dmc_1.R

zamirg13

Thu Apr 12 21:34:22 2018

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

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

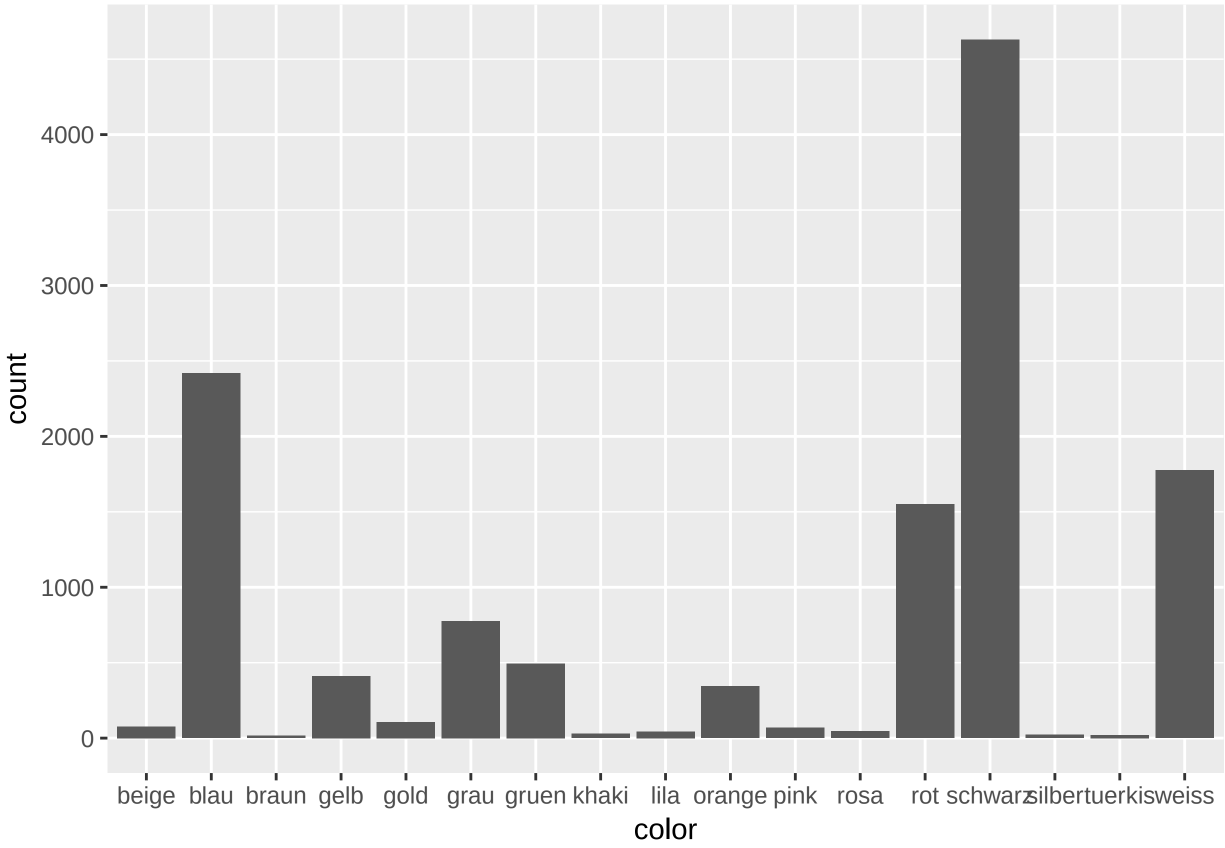
```r
library(lubridate)
```

```
##
## 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
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))



# 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()
```



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

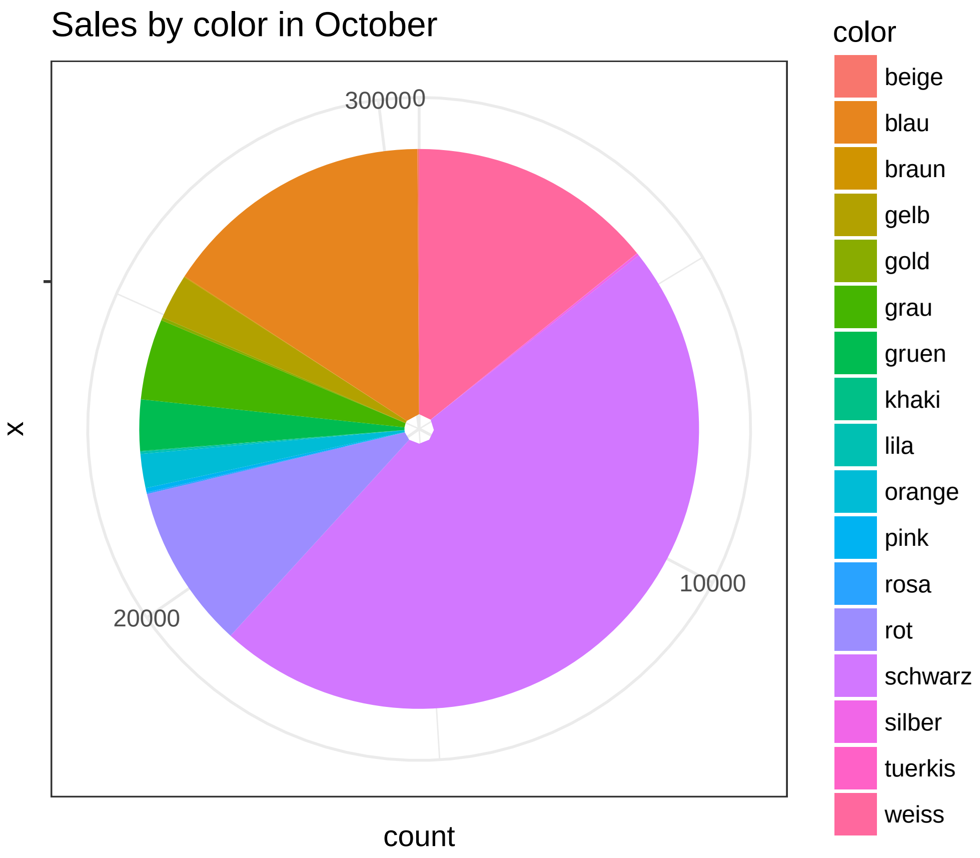
```
##      color frequency
## 14 schwarz      4629
## 2     blau      2418
## 17   weiss      1775
## 13     rot      1550
## 6     grau       777
## 7    gruen       494
## 4     gelb       411
## 10  orange       343
## 5     gold       107
## 1    beige        77
## 11    pink        68
## 12    rosa        45
## 9     lila        44
## 8    khaki        29
## 15  silber        22
## 16 tuerkis        20
## 3    braun        15
```

```r
# 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)))
# }
```

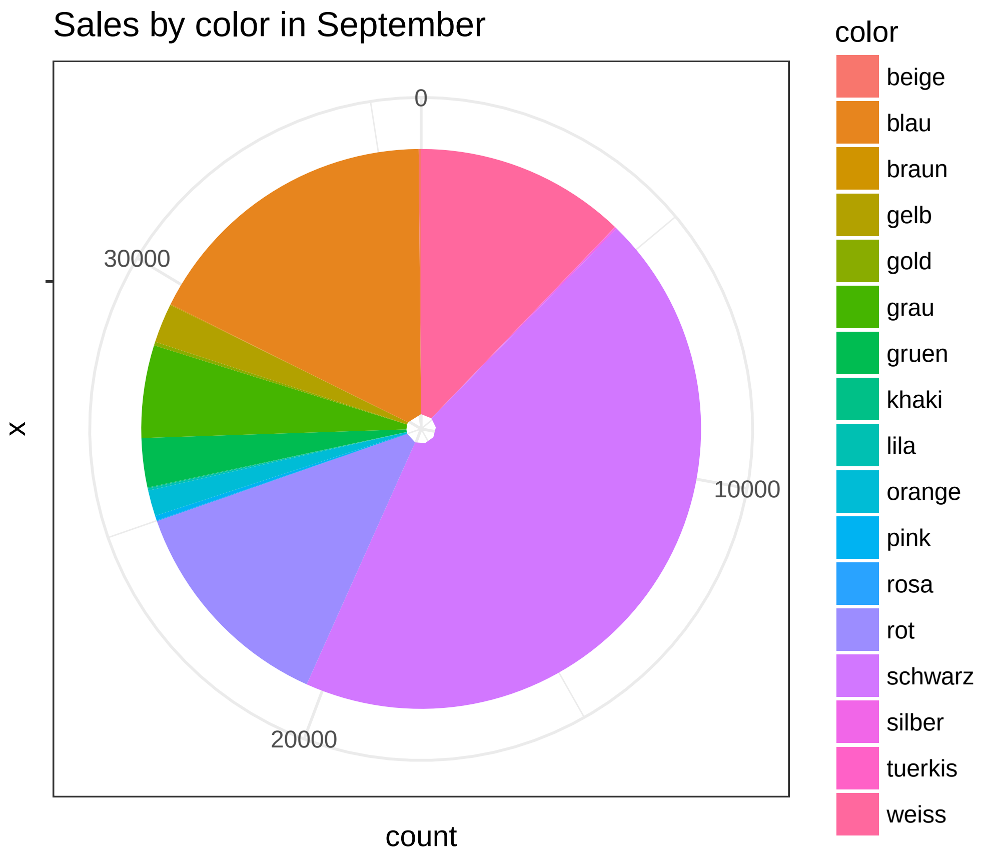
```r
m10 <- detailed_train[detailed_train$month == 10,]
s10 <- data.frame(sales_oct = tapply(m10$units, m10$color, sum))
m11 <- detailed_train[detailed_train$month == 11,]
s11 <- data.frame(sales_nov = tapply(m11$units, m11$color, sum))
m12 <- detailed_train[detailed_train$month == 12,]
s12 <- data.frame(sales_dec = tapply(m12$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
## schwarz     34582     49324     35489     39289
## blau         8765     12311     12544     13081
## weiss        8653      9492      8456     10160
## rot          6005      9295      9337     10696
## grau         2473      4387      4126      3731
## gruen        1816      1734      1590      2296
## gelb         1564      1293      1089      1845
## orange        835       696       603       669
## pink          121       144       148       225
## gold          114       125       300       260
## silber         56        47        38        23
## lila           53        54        56        82
## beige          32        54        62       193
## rosa           26        18        22        42
## braun          19        15         9         3
## khaki          17        22        11        16
## tuerkis         4        10         9        15
```

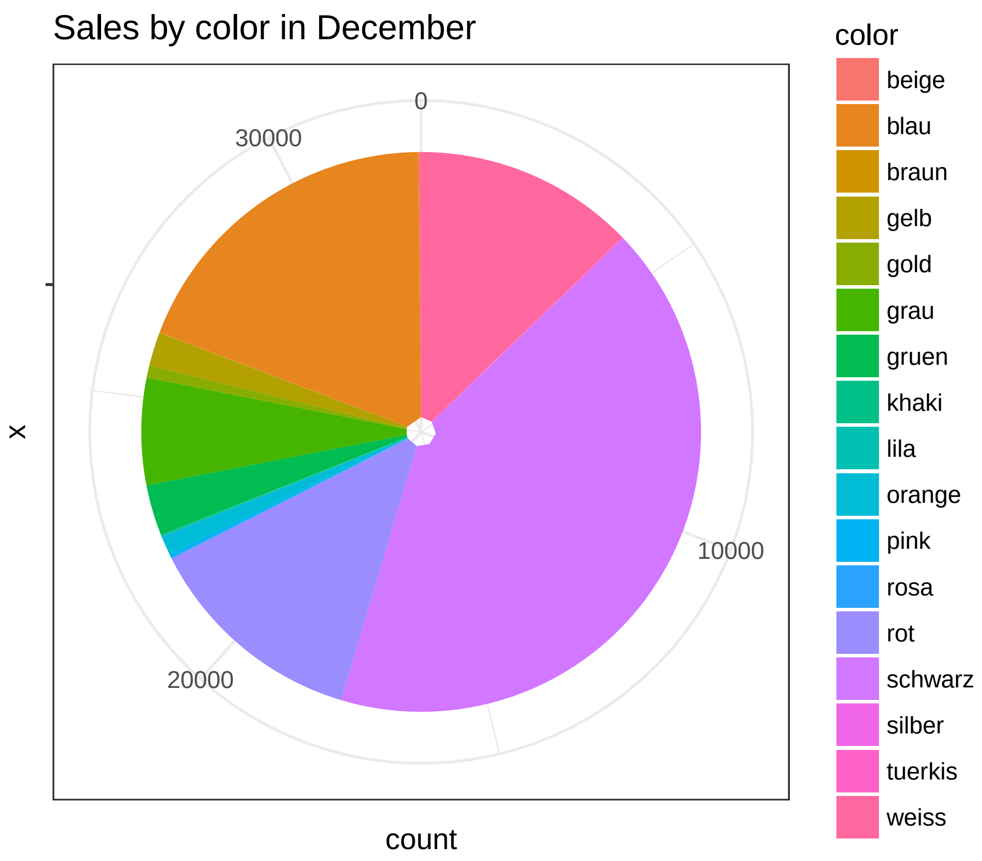
```r
# Items sold by color in different months
# boxplots did not work because of the small values and outliers
# ggplot(m10, aes(x = color, y = units)) + geom_boxplot()
ggplot(m10, aes(x = "", fill = color)) + geom_bar() + coord_polar("y") +
    theme_bw() + ggtitle("Sales by color in October")
```



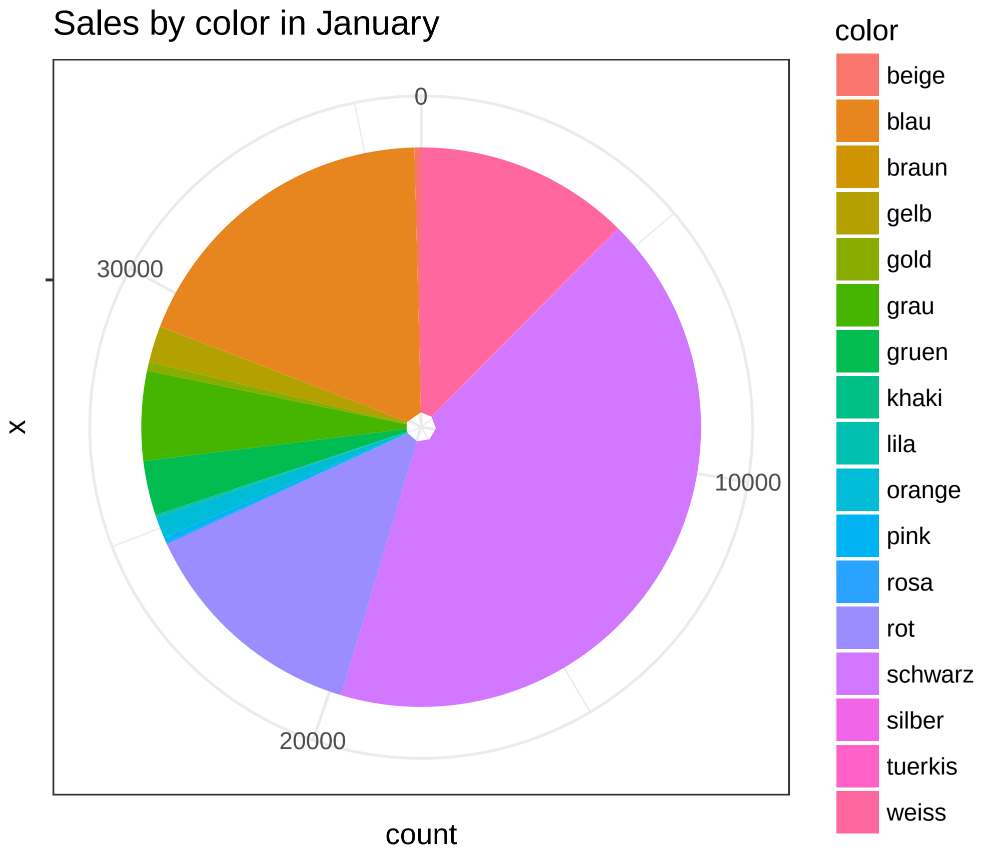
```
ggplot(m11, aes(x = "", fill = color)) + geom_bar() + coord_polar("y") +
    theme_bw() + ggtitle("Sales by color in September")
```



```
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
##   blau        684     2            0        6       0    25     0     15
##   braun         5     0            0        0       0     0     0      0
##   gelb        161     0            0        0       0     4     0      0
##   gold         82     0            0        0       0     0     0      0
##   grau        137     1            0       16       3     2     0      2
##   gruen       149     0            0        1       0     2     0      1
##   khaki         4     0            0        1       0     0     0      0
##   lila          4     0            0        0       0     0     0      0
##   orange       37     0            0        0       0     0     0      0
##   pink         12     0            0        0       0     0     0      1
##   rosa         12     0            0        4       0     0     0      0
##   rot         433     0            0        4       0     3     0     13
##   schwarz    1623     2            6       58       8    55     2     49
##   silber        2     0            0        2       0     0     0      0
##   tuerkis       0     0            0        0       0     0     0      5
##   weiss       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
##   blau      113      6         0     0     3      7          11 1369
##   braun       0      0         0     0     0      0           0   10
##   gelb      21      0         0     0     0      0           0   163
##   gold       0      0         0     0     0      0           0    23
##   grau      34     15         0     0     0      0           9   507
##   gruen     35      0         0     0     0      0           7   254
##   khaki      0      0         1     0     0      0           0    22
##   lila       6      0         0     0     0      0           0    22
##   orange     6      2         0     0     0      2           7   270
##   pink       1      0         0     0     0      0           0    47
##   rosa       0      0         0     0     0      0           2    22
##   rot       64      3         0     0     1      0           2   905
##   schwarz  318     70         2     1     4      2          19  1936
##   silber     0      0         0     0     0      0           1     7
##   tuerkis    0      0         0     0     0      0           0    13
##   weiss     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
##   blau           0  148      2      0     0         0      0       15
##   braun          0    0      0      0     0         0      0        0
##   gelb           0   55      0      0     0         0      0        3
##   gold           0    2      0      0     0         0      0        0
##   grau           0   25      6      0     0         5      0        5
##   gruen          0   33      0      0     0         0      0        6
##   khaki          0    1      0      0     0         0      0        0
##   lila           0   12      0      0     0         0      0        0
##   orange         0   16      1      0     0         0      0        2
##   pink           0    1      0      0     0         4      0        1
##   rosa           0    2      3      0     0         0      0        0
##   rot            0   93      0      0     0         0      0       12
##   schwarz        0  275     24      7     1        37      6       73
##   silber         0    7      0      0     0         0      0        1
##   tuerkis        1    0      1      0     0         0      0        0
##   weiss          0  100     67      0     0        12      2        8
##
##           Under Armour
##   beige              0
##   blau              12
##   braun              0
##   gelb               4
##   gold               0
##   grau              10
##   gruen              6
##   khaki              0
##   lila               0
##   orange             0
##   pink               1
##   rosa               0
##   rot               17
##   schwarz           51
##   silber             2
##   tuerkis            0
##   weiss             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.
```