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:**

Why do your team choose this topic?

1. **Virtual Reality in Education:**

What prior research and analysis have been done on this topic?

1. **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)

A snap of the variables and few rows from the dataframe is shared below:

A black and white image of a movie strip

Description automatically generated

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:**

We started with EDA to understand the structure, patterns, and anomalies in our dataset.

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.

*Image 5.1: VR Usage Percentages*

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 5.2: 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

*Image 5.3: VR Usage Patterns by Subject (Bar and Violin Plot)*

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

*Image 5.4: VR Usage Patterns by Field of Study (Bar and Box Plot)*

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

*Image 5.5: VR Usage Patterns by Grade Level (Bar and Violin Plot)*

From the bar plot on the left, we observe that the count of VR users ("Yes") and non-users ("No") is relatively balanced across all grade levels (Postgraduate, Undergraduate, and High School). However, the undergraduate group has slightly higher VR usage compared to other grade levels.

The violin plot on the right shows the distribution of hours of VR usage per week for each grade level. Across all groups, the median usage is close to zero, suggesting most students spend minimal time on VR weekly. However, the distribution is wider for undergraduates, indicating a more diverse usage pattern, including students who spend significant time using VR.

**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.

**7. Feature Importance:**

**8. SMART Questions:**

SMART questions help uncover underlying patterns and insights by providing clear, focused objectives that guide data exploration and analysis. We designed four SMART questions to examine relationships among different parameters, their statistical significance, and to determine whether using VR has had any significant impact from various perspectives.

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:

In an educational system, high-performing clusters consist of students who achieve consistently superior outcomes compared to others. These clusters often set benchmarks for academic excellence and demonstrate strong critical thinking, problem-solving, and collaborative skills.

Low-performing clusters, on the other hand, are characterized by consistently below-average outcomes. These groups often struggle with issues like lower student engagement, higher dropout rates, and poor performance on standardized tests.

Our objective was to investigate whether there are disparities in features between high-performing and low-performing clusters across various VR-related parameters such as VR usage (whether the student uses VR or not), hours spent using VR, and interest in continuing VR-based learning. This is an unsupervised learning method, as there are no predefined labels for high or low performance. So, these clusters were created based on patterns identified in the data. To assess performance, it is essential to define performance indicators. When we look at our dataset, we identified 3 parameters that we can consider:

* Engagement\_Level
* Improvement\_in\_Learning\_Outcomes
* Impact\_on\_Creativity

As the dataset is relatively small and the number of clusters is predetermined, we used the computationally efficient K-means clustering method, which is particularly suitable for such scenarios because it quickly partitions the data into distinct groups based on predefined cluster counts. Two clusters were created, and the quality of clustering was evaluated using the silhouette score, a metric that assesses how well each data point lies within its assigned cluster. The *silhouette score* was 0.35, indicating moderate separation with some overlap between clusters, meaning the clustering is better than random but not strongly distinct.

A screen shot of a graph

Description automatically generated

*Image 8.1: High and Low Performing Clusters using K-Means*

We then performed statistical t-tests on selected features to determine if there were significant differences between the high- and low-performing clusters. The resulting p-values were as follows:

|  |  |
| --- | --- |
| **Feature Name** | **p-value** |
| Usage\_of\_VR\_in\_Education | 0.82 |
| Hours\_of\_VR\_Usage\_Per\_Week | 0.86 |
| Instructor\_VR\_Proficiency | 0.97 |
| Access\_to\_VR\_Equipment | 0.57 |
| Collaboration\_with\_Peers\_via\_VR | 0.01 |
| Stress\_Level\_with\_VR\_Usage | 0.62 |
| Feedback\_from\_Educators\_on\_VR | 0.51 |
| Interest\_in\_Continuing\_VR\_Based\_Learning | 0.99 |
| Perceived\_Effectiveness\_of\_VR | 0.26 |

From the p-values, we are observing that "Collaboration with Peers via VR" is the only feature that significantly distinguishes the high-performing and low-performing clusters. Other features showed no significant differences.

Next, we conducted an in-depth analysis and statistical testing on various student attributes, such as gender, grade level, field of study, region, and school support systems, to determine if there were any significant differences between the high- and low-performing clusters.

From the resulting p-values, we observed a significant difference in VR usage for education based on gender. No significant differences were found in grade level. However, for field of study, school support systems, and region, VR usage significantly differed between the high- and low-performing clusters, similar to gender. Additionally, in the regional context, collaboration with peers via VR also showed a significant difference.

Although we expected there to be differences in features between high and low-performing clusters, the detailed analysis showed that no visual or statistical differences exist between them.

Question-3:

How do cluster characteristics vary across different regional and support system contexts?

Answer:

The objective was to analyze how cluster characteristics vary across different regional and support system contexts.

In our dataset, there are 6 regions representing 6 continents. Comparing from a regional context in statistical testing is important because regional factors, such as access to resources, cultural differences, and varying educational support systems, can significantly influence student performance and behavior. Regional disparities can uncover trends or patterns that might be masked in a broader analysis, helping to identify specific needs or opportunities for targeted interventions. Additionally, understanding regional context ensures that findings are more applicable and relevant to different areas, allowing for more effective policy decisions and resource allocation.

From the ANOVA test for regional context, we found that "Collaboration with Peers via VR" significantly varies across regions. Other variables showed no significant regional variation. This we can consider a good one because all the students of the region are getting similar opportunities and responding the same way. If the dataset had more granular breakdown, like sub-continents or countries of the regional data, we might have observed in-depth insights here.

|  |  |
| --- | --- |
| **Feature Name** | **p-value** |
| Usage\_of\_VR\_in\_Education | 0.81 |
| Hours\_of\_VR\_Usage\_Per\_Week | 0.31 |
| Instructor\_VR\_Proficiency | 0.61 |
| Access\_to\_VR\_Equipment | 0.77 |
| Collaboration\_with\_Peers\_via\_VR | 0.01 |
| Stress\_Level\_with\_VR\_Usage | 0.26 |
| Feedback\_from\_Educators\_on\_VR | 0.78 |
| Interest\_in\_Continuing\_VR\_Based\_Learning | 0.72 |
| Perceived\_Effectiveness\_of\_VR | 0.08 |

We also performed a t-test for the support system-based context and found that none of the variables exhibited significant variation, suggesting that support systems may not strongly differentiate the characteristics of VR usage in education.

|  |  |
| --- | --- |
| **Feature Name** | **p-value** |
| Usage\_of\_VR\_in\_Education | 0.74 |
| Hours\_of\_VR\_Usage\_Per\_Week | 0.76 |
| Instructor\_VR\_Proficiency | 0.2 |
| Access\_to\_VR\_Equipment | 0.5 |
| Collaboration\_with\_Peers\_via\_VR | 0.29 |
| Stress\_Level\_with\_VR\_Usage | 0.78 |
| Feedback\_from\_Educators\_on\_VR | 0.27 |
| Interest\_in\_Continuing\_VR\_Based\_Learning | 0.82 |
| Perceived\_Effectiveness\_of\_VR | 0.68 |

Question-4:

How do engagement levels in VR correlate with academic outcomes within each identified cluster?

Answer:

Limitations:

Conclusion: