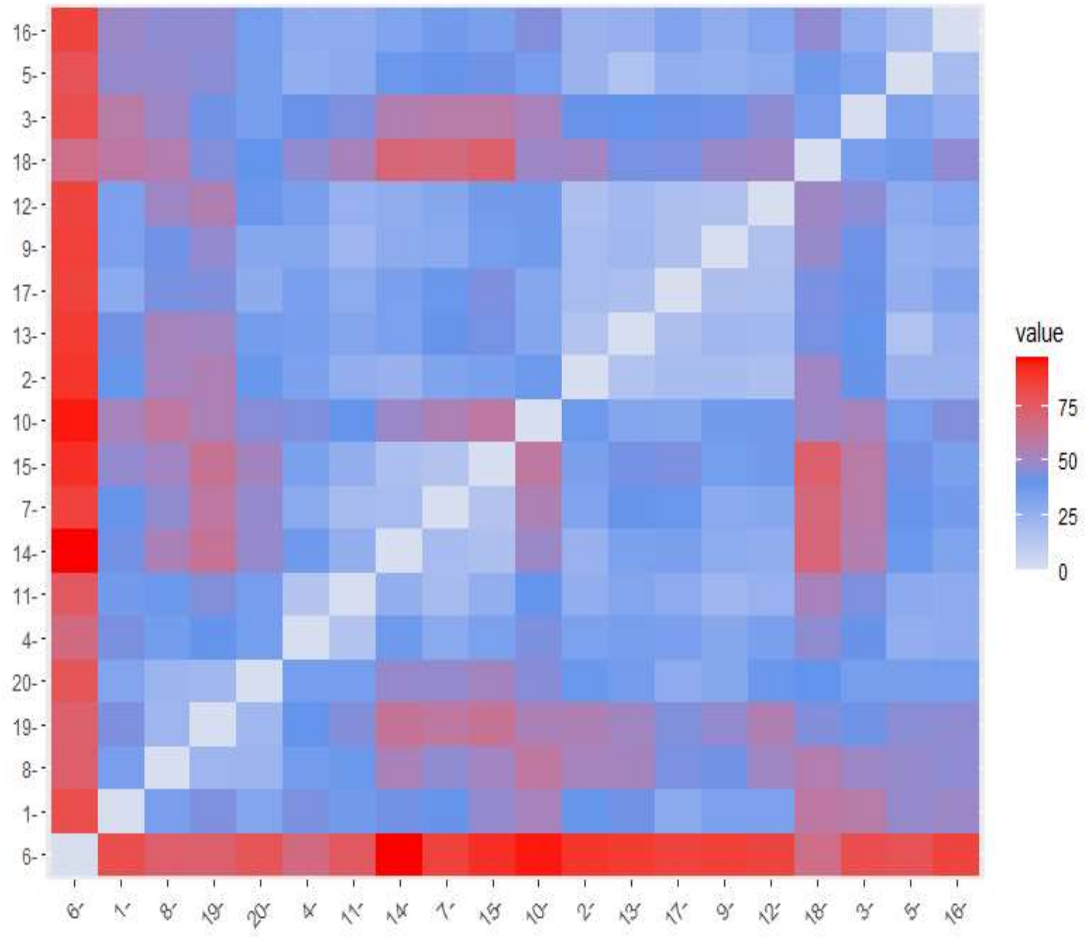
Furthermore, the correlation heatmap was computed using all the variables with scores between -1 and 1 stating a strong negative or strong positive relationship and 0 indicating no relationship at all. It turns out that PROT and ALB have a strong positive relationship with a correlation score of 0.56 followed by the couple GGT-ALP and GGT-AST with correlation scores of 0.46 and 0.49 respectively. In addition, patients with Cirrhosis also exhibit a strong relationship with AST and BIL with correlation scores of 0.5 and 0.55 respectively. In the meantime, Category_0 Blood Donor shows a strong negative correlation score of -0.64, -0.58, -0.51, -0.49, -0.48, and -0.37 with AST, Category_3Cirrhosis, Category_1Hepattities, GGT, Category_2 Fibrosis, and BIL.

4.2.1 Euclidean Distance Matrix

- **For individual**

Figure 11

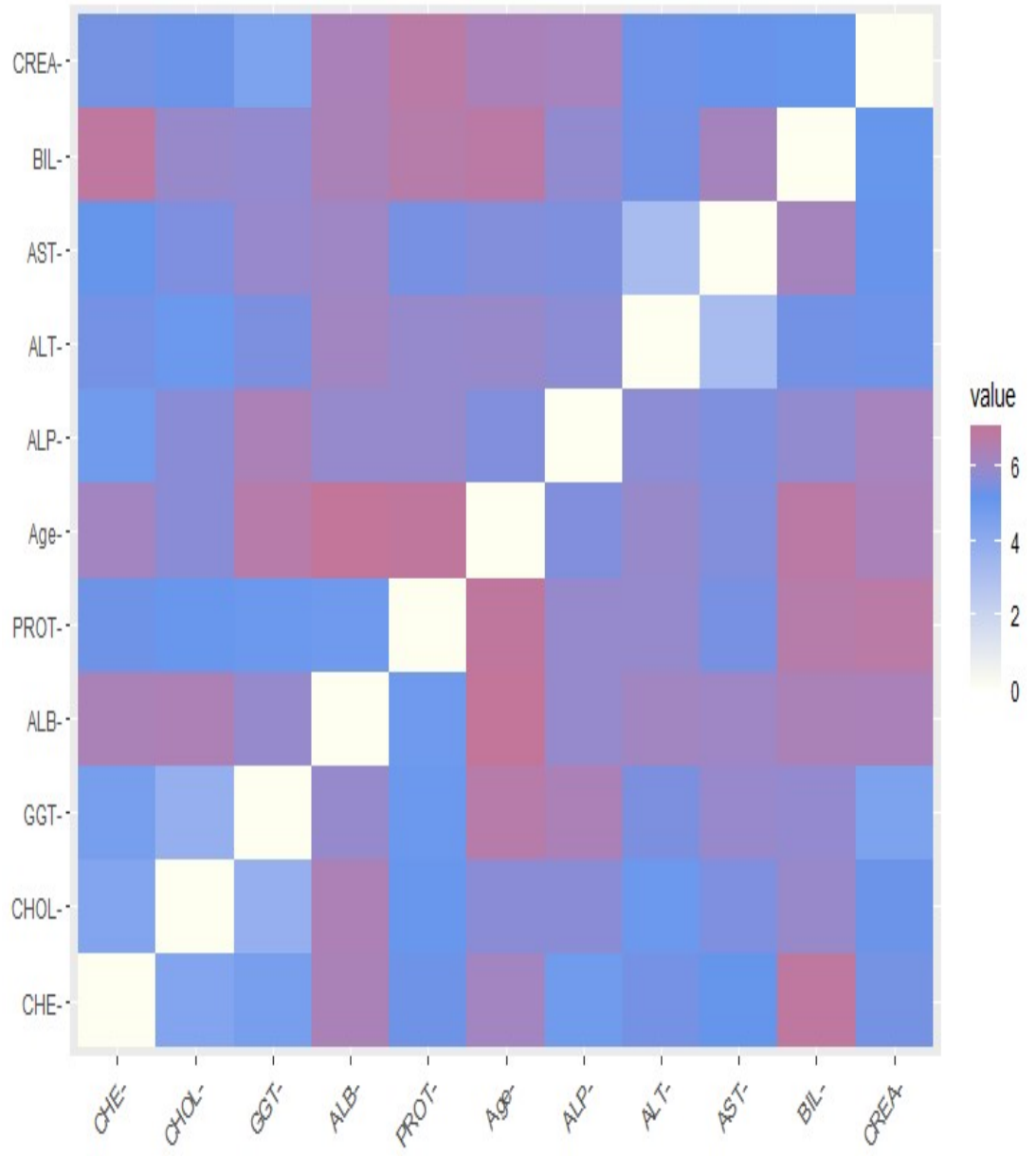
Euclidean Distance Matrix Heatmap for Individuals



- For variables

Figure 12

Euclidean Distance Matrix Heatmap for Variables

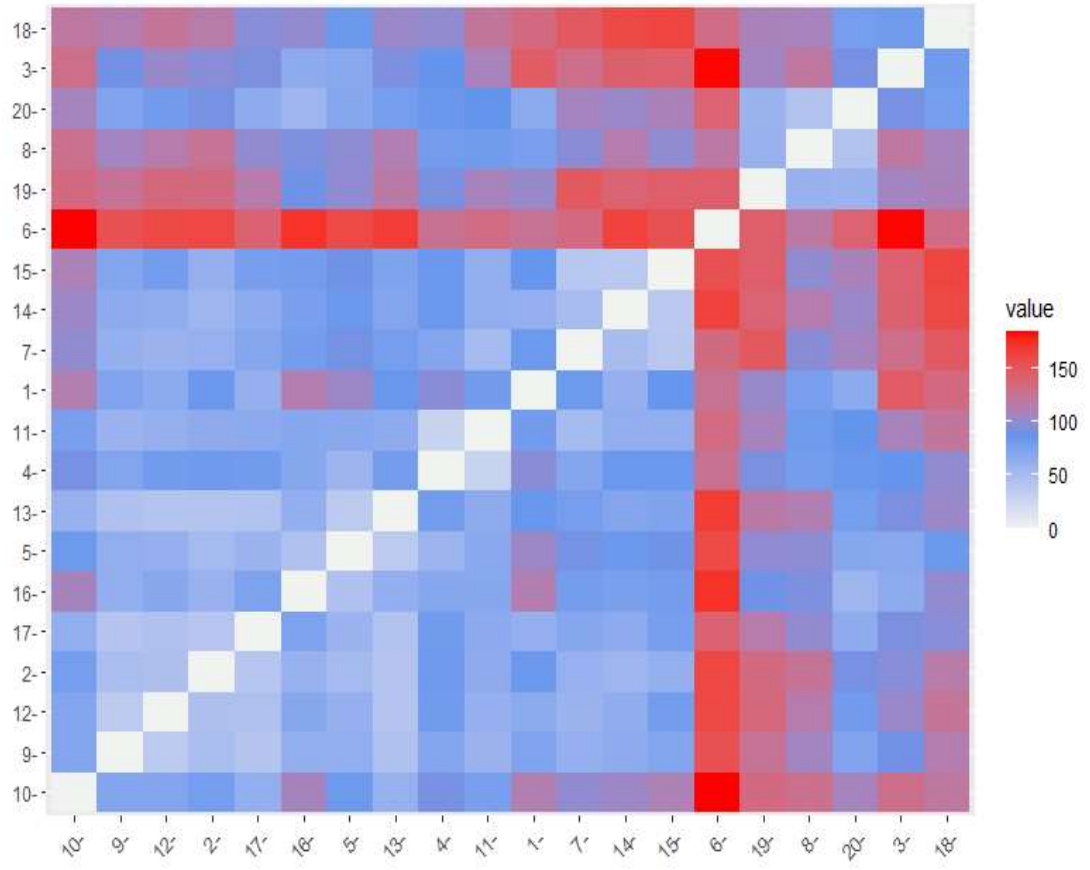


4.2.2 Manhattan Distance Matrix

- **Individuals**

Figure 13

Manhattan Distance Matrix Heatmap for Individuals

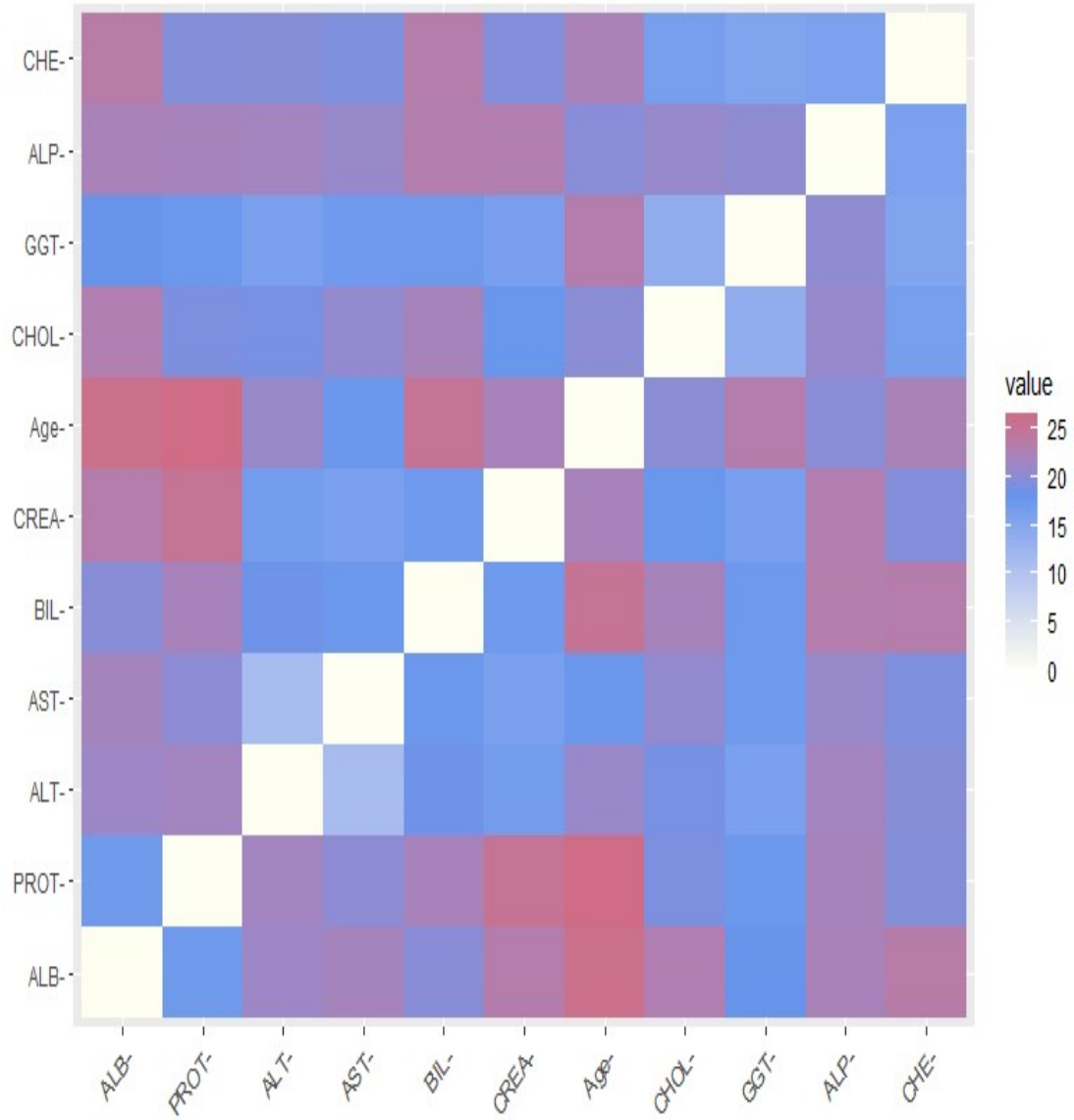


Computing the Manhattan distance matrix on 20 patients shows that patient 6 seems to be most dissimilar to the rest of the patients.

- Variables

Figure 14

Manhattan distance matrix for variables

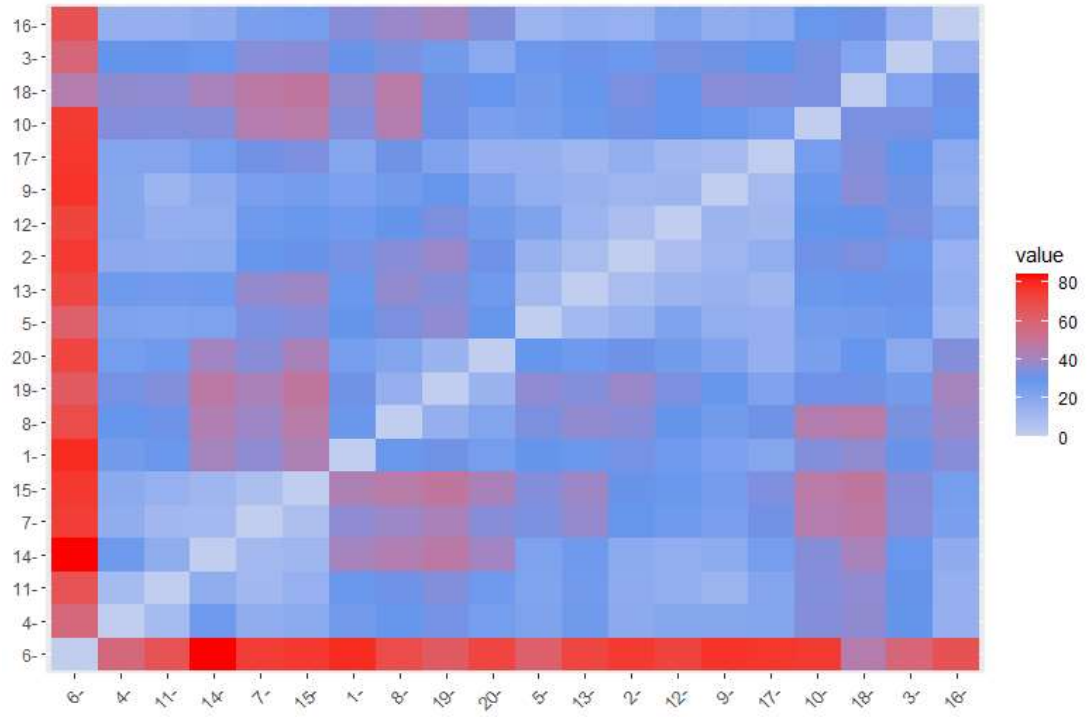


4.2.3 Maximum Distance Matrix

- **Individuals**

Figure 15

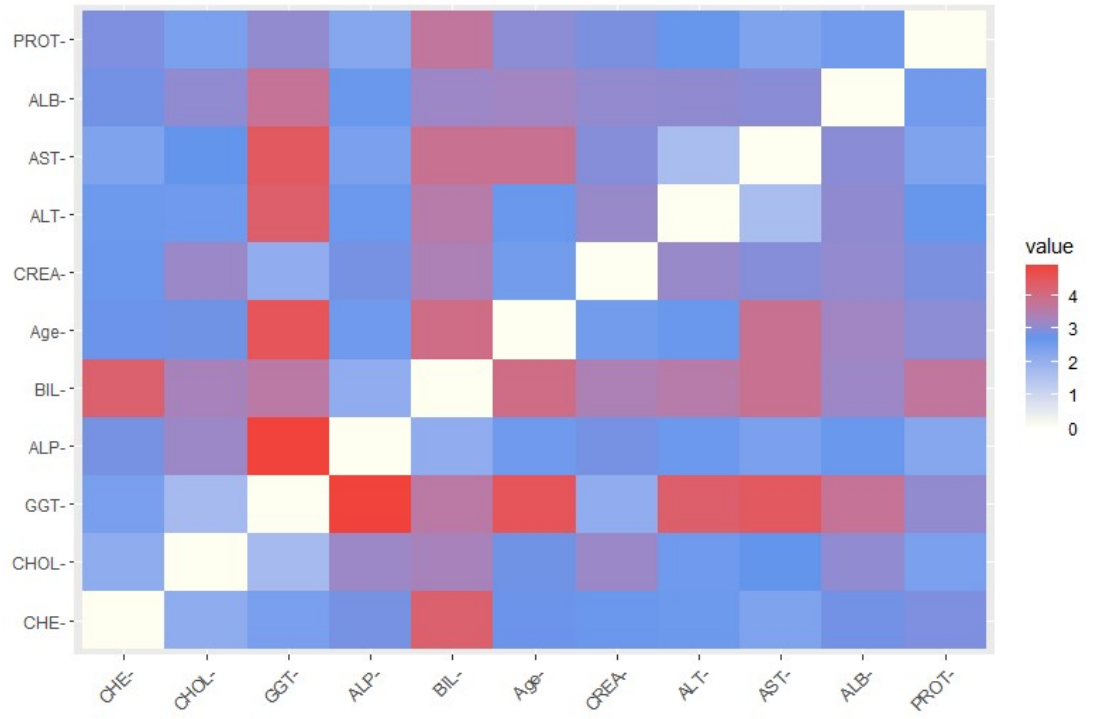
Maximum Distance matrix HeatMap for Individuals



- Variables

Figure 16

HeatMap of Maximum Distance Matrix for Variables

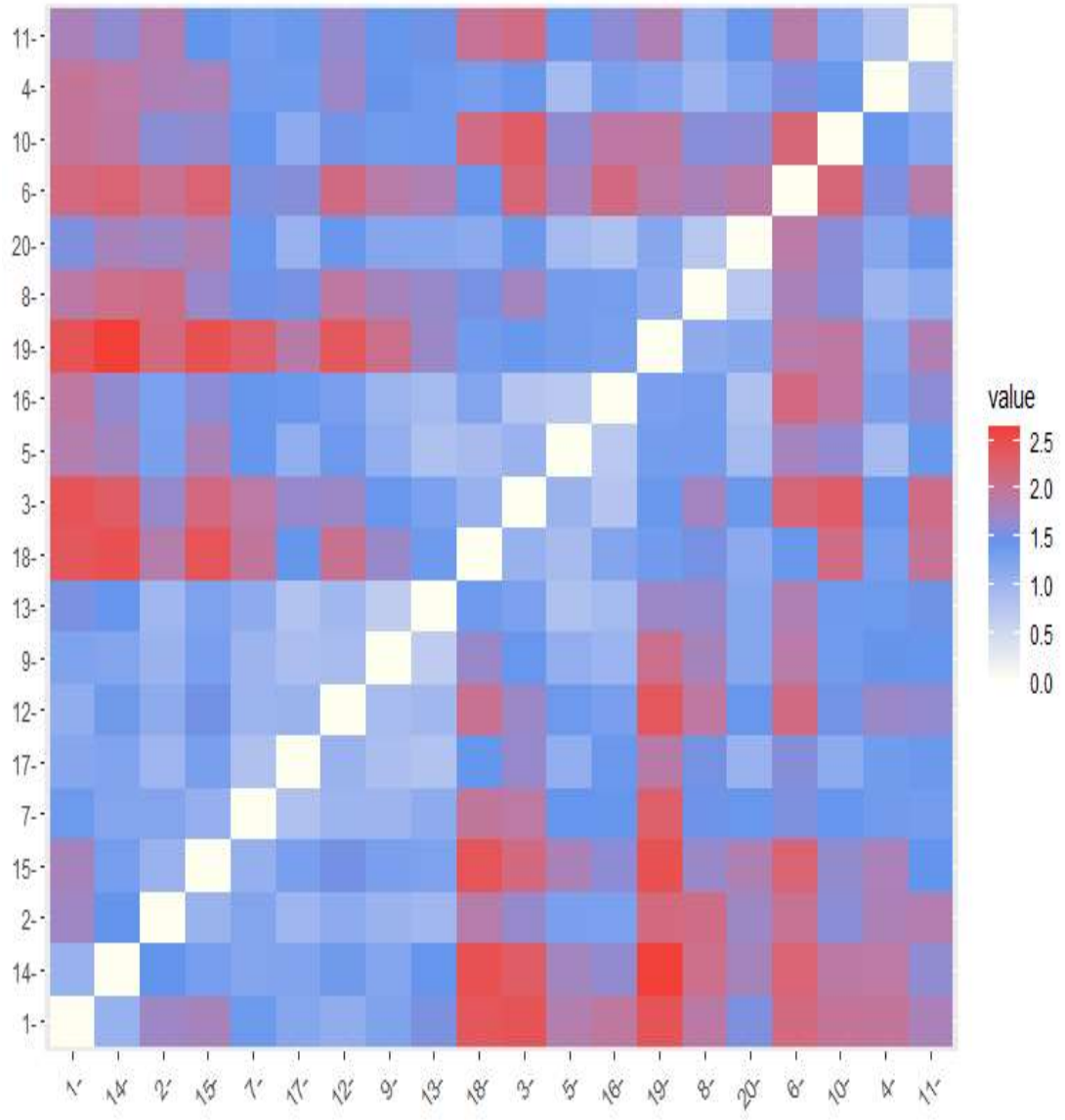


4.2.4 Canberra Distance Matrix

- **Individuals**

Figure 17

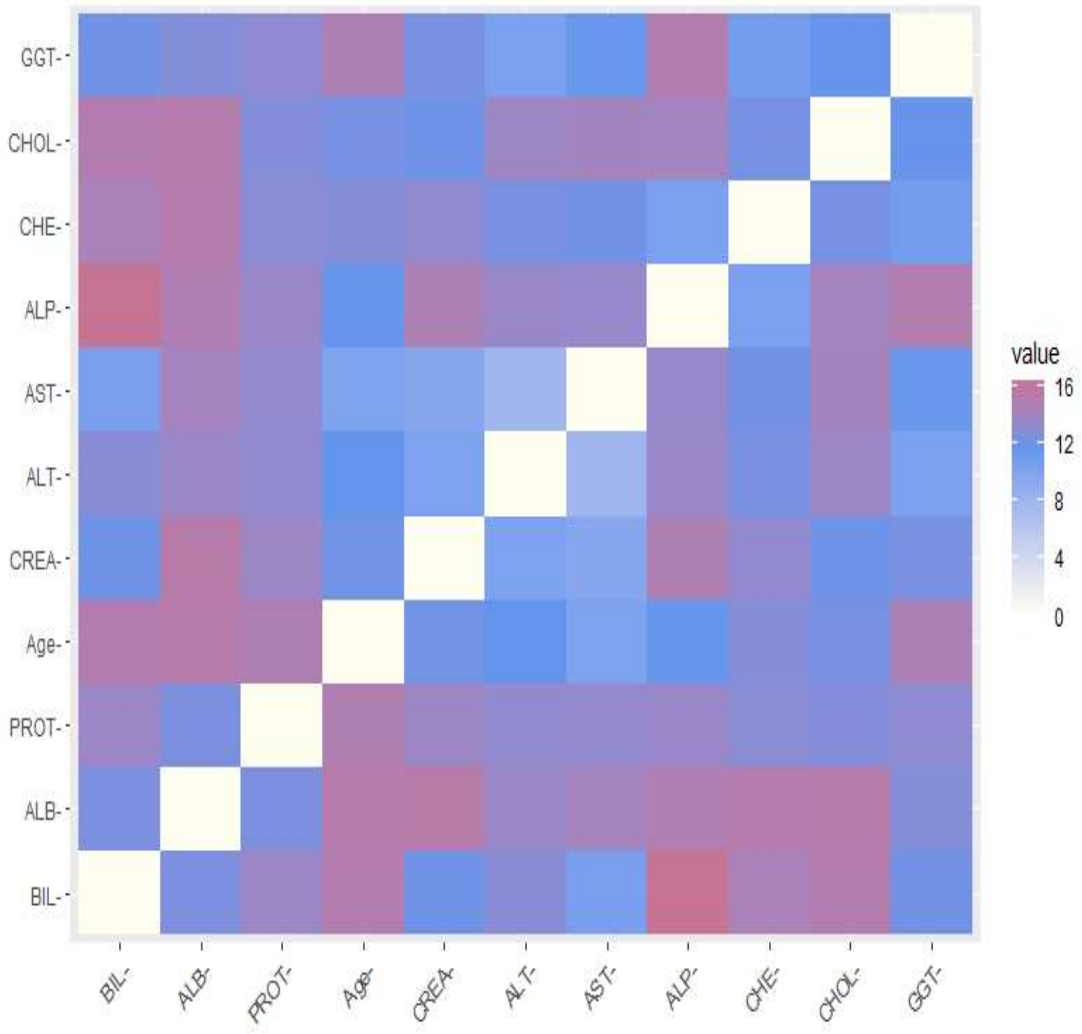
Canberra Distance Matrix Heatmap for Individuals



- **Variables**

Figure 18

Heatmap for Canberra Distance Matrix for Variables

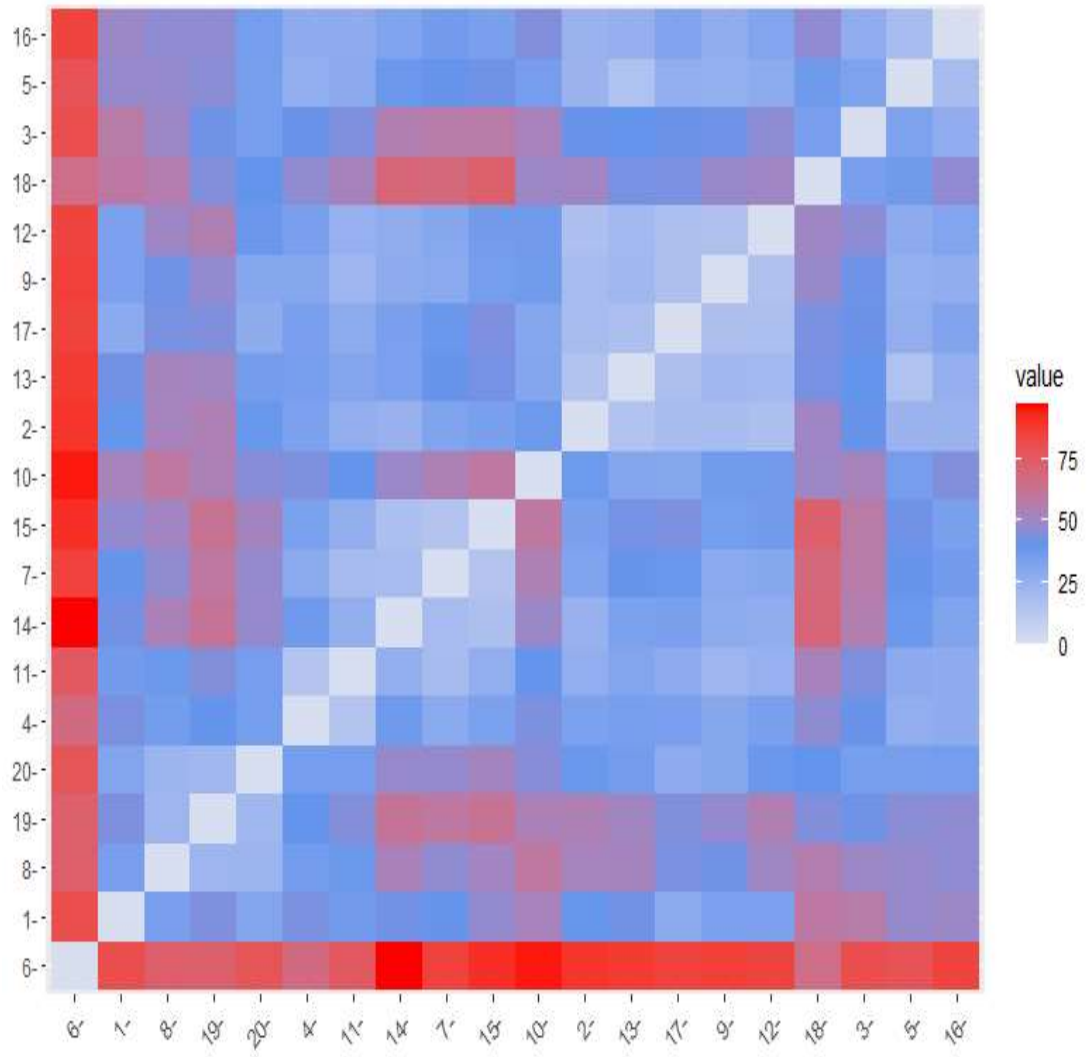


4.2.5 Minkowski Distance Matrix

- **Individuals**

Figure 19

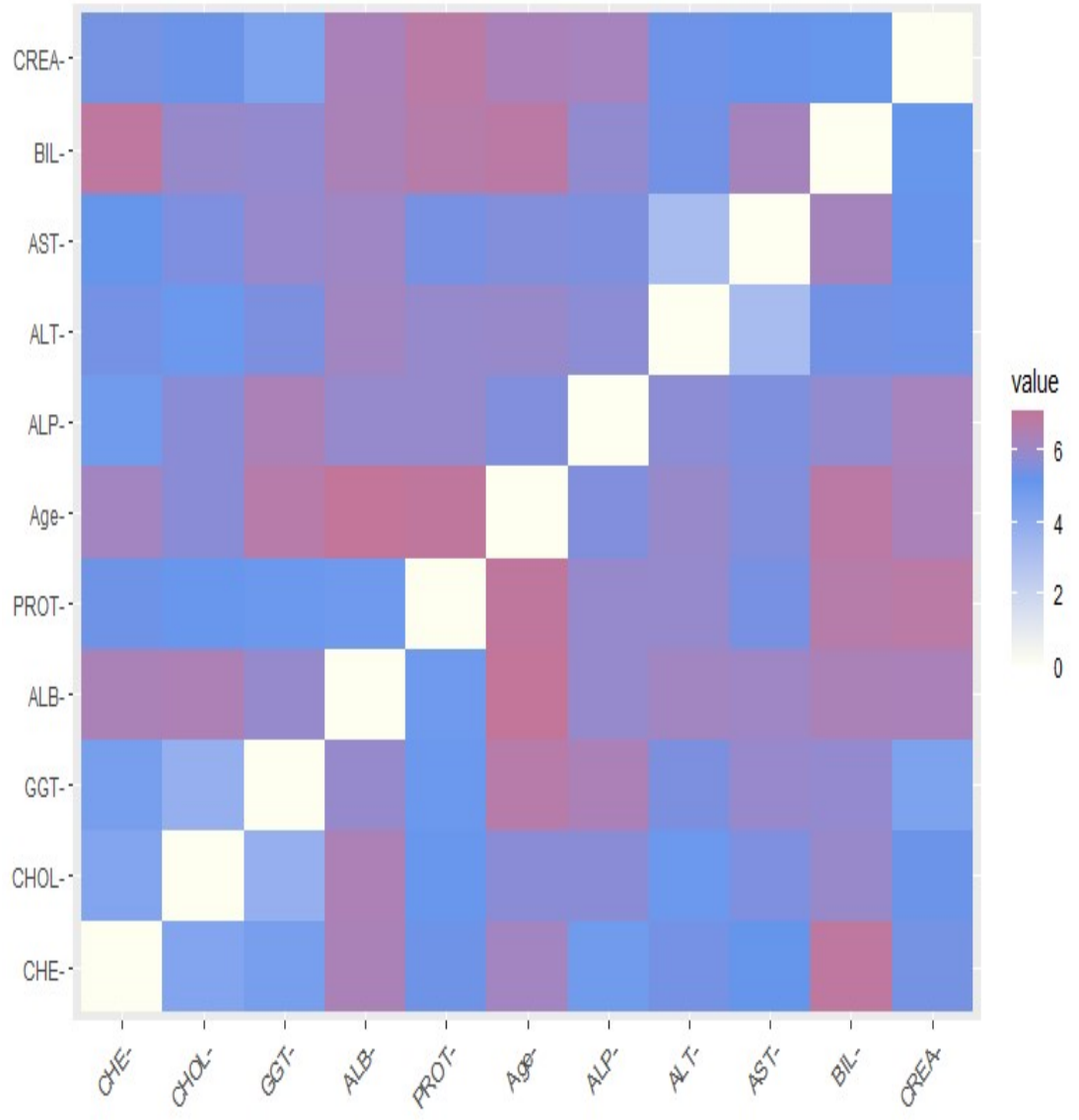
HeatMap using Minkowski distance matrix for individuals



- **Variables**

Figure 20

Minkowski Distance Matrix for Variables

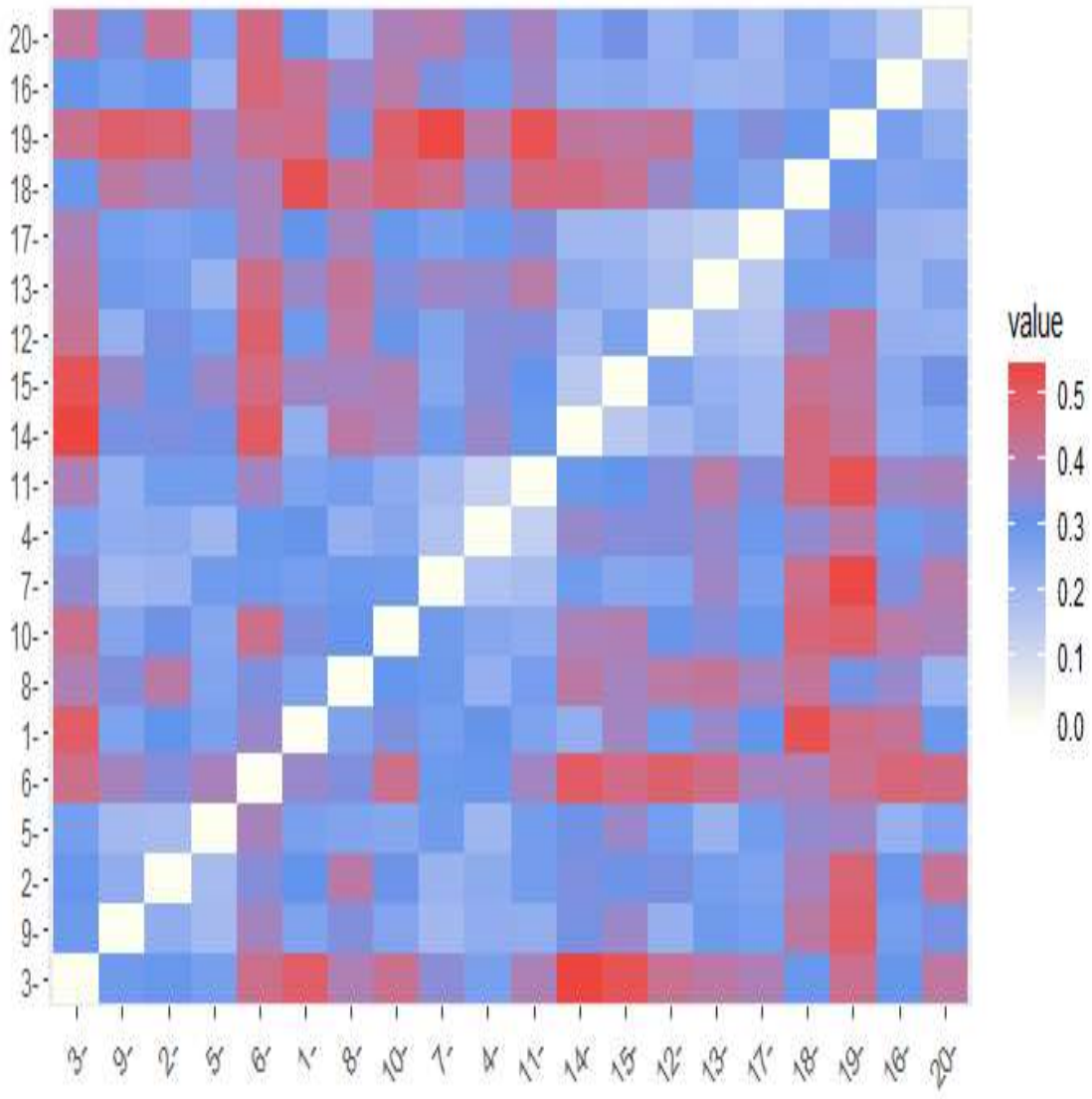


4.2.6 *Gower Distance Matrix*

The following heatmap is obtained by applying the Gower distance on the 20 first individuals of the dataset.

Figure 21

Heat map of the Gower Distance Matrix



The figure above only shows 20 patients and their similarities or dissimilarity. The evaluation of them all showed that patients 210 and 184 are most similar to any other group of two while patients 546 and 611 are the most dissimilar.

4.3 Mantel Correlation Score

The Mantel test is used to test the similarity of two matrices. In the following table is the observations score of the comparison of distance matrices. That is for example to explain whether or not the distance between two patients evaluated using the Euclidean distance is the same as evaluated with the Minkowski distance for our dataset.

Table 4

Correlation Scores of Distances Matrices Similarities

	EDM	**MANDM**	**MAXDM**	**CANDM**	**MINKDM**
EDM	1	0.9727947	0.9910533	0.6510335	1
MANDM		1	0.9379017	0.7601727	0.9727947
MAXDM			1	0.5958981	0.9910533
CANDM				1	0.6510335
MINKDM					1

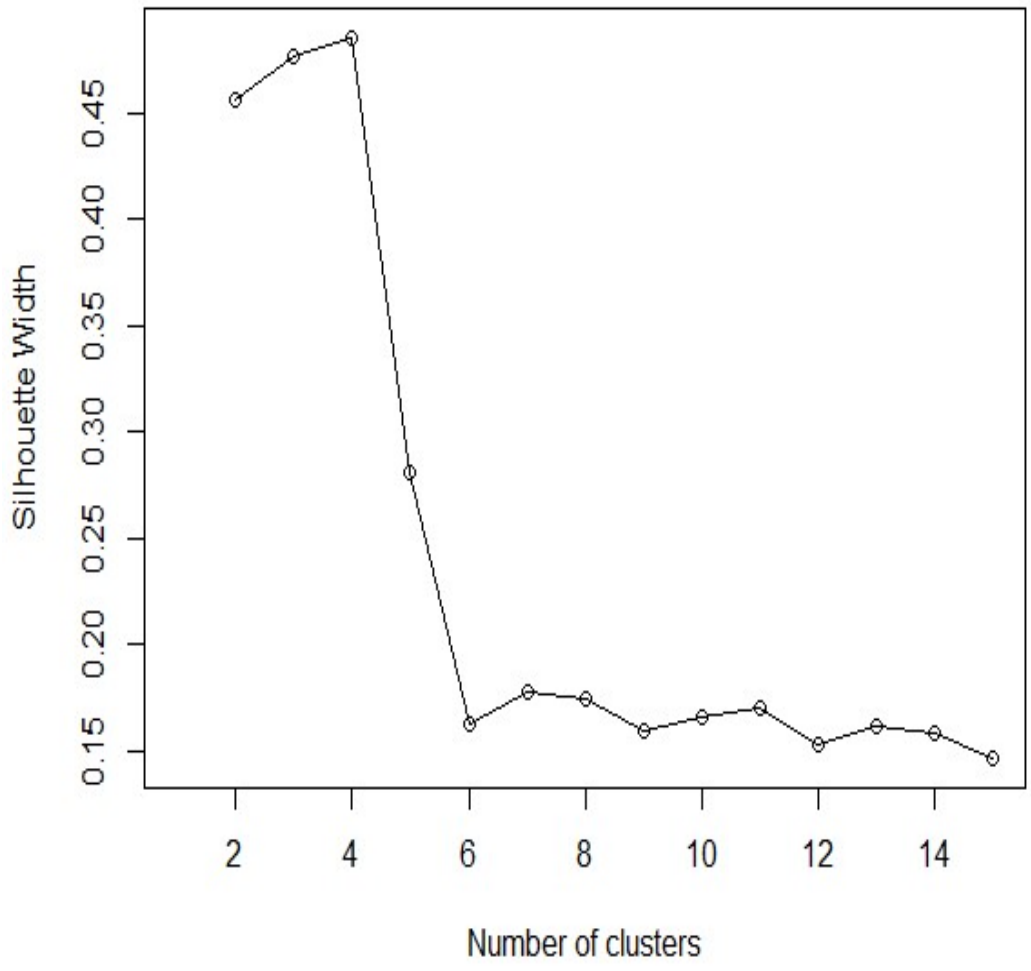
The tables show that the Euclidean distance matrix (EDM) and Minkowski Distance matrix (MINDM) are 100% correlated and account for the highest correlation between two different distances measures used to compute distance matrices. Besides, the Canberra distance matrix (CANDM) shares the least similarity score of 59.59% with the maximum distance matrix (MAXDM). In addition, the couple (EDM and MAXDM) & (MAXDM and MINKDM) have the same mantel correlation score of 99.10%.

4.4 Clustering

The figure below illustrates the number of clusters. It can be seen that all the data can be grouped in three clusters – those with the highest silhouette width. But we are going to select the 4th one.

Figure 22:

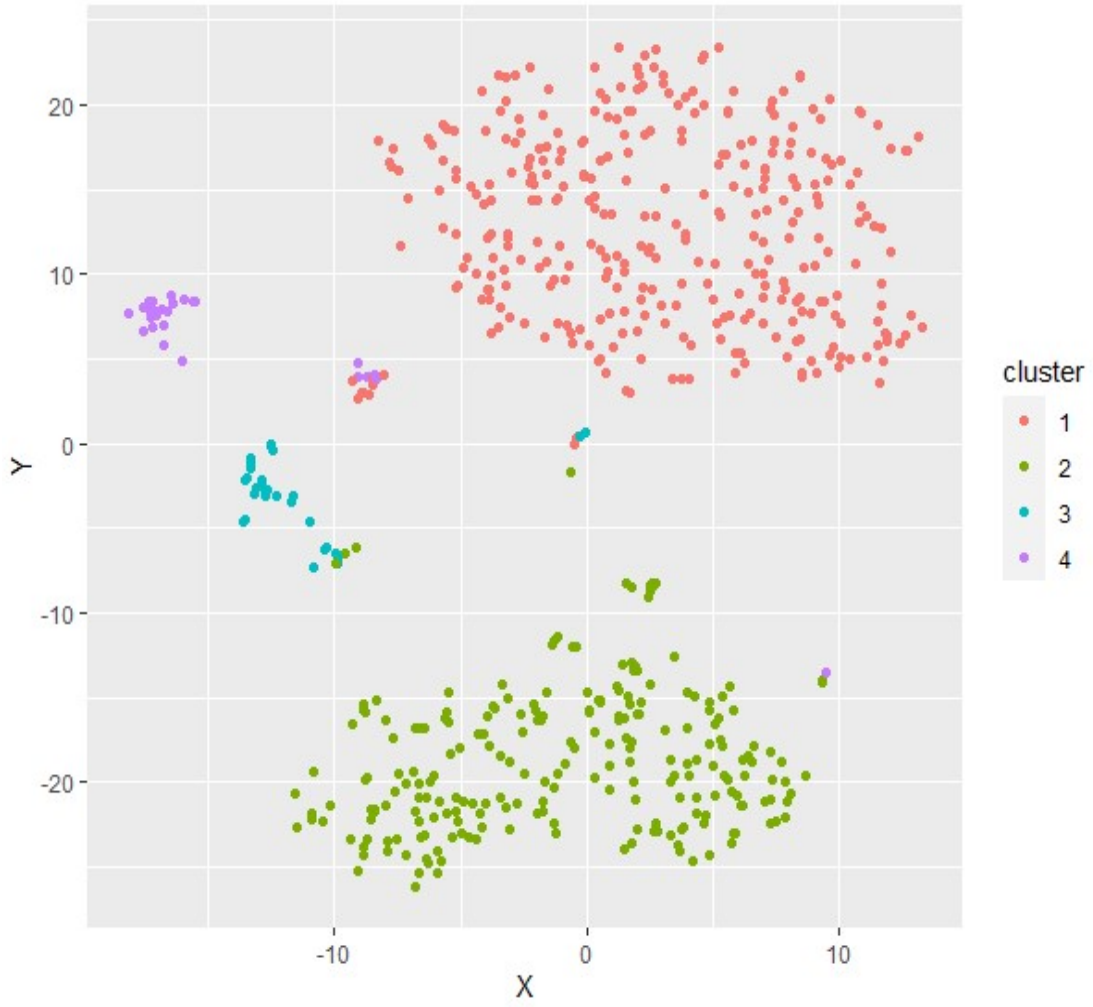
Silhouette Curve to find the Number of Clusters



After selecting 4 clusters, the clustering algorithm used in the process gives the figure below. Besides, clusters 1 and 2 are the ones most data are gathered around.

Figure 23

HCV Dataset Clustered in 4 Groups



4.5 Multivariate Distance Matrix Regression

In the following part, the Multivariate Distance Matrix Regression is computed with the Euclidean, Manhattan, maximum Canberra, Minkowski, and Gower distances. The results are described in tables like in table 5 to table 10. The lines are the variables' names and the columns are the results of the MDMR test. They are:

Statistic: which is the value of the corresponding MDMR test statistic

Number.DF is the numerator degrees of freedom for the corresponding effect

Pseudo.R2 is the size of the corresponding effect on the distance matrix

Analytic.p.Value is the p-value of each effect

4.5.1 MDMR with Euclidean Distance

Table 5

Multivariate Distance Matrix Regression results were obtained when using the Euclidean Distance.

	Statistic	Numer.DF	Pseudo.R2	Analytic.p.Value
(Omnibus)	-1.27e+14	11	1	1
Age	-1.36e+12	1	0.011	1
ALB	-2.94e+11	1	0.002	1
ALP	-7.27e+12	1	0.057	1
ALT	-8.11e+12	1	0.064	1
AST	-9.82e+12	1	0.077	1
BIL	-4.68e+12	1	0.037	1
CHE	-4.77e+10	1	0.000	1
CHOL	-1.43e+10	1	0.000	1

(Table 5 continued)

CREA	-3.54e+13	1	0.279	1
GGT	-2.55e+13	1	0.201	1
PROT	-2.79e+11	1	0.002	1

4.5.2 MDMR with Manhattan Distance

Table 6

Multivariate Distance Matrix Regression results obtained when using the Manhattan Distance.

	Statistic	Numer.DF	Pseudo.R2	Analytic.p.Value
(Omnibus)	14.050	11	0.934	< 1e-20
Age	0.486	1	0.032	< 1e-20
ALB	0.182	1	0.012	< 1e-20
ALP	0.128	1	0.075	< 1e-20
ALT	0.523	1	0.101	< 1e-14
AST	0.057	1	0.070	< 1e-20
BIL	0.803	1	0.053	< 1e-20
CHE	0.094	1	0.006	< 1e-20
CHOL	0.051	1	0.003	< 1e-20
CREA	0.321	1	0.154	< 1e-20
GGT	0.177	1	0.144	< 1e-20
PROT	0.211	1	0.014	< 1e-15

4.5.3 MDMR with Maximum Distance

Table 7

Multivariate Distance Matrix Regression results were obtained when using the Maximum Distance.

	Statistic	Numer.DF	Pseudo.R2	Analytic.p.Value
(Omnibus)	-137.1857	11	1.007	1
Age	-0.2965	1	0.002	1
ALB	0.1436	1	< 2e-16	< 1e-20
ALP	-3.0543	1	0.022	1
ALT	-5.6212	1	0.041	1
AST	-9.6672	1	0.070	1
BIL	-4.1086	1	0.003	1
CHE	-0.0738	1	0.000	1
CHOL	0.0738	1	< 1e-16	< 1e-15
CREA	-46.4500	1	0.341	1
GGT	-31.5977	1	0.232	1
PROT	-0.1359	1	0.000	1

4.5.4 MDMR with Canberra Distance

Table 8

Multivariate Distance Matrix Regression results were obtained when using the Canberra Distance.

	Statistic	Numer.DF	Pseudo.R2	Analytic.p.Value
(Omnibus)	1.6149	11	0.618	< 1e-20
Age	0.1297	1	0.050	< 1e-20
ALB	0.0511	1	0.020	< 1e-20

ALP	0.1296	1	0.050	< 1e-20
ALT	0.1601	1	0.061	< 1e-20
AST	0.0765	1	0.029	< 1e-20
BIL	0.1117	1	0.043	< 1e-20
CHE	0.1175	1	0.045	< 1e-20
CHOL	0.0954	1	0.037	< 1e-20
CREA	0.0507	1	0.019	< 1e-20
GGT	0.106	1	0.040	< 1e-20
PROT	0.039	1	0.015	< 1e-15

4.5.5 MDMR with Minkowski Distance

Table 9

Multivariate Distance Matrix Regression results were obtained when using the Minkowski distance.

	Statistic	Numer.DF	Pseudo.R2	Analytic.p.Value
Omnibus	-1.27e+14	11	1	1
Age	-1.36e+12	1	0.010	1
ALB	-2.94e+11	1	0.002	1
ALP	-7.27e+12	1	0.057	1
ALT	-8.11e+12	1	0.064	1
AST	-9.82e+12	1	0.077	1
BIL	-4.68 e+12	1	0.036	1
CHE	-4.77 e+10	1	0	1
CHOL	-1.43 e+10	1	0	1
CREA	-3.54 e+13	1	0.279	1
GGT	-2.55 e+13	1	0.201	1
PROT	-2.79 e+11	1	0.002	1

4.5.6 MDMR with Gower Distance Matrix

Table 10

Multivariate Distance Matrix Regression results were obtained when using the Gower distance.

	Statistic	Numer.DF	Pseudo.R2	Analytic.p.Value
Omnibus	-7.49	17	1.54	1
Category	-0.43	4	0.07	1
Age	-0.13	1	0.02	1
Sex	-0.54	1	0.08	1
ALB	-0.11	1	0.02	1
ALP	-0.09	1	0.01	1
ALT	-0.07	1	0.01	1
AST	-0.09	1	0.01	1
BIL	-0.05	1	0.01	1
CHE	-0.28	1	0.04	1
CHOL	-0.30	1	0.04	1
CREA	-0.02	1	0	1
GGT	-0.09	1	0.01	1
PROT	-0.24	1	0.03	1

5 CHAPTER V

Discussion

After computing the distance matrices, their similarities were tested using mantel but in this case on all the maintained patients and found that the Euclidean distance matrice and Minkowski distance matrice are 100% similar while Maximum distance matrice and Minkowski Distance matrix are 99% similar. In addition, the Canberra distance matrix and Maximum shared the least similarity score of 60%. Euclidean distance matrix and Manhattan distance matrix are similar at 97% while maximum and Manhattan Distance matrix are 94% similar. Canberra distance matrix shares a 65% In addition, the Multivariate Distance Matrix Regression (MDMR) is a hypothesis-testing method or multivariate data with the goals of finding a relationship between predictors and observations during experimentation by the mean of test-statistics. It was implemented to understand the relationship between experimental factors as input and the association of outputs variables. In this assignment, computed distance matrices (Euclidean, Manhattan, Minkowski, Canberra, and Maximum) were used for continuous variables and a Gower distance matrix for mixed variables. In *table 5* and *table 9*, the p-value obtained when computing the MDMR analysis with Euclidean and Minkowski distance matrice is 1 and all the statistics values are negative for all the variables and similarity with the Minkowski distance matrix. Therefore they are not statistically significant. That proves that Euclidean and Minkowski are not suitable for MDMR analysis. MUHANAD SHAB KALEIA (2017) [25] states in his thesis that Euclidean distance is not suitable for times series data in functional Magnetic Resonance Imaging (fMRI); this result corroborates with our HCV-data since the p-values of Euclidean is 1 and the statistics are all negative.

In the Multivariate Distance Matrix Regression computed with Gower distance matrix, the significance of the hypothesis is not suitable since all p-values are 1 and statistics are all negative. Therefore, for the current dataset comprised of mixed data the Gower distance is not suitable for similarity in between our matrices. In addition, Canberra and Manhattan distance matrices applied on the MDMR analysis showed that all variables have statistical significance and therefore prove the significance of their distance matrices in the application of the Multivariate Distance Matrix Regression. When using the maximum distance measures in the MDMR analysis, just

two variables namely ALB and CHOL with respective statistics of 0.14 and 0.07 plus p-values below 1e-15.

6 CHAPTER VI

Conclusion

The work aimed to study some distance measures and/or metrics and compute their matrices. For mixed variables (both categorical and continuous) their distance matrices were computed using the Gower distance. Besides, we used Maximum, Minkowski, Euclidean, Manhattan, and Canberra distance on the continuous variables of the dataset. However, Euclidean and Minkowski proved not to be suitable using MDMR analysis while Manhattan and Canberra proved to be appropriate for the MDMR test.

As to understand which methods are suitable for calculating distance matrices on categorical variables, there are not enough resources on the matter but trying to compute a distance metric on mixed data using one amongst the following distances Euclidean, Minkowski, Manhattan, Maximum and Canberra yields errors due to the type of data. However, only the Gower distance can perform the computation, and therefore, it seems to be the most appropriate one for mixed data. In the meantime, the data used in this work contained only two categorical values that are not enough to compute the matrix. Apart from Gower distance, the other explored distances were easily used to compute matrices on continuous variables.

For future work, it will be better to use and compare other measures to compute their matrices and therefore test them against each other in the given MDMR analysis. In addition, it will also be useful to investigate how distance metrics act for multivariate time series data. The literature on the subject is not rich enough and more investigation could be carried out on various datasets in different domains.

REFERENCES

[1] C. Nebeker, J. Torous, and R. J. Bartlett Ellis, “Building the case for actionable ethics in digital health research supported by artificial intelligence,” *BMC Medicine*, vol. 17, no. 1, pp. 1–7, 2019, DOI: 10.1186/s12916-019-1377-7.

[2] M. A. Zapala and N. J. Schork, “Statistical properties of multivariate distance matrix regression for high-dimensional data analysis,” vol. 3, no. September, pp. 1–10, 2012, DOI: 10.3389/fgene.2012.00190.

[3] M. W. Al-neama, N. M. Reda, and F. F. M. Ghaleb, “An Improved Distance Matrix Computation Algorithm for Multicore Clusters,” vol. 2014, 2014.

[4] J. Venna, “Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization,” vol. 11, pp. 451–490, 2010.

[5] A. Roman-Gonzalez and A. Roman-Gonzalez, “Compression Techniques for Image Processing Tasks,” 2013.

[6] I. Engineering, “A Novel Spectral Clustering Method Based on Pairwise Distance Matrix,” vol. 658, pp. 649–658, 2010.

[7] R. Hu, W. Jia, H. Ling, and D. Huang, “Multiscale Distance Matrix for Fast Plant Leaf Recognition,” vol. 21, no. 11, pp. 4667–4672, 2012.

[8] X. Hua, J. J. Goedert, M. T. Landi, and J. Shi, “Identifying host genetic variants associated with microbiome composition by testing multiple beta diversity matrices,” *Human Heredity*, 2017, DOI: 10.1159/000448733.

[9] Z. Shehzad *et al.*, “A multivariate distance-based analytic framework for connectome-wide association studies,” *NeuroImage*, vol. 93, no. P1, pp. 74–94, 2014, doi: 10.1016/j.neuroimage.2014.02.024.

[10] F. N. Baker and A. Porollo, “CoeViz: A web-based integrative platform for interactive visualization of large similarity and distance matrices,” *Data*, vol. 3, no. 1, Mar. 2018, DOI: 10.3390/data3010004.

[11] F. N. Baker and A. Porollo, “CoeViz: A web-based tool for coevolution analysis of protein residues,” *BMC Bioinformatics*, vol. 17, no. 1, 2016, DOI: 10.1186/s12859-016-0975-z.

[12] M. Markatou, Y. Chen, G. Afendras, and B. G. Lindsay, “Statistical Distances and Their Role in Robustness,” 2017. DOI: 10.1007/978-3-319-69416-0_1.

[13] M. Markatou and E. M. Sofikitou, “Statistical Distances and the Construction of Evidence Functions for Model Adequacy,” *Frontiers in Ecology and Evolution*, vol. 7, p. 447, Nov. 2019, DOI: 10.3389/fevo.2019.00447.

[14] P. Mulak and N. Talhar, “Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset,” 2013.

[15] H. Arafat Abu Alfeilat *et al.*, “Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review”, DOI: 10.1089/big.2018.0175.

[16] R. Todeschini, D. Ballabio, V. Consonni, and F. Grisoni, “A new concept of higher-order similarity and the role of distance/similarity measures in local classification methods,” *Chemometrics and Intelligent Laboratory Systems*, vol. 157, pp. 50–57.

[17] R. Ehsani and F. Drabløs, “Robust Distance Measures for kNN Classification of Cancer Data,” *Cancer Informatics*, vol. 19, 2020, DOI: 10.1177/1176935120965542.

[18] P. M. Dixon, L. Wu, M. P. Widrlechner, and E. Syrkin, “Weighted distance measures for metabolomic data,” *Iowa State University Department of Statistics Preprint Series*, no. 10, pp. 1–8, 2009.

[19] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, “The distance function effect on k-nearest neighbor classification for medical datasets,” 2016, DOI: 10.1186/s40064-016-2941-7.

[20] “(No Title).” http://www.fisica.edu.uy/~cris/teaching/Cha_pdf_distances_2007.pdf (accessed Mar. 03, 2021).

[21] G. N. Lance and W. T. Williams, “Mixed-Data Classificatory Programs I – Agglomerative Systems,” *Australian Computer Journal*, 1967.

[22] M. Faisal and E. M. Zamzami, “Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance,” *Journal of Physics: Conference Series*, vol. 1566, p. 12112, 2020, DOI: 10.1088/1742-6596/1566/1/012112.

[23] S. Ben Salem and S. Naouali, “Clustering Categorical Data: A Survey,” *Article in International Journal of Information Technology and Decision Making*, 2019, DOI: 10.1142/S0219622019300064.

[24] M. Angeletti, J.-M. Bonny, and J. Koko, “Parallel Euclidean distance matrix computation on big datasets Parallel Euclidean distance matrix computation on big datasets *.”

[25](Online)‘’https://shareok.org/bitstream/handle/11244/54305/2017_ShabKaleia_Muhanad_Thesis.pdf?sequence=3’’

7 Annex

Computation code in R programming language

```
library(funModeling)
library(tidyverse)
library(Hmisc)
library(mice)
library(DataExplorer)
library('vegan')
library(cluster)
library(Rtsne)
library(factoextra)

#IMPORT CSV FILE INTO A DATAFRAME
df = read.csv("C:\\Users\\hp\\Desktop\\Thesis\\hcvdat.csv")
glimpse(df)
status(df)

#frequency histogram of missings categorical variables.
freq(df)
#Visual descriptive statistics of the numerical variables
plot_num(df, path_out = "img.jpg")

# Impute missing data
imp <- mice(df, m = 1)
# Store imputed data as new data frame
df_imputed <- complete(imp)

summary(df_imputed)
status(df_imputed)


DataExplorer::create_report(df_imputed)

#Comuputing Distance Matrices
#Let get a dataframe newdf without categorical values and the patient's ID
newdf <- subset(df_imputed, select = -c(X,Category,Sex))
head(newdf)
sum(is.na(newdf))
#computation
Dist_E<- dist(newdf,method="euclidean")
Dist_M<- dist(newdf,method="manhattan")
Dist_Max<-dist(newdf,method="maximum")
Dist_Can<- dist(newdf,method="canberra")
Dist_Mink<- dist(newdf,method="minkowski")

#Heatmap of distance matrices computed without missing values
fviz_dist(Dist_E,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_M,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Max,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Can,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Mink,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Pear,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))


#computation
```

```
Dist_E1<-dist(newdf.scaled,method="euclidean")
head(as.matrix(Dist_M))
Dist_M1<- dist(newdf.scaled,method="manhattan")
Dist_Max1<- dist(newdf.scaled,method="maximum")
Dist_Can1<- dist(newdf.scaled,method="canberra")
Dist_Mink1<- dist(newdf.scaled,method="minkowski")

#Heatmap of distance matrices computed without missing values
fviz_dist(Dist_E1,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_M1,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Max1,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Can1,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Mink1,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))
fviz_dist(Dist_Pear1,gradient=list(low='ivory',mid='cornflowerblue',hight='red'))


#MDMR analysis computation
#computation
DistE<- as.matrix(dist(newdf,method="euclidean"))
DistM<- as.matrix(dist(newdf,method="manhattan"))
DistMax<-as.matrix( dist(newdf,method="maximum"))
DistCan<- as.matrix(dist(newdf,method="canberra"))
DistMink<- as.matrix(dist(newdf,method="minkowski"))

res1 <- mdmr(X = newdf , D = DistE)
res2 <- mdmr(X = newdf , D = DistM)
res3 <- mdmr(X = newdf , D = DistMax)
res4 <- mdmr(X = newdf , D = DistCan)
res5 <- mdmr(X = newdf , D = DistMink)

summary(res1)
summary(res2)
summary(res3)
summary(res4)
summary(res5)

 #Comparing distance matrices
mantel.rtest(Dist_E,Dist_M,nrepet = 99)
mantel.rtest(Dist_E,Dist_Max,nrepet = 99)
mantel.rtest(Dist_E,Dist_Can,nrepet = 99)
mantel.rtest(Dist_E,Dist_Mink,nrepet = 99)
mantel.rtest(Dist_M,Dist_Max,nrepet = 99)
mantel.rtest(Dist_M,Dist_Can,nrepet = 99)
mantel.rtest(Dist_M,Dist_Mink,nrepet = 99)
mantel.rtest(Dist_Max,Dist_Can,nrepet = 99)
mantel.rtest(Dist_Max,Dist_Mink,nrepet = 99)
mantel.rtest(Dist_Can,Dist_Mink,nrepet = 99)


#Converting character to factor
df_factor <-  df_imputed
class(as.factor(df_factor$Sex))
df_factor <- as.data.frame(unclass(df_factor),stringsAsFactors = TRUE)
glimpse(df_factor)

#computing the Gower distance
gower_dist <- daisy(df_factor[,-1],metric = "gower",type = list(logratio = 3))
summary(gower_dist)
```

```
#Computing the Gower distance matrix
gower_mat <- as.matrix(gower_dist)

#Show the most similar pair
df_factor[
  which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
        arr.ind = TRUE)[1, ], ]
#Let show the most dissimilar
df_factor[
  which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
        arr.ind = TRUE)[1, ], ]

# Calculate silhouette width for many k using PAM

sil_width <- c(NA)

for(i in 2:10){

  pam_fit <- pam(gower_dist,
                 diss = TRUE,
                 k = i)

  sil_width[i] <- pam_fit$silinfo$avg.width

}

# Plot silhouette width (higher is better)

plot(1:10, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:10, sil_width)

#cluster Visualization
tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)

tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering),
         name = df_factor$Category)

ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```