YouTube Video Transcript Notes: Handling Missing Data in Numerical Columns

Video Goal: This video series focuses on handling missing data, specifically in numerical columns. This video covers imputation techniques for numerical data. Subsequent videos will address categorical data.

I. Univariate vs. Multivariate Imputation:

* **Univariate:** Imputation using only values within the *same* column. Methods include using the mean, median, or randomly selecting from existing values. Simple, but can distort the distribution if many values are missing.

* **Multivariate:** Imputation considering values across *multiple* columns. This video will cover KNN imputation and other multivariate techniques in later videos. More complex but potentially more accurate.

II. Univariate Imputation Techniques for Numerical Data (Focus of this video):

The video details four methods for univariate imputation:

1. **Arbitrary Value Imputation:** Replacing missing values with an arbitrary constant (e.g., 999, -1, 0). Simple but can significantly distort the distribution and mislead models. Useful for creating a distinct marker for missing data.

2. **Mean/Median Imputation:** Replacing missing values with the mean or median of the existing values in the column.

* **Choosing Mean vs. Median:**

* Use the **mean** if the data is normally distributed.

* Use the **median** if the data is skewed (to mitigate the influence of outliers).

* **Advantages:** Simple to implement, computationally inexpensive, easy to reproduce in various environments (including production machine learning systems).

* **Disadvantages:**

* Distorts the distribution of the data (changes the shape).

* Can mask outliers, leading to inaccurate analysis.

* Alters the relationships between the imputed column and other columns (correlations can change). This is referred to as inducing spurious correlations.

* **When to Use:** Best suited for small percentages of missing data (less than 5%). Should be used cautiously and only when other, more sophisticated techniques are not feasible.

3. **End of Distribution Imputation:** Replacing missing values with values at the extreme ends of the data's distribution.

* **For Normally Distributed Data:** Use mean ± 3 * standard deviation. This captures values outside the typical range (outliers).

* **For Skewed Data:** Use interquartile range (IQR) based methods (e.g., Q1 - 1.5 * IQR, Q3 + 1.5 * IQR).

* **Advantages:** Explicitly marks missing data as outliers.

* **Disadvantages:** Similar to arbitrary value imputation, it can distort the distribution and relationships with other variables.

* **When to Use:** When data is not missing at random and you want to highlight the missingness.

4. **Random Sample Imputation:** Replacing missing values with randomly selected values from the existing data in the column. Preserves the distribution better than mean/median imputation but introduces randomness. Not explicitly covered in detail in this video, but mentioned as a future topic.

III. Practical Demonstration:

The presenter demonstrates mean and median imputation using the Titanic dataset in Python. The code is shown, and the impact on the distribution (using histograms) and correlation with other variables is analyzed before and after imputation. Key observations include changes in variance and potential spurious correlations.

IV. Selecting the Best Imputation Technique:

The presenter emphasizes the importance of evaluating the impact of imputation on the data's distribution, variance, and correlations with other variables. Automatic selection of the best technique will be covered in a future video.

V. Using SimpleImputer (Python):

The presenter shows how to use the `SimpleImputer` function in Python to easily apply different imputation strategies (mean, median, most frequent, constant) to multiple columns simultaneously. This is demonstrated by applying median imputation to one column and mean imputation to another within the same dataset.

VI. Additional Imputation Techniques (Future Videos):

The presenter mentions that future videos will cover:

* Random sample imputation.

* Automatic selection of the best imputation technique.

* Imputation techniques for categorical data.

Overall, the video provides a practical introduction to handling missing numerical data, emphasizing the importance of understanding the implications of each technique and choosing the most appropriate method based on the data's characteristics and the goals of the analysis.