Scraping (Outside of Written Functions)

Samad Patel

8/21/2018

This document will contain all the relevant code for scraping and data cleaning that is done outside of any user-written functions.

Much of this is completed using splashr and docker. My Python isn't configured such that R can access the needed containers directly, so there is some command line code being run outside of this document. That which is entered into the command line will be commented.

1) Abbreviations

```
page <- read_html('https://en.wikipedia.org/wiki/Wikipedia:WikiProject_National_Basketball_Association/
Abbs <- page %>% html_nodes('td') %>% html_text()
Abbrevs <- str_replace_all(string = Abbs[seq(3,length(Abbs), by=2)], pattern = '\\n',
                           replacement = '')
FullNames <- str_replace_all(Abbs[seq(4, length(Abbs), by = 2)], pattern = '\\n',
                             replacement = '')
abbs_df <- data.frame('Abbreviation' = Abbrevs, 'Franchise' = FullNames, stringsAsFactors = FALSE)
```

2) Possessions Per Game

```
# docker run -p 8050:8050 -p 5023:5023 scrapinghub/splash
page <- render_html(url = 'https://www.teamrankings.com/nba/stat/possessions-per-game')
PossPerGame <- page %>% html_nodes('#DataTables_Table_0') %>% html_table()
PossPerGame <- PossPerGame[[1]]
PossPerGame <- PossPerGame %>% select(Team, `2017`, Home, Away)
# Franchise names are very inconvenient for cleaning
PossPerGame$Team <- c('NOP', 'LAL', 'PHX', 'PHI', 'BKN', 'GSW', 'LAC', 'CHI', 'CHA', 'ATL',
          'ORL', 'TOR', 'DEN', 'OKC', 'NYK', 'WAS', 'HOU', 'POR', 'MIA', 'CLE',
          'MIL', 'DET', 'MIN', 'BOS', 'UTA', 'DAL', 'IND', 'MEM', 'SAC', 'SAS')
head(PossPerGame)
```

```
##   Team  2017  Home  Away
## 1  NOP 104.9 105.0 104.9
## 2  LAL 104.6 104.4 104.8
## 3  PHX 103.6 104.2 103.1
## 4  PHI 103.4 103.7 103.1
## 5  BKN 102.8 102.2 103.3
## 6  GSW 102.3 101.3 103.2
```

3) Opponent's Points Per Game

```
page <- render_html(url = 'https://www.basketball-reference.com/leagues/NBA_2018.html')
OPtsPerGame <- page %>% html_nodes('#opponent-stats-per_game') %>% html_table()
OPtsPerGame <- OPtsPerGame[[1]]
```

```r
OPtsPerGame <- OPtsPerGame %>% select(Team, PTS)
# Remove the stars
OPtsPerGame$Team <- str_replace_all(OPtsPerGame$Team, '\\*', '')
# Remove league average
OPtsPerGame <- OPtsPerGame[1:30, ]
# Replaces team names with abbreviations
OPtsPerGame$Team <- abbs_df$Abbreviation[match(OPtsPerGame$Team, abbs_df$Franchise)]
head(OPtsPerGame)
```

```
##   Team   PTS
## 1  UTA  99.8
## 2  SAS  99.8
## 3  BOS 100.4
## 4  MIA 102.9
## 5  POR 103.0
## 6  HOU 103.9
```

```r
rm(abbs_df, page, Abbrevs, Abbs, FullNames)
save.image(file = 'PreLoadedData.RData')
```