STAT 650 Homework 2

YiChia Wu(UIN: 132006360)

Problem 1

a. What fraction of the calls are cancelled?

```
a <- subset(ambulance, Precancel_Dur != '-1')

length(a$Precancel_Dur) / length(ambulance$Precancel_Dur)
```

```
## [1] 0.05886667
```

b. Provide a histogram for the time in seconds until cancellation for all of those calls that are eventually cancelled.
Convert the vector x of values to second.

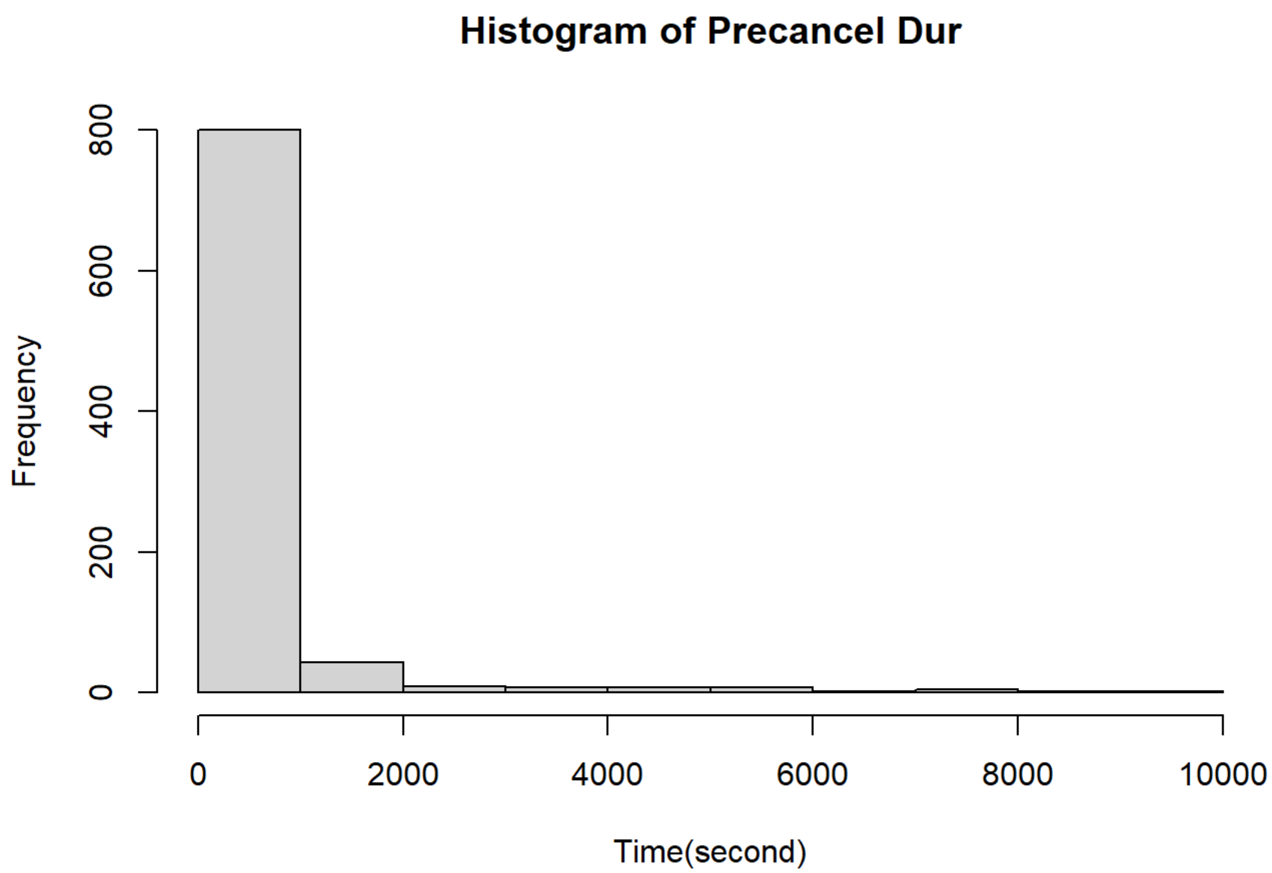
```
aaselect <- select(a, Precancel_Dur)
aaselect$Precancel_Dur <- aaselect$Precancel_Dur *24 *60 *60

knitr::kable(head(aaselect))
```

	Precancel_Dur
1	1332.201
2	6770.670
3	4366.794
5	5014.218
8	6856.097
11	4627.108

- making the full plot:

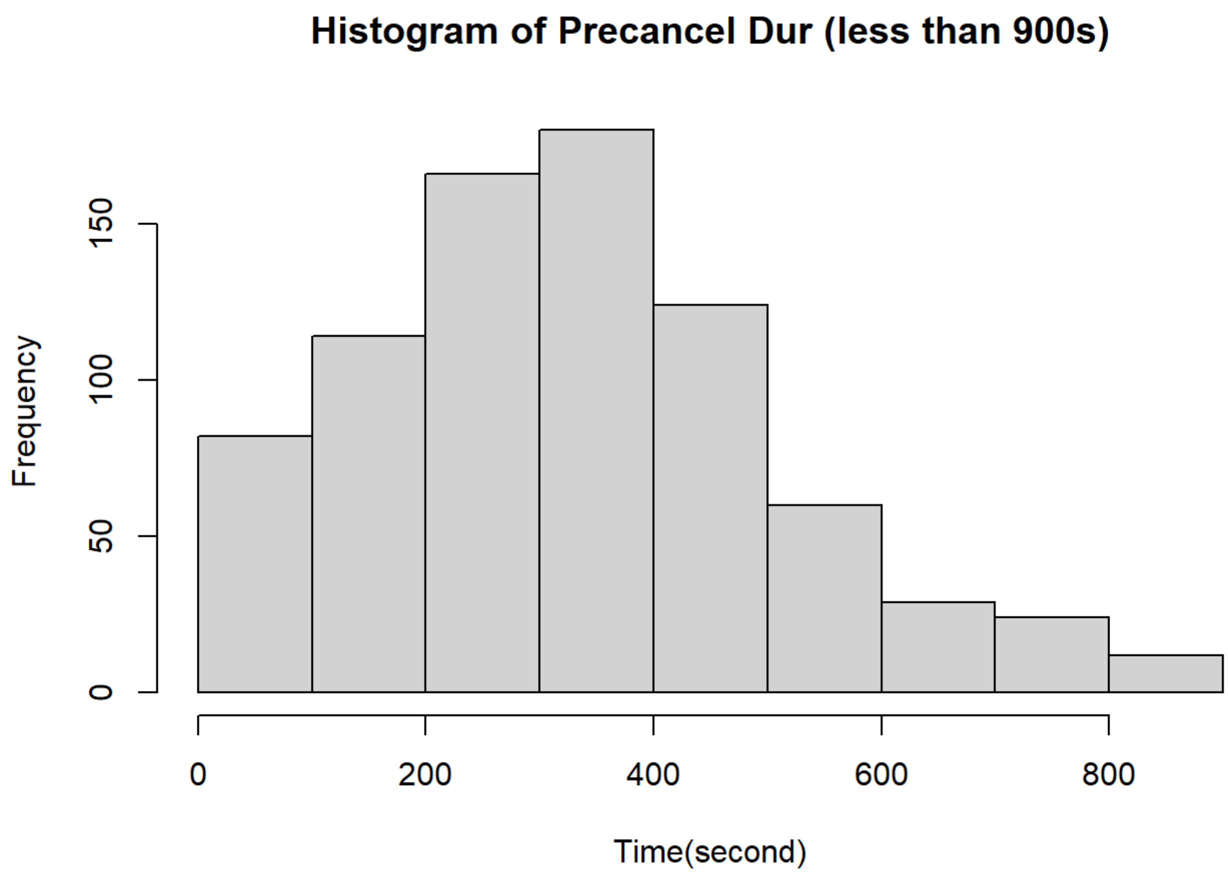
```
hist(x=aaselect$Precancel_Dur, main="Histogram of Precancel Dur", xlab="Time(second)", ylab=
"Frequency")
```



- making the plot of values $\leq$ 900 seconds:

```
bfilter<- filter(aaselect, Precancel_Dur<=900)

hist(x=bfilter$Precancel_Dur, main="Histogram of Precancel Dur (less than 900s)", xlab="Time
(second)", ylab="Frequency")
```



- How many (percentage) durations are > 900 seconds?

```
(length(aaselect$Precancel_Dur) - length(bfilter$Precancel_Dur)) / length(aaselect$Precancel_
Dur) * 100
```

```
## [1] 10.41903
```

c. What fraction of these calls require transport to a hospital?

```
c <- subset(ambulance, Precancel_Dur == '-1')
cc <- subset(c, Time_Arrive_Hosp != '-1')
cc <- select(cc, Precancel_Dur, Time_Arrive_Hosp)
knitr::kable(head(cc))
```

	Precancel_Dur	Time_Arrive_Hosp
4	-1	57.46363
6	-1	58.67119
7	-1	60.40635
9	-1	58.60423
10	-1	60.64163
12	-1	58.68276

- answer:

```
length(cc$Precancel_Dur) / length(c$Precancel_Dur)
```

```
## [1] 0.8385634
```

d. What are the mean and median times spent at the scene for calls that require transport the hospital?

- Mean: 1209.944 (second)
- Median: 1076.840 (second)

```
d <- subset(c, Time_Arrive_Hosp != '-1')
d <- select(d, Time_Arrive_Hosp, Time_Vehicle_At_Scene, Time_Depart_Scene)

d$Time_Vehicle_At_Scene <- d$Time_Vehicle_At_Scene *24 *60 *60
d$Time_Depart_Scene <- d$Time_Depart_Scene *24 *60 *60

d <- mutate(d,spent=Time_Depart_Scene-Time_Vehicle_At_Scene)
d <- select(d, spent)

summary(d)
```

```
##      spent
##  Min.   :    1.419
##  1st Qu.:  753.638
##  Median : 1076.840
##  Mean   : 1209.944
##  3rd Qu.: 1476.622
##  Max.   :12597.309
```

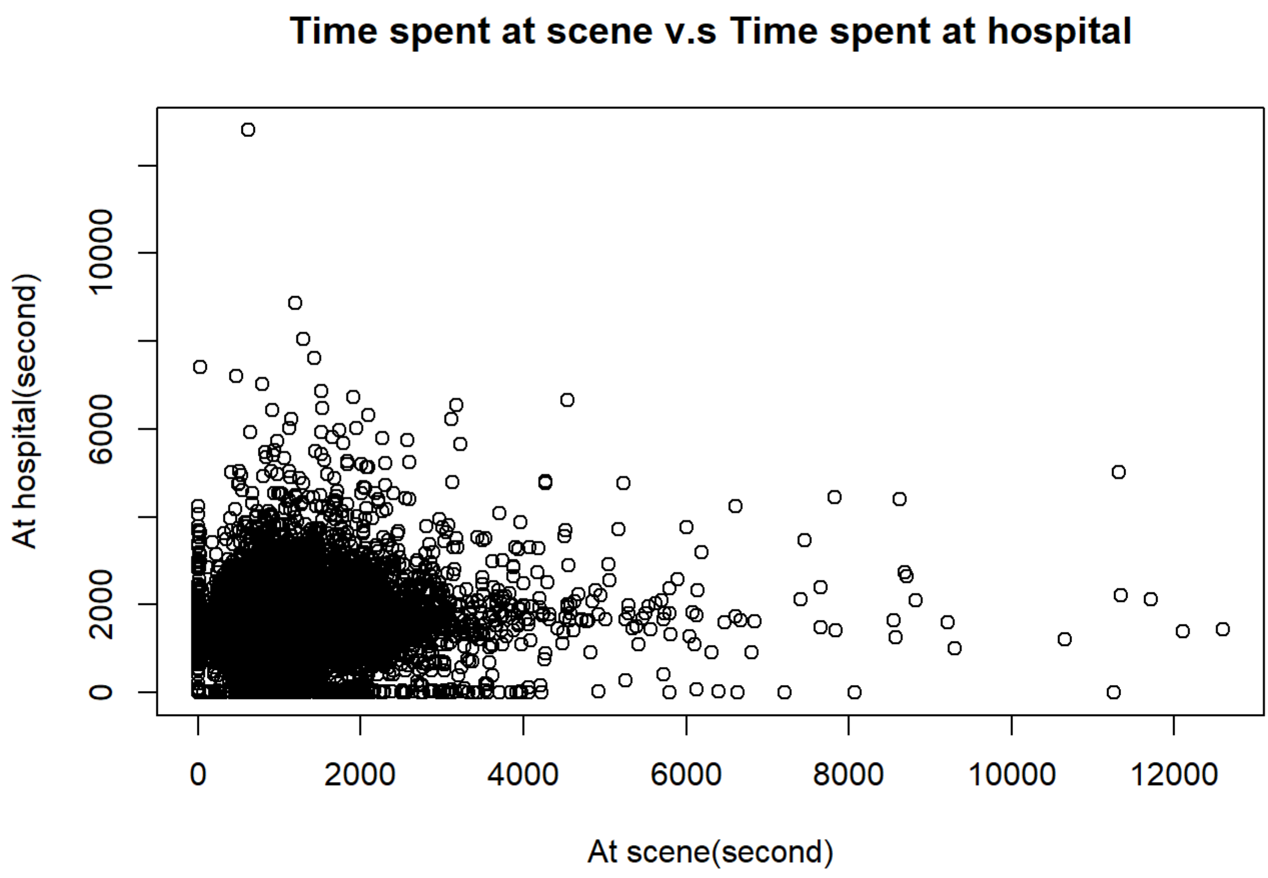
e. Generate a scatter plot of the time spent at scene versus the time spent at hospital (calls required transport to the hospital)

```
d <- subset(c, Time_Arrive_Hosp != '-1')
d <- select(d, Time_Vehicle_At_Scene, Time_Depart_Scene, Time_at_Hospital)

d$Time_Vehicle_At_Scene <- d$Time_Vehicle_At_Scene *24 *60 *60
d$Time_Depart_Scene <- d$Time_Depart_Scene *24 *60 *60
d$Time_at_Hospital <- d$Time_at_Hospital *24 *60 *60

d <- mutate(d,spent=Time_Depart_Scene-Time_Vehicle_At_Scene)
d <- select(d, spent, Time_at_Hospital)

plot(x=d$spent,
     y=d$Time_at_Hospital,
     main="Time spent at scene v.s Time spent at hospital",
     xlab="At scene(second)",
     ylab="At hospital(second)")
```



From the plot I found:

- Most of the data spent time at scene and hospital less than 4000 seconds(about one hour.) In specific, it's highly to spend time at scene less than 2000 seconds.
- It's unusual to spend lots of time at scene more than 8000 seconds(about two hour.)
- When spending more than 6000 seconds at scene, they usually spent less than 4000 seconds at hospital.
- Without outliers, it's usual to spend less than 8000 seconds at hospital.

f. What fraction of non-cancelled calls in this dataset have response times under 10 minutes?
10 minutes = 600 seconds
To find fraction that less than 600 seconds:

```
f <- subset(c, Time_Arrive_Hosp != '-1')
f <- select(f, Time_Rec, Time_Vehicle_At_Scene)

f$Time_Vehicle_At_Scene <- f$Time_Vehicle_At_Scene *24 *60 *60
f$Time_Rec <- f$Time_Rec *24 *60 *60

f <- mutate(f,spent=Time_Vehicle_At_Scene-Time_Rec)
f <- select(f, spent)

ffilter<- filter(f, spent<600)
length(ffilter$spent) / length(f$spent)
```

```
## [1] 0.5923298
```

Problem 2 (The game of craps)

a. Point $= 4$
   For each roll, P(win the game)$= \frac{3}{36}$, P(lose the game)$= \frac{6}{36}$

   By observing the probability in each roll:
   P(win the game in roll 2) $= \frac{3}{36}$
   P(win the game in roll 3) $= \frac{3}{36} \cdot \frac{27}{36}$
   P(win the game in roll 4) $= \frac{3}{36} \cdot \frac{27}{36} \cdot \frac{27}{36}$
   It will be a infinite Sum of geometric sequence to calculate the answer.

   By sum to infinity formula, with $a_1 = \frac{3}{36}, r = \frac{27}{36}$,
   We can get:
   P(win) = $\frac{a1}{1-r}$= $\frac{3/36}{9/36}$ = $\frac{1}{3}$

   In other words, the answer can also be calculated by $\frac{P(roll\ a\ 4)}{P(roll\ a\ 4)+P(roll\ a\ 7)}$

b. Calculate p_i, $i = 2\ to\ 12$
   Same as (a). Calculate each in same way.

   First, address special case:

- p_2, p_3, p_{12} will lose in first roll (with $\frac{4}{36}$ to get 0)
- p_7, p_{11} will win in first roll.(with $\frac{8}{36}$ to get 1)

   Other case:

- p_4 = p_{10} = $\frac{1}{3}$
- p_5 = p_9 = $\frac{4/36}{10/36}$ = $\frac{2}{5}$
- p_6 = p_8 = $\frac{5}{11}$

c. Compute P(Win)
   By summarize all above in (b), we can get the answer.
   P(Win) = $1 \cdot \frac{8}{36} + \frac{1}{3} \cdot \frac{3}{36} \cdot 2 + \frac{2}{5} \cdot \frac{4}{36} \cdot 2 + \frac{5}{11} \cdot \frac{5}{36} \cdot 2 + 0 \cdot \frac{4}{36}$
   = $\frac{244}{495}$

Problem 3 naive Bayes classifier

$P(play\ golf\ |sunny, hot, high, windy)$
= $\frac{P(sunny|play\ golf)P(hot|play\ golf)P(high|play\ golf)P(windy|play\ golf)P(play\ golf)}{P(sunny,hot,high,windy)}$

Calculate each feature:

Outlook_Play Golf	Yes	No
sunny	3/9	2/5
overcast	4/9	0
rainy	2/9	3/5

Temp_Play Golf	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

humidity_Play Golf	Yes	No
high	3/9	4/5
normal	6/9	1/5

Windy_Play Golf	Yes	No
True	3/9	3/5
False	6/9	2/5

Play Golf	Yes	No
Play Golf	9/14	5/14

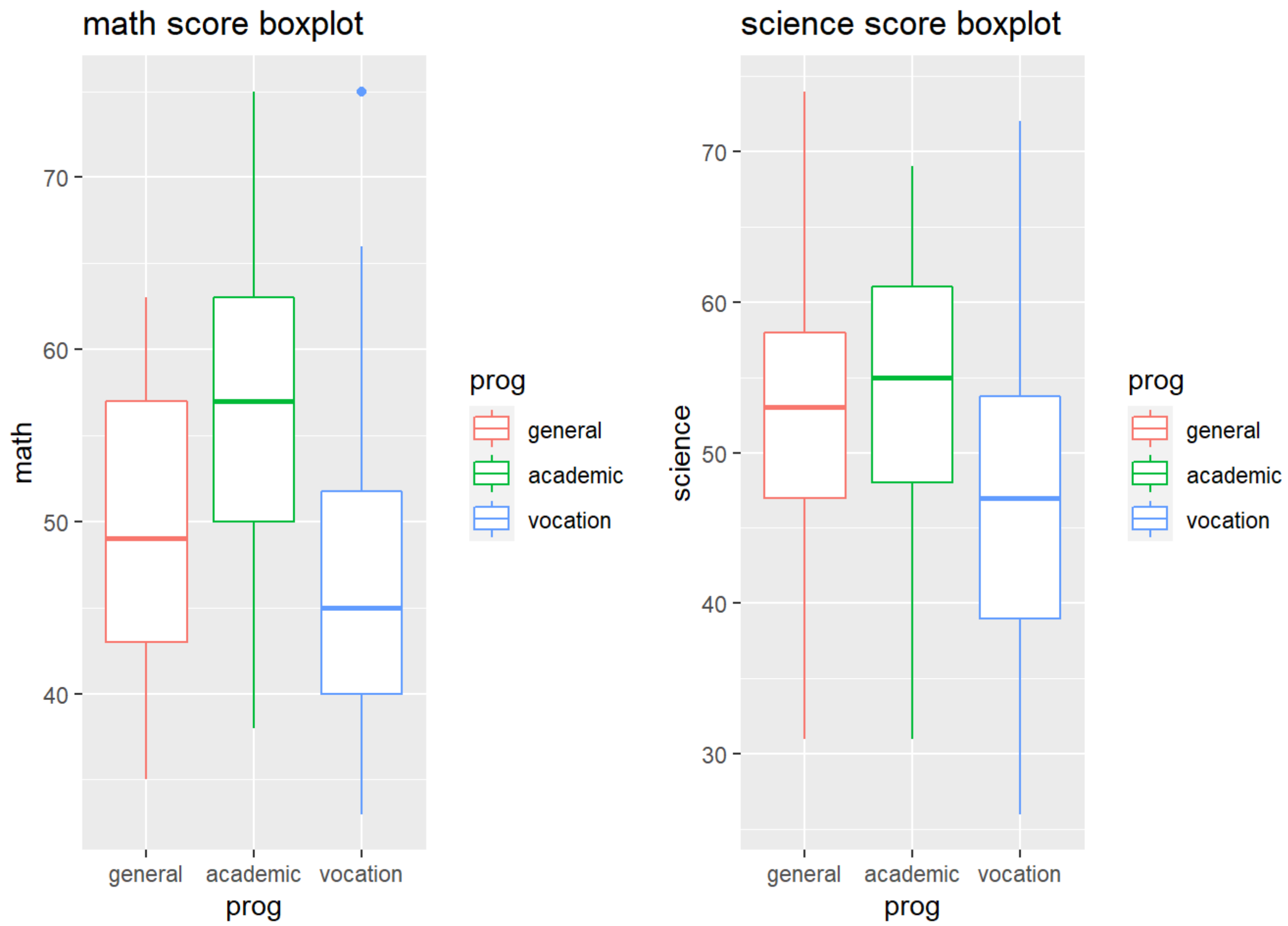
Therefore, we can get answer:
$P(play\ golf\ |sunny, hot, high, windy)$
= $\frac{3/9 \cdot 2/9 \cdot 3/9 \cdot 3/9 \cdot 9/14}{3/9 \cdot 2/9 \cdot 3/9 \cdot 3/9 \cdot 9/14 + 2/5 \cdot 2/5 \cdot 4/5 \cdot 3/5 \cdot 5/14}$
$\approx \frac{0.0053}{0.0053+0.0274} \approx 0.1621$

Problem 4 naive Bayes classifier

a. Plot the side by side boxplots for **math** and **science** scores with respect to different program choices using the package ggplot2.

```
a1 <- ggplot ( data = student , aes(x = prog , y = math , color = prog )) + geom_boxplot()+ y
lab ("math")+ xlab ("prog") + ggtitle("math score boxplot")
a2 <- ggplot ( data = student , aes(x = prog , y = science , color = prog )) + geom_boxplot()
+ ylab ("science")+ xlab ("prog") + ggtitle("science score boxplot")

grid.arrange(a1, a2, nrow = 1, ncol = 2)
```



b. Fit the Naive Bayes to classify the program choice prog using the three categorical variables gender **female**, school type **schtyp** and social economic status **ses**.

```
student_sub <- select(student,female, schtyp, ses, prog)

classifier <- naiveBayes(prog ~ ., data = student_sub )
classifier
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##  general academic vocation
##    0.225    0.525    0.250
##
## Conditional probabilities:
##           female
## Y               male    female
##   general  0.4666667 0.5333333
##   academic 0.4476190 0.5523810
##   vocation 0.4600000 0.5400000
##
##           schtyp
## Y             public   private
##   general  0.8666667 0.1333333
##   academic 0.7714286 0.2285714
##   vocation 0.9600000 0.0400000
##
##           ses
## Y                low    middle      high
##   general  0.3555556 0.4444444 0.2000000
##   academic 0.1809524 0.4190476 0.4000000
##   vocation 0.2400000 0.6200000 0.1400000
```

a. What is the fraction of the students who choose the academic program?
   academic program: 0.525

b. For a female student, what is the probability that she chooses the vocation program?
   P(vocation | female) = 0.2477064

```
student_sub <- select(student,female, prog)
classifier <- naiveBayes(female ~ ., data = student_sub )
classifier
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##   male female
##  0.455  0.545
##
## Conditional probabilities:
##         prog
## Y           general  academic  vocation
##   male   0.2307692 0.5164835 0.2527473
##   female 0.2201835 0.5321101 0.2477064
```

c. Are the “social economic status” and “whether the student chooses the academic program” independent from each other?

```
student_sub <- select(student,ses, prog)
classifier <- naiveBayes(ses ~ ., data = student_sub )
classifier
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##    low middle   high
##  0.235  0.475  0.290
##
## Conditional probabilities:
##         prog
## Y           general  academic  vocation
##   low    0.3404255 0.4042553 0.2553191
##   middle 0.2105263 0.4631579 0.3263158
##   high   0.1551724 0.7241379 0.1206897
```

No. If “social economic status” and “whether the student chooses the academic program” are independent from each other,
P(ses and academic) = $P(ses) \cdot P(academic)$ or P(ses | academic) = P(ses)

However, by checking the result with P(ses=low | academic):
P(ses=low | academic) = 0.1809524
P(ses=low) = 0.235
It’s not match for these two probability. And so do other features.
Therefore, “social economic status” and “whether the student chooses the academic program” are **not independent** from each other.