CSC8631 - Data Management and Exploratory Data Analysis Cyber Security Course In FutureLearn Provider Analysis

Sandra M Nino

2023-11-17

First Cycle of CRISP-DM

1. Business Understanding

1.1 Business objectives

Background

Massive Open Online Courses (MOOC) have become popular since a few years ago providing, according to Gartner, free access to online resources, social networking and interactions of experts in a field of study, and thus a massive number of registered students across the world who want to switch careers, gain a promotion, improve job prospects and/or just keep learning.

There are many MOOC providers such as Coursera, edX, Pluralsight, FutureLearn and so on, which facilitate the potential of launching online courses to a massive global audience. Therefore, these providers have reached out to universities to put their courses on their platforms with the aim of revolutionary education. However, "MOOC providers are not an alternative to traditional education, but a strategic partner" (Shah, 2017).

Newcastle University has not been the exception in trying this strategic partnership with MOOC providers. For instance, the university and the School of Computing launched for seven times a MOOC called "Cyber Security: Safety At Home, Online, and in Life" between 2016 and 2018 in the skills provider FutureLearn. This three-week free online course explored cyber security including privacy online, payment safety and security at home from different perspectives (the user's, a potential attacker's and a business's), with an expected weekly hours commitment of 3 hours. The course was intended for people with basic knowledge of cyber security, some IT background or a genuine interest in the practice of cyber security. The course was conducted by Steve Riddle, a Computing Science lecturer at Newcastle University and programme director for MSc in Computer Security and Resilience.

MOOC providers would not exist without students. Thus, in 2018, FutureLearn published the results of a survey of 7000 learners conducted by FutureLearn's User Experience Research Team to explore who its learners are (Pickard, 2019). In this way, they identified seven learner archetypes which group people with similar characteristics and behaviours to understand their needs when learning.

The seven archetypes are grouped into three types:

1. **Work and Study:**

- **Advancers**: They are in their chosen career path. Identified as long life learners and they want to stay up to date with the latest research and trends.
- **Explorers:** They are evaluating options to make informed decisions on their next step in their career path. They might be looking to change careers, where to start in their working lives, specialise in an area or return to education. Their needs are focused on engaging courses, ways to build self-confidence, to reassure their chosen career path, to being able to switch into different topics to see what suits best, offline courses, and to interact with educators to solve their questions.

- **Preparers:** These learners are in early careers or starting in a new job. They are clear on what to do next to improve their career paths and stand out from others.

2. **Personal Life:**

- **Fixers:** They seek to learn with the aim of fixing current aspects of their personal lives. This includes topics such as physical and mental health, life skills, political or cultural issues, life changes like maternity or retirement.
- **Flourishers:** Their aim is to learn to be happy and healthy in their personal and professional lives. They are interested in building self-steem, keeping brain active, sharing their thoughts or improving their physical and mental health.

3. **Leisure:**

- **Hobbyists:** They want to keep learning to support their current personal projects and leisure activities.They look for courses when they actually need it and like to network to share their findings.
- **Vitalisers:** These learners like to learn as a hobby and for the love of keep learning. They learn anything of their personal interest because they enjoy learning to keep their brains active and use time appropriately. Their primary interested is not focused on communicating with other people. Their needs are focused on diverse and interesting courses, to feel stimulated when learning, to have accessible content and good reference materials, and to feel the good use of time.

Business objectives

Having in mind the increasing demand of MOOCs to keep in the loop of revolutionary education, Newcastle University and the School of Computing, as our main stakeholders, want to understand who are their learners in the FutureLearn platform to design future cyber security courses for their principal learner archetypes based on the definitions made by the FutureLearn provider. Therefore, by considering the learner in the course design, there will be potentially more engaged learners with the course and the university because their needs, personal and professional goals, and career paths are taken into account.

Business success criteria

The success criteria for our business objective is to give useful insights about the types of learners who enrolled in the cyber security course over the seven runs. The judgement for this success criteria will be made by our main stakeholders.

1.2 Assess Situation

Inventory of resources

- Personel:
  - Sandra M Nino - Data expert
  - Joe Matthews - Business expert
- Data: We have available the raw data collected by FutureLearn for the seven runs of the cyber security course in CSV format which have information about the learners based on their profiles.
- Software: We are using RStudio Version 2023.06.2+561. We have available libraries such as Project-Template, dplyr and ggplot2.

Constraints

The archetype survey held by FutureLearn in 2018 just considered 7000 learners in the whole platform. Therefore, we do not have the results of the archetype surveys for all the learners who enrolled in the cyber security course over the seven runs, meaning that we have few data that was just recollected from the third run to the seventh run; we do not have data from the first two runs of the course.

1.3 Data Mining Goals

Data Mining Objective

The data mining objective is to compare the proportions of the learners in the cyber security course and to look the differences in demographic characteristics which identify each learner archetype.

Data Mining Success Criteria

The data mining success criteria is to provide an analysis with visual representations to our stakeholders to understand the proportion of the learner archetypes for the cyber security course in order to design more engaging courses in the FutureLearn platform. In this way, Newcastle university and the School of Computing can make informed decisions based on the report produced.

1.4 What Investigations Are We Carrying Out?

To achieve our business objective we will investigate the following:

- For our first cycle we are interested in answering how the archetype influences in the enrolment and certification purchase of the cyber security course over the different runs.
- The results from this first investigation will allow us to select the most common archetypes for the course, thus in the following cycle we will check the distribution of enrolment and certification purchase against gender, age range, employment status and education level.

1.5 Project Plan

We are going to execute the project based on agile practices. We are going to start by dividing our main task into smaller pieces to reach our goals. Thus, we will use the tool called Trello to track the progress of the tasks by updating their status between to do, in progress, blocked, and done. In this way, we can identify dependencies between tasks, blockers, and time spent on each task.

2. Data Understanding

2.1 Initial Data

We have raw data collected by the FutureLearn provider. In this way, we have six common CSV files with different type of data for all the seven runs of the cyber security course, which means we have available forty-two CSV files. Also, for the seven runs except the first one, we have files about the team members' information, which accounts for six CSV files. In addition, for the seven runs excluding the first two runs, we have files about the video statistics used in the course, which means five more files. In total, we have fifty-three files. We also have seven PDF files with the course overview for each run. However, they are just informative and they will not be used in the analysis. All these files (CSV and PDF) are stored in the data folder of the project.

2.2 Data Description

As mentioned previously we have different CSV files for each of the seven runs which contains diverse information about our learners in the cyber security course. The format of the files and what information includes are mentioned below:

- cyber-security-{number of run}_archetype-survey-responses: This includes information about the responses of the survey conducted by FutureLearn to identify the learner archetype.
- cyber-security-{number of run}_enrolments: Includes information about the enrolment in the course and demographic characteristics of the learner.
- cyber-security-{number of run}_leaving-survey-responses: Information about when the learner left the course and the status of their progress in the course, as well as the reason of leaving.
- cyber-security-{number of run}_question-response: Information about the quizzes in the course, the learner response and its correctness.

- cyber-security-{number of run}_step-activity: Information about the step of the activities and when it was completed.
- cyber-security-{number of run}_team-members: Information about the team members and their roles in the platform.
- cyber-security-{number of run}_video-stats: Information about the reproduction statistics of the video materials in the course.
- cyber-security-{number of run}_weekly-sentiment-survey-responses: Information about the rating of the learning experience and comments about it.

To achieve our business and data mining goals, we are going to explain in more detail the content of the files we are actually going to use for our analysis:

1. cyber-security-{number of run}_archetype_survey_responses:

- id: Identification of the survey response
- learner_id: Identification of the learner. This variable will be useful to combine the data with other files.
- responded_at: Date the survey was responded
- archetype: The result of the learner archetype based on the responses of the survey. This variable is the core of our analysis, so it is extremely important.

2. cyber-security-{number of run}_enrolments:

- learner_id: Identification of the learner. This variable will be useful to combine the data with other files.
- enrolled_at: Date of enrolment
- unenrolled_at: Date of unenrolment.
- role: Type of role in the platform. For our analysis, we have to make sure that the roles are just learners
- fully_participated_at: Date of participation in a specific activity
- purchased_statement_at: Date of certificate purchase. This variable will allow us to know if the learner purchased or not a certificate for the course.
- gender: Gender of the learner.
- country: Country of where the learner is doing the enrolment.
- age_range: The age range where the learner belongs to.
- highest_education_level: The highest education level achieved.
- employment_status: The status of employment.
- employment_area: The area in which the learner is employed.
- detected_country: Country of where the learner is doing the enrolment.

The variables gender, age_range, country, highest_education_level, employment_status and employment_area are considered demographic characteristics. Thus, they will be useful for our analysis in the second round of CRISP-DM.

2.3 Data Exploration

For our first investigation, we can see from the data available that we have for all seven runs of the course a total of 37,296 enrolments. However, having in mind the constraint explained in the business understanding phase, we just have 1,074 archetype survey responses, which accounts for 2,88 per cent of the total enrolments. Therefore, this will have an impact on the results of our investigation. Also, the analysis will be limited from the third run to the seventh run because we do not have survey responses for the first two runs of the course.

2.4 Data Quality

For this first investigation in which we want to analyse how the archetype influences the enrolment in the course and the certification purchase, we do not have many data quality issues for the variables we will require for this analysis. We do not have missing values for the archetype, learner id and role variables. They all have the correct data type: character. However, for the archetype variable, we have to make sure that the role of all the observations is learner and that the learner archetypes are the ones defined by FutureLearn, so we do not have to include values such as "Other" in our analysis. Regarding the date of purchase of the certificate, it is in the correct data type: character as well, and we have empty values which indicate that the learner did not buy a certificate, but this does not mean there is a quality issue with this variable. In this way, we can proceed with the data preparation phase without worrying a lot about the data cleaning phase.

3. Data Preparation

All the tasks and outputs described below are done in the `01-Enrolments.R` file located in the munge folder of the ProjectTemplate project directory.

3.1 Data Selection

First of all, we created a whole new data frame with all the enrolments by appending each enrolment data frame (cyber-security-{number of run}_enrolments) between the third and seventh run. The same approach was done to the archetype survey responses data frames (cyber-security-{number of run}_archetype.survey.responses) by appending each data frame for the same runs.

In this phase, we must select the variables and records to achieve our goals. In this way, for the enrolments data frame, we selected the records in which the role is learner to ensure that we are analysing this type of user. Also, for the archetype survey responses, we selected the records in which the archetype is one of the seven archetypes defined by FutureLearn in their research. Therefore, we are dropping out values such as "Other" for the archetype. Moreover, we removed any duplicates having the same learner and run. However, there are instances where the same learner participated in more than one run, which is fine and does not need to be removed.

Also, we merged both data frames by learner id and run (this column addition will be explained next), and removed unnecessary columns which are not useful for our analysis. Thus, we selected the columns learner id, archetype, which is the core of our analysis, purchased_statement_at for the certificate purchase analysis, and demographic variables (gender, highest_education_level, age_range, employment_status, employment_area, detected_country) which will be used in the following cycle. Also, we kept the column run to identify which run the learner participated in.

3.2 Data Construction

In the first instance, before appending each enrolment data frame described in the data selection phase, we created a new column for each one called run to identify which run the learner participated in. Moreover, we used the same approach for the archetype survey responses and we created the same column for each data frame. In consequence, we merged both data frames to see the variations in enrolments over the runs.

Additionally, regarding the certificate purchase analysis, we created a new custom column called purchased_certificate to transform the variable purchased_statement_at, which is a string containing the date of purchase, into a categorical variable with two classes: Yes and No. Consequently, if the purchased_statement_at had a date the value for the new column is Yes, otherwise, the value is No. In this way, it is more clear to make further modeling with this binary classification.

3.3 Merging and Transforming Data

As mentioned in the data selection step, we merged the enrolments and archetype survey responses data frames by learner id and run to have a big data frame that combined the different information in both data frames about the same learners. In this way, this will be the final data frame used for the modeling phase.

In addition, we created a new data frame needed for modeling which summarises the total count of each learner archetype in different runs. Therefore, this new data frame contains the following columns: archetype, run and count (the number of learners per archetype and run). Also, after doing this, we encountered that there were no learners with the Preparers archetype in the third run, so we added a new row to complete the data frame by setting the count to zero for the tuple of the Preparers archetype and run three. In conclusion, this will be the data frame to use in the following phase to analyse the proportions of learner archetypes.

4. Modeling

After preparing our data and getting it ready for the modeling phase, we are now able to produce some visual representations to analyse our data and make some conclusions to answer our first investigation.

In the line plot in Figure 1, we can analyse the learner archetypes who enrolled in the cyber security course over the different runs. In this way, it is clear from the graph that in the third run, there was not a big difference in the amount of learners for each archetype in the course. However, from the fourth run onwards, we can see an interesting and consistent domination of two archetypes which are the Vitalisers and the Explorers over the other types. We could also think that Advancers is a common archetype because the line is below of the both latter archetypes, but it just peaked in the fourth run and then it decreased significantly for the following runs. The less dominant archetypes in the course are the Flourishers, with an almost constant trend between the fourth and the following runs; Preparers, from no participation in the third run to an unexpected rise in the next run, but, it decreased gradually in the remaining runs; and the Hobbyists, which also had a rapid increase, but it remained almost constant for two runs followed by a decline in the last run.

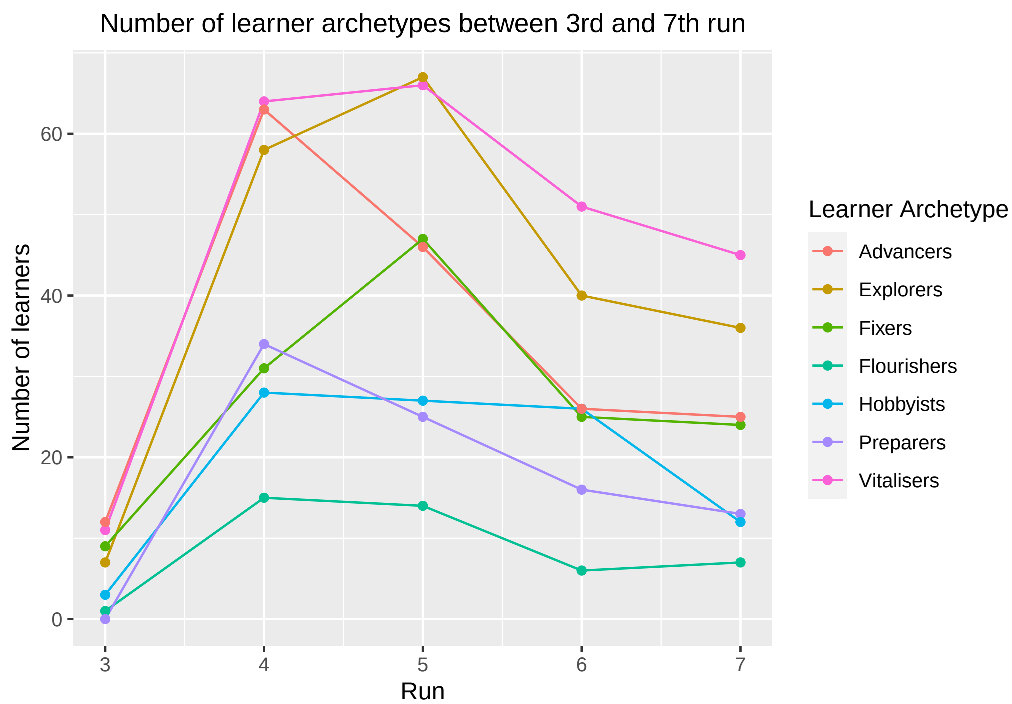

Figure 1: Distribution of learner archetypes between the third and seventh run

To analyse the purchase of a certificate for the cyber security course, the results are clear from the plot in Figure 2. It does not matter what archetype is the learner, simply people were not interested in purchasing a certificate and this has nothing to do with the learner archetype. The trend of not acquiring a certificate did not change over the analysed runs, so apparently there was no introduction of material to engage people to purchase a certificate.

One important thing to note is that for the most common learner archetypes identified in the line plot, some Explorers and Vitalisers did buy a certificate. However, the number of learners who did this is less than five. This does not give us insights to conclude that these specific learners are more interested in buying a certificate than other types.

5. Evaluation

From the modeling phase, we could see the distribution of learner archetypes in the cyber security course over the different runs with some types consistently more predominant than others. This gives our stakeholders

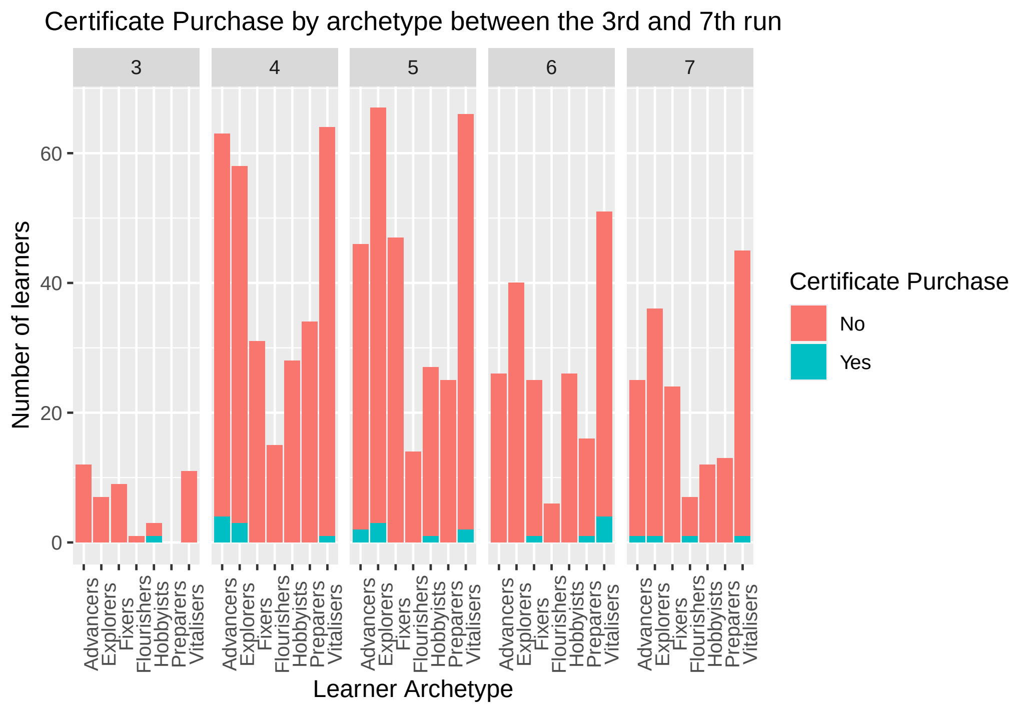

Figure 2: Distribution of the certificate purchase by archetype between the third and seventh run

useful insights about who their learners are and with whom they should focus their design for future courses. Additionally, we were able to answer if the certification purchase varies depending on the archetype, which the response is a resounding no. In this way, the initial business objectives were met.

These are the initial findings for our investigation, but we are aware that we had a small sample size of around 1000 learners who conducted the archetype survey for this course compared to more than 30000 enrolments in the seven runs. Therefore, as advised in the business understanding phase, it limited the conclusions of our analysis to be more accurate. Ideally, we should have a bigger sample of people who answered the archetype survey.

We will not move to the deployment step, but we will initiate a new iteration of CRISP-DM based on the results of this first cycle. This leads to our second investigation in which we now want to focus on the two most common learner archetypes, that is Explorers and Vitalisers, and analyse if there are differences in the distribution of enrolment between demographic variables such as gender, age range, employment status and education level. From the results of this cycle, it is not worth it to conduct further analysis for the certification purchase between demographic characteristics because the trend in this cyber security course is that learners are not interested in buying a certificate.

Second Cycle of CRISP-DM

1. Business Understanding

Thanks to the results of our previous cycle, we know that our initial second investigation changed slightly because of the findings about the certification purchase. One more time, it is not of interest for the learners of this cyber security course to acquire a certificate, thus, no demographic analysis will be intended for this aspect.

However, it is of strong interest for our stakeholders to focus the business objective on the two consistently dominant archetypes shown in the line plot of the previous cycle and explore the differences between Explorers and Vitalisers against the demographic variables of gender, age range, employment status and education level. This investigation will strengthen the results of our initial objectives, which remain the same, to give more insightful information and a more complete understanding of who the learners are for the course to consider in the design process of upcoming courses. Regarding, the data mining goals, they also remain the same for this iteration.

It is important to mention the constraints of this cycle regarding the sample size. By focusing the analysis on the two most common archetypes, the sample size is reduced significantly to 440 records. Thus, we have available less data than before to conduct the analysis and this might have an impact on the results of this iteration as well.

2. Data Understanding

The initial data described in the previous cycle remains the same because it is acquired from the same files. However, we will explain in more detail the demographic variables which are the core for this cycle.

2.1 Data Description

We are going to use the same two files (enrolments and archetype survey responses) to conduct this second investigation. Now, we are going to explain in more detail the content of the demographic variables to be used.

1. cyber-security-{number of run}_enrolments:

- gender: Gender of the learner. The values are female and male.
- age_range: The age range where the learner belongs to. The values are <18, 18-25, 26-35, 36-45, 46-55, and 56-65.
- highest_education_level: The highest education level achieved. The values are less_than_secondary, professional, secondary, tertiary, university_degree, university_doctorate, and university_masters.
- employment_status: The status of employment. The values are full_time_student, looking_for_work, retired, self_employed, not_working, umemployed, working_full_time, and working_part_time.

2.3 Data Exploration

We had 1,074 archetype survey responses from our first investigation. Now that we are going to focus on the two most common archetypes, we have available 440 rows with the archetype response and the demographic values from the learner's profiles. In this way, we are aware of the data availability constraint.

2.4 Data Quality

We do have data quality issues for the demographic variables. The columns of our data regarding gender, age range, highest education level and employment status, all have "Unknown" values in many of our observations. If one of the latter columns has an "Unknown" value, then all the other demographic variables have also an "Unknown" value.

We could think of replacing these "Unknown"s with the mode of each column, that is, the most frequent value. Nevertheless, by inspecting our data, we see that we have 355 rows with this value in the demographic variables of our analysis. That is the 80% of our current data. For this reason, we can potentially introduce some bias to our analysis by performing this action because there are many missing values. Consequently, we could instead follow the approach of removing the rows with these missing values with the concern of reducing our sample or implementing a model for each column that predicts the missing values.

3. Data Preparation

All the tasks and outputs described below are done in the `01-Enrolments.R` file located in the munge folder of the ProjectTemplate project directory.

The data preparation done in the first cycle gives us a strong basis to do minor work in this step for the current iteration. We are going to use the same big data frame where we merged the enrolments and archetype survey responses data frames and in which we appended the third to the seventh run. This big data frame is cleaned with the process done and described in the data preparation of the last cycle.

3.1 Data Selection

In this phase, we must select the variables and records to achieve our goals. First of all, for different runs, we had duplicates for the same learner id and archetype meaning that the same person was enrolled in more than one run of the course, so we removed these duplicates because we were not discriminating the analysis by run in this cycle. Second, we dropped all the unnecessary columns for this cycle, thus, the only columns that remained are the archetype, gender, age range, employment status and highest education level. Third, we selected the records that correspond to the two most common archetypes, that is Explorers and Vitalisers. Fourth, regarding the missing values described in the data quality section, considering the scope of the course we are not going to perform a model for each variable to predict the missing value. Therefore, we are going to remove the rows which have an "Unknown" value in each of the demographic variables of our analysis. This gives us a smaller sample size of 85 observations which will have an impact on the conclusions of the report. Lastly, we are going to combine two possible values (not_working and unemployed) of the column employment status to be just unemployed.

3.2 Data Construction

We created a new data frame needed for modeling which summarises the total count of each learner archetype by different employment status. Therefore, this new data frame contains the following columns: archetype, employment status and count (the number of learners per archetype and employment status).

After doing this minor changes to our initial data frame we are now able to continue to the modeling phase.

4. Modeling

After preparing our data and getting it ready for the modeling phase, we are now able to produce some visual representations to analyse our data and make some conclusions to answer our second investigation. We are going to create some plots for each demographic variable to see if there are differences between archetypes.

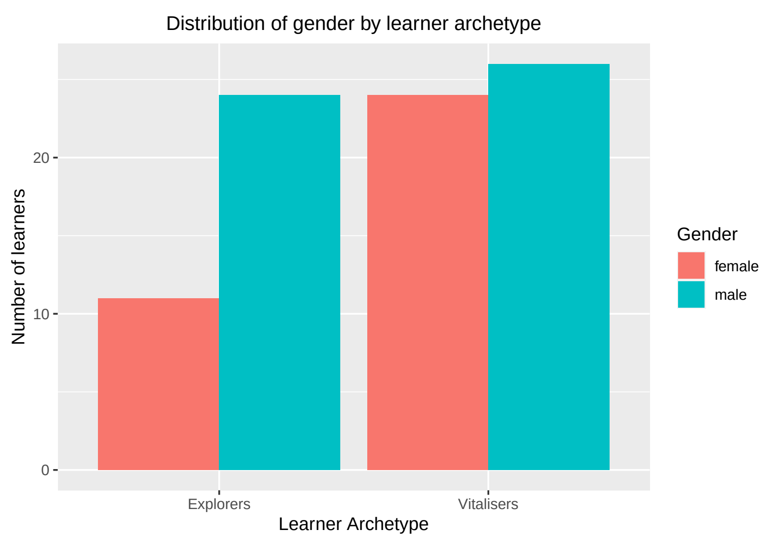

Figure 3: Distribution of gender between Explorers and Vitalisers

In the bar plot in Figure 3, we can see that there is not much difference between the number of males and females for the Vitalisers and the number of males between Explorers and Vitalisers. However, it is clear from the graph that more women belong to the Vitalisers archetype than the Explorers. Moreover, few women are Explorers compared to men.

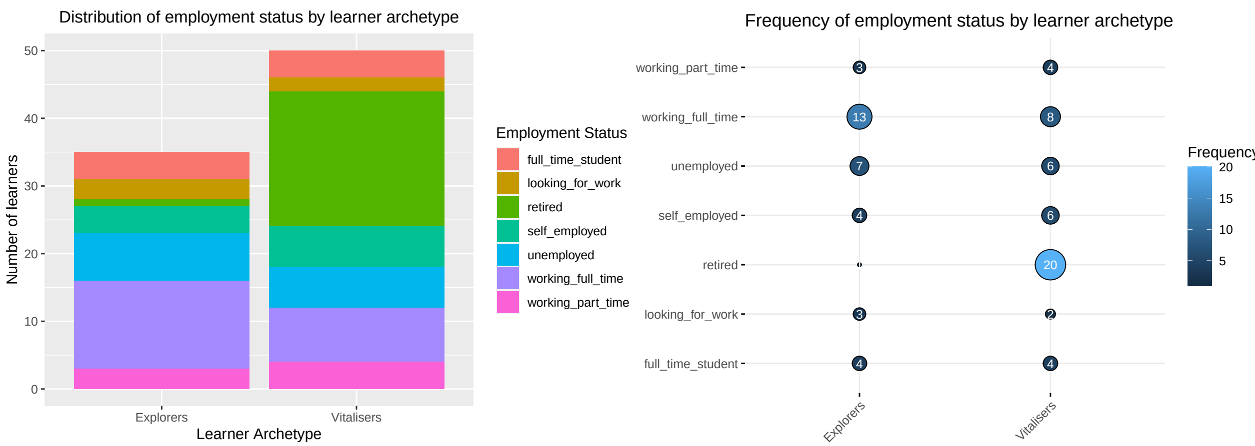

Figure 4: a) Bar plot of the distribution of employment status between Explorers and Vitalisers. b) Bubble plot with the frequency of employment status between the same learner archetypes

From the two plots in Figure 4, we can see that the biggest proportion of the employment status of Vitalisers corresponds to people who are retired, followed by those working full-time and self-employed. However, compared to the Explorers, the retired and self-employed people have the smallest proportion in this archetype. On the contrary, people working full time have the largest portion. This makes sense regarding the nature of both archetypes. Additionally, we can see that there is a similar proportion of full-time students for both learner archetypes. We should not make further conclusions with the very small numbers shown in the bubble plot for the other categories of employment status.

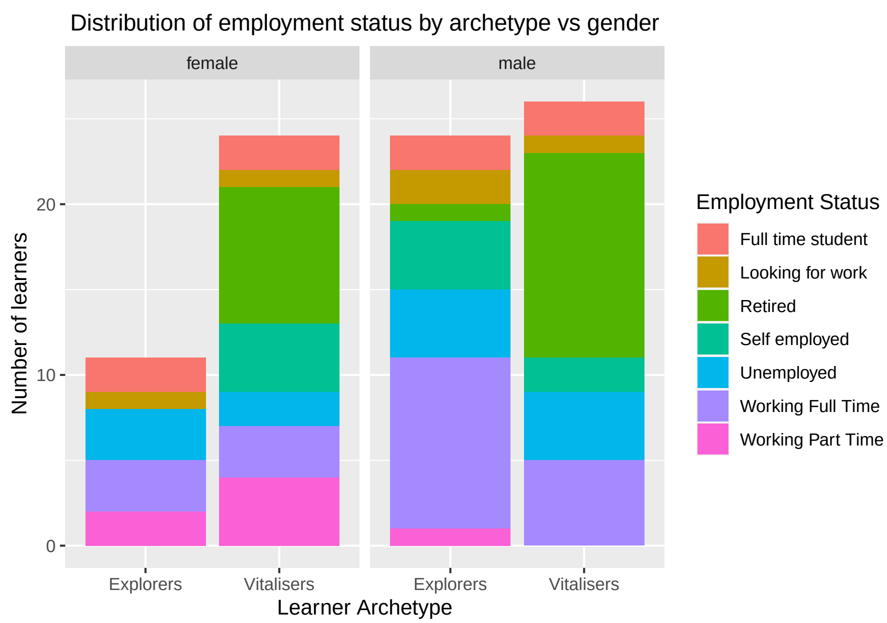

Figure 5: Distribution of employment status between Explorers and Vitalisers versus gender

In the plot in Figure 5, we want to compare the distribution of employment status by learner archetype versus gender. For the largest proportion of employment status of Vitalisers, we see that more men are retired than women. More men are working full-time who belong to the Explorers than the Vitalisers, and for both archetypes, fewer women are working full-time than men. Regarding self-employment, more females are part of the Vitalisers than males, and for the Explorers, all the data belongs to males. Interestingly, as we see in the last plots, the distribution between males and females for the full-time students are quite similar as well.

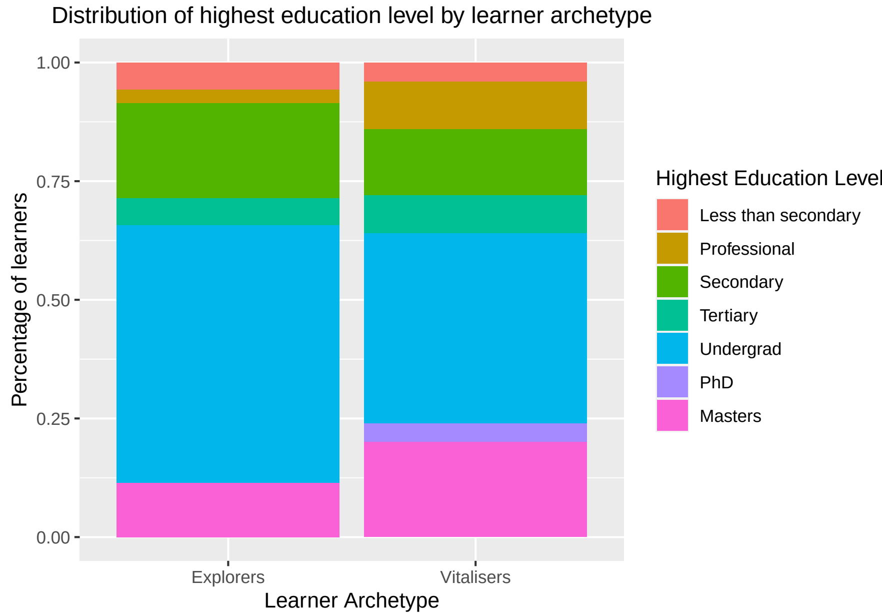

Figure 6: Distribution of highest education level between Explorers and Vitalisers

From the plot in Figure 6, we can see that the largest proportion of the highest education level achieved by the learners of this course is an undergraduate degree for both archetypes. In the Vitalisers, we can see that a master's is the following highest education achieved. However, in the Explorers, the second highest education level is secondary. There is a very small percentage of people who pursued a PhD and this one just belongs to the Vitalisers.

In the last plot in Figure 7, we want to compare the distribution of the age range by learner archetype versus gender. We see that in the Vitalisers, the largest proportion for both males and females is for people with ages higher than 65. This makes sense because the biggest portion of the Vitalisers are retired. Moreover, the following age ranges for female Vitalisers are between 18 and 25 years, and for male Vitalisers are between 46 and 65. For the Explorers, we see that the highest bars are for males between 36 and 45. For women who are Explorers, the distribution of ages is similar between the ranges from 18 to 45 years old.

Finally, it is important to mention that we did not create more complex plots that could have all the categorical variables compared into one plot. For instance, distribution of employment status versus learner archetype, by gender and age range, because we do not have enough data to produce a meaningful comparison, we would potentially have between one and three records per combination.

5. Evaluation

For this second cycle, 80% of our data had "Unknown" values in the columns we needed for our investigation. This limitation of the data was our main issue in providing strong conclusive insights to our stakeholders about the analysis of the demographic variables. We had to remove all the rows with these issues because building models to predict the missing values for each variable was out of the scope of this course. In this way, our data mining goals were restricted to the data available after performing this action.

Our business objectives were partially met because we had initial findings that allow our stakeholders to understand who their learners are based on the differences found in gender, age range, education level and employment status. We concluded that the distributions of the demographic variables considered in this second cycle are highly related to the definition of each archetype. But, we emphasize that the sample size was insufficient. Therefore, we suggest that future courses include more data for the demographic variables by finding a way to ensure that the learners' profiles are completed in the enrolment for the course.

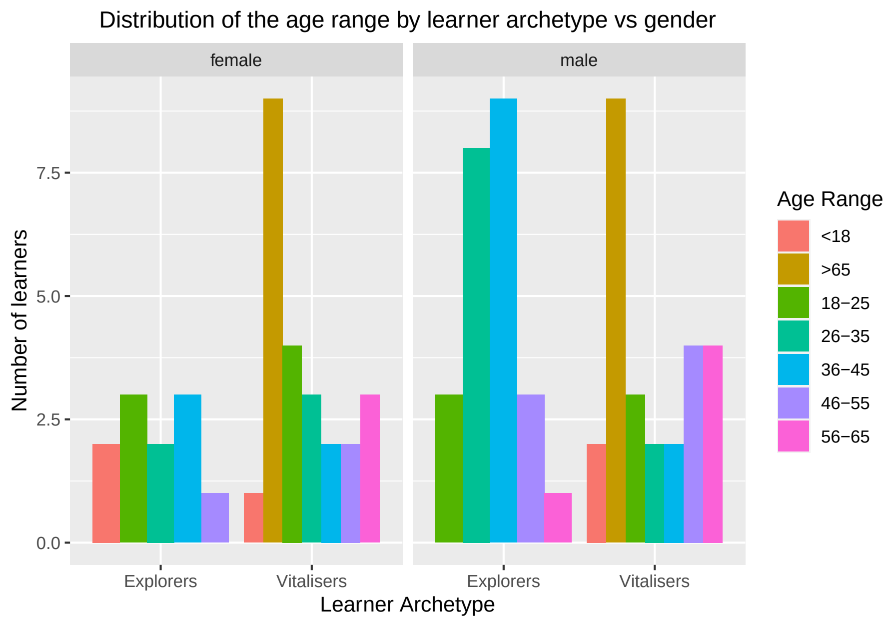

Figure 7: Distribution of age range between Explorers and Vitalisers versus gender

We could think about performing more cycles to strengthen our initial business objective of understanding the learners in the cyber security course such as how these learner archetypes behaved in the course by analysing the hours spent in each task, video reproductions and steps completed for the suggested activities. However, we will continue to the deployment phase to show these initial valuable findings to our stakeholders in which we used the correct data to conclude our analysis.

6. Deployment

In this final phase of our second cycle, we are going to deploy the results to our stakeholders. First of all, we will deliver this analysis report which includes all the steps, considerations and visual representations to answer the two investigations of our data to meet the business objectives. Moreover, we will do a presentation with the main findings of our investigation and the highlights of the analysis done with CRISP-DM to be presented to Newcastle University and the School of Computing.

We are aware that data is in constant change, therefore, the results of this investigation might not be accurate in the future. In this way, we need to monitor and maintain the strategy explained in the previous phases by recollecting new data of new cyber security courses (hopefully with a larger sample size to avoid the limitations of the data we had in this version of the analysis) and compare if the common archetypes are still the same if people still avoid purchasing a certificate and if the demographic variables for the common archetypes vary.

Consequently, the next steps for our stakeholders are to explore new methods and features to approach the most common learner archetypes in the course. Then, launch the new course with the new design. Following, conduct the archetype survey and compare if the results of this investigation are still valid. Finally, expand the analysis to other courses and platforms.

In terms of assessing the overall project, we will create a reflective report. Within this, I will discuss what went right, what went wrong and what needs to be improved for future projects similar to this one.

References

Pickard, L. (2019, March 9th). What Kind of Online Learner Are You?. *The Report*. https://www.classcentral.com/report/what-kind-of-online-learner-are-you/

Shah, D. (2017, July 24th). MOOCs Find Their Audience: Professional Learners and Universities. *The Report*. https://www.classcentral.com/report/moocs-find-audience-professional-learners-universities/

Information Technology. Garnert Glossary. (n.d). Massive Open Online Course (MOOC). https://www.gartner.com/en/information-technology/glossary/massive-open-online-course-mooc

Balloon plot. https://rpkgs.datanovia.com/ggpubr/reference/ggballoonplot.html

FutureLearn Learner Archetypes. https://ugc.futurelearn.com/uploads/files/2f/ce/2fce94dd-54e6-4614-972d-ce9c3a6d7d91/Learner-Archetypes.pdf