1. ***Data splitting:****Determine which subsets of the data to use for analysis, assessment, and final training of the model*
   * *Because your models will be trained on data from one time period in order to predict observations in another, splitting the data will not be as straightforward as in the Tutorials*
   * *A companion document, BM05BAM–2023 Group Project Part 2, provides details on dividing the data into analysis, assessment, final training, and prediction sets*

*Decisions:*

* *Split*
  + *Analysis set 🡪 2020-2021 (all)*
  + *Assessment set 🡪 2022 (full)*
    - *Subset 2022 after 15th 🡪 for estimating test error & making final decision about the choice of model*
  + *Final training set 🡪 2020-2022 (all)*

*Assumptions:*

* Students applying before and after March 15 are significantly different
* 2023: If offerdate is within two weeks from March 15 🡪 then: response = “accepted” else “rejected”
* We will overpredict to provide students more seats in the rooms than needed
* Remove outliers: response == NA ; status == enrolled

1. **Data exploration:** Explore the data to understand it better, and to identify any feature engineering that will improve your models
2. **Assessment metrics:** Choose an appropriate set of metrics for tuning, assessing, and selecting among candidate models, based on your understanding of the problem
   * What counts as a good prediction in this context is unclear, and you will need to bring clarity to this issue
   * Your team must decide whether over-predicting (making poor use of larger rooms) or under-predicting (not providing enough space for all students) is worse—or if they are equally bad
   * You will need to choose a primary (i.e. most important) evaluation metric reflecting this decision, and construct and assess your models accordingly
3. **Multiple methods:** Use a variety of supervised learning methods from the course. At a minimum, you must consider at least one model from each of these two groups:
   * **Group A**: k-nearest neighbors, regularized logistic regression
   * **Group B:** Random forest, boosting
4. **Model selection and assessment:** Select a final model and assess its performance against an appropriate test set
5. **Make predictions:** Retrain the final model, generate predictions for AY2023, and submit these predictions (along with your chosen metric for assessing prediction quality) as an interim deliverable
   * Initially, to prevent data leakage, the prediction set will not contain the target variable for AY2023
   * Before the final deadline, you will submit your final predictions and your primary evaluation metric
   * Afterward, you will gain access to the full data, so you can assess the final model’s true performance against the AY2023 data and finalize your presentation
   * In both your interim and final predictions, you must report, for each program, the average predicted probability of enrollment (soft predictions), and the predicted number of attending students (hard predictions), for the subset of AY2023 offers sent out prior to March 15
   * To produce hard predictions, you must select a suitable decision threshold
6. **Evaluate predictions:** Compare the final model’s performance against the full AY2023 data based on the primary metric (among others), and perform an error analysis
7. **Presentation:** Present your results in a recorded presentation of no more than 10 minutes
8. **Reproducible code:** Document all of the aforementioned steps in a well-commented and reproducible (by me) R script or R markdown file; all results in your presentation need to be reproducible from this code
9. **Work as a team:** You should strive to work together well and allocate the workload as fairly and efficiently as you can