615Midtermmmmmmm

Kaiyu Yan,Si Chen, Wenjia Xie, Siwei Hu

October 17, 2018

```
##### Basketball #####
library(xml2)
library(rvest)
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------------------------------- tidyverse 1.2.1

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ----------------------------------------------------------------- tidyverse_conflicts()
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```
site1 <- "https://www.basketball-reference.com/leagues/NBA_"
site2 <- "_games-"
site3 <- ".html"
year <- c(2013:2018)
month <- c("january","february","march","april","may","june","october","november","december")
month_index<-c(1:9)
year_index<-c(1:6)
name1<-c()
name2<-c()
total.date<-c()
total.attdence<-c()
for (i in year_index){
  for (j in month_index){
    site<-paste(site1,year[i],site2,month[j],site3,sep="")
    webpage<-read_html(site)
    name1<- webpage %>% html_nodes('.left:nth-child(1)') %>% html_attrs()
    name2<- webpage %>% html_nodes('.center+.right') %>% html_text()
    total.date <- c(total.date,name1)
    total.attdence <- c(total.attdence,name2)
    j<-j+1
  }
  i<-i+1
}

total.date1 <- t(data.frame((total.date)))[,4]
total.attdence1 <- data.frame(total.attdence)
b_data<-cbind(total.date1,total.attdence1)[-1,]
colnames(b_data)<-c("date","attendance")
rownames(b_data)<-rep(1:7934)
host<-substr(b_data$date, 10, 12)
b_data$date<-substr(b_data$date,1,9)
```

```
b_data3<-cbind(b_data,host)

Homegame <- filter(b_data3,str_detect(host,"BOS"))
Homegame$date <- substring(Homegame$date, 1,8)
Homegame$date <- as.Date(Homegame$date,"%Y%m%d")

Home_2012 <- filter(Homegame,str_detect(date,"2012"))
Home_2013 <- filter(Homegame,str_detect(date,"2013"))
Home_2014 <- filter(Homegame,str_detect(date,"2014"))
Home_2015 <- filter(Homegame,str_detect(date,"2015"))
Home_2016 <- filter(Homegame,str_detect(date,"2016"))
Home_2017 <- filter(Homegame,str_detect(date,"2017"))

weather_data <- read.csv("weather data.csv")
colnames(weather_data)[colnames(weather_data)=="DATE"] <- "date"
weather_data$date <- as.Date(weather_data$date,"%Y-%m-%d")
typelist = c("Fog","Mist","Drizze","Rain","Snow")
type_code = c("WT01","WT13","WT14","WT16","WT18")

Celtics_2012<- inner_join(Home_2012,weather_data,by = "date", match = all)
Celtics_2013<- inner_join(Home_2013,weather_data,by = "date", match = all)
Celtics_2014<- inner_join(Home_2014,weather_data,by = "date", match = all)
Celtics_2015<- inner_join(Home_2015,weather_data,by = "date", match = all)
Celtics_2016<- inner_join(Home_2016,weather_data,by = "date", match = all)
Celtics_2017<- inner_join(Home_2017,weather_data,by = "date", match = all)
Celtics_All <- do.call("rbind", list(Celtics_2012,Celtics_2013,Celtics_2014,Celtics_2015,Celtics_2016,Ce

Celtics_All$type<-NA

for (i in 1:length(typelist)) {
  colnames(Celtics_All)[which(colnames(Celtics_All)==type_code[i])] = typelist[i]
}

Celtics_All[is.null(Celtics_All)] <- NA

##Run through all types to get the weather of a certain day, add that to the "type" column
for (m in 1:dim(Celtics_All)[1]) {
  t<-0
  for (n in 1:length(typelist)) {
    if (is.null(Celtics_All[m,typelist[n]])) {
      Celtics_All[m,typelist[n]] = NA
    }
    if (!is.na(Celtics_All[m,typelist[n]])) {
      Celtics_All[m,"type"] =  typelist[n]
      t<-t+1
    }
  }
  if(t==0)
    Celtics_All[m,"type"] = "normal"
}
#######  Baseball Data  ######
site1 <- "https://www.basketball-reference.com/leagues/NBA_"
site2 <- "_games-"
```

```r
site3 <- ".html"
year <- c(2013:2018)
month <- c("january","february","march","april","may","june","october","november","december")
month_index<-c(1:9)
year_index<-c(1:6)
name1<-c()
name2<-c()
total.date<-c()
total.attdence<-c()
for (i in year_index){
  for (j in month_index){
    site<-paste(site1,year[i],site2,month[j],site3,sep="")
    webpage<-read_html(site)
    name1<- webpage %>% html_nodes('.left:nth-child(1)') %>% html_attrs()
    name2<- webpage %>% html_nodes('.center+.right') %>% html_text()
    total.date <- c(total.date,name1)
    total.attdence <- c(total.attdence,name2)
    j<-j+1
  }
  i<-i+1
}

total.date1 <- t(data.frame((total.date)))[,4]
total.attdence1 <- data.frame(total.attdence)
b_data<-cbind(total.date1,total.attdence1)[-1,]
colnames(b_data)<-c("date","attendance")
rownames(b_data)<-rep(1:7934)
host<-substr(b_data$date, 10, 12)
b_data$date<-substr(b_data$date,1,9)
b_data3<-cbind(b_data,host)

site1 <- "https://www.baseball-reference.com/teams/BOS/"
site2 <- "-schedule-scores.shtml"
year <- c(2012:2017)
site <- paste0(site1,year,site2)

Raw_2012<- as.data.frame(read_html(site[1]) %>% html_nodes("table") %>% html_table())
Raw_2013<- as.data.frame(read_html(site[2]) %>% html_nodes("table") %>% html_table())
Raw_2014<- as.data.frame(read_html(site[3]) %>% html_nodes("table") %>% html_table())
Raw_2015<- as.data.frame(read_html(site[4]) %>% html_nodes("table") %>% html_table())
Raw_2016<- as.data.frame(read_html(site[5]) %>% html_nodes("table") %>% html_table())
Raw_2017<- as.data.frame(read_html(site[6]) %>% html_nodes("table") %>% html_table())

weather_data <- read.csv("weather data.csv")
weather_data <- select(weather_data,DATE,TAVG,WT01,WT13,WT14,WT16,WT18)
typelist = c("Fog","Mist","Drizze","Rain","Snow")
type_code = c("WT01","WT13","WT14","WT16","WT18")
weather_data$type<-NA

for (i in 1:length(typelist)) {
  colnames(weather_data)[which(colnames(weather_data)==type_code[i])] = typelist[i]
}
```

```r
weather_data[is.null(weather_data)] <- NA

##Run through all types to get the weather of a certain day, add that to the "type" column
for (m in 1:dim(weather_data)[1]) {
  t<-0
  for (n in 1:length(typelist)) {
    if (is.null(weather_data[m,typelist[n]])) {
      weather_data[m,typelist[n]] = NA
    }
    if (!is.na(weather_data[m,typelist[n]])) {
      weather_data[m,"type"] =  typelist[n]
      t<-t+1
    }
  }
  if(t==0)
    weather_data[m,"type"] = "normal"
}



# Read and select home game from 2017 Dataset
Game_2017 <- select(Raw_2017, Date, Tm:Opp, contains("D/N"), Attendance)
Data_2017<- data.frame(do.call('rbind', strsplit(as.character(Game_2017$Date),',',fixed=TRUE)))
Data_2017<- data.frame(do.call('rbind', strsplit(as.character(Data_2017$X2),' ',fixed=TRUE)))
```

```
## Warning in rbind(c("", "Apr", "3"), c("", "Apr", "5"), c("", "Apr", "7"), :
## number of columns of result is not a multiple of vector length (arg 1)
```

```r
Game_2017 <- merge(Game_2017,Data_2017,by = 0)
Clean_2017 <- filter(Game_2017, !str_detect(Var.5,"@"))
Clean_2017 <- filter(Clean_2017,str_detect(Tm,"BOS"))
C_2017 <- select(Clean_2017, 6,8,9)
C_2017$X2 <- match(C_2017$X2,month.abb)
C_2017$year <- rep(2017,nrow(C_2017)) # make new column
A <- C_2017[,c(1,4,2,3)]
C_2017 <- unite(A, Date,2:4, sep = "-", remove = TRUE)
C_2017$Date <- as.Date(C_2017$Date,"%Y-%m-%d")


# Read and select home game from 2016 Dataset
Game_2016 <- select(Raw_2016, Date, Tm:Opp, contains("D/N"), Attendance)
Data_2016<- data.frame(do.call('rbind', strsplit(as.character(Game_2016$Date),',',fixed=TRUE)))
Data_2016<- data.frame(do.call('rbind', strsplit(as.character(Data_2016$X2),' ',fixed=TRUE)))
```

```
## Warning in rbind(c("", "Apr", "5"), c("", "Apr", "6"), c("", "Apr", "8"), :
## number of columns of result is not a multiple of vector length (arg 1)
```

```r
Game_2016 <- merge(Game_2016,Data_2016,by = 0)
Clean_2016 <- filter(Game_2016, !str_detect(Var.5,"@"))
Clean_2016 <- filter(Clean_2016,str_detect(Tm,"BOS"))
C_2016 <- select(Clean_2016, 6,8,9)
C_2016$X2 <- match(C_2016$X2,month.abb)
C_2016$year <- rep(2016,nrow(C_2016)) # make new column
A <- C_2016[,c(1,4,2,3)]
C_2016 <- unite(A, Date,2:4, sep = "-", remove = TRUE)
```

```r
C_2016$Date <- as.Date(C_2016$Date,"%Y-%m-%d")
C_2016 <- na.omit(C_2016)

# Read and select home game from 2015 Dataset
Game_2015 <- select(Raw_2015, Date, Tm:Opp, contains("D/N"), Attendance)
Data_2015<- data.frame(do.call('rbind', strsplit(as.character(Game_2015$Date),',',fixed=TRUE)))
Data_2015<- data.frame(do.call('rbind', strsplit(as.character(Data_2015$X2),' ',fixed=TRUE)))
```

```
## Warning in rbind(c("", "Apr", "6"), c("", "Apr", "8"), c("", "Apr", "9"), :
## number of columns of result is not a multiple of vector length (arg 1)
```

```r
Game_2015 <- merge(Game_2015,Data_2015,by = 0)
Clean_2015 <- filter(Game_2015, !str_detect(Var.5,"@"))
Clean_2015 <- filter(Clean_2015,str_detect(Tm,"BOS"))
C_2015 <- select(Clean_2015, 6,8,9)
C_2015$X2 <- match(C_2015$X2,month.abb)
C_2015$year <- rep(2015,nrow(C_2015)) # make new column
A <- C_2015[,c(1,4,2,3)]
C_2015 <- unite(A, Date,2:4, sep = "-", remove = TRUE)
C_2015$Date <- as.Date(C_2015$Date,"%Y-%m-%d")


# Read and select home game from 2014 Dataset
Game_2014 <- select(Raw_2014, Date, Tm:Opp, contains("D/N"), Attendance)
Data_2014<- data.frame(do.call('rbind', strsplit(as.character(Game_2014$Date),',',fixed=TRUE)))
Data_2014<- data.frame(do.call('rbind', strsplit(as.character(Data_2014$X2),' ',fixed=TRUE)))
```

```
## Warning in rbind(c("", "Mar", "31"), "April", c("", "Apr", "2"), c("",
## "Apr", : number of columns of result is not a multiple of vector length
## (arg 1)
```

```r
Game_2014 <- merge(Game_2014,Data_2014,by = 0)
Clean_2014 <- filter(Game_2014, !str_detect(Var.5,"@"))
Clean_2014 <- filter(Clean_2014,str_detect(Tm,"BOS"))
C_2014 <- select(Clean_2014, 6,8,9)
C_2014$X2 <- match(C_2014$X2,month.abb)
C_2014$year <- rep(2014,nrow(C_2014)) # make new column
A <- C_2014[,c(1,4,2,3)]
C_2014 <- unite(A, Date,2:4, sep = "-", remove = TRUE)
C_2014$Date <- as.Date(C_2014$Date,"%Y-%m-%d")


# Read and select home game from 2013 Dataset
Game_2013 <- select(Raw_2013, Date, Tm:Opp, contains("D/N"), Attendance)
Data_2013<- data.frame(do.call('rbind', strsplit(as.character(Game_2013$Date),',',fixed=TRUE)))
Data_2013<- data.frame(do.call('rbind', strsplit(as.character(Data_2013$X2),' ',fixed=TRUE)))
```

```
## Warning in rbind(c("", "Apr", "1"), c("", "Apr", "3"), c("", "Apr", "4"), :
## number of columns of result is not a multiple of vector length (arg 1)
```

```r
Game_2013 <- merge(Game_2013,Data_2013,by = 0)
Clean_2013 <- filter(Game_2013, !str_detect(Var.5,"@"))
Clean_2013 <- filter(Clean_2013,str_detect(Tm,"BOS"))
C_2013 <- select(Clean_2013, 6,8,9)
C_2013$X2 <- match(C_2013$X2,month.abb)
C_2013$year <- rep(2013,nrow(C_2013)) # make new column
```

```r
A <- C_2013[,c(1,4,2,3)]
C_2013 <- unite(A, Date,2:4, sep = "-", remove = TRUE)
C_2013$Date <- as.Date(C_2013$Date,"%Y-%m-%d")

# Read and select home game from 2012 Dataset
Game_2012 <- select(Raw_2012, Date, Tm:Opp, contains("D/N"), Attendance)
Data_2012<- data.frame(do.call('rbind', strsplit(as.character(Game_2012$Date),',',fixed=TRUE)))
Data_2012<- data.frame(do.call('rbind', strsplit(as.character(Data_2012$X2),' ',fixed=TRUE)))
```

```
## Warning in rbind(c("", "Apr", "5"), c("", "Apr", "7"), c("", "Apr", "8"), :
## number of columns of result is not a multiple of vector length (arg 1)
```

```r
Game_2012 <- merge(Game_2012,Data_2012,by = 0)
Clean_2012 <- filter(Game_2012, !str_detect(Var.5,"@"))
Clean_2012 <- filter(Clean_2012,str_detect(Tm,"BOS"))
C_2012 <- select(Clean_2012, 6,8,9)
C_2012$X2 <- match(C_2012$X2,month.abb)
C_2012$year <- rep(2012,nrow(C_2012)) # make new column
A <- C_2012[,c(1,4,2,3)]
C_2012 <- unite(A, Date,2:4, sep = "-", remove = TRUE)
C_2012$Date <- as.Date(C_2012$Date,"%Y-%m-%d")


#join with Weather


#2012
Weather_2012 <- filter(weather_data,str_detect(DATE,"2012"))
colnames(Weather_2012)[colnames(Weather_2012)=="DATE"] <- "Date"
Weather_2012$Date <- as.Date(Weather_2012$Date,"%Y-%m-%d")
Table_2012<- inner_join(C_2012,Weather_2012,by = "Date", match = all)


#2013
Weather_2013 <- filter(weather_data,str_detect(DATE,"2013"))
Weather_2013 <- filter(weather_data,str_detect(DATE,"2013"))
colnames(Weather_2013)[colnames(Weather_2013)=="DATE"] <- "Date"
Weather_2013$Date <- as.Date(Weather_2013$Date,"%Y-%m-%d")
Table_2013<- inner_join(C_2013,Weather_2013,by = "Date", match = all)

#2014
Weather_2014 <- filter(weather_data,str_detect(DATE,"2014"))
Weather_2014 <- filter(weather_data,str_detect(DATE,"2014"))
colnames(Weather_2014)[colnames(Weather_2014)=="DATE"] <- "Date"
Weather_2014$Date <- as.Date(Weather_2014$Date,"%Y-%m-%d")
Table_2014<- inner_join(C_2014,Weather_2014,by = "Date", match = all)

#2015
Weather_2015 <- filter(weather_data,str_detect(DATE,"2015"))
Weather_2015 <- filter(weather_data,str_detect(DATE,"2015"))
colnames(Weather_2015)[colnames(Weather_2015)=="DATE"] <- "Date"
Weather_2015$Date <- as.Date(Weather_2015$Date,"%Y-%m-%d")
Table_2015<- inner_join(C_2015,Weather_2015,by = "Date", match = all)
```

```
#2016
Weather_2016 <- filter(weather_data,str_detect(DATE,"2016"))
Weather_2016 <- filter(weather_data,str_detect(DATE,"2016"))
colnames(Weather_2016)[colnames(Weather_2016)=="DATE"] <- "Date"
Weather_2016$Date <- as.Date(Weather_2016$Date,"%Y-%m-%d")
Table_2016<- inner_join(C_2016,Weather_2016,by = "Date", match = all)

#2017
Weather_2017 <- filter(weather_data,str_detect(DATE,"2017"))
Weather_2017 <- filter(weather_data,str_detect(DATE,"2017"))
colnames(Weather_2017)[colnames(Weather_2017)=="DATE"] <- "Date"
Weather_2017$Date <- as.Date(Weather_2017$Date,"%Y-%m-%d")
Table_2017<- inner_join(C_2017,Weather_2017,by = "Date", match = all)


##merge

Table_all <- do.call("rbind", list(Table_2012,Table_2013,Table_2014,Table_2015,Table_2016,Table_2017))
```