HW 09

Group 8

11/13/2019

We are scraping from the College Basketball page of Sports Reference. One example url is https://sports-reference.com/cbb/players/carmelo-anthony-1.html We are interested in how many games they played and the total minutes to measure the correlation between college experience and rookie year performance.

```
library(XML)
```

```
## Warning: package 'XML' was built under R version 3.5.3
```

```
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.5.3
```

```
## Loading required package: bitops
```

```
## Warning: package 'bitops' was built under R version 3.5.2
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
# read all rookie stats in
data <- read.csv('AllRookieDataWithURL.csv')
```

```
# get urls from br_url column
all_urls <- as.vector(unique(data$br_url))
# get the length of urls
nurls <- length(all_urls)
# create a dataframe for players with NCAA experience
cbb_df <- data.frame(matrix(ncol = 3, nrow = nurls))
```

```
# Scrape college careers
for (i in 1:nurls) {
  # get the appopriate url address
  u <- paste0("https://sports-reference.com/cbb/players/",
              all_urls[i],".html")
  if (url.exists(url = u)) {
    download.file(u, destfile = paste0("cbb.html"))
    doc <- htmlParse("cbb.html")
    g <- getNodeSet(doc, "//*[@id='players_per_game']/tfoot/tr/td[3]")
    mp <- getNodeSet(doc, "//*[@id='players_per_game']/tfoot/tr/td[5]")
    (cbb_df[i, ] <- c(all_urls[i], as.numeric(xmlValue(g[[1]])),
                      as.integer(as.numeric(xmlValue(mp[[1]]))
                                            ) *
                                 as.numeric(xmlValue(g[[1]]))
                                            )
                                 )
                       )
      )
    free(doc)
  }
```

```
}
names(cbb_df)=c('nameurl','col_games','col_minutes')
```

An example link of a page is here, for James Harden: https://www.sports-reference.com/cbb/players/james-harden-1.html

```
# We have 1180 rows of data
nrow(cbb_df)
```

```
## [1] 1180
```

```
head(cbb_df)
```

```
##             nameurl col_games col_minutes
## 1 carmelo-anthony-1        35        1274
## 2    marcus-banks-1        63        2160
## 3              <NA>      <NA>        <NA>
## 4     matt-barnes-1       121        2734
## 5              <NA>      <NA>        <NA>
## 6       troy-bell-1       122        4355
```

```
library(RSQLite)
```

```
## Warning: package 'RSQLite' was built under R version 3.5.3
```

```
dcon <- dbConnect(SQLite(), dbname = "stat405_final_project.db")
```

```
# We have already merged cbb_df with main database, called AllRookieDataWithCBB, through DB Browser
dbListTables(dcon)
```

```
## [1] "AllRookieDataWithCBB"   "AllRookieDataWithDates"
## [3] "AllRookieDataWithURL"   "box_scores"
## [5] "cbb_df"
```

```
# We show first 1000 rows here
# Each row represents a game played by a rookie, so if a rookie played 80 games, his
# college stats will show up 80 times, but we decide to sacrifice redundancy for
# future simplicity on manipulations
res <- dbSendQuery(conn = dcon, "
SELECT *
FROM AllRookieDataWithCBB
LIMIT 1000;
")
mydf <- dbFetch(res, -1)
```

```
## Warning in result_fetch(res@ptr, n = n): Column `Total_Rebounds`: mixed
## type, first seen values of type integer, coercing other values of type
## string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `points_per_36`: mixed
## type, first seen values of type real, coercing other values of type string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `reb_per_36`: mixed type,
## first seen values of type real, coercing other values of type string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `ast_per_36`: mixed type,
## first seen values of type real, coercing other values of type string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `stl_per_36`: mixed type,
## first seen values of type real, coercing other values of type string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `blk_per_36`: mixed type,
## first seen values of type real, coercing other values of type string
```

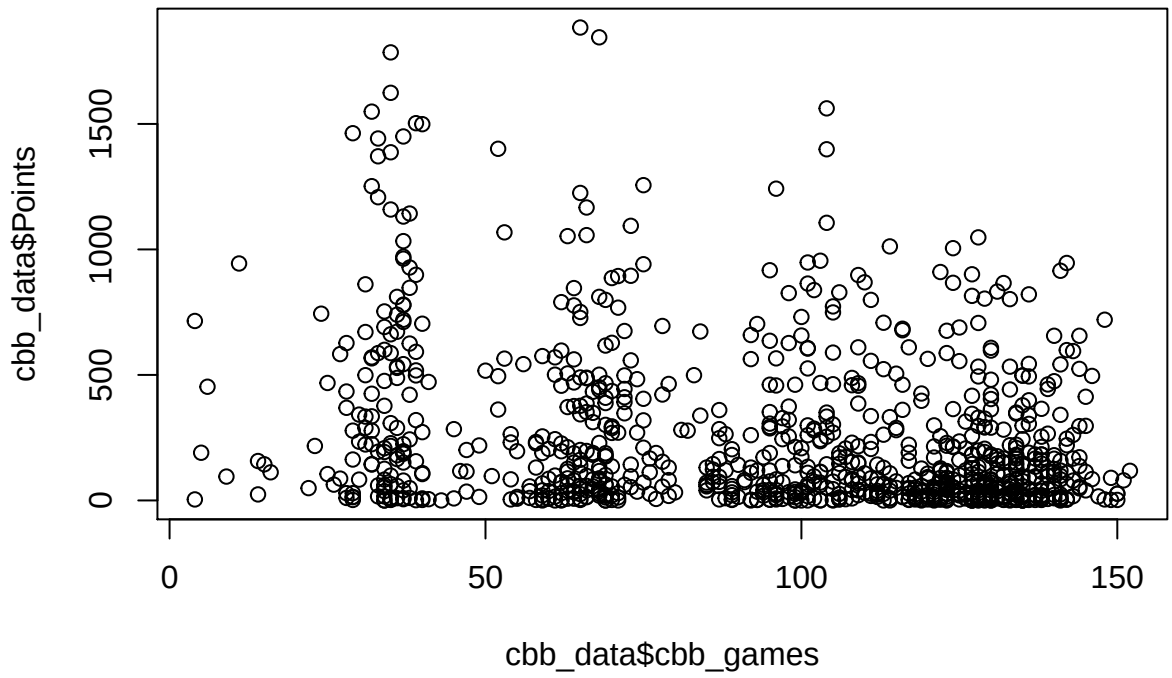
```
dbClearResult(res)
View(mydf)
```

```
dbDisconnect(dcon)
cbb_data=read.csv('GroupedRookieDataWithCBB.csv')
```

Using URLs like https://www.sports-reference.com/cbb/players/james-harden-1.html, we pulled the number of games played, and minutes played, by every player in our dataset. Our dataset consists of all rookies (first year players) from the 2003-04 season through the 2018-19 season, thus we have 16 seasons of approximately 60-70 rookies through which we loop and pull their number of games and minutes in college. The point of this is that our project is going to be analyzing the "readiness" of players for the NBA based on college and international experience. We'll be extracting similar info for international players later.

This data is then added to our dataframe cbb_data to identify players' stats, but also their number of games and minutes in college to run statistical tests and compare college and international experience with future NBA experience. Here is an example of a statistical test of correlation, of points per 36 minutes vs cbb_games played.

```
plot(cbb_data$cbb_games,cbb_data$Points)
```



What this shows is a clear group of four different sets of players, those who played only one college season through those who played four seasons. Traditionally players who are drafted after only one year of college basketball are the best of the best, and therefore it makes sense that a lot of the players with the most points played only one or two seasons in college (the highest values on the graph appear more towards the left, indicating more points scored.