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: Online Courses and Degrees from Top Universities). 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 (Shacklock, X., 2016). 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 the course infrastructure 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 was 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) 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 data frame was made, transforming the data so that this new data frame contained a count of the number of enrolments in each month of the course. Months that contained no enrolments were not included in this data frame but were to be accounted for on future graphs. Finally, another column was added to the data frame 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 code for this can be found in the eda.R file in the src folder of the FutureLearn repository.

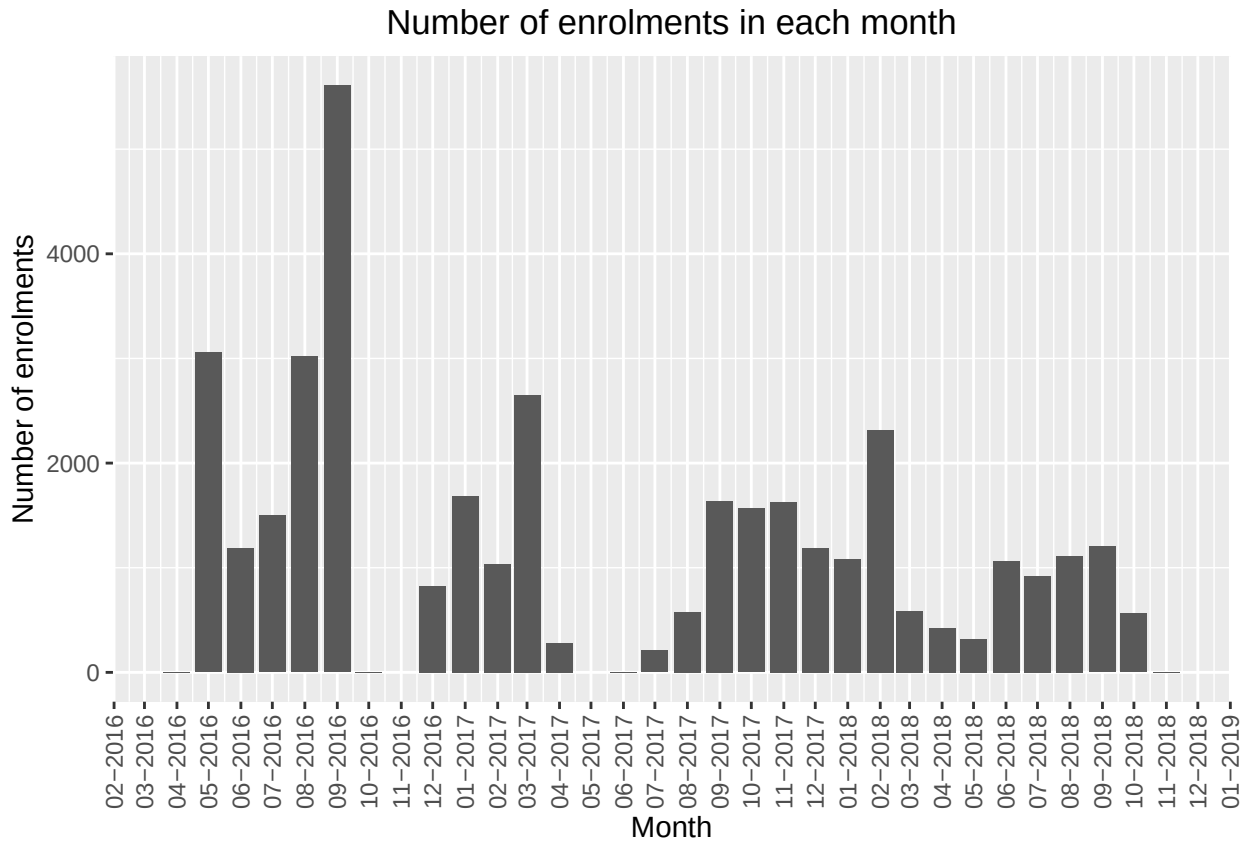

Figure 1: Bar chart showing the number of enrolments in each month from Ferbruary 2016 to January 2019

Evaluation

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.

In regard to the data mining goals, the knowledge that there exists periods of time both when enrolments were high and when they were low, helped to reach these goals however without further investigation it was difficult to make any concrete assumptions. It is for this reason that progression towards the business success criteria was also not at its most potential however these result did leave opportunity for a directed second cycle of investigation.

Focusing on the undulating pattern, this suggested that there may be a potential seasonal pattern to the enrolment numbers and therefore this was to be investigated in the succeeding cycle. In addition to this the fact that there were dipping in enrolment numbers in the same months of consecutive year suggests this pattern may align with global seasonal change for example, the Earth's seasons.

Cycle 2

Business Understanding

Due to the business objectives and success criteria not being achieved in the first cycle a second cycle of analysis was conducted. The business objectives and success criteria largely stayed consistent with the first cycle however with the knowledge of a potential pattern, FutureLearn would also like to gain some idea on why this pattern may have been occurring.

The data mining goals for this cycle carried some of the key themes from the previous cycle, although more concise as the goals were now to identify if the rise and fall of enrolment numbers aligned with that of the changing of the Earth's seasons. In order to investigate this an additional assumption had to be made. This assumption was that the Earth's seasons changed around the time of the equinoxes and solstices. That is to say that every year, at the spring equinox(20th March) the season changes from winter to spring, at the summer solstice (21st June) the season changes from spring to summer, at the autumn equinox (22nd September) the season changes from summer to autumn, and at the winter solstice (21st December) the season changes from autumn to winter. The success of this data mining would be judged by the ability to form sound conclusions about whether the pattern of enrolment numbers aligns or does not align with that of the Earth's seasons.

Data Understanding

The data did not already have the seasons allocated to the enrolment, however as the enrolments were dated this was possible to assign during the data preparation. In addition, the data did not span across an exact whole number of years and so in order to maintain a consistent number of each season across the investigation a two year window was selected (1st May 2016 - 1st May 2018). This was the largest number of whole years that could be extracted from the data and so was chosen to increase reliability.

Data Preparation

The transforming of data for this cycle can be found in the 02-A.R file in the munge folder of the FutureLearn repository. For this cycle a column indicating which season the enrolment took place had to be added to the original data frame that contained the enrolment data on all over the learners from the seven runs. The assignment of the seasons was done by creating a function that examines a date and returns the season based on the previously mentioned equinoxes and solstices. This function can be found in helpers.R file in the lib folder.

Next, a new data frame was created by filtering the enrolments from the larger original data frame to those that fell in the selected two year time period, and then grouping the enrolments by season and performing a count on the numbers in each season.

Now the data was ready to be plotted and was done so, again using the `ggplot2` package, to plot the bar chart seen in Figure 2. The code for this can again be found in the eda.R file in the src folder of the FutureLearn repository.

Evaluation

From looking at Figure 2 it was clear to see the undulating pattern seen from the first cycle with peaks in summer and winter (with the peak in summer being extremely prominent), and troughs in spring and autumn. It could be suggested from this that FutureLearn and Newcastle University could prepare for these increased and decreased numbers of learners by adjusting their staff counts for the course depending on the time of year so that regardless of when they are completing the course, each learner has a sufficient level of staff support. In addition to this there may be value to FutureLearn allocating a larger proportion of their

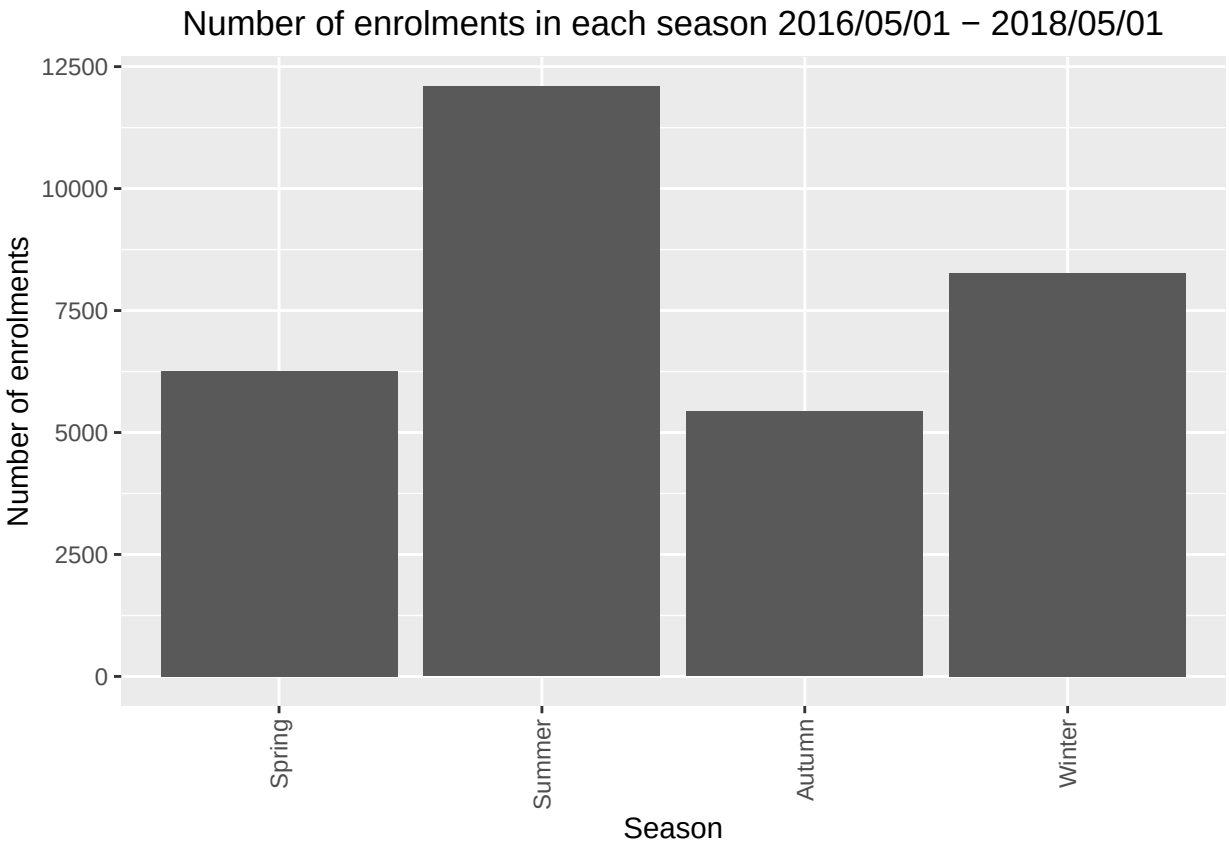

Figure 2: Bar chart showing the number of enrolments per season from 1st May 2016 to 1st May 2018

advertising budget towards attracting more consumer in these busier periods. Putting these results into context it was suggested that perhaps these seasonal patterns were due to the breaks people are allocated in these seasons for example summer holidays. However, due to the course being available internationally, this raised a question of "how do these results differ depending on the location of the individual enrolled on the course?" This question was raised with the knowledge that different parts of the globe experience the seasons at different periods of time and for this a new cycle was to be undertaken.

Cycle 3

Business Understanding

Although the general business aims of identifying a method to increase learner satisfaction and to increase the total number of enrolments have been met through the suggested plans of allocating of staff numbers depending on enrolment levels to accommodate for demand and allocate advertising funds to prepare for busier periods respectively, the question raised at the end of the previous cycle drew attention to the idea that these plans can be potentially refined and improved upon by taking location in to consideration.

Therefore the data mining goal of this cycle was to identify if there were any significant differences between enrolment demographics at different time periods through the year and then in turn use these results to potentially refine the previous recommendations to FutureLearn.

The demographics being investigated in this cycle will be geographic, more specifically continental. This was chosen under the assumption that countries in the same continent will be close enough in geographical distance to experience the different seasons at the roughly the same time.

Data Understanding

While the data did not directly specify in which continent each learner was located, there was a column that listed a detected country in the form of an ISO 3166-1 alpha-2 code. Using this a continent was assigned to the individual during the data preparation. In some instances there was no detected country and so these individuals will have an "NA" in their continent column. It is also assumed for this investigation that all detected countries are correct, however there is an understanding that some learners may be completing the course using a VPN and hence their detected location may be inaccurate.

Data Preparation

Prior to assigning each learner a continent, a vector was created for each continent containing the alpha-2 codes for every country in that continent. Next using the `dplyr` package, the larger enrolments data frame was transformed to create a new data frame that contained a count of the number of individuals enrolled from each continent in each month. Again, using the `lubridate` package, another column was added to the data frame that contained the date of the first day of each of the months for the x-axis positioning. This can be seen in the 03-A.R file in the munge folder. This data was then plotted using `ggplot2` as the stacked bar chart in Figure 3. The code for this plot can be seen in the eda.R file in the src folder.

Next, similar to the data preparation in the second cycle, a new data frame was created by filtering the enrolements from the larger enrolments data frame to those that fell in the selected two year time period, except this time the individuals were also assigned a continent. The learners were then grouped by the season they enrolled and the continent from which they were detected. A count was performed for each group. This was then plotted to give the stacked bar chart in Figure 4.

Finally, a data frame was made for each season by filtering the larger enrolments data frame to enrolments that fell in the two year period and that that were part of the specified season. These were then assigned a continent which they were then grouped by and a count for each was performed. These four data frames

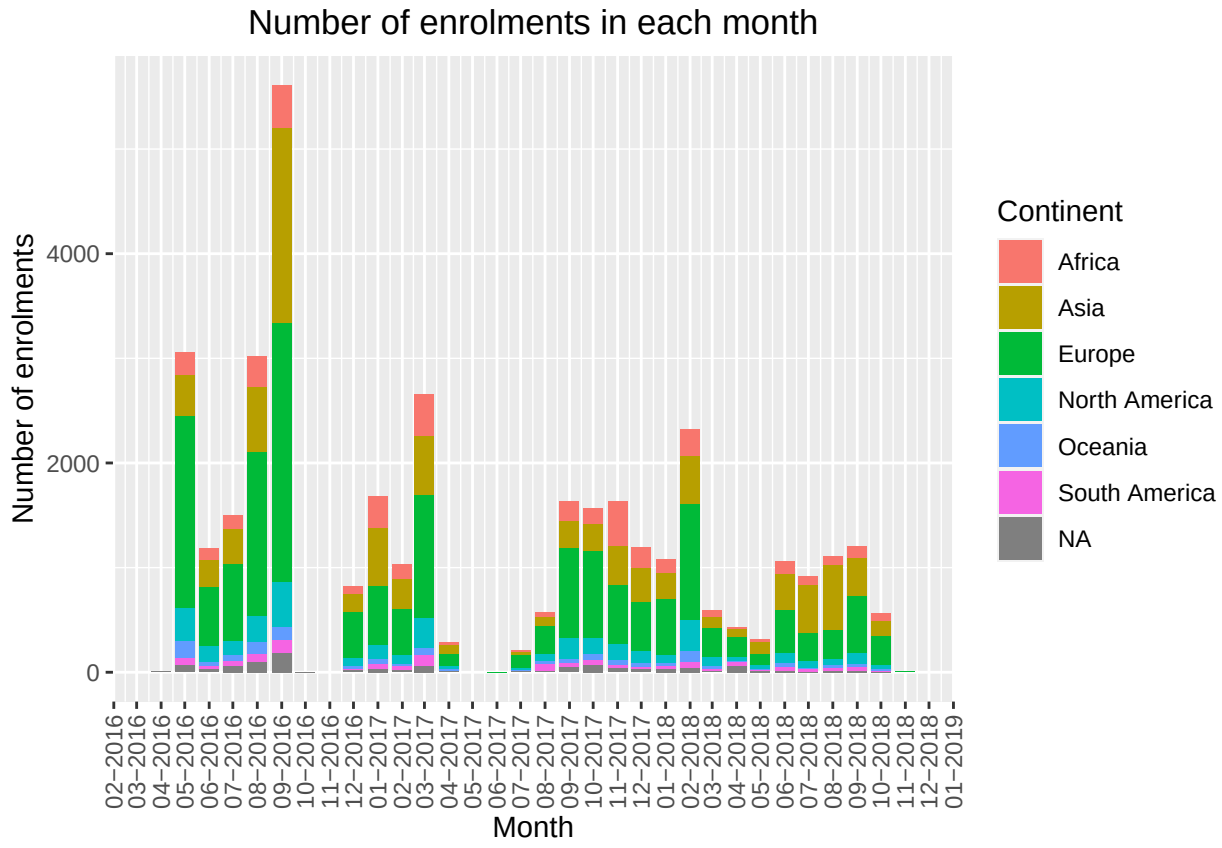

Figure 3: Stacked bar chart showing the number of enrolments from each continent in each month from Ferbruary 2016 to January 2019

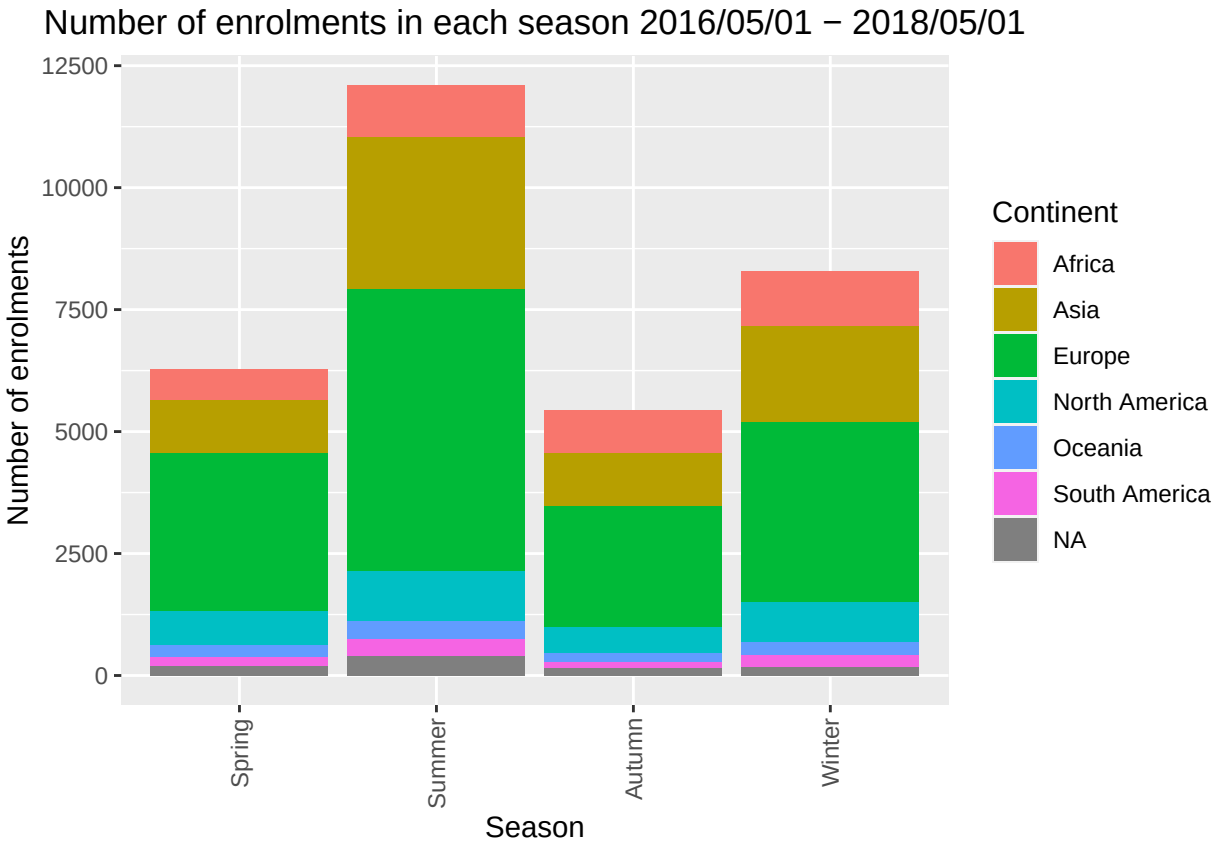

Figure 4: Stacked bar chart showing the number of enrolments from each continent per season from May 2016 to May 2018

were then plotted as the pie charts seen in Figure 5 with the percentage proportion for each continent in each season being calculated and included in the legend for each pie chart.

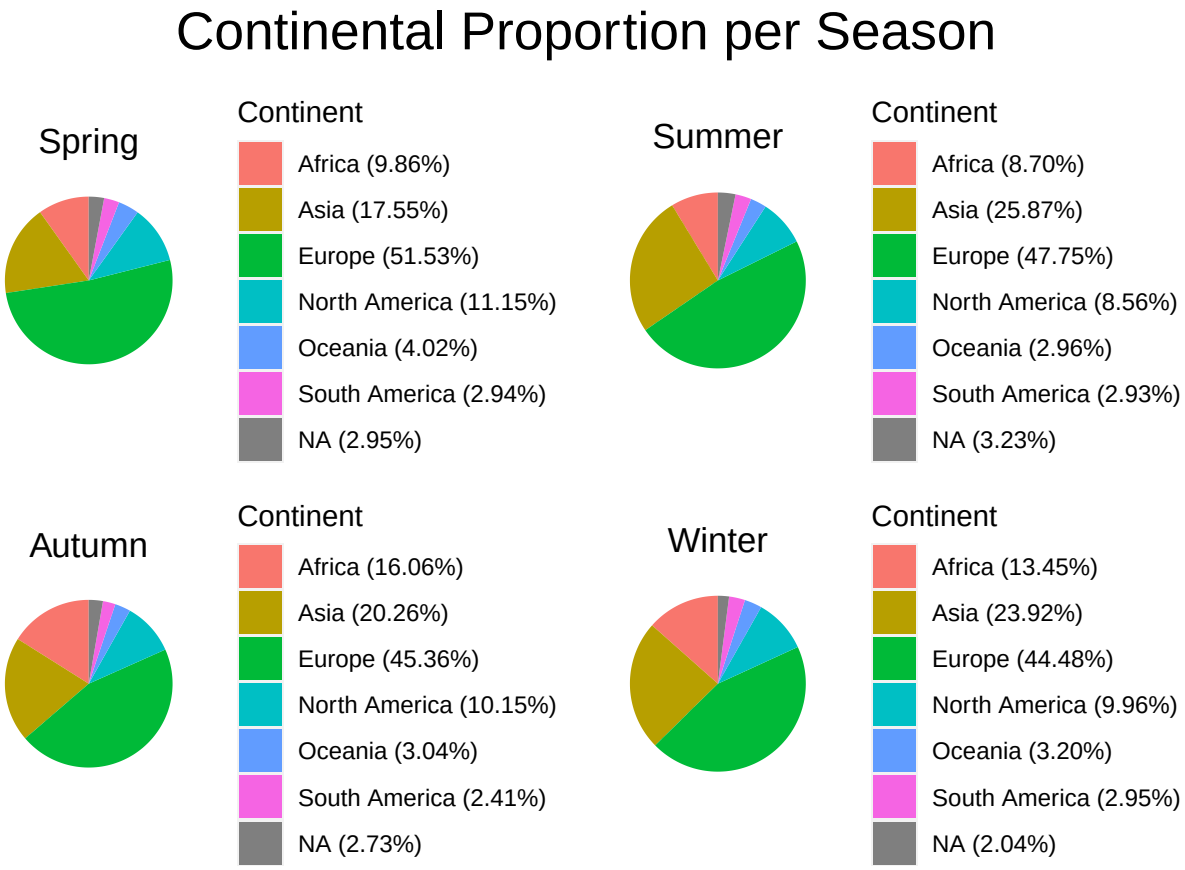

Figure 5: Pie charts showing the proportion of enrolments from each continent per season from May 2016 to May 2018

Evaluation

Looking at Figures 3, 4 and 5 it was clear to see that the largest proportion of learners were situated in Europe (green), followed by Asia (yellow). From this it was suggested that it could be useful to employ staff that are able to speak languages commonly spoken in these areas of the Earth for example English, Mandarin and Hindi in order to accommodate for the large number of students from these areas. Focusing on Figure 5, comparing the percentages between the seasons it could be seen that the proportion of learners from Asia was highest in summer (25.87%) and lowest in spring (17.55%) so it could be suggested that it may be sensible to employ more staff who speak languages commonly spoken in Asia during summer periods than during spring periods. Additionally, the proportion of individuals from Africa is largest during the autumn periods (16.06% compared with a low of 8.70% in summer periods) so it may be useful during autumn periods to employ some staff who speak languages commonly spoken in Africa such as Arabic. These changes were suggested with the intention of making the learners more comfortable on the course and potentially enable them the opportunity to ask questions in their native languages. Also if people knew they were to have this support on the course they may be more inclined to sign up to a FutureLearn course and hence increase the number of enrolments. Making students feel more comfortable and increasing enrolments both fall in line with FutureLearn's business goals. Moreover, this data could also be used by FutureLearn to target specific areas at different times when advertising their courses with knowledge of when people from those areas are more likely to sign up which would potentially both increase the number of enrolments, giving them a larger

share of the market and also increase the efficiency of their advertising budget, in turn possibly saving them money.

Some final limitations of this investigation are that data from only one course was used so results are limited as the observations from this project may be different for other courses, and also the length of time over which data was collected may not be long enough to accurately predict trends. To combat this in future, data collected from multiple courses over a longer period of time may be used. Finally, as a recommendation for additional cycles to this project it might be useful to investigate which individual countries have the largest proportion of enrolments.

References

- Shacklock, X., 2016. *From bricks to clicks: The potential of data and analytics in higher education.* London: Higher Education Commission.
- FutureLearn. n.d. FutureLearn: Online Courses and Degrees from Top Universities. [online] Available at: https://futurelearn.com/ [Accessed 29 November 2021].
- Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.