Clustering Student Engagement in Online Learning Environments

By Victoria Hernandez

Introduction

A student's educational success rely heavily on a student's engagement, especially in online educational environments. (Bergdahl et al., 2024) The COVID-19 pandemic significantly shifted educational practices worldwide and there has been an increase in the utilization of online learning. (Akpen et al., 2024) It is important to understand student engagement in online learning environments because it helps improve student performance and provides and opportunity for earl intervention for high risk students. Educators use this type of analysis to track student performance and student habits to adapt their teaching methods. The dataset used in this report was synthetically generated to stimulate real world student behaviors.

The dataset includes six attributes that represent different aspect student engagement. To analyze the data, exploratory data analysis was performed, and clustering methods were utilized. The exploratory analysis performed on the data was finding the minimum, maximum, mean, standard deviation, 1st quartile, and 3rd quartile. The data was also visualized using box plots. The preprocessing methods used on the data were checking for NA values and scaling each attribute on a scale from 0 to 1. The unsupervised machine learning method used in this report w K-means. The visualizations used were silhouette plots pairs pots, and an elbow plot.

Data

The dataset was synthetically generated using AI(ChatGPT) to reflect realistic student engagement behaviors in an online environment. The dataset contains 999 student records and has 6 attributes.

Attribute	Description	Type	Range
avg_session_time	Average time spent per session	Numeric	~14 to ~79 minutes
assignments_completed	Assignments completed in the past month	Numeric	0 to 22 assignments
video_watch_pct	Percentage of videos watch	Numeric	30% to 100%
forum_posts	Number of posts in the discussion forums	Numeric	0 to 15 posts

Quizzes_attempted	Number of quizzes attempted	Numeric	0 to 9 quizzes
Avg_score	Average quiz score	Numeric	40 to 100 percent

Methods

Data Preprocessing:

- The dataset was checked for missing value and there were none.
- Numeric Attributes were scaled to [0,1] using the rescale() function. Rescale uses min-max normalization.

Exploratory Data Analysis:

- Exploratory data analysis was performed using summary statistics using the summary() function R.
- Pairs Plots were created for visualization to show pattern in the data using the pairs() function in R.
- Box plots were utilized to visualize clustering results and to show different student engagement levels.

Clustering:

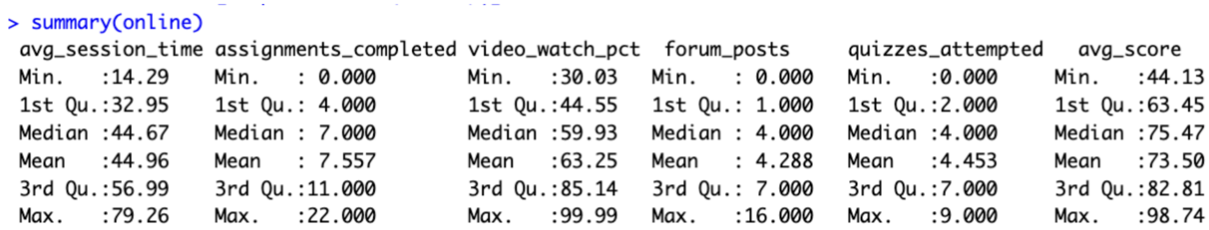
- **K-means:** K-means was used to cluster and analyze the dataset. K-means in an unsupervised machine learning method that partitions object. The k-means algorithm stores centroids that it uses to define clusters. (Piech, n.d.) In k-means a point is an individual cluster if it's closest to the centroid. (Piech, n.d.)
- **Silhouette coefficient:** This is a metric used to evaluate clustering quality. (Siemens, 2022) It uses a coefficient to determine the coherence of clusters. (Siemens, 2022) This method was used to find the most appropriate k value for our k -mean algorithm. A silhouette plot was used to visualize this method.
- **Elbow Method:** The elbow method helps find the ideal k for the number of clusters in K-means. (Zala, 2023) It calculates the total squared distances between datapoints and the data points cluster center. (Zala, 2023)

Use of generative AI: The AI tool OpenAI ChatGPT was used for this project. Generative AI was used to assist in creating a synthetic dataset about student online engagement. It was used to create a dataset with 6 attribute and 999 instances. The prompt used to generate the dataset was "Create a dataset that is well suited for clustering about a meaningful topic?".

Results

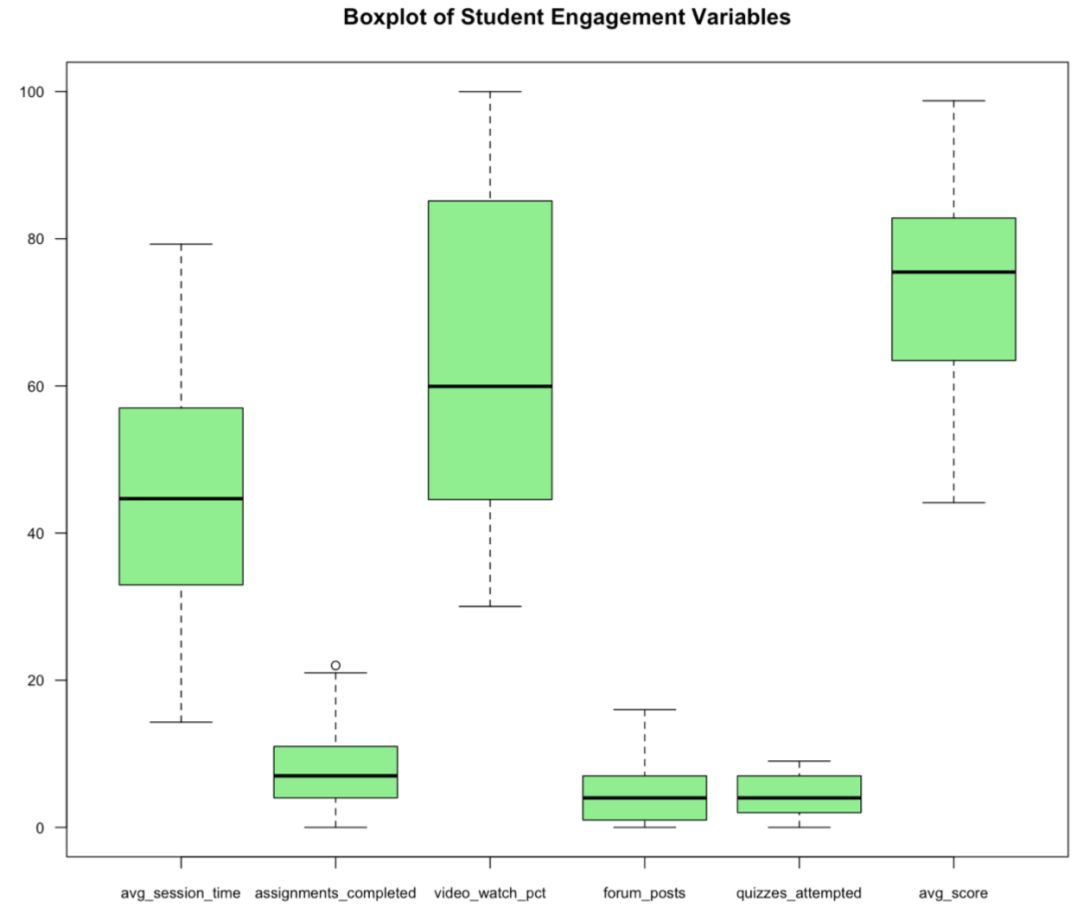
Exploratory analysis:

Summary:



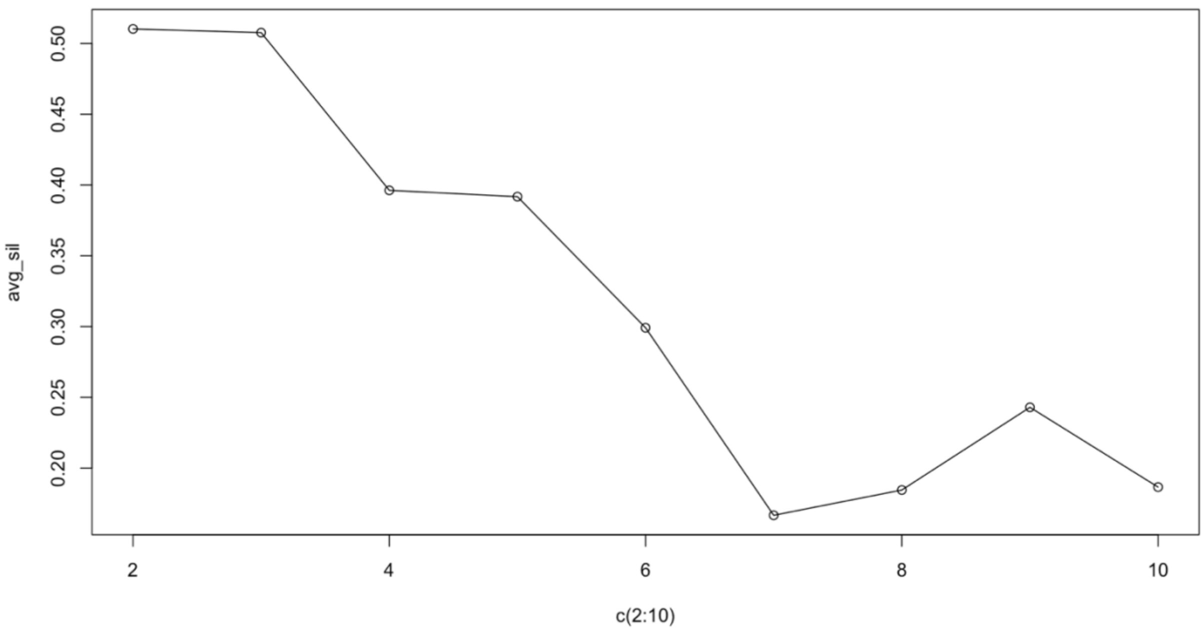
```
> summary(online)
 avg_session_time assignments_completed video_watch_pct  forum_posts     quizzes_attempted   avg_score
 Min.   :14.29    Min.   : 0.000        Min.   :30.03   Min.   : 0.000   Min.   :0.000     Min.   :44.13
 1st Qu.:32.95    1st Qu.: 4.000        1st Qu.:44.55   1st Qu.: 1.000   1st Qu.:2.000     1st Qu.:63.45
 Median :44.67    Median : 7.000        Median :59.93   Median : 4.000   Median :4.000     Median :75.47
 Mean   :44.96    Mean   : 7.557        Mean   :63.25   Mean   : 4.288   Mean   :4.453     Mean   :73.50
 3rd Qu.:56.99    3rd Qu.:11.000        3rd Qu.:85.14   3rd Qu.: 7.000   3rd Qu.:7.000     3rd Qu.:82.81
 Max.   :79.26    Max.   :22.000        Max.   :99.99   Max.   :16.000   Max.   :9.000     Max.   :98.74
```

Boxplot:



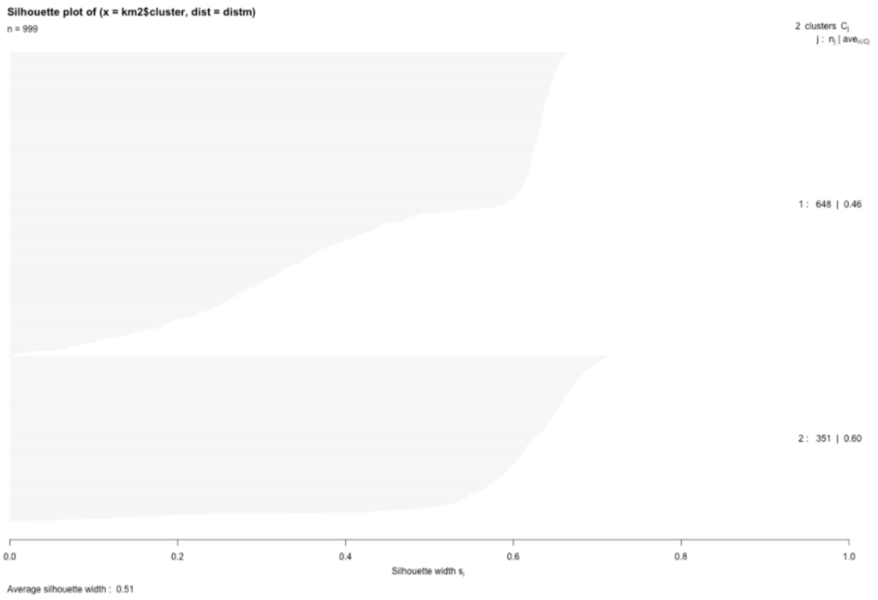
Clustering:

Figure 1. Elbow Plot using Average Silhouette Score



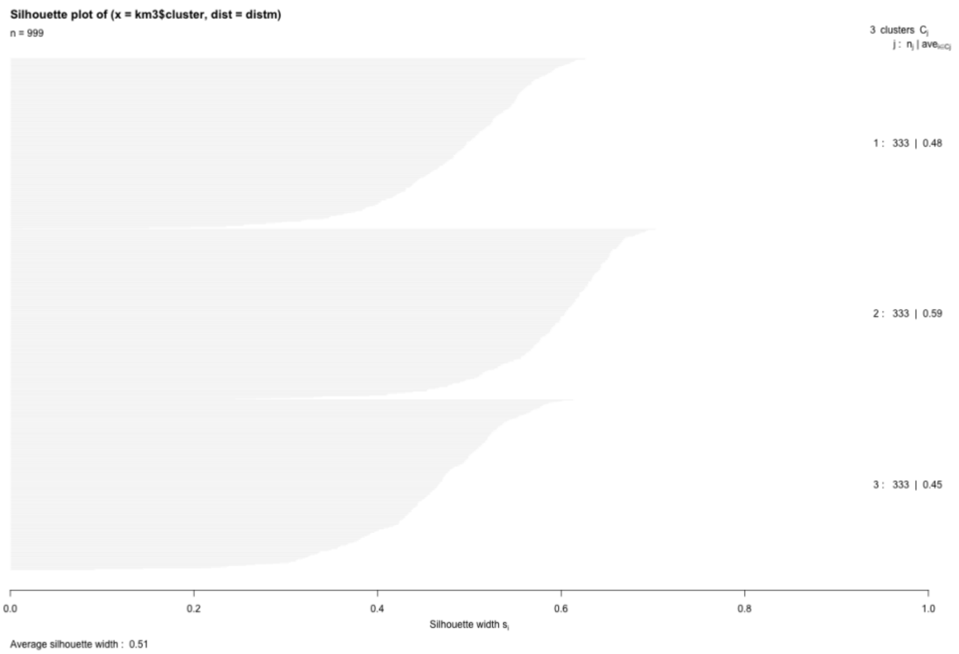
The elbow plot in Figure 1 shows the average silhouette width for k = 2 to 10. From the plot you can see that k = 2 and k = 3 resulted in the highest silhouette values at ~0.51. This indicates string clusters at those k values. The results from this elbow plot suggest that either k =2 or k =3 should be selected for analysis.

Figure 2. Silhouette plot of km2



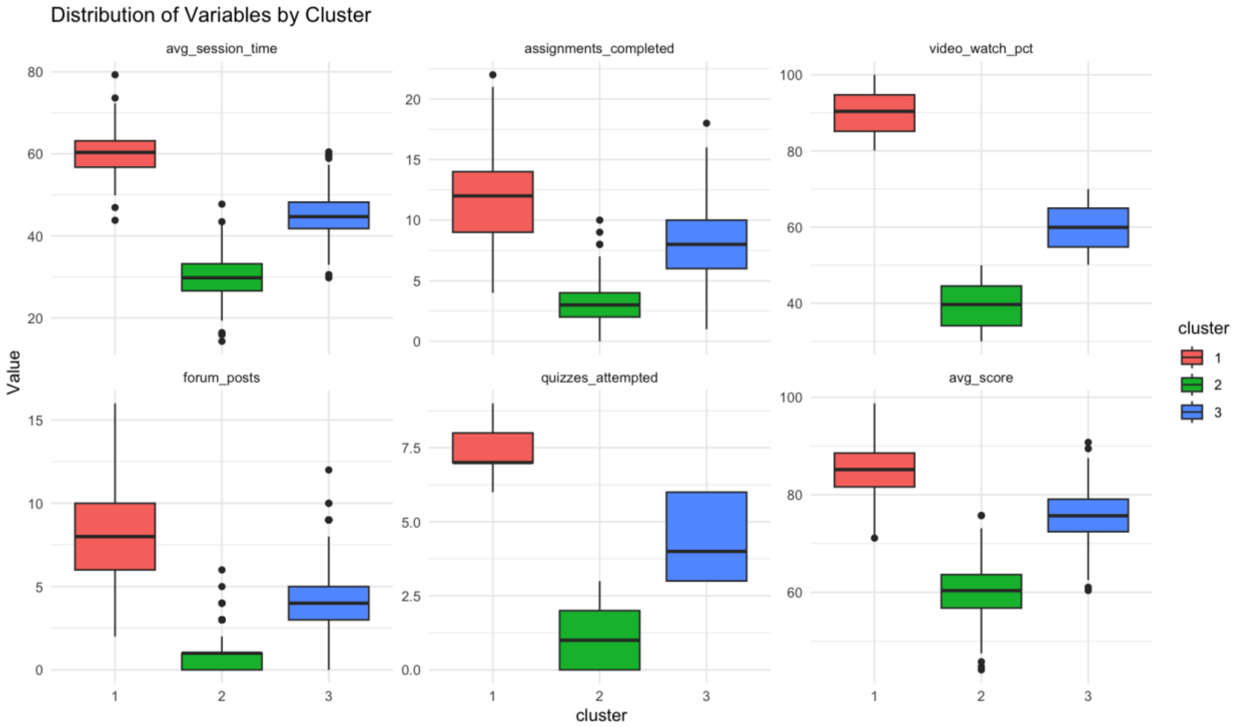
The silhouette plot for k = 2 shows a silhouette width of 0.51. Cluster 1 is the larger group with 648 data points and Cluster 2 is the smaller group with 351 data points. These silhouette widths show a clear separation between clusters. This k value would split the student groups in to two categories which could give us sufficient insights but may not provide a deeper evaluation.

Figure 3. Silhouette plot of km3.



The silhouette plot above in Figure 3 shows clustering for k = 3. The pot shows more equal balance between clusters with each cluster having 333 data points. This plot showed an average silhouette width of 0.51. The pot has nearly equal cluster size indicating well balanced clusters to represent each engagement level. The plot low, medium, and high student engagement groups. From this plot k = 3 seems to be a better fit for our analysis because it clear separation on different groups.

Figure 4. Boxplot of the Distribution of Variables by Clustering using k = 3.



The box plot in Figure 4 shows each variable and their clusters.

Cluster 1 (Red):

- This cluster show high student engagement.
- Highest session times, the most assignment completed, the most forum activity and the highest quiz scores.
- The students in this cluster most likely do very well academically and are actively engaged.

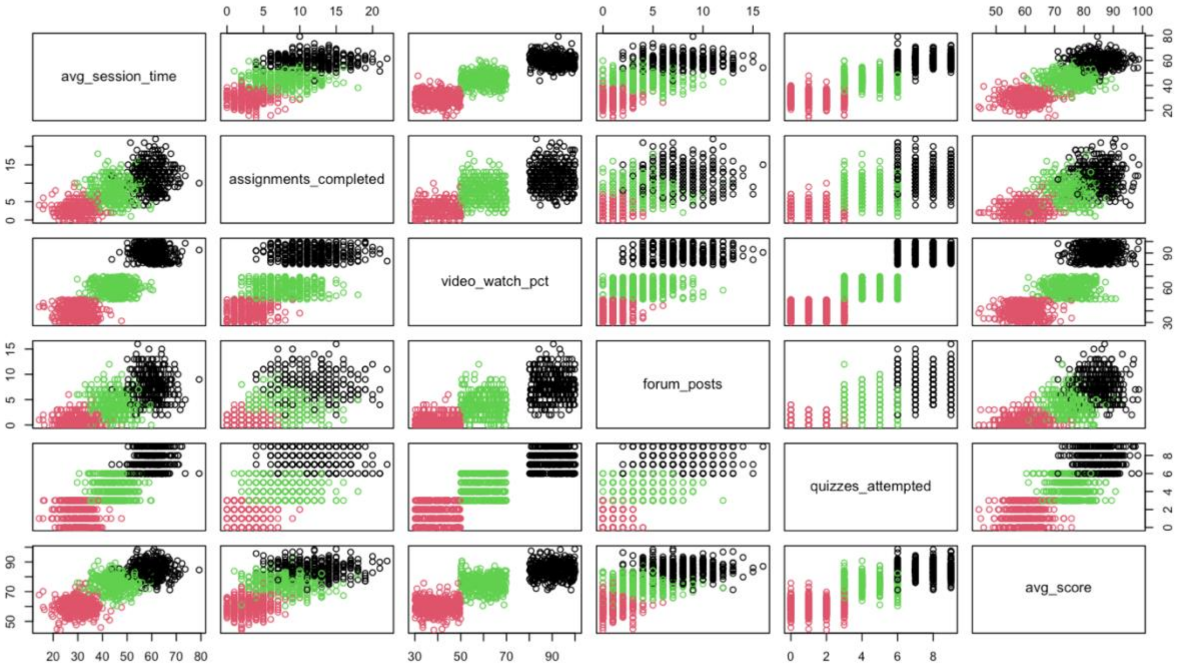
Cluster 2 (Green):

- This cluster shows low student engagement.
- This cluster had the lowest values in almost all the categories. It had low assignment completed and forum post. It also had the lowest student video consumption and average quiz score at ~60%.
- This cluster represents high risk students who may need some type of intervention.

Cluster 3(Blue):

- The shows to be in between clusters 1 and 2 for all attributes.
- Represents the students who perform average but could probably use some help to improve their engagement and scores.

Figure 5. Pairs plot of k = 3 clustering results.



This pair plot shows the relationship between the three clusters from the k-means model.

Cluster 1 (Red):

- Appear in lower left of all the plots, so represents low engagement.
- Student showed lowest session times, assignments completed, videos watched, forum posts, quizzes completed, and average quiz scores.

Cluster 2 (Green):

- In the middle of all the plots, so represent moder engagement.
- Students had moderate values across all attributes, including session times, assignments completed, videos watched, forum posts, quizzes completed, and average quiz scores.

Custer 3 (Black):

- In the upper right of across all the plots, so represents the high engagement students.
- Students had high scores across all attributes, including session times, assignments completed, videos watched, forum posts, quizzes completed, and average quiz scores.

Conclusion and Discussion:

During this project k-means was applied to the dataset that shows online student engagement and performance patterns. Using attributes like session time, videos watched, forum posts, quizzes completed, average quiz score, and assignment completion, we were able to analyze different student learning groups.

Key Findings:

After creating the silhouette plot and using the elbow method it could be seen that k = 3 was a better choice over k = 2. They both had a silhouette average of 0.51 but k = 3 allowed for more diversity in cluster by providing the option of creating 3 different categories as opposed to 2.

Using K = 3 during our k-means analysis revealed 3 student engagement levels. Those level being low, moderate, and high engagement. We also validated our chosen k value using silhouette, elbow plots, pair plot, and box plots.

- Cluster 1 represented high engagement and showed student with high session times, average quiz score, quizzes attempted, assignments completed, videos watched percentage, and forum posts
- Cluster 2 had moderate engagement and showed students with mid level session times, average quiz score, quizzes attempted, assignments completed, videos watched percentage, and forum posts.
- Cluster 3 showed students with low engagement with low session times, average quiz score, quizzes attempted, assignments completed, videos watched.

Limitations:

- The main limitation for this analysis would be using a synthetic dataset. There really isn't any way to find out if the data provided is an accurate representation of student engagement because it has no real world source.

Future Analysis:

- Future analysis for this analysis could include student demographics that would give deeper look at a possible correlation between whether certain demographical data might influence student online learning engagement.
- Other future analysis could include trying a different clustering method such a DBSCAN to see if there are any significant difference from our k-means results.

References

Bergdahl, N., Bond, M., Sjöberg, J. *et al.* Unpacking student engagement in higher education learning analytics: a systematic review. *Int J Educ Technol High Educ* **21**, 63 (2024). https://doi.org/10.1186/s41239-024-00493-y

Piech, C. (n.d.). *K-means clustering*. Stanford University. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Akpen, C.N., Asaolu, S., Atobatele, S. *et al.* Impact of online learning on student's performance and engagement: a systematic review. *Discov Educ* **3**, 205 (2024). https://doi.org/10.1007/s44217-024-00253-0

Siemens, G. (2022). *Foundations of learning analytics*. In C. Lang, G. Siemens, A. F. Wise, & D. Gašević (Eds.), *Handbook of Learning Analytics* (2nd ed., pp. 9–19). Elsevier. https://doi.org/10.1016/B978-0-12-821929-4.00002-0

OpenAI. (2025). *Dataset of synthetic student engagement data generated via ChatGPT* [Large language model]. https://chat.openai.com

Zala, R. (2023, January 24). *The Elbow Method: Finding the optimal number of clusters*. Medium. https://medium.com/@zalarushirajsinh07/the-elbow-method-finding-the-optimal-number-of-clusters-d297f5aeb189