Introduction to Data Science at the University of British Columbia

an accessible course with an emphasis on reproducible workflows

Tiffany Timbers, Ph.D.

Department of Statistics, UBC

2021/01/06 (updated: 2021-01-06)
Course title and calendar Description:

Introduction to Data Science

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

Prerequisites: Grade 12 Math

Syllabus: [https://ubc-dsci.github.io/dsci-100/README.html](https://ubc-dsci.github.io/dsci-100/README.html)
Course history

- Started Development in 2017 by the UBC Department of Statistics
- Course enrollment currently limited by seats offered
Adopted definition of data science

The process of obtaining value (i.e., insight) from data through reproducible and transparent methods.
Course structure

First third focuses on how to use the R programming language to load, wrangle/clean, and visualize data, while answering descriptive and exploratory data analysis questions.

The remaining two thirds of the course illustrate how to solve four common problems in data science, which are useful for answering predictive and inferential data analysis questions.
Statistical questions we focus on

- Predicting a class/category for a new observation/measurement (e.g., cancerous or benign tumour)

- Predicting a value for a new observation/measurement (e.g., 10 km race time for 20 year old females with a BMI of 25).

- Finding previously unknown/unlabelled subgroups in your data (e.g., products commonly bought together on Amazon)

- Estimating an average or a proportion from a representative sample (group of people or units) and using that estimate to generalize to the broader population (e.g., the proportion of undergraduate students that own an iPhone)
What is the question

Data analysis flowchart

Not a data analysis

Descriptive

Exploratory

Did you summarize the data?

Yes

Did you report the summaries without interpretation?

No

Did you quantify whether your discoveries are likely to hold in a new sample?

Yes

Are you trying to predict measurement(s) for individuals?

No

Are you trying to figure out how changing the average of one measurement affects another?

Yes

Is the effect you are looking for an average effect or a deterministic effect?

Average

Deterministic

Causal

Mechanistic

Peng & Leek, Science, 2015
Course outline

Data Science workflow:

Source: *R for Data Science* by Garrett Grolemund & Hadley Wickham

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A vignette</td>
</tr>
<tr>
<td>2</td>
<td>Reading data</td>
</tr>
<tr>
<td>3</td>
<td>Wrangling data</td>
</tr>
<tr>
<td>4</td>
<td>Visualising data</td>
</tr>
<tr>
<td>5</td>
<td>Collaboration &amp; start projects</td>
</tr>
<tr>
<td>6</td>
<td>Classification part I</td>
</tr>
<tr>
<td>7</td>
<td>Classification part II</td>
</tr>
<tr>
<td>8</td>
<td>Regression part I</td>
</tr>
<tr>
<td>9</td>
<td>Regression part II</td>
</tr>
<tr>
<td>10</td>
<td>Work on projects</td>
</tr>
<tr>
<td>11</td>
<td>Clustering</td>
</tr>
<tr>
<td>12</td>
<td>Introduction to statistical inference part I</td>
</tr>
</tbody>
</table>
Key aspects of the course outline for success:

- Learners see & do an entire data analysis in R in week 1!
- Programming skills are taught in the context of simple data analysis
- Predictive questions before inferential questions
- Repeated emphasis on what is the question, and choosing methods based on the question and data in hand.
Key aspects of course organization, mechanics and pedagogy that allow for success
Course organization and mechanics

- Two 90 min meetings a week (lecture + tutorial)
- 3/4 flipped classroom
- Paperless course
- ~60% of assessments are autograded

<table>
<thead>
<tr>
<th>Deliverable</th>
<th>% grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture worksheets</td>
<td>5</td>
</tr>
<tr>
<td>Tutorial problem sets</td>
<td>15</td>
</tr>
<tr>
<td>Group project</td>
<td>20</td>
</tr>
<tr>
<td>Two quizzes/exams</td>
<td>40</td>
</tr>
<tr>
<td>Final exam</td>
<td>20</td>
</tr>
</tbody>
</table>
No modern yet accessible textbooks available that are suitable for our list of topics, programming language, and target learners... at least that I am aware of...

... so we wrote our own using the bookdown R package!

URL: https://ubc-dsci.github.io/introduction-to-datascience
open source and licensed Attribution-NonCommercial-ShareAlike 4.0 International
Lecture worksheets & tutorial homework

- Jupyter notebooks are literate code documents similar to R Markdown
- Markdown and LaTeX rendering in developing environment makes them easier to read while editing
- notebooks can be manually or autograded using an open source tool, nbgrader

Examples of DSCI 100 worksheets:
- worksheet_01
- worksheet_08
Group project

End product is a self-contained reproducible data analysis and report inside a Jupyter notebook
Reducing barriers to entry and success

- Gender and cultural minorities are under represented in STEM
- Aim: remove as many barriers as possible for entry & success in DSCI 100
How?

- Minimal pre-requisites (MATH 12)
- Anonymous class discussion forum (Piazza)
- Formal and public course code of conduct
- **Web server to provide access to homework via the course learning management system (LMS)!**
Summary

DSCI 100 at UBC:

- a first experience for students to gain skills in the areas of assembling, analyzing, and interpreting data

- by the end of the course, students are able to implement an end-to-end data science workflow for "simple" questions

- emphasis is placed on making analysis reproducible and transparent through the use of code in literate code documents (i.e., Jupyter notebooks)

- emphasis is also placed on choosing an appropriate method based on "what is the question" and the data at hand
Acknowledgements

DSCI 100 Development:

- Paul Gustafson
- Matias Salibian-Barrera
- Will Welch
- Nancy Heckman
- Tiffany Timbers
- Melissa Lee
- Samuel Hinshaw
- Melissa Guzman
- Harmeet Gill
- Ian Flores Siaca

DSCI 100 Infrastructure:

- Ian Allison
- Samuel Hinshaw
- The Ha
- Calvin Leung
- Yuvi Pandas

DSCI 100 Teaching Team:

- Tiffany Timbers
- Trevor Campbell
- Melissa Lee
- DSCI 100 TAs
Thanks!

Resources:

- Syllabus: https://ubc-dsci.github.io/dsci-100/README.html
- Textbook: https://ubc-dsci.github.io/introduction-to-datascience/
- Past course work: https://github.com/UBC-DSCI/dsci-100-assets