CSIRSP_DM_1

Business Understanding

SMART principal

Specific

Look at the data overview, understand the structure and meaning of the data, and prepare for the next cycle

Measurable

We have data from 7 runs of a massive open online course (MOOC) entitled "Cyber Security: Safety At Home, Online, and in Life" made by Newcastle University and made available to the public by the online skills provider FutureLearn We have raw data collected by FutureLearn on leaners as they progressed through the course, along with some characteristic information collected from their profiles.

Attainable

We will conduct data set "cyber-security-1_archetype-survey-responses" through R language.

Relevant

There are always more or less relationships between data, and in this loop we will observe the distribution of data, the general situation.

Time-bouned

We will present our analysis report within 7 days

Data Understanding

1.load data set

```
data1 <- read.csv(file = 'cyber-security-1_enrolments.csv')
head(data1)
```

```
##                             learner_id             enrolled_at
## 1 160d6600-ea0e-4568-bfa9-5d7cd5b8e61b 2016-08-10 14:28:49 UTC
## 2 4dc22fed-63d4-4bf6-b162-bdf482e1ec38 2016-05-24 17:34:34 UTC
## 3 ecdd37db-0c75-496e-bff2-230553d0e38c 2016-05-19 00:52:38 UTC
## 4 988964c9-7410-40cc-addf-441f93e7a8b8 2016-05-19 21:40:01 UTC
## 5 f1493366-17a1-41b8-9de3-fbaad9d811d4 2016-09-19 15:35:35 UTC
## 6 25cc3b46-a955-4e2a-a71f-6b2025cc2787 2016-08-30 04:16:43 UTC
##             unenrolled_at    role   fully_participated_at
## 1                         learner
## 2 2018-10-30 20:20:51 UTC learner
## 3                         learner 2016-09-22 16:56:03 UTC
## 4                         learner
## 5                         learner
## 6                         learner 2016-10-25 12:44:14 UTC
##   purchased_statement_at  gender country age_range highest_education_level
## 1                       Unknown Unknown   Unknown                 Unknown
```

```
## 2                          male      PE     46-55       university_degree
## 3                       Unknown Unknown   Unknown                  Unknown
## 4                       Unknown Unknown   Unknown                  Unknown
## 5                       Unknown Unknown   Unknown                  Unknown
## 6                       Unknown Unknown   Unknown                  Unknown
##   employment_status        employment_area detected_country
## 1           Unknown                Unknown               GB
## 2 working_part_time teaching_and_education               PE
## 3           Unknown                Unknown               NG
## 4           Unknown                Unknown               UG
## 5           Unknown                Unknown               IM
## 6           Unknown                Unknown               NO
```

2.check data

Field Description

1.**learner_id**: Each learner has his or her own id, which can be used to query multiple tables. However, since only one table will be used in this analysis, it has little effect in this report. 2.**enrolled_at**: The start time for learners to attend the course 3.**unenrolled_at**: Time to cancel the account 4.**role**:The learner 5.**fully_participated_at**:The date a full participant completes the course,if this list of features has data, it means that the learner completed the course 6.**purchased_statement_at**:Time for learners to purchase certificates 7.**gender**:Gender of learners 8.**country**:The country the learner is from 9.**age_range**:Age distribution of learners 10.**highest_education**:The learner's highest degree 11.**employment_status**:The learner's working state 12.**employment_area**:The learner's field of work 13.**detected_country**:The system detects the country the learner is from

Scale of statistics.

```
Sys.setenv(LANGUAGE = "en")
library(psych)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#glimpse(data1)
summary(data1)
```

```
##   learner_id        enrolled_at        unenrolled_at          role
##  Length:14394       Length:14394       Length:14394       Length:14394
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  fully_participated_at purchased_statement_at    gender
##  Length:14394          Length:14394           Length:14394
##  Class :character      Class :character       Class :character
##  Mode  :character      Mode  :character       Mode  :character
```

```
##    country            age_range          highest_education_level
##  Length:14394       Length:14394       Length:14394
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  employment_status  employment_area    detected_country
##  Length:14394       Length:14394       Length:14394
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
str(data1)
```

```
## 'data.frame':    14394 obs. of  13 variables:
##  $ learner_id             : chr  "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b" "4dc22fed-63d4-4bf6-b162-bdf
##  $ enrolled_at            : chr  "2016-08-10 14:28:49 UTC" "2016-05-24 17:34:34 UTC" "2016-05-19 00:
##  $ unenrolled_at          : chr  "" "2018-10-30 20:20:51 UTC" "" "" ...
##  $ role                   : chr  "learner" "learner" "learner" "learner" ...
##  $ fully_participated_at  : chr  "" "" "2016-09-22 16:56:03 UTC" "" ...
##  $ purchased_statement_at : chr  "" "" "" "" ...
##  $ gender                 : chr  "Unknown" "male" "Unknown" "Unknown" ...
##  $ country                : chr  "Unknown" "PE" "Unknown" "Unknown" ...
##  $ age_range              : chr  "Unknown" "46-55" "Unknown" "Unknown" ...
##  $ highest_education_level: chr  "Unknown" "university_degree" "Unknown" "Unknown" ...
##  $ employment_status      : chr  "Unknown" "working_part_time" "Unknown" "Unknown" ...
##  $ employment_area        : chr  "Unknown" "teaching_and_education" "Unknown" "Unknown" ...
##  $ detected_country       : chr  "GB" "PE" "NG" "UG" ...
```

```
#head(describe(data1))
```

1.As we can see, most of the data types are character types, but they need to be converted to categorical or numeric types in later modeling.

2.The data set have a lot of "unknown",it is like NONE,we need remove it

Data Preparation

1.choose the data

because this is first cycle,so we choose all data that we can analysis.

2.data cleaning

omit the NA

```
Sys.setenv(LANGUAGE='en')
library(dplyr)
data1 <- read.csv(file = 'cyber-security-1_enrolments.csv')
data1[data1=='Unknown'] <- NA
data1 <- na.omit(data1)
```

3.data construction

In this cycle,i choose most of features except learner_id,roll.

```
data_new <- select(data1,-c(learner_id,role))
#head(data_new)
```

4.data processing

```
#turn gender into factor 0,1
#In this place, we need to remove nonbinary and other in our data set,because there are so few data tha
data_new$gender[data_new$gender=='other'] <- NA
data_1 <- na.omit(data_new)
data_1$gender[data_1$gender=='nonbinary'] <- NA
data_new <- na.omit(data_1)
data_new$gender <- as.factor(data_new$gender)
data_new$gender <- as.numeric(data_new$gender)-1
#turn all the date into factor 0 and 1,if have data is 1 or 0
data_new$unenrolled_at[data_new$unenrolled_at==''] <- NA
data_new$fully_participated_at[data_new$fully_participated_at==''] <- NA
data_new$purchased_statement_at[data_new$purchased_statement_at==''] <- NA
data_new$enrolled_at <- ifelse(is.na(data_new$enrolled_at),0,1)
data_new$unenrolled_at <- ifelse(is.na(data_new$unenrolled_at),0,1)
data_new$fully_participated_at <- ifelse(is.na(data_new$fully_participated_at),0,1)
data_new$purchased_statement_at <- ifelse(is.na(data_new$purchased_statement_at),0,1)
#Setting dummy variables
#Sys.setenv(LANGUAGE = "en")
#library("nnet")
#data_new <- read.csv("file_new.csv")
#dummy_detected_country <- class.ind(data_new$detected_country)
#data_new$dummy_detected_country <- dummy_detected_country
```

5.Export the processed data

```
write.csv(data_new,file = 'file_new.csv')
```

Modeling

1.Model choose

In this project, the construction of the model is mainly about the understanding and analysis of data

2.training model

```
data_new <- read.csv("file_new.csv")
#str(data_new)
```

Data distribution visualization

1.Barplot

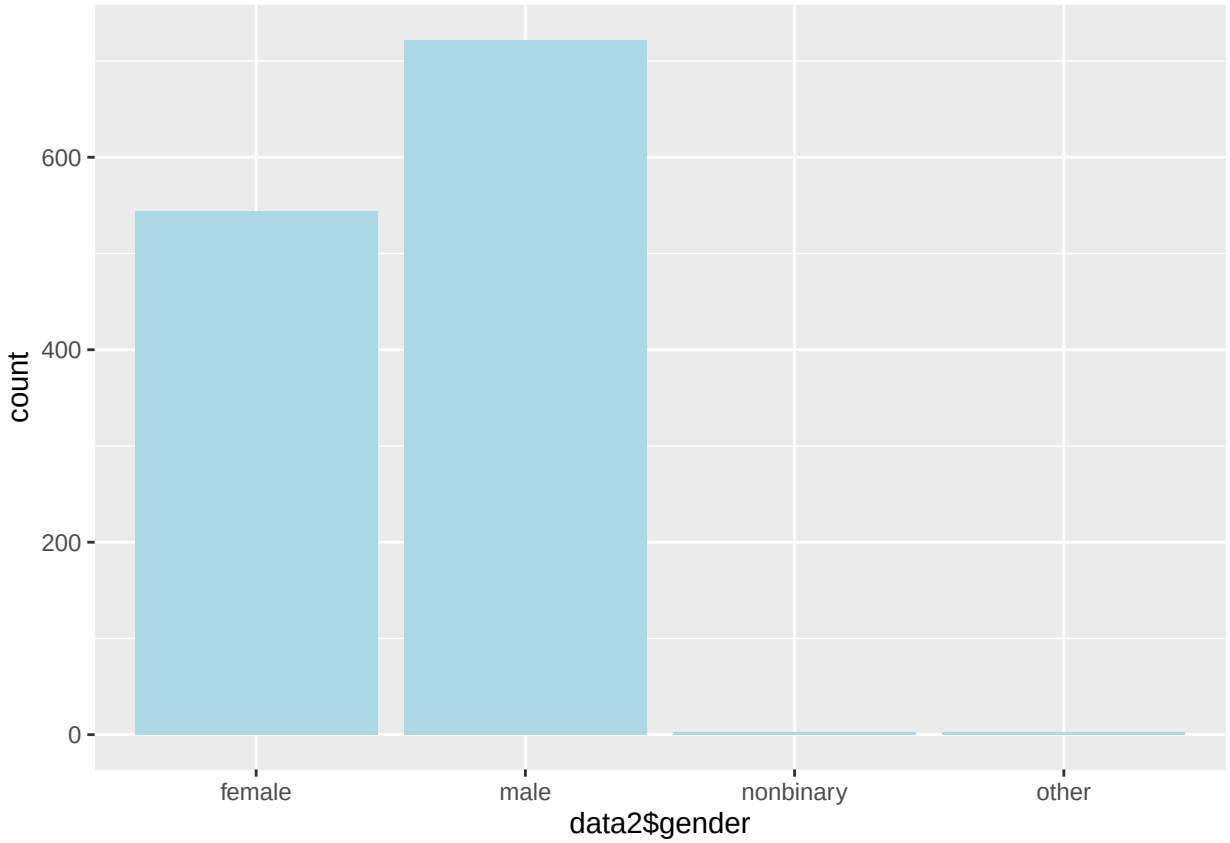
gender

```
data1 <- read.csv("cyber-security-1_enrolments.csv")
Sys.setenv(LANGUAGE = "en")
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
data1[data1=='Unknown'] <- NA
data2 <- na.omit(data1)
ggplot(data2, aes(data2$gender))+
geom_bar(fill="lightblue")
```



we can see there are slightly more men than women, but not by much,apart from that,there are few learners are nonbinary and other.

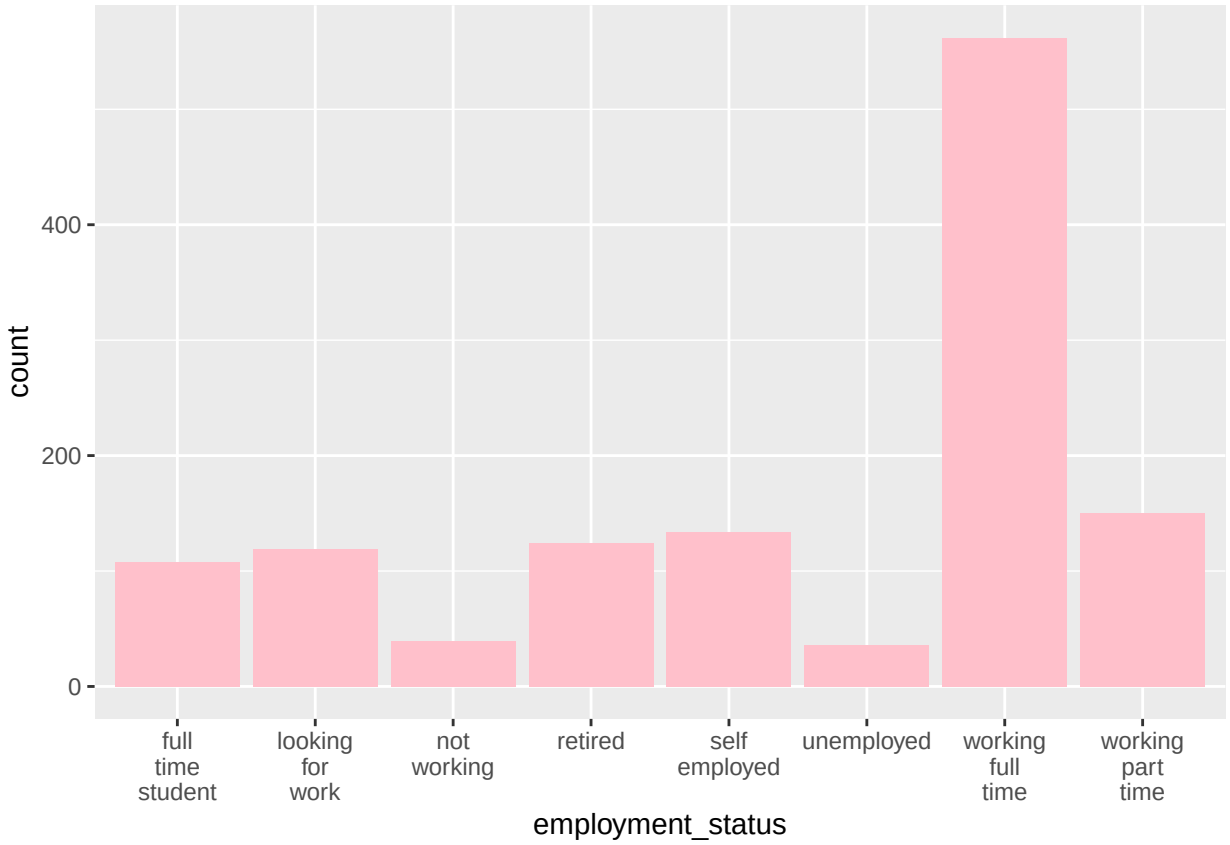
employment_status

```
Sys.setenv(LANGUAGE = "en")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   0.3.5
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
data2 %>%
  tibble() %>%
  select(employment_status) %>%
  mutate(employment_status = str_replace_all(employment_status, "_", " ")) %>%
  mutate(employment_status = str_wrap(employment_status, w=1)) %>%
  ggplot()+
  geom_bar(fill="pink", mapping = aes(x = employment_status))
```



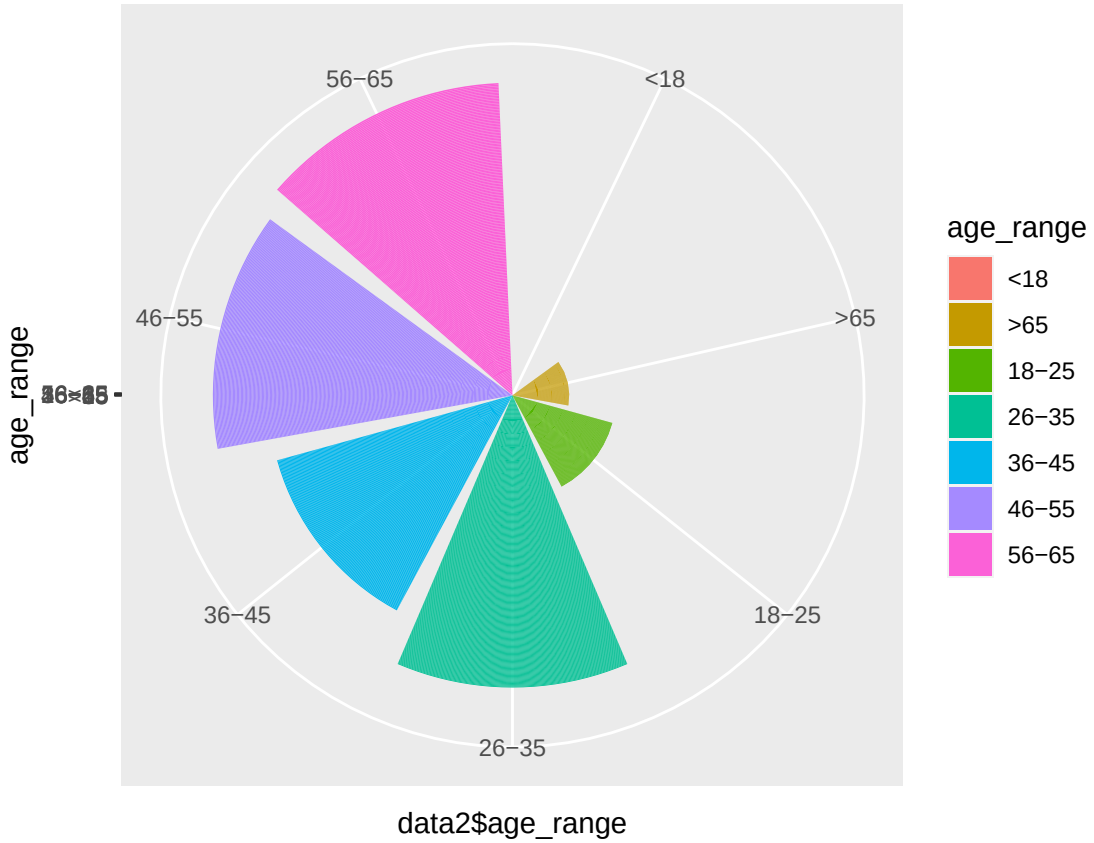

```
#ggplot(data2, aes(data2$employment_status))+
#geom_bar(fill="pink")
```

Full-time workers make up about 50 percent of the class's members.

2.Pie

age range

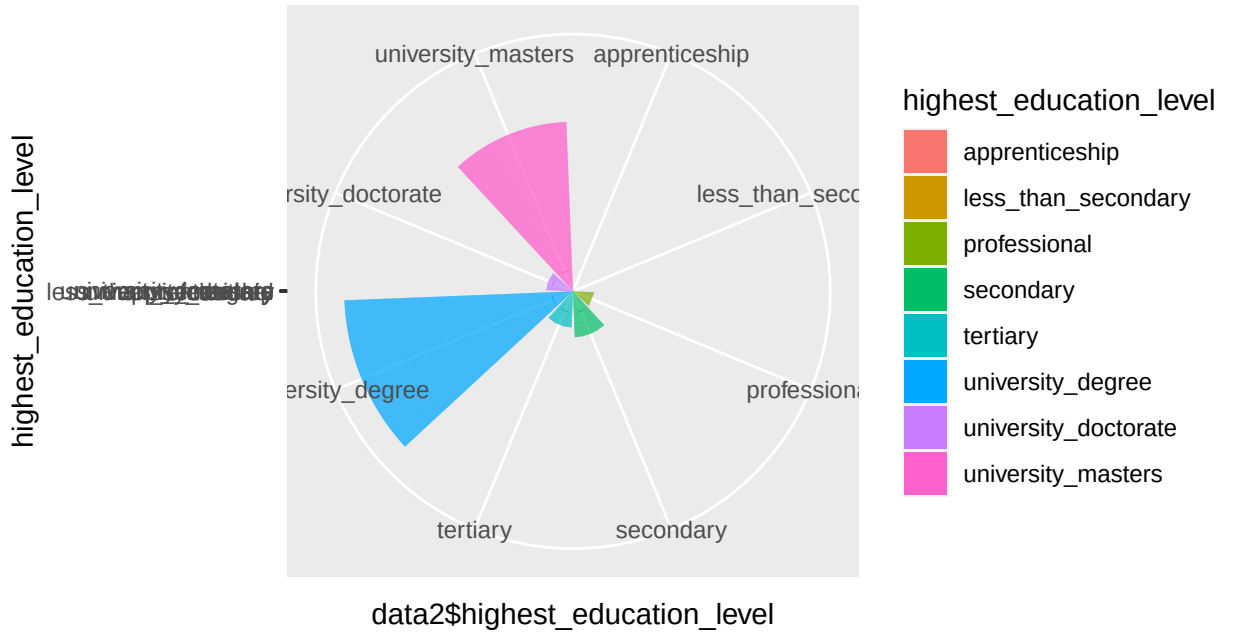
```
age_range <- data2$age_range
ggplot(data2,aes(x=data2$age_range,y=age_range,fill=age_range))+
  geom_bar(stat="identity")+
  coord_polar()
```



In this course,most of students age range are in 26-35,45-55,56-65,and the largest number of people are between 56 and 65 years old.This set of data shows that this course is more popular among middle-aged and elderly people.

highest_education_level

```
highest_education_level <- data2$highest_education_level
ggplot(data2,aes(x=data2$highest_education_level,y=highest_education_level,fill=highest_education_level
  geom_bar(stat="identity")+
  coord_polar()
```



We can clearly see that this course attracts most of the people with the highest qualifications, mostly undergraduate and master's students.

some data are hard to visualize

1.employment area

```
## [1] 289
```

```
## [1] 0.2272013
```

According to the result,we can see the learners come from all walks of life, computer, finance, sports rehabilitation and so on,among them, the computer industry learners account for the largest proportion, about 23 percent of the total number of students.

2.detected country

```
## [1] 436
```

```
## [1] 0.3427673
```

It is not difficult to see that the majority of learners are from the UK, accounting for about 34 percent of the total number(436).

Model Evaluation

In the first cycle ,our main goal is to observe and understand the data set for analysis in the next cycle.

In this place,it's my evaluation of the entire data set:

1.**learner_id**: It doesn't matter much in this project because I'm only using one table in this project, so I don't need ids for table to table joins. In the next loop, this feature will be removed. 2.**enrolled_at**: This is an important data point in this project, which can be used to assess the change in the number of participants as the course progresses. 3.**unenrolled_at**: Not important data in this project. 4.**role**: Not important,in the next cycle, this feature will be removed. 5.**fully_participated_at**: Very important indicators,and it can assess whether the student has completed the course. 6.**purchased_statement_at**: important indicators,and it can assess the importance learners place on certificates. 7.**gender**: Very important indicators,since they are categorical variables, any combination of variables can form an evaluation of the data 8.**country**: Is a good variable to see the country the learner comes from, and can also be combined with various indicators to see whether the region affects other indicators.However, as 30% of the learners are from the UK, which accounts for a large proportion, there may be errors in the assessment. 9.**age_range**: Very important indicators,the age group of the course can be analyzed. 10.**highest_education**: Very important indicators,it is possible to analyze which educational background this course mainly attracts. 11.**employment_status**: Very important indicators,it is used to assess the status of the learners and analyze which stage of the course is mainly helpful to the practitioners. 12.**employment_area**: Very important indicators,it can be analyzed that what industry do the learners come from 13.**detected_country**: important indicators,different from the eighth variable, this is the country address detected by the system, which should be more accurate

Summary

In the first cycle, we made clear all the data distribution and data situation, and preprocessed the data before. In the next cycle, we will aim to analyze the relationship between the completion rate (fully_participated_at) and other variables.

CRISP_DM_2

#BusinessUnderstanding

SMART principal

Specific

Which indicators of characteristics have a higher completion rate

Measurable

After the data in the previous loop is preprocessed, the data in the data set can be used for analysis

Attainable

We have the data that was preprocessed in the last cycle, file_new

Relevant

Often the completion rate is related to many indicators, for example, people who work full time may have less time to take the course, and people who are older tend to have more time to study.

Time-bouned

We will present our analysis report within 7 days

Data Understanding

1.load data set

```
data1 <- read.csv(file = 'file_new.csv')
```

2.check data

Field Description

1..**enrolled_at**: The start time for learners to attend the course,an in the previous round of data processing, we had changed all the dates to numbers, so if we had any, it would be 1, and if we didn't, it would be 0 2.**unenrolled_at**: Time to cancel the account,and in the previous round of data processing, we had changed all the dates to numbers, so if we had any, it would be 1, and if we didn't, it would be 0 3.**fully_participated_at**:The date a full participant completes the course,if this list of features has data, it means that the learner completed the course,and in the previous round of data processing, we had changed all the dates to numbers, so if we had any, it would be 1, and if we didn't, it would be 0 4.**purchased_statement_at**:Time for learners to purchase certificates,and in the previous round of data processing, we had changed all the dates to numbers, so if we had any, it would be 1, and if we didn't, it would be 0 5.**gender**:Gender of learners.In the last loop, we've changed it to a categorical variable, 1 for male and 0 for female 6.**country**:The country the learner is from 7.**age_range**:Age distribution of learners 8.**highest_education**:The learner's highest degree 9.**employment_status**:The learner's working state 10.**employment_area**:The learner's field of work 11.**detected_country**:The system detects the country the learner is from.and in the last cycle, we changed them into dummy variables. Later, after analysis, we found that they were not useful, so these dummy variables will be removed from the data processing in this cycle

Data Preparation

1.choose the data

Because our data set is fixed, we select features.

x=fully_participated_at,y=enrolled_at+purchased_statement_at+gender+age_range+highest_education_level+employmer
# 2.data cleaning

```
data1 <- read.csv("file_new.csv")
data_x <- data1$fully_participated_at
#head(data_x)
data_y <-data.frame(data1$enrolled_at,data1$purchased_statement_at,data1$gender,data1$age_range,data1$hi
```

3.data construction

for highest_education,we create new line,divide it into 3 part:

1.ordinary(less_than_secondary,secondary) 2.medium(university_degree,university_masters,apprenticeship) 3.high(professional,tertiary,universiety_doctorate)

```
table(data_y$data1.highest_education_level)
```

```
##
##       apprenticeship  less_than_secondary         professional
##                    3                   13                  101
```

```
##            secondary             tertiary    university_degree
##                  162                  102                  533
## university_doctorate   university_masters
##                   54                  298
```

```
data_y$data1.highest_education_level[data_y$data1.highest_education_level == 'secondary'] <- 'ordinary'
data_y$data1.highest_education_level[data_y$data1.highest_education_level =='less_than_secondary'] <- '
data_y$data1.highest_education_level[data_y$data1.highest_education_level == 'university_degree'] <- 'm
data_y$data1.highest_education_level[data_y$data1.highest_education_level == 'university_masters'] <- '
data_y$data1.highest_education_level[data_y$data1.highest_education_level == 'apprenticeship'] <- 'medi
data_y$data1.highest_education_level[data_y$data1.highest_education_level == 'professional'] <- 'high'
data_y$data1.highest_education_level[data_y$data1.highest_education_level == 'tertiary'] <- 'high'
data_y$data1.highest_education_level[data_y$data1.highest_education_level =='university_doctorate'] <-
head(data_y$data1.highest_education_level)
```

```
## [1] "medium"   "ordinary" "medium"   "high"     "medium"   "ordinary"
```

```
table(data_y$data1.highest_education_level)
```

```
##
##     high   medium ordinary
##      257      834      175
```

for age_range,we create new line,divide it into 4 parts:

1.young people(<18,18-25), 2.prime of life(26-35,36-45), 3.middle age(46-55,56-65), 4.old age(>65)

```
table(data_y$data1.age_range)
```

```
##
##   <18   >65 18-25 26-35 36-45 46-55 56-65
##     2   129   156   329   221   227   202
```

```
data_y$data1.age_range[data_y$data1.age_range == '<18'] <- 'young people'
data_y$data1.age_range[data_y$data1.age_range == '18-25'] <- 'young people'
data_y$data1.age_range[data_y$data1.age_range == '26-35'] <- 'prime of life'
data_y$data1.age_range[data_y$data1.age_range == '36-45'] <- 'prime of life'
data_y$data1.age_range[data_y$data1.age_range == '46-55'] <- 'middle age'
data_y$data1.age_range[data_y$data1.age_range == '56-65'] <- 'middle age'
data_y$data1.age_range[data_y$data1.age_range == '>65'] <- 'old age'
table(data_y$data1.age_range)
```

```
##
##    middle age       old age prime of life  young people
##           429           129           550           158
```

for employment_status,we create 3 parts:

1.no job(full_time_student,looking_for_work,not_working,unemployed) 2.in work(working_full_time,working_part_time,self
3.retired(retired)

```
table(data_y$data1.employment_status)
```

```
##
## full_time_student  looking_for_work       not_working           retired
##               107               119                38               124
##     self_employed        unemployed working_full_time working_part_time
##               133                35               561               149
```

```
data_y$data1.employment_status[data_y$data1.employment_status == 'full_time_student'] <- 'no job'
data_y$data1.employment_status[data_y$data1.employment_status == 'looking_for_work'] <- 'no job'
data_y$data1.employment_status[data_y$data1.employment_status == 'not_working'] <- 'no job'
data_y$data1.employment_status[data_y$data1.employment_status == 'unemployed'] <- 'no job'
data_y$data1.employment_status[data_y$data1.employment_status == 'working_full_time'] <- 'in work'
data_y$data1.employment_status[data_y$data1.employment_status == 'working_part_time'] <- 'in work'
data_y$data1.employment_status[data_y$data1.employment_status == 'self_employed'] <- 'in work'
data_y$data1.employment_status[data_y$data1.employment_status == 'retired'] <- 'retired'
table(data_y$data1.employment_status)
```

```
##
## in work  no job retired
##     843     299     124
```

for employment_area,we create 2 parts:

1.traditional industries(business_consulting_and_management, engineering_and_manufacturing,health_and_social_care, marketing_advertising_and_pr,property_and_construction, recruitment_and_pr,transport_and_logistics, armed_forces_and_emergency_services, charities_and_voluntary_work,environment_and_agriculture, hospitality_tourism_and_sport,law,media_and_publishing, public_sector,retail_and_sales,teaching_and_education)

2.Non-traditional industries(accountancy_banking_and_finance, creative_arts_and_culture,it_and_information_services, science_and_pharmaceuticals,energy_and_utilities)

```
data_y$data1.employment_area[data_y$data1.employment_area == 'business_consulting_and_management'] <- '
data_y$data1.employment_area[data_y$data1.employment_area == 'engineering_and_manufacturing'] <- 'tradi
data_y$data1.employment_area[data_y$data1.employment_area == 'health_and_social_care'] <- 'traditional
data_y$data1.employment_area[data_y$data1.employment_area == 'marketing_advertising_and_pr'] <- 'tradit
data_y$data1.employment_area[data_y$data1.employment_area == 'property_and_construction'] <- 'traditiona
data_y$data1.employment_area[data_y$data1.employment_area == 'recruitment_and_pr'] <- 'traditional indus
data_y$data1.employment_area[data_y$data1.employment_area == 'transport_and_logistics'] <- 'traditional
data_y$data1.employment_area[data_y$data1.employment_area == 'armed_forces_and_emergency_services'] <-
data_y$data1.employment_area[data_y$data1.employment_area == 'charities_and_voluntary_work'] <- 'tradit
data_y$data1.employment_area[data_y$data1.employment_area == 'environment_and_agriculture'] <- 'traditi
data_y$data1.employment_area[data_y$data1.employment_area == 'hospitality_tourism_and_sport'] <- 'tradi
data_y$data1.employment_area[data_y$data1.employment_area == 'law'] <- 'traditional industries'
data_y$data1.employment_area[data_y$data1.employment_area == 'media_and_publishing'] <- 'traditional in
data_y$data1.employment_area[data_y$data1.employment_area == 'public_sector'] <- 'traditional industrie
data_y$data1.employment_area[data_y$data1.employment_area == 'retail_and_sales'] <- 'traditional indust
data_y$data1.employment_area[data_y$data1.employment_area == 'teaching_and_education'] <- 'traditional
data_y$data1.employment_area[data_y$data1.employment_area == 'accountancy_banking_and_finance'] <- 'Non
data_y$data1.employment_area[data_y$data1.employment_area == 'creative_arts_and_culture'] <- 'Non-tradi
data_y$data1.employment_area[data_y$data1.employment_area == 'it_and_information_services'] <- 'Non-tra
data_y$data1.employment_area[data_y$data1.employment_area == 'science_and_pharmaceuticals'] <- 'Non-tra
data_y$data1.employment_area[data_y$data1.employment_area == 'energy_and_utilities'] <- 'Non-traditional
data_y$data1.employment_area[data_y$data1.employment_area == ''] <- 'Non-traditional industries'
table(data_y$data1.employment_area)
```

```
##
## Non-traditional industries     traditional industries
##                        433                        833
```

Modeling

1.Model choose

In this project, we will build a graph model and analyze the relationship between the completion rate and various parameters by drawing various graphs

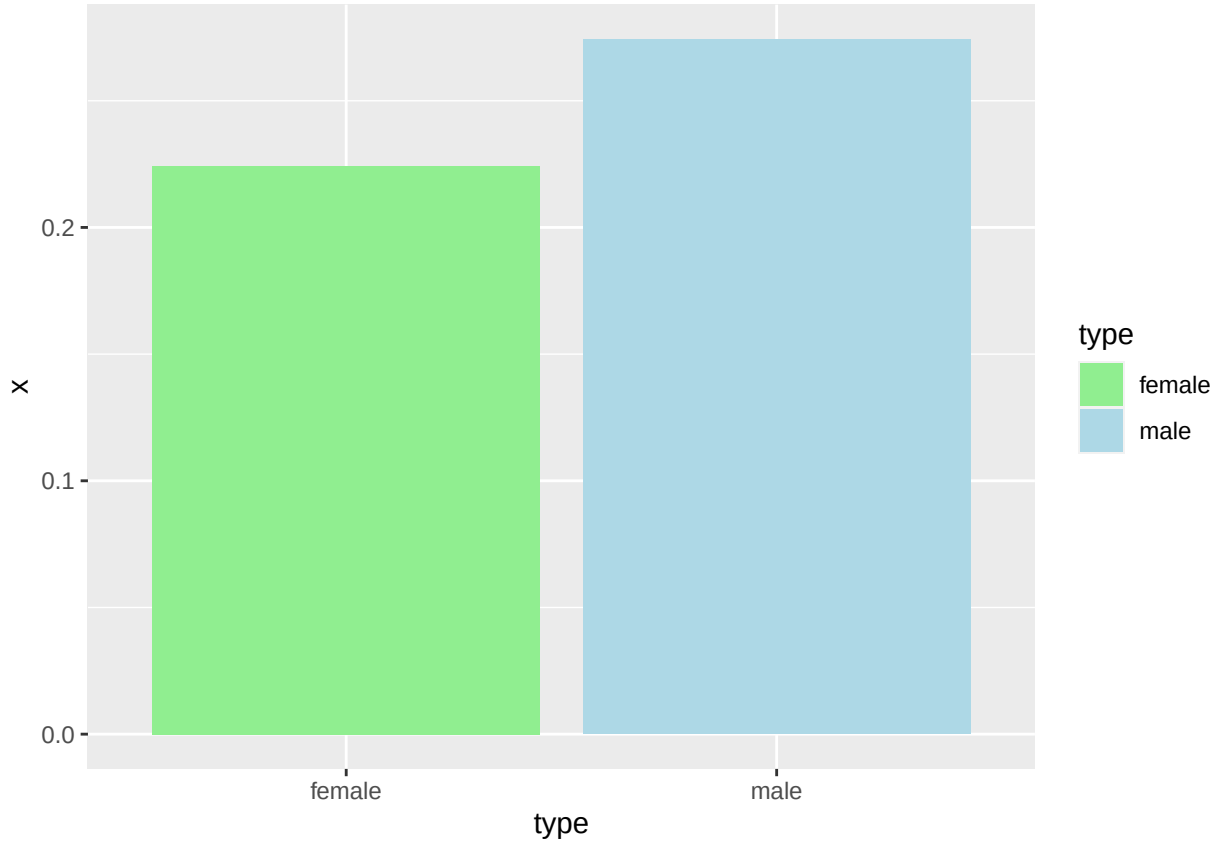
2.training model(data visualization)

```
x <- read.csv(file = "data_x.csv")
y <- read.csv(file = "data_y.csv")
y$completion_rate <- x$x
table(y$data1.gender)
```

```
##
##   0   1
## 544 722
```

gender & completion rate(fully_participated_at)

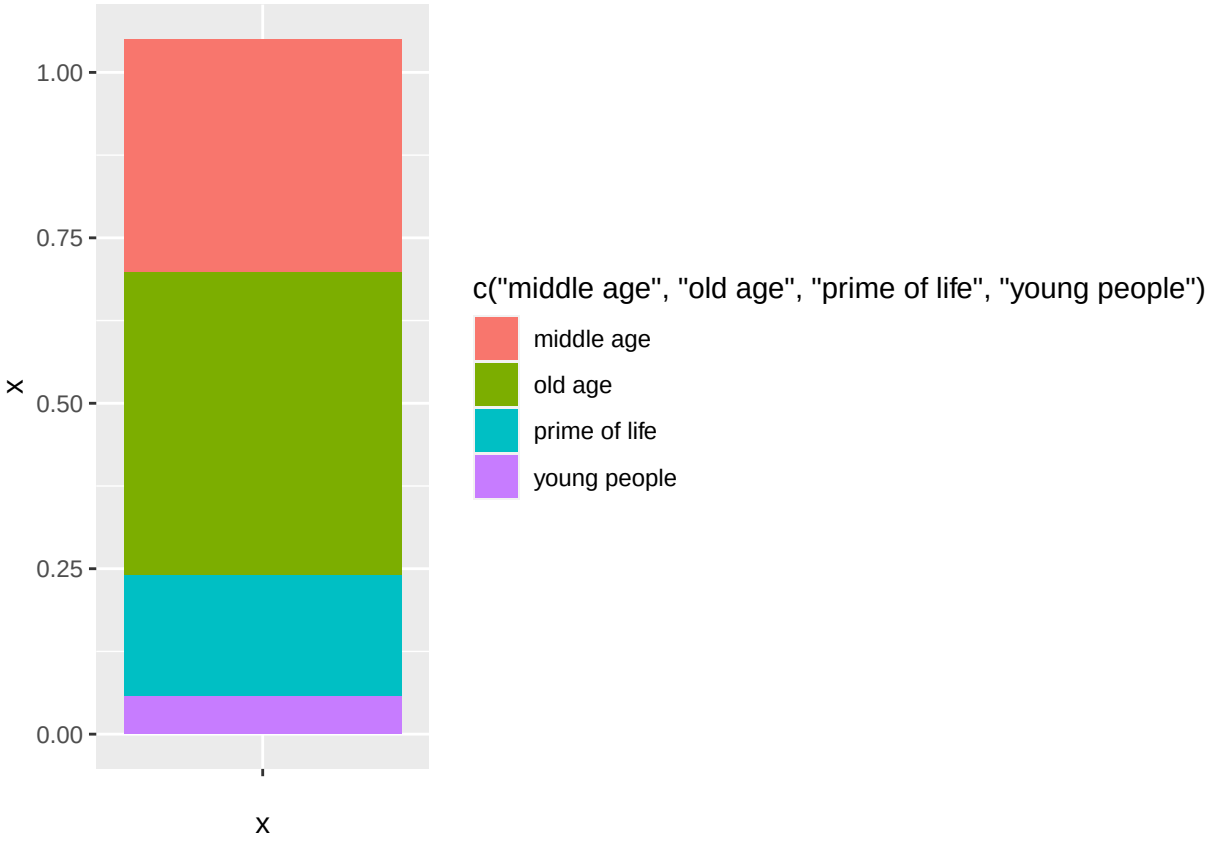
```
y$data1.gender <- ifelse(y$data1.gender=="0",'female','male')
data_g_2 <- data.frame(aggregate(y$completion_rate, by=list(type=y$data1.gender),sum))
data_g_2$type <- as.factor(data_g_2$type)
data_g_2$x <- as.numeric(data_g_2$x)
Sys.setenv(LANGUAGE="en")
library(ggplot2)
data_g_1 <- data.frame(table(y$data1.gender))
data_g_3 <- data_g_2/data_g_1
data_g_3$type[1] <- 'female'
data_g_3$type[2] <- 'male'
ggplot(data_g_3,aes(x=type,y=x,,fill=type)) +
  geom_bar(stat="identity")+
  scale_fill_manual(values = c("lightgreen", "lightblue") )
```



According to the situation of this plot, the completion rate of male is higher than that of female.This shows that men are more interested in Internet security, which is also more consistent with the law of reality. Men are slightly better at understanding and solving problems than women when it comes to computers and the Internet

age_range & completion rate(fully_participated_at)

```
data_a_1 <- data.frame(table(y$data1.age_range))
data_a_2 <- data.frame(aggregate(y$completion_rate, by=list(type=y$data1.age_range),sum))
data_a_3 <- data_a_2/data_a_1

ggplot(data_a_3, aes(x='', y=x,fill=c('middle age','old age','prime of life','young people'))) +
  geom_bar(stat="identity", width=1)
```



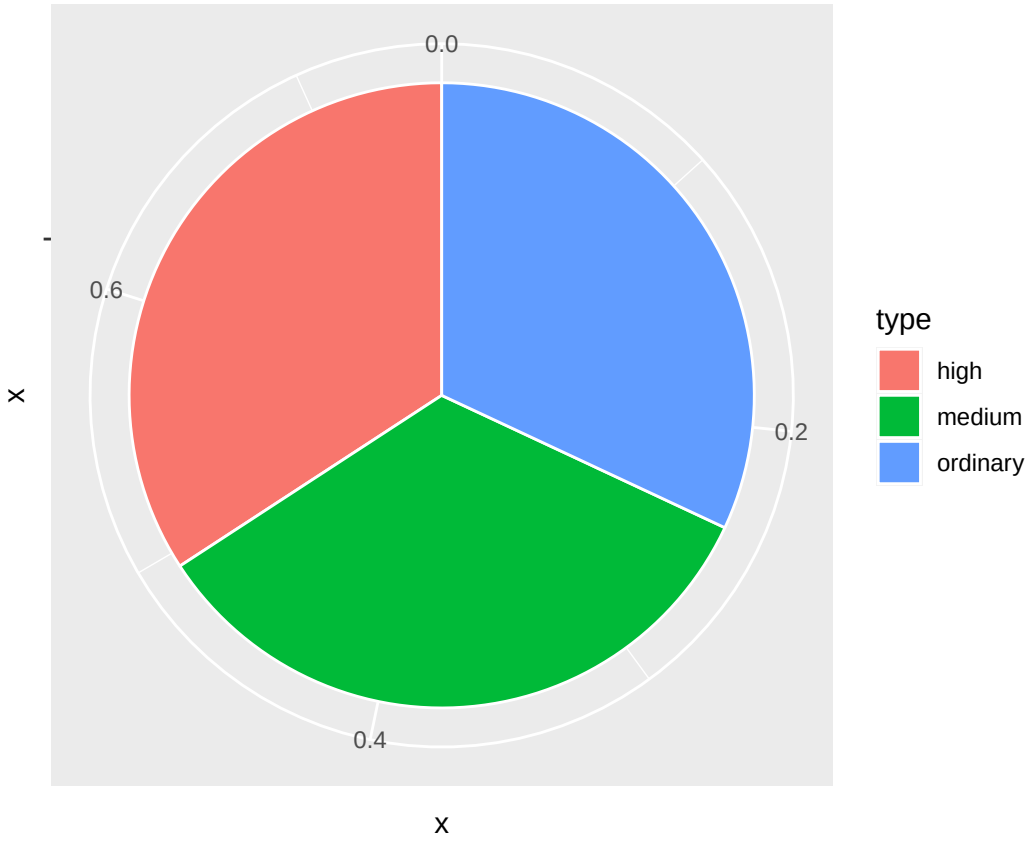
We can see that the completion rate of middle-aged and elderly people is higher than that of young and middle-aged people. It can be understood from two aspects: 1. Most middle-aged and elderly people have already married, so they pay more attention to family network security. 2. The middle-aged and the elderly have more time to study this course than the young. On the contrary, young adults may be busy with work or study.In the following courses, the content can be slightly biased towards the middle-aged and the elderly, so as to improve the completion rate of the course.

highest_education level & completion rate(fully_participated_at)

```
data_h_1 <- data.frame(table(y$data1.highest_education_level))
data_h_2<- data.frame(aggregate(y$completion_rate,by=list(type=y$data1.highest_education_level),sum))
data_h_3 <- data_h_2/data_h_1
data_h_3$type[1] <- 'high'
data_h_3$type[2] <- 'medium'
data_h_3$type[3] <- 'ordinary'
data_h_3
```

```
##       type         x
## 1     high 0.2568093
## 2   medium 0.2541966
## 3 ordinary 0.2400000
```

```
ggplot(data_h_3, aes(x='', y=x, fill=type)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0)
```



We can see that the completion rate of learners is similar no matter what degree they are in. This shows that home Internet security is a subject of great concern to the public, regardless of education level. Therefore, in my opinion, we should try to use less professional vocabulary in the course and more easy to understand vocabulary. At the same time, the content should not be too advanced and easy to understand.

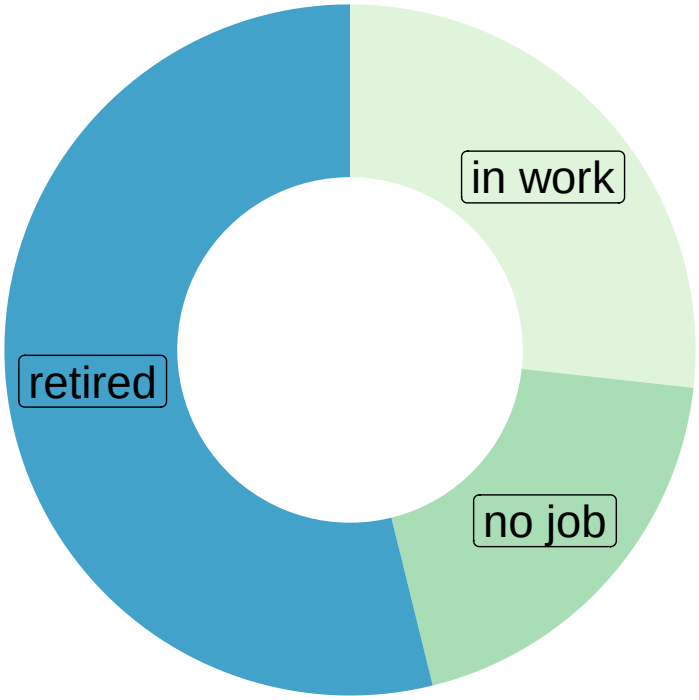
employment_status & completion rate(fully_participated_at)

```
data_e_1 <- data.frame(table(y$data1.employment_status))
data_e_2<- data.frame(aggregate(y$completion_rate,by=list(type=y$data1.employment_status),sum))
data_e_3 <- data_e_2/data_e_1
data_e_3$type[1] <- "in work"
data_e_3$type[2] <- "no job"
data_e_3$type[3] <- "retired"
data_e_3
```

```
##      type         x
## 1 in work 0.2443654
## 2  no job 0.1772575
## 3 retired 0.4919355
```

```
# Compute percentages
data_e_3$fraction <- data_e_3$x / sum(data_e_3$x)

# Compute the cumulative percentages (top of each rectangle)
data_e_3$ymax <- cumsum(data_e_3$fraction)

# Compute the bottom of each rectangle
```

```
data_e_3$ymin <- c(0, head(data_e_3$ymax, n=-1))

# Compute label position
data_e_3$labelPosition <- (data_e_3$ymax + data_e_3$ymin) / 2

# Make the plot
ggplot(data_e_3, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=type)) +
  geom_rect() +
  geom_label( x=3.5, aes(y=labelPosition, label=data_e_3$type), size=6) +
  scale_fill_brewer(palette=4) +
  coord_polar(theta="y") +
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "none")
```

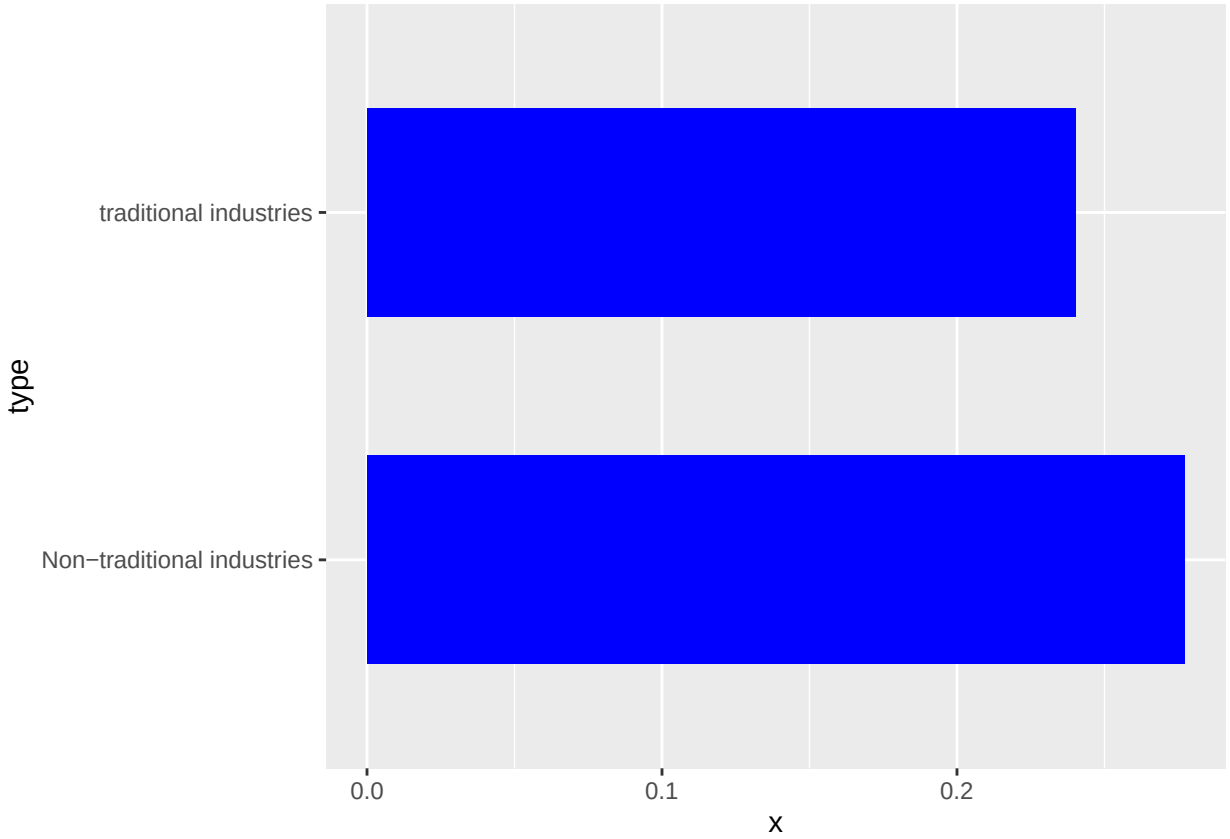


Retired people tend to have more time to study the course, so there is no doubt that retirees have a higher completion rate.But the interesting part is that the completion rate of employed people is even slightly higher than that of unemployed people, which suggests that there's a reason why these people can't find work.

employment_area & completion rate(fully_participated_at)

```
data_em_1 <- data.frame(table(y$data1.employment_area))
data_em_2<- data.frame(aggregate(y$completion_rate,by=list(type=y$data1.employment_area),sum))
data_em_3 <- data_em_2/data_em_1
data_em_3$type[1] <- 'Non-traditional industries'
data_em_3$type[2] <- 'traditional industries'
```

```
ggplot(data_em_3) +
  geom_col(aes(x, type), fill = 'blue', width = 0.6)
```



The completion rate of learners in non-traditional industries is higher than that of learners in traditional industries. In my opinion, people in non-traditional industries, such as IT and finance, have more access to computers and networks than those in traditional industries, so they know more about the importance of network security.

Model Evaluation

In this project,our main purpose is to look at the relationship between the completion rate and various indicators, so as to improve the course and improve the completion rate.Now let's take this visualization a step further.

gender & completion rate(fully_participated_at)

I don't think this parameter can be used as an indicator of our deployment goals. Since this course is a popular science course for the general public, we should not bias the content towards the part that men are more interested in just because men have a high completion rate.

age_range & completion rate(fully_participated_at)

I think this is a very important indicator for improving the course. Both the number and completion rate of young people are smaller than that of middle-aged and elderly people. It shows that the audience of this

subject is biased towards the older age. That is, if you need to improve the course later. I think there are two things that can be improved: 1.Font enlargement, because viewing is a problem for middle-aged and elderly people, font enlargement can make them feel comfortable. 2. The visual interface is simple, often only young people like gorgeous interface, for older people, practical is the key.If we want to attract more young users at the same time, I have the following suggestions: 1. Cooperate with more universities to attract more young people to our courses. 2. Add a community section that young people like, in which they can freely exchange what they have learned.

highest_education level & completion rate(fully_participated_at)

Important indicator,and we can see this from the data visualization in the previous step. The completion rate is about the same for learners regardless of their educational level. It shows that this course has the characteristics of popularization. If we need to improve, we need to develop this advantage. I have two suggestions: 1. Lower the threshold of courses and try not to use technical terms. 2. The content must be detailed. The course is divided into three parts: elementary, intermediate and advanced. Let each learner find their own position.

employment_status & completion rate(fully_participated_at)

Important indicator,This indicator is actually a little bit similar to the age indicator, because age and working status always correspond. In this section, retired learners had the highest completion rate. This also corresponds to the analysis we just made on the age index. But the completion rate of those who are working is also high. If we want to keep both types of customers. I think so: reduce the content of each episode to 15 minutes, because people who are busy at work usually have little time to study this course. And for the retired, a 15-minute lesson is not too short.

employment_area & completion rate(fully_participated_at)

Not very important indicator,since this index represents such a wide range of people, it is difficult to specifically change a course setting to improve the completion rate of these people. The reason why non-traditional companies pay more attention to cyber security is that these people spend more time with computers. If I had to make a suggestion to improve the completion rate, I think it would be helpful to insert some industry knowledge into the course, which would be helpful for learners in both traditional and non-traditional enterprises.

ResultDeployment

Deployment Plan

Course videos

1. Enlarge the font in the video and simplify the interface design. (from indicator age_range)
2. The speaker should speak at a moderate pace, not too fast. (from indicator age_range)
3. Adjust the picture quality clearly(from indicator age_range) ## Content
4. Lower the threshold of courses and use less technical terms.(from highest_education)
5. Keep the class time under 15 minutes.(from employment_status)
6. Divide the course into three parts: beginner, intermediate and advanced(from highest_education) ## Spread
7. Create a separate communication community in the video interface, where learners can share their views.(from age_range)
8. Cooperate with other universities to attract more young people.(from age_range)