DataScience101 - Term Project
Student Performance Analysis 2/28/2026

Contents

Introduction

The data for this project is downloaded from 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."

During the EDA part of the project, 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

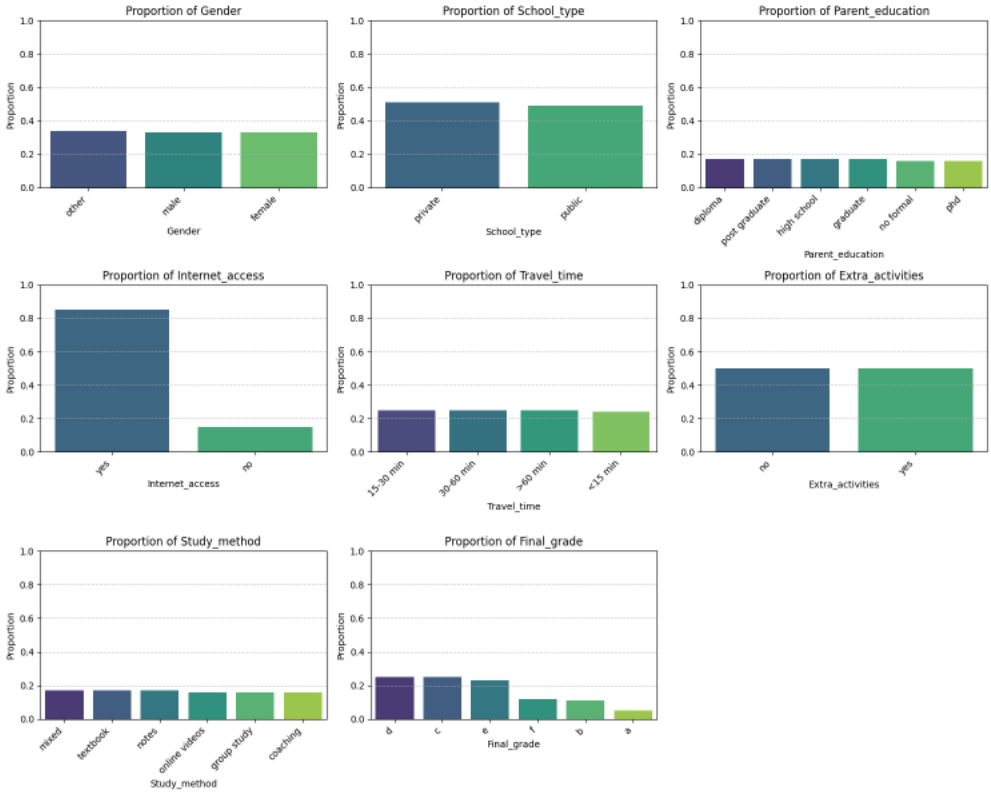
There is no missing data in any column. 10,000 Duplicate Rows – removed

Univariate Analysis

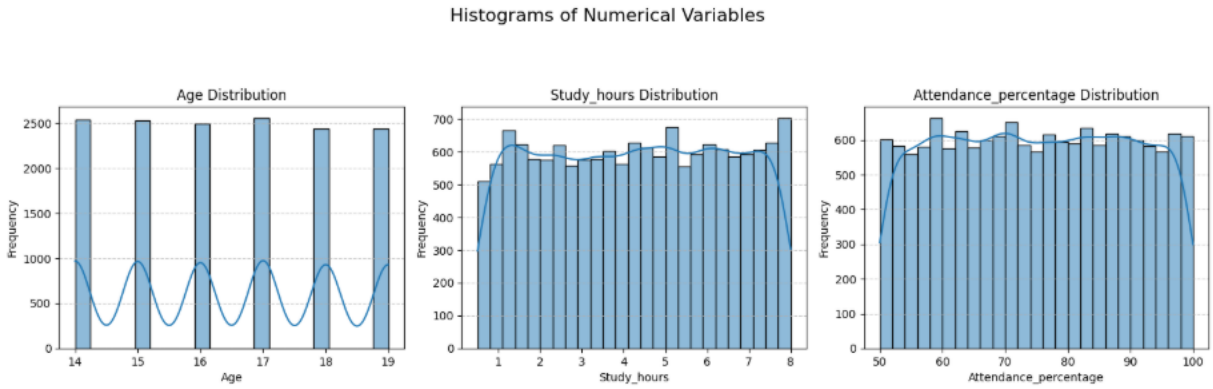
I did Univariate Analysis for both Categorical and Numerical Variables.

Univariate Analysis of Categorical Variables – Bar Charts

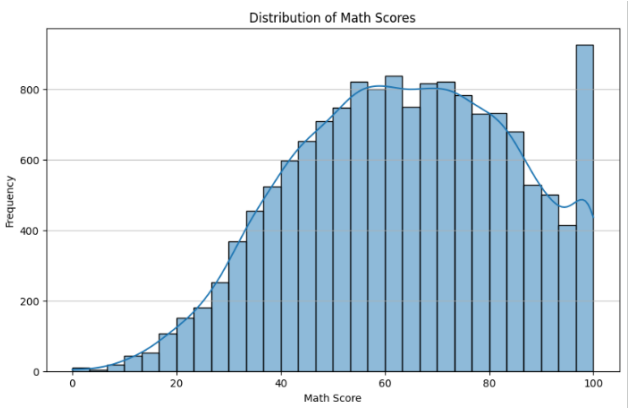
It seems that a stratified sampling method was used. Only internet access and final grades showed any potentially interesting proportional differences.



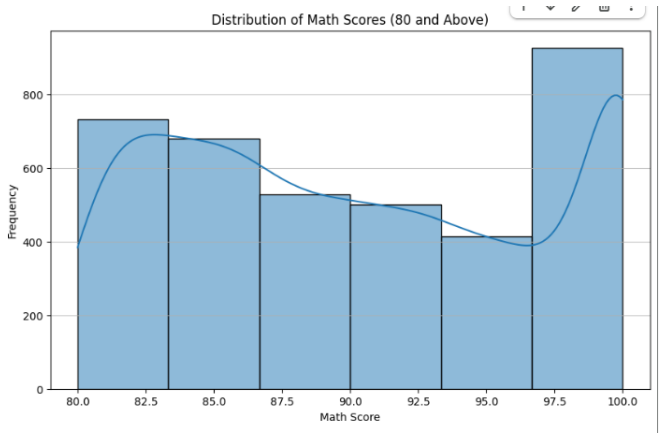
Univariate Analysis of Numerical Variables - Histograms



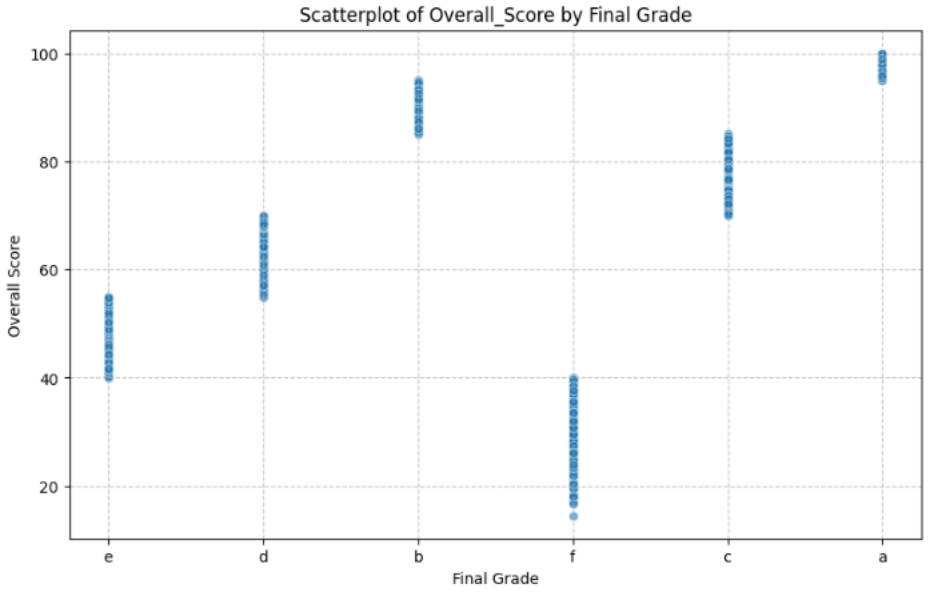
I generated histograms of the overall_scores, math_scores, english_scores, and science_scores. The distribution of the math, English, and science scores all look similar. Here is the one for math.



This is consistent with the high school model where it is expected that multiple students will get a perfect or near perfect score. Close to 1000 students are scoring perfect or near perfect scores.

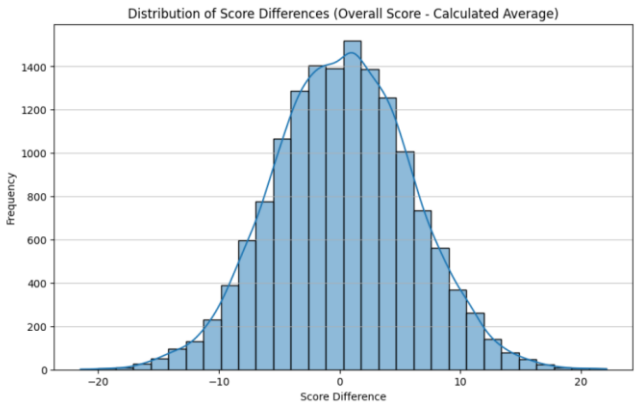


I determined the relationship Between the Overall_Scores and Grades. The grades are related to the overall_score. See below. The grade assignments are consistent with traditional practices, although I am unfamiliar with the "e" grade.



Overall Score Observations

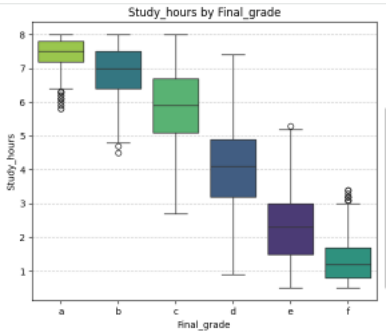
Interestingly, the overall_score is not simply derived from the average math, English, and science scores.

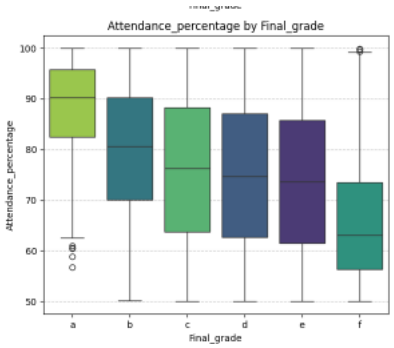


Bivariate Analysis

Box/Violin plots - Numerical VS Categorical Variables

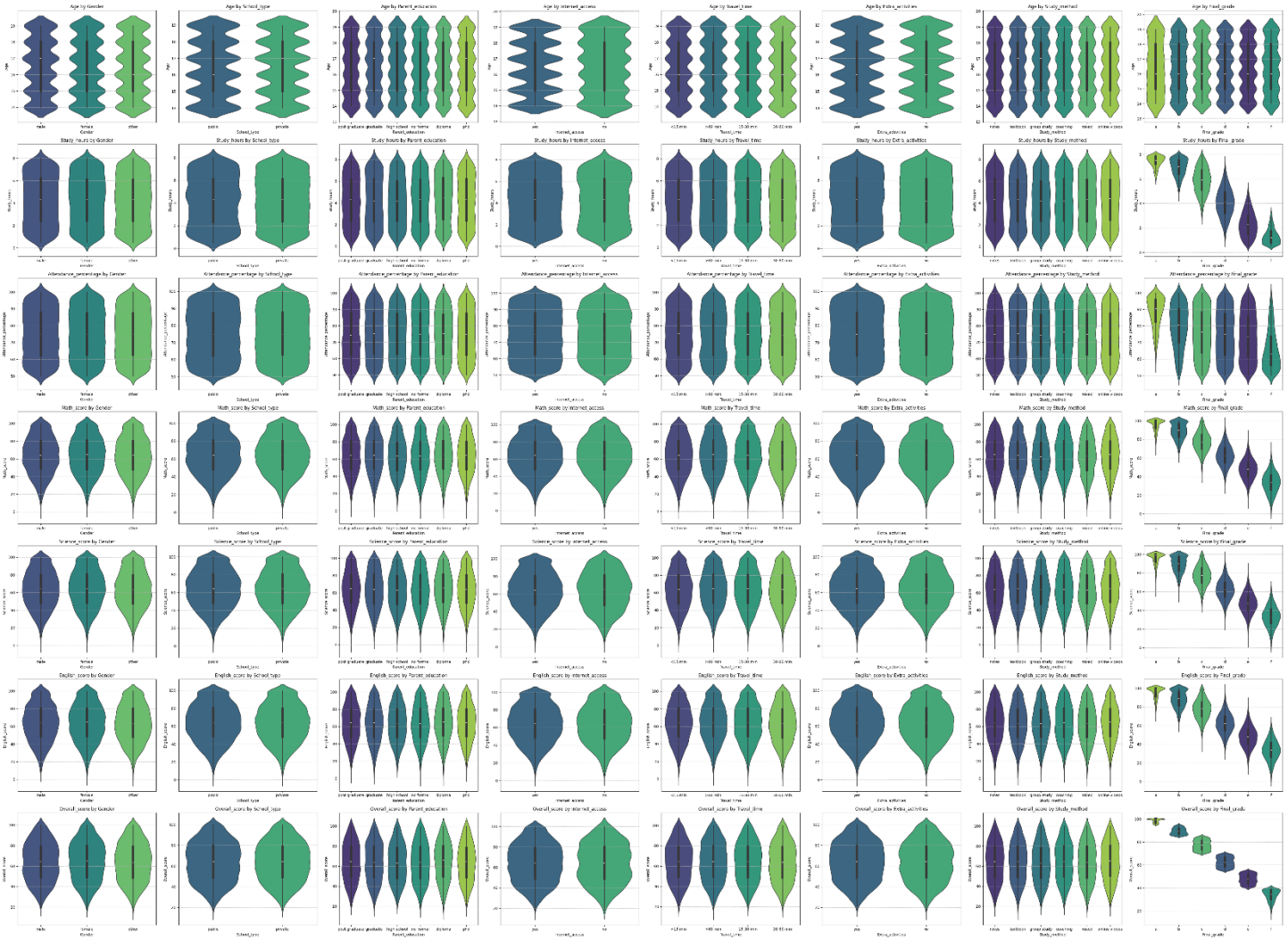
I generated a series of BoxPlots (and then later Violin Plots) to Illustrate the Relationship between Numerical and Categorical Variables. Both Study Hours and Attendance Percentage showed a correlation with the final grade.





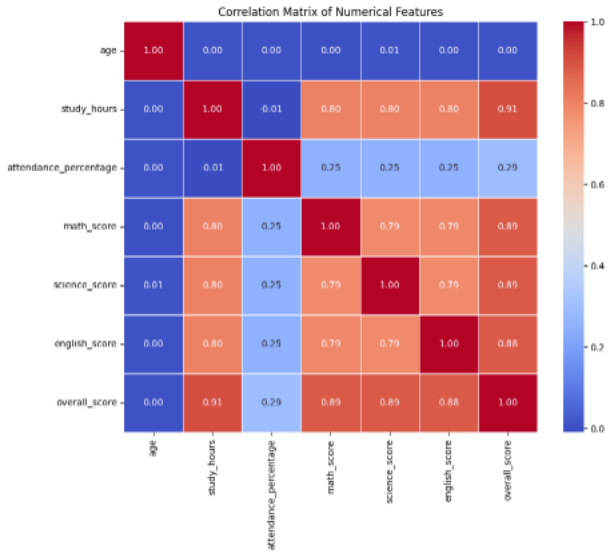
Most BoxPlots (and later ViolinPlots) were uninteresting:

--- Bivariate Analysis: Numerical vs. Categorical Variables (Violin Plots) ---

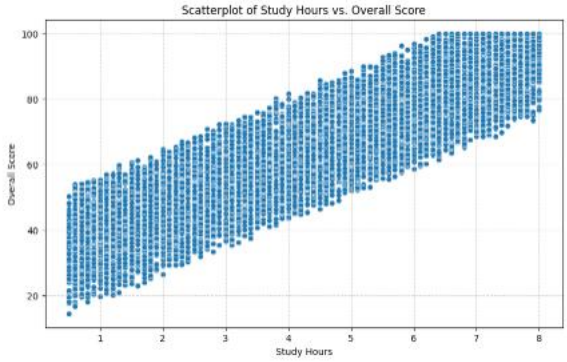


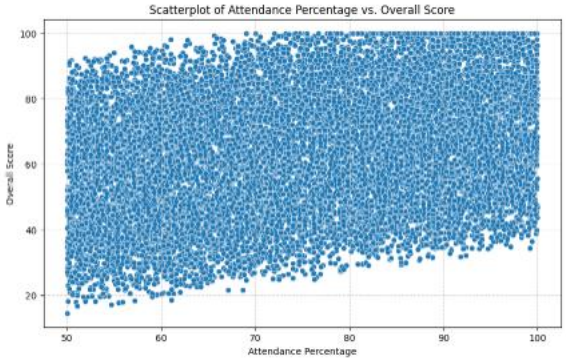
Correlation Matrix

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



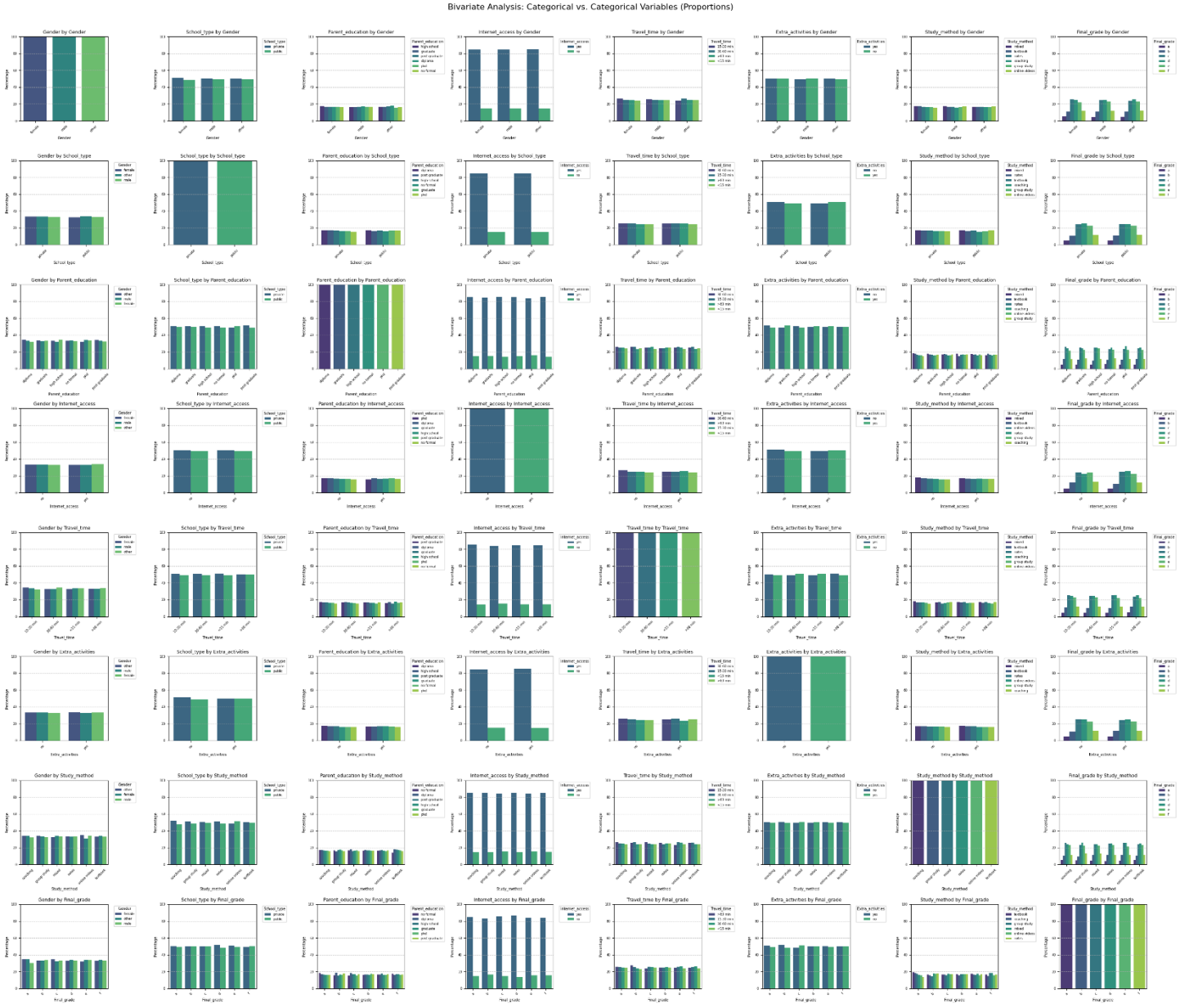
I generated some scatterplots of overall_score versus study_hours and attendance.





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.

Inference for a Correlation

Overall Score versus Attendance Percentage

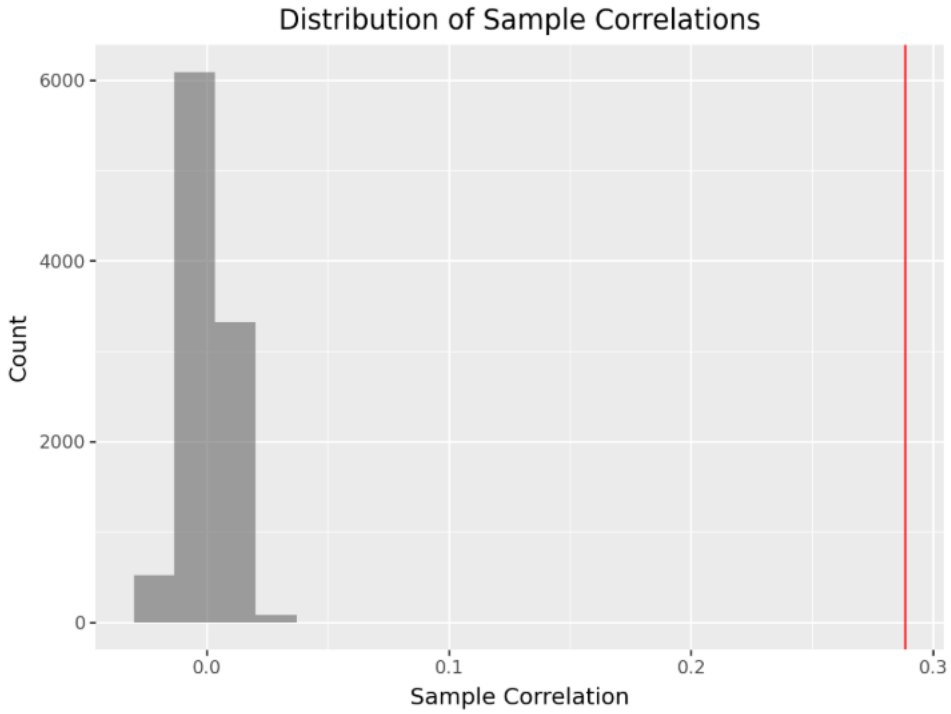
The scatterplot suggested a positive linear relationship. Is there a statistically significant correlation between attendance and score? Is better attendance associated with higher overall scores?

Null Hypothesis: The correlation between overall_score and attendance_percentage is 0.

Alternative Hypothesis: A positive correlation exists between overall_score and attendance_percntage.

The simulated correlation is 0.28847322776538425

Histogram of 10000 simulated sample Correlations



The proportion of results that are at least as extreme as our sample is essentially 0. With a p-value of 0 the is evidence to reject the Null Hypothesis in favor of the Alternative Hypothesis. There is evidence to conclude that the correlation between attendance percentage and overall score is greater than 0.

Overall Score versus Study Hours

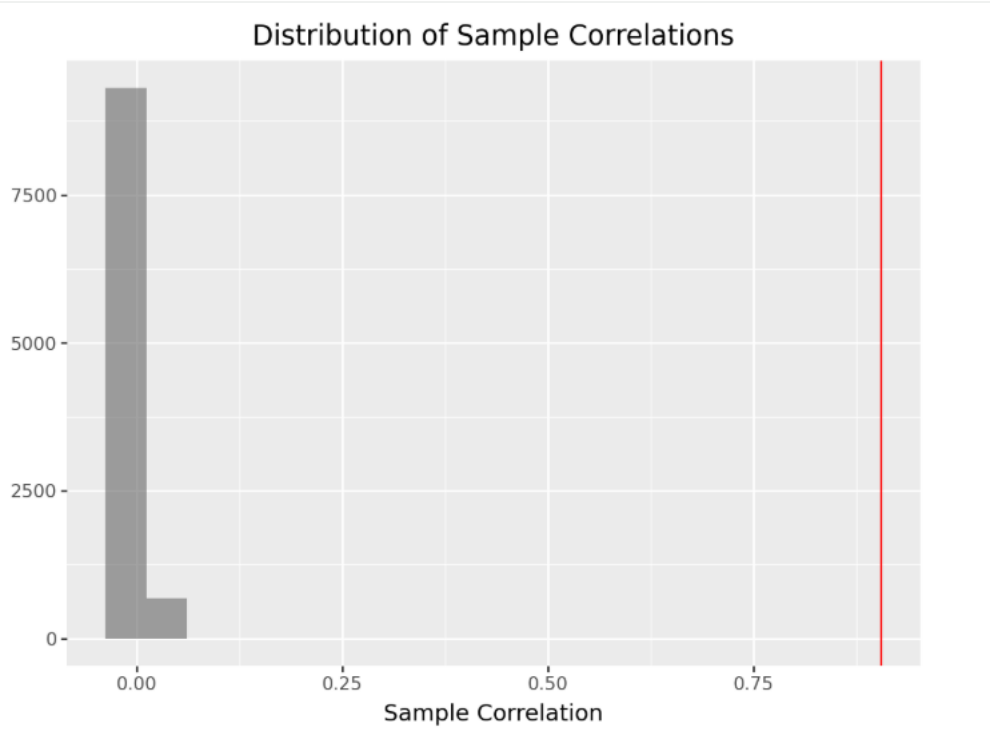
The scatterplot suggested a positive linear relationship. Is there a statistically significant correlation between study_hours and score? Are more study hours associated with higher overall scores?

Null Hypothesis: The correlation between overall_score and study_hours is 0.

Alternative Hypothesis: A positive correlation exists between overall_score and study_hours.

The simulated correlation is 0.9058848411126499

Histogram of 10000 simulated sample Correlations



The proportion of results that are at least as extreme as our sample is essentially 0. With a p-value of 0 the is evidence to reject the Null Hypothesis in favor of the Alternative Hypothesis. There is evidence to conclude that the correlation between attendance study_hours and overall score is greater than 0.