Impact of Gender on Educational Outcomes at Secondary and Higher Education level in the UK

CFG Project

By Adelaide Baron, Claire Evans, & Sian Steen

Introduction

For our final CFGdegree project on the data stream, we have used Python and analytical libraries to analyse the gender split in education. To do so, we thought of this from the point of view of data analysis consultants working on behalf of a company. Our (theoretical) client is a charity organisation who wishes to assess the impact of gender on educational outcomes. The organisation is working to improve access and opportunities for females in the UK. In particular, they are looking to increase the number of females entering the technology industry, which has historically been male-dominated.

Our main points of focus are:

1. What outcomes are being achieved by females at secondary and higher education levels, and how does this differ from outcomes for males?
2. Is there an increase in the number of females studying computer science-based subjects and in females entering the technology industry?
3. Is there a link between geographical location and the educational outcomes for females?

Background

Our group is investigating the outcome for female students leaving secondary education and higher education. We have analysed three stages of UK education:

i. 16-18 destination measures
ii. First-year enrolments at higher education providers
iii. Graduates of higher education providers

The aim is to identify what UK females decide to do on completion of secondary level education, and whether this differs to male students. How are female students performing and what do graduates go on to do after completing their degrees? There is an increased awareness of gender equality in the UK and ensuring equal opportunities for all. We expect to find that female graduates earn less than their male counterparts and that there are fewer female students enrolling on technology related degrees. Our analysis aims to provide an evidence base to answer the questions posed in the proposal. Are females attaining successful outcomes at each level? If so, in what subjects? And is there a difference in geographical location? The analysis we provide to our client will be used to identify where new opportunities can be created for females and where and how to promote options such as courses and qualifications in technology related areas.

Steps Specification

Data gathering

Initial data gathering was undertaken as a team in an agile planning session. To refine the project title and objectives, we searched for data sources online that could be analysed to answer our questions. At first, we wished to include widening participation markers in our analysis, but soon realised our project was too broad and needed to be more focused. Data wasn't comparable between higher and secondary education datasets and we chose to focus on gender, for which there is ample data. Our project objectives were refined to frame our questions set out in the introduction.

Data sources were identified through a desk-based literature review. Through existing knowledge within our team, we were able to identify reliable data sources such as the gov.uk website for secondary level data and the HESA website for Higher Education data.

Finding relevant REST APIs (or similar) was challenging, as most were either inaccessible to us (e.g. private, required payment, only for businesses), or were for app development. We did however find the UniDB API, which contains data on higher education providers and their programmes, with information directly related to our project aim. Additionally, to analyse the coverage of the UK from our data, we used a Geocoding API to map the locations we have analysed.

Preprocessing

Higher Education Statistics Agency (HESA) data:

Data from HESA is largely clean data as it has already been pre-processed due to its use in informing funding allocations for higher education providers and in other publications, such as league tables. A rounding strategy is applied to HESA data so as to prevent the identification of individuals (further information on this can be found at www.hesa.ac.uk). All pre-processing is fully documented in the HESA Analysis Jupyter Notebooks.

To note, we are looking at snapshot data for academic years for secondary and higher education levels. We are not tracking individuals or cohorts through their education journey.

API data

UniDB

In terms of preprocessing, the API didn't have detailed documentation on accepted universities or degrees as parameters, and as a result we were unsure if an error would be returned. To call the API and get information regarding each degree or university, we needed to start with a list of each. For degrees, we created a list of the most common and broad degrees from two online resources. For universities, we created a list using online lists such as the top 100 universities and, to ensure we try to cover a broad range of UK locations, we entered all universities named 'University of' <location>. Then, we created methods for each to test if they were successful in the API, and created lists of those that were. For universities we also tried adding 'The' to the university name to test if that brought back a result.

Additionally, following initial preprocessing, further data cleansing was performed throughout analysis. By requesting frequent outputs of any data gathered from the api, it was evident that even with valid calls (universities/degrees known to the API), some calls would return ‘none’ for the header we required. We used methods to remove ‘none’ values from a dataframe, and performed this on our dataframe of university female-male ratios.

In-depth analysis

Regarding datasets, gender data for some was only available for female/male split. An ‘Other’ category is available for some datasets; however, the reported numbers were very low, and were removed from the analysis so as not to skew the data. Additionally, there are also some ‘Not Knowns’ recorded in the data. These numbers were extremely low and have been excluded from the analysis as the focus of our project is looking at gender, and gender cannot be determined in these records. In the Graduate Outcomes data, we have removed the data for non-respondents to the Graduate Outcomes survey, as we do not know what activity they are undertaking following completion of their degrees.

Similarly, for API data, only female or male was available.

The team undertook an initial review of the educational outcomes of females in the UK. The available data was broad and varying in nature, and was not consistent across education levels. As the data relates to individual student outcomes, data is not published at the individual level to prevent identification of individual students. In addition, there would be millions of individual records to analyse in this scenario. Data is published by counts of students in particular categories, which did place some limitations on our options for analysis.

As we began cleaning and preprocessing our data, we became aware that data was only consistently available for England. We have targeted our analysis on schools and higher education providers in England to gain a fuller picture of educational outcomes.

HESA data

In-depth analysis was conducted on a number of datasets; first-year enrolments at higher education providers, first-year enrolments by subject area, graduate activities and graduate salaries. Pandas and MatplotLib were used extensively to analyse and plot the data and identify any trends present. There is limited time series analysis available for some data due to changes in data collection methods, such as the change in subject coding ahead of 2019/2020 student record data collection. As further data become available, further analysis should be undertaken to investigate patterns in the data. We attempted machine learning to perform linear regression in order to predict female student numbers in the future, however, more data are required as current data are sparse, and may mean currently our model will be undertrained.

Implementation and Execution

Agile Approach

Once our project title and questions were confirmed, Sian was assigned the role of scrum master. We split the work into 3 key areas:

- Secondary level data (gov.uk) – Claire
- higher education data (www.hesa.ac.uk) – Sian.
- API's – Adelaide

With just three members of our team, this was a large project to take on. We had an initial discussion to lay out our goals and steps to complete the project. We decided to work in short sprints of Monday to Thursday and Friday to Monday, with video calls on Slack on a Monday and Wednesday evening. Using a Kanban board we managed our backlog with group visibility, and managed code on a mutual GitHub repository. We set up our own Slack channel and communicated throughout the week with any questions, queries and progress updates.

Tech Stack Implemented

We have utilised the following tools and technologies in our project work:

- Jupyter Notebook - performing our analysis
- Pandas - analysis, dataframes
- MatplotLib - for plotting graphs
- Excel - data analysis
- GitHub - code management
- Trello - Kanban Board

Challenges

During our project, we have faced a number of challenges. Finding suitable datasets for the analysis, and which contained relevant data for our project was the first obstacle. It was very challenging to find an API which contained relevant and up-to-date data, which also had appropriate documentation for us to be able to utilise it appropriately. We faced the usual challenges of working with large datasets, including the time it takes to clean and pre-process the data before starting analysis.

Additionally, It has been difficult to include the various tools and techniques we have been learning in the specialisation stage of our course. These techniques are new to all of us and because of the time constraints for the project, we have not yet been able to fully understand and implement these techniques. We were only able to undertake limited time series analysis due to availability of data and this has hampered our ability to implement machine learning techniques such as linear regression.

Time has been the biggest issue in completing the project. Each team member works full-time and alongside the CFGDegree course commitments, it has been difficult to spend the required time on analysing our data and creating our project.

Results

Secondary Education

The initial dataset used was National 16-18 destination measures provided by Gov.UK. This had too many options to provide a meaningful insight, and recent reforms to technical and applied qualifications mean that the makeup of level 3 and level 2 approved qualification groups differ significantly between destination years 2017/18, 2018/19 and 2019/20. Comparisons across years and approved/other qualification groups should therefore not be made, so this wasn't much use.

We instead went on to use a smaller set of headline data, comparing a smaller number of destination options by gender. We looked at 2019/20 to see what the most recent trend is, and compared to the earlier 2017/18 set to see if there was any difference.

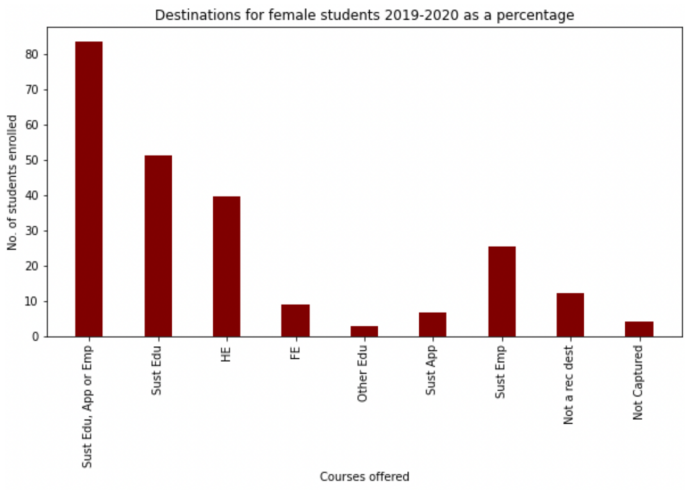

Figure 1

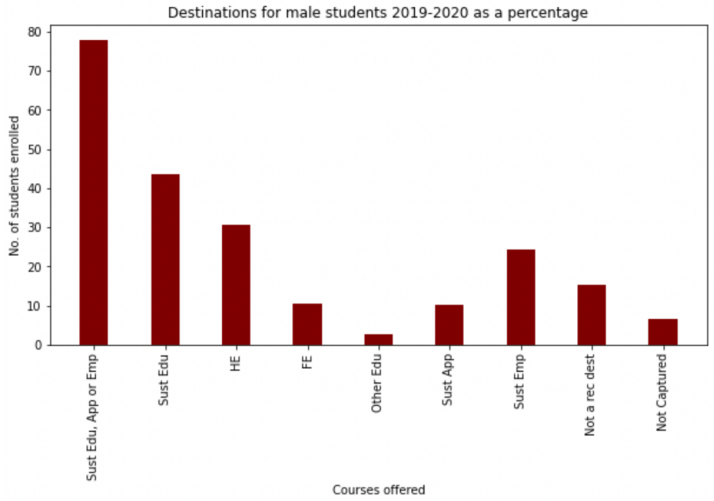

Figure 2

2019/20 shows a similar trend for both female and male students. The majority of students move into sustained education, apprenticeship or employment, with sustained education being the most popular option.

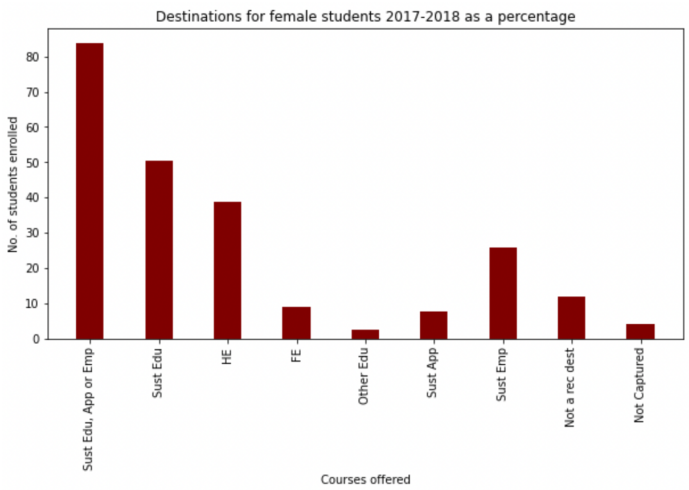

Figure 3

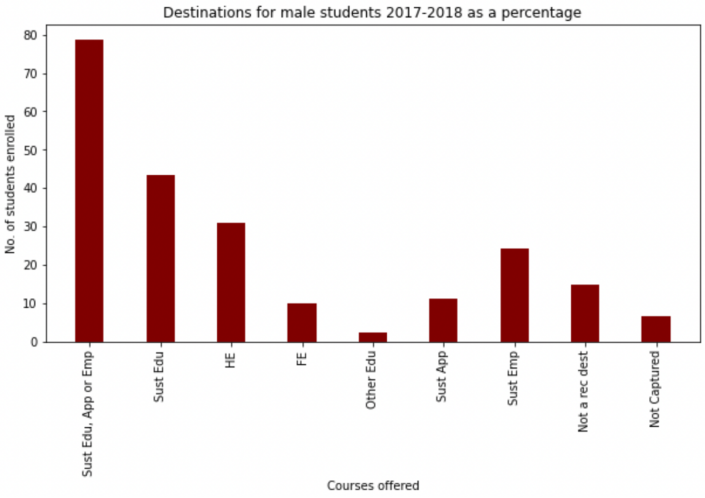

Figure 4

Higher Education

First year student enrolments

HESA data

Main Data Results:

- The percentage of female students as a percentage of all students was 55.18% in 2014/2015, increasing to 56.28% in 2020/1; a +1.06% in 7 years.
- The mean number of first-year enrolled students between 2011 and 2021 is 1,070,695 and the median value is 1,054,234.
- The percentage change in the number of female students from 2014/2015 to 2020/21 is +34.16%.

An analysis of the percentage for first-year enrolled students, as a proportion of total students, revealed a small number of institutions which have only female students (such as Royal Academy of Dance). However, a further analysis would need to be undertaken to establish the type of higher education provided and whether they are considered specialist, as this could provide reasoning for their student populations.

Top Ten Higher Education Providers by Number of First-year Enrolled Female Students				
	2014/5		**2020/1**	
Rank	**Higher Education Provider**	**Number of Female Students**	**Higher Education Provider**	**Number of Female Students**
1	The University of Manchester	7855	University College London	13705
2	University College London	6750	The University of Manchester	11130
3	The University of Leeds	6640	King's College London	10555
4	The Manchester Metropolitan University	6175	The University of Leeds	8755
5	King's College London	6085	The University of Birmingham	7765
6	The University of Birmingham	5945	The Manchester Metropolitan University	7725
7	University of Nottingham	5655	Anglia Ruskin University	7505
8	University of the Arts, London	5535	The University of Sheffield	7415
9	The University of Sheffield	5420	The Nottingham Trent University	7320
10	Coventry University	5410	Coventry University	7190

Table 1

Using table 1, and the Geocoding API, we have mapped the top 10 universities for both 2014/2015, and 2020/2021, shown in figures 5 and 6. By mapping English cities on our map for reference, we can see that for both academic years, the top 10 are all in the area between Manchester and London (North/North-West, Midlands). This is a region covering roughly 27,145 square kilometers, 20.84% of England's 130,279 square kilometers.

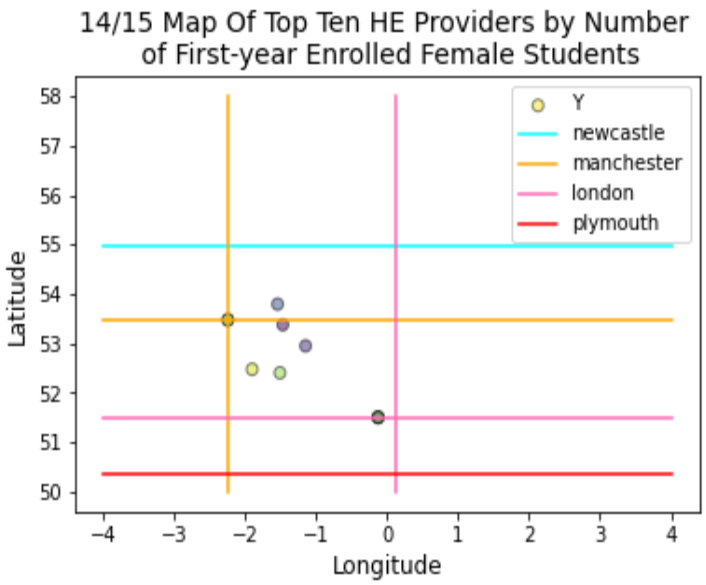

Figure 5

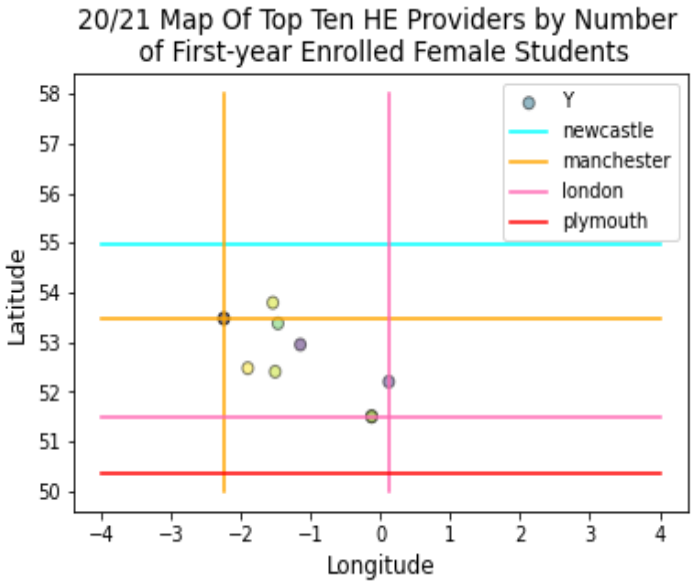

Figure 6

Percentages of female students in Universities

API data

The average percentage of female students of the universities analysed was 56.5%, indicating more females than males in higher education, supporting the HESA first year student enrolment data.

Considering universities individually, we see that the majority have more females than males, as displayed in figure 7. However, there are two that fall below the 50% mark, namely Coventry University and Newcastle University. Online resources show that Newcastle has a female population of 49.4%, and Coventry 49.05%, so this split is below the city average. Further studies may research the degrees offered at the individual universities, and compare this with the information returned from the degree endpoint. Perhaps there is a correlation, for example those with a low female percentage may offer a lower amount of female-dominated courses.

To analyse potential drives at each university with the data available from the UniDB api, we compared the staff population to students, as shown in figure 8. The female staff average population across all universities analysed was 43.94%, showing more men than women. There appears to be a weak positive correlation between the two, with more staff generally equaling more students, but there is not enough data to draw a conclusion from.

Locations

We are aware that the API was not able to capture all of the UK, so we have used a Geocoding API (Open Weather) to map the universities used for analysis. Using the demographics endpoint, figure 9 displays a Map of the locations analysed.

We have also displayed the female population percentages on a bubble chart, figure 10, to visualise the location and percentages. A larger bubble indicates a larger percentage; the smallest, most northern, bubble is Newcastle. This again shows the location split, but also shows that Universities around London (the

localisation) tend to have a larger percentage of females than the North. Keeping in mind however, from figure 9, that we may just not have data from universities in other locations.

Future studies may be beneficial to map out all UK universities using the geocoding API, and compare the universities we have analysed with the API to see if we have an accurate representation of the entire England University population.

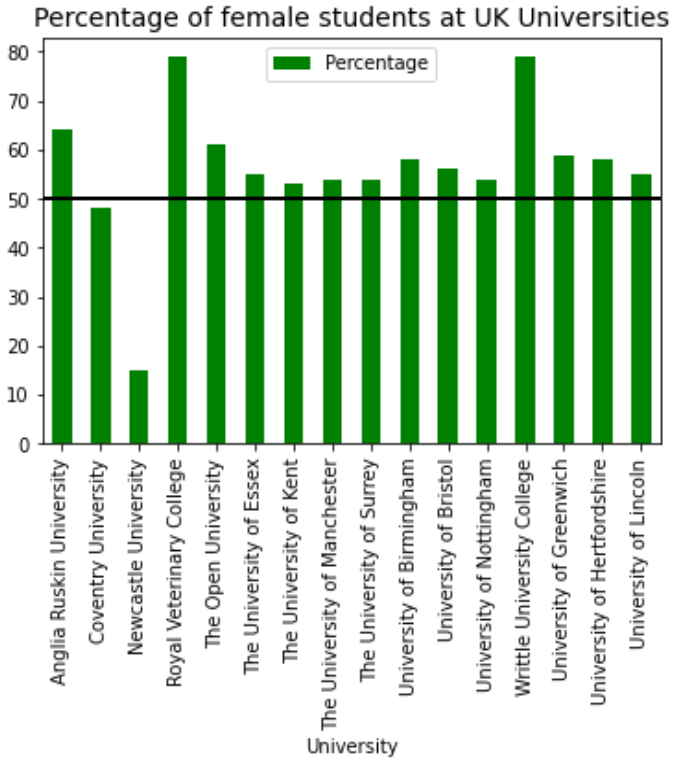

Figure 7

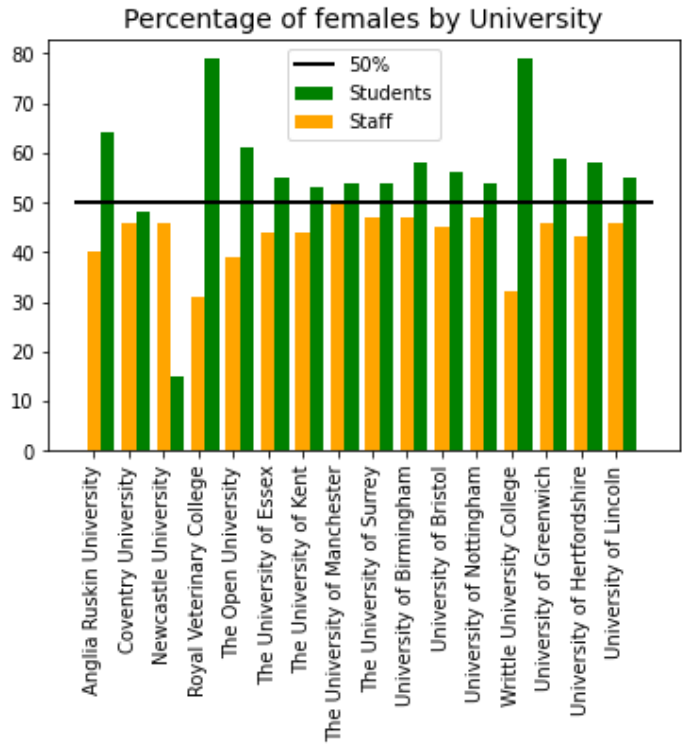

Figure 8

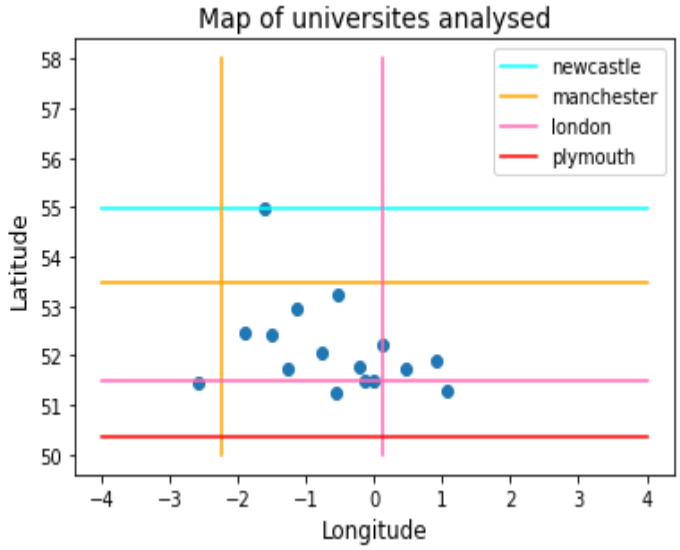

Figure 9

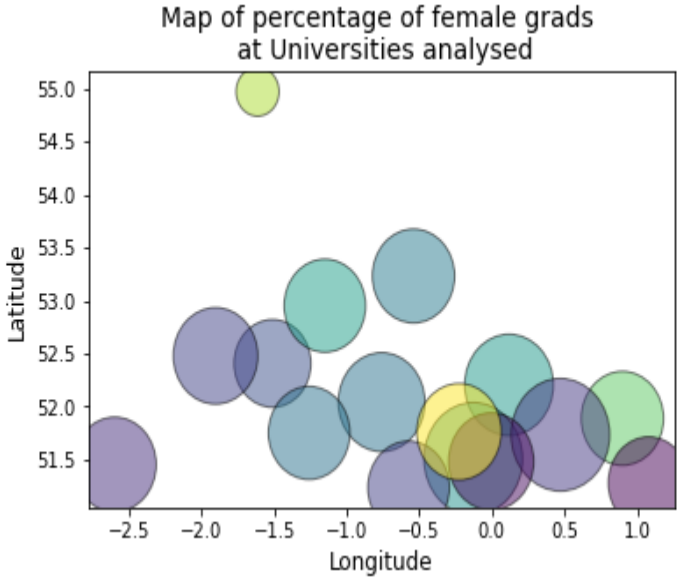

Figure 10

What are female students studying?

HE student enrolments by subject of study (Full-time) 2020/1

For first year enrolments at Higher Education providers in 2020/2021, the highest number of female students enrolled on courses assigned to the CAH subject ‘17- Business and management’, closely followed by ’02- Subjects allied to medicine’. The most popular subject area for male students that year was also ‘17- Business and management’, but there is then a considerable drop in numbers to the second most popular subject, ’10 - Engineering and technology’. The least popular subjects for female students were ’23 – Combined and general studies’, ’26 - Geography, earth and environmental studies (social sciences) and ’05 – Veterinary sciences’.

In science-related subjects, more male students are enrolled on courses in subject areas ‘10 - Engineering and technology’ and ’11 – Computing’, however, female enrolments are considerably higher in ’02- Subjects allied to medicine’ and higher than males for courses in ’01 – Medicine and dentistry’.

A similar pattern can be seen in the subject enrolment for part-time students.

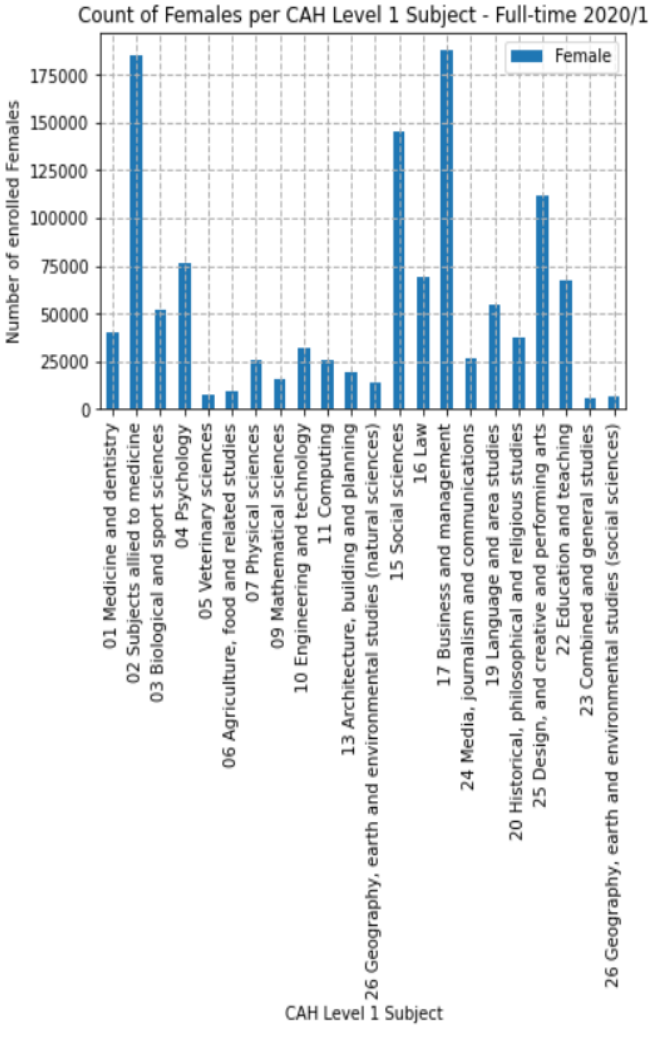

Figure 11

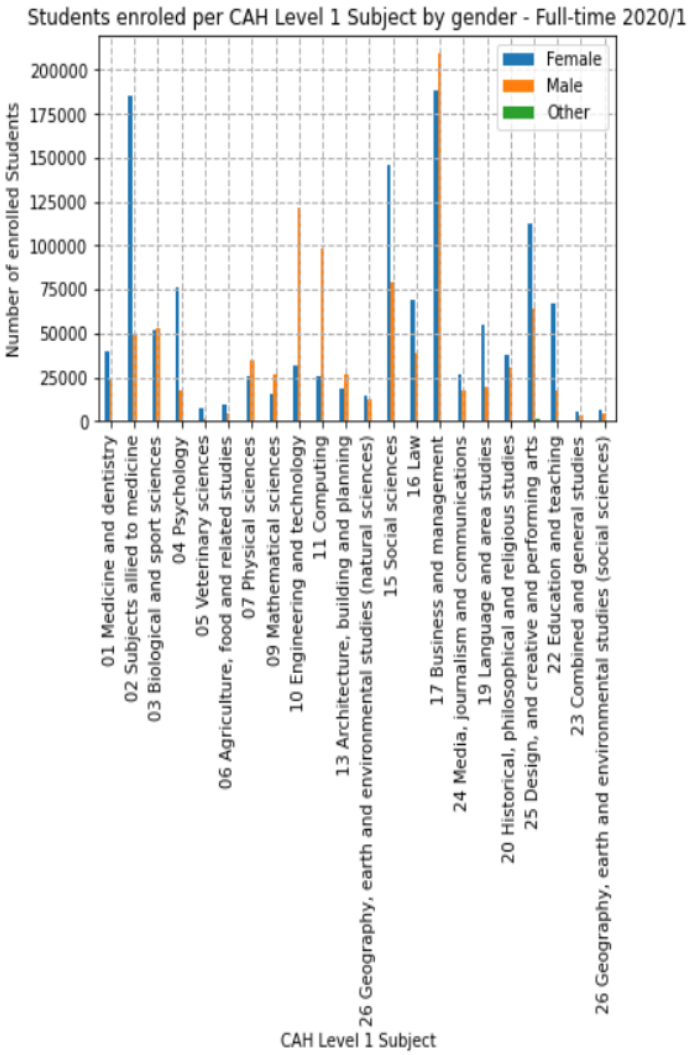

Figure 12

The only degrees we could analyse from the API are displayed in figure 13. Considering there are many degrees, in excess of 59 (viewing UCAS courses available) undergraduate degrees on offer in the UK, this is not a full representation. It does, however, give a rough overview of some of the most common degrees.

A ratio of 1 indicates an even split between male and female students. It is apparent that degrees historically male occupied, such as mathematics and computer science, have some of the lowest female to male ratios. Computer science has an average ratio of 0.240, around 6 women per 25 men, that's less than a third of the population being female, and the rest male. Medical Science, however, shows an average ratio over the five year period of 3.798, around 19 women per 5 men. However, both medical science and computer science are STEM degrees, which typically are stereotyped as a whole. These figures show that sciences are popular among women, for example biology has a ratio of above 1, meaning more females than males. However, it is the potentially computer-based degrees, as mathematics is too low, that are male occupied.

Figure 14 allows for better visualisation of the change year to year. All degrees, apart from mathematics and medical science, either remain relatively unchanged, or increase, as years go on.

As we are particularly focused on the gender split in computer science and mathematics, as they are more typically 'tech'/'IT' oriented degrees, we also predicted when they would reach a ratio of 1:1 if they continued at their current growth rate. For computer science, current data predicts a 1:1 ratio within 288 years. Mathematics is actually decreasing by an average of -0.0368 every four years, meaning that we cannot predict an increase to 1 with this data, as it only decreases.

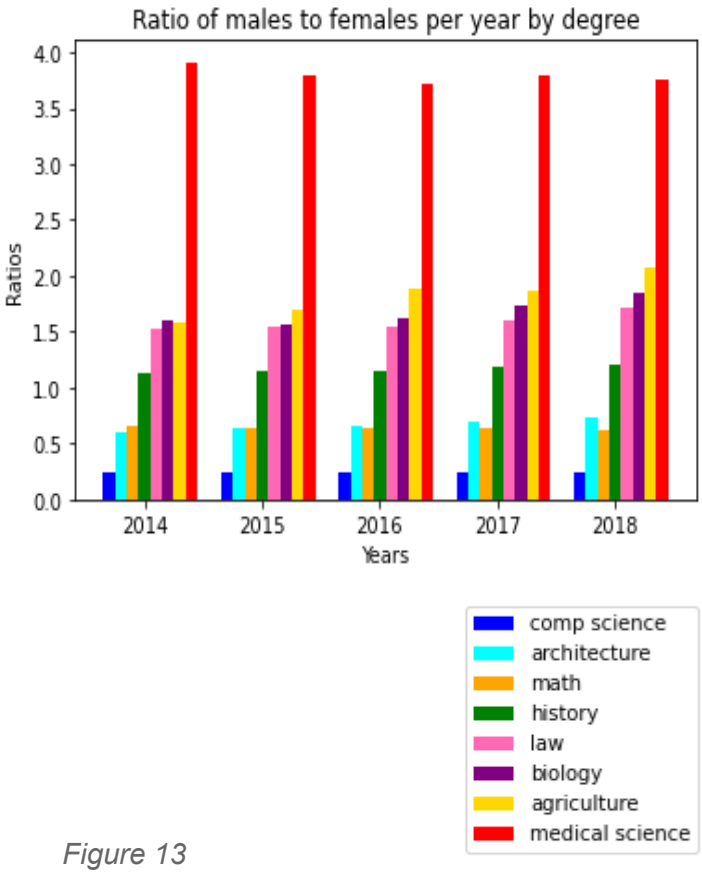

Figure 13

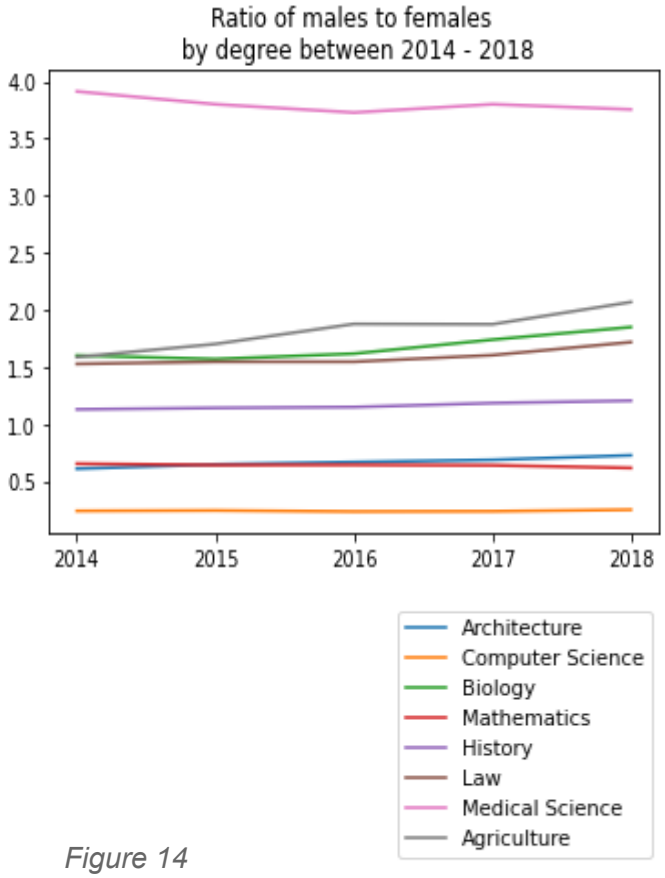

Figure 14

Graduate Outcomes

What are female graduates doing after completing their courses?

Graduate Outcomes by Activity Undergraduate Full-time 2019/20

59,935 female undergraduates were in full-time employment 6 months after completing their courses, whereas 42,040 males followed the same path. A higher number of females than males were also in employment or further study, full-time further study, and part-time further study. This shows that higher female populations went on to graduate outcomes than males.

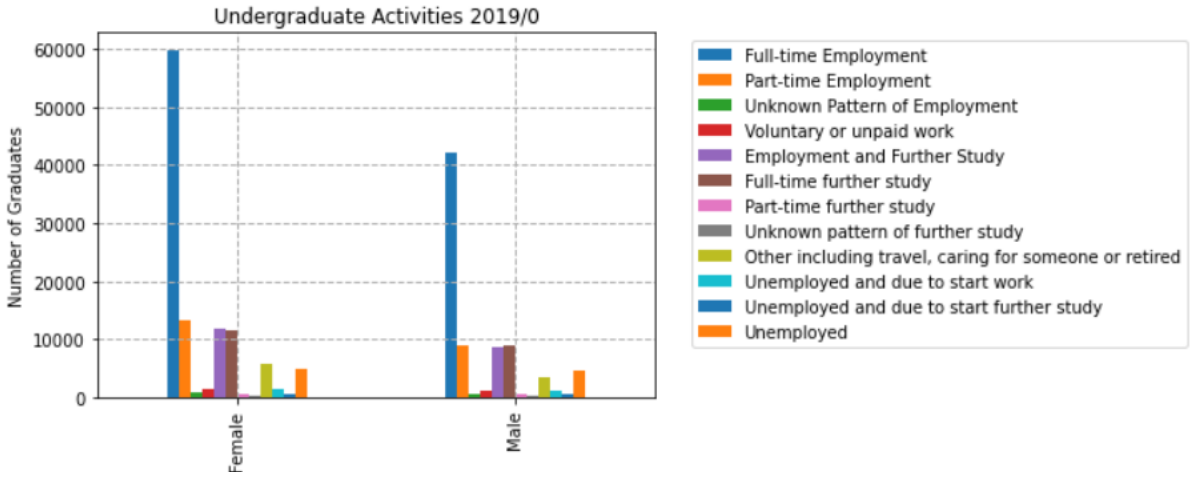

Figure 15

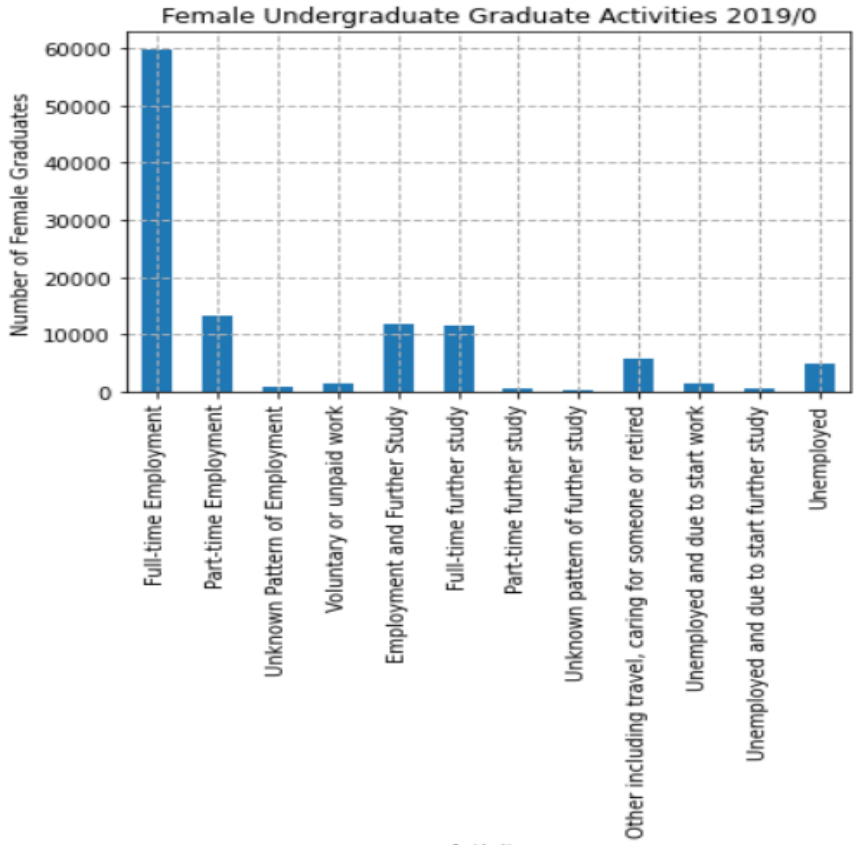

Figure 16

How much are female graduates earning?

Percentage of graduates in full-time paid employment in the UK by salary band and personal characteristics 2019/2020

The following data is for students in full, or part-time, employment only, as a percentage of all graduate activities.

For graduates (both male and female) in 2019/2020, the most common salary bracket was £24,000 - £26,999. The plot is right skewed, showing more students to be earning salaries of £32,999 or lower. There is a clear difference in the higher salary brackets, with more males earning the higher salaries of £30,000 - £32,999, upwards to £51,000+. This is in contrast to more females in the lower salary brackets of less than £15,000 and £24,000- £26,999.

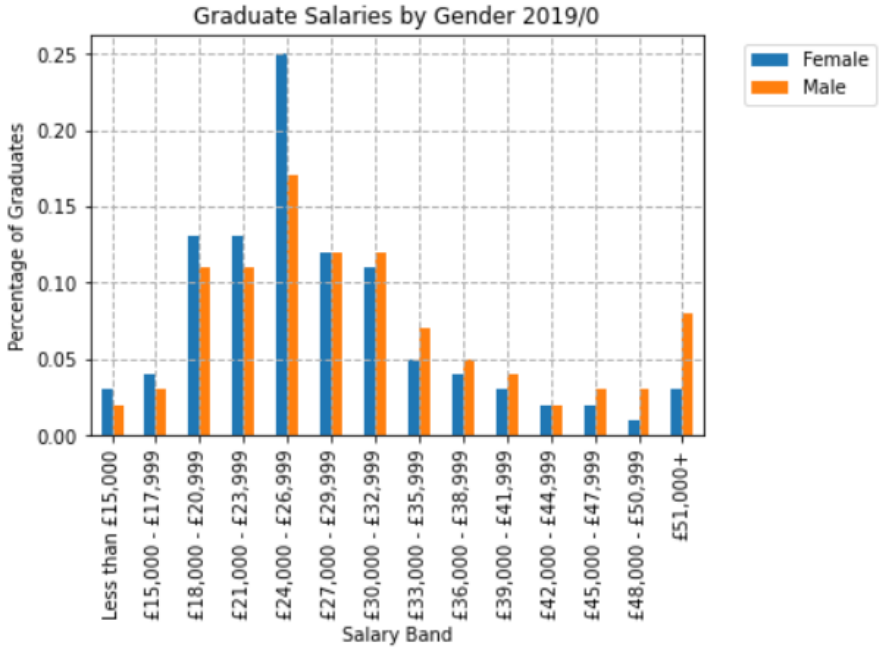

Figure 17

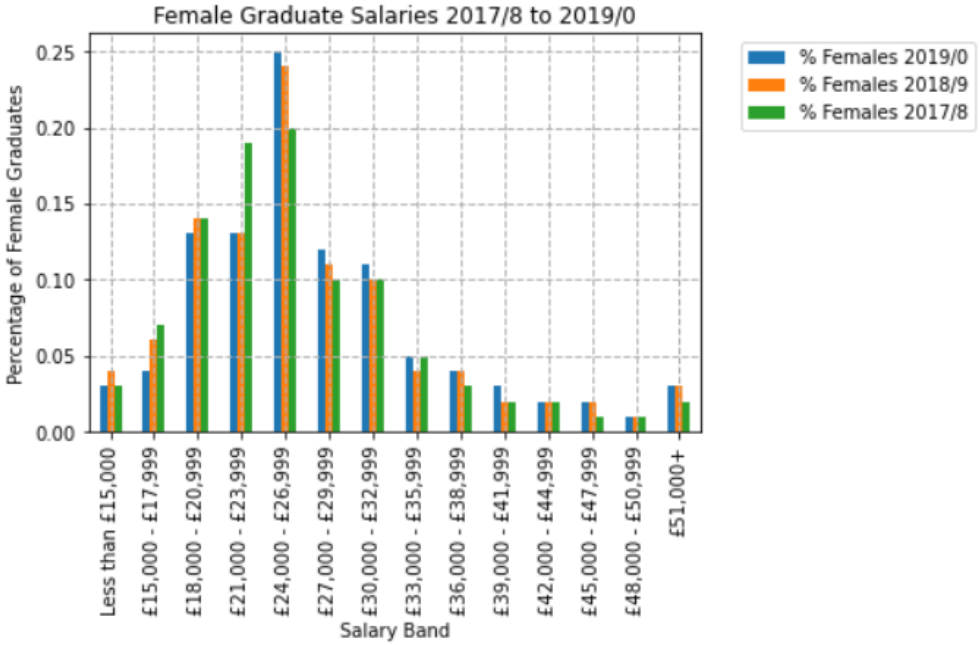

Figure 18

There were no significant changes in the salaries female graduates are earning from 2017/2018 to 2019/2020. There have been increases in the number of females earning the salary bands of £15,000 - £17,999 and £21,000 - £23,999.

For salary bands of £24,000 - £26,999 upwards, numbers have either remained consistent, or there has been a decrease in the number of female graduates earning higher salaries (such as, £51,000+, £45,000 - £47,999, and £36,000 - £38,999).

Conclusion

To summarise our findings, we will return to our project questions.

1. What outcomes are being achieved by females at secondary and higher education levels and how does this differ from outcomes for males?

Secondary Education

At a secondary level, the data for 2017/2018 is almost identical to 2019/2020, indicating that this pattern is a steady constant. In 2019/20, the number of female students completing 16-18 education was 258,569, compared to a similar figure of 262,548 males. This is likely because education is compulsory at this stage. 83.80% of female students went on to sustained education, apprenticeships or employment, compared to a similar 78.70% of males. At this stage in the education journey, gender appears to have zero impact on the destination.

Higher Education

- Female students represent more than half the student population in England (56.28% in 2020/2021).
- Female student numbers are increasing (percentage change in female student numbers from 2014/5 to 2020/1 is +34.16%).
- Female graduates (from undergraduate courses) are predominantly in full-time employment following completion of their programmes.
- More female graduates than male graduates are in full, or part-time, employment or study than male graduates.
- However, female graduates are earning less than male graduates. Female salaries are predominantly in the salary bands up to £32,999, and more male graduates are earning salaries upwards of £33,000 than female graduates.
- Female salaries have not experienced significant change between 2017/2018 and 2019/2020.

2. Is there an increase in the number of females studying computer science-based subjects and in females entering the technology industry?

Higher Education

- Fewer female students (than male students) are studying subjects in the subject areas of 10 - ‘Engineering and technology’ and ’11 – Computing’. However, female enrolments are considerably higher in ’02- Subjects allied to medicine’ and higher than males for courses in ’01 – Medicine and dentistry’

- APIs show that over a four year period, mathematics and computer science had more males than females. However, there was a steady increase for computer science - a one-to-one ratio to be achieved within 288 years at the current growth rate.

3. Is there a link between geographical location and the educational outcomes for females?

Higher Education

As seen from the maps created, using the UniDB we have poor representation of places above Manchester (e.g. Lancashire), and west of Manchester (e.g. Liverpool). However, with the data available, it would appear that the universities around London have, on average, a higher female population than male. Considering the lowest value is in Newcastle, North, this is supported. However, further analysis would be required to make a reliable comment on this question.

We must also note that locations for higher education may not be representative of young people's home locations. So, although our maps show a concentration in the area between the East Midlands and London, this may be representative of student populations. Extensions of this project would review the spread of students throughout the UK, and perform further comparisons. Additionally, should information become available regarding student's home regions in certain universities, we could perform a more accurate geographical analysis.

Recommendations

Our recommendations to our partner organisation are as follows:

- Target promotion of computing and science-based subjects to secondary level pupils to encourage them to apply to study subjects in these areas at higher education level. Increase support and initiatives to encourage female pupils to apply for science-based programmes, working with companies to promote work experience and to drive sponsored places on degrees.

- Encourage and support female graduates to apply for jobs in the technology sector and for roles with higher salaries. Create programmes to assist in writing job applications and interview skills to coach female graduates and improve their confidence.

- Work with companies to identify and create opportunities for females to enter the technology sector.

Next Steps and Lessons Learnt

Due to time constraints, there are many aspects we were not able to explore in this project. To further develop our analysis, we would like to review further data sets to undertake a more robust analysis. We used open data in our project but would have liked to investigate what data could be requested via an information request. As the data we are using is about individuals, we are limited to counts and categories of students, rather than individual records.

Regarding secondary school data, due to educational reforms in 2017, it was not possible to compare data before and after this date. As we were looking to see how trends followed through into higher

education, it made sense to focus on the later information available. This meant that our range was limited from 2017/2018 to 2020/2021, so it wasn't possible to see any trends over a long period of time. For future studies, we will look to obtain a more historical range of data.

Additionally, obtaining a wider (date) range of data would allow us to use machine learning to predict how female educational outcomes will change in the future. Part of the challenge here is the change in reporting of data. For example, the HESA data returns have changed over the past 10 years, which means that it is not possible to make direct comparisons between years without significant mapping.

We focused on education in England in this project, but in enhancing our geographical analysis, we would like to analyse data for Wales, Scotland and Northern Ireland to investigate whether there are any differences across the devolved administrations. We may encounter issues with differences in data collection and reporting across the country, so measures would have to be put in place to consider data fairly, or do direct comparisons.

Finally, given more time, we would like to stick to a stricter agile work process. For our API research we created a lot of methods for filtering and to avoid repeating ourselves. We didn't have time for a thorough test process, only testing along the way with printing. We would like to use unit testing in the future, and perform code reviews.