Data Mining Cup: Iowa State University Team

basePrice Over price

Spring 2015, Feature Matrix Creation

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 A few simple statistics

The ratio of price to basePrice

```
ratios = dmc %>% mutate(bPr2pr_ratio = basePrice/price, colname1 = paste0("bPr2pr_ratio",
    couponCol), bPr2pr_approx_ratio = round(bPr2pr_ratio, 1), colname2 = paste0("bPr2pr_approx_ratio",
    couponCol)) %>% select(orderID, colname1, bPr2pr_ratio, colname2, bPr2pr_approx_ratio) %>%
    arrange(orderID, colname1) %>% spread(colname1, bPr2pr_ratio) %>% arrange(orderID,
    colname2) %>% spread(colname2, bPr2pr_approx_ratio) %>% data.frame %>% saveRDS(file = "~/dmc2015/fe
```