:

,

Name:	Ian Mouzon
email:	imouzon@iastate.edu
Instructions:	
Assignment:	
Due Date:	

I am using the following packages to create this feature matrix:

```
library(ggplot2)
library(lubridate)
library(dplyr)
library(reshape2)
library(sqldf)
```

0.1 Getting the data

I read the raw and clean data into R using the following simple commands:

```
# training set ('historical data')
d = readRDS("~/dmc2015/data/clean_data/clean_simple.rds")
```

0.1.1 Adding Batch ID

The data set `clean_simple` does not have batch information. I will have to add it in: I read the batch ID file as follows: %– readBatch: R code (Code in Document)

```
# batch features in ~/dmc2015/features/feature_files/batchInfo_test.csv,
# batchInfo_train.csv
bit = readRDS("~/dmc2015/features/feature_files/batchInfo_train.rds")
bic = readRDS("~/dmc2015/features/feature_files/batchInfo_test.rds")
bi = rbind(bit, bic)
```

Fix the formatting:

```
bi$couponsReceivedTime = period_to_seconds(bi$couponsReceivedTime)/3600
bi$orderTimeTime = period_to_seconds(bi$orderTimeTime)/3600
bi = list(train = bi[which(bi$orderID <= 6053), ], class = bi[which(bi$orderID >
    6053), ])
```

Add it to the clean data:

```
d$train = d$train %>% left_join(bi$train, by = "orderID")
d$class = d$class %>% left_join(bi$class, by = "orderID")
```

0.2 Creating Historical, Validation, and Training Sets

0.2.1 Set 1: Random Sampling on Coupons Sets

```
sample.set = d$train[, c("orderID", "couponID1", "couponID2", "couponID3")] %>%
    gather(couponCol, couponID, -orderID) %>% arrange(orderID) %>% select(orderID,
    couponID) %>% mutate(couponID = factor(couponID))
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

I would like to use 1/3 of my data as historical information in this case. This means that I would like to get about 33% of each coupons total number.
That is, for n_i coupons of type i we want $\frac{1}{3}n_i$ coupons of type i in our historical pool.

```
set.seed(1999)
H1.orderIDs = d$train$orderID[sample(1:nrow(d$train))[1:round(nrow(d$train)/3)]]

H1 = d$train[which(d$train$orderID %in% H1.orderIDs), ]
T1 = d$train[which(!(d$train$orderID %in% H1.orderIDs)), ]

set.seed(7)
V1.orderIDs = T1$orderID[sample(1:nrow(T1))[1:round(nrow(T1)/3)]]
V1 = T1[which(T1$orderID %in% V1.orderIDs), ]
T1 = T1[which(!(T1$orderID %in% V1.orderIDs)), ]

# feature matrix
F1 = list(orderids = list(h = H1.orderIDs, v = V1.orderIDs), H = H1, V = V1,
    T = T1, C = d$class)
```

0.2.2 Set 2: Random Sampling on Users

```
set.seed(1979)
H2.userIDs = sample(unique(d$train$userID))[1:round(length(unique(d$train$userID))/3)]

H2 = d$train[which(d$train$userID %in% H2.userIDs), ]
T2 = d$train[which(!(d$train$userID %in% H2.userIDs)), ]

set.seed(17)
V2.userIDs = T2$userID[sample(1:nrow(T2))[1:round(nrow(T2)/3)]]
V2 = T2[which(T2$userID %in% V2.userIDs), ]
T2 = T2[which(!(T2$userID %in% V2.userIDs)), ]

# feature matrix
F2 = list(userids = list(h = H2.userIDs, v = V2.userIDs), H = H2, V = V2, T = T2,
    C = d$class)
```

0.2.3 Set 3: Batch Number

```
H3 = d$train[which(d$train$batchID %in% 1:7), ]
T3 = d$train[which(d$train$batchID == 8), ]
V3 = d$train[which(d$train$batchID == 9), ]
C3 = d$class
F3 = list(H = H3, V = V3, T = T3, C = d$class)
```

0.3 Write out the sets

```
saveRDS(F1, file = "../data/featureMatrix/HTVset1.rds")
saveRDS(F2, file = "../data/featureMatrix/HTVset2.rds")
saveRDS(F3, file = "../data/featureMatrix/HTVset3.rds")
```