interview_question

Najia Pan

5/24/2022

##Question 1:

#a. Simulate two data sets: 1.Data set called Demo with three variables: id, birth_dt, gender, county. id variable uniquely identifies rows. 2.The birth_dt will range from '01-01-1940' to '12-31-2021'. 3.The county should have Los Angeles county and other counties in California. 4.Data set called Medical_visit of medical records with these variables: id, service_date, service_code, service_num. The variables id, service_date and service_num uniquely identify rows. 5.Make sure that 20% of unique ids in the Medical_visit data set have multiple records. 6.Make sure that these two data sets share at least 90% of the values of the id variable.
7.The service_date will range from '01-01-2018' to '12-31-2019'.

```r
library(readxl)
library(data.table)
library(formattable)
# create Demo table
count_df <- read_excel("California Counties.xlsx")
Demo = data.table(id = sample(1:1000,size=1000,replace = T) )
Demo$county = sample(count_df$County, size=1000, replace = T)
Start <- as.Date("1940-01-01")
End <- as.Date("2021-12-31")
dates <- seq(from = Start, to = End, by = 1)
Demo$birth_dt <- sample(dates, size=1000, replace = T)
Demo[, gender := sample(rep(c('Males','Females'),5),replace = T, size = 1000)]
dim(Demo)
```

```
## [1] 1000    4
```

```r
formattable(head(Demo))
```

id

county

birth_dt

gender

302

Santa Clara County

1968-01-07

Males

473

San Diego County

1990-07-26

Males

869

Merced County

2017-09-06

Males

462

San Francisco County

1999-08-31

Males

734

Kings County

1984-04-06

Females

129

Amador County

1993-09-19

Females

```
## create Medical_visit table, Choose size = 900 to make sure that these two data sets share at least 9
Medical_visit = data.table(id =  sample(1:1000,size=900,replace = T) )
Start <- as.Date("2018-01-01")
End <- as.Date("2019-12-31")
dates <- seq(from = Start, to = End, by = 1)
Medical_visit[, service_date := sample(dates, size=900, replace = T)]
Medical_visit[, service_code := sample(c('AX','BY','HG','HT','DT'),replace = T, size = 900)]
Medical_visit[, service_num := sample(1:50,replace = T, size = 900)]
dim(Medical_visit)
```

```
## [1] 900   4
```

```
formattable(head(Medical_visit))
```

id

service_date

service_code

service_num

321

2018-08-19

HT

14

24

2019-09-21

AX

38

689

2019-10-02

DT

16

676

2019-06-10

HG

17

942

2018-04-09

HT

8

242

2019-09-22

DT

6

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

data.table is a faster way to manipulate data than using data.frame. Also learned there are 58 counties in CA.

#b.Subsetting Medical_visit to only female patients from Los Angeles county in the Demo data.

```
de_me<- merge(Demo,Medical_visit,by="id")
de_me
```

```
##        id                 county   birth_dt  gender service_date service_code
##   1:    3      Stanislaus County 1980-01-22   Males   2018-08-29           HG
##   2:    3          Merced County 1971-02-15   Males   2018-08-29           HG
##   3:    3          Colusa County 1997-01-23   Males   2018-08-29           HG
##   4:    4        Imperial County 2016-07-14 Females   2018-03-24           AX
##   5:    4            Mono County 1995-06-16   Males   2018-03-24           AX
##  ---
## 873:  995 San Luis Obispo County 1962-05-13 Females   2018-03-10           HT
## 874:  995 San Luis Obispo County 1962-05-13 Females   2018-07-16           BY
## 875: 1000   Santa Barbara County 1969-05-20   Males   2018-08-16           HG
## 876: 1000   Santa Barbara County 1969-05-20   Males   2019-09-11           BY
## 877: 1000   Santa Barbara County 1969-05-20   Males   2019-08-28           HT
##      service_num
##   1:          37
##   2:          37
##   3:          37
##   4:          16
##   5:          16
##  ---
## 873:          20
## 874:          10
## 875:          49
## 876:          26
## 877:          24
```

```
female_la<- de_me[gender=="Females"&county=="Los Angeles County"]
female_la
```

```
##      id             county   birth_dt  gender service_date service_code
##  1: 156 Los Angeles County 1942-08-14 Females   2018-06-24           HT
##  2: 321 Los Angeles County 1960-12-31 Females   2018-08-19           HT
##  3: 321 Los Angeles County 1960-12-31 Females   2018-05-22           DT
##  4: 468 Los Angeles County 1974-07-28 Females   2018-08-03           HT
##  5: 468 Los Angeles County 1974-07-28 Females   2018-11-21           DT
##  6: 561 Los Angeles County 1940-10-29 Females   2019-06-01           HG
##  7: 561 Los Angeles County 1940-10-29 Females   2018-05-02           HT
##  8: 561 Los Angeles County 1940-10-29 Females   2018-07-24           HT
##  9: 609 Los Angeles County 1967-10-23 Females   2019-01-27           BY
## 10: 640 Los Angeles County 1961-09-30 Females   2018-12-09           BY
## 11: 752 Los Angeles County 1985-11-10 Females   2019-07-24           HG
## 12: 752 Los Angeles County 1985-11-10 Females   2019-07-28           DT
## 13: 965 Los Angeles County 1951-11-05 Females   2018-04-03           HG
##     service_num
##  1:          36
##  2:          14
##  3:          15
##  4:          49
##  5:          48
##  6:          15
##  7:          37
##  8:          45
##  9:          23
## 10:           6
## 11:          34
## 12:          22
## 13:          15
```

```
setkey(Medical_visit,id)
 index<- Demo[gender=="Females"&county=="Los Angeles County",id]
 index
```

```
##  [1] 468 156   6 752 709 609 458 561 965 390 439 321 640
```

```
 female_la2<- Medical_visit[ id %in% index]
 female_la2
```

```
##      id service_date service_code service_num
##  1: 156   2018-06-24           HT          36
##  2: 321   2018-08-19           HT          14
##  3: 321   2018-05-22           DT          15
##  4: 468   2018-08-03           HT          49
##  5: 468   2018-11-21           DT          48
##  6: 561   2019-06-01           HG          15
##  7: 561   2018-05-02           HT          37
##  8: 561   2018-07-24           HT          45
##  9: 609   2019-01-27           BY          23
## 10: 640   2018-12-09           BY           6
## 11: 752   2019-07-24           HG          34
## 12: 752   2019-07-28           DT          22
## 13: 965   2018-04-03           HG          15
```

This question is asked for subsetting for two datasets. I've learned two ways of subsetting. One way is to merge to dataset into one, and subset based on the requirnments. Another way is to subset from the first dataset to get the id number(share variable), then use this common variable to subset from another dataset.

#c.Subsetting Medical_visit to those where the service_date is prior to '05-14-2018'.

```
setkey(Medical_visit,id)
certain_date<- as.Date("2018-05-14")

Medical_visit[ service_date < certain_date]
```

```
##       id service_date service_code service_num
##   1:   4   2018-03-24           AX          16
##   2:  10   2018-03-12           BY          39
##   3:  12   2018-03-07           BY          23
##   4:  22   2018-01-17           AX           9
##   5:  25   2018-03-19           BY          31
##  ---
## 151: 992   2018-02-01           HT          46
## 152: 995   2018-03-10           HT          20
## 153: 997   2018-01-18           HT          44
## 154: 998   2018-05-01           HG          28
## 155: 999   2018-05-05           HT           4
```

This question is to ask subset from a certain date. The most important part is to use as.Date(),otherwise way may not compare each date correctly.

#d. Add a variable called max_svc_date to the Medical_visit data set which is the service_date when a patient had the maximal service_num among all the medical records. For example, here are all medical records for a patient:

```
 max_ser_num<- Medical_visit[,Medical_visit[,max(service_num),id]]
setkey(Medical_visit, id, service_num)
 max_ser_dataset<-Medical_visit[.(max_ser_num[,1],max_ser_num[,2]), .(service_date)]
ans<-Medical_visit[ ,max(service_num), keyby = id]
max_ser_dataset
```

```
##      service_date
##   1:   2019-12-24
##   2:   2018-08-29
##   3:   2018-03-24
##   4:   2018-10-04
##   5:   2019-10-30
##  ---
## 600:   2018-03-10
## 601:   2018-01-18
## 602:   2019-06-03
## 603:   2018-05-05
## 604:   2018-08-16
```

```
Medical_visit[,max_svc_dat:=max(service_date), keyby = id]
Medical_visit
```

```
##        id service_date service_code service_num max_svc_dat
##   1:    1   2019-12-24           HG           1  2019-12-24
##   2:    3   2018-08-29           HG          37  2018-08-29
##   3:    4   2018-03-24           AX          16  2018-03-24
##   4:    5   2018-10-04           DT           8  2018-10-04
##   5:    7   2019-10-30           BY          32  2019-10-30
##  ---
## 896:  998   2019-06-03           BY          45  2019-06-03
## 897:  999   2018-05-05           HT           4  2018-05-05
## 898: 1000   2019-08-28           HT          24  2019-09-11
## 899: 1000   2019-09-11           BY          26  2019-09-11
## 900: 1000   2018-08-16           HG          49  2019-09-11
```

I know how to subset the max_num first, but I don't know how to connect them

#e.Add an integer variable called age_by_svc to the Medical_visit data set which represents a patient's age by the service_date. This variable can have missing values if no demographic information is available for a patient.

```
Medical_visit[,age_by_svc:=max(service_date)-min(service_date), keyby = id]
Medical_visit
```

```
##        id service_date service_code service_num max_svc_dat age_by_svc
##   1:    1   2019-12-24           HG           1  2019-12-24     0 days
##   2:    3   2018-08-29           HG          37  2018-08-29     0 days
##   3:    4   2018-03-24           AX          16  2018-03-24     0 days
##   4:    5   2018-10-04           DT           8  2018-10-04     0 days
##   5:    7   2019-10-30           BY          32  2019-10-30     0 days
##  ---
## 896:  998   2019-06-03           BY          45  2019-06-03   398 days
## 897:  999   2018-05-05           HT           4  2018-05-05     0 days
## 898: 1000   2019-08-28           HT          24  2019-09-11   391 days
## 899: 1000   2019-09-11           BY          26  2019-09-11   391 days
## 900: 1000   2018-08-16           HG          49  2019-09-11   391 days
```

This question is to Calculate Difference between dates by id in as.table. I've learned how to create a new column by using := and how to calculate duration of a time. Note that diff()is not working here.

#f.Please create a dataset called Medical_visit2018: Medical_visit2018 subsets Medical_visit data with added age_by_svc variable from part e) to those have county information in the Demo data and have service_date in 2018.

Create an age_by_svc by county distribution data called Age_distri based on Medical_visit2018. Please include Min, Max, Quartiles, Mean in the distribution and format the Age_distri like the following:

```
setkey(Medical_visit,id)
 index<- Medical_visit[service_date %like% 2018,id]
 index
```

```
##   [1]    3    4    5    9   10   12   18   19   22   24   25   26   27   27   30
##  [16]   35   43   43   47   47   51   54   55   56   60   62   64   66   68   69
##  [31]   70   78   79   80   84   87   90   92   93   96   99  100  100  102  102
##  [46]  105  106  108  110  111  121  122  124  128  129  133  136  144  145  145
##  [61]  146  155  156  159  159  164  167  167  169  172  172  174  177  178  178
##  [76]  182  183  186  187  192  194  198  201  201  203  208  211  212  218  219
##  [91]  222  225  227  227  232  233  239  245  247  249  250  252  254  255  256
## [106]  259  263  265  265  265  266  270  275  275  276  276  282  290  290  292
## [121]  294  294  294  295  297  302  303  309  310  313  315  316  316  317  318
## [136]  321  321  326  328  329  331  332  335  341  341  341  345  348  354  358
## [151]  361  361  364  364  368  370  370  373  376  377  377  379  381  382  383
## [166]  384  384  387  388  388  389  391  393  395  396  400  402  404  406  408
## [181]  410  410  414  418  431  435  440  441  443  444  446  447  451  455  457
## [196]  459  461  468  468  470  470  470  472  475  478  480  480  484  487  487
## [211]  487  490  490  495  501  503  505  506  507  508  513  515  516  519  521
## [226]  526  531  532  537  541  542  545  545  546  546  547  548  550  550  551
## [241]  554  558  561  561  563  564  565  568  568  569  573  574  574  574  577
## [256]  578  580  581  583  584  587  587  590  590  593  595  598  611  611  612
## [271]  615  616  618  618  619  623  625  627  630  630  635  636  637  640  641
## [286]  642  647  648  649  655  656  657  660  661  661  666  667  668  673  674
## [301]  675  675  677  677  678  678  681  686  686  687  688  690  691  692  697
## [316]  699  700  705  710  710  712  712  717  721  721  722  723  727  729  731
## [331]  734  735  735  737  740  744  745  747  747  747  750  754  757  758  759
## [346]  764  768  772  774  776  777  781  782  785  791  794  795  799  801  803
## [361]  811  811  820  821  826  826  826  829  829  830  832  835  835  836  842
## [376]  843  843  843  845  847  849  849  853  853  858  862  864  864  865  867
## [391]  867  868  877  882  883  884  885  886  897  905  905  906  909  911  913
## [406]  919  921  924  926  926  928  934  934  935  937  938  939  940  941  941
## [421]  942  947  947  949  951  952  953  957  957  961  964  965  967  968  970
## [436]  971  978  981  984  987  990  991  992  995  995  997  998  999 1000
```

```
Medical_visit2018<- Demo[ id %in% index]
Medical_visit2018
```

```
##       id             county   birth_dt  gender
##   1: 302 Santa Clara County 1968-01-07   Males
##   2: 734       Kings County 1984-04-06 Females
##   3: 129      Amador County 1993-09-19 Females
##   4: 843     Ventura County 1984-07-08 Females
##   5: 444       Butte County 2014-09-05   Males
##  ---
## 367: 155      Tehama County 1954-07-27   Males
## 368: 111      Shasta County 2006-12-11   Males
## 369: 156    Monterey County 2012-06-22   Males
## 370: 744   San Mateo County 1999-03-10   Males
## 371: 587       Glenn County 2009-08-30 Females
```