STATS 369: Assignment 1

20 points

Due 4 August by 23:59

Instructions

- **Submission requirements**: You must submit both the HTML and Rmd with your solutions. It is suggested that you use the template provided for labs and assignments.
    - Marks will be lost for poorly organised submissions — see Penalties to avoid. Use headings for each task and clearly indicate your answers to subtasks, as appropriate.
    - Comment code in the code chunks at a reasonable level so that another person with some R familiarity could easily follow your thinking and process. This doesn't mean every single line must be commented, but the overall **purpose** should be clear.
    - All code should be shown in your HTML (i.e., don't hide any code), but messages from loading packages and data should be suppressed. `message = F` and `warning = F` are useful for your libraries chunk.
    - Your first code chunk should load all your libraries (don't include libraries you're not using).
    - There should be no `install.packages()` code in your submission. Package installs can be done in the console and should not be run every time you knit.
- **Late submissions**: Late submissions are accepted for *up to three days* with a 10 percentage point-per-day penalty (pro-rated to an hourly penalty of 0.42 percentage points per hour). Make sure you submit BEFORE 23:59. A 1 hour late penalty may apply to submissions that are processed by Canvas at exactly the deadline.
    - Note: You can submit as many times as you like before the deadline.
- **Allowed libraries**: Students come to this course with different coding backgrounds. To help keep this course fair, you will be asked to generally rely on the packages and functions we use in class and labs to answer code questions in the main questions in these assignments. (Bonus questions, are fair game for showing-off and Googling).
    - Tidyverse: For example, as we have a focus on Tidyverse, you are expected to use Tidyverse to achieve the wrangling and visualisation solutions.
    - Other packages: If there are other packages you like to use and their use won't detract from our ability to assess your knowledge of the course content, you CAN use them, but please ensure you explain what they are for (comments on the code are appropriate). Doing a good job of this helps the markers see that you understand the code and have been thoughtful about it, and that you haven't just copied something from StackOverflow or ChatGPT without understanding what the course is about.
- Use a **referencing style** of your choice to reference relevant resources used as you work on your assignment. Class slides do not need to be referenced. There should be paired in-text citations and a references section at the end.

Task 1: Revision [3 marks]

From the Course Outline (https://courseoutline.auckland.ac.nz/dco/course/STATS/369/1235) the pre-requisites for this class are: STATS 220 and STATS 210 or 225 and 15 points from ECON 221, STATS 201, 208, or ENGSCI 314.

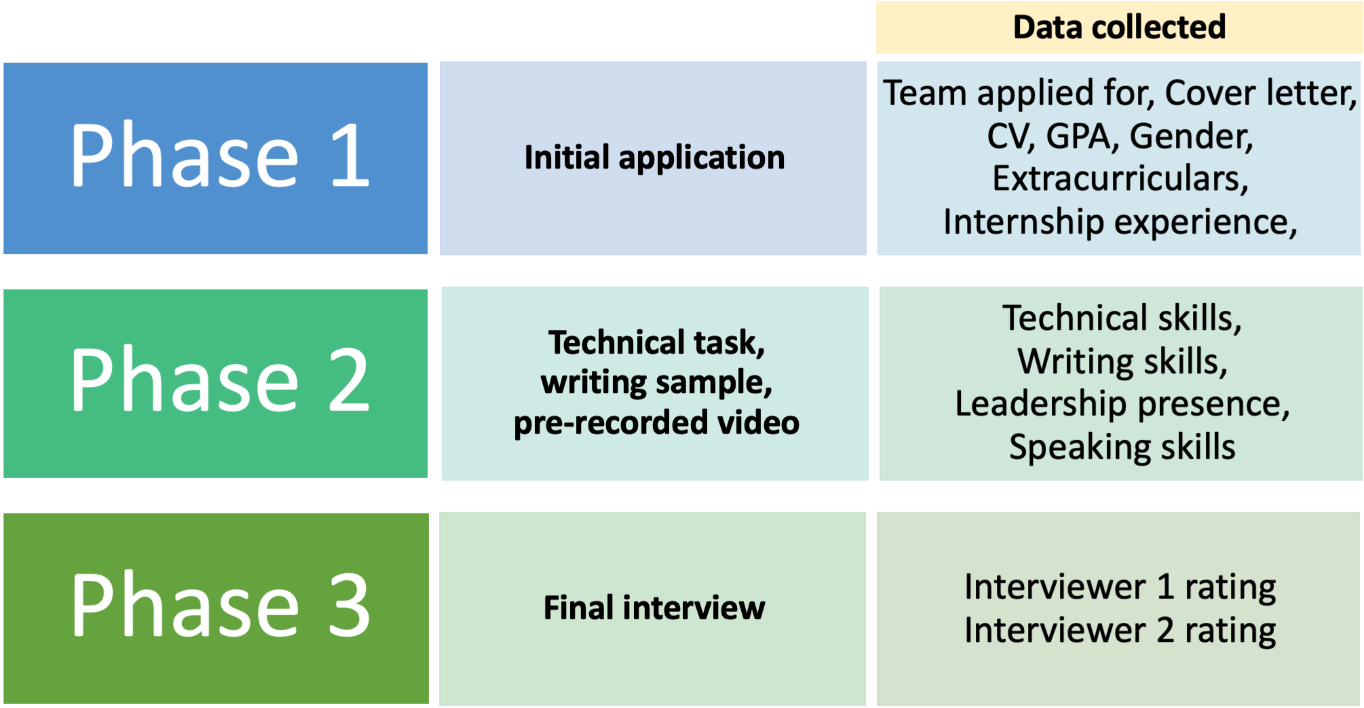
Choose one of these courses (presumably one you’ve taken) and write one multichoice question on a topic from that course. The audience should be your peers in this course who are aiming to revise topics from the prerequisites. *If I have time, I will make some of these into a practice quiz on Canvas. You can just write a note if you don’t want your question considered for inclusion, it won’t affect your mark.*

- Make it clear for which **course** this is revision.
- Your question should have **4 options** (1 correct and 3 distractors)
- Write an **answer key** that explains which answer is correct and why the others are wrong.

You will be marked on the correctness and quality of your question and explanation. The question does not have to be *hard*, per se, but should be **USEFUL** to you and your fellow students.

Task 2: Algorithmic fairness audit (mini-consulting project) [11 marks]

A company called Black Saber[1] has been trialling a new AI recruitment pipeline manager for their Data and Software teams. There are three phases, outlined below, each narrowing down the field of applicants. Based on advice from their legal team, they are not able to provide you with the original application data, but they can provide these anonymised indicators/ratings from each phase. `applicant_id` is consistent across phases.



		Data collected
Phase 1	Initial application	Team applied for, Cover letter, CV, GPA, Gender, Extracurriculars, Internship experience,
Phase 2	Technical task, writing sample, pre-recorded video	Technical skills, Writing skills, Leadership presence, Speaking skills
Phase 3	Final interview	Interviewer 1 rating Interviewer 2 rating

Data explanation

Phase 1

`phase1-new-grad-applicants-2022.csv`

In the first phase of the hiring pipeline applicants complete a form and are asked to submit a CV and cover letter. Extracurriculars and internship experience are auto-rated based on the descriptions applicants provide in the application form.

Variable	Description
`applicant_id`	A unique ID assigned to applicants in Phase 1
`team_applied_for`	Software or Data
`cover_letter`	0 if absent, 1 if present
`cv`	0 if absent, 1 if present
`gpa`	0.0 to 4.0 (American style)
`gender`	Gender of employee: ‘Man’, ‘Woman’, ‘Prefer not to say’ only options provided

Variable	Description
`extracurriculars`	The description of extracurricular involvement is assessed against a proprietary key term and phrase bank and given a 0, 1 or 2 for where 2 indicates several high relevance and/or skills building extracurriculars, 1 indicates some relevant and/or skills building extracurriculars and 0 indicates no extracurriculars describes or that those describe were not rated as high relevance or high skills building
`work_experience`	Similar to `extracurriculars`, the description applicants provided is assessed against a proprietary key term and phrase bank, that also considers company names and reputations, to give a 0, 1 or 2 score, with 2 being the best, 0 the worst

Phase 2

`phase2-new-grad-applicants-2022.csv`

We don't know exactly how these are being assessed by the AI, the algorithm is commercially sensitive but their demonstrations of the system were impressive.

Variable	Description
`applicant_id`	A unique ID assigned to applicants in Phase 1
`technical_skills`	Score from 0 to 100 on a timed technical task, AI autograded
`writing_skills`	Score from 0 to 100 on a timed writing task, AI autograded
`speaking_skills`	A rating of speaking ability based on pre-recorded video, AI autograded
`leadership_presence`	A rating of 'leadership presence' based on pre-recorded video, AI autograded

Phase 3

`phase3-new-grad-applicants-2022.csv`

This is the information from interview phase. Being listed as 'first' or 'second' interviewer is arbitrary and who the interviewers were is not available from our tracking system. Applicant IDs are listed across the top and then the two scores for the applicant are listed below their ID.

The average score of the two interviewers was used to determine final hires.

Final hires

`final-hires-newgrad_2022.csv`

This data set contains the applicant IDs of everyone who was sent an offer letter. In this cohort, everyone accepted.

Variable	Description
`applicant_id`	A unique ID assigned to applicants in Phase 1

Subtasks

1. Load, wrangle and join these datasets in to ONE appropriate tidy dataset where each applicant is an observation. Create 3 indicator variables for whether or not each applicant passed a given phase (e.g., final hires *passed* phase 3). These indicators can be 0/1 numerics or have text levels — up to you. Show the head (`head()`) of your finished dataset's first **10 rows** (do *not* print the whole thing!) [4 marks]
2. Create appropriate numeric summaries, basic statistical tests (think t-test or ANOVA/F-test) and at least ONE appropriate chart to explore whether their are any concerns about this new AI recruitment pipeline.[2] [4 marks]

3. Write ‘alt text’ for the chart you created above based on the advice from *Writing Alt Text for Data Visualization* by Amy Cesal (https://nightingaledvs.com/writing-alt-text-for-data-visualization/). If you created more than one chart, pick one to write about. Assume this alt text would also be geared towards an audience of Black Saber Executives — you do not need to link the data as it is their data, but it would be worth being clear about. [1 mark]
4. Write a short[3] conclusion for Black Saber on whether their new AI applicant processing system appears to be working well. Highlight any potential risks to your client. Address your comments to an audience of executives who may not have much statistical or data knowledge. [2 marks]

Bonus opportunity [+1 bonus]

Create the UGLIEST version of the graph you created about that you can. Dancing T-rex as the background or colours that make your eyes bleed? Go for it! Note: You are free to use any functions you can find, you don’t have to stick as closely to course content.

Task 3: Reprex critique [3 marks]

Consider the following code. It scrapes the text of a poem about statistics from a webpage.[4]

Original code

```
html_text(read_html("https://link.lizabolton.com/a_scrapable_poem.html"))
```

```
#> [1] "\n  In this vast age of data's endless stream, A science blooms with wonders to beho
ld, Where bits and bytes converge in seamless theme, Unveiling truths that were once left un
told.\n\nWith algorithms and models as our guide,\nWe journey through the realms of structur
ed lore,\nEach data point a star to be untied,\nTo find the patterns hidden deep in core.\nT
hrough clustering, we sort and classify,\nRegression leads us to predictive might,\nIn neura
l networks, connections amplify,\nEmerging knowledge, dazzling and bright.\nOh, data scienc
e, thou art a beacon rare,\nIlluminating paths to futures fair.\n\n    Written by ChatGPT -
The AI Poet | 2023\nI would like to make it clear that I take no responsibility for any crim
es against poetry committed here. - Liza\n\n"
```

```
html_text2(read_html("https://link.lizabolton.com/a_scrapable_poem.html"))
```

```
#> [1] "Inthisvastageofdata'sendlessstream,\nAsciencebloomswithwonderstobehold,\nWherebitsan
dbytesconvergeinseamlesstheme,\nUnveilingtruthsthatwereonceleftuntold.\n\nWith algorithms an
d models as our guide,\n\nWe journey through the realms of structured lore,\n\nEach data poi
nt a star to be untied,\n\nTo find the patterns hidden deep in core.\n\n\n\n\nThrough cluste
ring, we sort and classify,\n\nRegression leads us to predictive might,\n\nIn neural network
s, connections amplify,\n\nEmerging knowledge, dazzling and bright.\n\n\n\n\nOh, data scienc
e, thou art a beacon rare,\n\nIlluminating paths to futures fair.\n\n\n\n\nWritten by ChatGP
T - The AI Poet | 2023\n\n\n\n\nI would like to make it clear that I take no responsibility
for any crimes against poetry committed here. - Liza"
```

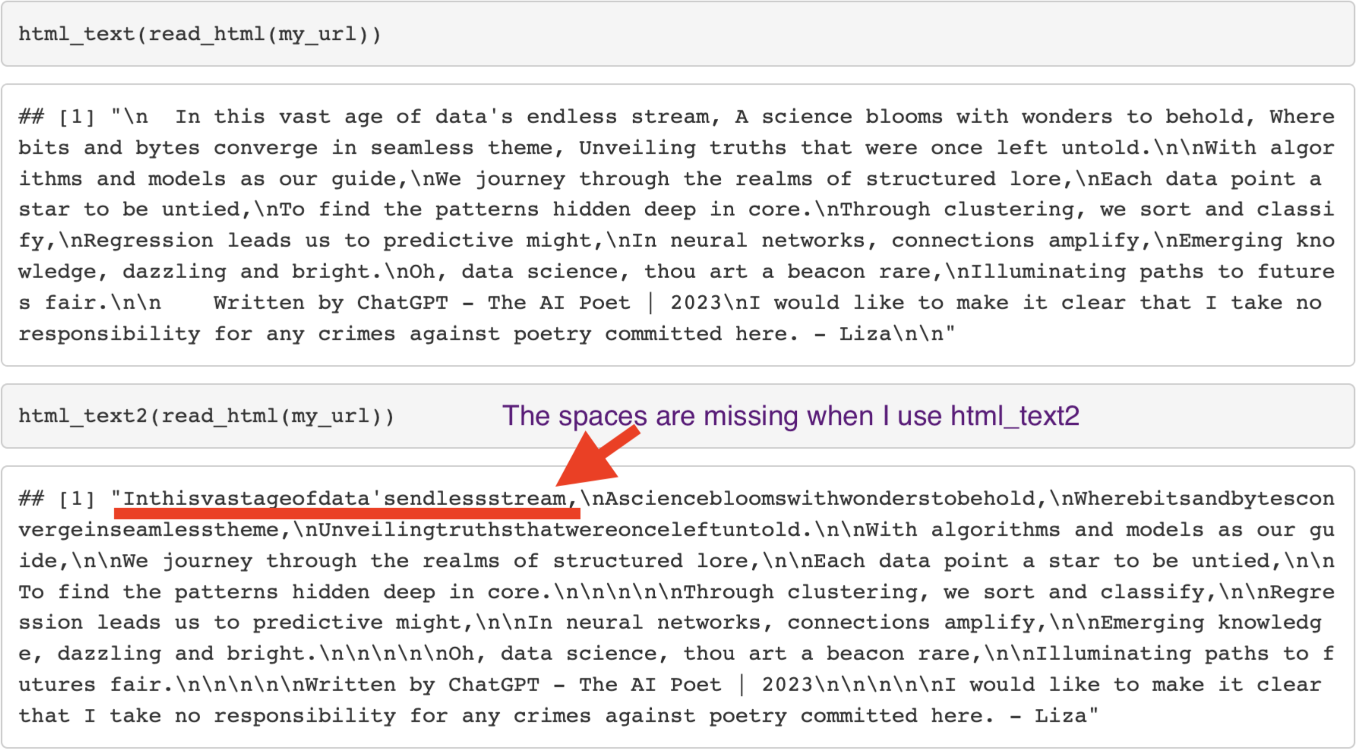
This works as expected when scraping using `html_text`, but has a problem when using `html_text2`. These are both functions from the `rvest` package (Wickham 2022) and `html_text2` provides more nicely formatted outputs, i.e., according to the help text: “html_text2() simulates how text looks in a browser, using an approach inspired by JavaScript’s innerText(). Roughly speaking, it converts <br /> to ``\n”, adds blank lines around <p> tags, and lightly formats tabular data.”

Our problem is that when using `html_text2`, some of the spaces are dropped and the words are all smushed together as part of this reformatting.

Suppose three students have each created an example to report this potential bug to the `rvest` development team. Using the article on reprex dos and don’ts (Bryan et al. 2022) and broader information about the reprex philosophy, choose THREE things to compare and contrast these three samples on.

Note: You do NOT need to be able to read the HTML to answer this question.

Bug report example A



```
html_text(read_html(my_url))
```

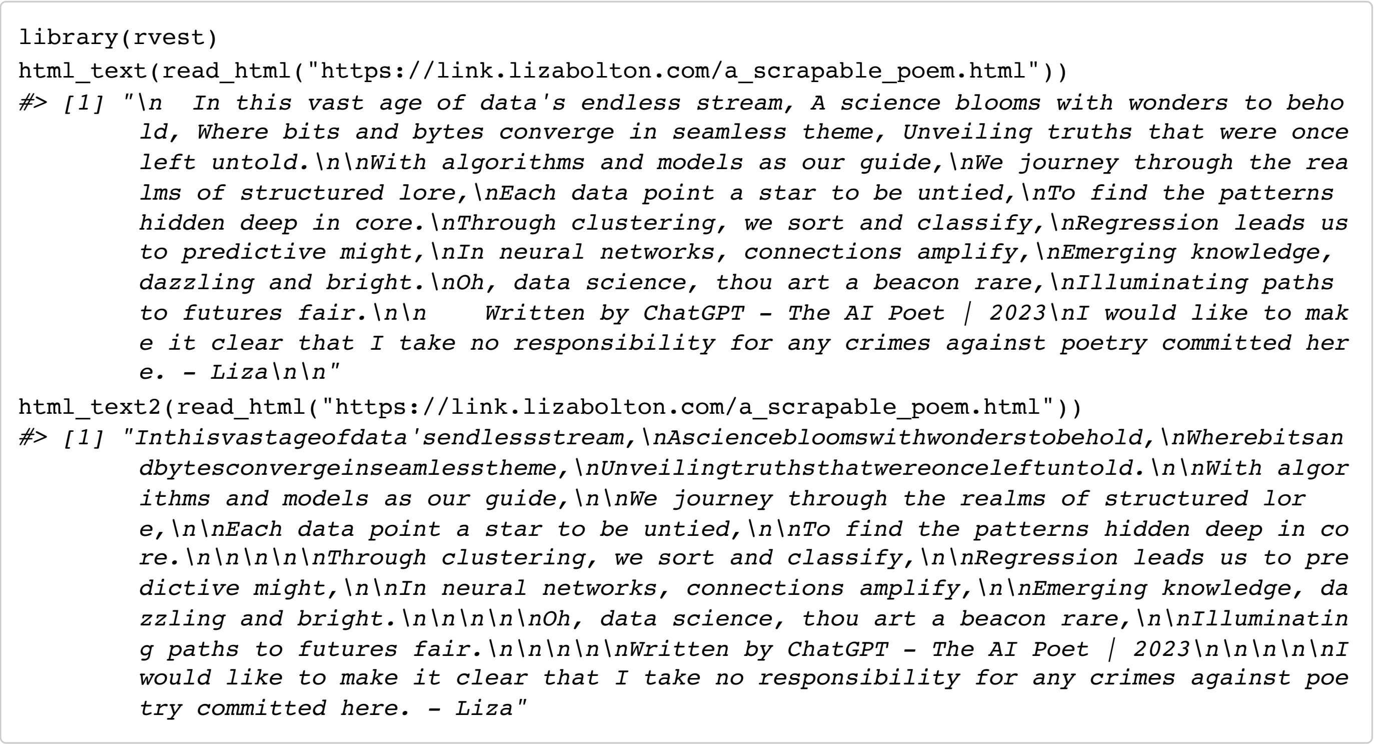
```
## [1] "\n  In this vast age of data's endless stream, A science blooms with wonders to behold, Where
bits and bytes converge in seamless theme, Unveiling truths that were once left untold.\n\nWith algor
ithms and models as our guide,\nWe journey through the realms of structured lore,\nEach data point a
star to be untied,\nTo find the patterns hidden deep in core.\nThrough clustering, we sort and classi
fy,\nRegression leads us to predictive might,\nIn neural networks, connections amplify,\nEmerging kno
wledge, dazzling and bright.\nOh, data science, thou art a beacon rare,\nIlluminating paths to future
s fair.\n\n    Written by ChatGPT - The AI Poet | 2023\nI would like to make it clear that I take no
responsibility for any crimes against poetry committed here. - Liza\n\n"
```

```
html_text2(read_html(my_url))
```

The spaces are missing when I use html_text2

```
## [1] "Inthisvastageofdata'sendlessstream,\nAsciencebloomswithwonderstobehold,\nWherebitsandbytescon
vergeinseamlesstheme,\nUnveilingtruthsthatwereonceleftuntold.\n\nWith algorithms and models as our gu
ide,\n\nWe journey through the realms of structured lore,\n\nEach data point a star to be untied,\n\n
To find the patterns hidden deep in core.\n\n\n\n\nThrough clustering, we sort and classify,\n\nRegre
ssion leads us to predictive might,\n\nIn neural networks, connections amplify,\n\nEmerging knowledg
e, dazzling and bright.\n\n\n\n\nOh, data science, thou art a beacon rare,\n\nIlluminating paths to f
utures fair.\n\n\n\n\nWritten by ChatGPT - The AI Poet | 2023\n\n\n\n\nI would like to make it clear
that I take no responsibility for any crimes against poetry committed here. - Liza"
```

Bug report example B



```
library(rvest)
html_text(read_html("https://link.lizabolton.com/a_scrapable_poem.html"))
#> [1] "\n  In this vast age of data's endless stream, A science blooms with wonders to beho
        ld, Where bits and bytes converge in seamless theme, Unveiling truths that were once
        left untold.\n\nWith algorithms and models as our guide,\nWe journey through the rea
        lms of structured lore,\nEach data point a star to be untied,\nTo find the patterns
        hidden deep in core.\nThrough clustering, we sort and classify,\nRegression leads us
        to predictive might,\nIn neural networks, connections amplify,\nEmerging knowledge,
        dazzling and bright.\nOh, data science, thou art a beacon rare,\nIlluminating paths
        to futures fair.\n\n    Written by ChatGPT - The AI Poet | 2023\nI would like to mak
        e it clear that I take no responsibility for any crimes against poetry committed her
        e. - Liza\n\n"
html_text2(read_html("https://link.lizabolton.com/a_scrapable_poem.html"))
#> [1] "Inthisvastageofdata'sendlessstream,\nAsciencebloomswithwonderstobehold,\nWherebitsan
        dbytesconvergeinseamlesstheme,\nUnveilingtruthsthatwereonceleftuntold.\n\nWith algor
        ithms and models as our guide,\n\nWe journey through the realms of structured lor
        e,\n\nEach data point a star to be untied,\n\nTo find the patterns hidden deep in co
        re.\n\n\n\n\nThrough clustering, we sort and classify,\n\nRegression leads us to pre
        dictive might,\n\nIn neural networks, connections amplify,\n\nEmerging knowledge, da
        zzling and bright.\n\n\n\n\nOh, data science, thou art a beacon rare,\n\nIlluminatin
        g paths to futures fair.\n\n\n\n\nWritten by ChatGPT - The AI Poet | 2023\n\n\n\n\nI
        would like to make it clear that I take no responsibility for any crimes against poe
        try committed here. - Liza"
```

Created with reprex v2.0.2 (https://reprex.tidyverse.org)

Bug report example C

```
library(rvest)
some_html <- '<p dir="ltr" style="text-align:left;"></p><span style="font-size:0.9375rem;">T
        he sentence starts this way,</span><span style="font-size:0.9375rem;"> </span><span
        style="font-size:0.9375rem;">then</span><span style="font-size:0.9375rem;"> </span><
        span style="font-size:0.9375rem;">spaces</span><span style="font-size:0.9375rem;">
        </span><span style="font-size:0.9375rem;">disappear</span>'
html_text(read_html(some_html)) # is correct
#> [1] "The sentence starts this way, then spaces disappear"
html_text2(read_html(some_html))  # not correct
#> [1] "The sentence starts this way,thenspacesdisappear"
```

Created with reprex v2.0.2 (https://reprex.tidyverse.org)

Task 4: Reflection [3 marks]

1. Read over the current graduate capability themes, LEVEL 2: Graduate Capabilities – Themes (https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/graduate-profile.html) or the refreshed version that connects to Taumata Teitei (https://www.auckland.ac.nz/en/on-campus/life-on-campus/latest-student-news/curriculum-framework-transformation-programme0/university-graduate&~_). Choose one specific task or subtask in this assignment or in lab 01 or lab 02 and discuss how your work demonstrates ONE of these specific capabilities. Make sure you explain the capability in your own words as well as referencing the graduate profile. Assume your audience is the HR team at a potential employer for the Data Science job of your dreams. Write ~100 to 300 words. [2 marks]
2. What is something you're proud of in the assignment? [0.5 marks]
3. What is something you're going to do differently for the next assignment? [0.5 marks]

⚠️⚠️⚠️ Penalties to avoid ⚠️⚠️⚠️

- **-1 mark** for not showing (don't `echo = F`!) a setup chunk at/near the beginning of your submission. It should include all the required libraries and suppress package loading messages and not have any `install.packages()` commands.
- **-1 mark** for missing/incomplete references. Make sure you have both an in-text citation and then also the association full reference in your references section. UoA has a resource called QuickCite (https://www.cite.auckland.ac.nz/2.html) to help you.
- **-2 marks** for not uploading one of the required files (HTML or Rmd).
- Up to **-2 marks** for poor formatting (unless extreme issues).
- Up to **-1 marks** for insufficient commenting of code.

References

Remember to include references if you use AI, and you should reference the Graduate Capabilities document use use in the reflection and the 'dos and don'ts' article.

Bryan, Jenny, Jim Hester, David Robinson, Hadley Wickham, and Cristophe Dervieux. 2022. *Reprex Dos and Don'ts*. https://reprex.tidyverse.org/articles/reprex-dos-and-donts.html (https://reprex.tidyverse.org/articles/reprex-dos-and-donts.html).

Wickham, Hadley. 2022. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest (https://CRAN.R-project.org/package=rvest).

[1] This isn't a real company↩

[2] Hint: The interview scores aren't based on AI. Of the previous phases, consider which of these parts of the pipeline might be most impacted by potential bias in the training data. E.g., GPA is just being read from the form so probably doesn't have bias issues. You might be interested to know that some studies suggest people (specifically American voters, but may be more generalisable) prefer leaders with lower-pitched voices (https://doi.org/10.1371/journal.pone.0133779 (https://doi.org/10.1371/journal.pone.0133779)) and that Amazon had to scrap it's AI recruitment system due to bias (https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10 (https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10))↩

[3] Approximately 100 to 300 words↩

[4] We should always consider the ethics of web scraping. In this case, our target is *my* site, and I've set it up for you to scrape so we don't have to do any other work.↩