# Methodology

**1. Importing Libraries:**

Aim: To load necessary Python packages for data manipulation, visualization, and machine learning.

How Obtained: Executing import statements like `import pandas as pd`.

Outcome: Availability of a wide range of functions and methods for data analysis and machine learning.

**2. Loading the Data:**

Aim: To read the educational data from a CSV file into a Python DataFrame for analysis.

How Obtained: Using `pd.read_csv(file_path)` where `file_path` is the location of the CSV file.

Outcome: A DataFrame containing the GCSE English and Maths national data ready for exploration and cleaning.

**3. Data Cleaning:**

Aim: To ensure data quality by converting data to numeric where necessary and handling missing values.

How Obtained: By applying `pd.to_numeric()` with `errors='coerce'` to convert data to numeric types and using `dropna()` to remove rows with NaN values.

Outcome: A cleaned DataFrame with numeric values for analysis and no missing data which could bias the results.

**4. Exploratory Data Analysis (EDA):**

Aim: To explore the main characteristics of the data and identify any patterns, relationships, or anomalies.

How Obtained: Through descriptive statistics with `describe()`, visualizations like histograms and bar charts using `matplotlib`, and correlation analysis with `sns.heatmap()`.

Outcome: Insights into the distribution of scores and potential relationships between different demographic factors and GCSE outcomes.

**5. Data Preparation for Machine Learning:**

Aim: To prepare the dataset for machine learning by encoding categorical variables and splitting the data into training and testing sets.

How Obtained: Selecting relevant features, encoding them with `OneHotEncoder`, and splitting the dataset using `train_test_split()`.

Outcome: A processed dataset with features in a format suitable for machine learning and separate training and testing sets.

**6. Model Training:**

Aim: To train a predictive model to estimate GCSE outcomes based on various features.

How Obtained: By constructing a `Pipeline` that includes `OneHotEncoder` and `LinearRegression` and calling the `fit()` method on the training data.

Outcome: A linear regression model trained to understand the relationship between students' characteristics and their GCSE performance.

**7. Model Evaluation:**

Aim: To evaluate the predictive performance of the model.

How Obtained: Using the `predict()` method on the test set and then applying evaluation metrics like `mean_squared_error()` and `r2_score()` to assess accuracy.

Outcome: Quantitative measures of model performance, specifically the MSE indicating prediction error and R-squared indicating the variance explained by the model.

Each point is methodically executed to move from raw data to a validated predictive model, providing a structured approach to data-driven decision-making in the educational context.