CrispDM_assignment

Jake Bonnici

2022-11-13

Introduction

A massive open online course (MOOC) entitled "Cyber Security: Safety At Home, Online and in Life", consists of 7 runs of data. For this assignment an investigation was done on two main business questions and a keen importance was given in applying the CRISP-DM procedure. ProjectTemplate was used to structure the investigation, mainly its data wrangling and analysis, to ensure reproducibility. Rstudio's ggplot/dplyr were the main tools used to carry out the analysis, whilst R Markdown was used to carry out this report. This report is aimed at stakeholders to generate insight at aspects of the data that can be useful towards stakeholders.

CRISP-DM

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (**CRISP-DM**) is mainly a model used for data science processes. It consists of 6 sequential phases;

1. Business understanding
2. Data understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

The upcoming sections are split in the sequential order of a CRISP-DM cycle and will briefly explain the purpose of each phase, whilst also adding the application done from the project on the MOOC data sets.

Business Understanding

The first phase of the procedure, focuses on understanding the objective of the project and any requirements from a business‘s perspective. MOOC would like to acquire more clients, and it‘s needs to understand which continents and countries are most popular with its course. Additionally, MOOC wants to understand which devices are preferred in order to follow their lectures. Finally, the drop rate from the MOOC can be asses over all the 7 runs and separately by each run.

This information will be collectively used to further improve the methods they use to cater for their clientele, with regard targeting other countries to increase their diversity and the form their lectures take to prioritize the most used devices. The second question might help in giving insight on whether the business is heading in the right direction and adjust accordingly, depending on the increasing/decreasing drop rates.

Data Understanding

The second phase of the CRISP-DM consists of collecting the data and start summarizing the data to familiarize ourselves with it and also identify and data quality issues and simple insights within the data.

This help us to get a good overall perception of the limitations of our data and what possible questions can/cannot be answered.

The quality of the data varied over the 7 runs. The first problem was that for the first and second runs their were a total of 6 and 7 data sets whilst the rest consist of 8 data sets. Although some of the column names such as age range and gender would have been interesting insight to work with within certain data sets, their quality was lacking, either by an overwhelming amount of data and occasionally whole missing data sets in some runs. The best data across the 7 runs was selected and worked with. The data sets that were used for this project are $cyber - security - \_video - stats$ and $cyber - security - \_enrolments$. The video statistics were available from run 3 onwards, whilst the enrolments were available for all runs. The video statistics was the most consistent when compared to the rest of the data sets. Additionally, it was the best data set to answer the first question mentioned above.

Data Preparation

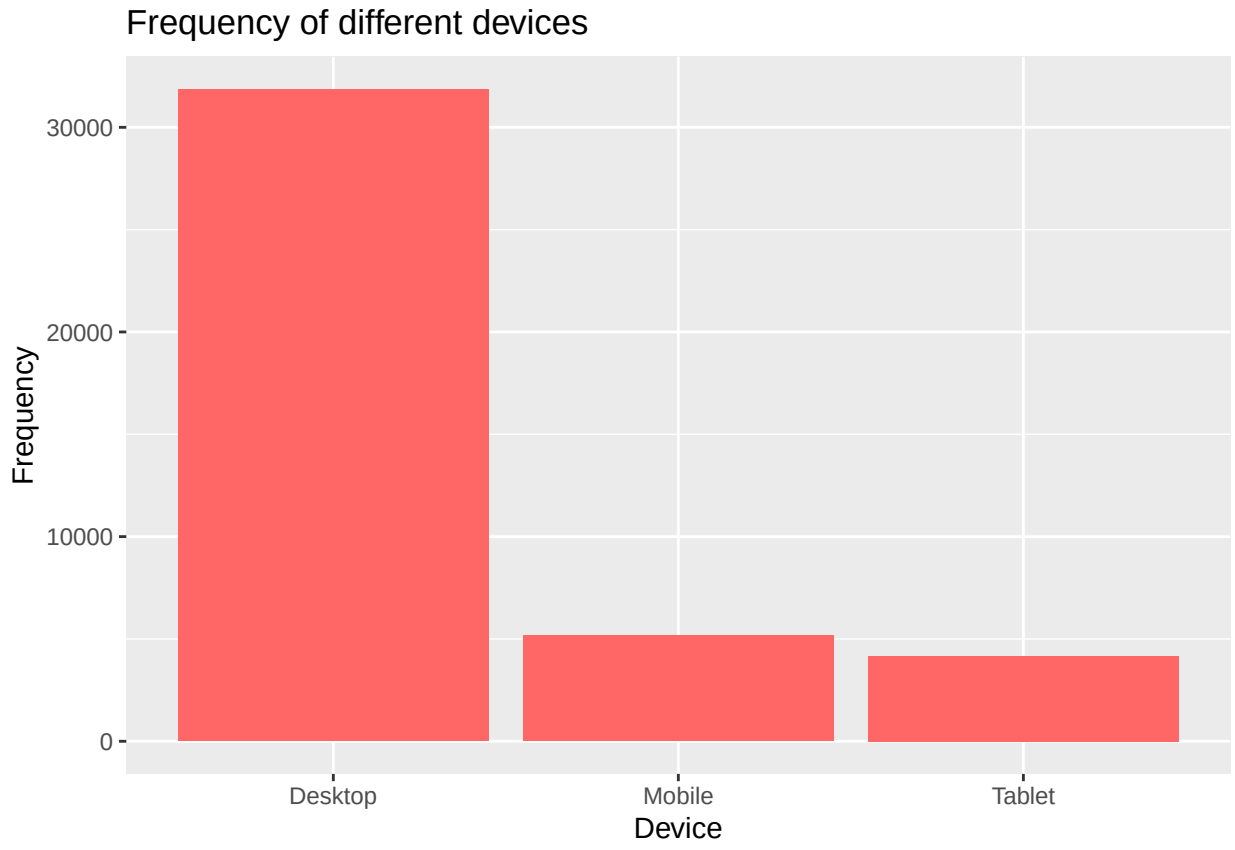
This phase collects the required data and forms it in a manner that can be used into our modelling tools from the initial raw data.

For the first cycle of the CRISP-DM, the data preparation was first performed on the video statistics data set. The data for all the available runs were binded together to have one data frame. The percentage numbers were converted back to total numbers for all columns describing frequency of different devices and and continents. Additionally, the means of the total views for each step lectures were obtained. For the second cycle, the data was also prepared using the enrollment data set. However, this dataset had data quality issues with a good amount of its columns and since they provided no useful insight they were removed from the data frame. Comparing both data frames the video data set had 65 number of rows while the enrollment data set had 37296 rows. Clearly, showing a large difference between the two data sets in terms of length.

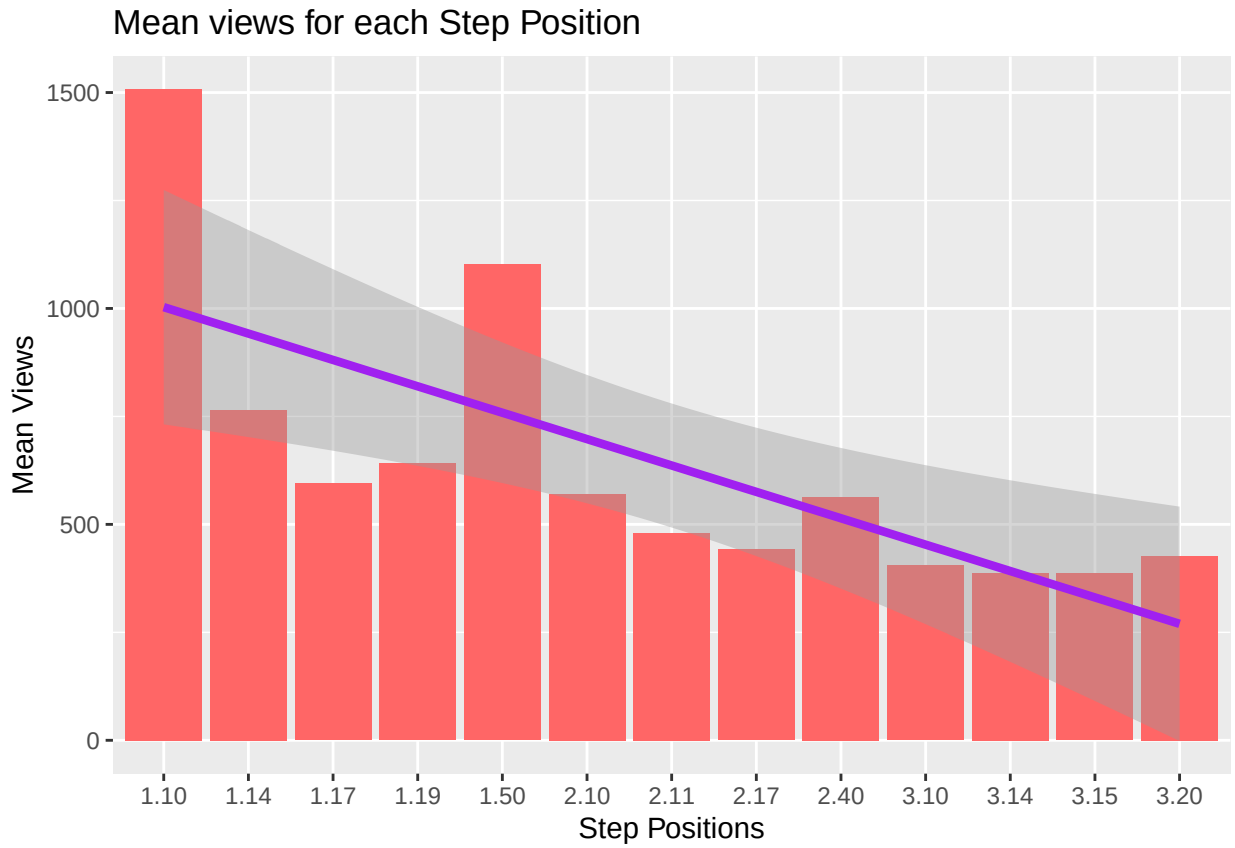
Modeling

At this stage the data is ready for use and a model is built for the use of the data. Understanding which analysis and visualizations best fit for the case. If some of the analysis required different needs in data formation then the process is reverted back to the data preparation phase.

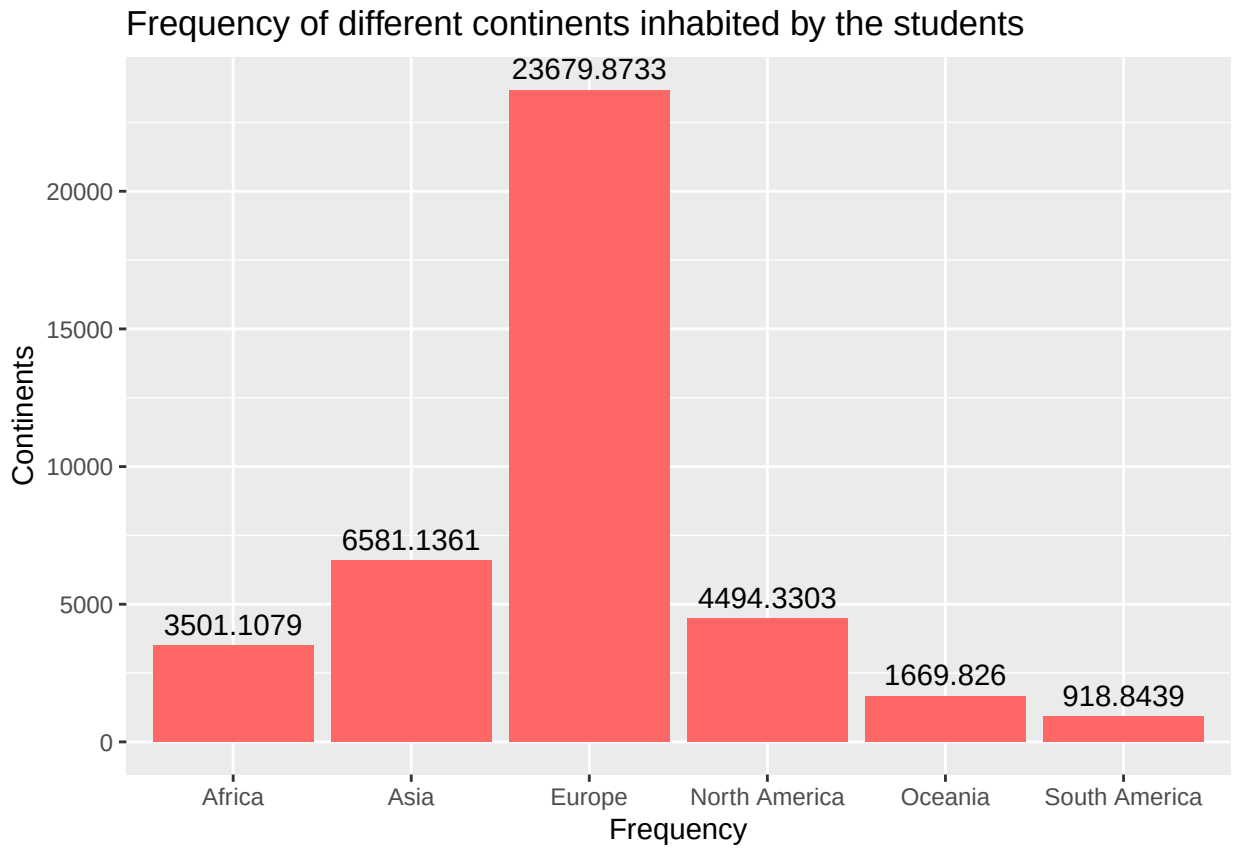
For the first cycle of our CRISP-DM, the exploratory analysis was first done on the devices that were used by the students. Only the three most popular once were shown, the rest were considered negligible due to high lack of users using them. As seen in the *Frequency of different devices plot*, the desktop was the dominant device being used, with a far second and third being the mobile and tablet, respectively. This information can be used to create interactive tools which are mostly catered towards desktop use with the confidence that the majority of the students will be able to access them. Additionally, the correlation between the total views of a lecture and its length was calculated to be -0.0633198, hence showing close to no significance between these two variables.



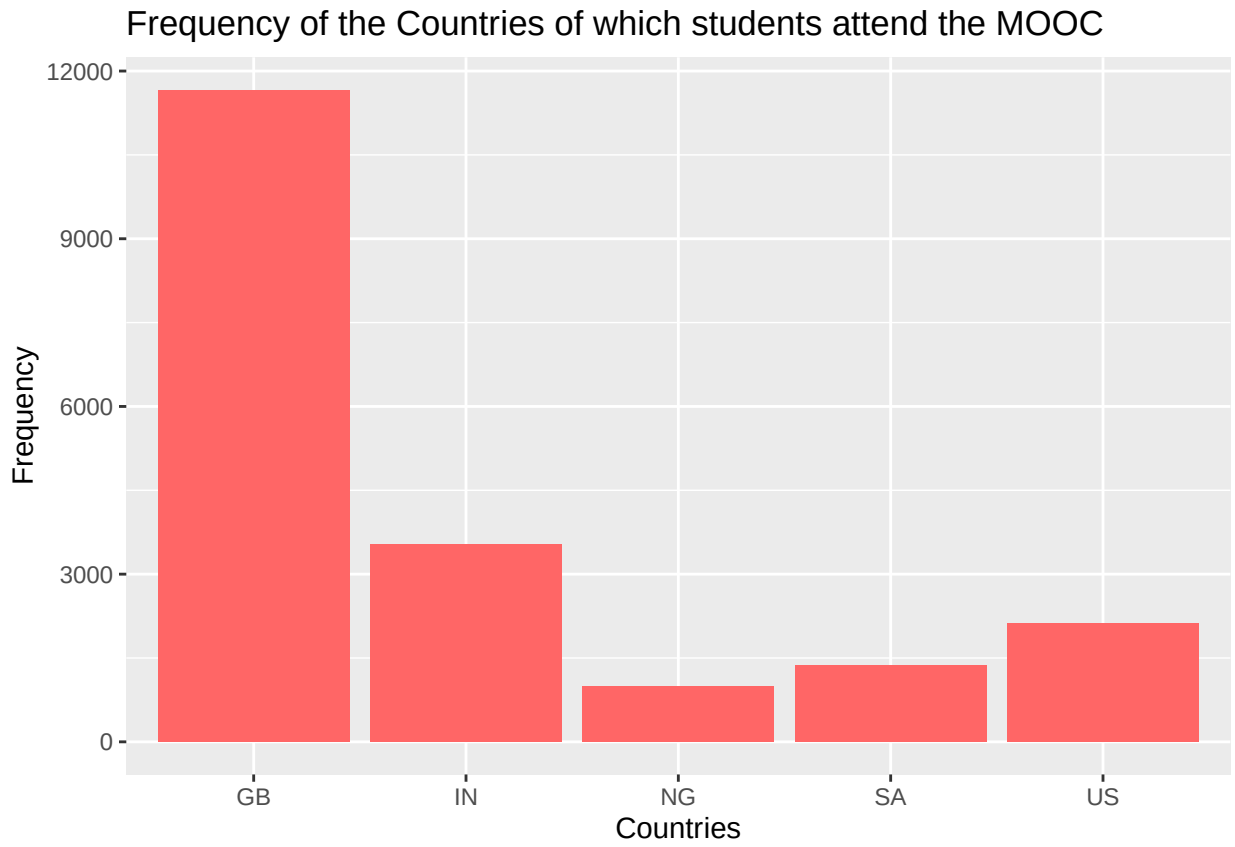
The second graph showing *Mean views for each Step Position*, shows a clear trend with decreasing views as the course progresses with their respective confidence intervals. The correlation of the graph's axes variables was calculated to be -0.684333 hence confirming this hypothesis. This lack of continuation might be to various of reasons. For example; the course might not feel satisfactory to the students either by being to hard or easy for them, hence demotivating their continuation in watching the lectures. The weekly sentiment survey data set might provide better insight in such a question, however due to time constraints it was not possible to explore this statement any further.



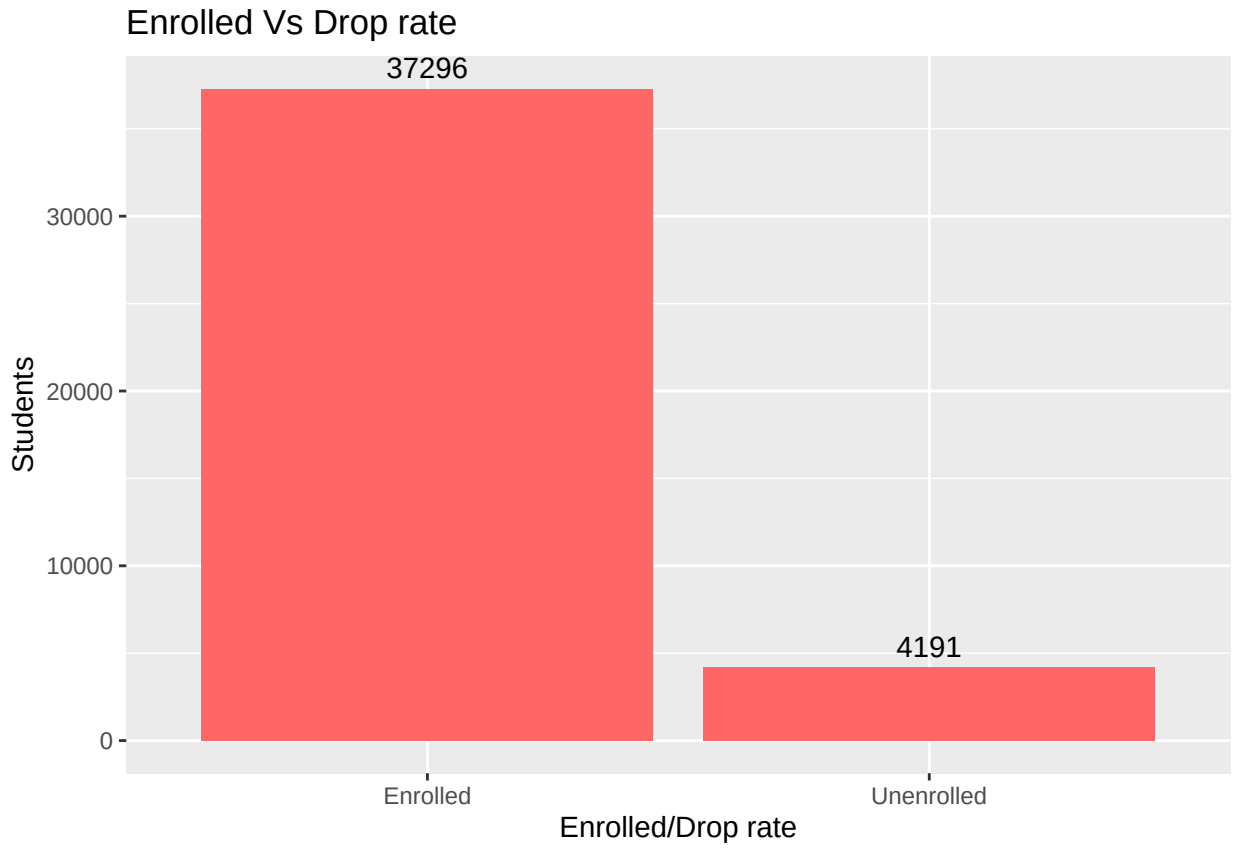
The figure below shows the *Frequency of different continents inhabited by the students*. Antacrctica was excluded from this plot since no students were attending the MOOC course from that continent. It seemed that the MOOC had its majority of attending students from Europe. Not closely followed by Asia and North America. This indicates could acquisition methods to Europeans, however for various reasons a lot of potential and scaling potential in the continents mentioned above. However, this does not address the possible different barriers to entry within other continents.



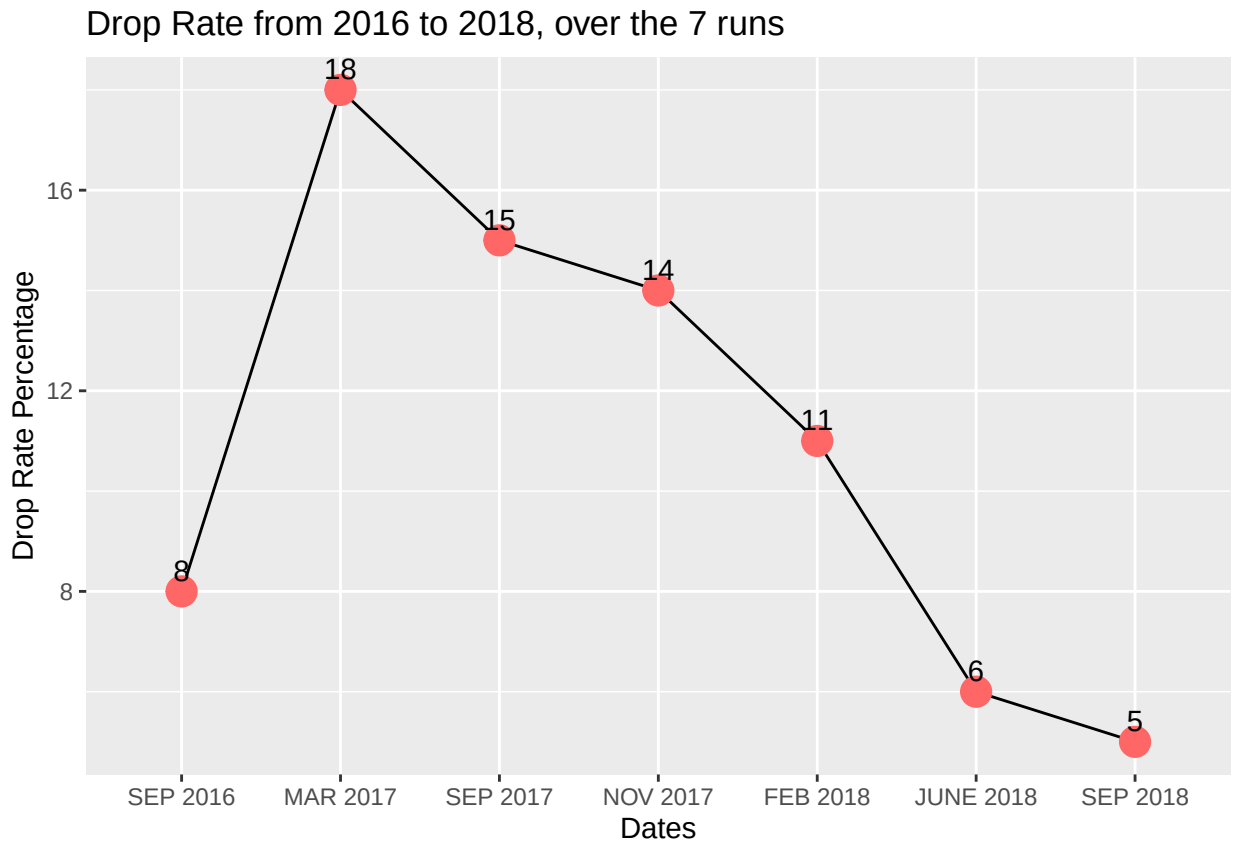
From the enrollment data set, the countries were retrieved to have a deeper understanding apart from than the continents. Therefore, first the data for this question was understood and wrangled in a manner to be modeled. From the figure below, Great Britain was the dominating country which followed the MOOC lectures. The closest two countries were India and the United States. This shows which parts of the Asian and North American logistics were most interested in Cyber Security, at the time.



The figure below shows the *Enrolled Vs Drop rate* with 37296 enrollments in the beginning of the year and 4191 drop outs, hence an 11 %, drop out rate. This shows a decent retention of students completing the MOOC.



The graph below describes the drop rate from 2016 to 2018, over the 7 runs. A sharp increase can be seen from its first to the second run of MOOC, however a consistent decline in drop rates follows afterwards, all the way from 2017 to the 2018. This means that the MOOC has adopted an adaptable method in improving the course run after run. Hence motivating and keeping their students mentally stimulated to their last drop rate of just 5 %.



Evaluation

The penultimate phase of the process consists of having the model of the data analysis ready (as shown above), however, before committing to the deployment phase, a check that all the objectives are reached and if not, the process is repeated from the necessary phase again. Additionally, a decision on how this data mining results are going to be used should be achieved. For example, the data for the student‘s sentiments can be used in tandem with the drop out rates, to particularly better understand which aspects of the course were bothering the students, and causing them to drop out. This would require reconsideration and understanding of a new data set, its wrangling and modeling in a proper manner.

Deployment

The knowledge that was gained from the above two cycles need to be organised and presented in a way the user on the other end, such as a customer can use it. Depending on what is required, this phase can be as simple as generating a report or as a complex as implementing a repeatable data mining process across the enterprise. The whole process here was organised and structured by using the *ProjectTemplate* library in RStudio, compressed to a zip file and deployed onto the Canvas page.

Conclusion

There is no definitve research showing how frequently data science teams use CRISP-DM. A poll was conducted by Data Science Process Alliance, showing which methodologies are most famous for data mining. Crisp-DM was the most frequently used followed by “SCRUM” and “Kanban”. This data was spanned over 12 years. The CRISP-DM method feels natural and logical, and has the ability to align business understanding with technical work.