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 the first and second academic semesters from 2016 through 2018. The tasks of the course were exclusively online, and could be completed asynchronously.

The FutrureLern system recorded a vast about of data, covering a range of variables, including demographics, and engagement with the course, and because each user had a unique identifier, and their demographic and engagement data was recorded against their ID, there is a wide range of analysis that can be done.

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 modeling, I will deploy some limited analysis techniques.

4 Data Understanding

4.1 Collect Inital Data

The FutureLearn system recorded an extensive amount of raw data and published them to CSV files, which have been made available for this report.

4.2 Decribe Data

The date is separated by each of the 7 runs of the course, and by the description of the data, for example enrollment is separated from engagement, and so on. Many of the files contain the learners unique ID, so a parameter from one file can be linked to the same user in a different file. Not all runs tracked the same parameter, for example a learners 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 most users, did not enter their country, but their country was also automatically detected, and recorded which had a much better level of data. // reword

I also found there to be 2052 duplicate learner IDs, so i removed those, as well as the 39 users, who's roles were administrators, as i only want to track the data of the learners. // don't say what i did, just metion the bad data.

Number of learners:	**37257**	**Percent of learners:**
who's country was detected	35205	94.5%
who self reported their country	3730	10%
who's reported and detected country was correct	3421	9.2%
who's reported and detected country was incorrect	221	0.6%
who's country was not Detected	873	2.3%

4.4 Verify Data Quality

I found the quality of the data collected around a learners country is likely to be quite accurate as only 6.4% of learners who reported their country, differed from their detected country. In all cases i used a learners detected country, as I had a larger data set, and there is no way of verifying if the detected county or reported country is the correct answer. There are a range of reasons, that one or the other may be correct, for example is a user is using a VPN, then their detected country could be incorrect.

5 Data Preparation

5.1 Select Data

For selecting my data, I started with creating new data frames from the provided CSV files provided by FutureLearn, using the enrollment data and the step data from the 7 runs. I also removed the unnecessary columns for clearer reading while working on the sets.

5.2 Clean Data

To clean the data, I removed the administrators, as I was only interested in the learners, and I 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, I which case the detected country would be more relevant. I did not include the 2.3% of learners who's country was not detected as only 84 of them reported their country.

5.3 Construct Data

didn't do this.

5.4 Intergrate Data

For my second round of CRISP-DM analysis, I built on the data i found in my first round, so i integrated the additional raw data required to my initial data frame.

5.5 Format Data

not done.

6 Modeling

6.1 Select Modeling Techniques

6.2 Generate Test Design

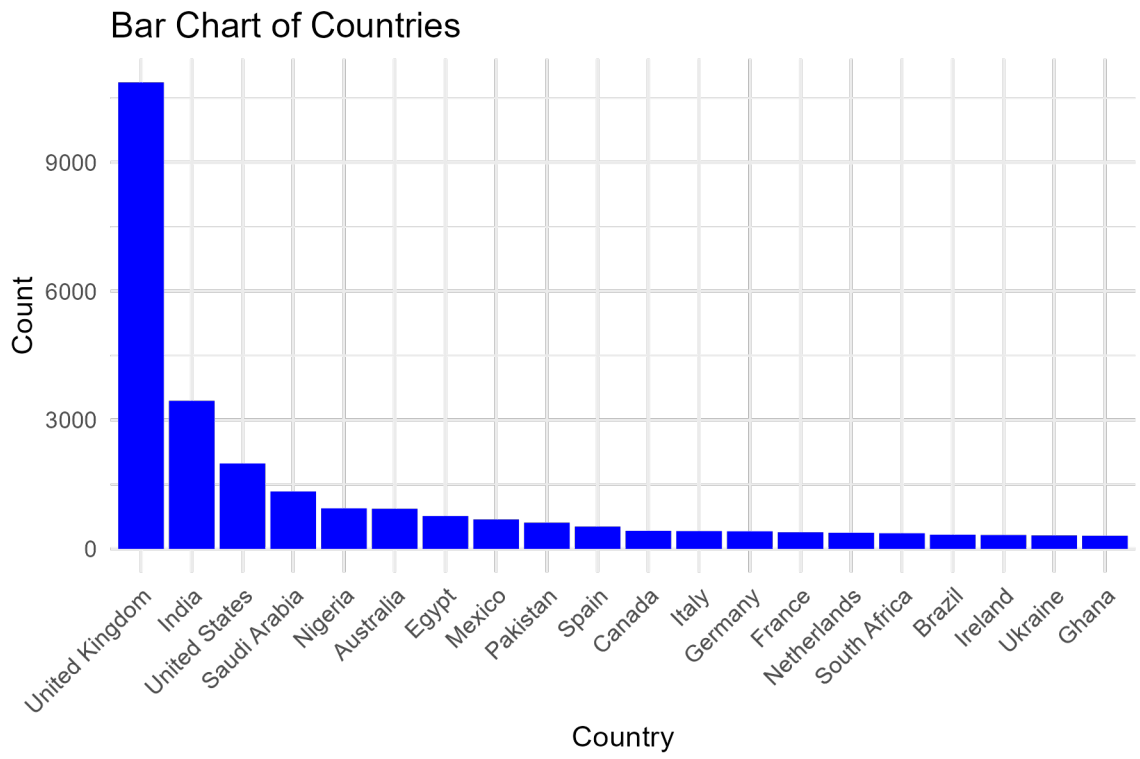
6.3 Build Model

6.4 Assess Model

7 Evaluation

7.1 Evaluate Results

Through the first round of CRISP-DM i found the found the number of learners who enrolled onto the course, and which country they were from. There were learners from 199 countries who participated in the course.



7.2 Review Proccess

7.3 Deterime next steps

8 Deployment

8.1 Plan Deployment

8.2 Plan Monitoringand Maintenance

8.3 Produce Final Report

8.4 Review Project

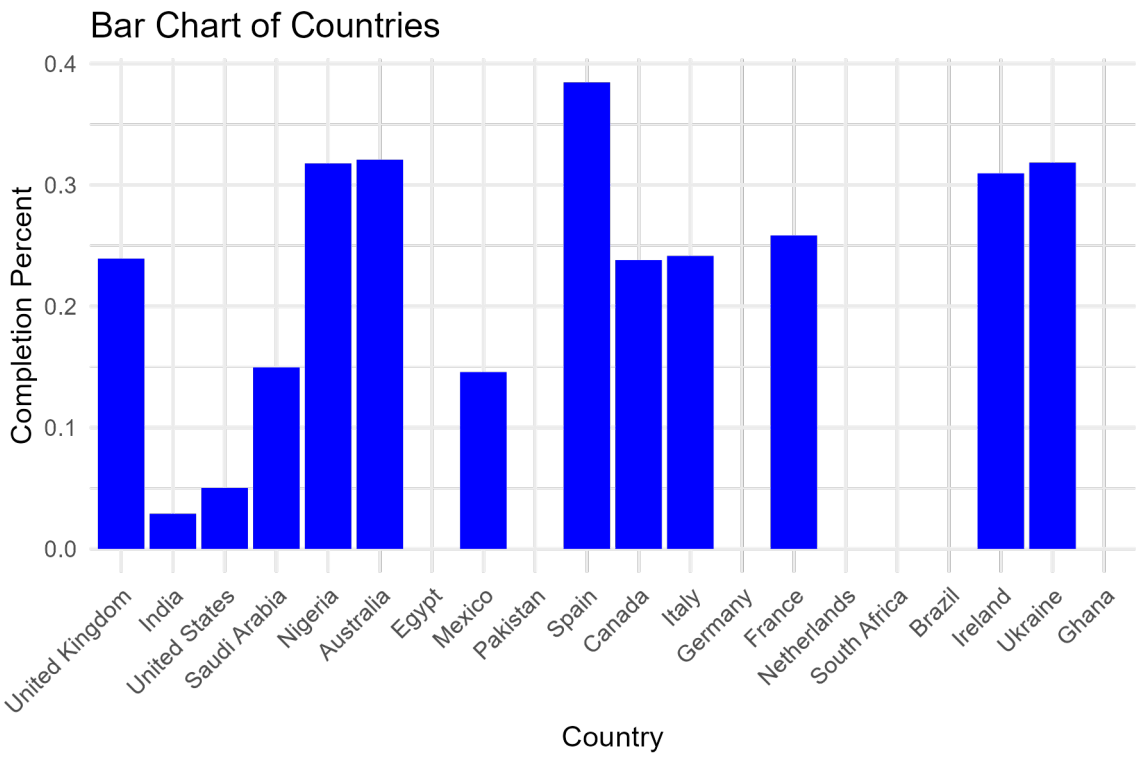


Figure 1: Alt text