MOOC Security Learning Insights - CRISP Report

Gabriel Medina Galicia

2022-11-07

Contents

1 Document Version

Date	Author	Description
2022-11-05	Gabriel M.	Initial draft version

Date	Author	Description
2022-11-06	Gabriel M.	Adding executive summary details and business understanding text for Enrollments Process
2022-11-07	Gabriel M.	Document is now in R Markdown
2022-11-11	Gabriel M.	Adding data description report
2022-11-13	Gabriel M.	Adding data analysis and R code integration
2022-11-14	Gabriel M.	Adding executive summary details, business understanding and data analysis for Leaving Reasons and Archetypes
2022-11-14	Gabriel M.	Refining figures, tables, deployment and unit testing details

2 Disclaimer

Fictional company names and situations have been used in this assignment only for educational purposes. Newcastle University will act as the customer company that is interested in learning more about their data. MG Tech Solutions LTD will act as the contractor company who is responsible of delivering a solution to Newcastle University.

3 Introduction

The aim of this report is to collect, describe and/or present all the items produced while creating a data project following CRISP-DM methodology. CRISP-DM methodology has been used for about 20 years in many data projects across the information technology industry as well as research and development. Most of the item considerations for this document were taken from the IBM SPSS Modeler CRISP-DM Guide and from the standard reference guide.

4 Intial assumptions and considerations

Newcastle University and MG Tech Solutions LTD could be refereed in this document as the customer and the contractor respectively. This document was produced as a result of 2 business understanding, data understanding and data preparation cycles. First, the understanding and data related to enrollments; second, the understanding and data related to leaving reasons and archetypes of learners leaving.

5 Business Understanding

Newcastle University has developed a Massive Open Online Course (MOOC) entitled “Cyber Security: Safety At Home, Online, and in Life”. The customer is interested in knowing how the MOOC course has performed throughout all the runs to be able to detect areas of opportunity and increase the number of new enrollments, while preserving the quality of the course at the same time.

As part of the process, the customer would like to have a overview of the current situation on the enrollments, including demographic data of those who enrolled into the course, given the following questions:

5.1 Enrollment Analysis

- How many enrollments have we had since the beginning?

- What countries do the enrollments come from?
- What are the demographics of the learners?
- What are the proportions of learners archetypes enroll the course?
- How many enrollments are we expected to have in future runs given the current action plan?

5.2 Leaving analysis

After having understanding the previous set of questions, as part of a re-iteration of the business understanding cycle, the customer would also want to have an overview of the reasons why people has decided to leave.

- What types of learners are likely to leave?
- What are the most common reasons to leave?

5.3 Inventory of Resources

- Personnel
  - Data Scientist (1)
  - Product Owner (Report reviewer) (1)
- Data
  - We have been provided with data sets for each course run (7). For each course we have CSV files for:
    - Archetype Survey Responses (cyber-security-n_archetype-survey-responses.csv)
    - Enrollments (cyber-security-n_enrolments.csv)
    - Leaving Survey Responses (cyber-security-n_leaving-survey-responses.csv)
    - Question Responses (cyber-security-n_question-response.csv)
    - Step Activities (cyber-security-n_step-activity.csv)
    - Weekly Sentiment Survey Responses (cyber-security-n_weekly-sentiment-survey-responses.csv)
    - Course Overview (run n - Course overview - FutureLearn Course Creator.pdf)
- Computing Resources
  - Newcastle University Computer Clusters
  - 1 Personal Computer
- Technology Stack (Software)
  - R (latest distribution up to date)
  - ProjectTemplate (latest distribution up to date)
  - GGplot (latest compatible version with the latest distribution of R)
  - Tidyverse (latest compatible version with the latest distribution of R)
  - RMarkdown (latest distribution up to date)
  - RStudio (latest distribution up to date)

5.4 Requirements

An MVP (Minimal Viable Product) must by delivered by 18 November 2022 16:30 BST.

The MVP should include:

- A full detailed report including elements considered in each stage of the CRISP-DM process
- A demo presentation
- A retrospective report

5.5 Assumptions

Staff works 4 hours a day, 20 hours per week.

5.6 Constraints

- Not all students enrolled in this course have disclosed their demographic data.

- Total development hours to deliver: 60 hours (4 hours per week)

5.7 Risks and Contingencies

In the event of any issue that might delay the delivery of the MVP, the contractor staff should be in contact immediately with the customer staff to be able to negotiate a later delivery date.

5.8 Business Success Criteria

The customer should be able to easily identify the areas where improvements can be done to attract more learners to enroll, by using summarized data presented in form of simplified tables and charts.

Furthermore, provided we have data from surveys by learners archetypes and leaving reasons, the customer should be able visualize what causes the learners not to fully complete the course and, possibly, not encouraging other people to take this course, as well as identify the learning profile of leavers.

5.9 Terminology

5.9.1 MVP (Minimal Viable Product)

- A version of a new product which allows a team to collect the maximum amount of validated learning about customers with the least effort. (Lean Startup: http://www.startuplessonslearned.com/2009/08/minimum-viable-product-guide.html)
- A version of a product with just enough features to be usable by early customers who can then provide feedback for future product development. (Wikipedia: https://en.wikipedia.org/wiki/Minimum_viable_product)

5.9.2 Learner archetype

- An archetype is a personification of a collection of traits. (https://www.laurastrudwick.com/blog/2020/2/16/archetypes-of-learners)
- Archetype types:
  - Advancers: Ambitious and self-motivated to do better, progress and not stagnate.
  - Explorers: Evaluators of their options and like to inform their decisions about what to do next.
  - Fixers: They understand or manage current aspects of ther personal life
  - Flourishers: Enjoyers of self learning in order to be happy and healthy in their personal and professional lives.
  - Hobbyists: They are learn to support their existing personal projects, leisure activities and pastimes
  - Prepares: They tend to be starting out in jobs, careers or related study, having chosen what they want to do.
  - Vitalisers: Learning is ther hobbie and they do things for the love of learning

6 Data Understanding

6.1 Data Description

In the following section, a summary of the number of records plus, its data types and a sample of the records from the 7th data set is presented in form of R code output:

```
display_description_report()
```

```
## [1] "Counting the rows and colums of all datasets:"
## [1] "Dataset dimensions:  cyber.security.1_archetype.survey.responses"
## [1] 0 4
## [1] "Dataset dimensions:  cyber.security.1_enrolments"
## [1] 14394    13
```

```
## [1] "Dataset dimensions:  cyber.security.1_leaving.survey.responses"
## [1] 0 8
## [1] "Dataset dimensions:  cyber.security.1_question.response"
## [1] 77002    10
## [1] "Dataset dimensions:  cyber.security.1_step.activity"
## [1] 143092      6
## [1] "Dataset dimensions:  cyber.security.1_weekly.sentiment.survey.responses"
## [1] 0 5
## [1] "Dataset dimensions:  cyber.security.2_archetype.survey.responses"
## [1] 0 4
## [1] "Dataset dimensions:  cyber.security.2_enrolments"
## [1] 6488   13
## [1] "Dataset dimensions:  cyber.security.2_leaving.survey.responses"
## [1] 0 8
## [1] "Dataset dimensions:  cyber.security.2_question.response"
## [1] 22463    10
## [1] "Dataset dimensions:  cyber.security.2_step.activity"
## [1] 64809     6
## [1] "Dataset dimensions:  cyber.security.2_weekly.sentiment.survey.responses"
## [1] 0 5
## [1] "Dataset dimensions:  cyber.security.3_archetype.survey.responses"
## [1] 47  4
## [1] "Dataset dimensions:  cyber.security.3_enrolments"
## [1] 3361   13
## [1] "Dataset dimensions:  cyber.security.3_leaving.survey.responses"
## [1] 0 8
## [1] "Dataset dimensions:  cyber.security.3_question.response"
## [1] 16520    10
## [1] "Dataset dimensions:  cyber.security.3_step.activity"
## [1] 46614     6
## [1] "Dataset dimensions:  cyber.security.3_weekly.sentiment.survey.responses"
## [1] 0 5
## [1] "Dataset dimensions:  cyber.security.4_archetype.survey.responses"
## [1] 319   4
## [1] "Dataset dimensions:  cyber.security.4_enrolments"
## [1] 3992   13
## [1] "Dataset dimensions:  cyber.security.4_leaving.survey.responses"
## [1] 67  8
## [1] "Dataset dimensions:  cyber.security.4_question.response"
## [1] 21116    10
## [1] "Dataset dimensions:  cyber.security.4_step.activity"
## [1] 54524     6
## [1] "Dataset dimensions:  cyber.security.4_weekly.sentiment.survey.responses"
## [1] 0 5
## [1] "Dataset dimensions:  cyber.security.5_archetype.survey.responses"
## [1] 326   4
## [1] "Dataset dimensions:  cyber.security.5_enrolments"
## [1] 3544   13
## [1] "Dataset dimensions:  cyber.security.5_leaving.survey.responses"
## [1] 173   8
## [1] "Dataset dimensions:  cyber.security.5_question.response"
## [1] 18752    10
## [1] "Dataset dimensions:  cyber.security.5_step.activity"
## [1] 54257     6
```

```
## [1] "Dataset dimensions:  cyber.security.5_weekly.sentiment.survey.responses"
## [1] 1 5
## [1] "Dataset dimensions:  cyber.security.6_archetype.survey.responses"
## [1] 208   4
## [1] "Dataset dimensions:  cyber.security.6_enrolments"
## [1] 3175   13
## [1] "Dataset dimensions:  cyber.security.6_leaving.survey.responses"
## [1] 83  8
## [1] "Dataset dimensions:  cyber.security.6_question.response"
## [1] 10533    10
## [1] "Dataset dimensions:  cyber.security.6_step.activity"
## [1] 31472     6
## [1] "Dataset dimensions:  cyber.security.6_weekly.sentiment.survey.responses"
## [1] 103   5
## [1] "Dataset dimensions:  cyber.security.7_archetype.survey.responses"
## [1] 174   4
## [1] "Dataset dimensions:  cyber.security.7_enrolments"
## [1] 2342   13
## [1] "Dataset dimensions:  cyber.security.7_leaving.survey.responses"
## [1] 80  8
## [1] "Dataset dimensions:  cyber.security.7_question.response"
## [1] 10077    10
## [1] "Dataset dimensions:  cyber.security.7_step.activity"
## [1] 28304     6
## [1] "Dataset dimensions:  cyber.security.7_weekly.sentiment.survey.responses"
## [1] 77  5
## [1] "Showing some rows and data description of the 7th run:"
## [1] "Dataset:  cyber.security.7_archetype.survey.responses"
## Rows: 174
## Columns: 4
## $ id           <int> 2564612, 2574521, 2579047, 2603632, 2638826, 2754856, 287~
## $ learner_id   <chr> "732b60fc-d132-4364-b37e-0e3a5c34f346", "a45deed2-ded4-49~
## $ responded_at <chr> "2018-06-26 23:51:56 UTC", "2018-06-28 09:03:05 UTC", "20~
## $ archetype    <chr> "Other", "Fixers", "Vitalisers", "Fixers", "Fixers", "Vit~
## # A tibble: 174 x 4
##        id learner_id                           responded_at            archetype
##     <int> <chr>                                <chr>                   <chr>
## 1 2564612 732b60fc-d132-4364-b37e-0e3a5c34f346 2018-06-26 23:51:56 UTC Other
## 2 2574521 a45deed2-ded4-4979-b3dc-f1519edeba79 2018-06-28 09:03:05 UTC Fixers
## 3 2579047 04d122eb-d1b8-4c9a-bf83-3480b9bd9101 2018-06-29 05:53:28 UTC Vitalise~
## # ... with 171 more rows
## [1] "Dataset:  cyber.security.7_enrolments"
## Rows: 2,342
## Columns: 13
## $ learner_id              <chr> "f0ebc6f6-0f25-407f-a528-834414186f59", "0fa1c~
## $ enrolled_at             <chr> "2018-10-30 15:14:09 UTC", "2018-10-25 12:23:4~
## $ unenrolled_at           <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ role                    <chr> "learner", "learner", "learner", "learner", "l~
## $ fully_participated_at   <chr> "", "", "", "", "", "", "", "2018-11-01 12:05:~
## $ purchased_statement_at  <chr> "", "", "", "", "", "", "", "2018-10-09 12:11:~
## $ gender                  <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ country                 <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ age_range               <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
```

```
## $ employment_status       <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_area         <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ detected_country        <chr> "GB", "GB", "IN", "GB", "IQ", "GB", "GB", "GB"~
## # A tibble: 2,342 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 f0ebc6f6~ 2018-1~ ""      lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 2 0fa1c614~ 2018-1~ ""      lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## 3 a0ac585a~ 2018-1~ ""      lear~ ""      ""      Unkno~ Unknown Unknown Unknown
## # ... with 2,339 more rows, 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
## [1] "Dataset:  cyber.security.7_leaving.survey.responses"
## Rows: 80
## Columns: 8
## $ id                         <int> 153711, 162741, 175430, 184295, 187244, 190~
## $ learner_id                 <chr> "72669fb8-cc20-4b69-ba0a-241ff767b4de", "66~
## $ left_at                    <chr> "2018-07-06 10:55:39 UTC", "2018-07-23 01:2~
## $ leaving_reason             <chr> "Other", "Other", "The course required more~
## $ last_completed_step_at     <chr> "", "", "", "", "", "", "2018-09-10 11:02:4~
## $ last_completed_step        <dbl> NA, NA, NA, NA, NA, NA, 2.23, 1.20, NA, 1.1~
## $ last_completed_week_number <int> NA, NA, NA, NA, NA, NA, 2, 1, NA, 1, NA, NA~
## $ last_completed_step_number <int> NA, NA, NA, NA, NA, NA, 23, 2, NA, 12, NA, ~
## # A tibble: 80 x 8
##       id learner_id              left_at leavi~1 last_~2 last_~3 last_~4 last_~5
##    <int> <chr>                   <chr>   <chr>   <chr>     <dbl>   <int>   <int>
## 1 153711 72669fb8-cc20-4b69-ba0~ 2018-0~ Other   ""           NA      NA      NA
## 2 162741 662e6b45-7695-4a2e-b39~ 2018-0~ Other   ""           NA      NA      NA
## 3 175430 a3d4ef1f-5d1e-4dfe-af3~ 2018-0~ The co~ ""           NA      NA      NA
## # ... with 77 more rows, and abbreviated variable names 1: leaving_reason,
## #   2: last_completed_step_at, 3: last_completed_step,
## #   4: last_completed_week_number, 5: last_completed_step_number
## [1] "Dataset:  cyber.security.7_question.response"
## Rows: 10,077
## Columns: 10
## $ learner_id      <chr> "77454a73-6b8b-46a2-8dee-35f36b6c4fc1", "62449cd5-916b~
## $ quiz_question   <chr> "1.8.1", "1.8.1", "1.8.1", "1.8.1", "1.8.1", "1.8.1", ~
## $ question_type   <chr> "MultipleChoice", "MultipleChoice", "MultipleChoice", ~
## $ week_number     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ step_number     <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, ~
## $ question_number <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ response        <chr> "1,2,3", "1,2", "1,2", "1,2,3", "3", "1,2", "1,2,3", "~
## $ cloze_response  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ submitted_at    <chr> "2018-07-31 15:44:17 UTC", "2018-09-10 02:16:21 UTC", ~
## $ correct         <chr> "true", "false", "false", "true", "false", "false", "t~
## # A tibble: 10,077 x 10
##   learner_id     quiz_~1 quest~2 week_~3 step_~4 quest~5 respo~6 cloze~7 submi~8
##   <chr>          <chr>   <chr>     <int>   <int>   <int> <chr>   <lgl>   <chr>
## 1 77454a73-6b8b~ 1.8.1   Multip~       1       8       1 1,2,3   NA      2018-0~
## 2 62449cd5-916b~ 1.8.1   Multip~       1       8       1 1,2     NA      2018-0~
## 3 62449cd5-916b~ 1.8.1   Multip~       1       8       1 1,2     NA      2018-0~
## # ... with 10,074 more rows, 1 more variable: correct <chr>, and abbreviated
```

```
## #   variable names 1: quiz_question, 2: question_type, 3: week_number,
## #   4: step_number, 5: question_number, 6: response, 7: cloze_response,
## #   8: submitted_at
## [1] "Dataset:  cyber.security.7_step.activity"
## Rows: 28,304
## Columns: 6
## $ learner_id        <chr> "77454a73-6b8b-46a2-8dee-35f36b6c4fc1", "20e6ec35-0f~
## $ step              <dbl> 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.~
## $ week_number       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ step_number       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ first_visited_at  <chr> "2018-08-10 08:39:26 UTC", "2018-09-05 13:57:38 UTC"~
## $ last_completed_at <chr> "", "", "2018-09-10 00:53:16 UTC", "", "", "2018-09-~
## # A tibble: 28,304 x 6
##   learner_id                            step week_number step_~1 first~2 last_~3
##   <chr>                                <dbl>       <int>   <int> <chr>   <chr>
## 1 77454a73-6b8b-46a2-8dee-35f36b6c4fc1   1.1           1       1 2018-0~ ""
## 2 20e6ec35-0f50-4819-9c2e-d1851fd54638   1.1           1       1 2018-0~ ""
## 3 62449cd5-916b-46a6-9710-441b68d2199f   1.1           1       1 2018-0~ "2018-~
## # ... with 28,301 more rows, and abbreviated variable names 1: step_number,
## #   2: first_visited_at, 3: last_completed_at
## [1] "Dataset:  cyber.security.7_weekly.sentiment.survey.responses"
## Rows: 77
## Columns: 5
## $ id                <int> 60491, 60882, 61034, 61062, 61746, 61818, 61981, 625~
## $ responded_at      <chr> "2018-09-10 10:50:47 UTC", "2018-09-11 07:16:55 UTC"~
## $ week_number       <int> 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 2, 2~
## $ experience_rating <int> 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3~
## $ reason            <chr> "", "Im paranoid about my online privacy, this week ~
## # A tibble: 77 x 5
##      id responded_at            week_number experience_rating reason
##   <int> <chr>                        <int>             <int> <chr>
## 1 60491 2018-09-10 10:50:47 UTC          1                 3 ""
## 2 60882 2018-09-11 07:16:55 UTC          1                 3 "Im paranoid abou~
## 3 61034 2018-09-11 13:42:18 UTC          1                 3 ""
## # ... with 74 more rows
```

We decided to only include the 7th data run in the description report as, based on complete data explorations,surveys appear to be recorded from the 5th run onwards.

6.2 Data Quality

6.2.1 Missing Data

For summary and report purposes, a country code datafile has been added to our data repository. Source: https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv.

NA and null values depending on what information they represent will be treated as “Unknown”, “Undetermined” or omitted from statistical reports.

As mention in Data Description section, not all the runs recorded survey data from the beginning. As such, only the runs that include data will be considered.

Table 2: Enrollments over all runs

Run Number	Enrolments
1	14394
2	6488
4	3992
5	3544
3	3361
6	3175
7	2342

6.3 Initial Data Exploration

7 Data Preparation

Not all data from the provided sources are going to be selected for the purposes of our analysis. The following list includes the data sets considered for our analysis:

- Enrollments
  - All runs have been condensed in one single data set per category, adding one column "run_nums" that identifies the run it belongs to, so that we could analyse and explore data considering all runs.
  - Categorical values that are in snake case are converted to pascal case.
- Archetype Survey Responses

7.1 Exploratory Data Analysis

7.1.1 Enrollment

```
df <- get_enrolments_count_per_run(enrolments_df)
df <- df %>% rename( c(Run.Number = run_num, Enrolments = n) )
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Enrollments over all runs")
```

```
p <- get_plot_enrolments_count_per_run(enrolments_df)
p
```

7.1.1.1 How many enrollments have we had since the beginning? As we can see from the previously presented table 2 and in figure 1, the first run has a considerable amount of enrollments, probably due to it being the first run.

```
get_plot_enrolements_count_per_country(enrolments_df)
```

7.1.1.2 What countries do the enrollments come from? As Newcastle University is an institution in the United Kingdom, it is evident that most enrollments come from this country as we can see in the chart from figure 2.

7.1.2 Learners demographics

```
df <- get_enrolments_count_per_age(enrolments_df)
df <- df %>% rename( c(Age.Range = age_range, Enrolments = n) )
```

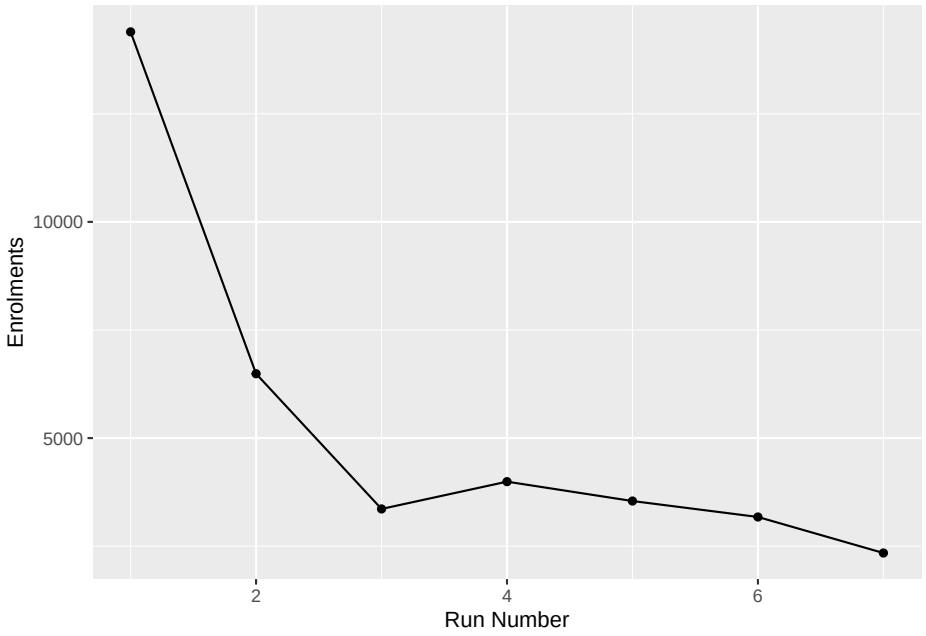

Figure 1: Enrollments over all runs

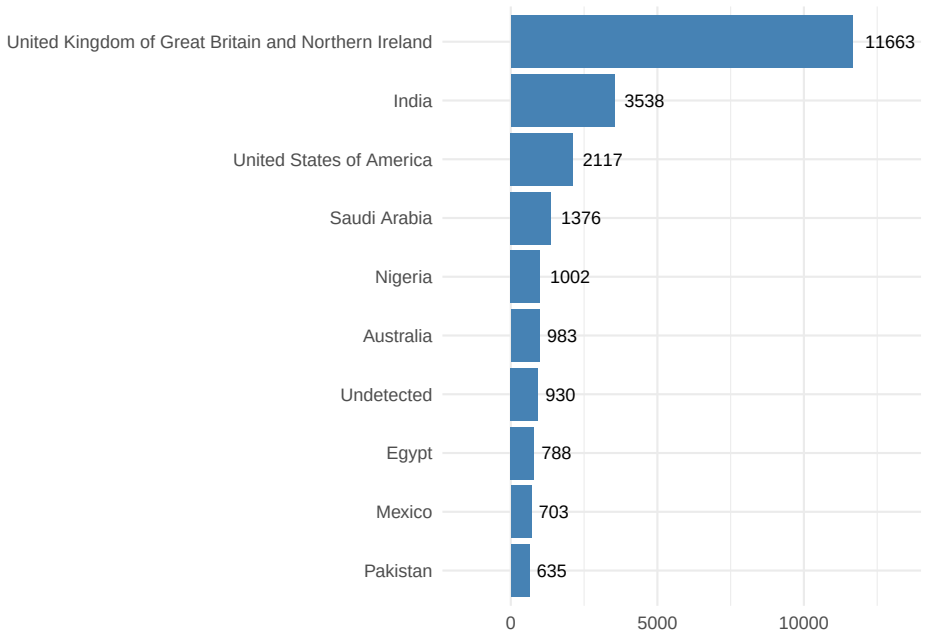

Figure 2: Total Enrollments per Country

Table 3: Total Enrollments per Age Range

Age Range	Enrolments
26-35	875
18-25	691
56-65	612
36-45	608
46-55	602
>65	598
<18	42

```
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Total Enrollments per Age Range")
```

```
get_plot_enrolments_count_per_age(enrolments_df)
```

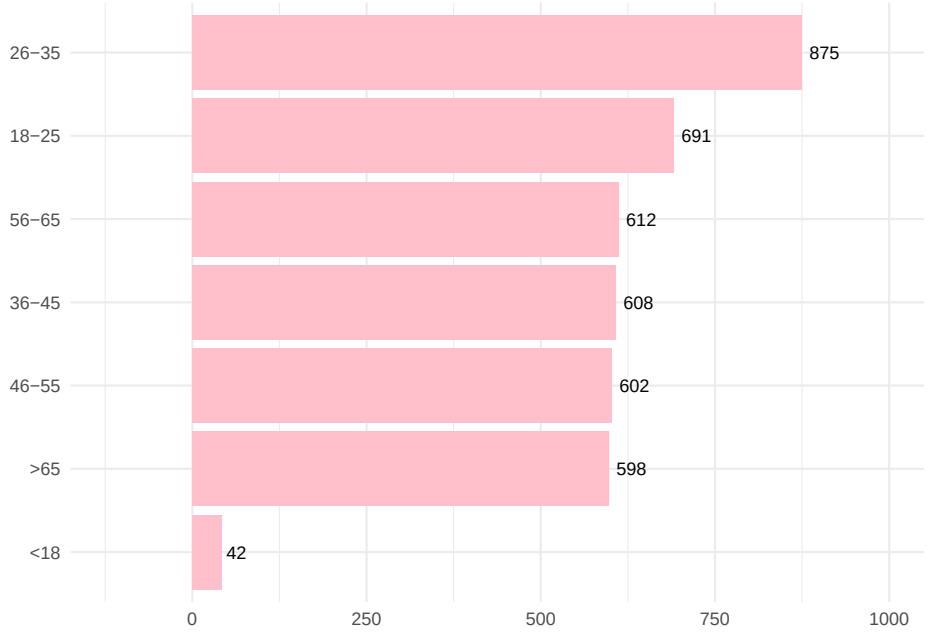

Figure 3: Total Enrollments per Age Range

7.1.2.1 Ages We can observe that in figure 3 and table 3 the most enthusiastic age group to be enrolled was 25-35.

```
df <- get_enrolments_count_per_edlvl(enrolments_df)
df <- df %>% rename( c(Highest.Education.Level = highest_education_level, Enrolments = n) )
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Total Enrollments per Highest Education Level")
```

```
get_plot_enrolments_count_per_edlvl(enrolments_df)
```

Table 4: Total Enrollments per Highest Education Level

Highest Education Level	Enrolments
University Degree	1729
University Masters	852
Secondary	605
Tertiary	361
Professional	339
University Doctorate	146
Less than Secondary	88
Apprenticeship	15

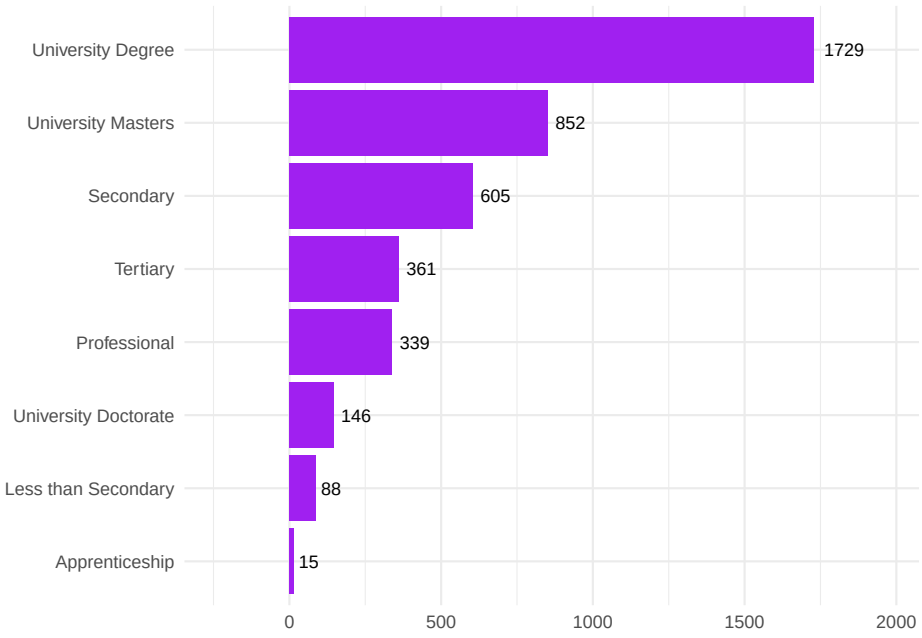

Figure 4: Total Enrollments per Highest Education Level

Table 5: Total Enrollments per Employment Status

Employment Status	Enrolments
Working Full Time	1402
Retired	717
Full Time Student	462
Self Employed	394
Looking for Work	373
Working Part Time	356
Not Working	230
Unemployed	171

7.1.2.2 Highest Education Level People with University degrees are an important proportion of the enrollments, followed by University masters as it can be observed in figure **??** and table 4.

```
df <- get_enrolments_count_per_employment_status(enrolments_df)
df <- df %>% rename( c(Employment.Status = employment_status, Enrolments = n) )
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Total Enrollments per Employment Status")
```

```
get_plot_enrolments_count_per_employment_status(enrolments_df)
```

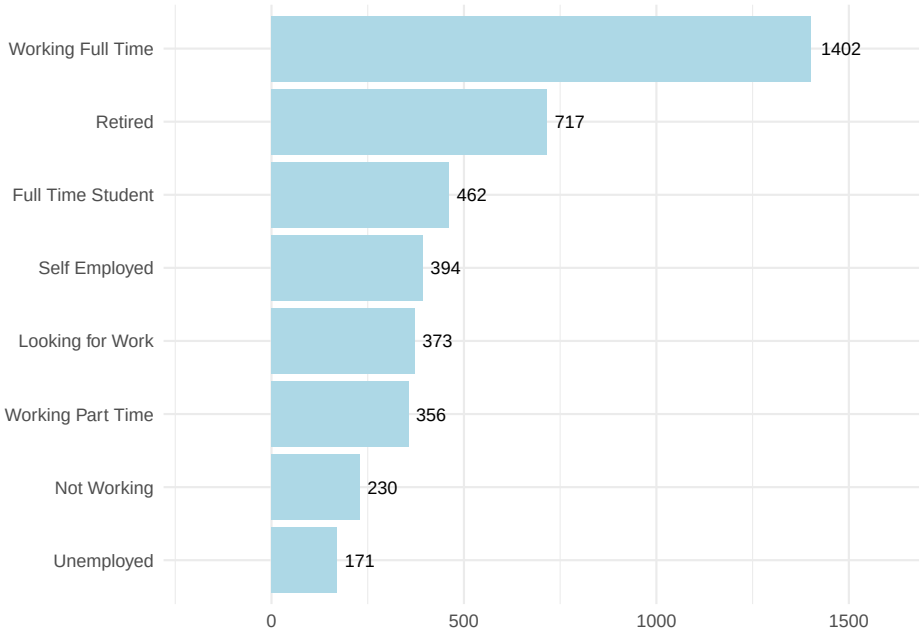

Figure 5: Total Enrollments per Employment Area

7.1.2.3 Employment Status From in figure 5 and table 5, we can see that those who are working full time make are the ones more attracted to be enrolled in this course.

Table 6: Total Enrollments per Employment Status

Employment Area	Enrolments
It and Information Services	765
Teaching and Education	485
Engineering and Manufacturing	244
Health and Social Care	208
Public Sector	207
Business Consulting and Management	186
Accountancy Banking and Finance	136
Charities and Voluntary Work	135
Law	111
Creative Arts and Culture	97
Retail and Sales	88
Marketing Advertising and Pr	80
Science and Pharmaceuticals	73
Media and Publishing	71
Hospitality Tourism and Sport	58
Transport and Logistics	57
Armed Forces and Emergency Services	55
Energy and Utilities	44
Environment and Agriculture	42
Property and Construction	41
Recruitment and Pr	23

```
df <- get_enrolments_count_per_employment_area(enrolments_df)
df <- df %>% rename( c(Employment.Area = employment_area, Enrolments = n) )
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Total Enrollments per Employment Status")
```

```
get_plot_enrolments_count_per_employment_area(enrolments_df)
```

7.1.2.4 Employment Area Not surprisingly, Information Technology is the predominant area, as this course provides tools and information that is crucial for this particular sector, as represented in figure 6 and table 6.

7.2 Leaving Surveys

Learners archetypes leaving

```
df <- get_archetypes_count_from_survey_responses(archetype_survey_df)
df <- df %>% rename( c(Archetype = archetype, Leaving.Count = n) )
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Total Leavings per Archetypes")
```

```
get_plot_archetypes_count_from_survey_responses(archetype_survey_df)
```

From table **??** and figure **??**, we found that vitalisers were the group that leave the most from the total runs.

```
df <- get_leaving_reasons_count(leaving_survey_df)
df <- df %>% rename( c(Leaving.Reason = leaving_reason, Leaving.Count = n) )
```

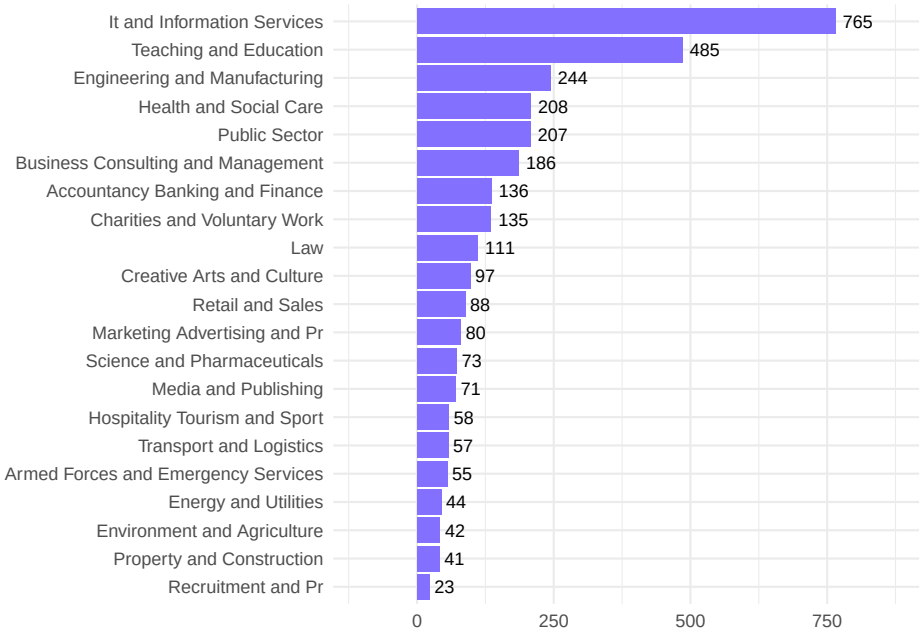

Figure 6: Total Enrollments per Employment Area

Table 7: Total Leavings per Archetypes

Archetype	Leaving Count
Vitalisers	237
Explorers	208
Advancers	172
Fixers	136
Hobbyists	96
Other	94
Preparers	88
Flourishers	43

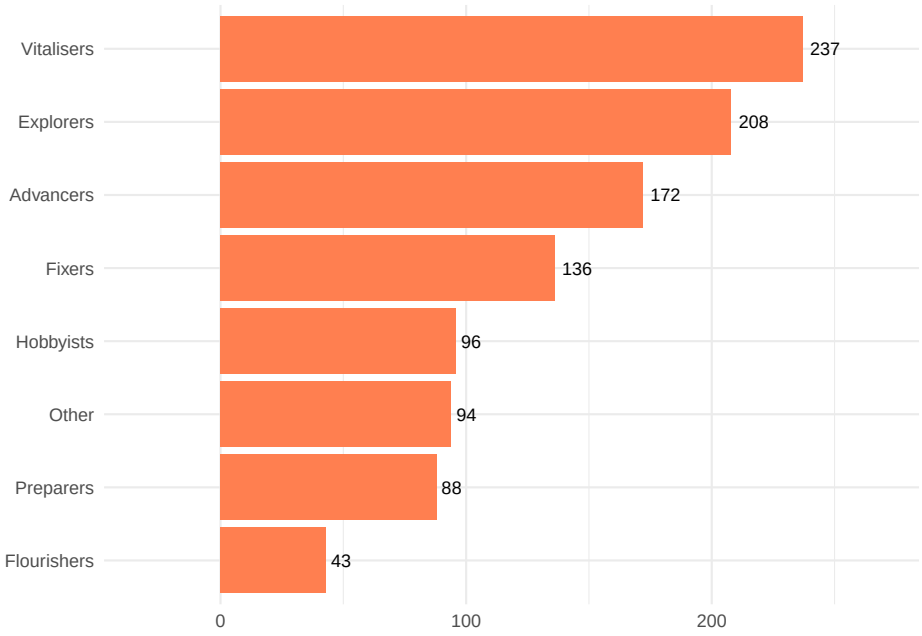

Figure 7: Total Leavings per Archetypes

Table 8: Leaving Reasons for all Runs

Leaving Reason	Leaving Count
I don't have enough time	103
Other	97
I prefer not to say	47
The course required more time than I realised	40
The course wasn't what I expected	36
The course won't help me reach my goals	36
The course was too hard	26
The course was too easy	18

```
knitr::kable(df, col.names = gsub("[.]", " ", names(df)),
             caption = "Leaving Reasons for all Runs")
```

```
get_plot_leaving_reasons_count(leaving_survey_df)
```

7.2.0.1 Leaving reasons It can be seen in **??** and figure **??**, that the most common leaving reason was that learners do not have enough time to complete or follow the pace of the course.

7.3 Modeling

The aim of this section is to present and summaries of any data mining or machine learning modeling analysis.

As for the initial efforts for the enrollment analysis, the customer is interested in knowing how many enrollments can they expect in the following runs given their current conditions (teaching techniques, marketing campaigns, etc.)

Based on our previous analysis, we can use the data from the enrollments per run to try to predict the

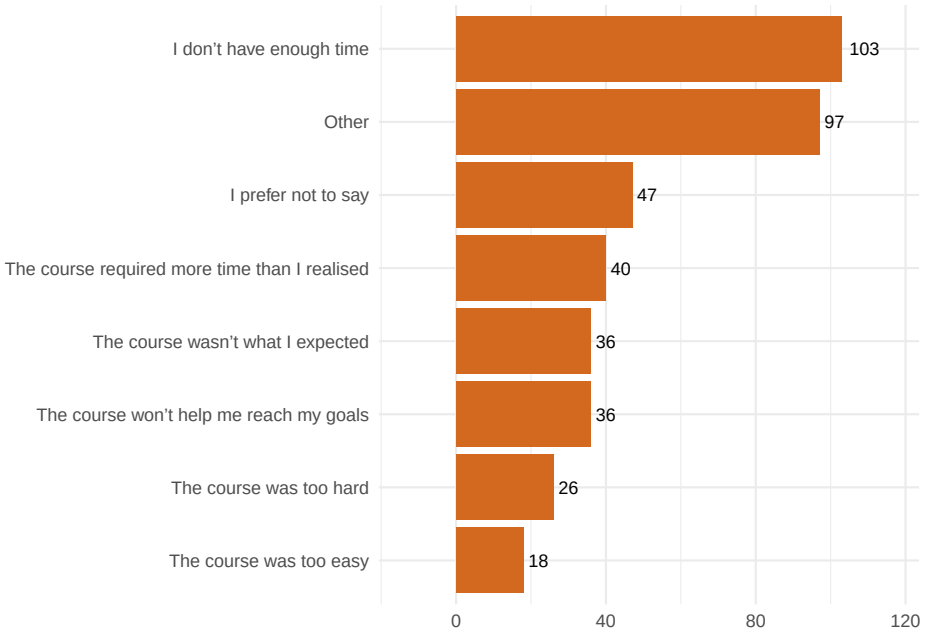

Figure 8: Leaving Reasons for all Runs

following number of runs in future project iteration. For this current deployment cycle, modeling is not currently a required step to obtain a deliverable product. In future iterations, we should consider that the customer would like us to do more research in this particular topic, particularly when looking at **??** and in figure 2.

8 Evaluation

We consider that the customer can get a reasonable insight of the current course situation by having explored and obtained run-condensed results from the first questions related to enrollments. The customer could then evaluate their marketing targets and make the necessary efforts in order to increase enrollments and stop the declining trend. Furthermore, there can also detect that there are there is a significant difference between the United Kingdom and the rest of the countries. In case in future runs the enrollment trend changes, we can create new analysis instances based on data from new runs.

9 Deployment

As for now, as there are no models that need to be deployed. However, we should consider this report itself and the data manipulation and analysis coding as a deployable units, all of them organized into a deployable unit using ProjectTemplate.

ProjectTemplate is an open source opinionated tool that allow us to set up a data science project by providing templates and organizing files in predefined directories, allowing the user to easily include data, set processing scripts in various steps and perform code profiling and unit testing.

To be able to generate a new version of this report, the one who decides to publish a new version should execute a pipeline that includes the generation (knitting) of this PDF document. The deployment execution pipeline using continuous deployment and continuous integration is planned to be implemented in future iterations.

As for now, the project directory should be loaded into RStudio, set as the working directory, have the

"report.rmd" file opened and proceed to execute the "Knit" command to generate a PDF in the "reports" directory.

Unit Testing can be run using "test.project()" command in RStudio, or in case this wants to be included in a CI/CD pipeline, this must be integrated as a command being run before the deployment of the document takes places.

9.1 Development phases and repository branching strategy

Code units, unit testing and documentation is all condensed in a single GIT repository that contains the root of a ProjectTemplate project.

For internal purposes, we used GIT to keep track of changes. The repository can be seen in the following URL: https://github.com/gmedina-mx/cs8631-datamngmt-ncl-final-assignment

As this project was only managed by one person of our staff, no branching strategy was followed. Once this project gets to production, the following GIT branching strategy should be used:

- **master**: Holds the final production deliverable
- **uat**: Deliverable ready for User Acceptance Test
- **dev**: Stable branch for developers and data scientist with latest code and report changes.
- Each developer or data scientist should create their own branch from dev to keep track of their own feature (i.e. dev_gabriel, dev_gabriel_regularization, etc) which will be later reviewed by other developers or data scientists via Pull Requests to dev branch.