|  |
| --- |
| **Data Set:** Grades1000 |
| **Question / Problem:**  Import the Grades1000 dataset and clean the data! |
| **Variables:**   * ID, unique id - numeric * Gender, factor – male or female * EthnicGroup, - factor – a, b, c, d, e * ParentEducationLevel, character – education level type value * LunchType, - factor – standard or free/reduced * PrepTest, - factor – completed or none * MathGrade, numeric 0 - 100 * ReadingGrade, numeric 0 - 100 * WritingGrade, numeric 0 - 100 |
| **Techniques used:**  Data Inspection: used to provide an understanding of the data structure and variable names  Removing Duplicates: Ensured no repeated observations which can mess with analysis  Handling Missing Values:   * Replaced blank values with NA. * Identified missing values which allowed for a quicker and more comprehensive check of mussing data. * Dropped columns with more than 70% missing data because it provided little value   Data Type Conversion: Converted variables to appropriate tyles to make it easier to analysis and avoid errors.  Standardizing Values; Cleaned text values like gender to ensure uniformity.  Handling Outliers: Used a boxplot to visualize outliers.  Imputing Missing Data:   * Replaced missing data with “unknown” to retain observations. * Replaced the missing numerical data with the mean to preserve distribution.   Exporting Clean Data: Saved the clean data to a new cvc file. |
| **Analysis and Visualisation:**  **1. Box Plots**   * **Description**: Used to identify outliers in numerical variables.   **2. Summary Statistics**   * **Description**: Provided key metrics such as mean, median, minimum, and maximum values for numerical variables.   **3. Using summarise(across())**   * **Description**: Allowed for quick validation of missing values across all variables.   **4. Saved Cleaned Data**   * **Description**: The final cleaned dataset was exported as a .csv file for future analysis.   A screenshot of a computer program  AI-generated content may be incorrect.  Above code allows for understanding of the data structure and the variable names    A screenshot of a computer  AI-generated content may be incorrect.  A screen shot of a computer  AI-generated content may be incorrect.  A screenshot of a computer  AI-generated content may be incorrect.A screenshot of a computer  AI-generated content may be incorrect.  Box plot of the outliers  A diagram of a diagram  AI-generated content may be incorrect.    Replacing missing data with “unknown”A screenshot of a computer  AI-generated content may be incorrect.  Replacing missing numerical data with the mean    The cleaned data being saved to a new file. |
| **Considerations:**  The missing data imputation methods can affect the variability (ie using the mean)  Dropping columns with high missing rates reduced the data  Encountered errors which required further inspection of the columns.  Outliers were visualized but not removed or transformed. |