



Overview

Introduction

Data in various digital forms originating from a variety of sources is all around us. This course introduces students to fundamental concepts in the data science workflow including data collection, cleaning, organization, exploratory analysis and visualization. The focus is on structured (tabular) data which can be found in a number of domains. The course introduces students to the programming and visualization concepts they need to perform exploratory data analysis. Our language of choice is Python; various Python data structures and libraries for data organization, cleaning and visualization will be introduced in a hands-on way. Students will further develop their critical thinking skills by completing a group project where they will explore and visualize a dataset from a domain of their choice.

Learning Outcomes

By the end of this courses, students should be able to:

- use programmatic tools and suitable programming paradigms to clean and transform data into a format suitable for subsequent analytics.
- use existing software libraries to summarize and aggregate structured data.
- use existing software libraries to explore data and create informative and engaging visualizations.
- present data stories in an interactive and collaborative data science environment.

Staff

Instructor

- Leanne Wu (lewu@ucalgary.ca): L01: MW 9 AM - 12 PM
Office Hours: MW 1-2 PM in ICT 734
- [Usman Alim](#) (ualim@ucalgary.ca): L02: MW 5 - 8 PM
Office Hours: MW 3 - 4 PM in ICT 503

Teaching Assistant

- Haysn Hornbeck (hhornbec@ucalgary.ca)
Haysn will hold drop-in hours on Fridays from 5-7 pm in ICT 517 starting Fri. Sep. 6th.

Software Support

- Kyle Otto Jorgensen (kojorgen@ucalgary.ca)
For all software support inquiries outside of class-time, please email Kyle.
Questions about the course content, assignments and projects should be directed to your TA and/or instructor.

Grading

- Assignments - 60%
- Final project - 40%

The following conversion table will be used to convert your final percentage score to a letter grade.

	>=95	90-94	85-89	80-84	75-79	70-74	65-69	60-64	55-59	50-55	45-49	0-39
A+	A	A-	B+	B	B-	C+	C	C-	D+	D	F	

Communication

We will be making use of the course news and discussion forums. The instructors will post all news related to the class to the news forum while the discussion forum will be used for questions related to the homework assignments, software issues and project ideas.

Prerequisites

In order to successfully complete the assignments and deliver a final project, you should have the following prerequisites.

- Mathematical maturity (Calculus and/or Linear Algebra).
- We will cover the basics of Python in the first week. Some prior exposure to computer programming is expected.

Course Materials

Texts

There is no required textbook for the course.

The lectures and examples covered in the first few weeks are based on the following text:

1. *Wes McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd edition), O'REILLY, 2017.
(available online via the library).

Data Science Environment

We will be making extensive use of the [Jupyter](#) notebook environment for lectures, in-class practice exercises and homework assignments. You have the following methods -- in order of preference -- to access the Jupyter notebook environment:

- Access it on the ARC cluster (preferred):
The Advanced Research Computing (ARC) cluster is managed by Research Computing Services at UCalgary. This is the preferred method as your notebook session will be running on a powerful machine with plenty of RAM. To launch a Jupyter notebook server on ARC, please go to:

<https://jupyter.ucalgary.ca:8000>

Use your UCalgary IT credentials to login.

Please note that if you are connecting from off campus, you will need to be on the UCalgary VPN. For more details on setting up a VPN connection, please consult the [VPN FAQ](#) page.

- Access it on the DATA 60x Jupyter Hub.
Another Jupyter Hub has been setup to support courses in the Data Science Certificate program. This hub is not as powerful as ARC and you will only have 512 MB RAM for your session. However, it gives you access to both Jupyter notebooks and RStudio. Access it at:

<https://jupyter.data60xucalgary.ca>

Sign up for an account using your UCalgary email address.

- Run it locally on your own machine. To do so, you can install the Anaconda environment. Anaconda is a free and open-source distribution of Python and R programming languages. Jupyter and other data science related packages and libraries can be easily installed and managed through

Anaconda. Please download and install Anaconda from [here](#). *Please ensure that you install the Python 3.7 version.*

In any case, it is expected that you will have access to a machine during class time. There are a few machines available for student use in ICT 517. The Data Science program has also purchased some laptops for students to sign out. If you would like to borrow one, please contact Kyle at kojorgen@ucalgary.ca.

Assignments

There will be five *individual* assignments in this course. The assignments will ask you to complete analysis and visualization tasks in Jupyter. You will submit Jupyter notebooks via D2L.

Some portion of class time will be set aside for you to work on your homework assignments.

Outside of class time, if you have questions of a general nature, please make use of the class discussion forum. For more specific questions, please contact your TA. For

The assignments will be weighted as follows:

HW0	HW1	HW2	HW3	HW4
6%	13%	14%	14%	13%

Term Project

A term project is a major component of this course and you are highly encouraged to start thinking about it early. Working in a team, you will identify a suitable dataset to explore and visualize. You will describe the specifics of the dataset and the analysis and visualization in a proposal, and present your results during the last class.

The project consists of the following components:

- Proposal - 25%
- Project report - 60%
- Project demo/presentation - 15%

More details will be made available on the [Project](#) page.