401_Project

Chuan Du (Sophie)

11/16/2018

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.4
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(tidyverse)
```

```
## ── Attaching packages ──────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.1.0     ✔ purrr   0.2.5
## ✔ tibble  1.4.2     ✔ dplyr   0.7.4
## ✔ tidyr   0.8.2     ✔ stringr 1.2.0
## ✔ readr   1.1.1     ✔ forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
## Warning: package 'forcats' was built under R version 3.4.3
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ lubridate::as.difftime() masks base::as.difftime()
## ✖ lubridate::date()        masks base::date()
## ✖ dplyr::filter()          masks stats::filter()
## ✖ lubridate::intersect()   masks base::intersect()
## ✖ dplyr::lag()             masks stats::lag()
## ✖ lubridate::setdiff()     masks base::setdiff()
## ✖ lubridate::union()       masks base::union()
```

```
library(dplyr)
library(tidyr)
```

```
#load data
book = read.csv("book.csv")
booktrain = read.csv("booktrain.csv")
train <- booktrain[, -3]
booktest = read.csv("booktest.csv")
test = booktest[, -3]
order = read.csv("ordersall.csv")
```

```
head(book, 5)
```

```
##     id logtargamt recency frequency    amount  tof Ffiction1 Fclassics3
## 1  914         NA     194         7 318.89478 1703         1          0
## 2  957         NA       3        14 368.05225 2364         1          0
## 3 1406         NA    1489        15 423.29834 2371         0          0
## 4 1414         NA     155         4  71.21704 1290         0          0
## 5 1546         NA     194         6 442.63818 2188         0          0
##   Fcartoons5 Flegends6 Fphilosophy7 Freligion8 Fpsychology9 Flinguistics10
## 1          1         0            0          0            0              0
## 2          2         0            0          0            0              0
## 3          0         0            0          0            0              0
## 4          0         0            0          0            0              0
## 5          1         0            3          1            0              1
##   Fart12 Fmusic14 Ffacsimile17 Fhistory19 Fconthist20 Feconomy21
## 1      1        7            0          0          17          0
## 2      0        4            0          4           4          0
## 3      0        0            0          1          34          0
## 4      0        0            0          0           7          0
## 5      1        3            0          6           7          1
##   Fpolitics22 Fscience23 Fcompsci26 Frailroads27 Fmaps30 Ftravelguides31
## 1           0          0          0            0       0               2
## 2           0          0          1            0       0               0
## 3           0          0          0            0       0               0
## 4           0          0          0            0       0               0
## 5           0          2          0            2       0               1
##   Fhealth35 Fcooking36 Flearning37 FGamesRiddles38 Fsports39 Fhobby40
## 1         2          1           0               0         0        0
## 2         7          0           4               0         0        4
## 3         0          0           0               0         0        0
## 4         0          0           0               0         0        0
## 5         3          0           2               0         0        2
##   Fnature41 Fencyclopaedia44 Fvideos50 Fnonbooks99 Mfiction1 Mclassics3
## 1         0                0         0           0  5.949997          0
## 2         2                1         0           0 12.680038          0
## 3         0                0         0           0  0.000000          0
## 4         0                0         0           0  0.000000          0
## 5         0                0         3           0  0.000000          0
##   Mcartoons5 Mlegends6 Mphilosophy7 Mreligion8 Mpsychology9 Mlinguistics10
## 1   7.643810         0      0.00000    0.00000           0       0.000000
## 2  20.236496         0      0.00000    0.00000           0       0.000000
## 3   0.000000         0      0.00000    0.00000           0       0.000000
## 4   0.000000         0      0.00000    0.00000           0       0.000000
## 5   8.899994         0     26.19841   19.89999           0       6.621227
##      Mart12 Mmusic14 Mfacsimile17 Mhistory19 Mconthist20 Meconomy21
## 1  8.589699 61.42996            0    0.00000   197.93555   0.000000
## 2  0.000000 29.41417            0   53.83902    60.74997   0.000000
## 3  0.000000  0.00000            0   10.12357   413.17480   0.000000
## 4  0.000000  0.00000            0    0.00000    71.21704   0.000000
## 5 15.313187 39.80405            0   64.74493    86.94574   5.112919
##   Mpolitics22 Mscience23 Mcompsci26 Mrailroads27 Mmaps30 Mtravelguides31
## 1           0    0.00000   0.000000     0.000000       0        12.27100
## 2           0    0.00000   5.899998     0.000000       0         0.00000
## 3           0    0.00000   0.000000     0.000000       0         0.00000
## 4           0    0.00000   0.000000     0.000000       0         0.00000
## 5           0   19.84999   0.000000     9.960655       0        10.12357
##   Mhealth35 Mcooking36 Mlearning37 MGamesRiddles38 Msports39 Mhobby40
## 1  14.89999   10.17471    0.000000               0         0  0.00000
## 2  74.15625    0.00000   22.749985               0         0 61.17603
## 3   0.00000    0.00000    0.000000               0         0  0.00000
## 4   0.00000    0.00000    0.000000               0         0  0.00000
## 5  41.51822    0.00000    7.949997               0         0 17.89520
##   Mnature41 Mencyclopaedia44 Mvideos50 Mnonbooks99
## 1   0.00000                0   0.00000           0
## 2  17.15027               10   0.00000           0
## 3   0.00000                0   0.00000           0
## 4   0.00000                0   0.00000           0
## 5   0.00000                0  61.79999           0
```

```
head(booktrain, 5)
```

```
##     id logtargamt    X
## 1 2062          0 7984
## 2 2232          0   NA
## 3 2623          0   NA
## 4 3000          0   NA
## 5 4693          0   NA
```

```
head(booktest, 5)
```

```
##     id logtargamt     X
## 1  914          0 24403
## 2  957          0    NA
## 3 1406          0    NA
## 4 1414          0    NA
## 5 1546          0    NA
```

```
head(order, 5)
```

```
##    id   orddate ordnum category qty     price
## 1 914  2-Dec-09 314037       20   1  9.203247
## 2 914  2-Dec-09 314037       20   1 10.200272
## 3 914 14-Dec-10 499719       36   1 10.174706
## 4 914 14-Dec-10 499719       20   1 10.200272
## 5 914 14-Dec-10 499719       31   1  6.135502
```

```
#train ratio
dim(booktrain)[1] / dim(book)[1]
```

```
## [1] 0.2465221
```

```
#test ratio
dim(booktest)[1] / dim(book)[1]
```

```
## [1] 0.7534779
```

change date in ordersall to datetime

```
order$orddate = lubridate::dmy(order$orddate)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'default/America/
## Chicago'
```

keep only order year

```
order$orddate = lubridate::year(order$orddate)
```

number of orders by id in each year

```
df <- order %>%
  group_by(id,orddate) %>%
  summarise(n = n_distinct(ordnum)) %>%
  #count(id, orddate) %>%
  spread(key = orddate, value = n)
df[is.na(df)] <- 0
```

sale amount by id in each year

needs debugging here

```
#df2 <- order %>%
#  mutate(P = qty*price) %>%
#  group_by(id,orddate) %>%
#  summarise(totalPrice = sum(P)) %>%
#  spread(key = orddate, value = totalPrice)
#df2[is.na(df2)] <- 0
```

```
d <- merge(book,df,all.x=TRUE)
d$`2007`[is.na(d$`2007`)] <- 0
d$`2008`[is.na(d$`2008`)] <- 0
d$`2009`[is.na(d$`2009`)] <- 0
d$`2010`[is.na(d$`2010`)] <- 0
d$`2011`[is.na(d$`2011`)] <- 0
d$`2012`[is.na(d$`2012`)] <- 0
d$`2013`[is.na(d$`2013`)] <- 0
d$`2014`[is.na(d$`2014`)] <- 0
```