Machine And Data Mining I

Final Report Of Group Project

**Student Performance Prediction**

**Hoàng Đức Nghĩa - 22BI13329** [NghiaHD.22BI13329@usth.edu.vn]

**Nguyễn Duy Nghĩa - 22BI13331** [NghiaND.22BI13331@usth.edu.vn]

**Trần Trọng Nghĩa - 22BI13332** [NghiaTT.22BI13332@usth.edu.vn]

**Lê Phúc Nguyên - 22BI13338** [NguyenLP.22BI13338@usth.edu.vn]

**Nguyễn Đăng Nguyên - 22BI13340** [NguyenND.22BI13340@usth.edu.vn]

# Table of Content

[Table of Content 2](#_Toc166891920)

[1. INTRODUCTION: 3](#_Toc166891921)

[1.1 Background and Problem Statement: 3](#_Toc166891922)

[1.2 Objective: 3](#_Toc166891923)

[1.3 Scope of the Research: 3](#_Toc166891924)

[1.4 Methodology Overview: 4](#_Toc166891925)

[2. DATA OVERVIEW: 5](#_Toc166891926)

[3. METHODOLOGY: 7](#_Toc166891927)

[3.1 Importing Necessary Libraries: 7](#_Toc166891928)

[3.2 Data Reading and Processing: 7](#_Toc166891929)

[3.2.1 Set out Categorical and Logical Columns in the Dataset: 8](#_Toc166891930)

[3.2.2 Unique values in Categorical Columns: 8](#_Toc166891931)

[3.2.3 Data Cleaning: 9](#_Toc166891932)

[3.2.4 Calculate Average Score: 9](#_Toc166891933)

[3.3 Data Analysist: 10](#_Toc166891934)

[3.4 Data Encoding: 17](#_Toc166891935)

[3.5 Initialize input and output: 18](#_Toc166891936)

[3.6 Building Test Model Using Linear Regression: 19](#_Toc166891937)

[3.6.1 Linear Regression Model: 19](#_Toc166891938)

[3.6.2 Model Training and Evaluating: 20](#_Toc166891939)

[3.6.3 Comparting Actual Values and Predicted Values: 23](#_Toc166891940)

[3.6.4 Residual Analysist: 27](#_Toc166891941)

[4. CONCLUSION: 29](#_Toc166891942)

[5. REFERENCE: 30](#_Toc166891943)

# INTRODUCTION:

## Background and Problem Statement:

Education plays a crucial role in shaping future opportunities. Integrating data analytics into education enables the identification of factors that influence student performance and enhances educational outcomes. Standardized tests offer a way to measure student capabilities across different subjects, and analyzing these results can help educators design targeted interventions to improve academic achievement.

Despite the availability of numerous educational resources, many students do not reach their full academic potential due to unaddressed diverse needs and overlooked influencing factors. Traditional methods often fall short in fully understanding these variables, which include gender, race/ethnicity, parental education, lunch type, and test preparation courses. A data-driven approach is essential to effectively identify and address the root causes of academic disparities.

## Objective:

This project aims to use data analytics to predict student performance based on demographic and scoring factors. By analyzing a dataset containing information on various factors and scores in speaking, reading, writing, and listening, this research seeks to identify key factors that significantly impact student performance.

Through this analysis, the project intends to provide actionable insights and recommendations for educators and policymakers.

## Scope of the Research:

This study focuses on analyzing student performance data from a specific dataset. It includes demographic factors such as gender, race/ethnicity, parental education levels, lunch types, and participation in test preparation courses. The analysis is limited to the scores in speaking, reading, writing, and listening, as recorded in the dataset.

## Methodology Overview:

This study employs a comprehensive data analytics approach to predict student performance and determine the influence of various factors. The methodology includes several key steps. First, the dataset is loaded, cleaned, and preprocessed by encoding categorical variables and splitting the data into training and testing sets. Next, exploratory data analysis (EDA) is conducted to visualize the distributions of scores, analyze correlations between variables, and identify key factors affecting performance. Separate predictive models are then developed for each of the four scores (speaking, reading, writing, and listening) using demographic and educational factors as inputs, applying machine learning algorithms such as Linear Regression. The performance of each model is evaluated using metrics like RMSE and R² to compare the influence of different factors on each score. Feature importance analysis is conducted to determine which factors have the most significant impact on each score and to compare their relative importance across the four models. Finally, the predicted and actual values are compared using plot diagrams and other methods to vividly illustrate the residuals of the models.

# DATA OVERVIEW:

The dataset used in this analysis is originally from an existing dataset on the Internet [1], but we make some changes on the datasets so that it can fit with our objectives. It consists of information on student performance, with a total of 1000 entries and 9 columns. The columns in the dataset include both categorical and numerical data:

**Categorical Columns:**

* gender: The gender of the student (male or female).
* race/ethnicity: The race or ethnicity of the student, categorized into groups (A, B, C, D, E).
* parental level of education: The highest education level achieved by the student's parents.
* lunch: The type of lunch received by the student (standard or free/reduced).
* test preparation course: Whether the student completed a test preparation course (none or completed).

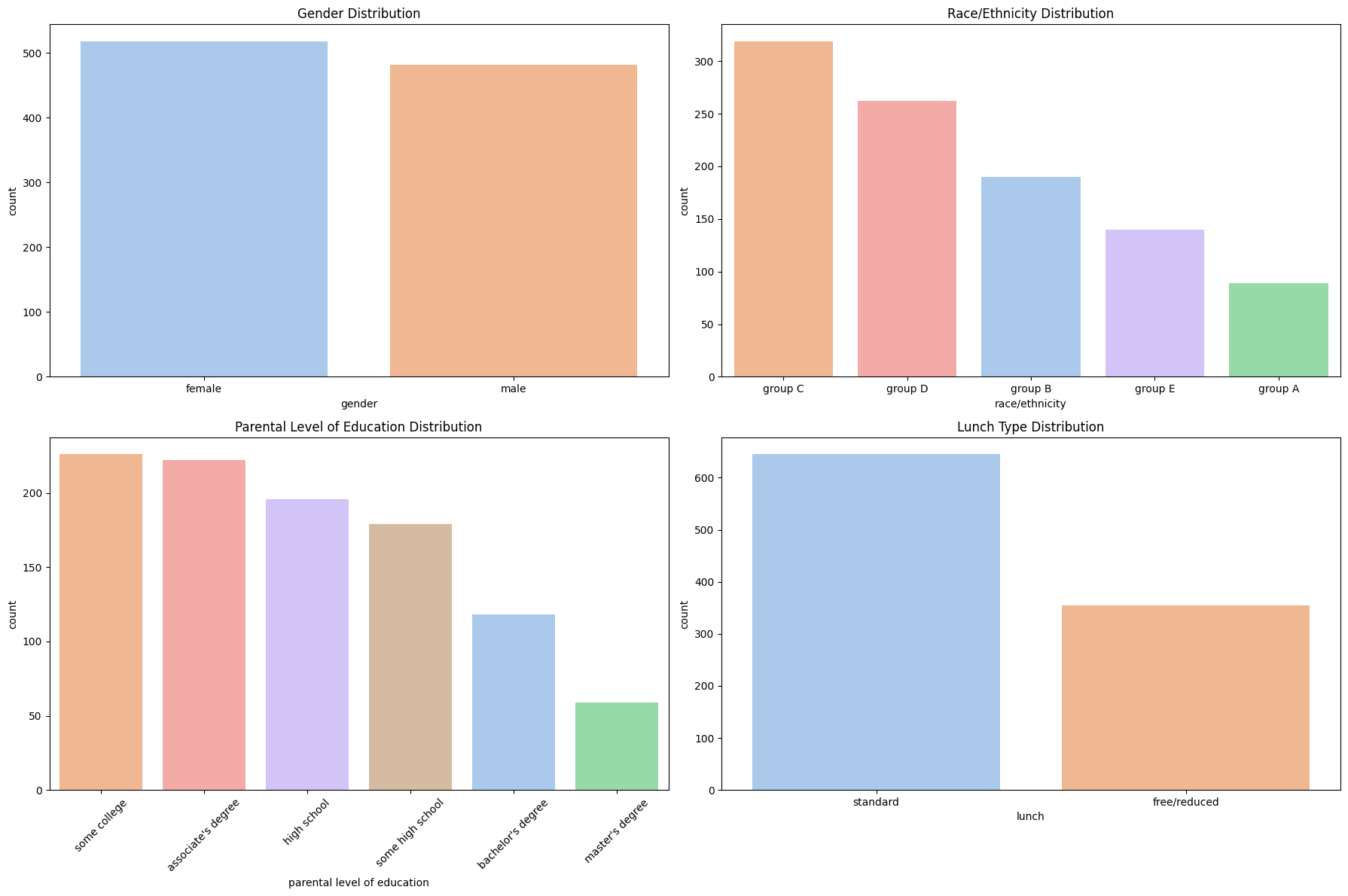
**Numerical Columns:**

* speaking score: The student's score in speaking.
* reading score: The student's score in reading.
* writing score: The student's score in writing.
* listening score: The student's score in listening.

**Sample Data**

****

**Count Plot for Categorical Values:**



This dataset provides a comprehensive overview of various factors related to student performance, including demographic details, parental education, and different scores across various skills. The rich variety of data points makes it suitable for regression analysis to identify key factors impacting student performance and predict average scores.

**Data Usage for Model Training and Testing:**

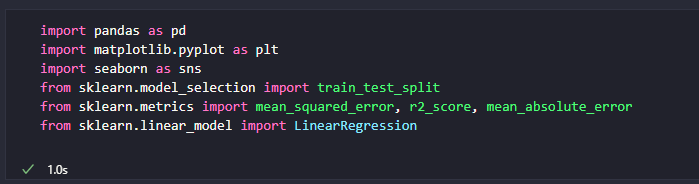
For the purpose of training and testing the linear regression model, the dataset is divided into two subsets:

* Training Set: Used to train the model, which take 80% of the dataset
* Testing Set: Used to evaluate the model’s performance on unseen data, which take up 20% of the dataset.

# METHODOLOGY:

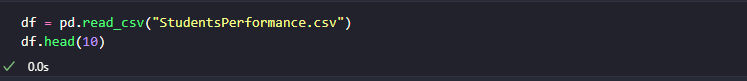
## Importing Necessary Libraries:

Before working with the dataset, we need to use the following libraries in Python:



## Data Reading and Processing:

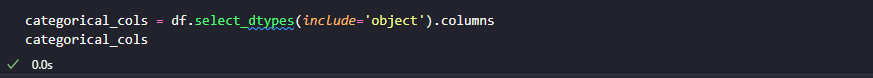
First, we need to read the data from our dataset which is the CSV file. After that, we will extract the first 10 rows of data and print it out using head() function with parameter 10.

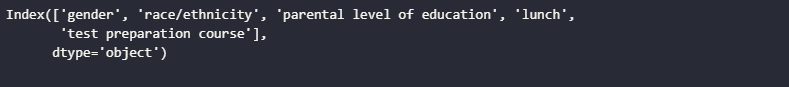




### Set out Categorical and Logical Columns in the Dataset:

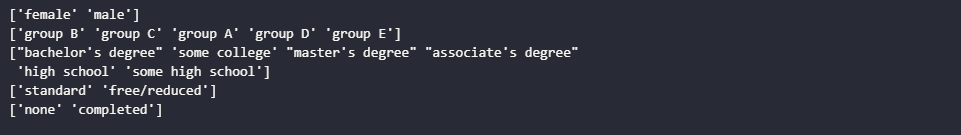
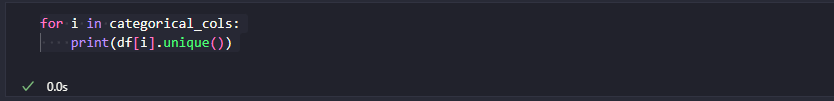
In this step, we select all columns in the DataFrame **‘df’** which have data type **‘object’**. As in **pandas** module, columns with data type **‘object’** typically contain categorical data. After that, **“.columns”** means extracting the names of the selected columns above. Then we assign it to **“categorical\_cols”** and print it out as you can see on the picture

.



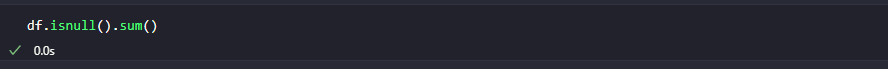
### Unique values in Categorical Columns:

Next, for each categorical column **“i”**, this code will retrieve the unique values present in that columns using the **“.unique()”** method and prints them out. For example, in **Gender** column , we will have two unique values which are ‘female’ and ‘male’ as you can see on the output below.



### Data Cleaning:

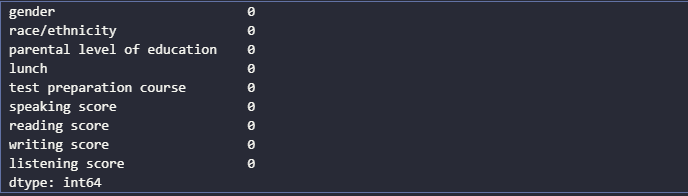
As we want to use this dataset to train our model, the first thing we need to check is the accuracy and stable dataset. Therefore, we have to verify that our dataset does not have any null values or missing values. With incomplete data on the dataset, the processing of the model training will be interrupted or not giving a good output.



First **“df.isnull()**”, it will create a same shape DataFrame with the ‘df’ where each element is Boolean. If there are any missing value, it will return **‘True’** and **‘False’** otherwise.

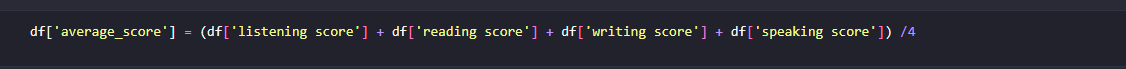
Then **“.sum()”** will sum up all the **‘True’** values which means it will try to find if there any missing value for each column.\

Result:



As our dataset has no missing or null values, we don’t need go through dropping phase which will drop the unnecessary or error features with missing values.

### Calculate Average Score:



As we need to predict the performance of the student, **“average\_score”** will be necessary for constructing output and it can be calculated by sum of four scores then divise by 4. Now we have another column name **“average\_score”.**

The new Dataset after constructing **“average\_score”** column:



## Data Analysist:

After successfully loading the data and make some initial changes to it, we need to analyze the dataset and giving some insight, identifying strengths, areas for improvement as well as how the other columns affect the whole dataset. Starting with the distribution of the score.

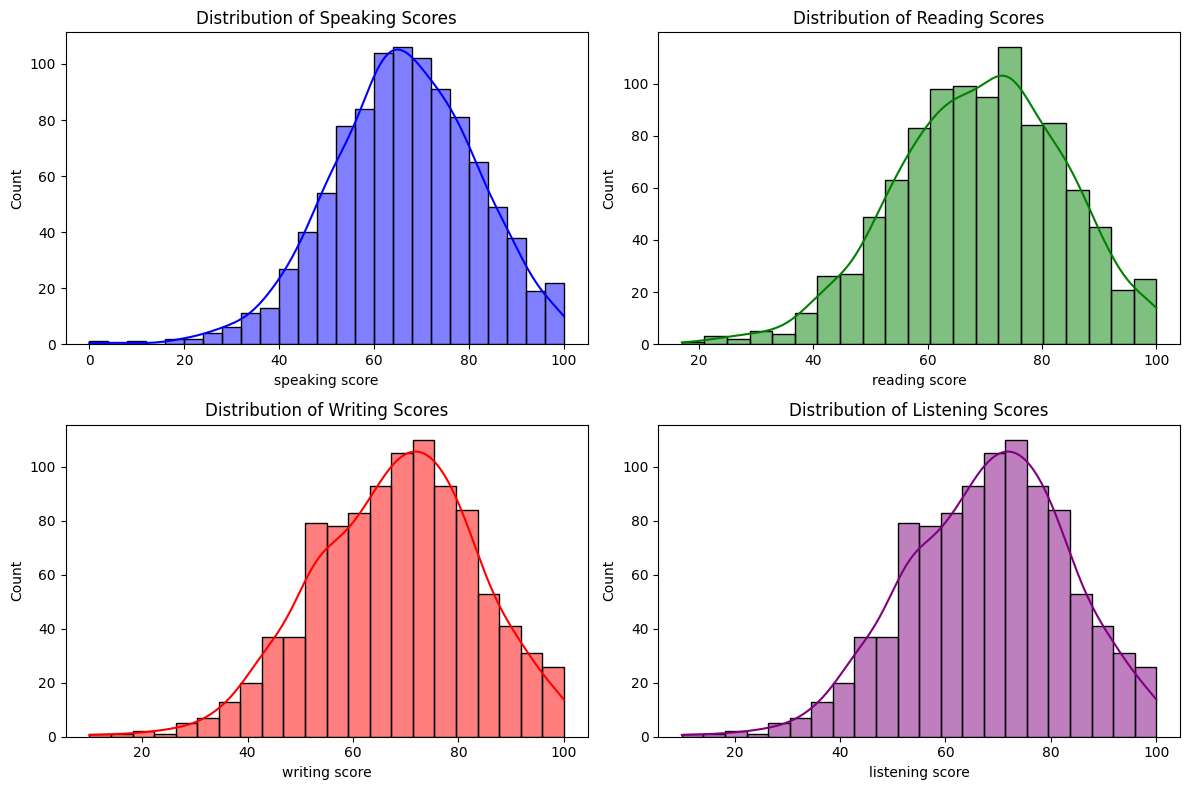


Figure 1: Distribution of each Score

The score distribution for speaking, reading, writing, and listening are similar in shape, central tendency, spread, and skewness, with a slight tendency for higher scores to be more spread out.

* All distributions exhibit a central peak around the 60-70 range, suggesting that this score range is common for all skills.
* Looking at the graph, the histograms of four score all show a slight positive skewness, with the right tail (higher score) being longer than the left tail (lower score).
* The range of scores is quite similar across all skills, spanning from 10 to 100 with a majority of scores between 40 or 80.

Next, we are going to analyze the distribution of average score according to each factor in the dataset.

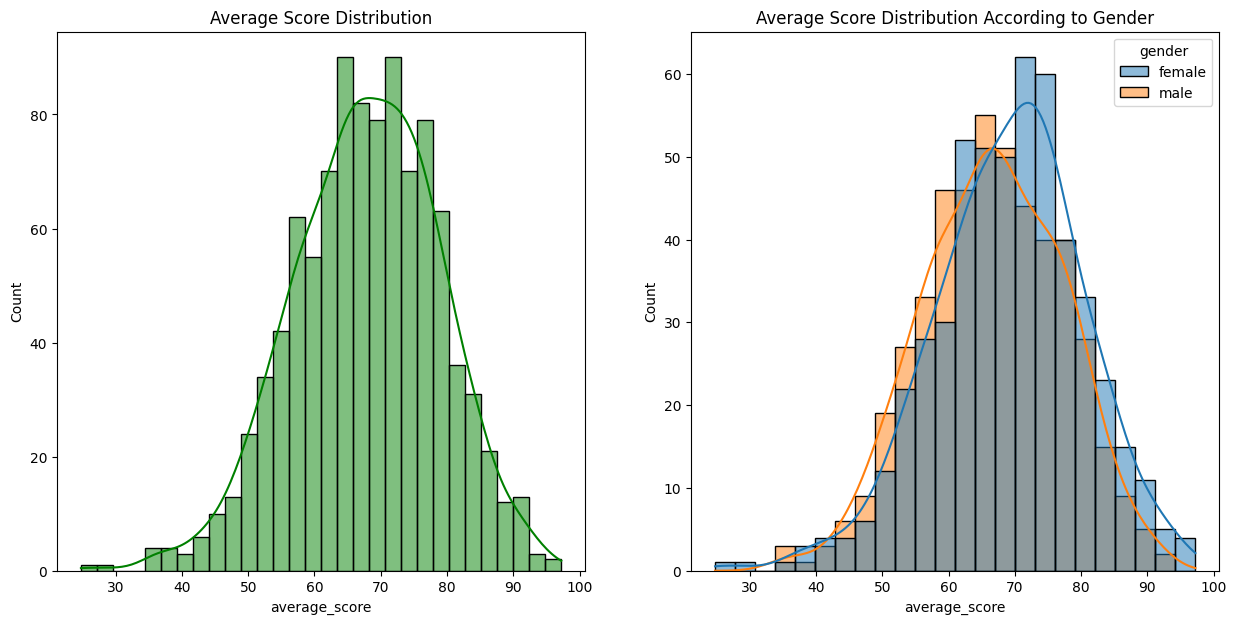


Figure 2: Distribution of Average Score According to Gender

**Overall Performance:** The general distribution of average scores is normal with a slight positive skew, indicating that most students score around the 70-75 range with some higher outliers.

**Gender Comparison:** Female students tend to have slightly higher average scores compared to male students, as indicated by the higher peak in their distribution. Female scores are more concentrated around the higher end (70-75), while male scores are more spread out with a noticeable skew towards lower scores.

I**mplications:** These observations suggest that female students, on average, perform slightly better than male students. The higher concentration of female scores around the 70-75 range and the more pronounced skew in male scores highlight this difference.

* **Female student tends to perform better than male student.**

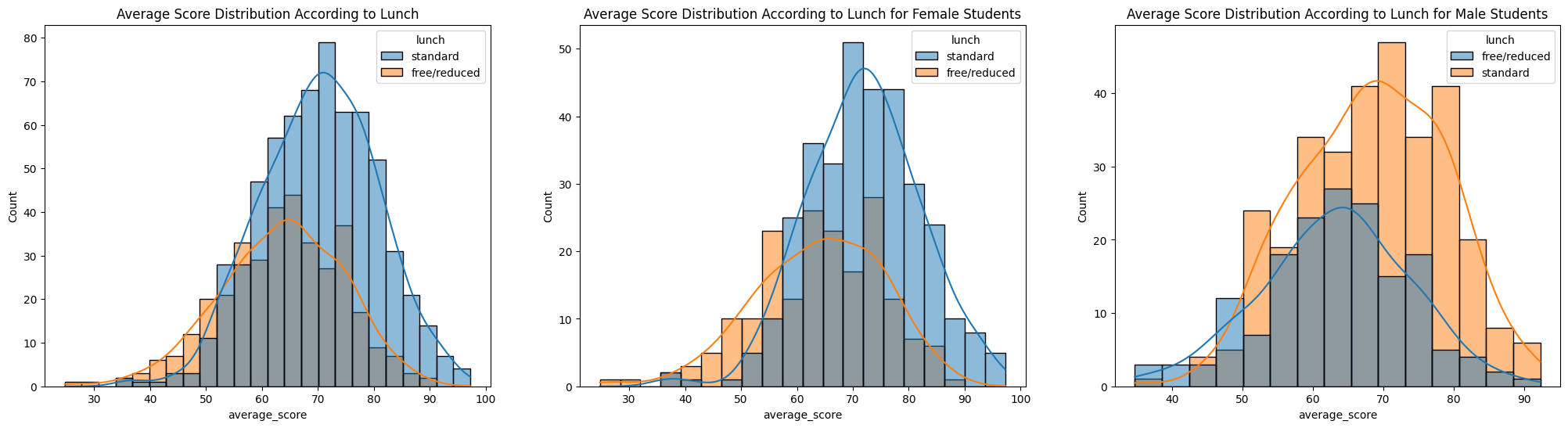


Figure 3: Distribution of Average Score According to Lunch

I**mpact of Lunch Type:** Across both genders, students who receive standard lunch tend to score higher on average compared to those receiving free/reduced lunch. This trend is consistent for both male and female students.

**Gender Comparison:** Both female and male students with standard lunch have higher average scores and less skewness in their distributions, indicating a more uniform performance. Conversely, students with free/reduced lunch have a wider spread of scores and more pronounced positive skewness, suggesting more variability and a tendency for lower scores.

**Overall Insight:** The type of lunch (standard vs. free/reduced) appears to be a significant factor in student performance, with standard lunch associated with higher average scores and less variability.

* **Standard lunch helps student perform well in exams.**

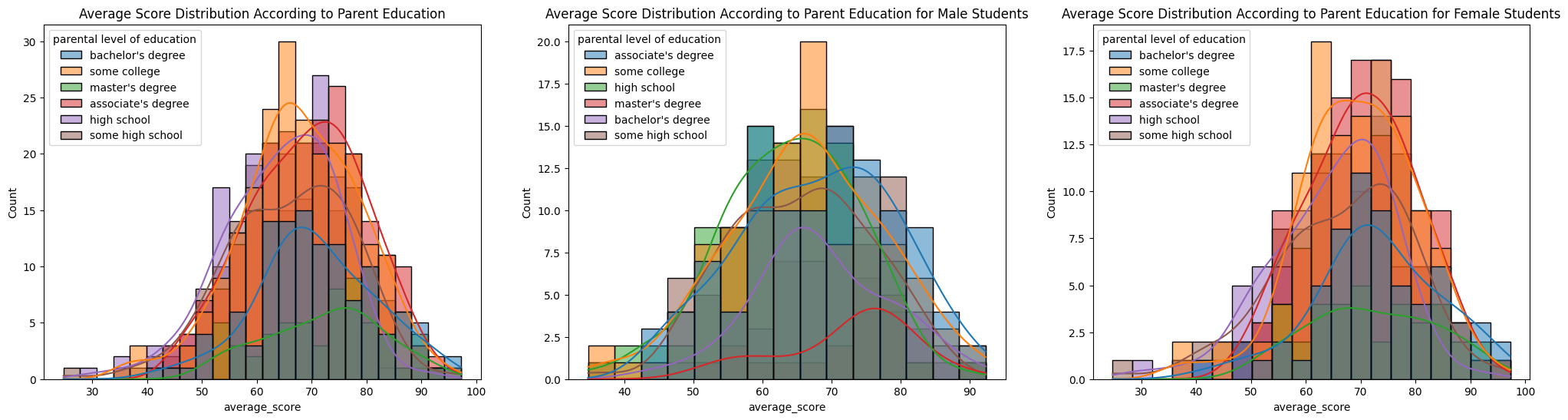


Figure 4: Distribution of Average Score According to Parent Education

**Impact of Parental Education:** Across all students, higher parental education levels (master's and bachelor's degrees) are associated with higher average scores. This trend is consistent for both male and female students.

**Gender Comparison:** Both male and female students show similar patterns in score distribution according to parental education, with slight differences in peaks and spread.

**Socioeconomic Influence:** The data suggests that parental education level is a significant factor in student performance. Students with parents who have higher education levels tend to score higher on average.

**Consistency:** The trend is consistent across genders, indicating that the influence of parental education on student performance is uniform.

* **Higher parental education levels will likely encourage their child to perform better on the exams.**

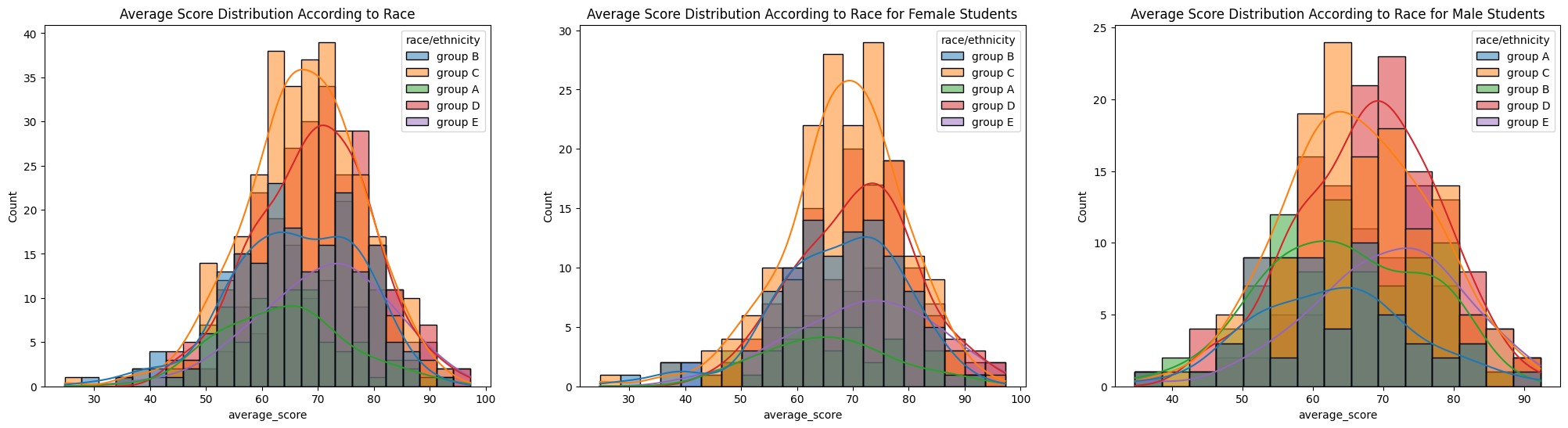


Figure 5: Distribution of Average Score According to Race/ Ethnicity

**Impact of Race/Ethnicity:** Across all students, Group E and Group D tend to have higher average scores compared to other groups. This trend is consistent for both male and female students.

**Gender Comparison:** Both male and female students show similar patterns in score distribution according to race/ethnicity, with Group E and Group D consistently performing better.

**Socioeconomic Influence:** The data suggests that race/ethnicity is a significant factor in student performance, with certain groups (E and D) showing higher average scores.

**Consistency:** The trend is consistent across genders, indicating that the influence of race/ethnicity on student performance is uniform.

* **Based on their race and ethnicity, the performance of the student is also different as group D and E will perform better than the others race.**

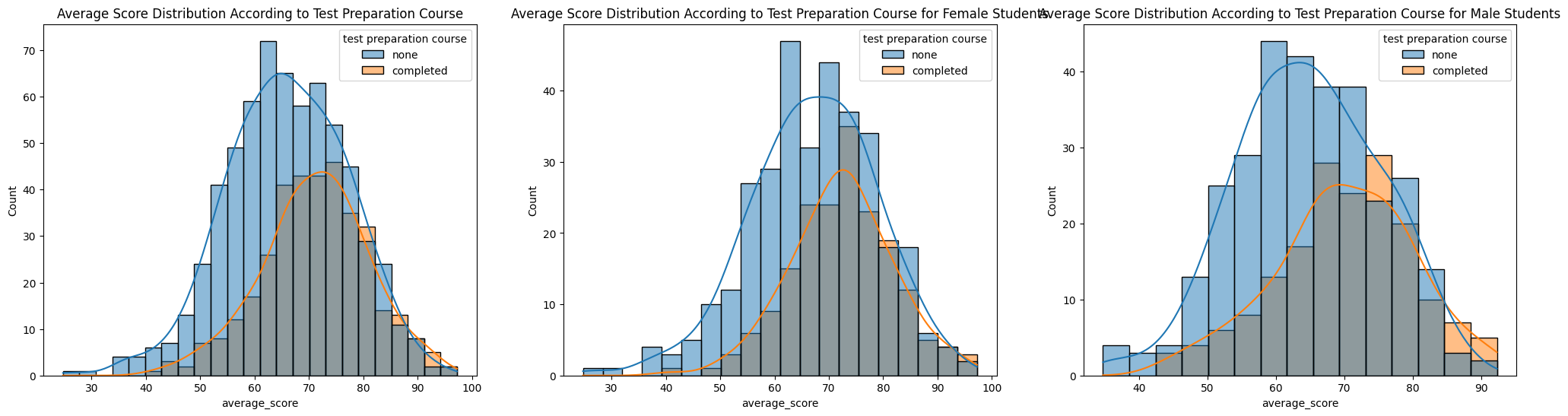


Figure 6: Distribution of Average Score According to Test Preparation Course

**First Histogram:** It has a trend that students who completed the test preparation course score higher on average. As students don’t take courses, scores are more spread out and skewed toward lower values, otherwise, scores are more concentrated around higher values.

**Second and Third Histogram:** Female and Male students who completed the test preparation course have higher scores. Without a course, scores are also more common into lower values. On the other hand, with the course, scores are concentrated more around higher values.

* **Student who completed the test preparation course tend to perform better than who don’t.**

**Correlation Matrix** is a table showing correlation coefficients between sets of variables. It allows us to study both strength and direction of the relationship between sets of variables.

**Type of Correlation:**

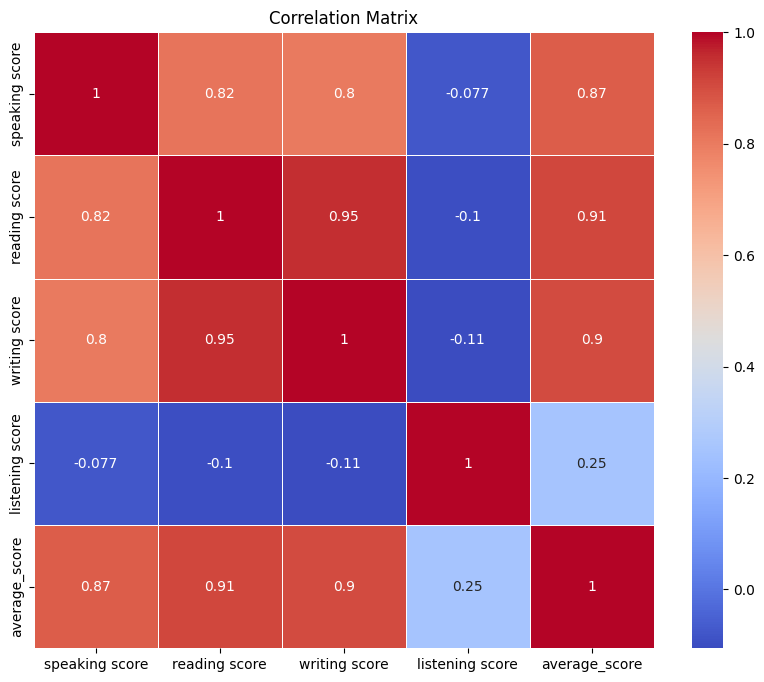
* **Positive Correlation (close to 1):** Indicates a direct relationship; as one variable increase, the other also increases.
* **Negative Correlation (close to -1):** Indicates an inverse relationship; as one variable increases, the other decreases.
* **No Correlation (close to 0):** Indicates no relationship between the variables.

Figure 7: Correlation Matrix for Scores

In this correlation matrix, it shows the correlation coefficients between speaking, reading, writing and listening scores, along with the overall average score.

* **Speaking Score:**
  + Strong positive correlation with average score (0.87), reading score (0.82), and writing score (0.8)
  + Minor negative correlation with listening score (-0.077)
* **Reading Score:**
  + Very strong positive correlation with writing score (0.95),

and average score (0.91).

* + High positive correlation with speaking score (0.82)
  + Minor negative correlation with listening score (-0.1)
* **Writing Score:**
  + Very strong positive correlation with reading score (0.95),

and average score (0.9).

* + High positive correlation with speaking score (0.8).
  + Slight negative correlation with listening score (-0.11).
* **Listening Score:**
  + Low positive correlation with average score (0.25).
  + Negligible correlations with speaking (-0.077), reading (-0.1), and writing scores (-0.11).
* **Average Score:**
  + Strong positive correlations with speaking (0.87), reading (0.91), and writing scores (0.9).
  + Low positive correlation with listening score (0.25).

#### **CONCLUSION:**

**High Inter-Correlation Among Core Skills:** There are strong positive correlations among speaking, reading, and writing scores, indicating that students who perform well in one of these areas tend to perform well in the others.

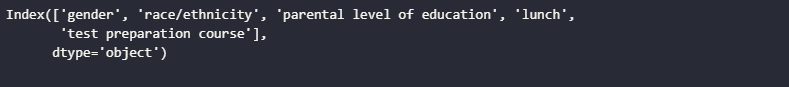
**Listening Score's Weak Correlation:** The listening score shows a low positive correlation with the average score and negligible correlations with the other individual scores. This suggests that listening skills might be influenced by different factors compared to speaking, reading, and writing skills.

**Overall Average Score:** The average score is most strongly influenced by the reading score (0.91) and least by the listening score (0.25), indicating that reading performance has the greatest impact on the overall average score.

## Data Encoding:

This step is a crucial preprocessing step in machine learning that transforms raw data into a suitable format for model training. Effective data encoding can significantly enhance the performance and accuracy of machine learning algorithms by ensuring that all data is numerical and standardized.

In step [**2.2.1**](#_Set_out_Categorical)**,** we have already set out the columns which are categorical and numerical.



Now we need to transform these categorical column’s data into a numerical one.

For **‘gender’**, ‘male’ will be 1 and ‘female’ will be 0.

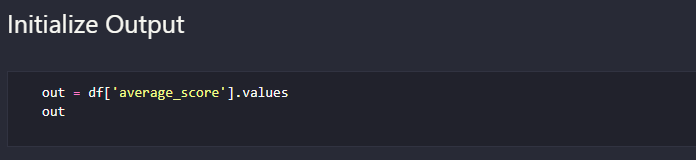
For **‘race’**, ‘group A’ is 0, ‘group B’ is 1, ‘group C’ is 2, ‘group D’ is 3, and ‘group E’ is 4.

For **‘education level’**, “bachelor's degree" is 0, 'some college' is 1,"master's degree” is 2, "associate's degree" is 3, "high school" is 4, "some high school" is 5.

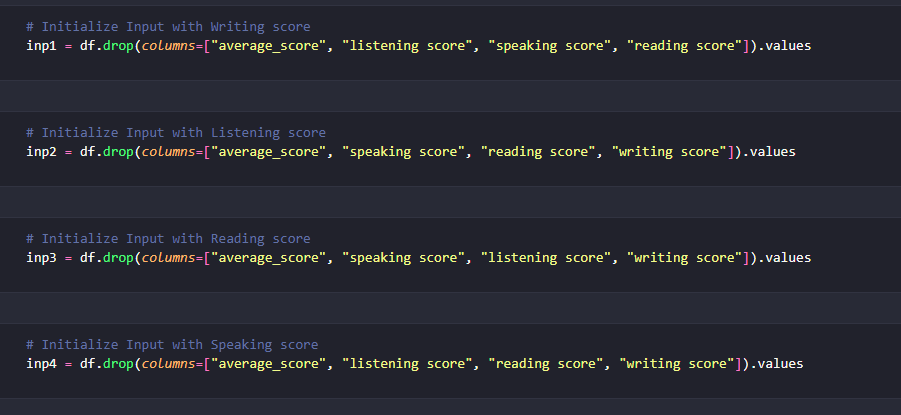
Now the dataset will be ready to apply in training and testing in machine learning model.

## Initialize input and output:

In this step, we will create four distinct input sets and one output for our machine learning model. Each input set will exclude one of the following: **‘average\_score’**, **‘listening score’**, **‘speaking score’**, or **‘reading score’**. This approach ensures that the model is trained using the categorical columns and the **‘writing score’**. The remaining inputs will include the **‘listening score’**, **‘reading score’**, and **‘speaking score’**. The output variable for the model will be the **‘average\_score’**, which was computed in step [2.2.4](#_Calculate_Average_Score:).



The decision to create multiple input sets stems from the need to identify key factors that significantly impact student performance. By evaluating model performance with different combinations of inputs, we aim to determine which elements are most influential in predicting student performance. This analysis will provide valuable insights into the relative importance of various skills and attributes, allowing for targeted educational interventions and improvements.



## Building Test Model Using Linear Regression:

### Linear Regression Model:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and on or more independent variables.

It aims to find the best-fitting linear relationship, expressed by the equation

**y=β0+β1.x1+β2.x2+...+βn.xn+ϵ**

where:

* y is the dependent value (the output we are trying to predict)
* x1, x2, … xn are the independent variables (the predictors)
* β0 is the intercept
* β1, β2, …βn are the coefficients of the independent variables.
* ϵ is the error term

A diagram of a line of regression

Description automatically generated

Figure 8: Linear Regression Model

In the above figure, we have:

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the ‘best fit line.’ The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Why we choose Linear Regression?

* As my project aims to identify key factors impacting student performance, linear regression is well-suited as it quantifies the relationship between predictors (categorical column, scores, …) and outcome (average score).
* By creating multiple input sets and evaluating their impact on the average score, linear regression helps determine which variables are most significant, from that to identify key performance factors.
* Last but not the least, it is easy to implement and interpret when dealing with multiple sets of predictors.

### Model Training and Evaluating:

The provided function **train\_and\_evaluate** is designed to train and evaluate a linear regression model using various sets of input features to predict the output variable, which in this case is the **average\_score.** The function begins by initializing an empty dictionary, results, to store the evaluation metrics and other results for each input set.

The function accepts four parameters: **inputs**, **output**, **test\_size**, and **random\_state**. The inputs parameter is a list of different input datasets (features), while output is the target variable (average\_score). The **test\_size** parameter defines the proportion of the dataset to include in the test split, with a default value of 0.2 (or 20%). The **random\_state** parameter ensures the reproducibility of the results by setting a seed for the random number generator, with a default value of 42.

Within the function, a loop iterates over each input dataset in the inputs list. For each dataset, the data is split into training and testing sets using the **train\_test\_split** function from the **sklearn.model\_selection** module. The training data consists of **inp\_train** and **out\_train**, while the testing data consists of **inp\_test** and **out\_test.**

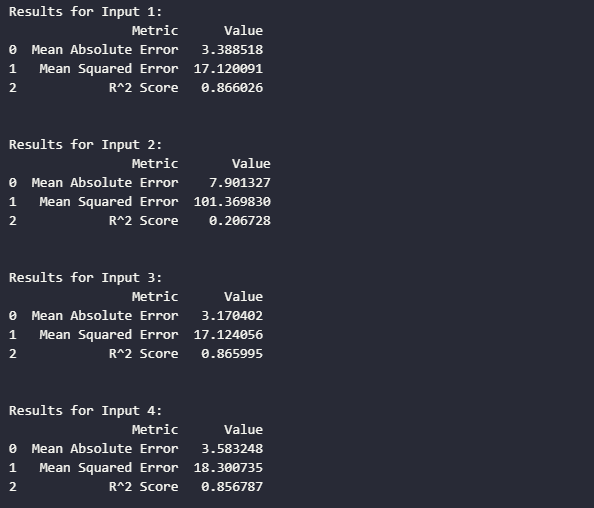
A linear regression model is then initialized and trained using the training data. The model is fitted with **inp\_train** and **out\_train** using the fit method. After training, the model makes predictions on the test data **(inp\_test)** using the predict method.

The function then evaluates the model's performance by calculating three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score. The MAE is the average of the absolute errors between the actual and predicted values, the MSE is the average of the squared errors, and the R^2 Score represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

These evaluation metrics, along with the actual values, predicted values, and residuals (differences between actual and predicted values), are stored in the results dictionary under a key corresponding to the current input set. A DataFrame metrics is created to display the evaluation metrics in a tabular format, which is then printed to provide a summary of the model's performance for each input set.

Finally, the function returns the results dictionary containing all the evaluation metrics and related data for each input set. This systematic approach allows for a comprehensive understanding of which input features are most significant in predicting the target variable, thus aiding in the identification of key factors that impact student performance.

Evaluating the model:



By analyzing the performance of the linear regression model with different input sets, we observed significant variations in the evaluation metrics. The models trained on inputs excluding **writing score, reading score, and speaking score** (Inputs 1, 3, and 4) performed similarly well, as indicated by their high R^2 Scores (around 0.86-0.87) and low MAE and MSE values. However, the model trained on input excluding **listening score** (Input 2) performed significantly worse, with a much lower R^2 Score (0.207) and higher MAE and MSE values. This suggests that the listening score is a critical factor in predicting the average\_score. Overall, these insights help identify the key predictors of student performance, guiding targeted educational interventions.

### Comparting Actual Values and Predicted Values:

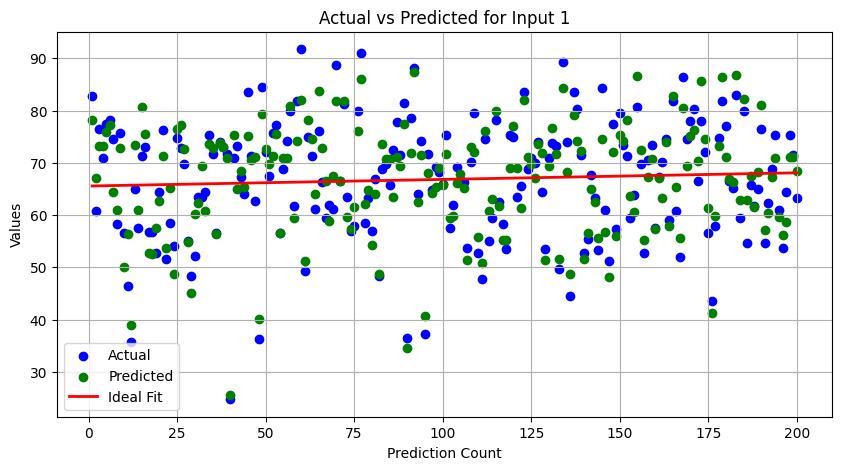


Figure 9: Actual and Predicted Values for Input 1(Excluding Writing score)

The scatter plot for Input 1 shows a moderate alignment between the actual and predicted values. The green dots (predicted) are generally close to the blue dots (actual), particularly in the mid-value range (50-80). However, there are some deviations where the predicted values do not align perfectly with the actual values. The ideal fit line in red demonstrates the expected linear relationship, indicating that while the model performs reasonably well, there is still room for improvement in capturing the variance in actual scores.

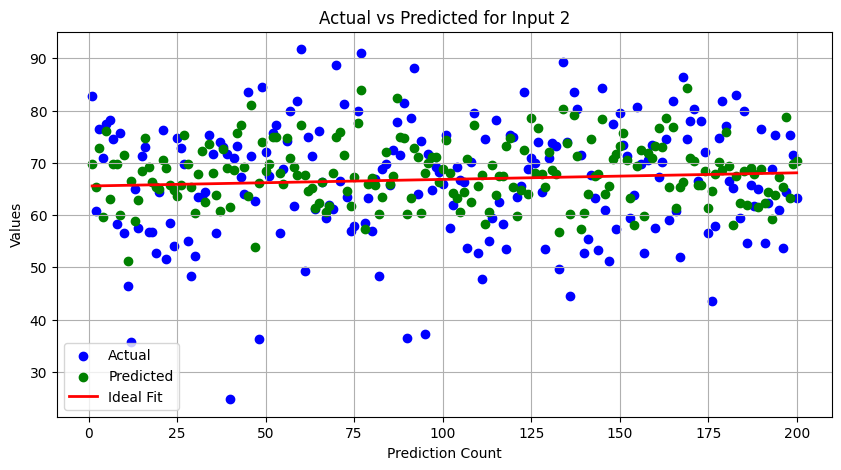


Figure 10: Actual and Predicted Values for Input 1(Excluding Listening score)

For Input 2, the scatter plot reveals a significant discrepancy between the actual and predicted values. The green dots are more scattered and do not align well with the blue dots, especially in the mid-range values (40-70). This scatter reflects the lower R^2 Score of 0.207 observed earlier, indicating that the model struggles to predict the average score accurately without the **listening score**. The ideal fit line is less effective in demonstrating a strong correlation, highlighting the model's poor performance with this input set.

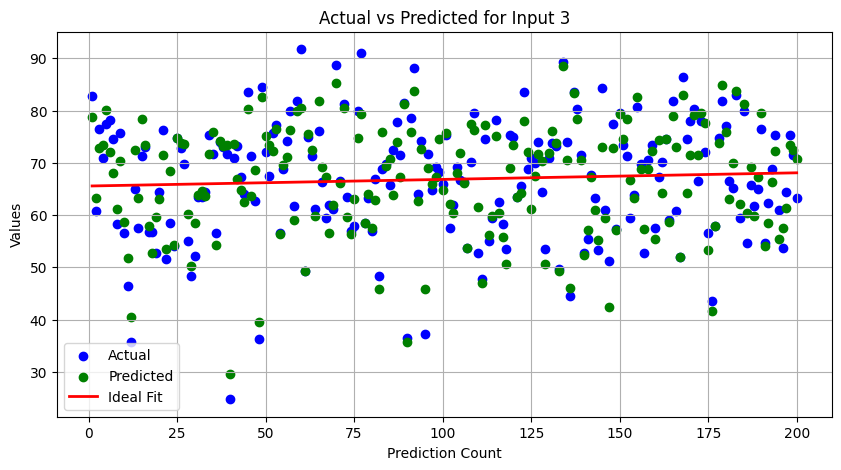
****

Figure 11: Actual and Predicted Values for Input 1(Excluding Reading score)

The scatter plot for Input 3 shows a stronger alignment between the actual and predicted values, similar to Input 1. The green dots closely follow the blue dots, particularly in the mid to high range (60-80). The ideal fit line reinforces the linear relationship, suggesting that the model can effectively predict the average score even when **reading score** is excluded. This alignment supports the high R^2 Score of 0.866, indicating strong predictive power

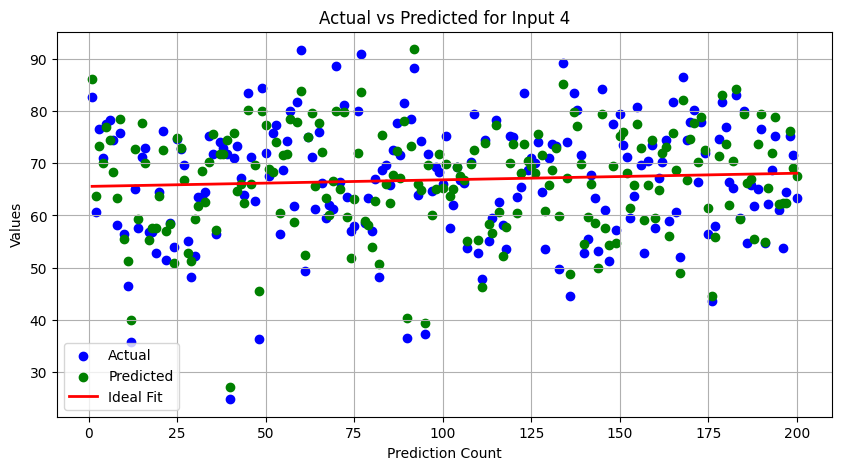
****

Figure 12: Actual and Predicted Values for Input 1(Excluding Speaking score)

In the scatter plot for Input 4, the actual and predicted values also exhibit a good alignment. The green dots follow the blue dots closely, especially in the mid to high range (60-80), indicating accurate predictions by the model. The ideal fit line further emphasizes the linear relationship, consistent with the R^2 Score of 0.857. This suggests that the model performs well in predicting the average score without the **speaking score.**

#### **CONCLUSION:**

Comparing the scatter plots, it is evident that the model performs well with Inputs 1, 3, and 4, where the predicted values align closely with the actual values, particularly in the mid to high score ranges. This alignment reflects the high R^2 Scores for these inputs, indicating strong predictive power. In contrast, the model's performance significantly drops for Input 2, where the predicted values are more scattered and less aligned with the actual values, corresponding to the much lower R^2 Score. This analysis confirms that the **listening score** is a crucial predictor for accurately determining the average score, while the other inputs can still yield strong predictions even when one of the scores is excluded.

### Residual Analysist:

Introduction to Residuals:

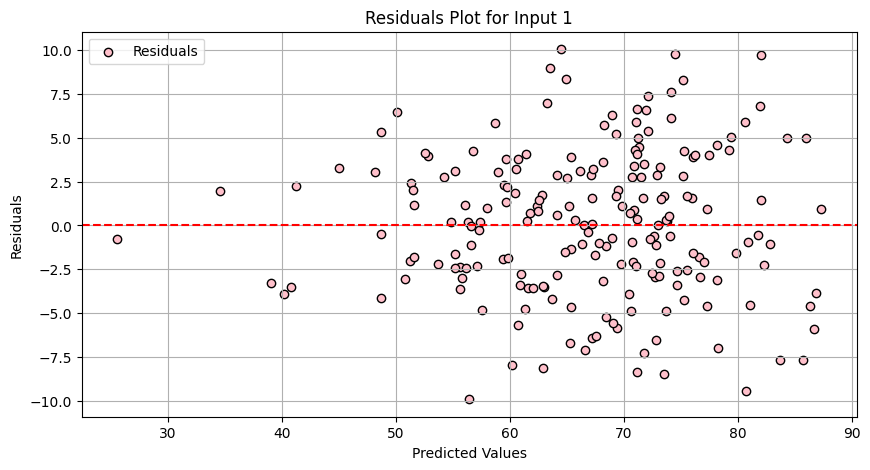
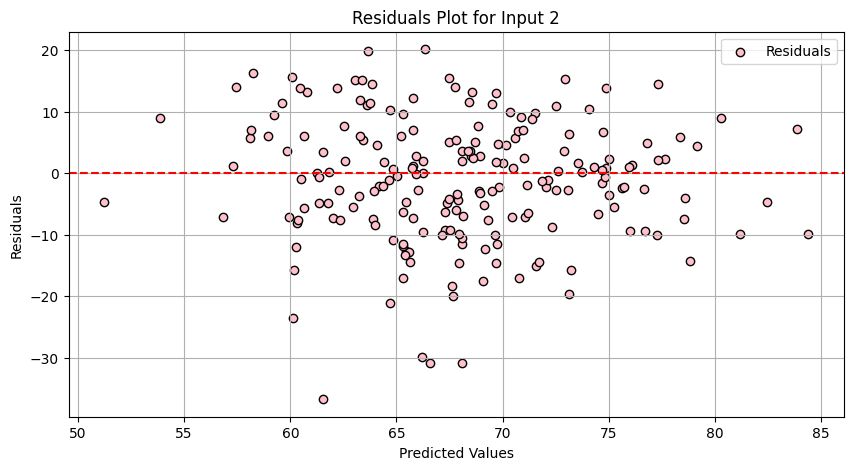
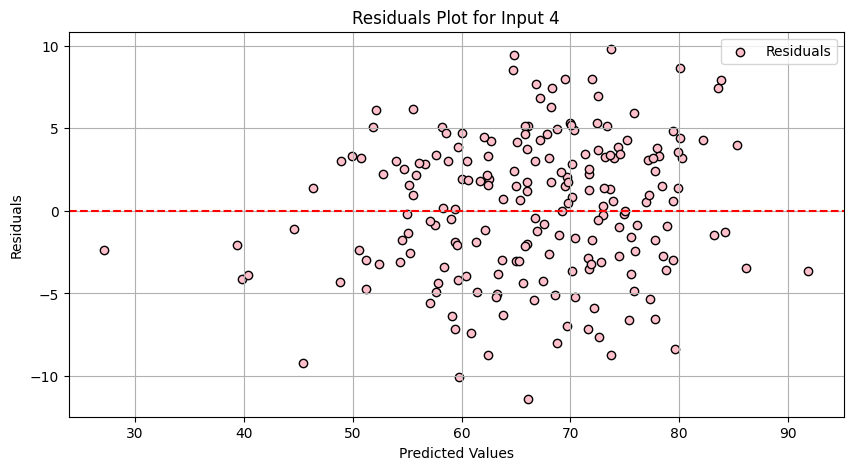
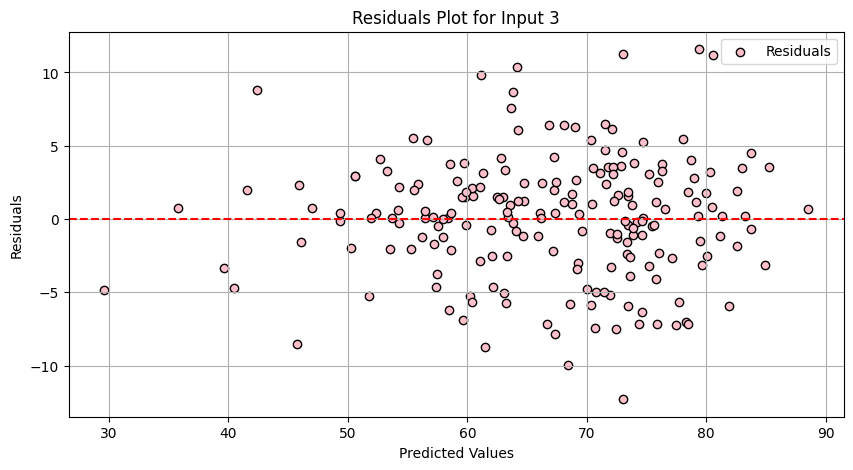
Residuals in regression analysis are the differences between the observed actual values and the values predicted by the model. They are crucial for diagnosing the fit of a regression model. If the residuals are randomly scattered around zero without any discernible pattern, it indicates that the model is well-fitted to the data. However, if there are patterns in the residuals, it suggests that the model is missing some information, indicating non-linearity, heteroscedasticity, or the presence of outliers.

Figure 13, 14, 15, 16: Residuals Plot for Input 1, 2, 3, 4

Comparing the residual plots for the four input sets, it is evident that the models for Inputs 1, 3, and 4 show residuals that are randomly distributed around zero, indicating a good fit. In contrast, the residual plot for Input 2 shows a pattern and a wider spread, suggesting that the model's performance is significantly impacted when listening score is excluded. This analysis highlights the importance of including **listening score** in the predictive model to achieve better accuracy and reliability.

# CONCLUSION:

This project effectively demonstrated how data mining and machine learning techniques can be utilized to predict student performance based on demographic and scoring factors. The findings emphasize the crucial role of the listening score in overall predictive accuracy, along with the significance of demographic factors such as gender, parental education, and lunch type. By identifying these key determinants, the study offers valuable insights for targeted educational interventions and enhancements.

The outcomes highlight the necessity of a data-driven approach in education to effectively address academic disparities and improve student performance. Educators and policymakers can use these insights to develop more informed strategies and initiatives that accommodate the diverse needs of students, ultimately promoting a more equitable and effective educational environment.

# REFERENCE:

[1][Student Performance Dataset Used](https://www.kaggle.com/datasets/spscientist/students-performance-in-exams)

[2][The Source Code I Reference From](https://github.com/harsh0703-harsh/students)

[3][Linear Regression Information And Residuals](http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm)