Competition 1

Tim Reznicek

2024-3-15

```
# Libraries
```

```
# orders dataframe
orders <- read.csv("contest1data/orders.csv")
orders$date <- as.Date(orders$date)
str(orders)
```

```
## 'data.frame':    816574 obs. of  4 variables:
##  $ date  : Date, format: "2020-08-01" "2020-08-01" ...
##  $ userID: int  14477 14477 33883 33883 33883 33883 19087 42266 4286 4286 ...
##  $ itemID: int  2375 14961 8927 27830 29700 27432 9798 18630 31949 10468 ...
##  $ count : int  1 2 1 2 1 4 1 3 1 1 ...
```

```
# items dataframe
items <- read.csv("contest1data/items.csv")
str(items)
```

```
## 'data.frame':    32776 obs. of  8 variables:
##  $ itemID        : int  22665 28640 13526 21399 8504 32122 31956 6237 16971 18385 ...
##  $ manufacturerID: int  861 1366 1090 1090 768 5 1388 1492 288 288 ...
##  $ f1            : int  4 10 10 10 4 4 4 4 6 6 ...
##  $ f2            : int  0 1 0 1 1 1 0 1 0 0 ...
##  $ f3            : int  490 537 511 511 484 491 491 491 314 314 ...
##  $ f4            : int  2 0 0 0 0 0 0 3 0 0 ...
##  $ f5            : int  66 101 0 0 66 66 66 66 45 45 ...
##  $ category      : chr  "[2890, 855, 3908, 3909]" "" "[3270, 163, 284, 1694, 12, 3837, 2422, 3595, 3
```

```
# categories dataframe
categories <- read.csv("contest1data/categories.csv")
str(categories)
```

```
## 'data.frame':    4332 obs. of  2 variables:
##  $ category       : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ parent_category: int  75 1499 1082 3498 1623 2478 1582 3027 2364 3590 ...
```

```
# test dataframe
test <- read.csv("contest1data/test.csv")
str(test)
```

```
## 'data.frame':    10000 obs. of  3 variables:
##  $ userID: int  0 0 13 15 15 20 24 34 34 46 ...
##  $ itemID: int  20664 28231 2690 1299 20968 8272 11340 21146 31244 31083 ...
##  $ ID    : chr  "0-20664" "0-28231" "13-2690" "15-1299" ...
```

Several

```
# Cleaning the data
```