

# Dataset Creation

Matthew Murray

2023-02-09

```
# 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
```

```
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
```

```
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

library(r2r)

## Warning: package 'r2r' was built under R version 4.0.5

library(purrr)
library(base)
library(stringr)
```

## Functions

```
create.df <- function(Players, Year) {  
  df <- data.frame(Players, Year, seq(1, length(Players)))  
  colnames(df) = c("Player", "Years", "Rank")  
  return(df)  
}
```

*# modifying strings so that 'present' is changed is 2022/2023*

```
presentto2022 <- function(df) {  
  for (i in 1:nrow(df)){  
    df[i, "Years"] <- str_replace(df[i, "Years"], "present", "2022")  
  }  
  return(df)  
}
```

*# changing the intervals from a string to actual series of integers with a row for each year  
# doing that for each sport and placing those rows into the new dataframe, df*

```
rowforyear <- function(big.df, sport.df, string){  
  rownum = 0  
  for (i in 1:nrow(sport.df)){  
    seq = seq(strtoi(substr(sport.df[i, 2], 1, 4)),  
              strtoi(substr(sport.df[i, 2], 6, 9)))  
    for (x in 1:length(seq)){  
      rownum = rownum + 1  
      big.df[rownum, ] = c(sport.df[i, 1], seq[x], string, sport.df[i, 3], sport.df[i, 2])  
    }  
  }  
  return(big.df)  
}
```

*# filter df for sport  
# take the max oldest year for each player  
# create empty dataframe for goats  
# if the player has the best ranking at that point in his career, put him into goats dataframe*

```
sportgoats <- function(df, string){  
  filter <- df[df$Sport == string, ]  
  filter <- filter %>% group_by(Player) %>% summarise(Year = max(Year),  
                                                       Rank = max(Rank),  
                                                       Years = unique(Interval))  
  
  goats <- data.frame(matrix(nrow = 0, ncol = 4))  
  
  for (i in 1:nrow(filter)){  
    filter.2 <- filter[filter$Year <= as.numeric(filter[i, 2])[[1]], ]  
    if(as.numeric(min(filter.2$Rank)) == as.numeric(filter[i, 3][[1,1]])){  
      goats <- rbind(goats, filter[i,])  
    }  
  }  
  return(goats)  
}
```

# Data Imputation

## 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")

soccer <- create.df(Players, Peak)
```

## 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")

cricket <- create.df(Players, Years.Active)
```

## Tennis

```
# 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")

tennis <- create.df(Players, Years.Active)
```

## Tennis 2

```
# taken from https://www.landoftennis.com/statistics_men/grand_slam_most_titles.htm
# Ranking Players by the numbers of majors they won
# including players who have won at least 7 majors
# if I am dealing with a very old player whose start and end years are ambiguous, I use the first and last

Players <- c("Rafael Nadal", "Novak Djokovic", "Roger Federer", "Pete Sampras",
  "Roy Emerson", "Bjorn Borg", "Rod Laver", "Bill Tilden",
  "Andre Agassi", "Jimmy Connors", "Ivan Lendl", "Fred Perry",
  "Ken Rosewall", "Henri Cochet", "Rene Lacoste", "William Larned",
  "John McEnroe", "John Newcombe", "William Renshaw", "Richard Sears",
  "Mats Wilander")
```

```

Years.Active <- c("2001-present", "2003-present", "1998-2022", "1988-2002",
                 "1953-1983", "1973-1983", "1956-1979", "1912-1946",
                 "1986-2006", "1972-1996", "1978-1994", "1929-1956",
                 "1956-1980", "1926-1933", "1922-1932", "1900-1911",
                 "1978-1994", "1960-1981", "1891-1890", "1881-1887",
                 "1981-1996")

# Majors <- c(22, 21, 20, 14,
#            12, 11, 11, 10,
#            8, 8, 8, 8,
#            8, 7, 7, 7,
#            7, 7, 7, 7,
#            7)

tennis.2 <- create.df(Players, Years.Active)

# making Djokovic's rank #1 because he and Nadal are TIED for the most majors
tennis.2[2, 3] = 1

```

## Basketball

```

# 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"
      )

basketball <- create.df(Players, Years.Active)

```

## Basketball 2

```

# 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("1984-1993, 1995-1998, 2001-2003", "2003-present", "1969-1989", "1979-1991, 1996",
  "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", "1950-1963, 1969-1970", "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"
)

basketball.2 <- create.df(Players, Years.Active)

```

## Basketball 3

```

# taken from Basketball Reference
# Value over Replacement Player
# https://www.basketball-reference.com/leaders/vorp_career.html

```

```

Players <- c("Lebron James", "Michael Jordan", "John Stockton", "Karl Malone",
             "Kevin Garnett", "Chris Paul", "Tim Duncan", "Kareem Abdul-Jabbar",
             "Dirk Nowitzki", "David Robinson", "Charles Barkley", "Kobe Bryant",
             "Julius Erving", "Magic Johnson", "Kevin Durant", "Larry Bird",
             "Shaquille O'Neal", "James Harden", "Hakeem Olajuwon", "Jason Kidd"
             )

Years.Active <- c("2003-present", "1984-2003", "1984-2003", "1985-2004",
                  "1995-2016", "2005-present", "1997-2016", "1969-1989",
                  "1998-2019", "1989-2003", "1984-2000", "1996-2016",
                  "1971-1987", "1979-1991", "2007-present", "1979-1992",
                  "1992-2011", "2009-present", "1984-2002", "1994-2013"
                  )

basketball.3 <- create.df(Players, 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")

```



```
baseball <- create.df(Players, Years.Active)
```

## Football

```
# taken from ClutchPoints
# https://clutchpoints.com/updated-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")

football <- create.df(Players, Years.Active)
```

```
# taken from Sporting News
# list with players who have won more than 1 NFL MVP
# https://www.sportingnews.com/us/nfl/news/nfl-mvp-winners-list-most-repeat/e17cir63vdd41u4gb0s8ahoby
```

```
Players <- c("Peyton Manning", "Aaron Rodgers", "Tom Brady", "Brett Favre",
  "Jim Brown", "Johnny Unitas", "Kurt Warner", "Steve Young",
  "Joe Montana")

Years.Active <- c("1998-2015", "2005-present", "2000-present", "1991-2010",
  "1957-1965", "1955-1973", "1998-2009", "1985-1999",
  "1979-1994")

football.2 <- create.df(Players, Years.Active)
```

```
# changing the rank column account for players with the same number of MVP's
football.2[3:6, 3] = 3
football.2[7:9, 3] = 4
```

## Ice Hockey

```
# 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")

hockey <- create.df(Players, Years.Active)
```

## Golf

```
# including players who have won at least 5 majors
# taken from https://en.wikipedia.org/wiki/List_of_men%27s_major_championships_winning_golfers
```

```
Players <- c("Jack Nicklaus", "Tiger Woods", "Walter Hagen", "Ben Hogan",
  "Gary Player", "Tom Watson", "Harry Vardon", "Bobby Jones",
  "Gene Sarazen", "Sam Snead", "Arnold Palmer", "Lee Trevino",
  "Nick Faldo", "Phil Mickelson", "James Braid", "John Henry Taylor",
```

```

      "Byron Nelson", "Peter Thomson", "Seve Ballesteros")

Winning.Span <- c("1962-1986", "1997-2019", "1914-1929", "1946-1953",
                  "1959-1978", "1975-1983", "1896-1914", "1923-1930",
                  "1922-1935", "1942-1954", "1958-1964", "1968-1984",
                  "1987-1996", "2004-2021", "1901-1910", "1894-1913",
                  "1937-1945", "1954-1965", "1979-1988")

golf <- create.df(Players, Winning.Span)

```

## Chess

```

# players who had a peak ELO rating of 2775 +
# taken from https://en.wikipedia.org/wiki/List_of_chess_players_by_peak_FIDE_rating#2800+

Players <- c("Magnus Carlsen", "Garry Kasparov", "Fabiano Caruana", "Levon Aronian",
             "Wesley So", "Shakhriyar Mamedyarov", "Maxime Vachier-Lagrave", "Viswanathan Anand",
             "Vladimir Kramnik", "Veselin Topalov", "Hikaru Nakamura", "Ding Liren",
             "Alexander Grischuk", "Alireza Firouzja", "Anish Giri", "Teimour Radjabov",
             "Ian Nepomniachtchi", "Alexander Morozevich", "Sergey Karjakin", "Vassily Ivanchuk",
             "Bobby Fischer", "Anatoly Karpov", "Boris Gelfand", "Richard Rapport")

Peak.Rating <- c(2881, 2851, 2844, 2830,
                 2822, 2820, 2819, 2817,
                 2817, 2816, 2816, 2816,
                 2810, 2804, 2798, 2793,
                 2793, 2788, 2788, 2787,
                 2785, 2780, 2777, 2776)

```

## Data Wrangling

```
df.list <- list(soccer, cricket, tennis, tennis.2, basketball, basketball.3, football, football.2, hock  
df.list <- lapply(df.list, presentto2022)
```

```
# new, empty dataframe  
df.new = data.frame(Player = character(),  
                    Year = integer(),  
                    Sport = character(),  
                    Rank = integer(),  
                    Interval = character())  
  
rownum = 0  
  
# names of sports and iterator for the list  
strings = list("Soccer", "Cricket", "Tennis", "Tennis 2", "Basketball", "Basketball 3", "Football", "Fo  
iterator = 0  
  
new.list <- list()  
  
for (df in df.list){  
  iterator = iterator + 1  
  new.list[[iterator]] = rowforyear(df.new, df.list[[iterator]], strings[iterator])  
}  
  
df.new <- bind_rows(new.list, .id = "column_label")  
df.new$column_label <- NULL
```

```
goats.list <- replicate(length(strings), data.frame())  
for (i in 1:length(strings)) {  
  goats.list[[i]] = sportgoats(df.new, strings[i])  
  goats.list[[i]]$Year <- NULL  
  goats.list[[i]] <- goats.list[[i]][, c("Player", "Years", "Rank")]  
}
```

```
# new, empty dataframe  
df.goats = data.frame(Player = character(),  
                    Year = integer(),  
                    Sport = character(),  
                    Rank = integer(),  
                    Interval = integer())  
  
rownum = 0  
  
# names of sports and iterator for the list  
strings = list("Soccer", "Cricket", "Tennis", "Tennis 2", "Basketball", "Basketball 3", "Football", "Fo  
iterator = 0  
  
goat.list <- list()  
  
for (df in goats.list){  
  iterator = iterator + 1  
  goat.list[[iterator]] = rowforyear(df.goats, goats.list[[iterator]], strings[iterator])  
}
```

```
# combine all the dataframes from goat.list into one big dataframe
df.goats <- bind_rows(goat.list, .id = "column_label")

# eliminate superfluous column
df.goats$column_label <- NULL

# combine same sports
df.goats$Sport <- str_replace(df.goats$Sport, "Tennis 2", "Tennis")
df.goats$Sport <- str_replace(df.goats$Sport, "Basketball 3", "Basketball")
df.goats$Sport <- str_replace(df.goats$Sport, "Football 2", "Football")

# eliminate identical rows
df.goats <- distinct(df.goats)
```

## Visualization

```
# group by and count number of players in each sport for each year
```

```
df.goats.subset <- df.goats %>%  
  group_by(Year, Sport) %>%  
  summarize(N = n()) %>%  
  ungroup() %>%  
  complete(Year, Sport,  
           fill = list(N = 0))
```

## 'summarise()' has grouped output by 'Year'. You can override using the '.groups' argument.

```
f1.data.dummy <- data.frame(matrix(nrow = 0, ncol = length(unique(df.goats.subset$Year))))  
colnames(f1.data.dummy) <- unique(df.goats.subset$Year)
```

```
for (i in 1:nrow(df.goats.subset)){  
  for (j in 1:length(unique(df.goats.subset$Year))){  
    if (df.goats.subset[i, 1] == unique(df.goats.subset$Year[j])){  
      f1.data.dummy[i, j] = 1  
    }  
    else{  
      f1.data.dummy[i, j] = 0  
    }  
    f1.data.dummy[i, "Sport"] = df.goats.subset[i, 2]  
  }  
}
```

```
# write visualization data to csv
```

```
write.csv(df.goats.subset, "data.csv", row.names=FALSE)
```