1. **Why do we choose boxplot method than other for outlier detection and removal?**

Outlier detection method depends upon the business process like how the data is generated or what is the business flow. But most of the business uses box plot as it is easy to analyze the data.

Box plot is a graphical method to detect outliers, which is easier as compare to other methods. It is depicted by quartiles and inter quartiles that helps in defining the upper limit and lower limit beyond which any data point lying will be considered as outliers.

The main purpose of this diagram is to identify outliers and discard it from the data series before making any further observation so that the conclusion made from the study gives more accurate result.

The procedure holds only 3 steps to get the result:

Lower quartile (Q1)

Median (Q2)

Upper quartile (Q3)

So 1st step will to get Q1, Q2 and Q3:

Q1 = ¼ (n+1) \*n = number of data points

Q2 = ½(n+1)

Q3 = ¾(n+1)

Interquartile range (IQR) = Q3-Q1

Lower limit = Q1 – 1.5 IQR

Upper limit = Q3 + 1.5 IQR

**How do we choose best method to impute missing value for a data?**

There are number of methods to impute missing value but choosing the best one depends upon the type of data. We have 3 procedures in which there are methods:

* Central statistics
* Distance based
* Prediction

So 1st we have to create small subset of data then we have to delete some values manually. Then use multiple methods to fill. See the closer match value with actual value and fix that method for imputation.

So by using central statistics method we get 2 type of data set categorical and numerical. So for categorical data we can use mode value for missing value imputation and for numerical data we can use median or mean value.

For distance based data we use KNN imputation (K nearest neighbor) method. In python and R there is library, by using that we can get KNN imputation. As this method uses nearest value for imputation it is also called lazy imputation method.