Homework1

YikunHan42

2022-04-04

homework1_part1

Importing libraries

```r
library(tinytex)
```

```
## Warning:   'tinytex' R 4.1.3
```

```r
library(tidyverse)
```

```
## Warning:   'tidyverse' R 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(stringr)
```

Reading files & Setting variables

```r
data <- read.csv("D://Study/DSBI/Task1/test_id_card_no.csv")
sex =c()
Birthday =c()
Valid = c()
Age = c()
Zone_code = c()
```

1

(1)

```
data %>% filter(str_detect(Id_Card_No,pattern = "^22"))
```

```
##   seq         Id_Card_No
## 1   6 222424195110306886
## 2   9 220182196410190862
```

(2)(3)

```
#flag
count = 0
#
for(i in 1:nrow(data)){
#
  newdata<-data[c(i),2]
  a = substring(newdata, 1:18,1:18)

  if( a[18] == 'X'){a[18]=10}
    sum_ = 0
    for(j in 1:17){
    sum_ = sum_ +  as.numeric(a[j]) * (2**(18-j) %% 11)
    }
    a1 = ((sum_ %% 11) + as.numeric(a[18]))%%11- 1
  if(a[18]=="10"){a[18]='X'}

  if(a1 != 0){
    Valid = append(Valid,0)
    count = count + 1
  print("    ")
  print(paste(a, sep = "",collapse=""))}
  if(a1 == 0){
      Valid = append(Valid,1)
  }
#50
  byear <- c(paste(a[7:10], sep = "",collapse=""))
  bmonth <- c(paste(a[11:12], sep = "",collapse=""))
  bday <- c(paste(a[13:14], sep = "",collapse=""))
  b <- paste(byear,bmonth,bday, sep = "-")
  today <- Sys.Date()
  gtd <- as.Date(b)
  differencetime = difftime(today, gtd, units="days")
  if(differencetime>(365*38+366*12)){
    print("  50 ")
    print(paste(a, sep = "",collapse=""))}
#
  Birthday<-append(x=Birthday,as.Date(b))
  Age<-append(Age,2022-as.numeric(byear))
  Zone_code = append(Zone_code,c(paste(a[1:2], sep = "",collapse="")))
  #
  if (as.numeric(a[17])%%2==1){
    sex = append(sex,"Male")
  }
  else{
    sex = append(sex,'Female')
  }
}
```

```
## [1] "  50 "
## [1] "431128197009055759"
## [1] "  50 "
## [1] "360731196804216811"
## [1] "  50 "
## [1] "150123195103126841"
## [1] "  50 "
## [1] "222424195110306886"
## [1] "  50 "
## [1] "61102319591227666X"
## [1] "  50 "
## [1] "141182195505236567"
## [1] "  50 "
## [1] "220182196410190862"
## [1] "  50 "
## [1] "14062219620604034X"
## [1] "  50 "
## [1] "341124196902230765"
```

```
if(count == 0) print("   ")
```

```
## [1] "   "
```

2

(1)

```
data <- data %>% mutate(Birthday)
data %>% arrange(Birthday)
```

```
##    seq         Id_Card_No   Birthday
## 1    5 150123195103126841 1951-03-12
## 2    6 222424195110306886 1951-10-30
## 3    8 141182195505236567 1955-05-23
## 4    7 61102319591227666X 1959-12-27
## 5   10 14062219620604034X 1962-06-04
## 6    9 220182196410190862 1964-10-19
## 7    4 360731196804216811 1968-04-21
## 8   11 341124196902230765 1969-02-23
## 9    2 431128197009055759 1970-09-05
## 10   1 431021197306142736 1973-06-14
## 11   3 440700197510019150 1975-10-01
```

(2)

```
sex = as.factor(sex)
data <- data %>% mutate(sex)
data %>% arrange(sex,desc(Birthday))
```

```
##    seq         Id_Card_No   Birthday    sex
## 1   11 341124196902230765 1969-02-23 Female
## 2    9 220182196410190862 1964-10-19 Female
## 3   10 14062219620604034X 1962-06-04 Female
## 4    7 61102319591227666X 1959-12-27 Female
## 5    8 141182195505236567 1955-05-23 Female
## 6    6 222424195110306886 1951-10-30 Female
## 7    5 150123195103126841 1951-03-12 Female
## 8    3 440700197510019150 1975-10-01   Male
## 9    1 431021197306142736 1973-06-14   Male
## 10   2 431128197009055759 1970-09-05   Male
## 11   4 360731196804216811 1968-04-21   Male
```

3

```
Valid = as.logical(Valid)
data <- data %>% mutate(Valid)
data
```

```
##    seq         Id_Card_No   Birthday    sex Valid
## 1    1 431021197306142736 1973-06-14   Male  TRUE
## 2    2 431128197009055759 1970-09-05   Male  TRUE
## 3    3 440700197510019150 1975-10-01   Male  TRUE
## 4    4 360731196804216811 1968-04-21   Male  TRUE
## 5    5 150123195103126841 1951-03-12 Female  TRUE
## 6    6 222424195110306886 1951-10-30 Female  TRUE
## 7    7 61102319591227666X 1959-12-27 Female  TRUE
## 8    8 141182195505236567 1955-05-23 Female  TRUE
## 9    9 220182196410190862 1964-10-19 Female  TRUE
## 10  10 14062219620604034X 1962-06-04 Female  TRUE
## 11  11 341124196902230765 1969-02-23 Female  TRUE
```

4

(1)

```
data <- data %>% mutate(Age)
data %>% filter(Valid=TRUE) %>% summarise(mean(Age))
```

```
##   mean(Age)
## 1  58.63636
```

```
data %>% filter(Valid=TRUE) %>% summarise(median(Age))
```

```
##   median(Age)
## 1          58
```

(2)

```
data %>% filter(Valid=TRUE) %>% summarise(n())
```

```
##   n()
## 1  11
```

(3)

```
data %>% filter(Valid=TRUE) %>% summarise(any(Age<30))
```

```
##   any(Age < 30)
## 1         FALSE
```

5

(1)

```
data  %>% filter(Valid=TRUE) %>% group_by(sex) %>% summarise(n(),mean(Age))
```

```
## # A tibble: 2 x 3
##   sex    `n()` `mean(Age)`
##   <fct>  <int>       <dbl>
## 1 Female     7        63.3
## 2 Male       4        50.5
```

(2)

```
data  %>% filter(Valid=TRUE) %>% group_by(sex) %>% summarise(n50=sum(Age>50),n_per=mean(Age>50))
```

```
## # A tibble: 2 x 3
##   sex      n50 n_per
##   <fct>  <int> <dbl>
## 1 Female     7   1
## 2 Male       2   0.5
```

6

(1)

```
data  %>% filter(Valid=TRUE)  %>% filter(Age<=65) %>% filter(!str_detect(Id_Card_No,pattern = "^220101"
```

```
##   seq         Id_Card_No   Birthday    sex Valid Age
## 1   7 61102319591227666X 1959-12-27 Female  TRUE  63
## 2  10 14062219620604034X 1962-06-04 Female  TRUE  60
## 3   9 220182196410190862 1964-10-19 Female  TRUE  58
## 4   4 360731196804216811 1968-04-21   Male  TRUE  54
## 5  11 341124196902230765 1969-02-23 Female  TRUE  53
## 6   2 431128197009055759 1970-09-05   Male  TRUE  52
## 7   1 431021197306142736 1973-06-14   Male  TRUE  49
## 8   3 440700197510019150 1975-10-01   Male  TRUE  47
```

```
newdata <- data[,c(2,3,4,5)]
write.csv(newdata,file = 'Out_Id_Card_Data.csv')
```

(2)

```
data = data %>% mutate(Zone_code)
data %>% filter(Valid=TRUE) %>% group_by(Zone_code,sex) %>% summarise(n_old=sum(Age>=60),n_old_per=mean
```

```
## `summarise()` has grouped output by 'Zone_code'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 8 x 6
## # Groups:   Zone_code [8]
##   Zone_code sex    n_old n_old_per n_not_old n_not_old_per
##   <chr>     <fct>  <int>     <dbl>     <int>         <dbl>
## 1 14        Female     2       1           0           0
## 2 15        Female     1       1           0           0
## 3 22        Female     1       0.5         1           0.5
## 4 34        Female     0       0           1           1
## 5 36        Male       0       0           1           1
## 6 43        Male       0       0           2           1
## 7 44        Male       0       0           1           1
## 8 61        Female     1       1           0           0
```

homework1_part2

Importing libraries

```
library(tidyverse)
library(Hmisc)
```

```
## Warning:  'Hmisc' R 4.1.3

##      lattice

##      survival

## Warning:  'survival' R 4.1.3

##      Formula

##
##    'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(dplyr)
library(plyr)
```

```
## Warning:  'plyr' R 4.1.3

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
##    'plyr'

## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(stringr)
library(lubridate)
```

```
##
##    'lubridate'

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

```
library(tidyr)
```

(1) Importing the file

```
## Sys.setlocale(category="LC_ALL",locale="en_US.UTF-8")
data <- read_csv('D://Study/DSBI/Task1/pharmacy_data.csv')
```

```
## Rows: 6574 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2):  ,
## dbl (5):   ,  ,  ,  ,
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data %>% as_tibble() ->data
```

(2) Summary view

```
summary(data)
```

```
##
##  Length:6574        Min.   :1.617e+06   Min.   : 236701   Length:6574
##  Class :character   1st Qu.:1.014e+08   1st Qu.: 861456   Class :character
##  Mode  :character   Median :1.002e+10   Median : 861507   Mode  :character
##                     Mean   :6.093e+09   Mean   :1016344
##                     3rd Qu.:1.005e+10   3rd Qu.: 869069
##                     Max.   :1.284e+10   Max.   :2367012
##                     NA's   :2           NA's   :1
##
##  Min.   :-10.000   Min.   :-374.00   Min.   :-374.00
##  1st Qu.:  1.000   1st Qu.:  14.00   1st Qu.:  12.32
##  Median :  2.000   Median :  28.00   Median :  26.60
##  Mean   :  2.386   Mean   :  50.48   Mean   :  46.32
##  3rd Qu.:  2.000   3rd Qu.:  59.60   3rd Qu.:  53.00
##  Max.   : 50.000   Max.   :2950.00   Max.   :2650.00
##  NA's   :1         NA's   :1         NA's   :1
```

(3) Renaming date & change the data type   &

```
data <- rename(data,c(" "="  "))
data$   <- ymd(data$  )
```

```
## Warning: 23 failed to parse.
```

```
data
```

```
## # A tibble: 6,574 x 7
##
##    <date>           <dbl>  <dbl> <chr>          <dbl> <dbl>    <dbl>
##  1 2018-01-15   101554328 236701             8 224      208
##  2 2018-01-20    13389528 236701             1  28       28
##  3 2018-01-31   101464928 236701             2  56       56
##  4 2018-02-17    11177328 236701             5 149      131.
##  5 2018-02-22 10065687828 236701             1  29.8     26.2
##  6 2018-02-24    13389528 236701             4 119.     105.
##  7 2018-03-05 10026389628 236701             2  59.6     59.6
##  8 2018-03-05   102285028 236701             3  84       84
##  9 2018-03-05 10077400828 236701          1  28       24.6
## 10 2018-03-07 10077400828 236701          5 140      112
## # ... with 6,564 more rows
```

(4) Deleting rows of missing data

```
data <- data[complete.cases(data[,1:2]),]
```

(5) Adding mean

```
data <- data %>% group_by( )
data$ [is.na(data$ )] <- mean(data$ ,na.rm=TRUE)
```

(6) Excluding rows

```
data <- data %>% filter( >0)
```

(7) Descending order

```
data %>% arrange(desc(  ))
```

```
## # A tibble: 6,505 x 7
## # Groups:     [78]
##
##    <date>           <dbl>   <dbl> <chr>                    <dbl> <dbl>    <dbl>
##   1 2018-07-19     1616528  236701                    1  28       28
##   2 2018-07-19 10013306428 2367011                          1  31       28
##   3 2018-07-19 10030713328 2367011                          4 124      118
##   4 2018-07-19 10059383628 2367011                          2  62       56
##   5 2018-07-19   101409528 2367011                          2  62       56
##   6 2018-07-19    13406628 2367011                          2  62       56
##   7 2018-07-19 10065621228  861435    ( )             2  71.6     64
##   8 2018-07-19 10081634128  861459    (  )          2  33       29.6
##   9 2018-07-19   101921828  861464     (  )     1   3.7      3.3
## 10 2018-07-19    13216828  861464     (  )     1   3.7      3.3
## # ... with 6,495 more rows
```

(8) Adding line

```
data <- data %>% mutate(  = ( -  )/ )
data
```

```
## # A tibble: 6,505 x 8
## # Groups:     [78]
##
##    <date>           <dbl>  <dbl> <chr>          <dbl> <dbl>    <dbl>  <dbl>
##   1 2018-01-15   101554328 236701             8 224      208   0.0714
##   2 2018-01-20    13389528 236701             1  28       28   0
##   3 2018-01-31   101464928 236701             2  56       56   0
##   4 2018-02-17    11177328 236701             5 149      131.  0.12
##   5 2018-02-22 10065687828 236701             1  29.8     26.2 0.120
##   6 2018-02-24    13389528 236701             4 119.     105.  0.120
##   7 2018-03-05 10026389628 236701             2  59.6     59.6 0
##   8 2018-03-05   102285028 236701             3  84       84   0
##   9 2018-03-05 10077400828 236701          1  28       24.6 0.12
## 10 2018-03-07 10077400828 236701          5 140      112   0.2
## # ... with 6,495 more rows
```

(9) Statistic about sale

```
data %>% summarise(n_sale=sum(  ), n_num=sum( ))
```

```
##     n_sale n_num
## 1 303898.3 15646
```

(10) Statistic by commodity

```
data %>% group_by( ) %>% dplyr::summarise(n_num=n(),n_sale=sum( ),n_average=sum( )/sum( ))
```

```
## # A tibble: 78 x 4
##                                  n_num n_sale n_average
##    <chr>                            <int>  <dbl>     <dbl>
##  1 **                              8   76.5       4.5
##  2 **    (  )         34 4040         40
##  3 D     (  )       1  453         15.1
##  4 D     ( )           1  132.        66.1
##  5 D                               1 2500       250
##  6 D                            3 1125.       34.1
##  7 G    (6 / )        72 2968.        12.5
##  8 G    (6 / )           9  576         32
##  9 G      (  )   195 9648.       19.0
## 10 G    (II)(6 / )       9  480         30
## # ... with 68 more rows
```

(11) Statistic by month

```
data %>%
  mutate( =year(  )) %>%
  mutate( =month(  )) %>%
  mutate( _ =str_c( , ,sep = "-")) -> data
data %>%
  group_by( _ ) %>%
  dplyr::summarise(
       = sum( ),
       = sum( ),
       = mean( ))
```

```
## # A tibble: 7 x 4
##     _
##   <chr>     <dbl>    <dbl>    <dbl>
## 1 2018-1     2517   53406      50.8
## 2 2018-2     1858   42029.     56.6
## 3 2018-3     2225   45318      45.8
## 4 2018-4     3010   54324.     44.0
## 5 2018-5     2225   51263.     53.8
## 6 2018-6     2328   52301.     57.5
## 7 2018-7     1483   32568      51.9
```

(12) Statistic by customer

```
data %>%
  mutate( _ =str_c(   , _ ,sep = "-"))-> data
data %>%
  group_by( _ ) %>%
  dplyr::summarise(
      =sum(   ))
```

```
## # A tibble: 4,375 x 2
##      _
##    <chr>              <dbl>
##  1 10000428-2018-2     17
##  2 10000528-2018-5     25
##  3 10001928-2018-1      2.2
##  4 10005028-2018-1    276.
##  5 10005028-2018-2     50
##  6 10005028-2018-4     23.6
##  7 10005028-2018-7     47.2
##  8 10006928-2018-3     12.3
##  9 10006928-2018-4      5.4
## 10 10006928-2018-6      6.4
## # ... with 4,365 more rows
```