MSIN0010 Data Analytics I
Term 1 2019

Lecturer:	Dr Adam Smith
Email:	a.smith@ucl.ac.uk
Office Hours:	Tuesdays 1-2pm (Engineering Front Building, Rm 4.01) and by appointment
TA:	Dr Viviana Culmone
Email:	v.culmone@ucl.ac.uk
Office Hours:	Wednesdays 4-5pm (Birkbeck Malet Street B30)

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.

Assessment

1. (Term 1) Coursework: 4 problem sets – 20%
2. (Term 2) Group coursework during scenario week – 20%
3. (Term 3) Exam – 60%

ATTENDANCE AND PARTICIPATION

You are expected to be present (both physically and mentally!) to weekly lectures and seminars. Being active on your laptop or mobile phone during lecture not only detracts from your own learning, but also inhibits the learning of others and will not be tolerated. Moreover, I encourage you to participate in class! Stop me by raising your hand whenever you have a question or want to contribute to the class discussion. However, asking your classmates and carrying on side conversations is distracting to your peers and will also not be tolerated.

TEXTBOOKS

There are no required textbooks, but the following may be useful as a reference.

Probability and Statistical Inference

- Wasserman (2004). *All of Statistics.* Springer, First Edition.
- Blitzstein and Hwang (2019). *Introduction to Probability*. Taylor & Francis Group, Second Edition. [http://probabilitybook.net/]
- Aczel and Sounderpandian (2008). *Complete Business Statistics.* McGraw-Hill/Irwin, Seventh Edition.

Statistical Models and Machine Learning

- Taddy (2019). Business Data Science. McGraw-Hill, First Edition.
- James, Witten, Hastie, and Tibshirani (2013). *An Introduction to Statistical Learning.* Springer, First Edition. [http://faculty.marshall.usc.edu/gareth-james/ISL/]
- Provost and Fawcett (2013). *Data Science for Business.* O'Reilly, First Edition.

COMPUTING

We will analyze data using both Microsoft Excel and R. R is a free, open-source statistical software package that is used by most firms on the frontier of data analytics. Unlike other point-and-click platforms (e.g., Excel, SPSS, Minitab), R has a command line interface that provides complete flexibility when carrying out statistical analyses. However, this comes with a bit of a learning curve as the user must enter code to execute core functions. I will provide in-class instruction and code when necessary. No prior knowledge of R is necessary/expected.

You have two options for running R.

1. You can download R on your personal machine.

   https://cran.ma.imperial.ac.uk

   After installing R, you also need to download RStudio (http://www.rstudio.com) which is an integrated development environment (IDE) used for running R. Among other things, it provides a code editor and visualization tools all in one environment, making for a better user experience.

2. You can run R using Jupyter Notebooks using the Microsoft Azure cloud computing platform.

   https://notebooks.azure.com/

   Sign in with your UCL email (in the form of ucaaxxx@ucl.ac.uk) and password.

SCHEDULE

Week	Unit	Topics	DataCamp Assigment
1. Oct 1	Data	purpose of analytics, data visualization, summary statistics	Spreadsheet Basics, Introduction to R, Data Visualization in R
2. Oct 8	Probability	probability rules, random variables, probability distributions	Foundations of Probability in R
3. Oct 15		expectations, inequalities, asymptotics	
4. Oct 22	Statistical Inference	point estimation, confidence intervals	Foundations of Inference
5. Oct 29		NO CLASS	
6. Nov 5		NO CLASS - READING WEEK	
7. Nov 12		hypothesis testing	
8. Nov 19	Statistical Models and Machine Learning	linear regression, regression trees, model selection I	Machine Learning Toolbox
9. Nov 26		k-NN, logistic regression, classification trees, model selection II	
10. Dec 3		k-means clustering, hierarchical clustering	Cluster Analysis in R
11. Dec 10		NO CLASS	
February 3-7 2020		Scenario Week	

UNIVERSITY POLICIES

Student Disability Services: http://ucl.ac.uk/disability/

Plagiarism: https://ucl.ac.uk/students/exams-and-assessments/plagiarism

Academic Calendar: https://www.ucl.ac.uk/estates/sites/estates/files/cal_2019_2020_-_amended.pdf