CSC8631: An Exploratory Data Analysis of Newcastle University's Future Learn Course in Cyber Security

By Ambreen Masud

02 December, 2021

1. Introduction

This project was commissioned by Newcastle University to gain insights into an online Future Learn course in cyber security. The aim of this project is to conduct an analysis on data provided by Newcastle University of the cyber security course, in order to gain an insight into course productivity, the learners, and better understand future course delivery. This course was delivered over 7 runs, with the initial run beginning in September 2016 and the final run beginning in September 2018. On average, each run had a total of 62 steps in order for students to complete the course.

The structure of this report follows the CRISP-DM approach carried out throughout the data analysis of this project. This methodology is as follows:

- **Business Understanding** - Gaining an understanding of Newcastle University's business needs and determining the goals of this project based on those needs
- **Data Understanding** - Understanding the properties of the data provided, including assessing data quality
- **Data Preparation and Modeling** - Cleaning, constructing and formatting the data, readying it for analysis and visually displaying the results and outputs with a descriptive analysis
- **Deployment** - Providing concluding thoughts on the project in the form of this report.

This report answers key questions which were based on Newcastle University's business needs, with a focus on developing an understanding of the learners. The findings of the initial analysis also led to further analysis in order to further develop the findings of this report. This report will conclude that whilst the typical student is educated and employed, there are opportunities for wider participation as atypical students also enrol on the course such as those that are unemployed.

2. Business Understanding

Newcastle University is a leading university in the United Kingdom, with a global reputation for academia, research and employability. The university collaborates closely with Future Learn, an online learning platform, to deliver a range of courses to a wide variety of students globally. This project analysed a Future Learn course run by Newcastle University entitled *"Cyber Security:Safety at Home, Online, in Life"*.

2.1 Assumptions

Before analysing the data, a number of assumptions were made, which were based on understanding the business:

- The data will display how many students enrolled on the course and will likely show that more students enrolled as each run progressed.
- The data will show where students are based geographically, which will provide insights into which countries to target in future course promotional activities.
- Analysing the student demographics will provide insights into understanding who to market future course promotional activities to.

2.2 Business Questions

The data assumptions led to a number of questions during this project:

1. How many students enrolled per run?
2. Where are the majority of students situated?
3. What are the student demographics and who should be targeted for future marketing perspectives?

These questions can not only determine analytical insights into the development of the course, but it can also support future marketing drives in increasing enrolments. In understanding the consumers, which in this case are the students, there are opportunities to put measures in place to ensure more effective courses are being run with improved student attainment. As an educational institution and a business, these are key goals in developing the business objectives of the university. This is aligned with the recommendations outlined by the Higher Education Commission (2016), who suggest that learning analytics should inform student development.In answering the business questions effectively, the data mining process included cleaning and transforming the data in order to calculate totals for the purposes of plotting the data and statistics.

3. Data Understanding

The dataset provided by Newcastle University comprised of 7 PDF files and 52 CSV files. Each dataset is differentiated by which run the data was collected from, so the records vary depending on the run. The PDF files provide information on the course overview of each run, highlighting each step that students need to complete as part of the course. The CSV files included within the dataset are:

- **Archetype Survey Responses** - Responses from surveys conducted on better understanding the learners, their needs and motivations. Runs 1 to 2 contain no data, whereas runs 3 to 7 contain data.
- **Enrolments** - Student enrolment records, including enrolment dates and times, and information on student profile. Data is available for all runs.
- **Leaving Survey Responses** - Responses from a survey conducted when students left the course, including information date and time information, the reasons why students left and what step they ended the course on. Runs 1 to 3 contain no data, whereas runs 4 to 7 have data.
- **Question Response** - Responses from quizzes that the students took as part of the course. Data is available for all runs.
- **Step Activity** - Records on step completion, including information on start and finish dates and times. Data is available for all runs.
- **Team members** - Information on the members of the course organising team. Data is available from run 2 to 7.
- **Video Stats** - Information on the videos watched as part of the course, including statistics on views, downloads and the type of devices that the videos were watched on. Data is available from run 3 to 7.
- **Weekly Sentiment Survey Responses** - Student responses on their point of view on the course on a week by week basis. Data is available from run 5 to 7.

For the purposes of this analysis and in line with the business objectives and key questions, the enrolments data will be assessed in further detail. Further processing on other datasets may be required, depending on the results of the analysis on enrolments.

There are 7 CSV files entitled *Enrolments*, each with a number from 1 to 7 indicating which run each dataset belongs to. Each dataset has 14 identically named columns as can be seen below.

```
##  [1] "learner_id"              "enrolled_at"
##  [3] "unenrolled_at"           "role"
##  [5] "fully_participated_at"   "purchased_statement_at"
##  [7] "gender"                  "country"
##  [9] "age_range"               "highest_education_level"
## [11] "employment_status"       "employment_area"
## [13] "detected_country"        "run"
```

Column 1 provides information about the student's ID number, whilst columns 2 and 3 provide information about when students enrolled and when they unenrolled, with column 4 highlighting student status. Columns 5 and 6 provide information about student's participation on the course, whilst columns 7 to 13 provide personal information about each student. An additional column was added to this data to clearly highlight which run is being looked at, which is the final column in the dataset.

In assessing the data quality, some of the data has missing values with empty fields and through glancing at the data, much of it is with fields labelled *Unknown*. This can be seen in table 1 which highlights a portion of the dataset and displays that the data is incomplete in some parts. This will be further assessed and cleaned in the data preparation portion of this report. However, much care will be taken to avoid losing data. Therefore, data cleaning will be carried out individually, depending on the project aims.

Table 1: Sample of the enrolments dataset

Fully Participated	Gender	Country	Age Range	Education Level
	Unknown	Unknown	Unknown	Unknown
	male	PE	46-55	university_degree
2016-09-22 16:56:03 UTC	Unknown	Unknown	Unknown	Unknown
	Unknown	Unknown	Unknown	Unknown
	Unknown	Unknown	Unknown	Unknown
2016-10-25 12:44:14 UTC	Unknown	Unknown	Unknown	Unknown

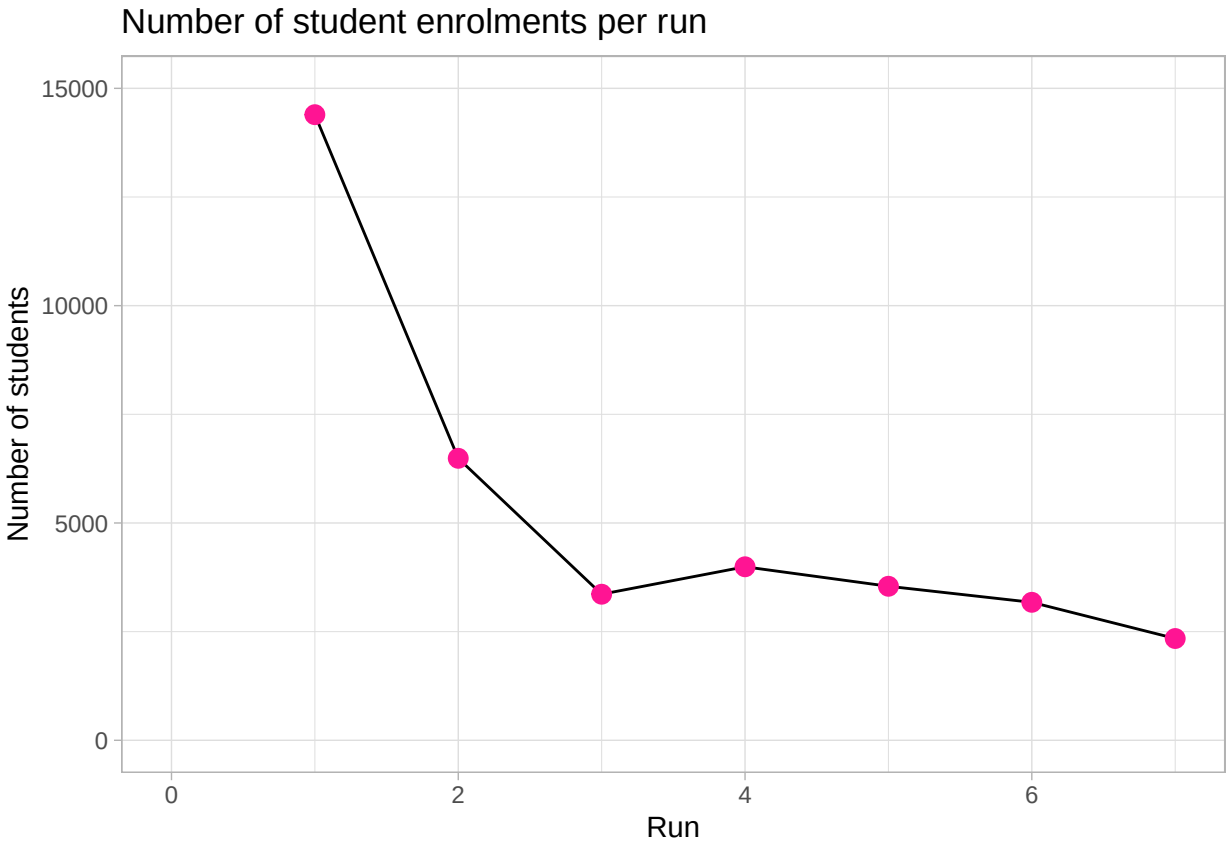
4. Data Preparation and Modelling

This section will be divided according to the business questions, and will discuss the data cleaning process per question in greater depth, and present the results of the data analysis.

In order to process the data easily, the enrolments data from each run was given an extra column to indicate which run the data belonged to. This was then combined into a new dataset entitled *all_enrolments*. The new dataset has a total of 14 columns and 37296 rows. This combined dataset has been used as the basis for all of the analysis on the business questions outlined in section 1, and the data will be cleaned according to each question as to avoid data gaps as much as possible. The types of data in this dataset are categorical and datetime, therefore a descriptive analysis will be used when analysing the data.

4.1 How many students enrolled per run?

In analysing the number of student enrolments per run, the data was grouped by each run before counting the number of enrolments per run. The plot below highlights the findings from this.



As can be seen in the table and plot, the number of students enrolling on the course has reduced after every session, except for a small increase in run 4. There is a large difference between the first run and the last run, with a 83.73% decrease between them. This is surprising as it defies the assumption made during the business understanding stage of this project, which assumed that as each run progresses, more students will enrol on the course. However, the data shows the opposite is true. This raises further questions as to why this may be, yet there is no data in the *all_enrolments* dataset which answers this question. Instead, this will be explored towards the latter portion of this report using another another file from the raw dataset provided by Newcastle University. The plot further strengthens the need to better understand the learners in order to improve student attainment.

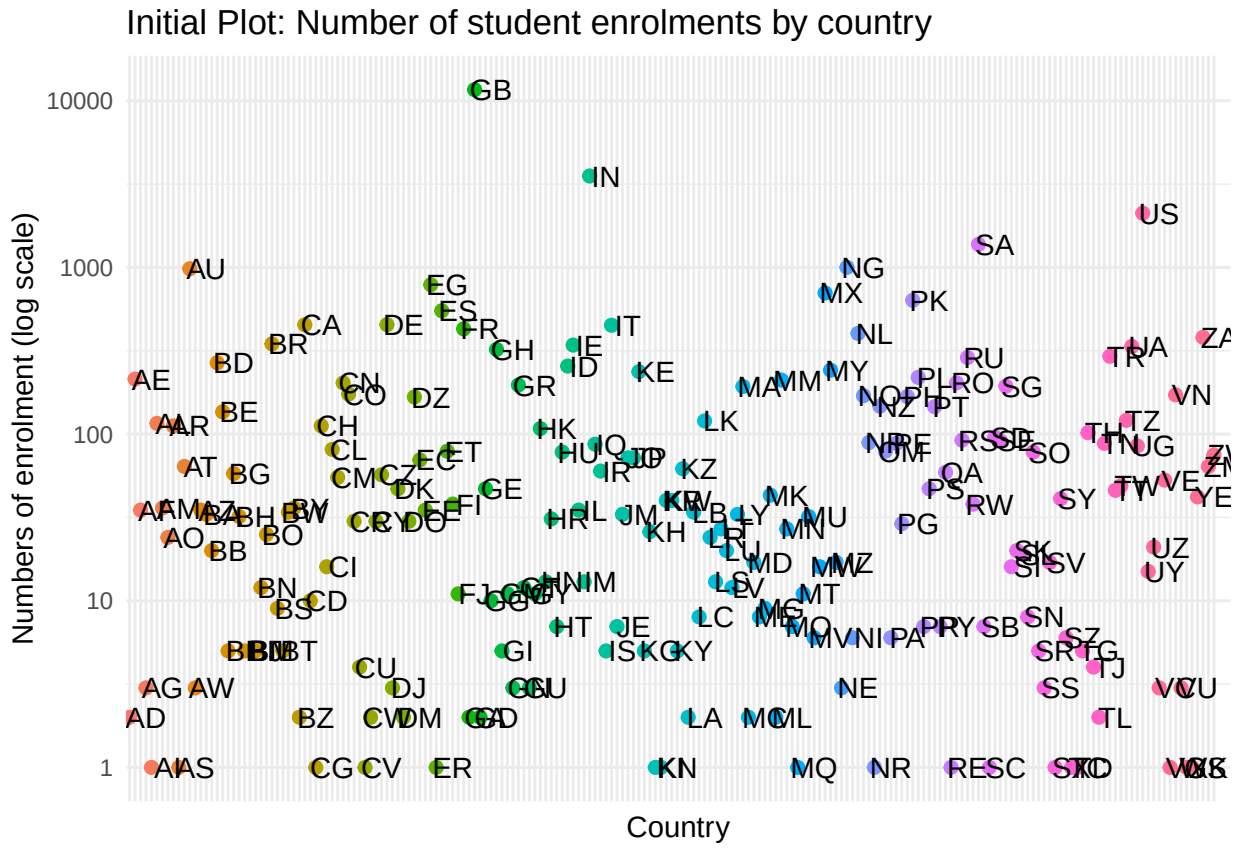
4.2. Where are the majority of students situated?

When considering this question, there are 2 columns in the dataset that can help provide the answer; *country* and *detected_country*. Based on assumption, *country* is information gathered from the student when enrolling, whilst *detected_country* is based on analytics provided by Future Learn. However, when delving further into the data and assessing missing data, *country* has 33142 missing values and *detected_country* has missing values labelled *"–"*, of which there are 930 in total. Due to the large discrepancy in missing data between each column, the *detected_country* data is used in this analysis.

A new dataframe was created which included the total number of students enrolled per country across all runs, a sample of which can be seen in table 2. The data was not divided per run as 199 countries were detected across the dataset. Whilst this is an interesting observation as the reach of the course is vast across many countries, it is difficult to visualise a large dataset across multiple runs.

Table 2: Sample of number of Student Enrolments by Country

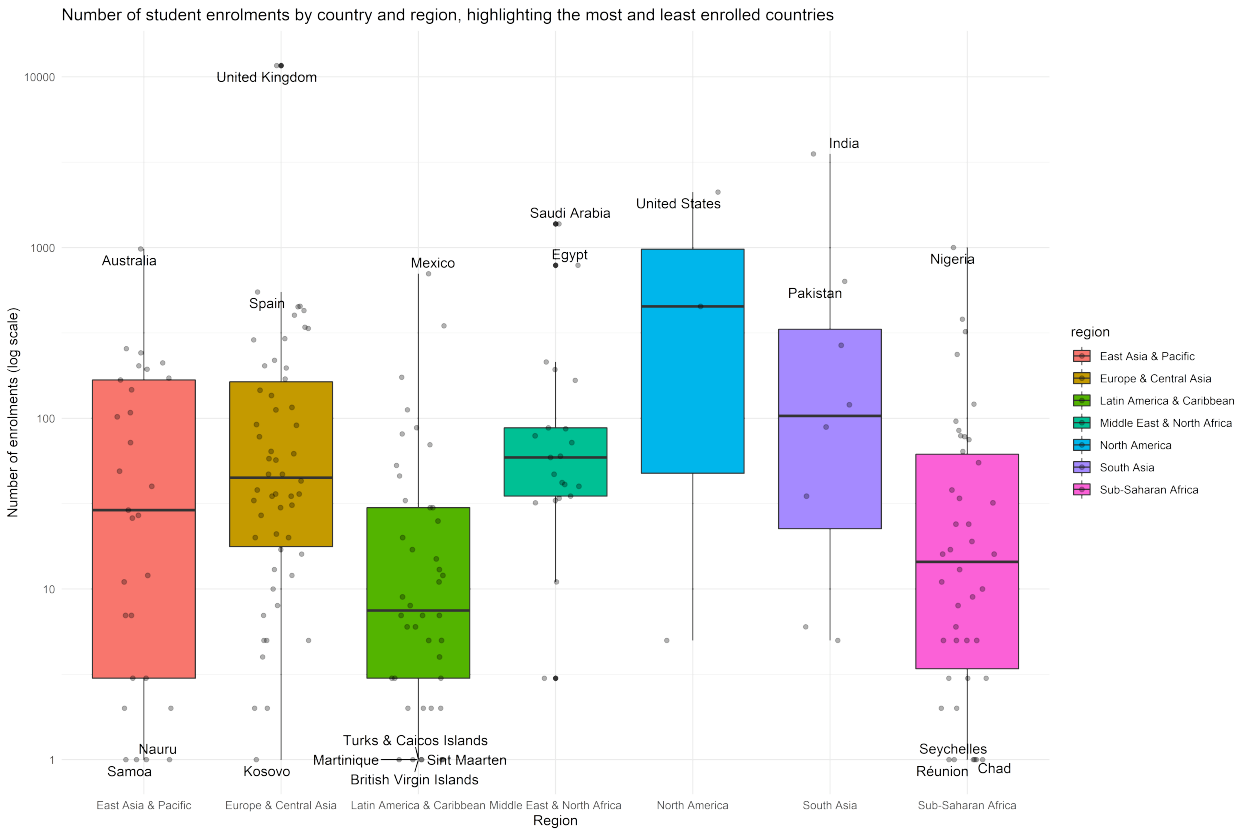
Detected Country	Number of Students
GB	11663
IN	3538
US	2117
SA	1376
NG	1002



The data was plotted which can be seen above, although it did not provide any insights into the data as it was difficult to see where each student is actually from. For example, it would be possible to guess from *IN* that this country was India. However, it is less clear which country *NG* is. In order to overcome this and get the country name, the R package countrycode was used. This allowed matching up the detected country code with the name of the country that the code belongs to and its region. Any missing countries or regions were checked online and custom added to the data. For example, the package could not match a country name for the code *XK*. A quick search online matched this code with *Kosovo*, which was then added to the dataframe. An example of the new table can be seen in table 2. This was then used to plot the data.

Table 3: Sample of Student Enrolments by Country and Region

Detected Country	Number of Students	Country	Region
GB	11663	United Kingdom	Europe & Central Asia
IN	3538	India	South Asia
US	2117	United States	North America
SA	1376	Saudi Arabia	Middle East & North Africa
NG	1002	Nigeria	Sub-Saharan Africa

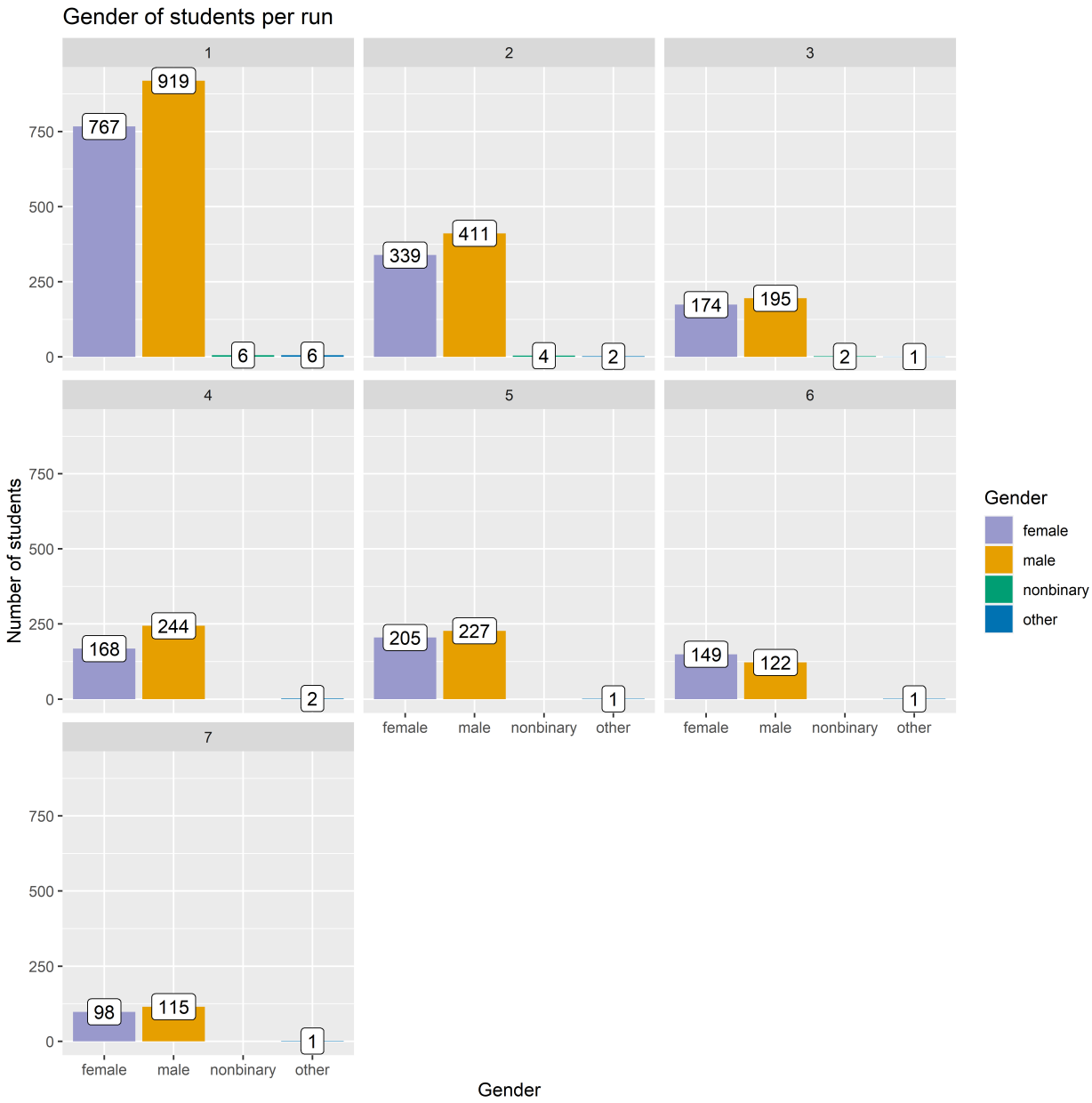


The second plot provides more information in that it displays that students from all regions of the world have enrolled on the course, and whilst there are slightly more students from North America, there is little variation between the regions. There are however, many variations between countries. The top 10 most and bottom 10 least enrolled countries have been labelled on the map. From Australia to Mexico to Nigeria, the countries with the largest numbers of enrollees are from all regions. However, the largest proportion of students derive from the United Kingdom (UK) at 32.09%. The global reach of the course is positive from a business perspective and in highlighting the countries with the most enrolments, there are opportunities for targeted marketing campaigns to students in these countries. However, the university is also strongly encouraged to promote courses widely in order to continue their global reach.

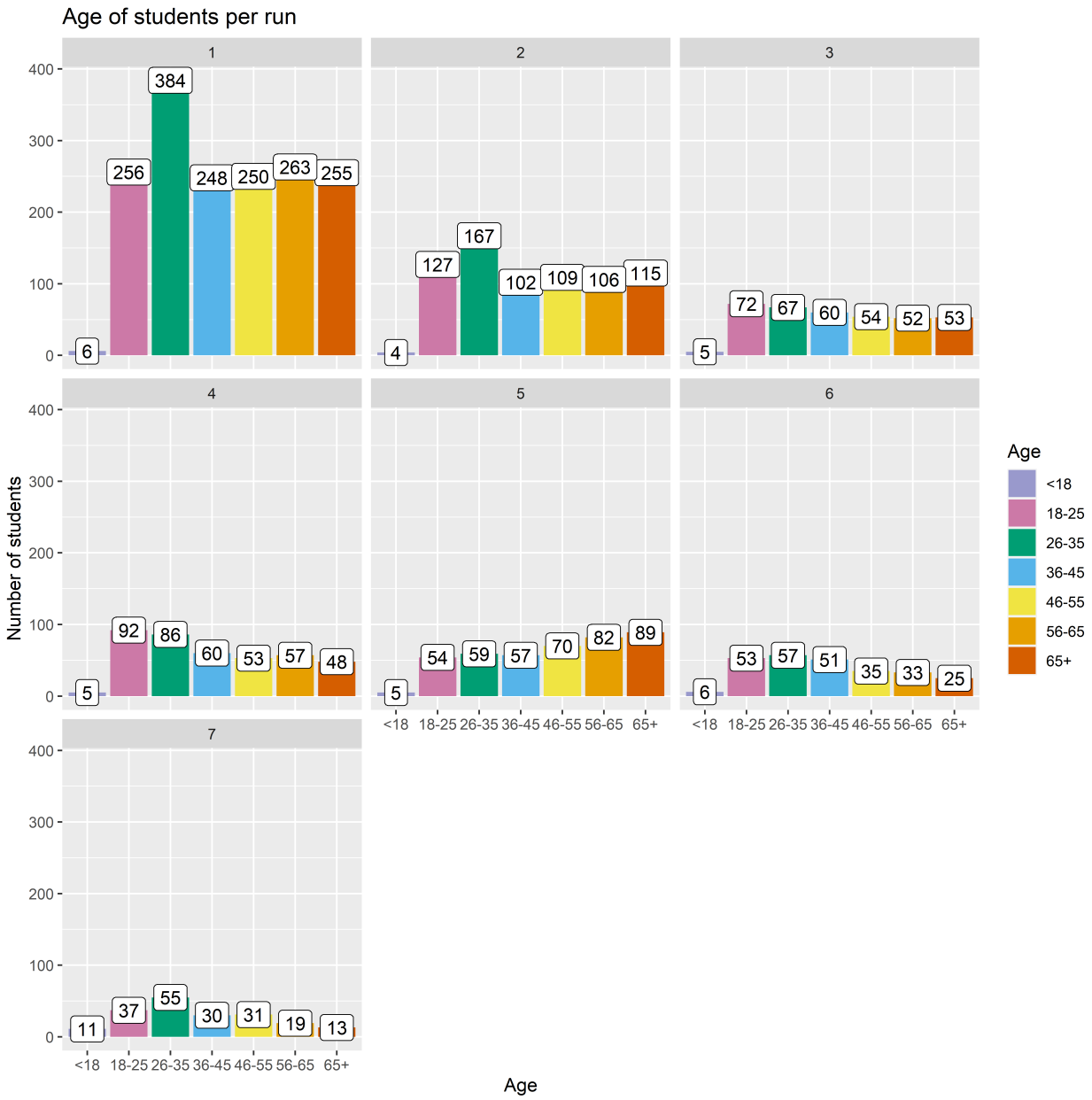
4.3. What are the student demographics and who should be targeted for future marketing perspectives?

When glancing through the *all_enrolments* dataset, there are four columns that are of interest when considering student demographics. These are *gender*, *age_range*, *highest_education_level* and *employment_status*. Whilst education level and employment do not typically display demographics, they can be seen as indicators of the demographics of students and provide a broader picture of who the learners are. Therefore, this information has been included for analysis. When analysing student demographic data, the data was grouped by run. This was to observe any changes in demographics between runs, for it may provide deeper insights as to why student enrolments have decreased over time. Additionally, all *Unknown* data was removed as pat of the data cleaning process.

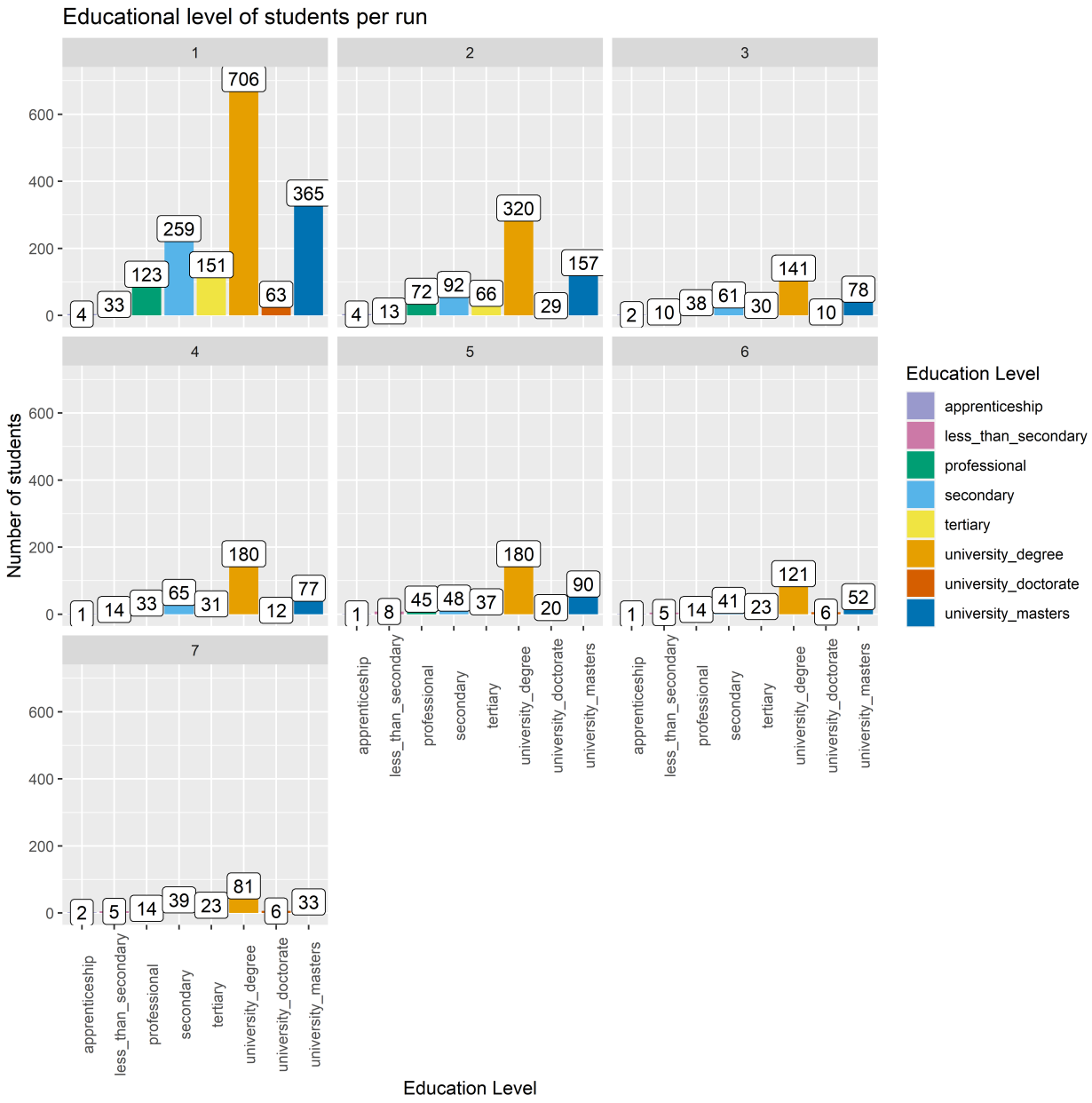
The first demographic which was explored was gender. After removing any unknown data, 4159 observations were found. Student gender had been split into four categories: *male*, *female*, *non-binary* and *other*. The plot below highlights this across each run. The majority of the students identified as male or female, with very few numbers identifying as non-binary or other. In all runs except for run 6, there were more males and than females. This displays that generally, there are likely to be more males than females on the course. However, this is not a significantly large difference.



Next, the age range of students was explored. After removing any unknown data, 4028 observations were found. This displays that 131 more students provided their gender than their age. This was then plotted, which can be seen below. The plot displays that all age ranges were represented in the course of all of the runs, with the lowest numbers of students aged under 18. In the first run, the majority of students were aged 26-35. However, this is not indicative of any of the other runs. For example, in run 5, the majority of students were aged 65+. However, again, this is not indicative of any of the other runs.



When analysing age range per run, there aren't any clear significant differences. Instead, the data has been collated into all runs to assess if that provides any further insights into age range. This is highlighted in table 4 below. This reinforces what the plot above shows, highlighting that only 1% of students were aged under 18 and that 21.72% of students were aged between 26-35 which was the largest proportion, but that all other age groups were fairly evenly distributed.



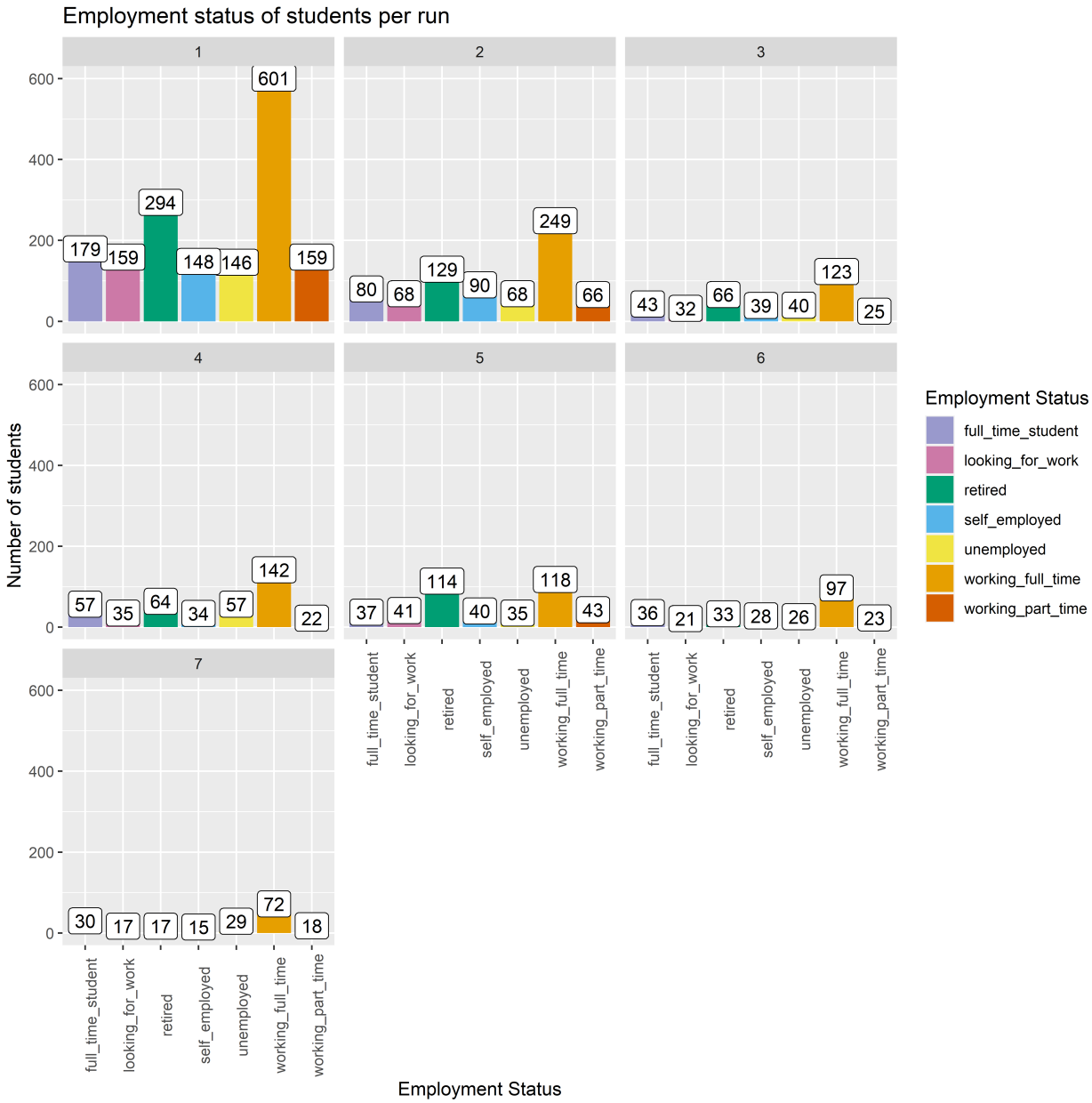
In analysing education level of students, 4135 observations were found after removing missing data. The data shows that the majority of students are educated from secondary up to doctorate level, with a very small proportion of students with less than secondary school education and at apprenticeship level. The vast majority of students are degree educated. This is consistent across all runs, and is reinforced in table 5, which highlights the proportions across all runs.

Table 5: Educational level of students across all runs

Education Level	Number of Students	Percentage %
apprenticeship	15	0.36
less_than_secondary	88	2.13
professional	339	8.20
secondary	605	14.63
tertiary	361	8.73

university_degree	1729	41.81
university_doctorate	146	3.53
university_masters	852	20.60

Lastly, 4105 observations were found of employment status. When analysing the levels of education, *not working* and *unemployed* were categorised as the same thing. Therefore these two were combined when preparing the data for analysis. However, *looking for work* was not considered as the same as being unemployed, as students can be employed and also looking for a job. The majority of students were in employment, whether working full-time, part-time or self-employed, with small proportions of students retired, unemployed and studying full time.



Whilst the data was significantly reduced through removing the missing values, there was still enough data to provide a sample of student demographics. However, the data has not provided many profound insights into the differences per run, nor who should be targeted for future marketing strategies.The typical student

is educated and employed, with slightly more male students than females. With the exception of under 18s, there were small differences across all other age groups. This is positive thing in that the students are not typical to one type of group. Although, non-binary students were disproportionately represented. In answering the question, who should be targeted for future marketing activities, the answer may lie with targeting educational institutions and places of employment. However, this would not be very inclusive and the data shows that students that do not represent the majority also enrol on the course. Therefore, for wider participation, there is an opportunity to also market course activities to non-binary students, those that are unemployed and looking for work and those with less than secondary school education.

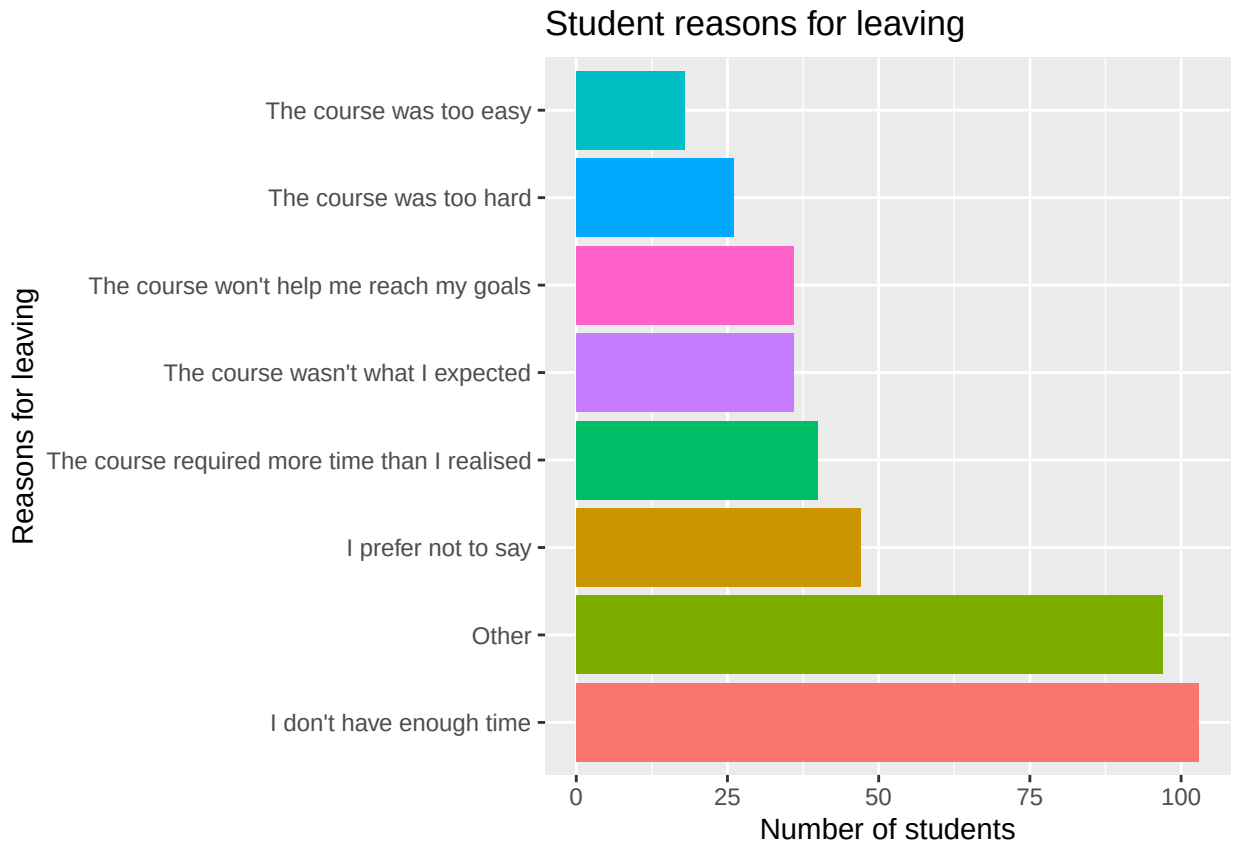
4.4. Further analysis: Reasons for leaving

An assumption of this analysis was that students would increase with each run. However, the data analysis found that student enrolment had decreased by 83.73%. Therefore, it is important to understand why students may not complete the course. The *Leaving Survey Responses* dataset may be able to provide that information, as it may provide an indication as to why student enrollment is low. There are 8 columns and the heading names are as follows:

```
## [1] "id"                          "learner_id"
## [3] "left_at"                     "leaving_reason"
## [5] "last_completed_step_at"      "last_completed_step"
## [7] "last_completed_week_number" "last_completed_step_number"
```

The *leaving_reason* column may be able to provide this information. The data available for *Leaving Survey Responses* is from runs 4 to 7, therefore this data was combined together with an additional column to differentiate between each run. This combined data had 403 observations. This is just 3% of the total number of enrolments, which is an extremely low percentage. Nonetheless, it may still provide some insights.

There were a total of 8 leaving reasons. However, the data responses were messy. For example, *I donâ€™t have enough time*. The data was then cleaned to remove the incorrect characters and input the correct characters. This was then plotted in the table below.



25.56% of students stated that they did not have enough time, whilst 35.73% of students did not provide reasons, choosing other and preferring not to say. The rest of the reasons for leaving are related to the course itself. As this was a very small sample set, more analysis will need to be carried out to further understand why students left the course. Unfortunately, this is outside of the timeline of this project, therefore would need to be considered further in future. However, if many students are concerned about time, the length of the steps of course may need to be a consideration in future.

5. Conclusion

The data analysis carried out in this project assessed a cyber security course run by Newcastle University on the platform Future Learn, and provided insights for Newcastle University on the course and their future marketing activities. This analysis found that student enrolment has decreased over 7 runs carried out during a 2 year process, thus highlighting how important it is to understand the learners and their needs in order to improve student recruitment and attainment. The course has a global reach and students have joined from 199 countries, which is very positive. Whilst the typical student tends to be educated and in employment, there are opportunities for wider participation as atypical students also enrol on the course such as those that are unemployed. Therefore, the university can market the course to a wider audience. There is also a need to further understand why students are leaving the course, or not enrolling on courses, which is something that the university should consider investigating further.

6. References

Higher Education Commission (2016) *From bricks to clicks: The potential of data analytics in higher education.* Policy Connect, London. Available at: https://www.policyconnect.org.uk/research/report-bricks-clicks-potential-data-and-analytics-higher-education (Last Accessed: 01/12/2021)