Team-1

DATS 6103: Final Project Paper

Professor Ning Rui

December 11, 2024

***Exploring the Parameters of VR in Education​: Clustering Analysis of VR Learning Patterns and Student Profiles​***

Abstract:

1. Introduction:
2. Virtual Reality in Education:
3. Description of Data:

We got the relevant dataset from Kaggle(https://www.kaggle.com/datasets/waqi786/impact-of-virtual-reality-on-education/data). It contains some relevant parameters which help to analyze the impact of virtual reality (VR) on education, highlighting its potential to enhance learning experiences across various subjects.

Total Observations: 5000

Total Variables: 20(Categorical: 18, Numeric: 2)   
  
\*\*\*\*\* Snap of the variables and few rows from the df

1. Data Preprocessing:

To avoid getting inaccurate results due to inconsistencies, errors or irrelevant information in the raw data, we followed the following crucial steps to pre-process data:

* Missing Values: The dataset has 0 missing values.
* Noise: There were only 2 numerical data which we checked and didn’t get any outlier that can alter the actual insights.
* Inconsistencies: No error found due to wrong formatting or data entry which could create mismatches.
* Data Suitability: Few of the columns contain numerical values, which are basically ordinal variables. We ensured.

The final summary of the data set:

1. A blue circle with black text

   Description automatically generatedExploratory Data Analysis:

At first, our goal was to identify what percentage of the total students are using VR for their educational purposes. We discovered that the segment is quite good with more than 50% of the students.

Then, we wanted to understand the underlying distribution of VR usage hours. And, we got this,

A graph of a person using vr usage

Description automatically generated A graph with a red line

Description automatically generated

*Image 3.1: Distribution of VR Usage Hours (histogram and Q-Q Plot)*

The histogram shows a non-symmetric distribution, suggesting potential skewness or that a large number of users have very high VR usage. The Q-Q plot further supports this, as points significantly deviate from the red diagonal line, particularly in the tails, indicating that the data does not follow a normal distribution.

5.1: VR Usage Patterns by Subject:

A bar graph with green and orange bars

Description automatically generated A row of colorful objects

Description automatically generated with medium confidence

The bar plot shows that the count of students using VR is consistent across subjects, with a slight preference for non-VR usage in most subjects. The violin plot reveals that the distribution of hours of VR usage per week is fairly similar across subjects, with a median usage of around 4-5 hours. Some variability exists in usage patterns, as indicated by the spread of the distributions, but no significant outliers or extreme values are visible.

5.2: VR Usage Patterns by Field of Study:

A bar graph with green and orange bars

Description automatically generated A chart of different colored rectangular shapes

Description automatically generated

The bar plot shows that VR usage is evenly distributed across fields of study, with slightly more users in fields like Medicine and Law. The box plot reveals that the hours of VR usage per week are consistent across all fields, with medians around 4-6 hours. There is a similar spread of VR usage across all fields, with no significant outliers or extreme variations.

5.3: VR Usage Patterns by Grade level:

A bar chart of a bar graph

Description automatically generated A diagram of a box plot of grade level

Description automatically generated

**6. Hypothesis Testing to Analyze Relationships Among Variables and Assess Their Statistical Significance**

We performed hypothesis testing is to explore and validate the relationships between different variable pairs in the dataset to determine whether these relationships are statistically significant and to evaluate patterns or associations, ensuring they are not due to random chance.

6.1: Relationship between VR Usage and Gender

To determine whether VR usage is associated with gender, we conducted a Chi-Square test. The analysis resulted in a p-value of 0.39, which is greater than the significance level of 0.05. Based on this result, we concluded that there is no statistically significant association between VR usage and gender.

To examine whether there is a significant difference in the mean VR usage hours across genders, we performed an ANOVA test. The test yielded a p-value of 0.23, which is greater than the significance level of 0.05. Thus, we concluded that there is no statistically significant difference in VR usage hours across genders.

6.2: Relationship between VR Usage and Subject

A Chi-Square test was conducted to determine if VR usage is associated with the subject of study. The test resulted in a p-value of 0.63, which is greater than the significance level of 0.05. This indicates that there is no statistically significant association between VR usage and the subject of study.

An ANOVA test was performed to examine whether there is a significant difference in the mean VR usage hours across subjects. The test produced a p-value of 0.395, which is greater than the significance level of 0.05. This suggests that there is no statistically significant difference in VR usage hours across subjects.

6.3: Relationship between VR Usage and Academic Outcome

A Chi-Square test was conducted to assess whether academic outcome is associated with VR usage. The test yielded a p-value of 0.84, which is greater than the significance level of 0.05. Therefore, we conclude that there is no statistically significant association between academic outcome and VR usage.

A logistic regression analysis was conducted to determine whether VR usage hours significantly predict academic outcomes. The resulting p-value of 0.856, which is greater than the significance level of 0.05, indicates that the number of hours of VR usage per week does not significantly predict academic outcomes in this dataset.

6.4: Relationship between VR Usage and Engagement Level

A Chi-Square test was performed to evaluate whether engagement level is associated with VR usage. The test resulted in a p-value of 0.75, which is greater than the significance level of 0.05. Thus, we conclude that there is no statistically significant association between engagement level and VR usage.

A linear regression analysis was conducted to determine if VR usage hours significantly influence engagement levels. The resulting p-value of 0.679, which is greater than the significance level of 0.05, indicates that VR usage hours do not have a statistically significant influence on engagement levels.

6.5: Relationship between Instructor VR Efficiency and VR Usage

A Chi-Square test was conducted to assess whether instructor VR efficiency is associated with the perceived effectiveness of VR. The test produced a p-value of 0.41, which is greater than the significance level of 0.05. Therefore, we conclude that there is no statistically significant association between instructor VR efficiency and perceived effectiveness of VR.

A Chi-Square test was conducted to determine if instructor VR efficiency is associated with students’ interest in continuing VR-based learning. The test yielded a p-value of 0.54, which is greater than the significance level of 0.05. Therefore, we conclude that there is no statistically significant association between instructor VR efficiency and students’ interest in continuing VR-based learning.

Feature Importance:

SMART Questions:

1. How does the instructor's VR proficiency affect students' improvement in learning outcomes?
2. What are the key distinguishing features between high-performing and low-performing clusters?
3. How do cluster characteristics vary across different regional and support system contexts?
4. How do engagement levels in VR correlate with academic outcomes within each identified cluster?

Question-1:

How does the instructor's VR proficiency affect students' improvement in learning outcomes?

Answer:

With the objective of examining the relationship between instructor's VR proficiency and students' improvement in learning outcomes, a one-way ANOVA test was performed. The data was first grouped by instructor VR proficiency, and the improvement in learning outcomes was analyzed across these groups.

The ANOVA test resulted in a p-value of 0.972 and an F-statistic of 0.280. Since the p-value is much greater than the significance level of 0.05, we conclude that there is no statistically significant relationship between instructor's VR proficiency and students' improvement in learning outcomes. Therefore, instructor VR proficiency does not appear to significantly influence students' improvement in learning outcomes in this dataset.

Question-2:

What are the key distinguishing features between high-performing and low-performing clusters?

Answer:

Limitations:

Conclusion: