dmc_1.R

zamirg13

Mon Apr 23 15:48:34 2018

```r
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
library(reshape2)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:reshape2':
##
##     dcast, melt
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```r
library(plyr)
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:lubridate':
##
##     here
```

```r
prices <- read.csv("prices.csv", sep = "|")
items <- read.csv("items.csv", sep = "|")
train <- read.csv("train.csv", sep = "|")

# separate dates (123 days, last date: 01/31/18)
train$year <- year(ymd(train$date))
train$month <- month(ymd(train$date))
train$day <- day(ymd(train$date))
train$weekday <- weekdays(ymd(train$date))

# add id for the unique combination of the "pid" and "size"
items$id <- as.factor(seq(1,length(items$pid)))

# combine new id with the train data_set
```

```r
train_items <- merge(train, items, by.y = c("pid","size"))

# translate to english
levels(items$color) <- c("beige", "blue", "brown", "yellow", "gold",
                         "gray", "green", "khaki", "purple", "orange",
                         "dark_pink", "pink", "red", "black", "silver",
                         "turquoise", "white")

# explore sizes(many contain the same information)
levels(train$size)
```

```
##   [1] ""                "0 ( 128 )"       "0 ( 31-33 )"
##   [4] "0 ( Bambini )"   "00 ( 27-30 )"    "01 Junior"
##   [7] "02 Senior"       "1 ( 140 )"       "1 ( 25-30 )"
##  [10] "1 ( 31-34 )"     "1 ( 33-36 )"     "1 ( 34-36 )"
##  [13] "1 ( Junior)"     "10"              "10 (140)"
##  [16] "10 (36-40)"      "10/12 (140-152)" "102 (M)"
##  [19] "104"             "11"              "116"
##  [22] "116-122"         "116/128"         "12 (41-45)"
##  [25] "128"             "134"             "14 (164)"
##  [28] "14 (46-48)"      "14/16 (164-176)" "140"
##  [31] "140/152"         "146"             "152"
##  [34] "158"             "16 (176)"        "164"
##  [37] "164/176"         "176"             "19 (38)"
##  [40] "2"               "2 ( 152 )"       "2 ( 31-34 )"
##  [43] "2 ( 35-38 )"     "2 ( 37-39 )"     "2 ( 37-40 )"
##  [46] "2 ( Senior )"    "24 (M)"          "28 (3XL)"
##  [49] "29"              "2XL"             "2XL/T"
##  [52] "3"               "3 ( 164 )"       "3 ( 39-42 )"
##  [55] "3 ( 40-42 )"     "3 ( 41-43 )"     "3 (35-38 )"
##  [58] "30"              "30 (5XL)"        "31"
##  [61] "31,5"            "32"              "33"
##  [64] "33,5"            "34"              "35"
##  [67] "35 - 38"         "35,5"            "35/38"
##  [70] "36"              "36 2/3"          "36,5"
##  [73] "37"              "37 - 40"         "37 1/3"
##  [76] "37,5"            "38"              "38 2/3"
##  [79] "38,5"            "38/40 ( M / L )" "39"
##  [82] "39 - 42"         "39 1/3"          "39-42"
##  [85] "39,5"            "39/42"           "3XL"
##  [88] "3XL/T"           "4"               "4 ( 39-42 )"
##  [91] "4 ( 43-45 )"     "4 ( 43-46 )"     "4 ( 44-46 )"
##  [94] "40"              "40 2/3"          "40,5"
##  [97] "41"              "41 - 44"         "41 1/3"
## [100] "41,5"            "42"              "42 2/3"
## [103] "42,5"            "43"              "43 - 46"
## [106] "43 1/3"          "43-46"           "43,5"
## [109] "43/46"           "44"              "44 2/3"
## [112] "44,5"            "45"              "45 - 47"
## [115] "45 1/3"          "45-48"           "45,5"
## [118] "46"              "46 2/3"          "46,5"
## [121] "47"              "47 - 50"         "47 1/3"
## [124] "47,5"            "47/49"           "48"
## [127] "48 2/3"          "48,5"            "4XL"
## [130] "5"               "5 ( 43-46 )"     "5 ( 46-48 )"
## [133] "5 ( 47-49 )"     "6"               "6 ( 47-50 )"
## [136] "6/8 (116-128)"   "7"               "7 ( L )"
## [139] "8"               "8 ( XL )"        "9"
## [142] "L"               "L ( 152-158 )"   "L ( 40/42 )"
## [145] "L ( 42-46 )"     "L ( 42-47 )"     "L ( 44 )"
## [148] "L (43 - 46)"     "L/K"             "L/T"
## [151] "L/XL ( 39-47 )"  "M"               "M ( 140-152 )"
## [154] "M ( 38-42 )"     "M ( 38/40 )"     "M ( 40 )"
## [157] "M (38 - 42)"     "M/L"             "S"
## [160] "S ( 128-140 )"   "S ( 34-38 )"     "S ( 34/36 )"
## [163] "S ( 36 )"        "XL"              "XL ( 158-170 )"
## [166] "XL ( 44/46 )"    "XL (46-48,5)"    "XL (46-50 )"
## [169] "XL/T"            "XS"              "XS ( 116-128 )"
## [172] "XS ( 30-34 )"    "XS ( 32 )"       "XS ( 32/34 )"
## [175] "XS/S"            "YLG 147,5-157,5" "YM 135-147,5"
## [178] "YSM 125-135"     "YXL 157,5-167,5"
```

```r
# sum of sold items by id
sold_by_id <- ddply(train_items, "id", summarise, sum = sum(units))
ord <- sold_by_id[order(sold_by_id$sum, decreasing = TRUE),]
head(ord, 20)
```

```
##          id  sum
## 3023   3023 2979
## 5886   5886 2643
## 5885   5885 2411
## 6865   6865 1819
## 8189   8189 1694
## 3034   3034 1562
## 9306   9306 1439
## 8188   8188 1427
## 8508   8508 1388
## 426     426 1358
## 2954   2954 1319
## 7243   7243 1289
## 9305   9305 1280
## 4121   4121 1259
## 8509   8509 1237
## 427     427 1224
## 9129   9129 1146
## 7242   7242 1113
## 12041 12041 1044
## 3060   3060 1012
```

```r
# There are 2263 items that were sold only one times
sum(as.numeric(sold_by_id$sum == 1))
```

```
## [1] 2263
```

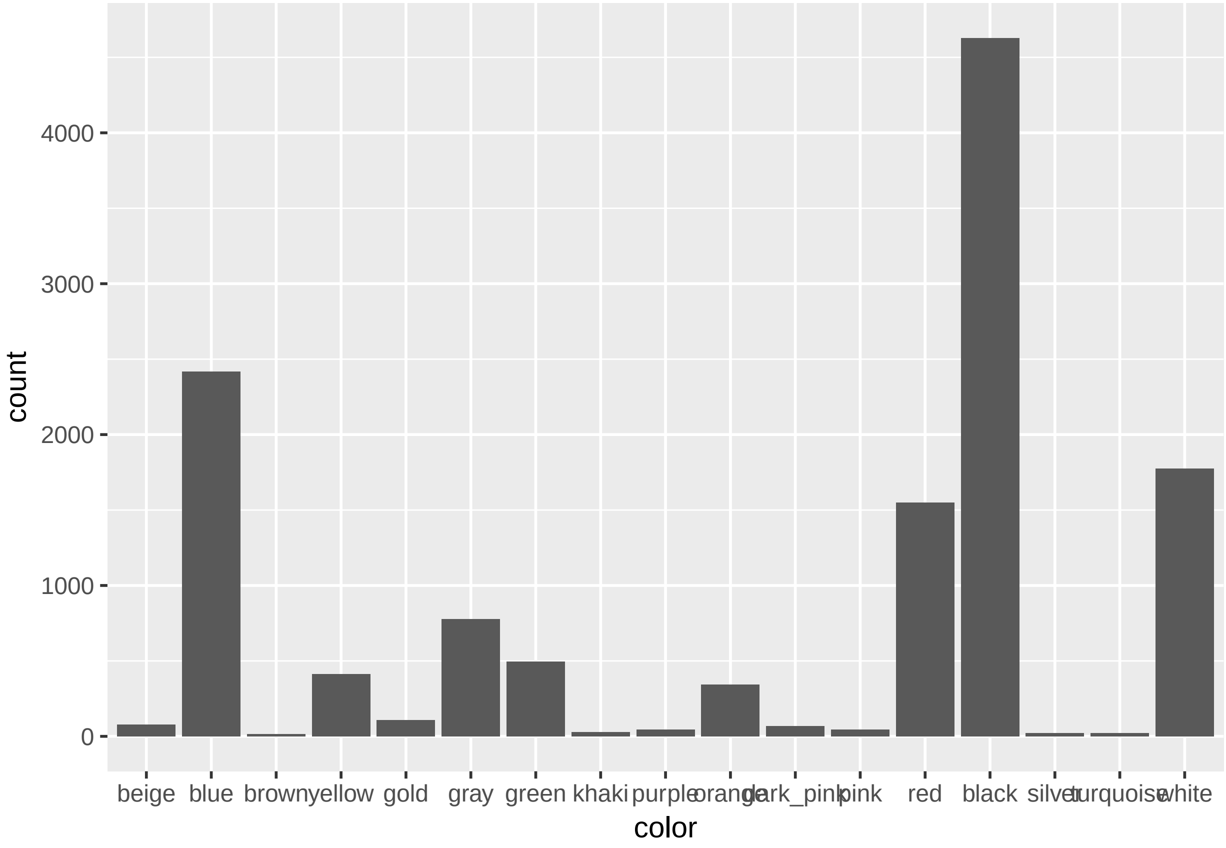
```r
# which items were sold only one times?
ids <- which(sold_by_id$sum == 1) # id is consistent with the raw number
sum(as.numeric(train_items[ids,]$stock != 0)) # all have non-zero stocks
```

```
## [1] 2263
```

```
table(train_items[ids,]$month) # rare sold products were sold in average
```

```
##
##   1  10  11  12
## 608 507 620 528
```

```
# equal amount each month. But the whole other stock: 2263 items were
# sold in 28 days on February. So there is effect of discounts(probably)
# on the sale of these.

# chosen: Color
color <- data.frame(table(items$color))
colnames(color) <- c("color", "frequency")


# 17 colors in total
# there are 4 major colors: black, blue, white and red
# 4 submajor: grey, green, gold and orange
ggplot(items, aes(color)) + geom_bar()
```



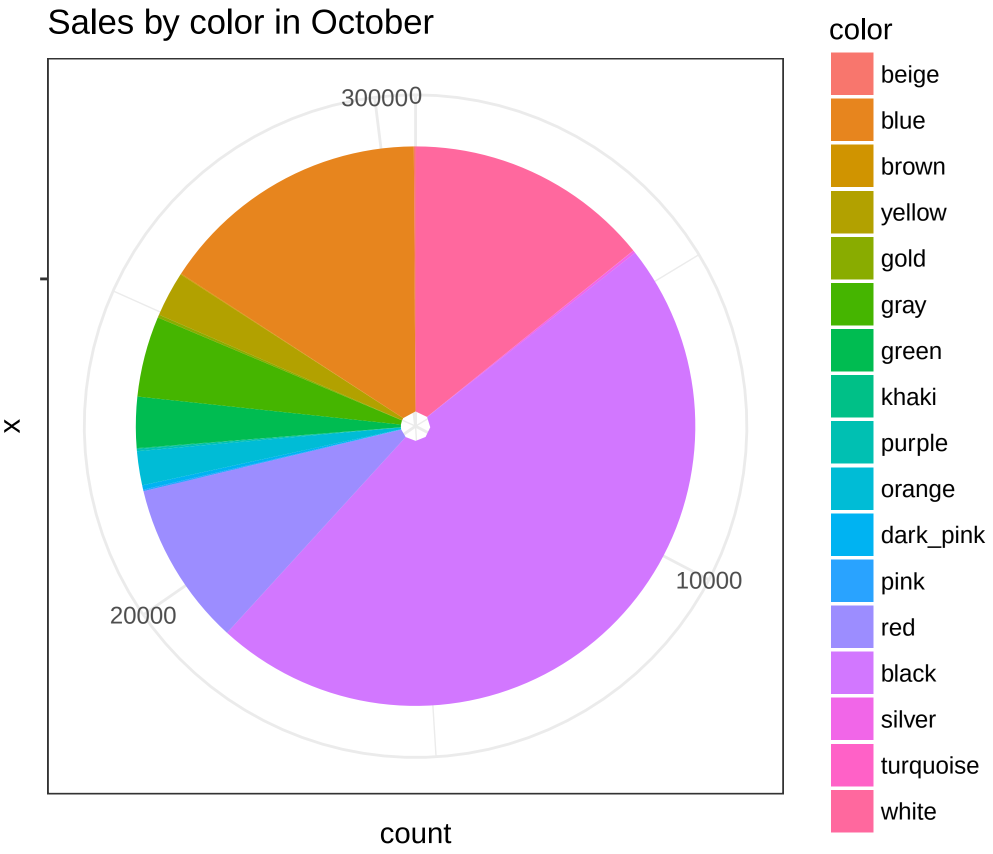
```
color[order(color$frequency, decreasing = TRUE),]
```

```
##        color frequency
## 14     black      4629
## 2       blue      2418
## 17     white      1775
## 13       red      1550
## 6       gray       777
## 7      green       494
## 4     yellow       411
## 10    orange       343
## 5       gold       107
## 1      beige        77
## 11 dark_pink        68
## 12      pink        45
## 9     purple        44
## 8      khaki        29
## 15    silver        22
## 16 turquoise        20
## 3      brown        15
```

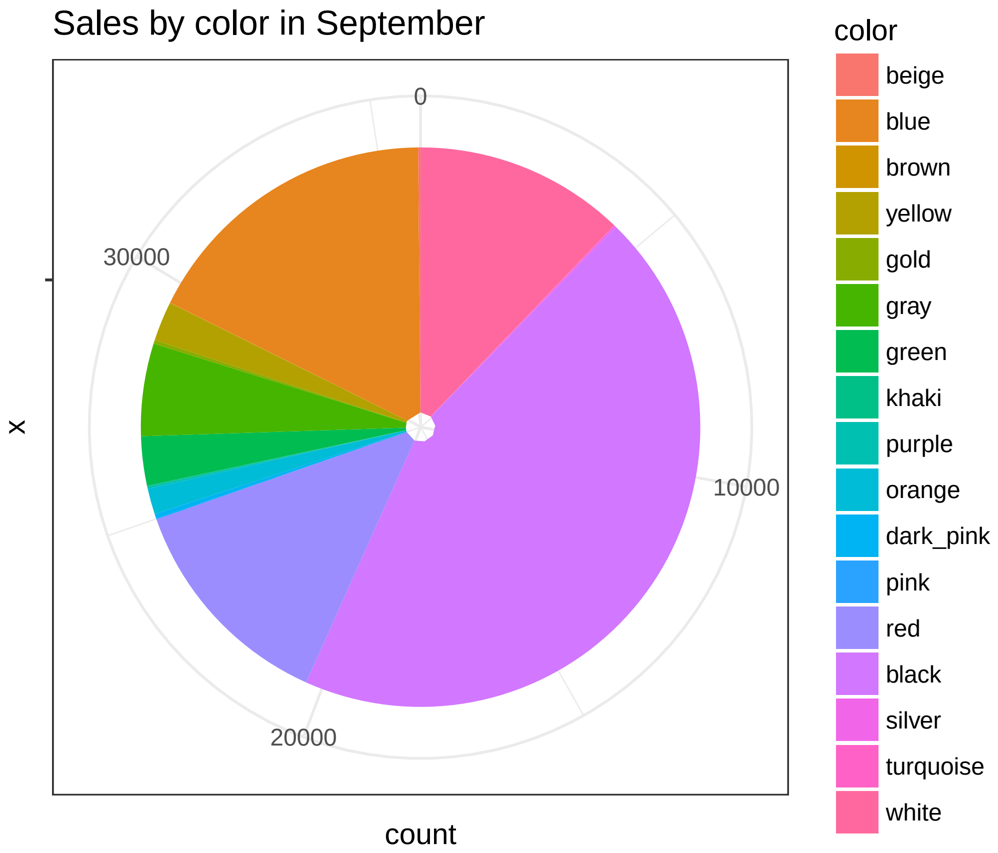
```
# merge datasets:
detailed_train <- merge(items, train, by = c("pid", "size"))

# extract_per_month <- function(m) {
#    m10 <- detailed_train[detailed_train$month == m,]
#    return(data.frame(sold_oct = tapply(m10$units, m10$color, sum)))
# }

m10 <- detailed_train[detailed_train$month == 10,]
s10 <- data.frame(sales_oct = tapply(m10$units, m10$color, sum))
m11 <- detailed_train[detailed_train$month == 11,]
s11 <- data.frame(sales_nov = tapply(m11$units, m11$color, sum))
m12 <- detailed_train[detailed_train$month == 12,]
s12 <- data.frame(sales_dec = tapply(m12$units, m12$color, sum))
m01 <- detailed_train[detailed_train$month == 01,]
s01 <- data.frame(sales_jan = tapply(m01$units, m01$color, sum))
sales_by_col <- cbind(s10, s11, s12, s01)


# relationship with sales
# sold units by color per month
sales_by_col[order(sales_by_col$sales_oct, decreasing = TRUE),]
```

```
##           sales_oct sales_nov sales_dec sales_jan
## black         34582     49324     35489     39289
## blue           8765     12311     12544     13081
## white          8653      9492      8456     10160
## red            6005      9295      9337     10696
## gray           2473      4387      4126      3731
## green          1816      1734      1590      2296
## yellow         1564      1293      1089      1845
## orange          835       696       603       669
## dark_pink       121       144       148       225
## gold            114       125       300       260
## silver           56        47        38        23
## purple           53        54        56        82
## beige            32        54        62       193
## pink             26        18        22        42
## brown            19        15         9         3
## khaki            17        22        11        16
## turquoise         4        10         9        15
```

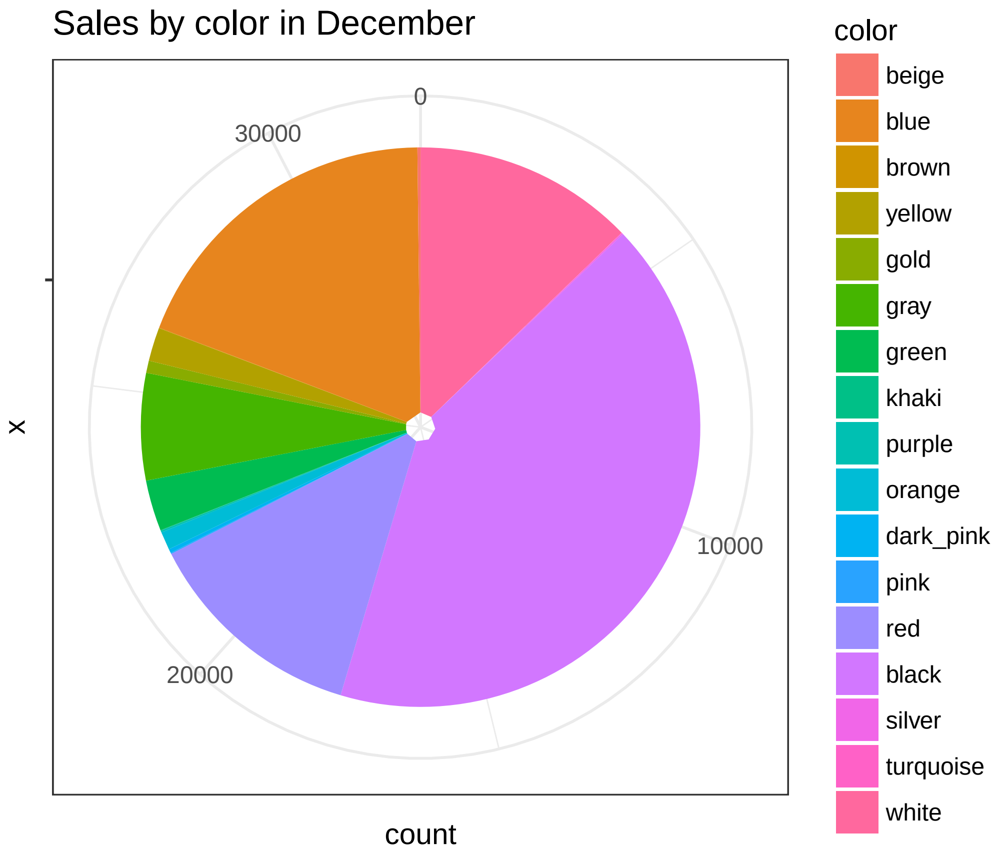
```r
# Items sold by color in different months
# boxplots did not work because of the small values and outliers
# ggplot(m10, aes(x = color, y = units)) + geom_boxplot()
ggplot(m10, aes(x = "", fill = color)) + geom_bar() + coord_polar("y") +
    theme_bw() + ggtitle("Sales by color in October")
```



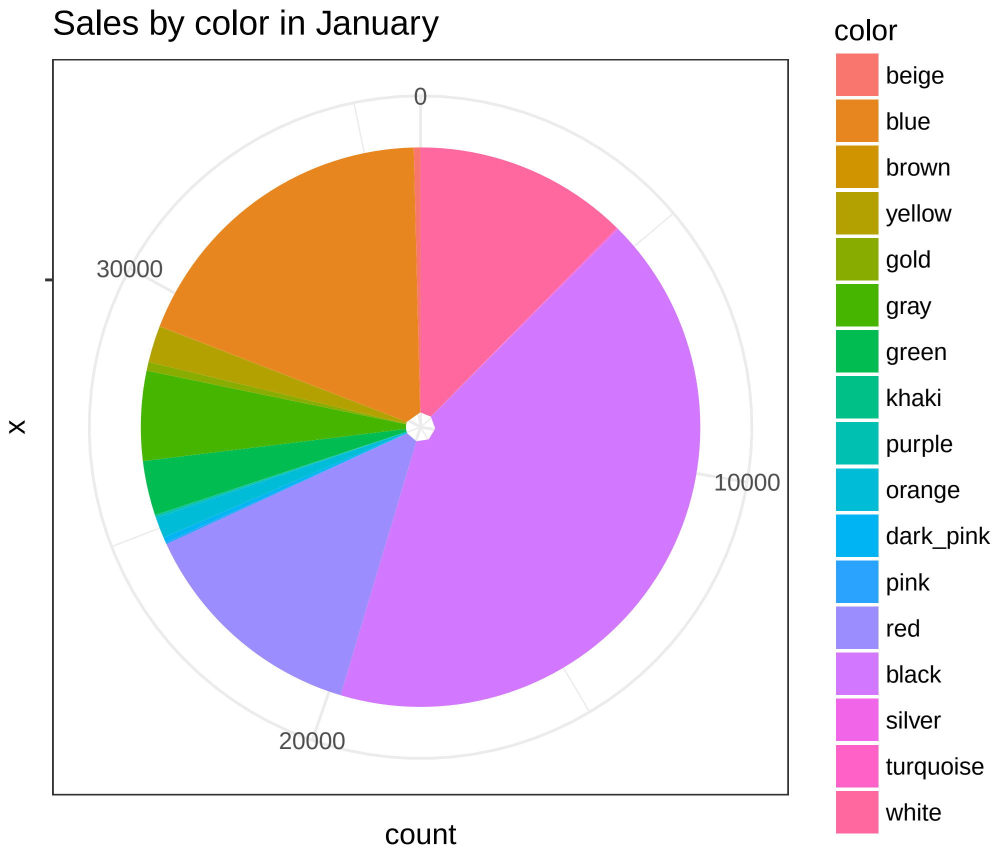
```r
ggplot(m11, aes(x = "", fill = color)) + geom_bar() + coord_polar("y") +
    theme_bw() + ggtitle("Sales by color in September")
```



```
ggplot(m12, aes(x = "", fill = color)) + geom_bar() + coord_polar("y") +
    theme_bw() + ggtitle("Sales by color in December")
```



```
ggplot(m01, aes(x = "", fill = color)) + geom_bar() + coord_polar("y") +
    theme_bw() + ggtitle("Sales by color in January")
```



```r
# relation with the other categorical variables:
table(items$color, items$brand)
```

```
##
##              adidas Asics Cinquestelle Converse Diadora Erima FREAM Hummel
##   beige          53     0            0        0       0     0     0      0
##   blue          684     2            0        6       0    25     0     15
##   brown           5     0            0        0       0     0     0      0
##   yellow        161     0            0        0       0     4     0      0
##   gold           82     0            0        0       0     0     0      0
##   gray          137     1            0       16       3     2     0      2
##   green         149     0            0        1       0     2     0      1
##   khaki           4     0            0        1       0     0     0      0
##   purple          4     0            0        0       0     0     0      0
##   orange         37     0            0        0       0     0     0      0
##   dark_pink      12     0            0        0       0     0     0      1
##   pink           12     0            0        4       0     0     0      0
##   red           433     0            0        4       0     3     0     13
##   black        1623     2            6       58       8    55     2     49
##   silver          2     0            0        2       0     0     0      0
##   turquoise       0     0            0        0       0     0     0      5
##   white         571     3            0       31       2    16     0      9
##
##              Jako Jordan KangaROOS Kempa Lotto Mizuno New Balance Nike
##   beige         0      3         0     0     0      0           0   19
##   blue        113      6         0     0     3      7          11 1369
##   brown         0      0         0     0     0      0           0   10
##   yellow      21      0         0     0     0      0           0  163
##   gold         0      0         0     0     0      0           0   23
##   gray        34     15         0     0     0      0           9  507
##   green       35      0         0     0     0      0           7  254
##   khaki        0      0         1     0     0      0           0   22
##   purple       6      0         0     0     0      0           0   22
##   orange       6      2         0     0     0      2           7  270
##   dark_pink    1      0         0     0     0      0           0   47
##   pink         0      0         0     0     0      0           2   22
##   red         64      3         0     0     1      0           2  905
##   black      318     70         2     1     4      2          19 1936
##   silver       0      0         0     0     0      0           1    7
##   turquoise    0      0         0     0     0      0           0   13
##   white       75     37         0     0     2      1           7  800
##
##            Onitsuka PUMA Reebok Reusch Sells Sport2000 Stance Uhlsport
##   beige            0    2      0      0     0         0      0        0
##   blue             0  148      2      0     0         0      0       15
##   brown            0    0      0      0     0         0      0        0
##   yellow           0   55      0      0     0         0      0        3
##   gold             0    2      0      0     0         0      0        0
##   gray             0   25      6      0     0         5      0        5
##   green            0   33      0      0     0         0      0        6
##   khaki            0    1      0      0     0         0      0        0
##   purple           0   12      0      0     0         0      0        0
##   orange           0   16      1      0     0         0      0        2
##   dark_pink        0    1      0      0     0         4      0        1
##   pink             0    2      3      0     0         0      0        0
##   red              0   93      0      0     0         0      0       12
##   black            0  275     24      7     1        37      6       73
##   silver           0    7      0      0     0         0      0        1
##   turquoise        1    0      1      0     0         0      0        0
##   white            0  100     67      0     0        12      2        8
##
##            Under Armour
##   beige               0
##   blue               12
##   brown               0
##   yellow              4
##   gold                0
##   gray               10
##   green               6
##   khaki               0
##   purple              0
##   orange              0
##   dark_pink           1
##   pink                0
##   red                17
##   black              51
##   silver              2
##   turquoise           0
##   white              32
## some brands have only one of the 4 major colors in stock
## (for unique product). Adidas has the largest number of color variations.
## Nike is second, then PUMA and Jako.
```