Introduction

Max Turgeon

SCI 2000–Introduction to Data Science

Lecture schedule

We meet twice a week on WebEx:

- **Tuesday** 11:30am to 12:45pm
- **Thursday** 11:30am to 12:45pm

Both weekly meetings will be **recorded** and available on UM Learn.

Assessments

- 4 assignments (10% each)
- 1 final project (50% each)
- Summaries (5%)
- Class participation (5%)

Assignments

- Assignments will provide an opportunity to analyse data and practice the skills from the lectures.
    - They will require using a programming language: R or Python
    - The lectures will use **R**, so if you want to use Python, you're on your own!

By the way, should I already know R?

- No, we'll learn together as needed.
  - Concepts will be introduced as needed, and through examples.
  - See UM Learn for extra material on R.
- **Important**: Let me know if some of the code isn't clear!

Final project

- In teams of 2-3, you will have to find a dataset, analyse it, and summarise your findings.
- I will provide more details later, but you can start teaming up.

Summaries

- Three times during the semester, we will analyse a dataset together.
- After the lecture, you will have to write a short summary about
    - what we learned about the data
    - what you learned about data science

Class participation

- Participation can be during lectures or in discussion groups.
  - Monthly participation gives you 4/5
  - Weekly participation gives you 5/5

What is data science? i

- **Disclaimer**: Data science is not a well-defined field, so this is only my personal definition.
- Traditional statistics has focused on using statistical tests or models to understand data.
- Data science is concerned with the whole process of data analysis:
    - Store and retrieve data.
    - Visualizing data effectively.
    - Communicating results.

What is data science? ii

- It also blurs the line between traditional statistical models (e.g. linear regression) and machine learning (e.g. neural networks).
    - If a model does the job, who cares where it comes from!
- For these reasons, a good data scientist has good knowledge of both statistics and computer science.

What will we learn? i

- Here are the three main course objectives:
- Become proficient in R, to the level that you can analyse data using the tools from this class.
- Be able to describe and analyze data through visualization and simple statistical procedures.
- Be introduced to statistical thinking and be able to think critically about variation and biases.
- We will cover the following tools and techniques:
    - Data visualization
    - Data wrangling
    - Relational data

What will we learn? ii

- Web scraping
- Introduction to regular expressions
- Linear regression
- (If time permits) Automation and version control

What will we not learn?

- Machine learning
- Cloud computing
- Big data technologies (e.g. Spark and Hadoop)
- Advanced statistical techniques.

But we have courses on all of them if you're interested!