CSC8631 Report

Thomas Richardson

05/11/2021

Cycle 1

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 was 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.

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

This project follows the CRISP-DM process model (Chapman et al., 2000, CRISP-DM 1.0 Step-by-step data mining guide) and all data processing was completed in R, incorporating the package `ProjectTemplate`. Other R libraries used include `ggplot2` which was to 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 was to be used to transform the data effectively.

Data Understanding

As previously mentioned the data collected was 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 had been grouped into (up to) eight different csv files per run,

- **Archetype Survey Responses:** This contained learner IDs, the time and date they completed the survey, and the archetype they were given based on their results.
- **Enrolments:** This contained learner IDs, 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:** This contained learner IDs, the date and time they left the course as well as details on their last completed step, and a reason for leaving.
- **Question Responses:** This contained learner IDs, information to identify the question (e.g. question number), the type of question (e.g. multiple choice), the learners' answers and whether they answered correctly.
- **Step Activity:** This contained learner IDs, and when they first visited and last completed each step.
- **Team Members:** This contained information on the different team members such as their team role and user role.
- **Video Stats:** This contained statistics on each of the videos on the course including video length, views, percentage of viewers who reached different percentage lengths of the videos, the proportion of device types used to watch and the proportion of viewers per continent.
- **Weekly Sentiment Survey Responses:** This contained feedback responses, including the week they were submitting, an experience rating and a reason for that rating.

Due to the differing lengths of each of the runs there were some runs with significantly less data than others and due to learners dropping out at various points throughout the course there was henceforth less data on these later stages. Not all of the data collected was quantitative, as data such as feedback also takes qualitative responses and while these can offer greater insights into learner attitudes and opinions they also allow for brief non-descriptive opinions such as a response found in the **Weekly Sentiment Survey Responses** of the seventh run which was simply "DEAD TING". In spite of this the overall scope of the data was wide and allowed for a variety of investigative approaches in a variety of areas.

As per the goals of this investigation it appeared that the most useful data would be found in the **Enrolment** files across all 7 of the runs as from this it would be possible to investigate the number of learners enrolled since the course began and at which points in time the number of enrolments was at a greater level that others.

Data Preparation

Following on from the data understanding the first act to performed on the data was to merge the 7 **Enrolment** files from the different runs. This was done using the `merge` function in R and can be seen in the 01-A.R file in the munge folder of the FutureLearn repository. It was felt important to use data from all 7 of the runs so as to investigate across as much data as possible in order to gain more reliable results. Following this the data was cut down to only contain the rows where the learner's role was labelled as "learner" so as to not potentially skew results by including data of other roles. Next, making use of the `dplyr` package, a new tibble was made, transforming the data so that this new tibble contained a count of the number of enrolments in each month of the course. Months that contained no enrolments were not included in this tibble but were to be accounted for on future graphs. Finally, another column was added to the tibble that, with help from the `lubridate` package, contained the date of the first day of each of the months so as to identify a position on an axis for this data to be plotted.

A similar investigation could have been performed but instead using data only from the learners who completed to course in order to gain an understanding of who may most suit this course at particular times

however after a short inspection it was clear the number of completion was not at a level high enough to infer any reliable conclusion and for this reason enrolments remained the focus.

The next task was to plot the data so that any potential trends or patterns could be seen. This was done using the `ggplot2` package to plot the bar chart seen in Figure 1. The bars in Figure 1 appeared to follow an undulating pattern with the highest number of enrolments to be in September 2016. There was also common lows in the months of April and May in 2017 and 2018. Finally there was also a low number of enrolments in October and November 2016.

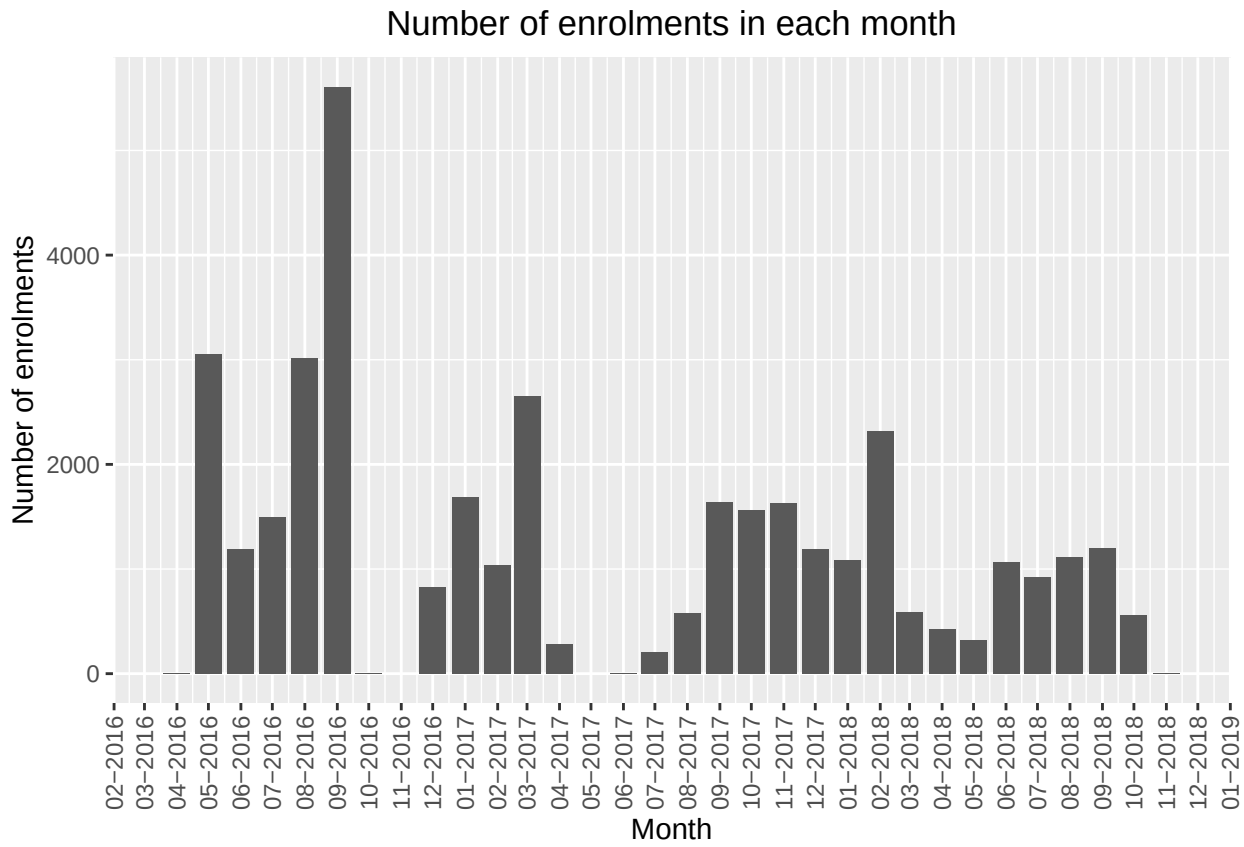

Figure 1: Bar chart showing the number of enrolments in each month from Ferbruary 2016 to January 2019

Cycle 2

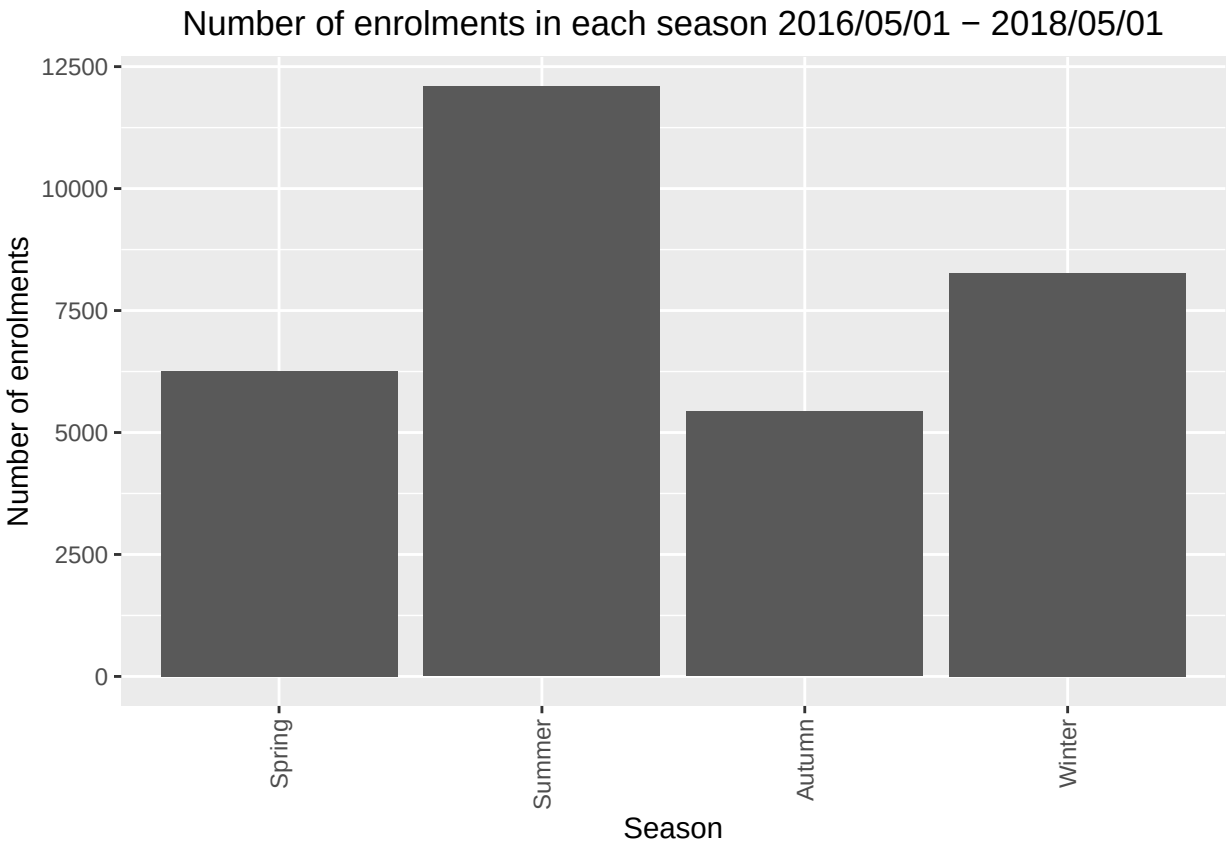

Figure 2: Bar chart showing the number of enrolments per season from May 2016 to May 2018

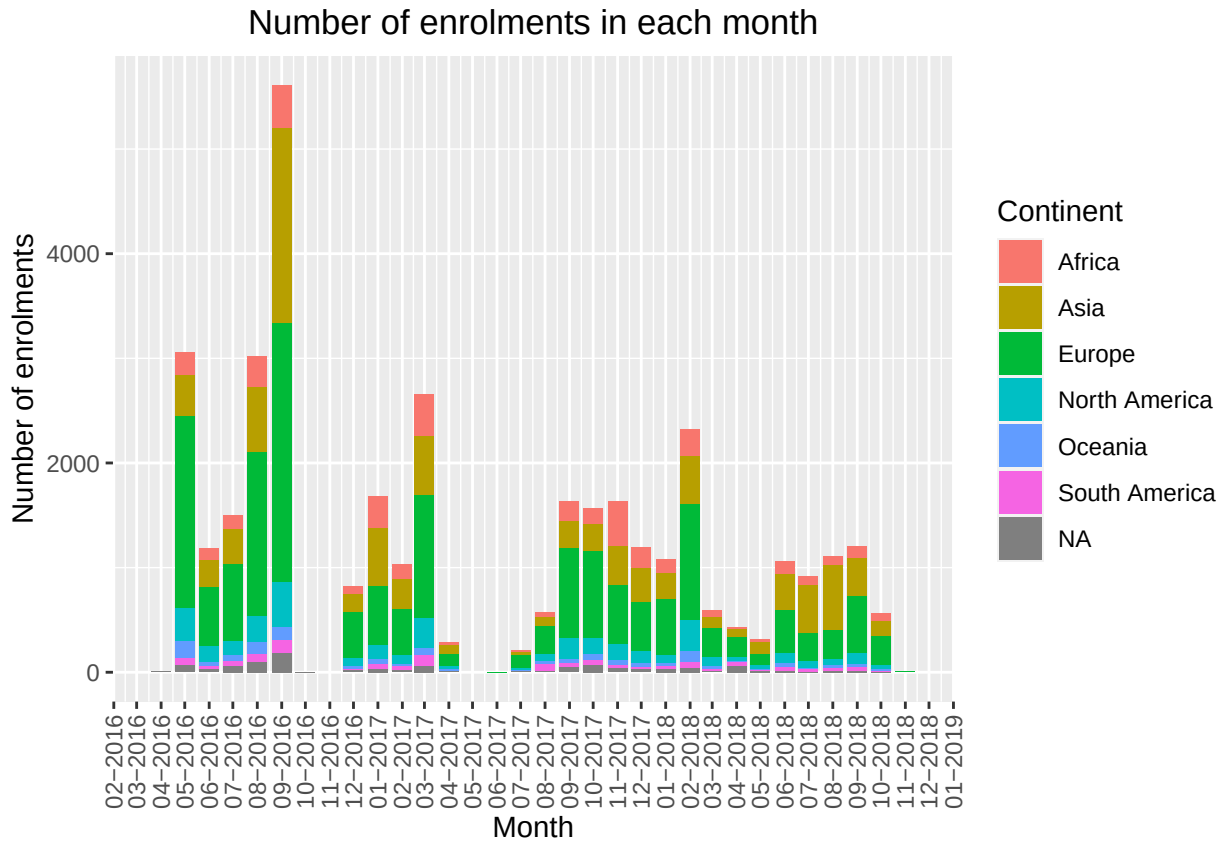

Figure 3: Stacked bar chart showing the number of enrolments from each continent in each month from Ferbruary 2016 to January 2019

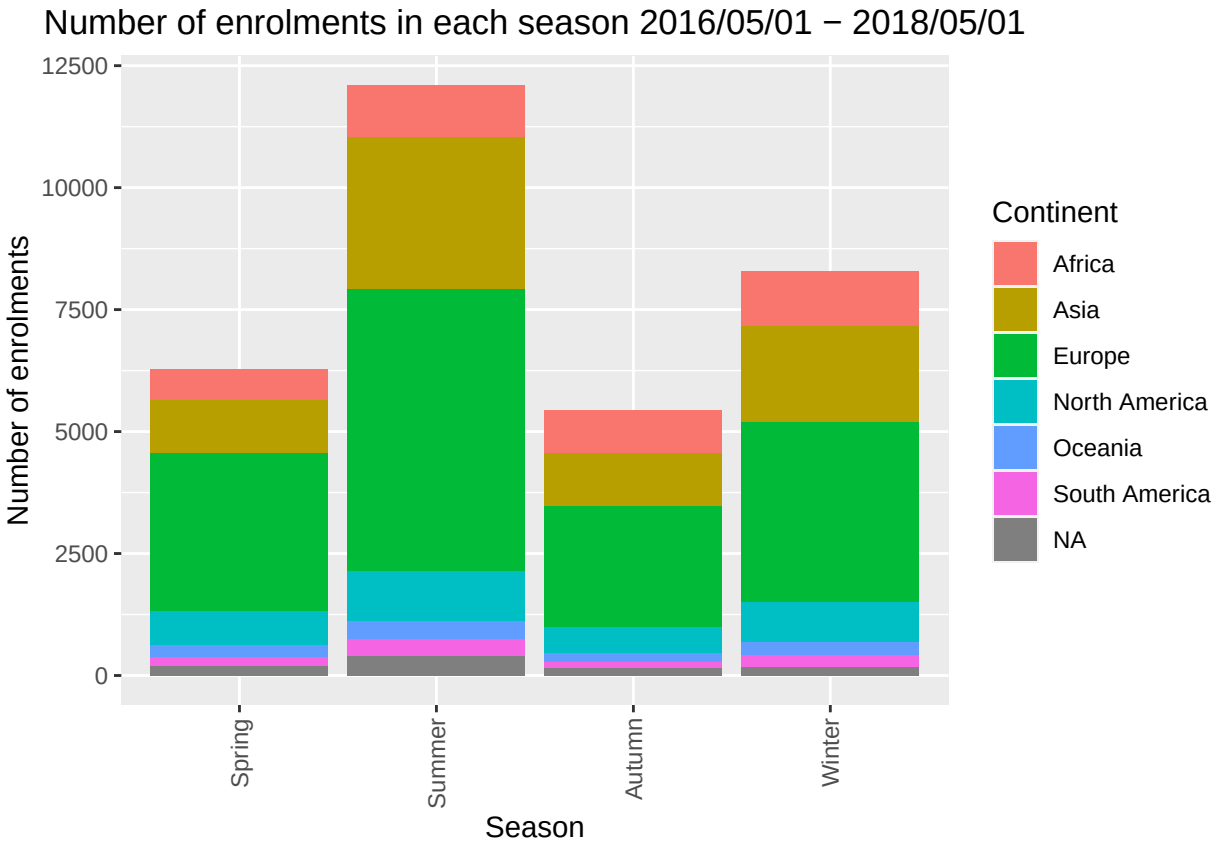

Figure 4: Stacked bar chart showing the number of enrolments from each continent per season from May 2016 to May 2018

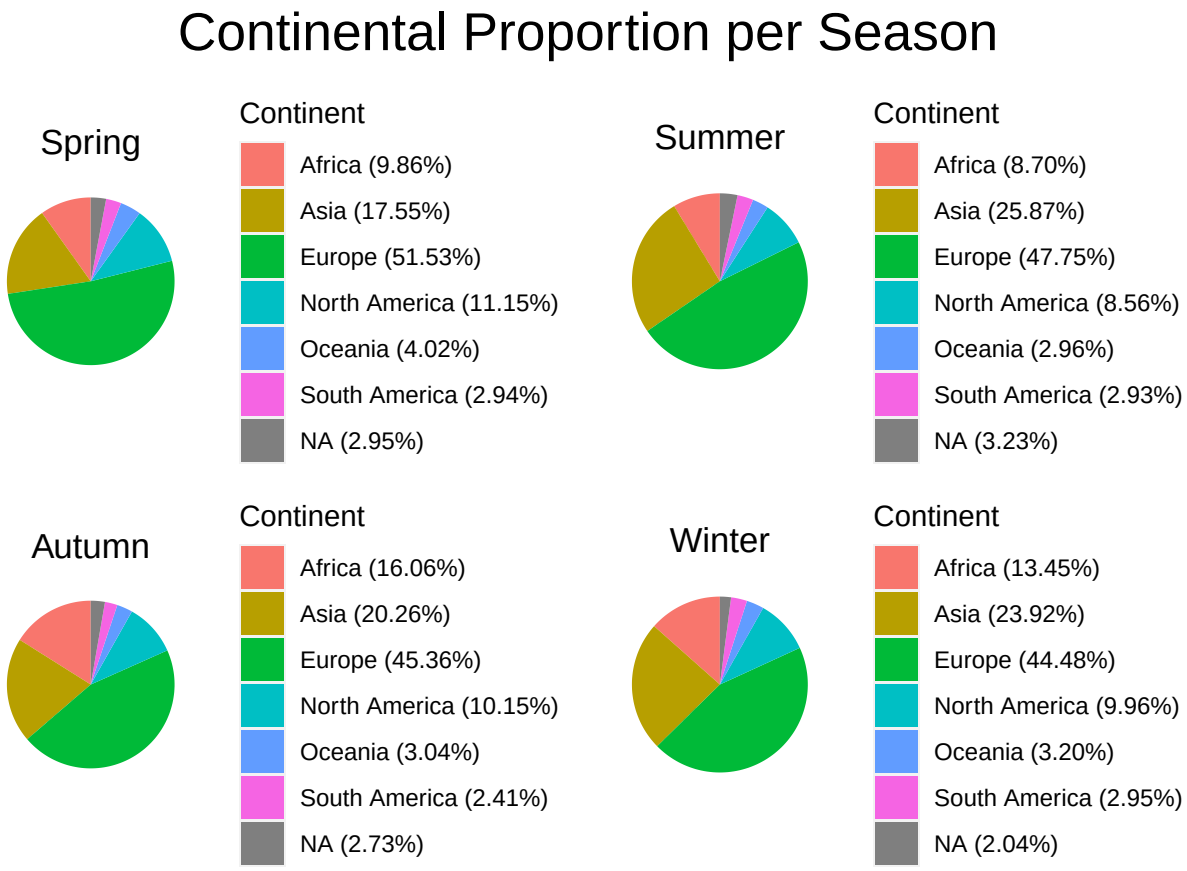

Figure 5: Pie charts showing the proportion of enrolments from each continent per season from May 2016 to May 2018