CSC8631- Data Management and Explanatory Data Analysis Project

SATISH PILLA

5/20/2022

Contents

1 BUSINESS UNDERSTANDING

Future Learn, an online learning platform, offers a course called Cyber Security: Safety at Home, Online. This three-week self-paced course focuses on critical cyber security issues. The courses are divided into three parts, each of which is studied over three weeks: internet privacy, payment security, and home security. From 2016 to 2018, the course was offered seven times. Throughout the course, a variety of data was collected. This project aims to evaluate the data and get some useful insights that can be utilized to improve the course and establish a viable business model for future courses.

2 DATA UNDERSTANDING

Each time the course is run, around 8 data files are generated.

Datasets:

1. cyber-security-enrolments
2. cyber-security-leaving-survey-response
3. cyber-security-step-activity
4. cyber-security-video-stats
5. cyber-security-step-activity
6. cyber-security-archtype-survey-responses
7. cyber-security-question-response
8. cyber-security-weekly sentiment-survey

The enrollments data set contains the learners' information, including their unique IDs, the course's enrolled and unenrolled dates and times, and additional fields such as gender, nation, age, highest education level, employment location, and current employment status. The vast majority of the students' data was not captured. A figure or action that symbolizes universal human nature patterns is known as an archetype. The information that categorizes the students is contained in the archetype data collection.

The details of the learners who left the course at what stage and for what reason can be found in the leaving survey response dataset. Each learner's responses to quiz questions conducted at a certain moment in each week are collected in the question response data collection. The quiz's results are also included. Learners who started and left the step at particular times are recorded in the step activity data collection. The video statistics data set includes videos of certain steps with titles and information such as time, views, downloads, viewed percentage, and learners' viewed continents. The weekly survey replies data collection contains the learner's input on the course.

3 OBJECTIVE

This investigation has primarily two objectives:

3.1 Selecting the Right Audience.

To select the right audience we must know from which location or country the students are getting most enrolled into the course, Once we know from which location the most people are getting enrolled we can do further analysis on learners gender, education qualification, employment background, status and age etc. All this analysis will help the course provider in targeting the right audience and helps us in better understanding of the students or people showing a strong interest in the course.

3.2 Investigating the delivery methods.

Once we know who the target audience are, we can make course interesting and appealing to the learners. We can do this by investigating the delivery methods. We know that the course is primarily offered in four formats: video, articles, discussion, and a quiz which is conducted at the end of each week. Once we know which format is most popular among the students we can use that format to attract more audience to learn the course.

4 DATA PREPRATION

For my first analysis(Selecting the Right Audience) I have considered cyber security enrollment data files from all of the runs from 1 to 7. For the second analysis(Investigating the delivery methods) I am considering using cyber security stats dataset to know which delivery method is more efficient.

4.1 Merging data:

Combining data from all iterations of the same file genre. For example, enrollment data from each iteration is combined row by row (rbind) on top of each other and similarly it is done for stat dataset. The goal of merging data from several iterations is to create a more comprehensive picture of how the course is functioning, who is enrolling in it, and how they are using it.

4.2 Independent data:

To see how the data changes between different runs of the course, all different file genres from various iterations are also kept separately. For example to know the unique count of learner, we will use each individual enrollments dataset and plot how the course is changing over time form when the course was for 1st to 7th time(2016 to 2018).

4.3 Removing Unkowns:

Since the majority of the fields are unknown, all of the analysis are done by deleting “unknown” values. As a result, this study may or may not represent the genuine population distribution. Since most of the values in the country column are “unknown,” we are using “detected country” rather than “country” from enrollment data to determine where the majority of learners are enrolling.

5 MODELING

This research was conducted using NUMERICAL and GRAPHICAL summaries as modelling tools. This was accomplished by combining R-markdown with a variety of supporting libraries such as ggplot, dplyr, and many others. All of these are combined utilizing the Project template for better project management and reproducibility. The CRISP-DM approach is used for the analysis.

6 DESIGN AND IMPLEMENTATION

Data Analysis:

How has the popularity of a course changed over the last seven years?

To begin, we can look at how the popularity of the course has changed during the last seven times it has been run. This can be investigated by looking at the cyber-security-enrollments data set from all of the Runs and seeing how the numbers have evolved over time.

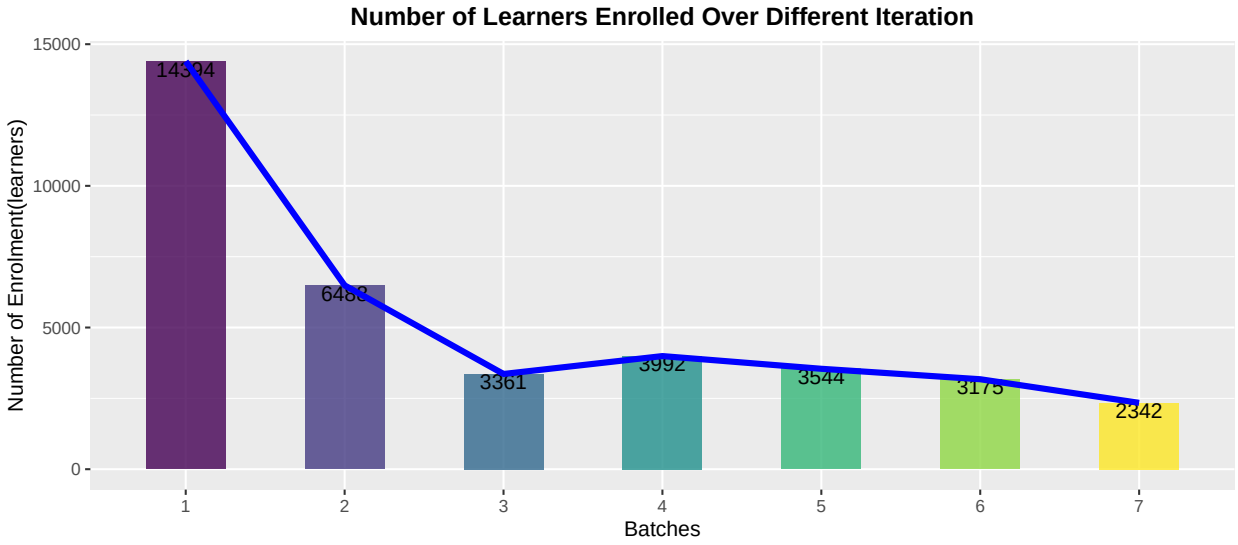

Figure 1: Number of learners enrolled over different iteration

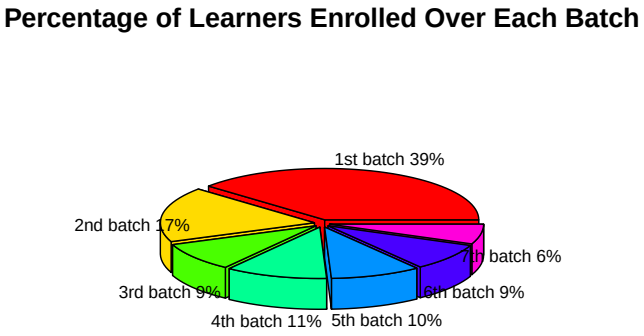

Figure 2: Percentage of Learners Enrolled Over Each Batch

From figure 1 and 2 we can clearly see that there were around 14398 which is 39% of total enrollments over all the course run from 2016 to 2018. The enrollments got decreased as the course run and became less than half by the end of last run. In the 7th run the total enrollments were only 2342 which is just 6% of the total enrollments over all the course run. We can say that as time progressed the popularity has decreased for the course. There may be n numbers of reason for less enrollments over the period of time so We'll try to figure out what's going wrong based on a variety of indicators, such as why students are dropping out and how they feel about the course.

But our primary objective is to investigate the location where course is most popular, so let's now check for the continents and countries were the course is most popular.