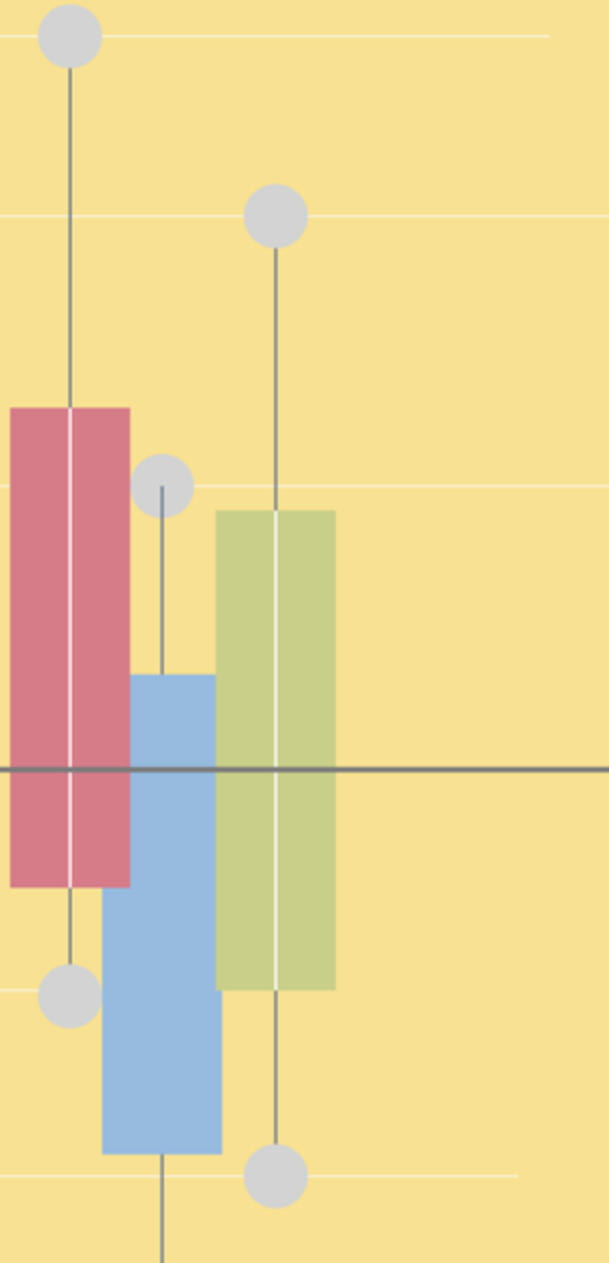
Splitting the load

Making students comfortable with R
before throwing t-tests at them

Maria Christodoulou and Mariagrazia Zottoli

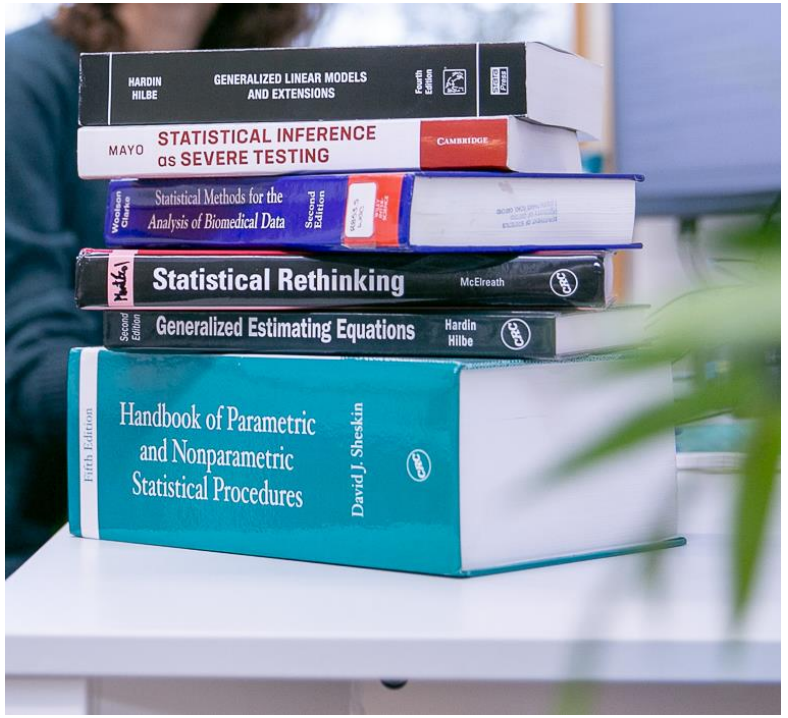


Who we are and who do we teach

- We are based in the Department of Statistics but we very rarely teach statisticians or even mathematicians
- Our main audience is graduate researchers in Life Sciences, Earth Sciences, and Psychiatry

Who we are and who we teach

- The content we deliver can range from fundamentals of inference to relatively complex models
- R teaching is rarely the main focus of training but an absolutely necessary skill to be able to work through the materials available



What makes our audience unique?

- Diverse backgrounds in terms of quantitative experience
- May have had courses during undergraduate on R and stats
- Some can be quite anxious about the course and their ability to do well

What makes our audience unique?

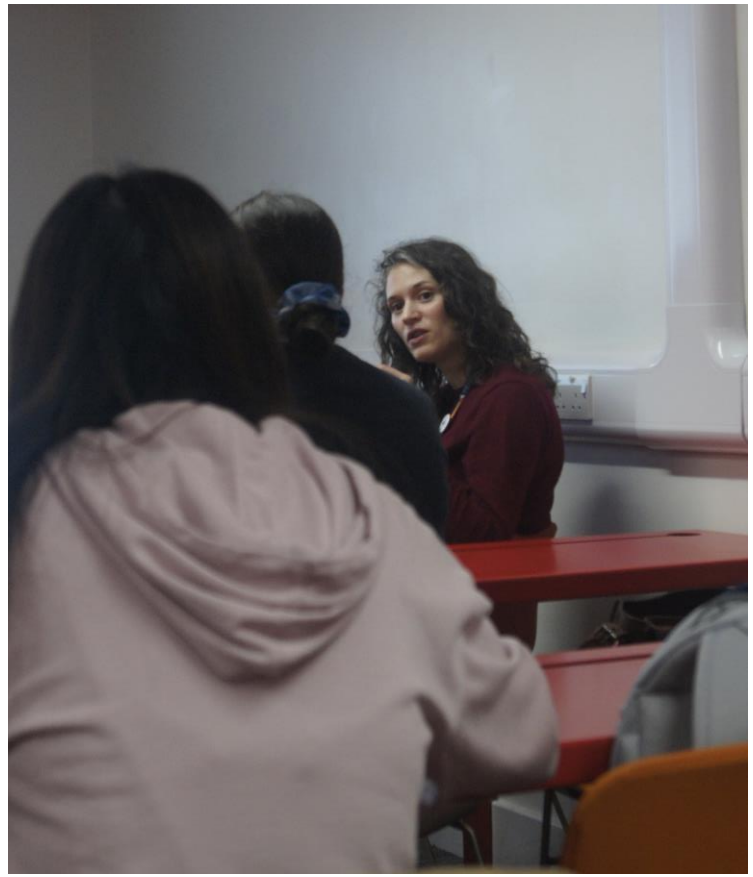
- Their primary interest is how they can use what they learn rather than stats for their own sake
- They may not respond as well to certain teaching techniques that are common in mathematical and physical sciences

Delivery setup

- Graduate training in our institution is often in intensive blocks
- For beginners, delivery is often in one or two week units, from 9.30-17.30
- It is also during the first term of their programme

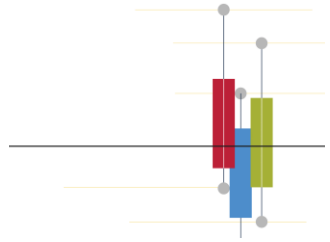
Issues with delivery

- Too intensive, limited processing time
- Large amount of content on topics they are already anxious about
- They have just started a new programme and may be dealing with a lot of anxiety



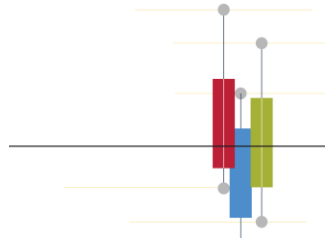
The most common setup we encounter

Time	Monday 15th November	Tuesday 16th November	Wednesday 17th November	Thursday 18th November	Friday 19th November	Time
09:30	Lecture: Course Introduction	Introduction to formal statistical tests: t-test and chi-squared	Lecture: Meet your demonstrators & data science tips	Lecture: ANOVA	Lecture: Plotting models and statistical reporting	09:30
09:45						09:45
10:00	Break					10:00
10:15	Lecture: Introduction to R - Objects				Break	10:15
10:30		Break		Break	Lecture: Effects and Effect sizes	10:30
10:45		Practical: Basic statistical tests- t-tests and Chi-squared test		Practical: ANOVA		10:45
11:00	Break		Break			11:00
11:15	Lecture: Introduction to R - Object				Break	11:15
11:30			Lecture: Tips and Tricks for Surviving and Thriving in your Dphil		Practical: Effects and power	11:30
11:45						11:45
12:00	Break	Break				12:00
12:15	Practical: How to manipulate data objects in R	Practical: Catch up session				12:15
12:30			Introduction to projects week			12:30
12:45						12:45
13:00	Lunch	Lunch	Free time	Lunch	Lunch	13:00
13:15						13:15
13:30						13:30
13:45						13:45
14:00	Lecture: Hypothesis testing, Descriptive statistics, and Exploratory Data Analysis PART I	Lecture: Correlations		Lecture: Regression and linear models PART I	Lecture: Data Wrangling PART I	14:00
14:15						14:15
14:30						14:30
14:45	Break			Break	Break	14:45
15:00	Lecture: Descriptive statistics, and Exploratory Data Analysis PART II, Confidence Intervals	Break		Lecture: Regression and linear models PART II	Lecture: Data Wrangling PART II	15:00
15:15		Practical: Correlations				15:15
15:30						15:30
15:45	Break			Break	Break	15:45
16:00	Practical: Summary Statistics, basic visualisations and Exploratory Data Analysis			Practical: Regression)	Practical: Data wrangling using tidyverse	16:00
16:15						16:15
16:30						16:30
16:45						16:45
17:00						17:00
17:15						17:15
17:30	Finish	Finish		Finish	Finish	17:30



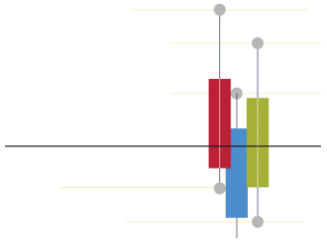
The most common setup we encounter

Time	Monday 22nd November	Tuesday 23rd November	Wednesday 24th November	Thursday 25th November	Friday 26th November	Time
09:30	Lecture: Experimental Design	Lecture: Introduction to reproducibility	Lecture: So now you have data	Lecture: Introduction to machine learning I	Lecture: The FAIR Principle and Data Readiness	09:30
09:45						09:45
10:00						10:00
10:15	Break		Break			10:15
10:30	Lecture: Graphics and Advanced Graphics	Break	Practical: Practising data through to analysis	Break	Break	10:30
10:45		Simulation of data and data analyses		Practical: Introduction to machine learning I		10:45
11:00					Lecture: Data FAIRification in Practice: Example	11:00
11:15	Break					11:15
11:30	Practical: Data visualisations using R					11:30
11:45						11:45
12:00					Q & A Session	12:00
12:15						12:15
12:30					Lunch	12:30
12:45						12:45
13:00	Lunch	Lunch	Free Time	Lunch		13:00
13:15						13:15
13:30					Practical 1: Reproducible workflow	13:30
13:45						13:45
14:00	Lecture: Multivariate methods	Lecture: Rmarkdown		Lecture: Introduction to machine learning II		14:00
14:15						14:15
14:30						14:30
14:45	Break					14:45
15:00	Practical: Multivariate methods	Break		Break		15:00
15:15		Practical: RMarkdown		Practical: Introduction to machine learning II	Break	15:15
15:30						15:30
15:45					Practical 2: Collaborative coding with Github	15:45
16:00						16:00
16:15						16:15
16:30						16:30
16:45						16:45
17:00						17:00
17:15						17:15
17:30	Finish	Finish		Finish	Finish	17:30



What was our experience teaching this way?

- We noticed that some students were already overwhelmed by late Monday/ early Tuesday
  - Hard to regain their trust and de-escalate
- By the time we introduced tidyverse, on Friday, some had already decided they “hated R”
  - Those who did enjoy tidyverse felt particularly betrayed that the majority of that week’s content was delivered in base R



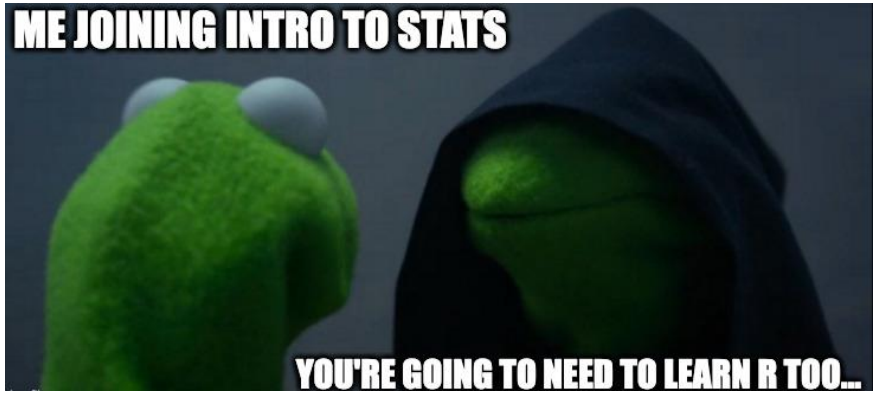
What was our experience teaching this way?

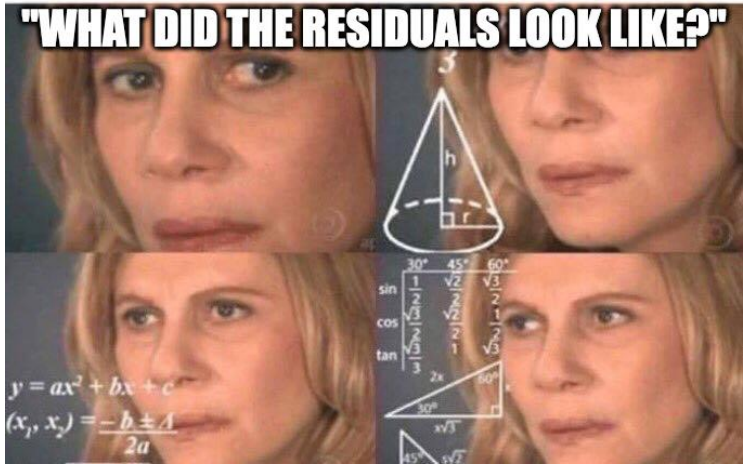
- Because of the mix of R and stats training it was hard for us to disentangle whether someone was struggling with:
  - the coding
  - understanding the methods
  - both

What did our students think?

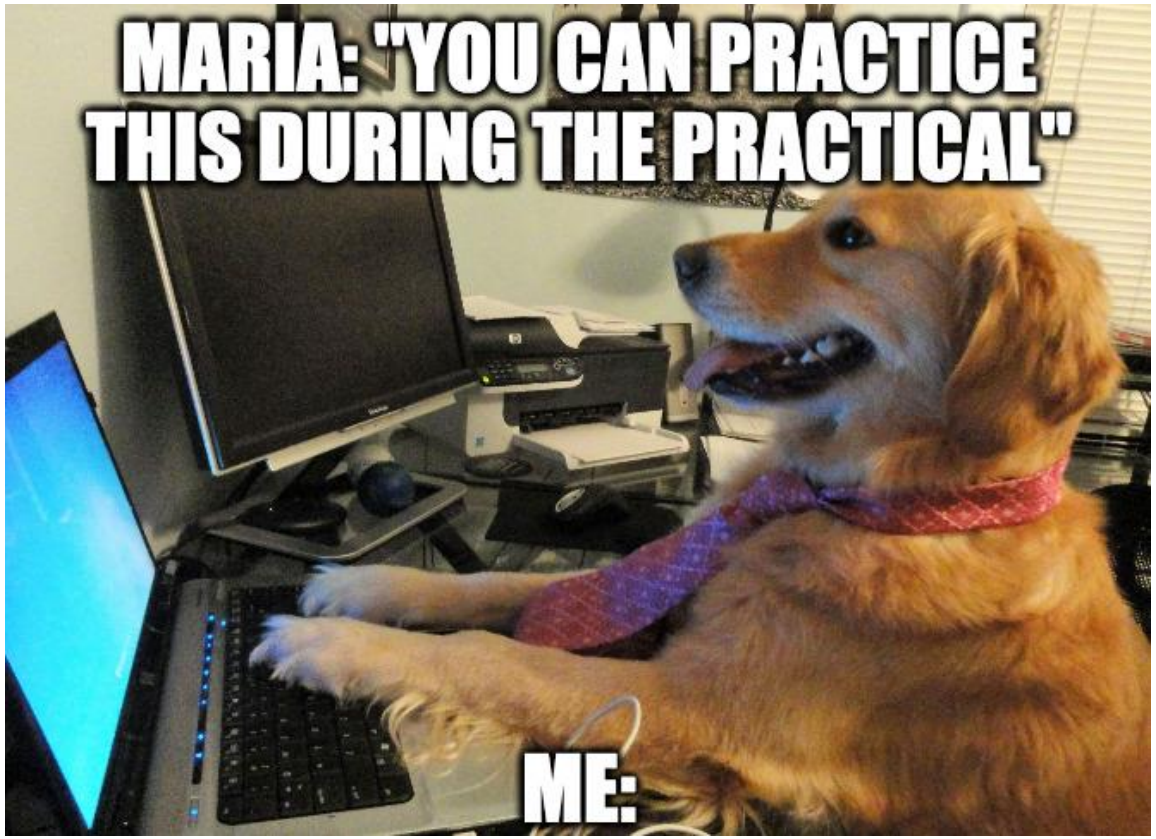
- Positive feedback on our style, delivery, empathy, and approachability
- Students expressed concerns on intensity, amount of material, content, and scheduling
- Many expressed feelings of frustration and being overwhelmed
- They also sent memes…

So...many...memes...



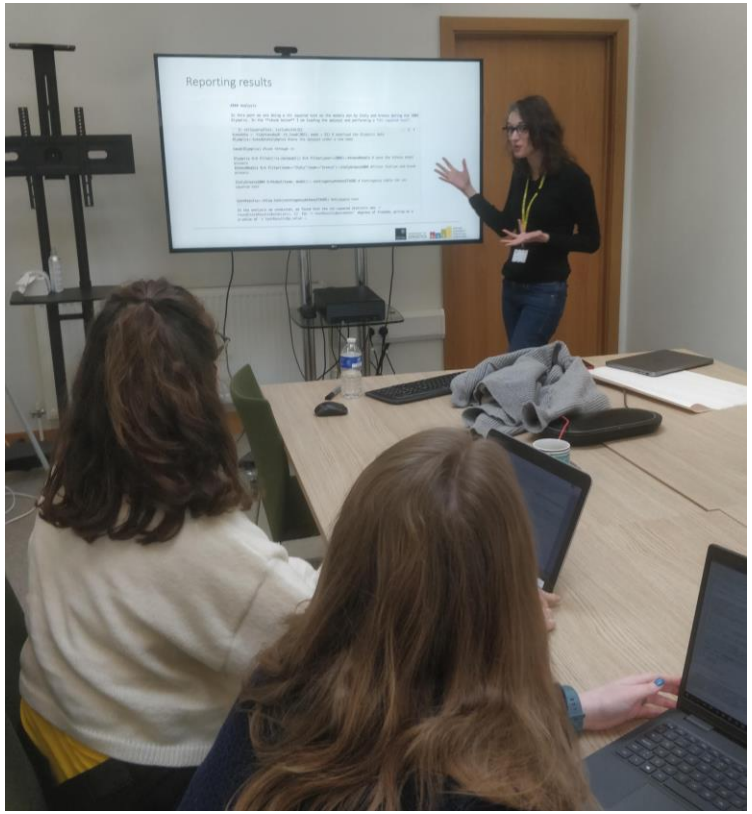


So...many...memes...



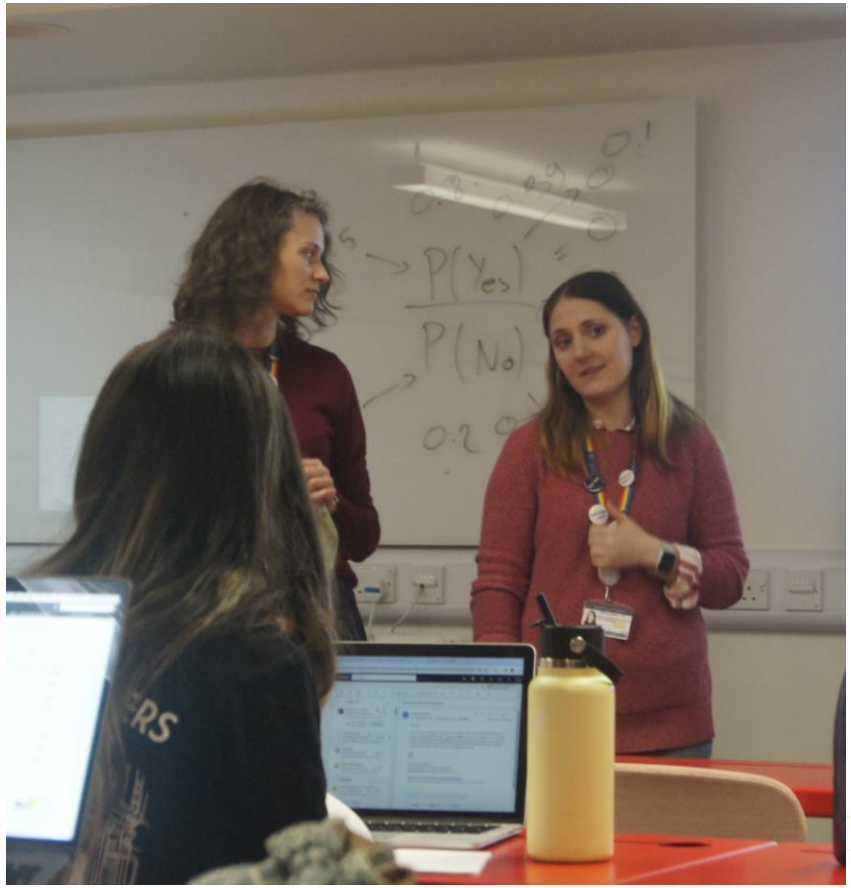
What can we change?

- We have limited control over the syllabus, but we can change the timings, and format
- Things we wanted to attempt:
  - Change the format of worksheets
  - Separate R from stats to begin with
  - Bring the tidyverse earlier



Separating stats from R

- Stops them from multitasking and focusing on learning one skill before introducing the next
- Allows us to troubleshoot what may be causing anxiety



Bringing the tidyverse earlier

- We can get them coding faster
- A lot of data wrangling tasks can be conducted with the 5 verbs of dplyr
- Builds their confidence in a sandbox

Changing the format of worksheets

- Tricky balance of scaffolding but allowing independent thinking
- We started building our worksheets like lab protocols
- As the course progressed, we made our worksheets less prescriptive

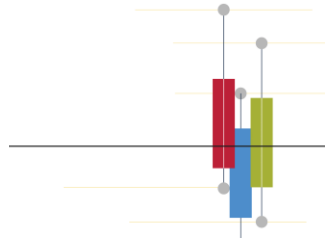
First worksheet

Part III - Indexing

Selecting objects can be done in many ways. Here we will guide you through indexing. Indexing is using square brackets to select by position. For a dataframe we have two dimensions, rows and columns. If for example if we wanted to select the 4th row (Batman) we would:

```
> superheroes[4,]
# A tibble: 1 × 5
  Name   Alignment `Gender (self identified)` Publisher Height
  <chr>  <chr>     <chr>                      <chr>       <dbl>
1 Batman Bad       Male                       DC Comics    1.95
```

Notice that we specified the dataset, then that we are indexing (square brackets), then row (4) and then the column (comma followed by closing of indexing brackets). What this says is: "take the superheroes dataset, then show me all columns for row 4."



Third worksheet

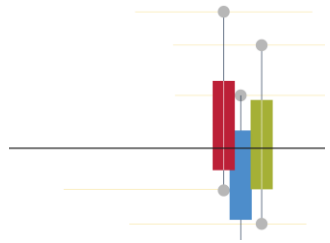
Part I – Summary statistics

For this part we will be exploring the `datasauRus` package in R.

Tasks:

1. Install and load the `datasauRus` package
2. We will be working with the `datasaurus_dozen` tibble, which contains multiple datasets. How many datasets are in the `datasaurus_dozen` and how many entries per dataset are there?
3. Create a table with the summary statistics by dataset in the `datasaurus_dozen` tibble. Ideally, we want a table or tibble that has a dataset by row, and then the mean and standard deviations for x and y as columns, as well as any other summary statistics you may like (e.g. interquartile range etc). A couple of functions that you may want to consider using in R are:
    - `group_by()`
    - `summarise()`
4. What do you notice about the summary statistics?
5. Create scatterplots for x against y for each dataset. To do so use the function plot(). Some instructions on very basic scatterplot can be found in the resources below. Is this what you expected from the summary statistics?

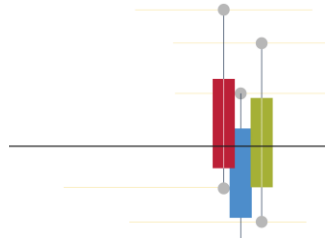
There is no big lesson here other than summary statistics alone are not the entire story. When you explore your data always make sure you plot them as well as check summary statistics.



Fifth worksheet

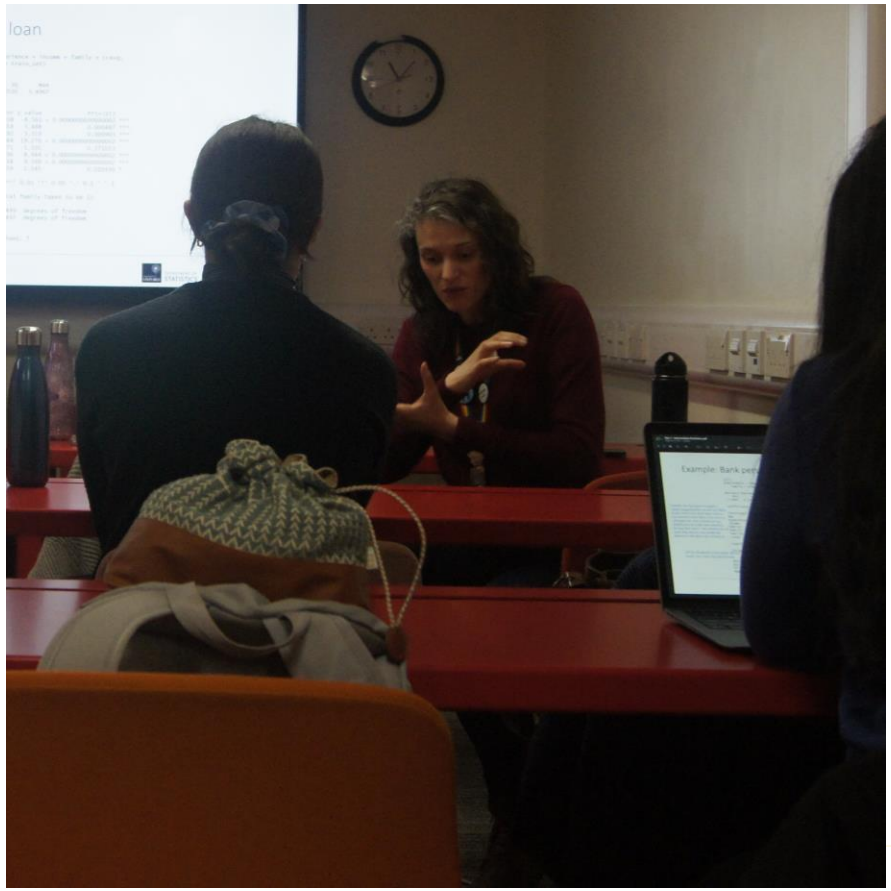
Tasks:

1. Use any of the exploratory data analysis tools to learn the dataset and get an understanding of how the variables behave
2. Using data wrangling tools, create a new dataset that contains **all medal winners** for **Italy** and **Greece** during the **Athens 2004 summer Olympics**.
    a. Checkpoint: the filtered dataset should have 135 entries and 15 variables.
3. Use the `janitor` package (vignette link below) to create a contingency table for medal type between Italy and Greece (the function you should look up is "`tabyl`") and save it under a name
4. Did Italy win a significantly different number of medals than Greece? How would you test this? Look through the `janitor` package for a version of this test that would take as input the `tabyl` contingency table
5. How do you interpret the results?
6. Is the height of the Italian medal winners competing in the "M" (male) categories significantly different to the height of the Greek medal winners competing in the "M" categories for the summer 2004 Olympics? What test would you conduct? How would you test the assumptions? If you need inspiration look through the resources below
7. Filter the **Athens medal winners** dataset to keep only the athletes competing in the "**F**" (female) category for **Brazil**, **Canada**, **France**, **Great Britain**, **Poland**, and **Spain**, and save this under a new name.
    a. Checkpoint: the filtered dataset should have 81 entries and 15 variables
8. Are there any significant differences in height between these athletes when grouped by country? How would you test for this? What assumptions do you need to verify?



What worked…

- Gently builds confidence
- Easy to pinpoint where they need more support and improve materials
- Requires fewer demonstrators than the original worksheet



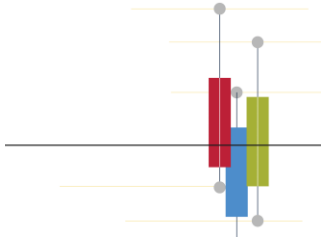
…and what didn’t

- Time consuming to make
- “Looks like a lot” and may put students off
- Students may blindly follow steps for early practicals and not learn as much

Did we decide to keep the changes?

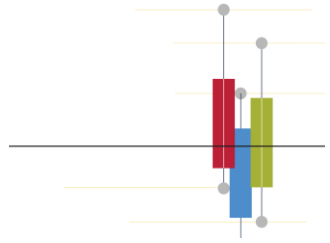
Somewhat…

- We liked the scaffolding
- We liked how it gave us an opportunity to give them worked examples that they can refer back to in years to come
  - e.g. our regression practical doubles as lecture notes for how to interpret output and diagnostics
- The amount of labour it required means that refreshing them is a big commitment
- We want to develop more dynamic practical sheets that we can update more regularly



Let’s start with R...

Time	Tuesday 5th July	Wednesday 6th July	Thursday 7th July	Friday 8th July	Time
09:30	Programme Introduction	Lecture: Data Wrangling	Lecture: Introduction to formal statistical tests: t-tests and Chi-squared test	Lecture: ANOVA	09:30
09:45					09:45
10:00					10:00
10:15		Break			10:15
10:30		Lecture: Data Wrangling	Break	Break	10:30
10:45			Practical: Basic statistical tests - t-tests and Chi-squared test	Practical: ANOVA	10:45
11:00					11:00
11:15		Break			11:15
11:30		Practical: Data wrangling using tidyverse			11:30
11:45					11:45
12:00					12:00
12:15					12:15
12:30		Catch up session OR leave early to meet with supervisor	Catch up session OR leave early to meet with supervisor	Catch up session OR leave early to meet with supervisor	12:30
12:45					12:45
13:00		Lunch	Lunch	Lunch	13:00
13:15					13:15
13:30					13:30
13:45					13:45
14:00	Lecture: Course Introduction	Lecture: Advanced graphs	Lecture: Hypothesis testing and power and effect	Lecture: Correlation	14:00
14:15					14:15
14:30	Break				14:30
14:45	Lecture: Introduction to R - Objects	Break	Break	Break	14:45
15:00		Lecture: Advanced graphs	Lecture: Hypothesis testing and power and effect	Lecture: Introduction to regression	15:00
15:15					15:15
15:30	Break				15:30
15:45	Lecture: Introduction to R - Objects	Break	Break	Break	15:45
16:00		Practical: Data visualisations using R	Practical: Summary Statistics, basic visualisations and Exploratory Data Analysis	Practical: Regression	16:00
16:15	Break				16:15
16:30	Practical: How to manipulate data objects in R				16:30
16:45					16:45
17:00					17:00
17:15					17:15
17:30	Finish	Finish	Finish	Finish	17:30



What worked…*and what didn't*

- It helped them focus on learning R
- They engaged better with the stats content
- *Some students who knew R felt bored for one day*
- *Some didn't bond with base R*

FIRST STEPS WITH R

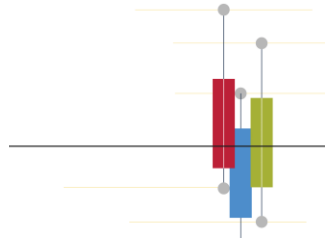
Mariagrazia Zottoli and Maria Christodoulou

Part 0 – R and RStudio

If you already have R and RStudio installed on your machines, feel free to skip to **Part I**. Below are instructions on where to find R, and RStudio and how to download and install them.

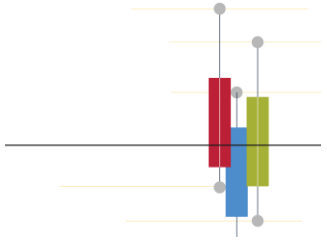
We will be working on RStudio. For this you will need a fully functioning version of R installed on your machines. To get a copy of R, visit:

https://www.r-project.org



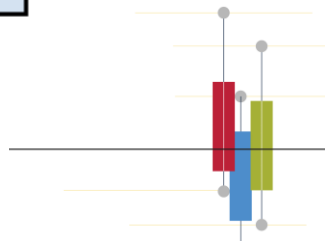
Did we decide to keep the changes?

- Yes!
  - Feedback was overwhelmingly positive
  - Our rates of students leaving early from practicals due to frustration dropped to one or two students for the entirety of the course
- But…
  - We got some sassy comments about leaving tidyverse too late
  - We should have prepared advanced R materials for those who knew the basics so they don’t get bored



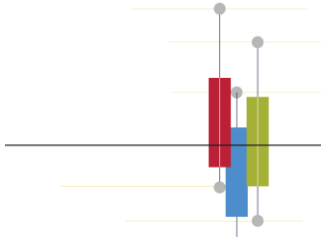
What if we started with tidyverse?

Week	Date	Oxford Term week	Topic
1	02/11/2023	MT4	Introduction to tidyverse and base R
2	09/11/2023	MT5	Advanced data handling using R
3	16/11/2023	MT6	Probability refresher and introduction to inference
4	23/11/2023	MT7	Hypothesis testing and introduction to statistical testing
5	30/11/2023	MT8	Statistical testing
6	07/12/2023	MT9	Correlations and introduction to linear models
7	14/12/2023	MT10	Linear models
8	11/01/2024	HT0	Linear model (interactions) and revision



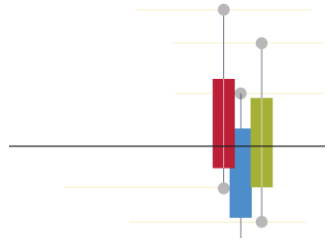
What worked…*and what didn't*

- The 5 verbs of dplyr are really intuitive for complete beginners
- Introducing base R after the 5 verbs actually got them to engage with a more positive attitude and appreciate it more
- *We still got questions about why we haven't started talking about stats*
- *We…still haven't made any advanced materials for the more experienced students*



Plans for the future

- Create more dynamic and interactive practicals, while maintaining the scaffolding
- Prepare resources for those more familiar with R, so they don’t get bored
- Setup some video resources for asynchronous learning to support complete beginners with revision



Thank you for your time!