```r
#=============================================================================
#                    Using R: Manipulating data in data frames
#=============================================================================
#(a) Load the data frame baseball in the plyr package. Use ?baseball to get
information about the data set and definitions for the variables.
library(plyr)
data(baseball)
help("baseball")
```

```
## starting httpd help server ... done
```

```r
#=============================================================================
#(b) You will calculate the on base percentage for each player, but first clean up the
data:
#* Before 1954, sacrifice flies were counted as part of sacrifice hits, so for players
before 1954, sacrifice flies (i.e. the variable sf) should be set to 0.
baseball[baseball$year<1954,"sf"]=0

#* Hit by pitch (the variable hbp) is often missing - set these missings to 0.
baseball[is.na(baseball$hbp),"hbp"]=0

#* Exclude all player records with fewer than 50 at bats (the variable ab).
baseball=baseball[baseball$ab>50,]

#=============================================================================
#(c) Compute on base percentage in the variable obp according to the formula:
#                obp =(h + bb + hbp)/(ab + bb + hbp + sf)
obp=(baseball$h+baseball$bb+baseball$hbp)/(baseball$ab+baseball$bb+baseball$hbp+baseba
ll$sf)

#=============================================================================
#(d) Sort the data based on the computed obp and print the year, player name, and on
base percentage for the top five records based on this value.
baseball <- data.frame(baseball, obp)
baseballsorted <- baseball[order(-obp),]
View(baseballsorted[1:5,c("year","id","obp")])
```