Department of Computer Science and Software Engineering
Concordia University

COMP 499 — Introduction to Data Analytics
Course Outline
Summer 2019 Semester 1

Web page: `http://users.encs.concordia.ca/~gregb/home/comp499-s2019.html`

Course Calendar Description

COMP 499 Introduction to Data Analytics (4 credits) Data wrangling: Extract-Transform-Load, data cleaning, data integration. Data analytics: description, prescription, modeling, simulation, optimization. Discipline specific datasets, tools, and analysis approaches. Python ecosystem: language; libraries `numpy`, `scipy`, `pandas`; Jupyter environment; `R` data analysis package. A project. *Prerequisite*: COMP 233 or ENGR 371; COMP 352.

Note: This semester will not include material on `R` data analysis package. It will include material on visualization.

Course website: Check frequently the website for announcements, course material, assignments, etc.

IMPORTANT NOTE: In the event of extraordinary circumstances beyond the University's control, the content and/or evaluation scheme in this course is subject to change.

Course Objectives

Big Data and Data Analytics has permeated into every industry, government, and business function. The future will need data-driven approaches for all fields of human endeavour. The challenges in handling massive datasets and performing the computations for analysis to transition from raw data to information to knowledge and to application are many and varied. Data analytics is at the core of interdisciplinary collaboration with the social sciences, humanities, health and life sciences, ecology and the environment, culture and heritage, and engineering.

The aim of this course is to introduce students to Python programming language and related tools for data analytics; and to expose them to a broad range of data analysis problems across a range of disciplines.

Recommended Books:

(1) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd Edition, by Wes McKinney, O'Reilly Media, 2017.

(2) *Data Crunching: Solve Everyday Problems using Java, Python and More*, by Greg Wilson, The Pragmatic Bookshelf, 2005. This book is out of print, but can be found in the Library.

(3) *Data Science from Scratch: First Principles with Python*, by Joel Grus, O'Reilly, 2015.

(4) *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, by Hadley Wickham, Garrett Grolemund, O'Reilly Media, 2016.

(5) Material on available data resources and data analysis problems from a range of disciplines will be provided.

Grading Scheme

The grade will be determined by the following components:

10% Attendance
15% Assignments
25% Course project
15% Midterm
35% Final Examination

There is no standard correspondence between the numerical marks and the final letter grades. Students must pass the examinations, combined, in order to pass the course. Students must pass the assignment and project components, combined, in order to pass the course.

Course Components

This course is a 4-credit course with 2.5 class hours of lecture and 2 hours of lab.

This is an accelerated six-and-a-half week semester, so you should expect to **average** a total of 20–24 hours per week on this course.

Attendance at all lectures and labs is compulsory.

Lectures

Attendance at the lectures is compulsory.

Lecture 1: Introduction: Course, Python, Data Analytics
Lecture 2: Numbers, Data, and Experimental Design
Lecture 3: Data Warehouses, OLAP, and Business Intelligence
Lecture 4: Data Formats and Schemas; Python `pandas`
Lecture 5: Data Wrangling; OpenRefine
Lecture 6: Descriptive Data Analysis: Python `numpy`, `scipy`, `matplotlib`, `seaborn`
Lecture 7: Correlation; Causality
Lecture 8: Midterm
Lecture 9: Exploratory Data Analysis
Lecture 10: Visualization; Dashboards; Story Telling
Lecture 11: Models: Regression, Classification, Prediction, Simulation
Lecture 12: Project presentations
Lecture 13: Project presentations

Laboratory

Attendance at the labs is compulsory.

The lab session for the course will address practical issues with the Python ecosystem, provide a series of running examples and assignments for students to do hands-on data analytics, and provide students with data resources from a range of disciplines.

The Python ecosystem to be used is the anaconda distribution of Jupyter with Python v3.x with its standard libraries including `numpy`, `scipy`, `matplotlib`, `seaborn` and `pandas`.

Lab 1: There is no lab 1
Lab 2: Jupyter installation; Python introduction (software-carpentry.org)
Lab 3: Python Object-Oriented Programming (Corey Schafer: youtube)
Lab 4: Python pandas example (datacarpentry.org)
  Python for Social Science: pandas (steps 8-12), matplotlib (step 13), SQLite (step 14)
Lab 5: Python Exploratory Data Analysis example (datacarpentry.org)
  Python for Ecologists: Focus on steps 4-8, ignoring challenges and exercises
Lab 6: Data wrangling example (biggorilla.org)
  Kaggle 5000 Movie Dataset (imdb)
Lab 7: continuation of Lab 6
Lab 8: Midterm Preparation
Lab 9: Data wrangling with OpenRefine (datacarpentry.org)
  Data Cleaning with OpenRefine for Ecologists
Lab 10: Python machine learning with `scikit-learn` (scikit-learn.org)
  An introduction to machine learning with scikit-learn
Lab 11 – 13: Your project

Assignments

Assignment 1 is to complete the exploratory data analysis task of **Lab 5** as a Jupyter notebook, show the work to the TA, and submit the notebook to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `programming_assignment 1`.

Assignment 2 is to complete the data wrangling task using Python in **Lab 6–7**, show the work to the TA, and to submit one (combined) Jupyter notebook to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `programming_assignment 2`.

Project

Each student is to investigate a selected large dataset; carry out data wrangling, and exploratory data analysis; and develop models and stories to communicate their findings. The student will present their work in class at the end of the course.

Each student will submit a preliminary report as a pdf file, based on their Jupyter notebook, to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `project 1`.
The report should identify the dataset, the data wrangling, and the descriptive data analysis. It is due in Lab 10.

Each student will submit a final report as a pdf file, based on their Jupyter notebook, to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `project 2`.
The report should identify the dataset, the data wrangling, and the descriptive data analysis, the exploratory data analysis, and any model building, validation, and story-telling visualizations. It is due after the presentation at the end of the course.

The choice of dataset, and the questions to be explored using the dataset, should be discussed with the course instructor as early as possible.

Examinations

The midterm exam will be a one-hour exam covering data wrangling, exploratory data analysis, and the use of Python and the libraries. It will cover both theory and practice. It may have a mixture of true-false questions, multiple choice questions, short answer questions, and programming tasks.

The final examination will be a three-hour examination covering all theoretical and practical parts of the course. It may have a mixture of true-false questions, multiple choice questions, short answer questions, and programming tasks.

Academic Honesty

Violation of the Academic Code of Conduct in any form will be severely dealt with. This includes copying (even with modifications) of program segments. You must demonstrate independent thought through your submitted work. The Academic Code of Conduct is available at: http://www.concordia.ca/students/academic-integrity/code.html

Graduate Attributes

(GA1) A Knowledge Base for Engineering: *Demonstrated competence in university level mathematics, natural sciences, engineering fundamentals, and specialized engineering knowledge appropriate to the program.*

Knowledge-base: Data wrangling: Extract-Transform-Load, data cleaning, data integration. Data analytics: description, prescription, modeling, simulation, optimization. Python ecosystem: language; libraries `numpy`, `scipy`, `pandas`, `matplotlib`, `seaborn`; Jupyter environment.

(GA2) Problem analysis: *An ability to use appropriate knowledge and skills to identify, analyze, and solve complex engineering problems in order to reach substantiated conclusions.*

To perform data wrangling, exploratory data analysis, model building and visualization on a relatively complex dataset through selection and application of appropriate tools.

(GA5) Use of Engineering tools *is the ability to create, select, apply, adapt, and extend appropriate techniques, resources, and modern engineering tools to a range of engineering activities, from simple to complex, with an understanding of the associated limitations.*

To perform data wrangling, exploratory data analysis, model building and visualization on a relatively complex dataset through selection and application of appropriate tools.

(GA6) Individual and Team Work: *An ability to work independently and as a member and leader in diverse teams and in multi-disciplinary settings.*

Work individually to analysis a relatively complex dataset.

(GA7) Communication Skills: *An ability to communicate complex engineering concepts within the profession and with society at large. Such abilities include reading, writing, speaking and listening, and the ability to comprehend and write effective reports and design documentation, and to give and effectively respond to clear instructions.*

To present a report on a data analytics project.