CSC 8631 - EDA Report

Syed Mohib Raza || Student ID: 200740241

December 10, 2021

Abstract

This document provides an analysis of massive open online certificate (MOOC) learning data through Newcastle University. And along with it shows the benefits of reproducible data science using r markdown. The data set contains 52 .csv files, of which three files have been used for performing analysis. The analysis performed is a freehand analysis of the data set with the liberty to generate our own questions and find their answers.

keywords: MOOC data set, Exploratory Data Analysis (EDA)

Introduction

With the advent of the internet, online education adoption has been a much debatable area but nevertheless, it has been rising. The world is now more connected than ever and online education is enabling thousands of individuals who aspire to study world-class education and all that at the comfort of their homes with flexible times. Thus virtually removing the barriers of inaccessible education and promoting free and fair resource sharing.

Due to the unfortunate impact of the covid-19 pandemic in 2020, there has been a significant increase in the use of online education platforms also called massive open online certificates or MOOCs. The dataset contains of 52 .csv files, of which 3 files have been chosen which have more number of features than others and can be used to performing data analysis for business understanding. The analysis & results are presented after the data set details section.

Data Exploration and Preparation

The first data set comprises of enrollment information, had several unknown values and had to be cleaned to generate a clean sample. All the data uses ETS that is Extract, Transform and Load to perform EDA. Enrollment data set has 13 columns listed below.

```
## # A tibble: 6 x 13
##   learner_id  enrolled_at  unenrolled_at role  fully_participa~ purchased_state~
##   <chr>       <chr>        <chr>         <chr> <chr>            <chr>
## 1 160d6600-e~ 2016-08-10 ~ ""            lear~ ""               ""
## 2 4dc22fed-6~ 2016-05-24 ~ "2018-10-30 ~ lear~ ""               ""
## 3 ecdd37db-0~ 2016-05-19 ~ ""            lear~ "2016-09-22 16:~ ""
## 4 988964c9-7~ 2016-05-19 ~ ""            lear~ ""               ""
## 5 f1493366-1~ 2016-09-19 ~ ""            lear~ ""               ""
## 6 25cc3b46-a~ 2016-08-30 ~ ""            lear~ "2016-10-25 12:~ ""
## # ... with 7 more variables: gender <chr>, country <chr>, age_range <chr>,
## #   highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

The second data set contains data for the video statistics. It can be useful to get deep insight on how the enrolled students are utilizing the online courses. Details below are about some of the columns used in the analysis.

1. step_position - Course is divided into 3 main sections and each of them has some sub sections, these divisions are called as step position.
2. total_transcript_views - Total number of students that used transcripts to understand the course.
3. total_downloads - Total videos downloaded by the students.
4. north_america_view_percentage - Percentage of students that are from North America.

```
##  [1] "step_position"                 "title"
##  [3] "video_duration"                "total_views"
##  [5] "total_downloads"               "total_caption_views"
##  [7] "total_transcript_views"        "viewed_hd"
##  [9] "viewed_five_percent"           "viewed_ten_percent"
## [11] "viewed_twentyfive_percent"     "viewed_fifty_percent"
## [13] "viewed_seventyfive_percent"    "viewed_ninetyfive_percent"
## [15] "viewed_onehundred_percent"     "console_device_percentage"
## [17] "desktop_device_percentage"     "mobile_device_percentage"
## [19] "tv_device_percentage"          "tablet_device_percentage"
## [21] "unknown_device_percentage"     "europe_views_percentage"
## [23] "oceania_views_percentage"      "asia_views_percentage"
## [25] "north_america_views_percentage" "south_america_views_percentage"
## [27] "africa_views_percentage"       "antarctica_views_percentage"
```

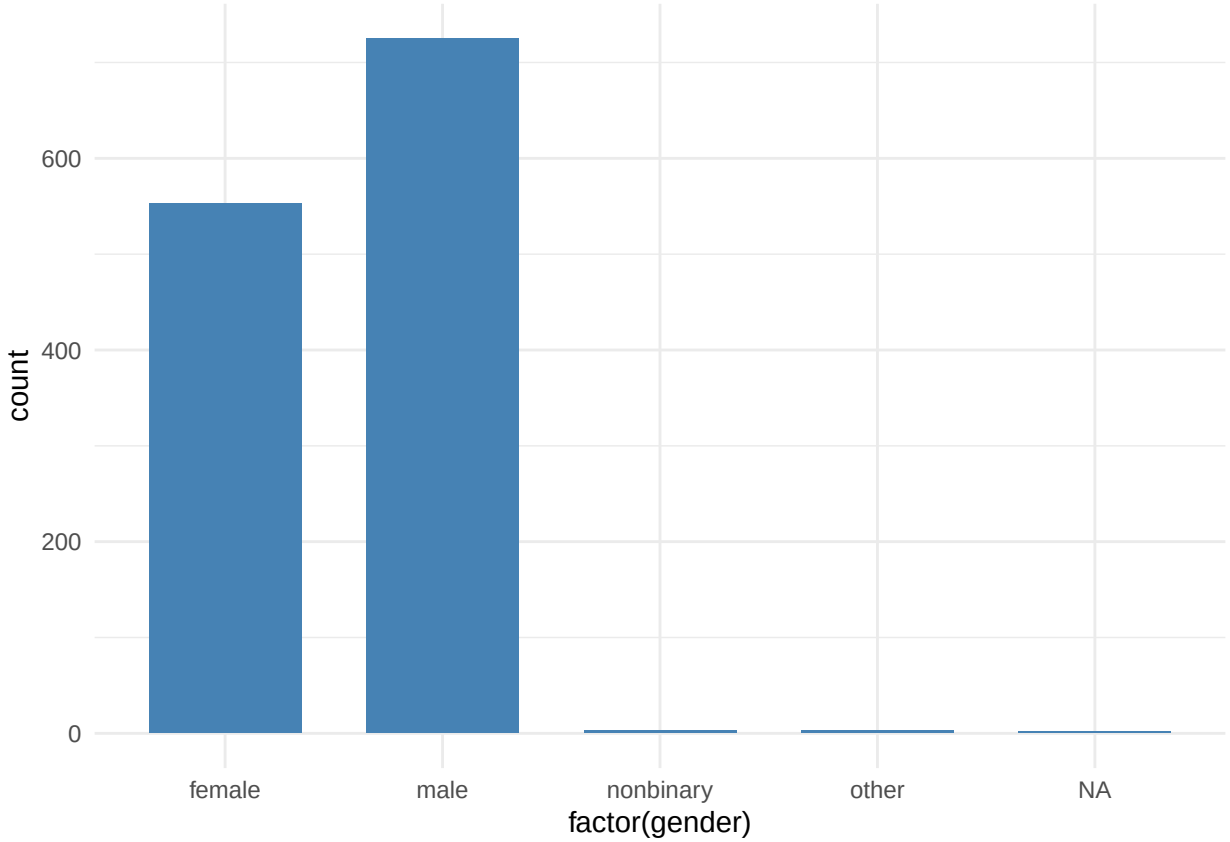
The quiz taken by the registered students is the subject of the third data collection. The data can be used to examine the performance of the students taking the course, and the feedback collected can be used to improve the process.

```
## # A tibble: 6 x 10
##   learner_id quiz_question question_type week_number step_number question_number
##   <chr>      <chr>         <chr>               <int>       <int>           <int>
## 1 77454a73-~ 1.7.1         MultipleChoi~           1           7               1
## 2 77454a73-~ 1.7.1         MultipleChoi~           1           7               1
## 3 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 4 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 5 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 6 f27eec8c-~ 1.7.1         MultipleChoi~           1           7               1
## # ... with 4 more variables: response <chr>, cloze_response <lgl>,
## #   submitted_at <chr>, correct <chr>
```

Data Set 1

Plot - Gender

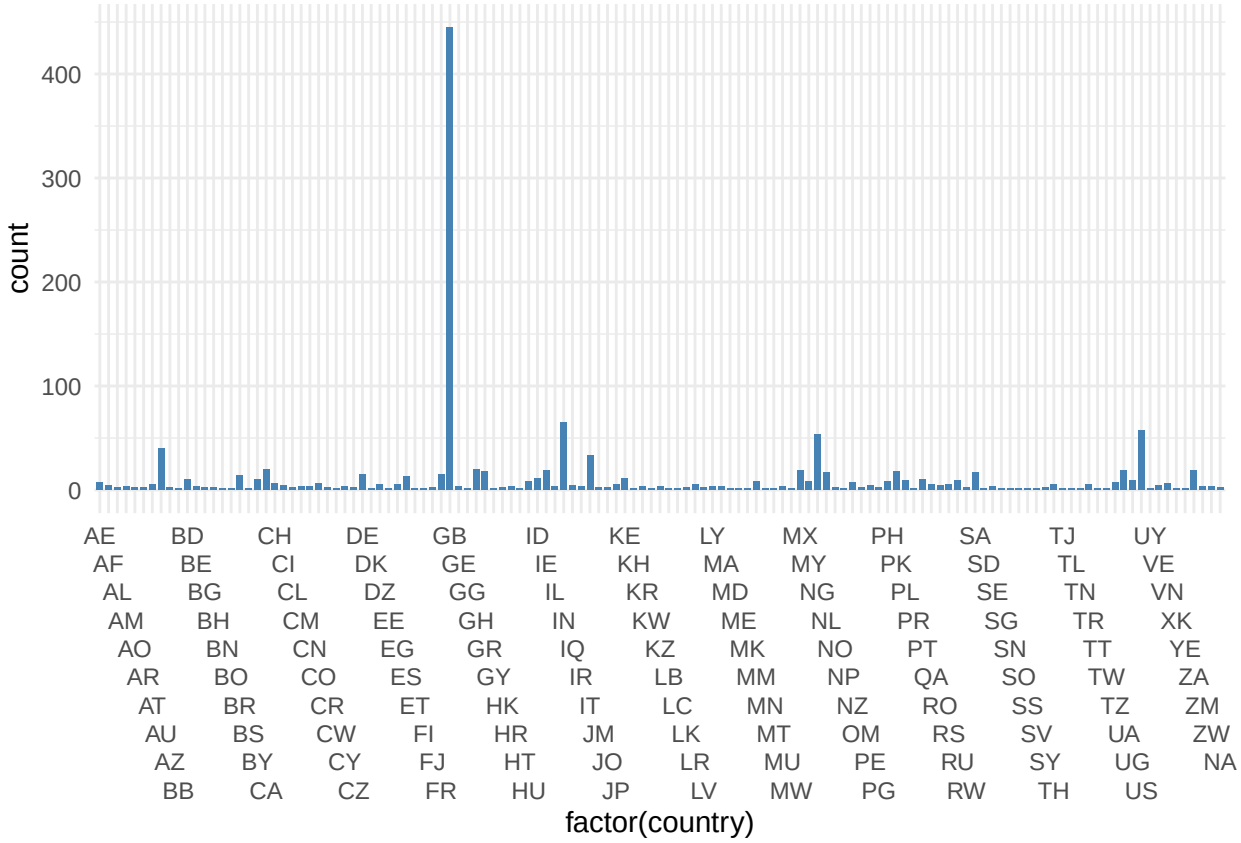
In the below plot we are identifying the enrollments based on the gender.



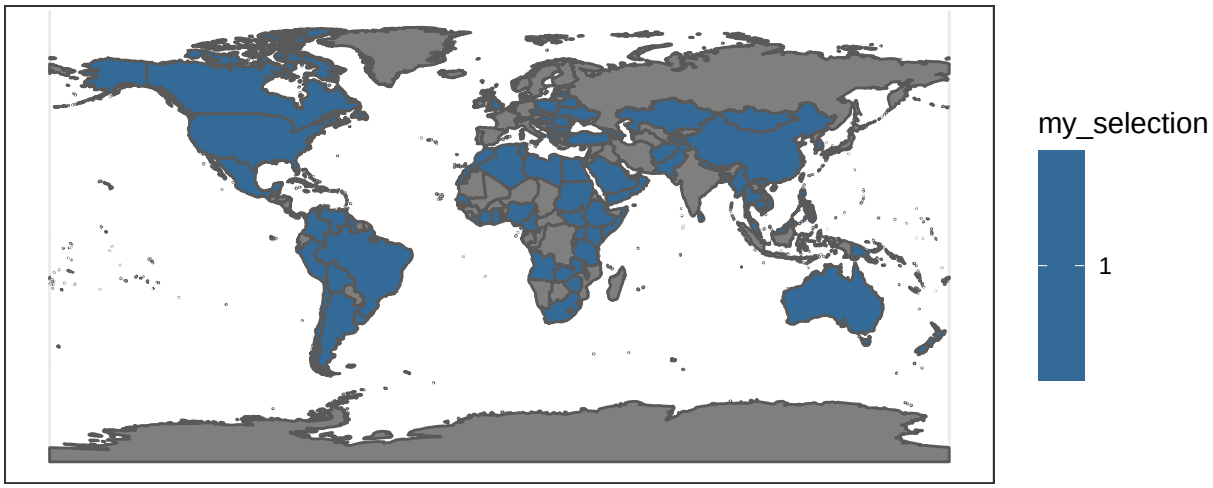
From the graph we can observe that the number of "Males students" enrolled in the course are comparatively more than the "females".

Plot - Country

The country plot helps in understanding the demographics of the audience. From the bar plot of the count vs countries we can observe that most number of participants have been from GB (Great Britain - 445 count).



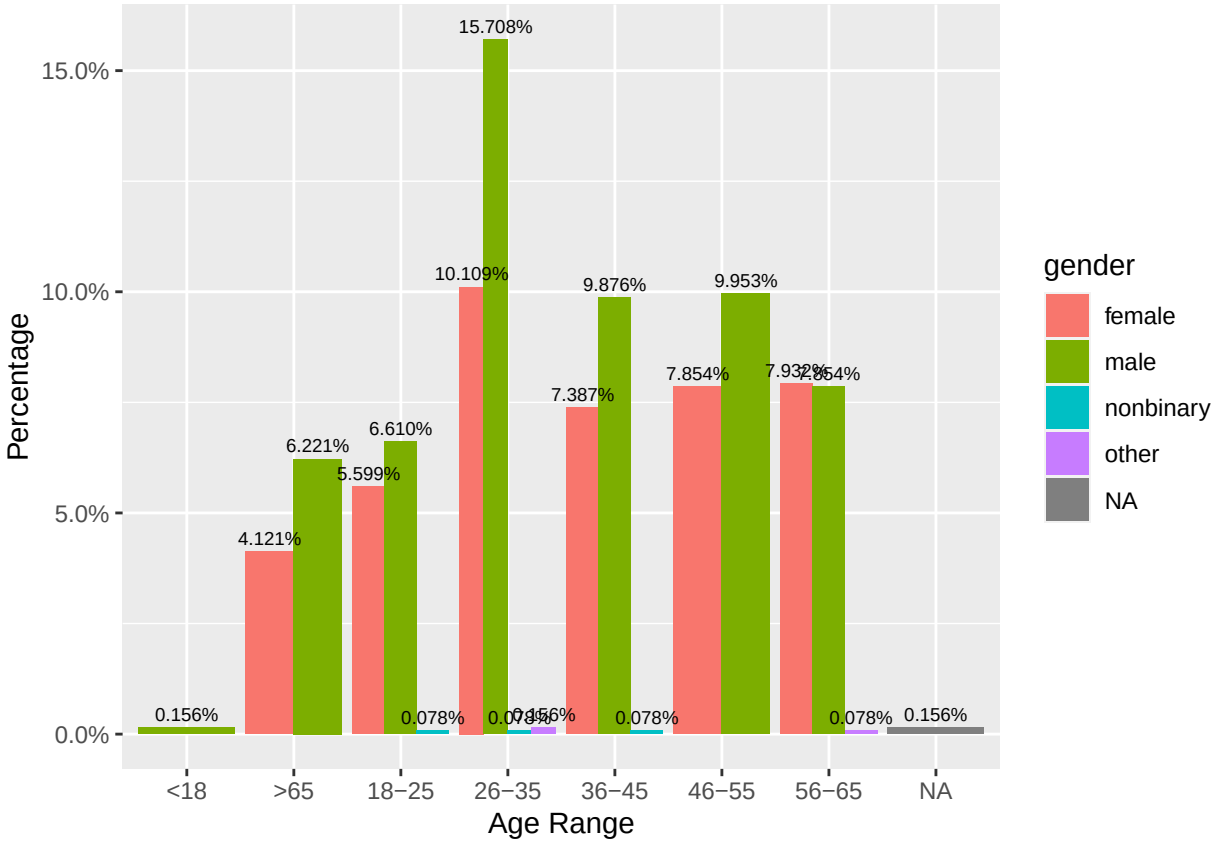
World map



The map helps us to identify the areas of no coverage. We can see that places from Africa, Middle east & south east countries just above Australia, have little to no engagement with the course moreover surprisingly India with its rich population has zero number of students that have taken this course.

Plot - Age range vs gender

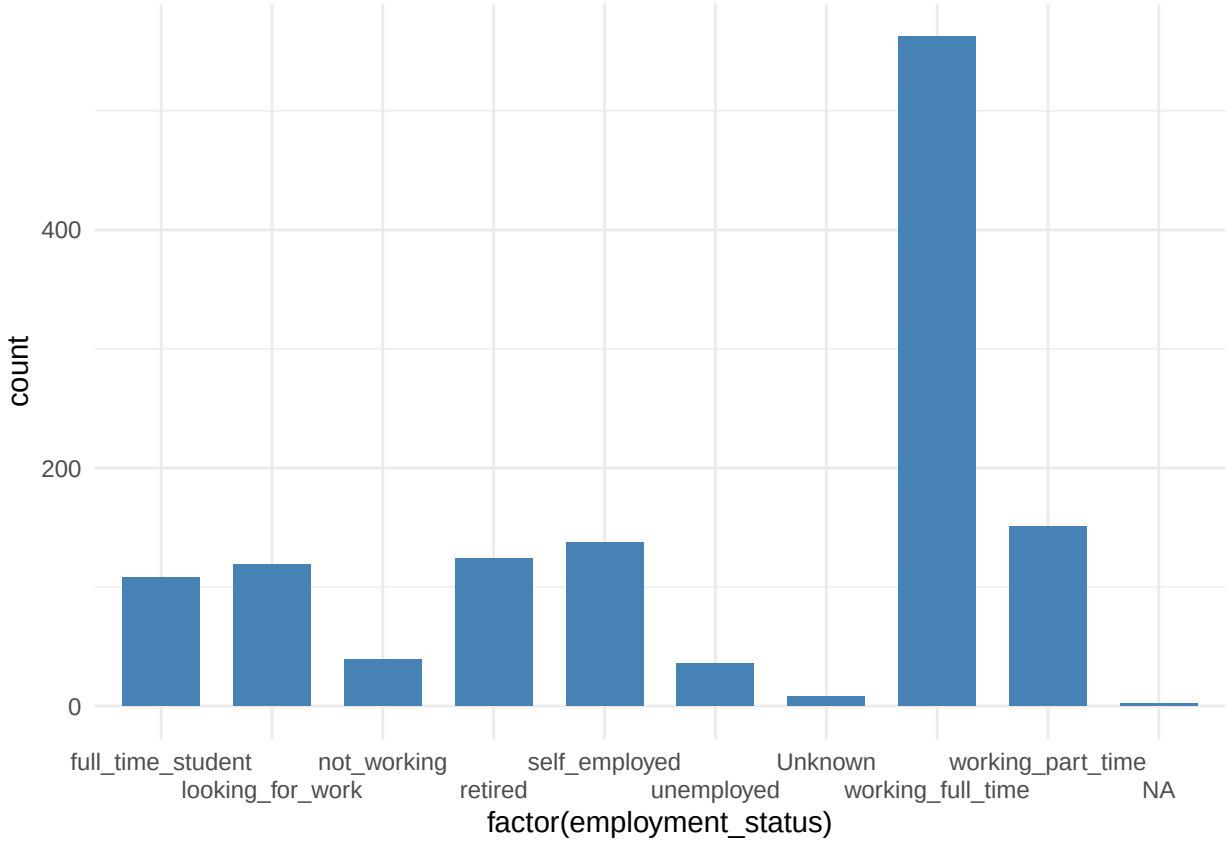
For identifying various age groups and their respective frequency of engagement with the course we can best customize the course curriculum and teaching methods so as to make better and more interactive course for all the respective age groups. From the plot we can also analyze the trends among various age groups.



Plot shows us that the highest number of enrollments have been from youth which are from the age group of 26 - 35. This age group is mostly out of college & typically requires a lot of skill development for filling up proper job skills. Males (15.7%) are the highest enrollments for this age group which may imply greater need of skill development because of much more competition. However, it is exciting to see middle age ranging from 36-45 (9.8%) & 46-55 (9.9%) in significant numbers with male to female ratio (1.3% & 1.2% for respective age ranges) relatively same as compared to age range 26-35(1.5%).

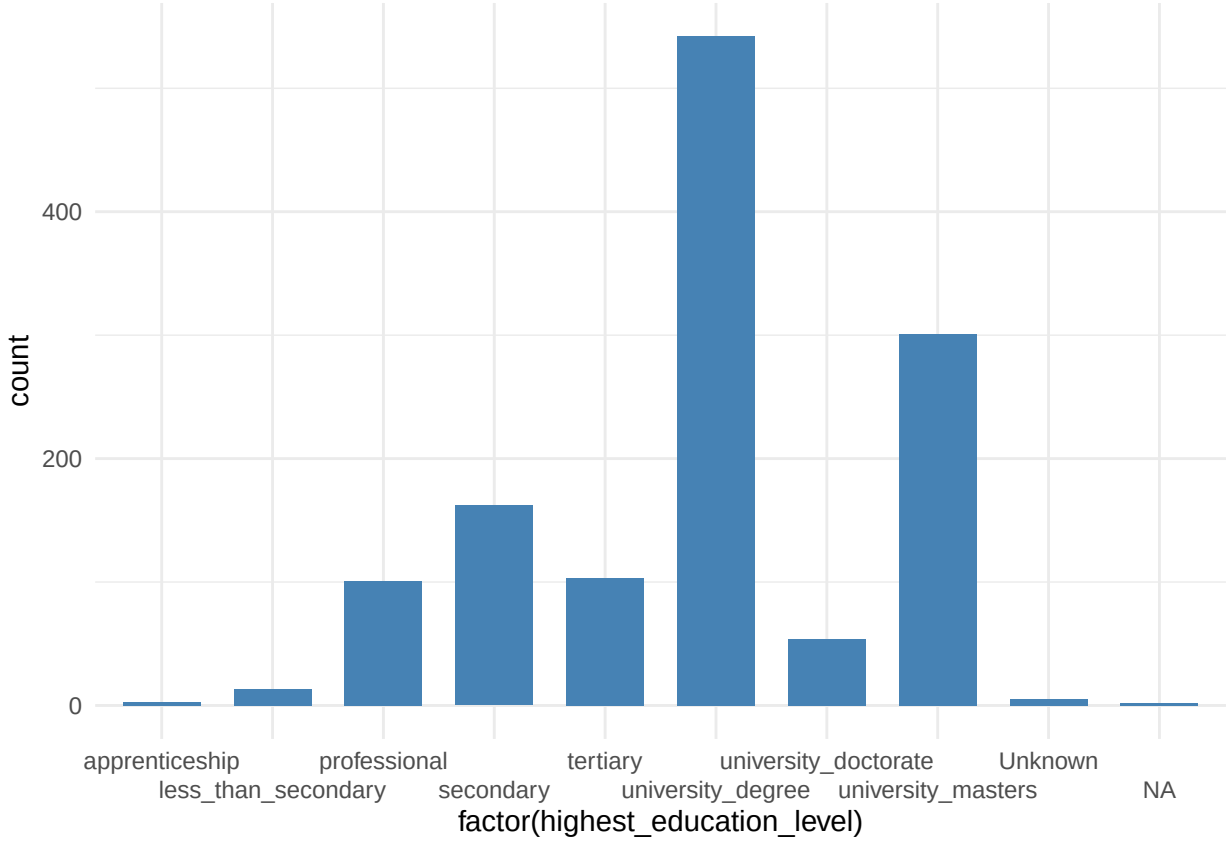
Employment plot

The data is useful in understanding the job status and curate the curriculum in a way to up skill the learners.



From the above plot we can observe that the full time working population have been the most attracted to this course which intern suggests high demand for skill improvement in professional environment. They can obtain new & desired professional skills while being employed full time, as it helps to fund their education thereby maximizing their results.

Education plot

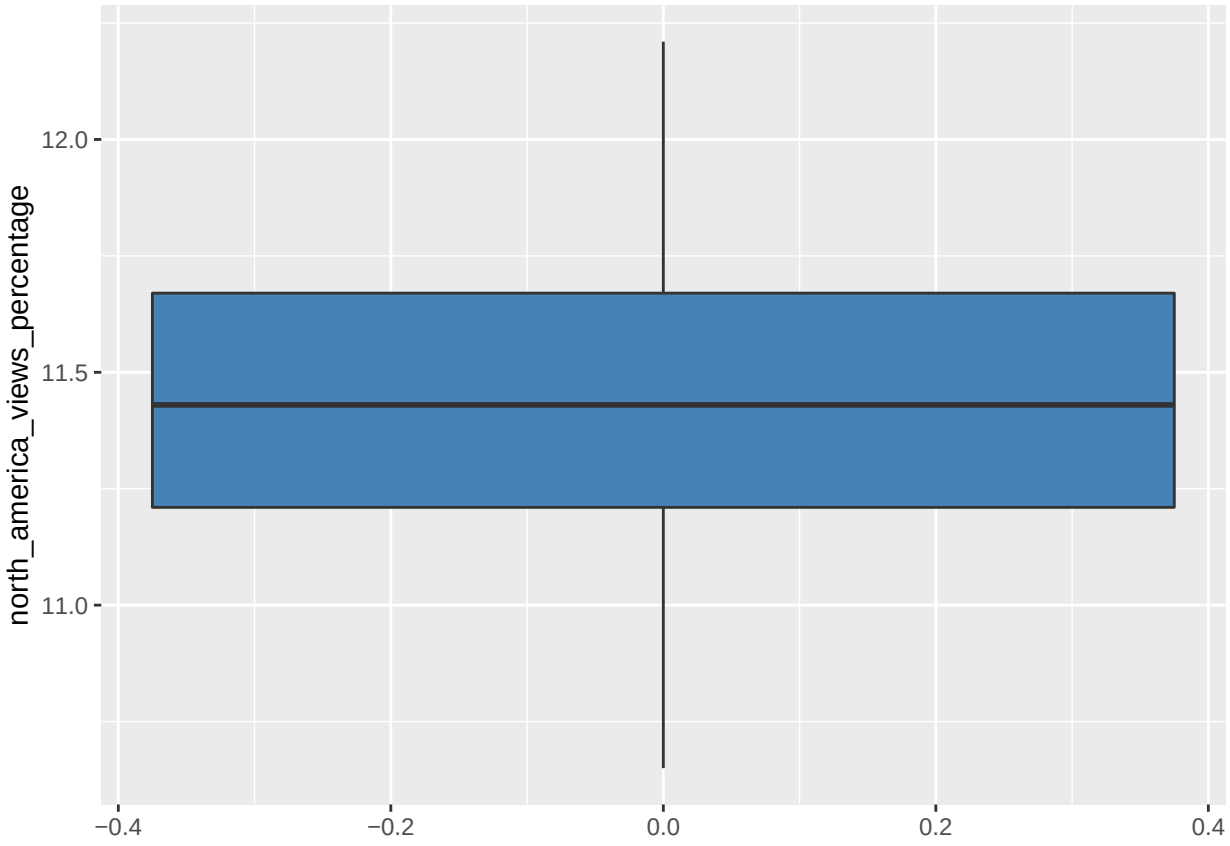


The plot indicates that the students with the University degree have the highest enrollments count which in turn is a sign of the need to improve the quality and practicality of education at the university level. This is further proved by the second-highest enrollments count which is also from students who have completed their Masters from a University.

Dataset 2

Box Plot - Videowatch time

The below boxplot represents the statistical summary for the video watch time for students in North America. from above graph we can infer that the median is about 11.43, mean about 11.45. The above statistics can also be obtained via `summary` function like below.



North America views percentage summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.65   11.21   11.43   11.45   11.67   12.21
```

Asia views percentage summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.24    9.11    9.51   10.03    9.92   16.09
```

Europe views percentage summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.15   64.90   65.60   64.73   66.25   67.25
```

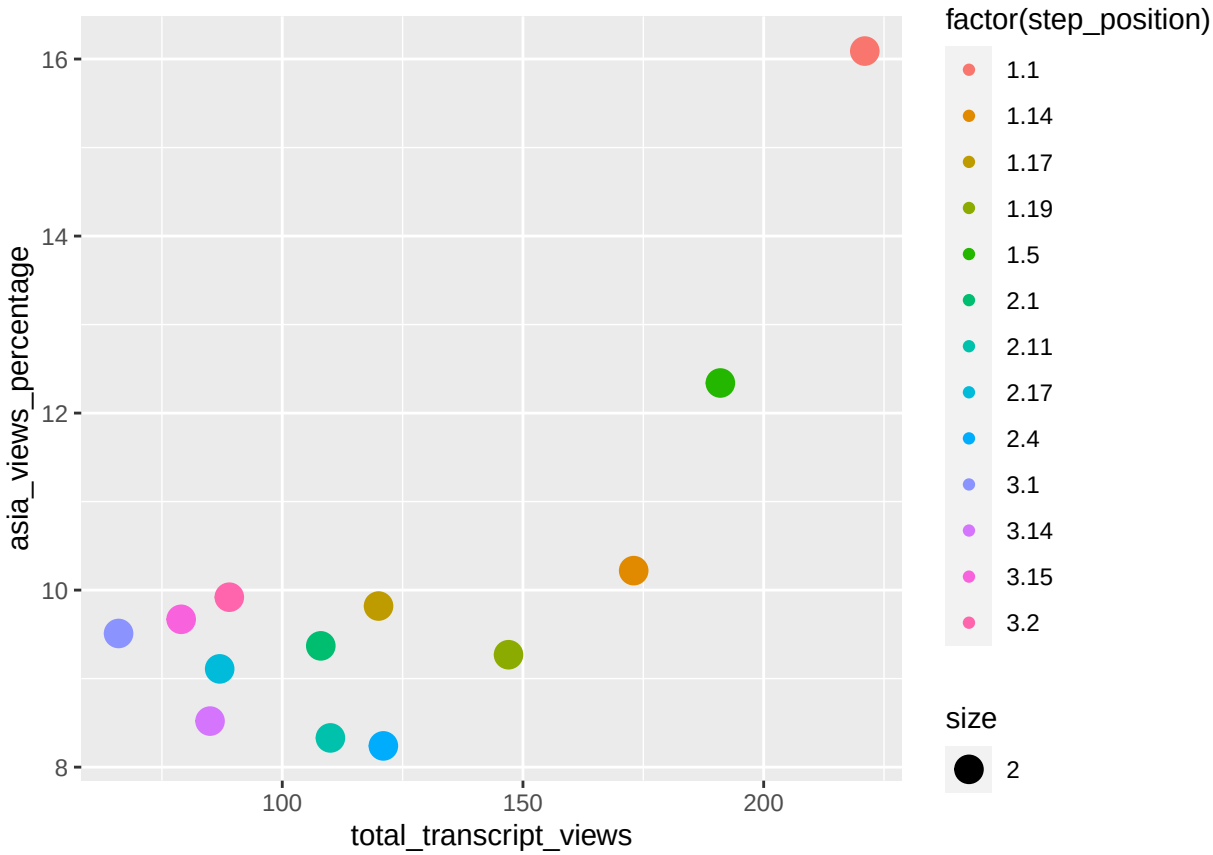
Oceania view percentage summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.240   3.170   3.240   3.265   3.550   4.070
```

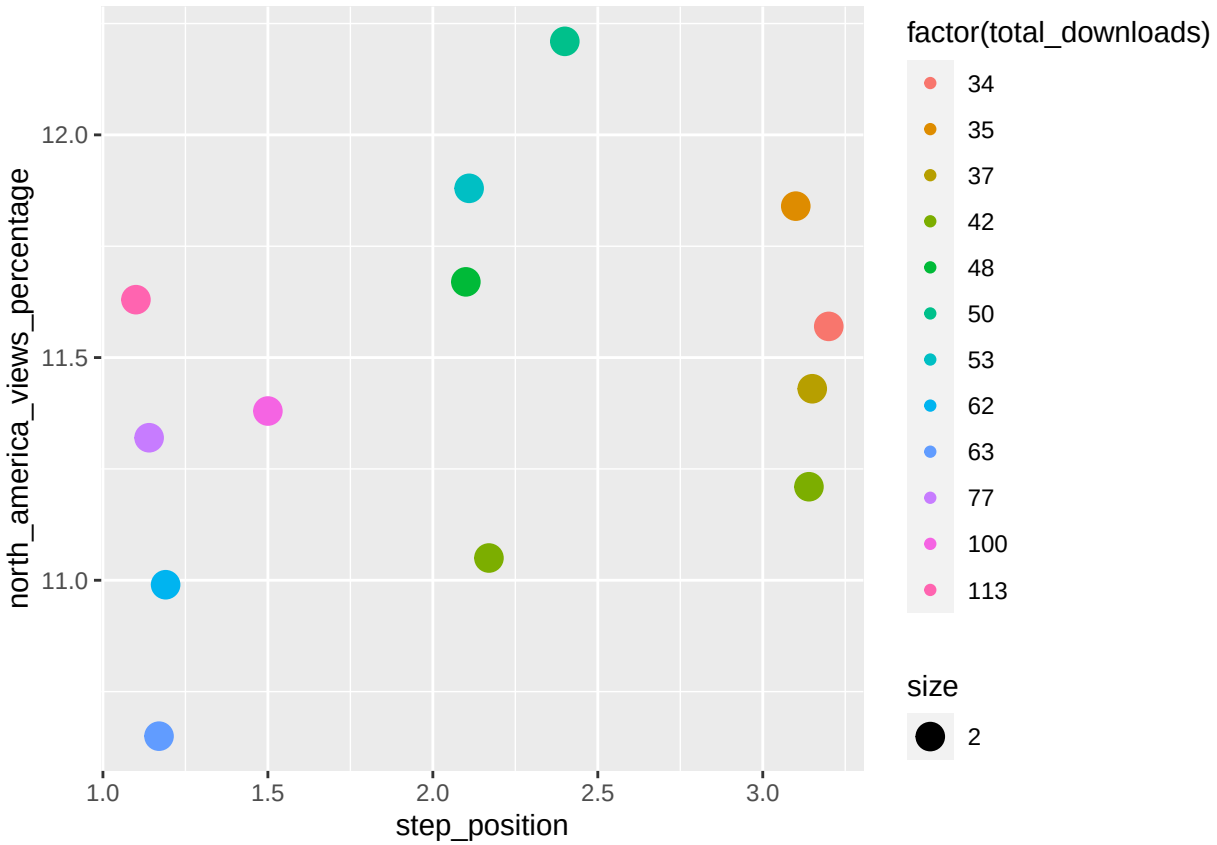
The boxplot presents the statistical five number summary information visually and is a better alternative to quickly identify the mean value and dispersion in data, through above summary we can observe precise numerical values of each of the statistical perimeters used in box plot for the respective countries. We can observe that europe has the highest view percentage out of all the options in our data set. Indicating the course strength in the eyes of European population.

Scatter plot

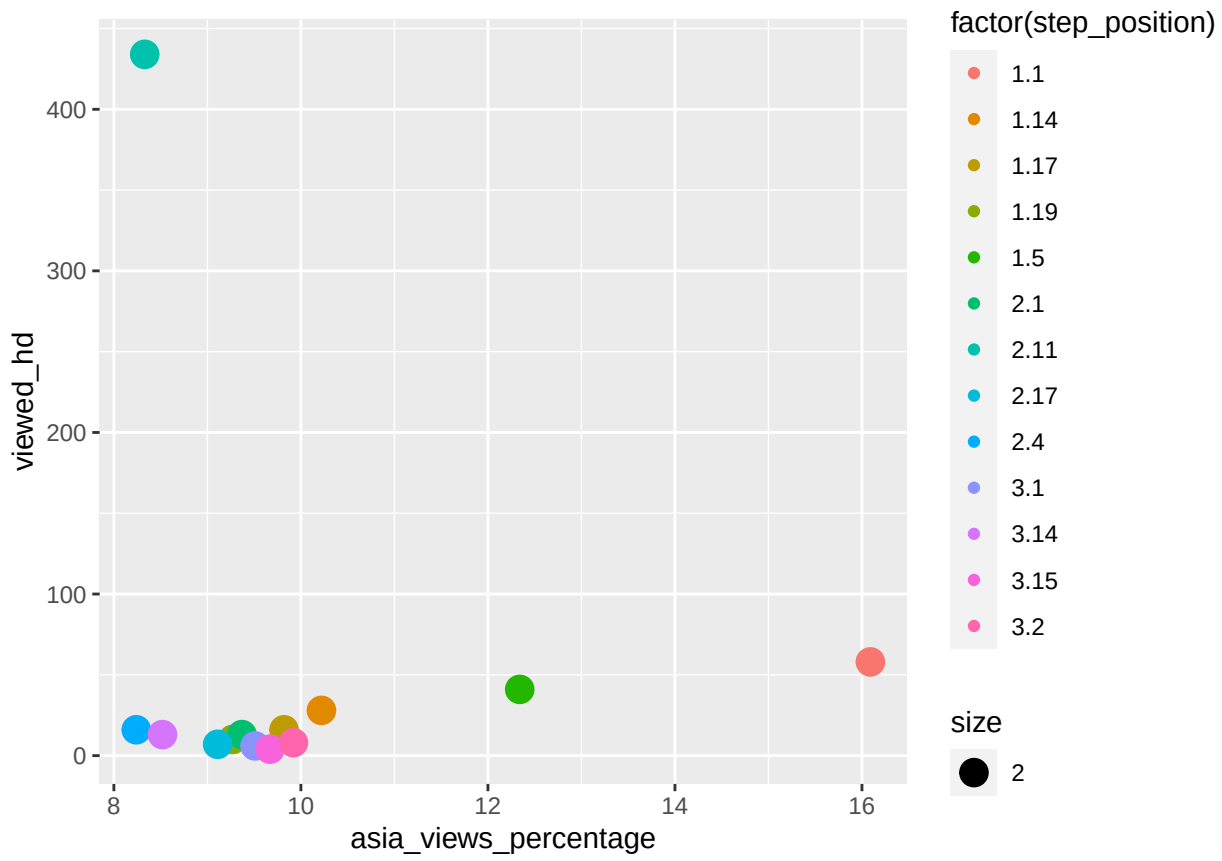
The plot is a transcript views vs asian view percentage graph with respect to step position fill.



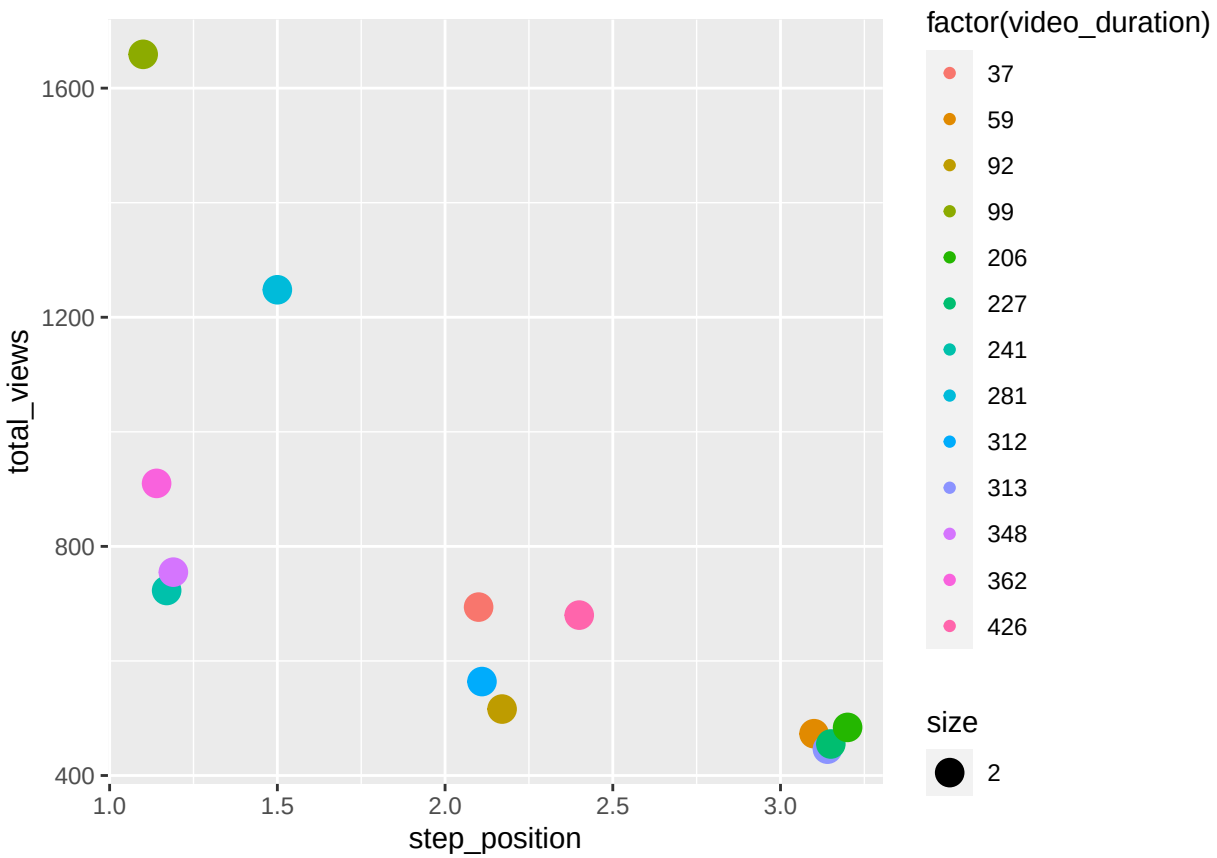
Non native-english speakers has higher usage of transcripts, as evident from graph, with increase in views, there is increase in the usage of transcript



The above graph is a scatter of plot for north America view % to step position with respect to total downloads. The graph can be used to understand at what position, the learner considered downloading. Here, we could see the most downloads are at the start therefore representing that learners often downloaded the videos to view later as per their convenience.



From the graph we can observe that number of people watching in high definition are very low in numbers.



From the above plot we can infer that number of views are declined as the viewers move forward in the course, with 1st video watched as the most number of times and the last videos from chapter 3 watched least number of times. This indicates a decline in interest experienced by the students.

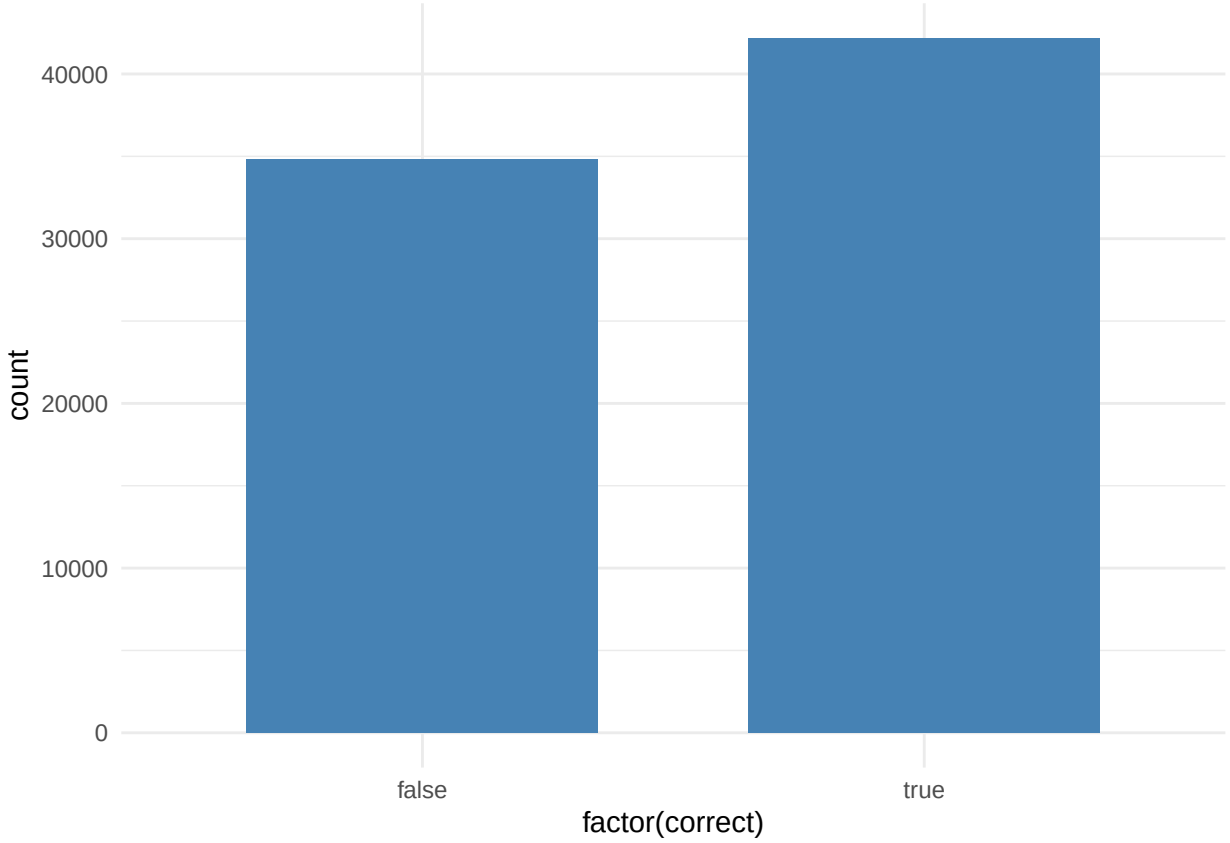
```
## [1] 1659
```

The maximum number of people watched

```
## [1] 446
```

The minimum number of people watched

Dataset 3



Visualizing the overall scores by a bar graph.

```
##
## false  true
## 34817 42185
```

Obtaining the numbers for each category

```
## [1] 42185
```

Count of learners who scored “Correct” is 5739

```
## [1] 34817
```

Count of learners who scored “Incorrect” is 4338

```
## [1] 77002
```

```
## [1] 54.78429
```

The percentage of people who scored “correct”, I have rounded the decimals.

```
## [1] 45.21571
```

The percentage of people who scored “incorrect”

Conclusion Enrollments Data

From the patterns emerging from the plots, we can easily conclude that there is a significant necessity to improve the practical nature of the university courses because the people between the ages 26-35 and the people with university degrees have the highest enrollments in these courses. In terms of employment, the need to up-skill oneself is the latest and growing need, thereby paving the way for the applied programs both at University levels and MOOCs. Furthermore for this course based on the data we can more orient the course for the Asian population which holds a significant future growth area.

I look forward to discussing this report in the Presentation.

Notes:

From a large Data set, a single data set is collected from each category i.e 1) Enrollment Data Set 2) Video Stats Data Set 3) Question Response Data Set