Original Research Article

Inputting Missing Values in Mechanical Materials Data: Accuracy and Statistical Effect of Mean, Median, and KNN Methods

Jaydn N. Su

Eastvale STEM Academy at Eleanor Roosevelt High School, 7447 Scholar Way, Eastvale, CA 92880 United States

ABSTRACT

This study explores how different methods for handling missing data change the accuracy of mechanical property datasets. A dataset containing ultimate tensile strength (Su) and related mechanical properties was used, with 100 Su values randomly removed to simulate realistic data loss. Three commonly used methods were tested: mean substitution, median substitution, and k-nearest neighbor (KNN) imputation. Each completed dataset was then compared to the original to see how well statistical relationships were maintained. The results indicated that the median imputation method produced the most accurate reconstruction in this dataset and under MCAR simulation among the tested methods, maintaining an almost exact correlation with the original Su values with a Pearson correlation coefficient (r) value of 0.9987 and a coefficient of determination (R^2) value of 0.9487 in the linear regression model. Both mean and KNN imputation performed sufficiently, but introduced larger deviations from the original relationships. Overall, the findings show that under the MCAR missingness simulation used here and within this specific mechanical-property dataset, the median imputation method provides the most effective balance between accuracy and preservation of statistical structure among the three methods tested, suggesting that median imputation may serve as a practical solution for researchers and engineers who regularly work with incomplete mechanical property data.

Keywords: Imputation; K-Nearest Neighbor; Simulation; Mechanical Property; Missing Values

INTRODUCTION

In the field of materials science and mechanical engineering, accurate characterization and analysis of mechanical properties such as ultimate tensile strength (*Su*), yield strength (*Sy*), elastic modulus (*E*), shear modulus (*G*), and Poisson's ratio (μ) are essential for producing safe, efficient, and sustainable designs. These properties form the foundation for structural analysis, product development, and performance evaluation across a wide range of engineering applications. As the field continues to emphasize optimization and sustainability, the ability to perform reliable data analysis has become an essential component of innovation and design integrity.

Despite their importance, obtaining complete and in-depth datasets on mechanical properties is often challenging. Experimental testing of materials requires considerable time, specialized equipment, and significant financial resources, limiting the scale of data collection. As a result, missing data is a common issue appearing in both laboratory experiments and industrial databases (1, 2). When even a singular property within a set is absent, researchers are often forced to remove the entire record,

Corresponding author: Jaydn N. Su, E-mail: jaydnsu335006@gmail.com.

Accepted April 2, 2026
https://doi.org/10.70251/HYJR2348.42319326

reducing sample size and wasting valuable information. These limitations both reduce efficiency and restrict the development of predictive models that rely on complete and consistent datasets.

Researchers have long acknowledged the issue of missing data in statistics and data science (1). Three primary mechanisms of missingness have been identified: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The underlying cause of missing data strongly influences which imputation method should be applied (1). Later work expanded the practical application of imputation by introducing techniques that balance computational efficiency and statistical accuracy (2). Several approaches have been developed to address this issue. Among the most common are mean and median substitution, which replace missing values with a single representative statistic. These methods are simple to apply and computationally efficient, but they can distort data distributions and reduce correlations by artificially lowering variance (1, 3). Despite these drawbacks, mean and median imputation remain widely in use because of their accessibility and low computational cost.

More complex methods such as KNN imputation have been developed to improve accuracy by accounting for relationships among multiple variables. This method estimates missing values by identifying the most similar observed data points and using their values to generate predictions. KNN performs particularly well within datasets containing strong multivariate relationships (4). However, its accuracy depends heavily on the consistency of predictor variables and the absence of outliers (5). While these findings demonstrate the flexibility of KNN, they also reveal its dependence on data quality and structure, making it less reliable for smaller or inconsistent engineering datasets.

In materials science and mechanical engineering, the problem of missing data is both practical and statistical. Obtaining measurements for properties such as *Su*, *Sy*, and *E* is both expensive and time-consuming, which often leads to incomplete or limited datasets. Although imputation is commonly used in disciplines such as medicine and economics, it has not been widely studied in the context of mechanical property data. Maintaining realistic correlations and regression relationships is especially important in this field because the variables involved are physically interdependent (6, 2).

This study compares three widely used imputation methods, mean substitution, median substitution, and KNN imputation, to determine how each affects correlation structure and regression accuracy in mechanical property data. First, each method's reconstructed *Su* values are compared directly with the original data to measure accuracy. The analysis then examines how each method alters correlations between *Su* and the other mechanical properties (*Sy*, *E*, *G*, μ, and ρ). Finally, linear regression models are used to evaluate the influence of each approach on predictive relationships, with *Su* treated as the dependent variable. These combined analyses provide a comprehensive view of how each approach influences both the numerical precision and the structural consistency of the dataset.

This study compares several imputation methods specifically using mechanical property data, rather than treating missing data methods only in a general statistical context. This helps handle a gap in existing research, where imputation methods are usually tested on abstract or simulated datasets instead of real engineering measurements. The findings are intended to give engineers and researchers clearer guidance when choosing how to handle incomplete mechanical data, supporting more reliable analysis and better-informed design decisions.

METHODS AND MATERIALS

Data Description

The dataset used in this study was obtained from Kaggle (7) and is publicly available for research use. It contains 1,553 observations and 15 variables, and is a real-world dataset describing a range of mechanical and physical properties for different engineering materials. The variables include a material standard (*Std*), a unique identification code (*ID*), material name, heat treatment method, ultimate tensile strength (*Su*), yield strength (*Sy*), elongation at break or strain (*A5*), Brinell hardness number (*BHN*), elastic modulus (*E*), shear modulus (*G*), Poisson's ratio (μ), density (ρ), pressure at yield (*pH*), a brief material description (*Desc*), and Vickers hardness number (*HV*).

This dataset was chosen because it represents the type of data typically encountered in real engineering settings. The values are derived from experimental testing rather than simulated processes, which introduces natural variability and realistic measurement conditions. Since the goal of this study is to evaluate how imputation methods perform under practical constraints, using an authentic materials dataset strengthens the relevance of the results.

A key strength of the dataset is the presence of several

mechanically related numeric variables, particularly *Su*, *Sy*, *E*, and *G*. These properties are physically linked and commonly used together in structural analysis and materials modeling. Their relationships make the dataset well-suited for examining whether imputation methods preserve meaningful correlations and regression behavior, and the sample size is also large enough to support statistical analysis while remaining manageable for computational processing.

At the same time, the dataset has certain limitations. Some variables are categorical or descriptive, such as material name, heat treatment method, and the textual description field, which cannot be directly included in numeric correlation or regression models. In addition, as with many experimental datasets, the measurements may include outliers or uneven distributions, reflecting differences in material classes and testing conditions. These features can affect the performance of more complex imputation methods, especially those based on distance calculations.

Overall, the dataset aligns well with the objectives of this study. It allows for realistic simulation of missing data, meaningful comparison of imputation strategies, and evaluation of how reconstructed values influence both correlation structures and predictive models. As such, it provides a practical and appropriate foundation for assessing imputation methods in the context of mechanical engineering data.

This study examines the influence of different methods of data imputation on the correlation structure of a dataset containing missing values. The dataset, titled Data.csv, contained the variable *Su* along with additional numeric predictors. A complete summary of the dataset variables, including their definitions, data types, and descriptive statistics, is provided in Table 1. To simulate a realistic scenario of data loss, 100 entries were randomly removed from the *Su* column. Because the removed Su entries were selected uniformly at random, the simulated missingness mechanism is MCAR, meaning the probability of a value being missing is independent of both the observed data and the unobserved (missing) values. This process created a modified dataset, referred to as missing_df, in which the designated values of *Su* were replaced with missing entries (NaN). Choosing random entries ensured that the missing data was distributed evenly across the dataset without bias toward any specific cases. This study therefore evaluates imputation performance only under MCAR conditions; real engineering datasets may exhibit Missing at Random (MAR) or Missing Not at Random (MNAR) mechanisms, and results may differ under those patterns.

Table 1. *Variables included in the mechanical materials dataset, with their definitions, data types, and descriptive statistics used in the present analysis.*

Variable	Type	Definition	Descriptive Statistics
Std	Categorical	Material standard classification	n = 1552; unique = 8
ID	Identifier	Unique identification code for each material	n = 1552; unique = 1552
Material	Categorical	Material name	n = 1552; unique = 1225
Heat Treatment	Categorical	Heat treatment method applied to the material	n = 802; unique = 44; missing = 750
Su	Numeric (Continuous)	Ultimate tensile strength (MPa)	mean = 572.75; SD = 326.83; min = 69; max = 2220
Sy	Numeric (Continuous)	Yield strength (MPa)	mean = 387.76; SD = 290.04; min = 28; max = 2048

Continued Table 1. *Variables included in the mechanical materials dataset, with their definitions, data types, and descriptive statistics used in the present analysis.*

Variable	Type	Definition	Descriptive Statistics
A5	Numeric (Continuous)	Elongation at break or strain (%)	mean = 19.33; SD = 12.42; min = 0.5; max = 70
BHN	Numeric (Continuous)	Brinell hardness number	mean = 177.14; SD = 113.51; min = 19; max = 627
E	Numeric (Continuous)	Elastic modulus (MPa)	mean = 164571.52; SD = 56135.41; min = 73000; max = 219000
G	Numeric (Continuous)	Shear modulus (MPa)	mean = 85598.84; SD = 125326.80; min = 26000; max = 769000
μ (mu)	Numeric (Continuous)	Poisson's ratio	mean = 0.303; SD = 0.025; min = 0.20; max = 0.35
ρ (Ro)	Numeric (Continuous)	Density (kg/m³)	mean = 6929.84; SD = 2115.17; min = 1750; max = 8930
pH	Numeric (Continuous)	Pressure at yield (MPa)	mean = 627.39; SD = 370.53; min = 190; max = 1360
HV	Numeric (Continuous)	Vickers hardness number	mean = 328.48; SD = 202.76; min = 105; max = 800
Desc	Text	Material description	n = 981; missing = 571

Three imputation strategies were applied to replace the missing values in *Su*: mean imputation, median imputation, and k-nearest neighbor (KNN) imputation. In mean imputation, each missing *Su* value was replaced with the average of the observed ones, forming the dataset m1_df. For median imputation, each missing entry was filled with the median of the available *Su* values, which produced m2_df. KNN imputation is a distance-based method that estimates missing values by identifying the most similar observations (neighbors) in the dataset and deriving imputed values from them, often through a weighted average (8, 4).

Similarity between two observations, p and q, was measured using the Euclidean distance across all standardized numeric predictors, defined as

$$d(p,\ q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

where p_k and q_k represent the standardized values of predictor k for observations p and q, respectively.

Standardization of numeric variables ensured that each predictor contributed equally to this distance calculation (9). The k observations with the smallest distances to the incomplete case were selected as neighbors, and the missing *Su* value was computed as a weighted average of those neighbors' *Su* values, with closer observations receiving greater weight. For each missing value, up to five nearest neighbors were used, and the resulting KNN-imputed dataset was saved as m3_df.

After the three imputed datasets were generated, a series of statistical analyses was performed to compare each with the original dataset. Pearson correlation matrices were calculated for both the original and imputed datasets to examine the linear relationships between *Su* and the other numeric variables. To determine how each imputation method influenced these relationships, difference matrices were then calculated by subtracting the original correlations from the imputed correlations. These matrices could be visualized as heatmaps to allow for a clear comparison of correlation structures and any changes that occurred following imputation. However, in this study, for simplicity and consistency, only the correlation coefficients and corresponding p-values were used for analysis. Additionally, summary measures such as the mean, median, and maximum absolute difference in correlation values were computed to provide a more concrete indication of how strongly each imputation method altered the relationships among variables.

Finally, multiple linear regression models were fit using ultimate tensile strength (Su) as the dependent variable and yield strength (Sy), elastic modulus (E), shear modulus (G), Poisson's ratio (μ), and density (ρ/Ro) as predictor variables. The model form was

$$Su = \beta_0 + \beta_1(Sy) + \beta_2(E) + \beta_3(G) + \beta_4(\mu) + \beta_5(\rho) + \varepsilon$$

The same model specification was applied to the original dataset and to each imputed dataset to allow direct comparison of regression coefficients, statistical significance, adjusted and unadjusted R^2 values, and overall model fit. To assess multicollinearity among predictors, variance inflation factors (VIFs) were examined. Regression diagnostics were also evaluated using residual-versus-fitted plots for homoscedasticity, Q–Q plots of residuals for normality, and residual-order inspection or an independence test for residual independence. In addition, standardized coefficients were compared so that predictor importance could be evaluated across variables measured in different units.

RESULTS

Three sets of analyses were performed to evaluate how the imputation methods affected the dataset. These analyses included comparing imputed Su values with the original data, examining changes in correlations between Su and other variables, and assessing how imputation influenced linear regression outcomes.

Comparison with Original *Su*

The first analysis compared each imputed version of *Su* with the original variable using correlation coefficients and mean absolute differences. As shown in Table 2, the median-imputed variable (Su_M2) achieved the highest similarity to the original data, with a correlation coefficient of 0.9987 and a mean absolute difference of 2.96. The mean-imputed (Su_M1) and KNN-imputed (Su_M3) versions showed slightly lower correlations (0.9703 and 0.9694, respectively) and larger mean absolute differences (14.92 and 14.58). These findings suggest that median imputation most closely preserved *Su* in this MCAR simulation, while mean and KNN imputation introduced greater variation.

Table 2. *Comparison of imputed Su values with the original dataset, including correlation coefficients, p-values, and mean absolute differences for each imputation method.*

Variable	Correlation Coefficient	P-Value	Mean Absolute Difference
Su_M1	0.970324	0.0	14.915516
Su_M2	0.998666	0.0	2.959684
Su_M3	0.969387	0.0	14.577320

Effect on Correlations with Other Variables

Pearson correlation matrices were computed to evaluate how each imputation method affected relationships between *Su* and the other variables (*Sy*, *E*, *G*, *μ*, and *Ro*). The results indicated that all three methods produced relatively minor changes in correlation strength and direction.

In the original dataset, *Su* demonstrated strong positive correlations with *Sy* (r = 0.957), moderate positive correlations with *E* (r = 0.594), and weak positive correlations with *G* (r = 0.255) and *Ro* (r = 0.408). A weak negative correlation was observed with μ (r = –0.240). After mean imputation, correlations between *Su* and other variables slightly decreased, most notably with *Sy* (r = 0.931) and *E* (r = 0.572). KNN imputation produced nearly identical results to the mean method (*Su–Sy* r = 0.930; *Su–E* r = 0.572), showing that both techniques moderately weakened linear relationships. On the other hand, median imputation preserved almost all of the original correlations, maintaining *Su–Sy* (r = 0.958) and *Su–E* (r = 0.595) nearly unchanged. These findings suggest that median imputation caused the least disruption to inter-variable relationships, while mean and KNN slightly reduced correlation strength across variables.

Impact on Linear Regression Results

Linear regression analyses were conducted for the original and each imputed dataset, using *Su* as the dependent variable and (*Sy*, *E*, *G*, μ, *Ro*) as predictors. Table 3 summarizes the R^2 values and key coefficients.

Table 3. *Comparison of regression model fit across the original and imputed datasets, using R^2 and adjusted R^2 values.*

Dataset	R^2	Adj. R^2
Su_Ori	0.9478	0.9476
Su_M1	0.8959	0.8956
Su_M2	0.9487	0.9486
Su_M3	0.8945	0.8941

The median-imputed dataset (*Su_M2*) again demonstrated the strongest alignment with the original model, achieving an R^2 value (0.9487) nearly identical to the original (0.9478). In contrast, both mean (*Su_M1*) and KNN (*Su_M3*) imputations reduced model fit, with R^2 values around 0.895. Coefficient comparisons further supported these trends. The regression parameters for *Su_M2* remained close to the original coefficients across all predictors, while *Su_M1* and *Su_M3* showed lower coefficients for *Sy* and *E*, indicating reduced predictive strength. Although the direction of relationships (positive or negative) remained consistent, the magnitude of effects diminished slightly for the mean and KNN imputations, aligning with the observed decrease in overall correlation.

DISCUSSION

The results of this study demonstrate that the choice of imputation method has a meaningful influence on how well a dataset retains its original statistical structure. Across the three methods evaluated, which included mean, median, and k-nearest neighbor (KNN) methods, the median approach consistently showed the closest alignment with the original data in this dataset under MCAR missingness. Its correlation with the unaltered Su values remained nearly perfect (r = 0.9987), and the regression model built from the median-imputed dataset closely matched the performance of the original with an R^2 of 0.9487. Both the mean and KNN methods were effective but produced slightly weaker fits and larger deviations from the original values, a pattern consistent with prior findings on the sensitivity of imputation methods to data structure (6, 2).

The success of the median method can be attributed to its resistance to extreme values and its ability to reflect the central tendency of the data without artificially altering overall variability. Because the median is not influenced by unusually high or low observations, it remains stable even when the dataset includes outliers. On the other hand, mean imputation is more sensitive to extreme values and can shift the distribution toward those points, leading to reduced variance and weakened correlations (3, 6). KNN imputation depends heavily on the consistency of the neighboring predictors used for estimation (4, 5) meaning its accuracy can vary when the predictor variables differ substantially across cases. This interpretation is supported by the results, as observed in the modest decline in correlation between Su and Sy after applying KNN and mean imputation.

These results align closely with previous studies comparing statistical and algorithmic approaches to handling missing data. Even small proportions of imputed values can distort correlation coefficients when variance is artificially reduced (6), while simpler imputation strategies often perform well when relationships among variables are strong and approximately linear (2). In a dataset characterized by consistent numeric relationships, the added complexity of KNN yielded little additional benefit. In this setting, a straightforward median substitution retained relational integrity more effectively than more elaborate models.

The regression analysis offers a complementary view of these effects. Although all imputed models produced coefficients in the same general direction as the original model, the mean and KNN versions showed small reductions in magnitude and overall explanatory power. This suggests that the act of imputation, particularly when using methods sensitive to data distribution, can alter not only individual values but the relative influence of predictors. In engineering contexts, such shifts can have practical consequences for material selection, performance prediction, and safety assessment (10).

It is important, however, to consider the limits of these findings. The missing values introduced here were MCAR by construction (random deletions from *Su*), so conclusions are restricted to MCAR conditions. In real-world research, missingness is often systematic (i.e., MAR or MNAR), and combined with nonlinear relationships among material properties, particularly in systems influenced by processing conditions or microstructural variability (11). Under MAR/MNAR mechanisms, relative imputation performance and downstream correlation/regression effects may change. Future studies could expand this work by applying the same comparison to multiple datasets, testing other imputation techniques such as multiple imputation or regression-based estimation, and examining how these choices affect more advanced models beyond simple correlations and regressions.

Ultimately, these results suggest that median imputation offers a balanced and statistically sound approach when missingness is random and the underlying relationships in the data are stable. Its simplicity and resistance to distortion make it a practical choice for many analytical contexts. While algorithmic methods like KNN remain valuable, their effectiveness depends on the consistency and scale of the data.

CONCLUSIONS

Across all three imputation methods tested, the median approach preserved the statistical structure of the data most effectively for this dataset under the MCAR deletion experiment. It produced results that were almost identical to the original dataset in both correlation and regression analyses. The mean and KNN methods, while still functional, introduced slightly greater differences, indicating a small but measurable loss of accuracy.

The median method performed well because it is naturally resistant to extreme values and continues to represent the center of the data without distorting overall variability. Since unusually large or small observations do not strongly affect the median, it remains stable even when outliers are present. By comparison, mean imputation can be pulled toward extreme values, and KNN imputation relies on how consistent the surrounding predictor variables are, which can introduce additional uncertainty when those variables vary across the dataset.

This study also highlights the broader importance of handling missing data carefully in engineering research. Even small inaccuracies introduced during imputation can affect regression coefficients and reduce the reliability of predictive models. Over time, these distortions can influence engineering decisions related to material selection and performance evaluation, directly affecting safety, structural integrity, and design performance.

Future studies could extend this analysis by using larger and more diverse datasets and by exploring additional imputation strategies, such as multiple imputation and regression-based methods. It would also be valuable to test how these techniques perform when data are not missing at random, since real-world datasets often include systematic gaps. Continued investigation in this area will help refine data reconstruction methods and support the development of more reliable, data-driven approaches to materials design and analysis.

ACKNOWLEDGEMENTS

The feedback and critique provided by the anonymous reviewers are deeply appreciated. The author also extends gratitude to the contributors of the Kaggle dataset used in this study.

CONFLICT OF INTEREST

The author declares that there are no conflict of interests related to this work.

REFERENCES

1. Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. Wiley; 2019. https://doi.org/10.1002/9781119482260
2. van Buuren S. Flexible imputation of missing data. 2nd ed. Chapman and Hall/CRC; 2018. https://doi.org/10.1201/9780429492259
3. Wicklin R. How mean imputation affects the variance and correlation of your data. The DO Loop (SAS Blog). 2017 May 31. Available from: https://blogs.

sas.com/content/iml/2017/05/31/mean-imputation-variance-correlation.html (accessed on 2025-10-05).
4. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak*. 2016; 16 (Suppl 3): 74. https://doi.org/10.1186/s12911-016-0318-z
5. Li X, Fang H, Zhang H, Chen C. A comparative study of data imputation methods for engineering datasets. *J Stat Comput Simul*. 2024; 94 (2): 289-305. https://doi.org/10.1080/00949655.2023.2257112
6. Taylor J, Baggaley A, Hodge V. The effect of missing data imputation on correlation: a simulation study. *Comput Stat Data Anal.* 2016; 102: 72-84. https://doi.org/10.1016/j.csda.2016.03.001
7. Nawale P. Materials: Mechanical properties dataset (Version 1.0) [dataset]. Kaggle; 2023. Available from: https://www.kaggle.com/datasets/purushottamnawale/materials/data (accessed on 2025-10-05).
8. Troyanskaya O, Cantor M, Sherlock G, Brown P, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001; 17 (6): 520-525. https://doi.org/10.1093/bioinformatics/17.6.520
9. Zhang Z. A k-nearest neighbor based imputation procedure. *Stat Probabil Lett*. 2020; 139: 1-10. https://doi.org/10.1016/j.spl.2018.03.022
10. Agrawal A, Choudhary A, Kalidindi SR. Materials informatics for accelerated materials discovery and development. *Data-Centric Engineering.* 2021; 2: e19. https://doi.org/10.1017/dce.2021.19
11. Alasfar RH, Ahzi S, Barth N, Kochkodan V, Khraisheh M, Koç M. A review on the modeling of the elastic modulus and yield stress of polymers and polymer nanocomposites: effect of temperature, loading rate and porosity. *Polymers*. 2022; 14 (3): 360. https://doi.org/10.3390/polym14030360