CSC8643: Data Management and Exploratory Data Analysis

by Matthew Taylor

20th November, 2023

Contents

1 Abstract

2 Introduction

3 Business Understanding

3.1 Business Objective

Newcastle University designed a massive open online course entitled "Cyber Security: Safety At Home, Online, and in Life". The Primary objective of this course was to provide a free resource that was easily accessible via the online provider FutureLearn on the topic of cyber security, and how the learners can protect their digital data.

The course was ran 7 times, with the majority of users signing up at the beginning of the first and second academic semesters from 2016 through 2018. The tasks set out in the course were exclusively online and could be completed asynchronously.

A main aim of this course was to have as large a reach around the world as possible, and this was facilitated by making the course free, online and the participation asynchronous. The FutureLearn system recorded a vast about of data, covering a range of variables, and because each user had a unique identifier, there is a wide range of analysis that can be done, bringing together different parameters.

With this aim in mind, I have decided to analysis the geographic demographic of the learners, as a greater understanding of this data could allow for identifying countries with a high number of enrolment. With country specific data, future course content could be designed with cultural nuance, language preference and regional challenges. This data could also help with resource allocation, based on learner's geographic demographic. For example, it could be found that translating the course into different languages could increase the number of enrolments.

For the second round of analysis, I will build on the learner's geographic demographic, and look at their engagement with the course, by looking at the learners completion rate compared to their country. This data could be helpful as it could demonstrate for example, that having course administrators who are familiar with the differing cultures, could make the course more relevant, which could increase the engagement.

3.2 Access Situation

3.2.1 Inventory of Resources

This project is being undertaken with a limited number of resources. There is a vast amount of data including demographics, and engagement with the course, as well as survey responses. Having an extensive amount of raw data is helpful to the aims but requires additional resources to handle. For this project, I am a team of one, with no data science or modelling expertise.

3.3 Determine Data Mining Goals

The specific goal of this analysis is to dig into the country demographics of the learners, and determine if there is a pattern that could lead to increasing the number of people who take the course, or increase the engagement of current users.

3.4 Produce Project Plan

Due to my limited experience in data modelling, I will deploy some limited analysis techniques.

4 Data Understanding

4.1 Collect Initial Data

The FutureLearn system recorded an extensive amount of raw data and recorded them to CSV files, which have been made available for this report. The data is mostly gathered automatically by the FutureLearn system, but some of the data was gathered though asking the learner. Many of the data points that were requested from the user were optional, including parameters like country, age range, gender, education level, and employment status, which resulted in a low response rate, shown in the table 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 %

4.2 Describe Data

The date is separated by each of the 7 runs of the course, and by which step the data was gathered. for example, during enrolment, the date and the learner's demographics were recorded. There is a separate file for engagement, as well as for the feedback survey, for an approximate total of 8 files for each run. Many of the files contain the learners unique ID, so a parameter from one file can be associated to the same learner in a different file. Not all runs tracked the same parameters, for example a learner's archetype was not determined for the first 2 runs.

4.3 Explore Data

As part of my exploration, I firstly looked at the data gathered on the learners country. I found that only 10.6% of learners reported their country, but unlike the other demographics, the FutureLearn platform would detect the country the learner connected from at the time of enrolment.

	Total:	**Percentage:**
Number of learners	35205	100.0 %
whose country was detected	34310	94.5 %
who self reported their country	3726	10.0 %
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 %

4.4 Verify Data Quality

I would assume the data of the learners country would be sufficiently accurate, as it was automatically gathered when the learner enrolled. I tested this assumption by looking at the learners who reported their country and compared it to their detected country and I found that 6.07% of learners who reported their country, reported a different country to the one detected. Therefor it could be assumed that the detected country parameter has an accuracy of greater than 93%. There are a range of reasons that one or the other may be correct, for example if a learner is using a VPN, then their detected country could be incorrect, or

a learner may have reported their home nation while living abroad, resulting in the reported country being incorrect.

5 Data Preparation

5.1 Select Data

As I found the accuracy of the reported country to be greater than 93%, and there are over 35 thousand learner's whose country was detected, so I chose to use that parameter over the reported country as only 3726 learners reported their country, giving a larger data set to work from. I also needed to work with the progress data from the step files provided as well as the learner IDs from each of the files to connect a learner's country and engagement.

5.2 Clean Data

To clean the data, I removed the 2052 duplicate learner IDs as well as the 39 administrators, as I was only interested in the learners. I also removed users who's country wasn't detected. I decided to keep users who's detected and reported countries don't match, as I envision the findings of this analysis will most likely be used to set up targeted advertising, In which case the detected country would be more relevant. I did not include the 2.3% of learners whose country was not detected as only 84 of them reported their country.

5.3 Construct Data

didn't do this.

5.4 Integrate Data

For my second round of CRISP-DM analysis, I built on the geographic data found in my first round, so I integrated the additional raw data from the step file, so that I could incorporate the engagement data from each learner.

5.5 Format Data

not done.

6 Modelling

6.1 Select Modelling Techniques

compere

6.2 Generate Test Design

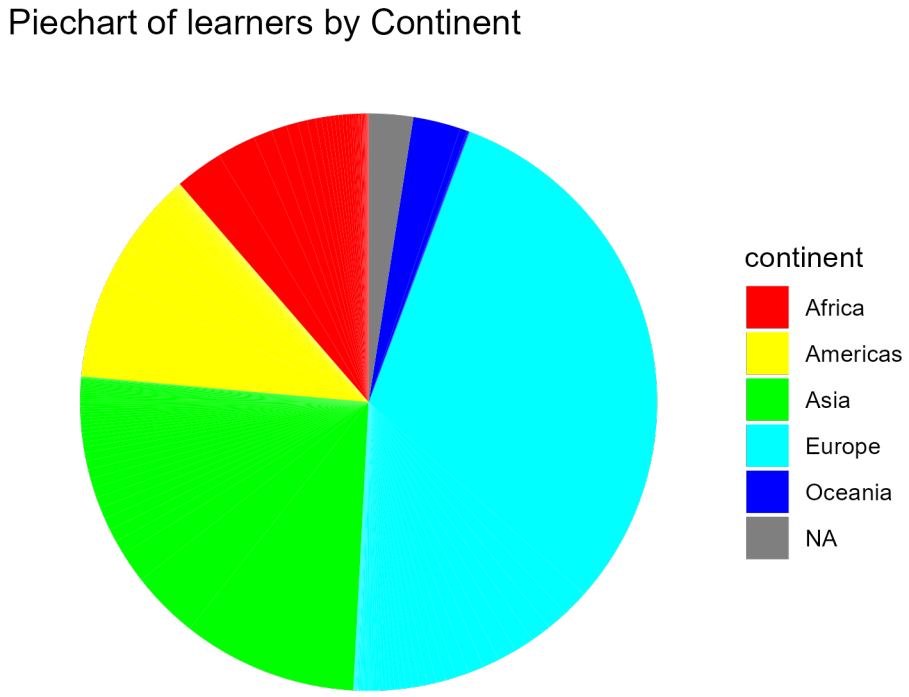
6.3 Build Model

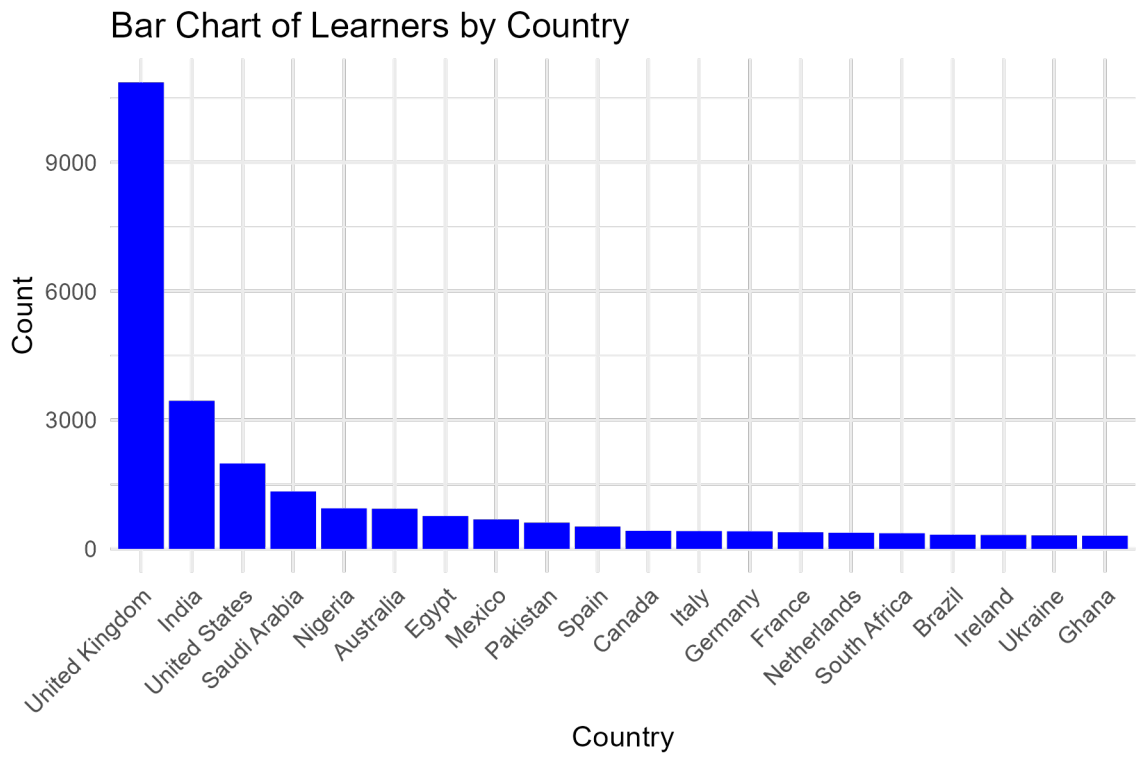
6.4 Assess Model

7 Evaluation

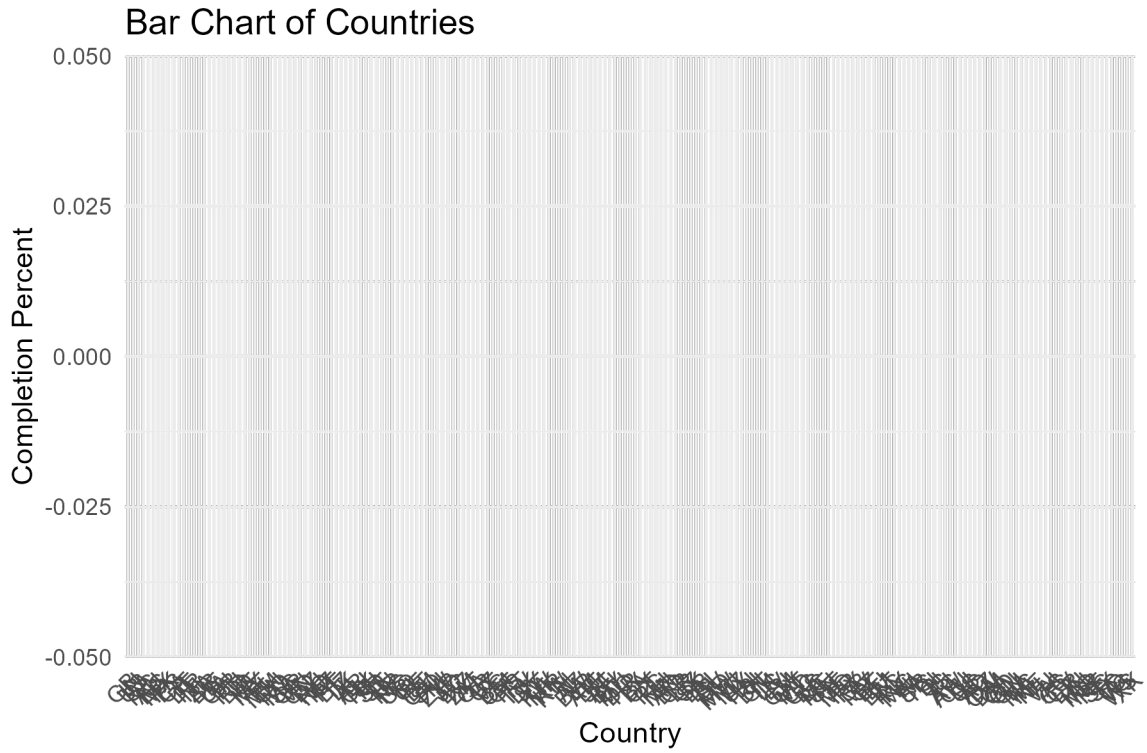
7.1 Evaluate Results

For my first cycle of the CRISP-DM analysis, I looked at which counties the learners were detected in. As we can see from the pie chart, most learners were located in Europe With Asia and the Americas coming in second and third respectively. Looking closer at the individual countries shown in the bar chart below, The vast majority of learners are from the UK, with India and the United States being second and third respectively.





For the second cycle of the CRISP-DM analysis, I decided to look at how many of the learners from each country completed the course, to get a better idea of each geograghic demographics engagement with the course.



7.2 Review Process

7.3 Determine next steps

8 Deployment

8.1 Plan Deployment

8.2 Plan Monitoring and Maintenance

8.3 Produce Final Report

8.4 Review Project