For final project, you will be working in a group with up to 5 people.

- You will find a large dataset or multiple datasets (**min. 200MB in total**) to work with and pose 1-2 research questions.
- For the research questions, it is supposed to be **decision-centric** and try to think as if it is a consulting project.

  Say, for the marketing examples covered in class, we try to understand how can we improve sales/click rates/profits, rather than finding the best predictive model from a set of candidates.
- You will use the techniques learned in the course (but feel free to extend beyond if you can explain to me what you do during presentation), and present empirical answers to your question.
- Only put in results that are **relevant**. Avoid overloading your report or presentation with unnecessary details (e.g. marginal distribution of every variable).
- While smaller groups are not expected to provide the same depth of analysis as larger groups, all analysis should have the components listed in the grading criteria.

Here is a rough breakdown of our grading criteria.

1. Statement of the question (15 pts)
   - Clearly identify the research question
   - Feasibility of addressing the question with available data
   - A sketch of the rest of the report (findings and conclusion)
2. Data Exploration (15 pts)
   - Describe what dataset(s) you use, quality of the dataset, why it is “Big” and how you plan to deal with it
   - Data preprocessing if necessary

- Exploratory data analysis (visualization, pattern, summary statistics…), only put in results that are relevant

- Any insights you want to carry over to analysis later

3. Model Building (25 pts)

- Appropriateness of the model(s) selected for the problem

- Implementation of models

- Evaluation of model performance

4. Conclusion (20 pts)

- Clear summary of your findings and model outcomes

- Interpretation of results to answer your research questions

- Reflection of the strengths and limitations of your model(s), what could be an improvement

- Discussion of future work

5. Efficient Computing (10 pts)

- Describe the computation bottleneck in your analysis (storage, memory, time)

- What you have implemented to relieve the computation burden

6. Clarity and novelty (15 pts)

- Project is well-structured. Writing is clear and in a logical manner

- Good visualizations and explanations of findings (to effectively communicate your results)

- Novelty

Here are some ideas on where to get data for your final projects. You are welcome to use data from any other source as well. You can always ask ChatGPT to point you to some big public datasets.

- City of Chicago data https://data.cityofchicago.org/
- Airlines data set http://www.stat.purdue.edu/~sguha/rhipe/doc/html/airline.html
- Kaggle Competitions http://www.kaggle.com/
- UCI Machine Learning Library https://archive.ics.uci.edu/ml/index.php
- Stanford Large Network Dataset Collection http://snap.stanford.edu/data/index.html
- Million Songs Database http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset
- PGA Tour (contributed by Tyler Burkett) http://www.pgatour.com/stats.html
- Big list of publicly available data sets http://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/
- Quora answer on “Where can I find large datasets open to the public?”https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public