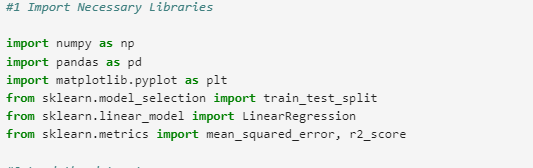
## Import Necessary Libraries



**Explanation**:

Think of these libraries as your toolbox. Each library has specific tools (functions) that help you perform different tasks:

**NumPy**: For numerical operations, like a calculator.

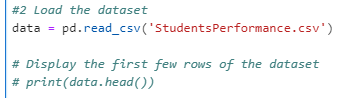
**Pandas**: For data manipulation, like a spreadsheet.

**Matplotlib**: For plotting graphs, like a drawing tool.

**Scikit-learn**: For machine lear2. Load the Datasetning algorithms, like a set of specialized tools for

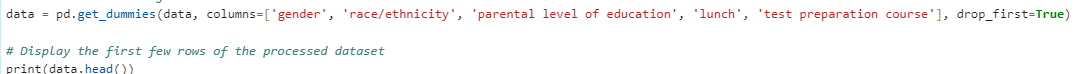
1. **Load the Dataset**

Explanation: Loading the dataset is like opening a book. You need to read the data into a format you can work with. Here, we use Pandas to read a CSV file and display the first few rows to understand its structure.



## Preprocess the Data

Explanation: Preprocessing is like preparing ingredients before cooking. You need to convert categorical variables (like gender, race) into numerical values so the model can understand them. One-hot encoding is a method to do this, similar to labeling different ingredients in a recipe.



**Explanation of pd.get\_dummies**

The pd.get\_dummies function in Pandas is used for one-hot encoding categorical variables. One-hot encoding transforms categorical data into a binary (0 or 1) format, which is essential for many machine learning algorithms that require numerical input.

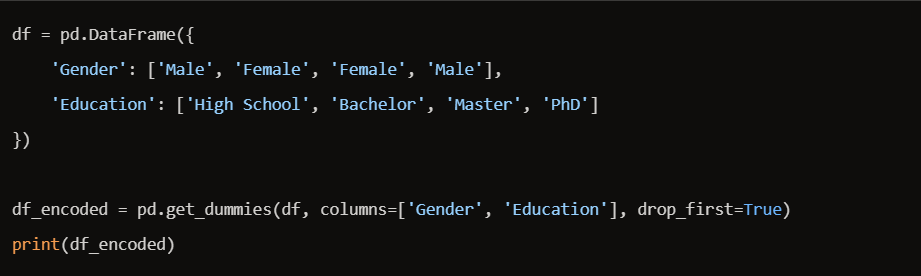
**Here’s a breakdown of what pd.get\_dummies does:**

Converts categorical variables into multiple binary columns: Each unique category in a column becomes a new column.

Binary representation: For each row, the column corresponding to the category is set to 1, and all other columns are set to 0.

Avoids multicollinearity: By setting drop\_first=True, it drops the first category to prevent multicollinearit

***EXAMPLE***

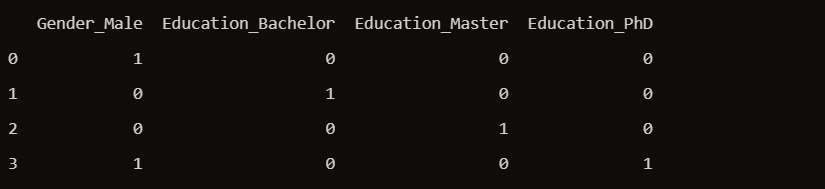


***OUTPUT***

**Multicollinearity**

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, meaning one can be linearly predicted from the others with a substantial degree of accuracy. This can cause problems in regression analysis, such as:

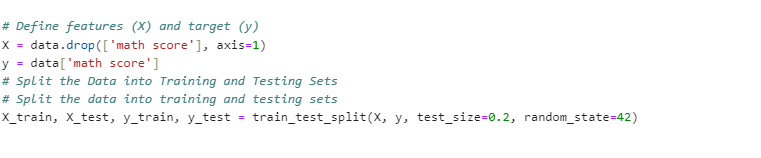
Unstable estimates: Coefficients can change erratically in response to small changes in the model.

Reduced precision: The standard errors of the coefficients increase, making it harder to determine the effect of each predictor.

## Define Features and Target Variable

Explanation: Defining features and the target variable is like deciding what ingredients (features) you need and what dish (target) you want to make. Here, features are the independent variables, and the target is the dependent variable we want to predict (math score).

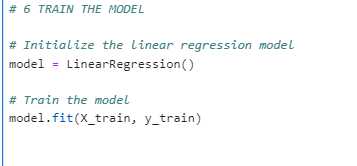
## Split the Data into Training and Testing Sets



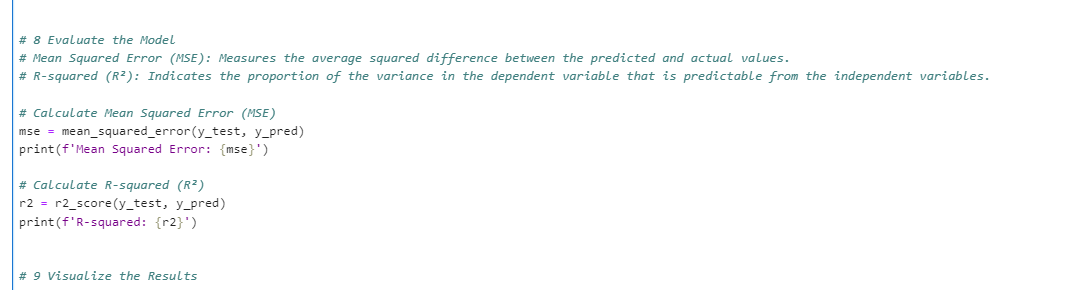
Splitting the data is like dividing your ingredients into two sets: one for practice (training) and one for testing. This helps ensure that your model can generalize well to new, unseen data.

## Train the Linear Regression Model

Explanation: Training the model is like practicing a recipe. You use the training data to teach the model how to predict the target variable. The model learns the relationship between the features and the target.



## Evaluate the Model



Explanation: Evaluating the model is like tasting the dish to see how well you did. Mean Squared Error (MSE) measures how far off your predictions are from the actual values, like checking if the dish tastes right. R-squared (R²) indicates how well the model explains the variability of the target variable, like seeing if the dish meets your expectations.

**Mean Squared Error (MSE) and R-squared (R²)**

Mean Squared Error (MSE)

Definition: MSE measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

**Interpretation: Lower** MSE values indicate a better fit of the model to the data. An MSE of 0 means the model perfectly predicts the data

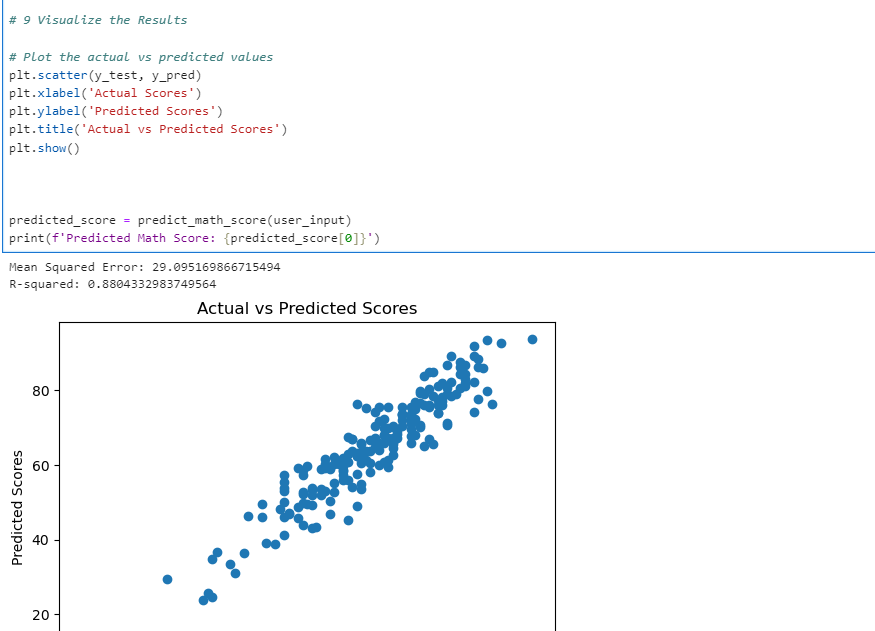
**R-squared (R²)**

Definition: R² is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable or variables in a regression model.

**Interpretation**: R² values range from 0 to 1. An R² of 1 indicates that the regression predictions perfectly fit the data. An R² of 0 indicates that the model does not explain any of the variability of the response data around its mean



## Visualize the Results



xplanation: Visualizing the results is like presenting your dish. You create a plot to compare the actual scores with the predicted scores, helping you visually inspect the model’s performance.