EDA_Project Report 2

Stanly Wilson Palathingal

12/11/2019

This report makes an attempt to look into the enrollment and the course completion. The enrollment file is taken to see how may are enrolled for the course. Also the question response file is taken to see the number of people starting the course and completing the course.

```
library(ProjectTemplate)
library(rmarkdown)
library(dplyr)
library(ggplot2)
# create.project("EDA_Project")
```

```
load.project()
```

The question response files of all the seven years are loaded and assigned.

```
q_response_7= read.csv("data/cyber-security-7_question-response.csv")
q_response_6= read.csv("data/cyber-security-6_question-response.csv")
q_response_5= read.csv("data/cyber-security-5_question-response.csv")
q_response_4= read.csv("data/cyber-security-4_question-response.csv")
q_response_3= read.csv("data/cyber-security-3_question-response.csv")
q_response_2= read.csv("data/cyber-security-2_question-response.csv")
q_response_1= read.csv("data/cyber-security-1_question-response.csv")
```

The enrollment files of all the seven years are loaded and assigned.

```
enrollment_7= read.csv("data/cyber-security-7_enrolments.csv")
enrollment_6= read.csv("data/cyber-security-6_enrolments.csv")
enrollment_5= read.csv("data/cyber-security-5_enrolments.csv")
enrollment_4= read.csv("data/cyber-security-4_enrolments.csv")
enrollment_3= read.csv("data/cyber-security-3_enrolments.csv")
enrollment_2= read.csv("data/cyber-security-2_enrolments.csv")
enrollment_1= read.csv("data/cyber-security-1_enrolments.csv")
```

For the analysis a data frame is created using the length of unique learners id which uniquely identifies a candidate. They are binded to another data frame for the analysis.

```
e_7_id=data.frame("year"= 2017,"enrolled"= length(unique(enrollment_7$learner_id)))
e_6_id=data.frame("year"= 2016,"enrolled"= length(unique(enrollment_6$learner_id)))
e_5_id=data.frame("year"= 2015,"enrolled"= length(unique(enrollment_5$learner_id)))
e_4_id=data.frame("year"= 2014,"enrolled"= length(unique(enrollment_4$learner_id)))
e_3_id=data.frame("year"= 2013,"enrolled"= length(unique(enrollment_3$learner_id)))
e_2_id=data.frame("year"= 2012,"enrolled"= length(unique(enrollment_2$learner_id)))
e_1_id=data.frame("year"= 2011,"enrolled"= length(unique(enrollment_1$learner_id)))

enrolled=rbind(e_1_id,e_2_id,e_3_id,e_4_id,e_5_id,e_6_id,e_7_id)
```

From the question response, data frame is created having weeks 1 to three and they are assigned to another data frame. This process is done for all seven years

```
q7_week1 = filter(q_response_7,q_response_7$week_number==1)
q7_week2 = filter(q_response_7,q_response_7$week_number==2)
q7_week3 = filter(q_response_7,q_response_7$week_number==3)
```

```
q7_1=data.frame("year"= 2017,"week"="week1","week_n"= length(unique(q7_week1$learner_id)))
q7_2=data.frame("year"= 2017,"week"="week2","week_n"= length(unique(q7_week2$learner_id)))
q7_3=data.frame("year"= 2017,"week"="week3","week_n"= length(unique(q7_week3$learner_id)))
```

```
q6_week1 = filter(q_response_6,q_response_6$week_number==1)
q6_week2 = filter(q_response_6,q_response_6$week_number==2)
q6_week3 = filter(q_response_6,q_response_6$week_number==3)

q6_1=data.frame("year"= 2016,"week"="week1","week_n"=length(unique(q6_week1$learner_id)))
q6_2=data.frame("year"= 2016,"week"="week2","week_n"=length(unique(q6_week2$learner_id)))
q6_3=data.frame("year"= 2016,"week"="week3","week_n"=  length(unique(q6_week3$learner_id)))
```

```
q5_week1 = filter(q_response_5,q_response_5$week_number==1)
q5_week2 = filter(q_response_5,q_response_5$week_number==2)
q5_week3 = filter(q_response_5,q_response_5$week_number==3)

q5_1=data.frame("year"= 2015,"week"="week1","week_n"=  length(unique(q5_week1$learner_id)))
q5_2=data.frame("year"= 2015,"week"="week2","week_n"=  length(unique(q5_week2$learner_id)))
q5_3=data.frame("year"= 2015,"week"="week3","week_n"= length(unique(q5_week3$learner_id)))
```

```
q4_week1 = filter(q_response_4,q_response_4$week_number==1)
q4_week2 = filter(q_response_4,q_response_4$week_number==2)
q4_week3 = filter(q_response_4,q_response_4$week_number==3)

q4_1=data.frame("year"= 2014,"week"="week1","week_n"=  length(unique(q4_week1$learner_id)))
q4_2=data.frame("year"= 2014,"week"="week2","week_n"=  length(unique(q4_week2$learner_id)))
q4_3=data.frame("year"= 2014,"week"="week3","week_n"=  length(unique(q4_week3$learner_id)))
```

```
q3_week1 = filter(q_response_3,q_response_3$week_number==1)
q3_week2 = filter(q_response_3,q_response_3$week_number==2)
q3_week3 = filter(q_response_3,q_response_3$week_number==3)

q3_1=data.frame("year"= 2013,"week"="week1","week_n"=  length(unique(q3_week1$learner_id)))
q3_2=data.frame("year"= 2013,"week"="week2","week_n"=  length(unique(q3_week2$learner_id)))
q3_3=data.frame("year"= 2013,"week"="week3","week_n"=  length(unique(q3_week3$learner_id)))
```

```
q2_week1 = filter(q_response_2,q_response_2$week_number==1)
q2_week2 = filter(q_response_2,q_response_2$week_number==2)
q2_week3 = filter(q_response_2,q_response_2$week_number==3)

q2_1=data.frame("year"= 2012,"week"="week1","week_n"=  length(unique(q2_week1$learner_id)))
q2_2=data.frame("year"= 2012,"week"="week2","week_n"=  length(unique(q2_week2$learner_id)))
q2_3=data.frame("year"= 2012,"week"="week3","week_n"=  length(unique(q2_week3$learner_id)))
```

```
q1_week1 = filter(q_response_1,q_response_1$week_number==1)
q1_week2 = filter(q_response_1,q_response_1$week_number==2)
q1_week3 = filter(q_response_1,q_response_1$week_number==3)

q1_1=data.frame("year"= 2011,"week"="week1","week_n"=  length(unique(q1_week1$learner_id)))
q1_2=data.frame("year"= 2011,"week"="week2","week_n"=  length(unique(q1_week2$learner_id)))
q1_3=data.frame("year"= 2011,"week"="week3","week_n"=  length(unique(q1_week3$learner_id)))
```

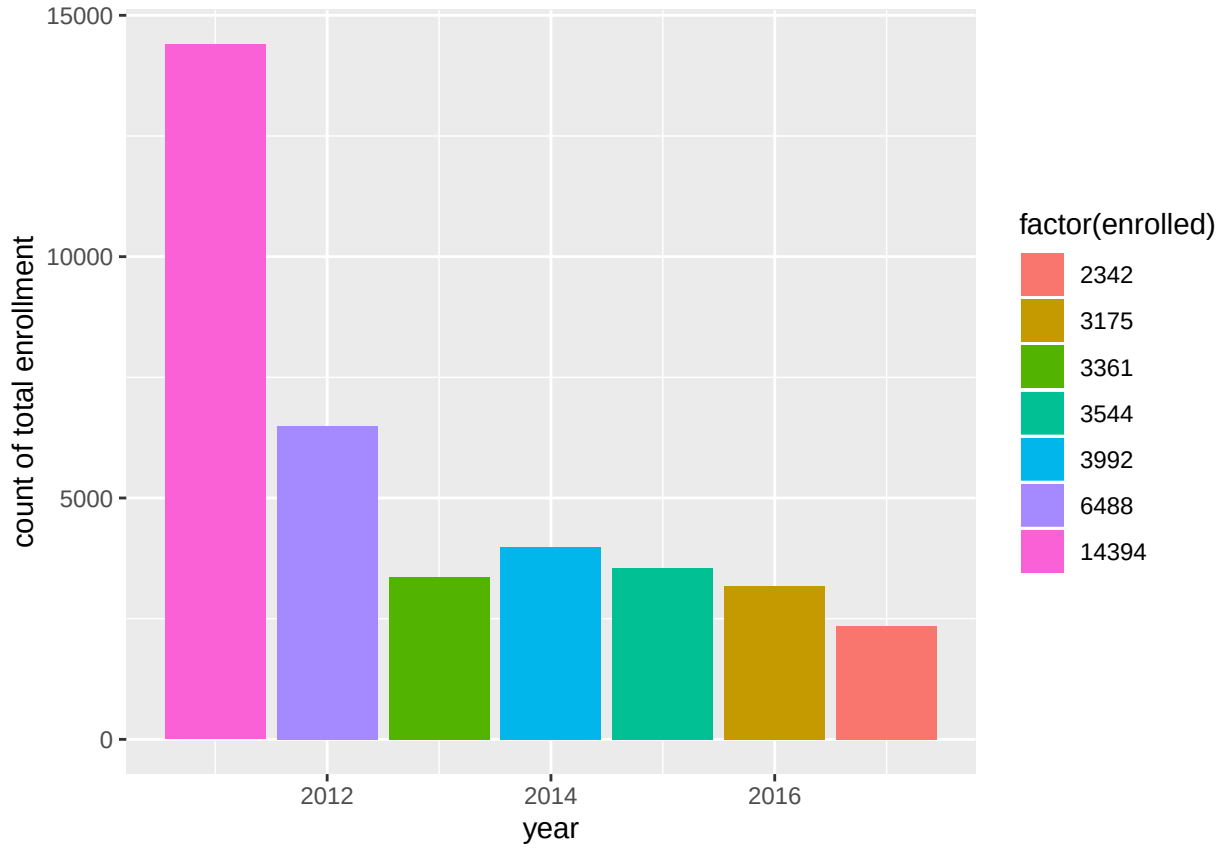
The data frames are combined here and made into one.

```
q=rbind(q1_1,q1_2,q1_3,q2_1,q2_2,q2_3,q3_1,q3_2,q3_3,q4_1,q4_2,q4_3,q5_1,q5_2,q5_3,q6_1
        ,q6_2,q6_3,q7_1,q7_2,q7_3)
head(q)
```

```
##   year  week week_n
## 1 2011 week1   3269
## 2 2011 week2   2022
## 3 2011 week3   1711
## 4 2012 week1   1342
## 5 2012 week2    889
## 6 2012 week3    668
```

Two bar graphs are generated here. One is that depicts the enrollment of each year.

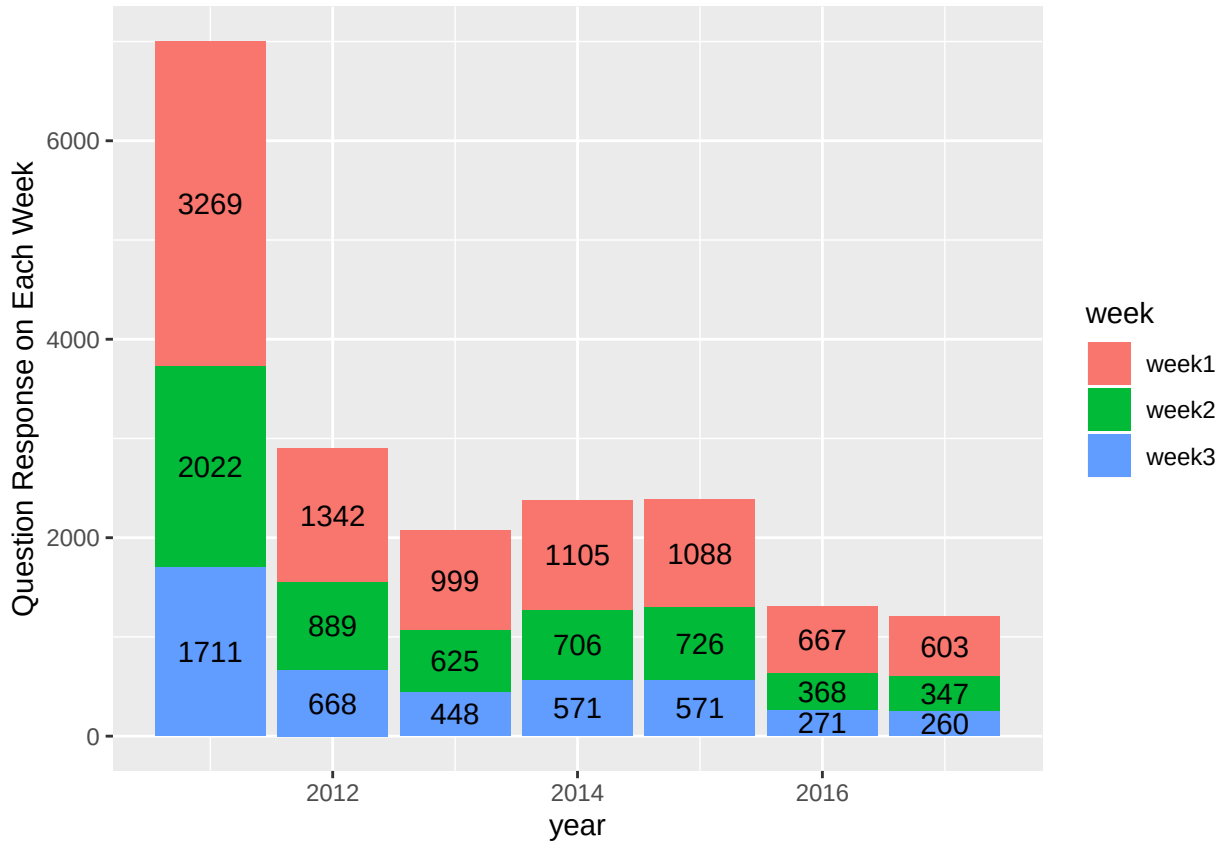
```
ggplot(data=enrolled) +
  geom_bar(aes(x=year,y= enrolled,fill=factor(enrolled)), stat = "identity") +
  xlab("year") + ylab("count of total enrollment")
```



From the plot is seen that in the first year there were 14394 candidates enrolled and in 2012 it is reduced by more than 50% having a count of 6488. In the following years, it declined further and reached 2342 in 2017.

The graph below is generated from the question response of seven years. It is an attempt to see how many actually start the course and reach the last week

```
ggplot(data=q,aes(x=year,y= week_n,fill=week)) +
  geom_bar(position = "stack", stat = "identity") +
  geom_text(aes(y=week_n, label = week_n),position = position_stack(vjust = 0.5)) +
  xlab("year") + ylab("Question Response on Each Week")
```



Comparing the two graphs it can be inferred that all those who are enrolled do not start the actual course or give a question response. In year 2011 there were 14394 candidates enrolled while only 3269 have started the course and only half the same reached till week 3. Similarly, 2342 have enrolled in 2017 and only 603 have started in week 1 and only obe third completed the last week.