IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 01-08
www.iosrjournals.org

A Comparison of Statistical Packages in R Tool to Impute Missing Values

Mrs.D.Suganthi[1], Mr.K.Dheenathayalan[2]
[1](Department of Computer Science, Sankara college of science and commerce, Coimbatore, India)
[2](Department of Computer Science, Kamban college of arts and science, Coimbatore, India)

Abstract: *Data mining has pushed the realm of information technology beyond predictable limits. Missing value is one of the major factor, which can render the obtain result beyond use attained from specific data set by applying data mining technique. There could be numerous reasons for missing values in a data set such as human error, hardware malfunction etc. It is imperative to tackle the labyrinth of missing values before applying any technique of data mining; otherwise, the information extracted from data set containing missing values will lead to the path of wrong decision making. Due to improper handling, the result obtained by the researcher will differ from ones where the missing values are present. Several methods have been, and continue to be, developed to draw inferences from data sets with missing values. In this work, we experimented and results are compared for three methods of imputation of missing values in numerical dataset. We compare MICE (PMM, CART and SAMPLE), HMISC, HOT.DECK, AMELIA, kNN and MISSFOREST, which are likely the most sophisticated imputation methods currently employed for imputing missing values.*
Keywords: *Data mining, distance measure, Imputation, missing values*

I. Introduction

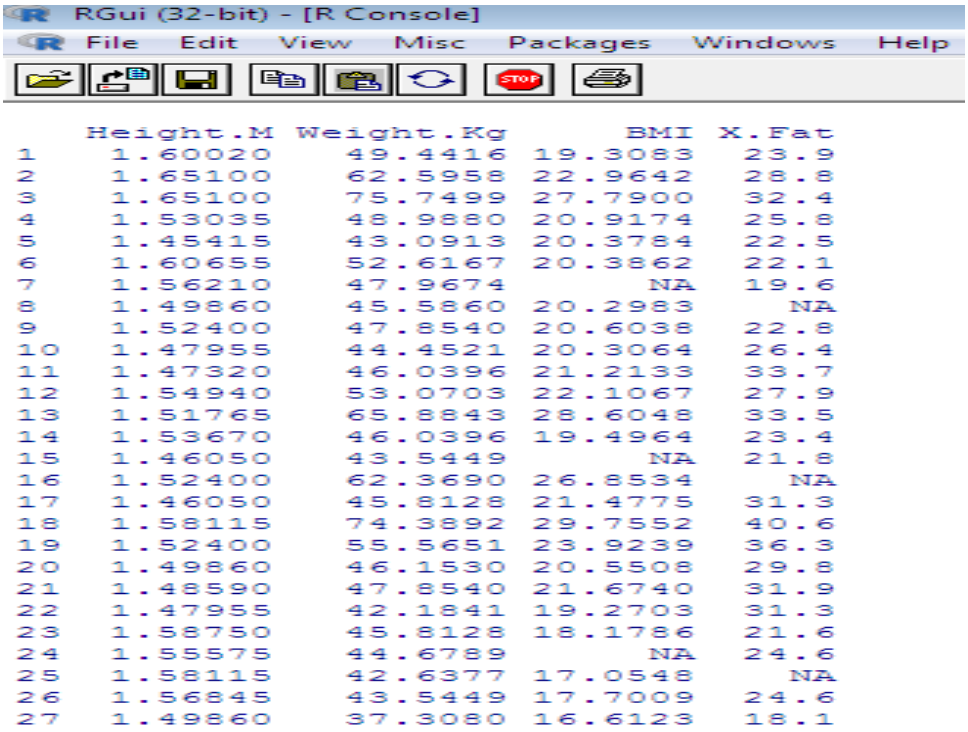
Missing data is one of the issues which are to be fathomed for real-time application. Improper imputation produces predisposition result. For example, manual information entrances system, inaccurate estimations, gear blunders, and numerous others. Hence legitimate consideration is expected to credit the missing values. Research data are prone to missing data. Missed observations may occur due to human error. For example, a researcher may forget to take a measurement such as the patient's pulse. Or, a malfunction which eliminates a measurement [1].

Databases also have missing data. Whenever there is a mismatch of variables between databases, there are missing occurrences. For example, a database analyst is analyzing sales databases consolidated from three regions. If the central region did not record one variable such as the educational background of sales representatives, then this variable would have missing occurrences when the three databases are merged [2]. Missing data can be serious Missing data can be treacherous because it is difficult to identify the problem. Each question or variable may only have a small number of missing responses, but in combination, the missing data could be numerous. Only thorough analysis on missing data can determine whether missing data are problematic [3].

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. The concept of missing values is important to understand in order to successfully manage data. Until now, this analysis has been time consuming and error prone. Missing data can cause serious problems. First, most statistical procedures automatically eliminate cases with missing data. This means that in the end, we may not have enough data to perform the analysis. Second, the analysis might run but the results may not be statistically significant because of the small amount of input data. Third, results may be misleading if the cases you analyze are not a random sample of all cases. Also, it is not always obvious when missing data will cause a problem.

II. Dataset

The example dataset taken is health dataset, which contains Height in meter, Weight in Kg, BMI and Fat percentage of individual person. Totally there are 4 variables with 92 observations. Initially there were no missing values, for testing purpose some values are randomly deleted. After random deletion there are 18 missing values (9 values in BMI and 9 values in Fat percentage). Figure 1 shows the snapshot of the dataset contains missing values.



	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.3083	23.9
2	1.65100	62.5958	22.9642	28.8
3	1.65100	75.7499	27.7900	32.4
4	1.53035	48.9880	20.9174	25.8
5	1.45415	43.0913	20.3784	22.5
6	1.60655	52.6167	20.3862	22.1
7	1.56210	47.9674	NA	19.6
8	1.49860	45.5860	20.2983	NA
9	1.52400	47.8540	20.6038	22.8
10	1.47955	44.4521	20.3064	26.4
11	1.47320	46.0396	21.2133	33.7
12	1.54940	53.0703	22.1067	27.9
13	1.51765	65.8843	28.6048	33.5
14	1.53670	46.0396	19.4964	23.4
15	1.46050	43.5449	NA	21.8
16	1.52400	62.3690	26.8534	NA
17	1.46050	45.8128	21.4775	31.3
18	1.58115	74.3892	29.7552	40.6
19	1.52400	55.5651	23.9239	36.3
20	1.49860	46.1530	20.5508	29.8
21	1.48590	47.8540	21.6740	31.9
22	1.47955	42.1841	19.2703	31.3
23	1.58750	45.8128	18.1786	21.6
24	1.55575	44.6789	NA	24.6
25	1.58115	42.6377	17.0548	NA
26	1.56845	43.5449	17.7009	24.6
27	1.49860	37.3080	16.6123	18.1

Figure1. Dataset with missing values(NA)

In the above figure NA represents missing value (Not Available). This work is carried out using R tool. Figure 2 Represents missing values map in dataset.

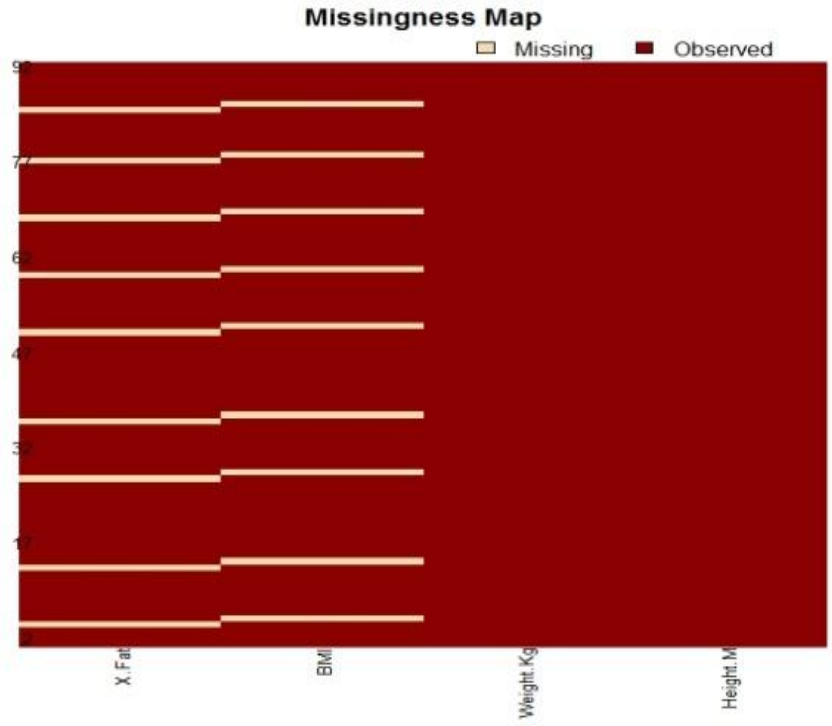


Figure2. Missingness map of the Dataset

The figure shows that there are no missing values in Weight.Kg and Height.M. Variable X.Fat contains 9 missing values and variable BMI also contains 9 missing values. The numbers on left side of the figure represents number of observations, totally 92 observations are there.

III. Experimental Setup

3.1 Packages Used For Imputation

a. MICE

Multiple imputation is the method of choice for complex incomplete data problems.Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). Methods of MICE are PMM (predictive mean matching), CART, Sample [4].

b. Hmisc

Contains many functions useful for data analysis, high-level graphics, utility operations, functions for computing sample size and power, importing and annotating datasets, imputing missing values, advanced table making, variable clustering, character string manipulation, conversion of R objects to LaTeX and html code, and recoding variables.

c. Hot.Deck

Historically, the term "hot deck" comes from the use of computer punch cards for data storage, and refers to the deck of cards for donors available for a non-respondent [5]. The deck was "hot" since it was currently being processed, as opposed to the "cold deck" which refers to using pre-processed data as the donors, i.e. data from a previous data collection or a different data set. Hot deck imputation methods to resolve missing data. Hot deck imputation involves replacing missing values of one or more variables with observed values.

d. Knn Impute

k *Nearest Neighbor* (kNN) is a data mining technique to impute missing expression data in microarray. For each gene with missing values, it finds the *k* nearest neighbors using a *distance metric*, confined to the columns for which that gene is NOT missing. Each candidate neighbor might be missing some of the coordinates used to calculate the distance. In this case we average the distance from the non-missing coordinates. Having found the k nearest neighbors for a gene, it imputes the missing elements by averaging those (non-missing) elements of its neighbors.

e. Amelia

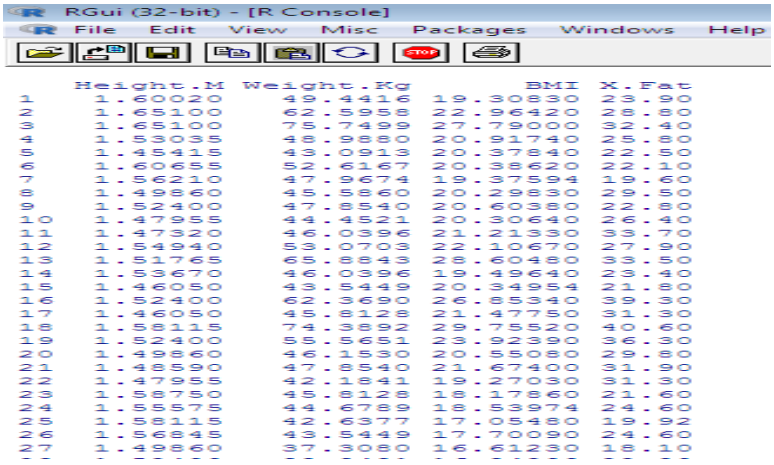
Amelia II performs multiple imputation, a general-purpose approach to data with Missing values. Multiple imputation has been shown to reduce bias and increase efficiency compared to list wise deletion. Multiple imputation involves imputing m values for each missing cell in your data matrix and creating m\completed data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data.

f. Missforest

One of the nice "extras" of the random forest algorithm is its use for mixed data type (numeric and categorical) imputation. Unlike other modeling approaches such as generalized linear models which employ a case-wise deletion, random forests will throw an error when it encounters missing values [6]. Out of the box, the random forest algorithm will perform imputation by leveraging a "proximity matrix" which also allowing for random forests to be used for outlier detection and clusters. As each tree in the RF is built, all cases are "run down" the tree and if two cases (i and j) end up in the same terminal node – they are considered to be relatively alike according to the model. A NxN matrix is ultimately produced with element (i,j) incremented each time case i and j end up in the same node. This is normalized for the number of trees in the forest.

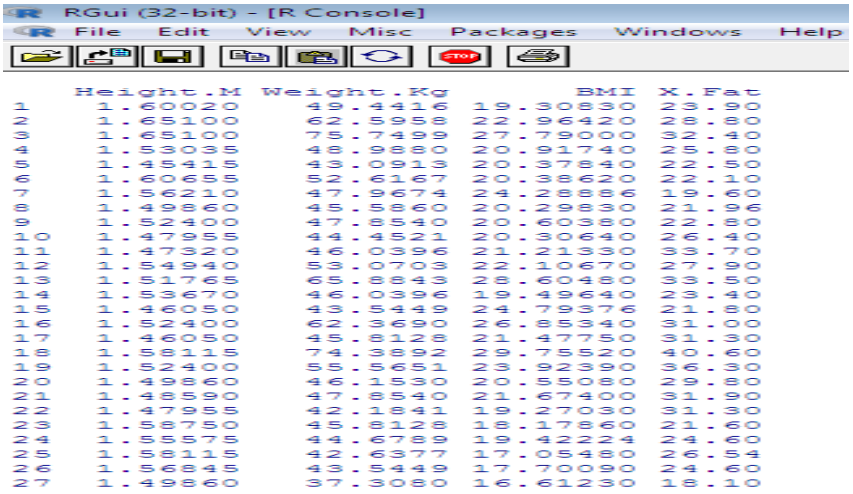
IV. Results And Discussion

a. Mice



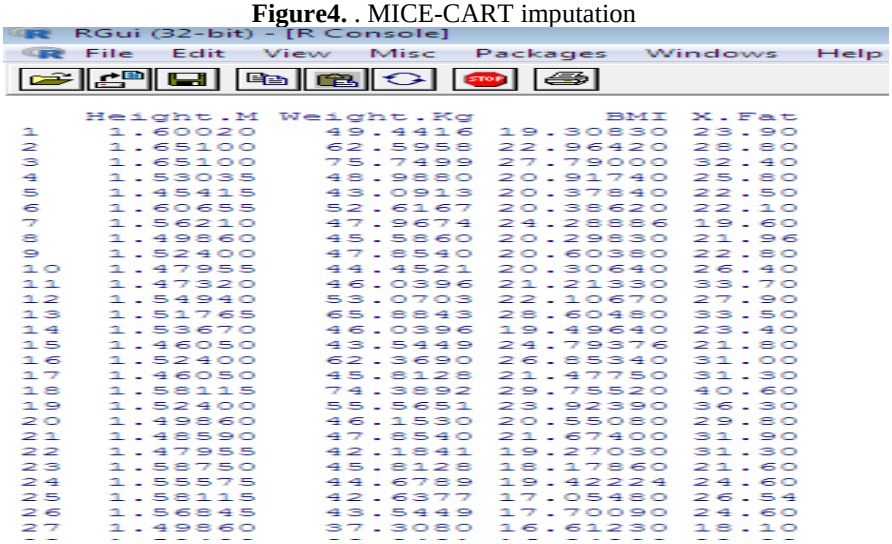
	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.30830	23.90
2	1.65100	62.5958	22.96420	28.80
3	1.65100	75.7499	27.79000	32.40
4	1.53035	48.9880	20.91740	25.80
5	1.45415	43.0913	20.37840	22.50
6	1.60655	52.6167	20.38620	22.10
7	1.56210	47.9674	19.37594	19.60
8	1.49860	45.5860	20.29830	29.50
9	1.52400	47.8540	20.60380	22.80
10	1.47955	44.4521	20.30640	26.40
11	1.47320	46.0396	21.21330	33.70
12	1.54940	53.0703	22.10670	27.90
13	1.51765	65.8843	28.60480	33.50
14	1.53670	46.0396	19.49640	23.40
15	1.46050	43.5449	20.34954	21.80
16	1.52400	62.3690	26.85340	39.30
17	1.46050	45.8128	21.47750	31.30
18	1.58115	74.3892	29.75520	40.60
19	1.52400	55.5651	23.92390	36.30
20	1.49860	46.1530	20.55080	29.80
21	1.48590	47.8540	21.67400	31.90
22	1.47955	42.1841	19.27030	31.30
23	1.58750	45.8128	18.17860	21.60
24	1.55575	44.6789	18.53974	24.60
25	1.58115	42.6377	17.05480	19.92
26	1.56845	43.5449	17.70090	24.60
27	1.49860	37.3080	16.61230	18.10

Figure3. MICE-PMM imputation



	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.30830	23.90
2	1.65100	62.5958	22.96420	28.80
3	1.65100	75.7499	27.79000	32.40
4	1.53035	48.9880	20.91740	25.80
5	1.45415	43.0913	20.37840	22.50
6	1.60655	52.6167	20.38620	22.10
7	1.56210	47.9674	24.28886	19.60
8	1.49860	45.5860	20.29830	21.96
9	1.52400	47.8540	20.60380	22.80
10	1.47955	44.4521	20.30640	26.40
11	1.47320	46.0396	21.21330	33.70
12	1.54940	53.0703	22.10670	27.90
13	1.51765	65.8843	28.60480	33.50
14	1.53670	46.0396	19.49640	23.40
15	1.46050	43.5449	24.79376	21.80
16	1.52400	62.3690	26.85340	31.00
17	1.46050	45.8128	21.47750	31.30
18	1.58115	74.3892	29.75520	40.60
19	1.52400	55.5651	23.92390	36.30
20	1.49860	46.1530	20.55080	29.80
21	1.48590	47.8540	21.67400	31.90
22	1.47955	42.1841	19.27030	31.30
23	1.58750	45.8128	18.17860	21.60
24	1.55575	44.6789	19.42224	24.60
25	1.58115	42.6377	17.05480	26.54
26	1.56845	43.5449	17.70090	24.60
27	1.49860	37.3080	16.61230	18.10

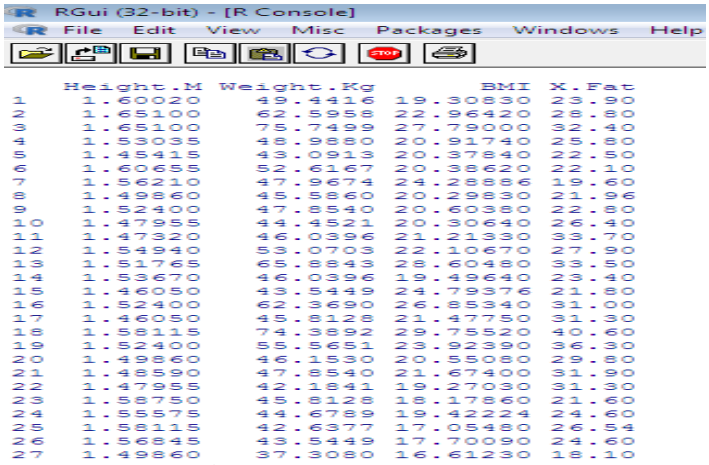
Figure4. . MICE-CART imputation



	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.30830	23.90
2	1.65100	62.5958	22.96420	28.80
3	1.65100	75.7499	27.79000	32.40
4	1.53035	48.9880	20.91740	25.80
5	1.45415	43.0913	20.37840	22.50
6	1.60655	52.6167	20.38620	22.10
7	1.56210	47.9674	24.28886	19.60
8	1.49860	45.5860	20.29830	21.96
9	1.52400	47.8540	20.60380	22.80
10	1.47955	44.4521	20.30640	26.40
11	1.47320	46.0396	21.21330	33.70
12	1.54940	53.0703	22.10670	27.90
13	1.51765	65.8843	28.60480	33.50
14	1.53670	46.0396	19.49640	23.40
15	1.46050	43.5449	24.79376	21.80
16	1.52400	62.3690	26.85340	31.00
17	1.46050	45.8128	21.47750	31.30
18	1.58115	74.3892	29.75520	40.60
19	1.52400	55.5651	23.92390	36.30
20	1.49860	46.1530	20.55080	29.80
21	1.48590	47.8540	21.67400	31.90
22	1.47955	42.1841	19.27030	31.30
23	1.58750	45.8128	18.17860	21.60
24	1.55575	44.6789	19.42224	24.60
25	1.58115	42.6377	17.05480	26.54
26	1.56845	43.5449	17.70090	24.60
27	1.49860	37.3080	16.61230	18.10

Figure5. MICE-Sample imputation

b. Hmisc

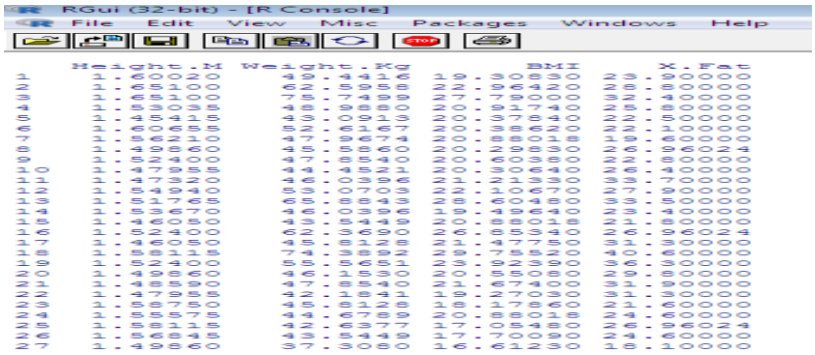


	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.30830	23.90
2	1.65100	62.5958	22.96420	28.80
3	1.65100	75.7499	27.79000	32.40
4	1.53035	48.9880	20.91740	25.80
5	1.45415	43.0913	20.37840	22.50
6	1.60655	52.6167	20.38620	22.10
7	1.56210	47.9674	24.28886	19.60
8	1.49860	45.5860	20.29830	21.96
9	1.52400	47.8540	20.60380	22.80
10	1.47955	44.4521	20.30640	26.40
11	1.47320	46.0396	21.21330	33.70
12	1.54940	53.0703	22.10670	27.90
13	1.51765	65.8843	28.60480	33.50
14	1.53670	46.0396	19.49640	23.40
15	1.46050	43.5449	24.79376	21.80
16	1.52400	62.3690	26.85340	31.00
17	1.46050	45.8128	21.47750	31.30
18	1.58115	74.3892	29.75520	40.60
19	1.52400	55.5651	23.92390	36.30
20	1.49860	46.1530	20.55080	29.80
21	1.48590	47.8540	21.67400	31.90
22	1.47955	42.1841	19.27030	31.30
23	1.58750	45.8128	18.17860	21.60
24	1.55575	44.6789	19.42224	24.60
25	1.58115	42.6377	17.05480	26.54
26	1.56845	43.5449	17.70090	24.60
27	1.49860	37.3080	16.61230	18.10

Figure: 6 HMISC imputation

Figure 6 Shows result of HMISC method after imputing the missing values in the health dataset

c. Hot.Deck



	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.30830	23.900000
2	1.65100	62.5958	22.96420	28.800000
3	1.65100	75.7499	27.79000	32.400000
4	1.53035	48.9880	20.91740	25.800000
5	1.45415	43.0913	20.37840	22.500000
6	1.60655	52.6167	20.38620	22.100000
7	1.56210	47.9674	24.28886	19.600000
8	1.49860	45.5860	20.29830	26.960224
9	1.52400	47.8540	20.60380	22.800000
10	1.47955	44.4521	20.30640	26.400000
11	1.47320	46.0396	21.21330	33.700000
12	1.54940	53.0703	22.10670	27.900000
13	1.51765	65.8843	28.60480	33.500000
14	1.53670	46.0396	19.49640	23.400000
15	1.46050	43.5449	20.88018	21.800000
16	1.52400	62.3690	26.85340	26.960224
17	1.46050	45.8128	21.47750	31.300000
18	1.58115	74.3892	29.75520	40.600000
19	1.52400	55.5651	23.92390	36.300000
20	1.49860	46.1530	20.55080	29.800000
21	1.48590	47.8540	21.67400	31.900000
22	1.47955	42.1841	19.27030	31.300000
23	1.58750	45.8128	18.17860	21.600000
24	1.55575	44.6789	20.88018	24.600000
25	1.58115	42.6377	17.05480	26.960224
26	1.56845	43.5449	17.70090	24.600000
27	1.49860	37.3080	16.61230	18.100000

Figure: 7 Hot.Deck imputation

Figure 7 Shows result of Hot.Deck method after imputing the missing values in the health dataset

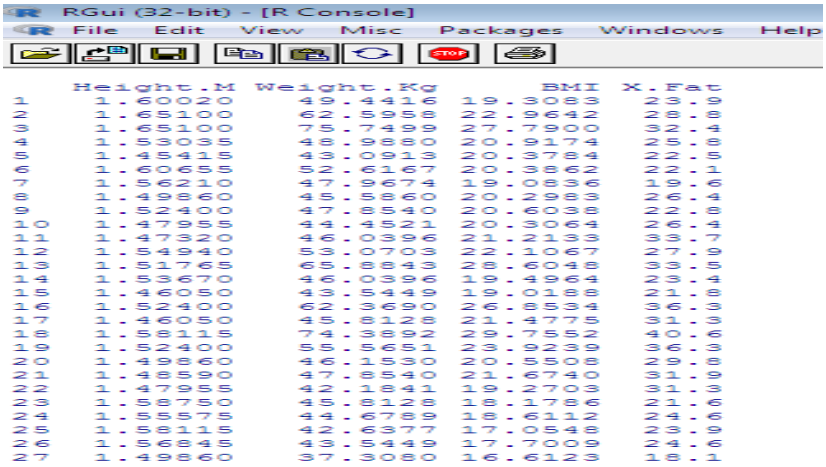
d. Knn Impute

	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.3083	23.9
2	1.65100	62.5958	22.9642	28.8
3	1.65100	75.7499	27.7900	32.4
4	1.53035	48.9880	20.9174	25.8
5	1.45415	43.0913	20.3784	22.5
6	1.60655	52.6167	20.3862	22.1
7	1.56210	47.9674	19.0836	19.6
8	1.49860	45.5860	20.2983	26.4
9	1.52400	47.8540	20.6038	22.8
10	1.47955	44.4521	20.3064	26.4
11	1.47320	46.0396	21.2133	33.7
12	1.54940	53.0703	22.1067	27.9
13	1.51765	65.8843	28.6048	33.5
14	1.53670	46.0396	19.4964	23.4
15	1.46050	43.5449	19.0188	21.8
16	1.52400	62.3690	26.8534	36.3
17	1.46050	45.8128	21.4775	31.3
18	1.58115	74.3892	29.7552	40.6
19	1.52400	55.5651	23.9239	36.3
20	1.49860	46.1530	20.5508	29.8
21	1.48590	47.8540	21.6740	31.9
22	1.47955	42.1841	19.2703	31.3
23	1.58750	45.8128	18.1786	21.6
24	1.55575	44.6789	18.6112	24.6
25	1.58115	42.6377	17.0548	23.9
26	1.56845	43.5449	17.7009	24.6
27	1.49860	37.3080	16.6123	18.1

Figure: 8 kNN imputation

Figure 8 Shows result of kNN-impute method after imputing the missing values in the health dataset

e. Amelia

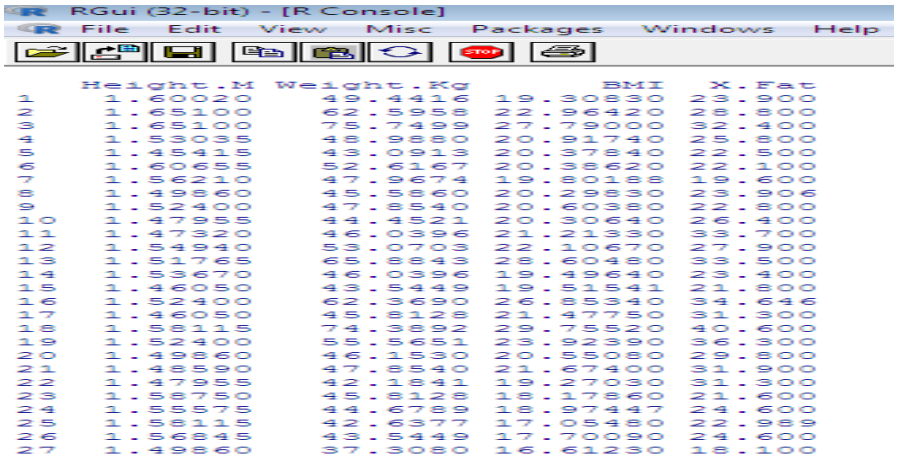


	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.3083	23.9
2	1.65100	62.5958	22.9642	28.8
3	1.65100	75.7499	27.7900	32.4
4	1.53035	48.9880	20.9174	25.8
5	1.45415	43.0913	20.3784	22.5
6	1.60655	52.6167	20.3862	22.1
7	1.56210	47.9674	19.0836	19.6
8	1.49860	45.5860	20.2983	26.4
9	1.52400	47.8540	20.6038	22.8
10	1.47955	44.4521	20.3064	26.4
11	1.47320	46.0396	21.2133	33.7
12	1.54940	53.0703	22.1067	27.9
13	1.51765	65.8843	28.6048	33.5
14	1.53670	46.0396	19.4964	23.4
15	1.46050	43.5449	19.0188	21.8
16	1.52400	62.3690	26.8534	36.3
17	1.46050	45.8128	21.4775	31.3
18	1.58115	74.3892	29.7552	40.6
19	1.52400	55.5651	23.9239	36.3
20	1.49860	46.1530	20.5508	29.8
21	1.48590	47.8540	21.6740	31.9
22	1.47955	42.1841	19.2703	31.3
23	1.58750	45.8128	18.1786	21.6
24	1.55575	44.6789	18.6112	24.6
25	1.58115	42.6377	17.0548	23.9
26	1.56845	43.5449	17.7009	24.6
27	1.49860	37.3080	16.6123	18.1

Figure: 9 Amelia imputation

Figure 9 Shows result of Amelia method after imputing the missing values in the health dataset

f. Missforest



	Height.M	Weight.Kg	BMI	X.Fat
1	1.60020	49.4416	19.30830	23.900
2	1.65100	62.5958	22.96420	28.800
3	1.65100	75.7499	27.79000	32.400
4	1.53035	48.9880	20.91740	25.800
5	1.45415	43.0913	20.37840	22.500
6	1.60655	52.6167	20.38620	22.100
7	1.56210	47.9674	19.80188	19.600
8	1.49860	45.5860	20.29830	23.906
9	1.52400	47.8540	20.60380	22.800
10	1.47955	44.4521	20.30640	26.400
11	1.47320	46.0396	21.21330	33.700
12	1.54940	53.0703	22.10670	27.900
13	1.51765	65.8843	28.60480	33.500
14	1.53670	46.0396	19.49640	23.400
15	1.46050	43.5449	19.51541	21.800
16	1.52400	62.3690	26.85340	34.646
17	1.46050	45.8128	21.47750	31.300
18	1.58115	74.3892	29.75520	40.600
19	1.52400	55.5651	23.92390	36.300
20	1.49860	46.1530	20.55080	29.800
21	1.48590	47.8540	21.67400	31.900
22	1.47955	42.1841	19.27030	31.300
23	1.58750	45.8128	18.17860	21.600
24	1.55575	44.6789	18.97447	24.600
25	1.58115	42.6377	17.05480	22.989
26	1.56845	43.5449	17.70090	24.600
27	1.49860	37.3080	16.61230	18.100

Figure: 10 MissForest imputation

Figure 10 Shows result of MissForest method after imputing the missing values in the health dataset
The mentioned methods are used to impute the missing values in the given health dataset. All of the missing values are imputed and replaced in their appropriate position.

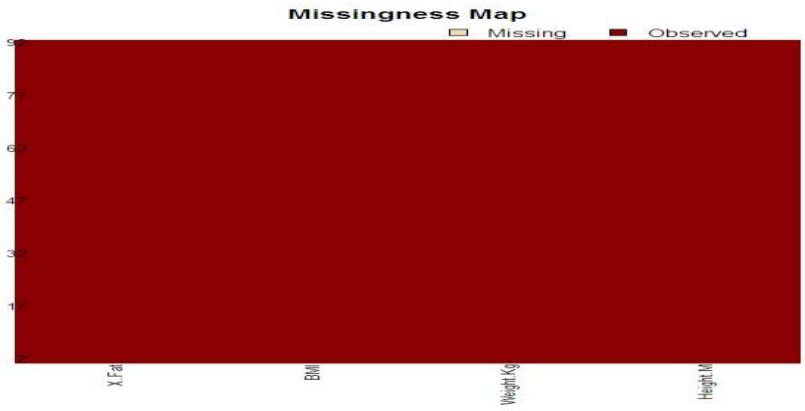


Figure: 11

Figure 11 shows the missingness map of the health dataset after imputing the missing values
From the figure 11, it is observed that there is no more missing values after imputation.

V. Discussion

a. Distance measure-Euclidean distance method

The results from all methods are analyzed using distance measure (Euclidean distance method). Here the method Amelia has the minimum distance between the original values and imputed values (Figure12 and Figure13). Detailed After the analyses it is come to know that Amelia is the best method to impute missing values in numerical dataset.

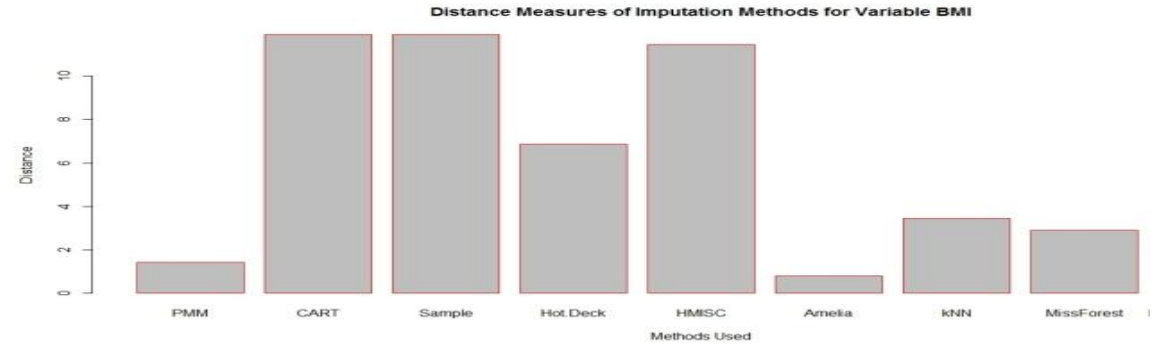


Figure: 12

Figure 12 Represents bar chart of distance measures of imputation methods for variable BMI in health dataset

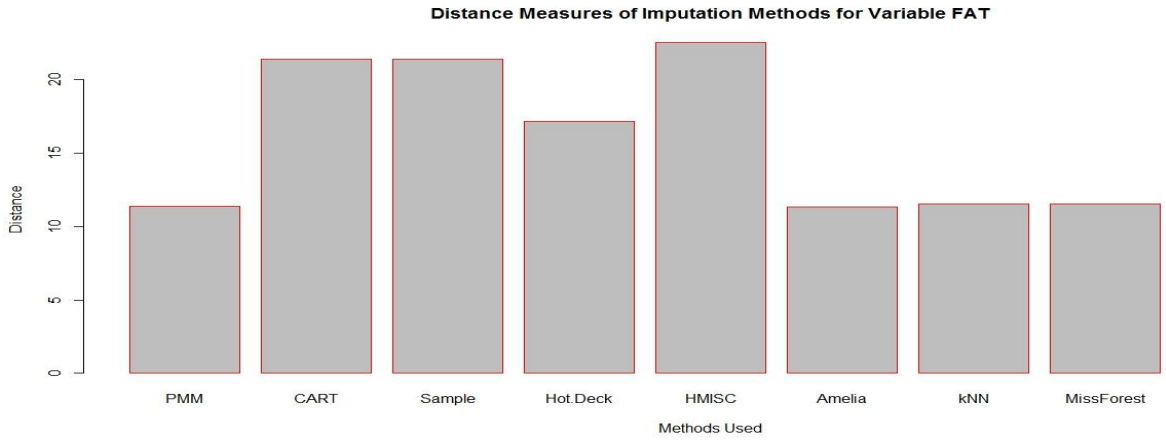


Figure: 13

Figure 13 Represents bar chart of distance measures of imputation methods for variable FAT in health dataset.

VI. Conclusion And Future Work

Incomplete data comes (missing values) from NA data value when collected Different consideration between the times when the data was collected and when it is analyzed. Human / hardware / software problems. Missing values can cause major risks, so it is necessary to impute the missing values. There are many statistical methods and tools available for imputation.

Among only few familiar methods are selected and tested in R using health dataset. From the analyses made it shows Amelia is the best method to impute missing values. In future this work can be extended by using other statistical methods and large datasets.

References

[1] Horton NJ, Kleinman KP (2007) *Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models*. The American Statistician 61: 79-90.

[2] Saunders JA, Morrow-Howell N, Spitznagel E,Dor P, Proctor EK, et al. (2006) *Imputing missing data: A comparison of methods for social work researchers*. Social Work Research 30: 19-31.

[3] Luengo J, GarcaS, Herrera F (2012) *On the choice of the best imputation methods for missing values considering three groups of classification methods*. Knowledge and Information Systems 32:77-108

[4] Brock G, Shaffer J,Blakesley R,Lotz M, Tseng G (2008) *Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes*. BMC Bioinformatics 9: 1-12.

[5] Celton M, MalpertuyA, Lelandais G, Brevern A (2010) *Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments*. BMC Genomics 11: 1-16.

[6] Ibrahim JG,Chen MH, LipsitzSR, HerringAH (2005) *Missing-data methods for generalized linear models. Journal of the American Statistical Association* 100: 332-346.