diamonds exploration

Adam Bozman

September 28, 2016

Reading Data

```
mydiamonds=read.csv("diamonds.csv")
```

Dimensions and names of variables

```
dim(mydiamonds)
```

```
## [1] 53940    11
```

```
nrow(mydiamonds)
```

```
## [1] 53940
```

```
ncol(mydiamonds)
```

```
## [1] 11
```

There are 53940 rows in our dataset.

```
colnames(mydiamonds)
```

```
##  [1] "X"       "carat"   "cut"     "color"   "clarity" "depth"   "table"
##  [8] "price"   "x"       "y"       "z"
```

```
mydiamonds=mydiamonds[,2:11]
```

After knocking off the first colum, I am now left with 10 columns.

Saving my truncated file

```
save(mydiamonds,file="mydiamonds.Rda")
```

data structure

```r
str(mydiamonds)
```

```
## 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Factor w/ 5 levels "Fair","Good",..: 3 4 2 4 2 5 5 5 1 5 ...
##  $ color  : Factor w/ 7 levels "D","E","F","G",..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Factor w/ 8 levels "I1","IF","SI1",..: 4 3 5 6 4 8 7 3 6 5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Summary

```r
summary(mydiamonds[,c("carat","cut")])
```

```
##      carat               cut
##  Min.   :0.2000   Fair     : 1610
##  1st Qu.:0.4000   Good     : 4906
##  Median :0.7000   Ideal    :21551
##  Mean   :0.7979   Premium  :13791
##  3rd Qu.:1.0400   Very Good:12082
##  Max.   :5.0100
```

number of levels of a factor variable

```r
nlevels(mydiamonds$clarity)
```

```
## [1] 8
```

```r
levels(mydiamonds$clarity)
```

```
## [1] "I1"   "IF"   "SI1"  "SI2"  "VS1"  "VS2"  "VVS1" "VVS2"
```

```r
fairdiamonds=mydiamonds[mydiamonds$cut=="Fair",]
```

```r
levels(fairdiamonds$cut)
```

```
## [1] "Fair"      "Good"      "Ideal"     "Premium"   "Very Good"
```

```r
summary(fairdiamonds$cut)
```

```
##      Fair      Good     Ideal   Premium Very Good
##      1610         0         0         0         0
```

refactoring after a subset using a factor variable

```
fairdiamonds$cut=factor(fairdiamonds$cut)
summary(fairdiamonds$cut)
```

```
## Fair
## 1610
```