Data Management & Exploratory Analysis

Newcastle University - CSC8631 Summative Assignment

December 2020

Summary

This report provides additional documentation detailing the findings from my exploratory analysis. It is intended to be read alongside the Design and Implementation Report.

A dataset of numerous files in a variety of formats, compressed into a single folder, was provided by the Newcastle programme team for analysis. The material related to a series of online cyber security training courses, entitled Cyber Security: Safety at Home, Online, in Life. This report summarises the data management and exploratory analysis undertaken using the enrolment files. Given that the training subject was online security I was interested to determine students' willingness to share information relating to gender, employment and education.

Seven enrolment files were provided for the courses that ran over a period of two years. The first student enrolled on 29 March 2016, the last student enrolled on 01 November 2018 indicating that the course series for a period of over two and a half years.

Business Understanding

As the online training course contents related to Cyber Security I was particularly interested to determine students' attitudes to sharing data. Although the file contents were anonymised, items captured within the enrolment survey file included gender, age range, employment status, employment area and highest education level.

I returned to this phase once I had looked at the data and determined that I was interested in the enrolments files for the online cyber security course, particularly information provided on personal data. I assumed this data had been provided by students before the training had commenced and was truthful. I removed duplicate records from the combined file, assuming learner_id was unique.

Data Understanding

During the Data Preparation it became clear there were duplicate learner_id records when the seven files were combined.I returned to this phase to determine how to handle duplicate records and decided the analysis would benefit from additional information about the number of students on each course, dates for first and last enrolment record and course duration.

Data Preparation

I initially used R and RStudio to start reviewing the data contained within the enrolment files. I created additional columns, some of which were later discarded, combined the files and generated initial statistics. Once I was comfortable with the initial data pre-processing steps I created a new R Project document, loaded the ProjectTemplate library and created a new project. Running the load.project() script in R created a project folder structure and loaded R packages.

The ProjectTemplate was particularly helpful in organising the project files and determining important content. I found the config file particularly helpful to manage the loading of R packages, project settings, e.g. caching data and running pre-processing scripts. I didn't use all of the folders though and an element of personal discretion was required to determine what was stored where.

The pre-processing scripts (or data munging code) for the project are stored in the 'munge' folder. These pre-processing scripts stored in `munge` will be executed sequentially when load.project()' is called. Numbers within filenames indicate the sequential order of the scripts. The scripts:

- add columns at runtime, merge data sets and adjust data formats
- remove duplicate records where learner_id is not unique, calculating course duration between first and last enrolment date
- filters the data by groups of students willing to share data
- categorises data
- provides additional data items on duplicate records
- calculates volume of shared data items
- determines counts and percentages of data that it shared by student, by survey
- and calculates counts and percentage of data that is not shared

Modelling

I used the RMarkdown file to generate a report to be stored in the Reports section of the project file structure. During the generation of graphs and tables I returned to the data preparation phase frequently. After working for some time on the enrolment files I realised that it was necessary to calculate the number of students to determine the percentage of students sharing data. I added the 'Key Findings' section of this report following this discovery. Regularly, I would update the Git version control by using a Commit and Push commands. Particular attention was needed to ensure titles and axis were labelled consistently.

Evaluation

At this stage I took time to thoroughly evaluate my work, creating the summary reports and presentation which reviews the steps executed. I checked there was a sufficient story to tell, regarding all the tools and techniques used. During this phase, I decided to update some of my analysis. At this stage I amended colour templates to be colour-blind friendly and made sure my graph formatting was consistent.

Deployment

The deployment phase of this project included publishing a final version of the project to GitHub and ensuring my project files could be compressed and submitted in advance of the assessment deadline. The use of RMarkdown reports, README files and comments within the R scripts increases transparency, allows the analysis to be reproduced or repeated.

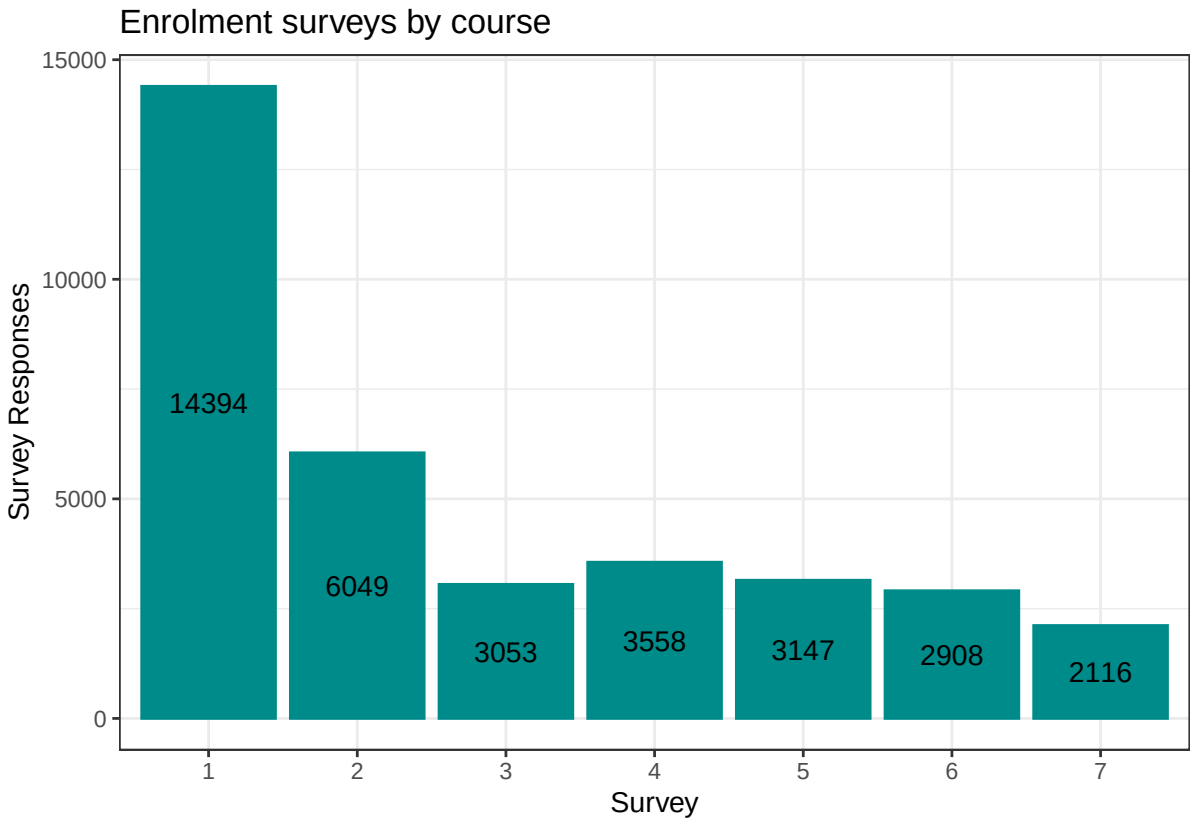
Key Findings

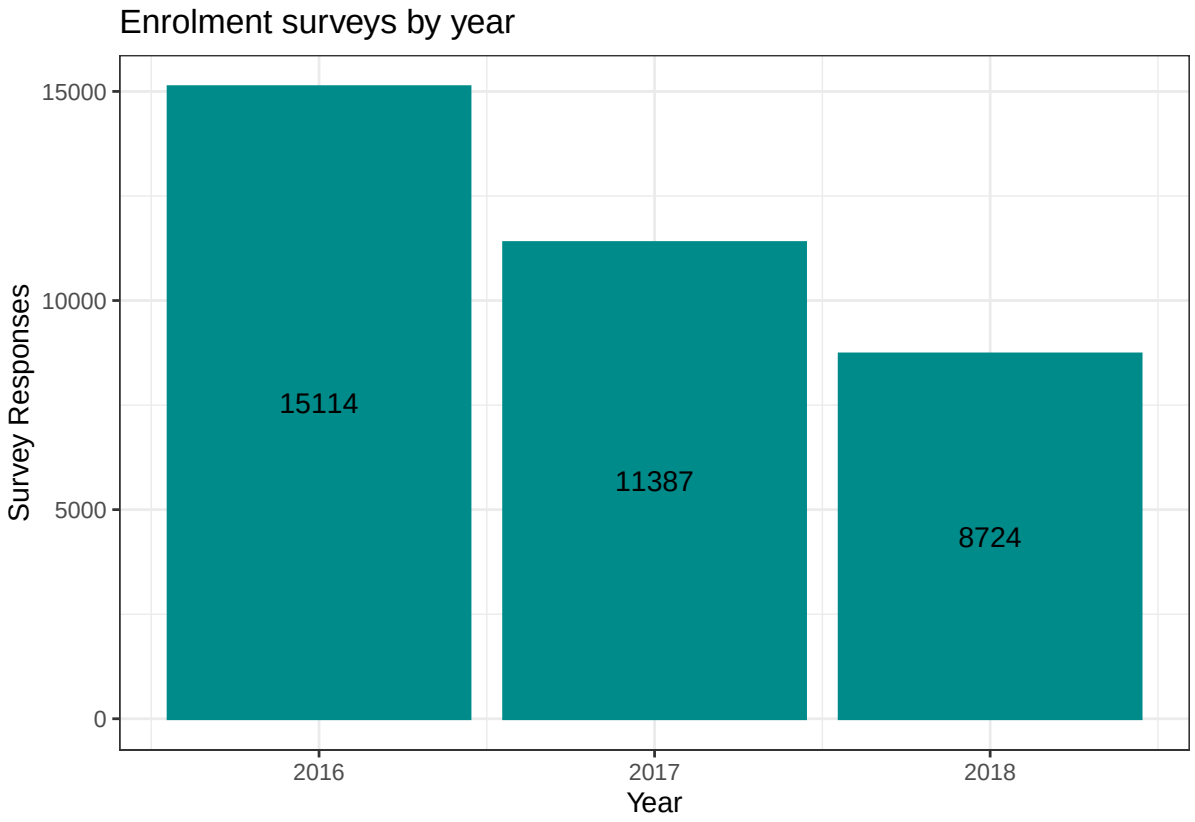
Course timings, duration and student numbers

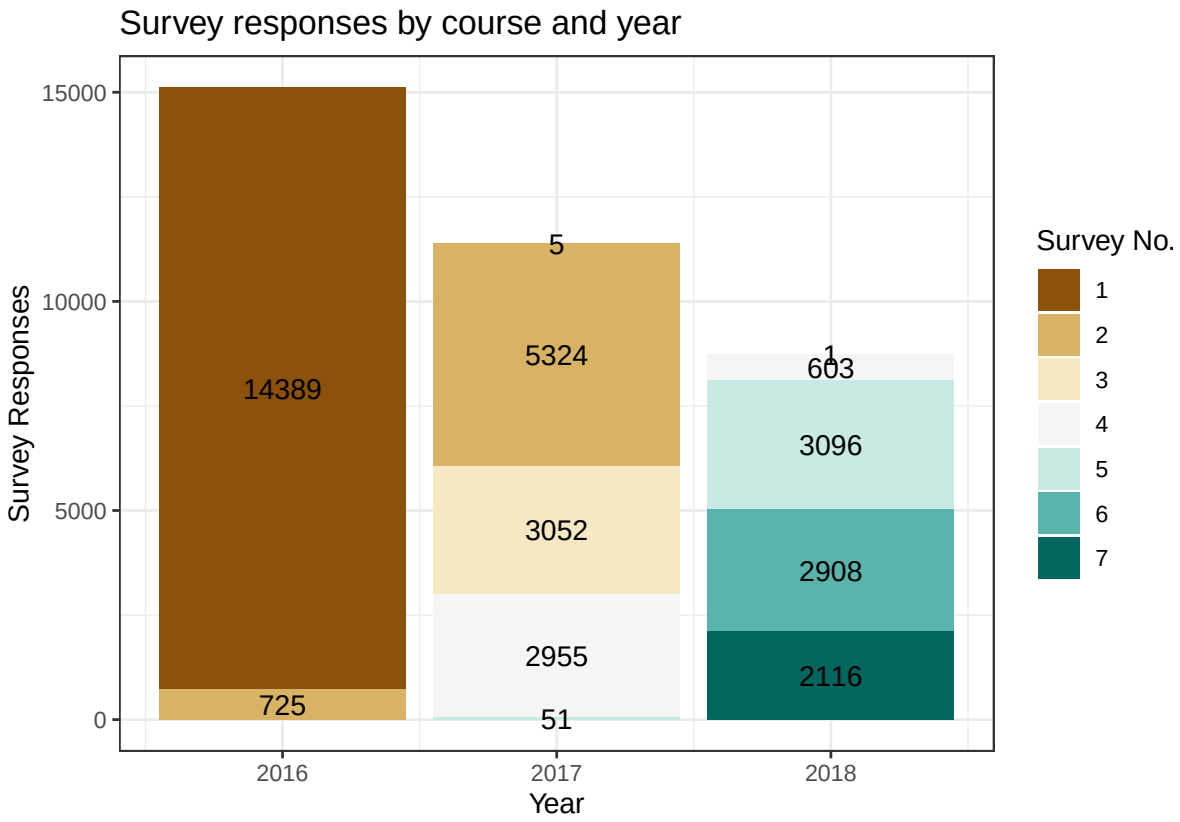
In total, 35225 unique student enrolment records were assessed from seven survey files.

Survey No.	Date First Student Enrolled	Date Last Student Enrolled	Duration (days)	No. of Students
1	29 March 2016	07 September 2017	527 days	14394
2	05 December 2016	13 July 2017	220 days	6049
3	02 July 2017	26 February 2018	239 days	3053
4	27 July 2017	25 January 2018	182 days	3558
5	15 December 2017	09 September 2018	268 days	3147
6	08 April 2018	11 August 2018	125 days	2908
7	25 June 2018	01 November 2018	129 days	2116

The ‘enrolled at’ dates within the files indicate that the courses overlapped in duration and varied in length, with the volumes of students applying for each course gradually reducing.



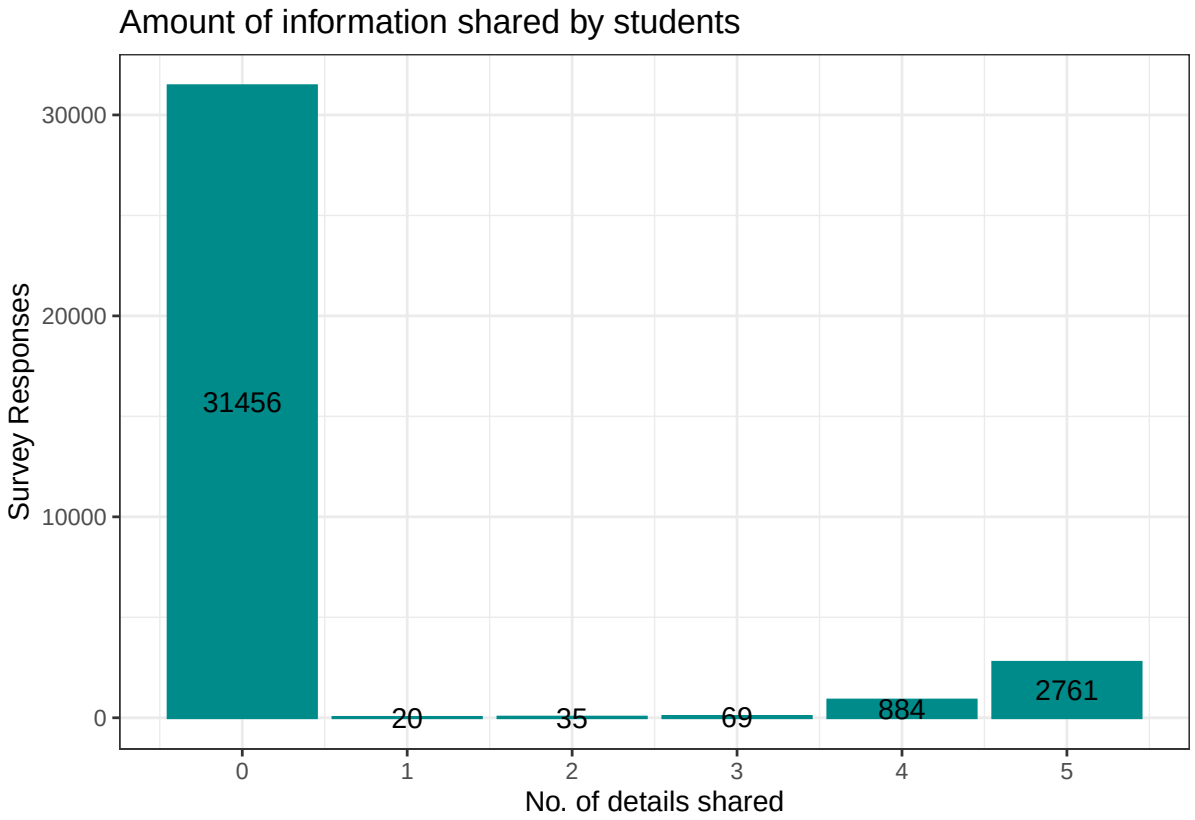




Course intake dropped by c25% annually.

Amount of Data Shared by Students

Students were asked for information on gender, age range, employment status, employment area and highest education level. Analysis was conducted on these five details to determine willingness to share.



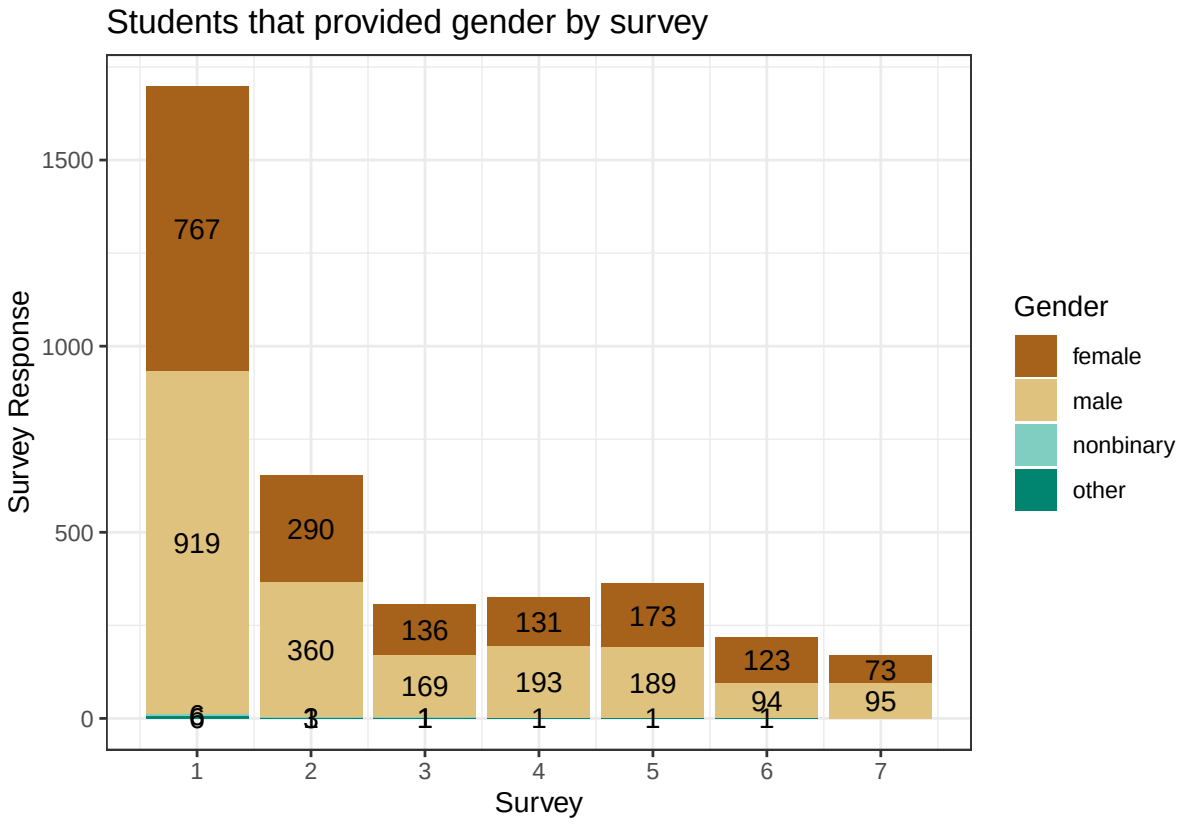
The total number of students that shared data at least one detail was 3769, which equates to 10.70% of total students.

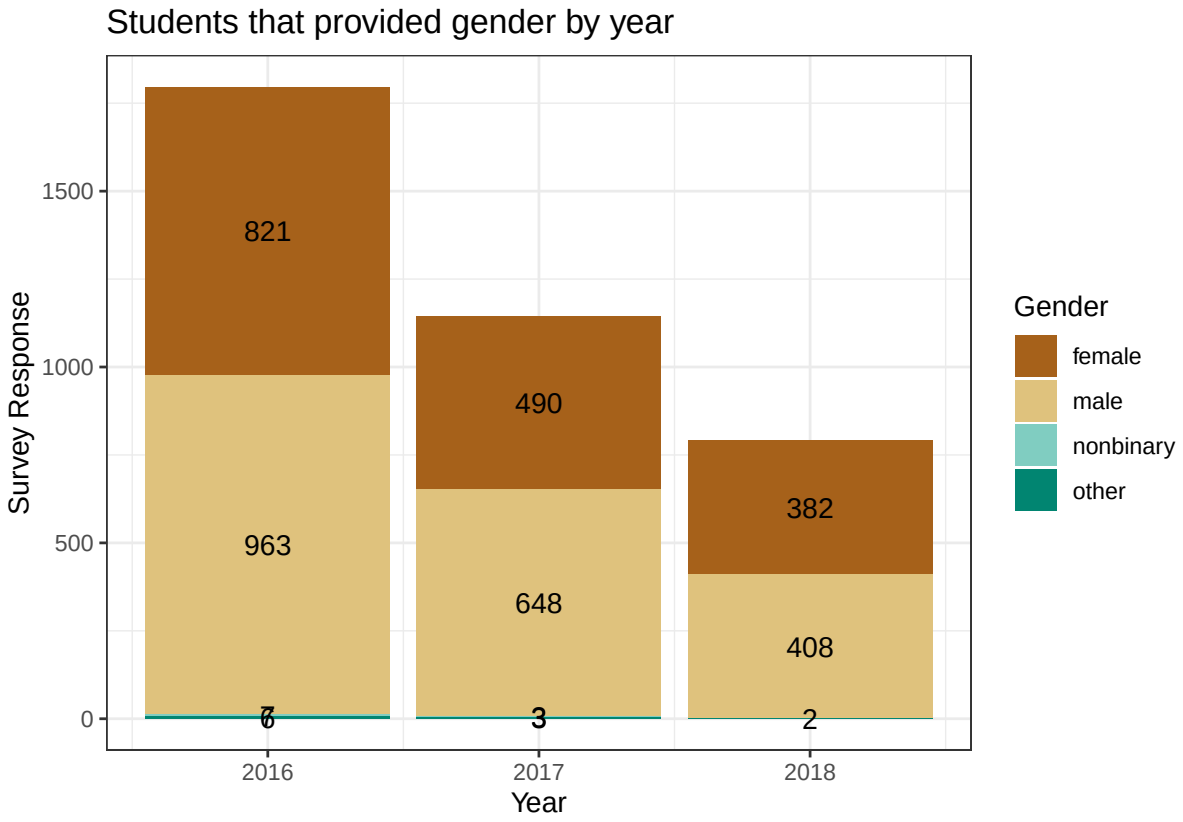
Survey No.	% of Students *	Ave. Items *
1	11.98%	4.69
2	10.91%	4.68
3	10.15%	4.68
4	9.13%	4.71
5	11.57%	4.65
6	7.50%	4.67
7	7.94%	4.57

* where students shared at least one of the five lifestyle details

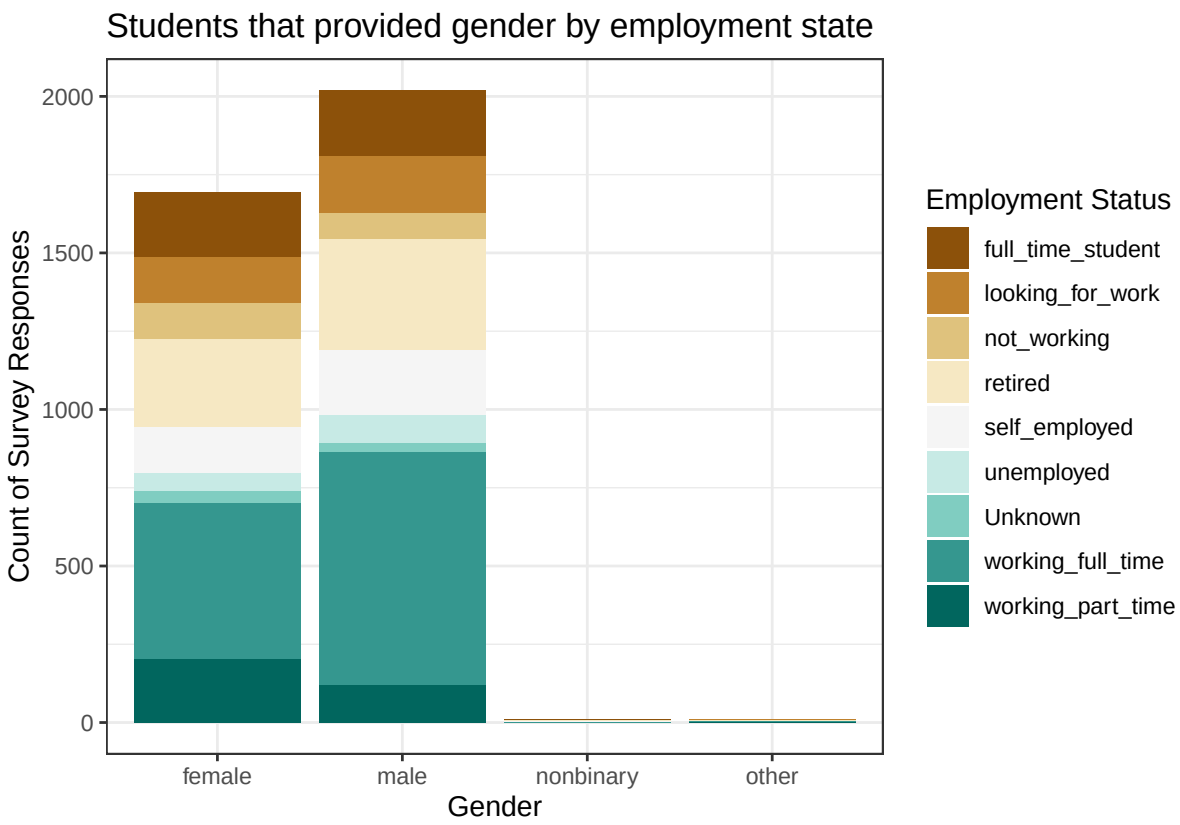
Over the seven enrolments, the percentage of students sharing data dropped from almost 12% to under 8%. However for those students that did share data the amount of details they provided stayed fairly level at around an average of 4.7 items per student, dropping very marginally in the last intake to 4.6. Where students did share at least one, 73.26% of students provide all five requested details.

Students willing to share data dropped over time. Males were slightly more likely to share details than females, although this changed slightly in the last two course intakes.

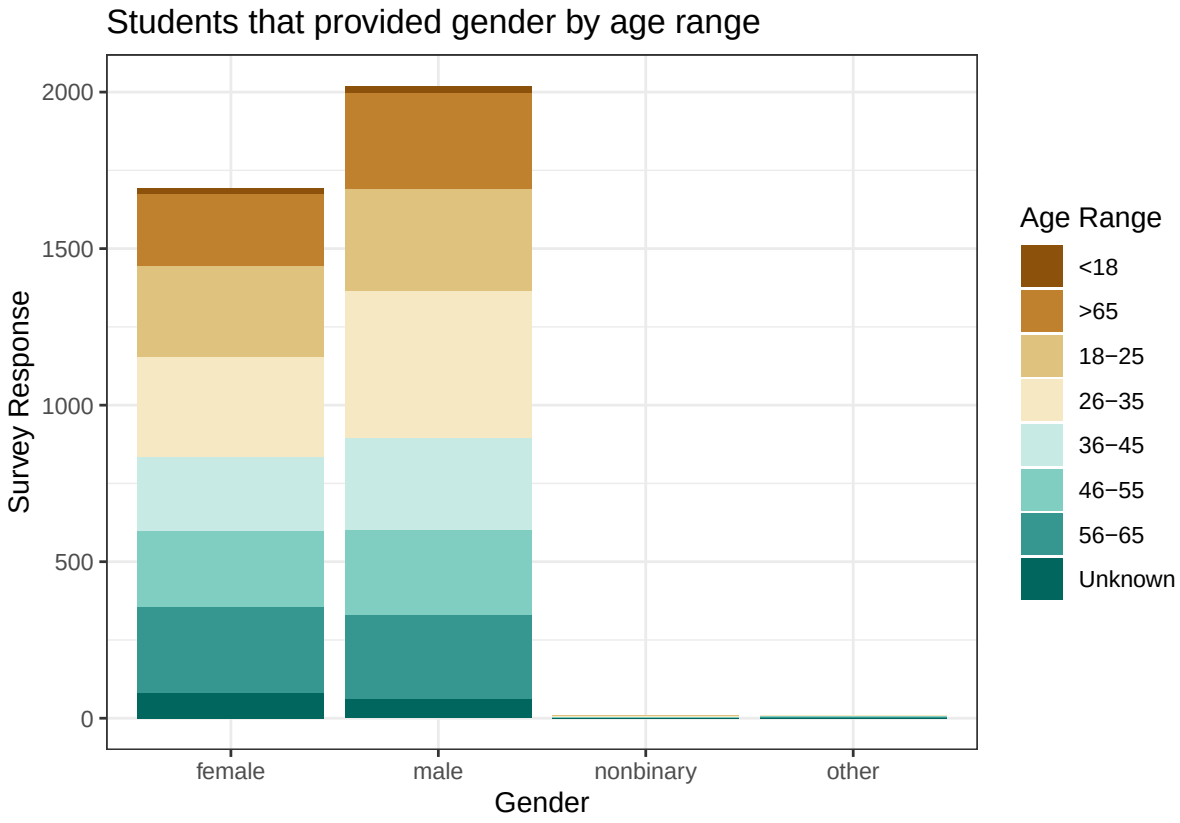




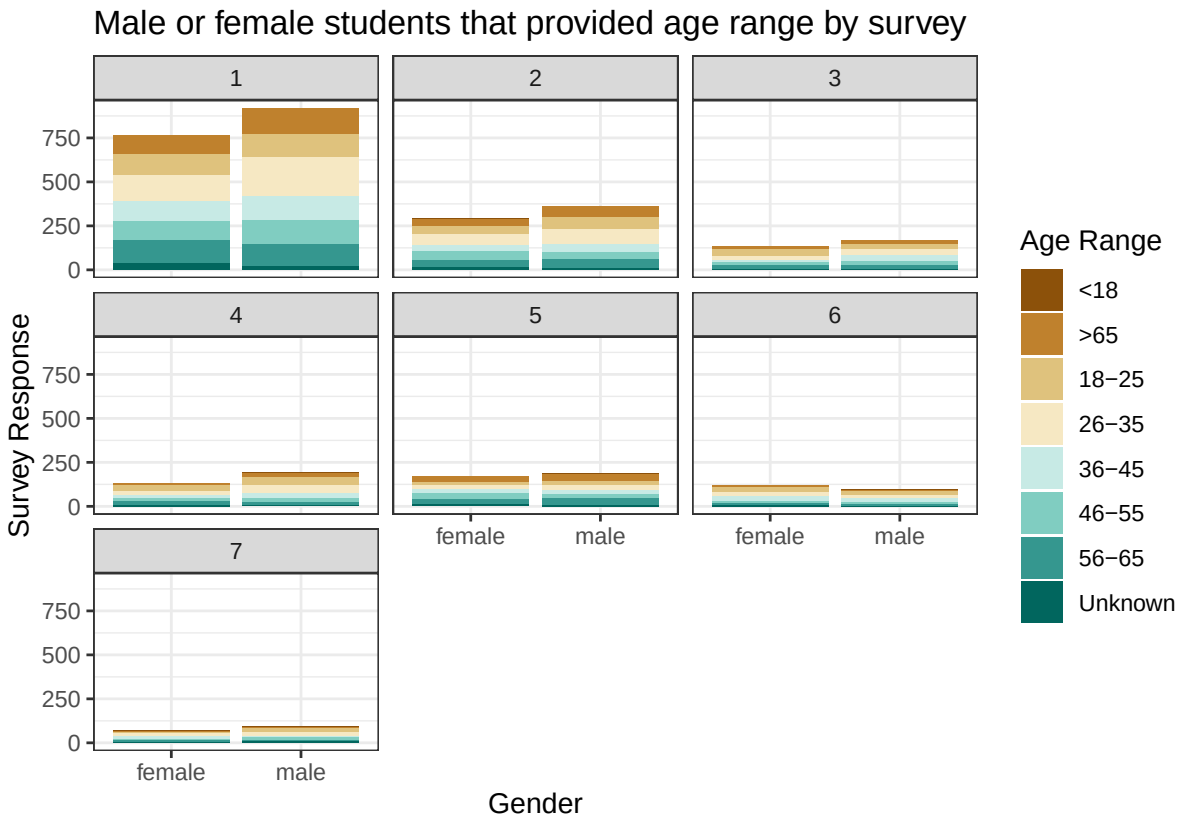
Male and females working for home were the largest groups willing to share gender.

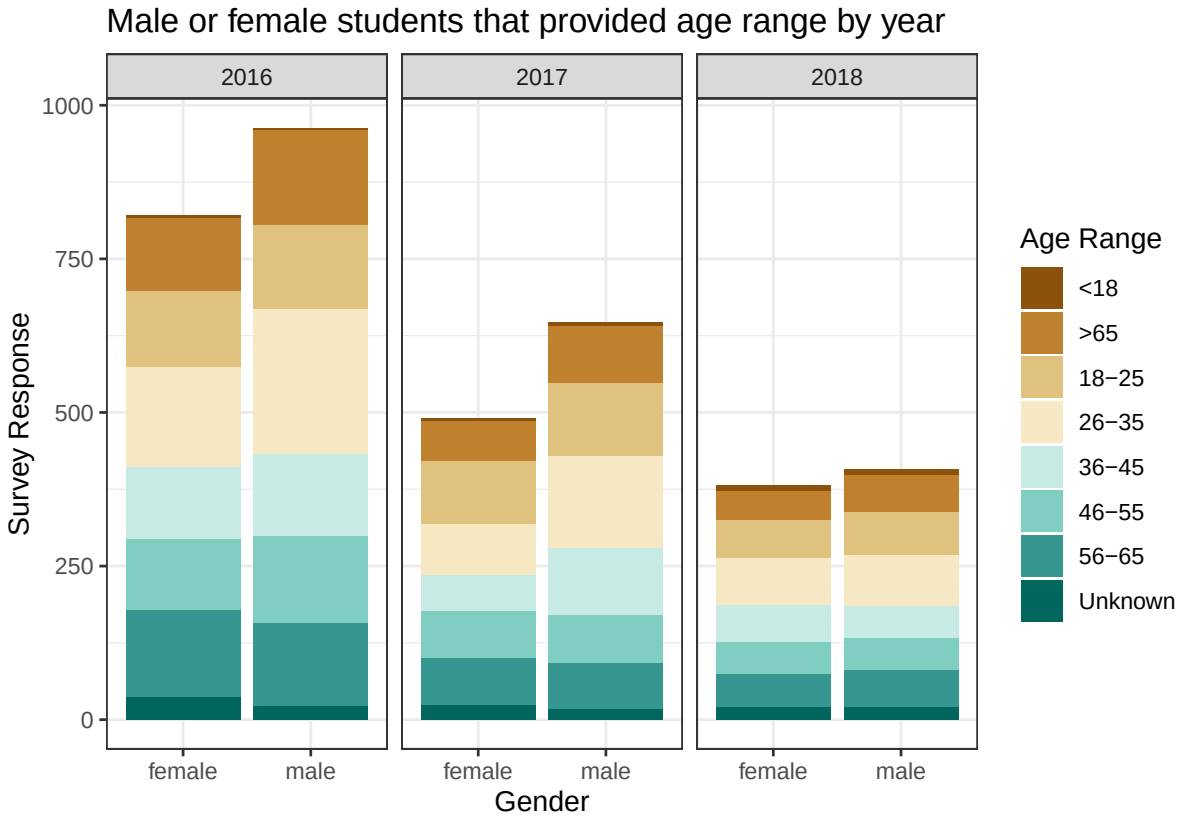


Students willing to share information about gender were split evenly across age range group, with males 26 -35 slightly more likely to share. Noticably there was a subset willing to share information on gender but not age range.

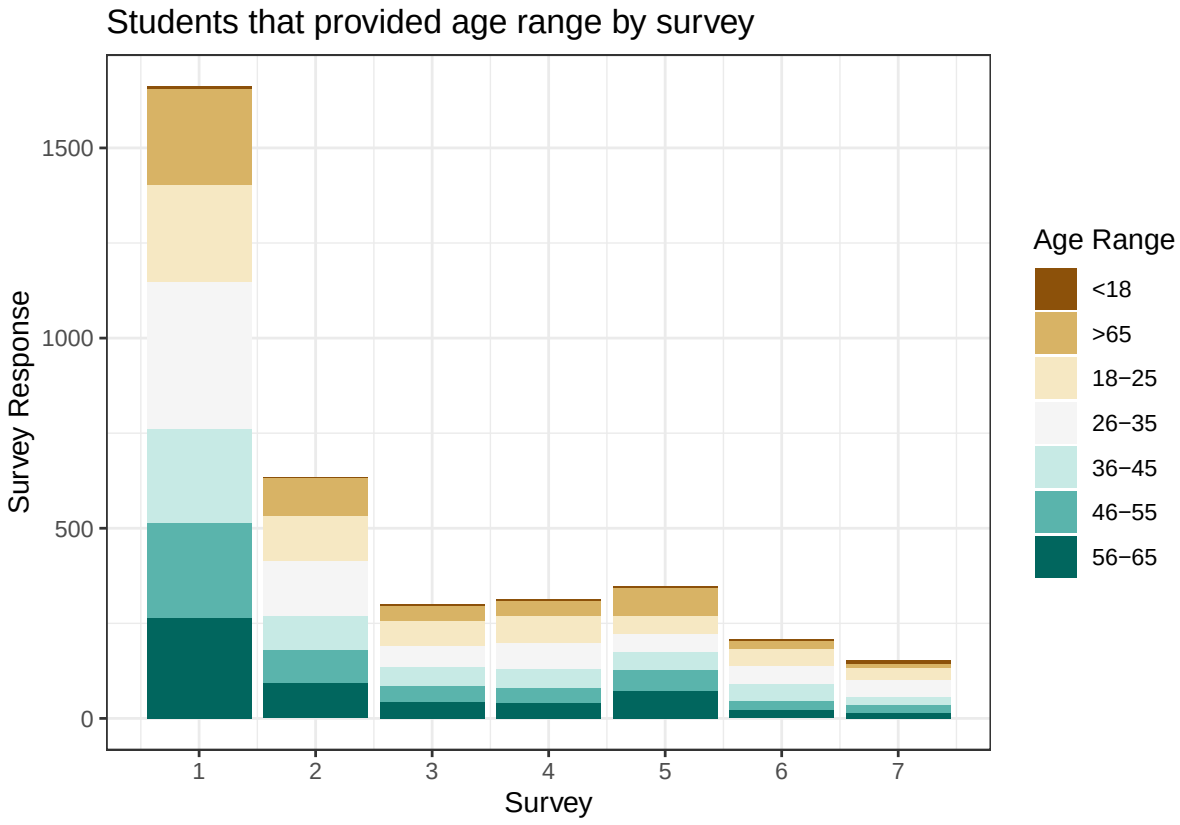


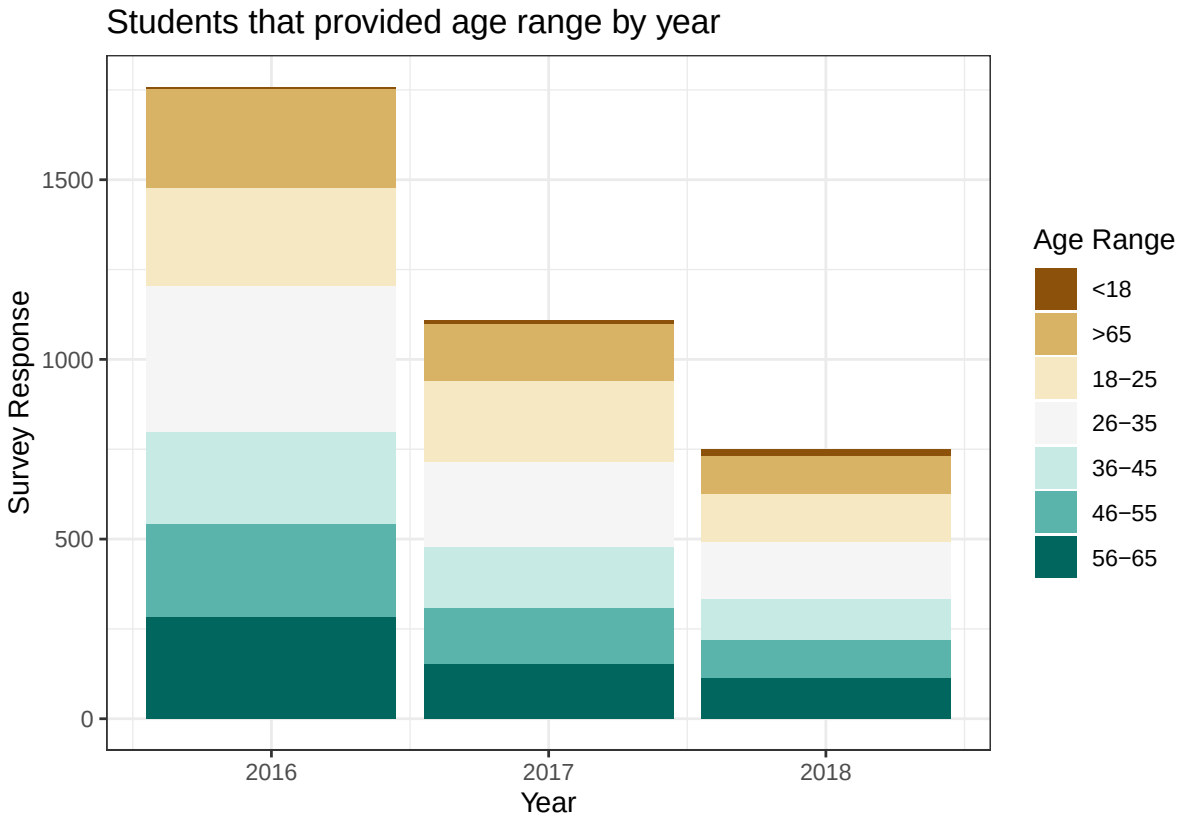
Attitudes on sharing age and gender declined by year and survey. Initial analysis would indicate data sharing dropped more than course enrolment by year, i.e. greater than a 25% decline. However more analysis would be required.



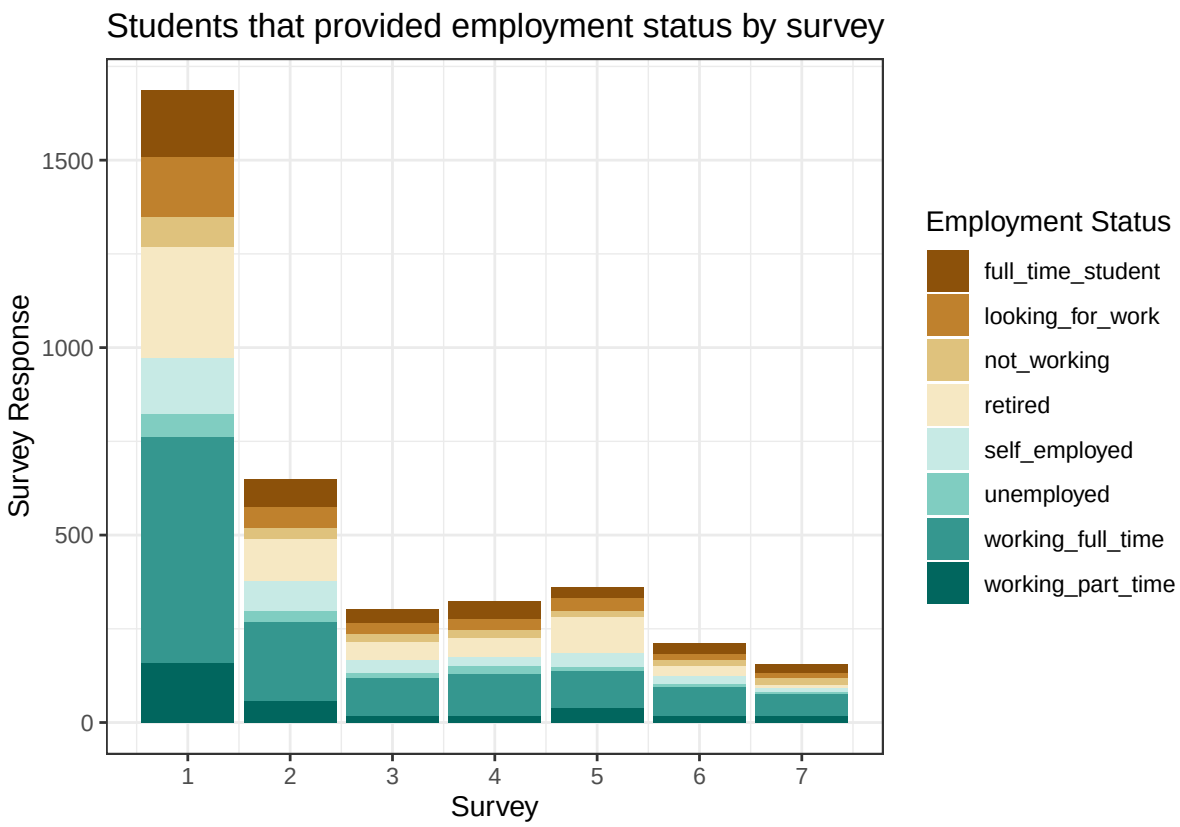


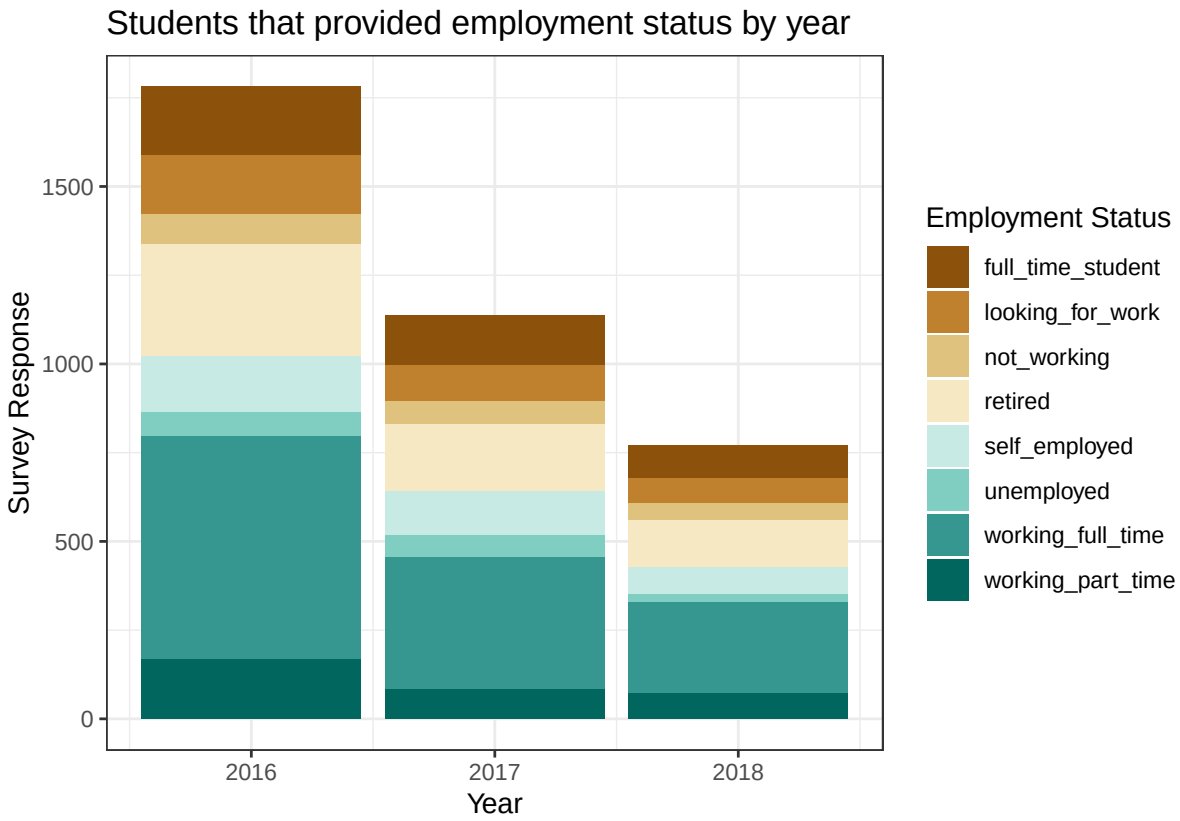
The age information shared by students showed attitudes to sharing were evenly spread across age brackets and dropped annually.



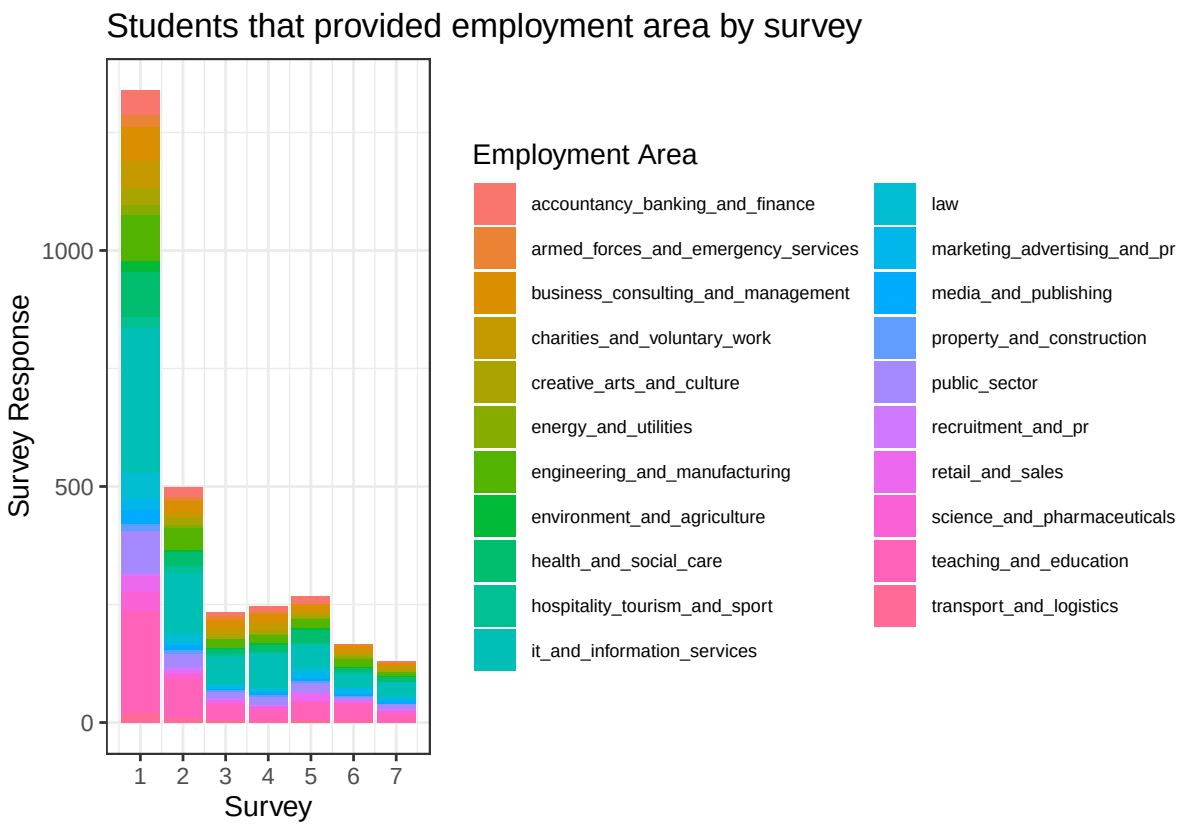


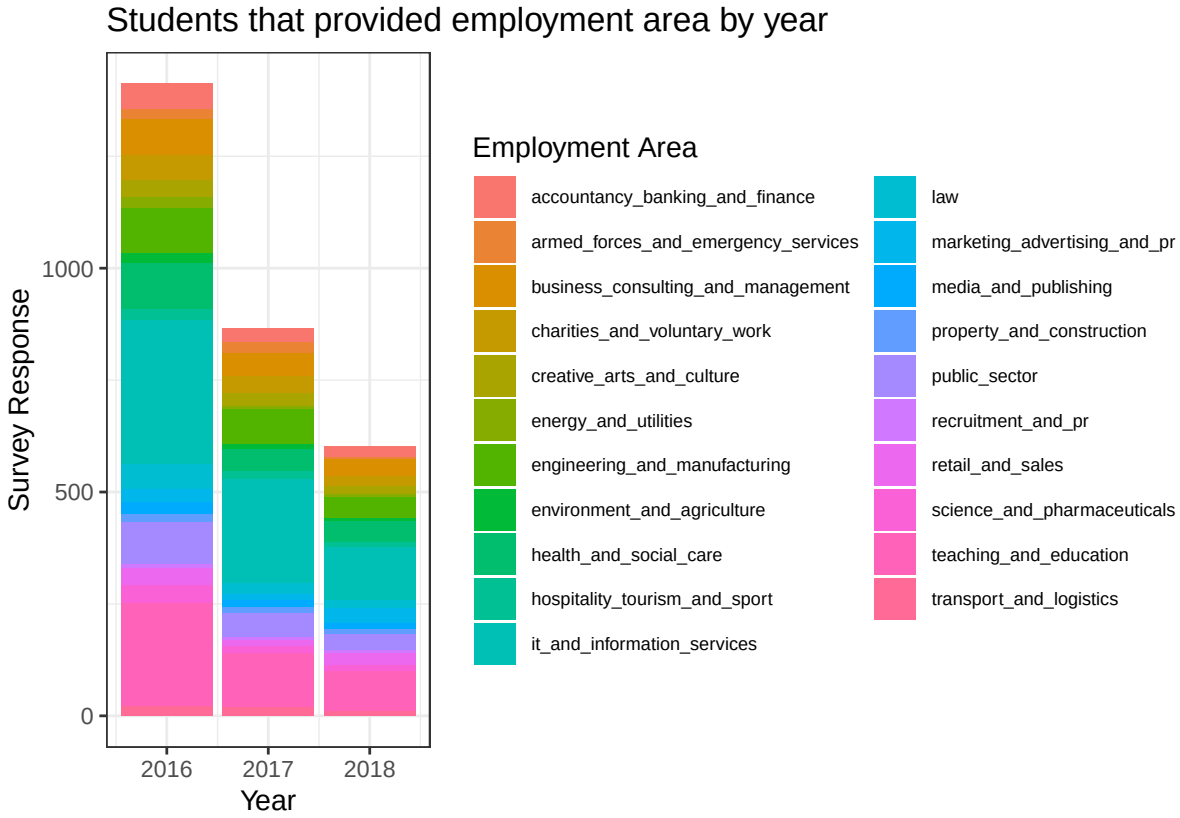
The employment status information shared by students was particularly varied and dropped annually.



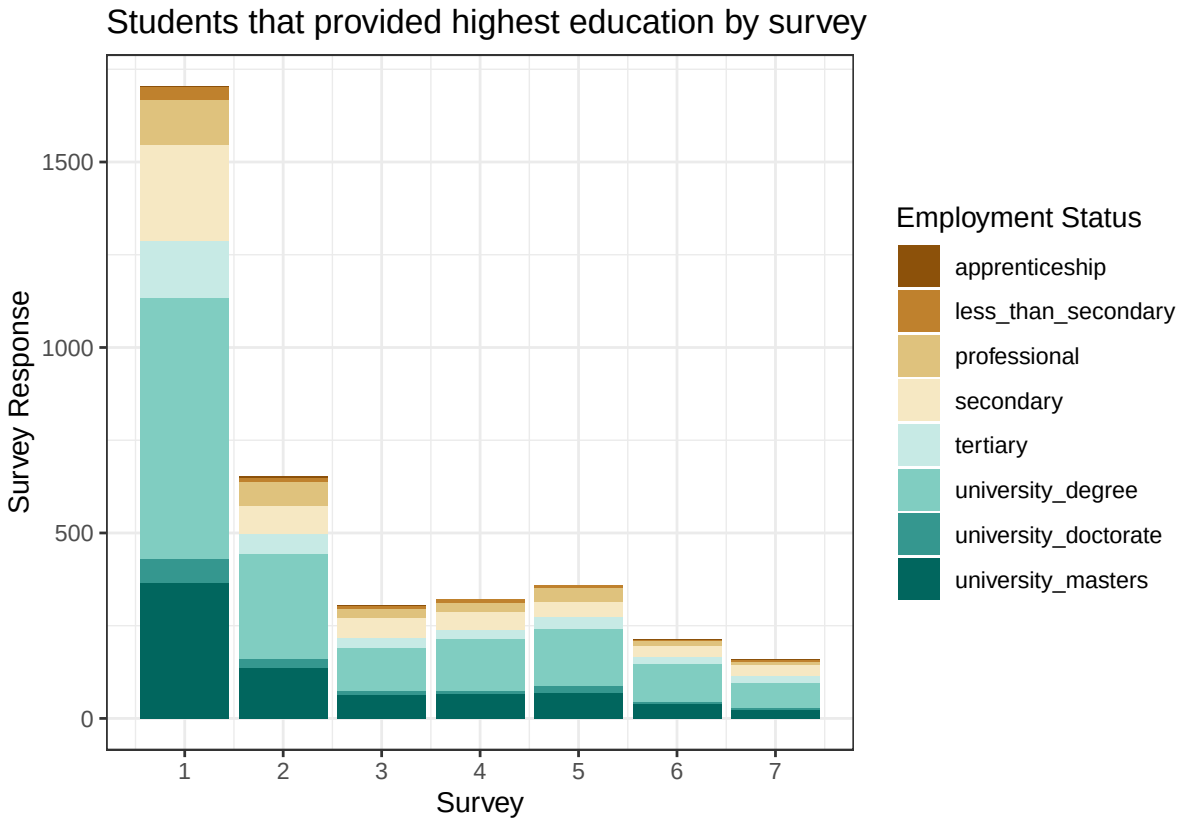


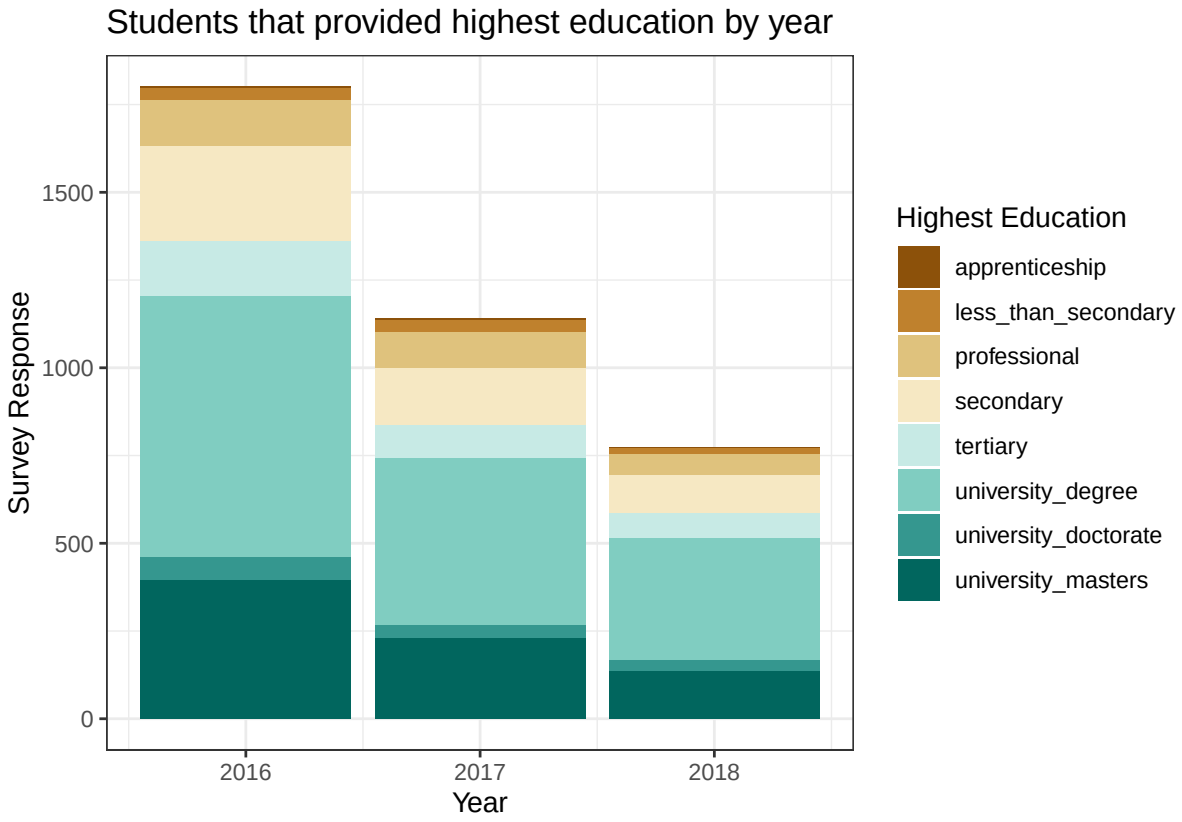
The employment area information shared by students was particularly varied and dropped annually.





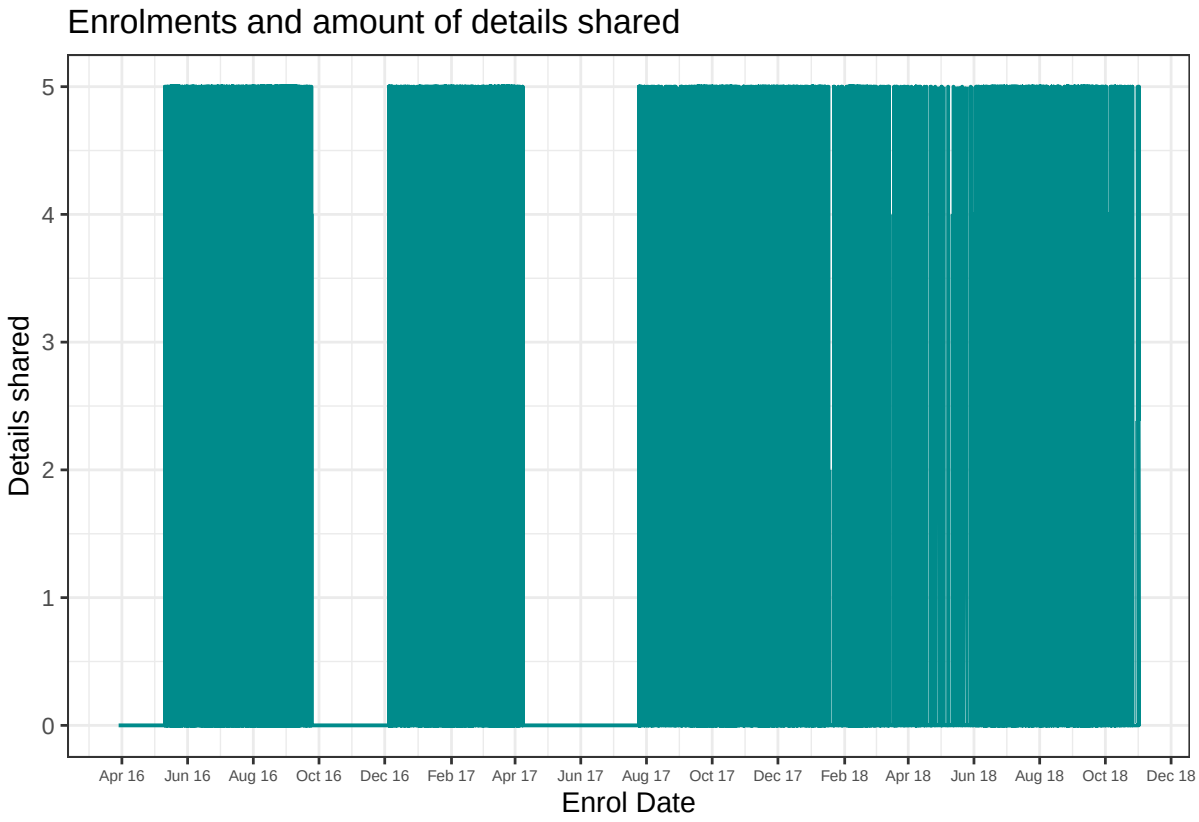
The highest level of education information shared by students indicated a larger group with degree education. Students sharing information on education dropped annually.

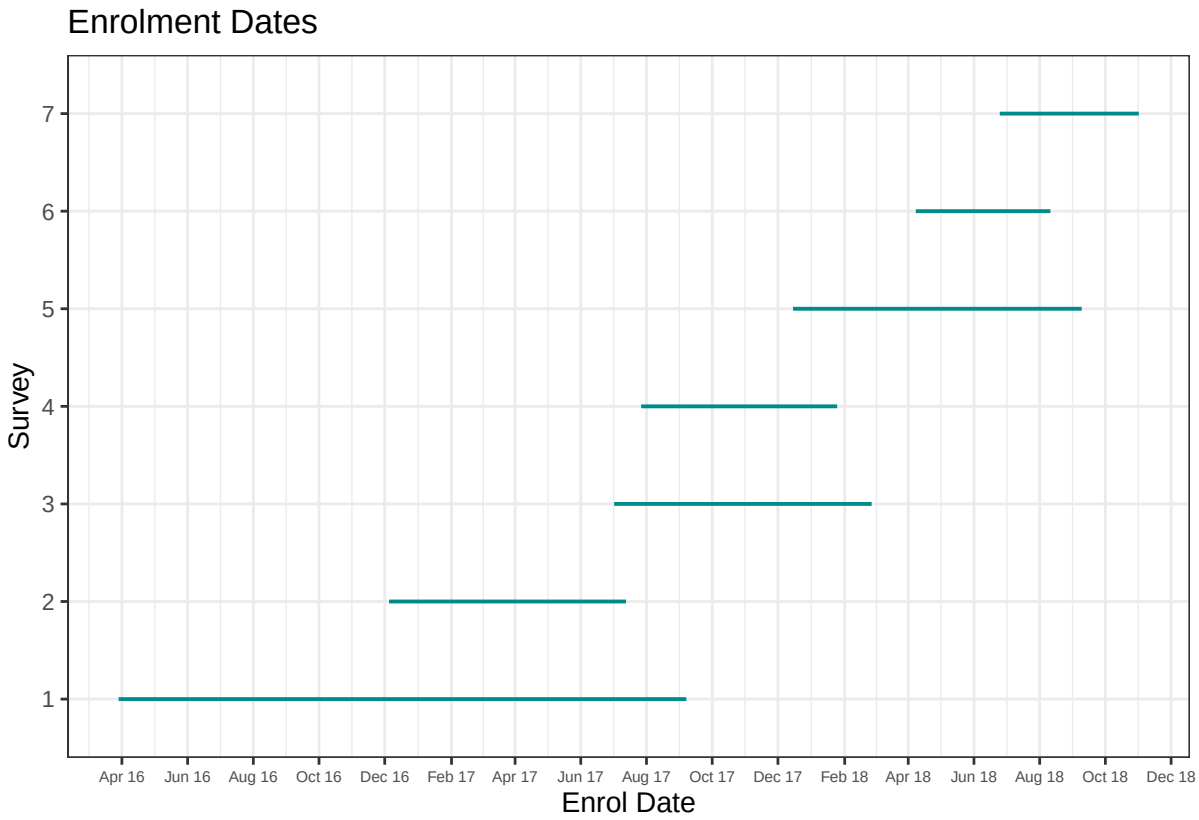




#

The pattern of students sharing information might indicate that there were changes to the enrolment survey settings, either to encourage students to provide information or that these fields were mandatory.



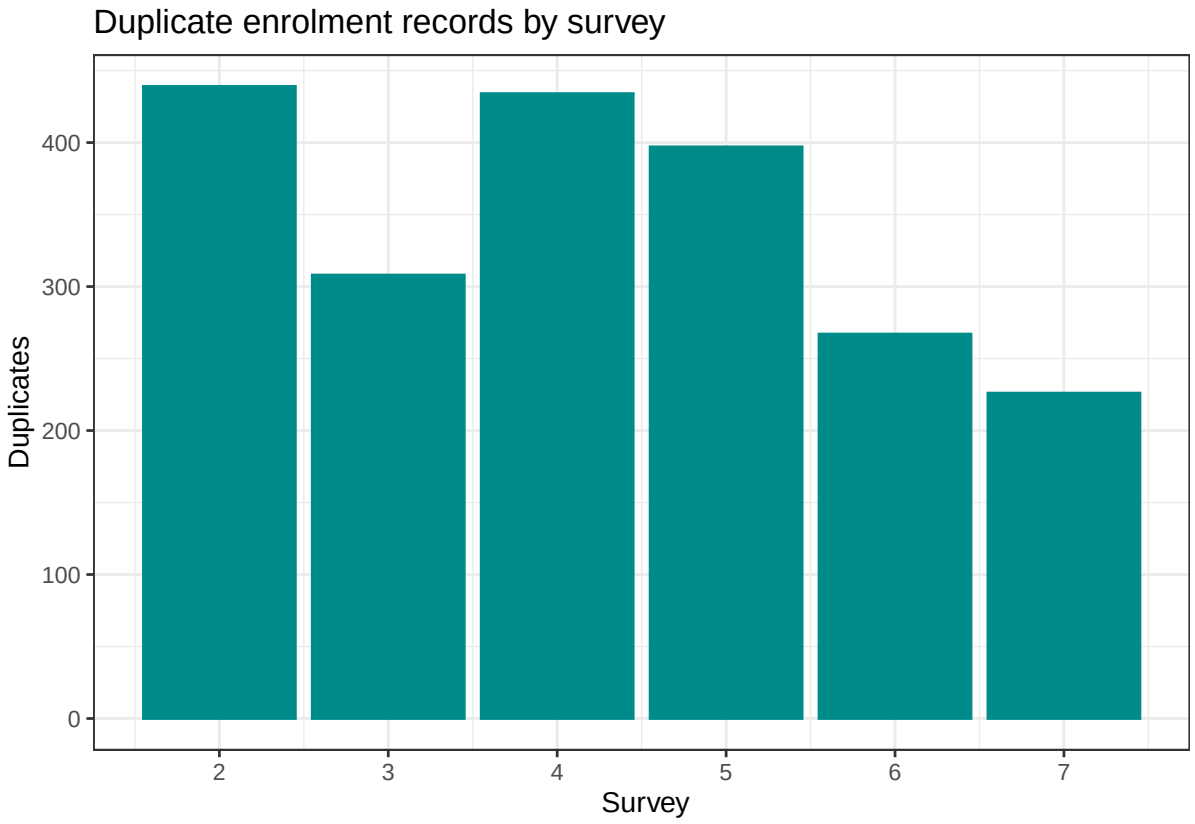


Conclusions

Over the course series, students shared less data. The most commonly shared data item was gender with 3733 students sharing this information, 10.60% of the total students enrolled. 3715 students shared highest education level, 3689 shared employment status,3620 shared age range and 2881 shared employment area.

Assumptions

- It is assumed that duplicated learner_id records are exact copies of the same survey response. In total 2071 non-unique learner_id records were removed from the dataset prior to analysis. It was expected that the learner_id being 36 characters in length was expected to be unique and represents a single student. Example, 233b9253-e8e6-4734-8a7f-cb2c7845c85a.



- It is assumed that each record within the cyber security enrolments files corresponds to a student who provided details that truthfully reflected their gender, employment status and area, age range and highest educational level.
- It is assumed that the data contained within the cyber security enrolments files was provided before the training had commenced.
- It is assumed that each enrolment file relates to a course intake.
- It is assumed that the order of questions was not changed over time and the sequence of the request had no impact on the students' willingness to share data.
- It is assumed that the earliest 'enrolled at' dates within the file indicate the intake start date and the latest date is the course closure. This would indicate that whilst enrolments within file four started after the third intake it was a shorter course.

Notes

I was mindful to apply the learning from the first module of the Masters degree course, Data Visualization - CSC8626. In particular the Gestalt design principles of Proximity, Repetition, Alignment and Contrast.

“Colors are an effective medium for communicating meaning. Well-chosen colors reduce the time to insight for your viewers and helps them understand your message sooner and more easily.” Source: https://www.forbes.com/sites/evamurray/2019/03/22/the-importance-of-color-in-data-visualizations/?sh=133b1c7a57ec Forbes, published 22 March 2019, Eva Murray, article title “The Importance Of Color In Data Visualizations

“Titles and text are key elements in a visualization and help recall the message.” Beyond Memorability: Visualization Recognition and Recall Michelle A. Borkin

“A visual channel is a way to control the appearance of marks, independent of the dimensionality of the geometric primitive. There are many different visual channels […], such as position, color, shape, and size, etc..” Source: https://jenniewblog.wordpress.com/2016/03/08/marks-and-channels-chapter5/ Published March 8, 2016

Colour, contrast and brightness are very important as “visual information is processed long before it reaches the brain” by the retina. “The very first level of the visual system, the retina, already provides information on colour, contrast, movement, and brightness. We notice individual objects ‘at a glance’ because they jump out from what we see as mere background.” Source:https://www.sciencedaily.com/releases/2017/02/170208151226.htm Science Daily, published Feburary 8, 2017, article title “Zapping between channels in the retina”