Cyber Security Data Analysis Report

Roshan Pandey

11/11/2021

1. Introduction

Online learning platforms are getting popularity day by day and hence, generating tons of data and we all know data is the fuel for today's world. We can leverage this data to it's true potential to help the course or the platform to grow and gain more traction by improving the course even further. So, in this report we will try to answer questions like, which device learners use frequently to access the content, from which region more learners are taking up the course, over the various iterations of the course number of enrolled learners are increasing or decreasing.

2. Objective

There are mainly two objectives for this analysis:

2.1. Target Right Audience.

> We want to know from which **location** mostly people are enrolling for the course, what are their employment background and status, their age range etc. This will allow the course provider to target right audience. First, we will try to analyze which continent has most enrolled learners, then we will pin point the country. This way it will be easy to target right locations with precision for advertisement. After this, analysis will be done on learners gender, age range, heights education quaification, employment area and status to understand better which type of people are showing keen interest in the course and hence, personalized ads can be pushed to the right audience.

2.2. Course Improvement.

> Once right audience has been singled out, course can be improved to make it more appealing to the potential learers. This can be done in various ways such as, analysing which type of devices (**Desktop, Mobile Phone, Tablet**) are being used most often than others, is there any trend in usability of a particular device over different runs of the course. Analysis can also be done on sentiments of the learners who have taken the course in the past and peaking into the reasons why they chose to leave the course.

3. Data Description

Data consists of 7 iterations of the course, Cyber Security: Safety at Home, Online, in Life.

- "Cyber-security-1_archetype-survey-responses"- This file consists of responses of learners for archetype surveys for 1st iteration of the course.

- "Cyber-security-2_enrolments"- This file consists of Enrollment details of the learner for 2nd iteration of the course.
- "Cyber-security-3_leaving-survey-responses"- This file consists of responses of learners for course leaving survey for 3rd iteration.
- "Cyber-security-4_question-response"- This file consists of responses of learners for assessment questions for 4th iteration of the course.
- "Cyber-security-5_step-activity"- This file consists of step wise first visit and last completion time of learners for 5th iteration of the course.
- "Cyber-security-6_team-members"- This file consists of the team and user role of people for the 6th iteration of the course.
- "Cyber-security-6_video-stats"- This file consists of data related to the length of the video, no. of views, no. of downloads, device viewed on, part of world viewed from etc. for 6th iteration of the course.
- "Cyber-security-7_weekly-sentiment-survey-responses"- This file consists of responses of learners about their sentiments for the week's learning.

4. Data Processing

We are considering data into two different formats, mentioned bellow:

- Merging data: Combining data of same file genre from all the iterations. For example, enrollment data from each iteration is been combined row wise (rbind) on top of each other ad like wise for archetype, video stats, leaving survey, weekly sentiments, question response, etc.
  The idea behind combining the data from different iterations is to get a broader picture of how the course is performing, what type of people are joining the course, and how are they utilizing the course.
- Separate data: All different file genres from various iterations are kept separately to see how the data is changing between different runs of the course. This analysis will assist us to see whether there is a shift in trend as to how people are making use of the course. For example, are people preferring mobile devices more in recent batched or they still like to use desktop computers.

Text processing: For analyzing the overall sentiments of learners about the course, all the text from weekly sentiment survey from different iterations are combined into one string of text and then it was cleaned(removal of stop words, punctuation, every word to lower case, removal of special chars and worlds of length 3 or less). After this cleaned text was split into words and then sentiment score was calculated for each words.

IMPORTANT:

- All the analysis are done by removing "unknown" values and hence, this analysis may or may not represent the true population distribution because majority of the fields are unknown.
- To analyze from which country most learners are enrolling, we are considering "detected_country" rather than "country" from enrollment data because, most of the rows in country column is "unknown". We matched both the columns and most of the values that are not "unknown" in country column are same as detected_country column except for few(approx 13-14 values). Hence, we are assuming detected_country can be a good measure of figuring out which country has most learners enrolled.

5. Data Analysis

How is course performing over different iterations?

From figure 1 we can clearly observe that number of learners are decreasing over different iterations of the course so, we can say that course is not performing well in terms of attracting potential learners. We will

try to analyze what is going wrong based on various factors like, whats the reason of leaving the course and what kind of sentiments they have towards the course.

About the figure 1:

- x-axis: Different runs of the course.
- y-axis: Number of Learners Enrolled.

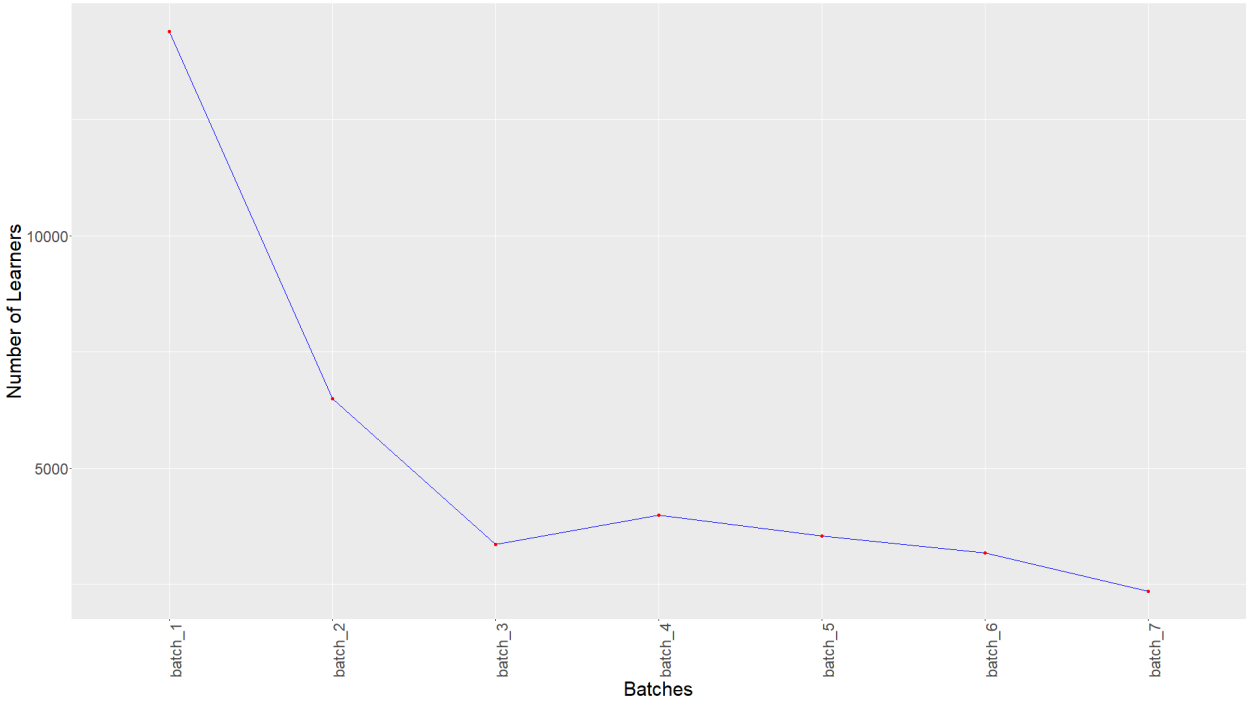

Figure 1: Number of Learners Enrolled over Different Iteration

5.1 Targeting right audience.

5.1.1 From which region most learners are enrolling?

From figure 2, we can observe most learners are coming from Europe accounting for around 59% of the total learners and from Asia and North America, 15% and 10% respectively. So, it will be a wise decision to tailor the course as per European countries, for example adding transcripts and captions in other languages like French, Spanish, Italian, and German etc. This will allow learners to understand the course much more comfortably and hence it will enhance their experience with course. We can also try to analyze from which country most learners are enrolling.

About the figure 2:

- x-axis: Latitude.
- y-axis: Longitude.
- Legend: Dark blue denotes less number of people, white denotes more number of people.

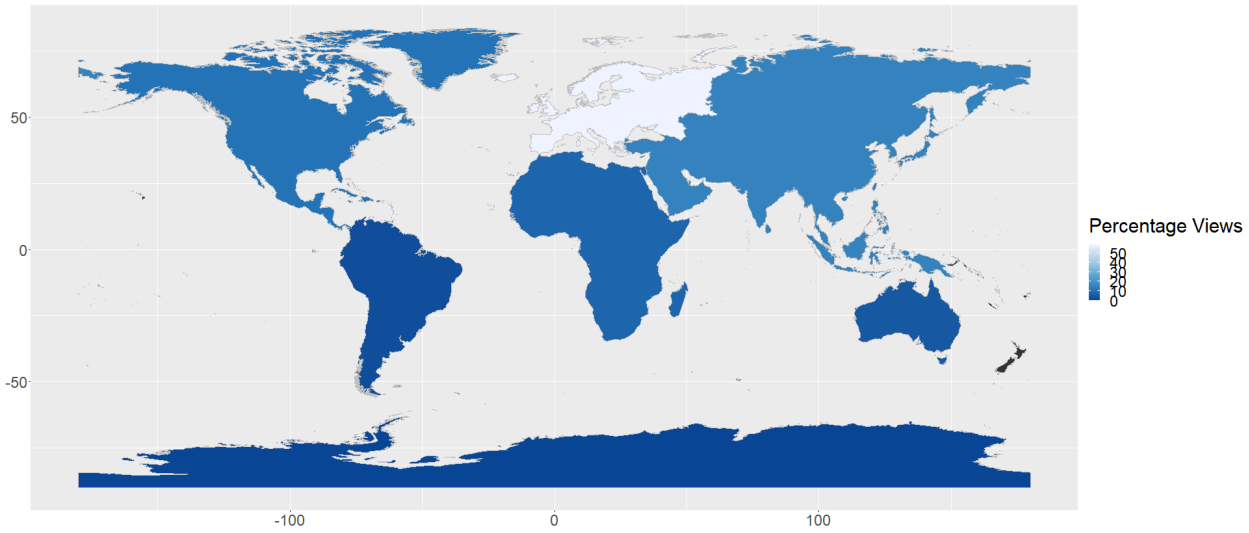

Figure 2: Continent from where most learners are enrolling

From figure 3, we can see that mostly learners are enrolled from the UK, India, and USA and there are considerable number of people from countries like Australia, Saudi Arabia, Nigeria, Mexico, and Russia. So, based on the number of people from various locations advertisement can be optimized. For example, rather than advertising the course or the platform with same intensity in all the regions, importance can be given to regions like the UK and India. This way cost of investment in advertising can be reduced and profits can be increased and that cost can also be used for reinvesting in the course or platform to improve it even further.

5.1.2 Which type of people are enrolling for the course?

From figure 4, we can observe following key information:

1. Most learners are male and females are close second. Nonbinary and others are very less.
2. Most learners are between age range 26 to 35 years. However, there are considerable number of learners in all age range except for less than 18 years.
3. People working full time and in IT and information services tend to take the course more often as compared to people from any other employment status and area.
4. Significant number of learners are holding university degree.

5.2 Course Improvement

From Figure 5, we can see that most of the times course is being accessed on desktop and hence, the learning platform and content can be optimized further for desktop and then focus can be shifted to other devices like mobile and tablets.

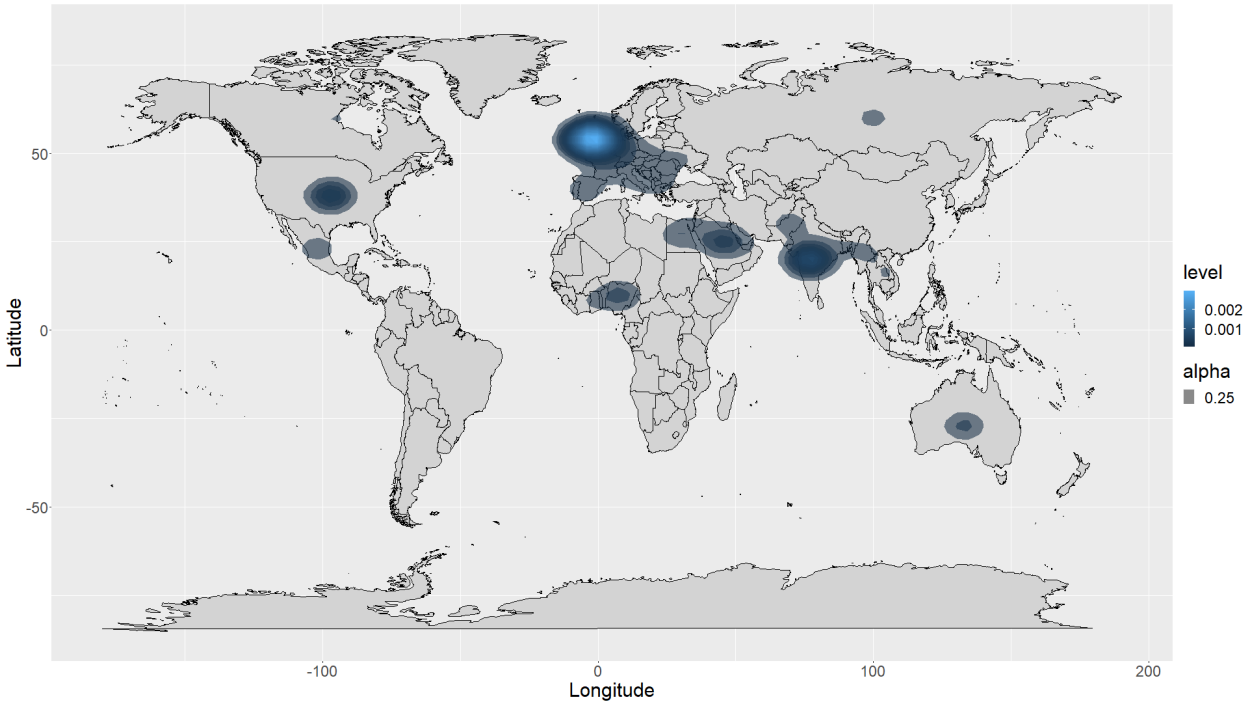

Figure 3: Countries from where most learners are enrolling

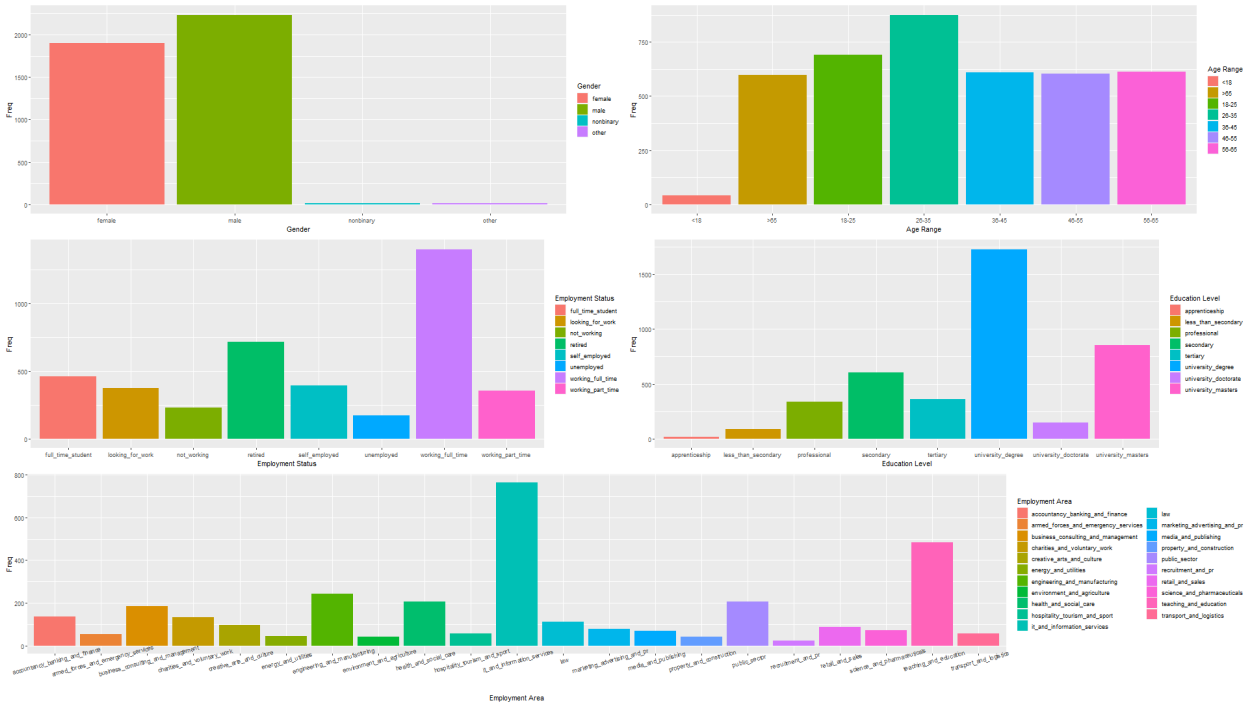

Figure 4: Which type of people are enrolling for the course?

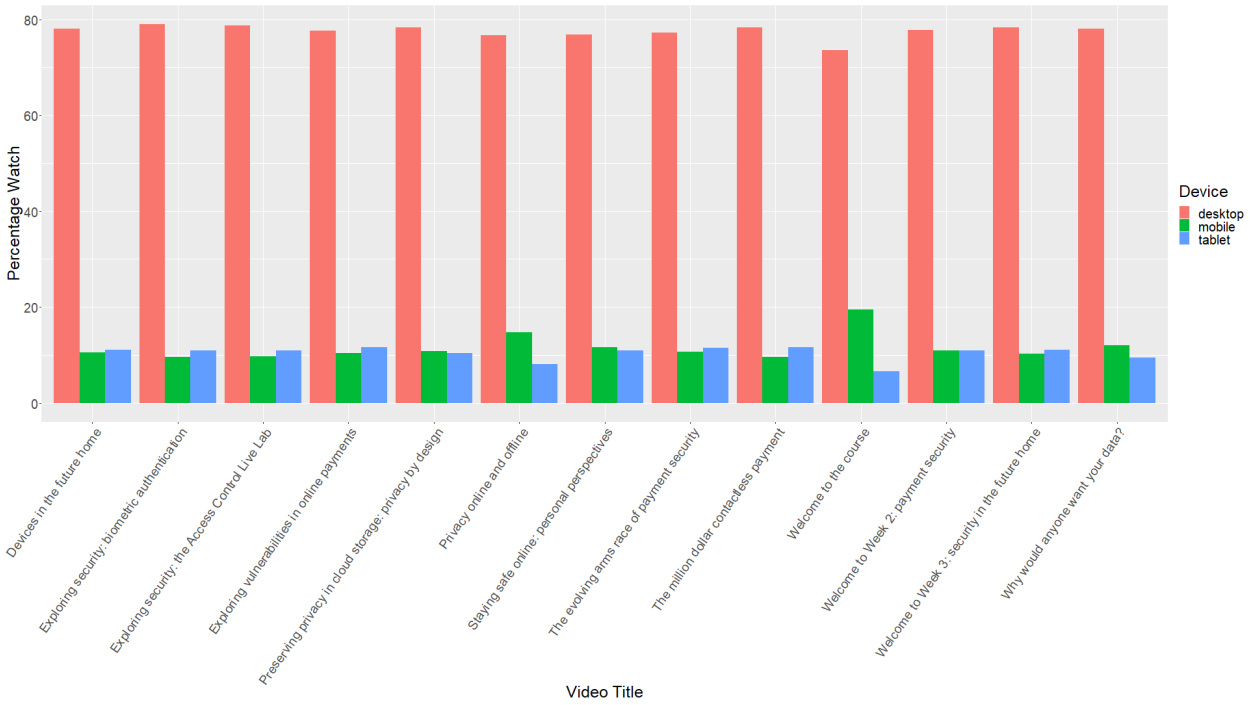

Figure 5: Most widely used device to access the content in each iteration

Figure 6 shows, although mobile phones are being used significantly less as compared to desktops, they are increasing with different iterations of the course. So, platform and course content optimization for mobile phone should also be given equally important.

Overall weekly sentiments of learners.

From figure 7, we can mostly words are positive and hence we can say that learners have had good experience with the course content. So, decrease in number of enrollments over the time is clearly not because of course content. Further analysis needs to be carried out to pin point the exact reason behind people leaving the course and decrement in number of learners over different runs of the course.

Reason for leaving the course.

From figure 8, we can clearly see people are leaving the course mainly because either they don't have sufficient time to complete the course or they expected the course to take less time to complete but that's not the case. So, if possible course can be either shortened or divided into multiple courses.

Optimal length of the videos.

Figure 9 suggests, when the video length is around 5 minutes (313 seconds) people tend to stay till the end of the video so, if possible keeping the length of most videos close to 5 minutes will be optimal because learners will be attentive through out the video and it is totally possible to cover good quality and quantity of content in this video length. However, it is also possible something very important is being covered in that specific video (**Exploring security: biometric authentication**) that has length 313 seconds and hence learners are watching that video till the end.

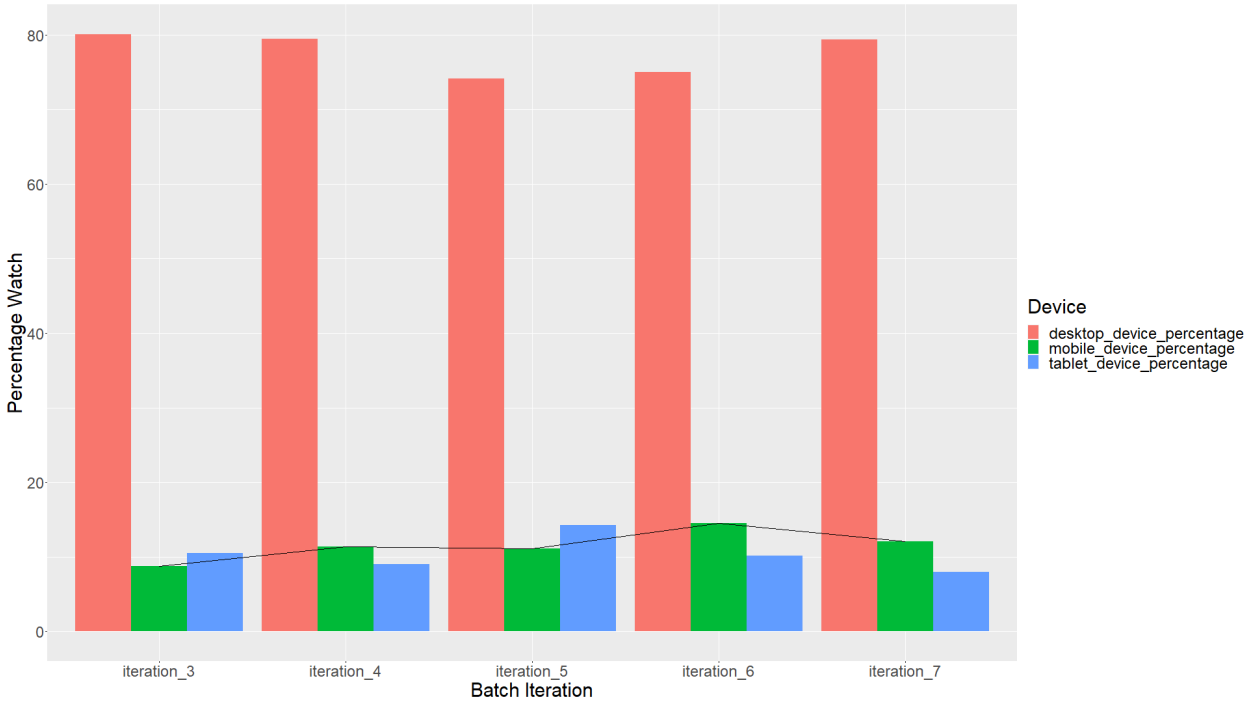

Figure 6: Most widely used device to access the content in each iteration

Figure 7: Overall weekly sentiments of learners

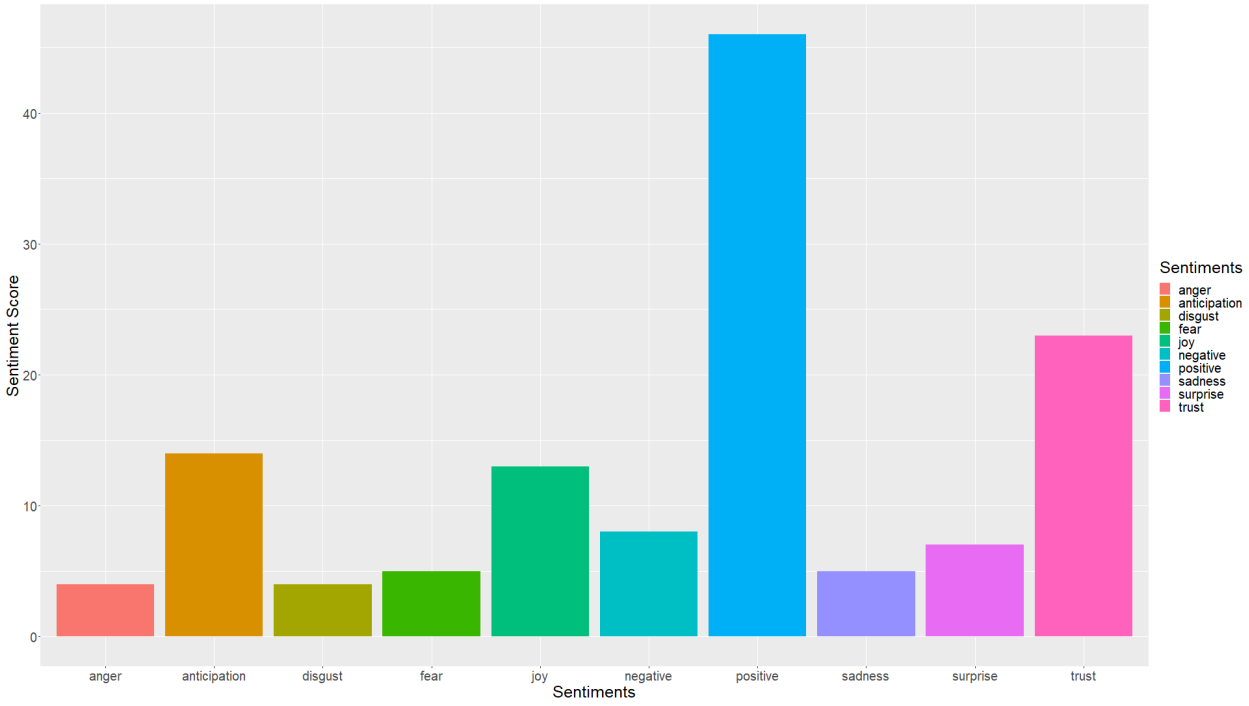

Figure 8: Overall weekly sentiments of learners

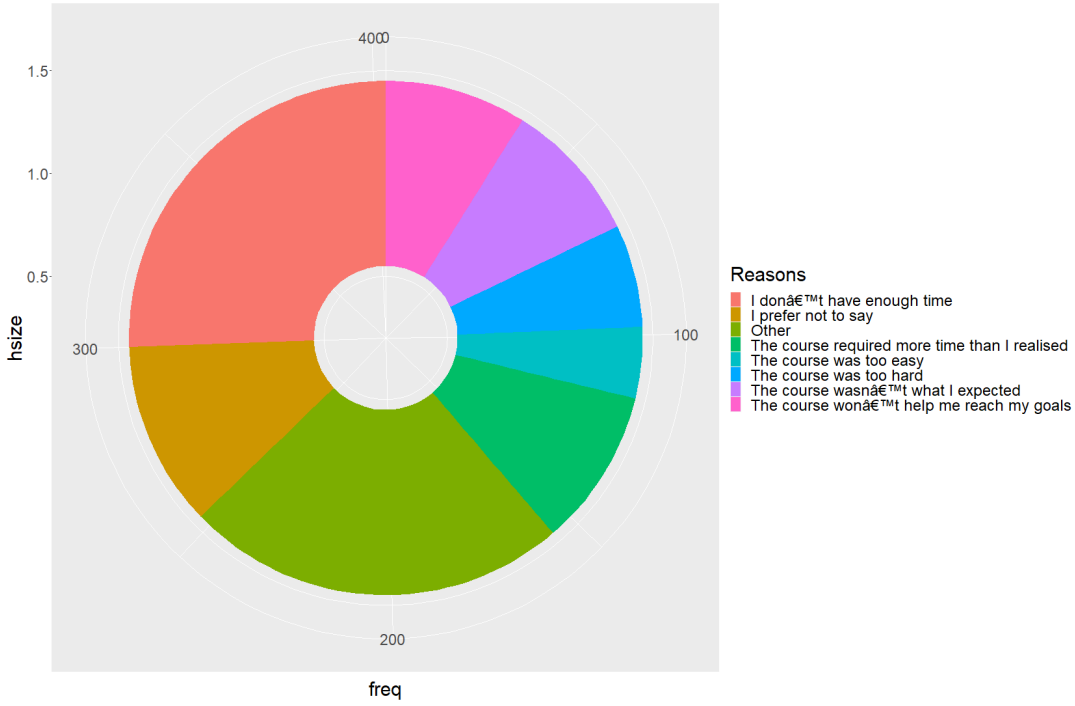

Figure 9: Reason for leaving the Course

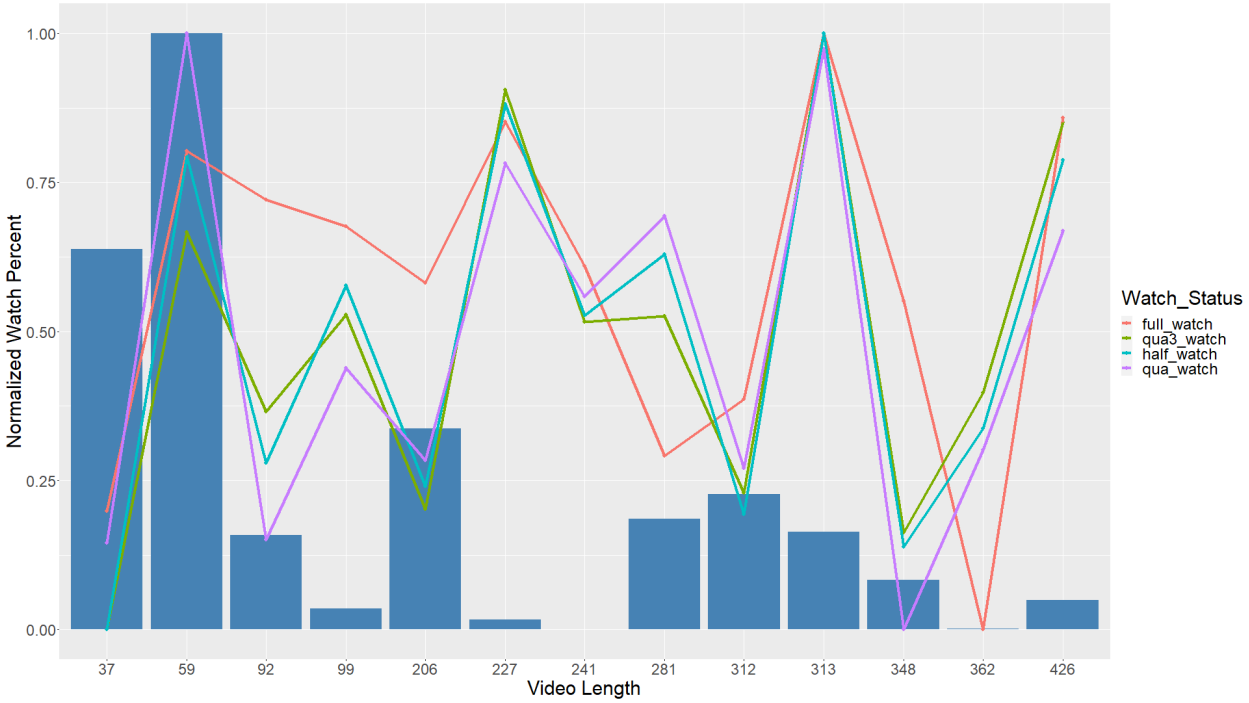

Figure 10: Optimal Video Length

- weekly sentiments are positive means course is good, people dont have time thats why leaving and hence course can be modified in such a way that quality can be maintained but can be covered in less time. . .