finding in category

Jing Li

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")
price <- read.csv("~/../Desktop/DMC/raw_data/prices.csv",sep = "|")
train <- read.csv("~/../Desktop/DMC/raw_data/train.csv",sep="|")
items <- read.csv("~/../Desktop/DMC/raw_data/items.csv",sep='|')
```

```
#Cross tabulation for mainCategory and category
#Only category 37 belongs to both mainCategory 1 and 9, the other categories are unique for each mainCa
xtabs(~mainCategory+category,items)
```

```
##              category
## mainCategory    2    7   10   16   18   24   30   33   36   37
##            1  3705 5276    0    0    0    0    0    0    0  624
##            9     0    0  873    0  918    0    0    0  244   50
##            15    0    0    0  491    0  193   11  439    0    0
```

```
main1 <- subset(items,mainCategory==1)
main9 <- subset(items,mainCategory==9)
main15 <- subset(items,mainCategory==15)
#Cross tabulation for category and subCategory in mainCategory 1
crosstab(main1, col.vars = "category", row.vars = "subCategory", type = "f")
```

```
##              category    2    7   37  Sum
## subCategory
## 3                     2006    0  366 2372
## 4                      326    0    0  326
## 5                      465    0   94  559
## 6                      518    0  105  623
## 8                        0  725    0  725
## 11                       0   89    0   89
## 12                       0   88    0   88
## 13                       0  460    0  460
## 14                       0  533    0  533
## 16                       0  665    0  665
## 17                       0   96    0   96
## 20                       0  149    0  149
## 21                       0 1082    0 1082
## 22                       0  385    0  385
## 23                       0  173    0  173
## 25                       0  347    0  347
## 26                       0   38    0   38
## 27                       9    0    0    9
## 28                       0  124    0  124
## 29                       0   10    0   10
## 31                       0  252    0  252
## 34                       0    4    0    4
## 39                     375    0   59  434
## 40                       0   35    0   35
```

```
## 41                      0    7    0    7
## 43                      0   14    0   14
## 44                      6    0    0    6
## Sum                  3705 5276  624 9605
```

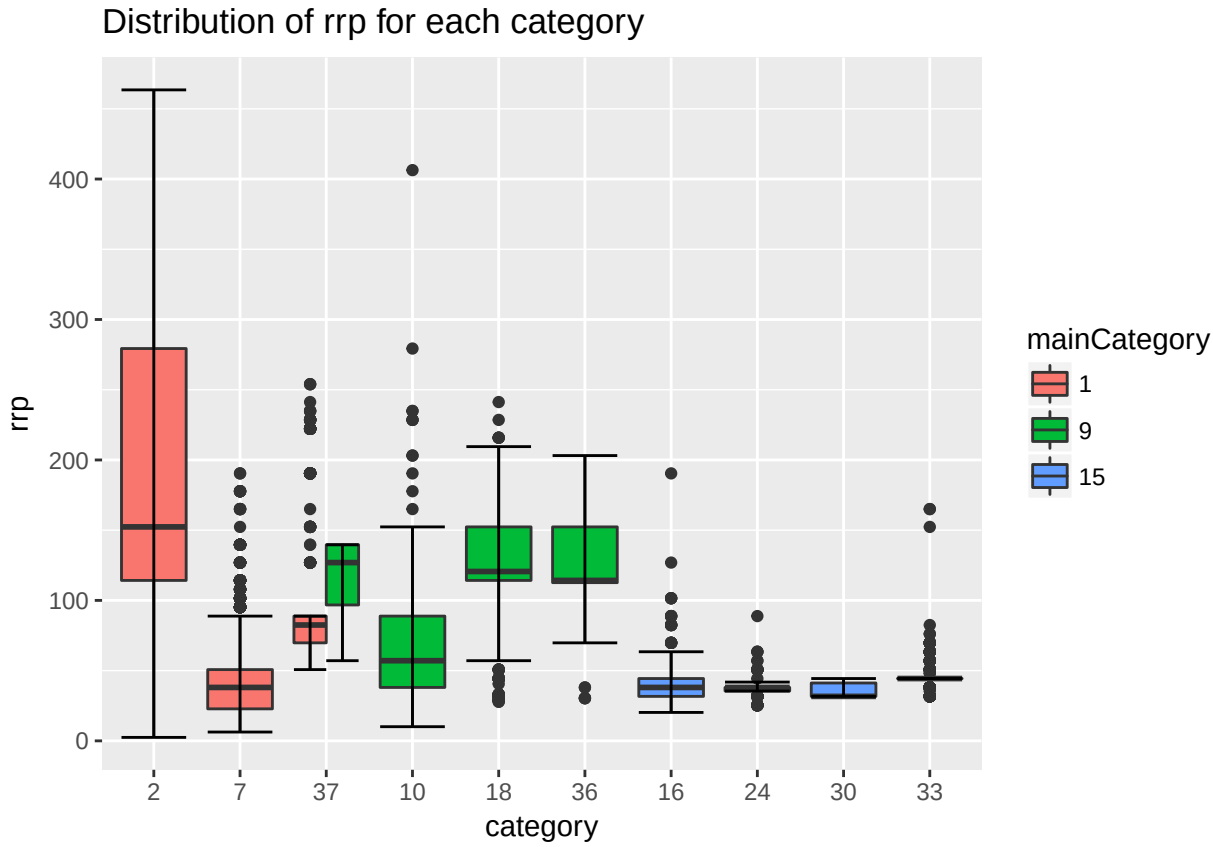
```
#Cross tabulation for category and subCategory in mainCategory 9
crosstab(main9, col.vars = "category", row.vars = "subCategory", type = "f")
```

```
##             category   10   18   36   37  Sum
## subCategory
## 11                    129    0    0    0  129
## 14                    201    0    0    0  201
## 19                      0    2    0    0    2
## 22                    129    0    0    0  129
## 23                      5    0    0    0    5
## 25                    163    0    0    0  163
## 29                     15    0    0    0   15
## 32                      0  874  240   50 1164
## 35                    210    0    0    0  210
## 38                      0   42    4    0   46
## 42                     21    0    0    0   21
## Sum                   873  918  244   50 2085
```

```
#mainCategory 15 have no subCategory
table(main15$category)
```

```
##
##  16  24  30  33
## 491 193  11 439
```

```
#Boxplot of rrp for each category
#mainCategory 1 havs large variance, especially for category 2.
#rrp for mainCategory 15 has the smallest varience, and is cheapest.
items$category <- factor(items$category, levels = c(2,7,37,10,18,36,16,24,30,33))
ggplot(items,aes(x=as.factor(category),y=rrp,fill=as.factor(mainCategory)))+
  geom_boxplot()+stat_boxplot(geom ='errorbar')+
  scale_fill_discrete(name="mainCategory")+
  labs(x="category",title="Distribution of rrp for each category")
```


Distribution of rrp for each category
rrp
0
100
200
300
400
2
7
37
10
18
36
16
24
30
33
category
mainCategory
1
9
15