Homework Week 7

Athithyaa Muthuraman, Brighton Matibe, Vijay Athithyaa, Venkatesh Prasad, Harish Kumar, Ragul Velmurugan
Foundations of Programming in Business Analytics
BUAN 6333.501
Dr. Waseem Shahid
Fall-2024 / Tuesday 7:00 pm - 9:45 pm

Exercise 1:
Choose a leader for your group a submit the name and email of the chosen leader

Name: Athithyaa Muthuraman
Email: axm190160@utdallas.edu

Exercise 2:
a) Download the data file x.csv, where x is a number, from your group project dataset. Get an overview of the data by using str(). This data file contains transaction records for a grocery store.

```
> data2 <- read.csv("2.csv", header = TRUE, sep = ",")
> View(data2)
> str(data2)
'data.frame':   6000 obs. of  36 variables:
 $ id          : int  6344 9445 8441 1431 6761 7362 1447 5685 5464 2253 ...
 $ basket_value: num  2.5 1.6 1.1 1.2 3.4 1.2 1.4 13.2 2.7 7.9 ...
 $ date        : chr  "5/19/2023" "5/17/2023" "7/31/2023" "7/26/2023" ...
 $ day         : chr  "Friday" "Wednesday" "Monday" "Wednesday" ...
 $ Item_1      : chr  "other vegetables" "root vegetables" "citrus fruit" "whole milk" ...
 $ Item_2      : chr  "whole milk" "butter milk" "" "shopping bags" ...
 $ Item_3      : chr  "" "specialty cheese" "" "" ...
 $ Item_4      : chr  "" "specialty bar" "" "" ...
 $ Item_5      : chr  "" "" "" "" ...
 $ Item_6      : chr  "" "" "" "" ...
 $ Item_7      : chr  "" "" "" "" ...
 $ Item_8      : chr  "" "" "" "" ...
 $ Item_9      : chr  "" "" "" "" ...
 $ Item_10     : chr  "" "" "" "" ...
 $ Item_11     : chr  "" "" "" "" ...
 $ Item_12     : chr  "" "" "" "" ...
 $ Item_13     : chr  "" "" "" "" ...
 $ Item_14     : chr  "" "" "" "" ...
 $ Item_15     : chr  "" "" "" "" ...
 $ Item_16     : chr  "" "" "" "" ...
 $ Item_17     : chr  "" "" "" "" ...
 $ Item_18     : chr  "" "" "" "" ...
 $ Item_19     : chr  "" "" "" "" ...
 $ Item_20     : chr  "" "" "" "" ...
 $ Item_21     : chr  "" "" "" "" ...
 $ Item_22     : chr  "" "" "" "" ...
 $ Item_23     : chr  "" "" "" "" ...
 $ Item_24     : chr  "" "" "" "" ...
 $ Item_25     : chr  "" "" "" "" ...
 $ Item_26     : chr  "" "" "" "" ...
 $ Item_27     : chr  "" "" "" "" ...
 $ Item_28     : chr  "" "" "" "" ...
 $ Item_29     : chr  "" "" "" "" ...
 $ Item_30     : chr  "" "" "" "" ...
 $ Item_31     : chr  "" "" "" "" ...
 $ Item_32     : chr  "" "" "" "" ...
```

b) These records are not clean, there might be duplicates or invalid records. Invalid records include empty ids, date, day, or negative basket value. How many duplicates are there? How many invalid records are there? Clean up the data so you have only the valid records without any duplicates.
c) How many unique grocery items are in the data file?

```
> which(duplicated(data2))
  [1]  285  402  734  749  817  824  842 1139 1148 1268 1274 1289 1323 1335 1389 1402 1407 1412 1417 1430
 [21] 1449 1557 1580 1737 1762 1797 1841 1866 1907 1944 2056 2065 2161 2166 2187 2200 2203 2272 2277 2295
 [41] 2301 2302 2358 2564 2669 2717 2787 2825 2851 2872 2878 2922 2950 2966 3032 3037 3067 3146 3186 3200
 [61] 3229 3267 3271 3281 3295 3312 3313 3332 3337 3366 3375 3386 3397 3445 3491 3492 3548 3552 3576 3596
 [81] 3599 3617 3665 3678 3697 3711 3743 3745 3756 3790 3792 3829 3845 3848 3868 3871 3881 3888 3970 4027
[101] 4101 4157 4167 4180 4184 4203 4214 4241 4258 4270 4274 4276 4309 4327 4361 4368 4369 4371 4384 4395
[121] 4399 4433 4441 4456 4483 4495 4509 4518 4567 4594 4600 4616 4634 4636 4674 4676 4681 4695 4703 4740
[141] 4758 4765 4813 4815 4816 4844 4855 4859 4866 4893 4906 4938 4942 4943 4977 4984 4999 5007 5010 5045
[161] 5058 5071 5079 5081 5092 5119 5124 5139 5143 5153 5167 5175 5194 5224 5242 5365 5382 5389 5400 5416
[181] 5421 5427 5428 5434 5447 5457 5478 5481 5485 5505 5534 5571 5577 5593 5594 5614 5617 5627 5636 5672
[201] 5677 5685 5707 5711 5714 5725 5750 5761 5764 5789 5816 5822 5830 5837 5842 5845 5869 5885 5897 5905
[221] 5939 5949 5991 5998
> data2 <- data2[!duplicated(data2),]
> View(data2)
> data2 <- data2[data2$basket_value>0,]
> View(data2)
> which(is.na(data2$id))
 [1]   29   41   64  145  168  241  403  408  870  883  942 1032 1034 1107 1198 1359 1472 1488 1575 1657
[21] 1674 1768 1786 1851 1901 1932 1945 1958 1979 2213 2538 2547 2602 2633 2658 2666 2773 2836 2838 2906
[41] 3017 3041 3066 3085 3128 3152 3235 3575 3722 3955 3977 4071 4076 4091 4330 4344 4636 4639 4749 4915
[61] 5063 5157 5194 5202 5203 5327 5360 5453 5487 5580 5632 5640 5641
> data2 <- data2[!is.na(data2$id),]
> which(is.na(data2$date))
integer(0)
> data2[data2$date=="",]
       id basket_value date       day             Item_1                Item_2           Item_3
213  1767          3.5       Thursday    other vegetables             rolls/buns
573  3404          2.6       Saturday          whole milk           bottled water
599  1262          1.5         Friday              yogurt               softener
702  1817          1.4         Friday                pork           bottled water     bottled beer
1130 3895         11.2       Saturday         frankfurter                   beef other vegetables
1152 1503          1.5       Saturday              yogurt
1251 2459          1.5         Friday              yogurt                waffles
1348 9718          6.9       Thursday           pip fruit        other vegetables           yogurt
1375 6292          2.2         Sunday          rolls/buns
1386 2947          3.7      Wednesday              butter                yogurt       rolls/buns
1515 7978          1.7       Thursday                soda           bottled beer           liquor
1707 9468          2.5         Friday    other vegetables             whole milk      brown bread
1751  588          4.8         Sunday                pork          tropical fruit       whole milk
1758  520          2.9         Sunday              yogurt           bottled water           dishes
1941 7593          2.5       Thursday    other vegetables             whole milk      brown bread

 [ reached 'max' / getOption("max.print") -- omitted 35 rows ]
> data2 <- data2[!data2$date=="",]
> View(data2)
> which(is.na(data2$day))
integer(0)
> data2[data2$day=="",]
 [1] id          basket_value date        day         Item_1      Item_2      Item_3
 [8] Item_4      Item_5       Item_6      Item_7      Item_8      Item_9      Item_10
[15] Item_11     Item_12      Item_13     Item_14     Item_15     Item_16     Item_17
[22] Item_18     Item_19      Item_20     Item_21     Item_22     Item_23     Item_24
[29] Item_25     Item_26      Item_27     Item_28     Item_29     Item_30     Item_31
[36] Item_32
<0 rows> (or 0-length row.names)
> data2 <- data2[!data2$day=="",]
> View(data2)
> library(dplyr)
> length(unique(unlist(data2 %>% select(starts_with("Item_")))))
[1] 169
```