EDA\_Project Report 3

Stanly Palathingal

17 November 2019

Having seen the weekly response, it is good to see in which part of the week the candidates leave. Every week there are various sections or step\_number indicating some modules. So the attempt is made to see how many candidates go through the steps or modules.

#install.packages("ProjectTemplate")  
#install.packages("rmarkdown")  
#install.packages("dplyr")

library(ProjectTemplate)  
library(rmarkdown)  
library(dplyr)  
library(ggplot2)  
# create.project("EDA\_Project")

To analyze, step activity files of all the seven years are loaded.

step\_act\_7= read.csv("data/cyber-security-7\_step-activity.csv")  
step\_act\_6= read.csv("data/cyber-security-6\_step-activity.csv")  
step\_act\_5= read.csv("data/cyber-security-5\_step-activity.csv")  
step\_act\_4= read.csv("data/cyber-security-4\_step-activity.csv")  
step\_act\_3= read.csv("data/cyber-security-3\_step-activity.csv")  
step\_act\_2= read.csv("data/cyber-security-2\_step-activity.csv")  
step\_act\_1= read.csv("data/cyber-security-1\_step-activity.csv")

Group by command from the dplyr package is used to extract the data and they are plotted using a line graph. It can be seen that the one value falls to the lowest with a steep decline in week 3

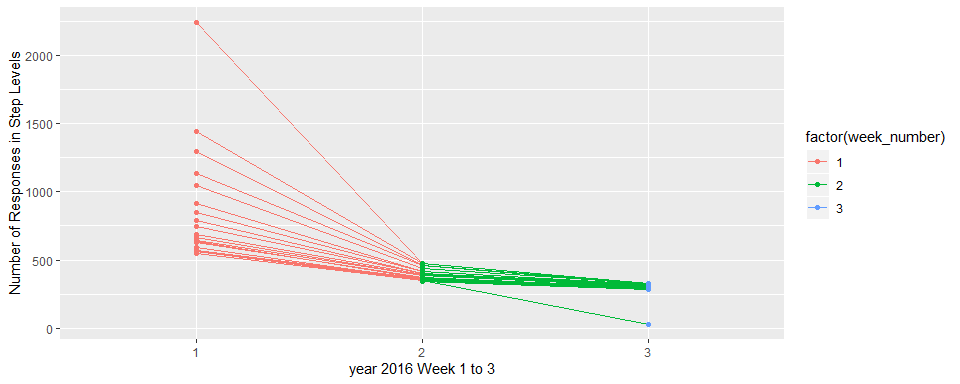
step\_7= step\_act\_7 %>% group\_by(week\_number, step\_number) %>% count()  
head(step\_7)

## # A tibble: 6 x 3  
## # Groups: week\_number, step\_number [6]  
## week\_number step\_number n  
## <int> <int> <int>  
## 1 1 1 1594  
## 2 1 2 1147  
## 3 1 3 1068  
## 4 1 4 949  
## 5 1 5 873  
## 6 1 6 782

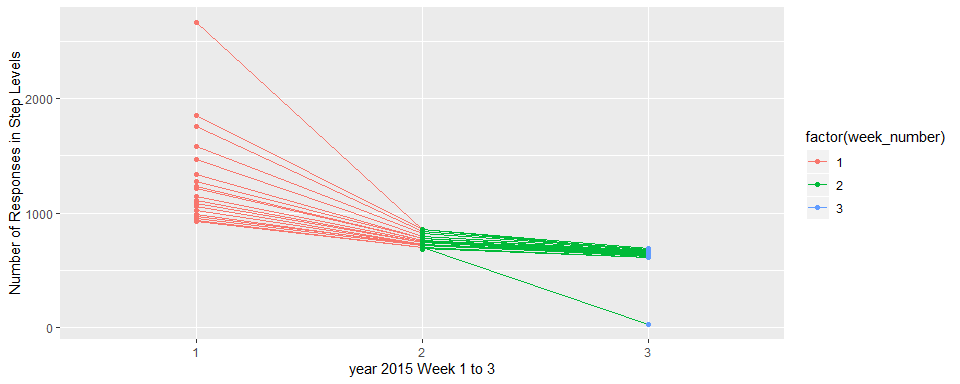
ggplot(data=step\_7,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2017 Week 1 to 3") + ylab("Number of Responses in Step Levels")



step\_6= step\_act\_6 %>% group\_by(week\_number, step\_number) %>% count()  
  
ggplot(data=step\_6,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2016 Week 1 to 3") + ylab("Number of Responses in Step Levels")



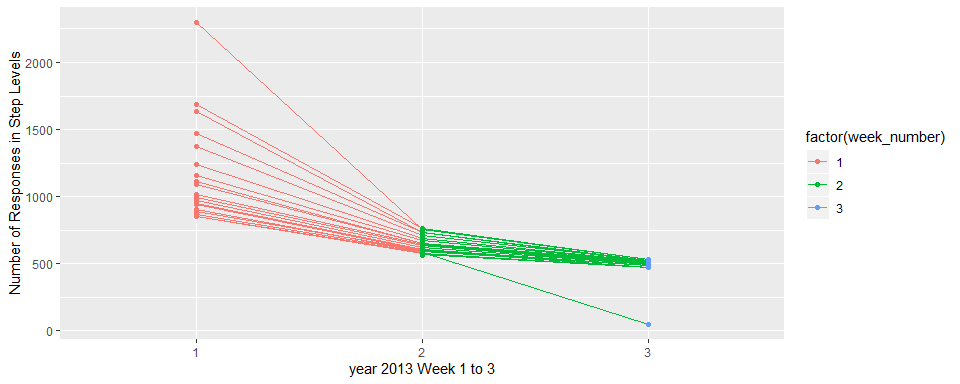
step\_5= step\_act\_5 %>% group\_by(week\_number, step\_number) %>% count()  
  
ggplot(data=step\_5,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2015 Week 1 to 3") + ylab("Number of Responses in Step Levels")



step\_4= step\_act\_4 %>% group\_by(week\_number, step\_number) %>% count()  
  
ggplot(data=step\_4,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2014 Week 1 to 3") + ylab("Number of Responses in Step Levels")



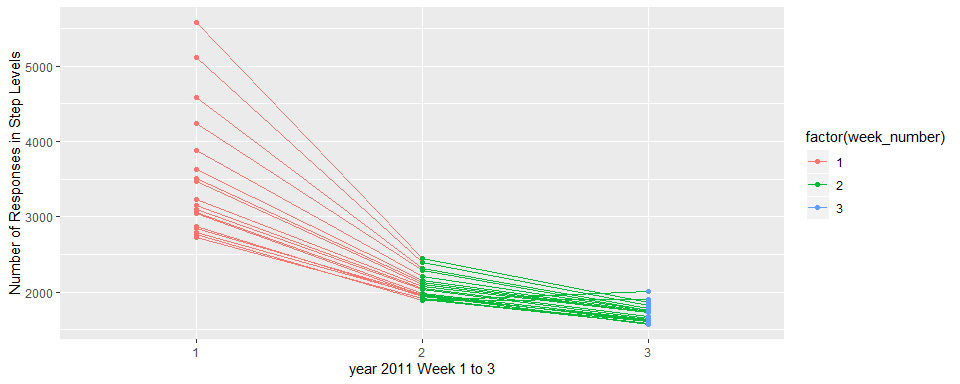
step\_3= step\_act\_3 %>% group\_by(week\_number, step\_number) %>% count()  
  
ggplot(data=step\_3,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2013 Week 1 to 3") + ylab("Number of Responses in Step Levels")



step\_2= step\_act\_2 %>% group\_by(week\_number, step\_number) %>% count()  
  
ggplot(data=step\_2,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2012 Week 1 to 3") + ylab("Number of Responses in Step Levels")



step\_1= step\_act\_1 %>% group\_by(week\_number, step\_number) %>% count()  
  
ggplot(data=step\_1,aes(x=factor(week\_number),y=n,group =step\_number)) + geom\_line(aes(color =factor(week\_number))) + geom\_point(aes(color = factor(week\_number))) + xlab("year 2011 Week 1 to 3") + ylab("Number of Responses in Step Levels")

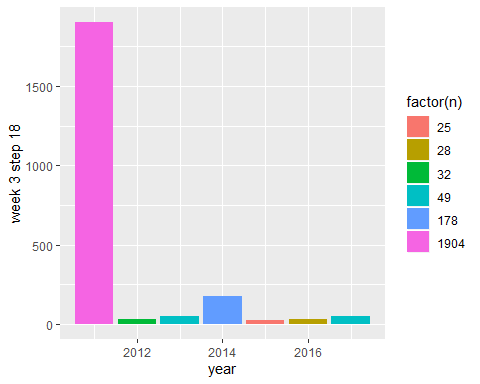


From the line plot, except in year 2011 all the other years there is a sudden fall of value in week 3. The analysis made us understand that it is in week 3 step 18. Those values are taken separately and then plotted as below.

step\_7\_1=step\_act\_7 %>% select(learner\_id) %>% filter(step\_act\_7$week\_number==3 & step\_act\_7$step\_number==18)  
st7=data.frame("year"= 2017,"n"= length(step\_7\_1$learner\_id))  
  
step\_6\_1=step\_act\_6 %>% select(learner\_id) %>% filter(step\_act\_6$week\_number==3 & step\_act\_6$step\_number==18)  
st6=data.frame("year"= 2016,"n"= length(step\_6\_1$learner\_id))  
  
step\_5\_1=step\_act\_5 %>% select(learner\_id) %>% filter(step\_act\_5$week\_number==3 & step\_act\_5$step\_number==18)  
st5=data.frame("year"= 2015,"n"= length(step\_5\_1$learner\_id))  
  
step\_4\_1=step\_act\_4 %>% select(learner\_id) %>% filter(step\_act\_4$week\_number==3 & step\_act\_4$step\_number==18)  
st4=data.frame("year"= 2014,"n"= length(step\_4\_1$learner\_id))  
  
step\_3\_1=step\_act\_3 %>% select(learner\_id) %>% filter(step\_act\_3$week\_number==3 & step\_act\_3$step\_number==18)  
st3=data.frame("year"= 2013,"n"= length(step\_3\_1$learner\_id))  
  
step\_2\_1=step\_act\_2 %>% select(learner\_id) %>% filter(step\_act\_2$week\_number==3 & step\_act\_2$step\_number==18)  
st2=data.frame("year"= 2012,"n"= length(step\_2\_1$learner\_id))  
  
step\_1\_1=step\_act\_1 %>% select(learner\_id) %>% filter(step\_act\_1$week\_number==3 & step\_act\_1$step\_number==18)  
st1=data.frame("year"= 2011,"n"= length(step\_1\_1$learner\_id))  
  
(step\_w3\_st18 = rbind(st1,st2,st3,st4,st5,st6,st7))

## year n  
## 1 2011 1904  
## 2 2012 32  
## 3 2013 49  
## 4 2014 178  
## 5 2015 25  
## 6 2016 28  
## 7 2017 49

ggplot(data=step\_w3\_st18) + geom\_bar(aes(x=year,y= n,fill=factor(n)), stat = "identity") +xlab("year")+ ylab("week 3 step 18")



Comparing these with the graph of candidates attending week 3, there is very huge difference. In week 3 year 2017 had 260 candidates and only 49 of them did step 18. By looking into the course content it will be evident that week 3 step 18 is the ‘test’ module and many did not go through it.