Cyber Security Data Analysis Report

Roshan Pandey (210113925)

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

3.1 Whole dataset of Cyber Security: Safety at Home, Online, in Life course.

Data is consist of 7 different iterations(runs) of the course.

- "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.

3.2 Subset of data taken into consideration.

- **Enrolment data** has numerous variables in this dataset but we are only considering following features of the learners:
  Gender has sex of the learners, in this we have male, female, nonbinary and others. There are majority of the fields as unknown.
  Age Range has several age intervals: less than 18, 19-25, 26-35, 36-45, 46-55, 56-65 and above 65, here also most fields are unknown.
  Highest Education Level tells about the highest education qualification that learners have achieved before joining the course. we have unknowns in this field too.
  Employment Area has the record of from which background learners are coming from. For example, health care, IT, education etc. Unknowns can be seen here as well.
  Employment Status depicts the current state of learners' working state. Majority of the fields are unknown.
  Detected Country tells about the location from where learners are enrolling.
- **Video Stats data** has details about the number of views each video has got, length of the video, devises used and many more, however, we have considered subset of features mentioned below:
  Title column has all the video titles.
  Video Duration has total length of the video in seconds.
  Viewed x Percent: There are multiple columns with percent of learners watched 10%, 25%, 50%, or 100% of a particular video length.
  x Device Percentage: There are different columns that has data of percent of learners watched the content on which device(mobile, desktop, console, tv, tablet).
  x View Percentage: There are several columns with percent of learners watched the video from which continent.
- **Weekly Sentiments data** has multiple fields but we are only interested in learners sentiment reasons for a particular week which is stored in *reasons* column.
- **Leaving Survey data** has data related to when an individual left the course, last completed step and many more, however, we are only considering the *leaving reason* column which will also allow us to get better understanding of why people left the course.

4. Data Processing

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

4.1 **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 and 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.

4.2 **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 batches or they still like to use desktop computers.

4.3 **Enrollment data with geographical coordinates**: Enrollment data from section 4.1 was joined with data downloaded from github(Link in reference section) based on alpha 2 code of the countries.

4.4 To analyse which device is being used the most, video stat data from section 4.1 was used and mean of different device columns was calculated.

4.5 **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 by removing the stop words, punctuation, converted every word to lower case, removal of special characters and worlds of length 3 or less, all these operations were done by helper function string_cleaner() present in ./lib/helpers.R file. After this cleaned text was split into words and then sentiment score was calculated for each word.

4.6 **IMPORTANT Assumptions:**

- 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 fields 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 from.

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 figure 1:

- x-axis: Different runs of the course.
- y-axis: Number of learners enrolled.
- Separate data from each iteration is taken into consideration.

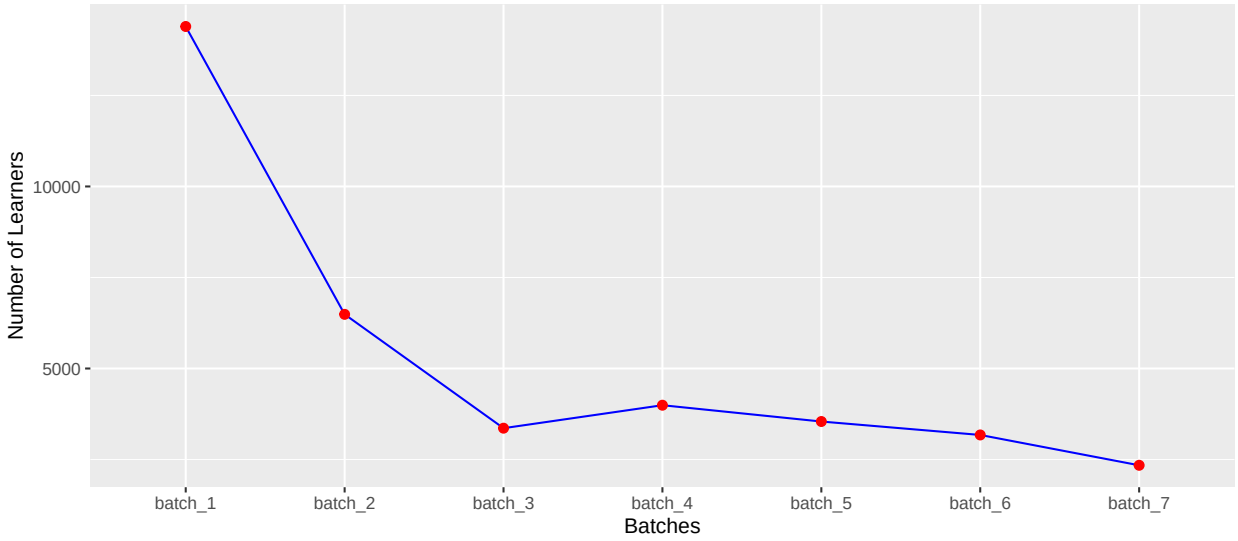

Figure 1: Number of learners enrolled over different iteration

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 figure 2:

- x-axis: Latitude.
- y-axis: Longitude.
- Legend: Dark blue denotes less number of people, white denotes more number of people.
- Merged data is taken into consideration.

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.

About figure 3:

- x-axis: Latitude.
- y-axis: Longitude.

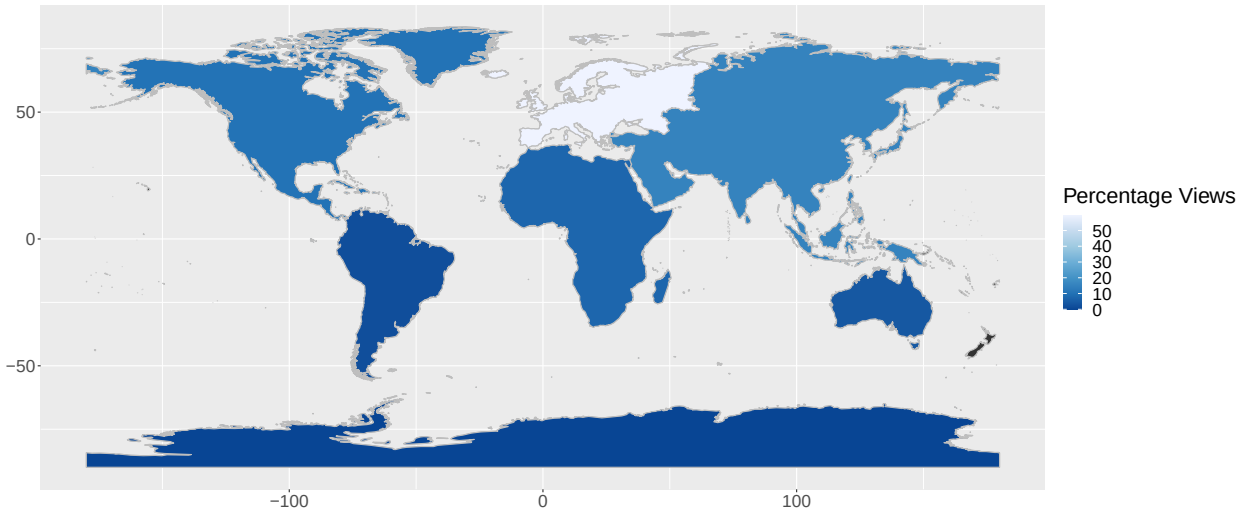

Figure 2: Continent from where most learners are enrolling

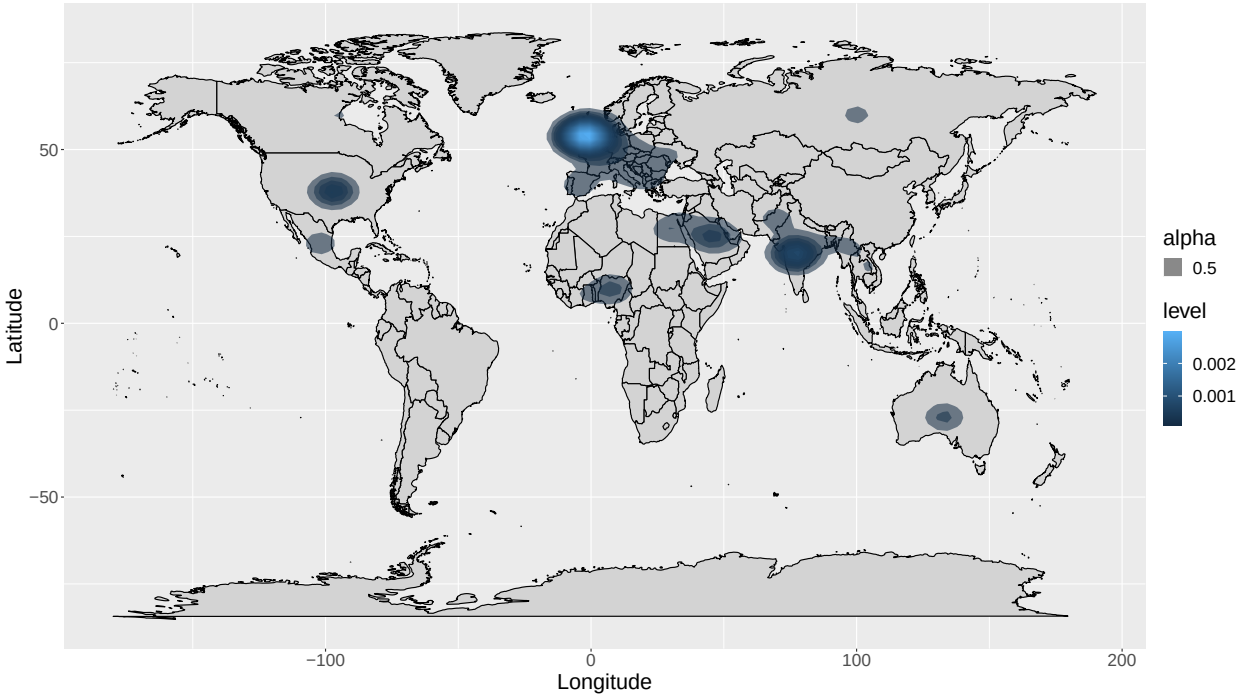

Figure 3: Countries from where most learners are enrolling

- Legend: *Level* tells about the intensity of a point on map. Higher level(Light blue) represents more number of people. **Alpha** is used for the hardness/softness of the marked area as we move away from the peak center region.
- Merged data is taken into consideration.

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. Overall, from the analysis that has been carried out so far, we can say that advertisement can be pushed to people from all age range but certainly focus can be from 26 to 35 years age range, people who are working full time are also seemed to be interested in the course and want to learn how their data is being used, university degree holders are also extremely interested in the course, in terms of employment area, people from information technology services and educational background have shown tremendous interest in the course. So people from all age range, having university degree, with full time job, or from IT and educational background can be attracted via suitable ads.

About figure 4:

- y-axis of all the plot is the count of individuals(learners).
- x-axis is different categorical variables(features) of the learners such as gender, age range, employment status, employment area, and highest education level.
- Merged data is taken into consideration.
- Please zoom in on this plot to see the labels, when font size of the text was increased, the plot was completely distorted and hard to understand.

5.2 Course Improvement

5.2.1 Which device is widely used?

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. In recent times with exponential advancements in mobile technology most people might want to access the content on their smartphones as it allows learners to access the content on the go with much ease. So we will try to analyse whether there is an upward trend in usage of mobile phone or not.

About figure 5:

- y-axis: Percent of content watched on a particular device. For example, 78% of the learners watched the video on desktop, 12% on mobile phone, and 10% on tablets.
- x-axis: Title of different videos.
- Legends: Different devices.
- Merged data is taken into consideration.

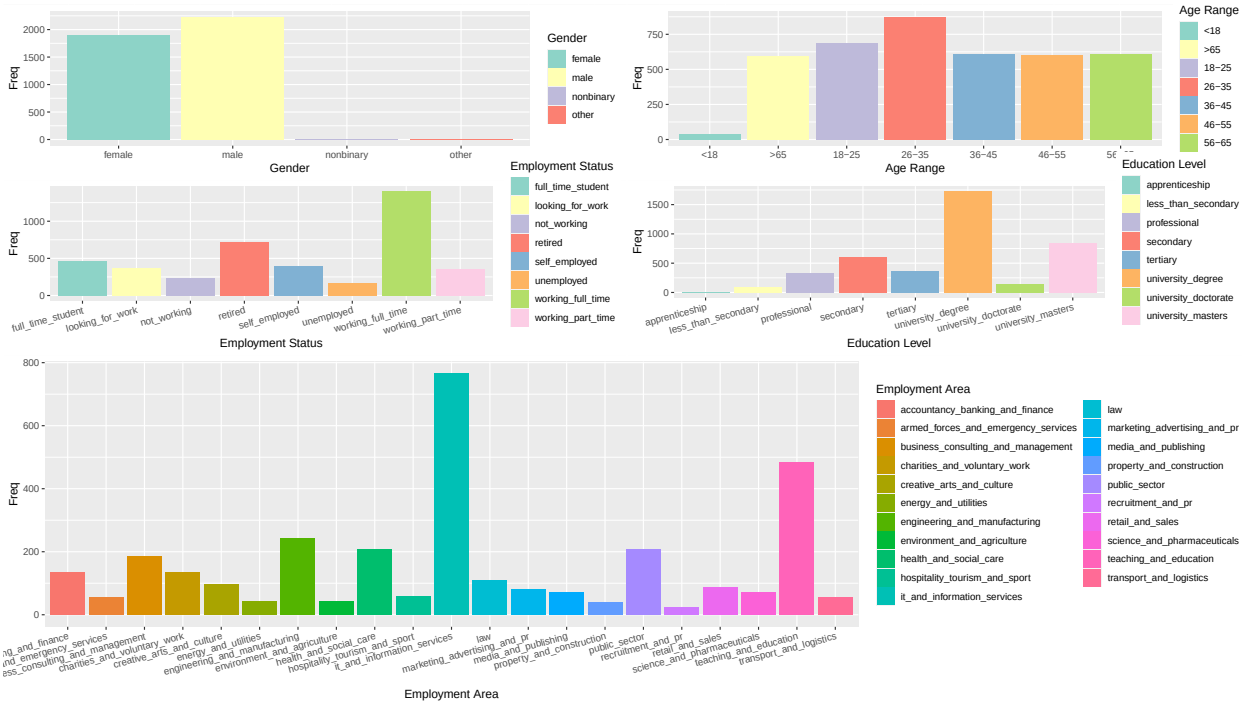

Figure 4: Which type of people are enrolling for the course?

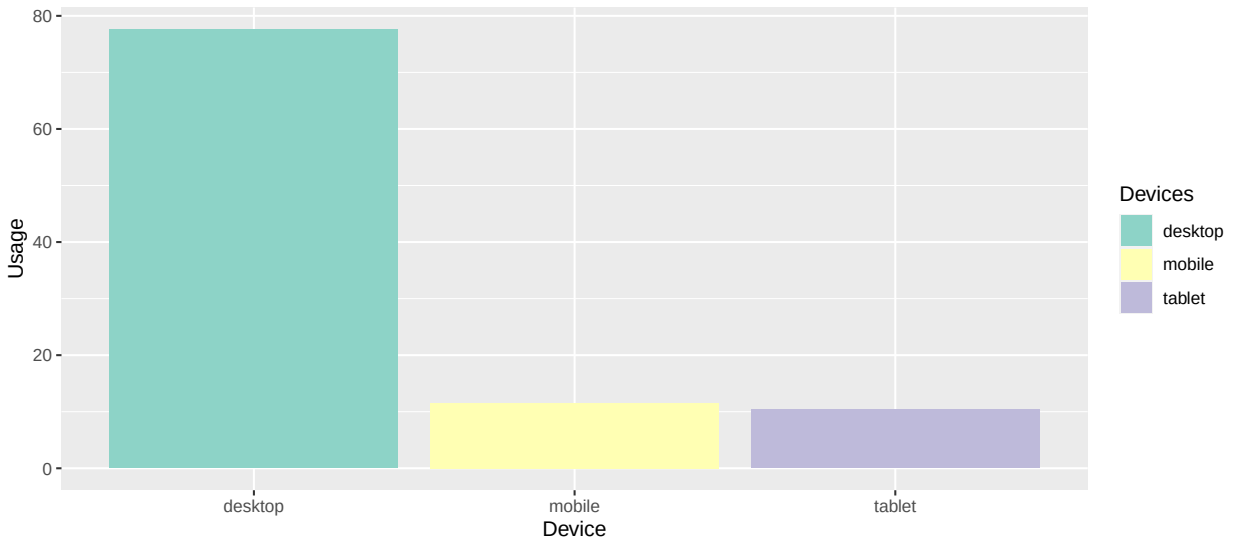

Figure 5: Most widely used device to access the content

NOTE: Consoles and TVs are not taken into consideration as they were very less or zero percent in use.

Figure 6 shows, although mobile phones are being used significantly less as compared to desktops, there is a subtle increase in the popularity of the mobile phones as the course is being run over different iterations. So, platform and course content optimization for mobile phone should also be taken into consideration as it will enhance the user experience of the learners. Moreover, from section 5.1.2 we found out there are significant number of learners who are senior citizens and in several studies(reference 4) it has been found more and more older people are starting to use smartphone as it is easy to carry. Hence, content viewing and readability on phone should also be given utmost importance.

About figure 6:

- y-axis: Percent of content watched on a particular device. For example, 80% of the learners watched the video content on desktop, 9% on mobile phone, and 11% on tablets in the first iteration of the course.
- x-axis: Different iterations of the Course.
- Legends: Different devices.
- Separate data is taken into consideration.
- Relevant data is not available for 1st two iterations thats why analysis is being done from 3rd to 7th iteration.

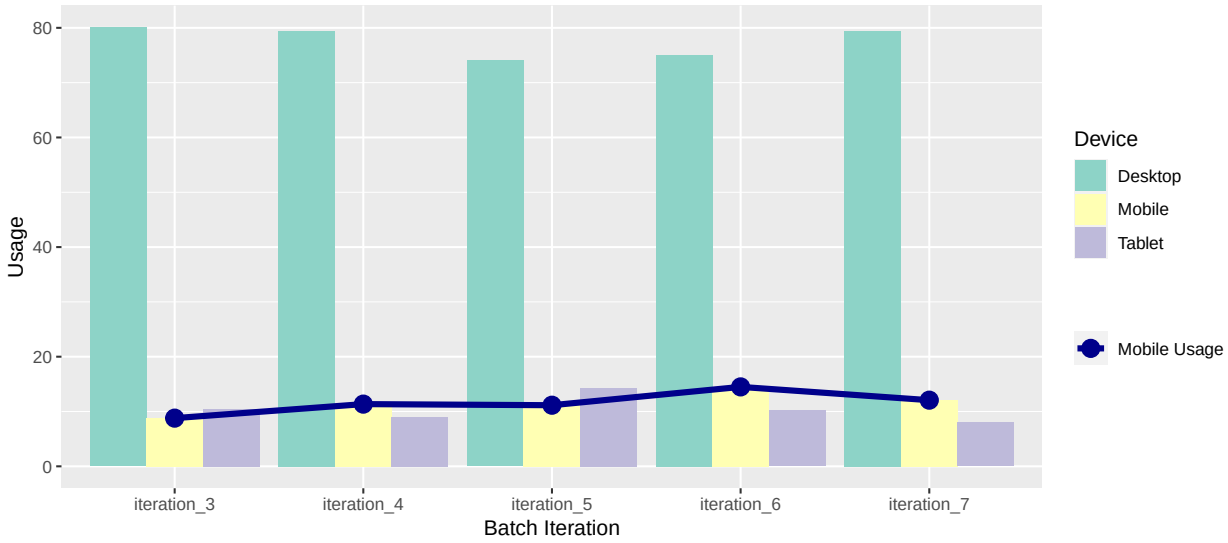

Figure 6: Most widely used device to access the content in each iteration

5.2.2 Overall weekly sentiments of learners.

From figure 7, we can see mostly words are positive and hence we can say that learners have had good experience with the course content. They have found the course to be useful, interesting, informative, practical, etc. 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.

About figure 7: All the weekly sentiment reasons from different batches has been merged and text processing mentioned in section 4 of the report was carried out for cleaner and better insights.

Figure 7: Overall sentiments of learners

Figure 8 represents the sentiment score of different sentiments like positive, negative, anticipation, joy, trust, etc. Syuzhet package was used for generating sentiment scores, which has sentiment dictionaries and offers a method for accessing the sentiment extraction tool developed in the NLP group at Stanford. Refer reference section for more details about the package and its usability. From figure 8 we can support our analysis in the previous step (Fig 7) and state that overall sentiment of learners was positive and need to spot the exact reason for leaving the course.

About figure 7:

- x-axis: Sentiments
- y-axis: Sentiments scores
- All the weekly sentiment reasons from different batches has been merged and text processing mentioned in section 4 of the report was carried out for cleaner and better insights.

5.2.3 Reason(s) for leaving the course.

From figure 9, 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 this way learners can choose which part of the course they want to take and they will be able to finish the course in less time which will lead them to buy the next part of the course.

About figure 9:

- Numbers on the donut chart represent the number of learners given a particular reason. For example, 103 learners gave reason as "I don't have enough time".
- Legend: Reasons for leaving the course.

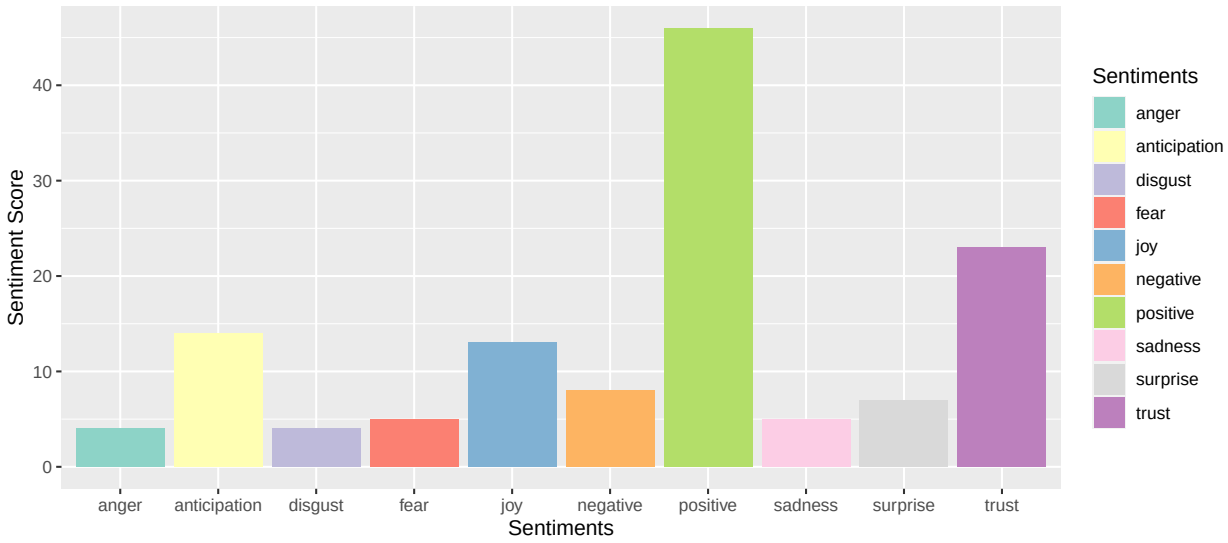

Figure 8: Overall sentiment scores of learners

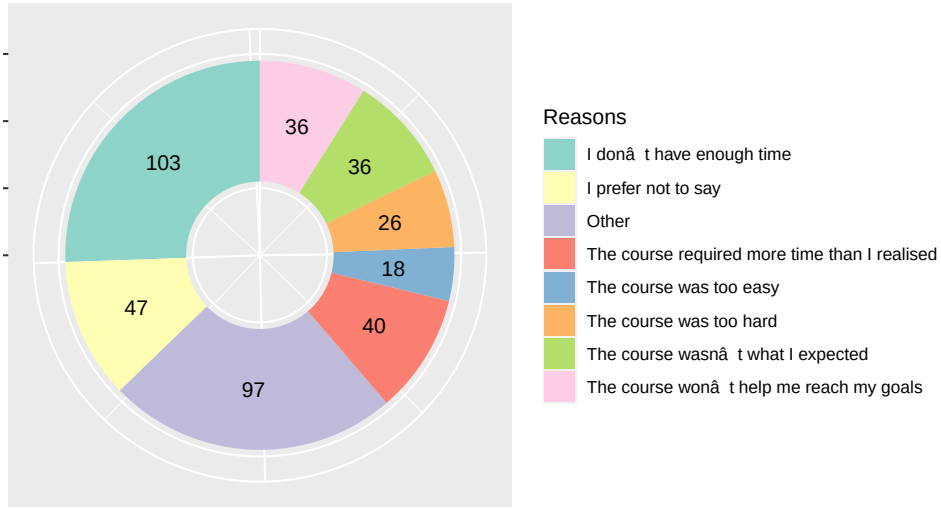

Figure 9: Reason for leaving the course

5.2.4 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 the videos close to 5 minutes will be optimal because learners will be attentive through out the video and it is possible to cover good quality and quantity of content in this video length. Some researches and surveys also back this(best video length: 5-6 mins) finding (Section 8 reference number 2 for more details)
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. This statement can also be supported by observing the video of length 426 seconds. Although there are relatively less people watching the video till the end, they are comparable with number of people watching video of length 313 and hence we can say people tend to stick to the video till the end when something important is being covered in the video. So, it is hard to comment on optimal length of the video but we can make an estimation of around 5 to 6 mins or 300 to 360 seconds.

About figure 10:

- x-axis: Video Length
- y-axis: Normalized Watch percentage and view count for each video length.
- Legend: full_watch represents video watched 100%, qua3_watch represents video watched 75% , half_watch represents video watched 50%, qua_watch represents video watched 25%

Data has been column normalized because total number of views were very high and can go up to very large number where as percentage of people who watched whole video or half video can only go up to 100 so, it is really hard to plot them on one plot, hence, total views and different watch percentages were normalized.

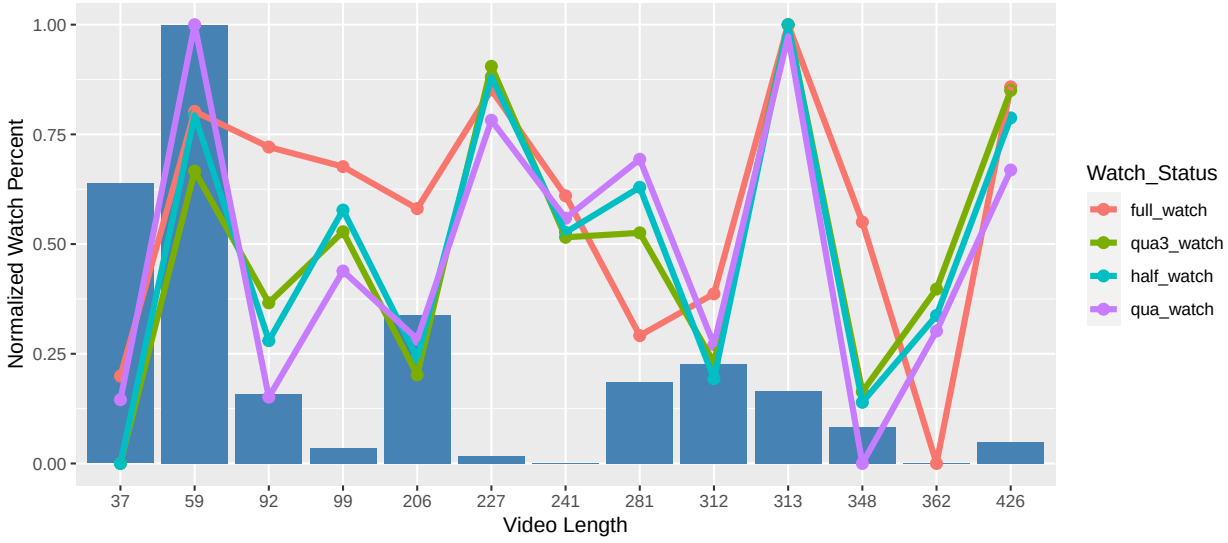

Figure 10: Optimal Video Length

6 Modeling

We have used pre-trained NLP model from Syuzhet package which was developed at Stanford as mentioned in section 5.2.2 figure 8 description. We used get_sentiment function which accepts two parameters, a character vector and a method. The selected method determines which of the available sentiment extraction

methods will be used. The methods are bing, afinn and nrc.

nrc: This method categorizes the words into binary mode (yes, no) for positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. For example, if a word is positive, 1 is assigned to positive and 0 to rest, if a word is of trust sentiment, 1 is assigned to trust and 0 to others.
bing: This method categorizes words in positive and negative words.
AFFINN: This method assigns score ranging from -5 to 5 to each word. Negative score means negative sentiment and positive score means positive sentiments.

6.2 Method selection

nrc method was selected because it not only gives positive and negative sentiment scores but also few other sentiments like anger, joy, trust, etc. This allowed us to analyze learners' feelings towards the course in greater detail and we observed most people had positive sentiments towards the course but also they trusted the course and enjoyed the learning. Which clearly suggests course content is really good.

7 Evaluation of the Analysis

All the analysis is carried out by taking assumptions into the consideration discussed in section 4. This may result in uncertainty however, we performed various analysis for same objective on different data so that analysis can be supported.

- To figure out from which part of the world most learners are enrolling, we considered video stats data to find which continent has more learners then we analysed detected_country data from enrollment dataset and they both supported each other. Most learners are coming from Europe and they are residing in the UK.
- To analyze what kind of people enroll for the course more often, we checked people from each iteration and then from overall data(merged data, for detail refer section 4) final plot was constructed. Plot from different iteration was not included in the report to avoid redundancy in the graphical representation of the information.
- To make estimations on which device is being used the most, data was available only from 3rd to 7th iteration so, analysis was done over 5 out of 7 runs of the course.
- To observe the sentiments of the learners we constructed the word cloud then eye balled which type of words are used most often and we stated positive sentiment. To support this finding we also calculated sentiment scores(for detail, refer section 6) and it also suggested positive sentiments.
- To pin point the reason we analysed the leaving survey, here considerable number of fields were null so findings might be biased. Also, significant number of people gave reason as "Other" so, we don't know what that can be.

8 Conclusion

After various analysis that has been performed in this report we can say that course content is good and that is not the reason why learners are leaving the course but it is talking more time for them to complete the course. There are several ways this can be tackled such as, dividing the whole course into multiple courses or assessment questions can be reduced or made optional so that serious learners can take the assessments as they like and casual learners can go through the content, learn from the course without having to spend much time in assessments.
From the analysis for targeting right audience we can say that people from all age range, from IT and Education sector, and working full time are interested so, ads can be optimized by keeping these criteria into consideration.

9 References

[1] syuzhet package for NLP: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
[2] Video length research: https://www.techsmith.com/blog/video-length/
[3] Sentiment Extraction Methods: https://www.tidytextmining.com/sentiment.html
[4] Old people using smartphones: https://www.pewresearch.org/internet/2017/05/17/tech-adoption-climbs-among-older-adults/pi_2017-05-17_older-americans-tech_1-01/
[5] Alpha 2 code to Coordinates: https://gist.github.com/tadast/8827699