rare_sales.R

zamirg13

Fri Apr 20 12:48:12 2018

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```
library(lubridate)
```

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

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

```
library(reshape2)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:reshape2':
##
##     dcast, melt
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
library(plyr)
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:lubridate':
##
##     here
```

```
prices <- read.csv("prices.csv", sep = "|")
items <- read.csv("items.csv", sep = "|")
train <- read.csv("train.csv", sep = "|")

# separate dates (123 days, last date: 01/31/18)
train$year <- year(ymd(train$date))
train$month <- month(ymd(train$date))
train$day <- day(ymd(train$date))
train$weekday <- weekdays(ymd(train$date))

# add id for the unique combination of the "pid" and "size"
items$id <- as.factor(seq(1,length(items$pid)))

# combine new id with the train data_set
```

```r
train_items <- merge(train, items, by.y = c("pid","size"))

# explore sizes(many contain the same information)
#levels(train$size)

# sum of sold items by id
sold_by_id <- ddply(train_items, "id", summarise, sum = sum(units))
ord <- sold_by_id[order(sold_by_id$sum, decreasing = TRUE),]
head(ord, 20)
```

```
##          id  sum
## 3023   3023 2979
## 5886   5886 2643
## 5885   5885 2411
## 6865   6865 1819
## 8189   8189 1694
## 3034   3034 1562
## 9306   9306 1439
## 8188   8188 1427
## 8508   8508 1388
## 426     426 1358
## 2954   2954 1319
## 7243   7243 1289
## 9305   9305 1280
## 4121   4121 1259
## 8509   8509 1237
## 427     427 1224
## 9129   9129 1146
## 7242   7242 1113
## 12041 12041 1044
## 3060   3060 1012
```

```r
# There are 2263 items that were sold only one times
sum(as.numeric(sold_by_id$sum == 1))
```

```
## [1] 2263
```

```r
# which items were sold only one times?
ids <- which(sold_by_id$sum == 1) # id is consistent with the raw number
sum(as.numeric(train_items[ids,]$stock != 0)) # all have non-zero stocks
```

```
## [1] 2263
```

```r
table(train_items[ids,]$month) # rare sold products were sold in average
```

```
##
##   1  10  11  12
## 608 507 620 528
```

```r
# equal amount each month. But the whole other stock: 2263 items were
# sold in 28 days on February. So there is effect of discounts(probably)
# on the sale of these.
```