instacart

09/03/2021

Abstract

The main goal of recommender systems is to provide suggestions to online users to make better decisions from many alternatives available over the Web. A better recommender system is directed more towards personalized recommendations by taking into consideration information about a product, such as specifications, purchase histor the users, comparison with other products, and so on, before making recommendations.


```
library(dplyr)
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(ggplot2)
```

```
#Import data

orders<-read.csv("orders.csv")
products<-read.csv("products.csv")
departments<-read.csv("departments.csv")
prior<-read.csv("order_products__prior.csv")
train<-read.csv("order_products__train.csv")
test<-read.csv("sample_submission.csv")
```

```
#merge prior & order

prior_order<-orders %>%
  filter(eval_set=="prior") %>%
    left_join(prior,orders,by=c("order_id"))
```

```r
#merge train and order

train_order<-orders %>%
  filter(eval_set=="train") %>%
    left_join(train,orders,by=c("order_id"))
```

```r
#merge test and order

test_order<-orders %>%
  filter(eval_set=="test") %>%
    left_join(test,orders,by=c("order_id"))
```

```r
dim(prior_order)
```

```
## [1] 32434489       10
```

```r
dim(train_order)
```

```
## [1] 1384617      10
```

```r
dim(test_order)
```

```
## [1] 75000     8
```

```r
## Total users

user_count<-unique(orders$user_id)
length(user_count)
```

```
## [1] 206209
```

```r
#total products
product_count<-unique(products$product_id)
length(product_count)
```

```
## [1] 49688
```

```r
orders_count<-unique(orders$order_id)
length(orders_count)
```

```
## [1] 3421083
```

```r
#total products in prior
product_prior_count<-unique(prior_order$product_id)
length(product_prior_count)
```

```
## [1] 49677
```

```
#total products in train

product_train_count<-unique(train_order$product_id)
length(product_train_count)
```

```
## [1] 39123
```

```
# top 50  products prior
top_products_prior<-prior_order %>%
 group_by(user_id,product_id) %>%
 summarise(tot=n()) %>%
  ungroup() %>%
  group_by(product_id) %>%
  summarise(count1=n()) %>%
  ungroup() %>%
arrange(desc(count1)) %>%
top_n(50)
```

```
## `summarise()` regrouping output by 'user_id' (override with `.groups` argument)

## `summarise()` ungrouping output (override with `.groups` argument)

## Selecting by count1
```

```
#top_products_prior
```

```
#prior user for top 50

prior_users<-prior_order %>%
  filter(product_id %in% top_products_prior$product_id ) %>%
 group_by(user_id,product_id) %>%
   summarise(tot=n()) %>%
  ungroup() %>%
  group_by(user_id) %>%
  summarise(count2=n()) %>%
 arrange(desc(count2))
```

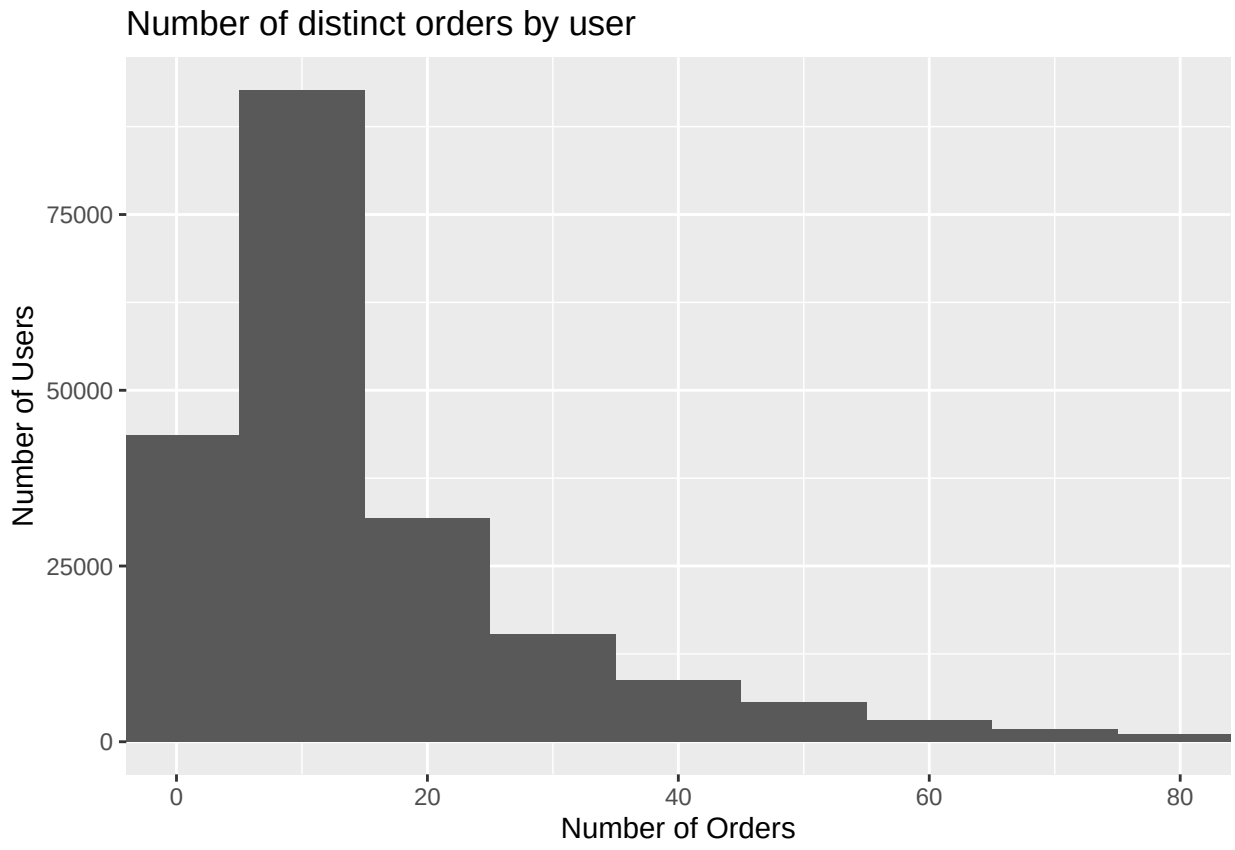
```
## `summarise()` regrouping output by 'user_id' (override with `.groups` argument)

## `summarise()` ungrouping output (override with `.groups` argument)
```

```
#prior_users
```

```
#Number of distinct orders by user

p1<-orders %>%
  group_by(user_id) %>%
  summarise(count_order=n()) %>%
  ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# %>% group_by(count_order) %>%
#   summarise(count_user=n())

ggplot(p1, aes(count_order)) + geom_histogram(binwidth = 10)+labs(title="Number of distinct orders by us
       x ="Number of Orders", y = "Number of Users")+coord_cartesian(xlim = c(0, 80))
```




```
#combine prior and train

all<-rbind(train_order,prior_order)
head(all,100)
```

```
##     order_id user_id eval_set order_number order_dow order_hour_of_day
## 1    1187899       1    train           11         4                 8
## 2    1187899       1    train           11         4                 8
## 3    1187899       1    train           11         4                 8
## 4    1187899       1    train           11         4                 8
## 5    1187899       1    train           11         4                 8
## 6    1187899       1    train           11         4                 8
## 7    1187899       1    train           11         4                 8
## 8    1187899       1    train           11         4                 8
## 9    1187899       1    train           11         4                 8
## 10   1187899       1    train           11         4                 8
## 11   1187899       1    train           11         4                 8
## 12   1492625       2    train           15         1                11
## 13   1492625       2    train           15         1                11
```

```
## 14   1492625       2    train          15         1                11
## 15   1492625       2    train          15         1                11
## 16   1492625       2    train          15         1                11
## 17   1492625       2    train          15         1                11
## 18   1492625       2    train          15         1                11
## 19   1492625       2    train          15         1                11
## 20   1492625       2    train          15         1                11
## 21   1492625       2    train          15         1                11
## 22   1492625       2    train          15         1                11
## 23   1492625       2    train          15         1                11
## 24   1492625       2    train          15         1                11
## 25   1492625       2    train          15         1                11
## 26   1492625       2    train          15         1                11
## 27   1492625       2    train          15         1                11
## 28   1492625       2    train          15         1                11
## 29   1492625       2    train          15         1                11
## 30   1492625       2    train          15         1                11
## 31   1492625       2    train          15         1                11
## 32   1492625       2    train          15         1                11
## 33   1492625       2    train          15         1                11
## 34   1492625       2    train          15         1                11
## 35   1492625       2    train          15         1                11
## 36   1492625       2    train          15         1                11
## 37   1492625       2    train          15         1                11
## 38   1492625       2    train          15         1                11
## 39   1492625       2    train          15         1                11
## 40   1492625       2    train          15         1                11
## 41   1492625       2    train          15         1                11
## 42   1492625       2    train          15         1                11
## 43   2196797       5    train           5         0                11
## 44   2196797       5    train           5         0                11
## 45   2196797       5    train           5         0                11
## 46   2196797       5    train           5         0                11
## 47   2196797       5    train           5         0                11
## 48   2196797       5    train           5         0                11
## 49   2196797       5    train           5         0                11
## 50   2196797       5    train           5         0                11
## 51   2196797       5    train           5         0                11
## 52    525192       7    train          21         2                11
## 53    525192       7    train          21         2                11
## 54    525192       7    train          21         2                11
## 55    525192       7    train          21         2                11
## 56    525192       7    train          21         2                11
## 57    525192       7    train          21         2                11
## 58    525192       7    train          21         2                11
## 59    525192       7    train          21         2                11
## 60    525192       7    train          21         2                11
## 61    880375       8    train           4         1                14
## 62    880375       8    train           4         1                14
## 63    880375       8    train           4         1                14
## 64    880375       8    train           4         1                14
## 65    880375       8    train           4         1                14
## 66    880375       8    train           4         1                14
## 67    880375       8    train           4         1                14
```

```
## 68    880375       8    train           4         1                14
## 69    880375       8    train           4         1                14
## 70    880375       8    train           4         1                14
## 71    880375       8    train           4         1                14
## 72    880375       8    train           4         1                14
## 73    880375       8    train           4         1                14
## 74    880375       8    train           4         1                14
## 75    880375       8    train           4         1                14
## 76    880375       8    train           4         1                14
## 77    880375       8    train           4         1                14
## 78    880375       8    train           4         1                14
## 79   1094988       9    train           4         6                10
## 80   1094988       9    train           4         6                10
## 81   1094988       9    train           4         6                10
## 82   1094988       9    train           4         6                10
## 83   1094988       9    train           4         6                10
## 84   1094988       9    train           4         6                10
## 85   1094988       9    train           4         6                10
## 86   1094988       9    train           4         6                10
## 87   1094988       9    train           4         6                10
## 88   1094988       9    train           4         6                10
## 89   1094988       9    train           4         6                10
## 90   1094988       9    train           4         6                10
## 91   1094988       9    train           4         6                10
## 92   1094988       9    train           4         6                10
## 93   1094988       9    train           4         6                10
## 94   1094988       9    train           4         6                10
## 95   1094988       9    train           4         6                10
## 96   1094988       9    train           4         6                10
## 97   1094988       9    train           4         6                10
## 98   1094988       9    train           4         6                10
## 99   1094988       9    train           4         6                10
## 100  1094988       9    train           4         6                10
##     days_since_prior_order product_id add_to_cart_order reordered
## 1                       14        196                 1         1
## 2                       14      25133                 2         1
## 3                       14      38928                 3         1
## 4                       14      26405                 4         1
## 5                       14      39657                 5         1
## 6                       14      10258                 6         1
## 7                       14      13032                 7         1
## 8                       14      26088                 8         1
## 9                       14      27845                 9         0
## 10                      14      49235                10         1
## 11                      14      46149                11         1
## 12                      30      22963                 1         1
## 13                      30       7963                 2         1
## 14                      30      16589                 3         1
## 15                      30      32792                 4         1
## 16                      30      41787                 5         1
## 17                      30      22825                 6         1
## 18                      30      13640                 7         0
## 19                      30      24852                 8         1
## 20                      30      45066                 9         1
```

```
## 21          30        9387          10          0
## 22          30        5450          11          1
## 23          30       24838          12          0
## 24          30       38547          13          0
## 25          30       19019          14          0
## 26          30       12007          15          0
## 27          30       26352          16          0
## 28          30       22559          17          1
## 29          30       45613          18          1
## 30          30       31883          19          0
## 31          30       12324          20          0
## 32          30       33957          21          1
## 33          30        5699          22          0
## 34          30       31612          23          0
## 35          30       34284          24          0
## 36          30       48523          25          0
## 37          30        2361          26          0
## 38          30       48821          27          0
## 39          30       11913          28          0
## 40          30       45645          29          0
## 41          30        1757          30          0
## 42          30       21329          31          0
## 43           6       15349           1          1
## 44           6       19057           2          0
## 45           6       16185           3          0
## 46           6       21413           4          1
## 47           6       20843           5          0
## 48           6       20114           6          0
## 49           6       48204           7          0
## 50           6       40706           8          1
## 51           6       21616           9          1
## 52           6       12053           1          0
## 53           6       47272           2          1
## 54           6       37999           3          1
## 55           6       13198           4          1
## 56           6       43967           5          1
## 57           6       40852           6          1
## 58           6       17638           7          1
## 59           6       29894           8          1
## 60           6       45066           9          1
## 61          10       15937           1          1
## 62          10        5539           2          0
## 63          10       10960           3          0
## 64          10       23165           4          1
## 65          10       22247           5          0
## 66          10        4853           6          0
## 67          10       27104           7          0
## 68          10        7058           8          0
## 69          10       41259           9          0
## 70          10       37803          10          0
## 71          10       48230          11          0
## 72          10       47766          12          0
## 73          10       31717          13          0
## 74          10       21903          14          1
```

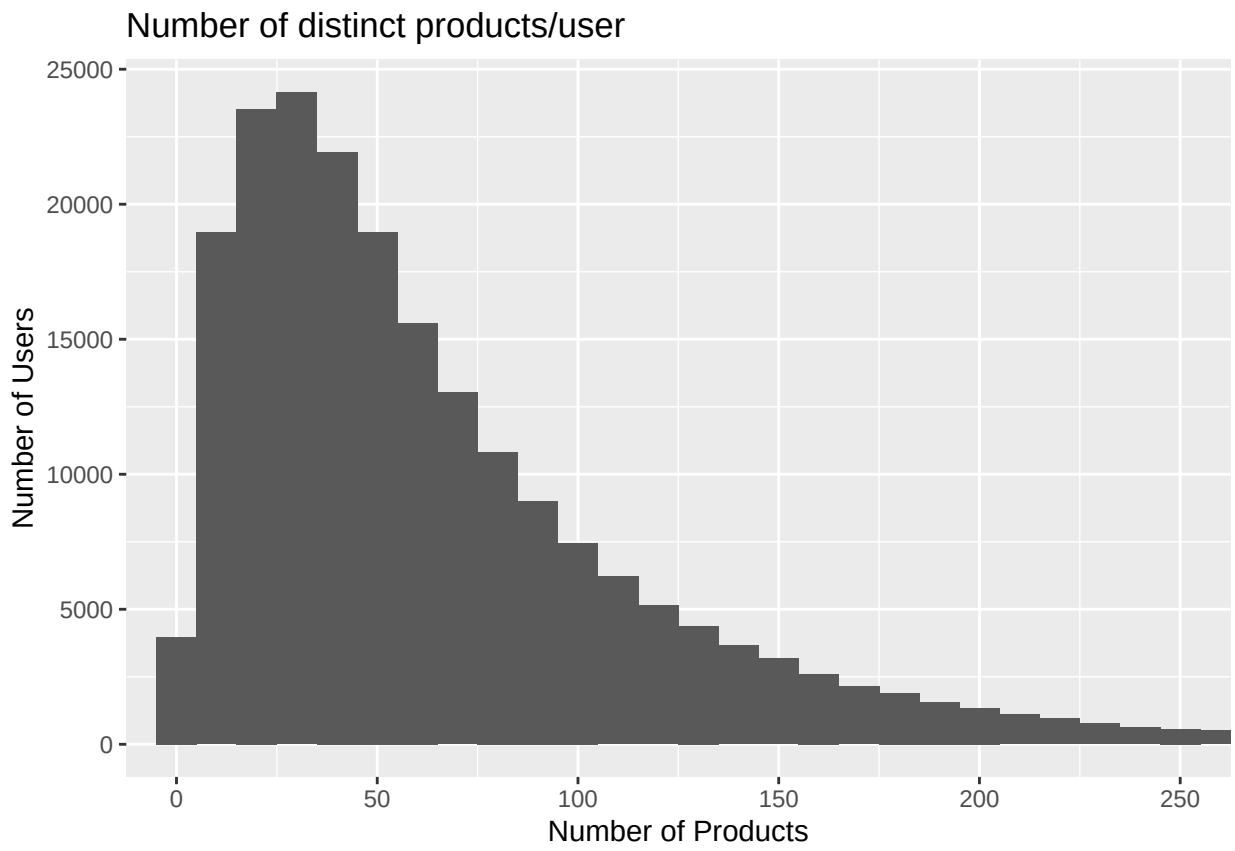
```
## 75                      10      25659                15         0
## 76                      10      41540                16         1
## 77                      10      48121                17         0
## 78                      10       2846                18         0
## 79                      30      27555                 1         1
## 80                      30      42347                 2         1
## 81                      30      27596                 3         1
## 82                      30       8834                 4         1
## 83                      30      26604                 5         1
## 84                      30      12075                 6         1
## 85                      30       8467                 7         1
## 86                      30      38988                 8         1
## 87                      30      30252                 9         1
## 88                      30      18926                10         1
## 89                      30      24954                11         1
## 90                      30      40571                12         1
## 91                      30       1559                13         1
## 92                      30      33754                14         1
## 93                      30      29594                15         1
## 94                      30      17600                16         1
## 95                      30      42828                17         1
## 96                      30      10132                18         1
## 97                      30      20899                19         1
## 98                      30      27973                20         1
## 99                      30      41844                21         1
## 100                     30      30967                22         1
```

```
#Number of distinct products/user

p2<-all %>%
  group_by(user_id,product_id) %>%
  summarise(count3=n()) %>%
  select(user_id,product_id,count3) %>%
  ungroup() %>%
  group_by(user_id) %>%
  summarise(count_product=n()) %>%
  ungroup()
```

```
## `summarise()` regrouping output by 'user_id' (override with `.groups` argument)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# %>% group_by(count_product) %>%
#   summarise(count_user=n())



ggplot(p2, aes(count_product)) + geom_histogram(binwidth = 10)+labs(title="Number of distinct products/u
       x ="Number of Products", y = "Number of Users")+coord_cartesian(xlim = c(0, 250))
```

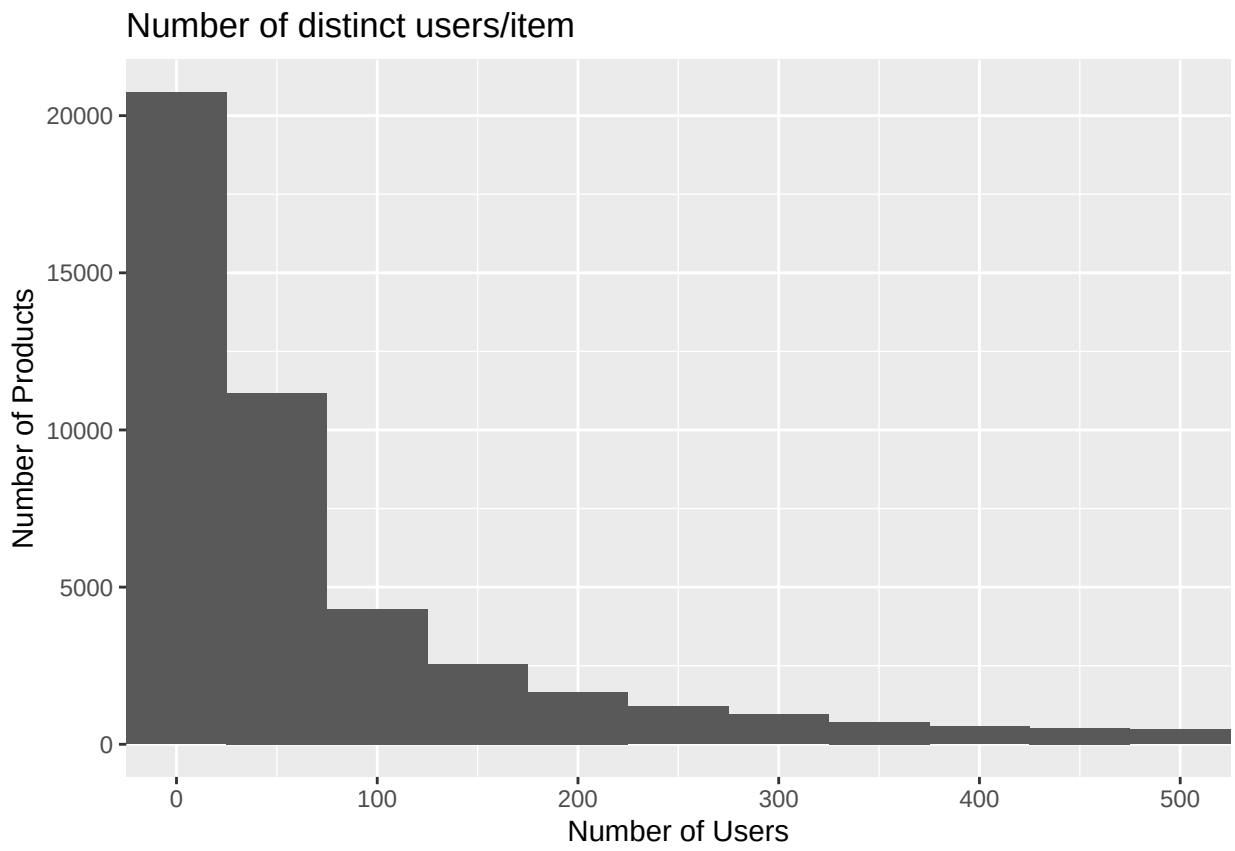



```
#Number of distinct users/item



p3<-all %>%
  group_by(product_id,user_id) %>%
  summarise(count4=n()) %>%
   select(user_id,product_id,count4) %>%
  ungroup() %>%
  group_by(product_id) %>%
  summarise(count_user=n()) %>%
  ungroup()
```

```
## `summarise()` regrouping output by 'product_id' (override with `.groups` argument)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# %>% group_by(count_user) %>%
#   summarise(count_product=n())


ggplot(p3, aes(count_user)) + geom_histogram(binwidth = 50)+labs(title="Number of distinct users/item",
        x ="Number of Users", y = "Number of Products")+coord_cartesian(xlim = c(0, 500))
```




```
#Extract columns for matrix
transactions<-prior_order[,c("user_id","product_id","order_id")]


#calculate sparsity of the dataframe
sparsity<-1-sum(transactions==0)/prod(dim(transactions))
sparsity
```

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

```
#Sample dataset for Processing (top 50 products)

transactions_sample<-transactions %>%
                     filter(product_id %in%top_products_prior$product_id)
```

```
dim(transactions_sample)
```

```
## [1] 5270142       3
```

```
sparsity<-1-sum(transactions_sample==0)/prod(dim(transactions))
sparsity
```

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