CSC8631 Report

Yahan Wang 200784463

11/23/2021

Business Understanding

The task of the business understanding phase is to articulate the goals and requirements of data mining from a business perspective and translate them into specific data mining problems. The objective of this paper is to measure, collect, analyse and report data about learners and their environment in order to understand and optimise their learning environment. For learners who are less motivated to attend classes, this may be influenced by a number of factors such as gender, difficulty of the course taken, country, major, and age. The key to optimising its learning environment is therefore to portray the factors that influence learner motivation based on learner characteristics and other supplementary data sources (e.g. access to on-campus facilities, Virtual Learning Environment (VLE) and Re-Cap visits, and student welfare referrals).

Data Understanding

Collect initial data

Data quality checking and initial characterisation, which is essentially a process of data collection and familiarisation, Found in practice, age range of learners, gender, highest level of education, employment status, region of employment, country, , survey responses, leaving survey responses, step activity, question responses, video statistics .... etc,can affect learners' motivation to learn. The data used in this paper are derived from a school's student information management warehouse, with detailed records of learner characteristics and other complementary data sources (e.g. use of on-campus facilities, Virtual Learning Environment (VLE) and Re-Cap access, and student welfare referrals). In order to facilitate student management, schools set up learner ids for each student, record enrolment information (e.g. gender, country, major, age, etc.) and keep records of step activities, question answers, etc. This paper will divide students by the number of questions they answer, the difference in the number of questions answered by students is more significant, so this paper will model the factors that influence the number of questions answered by students to be explored

Describe data

The data set cyber.security.enrolments contains a total of 13 variables, each with 35,225 data of the data type character

```
str(cyber.security.enrolments,vec.len  =1)
```

```
## tibble [35,225 x 13] (S3: tbl_df/tbl/data.frame)
##  $ learner_id             : chr [1:35225] "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b" ...
##  $ enrolled_at            : chr [1:35225] "2016-08-10 14:28:49 UTC" ...
```

```
##  $ unenrolled_at          : chr [1:35225] "" ...
##  $ role                   : chr [1:35225] "learner" ...
##  $ fully_participated_at  : chr [1:35225] "" ...
##  $ purchased_statement_at : chr [1:35225] "" ...
##  $ gender                 : chr [1:35225] "Unknown" ...
##  $ country                : chr [1:35225] "Unknown" ...
##  $ age_range              : chr [1:35225] "Unknown" ...
##  $ highest_education_level: chr [1:35225] "Unknown" ...
##  $ employment_status      : chr [1:35225] "Unknown" ...
##  $ employment_area        : chr [1:35225] "Unknown" ...
##  $ detected_country       : chr [1:35225] "GB" ...
```

The data set cyber.security.step.activity_1 contains six variables, each with 423072 data. learner_id, first_visited_at, last_completed_at are of type character, week_number, step_number are of type int, and step is of type numeric.

```
str(cyber.security.step.activity_1,vec.len  =1)
```

```
## tibble [423,072 x 6] (S3: tbl_df/tbl/data.frame)
##  $ learner_id       : chr [1:423072] "77454a73-6b8b-46a2-8dee-35f36b6c4fc1" ...
##  $ step             : num [1:423072] 1.1 1.1 ...
##  $ week_number      : int [1:423072] 1 1 ...
##  $ step_number      : int [1:423072] 1 1 ...
##  $ first_visited_at : chr [1:423072] "2016-08-02 13:45:37 UTC" ...
##  $ last_completed_at: chr [1:423072] "" ...
```

The data set cyber.security.question.response_1 contains a total of 10 variables with 176463 data each, learner_id, quiz_question, question_type, response, submitted_at and correct are of type character, week_number, step_number, question_number are of type integer, and cloze_response is of type logical.

```
str(cyber.security.question.response_1,vec.len  =1)
```

```
## tibble [176,463 x 10] (S3: tbl_df/tbl/data.frame)
##  $ learner_id     : chr [1:176463] "77454a73-6b8b-46a2-8dee-35f36b6c4fc1" ...
##  $ quiz_question  : chr [1:176463] "1.7.1" ...
##  $ question_type  : chr [1:176463] "MultipleChoice" ...
##  $ week_number    : int [1:176463] 1 1 ...
##  $ step_number    : int [1:176463] 7 7 ...
##  $ question_number: int [1:176463] 1 1 ...
##  $ response       : chr [1:176463] "1,2" ...
##  $ cloze_response : logi [1:176463] NA ...
##  $ submitted_at   : chr [1:176463] "2016-07-06 10:37:05 UTC" ...
##  $ correct        : chr [1:176463] "false" ...
```

The dataset cyber.security_video.stats_1 contains a total of 28 variables, each with 65 data, title is of type character, video_duration,total_views, total_downloads,total_caption video_duration, total_views, total_downloads, total_caption_views, total_transcript_views, viewed_hd are of type integer The other variables are of type numeric.

```
str(cyber.security_video.stats_1,vec.len  =1)
```

```
## tibble [65 x 28] (S3: tbl_df/tbl/data.frame)
##  $ step_position                 : num [1:65] 1.1 1.14 ...
##  $ title                         : chr [1:65] "Welcome to the course" ...
##  $ video_duration                : int [1:65] 99 362 ...
##  $ total_views                   : int [1:65] 1659 910 ...
##  $ total_downloads               : int [1:65] 113 77 ...
##  $ total_caption_views           : int [1:65] 36 8 ...
##  $ total_transcript_views        : int [1:65] 221 173 ...
##  $ viewed_hd                     : int [1:65] 58 28 ...
##  $ viewed_five_percent           : num [1:65] 77 ...
##  $ viewed_ten_percent            : num [1:65] 75.3 ...
##  $ viewed_twentyfive_percent     : num [1:65] 73.4 ...
##  $ viewed_fifty_percent          : num [1:65] 70.4 ...
##  $ viewed_seventyfive_percent    : num [1:65] 68.2 ...
##  $ viewed_ninetyfive_percent     : num [1:65] 66.4 ...
##  $ viewed_onehundred_percent     : num [1:65] 63.7 ...
##  $ console_device_percentage     : num [1:65] 0.06 0.11 ...
##  $ desktop_device_percentage     : num [1:65] 78.6 ...
##  $ mobile_device_percentage      : num [1:65] 13.3 ...
##  $ tv_device_percentage          : num [1:65] 0.06 0 ...
##  $ tablet_device_percentage      : num [1:65] 7.72 ...
##  $ unknown_device_percentage     : num [1:65] 0 0 ...
##  $ europe_views_percentage       : num [1:65] 55.1 ...
##  $ oceania_views_percentage      : num [1:65] 2.29 2.86 ...
##  $ asia_views_percentage         : num [1:65] 16.1 ...
##  $ north_america_views_percentage: num [1:65] 11.6 ...
##  $ south_america_views_percentage: num [1:65] 3.07 2.53 ...
##  $ africa_views_percentage       : num [1:65] 10.3 ...
##  $ antarctica_views_percentage   : num [1:65] 0 0 ...
```

The data set cyber.security_weekly.sentiment.survey.responses_1 contains a total of 5 variables, each with 181 data, with the data types responded_at and reason being character, id, week_number and experience_rating are of type int

```
  str(cyber.security_weekly.sentiment.survey.responses_1)
```

```
## tibble [181 x 5] (S3: tbl_df/tbl/data.frame)
##  $ id               : int [1:181] 393 16810 17388 17505 17807 17819 18277 18291 18439 18706 ...
##  $ responded_at     : chr [1:181] "2018-04-16 20:17:14 UTC" "2018-06-11 18:28:04 UTC" "2018-06-12 16
##  $ week_number      : int [1:181] 2 1 2 2 1 1 1 1 2 1 ...
##  $ experience_rating: int [1:181] 3 3 3 3 3 3 3 3 3 3 ...
##  $ reason           : chr [1:181] NA "" "It was good i have learnt new measure in protecting my money
```

Data preparation.

The data preparation phase covers all the work involved in constructing the final dataset (for modelling analysis) from the raw rough data, including steps such as data cleaning and variable selection.

Selecting data

Task:

First select the seven enrolments form.This is because the seven forms contain a lot of information about the learner, such as age range, gender, highest level of education, employment status, region of employment, country, etc.This may have an impact on the number of questions answered, the number of step activities, and the number of videos viewed by learners.Seven question response forms were then selected, which allowed for statistics on the number of responses per learner and the number and frequency of responses per question.Then select the seven Step activity tables, which will give you an idea of the number of people who completed and did not complete the activity at each step.Based on these video stats tables, you can see the popularity of each of the step videos.All this data has a crucial impact on portraying the factors that influence learners' motivation and optimising its learning environment.

Output:

In the seven enrolments tables, select age_range, highest_education_level, employment_status, employment_area,country and detected_country as the data to be used later, because these data can be used to explore the correlation with learner motivation, the enrolled at and unenrolled at cannot be used to explore learner motivation, and all the roles are learners, so they are removed. The seven question response forms leave the learner id and quiz_question, and correct.Because the question type is only one type of question that is MultipleChoice. The week number, step number and question number are all included in the quiz_question, The answer in the response is used to determine whether the correct is TRUE or FALSE, and the final result only depends on the correct so all this data needs to be removed. The seven Step activity tables need to leave the learner_id and step columns, and also need to leave first_visited_at, last_completed_at to determine if the learner is responding to the answer, based on the observation that as long as first_visited_at, last_completed_at both information is available. In the question response form there must be a response and a correct(TRUE or FALSE).However, the information on week_number and step_number is contained in the step so these data need to be deleted.

Clean data

Task

First select the seven enrolments table and then delete the rows with duplicate learner_id's from each of the seven enrolments leaving only one row for each duplicate learner_id, In each enrolments table, the six variables age_range, highest_education_level, employment_status, employment_area, country, detected_country are filtered to remove variables equal to "Unknown", and because the country and detected_country variables have too many categories, countries with a small number of people are filtered out, leaving only the top ten

Output

In the seven enrolments table, variables equal to "Unknown" will have an impact on the statistical results because they are highly uncertain. Although filtering out variables equal to "Unknown" will also have an impact on the results, the impact is much smaller than if the variables equal to "Unknown" were not filtered out.

Construct data

Task

First select the seven enrolments tables, use the table function on the country and detected_country variables to generate two tables with statistics on the frequency of each country, then use the sort function to sort these two tables to generate two new frames.

Output

Generate two tables with the frequency of each country to find out the number of learners from each country, then use the sort function to sort the two tables to generate two new frames to find out the top 10 countries in terms of the number of learners

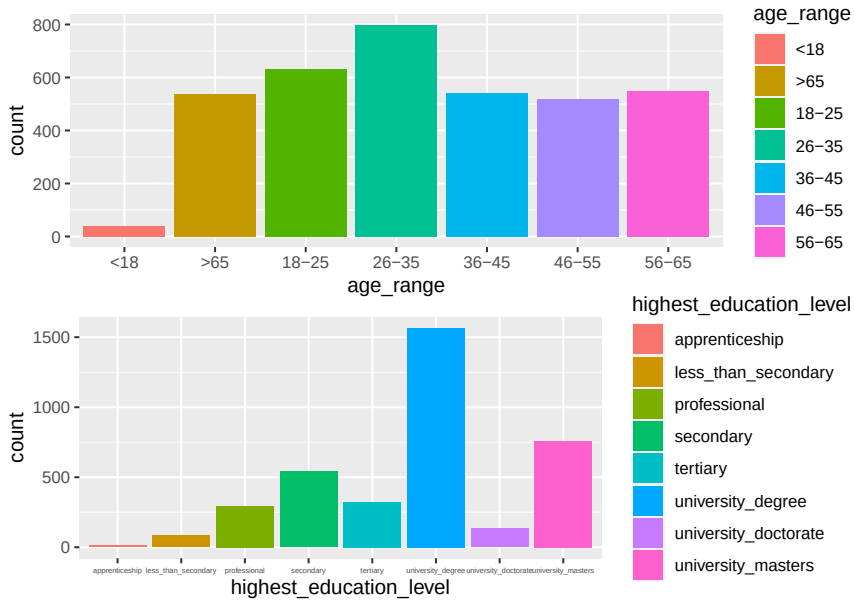
Integrate data

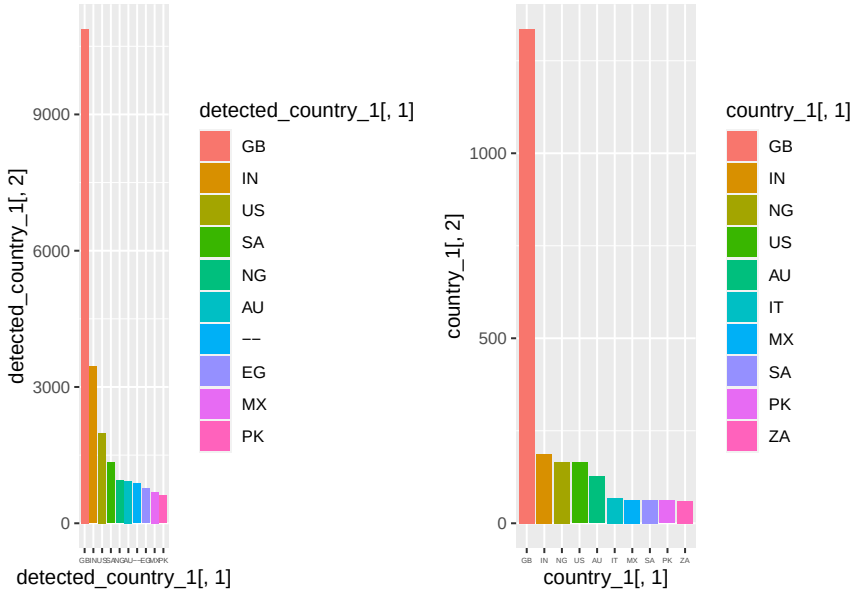
Task

Synthesize the seven enrolments tables into a single table, cyber.security.step.activity_1, and remove all duplicate learner_id's from the sequence (keeping only one) to become the table cyber.security.step.activity.

Output

Combine into a single table the statistics for all learners, e.g. age_range, highest_education_level, etc. The reason for removing all duplicate learner_id's from the sequence (keeping only one) is to prevent wasting resources by counting the same learner multiple times.





Charting

The first graph shows a histogram of the age distribution of all enrolments, from which it can be seen that the largest number of enrolments was between 26-35 years old, with around 800, a slightly smaller number between 18-25 years old, with around 620, and the smallest number of enrolments was under 18 years old, with around 20.

The second graph shows the distribution of the highest education of all registrants, from the graph we can see that the highest education is university_degree, there are more than 1500 people, the next highest education is university_masters, there are about 750 people, the highest education is apprenticeship, the least number of people, basically belong to single digits, this graph fully illustrates that the registrants' education is on the high side

The country with the most registrants in the third and fourth charts are both British nationals, while the country with relatively few registrants is Pakistan.

Clean data

Task

In the seven question response tables, filter according to quiz_question, and after filtering the remaining variables according to quiz_question (1.7.1, 1.7.2, 1.7.3, 1.7.4, 1.7.5, 1.7.6, 1.8.1, 1.8.2, 1.8.3, 1.8.4 1.8.5, 1.8.6, 2.8.1, 2.8.2, 2.8.3, 2.19.1, 2.20.1, 3.11.1, 3.11.2, 3.11.3, 3.18.1, 3.18.2, 3.18.3, 3.18.4, 3.18.5, 3.18.6, 3.18.7, 3.18.8, 3.18.9), the remaining variables were divided into different groups.

Output

In the seven question response tables, filtered by quiz_question, the number of people who answered each question and the number of times each question was answered were counted later

Construct data

Task

In the seven question response tables, the learner_id variable in the data set filtered by quiz_question(1.7.1, 1.7.2, 1.7.3, 1.7.4, 1.7.5, 1.7.6, 1.8.1, 1.8.2, 1.8.3, 1.8.4, 1.8.5, 1.8.6, 2.8.1, 2.8.2, 2.8.3 2.19.1, 2.20.1, 3.11.1, 3.11.2, 3.11.3, 3.18.1, 3.18.2, 3.18.3, 3.18.4, 3.18.5, 3.18.6, 3.18.7, 3.18.8, 3.18.9) is removed using the unique function, i.e. all duplicate learner_ids in the sequence are removed (only one is retained), the length function is used to count the number of learner_ids in each filtered data set, and a frame is generated for each quiz_question and the corresponding number of respondents

Output

In the seven question response forms, the learner_id variable in the data set filtered by quiz_question is removed using the unique function, which removes all duplicate learner_ids in the sequence (only one is retained), in order to find the learners for each quiz_question and avoid duplication, and then The length function is used to count the number of learners in each filtered dataset and to generate a frame for each quiz_question and the corresponding number of answers, for the purpose of the graphing operation later on.

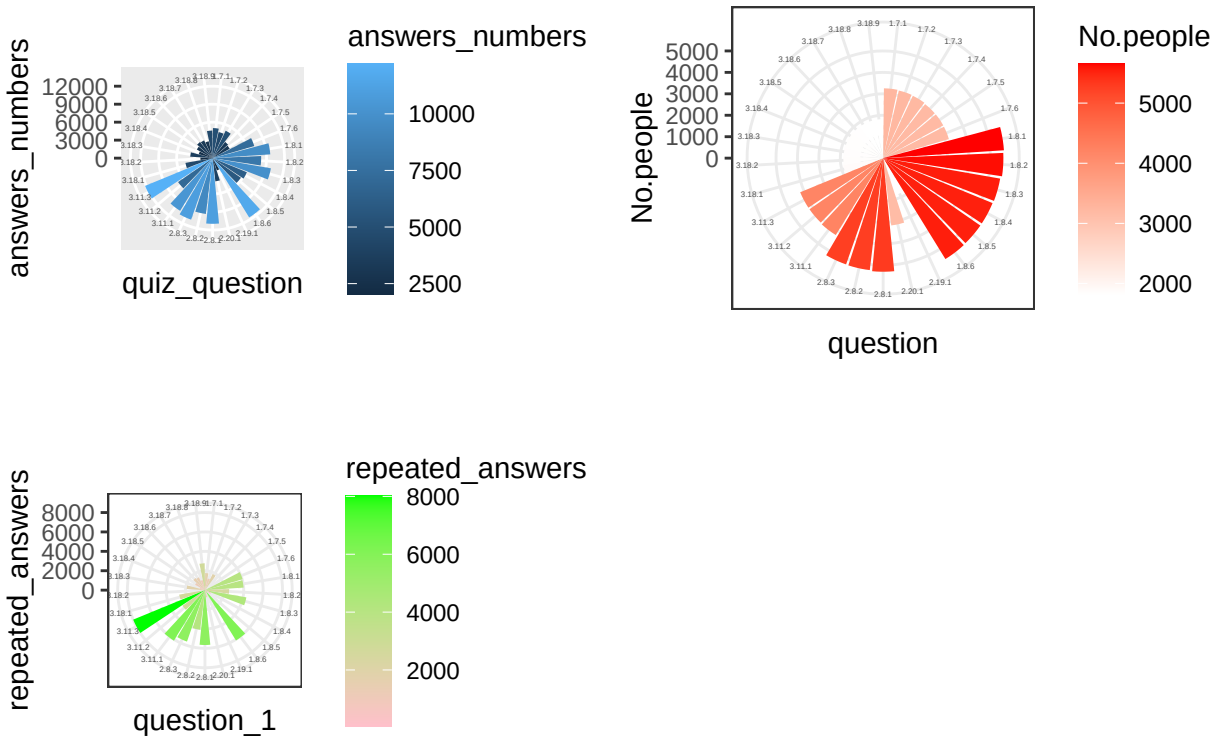
Integrate data

Task

Synthesize the seven question response tables into a single table cyber.security.question.response_1.

Output

The seven question response forms were combined into a single table to facilitate the counting of all students' responses and the number of responses to each question.



Charting

The first graph shows the number of answers to each question, with the highest number of answers to question 3.11.3, nearly 12,000, and the lowest number of answers to 3.18.2, less than 2,500.

The second chart shows the number of people who answered each question. The number of people who answered each question in 1.7 was very close, around 2,000, the number of people who answered each question in 1.8 was also very close, around 4,500, the number of people who answered 2.19 was very small and almost non-existent, the number of people who answered each question in 2.8 was also very close, around 4,000, and the number of people who answered each question in 3.11 was also very close, around 3,000. The number of respondents per question in 3.11 was also very close, at around 3,000. All in all, the number of respondents to each question in each module was very similar.

The third graph shows the number of extra answers each person gave after answering each question once, the more the number of extra answers, the more difficult the question is.According to the graph, 3.11.3 was the most difficult question, with 6,000 more answers, while the second graph shows that just over 3,000 people answered this question, with an average of two more answers per person.The question 2.20.1 was the easiest, with basically no more answers.

Clean data

Task

For the seven Step activity tables, filter out rows with last_completed_at=null

Output

Seven Step activity tables, filtering out rows with last_completed_at=null, to find learners with response answers.

Construct data

Task

In the seven Step activity tables,the cyber.security.step.activity and the cyber.security.step.activity with "last_completed_at=null" filtered out were converted to frames using the table function. The table function is used to generate two tables with statistics on the frequency of each step, and then the as.data.frame function is used to convert the two tables into a frame, generating two new frames.Divide the number of occurrences of cyber.security.step.activity using the table function by the number of occurrences of cyber.security.step.activity using the table function by filtering out last_completed_at=null to find the percentage.The resulting percentage and the step variable in a table like cyber.security.step.activity with "last_completed_at=null" are then filtered out to create another frame

Output

In the seven Step activity tables,the cyber.security.step.activity and the cyber.security.step.activity with "last_completed_at=null" filtered out were converted to frames using the table function. The table function is used to generate two tables with statistics on the frequency of each step„in order to count The number of people who participated in the activity and the number of people who participated but did not complete the activity (i.e. did not answer the question), and then the as.data.frame function is used to convert the two tables into a frame, generating two new frames,in order to facilitate future graphing operations.Divide the number of occurrences of cyber.security.step.activity using the table function by the number of occurrences of cyber.security.step.activity using the table function by filtering out last_completed_at=null to find the percentage.In order to find the percentage of participants who did not complete the activity (i.e. did not answer the question).The resulting percentage and the step variable in a table like cyber.security.step.activity with "last_completed_at=null" are then filtered out to create another frame for later graphing.

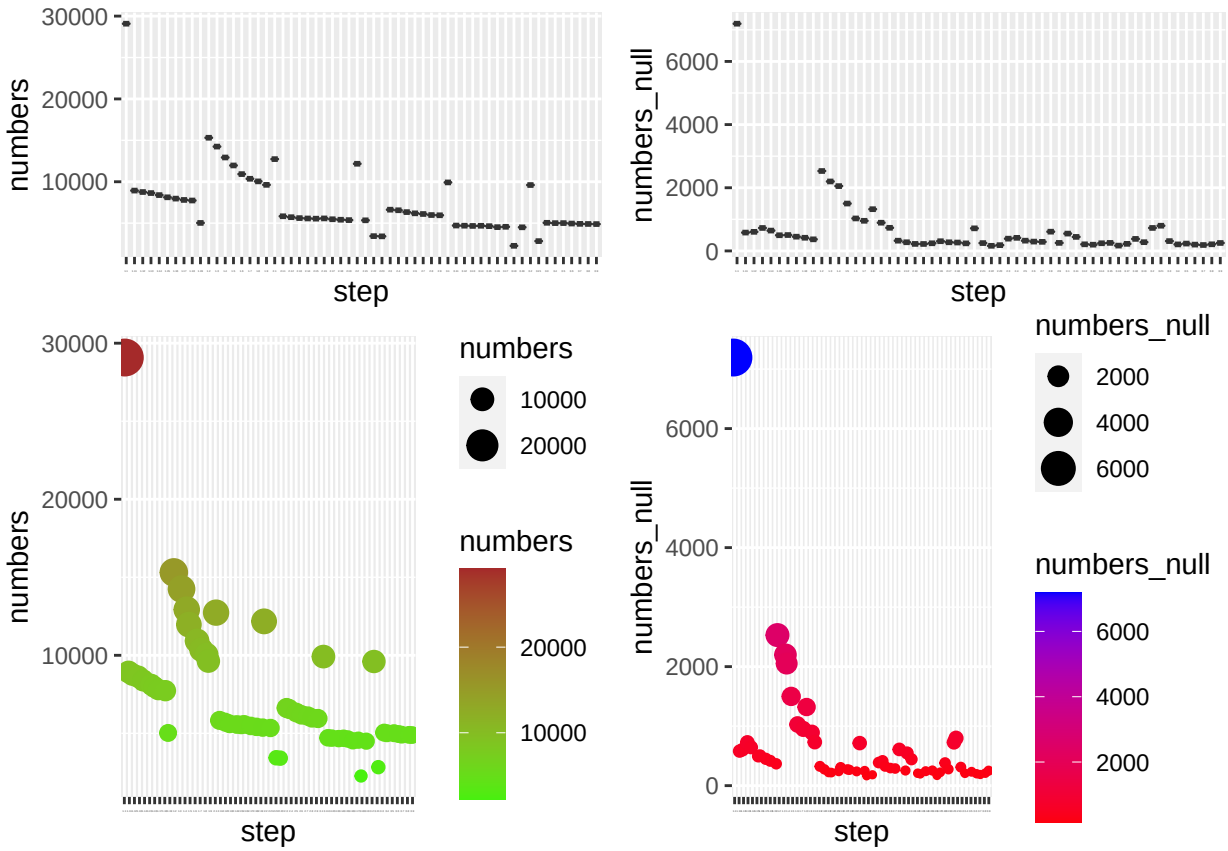
Integrate data

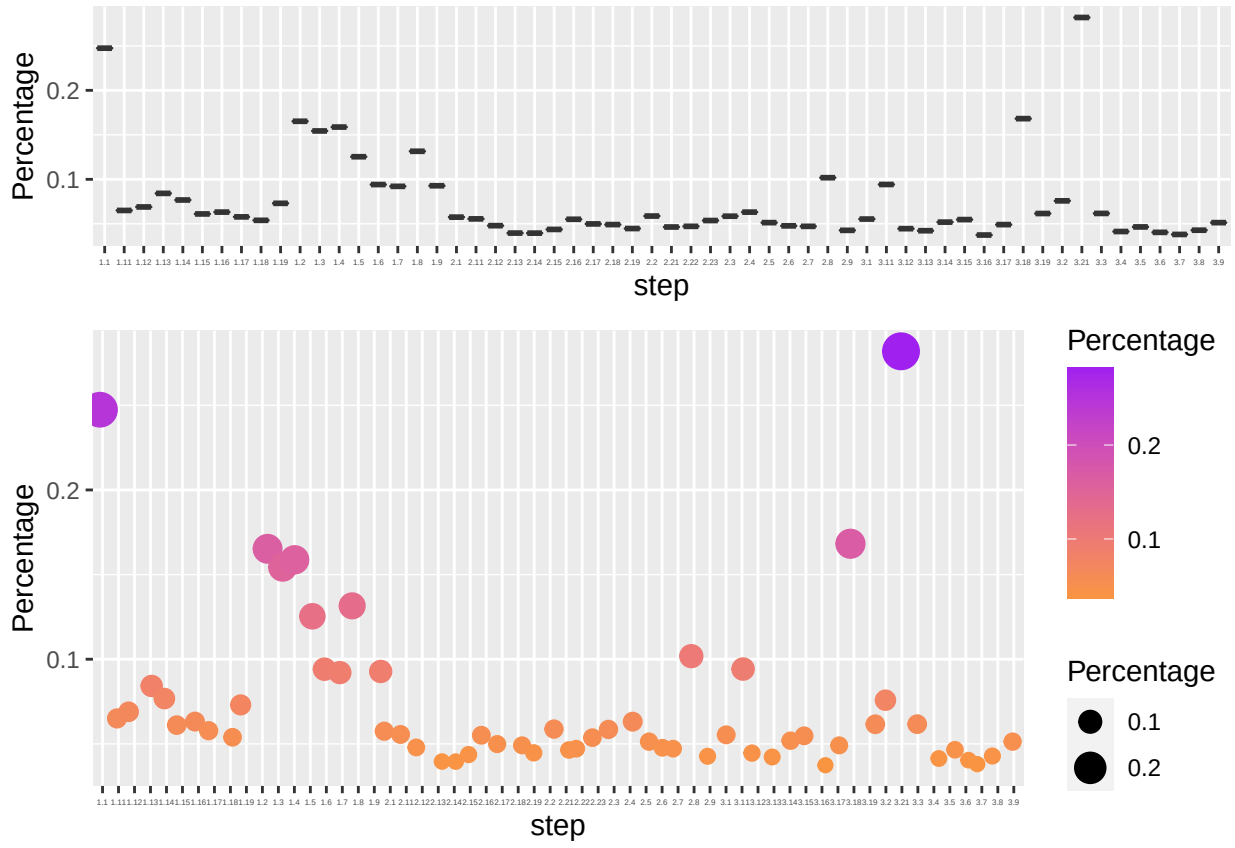
Task

Synthesize the seven cyber.security_step.activity tables into one table cyber.security.step.activity_1

Output

The seven cyber.security_step.activity tables were merged into one table in order to count the number of people who participated in all activities and the number of people who participated in activities but did not complete them (i.e. did not answer the questions) and to count the number of people who participated and completed them


numbers
30000
20000
10000
step
numbers_null
6000
4000
2000
0
step
numbers
10000
20000
numbers
20000
10000
numbers
30000
20000
10000
step
numbers_null
2000
4000
6000
numbers_null
6000
4000
2000
numbers_null
6000
4000
2000
0
step



Charting

The two charts on the left of the first graph show the number of visits for each step, with the highest number of visits for step being 29,081.The two graphs on the left show a gradual decrease in the number of visits from 1.1 to 1.19, but a sudden rise in the number of visits to the question at 2.1.Then from 2.1 to 2.19 the number of people visiting the steps started to decrease again, and then on 3.1 there was a sudden increase in the number of people visiting the steps,From 3.1 to 3.21 the numbers started to fall again.This shows that the enthusiasm at the beginning of each section is very high, but as time goes by the number of people who stick with it gradually decreases.

The two graphs on the right of the first chart show the number of people who did not answer the question at each step, with the highest number of people not answering the question at step 1.1 being 7194,The two graphs on the right show a gradual decrease in the number of non-respondents from 1.1 to 1.19, but a sudden rebound in the number of non-respondents by 2.1.Then from 2.1 to 2.19 the number of non-respondents started to gradually decrease again, and on 3.1 there was a sudden increase in the number of non-respondents,From 3.1 to 3.21 the numbers started to fall again. This is a good indication that the higher the number of vist, the higher the number of unanswered questions not completed

The second graph counts the number of people who did not answer the question as a percentage of the number of vist and finds that all are below 0.1 except for steps 1.1, 1.2-1.5, 1.8, 2.8, 3.18, 3.21 which account for more than 0.1

Clean data

Task

In the United_1 table, the six variables age_range, highest_education_level, employment_status, employment_area, country and gender are filtered. Filter out variables equal to “Unknown”, and filter out gender equal to “nonbinary” and “other”.

Output

In the United_1 table, variables equal to “Unknown” will have an impact on the statistical results because they are highly uncertain. Although filtering out variables equal to “Unknown” will also have an impact on the results, the impact is much smaller than if the variables equal to “Unknown” were not filtered out.Also filter out gender equals “nonbinary” and “other” because the number of learners in these two cases is very small and not statistically significant

Construct data

Task

In the United_1 table, use the table function on the leaner_id variable to generate a table with statistics on the frequency of each learner, and then use the as.data.frame function to transform this table to generate a new frame

Output

In the United_1 table, use the table function on the leaner_id variable to generate a table with statistics on the frequency of each learner, this operation is used to count the number of courses attended by each learner and then use the as.data.frame function to transform this table to generate a new frame this operation is for later drawing and modelling.

Integrate data

Task

Merge cyber.security.question.response_1 and cyber.security.enrolments through the “learner_id.”

Output

In the United_1 table, merge cyber.security.question.response_1 and cyber.security.enrolments by “learner_id” to get a count of the number of courses each learner has attended

Format data

Task

In the United_1 table, put male=0, female=1 in the gender variable,put “<18”=0” “18-25”=1, “26-35”=2, “36-45”=3, “46-55” = 4 , “56-65” = 5, “>65” = 6, in the age_range variable, put ” university_masters”=1, “university_degree”=2, “university_doctorate”=3, ” secondary”=4, “less_than_secondary”=5, “professional”=6, “tertiary”=7, “apprenticeship”=8 in the highest_education_level variable

Output

Because gender is a character, age_range is a character and highest_education_level is also a character, these data types should be converted to facilitate the subsequent statistical work

Modeling

In the modelling phase, data and information will be sorted, simplified, extracted and summarised using a variety of modelling methods to produce a rationalised model that works well. The objective of this part of the study is to determine whether the number of courses chosen by a learner is related to his or her gender, age and education level, which is a typical dichotomous problem.Logistic regression models can be used to predict the probabilities of interest and to find the most relevant characteristic variables.The correlation index obtained from the logistic regression is distributed in the interval [0, 1].The smaller the correlation index, the more relevant.So we can see that gender is most relevant to the number of courses taken.

Evaluation

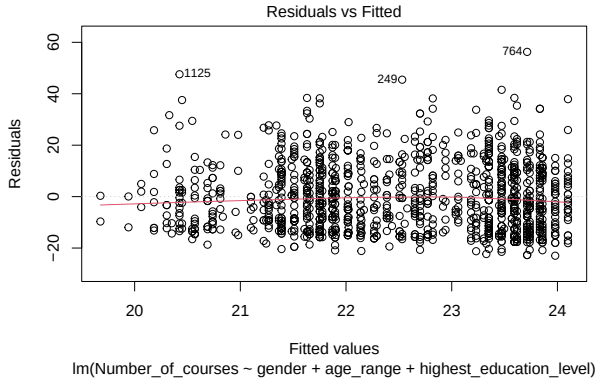
Before a model can be deployed, its effectiveness needs to be evaluated. It is also necessary to revisit each step of the modelling process to ensure that the model has met the objectives of this data analysis task. In this paper, the predict function is used to evaluate and graph the model, which is not good because the image drawn is scattered and not significant.

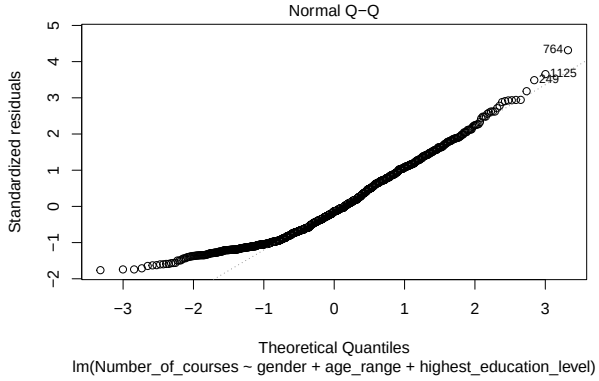
Deployment

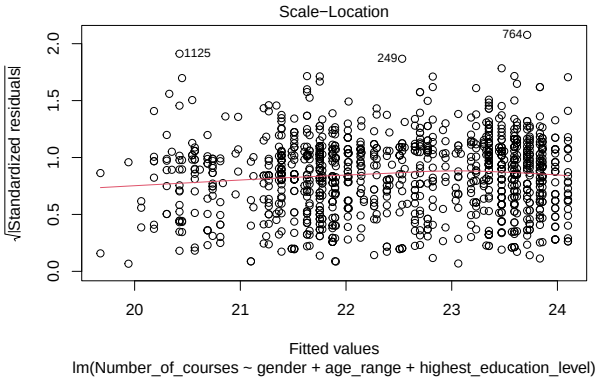
Model deployment refers to the application of data models to real-world scenarios.School managers can provide feedback on the model in relation to the actual application process. The model designed for this study does not allow for a good identification of the model that determines the number of lessons taken by learners, and it remains to be explored what strategies should be used to use the regular characteristic variables as a dimension of analysis to assist school administrators in capturing the characteristics of the target group of users.

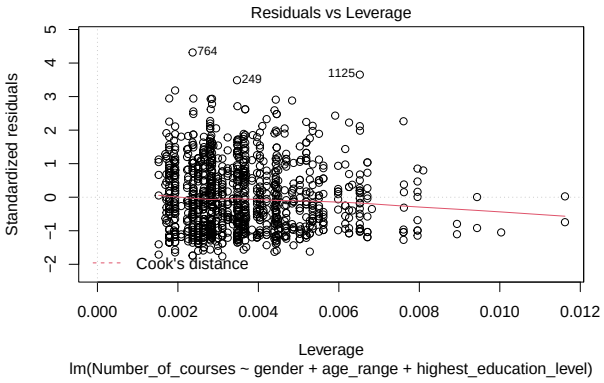
```
##
## Call:
## lm(formula = Number_of_courses ~ gender + age_range + highest_education_level,
##     data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.979 -10.840  -1.773   9.234  56.286
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               23.6412     1.2007  19.689   <2e-16 ***
## gender                    -1.9619     0.7979  -2.459   0.0141 *
## age_range                  0.1207     0.2489   0.485   0.6278
## highest_education_level   -0.2655     0.1972  -1.346   0.1785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

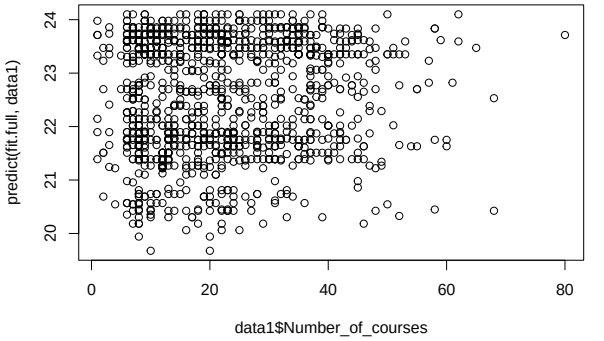
```
## Residual standard error: 13.06 on 1116 degrees of freedom
## Multiple R-squared:  0.006769,   Adjusted R-squared:  0.004099
## F-statistic: 2.535 on 3 and 1116 DF,  p-value: 0.05546
```








Residuals vs Leverage
764
249
1125
Cook's distance
Standardized residuals
Leverage
lm(Number_of_courses ~ gender + age_range + highest_education_level)


predict(fit.full, data1)
data1$Number_of_courses