Unit Guide: Optus U Data Science

MQ Unit code: MCMP6200

Base MQ Unit: COMP6200 Data Science

Course developer: Steve Cassidy – Associate Professor, Department of Computing

ProLearn link: **https://prolearn.mq.edu.au/course/view.php?id=181**

Learning outcomes

A *Develop superior skills using spreadsheets to summarise, model and visualise data to make decisions.*

B *Understand how data can be interpreted in different ways in line with an agenda and ask appropriate questions.*

Overview

This micro-credential introduces Data Science methods using Python and Jupyter Notebooks. The focus is on the typical data science workflow involving reading data in different formats, performing some analysis, and communicating the results. Notebooks are used to document this workflow in a reproducible manner and develop a story around the analysis you perform. You will complete a number of small data analysis projects to build into a portfolio of work at the end of your course of work.

Course duration

6 weeks (12.5hrs learning/week): 2hrs self-directed (pre-recorded lectures and written content); 2hrs real-time class (presentation, discussion, & exercises); 2hrs own-time reading and practice; and 6.5hrs own-time assessment exercise.

Texts

Selected readings from:

- *IDS*: *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*; Laura Igual, Santi Seguí.
- *Other readings will be linked from the ProLearn pages each week.*

Software

The course is based around the use of Jupyter Notebooks (Python) and the Git version management system. To complete the course you will need to access to installed versions of:

- Python – Latest version from https://www.python.org/ or at least version 3.5 or greater
- The following Python packages: notebook, pandas, numpy, matplotlib, seaborn, sklearn, scipy
- A Git client, e.g. https://desktop.github.com/ or https://git-scm.com/downloads

Instructions on installing these packages will be provided.

Getting started with Analytics

Curriculum

Topic 1: Introducing Python and Jupyter Notebooks

Goals and activities:

- Defining Data Science
- Review the basics of Jupyter Notebooks and Python
- Introduce Version Control with Git

Reading:

- Textbook - Chapter 1 (1.1) and 2 (2.1-2.5) of the textbook

Topic 2: Data Formats and Data Structures

Goals and activities:

- Explore the main data structures in Python, NumPy and Pandas
- To understand how to read common data formats into Python
- Discuss the use of different data

Reading:

- Textbook - from Chapter 2, section 2.6
- Article - How to read most commonly used file formats in Data Science by Ankit Gupta

Topic 3: Descriptive Statistics

Goals and activities:

- begin exploring data with descriptive statistics
- consider some privacy concerns around data sharing
- introduce a standard methodology for DS projects

Reading:

- Textbook - Chapter 3 of the text.
- Additional reading: The Ultimate Guide to Data Cleaning by Omar Elgabry
- Article - The Data Science Process by Chanin Nantasenamat
- Article - How do Data Professionals Spend Their Time by Bob Hayes

Topic 4: Causality and Correlation; Visualisation of data

Goals and activities:

- looking at correlation between variables
- discuss the difference between correlation and causality
- look at some techniques for visualisation of data
- discuss what it means for a graph to be 'good'

Reading:

- Article - Good Graphics? (PDF) Chapter from the Handbook of Data Visualisation
- Article - Causality and Experiments from Computational and Inferential Thinking

Getting started with Analytics

Topic 5: Predictive Models; Reproducibility

Goals and activities:

- Introduce a simple predictive model: linear regression
- Begin looking at machine learning workflow
- Classifying categorical variables with logistic regression
- Discuss what makes an analysis reproducible

Reading:

- Textbook - Chapter 6: Regression
- Textbook - Chapter 15: Prediction from Computational and Inferential Thinking
- Article - Source of the Birthweight data: University of Sheffield

Topic 6: Clustering

Goals and activities:

- Look at unsupervised learning algorithms
- How do we measure the similarity of two observations
- Two different clustering algorithms

Reading:

- Textbook - Chapter 7 Unsupervised Learning

Assessment information

Assessments	Overview	Due Date
Progress Task 1 (20%):	• A notebook completed each week demonstrating mastery of the techniques and methods introduced. • You will be given a notebook each week to add to your workshop repository. You should complete the work, commit, and push to your Github Repository by the end of Sunday each week. It will then be marked, and feedback provided before class on Tuesday.	Sunday of each class week
Progress Task 2 (20%):	• A critical analysis of a notebook analysis provided to the student to provide: ○ feedback as a Word/PDF document ○ An updated version of the Python notebook file (.ipynb) • Submitted to ProLearn	5 August
Data Science Portfolio (50%):	• A collection of three data analysis tasks collected into a portfolio reflecting the knowledge gained over the course of the unit. • Feedback will be provided on work in progress: ○ during week 5 (20-23 June) ○ during week 6 (27-30 June) ○ during the week 2-5 August • Submission via push to Github Portfolio repository	23 August
Learner Reflection (10%):	• Written reflections on the application of the knowledge gained in the unit to the learner's role in the organisation. • Submitted to ProLearn	30th August

Getting started with Analytics

Major Assessment [Hurdle]

There is a hurdle requirement for this course (Data Science portfolio). This means that you must complete this task and receive a mark above 65% to be able to pass the course. If you do not achieve 65% or greater, you will not pass this course.

Late Submission

Extensions will only be granted with an approved application for Special Consideration.

Information on the Special Consideration process is provided in the Late Submission – Special Consideration form in your course ProLearn site. A Special Consideration application must be made within five (5) working days of the assessment task due date. Lodging an application for Special Consideration does not guarantee that you will be granted an additional/alternative assessment.

Note that applications for Special Consideration should be emailed directly to the Lecturer, using the form available via the course ProLearn site.

If you submit your work late without an approved Special Consideration, there will be a deduction of 5% of the total available marks made from the total awarded mark for each 24-hour period or part thereof that the submission is late.

Results

Once approved, an email will be provided to students informing of the release of final results on ProLearn. Results published on ProLearn or other platforms or released directly by your Lecturer, are not confirmed until this notification is provided as they are subject to final approval by the University.

Pass or fail

The pass mark for the Macquarie University Optus U courses is **65% and above**.

In line with University's academic policy for these courses, if students do not achieve a pass mark, they will be provided with an opportunity to resubmit the major assessment task based on a deadline set by the University.

Academic Integrity Policy

The values of academic integrity provide an overarching declaration that informs the University's staff and students involved in learning, teaching, and research.

All students are required to uphold the academic integrity values explained in the Academic Integrity Module provided to students via ProLearn. Completion of the Academic Integrity Module is mandatory to complete the course.

Macquarie University students have a responsibility to be familiar with the Student Code of Conduct.