CSC 8631 - Future Learn Cuber Security Data Analysis Report

200709330 Hammad Mir

24/11/2021

Abstract

This article analyses the learner data for Cyber Security massive open online certificate (MOOC) by Newcastle University. It also demonstrates the advantages of repeatable data research with r markdown. The data collection comprises 62.csv files, from which enrollment data has been utilized for analysis with target demographic understanding as the overall objective. The study done is a freehand investigation of the data set, with the freedom to construct and answer our own questions.

Introduction

A massive open online course (MOOC) is a type of online course that allows for limitless participation and unrestricted access over the Internet. Many MOOCs offer interactive courses with user forums or social media discussions to support community interactions among students, professors, and teaching assistants (TAs). They also immediate feedback to quick quizzes and assignments, in addition to traditional course materials such as filmed lectures, readings, and problem sets. MOOCs are a well-studied trend in online education that was initially launched in 2008 and became a popular way of learning in 2012. With MOOCs becoming more and more popular, a huge amount of data is generated by the users/learners which can be leveraged to optimize the course to meet the business requirements.

Objective

The main objective of this analysis is to understand the target demographic. Our aim is to develop insights of the learners enrolled in the course. This analysis will helpful in developing the course to have a wider reach and be more appealing to the learners. We analyse the user information such as location data, gender, age, education, etc. to generate these insights. This gives us a better understanding of "Who" the learners are so that the course can be developed to the business needs and market demand.

Methodology

We followed the Cross-industry standard process for data mining (CRISP-DM) methodology for our analysis. CRISP-DM is an open standard process model that describes common approaches for data mining and analysis and is the most widely-used analytics model. CRISP-DM breaks the process of data mining into six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

The stage order is not strict, and travelling back and forth between stages is always needed. The process diagram's arrows represent the most essential and common relationships between phases. The diagram's outer circle represents the cyclical nature of data mining. After a solution has been installed, the data mining

process continues. Lessons learnt during the process might prompt new, more focused business inquiries, and succeeding data mining operations will profit from prior ones' experiences.

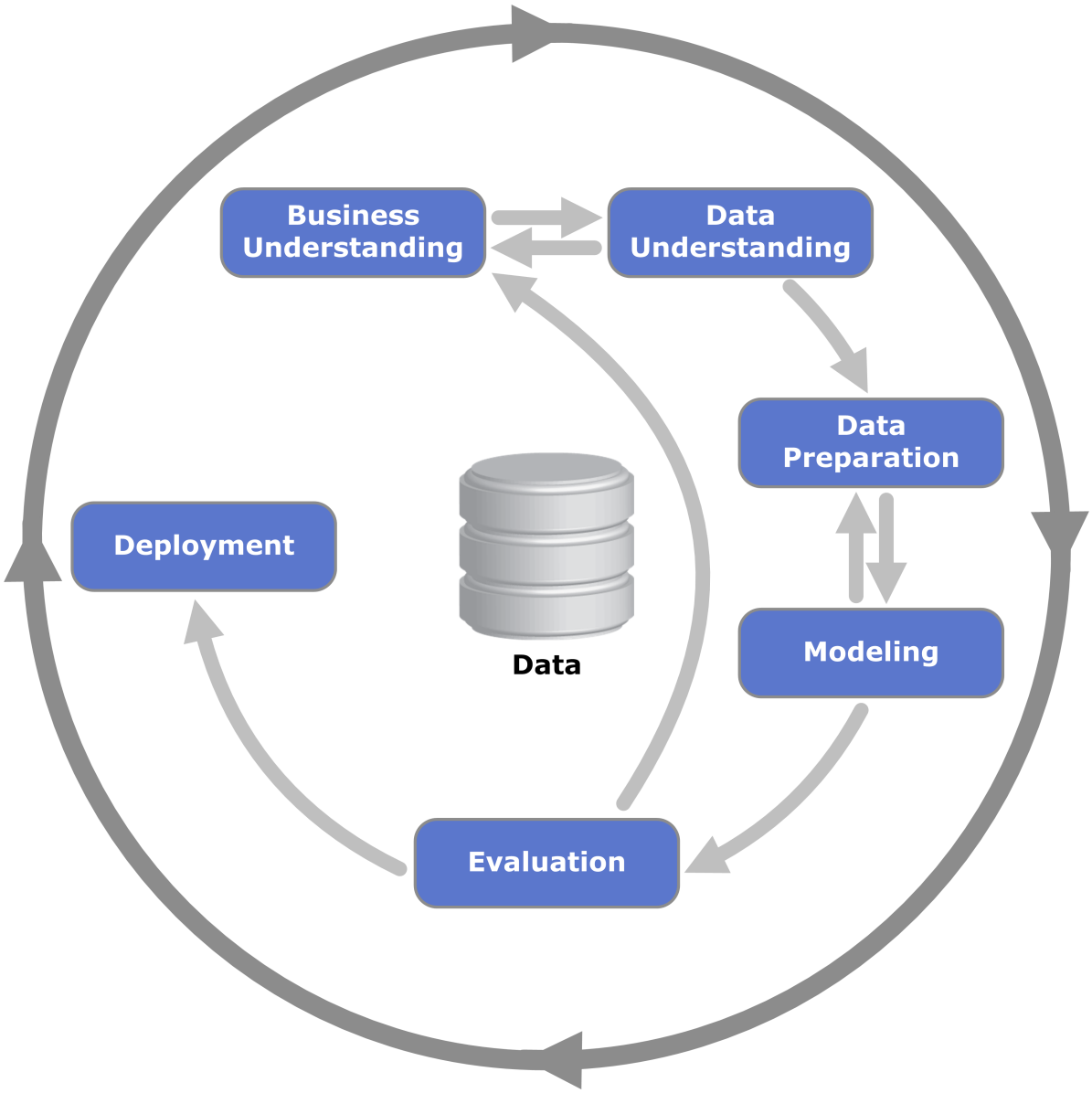


Figure 1: CRISPDM_Process_Diagram

Dataset

The dataset for Cyber Security: Safety at Home, Online, in Life course includes 62 with files containing user data such as follows: - Archetype survey - Enrolments - Leaving survey response - Question response - Step activity - Team members - Video stats - Weekly sentiment survey responses

The dataset comprises data gathered from seven runs of the course. During our analysis, we concentrated solely on the enrolment data, which contains information about each learner such as gender, age, education, nation, and so on.

Data Processing

We pre-processed the enrolment data to ensure that the country data for each user is retained either from the "detected country" or from "country" (set by learner), with priority being given to "country" field. The file contained many missing values, so the data was filtered during analysis, dropping rows where gender, age, highest education and employment status were missing.

Assumption

- Since the bulk of the fields are "unknown", all of the analyses is performed by deleting the rows with "unknown" values. As a result, we are assuming this subset represents the genuine population distribution.
- We are assuming "detected country" is a good enough measure for the location of the learner as most of the values in the "country" field are "Unknown".

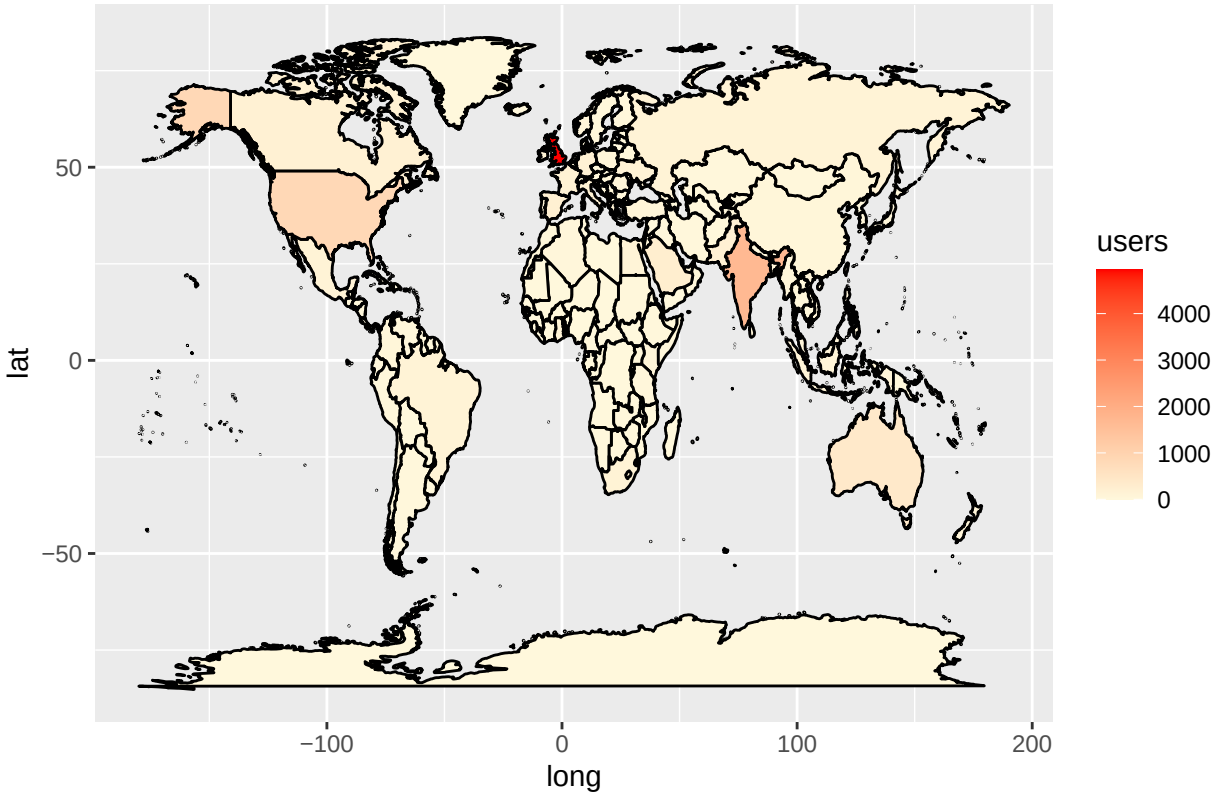
Analysis

Following the CRISP-DM model, we performed our analysis in more than one cycle as explained below, with our main business objective being target demographic understanding:

Cycle 1: Where learners are from.

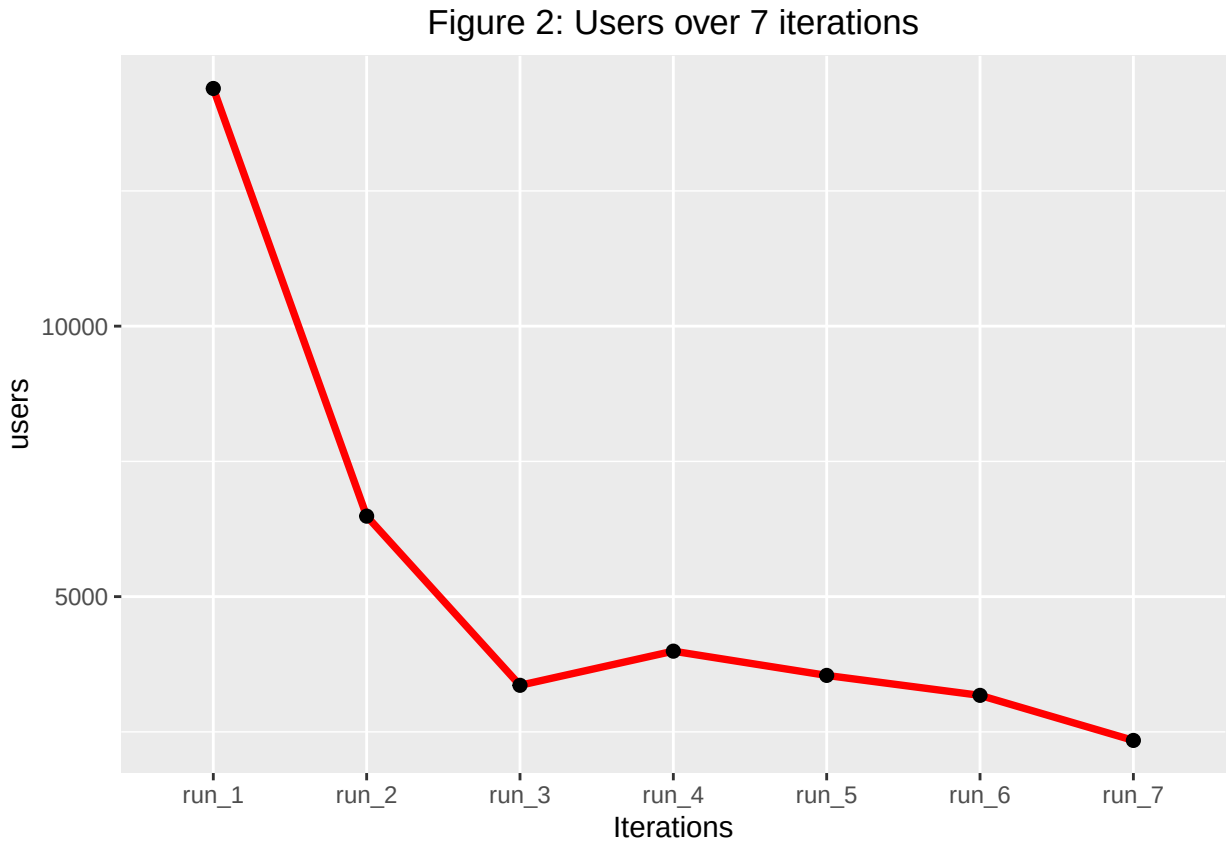
We analysed the country the learners are from and generated a map plot with heatmap describing the number of users per country as shown below. As can be seen from the generated map, most of the users are from the UK, followed by India and US. Observing the map over 7 runs of the course, we find the similar findings/distribution as stated above.

Figure 1: Country−wise users



Cycle 2: Enrolments over 7 runs.

To understand how the course has been performing over the 7 iterations, we analysed the data and generated a plot showing enrolments over the 7 runs. From the figure we clearly observe that the enrolment over 7 runs has been decreasing.



Cycle 3: Who the learners are.

Following the location and enrolment analysis, we decided to analyse the learner data further to understand who they are. The inference from these features is as follows:

Learner Gender Distribution From the gender distribution plot of the learners below, we clearly observe that the number of learners that are males is the highest, followed closely by female learners. Those identifying as non-binary and other are just a few. We observe that overall throughout the 7 runs of the course, the gender distribution trend remains almost the same.

Learners age distribution The age distribution plot of the learners shows the maximum number of learners belong to 26 to 35 years age group and the least number of learners are under the age of 18. Overall we observe the same trend through the 7 iterations of the course.

Learners highest education The education distribution plot shows clearly that the number of learners with a university degree is the highest followed by those with a masters degree. We also observe the least number of learners have less than secondary, university doctorate and apprenticeship level educational background. In general we observe the same trend all through the 7 runs of the course.

Learners employment status The employment status plot shows that the number of learners with a full time job is the highest, with un employed and non working learners being the lowest in number. As we have been observing so fat, the same trend can be observed for all of the 7 runs of the course.

Learners employment area The employment area plot below suggests that the highest number of learners belong to the it sector followed closely by those in the teaching and education sector. We observe a similar trend over 7 runs of the course.

Figure 3: Gender Distribution Plot

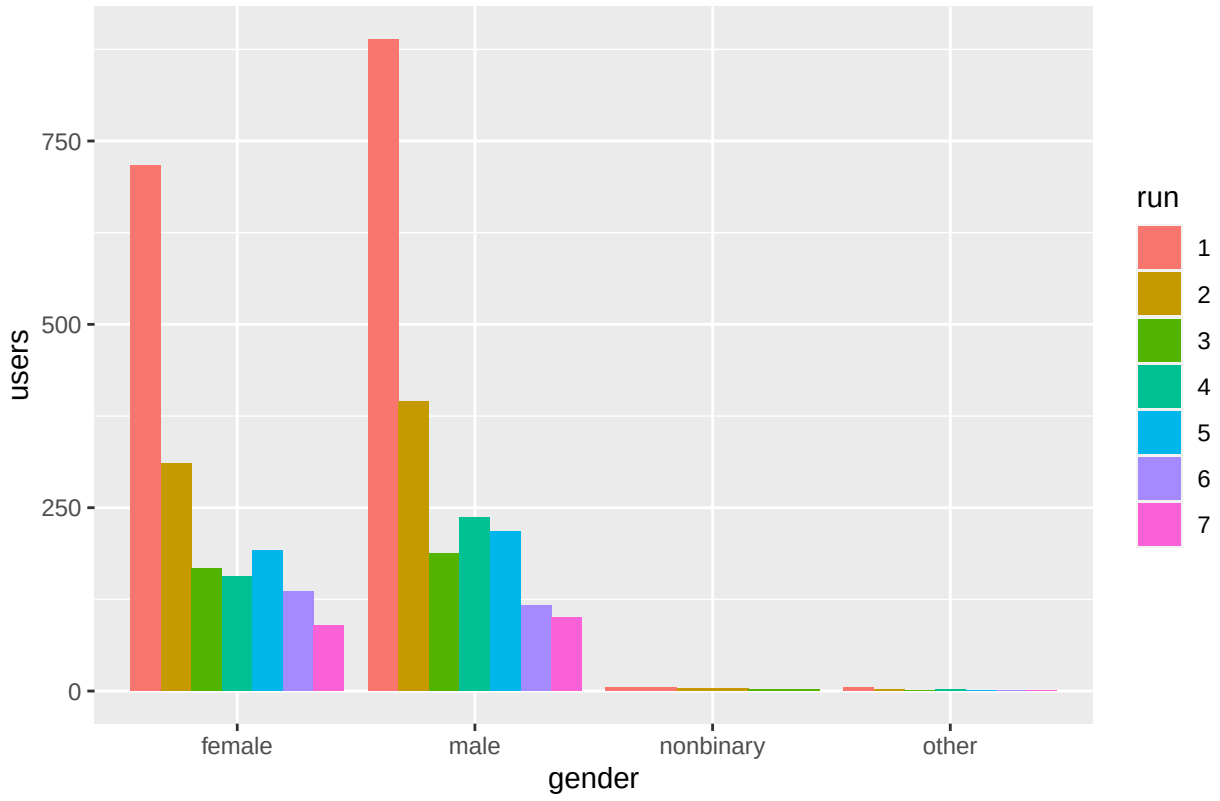


Figure 3: Age Distribution Plot

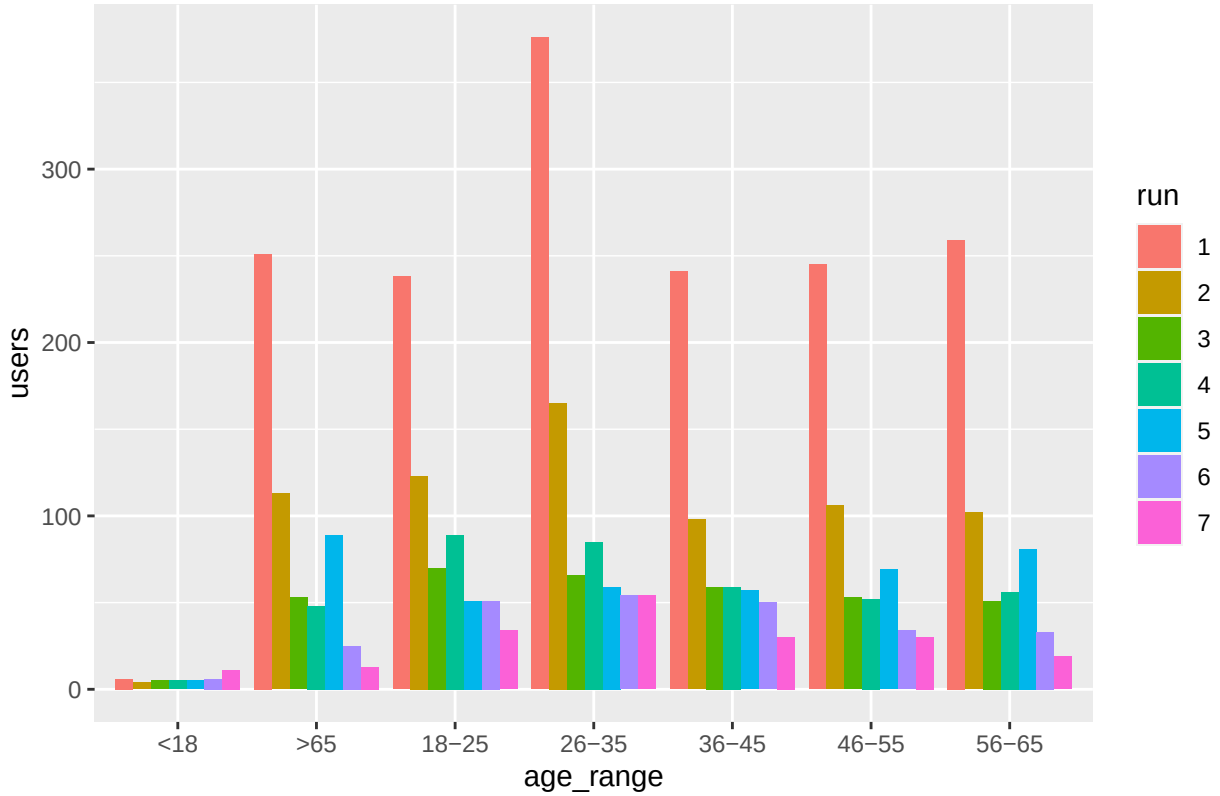


Figure 4: Education Distribution Plot

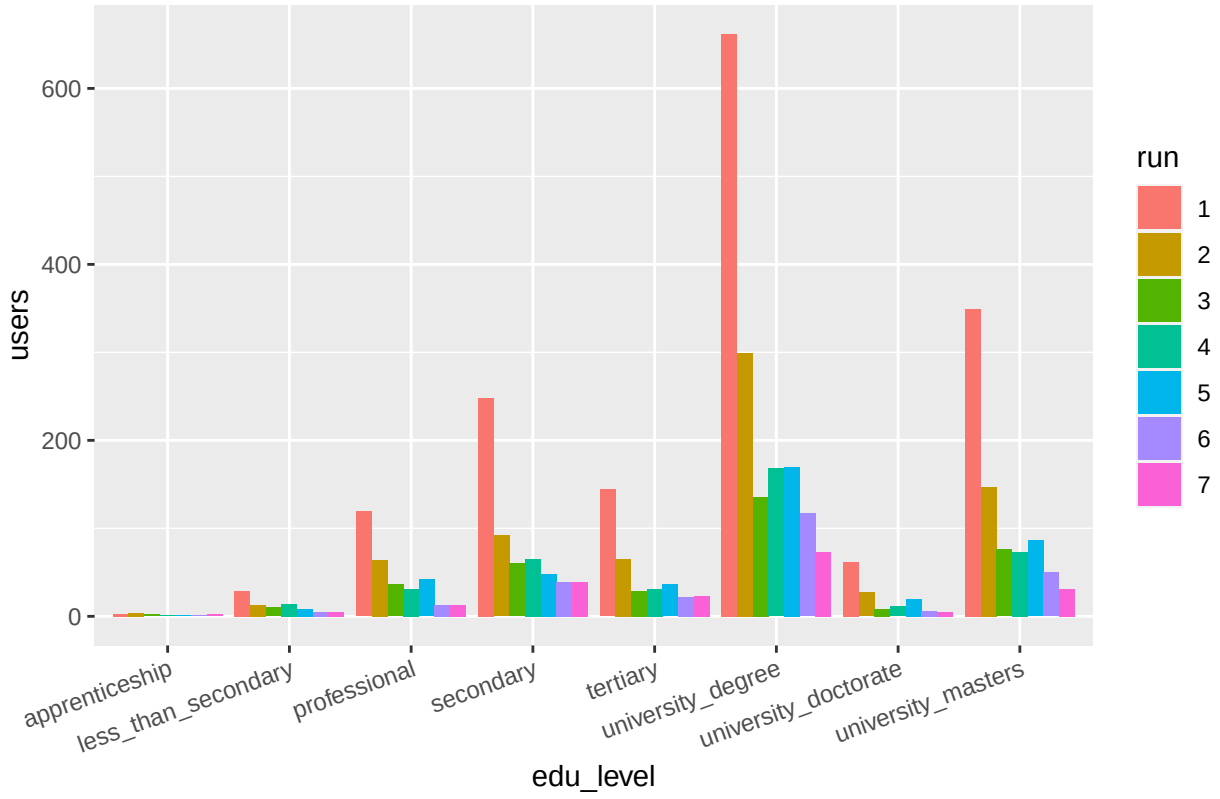


Figure 5: Employmetn Status Plot

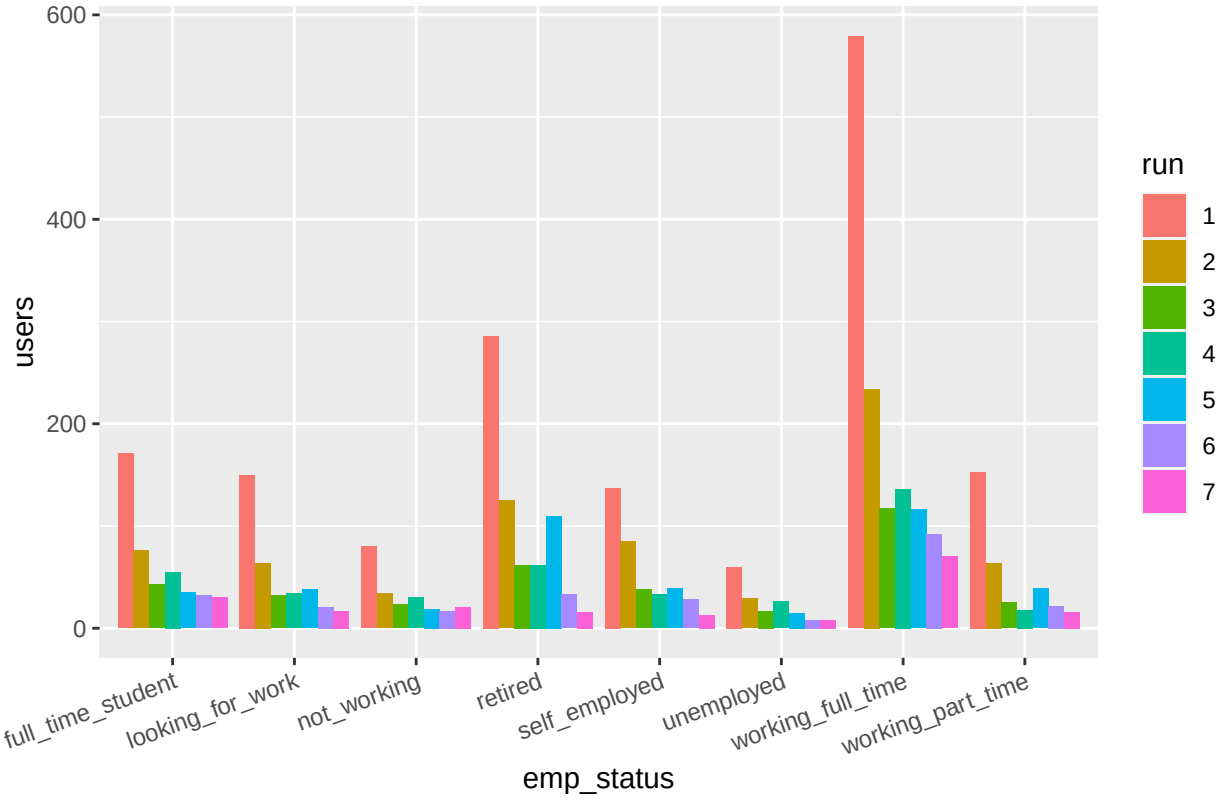
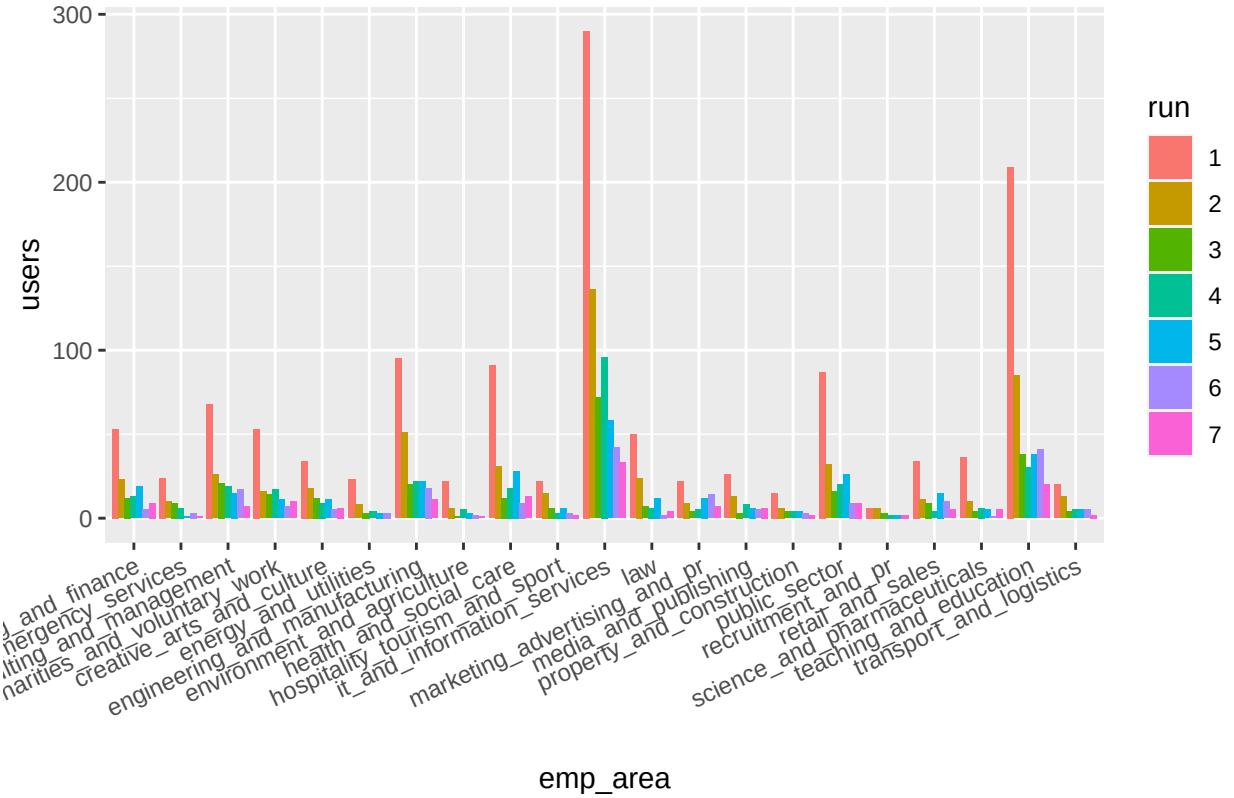


Figure 6: Employment area Plot



Conclusion

The analysis suggests that for targeting right audience we can say that people from all age range, in IT and Education sector, and working full time are interested so, ads can be optimized by keeping these criteria into consideration.