Homework10

Ivan Ojiambo

December 2, 2017

Exe 1

a

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

```
## Warning: package 'data.table' was built under R version 3.4.2
```

```
data <- fread("instacart_1m.csv")
```

```
##
Read 97.0% of 1000000 rows
Read 1000000 rows and 15 (of 15) columns from 0.088 GB file in 00:00:03
```

```
A= subset(data, order_id==2539329, user_id =1)
ids <- A$product_id
A$product_name
```

```
## [1] "Soda"
## [2] "Organic Unsweetened Vanilla Almond Milk"
## [3] "Original Beef Jerky"
## [4] "Aged White Cheddar Popcorn"
## [5] "XL Pick-A-Size Paper Towel Rolls"
```

b

```
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
filitered_data <- filter(data, !is.na(product_id)) %>% filter( product_id %in% ids, user_id != 1)
nrow(filitered_data)
```

```
## [1] 1829
```

1829 Rows remained

c

```
filitered_data[is.na(filitered_data)] <- 0

df <- dcast(filitered_data, order_id~product_id, value.var="product_id", fill=0)
df[, 2:6][df[, 2:6]>0] <- 1
dim(df)
```

```
## [1] 1814    6
```

```
#matrix_A <- dcast(A, order_id~product_id, value.var="product_id",fill=0)
#matrix_A[, 2:6][matrix_A[, 2:6]>0] <- 1
```

Dimension of the matrix is 1814 by 6

(d) Now compute euclidean distances between the initial order of A to the orders from the matrix.

```
dist <- function(x, y){
  sqrt(sum((x-y)^2))
}

distance <- NULL
for (i in 1:nrow(df)){
  distance[i] <- dist(df[i, 2:6],  c(1, 1, 1, 1, 1))
}

output <- data.frame(order_id= df$order_id, dist= distance)
```

(e) Select top N closest (in terms of euclidean distance) orders to the first order by A. Explain your choice of N.

```
selected_order <-  filter(output, dist == min(dist))$order_id

common_products <-
  filter(data, order_id %in% selected_order) %>% filter(!product_name %in% A$product_name)

group_by(common_products, product_name) %>% summarise(count= n())%>% arrange(desc(count)) %>% top_n(5, c
```

```
## # A tibble: 5 x 2
##               product_name count
##                      <chr> <int>
## 1   Bag of Organic Bananas     4
## 2 0% Greek Strained Yogurt     3
## 3             Dried Mangos     3
## 4     Organic Baby Carrots     3
## 5              Raspberries     3
```

(f) Based on these N closest orders, say, which product we should advertise to A next time he/she comes to the shop (NB! apart from the ones that were bought first time)? Why?

We should advertise to him *** Organic Bananas** because a good number of client who buy the same product as him often buy Organic Bananas.

g) Look at the next shopping basket of A. Do you see this product among purchases?

```
Next_order_A= filter(data,  user_id ==1, order_number == 2) %>% group_by(product_name) %>% summarise(cou
top_n(Next_order_A, 5, count)
```

```
## # A tibble: 6 x 2
##                 product_name count
##                        <chr> <int>
## 1 Aged White Cheddar Popcorn     1
## 2     Bag of Organic Bananas     1
## 3      Cinnamon Toast Crunch     1
## 4        Original Beef Jerky     1
## 5                  Pistachios     1
## 6                       Soda     1
```

- Yes, its is there
- KNN
- ***Content-based filtering*** Is based on a description of the item and a profile of the user's preferences.In a content-based recommender system, keywords are used to describe the items and a user profile is built to indicate the type of item the user likes
- ***Hybrid recommender systems*** combines collaborative filtering and content-based filtering [wikipedia]

Exe2

What is the overall most popular product? Can you guess the answer before finding it out?

```
select(data, product_name) %>% group_by(product_name) %>% summarise(count = n()) %>%
   arrange(desc(count)) %>% top_n(3, count)
```

```
## # A tibble: 3 x 2
##             product_name count
##                    <chr> <int>
## 1                 Banana 13731
## 2 Bag of Organic Bananas 11696
## 3   Organic Strawberries  8295
```

Banana is the most popular product

b What is the most popular product at hour 00? At hour 01? ... At hour 23? Provide the answer as a list or vector of 24 product names, one corresponding to each hour.

```
count_prod <- function(dataframe){
  grouped_data <- group_by(dataframe, product_name)
  summarise(grouped_data, count = n())
}


products <- NULL
for(i in 0:23){

  result1 <- filter(data, order_hour_of_day == i) %>% count_prod() %>% filter(count==max(count))
  #print(paste(result))
  products[i+1]<- result1$product_name[1]

}
data.frame(hour=0:23, prod = products)
```

```
##    hour                    prod
## 1     0 Bag of Organic Bananas
## 2     1 Bag of Organic Bananas
## 3     2 Bag of Organic Bananas
## 4     3                  Banana
## 5     4 Bag of Organic Bananas
## 6     5                  Banana
## 7     6                  Banana
## 8     7                  Banana
## 9     8 Bag of Organic Bananas
## 10    9                  Banana
## 11   10                  Banana
## 12   11                  Banana
## 13   12                  Banana
## 14   13                  Banana
## 15   14                  Banana
## 16   15                  Banana
## 17   16                  Banana
## 18   17                  Banana
## 19   18                  Banana
## 20   19                  Banana
## 21   20                  Banana
## 22   21                  Banana
## 23   22                  Banana
## 24   23 Bag of Organic Bananas
```

c

```
relative_popularity <- function(df, hr){
  grouped_data_hour <- filter(df, order_hour_of_day==i) %>% count_prod()
  grouped_data_prod <- filter(df, product_name %in% grouped_data_hour$product_name)%>% count_prod()
  mutate(grouped_data_hour,rel_dist=grouped_data_hour$count/grouped_data_prod$count)
```

```
}

output <- NULL
for (i in 0:23) {
  grouped_data_ratio <- relative_popularity(data, i)
  output[i+1] <- filter(grouped_data_ratio, rel_dist == max(grouped_data_ratio$rel_dist))$product_name
}
output1 <- data.frame(hour =0:23, product_name = output )
as.data.table(output1)
```

```
##     hour
##  1:    0
##  2:    1
##  3:    2
##  4:    3
##  5:    4
##  6:    5
##  7:    6
##  8:    7
##  9:    8
## 10:    9
## 11:   10
## 12:   11
## 13:   12
## 14:   13
## 15:   14
## 16:   15
## 17:   16
## 18:   17
## 19:   18
## 20:   19
## 21:   20
## 22:   21
## 23:   22
## 24:   23
##     hour
##                                                                product_name
##  1:                                            3 Ply Wheat Straw Bath Tissue
##  2: 3D White Brilliance Vibrant Peppermint Flouride Anticavity Toothpaste
##  3:                                  Aloe & Green Tea Natural Room Freshener
##  4:                                                  CafÃ© Caramel Shake
##  5:                                       Aged Vermont White Cheddar Cheese
##  6:                                                 85% Dark Chocolate Bar
##  7:                     1000 Roses Heavenly Night Cream for Sensitive Skin
##  8:                                           100% Juice Orange Pineapple
##  9:                                                   100% Pure Jojoba Oil
## 10:                                         100% Beeswax Hand Dipped Tapers
## 11:                                \\""Mies Vanilla Rohe\\"" Ice Cream Bars
## 12:                                      \\""Constant Comment\\"" Black Tea
## 13:                                         1,000 Mg Vitamin C Super Orange
## 14:                                               0% Fat Peach Greek Yogurt
## 15:                                  10.25\\"" Elegant Fluted Party Plates
```

```
## 16:                          100% Cotton 16 Ply Strength 25 ft Cooking Twine
## 17:                                                              1 Ply Napkins
## 18:                                        0% Milkfat Greek Plain Yogurt
## 19:                                1/3 Less Fat Chive & Onion Cream Cheese
## 20:                                                    100% Apple Cider
## 21:                                   100% Natural Zero Calorie Sweetener
## 22:                         100 Calorie Healthy Pop Butter Microwave Pop Corn
## 23:                         1% Hydrocortisone Anti-Itch Cream, Tube Anti-Itch
## 24:                                       9 Inch Graham Cracker Pie Crust
##                                                                 product_name
```

d

```
draw_table <- function(dataframe){
  output <- as.matrix(setNames(data.frame(matrix(ncol = 24, nrow = 24)), c(0:23)))
  for(i in 1:24){
    prod <- dataframe$product_name[i]
    r2 <- filter(data, product_name ==prod)
    for(j in 0:23) {
      grouped_data_hour <- filter(data, order_hour_of_day==j, product_name == prod)
      output[i, j+1] <- nrow(grouped_data_hour)/nrow(r2)
    }

   }
   return(output)

}
```

e

```
data_300 <- group_by(data, product_name) %>% summarise(count=n()) %>% arrange(desc(count)) %>% top_n(300
output3 <- NULL
for (i in 0:23) {
  grouped_data_ratio <-filter(data, product_name %in% data_300$product_name )%>% relative_popularity(i)
  output3[i+1] <- filter(grouped_data_ratio, rel_dist == max(grouped_data_ratio$rel_dist))$product_name
}
(output3 <- data.frame(hour = 0:23, product_name = output3 ))
```

```
##    hour                              product_name
## 1     0                Grapefruit Sparkling Water
## 2     1                          Watermelon Chunks
## 3     2                Roasted Red Pepper Hummus
## 4     3                     Natural Spring Water
## 5     4                Roasted Red Pepper Hummus
## 6     5                     Organic Sunday Bacon
## 7     6                     Organic Sunday Bacon
## 8     7    Organic Plain Greek Whole Milk Yogurt
## 9     8                   Organic Tortilla Chips
## 10    9                                Trail Mix
## 11   10                                     Soda
## 12   11                 0% Greek Strained Yogurt
```

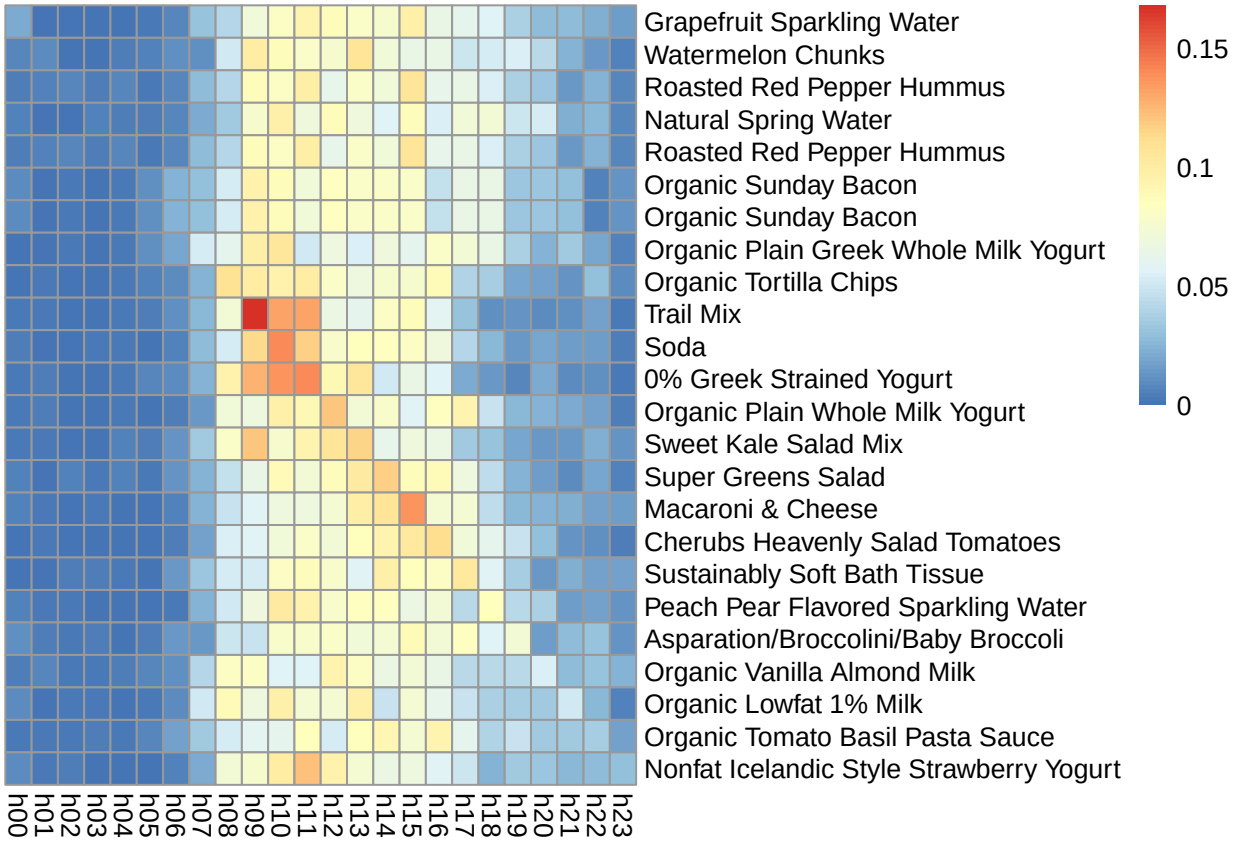
```
## 13   12          Organic Plain Whole Milk Yogurt
## 14   13                     Sweet Kale Salad Mix
## 15   14                       Super Greens Salad
## 16   15                        Macaroni & Cheese
## 17   16          Cherubs Heavenly Salad Tomatoes
## 18   17             Sustainably Soft Bath Tissue
## 19   18      Peach Pear Flavored Sparkling Water
## 20   19      Asparation/Broccolini/Baby Broccoli
## 21   20              Organic Vanilla Almond Milk
## 22   21                   Organic Lowfat 1% Milk
## 23   22         Organic Tomato Basil Pasta Sauce
## 24   23 Nonfat Icelandic Style Strawberry Yogurt
```

f

```
library(pheatmap)
```

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

```
##drawing table for visulaisation
X <- draw_table(output3)

rownames(X)<-output3$product_name
colnames(X)<-sprintf("h%02d",0:23)

pheatmap(X,cluster_rows=F,cluster_cols=F)
```



- I think the diagonal illustrates that each of the 24 products was more popular than the others at a certain hour

g

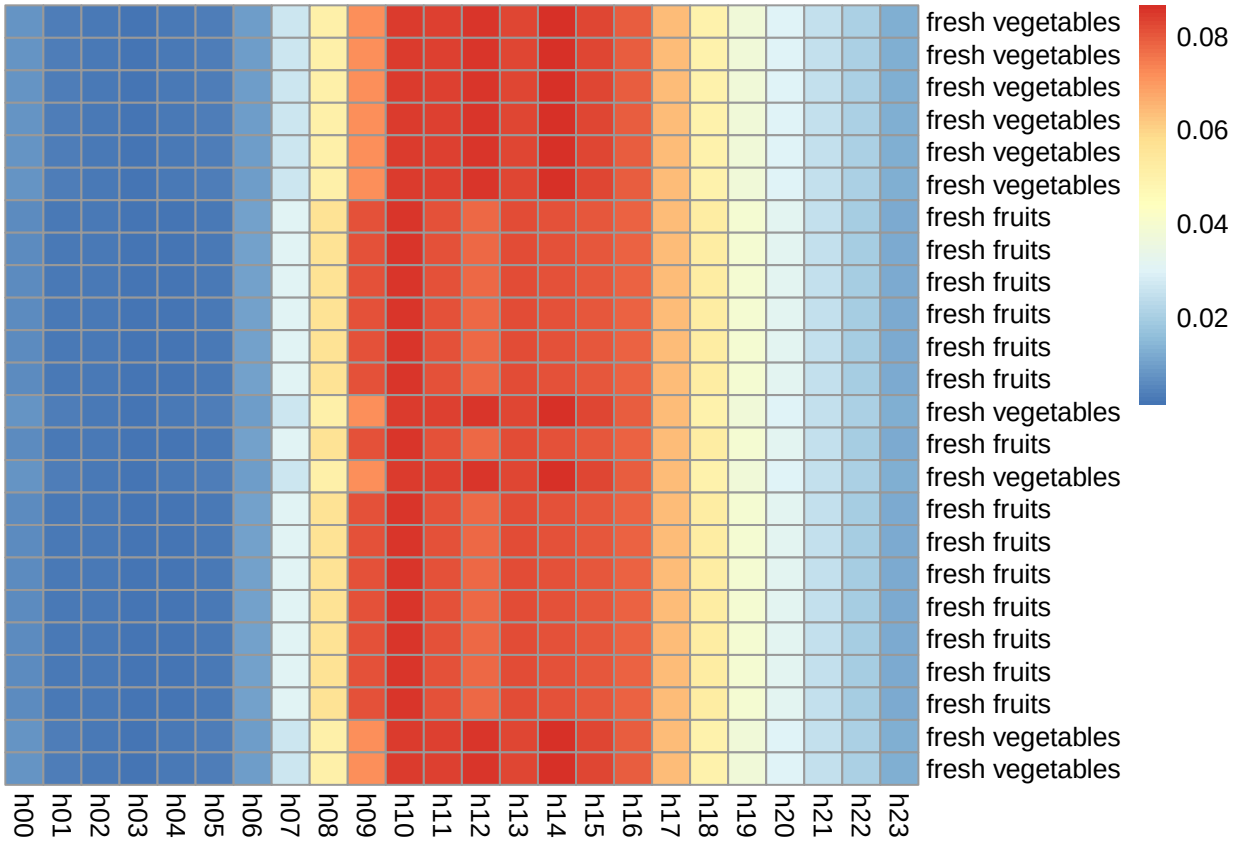
```
products2 <- NULL
for(i in 0:23){
  result2 <- filter(data, order_hour_of_day == i) %>%
    group_by(aisle)%>% summarise(count=n()) %>% filter(count==max(count))
  products2[i+1]<- result2$aisle[1]

}
(aisle <- data.frame(hour=0:23, aisle = products2))
```

```
##    hour            aisle
## 1     0 fresh vegetables
## 2     1 fresh vegetables
## 3     2 fresh vegetables
## 4     3 fresh vegetables
## 5     4 fresh vegetables
## 6     5 fresh vegetables
## 7     6     fresh fruits
## 8     7     fresh fruits
## 9     8     fresh fruits
## 10    9     fresh fruits
```

```
## 11   10     fresh fruits
## 12   11     fresh fruits
## 13   12 fresh vegetables
## 14   13     fresh fruits
## 15   14 fresh vegetables
## 16   15     fresh fruits
## 17   16     fresh fruits
## 18   17     fresh fruits
## 19   18     fresh fruits
## 20   19     fresh fruits
## 21   20     fresh fruits
## 22   21     fresh fruits
## 23   22 fresh vegetables
## 24   23 fresh vegetables
```

h

```
output <- as.matrix(setNames(data.frame(matrix(ncol = 24, nrow = 24)), c(0:23)))
for(j in 1:24){
   a <- aisle$aisle[j]
  for(i in 0:23){
    count1 <- filter(data, aisle == a) %>%
      group_by(aisle)%>% summarise(count=n())

    count2 <-  filter(data, aisle == a, order_hour_of_day == i) %>%
      group_by(aisle)%>% summarise(count=n())

    output[j, i+1] <- count2$count/count1$count

  }

}
Y <- output

rownames(Y)<-aisle$aisle
colnames(Y)<-sprintf("h%02d",0:23)

pheatmap(Y,cluster_rows=F,cluster_cols=F)
```



- For all the popular aisle product , almost none is bought from 00hours to 700 hours.
- For all the popular aisle product , there demand is highes from 9:00 hours to 16:00 hours.
- For all the popular aisle product, the demand begins going low from 17:00 hours and at 23:00 hrs, there is no more sales for any of the aisle products.
- The diagonal no longer stands out.

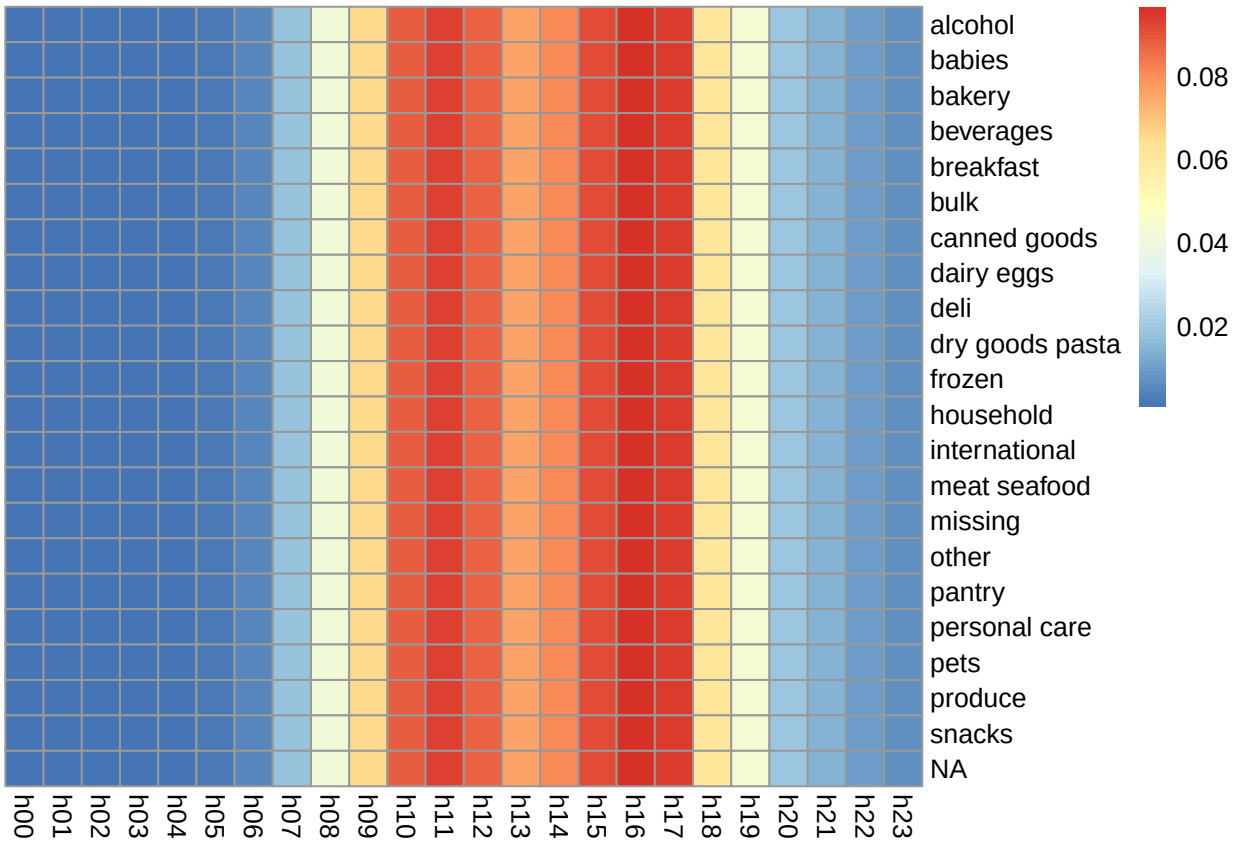
##i

```
depatments <- group_by(data, department) %>% summarise(count = n())
depatments$department[1]
```

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

```
output <- as.matrix(setNames(data.frame(matrix(ncol = 24, nrow = 22)), c(0:23)))
for(j in 1:22){
   a <-depatments$department[1]
  for(i in 0:23){

    count1 <- filter(data, department == a) %>%
      group_by(department)%>% summarise(count=n())

    count2 <-  filter(data, department == a, order_hour_of_day == i) %>%
      group_by(department)%>% summarise(count=n())

    output[j, i+1] <- count2$count/count1$count
```

```
  }

}
Z <- output

rownames(Z)<-depatments$department
colnames(Z)<-sprintf("h%02d",0:23)

pheatmap(Z,cluster_rows=F,cluster_cols=F)
```



- Its also true for department, no department sells from 00 hrs to 6:00 hrs
- most of the department make high sells at 11:00hrs, 16:00 hrs and 17:00 hrs.