Simple vs Advanced Missing Variable Imputation For A Continuous Numeric Variable

Mo Abulyusr
Connect With Me

Introduction:

In this guide, we will explore various methods for imputing continuous numerical variables and illustrate the outcomes of each technique. Our focus will be on simple imputation approaches like mean and median imputation, as well as more advanced strategies like K-nearest neighbors (KNN).

What Does Each Technique Do?

Mean imputation: Mean imputation involves filling in missing data with the average value (mean) of the observed data, where the mean is the sum of all observed values divided by the number of observations

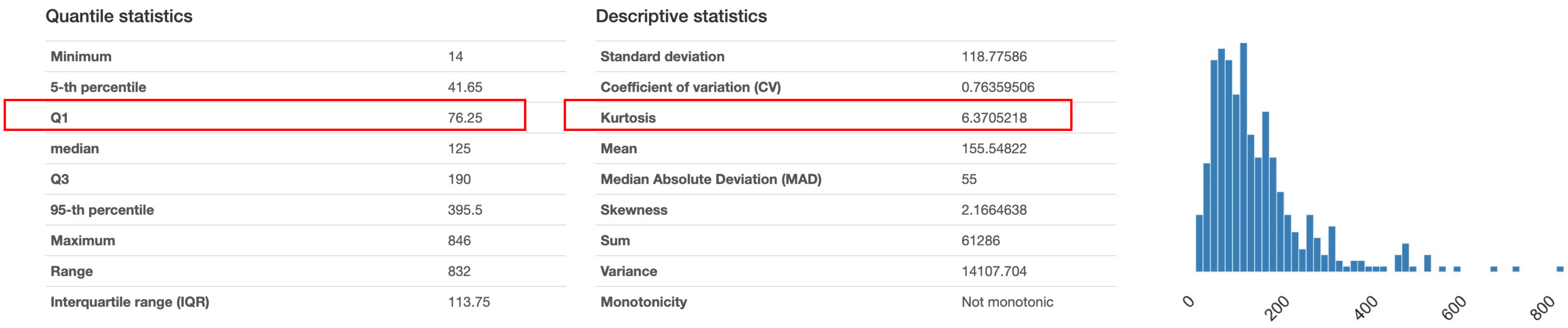
Median imputation: Median imputation, on the other hand, replaces missing entries with the median value, which is the middle number in a sorted list of the observed values

KNN imputation: KNN imputation infills missing data by identifying the 'nearest neighbors'—data points with similar features—and then estimates the missing values using the known values of those neighbors

Dataset Info:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage. Variables in the dataset include number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age, etc..

Variable of Interest 'Insulin' Stats:



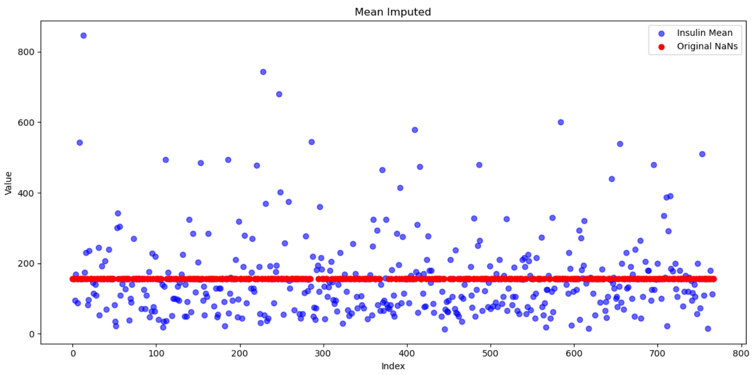
Quantile statistics

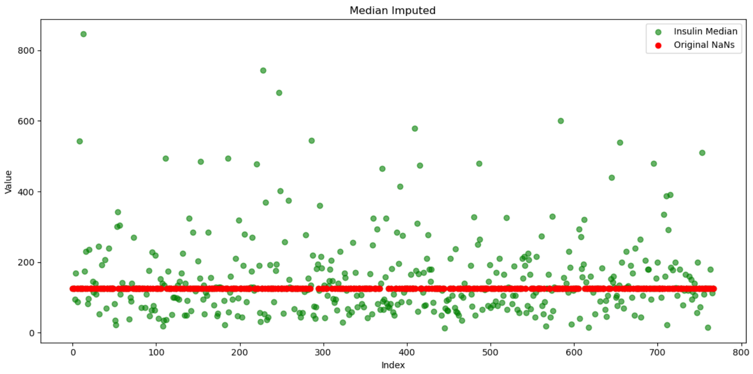
Minimum	14
5-th percentile	41.65
Q1	76.25
median	125
Q3	190
95-th percentile	395.5
Maximum	846
Range	832
Interquartile range (IQR)	113.75

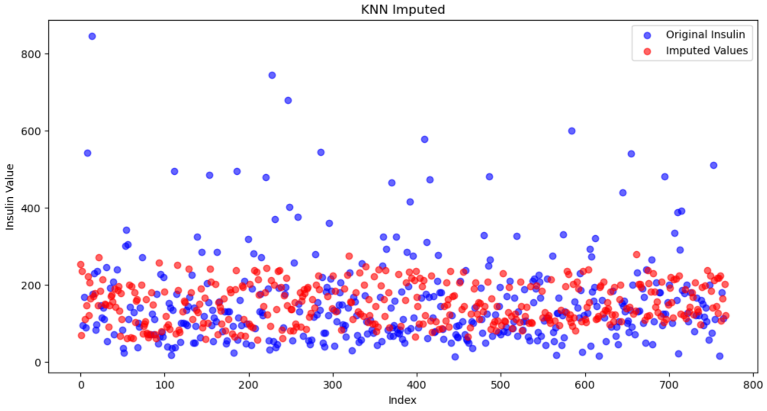
Descriptive statistics

Standard deviation	118.77586
Coefficient of variation (CV)	0.76359506
Kurtosis	6.3705218
Mean	155.54822
Median Absolute Deviation (MAD)	55
Skewness	2.1664638
Sum	61286
Variance	14107.704
Monotonicity	Not monotonic

Scatter Plot of 'Insulin' After Imputing Using the Mean, Median, and KNN





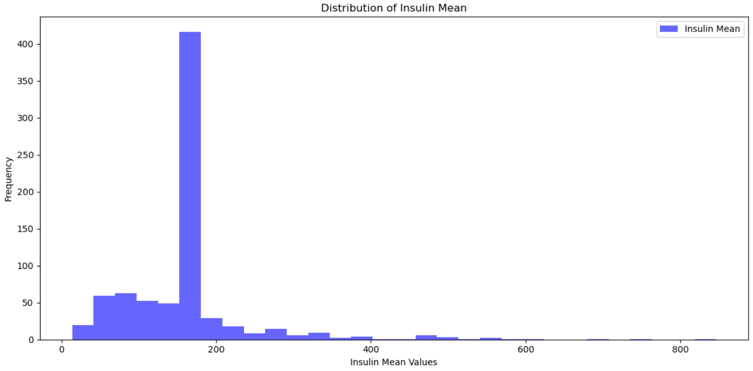


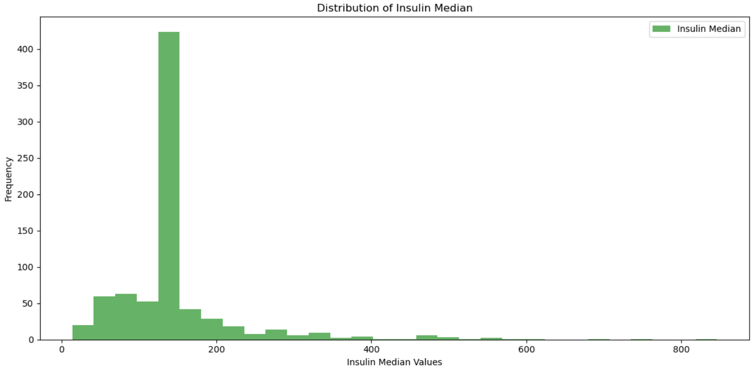
The provided scatter plots illustrate the outcomes of different imputation methods on 'Insulin' data.

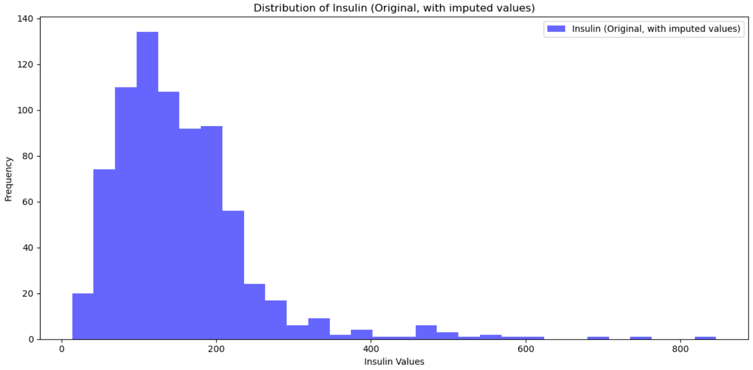
Mean and median imputation create uniform lines of points, as all missing values are filled with a single central value, potentially oversimplifying the dataset.

KNN imputation, however, introduces a more varied set of values, reflecting the natural variability in the data by using the most similar observations to estimate the missing points, resulting in a more nuanced and representative imputation.

Histogram Showing Distribution of 'Insulin' After Imputing Using the Mean, Median, and KNN







The first histogram shows a sharp peak at the mean value with a right-skewed distribution, signifying that all missing values have been replaced with the mean, which doesn't reflect the original data's variability.

The second histogram, for the median imputation, also shows a sharp peak, indicating a single imputation value, the median, which similarly ignores the data's diversity.

The third histogram, representing the original data with imputed values using KNN, demonstrates a more natural distribution with varying frequencies across the range, suggesting that KNN imputation maintains a distribution closer to what one might expect from the unimputed data. This method tends to preserve the original structure and variability of the dataset more effectively than mean or median imputation.