DATA 301
Introduction to Data Analytics
Course Introduction

Dr. Ramon Lawrence
University of British Columbia Okanagan
ramon.lawrence@ubc.ca

The Essence of the Course

The overall goal of this course is for you to:

Understand data analytics and be able to apply data analysis to data sets using a variety of software tools and techniques

This course will provide the tools for you to perform your own data analysis when encountering problems in the real-world.

My Course Goals

1) Provide the information in a simple, concise, and effective way for learning.
2) Strive for ***all*** students to understand the material and pass the course.
3) Be available for questions during class time, office hours, and at other times as needed.
4) Provide an introduction to data analytics tools and techniques so that students are able to apply data analysis to their own data sets.
5) Encourage students to continue with other data analytics or computer science courses.

Course Objectives

1) Understand data representation formats and techniques and how to use them.
2) Experience a wide-range of data analytics tools include Excel, SQL databases, R, and visualization and reporting software.
3) Develop a computational thinking approach to problem solving and use programs and scripting to solve data tasks.
4) Apply techniques to representative problems involving geographical (GIS), business, and scientific data.

Academic Dishonesty

Cheating in all its forms is strictly prohibited and will be taken very seriously by the instructor.

A guideline to what constitutes cheating:

- Assignments
  - Working in groups to solve questions and/or comparing answers to questions once they have been solved (except for group assignments).
  - Discussing HOW to solve a particular question instead of WHAT the question involves.
- Exams
  - All exams are closed book, so no course materials should be present.

Academic dishonesty may result in a "F" for the assignment or course and all instances are recorded in the Dean's office.

How to Pass This Course

The most important things to do to pass this course:

- Attend class
  - Read notes ***before*** class as preparation and try the questions.
  - Participate in class exercises and questions.
- Attend the labs and do all lab assignments
  - Labs are for marks and are practice to learn the material for the exams.

To get an "A" in this course do all the above plus:

- Practice on your own. Practice makes perfect.
  - Do more questions than in the labs. Try the techniques on your own data sets.

Systems and Tools

Connect is used for submitting assignments, posting marks, discussions, and for anonymous feedback.

All software is available in the laboratory in SCI 234.

My Expectations

You should ***SHOW UP TO CLASS AND LABS*** and put in the effort to learn the material. **Attendance ⇨ Success**

There is a wide variety in previous experience.

- Some material you may already know. Help others!!
- Build up your computer experience in labs and outside of class.
- Third-year standing means that you know how to "figure things out."

The course will be very straightforward:

Do the work and practice the techniques to do well.

The Lab Assignments

There are weekly lab assignments using computer software.

Lab assignments are worth ***30%*** of your overall grade.

Lab assignments may take more than the two hours lab time.

You have at least one week after your lab to complete it.

- No late assignments will be accepted.
- An assignment may be handed in any time before the due date.

Lab assignments are done individually or in groups of two depending on the assignment.

The lab assignments are critical to learning the material and are designed both to prepare you for the exams and build up your skills!

The In-Class Quizzes

To encourage attendance and effort, 5% of your overall grade is allocated to answering in-class questions.

These questions are answered electronically using a clicker.

- The clicker can be purchased at the bookstore.
- The clicker is personalized to you with your student number.
- At different times during all the lectures, questions reviewing material will be asked. Reponses are given using clickers.

There will be at least 75 questions throughout the semester. Each question is worth 1 mark, and you need at least 60 right answers to get the full 5%.

- That is, if you answer 45 questions right, you get 45/60 or 75%. Thus, do not worry if you must miss a class or two or forget your clicker one day!

What is Data Analysis?

Data analysis is the processing of data to yield useful insights or knowledge.

- Data processing involves finding, loading, cleaning, manipulating, transforming, modeling, and visualizing the data.
- The knowledge may be used for scientific discovery, business decision-making, or a variety of other applications.

A ***data analyst*** is a person who uses tools and applications to transform raw data into a form that will be useful.

- Data analyst jobs are projected to be one of the top jobs over the next 10 years.
  - See: http://blog.udacity.com/2014/11/data-analysts-what-youll-make.html

Why is Data Analytics Important?

Data analytics is important as society is collecting more and larger data sets all the time:

- Web: All web pages visited and links clicked, searches made, images and posts
- Business: Items purchased by date, supply chain/customers, industrial sensors
- Science: Massive data sets (biological/genomic, astronomy, physics)
- Environmental: Sensors and monitors (temperature, etc.)

and transforming this raw data into useful insights has major value:

- Web: Online advertising driven by understanding customer behaviour
- Business: Sales predictions, marketing promotions, manufacturing improvement
- Science: Scientific discoveries, new medical treatments and drugs
- Environmental: Understanding of environmental processes to allow for changing policies and behaviours

Data Analytics Toolkit

A data analyst has expertise in programming, statistics, data *munging* (transformation), and data visualization.

In this course, you will learn industrial tools and build competency in each one of these skills.

As an introductory course, the goal is to get exposure to the skills and techniques as there will not be time for mastery.

This toolkit of systems and techniques will be useful in many jobs even if they are not considered data analyst positions.

Why are you here?

A) I want to learn more about data analytics.

B) I know how important data is to my work or future work.

C) I need an upper-year elective course.

D) I already have training in computer science/statistics and want to expand my knowledge further.

E) I want an easy credit.

What Topic are You Most Interested In?

A) Excel and SQL Databases

B) Programming and Python

C) Data Visualization and GIS

D) R and Applied Statistics

E) None of the above

What is Your Major?

A) Math/Stat/Computer Science/Engineering

B) Business

C) Science (biology, chemistry, physics, environmental)

D) Arts

E) Other

What is Your Statistics Background?

A) I have taken no statistics courses.

B) I have taken a statistics course – not sure what I remember though.

C) I have taken a statistics course and can explain what a confidence interval is.

D) I have taken multiple statistics courses.

What is Your Computer Background?

A) I can use computer and mobile applications

B) I can write a formula in Excel

C) I can write a simple program in some programming language

D) I can write a query in SQL

E) I am a CS major or have taken several CS courses

What Grade are You Expecting to Get?

A) A

B) B

C) C

D) D

E) F

Why This Course is Important

Many professional jobs of the future will involve collecting, manipulating, and analyzing data. People who can understand how data can be used will have better employment opportunities.

Important results:

- Excel Proficiency – Everyone should know how to use Excel as a general data analysis and productivity software.
- Databases – Understand how they work and how to use them.
- Programming and Computational Thinking – The ability to clearly articulate a problem in a systematic way has applications beyond data analytics.
- Applied Statistics – Using R and other software makes your statistics training useful for real-world problems.
- Real-world problem solving – Your toolkit will allow you to tackle real-world data analysis problems and understand what tool to use and how to proceed.