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

By Ambreen Masud

30 November, 2021

Introduction

This project was commissioned by Newcastle University to gain insights into an online Future Learn course in cyber security and assess the productivity of the course. 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 in the data analysis. This methodology is as follows:

Gaining an understanding of Newcastle University's business needs and determining the goals of the project based on these needs Understanding the properties of the data provided, including assessing data quality *Preparing the data through cleaning, constructing and formatting the data, readying it for data analysis* Producing a report as an output of the analysis covered in this project

This report focused on the business

This report found that student attainment had dropped from _____% in run 1 to _____% in run 7.

The findings of this report led to a further analysis

This report will conclude with... ————

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"*.

Before analysing this project, a number of assumptions were made on the dataset:

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.

This 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 current courses, but 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 these questions effectively, the data mining process included:

1. Calculating and plotting the numbers of student enrolled per run.
2. Calculating and plotting the numbers of students per country and region over the total run of courses.
3. Calculating and plotting demographic data over each run.

In assessing this _____

Data Understanding

The dataset provided by Newcastle University comprised of 7 PDF files and 52 CSV files. The PDF files provide information on the course overview of each run, highlighting each step that students need to complete to course. The CSV files included within the dataset are:

Archetype Survey Responses *- Responses from a survey conducted on student archetype. Runs 1 to 2 contain no data, whereas runs 3 to 7 contain data.* **Enrolments** - Student enrolment records, including date enrolled and information on student profile. Data is available for all runs. ***Leaving Survey Responses*** *- Responses from a survey conducted as students want to leave the course, including information on when students left, the reason why and what step they ended the course on. Runs 1 to 3 contain no data, whereas runs 4 to 7 have data.* **Question Response** - ***Step Activity*** **Team members** ***Video Stats*** **Weekly Sentiment Survey Responses**

The records vary depending on the run. For the purposes of this analysis and in line with the business objectives and key questions, the enrolments data was assessed in further detail so that will be discussed further here. 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 13 identically named columns as can be seen below.

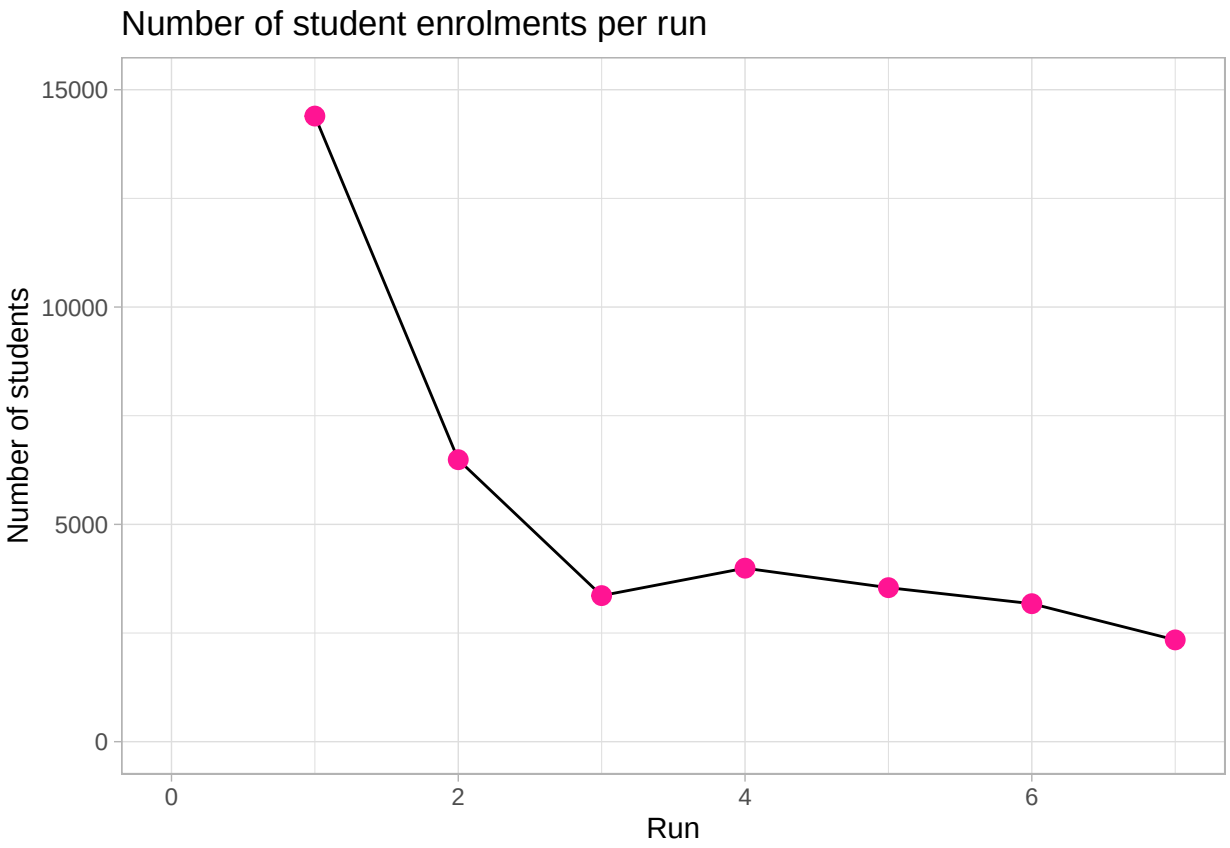
```
##  [1] "learner_id"              "enrolled_at"
##  [3] "unenrolled_at"           "role"
##  [5] "fully_participated_at"   "purchased_statement_at"
##  [7] "gender"                  "country"
##  [9] "age_range"               "highest_education_level"
## [11] "employment_status"       "employment_area"
## [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 their participation on the course whilst columns 7 to 12 provide demographic information about each student.

In assessing the data quality, some of the data has missing values with empty fields and some of the data is missing with fields labelled *"Unknown"*. This suggests that the data is in some parts incomplete. This will be further assessed in the data preparation portion of this report.

```
## # A tibble: 6 x 14
##   learner_id  enrolled_at  unenrolled_at role  fully_participa~ purchased_state~
##   <chr>       <chr>        <chr>         <chr> <chr>            <chr>
## 1 160d6600-e~ 2016-08-10 ~ ""            lear~ ""               ""
## 2 4dc22fed-6~ 2016-05-24 ~ "2018-10-30 ~ lear~ ""               ""
## 3 ecdd37db-0~ 2016-05-19 ~ ""            lear~ "2016-09-22 16:~ ""
## 4 988964c9-7~ 2016-05-19 ~ ""            lear~ ""               ""
## 5 f1493366-1~ 2016-09-19 ~ ""            lear~ ""               ""
## 6 25cc3b46-a~ 2016-08-30 ~ ""            lear~ "2016-10-25 12:~ ""
## # ... with 8 more variables: gender <chr>, country <chr>, age_range <chr>,
## #   highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>, run <dbl>
```

Data Preparation

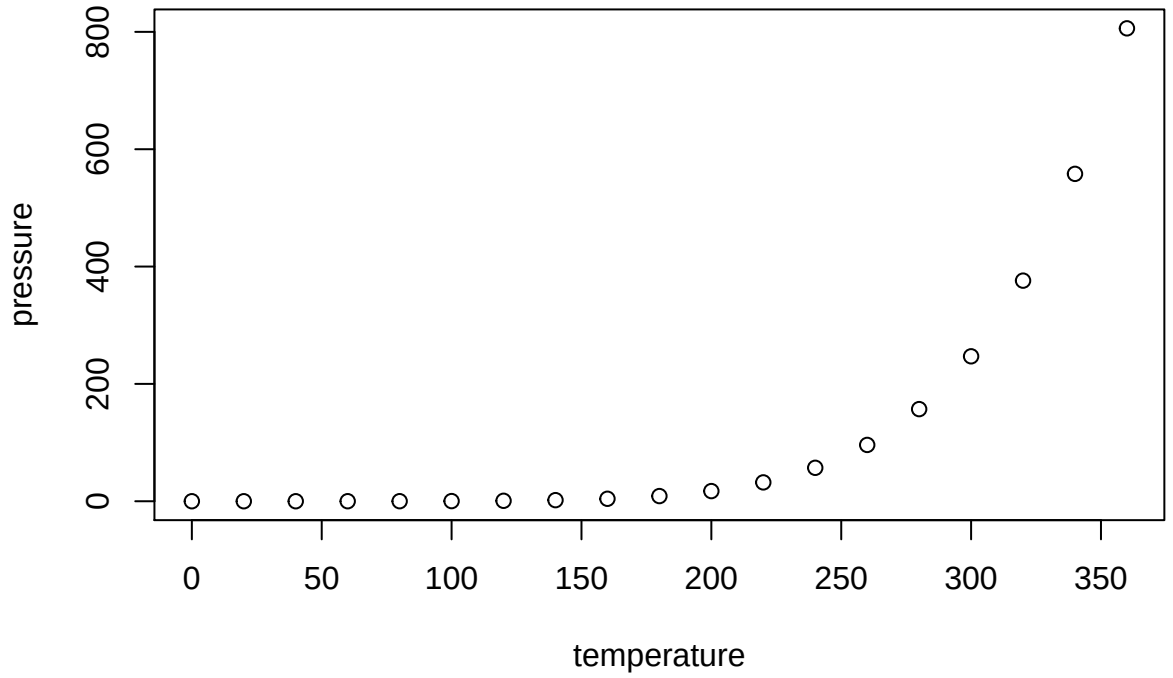


The results display that the number of students enrolling on the course has reduced over the number of sessions. This leads onto further questions as to why this may be.

Recognise the challenges posed by the data

Data Preparation

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

References

Higher Education Commission (2016) From bricks to clicks: The potential of data analytics in higher education.
https://www.policyconnect.org.uk/research/report-bricks-clicks-potential-data-and-analytics-higher-education