Baseball_2

Thomas Young

2025-03-21

Question: Which players started the most All-Star games at each position over a specified time period (e.g., 1990-2000)?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(httr)
library(jsonlite)
```

```
## Warning: package 'jsonlite' was built under R version 4.4.1
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(baseballr)
```

```
## Warning: package 'baseballr' was built under R version 4.4.1
```

```
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 4.4.1
```

```
##
## Attaching package: 'RCurl'
##
## The following object is masked from 'package:tidyr':
##
##     complete
```

```r
library(Lahman)
```

```
## Warning: package 'Lahman' was built under R version 4.4.1
```

```r
url <- "https://www.kaggle.com/api/v1/datasets/download/open-source-sports/baseball-databank/AllstarFull

kaggle_username <- "your_kaggle_username" # for replicability
kaggle_key <- "your_kaggle_key" # for replicability
headers <- add_headers(Authorization = paste("username", kaggle_username, "key", kaggle_key, sep = ":"))

# Download directly
response <- GET(url, headers)
allstar_data <- read.csv(text = content(response, "text"))
```

```
## No encoding supplied: defaulting to UTF-8.
```

```r
# View data
glimpse(allstar_data)
```

```
## Rows: 5,069
## Columns: 8
## $ playerID    <chr> "gomezle01", "ferreri01", "gehrilo01", "gehrich01", "dykes~
## $ yearID      <int> 1933, 1933, 1933, 1933, 1933, 1933, 1933, 1933, 1933, 1933~
## $ gameNum     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ gameID      <chr> "ALS193307060", "ALS193307060", "ALS193307060", "ALS193307~
## $ teamID      <chr> "NYA", "BOS", "NYA", "DET", "CHA", "WS1", "NYA", "CHA", "N~
## $ lgID        <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL"~
## $ GP          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1~
## $ startingPos <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```r
# get people data
data(People)
```

```r
# Filter Data
start_year <- 1990
end_year <- 2000
filtered_df <- allstar_data |>  filter(yearID >= start_year & yearID <= end_year)
```

```r
# Group players
starts_by_player_position <- filtered_df |>
  group_by(playerID, startingPos) |>
  summarize(total_starts = sum(GP)) |>
  ungroup()
```

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

```r
# Top Starters
top_starters <- starts_by_player_position |>
  group_by(startingPos) |>
```

```
  filter(total_starts == max(total_starts)) |>
  ungroup()

top_starters
```

```
## # A tibble: 12 x 3
##    playerID  startingPos total_starts
##    <chr>           <int>        <int>
##  1 alomaro01           4            9
##  2 boggswa01           5            7
##  3 bondsba01           7            7
##  4 griffke02           8            8
##  5 gwynnto01           9            3
##  6 johnsra05           1            3
##  7 maddugr01           1            3
##  8 martied01          10            2
##  9 mcgwima01           3            4
## 10 ripkeca01           6            7
## 11 rodriiv01           2            8
## 12 sheffga01          NA            5
```

```
# Join top starts
people_subset <- People |>  select(playerID, nameFirst, nameLast)

top_starters_with_names <- top_starters |>
  right_join(people_subset, by = "playerID") |>
  mutate(playerName = paste(nameFirst, nameLast)) |>
  select(playerID, playerName, everything()) |>
  select(-nameFirst, -nameLast) # Remove unnecessary columns

# Print the updated table
print(top_starters_with_names)
```

```
## # A tibble: 21,010 x 4
##    playerID  playerName     startingPos total_starts
##    <chr>     <chr>                <int>        <int>
##  1 alomaro01 Roberto Alomar           4            9
##  2 boggswa01 Wade Boggs               5            7
##  3 bondsba01 Barry Bonds              7            7
##  4 griffke02 Ken Griffey              8            8
##  5 gwynnto01 Tony Gwynn               9            3
##  6 johnsra05 Randy Johnson            1            3
##  7 maddugr01 Greg Maddux              1            3
##  8 martied01 Edgar Martinez          10            2
##  9 mcgwima01 Mark McGwire             3            4
## 10 ripkeca01 Cal Ripken               6            7
## # i 21,000 more rows
```