

# Student Performance Prediction: Data & EDA

---

Date of Presentation : 3/19/2024

Student Name : Alwala Chandrahas Reddy

Student's Pace Email : ca13115n@pace.edu

Address Class Name : CS667

Program Name : MS in Data Science

School Name, University Name : Seidenberg , Pace University




# Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Recommendations and technical next steps

# Executive summary

- There are students who are having different capabilities towards studies and also few students need extra care to do well in studies
- Student performance prediction helps us to know how students might perform well in studies and also how student scores during exams.
- The Student performance prediction is calculated by using Linear regression model by using this model we can predict student math scores.

# Project plan recap

Deliverable	Due Date	Status
Data & EDA	03/19/2024	Complete 
Methods, Findings, and Recommendations	04/02/2024	Complete 
Final presentation		In Progress 

# Data

---

# Data

- **Data details**

- Data source : The dataset used for student performance prediction was collected from Kaggle It contains information about student performance in different subjects
- Sample size : The dataset consists of 1000 rows and 9 columns
- Time period The dataset is from the year 2022
- Data that was included to know the student performance
- Important notes about the data is that data can undergo operations to solve the problem

- **Assumptions**

- If looking at student data we need the parents details about their highest education.

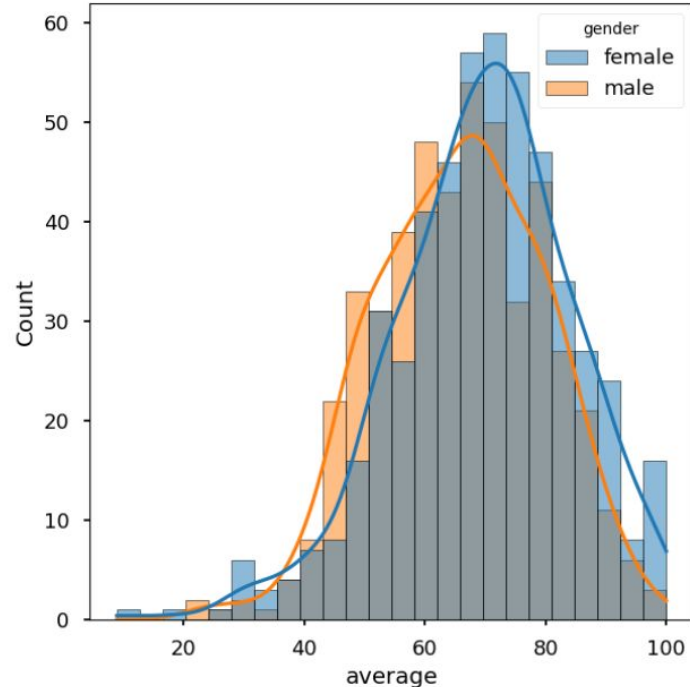
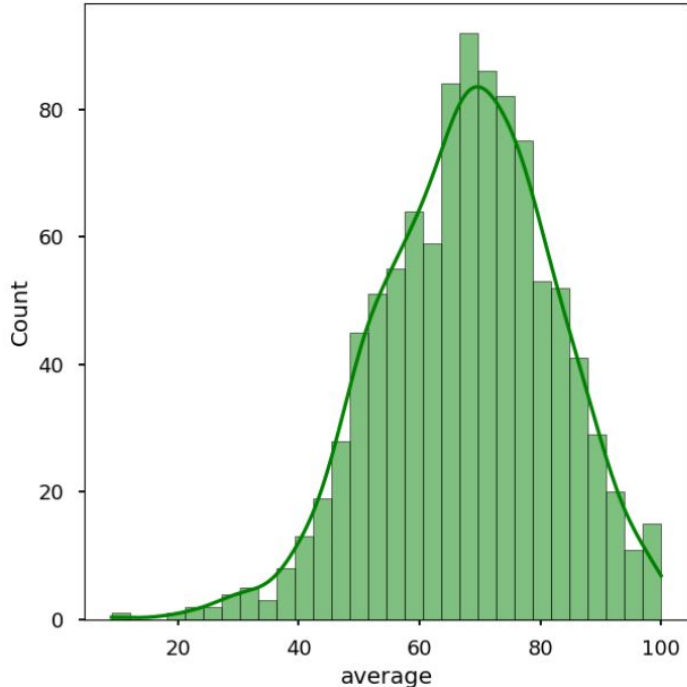
- Dataset link : <https://www.kaggle.com/datasets/abhyudayadubey/students-performance>

# Exploratory Data Analysis

---

# Marks Distribution for all Genders

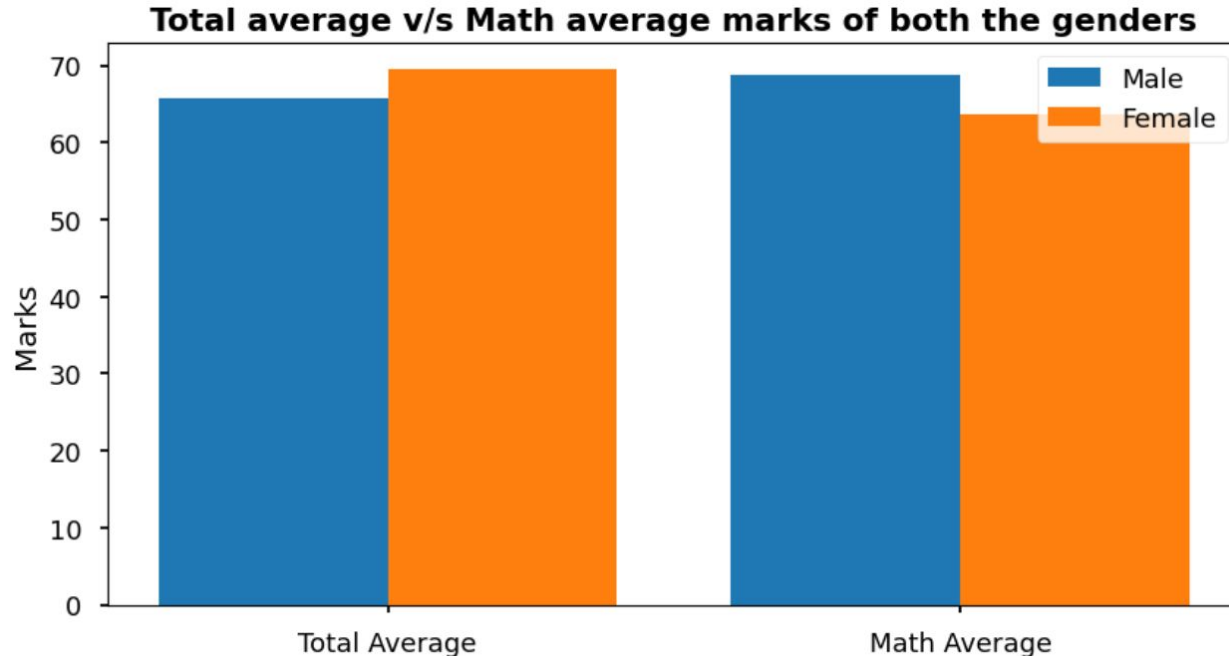
- This plot shows distribution of the average scores of all students
- The second plot shows the distributions of the both genders
- In the first plot it says about the average score of the class and the average is 80
- In the second plot it shows about the average marks of both the genders the average marks is more for female than male





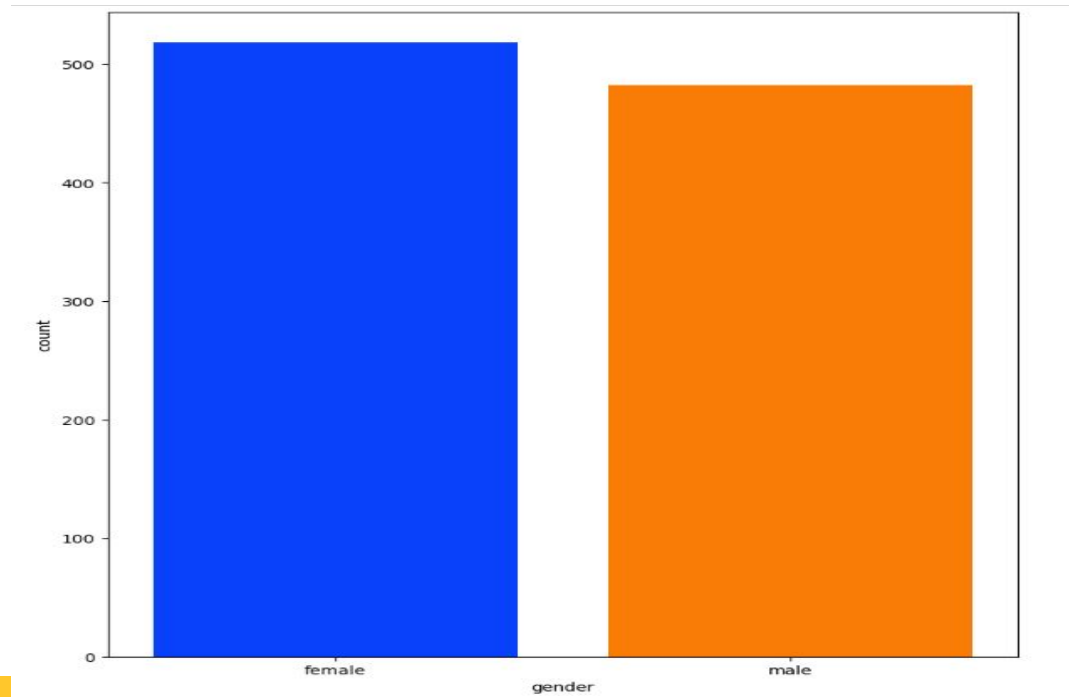
# Math average and Total Average between Genders

- This visualisation shows the total average marks between both genders
- The second chart shows the Math average marks of both genders
- Total average is higher for female than male and in second chart the math marks is higher for male than female



# Gender Average Comparison

- This is the Chart that shows about the count of Male and Female who wrote the test
- The Bar Chart in Blue colour shows the count of Female and the orange colour shows male count



# Test Completion Course Overview

- The Chart shows the Percentage of students who completed test\_course and also it shows the percentage of students not completed the test\_course

(-1.25, 1.25, -1.25, 1.25)

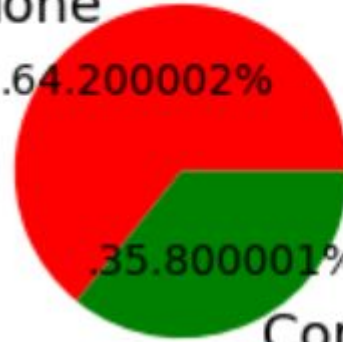
Test\_Course

None

.64.200002%

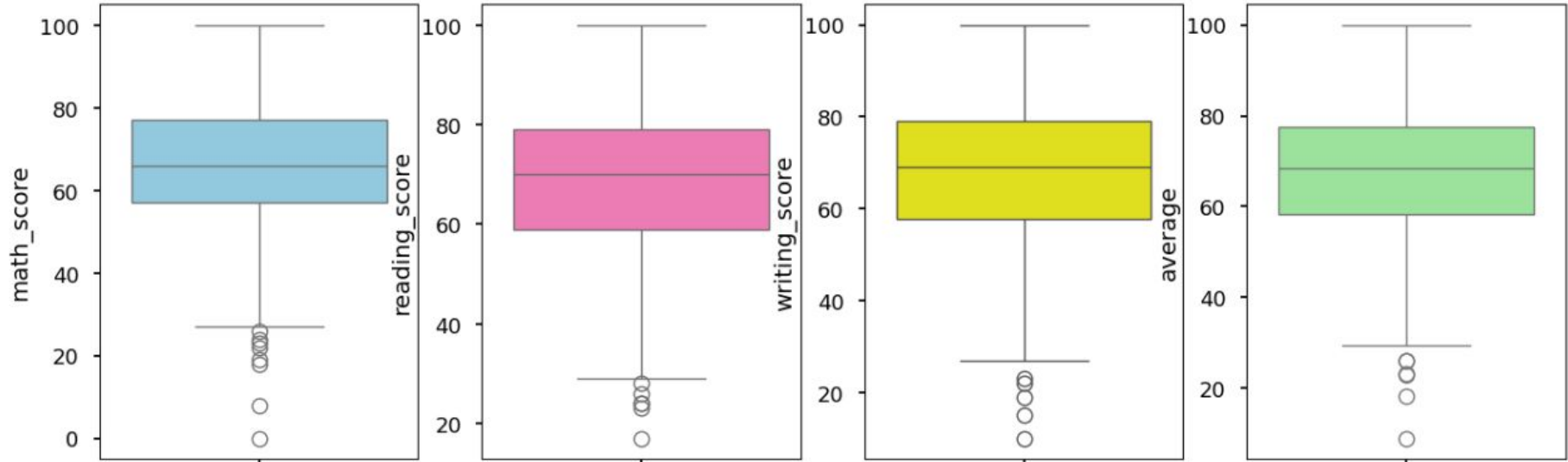
.35.800001%

Completed



# Identifying Unusual Data Points

- This visualisation shows about the unusual data points in the attributes of math\_score, reading\_score, writing\_score, average
- The unusual points are those which are not related and are different from the rest of the data



# Modeling methods

---

# Modelling Methods

Feature Table

Predictor Name	Predictor Description	Rationale
Student Background	Gender and ethnic background	
Parental level of education	Parental education level are masters,bachelor	Parental education reflects learning resource at home
School Resources	Access to lunch program are standard,free/reduced	The lunch program could indicate a student economic background which could be linked to educational opportunities
Test Preparation course	Whether student took a preparation course or not	Test preparation course are for those which impacts student performance in math subject

# Modeling methods

## Outcome Variable:

- In this the problem the outcome variable is the `math_score` the other attributes in the dataset helps in predicting the maths score of a gender.

To predict math scores, we considered several factors that might be related to a student's performance. These include:

- Student background: Gender and racial/ethnic background
- Parental\_level\_education : Parental education level
- School resources: Access to lunch program (free or paid)
- Preparation: Whether the student took a test preparation course

# Features

- The main features that I selected in this are the gender, parental\_level\_of\_education, test preparation course, reading score, writing score
- Parental education level might reflect access to learning resources at home.
- The lunch program could indicate a student's socioeconomic background, which can sometimes be linked to educational opportunities.
- Additionally, test preparation courses, while potentially targeted towards math, might also develop general study skills that benefit performance in reading and writing



## Model type and rationale for choosing this type of model

- We used a technique called **linear regression**. Imagine a separate straight line is created for each outcome variable (math score, reading score, writing score).
- In other terms linear regression is a way to explain the relationship between the dependent variable and one or more explanatory variable using a straight line .
- Each line helps us see the general trend between the features we considered (like parental education) and the specific score we're trying to predict.
- The steeper the slope of the line, the stronger the relationship between the feature and that particular score.

# Findings

---

# Findings

- From my analysis I find that the marks are more for female and from my visualisation plots the maths marks are more for the male and when the total average score are more for female when compared to men
- Factors such as gender, race\_ethnicity, parental level of education, lunch type, test preparation course, as well as reading and writing scores, significantly influenced the accuracy of our predictions.
- Students who have completed the Test Preparation Course have scores higher in all three categories than those who haven't taken the course
- The score of student whose parents possess master and bachelor level education are higher than others
- Terms used for calculating student performance is  $R^2$  score

# Recommendations & Data Science

## next steps

---

# Recommendations and Data Science next steps

## Recommendations

- Factors such as gender, race\_ethnicity, parental level of education, lunch type, test preparation course, as well as reading and writing scores, significantly influenced the accuracy of our predictions.
- The recommendation for this is the focus on implementing inclusive policies and initiatives aimed at providing equitable access to resources and opportunities for all students. This could involve providing support services, resources, and educational programs tailored to the needs of diverse student populations. Additionally, fostering a supportive and inclusive school culture can help create an environment where all students feel valued and empowered to succeed.
- From my analysis I find that the marks are more for female and from my visualisation plots the maths marks are more for the male and when the total average score are more for female when compared to men
- For this the recommendation is the encourage teachers to use a variety of instructional methods and teaching materials that appeal to different learning styles and interests. Promote collaborative learning activities and group projects that foster teamwork and communication skills among students of all genders.

# Technical next steps for Data Science

- As my dataset has 1000 rows and 8 columns in my view this is the less data but apart from it I think that the data need to be collected more in order to solve this project.
- The reason for collecting more data on this project so that we can solve the other business problems based on this dataset in the future
- The model I choose based on the accuracy and the percentage of the correct predictions also I think that there might be a chance to use other models which are capable of showing better results.
- The more advanced model can be used such as XGBoost ,Gradient Boosting

# Appendix

---

# Additional Information:

- <https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/#:~:text=Linear%20regression%20is%20a%20fundamental,make%20predictions%20for%20new%20inputs.>
- Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables ( $X_1, X_2, X_3, \dots, X_n$ ). It assumes that this relationship can be approximated by a linear equation.

## Model Assumptions:

Linear regression relies on several key assumptions:

- **Linearity:** The relationship between the independent variables and the dependent variable is linear. This means that the change in the dependent variable is proportional to changes in the independent variables.
- **Independence:** The observations are independent of each other. The value of one observation does not influence the value of another observation.
- **Homoscedasticity:** The variance of the residuals (the differences between the observed and predicted values) is constant across all levels of the independent variables. In other words, the spread of the residuals should be consistent.
- **Normality:** The residuals follow a normal distribution. This assumption implies that the errors are symmetrically distributed around zero.



