Data Management and Exploratory Data Analysis

Buisness understanding

Business objectives

This report is focusing on Learning Analytics a fast developing domain of Data Science. In Learning Analytics organizations can extract context and better understanding of their programs. The goal is to derive suitable business results while improving employee productivity. The discoveries made could allow learning organizations to pick and choose learning models that works best, resulting to their organizations increase profit while lowering the time spent in developing the learning material. The organization in question provides an online course on cyber security, they have conducted 7 runs for this course. Further discussion on the course as a whole and differences between some of the runs will take place. The main goal is to produce findings derived from the data collected as well as further investigation on some of the most noticeable and interesting topics.

Assess situation

Learning Analytics refers to the collection and analysis of data about learners and their environments. Therefore, it is significant to create tools that provide and in depth understanding of the learners background undertaking the course. The scope of this analysis is improve the learning outcomes of the course both as s business standpoint and how its best received by the learners.

Determining data mining goals

In educational data mining it is important to classify learner performance and engagement. Furthermore, it is important to assess the material provided and the levels of engagement their provided. Some of the data mining goals at this stage are:

1. Learner background information.
2. Extraction of learner course completion across all runs.
3. Assessing the course material that provide a sense of performance.
4. Analysis on the data that provide further context in terms of learner engagement.
5. Various useful information extracted by the data mining process

In addition, a successful Data Mining process will be derived by:

- The reliability of the data.
- The importance of the results in terms of context, clarity and usefulness.
- The reliability of the processes conducted to extract results.
- The appropriate method used to provide a coherent discussion,

Project plan

The tools and techniques that are going to be used include Git version control, ProjectTemplate for R and RMarkdown. Git version control is a system that allows to keep track the changes made to the code over time. As such, collaboration is made possible because specific changes on the work are tracked and tagged by various contributors. The work for this project will be conducted by the author of this report, allowing for further expansion of the work with other collaborators in the future. ProjectTemplate in R allows for the automation of new statistical analysis projects. In addition, it provides a directory structure that automate data loading, preprocessing, library importing and testing. Finally, RMarkdown is a versatile open source markup language and can be used to format plain text while enabling the user to directly post their work on web pages if they must.

The project plan consists of the tools and techniques mentioned above to dwell into the data provided from the organization and produce a coherent structured report. The project plan includes the data understanding divided by some appropriate exploration and discussions. Next,is data preparation where the cleaning and construction of the data will take place. Finally, the modelling where in depth analysis will commence. The project plan will be layed out in detail in the next chapters of this report.

Data Understading

Collection of initial data

The data are collected and loaded in the ProjectTemplate environment which has been set as a working directory using Git Bash. With the data loaded comes 7 pdf files containing the format of the material of each run commenced.

Data Description

The data that will be utilized consists of 53 csv files. The runs does not share the same number of csv files it noticeable that are inclusions and exclusions when each run commenced. This will be further explored in the data exploration.

The distinct csv files across all runs are as follows:

1. Archetype Survey responses
2. Enrollments
3. Leaving survey responses
4. Question responses
5. Step Activity
6. Weekly sentiment survey responses
7. Team members
8. Video stats

Decription of the data sets content

The archetype data set consist of the learner and an archetype answer regarding their general behavior. The enrollment data set consists of enrolled learners and 13 variables that provide numerous information such as enroll dates, fully completion dates, gender and more. Leaving survey responses includes of course the reason the learners gave for abandoning the course. The question responses data set consists of detailed learner performance regarding quizzes provided throughout the course. The step activity includes step progression for each week as well as started and completion dates for each step in the course. The weekly sentiment responses include weekly feedback from the learners regarding weekly progress. The team members data set consists of description regarding the roles of the teaching team. Lastly, The video stats data set consist of various statistics regarding the video material on the course for example views, percentages watched, devices, geolocation and more.

Data exploration

By exploring the data several questions surfaced for a portion of the data sets provided. Although, all data sets where explored and considered for further analysis. The data sets that was deemed appropriate for the outlined goals and business objectives includes the enrollment, question responses, step activity, and video stats. The reasoning will be discussed in detail on the the data selection and construction chapter of this report.

Enrollement data set (Run 1)

The enrollment data set consists of the learners as our observations number and each learner has 13 variables.

```
head(cyber.security.1_enrolments)
```

```
## # A tibble: 6 x 13
##   learner_id  enrolled_at  unenrolled_at role  fully_participa~ purchased_state~
##   <chr>       <chr>        <chr>         <chr> <chr>            <chr>
## 1 160d6600-e~ 2016-08-10 ~ ""            lear~ ""               ""
## 2 4dc22fed-6~ 2016-05-24 ~ "2018-10-30 ~ lear~ ""               ""
## 3 ecdd37db-0~ 2016-05-19 ~ ""            lear~ "2016-09-22 16:~ ""
## 4 988964c9-7~ 2016-05-19 ~ ""            lear~ ""               ""
## 5 f1493366-1~ 2016-09-19 ~ ""            lear~ ""               ""
## 6 25cc3b46-a~ 2016-08-30 ~ ""            lear~ "2016-10-25 12:~ ""
## # ... with 7 more variables: gender <chr>, country <chr>, age_range <chr>,
## #   highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

All 13 variable are stored as characters it noticeable that numerous rows are empty and some are stored as Unknown.

Question response data set (Run 1)

The question response data set consists of the learners as our observations number and for each learner there are 10 variables

```
head(cyber.security.1_question.response)
```

```
## # A tibble: 6 x 9
##   learner_id quiz_question question_type week_number step_number question_number
##   <chr>      <chr>         <chr>               <int>       <int>           <int>
## 1 77454a73-~ 1.7.1         MultipleChoi~           1           7               1
## 2 77454a73-~ 1.7.1         MultipleChoi~           1           7               1
## 3 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 4 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 5 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 6 f27eec8c-~ 1.7.1         MultipleChoi~           1           7               1
## # ... with 3 more variables: response <chr>, submitted_at <chr>, correct <chr>
```

Some variables are stored as characters namely learner_id, quiz_question, question_type response, submitted_at, correct. The rest are stored as integers expect cloze_response that is logical and is filled with NAs.It is advicable to further investigate regarding inconsistencies.

Step activity data set (Run 1)

The step activity data set consists 143092 observations with 6 variables.

```
head(cyber.security.1_step.activity)
```

```
## # A tibble: 6 x 6
##   learner_id     step week_number step_number first_visited_at  last_completed_~
##   <chr>         <dbl>       <int>       <int> <chr>             <chr>
## 1 77454a73-6b8~   101           1           1 2016-08-02 13:45~ ""
## 2 c1a75ae7-c76~   101           1           1 2016-08-02 15:40~ ""
## 3 a4fa6f89-a59~   101           1           1 2016-08-03 07:14~ ""
## 4 60b56cea-ad2~   101           1           1 2016-08-03 08:45~ "2016-08-03 08:~
## 5 05a815ce-3c4~   101           1           1 2016-08-03 13:24~ "2016-08-03 13:~
## 6 5553e67a-178~   101           1           1 2016-08-03 14:49~ "2016-08-04 15:~
```

Some variables are stored as characters namely learner_id, first_visited_at and last_completed_at the others are stored as integers except the step variable which is double.

Video stats data set (Run 3)

The video stats data set consists of 13 observation which refer to the video step position over the 3 weeks of the course and 28 variables. It's important to mention that the video stats data is not present in the first 2 runs.

```
head(cyber.security.3_video.stats)
```

```
## # A tibble: 6 x 28
##   step_position title                 video_duration total_views total_downloads
##           <dbl> <chr>                          <int>       <int>           <int>
## 1          1.1  Welcome to the course             99        1659             113
## 2          1.14 Why would anyone wan~            362         910              77
## 3          1.17 Preserving privacy i~            241         723              63
## 4          1.19 Staying safe online:~            348         755              62
## 5          1.5  Privacy online and o~            281        1248             100
## 6          2.1  Welcome to Week 2: p~             37         694              48
## # ... with 23 more variables: total_caption_views <int>,
## #   total_transcript_views <int>, viewed_hd <int>, viewed_five_percent <dbl>,
## #   viewed_ten_percent <dbl>, viewed_twentyfive_percent <dbl>,
## #   viewed_fifty_percent <dbl>, viewed_seventyfive_percent <dbl>,
## #   viewed_ninetyfive_percent <dbl>, viewed_onehundred_percent <dbl>,
## #   console_device_percentage <dbl>, desktop_device_percentage <dbl>,
## #   mobile_device_percentage <dbl>, tv_device_percentage <dbl>, ...
```

In the stat video we have variables stored as characters for the description of the videos,integers for counts and doubles for percentages.

Leaving survey responses (Run 5)

In the leaving survey response data set there are 174 learners with 8 variables.

```
head(cyber.security.5_leaving.survey.responses)
```

```
## # A tibble: 6 x 8
##      id learner_id   left_at  leaving_reason   last_completed_~ last_completed_~
##   <int> <chr>        <chr>    <chr>            <chr>                       <dbl>
## 1 34003 8853543d-b9~ 2018-01~ I prefer not to~ ""                             NA
## 2 38604 b170480c-7e~ 2018-02~ The course requ~ ""                             NA
## 3 39016 92b8485e-b2~ 2018-02~ Other            ""                             NA
## 4 39241 f4cae359-09~ 2018-02~ Other            ""                             NA
## 5 39604 3429e915-a6~ 2018-02~ I prefer not to~ ""                             NA
## 6 39657 a04b5f29-c2~ 2018-02~ Other            "2018-02-05 18:~              1.2
## # ... with 2 more variables: last_completed_week_number <int>,
## #   last_completed_step_number <int>
```

By reviewing the data it is noticeable that learners that left the course are assigned new IDs regarding their leaving survey that are stored as integers. Furthermore, The last_completed step variables are stored as doubles while the last_completed_step_number are stored as integers. NAs can be spotted instantly further investigation will be required to assess the quality of the set.

Weekly survey responses data set (Run 6)

The weekly survey responses consist of 80 observations (learners) and 5 variables.

```
head(cyber.security.6_weekly.sentiment.survey.responses)
```

```
## # A tibble: 6 x 5
##      id responded_at            week_number experience_rating reason
##   <int> <chr>                         <int>             <int> <chr>
## 1 16810 2018-06-11 18:28:04 UTC           1                 3 ""
## 2 17388 2018-06-12 16:38:12 UTC           2                 3 "It was good i ha~
## 3 17505 2018-06-12 20:24:38 UTC           2                 3 "Nice Content !!"
## 4 17807 2018-06-13 12:05:35 UTC           1                 3 ""
## 5 17819 2018-06-13 12:25:05 UTC           1                 3 ""
## 6 18277 2018-06-14 10:31:48 UTC           1                 3 ""
```

Similarly with the leaving responses they are assigned a survey ID, variables are stored as characters for the responded_at and the reason. The week_number and the experience_rating are stored as integers. It noticeable that some responses are missing for rating reason.

Verification of data quality

From the exploration in most data sets inconsistencies were noticed. Specifically, same type of variables where stored as different types, multiple empty rows where discovered, some rows where labeled Unknown. In addition, multiple data sets contain NAs, some columns were empty and in some of the responses surveys the characters contained symbols.

This report is going to cover specific topics and will be using a portion of the data sets available to explore and perform an in depth analysis. After the selection of relevant data sets and rationale for inclusion and exclusion, the cleaning process will be documented where is applicable. In the question response data set repetition of entries were discovered for some of the questions, the assumption is that the learners were able to resubmit the quiz question if they made a mistake so the focus is shifted to the quiz quality rather than learner performance. In the step activities data set a mistake was discovered regarding how the steps number variable were recorded and stored, appropriate transformation will take place.

Data preparation

Selection of data

In this chapter an explanation will be given for the selection of the data that was previous mentioned in data exploration. Each data set will be discussed and goals will be outlined for the production of the analysis. Furthermore, steps taken for cleaning, construction, integration and reformation will be discussed. It is important to state that the data preparation takes into account all the goals and methods outlined above. Namely, general statistics, engagement, performance and various elements that will be derived from the analysis.

Enrollment

In the enrollment data set it is possible to extract a count of the people enrolled and their background information. In addition, calculation of the percentage of the learners that fully participated in the course for the each run is possible. This will show a scale of popularity among the runs and completion rates. Also, some generic statistics about the learners will be generated

Analysis goals:

1. Learners information
2. Course Engagement rates

Questions responses

The question response data set contain variables regarding quiz performance for each learner. Quizzes are a satisfactory way to access learner performance regarding the teaching material provided across 3 weeks for each run were the data quality and volume is sufficient.

Analysis goals:

1. Success rate based on each week.
2. Correlation and failure statistics regarding the material.

Step activity

Step activity is a particularly useful data set because it contains details about each step across the three weeks of the course. It is possible to derive results that would provide vital information on the teaching material and the reception it received from the learners.

Analysis goals:

1. Time it took learners to complete each week.
2. Percentage of learners that completed every step.
3. Percentage of learners that completed each each week. 4 Completion ratio for each step.
4. Most popular step type discussion.

Video stats

The video stats data set enables the analysis to dwell deeper into this specific step type, getting context and statistics of how well this step type was received by the learners as well as how it was conducted by the teaching team. Several assumptions could be taken beforehand that will be discussed in the analysis section.

Analysis goals:

1. Percentages of video views for each step.
2. Percentages of viewers filtered by device utilized
3. Percentages of viewers from each continent.

Data cleaning and data construction

###Enrollement cleaning

By checking for NAs in the enrollment data set it returns 13 NAs.By running the code:

```
cyber.security.1_enrolments= cyber.security.1_enrolments[rowSums(is.na(cyber.security.1_enrolments)) ==
sum(is.na(cyber.security.1_enrolments))
```

```
## [1] 0
```

All NAs from the data set were omitted. in addition where applicable its possible to remove only rows that are fully NA using:

```
cyber.security.1_enrolments[rowSums(is.na(cyber.security.1_enrolments)) != ncol(cyber.security.1_enrolm
```

```
## # A tibble: 14,382 x 13
##    learner_id  enrolled_at unenrolled_at role  fully_participa~ purchased_state~
##    <chr>       <chr>       <chr>         <chr> <chr>            <chr>
##  1 160d6600-e~ 2016-08-10~ ""            lear~ ""               ""
##  2 4dc22fed-6~ 2016-05-24~ "2018-10-30 ~ lear~ ""               ""
##  3 ecdd37db-0~ 2016-05-19~ ""            lear~ "2016-09-22 16:~ ""
##  4 988964c9-7~ 2016-05-19~ ""            lear~ ""               ""
##  5 f1493366-1~ 2016-09-19~ ""            lear~ ""               ""
##  6 25cc3b46-a~ 2016-08-30~ ""            lear~ "2016-10-25 12:~ ""
##  7 9c23a086-f~ 2016-06-22~ ""            lear~ "2016-10-10 11:~ ""
##  8 8851dc49-0~ 2016-08-07~ ""            lear~ "2018-10-17 18:~ ""
##  9 a59b0a12-a~ 2016-08-02~ "2018-10-17 ~ lear~ ""               ""
## 10 198c1017-5~ 2016-09-09~ ""            lear~ ""               ""
## # ... with 14,372 more rows, and 7 more variables: gender <chr>, country <chr>,
## #   age_range <chr>, highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

Next checking for duplicates in the learner_id column:

```
cyber.security.1_enrolments$learner_id[duplicated(cyber.security.1_enrolments$learner_id)]
```

```
## character(0)
```

Enrolment data constuction for analysis

For the enrollment a new data frame is constructed containing the learners that fully participated in the course. Furthermore, this new subset is filtered by gender, education and employment status. Lastly, the same filters are applied in the full data set.

Question response cleaning

Firstly the empty cloze_response column is removed. Then we check for NAs it appears there are none:

```
sum(is.na(cyber.security.1_question.response))
```

```
## [1] 0
```

QUestion response data constuction for analysis

For the question response data set 3 data frames are constructed filtered by the correct answers for the questions for each week respectively. This will later utilized to compare performance in terms of success. Furthermore, the worst performing week will be analyzed further. The results will be stored inside a table.

Step activity Cleaning

The step activity has an issue on the step number variable. Specifically is not possible to distinguish the 1.1 step from 1.10, to fix that problem multiplication by 100 took place so now 101 refers to step 1.1 and 110 to step 1.10. The data set does not contain any NA rows. In terms of quality the step activity can be considered appropriate for the work intended.

Step activity data construction for analysis

The step activity data sets are split into 3 sets for each weak, the goal is to derive completion percentages for all steps in each week and then investigate all steps on the best performing week. The resulted percentages will be stored inside a table to be utilized in the analysis.

###Video stats cleaning In video stats data set the last column is empty and will be removed but in general the set does not contain any NA rows or other inconsistencies that were discovered previously the data stored are ready to be utilized for analysis.

Video stats data construction for analysis

For the analysis the total_views column is needed alongside the video_duration and the step_position to investigate engagement e.g. the relation of the video lengths and the views for each step. Also, the columns regarding the devices used for viewing will be kept to provide an understanding of the viewing circumstances. Lastly, the columns regarding the continents will be used to provide general statistics regarding the videos.

Three new data frames were constructed for the views, devices and continents.