HomeWork five

Yi Chen(yc3356)

December 2, 2017

HomeWork Five

part one

problem 1

```
setwd("C:/Users/cheny/Desktop/study/statistical computing and intro to data science/homework/
homework 5")

nodes <- read.csv('ckm_nodes.csv',header = TRUE)
dim(nodes)
```

```
## [1] 246  13
```

(a) How many doctors prescribing tetracycline in each month of the study?

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
nodes %>%
        select(adoption_date) %>%
        filter(adoption_date != Inf) %>%
        group_by(adoption_date) %>%
        summarize(number = n())
```

```
## # A tibble: 17 x 2
##    adoption_date number
##            <dbl>  <int>
##  1             1     11
##  2             2      9
##  3             3      9
##  4             4     11
##  5             5     11
##  6             6     11
##  7             7     13
##  8             8      7
##  9             9      4
## 10            10      1
## 11            11      5
## 12            12      3
## 13            13      3
## 14            14      4
## 15            15      4
## 16            16      2
## 17            17      1
```

(b) How many never prescribed it during the study?

```
# solution one
nodes %>%
        select(adoption_date) %>%
        filter(adoption_date == Inf) %>%
        summarize(number = n())
```

```
##   number
## 1     16
```

```
# solution two

sum(nodes$adoption_date==Inf,na.rm = TRUE)
```

```
## [1] 16
```

(c) How many are NAs?

```
# solution one
nodes %>%
        select(adoption_date) %>%
        filter(is.na(adoption_date) == TRUE) %>%
        summarize(number = n())
```

```
##   number
## 1    121
```

```
# solution two
sum(is.na(nodes$adoption_date) == TRUE)
```

```
## [1] 121
```

** note ** this three problem can also be solve together

```
nodes %>%
        select(adoption_date) %>%
        group_by(adoption_date) %>%
        summarize(number = n())
```

```
## # A tibble: 19 x 2
##    adoption_date number
##            <dbl>  <int>
##  1             1     11
##  2             2      9
##  3             3      9
##  4             4     11
##  5             5     11
##  6             6     11
##  7             7     13
##  8             8      7
##  9             9      4
## 10            10      1
## 11            11      5
## 12            12      3
## 13            13      3
## 14            14      4
## 15            15      4
## 16            16      2
## 17            17      1
## 18           Inf     16
## 19            NA    121
```

problem 2

```
# solution one

nodes <- nodes %>% mutate(index_number=(is.na(adoption_date))) %>%
                   filter(index_number == FALSE)
dim(nodes)
```

```
## [1] 125  14
```

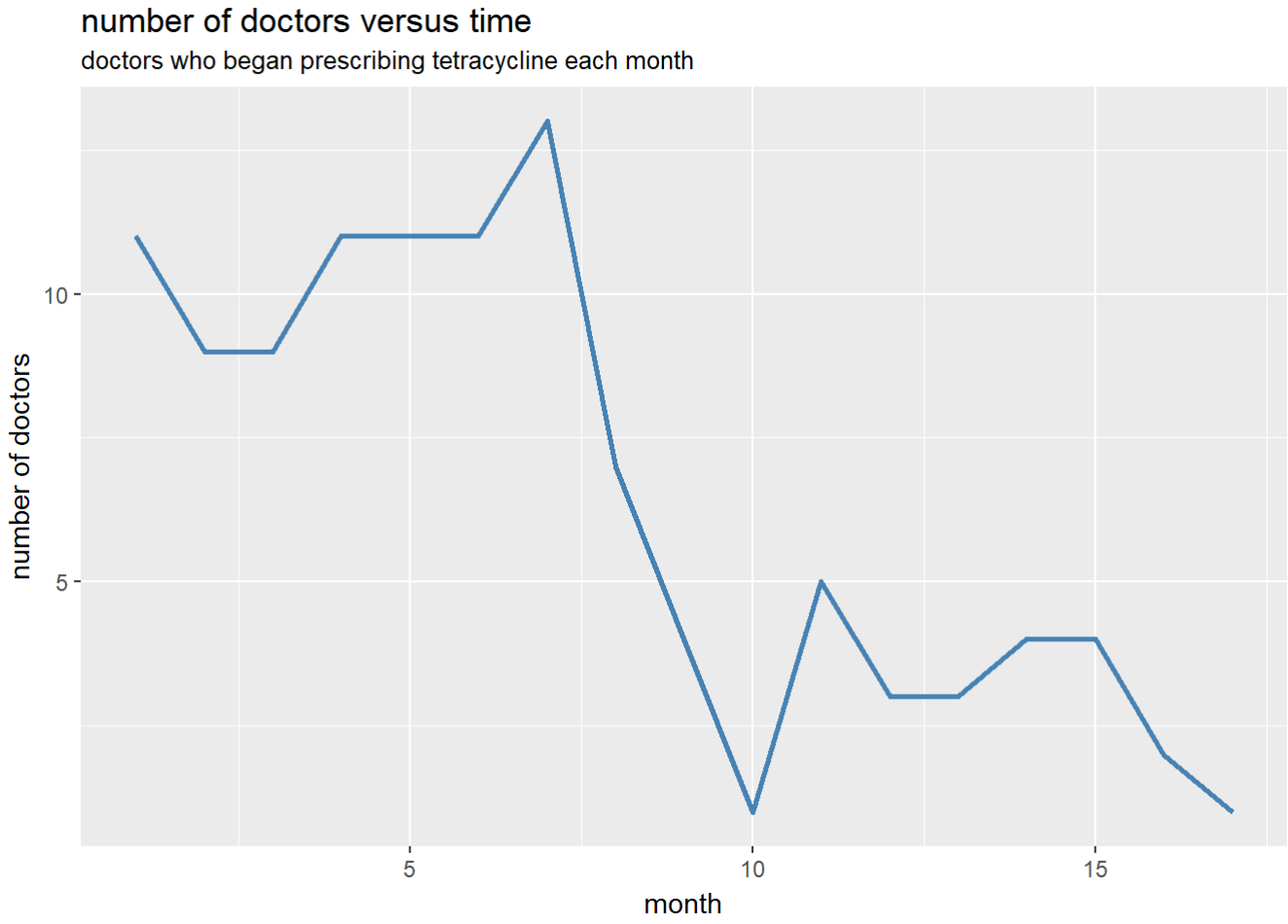
```
# solution two

# nodes$index_number = is.na(nodes$adoption_date))
# nodes = nodes[nodes$index_number == FALSE,]
```

problem 3

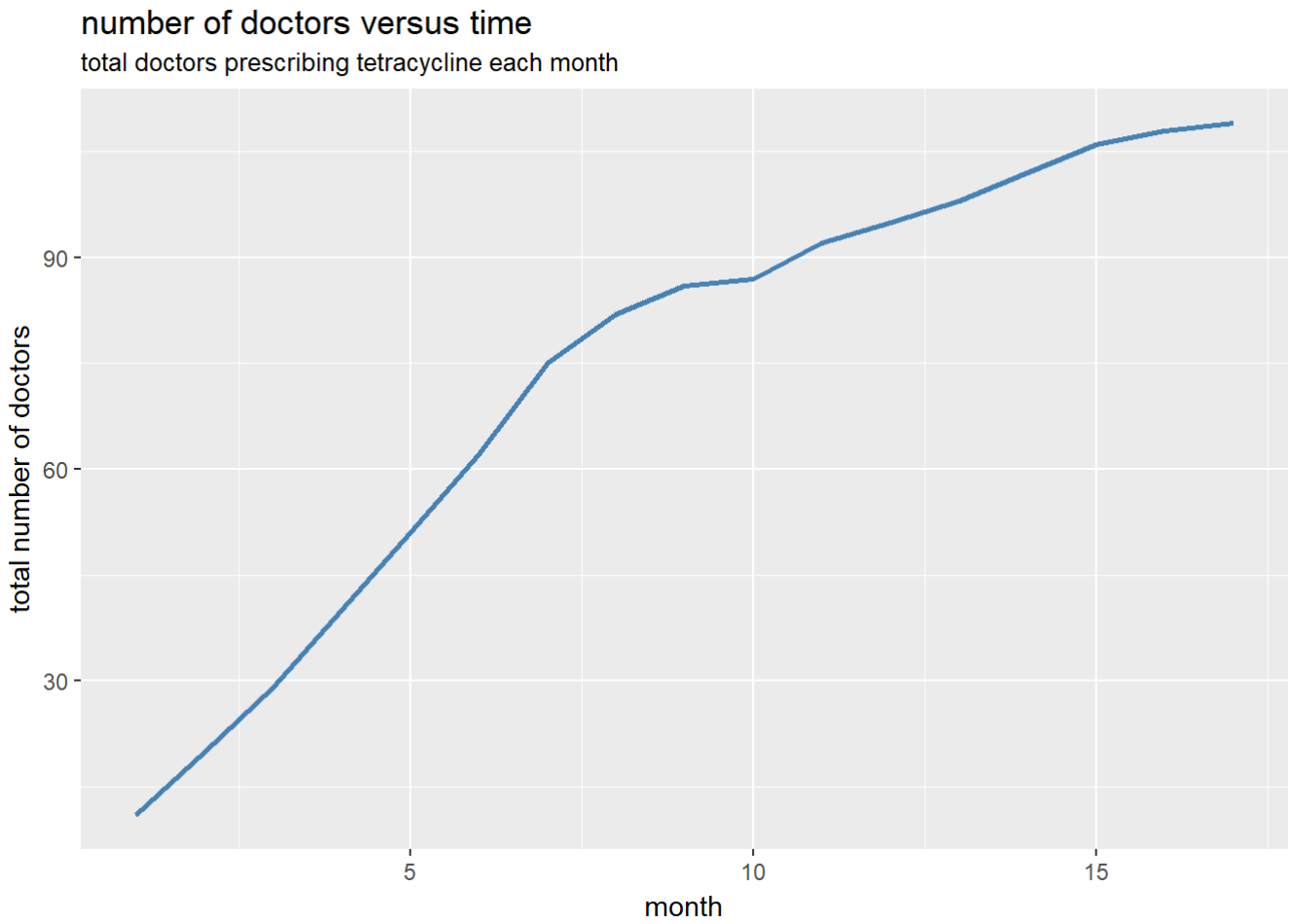
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
# copy from the problem one (a)
plot_data1 <- nodes %>%
        select(adoption_date) %>%
        filter(adoption_date != Inf) %>%
        group_by(adoption_date) %>%
        summarize(number = n())

ggplot(data=plot_data1)+
        geom_line(aes(x=adoption_date,y=number),lwd=1,col='steelblue')+
        labs(title='number of doctors versus time',subtitle='doctors who began prescribing te
tracycline each month',x='month',y='number of doctors')
```




```
plot_data1 = plot_data1 %>%
             mutate(cumsum = cumsum(number))

ggplot(data=plot_data1)+
        geom_line(aes(x=adoption_date,y=cumsum),lwd=1,col='steelblue')+
        labs(title='number of doctors versus time',subtitle='total doctors prescribing tetrac
ycline each month',x='month',y='total number of doctors')
```



problem 4

```
# begun by 2
begun_by_2 <- nodes$adoption_date <= 2
begun_by_2 <- which(begun_by_2)
length(begun_by_2)
```

```
## [1] 20
```

```
head(begun_by_2)
```

```
## [1]  1 10 13 20 27 45
```

```
# after 14
begun_after_14 <- nodes$adoption_date > 14
begun_after_14 <- which(begun_after_14)
length(begun_after_14)
```

```
## [1] 23
```

```
head(begun_after_14)
```

```
## [1]  7 14 16 17 30 39
```

problem 5

```
adopters <- function(month,not.yet=FALSE){
        if(not.yet==FALSE){
                return(sum(nodes$adoption_date == month))
        }

        if(not.yet==TRUE){
                return(sum(nodes$adoption_date > month))
        }
}

adopters(2)
```

```
## [1] 9
```

```
adopters(14,not.yet = TRUE)
```

```
## [1] 23
```

part two

problem 6

```
network <- as.matrix(read.table('ckm_network.txt',header = FALSE))

dim(network)
```

```
## [1] 246 246
```

```
nodes_old <- read.csv('ckm_nodes.csv',header = TRUE)
index_number=(is.na(nodes_old$adoption_date)==FALSE)

network <- network[index_number,index_number]
dim(network)
```

```
## [1] 125 125
```

```
colnames(network) <- 1:125
row.names(network) <- 1:125
```

problem 7

```
number_of_connect <- colSums(network)
number_of_connect[41]
```

```
## 41
##  3
```

problem 8

```
logical_vector <- (network[,37]==1 & nodes$adoption_date <= 5 )
sum(logical_vector)
```

```
## [1] 3
```

```
sum(logical_vector)/number_of_connect[37]
```

```
##  37
## 0.6
```

problem 9

```
count_peer_pressure <- function(doctor,month){
        return(sum(network[,doctor]==1 & nodes$adoption_date <= month))
}

count_peer_pressure(doctor = 37,month = 5)
```

```
## [1] 3
```

problem 10

```
prop_peer_pressure <- function(doctor,month){
        ifelse(sum(network[,doctor]==1)==0,NaN,count_peer_pressure(doctor = doctor, month = m
onth)/number_of_connect[doctor])
}

prop_peer_pressure(doctor = 37,month = 5)
```

```
## [1] 0.6
```

```
prop_peer_pressure(doctor = 102,month = 4)
```

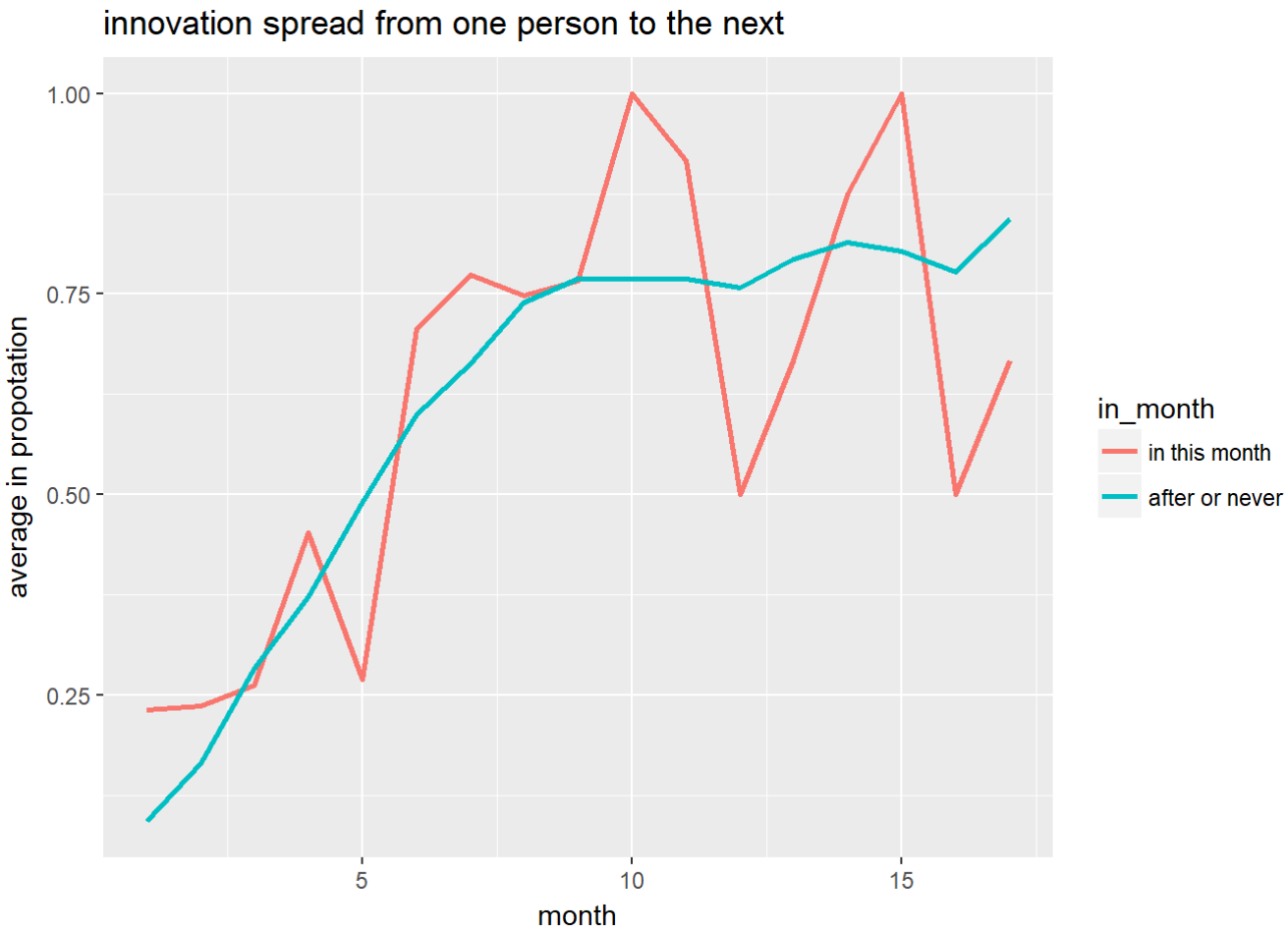
```
## [1] NaN
```

problem 11

```
average <- function(month){
        vector <- vector(length = 2)

        # find out the doctor index of each subproblem
        begun_in_month <- which(nodes$adoption_date == month)
        begun_after_never <- which(nodes$adoption_date >= month)

        # calculate the first element in the vector

        average_in_month <- mean(sapply(begun_in_month,FUN = prop_peer_pressure,month=month),
na.rm = TRUE)

        average_after_month <- mean(sapply(begun_after_never,FUN = prop_peer_pressure,month=m
onth),na.rm = TRUE)



        vector <- c(average_in_month,average_after_month)

        return(vector)
}
```

problem 12

```
average_in_month <- sapply(1:17,FUN = average)[1,]
average_after_month <- sapply(1:17,FUN = average)[2,]

in_month <- data.frame(average=average_in_month,in_month=rep('in this month',17),month=1:17)
after_never <- data.frame(average=average_after_month,in_month=rep('after or never',17),month
=1:17)
plot_data2 <- rbind(in_month,after_never)

ggplot(data=plot_data2)+
        geom_line(aes(x=month,y=average,col=in_month),lwd=1)+
        geom_line(aes(x=month,y=average,col=in_month),lwd=1)+
        labs(y='average in propotation',x='month',title='innovation spread from one person to
 the next')
```


innovation spread from one person to the next
1.00
0.75
0.50
0.25
average in propotation
5
10
15
month
in_month
in this month
after or never