CSC8631 | Data Management and Exploratory Data Analysis | FutureLearn Dataset Analysis Report

Deepika Chandramouli

2023-11-08

Introduction

The FutureLearn Massive Open Online Course (MOOC) was run by Newcastle University in the course titled "Cyber Security: Safety At Home, Online and in Life" [1]. The data set was collected from 7 runs of the same course to analyse the audience. It gives insights on the enrollments (male, female, country, age range), survey responses, archetypes, responses to quizzes in the module and view percentage across continents. The cyber security course is provided for 3 weeks and covers the basics of privacy, information theft, payment safety and security at home. The task in hand is to analyse the data set based on viewer profile to improve the dynamics of online learning provided by FutureLearn platform.

Figure 1: FutureLearn Cybersecurity course by Newcastle University [1]

Round 1 of the CRISP DM Cycle

Business Understanding

The first step of the CRISP DM cycle aims at gathering requirements from a business perspective. It is also important to look into the assumptions, constraints and risks imposed on the stakeholders helping to deduce a successful project plan. The above business criteria have to be achieved with cost benefits while improving efficiency of the online platform for the audience.

Requirements, Assumptions and Constraints

In this phase, the objective is to target on the Cyber security MOOC course and gain valuable insights. A thorough understanding of the course will support in identifying its target audience. The requirements need to be gathered in line with the business needs helping them grow as a sector. In this case, since its an online learning platform, business interests might be assumed in the areas of number of enrollments in the course, demographic areas of the audience and their occupation. For this purpose, the information has to be collected by running the course several times. The main constraint would be identifying user interaction because participants are not required to complete the course. The course is completed at the student's own pace and willingness.

Risks, contingencies and Successful project plan

In any business use-case, there are potential risks that could jeopardize the successful delivery of the online course. Risks may be encountered with inadequate learner engagement, geographical reasons, time-zones, and budget. The improvement of learner engagement could be facilitated by promoting the course to bring in a large user base. Since the course is online-based, it can be accessed by end-users in their comfortable time-zones. Finally, the cost might need to be tracked regularly so that it is not overrun by the initial budget proposed by the stakeholders. A successful project plan should address all the above points of constraints. Further, the project plan should highlight keys areas of business like cost, risk management, tools and techniques involved in the development of the course. In addition to this, the data collected has to be as accurate and relevant as possible.

What are our Areas of interest?

The report focuses on analyzing the target audience enrollment in the cyber security course considering key demographic factors such as gender, occupation, geographical location and age range. The output generated will help the business in enhancing the course to meet the expectations of learners. The aim is to craft a course so that it reaches a wider range of learners.

Data Understanding

The primary step in satisfying the business needs is to understand the data collected from key sources of information. various factors have to be analyzed to gain insights on potential learners. The source of data is very important in order to be accurate and produce significant results. Therefore, data understanding focuses on exploration and collection of data thus improving data quality.

Data collection and description

In this phase, the raw data highlighting enrollment details from 7 Cybersecurity course runs is collected from the FutureLearn online platform. The entire data is available as a .ZIP file as "FutureLearn MOOC Dataset". The project template folder contains the unzipped data in raw manner (.CSV files). These include data over 7 runs as described below:

- cyber.security.1_enrolments- Data contains audience details based on gender, age, occupation, university, country and enrollment timings.
- cyber.security.1_leaving.survey.responses- Data contains information on learners who are leaving the course, the reason for their leave, timings and course completion steps.
- cyber.security.1_question.response- This data contains quiz response to MCQ indicating the correct and wrong answer.

- cyber.security.1_archetype.survey.responses- Data contains information about learner archetypes indicating who they are. These include: Advancers, Explorers, Fixers, Flourishers, Hobbyists, Preparers and Vitalisers.
- cyber.security.1_step.activity- Data contains course step information based on week, first and last visited date and time.
- cyber.security.1_weekly.sentiment.survey.responses- This file contains feedback and experience of learners about the course including rating.
- cyber.security.2_team.members- This file contains various team roles and user roles

Data exploration and quality

Data exploration is the next step after collection that identifies the variables best suited for our needs. By exploring the data you tend to analyze patterns that improve data quality. For instance, in each run, a particular learner is uniquely identified using the learner ID. As such, several CSV files are included in the raw data over 7 runs. For each run there are enrollments thus making up to 7 enrollment CSV files. Although, the enrollment data is collected for different learner IDs, we take into account only one run to analyze the target audience. This is because we assume that our subsequent runs might also follow the same pattern and trends as our first run. Thus, it is very vital to explore the data before making a concrete decision. This avoids data quality issues in the longer run. Here, the main reason we are targeting the enrollment data is because it is consistent throughout 7 runs. For instance, the "Leaving survey response" file do not have data for the first 3 runs which might pose potential risks on data quality. The data gathered will then be pre-processed to enhance the learner experience in our areas of interest.

Data Preparation

The next phase focuses on preparing the data which is then used for modelling and evaluation. In this phase, we are interested in selecting the data followed by cleaning and transformation. The end goal is to produce a final data frame that contains the most useful set of information.

Data selection

The data is carefully selected taking into account quality improvements. For our analysis, we select data on enrollments giving us maximum details about the end users. The data is populated and is also uniform across different runs giving us the most information and yet not compromising on quality. Any run data could be taken for our analysis as they contain the same variables. For instance, here we take the Enrollment data from run 6. We name the entire data frame as "EnrollData".

```
## [1] 3175   13
```

We could understand that the enrollment data contains 3175 rows and 13 columns.

Data Cleaning

As such the data contains lot of missing and unknown values which might not be useful for our analysis. This could be identified using the below R code which highlights missing values by " " or "–"and unknown values by "Unknown". Both these values have to be removed to avoid misleading results. These values also impose biasness and inconsistency to our data interrupting significant outcomes.

```
##                             learner_id             enrolled_at
## 1 c6f25ce6-0b07-4401-954c-92338d6a3cbb 2018-07-19 17:15:30 UTC
## 2 8a3386e2-444a-4cc5-b758-6255bc68a59b 2018-06-22 23:38:37 UTC
## 3 6e6fa9df-15c4-45c9-a0fc-a210c623cbef 2018-07-10 13:14:49 UTC
##             unenrolled_at    role   fully_participated_at
## 1 2018-10-29 09:42:30 UTC learner
## 2 2018-10-30 00:18:12 UTC learner
## 3                         learner 2018-10-27 16:19:19 UTC
##    purchased_statement_at  gender country age_range highest_education_level
## 1                         Unknown Unknown   Unknown                 Unknown
## 2                         Unknown Unknown   Unknown                 Unknown
## 3 2018-09-21 13:35:43 UTC Unknown Unknown   Unknown                 Unknown
##   employment_status employment_area detected_country
## 1           Unknown         Unknown               TR
## 2           Unknown         Unknown               GB
## 3           Unknown         Unknown               GB
```

Also, for instance, if you take the age_range column, it has 2915 rows that are having Unknown data which needs to be filtered. This removes most of our unknown records. But we could still find few missing and unknown values in country, gender and other columns. We do a filter to remove all the missing and unknown values to the columns of our interest.

Columns of interest: gender, country, age_range, highest_education_level, employment_status. The munge file "01-A.R" contains scripts used to preprocess our data. After cleaning the data we are left with 252 rows and 13 columns.

Data transformation

Following the data selection and cleaning, we do a data transformation to get the most useful information about data. This includes applying various data wrangling techniques including normalizing variables to suit modelling. At first we have handled the imbalance in the data by filtering the unknown and missing values from the columns of our interest. Further, we group the columns based on gender to know the range of data spread across each gender.

```
## # A tibble: 3 x 2
##   gender count
##   <chr>  <int>
## 1 female   135
## 2 male     116
## 3 other      1
```

We could see that there are more female enrollments (135) than male (116). This indicates that females are more interested in taking the cyber security course over male. In addition to the gender, we could also group the enrollments based on country to know the top 5 enrolled countries. This is done by grouping and then sorting the country count in decreasing order.

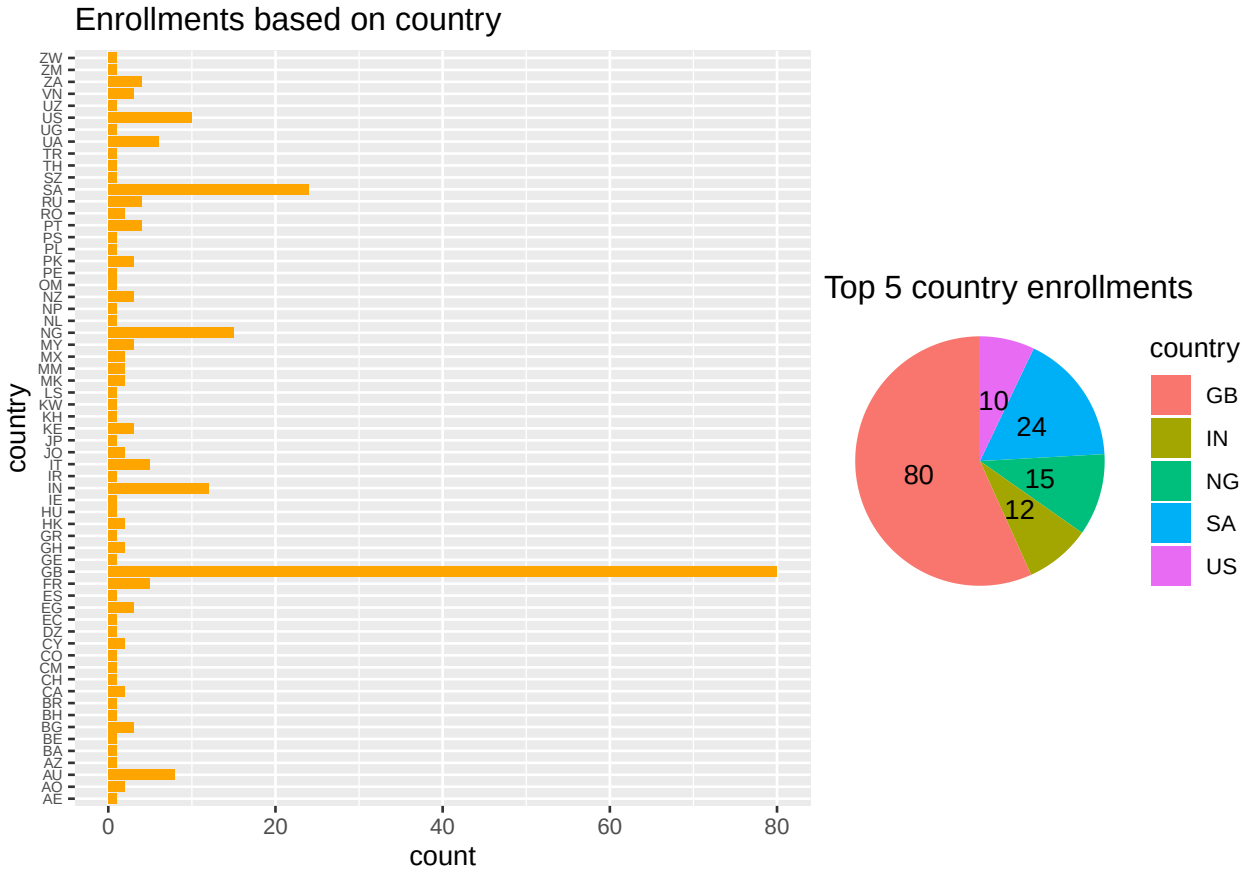
```
## # A tibble: 63 x 2
##    country count
##    <chr>   <int>
##  1 GB         80
##  2 SA         24
##  3 NG         15
##  4 IN         12
```

```
##  5 US          10
##  6 AU           8
##  7 UA           6
##  8 FR           5
##  9 IT           5
## 10 PT           4
## # i 53 more rows
```

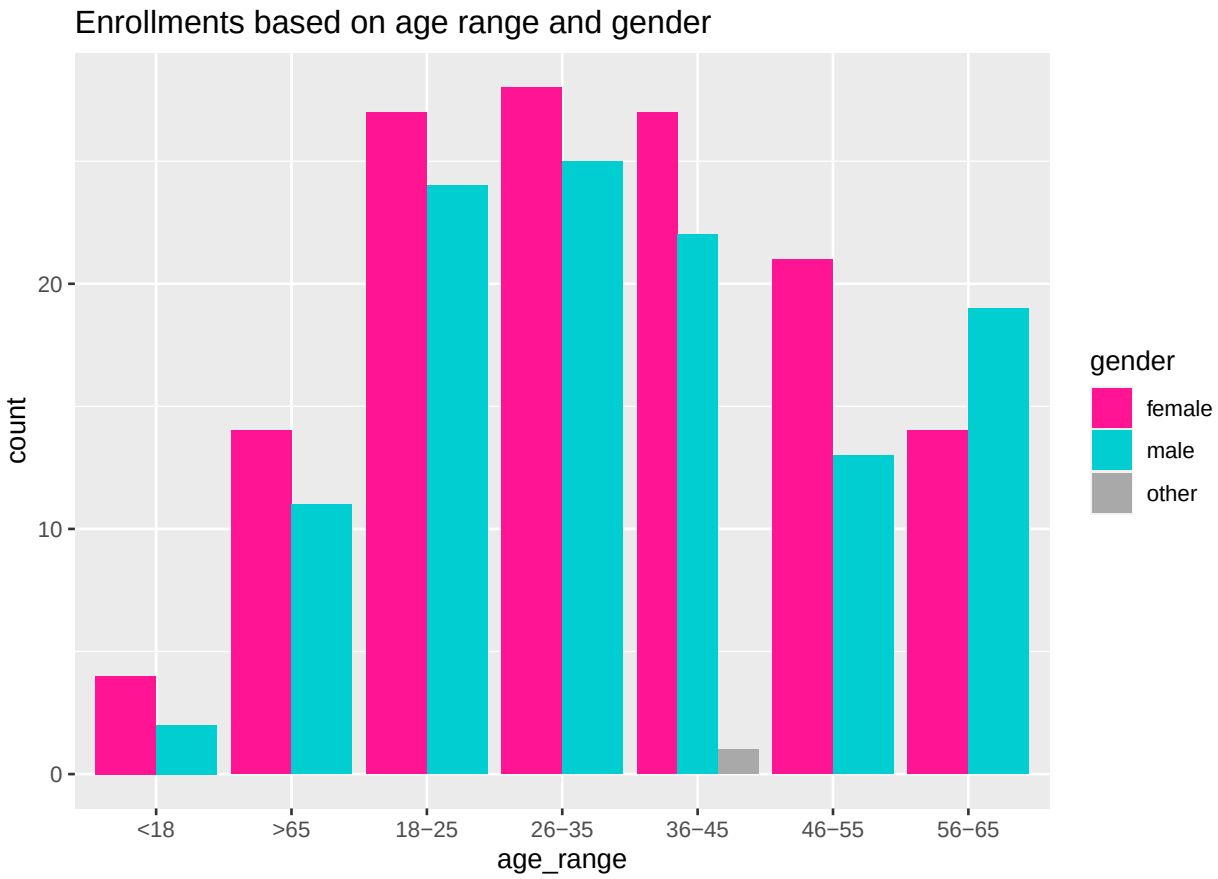
We could see that the maximum enrollments are from Great Britain (80), followed by South Africa (24), Nigeria (15), India (12) and United States (10). It is obvious that since the course is offered from GB, we could see more enrollments from Great Britain. Thus the data is reduced based on country and gender for further modelling. We cache this data to run from cache folder which makes computations quicker than munge folder. Munging is enabled FALSE and Cache loading is enabled TRUE in the *config/global.dcf* file to achieve this.

Modelling

The next phase of our CRISP DM cycle is Modelling, where we perform exploratory analysis by visual representation of data. This brings out the major insights about data in a more understanding form. Lets first look at the overview of the enrollment data based on country.

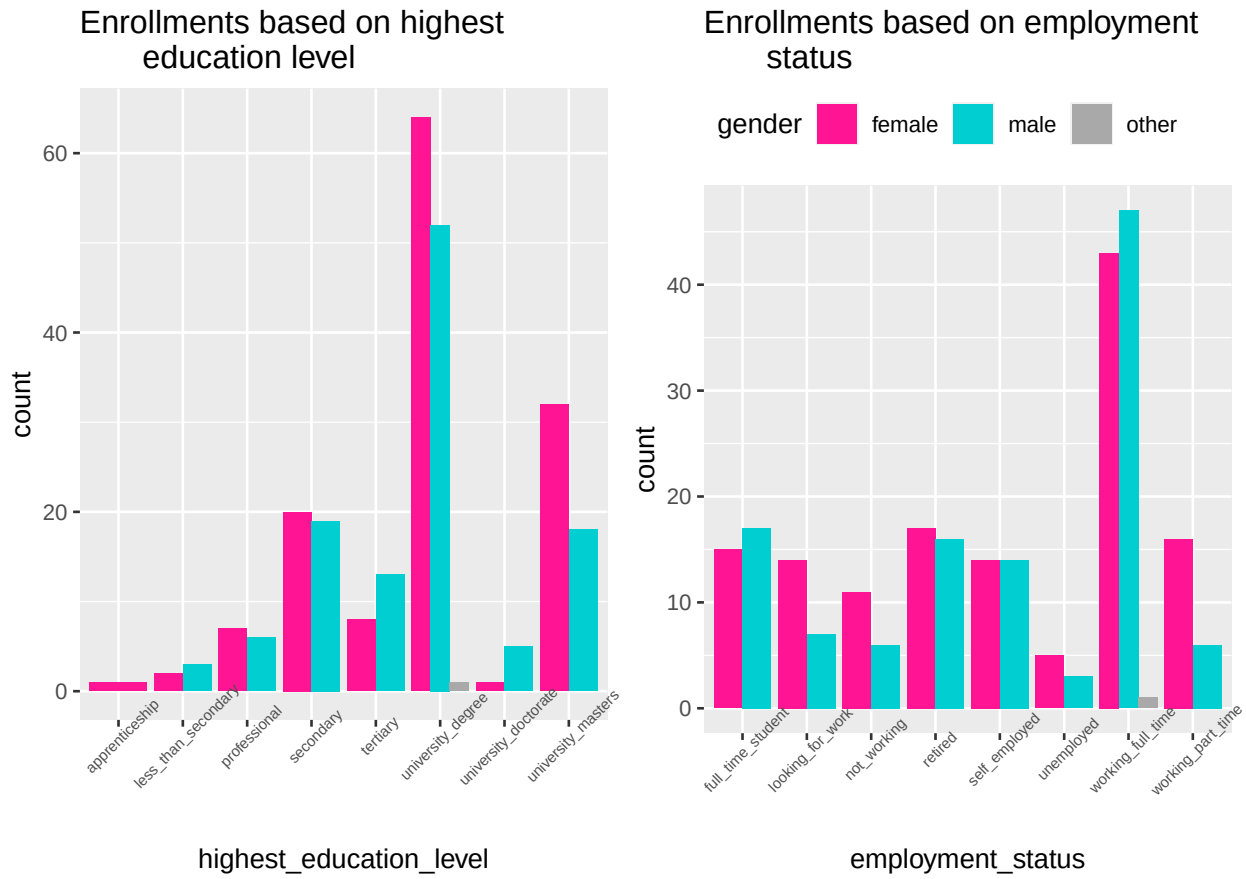


This clearly explains our previous interpretation where Great Britain has the maximum enrollments. We can also examine the distribution of enrollment data among the top five countries. This is visualized using the above pie chart that rightly splits the countries based on enrollments



The above model focuses on age ranges to understand the demographic age groups of the audience. Additionally, we categorize the age ranges by gender to explore the distribution of data across different gender groups within various age brackets. Age groups between 26 to 35 has the maximum enrollments (53) to the course where female participants take the lead of around 28 enrollments. Alternatively, we have very few participants from age groups less than 18.

Next up, we do a comparison of enrollments based on highest education level and employment status over gender to analyse the maximum targeted participant areas. We do this analysis to understand where the audience come from.



We could find that the maximum enrollment shoots up with University degree students where female enroll the most over male. On the other hand, full time employees contribute to the maximum enrollments where male being the majority over female. To conclude, overall we could see maximum enrollments from female university degree students between age groups 26-35 over others.

Evaluation

The next phase of CRISP DM cycle evaluates all our previous understanding and analysis. This involves evaluating the results gathered from the insights. Our models and analyzes should also meet the data quality in order to meet the success criteria. From our evaluation, we could clearly understand that our end goal is to know the where the enrollments come from in order to maximize our business needs. So far from the models, we could deduce that the maximum target enrollments are from Great Britain. Female university degree students between the age groups 26-35 contribute maximum towards the enrollment over male university degree students. However, male full time employees do tune in to the cyber security course over others. The results could be easily understood from the clustered bar plot of enrollments across various criteria. Almost in all the graphs, female enrollments are higher when compared to male enrollments. This could be inferred using the legend for gender. This moves our focus to less engaged segments, notably targeting age groups between 18-25 or male students as a whole. There is a chance to boost advertising efforts in this area, potentially increasing enrollment among these populations. Even engagement could be maximized to Doctorate students motivating them to take up the course.

Deployment

The final phase of the CRISP DM cycle focuses on enhancing our model to best align with the enrollment output data. Though we collected maximum enrollment insights via age, gender, country, employment status and education level, we could still not get what type of people (archetypes) are tuning into the course.

- Are they first time learners?
- Are they explorers who are taking a career diversion?
- Are they advancers who want to enhance their prior knowledge?

To answer all these questions, we do a second investigation where we review all our understanding to bring out a concrete information. This brings out a much more refined understanding of business requirements and motivates in deducing accurate results involved in the enrollment data.

Round 2 of CRISP DM Cycle

Business Understanding Review

Following the first round of CRISP DM Cycle, we review our analysis to further analyze the data to best suit the business requirements. Our initial goal still remains the same where we want to get know our target audience from the enrollments made to the course. But this time, we expand our requirement to know who are our audience apart from gender and nationality. To know what kind of an online learner our audience are, we would be interested in knowing the archetypes of the participants [3].

What is an archetype? * According to our context, archetype is a behaviour or patterns observed in an online learner [3]

FutureLearn's research team have identified archetypes and grouped them into three: [3]

- Work and Study- Advancers, Explorers, and Preparers
- Personal Life- Fixers and Flourishers
- Leisure- Hobbyists and Vitalizers

[3]

That makes up around 7 archetypes where our audience fit in. This type of data is important to analyze as it directly conveys the nature of our audience. Hence, it would be an interest to business to target these areas. The risk here would be a person unknowingly submitting his archetype. Thus, clear understanding of what each archetype is, has to be given to the audience before collecting the survey responses. That concludes our review of business understanding where the project plan is set out. Once the project plan is set out, we work on achieving the success criteria aligned with the model.

Data Understanding Review

In this phase, we try and expand our data so that it meets the archetype requirements as per the business needs. The source of data still remains the same enrollment data alongside the archetypes data that explains the learner types. The data is carefully taken after analysis and run 6 is preferred as our first review was also done on run 6. Other runs like run 1 and run 2 do not specify the archetype details, hence it is very vital to choose a run that contains information about archetypes. Thus our data quality guidelines are still met. As discussed above, the enrollments data contains information about people enrolled to the cyber security course with criteria such as gender, age_range, employment status, country and highest education level. On

top of this, we impose a new criteria "archetype" available in the archetype file. Both these files are available in .CSV format frm the data folder.

Data considered:

- cyber.security.6_enrolments
- cyber.security.6_archetype.survey.responses

The archetype survey responses describe the data collected by the FutureLearn user experience research team corresponding to seven patterns of behaviour by the learner [3].

Data preparation

Once our data source is fixed, we move on to our next phase of CRISP DM cycle where we prepare our data for further analysis and modelling. Firstly we stick on to our data selection to be archetype survey responses from run 6. The file is placed in the data folder along with other files. Now, let us have a look at the various columns (variables) contained in the data file

```
## [1] 208   4
```

We could understand that the archetype data contains 208 rows and 4 columns. As such the data would not be used as we need to clean our data before observing the insights.

```
##        id                            learner_id            responded_at
## 1 1880534 2d7e0b90-0ca3-48c9-a8e2-7ec870f04337 2018-04-09 20:18:52 UTC
## 2 1880823 a4d27ac3-ebdf-42c6-b1f2-50731d71f1e6 2018-04-13 18:38:44 UTC
## 3 1882371 435fefd4-29ee-4707-ba5a-f74f12130f19 2018-06-13 03:45:39 UTC
## 4 1889048 e0583ed1-f581-4eee-8f0a-016f5ec4d67c 2018-04-11 07:56:03 UTC
## 5 1891541 1e524488-26de-4101-8545-bf082a777671 2018-04-09 21:10:03 UTC
## 6 1893275 fc78d726-ec17-4c40-bc84-f5abf87a7991 2018-04-10 00:06:15 UTC
##    archetype
## 1     Fixers
## 2     Fixers
## 3     Fixers
## 4      Other
## 5 Vitalisers
## 6  Hobbyists
```

If you take the archetype column, we could find "Other" indicating a misleading observation. There are 18 observations accounting for "Other" archetypes. These data have to be removed or filtered as they pose an outlier characteristic over the weighted data.

The munge file "02-A.R" contains scripts used to preprocess our data. After cleaning the data we are left with 190 rows and 4 columns. During project load, the munge files are executed in order based on file names ordering.

Next up, we group the archetype data to understand where each of the seven archetypes lie. This provides more insight about the data and gives us an understanding of patterns and trends within the archetypes.

```
## # A tibble: 7 x 2
##   archetype   count
##   <chr>       <int>
## 1 Advancers      26
```

```
## 2 Explorers      40
## 3 Fixers         25
## 4 Flourishers     6
## 5 Hobbyists      26
## 6 Preparers      16
## 7 Vitalisers     51
```

We could see that the maximum archetype count is from Vitalisers contributing around 51 followed by Explorers of 40 and Flourishers are found the least with just 6 learners.

Followed by our analysis of archetypes, we want to know where each archetypes fit in the enrollments data to understand the interests of our target audience. We do this analysis by combining both the data files and arriving at a new data frame. "Learner ID" column is used as a unique value as it indicates one unique learner per ID. This Learner ID column will be used to combine both the enrollment and archetype data files. Merge functions under dplyr package are used to perform this transformation. From archetypes data file, the exact column match is identified using Learner ID field and this is used to merge the file with enrollment data file. After merging the data files, we now have a dimension as below.

```
## [1] 33 16
```

After merge, We could see that there are 33 rows and 16 columns. The 4 columns of Archetype data are combined with 13 columns of Enroll data and "Learner ID" is a common column making up to 16columns in the merged data. Lets have a look at the archetypes after merge. The archetype data and merged data are both cached for further computations

```
## # A tibble: 7 x 2
##   archetype   count
##   <chr>       <int>
## 1 Advancers       9
## 2 Explorers       7
## 3 Fixers          4
## 4 Flourishers     2
## 5 Hobbyists       5
## 6 Preparers       1
## 7 Vitalisers      5
```

Here Avancers (9) take the lead followed by Explorers (7) whereas Preparers are very few in number (1). Hobbyists and Vitalisers are same in number (5). We could also observe the gender ratio after merge where male takes the lead over female. This is contrary to our analysis in the first cycle where female showed majority. Although male seems to dominate, it could be neglegible as the change is not predominant.

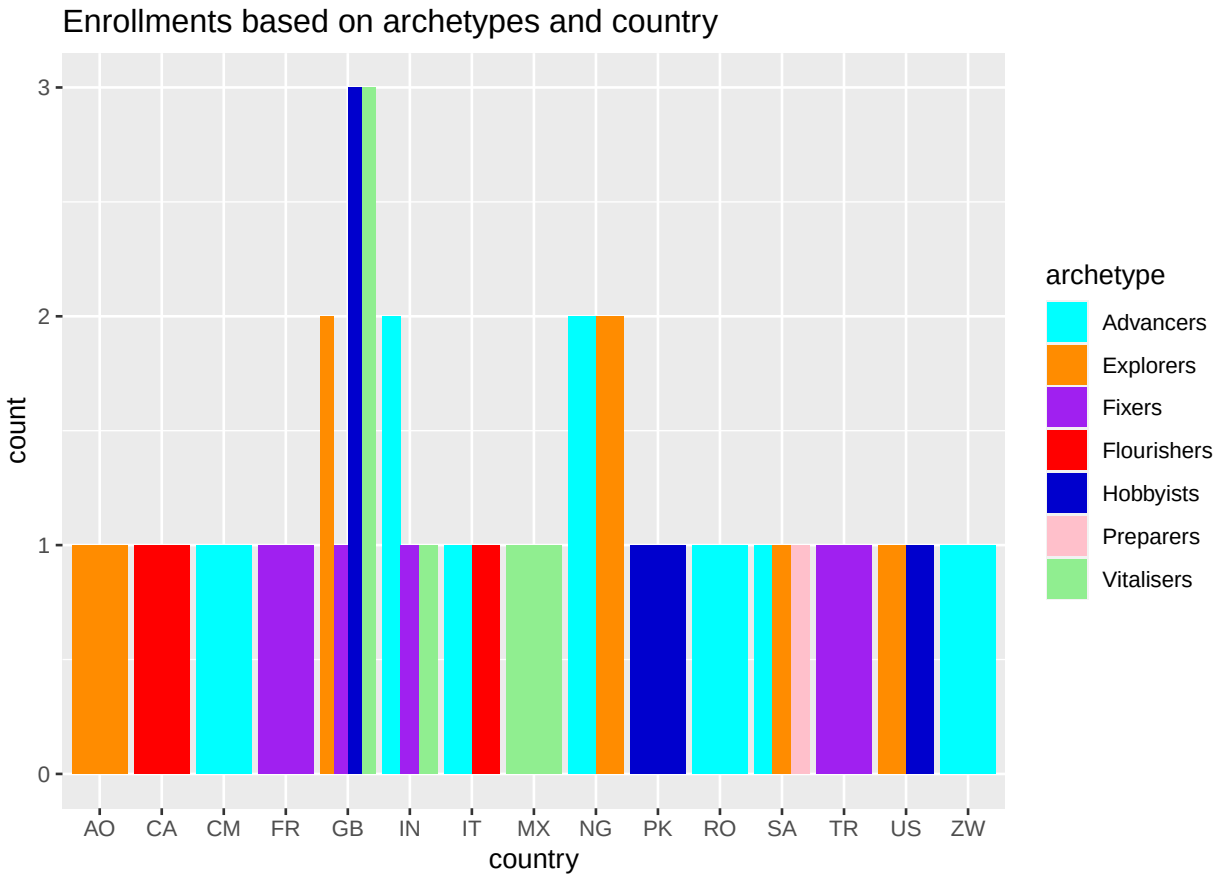
```
## # A tibble: 2 x 2
##   gender count
##   <chr>  <int>
## 1 female    15
## 2 male      18
```

Followed by our archetype and gender analysis, we also want to know where our country data stands.

Great Britain still tops in the number of enrollments making up to (9). South Africa is pushed to third position with over (3) enrollments.India and Nigeria are in second position with (4). We have a new country in addition IT-"Italy" with (2) enrollments. This observation partially matches with our previous review. There is a change in ordering and Italy takes up the spot instead of the United States. Thus the data is reduced based on gender and archetypes for further modelling.

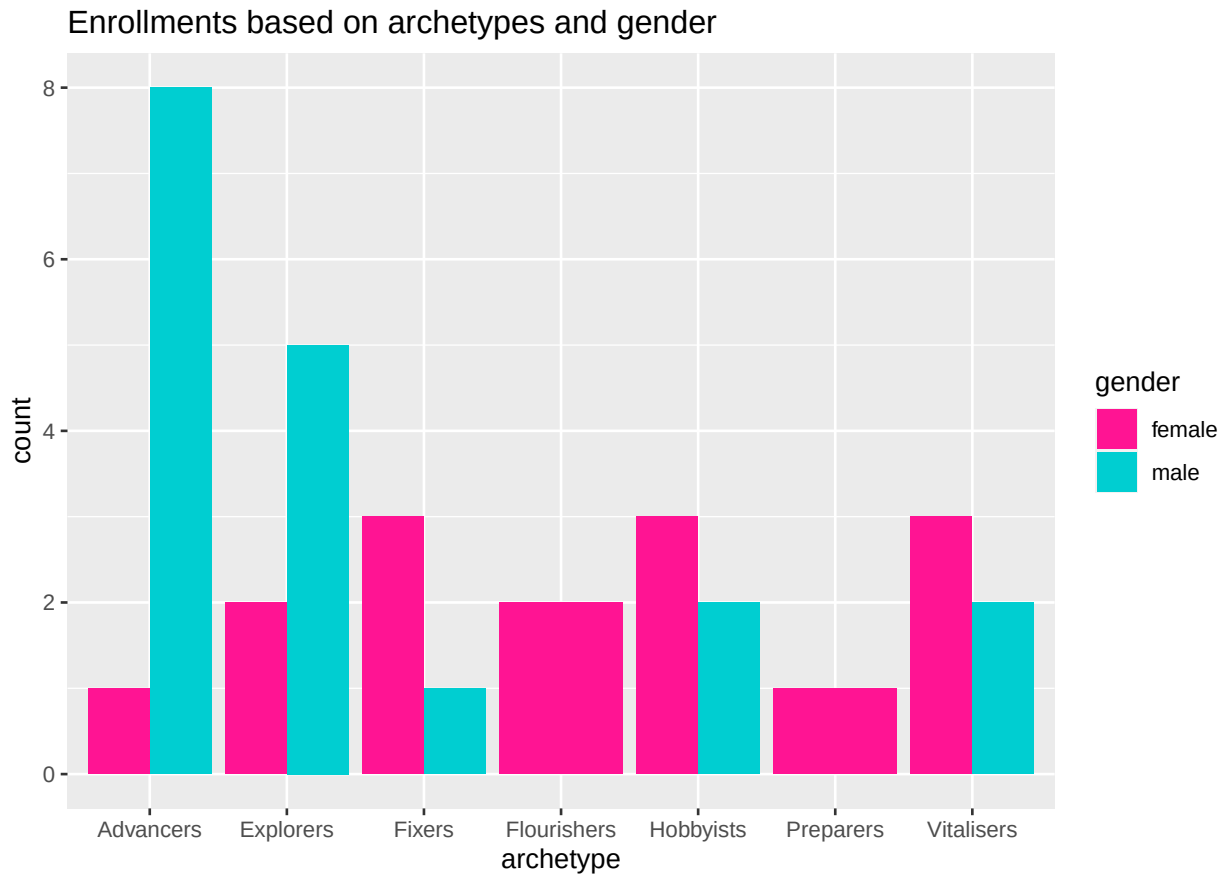
Modelling

The next phase of CRISP DM Cycle is modelling where we review our previous models to archetype data. Here we come across 7 archetypes represented by 7 fill colours. To understand which country each of our archetypes are from, we make a clustered bar chart. Bar chart works well with categorical variables and thats why it is preffered over other charts. More number of bars in a country indicates a diversified group of learners. On the other hand, a single bar indicates that only one particular learner is from that country.



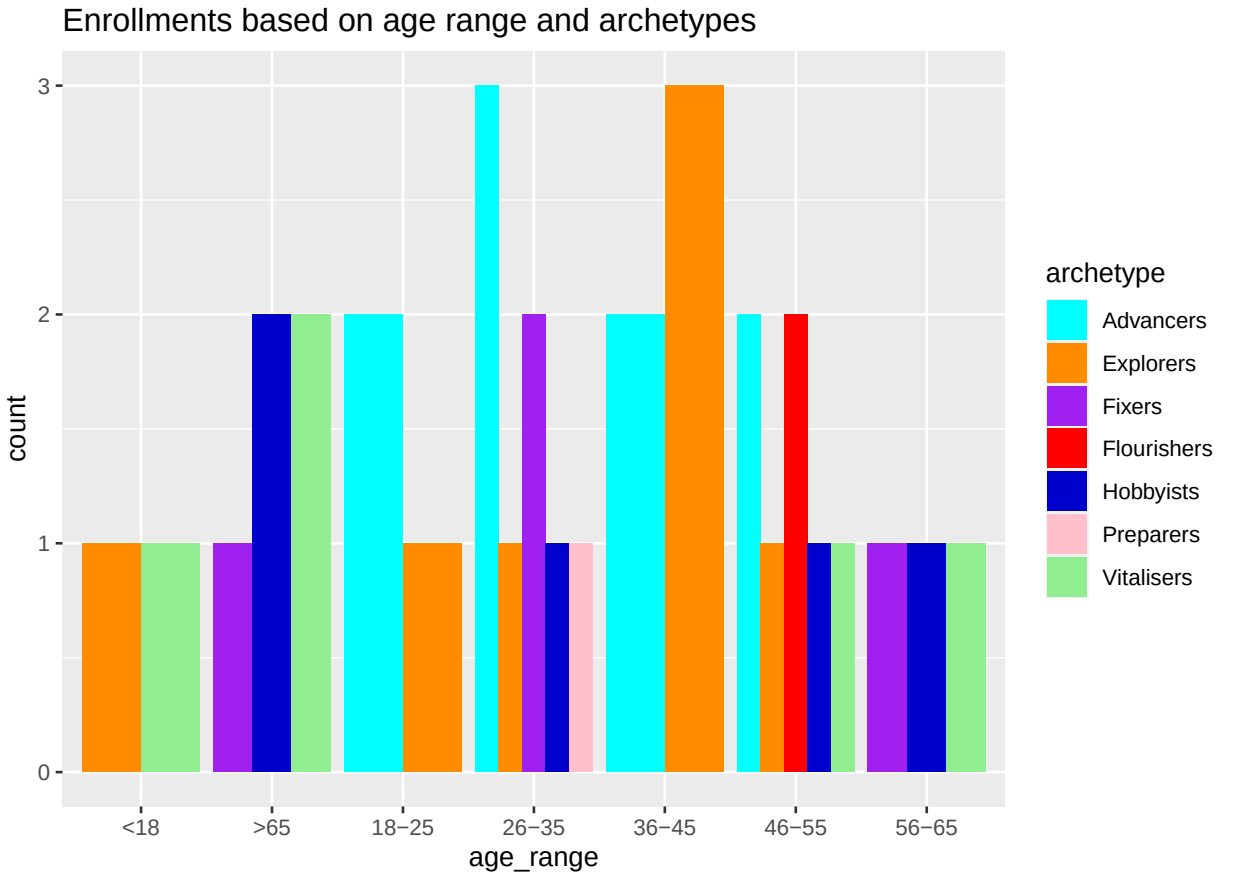
From the above plot, we could clearly see that only in the top 5 countries, we see more bars expressing more archetypes. In less enrolled countries we just see one archetype. This justifies our previous analysis of taking the top 5 countries into account for visualisation. Further, we see that Great Britain has the highest number of bars over other countries indicating a diverse range of learners from Explorers, Fixers, Hobbyists to Vitalisers. India and Nigeria also show the trend with Advancers being majority. Thus to conclude, we could observe more "Advance learners" from India and Nigeria and "leisure learners" from Great Britain.

Further, we analyze archetypes based on gender to get a more concrete picture.



We could observe that Males are generally Explorers and Advancers giving more time to work and study whereas females are fixers, flourishers, hobbyists, preparers and vitalisers giving more time to personal life and leisure. This is obvious from our day to day life as majority of women prioritize personal life over work.

To understand what is the age ranges of various learner archetypes, we could again do a clustered bar plot indicating the age ranges. The bar with the highest peak indicates that the archetype belong to the bar's age range.



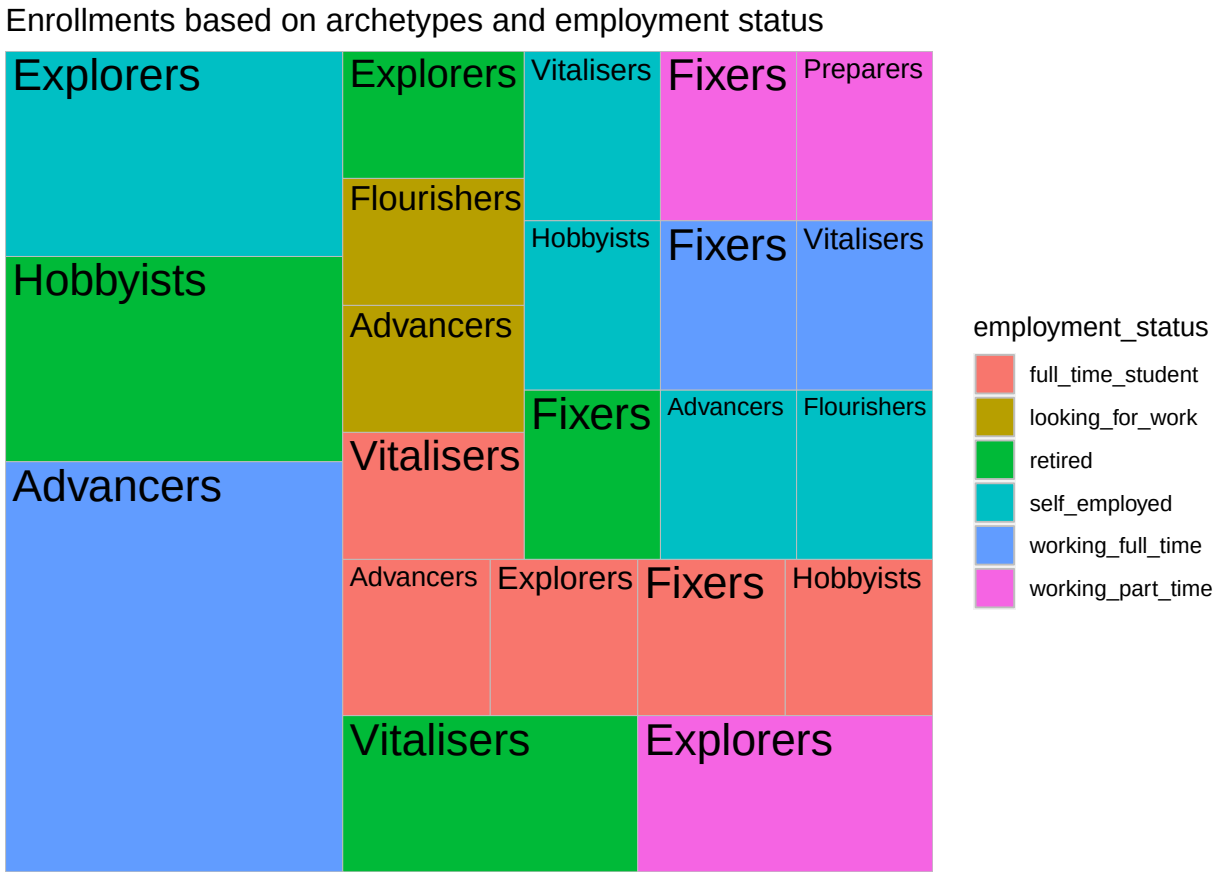
Here we could infer that Advancers are majority in the age groups 26-35.Flourishers on the other hand are between 46-55. This adds on to the gender analysis where Adavancer males are generally between 26-35 age groups and flourisher females are in 46-55 range.

Lets explore the archetypes based on employment status. We do this in order to know the learner types alongside employment status. A treemap is used to compute this as it works best with categorical variables.

We have two categories:

- Employment status
- Archetype

We could consider tree map as a hierarchical representation of our data. The entire rectangle specifies the total enrollments made after combining both data files. The size of the rectangle indicates number of archetypes enrolled. Additionally, here, color indicates employement status given in the legend.



We could see that full time employees are majority in number accessing the course who are Advancers, Fixers and Vitalisers. People looking for work contribute least towards the course who are flourishers and advancers. Hence, we could generalize full time employees as Advancers , self employed as Explorers and Retired as Hobbyists as they are the majority.

Evaluation reviewed

We have arrived at the evaluation phase where we compare multiple business criteria to improve enrollments to the cyber security course. We could find that archetypes make a significant impact in enrollment as they group the criteria to a specific function. We could deduce that:

1. Advancers are life long learners who love to study and improve their acumen. They make up the majority of enrollments to the cyber security course up to 27.27%.
2. Male enrollments are more when compared to female after merging of data.
3. Enrollments from “Advance learners” from India and Nigeria and “leisure learners” from Great Britain are observered in general.
4. Female enrollments are from people who give importance to personal life and leisure than work and study like male enrollments. Thus learner types from fixers, hobbyists and Vitalisers are female making upto 20% each whereas Advancers contribute the majority in male enrollments 44.4%
5. Adavancer males are generally between 26-35 age groups and flourisher females are in 46-55 range.
6. Full time employees are majorly Advancers , self employed are Explorers and Retired are Hobbyists.

```
##     archetype count Archetype_Percentage female_percentage
## 1   Advancers     9            27.272727          6.666667
## 2   Explorers     7            21.212121         13.333333
## 3      Fixers     4            12.121212         20.000000
## 4 Flourishers     2             6.060606         13.333333
## 5   Hobbyists     5            15.151515         20.000000
## 6   Preparers     1             3.030303          6.666667
## 7  Vitalisers     5            15.151515         20.000000
```

```
## # A tibble: 5 x 3
##   archetype  count Male_Percentage
##   <chr>      <int>           <dbl>
## 1 Advancers      8           44.4
## 2 Explorers      5           27.8
## 3 Fixers         1            5.56
## 4 Hobbyists      2           11.1
## 5 Vitalisers     2           11.1
```

Deployment

That brings us to final phase of the CRISP DM cycle focusing on enhancing our model to best align with the enrollment output data. Though we received enrollments from across the world, we still have maximum data in the Great Britain. Hence to improve the course to reach a wider audience, we need to model the course to more huge geographic locations. The language in which the course is modeled could be a barrier to other non-English speaking countries. To improve this, include subtitles to the course to attract countries like India, Nigeria and so on. To improve enrollments based on archetypes, we need to focus on archetypes that has less enrollment such as Preparers, flourishers and fixers. The course must also welcome more female who focus on personal life over studies. Advertising and having a separate section like "Security At Personal Life" would welcome more flourishers and fixers to take up the course. The course could also focus on people looking for work. Cybersecurity related job questionaaire with scenario based questions can be inculcated in the course to attract learners looking for work.

References

[1] Cyber Security: Safety at Home, Online, in Life (FutureLearn); Available at https://www.mooc-list.com/course/cyber-security-safety-home-online-life-futurelearn

[2] CRISP DM Process model- https://www.sciencedirect.com/science/article/pii/S1877050921002416

[3] Archetype- https://www.classcentral.com/report/what-kind-of-online-learner-are-you/