



“ChatGPT for your data”, using Deepnote and other AI tools for data analysis



Data and Decision Science Network

Part of the UOW Data and Decision Science Initiative

Dr Colin Cortie, Graduate School of Medicine

Professor Marijka Batterham, UOW Statistical Consulting Centre



UNIVERSITY
OF WOLLONGONG
AUSTRALIA



Outline

"ChatGPT for your Data" using Deepnote & other AI tools for data analysis

- Introductions
- Data and Decision Science Network – why are we giving this talk?
- “Deepnote” – Dr Colin Cortie
- GitHub Copilot – AI pair programmer with autocomplete function in RStudio
- Hacky hour



We acknowledge that Country for Aboriginal peoples is an interconnected set of ancient and sophisticated relationships.

The University of Wollongong spreads across many interrelated Aboriginal Countries that are bound by this sacred landscape, and intimate relationship with that landscape since creation.

From Sydney to the Southern Highlands, to the South Coast.

From fresh water to bitter water to salt. From City to Urban to Rural.

The University Acknowledges the devastating impact of colonisation on our campuses' footprint and commit ourselves to truth-telling, healing and education.

Flame Tree artwork by Samantha Hill, Dharawal/Wandandian woman

Introductions

- Dr Colin Cortie
- Postdoctoral researcher in the Graduate School of Medicine
- Uses AI-written code to speed up data analysis and make beautiful graphs.
- When it comes to AI, he considers himself an enthusiastic early adopter rather than an expert.

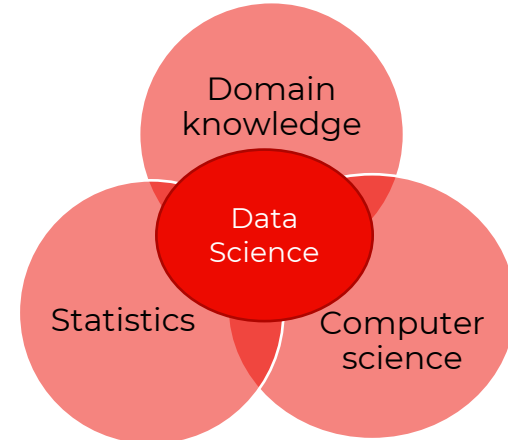


- Professor Marijka Batterham
- Co-Ordinator Data & Decision Science Initiative
- Director NIASRA
- Director Stats Consulting Centre
- Passionate about data literacy
- Use RStudio/SPSS most often
- Favourite analysis: logistic regression
- Mostly use: mixed models
- Like learning ML & exploring new packages
- Use ChatGPT4 to help with code, am now trialing github copilot.

UOW Data & Decision Science Initiative

- The Data and Decision Science Initiative is part of the UOW strategic Plan (2.5 Transformative technologies)
- Developed from a 2019 review and recommendations of “Big Data” and Health Informatics at UOW
- Commenced July 2021
- Led by NIASRA (Marijka Batterham Co-Ordinator)

Data Science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formation and hypothesis testing



Data & Decision Science Initiative

four key areas of focus

Research: virtual network and working groups of Data and Decision Science researchers

- Focal point for coordinating the development of Data Science at UOW
- Composed of researchers actively using or interested in Data Science methods
- Themed meetings emphasising translation: Data and Decision Science Network (DDSN)
- Strategically collaborations through the DDSI give a competitive advantage in translation

Education: Training in data science and reproducibility of research.

- Internal and external training and education in data science
- Upskilling research students & staff (particularly ECRs) in data & decision science methods
- Workshops (GRS, Statistical Consulting Centre)

T shaped graduates: Reviewing service subjects to refocus on data science.

- Review of service subjects in statistics and quantitative methods to give data science focus
- Graduates literate in data science and reproducible research

External/Industry engagement: Capitalising on existing links

- Provide enhanced opportunities for external engagement



Using AI for data analysis



Dr Colin Cortie

Graduate School of Medicine



UNIVERSITY
OF WOLLONGONG
AUSTRALIA




Why would we do this?

To make data analysis

- Easier
- Faster (*maybe*)
- More reproducible
- Because using AI is cool

Reproducible data analysis

PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS

EDITORIAL

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve , Anton Nekrutenko, James Taylor, Eivind Hovig

Published: October 24, 2013 • <https://doi.org/10.1371/journal.pcbi.1003285>

Rule 1: For Every Result, Keep Track of How It Was Produced

Rule 2: Avoid Manual Data Manipulation Steps

Rule 7: Always Store Raw Data behind Plots

Rule 10: Provide Public Access to Scripts, Runs, and Results

Why don't we do this?

- There is **no requirement** to do it

Reporting standards and availability of data, materials, code and protocols

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature Portfolio journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.** Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must also be disclosed in the submitted manuscript.

[nature.com/nature-portfolio/editorial-policies/reporting-standards](https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards)

Why don't we do this?

- There is **no requirement** to do it
- There is **no reward** for doing it – but its good science
- ~~The lack of transparency is intentional to **prevent data being checked / queried** – this is **bad form**~~

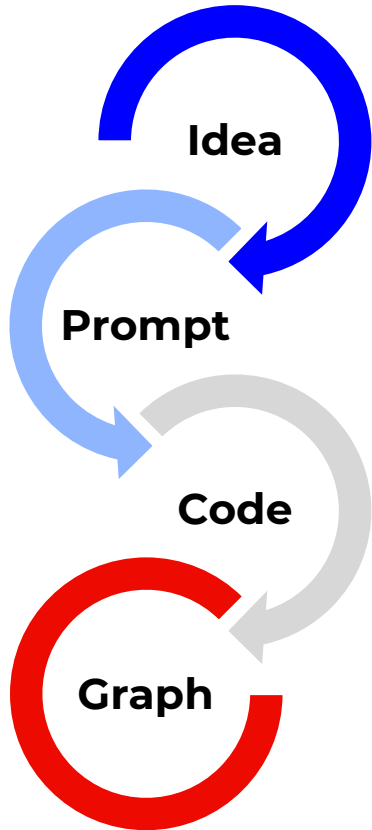
Why don't we do this?

- There is **no requirement** to do it
- There is **no reward** for doing it – but its good science
- Mostly... it can be hard – writing code takes **skill** and **time**

What if this was easy?



AI to the rescue?



- ChatGPT and other large language models can write code
- They can also edit and correct code
- They can clean data, calculate variables, make graphs etc.
- You don't need to know how to code (although it helps)

Would you like to know more?



Luke Barousse ✓

@LukeBarousse · 426K subscribers · 156 videos

What's up, Data Nerds! I'm Luke, a data analyst who is exploring how to use AI for analysis. >

lukebarousse.com and 4 more links

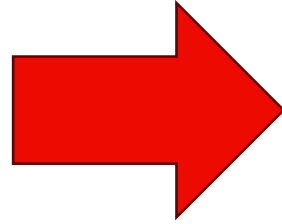
 Subscribed ▾

For you



Should you still learn to code? (ft. Devin)

38K views · 5 days ago



Typing becomes easier, but you still have to know what you want to say

Practical ethics of AI use (as of today)

We do have to **be careful**:

- Where is the data stored?
- Who has access to it?
- Is it sensitive data?

Ethically exempt data:

- Publicly available
- Not re-identifiable
- Not vulnerable populations
- Put in an application

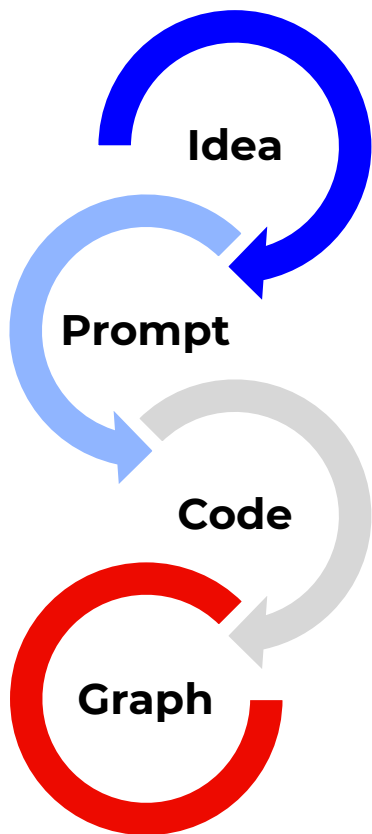
Blog

March 20 ChatGPT outage: Here's what happened

An update on our findings, the actions we've taken, and technical details of the bug.

We took ChatGPT offline earlier this week due to a bug in an open-source library which allowed some **users to see titles from another active user's chat history**. It's also possible that the first message of a newly-created conversation was visible in someone else's chat history if both users were active around the same time.

What should we be using?



	ChatGP	Deeptime
LLM	OpenAI	OpenAI
Runs	Python	Python, SQL
Easy to use	Very	Fairly
Records prompts	Yes	Yes
Records code	Yes	Yes
Allows edits to code	No	Yes
Share	No	Yes

Using prompts

How would you talk to a **research assistant**?

Prompts

- Clean the data
- Perform exploratory data analysis
- Why is that number so high, are you sure its right?
- Make some graphs for XXXX
- Make the graphs beautiful
- Make different graphs
- What stats would be best? Can you do them for me?



U

Live demo

O



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

W

what is GitHub?

Let's build from here

The complete developer platform to build, scale, and deliver secure software.

100+ million Developers 4+ million Organizations 420+ million Repositories 90% Fortune 100

What is GitHub Copilot?

- GitHub is an online software development platform. It helps users store and manage their code, as well as track and control changes to their code.
- Widely used to manage version control on research projects involving code and analysis.
- Public and private repositories (projects)
-

- A code completion tool developed by GitHub and OpenAI
- Trained on GitHub data
- Assists code completion in an Integrated Development Environment, Rstudio, Visual Studio.
- Autocompletes the code
- Starting to code can be intimidating
- Unless you are persistent hard to remember the exact commands
- GitHub can help by making suggestions “ghost text”
- It is often slow to get started
- Improves as you work through the analysis.
- \$10/month

Dataset









Research question

Sample dataset compares Body Mass Index BMI(kg/m²) between people with and without diabetes. Simulated from the Pima Indian Dataset. Dabelea et al. Journal of Maternal Fetal Medicine 2000;9:83-88.

BMI is continuous and reasonable to assume normally distributed,

Diabetes is categorical(binary) 0= no diabetes, 1=diabetes

Research question: Is there a significant difference in mean BMI between those with and those without diabetes?

	 npreg	 gluc	 bp	 skin	 ped	 age	 Diabetes	 BMI
1	1	4.94	24	19	.559	21	0	28.13
2	1	4.88	30	42	.496	26	1	45.22
3	1	5.72	30	38	.183	33	0	40.85
4	3	5.66	44	20	.400	26	0	31.26
5	6	5.99	44	20	.813	35	0	24.83
6	3	6.27	44	13	.140	22	0	23.20
7	1	6.60	44	47	.280	25	0	34.59

```
github copilot t test script.R x ChatGPT_t_test_and_boxplot.R x Untitled1 x
```

```
1
```

1:1 (Top Level)

R Script

Environment History Connections Tutorial

Import Dataset 103 MiB

R Global Environment

Environment is empty

Console Terminal Background Jobs

```
R 4.3.2 · ~/
```

```
R version 4.3.2 (2023-10-31 ucrt) -- "Eye Holes"  
Copyright (C) 2023 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
Natural language support but running in an English locale
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or
```

Files Plots Packages Help Viewer Presentation

Zoom Export

GitHub Copilot

what is it?

- Within R Studio, start typing and generative ideas presented in ghost text.

```
8 # perform a t test | to compare the mean age of the patients with diabetes and without diabetes
8 # perform a t test to compare bmi between those |with and without diabetes
9 t.test(bmi ~ diabetes, data = tdiabetes)
Error in eval(predvars, data, env) : object 'bmi' not found
9 t.test(BMI ~ Diabetes, data = tdiabetes)

welch Two Sample t-test
```

```
data: BMI by Diabetes
t = -7.5898, df = 240.09, p-value = 7.062e-13
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -5.851297 -3.439828
sample estimates:
mean in group 0 mean in group 1
    31.21093         35.85649
```


GitHub Copilot

Demonstration

```
12 # perform the t test assuming unequal variances
12 # perform the t test assuming equal variances
13 t.test(tdiabetes$BMI ~ tdiabetes$Diabetes, var.equal = TRUE)
```

Two sample t-test

```
data: tdiabetes$BMI by tdiabetes$Diabetes
```

```
t = -7.2715, df = 330, p-value = 2.599e-12
```

```
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
95 percent confidence interval:
```

```
-5.902336 -3.388790
```

```
sample estimates:
```

```
mean in group 0 mean in group 1
      31.21093      35.85649
```

```
14 # perform a test to determine if the variances are equal
15 var.test(tdiabetes$BMI ~ tdiabetes$Diabetes)
```

F test to compare two variances

```
data: tdiabetes$BMI by tdiabetes$Diabetes
```

```
F = 1.2839, num df = 222, denom df = 108, p-value = 0.1438
```

```
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
```

```
0.9176251 1.7632982
```

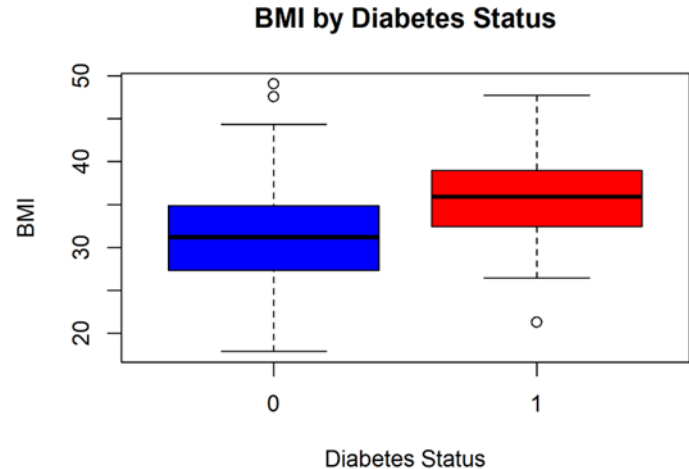
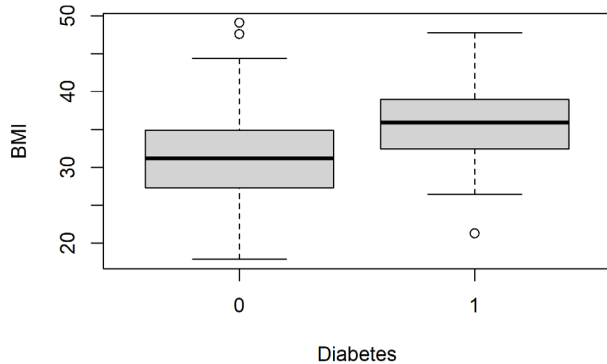
```
sample estimates:
```

```
ratio of variances
      1.283936
```

GitHub Copilot

Demonstration

```
16 # plot a side by side boxplot of bmi by diabetes status
17 boxplot(BMI ~ Diabetes, data = tdiabetes)
18 # make the boxplot beautiful
19 boxplot(BMI ~ Diabetes, data = tdiabetes, col = c("blue", "red"),
20         xlab = "Diabetes Status", ylab = "BMI", main = "BMI by Diabetes Status")
```



```

3 # load the diabetes dataset from the working directory
4 library(readr)
5 tdiabetes <- read_csv("ttestdiabetes.csv")
6 # show the dataset
7 View(tdiabetes)
8 # perform a t test to compare bmi between those with and without diabetes
9 t.test(BMI ~ Diabetes, data = tdiabetes)
10 # perform the t test assuming equal variances
11 t.test(tdiabetes$BMI ~ tdiabetes$Diabetes, var.equal = TRUE)
12 # perform a test to determine if the variances are equal
13 var.test(tdiabetes$BMI ~ tdiabetes$Diabetes)
14 # plot a side by side boxplot of bmi by diabetes status
15 boxplot(BMI ~ Diabetes, data = tdiabetes)
16 # make the boxplot beautiful
17 boxplot(BMI ~ Diabetes, data = tdiabetes, col = c("blue", "red"),
18         xlab = "Diabetes Status", ylab = "BMI", main = "BMI by Diabetes Status")
19:1 (Top Level)
  
```

Console Terminal Background Jobs

R 4.3.2 · C:/Users/marijka/OneDrive - University of Wollongong/ChatGPT/github copilot course/

```

> t.test(tdiabetes$BMI ~ tdiabetes$Diabetes, var.equal = TRUE)

Two Sample t-test

data: tdiabetes$BMI by tdiabetes$Diabetes
t = -7.2715, df = 330, p-value = 2.599e-12
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-5.902336 -3.388790
sample estimates:
mean in group 0 mean in group 1
31.21093 35.85649
  
```

Environment History Connections Tutorial

Import Dataset 255 MiB

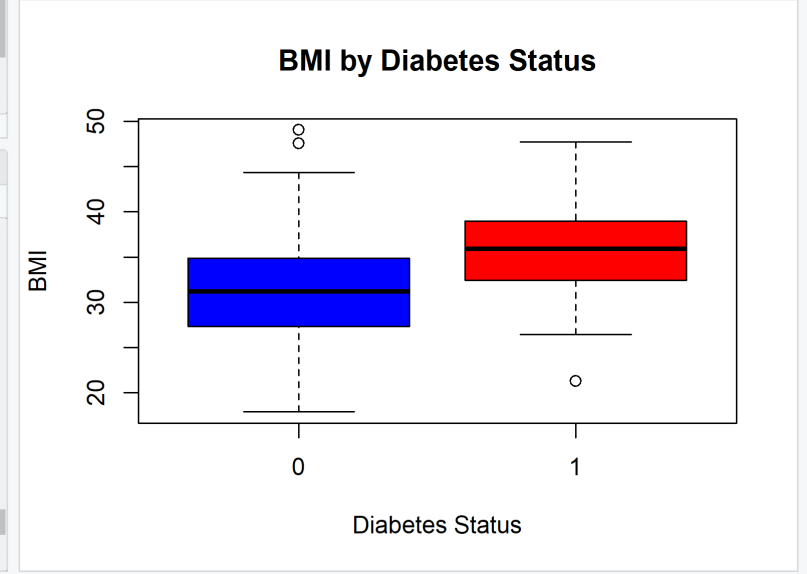
R Global Environment

Data

tdiabetes	332 obs. of 8 variables
-----------	-------------------------

Files Plots Packages Help Viewer Presentation

Zoom Export Publish



Chat GPT4 Advanced Data Analysis

Demonstration

ChatGPT 4 ▾

MA You



perform a t test to determine if there is a difference in bmi between those with and those without diabetes

ChatGPT

The dataset contains several columns, with `BMI` representing the body mass index of individuals, and `Diabetes` indicating whether the person has diabetes (1) or not (0). To determine if there's a difference in BMI between those with and without diabetes, I'll perform a t-test on these two groups. Let's proceed with the analysis. [-]

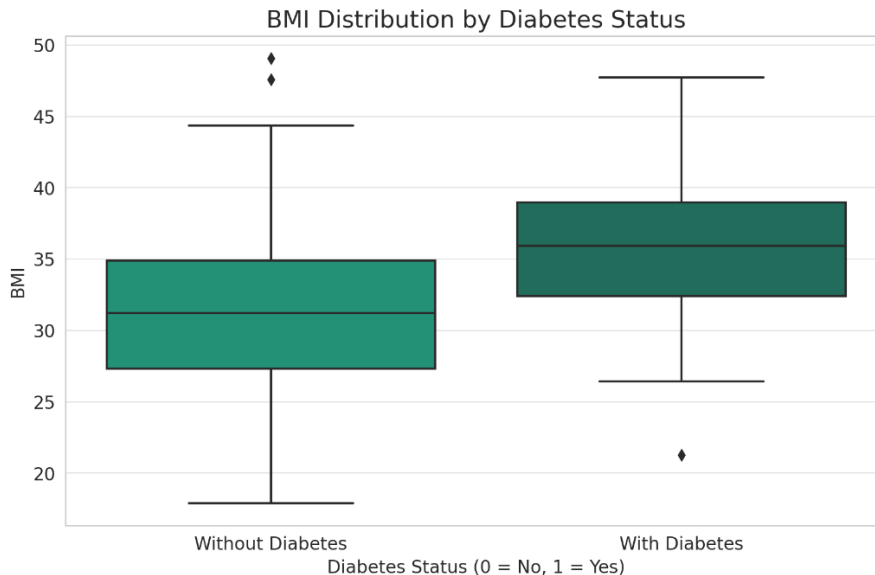
The t-test resulted in a t-statistic of 7.27 and a p-value of approximately 2.6×10^{-12} . Given the very small p-value, we can reject the null hypothesis that there's no difference in BMI between those with and without diabetes. This indicates a statistically significant difference in BMI between the two groups. [-]

ChatGPT4

Demonstration

You
present these results as a side by side boxplot

ChatGPT



Here's a side-by-side boxplot comparing the BMI distribution of individuals with and without diabetes. The boxplot illustrates the central tendency, spread, and outliers for each group, visually confirming the statistical analysis findings that there's a noticeable difference in BMI between those with diabetes and those without. [\[↗\]](#)

You
output the code used to perform the t test and generate the plot in an r script
[↗](#)

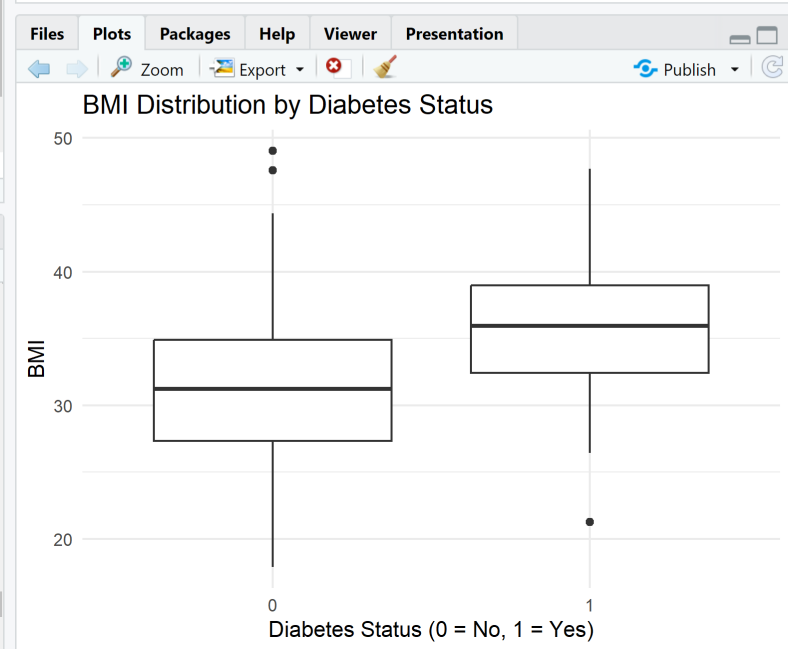
ChatGPT
I've created the R script that includes the code to perform the t-test and generate the boxplot. You can download it using the link below:

[Download R Script \[↗\]](#)

```
RStudio  
File Edit Code View Plots Session Build Debug Profile Tools Help  
Go to file/function Addins  
o_copilot.R x diabetes copilot trial script.R x github copilot t test script.R x ChatGPT_t_test_and_boxplot.R* x  
1  
2 # Load necessary libraries  
3 library(ggplot2)  
4 # Perform t-test  
5 t_test_result <- t.test(BMI ~ Diabetes, data = tdiabetes)  
6 print(t_test_result)  
7 # Generate boxplot  
8 p <- ggplot(tdiabetes, aes(factor(Diabetes), BMI)) +  
9   geom_boxplot() +  
10  labs(title = "BMI Distribution by Diabetes Status", x = "Diabetes Status (0 = No,  
11  theme_minimal())  
12 p  
13 # Save the plot as PDF  
14 ggsave("BMI_Distribution_by_Diabetes_Status.pdf", plot = p, width = 10, height = 6)  
15  
16  
15:1 (Top Level) Copilot: Completion response received. R Script
```

```
Console Terminal x Background Jobs x  
R 4.3.2 · C:/Users/marijka/OneDrive - University of Wollongong/ChatGPT/github copilot course/  
> library(ggplot2)  
> t_test_result <- t.test(BMI ~ Diabetes, data = tdiabetes)  
> print(t_test_result)  
  
welch Two sample t-test  
  
data: BMI by Diabetes  
t = -7.5898, df = 240.09, p-value = 7.062e-13  
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
95 percent confidence interval:  
-5.851297 -3.439828  
sample estimates:  
mean in group 0 mean in group 1  
31.21093 35.85649
```

Environment History Connections Tutorial
Import Dataset 24 MiB
R Global Environment
Data
p List of 9
t_test_result List of 10
tdiabetes 332 obs. of 8 variables





Limitations

It's all about the data and the prompts – garbage in garbage out

- Although generative AI makes assumptions about your data it relies on the information you give it about the context and structure (eg multilevel data, time series data)
- If you ask it to run a test it will, consider the prompts used to make sure the assumptions are checked. For the t test, we could ask if the data meets the assumption of the t test and **we should always explore the data first.**

Using your own data

Beware

- Do not upload sensitive or identifiable data to third party applications
- Consider obtaining consent when you develop your ethics submission if you intend to use generative AI to analyse your data
- The Ethics committee will consider the use of generative AI on an individual basis

- Ideally, we will have an Enterprise solution for the University where we can use this technology in a secure environment.

Introducing Hacky Hour

Co presentation with Molecular Horizons

- Launch 1130-1330 Thursday the 2nd May
- Dr Dezeræe Cox, Lecturer in School of Chemistry & Molecular Bioscience, Molecular Horizons
- “a self-taught computer programming enthusiast”
- Works in Python

- Introduction of Hacky Hours, fortnightly drop in sessions to receive help and support with programming in Python and R (Brad Wakefield)
- Stay tuned to Universe for more information or contact us directly marijka@uow.edu.au bradleyw@uow.edu.au

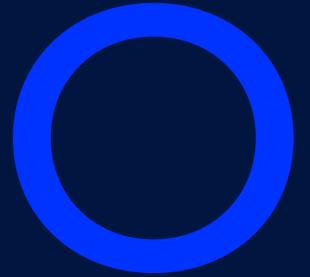
- Visit the Data and Decision Science Initiative website for presentations and slides <https://www.uow.edu.au/niasra/data-and-decision-science-initiative/>

The Data Science and Statistics CoP

INFORMATION SLIDES 2024



UNIVERSITY
OF WOLLONGONG
AUSTRALIA



The

Data Science and Statistics Community of Practice

An online space to:

- Foster Knowledge Sharing
- Access Resources
- Collaborate with Peers.
- Get Data Science and Stats Support.
- Learn about training opportunities.



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

A screenshot of a Microsoft Teams channel page. The channel is named "General" and has tabs for "Posts", "Files", "Home", and "1 more". The page shows a "WELCOME TO" message for the "Data Science and Statistics Community of Practice (CoP)". Below the message is a large image of a person sitting at a desk with a computer monitor displaying a data analysis dashboard. The dashboard includes a bar chart, a line graph, and several circular progress indicators. A "Learn more" link with a right-pointing arrow is located below the image. The left sidebar of the Teams interface is visible, showing icons for Activity, Chat, Teams, Calls, Calendar, Files, Bookings, and Apps.

Contact: bradleyw@uow.edu.au for link



OF WOLLONGONG
AUSTRALIA



The Statistical Consulting Centre

INFORMATION SLIDES 2024



UNIVERSITY
OF WOLLONGONG
AUSTRALIA



A BIT ABOUT US ...

Marijka Batterham

- **Director**

Aim

The service aims to improve the statistical content of research carried out by members of the University. Researchers from all disciplines may use the Centre. Priority is currently given to staff members and postgraduate students undertaking research for Doctor of Philosophy or Masters' degrees.

How we can help

Currently the Statistical Consulting Centre provides each academic or post-graduate student with a free initial consultation. Up to ten hours per calendar year of consulting time is provided without charge if research funding is not available. When researchers require more consulting time, or receive external funding, a service charge may be necessary.



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Brad Wakefield

- **Statistical Consultant**

To book an appointment, simply go to our website and select

<https://www.uow.edu.au/niasra/our-research/statistical-consulting-centre/>

MAKE AN APPOINTMENT



UNIVERSITY
OF WOLLONGONG
AUSTRALIA