DATA EXPLORATION & VISUALISATION

Table of Contents

1. Introduction:

Problem:

Australia's economy relies heavily on a skilled workforce to drive its industries and maintain competitiveness on a global scale. However, like many nations, it faces significant skill shortages in critical areas such as Engineering, Information Technology, and Health. Addressing these shortages is not only a matter of domestic education policy but also involves managing the contributions of international students who bring essential skills to the workforce.

Question:

1. What is the distribution of domestic and international students enrolled in courses that address skill shortages in Australia?

2. Does the course completion among students correlate with demographic factors such as gender ratios?

Motivation:

As a tech, sports and geopolitics enthusiast, I always had curiosity to understand the reasons to learn a new technology, the reasons behind the loss and wins in a game and the reason behind the conflicts among super power nations. And now as an overseas student it sparks my interest to understand how well the student mass is distributed in contributing towards current skill shortage crisis in Australia which gives a clarity on how far or close are students contributing towards this crisis.

Implications: This Report focuses on how enrolment of domestic and overseas students is distributed to contribute towards skill shortage fields such as Agriculture, Information Technology, Engineering Technologies, Health, Architecture and Construction, Education which has major effects in Australia. [3]

2. Data Wrangling

Data Sources: The project utilizes the latest data sourced from the Australian Government Department of Education[1], as of the most recent update on October 30th, 2023. This data serves as the primary foundation for conducting the comprehensive analysis and addressing the outlined objectives.

Data Source one: Student Enrolment Pivot Table.[1]

The above data source gives tabular information regarding student enrolment count between 2018 to 2022which has numerical field and many sets of categorical fields.

Data source two: Award Course Completion Pivot table.[2]

The above data source gives tabular information regarding student course completion count between the year 2018 to 2022.

Data source three: Reference of skill shortage fields. [3]

The Above data source gives huge set of information on skill shortages. It mainly highlights skill shortage areas like Information Technology, Health, Management and Commerce, Engineering related technologies, Agriculture and Environmental studies and Education. [3] These are major skill shortage fields in Australia and this exploration will keep this area of skill shortage as a single point to see how students contribute towards these skill shortages.

Data Set Information:

	Enrolment Data Set	Course Completion Data set
Number of Rows :	1,385,65	56,888
Number of Columns :	Number of Columns : 10	Number of Columns : 10
Numerical Fields:	ENROLMENT_COUNT	COMPLETION_COUNT
Categorical Fields:	YEAR, INSITITUTION, STATE, CITIZENSHIP, BROAD_COURSE_LEVEL, DETAILED_COURSE_LEVEL, GENDER, TYPE_OF_ATTENDANCE, EDUCATION_FIELD, ENROLLMENT_COUNT	Categorical Fields: Simillar to Enrolment Data

Step 1: Data Transformation

The given data set was in a pivoted table form which had to be unpivoted to bring it back to a readable tabular form. This was done with help Excel pivot tool. Figure 2.1 below shows the table before transformation and figure 2.2 shows the after transformation tabular data.

Fig 2.1 Data before Transformation

	A	B	C	D	E	F	G	H	I	J
1	Year	Institution	State	Citizenship	Broad Cours	Detailed Cours	Gender	Type Of Att	Broad Field	Completions
2	2018	Non-Universi	Australian Ca	Domestic	Postgraduate	Postgraduate	Female	Part-time	Health	325
3	2018	Non-Universi	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Part-time	Health	30
4	2018	Non-Universi	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Informatio	6
5	2018	Non-Universi	Australian Ca	Domestic	Postgraduate	Postgraduate	Female	Part-time	Health	89
6	2018	Non-Universi	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Part-time	Health	3
7	2018	Non-Universi	Australian Ca	Domestic	Undergradua	Bachelor	Male	Part-time	Informatio	3
8	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Female	Part-time	Manageme	0
9	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Engineering	3
10	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Female	Full-time	Engineering	4
11	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Female	Full-time	Manageme	5
12	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Health	1
13	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Health	50
14	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Manageme	37
15	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Engineering	11
16	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Manageme	6
17	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Manageme	3
18	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Female	Full-time	Engineering	14
19	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Female	Full-time	Manageme	13
20	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Engineering	4
21	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Health	1
22	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Female	Full-time	Engineering	3
23	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Female	Part-time	Engineering	0
24	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Engineering	16
25	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Female	Full-time	Agriculture	8
26	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Male	Full-time	Agriculture	12
27	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Informatio	3
28	2018	The Australia	Australian Ca	Domestic	Undergradua	Bachelor	Male	Full-time	Engineering	55
29	2018	The Australia	Australian Ca	Domestic	Postgraduate	Postgraduate	Female	Part-time	Agriculture	0

Fig 2.2Data After Transformation

3. Data Checking

Removal of redundant columns: Indvidual education field columns where removed to reduce the complex structute as they could be readily identified by 2 exsisting coulmns which are"broad field of education" and "broad course level".

Generating consistency in Categorical data: All the categorical coulmns and it's rows were converted to upper case to maintain uniformity thus making things easier during exploration. This has been done by using excel upper() function and with the help of special paste tool which made sure all values were pasted to the respective coulmn without any redundancies.

Removal of possible leading or trailing spaces: By uploading the data set to R enivironment and using trimws() function ensured removal of possible leading and trailing spaces.

Duplicate Values: This data set doesn't constitute to duplicate values since it has been generated based on the final total enrolment and completion count from the original dataset. With this, the main dataset doesn't provide unique identitifiers like student id and thus identification of duplicate values doesn't hold true and can give different interpretations if duplicates are removed has repetiton of digits in enrolment and completion count is possible.

Treating Missing values and Blanks: A simple method to identify would be to check filters once the data is uploaded to excel. No missing data were identitified in the Enrolment data however there were missing data identified in the type_off _attendance coulmn of Completion data when they were observed for missing and balnk values using ifelse condition inside a mutuate function in R programming , these missing categorical data are repalced by UNKNOWN name tag by using a ifelse condition inside a mutate function in the R enviroment.

	A	B	C	D	E	F	G	H	I	J
1	YEAR	INSTITUTION	STATE	CITIZENSHIP	BROAD_COURSE	DETAILED_CO(	GENDER	TYPE_OF_ATTEND,	EDUCATION	ENROLLMENT_COUI
2	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	MALE	PARTTIME	AGRICULTURE	2
3	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	MALE	PARTTIME	ENGINEERING	15
4	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	FEMALE	PARTTIME	EDUCATION	2
5	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	FEMALE	FULLTIME	HEALTH	2
6	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	HEALTH	9
7	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	HEALTH	35
8	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	MALE	FULLTIME	HEALTH	0
9	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	EDUCATION	32
10	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	ARCHITECTURI	65
11	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	HEALTH	4
12	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	FEMALE	PARTTIME	MANAGEMEN1	3
13	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	ARCHITECTURI	47
14	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	EDUCATION	17
15	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	HEALTH	33
16	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	MANAGEMEN1	8
17	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	FEMALE	PARTTIME	MANAGEMEN1	134
18	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	UNDERGRADUAT	FEMALE	PARTTIME	ARCHITECTURI	3
19	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	FEMALE	FULLTIME	HEALTH	17
20	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	EDUCATION	0
21	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	AGRICULTURE	11
22	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	FEMALE	PARTTIME	ARCHITECTURI	45
23	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	PARTTIME	ENGINEERING	152
24	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	HEALTH	80
25	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	MALE	FULLTIME	INFORMATION	10
26	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	MALE	PARTTIME	MANAGEMEN1	0
27	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE (	FEMALE	FULLTIME	HEALTH	74
28	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	POSTGRADUATE	POSTGRADUATE I	FEMALE	PARTTIME	HEALTH	75
29	2018	THE UNIVERSITY OF W	WESTERN AUSTRALIA	DOMESTIC	UNDERGRADUATE	BACHELOR	MALE	PARTTIME	HEALTH	0

Fig 3.1 dataset ready for exploration

4. Data Exploration:

At intial stage let's perform descriptive statistics to understand the distribution of our student enrollemnt data. Below are the summary statstics of total and average enrolement count details generated using tableau.

Total Enrolment Count in Education fields (2018-2022)		Average Enrolment Count in Education fields (2018-2022)	
Education Field		**Education Field**	
AGRICULTURE STUDIES	96,364	AGRICULTURE STUDIES	10.98
ARCHITECTURE AND BUILDING	208,903	ARCHITECTURE AND BUILDING	21.69
EDUCATION	620,985	EDUCATION	28.31
ENGINEERING TECHNOLOGIES	565,780	ENGINEERING TECHNOLOGIES	30.08
HEALTH	1,357,867	HEALTH	44.51
INFORMATION TECHNOLOGY	546,375	INFORMATION TECHNOLOGY	29.68
MANAGEMENT AND COMMERCE	1,760,175	MANAGEMENT AND COMMERCE	57.73

Figure 4: Descriptive statistics for student enrollment count for both domestic and overseas students

From the above table of figure 4 we can see that there's a high enrolment count and corresponding average for the Health and Management commerece, and next notable fields are Education followed by Engineering, IT and the rest. This gives us an rough idea on how enrolment has been distributed in these education fields which contribute towards skill shortages. Now Let's dive in deep to understand the trend of student enrolement in these education fields that contribute to skill shortage crisis in Australia.

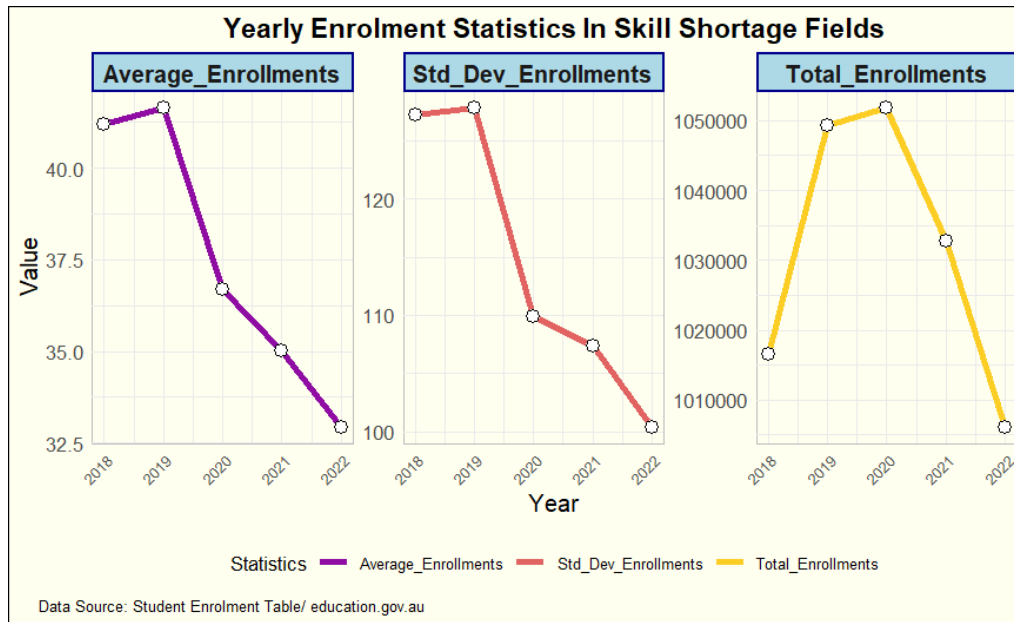To do this we will be using a facted line chart generated with R programming.

Figure 4.1: Descriptive statistics for student enrollment count for both domestic and overseas students

By using summaries () function for enrolment count in R which gives the numerical description on how data is distributed through mean and standard deviation and this summary is plotted with respect to years using ggplot library [4][5].

Total Enrollment Changes: The third chart, in yellow of fig: 4.1, illustrates fluctuations in total enrollments, peaking between 2018 and 2020, then dropping significantly in 2021, a year marked by the pandemic's severe impact on education, causing disruptions like deferrals and dropouts.

Enrollment Trends Overview: The first chart, shown in purple of fig: 4.1, reveals a steady decline in average enrollments over the past five years, with numbers hitting a low in 2022. This decline likely stems from several factors, including major disruptions from the COVID-19 pandemic affecting educational norms.

Enrollment Variability: The second chart, colored red of fig: 4.1, tracks enrollment variability, a high standard deviation constitutes to values which are far from mean showing a decrease in the enrolment trend.

Now that we have an idea about the distribution of enrolments in education fields lets dive into the preferences of domestic and overseas students to understand their enrolment contribution in education fields that address skill shortage in Australia.
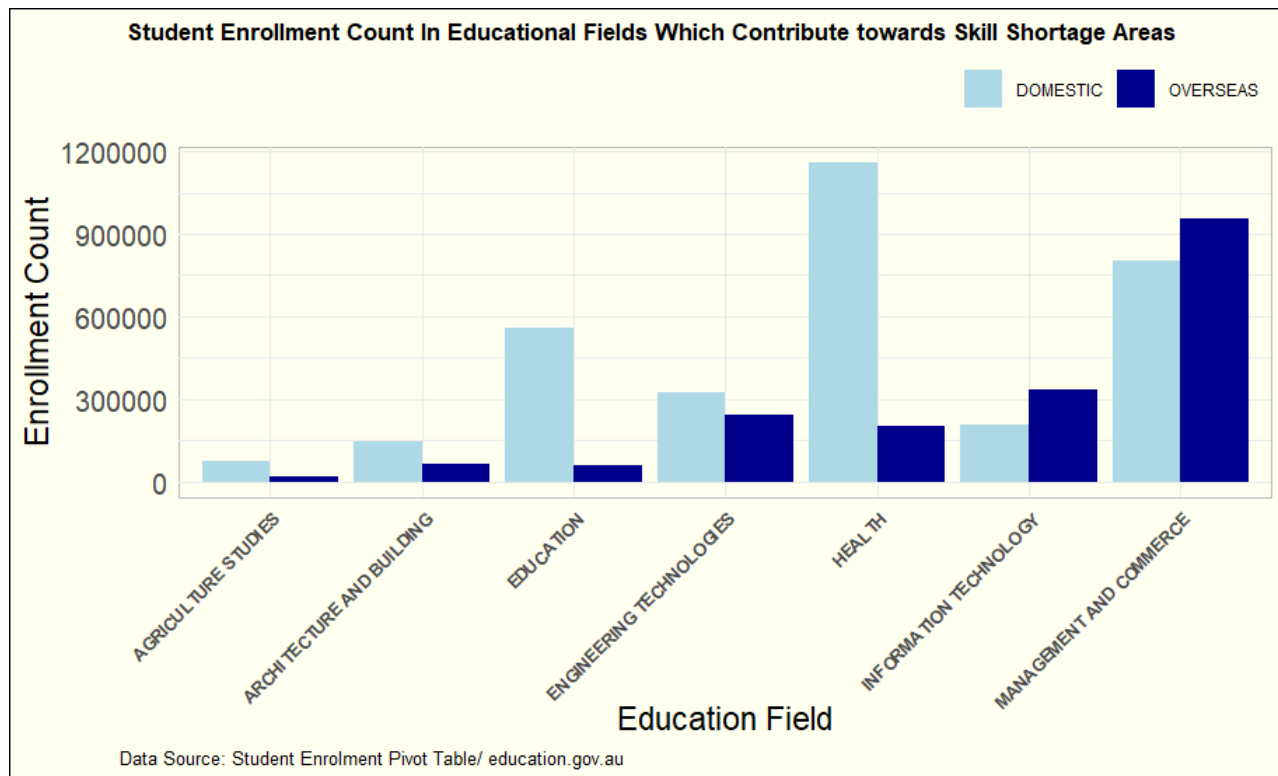
Figure 4.2: Student's preference in educational fields that accounts to skill shortage.

The above visualization in fig: 4.2is a simple boxplot created using geombar() function in R programming.

The analysis of student enrollments from the graph reveals distinct preferences: watching the visualization it is very clear that domestic students predominantly prefer Health, Education, Management & commerce over other educational fields whereas for overseas students, fields like health and education have lesser enrolments showing lesser interest in these fields, this can be due to multiple factors a possible factor could be the standardized admission process with certain criteria.

For overseas students, most of the enrolments are in the field of management and commerce and even domestic students prefer this educational field and the possible reason could be the job opportunities here in Australia due to skill shortage crisis and high pay rates. Domestic students also prefer IT and Engineering related fields due to the same reason. Education fields such as agriculture and architecture have least enrolment which show low interest rates by both domestic and overseas students in these fields.

Now we know the preferences of domestic and overseas students in educational fields that contribute towards skill shortage let's try to understand how the enrolment rates has varied from year to year which will give us an idea about the enrolment patterns
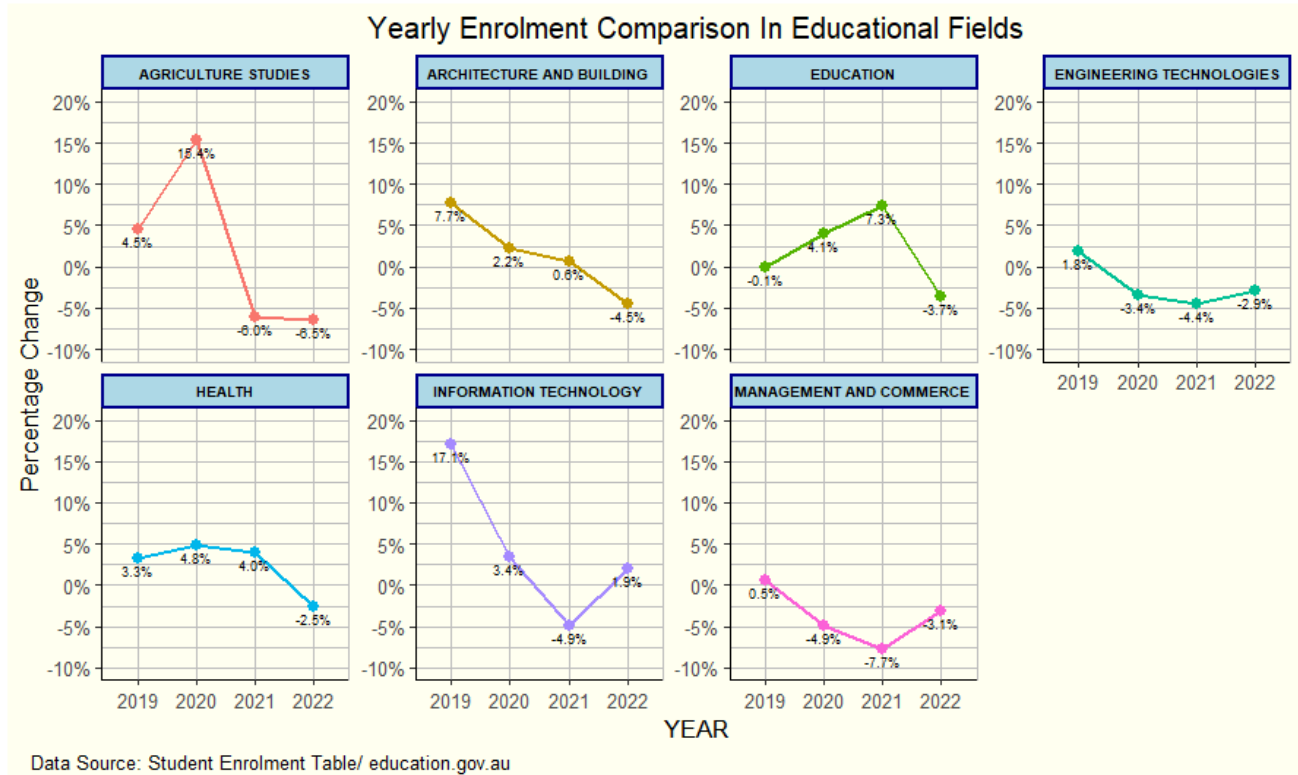
Figure 4.3: Explains enrolment fluctuations in skill shortage fields.

Figure 4.3 is a faceted line charts that gives an idea about enrolment rate from year to year explaining how the enrolments have evolved. This visualization has been created using R programming with the help of ggplot library and facet_wrap() functions to plot the visualization for individual education fields by calculating the percentage change of enrolments between the years.

Figure 4.3 gives a brief explanation on the enrolment rates yearly which will help us understand if the respective skill shortage field had an increase or decrease in the number of enrolments.

As we can see from figure 4.3, fields like Health and education follow some similar patterns of enrollment rates, Agricultural studies had its boom in 2020 and the enrolments rates crashed down from 2021 these can be due to external factors and impacts of COVID -19. Architecture field has seen down fall in enrollment showing least preference to it. Enrolment rates in information technology, management and commerce and engineering related technologies are seen crashing until 2021 and has started to rise from 2022 this can be due to several factors. One reason for enrollment fall would be because of reduction of overseas enrolment due to covid impact as overseas traveling was restricted during this period.
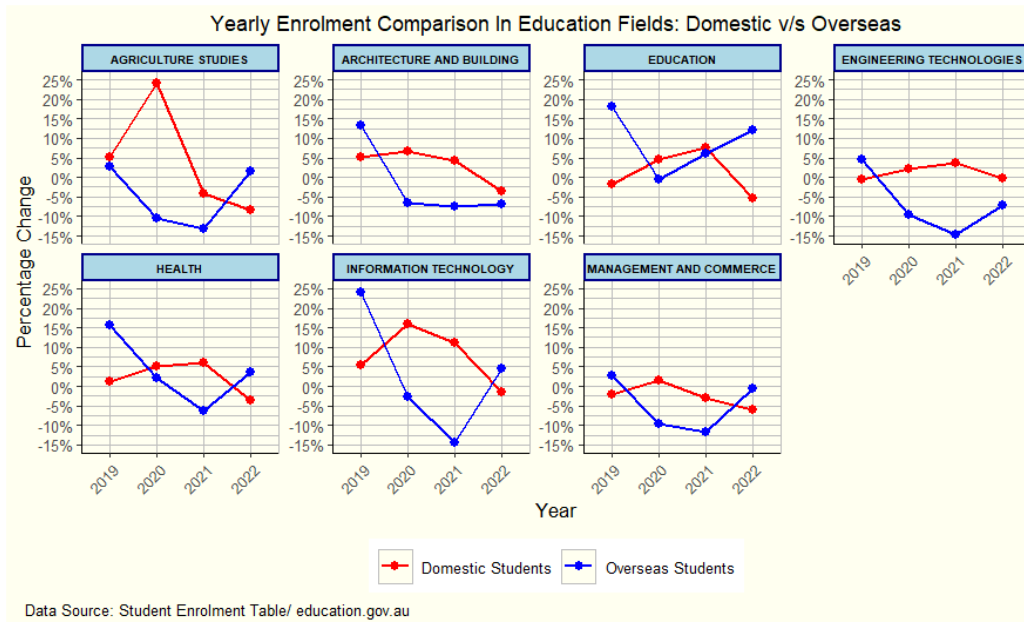
Figure 4.4: Explains enrolment fluctuations with respect to student type.

Figure 4.4 is a faceted line charts that gives an idea about enrolment rate from year to year based on student type. This visualization has been created using R programming with the help of ggplot library and facet_wrap() functions to plot the visualization for individual education fields with respect to student type.

The above figure gives us a rough idea of correlation between student type and enrolment rates. Comparing figure 4.4 with 4.3 we can see that the fluctuations of enrollment rates is directly proportional to the domestic student's enrolment trend, the same pattern follows for architecture and building and same follows for health and education, in all these fields, the enrolment rates correlate with domestic student enrolment rates. Where as for fields like Information technology with management and commerce graph pattern correlates with overseas student enrolment where the enrollment rate fluctuations follow the same patterns. In case of Engineering technologies field, the enrollment rates depend on both domestic and overseas students has they share equal interest in enrollment towards this field

This can be verified by chi-square [6] statical test to determine relationship between student type and their enrollment count in educational fields is Independent or dependent. To perform chi-square test, variables which are being analyzed must be categorical in our case both citizenship and educational field columns are categorical. Selection of one observation doesn't influence the selection of other, the sample size must be large, frequency of the cell must be greater than 5 and data must be collected in a random fashion. Out data set satisfies all these assumptions.

The Chi-square test was performed in R-programming environment using readxl library with the help of chisq.test(). At First contingency table consisting of CITIZENSHIP EDUCATION_FIELD column were created and this table was fed to chisq.test() function and the results gave a high x-squared value of 1523, degree of freedom value as 6 with which an extremely small p-value of 2.2e-16 was determined. With this small p-value we can reject the null hypothesis of independence

between the chosen variable which inversely means there is a strong statistical significance between student type and the educational fields which they opt for.

Now Let's see if course completion rate has a correlation with demographic factors such as gender ratios. At first let's understand the course completion count between the genders over the years.
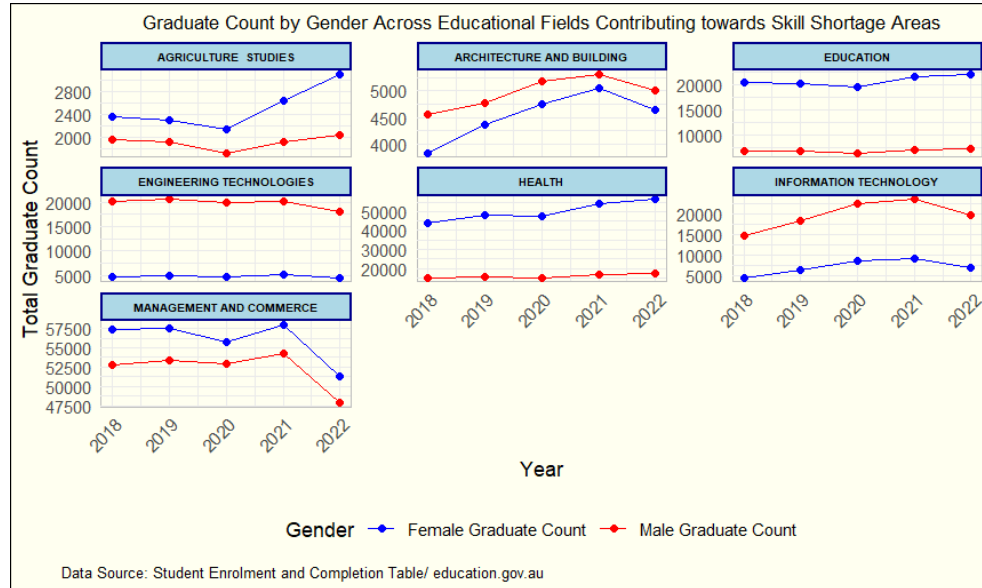

Figure 4.5: Explains graduate trend between genders over time.

Figure 4.5 is a faceted line charts which has been created using R programming with the help of ggplot library and facet_wrap() functions to plot the visualization for course completion with respect to student type.

From the derived facet line graph, it is very clear that the female dominated the males in terms of graduation count. The obvious reason for this is dominating female enrolment rates in the education field when compared to male enrolment. Now let's see if the course completion is corelated with the gender ratios. In order to test this, Pearson correlation [7] statistical test will be implemented in R

R packages dplyr and tidyr were used to process completion and enrollment data, focusing on male and female entries. Data from both the table was grouped by institution, year, field, and gender, and aggregated for total figures. The pivot_wider function converted data to a wide format with separate columns for each gender. To address divide-by-zero errors in ratio calculations, NA and Inf values were replaced with the calculated median values thus making a clear path for visualization of enrolment trend with respect gender over time.
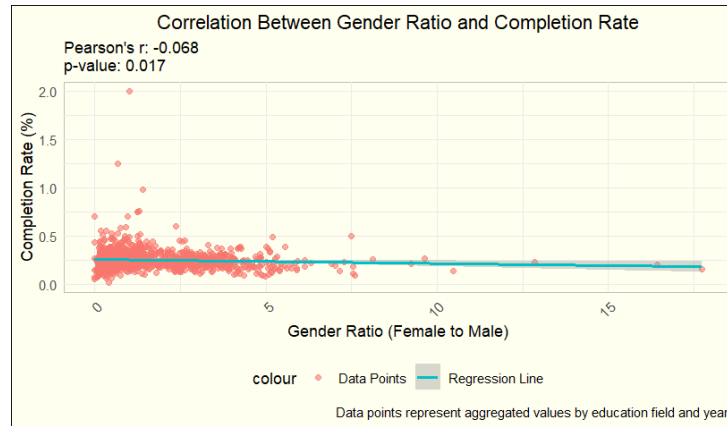
Figure 4.5: Pearson Correlation test between Gender Ratio and course completion.

The Pearson's correlation analysis shows a negative correlation (-0.068) between gender ratios and completion rates. This analysis suggests that there is a weak but statistically significant relationship between the proportion of females completing a course compared to males (gender ratio) and the overall completion rates. The negative correlation coefficient indicates that as the gender ratio shifts towards more females relative to males, completion rates tend to slightly decrease. However, the magnitude of this effect is quite small. Essentially, while there is a connection between gender ratios and completion rates, it's not strong, and other factors likely play a more significant role in determining completion outcomes in educational settings.

5. Conclusion:

1.While domestic students also contribute significantly to alleviating skill shortages in professions like management, commerce, and IT, overseas students predominate in these areas as the enrolment rates are high. Domestic students, on the other hand, have greater enrollment rates in the health education sector, indicating that they play a significant role in reducing skill shortages when compared to overseas student. enrollment rates in architecture and construction programs are trending downward for both local and international students, which makes it difficult to solve the skills shortages in these industries. The disciplines of education and agriculture continue to have respectable enrollment rates among domestic students.

2. By conducting Chi-square test it is evident that education fields like health, agriculture and education studies enrolment count dependent on domestic student enrolment rates and education fields like IT, Management and commerce enrolment count dependent on overseas student enrolment rate.

3. Female graduation rates are higher than male graduation rates in the education fields that address skill shortage. A Pearson correlation analysis shows a negative correlation coefficient which is weak, but indicates that completion rates marginally decline when the gender ratio moves in favor of more females than males.

6. **Reflection:**

With this exploration project I have learnt the steps to be taken when exploring a public dataset. Before jumping into any exploration, it is necessary to list down the data types that are present in the dataset followed by cleaning the data from blanks, outliers or nulls but at same time one has to be cautious in deciding whether the released outlier is a valid one or if it is the nature of data set. Aggreations always gives an abstract view of the data but not a in detail one but it definitely helps to identify trends. In-depth analysis with the help of statical test will help in identifying the patterns between the data types. My future research will concentrate on determining whether the skills of recently graduated students satisfy the necessary industrial standards.

7. **Bibliography**

1. Reference:[1] Student Enrolments Pivot Table 2022 - Department of Education, Australian Government

2. Reference:[2] https://www.education.gov.au/higher-education-statistics/resources/award-course-completions-pivot-table-2022

3. Reference: [3] Skills Priority List (SPL) | Anzscosearch

4. Reference: [4] https://ggplot2.tidyverse.org/reference/ggtheme.html

5. Reference: [5] https://ggplot2.tidyverse.org/reference/ggtheme.html

6. Reference:[6] https://www.geeksforgeeks.org/chi-square-test-in-r/

7. Reference:[7] https://www.r-bloggers.com/2021/10/pearson-correlation-in-r/