Project Work Document  
Author: Tony Nguyen

# Ask

What topic are you exploring?  
*Global university rankings and performance in research and teaching.*

What is the problem you are trying to solve?  
*In what aspects could Australian universities improve their performance and in what way?*

What metrics will you use to measure your data to achieve your objective?  
*Teaching Score, Research Score, Citations Score, Industry Income Score, Intl Outlook Score, number of students, student: staff ratio, and % international students*

Who are the stakeholders?

*The Australian Federal Government Department of Education, Universities Australia, Australian university staff and students.*

Who is your audience?

*Potential employers*

How can your insights help your client make decisions?

*It can help show how our universities can be changed to become more internationally competitive and performing and how public funding can be better focused to achieve these objectives.*

**Deliverable**: A clear statement of the business task you have selected to investigate

***Investigate the rankings of Australian universities and determine what impact their staff: student ratio and current student enrolment numbers impact their ability to teach and research. In addition, compare the average rankings & scores of Australian universities with their international counterparts to determine their international competitiveness.***

# Prepare

Where is your data located?  
*The dataset for Times Higher Education World University Rankings 2021 were downloaded from* [*Kaggle*](https://www.kaggle.com/datasets/matheusgratz/world-university-rankings-2021?select=universities_ranking.csv)*.*

How is the data organized?  
*The scores and rankings of universities are stored in two separate csv files, sorted by ranking.*

Are there issues with bias or credibility in this data? Does your data ROCCC?  
*No, the data is collated by Times Higher Education which is recognized as one of the leading publications on university rankings.*

How are you addressing licensing, privacy, security, and accessibility?  
*The author(MATHEUS GRATZ) who uploaded the dataset I used on Kaggle has placed it under a CC0: Public Domain license, and Times Higher Education gives permission to use their intellectual property under their* [*terms and conditions*](https://www.timeshighereducation.com/terms-and-conditions)*. There is inherently no PII located in this dataset and is publicly accessible information so extra security measures aren’t necessary.*

How did you verify the data’s integrity?

*As it’s a relatively small dataset, I’ve skimmed through the cells in Excel to find any abnormal values.*

How does it help you answer your question?

*The provides a score to determine the university’s teaching and research capabilities as well as its student makeup which can directly be used to answer the business question.*

Are there any problems with the data?*The percentage of students column is written in %, and currently isn’t recognized as a numeric outside of Excel. In addition, one row is missing one of its values. Also, the gender ratio column isn’t formatted correctly however it is outside the scope of the business question.*

**Deliverable:** A description of all data sources used

***I used a dataset found on Kaggle that contains the THE University Rankings 2021, with both the dataset publisher as well as THE allowing the use of the data gathered.***

# Process

What tools are you choosing and why?  
*I will be using Excel to clean the dataset as it’s relatively small, and RStudio to join the 2 datasets and create the visualizations.*

Have you ensured your data’s integrity?  
*I’ve skimmed the dataset by hand to find errors and checked for NULL values.*

What steps have you taken to ensure that your data is clean?  
*I’ve formatted any columns so that they’ll be recognized by RStudio and fixed any missing values.*

How can you verify that your data is clean and ready to analyze?  
*If the datasets are added in RStudio without any issues, then it’s clean to work with.*

Have you documented your cleaning process so you can review and share those results?

*Cleaning process:*

*Removed the commas in the ‘number students’ column in the ‘university\_ranking’ dataset.  
Changed the perc intl students column datatype from percentage to float.  
Manually changed ‘Karlstad University’ perc intl students field from % to 0.  
Starting at rank 203 and below, THE doesn’t provide the exact overall score of each university, only score bands, which initially sets the overall scores column in university\_scores.csv as a string datatype, so I converted each score-band into a score-average (e.g. score band 50.6 to 54.2 into 52.4 for the purpose of converting the datatype into a float and enable analysis. This should affect the accuracy of the data as we can use ranking as a proxy for overall score.*

# Analyse

How should you organize your data to perform analysis on it?  
*I needed to combine the two datasets together so I can compare variables across them. In order to compare Australian university’s average performance with the rest of the world, I also needed a Boolean variable to sort between the two.*

Has your data been properly formatted?

*I needed to convert some number formats to be compatible with RStudio.*

What surprises did you discover in the data?

*That despite how Australian university’s overall score and ranking are on average higher than average, their teaching score is only around the global average.*

What trends or relationships did you find in the data?

*We see that the there is a small correlation between lower student enrollment total and lower student to staff ratio with higher teaching and research scores. In addition, Australian universities outperform the global average when it comes to international outlook and research but on average with the rest.*

How will these insights help answer your business questions?

*There needs to be a greater focus on improving teaching quality at Australian universities as well as a greater connection between university and the private industry (to increase private funding for university research).*

# Share

Were you able to answer the business question?

*Yes*

What story does your data tell?

*While Australian universities perform overall above average, there are areas of improvement.*

How do your findings relate to your original question?

*Australian universities need to work on specifically improving teaching quality as well as connections with local industry.*

Who is your audience? What is the best way to communicate with them?

*As the audience are a government department as well as potential employers, an R Markdown rendered into HTML is the best option.*

Can data visualization help you share your findings?

*A scatterplot can show correlation between two variables whereas a column chart can directly compare Australian universities with global ones.*

Is your presentation accessible to your audience?

*Yes, the graphs and written analysis are written to be accessible whereas the underlying code is available for those who can read it.*

# Act

What is your final conclusion based on your analysis?

*Australian universities can further improve specifically on its offering to prospective and current students.*

How could your team and business apply your insights?

*It can show how to direct funding to specifically target areas of improvement.*

What next steps would you or your stakeholders take based on your findings?

*Take further research into programs and collaborations between government and universities, and universities with local industry.*

Is there additional data you could use to expand on your findings?*More research into specifically teaching qualities at Australian universities could be helpful to supplement this analysis.*