

## Contents

Introduction.....	1
Data Preparation .....	2
Univariate Analysis .....	2
Overall Score Observations .....	3
Difference Between Overall Score and the Calculated Average.....	4
Boxplots - Numerical VS Categorical Variables .....	5
Correlation Matrix.....	6
Relationships between Categorical Variables .....	6
Summary.....	7

# DataScience101 - Term Project

## Student Performance Analysis

## Introduction

The data for this project is downloaded from [www.kaggle.com](http://www.kaggle.com). At the time it was a trending data set on the kaggle site. It has good reviews for being clean and usable. It is described:

"Filename: Student\_Performance.csv Rows: 15,000 Columns: 16

This file contains individually structured student records, where each row represents a single student along with their demographic profile, educational background, learning habits, and academic performance. The dataset combines behavioral, environmental, and academic factors, making it suitable for a wide range of educational and analytical applications.

The file includes information on:

Demographics: age, gender, school type Family background: parent education level Study-related habits: daily study hours, study method, internet access School engagement: attendance percentage, travel time, participation in extra activities Academic records: marks in Math, Science, and English Final outcomes: overall performance score and assigned grade

All values follow consistent formatting, column naming conventions, and realistic ranges to ensure ease of use. The dataset is clean, balanced, and ready for immediate download and analysis."

I looked more closely at the description of the dataset on Kaggle. It is synthetically generated. Ugh.

# Data Preparation

The original dataset contained 25000 Rows

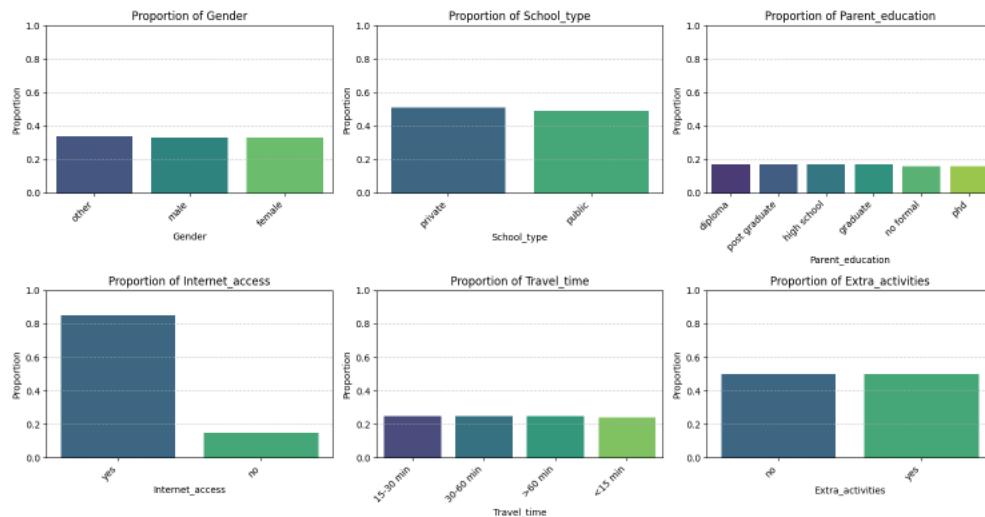
The fields:

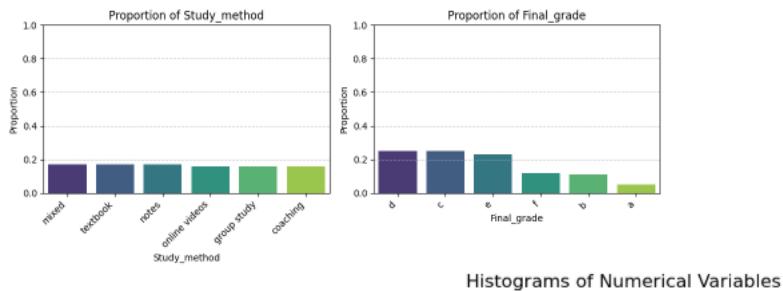
student\_id # of unique student id's is 15000  
age 14 -19  
gender - male, female, other  
school\_type 'public' 'private'  
parent\_education 'post graduate' 'graduate' 'high school' 'no formal' 'diploma' 'phd'  
study\_hours 0.5 – 8.0  
attendance\_percentage 50 – 100 %  
internet\_access 'yes' 'no'  
travel\_time '<15 min' '>60 min' '15-30 min' '30-60 min'  
extra\_activities 'yes' 'no'  
study\_method 'notes' 'textbook' 'group study' 'coaching' 'mixed' 'online videos'  
math\_score high score = 100, low score = 0  
science\_score high score = 100, low score = 0  
english\_score high score = 100, low score = 0  
overall\_score high score = 100, low score = 14.5  
final\_grade ['e' 'd' 'b' 'f' 'c' 'a'] - I am not sure what an “e” grade is

There is no missing data in any column. 10,000 Duplicate Rows – removed

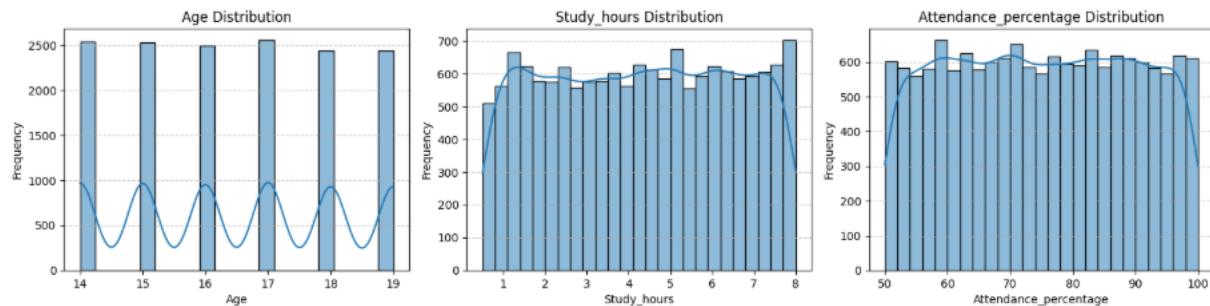
## Univariate Analysis

I did Univariate Analysis for both Categorical and Numerical Variables. It seems that a stratified sampling method was used. Only internet access showed any proportional differences. Only the variables associated with scores showed anything interesting.



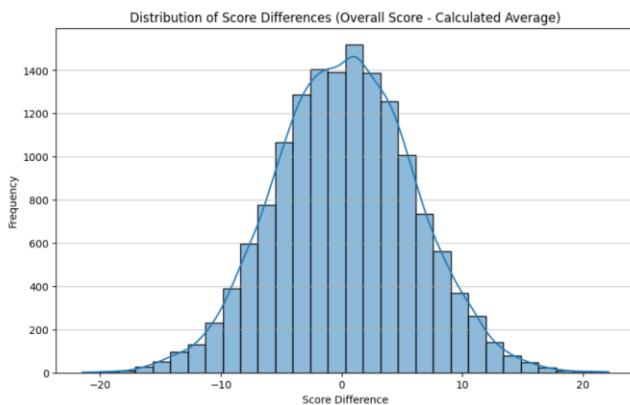


Histograms of Numerical Variables

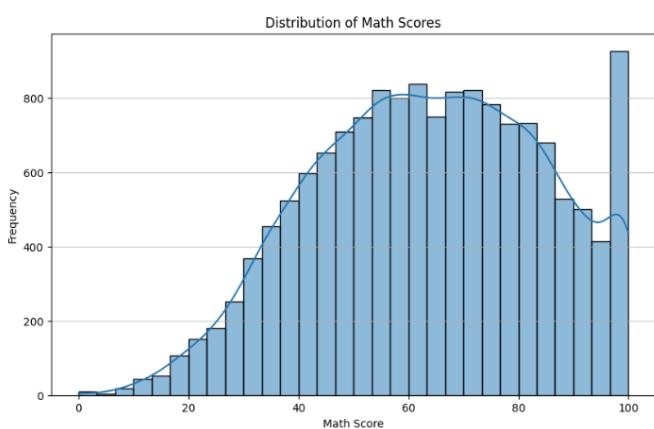


## Overall Score Observations

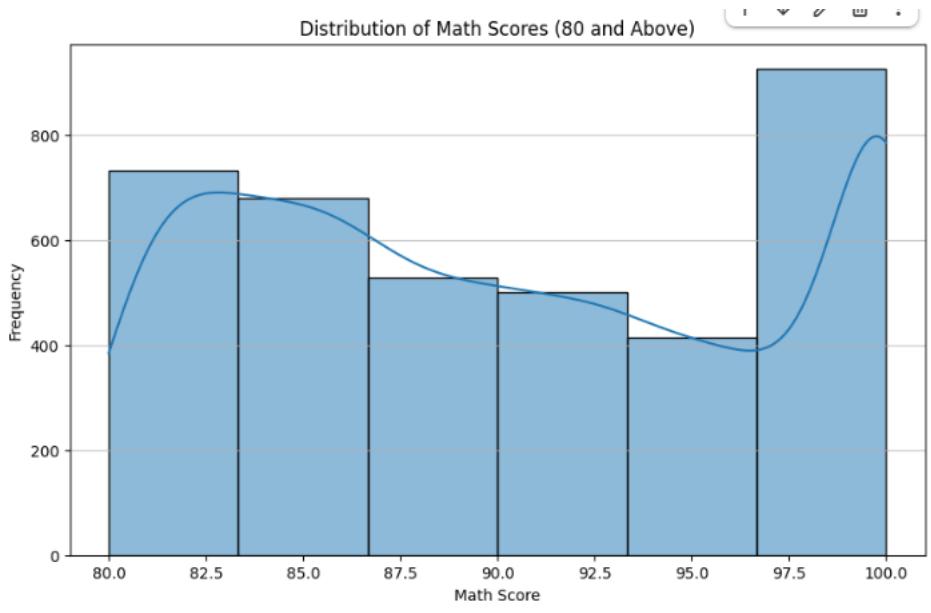
Interestingly, the overall\_score is not simply derived from the average math, English, and science scores.



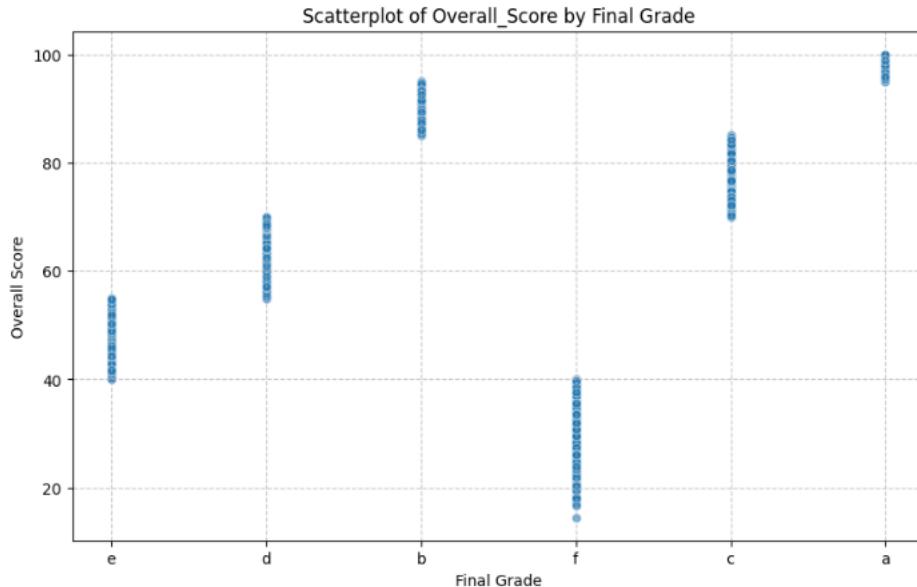
I generated histograms of the overall\_scores, math\_scores, english\_scores, and science\_scores. The distribution of the math, English, and science scores all look similar. Here is the one for math.



This is consistent with the high school model where it is expected that multiple student will get a perfect or near perfect score. Breaking down it can be seen that close to 1000 students are scoring perfect or near perfect scores.

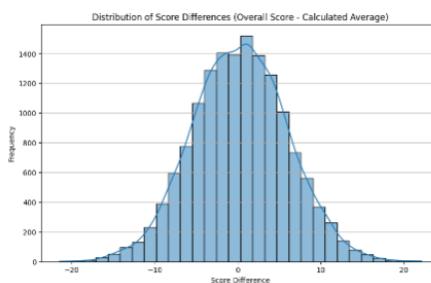


I determined the relationship Between the Overall\_Scores and Grades. The grades are related to the overall\_score. See below.



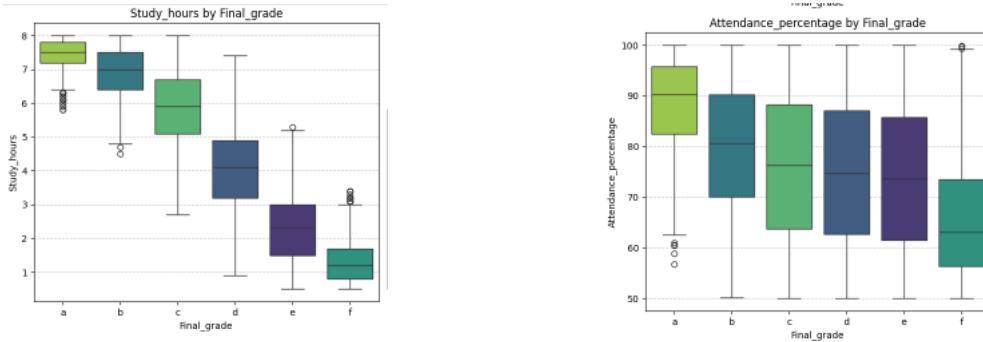
## Difference Between Overall Score and the Calculated Average

The overall\_score does not seem to be the exact calculated average of the math, English, science scores.

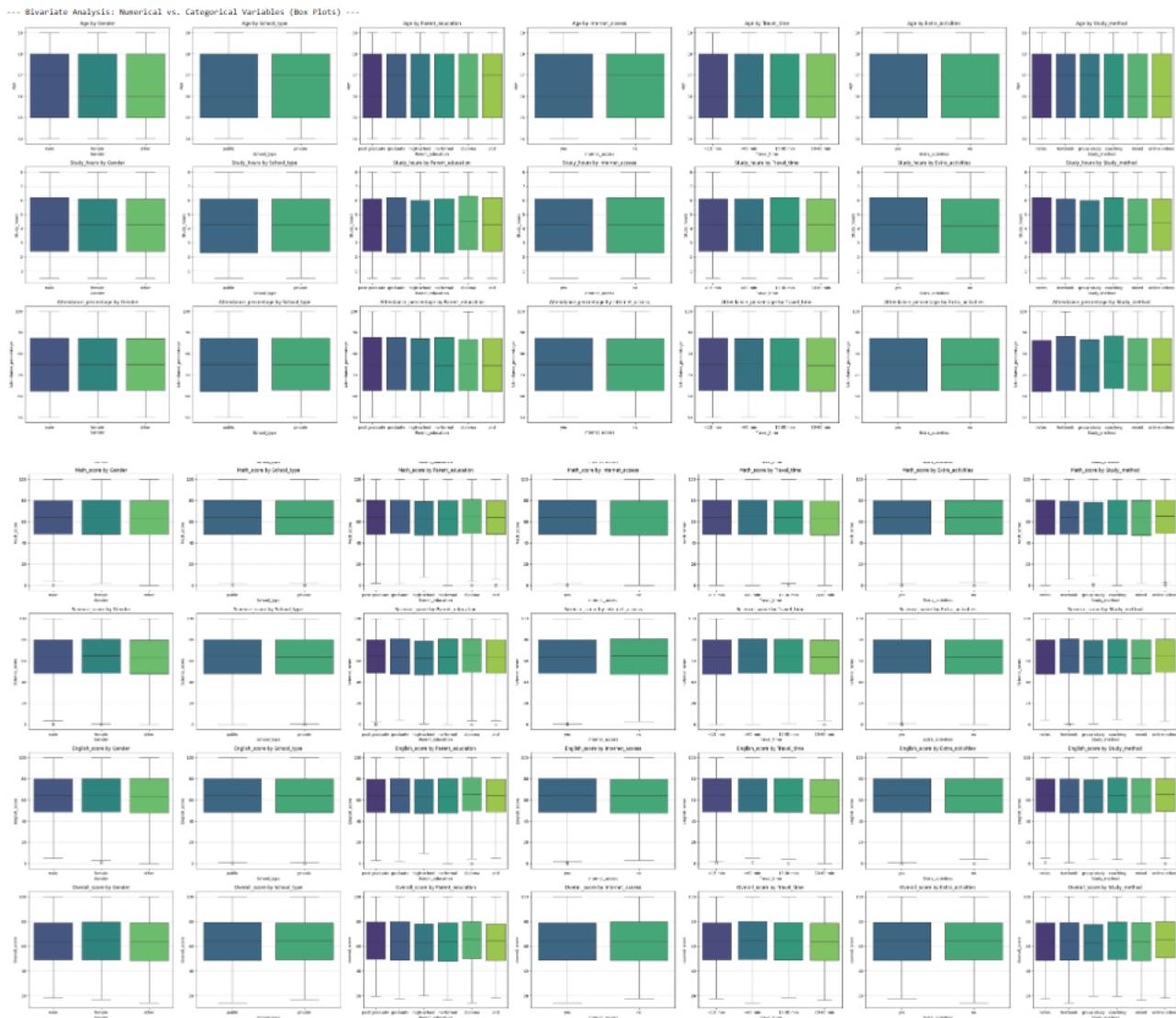


# Boxplots - Numerical VS Categorical Variables

I generated a series of BoxPlots to Illustrate the Relationship between Numerical and Categorical Variables. Both Study Hours and Attendance Percentage affected the final grade.

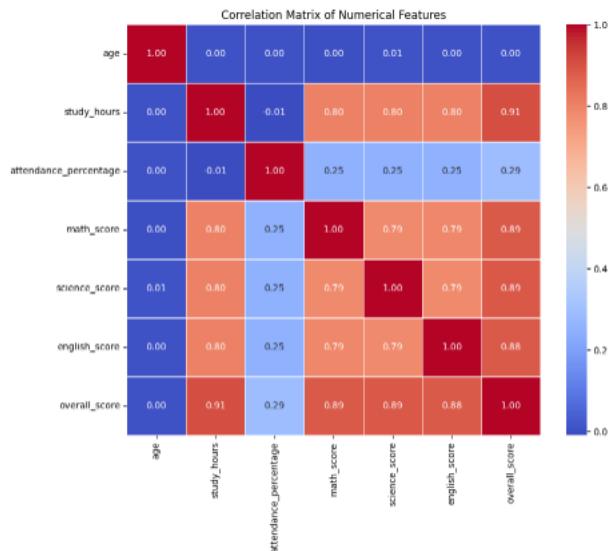


Most BoxPlots showed even distributions. Some examples:

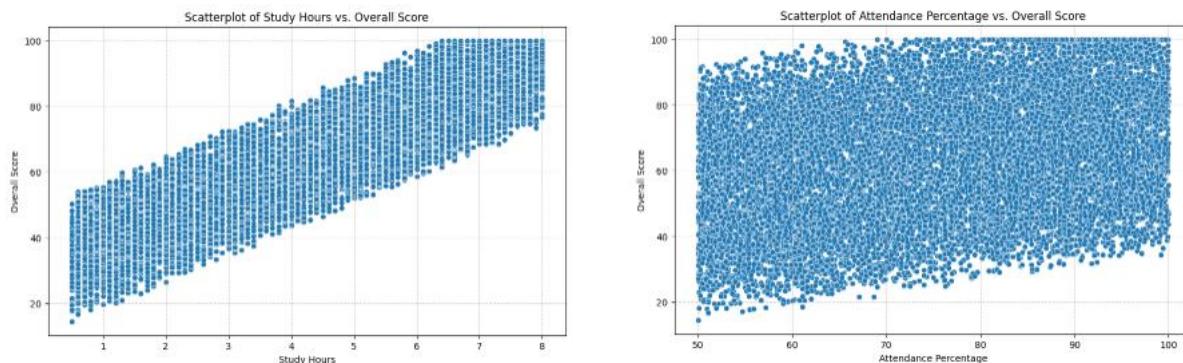


# Correlation Matrix

I generated a Correlation Matrix for the Numerical Variables. The student performance was once again linked to study hours and attendance percentage.

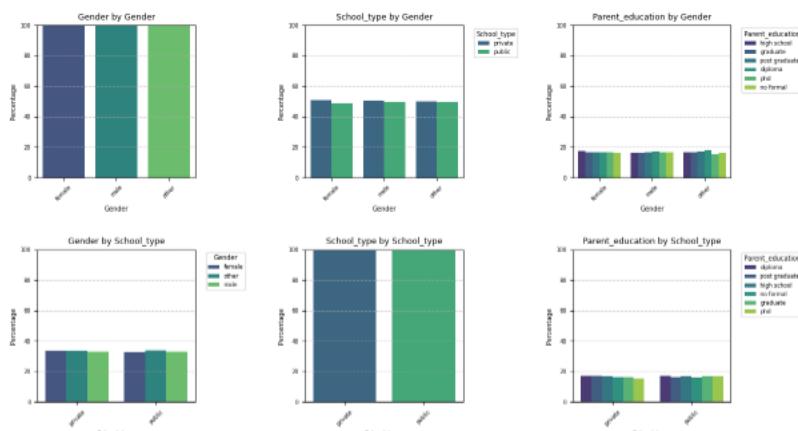


I generated some scatterplots of overall\_score versus study\_hours and attendance.



# Relationships between Categorical Variables

Lastly, I looked for relationship between the Categorical Variables by using grouped bar charts. Nothing seemed particularly interesting. Here are some examples:



## Summary

Going forward, I will be focusing on the predictive value of study\_hours and attendance\_percentage on scores.