Data management and exploratory data management

* CRISP DM
* Business understanding – we are looking at **learning analytics**
* Ie data collected about students
* Use this data to gain insights about students
* Develop a set of tools to extract interesting insights in a quick, reliable and repeatable manner
* Use GIT for version control
* All source code/programs should be well documented
* Use software testing
* **Reproducibility** – use ProjectTemplate for R
* Create written documentation using RMarkdown

Ideas of exploratory analyses

* Could look at question responses – see how many each got correct
* Ie see what the mean number of correct answer is – calculate STD – if they are 2 STD below mean ‘flag’ the student
* Also ‘flag’ them if they haven’t answered the question
* Look at the video stats – check if they have watched the videos or not

CRISPDM – 1

* Started by browsing all of the data
* I knew that to some extent, to know what kind of business questions I could ask, I’d have to know what data I had available
* Data sets I ruled out for using
* Enrolments – too many unknown variables – not useful
* Team members – small datafiles with no particularly useful information

Video data

* Look at the length of the video and see what proportion of people have watched the whole video
* Why did I pick this data set initially?
* This is also data that is probably very accurate – it isn’t relying on more subjective quantities such as survey response
* However, it doesn’t give a whole picture
* People might just switch it on but not actually watch it
* Given that videos probably take a lot of effort to make
* Look at the longer videos, if people are less likely to watch longer videos this might make a big difference
* Ie if longer videos mean that people watch less then could look at the longest video and see what effect that has
* There is only a slight downward trend
* Look at the percentage watched 100

CRISP\_DM -2

* Ideas – maybe leaving survey responses – useful to try and see why people decided to leave the program
* Or weekly sentiment responses – can see what the people are bothering to do
* Although these, again, can not be fully trusted to be representative of the whole group
* Surveys are normally optional, so only the people who have strong opinions (negative or positive are likely to answer the survey)

For CRISP\_DM 2 I will look at the question response data

* Can see the mean number of correct questions
* Can then count the number of questions that each student got correct