Least Squares Rating System

```
library(data.table)
options(width = 90)
```

Loading the data with American College Football matches from 1994

The match data can be downloaded as a *csv* file from the site sports-reference.com, this page gives the match schedule, while this page gives the final rankings by the poll of sports journalists (AP) and by a simple sports rating system (SRS).

Before displaying the loaded data table, we remove some unnecessary columns.

```
d94 = fread("../../data/1994-College-Football.csv")
d94[, Rk := NULL]
```

Cleaning the column V7 Add a column `HA` for home advantage. The rows in `V7` with the sign "@" indicate that the second team is the host, while those that are empty indicate that the first team is the host.

```
d94[, HA := ifelse(V7 == "@", -1, 1)]
```

However, this rule breaks for the bowl games. The bowl games can be guessed from that they have a none empty `Notes` column. If the second team is a host, then `V7` contains the sign "@", but otherwise `V7` is empty. Most bowl games are played on a neutral ground, but there are a few exceptions. The `Notes` says where the match has been played.

```
d94[Wk >= 16, HA := ifelse(V7 == "@", -1, 0)]
d94[617, HA := 0]
```

Mark individual bowl games where the the winner has played at home (Las Vegas, Texas)

```
d94[c(618,625), HA := 1]
```

Deleting the `Notes` and `V7` columns.

```
d94[, Notes := NULL]
d94[, V7 := NULL]
d94
```

```
##      Wk         Date Day                   Winner Pts              Loser Pts HA
##   1:  1 Aug 28 1994 Sun            (4) Nebraska  31  (24) West Virginia   0  1
##   2:  1 Aug 29 1994 Mon          (20) Ohio State  34        Fresno State  10  1
##   3:  2  Sep 1 1994 Thu             (7) Arizona  19        Georgia Tech  14 -1
##   4:  2  Sep 1 1994 Thu                 Kansas  35             Houston  13 -1
##   5:  2  Sep 1 1994 Thu    North Carolina State  20 Bowling Green State  15  1
##  ---
## 632: 19  Jan 2 1995 Mon        (7) Florida State  23         (5) Florida  17  0
## 633: 19  Jan 2 1995 Mon           (2) Penn State  38         (12) Oregon  20  0
## 634: 19  Jan 2 1995 Mon           South Carolina  24       West Virginia  21  0
## 635: 19  Jan 2 1995 Mon (21) Southern California  55          Texas Tech  14  0
## 636: 19  Jan 2 1995 Mon                Wisconsin  34           (25) Duke  20  0
```

Cleaning the team names from the added ranks The added numbers before the team names indicate the current or last season ranks of the teams. We need to get rid of them.

```
teams = unique(c(d94$Winner, d94$Loser))
teams[1:30]
```

```
##  [1] "(4) Nebraska"         "(20) Ohio State"      "(7) Arizona"
##  [4] "Kansas"               "North Carolina State" "Oklahoma State"
##  [7] "Washington State"     "(11) Alabama"         "Arizona State"
## [10] "Arkansas"             "(12) Auburn"          "Baylor"
## [13] "Brigham Young"        "(24) Clemson"         "(8) Colorado"
## [16] "Colorado State"       "Duke"                 "(1) Florida"
## [19] "(4) Florida State"    "Fresno State"         "Georgia"
## [22] "Indiana"              "Iowa"                 "Kansas State"
## [25] "Kentucky"             "(6) Miami (FL)"       "(5) Michigan"
## [28] "Mississippi State"    "Nevada"               "Nevada-Las Vegas"
```

We employ `gsub` to modify strings. The pattern we need to find and erase should be described with Perl-style regular expressions.

```
s1 = teams[1]

s2 = gsub("\\([0-9]+\\) ", '', s1)

cat(paste0(s1, " -> ", s2))
```

```
## (4) Nebraska -> Nebraska
```

Let us first clean the variable `teams`, where we keep all unique team names.

```
teams = gsub("\\([0-9]+\\) ", '', teams)
teams = unique(teams)
teams = sort(teams)
```

The variable teams is useful in order to have all team names sorted alphabetically and now their number.

```
head(as.data.table(teams), 10)
```

```
##                  teams
##  1:          Air Force
##  2:              Akron
##  3:            Alabama
##  4: Alabama-Birmingham
##  5:  Appalachian State
##  6:            Arizona
##  7:      Arizona State
##  8:           Arkansas
##  9:     Arkansas State
## 10:               Army
```

Now, we clean the columns in the data table.

```
d94[, Winner := gsub("\\([0-9]+\\) ", '', Winner)]
d94[, Loser := gsub("\\([0-9]+\\) ", '', Loser)]
```

Let us check that we get the same number of unique team names as in `teams`:

```
length(unique(c(d94$Winner, d94$Loser)))
```

```
## [1] 143
```

Let us check now all the matches for some individual teams. We start with Kansas.

```
d94[Winner == "Kansas" | Loser == "Kansas",]
```

```
##     Wk        Date Day          Winner Pts              Loser Pts HA
##  1:  2  Sep 1 1994 Thu          Kansas  35            Houston  13 -1
##  2:  3 Sep 10 1994 Sat          Kansas  17     Michigan State  10  1
##  3:  4 Sep 17 1994 Sat Texas Christian  31             Kansas  21  1
##  4:  5 Sep 24 1994 Sat          Kansas  72 Alabama-Birmingham   0  1
##  5:  7  Oct 6 1994 Thu    Kansas State  21             Kansas  13 -1
##  6:  8 Oct 15 1994 Sat          Kansas  41         Iowa State  23 -1
##  7:  9 Oct 22 1994 Sat        Oklahoma  20             Kansas  17 -1
##  8: 10 Oct 29 1994 Sat          Kansas  24     Oklahoma State  14  1
##  9: 11  Nov 5 1994 Sat        Nebraska  45             Kansas  17  1
## 10: 12 Nov 12 1994 Sat        Colorado  51             Kansas  26 -1
## 11: 13 Nov 19 1994 Sat          Kansas  31           Missouri  14 -1
```

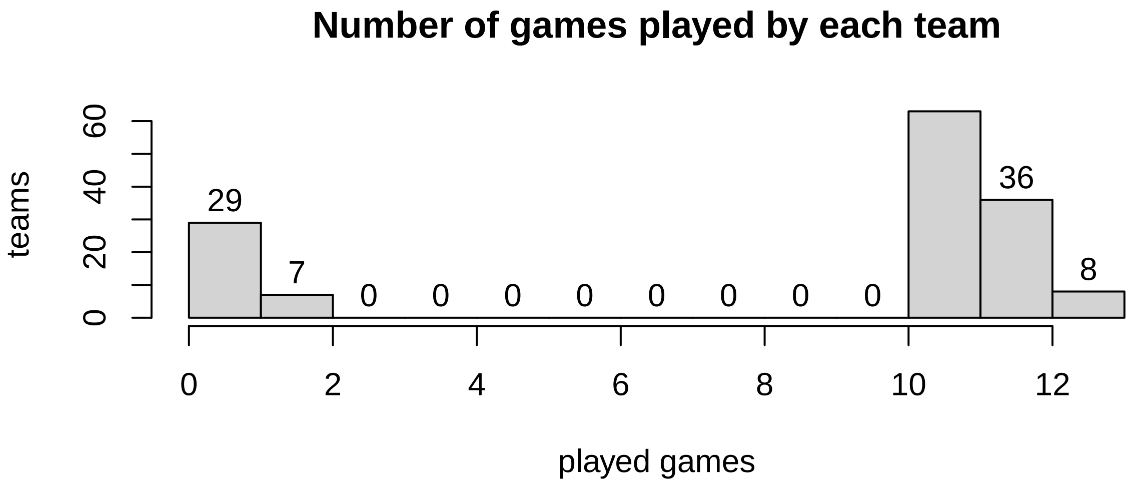
We now consider Nebraska.

```
d94[Winner == "Nebraska" | Loser == "Nebraska",]
```

```
##     Wk        Date Day   Winner Pts          Loser Pts HA
##  1:  1 Aug 28 1994 Sun Nebraska  31  West Virginia   0  1
##  2:  3  Sep 8 1994 Thu Nebraska  42     Texas Tech  16 -1
##  3:  4 Sep 17 1994 Sat Nebraska  49           UCLA  21  1
##  4:  5 Sep 24 1994 Sat Nebraska  70        Pacific  21  1
##  5:  6  Oct 1 1994 Sat Nebraska  42        Wyoming  32  1
##  6:  7  Oct 8 1994 Sat Nebraska  32 Oklahoma State   3  1
##  7:  8 Oct 15 1994 Sat Nebraska  17   Kansas State   6 -1
##  8:  9 Oct 22 1994 Sat Nebraska  42       Missouri   7 -1
##  9: 10 Oct 29 1994 Sat Nebraska  24       Colorado   7  1
## 10: 11  Nov 5 1994 Sat Nebraska  45         Kansas  17  1
## 11: 12 Nov 12 1994 Sat Nebraska  28     Iowa State  12 -1
## 12: 14 Nov 25 1994 Fri Nebraska  13       Oklahoma   3 -1
## 13: 18  Jan 1 1995 Sun Nebraska  24     Miami (FL)  17 -1
```

We see Kansas played only 11 games in regular season. Nebraska played 12 games in regular season and 1 bowl game post-season. Let us check how many games each team has played in season 1994.

```
ngames = table(c(d94$Winner, d94$Loser))
names(ngames) = teams
hist(ngames, labels = T, breaks = 0:13, main = paste("Number of games played by each team"),
     xlab = "played games", ylab = "teams")
```



We see that 36 teams have played only one game or two games. These teams must be from lower divisions. Let us separate the team names into `top_teams` and `low_teams` according to their division.

```
low_teams = names(which(ngames == 1 | ngames == 2))
top_teams = setdiff(teams, low_teams)
```

The teams playing in lower divisions are:

```
low_teams
```

```
##  [1] "Alabama-Birmingham"       "Appalachian State"        "Boise State"
##  [4] "Boston University"        "California-Davis"         "Central Florida"
##  [7] "Chattanooga"              "Citadel"                  "East Tennessee State"
## [10] "Eastern Illinois"         "Eastern Washington"       "Furman"
## [13] "Georgia Southern"         "Holy Cross"               "Idaho"
## [16] "Idaho State"              "Indiana State"            "Jacksonville State"
## [19] "Lafayette"                "Liberty"                  "Missouri State"
## [22] "North Texas"              "Northern Arizona"         "Northern Iowa"
## [25] "Northwestern State"       "Portland State"           "Samford"
## [28] "Southeast Missouri State" "Southern Illinois"        "Texas State"
## [31] "Troy"                     "Weber State"              "Western Carolina"
## [34] "Western Illinois"         "William & Mary"           "Youngstown State"
```

We now add a new logical column `top` to the data table `d94`, having the value `T` if both teams are from division `1-A` and `F` otherwise. Let us also the percentage of games in which both teams are from the top division.

```
d94[, top := !(Winner %in% low_teams | Loser %in% low_teams)]
formattable::percent(mean(d94$top))
```

```
## [1] 93.24%
```

Loading the season's rankings

```
r94 = fread("../../data/1994-College-Football-Rankings.csv")
r94
```

```
##       Rk          School        Conf  AP Rank  W   L T   OSRS    DSRS    SRS
##   1:   1     Penn State    Big Ten       2 12  0 0  20.15   5.87  26.02
##   2:   2         Florida SEC (East)       7 10  2 1  12.94   8.84  21.79
##   3:   3   Florida State         ACC       4 10  1 1  11.24   9.74  20.98
##   4:   4        Nebraska       Big 8       1 13  0 0   9.54  11.11  20.65
##   5:   5        Colorado       Big 8       3 11  1 0  13.05   5.35  18.40
##  ---
## 103: 103         Houston         SWC      NA  1 10 0 -11.34  -6.89 -18.23
## 104: 104  Arkansas State    Big West      NA  1 10 0 -15.89  -2.51 -18.40
## 105: 105      Kent State         MAC      NA  2  9 0 -17.94  -5.78 -23.72
## 106: 106            Ohio         MAC      NA  0 11 0 -23.51  -5.26 -28.78
## 107: 107           Akron         MAC      NA  1 10 0 -17.78 -13.39 -31.17
```

We now can see the number of teams in each Conference.

```
table(r94$Conf)
```

```
## 
##        ACC      Big 8   Big East    Big Ten   Big West        Ind        MAC     Pac-10 
##          9          8          8         11         10         11         10         10 
## SEC (East) SEC (West)        SWC        WAC 
##          6          6          8         10
```

Cleaning differences in team names between files

Seven teams are represented by different abbreviations of their names in the source *csv* files. In the schedule their names appear as:

```
top_teams[which(!top_teams %in% r94$School)]
```

```
## [1] "Brigham Young"       "Louisiana State"     "Mississippi"        
## [4] "Pittsburgh"          "Southern California" "Southern Methodist" 
## [7] "Texas-El Paso"
```

In the rankings their names appear as:

```
r94$School[which(!r94$School %in% top_teams)]
```

```
## [1] "USC"      "BYU"      "LSU"      "Ole Miss" "Pitt"     "UTEP"     "SMU"
```

```
replace_dict = data.table(r94 = c("USC", "BYU", "LSU", "Ole Miss", "Pitt", "UTEP", "SMU"),
                          d94 = c("Southern California", "Brigham Young", "Louisiana State",
replace_dict
```

```
##         r94                 d94
## 1:      USC Southern California
## 2:      BYU       Brigham Young
## 3:      LSU     Louisiana State
## 4: Ole Miss         Mississippi
## 5:     Pitt          Pittsburgh
## 6:     UTEP       Texas-El Paso
## 7:      SMU  Southern Methodist
```

```
set(r94, which(!r94$School %in% top_teams), "School", replace_dict$d94)
```

We create an enhanced table of match schedule with the following added or modified columns: PD (point difference), WS (winner score), LS (loser score), HA (home advantage).

```
x = d94[, c("Wk", "Winner", "Loser")]
x[, WS := d94[[5]]]
```

```
x[, LS := d94[[7]]]
x[, PD := d94[[5]] - d94[[7]]]
x[, HA := d94$HA]
```

We finally save the clean data into two *csv* files.

```
fwrite(x, "../../data/1994-season.csv")
fwrite(r94, "./../../data/1994-rankings.csv")
```