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 face 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 were recorded which limits our ability to effectively compare these runs in term of each source of recorded data. Another constraint is various data sets, there are entries with missing values for various variables. This is a problem because it means that the samples, we can extract for some variables are not really representative of the whole dataset.

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 with 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 how the data composition and how the methods that they can be manipulated to best extract answers for my data mining goals. Furthermore, 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. From there on, I will commence by picking out the data that is relevant to my data mining goals and work exclusively with that. 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 use by the models I am going to build. The next step is to take advantage for these newly prepared data sources to help me build my suite of reproducible models for analysis regarding the 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

Describe Data

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 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

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

Describe Data

This data set has been recorded for all seven runs 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.

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_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.
- Further investigate the trends among steps (if relevant to the course) at which they left.

Verify Data Quality

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

Weekly sentiment survey Dataset

Describe Data

This data set has been recorded for all seven runs and it hold 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 each week 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. However this is something that can be bypassed if we simply ignore that "reason" column.

Question response Dataset

Describe Data

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 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 week.
- 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

Describe Data

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:

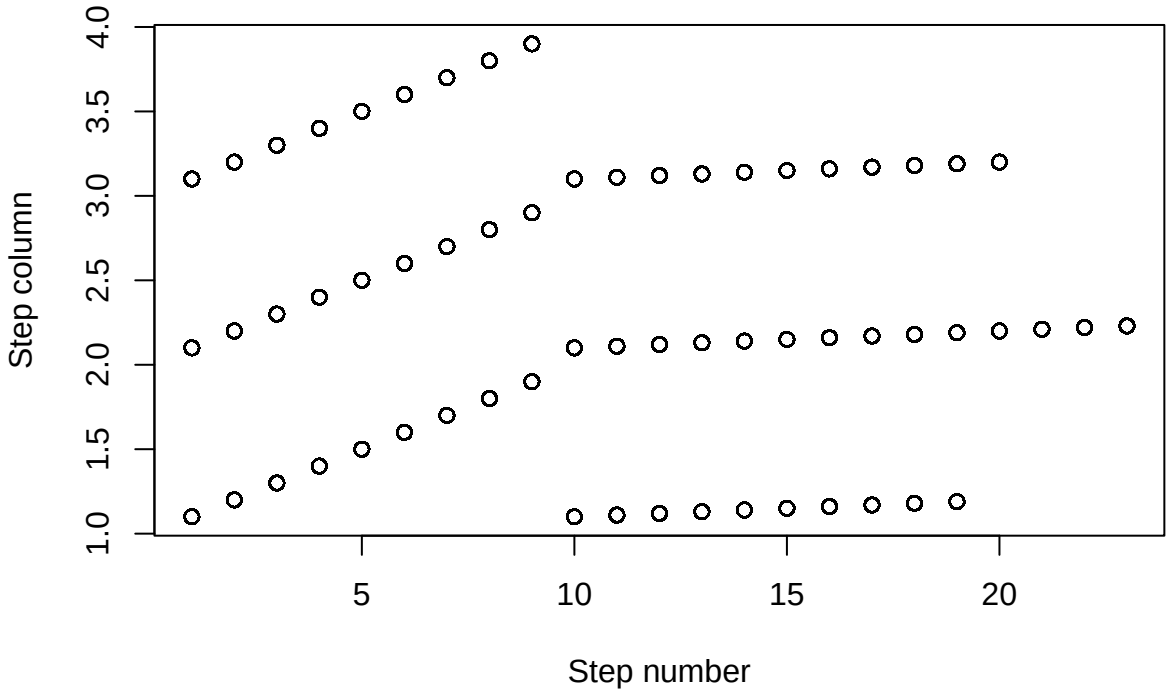
Learner Engagement:

- Percentage of learners that completed each step
- Percentage of learners that completed each week
- Percentage of learners that completed each type of learning material

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.

Video stats Dataset

Describe Data

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 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:

Engagement

- Percentage of video watched according to their duration.

Target Audience

- Percentage of viewers that used mobile, desktop, tv, tablet.
- Percentage of viewers from each continent.

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

Describe Data

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. Only thing to add is that it is empty in runs one and two.

Team members Dataset

Describe Data

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.

Data preprocessiong and cleaning

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 analyis 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.

Data Analysis