MSIN0010 Data Analytics I
2021-2022

Lecturer:	Dr Adam Smith
Email:	a.smith@ucl.ac.uk
Office Hours:	Tuesday 3 - 4pm (Zoom) or by appointment (https://calendly.com/adamnsmith)
TA:	Mr Xitong (Tony) Liang
Email:	xitong.liang.18@ucl.ac.uk

Module Description

Data Analytics I introduces students to how organizations use data and analytics to create value and improve performance, trains them to use selected statistical data analytics and data mining tools, and introduces them to elements of the statistical theory and algorithms that underpin those tools.

The context for the module is management in complex, innovation-intensive, data-driven environments. The explosion in the volume and range of internal and external data available to managers and the development of new data analytics tools is having a major impact on how people identify, formulate, and solve management problems.

During the module, students will manipulate example data sets and use basic data collection tools and APIs to source data from publicly available data sources.

Learning Outcomes

Upon successful completion of the module, students will be able to:

- Understand how organizations use data and analytics to create value and improve performance.
- Understand and apply founding probability and statistical theory to data analysis.
- Understand and apply information theory and data mining theory to data classification and data clustering problems.
- Characterize and critically assess the quality of data sets and their limitations in the context of data-driven decision-making.
- Use selected tools (Excel and R) to analyze and visualize data.
- Understand key elements of the theory, technology, and algorithms that underpin the tools used.

Meeting Times

- Online Lecture: Tuesday 1:30 - 3pm
- In-Person Seminar: Friday 11am - 12:30pm (Group A), 1:30 - 3pm (Group B), 3:30 - 5pm (Group C)

Assessment

1. (Term 1) Coursework – 20%
2. (Term 2) Scenario week group coursework– 20%
3. (Term 3) Exam – 60%

Required Materials

1. Online Textbook: https://www.adamnsmith.com/MSIN0010/
2. Data Camp: https://www.datacamp.com/
3. RStudio: https://rstudio.cloud
4. All lecture notes and additional readings will be posted on Moodle: https://moodle.ucl.ac.uk

Additional Resources

Textbooks on Probability and Mathematical Statistics

- *All of Statistics* (2004) by L. Wasserman
- *Introduction to Probability* (2019) by J. Blitzstein and J. Hwang [http://probabilitybook.net/]

Textbooks on Regression and Machine Learning

- *An Introduction to Statistical Learning* (2013) by G. James, D. Witten, T. Hastie, and R. Tibshirani [https://www.statlearning.com]
- *Business Data Science* (2019) by M. Taddy

Schedule

UNIT I: DATA

1. **Introduction to Data Analytics** (Oct 4-8)
   - Topics: brief history, data sets, data visualization, summary statistics
   - Reading: *Big Data: The Management Revolution, How R Helps Airbnb Make the Most of its Data*
   - DataCamp: Introduction to Spreadsheets (ch. 1 getting started), Introduction to R (ch. 1 intro to basics), Introduction to Data Visualization with ggplot2

UNIT II: PROBABILITY

2. **Probability** (Oct 11-15)
   - Topics: probability theory, Bayes' theorem, random variables, probability distributions
   - DataCamp: Foundations of Probability in R (ch. 2 laws of probability)
3. **Probability** (Oct 18-22)
   - Topics: expectations, law of large numbers, central limit theorem

UNIT III: STATISTICAL INFERENCE

4. **Estimation** (Oct 25-29)
   - Topics: point estimation, confidence intervals
5. NO CLASS (Nov 1-5)
6. READING WEEK (Nov 8-12)
7. **Testing** (Nov 15-19)
   - Topics: hypothesis testing

UNIT IV: STATISTICAL MODELS AND MACHINE LEARNING

8. **Regression** (Nov 22-26)
   - Topics: linear regression, regression trees, model selection
   - DataCamp: Supervised Learning in R - Regression (ch. 1 what is regression? + ch. 2 training and evaluating regression models + ch. 5 tree-based methods)
9. **Classification** (Nov 29 - Dec 3)
   - Topics: KNN algorithm, logistic regression, classification trees, model selection
   - DataCamp: Supervised Learning in R - Classification (ch. 1 KNN + ch. 3 logistic regression + ch. 4 classification trees)
10. **Clustering** (Dec 6-10)
    - Topics: K-means, hierarchical clustering
    - DataCamp: Unsupervised Learning in R (ch. 1 unsupervised learning in R)
11. NO CLASS (Dec 13-17)