ECE 20875

PATH 1: Student performance related to video-watching behavior

FAHEEM MEKRANI (FMEKRANI)
PUID: 0037465172

mini project repository link:

https://github.com/ECEDataScience/project-s25-fmekrani

Describing the dataset:

For Path 1 of the mini project, we are given a dataset containing information about students who watched educational videos and their performance on in-video quizzes. Each row in the dataset represents one instance of a student watching a video and contains 10 fields. The first field is userID, which is a unique identifier for each student. The second is videoID, indicating the ID of the video watched, ranging from 0 to 92 (since there are 93 videos in total). The third column is fracSpent, which denotes the fraction of the video duration that the student spent watching. The next field, fracComp, shows the fraction of the video completed by the student, with values ranging from 0 to 0.9. Following that is fracPaused, which indicates the fraction of time the student spent paused during the video relative to its length.

The dataset also includes numPauses, the number of times the student paused the video, and avgPBR, the average playback rate used, which ranges from 0.5x to 2x speed. Additionally, numRWs records how many times the student rewound the video, and numFFs shows how many times they fast-forwarded. Finally, the last field is s, which reflects whether the student answered the in-video quiz question correctly on their first attempt (s = 1) or not (s = 0). This dataset provides comprehensive behavioral and performance data for analyzing student engagement and learning outcomes.

1. How well can the students be naturally grouped or clustered by their video-watching behavior?
2. Can student's video- watching behavior be used to predict a student's performance? Will adding the cluster information help the model improve the performance and makes the model become more/less underfit/overfit? Explain your conclusion and discuss the reason why such results could happen.
3. How well can you predict a student's performance on a particular in-video quiz question based on their video-watching behaviors while watching the corresponding video?

Why did I choose the analysis I chose?

Problem 1:

For Question 1, I performed unsupervised clustering to identify natural groupings of students by their video-watching behavior. To ensure that each student had enough interaction data to meaningfully analyze, I filtered out users who had watched fewer than five videos. used the KMeans algorithm because it is simple and effective at grouping similar data points. I employed the elbow method to determine the optimal number of clusters, plotting the within-cluster sum of squares (WCSS) for various numbers of clusters and looking for the point at which the improvements level off. For evaluating cluster quality, I calculated the silhouette score, which measures how well each data point fits within its assigned cluster. This analysis should reveal distinctive viewing habits, adding to uncovering whether students naturally fall into behavior-based segments—such as those who watch videos holistically versus those who pause or skip frequently—perhaps informing targeted learning efforts.

Problem 2:

To see whether students' video watching could be used as a predictor for their study performance, I did a linear regression on aggregated user behavior per user. I first ordered the data by user ID to compute summary statistics: average fraction of video spent watching (avg_fracSpent), average fraction of time spent paused (avg_fracPause), and student's last performance, which was the mean score (final_score) for all quizzes. This dataset allows us to examine the overall trends in the ways in which a particular student interacts with the material to be mastered and how it relates to performance.

I then built two separate linear regression models. The first model employs avg_fracSpent as a predictor of final_score, whether students who view more of each video score higher. The second model uses avg_fracPause as the predictor to examine whether pause behavior is linked to outcomes. For both models, I considered the strength of the relationship based on the R-squared metric (indicating how well the predictor explains the variance in performance) and mean squared error (indicating how accurate predictions are). This enables consideration of whether simple behavioral metrics are strong predictors of student performance or if additional factors are stronger predictors. Ultimately, it can guide whether instructors should place emphasis on developing specific video habits to facilitate enhanced learning.

problem 3:

To further investigate the interaction between video-watching habits and in-video quiz performance, I made a set of scatter plots to see these interactions at a more detailed scale. As a first step, I graphed each student's average fraction of video watched (avg_fracSpent) and average fraction of time paused (avg_fracPause) against their final score. This approach enables us to see whether students who are more engaged with video content—by watching more or pausing to reflect—tend to do better on quiz questions. I also plotted the correlation between fracSpent and fracPaused for all video-watching sessions, which helps to find clusters or common interaction patterns in how students go through the material on a per-video basis.

These plots are helpful in that they can identify trends, outliers, or special patterns that may not be apparent from summary statistics. Scatter plots also make both linear and non-linear relationships easier to discern, and any clustering of behaviors easier to spot. If strong correlations are apparent, this adds credibility to the hypothesis that some video engagement metrics are good predictors of quiz performance. If so, it means that other variables have more impact. This analysis guides further feature selection and modeling.

Results for my analysis:

Analysis Problem 1:

To identify the appropriate number of clusters, I applied both the elbow method and the silhouette score. The **elbow plot** (see Figure 1) suggests diminishing returns for additional clusters beyond 4 to 6, indicating a natural grouping around this range. To further validate the number of clusters, I calculated silhouette scores for 6, 7, and 8 clusters:

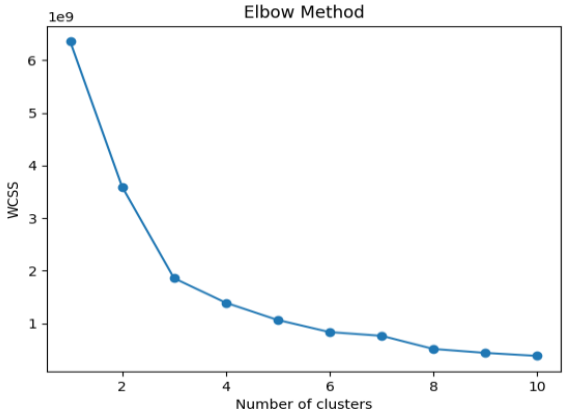
- **6 clusters:** Silhouette score = **0.9638**
- **7 clusters:** Silhouette score = **0.9438**
- **8 clusters:** Silhouette score = **0.9516**

The highest silhouette score was achieved with **6 clusters**. A silhouette score close to 1 indicates that clusters are both well-separated and internally coherent, meaning students in the same cluster have very similar behaviors, while those in different clusters behave quite differently. While silhouette scores for 7 and 8 clusters also remained high, the highest value was observed at 6 clusters, supporting this choice as the most effective grouping

Conclusion:

Based on these results, students' video-watching behaviors can be very effectively clustered into 6 distinct groups, with minimal overlap between clusters. This high level of separation suggests strong, naturally occurring patterns in how students engage with video content, which could help in identifying distinct learning strategies or tailoring interventions.

Visual Aid:
The attached elbow plot visually supports the selection of 6 clusters as an optimal balance between tight clusters and model simplicity.



Analysis Problem 2:

To determine if students' video-watching behavior could be used to predict their overall performance (average score across all quizzes), I built two simple linear regression models. The first model predicted final score using each student's average fraction of video watched (avg_fracSpent), while the second used their average fraction of time spent paused (avg_fracPause). The results were as follows:

Model 1	R^2	0.0008
	MSE	0.0436

Model 2	R^2	0.0013
	MSE	0.0436

Both models produced very low R^2 values (close to zero), indicating that neither the average fraction of video watched nor the average time paused explains much of the variance in students' quiz performance. The nearly identical mean squared error values further suggest that these video-watching behaviors have little predictive value for quiz outcomes when used alone.

These results show that, at least with simple linear models and the available behavioral data, video-watching metrics do not meaningfully predict students' average quiz scores. This may be because quiz performance depends on factors such as prior knowledge, study habits, or engagement outside of video content, which are not captured by these features.

It is also possible that quizzes test broader concepts or skills beyond those directly addressed in the videos.

Adding cluster information (e.g., group membership from KMeans clustering) as an additional feature could potentially improve prediction if specific behavioral types are associated with performance. However, given the very low R^2 values found here, it is likely that even with cluster information, predictive accuracy would remain limited unless combined with other relevant variables.

In summary, students' video-watching behavior alone does not appear to be a strong predictor of quiz performance in this dataset, highlighting the need to consider additional factors or more complex modeling approaches.

Analysis problem 3:

To evaluate whether a student's video-watching behaviors could predict their performance on specific in-video quiz questions, I visualized the relationship between key behavioral metrics and quiz outcomes using scatter plots.

Figures 1 and 2 show the relationship between a student's average fraction of video watched (avg_fracSpent) and average fraction of time paused (avg_fracPause) versus their final quiz score. Both plots reveal a large concentration of data points clustered near lower values on the x-axis, with scores ranging broadly across the y-axis. There is no clear trend or linear relationship between either behavioral metric and quiz performance. Notably, students who spent more time watching or pausing the videos did not consistently achieve higher scores, and there are substantial outliers in both directions.

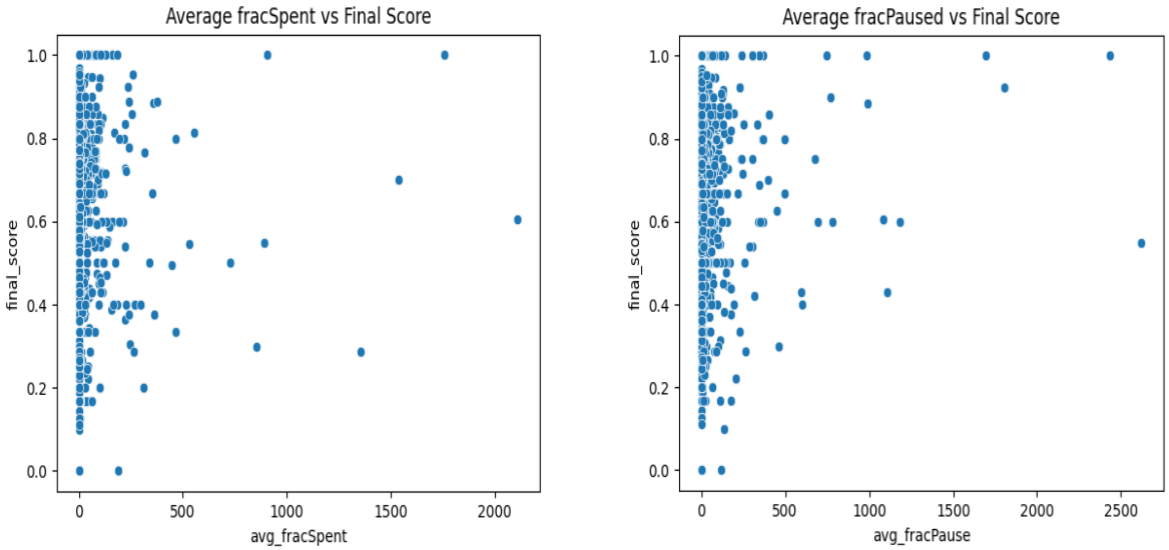
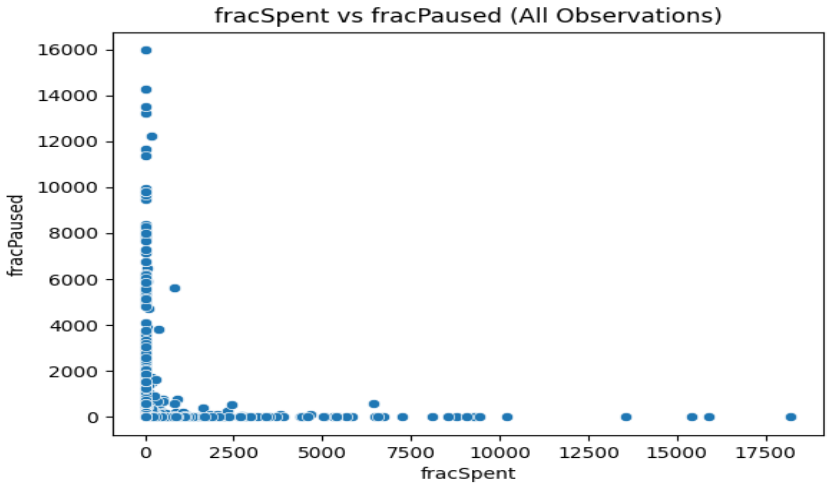


Figure 3 presents the relationship between fraction watched (fracSpent) and fraction paused (fracPaused) for all observations. Most data points cluster close to the origin, suggesting that the majority of students do not spend excessive time watching or pausing. However, the lack of clear structure or grouping further supports the absence of a strong relationship between these behaviors.



Conclusion :

The visual evidence proves that video-viewing behavior is not highly related to individual question performance on a quiz. Without the apparent patterns or correlations

discovered in the scatter plots, these results suggest other variables—prior knowledge, study habits, and outside learning sources—may carry more influence when calculating quiz scores. Particularly, neither the average fraction of video watched (avg_fracSpent) nor the average fraction of paused time (avg_fracPause) display any consistent relationship with in-video quiz performance in this data. These results highlight the richness of learning processes and suggest that simple behavioral measures like these may not be revealing the full picture of student performance.