CSC8631 Report

Morgan Frodsham

02/12/2021

FutureLearn Learning Analytics for Newcastle University

Business understanding

Background

This report details the process and findings of exploratory data analysis for Newcastle University's FutureLearn course called "Cyber Security: Safety at Home, Online, in Life." Data analysis of educational platforms, such as FutureLearn, is a key part of learning analytics, which is defined as *"the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs."* (*1st International Conference on Learning Analytics and Knowledge* (2011)) Learning analytics has the potential to help educational institutions undertake data-driven decision-making (Long and Siemens (2011)) that better support students through their educational journey (Shacklock (2016)).

Business objectives and success criteria

FutureLearn is a massive open online course (MOOC) provider. It promotes itself as *"a powerful new way to learn online"* (FutureLearn (n.d.)), designed with the principles of effective learning. Learning effectiveness can be indicated by learner engagement, their success rate, and completion time (S. Hubalovsky and Musilek (2018)). A core objective of FutureLearn is to improve learning effectiveness.

The objective of the exploratory data analysis outlined by this report is to better understand the learning effectiveness of Newcastle University's Future Learn Cyber Security course. The process and findings described by this report explore the question: "Do learners have a higher chance of success the more data they share with FutureLearn?" Here, "a higher chance of success" means that learners are more likely to complete the course. The data shared by learners refers to the optional data requested by FutureLearn; specifically, the provision of personal identifiers like demographic information. The provision of this data is taken to be an indicator of learner engagement with the course. For this analysis the business success criteria is to give useful insights into the relationship between learner engagement and learner success rate.

Other objectives that were considered included:

- Do learners complete the course more quickly the more data they share with FutureLearn?
- From which countries are learners more likely to engage with the course?
- Do learners have a higher chance of success the more times they interact with the future learn platform?

These objectives have not been pursued by this project because they are concerned with later stages of the learners' journey through the course. Instead, this project focused on the impact of learners initial engagement with FutureLearn on their completion of the course, and therefore their ultimate success.

Costs and benefits

This is a low cost project because it is not currently resource intensive (as outlined by the resource inventory below). If the analysis and findings outlined by this report hold true, there could be significant business benefits. For example, it could help Newcastle University and FutureLearn support learners to succeed, increasing learner satisfaction and, if desired, could lead to increase of learners for both organisations.

Project plan

Started on 8 November 2021, the deadline for this project is 16:30, 3 December 2021. The first iteration, between 8 and 18 November, focused on project set up, understanding what data was provided, learning R and practicing with the technical suite. The second and third iteration are highly dependent on the success of the first iteration. The second iteration, between 19 and 25 November, focused on tidying the data, exploring potential routes of analysis and setting up the ProjectTemplate configurations for this project. The third iteration, between 29 November and 3 December focused on deeper exploratory data analysis and creating project outputs, such as the report and graphs desired.

The resources required for this project, as well as inputs (resources, tools and techniques) and outputs, are outlined in other sections of this report.

Terminology

The terminology used throughout this report includes:

- R, which is a programming language for statistical computing and graphics.
- RStudio, which is an Interated Development Environment for R.
- R packages, which are extensions to the R statistcial programming language. R packages contain code, data, and documentation in a standardised collection format that can be installed by users of R. The R packages used by this report are listed in the section below.
- Git, which is software for tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code. GitHub is a provider of internet hosting for software development and version control using Git. Git Bash is an application for Microsoft Windows environments which provides an emulation layer for a Git command line experience.
- Tibble, which is a data frame that stores data in R and appears as a table.
- Success rate, which is the number of learners who have completed the course divided by the total number of learners as a percentage.

Inventory of resources

This project has been undertaken by one part-time person, on a Dell G7 17" laptop. Git, RStudio and a selection of R packages (detailed in the next section) have been used to deliver the project. Newcastle University also provided data from its FutureLearn cyber security course for analysis.

Initial assessment of tools and techniques

This project utilises the following tools and techniques:

- CRISP-DM (Chapman et. al. and Wirth (2000)) is the methodology used to deliver data workflow best practice;

- Tidyverse for R is the R package used for the analysis;
- ggplot2 for R is the R package used for visualisations;
- RColorBrewer is the R package used to provide accessible colour palettes for visualisations;
- ProjectTemplate is the R package used to improve the reproducibility of this analysis;
- Git (specifically GitHub and GitBash) is used for version control;
- all written documentation is created with the R package, RMarkdown; and
- natbib is an R package used for the bibliography.

Requirements, assumptions and constraints

This project, as well as the process and findings outlined in this report, is limited to the data analysis pipeline. The modelling section of CRISP-DM presents the graphical outputs of the exploratory data analysis as creating a model is beyond the scope of this project. Consequently, there is no deployment section in this report either. The type of project results to be expected are initial findings from exploratory analysis, graphical summaries of these findings and explanatory text.

To conduct the analysis outlined by this report, it is assumed that the data is accurate. Where there is a missing value (often represented as NA), it is assumed that this data value does not apply to the learner or the learner did not provide FutureLearn with that type of data. It is assumed that the data vales of fully_participated_at (inside the enrolments data set) demonstrate whether learners have completed the course or not. If this assumption is incorrect, it will have a significant impact on the project findings. It is also assumed that it is optional for the learners to complete the demographic data (inside the enrolments data set, like gender or country, or the archetype data set) as there is a high number of missing values. If this assumption is incorrect, the data should not be interpreted as a proxy of learner engagement.

Risks and contingencies

Due to the limited nature of this project, there is a strong risk that there are confounding variables, which could cause a spurious association. More comprehensive analysis of all the data sets would be required to determine this and provde the correlations uncovered in the analysis findings. Therefore, all findings outlined by this report are to be interpreted with caution.

Similarly, the assumptions outlined in the section above pose a risk to this project. Regular communication with key Newcastle University staff members has helped to mitigate this risk. However, it should be noted that FutureLearn has not been contacted for this report. This will be vital to mitigating the risk if further analysis and modelling is undertaken beyond this project.

This project has also attempted to address common programming risks such as unrepeatable code and opaque decision-making by using ProjectTemplate and Git for current best practice in programming transparency and reproducibility.

Data mining goals and success criteria

The exploratory data analysis outlined by this report explores the question: "Do learners have a higher chance of success the more data they share with FutureLearn?" To understand this, the data mining aims to determine:

1. What is the gross success rate of learners? The data mining success criteria is the percentage of all learners who complete the course.

2. How many learners provide any demographic data? The data mining success criteria is the percentage of all learners who provide optional demographic data such as gender, age range or employment status from the enrolments data set.
3. What is the success rate of learners who provide demographic data? The data mining success criteria is the percentage of learners who complete the course and provide demographic data (as outlined in point 2).
4. What is the success rate of learners who do not provide demographic data? The data mining success criteria is the percentage of learners who complete the course and do not provide any demographic data (as outlined in point 2).
5. Is there any correlation between learners' success and the provision of demographic data? The data mining success criteria is the difference between the success rates od learners who provided or did not provide demographic data (using the findings from point 3 and 4).

Data understanding

Initial data collection

Newcastle University's FutureLearn cyber security course has run seven times. The first run provided six different data sets, the second run provided seven, and the others provided eight. There are 53 different data sets for potential exploration. In total, 53 data sets have been provided by Newcastle University.

It is unclear whether there have been any problems in data acquisition or extraction as Newcastle University has provided the data.

Data description

The first run of Newcastle University's FutureLearn cyber security course provides the following data sets:

- archetype survey responses (id, learner_id, responded_at, archetype);
- enrolments (learner_id, enrolled_at, unenrolled_at role, fully_participated_at, purchased_statement_at gender, country age_range, highest_education_level, employment_status, employment_area detected_country);
- leaving survey responses (id, learner_id, left_at leaving_reason, last_completed_step_at, last_completed_step, last_completed_week_number, last_completed_step_number);
- question responses (learner_id, quiz_question, question_type, week_number, step_number, question_number, response, cloze_response, submitted_at, correct);
- step activity (learner_id step, week_number step_number, first_visited_at, last_completed_at); and
- weekly sentiment survey (id responded_at, week_number, experience_rating, reason).

The second run provides the same data sets as well as a data sets specifying:

- team members (id, first_name, last_name, team_role, user_role).

The other five runs provide the same data sets as the second run as well as a data set on:

- video stats (step_position, title video_duration, total_views, total_downloads, total_caption_views, total_transcript_views, viewed_hd, viewed_five_percent, viewed_ten_percent, viewed_twentyfive_percent, viewed_fifty_percent, viewed_seventyfive_percent, viewed_ninetyfive_percent, viewed_onehundred_percent, console_device_percentage, desktop_device_percentage, mobile_device_percentage, tv_device_percentage, tablet_device_percentage, unknown_device_percentage, europe_views_percentage, oceania_views_percentage, asia_views_percentage, north_america_views_percentage, south_america_views_percentage, africa_views_percentage, antarctica_views_percentage).

Data quality

The data provided by Newcastle University is raw, and so its accuracy is assumed. The completeness of the data varies by the run of the course; for example, the data collected by the first run of the course is relatively sparse compared to the seventh run of the course. In runs three and four more data is being collected; this could be because there are fewer optional sections (such as the archetype or leaving survey), but it is unknown why this change has occurred. There are also missing values in data sets where learners have not completed the course (or particular parts of it) and when learners have not provided optional data (as outlined above).

Data exploration

To address the business objective and success criteria, the exploratory analysis outlined by this report uses the enrolments data set because it includes the field fully_participated_at, which is considered to mean course completion, as well as learners' demographic data. This section is structured by the data mining goals.

Goal 1 (Gross Success Rate)

To understand the success rate of all learners who undertake the course, the total number of learners who have undertaken the course across all seven runs needs to be determined. All seven runs were combined, two filters were created to show how many learners completed (complete) and did not complete the course (incomplete).

```
# Identify how many learners there are in all seven runs.
total_learners <- (2154+35142) %>% # Adding complete and incomplete totals.
  print()
```

```
## [1] 37296
```

```
# Calculate the success rate of all learners who enrolled in the course.
SR_gross <- (100*(2154/total_learners)) %>%
  round(., digits = 2) %>%
  print()
```

```
## [1] 5.78
```

This allows us to determine that there are 37,296 learners who have enrolled in Newcastle University's FutureLearn cyber security course, and 5.78% of them have completed it, which is the gross success rate.

Goal 2 (Demographic Data)

For the enrolments data set, demographic data refers to the following fields that learners could provide:

- gender,

- country,
- age_range,
- highest_education_level,
- employment_status, and
- employment_area.

To investigate whether learners had provided any of this demographic data, a data set called 'FL1' was created with new columns (that corresponded to the aforementioned demographic data) showing TRUE or FALSE for whether the learner had shared that type of demographic data.

```
# Create a new column called "count" in FL1 containing the value (count of
# columns) where there is a a vale != FALSE.
true_counts <- FL1 %>%
  mutate(num_true = rowSums(.[ids_of_declared_cols] != FALSE)) # Count up how
# many of the columns have a value != FALSE.

# Count the number of demographic columns filled in by the learner who completed
# and did not complete the course.
learner_counts <- true_counts %>%
  group_by(num_true, completed) %>%
  count()

# Display a tibble of learners who have completed or not completed the course,
# counting number of demographic columns filled in by the learner.
learner_counts %>%
  print(ids_of_declared_cols, true_counts, n = 15, width = Inf)
```

```
## # A tibble: 15 x 3
## # Groups:   num_true, completed [15]
##    num_true completed     n
##       <dbl> <lgl>     <int>
##  1        0 FALSE     31467
##  2        0 TRUE       1626
##  3        1 FALSE        14
##  4        2 FALSE        15
##  5        2 TRUE          4
##  6        3 FALSE        31
##  7        3 TRUE          9
##  8        4 FALSE        79
##  9        4 TRUE          7
## 10        5 FALSE       858
## 11        5 TRUE        134
## 12        6 FALSE      2676
## 13        6 TRUE        372
## 14       NA FALSE         2
## 15       NA TRUE          2
```

This table shows us how many learners who have completed (TRUE) or not completed (FALSE) the course have provided demographic data and, if so, how many demographic data fields they have provided.

There are four instances of NA in the table that should be investigated further; for now, it is assumed that these learners did not provide demographic data.

Below shows that 4,199 learners provided demographic data.

```
# Calculate the number of learners who provided demographic data
d_count <- (37296 - 31467 - 1626 - 4) %>%
  print()
```

```
## [1] 4199
```

Goal 3 (Demographic Data Success Rate)

The table in Goal 2 shows us how many learners provided demographic data and completed the course. This enables us to calculate the success rate of learners who have provided demographic data.

```
# Calculate how many learners provided demographic data.
d_complete_count <- (4+9+7+134+372) %>%
  print ()
```

```
## [1] 526
```

```
# Calculate the success rate of learners who provided demographic data.
SR_d <- (100*(as.numeric(d_complete_count))/(as.numeric(d_count))) %>%
  round(., digits = 2) %>%
  print()
```

```
## [1] 12.53
```

This allows us to determine that of the 4,199 learners who provided demographic data, 526 completed the course. The success rate of learners who provide demographic data is 12.53%.

Goal 4 (No Demographic Data Success Rate)

The table created as part of Goal 2 shows how many learners did not complete the course and did not provide demographic data.

```
# Calculate how many learners did not provide demographic data.
no_d_count <- (31467+1626+2+2) %>%
  print()
```

```
## [1] 33097
```

```
# Calculate how many of those learners completed the course.
no_d_complete_count <- (1626+2) %>%
  print()
```

```
## [1] 1628
```

```
# Calculate success rate of learners who did not provide demographic data.
SR_no_d <- (100*(as.numeric(no_d_complete_count))/(as.numeric(no_d_count))) %>%
  round(., digits = 2) %>%
  print()
```

```
## [1] 4.92
```

This allows us to determine that of the 33,097 learners who did not provide demographic data, 1,628 learners did not complete the course. The success rate of learners who do not provide demographic data is 4.92%.

Goal 5 (Potential Correlation)

We have identified three different success rates for learners:

```
## # A tibble: 3 x 2
##   Success_Rate        Percentage
##   <chr>                    <dbl>
## 1 Gross                     5.78
## 2 No Demographic Data       4.92
## 3 Demographic Data         12.5
```

It appears there may be a correlation between learners' success rate and their provision of demographic data (as a proxy for learner engagement); learners who provide demographic data are over twice as likely (2.55x) to complete the course than learners who do not provide demographic data.

Business understanding (second iteration)

Business objectives and success criteria

Our core objective is to determine whether learners have a higher chance of success the more data they share with FutureLearn. An initial investigation into the data provided by Newcastle University suggests that learners do have a higher chance of success the more data they share with FutureLearn. However, the business success criteria is to give useful insights into the relationship between learner engagement and learner success rate. A better understanding of the correlation between learners' success rate and their provision of demographic data is therefore needed.

Data mining goals and success criteria

To better understand the relationship between learners' engagement and success rate. The gender, age, level of education and employment status of learners have been chosen as the demographic data fields for further exploration as there are fewer categories in each field by comparison to country and employment status. The second iteration of data mining aims to determine:

2.1. Which genders declared by learners have a higher chance of success? The data mining success criteria is the percentage of learners who have completed the course for each gender declared.

2.2. Which ages declared by learners have a higher chance of success? The data mining success criteria is the percentage of learners who have completed the course for each age range declared.

2.3. Which levels of education declared by the learners have a higher chance of success? The data mining success criteria is the percentage of learners who have completed the course for each age range declared.

2.4. Which employment statuses declared by learners have a higher chance of success? The data mining success criteria is the percentage of learners who have completed the course for each employment status declared.

Data understanding (second iteration)

The details of data collection, description and quality outlined in the first iteration of data understanding also apply to this second iteration.

Data exploration

This second iteration of data exploration also uses the enrolments data set, with the same caveats as the first iteration. This section is also structured by the data mining goals.

Goal 2.1 (Gender Success Rate)

Using true_counts from Goal 2 in the first iteration of data exploration we can determine that 4159 learners declared their gender to FutureLearn.

```
# Identify which genders completed or did not complete the course.
gender_counts <- true_counts %>%
  group_by(num_true, gender, completed) %>%
  count() %>%
  filter(!(gender == "Unknown"))

# Total learners who declared their gender.
total_gender <- sum(gender_counts$n) %>%
  print()
```

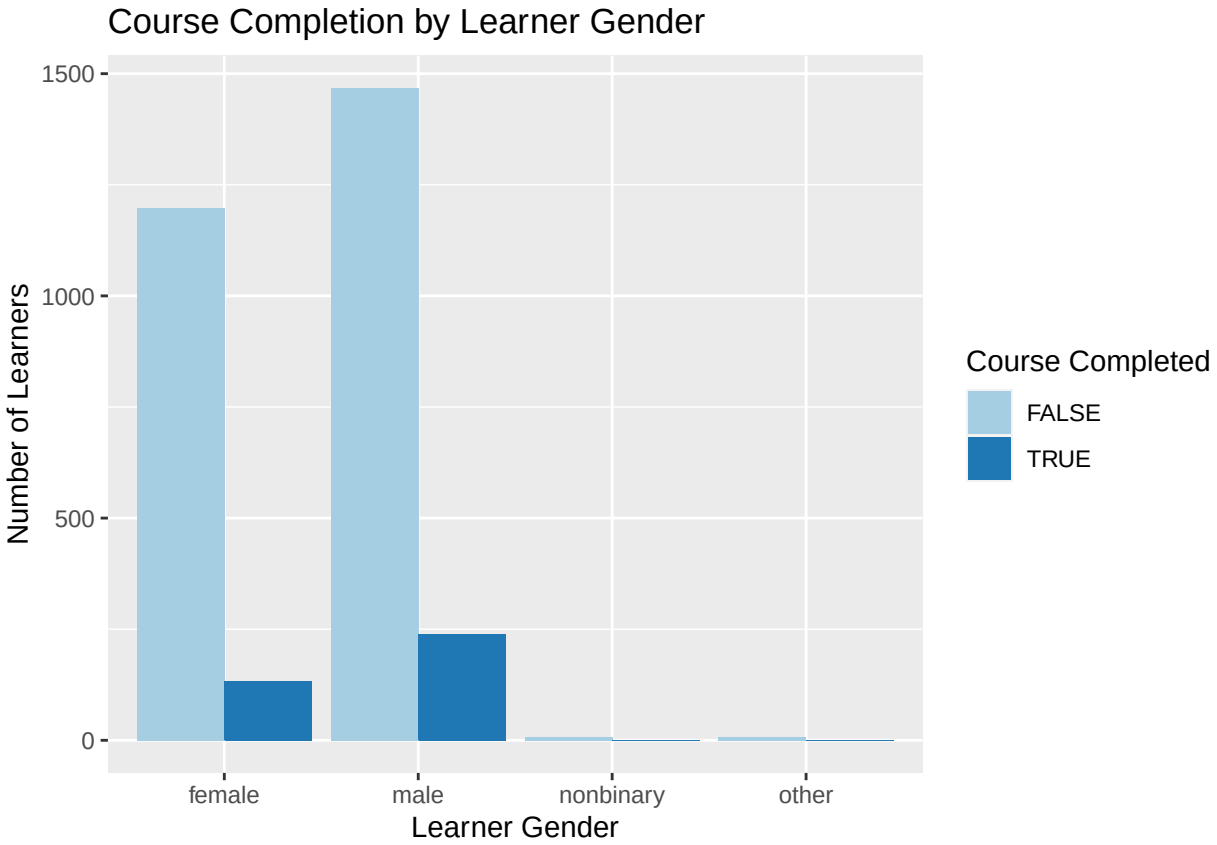
```
## [1] 4159
```

```
# Total learners who declared their gender and completed the course.
c_total_gender <- (3 + 1 + 1 + 4 + 6 + 60 + 60 + 1 + 132 + 239 + 1 + 2 )

# Calculate success rate of those who provided their gender.
SR_total_gender <- (100*(c_total_gender/total_gender)) %>%
  round(., digits = 2) %>%
  print()
```

```
## [1] 12.26
```

This allows us to determine that the success rate of learners who declare their gender is 12.26%. The graph below was created to understand whether the success rate varies for different genders declared by learners.



This graph indicates that those who identify as male may be more likely to complete the course. This is unclear, however, because there seems to be a proportionate number of learners who identify as male, but do not complete the course.

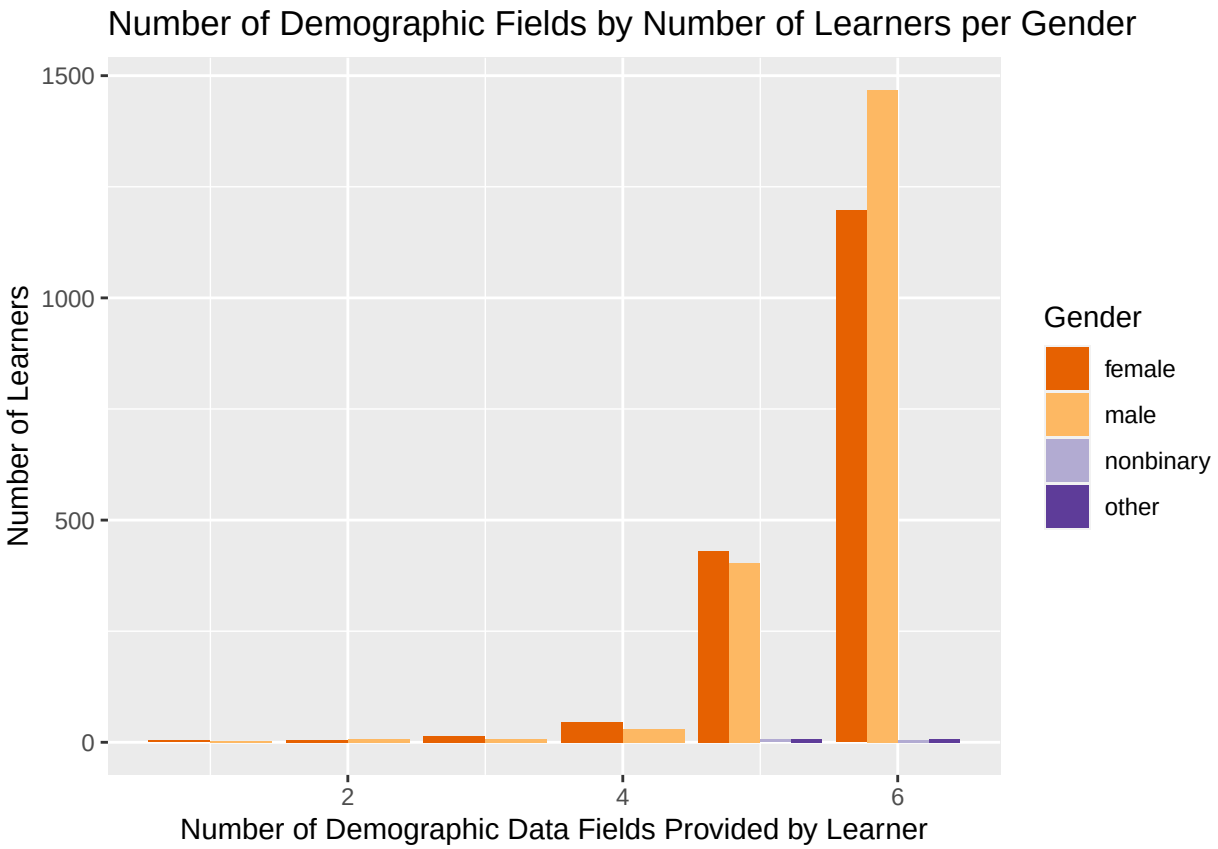
To better understand the results of this graph, the table below was created. The table was created through the same process used to create the gender success rate (as above) by applying it to each declared gender (the code can be found in the file `eda2.R` located in the `src` folder).

```
## # A tibble: 5 x 2
##   Success_Rate Percentage
##   <chr>              <dbl>
## 1 Gender             12.3
## 2 Male               14.0
## 3 Female             10.7
## 4 Nonbinary           8.33
## 5 Other               7.14
```

The table shows us that of learners who declared their gender, those who identified as male had a higher chance of success (14.02%) than people who identified as another gender. Those who identified as other had the lowest chance of success (7.14%).

The graph below was created to better understand the provision of demographic data by learners who declared their gender.

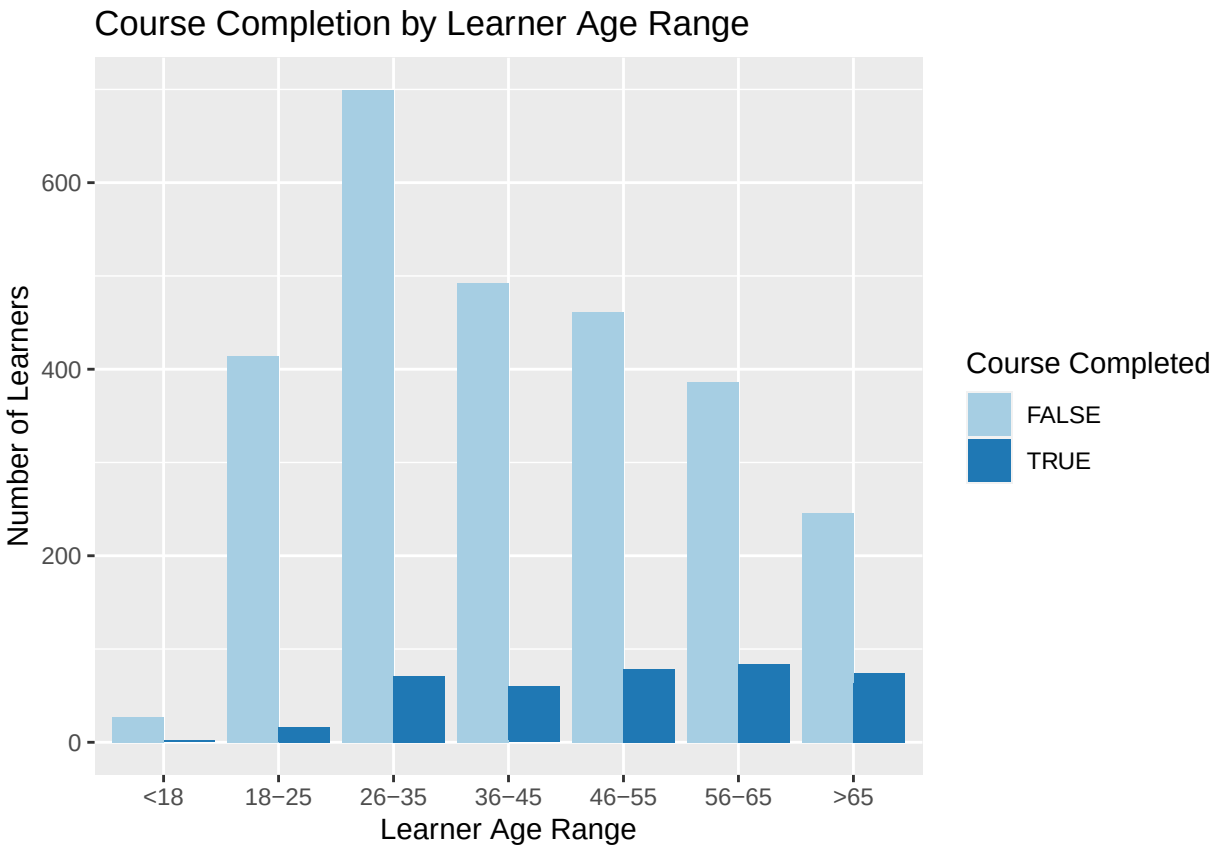
```
## Warning: Removed 3 rows containing missing values (geom_bar).
```



This graph indicates that learners who identify as male are more likely to declare all six demographic data values than women. Similarly, if a learner declares that they identify as nonbinary or other, five or more demographic fields are provided. This graph also suggests that learners who share their gender with FutureLearn are also likely to share more demographic data. To better understand these results, further exploration is required.

Goal 2.2 (Age Success Rate)

The same analysis process used for Goal 2.1 has been applied for Goal 2.2. We can determine that 4028 learners declared their age range to FutureLearn. In the graph below it appears that learners have a higher chance of success the older they are.



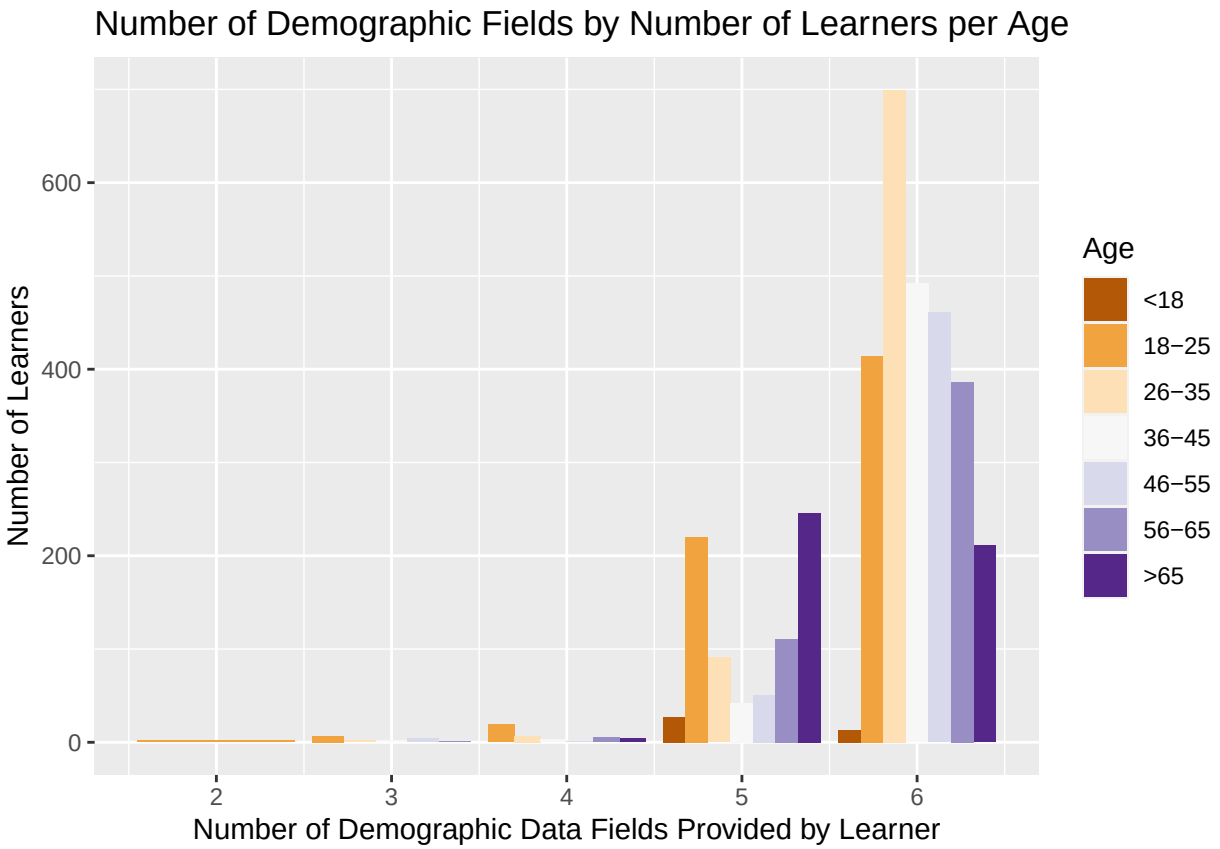
To better understand the results of this graph, the table below was also created (the code for calculating the percentages is located in the file `eda2.R` which is in the `src` folder).

```
## # A tibble: 8 x 2
##   Success_Rate Percentage
##   <chr>              <dbl>
## 1 Age                12.6
## 2 <18                 4.76
## 3 18-25               4.2
## 4 26-35               8.8
## 5 36-45              11.0
## 6 46-55              14.1
## 7 56-65              18.0
## 8 >65                22.9
```

This table shows us that of learners who declared their age 12.59% successfully completed the course. Those who are 65 or older had the highest chance of success (22.91%). Learners' chance of success appears to increase after the age of 26 with learners aged between 18 and 25 having the lowest chance of success (4.20%).

The graph below was created to better understand the provision of demographic data by learners who declared their age.

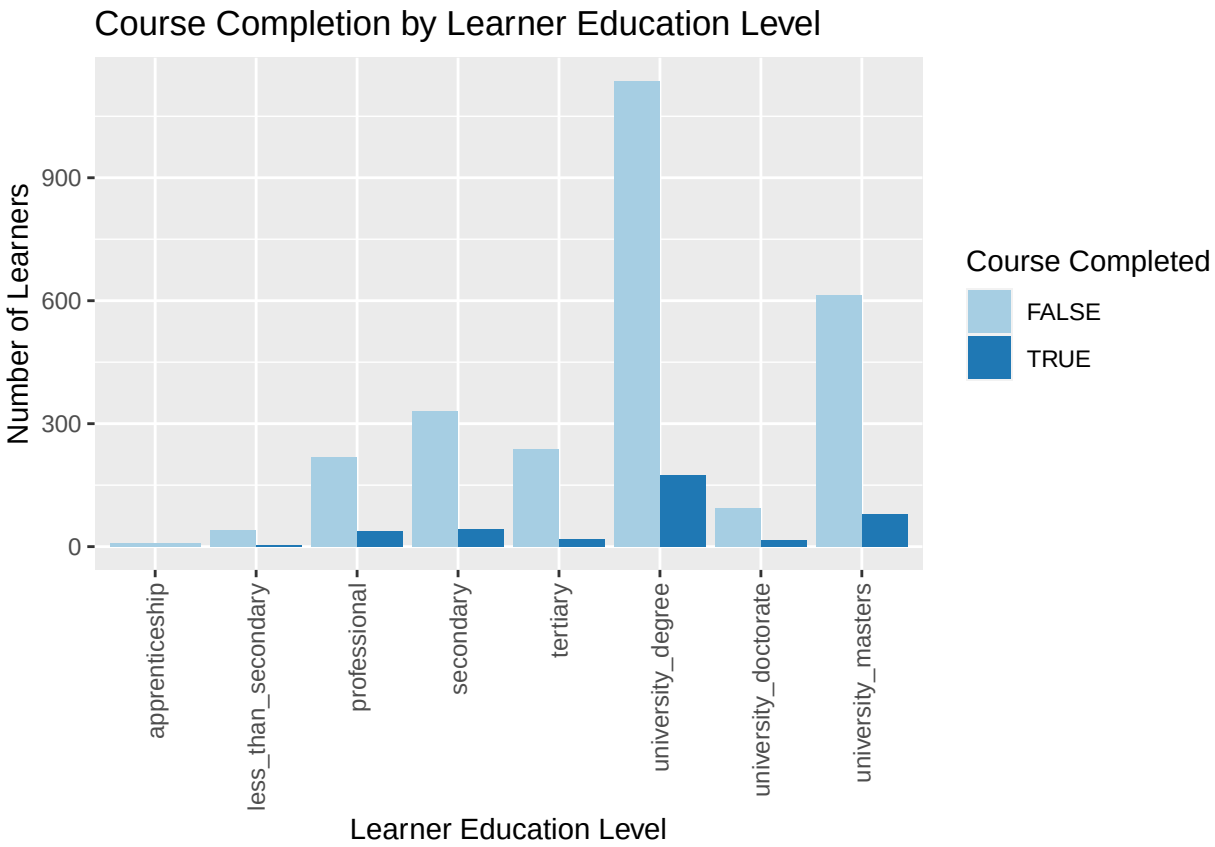
```
## Warning: Removed 3 rows containing missing values (geom_bar).
```



This graph indicates that learners in the 26-25 age range are more likely to share all six demographic data values. Comparatively, if a learner declares they are over 65, they appear more likely to provide 4 or more demographic data values. This graph also suggests that learners who share their age range with FutureLearn are also likely to share more demographic data. To better understand these results, further exploration is required.

Goal 2.3 (Education Level)

Applying the same analysis process as Goal 2.1 and 2.2, we can determine that 4135 learners declared their highest education level to FutureLearn.



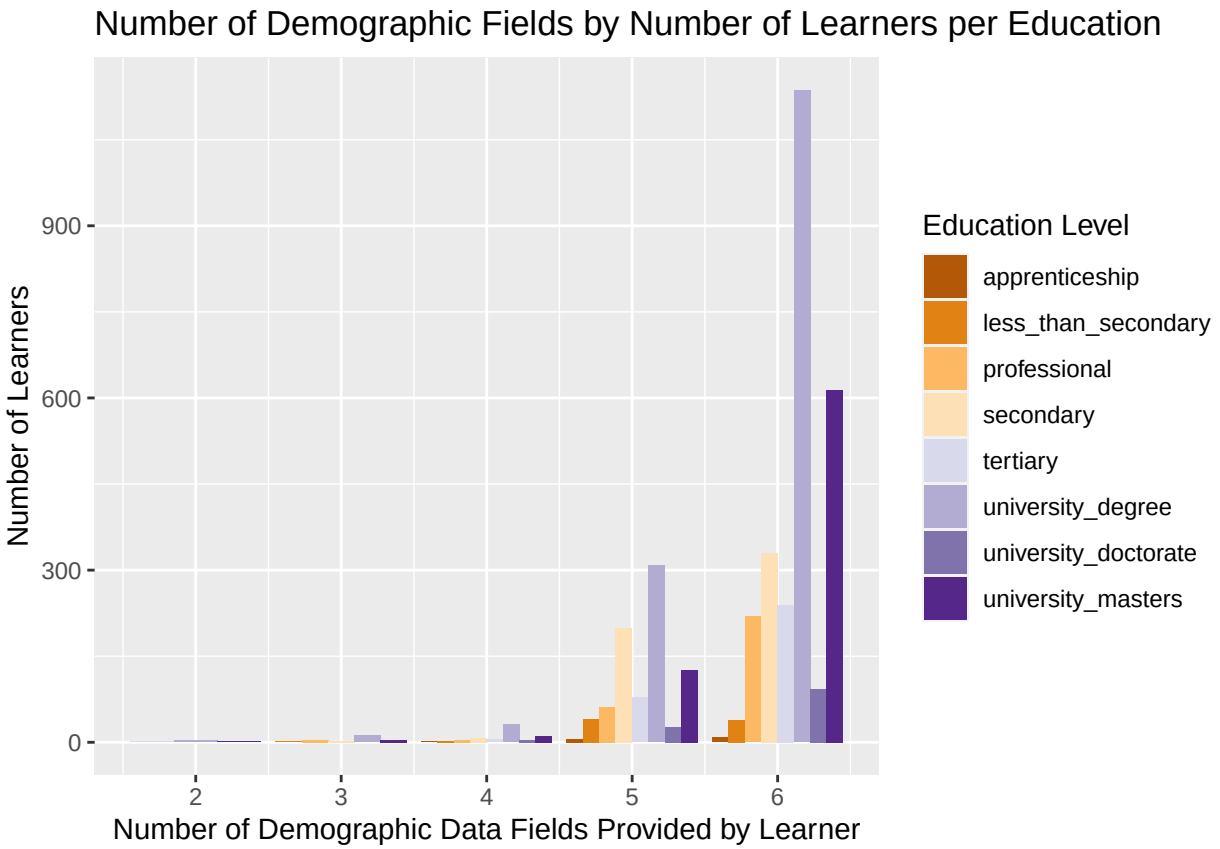
From this graph, it is unclear which education level correlates to a higher chance of success for learners. It is clear that there are more students who enrol in the course who have either a tertiary qualification of some kind and specifically a university degree or university masters. To better understand the results of this graph, the table below was also created (the code for calculating the percentages is located in the file `eda2.R` which is in the `src` folder).

```
## # A tibble: 9 x 2
##   Success_Rate          Percentage
##   <chr>                      <dbl>
## 1 Education                   12.6
## 2 Apprenticeship               0
## 3 Less_Secondary              10
## 4 Secondary                   10.9
## 5 Professional                15.0
## 6 Tertiary                    10.5
## 7 University_degree           13.8
## 8 University_masters          11.4
## 9 University_doctorate        15.1
```

This table shows us that of the learners who declared their highest education level, 12.55% successfully completed the course. Those who held university doctorates (15.07%) or professional qualifications (15.04) had the highest chance of success. Those who held apprenticeships had the lowest chance of success (0%).

The graph below was created to better understand the provision of demographic data by learners who declared their highest education level.

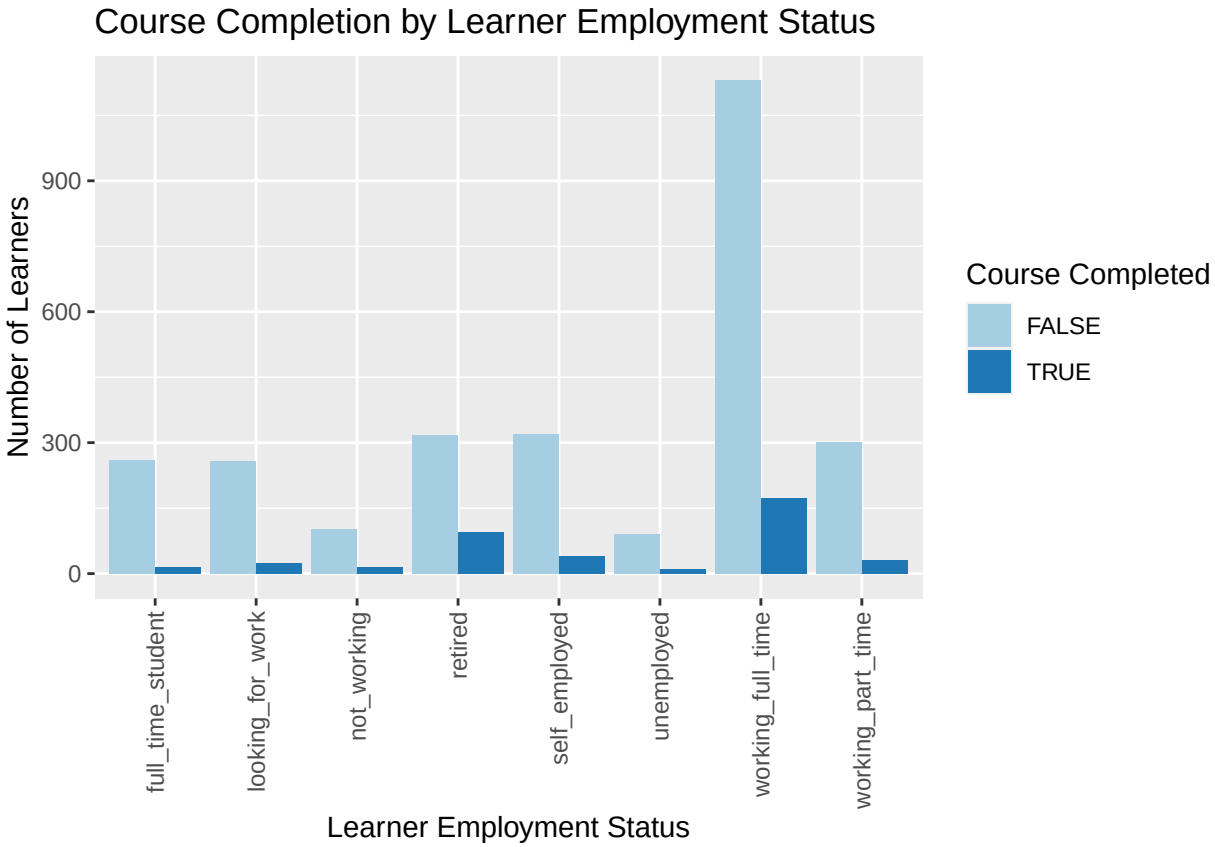
```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



This graph suggests that learners with a tertiary qualification of some kind (tertiary, university_degree, university_masters, university_doctorate) are more likely to share more demographic data with FutureLearn. It also indicates that those with apprenticeships and professional qualifications are less likely to share more demographic data with FutureLearn. To better understand these results, further investigation is required.

Goal 2.4 (Employment Status)

Applying the same analysis process as Goal 2.1, 2.2, and 2.3, we can determine that 4105 learners declared their employment status to FutureLearn. In the graph below it appears that learners have a higher chance of success if they are retired.



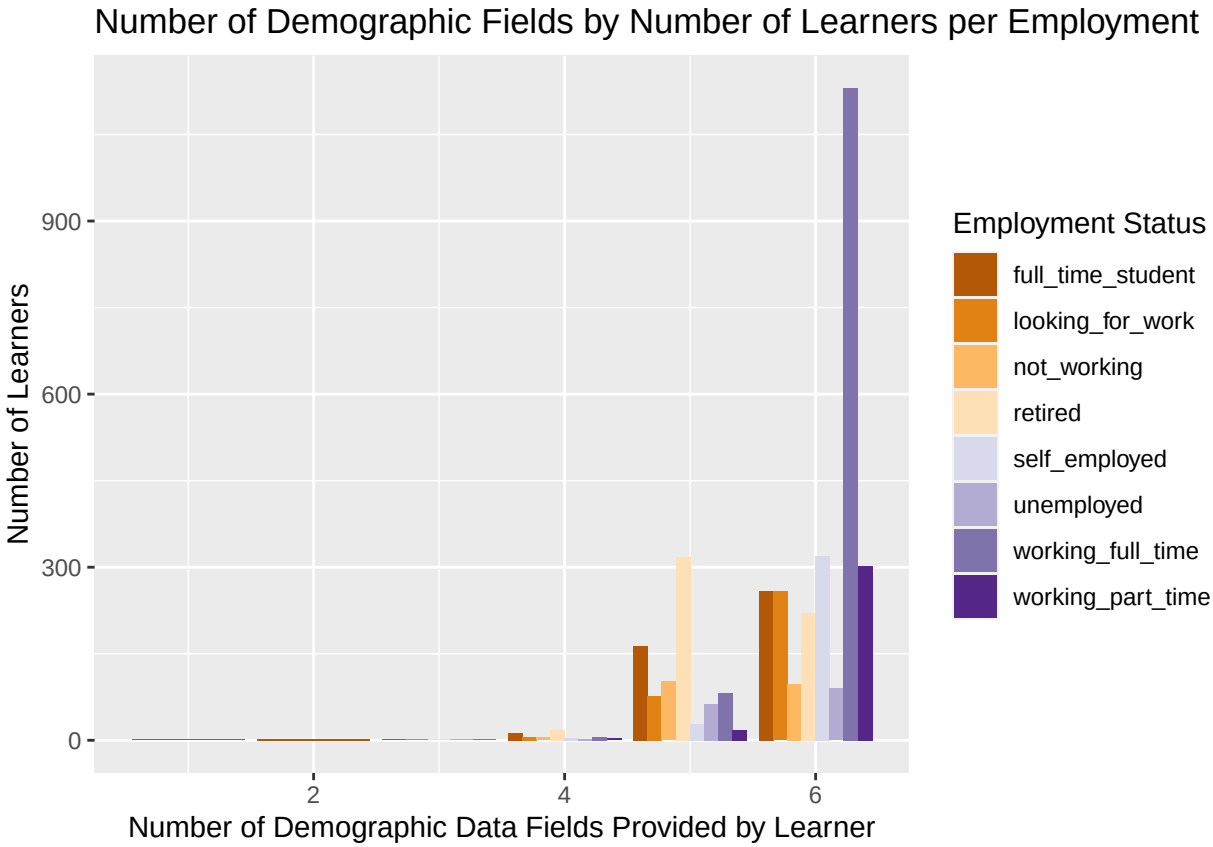
To better understand the results of this graph, the table below was also created (the code for calculating the percentages is located in the file `eda2.R` which is in the `src` folder).

```
## # A tibble: 9 x 2
##   Success_Rate  Percentage
##   <chr>              <dbl>
## 1 Employment         12.6
## 2 Student_FT          4.76
## 3 Looking             8.31
## 4 Not_working        11.3
## 5 Retired            22.4
## 6 Unemployed          9.36
## 7 Self-employed      10.9
## 8 Working_FT         13.1
## 9 Working_PT          9.27
```

This table shows us that of the 4105 learners who declared their employment status, 12.57% completed the course. Learners who are retired have the highest chance of success (22.45%) and learners who are also full time learners have the lowest chance of success (4.76%).

The graph below was created to better understand the provision of demographic data by learners who declared their employment status.

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



This graph indicates that learners who are working full time may be more likely to provide FutureLearn with more demographic data (although this could seem this was as there are more learners who are working full time). The graph also suggests that those who are retired are more likely to share five instead of six demographic data values with FutureLearn, and it would be interesting investigate if these learners do not share the same demographic data value. To better understand these results, further exploration is needed.

Data preparation

This section outlines the pre-processing steps that produce the data utilised by this report. The pre-processing code can be found in the folder called `munge`. All pre-processing code is included so that anyone can run the code of this project directly from this report. The code includes:

1. All seven enrolments data sets are combined into one single data set. The enrolments data set is chosen as the majority of demographic data provided by learners can be found here.
2. This data set is formatted so that all empty cells appear as NA.
3. Two filters are created that divide this data by learners who completed (complete) and learners who did not complete (incomplete) the course.
4. A new data set is created for the initial investigation (FL1) so that any data transformation or tidying is not applied to the raw data. In this new data set, there are seven new columns apply TRUE or FALSE based on whether there data in the aforementioned demographic data fields and in fully_participated_at.
5. The learner_ids for these new columns are identified and are pulled together as 'ids_of_declared_cols.'
6. A new column is created in FL1 that counts the number of TRUEs in the aforementioned columns.

7. There is also some pre-processing data for the third iteration of exploratory data analysis, which is not yet finished and, therefore, not included in this report. This code combined the seven archetype data sets, joins these to FL1 by learner_id, and creates a new data set of this called FL2. The data needed to assign learners with archetypes is potentially demographic data, and so potential insights from archetype data sets could offer an interesting comparison to the insights from the demographic data in the enrolments spreadsheet.

Evaluation

Assessment of data mining results with respect to business success criteria

The process and findings described by this report explore the question: "Do learners have a higher chance of success the more data they share with FutureLearn?" The business success criteria of this project is to give useful insights into the relationship between learner engagement and learner success rate. Learners' provision of optional demographic data is regarded as a proxy for learner engagement. Learner success rate is regarded as the likelihood a learner completes the course they enrolled in. Learner engagement and learner success rate are indicators of learning effectiveness, and the results of the data mining are interpreted with that context in mind.

```
## # A tibble: 15 x 2
##    Success_Rate        Percentage
##    <chr>                    <dbl>
##  1 Gross                     5.78
##  2 No Demographic Data       4.92
##  3 Demographic Data         12.5
##  4 Gender                   12.3
##  5 Male                     14.0
##  6 Other                     7.14
##  7 Age                      12.6
##  8 >65                      22.9
##  9 18-25                     4.2
## 10 Education                12.6
## 11 Doctorate                15.1
## 12 Apprenticeship            0
## 13 Employment Status        12.6
## 14 Retired                  22.4
## 15 Student                   4.76
```

This table shows the overarching success rates as well as the best and worst success rate for each field of demographic data explored through this report. The success rate of learners who share demographic data with FutureLearn is significantly higher (2.55x) than learners who do not. Interestingly, the success rate for learners who provide demographic data (12.53%) is not dissimilar for the overarching success rate of different demographic fields, such as gender (12.26%), age (12.59%), education (12.55%), employment status (12.57%). However, within these fields of demographic data, the results vary more widely.

The data mining results suggest that FutureLearn is most effective for people who identify as male (14.02%), are aged over 65 (22.91%), hold a doctorate (15.07%) or are retired (22.45%). The results also indicate that FutureLearn is least effective for people who identify as other (7.14%), are between the ages of 18 and 25 (4.20%), who hold an apprenticeship (0%) or are a full time student (4.76%). The similarity between some of these data variables suggests that FutureLearn may be more effective for experienced learners with more time on their hands. Further investigation is, however, needed to understand whether these variables refer to particular learner_ids or pattern for learners.

Initial insights were also collected to better understand the engagement of those who provided demographic data. The data mining results indicate that those who share demographic data are likely to provide FutureLearn with more demographic data. Further investigation is needed to conclusively determine whether learners have a higher chance of success the more data they share with FutureLearn; the findings in this report, however, suggest this holds true.

The findings of the exploratory analysis in this report do produce useful insights into the relationship between learner engagement and learner success rate. There is a clear correlation between these indicators of learning effectiveness, and it does appear that learners who engage by providing demographic data are more likely to succeed.

Review of process

This project has been limited by the time available, programming capabilities, choice of data sets, and scope of the project to focus on data exploration. The process has been sufficiently effective to somewhat address the business objective and success criteria. This project, however, focused more on learners success rate than learner engagement. If the project should be repeated, further attention should be given to understanding the impact of the amount of data shared by students as well as different indicators of learner engagement, such as archetypes, step activity and leaving survey responses.

List of possible actions

If the project were to be continued, the possible actions recommended are:

1. Calculate the success rate for learners of demographic data field that also provide one, two, three, four, five and six demographic data fields. This will allow us to confirm whether or not there are spurious associations.
2. Explore the relationship between success rate and other indicators of learner engagement (as mentioned above).
3. The finding of point 2 could then be compared with the findings for demographic data fields outlined by this report. This would help FutureLearn and Newcastle University better understand which proxies best reflect learner engagement.
4. Where groups of learners' success rate are lower, we should explore whether or not they responded to the leaving survey and what the most common responses are. This will help us to identify how we might improve FutureLean's learning effectiveness for them.
5. Another interest avenue for exploration is the relationship between the learners' success rate and completion time, as this could be another indicator of learner engagement.

Bibliography

1st International Conference on Learning Analytics and Knowledge. 2011. Banff, Alberta. https://tekri.athabascau.ca/analytics/.

Chapman et. al., Juian Clinton, Pete Chapman, and Rudiger Wirth. 2000. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. https://www.the-modeling-agency.com/crisp-dm.pdf.

FutureLearn. n.d. "The Power of Social Learning: An Effective Way to Learn." https://www.futurelearn.com/using-futurelearn/why-it-works.

Long, Phil, and George Siemens. 2011. "Penetrating the Fog: Analytics in Learning and Educaton." *EDUCAUSE Review* September/October. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education.

S. Hubalovsky, M. Hubalovska, and M. Musilek. 2018. "Assessment of the Influence of Adaptive e-Learning on Learning Effectiveness of Primary School Puplis." *Computers in Human Behaviour* 92. https://doi.org/10.1016/j.chb.2018.05.033.

Shacklock, Xanthe. 2016. "From Bricks to Clicks." Policy Connect; Higher Education Commission.