CSC8643: Data Management and Exploratory Data Analysis

by Matthew Taylor

17th November, 2023

Contents

1 Abstract

This report covers two rounds of CRISP-DM analysis on the FutureLearn "Cyber Security: Safety At Home, Online, and in Life" course ran by Newcastle University. The report covers the relevant steps of the analysis process, and well as an evaluation of the findings. This report aims to show the use of the R language and the tools provided in Rstudio like "ProjectTemplate", as well as additional libraries like ggplot2.

2 Business Understanding

2.1 Business Objective

Newcastle University developed a massive open online course titled "Cyber Security: Safety At Home, Online, and in Life" with the primary goal of providing a free and easily accessible resource on the topic of cyber security, specifically on how learners can safeguard their digital data. The course was offered seven times, with the majority of users enrolling during the first and second academic semesters from 2016 to 2018. All course tasks were conducted online and asynchronously.

The main objective of the course was to reach as many learners as possible worldwide, which was facilitated by offering the course for free and making participation asynchronous. The FutureLearn system captured an extensive amount of data, covering a range of variables. Since each user had a unique identifier, it allowed for a wide range of analyses that can be conducted by combining different parameters.

This study aims to analyse the geographic demographics of the learners with the objective of identifying countries with a high number of enrolments. This analysis would provide a better understanding of the data, allowing future course content to be designed with cultural nuance, language preference, and regional challenges in mind. The data could also aid in allocating resources based on the geographic demographics of the learners. For instance, translating the course into different languages could increase enrolments.

In the second round of analysis, the study investigated the engagement of the learners with the course by examining their completion rates relative to their detected country. This analysis could help to identify ways to make the course more relevant to the learners' differing cultures and increase their engagement. For example, having course administrators who are familiar with the learners' cultures could be beneficial. The report's success criteria are to successfully explore the geographic demographics of the user base and to examine the engagement and completion rates of the learners on the course.

2.2 The Project Plan and Data Mining Goals

The project plan entails conducting two rounds of CRISP-DM analysis, with the second round building upon the findings of the first. The utilization of the CRISP-DM structure, a Cross-Industry Standard Process in Data Mining, offers a highly structured approach for a data mining-driven report, encouraging the completion of integral processes and the consideration of important factors.

The analysis aims to delve into the demographics of the learners' respective countries and discern any patterns that may contribute to an increase in the number of individuals taking the course or improve the engagement levels of current users.

3 Data Understanding

3.1 Collect Initial Data

The FutureLearn system amassed a significant amount of raw data that has been logged in CSV files, and these files have been made available for the purposes of this report. The data is predominantly collected automatically by the FutureLearn system, although certain data points were obtained from the learners themselves. Notably, many of the data points that were requested from the users were optional, such as their country, age range, gender, level of education, and employment status, and this resulted in a low response rate, as illustrated in the table presented below.

	Number who reported:	**Percentage who reported:**
Gender	3733	10.6 %
Country	3726	10.6 %
Age Range	3620	10.3 %
Highest Education Level	3715	10.6 %
Employment Status	3689	10.5 %
Employment Area	2881	8.2 %

3.2 Describe Data

The data is categorized by each of the 7 runs of the course and by the specific phase in which the data was collected. For instance, during enrolment, the date and demographics of the learner were documented. There are eight separate documents for each run, which include engagement and feedback survey responses. Many of these files contain the unique ID of the learner, allowing for the association of parameters from one file to another. It is important to note that not all runs monitored the same parameters, such as the learner's archetype, which was not determined during the first two runs.

3.3 Explore Data

During the exploration, the data on the country of the learners was initially examined. It was discovered that only 10.6% of the learners indicated their country. However, unlike the other demographic information, the detected country of the learners was automatically detected by the FutureLearn platform upon enrolment.

	Total:	**Percentage:**
Number of learners	35205	100.0 %
whose country was detected	34310	94.5 %
who self reported their country	3726	10.6 %
whose reported and detected country was correct	3421	9.2 %
whose reported and detected country was incorrect	221	0.6 %
whose country was not Detected	873	2.3 %

3.4 Verify Data Quality

The accuracy of a learner's reported country was assumed to be sufficiently accurate, given that it was automatically collected upon enrolment. To test this assumption, a comparison was made between the reported country and the detected country of learners who reported their country. The analysis revealed that 6.07% of learners who reported their country, provided a different country from the one detected. As a result, it is assumed that the detected country parameter has an accuracy rate exceeding 93%. Factors, such as a learner's use of a VPN or reporting their home nation while residing abroad, could be deemed to influence the accuracy of either parameter.

4 Data Preparation

4.1 Select Data

The accuracy in the identification of reported countries among the 35,000 learners exceeded 93%. Consequently, this parameter over the self-reported country was used, considering that only 3726 learners provided such information. This decision was driven by the desire to work with a more substantial data set for the analysis. Additionally, the investigation required the integration of progress data from the supplied step files along with learner IDs from each file to connect a learner's country and their level of engagement.

4.2 Clean Data

Data cleaning involved the removal of the 2052 duplicate learner IDs and the 39 administrators, as well as users whose country was not detected. Users whose detected and reported countries did not match were retained, as the findings of the analysis were envisioned to be used for targeted advertising, where the detected country was deemed more relevant. Only 2.3% of learners whose country was not detected were excluded, as only 84 of them reported their country.

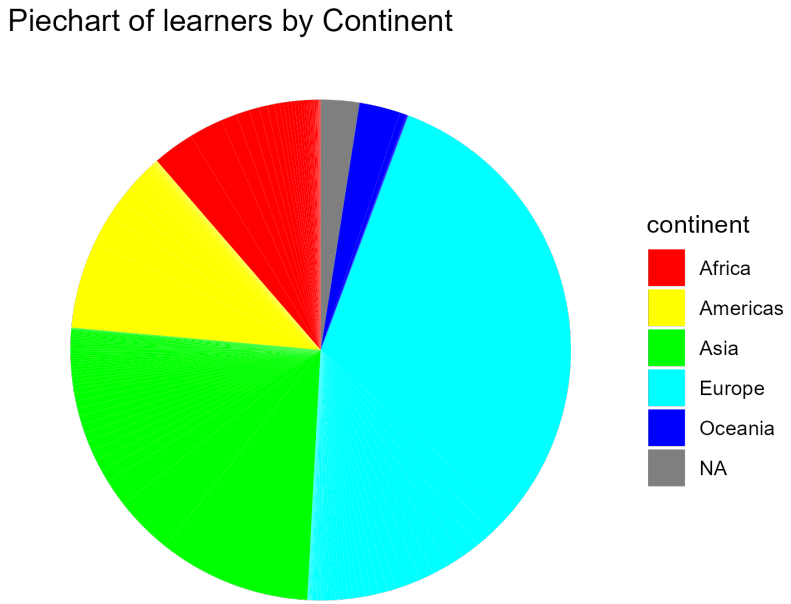
4.3 Integrating and Constructing Data

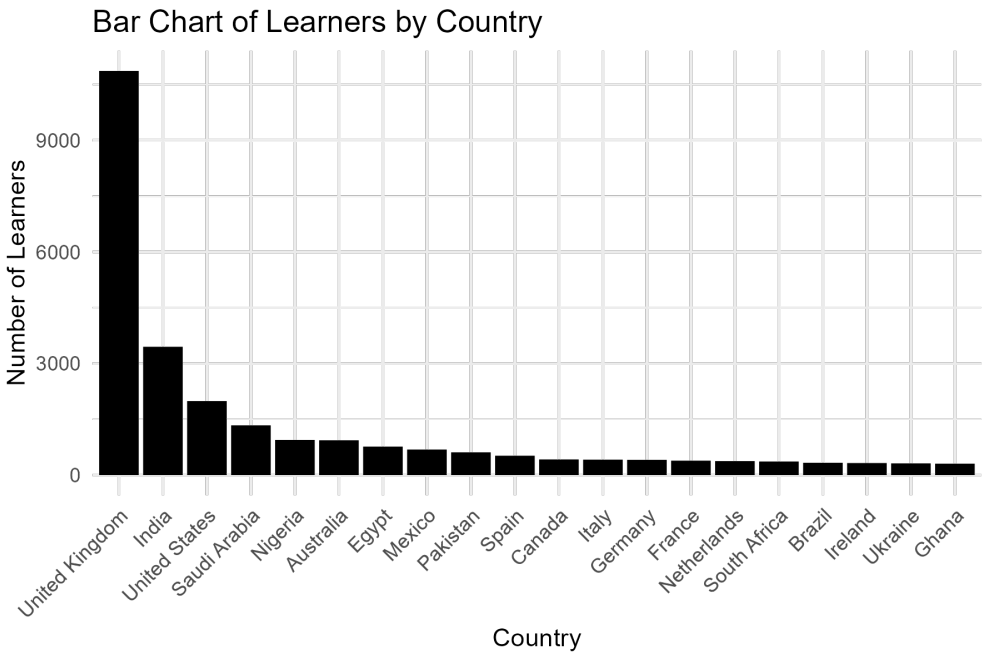
For the second round of CRISP-DM analysis, the geographic data from the first round was integrated with additional raw data from the step file, which enabled the incorporation of engagement data from each learner. Once the data was collated and associated with each learner, additional data points were constructed. Specifically, the number of learners from each country and their engagement were used to identify the learners who completed the course. This information was then used to calculate the engagement and course completion percentage for each country.

5 Modeling and Evaluation

5.1 Modeling and Evaluation of the First Round

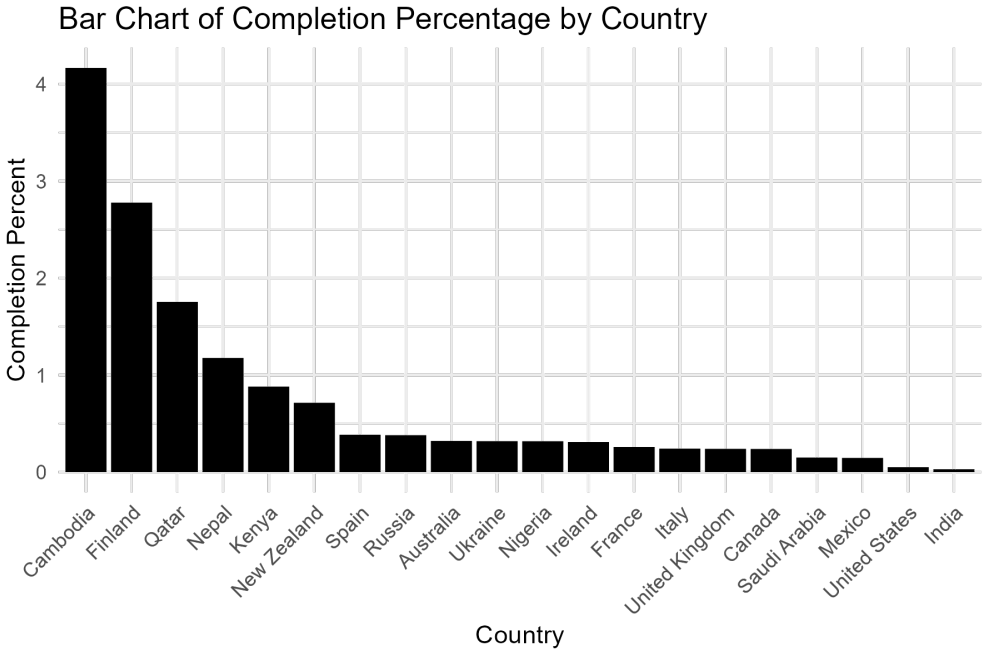
In the first cycle of the CRISP-DM analysis, an investigation was conducted into the countries where the learners were detected. Based on the pie chart, the majority of learners were located in Europe, with Asia and the Americas following in second and third place, respectively. The bar chart below provides a closer look at individual countries. It is observed that the UK had the vast majority of learners, with India and the United States following in second and third place, respectively.





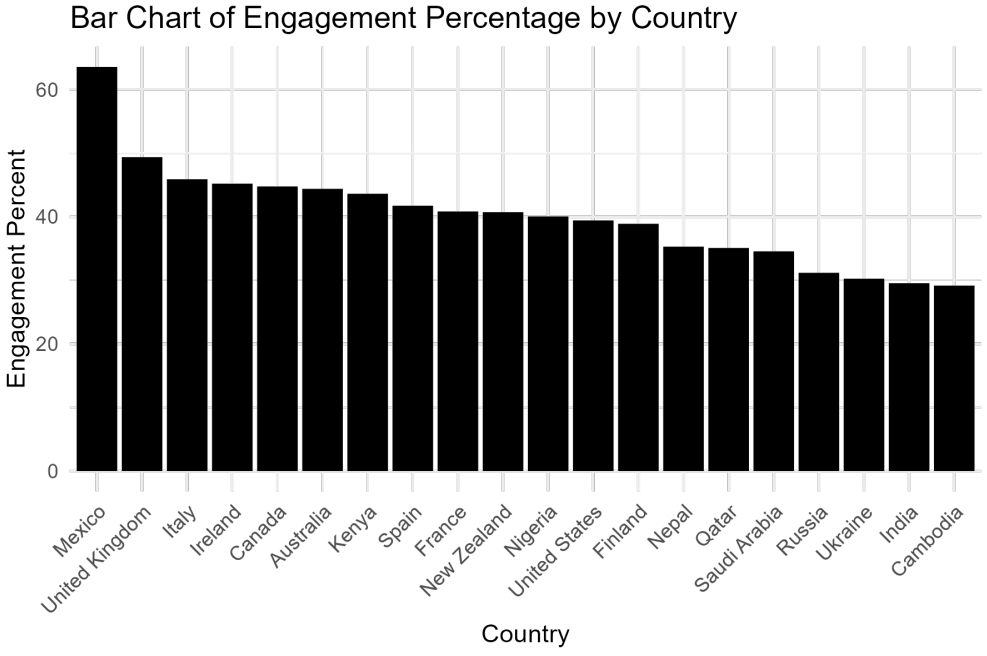
5.2 Modeling and Evaluation of the Second Round

For the second cycle of the CRISP-DM analysis, the number of learners who completed the course from each country was analyzed to gain a better understanding of their engagement with the course. The country with the highest number of learners who completed the course was the UK, with 26 learners, which represents 0.24% of all learners from the UK. The country with the highest completion rate was Cambodia with 4.2%, which was one of their 24 learners.

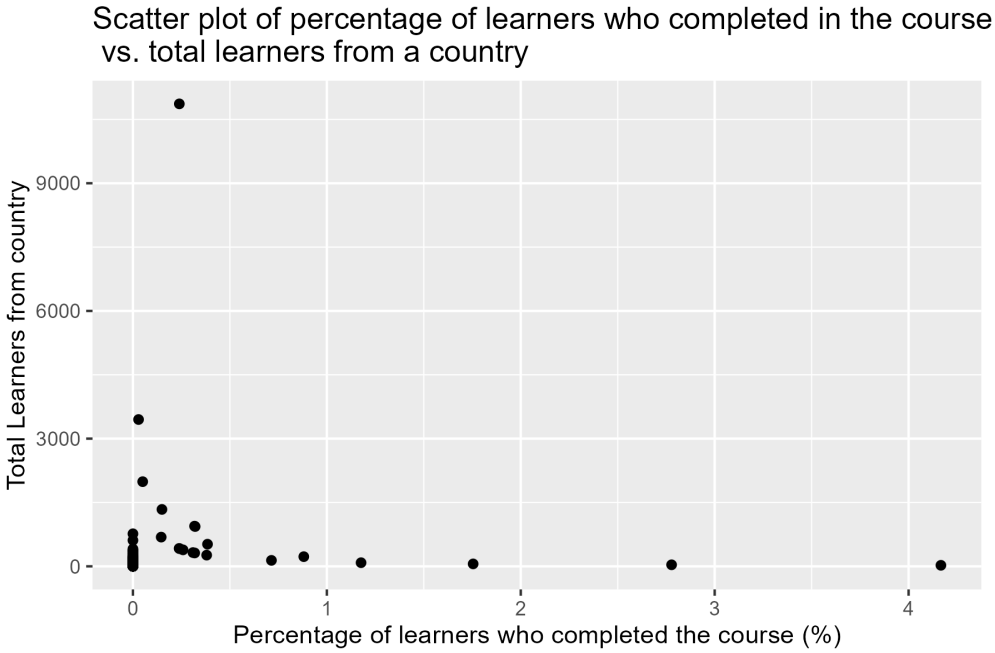


It was found that the number of learners who completed the course was quite low, with only 0.15% of learners completing it. Therefore, engagement was also analysed. It was observed that 40.2% of learners engaged with the course after enrolment, resulting in a considerably larger data set when analyzing engagement rather than completion.

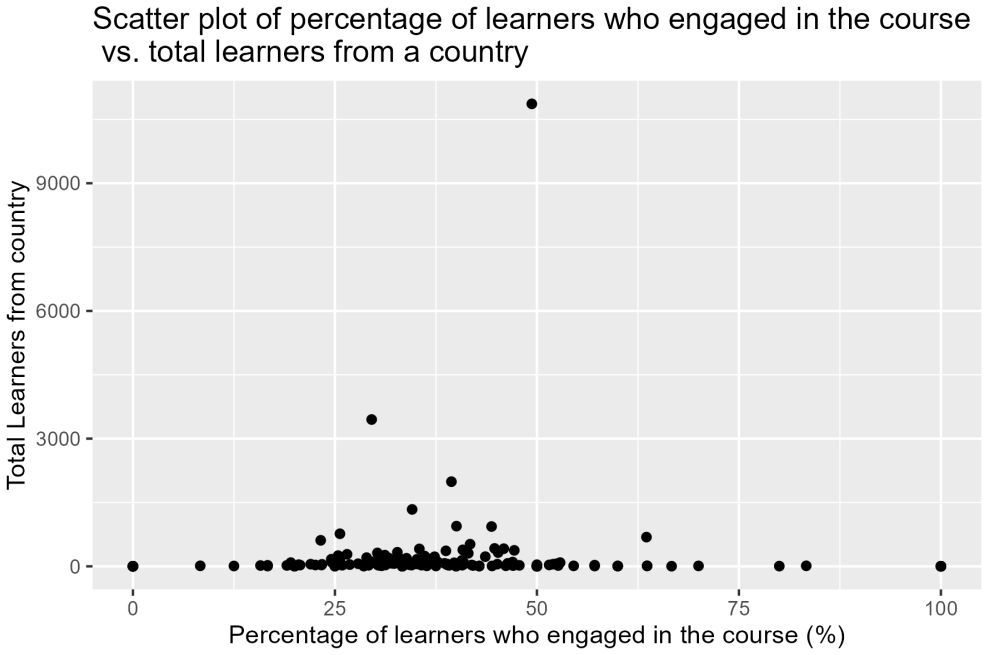
This bar chart shows the percentage of learners who engaged in the course from the top 20 countries.



A scatter plot was created for both engagement in the course and completing the course to determine if there was a correlation between the number of learners from a country and their completion or engagement rate. The first scatter plot shows a loose negative monotonic relationship, indicating that as the number of learners from a country increases, the percentage of learners who complete the course decreases. Of the 20 countries that had a learner complete the course, 16 of those countries only had one user complete the course. Therefore, the quality and quantity of the data is insufficient to draw any conclusions.



The second scatter plot, there is no correlation between the number of learners from a country and how many users continued to engage with the course after enrolment.



5.3 Next Steps

Following this report, the stakeholders will need to decide on the next steps. While the first round of analysis successfully met its objective of providing the geographic demographics of the learners, which the stakeholders can action upon. The second round was inconclusive due to the limited data of learners who completed the course, and no correlation between number of users in a country, and their engagement. Further analysis could be beneficial, pulling in additional data points, like testing for a correlation between declaring their demographic data and engagement. If the course is run again in the future, there should be consideration into making the learners declaring their demographic data mandatory as those fields only had a response rate of around 10%.

6 Deployment

The last process in the CRISP-DM framework is for the analysis to be deployed. The aim of this analysis was to share the geographic insight into the learners to the stakeholders, which was the goal of this report. The report included all the data that was available at the time of publishing, and as the course ceased in 2018, there was no attempt in the structure of the analysis to include future data. Although in practical turns it would be simple to modify the munging process to append future rounds of data as long as it was structured the same way as previous rounds. Analytical scripts that follow the munging process were sufficiently generalised that additional data points would still be processed.

7 Conclusion

In Conclusion, this report shows the countries with the greatest number of enroled learners, and the countries with the highest percentage of learners who engaged or completed the course. The evaluation was not able to find a correlation between the chosen parameters, so conclusions can't be drawn about the number of learners in a country, and their engagement or completion rate. The number of learners per country, which was the aim of the first round of CRISP-DM analysis, did successfully find where most of the learners resided, and could provide valuable insight into the languages and cultures of the user base.

The second round of CRISP-DM analysis was applied with the creation of the two scatter plots. The scatter plots had the opportunity to show if there was a correlation between the number of learners from a country and the engagement or completion rate. There was no correlation in the engagement rate, and there was not enough data to generate the second scatter plot to draw any conclusions.

8 Refrences

P. Bhandari, “Correlation Coefficient | Types, Formulas & Examples,” Scribbr, Jun. 22, 2023. https://www.scribbr.com/statistics/correlation-coefficient/