CSC8631 Report

Antreas Kasiotis

07/11/2021

Business Understanding

Determine Business Objectives

Background

This project concerns an online course. As part of this project, we seek to carry out learning analytics to measure and analyse the data that regularly get collected about the learners of this course. The purpose of this analysis is to understand and optimize learning and the environment in which it occurs.

Business Objectives

1. Investigation into how the students engaged with the course. Which variables were responsible for their engagement with the course (resources used, steps completed, videos watched, etc.)? Looking at the changes that could be made (what worked well, what didn't). This could help improve the curriculum design.
2. Investigation into student performance. Understand which variables may have influenced their performance and progress (step quiz, grades, background characteristics, etc.). This could help better inform how to better support students.
3. Investigation into student background. Understand the target audience of the learners that participate in this course and identify the characteristics of those that were successful and those that were not.

Assess Situation

Resources

The resources that were provided for this project are a series of recorded log data (enrollments, arch-type survey responses, leaving survey response, weekly sentiment survey, team members, Quiz performance, step attendance, video coverage) for 7 different runs. In total, that is eight distinct data sources that could be used to draw inferences from regarding the business objectives.

Requirements

The main requirement is to use some or all of the provided sets of data to create a data analysis pipeline for a set of reliable, quick and reproducible set of tools that will be used to derive information regarding my business questions.

Assumptions and constraints

Multiple assumptions will have to be made as part of this project in terms of indicating the business objectives. Since I am not in contact with the company to discuss the aims and goals of the analysis with these data sets, I will have to determine some business objectives which I deem important and useful to investigate within the data sets.

The constraints faced by the exploratory analysis are multiple. Firstly, the eight aforementioned data sources were not recorded for every run, in some runs only a small number of sources and/or entries were recorded which limits our ability to effectively compare these runs in term of each source of recorded data. Another constraint is that is various data sets there are entries with missing values for various variables. This is a problem because it means that the samples from which we can extract for some variables are not really representative of the whole dataset and any such finding may be unreliable.

Determine data mining goals

Goals

1. Learner Background:

- Finding the educational background and personal characteristics of the learners.
- Finding if and how their background and characteristics affected their learning.

2. Learner Engagement:

- Finding methods to quantify the engagement of the students within the course.
- Finding where their engagement was attributed.

3. Learner Performance:

- Finding how the students performed in this course.
- Finding where their performance was attributed.

Success Criteria

The successfulness of my data mining project is dependent on the a variety of criteria, such as:

1. Having structured and clear business goals for what it is that I am trying to achieve from my data mining process.
2. Further dividing the business drivers down to straightforward hypotheses that can be addressed directly.
3. Thorough questioning of the reliability, completeness and relevance of the data sources used to avoid producing biased, unreliable or unimportant results.

Project Plan

Plan

My plan for this project is going to be based around the steps that formulate the CRISP-DM methodology. To begin with, I will look at all eight data sources methodically to try and understand what it is that lays within each set and how it can be used. I will discuss the data composition and how the methods that they can be used to best extract answers for my data mining questions. Furthermore, I will commence by

picking out the data that is relevant to my data mining goals and work exclusively with that. From there on I look for issues with the data that need cleaning and reformatting, this will be done to verify its quality before moving on to the next steps. I will clean that data, construct new records of it, merge it with other pre-existing data and reformat it wherever necessary so that it is ready for analysis. The next step is to take advantage for these newly prepared data sources to help me build my suite of reproducible models for analysis which will address my data mining goals.

Assessment of tools and techniques

I am going to be using multiple tools for this project.

1. Firstly, I will be using the ProjectTemplate library which allows me to automate multiple parts of the data analysis process such as organizing files, loading packages and data sets into memory and munging and preprocessing my data is a form that renders them ready for analysis(Git, N/A).
2. Another tool that I am going to making use of is Git version control that allows me to save copies of my work in progress in case I need to go back to specific point of my work (ProjectTemplate, N/A).
3. Next tool I am going to be using is the dplyr package as this is a tool that is very commonly used for data manipulation and analysis (Zev, 2014).
4. Lastly the visualization package called ggplot2 will be used to create plots at a high level of abstraction (Tidyverse, N/A).

Data Understanding

Introduction

This section seeks to describe the data that have been provided to me for this analysis as part of this project. This process will include describing, exploring and verifying the quality of each data set individually.

Initial Data Collection

The data was initially stored in a zip folder. This folder includes eight distinct data sets that each recorded an aspects of the course and/or the learner's interaction with the course. Additionally, we observe an number of repetitions in some of these data sets which represent different runs of this course. In total there are seven runs. In runs three to seven we see that all types of data sets are apparent. While on the other, in run one only six are present, with sets labeled as "team-members" and "video-stats" are missing. Moreover, in run two we see that only seven are present, with the set labeled as "video-stats" is missing again.

Enrolments Dataset

This data set has been recorded for all seven runs and it is a data set that holds information about multiple characteristics of the learner such the time they enrolled in the course, the time they finished it, their previous education and much more.

Explore Data

To further explore the contents of this data set we can see below a few rows along with their respective column names.

```
str(cyber.security.7_enrolments, strict.width = "wrap", vec.len=2)
```

```
## tibble [2,342 x 13] (S3: tbl_df/tbl/data.frame)
## $ learner_id : chr [1:2342] "f0ebc6f6-0f25-407f-a528-834414186f59"
##    "0fa1c614-8a49-42a7-a02a-8b866076d552" ...
## $ enrolled_at : chr [1:2342] "2018-10-30 15:14:09 UTC" "2018-10-25 12:23:45
##    UTC" ...
## $ unenrolled_at : chr [1:2342] "" "" ...
## $ role : chr [1:2342] "learner" "learner" ...
## $ fully_participated_at : chr [1:2342] "" "" ...
## $ purchased_statement_at : chr [1:2342] "" "" ...
## $ gender : chr [1:2342] "Unknown" "Unknown" ...
## $ country : chr [1:2342] "Unknown" "Unknown" ...
## $ age_range : chr [1:2342] "Unknown" "Unknown" ...
## $ highest_education_level: chr [1:2342] "Unknown" "Unknown" ...
## $ employment_status : chr [1:2342] "Unknown" "Unknown" ...
## $ employment_area : chr [1:2342] "Unknown" "Unknown" ...
## $ detected_country : chr [1:2342] "GB" "GB" ...
```

One thing that we can immediately see is that there are 13 columns in this set, all of which are of class character. Additionally, there are 2342 rows in this run which indicates that this many students enrolled in this course. By looking at the information held in this data set I believe that I could use it to extract some insights that would help me address my data mining goals. The kinds of questions that I could answer include:

Learner Background:

- Number of people enrolled categorized by age, gender, education, country, employment.

Learner Performance:

- Percentage of people who completed the course
- Percentage of people who completed the course categorized by age, gender, education, country, employment.

Verify Data Quality

In this data set we see that some columns have a lot of empty cells or cells that are labeled as unknown or with "–". This may cause some issues later on when we are processing the data and should be taken into consideration.

Leaving survey response Dataset

This data set has been properly recorded for runs 4 to 7 and is tracking data about the learners that leave the course. It holds information about when they left, why they left and at which point in their learning curriculum.

Explore Data

To further explore the contents of this data set we can see below a few rows along with their respective column names.

```
str(cyber.security.7_leaving.survey.responses, strict.width = "wrap", vec.len=2)
```

```
## tibble [80 x 8] (S3: tbl_df/tbl/data.frame)
## $ id : int [1:80] 153711 162741 175430 184295 187244 ...
## $ learner_id : chr [1:80] "72669fb8-cc20-4b69-ba0a-241ff767b4de"
##    "662e6b45-7695-4a2e-b395-d8c39b493d14" ...
## $ left_at : chr [1:80] "2018-07-06 10:55:39 UTC" "2018-07-23 01:27:36 UTC" ...
## $ leaving_reason : chr [1:80] "Other" "Other" ...
## $ last_completed_step_at : chr [1:80] "" "" ...
## $ last_completed_step : num [1:80] NA NA NA NA NA ...
## $ last_completed_week_number: int [1:80] NA NA NA NA NA ...
## $ last_completed_step_number: int [1:80] NA NA NA NA NA ...
```

In the snippet above we can see that eight metrics were recorded for this data set, four of which are of class character and the other four are Integer. In terms of my data mining goals, this set can help me extract some useful information, such as:

Engagement:

- Percentages of the leaving reasons for which learners left.
- Percentages of the leaving steps at which learners left.

Verify Data Quality

The data set has a many NA values in the last four columns. Additionally, we can also see that the column "leaving_reason" has some non-UTF characters.

Weekly sentiment survey Dataset

This data set has been recorded for all seven runs and it holds information about the sentiment of the learners at the end of each week in their curriculum. The recorded data include the week of the survey, the rating given, the time of response and the reason for their rating.

Explore Data

To further explore the contents of this data set we can see below a few rows with their respective column names.

```
str(cyber.security.7_weekly.sentiment.survey.responses, strict.width = "wrap", vec.len=2)
```

```
## tibble [77 x 5] (S3: tbl_df/tbl/data.frame)
## $ id : int [1:77] 60491 60882 61034 61062 61746 ...
## $ responded_at : chr [1:77] "2018-09-10 10:50:47 UTC" "2018-09-11 07:16:55 UTC"
##    ...
## $ week_number : int [1:77] 1 1 1 1 2 ...
```

```
## $ experience_rating: int [1:77] 3 3 3 2 3 ...
## $ reason : chr [1:77] "" "Im paranoid about my online privacy, this week
##    confirmed I am doing most things right and I learned a few more "|
##    __truncated__ ...
```

In the snippet above we can see the five metrics that were recorded in total. The columns "id", "week_number" and "rating" are integers, while the other two are characters. I can clearly see that this data set can help my analysis in term of understanding learner engagement. The question that can be asked about this data is:

Learner Engagement:

- How well was the course received by the learners?

Verify Data Quality

The quality of this data set seems to be fine. The only thing that may be a problem is that not all students have given a reason for their rating.

Question response Dataset

This data set has been recorded for all seven runs and it includes information about quizzes that were taken by the students.

Explore Data

To further explore the contents of this data set we can see below a few rows along with their respective column names.

```
str(cyber.security.7_question.response, strict.width = "wrap", vec.len=2)
```

```
## tibble [10,077 x 10] (S3: tbl_df/tbl/data.frame)
## $ learner_id : chr [1:10077] "77454a73-6b8b-46a2-8dee-35f36b6c4fc1"
##    "62449cd5-916b-46a6-9710-441b68d2199f" ...
## $ quiz_question : chr [1:10077] "1.8.1" "1.8.1" ...
## $ question_type : chr [1:10077] "MultipleChoice" "MultipleChoice" ...
## $ week_number : int [1:10077] 1 1 1 1 1 ...
## $ step_number : int [1:10077] 8 8 8 8 8 ...
## $ question_number: int [1:10077] 1 1 1 1 1 ...
## $ response : chr [1:10077] "1,2,3" "1,2" ...
## $ cloze_response : logi [1:10077] NA NA NA ...
## $ submitted_at : chr [1:10077] "2018-07-31 15:44:17 UTC" "2018-09-10 02:16:21
##    UTC" ...
## $ correct : chr [1:10077] "true" "false" ...
```

In total we see that there are ten columns in this set. There are many interesting metrics being recorded in this set such as the type of the question, the week it was about, the step number that the quiz held as part of the curriculum, the time of submission and lastly whether or not the answer give by the student was the correct one or not. In regards to my data mining goals, this data set is very useful as it can provide my analysis with a lot of insight. The information that can be extracted includes:

Learner Performance:

- Success rate for every quiz.
- Success rate based on the number of questions

Verify Data Quality

The only issue with that can be immediately spotted with this data set is that the column "cloze_response" is almost completely empty.

Step activity Dataset

This data set has been recorded for all seven runs and is holding information about the curriculum of each week.

Explore Data

To further explore the contents of this data set we can see below a few rows with their respective column names.

```
str(cyber.security.7_step.activity, strict.width = "wrap", vec.len=2)
```

```
## tibble [28,304 x 6] (S3: tbl_df/tbl/data.frame)
## $ learner_id : chr [1:28304] "77454a73-6b8b-46a2-8dee-35f36b6c4fc1"
##    "20e6ec35-0f50-4819-9c2e-d1851fd54638" ...
## $ step : num [1:28304] 1.1 1.1 1.1 1.1 1.1 ...
## $ week_number : int [1:28304] 1 1 1 1 1 ...
## $ step_number : int [1:28304] 1 1 1 1 1 ...
## $ first_visited_at : chr [1:28304] "2018-08-10 08:39:26 UTC" "2018-09-05
##    13:57:38 UTC" ...
## $ last_completed_at: chr [1:28304] "" "" ...
```

The are six metrics being recorded in this set, of which the two are integers, the other three are characters and there is also one that is a numeric. In this data set I see that there is a great amount of information that can be extracted and used to help me inform my data mining goals. Some of the questions that can be answered include:

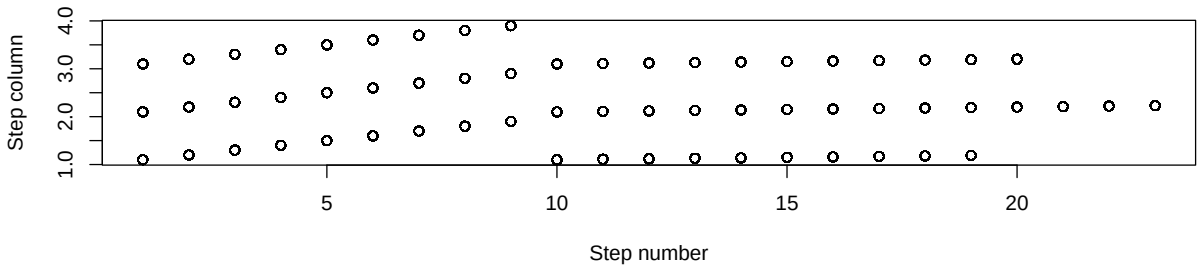
Learner Engagement:

- Percentage of learners that completed each step
- The time it took to complete each step
- The difference of both metrics across the runs

Verify Data Quality

Overall the data set looks to be of good quality. However, I have spotted that the "step" column seems a little bit problematic, so lets investigate a bit further.

```
#plotting the step (which consists of week number followed by a dot and then the step number)
#against the actual recorded step number to see if the data is correct
plot(cyber.security.7_step.activity$step_number,cyber.security.7_step.activity$step, xlab = "Step numbe
```



By looking at the plot above we clearly see that the step numbers are not correctly recorded in the "step" column. All values of step number below 10 get recorded at the wrong position after the dot.

Video stats Dataset

This data set has only been recorded for runs three to seven and it holds statistical metrics about the video material coverage of the learners.

Explore Data

To further explore the contents of this data set we can see below a few rows along with their respective column names.

```
str(cyber.security.7_video.stats, strict.width = "wrap", vec.len=2)
```

```
## tibble [13 x 28] (S3: tbl_df/tbl/data.frame)
## $ step_position : num [1:13] 1.1 1.14 1.17 1.19 1.5 ...
## $ title : chr [1:13] "Welcome to the course" "Why would anyone want your data?"
##    ...
## $ video_duration : int [1:13] 99 362 241 348 281 ...
## $ total_views : int [1:13] 1041 489 362 476 777 ...
## $ total_downloads : int [1:13] 43 16 21 21 55 ...
## $ total_caption_views : int [1:13] 14 8 8 4 12 ...
## $ total_transcript_views : int [1:13] 196 112 75 102 164 ...
## $ viewed_hd : int [1:13] 41 15 11 10 20 ...
## $ viewed_five_percent : num [1:13] 80.9 73.6 ...
## $ viewed_ten_percent : num [1:13] 79.6 71.8 ...
## $ viewed_twentyfive_percent : num [1:13] 76.2 68.9 ...
## $ viewed_fifty_percent : num [1:13] 72.4 64.6 ...
## $ viewed_seventyfive_percent : num [1:13] 69.8 61.4 ...
## $ viewed_ninetyfive_percent : num [1:13] 68.3 60.3 ...
## $ viewed_onehundred_percent : num [1:13] 66.3 57.5 ...
## $ console_device_percentage : num [1:13] 0 0 0 0 0 ...
## $ desktop_device_percentage : num [1:13] 75.3 82 ...
## $ mobile_device_percentage : num [1:13] 20.5 10.6 ...
## $ tv_device_percentage : num [1:13] 0 0 0 0 0 ...
## $ tablet_device_percentage : num [1:13] 4.03 7.16 9.12 8.19 6.05 ...
## $ unknown_device_percentage : num [1:13] 0 0 0 0 0 ...
## $ europe_views_percentage : num [1:13] 52.2 63.4 ...
## $ oceania_views_percentage : num [1:13] 2.79 4.7 4.7 3.57 4.12 ...
```

```
## $ asia_views_percentage : num [1:13] 25.6 17.2 ...
## $ north_america_views_percentage: num [1:13] 8.07 7.57 7.73 7.56 6.69 ...
## $ south_america_views_percentage: num [1:13] 2.31 2.04 2.49 2.94 2.96 ...
## $ africa_views_percentage : num [1:13] 8.65 4.5 5.25 4.62 5.79 ...
## $ antarctica_views_percentage : num [1:13] 0 0 0 0 0 ...
```

In this set we have twenty-eight metrics being recorded in total. Most of these metrics are pre-calculated percentage statistics and are classed are numerical values. While there are also six integer classes that represent amounts of learners. Lastly, there is also one character column for the title of the video. The statistics show us many things such as the percentage of people that watched a certain amount of each video, the countries they watched from and some options they used while watching such as subtitles, hd video quality, etc. This data set can be exploited to answer the following questions:

Learner Engagement:

- Percentage of video watched according to their duration.

Verify Data Quality

This appears to be okay. There are however a few things we may assume about it. First is the duration watched by learners for each video cannot account for people that downloaded that video since however much they watched, they did offline. Second assumption is videos with high percentage of watching are those with at least 95%. The reason behind this is that people may pause and close the video just a few seconds before it ends if there is nothing more to watch.

archtype survey responses

This data set was recorded for all seven runs and it holds the responses of the learners about how they view themselves in regards to what archetype the feel like.

Explore Data

To further explore the contents of this data set we can see below a few rows with their respective column names.

```
str(cyber.security.7_archetype.survey.responses, strict.width = "wrap", vec.len=2)
```

```
## tibble [174 x 4] (S3: tbl_df/tbl/data.frame)
## $ id : int [1:174] 2564612 2574521 2579047 2603632 2638826 ...
## $ learner_id : chr [1:174] "732b60fc-d132-4364-b37e-0e3a5c34f346"
##    "a45deed2-ded4-4979-b3dc-f1519edeba79" ...
## $ responded_at: chr [1:174] "2018-06-26 23:51:56 UTC" "2018-06-28 09:03:05 UTC"
##    ...
## $ archetype : chr [1:174] "Other" "Fixers" ...
```

There are four columns in this set, threes character and one integer. There seems to be nothing valuable in this set that would help in my analysis.

Verify Data Quality

The data held in this set seems fine.

Team members Dataset

This data set has been recorded for runs two to seven. The data stored in this set describe how the staff member teams were set up for the course.

Explore Data

To further explore the contents of this data set we can see below a few rows with their respective column names.

```
str(cyber.security.7_team.members, strict.width = "wrap", vec.len=2)
```

```
## tibble [13 x 5] (S3: tbl_df/tbl/data.frame)
## $ id : chr [1:13] "f27eec8c-eaf1-4e6a-90f0-d6d5b653285d"
##    "77454a73-6b8b-46a2-8dee-35f36b6c4fc1" ...
## $ first_name: chr [1:13] "FIRST" "FIRST" ...
## $ last_name : chr [1:13] "LAST" "LAST" ...
## $ team_role : chr [1:13] "host" "host" ...
## $ user_role : chr [1:13] "organisation_admin" "organisation_admin" ...
```

There are five columns in total that describe the roles of the team members and their personal information. This set also seems like it cannot help me very much in my analysis.

Verify Data Quality

The personally identifiable pieces of information have been retracted from this data set. Other than that the structural integrity of the data seems alright.

Data Preparation

Data Selection

The data sources that I am going to include are the ones that can provide me with the most relevant and useful information regarding my business objectives. To extract that information I am going to have to work with a variety of data sources. The data sources which I have found to be the most relevant for informing the engagement, performance and background of the learners are the data sets of enrollments, leaving survey response, weekly sentiment survey, quiz performance, step attendance, video coverage. The only ones that I am going to exclude are the archtype survey responses and team members as they do not hold any data that would help me address my business goals. The columns which I am going to be using for each of my sets has been mentioned in the previous section where I explain how I plan on using each set to extract information that informs my business question.

Data preprocessing, cleaning, merging and constructing

Enrolments Dataset

For this dataset I have created six new types of data sets that hold information about the enrollments and graduations separated by the age, gender, education, employment, country and a general one that included

all people. To create this six types of data sets, I have created six functions that take in a specific run's dataset and return a new one for the type of information I need. I have called each of these six functions seven times so that all of the necessary data (new data sets) for all seven runs is automatically created in the munge file for enrollments. The process of creating the functions for new data sets mainly involved separating the rows and columns of the original data according to the needs of my data mining goals and extracting and calculating only what was deemed useful. However, a big part of the dataset was taken advantage of and only a small amount of information was excluded such as the columns for purchase of statement and the country.

Question Response Dataset

For question response I created one function in the munge that takes in a data set and carries out all of the necessary preprocessing and clean automatically to return a new clear and concise data set that includes all of the needed data for answering my data mining questions. I called this function seven times in the munge file, one for each run and produced seven new data sets that are ready for analysis. The function takes care of a variety of things that are required to extract the appropriate information from the original data set. Firstly it creates two separate data frames for the numbers of answers given and the number of answers that were correct along with their respective week and step number. From these two data frames it then creates another data frame of the correct answer percentages for each step and week. Next I iterate through all of the found unique steps (quizzes) and find and store in a vector how many questions there were in each one of those quizzes. Another thing that it takes care of is the issue of the wrongful step values which need to be reformatted. To reformat the step numbers, I multiply the weeks by a hundred and add the the step numbers. Lastly, I return a newly created data frame consisting of all the vectors that are needed from my previously created vectors and data frames.

Weekly sentiment survey Dataset

For the weekly sentiment dataset I decided to create a function that creates a plot. This function only takes into consideration the qualitative data that the users gave in their survey as a reason for their rating. Essentially the function takes in a vector of character responses, cleans them by getting rid of all the stopwords that are very common and meaningless and then sorts the remaining words according to their frequency. The plot that it returns is a "wordcloud" with the most frequent words in the center with a larger font and the rest of the words around them with a smaller font and placed outwards in the plot according to their frequency levels. This function allows me to then compare the sentiment of the learners across runs.

Video stats Dataset

This particular data set is actually ready for analysis as it is. It does not require any further cleaning or preprocessing in order to answer the data mining goals set out earlier. It already has pre-calculated values for the percentages for all aspects of the variables I want to investigate.

Step Activity Dataset

As we saw in the previous chapter in the section of Verifying data quality, this particular data set had an issue with the column labeled as "step". To solve this issue in a reproducible and clean way I created a function that takes in a data set, modifies the step column and return a new clean dataset. The modification is essentially a re-calculation of the step column but this time the weeks are represented by the hundreds and the steps by the rest of the number. To prepare my data for the analysis I want to do, I created another function for this data set that again takes in a step.activity set and returns a new more concise and processed data frame. This data frame consists of all steps, the number of learner that started each step and the number that finished it, their completion percentage, and lastly the time it took them on average to

complete each step. The returned data frame provides me with all of the information I need for my analysis in order to answer my data mining questions. To make things even easier, I called both functions together for all seven runs, to clean and preprocess my data so that it is ready for when I start my analysis.

Leaving survey responses Dataset

This dataset was in need of some clean before I could extract any data from it. To clean it I had to reformat and recalculate the "leaving_step" column just as before; by representing the weeks with the hundreds and the step number with the rest of the value. This was achieved by creating a function that took in a leaving survey responses data set and return a clean one with the aforementioned changes. The was one more issues with this dataset that I had to clean, that is the "leaving_reason" column which has some typos in a few of the reasons given. To fix this issues, I tried finding each one of these problematic reasons and changing their values to the proper piece of text. However, because the typos are not of UTF-8 type, the R language could not understand them and therefore I could not select them in order to change them. Therefore, to bypass this issue, I added another line of code in my cleaning method where I used the "stri_enc_toutf8" function that replaces non-UTF-8 characters with the "REPLACEMENT CHARACTER". I then when and replaced all my faulty reasons with their now searchable "REPLACEMENT CHARACTERS" for all their non_UTF-8 values and turned them into the correct piece of text for the leaving reasons. To preprocess my information in order to create new data, I created two functions in total. The first function takes in a leaving survey response data set and finds and returns the number of people that left at each step. The second function again takes in the same type of data set and returns the number of times each leaving reason was given by the learners.

Merging

All newly constructed data sets that will derive from each run for each type of set will also have a merged version where I will have either a total or averaged version of all other runs for that particular type of set. That will be done so that analysis can be done for all available runs without having to use multiple sets. For comparisons between runs I am also going to create combined versions of each set that will hold all of the newly created information about all runs for each type of set.

Data Analysis and results

Introduction

My analysis was solely based off of my initial business questions and my Data mining goals. The business questions that I set out to investigate as part of this project included the Learner background, engagement and performance in this course. From the data sets which I had to work with for this project, I derived quite a few questions that could be investigated that concern the aforementioned business goals. These questions I decided to work with can be found in the "Data Exploration" section of this report along with a more detailed explanation of which data sets were used and why. The following sections will include explanations of the processes that were followed for my analysis and the results that were produced as part of it, as well as the cases were the analysis was not particularly fruitful. The structure of the analysis will be a step by step walk through of the work that was done for each of the data sets that I used.

Enrollments Dataset

For the enrollments data set I had stated that my investigation would look at the Learner background and the Learner performance. Below we can see two sections that detail the analysis and its results about these

two business goals. In order to make it easier for me to get a better picture of what happend in all runs, I have taken the values below by merging all runs and finding the sums of each variable. Essentially what that means is that the numbers of enrollments and graduations represent all seven runs, and so does the graduate percentage.

When looking to investigate the learner background and performance of the course I set out to look at the following questions:

Learner Background:

- Number of enrollments categorized by age, gender, education, country, employment.

Learner Performance:

- Percentage of people who completed the course
- Percentage of people who completed the course categorized by age, gender, education, country, employment.

Enrollments and Graduation percentage in all runs

```
##   run enrollments graduations graduation.percent
## 1   1       14394        1803         12.5260525
## 2   2        6488          33          0.5086313
## 3   3        3361          56          1.6661708
## 4   4        3992         166          4.1583166
## 5   5        3544          22          0.6207675
## 6   6        3175          31          0.9763780
## 7   7        2342          43          1.8360376
```

The first thing we can see is that the learners did considerably better in terms of performance in run 1, with a graduation percentage of 12.5%. Whereas, in the rest of the runs only less that 4% completed the course. Additionally, the number of enrollments was decreasing as the runs went on. This means that not a lot people were signing up for the course.

Enrollments and Graduation percentage based on age groups

```
##       age enrollments graduations grad.percent
## 1     <18          42           2     4.761905
## 2   18-25         691          29     4.196816
## 3   26-35         875          77     8.800000
## 4   36-45         608          67    11.019737
## 5   46-55         602          85    14.119601
## 6   56-65         612         110    17.973856
## 7     >65         598         137    22.909699
## 8 Unknown       33268        1647     4.950703
```

By observing the estimates for enrollments and graduations by age groups we see that the is a significant amount of people that did not provide any information about their age which may affect the reliability of the analysis. However, from the data that we have we see that the age groups with the most enrollments is that of 26-35. This is the group which seems to show the most interest in the course. Their performance however is mediocre, with just 8.8% graduating the course. The best performers were surprisingly the people who were over 65 years old, with an average across runs graduate percentage of 22.9%. When also looking at the rest of the age groups we can see that there may be a trend in our results. It seems that as the age groups go down (the younger the learners), so does the percentage of graduates.

Enrollments and Graduation percentage based on Gender

```
##      gender enrollments graduations grad.percent
## 1    female        1900         204    10.736842
## 2      male        2233         313    14.017017
## 3 nonbinary          12           1     8.333333
## 4     other          14           1     7.142857
## 5   Unknown       33137        1635     4.934062
```

In regards to gender we seem to have the same issue that might affect the reliability of the results. Just as with the age groups, where we have a significant amount of learners that did not declare their gender. From looking at the table above, the first thing we see is that by slight majority the gender with the most enrollments is that of males, with 2233 learners across runs. The gender with the best performance in regards to graduations is also the males, with an average rate of 14% graduating the course.

Enrollments and Graduation percentage based on background education

```
##       education.status enrollments graduations grad.percent
## 1       apprenticeship          15           0     0.000000
## 2  less_than_secondary          88           8     9.090909
## 3         professional         339          51    15.044248
## 4            secondary         605          66    10.909091
## 5             tertiary         361          38    10.526316
## 6    university_degree        1729         238    13.765182
## 7 university_doctorate         146          22    15.068493
## 8   university_masters         852          97    11.384977
## 9              Unknown       33161        1634     4.927475
```

In terms of previous education, we also see the same issues as before with the unrecorded entries. Apart from that, the data shows us that there is strong interest in the course by people with a university degree, with 1729 people having one. The university graduates actually also performed quite well with a graduation rate of 13.7%. When it comes to the best performance though it seems that there are two groups of people that performed a little bit better, the professionals and those with a university doctorate, both with a 15% graduation rate. This indicates just how useful it can be for a learner to have professional experience. The people who performed the worst were those who had an apprenticeship as a background education, having 0% graduation percentage.

Enrollments and Graduation percentage based on employment

```
##   employment.status enrollments graduations grad.percent
## 1 full_time_student         462          22     4.761905
## 2  looking_for_work         373          31     8.310992
## 3       not_working         230          26    11.304348
## 4           retired         717         161    22.454672
## 5     self_employed         394          43    10.913706
## 6        unemployed         171          16     9.356725
## 7           Unknown       33191        1638     4.935073
## 8 working_full_time        1402         184    13.124108
## 9 working_part_time         356          33     9.269663
```

In terms of employment, we once again see the same issue with the unknown entries. When looking at the information that we have, we see that the people who were most interested in the course are those who are

working full time, with 1402 people enrolling in the course across the runs, double the number of the second most interested group. The best performing group was the retired people with a 22.4% graduation rate. This makes sense because as we saw earlier, the over 65 were also the best performers in terms of age groups. The full time students were the ones that performed the worst, with 4.7% graduation rate.

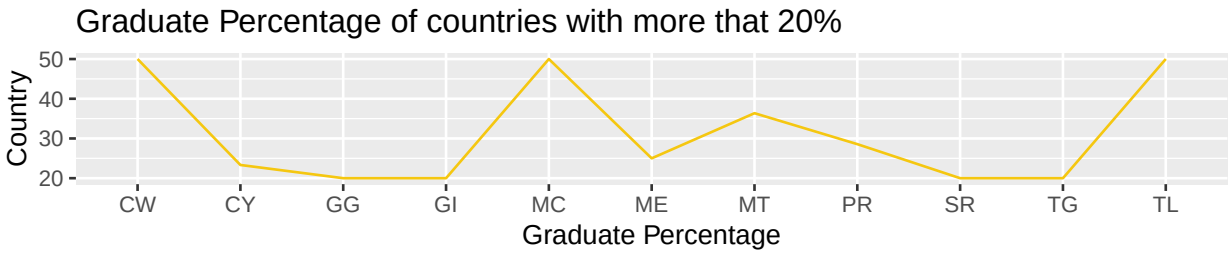
Enrollments and Graduation percentage based on country of origin

In total there were learners from 200 different countries that enrolled in the course throughout the runs. The top 10 countries with the biggest number of enrollments have been extracted and are shown below.

```
##     country enrollments
## 59       GB       11663
## 79       IN        3538
## 173      US        2117
## 144      SA        1376
## 123      NG        1002
## 10       AU         983
## 1        --         930
## 53       EG         788
## 119      MX         703
## 133      PK         635
```

We see that Great Britain, India and the US dominate the ranks of the countries from which the learner came from for this course. Learners from these 10 countries have shown the greatest interest in enrolling in the course. Additionally, we also see that the unknown entries are again quite a lot, causing the same issue as with the other metrics.

To effectively understand the performance of the learners from all the countries I will investigate the graduate percentage of countries that were more that 20%.



Above we can see 11 countries with more than 20% graduation rate. However, after further inspecting these countries I have discovered that some of them had really small number of enrollments, which can cause the graduate percentage to spike with just a few graduates. To get an ever better picture of the list of the top performing countries while also avoiding bias, I will set a minimum amount of enrollments to be 15 and bring the graduate percentage threshold down to 10%. This helps avoid presenting cases where the number of enrollments was really small and also bringing in countries were the graduate rate was above the average of run 1 which was the best one (run 1 as a benchmark).

```
##     country enrollments graduations grad.percent
## 43       CY          30           7     23.33333
## 66       GR         197          25     12.69036
## 7        AO          24           3     12.50000
## 101      LT          27           3     11.11111
## 9        AT          64           7     10.93750
```

In the list above we see a very detailed table of the countries with the best performers. Lets take a look at the top 3. At the first place we see Cyprus, with a relatively significant amount of enrollments, namely 30 but and an excellent graduate rate of just over 23%. In the second place is Greece with a much most significant number of enrollments, 197 learners and a close to average graduate rate of 12.6%. At the third place we see Angola, with 24 enrollments and a graduate rate similar to Greece at 12.5%. After that we see that the rest of the countries start to fall below the the best performing run, but still having a good overall rate.

Conclusion

The analysis from the enrollments dataset provides us with some really useful information about the characteristics of learners and their performance. To bring everything into a short summary lets create some profiles of the learners that showed the most interest in the course and those that performed the best.

Most Interested:

The characteristics of the learners that were the most interested in the course are that of a male, between the ages of 26 to 35, with a university degree, who has a full time job and those that are from either the UK, USA or India.

Best Performer:

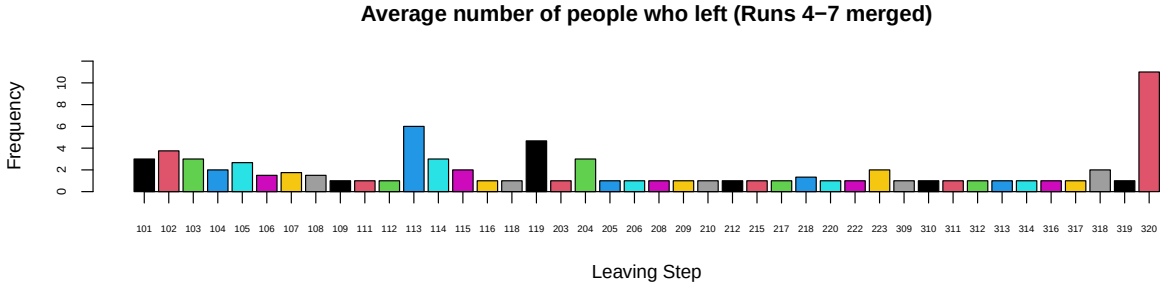
The characteristics of the learners that were the best performers in the course are primarily that of a males, people above 65, either a doctorate or professional, people who are retired and people who are from Cyprus, Greece and Angola

Leaving surevey response dataset

To make it easier for me to get a better picture of what happend in all runs, I have taken the values below by merging runs 4 to 7 and producing the means of each variable I was investigating. This was done because of the differences in the previous runs and the difficulty in merging sets with different steps. When looking to investigate the learner engagement of the course I set out to look at the following questions for this dataset:

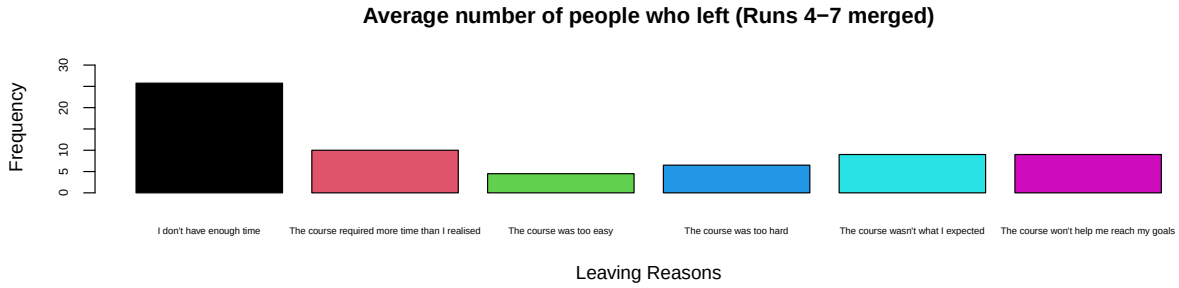
- Percentages of the leaving reasons for which learners left.
- Percentages of the leaving steps at which learners left.

Leaving Steps



The numbers that represent the steps are slightly modified. The hundreds represent the week and the rest of the number is the step number. The step at which by far the most students left is step 320 with more than 11 people leaving on average. This is really surprising since step 320 is the last step in the course. The second most usual one is 113 and then 119 which are in the middle and at the end of week 1 respectively.

Leaving Reasons



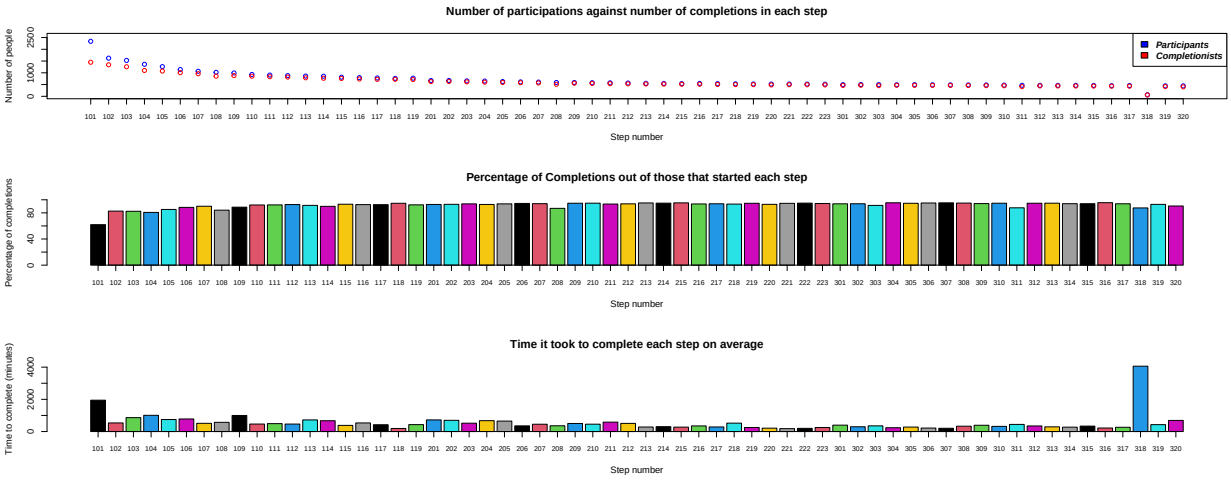
When we look at the leaving reasons it is clear that most people either mentioned that they don't have enough time or that the course required more time that they thought, which both are very similar since they are both referring to time limitations. Therefore it is safe to say that the course would benefit by being faster. To further inspect just how much time the course it taking students, I will next look at the step activity dataset which hold information about the duration of each step.

Step Activity dataset

For the steps activity dataset I have used a merged version of runs 3 to 7 as they were the only ones with the same steps. This was again done because of the differences in the previous runs and the difficulty in merging sets with different steps. I have also performed some analysis across the runs to look at the differences between them. The main data mining goals for this dataset encompassed the following questions that concern Learner Engagement:

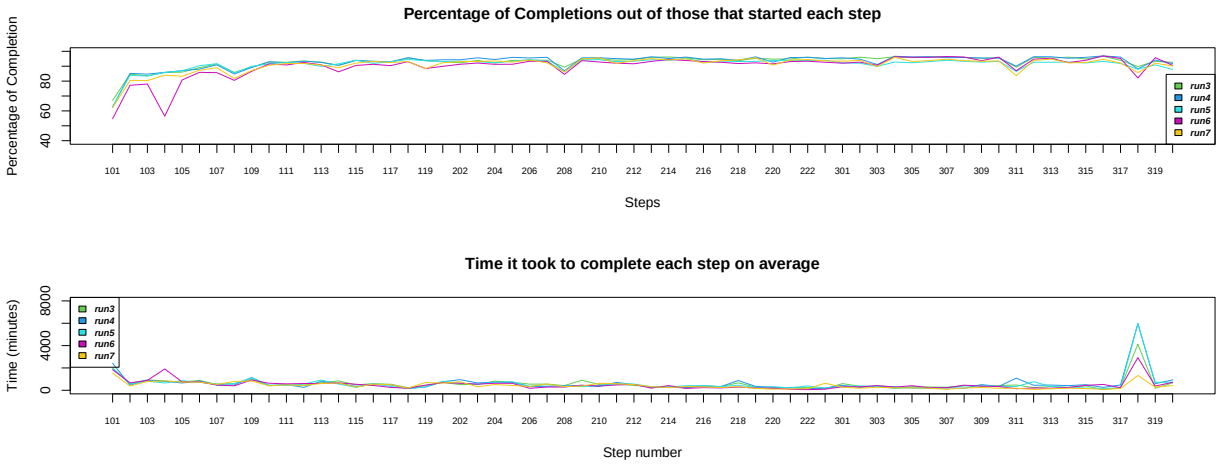
- Percentage of learners that completed each step
- The time it took to complete each step
- The difference of both across the runs

Step engagement and time to complete



The plot above shows us three different things. First is the number of people that started each step and those that completed it. Here we see that as the steps progressed, less and less people were covering them. The second figure show a more clear view of the percentage of what is shown in the first one, the percentage of people that completed each step. This essentially gives us a value that shows how engaging each step was, because if a learner gave up midway it means that something was wrong with the step. It could be that was too hard to understand, too long or simply boring, among other things. In the second plot we observe that overall the steps in the beginning and those in the ends were the ones where most people were giving up. Although we have to mention that the difference in engagement in these steps is not huge. Lastly, we can see the third figure which shows the time on average that it took students to complete each step. In this plot, most of the steps have relatively similar values, except for a few that really stand out. Steps 318, 101, 109 and 104 have significantly higher values in term of the time it took student to complete them. When we compare these steps with the second figure we can also see that the engagement levels were also lower. This would indicate that when a step takes too long to finish it leads students to give up on it more easily.

Step engagement and time to complete (runs 3 - 7)



Another interesting observation I have made during my analysis is that when we check these same metrics across multiple runs we see that run 6 has some significantly outlying values for step 104. When looking into this step, I found that it was an exercise which could mean that this exercise was altered in run 6 and caused some problems.

Weekly Sentiment dataset

Not a lot of data was available for this dataset so I decided to use only run 6 for my analysis. My data mining goal for this dataset was to investigate the following question about Learner Engagement:

- How well was the course received by the learners?

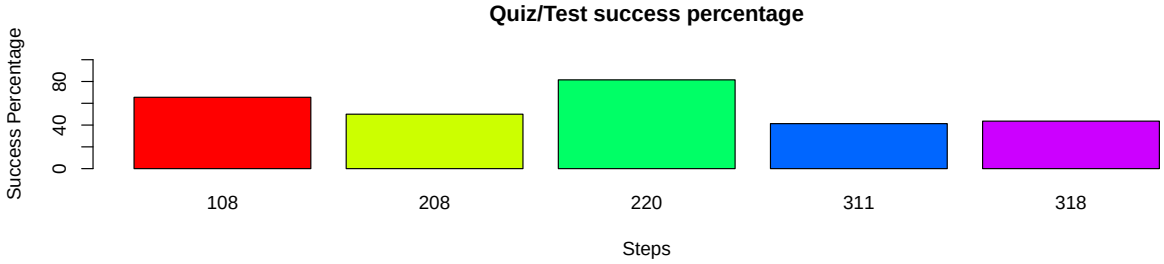
The way I decided to approach my analysis was by investigating the text reason that the students gave. To analyse qualitative data I manipulated the data in such a way that allowed me to see the most frequent words from the responses in the form of a “wordcloud”.

After removing a few stopwords that do not give real meaning, the remaining word could seems to show a positive sentiment. Learners have been giving nice adjectives and comments about the course.

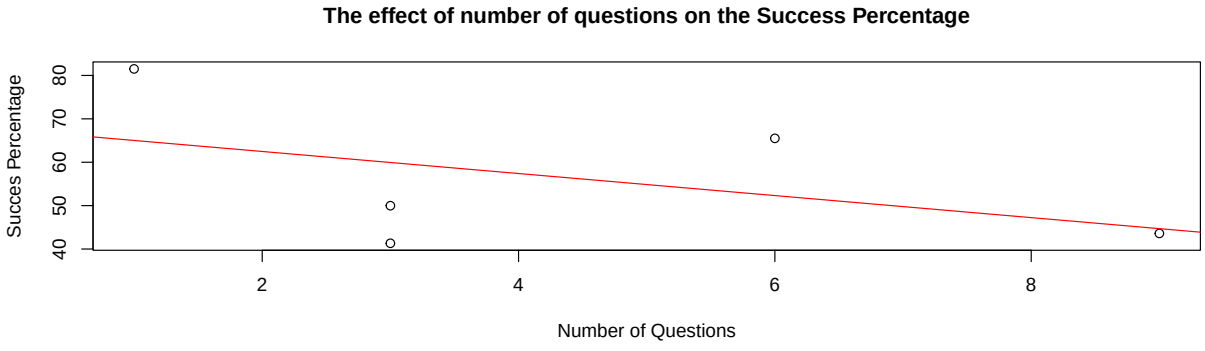
Question response dataset

For the question response dataset I have checked the percentage of correct answers given to investigate the difficulty of each quiz for runs 2 to 7. The cross-run analysis showed that the runs had similar results. The primary goals of my analysis was to answer the following questions that concern the Learner Performance:

- Success rate for every quiz.
- Success rate based on the number of questions



The most successful quizzes seem 220 and 108, which had 80% and 65% of their responses correct. Whereas the the quizzes 311 and 316 which are both in Week 3 have very low success rates. This could be an indication that the material covers before the successful quizes was better understood by the learner, in contrast to that of week 3.

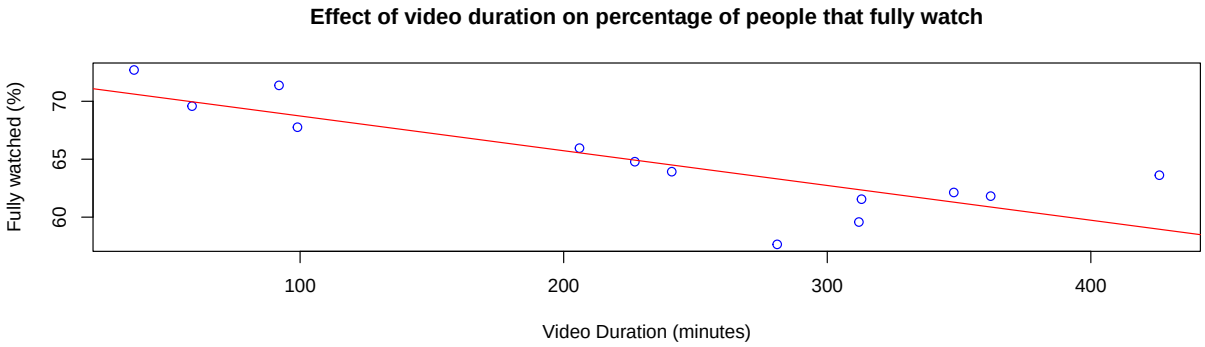


When investigating the the success percentage according to the number of questions a quiz had, we observe that there seems to be a relationships that indicates that the more questions a quiz has, the less the success percentage. Although, it is worth mentioning that the points lay quite far from our trend line, therefore the relationship is not that strong.

Video Stats dataset

To make my analysis more reliable, I have merges the runs 3 to 7 into one averaged one. As per my data mining goals I had decided to investigate the learner engagement from the video statistics dataset by looking at the following question:

- Percentage of video watched according to their duration.



As we can see from the plot above, there is a clear correlation between the duration of the videos and the percentage of people who watched the whole thing. This relationship shows us that the longer the video is, the more the chances that a learner will stop watching.

Conclusion

With the analysis of the data sets now complete we have come to the end of this project. Through the data understanding, preparation and analysis I have managed to extract some really useful insights about the Future Learn Cyber Security Course. The business goals which I had originally set out as the framework of my work have now been fully satisfied. Information has been produced to answer questions from 6 out of the 8 different types of dataset, for a wide selection of runs. The data mining goals of Investigating how and why the learners engagement and performance was attributed has been a complete success. Furthermore, a thorough investigation into the background characteristics and the effect has provided us with eye-opening results about the target audience of this company.