ETC1010: Data Modelling and Computing

Week of introduction

Professor Di Cook & Dr. Nicholas Tierney

EBS, Monash U.

2019-07-31

What is this song?

(Discuss with your neighbour)


Welcome

What is this course?

This is a course on introduction to **data**, **modelling**, and **computing**.

You can also think of it as introduction to data science or introduction to data analysis.

What is this course?

Q - What data analysis background does this course assume?

A - None.

Q - Is this an intro stat course?

A - Statistics $\neq$ data science. BUT they are closely related. This course is a great way to get started with statistics. But is **not** your typical high school statistics course.

What is this course?

Q - Will we be doing computing?
A - Yes.

Q - Is this an intro Computer Science course?
A - No, but there are some shared themes.

What is this course?

Q - What computing language will we learn?
A - R.

Q: Why not language X?
A: We can discuss that over ☕.

What is this course?

Taught as a **lectorial** (Lecture + Tutorial)

It is **not** recorded because **you** are doing work

About your instructors

Nick

- 🎓 Bachelor of Psychological Sciences UQ
- 🎓 PhD in Statistics at QUT.
- Research: missing data, data visualisation, statistical computing
- R 📦: `naniar`, `visdat`,
- `#rstats` 🎤: Credibly Curious w Saskia Freytag
- ❤️ outdoors, especially: 🥾, 🏃, and 🧗.

Steph

- 🎓 Bachelor of Economics and Bachelor of Commerce from Monash
- Studying a Masters of Statistics at QUT, based at Monash.
- Loves to read 📖, any and all recommendations are welcome.
- Has an R package called taipan, and another called sugarbag.

Mitch

- A data science exploreR
- 🎓 Econometrics, Math Statistics & Computational Science from Monash.
- Compulsively collects and uses data to automate life at home, loves his bees and chickens 🐝🐣.
- Lots of R packages including vitae, icon, and tidy time series forecasting packages.

Sayani

- A statistician and currently in her second year of PhD.
- 🎓 Masters and Honors in Statistics back in India,
- Worked as a consultant and senior analyst in firms like KPMG and American Express.
- Trained in Indian classical vocal music for more than 10 years and loves to nurture that in her spare time.

- Currently an intern with Google Summer of Code 2019 and also on

Sherry

- 🎓 Bachelor of Commerce 2018
- Doing Honours in Econometrics this year with Di Cook
- On her way to have her first ever package, whose name is still a mystery 🤔
- Loves puzzles games like jigsaws 🧩.

Di

- Professor at Monash University in Melbourne Australia, doing research in statistics, data science, visualisation, and statistical computing.
- Created the current version of the course
- I like to play all sorts of sports, tennis, soccer, hockey, cricket, and go boogie boarding.

Your Turn: Turn to the people next to you and ask 2 questions:

- Are you more of a dog or a cat person?

03:00

The *language* of data analysis

This course is
brought to you

today by the

What is R?

> R is a language for data analysis. If R seems a bit confusing, disorganized, and perhaps incoherent at times, in some ways that's because so is data analysis.
>
> -- Roger Peng, 12/07/2018

Why R?

- **Free**
- **Powerful**: Over 14600 contributed packages on the main repository (CRAN), as of July 2019, provided by top international researchers and programmers.
- **Flexible**: It is a language, and thus allows you to create your own solutions

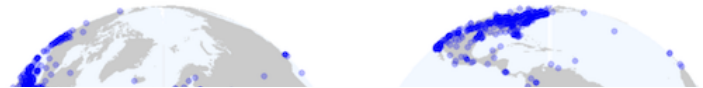
- **Community**: Large global community friendly and helpful, lots of

Community

R Consortium conducted a survey of users 2017.

These are the locations of respondents to an R Consortium survey conducted in 2017.

8% of R users are between 18-24 BUT 45% of R users are between 25-34!



Sample of Australian organisations/companies that sent employees to useR! 2018

> ABS, **CSIRO**, ATO, **Microsoft**, Energy Qld, Auto and General, Bank of Qld, BHP, AEMO, Google, Flight Centre, Youi, Amadeus Investment Partners, Yahoo, Sydney Trains, Tennis Australia, Rio Tinto, Reserve Bank of Australia, PwC, Oracle, **Netflix**, NOAA Fisheries, NAB, Menulog, Macquarie Bank, Honeywell, Geoscience Australia, DFAT, DPI, CBA, Bank of Italy, Australian Red Cross Blood Service, **Amazon**, **Bunnings**.

Traffic Light System

Traffic Light System

Red Post it

--

- I need a hand
- Slow down

Green Post it

--

- I am up to speed
- I have completed the thing

Let's start writing...

Go to bit.ly/LINK SHARED IN CLASS to log in to RStudio cloud.

Log in with Google / GitHub / other credentials.

If you have questions, place a red sticky note on your laptop.

If you are done, place a green sticky on your laptop

This section is based on an exercise from data science in a box by Mine Çetinkaya-Rundel

Create your first data visualisation

- Once you log on to RStudio Cloud, click on this course's workspace "ETC1010 2019 semester 2"
- You should see a project called UN Votes, fork it by clicking on the icon. This will create your copy of the project and launch it.
- In the Files pane in the bottom right corner, spot the file called `unvotes.Rmd`. Open it, and then click on the "Knit" button.
- Go back to the file and change your name on top (in the `yaml` -- we'll talk about what this means later) and knit again.
- Change the country names to those you're interested in. Spelling and capitalization should match the data so take a peek at the Appendix to see how the country names are spelled. Knit again. And voila, your first data visualization!

What can you do at the end of semester?

Some of our best final projects:

- instagram
- babynames
- oztourism
- salary gaps
- FantasyAFL

What you need to learn

> Data preparation accounts for about 80% of the work of data scientists
>
> -- Gil Press, Forbes 2016

Data Preparation

- This is one of the least taught parts of data science, and business analytics, and yet it is what data scientists spend most of their time on.
- By the end of this semester, you will have the tools to be more efficient and effective in this area, so that you have more time to spend on your mining and modeling.

Learning objectives

The learning goals associated with this unit are to:

1. Learn to read different data formats, learn about tidy data and wrangling techniques
2. Apply effective visualisation and modelling to understand relationships between variables, and make decisions with data
3. Develop communication skills using reproducible reporting.

What could this image say about R?

03:00

Philosophy

> If you feed a person a fish, they eat for a day. If you teach a person to fish, they eat for a lifetime.

Whatever I do in the data analysis that is shown to you during the

Course Website: http://dmac.dicook.org

- "dmac" = Data Modeling and Computing
- unit guide (authority on course structure).
- Lecture notes for each class
- Assignment and project instructions
- Textbook + other online resources related to topics
- Consultation times (6 x 1Hr consultations)

MoVE unit

You can use the rstudio cloud server.

In the future we will have R and Rstudio installed locally.

When this happens, you can use USB stick, attach to the borrowed laptop, and install R, RStudio and all your packages on this. Use can then use the USB stick as your working environment, with the borrowed laptop simply as the computing engine.

Grading

Assessment	Weight	Task
Reading Quiz	5%	Complete prior to each class, for the first 8 weeks on ED. Quiz needs to be completed by class time. No mulligans. One can be missed without penalty.
Lab Exercise	5%	Each class period will have a quiz to be completed individually. Two can be missed without penalty.

Grading Example: Reading Quiz

- Before 8am on Friday, you need to complete the 5 question **reading quiz** on ED
- Before 3pm **next Wednesday** You need to complete the 5 question **reading quiz** on ED.

Grading Example: Lab Exercise

There is time at the end of class to complete **lab exercise on ED**:

- Before 5pm Today, you need to complete the 10 question **Lab Exercise** on ED
- Before 10am **This Friday** you need to complete the 10 question **Lab Exercise** on ED.

Grading

Assessment	Weight	Task
Assignment	12%	Teamwork, data analysis challenge, due in weeks 3, 5, 9
Mid-Sem Theory + Concept exam	8%	Due week 6
Data Analysis Exam	10%	Due week 11
Project	10%	Due week 11
Final Exam	50%	TBA

Textbook

- Free

Remember:

All information is on the website 😄

Post questions on ED over email

How do you do well in this class

- Do the reading prior to each class period.
- Participate actively in this class.
- Ask questions on the **ed**.

How do you do well in this class

- Come to consultation if you have questions.
- Practice the materials taught in each lectorial by doing more exercises from the textbook.
- Be curious, be positive, be engaged.

Ed System

- Online quizzes
- Conduct discussions
- Ask questions about the course material and exercises, and turn in assignments and project. *Only your name and email address are recorded in the ED systems.*

Ed System

DEMO

Tips for asking questions

- First search existing discussion for answers. If the question has already been answered, you're done! If it has already been asked but you're not satisfied with the answer, add to the thread.
- Give your question context from course concepts not course assignments.
  - Good context: "I have a question on filtering"
  - Bad context: "I have a question on Assignment 1"

Tips for asking questions

- Be precise in your description:
  - Good description: "I am getting the following error and I'm not sure how to resolve it - `Error: could not find function "ggplot"`"
  - Bad description: "R giving errors, help me! Aaaarrrrrgh!"
- Remember: you can edit a question after posting it.

Diversity & Inclusiveness:

- Intent: Students from all diverse backgrounds and perspectives be well-served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that the students bring to this class be viewed as a resource, strength and benefit. It is my intent to present materials and activities that are respectful of diversity: gender identity, sexuality, disability, age, socioeconomic status, ethnicity, race, nationality, religion, and culture. Let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups.

- If you have a name and/or set of pronouns that differ from those that appear in your official Monash records, please let me know!

Diversity & Inclusiveness:

- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. I want to be a resource for you. If you prefer to speak with someone outside of the course, talk to Di Cook, or look at the services available to you in the Monash student support services.

- I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to me about it.

Sharing / Reusing code

- I am well aware that a huge volume of code is available on the web to solve any number of problems.
- Unless I explicitly tell you not to use something the course's policy is that you may make use of any online resources (e.g. StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). This can be as simple as pasting the link in a references section.

Sharing / Reusing code

- Any recycled code not explicitly cited will be treated as plagiarism.
- Assignment groups may not directly share code with another group.
- You are welcome to discuss the problems together and ask for advice, but you may not make direct use of code from another team.

Group Assignments

Assignment 1 will be announced this Friday.

Group Assignments

Conducted according to the Monash policies.

What we expect:

- Each member of the group completes the entire assignment, as best they can.
- Group members compare answers and combine it into one document for the final asubmission.

Group Assignments

- 25% of the assignment grade will come from peer evaluation.
- Peer evaluation is an important learning tool.

Each student will be randomly assigned another team's submission to provide feedback on three things:

1. Could you reproduce the analysis?
2. Did you learn something new from the other team's approach?
3. What would you suggest to improve their work?

Group Assignments

- Conflicts can arise in group work.
- They can be both productive and destructive.
- Teams need to work on managing conflicts and building on the strengths of all team members.

Group Assignments

- For each assignment, you will be given the option to comment on the efforts of your other group members.
- If a team member has not contributed to an assignment submission, they might score a 0.
- In this situation the team will need to discuss team function and dysfunction with the instructor.

R and RStudio

What is R/RStudio?

- R is a statistical programming language
- RStudio is a convenient interface for R (an integrated development environment, IDE)

What is R/RStudio?

> If R were **an airplane**, RStudio would be **the airport**, providing many, many supporting services that make it easier for you, the pilot, to take off and go to awesome places. Sure, you can fly an airplane without an airport, but having those runways and supporting infrastructure is a game-changer
>
> -- Julie Lowndes

Let's take a tour - R / RStudio

DEMO

R essentials: A short list (for now)

- Functions are (most often) verbs, followed by what they will be applied to in parentheses:

```
do_this(to_this)
do_that(to_this, to_that, with_those)
```

- Columns (variables) in data frames are accessed with `$`:

```
dataframe$var_name
```

- Packages are installed with the `install.packages` function and loaded with the `library` function, once per session:

```
install.packages("package_name")
library(package_name)
```

Concepts introduced:

- How to edit R code
- Creating Data Visualisations
- R
- RStudio
- Console
- Using R as a calculator
- Environment
- Loading and viewing a data frame
- Accessing a variable in a data frame
- R functions

Lab Exercise

Check your knowledge and comprehension by taking your first lab quiz on Ed

Go to the ED page, and complete the lab quiz.

NOTE: Reading assignment on ED site due by 8am before Class on Friday

Share and share alike