Department of Computer Science and Software Engineering
Concordia University
COMP 333 — Data Analytics: Summer 2020 Course Outline
COVID-19 Emergency Remote Learning

These are difficult times. They require changes to our normal methods of instruction, learning, communication, and assessment. First, everything is being done *remotely*:

> All students are required to have access to the internet and the hardware equipment (computer, webcam, microphone).

This is particularly important so you are capable of performing the online timed quizzes. However, it is also important for access to the course material, and communication with the lab demonstrators and instructor using `Zoom`, `slack`, and email. Also

> All students are required to perform the timed quizzes during the scheduled 30-minute period at the end of lecture timeslots.

and

> All students are expected to complete each lab session, submit each lab exercise at the end of the lab session, and preferably be available to communicate with the lab demonstrator during the lab session timeslot.

Second, everything is still operating under *emergency* conditions where major changes to our normal methods of instruction, learning, communication, and assessment are happening over a very short period and this necessitates significant re-organization of the course. So

> I apologize now for any hiccups that may occur during the course and hope everyone can show understanding and tolerance during this difficult time.

But

> note that there may still be circumstances beyond our control due to COVID-19 that force even more changes to how the course runs for this semester.

Unlike Winter semester, when an emergency also existed, you will **not** be able to

1. opt to have the course graded on a pass/fail basis; and
2. opt to DISC the course after you see your final letter grade.

I repeat these options are **not** available to you.

It is important to maintain academic integrity as all assessment is individual assessment of your assignments and quizzes.

The instructor reserves the right to conduct an individual oral examination after each timed quiz to verify the student's response to specific questions.

Read this course outline carefully, as the course is not offered in the normal way this semester. **If it will not work for you then you should drop the course now.**

Calendar Description

COMP 333 Data Analytics (3 credits)

Prerequisite: ENCS 282; COMP 233 or ENGR 371; COMP 352.

The course introduces the process of data analytics with the aid of examples from several disciplines. It covers data wrangling: extract-transform-load (ETL), cleaning, structuring, integration; data analytics activities: description, prescription, modeling, simulation, optimization, story-telling; and the Python ecosystem: language, libraries, and Jupyter environment.

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 the 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 Book

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

Learning Outcomes

The learning outcomes of the course are:

- To know, and be able to carry out, the data analytics process from beginning to end.
- To know the terminology of the field.
- To understand the various types of data, and the issues in analysing each type of data.
- To know the techniques used for each step, and when a technique is appropriate.
- To know how to use the tools available in the Python ecosystem.

Course Components

The web page is `http://users.encs.concordia.ca/~gregb/home/comp333-s2020.html`.

The course consists of lecture material, lab sessions with exercises and assignments, and quizzes, both self-test quizzes for feedback to you through self-learning, and timed quizzes for assessment.

Labs are compulsory.

Lectures

Each week's lecture will be given as four 30-minute chunks provided as written texts. Read these texts carefully; maybe read then out loud; and read them again after doing the associated lab or self-test quiz.

There will be no Zoom lectures, nor Zoom videos. There may be links to related videos and other material.

You will be provided with ample resources on the web for each of the topics and each technology of the course to guide your learning.

Labs

Labs start in Week 2 of the semester (but note this is an accelerated summer course where 13 "semester weeks" are compressed into 6.5 actual weeks).

Labs will use Zoom to connect you with your lab demonstrator during lab sessions, and a permanent `slack` channel for you to discuss lab exercises and the Python ecosystem technology with the lab demonstrators and amongst yourselves.

Do not expect lab demonstrators to be available outside the timeslots for the lab sessions.

Lab Exercises

Each lab will feature a guided lab exercise that you must work through and submit. The submission is essentially an attendance check for the labs, as the labs are the most important component of the course when it comes to learning data analytics.

Collectively the lab exercises are worth 20% of your mark.

Lab Assignments

The lab assignments will require you to apply what you have learned in the lab exercises to other datasets. They will be unguided, and are individual work for which you will be assessed.

There will be four lab assignments. Each lab assignment is worth 10% of your mark.

Quizzes

There will be self-test quizzes for you to do and obtain feedback yourself as part of your self-learning. There will also be "real" quizzes for assessment purposes.

Self-Testing Quizzes

I plan to provide links to quizzes with answers that exist on the web and cover each of the basic topics of the course. These will let you know how you are progressing with the course material.

Timed Quizzes for Assessment

While there is no final examination, and no mid-term examination, there will be four quizzes done in the lecture timeslot as a 30-minute timed quiz.

You may use the provided "cheat sheets" for technology that are on the course web site, but otherwise you should regard these quizzes as closed-book examinations to be done as individually without any outside assistance.

Each timed quiz will be worth 10% of your mark.

You can use the cheat sheets because the quizzes are not meant to be tests of your memory. Rather they aim to test your understanding of the lecture material, of the use of the technology, and of the steps taken in lab exercises that are central to data analytics. To be honest, the cheat sheets will not be of much use to you during a timed quiz.

I plan to have each quiz consist of ten true/false questions and ten multiple choice questions to be completed in 30 minutes. Each question will score 1 mark for a correct answer, with no penalty for an incorrect answer (but 0 marks). Each quiz will be scheduled in the last 30 minutes of a lecture timeslot.

I plan to use Moodle for the timed quizzes.

For these quizzes, all students are expected to have access to the internet and the hardware equipment (computer, webcam, microphone) so they are capable of performing the online timed quizzes.

The instructor reserves the right to conduct an individual oral examination after each timed quiz to verify the student's response to specific questions.

Self-Learning

The course is designed for your self-learning of data analytics, primarily through doing!

So focus on the lab exercises to learn the Python ecosystem: Jupyter (formerly iPython), Python, `pandas` and the related Python libraries; and to learn the major techniques of data analytics: descriptive data analysis, data wrangling (especially data cleaning), and exploratory data analysis.

The lab exercises will also provide an introduction to modeling, machine learning, story telling, and visualization.

Use the Self-Test Quizzes to test your understanding of the material. And if your understanding is not as good as it should be, then consult the supplementary material provided, or consult a resource from the list of resources.

Evaluation

This is the evaluation scheme for the Summer 2020 semester which will be given remotely.

Component	Number	Mark per Contribution	Total Mark
Lab Exercises			20
Lab Assignments	4	10	40
Timed Quizzes	4	10	40
Total			100

Table 1: Evaluation Scheme

There is no final examination, and no midterm examination.

There is no standard correspondence between the numerical marks and the final letter grades.

Students must pass the timed quizzes, combined, in order to pass the course.

Students must pass the lab exercises, combined, in order to pass the course.

Students must pass the lab assignments, combined, in order to pass the course.

We reserve the right to modify the evaluation schema in the light of circumstances outside the control of the University.

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, story-telling. Python ecosystem: language; libraries `numpy`, `scipy`, `pandas`, `matplotlib`, `seaborn`, `scikit-learn`; 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.* To 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 task as a Jupyter notebook, that tells a story, delivers a message, connects to the audience, with appropriate visualizations.

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