**Section 1: Motivation -** The overarching goal of this thesis is to gain a better understanding of how K-12 students progress throughout their courses on the online coding platform, Grok Academy. Specifically, this thesis aims to look at how certain drivers and student behaviours affect course progress, and identify points within the course where teachers might be able to intervene.

**Section 2: Data:**

* Provide general insights about the students
  + Distributions: challenges, grades, final scores, gender

**Initial data exploration**

* **Interaction sequences and how to define them**
  + **Initial attempt: recording every single slide behaviour chronologically** 
    - Initially, we had the idea of recording every single behaviour chronologically for each student’s attempt on a module e.g. all slide views, slide attempts and slide submissions.
    - We quickly realised that not only was this not scalable, but that a lot of the data was noisy as the way by which students interact with slides. For instance, students might view a few slides, skip to the end of the module, and then view the same slides again.
  + **Second attempt and refinement: recording no submission, slide attempt and slide completion** 
    - Then, we reasoned that what really mattered wasn’t necessarily the order in which students completed certain slides or how many times that they went back to a slide, but rather whether they had made a significant attempt at a slide and successfully passed it.
    - We justified losing that level of granularity afforded by the initial attempt for more flexible analysis and less variability in the data.
* **Which interaction sequences were most commonly occurring** 
  + Most commonly occurring interaction sequences involved:
    - Completing most if not all content slides and problem slides
    - Completing all problem slides but no content slides
    - For certain modules, there was also quite a lot of variation in student interaction sequences.
* **Which slides were most frequently completed and which ones were skipped** 
  + Content slides typically skipped; certain content slides completed (worth matching this up with the module outcome prediction results)
* **Types of students**
  + Plotted the number of slides students completed in the course against their final score and found three types of students
    - Students who complete most if not all slides and score highly
    - Students who barely complete any slides and score highly
    - Students where slide completion is proportional to their score

**Outcome variables:** To answer our key question of how certain drivers and student behaviours affect course questions, we chose to focus on and predict two key outcome measurements 1) student performance at the end of each module and course and 2) student dropout.

**Predictor variables:** Our hypothesis is that the way by which students interact with certain slides (contained within course modules) has an impact on their overall performance (both at the end of each module and course) as well as tendency to dropout. Specifically, we posit that there is a subset of key slides within each module where completion is highly recommended to avoid failure.

**Data processing steps to retrieve the outcome variables**

* **Creation of interaction sequences**
  + Interaction sequences for each module and course (determined for each student and justification
  + Label encodings and justification

**Experiments**

1. **Predicting student performance on the last slide of the final module for each course**
   1. **Predictor**
      1. Entire interaction sequence, including previous content and problem slides.
      2. Interaction sequence including only previous content slides
   2. **Classifiers** 
      1. Logistic regression (and justification)
   3. **Results** 
      1. Low predictive accuracy
      2. We were able to determine which slides were predictors of good performance on the last slide of the final module. These slides tended to fall within the latter modules.
      3. **Reflection**
         1. Only a subset of slides is useful in the prediction. Other slides might be correlated with each other.
         2. Difficult to predict this as there was a heavy skew in the data towards not completing the final slide of the final module, as most students would have dropped out, particularly in earlier challenges e.g. newbies.
2. **Predicting student performance on the last slide of the final module for each module in challenge-beginners-2018**

We reasoned that a more effective and insightful thing to look at might be individual modules, and how the slides affect end of module performance, because the slides of a module are more likely to be directly related and help the student answer the final problem of the module.

* 1. **Predictor**
     1. Interaction sequence for the module, including previous content and problem slides.
  2. **Classifiers** 
     1. Logistic regression
     2. Naïve Bayes
     3. Decision tree
  3. **Results** 
     1. On average, >80% accuracy across all three classifiers, with Naïve Bayes performing less than average in cases where there are highly correlated attributes (i.e. highly correlated slides).
     2. Decision tree has a slight edge over the other algorithms.
  4. **Reflection**
     1. In this experiment, we looked at three classifier models – logistic regression, naïve Bayes and decision trees (explain why).
     2. As naïve Bayes can suffer from violation of the independence assumption when there are correlated attributes in the feature set, we chose to discontinue using it in further experiments.
     3. Comparing decision trees and logistic regression, the former has more sophisticated algorithms at 1) eliminating correlated features and 2) identifying key attributes and ranking them, and so we chose to continue with decision trees, which are also more interpretable than logistic regression and can help identify pathways to success/failure.

1. **Predicting student performance on the last slide of the final module for each module in all challenges**
   1. **Predictor**
      1. Interaction sequence for the module, including previous content and problem slides.
   2. **Classifiers** 
      1. J48
      2. J48 with CFS selected attributes
      3. J48 with gain-ratio selected attributes
   3. **Metrics**
      1. Overall accuracy
      2. Accuracy (passed)
      3. Accuracy (failed)
      4. Accuracy (no submission)
   4. **Results** 
      1. **Accuracy**
         1. High overall accuracy (>80% in on average in all cases, and sometimes reaching >90%)
         2. High accuracy (passed) – in most cases > 90% on average
         3. Low accuracy (failed) – close to 0
         4. Moderately high accuracy (no submission) (>70% on average in all cases)
      2. **Important slides identified for each module** 
         1. Most important slides at predicting performance on the final problem slide were preceding problem slides
         2. There were other content slides directly before/after those problem slides that were also determined as important, potentially through providing necessary content and instructions for students.
      3. **Reflection** 
         1. Overall accuracy can be biased towards passed cases, as there is a high level of passing rates (anywhere from 50-80% in most cases)
         2. Important slides finding is important but not surprising/sophisticated.
2. **Predicting dropout (still need to flesh out)**
   1. Previously we looked at predicting dropout and using three different categories - early late and no dropout with considerable accuracy.
   2. One key finding was that we could use the first weeks or so of data and predict later dropout with considerable accuracy - meaning that first weeks behaviour was particularly important.