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

Modelling

Evaluation

Deployment

Key Findings

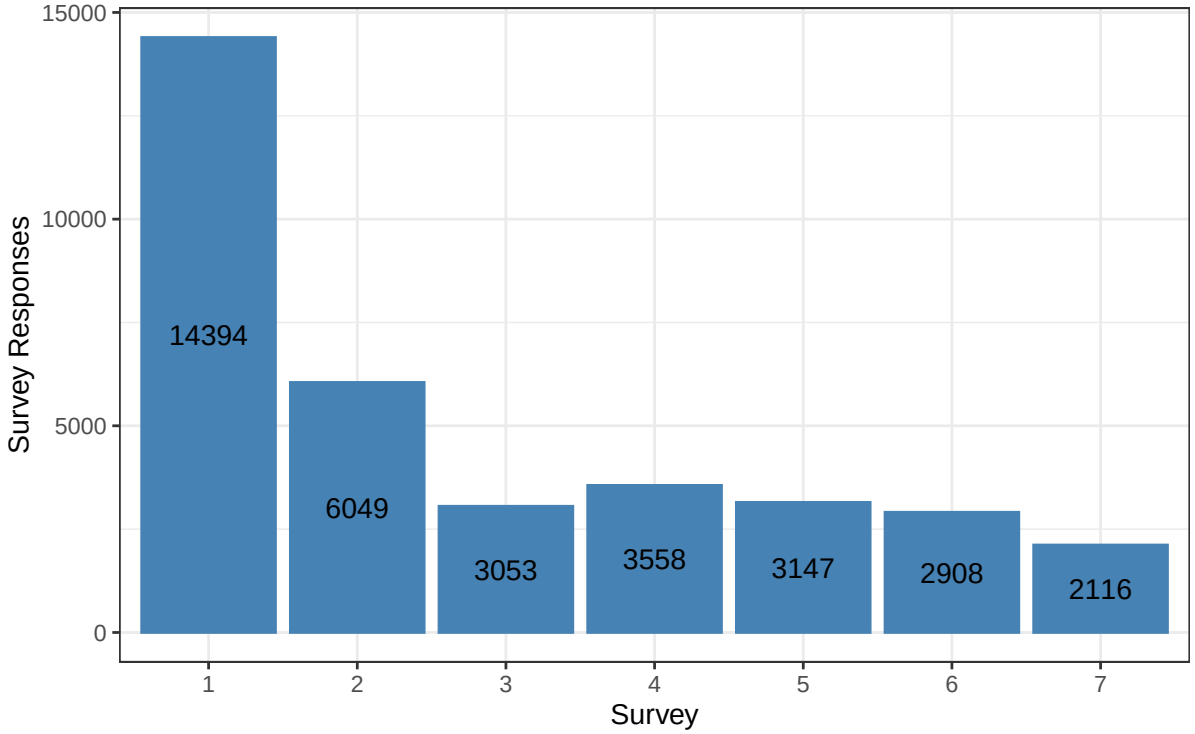
Course timings, duration and student numbers

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

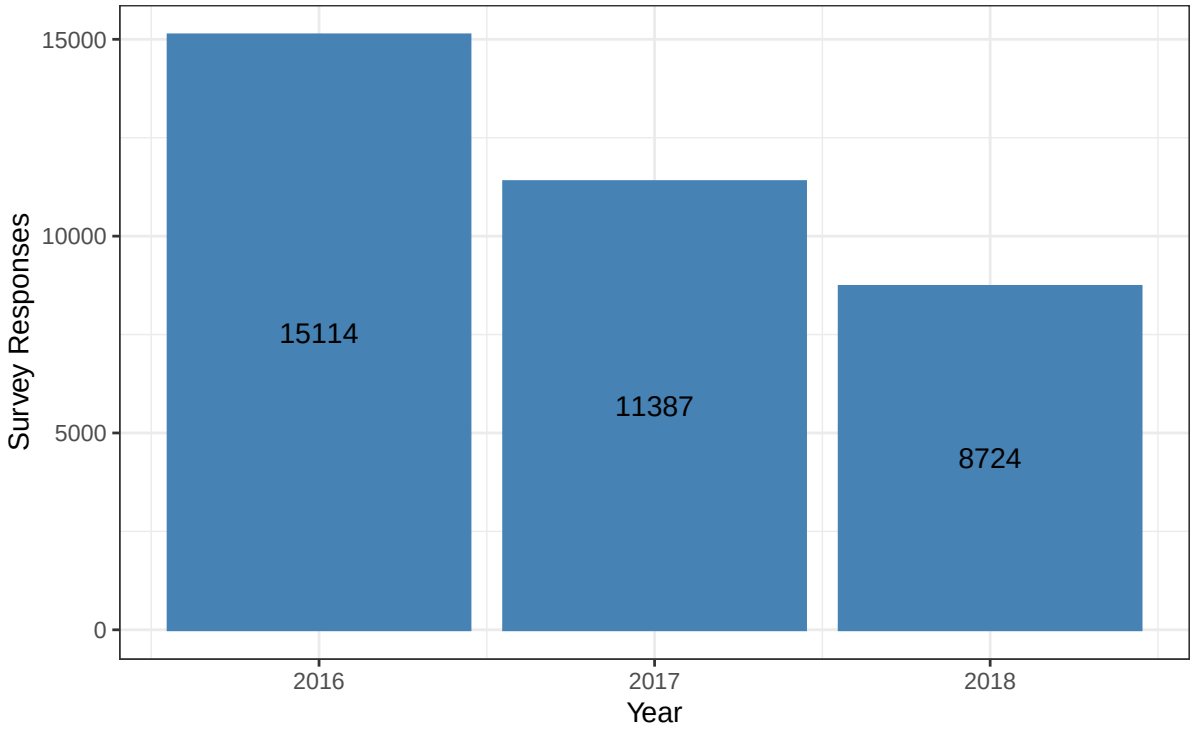
Survey No.	Date First Student Enrolled	Date Last Student Enrolled	Duration (days)	No. of Students
1	29 March 2016	07 September 2017	527 days	14394
2	05 December 2016	13 July 2017	220 days	6049
3	02 July 2017	26 February 2018	239 days	3053
4	27 July 2017	25 January 2018	182 days	3558
5	15 December 2017	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

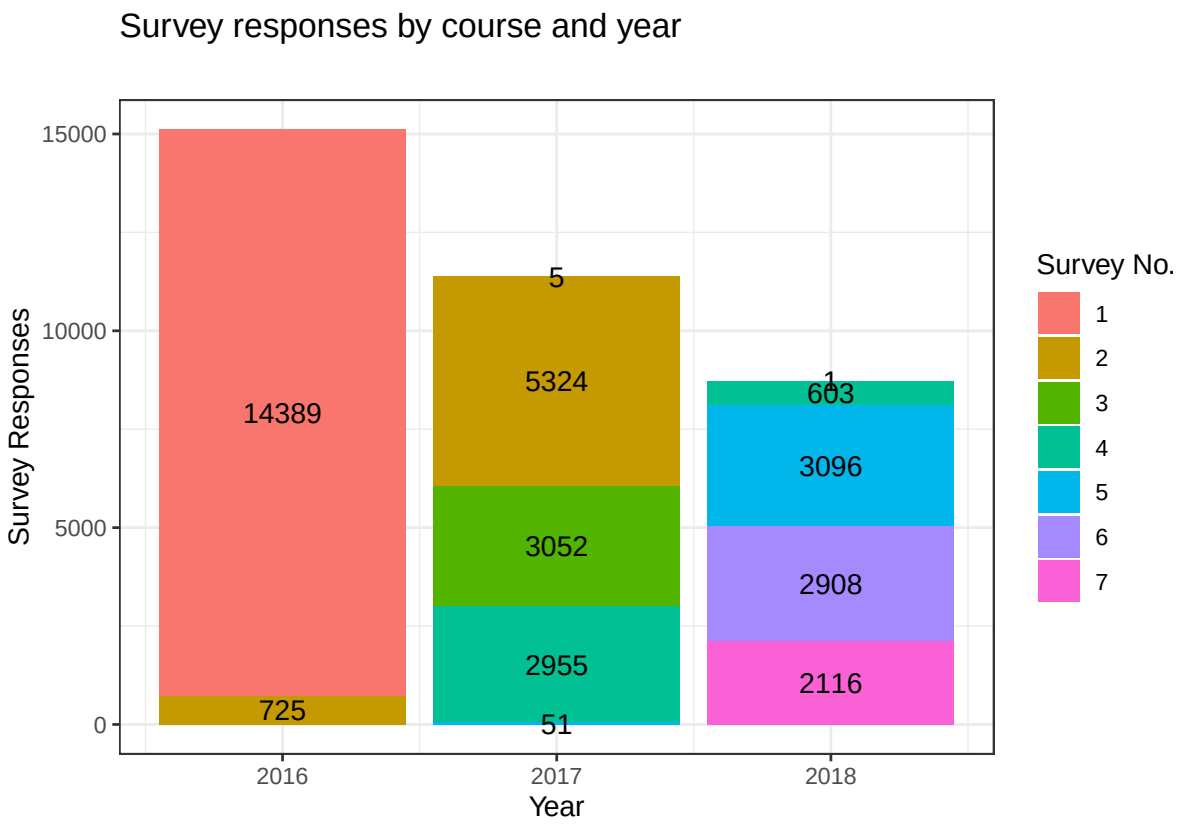
The ‘enrolled at’ dates within the files indicate that the courses overlapped in duration and varied in length, with the volumes of students applying for each course gradually reducing.

Enrolment surveys by course



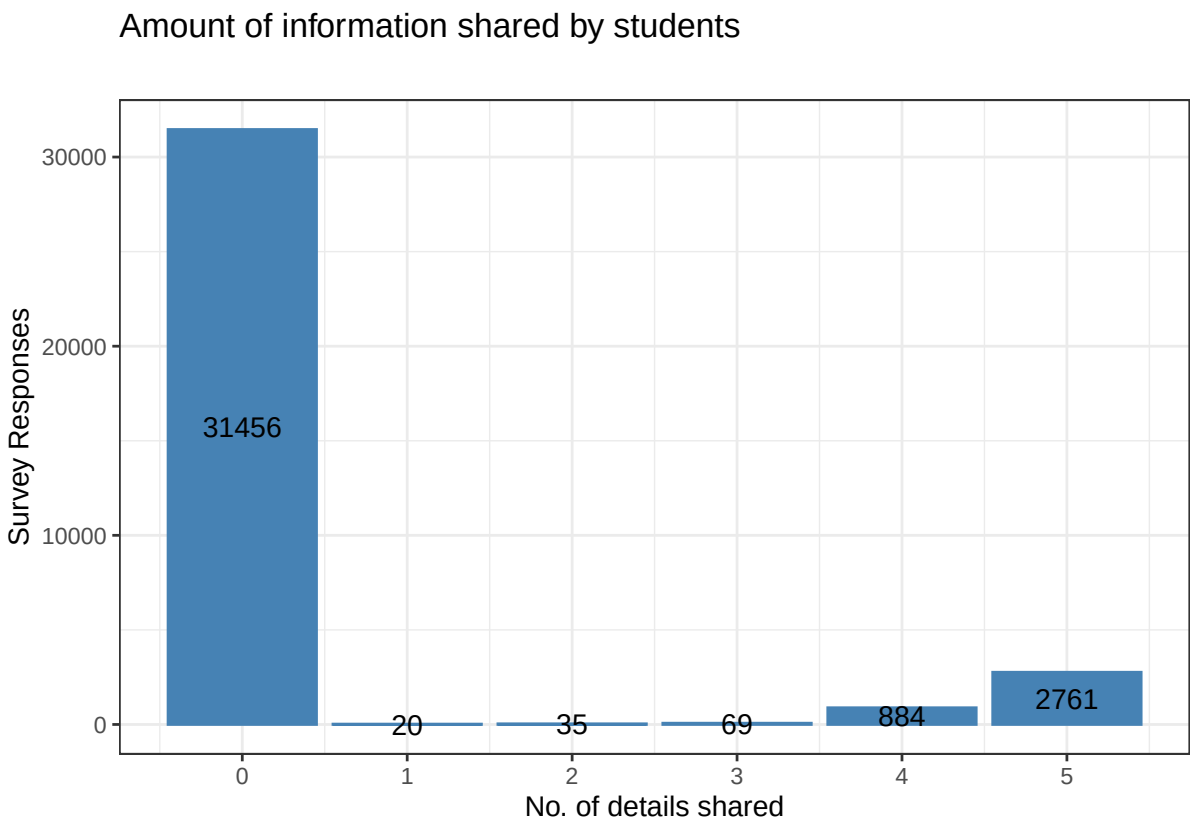
Enrolment surveys by year





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.



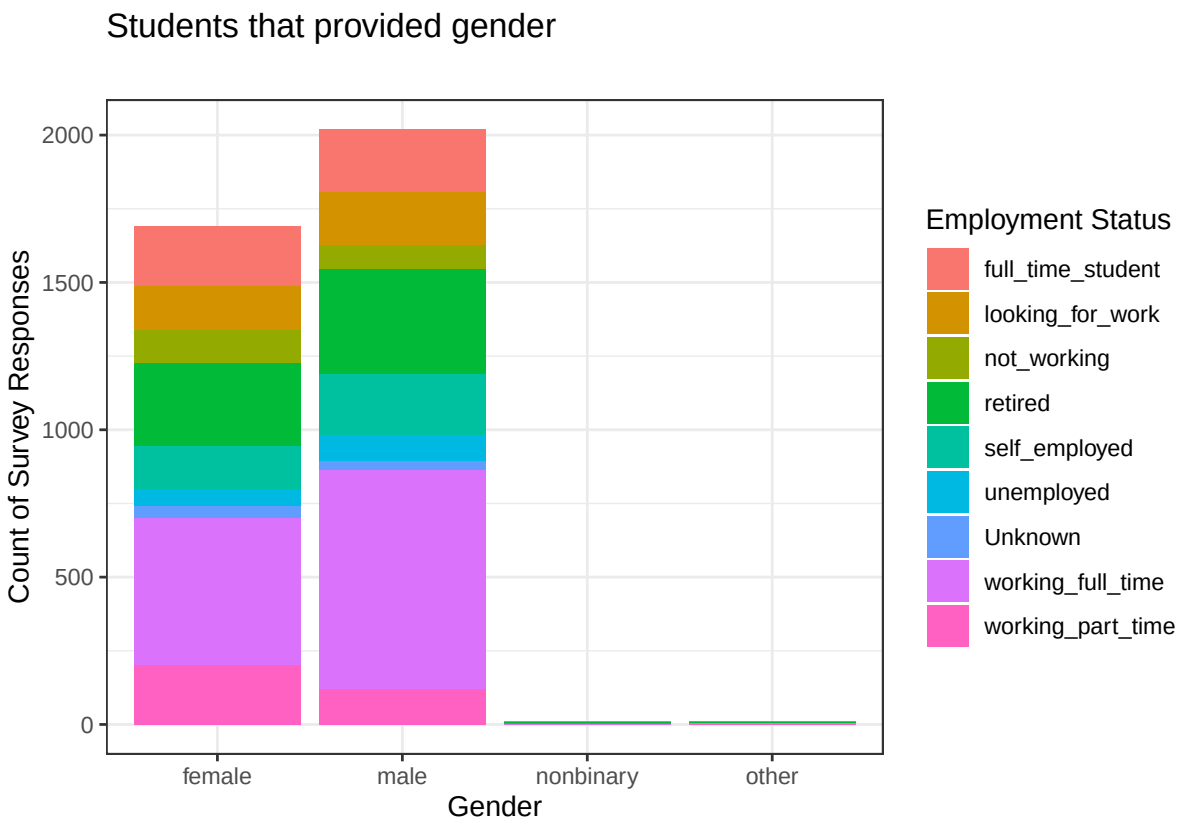
The number of students that shared data was 3769, which equates to 10.70% of total students.

Surprisingly the most commonly shared data item was gender with 3733 students sharing this information. This detail was followed closely by 3715 students sharing highest education level, 3689 sharing employment status,3620 sharing age range and 2881 sharing employment area.

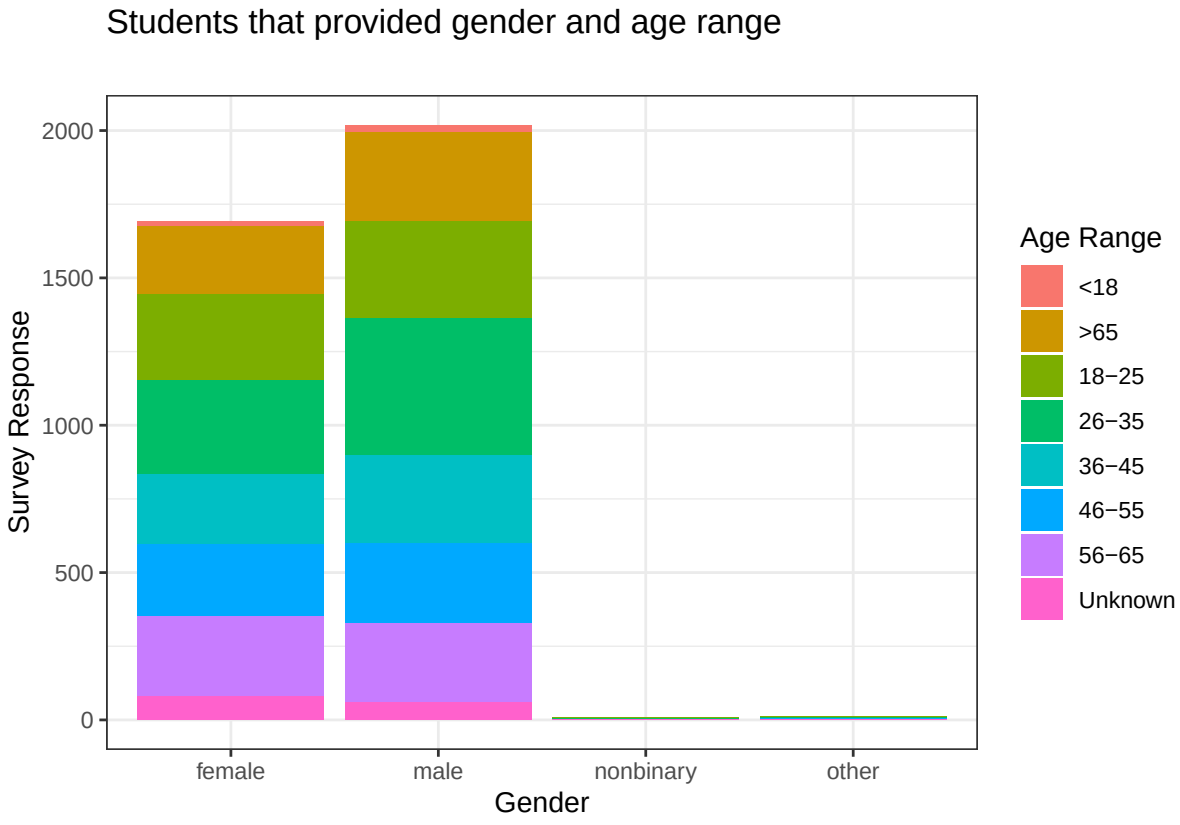
Survey No.	% of Students that Shared Data	Ave. Items Provided by Sharing Students
1	11.98%	4.69
2	10.91%	4.68
3	10.15%	4.68
4	9.13%	4.71
5	11.57%	4.65
6	7.50%	4.67
7	7.94%	4.57

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.

The average amount of data shared was low, at less than one item per student (0.50 items). Students that shared data provided an average of 4.68 items with 73.26% of students that shared information providing all five requested details.



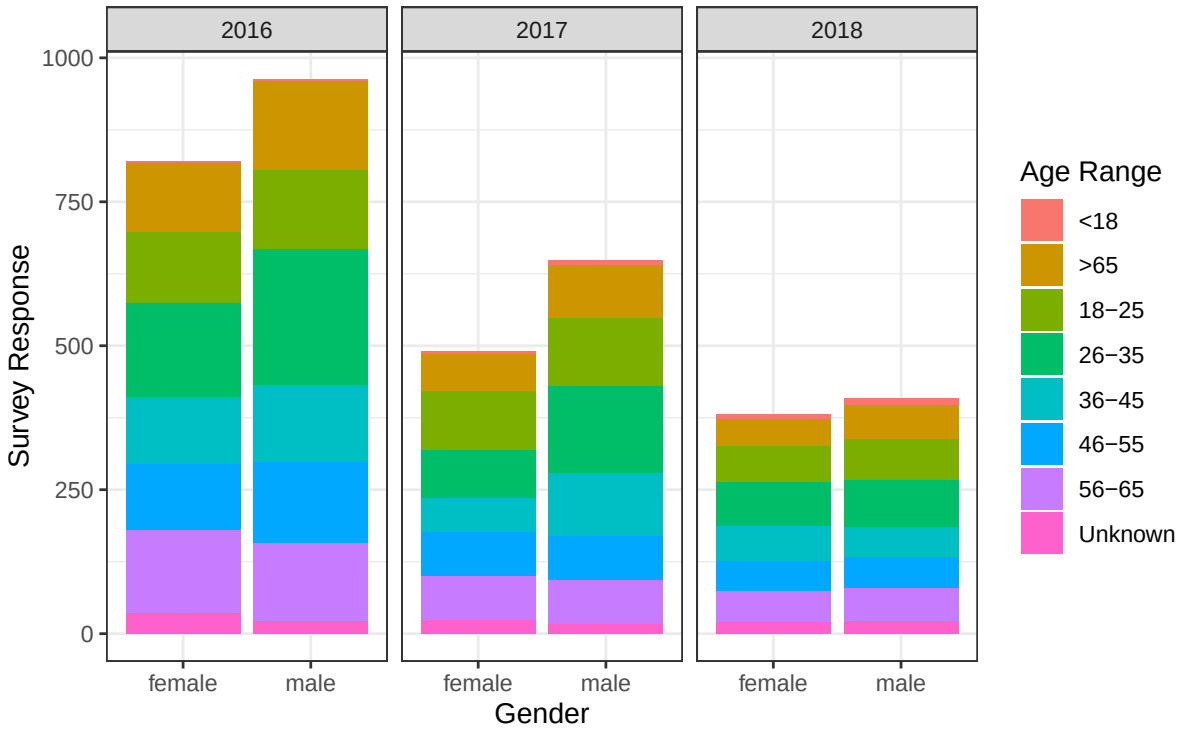
Male and females working for home were the largest groups willing to share gender.



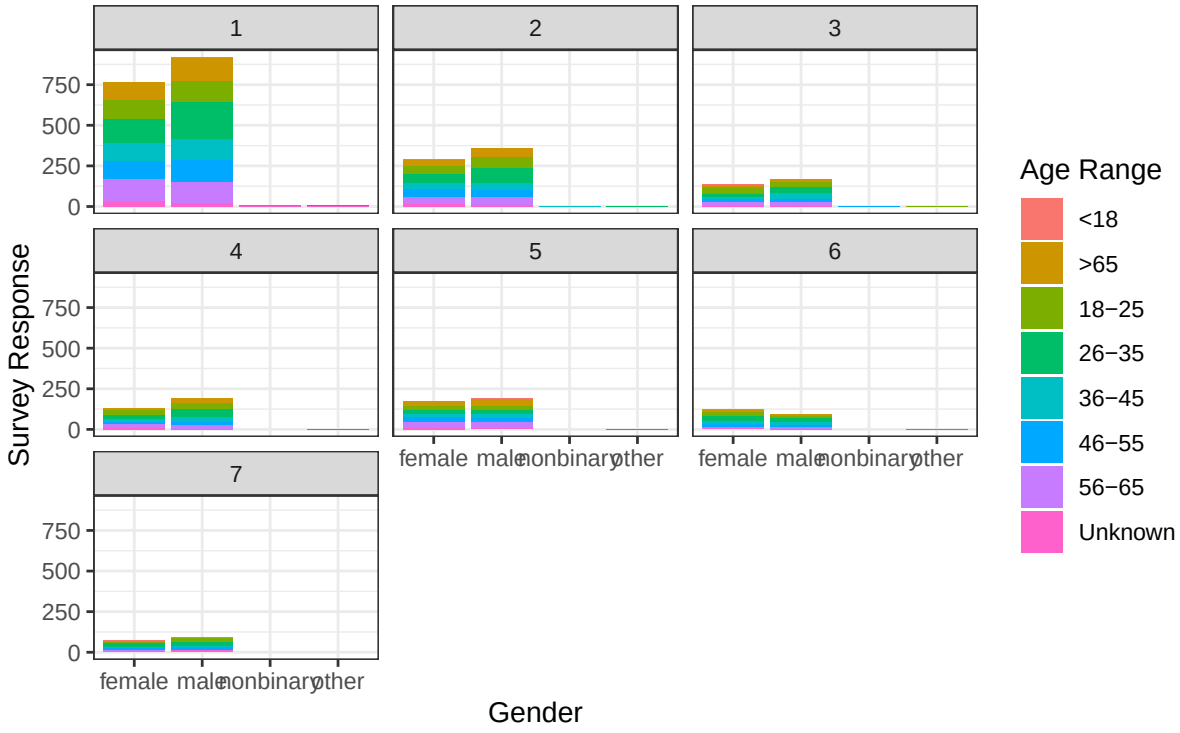
Students willing to share information about gender were split very evenly across age range group. There was a subset of individuals willing to share information on gender but not age

range.

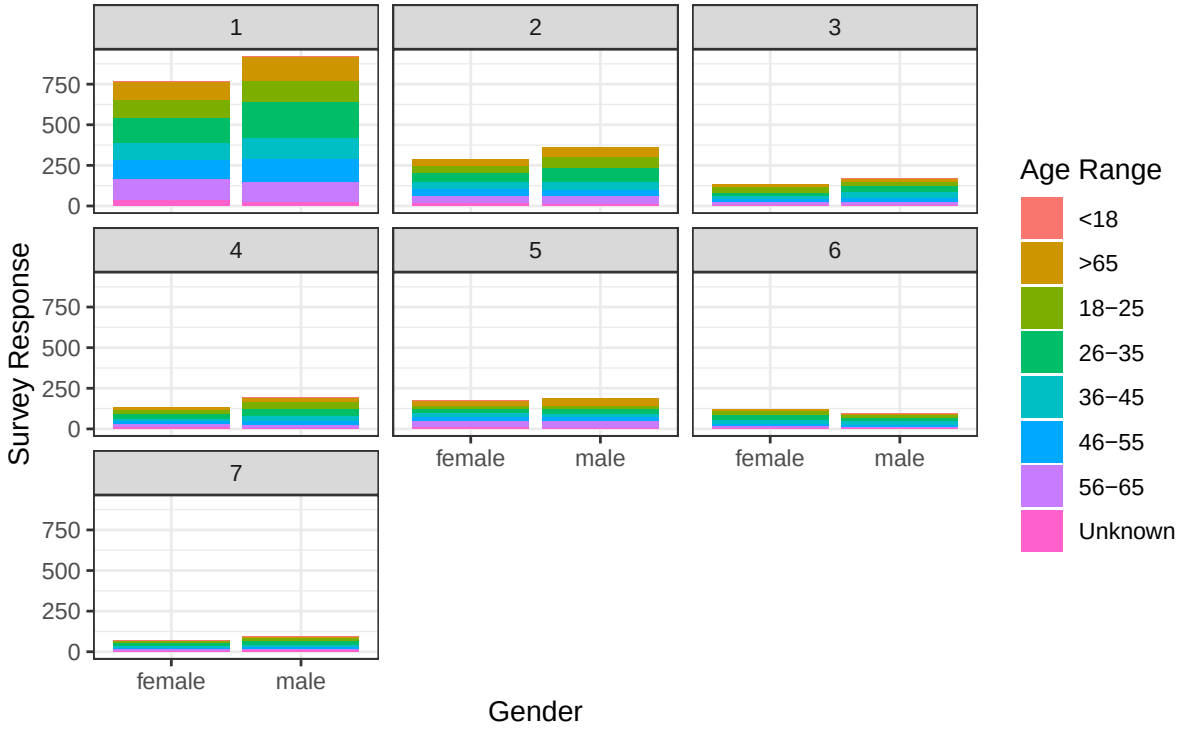
Male or female students that provided age range



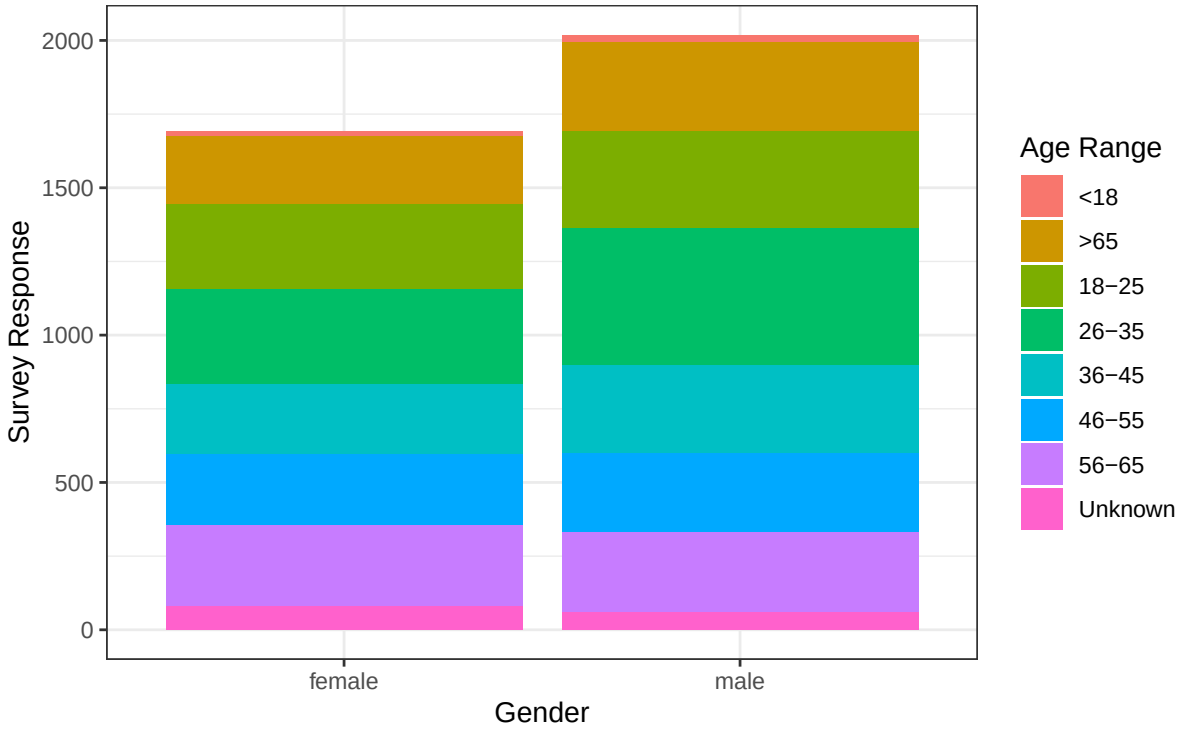
Students that provided gender and age



Male or female students that provided age range

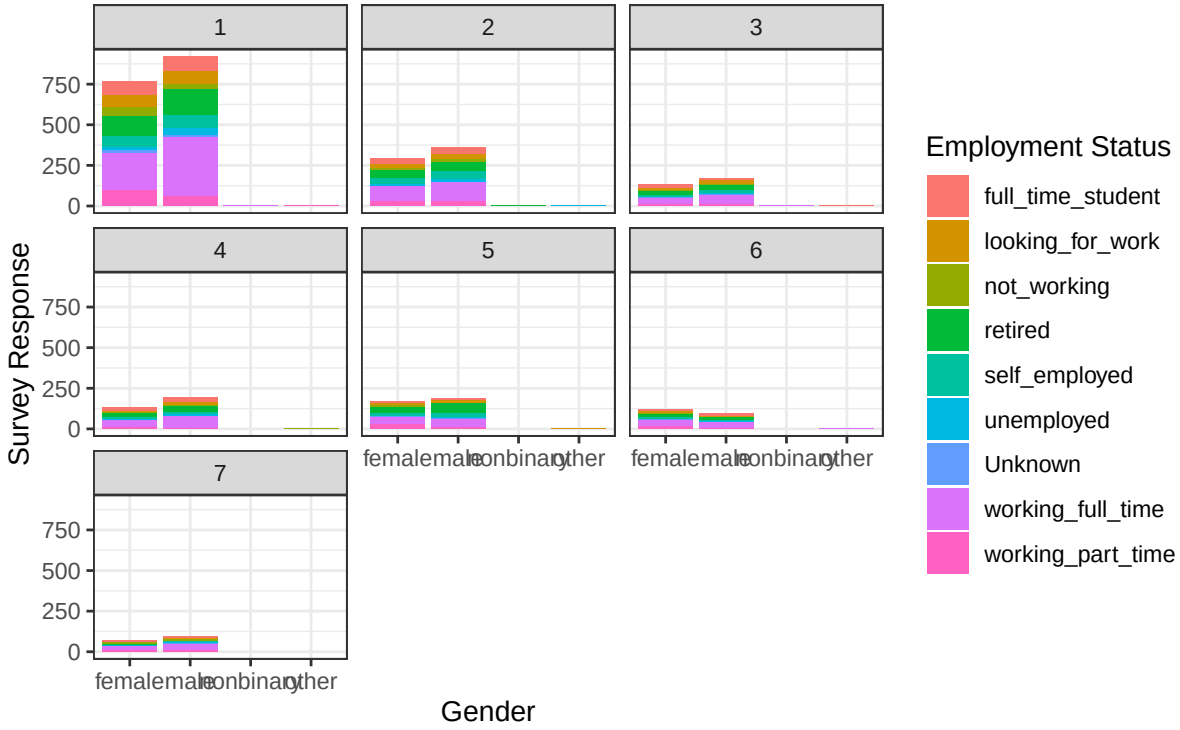


Male or female students that provided age range



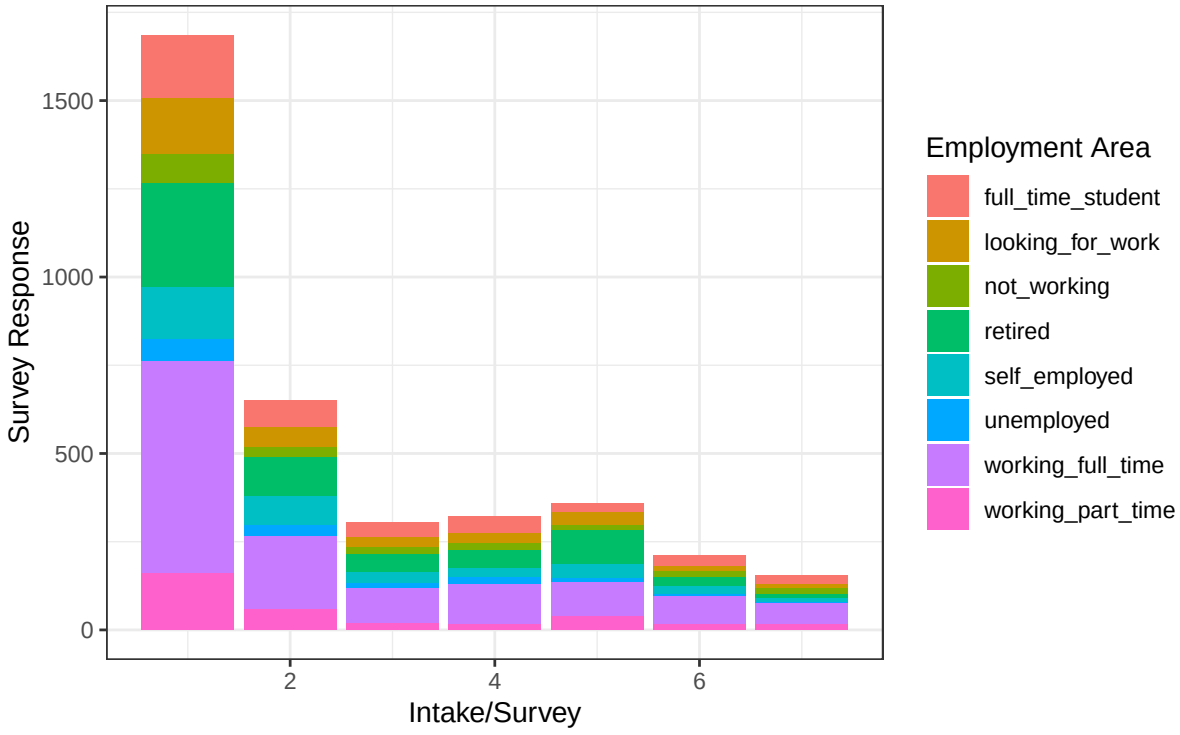
Attitudes on sharing age and gender declined by year and survey but further analysis was required to determine if this was disportionate, course duration changed and student numbers varied.

Students that provided gender and employment status

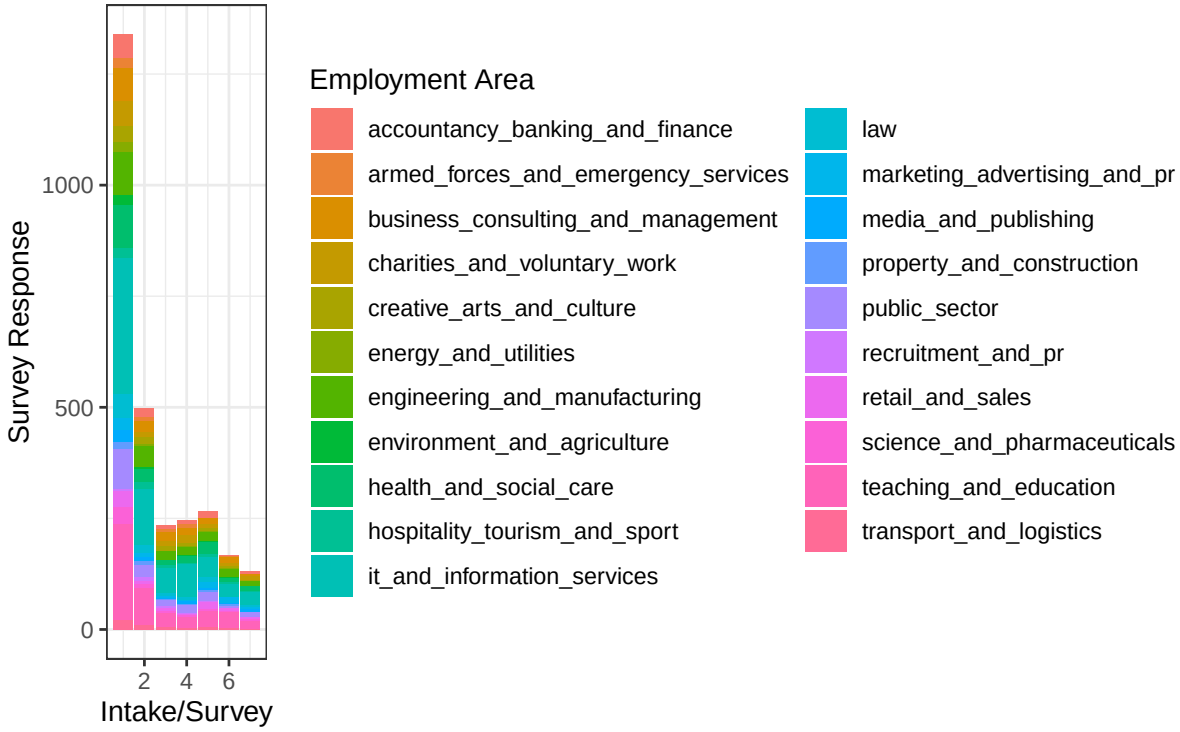


Words about employment graphs to follow

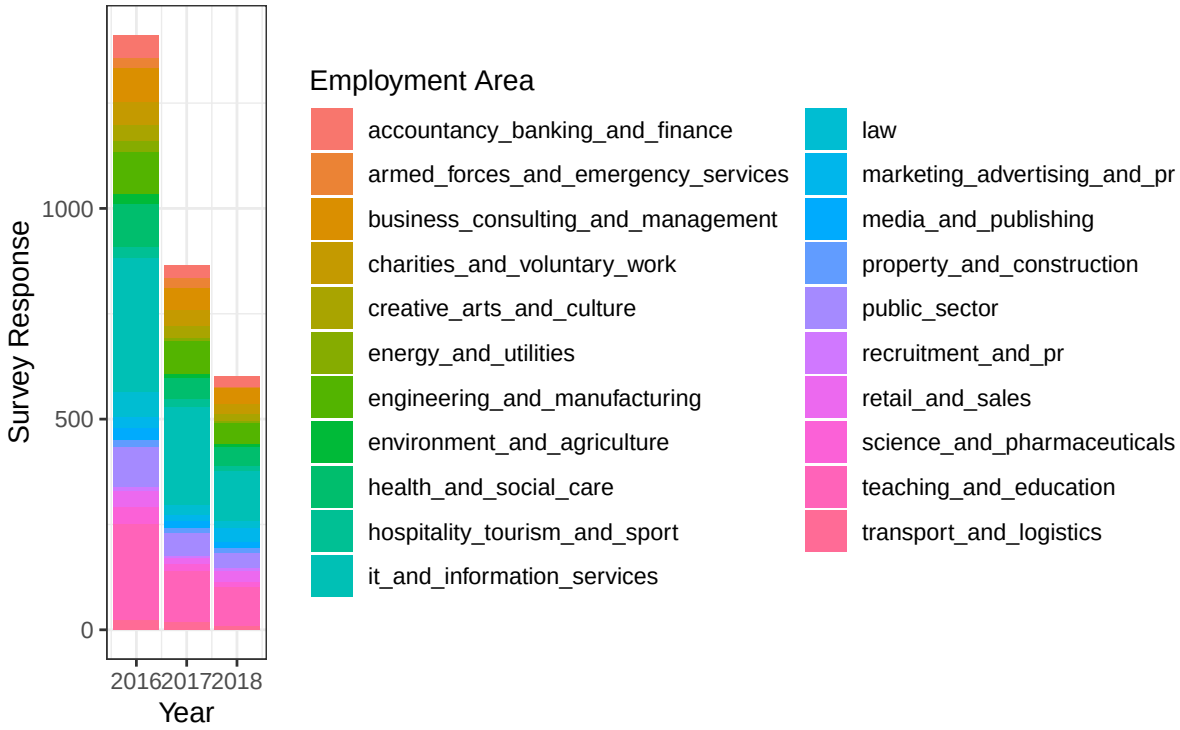
Students that Provided Information on employment status



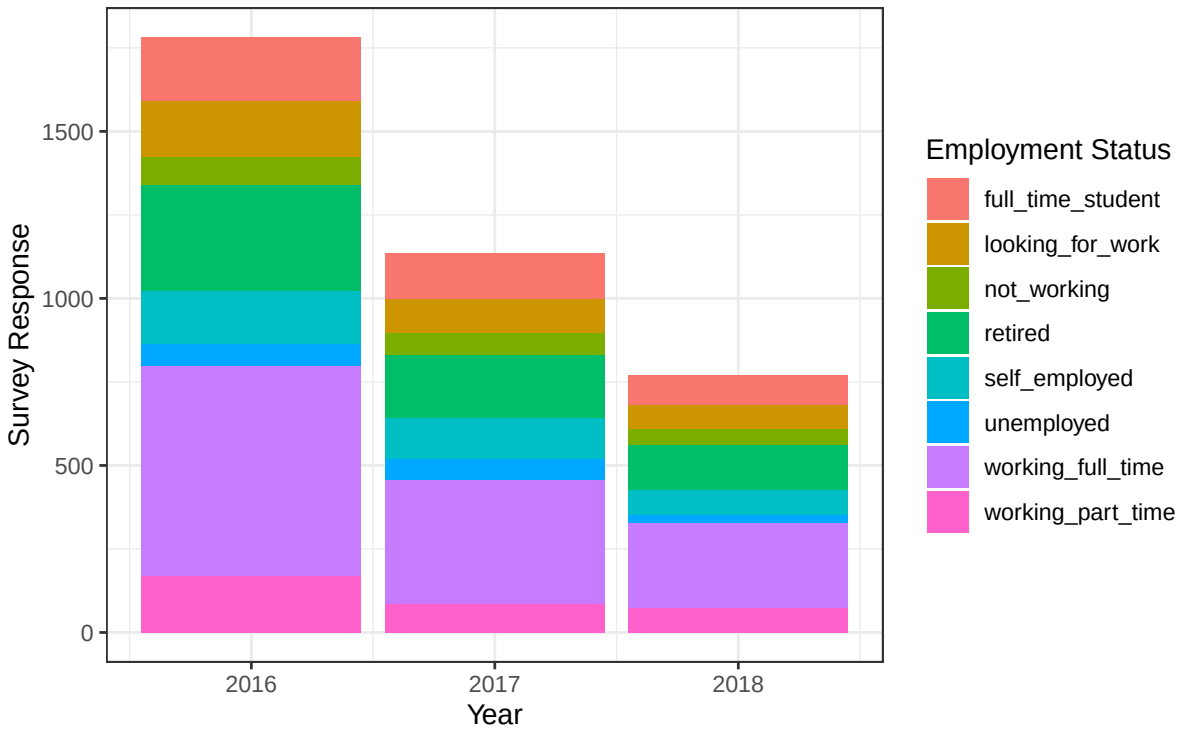
Students that provided employment area



Students that provided information on employment area



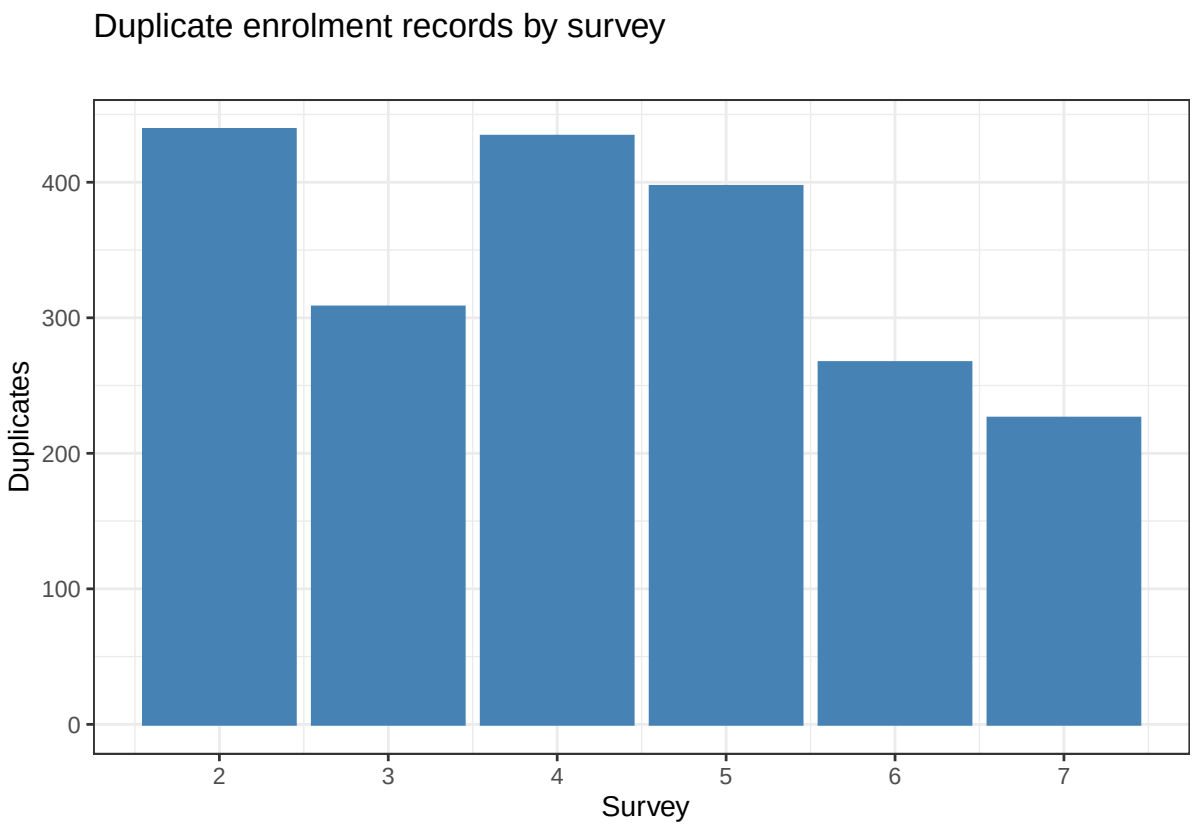
Students that provided information on employment status



Data and assumptions

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.

Data summary

Issy Middleton - C1000051
Newcastle University - CSC8631 Summative Assignment
December 2020

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”