CSC8631 Report

Thomas Richardson

05/11/2021

Business Understanding

FutureLearn is a Massive Online Open Course (MOOC) platform, and since launching in 2013 they have partnered with a range of universities and businesses to provide a variety of online courses and degrees (futurelearn.com). They are situated in the Higher Education space and are hence competing with a number of new entrants in the marketplace many of which have an emphasis on attracting consumers in the international market (bricks to clicks Higher Education, Foreword). In order to gain an advantage in this area, FutureLearn are looking to incorporate Learning Analytics to both improve their courses and develop a strategy to attract more consumers. The course being investigated for this project is a Cyber Security course led by Newcastle University with specific success criteria of finding an accessible method to encourage consumers to sign up to a FutureLearn course and determining methods to adjust course structure so that these people feel more satisfied throughout.

The first data mining goal for this investigation is to identify at which times throughout the year the course was most popular in an attempt to gain an understanding of why this may be the case and also allow for future planning so that increased support can be made available to the participants to help keep up with the increasing demand during these periods. In addition to this a by-product of this investigation may enable FutureLearn to identify periods to focus their advertising budget in an attempt to optimise their resources and potentially gain an advantage over their competition. If successful there will be clear recurring time period(s) of increased engagement that can be seen from the data.

Some constraints to the investigation is that there is no data in regard to previous advertising of the course and how this may have affected enrolments so for this reason it is assumed throughout that advertising has remained at a constant level spread evenly across all regions and mediums for the entirety of the course's lifespan. In addition it is to be assumed that the course was available to enrol upon at moment throughout the course history.

This project will follow the CRISP-DM process model (Chapman et al., 2000, CRISP-DM 1.0 Step-by-step data mining guide) and all data processing will be completed in R, incorporating the package `ProjectTemplate`. Other R libraries intended to be used include `ggplot2` which will be used to create plots due to its flexibility in creating plots allowing for ease of use when layering and using position adjustment in plots, and also `dplyr` which will be used to transform the data effectively.

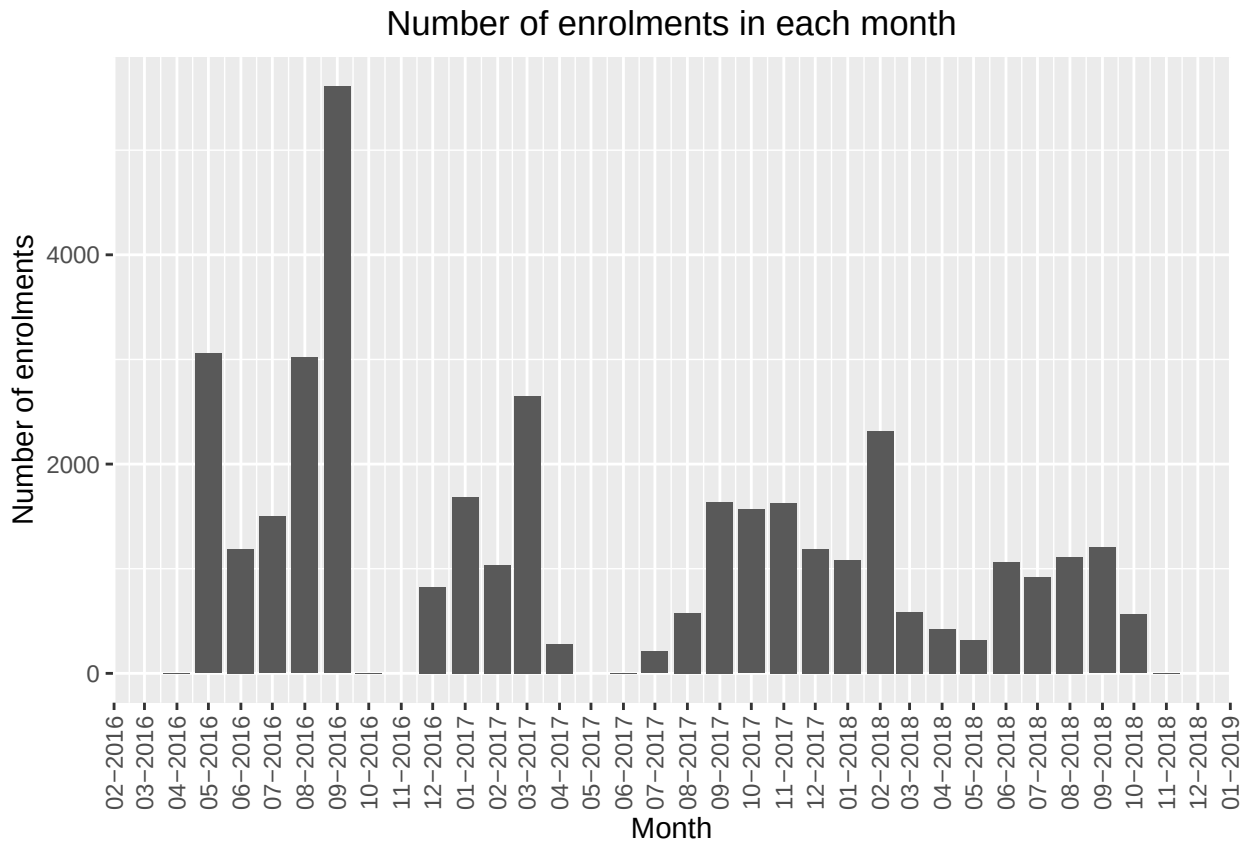
Data Understanding

As previously mentioned the data collected is from the Newcastle University Cyber Security MOOC. It was collected over 7 consecutive runs of the course with each run spanning a different length of time ranging from roughly two months in duration to over sixth months. There were often slight changes made to the course between runs and there were also sometimes some data collection changes between runs, for example data collection on team members did not start until the second run and hence this csv file for the first run does not exist.

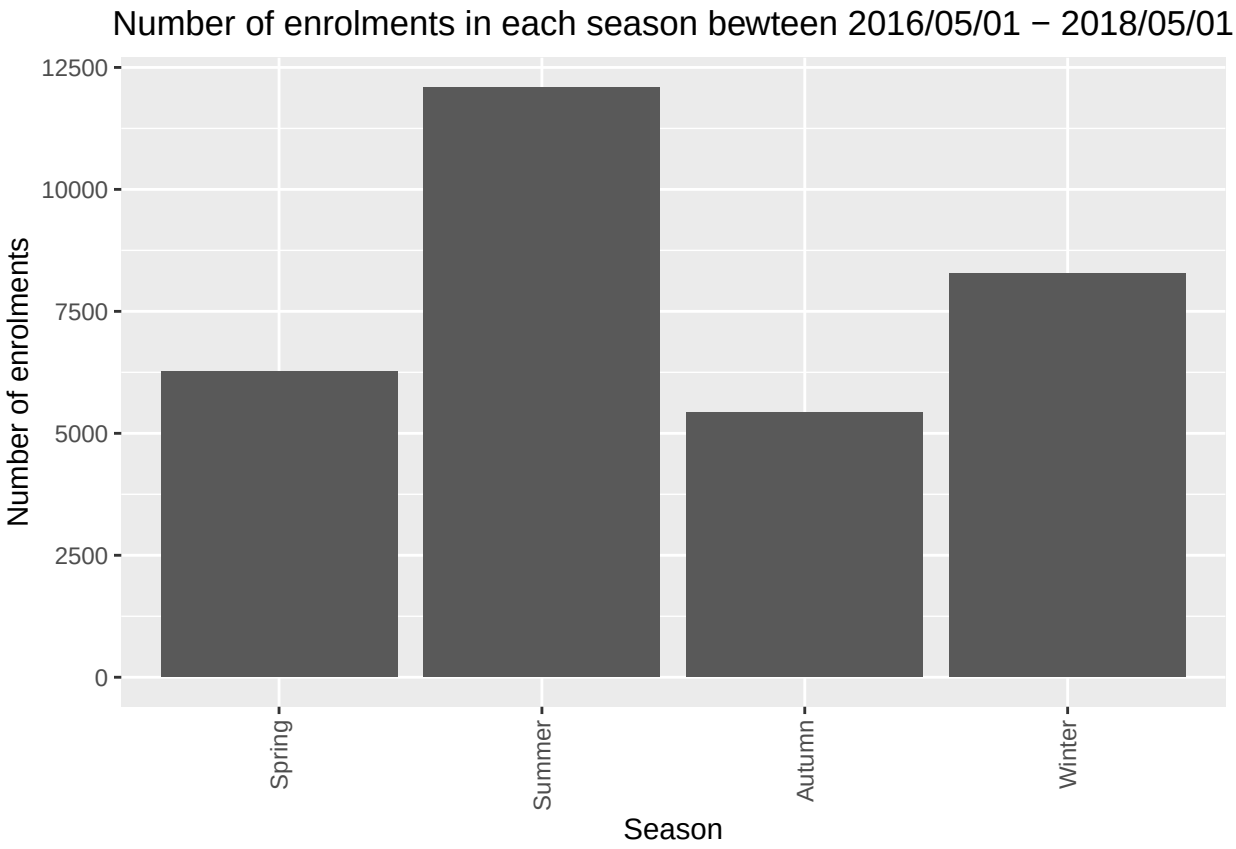
The data has been grouped into (up to) eight different csv files per run,

- **Archetype Survey Responses:** This contains the learner ID, the time and date they completed the survey, and the archetype they were given based on their results.
- **Enrolments:** This contains the learner ID, the date and time they enrolled, unenrolled, completed the course and purchased a certificate. It also contains information such as the individuals role on the course (e.g. learner), their age range, gender, country, highest education level, employment stauts and employment area, and finally the country they were detected from.
- **Leaving Survey Responses:** "DEAD TING"
- **Question Responses:**
- **Step Activity:**
- **Team Members:**
- **Video Stats:**
- **Weekly Sentiment Survey Responses:**

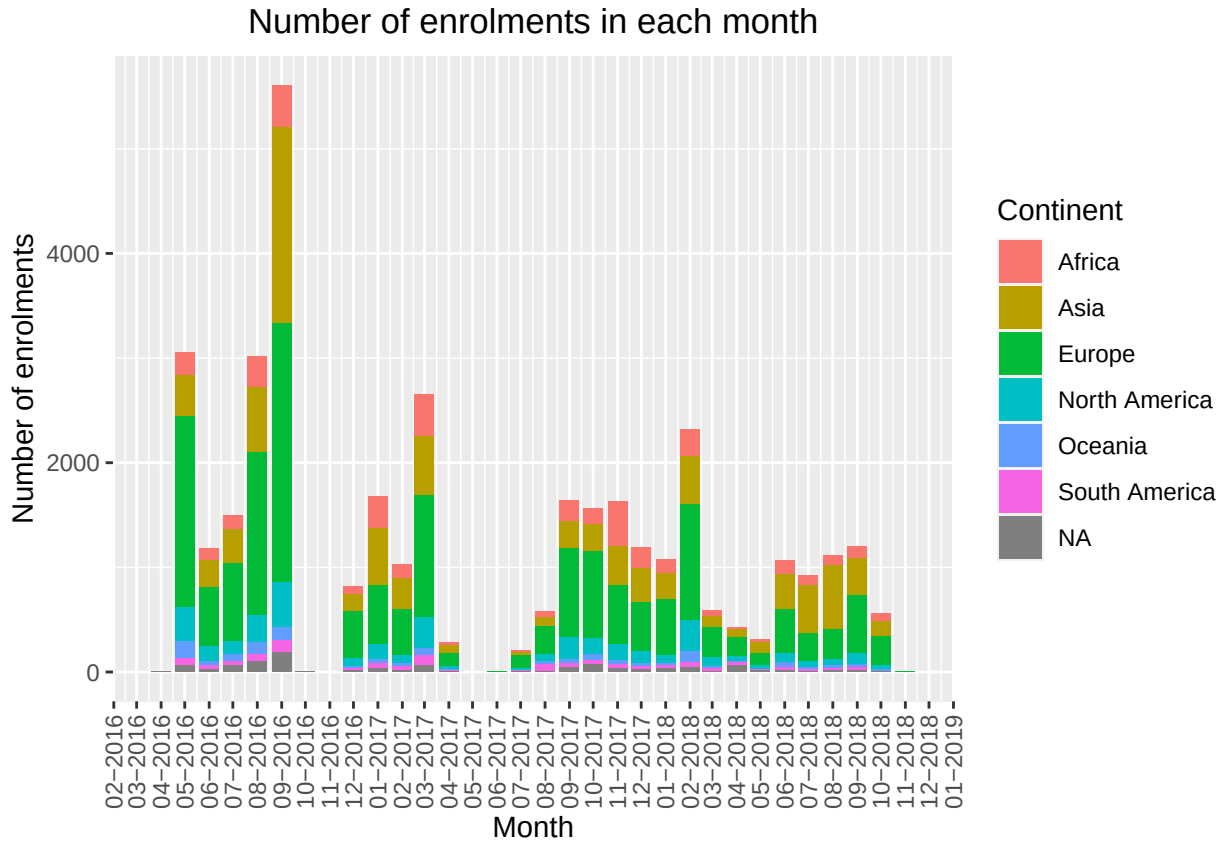
```
months_plot
```



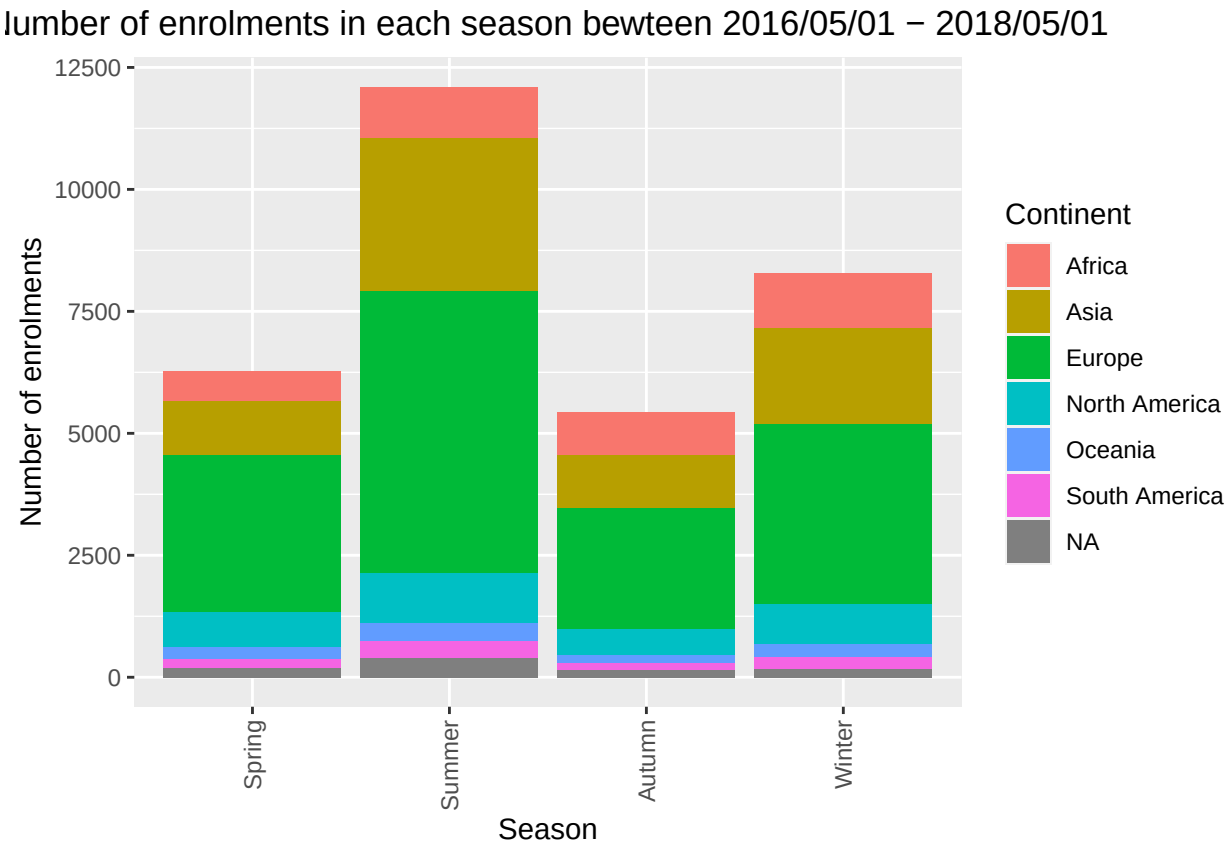

```
seasons_plot
```




```
months_continents_plot
```



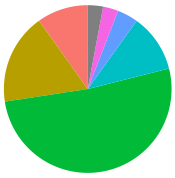

```
seasons_continents_plot
```




```
grid.arrange(spring_pie, summer_pie, autumn_pie, winter_pie, ncol=2,
             top=textGrob("Continental Proportion per Season",
                          gp=gpar(fontsize=20,font=1)))
```

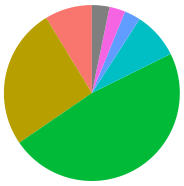
Continental Proportion per Season

Spring



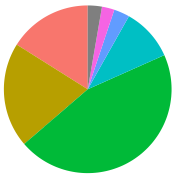
Continent
- Africa (9.86%)
- Asia (17.55%)
- Europe (51.53%)
- North America (11.15%)
- Oceania (4.02%)
- South America (2.94%)
- NA (2.95%)

Summer



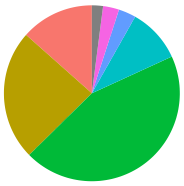
Continent
- Africa (8.70%)
- Asia (25.87%)
- Europe (47.75%)
- North America (8.56%)
- Oceania (2.96%)
- South America (2.93%)
- NA (3.23%)

Autumn



Continent
- Africa (16.06%)
- Asia (20.26%)
- Europe (45.36%)
- North America (10.15%)
- Oceania (3.04%)
- South America (2.41%)
- NA (2.73%)

Winter



Continent
- Africa (13.45%)
- Asia (23.92%)
- Europe (44.48%)
- North America (9.96%)
- Oceania (3.20%)
- South America (2.95%)
- NA (2.04%)