**Assignment - I**

**Question1 : Steps needed to perform imputation on a dataset.**

There are 3 types of ways, normally imputation is used

1. Imputation using Statistical methods(Mean, Median and Model)
2. Imputation using Machine Learning Algorithms
3. Imputation using Datawig Package
4. **Imputation using Statistical methods**

Imputation is one of the most common practices usually followed for treating the missing values in the dataset. For using statistical method for imputation distribution of the data need to be checked before replacing the missing values. Data could be sometimes skewed to left or right, or it could be symmetrically distributed. Also, for numerical and categorical variables we use different approaches. Mode is usually used for categorical variables, mean and median are considered for treating missing values in numerical variables. If the data is normally distributed mean is used replace the missing values. If the data is asymmetrical using median makes more sense.

1. **Imputation using Machine Learning Machine Learning Algorithms**

Machine Learning algorithms are strong tools for solving any complex computational tasks, likewise it can also be used in treating outliers in data analysis. The first step of imputation is to create subset of the original dataset. After that choose a Model generating algorithm like Simple Imputer for example. Train a machine learning model then using trained model we will be predicting missing values of the original dataset.

1. **Imputation using Datawig Package**

The first step of the Datawig, install the package from the Datawig official documentation page, Secondly, data is loaded. After that imputer needs to be initialized. Fitting the model is the next step. Once the models are fitted, data will be trained using the model. Like other machine learning algorithms, trained model, testing will be carried on the rest of the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl No** | **Student Name** | **English** | **Irish** | **Maths** | **Results** |
| 1 | Robert | 96 | 66 | NA | Pass |
| 2 | Vanda | 68 | 70 | 0 | Fail |
| 3 | Ken | NA | 45 | 56 | Pass |
| 4 | Robin | 90 | NA | 48 | Pass |
| 5 | Mrunali | 56 | 50 | 51 | Pass |
| 6 | Kate | 98 | 74 | 29 | Fail |
| 7 | Lidija | 0 | NA | NA | Fail |

Suppose we are working with the above dataset, First, the data need to be described and the target variable to be established. As per the dataset we have 7 records and 5 features. Please note that this dataset is only for the example, practically we take n\*100 or n\*1000 records for empirical analysis where n is the number of features. Now we know that ‘results’ is categorical missing values in this feature needs to be treated with mode in statistical imputation method. English, Irish and Maths are continuous numerical variable in which the missing values will be treated by replacing with mean or median.

|  |  |  |
| --- | --- | --- |
| Chart, histogram  Description automatically generated |  |  |

Looking at the distplot, English and maths seems to have skewness to the left side where as irish has almost symmetrical distribution. In this case we will be taking Median for Both English and maths where as irish will be treated with median.

we will make a distplot or barplot from the data and see wat kind of distribution is the data is having.

**Question2 : Why do you think having mean as imputed value not good for asymmetric distribution?**

In Data Imputation we have three values considered for missing values replacement.

Mean median and Mode. If you have below 5 values for example. If this is the data

X = { 25, 10, 6, 18, 6 }

In order to get the mean, we need to sum the numbers and divide them with number of occurrences.

= =

So, the mean of the series x is 13. However, to find the median, data must be sorted.

So, the series will be changed as follows

X = { 6, 6, 10, 18, 25 }

If the value of n is odd, we can take the element in n-1/2th position. If the n value is even then you take n/2 and n/2+1 element and take there mean. Please note that n represents number of elements in a series. In the above example, mean is 13 and mode is 10.

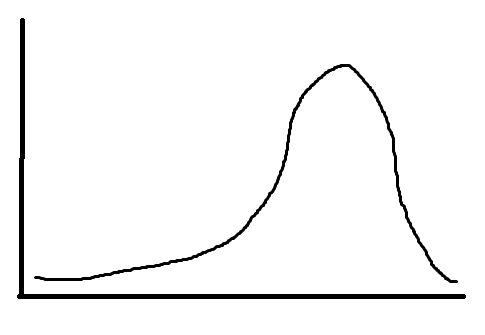
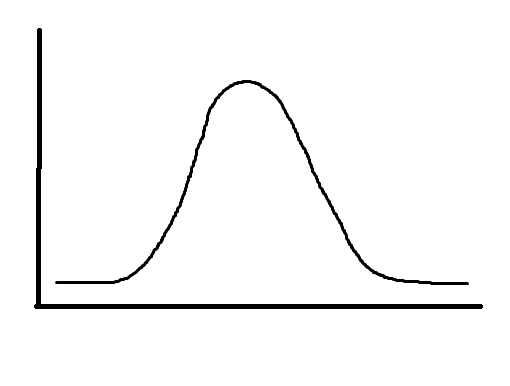


fig 1 Symmetric fig 2 Asymmetric(skewed)

If you look at fig1 it is symmetrically distributed. The idea of imputation is to replace missing values with the average of the variable. In this kind of distribution, taking mean will be meaningful for example in the previously shown series x, 13 is the mean. Meanwhile mode calculated gave 10 as the value. The series we took is skewed to the left because as we can see the values having occurrence less or equal to ten is occurred 3 times and other two numbers occurred two times(18 and 25). So, this is an asymmetric data and requires media to work with. histogram represents the weight of a value as height of each bar. So, when you choose mean in asymmetric data the weight of data is not evenly distributed and it creates a bias in the data and make in unbalanced. So, if you choose mean in the series x which skewed to the right, it will use 13 as value for replacement instead of taking the median which is 10.

Similarly, When it comes to Categorical variable like, show size, age group, color or result using mode will be more appropriate. Because using mode we are checking the number of occurrences of the value in the dataset. So, using mode will make more sense on categorical data.