ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

TOÁN RỜI RẠC 1
BÁO CÁO BÀI TẬP LỚN ĐỀ TÀI 56

Thống kê
&
Phân tích dữ liệu bằng R

GVHD: NGUYỄN AN KHƯƠNG
Mã nhóm: 23
Nhóm SV: Nguyễn Công Anh - 1710477
Phan Tấn Quốc - 1712855
Nguyễn Đăng Hà Nam - 1710195
Đinh Gia Khiêm - 1711747
Lê Thành Nhơn - 1712516

TP. HỒ CHÍ MINH, THÁNG 12/2017

Mục lục

Nhật ký làm việc nhóm

(*). Thời gian nhận đề bài tập lớn: *November 25, 2017 5:00 pm*
(*). Due to: *December 25, 2017 5:00 pm*
(*). Thời gian hoàn thành: *December 22, 2017*

(1). Phân công nhiệm vụ:
- Đinh Gia Khiêm: Lấy dataset mã đề 56 từ tập dữ liệu, làm bài 2.
- Lê Thành Nhơn: Làm bài 3, 4.
- Nguyễn Công Anh: Làm bài 5, 6, 7
- Nguyễn Đăng Hà Nam: Làm bài 8, 10.
- Phan Tấn Quốc: Làm bài 9, 11.
(2). Cách tổ chức làm việc nhóm:
- Về tập dữ liệu: Tập dữ liệu được lưu trên Driver mà mỗi thành viên trong nhóm đều có thể truy cập đến bất cứ lúc nào
- Về nơi thảo luận, hình thức thỏa luận nhóm:
Online: Nhóm thảo luận về các vấn đề của các bài tập trong nhóm kín trên Facebook, trên Asana, và nhóm chat riêng của nhóm.
Offline: Nhóm thường tập trung các thành viên vào các ngày thứ 6, thứ 7, Chủ nhật hàng tuần để cùng thảo luận trực tiếp.

(3). Quá trình xử lý dữ liệu và giải quyết các yêu cầu:
- Nhóm làm việc trong 4 tuần, lần lượt giải quyết các yêu cầu của bài tập lớn và thu về các kết quả
- Nhóm làm việc dựa trên:
- Cơ sở dữ liệu: Dữ liệu về hệ thống xe buýt của Tp.Hồ Chí Minh.
- Cơ sở lý thuyết: R và LaTex
- Tìm hiểu cơ sở lý thuyết trên Google.
- Các thành viên tích cực đưa ra các câu hỏi để cả nhóm cùng tìm ra hướng giải quyết

(4). Kết quả sơ lược của quá trình làm việc nhóm:
- Nhóm đã hoàn thành hầu hết các yêu cầu của bài tập lớn và thu được các kết quả thực nghiệm.
- Các thành viên trong nhóm tích cực tìm hiểu và giải quyết các bài tập cá nhân đồng thời hỗ trợ các thành viên khác, đảm bảo mỗi thành viên đều có thể hoàn thành bài tập được giao đồng thời có thể nắm bắt được các bài tập của các thành viên khác.
- Các thành viên đều nắm được yêu cầu và bản chất của các bài toán, các dòng lệnh R, được rèn luyện khả năng làm việc nhóm, khả năng lập trình cơ bản trên R và LaTex

1 Giới thiệu bài toán

Ta cần phải phân tích dữ liệu để cung cấp các thông tin xác thực, trực quan, mô tả cụ thể, dễ hiểu vấn đề đang phân tích để phục vụ nghiên cứu khoa học. Đặc biệt trong các vấn đề kinh tế-xã hội và khi nghiên cứu số lớn chúng ta cần phải quan tâm đến các công cụ kỹ thuật về phân tích số liệu và biểu đồ.

Phân tích số liệu và biểu đồ thường được tiến hành bằng các phần mềm thông dụng như SAS, SPSS, Stata, Statistica, và S-Plus. Đây là những phần mềm được các công ti phần mềm phát triển và giới thiệu trên thị trường khoảng ba thập niên qua, và đã được các trường đại học, các trung tâm nghiên cứu và công ti kĩ nghệ trên toàn thế giới sử dụng cho giảng dạy và nghiên cứu. Nhưng vì chi phí để sử dụng các phần mềm này tương đối đắt tiền (có khi lên đến hàng trăm ngàn đô-la mỗi năm). Do đó, các nhà nghiên cứu thống kê trên thế giới đã hợp tác với nhau để phát triển một phần mềm mới, với chủ trương mã nguồn mở, sao cho tất cảcác thành viên trong ngành thống kê học và toán học trên thế giới có thể sử dụng một cách thống nhất và hoàn toàn miễn phí.

Năm 1996, trong một bài báo quan trọng về tính toán thống kê, hai nhà thống kê học Ross Ihaka và Robert Gentleman [lúc đó] thuộc Trường đại học Auckland, New Zealand phát hoạ một ngôn ngữ mới cho phân tích thống kê mà họ đặt tên là R. Nói một cách ngắn gọn, R là một phần mềm sử dụng cho phân tích thống kê và vẽ biểu đồ. Thật ra, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

Sơ lược về đề tài : Hiện nay trong nước, đặt biệt là thành phố Hồ Chí Minh. Lượng người sử dụng phương tiện công cộng, cụ thể là xe buýt ngày một tăng cao. Chính vì thế, để đáp ứng nhu cầu người dân, số lượng xe buýt đã được gia tăng đáng kể, đồng thời kéo số lượng trạm xe buýt tăng theo. Nhưng xe buýt là một trong những nguyên nhân gây kẹt xe hàng đầu: đón trả khách, kích thước lớn, tần suất cao,... Những năm gần đây, Chính Phủ đang cố gắng giảm thiểu xe máy, điều này cũng khiến cho lượng xe buýt ngày một nhiều. Vì thế vấn đề kẹt xe hay hiệu suất sử dụng trạm trở nên đáng quan tâm hơn bao giờ hết. Các giải pháp hiện nay được đưa ra chỉ là tăng cường phát triển dịch vụ đường sắt, phân luồng giao thông, tăng giá xe hay giải quyết vấn đề lấn chiếm lòng lề đường, v.v... Các giải pháp này chỉ mang tính tạm thời, giảm quyết được phần nào vấn nạn kẹt xe cục bộ hoặc là cần có thời gian triển khai lâu dài, không thể áp dụng trong tương lai gần. Vì thế, cần có một giải pháp để có thể dự đoán điểm kẹt xe để thông báo kịp thời nhằm điều tiết giao thông hiệu quả. Ngoài ra, việc này cũng chỉ ra được các trạm ít sử dụng để đưa ra đề xuất cắt bỏ để giảm chi phí.

Trong bài tập lớn này, sinh viên được yêu cầu làm một số phép tính toán, thống kê dựa trên dữ liệu thực về xe buýt tại TP. HCM. Qua đó tìm các điểm hay bị kẹt xe để giải quyết vấn đề kẹt xe trong thực tiễn ở Tp.Hồ Chí Minh

2 Cơ sở lý thuyết

2.1 Thống kê mô tả

Nói đến thống kê mô tả là nói đến việc mô tả dữ liệu bằng các phép tính và chỉ số thống kê thông thường mà chúng ta đã làm quen qua từ thuở trung học như số trung bình (mean), số trung vị (median), số lớn nhất (max), số nhỏ nhất (min), phương sai (variance), độ lệch chuẩn (standard deviation)...

Trong đó ta làm quen các định nghĩa chưa biết :

- Phương sai của một biến ngẫu nhiên là một độ đo sự phân tán thống kê của biến đó, nó hàm ý các giá trị của biến đó thường ở cách giá trị kỳ vọng bao xa.

- Độ lệch chuẩn, hay độ lệch tiêu chuẩn, là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số. Có thể tính ra độ lệch chuẩn bằng cách lấy căn bậc hai của phương sai.

- số trung vị (tiếng Anh: median) là một số tách giữa nửa lớn hơn và nửa bé hơn của một mẫu, một quần thể, hay một phân bố xác suất. Nó là giá trị giữa trong một phân bố, mà số số nằm trên hay dưới con số đó là bằng nhau. Điều đó có nghĩa rằng $1/2$ quần thể sẽ có các giá trị nhỏ hơn hay bằng số trung vị, và một nửa quần thể sẽ có giá trị bằng hoặc lớn hơn số trung vị.

2.2 Công cụ R

Như đã nói ở trên, R là một công cụ miễn phí dùng để phân tích dữ liệu. Chúng ta có thể sử dụng R để thực hiện các phép toán từ đơn giản đến phức tạp. Những bài toán tiêu biểu: các phép kiểm định thống kê, tính toán trên ma trận, hồi quy tuyến tính, gom cụm dữ liệu, bài toán phân lớp... Và vì R là một ngôn ngữ nên chúng ta có thể viết ứng dụng trên R để giải quyết các vấn đề cụ thể.

- Các hàm của R để tính toán thống kê mô tả:

```
> option (width=100)
# chuyển directory
> setwd ("C:/works/stats")

# đọc dữ liệu vào R
> igfdata <- read.table ("igf.txt", hearder = TRUE, na.string = ".")
> attach (igfdata)

# xem xét các cột số trong dữ liệu
> names (igfdata)
hoặc
> igfdata

# tính trung bình
> mean (age)

# phương sai và độ lệch chuẩn
> var (age)
> sd (age)
```

3 Kết quả phân tích dữ liệu

3.1 Tập dữ liệu

- Tập dataset của mã đề 56 chứa:
- Dữ liệu hành trình một ngày của xe buýt gồm:
393 file JourneyCell40
393 file JourneyCell60
393 file JourneyGPS
- Dữ liệu tuyến xe buýt gồm:
42 file RouteCell40
42 file RouteCell60
42 file RouteGPS

- Dữ liệu hành trình một ngày của xe buýt: Dữ liệu mô tả sự di chuyển trong một ngày của xe buýt TP HCM (tạm gọi là hành trình). Hành trình được biểu diễn dưới dạng tập hợp các điểm GPS theo tọa độ Latitude và Longitude. Các điểm này được hộp đen xe buýt ghi lại. Dữ liệu này bao gồm ba cột:
• Lat: Vĩ độ
• Long: Kinh độ
• Receiving time: Thời gian nhận được tín hiệu GPS mà hộp đen xe buýt gửi lên, tính theo hệ Unix Epoch.
- Ví dụ:

```
> library(readxl) # Khai bao thu vien readxl
> Jc <- NULL # Khoi tao data.frame Jc la rong
> setwd("C:/BTLMade56/DTB56JourneyCell40") # Dan R den thu muc chua file JourneyCell40
> Jc40 <- list.files(pattern = "*.xlsx")
> Jc <- lapply(Jc40,read_excel) # Tao 1 list chua cac file JourneyCell40
> A <- data.frame(Jc[[1]]) # Gan du lieu cua xe thu nhat cho A
```

- Xem thử dữ liệu trong A:

```
> options(max.print=60)
> A
```

- Kết quả:

Lat	Long	Send.time
1	1061	738	1472749255
2	1061	738	1472749265
3	1061	738	1472749275
4	1061	738	1472749285
5	1061	738	1472749295
6	1061	738	1472749305
7	1061	738	1472749315
8	1061	738	1472749325
9	1061	738	1472749335
10	1061	738	1472749345
11	1061	738	1472749355

12	1061	738	1472749365
13	1061	738	1472749375
14	1061	738	1472749385
15	1061	738	1472749395
16	1061	738	1472749405
17	1061	738	1472749415
18	1061	738	1472749425
19	1061	738	1472749435
20	1061	738	1472749445

```
> A <- na.omit(A) # Loai bo nhung dong co gia tri NA
> save(A, file="A.rda") # Luu A duoi dang R
> attach(A) # Dan cho R biet ta muon xu ly A
```

- Dữ liệu tuyến xe buýt:
Dữ liệu mô tả các tuyến xe buýt của TPHCM. Các tuyến được biểu diễn bởi tập hợp các trạm, mỗi trạm được miêu tả bằng một điểm GPS theo tọa độ Latitude và Longitude. Một dữ liệu tuyến bao gồm các cột sau:
• Route_Id:Số hiệu của tuyến
• Station_Id: Số hiệu của trạm (trong toàn bộ các trạm TPHCM)
• Station_Code: Mã trạm
• Station_Direction: Hướng trạm (0 – Trạm nằm trên chiều đi, 1 – Trạm nằm trên chiều về)
• Station_Order: Thứ tự của trạm trong mỗi chiều của tuyến
• Station_Name: Tên trạm
• Station_Address: Địa chỉ trạm
• Lat: Vĩ độ • Lng: Kinh độ
• Polyline: Tập hợp các điểm GPS biểu diễn lộ trình di chuyển từ trạm liền trước tới trạm hiện tại.
• Distance: Khoảng cách từ trạm hiện tại tới trạm liền trước (tính theo mét).

- Các thư mục Route* chứa các thông tin về tất cả các tuyến xe buýt ở TP. HCM.
Các thư mục Journey* chứa các thông tin về hành trình của một số xe buýt tại một ngày.
Các thư mục *GPS chứa dữ liệu theo đơn vị GPS.
Các thư mục *Cell* chứa dữ liệu khi đã mapping vào lưới 40 hoặc 60 mét, tức là chia TP. HCM thành các lưới ô vuông có cạnh là 40m hoặc 60m và điểm GPS được chuyển thành dòng, cột của lưới.

3.2 Kết quả phân tích

(1). Trích tập dataset ứng với mã đề 56
(2). Xác định số lượng xe buýt trong tập mẫu:

Code:
- Truy cập đến thư mục chưa thông tin của tất cả các xe, gộp các file thành 1 list, rồi đếm số lượng file trong list ứng với số lượng xe buýt

```
# Gop cac file JourneyCell40 thanh mot list la Jc
> library(readxl)
```

```
> Jc <- NULL
> setwd("C:/BTLMade56/DTB56JourneyCell40")
> Jc40 <- list.files(pattern = "*.xlsx")
> Jc <- lapply(Jc40,read_excel)
# Dem so xe
> length(Jc)
[1] 393
```

Kết quả: Có 393 xe buýt trong tập dataset.

(3).Xác định số lượng tuyến trong tập mẫu

Code:
- Truy cập đến thư mục chưa thông tin của tất cả các tuyến, gộp các file thành 1 list, rồi đếm số lượng file trong list ứng với số tuyến xe buýt.

```
# Gop cac file RouteCell40 thanh 1 list la Rc
> library(readxl)
> Rc <- NULL
> setwd("C:/BTLMade56/DTB56RouteCell40")
> Rc40 <- list.files(pattern = "*.xlsx")
> Rc <- lapply(Rc40,read_excel)
# Dem so tuyen
> length(Rc)
[1] 42
```

Kết quả: Có 42 tuyến xe buýt trong tập dataset

(4). Nhóm câu hỏi liên quan đến hành trình của một tuyến xe buýt
a) Số lượng cell mà một tuyến xe buýt đi qua
b) Tổng quãng đường di chuyển của một tuyến xe buýt
c) Danh sách cell mà một tuyến xe buýt chứa nhiều lần
d) Khoảng cách trung bình giữa các trạm liên tiếp trên hành trình của tuyến xe buýt là bao nhiêu?

Code:
- Trong list chứa tất cả các file của các tuyến xe buýt, chọn ra data.frame có mã tuyến 55.

```
> vidu <- data.frame(Rc[55-46])
# Cau a) So cell ma tuyen nay di qua
> attach(vidu)
> nrow(unique(vidu[,c(8,9)]))
[1] 76

# Cau b) Tinh tong quang duong di chuyen cua tuyen ngau nhien nay (km)
> x <- subset(Distance, Distance>0)
> sum(as.numeric(x))/1000
[1] 36.344

# Cau c) Danh sach cac cell ma tuyen xe nay chua nhieu lan
> latlong <- vidu[,c(8,9)]
> X <- NULL
> A <- NULL
```

```
> B <- NULL

# Dem so luong lap lai cua moi cell
> for (i in 1:nrow(latlong))
+ X[i] <- nrow(subset(latlong, Lat==Lat[i] & Lng==Lng[i]))
> for (i in 1:length(X))
{if (X[i]>1)
     { A <- unique(rbind(A,Lat[i]))
       B <- unique(rbind(B,Lng[i]))
     }
     else{}
}
> cbind.data.frame(A,B)
     A   B
1   731 611
2   733 607
3   735 606
4   730 681
5   729 680
6   720 674
7   717 662
8   716 658
9   711 659
10  707 666
11  706 672
12  693 673
13  686 664
14  680 633

# Cau d) Tinh khoang cach trung binh giua cac tram lien tiep tren hanh trinh cua tuyen
    xe buyt (km)
> x <- subset(Distance, Distance>0)
> mean(as.numeric(x))/1000
[1] 0.4038222
```

Kết quả:
- Tuyến này đi qua 76 cell khác nhau
- Tổng quãng đường di chuyển của tuyến này là 36.344 km
- Các cell mà tuyến này chứa nhiều lần là (A[i],B[i])
- Khoảng cách trung bình giữa các trạm liên tiếp là 0.4038222 km

(5). Nhóm câu hỏi liên quan đến một tập các tuyến xe buýt
a) Số lượng cell mà một tuyến đi qua
b) Số lượng tuyến xe đi qua một cell cho trước
c) Tuyến nào dài nhất
d) Tuyến nào dài nhì
e) Danh sách các tuyến thuộc một phần ba đầu theo thứ tự chiều dài tuyến giảm dần
f) Tuyến nào chứa nhiều cell nhất
g) Tuyến nào chứa nhiều cell nhì
h) Danh sách các tuyến thuộc một phần ba đầu theo thứ tự số lượng cell đi qua giảm dần
i) Xác định phổ phân bố theo số lần chứa cell của các tuyến xe buýt
j) Danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số

lượng tuyến xe buýt đi qua mỗi cell trong nhóm là như nhau.
k) Phân tích khoảng cách giữa 2 trạm liên tiếp trên mỗi tuyến: tính giá trị trung bình, phương sai, độ lệch chuẩn, giá trị trung vị, giá trị lớn nhất, nhỏ nhất. Hãy vẽ histogram cho biến ngẫu nhiên này. Đưa ra các nhận xét.

Code:

```
# Gop cac data.frame trong list Rc thanh 1 data.frame duy nhat la "gop".
> gop <- NULL
> for (i in 1:length(Rc))
+ gop <- rbind.data.frame(gop, (Rc[[i]]))

# Tao ra mot data.frame gom cac cell ma moi tuyen di qua
> Cell <- NULL
> for (i in 1:length(Rc))
+ Cell <- rbind.data.frame(Cell, unique(Rc[[i]][ ,c(1,8,9)]))
```

```
# Cau a) So luong cell ma mot tuyen di qua
> attach(Cell)
> dem <- data.frame(table(Route_Id))
> View(dem)

# Cau f) Tim tuyen chua nhieu cell nhat
> for (i in 1:nrow(dem))
+ if (Freq[i]==max(Freq)) {print(Route_Id[i])} else{}
[1] 75

# Cau g) Tim tuyen chua nhieu cell nhi
> sapxep <- sort(Freq)
> for (i in 1:nrow(dem))
+ if (Freq[i]==sapxep[nrow(dem)-1]) {print(Route_Id[i])} else{}
[1] 49

# Cau h) Danh sach cac tuyen thuoc mot phan ba dau theo theo thu tu so luong cell di
    qua giam dan
> round((1/3)*42)-1
[1] 13
> for (i in 1:nrow(dem))
+ if (Freq[i] <= max(Freq) & Freq[i] >= sapxep[nrow(dem)-13]) {print(Route_Id[i])}
    else{}
[1] 101 105 49 50 51 52 53 54 57 62 75 76 79 91

# Cau i) Xac dinh pho phan bo theo so lan chua cell cua cac tuyen xe buyt
> attach(dem)
> plot(dem)
```

```
# Cau b) So luong tuyen xe di qua mot cell cho truoc
> attach(Cell)
> x <- sample(Lat,1)
> y <- sample(Lng,1)
> table(Lat==x & Lng==y)
```

```
FALSE      TRUE
value1     value2
# value2 la so tuyen di qua cell (x,y)
```

```
# Tim so tuyen xe di qua mot cell bat ki
> attach(Cell)
> a <- Lat
> b <- Lng
> N <- NULL
> for (i in 1:nrow(Cell))
+ N[i] <- nrow(subset(Cell, Lat==a[i] & Lng==b[i]))
> Num <- data.frame(N)
> View(Num)
> attach(Num)

# Cau j) Danh sach nhom cell co so luong lon nhat ma trong do cac cell duoc gom nhom
    sao cho so luong tuyen xe buyt di qua moi cell trong nhom la nhu nhau
> khaosat <- data.frame(table(Num))
> attach(khaosat)
> for (i in 1:nrow(khaosat))
+ if (Freq[i]==max(Freq)) {z <- i} else{}
> A <- NULL
> B <- NULL
> Vitri <- NULL
> for (i in 1:nrow(Num))
+ if (N[i]==z) {Vitri <- rbind(Vitri, i)} else{}
> for (i in 1:length(Vitri))
  {A[i] <- a[Vitri[i]]
   B[i] <- b[Vitri[i]]}
> danhsach <- cbind.data.frame(A,B)
> options(max.print = 100000)
> unique(danhsach)
       A   B
1    607  435
2    611  438
3    607  443
4    641  451
5    645  452
6    651  454
7    656  456
8    663  458
9    667  460
10   672  461
..............
1383  60  273
1384  54  266
1385  44  253
```

```
# Tinh do dai di chuyen cua moi tuyen xe
> for (i in 1:length(Rc))
{ attach(Rc[[i]])
```

```
  X[i] <- round(sum(as.numeric(subset(Distance, Distance>0)))/1000,3)
}
> S <- data.frame(X)
> attach(S)
> sapxep <- sort(X)
> attach(gop)
> RtID <- unique(Route_Id)

# Cau c) Tim tuyen dai nhat
> attach(S)
> for (i in 1:nrow(S))
+ if (X[i]==sapxep[42]) {print(RtID[i])} else{}
[1] "73"

# Cau d) Tim tuyen dai nhi
> for (i in 1:nrow(S))
+ if (X[i]==sapxep[41]) {print(RtID[i])} else{}
[1] "51"

# Cau e) Tim danh sach cac tuyen thuoc mot phan ba dau theo thu tu chieu dai tuyen giam
    dan
> for (i in 1:nrow(S))
    {for (j in (42-13):42) {
    if (X[i]==sapxep[j]) {(print(RtID[i]))} else {}
 }
}
[1] "46" "49" "51" "53" "54" "70" "72"
[7] "73" "75" "76" "78" "79" "91" "106"
```

```
# Cau k) Phan tich khoang cach giua hai tram lien tiep tren moi tuyen

# Lay mot tuyen ngau nhien trong so cac tuyen (tuyen 104)
> vidu <- data.frame(sample(Rc,1))
> attach(vidu)
> x <- as.numeric(subset(Distance, Distance>0))

# Gia tri trung binh, trung vi, gia tri lon nhat, gia tri nho nhat
> summary(x)
   Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
   98.0   276.8    362.5   456.7  537.5    1526.0

# Phuong sai
> var(x)
[1] 82127.34

# Do lech chuan
> sd(x)
[1] 286.5787

# Ve histogram
> hist(x, main='Khoang cach giua cac tram')
```

```
# Nhan xet: " Cac tram cach nhau trung binh la 456.7 m, hai tram xa nhat cach nhau 1526
    m, hai tram gan nhat cach nhau 98m."
```

Kết quả:
- Số lượng cell mà mỗi tuyến đi qua nằm trong data.frame "dem"
- Tuyến chứa nhiều cell nhất là tuyến 75
- Tuyến chứa nhiều cell nhì là tuyến 49
- Danh sách các tuyến thuộc một phần ba đầu theo thứ tự số lượng cell đi qua giảm dần là
101 105 49 50 51 52 53 54 57 62 75 76 79 91
- Phổ phân bố được vẽ bằng R
- Số lượng tuyến đi qua 1 cell cho trước là value2
- Danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lượng tuyến xe buýt đi qua mỗi cell trong nhóm là như nhau là: (A[i],B[i]).
- Tuyến dài nhất là tuyến 73
- Tuyến dài nhì là tuyến 51
- Danh sách các tuyến thuộc một phần ba đầu theo thứ tự chiều dài tuyến giảm dần là:
46 49 51 53 54 70 72 73 75 76 78 79 91 106
- Phân tích khoảng cách giữa hai trạm liên tiếp trên mỗi tuyến:(m)
Giá trị nhỏ nhất: 98 m
Giá trị lớn nhất: 1526 m
Trung bình: 456.7 m
Trung vị: 362.5 m
Phương sai: 82127.34
Độ lệch chuẩn: 286.5787

(6). Nhóm câu hỏi liên quan đến hành trình thực tiễn của một xe buýt
a) Số lượng cell mà một xe buýt đi qua
b) Tổng quãng đường di chuyển của xe buýt
c) Vận tốc trung bình của xe buýt
d) Xác định phổ phân bố theo số lần xe buýt đi qua các cell
e) Danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lần xe buýt đi qua mỗi cell trong nhóm là như nhau

Code:

```
# Gop cac file JourneyCell40 thanh mot list la Jc
> library(readxl)
> Jc <- NULL
> setwd("C:/BTLMade56/DTB56JourneyCell40")
> Jc40 <- list.files(pattern = "*.xlsx")
> Jc <- lapply(Jc40,read_excel)
```

```
# Chon xe buyt 280 trong list Jc (xe buyt thu 113)
> vidu <- data.frame(Jc[113])

# Cau a) Tinh so luong cell ma xe buyt nay di qua
> nrow(unique(vidu[,c(1,2)]))
[1] 635
```

```
# Cau b) Tinh tong quang duong cua xe buyt (km)
> attach(vidu)
> a <- Lat
> b <- Long
> d <- NULL
> for (i in 1:nrow(vidu))
+ d[i] <- (sqrt((a[i+1]-a[i])^2 + (b[i+1]-b[i])^2)*40)/1000
> d[nrow(vidu)] <- 0
> sum(d)
[1] 204.7922
```

```
# Cau c) Tinh van toc trung binh cua xe buyt
> attach(vidu)
> t <- Send.time
> T <- NULL
> for(i in 1:nrow(vidu))
+ if (d[i]!=0) {T[i] <- t[i+1]-t[i]} else{T[i] <- 0}
> v <- sum(d)/(sum(T)/3600)
> v
[1] 27.58041
```

```
# Cau d) Xac dinh pho phan bo theo so lan xe buyt di qua cac cell

# Dem so lan xe buyt di qua mot cell bat ki
> attach(vidu)
> a <- Lat
> b <- Long
> N <- NULL
> for (i in 1:nrow(vidu))
+ N[i] <- nrow(subset(vidu, Lat==a[i] & Long==b[i]))
> Num <- data.frame(N)

# Ve bieu do
> attach(Num)
> plot(Num)
> hist(N, main='Do thi phan bo so lan xe buyt di qua cac cell')
```

```
# Cau e) Danh sach nhom cell co so luong lon nhat ma trong do cac cell duoc gom nhom
    sao cho so lan xe buyt di qua moi cell trong nhom la nhu nhau
> Vitri <- NULL
> A <- NULL
> B <- NULL
> for (i in 1:nrow(Num))
> if (N[i]==1) {Vitri <- rbind(Vitri, i)} else{}
> for (i in 1:length(Vitri))
  {A[i] <- a[Vitri[i]]
   B[i] <- b[Vitri[i]]}
> danhsach <- cbind.data.frame(A,B)
> options(max.print = 100000)
> unique(danhsach)
```

```
     A    B
1   737 1179
2   737 1187
3   738 1188
4   739 1193
5   739 1199
6   732 1224
7   739 1231
8   744 1233
9   763 1241
10  774 1227
.............
361 736 1218
362 738 1199
363 734 1178
```

Kết quả:
- Số cell mà xe buýt này đi qua là 635
- Tổng quãng đường của xe này trong ngày là 204.7922 km
- Vận tốc trung bình của xe này là 27.58 km/h
- Danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lần xe buýt đi qua mỗi cell trong nhóm là như nhau là:(A[i],B[i]).

(7). Nhóm câu hỏi liên quan đến hành trình thực tiễn của một tập các xe buýt
a) Tổng số lần di chuyển qua một cell cho trước
b) Quãng đường di chuyển trung bình của các xe buýt
c) Vận tốc trung bình di chuyển của các xe buýt
d) Số lượng cell trung bình đi qua của xe buýt
e) Xe buýt nào di chuyển dài nhất
f) Xe buýt nào di chuyển dài nhì
g) Danh sách các xe buýt thuộc một phần ba đầu theo thứ tự chiều dài di chuyển giảm dần
h) Danh sách các cell có lượng xe buýt qua nhiều nhất
i) Danh sách các cell có lượng xe buýt qua nhiều nhì
j) Danh sách các cell có lượng xe buýt qua nhiều nhất hoặc nhiều nhì
k) Danh sách các cell thuộc một phần ba đầu theo thứ tự số lượng xe buýt đi qua giảm dần
l) Danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lần xe buýt đi qua mỗi cell trong nhóm là như nhau.
m) Khảo sát thời gian trung bình của một xe buýt để đi từ điểm đầu tới điểm cuối của một tuyết nào đó theo đơn vị phút. Hãy tính giá trị trung bình, trung vị, giá trị lớn nhất, nhỏ nhất. Hãy vẽ histogram cho biến ngẫu nhiên này. Đưa ra các nhận xét.
n) Hãy tìm tuyến (mã tuyến) mà trong tập dataset này có nhiều hành trình nhất chạy trên tuyến đó.
o) Với kết quả ở câu 7n, chọn 3 trạm bất kỳ ở trên tuyến đó không phải là 2 điểm đầu cuối để khảo sát, mỗi trạm chỉ xét theo một chiều (đi hoặc về). Khảo sát thời gian để 2 chiếc xe buýt liên tiếp nhau qua trạm (tính theo phút). Tính trung bình, trung vị, lớn nhất, nhỏ nhất. Hãy vẽ histogram cho biến ngẫu nhiên này và đưa ra các nhận xét.
p) Với kết quả khảo sát ở 7o, giả sử một người hoàn toàn không có thông tin gì về thời gian chạy của xe buýt, đi ra trạm để đón xe buýt vào một thời điểm bất kỳ và ngẫu nhiên (trong khoảng thời gian có xe buýt chạy, tức từ lúc có chuyến sớm nhất cho tới khi cho chuyến muộn nhất theo dataset chạy qua trạm đó). Hãy tính thời gian trung bình mà người đó phải chờ để bắt được xe

buýt.

Code:

```
# Gop cac datsets thanh 1 data.frame duy nhat la "gop".
> library(readxl)
> Jc <- NULL
> setwd("C:/BTLMade56/DTB56JourneyCell40")
> Jc40 <- list.files(pattern = "*.xlsx")
> Jc <- lapply(Jc40,read_excel)
> gop <- NULL
> for (i in 1:393)
+ gop <- rbind.data.frame(gop, (Jc[[i]]))
> View(gop)
```

```
# Cau a) Tong so lan di chuyen qua 1 cell cho truoc
> attach(gop)
> x <- sample(Lat,1)
> y <- sample(Long,1)
> table(Lat=="x" & Long=="y")
  [1] FALSE =value1 TRUE =value2

# TRUE =value2 la so lan di chuyen qua cell (x,y)
```

```
# Cau b) Quang duong di chuyen trung binh cua cac xe trong mot ngay.
> rm(x, y)
> attach(gop)
> a <- Lat
> b <- Long

# Tinh do quang duong qua moi cell (km)
> d <- NULL
> for (i in 1:nrow(gop))
+ d[i] <- (sqrt((a[i+1]-a[i])^2 + (b[i+1]-b[i])^2)*40)/1000
> d[nrow(gop)] <- 0

# Tao ra cac khoang gia tri cho moi xe
> x <- NULL
> y <- NULL
> for (i in 1:393)
+ x[i] <- nrow(Jc[[i]])
> y[1] <- 0
> for (i in 1:393)
+ y[i+1] <- y[i] + x[i]

# Tinh quang duong di chuyen cua moi xe (km)
> xe <- NULL
> for (i in 1:393)
+ xe[i] <- sum(d[(y[i]+1):(y[i+1]-1)])
> S <- data.frame(xe)
> View(S)
```

```
# Tinh quang duong trung binh cua cac xe trong mot ngay (km)
> attach(S)
> mean(xe)
[1] 202.007
```

```
# Cau e,f) Xe nao di chuyen nhieu nhat, xe nao di chuyen nhieu nhi?
> sx <- sort(xe)

# Xe di chuyen nhieu nhat la xe: 353
> for (i in 1:393)
+ if (xe[i] == sx[393]) {print(i+168-1)} else{}
[1] 353

# Xe di chuyen nhieu nhi la xe: 321
> for (i in 1:393)
+ if (xe[i] == sx[392]) {print(i+168-1)} else{}
[1] 321

# Trong tap datasets, xe thu nhat la xe 168.
```

```
# Cau g) Danh sach cac xe buyt thuoc mot phan ba dau theo thu tu chieu dai di chuyen
    giam dan
> round((1/3)*393,0) - 1
[1] 130
> for (i in 1:393)
+ if (xe[i]<= sx[393] & xe[i] >= sx[393-130]) {print(i+168-1)} else{}
[1] 171 176 177 183 184 185 186 188 190 194 195
[12] 199 200 202 204 207 208 212 214 220 227 228
[23] 234 238 239 252 254 255 256 259 260 263 265
[34] 273 276 277 281 291 298 299 303 305 307 308
[45] 309 311 312 313 315 316 319 321 322 323 326
[56] 328 330 331 332 333 334 342 344 353 359 360
[67] 361 366 369 370 372 373 383 384 385 386 388
[78] 389 390 392 395 397 401 407 409 415 422 429
[89] 432 435 456 461 463 464 465 469 471 481 482
[100] 483 500 506 507 509 512 513 515 516 518 519
[111] 520 521 522 524 525 526 528 529 530 531 532
[122] 538 542 543 547 548 549 550 551 552 559
```

```
# Cau c) Tinh van toc trung binh cua cac xe buyt
> rm(a, b, i, sx, x)
> attach(gop)
> t <- 'Send time'
> T <- NULL
> tim <- NULL
> v <- NULL

# Tinh thoi gian di qua moi cell
> for(i in 1:nrow(gop))
+ if (d[i]!=0) {T[i] <- t[i+1]-t[i]} else{T[i] <- 0}
```

```
# Tinh thoi gian di chuyen cua moi xe(h),(khong tinh luc dung yen)
> for(i in 1:393)
+ tim[i] <- sum(T[(y[i]+1):(y[i+1]-1)])/3600

# Tinh van toc cua moi xe
> for(i in 1:393)
+ if(tim[i]!=0) {v[i] <- xe[i]/tim[i]} else {v[i] <- 0}

# Tinh van toc trung binh cua cac xe (km/h)
> mean(v)
[1] 23.1853
```

```
# Cau h), i), j), k), l).
> Cell <- NULL

# Tao ra mot data.frame gom cac cell ma moi xe di qua
> for (i in 1:393)
+ Cell <- rbind.data.frame(Cell, unique(Jc[[i]][ ,c(1,2)]))

# Dem so xe buyt di qua mot cell bat ki
> attach(Cell)
> a <- Lat
> b <- Long
> N <- NULL
>for (i in 1:nrow(Cell))
> N[i] <- nrow(subset(Cell, Lat==a[i] & Long==b[i]))
> Num <- data.frame(N)
> attach(Num)

# Liet ke cac cell co so xe buyt di qua nhieu nhat (h)
> A <- NULL
> B <- NULL
> dem <- NULL
> Vitri <- NULL
> attach(Num)
> dem <- unique(sort(N))
> for (i in 1:nrow(Num))
+ if (N[i]==dem[length(dem)]) {Vitri <- rbind(Vitri, i)} else{}
> for (i in 1:length(Vitri))
 {A[i] <- a[Vitri[i]]
  B[i] <- b[Vitri[i]]}
> c(unique(A), unique(B))
[1] 1061 853

# Liet ke cac cell co so xe buyt di qua nhieu nhi (i)
> A <- NULL
> B <- NULL
> Vitri <- NULL
> for (i in 1:nrow(Num))
+ if (N[i]==dem[length(dem)-1]) {Vitri <- rbind(Vitri, i)} else{}
> for (i in 1:length(Vitri))
```

```
 {A[i] <- a[Vitri[i]]
  B[i] <- b[Vitri[i]]}
> c(unique(A), unique(B))
[1] 832 765

# Liet ke cac cell co so xe buyt di qua nhieu nhat hoac nhieu nhi (j)
[1] 1061 853
[2] 832 765

# Liet ke danh sach cac cell thuoc mot phan ba dau theo thu tu so luong xe buyt di qua
    giam dan (k)
> f <- round((1/3)*length(dem))
> A <- NULL
> B <- NULL
> Vitri <- NULL
> for (i in 1:nrow(Num))
+ if (N[i]<=dem[length(dem)] & N[i]>=dem[length(dem)-f]) {Vitri <- rbind(Vitri, i)}
    else{}
> for (i in 1:length(Vitri))
 {A[i] <- a[Vitri[i]]
  B[i] <- b[Vitri[i]]}
> danhsach <- cbind.data.frame(A,B)
> unique(danhsach)
      A    B
1   1095  763
2   1086  772
3   1085  773
4   1084  774
5   1061  856
6   1061  855
7   1061  854
8   1061  853
9   1061  852
10  1084  773
.............
762 1050 882
827 1052 873

# Liet ke danh sach nhom cell co so luong lon nhat ma trong do cac cell duoc gom nhom
    sao cho so lan xe buyt di qua moi cell trong nhom la nhu nhau (l)
> khaosat <- data.frame(table(Num))
> attach(khaosat)
> for (i in 1:nrow(khaosat))
> if (Freq[i]==max(Freq)) {z <- i} else{}
> A <- NULL
> B <- NULL
> Vitri <- NULL
> for (i in 1:nrow(Num))
> if (N[i]==z) {Vitri <- rbind(Vitri, i)} else{}
> for (i in 1:length(Vitri))
  {A[i] <- a[Vitri[i]]
   B[i] <- b[Vitri[i]]}
> danhsach <- cbind.data.frame(A,B)
```

```
> options(max.print = 100000)
> unique(danhsach)
          A    B
1      1059  743
2      1059  733
3      1061  721
4      1063  717
5      1065  709
6      1068  686
7      1088  675
8      1097  793
9      1088  796
10     1095  786
.................
33465   938 1022
33466   964 1008
33467   885 1049
```

Kết quả:
- Tổng số lần di chuyển qua 1 cell cho trước là value2
- Quãng đường di chuyển trung bình của các xe trong 1 ngày là 202.007 km
- Xe di chuyển nhiều nhất là xe 353
- Xe di chuyển nhiều nhì là xe 321
- Danh sách các xe buýt thuộc một phần ba đầu theo thứ tự chiều dài di chuyển giảm dần là: 171 176 177 183 184 185 186 188 190 194 195 199 200 202 204 207 208 212 214 220 227 228 234 238 239 252 254 255 256 259 260 263 265 273 276 277 281 291 298 299 303 305 307 308 309 311 312 313 315 316 319 321 322 323 326 328 330 331 332 333 334 342 344 353 359 360 361 366 369 370 372 373 383 384 385 386 388 389 390 392 395 397 401 407 409 415 422 429 432 435 456 461 463 464 465 469 471 481 482 483 500 506 507 509 512 513 515 516 518 519 520 521 522 524 525 526 528 529 530 531 532 538 542 543 547 548 549 550 551 552 559.
- Vận tốc trung bình của các xe là 23.1853 km/h
- Các cell có số xe buýt đi qua nhiều nhất là: (1061,853)
- Các cell có số xe buýt đi qua nhiều nhì là: (832,765)
- Danh sách các cell thuộc một phần ba đầu theo thứ tự số lượng xe buýt đi qua giảm dần là: (A[i],B[i])trong câu k)
- Danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lần xe buýt đi qua mỗi cell trong nhóm là như nhau: (A[i],B[i]) trong câu l).

(8). Điểm ùn tắc: có ba định nghĩa như dưới đây.
a) Có nhiều tuyến xe buýt giao nhau tại một cell
b) Nhiều hơn hai xe buýt xuất hiện tại một cell trong cùng một khoảng thời gian $\triangle t$
c) Có nhiều hơn bốn xe buýt di chuyển trong các cell liền kề trong cùng một khoảng thời gian $\triangle t$.
- Định nghĩa đầu tiên có ý nghĩa xét về mặt lý thuyết, hai định nghĩa sau cần thông qua dữ liệu di chuyển thực tế của các xe buýt.
Theo tập dữ liệu nhận được, hãy xác định các điểm ùn tắc theo từng định nghĩa trên.

Code:

```
#--------# 8a ------------------------------------------------------------------------
# tao tap mau, dataset
```

```
> library(readxl)
> q <- NULL
> setwd("C:/RR deadline 2512/drive-download-20171219T155036Z-001/DTB56RouteGPS")
> q2 <- list.files(pattern = "*.xlsx")
> q <- lapply(q2,read_excel)
-------------------------------------------------------------------------------
# gop cac list trong q2 thanh q3
>  for (i in 1:42){
+ q3 <- q[[i]]}
> for (i in 1:42){
+     q3 <- rbind.data.frame(q3, (q[[i]]))}
-------------------------------------------------------------------------------
# trich cac Station_Code ra q4 (vi moi tram deu co 1 Station_Code khac nhau)
> q4 <- data.frame(q3$Station_Code)
# tan suat xuat hien cua cac tram
> q5 <- table(q4)
$ tan xuat xuat hien cua cac so 1, 2, 3,........
> table(q5)
# ket qua la n = sum(q5) - so lan xuat hien cua so 1 = 2366
-------------------------------------------------------------------------------
#--------#8b ---------------------------------------------------------------------
# tao tap mau, dataset
> library(readxl)
> Jc <- NULL
>  setwd("C:/RR deadline 2512/drive-download-20171219T155036Z-001/DTB56JourneyCell40")
> Jc40 <- list.files(pattern = "*.xlsx")
> Jc <- lapply(Jc40,read_excel)
>
> for (i in 1:393){
+ data_journey <- Jc[[i]]
+ }
> for (i in 1:393){
+     data_journey <- rbind.data.frame(data_journey, (Jc[[i]]))}
-------------------------------------------------------------------------
# luot bo cac thoi diem khong can thiet, chi lay thoi gian dau xuat hien (time) va thoi
    gian cuoi xuat hien (end)
> for(a in 1:393){
+ {journey_tempo <- data.frame(Jc[[a]])
+ journey_tempo$time = 0
+ for(i in 1:(nrow(journey_tempo)-1)){
+ if(journey_tempo$Lat[[i]] != journey_tempo$Lat[[i+1]] | journey_tempo$Long[[i]] !=
    journey_tempo$Long[[i+1]]){journey_tempo$time[[i]] = journey_tempo$Send.time[[i]]
+ journey_tempo$time[[i+1]] = journey_tempo$Send.time[[i+1]]}
+ journey_tempo$time[[1]] = journey_tempo$Send.time[[1]]
+ journey_tempo$time[[nrow(journey_tempo)]] =
    journey_tempo$Send.time[[nrow(journey_tempo)]]
+ }
+ journey_tempo <- subset(journey_tempo, journey_tempo$time != 0)
+ journey_tempo$end = 0
+ for(i in 1:(nrow(journey_tempo)-1)){
+ if(journey_tempo$Lat[[i]] == journey_tempo$Lat[[i+1]] & journey_tempo$Long[[i]] ==
    journey_tempo$Long[[i+1]]){journey_tempo$end[[i]] = journey_tempo$time[[i+1]]
+ journey_tempo$end[[i+1]] = 1}
```

```
+ }
+ journey_tempo <- subset(journey_tempo, journey_tempo$end != 1)}
+ ketqua[[a]] <- journey_tempo                              # ghi ra cac journey_tempo
    thanh cac list trong ket qua
+ }
-------------------------------------------------------------------------------------
# gop tat ca cac list ketqua lai thanh 1 file gop
> gop <- NULL
> for (i in 1:393){
+     gop <- rbind.data.frame(gop, (ketqua[[i]]))}
> gop_test <- NULL
> gop_test <- unique(gop)
-----------------------------------------------------------------
# lay cot 1, cot 2 cua file gop, Lat, Long
> gop_test <- gop[,1:2]
> l <- NULL
> l <- unique(gop_test)
-----------------------------------------------------------------
# ghi tat ca cac Lat Long giong nhau theo tung list, cac list trong list_chung2 (khong
    co list nao chi 1 phan tu, neu no la 1 phan tu thi bo)
> number <- 0
> list_chung2 <- list()
> for (i in 1:(nrow(l))){
+    a <- subset(gop, gop$Lat == l$Lat[[i]] & gop$Long == l$Long[[i]])
+    if (nrow(a) > 1){ number <- number +1
+    list_chung2[[number]] <- a
+    }
+    }
> View(list_chung2)
-----------------------------------------------------------------
# gop tat ca cac list trong list_chung2 thanh 1 file co_len
> co_len <- NULL
> for (i in 1:393){
+      co_len <- rbind.data.frame(co_len, (list_chung2[[i]]))}
-----------------------------------------------------------------
# sap xep colen2 theo thu tu tang dan | neu Lat & Long giong nhau, va end[[i]]==0 thi
    end[[i+1]] = end[[i]]
> colen2 <- colen
> colen2 <- arrange(colen2, Lat, Long, time, end, decreasing = FALSE)
> for(i in 1:(nrow(colen2)-1)){
+     if(colen2$Lat[[i+1]]==colen2$Lat[[i]] & colen2$Long[[i+1]]==colen2$Long[[i]]){
+         if(colen2$end[[i+1]]==0){
+             colen2$end[[i+1]]=colen2$end[[i]]
+            }
+        }
+     }
-----------------------------------------------------------
# trich file colen2 va bo cac phan tu end==0, thanh colen3
# neu Lat, Long giong nhau & end[[i]] > time[[i+1]] thi danhdau = 1
> colen3 <- subset(colen2, colen2$end != 0)
> colen3$danhdau=0
> for(i in 1:(nrow(colen3)-1)){
+     if(colen3$Lat[[i]]==colen3$Lat[[i+1]] & colen3$Long[[i]]==colen3$Long[[i+1]]){
```

```
+      if(colen3$end[[i]] > colen3$time[[i+1]]){colen3$danhdau[[i]]=1}
+      }
+      }
-------------------------------------------------------------
# danhdau2 cho cac cell[[i+1]] ma colen3$danhdau[[i]] == 1 & colen3$danhdau[[i+1]] == 0
> for(i in 1:(nrow(colen3)-1)){
+     if(colen3$Lat[[i]]==colen3$Lat[[i+1]] & colen3$Long[[i]]==colen3$Long[[i+1]]){
+      if(colen3$danhdau[[i]] == 1 & colen3$danhdau[[i+1]] == 0){colen3$danhdau[[i+1]]=2}
+      }
+      }
-------------------------------------------------------------
# tim cac cell co nhieu hon 2 xem trong cung khoang thoi gian
> colen4 <- subset(colen3, colen3$danhdau!=0)
> colen4$danhdau1=0
> for (i in 1:(nrow(colen4)-2)){
+     if(colen4$danhdau[[i]]==1 & colen4$danhdau[[i+1]]==1){
+     if(colen4$end[[i]] > colen4$end[[i+2]]){colen4$danhdau1[[i]]=1}}
+      }
-------------------------------------------------
$ tim cac cell khac nhau sau khi da biet cell nao co diem un tac
> colen5 <- subset(colen4, colen4$danhdau1!=0)
> for (i in 1:(nrow(colen5)-1)) {
+     if(colen5$Lat[[i]] != colen5$Lat[[i+1]] | colen5$Long[[i]] !=
     colen5$Long[[i+1]]){colen5$danhdau2[[i]]=1}
+      if(colen5$Lat[[nrow(colen5)]]!=colen5$Lat[[nrow(colen5)-1]] |
     colen5$Long[[nrow(colen5)]] !=
     colen4$Long[[nrow(colen5)-1]]){colen5$danhdau2[[nrow(colen5)]]=1}
+      }
> View(colen5)
-------------------------------------------------------------
$ loc ra cac cell vao colen6
> colen6 <- subset(colen5, colen5$danhdau2==1)
> sum(colen6$danhdau2, na.rm = FALSE)
# Ket qua la 6
```

Kết quả:
- Có 2366 điểm ùn tắc theo định nghĩa a)
- Có 6 điểm ùn tắc theo định nghĩa b)

(9). Điểm kẹt xe: là nơi mà xe buýt di chuyển chậm. Đặc điểm là xe buýt di chuyển chậm qua hai cell liền kề, cụ thể là hơn 60 giây để di chuyển qua một cell.
Lưu ý rằng có nhiều trường hợp xe buýt dừng lâu tại một cell, ví dụ như là trạm, hoặc gặp đèn đỏ hoặc gặp sự cố khác. Do vậy, không chắc chắn rằng xe buýt đang tại vị trí kẹt xe nếu chỉ có thông tin tại một cell.
Theo tập dữ liệu nhận được, hãy xác định các điểm kẹt xe.

Code:

```
> setwd("C:/work/journeycell60")
> journey_list <- list.files(pattern = ".xlsx")
> library("readxl", lib.loc="~/R/win-library/3.4")
> journey_data <- lapply(journey_list,read_excel )
```

```
> a<- NULL
> many <- 0
> journey_gop <- NULL
> for (b in 1:393){
+ a <- data.frame(journey_data[[b]])
+ a$preq = 0
+ for (i in 1:(nrow(a) - 1)) {
+ if (a$Lat[[i]] == a$Lat[[i+1]] & a$Long[[i]] == a$Long[[i+1]]){
+ many <- many +1
+ if (i == nrow(a) -1) { a$preq[[i +1]] = many + 1}
+ } else{
+ a$preq[[i]] = many + 1
+ many <- 0
+ }
+ }
+ journey_one <- subset(a, a$preq > 5)
+ journey_one <- unique(journey_one)
+ journey_gop <- rbind(journey_gop, journey_one)
+ }
+ journey_gop_LatLong <- journey_gop[,1:2]

> journey_gop_uni <- unique(journey_gop_LatLong)
> journey_gop_uni$Preq = 0
> for (i in 1:nrow(journey_gop_uni)){
+ a <- subset(journey_gop, journey_gop$Lat == journey_gop_uni$Lat[[i]] &
    journey_gop$Long == journey_gop_uni$Long[[i]])
+ journey_gop_uni$Preq[[i]] = nrow(a)
}
> journey_gop_use <- subset(journey_gop_uni, journey_gop_uni$Preq > 3)


>  setwd("C:/work/routecell60")
> route_list <- list.files(pattern = ".xlsx")
> route_data <- lapply(route_list, read_excel)
> route_gop <- NULL
> for (i in 1:42) {
+ route_gop <- rbind.data.frame(route_gop, route_data[[i]])
+ }
> route_gop_LatLong <- route_gop[,8:9]
> route_gop_LatLong_uni <- unique(route_gop_LatLong)
> journey_LatLong_use <- journey_gop_use[,1:2]
> names(route_gop_LatLong_uni) <- c("Lat","Long")
> processfile <- rbind(journey_LatLong_use,route_gop_LatLong_uni)
> processfile_uni <- unique(processfile)
> processfile_uni$Preq = 0

> for ( i in 1: nrow(processfile_uni)){
+ a <- subset(processfile, processfile$Lat == processfile_uni$Lat[[i]] &
    processfile$Long == processfile_uni$Long[[i]])
+ processfile_uni$Preq[[i]] <- nrow(a)
+ }
> choose_station <- subset(processfile_uni, processfile_uni$Preq >1)
```

```
> choose_station <- choose_station[,1:2]

> processfile2 <- rbind(journey_LatLong_use, choose_station)
> processfile2_uni <- unique(processfile2)
> processfile2_uni$Preq = 0

> for (i in 1:nrow(processfile2_uni)){
+ a <- subset(processfile2, processfile2$Lat == processfile2_uni$Lat[[i]] &
    processfile2$Long == processfile2_uni$Long[[i]])
+ processfile2_uni$Preq[[i]] = nrow(a)
+ }

> ketqua <- subset(processfile2_uni, processfile2_uni$Preq == 1)
> ketqua <- ketqua[,1:2]
```

Kết quả:
- Các điểm kẹt xe được lưu lại trong diemketxe2.csv

	Lat	Long	Freq
1943	708	492	12
2130	720	385	31
2255	742	446	4
2325	755	485	6
2398	723	516	20
2630	707	570	83
2797	722	516	6
2868	746	495	4
3134	719	384	50
3348	746	496	5
...	...	...	...

(10). Điểm thông thoáng: là điểm không hề bị ùn tắc hoặc kẹt xe tại bất kỳ thời điểm nào.

Code:

```
# Su dung cac ket qua cua bai 8:
# Trong bai 8 co gan l <- unique(gop_test), tong so phan tu trong l chinh la tong tat
    ca cac cell
# Va co su dung ket qua bai 9, diemketxe
# Bai lam:
#Lay cot Lat va Long cua ket qua bai 8
> colen7 <- colen6[,1:2]
#Lay cot Lat va Long cua ket qua bai 9
> bai9 <- diemketxe[,2:3]
#gop 2 file lai
> bai10 <- rbind(bai9, colen7)
#Luu y: se co nhung diem trung lap (vua ket xe, vua un tac), nen ta phai lay unique
> bai10_1 <- unique(bai10)
#lay tat ca nhung diem thuoc l nhung khong thuoc bai10_1 la ra diem thong thoang
> bai10_2 <- NULL
> for(i in 1:559){
+ bai10_2 <- subset(l, l$Lat != bai10_1$Lat[[i]] & l$Long != bai10_1$Long[[i]])}
> View(bai10_2)
# bai10_2 chinh la ket qua (diemthongthoang)
```

Kết quả:
- Các điểm thông thoáng được lưu trong data.frame bai10_2

(11). Điểm bất thường: là điểm thường xuyên bị ùn tắc hoặc kẹt xe.

Code:

```
# Bai nay co su dung 1 vai ket qua cua bai 8, 9 de tinh
# Tim so lan bi un tac trong cac cell cua cau 8:
> colen8 <- colen5[,1:2]        # Lay cot Lat, Long cua colen5
> colen9 <- unique(colen8)      # Lay unique colen8
# tim so lan lap cua cac cell trong cau 8
> a <- NULL
> colen9$Freq = 0
> for (i in 1:nrow(colen9)){
+     a <- subset(colen8, colen8$Lat == colen9$Lat[[i]] & colen8$Long ==
    colen9$Long[[i]])
+     colen9$Freq[[i]] = nrow(a)}
# trich cac cell hay bi un tac (neu nhu lan lap lon hon 2 duoc cho la hay bi un tac)
> colen10 <- subset(colen9, colen9$Freq > 2)
> colen10 <- colen10[,1:2]
# colen10 la cac cell thuong xuyen un tac
# diemketxe2 la cac cell thuong xuyen ket xe
# hop cua colen10 va diemketxe2 la ket qua cau 11, nhung diem bat thuong
> bai11 <- rbind(colen10, diemketxe2)
> bai11 <- unique(bai11)
# bai11 chinh la tap hop cac cell la nhung diem bat thuong, cung la dap an cau 11
```

Kết quả:
- Các điểm bất thường được lưu trong data.frame bai11

Kết quả:

4 Kết luận

- Thông qua đề tài Thống kê và Phân tích dữ liệu bằng R, nhóm chúng em đã được tiếp cận và làm quen với R - một ngôn ngữ lập trình đa năng dùng trong thống kê, sử dụng R cho các phép toán đơn giản như tính tổng, tính trung bình, tìm max, min, tính phương sai, độ lệch chuẩn... và tiến hành một dự án nhỏ: Nghiên cứu vấn đề kẹt xe buýt của thành phố Hồ Chí Minh.
- Thông qua dự án này chúng em nhận thấy rằng thành phố Hồ Chí Minh vẫn còn rất nhiều các điểm ùn tắc và kẹt xe gây ra bởi các xe buýt. Không thể phủ nhận về lợi ích của hình thức di chuyển công cộng này nhưng kẹt xe là một vấn đề lớn cần được giải quyết để đảm bảo cho thành phố Hồ Chí Minh có một nền giao thông an toàn và văn minh.
- Đề xuất của nhóm cho vấn đề kẹt xe này là:
Trung tâm quản lý xe buýt thành phố nên phân phối các tuyến đường ưu tiên riêng dành cho xe buýt, sắp xếp lại các lộ trình và thời gian di chuyển của các xe buýt một cách hợp lý hơn.

Tài liệu

[1] Giáo sư Nguyễn Văn Tuấn "<**http://www.nguyenvantuan.net/**>", *xem ngày :* 26-30/11/2017.

[2] Chuyển đổi Convert Unix Timestamp "**link: https://xuanthulab.net/unix-timestamp-chuyen-doi-thoi-gian-unix.html**", *Chuyển đổi thời gian*, lần truy cập cuối: 24/12/2017.

[3] Môi trường làm việc R "**link: https://www.rstudio.com/wp-content/uploads/2016/07/Base-R-Vietnamese.pdf**", *Base R Cheat Sheet*, lần truy cập cuối: 24/12/2017.