**Data Science Project (Dawie)**

Below are some details for the project that you have to undertake for the Data Science module. This project counts toward 50% of your mark for the module, so please take the appropriate amount of time to think through the project.

**Format**

For the project you are required to write a report that implements some machine learning techniques and SQL tools. For the machine learning component, you can utilize what we discussed in the lecture sessions, or you can bring in new techniques. However, if you want to use a technique that we have not covered, make sure to contact me so that I can advise on whether the technique is too advanced for the module.

You are welcome to use any of the techniques on parallel programming and cloud computing to perform / speed up your computations. Include these elements in your code and discuss them in your project for extra credit.

In terms of structure, you can write a brief introduction, discuss the manipulations that you performed on the data, talk about your research question and the methodology that you are utilizing to answer the question and then finally discuss the results. This is like the format of a thesis, but you do not have to do an extended literature review. The goal is to focus on execution of techniques.

With respect to data, there are two approaches that you can follow for the project.

**Approach #1**

Data has been provided via email. The dataset must be transformed into the format of a relational database (by you).

**Approach #2**

Find your own data and put it into the format of a relational database. With this option you are free to pursue any topic you want. Try to not be too ambitious for this project. I am simply looking to see whether you can execute techniques and not too concerned about making contributions to the field. If you want to follow this approach, please contact me to make sure that you are attempting something that is realistic.

For your project you need to attach your data and code so that I can run it on my computer to see if it works. I will not judge your code on the style, but I would like to see that the results are easy to replicate. If you are using the data that I have provided you do not need to attach the data.

**Due Date**

The due date for the project is the 30th of June 2021 at midnight. No late submissions will be accepted. I will create a link on SUNLearn where you can submit the document. Please do not submit via email.

Given the due date, you have more than a month to find data, clean the data, transform data into database format, identify a question, identify the techniques that are most appropriate to your question and write up the results.

**Appointments**

Due to the size of the class I won’t be able to have a discussion with everyone regarding their projects. Please email me if you have particular questions so that I can determine whether your questions require a quick reply or a more lengthy meeting. You can also come and talk to me after lectures about your ideas.