AIDM7400 Data Analysis and Visualization Studio

Group Project Report

Date: May 02, 2025

Group Member	Student ID
HUANG Xiaoqi	24483494
JIANG Hanbing	24456276
SHEN Lan	24448850

Topic: Data Analytics on Students' Adaptive Learning

Background

Nowadays, with the deepening of digital transformation of education, adaptive learning systems and personalized education models are reshaping the modern education ecology. On the online platform, people of different ages can freely choose the subjects to study, the length of study time, what exercises to do, etc. But does this situation really achieve the initial goal of personalized learning and help everyone gain learning results? This study is based on 10,000 student interaction data collected from an online learning platform, covering dynamic behavioural indicators such as time investment, resource access frequency, forum participation, as well as demographic characteristics such as gender, age, region and final learning outcome data, to explore the correlation patterns between indicators of various dimensions. The study aims to reveal the key factors that affect the effectiveness of personalized education, provide data support for optimizing the algorithm design of intelligent education systems, and formulate differentiated teaching intervention strategies, which has practical significance for promoting educational equity and quality improvement.

Data Cleaning and Preprocessing (Exploratory Data Analysis, EDA)

Data Description

Before conducting data analysis, the quality of the data and the understanding of the data content are effective steps to raise questions and hypotheses.

```
[ ] # Get Data Frame information (number of non-null values, data type etc)
    df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 15 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Student_ID                  10000 non-null  object
 1   Age                         10000 non-null  int64
 2   Gender                      10000 non-null  object
 3   Education_Level             10000 non-null  object
 4   Course_Name                 10000 non-null  object
 5   Time_Spent_on_Videos        10000 non-null  int64
 6   Quiz_Attempts               10000 non-null  int64
 7   Quiz_Scores                 10000 non-null  int64
 8   Forum_Participation         10000 non-null  int64
 9   Assignment_Completion_Rate  10000 non-null  int64
 10  Engagement_Level            10000 non-null  object
 11  Final_Exam_Score            10000 non-null  int64
 12  Learning_Style              10000 non-null  object
 13  Feedback_Score              10000 non-null  int64
 14  Dropout_Likelihood          10000 non-null  object
dtypes: int64(8), object(7)
memory usage: 1.1+ MB
```

Figure 1: the descriptive information of the data set

According to the content of EDA, we first observed the overall data characteristics and determined the data type of each variable feature, including integer class and object class, to avoid errors caused by misusing data in subsequent analysis. This data set has no null values or duplicate values, and there are no obvious outliers in the box plots of all numerical features, and the distribution is relatively uniform overall.

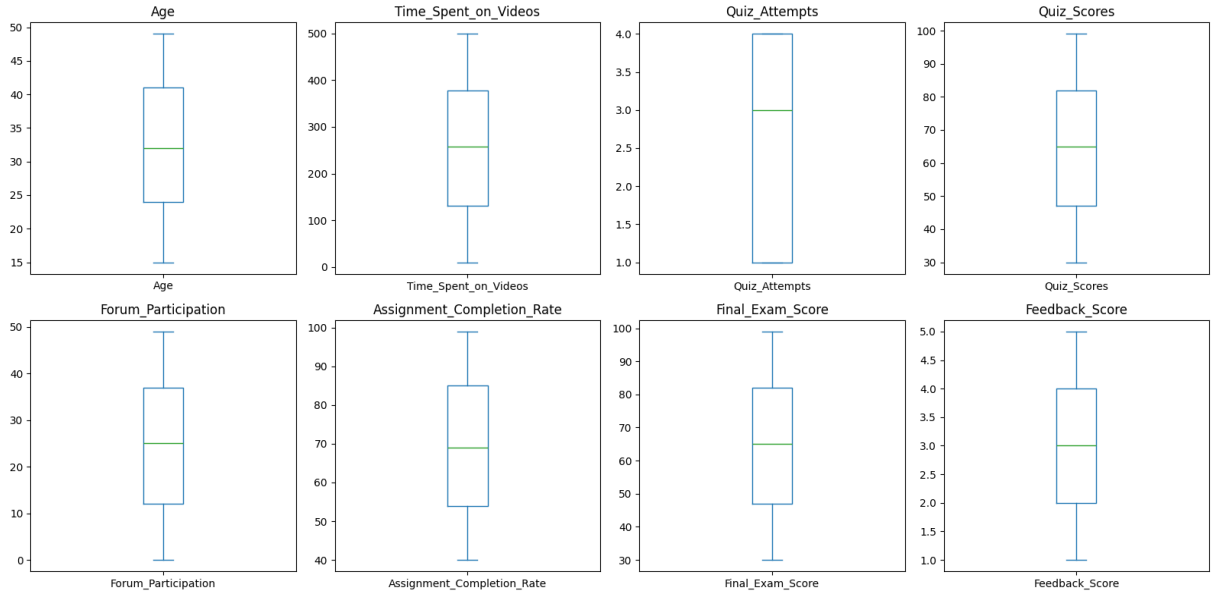


Figure 2: the box figures of each digital features

Then, the distribution characteristics of each variable are refined through density maps and histograms, which can also identify outliers and understand the data structure. For example, the image of *Time_Spent_on_Videos* explains that these data are concentrated in certain intervals, and

the several obvious peaks shown in the density map indicate that the specific length of time is the length of time that users are accustomed to watching videos. The *Forum_Participation* table shows that participation may be higher in certain intervals, showing a multi-peak distribution, and the density map shows the concentration trend of participation.

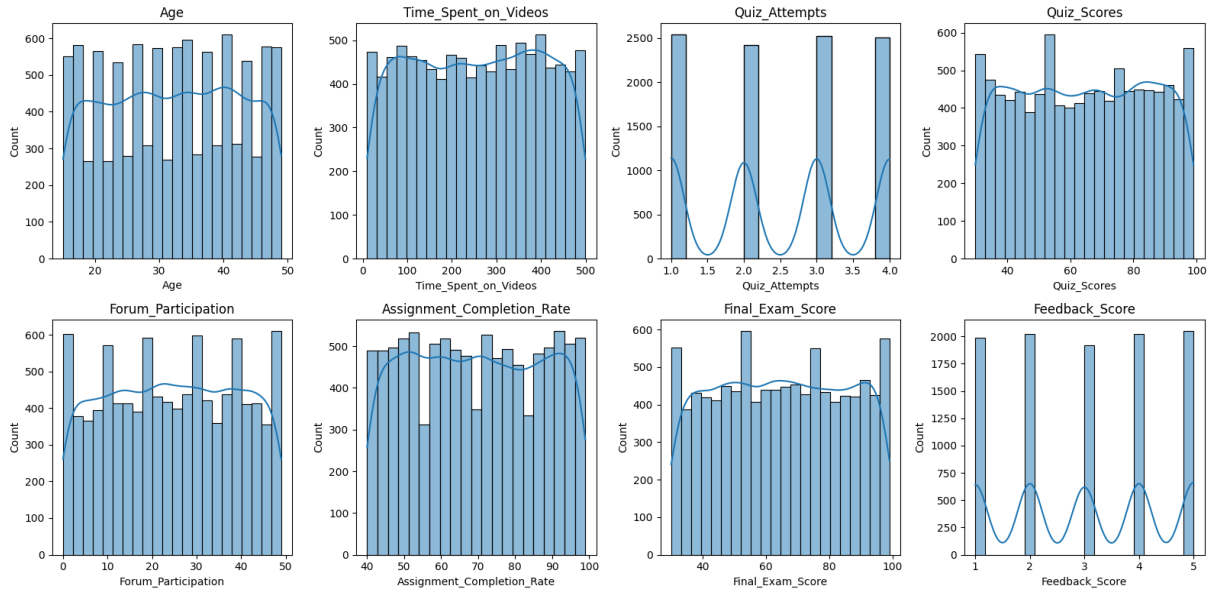


Figure 3: the distribution characteristics of variables

In addition to numerical variables, categorical variables also need to further study the data model between them, so we use frequency calculation to summarize the category proportion of each feature, including average ones, such as *Gender*, *Course_Name*, anId *Learning_Style*, and those with large skewness, such as *Engagement_Level*, and the final result *Dropout_Likelihood*.

```
Gender
Female    4886
Male      4699
Other      415
Name: count, dtype: int64
========================================
Education_Level
Undergraduate    5070
High School      2923
Postgraduate     2007
Name: count, dtype: int64
========================================
Course_Name
Machine Learning    2043
Cybersecurity       2026
Python Basics       1994
Data Science        1984
Web Development     1953
Name: count, dtype: int64
========================================
```

Figure 4: Examples for categories' frequency outcome

Based on the previous steps, we visualized the correlation between digital feature variables using a heat map. From the results that both positive and negative values are close to 0, we can conclude that the relationship between several features is not as close as we expected. The relatively obvious ones are *Age* and *Quiz_Scores*, but they are negative numbers, which means that in real life, the older you are, the lower your test scores may be due to decreased attention and insufficient brain power.

Figure 5: correlation between digital feature variables

At the end of EDA, we first try to use some data visualization to discover the potential relationship between data. For example, the average score of students in different online courses is close to 65 points, but the difference is not very large. Web Development has the highest average score by a slight advantage, indicating that students who choose this online course can master the subject content better. Another example is to observe the changes in the time students are accustomed to or can spend on watching video online courses under age trends by finding the average. The conclusion is that 27 years old is the best age for students to self-study through videos, and they

can persist for a long time. On the contrary, 33 years old reaches the lowest point, which may be a stage where they are busy with their careers and families, and do not pay attention to self-improvement and self-discipline. After a certain age, such as 40 years old, they gradually start to have free time to watch teaching videos.

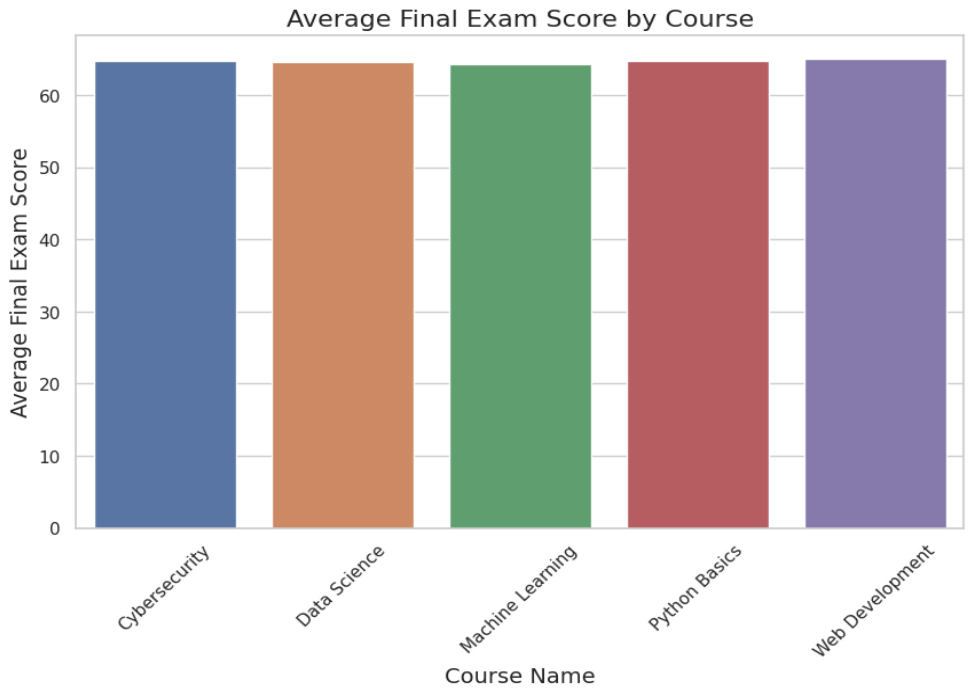


Figure 6: Average Final Exam Score by Course

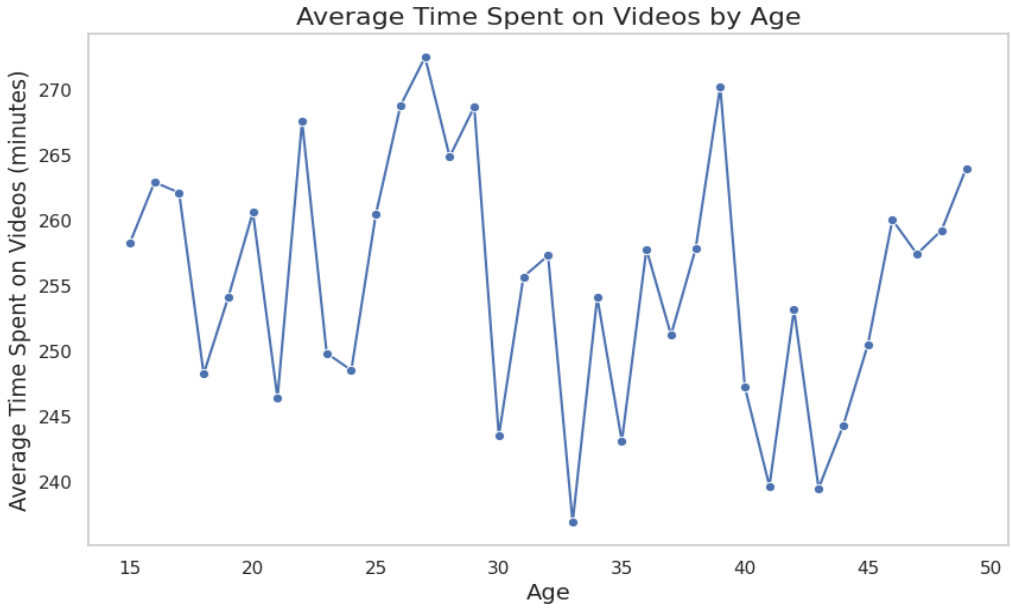


Figure 7: Average Time Spent on Videos by Age

Research Questions (Progressive generation in the analysis process**)**

- *Is there any significant relationship between the time spent watching videos and the final score?*
- *Is there any significant relationship between the engagement level and the final score?*
- *Are there any differences in the final scores among students with different learning styles?*
- *Is there any significant relationship between the time students spend watching videos and their final exam scores? if this relationship differs across various learning styles?*
- *Is there a difference in dropout rates among different types of students based on combinations of engagement and assignment completion levels?*

Data Analysis and Visualization

We primarily use *Final_Exam* (continuous variable) and *Dropout_Likelihood* (binary variable) as dependent variables. Through histogram visualization, we observed that the distribution of *Final_Exam* did not exhibit distinct clustering characteristics and did not fully conform to the assumption of normality. Nevertheless, given the large sample size, the Central Limit Theorem (CLT) justifies the reasonable application of parametric tests such as *t-tests* and *one-way ANOVA* in practical analyses to assess the presence of statistically significant differences between these variables.

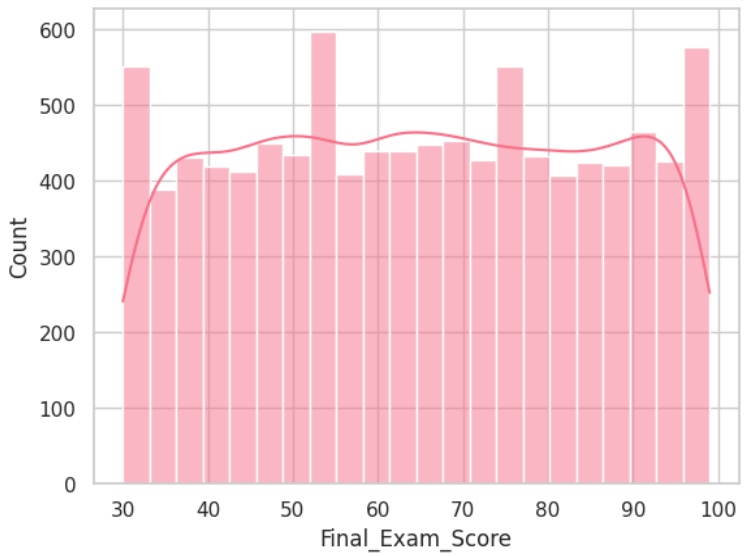


Figure 8: Distribution Analysis of Final Exam Scores

T-test: Examining the Relationship Between Dropout Likelihood and Other Variables

Dropout_Likelihood was used as a binary independent variable (No vs Yes), and Independent Samples *t*-test was used with a significance level of $\alpha = 0.05$ to assess whether there was a significant difference in the means of the two groups for each continuous variable. The test results showed that there was no statistically significant difference between the *Dropout_Likelihood* groups ($p > 0.05$).

	Group Var	Group 1	Group 2	Numeric Var	n1	n2	t-stat	p-value	Significant (p<0.05)
0	Dropout_Likelihood	No	Yes	Age	8043	1957	1.240	0.2149	
1	Dropout_Likelihood	No	Yes	Time_Spent_on_Videos	8043	1957	-0.651	0.5151	
2	Dropout_Likelihood	No	Yes	Quiz_Attempts	8043	1957	-0.282	0.7782	
3	Dropout_Likelihood	No	Yes	Quiz_Scores	8043	1957	-1.447	0.1481	
4	Dropout_Likelihood	No	Yes	Forum_Participation	8043	1957	1.370	0.1709	
5	Dropout_Likelihood	No	Yes	Assignment_Completion_Rate	8043	1957	1.015	0.3103	
6	Dropout_Likelihood	No	Yes	Final_Exam_Score	8043	1957	-0.436	0.6629	
7	Dropout_Likelihood	No	Yes	Feedback_Score	8043	1957	-0.396	0.6919	

Table 1: The relationship between Dropout Likelihood and Other Variables

One-Way ANOVA: Examining the Relationship Between Non-Binary Categorical Variables and Continuous Variables

With the significance level set at $\alpha = 0.05$, it was observed that out of a total of 40 variable combinations, only 4 showed statistically significant differences.

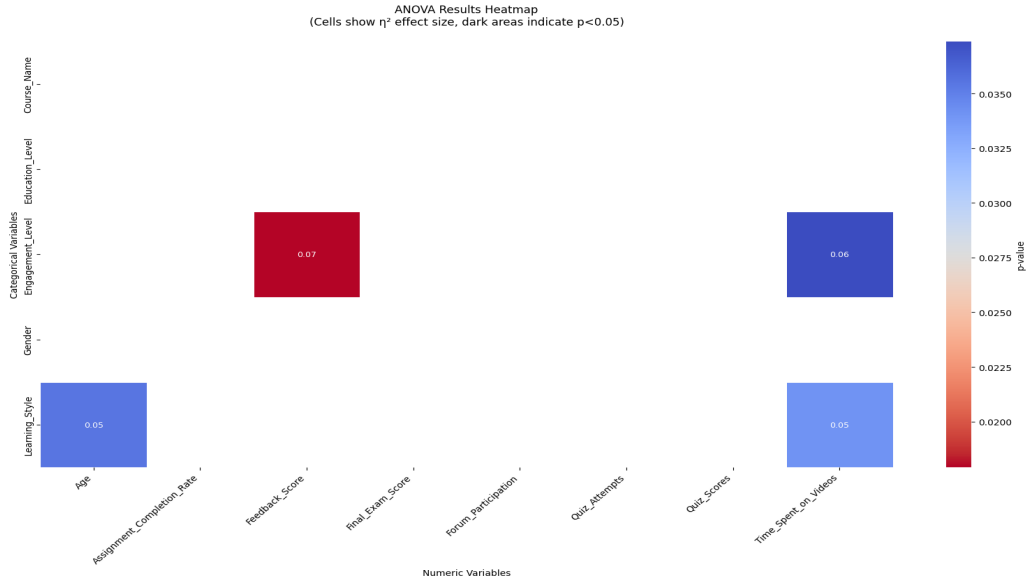


Figure 9: One-Way ANOVA Significance Heatmap Across Categorical and Numeric Variables

The results show that different Engagement Levels have significant differences in *Time_Spent_on_Videos* and *Feedback_Score* ($p < 0.05$). Different Learning Styles are also significantly associated with *Age* and *Time_Spent_on_Videos.* Therefore, we further used Tukey's honestly significant difference post hoc test to explore which groups have significant differences. The test outputs mean difference, adjusted *p*-value, confidence interval bounds for each group comparison, and after visualization, the following results were obtained:

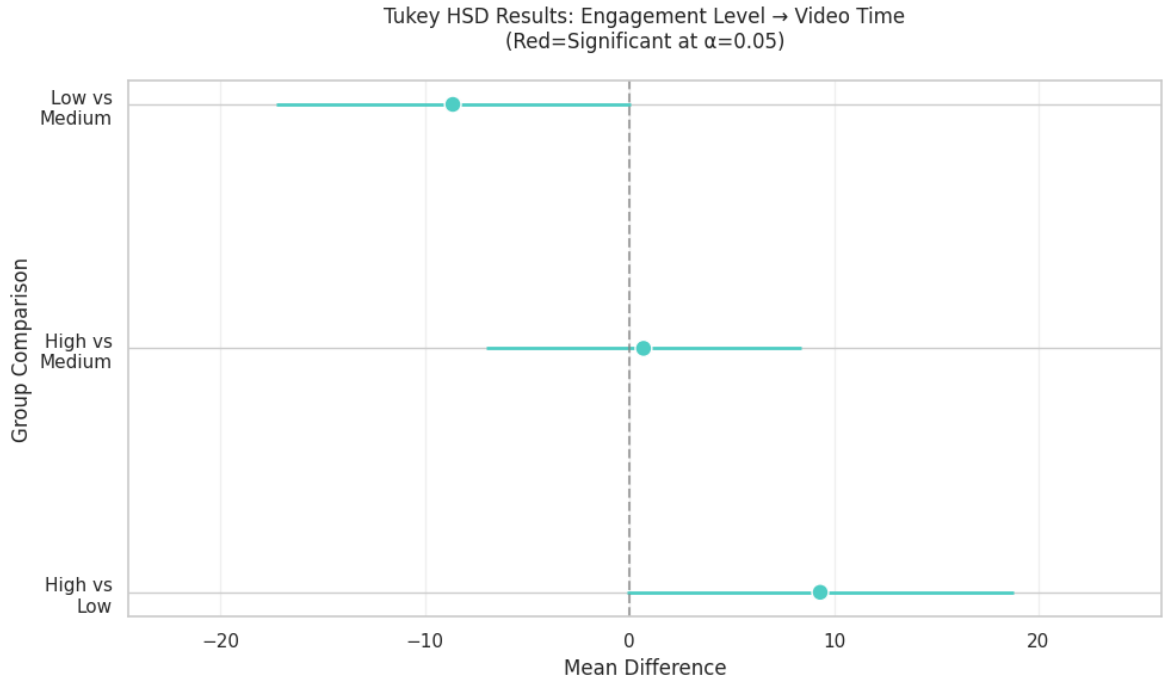


Figure 10: Tukey HSD Pairwise Comparisons (Engagement Level vs. Time Spent on Videos)

In the comparison between Engagement Level and Time Spent on Videos, no significant differences were found between the two groups. However, the *p*-values for the comparison between Low and Medium were 0.0508, and the *p*-values for the comparison between High and Low were 0.0547, both close to the significance threshold of 0.05, showing a marginally significant trend. The results show that the low engagement group may spend more time watching videos than the medium engagement group, but the difference has not yet reached a statistically significant level.

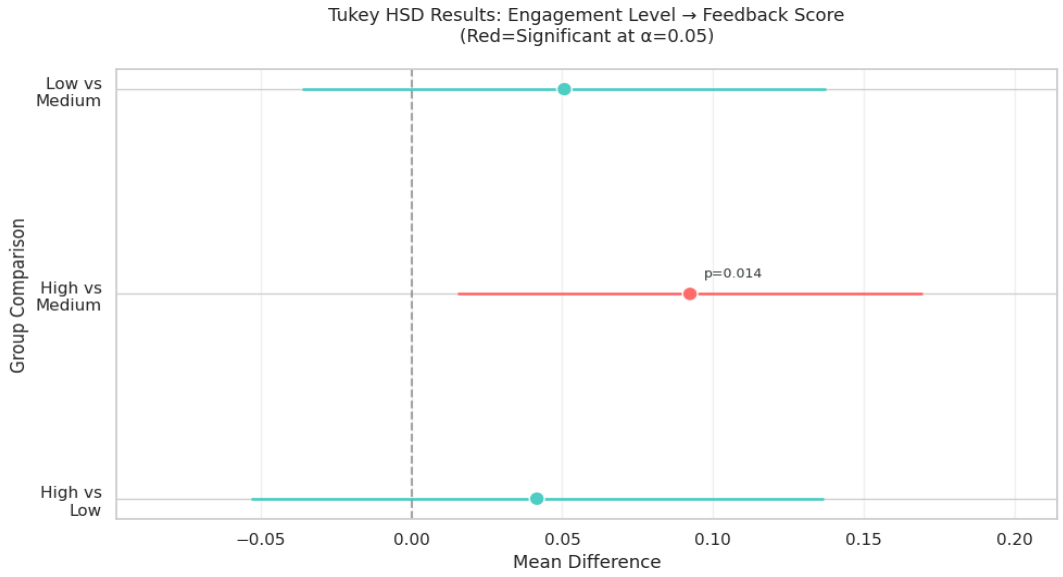


Figure 11: Tukey HSD Pairwise Comparisons (Engagement Level vs Feedback Score)

In the comparison between Engagement Level and Feedback_Score, only one comparison showed a significant difference: High vs Medium p (0.014) < 0.05. This indicates that students with high engagement gave significantly higher course feedback scores than students with medium engagement . There was no significant difference between the Medium and Low groups, indicating that the improvement in feedback scores was mainly concentrated in the high engagement group, which may reflect their higher recognition or satisfaction with the course content.

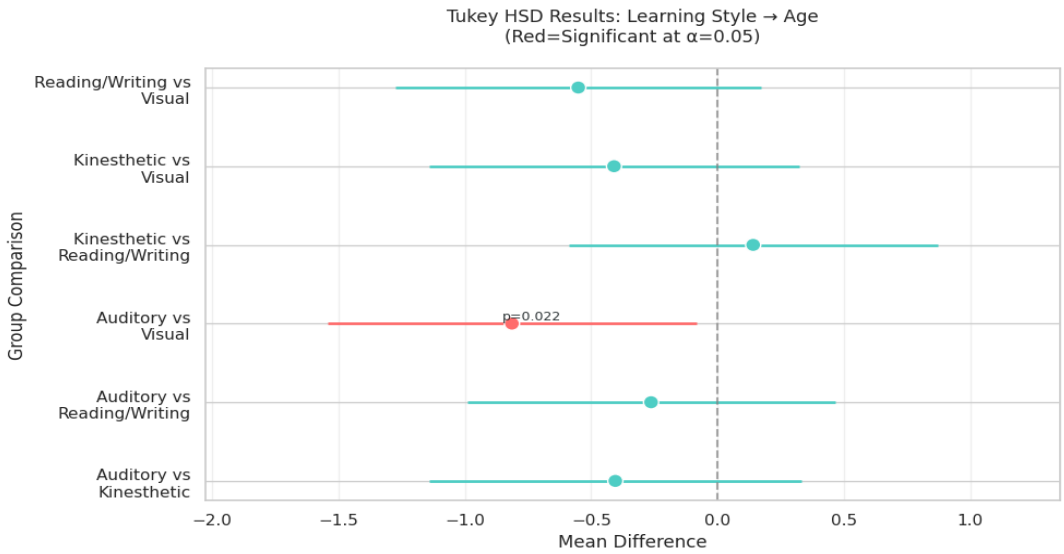


Figure 12: Tukey HSD Pairwise Comparisons: Learning Style vs Age

In the comparison between Learning Style and Age, only one set of comparison results reached statistical significance: the comparison between Auditory and Visual with a p value of 0.022 (< 0.05). The results showed that the age of students who preferred auditory learning (Auditory) was significantly younger than that of students who preferred visual learning (Visual). This finding may reveal that younger students are more inclined to learn through auditory methods, while older students prefer to use visual materials.

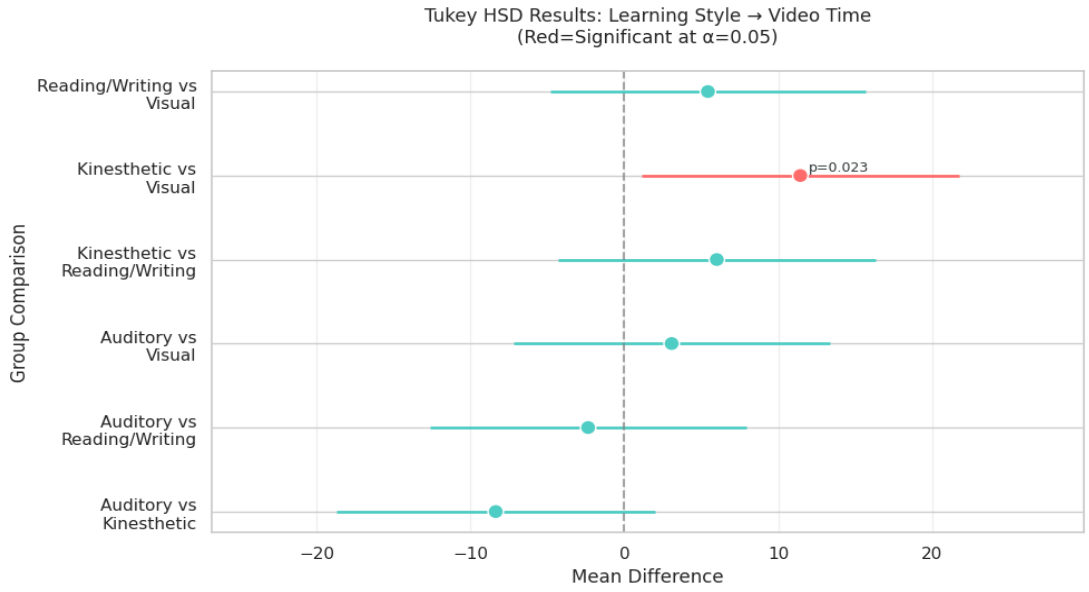


Figure 13: Tukey HSD Pairwise Comparisons: Learning Style vs Video Time

In the comparison between *Learning Style* and *Time Spent on Videos*, only one set of results was statistically significant: the p-value of the comparison between *Kinesthetic* and *Visual* was 0.023 (< 0.05). This result shows that students who prefer a kinesthetic learning style spend significantly more time watching videos than students who prefer a visual learning style.

Using Multiple Linear Regression to Analyze Factors Affecting Final Grades

Since the previous ANOVA analysis only revealed the differences in usage preferences and habits among different types of users, but did not directly test which factors would significantly affect the final exam score, we further used the multivariate linear regression method to explore the combined effect of multiple variables.

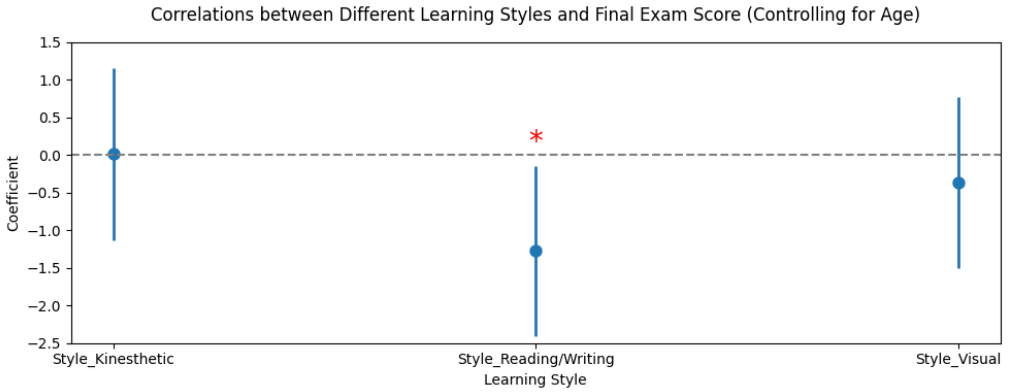
In this stage, we focus on the relationships of core variables on the final exam score under the premise of controlling other variables. We build regression models around the following three research questions:

1. ***Is there any significant relationship between the time spent watching videos and the final score?***
2. ***Is there any significant relationship between the engagement level and the final score?***
3. ***Are there any differences in the final scores among students with different learning styles?***

In the model, we set *Final_Exam_Score* as the dependent variable, *Time_Spent_on_Videos*, *Engagement_Leve*l and *Learning_Style* as the main independent variables, and included the corresponding control variables to establish a regression model.

The regression results show that in research questions 1 and 2, the influence of the main variables on the final score did not reach the statistical significance level, so no clear conclusion can be drawn.

For the third question, since *Learning_Style* is a categorical variable, we use One-Hot Encoding to process it and convert it into a numerical form that the model can recognize. To avoid the dummy variable trap, we set the *Auditory* type as the reference group, so the following three dummy variables are introduced into the model: *Style_Kinesthetic*, *Style_Reading/Writing* and *Style_Visual*.



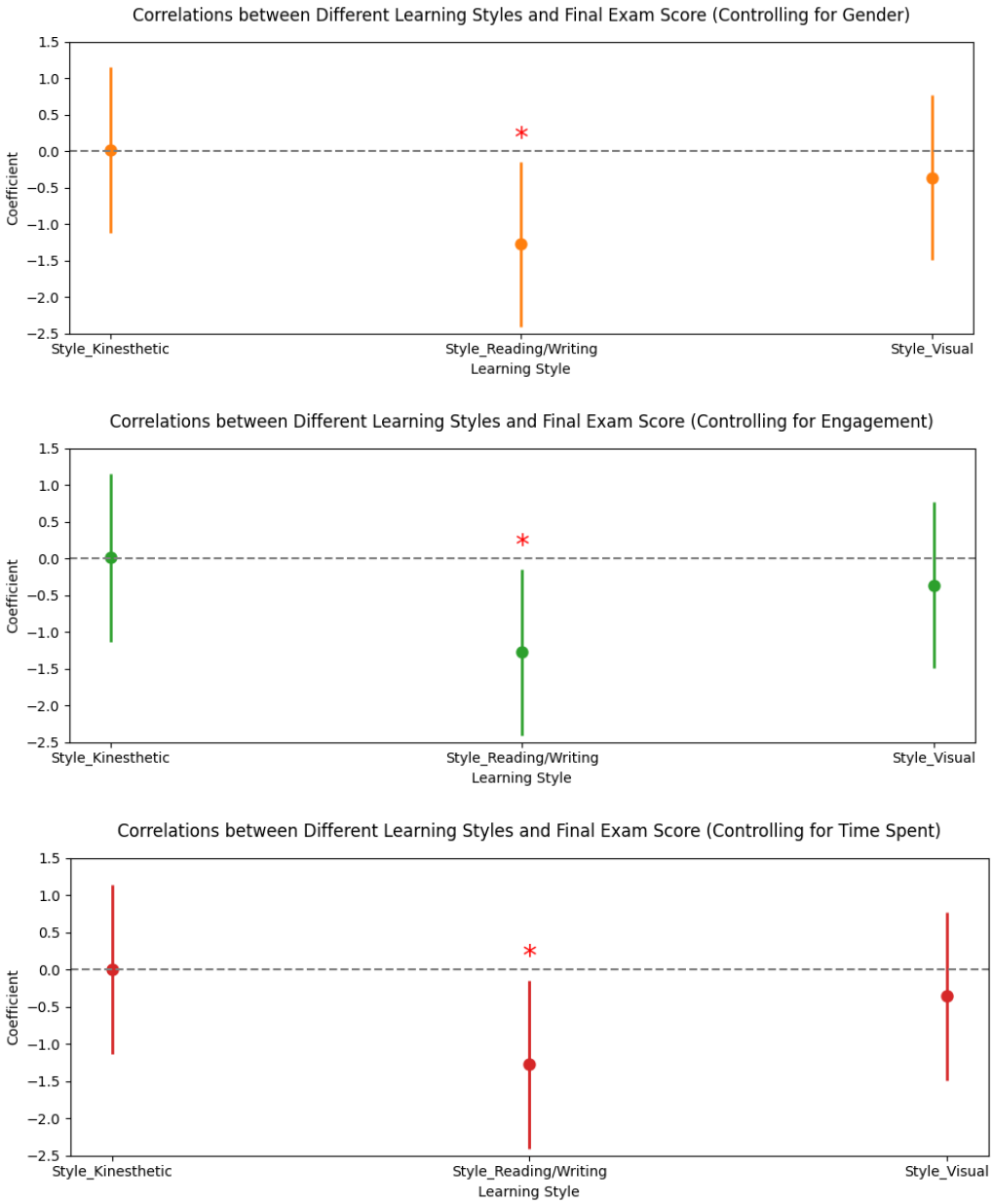


Figure 14: Correlations between Learning Style and Final Exam Score (with Different Control Variables)

The results of regression analysis show that the final grades of students with a reading/writing learning style are significantly lower than those of auditory students, with a difference of about -1.27 points, and the p value in all models is 0.027, reaching the statistical significance level ($p < 0.05$).

Since multiple potential interference variables (including age, gender, engagement level, and time spent on videos) were controlled in each regression model, the same conclusion was consistently drawn. Therefore, we can safely believe that the result is not caused by the synergy of other variables, but the independent influence of learning style itself on the final grade.

In summary, even without controlling any variables, the final grades of reading/writing students are significantly lower than those of auditory students, and this difference is statistically significant.

Self-defined Research Questions

Given the large number of post-hoc comparisons between pairs of variables in the previous analyses, we proposed the following exploratory research questions to further investigate potential interaction relationships:

1. ***Is there a significant relationship between video time and final score for students with different learning behaviour patterns?***
2. ***Is there a significant relationship between video time and final score for students with different learning styles?***

To address the first question, we combined two key dimensions — *Learning Style* and *Engagement Level* — and applied the median split method to categorize each student as either "high" or "low" on each dimension. Based on this classification, students were grouped into several distinct learning behaviour patterns, which were then used for subsequent comparative analysis.

Student Type	Completion Rate	Engagement Level
Disengaged Type	Low	Low
Active Type	High	High
Efficient Completer	High	Low
Independent Explorer	Low	High

Table 2: Classification of Student Types Based on Completion Rate and Engagement Level

Finally, we performed regression analysis and visualized it:

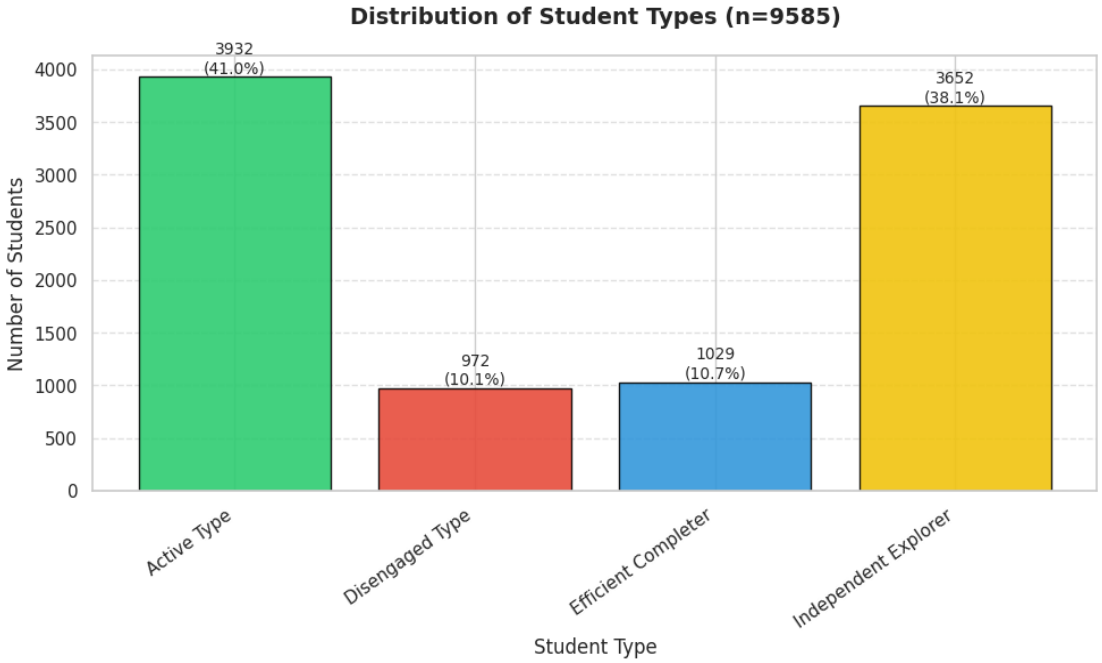


Figure 15: Distribution of Student Behavioural Types

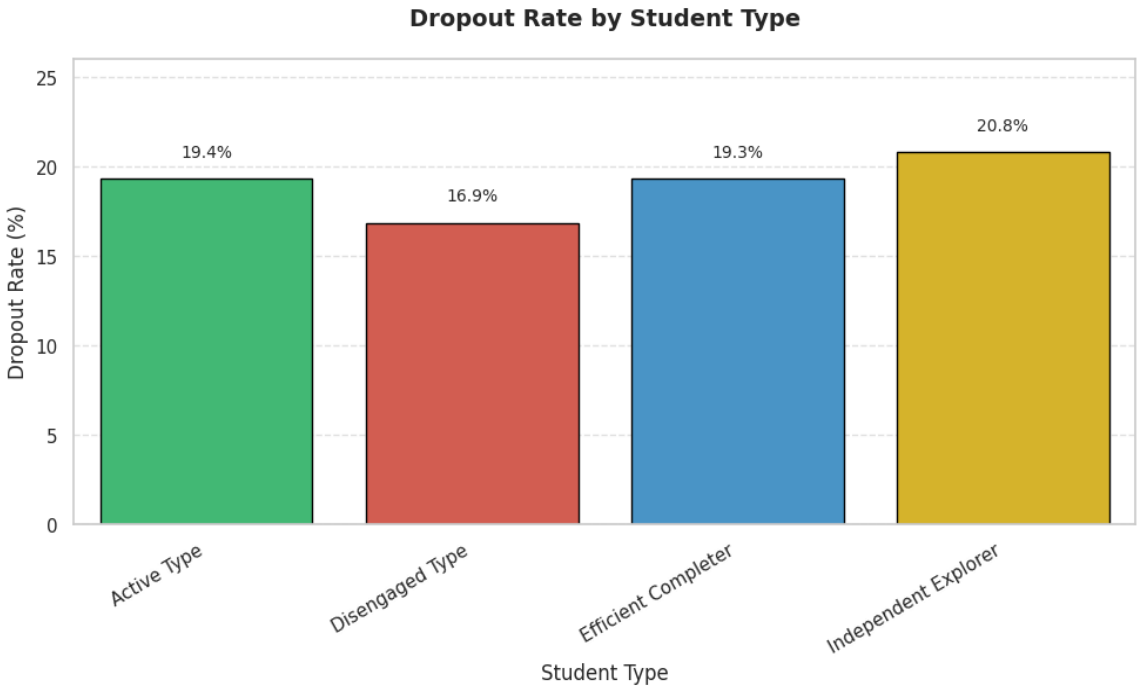


Figure 16: Dropout Rates Across Student Behavioural Types

Through regression analysis, we found that the dropout rate of students defined as “disengaged type” was significantly lower than that of other types of students ($Z = -2.303$, $p = 0.0213 < 0.05$).

This result is contrary to traditional cognition, indicating that such students may still persist in completing the course even when their class participation and homework completion rates are low.

In contrast, the dropout rate of "independent explorer" students is the highest. Although they are active in class interaction, their homework completion rate is relatively low, suggesting that such students may have certain learning motivation, but lack sufficient task execution, or find it difficult to maintain continuous learning investment without external feedback support.

To address the second research question, we conducted four separate linear regression analyses using learning style as the grouping variable, examining whether video viewing time predicted final exam scores within each group. The results showed that only among students with a reading/writing learning style, video viewing time was negatively associated with final scores—that is, the more time they spent watching videos, the lower their performance. This suggests that video-based learning, which primarily relies on visual and auditory modalities, may not align well with the preferences of reading/writing learners and could even hinder their learning effectiveness.

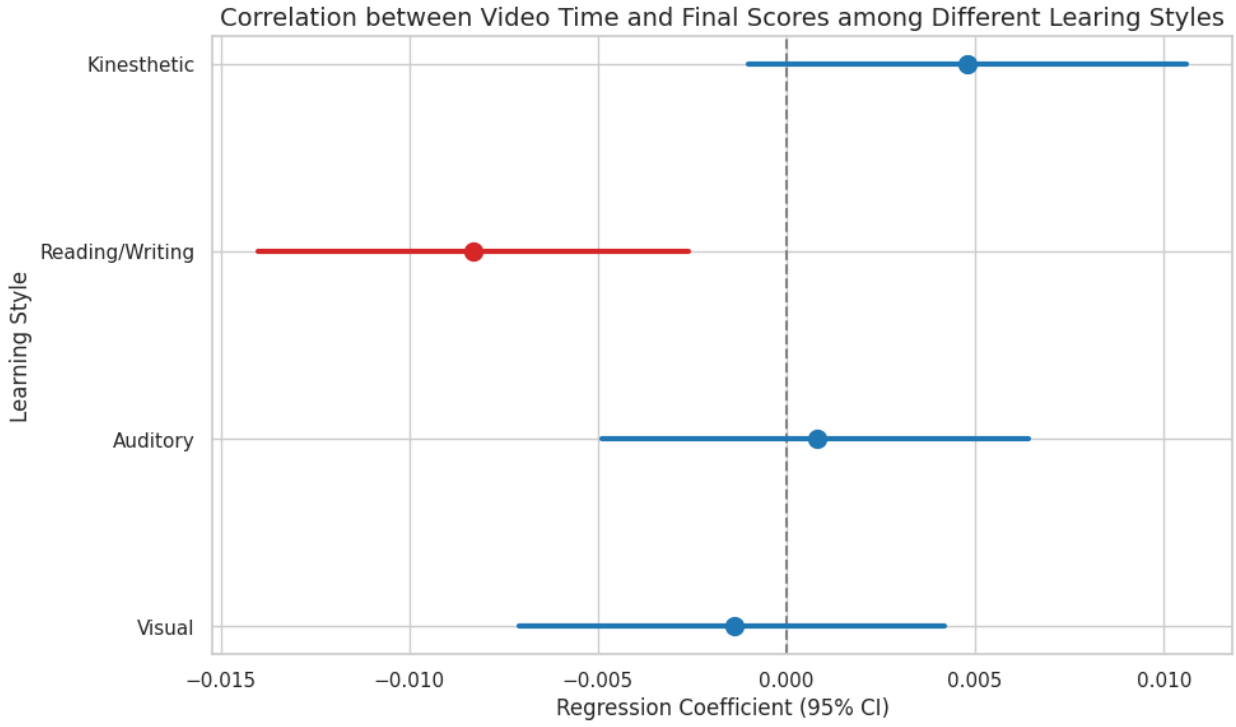


Figure 17: Effect of Video Viewing Time on Final Scores Across Learning Styles

Conclusion

This project investigated the relationship between adaptive learning behaviours and academic outcomes through five research questions. First, while overall video-watching time showed no significant association with final exam scores, a notable negative correlation emerged among reading/writing-preference learners, suggesting that excessive reliance on visual/auditory resources may hinder their performance. Second, engagement levels did not directly predict final scores, though highly engaged students reported significantly higher course satisfaction compared to moderately engaged peers. Third, learning styles exerted a measurable influence: reading/writing learners scored lower than auditory learners, even after controlling for variables like age and participation patterns. Fourth, the interaction between learning style and video time revealed that only reading/writing learners experienced a decline in scores with increased video consumption, highlighting a mismatch between their preferred modalities and resource design. Finally, behavioural typologies demonstrated divergent dropout risks: "independent explorers" (high engagement, low assignment completion) exhibited the highest attrition, whereas "disengaged" students (low engagement and completion) paradoxically showed lower dropout rates, challenging conventional assumptions about persistence. These findings underscore the need for personalized resource allocation and adaptive interventions tailored to learning preferences and behavioural patterns to optimize educational outcomes.

(word count 2452)

Declaration

The members of this group did knowingly use generative Al tools in this assignment task.

Acknowledgment

- In this assignment, we followed the University's guidelines for students on academic integrity. No content generated by generative Al tools has been presented as our own work. We take responsibility for the work submitted.
- Process: In this assignment preparation, we acknowledge the use of [Chat GPT-4o at https://genai.hkbu.edu.hk/] to [paraphrase elements of texts] for [clarifying language and accuracy], and [fix incorrect codes] for [running the programme correctly].