Dataset Creation

Matthew Murray

2022-12-28

```r
# libraries

library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.0.5
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages ----------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.4

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.4

## Warning: package 'forcats' was built under R version 4.0.4

## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.4

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(stringi)
```

```
## Warning: package 'stringi' was built under R version 4.0.5
```

Soccer

```
# taken from Sports Illustrated
# https://www.si.com/soccer/2019/05/21/50-greatest-footballers-all-time

Players <- c("Diego Maradona", "Pele", "Lionel Messi", "Franz Beckenbauer",
             "Johan Cruyff", "Zinedine Zidane", "Ronaldo (Brazil)", "Michel Platini",
             "Roberto Baggio", "Alfredo di Stefano", "Mane Garrincha", "Gerd Muller",
             "Paolo Maldini", "Ferenc Puskas", "Cristiano Ronaldo", "Franco Baresi",
             "Zico", "George Best", "Marco van Basten", "Eusebio",
             "Romario", "Raymond Kopa", "Giuseppe Meazza", "Bobby Charlton",
             "Ruud Gullit", "Ronaldinho", "Lothar Matthaus", "Sandor Kocsis",
             "Socrates", "Bobby Moore", "Rivelino", "Andres Iniesta",
             "Karl-Heinz Rummenigge", "Luis Suarez Miramontes", "Xavi Hernandez", "Johan Neeskens",
```

```
                "Gianluigi Buffon", "Hristo Stoichkov", "Kevin Keegan", "Gunnar Nordahl",
                "Lev Yashin", "Kakka", "George Weah", "Paul Breitner",
                "Paolo Rossi", "Omar Sivori", "Jairzinho", "Hugo Sanchez",
                "John Charles", "Luka Modric")

Peak <- c("1985-1990", "1958-1964", "2009-2012", "1966-1976",
          "1971-1975", "1997-2002", "1995-1998", "1982-1985",
          "1990-1994", "1956-1962", "1958-1962", "1970-1974",
          "1994-2003", "1950-1954", "2012-2016", "1987-1994",
          "1979-1982", "1966-1971", "1988-1992", "1962-1968",
          "1990-1994", "1956-1959", "1934-1938", "1963-1967",
          "1986-1990", "2004-2006", "1988-1992", "1950-1955",
          "1976-1984", "1964-1970", "1968-1974", "2008-2012",
          "1979-1984", "1960-1965", "2008-2012", "1971-1978",
          "2002-2006", "1990-1994", "1977-1979", "1950-1955",
          "1956-1964", "2005-2009", "1993-1996", "1974-1981",
          "1982-1982", "1958-1961", "1970-1970", "1986-1990",
          "1957-1960", "2014-2018")

rk <- seq(1, 50)



soccer <- data.frame(Players, Peak, rk)
colnames(soccer) = c("Player", "Peak", "Rank")
```

Cricket

```
# taken from BBC
# https://www.bbc.co.uk/programmes/articles/2V6BjFgdJ5KcfVHhR3bwBLz/the-greatest-cricketer-of-all-time-

Players <- c("Sir Donald Bradman", "Sachin Tendulkar", "Sir Garfield Sobers", "Imran Khan",
             "Sir Ian Botham", "Shane Warne", "Sir Viv Richards", "Brian Lara",
             "Jaques Kallis", "MS Dhoni", "Wasim Akram", "Virat Kohli",
             "James Anderson", "Sir Alastair Cook", "Muttiah Muralitharan", "Kumar Sangakkara",
             "Kapil Dev", "Sir Richard Hadlee", "Adam Gilchrist", "Chris Gayle",
             "Glenn McGrath", "Ricky Ponting", "Steve Waugh", "Rahul Dravid",
             "Sunil Gavaskar", "Shoaib Akhtar", "Sir Curtly Ambrose", "Mahela Jayawardene",
             "Dale Steyn", "Allan Donald")

Years.Active <- c("1927-1949", "1988-2013", "1952-1974", "1971-1992",
                  "1973-1993", "1991-2007", "1974-1991", "1987-2007",
                  "1996-2014", "1999-2000", "1985-2003", "2006-present",
                  "2003-present", "2006-2018", "1992-2011", "2000-2015",
                  "1978-1994", "1971-1990", "1991-2008", "1998-2021",
                  "1992-2007", "1992-2012", "1985-2004", "1996-2011",
                  "1971-1987", "1997-2011", "1988-2000", "1997-2015",
                  "2004-2020", "1991-2003")

rk <- seq(1, 30)
```

```r
cricket <- data.frame(Players, Years.Active, rk)
colnames(cricket) <- c("Player", "Years Active", "Rank")
```

Tennis

```r
# taken from Stadium Talk
# https://www.stadiumtalk.com/s/greatest-mens-tennis-players-1e13282683434178

Players <- c("Roger Federer", "Rafael Nadal", "Novak Djokovic", "Pete Sampras",
             "Rod Laver", "Bjorn Borg", "John McEnroe", "Ivan Lendl",
             "Andre Agassi", "Jimmy Connors", "Don Budge", "Doris Becker",
             "Andy Murray", "Bill Tilden", "Roy Emerson", "Ken Rosewall",
             "Fred Perry", "Guillermo Vilas", "Jim Courier", "Ilie Nastase",
             "John Newcombe", "Arthur Ashe", "Rene Lacoste", "Mats Wilander",
             "Stefan Edberg", "Gustavo Kuerten", "Lleyton Hewitt", "Marat Safin",
             "Patrick Rafter", "Andy Roddick")

Years.Active <- c("1998-2022", "2001-present", "2003-present", "1988-2002",
                  "1956-1979", "1973-1983", "1978-1994", "1978-1994",
                  "1986-2006", "1972-1996", "1932-1955", "1984-1999",
                  "2005-present", "1912-1946", "1953-1983", "1956-1980",
                  "1929-1956", "1969-1992", "1988-2000", "1966-1985",
                  "1960-1981", "1959-1980", "1922-1932", "1981-1996",
                  "1983-1996", "1995-2008", "1998-2016", "1997-2009",
                  "1991-2002", "2000-2012")

rk <- seq(1, 30)

tennis <- data.frame(Players, Years.Active, rk)
colnames(tennis) <- c("Player", "Years Active", "Rank")
```

Basketball

```r
# taken from ESPN
# https://www.espn.com/nba/story/_/id/33297498/the-nba-75th-anniversary-team-ranked-where-76-basketball

Players <- c("Michael Jordan", "LeBron James", "Kareem Abdul-Jabbar", "Magic Johnson",
             "Wilt Chamberlain", "Bill Russell", "Larry Bird", "Tim Duncan",
             "Oscar Robertson", "Kobe Bryant", "Shaquille O'Neal", "Kevin Durant",
             "Hakeem Olajuwon", "Julius Erving", "Moses Malone", "Stephen Curry",
             "Dirk Nowitzki", "Giannis Antetokounmpo", "Jerry West", "Elgin Baylor",
             "Kevin Garnett", "Charles Barkley", "Karl Malone", "John Stockton",
             "David Robinson", "John Havlicek", "Isiah Thomas", "George Mikan",
             "Chris Paul", "Dwyane Wade", "Allen Iverson", "Scottie Pippen",
             "Kawhi Leonard", "Bob Cousy", "Bob Pettit", "Dominique Wilkins",
             "Steve Nash", "Rick Barry", "Kevin McHale", "Patrick Ewing",
             "Walt Frazier", "Gary Payton", "Jason Kidd", "Bill Walton",
             "Bob McAdoo", "Jerry Lucas", "Ray Allen", "Wes Unseld",
```

```
                "Nate Thurmond", "James Harden")


Years.Active <- c("1984-2003", "2003-present", "1969-1989", "1979-1991",
                  "1959-1973", "1956-1969", "1979-1992", "1997-2016",
                  "1960-1974", "1996-2016", "1992-2011", "2007-present",
                  "1984-2002", "1971-1987", "1974-1995", "2009-present",
                  "1998-2019", "2013-present", "1960-1974", "1958-1972",
                  "1995-2016", "1984-2000", "1985-2004", "1984-2003",
                  "1989-2003", "1962-1978", "1981-1994", "1948-1956",
                  "2005-present", "2003-2019", "1996-2010", "1987-2004",
                  "2011-present", "1950-1970", "1954-1965", "1982-1999",
                  "1996-2014", "1965-1980", "1980-1993", "1985-2002",
                  "1967-1979", "1990-2007", "1994-2013", "1974-1987",
                  "1972-1986", "1963-1974", "1996-2014", "1968-1981",
                  "1963-1977", "2009-present"
                  )

rk <- seq(1, 50)

basketball <- data.frame(Players, Years.Active)
colnames(basketball) <- c("Player", "Years Active")
```

Baseball

```
# taken from ESPN
# https://www.espn.com/mlb/story/_/id/33158613/top-100-mlb-players-all-nos-25-1
# https://www.espn.com/mlb/story/_/id/33145627/top-100-mlb-players-all-nos-50-26
Players <- c("Babe Ruth", "Willie Mays", "Hank Aaron", "Ty Cobb",
             "Ted Williams", "Lou Gehrig", "Mickey Mantle", "Barry Bonds",
             "Walter Johnson", "Stan Musial", "Pedro Martinez", "Honus Wagner",
             "Ken Griffey Jr.", "Greg Maddux", "Mike Trout", "Joe DiMaggio",
             "Roger Clemens", "Mike Schmidt", "Frank Robinson", "Roger Hornsby",
             "Cy Young", "Tom Seaver", "Rickey Henderson", "Randy Johnson",
             "Christy Mathewson", "Alex Rodriguez", "Roberto Clemente", "Derek Jeter",
             "Johnny Bench", "Albert Pujols", "Mariano Rivera", "Sandy Koufax",
             "Bob Gibson", "Pete Rose", "Josh Gibson", "Tris Speaker",
             "Joe Morgan", "Jackie Robinson", "Yogi Berra", "Jimmie Foxx",
             "Satchel Paige", "Nolan Ryan", "George Brett", "Tony Gwynn",
             "Wade Boggs", "Ichiro Suzuki", "Warren Spahn", "Nap Lajoie",
             "Frank Thomas", "Bob Feller")
Years.Active <- c("1914-1935", "1948, 1951-1952, 1954-1973", "1951, 1954-1976", "1905-1928",
                  "1939-1942, 1946-1960", "1923-1939", "1951-1968", "1986-2007",
                  "1907-1927", "1941-1963", "1992-2009", "1897-1917",
                  "1989-2010", "1986-2008", "2011-present", "1936-1951",
                  "1984-2007", "1972-1989", "1956-1976", "1915-1937",
                  "1890-1911", "1967-1986", "1979-2003", "1988-2009",
                  "1900-1916", "1994-2016", "1955-1972", "1995-2014",
                  "1967-1983", "2001-2022", "1995-2013", "1955-1966",
                  "1959-1975", "1963-1986", "1930-1946", "1907-1928",
                  "1963-1984", "1945, 1947-1956", "1946-1965", "1925-1945",
```

```
                        "1927-1949, 1951-1953, 1965", "1966-1993", "1973-1993", "1982-2001",
                        "1982-1999", "2001-2019", "1942, 1946-1965", "1896-1916",
                        "1990-2008", "1936-1941, 1945-1956")
rk <- seq(1, 50)
baseball <- data.frame(Players, Years.Active, rk)
colnames(baseball) <- c("Player", "Years Active", "Rank")
```

Football

```
# taken from ClutchPoints
# https://clutchpoints.com/updating-and-ranking-the-50-greatest-nfl-players-of-all-time

Players <- c("Tom Brady", "Jerry Rice", "Lawrence Taylor", "Jim Brown",
             "Joe Montana", "Walter Payton", "Reggie White", "Johnny Unitas",
             "Peyton Manning", "Emmitt Smith", "Joe Green", "Ronnie Lott",
             "John Elway", "Dick Butkus", "Ray Lewis", "Barry Sanders",
             "Deion Sanders", "Dan Marino", "Anthony Munoz", "Deacon Jones",
             "Otto Graham", "Gale Sayers", "Brett Favre", "Randy Moss",
             "Jack Lambert", "Alan Page", "Bruce Smith", "Don Hutson",
             "Drew Brees", "Ed Reed", "Sammy Baugh", "Bob Lilly",
             "Dick Lane", "Aaron Rodgers", "Gino Marchetti", "Tony Gonzalez",
             "Rod Woodson", "Mel Blount", "Eric Dickerson", "Ray Nitschke",
             "John Hannah", "Mike Singletary", "Early Campbell", "Jim Thorpe",
             "Roger Staubach", "Chuck Bednarik", "OJ Simpson", "Forrest Gregg",
             "Steve Young", "Terrell Owens")

Years.Active <- c("2000-present", "1985-2004", "1981-1993", "1957-1965",
                  "1979-1994", "1975-1987", "1985-2000", "1955-1973",
                  "1998-2015", "1990-2004", "1969-1981", "1981-1994",
                  "1983-1998", "1965-1973", "1996-2012", "1989-1998",
                  "1989-2000, 2004-2005", "1983-1999", "1980-1992", "1961-1974",
                  "1946-1955", "1965-1971", "1991-2010", "1998-2012",
                  "1974-1984", "1967-1981", "1985-2003", "1935-1945",
                  "2001-2020", "2002-2013", "1937-1952", "1961-1974",
                  "1952-1965", "2005-present", "1952-1966", "1997-2013",
                  "1987-2003", "1970-1983", "1983-1993", "1958-1972",
                  "1973-1985", "1981-1992", "1978-1985", "1920-1928",
                  "1969-1979", "1949-1962", "1969-1979", "1956-1971",
                  "1985-1999", "1996-2010")

rk <- seq(1, 50)

football <- data.frame(Players, Years.Active, rk)
colnames(football) <- c("Players", "Years Active", "Rank")
```

Ice Hockey
```

```r
# players taken from Bleacher Report
# https://bleacherreport.com/articles/630824-nhl-power-rankings-the-50-greatest-players-in-nhl-and-hock
# years active taken from Wikipedia

Players <- c("Wayne Gretzky", "Mario Lemieux", "Gordie Howe", "Maurice Richard",
             "Bobby Orr", "Jean Beliveau", "Joe Malone", "Bobby Hull",
             "Doug Harvey", "Mike Bossy", "Jaques Plante", "Mark Messier",
             "Eddie Shore", "Terry Sawchuck", "Denis Potvin", "Guy Lafleur",
             "Ray Bourque", "Phil Esposito", "Howie Morenz", "Jaromir Jagr",
             "Glenn Hall", "Martin Brodeur", "Stan Mikita", "Steve Yzerman",
             "Frank Mahovlich", "Larry Robinson", "Dominik Hasek", "Niklas Lidstrom",
             "Joe Sakic", "Henri Richard", "Bryan Trottier", "Dickie Moore",
             "Newsy Lalonde", "Paul Coffey", "Syl Apps", "Patrick Roy",
             "Brendan Shanahan", "Marcel Dionne", "Charlie Conacher", "Brett Hull",
             "Bill Durnan", "Johnny Bucyk", "Sidney Crosby", "Dit Clapper",
             "Mike Gartner", "Chris Chelios", "Bobby Clarke", "Jari Kurri",
             "Alexander Ovechkin", "Gilbert Perreault")

Years.Active <- c("1978-1999", "1984-1997, 2000-2006", "1946-1971, 1973-1980", "1942-1960",
                  "1966-1978", "1950-1971", "1910-1924", "1957-1980",
                  "1945-1969", "1977-1987", "1947-1965, 1968-1973, 1974-1975", "1978-2004",
                  "1924-1943", "1949-1970", "1973-1988", "1971-1985, 1988-1991",
                  "1979-2001", "1963-1981", "1923-1937", "1988-present",
                  "1951-1971", "1991-2015", "1958-1980", "1983-2006",
                  "1957-1978", "1973-1992", "1980-2011", "1987-2012",
                  "1988-2009", "1955-1975", "1975-1994", "1951-1968",
                  "1904-1927", "1980-2001", "1936-1948", "1984-2003",
                  "1987-2009", "1971-1989", "1929-1941", "1986-2005",
                  "1943-1950", "1955-1978", "2005-present", "1927-1947",
                  "1978-1998", "1984-2010", "1969-1984", "1977-1998",
                  "2001-present", "1970-1986")

rk <- seq(1, 50)

hockey <- data.frame(Players, Years.Active, rk)
colnames(hockey) <- c("Players", "Years Active", "Rank")
```

Golf

```r
# taken from Athlon Sports
# https://athlonsports.com/golf/greatest-golfers-all-time

# Players <- c("Tiger Woods", "Jack Nicklaus", "Sam Snead", "Arnold Palmer",
#             "Ben Hogan", "Bobby Jones", "Tom Watson", "Gary Player",
#             "Gene Sarazen", "Phil Mickelson", "Seve Ballesteros", "Byron Nelson",
#             "Lee Trevino", "Nick Faldo", "Walter Hagen", "Ernie Els",
#             "Billy Casper", "Vijay Singh", "Rory McIlroy", "Greg Norman")
#
# Years.Active <- c("1996-present", "1962-1986", )
#
```

```
# rk <- seq(1, 20)
#
# golf <- data.frame(Players, Years.Active, rk)
# colnames(golf) <- c("Player", "Years Active", "Rank")
```

Visualization

```r
cricket.revised <- cricket
for (i in 1:nrow(cricket.revised)){
  cricket.revised[i, "Years Active"] <- str_replace(cricket.revised[i, "Years Active"], "present", "2022
}

tennis.revised <- tennis
for (i in 1:nrow(tennis.revised)){
  tennis.revised[i, "Years Active"] <- str_replace(tennis.revised[i, "Years Active"], "present", "2022")
}

basketball.revised <- basketball
for (i in 1:nrow(basketball.revised)){
  basketball.revised[i, "Years Active"] <- str_replace(basketball.revised[i, "Years Active"], "present"
}

football.revised <- football
for (i in 1:nrow(football.revised)){
  football.revised[i, "Years Active"] <- str_replace(football.revised[i, "Years Active"], "present", "2
}

hockey.revised <- hockey
for (i in 1:nrow(football.revised)){
  hockey.revised[i, "Years Active"] <- str_replace(hockey.revised[i, "Years Active"], "present", "2022")
}
```

```r
df = data.frame(Player=character(),
                Year=integer(),
                Sport = character())
rownum = 0

for (i in 1:nrow(soccer)){
  seq = seq(strtoi(substr(soccer[i, 2], 1, 4)),
            strtoi(substr(soccer[i,2], 6, 9)))
  for (x in 1:length(seq)){
    rownum = rownum + 1
    df[rownum, ] = c(soccer[i,1], seq[x], "Soccer")
  }
}

for (i in 1:nrow(cricket.revised)){
  seq = seq(strtoi(substr(cricket.revised[i, 2], 1, 4)),
            strtoi(substr(cricket.revised[i,2], 6, 9)))
  for (x in 1:length(seq)){
    rownum = rownum + 1
    df[rownum, ] = c(cricket.revised[i,1], seq[x], "Cricket")
  }
}

for (i in 1:nrow(tennis.revised)){
  seq = seq(strtoi(substr(tennis.revised[i, 2], 1, 4)),
            strtoi(substr(tennis.revised[i,2], 6, 9)))
```

```r
  for (x in 1:length(seq)){
    rownum = rownum + 1
    df[rownum, ] = c(tennis.revised[i,1], seq[x], "Tennis")
  }
}

for (i in 1:nrow(basketball.revised)){
  seq = seq(strtoi(substr(basketball.revised[i, 2], 1, 4)),
            strtoi(substr(basketball.revised[i,2], 6, 9)))
  for (x in 1:length(seq)){
    rownum = rownum + 1
    df[rownum, ] = c(basketball.revised[i,1], seq[x], "Basketball")
  }
}

for (i in 1:nrow(football.revised)){
  seq = seq(strtoi(substr(football.revised[i, 2], 1, 4)),
            strtoi(substr(football.revised[i,2], 6, 9)))
  for (x in 1:length(seq)){
    rownum = rownum + 1
    df[rownum, ] = c(football.revised[i,1], seq[x], "Football")
  }
}

for (i in 1:nrow(hockey.revised)){
  seq = seq(strtoi(substr(hockey.revised[i, 2], 1, 4)),
            strtoi(substr(hockey.revised[i,2], 6, 9)))
  for (x in 1:length(seq)){
    rownum = rownum + 1
    df[rownum, ] = c(hockey.revised[i,1], seq[x], "Hockey")
  }
}

f1.data <- df %>%
  group_by(Year, Sport) %>%
  count()

f1.data
```

```
## # A tibble: 570 x 3
## # Groups:   Year, Sport [570]
##    Year  Sport        n
##    <chr> <chr>    <int>
##  1 1904  Hockey       1
##  2 1905  Hockey       1
##  3 1906  Hockey       1
##  4 1907  Hockey       1
##  5 1908  Hockey       1
##  6 1909  Hockey       1
##  7 1910  Hockey       2
##  8 1911  Hockey       2
##  9 1912  Hockey       2
## 10 1912  Tennis       1
```

```
## # ... with 560 more rows

f1 <- ggplot(data = f1.data,
       aes(x = as.numeric(Year), y = n, group = Sport, color = Sport)) +
  geom_line() +
  labs(x = "Year",
       y = "Number of All-Time Greats",
       title = "Number of All-Time Greats by Year",
       caption = "Figure 1") +
  scale_x_continuous(limits = c(1900, 2025)) +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  theme(plot.title=element_text(family = "serif", face = "bold", hjust = 0.5, size = 12))
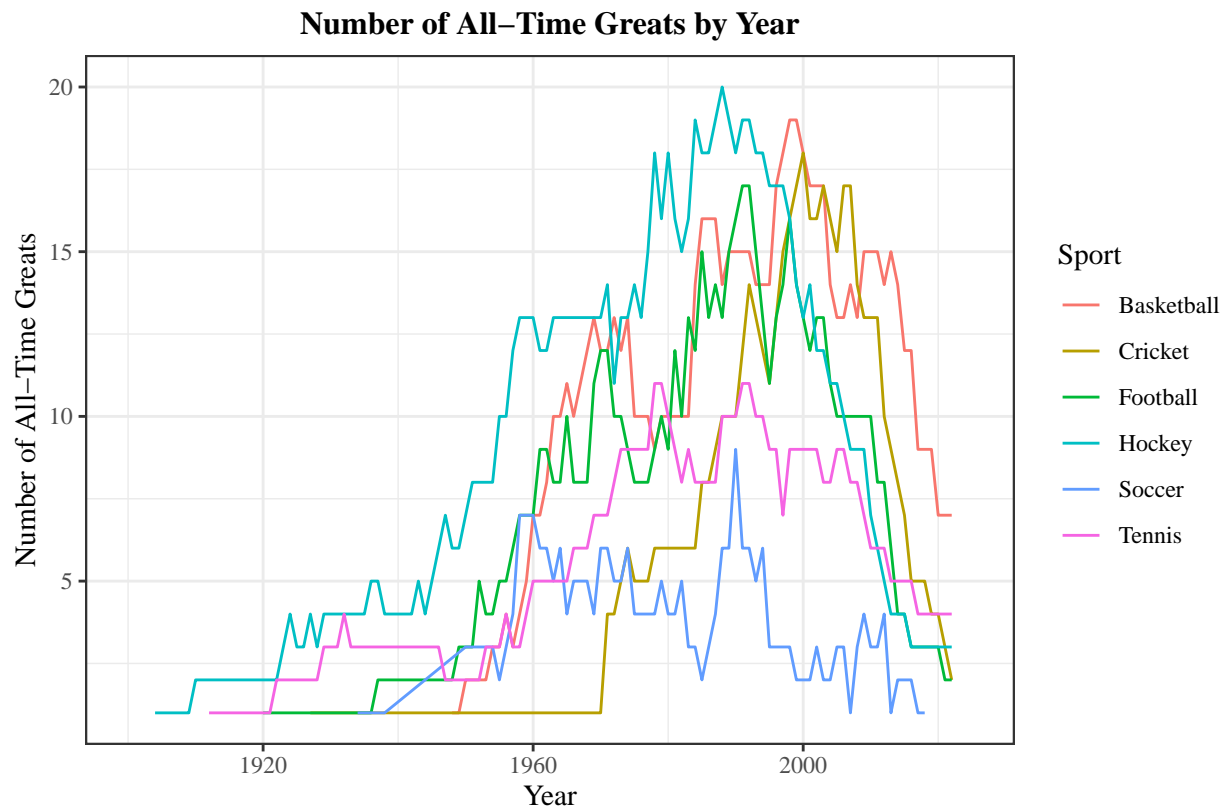
f1
```


Figure 1