CSC 8631 - EDA Report

Syed Mohib Raza || Student ID: 200740241

December 03, 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, Data Analysis, reproduciblity

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 for performing data analysis and the results are presented in this report.

Data Set information

The first data set comprises of enrollment information, had several unknown values and had to be cleaned to generate a clean sample from a population. Below is the overview of the enrollment data.

```
## # A tibble: 6 x 13
##   learner_id  enrolled_at  unenrolled_at role  fully_participa~ purchased_state~
##   <chr>       <chr>        <chr>         <chr> <chr>            <chr>
## 1 4dc22fed-6~ 2016-05-24 ~ "2018-10-30 ~ lear~ ""               ""
## 2 7a44b170-7~ 2016-05-19 ~ "2018-10-16 ~ lear~ "2016-10-06 04:~ ""
## 3 3fc06ecd-3~ 2016-09-05 ~ "2018-10-12 ~ lear~ ""               ""
## 4 51c61184-8~ 2016-05-18 ~ "2018-09-23 ~ lear~ ""               ""
## 5 e7dc43d0-a~ 2016-09-05 ~ ""            lear~ ""               ""
## 6 d8d3f0a0-9~ 2016-09-04 ~ ""            lear~ ""               ""
## # ... with 7 more variables: gender <chr>, country <chr>, age_range <chr>,
## #   highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

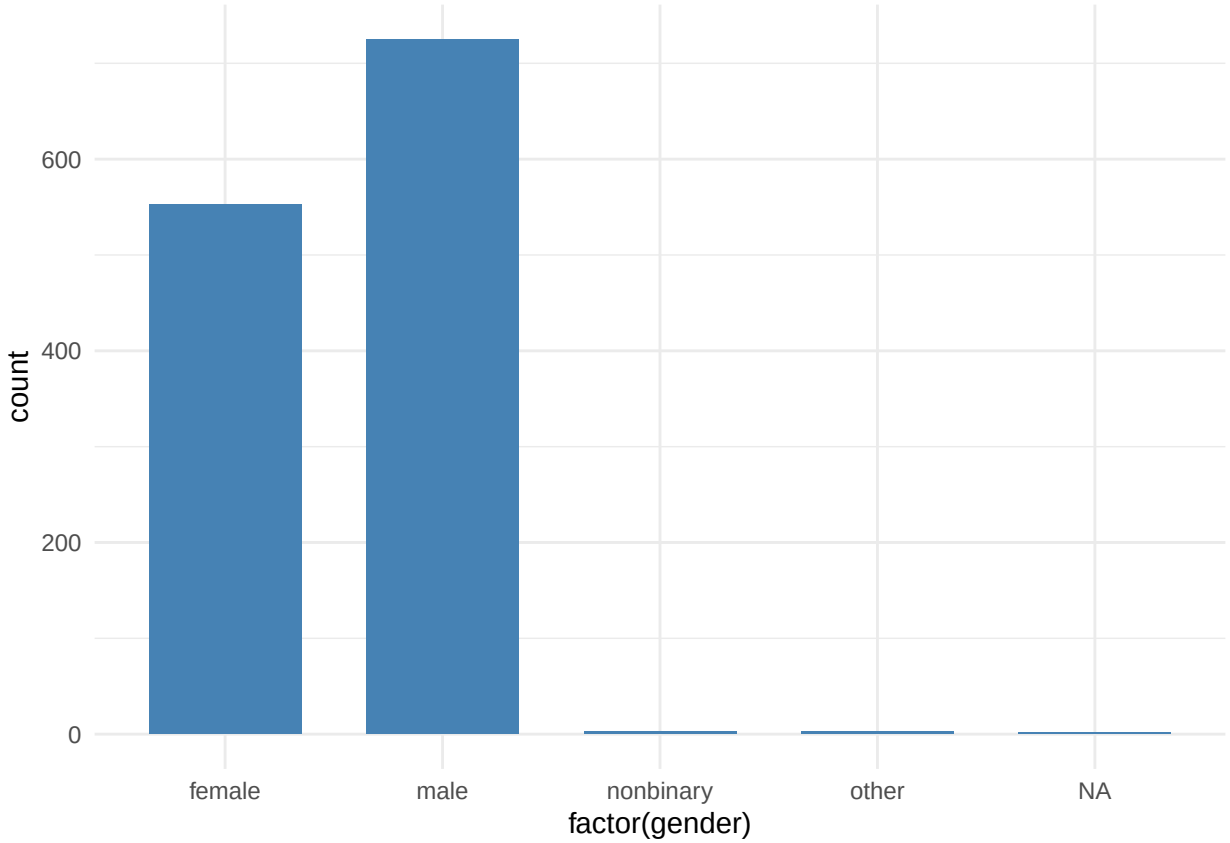
```
##   step_position                                                title
## 1          1.10                                 Welcome to the course
## 2          1.14                          Why would anyone want your data?
## 3          1.17 Preserving privacy in cloud storage: privacy by design
## 4          1.19              Staying safe online: personal perspectives
## 5          1.50                                Privacy online and offline
## 6          2.10                     Welcome to Week 2: payment security
##   video_duration total_views total_downloads total_caption_views
## 1             99        1659             113                  36
## 2            362         910              77                   8
## 3            241         723              63                   5
## 4            348         755              62                   2
## 5            281        1248             100                  15
## 6             37         694              48                   1
##   total_transcript_views viewed_hd viewed_five_percent viewed_ten_percent
## 1                    221        58               76.97              75.35
## 2                    173        28               72.53              70.88
## 3                    120        16               73.72              73.86
## 4                    147        10               72.85              71.92
## 5                    191        41               78.45              75.64
## 6                    108        13               76.37              75.07
##   viewed_twentyfive_percent viewed_fifty_percent viewed_seventyfive_percent
## 1                     73.42                70.40                      68.17
## 2                     68.57                65.38                      63.08
## 3                     71.92                69.71                      66.11
## 4                     69.27                64.90                      63.44
## 5                     69.87                65.63                      62.66
## 6                     74.93                73.49                      72.91
##   viewed_ninetyfive_percent viewed_onehundred_percent console_device_percentage
## 1                     66.43                     63.71                      0.06
## 2                     61.54                     56.81                      0.11
## 3                     61.83                     44.67                      0.14
## 4                     61.59                     49.40                      0.13
## 5                     59.05                     44.87                      0.00
## 6                     71.18                     69.45                      0.14
##   desktop_device_percentage mobile_device_percentage tv_device_percentage
## 1                     78.60                   13.26                 0.06
## 2                     79.23                   10.33                 0.00
## 3                     79.67                    8.71                 0.00
## 4                     78.54                    9.40                 0.00
## 5                     80.37                   11.38                 0.00
## 6                     79.11                    9.37                 0.00
##   tablet_device_percentage unknown_device_percentage europe_views_percentage
## 1                     7.72                        0                    55.15
## 2                    10.11                        0                    65.38
## 3                    11.07                        0                    66.25
## 4                    11.39                        0                    67.15
## 5                     7.93                        0                    61.62
## 6                    10.95                        0                    64.27
##   oceania_views_percentage asia_views_percentage north_america_views_percentage
## 1                     2.29                16.09                          11.63
## 2                     2.86                10.22                          11.32
## 3                     3.18                 9.82                          10.65
## 4                     3.18                 9.27                          10.99
```

```
## 5                    2.24                12.34                          11.38
## 6                    3.17                 9.37                          11.67
##   south_america_views_percentage africa_views_percentage
## 1                           3.07                  10.31
## 2                           2.53                   6.26
## 3                           2.21                   6.36
## 4                           2.12                   5.56
## 5                           2.72                   8.17
## 6                           3.75                   6.20
##   antarctica_views_percentage
## 1                           0
## 2                           0
## 3                           0
## 4                           0
## 5                           0
## 6                           0
```

```
## # 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>
```

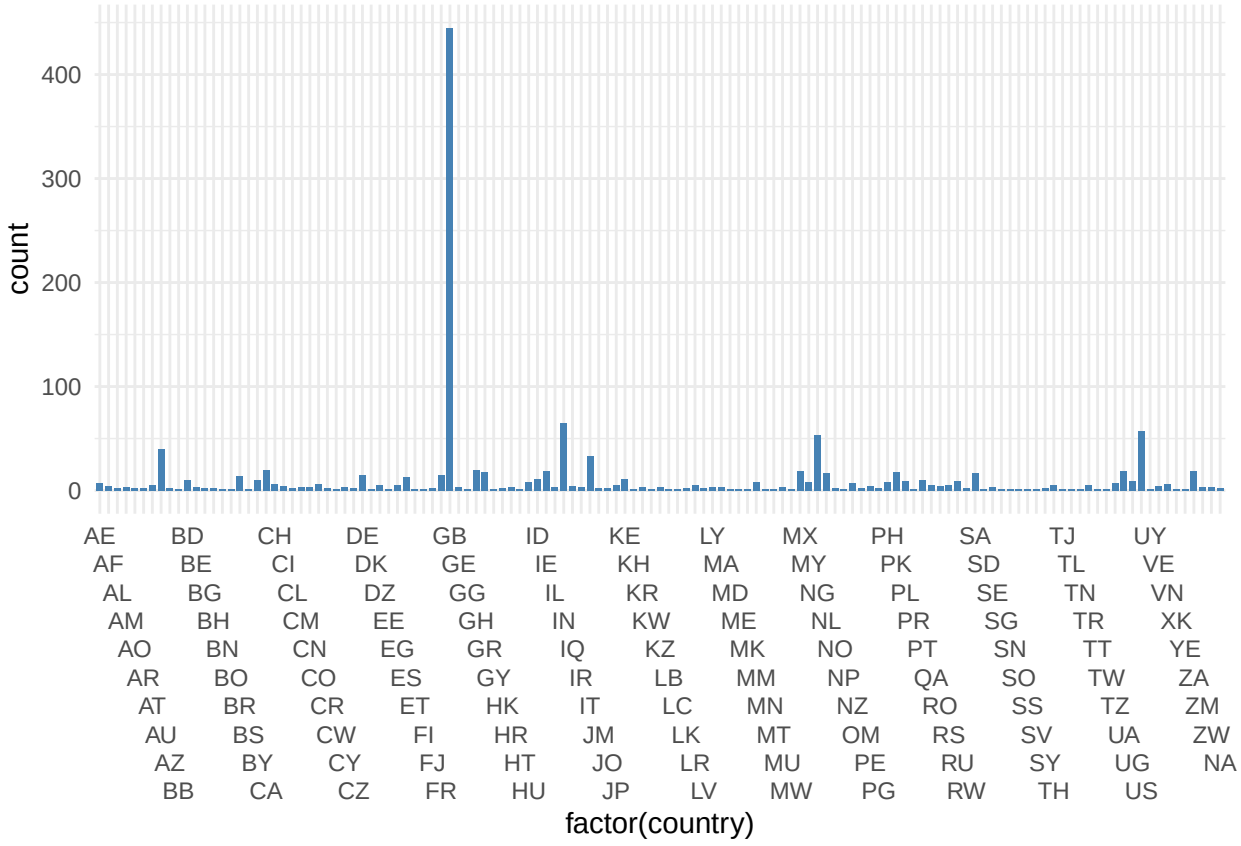
Data Set 1

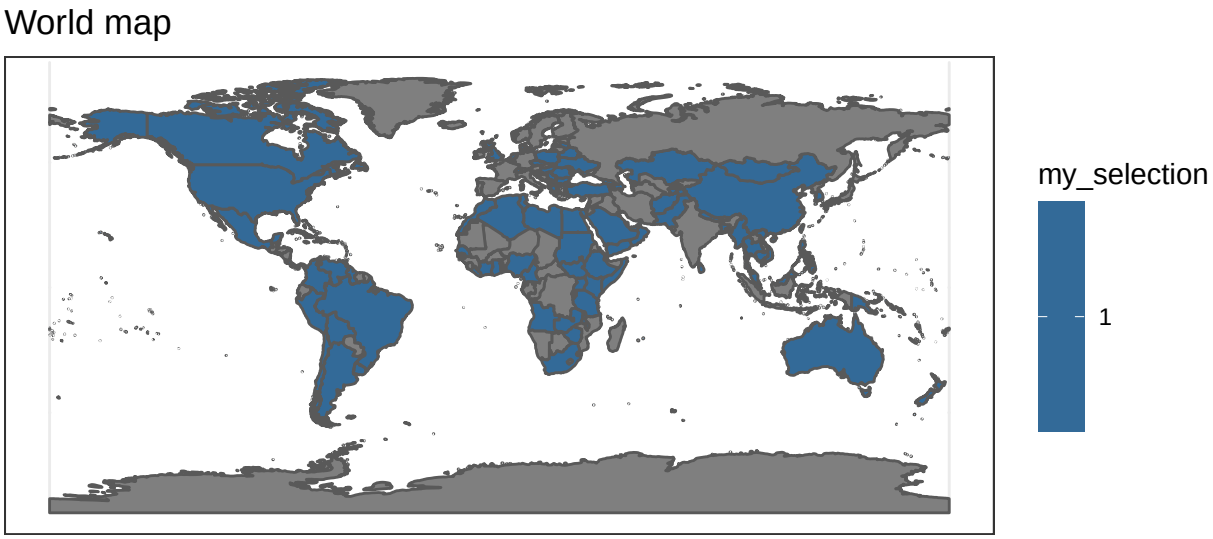
Plot - Gender In the below plot we are identifying the enrollments based on the gender.



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).

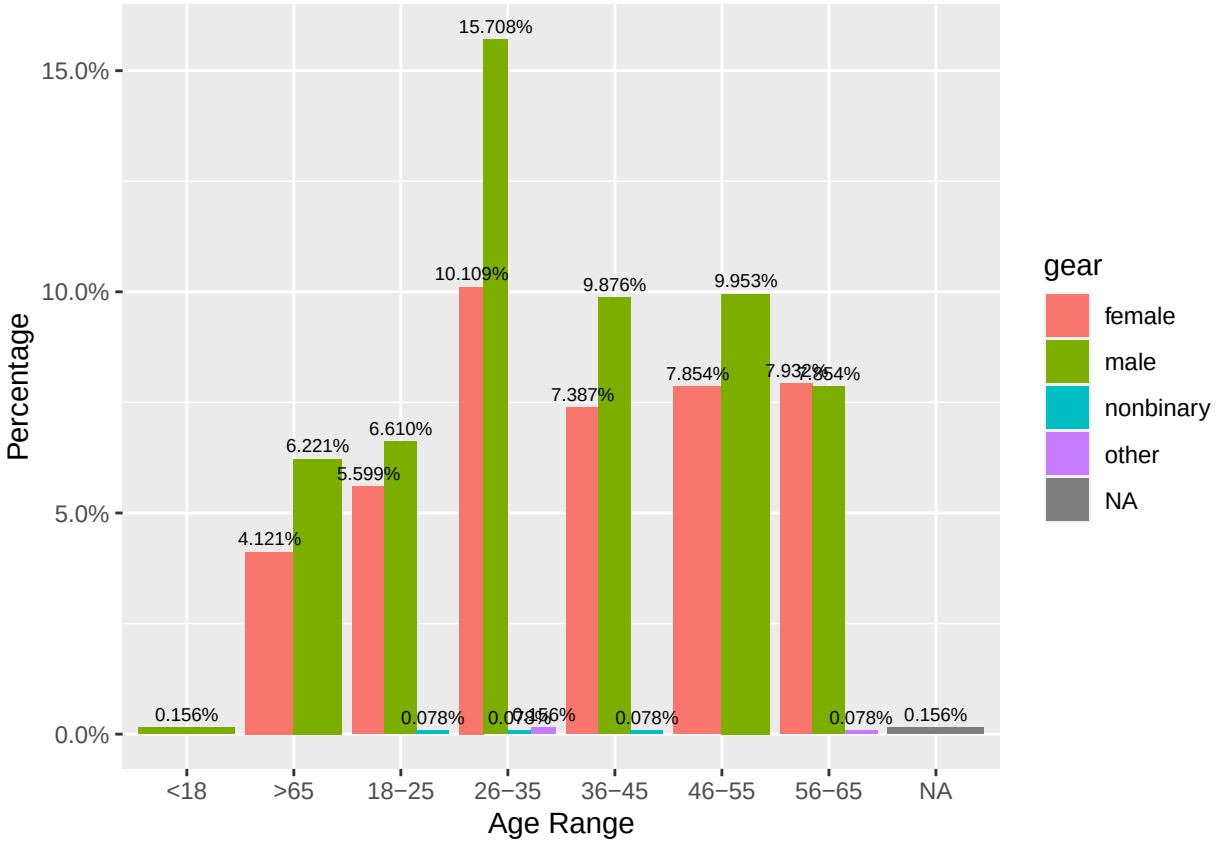

count
400
300
200
100
0
AE BD CH DE GB ID KE LY MX PH SA TJ UY
AF BE CI DK GE IE KH MA MY PK SD TL VE
AL BG CL DZ GG IL KR MD NG PL SE TN VN
AM BH CM EE GH IN KW ME NL PR SG TR XK
AO BN CN EG GR IQ KZ MK NO PT SN TT YE
AR BO CO ES GY IR LB MM NP QA SO TW ZA
AT BR CR ET HK IT LC MN NZ RO SS TZ ZM
AU BS CW FI HR JM LK MT OM RS SV UA ZW
AZ BY CY FJ HT JO LR MU PE RU SY UG NA
BB CA CZ FR HU JP LV MW PG RW TH US
factor(country)



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.

```
enrol_age
```

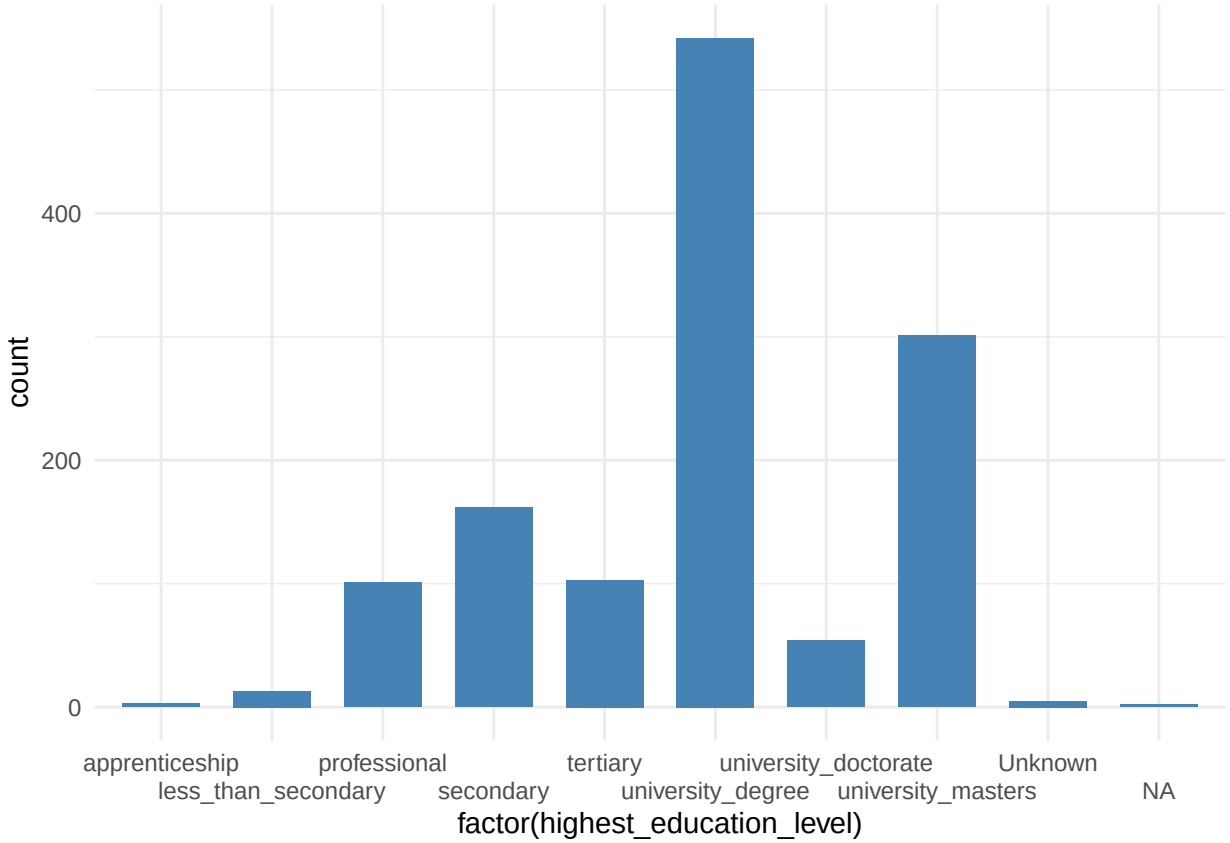


Plot shows us that the highest number of enrollments have been from youth which are from the age group of 26 - 35. This age group is mostly out of college & typically requires a lot of skill development for filling up proper job skills. Males (15.7%) are the highest enrollments for this age group which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

```
## [1] 1.553863
```

) ## Employment Plot The employment data is useful in understanding the job status and curate the curriculum in a way to upskill the learners.

```
employment_plot
```



From above, we can infer that, the full time working have been attracted to online education at most. It is legit in a way that, they can obtain new desire professional skills while being employed full time, as it helps to fund their education.