Data-Wrangling

Ziji Zhou and Rohil Bathija

October 14, 2021

NBA Dataset

Get the payroll of each team for a certain year range.

things we need: total payroll of team by year total team % for the year https://www.basketball-reference.com/ each teams zipcode DONE any data about zipcode DONE highest paid players: DONE http://www.espn.com/nba/salaries/_/year/2020/seasontype/1 stats for those players.

Historical Salaries

Workflow

Seasons -> teams -> player salaries. Seasons url: https://www.basketball-reference.com/leagues/

```
# check paths
nba_url <- "https://www.basketball-reference.com/leagues/"

paths_allowed(nba_url)
```

```
## [1] TRUE
```

```
# Team historical win% and total salary

# get table

nba_data <- nba_url %>%
  read_html() %>%
  html_nodes("table") %>%
  purrr::pluck(1) %>%
  html_table() %>%
  janitor::clean_names() %>%
  # filter years wanted
  filter(str_detect(x,"^20|199")) %>%
  filter(!str_detect(x,"^2021")) %>%
  # rename
  rename(season = x) %>%
  # add link column and delete extra columns
  select("season") %>%
  mutate(link_bref = "")

# initialize big data frame
team_info <- data_frame(season = character(),
                        team = character(),
                        #abr = character(),
                        w = integer(),
                        l = integer(),
                        winper = numeric(),
                        total_salary = numeric()
                        )
#initialize all team names vector
all_team_names = c()

# manual urls per season
rows = 1
years = 2021:1991
for(i in years){
  nba_data$link_bref[rows] = paste("https://www.basketball-reference.com/leagues/NBA_",i,".html",sep =
  rows = rows + 1
}

# add link for each team on each season
for(i in 1:nrow(nba_data)){
  # set url
  url <- nba_data$link_bref[i]

  # get the team names as a vector
  team_name <- url %>%
    read_html() %>%
    html_elements("#per_game-team > tbody > tr > td.left > a") %>%
    html_text()

  all_team_names = c(all_team_names,team_name)

  # get the links of each team
  team_link <- url %>%
    read_html() %>%
    html_elements("#per_game-team > tbody > tr > td.left > a") %>%
    html_attr("href")

  # adjust the links
  for(x in 1:length(team_link)){
    team_link[x] <- paste("https://www.basketball-reference.com",team_link[x],sep = "")
  }

  teams <- data_frame(name = team_name, link = team_link)

  # go through each team
  for(x in 1:nrow(teams)){
    # get link
    url <- teams$link[x]

    #get team name
    teamname <- teams$name[x]

    page <- url %>%
      read_html()

    # get win % for each team
```

```r
# take the right element
team_stats <- page %>%
  html_elements("#meta > div:nth-child(2) > p:nth-child(3)") %>%
  html_text() %>%
  str_split("\n") %>%
  pluck(1) %>%
  str_split(",") %>%
  pluck(4) %>%
  pluck(1)
#formatting

team_stats <- sub("      ", "", team_stats)

#calculate
win = as.numeric(team_stats %>%
  str_split("-") %>%
  pluck(1) %>%
  pluck(1))
lose = as.numeric(team_stats %>%
  str_split("-") %>%
  pluck(1) %>%
  pluck(2))

team_winper = win/(win+lose)

#salary total
#get elements
salary <- page %>%
  html_elements("#all_salaries2")

#manualy sift through the data since selector won't work for some reason
salary <- as.character(salary)%>%
  substring(gregexpr("<!--",as.character(salary))) %>%
  substring(5) %>%
  minimal_html()%>%
  html_nodes("table") %>%
  pluck(1) %>%
  html_table()

salary_clean <- salary %>%
  rename(name = "") %>%
  mutate(Salary = as.numeric(
    str_remove_all(
      substring(Salary,2), ",")
    )
  ) %>%
  janitor::clean_names()

total_salary = sum(salary_clean$salary)

#add to big data frame
df <- data_frame(
      season = nba_data$season[i],
          team = teamname,
          #abr = character(),
          w = win,
          l = lose,
          winper = team_winper,
          total_salary = total_salary
          )

    team_info <- rbind(team_info,df)
  }

}

#clean out data and added start year column
team_info_clean <- team_info %>%
  unique() %>%
  arrange(season,desc(winper)) %>%
  mutate(start_year = as.integer(
    substring(season,
              first = 1,
              last = 4)
    )
  )

write_csv(team_info_clean, "data/team_info.csv")
```

```
# nba city and median income

# convert the excel zip codes into csv for zip code info

library(readxl)

nba_zip_codes <- read_excel("data/nba-zip-codes.xlsm")

# source: incomebyzipcode.com
zipcode_income <- read_excel("data/zipcodeavgsal.xlsm")


# tidy and join data

zipcode_income <- zipcode_income %>%
  janitor::clean_names() %>%
  rename(zip = zip_codes)

nba_zip_codes <- nba_zip_codes %>%
  janitor::clean_names()

nba_city_income <- nba_zip_codes %>%
  inner_join(zipcode_income, by = c("zip" = "zip"))
```

```
#NBA top player salary

# from basketball reference
```

```
nba_salary <- read_excel("data/nba_salary.xlsm")

#to create the codes
nba_salary$codes <- sapply(strsplit(as.character(nba_salary$Player), "\\\\"),"[", 3)
#to rename the players to just their name
nba_salary$Player <- sapply(strsplit(as.character(nba_salary$Player), "\\\\"),"[", 1)
head(nba_salary)
```

```
# A tibble: 6 x 12
  Rk    Player Tm    ‘2021-22‘ ‘2022-23‘ ‘2023-24‘ ‘2024-25‘ ‘2025-26‘ ‘2026-27‘
  <chr> <chr>  <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 1     Steph~ GSW    45780966  48070014  51915615  55761216  59606817        NA
2 2     John ~ HOU    44310840  47366760        NA        NA        NA        NA
3 3     Russe~ LAL    44211146  47063478        NA        NA        NA        NA
4 4     James~ BRK    43848000  46872000        NA        NA        NA        NA
5 5     Damia~ POR    43750000  47250000  50750000  54250000        NA        NA
6 6     LeBro~ LAL    41180544  44474988        NA        NA        NA        NA
# ... with 3 more variables: Signed Using <chr>, Guaranteed\ <chr>, codes <chr>
```

```
#only need certain columns
nba_salary_clean <- nba_salary %>%
  select(Rk, Player, Tm, "2021-22", codes) %>%
  #renamed the salary column
  rename(Salary = "2021-22") %>%
  #taking out Toronto because of a lack of data
  filter(Tm != "TOR") %>%
# adding in links
  mutate( links =
            paste("https://www.basketball-reference.com/players/",
                  substring(codes, 1, 1),
                  "/",
                  codes,
                  ".html",
                  sep = ""),
          #adding in variables for player stats
          games = 0,
          points = 0,
          rebounds = 0,
          assists = 0
          )
#creating vectors for each stat
games = c()
points = c()
rebounds = c()
assists = c()
#using a for loop to go through each player and get his stats
for(i in 1:nrow(nba_salary_clean)){
  #getting url and storing the finsihed html in a variable
  url <- nba_salary_clean$links[i]
  page <- url %>%
    read_html()
  #finding each statistic
  games[i] <- page %>%
    html_elements("#info > div.stats_pullout > div.p1 > div:nth-child(1) > p:nth-child(2)") %>%
    html_text()
  #checking if the value is blank and setting to 0 if so.
  if(games[i] == ""){
    games[i] = "0"
  }
  #adding it to our main dataset
  nba_salary_clean$games[i] = as.integer(games[i])

  #points (same as game just for points)
  #only thing that is different is url
  points[i] <- page %>%
    html_elements("#info > div.stats_pullout > div.p1 > div:nth-child(2) > p:nth-child(2)") %>%
    html_text()
  if(points[i] == ""){
    points[i] = "0"
  }
  nba_salary_clean$points[i] = as.integer(points[i])

  #rebounds (same as game just for rebounds)
  #only thing that is different is url
  rebounds[i] <- page %>%
    html_elements("#info > div.stats_pullout > div.p1 > div:nth-child(3) > p:nth-child(2)") %>%
    html_text()
  if(rebounds[i] == ""){
    rebounds[i] = "0"
  }
  nba_salary_clean$rebounds[i] = as.integer(rebounds[i])

  #assists (same as game just for assists)
  #only thing that is different is url
  assists[i] <- page %>%
    html_elements("#info > div.stats_pullout > div.p1 > div:nth-child(4) > p:nth-child(2)") %>%
    html_text()
  if(assists[i] == ""){
    assists[i] = "0"
  }
  nba_salary_clean$assists[i] = as.integer(assists[i])
}
nba_salary_master <- nba_city_income %>%
  inner_join(nba_salary_clean, by = c("abbreviations" = "Tm"))
```

```
#code chunk to make final edits to dataset and have it be the one used in shiny
nba_salaries <- nba_salary_master
#creating a final dataset with desired columns and values for shiny implementation
nba_salaries <- nba_salaries %>%
  #adding in columns to have how many times the average salary a players salary is (by area)
  mutate(salaryProportion = Salary/med_sal) %>%
  #selecting only certain columns
  select(c("team", "abbreviations", "Player", "games", "points", "rebounds", "assists", "salaryProporti
write_csv(nba_salaries, "data/nba_salaries.csv")
```