|  |
| --- |
| Shows plot by 16 courses of num forum posts (each dot represents multiple posts)—basically just showing the range |
| Number of forum posts (for all classes)-- scatterplot |
| FOR CS50: NO FORUM POSTS FOR ANYONE!! |
| In dataset: 98.83% have no forum posts aka only 1.16% of students posted in forums  (633,677 out of 641,138)  The following shows a histogram for those that did post at least 1 |
|  |
| Doesn’t seem to be a clear relationship b/w forumposts, grade (for entire dataset)—lots of forum posts, higher grades—but also seeing a varied grade distribution for other numbers of forum posts |
| Need to look for in depth—but seem to be differences by country |
|  |
| summary(mit)  course\_id userid\_DI  HarvardX/CS50x/2012 :169621 MHxPC130027283: 16  MITx/6.00x/2012\_Fall : 66731 MHxPC130103410: 16  MITx/6.00x/2013\_Spring : 57715 MHxPC130121287: 16  HarvardX/ER22x/2013\_Spring: 57406 MHxPC130126780: 16  HarvardX/PH207x/2012\_Fall : 41592 MHxPC130183602: 16  MITx/6.002x/2012\_Fall : 40811 MHxPC130200926: 16  (Other) :207262 (Other) :641042  registered viewed explored  Min. :1 Min. :0.0000 Min. :0.0000  1st Qu.:1 1st Qu.:0.0000 1st Qu.:0.0000  Median :1 Median :1.0000 Median :0.0000  Mean :1 Mean :0.6243 Mean :0.0619  3rd Qu.:1 3rd Qu.:1.0000 3rd Qu.:0.0000  Max. :1 Max. :1.0000 Max. :1.0000    certified final\_cc\_cname\_DI  Min. :0.00000 United States :184240  1st Qu.:0.00000 India : 88696  Median :0.00000 Unknown/Other : 82029  Mean :0.02759 Other Europe : 40377  3rd Qu.:0.00000 Other Africa : 23897  Max. :1.00000 United Kingdom: 22131  (Other) :199768  LoE\_DI YoB gender  : 47876 Min. :1931 : 28674  Bachelor's :219768 1st Qu.:1982 f :142795  Doctorate : 13387 Median :1988 m :411520  Less than Secondary: 14092 Mean :1985 o : 17  Master's :118189 3rd Qu.:1991 NA's: 58132  Secondary :169694 Max. :2013  NA's : 58132 NA's :96605  grade start\_time\_DI last\_event\_DI  Min. :0.00 2012-08-17: 10165 :178954  1st Qu.:0.00 2013-01-23: 8368 2013-11-17: 7046  Median :0.00 2012-10-15: 6766 2013-03-13: 4747  Mean :0.03 2012-08-16: 6369 2013-03-14: 3470  3rd Qu.:0.00 2012-12-20: 5858 2012-10-16: 3339  Max. :1.01 2013-02-14: 5810 2012-10-15: 3125  NA's :57400 (Other) :597802 (Other) :440457  nevents ndays\_act nplay\_video  Min. : 1 Min. : 1.00 Min. : 1.0  1st Qu.: 3 1st Qu.: 1.00 1st Qu.: 5.0  Median : 24 Median : 2.00 Median : 18.0  Mean : 431 Mean : 5.71 Mean : 114.8  3rd Qu.: 158 3rd Qu.: 4.00 3rd Qu.: 73.0  Max. :197757 Max. :205.00 Max. :98517.0  NA's :199151 NA's :162743 NA's :457530  nchapters nforum\_posts roles  Min. : 1.00 Min. : 0.00000 Mode:logical  1st Qu.: 1.00 1st Qu.: 0.00000 NA's:641138  Median : 2.00 Median : 0.00000  Mean : 3.63 Mean : 0.01897  3rd Qu.: 4.00 3rd Qu.: 0.00000  Max. :48.00 Max. :20.00000  NA's :258753  incomplete\_flag  Min. :1  1st Qu.:1  Median :1  Mean :1  3rd Qu.:1  Max. :1  NA's :540977 |
| **NOTE!!! Need to make sure we’re working with only the complete data! My plots above did not filter out the incomplete flags!!**    For country—almost as many unknown/other as for US in total. Lots of missing values! |
| **Country distribution for CS50 (complete only)**  **Australia**  **1127**  **Bangladesh**  **378**  **Brazil**  **2021**  **Canada**  **2496**  **China**  **511**  **Colombia**  **453**  **Egypt**  **1021**  **France**  **654**  **Germany**  **851**  **Greece**  **415**  **India**  **9473**  **Indonesia**  **396**  **Japan**  **176**  **Mexico**  **613**  **Morocco**  **603**  **Nigeria**  **1165**  **Other Africa**  **3018**  **Other East Asia**  **545**  **Other Europe**  **4965**  **Other Middle East/Central Asia**  **2043**  **Other North & Central Amer., Caribbean**  **716**  **Other Oceania**  **106**  **Other South America**  **1105**  **Other South Asia**  **1607**  **Pakistan**  **1460**  **Philippines**  **925**  **Poland**  **533**  **Portugal**  **345**  **Russian Federation**  **1450**  **Spain**  **946**  **Ukraine**  **728**  **United Kingdom**  **3043**  **United States**  **28350**  **Unknown/Other**  **26521** |
| > table(cs50\_mit\_c$gender)  f m o  3400 17215 73314 8  Only 19% female in CS50 |
| table(cs50\_mit\_c$grade)  0 1  98806 1286  Weird… other courses had a range of data (also checked for incomplete—still only 0 and 1—pass/fail? |
| For CS50 FB interactions… not much data here |
| Also missing a lot for country data… |
| Colored by locale… really disappointed here! But does show—small number of users with high number of interactions—do we see similar things in the forums?? Large amount of missign data for gender and locale |
|  |

**My R log:**

R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"

Copyright (C) 2014 The R Foundation for Statistical Computing

Platform: x86\_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.

Type 'contributors()' for more information and

'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or

'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

Loading required package: tuneR

tuneR >= 1.0 has changed its Wave class definition.

Use updateWave(object) to convert Wave objects saved with previous versions of tuneR.

> Harvard.MITdataset <- read.csv("~/Desktop/Data Pipeline/DataPipelineProject1/Harvard-MITdataset.csv")

> View(Harvard.MITdataset)

> mit <- Harvard.MITdataset

> plot(mit$nforum\_posts, type= "h")

> plot(mit$nforum\_posts, course\_id[1])

Error in plot(mit$nforum\_posts, course\_id[1]) :

error in evaluating the argument 'y' in selecting a method for function 'plot': Error: object 'course\_id' not found

> plot(mit$nforum\_posts, mit$course\_id[1])

Error in xy.coords(x, y, xlabel, ylabel, log) :

'x' and 'y' lengths differ

> plot(mit$nforum\_posts, mit$course\_id)

> pairs(~nforum\_posts + gender + grade + nevents + final\_cc\_cname\_DI,data=mit,

+ main="Simple Scatterplot Matrix")

> with(mit[(mit$course\_id == "HarvardX/CS50x/2012"),], plot(mit$nforum\_posts))

> plot(mit$nforum\_posts)

> hist(mit$nforum\_posts)

> max(mit$nforum\_posts)

[1] 20

> hist(mit$nforum\_posts>0)

Error in hist.default(mit$nforum\_posts > 0) : 'x' must be numeric

> anyposts <- mit$nforum\_posts>0

> hist(anyposts)

Error in hist.default(anyposts) : 'x' must be numeric

> with(mit$nforum\_posts>0, hist(mit$nforum\_posts))

Error in eval(substitute(expr), data, enclos = parent.frame()) :

invalid 'envir' argument of type 'logical'

> cs50\_mit <- mit[ which(mit$course\_id=='HarvardX/CS50x/2012', ]

Error: unexpected ']' in "cs50\_mit <- mit[ which(mit$course\_id=='HarvardX/CS50x/2012', ]"

> cs50\_mit <- mit[ which(mit$course\_id=='HarvardX/CS50x/2012'), ]

> hit(cs50\_mit$nforum\_posts)

Error: could not find function "hit"

> hist(cs50\_mit$nforum\_posts)

> View(cs50\_mit)

> max(cs50\_mit$nforum\_posts)

[1] 0

> hist(cs50\_mit$final\_cc\_cname\_DI)

Error in hist.default(cs50\_mit$final\_cc\_cname\_DI) : 'x' must be numeric

> mit\_posts <- mit[which(mit$nforum\_posts >0),]

> hist(mit\_posts$nforum\_posts)

> 10-2

[1] 8

> 641138-7461

[1] 633677

> (641138-7461)/641138

[1] 0.9883629

> 7461/641138

[1] 0.01163712

> plot(mit$nforum\_posts, grade)

Error in plot(mit$nforum\_posts, grade) :

error in evaluating the argument 'y' in selecting a method for function 'plot': Error: object 'grade' not found

> plot(mit$nforum\_posts, mit$grade)

> plot(mit$nforum\_posts, mit$final\_cc\_cname\_DI)

> sum(cs50\_mit$final\_cc\_cname\_DI)

Error in Summary.factor(c(33L, 33L, 33L, 33L, 33L, 34L, 33L, 14L, 33L, :

‘sum’ not meaningful for factors

> levels(cs50\_mit$final\_cc\_cname\_DI)

[1] "Australia" "Bangladesh"

[3] "Brazil" "Canada"

[5] "China" "Colombia"

[7] "Egypt" "France"

[9] "Germany" "Greece"

[11] "India" "Indonesia"

[13] "Japan" "Mexico"

[15] "Morocco" "Nigeria"

[17] "Other Africa" "Other East Asia"

[19] "Other Europe" "Other Middle East/Central Asia"

[21] "Other North & Central Amer., Caribbean" "Other Oceania"

[23] "Other South America" "Other South Asia"

[25] "Pakistan" "Philippines"

[27] "Poland" "Portugal"

[29] "Russian Federation" "Spain"

[31] "Ukraine" "United Kingdom"

[33] "United States" "Unknown/Other"

> mytable <-table(mit)

Error in table(mit) : attempt to make a table with >= 2^31 elements

> summary(mit)

course\_id userid\_DI

HarvardX/CS50x/2012 :169621 MHxPC130027283: 16

MITx/6.00x/2012\_Fall : 66731 MHxPC130103410: 16

MITx/6.00x/2013\_Spring : 57715 MHxPC130121287: 16

HarvardX/ER22x/2013\_Spring: 57406 MHxPC130126780: 16

HarvardX/PH207x/2012\_Fall : 41592 MHxPC130183602: 16

MITx/6.002x/2012\_Fall : 40811 MHxPC130200926: 16

(Other) :207262 (Other) :641042

registered viewed explored

Min. :1 Min. :0.0000 Min. :0.0000

1st Qu.:1 1st Qu.:0.0000 1st Qu.:0.0000

Median :1 Median :1.0000 Median :0.0000

Mean :1 Mean :0.6243 Mean :0.0619

3rd Qu.:1 3rd Qu.:1.0000 3rd Qu.:0.0000

Max. :1 Max. :1.0000 Max. :1.0000

certified final\_cc\_cname\_DI

Min. :0.00000 United States :184240

1st Qu.:0.00000 India : 88696

Median :0.00000 Unknown/Other : 82029

Mean :0.02759 Other Europe : 40377

3rd Qu.:0.00000 Other Africa : 23897

Max. :1.00000 United Kingdom: 22131

(Other) :199768

LoE\_DI YoB gender

: 47876 Min. :1931 : 28674

Bachelor's :219768 1st Qu.:1982 f :142795

Doctorate : 13387 Median :1988 m :411520

Less than Secondary: 14092 Mean :1985 o : 17

Master's :118189 3rd Qu.:1991 NA's: 58132

Secondary :169694 Max. :2013

NA's : 58132 NA's :96605

grade start\_time\_DI last\_event\_DI

Min. :0.00 2012-08-17: 10165 :178954

1st Qu.:0.00 2013-01-23: 8368 2013-11-17: 7046

Median :0.00 2012-10-15: 6766 2013-03-13: 4747

Mean :0.03 2012-08-16: 6369 2013-03-14: 3470

3rd Qu.:0.00 2012-12-20: 5858 2012-10-16: 3339

Max. :1.01 2013-02-14: 5810 2012-10-15: 3125

NA's :57400 (Other) :597802 (Other) :440457

nevents ndays\_act nplay\_video

Min. : 1 Min. : 1.00 Min. : 1.0

1st Qu.: 3 1st Qu.: 1.00 1st Qu.: 5.0

Median : 24 Median : 2.00 Median : 18.0

Mean : 431 Mean : 5.71 Mean : 114.8

3rd Qu.: 158 3rd Qu.: 4.00 3rd Qu.: 73.0

Max. :197757 Max. :205.00 Max. :98517.0

NA's :199151 NA's :162743 NA's :457530

nchapters nforum\_posts roles

Min. : 1.00 Min. : 0.00000 Mode:logical

1st Qu.: 1.00 1st Qu.: 0.00000 NA's:641138

Median : 2.00 Median : 0.00000

Mean : 3.63 Mean : 0.01897

3rd Qu.: 4.00 3rd Qu.: 0.00000

Max. :48.00 Max. :20.00000

NA's :258753

incomplete\_flag

Min. :1

1st Qu.:1

Median :1

Mean :1

3rd Qu.:1

Max. :1

NA's :540977

> mit\_complete <- mit[which(mit$incomplete\_flag is.na),]

Error: unexpected symbol in "mit\_complete <- mit[which(mit$incomplete\_flag is.na"

> mit\_complete <- mit[which(is.na(mit$incomplete\_flag) ),]

> cs50\_mit\_c <- mit\_complete[ which(mit\_complete$course\_id=='HarvardX/CS50x/2012'), ]

mit\_complete <- mit[which(is.na(mit$incomplete\_flag) ),]

> cs50\_mit\_c <- mit\_complete[ which(mit\_complete$course\_id=='HarvardX/CS50x/2012'), ]

> table(mit\_complete$final\_cc\_cname\_DI)