# DSCI 100: Introduction to Data Science

## Time and Place

### R Sections

| Section | Date | Time | Room | Instructor |
| --- | --- | --- | --- | --- |
| DSCI 100 - 002 | Tues/Thurs | 15:30 - 17:00 | Leonard S Klinck 201 | Melissa Lee |
| DSCI 100 - 005 | Tues/Thurs | 12:30 - 14:00 | Hybrid | Rodolfo Lourenzutti |
| DSCI 100 - 006 | Tues/Thurs | 16:00 - 17:30 | Hennings 200 | Lasantha Premarathna |
| DSCI 100 - 007 | Mon/Fri | 15:00 - 16:30 | Henning 200 | Anthony Christidis |

## Course Description

Use of data science tools to summarize, visualize, and analyze data. Sensible workflows and clear interpretations are emphasized.

**Long Version:** In recent years, virtually all areas of inquiry have seen an uptake in the use of data science tools. Skills in the areas of assembling, analyzing, and interpreting data are more critical than ever. This course is designed as a first experience in honing such skills. Students who have completed this course will be able to implement a data science workflow in the R programming language, by “scraping” (downloading) data from the internet, “wrangling” (managing) the data intelligently, and creating tables and/or figures that convey a justifiable story based on the data. They will be adept at using tools for finding patterns in data and making predictions about future data. There will be an emphasis on intelligent and reproducible workflow, and clear communications of findings. No previous programming skills necessary; beginners are welcome!

## Dealing With COVID-19

**In-Person Section: please see Canvas for rules we are implementing this semester to limit the spread of COVID-19.**

The [COVID-19 pandemic](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen) has affected us all in different ways. It’s okay to not be okay, and you should never hesitate to reach out to your instructor if you need support. Just ask! UBC also has [great student support resources](https://students.ubc.ca/covid19) related to COVID-19 (and otherwise).

Also, please keep in mind that running a course during a pandemic is a new experience for your teaching team. The way this course usually runs (with a lot of close interaction between students and instructors) is not safe when we are trying to limit the spread of COVID-19. So a lot of how we run things in-class will be a bit “experimental,” and we will have to adjust on the fly; please do not hesitate to provide feedback on how we can improve your learning experience.

## Textbook

This course uses [Data Science: A First Introduction](https://ubc-dsci.github.io/introduction-to-datascience/). This textbook is open source and will always be freely available on the web.

## Hardware & Software

Students are required to bring a laptop, chromebook or tablet to both lectures and tutorials. Students who do not own a laptop, chromebook, or tablet may be able to [loan a laptop from the UBC library](https://services.library.ubc.ca/computers-technology/technology-borrowing/).

All other required software will be provided by the instructors. Students will learn to perform their analysis using the [R programming language](https://cran.r-project.org/). Worksheets and tutorial problem sets as well as the final project analysis, development, and reports will be done using [Jupyter Notebooks](http://jupyter.org/) accessed via [Canvas](https://canvas.ubc.ca).

## Prerequisite Knowledge

* distance between points on a graph
* percentages, average
* powers, roots, basic operations, logarithm, exponential
* equation of a line / plane

As an example, British Columbia’s Math 12 or Pre-Calculus 12 courses would satisfy the prerequisite.

## Learning Outcomes

By the end of the course, students will be able to:

* Read data using computation from various sources (local and remote plain text files, spreadsheets and databases)
* Wrangle data from their original format into a fit-for-purpose format.
* Identify the most common types of research/statistical questions and map them to the appropriate type of data analysis.
* Create, and interpret, meaningful tables from wrangled data.
* Create, and interpret, impactful figures from wrangled data.
* Collaborate with others using version control.
* Apply, and interpret the output of simple classifier and regression models.
* Make and evaluate predictions using a simple classifier and a regression model.
* Apply, and interpret the output of, a simple clustering algorithm.
* Distinguish between in-sample prediction, out-of-sample prediction, and cross-validation.
* Calculate a point estimate in the context of statistical inference and explain how that relates to the population quantity being estimated.
* Accomplish all of the above using workflows and communication strategies that are sensible, clear, reproducible, and shareable.

## Teaching Team

*Note that your TAs are students too; they may have class right before their office hours, and they may run a few minutes late. Please be patient!*

| Section | Position | Name | Email | Office Hours | Office Hour Location |
| --- | --- | --- | --- | --- | --- |
| All | Course coordinator | Julia Peng | courses[-at-]stat.ubc.ca | n/a | n/a |
| 002 | Instructor | Melissa Lee | melissa.lee[-at-]stat.ubc.ca | Tuesdays 5 - 6 PM | LSK 201 |
| 005 | Instructor | Rodolfo Lourenzutti | lourenzutti[-at-]ubc.ca | Wednesdays 1 - 2 PM | Zoom |
| 006 | Instructor | Lasantha Premarathna | wpremara[-at-]stat.ubc.ca | Wednesdays 10 - 11 AM | ESB 3174 |
| 007 | Instructor | Anthony Christidis | anthony.christidis[-at-]stat.ubc.ca | Mondays 6 - 7 PM | Zoom |
| All | TA | TBA | n/a | TBA | TBA |

**Please contact the course coordinator about any administrative questions. Please read the course policy (e.g., late registration, missing quiz/assignment due to sickness) below before contacting**.

When sending emails, please include DSCI 100 in the subject line.

## Assessment

* Quiz 1: (invigilated in-person) Same time & location of week 5’s tutorial (Friday Feb 10 for section 007, and Thursday Feb 9 for 002, 005, 006, 100)
* Quiz 2: (invigilated in-person) Same time & location of week 10’s lecture (Friday Mar 24 for section 007, and Thursday Mar 23 for 002, 005, 006, 100)
* Quiz 3: (invigilated in-person) To be scheduled by Classroom Services

**Note**: Since DSCI 100 is a large course with multiple sections (hence, multiple versions of quizzes), the instructors reserve the rights to scale grades in order to maintain equity among sections according the [UBC campus wide policies and regulations](https://www.calendar.ubc.ca/Vancouver/index.cfm?tree=3,42,96,0).

In each class (lecture and tutorial) there will be an assignment:

* Lecture and tutorial worksheet **due dates are posted on Canvas.**
* To open the assignment, click the link (e.g. worksheet\_intro) from Canvas.
* To submit your assignment, just make sure your work is saved **on our server** (File -> Save Notebook to be sure).
* At the deadline, our server will automatically snapshot your work.
* You **must access the lecture and tutorial worksheets through our Canvas course page** (as opposed to the worksheets publicly available via Github). Otherwise your worksheets may not be marked!

### Course breakdown

| Deliverable | Percent Grade |
| --- | --- |
| Lecture worksheets | 5 |
| Tutorial problem sets | 14 |
| Group project | 20 |
| Three quizzes | 60 |
| Bonus regrade percent | 1 |

### Group project breakdown

| Deliverable | Percent Grade |
| --- | --- |
| Proposal | 3 |
| Final report | 11 |
| Team work | 5 |
| Group contract | 1 |

## Schedule

| Week | Topic | Description |
| --- | --- | --- |
| 1 | Introduction | Learn to use the R programming language and Jupyter notebooks as you walk through a real world data Science application that includes downloading data from the web, wrangling the data into a useable format and creating an effective data visualization. |
| 2 | Reading in data locally and from the web | Learn to read in various cases of data sets locally and from the web. Once read in, these data sets will be used to walk through a real world data Science application that includes wrangling the data into a useable format and creating an effective data visualization. |
| 3 | Cleaning and wrangling data | This week will be centered around tools for cleaning and wrangling data. Again, this will be in the context of a real world data science application and we will continue to practice working through a whole case study that includes downloading data from the web, wrangling the data into a useable format and creating an effective data visualization. |
| 4 | Effective data visualization | Expand your data visualization knowledge and tool set beyond what we have seen and practiced so far. We will move beyond scatter plots and learn other effective ways to visualize data, as well as some general rules of thumb to follow when creating visualations. All visualization tasks this week will be applied to real world data sets. Again, this will be in the context of a real world data science application and we will continue to practice working through a whole case study that includes downloading data from the web, wrangling the data into a useable format and creating an effective data visualization. |
| 5 | Version control | Collaboration with version control |
| 5 | Quiz 1 | Cover week 1-4 concepts |
| 5 | Group contract due |  |
| 6 | Classification | Introduction to classification using K-nearest neighbours (k-nn) |
| 7 | Classification, continued | Classification continued |
| 8 | Regression | Introduction to regression using K-nearest neighbours (k-nn). We will focus on prediction in cases where there is a response variable of interest and a single explanatory variable. |
| 8 | Group proposal due |  |
| 9 | Regression, continued | Continued exploration of k-nn regression in higher dimensions. We will also begin to compare k-nn to linear models in the context of regression. |
| 10 | Quiz 2 | Cover classification 1 & 2 and regression 1 & 2 |
| 11 | Clustering | Introduction to clustering using K-means |
| 12 | Introduction to statistical inference | Introduce sampling and estimation for sample means and proportions. |
| 13 | Introduction to statistical inference, continued | Introduce confidence intervals, and calculating them via boostrapping. |
| 13 | Group report due |  |
| TBD | Quiz 3 | To be Scheduled by To be scheduled by Classroom Services |

## Policies

### Late Registration

Students who register for the class late have 1 week from their registration date on Canvas to complete all prior assignments.

### Late Assignments / Quiz Absence

Students **must be present** at the invigilation venue (in class, examination centre, etc) to take quizzes; otherwise they will be considered to have missed the quiz and will be assigned a grade of zero.

Students who will miss a quiz **must provide a self-declaration of academic concession prior to the quiz** (see Canvas homepage for the academic concession form) and make arrangements with the Instructor. Failing to present a declaration within a reasonable timeframe before the quiz will result in a grade of zero.

There will be **no extensions for the lecture and tutorial worksheets**; late assignments will receive a grade of zero. Instead, we will drop the lowest grade on tutorials and worksheets for the semester. However, if you have extenuating circumstances and require further accommodations, please contact the course coordinator.

For all other assignments and the course project, a **late submission will receive a 50% penalty**.

### Autograder Policy

Many of the questions in assignments are graded automatically by software. The grading computer has exactly the same hardware setup as the server that students work on. No assignment, when completed, should take longer than 5 minutes to run on the server. The autograder will automatically stop (time out) for each student assignment after a maximum of 5 minutes; **any ungraded questions at that point will receive a score of 0.**

Students are responsible for making sure their assignments are *reproducible*, and run from beginning to end on the autograding computer. In particular, **please ensure that any data that needs to be downloaded is done so by the assignment notebook with the correct filename to the correct folder.** A common mistake is to manually download data when working on the assignment, making the autograder unable to find the data and often resulting in an assignment grade of 0.

In short: whatever grade the autograder returns after 5 minutes (assuming the teaching team did not make an error) is the grade that will be assigned.

### Re-grading

Students get a free 1% at the end of the course. If you want questions regraded, you can fill out a form at the end of the term documenting in detail what you want regraded. Regrading is only offered if the teaching team makes a mistake. If the total sum of value of possible points via regrading is greater than 1%, you can submit (please do so before the last day of class); otherwise the regrade request is rejected. Note: If you submit a regrade request, you may get less than 1% via regrade.

Note if you received 0 on an assignment when you shouldn’t have, please contact the course coordinator as soon as possible rather than waiting until the end of the term.

### Device/Browser

Students are responsible for using a device and browser compatible with all functionality of Canvas. Chrome or Firefox browsers are recommended; Safari has had issues with Canvas quizzes in the past.

### Missed Final Exam

Students who miss the final quiz must report to their faculty advising office within 72 hours of the missed exam, and must supply supporting documentation. Only your faculty advising office can grant deferred standing in a course. You must also notify your instructor prior to (if possible) or immediately after the exam. Your instructor will let you know when you are expected to write your deferred exam. Deferred exams will ONLY be provided to students who have applied for and received deferred standing from their faculty.

### Academic Concession Policy

Please see [UBC’s concession policy](http://www.calendar.ubc.ca/vancouver/index.cfm?tree=3,329,0,0) for detailed information on dealing with missed coursework, quizzes, and exams under circumstances of an acute and unanticipated nature.

See our Canvas homepage for the academic concession form.

### Academic Integrity

The academic enterprise is founded on honesty, civility, and integrity. As members of this enterprise, all students are expected to know, understand, and follow the codes of conduct regarding academic integrity. At the most basic level, this means submitting only original work done by you and acknowledging all sources of information or ideas and attributing them to others as required. This also means you should not cheat, copy, or mislead others about what is your work. Violations of academic integrity (i.e., misconduct) lead to the breakdown of the academic enterprise, and therefore serious consequences arise and harsh sanctions are imposed. For example, incidences of plagiarism or cheating may result in a mark of zero on the assignment or exam and more serious consequences may apply if the matter is referred to the President’s Advisory Committee on Student Discipline. Careful records are kept in order to monitor and prevent recurrences.

A more detailed description of academic integrity, including the University’s policies and procedures, may be found in the Academic Calendar at <http://calendar.ubc.ca/vancouver/index.cfm?tree=3,54,111,0>.

### Plagiarism

Students must correctly cite any code or text that has been authored by someone else or by the student themselves for other assignments. Cases of plagiarism may include, but are not limited to:

* the reproduction (copying and pasting) of code or text with none or minimal reformatting (e.g., changing the name of the variables)
* the translation of an algorithm or a script from a language to another
* the generation of code by automatic code-generation software

An “adequate acknowledgement” requires a detailed identification of the (parts of the) code or text reused and a full citation of the original source code that has been reused.

The above attribution policy applies only to assignments. **No code or text may be copied (with or without attribution) from any source during a quiz or exam. Answers must always be in your own words. At a minimum, copying will result in a grade of 0 for the related question.**

**Repeated plagiarism of any form could result in larger penalties, including failure of the course.**

## Attribution

Parts of this syllabus (particularly the policies) have been copied and derived from the [UBC MDS Policies](https://ubc-mds.github.io/policies/).