Part 1

Emergency! There is a zombie apocalypse and no one can go outside! However, school must go on!

You are a data analyst working for a large school district. The district superintendent wants to roll out a new online course for all students called “Cool Schools Online 2.0 Super Funtime”. The superintendent is worried that some students will not engage with the materials in the course though. She has asked you and your team to advise her on what she should expect in terms of student engagement. She will be calling you in 1 hour and 20 min so you don’t have a lot of time!

Don’t panic though luckily, approximately 1000 students already took the course last year so you have some data to investigate before you have to advise her. [The data is located here](https://github.com/core-methods-in-edm/hackathon3) but is in two files that need to be combined and needs some tidying up. You will also find a codebook for the data in the same repo. Familiarize yourself with the codebook, download the data, combine it and tidy it up and then structure it so each student is represented by a single row. Report back when you think you have a workable, tidy data set.

Part 2

Your key goal is to categorize students with respect to how they engage with the “Cool Schools Online 2.0 Super Funtime” online course. Your first step is to see if you can identify any patterns in your data visually. Create some visualizations that show how students are engaging with the course materials and roughly categorize students according to the patterns you observe. How many groups of students do you think are in the data?

**Report back when you have identified your groups and have a rough idea of what they represent.**

# Part 3

The superintendent is not only interested in the types of engagement but being able to identify and classify these types automatically. You will need to develop an algorithm to automate the classification of students according to how they engage with the course.

Luckily you have heard of a clustering algorithm called ***k-means*** that you think will do the job but you can’t remember the details. Familiarize yourself with this algorithm using the following sources:

* [Josh Starmer K-means video](https://youtu.be/4b5d3muPQmA)
* [Lindholm, Wahlström, Lindsten, and Schön, The Supervised Machine Learning Book (2020), pp.198-202](https://drive.google.com/file/d/1WUL5tu1e4IMNsMAeQQAhPA4dnivSed9A/view?usp=sharing)
* [Wikipedia K-means explanation](https://en.wikipedia.org/wiki/K-means_clustering)

## Questions to think about

* How does k-means work?
* Is your data appropriate for this method?
* Do you need to process your data before you apply the k-means algorithm to it?
* Can you identify issues that can arise when using K-means?

Once you have answered these questions look up the help file for kmeans command in RStudio (?kmeans)

Apply k-means to your data to try to identify the groups of students you had observed in your visualization.

To run the k-means algorithm:

fit <- kmeans(DF, number of clusters you think exist in the data)

To append clusters to your dataframe

DF <- data.frame(DF, fit$clusters)

**Report back when you have assigned you k-means clusters to your data**

# Part 4

Create a single visualization that presents your clusters and their relationship to student engagement with the course, explain what each represents for the superintendent.