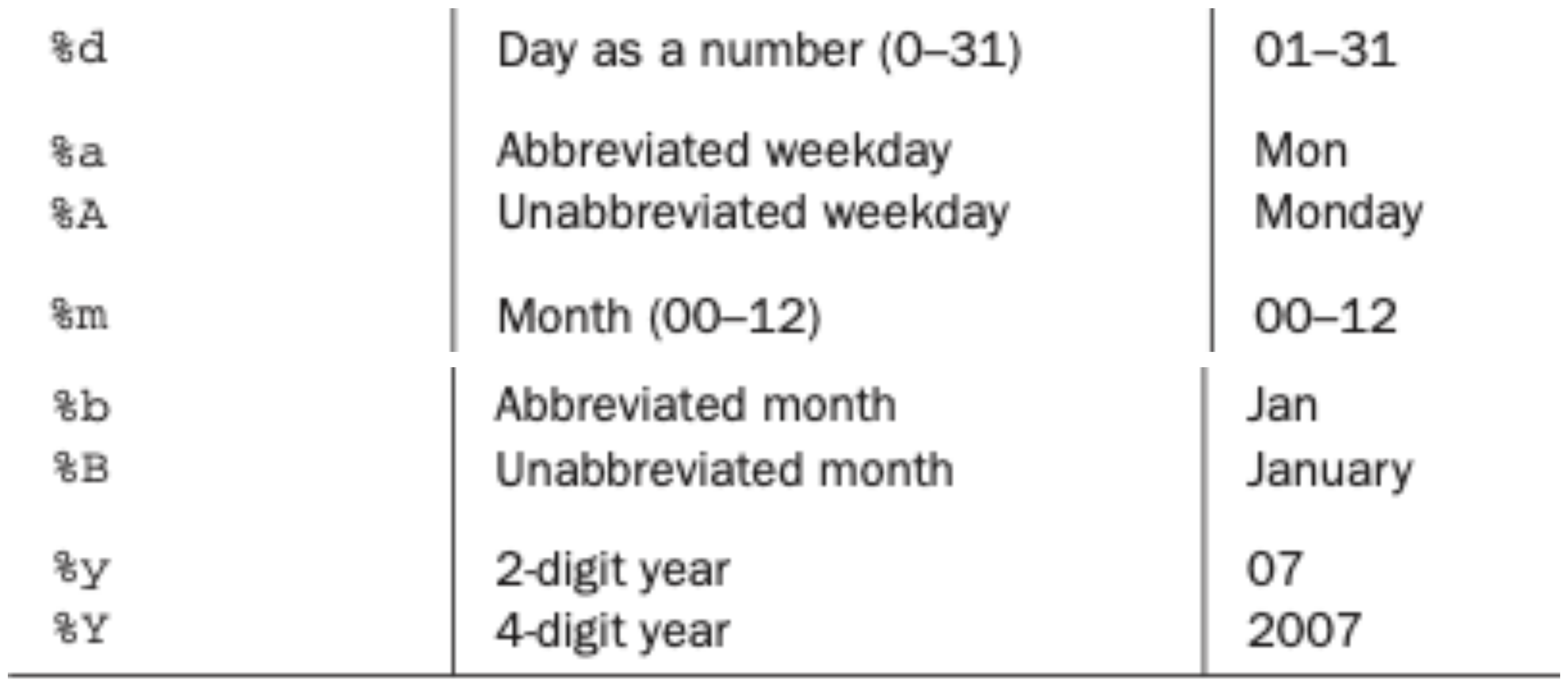
Some useful Data transformation functions

Date functions

- as.Date(x, format)
- tab2 <- read.table("house_copy.txt",header=TRUE,colClasses = c('integer','double','factor','factor','character'))
- tab2$last.sale.date <- as.Date(tab2$last.sale.date,"%d/%m/%Y")
- tab2$last.sale.date[15]-tab2$last.sale.date[13]



%d	Day as a number (0–31)	01–31
%a	Abbreviated weekday	Mon
%A	Unabbreviated weekday	Monday
%m	Month (00–12)	00–12
%b	Abbreviated month	Jan
%B	Unabbreviated month	January
%y	2-digit year	07
%Y	4-digit year	2007

Character Functions

- nchar(x) : Counts the number of characters of *x*
  - x <- c("ab", "cde", "fghij"); nchar(x[3]) returns 5
- substr(x, start, stop) : Extract or replace substrings in a character vector
  - x <- "abcdef"; substr(x, 2, 4) returns “bcd”.
  - substr(x, 2, 4) <- "22222" (x is now "a222ef")
- grep(pattern, x, ignore. case=FALSE, fixed=FALSE) : Search for pattern in `x`. If `fixed=FALSE`, then *`pattern`* is a regular expression. If `fixed=TRUE`, then *`pattern`* is a text string. Returns matching indices
  - grep("A", c("b","A","c"), fixed=TRUE) returns 2
- strsplit(x, split, fixed=FALSE) : Split the elements of character vector x at split. If fixed=FALSE, then pattern is a regular expression. If fixed=TRUE, then pattern is a text string
  - strsplit(c("abc",’cbc','dabccdbde'),'b',TRUE) returns
  - [[1]]
  - [1] "a" "c"
  - [[2]]
  - [1] "c" "c"
  - [[3]]
  - [1] "da" "ccd" "de"

Character Functions

- y <- strsplit(c("abc",’cbc','dabccdbde'),'b',TRUE)
- sapply(y,"[",2) return the character vector "c"   "c"   "ccd" (NOTE: "[" is an extraction operator and extracts by index number)
- paste and paste0 functions already covered
- toupper(x) : returns uppercase (similarly tolower)
  - toupper("abc") returns “ABC”

Convert numeric to factor

- cut(x, n) : Divide continuous variable x into factor with n levels.
  - tab2_breaks <- with(tab2, seq(min(area),max(area),(max(area)-min(area))/10))
  - with(tab2, cut(area,tab2_breaks,labels=LETTERS[1:10],include.lowest=TRUE))

Data Summarization - descriptives

- summary
- mean
- median
- quantile - with(tab2, quantile(area,seq(0,1,0.2)))
- sd
- variance
- cor
- table / prop.table / xtabs
  - with(tab2, table(availability,region))
  - xtabs(~availability+region,tab2)
  - xt <- xtabs(~availability+region,tab2); prop.table(xt,1)

which functions

- which(tab2$area>1000)
- which.max(tab2$area)