Starters and Finishers

Rachael Sanderson

6 November 2019

# **Who is completing the course?**

## *The Intended Analysis*

This initial analysis will reflect on the decline of participants within the course, to consider which steps have the steepest decline. With this analysis, ideally a further understanding will be gained about which stages of teaching attract the most participents, and so which elements need to be improved to maintain student participation and so overall cohort attainment.

## *The Applied Datasets*

The dataset used was the recorded “Step Activity” for each run of the course. For consistancy of the temporally based analysis, this was undertaken for runs 3-7 of the course, as these have identical step progressions through and so allow a more direct comparison.

## *Data Preparation*

The following function was applied to produce a dataset for analysis. It was applied to produce each dataframe needed within the report, with minor changes to extract the appropriate data.

*Function: Making a data frame of starters*

QuantitiesStarting.function = function(x){ #x = file selected to analyse  
 StepActivity = x  
 Steps=c((StepActivity$week\_number)+((StepActivity$step\_number)/100)) #acknowledging data complication (3.1 matches 3.10)  
 StepActivity = cbind(StepActivity, Steps) #editing the data with the new column  
 S= unique(StepActivity$Steps, incomparables = FALSE) #list the unique steps  
 L =length(S) #How many steps are there?  
   
#making the vector of quantities for each step  
   
 total\_values = vector() #making the vector  
 for(i in 1:L){ #for loop: L = how many steps will be assessed/length of vector  
 Step= StepActivity %>% filter(Steps==S[i]) #filter: group the rows for a single step  
 Quant = nrow(Step) #how many rows were collected for that step?  
 total\_values[i]= Quant} #compile a vector of how many people started each step  
  
#making the dataframe to be able to chart: Step against how many participants.  
   
DFActivity = data.frame(Step = S, Total = total\_values)  
   
 return(DFActivity)} #return: the final dataframe for further analysis

This function then counts how many people started each step, to produce a data frame that can then be evaluated. It can further be manipulated to identify the total finishers, or to divide the participants weekly, as will be demonstrated further.

## *Analysis*

### *Part One: How many are starting at each step?*

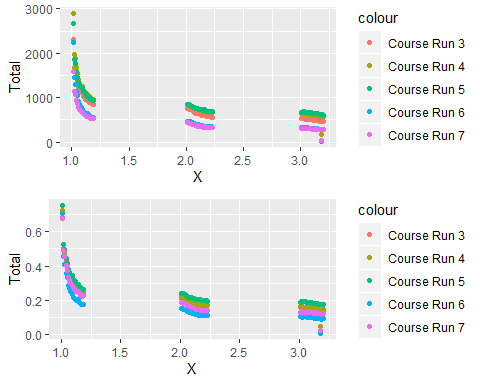
The first analysis was to record how many people are recorded as starting at each step, to reflect on how this declines. This was considered for each course, to consider if the trends were comparable for each run. With an awareness that the number of enrolments varies each year, there was a realisation that this data needs to be standardised to acknowledge this. Therefore, the data was represented as a proportion. Within this process, the code as above was repeated, with the y axis instead being the Y value / number of course starters.

The two can be modelled as below, with exact figures (left) and proportions (right):

Plot = ggplot (data =CourseS3,aes (x = X, y = Total))  
CourseS3Graph =Plot +geom\_point (aes (x= Step, y = CourseS3$Total, colour = "Course Run 3"))  
CourseS4Graph = CourseS3Graph + geom\_point (aes(x = Step, y = CourseS4$Total, colour = "Course Run 4"))  
CourseS5Graph = CourseS4Graph +geom\_point (aes (x= Step, y = CourseS5$Total, colour = "Course Run 5"))  
CourseS6Graph = CourseS5Graph + geom\_point (aes(x = Step, y = CourseS6$Total, colour = "Course Run 6"))  
CourseS7Graph = CourseS6Graph + geom\_point (aes(x = Step, y = CourseS7$Total, colour = "Course Run 7"))  
  
#Alternatively, a proportion graph can be formed as below;  
  
StartProp = ggplot (data = CourseS3,aes (x = X, y = Total))  
CourseSP3Graph =StartProp +geom\_point (aes (x= Step, y = CourseS3$Total/(Cohort\_Summaries$Entries[3]), colour = "Course Run 3"))  
CourseSP4Graph = CourseSP3Graph + geom\_point (aes(x = Step, y = CourseS4$Total/(Cohort\_Summaries$Entries[4]), colour = "Course Run 4")) #.....

The two can be modelled as below, with exact figures (left) and proportions (right):

grid.arrange(CourseS7Graph, CourseSP7Graph)



When comparing these graphs, the trends show limited differences. The trends show a gradual decline of starters at each step. DISTRIBUTION OF DECLINE? HYPOTHESIS TEST? NEED TO NEATEN GRAPHS

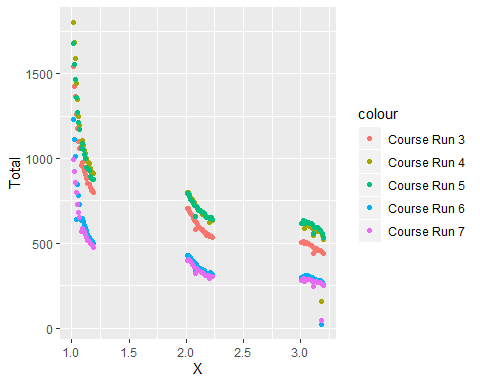
### *Part Two: How many are finishing at each step?*

When reflecting on analysis that focused on the starters, it was realised that there is an additional column of ‘last\_completed\_at’. The assumption was made that when this ‘completed’ column is empty, it refers to the people who *did not finish a step* and so would have not completed the course, despite starting that section. If a step has a lower number of finishers, it is potential responsible for a higher number of dropouts - something that the educational programme wants to reduce. The data frame of finishers was produced in the same way as for the starters, with the addition of a single line of code at the start:

StepActivity = cyber.security.7.step.activity  
 Filter = StepActivity %>% filter(last\_completed\_at !="")  
 StepActivity=data.frame(Filter)

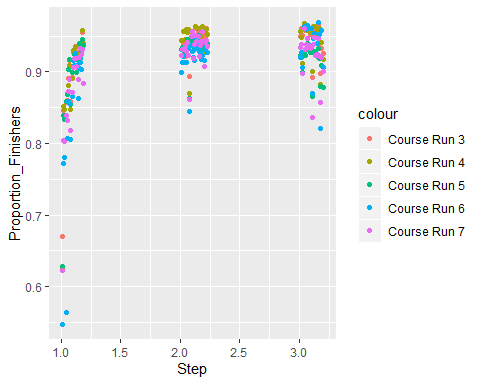
This code removes the rows that are empty in the “last\_completed\_at” to produce a data sample that only contains the people that are recorded as finishing each stage. The following graph was hence produced:

Finishers = ggplot (data =CourseF3,aes (x = X, y = Total))  
#CourseF1Graph = Plot +geom\_point (aes (x= Step, y = CourseF1$Total, colour = "Course Run 1"))  
#CourseF2Graph = Course1Graph + geom\_point (aes(x = Step, y = CourseF2$Total, colour = "Course Run 2"))  
CourseF3Graph =Finishers +geom\_point (aes (x= Step, y = CourseF3$Total, colour = "Course Run 3"))  
CourseF4Graph = CourseF3Graph + geom\_point (aes(x = Step, y = CourseF4$Total, colour = "Course Run 4"))  
CourseF5Graph = CourseF4Graph +geom\_point (aes (x= Step, y = CourseF5$Total, colour = "Course Run 5"))  
CourseF6Graph = CourseF5Graph + geom\_point (aes(x = Step, y = CourseF6$Total, colour = "Course Run 6"))  
CourseF7Graph = CourseF6Graph + geom\_point (aes(x = Step, y = CourseF7$Total, colour = "Course Run 7"))  
  
CourseF7Graph

 Once again, it is acknowledged that the exact number of starters is a poor statistic to use for comparison of decline, as certain courses may have started with less entrants. Therefore, the Y axis was adapted to represent the number of finishers as a proportion of the number of starters;

FinishProp = ggplot (data = CourseF3,aes (x = Step, y = Proportion\_Finishers))  
CourseFP3Graph = FinishProp + geom\_point (aes (x= Step, y = CourseF3$Total/CourseS3$Total, colour = "Course Run 3"))

The final graph produced can be seen below;



The graph shows limited correlation for each week, with each grouping being fairly clustered. There does not seem to be a particular *step* with a higher proportion of non-finishers throughout every run of the course, although noticably the cluster for stage 1 has a more linear trend than stages 2 and 3, implying that more people tend to drop out of the course as they are introduced to the content in the earlier stages, with the cohort settling as they are embedded in the course. In reflecting on the difference in the chart when the proportion of finishers is considered, the analysis led to the consideration of both starters and finishers.

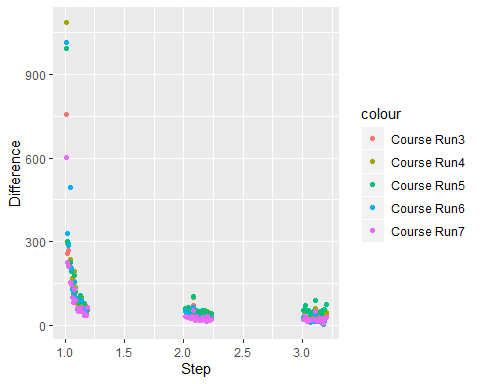
### *Part Three: Contrasting Starters and Finishers*

The analysis focused on the difference between the two columns, and so began with the following data frame;

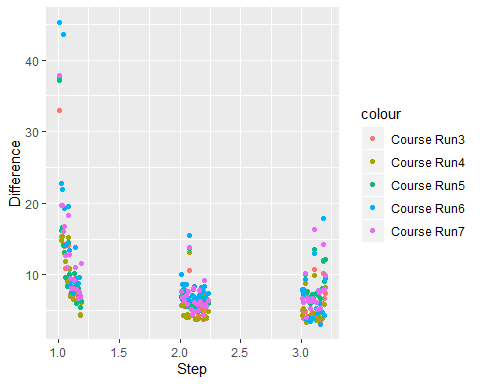
DifferenceDF = data.frame(Step = CourseS3$Step,   
 Difference3 = (CourseS3$Total-CourseF3$Total),  
 Difference4 = (CourseS4$Total-CourseF4$Total),  
 Difference5 = (CourseS5$Total-CourseF5$Total),  
 Difference6 = (CourseS6$Total-CourseF6$Total),  
 Difference7 = (CourseS7$Total-CourseF7$Total))

To plot the following graph;

D= ggplot (data=DifferenceDF, aes (x=Step, y=Difference))  
D3 = D + geom\_point(aes(x=Step, y= Difference3, colour = "Course Run3"))  
D4 = D3 + geom\_point(aes(x=Step, y= Difference4, colour = "Course Run4"))  
D5 = D4 + geom\_point(aes(x=Step, y= Difference5, colour = "Course Run5"))  
D6 = D5 + geom\_point(aes(x=Step, y= Difference6, colour = "Course Run6"))  
D7 = D6 + geom\_point(aes(x=Step, y= Difference7, colour = "Course Run7"))  
  
D7



To adjust the graph, the difference between the two was calculated as a % of the number of starters:



The graph of the percentages reflects a similar scatter to the number of finishers, however as a visualisation it is more effective. The difference calculated is the number of people that do not finish a step: this is the quantity that the educational provider wishes to reduce. However, in this graph the higher values are the ones that demand attention as the steps with the largest proportion of students dropping out. Therefore, as a final model to reflect on the proportion of students that do not complete each step, it is the most useful.

### *Discrepencies: What does it mean to “Finish”?*

Returning to the finishers graph (see below), an anomoly was revealed, as certain steps noticably have less finishers than a step after it, implying that ‘not completing’ a step may not mean completing the course; this was investigated further.

