How to boost Your exam score

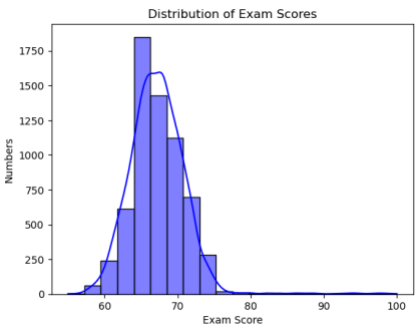


Fig1: Distribution of Exam Scores

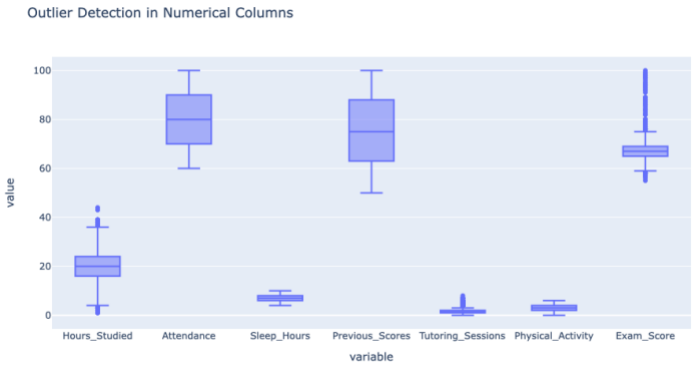


Fig 2: Detection in Numerical attributes

Figure 1 shows the distribution of exam scores, from the picture, we can find out that most students gain exam scores among 68.
Figure 2 is the abstract of outliers and the distribution of values across multiple numerical variables.

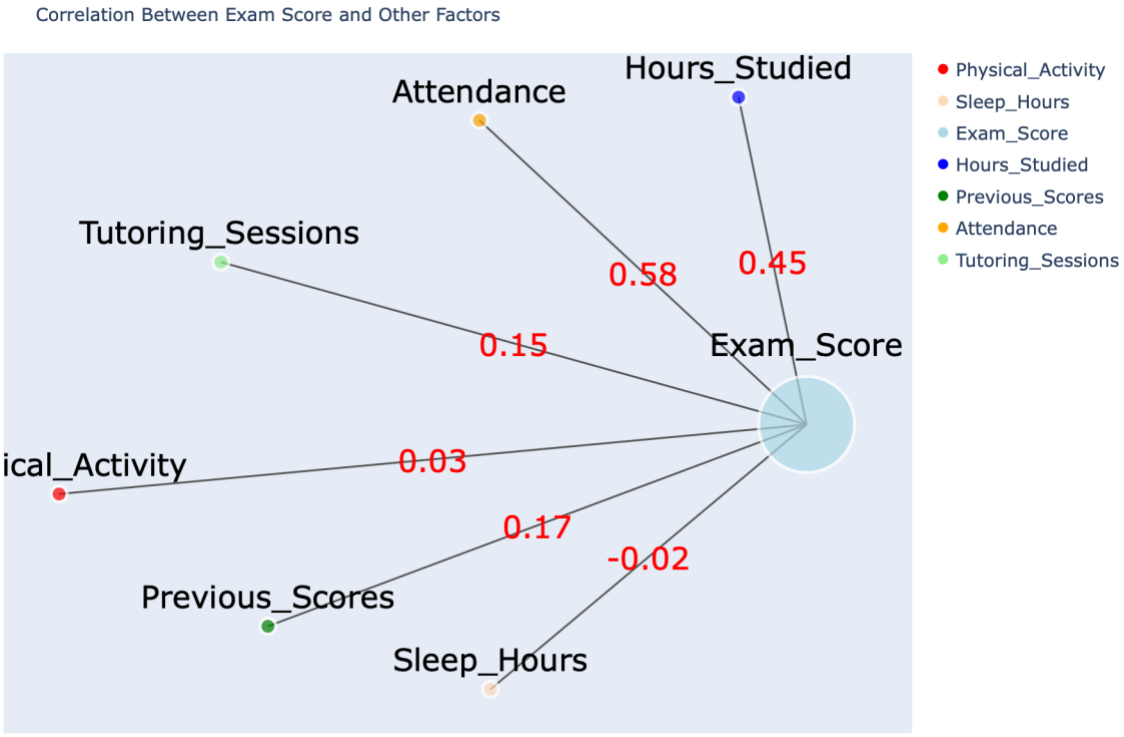
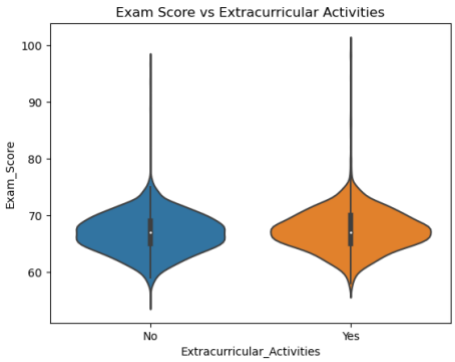


Fig 3: Correlation between exam scores and numerical factors

This network graph illustrates the correlations between Exam Score and various factors. Each node represents a variable, with Exam Score as the central node. The red numbers represent for edge weights, which indicate the strength of the correlation. We can find that study hours and attendance are the most factors.

After exploring the relationship between Exam_Score and numerical factors, we will now investigate its association with categorical factors.



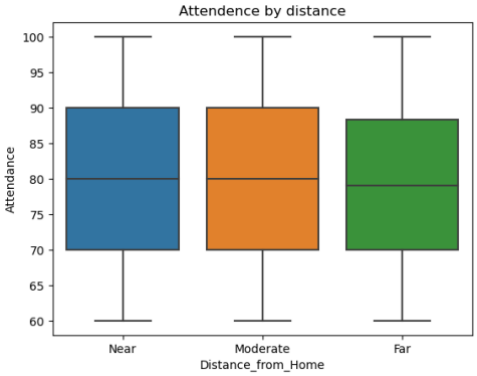
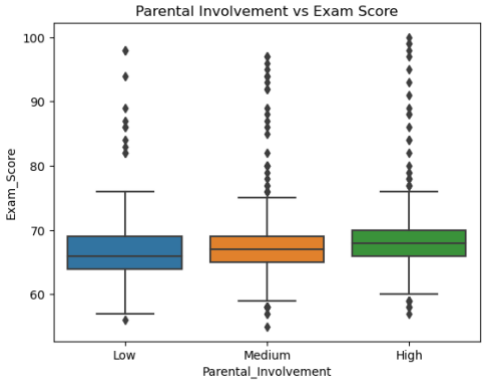


Fig 4: Correlation between exam score and activities Fig 5: Correlation between attendance and distance

Since study hours and attendance both contribute a lot to exam scores, we go further on these factors. Figure 4 focuses on whether students have more time to study after class, however, those students who have extracurricular activities perform even a little better than those who don't.
In the Figure 5, three boxplots stand for different distance from home, blue for near distance, orange for moderate distance and green for the far distance.



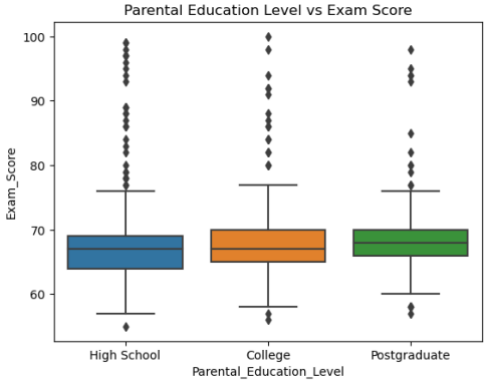
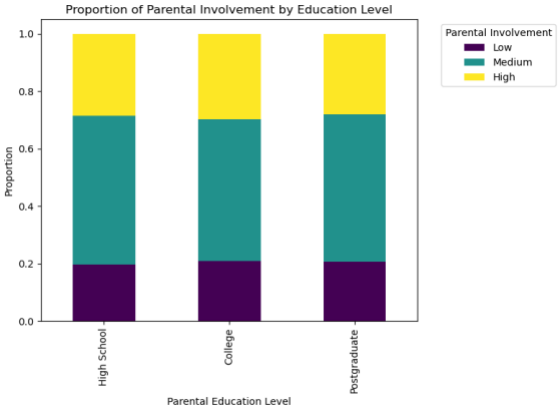


Fig 6: Correlation between exam score and parents Fig 7: Correlation between parental education level and scores



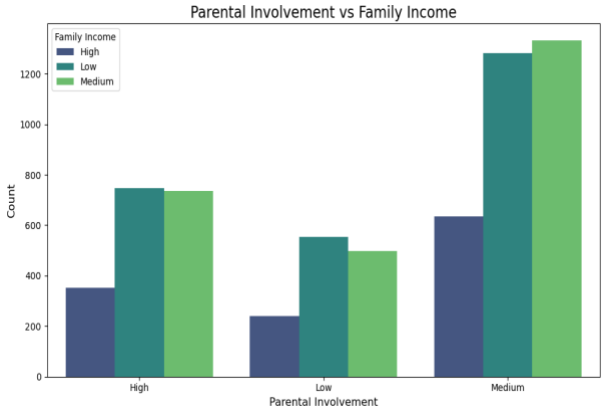
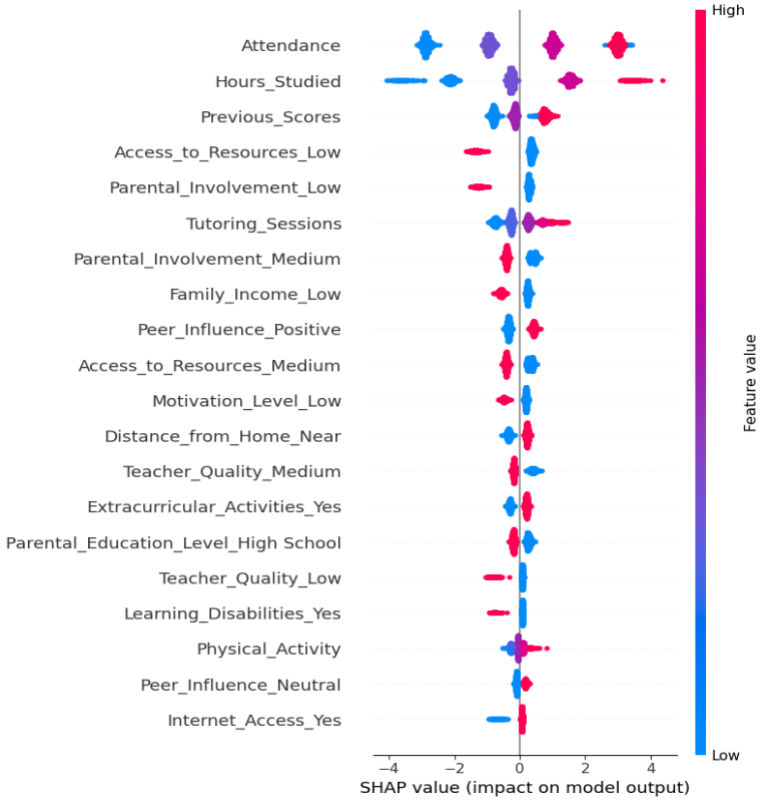


Fig 8: Parental involvement vs their education Fig 9: Correlation between parents and family income

These four pictures show factors correlated with parents. In figure 6, the box plot shows the distribution of Exam_Score by Parental Involvement levels, blue stands for low involvement, while orange means medium and green for high. In figure 7, different parents education level uses

different colors in the boxplot. Figure 8 shows whether parental involvement is affected by their own education level. In figure 9, it focuses on whether the family income will influence parental involvement, purple stands for high family income, dark-green for low while light-green for medium income.

Importance: In this survey, my goal is to explore how to achieve higher exam scores through the analysis of both numerical and classified factors. First, I found that study time and attendance have a significant impact on achieving higher scores. From the network graph, the specific correlation coefficients between numerical factors and exam scores can be observed. Next, I further explored the influence of categorical factors and found that extracurricular activities do not affect students' performance, while longer commuting distances negatively impact attendance, which in turn affects their scores. Parental involvement is an essential factor, as shown in the box plot, where higher parental involvement contributes to better student performance. Additionally, parents' education levels also have a positive impact on their children's academic achievements. However, parental education levels are generally unrelated to their involvement, with most showing a medium level of involvement. Family income is also largely unrelated to parental involvement, but families with higher incomes tend to have less involvement, likely due to parents spending more time at work.



I used XGBoost for model training and visualized the results with a SHAP plot. The y-axis shows the contribution of each feature to the model's predictions, demonstrating that my previous survey was effective, as the factors I explored significantly influence the outcome. The x-axis represents the magnitude of the feature's impact, while the color transition from blue to red indicates increasing feature values.

Data and method: The data I used for this assignment came from the open-source Airbnb Hawaii dataset. You may also download it directly from my GitHub repository. All the graphs were generated using python. Depending on the type of the graph, different python packages were installed. I also use XGBoost to modify my train data set and use SHAP summary plot.

GitHub link:https://github.com/Yif50/2410-Final.git