CSC8631- Data Management and Explanatory Data Analysis Project

SATISH PILLA-B210472095

5/20/2022

Contents

1 BUSINESS UNDERSTANDING 2

2 DATA UNDERSTANDING 2

2.1 Exploratory Data Analysis (Numerical Analysis) 2

3 OBJECTIVE 4

3.1 Selecting The Right Audience. 4

3.2 Investigating The Delivery Methods. 4

4 DATA PREPRATION 4

4.1 Merging Data: 4

4.2 Independent Data 5

4.3 Removing Unkowns 5

5 MODELING 5

6 DESIGN AND IMPLEMENTATION 5

6.1 Exploratory Data Analysis (Graphical Analysis) 5

6.2 Selecting The Right Audience 9

6.3 Investigating The Delivery Methods 13

7 EVALUATION 14

8 DEPLOYMENT 15

9 CONCLUSION 15

10 FUTURE SCOPE 15

1 BUSINESS UNDERSTANDING

Future Learn, an online learning platform, offers a course called Cyber Security: Safety at Home, Online. This three-week self-paced course focuses on critical cyber security issues. The courses are divided into three parts, each of which is studied over three weeks: internet privacy, payment security, and home security. From 2016 to 2018, the course was offered seven times. Throughout the course, a variety of data was collected. This project aims to evaluate the data and get some useful insights that can be utilized to improve the course and establish a viable business model for future courses.

2 DATA UNDERSTANDING

Each time the course is run, around 8 data files are generated.

Datasets:

1. cyber-security-enrolments
2. cyber-security-leaving-survey-response
3. cyber-security-step-activity
4. cyber-security-video-stats
5. cyber-security-step-activity
6. cyber-security-archtype-survey-responses
7. cyber-security-question-response
8. cyber-security-weekly sentiment-survey

The enrollments data set contains the learners' information, including their unique IDs, the course's enrolled and unenrolled dates and times, and additional fields such as gender, nation, age, highest education level, employment location, and current employment status. The vast majority of the students' data was not captured. A figure or action that symbolizes universal human nature patterns is known as an archetype. The information that categorizes the students is contained in the archetype data collection.

The details of the learners who left the course at what stage and for what reason can be found in the leaving survey response dataset. Each learner's responses to quiz questions conducted at a certain moment in each week are collected in the question response data collection. The quiz's results are also included. Learners who started and left the step at particular times are recorded in the step activity data collection. The video statistics data set includes videos of certain steps with titles and information such as time, views, downloads, viewed percentage, and learners' viewed continents. The weekly survey replies data collection contains the learner's input on the course.

2.1 Exploratory Data Analysis (Numerical Analysis)

Overview of Enrollment Dataset

Descriptions	Value
Sample size (nrow)	37296
No. of variables (ncol)	13
No. of numeric/interger variables	0
No. of factor variables	0
No. of text variables	13
No. of logical variables	0
No. of identifier variables	0
No. of date variables	0

Descriptions	Value
No. of zero variance variables (uniform)	0
%. of variables having complete cases	61.54% (8)
%. of variables having >0% and <50% missing cases	15.38% (2)
%. of variables having >=50% and <90% missing cases	0% (0)
%. of variables having >=90% missing cases	23.08% (3)

Overview of Leaving Survey Dataset

Descriptions	Value
Sample size (nrow)	403
No. of variables (ncol)	8
No. of numeric/interger variables	4
No. of factor variables	0
No. of text variables	4
No. of logical variables	0
No. of identifier variables	2
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	50% (4)
%. of variables having >0% and <50% missing cases	0% (0)
%. of variables having >=50% and <90% missing cases	37.5% (3)
%. of variables having >=90% missing cases	12.5% (1)

Overview of Step Activity Dataset

Descriptions	Value
Sample size (nrow)	423072
No. of variables (ncol)	6
No. of numeric/interger variables	3
No. of factor variables	0
No. of text variables	3
No. of logical variables	0
No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	83.33% (5)
%. of variables having >0% and <50% missing cases	16.67% (1)
%. of variables having >=50% and <90% missing cases	0% (0)
%. of variables having >=90% missing cases	0% (0)

Overview of Video Stat Dataset

Descriptions	Value
Sample size (nrow)	65494
No. of variables (ncol)	6
No. of numeric/interger variables	3
No. of factor variables	0
No. of text variables	3

Descriptions	Value
No. of logical variables	0
No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	83.33% (5)
%. of variables having >0% and <50% missing cases	16.67% (1)
%. of variables having >=50% and <90% missing cases	0% (0)
%. of variables having >=90% missing cases	0% (0)

3 OBJECTIVE

This investigation has primarily two objectives:

3.1 Selecting The Right Audience.

To select the right audience we must know from which location or country the students are getting most enrolled into the course, Once we know from which location the most people are getting enrolled we can do further analysis on learners gender, education qualification, employment background, status and age etc. All this analysis will help the course provider in targeting the right audience and helps us in better understanding of the students or people showing a strong interest in the course.

3.2 Investigating The Delivery Methods.

Once we know who the target audience are, we can make course interesting and appealing to the learners. We can do this by investigating the delivery methods. We know that the course is primarily offered in four formats: video, articles, discussion, and a quiz which is conducted at the end of each week. Once we know which format is most popular among the students we can use that format to attract more audience to learn the course.

4 DATA PREPRATION

For my first analysis(Selecting the Right Audience) I have considered cyber security enrollment data files from all of the runs from 1 to 7. For the second analysis(Investigating the delivery methods) I am considering using cyber security stats dataset to know which delivery method is more efficient.

4.1 Merging Data:

Combining data from all iterations of the same file genre. For example, enrollment data from each iteration is combined row by row (rbind) on top of each other and similarly it is done for stat dataset. The goal of merging data from several iterations is to create a more comprehensive picture of how the course is functioning, who is enrolling in it, and how they are using it.

4.2 Independent Data

To see how the data changes between different runs of the course, all different file genres from various iterations are also kept separately. For example to know the unique count of learner, we will use each individual enrollments dataset and plot how the course is changing over time form when the course was for 1st to 7th time(2016 to 2018).

4.3 Removing Unkowns

Since the majority of the fields are unknown, all of the analysis are done by deleting "unknown" values. As a result, this study may or may not represent the genuine population distribution. Since most of the values in the country column are "unknown," we are using "detected country" rather than "country" from enrollment data to determine where the majority of learners are enrolling.

5 MODELING

This research was conducted using NUMERICAL and GRAPHICAL summaries as modelling tools. This was accomplished by combining R-markdown with a variety of supporting libraries such as ggplot, dplyr, and many others. All of these are combined utilizing the Project template for better project management and reproducibility. The CRISP-DM approach is used for the analysis.

6 DESIGN AND IMPLEMENTATION

6.1 Exploratory Data Analysis (Graphical Analysis)

How has the popularity of a course changed over the last seven years?

To begin, we can look at how the popularity of the course has changed during the last seven times it has been run. This can be investigated by looking at the cyber-security-enrollments data set from all of the Runs and seeing how the numbers have evolved over time.

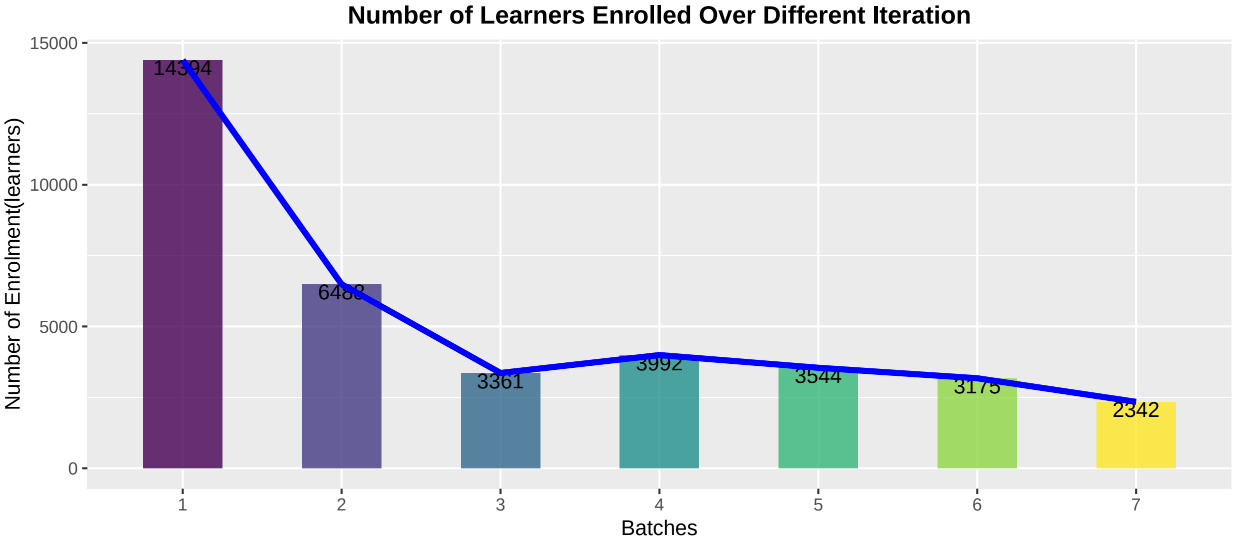

Figure 1: Number of learners enrolled over different iteration

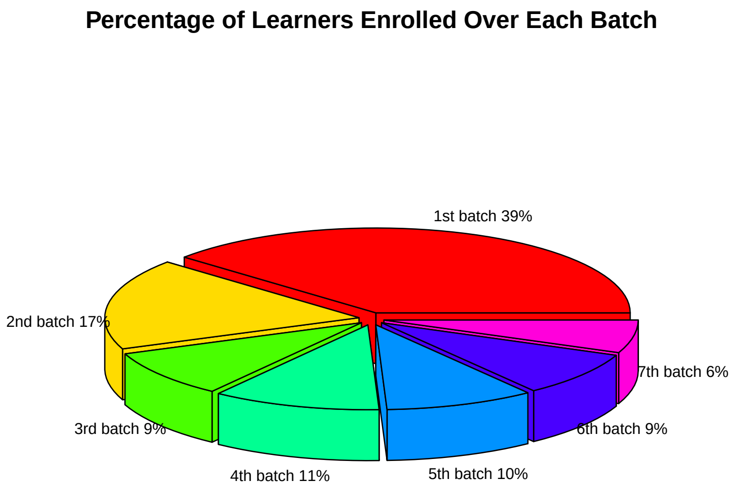

Figure 2: Percentage of Learners Enrolled Over Each Batch

From figure 1 and 2 we can clearly see that there were around 14398 which is 39% of total enrollments over all the course run from 2016 to 2018. The enrollments got decreased as the course run and became less than half by the end of last run. In the 7th run the total enrollments were only 2342 which is just 6% of the total enrollments over all the course run. We can say that as time progressed the popularity has decreased for the course. There may be n numbers of reason for less enrollments over the period of time so We'll try to figure out what's going wrong based on a variety of indicators, such as why students are dropping out and how they feel about the course.

Reason(s) for Un-enrollment?

We'd like to know how many students have registered up for the course and how many have dropped out. UN-enrollment reasons and the percentage of students who enrolled but did not start the course, So at first we are binding all the unerollment dataset and step-activity datase, then we are calculating total number of students from enrollment dataset, number of students unregistered from the course from leaving survey response, number of students who started the course from step-activity dataset. From below table we can see that total 35225 students where enrolled over the period of 7 runs. Over the 7 runs only 370 people un-registered the course which is very good number, but at the same time 20285 students did not start the course, So mostly they will fall under un-register category. One of the positive sign is that out 35225 students 14570 student started the course.

Summary	Count_number
Total students Register	35225
Students Un-register	370
Student did not start the Course	20285
Students Started the Course	14570

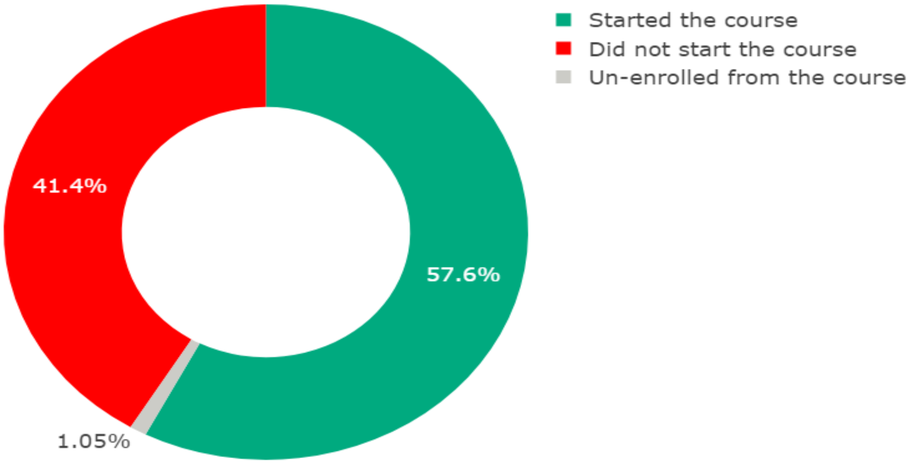

Figure 3: Student Started ,Un-enrolled And Did not Start The Course

From figure 3 we can get an overall view of the the students who started the course, did not start the course and un-enrolled from the course. So according to the figure, 57.6 percent have only begun the course. More than a 40% of students who enrolled in the course never started it, and only 1% percent dropped out of the course when the course was run over 7 times.

Let's check the reasons for Un-enrollment which can be found in the leaving.survey.responses data set.

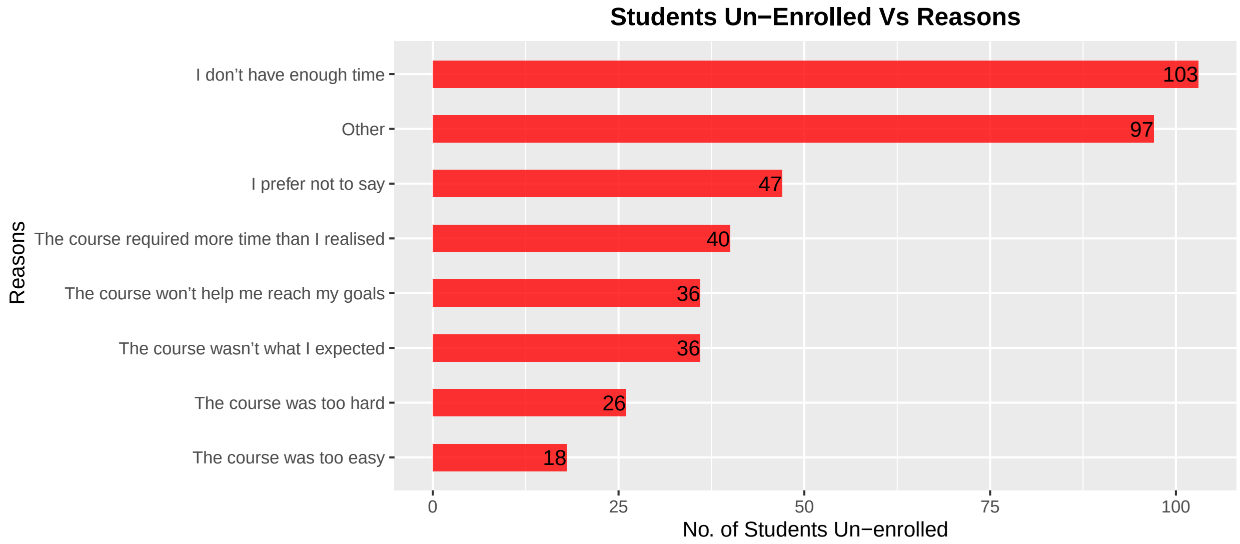

Figure 4: Student Un-Enrolled vs Reasons

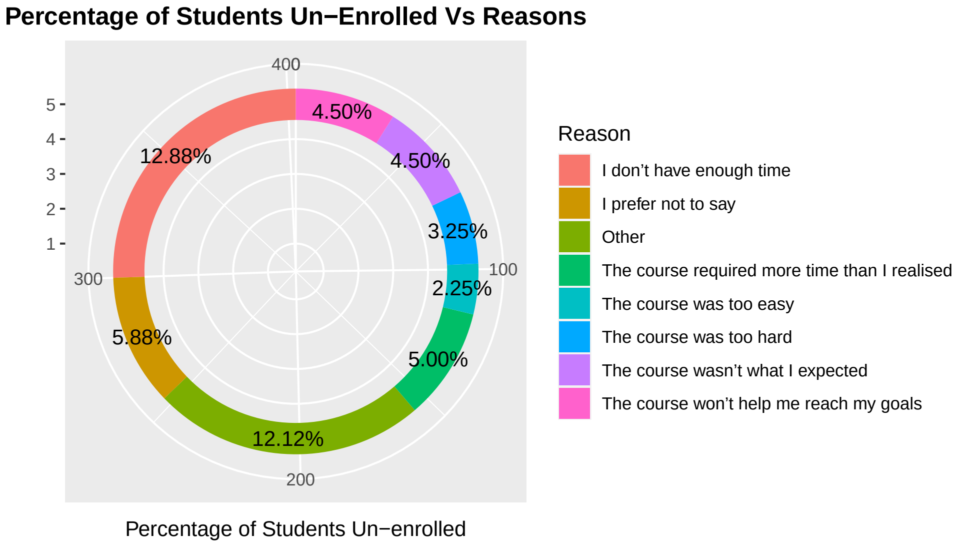

Figure 5: Percentage of Student Un-Enrolled vs Reasons

The figure 4 and 5 depicts the most common reasons for un-enrollment. With 103 students, the most common reason for dropping out is a lack of time, which is 13% of the total students un-enrolled from the course. Second most reason for un-enrollment was other with 97 students which is around 12%. Whereas 18 student dropped out because the course was too simple which constitute of 2.25% of total students unenrolled.

6.2 Selecting The Right Audience

But our primary objective is to investigate the location where course is most popular, so let's now check for the countries were the course is most popular.

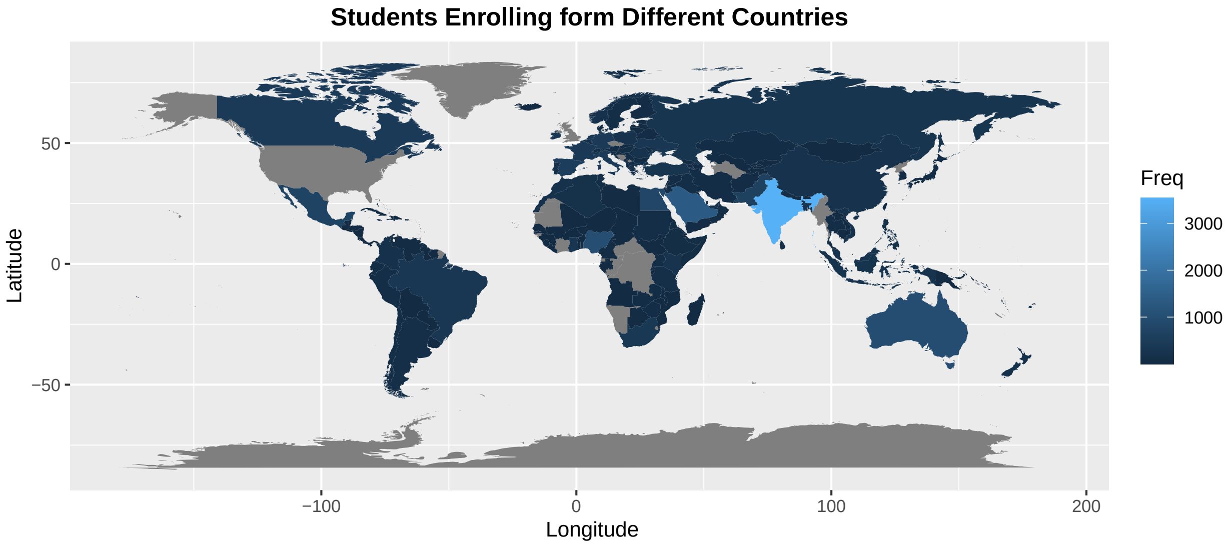

Figure 6: Students Enrolling form Different Countries

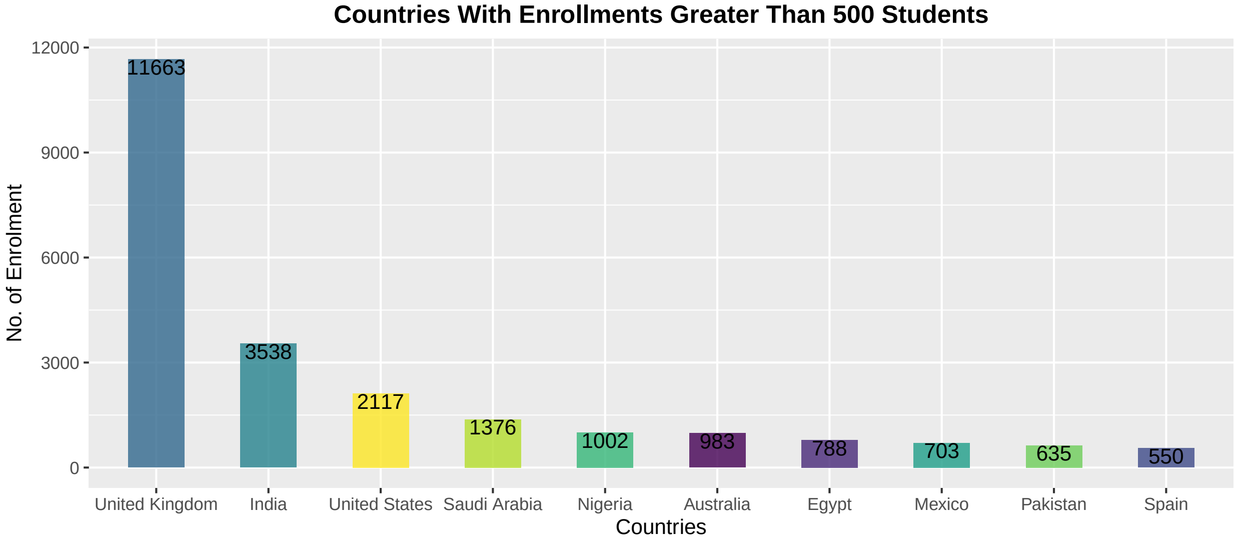

Figure 7: Countries With Enrollments Greater Than 500 Students

Evaluation based on 1st Analysis: At first we are plotting a heat map to know from which part of the world most number of students are getting enrolled. Figures 7 show that the majority of students are from the United Kingdom, India, and the United States, with a small number from Australia, Saudi Arabia, Nigeria, Mexico, and Russia. From heat map we know from which countries students are getting enrolled, let try to get top 10 countries with more than 500 enrollments. So we are plotting a bar graph, Figure 4 show the top 10 countries with most number of enrollments. From the figure we can clearly see that more than 10000 students got enrolled form united kingdom , followed by India and united states with 3538 and 2117 students. So based on above data the course provider may focus on these countries to attract more students and build a strong business model in these countries.

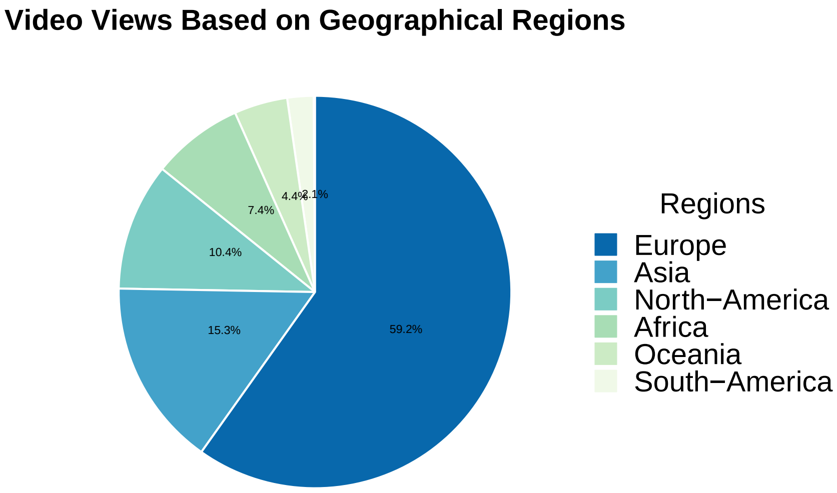

Figure 8: Video views based on geographical regions

The above pie chart with video views based on geographical Region and the countries with high enrollment are highly correlated .So based on this we can see than about 85 % of the enrollments are from Europe , Asia and North-America. These are the regions who use the internet the most, so we could develop marketing strategies which could get more students enrolled.

Now we know that from which location the people are getting most enrolled, let's now check and focus on learner's gender, education background and various other factors which will help us understand our audience in better way.

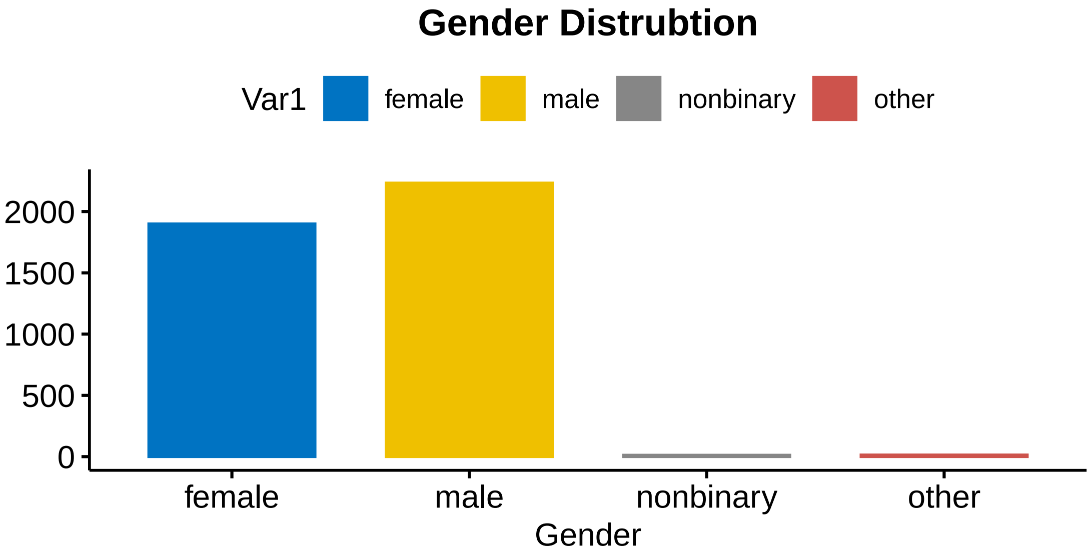

Figure 9: Gender Distrubtion

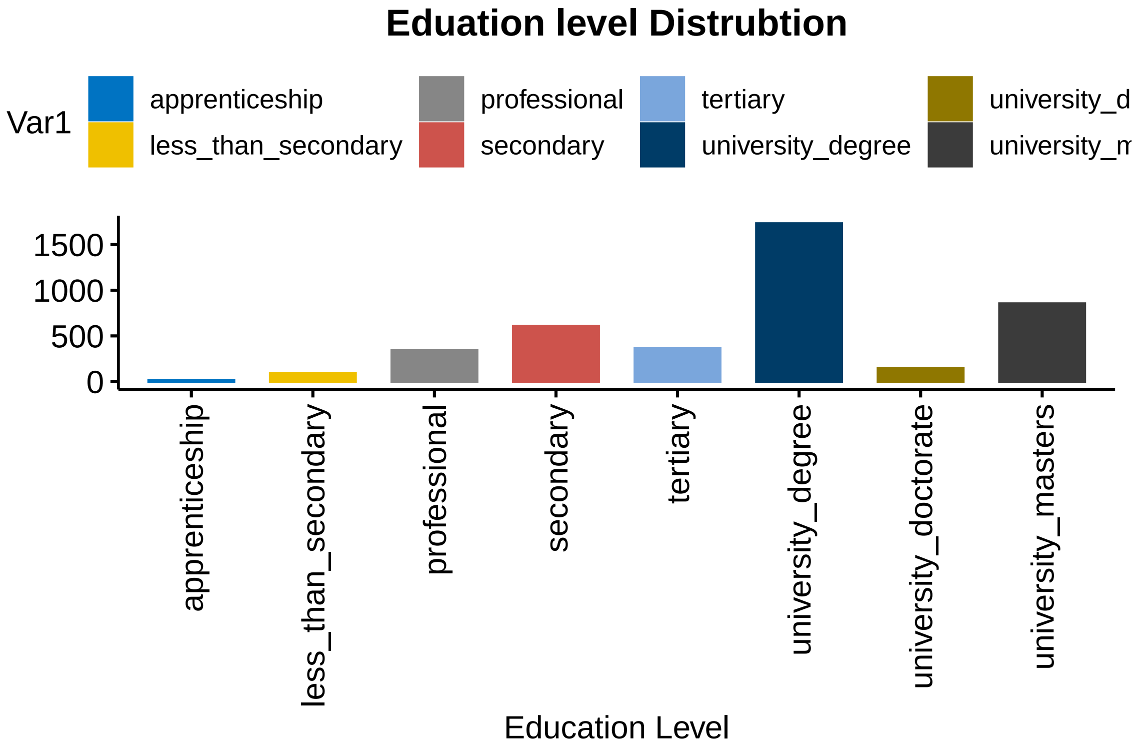

Figure 10: Education level Distrubtion

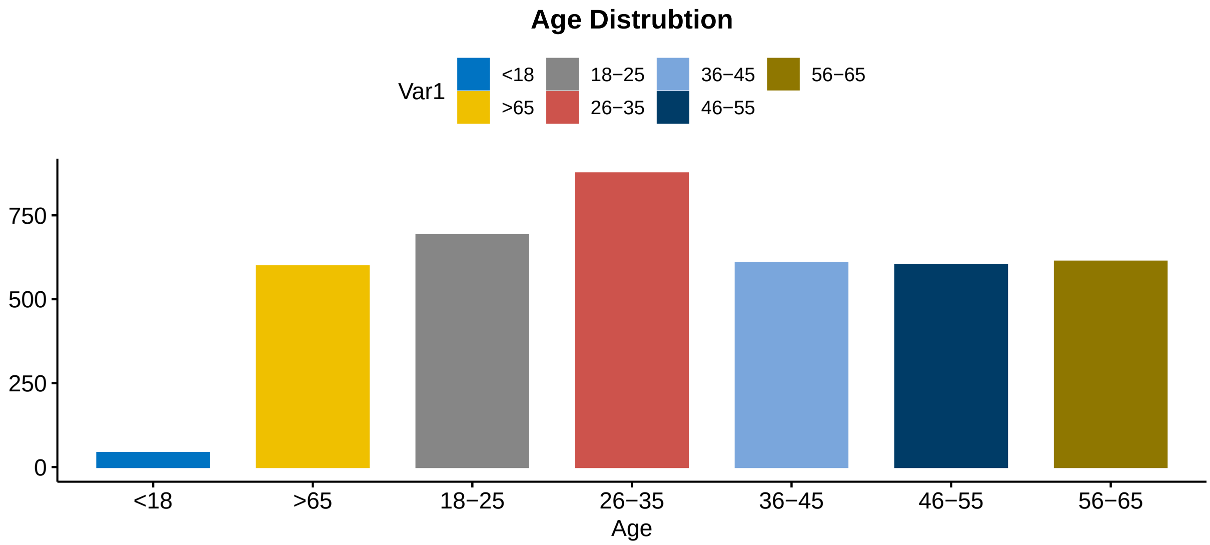

Figure 11: Age Distrubtion

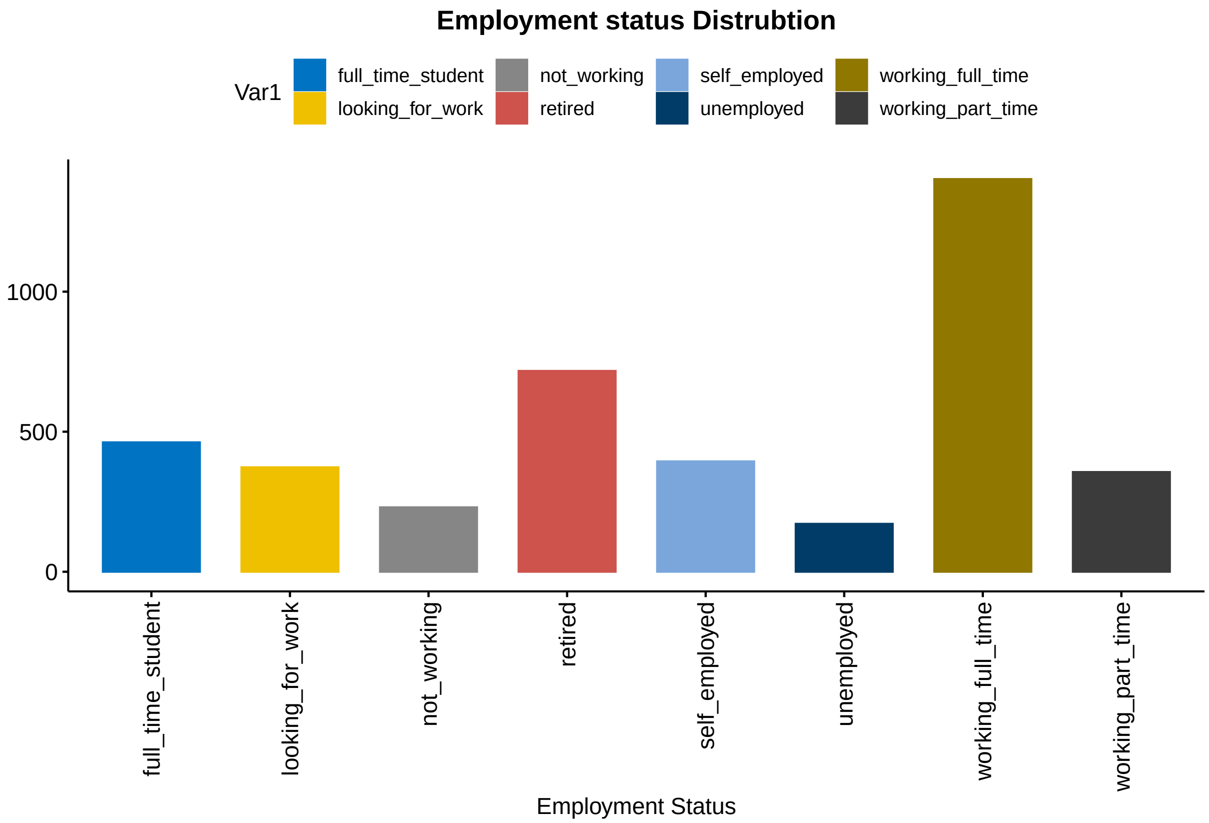

Figure 12: Employment status Distrubtion

From above figures we can clearly see that male learners are more than female learners. Most learners age range is in between 26 to 35. There are significant number of learner in every age group except for all less than 18 years. Most learners who enrolled the course hold university degree at last people who are working full time are showing very keen interest in learning the course. So a course provider may focus on the learners from all age range, having university degree, with full time job.

6.3 Investigating The Delivery Methods

From video stats data we know that the cyber security course can be accessed on multiple devices, let's see most widely used devices to access the content.

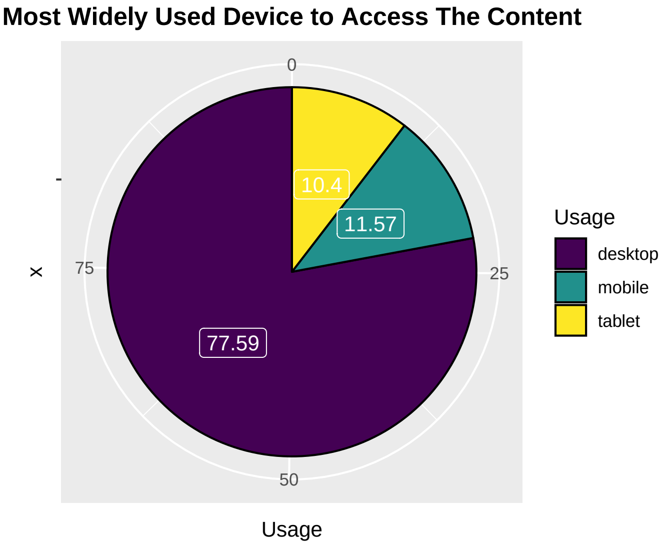

Figure 13: Most Widely Used Device to Access The Content

So from figure 13 we can clearly see that most learners accessed the course content through desktop. More than 70% student are accessing the course content through desktop. Less than 13% learners are accessing the course content through mobile and tablet.

TO know which delivery method is most efficient we will divide the course into three modules and find the numbers of unique students going through each module.

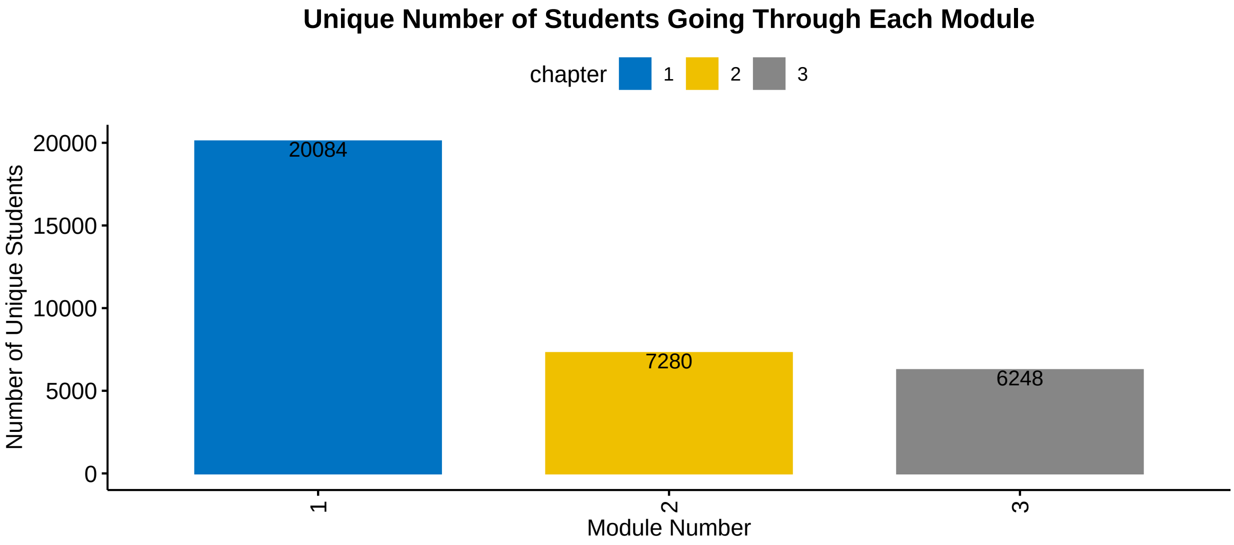

Figure 14: Unique Number of Students Going Through Each Module

From figure 13 we can see that around 20000 students visited the course module when the course was run over 7 iteration's. As the module progressed there is decline in the number of students participating in the modules. Only 6248 students visited the module 3.

Each module contains a number of steps. Each lesson is primarily delivered through one of the following methods: videos, quizzes, articles, and discussions. As the course progresses, we'd like to know how many unique students have proceeded through each of the delivery methods. I used the step activity data set for this analysis and did some pre-processing to separate out the step numbers based on delivery methods.

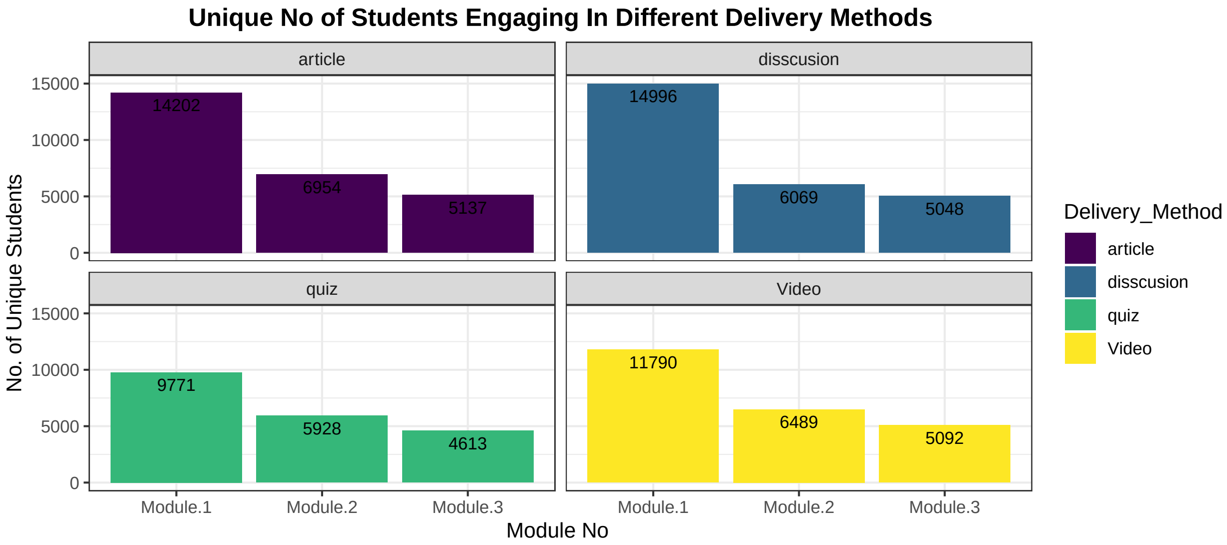

Figure 15: Unique No of Students Engaging In Different Delivery Methods

From the figure 15 we can see that the unique number of students engaging in different delivery method is highly correlated with the data student visiting each modules. Almost all the delivery methods are highly correlated with student visiting each module. Most preferable method of delivery is either video or articles.

Evaluation based on 2nd Analysis: At first we find out how students or learners are accessing the course. We saw that more than 75% percent learners are accessing the course through desktop. After that we find out which delivery method is most preferable by learners, we find out that most preferable method is either video or articles. The course content creator can focus making video more interesting to learners.

7 EVALUATION

All of the analysis is carried out with the assumptions discussed in section 4 in mind. This may cause uncertainty, but we conducted numerous analyses for the same purpose on different data to back up our findings.

1. We verified people from each iteration and then using the entire data (merged data), a final plot was produced to analyse what kind of people enroll for the course more frequently.
2. To pinpoint the problem, we looked at the leaving survey, where a large number of fields were null, implying that the results could be skewed. Also, a large percentage of respondents stated "Other," so we have no idea what that means.
3. We used video stats data to determine which continent has the most learners, then we analysed detected country data from the enrollment dataset, and they both complemented each other. The majority of students are from Europe and live in the United Kingdom.
4. Since data was only provided from the third to the seventh iteration to generate estimates on which device was being utilised the most, analysis was done across 5 of the 7 runs of the course. We find out that most learners accessed the course through desktop.

5. We used the step activity data set for this analysis and did some pre-processing to separate out the step numbers based on delivery methods. We found that the most preferable delivery method is either video or articles.

8 DEPLOYMENT

The analysis was done in R, and it can be used to evaluate and compare subsequent runs of the course in the future.

9 CONCLUSION

We can conclude from the numerous analyses performed in this report that the course content is good. To improve the course the course provider may focus on the countries with most enrollments.People of all ages, from the IT and education sectors, and working full-time are among the people who the course provider may target for the ideal audience, according to the analysis. The course can be divided into multiple courses or reduce the quiz questions so that serious learners can take the assessments whenever they want and casual learners can go through the content and learn from the course without having to spend a lot of time on quiz. At the end the course content creator can focus making video and articles more interesting to learners.

10 FUTURE SCOPE

While the research conducted for this paper yielded some encouraging outcomes, there is still need for more research to be done. Furthermore, if the enrollment file had not been so sparse, the entire study of un-enrollments in particular would have portrayed the most precise intuitions. Another proposed option is to create a dash board using R shiny that displays details of student activity based on real-time feed data, allowing module course module leaders to better understand the students’ actions, perhaps lowering enrollments and increasing student engagement with the course.