ETC1010: Introduction to Data Analysis

Week 1

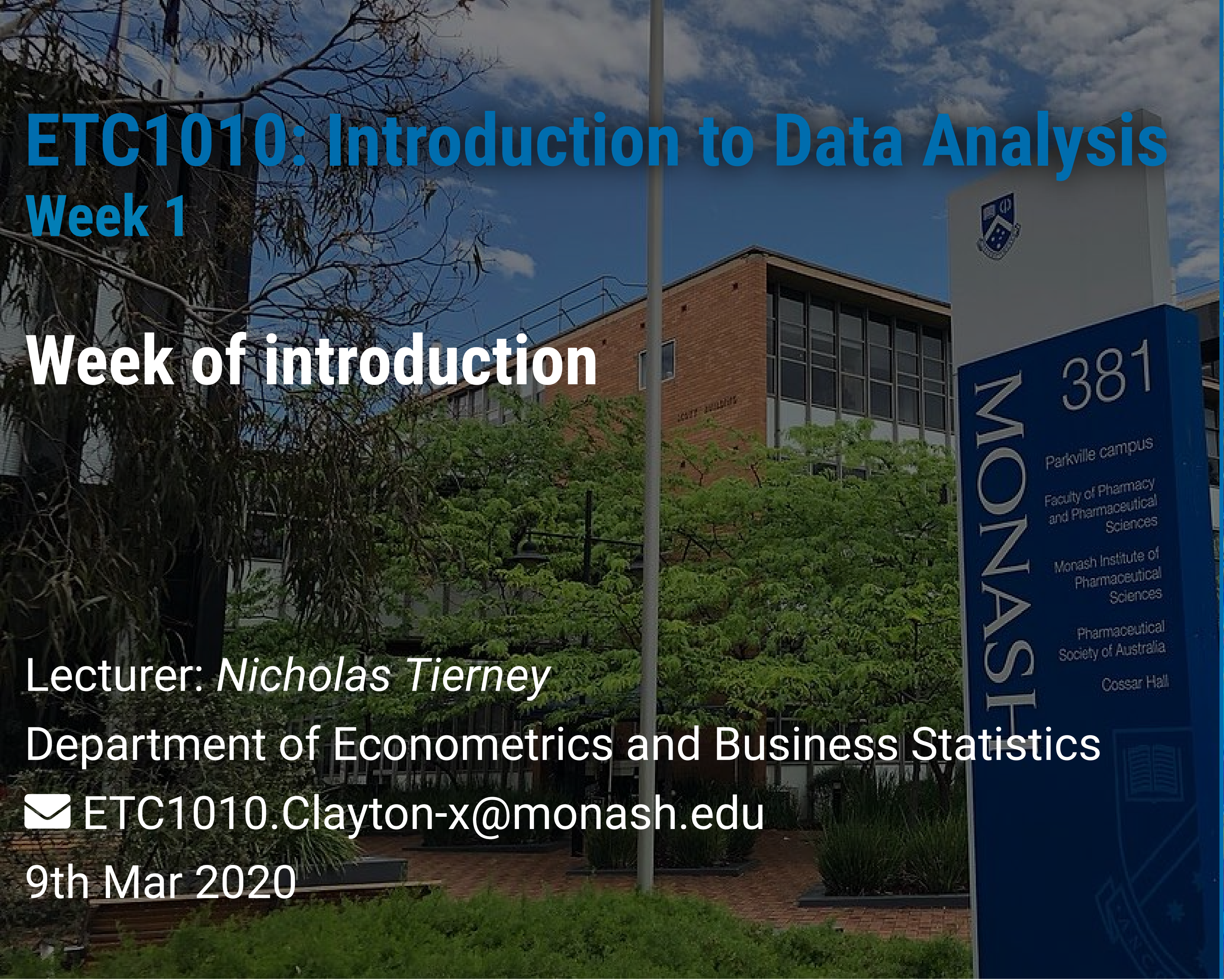
Week of introduction

Lecturer: *Nicholas Tierney*

Department of Econometrics and Business Statistics

✉ ETC1010.Clayton-x@monash.edu

9th Mar 2020


MONASH
381
Parkville campus
Faculty of Pharmacy and Pharmaceutical Sciences
Monash Institute of Pharmaceutical Sciences
Pharmaceutical Society of Australia
Cossar Hall

Welcome

What is this course?

This is a course on introduction to **data analysis**.

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

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.

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

What is this course?

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

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

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

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

It is **not** (typically) recorded because **you** are doing work

You have to show up to class to practice!

The *language* of data analysis

This course is brought to you today by the letter "R"!

Grover image sourced from https://en.wikipedia.org/wiki/Grover.

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?

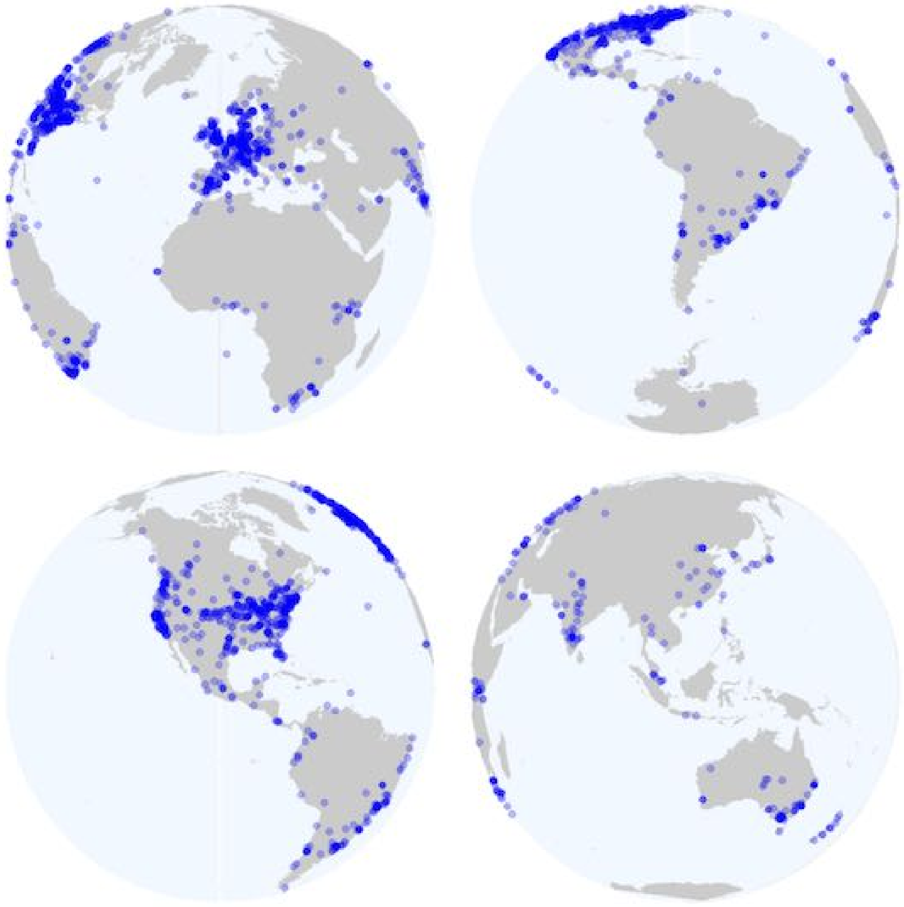
- **Free**
- **Powerful**: Over 15000 contributed packages on the main repository (CRAN), as of March 2020, provided by top international researchers and programmers.
- **Flexible**: It is a language, and thus allows you to create your own solutions
- **Community**: Large global community friendly and helpful, lots of resources

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**.

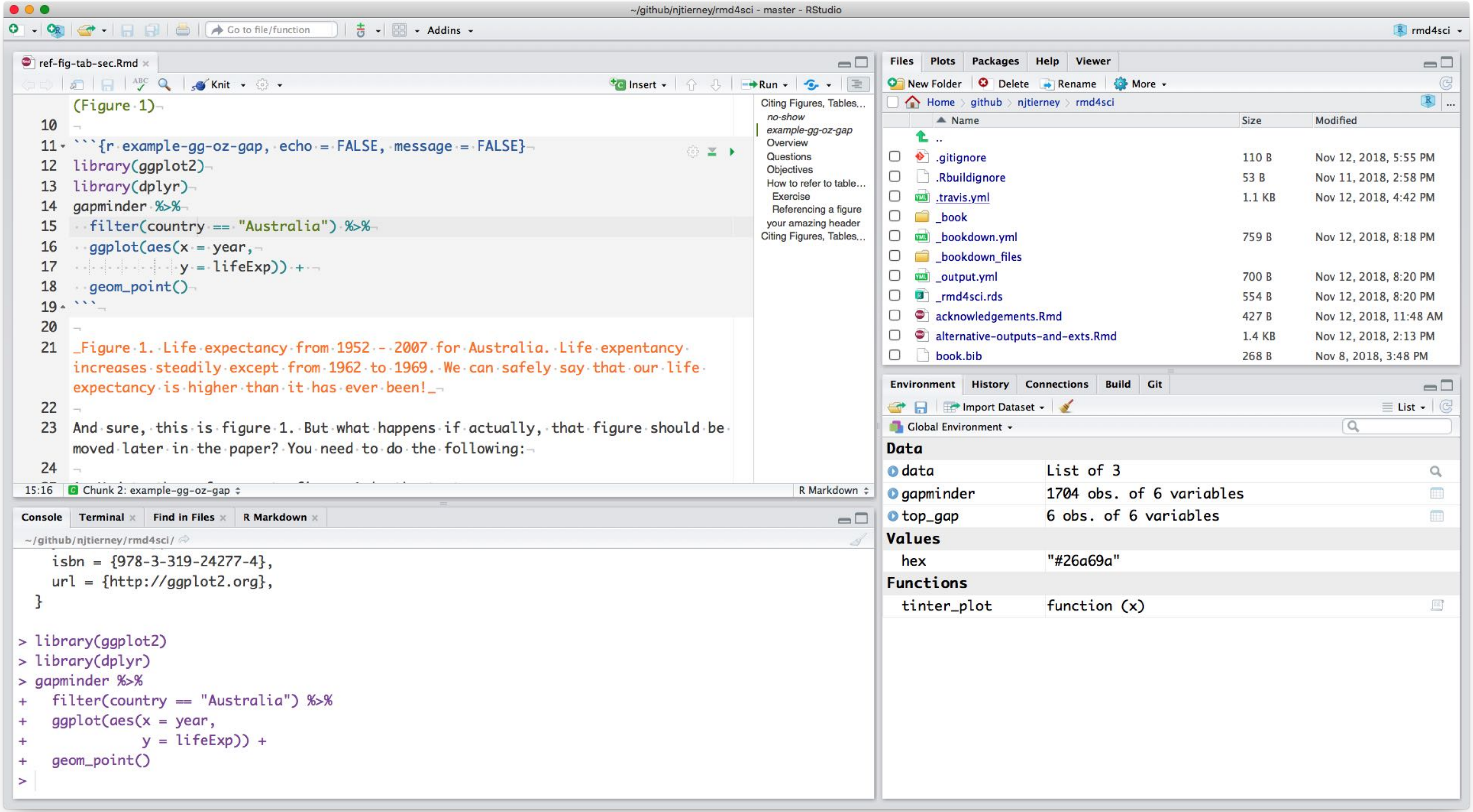
R and RStudio

What is R/RStudio?

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

> 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 of R and RStudio



~~~
(Figure 1)

```{r example-gg-oz-gap, echo = FALSE, message = FALSE}
library(ggplot2)
library(dplyr)
gapminder %>%
  filter(country == "Australia") %>%
  ggplot(aes(x = year,
             y = lifeExp)) +
  geom_point()
```

_Figure 1. Life expectancy from 1952 - 2007 for Australia. Life expentancy increases steadily except from 1962 to 1969. We can safely say that our life expectancy is higher than it has ever been!_

And sure, this is figure 1. But what happens if actually, that figure should be moved later in the paper? You need to do the following:
~~~

Name	Size	Modified
..		
.gitignore	110 B	Nov 12, 2018, 5:55 PM
.Rbuildignore	53 B	Nov 11, 2018, 2:58 PM
.travis.yml	1.1 KB	Nov 12, 2018, 4:42 PM
_book		
_bookdown.yml	759 B	Nov 12, 2018, 8:18 PM
_bookdown_files		
_output.yml	700 B	Nov 12, 2018, 8:20 PM
_rmd4sci.rds	554 B	Nov 12, 2018, 8:20 PM
acknowledgements.Rmd	427 B	Nov 12, 2018, 11:48 AM
alternative-outputs-and-exts.Rmd	1.4 KB	Nov 12, 2018, 2:13 PM
book.bib	268 B	Nov 8, 2018, 3:48 PM

```
    isbn = {978-3-319-24277-4},
    url = {http://ggplot2.org},
  }

> library(ggplot2)
> library(dplyr)
> gapminder %>%
+   filter(country == "Australia") %>%
+   ggplot(aes(x = year,
+              y = lifeExp)) +
+   geom_point()
>
```

Data

data	List of 3
gapminder	1704 obs. of 6 variables
top_gap	6 obs. of 6 variables

Values

hex	"#26a69a"

Functions

tinter_plot	function (x)

End of part 1 of Lecture 1A

Start of part 2 of Lecture 1A

Let's start writing...

Go to http://bit.ly/etc1010-s1-2020 to log in to RStudio cloud.

Log in with Google / GitHub / other credentials.

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 2020 semester 1"

📊 You should see a project called "UN Votes", click on the icon to create a copy of the project, and launch it.

📊 In the Files pane in the bottom right corner, open the file called `unvotes.Rmd`. 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!

End of part 2 of Lecture 1A

Start of part 3 of Lecture 1A

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)
```

For example:

```
mean(c(1,2,1,2))
## [1] 1.5
```

R essentials: A short list (for now)

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

```
dataframe$var_name
```

For example:

```
starwars$name
##  [1] "Luke Skywalker"        "C-3PO"                 "R2-D2"
##  [4] "Darth Vader"           "Leia Organa"           "Owen Lars"
##  [7] "Beru Whitesun lars"    "R5-D4"                 "Biggs Darklighter"
## [10] "Obi-Wan Kenobi"        "Anakin Skywalker"      "Wilhuff Tarkin"
## [13] "Chewbacca"             "Han Solo"              "Greedo"
## [16] "Jabba Desilijic Tiure" "Wedge Antilles"        "Jek Tono Porkins"
## [19] "Yoda"                  "Palpatine"             "Boba Fett"
## [22] "IG-88"                 "Bossk"                 "Lando Calrissian"
## [25] "Lobot"                 "Ackbar"                "Mon Mothma"
## [28] "Arvel Crynyd"          "Wicket Systri Warrick" "Nien Nunb"
## [31] "Qui-Gon Jinn"          "Nute Gunray"           "Finis Valorum"
## [34] "Jar Jar Binks"         "Roos Tarpals"          "Rugor Nass"
## [37] "Ric Olié"              "Watto"                 "Sebulba"
```

R essentials: A short list (for now)

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

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

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

- 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.

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 class, you can do it, too.

Course Website: ida.numbat.space

- "ida" = Introduction to Data Analysis
- "numbat" = Non-Uniform-Monash-Business-Analyics-Team
- unit guide (authority on course structure).
- Lecture notes for each class
- Assignment and project instructions
- Textbook + other online resources related to topics
- Consultation times (7 x 1Hr consultations)
- demo

Using laptops

- We will start out using the rstudio cloud server.
- In the future we will have R and Rstudio installed locally.
- This course is also set up as a "MoVE unit", which means you can borrow a laptop from the university for class hours.
- It is also possible to set up R and RStudio onto a USB stick to use with your borrowed laptop.

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 12pm (noon) on Wednesday, you need to complete the 5 question **reading quiz** on ED
- Before 4pm **next Monday** You need to complete the 5 question **reading quiz** on ED.

Grading Example: Lab Exercise

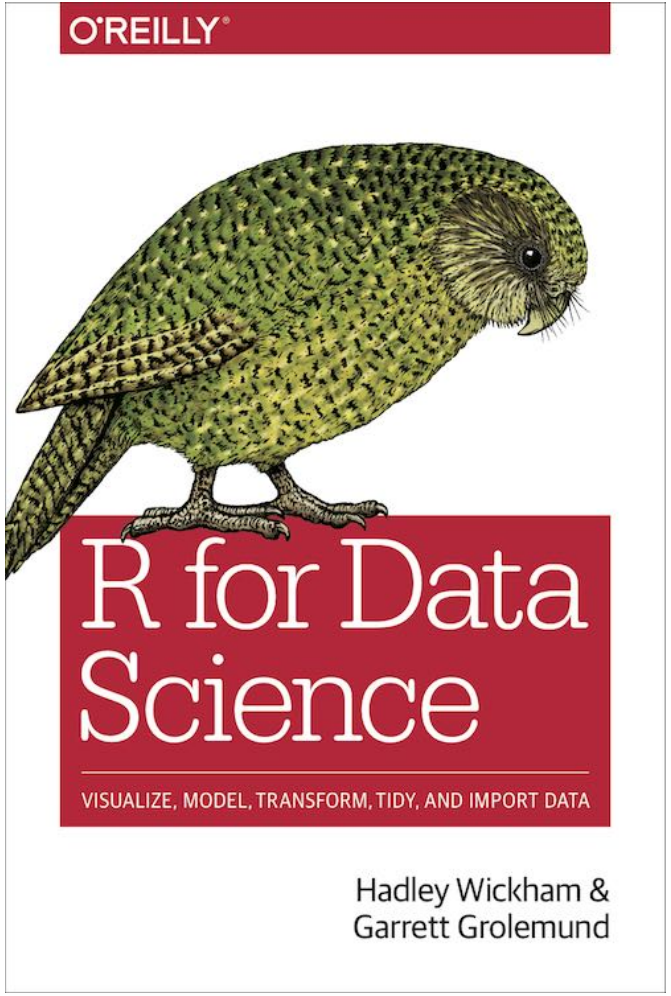
There is time at the end of class to complete **lab exercise on ED**:

- Before 6pm **Next Monday (16th March)**, you need to complete the 10 question **Lab Exercise** on ED
- Before 4pm **Mext Wednesday (18th March)** you need to complete the 10 question **Lab Exercise** on ED.

Grading

Assessment	Weight	Task
Assignment	12%	Teamwork, data analysis challenge, due in weeks 4, and 8
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

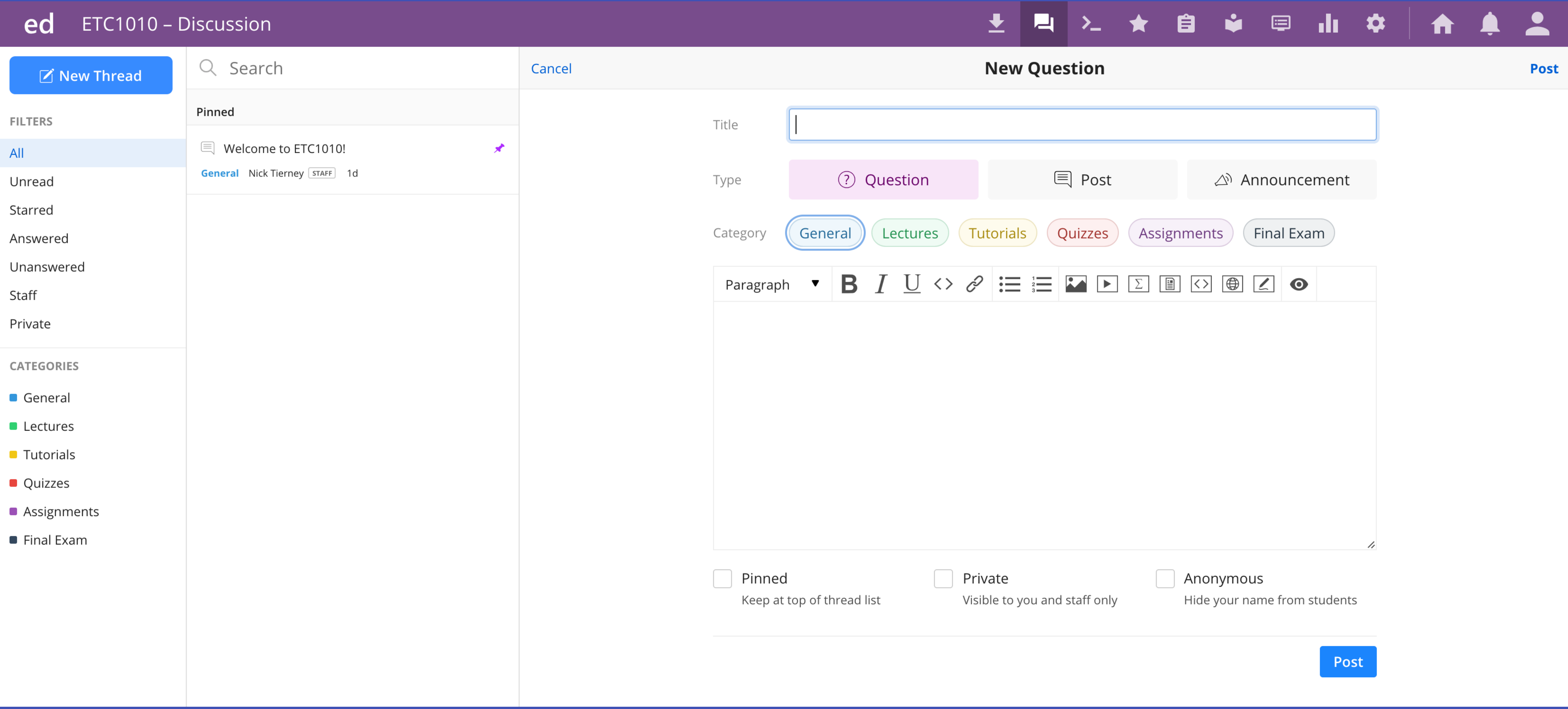


- Free
- Written by authors of Tidyverse R packages

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.*

(DEMO)



Welcome to ETC1010!

General Nick Tierney STAFF 1d

New Question

Title

Type: [✓ Question] Post Announcement

Category: [✓ General] Lectures Tutorials Quizzes Assignments Final Exam

□ Pinned
Keep at top of thread list

□ Private
Visible to you and staff only

□ Anonymous
Hide your name from students

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 data"
  - 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.

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.

Remember:

All information is on the website 😄

Post questions on ED **instead of** questions over email

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.

Diversity & Inclusiveness:

- 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!

- 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.

Diversity & Inclusiveness:

- 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

What we expect:

- Conducted according to the Monash policies.
- 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.
- 25% of the assignment grade will come from peer evaluation.
- Peer evaluation is an important learning tool.

Group Assignments: Peer evaluation

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: Working in groups

- 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: Working in groups

- 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.

Group Assignments

Assignment 1 will be announced at class on Monday Week 2

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

MONASH
381
Parkville campus
Faculty of Pharmacy and Pharmaceutical Sciences
Monash Institute of Pharmaceutical Sciences
Pharmaceutical Society of Australia
Cossar Hall

That's it!

Lecturer: Nicholas Tierney

Department of Econometrics and Business Statistics

✉ ETC1010.Clayton-x@monash.edu

9th Mar 2020