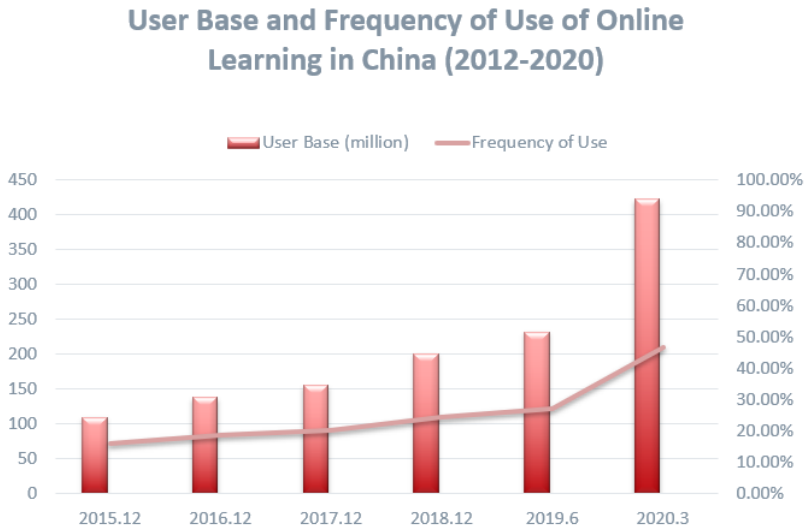
Transportation Systems Analysis

A Study of the HarvardX-MITx Person-Course Dataset with Discrete Choice Models

2017010914 Wenxin Zhang

Background and Significance

Since the beginning of 2020, COVID-19 has developed into a pandemic, influencing millions of people around the world. Surely the pandemic has interrupted people's life, especially for students who rely heavily on face-to-face learning. However, as the pandemic has accelerated the digital transformation, online education has been made more easily assessed, offering another opportunity. Based on the data from China Internet Network Information Center, by June 2019, there are 854 million netizens in China. Among them, 232 million use online learning. During the pandemic, 265 million students moved to online, boosting both user base and frequency of use. By March 2020, the user base has reached 423 million rapidly and the use frequency increases from 27.2% to 46.8%. Meanwhile, since many colleges like Tsinghua offer free online courses to the public, online learning became even more popular.



Source: China Internet Network Information Center (CNNIC)

One popular online learning form is MOOC. MOOC is the abbreviation of massive open online course, which is an online course aimed at unlimited participation and open access via the web (Kaplan et al., 2016). It was first introduced in 2008 and emerged as a popular mode of learning in 2012 (Siemens, 2013). MIT and Harvard, as university pioneers, joined in the initiative. In the year from the fall of 2012 to the summer of 2013, the first 17 HarvardX and MITx courses launched on the edX platform (Ho et al., 2014). During the past decade, millions of students from around the globe have enrolled; thousands of courses have been offered; and hundreds of universities have joined hands to revolutionize education. Therefore, it is vital to understand people's choice behaviors concerning online education so that we can better predict demands for online learning and offer attractive courses.

Prior researches have been focusing on MOOCs as a phenomenon and perform quantitative analysis. One possible reason could be a lack of data since MOOCs platforms like edX may rely on data for profit. Privacy concerns could be another reason. One qualitative research was conducted by Christensen et al. in 2014. They used the survey data based on students enrolled in the University of Pennsylvania's MOOCs to analyze who takes MOOCs and why they take MOOCs. Based on stated preferences data, they found that students' main reasons for taking a MOOC are advancing in their current job and satisfying curiosity (Christensen et al., 2014). In our research, we use the HarvardX-MITx Person-Course Dataset, which involves revealed preferences data collected on the edX platform to further analyze people's choice behavior in terms of MOOCs. Specifically, we want to understand how people choose from multiple courses and how they choose to get certificates. As far as we know, we are the first to use discrete choice models to analyze the problem.

HarvardX

Data

We use (1) HarvardX-MITx Person-Course Dataset, (2) Course Info Data, and (3) Human Development Index Data.

- https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147 (HarvardX, n.d.)
- https://www.kaggle.com/edx/course-study
- http://hdr.undp.org/en/home

(1) HarvardX-MITx Person-Course Dataset

The HarvardX-MITx Person-Course Dataset (online open version) contains re-identified data from the first year of five HavardX courses on the edX platform (Ho et al., 2014). There are 338,223 registration activities by 301,609 users. Note that users can register for multiple courses as long as the registration open. Each record represents one individual's registration activity in one course. After dropping missing data, we 290,948 records from 261,151 users. Further data preparation details are shown later.

(2) Course Info Data

Course Info Data was published along with the report HarvardX and MITx: Four Years of Open Online Courses (Chuang, 2017). This "Year 4 Report" involves basic course information and user activity information of HarvardX and MITx courses on edX over four complete years. We then find the course info of the five courses we focus on. The detailed course info is as follows.

Course ID	Registration Open	Launch Date	Course Title	Subject	# participants	Participants median age	% Male	% Female	% Bachelor's Degree or Higher
CB22x	12/19/2012	03/13/2013	The Ancient Greek Hero	Humanities, History, Design, Religion, and Education	25,873	32	53.31	46.69	71.95
CS50x	07/24/2012	10/15/2012	Introduction to Computer Science	Computer Science	129,400	28	80.02	19.98	58.78
ER22x	12/19/2012	03/02/2013	Justice	Humanities, History, Design, Religion, and Education	58,779	30	60.42	39.58	69.78
PH207x	07/24/2012	10/15/2012	Health in Numbers: Quantitative Methods in Clinical and Public Health Research	Government, Health, and Social Science	52,521	32	56.78	43.22	88.33
PH278x	12/19/2012	05/15/2013	Human Health and Global Environmental Change	Government, Health, and Social Science	23,179	30	51.15	48.85	75.19

We didn't include course info like duration, time commitment, and difficulty for three reasons. First, they were not involved in the original datasets so we need to find them elsewhere. Second, only CS50x and PH278x are still available online, which means we cannot find information related to the other three courses. Third, even if we find the courses online, it is very likely the courses have been updated since 2012 so that the information is not eligible. However, as we focus on people's choice behavior to further discuss future demands, not the specific design of a given course, it is fair to say the data satisfy the need of our project.

(3) Human Development Index Data

As part of the UNITED NATIONS DEVELOPMENT PROGRAMME, Human Development Report Office (HDRO) aims to advance human development. Human Development Index (HDI) is a composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge, and a decent standard of living. See Technical note 1 at http://hdr.undp.org/sites/default/files/hdr2019_technical_notes.pdf for details on how the HDI is calculated.

The three datasets are merged to estimate the parameters of discrete choice models. Note that in the HarvardX-MITx Person-Course Dataset, during de-identification, some country names were replaced with the corresponding continent/region name, like "Other South Asia". For those, we use the average HDI for the corresponding region.

Choose from Multiple Courses

Choice Set

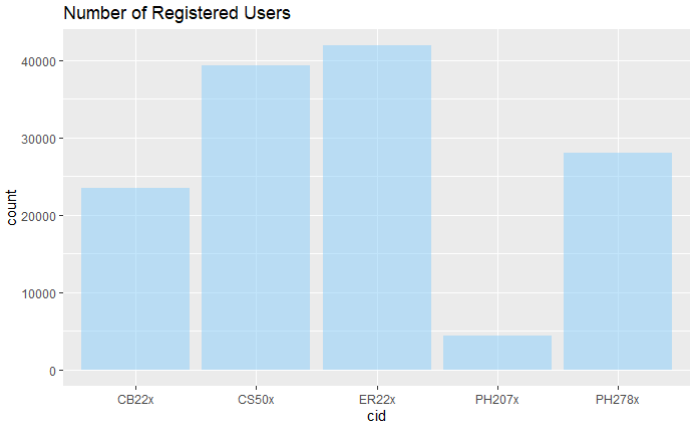
Choice set needs to exhibit three characteristics. First, the alternatives must be mutually exclusive from the decision maker's perspective. Second, the choice set must be exhaustive, in that all possible alternatives are included. Third, the number of alternatives must be finite (Train, 2003). Since users can choose several courses, to obtain mutually exclusive alternatives, one approach is to list every possible combination of courses as an alternative. However, as there are five courses in total, the large number of possible combinations makes it difficult to estimate. Meanwhile, we also need data that distinguish the alternatives, for example, the benefit of taking

both CS50x and PH207x versus the benefit of taking CS50x alone. Another approach is to define the "primary" course. In other words, we need to develop a rule for determining which course is primary when the user chooses multiple courses.

In this specific problem, we believe the second approach is more tractable, though not perfect. As the platform launched in 2012 and the data was collected shortly after that, people may not have enough time to finish multiple courses during such a short time period. Therefore, as we know when the user registered for a given course, we define the first course the user chooses to be the primary one. In order to make the choice set exhaustive, we need to make sure all five courses are available to the user when he makes the choice. Based on the registration open date of the courses, we keep the records after 12/19/2012. Note that the registration opened before the courses were successfully launched. Then we keep the first course the user registers for, yielding 137,004 records, each corresponding to one user.

Further data cleaning involves drop outliers, for example, a user with a Bachelor's degree but reports his age to be 1. Then we have 136,878 records.

Then the choice set is **{CB22x, CS50x, ER22x, PH207x,PH278x}**. Therefore, we will use Multinomial Choice Models and Nested Logit Model.



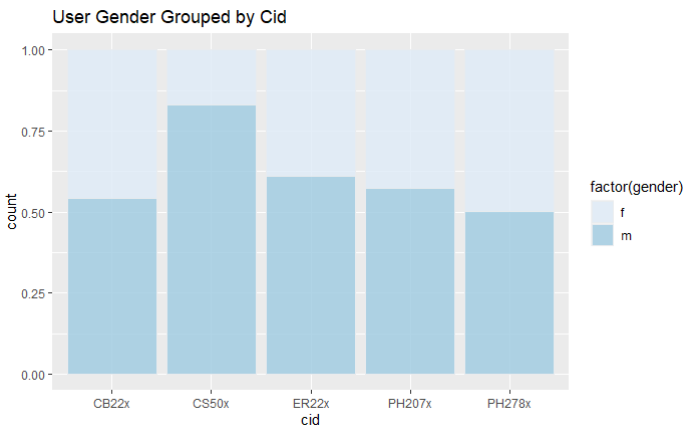
Based on the figure, we can find that the number of registered users is unbalanced across the five courses. ER22x is the most popular course while PH207x is the least popular.

See Appendix for data summary and statistics description.

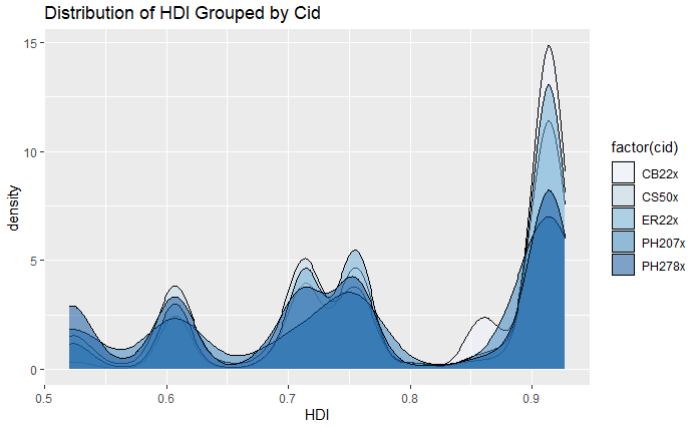
Multinomial Choice Models

User Info

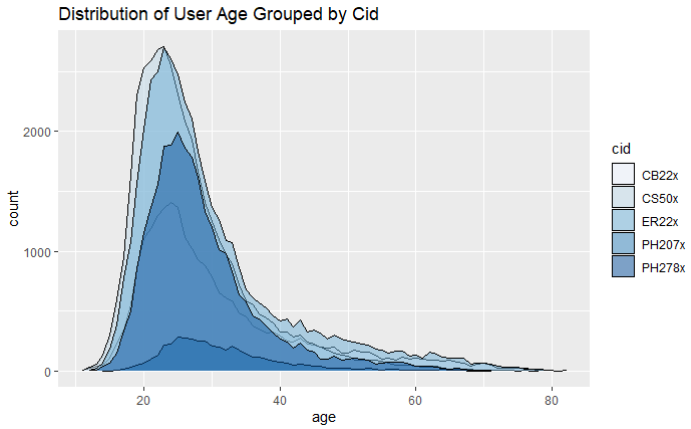
In this part, we will explore how socioeconomic variables influence user's choice behavior. First, let us visualize the users' course choice across gender, human development level, age, and education background.



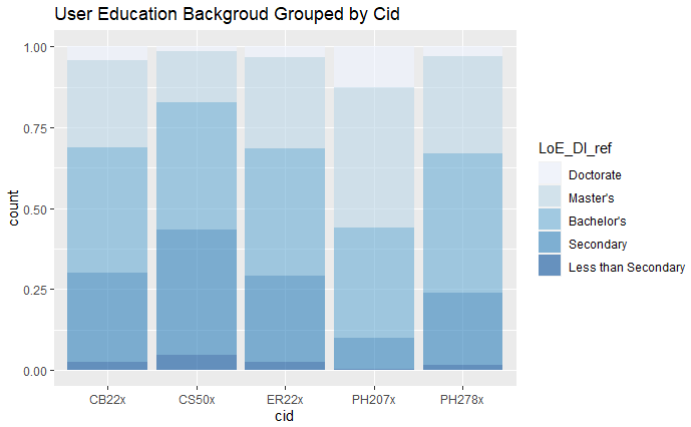
Based on the above figure, we can see that compared with other courses, the proportion of female students in CS50x is the lowest and the proportion in PH278x is the highest. Overall, we can see that gender is relatively balanced in CB22x, ER22x, PH207x, and PH278x.



Based on the above figure, we can see that the average HDI of students in CB22x is the highest and the other four are similar. Meanwhile, PH207x and PH278x cover people from less developed countries.



Based on the above figure, we can see that compared with other courses, the age distribution of PH278x is relatively balanced. Overall, we can see that the age distributions are similar and peak at approximately 25.



Based on the above figure, we can see that the average education level of students in PH207x is the highest while the one in CS50x is the lowest. One possible explanation could be prerequisite demands are different, and CS50x is welcome to the public. The majority of MOOC users have a Bachelor's degree.

The above figures indicate that user's socioeconomic status does influence his course choice. Therefore, we want to how those information influences their course choice.

Model 1

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
V_{CB22x}	1				gender				HDI			
V_{CS50x}												
V_{ER22x}		1				gender				HDI		
V_{PH207x}			1				gender				HDI	
V_{PH278x}				1				gender				HDI

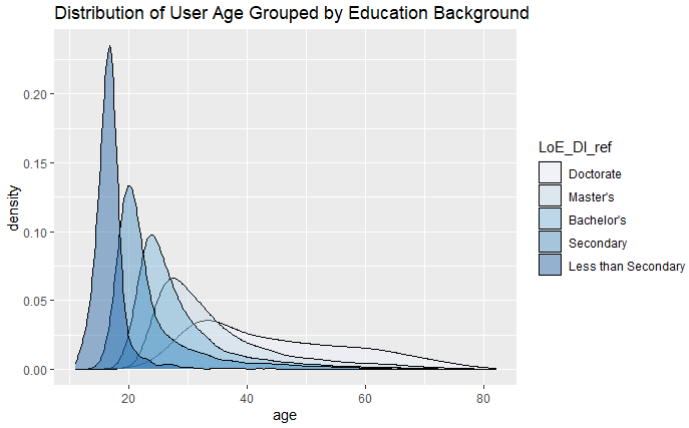
Results

```
Call:
mlogit(formula = cid ~ 1 | gender + HDI, data = H, reflevel = "CS50x",
    method = "nr")

Frequencies of alternatives:choice
   CS50x    CB22x    ER22x   PH207x   PH278x
0.286949 0.171087 0.305871 0.031722 0.204372

nr method
6 iterations, 0h:0m:28s
g'(-H)^-1g = 0.00252
successive function values within tolerance limits

Coefficients :
                    Estimate Std. Error  z-value  Pr(>|z|)
(Intercept):CB22x  -1.398071   0.059990 -23.3051 < 2.2e-16 ***
(Intercept):ER22x   0.312776   0.047876   6.5331 6.442e-11 ***
(Intercept):PH207x -0.517782   0.102484  -5.0523 4.365e-07 ***
(Intercept):PH278x  2.487185   0.051647  48.1577 < 2.2e-16 ***
genderm:CB22x      -1.345066   0.018918 -71.1002 < 2.2e-16 ***
genderm:ER22x      -1.116341   0.016862 -66.2044 < 2.2e-16 ***
genderm:PH207x     -1.330150   0.033849 -39.2965 < 2.2e-16 ***
genderm:PH278x     -1.662774   0.018292 -90.9018 < 2.2e-16 ***
HDI:CB22x           2.271642   0.070133  32.3907 < 2.2e-16 ***
HDI:ER22x           0.716103   0.056114  12.7616 < 2.2e-16 ***
HDI:PH207x         -0.949713   0.124240  -7.6442 2.109e-14 ***
HDI:PH278x         -2.210737   0.062198 -35.5437 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -192050
McFadden R^2:  0.036776
Likelihood ratio test : chisq = 14664 (p.value = < 2.22e-16)
```

We can see that compared to CS50x, the intercepts of CB22x and PH207x are negative and statistically significant while the intercepts of ER22x and PH278x are positive and statistically significant. For user gender, compared to CS50x, the coefficients of other courses are negative and statistically significant. This is in accordance with the fact that CS50x has the highest male proportion. Based on the coefficients of alternative specific variable HDI, we can see that compared to CS50x, users from more developed countries will be more likely to choose CB22x and ER22x which talk about Humanities, History, Design, Religion, and Education. Meanwhile, people from less developed countries will be more likely to choose Government, Health, and Social Science related courses.

Model 2

Similar to Model 1, we further add other socioeconomic variables in Model 2. Intuitively, education background is positively correlated with age since it takes time to get a high degree. Based on the figure, we can see that the peaks of the distribution move to the right as the degree level gets higher. Therefore, we only add education background in Model 2.



For simplicity, we do not show the model specification table here.

Results

```
Call:
mlogit(formula = cid ~ 1 | gender + HDI + LoE_DI, data = H, reflevel = "CS50x",
    method = "nr")

Frequencies of alternatives:choice
   CS50x    CB22x    ER22x   PH207x   PH278x
0.286949 0.171087 0.305871 0.031722 0.204372

nr method
8 iterations, 0h:0m:47s
g'(-H)^-1g = 1.25E-05
successive function values within tolerance limits

Coefficients :
                          Estimate Std. Error  z-value  Pr(>|z|)
(Intercept):CB22x       -1.900950   0.079661 -23.8630 < 2.2e-16 ***
(Intercept):ER22x       -0.176940   0.064634  -2.7376  0.006189 **
(Intercept):PH207x      -2.434975   0.299668  -8.1256 4.441e-16 ***
(Intercept):PH278x       1.585368   0.079842  19.8563 < 2.2e-16 ***
genderm:CB22x           -1.338466   0.019037 -70.3091 < 2.2e-16 ***
genderm:ER22x           -1.106060   0.017003 -65.0521 < 2.2e-16 ***
genderm:PH207x          -1.348416   0.034501 -39.0839 < 2.2e-16 ***
genderm:PH278x          -1.648303   0.018509 -89.0564 < 2.2e-16 ***
HDI:CB22x                2.124246   0.071210  29.8306 < 2.2e-16 ***
HDI:ER22x                0.582192   0.057280  10.1639 < 2.2e-16 ***
HDI:PH207x              -1.810112   0.128903 -14.0424 < 2.2e-16 ***
HDI:PH278x              -2.375208   0.063834 -37.2094 < 2.2e-16 ***
LoE_DIBachelor's:CB22x   0.602494   0.050315  11.9745 < 2.2e-16 ***
LoE_DIBachelor's:ER22x   0.584446   0.041394  14.1190 < 2.2e-16 ***
LoE_DIBachelor's:PH207x  2.411142   0.279991   8.6115 < 2.2e-16 ***
LoE_DIBachelor's:PH278x  1.057970   0.059058  17.9142 < 2.2e-16 ***
LoE_DIDoctorate:CB22x    1.541172   0.071860  21.4468 < 2.2e-16 ***
LoE_DIDoctorate:ER22x    1.361383   0.063851  21.3214 < 2.2e-16 ***
LoE_DIDoctorate:PH207x   4.939188   0.285025  17.3290 < 2.2e-16 ***
LoE_DIDoctorate:PH278x   1.973869   0.079947  24.6896 < 2.2e-16 ***
LoE_DIMaster's:CB22x     1.115758   0.051689  21.5859 < 2.2e-16 ***
LoE_DIMaster's:ER22x     1.145338   0.042769  26.7793 < 2.2e-16 ***
LoE_DIMaster's:PH207x    3.599094   0.279840  12.8612 < 2.2e-16 ***
LoE_DIMaster's:PH278x    1.677747   0.060119  27.9072 < 2.2e-16 ***
LoE_DISecondary:CB22x    0.306880   0.050772   6.0443 1.500e-09 ***
LoE_DISecondary:ER22x    0.230050   0.041752   5.5099 3.590e-08 ***
LoE_DISecondary:PH207x   1.194769   0.283082   4.2206 2.437e-05 ***
LoE_DISecondary:PH278x   0.488298   0.059730   8.1752 2.220e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -188790
McFadden R^2:  0.053097
Likelihood ratio test : chisq = 21173 (p.value = < 2.22e-16)
```

We can see the results are similar to Model 1 in terms of HDI and gender. As for user education level, we can see that all coefficients are positive, which means that compared to CS50x, people of high education level will be more likely to register liberal arts courses. This is also in accordance with the fact that students in CS50s have the lowest average education level.

Comparison of Model 1 and Model 2

Model	Log-Likelihood	McFadden R^2	Likelihood Ratio Test
Model 1	-192,050	0.036776	p.value = < 2.22e-16
Model 2	-188,790	0.053097	p.value = < 2.22e-16

Based on the table, we can see that Model 2 improved. Therefore, we want to conduct a likelihood ratio test to see if the improvement is significant.

```
Likelihood ratio test

Model 1: cid ~ 1 | gender + HDI
Model 2: cid ~ 1 | gender + HDI + LoE_DI
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  12 -192046
2  28 -188792 16 6508.1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the test result, we should reject the null hypothesis at a 95% level of confidence. Therefore, compared to Model 1, Model 2 has a significant improvement in fit.

Course Info

Now let us look at how course-related variables influence people's choice.

Note that course-related variables are highly correlated (see Appendix), so we only include *np* (number of participants) as the indicator of the popularity of the course.

Model 3

	β_1	β_2	β_3	β_4
V_{CB22x}	np_{CB22x}	lyr_{CB22x}	1	
V_{CS50x}	np_{CS50x}	lyr_{CS50x}		
V_{ER22x}	np_{ER22x}	lyr_{ER22x}	1	
V_{PH207x}	np_{PH207x}	lyr_{PH207x}		1
V_{PH278x}	np_{PH278x}	lyr_{PH278x}		1

Note: β_3 and β_4 are for categorical variable course subject Humanities, History, Design, Religion, and Education and Government, Health, and Social Science respectively. We expect β_1 to be positive because the more popular the course is, the more likely the user will choose the course.

Results

```
Call:
mlogit(formula = cid ~ np + c_sub + l_yr | 0, data = H, method = "nr")

Frequencies of alternatives:choice
   CB22x    CS50x    ER22x   PH207x   PH278x
0.171087 0.286949 0.305871 0.031722 0.204372

nr method
6 iterations, 0h:0m:24s
g'(-H)^-1g = 0.00189
successive function values within tolerance limits

Coefficients :
                                                      Estimate  Std. Error z-value  Pr(>|z|)
np                                                1.7656e-05  2.4798e-07  71.199 < 2.2e-16 ***
c_subGovernment, Health, and Social Science      -8.4492e-01  2.4884e-02 -33.954 < 2.2e-16 ***
c_subHumanities, History, Design, Religion, and Education -1.0703e+00  2.1911e-02 -48.844 < 2.2e-16 ***
l_yr2013                                          2.3810e+00  1.7861e-02 133.311 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -199380
```

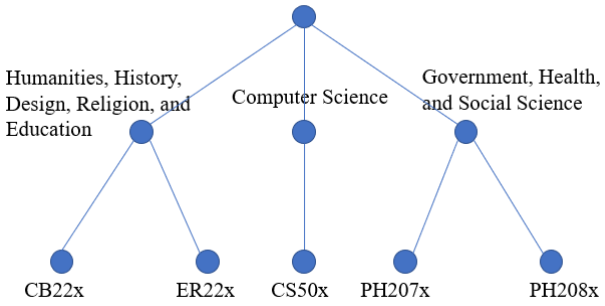
We can find that β_1 is indeed positive and statistically significant. Moreover, β_2 is also positive and statistically significant, indicating that newly launched courses are more attractive, which is also plausible. As for the course subject, we can see that compared to CS, the other two subjects are less attractive. There are two possible interpretations for the popularity of CS. First, as the

information revolution sweeps the globe, it calls for a workforce with CS skills. Second, as liberal arts courses stress interactions, the teaching efficiency of tech-related courses are less influenced when moving from offline to online.

The loglikelihood of the model is -199,380, and we can reject the null hypothesis that all the parameters are zero.

Nested Logit Model

Intuitively, courses of the same subject may have shared observed and unobserved attributes. For example, both PH278x and PH207x talk about Government, Health, and Social Science, and we can find similar modes in the figures and model results in the previous part. Therefore, in this part, we use a Nested Logit model to account for the inter-relationship between the class of course subjects. Specifically, the structure is as follows.



Model 4

	β_1	β_2
V_{CB22x}	np_{CB22x}	lyr_{CB22x}
V_{CS50x}	np_{CS50x}	lyr_{CS50x}
V_{ER22x}	np_{ER22x}	lyr_{ER22x}
V_{PH207x}	np_{PH207x}	lyr_{PH207x}
V_{PH278x}	np_{PH278x}	lyr_{PH278x}

Results

```
Call:
mlogit(formula = cid ~ np + l_yr | 0, data = H, nests = list(humanity = c("CB22x",
    "ER22x"), CS = c("CS50x"), government = c("PH207x",
    "PH278x")), un.nest.el = TRUE)

Frequencies of alternatives:choice
   CB22x    CS50x    ER22x   PH207x   PH278x
0.171087 0.286949 0.305871 0.031722 0.204372

bfgs method
8 iterations, 0h:0m:50s
g'(-H)^-1g = 3.4E-08
gradient close to zero

Coefficients :
           Estimate Std. Error z-value  Pr(>|z|)
np       1.4988e-05 1.7329e-07  86.491 < 2.2e-16 ***
l_yr2013 1.4047e+00 1.3729e-02 102.313 < 2.2e-16 ***
iv       6.1383e-01 6.5836e-03  93.236 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -200050
```

Testing of Nesting Structure

We can use a Quasi t-test of the hypothesis that the log-sum coefficient equal to 1. Since the test statistics is -58.66, we should reject H_0.

Within nest correlation is $1 - 0.61^2 = 0.6279$.

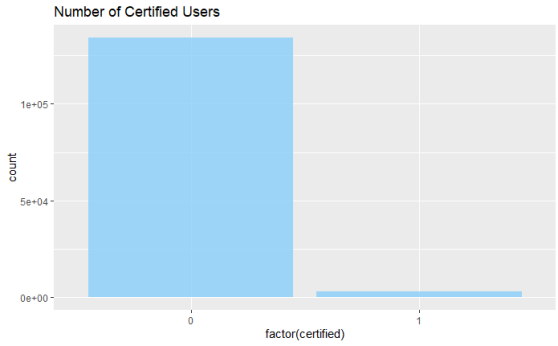
We can also use a likelihood ratio test because the multinomial logit is a special case of the nested model. Since p-value < 2.2e-16, we reject the null hypothesis. The nested structure is reasonable.

```
Likelihood ratio test

Model 1: cid ~ np + l_yr | 0
Model 2: cid ~ np + l_yr | 0
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -201130
2   3 -200049  1 2162.1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Choose to Get Certificates

Choice Set

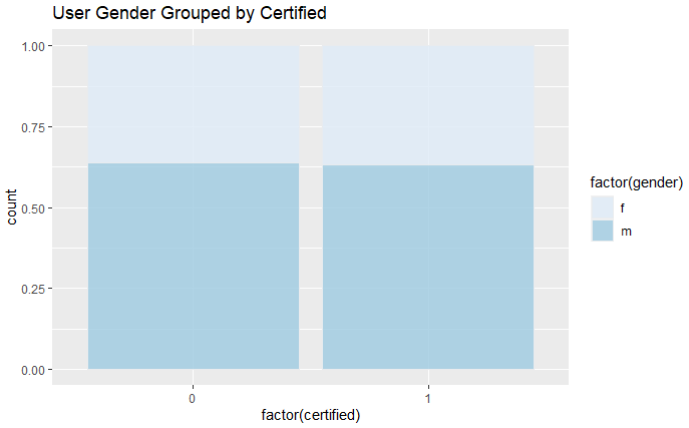
In the previous section, we discussed how users choose from multiple courses. Now we want to further explore people's choice behavior when it comes to certificates. A certificate normally can provide proof that the user has successfully completed an online course, and you may need to pay for such a certificate. Redefine the choice situation as choosing to earn a certificate conditional on having registered for a given course. Choice Set is **{Certified, Not Certified}.** As there are two choices, we will use Binary Logit Choice Model.



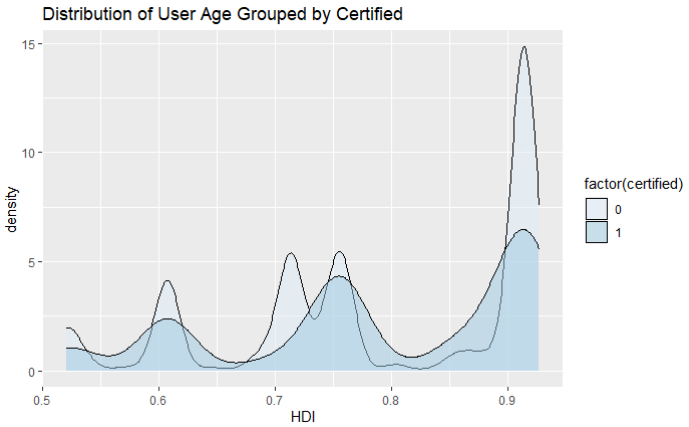
Binary Logit Choice Model

User Info

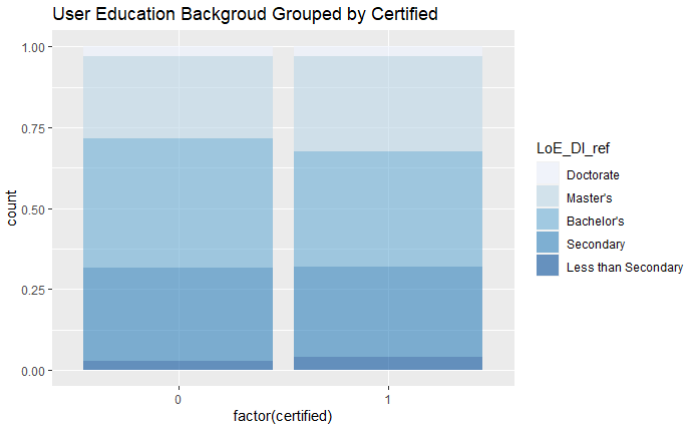
First, let us have a look at the distribution of user gender, HDI, education background, and age grouped by certified and not certified.



Based on the above figure, we can see gender is relatively balanced between certified and not certified.



Based on the above figure, we can see people from well-developed countries will be less likely to choose to get a certificate. The average HDI of people who choose to get a certificate is lower.



Based on the above figure, we can see among users who choose to get a certificate, the majority have a Bachelor's degree.

Model 5

In Model 5, we focus on user info.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
$V_{Certified}$	1	1 if Secondary	1 if Bachelor	1 if Master	1 if Doctorate	1 if Male	HDI
$V_{NotCertified}$							

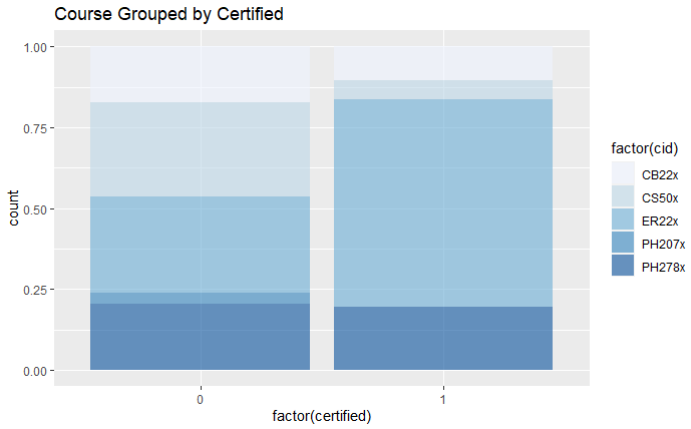
Result

```
Call:
glm(formula = certified ~ LoE_DI + gender + HDI, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2416  -0.2119  -0.1936  -0.1866   2.8545

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.49717    0.16831 -20.778  < 2e-16 ***
LoE_DIBachelor's  -0.49543    0.10394  -4.767 1.87e-06 ***
LoE_DIDoctorate   -0.40726    0.14962  -2.722  0.00649 **
LoE_DIMaster's    -0.22365    0.10483  -2.133  0.03289 *
LoE_DISecondary   -0.40676    0.10518  -3.867  0.00011 ***
genderm           -0.02983    0.04106  -0.726  0.46763
HDI               -0.03726    0.15555  -0.240  0.81072
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26217  on 136877  degrees of freedom
Residual deviance: 26172  on 136871  degrees of freedom
AIC: 26186

Number of Fisher Scoring iterations: 6
```

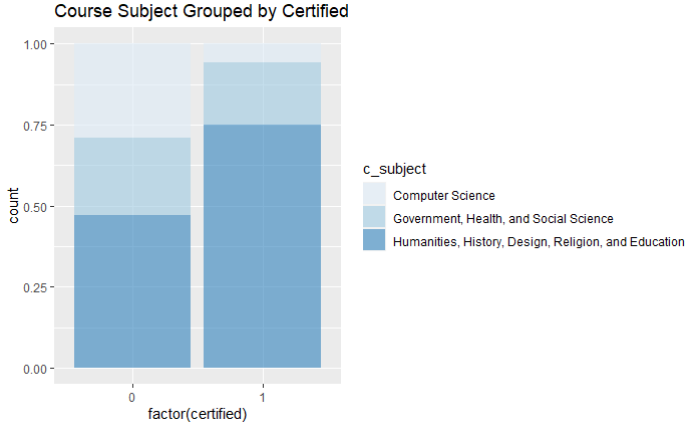
We can see that the intercept is negative and statistically significant, which could be seen as a baseline. For education background, the coefficients are all negative and statistically significant, indicating that people with high education levels are less likely to choose to get a certificate. One explanation is that people with high degrees do not have urgent needs for such certificates whereas people with low degrees need certificates as a sign of capability. Moreover, the coefficients of gender and HDI are negative but not statistically significant.

Course Info

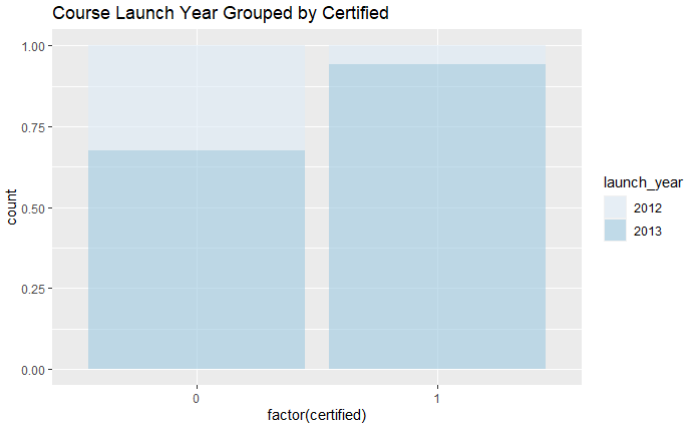
Course information may also influence people's choices. For example, a certificate in computer science may help the user with job applications. Therefore, we want to further examine the effects of course info.



Based on the figure, we can see that ER22x accounts for the biggest portion in people who choose to get a certificate. Surprisingly, few people in CS50x choose to get certificates.



Based on the figure, we can see that Humanities, History, Design, Religion, and Education accounts for the biggest portion of people who choose to get a certificate. This is in accordance with the Course Grouped by Certified figure.



Based on the figure, we can see that courses launched in 2013 account for the biggest portion of people who choose to get a certificate. This is partly because the courses are relatively new so that people are more willing to get a certificate.

Model 6

In Model 6, we add course info like course subject, launch year, and the number of participants. For simplicity, we do not show the model specification table here.

Result

```
Call:
glm(formula = certified ~ LoE_DI + gender + HDI + launch_year +
    c_subject + num_participants, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3963  -0.2634  -0.1689  -0.0912   4.0700

Coefficients:
                                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                                           -9.883e+00  3.226e-01 -30.636  < 2e-16 ***
LoE_DIBachelor's                                      -6.702e-01  1.052e-01  -6.368 1.92e-10 ***
LoE_DIDoctorate                                       -5.893e-01  1.512e-01  -3.897 9.73e-05 ***
LoE_DIMaster's                                        -4.804e-01  1.062e-01  -4.523 6.09e-06 ***
LoE_DISecondary                                       -4.934e-01  1.065e-01  -4.635 3.57e-06 ***
genderm                                                1.551e-01  4.170e-02   3.719   0.0002 ***
HDI                                                   -1.980e-01  1.606e-01  -1.233   0.2176
launch_year2013                                        5.516e+00  1.002e+00   5.507 3.65e-08 ***
c_subjectGovernment, Health, and Social Science        1.375e-01  1.014e+00   0.136   0.8922
c_subjectHumanities, History, Design, Religion, and Education -4.082e-01  1.008e+00  -0.405   0.6854
num_participants                                       3.779e-05  1.981e-06  19.081  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26217  on 136877  degrees of freedom
Residual deviance: 24410  on 136867  degrees of freedom
AIC: 24432

Number of Fisher Scoring iterations: 10
```

We can see the results are similar to Model 1 in terms of user info. As for launch year, the coefficient is positive, which means that people are more willing to choose to get a certificate from newly launched courses. The coefficient of the number of participants is also positive and statistically significant, indicating that the more popular the course is, the more likely people will choose to get a certificate. As for course subject, the coefficients are not statistically significant, which means that controlling for other variables, people are indifferent about the course subject when deciding whether or not to get a certificate.

Comparison of Model 5 and Model 6

Model	Log-Likelihood	McFadden R^2	Likelihood Ratio Test
Model 4	-13086.12	0.0017	p.value=4.890432e-08
Model 5	-12205.14	0.0689	p.value = < 2.22e-16

Based on the table, we can see that Model 5 improved. Therefore, we want to conduct a likelihood ratio test to see if the improvement is significant.

```
Likelihood ratio test

Model 1: certified ~ LoE_DI + gender + HDI
Model 2: certified ~ LoE_DI + gender + HDI + launch_year + c_subject +
    num_participants
  #Df LogLik Df Chisq Pr(>Chisq)
1   7 -13086
2  11 -12205  4  1762  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the test result, we should reject the null hypothesis at a 95% level of confidence. Therefore, compared to Model 4, Model 5 has a significant improvement in fit.

Conclusion

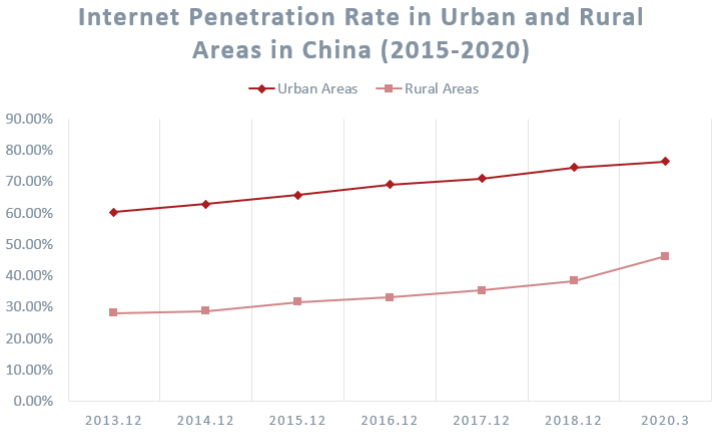
In this project, we use Multinomial Choice Models and Nested Logit Model to examine people's choice behavior when faced with five Harvardx courses of different subjects. The main findings are that CS course is overall more popular, females are more likely to choose liberal arts while people with low education level are more likely to choose CS course. We further examine people's choice behavior in terms of whether or not to get a certificate. We find that people with high education levels and people who register in CS course are less likely to choose to get a certificate. Detailed statistical tests are also adopted to evaluate the models. Based on the models, we can predict future demands and help improve MOOCs.

The main limitation of the project is due to the lack of latest data. Indeed, the data may only partly reflect what is happening right now. However, we believe in terms of liberal arts education, unlike tech-related courses, the research trend are relatively constant, so do the demand and characteristics. Therefore, the results are valid.

Discussion

Scholars claim that MOOCs may have potential to mitigate some of the world's educational disparities by facilitating access to high-quality education (Siemens, 2013; Christensen *et al.*, 2014; Kaplan and Haenlein, 2016). Similarly, one survey by China Internet Network Information Center indicates that 69.7% of the interviewees believe that online learning offers high-quality learning resources to people in less-developed areas, which helps them overcome poverty. However, prior researches on MOOC also find that across all geographic regions, MOOC students have very high levels of educational attainment and the student population tends to be young, male, and employed, with a majority from developed countries (Christensen et al., 2014). This is also the typical user in our dataset. Therefore, while online education seems promising in bridging the gap between the rich and the poor, the typical users are the same group of people who already have access to high-quality education resources. The individuals that online education should help the most take only a small proportion among early adopters.

The mission is not yet complete. One possible reason could be difficult access to the Internet itself. According to the 45th China Statistical Report on Internet Development by China Internet Network Information Center, there still lies a huge gap between the internet penetration rate in urban and rural areas. This is also the case around the world.



Source: China Internet Network Information Center (CNNIC)

Therefore, there is still a long way to go. To avoid further widening the gap, not only do we need to improve online learning menus based on predicted demands, but we also need to notice and keep developing regions from lagging behind.

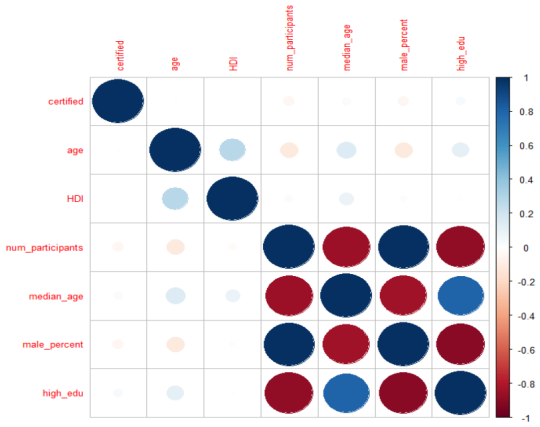
Appendix

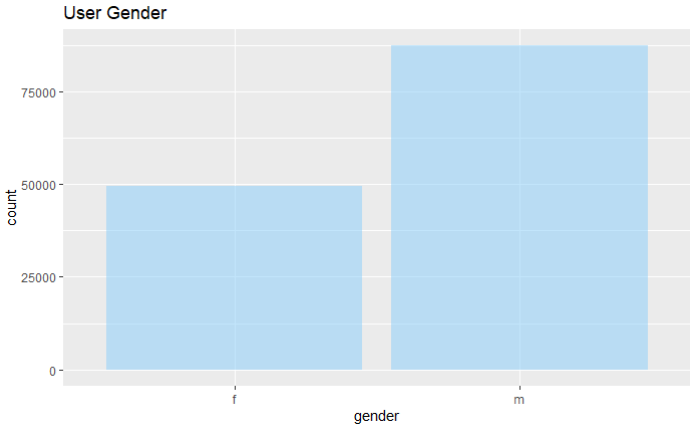
Variables Description

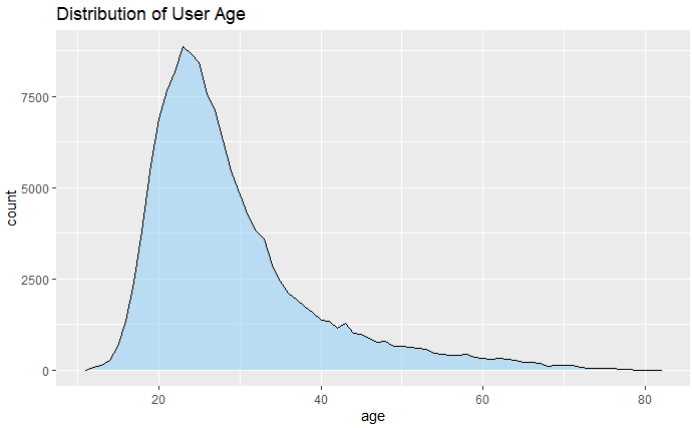
Variable Name	Description
cid	Course Name, e.g. CB22x
userid_DI	User ID, e.g. MHxPC130442623
certified	Dummy, 1 if the user chooses to earn a certificate, 0 otherwise.
country	Country name, some country names were replaced with the corresponding continent/region name.
HDI	Human Development Index
LoE_DI	Highest level of education complete, e.g. Bachelor's
age	User's age, year of course registration - year of born
gender	Gender of the user, e.g. m(male), f(female)
np	Number of participants
lyr	Launch year of the course
csub	Subject of the course, e.g. Computer Science

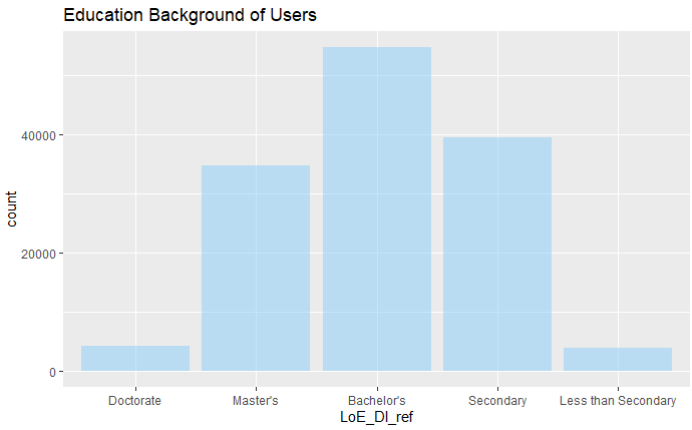
Data Summary and Related Figures

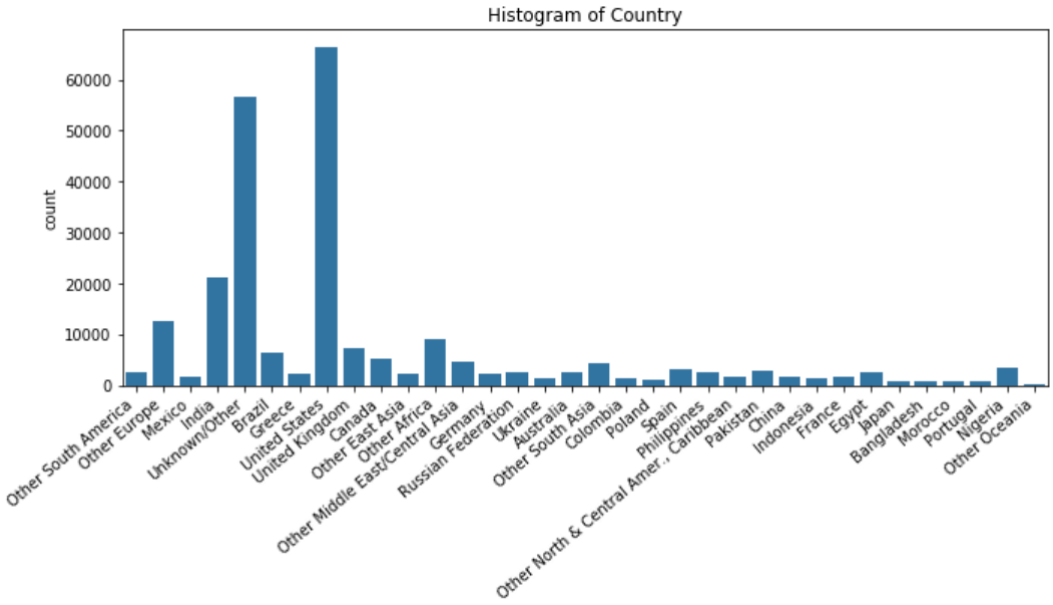
	certified	age	HDI
Count	136878	136878	136878
Mean	0.019	28.995	0.787
Std. Dev	0.138	10.260	0.129
Min	0	11	0.520
25%	0	22	0.713
50%	0	26	0.759
75%	0	33	0.914
Max	1	82	0.927

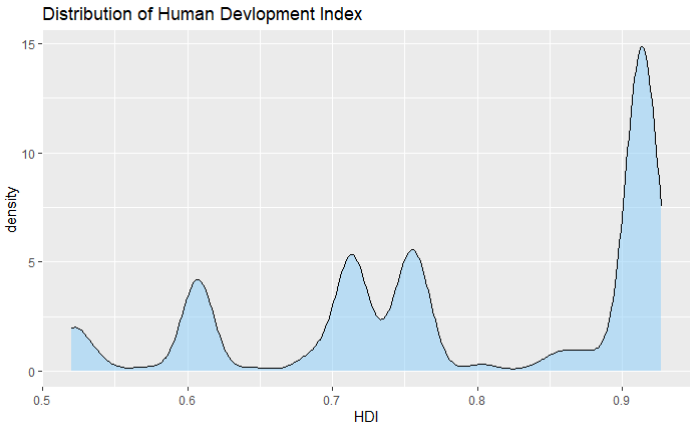

certified
age
HDI
num_participants
median_age
male_percent
high_edu
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


User Gender
count
75000
50000
25000
0
f
m
gender


Distribution of User Age
count
7500
5000
2500
0
20
40
60
80
age


Education Background of Users
count
40000
20000
0
Doctorate
Master's
Bachelor's
Secondary
Less than Secondary
LoE_DI_ref





References

Christensen, G. *et al.* (2014) 'The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why?', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2350964.

Chuang, I. (2017) 'HarvardX and MITx: Four Years of Open Online Courses -- Fall 2012-Summer 2016', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2889436.

HarvardX (no date) 'HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0'. Harvard Dataverse. doi: doi:10.7910/DVN/26147.

Ho, A. D. *et al.* (2014) 'HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2381263.

Kaplan, A. M. and Haenlein, M. (2016) 'Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster', *Business Horizons*. doi: 10.1016/j.bushor.2016.03.008.

Siemens, G. (2013) 'Massive Open Online Courses: Innovation in Education?', in *Open educational resources: innovation, research and practice*.

Train, K. and Croissant, Y. (2012) 'Kenneth Train's exercises using the mlogit package for R', *Mineo*.

Train, K. E. (2003) *Discrete choice methods with simulation*, *Discrete Choice Methods with Simulation*. doi: 10.1017/CBO9780511753930.