

Review Article

# Single and Multiple Imputation Techniques to Treat Missing Numerical Variables (MNV) in Perspectives of Data Science Project - A Case Study

Dharmendra Patel<sup>1</sup>, Octavio Loyola-González<sup>2</sup>, Arpit Trivedi<sup>3</sup>, Hardik Rajgor<sup>4</sup>, Tushar Mehta<sup>5</sup>,  
Sanskriti Patel<sup>6</sup>, Pranav Vyas<sup>7</sup>, Nilay Ganatra<sup>8</sup>, Hardik I Patel<sup>9</sup>

<sup>1,3,4,5,6,7,8,9</sup>Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, CHARUSAT, Changa, Gujarat, India.

<sup>2</sup>Artificial Intelligence Manager, Altair Management Consultants Corp., USA.

<sup>1</sup>dharmendrapatel.mca@charusat.ac.in

Received: 24 February 2022

Revised: 07 April 2022

Accepted: 24 April 2022

Published: 19 May 2022

**Abstract** - Data Science is extensively used in various industrial domains to understand the enormous amount of data and derive meaningful and valuable insights to make smarter business decisions. The quality of data plays a vital role in insights generations. Data quality can be enhanced by imputing appropriate values in place of missing data. Data imputation plays a critical role in a data science project. In this paper, we have described single and multiple imputation techniques in the context of missing numerical variables with proper cases. We have explained different scenarios to select appropriate imputation techniques for any data science project. We also produce results based on imputation techniques by taking simple and meaningful examples.

**Keywords** - Single Imputation, Data Science, Numerical Variables, Missing Completely at Random(MCAR), Regression, Multiple Imputation.

## 1. Introduction

The Data Science project works with abundant data from multiple sources. The common observation from almost all projects is missing values in dataset. There is number of reasons for missing values, such as undefined values, human error, incorrect implementation of joins while integrating multiple data sources etc. Training Data Science model comprises significant missing values that may profoundly influence the quality of the model. Hence, the treatment of missing values in the appropriate direction is vital. Data is mainly of two categories: numerical and categorical. In any dataset, majority of variables are either numerical or categorical. If, as Data scientists, we are able to know the different missing data imputation techniques [1] in the context of numerical, categorical or mixture of both kinds of variables, then our job becomes easy to treat them efficiently. This proposed research aims to describe missing data imputation techniques for numerical, categorical, and variables with appropriate scenarios.

There are two types of variables: numerical and categorical. The figure-1 depicts the bifurcation of types of variables.



Fig. 1 Bifurcation of types of variables

A numerical variable contains information that is measurable and represented in numbers. It can be continuous or discrete. The continuous numerical variable represents a measurement, whereas the discrete numerical variable can't be measured but can be counted. The continuous numerical variable is further divided into interval and ratio data. Interval values represent ordered units with the same differences. The main problem with Interval data is they do not have true zero values; therefore, numerous descriptive and inferential statistics can't be applied. The other thing is, we can apply only addition and subtraction to Interval data. Ratio data is the same as Interval, with the difference that they have zero value.



The categorical variable represents the characteristics of the data. They can also take numerical values but do not have mathematical meaning. Categorical variables are divided into three main categories: Nominal, Ordinal and Binary. Nominal data is used to label variables without any intrinsic order among labels. Ordinal data is the same as nominal data with intrinsic order among labels. Binary data has only two values.

Datatypes of variables play an important role in attributing numerical or categorical data [2]. Several statistical methods can only be used with specific kinds of data. This paper will deal with single and multiple imputing techniques for numerical data with their pros and cons. Section II will describe the Mean or Median imputation technique with suitable examples. The section-II will describe the regression imputation technique with a suitable case study. This section also describes the comparison with the previous technique. The section-IV will use a suitable case study to describe multiple imputation techniques with a well-known MICE algorithm.

## 2. Mean or Median Imputation (Single Imputation)

There are two main techniques to impute single numerical values in literature [3] [4] [5]: Mean or Median imputation and regression imputation technique.

Mean and median imputation [6] is only suitable for continuous and discrete numerical variables. This technique replaces all missing values (NA) occurrences by mean or median. The technique is based on certain assumptions:

- It is applicable when data is missing [7] [8]. MCAR means the causes of missing data are unrelated to the data.
- It gives a better result when very fewer data (usually less than 5%) of the variable contains missing data.
- When data is normally distributed, means not have any skewness, then mean imputation is better than median imputation. If the data exhibits some skewness, then the median would be better.

Mean imputation applies to classes or categories and can be expressed as  $M^{\wedge}$ .

$$M^{\wedge} = \sum_{i=1}^N (xi) / N \text{ -----(1)}$$

Where xi is  $x_1, x_2, \dots, x_i$  observations of the dataset without missing values

N is the total number of observations excluding missing values. It is a very simple and effective technique when data is Missing Completely at Random (MCAR) [9] [10]. Figure-2 describes the mean imputation on price data. There are 11 dealers for particular product X. Several dealers have not

disclosed their price on the product. The mean imputation technique is best suited for this kind of situation to treat missing values.

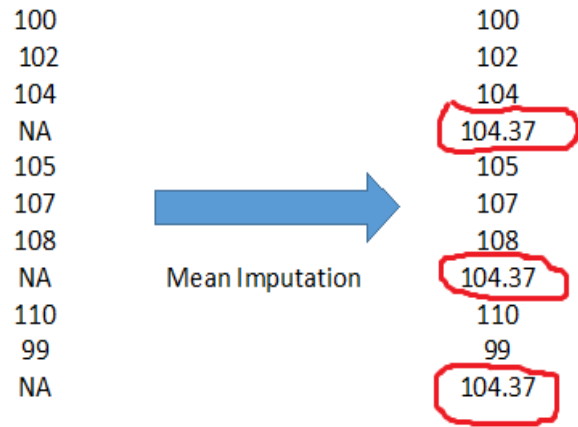


Fig. 2 Mean Imputation Technique

It is very easy and fast. However, the main limitation is, it does not consider the correlation among features. Figure-3 depicts the limitation of the mean imputation technique in connection with correlation.



Fig. 3 Correlation Problem in Mean Imputation

The feature's age and fitness scores are correlated with each other. The mean imputation does the same treatment for all age groups. The figure showed that a 5.75 fitness score is imputed for ages 40 and 60. The other limitation is it reduces the variance value. The table-1 describes the variance value of original and imputed data as per the previous example. The variance value of imputed data is almost half that of original data. A smaller variance leads to narrower confidence in the probability distribution, and as a result, bias is introduced in the model.

**Table 1. Variance of Original Vs Imputed Data in Mean Imputation**

	Original	Imputed
Fitness Score	2.7857	1.7727

The other critical drawback of this technique is; it treats all values differently. It does not consider correlation among variables while imputing values. Table-2 describes the correlation values of before and after mean imputation.

**Table 2. correlation before and after mean imputation**

Correlation Between	Correlation Before Imputation	Correlation After Imputation
Age and Fitness Score	-0.9817252	-0.9035089

The figure in table-2 indicates that the mean imputation does not consider correlation among variables. After imputation correlation between age and fitness score has reduced a lot.

The median imputation can be used in place of mean imputation on numerical data. There are various advantages of median over mean. The prominent advantage is, it is robust against outliers [9] [10]. The other one is, the median gives a more appropriate idea of data distribution when data are skewed. [11].

**3. Regression Imputation (Single Imputation)**

Regression imputation [14] [15] is the opposite of mean/median imputation. It considers correlation among features and predicts observed value based on other correlated variables [12, 13]. The equation of the simple linear regression model is depicted in equation-2.

$$Y=A+BX \text{ -----(2)}$$

Where Y is the dependent variable,  
 X is the independent variable,  
 A is the Y-intercept,  
 B is the slope.

When dealing with the number of independent variables, let's say  $X_i=1, 2, n$ , then a certain deviation may occur in the result and  $\epsilon$  denotes it. The modified version of equation-2 is depicted in equation-3.

$$Y_i = A + BX_i + \epsilon \text{ -----(3)}$$

Where  $X_i$  is dependent variables  $X_1, X_2, \dots, X_n$   
 $\epsilon$  is the deviation or error in the result

Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\hat{Y}_i = \hat{A} + \hat{B}X_i \text{ -----(4)}$$

The residual values are the difference between the

predicted and actual values. The important method is SSR which obtains parameter estimates by minimizing the sum of squared residuals [14].

$$\text{Sum of Square Residual(SSR)} = \sum_{i=1}^n e_i^2 \text{ -----(5)}$$

Where  $e_i$  is the residual value.

In the case of simple regression, equations-6 and 7 are used as least square estimates

$$\hat{B} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ -----(6)}$$

Where  $\bar{x}$  and  $\bar{y}$  are mean values.

$$\hat{A} = \bar{y} - \hat{B}\bar{x} \text{ -----(7)}$$

The estimate of variance depicts in equation-8. It is based on the assumption that the population error term has constant variance.

$$\hat{\sigma}_\epsilon^2 = (SSR) / n - 2 \text{ -----(8)}$$

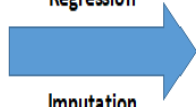
The equation-8 is termed the regression's mean square error(MSE).

The standard errors of the parameter estimates are depicted in equation-9 and 10.

$$\hat{\sigma}_B = \hat{\sigma}_\epsilon \sqrt{1 / \sum (x_i - \bar{x})^2} \text{ -----(9)}$$

$$\hat{\sigma}_A = \hat{\sigma}_B \sqrt{\sum (x_i^2) / n} \text{ -----(10)}$$

The regression imputation technique is complex compared to the mean/median; however, it predicts the value based on correlation more effectively than the mean/median imputation. The figure-4 describes the regression imputation on the same data applied for mean/median imputation.

Age	Fitness Score		Age	Fitness Score
20	8	Regression  Imputation	20	8
25	7		25	7
55	NA		55	4.99
60	5		60	5
70	4		70	4
80	3		80	3
30	7		30	7
35	6		35	6
38	6		38	6
40	NA		40	6.09
57	NA		57	4.84
60	NA		60	4.62

**Fig. 4 Regression Imputation**

The mean/median imputation does not consider correlation among fields while imputing values. It imputes the same values for all missing elements. On the other hand, the regression imputation imputes values based on correlation. The regression imputation value is 4.99 and 6.09 for 55 and 40, respectively. The table-3 depicts the correlation values between age and fitness score before and after regression imputation.

**Table. 3 Correlation before and after regression imputation**

Correlation Between	Correlation Before Imputation	Correlation After Imputation
Age and Fitness Score	-0.9817252	-0.979475

The figure in table-3 indicates that there is not much difference between correlation values before and after regression imputation. Regression Imputation preserves the relationship between variables while imputing missing values. However, this technique might lead to biased parameters estimates. It generates biased parameters, especially for MNAR and MAR missing values. [15].

**4. Multiple Imputation Techniques**

Single imputation techniques like mean/median and regression lead to biased parameter estimates when the missing data is systematically related to the observed and unobserved data, i.e. MAR and MNAR. Regression imputation is effective in handling correlation compared to mean/median imputation; however, the error of this technique is not incorporated. Some data points may deviate from the regression line, causing error variance. Another important limitation of single imputation is that it does not represent uncertainty allied with missing values [16].

Multiple imputation solves the limitations of single imputation techniques ([17] [18] [19]. MICE is the well-known algorithm to deal with missing values in this category [20] [21]. The MICE algorithm derives the posterior distribution of  $\theta$  from conditional distributions of the form described in equation-11.

$$P(X_q | X_{-1, \dots, -q}, \theta_1, \dots, \theta_q) \text{ -----(11)}$$

Where  $\theta_1, \dots, \theta_q$  are specific conditional densities  
 The  $j^{th}$  iteration of the chained equation that successively draws as per equation-12 and 13.

$$\theta_1, \dots, \theta_q^{*(j)} \sim P(\theta_1, \dots, \theta_q | X_1, \dots, X_q^{obs}, X_2^{(j-1)}, \dots, X_{q-1}^{(j-1)}) \text{ ----(12)}$$

$$X_1, \dots, X_q^{*(j)} \sim P(X_1, \dots, X_q | X_1, \dots, X_q^{obs}, X_2^{(j)}, \dots, X_q^{(j)}, \theta_1, \dots, \theta_q^{*(j)}) \text{ ---(13)}$$

Where  $X_i^{(j)} = (X_i^{obs}, X_i^{*(j)})$  is the  $i^{th}$  imputed variable at iteration  $j$ . The previous imputation  $X_i^{*(j-1)}$  only enter  $X_i^{*(j)}$ s through its relation with other variables.

Let's say there is one dataset of the health record of people. There are 20 observations and 5 variables. The dataset is unprocessed as it contains several missing values. The statistics of missing values of the dataset are depicted in Table-4.

**Table. 4 Missing values statistics in the Fitness dataset**

Variable Name	Total Missing Values	Observations number having missing values
Age	0	-
Sex	4	5,11,14,18
Score	6	3,6,8,12,17,19
Systolic(BP)	8	3,5,8,10,12,13,17,19
Diastolic(BP)	9	2,6,7,10,11,14,16,18,20

Table-4 indicates that only the Age variable contains all values. All other variables have missing values. All variables are correlated with each other. The multiple imputation techniques play a vital role in it as the prediction of only a single value is misleading. MICE algorithm generates multiple imputation values for all missing values based on appropriate machine learning algorithms. In our example, different machine learning methods are applied for different variables. Sex is categorical data and only needs binary classification, so Logistic Regression(LR) is selected. Score and two types of blood pressure are numerical, so Prediction by Partial Matching (PPM) would be appropriate for them. For the experimental study, five imputation values are generated for all variables. Table-5 describes the sex imputation values for all missing observations.

**Table. 5 Imputation values for Sex variable missing observations**

Missing Observation	Imputation Value-1	Imputation Value-2	Imputation Value-3	Imputation Value-4	Imputation Value-5
5	M	F	F	M	F
11	F	F	F	F	M
14	M	M	F	F	M
18	M	M	F	M	F

The multiple imputation techniques generate multiple imputation values based on the algorithm. Now data scientist has to decide the appropriate value for the imputation. As per the above table values, F is suitable for the 5<sup>th</sup> observation as 60% of prediction values have F. Similarly, F, M, M are suitable for observations 11,14 and 18, respectively. The most suitable values for imputations are F, F, M, M, which is the second value.

Table-6, Table-7 and Table-8 describe the imputation values based on Score, Systolic(BP) and Diastolic(BP), respectively.

**Table. 6 Imputation values for Score variable missing observations**

Missing Observation	Imputation Value-1	Imputation Value-2	Imputation Value-3	Imputation Value-4	Imputation Value-5
3	7	7	7	7	8
6	3	3	3	3	3
8	5	5	5	5	5
12	7	7	8	7	7
17	5	5	5	5	5
19	7	7	8	7	7

Imputation value-1,3 and 4 are suitable for this context.

**Table. 7 Imputation values for Systolic(BP) variable missing observations**

Missing Observation	Imputation Value-1	Imputation Value-2	Imputation Value-3	Imputation Value-4	Imputation Value-5
3	110	118	118	110	118
5	125	120	125	118	135
8	120	125	120	110	115
10	125	135	135	152	155
12	115	110	112	112	110
13	118	118	110	112	115
17	135	125	118	120	120
19	118	112	110	115	118

The values generated by the algorithm for this variable are not close to each other, so difficult to decide which imputation value is suitable based on the variable. In such a case, calculate the average of original values after eliminating missing values. The mean value of original data is 127.75. Now, calculate the mean values of all imputation

values. Decide the imputation value, which means the value is closer to the mean value of the original values. The mean values of imputation values 1,2,3,4 and 5 are 120.75, 120.37,118.5,118.62,123.8 respectively. The value 123.8 is close to 127.75, so decide the fifth value for the imputation.

**Table. 8 Imputation values for Diastolic(BP) variable missing observations**

Missing Observation	Imputation Value-1	Imputation Value-2	Imputation Value-3	Imputation Value-4	Imputation Value-5
2	76	76	79	75	72
6	80	85	92	80	80
7	76	76	76	79	75
10	85	79	85	82	92
11	85	75	75	80	85
14	75	79	75	92	75
16	82	79	80	85	75
18	76	75	76	76	76
20	92	82	82	80	75

The imputation value-2 is closer to the average of the original values. After eliminating missing values, the average of original values is 78.91, and the average of imputation value-2 is 78.44. Imputation value-2 is the right choice for the imputation.

### 5. Conclusion

Missing values treatment is vital for the success of any data science project. There are two main ways to treat missing values, i.e. eliminate rows having missing values or impute with proper values. The elimination technique is not suitable for any real data science project as it reduces the

data. Imputation with proper value makes sense. However, selecting appropriate techniques is most challenging for the data scientist. Many data science projects fail due to improper imputation. In this paper, we have systematically discussed imputation techniques with proper examples. We also justified the selection of the technique. We have discussed techniques by keeping missing numerical values in mind as numerical values are vital for computation. We have discussed single value and multiple values imputation techniques by providing suitable cases. We have concluded that the imputation process is vital for data science projects, and no value is attributed without understanding the nature of

data types and applications. The appropriate imputation technique is crucial for the success of any data science project. Simple imputation techniques are straightforward, but selection of proper techniques is very important. The

multiple imputation techniques are complex, and after generating multiple values, which value is suitable needs proper technique and justification.

## REFERENCES

- [1] M. C. P. A. J. A. I. G. C. De Souto, Impact of Missing Data Imputation Methods on Gene Expression Clustering and Classification, *BMC Bioinformatics*. (2015) 1-9.
- [2] A. R. T. V. D. H. G. J. S. T. & M. K. G. Donders, Review: A Gentle Introduction to Imputation of Missing Values, *Journal of Clinical Epidemiology*. (2006) 1087–1091.
- [3] C. P. E. S. A. E. Graham John W, *Methods for Handling Missing Data*, Wiley. (2012).
- [4] K. J. H. Kwak Sang Kyu, Statistical Data Preparation: Management of Missing Values and Outliers. *Korean J Anesthesiol*. (2017) 407-411.
- [5] G.-B. W. J. Grzymala-Busse Jerzy W, *Handling Missing Attribute Values*, Berlin: Springer. (2009).
- [6] L. A. C. J. Peyre H, Missing Data Methods for Dealing with Missing Items in Quality of Life Questionnaires. A Comparison by Simulation of Personal Mean Score, Full Information Maximum Likelihood, Multiple Imputations, and Hot Deck Techniques Applied to the SF-36 in the French, *Quality of Life Research*. (2011) 287-300.
- [7] C. Li, Little's Test of Missing Completely at Random, *The Stata Journal*. (2013) 795–809.
- [8] K. B. A. L. Smeeth, What is the Difference Between Missing Completely at Random and Missing at Random? *International Journal of Epidemiology*. (2014) 1336–1339.
- [9] C. Li, Little's Test of Missing Completely at Random, *The Stata Journal*. 13(4) (2013) 795-809.
- [10] N. Shutoh, T. Nishiyama and M. Hyodo, Bartlett Correction to the Likelihood Ratio Test for MCAR with Two-Step Monotone Sample, *Statistica Neerlandica*. 71(3) (2017) 184-199.
- [11] D. S. G. A. Z. Yu, A Find Out: Finding Outliers in Very Large Datasets, *Knowledge and Information Systems*. (2002) 387 - 412.
- [12] C. C. Y. S. P. Aggarwal, Outlier Detection for High Dimensional Data, *Sigmod'01*. (2001) 37-46.
- [13] Ö. Senger, Impact of Skewness on Statistical Power, *Modern Applied Science*. (2013) 49-56.
- [14] M. Templ, A. Kowarik And P. Filzmoser, Iterative Stepwise Regression Imputation Using Standard and Robust Methods, *Computational Statistics & Data Analysis*. 55(10) (2011) 2793-2806.
- [15] J. Shao And H. Wang, Sample Correlation Coefficients Based on Survey Data Under Regression Imputation, *Journal of the American Statistical Association*. 97(458) (2002) 544-552.
- [16] D. P. R. Anil Jadhav, Comparison Of Performance Of Data Imputation Methods For Numeric Dataset, *Applied Artificial Intelligence*. (2019) 913-933.
- [17] Y. H. Christophe Crambes, Regression Imputation in the Functional Linear Model with Missing Values in the Response, *Journal of Statistical Planning and Inference*. (2019) 103-109.
- [18] S. Xu, Predicted Residual Error Sum of Squares of Mixed Models: An Application for Genomic Prediction, *G3 (Bethesda)*. (2017) 895–909.
- [19] J. L. & G. J. W. Schafer, Missing Data: Our View of the State of the Art, *Psychological Methods*. (2002) 147–177.
- [20] J. W. Graham, Missing Data Analysis: Making it Work in the Real World, *Annual Review of Psychology*. (2009) 549–576.
- [21] W. I. C. J. S. M. R. P. K. M. W. A. C. J. Sterne Jac, Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls, *BMJ*. (2009) 157–160.
- [22] G. D. Garson, *Missing Values Analysis and Data Imputation*, Asheboro, NC: Statistical Associates Publishers. (2015).
- [23] G. A. L. M. Abayomi K, Diagnostics for Multivariate Imputations, *Journal of the Royal Statistical Society*. (2008) 273–291.
- [24] K. G.-O. Stef Van Buuren, Mice: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*. (2011) 1-67.
- [25] Mera-Gaona, N. M., V.-C. U. And & L. D. M. R., Evaluating the Impact of Multivariate Imputation by Mice in Feature Selection, *Plos One*. 16(7) (2021).
- [26] Missing Data Methods for Dealing with Missing Items in Quality of Life Questionnaires. A Comparison by Simulation of the Personal Mean Score, Full Information Maximum Likelihood, Multiple Imputation, and Hot Deck Techniques Applied to the SF-36 in the French, *Quality of Life Research*. (2011) 287–300.