:

Coupon Similarity Cluster Categories

Name:	Ian Mouzon
email:	imouzon@iastate.edu

I am using the following packages:

```
library(ggplot2)
library(lubridate)
library(xtable)
library(foreach)
library(rCharts)
library(magrittr)
library(tidyr)
library(dplyr)
library(reshape2)
library(gtools)
library(sqldf)
library(missForest)
source("./R/renm.R")
```

and our working directory is set to `dmc2015/ian`.
Getting the Data and Manipulations
I am using our new clean data - so should you

```
d = readRDS("~/dmc2015/data/clean_data/universalCleanData.rds")
```

I can melt the columns by coupon using the following:

```
source("~/dmc2015/ian/r/stackCoupons2.R")
dm = stackCoupons2(d, idcols = c(1:4, 32:49))
```

I and can split the columns of product group using:

```
source("~/dmc2015/ian/r/splitColumn.R")
dmc = splitColumn(dm, "categoryIDs", "orderID", splitby = ":")
```

```
## Loading required package: tcltk
```

0.1 Creating the Item ID

```
d$item_id1 = with(d, paste(brand1, productGroup1, categoryIDs1, sep = "|"))
d$item_id2 = with(d, paste(brand2, productGroup2, categoryIDs2, sep = "|"))
d$item_id3 = with(d, paste(brand3, productGroup3, categoryIDs3, sep = "|"))
```

0.1.1 Creating order_match class

```
d$order_match_class = factor("000", levels = c("111", "110", "101", "011", "000"))
d$order_match_class[(d$item_id1 != d$item_id2 & d$item_id2 == d$item_id3)] = "011"
d$order_match_class[(d$item_id1 != d$item_id2 & d$item_id1 == d$item_id3)] = "101"
d$order_match_class[(d$item_id1 == d$item_id2 & d$item_id1 != d$item_id3)] = "110"
d$order_match_class[(d$item_id1 == d$item_id2 & d$item_id1 == d$item_id3)] = "111"
```

0.1.2 Save the category feature

```
d %>% select(orderID, order_match_class) %>% saveRDS(file = "~/dmc2015/features/feature_files/universal/
```

0.2 Results

```
chk = d %>% mutate(order_res_paste = paste0(coupon1Used, coupon2Used, coupon3Used),
    nCouponUsed = coupon1Used + coupon2Used + coupon3Used) %>% select(order_res_paste,
    coupon1Used, coupon2Used, coupon3Used, nCouponUsed, order_match_class) %>%
    gather(tmp, couponUsed, -order_res_paste, -order_match_class, -nCouponUsed) %>%
    mutate(couponCol = as.numeric(gsub("coupon", "", gsub("Used", "", tmp)))) %>%
    select(-tmp) %>% filter(order_res_paste != "NANANA")

summary(chk$order_match_class)
```

```
##   111   110   101   011   000
##    69   225    69   420 17376
```

```
chk$couponMatch = 0
for (i in 1:nrow(chk)) chk$couponMatch[i] = 1 * (as.numeric(unlist(strsplit(as.character(chk$order_matc
    ""))[chk$couponCol[i]]) == chk$couponUsed[i])

ggplot(data = chk, aes(x = order_match_class, fill = couponMatch)) + geom_bar() +
    facet_grid(couponCol ~ couponUsed)
```