# MOOC Cyber Security Course Analysis Report

Jashwant Anandan - 220120148

2022-11-18

# 1. Introduction:

This is a basic data analysis report for the marketing and sales team of FutureLearn Online Learning platform, a company that partners with numerous universities to offer online courses. Depending on specific regions, relevant demographics, types of devices used, and analysis of each stage of the course, the company's marketing team will be able to make decisions for a variety of course recommendations and advertising campaigns. These decisions will be based on how well the courses are working. This report will include two analysis cycles using the *CRISP-DM* reporting framework, the respective analysis are:

- Primary Analysis - Country and Demographic analysis for the MOOC dataset.
- Secondary Analysis - Analysis on Devices used by Users and Views on Each step of the Course.

# 2. Primary Analysis - Country and Demographic Analysis:

## 2.1 Business Understanding and Assessing the Situation:

### 2.1.1 Stakeholder Requirement:

The knowledge around the online course *Cyber Security: Safety at Home, Online, and in Life* requires some fundamental study, according to the marketing and sales teams of the FutureLearn website. The following analysis will be helpful in providing the relevant teams with the knowledge they need to make knowledgeable decisions about how to use their advertising and marketing expenditures. This research is based on the nations that people have enrolled in as well as their individual demographic information based on employment and education.

### 2.1.2 Questions Answered in the Analysis:

- Which Locations (Countries) has the most enrollments for the particular course ?
- What are the educational and occupational characteristics of people from the top-ranking countries for course enrollments ?
- What recommendations can be drawn from the analysis's findings ?

### 2.1.3 Available Resources:

The resources available for the following data analysis and report are as follows:

- Data : Cybersecurity enrollments dataset in multiple `.csv` files from MOOC collection of data.

- Computing Resources : Laptop or Personnel Computer to carry out basic analysis and report generation.

- Software :Microsot OS, R and R Studio.

### 2.1.4 Requirements, assumptions and constraints:

- Finishing the analysis and delivering the report to the stakeholders by no later than November 18th, 2022.
- The results of the study could be hampered by any missing data or data tampering that occurred during the data recording.
- There might not be enough accurate data to produce a persuasive conclusion.

### 2.1.5 Risks and Contingencies:

- Poor data quality.
- Improper understanding of analysis by stakeholders.
- Improper data understanding and analysis performed by data analyst.

### 2.1.6 Terminology:

CRISP-DM : Cross-Industry Standard Process for Data Mining

MOOC : Massive Open Online Course

### 2.1.7 Analysis and Data Mining Goals:

- Clean up the data collection by removing any empty(`NA`) values.
- Only include columns for countries, education, and employment that are relevant to the analysis.
- Create useful graphs that will more comfortably illustrate the conclusions.

### 2.1.8 Project Plan:

- Look through the MOOC dataset and just collect the necessary information.
- To complete all of the following, use the `R` programming language and *RStudio*:
  - Before cleaning, understand and examine the available data.
  - Clean and retain only the data that is necessary for the planned analysis.
  - Explain the created plots and answer the stakeholder's question
- On the basis of the analysis and plots produced, offer business suggestions.
- Create a report using `RMarkdown` that details every step.
- Make a presentation that explains the report's outline.

## 2.2 Data Understanding:

### 2.2.1 Initial Data Collection:

The following *.csv* files were collected for the Analysis because it is based on enrollment data : `cyber.security.3_enrolments.csv`, `cyber.security.4_enrolments.csv`,`cyber.security.5_enrolments.csv`, `cyber.security.6_enrolments.csv`,`cyber.security.7_enrolments.csv`.

- The FutureLearn website's MOOC datasets section provided the relevant data.
- There were no problems with data collection.
- The Project Template's *data* folder contains the information gathered.

### 2.2.2 Data Description:

Data from various *.csv* files have been appended together into a single dataframe called `enrolments` for basic quality checking.

```
# Appending different .csv files data into single Dataframe
# Code executed in 01 - A.R file in munge folder of Project Template
enrolments <- rbind(cyber.security.3_enrolments, cyber.security.4_enrolments,
                    cyber.security.5_enrolments, cyber.security.6_enrolments,
                    cyber.security.7_enrolments)
```

The dataframe has the following column names otherwise called as *variables*:

```
# Get column names
colnames(enrolments)
```

```
##  [1] "learner_id"             "enrolled_at"
##  [3] "unenrolled_at"          "role"
##  [5] "fully_participated_at"  "purchased_statement_at"
##  [7] "gender"                 "country"
##  [9] "age_range"              "highest_education_level"
## [11] "employment_status"      "employment_area"
## [13] "detected_country"
```

```
# Get number of rows and columns
dim(enrolments)
```

```
## [1] 16414    13
```

The code chunks above indicate the column names and the uncleaned dataframe's dimensions (*rows=16414, columns=13*); this type of dataset is known as a *tall* dataset.

### 2.2.3 Data Exploration and Data Quality Report:

```
# Display first 7 rows
head(enrolments, 7)
```

```
## # A tibble: 7 x 13
##   learner~1 enrol~2 unenr~3 role  fully~4 purch~5 gender country age_r~6 highe~7
##   <chr>     <chr>   <chr>   <chr> <chr>   <chr>   <chr>  <chr>   <chr>   <chr>
## 1 1f5166ea~ 2017-0~ 2018-1~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 2 1154e656~ 2017-1~ 2018-1~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 3 455709fb~ 2017-0~ 2018-1~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 4 2f8932ed~ 2017-1~ 2018-1~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 5 7f57e1b0~ 2017-0~ 2018-1~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 6 14934968~ 2017-1~ 2018-1~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 7 0a303d0a~ 2017-1~ 2018-0~ lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## # ... with 3 more variables: employment_status <chr>, employment_area <chr>,
## #   detected_country <chr>, and abbreviated variable names 1: learner_id,
## #   2: enrolled_at, 3: unenrolled_at, 4: fully_participated_at,
## #   5: purchased_statement_at, 6: age_range, 7: highest_education_level
```

The displayed dataframe appears to have several *“Unknown”* and *missing* data values at first glance. These unknown cells could interfere with the analysis process, but the report is nevertheless produced with just the “Unknown” data to estimate their amount and to encourage stakeholders to collect data correctly. Additionally, a lot of essential information will be lost by removing all the rows with Unknown values because the said data is in distinct rows for each demographic. Moreover, some columns might not be necessary for the analysis, and it’s crucial to remove any missing values. These tasks will be completed in the process’s next significant step, known as **Data Preparation**.

## 2.3 Data Preparation:

### 2.3.1 Data Selection:

Many columns or variables can be left out of the main data frame because the analysis’s primary goal is to learn more about the location and particular demographics of the enrolled sample individuals. From the output in Section 3.2 that has the column names the columns that can be removed are as follows:

- `learner_id` -> Individual enrollment information is not taken into account in the analysis.
- `unenrolled_at` -> This is not necessary because the report concentrates on enrollment data rather than the opposite.
- `fully_participated_at` and `purchased_statement_at` -> Most of the useful values in these columns are missing, making them useless for the analysis.
- `country` -> Due to the large number of missing values in this column as well, the emphasis is on the column `detected country` rather than this one.
- `role`, `gender`and `age_range` -> This report does not include any analysis of these demographics.

```
# Remove columns and create new data subset
# Code executed in 01 - A.R file in munge folder of Project Template
enrolments_main = subset(enrolments, select = -c(learner_id, unenrolled_at, fully_participated_at, purch
```

```
# Get column names
colnames(enrolments_main)
```

```
## [1] "enrolled_at"             "highest_education_level"
## [3] "employment_status"       "employment_area"
## [5] "detected_country"
```

Therefore, the new dataframe `enrolments_main` contains all the necessary columns to conduct analysis.

```
# Get the number of rows and column
dim(enrolments_main)
```

```
## [1] 16414     5
```

There are *16414 rows* and *5 columns* in the example dataframe, which needs to be further cleaned to get a flawless dataset for analysis.

### 2.3.2 Data Cleaning:

```
# Count the number of "--" values
table(enrolments$detected_country == '--')
```

```
##
## FALSE  TRUE
## 16031   375
```

The total number of `--` entries in the `detected_country` column is 375. This entry can be changed to *“Not Available”* for clearer understanding.

```
# Replace "--" with "Not Available"
# Code executed in 01 - A.R file in munge folder of Project Template
enrolments_main <- enrolments_main %>%
  mutate(detected_country = replace(detected_country,
                                    detected_country == "--", "NotAvailable"))
```

```
# Display number of "--" entries
table(enrolments_main$detected_country == '--')
```

```
##
## FALSE
## 16406
```

```
# Display number of "Not Available" entries
table(enrolments_main$detected_country == 'Not Available')
```

```
##
## FALSE  TRUE
## 16031   375
```

The values for the `detected_country` column have been successfully updated.

The `enrolled_at` column needs to be changed from *Character* to *Date* datatype, since this will make it easier to retrieve the month and year from the column whenever needed.

```
# Code executed in 01 - A.R file in munge folder of Project Template
enrolments_main$enrolled_at <- as.Date(enrolments_main$enrolled_at)
```

```
# Display Datatype of all columns
sapply(enrolments_main, class)
```

```
##             enrolled_at highest_education_level       employment_status
##                  "Date"             "character"             "character"
##         employment_area        detected_country
##             "character"             "character"
```

```
# Display first few months from date type column
head(month(enrolments_main$enrolled_at))
```

```
## [1]  9 10  9 10  8 10
```

To make dealing with the dataframe easier, multiple *.csv* files were already integrated during the Data Understanding Process (Section 2).

```
# Display first ew rows
head(enrolments_main)
```

```
## # A tibble: 6 x 5
##   enrolled_at highest_education_level employment_status employment_area detect~1
##   <date>      <chr>                   <chr>             <chr>           <chr>
## 1 2017-09-14  Unknown                 Unknown           Unknown         GB
## 2 2017-10-14  Unknown                 Unknown           Unknown         GH
## 3 2017-09-24  Unknown                 Unknown           Unknown         GB
## 4 2017-10-19  Unknown                 Unknown           Unknown         GB
## 5 2017-08-08  Unknown                 Unknown           Unknown         GB
## 6 2017-10-19  Unknown                 Unknown           Unknown         GB
## # ... with abbreviated variable name 1: detected_country
```

The `enrolments_main` dataframe has now been cleaned and is prepared for in-depth analysis.

### 2.4 Data Analysis:

The marketing team could advance the course even further by focusing and targeting the advertising and other related campaigns to the respective countries and groups with the aid of a general analysis of the locations where the enrolled individuals were based and their respective educational and employment details.

```
# Group By and Summarise other columns
count_of_Countries <- enrolments_main %>%
    group_by(detected_country) %>%
    summarise(count = n()) %>%
    arrange(desc(count), .by_group = TRUE)
```

```
# Count Number of "Not Available" entries
(count_of_Countries[count_of_Countries$detected_country == "Not Available", "count"])
```

```
## # A tibble: 2 x 1
##   count
##   <int>
## 1   375
## 2    NA
```

It is also evident that *Not Available* has 375 entries, this is an issue because a considerable amount of individuals cannot be analysed for their location. It is important to note that in the future data collection proper detection of countries is required. Now for the further analysis the *Not Available* entry is removed.

| detected_country | count |
|---|---:|
| GB | 4877 |
| IN | 1310 |
| SA | 1020 |
| US | 869 |
| NG | 473 |
| EG | 467 |
| MX | 439 |
| AU | 434 |
| ES | 251 |
| PK | 250 |

| detected_country | count |
|---|---:|
| NE | 1 |
| NR | 1 |
| PA | 1 |
| RE | 1 |
| SR | 1 |
| SS | 1 |
| SX | 1 |
| VG | 1 |
| XK | 1 |

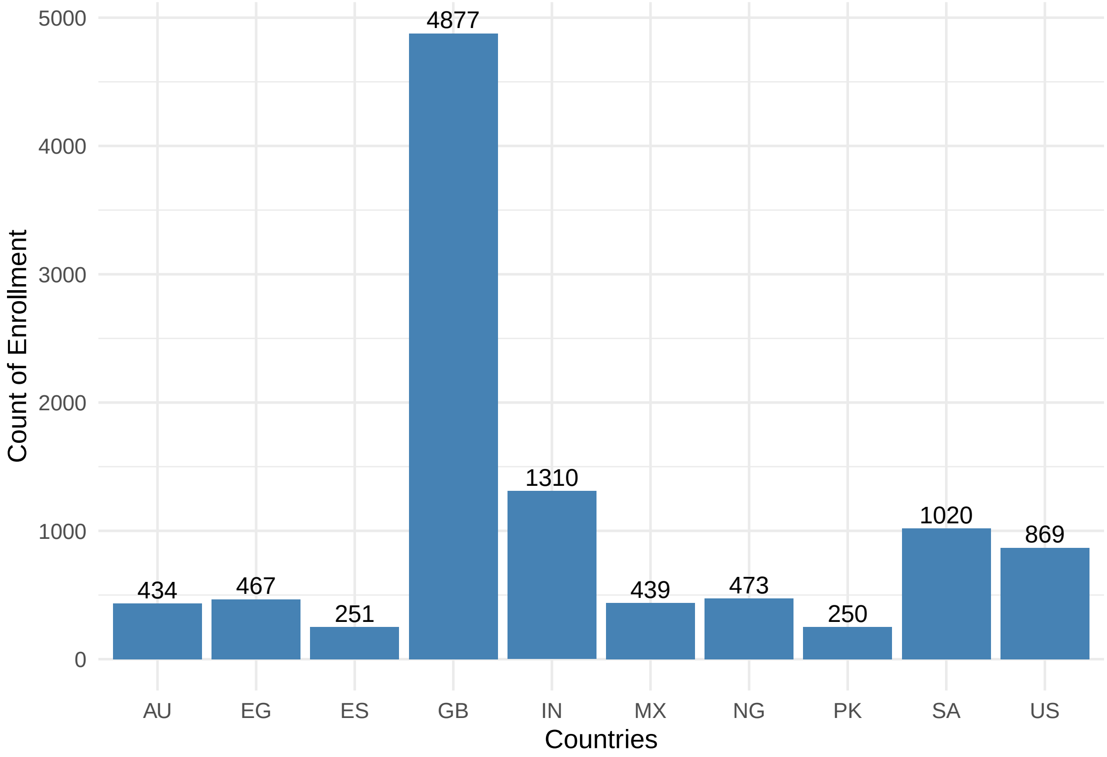
```r
# Remove the "Not Available Entry"
count_of_Countries <- filter(count_of_Countries, detected_country != "Not Available")
```

Above are the Tables with Top 10 and Bottom 10 Enrolled Countries for the Cybersecurity course in the left and right respectively

```
## Selecting by count
## Selecting by count
```

From the above tables it is evident that *GB* is the most enrolled country with a total count of 4877, followed by *IN* 1310, the difference between GB and IN is very huge compared to the other top 10 countries. This particular Cyber security course is very popular in GB. Following IN the third highest is *SA* with a count of 1020. On the right the table shows some of the country which have least enrollment rate of only one individual. The Below bar graph shows the visual representation of the Top 10 highest enrolled countries.

```r
# Plot a Bar Chart
p<-ggplot(data=t1, aes(x=detected_country, y=count)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=count), vjust=-0.3, size=3.5)+
  theme_minimal()+
  labs(x="Countries", y = "Count of Enrollment")
p
```



`Fig 1: Bar graph displaying the Count for Top 10 Countries.`

Now that the top 10 countries are figured out, a further analysis on the educational and employment details for the enrolled individuals in the respective countries can be carried out.

```
# Create new dataframe
enrolments_top_ten <- filter(enrolments_main, detected_country %in% (t1$detected_country))
```

A new dataframe called `enrolments_top_ten` with all the required demographics of the top 10 countries is created to create plots for the same.

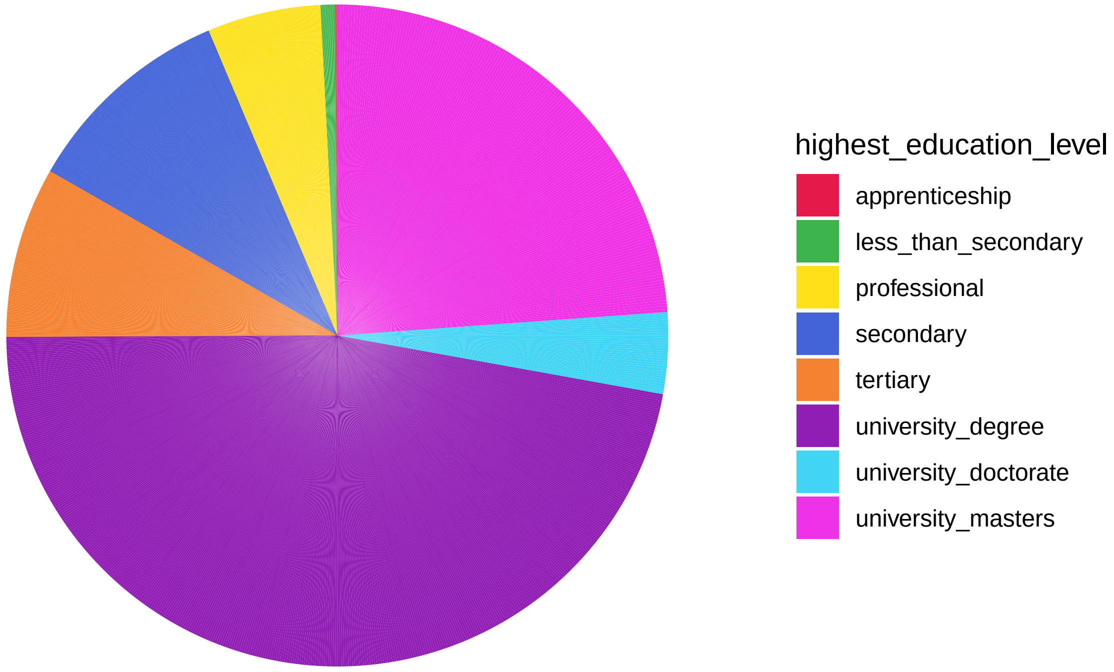
There are a lot of *Unknown* values for all the three demographics considered for the analysis. But all the unknown will not be removed at once because the number of these values change for different rows, so removing them whole will cause a lot of data loss, instead it will be removed seperately during each individual demographic analysis.

```
# Create a Blank theme for plots
blank_theme <- theme_minimal()+
  theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank(),
  plot.title=element_text(size=14, face="bold")
  )
```

```
# Display Count of each entry in the Column
(table(enrolments_top_ten$highest_education_level))
```

```
##
##      apprenticeship  less_than_secondary        professional
##                   6                   20                 104
##           secondary             tertiary   university_degree
##                 146                   94                 442
## university_doctorate  university_masters             Unknown
##                  32                  168                9378
```

```
# Create new dataframe without "Unknown" entry
enrolments_ten_edu <- enrolments_top_ten %>% filter(highest_education_level != "Unknown")
# Plot a Pie Chart
bp_edu<- ggplot(enrolments_ten_edu, aes(x="", y=highest_education_level, fill=highest_education_level))
geom_bar(width = 1, stat = "identity")+coord_polar("y", start=0) + blank_theme +
  scale_fill_manual(values=c("#e6194B", "#3cb44b", "#ffe119", "#4363d8", "#f58231", "#911eb4", "#42d4f4
  theme(axis.text.x=element_blank())
bp_edu
```



```
Fig 2: Pie Chart showing the ratio of the types of Education Levels by the
individuals enrolled in the course.
```

The bulk of the enrollees are clearly in their *university degree years* when one first looks at the pie chart in Fig. 3, which shows the data on the highest educational information. With the highest enrollment of any group, approximately *70%* of them are pursuing undergraduate and graduate degrees. The enrollments of those with secondary or higher education are listed following the prior mentioned groups. The portions for the rest are all significantly smaller.

```
# Display Count of each entry in the Column
(table(enrolments_top_ten$employment_status))
```

```
##
## full_time_student  looking_for_work       not_working           retired
##               101                80                74               242
##     self_employed        unemployed           Unknown working_full_time
##                83                44              9385               294
## working_part_time
##                87
```

```
# Create new dataframe without "Unknown" entry
enrolments_ten_emp <- enrolments_top_ten %>% filter(employment_status != "Unknown")
# Plot a Pie Chart
bp_emp<- ggplot(enrolments_ten_emp, aes(x="", y=employment_status, fill=employment_status))+
geom_bar(width = 1, stat = "identity")+ coord_polar("y", start=0)+ blank_theme +  scale_fill_manual(valu
  theme(axis.text.x=element_blank())

bp_emp
```

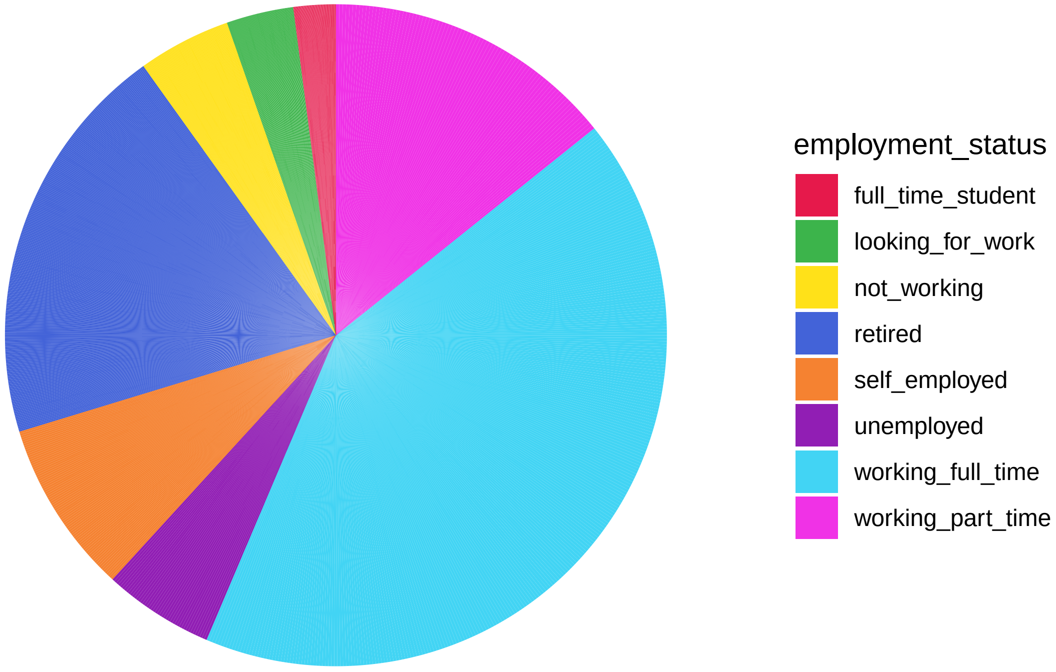


Fig 3: Pie Chart showing the ratio of the types of Employment status by the individuals enrolled in the course.

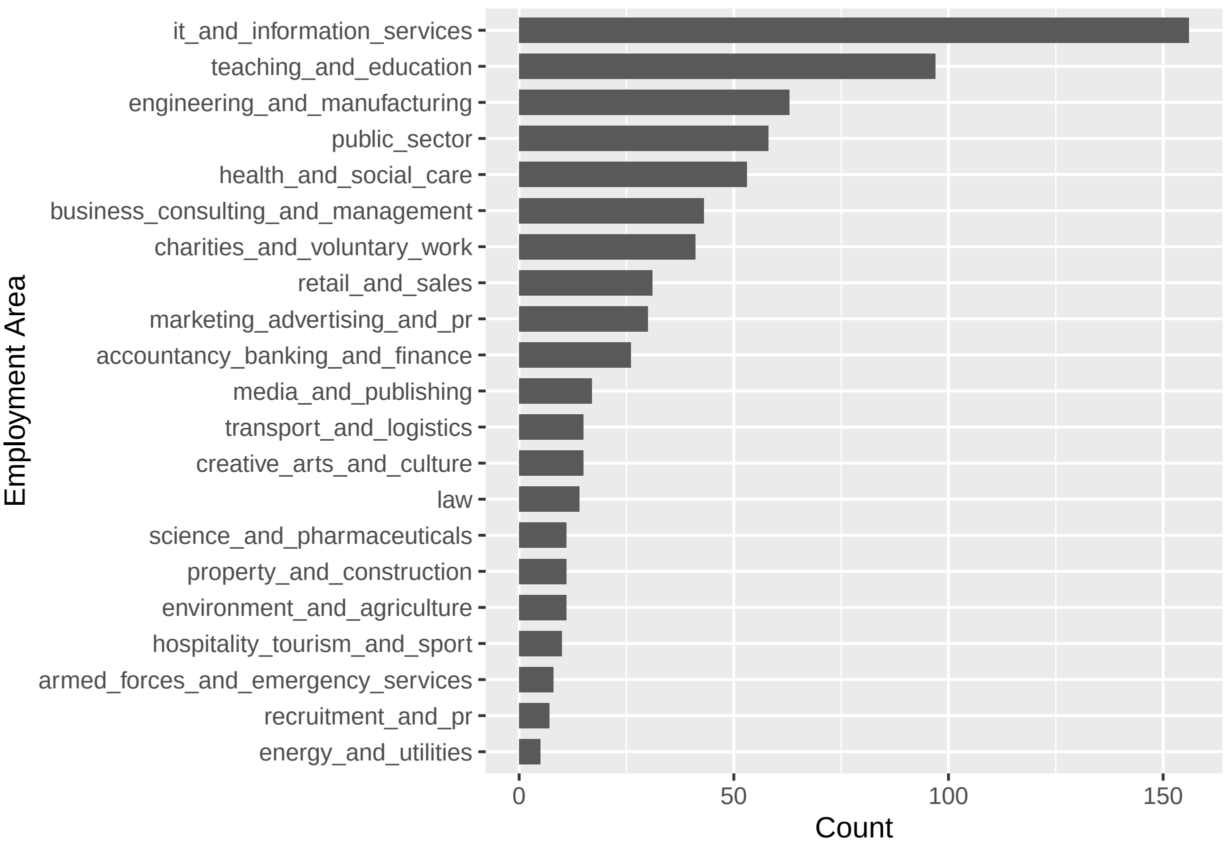
The information on the participants' employment status is then displayed in Fig. 3 as a pie chart. The majority of people—nearly 40%—are *full-time employees*, followed by *retirees* and *part-timers*. The portions for the next items are all somewhat smaller.

```
# Display Count of each entry in the Column
(table(enrolments_top_ten$employment_area))
```

```
##
##     accountancy_banking_and_finance armed_forces_and_emergency_services
##                                  26                                   8
##  business_consulting_and_management        charities_and_voluntary_work
##                                  43                                  41
##           creative_arts_and_culture                energy_and_utilities
##                                  15                                   5
##       engineering_and_manufacturing         environment_and_agriculture
##                                  63                                  11
##              health_and_social_care       hospitality_tourism_and_sport
##                                  53                                  10
##         it_and_information_services                                 law
##                                 156                                  14
##        marketing_advertising_and_pr                media_and_publishing
##                                  30                                  17
##           property_and_construction                       public_sector
##                                  11                                  58
##                  recruitment_and_pr                    retail_and_sales
##                                   7                                  31
##         science_and_pharmaceuticals              teaching_and_education
##                                  11                                  97
##             transport_and_logistics                             Unknown
##                                  15                                9668
```

```
# Create new dataframe without "Unknown" entry
enrolments_ten_area <- enrolments_top_ten %>% filter(employment_area != "Unknown")
# Create new dataframe by Gouping and Summarising
count_of_area <- enrolments_ten_area %>%
    group_by(employment_area) %>%
    summarise(Count = n())%>%
    arrange(desc(Count), .by_group = TRUE)
```

```
# Plot a bar chart
bp_area<- ggplot(count_of_area, aes(x =fct_reorder(employment_area,Count) , y = Count))+
  geom_col(width = 0.7)+ coord_flip()+labs(x="Employment Area", y = "Count")
bp_area
```



`Fig 4: Barchart representing the Count of different fields(area) of employment by the enrolled people.`

The industries that the enrolled persons are from are depicted in the bar charts above. The majority of participants come from the *IT and Information Services* industry, which is directly tied to the relevant course. *Engineering and manufacturing* come in second, almost half as high as the preceding sector, followed by *Teaching and Education.* However, based on the graph, people from practically all the major industries have enrolled in this course in some capacity.

**Additional Note**: The Unknown data is removed because, for few demographics more than 9000 entries are "Unknown" entry and this may make the visual very bad. But before output the number of entries for each value including "Unknown" is shown to understand the imporatce of proper data accquiring

## 2.5 Analysis Evaluation:

The top 10 nations with the most enrollment were thoroughly analysed in Section 2.5, along with information on the people's backgrounds in terms of education and occupation. As a result, the marketing team can concentrate on developing their course and attracting new students to the Cybersecurity course.

- The course is especially well-liked in nations like *GB, IN, SA, US,* and *NG*; many people from these nations are already actively enrolling in it. In these nations, there is a higher likelihood that more people will enlist.
- Since more than 50% of those enrolled are in universities, more attention can be paid to them in order to increase enrollment rates. This is because those studying *university bachelor's degrees* and *master's*

*degrees* may be keen to learn more about their courses or as a supplementary talent. This can be accomplished by working with the university's career service and emphasising the course's value to students during career fairs and other related events.

- On the other hand, it is clear from Fig. 3 that a significant portion of the working population has also enrolled in the course. Based on the study, it can be inferred that this course may help those who are *working full time* or *working part time* advance in their careers. People who have retired come in third; this is an unconventional finding, and more research may provide insightful results.
- The highest enrollment rate was from those working in *IT and Information services*, which may be due to the course's computer-related content, as shown by Fig. 5. *Engineering and manufacturing* are the next-highest sectors, followed by *Teaching and Education.*
- By focusing on the *top 10 nations* and the corresponding individual's *educational and job* background, the established results can be used by the stakeholders to organise their way of advertising and investing in the right way.
- Advertising can be concentrated on social media platforms targeted at the above- mentioned demographics and countries, career fairs at universities where students have some of the highest enrollment rates, and FutureLearn's website, which can partner with *IT*, *Engineering*, and *Educational* departments to provide these courses for their employees so they can advance their skills in case a Cybersecurity skill is required.
- The course can charge a price so that the FutureLearn firm can profit from individual students, and for employees, the course costs could be charged to the corporations so that the companies could profit from the knowledge the employees acquire.

The analysis that was done is a general examination of the dataset that was provided; further data may be needed for a more in-depth analysis.

### 2.5.1 Analysis Review:

The MOOC dataset contained a large amount of unknown and missing data, so an improvement in the data acquisition procedure could enhance the analysis. However, the analysis performed using the data that was available provided a comprehensive insight and a solution to the business demand.

### 2.5.2 Next Steps:

- A secondary CRISP-DM cycle will be performed to provide the marketing and sales team with further information.
- The secondary cycle will examine the view analysis of the course's steps (Chapters) as well as an examination of the electronic devices utilised to study the course.
- Finally, a joint deployment plan will be created for both analyses.

# 3 Secondary Analysis - Analysis on Devices used by Users and Views on Each step of the Course.

## 3.1 Business Understandind and Assessing the Situation :

### 3.1.1 Stakeholder Requirement:

The *Marketing and Sales Teams* of the FutureLearn website are the same stakeholders for the second portion of the report analysis, which is for the same *Cybersecurity course.* The marketing team is seeking ideas on how to promote the course to users of a specific device type. However, there is another application for this research; depending on the subsequent analysis, the development team will utilise it to further examine the views of each course step (Chapter) and determine whether the user interfaces are compatible and the UI design is pleasant for all types of devices.

### 3.1.2 Questions Answered in the Analysis:

- What kind of gadgets are most commonly utilised by consumers to study the course?
- Analyze which sections of the courses are most frequently seen by users and which devices are used the least to study the material, then advise to the development team that the course content and user interface be improved for those devices.

### 3.1.3 Available Resources:

The resources available for the following data analysis and report are as follows:

- Data : Cybersecurity enrollments dataset in multiple `.csv` files from MOOC collection of data.
- Computing Resources : Laptop or Personnel Computer to carry out basic analysis and report generation.
- Software : Windows OS, R and R Studio.

### 3.1.4 Requirements, assumptions and constraints:

- Completing the analysis and providing the stakeholders with the report no later than November 18th, 2022.
- Any missing data or data tampering that took place during the data recording could compromise the study's conclusions.
- There might not be sufficient reliable information to draw a convincing conclusion.

### 3.1.5 Risks and Contingencies:

- Poor data quality.
- Incorrect stakeholder interpretation of the analysis.
- Incorrect data analysis and understanding by the data analyst.

### 3.1.6 Terminology:

CRISP-DM : Cross-Industry Standard Process for Data Mining

MOOC : Massive Open Online Course

### 3.1.7 Analysis and Data Mining Goals:

- Remove any empty values (`NA`) from the data collection to tidy it up.
- Include columns for Steps (Chapters), view totals, and information on the viewing devices only.
- Create informative graphs that will help to more clearly convey the findings.

## 3.2 Data Understanding:

### 3.2.1 Initial Data Collection:

The data used for the following Analysis is present in the *MOOC* Dataset provided by the FutureLearn website for the course *Cyber Security: Safety at Home, Online, and in Life*. From the multiple `.csv` files present the files considered for this analysis are `cyber-security-3_video-stats.csv`, `cyber-security-3_video-stats.csv`, `cyber-security-4_video-stats.csv`, `cyber-security-5_video-stats.csv`, `cyber-security-6_video-stats.csv`, `cyber-security-7_video-stats.csv`. The above mentioned files are stored in the *data* folder of the Project Template. There was no issue with data accuring as this was provided by the Stakeholders themselves.

### 3.2.2 Data Description:

```
# Appending different .csv files data into single Dataframe
# Code executed in 02 - A.R file in munge folder of Project Template
video_stats <- rbind(cyber.security.3_video.stats, cyber.security.4_video.stats,
                     cyber.security.5_video.stats, cyber.security.6_video.stats,
                     cyber.security.7_video.stats)
```

To make working on the analysis easier, the preceding code chunk automatically combines many *.csv* files at the beginning.

```
# Display number of rows and columns
dim(video_stats)
```

```
## [1] 65 28
```

The data from all 5 *.csv* files are combined into the dataframe called `video_stats`. There are a total of *n=65* rows and *p=28* columns in this combined data set.

```
# Display all the column names
colnames(video_stats)
```

```
##  [1] "step_position"                 "title"
##  [3] "video_duration"                "total_views"
##  [5] "total_downloads"               "total_caption_views"
##  [7] "total_transcript_views"        "viewed_hd"
##  [9] "viewed_five_percent"           "viewed_ten_percent"
## [11] "viewed_twentyfive_percent"     "viewed_fifty_percent"
## [13] "viewed_seventyfive_percent"    "viewed_ninetyfive_percent"
## [15] "viewed_onehundred_percent"     "console_device_percentage"
## [17] "desktop_device_percentage"     "mobile_device_percentage"
## [19] "tv_device_percentage"          "tablet_device_percentage"
## [21] "unknown_device_percentage"     "europe_views_percentage"
## [23] "oceania_views_percentage"      "asia_views_percentage"
## [25] "north_america_views_percentage" "south_america_views_percentage"
## [27] "africa_views_percentage"       "antarctica_views_percentage"
```

The dataset's column names or *variables* are listed above, along with the number of views, duration of viewing for a specific video, and device types used for viewing.

### 3.2.3 Data Exploration and Data Quality Report:

```
# Display first 5 rows
head(video_stats, 5)
```

```
## # A tibble: 5 x 28
##   step_p~1 title video~2 total~3 total~4 total~5 total~6 viewe~7 viewe~8 viewe~9
##      <dbl> <chr>   <int>   <int>   <int>   <int>   <int>   <int>   <dbl>   <dbl>
## 1     1.1  Welc~      99    1659     113      36     221      58    77.0    75.4
## 2     1.14 Why ~     362     910      77       8     173      28    72.5    70.9
## 3     1.17 Pres~     241     723      63       5     120      16    73.7    73.9
## 4     1.19 Stay~     348     755      62       2     147      10    72.8    71.9
## 5     1.5  Priv~     281    1248     100      15     191      41    78.4    75.6
## # ... with 18 more variables: viewed_twentyfive_percent <dbl>,
## #   viewed_fifty_percent <dbl>, viewed_seventyfive_percent <dbl>,
## #   viewed_ninetyfive_percent <dbl>, viewed_onehundred_percent <dbl>,
## #   console_device_percentage <dbl>, desktop_device_percentage <dbl>,
## #   mobile_device_percentage <dbl>, tv_device_percentage <dbl>,
## #   tablet_device_percentage <dbl>, unknown_device_percentage <dbl>,
## #   europe_views_percentage <dbl>, oceania_views_percentage <dbl>, ...
```

```
# Display number of NA values in each column
(video_stats %>% summarise(across(everything(), ~ sum(is.na(.x)))))
```

```
## # A tibble: 1 x 28
##   step_p~1 title video~2 total~3 total~4 total~5 total~6 viewe~7 viewe~8 viewe~9
##      <int> <int>   <int>   <int>   <int>   <int>   <int>   <int>   <int>   <int>
## 1        0     0       0       0       0       0       0       0       0       0
## # ... with 18 more variables: viewed_twentyfive_percent <int>,
## #   viewed_fifty_percent <int>, viewed_seventyfive_percent <int>,
## #   viewed_ninetyfive_percent <int>, viewed_onehundred_percent <int>,
## #   console_device_percentage <int>, desktop_device_percentage <int>,
## #   mobile_device_percentage <int>, tv_device_percentage <int>,
## #   tablet_device_percentage <int>, unknown_device_percentage <int>,
## #   europe_views_percentage <int>, oceania_views_percentage <int>, ...
```

The data cleaning process will continue in next part by deleting the unneeded columns since it appears from the initial analysis of the data that the dataset contains no *NA* values.

## 3.3 Data Preparation:

### 3.3.1 Data Selection:

Only information on the views on each section of the video and the content viewed on various devices is needed for this particular analysis. The only fields that must be included are `step_position`, `title`, `video_duration`, `total_views`, `console_device_percentage`, `desktop_device_percentage`, `mobile_device_percentage`, `tv_device_percentage`, `tablet_device_percentage`and `unknown_device_percentage`.

```
# Removing Unwanted Columns
# Code executed in 02 - A.R file in munge folder of Project Template
video_stats_main = subset(video_stats, select = c(step_position, title, video_duration, total_views, co
```

`video_stat_main` is regarded as the primary dataframe for analysis after the preceding step.

### 3.3.2 Data Cleaning:

The primary dataframe was created by combining several *.csv* files, thus there are multiple rows of data for the same course stages. The initial phase of the data cleaning procedure would be to group the dataframe based on the step position. The aforementioned action is carried out via the following code piece.

```
# Grouping data based on a column and Summarising other columns
# Code executed in 02 - A.R file in munge folder of Project Template
(video_stats_main <- video_stats_main %>%
    group_by(step_position, title, video_duration) %>%
    summarise(total_views = sum(total_views),
              console_device_percentage = mean(console_device_percentage),
              desktop_device_percentage = mean(desktop_device_percentage),
              mobile_device_percentage = mean(mobile_device_percentage),
              tv_device_percentage = mean(tv_device_percentage),
              tablet_device_percentage = mean(tablet_device_percentage),
              unknown_device_percentage = mean(unknown_device_percentage)))
```

The grouping is done in such a way that the percentage numbers are considered as an average while the views over several rows of the same step course are added.

```
# Changing data type to Character
# Code executed in 02 - A.R file in munge folder of Project Template
video_stats_main$step_position <- as.character(video_stats_main$step_position)
```

```
# Display Datatype of Each Column
sapply(video_stats_main, class)
```

```
##             step_position                     title            video_duration
##               "character"               "character"                 "integer"
##               total_views console_device_percentage desktop_device_percentage
##                 "integer"                 "numeric"                 "numeric"
##  mobile_device_percentage      tv_device_percentage  tablet_device_percentage
##                 "numeric"                 "numeric"                 "numeric"
## unknown_device_percentage
##                 "numeric"
```

The change of the `step_position` column from `numeric` to `character` is important during data visualisation as the visualisation code may perform unnecessary mathematical calculations and also the values in the above mentioned column do not represent a numerical value but is present as an index for the title of the course.

```
# Display Number of Rows and Columns
dim(video_stats_main)
```

```
## [1] 13 10
```

Finally, the cleaned dataset contains *n=13* rows and *p=10* columns, these type of datasets are referred to as *wide* dataset because it has almost equal number of rows as columns.

## 3.4 Data Analysis:

```
# Plot a Bar Chart
ggplot(data=video_stats_main, aes(x=step_position, y=total_views)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=total_views), vjust=-0.3, size=3.5)+
  theme_minimal()+
  labs(x="Step Title", y = "View Count")+
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```

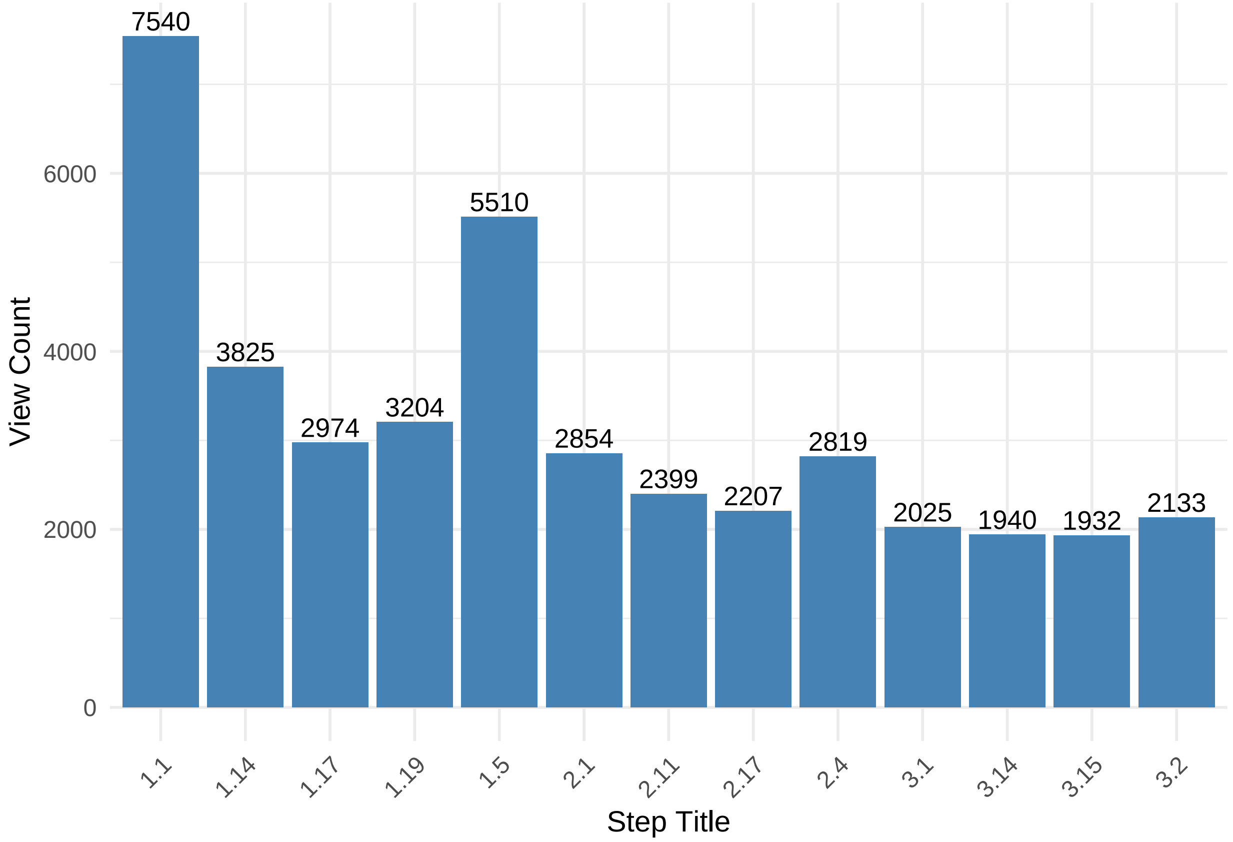


Fig 5 : Bar chart representing the Total number of views for each step of the course.

With a total view count of *7548*, Step 1.1 is the Step or Chapter that has had the most views, followed by Step 1.5 with *5510*. Generally speaking, the Steps from 1.1 to 1.5 each have more individual views than any other Steps. Therefore, it is clear that Chapter 1 overall has had the most views, with a total of *23053*, followed by Chapters 2 and 3 (*10279* and *8030*). Furthermore, *41362* is the total number of views for the cybersecurity course.

```
# Copy dataframe to new variable
devices <- video_stats_main
# Create new dataframe with Selected Columns
devices <- subset(devices, select = c(total_views, console_device_percentage, desktop_device_percentage
# Create Dataframe by Summarising columns.
devices <- devices %>% summarise(total_views = sum(total_views),
              console_device_percentage = mean(console_device_percentage),
              desktop_device_percentage = mean(desktop_device_percentage),
```

| total_views | console_device_percentage | desktop_device_percentage |
|---:|---:|---:|
| 41362 | 0.0301538 | 77.58908 |

| mobile_device_percentage | tv_device_percentage | tablet_device_percentage |
|---:|---:|---:|
| 11.56954 | 0.0009231 | 10.39785 |

```
              mobile_device_percentage = mean(mobile_device_percentage),
              tv_device_percentage = mean(tv_device_percentage),
              tablet_device_percentage = mean(tablet_device_percentage))
# Create new dataframe with Selected Columns
devices_1 <- devices[,1:3]
devices_2 <- devices[,4:6]
# Display Table
knitr::kable(list(devices_1, devices_2))
```

The table shows that out of the 41362 views almost *77.5%* (32055 views) are from *desktop* device. The next is *mobile* and *tablet* devices with each *11.6%* (4798 views) and *10.4%* (4301 views) respectively. The *Console* and *TV* share minute percentage of the total share.

## 3.5 Analysis Evaluation:

- It is clear from Fig. 5 that Steps *1.1 (Welcome to the course)*, *1.5 (Privacy online and offline)*, and *1.14 (Why would anyone want your data?)* have received the most views. Since Chapter 1 has had the most views overall and all other Chapters have received comparably less views, only about half of those who enrol in and begin the course actually finish it. This can have an adverse effect on those taking the course as well as the business operations of FutureLearn. Users may give the course a poor review as a result of this. The marketing and sales team should immediately address the issue by getting in touch with the development team, improving the course videos, or addressing any other specific reason, and one of the main recommendations would be to conduct further analysis on the reason why users don't finish the course.
- Table in Section 3.4 shows that, of the total 41362 views, roughly 77.7% are from desktop or personal computers. This is consistent with the earlier analysis, which showed that most of the enrollees were university students and people who were employed. However, the utilisation of other devices, such as smartphones and tablets, is fair at 11.4% and 10.5%, respectively. Therefore, the marketing team should work with the development team to ensure that both the FutureLearn website in general and the Cybersecurity course specifically are suitable with mobile and tablet devices, with ongoing updates and improvements to the desktop user interface as well. This might allow new user enrolment and encourage returning users.
- The idea of creating a nice user interface for mobile devices is a smart one because mobile devices are a major source of internet traffic in nations like IN and PK. Given that these nations are among the top 10 for student enrollment, there may be a lot of opportunity for profitable business.

### 3.5.1 Analysis Review:

- Even though the analysis couldn't provide a particularly in-depth study, it provided a general overview of the course steps so that each team's material could be improved.
- Based on the prior analysis in cycle 1, a comprehensive examination of the different types of devices was conducted, as well as recommendations for improving the user interface.

### 3.5.2 Next Steps:

- For both analyses, a collaborative deployment strategy will be developed.
- A presentation incorporating both the *Primary and Secondary analyses* from will be developed.

# 4. Overall Project Deployment:

This deployment is common, as was stated above in both analyses.

- As this is not a Machine learning Model but an analysis the deployment would be the stakeholders understanding the produced report and presentation and taking necessary actions.
- The two analyses - Primary and Secondary are a very good source of input for improving the Marketing and Business strategy to develop the *Cyber Security: Safety at Home, Online, and in Life* and also performing more in-depth analysis on the suggested ideas.
- The benefits of the analyses can be seen and measured based on the improvement in the enrolment count for the course and also enrolled users completing the full course. This will inturn increase the income and also positive feedback and reviews form the enrolled users.

## 4.1 Monitoring and Maintainance:

- The analysis can be manually updated since it is a general comprehension of the data and will only be helpful if there is a significant change in the trend. Therefore, dynamic analysis will be of little help and may result in resource waste.
- A frequent update of the report is necessary since there is a very good likelihood that the trend of the analysis's findings will alter. This update may take place every six months or once a year. There is a possibility that the enrolment rate will grow from other elements as well if the marketing campaign is conducted based on the study. Therefore, to follow the changes periodic update of the analysis and report is required.