4/25/2022

Data Analysis of Online Courses from Harvard and MIT

INFX412 Semester Project

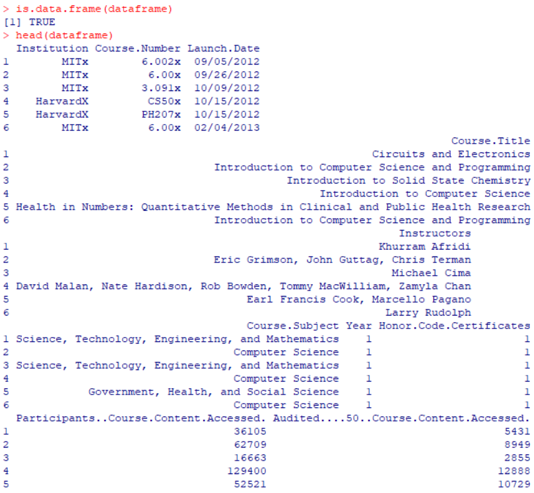
Jaylen Gordon
C00401520

Table Of Contents

I. Dataset

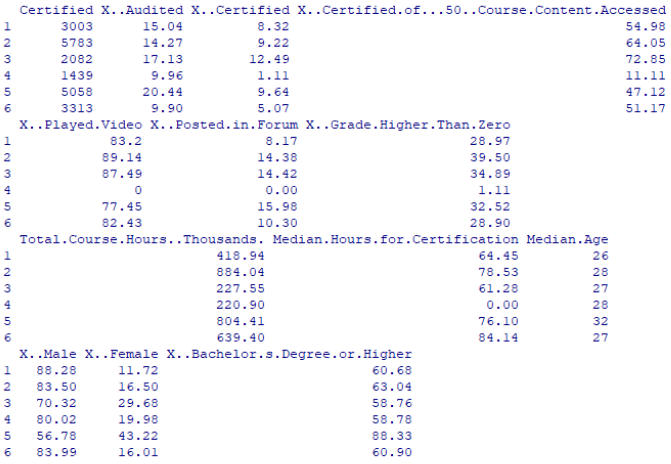
Dataset Description

The dataset selected is titled “Online Courses from Harvard and MIT.” Both Harvard and Massachusetts Institute of Technology (MIT) created this enormous open online course provider. It offers online university-level courses in a variety of fields to a global student base, with some courses available for free. Since 2012, the edX platform has curated 290 Harvard and MIT online courses, 250 thousand certifications, 4.5 million participants, and 28 million participant hours. This data was provided as an appendix to MIT professor Isaac Chuang's and Harvard University professor Andrew Ho's paper "HarvardX and MITx: Four Years of Open Online Courses." The original dataset was loaded into R from a Microsoft Excel CSV file. I used the is.data.frame command (figure I) to check that the dataset was loaded into R correctly and with no errors. The head command (figure I-II) shows the first 6 rows for each column of the dataset. From using the dim command (figure III), we can see that this dataset has 290 rows and 23 columns making this a very rich dataset.



```
> is.data.frame(dataframe)
[1] TRUE
> head(dataframe)
  Institution Course.Number Launch.Date
1        MITx        6.002x  09/05/2012
2        MITx         6.00x  09/26/2012
3        MITx        3.091x  10/09/2012
4    HarvardX         CS50x  10/15/2012
5    HarvardX        PH207x  10/15/2012
6        MITx         6.00x  02/04/2013
                                                                    Course.Title
1                                                       Circuits and Electronics
2                               Introduction to Computer Science and Programming
3                                          Introduction to Solid State Chemistry
4                                               Introduction to Computer Science
5 Health in Numbers: Quantitative Methods in Clinical and Public Health Research
6                               Introduction to Computer Science and Programming
                                                            Instructors
1                                                        Khurram Afridi
2                               Eric Grimson, John Guttag, Chris Terman
3                                                          Michael Cima
4 David Malan, Nate Hardison, Rob Bowden, Tommy MacWilliam, Zamyla Chan
5                                    Earl Francis Cook, Marcello Pagano
6                                                         Larry Rudolph
                                     Course.Subject Year Honor.Code.Certificates
1 Science, Technology, Engineering, and Mathematics    1                       1
2                                  Computer Science    1                       1
3 Science, Technology, Engineering, and Mathematics    1                       1
4                                  Computer Science    1                       1
5            Government, Health, and Social Science    1                       1
6                                  Computer Science    1                       1
  Participants..Course.Content.Accessed. Audited....50..Course.Content.Accessed.
1                                  36105                                    5431
2                                  62709                                    8949
3                                  16663                                    2855
4                                 129400                                   12888
5                                  52521                                   10729
```

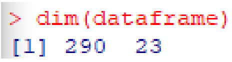
I.



```
  Certified X..Audited X..Certified X..Certified.of...50..Course.Content.Accessed
1      3003      15.04         8.32                                         54.98
2      5783      14.27         9.22                                         64.05
3      2082      17.13        12.49                                         72.85
4      1439       9.96         1.11                                         11.11
5      5058      20.44         9.64                                         47.12
6      3313       9.90         5.07                                         51.17
  X..Played.Video X..Posted.in.Forum X..Grade.Higher.Than.Zero
1            83.2               8.17                     28.97
2           89.14              14.38                     39.50
3           87.49              14.42                     34.89
4               0               0.00                      1.11
5           77.45              15.98                     32.52
6           82.43              10.30                     28.90
  Total.Course.Hours..Thousands. Median.Hours.for.Certification Median.Age
1                         418.94                          64.45         26
2                         884.04                          78.53         28
3                         227.55                          61.28         27
4                         220.90                           0.00         28
5                         804.41                          76.10         32
6                         639.40                          84.14         27
  X..Male X..Female X..Bachelor.s.Degree.or.Higher
1   88.28     11.72                          60.68
2   83.50     16.50                          63.04
3   70.32     29.68                          58.76
4   80.02     19.98                          58.78
5   56.78     43.22                          88.33
6   83.99     16.01                          60.90
```

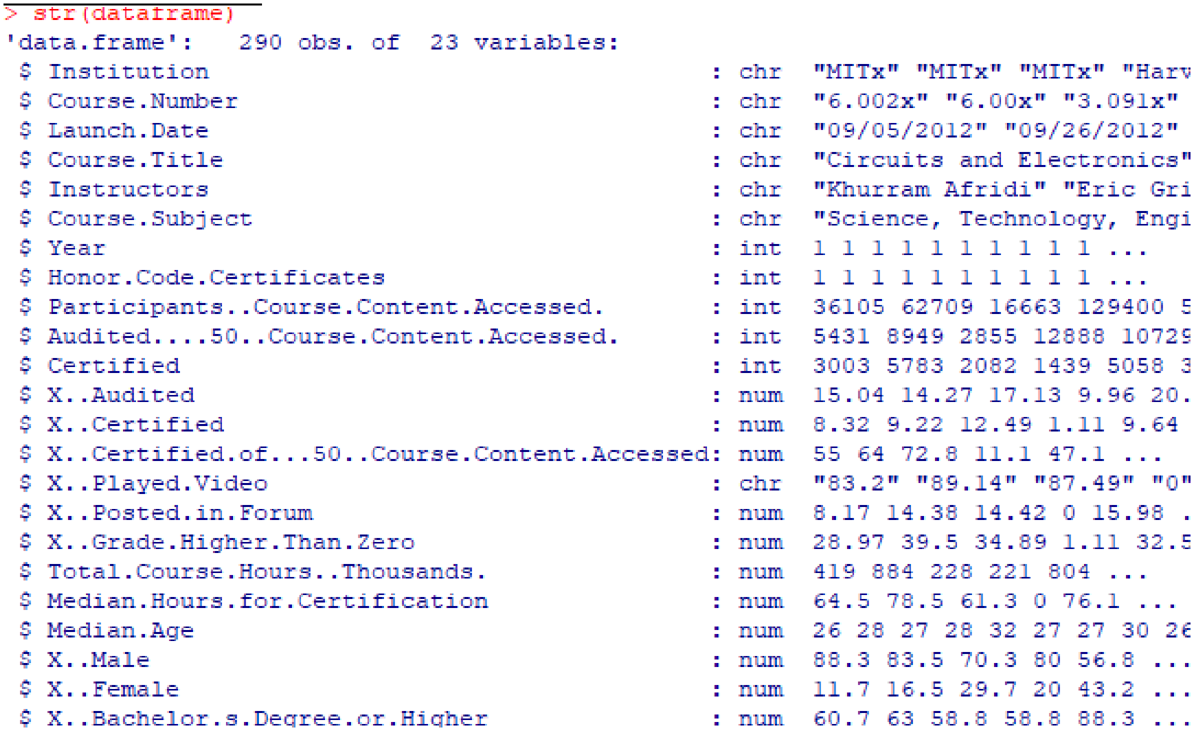
II.

III.



```
> dim(dataframe)
[1] 290  23
```

Dataset Structure



```
> str(dataframe)
'data.frame':   290 obs. of  23 variables:
 $ Institution                                  : chr  "MITx" "MITx" "MITx" "Harv
 $ Course.Number                                : chr  "6.002x" "6.00x" "3.091x"
 $ Launch.Date                                  : chr  "09/05/2012" "09/26/2012"
 $ Course.Title                                 : chr  "Circuits and Electronics"
 $ Instructors                                  : chr  "Khurram Afridi" "Eric Gri
 $ Course.Subject                               : chr  "Science, Technology, Engi
 $ Year                                         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Honor.Code.Certificates                      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Participants..Course.Content.Accessed.       : int  36105 62709 16663 129400 5
 $ Audited....50..Course.Content.Accessed.      : int  5431 8949 2855 12888 10729
 $ Certified                                    : int  3003 5783 2082 1439 5058 3
 $ X..Audited                                   : num  15.04 14.27 17.13 9.96 20.
 $ X..Certified                                 : num  8.32 9.22 12.49 1.11 9.64
 $ X..Certified.of...50..Course.Content.Accessed: num  55 64 72.8 11.1 47.1 ...
 $ X..Played.Video                              : chr  "83.2" "89.14" "87.49" "0"
 $ X..Posted.in.Forum                           : num  8.17 14.38 14.42 0 15.98 .
 $ X..Grade.Higher.Than.Zero                    : num  28.97 39.5 34.89 1.11 32.5
 $ Total.Course.Hours..Thousands.               : num  419 884 228 221 804 ...
 $ Median.Hours.for.Certification               : num  64.5 78.5 61.3 0 76.1 ...
 $ Median.Age                                   : num  26 28 27 28 32 27 27 30 26
 $ X..Male                                      : num  88.3 83.5 70.3 80 56.8 ...
 $ X..Female                                    : num  11.7 16.5 29.7 20 43.2 ...
 $ X..Bachelor.s.Degree.or.Higher               : num  60.7 63 58.8 58.8 88.3 ...
```

By looking at the output of this command we can see that it needs some cleaning. To begin with, we can rename our variable names. After close observation, a conversion is also needed for a proper analysis.

Variable Description

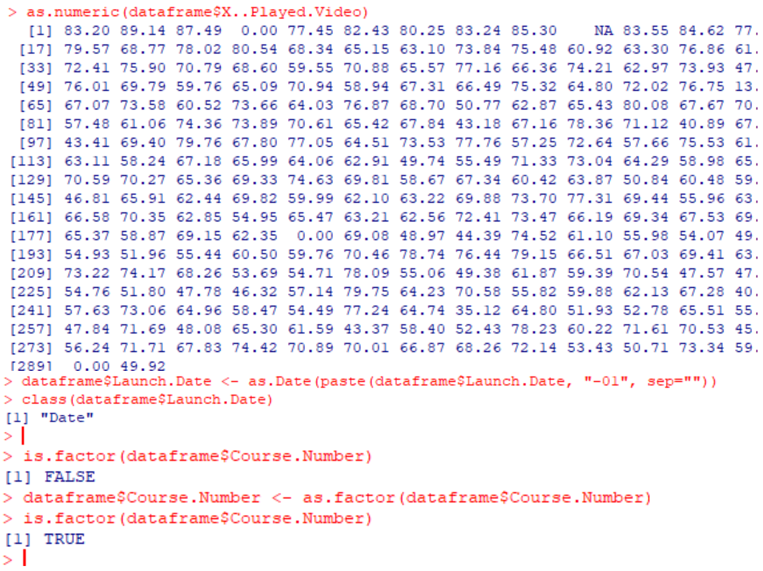
Institution	online course holders
Course Number	the unique id of each course
Launch Date	the launch date of each course
Course Title	the title of each course
Instructor	the instructors of each course
Course Subject	the subject of each course
Year	the last time of each course
Honor Code	with (1), without (0).
Participants (Course Content Accessed)	the number of participants who have accessed the course
Audited (> 50% Course Content Accessed)	the number of participants who have audited more than 50% of the course
Certified	the number of participants who have been certified
% Certified	the percent of the certified
% Audited	the percent of the audited
% Played Video	the percent of playing video
% Posted in Forum	the percent of posting in forum
% Grade Higher Than Zero	the percent of grade higher than zero
xTotal Course Hours (Thousands)	total course hours(per 1000)
Median Hours for Certification	median hours for certification
Median Age	median age of the participants
% Male	the percent of the male
% Female	the percent of the female
% Bachelor's Degree or Higher	the percent of bachelor's degree of higher

% appears as X in data*

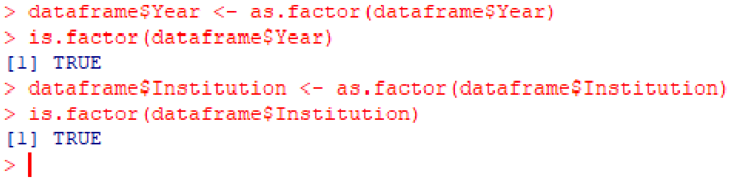
Data Cleaning

According to Online Courses from Harvard and MIT | Kaggle, this data set should contain 11 Decimal, 6 String, 5 Integer, and 1 Other variables. However, we will observe this dataset and make any conversions if needed. This dataset has Percent_Played_Video as a character. It should be a numeric variable. Therefore, a conversion was done from character to a numeric variable using the as.numeric function.(figure VI). After observing the dataset, we could also convert Launch_Date to a Date class instead of chr. Course_Date was also converted from chr using as.factor. as well as Year and Institution. Figure VIII shows the code and output to get rid of duplicate data within the Course_Hourse_Thousands. column.

VI

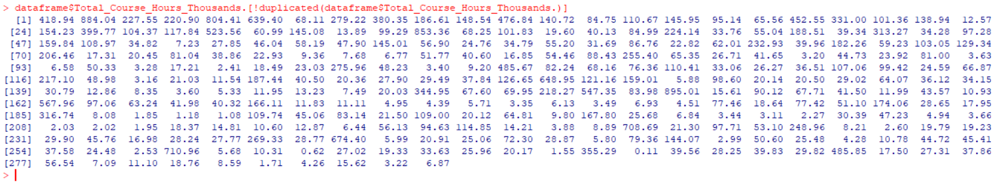


```
> as.numeric(dataframe$X..Played.Video)
  [1] 83.20 89.14 87.49  0.00 77.45 82.43 80.25 83.24 85.30    NA 83.55 84.62 77.
 [17] 79.57 68.77 78.02 80.54 68.34 65.15 63.10 73.84 75.48 60.92 63.30 76.86 61.
 [33] 72.41 75.90 70.79 68.60 59.55 70.88 65.57 77.16 66.36 74.21 62.97 73.93 47.
 [49] 76.01 69.79 59.76 65.09 70.94 58.94 67.31 66.49 75.32 64.80 72.02 76.75 13.
 [65] 67.07 73.58 60.52 73.66 64.03 76.87 68.70 50.77 62.87 65.43 80.08 67.67 70.
 [81] 57.48 61.06 74.36 73.89 70.61 65.42 67.84 43.18 67.16 78.36 71.12 40.89 67.
 [97] 43.41 69.40 79.76 67.80 77.05 64.51 73.53 77.76 57.25 72.64 57.66 75.53 61.
[113] 63.11 58.24 67.18 65.99 64.06 62.91 49.74 55.49 71.33 73.04 64.29 58.98 65.
[129] 70.59 70.27 65.36 69.33 74.63 69.81 58.67 67.34 60.42 63.87 50.84 60.48 59.
[145] 46.81 65.91 62.44 69.82 59.99 62.10 63.22 69.88 73.70 77.31 69.44 55.96 63.
[161] 66.58 70.35 62.85 54.95 65.47 63.21 62.56 72.41 73.47 66.19 69.34 67.53 69.
[177] 65.37 58.87 69.15 62.35  0.00 69.08 48.97 44.39 74.52 61.10 55.98 54.07 49.
[193] 54.93 51.96 55.44 60.50 59.76 70.46 78.74 76.44 79.15 66.51 67.03 69.41 63.
[209] 73.22 74.17 68.26 53.69 54.71 78.09 55.06 49.38 61.87 59.39 70.54 47.57 47.
[225] 54.76 51.80 47.78 46.32 57.14 79.75 64.23 70.58 55.82 59.88 62.13 67.28 40.
[241] 57.63 73.06 64.96 58.47 54.49 77.24 64.74 35.12 64.80 51.93 52.78 65.51 55.
[257] 47.84 71.69 48.08 65.30 61.59 43.37 58.40 52.43 78.23 60.22 71.61 70.53 45.
[273] 56.24 71.71 67.83 74.42 70.89 70.01 66.87 68.26 72.14 53.43 50.71 73.34 59.
[289]  0.00 49.92
> dataframe$Launch.Date <- as.Date(paste(dataframe$Launch.Date, "-01", sep=""))
> class(dataframe$Launch.Date)
[1] "Date"
> |
> is.factor(dataframe$Course.Number)
[1] FALSE
> dataframe$Course.Number <- as.factor(dataframe$Course.Number)
> is.factor(dataframe$Course.Number)
[1] TRUE
> |
```



```
> dataframe$Year <- as.factor(dataframe$Year)
> is.factor(dataframe$Year)
[1] TRUE
> dataframe$Institution <- as.factor(dataframe$Institution)
> is.factor(dataframe$Institution)
[1] TRUE
> |
```

VII

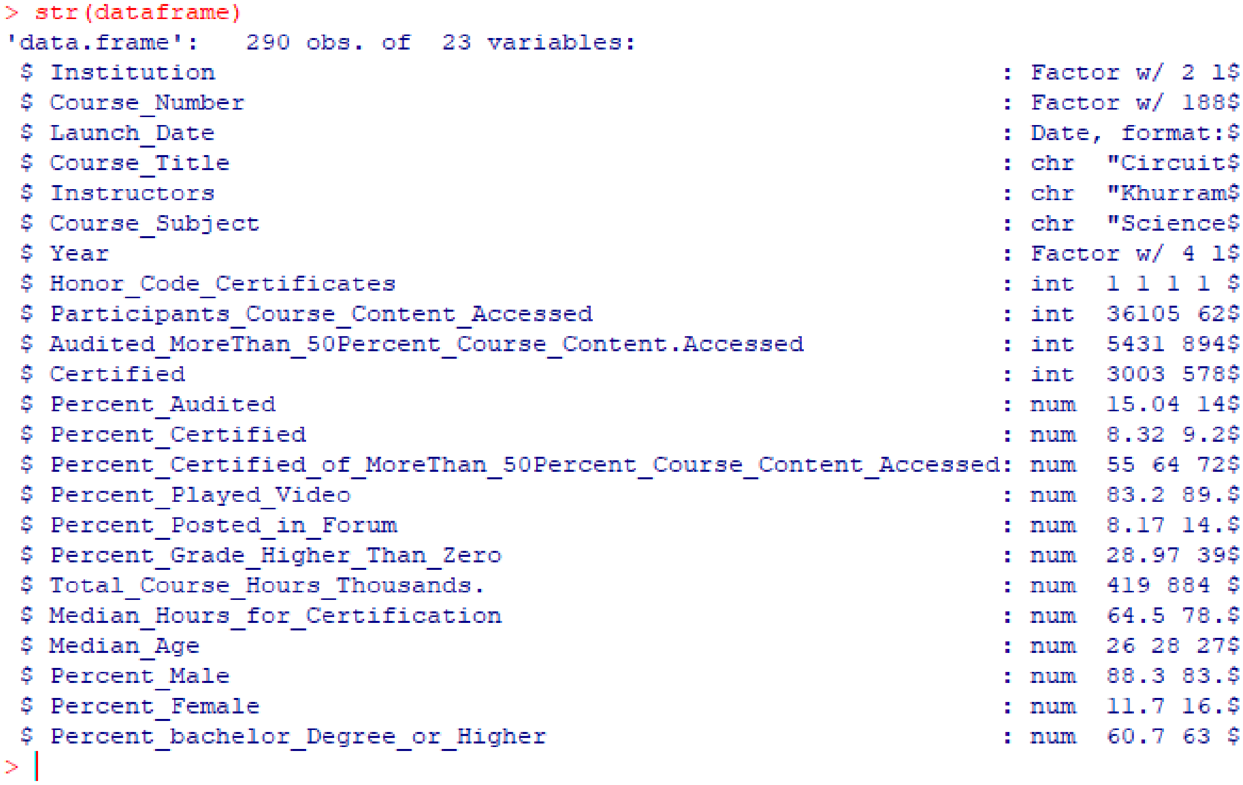


```
> dataframe$Total_Course_Hours_Thousands.[!duplicated(dataframe$Total_Course_Hours_Thousands.)]
  [1] 418.94 884.04 227.55 220.90 804.41 639.40  68.11 279.22 380.35 186.61 148.54 476.84 140.72  84.75 110.67 145.95  95.14  65.56 452.55 331.00 101.36 138.94  12.57
 [24] 154.23 399.77 104.37 117.84 523.56  60.99 145.08  13.89  99.29 853.36  68.25 101.83  19.60  40.13  84.99 224.14  33.76  55.04 188.51  39.34 313.27  34.28  97.28
 [47] 159.84 108.97  34.82   7.23  27.85  46.04  58.19  47.90 145.01  56.90  24.76  34.79  55.20  31.69  86.76  22.82  62.01 232.93  39.96 182.26  59.23 103.05 129.34
 [70] 206.46  17.31  20.45  81.04  38.86  22.93   9.36   7.68   6.77  51.77  40.60  16.85  54.46  88.43 255.40  65.35  26.71  41.65   3.20  44.73  23.92  81.00   3.63
 [93]   6.58  50.33   3.28  17.21   2.41  18.49  23.03 275.96  48.23   3.40   9.20 485.67  82.24  68.16  76.36 110.41  33.06  26.27  36.51 107.06  99.42  24.59  66.87
[116] 217.10  48.98   3.16  21.03  11.54 187.44  40.50  20.36  27.90  29.49  37.84 126.65 648.95 121.16 159.01   5.88  98.60  20.14  20.50  29.02  64.07  36.12  34.15
[139]  30.79  12.86   8.35   3.60   5.33  11.95  13.23   7.49  20.03 344.95  67.60  69.95 218.27 547.35  83.98 895.01  15.61  90.12  67.71  41.50  11.99  43.57  10.93
[162] 567.96  97.06  63.24  41.98  40.32 166.11  11.83  11.11   4.95   4.39   5.71   3.35   6.13   3.49   6.93   4.51  77.46  18.64  77.42  51.10 174.06  28.65  17.95
[185] 316.74   8.08   1.85   1.18   1.08 109.74  45.06  83.14  21.50 109.00  20.12  64.81   9.80 167.80  25.68   6.84   3.44   3.11   2.27  30.39  47.23   4.94   3.66
[208]   2.03   2.02   1.95  18.37  14.81  10.60  12.87   6.44  56.13  94.63 114.85  14.21   3.88   8.89 708.69  21.30  97.71  53.10 248.96   8.21   2.60  19.79  19.23
[231]  29.90  45.76  16.98  28.24  27.77 269.33  28.77 674.40   5.99  20.91  25.06  72.30  28.87   5.80  79.36 144.07   2.99  50.60  25.48   4.28  10.78  44.72  45.41
[254]  37.58  24.48   2.53 710.96   5.68  10.31   0.62  27.02  19.33  33.63  25.96  20.17   1.55 355.29   0.11  39.56  28.25  39.83  29.82 485.85  17.50  27.31  37.86
[277]  56.54   7.09  11.10  18.76   8.59   1.71   4.26  15.62   3.22   6.87
> |
```

VIII

Cleaned Data Structure

The final step in cleaning data was renaming the colums for an easier read, leading the final cleaned data structure to look like the following:



```
> str(dataframe)
'data.frame':   290 obs. of  23 variables:
 $ Institution                                           : Factor w/ 2 1$
 $ Course_Number                                         : Factor w/ 188$
 $ Launch_Date                                           : Date, format:$
 $ Course_Title                                          : chr  "Circuit$
 $ Instructors                                           : chr  "Khurram$
 $ Course_Subject                                        : chr  "Science$
 $ Year                                                  : Factor w/ 4 1$
 $ Honor_Code_Certificates                               : int  1 1 1 1 $
 $ Participants_Course_Content_Accessed                  : int  36105 62$
 $ Audited_MoreThan_50Percent_Course_Content.Accessed    : int  5431 894$
 $ Certified                                             : int  3003 578$
 $ Percent_Audited                                       : num  15.04 14$
 $ Percent_Certified                                     : num  8.32 9.2$
 $ Percent_Certified_of_MoreThan_50Percent_Course_Content_Accessed: num  55 64 72$
 $ Percent_Played_Video                                  : num  83.2 89.$
 $ Percent_Posted_in_Forum                               : num  8.17 14.$
 $ Percent_Grade_Higher_Than_Zero                        : num  28.97 39$
 $ Total_Course_Hours_Thousands.                         : num  419 884 $
 $ Median_Hours_for_Certification                        : num  64.5 78.$
 $ Median_Age                                            : num  26 28 27$
 $ Percent_Male                                          : num  88.3 83.$
 $ Percent_Female                                        : num  11.7 16.$
 $ Percent_bachelor_Degree_or_Higher                     : num  60.7 63 $
> |
```

Data Citations

edX. (2017, January 27). *Online courses from Harvard and MIT*. Kaggle. Retrieved April 24, 2022, from https://www.kaggle.com/datasets/edx/course-study?resource=download

Expectations

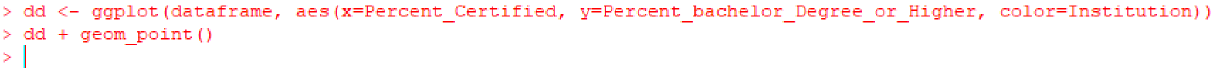
From the description of this dataset, we can compare the two programs we have information about. This will be a useful dataset in locating general information on MIT and

Harvard’s online programs. Being a student looking into higher education after graduation, I believe research for this dataset would be very beneficial. I am interested in comparing the amount of student with bachelor’s degrees already obtained as opposed to those who are only certified. Another variable that I’ll be analyzing is the number of courses in each course subject. My hypothesis is that computer science will have the greatest amount of participants. My second hypothesis is that the data for both MITx and Harvardx programs are similar to each other.

II. Analysis

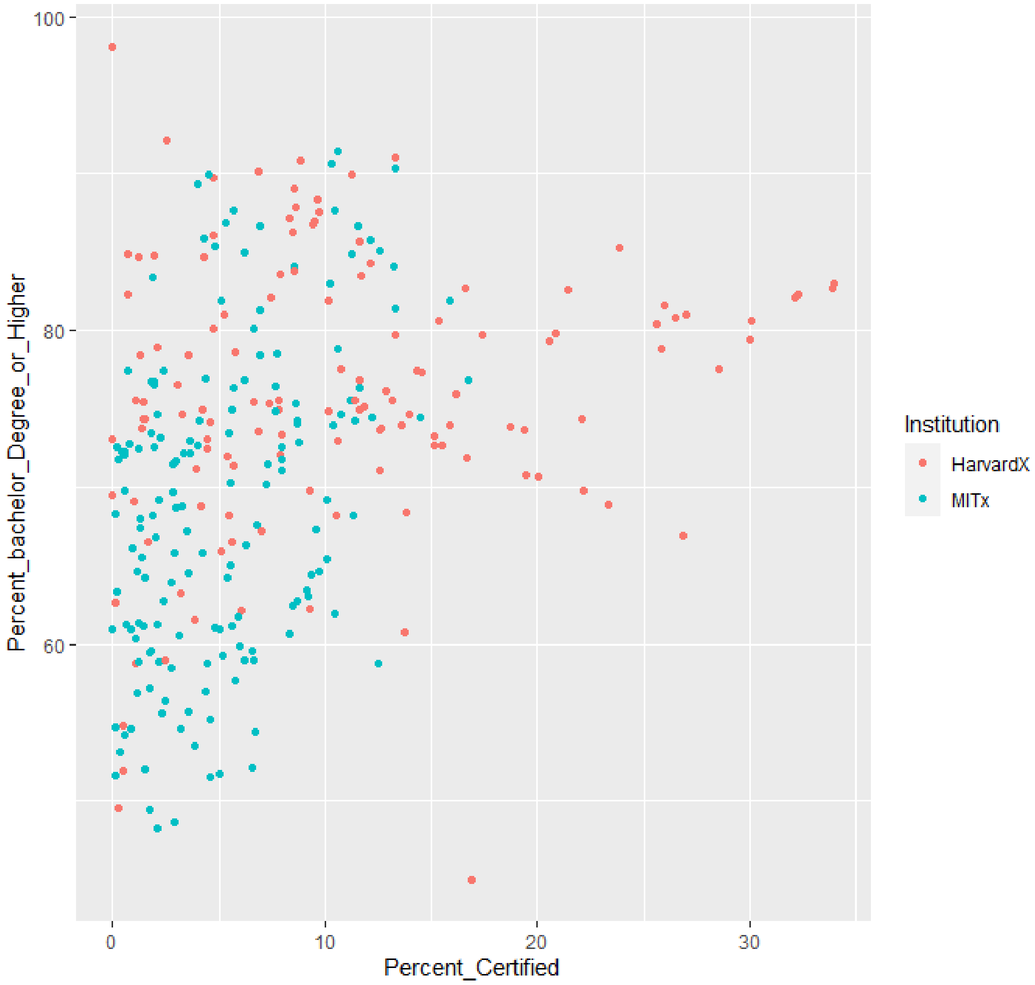
For this dataset,the first thing I was interested in comparing was the percent of students who have already obtained a bachelor's degree or higher education to the percent of students who were just certified and from which program they are attending. As you can see by the data, we have the orange markers representing Harvardx program and blue markers indicating MITx program. According to the data, we see a greater variability with Harvardx students who are certified. Most of our data is located on the left side of our chart indicating less students who are certified, however still have obtained a bachelor's degree.

Code:



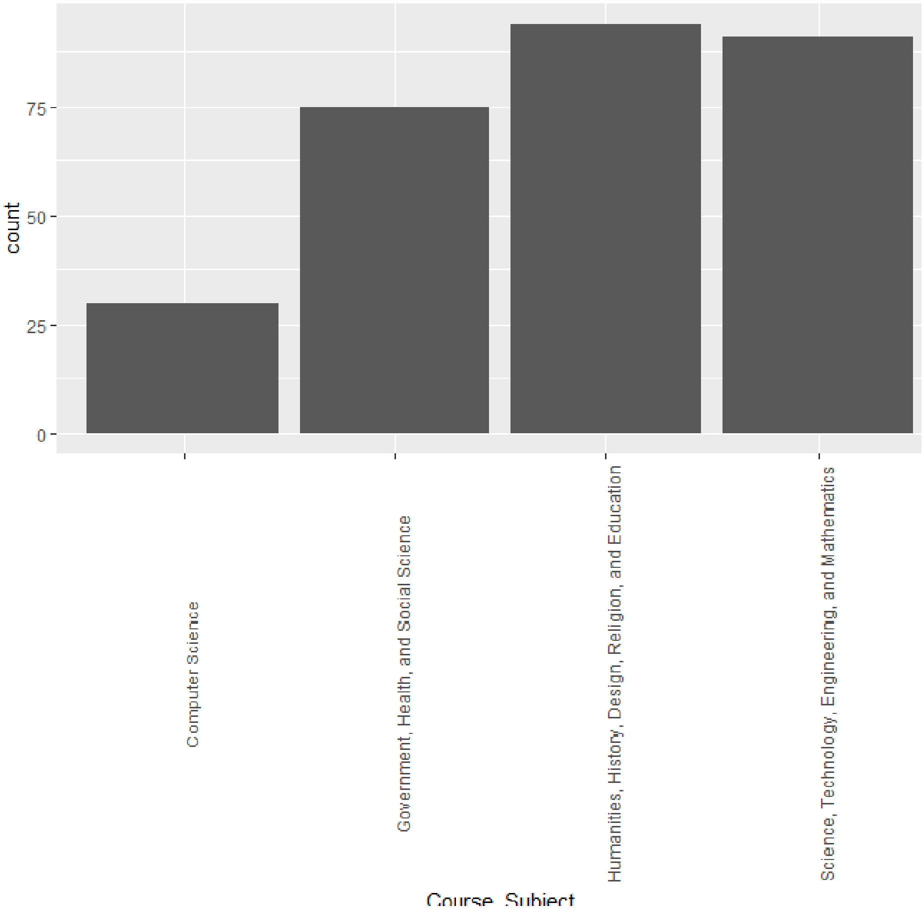
```
> dd <- ggplot(dataframe, aes(x=Percent_Certified, y=Percent_bachelor_Degree_or_Higher, color=Institution))
> dd + geom_point()
> |
```

Output:



Another interest of mine was simply which course subject had the most courses available. According to the data, Humanities, History, Design, Religion, and Education have the most available courses. To view this barchart, the following code was used:

```
ggplot(data=dataframe, mapping=aes(Course_Subject)) + geom_bar() + theme(axis.text.x =
element_text(angle = 90))
```



III. Summary

The analysis of this dataset has told us a few things. There are many other things that can be revealed if we were given an ample amount of time to do the research. I found it interesting and surprising that Computer Science or Engineering didn't have the most amount of courses offered. Another thing I also found surprising was how spread out the data for the point plot chart. I was expecting the data for Harvardx and MITx to be more similar. The data has shown us that there are more students who have a bachelors degree and a certificate who are attending the Harvardx program. I believe this data would be very useful in someone who wanted to know how many students who participate in these programs get certified and in which courses. It is also beneficial to see the amount of students we have participating in each course labeled by gender.