**Ishan Sharma**

**Is there a connection between hours spent in a course and completion of a course?**

**For Dr. Noukhovitch**

**ICS4UI**

**Table of Contents**

Introduction and Purpose 3

Hypothesis 3

Required Variables 3

Code 3

Raw data from PROC PRINT step 6

Graphs from PROC SGPLOT steps 8

Analysis 10

Conclusion 10

# Introduction and Purpose

Various courses have varying amount of time necessary for completion. For instance, according to the datasets given, in order to be an Electrician (Signal Maintenance), no in-class time whatsoever is required, while in order to be a Terrazzo, Tile and Marble Setter, 820 hours, or just over 34 days of non-stop studying, of in-class time are required. They have varying amounts of time necessary for completion, and require varying amounts of effort in order to be able to pass the course. As such, the hours spent within a course may have an effect on the completion of a course. Do the numbers of in-class hours and on-the-job hours play a role in completing an apprenticeship?

# Hypothesis

The required amounts of in-class and on-the-job hours is inversely proportional to the continuation rate of a course, and proportional to the discontinuation rate. Because courses with high amounts of time require high amounts of effort on the part of the apprentice to pass, courses that require a high amount of time will be too hard for many apprenticeships. As well, many apprentices will not have enough time in order to pass the course, and will be forced to drop it before completing it.

# Required Variables

The CSV source file named OCTAA, which contains the numbers of hours, both required for the courses, will need to be used in order to retrieve the variables named otj\_hours and in\_class\_hours.

This will need to be compared to the continuation and discontinuation rates of a course, which are named total\_cont and total\_disc. However, as this data is separated by year with many entries per course, in order to be compared to the other two variables, it’ll need to be accumulated into single variables for each course, and as the number of people enrolled in each course differs drastically and so the raw numbers of those continuing and discontinuing courses cannot be compared directly, they need to be converted into percentages of the total. The variables for the percentages are known as perc\_cont and perc\_disc.

# Code

/\*Ishan Sharma

March 9, 2017

for Dr. Noukhovitch

comparison\_analysis.sas

Compiling raw data into a readable sheet

Input: RAIS and OCTAA CSV files file

Output: Excel sheet of formatted data\*/

/\*Creating an excel file, with preferred style and sheet name\*/

ods excel file="/home/ishansharma30/summative/comparison.xlsx"

style=pearl

options(

sheet\_name="comparison"

);

/\*Converting the RAIS file into a readable dataset\*/

data rais;

infile "/home/ishansharma30/summative/RAIS\_dataset\_suppressed.csv" dsd delimiter=",";

input repyr reg\_status $ appr\_trade\_code $ appr\_trade\_name $ comp\_vol $ gender $ total\_reg mean\_age sd\_age total\_NA total\_cont total\_disc mean\_reg\_dur mean\_date\_reg;

if total\_cont=. then total\_cont=0;

if total\_disc=. then total\_disc=0;

run;

/\*Converting the OCTAA file into a readable dataset\*/

data octaa;

length appr\_trade\_name $ 200.;

infile "/home/ishansharma30/summative/OCTAA\_chart.csv" dsd delimiter=",";

input appr\_trade\_name $ appr\_trade\_code $ noc\_code appr\_sector $ appr\_ratio $ tax\_train\_cred $ fact\_sheets $ red\_seal $ cofq $ otj\_hours in\_class\_hours train\_std\_year curr\_std\_year trade\_board $ acad\_entry\_req;

if in\_class\_hours=. then in\_class\_hours=0;

if otj\_hours=. then otj\_hours=0;

run;

These steps first create the Excel file in which everything will be printed within, and then they convert the CSV files for RAIS and OCTAA into easily readable datasets, while also turning the missing points into zeroes so that further operations can be done with them in later steps without carrying over the missing data to those steps

/\*Sorting the datasets in order to create easily-mergeable sets\*/

proc sort data=rais;

by appr\_trade\_code;

run;

proc sort data=octaa;

by appr\_trade\_code;

run;

In preparation of the merge step that follows, the data sheets are simplified and sorted so that the merge can go as smoothly as possible.

/\*Merge the datasets, strip the unnecessary variables from the final data set, remove the records with unknown hours and/or students\*/

data comparison (keep=perc\_cont perc\_disc appr\_trade\_code in\_class\_hours otj\_hours);

merge rais octaa;

by appr\_trade\_code;

retain last\_cont 0;

last\_cont + total\_cont;

retain last\_disc 0;

last\_disc + total\_disc;

first\_code = first.appr\_trade\_code;

last\_code = last.appr\_trade\_code;

if first\_code=1 then last\_cont=total\_cont;

if first\_code=1 then last\_disc=total\_disc;

if last\_code=1;

if in\_class\_hours=. then in\_class\_hours=0;

if otj\_hours=. then otj\_hours=0;

if last\_cont=. then last\_cont=0;

if last\_disc=. then last\_disc=0;

perc\_cont=100\*last\_cont/(last\_cont + last\_disc);

perc\_disc=100\*last\_disc/(last\_cont + last\_disc);

if last\_cont > 0 or last\_disc > 0;

run;

Here, the data sheets are merged and bound together, and all unnecessarily variables are filtered out in order to simplify the printouts that follow. In addition, the RAIS sheet is simplified by using the retain statement for both its total\_cont and total\_disc variables, so those values can be added up for every single existing value, and these raw totals are easily converted to percentages. Finally, the values for which no data exists for the students that continue and leave an apprenticeship are removed.

/\*Print the dataset\*/

proc print data=comparison label split=" " noobs;

var appr\_trade\_code in\_class\_hours otj\_hours perc\_cont perc\_disc;

title "Tabular data of apprenticeship course";

footnote "By Ishan Sharma";

options nodate pageno=2;

label appr\_trade\_code="Program code" in\_class\_hours="In-class hours" otj\_hours="On-job hours" perc\_cont="Percentage continued" perc\_disc="Percentage discontinued";

run;

title "The in-class hours of an apprenticeship in relation to the percent of students who continue the course";

/\*Print the graphs of the data\*/

proc sgplot data=comparison;

scatter y=perc\_cont x=in\_class\_hours;

run;

title;

title "The in-job hours of an apprenticeship in relation to the percent of students who continue the course";

proc sgplot data=comparison;

scatter y=perc\_cont x=otj\_hours;

run;

title;

title "The in-class hours of an apprenticeship in relation to the percent of students who discontinue the course";

proc sgplot data=comparison;

scatter y=perc\_disc x=in\_class\_hours;

run;

title;

title "The in-job hours of an apprenticeship in relation to the percent of students who discontinue the course";

proc sgplot data=comparison;

scatter y=perc\_disc x=otj\_hours;

run;

title;

ods excel close;

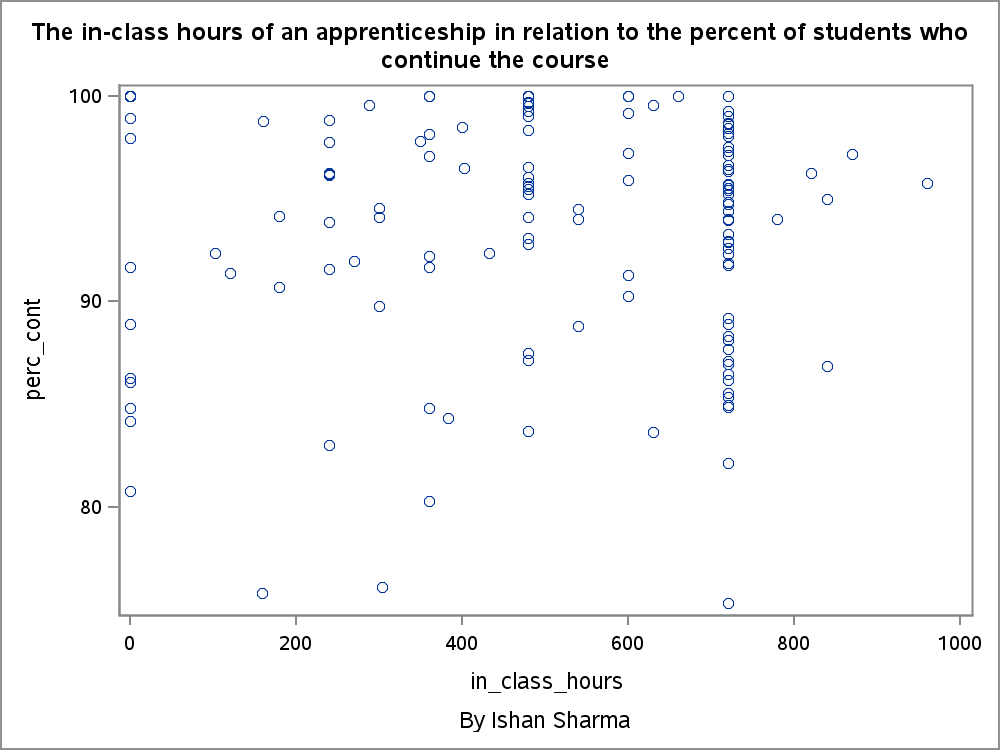
Finally, the table is printed out, with appropriate labels and footnotes, while a number of graphs are printed, each of which compare either the continuing or discontinuing rate, now adjusted for percentage, with either the hours spent in class or the hours spent on the job. As the number of hours spent is what, according to the hypothesis, determines whether an apprentice continues a course, it is therefore the independent variable for every graph, and in accordance to convention is therefore the x-axis for every graph

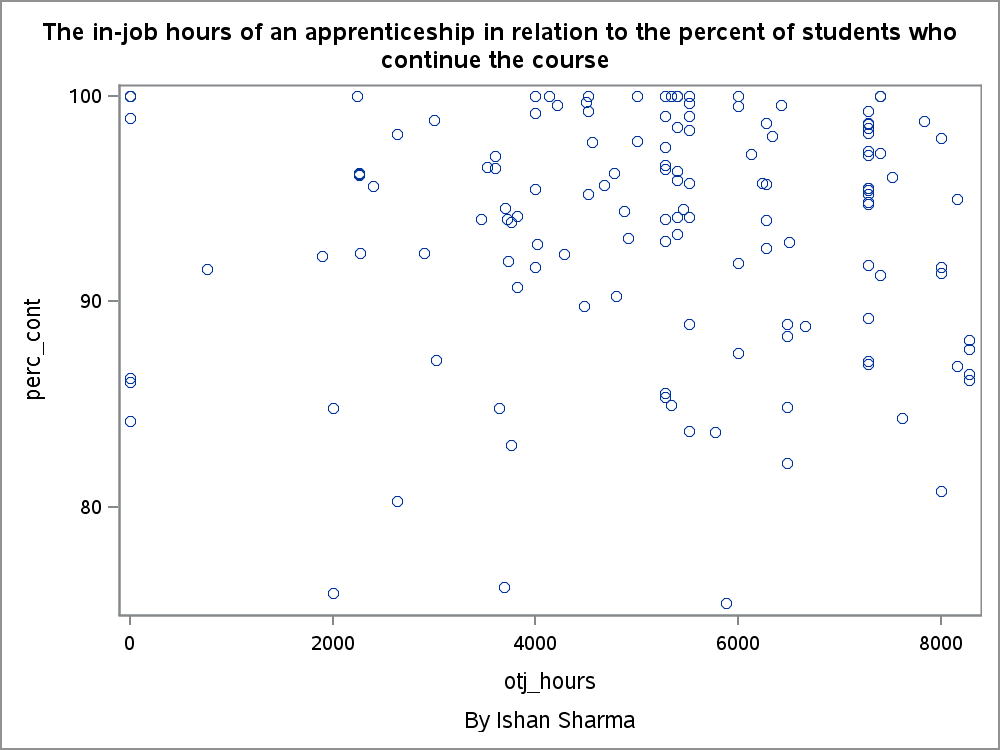
# Raw data from PROC PRINT step

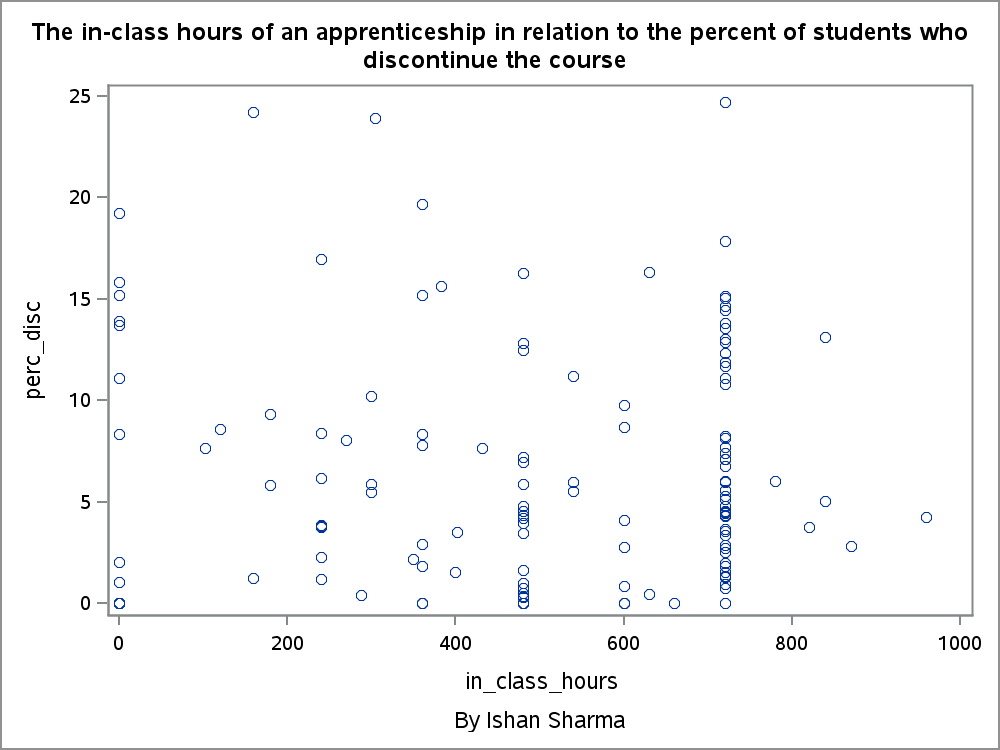
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Program code** | **In-class hours** | **On-job hours** | **Percentage continued** | **Percentage discontinued** |
| 200G | 720 | 7280 | 98.182 | 1.8182 |
| 207S | 120 | 8000 | 91.398 | 8.6022 |
| 211W | 0 | 8000 | 91.667 | 8.3333 |
| 225A | 0 | 8000 | 80.769 | 19.2308 |
| 237S | 480 | 5520 | 99.636 | .3636 |
| 237T | 480 | 5520 | 95.783 | 4.2169 |
| 240P | 540 | 5460 | 94.489 | 5.5110 |
| 241A | 820 | 4780 | 96.260 | 3.7402 |
| 244G | 480 | 4500 | 99.705 | .2950 |
| 244H | 0 | 6000 | 100.000 | .0000 |
| 244K | 0 | 5400 | 100.000 | .0000 |
| 245R | 540 | 3460 | 94.027 | 5.9730 |
| 246F | 720 | 6480 | 84.867 | 15.1333 |
| 246R | 720 | 6480 | 82.151 | 17.8487 |
| 246T | 300 | 4480 | 89.773 | 10.2273 |
| 253A | 630 | 5770 | 83.667 | 16.3327 |
| 253H | 102 | 2898 | 92.341 | 7.6591 |
| 255B | 720 | 7280 | 98.441 | 1.5595 |
| 255W | 480 | 5520 | 98.998 | 1.0020 |
| 259L | 480 | 6000 | 87.500 | 12.5000 |
| 263F | 360 | 4000 | 91.667 | 8.3333 |
| 268A | 0 | 0 | 100.000 | .0000 |
| 268R | 720 | 7280 | 95.532 | 4.4681 |
| 269E | 480 | 4520 | 100.000 | .0000 |
| 274L | 240 | 3760 | 93.846 | 6.1538 |
| 277Z | 600 | 7400 | 91.304 | 8.6957 |
| 278B | 360 | 2240 | 100.000 | .0000 |
| 282E | 720 | 6280 | 95.699 | 4.3011 |
| 289F | 0 | 0 | 84.163 | 15.8371 |
| 295A | 300 | 3700 | 94.545 | 5.4545 |
| 296A | 0 | 4000 | 100.000 | .0000 |
| 297B | 350 | 5000 | 97.814 | 2.1858 |
| 306A | 720 | 8280 | 86.193 | 13.8068 |
| 307A | 720 | 8280 | 86.456 | 13.5437 |
| 308A | 720 | 8280 | 87.667 | 12.3328 |
| 308R | 360 | 4140 | 100.000 | .0000 |
| 309A | 840 | 8160 | 86.878 | 13.1223 |
| 309C | 540 | 6660 | 88.823 | 11.1773 |
| 310B | 720 | 7280 | 95.206 | 4.7943 |
| 310D | 480 | 6000 | 99.509 | .4914 |
| 310E | 480 | 5400 | 100.000 | .0000 |
| 310G | 480 | 5520 | 94.108 | 5.8924 |
| 310J | 480 | 4000 | 95.455 | 4.5455 |
| 310K | 240 | 3000 | 98.810 | 1.1905 |
| 310Q | 480 | 5520 | 98.361 | 1.6393 |
| 310S | 720 | 6500 | 92.896 | 7.1043 |
| 310T | 720 | 6000 | 91.862 | 8.1381 |
| 313A | 720 | 8280 | 88.111 | 11.8885 |
| 313D | 480 | 4020 | 92.818 | 7.1819 |
| 332A | 480 | 3020 | 87.162 | 12.8382 |
| 339A | 480 | 5520 | 83.722 | 16.2781 |
| 339B | 360 | 2640 | 80.312 | 19.6884 |
| 339C | 240 | 760 | 91.597 | 8.4034 |
| 401A | 720 | 4880 | 94.418 | 5.5821 |
| 403A | 720 | 6480 | 88.912 | 11.0881 |
| 404C | 720 | 5280 | 85.552 | 14.4483 |
| 404D | 720 | 5280 | 92.918 | 7.0815 |
| 410K | 660 | 5340 | 100.000 | .0000 |
| 410N | 240 | 4560 | 97.748 | 2.2523 |
| 415A | 720 | 5280 | 96.633 | 3.3666 |
| 415B | 360 | 2640 | 98.167 | 1.8335 |
| 415C | 360 | 1900 | 92.208 | 7.7922 |
| 415D | 360 | 3600 | 97.087 | 2.9126 |
| 416E | 720 | 7280 | 98.639 | 1.3605 |
| 419A | 0 | 0 | 100.000 | .0000 |
| 420A | 720 | 5280 | 85.358 | 14.6417 |
| 421A | 720 | 6280 | 93.985 | 6.0145 |
| 421C | 720 | 5000 | 100.000 | .0000 |
| 423A | 720 | 5280 | 99.037 | .9627 |
| 423C | 870 | 6130 | 97.166 | 2.8338 |
| 424A | 720 | 7280 | 89.211 | 10.7892 |
| 425A | 720 | 6280 | 92.612 | 7.3880 |
| 426A | 720 | 7280 | 86.968 | 13.0316 |
| 427A | 720 | 6480 | 88.304 | 11.6955 |
| 428A | 720 | 5880 | 75.324 | 24.6762 |
| 429A | 720 | 7280 | 95.433 | 4.5670 |
| 430A | 720 | 7280 | 95.529 | 4.4707 |
| 430M | 720 | 7280 | 97.133 | 2.8674 |
| 431A | 720 | 7280 | 99.252 | .7481 |
| 433A | 720 | 7280 | 91.781 | 8.2186 |
| 434A | 383 | 7617 | 84.347 | 15.6533 |
| 435A | 480 | 4520 | 95.205 | 4.7955 |
| 435B | 720 | 4280 | 92.301 | 7.6987 |
| 437A | 720 | 4680 | 95.667 | 4.3334 |
| 438A | 720 | 7280 | 97.295 | 2.7049 |
| 441C | 720 | 5400 | 96.355 | 3.6446 |
| 442A | 840 | 8160 | 94.978 | 5.0225 |
| 443A | 720 | 7280 | 98.701 | 1.2987 |
| 444A | 720 | 5400 | 93.258 | 6.7416 |
| 444B | 720 | 5340 | 84.981 | 15.0193 |
| 445A | 720 | 5280 | 97.500 | 2.5000 |
| 446A | 480 | 7520 | 96.067 | 3.9326 |
| 447A | 720 | 7280 | 94.725 | 5.2749 |
| 448A | 600 | 5400 | 95.916 | 4.0842 |
| 449A | 480 | 3520 | 96.533 | 3.4670 |
| 450A | 480 | 2400 | 95.640 | 4.3605 |
| 451A | 600 | 4800 | 90.254 | 9.7464 |
| 452A | 360 | 3640 | 84.822 | 15.1782 |
| 453A | 480 | 4920 | 93.066 | 6.9336 |
| 455A | 400 | 5400 | 98.467 | 1.5326 |
| 456A | 720 | 5280 | 96.466 | 3.5335 |
| 456P | 480 | 5280 | 100.000 | .0000 |
| 600H | 402 | 3598 | 96.471 | 3.5294 |
| 605B | 480 | 4520 | 99.242 | .7576 |
| 606W | 288 | 4212 | 99.574 | .4255 |
| 609C | 160 | 7840 | 98.780 | 1.2195 |
| 610C | 0 | 0 | 98.945 | 1.0554 |
| 611B | 0 | 8000 | 97.959 | 2.0408 |
| 614A | 600 | 7400 | 100.000 | .0000 |
| 614B | 600 | 7400 | 97.222 | 2.7778 |
| 614C | 600 | 7400 | 100.000 | .0000 |
| 617A | 300 | 5400 | 94.118 | 5.8824 |
| 620A | 960 | 6240 | 95.771 | 4.2289 |
| 620C | 720 | 5280 | 94.009 | 5.9905 |
| 620D | 780 | 3720 | 94.003 | 5.9971 |
| 620E | 432 | 2268 | 92.358 | 7.6417 |
| 620G | 0 | 0 | 86.096 | 13.9037 |
| 630A | 304 | 3696 | 76.101 | 23.8994 |
| 630T | 720 | 7280 | 94.853 | 5.1471 |
| 631A | 600 | 4000 | 99.149 | .8508 |
| 634A | 270 | 3730 | 91.967 | 8.0333 |
| 634B | 630 | 6420 | 99.543 | .4570 |
| 634C | 720 | 6340 | 98.027 | 1.9729 |
| 634D | 180 | 3820 | 94.156 | 5.8436 |
| 634E | 180 | 3820 | 90.684 | 9.3161 |
| 636A | 240 | 2260 | 96.166 | 3.8339 |
| 636B | 240 | 2260 | 96.185 | 3.8155 |
| 636C | 240 | 2260 | 96.270 | 3.7304 |
| 636E | 720 | 7280 | 87.106 | 12.8936 |
| 636G | 0 | 0 | 100.000 | .0000 |
| 637C | 159 | 2000 | 75.817 | 24.1830 |
| 638A | 0 | 2000 | 84.815 | 15.1849 |
| 640D | 0 | 5520 | 88.925 | 11.0749 |
| 640S | 480 | 5520 | 100.000 | .0000 |
| 670C | 240 | 3760 | 83.041 | 16.9591 |
| 690H | 720 | 6280 | 98.680 | 1.3201 |
| 800A | 0 | 0 | 86.290 | 13.7102 |

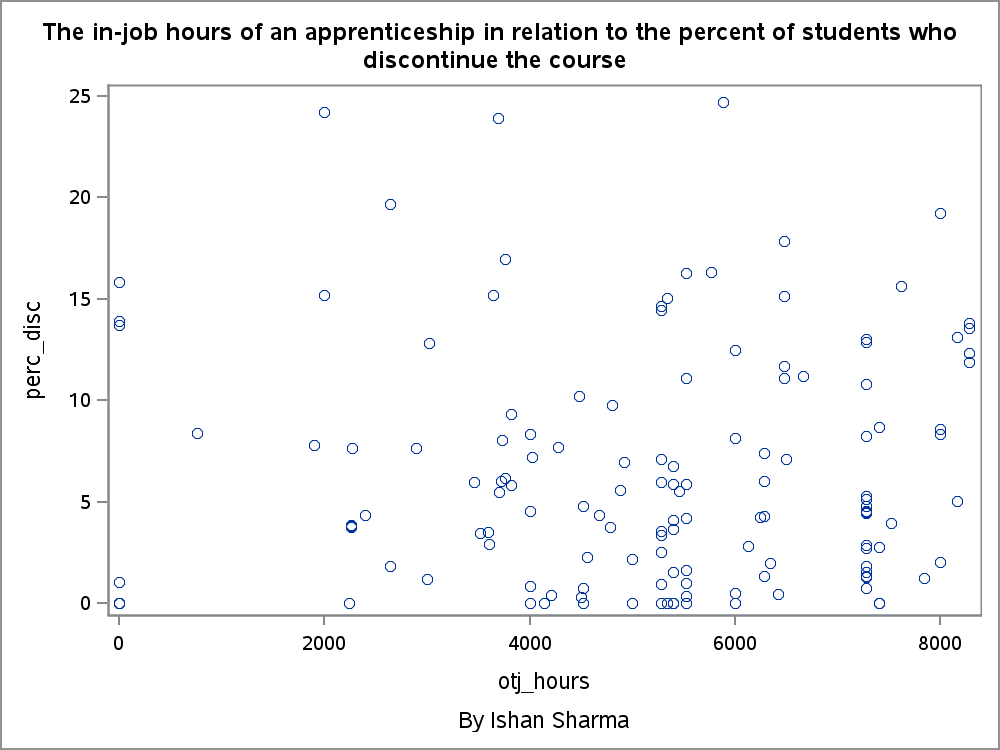
This raw data, however, is impossible to analyze alone. Another representation needs to exist in order to analyze it more easily.

# Graphs from PROC SGPLOT steps









# Analysis

Here, it is apparent that the points are all over the place in all four graphs. For many given numbers of hours, dots are spread just about everywhere in all of the graphs. For instance, for the in\_class\_hours value of 720, values exist everywhere from, in the graph that compare it to the continuing rate, points exist for values everywhere from 0 to 25%, reaching both the ceiling and the floor for that piece of data despite similar amounts of in-class time required for each course. Such circumstances are replicated in every other graph.

Indeed, the only trends that are visible in the above graphs is that the percentage of people who continue a course is on the higher end, while the percentage of people who discontinue the course is on the low end, and that the number of hours required, both on-the-job and in-class, is towards the higher end for most courses. However, for the hypothesis given, these trends are absolutely useless, and what is important is how the points increase or decrease.

Indeed, in this regard, the points are absolutely random and that no trend exists between hours spent and students that continue a course is clear. There could be a number of factors behind this. One reason could be that would-be apprentices are smart, and are able to determine which courses that require more effort and time than they want, and thus do not pick them. This, of course, would make most would-be apprentices good planners, able to determine whether a certain course is right for them. This may or may not be corroborated by other evidence from apprentice masters and from other reports. However, that falls out of the realm of this report. Another possibility is that, when an apprentice is required to put more time and effort into a job, they are willing to drop their other activities, as they feel that apprenticeships are sufficiently important to do that. This may mean that apprenticeships are well-organized and are fun enough that participants are willing to sacrifice their other activities, but that again falls outside the parameters of this report, and requires further research.

# Conclusion

Though the time taken for a course may seem to have an effect on the completion rate of that course, no relationship exists between the two at all in any way whatsoever. This may be caused by a variety of other factors, but what those factors are requires further research.