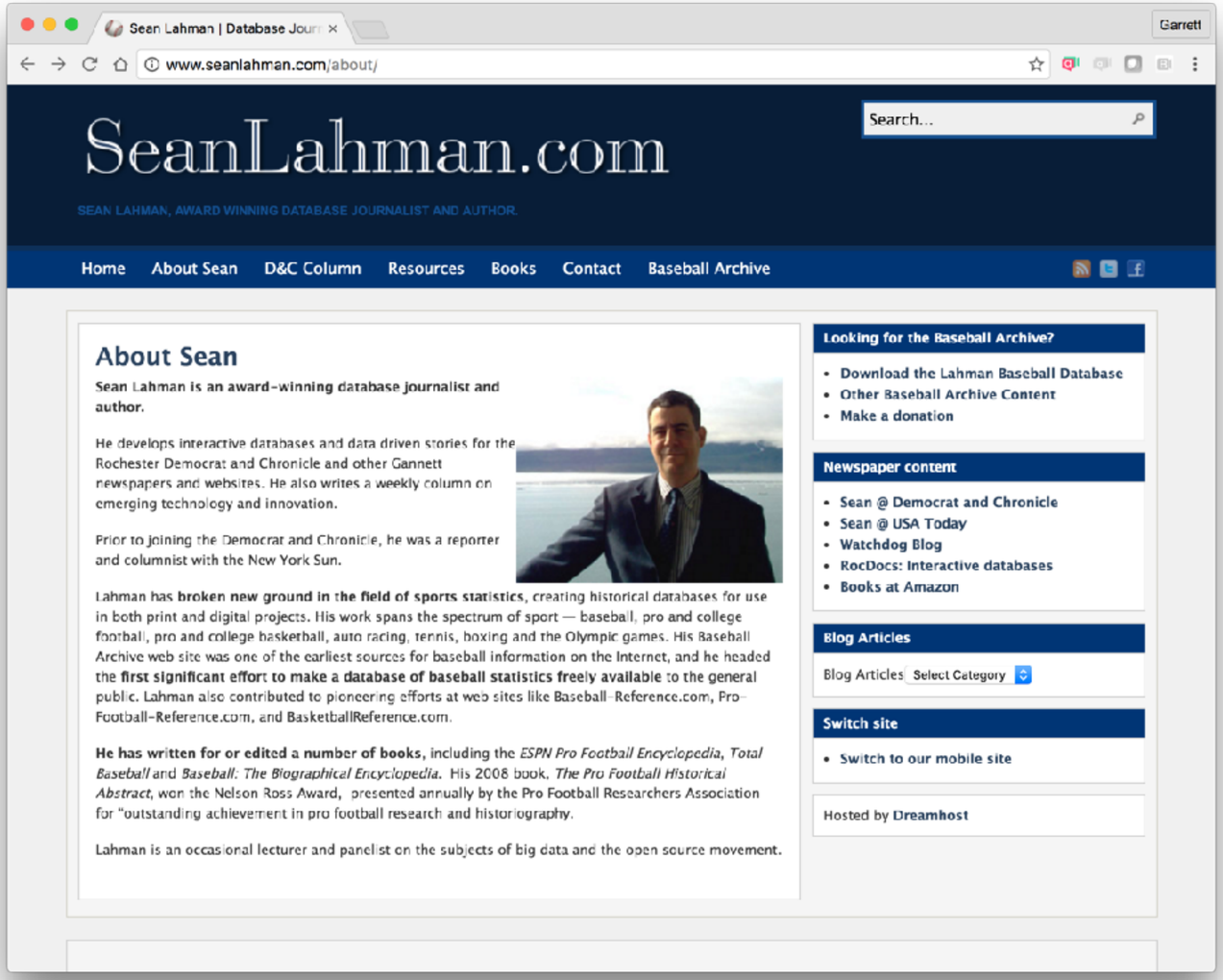
JOINING DATA IN R WITH DPLYR

Lahman's Baseball Database

Lahman R package

- `library(Lahman)`
- Sean Lahman's Baseball Database
- 26 tables

Lahman R package



```
> library(Lahman)
> as.tibble(Teams)
# A tibble: 2,775 × 48
   yearID   lgID teamID franchID divID  Rank     G Ghome     W
    <int> <fctr> <fctr>   <fctr> <chr> <int> <int> <int> <int>
1    1871     NA    BS1      BNA  <NA>     3    31    NA    20
2    1871     NA    CH1      CNA  <NA>     2    28    NA    19
3    1871     NA    CL1      CFC  <NA>     8    29    NA    10
4    1871     NA    FW1      KEK  <NA>     7    19    NA     7
5    1871     NA    NY2      NNA  <NA>     5    33    NA    16
6    1871     NA    PH1      PNA  <NA>     1    28    NA    21
7    1871     NA    RC1      ROK  <NA>     9    25    NA     4
8    1871     NA    TRO      TRO  <NA>     6    29    NA    13
9    1871     NA    WS3      OLY  <NA>     4    32    NA    15
10   1872     NA    BL1      BLC  <NA>     2    58    NA    35
# ... with 2,765 more rows, and 39 more variables:
```

```
> library(Lahman)
> as.tibble(Batting)
# A tibble: 99,846 × 22
    playerID yearID stint teamID   lgID      G    AB      R
       <chr>  <int> <int> <fctr> <fctr> <int> <int> <int>
1  abercda01   1871     1    TRO     NA     1     4     0
2   addybo01   1871     1    RC1     NA    25   118    30
3  allisar01   1871     1    CL1     NA    29   137    28
4  allisdo01   1871     1    WS3     NA    27   133    28
5  ansonca01   1871     1    RC1     NA    25   120    29
6  armstbo01   1871     1    FW1     NA    12    49     9
7  barkeal01   1871     1    RC1     NA     1     4     0
8  barnero01   1871     1    BS1     NA    31   157    66
9  barrebi01   1871     1    FW1     NA     1     5     1
10 barrofr01   1871     1    BS1     NA    18    86    13
# ... with 99,836 more rows, and 14 more variables:
```

```
> library(Lahman)
> as.tibble(Salaries)
# A tibble: 24,758 × 5
   yearID teamID   lgID  playerID salary
    <int> <fctr> <fctr>     <chr>  <int>
1    1985    ATL     NL barkele01 870000
2    1985    ATL     NL bedrost01 550000
3    1985    ATL     NL benedbr01 545000
4    1985    ATL     NL  campri01 633333
5    1985    ATL     NL ceronri01 625000
6    1985    ATL     NL chambch01 800000
7    1985    ATL     NL dedmoje01 150000
8    1985    ATL     NL forstte01 483333
9    1985    ATL     NL garbege01 772000
10   1985    ATL     NL harpete01 250000
# ... with 24,748 more rows
```

JOINING DATA IN R WITH DPLYR

Let’s practice!

JOINING DATA IN R WITH DPLYR

Salaries

```
> library(Lahman)
> as.tibble(Salaries)
# A tibble: 24,758 × 5
   yearID teamID   lgID  playerID salary
    <int> <fctr> <fctr>     <chr>  <int>
1    1985    ATL     NL barkele01 870000
2    1985    ATL     NL bedrost01 550000
3    1985    ATL     NL benedbr01 545000
4    1985    ATL     NL  campri01 633333
5    1985    ATL     NL ceronri01 625000
6    1985    ATL     NL chambch01 800000
7    1985    ATL     NL dedmoje01 150000
8    1985    ATL     NL forstte01 483333
9    1985    ATL     NL garbege01 772000
10   1985    ATL     NL harpete01 250000
# ... with 24,748 more rows
```

distinct()

```
> multinames
   name surname
1  Mick  Jagger
2  John  Lennon
3  John  Lennon
4 Ringo   Starr
```

```
> distinct(multinames)
   name surname
1  Mick  Jagger
2  John  Lennon
```

count()

```
> names
  name    band
1 Mick  Stones
2 John Beatles
3 Paul Beatles
```

```
> count(names)
# A tibble: 1 × 1
      n
  <int>
1     3
> count(names, vars = band)
# A tibble: 2 × 2
     vars     n
    <chr> <int>
1 Beatles     2
2  Stones     1
```

```
g_names <- group_by(names, band)
count(g_names)
```

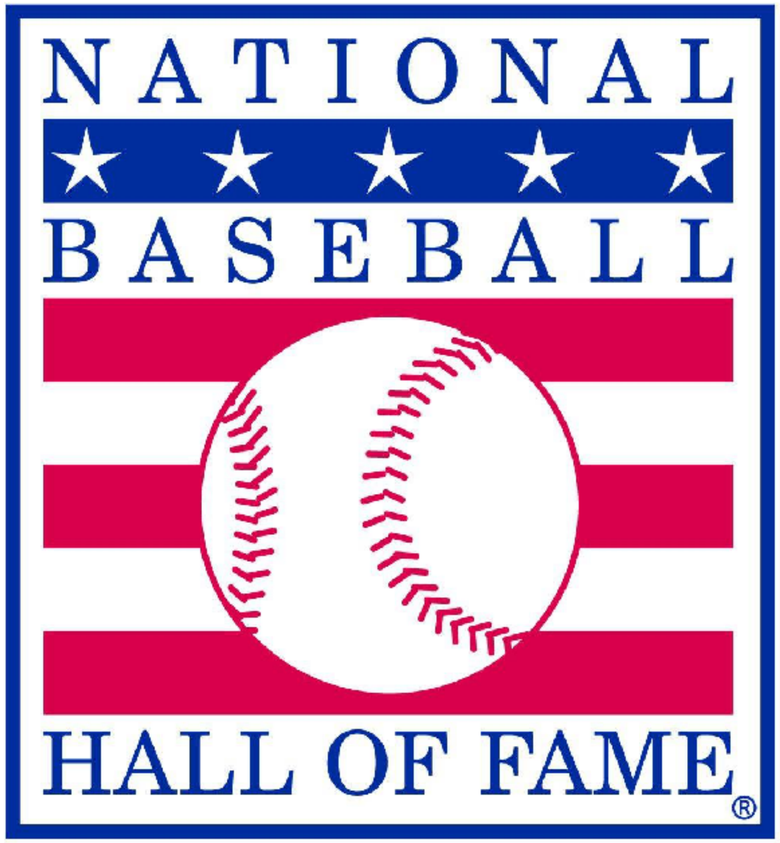
JOINING DATA IN R WITH DPLYR

Let’s practice!

JOINING DATA IN R WITH DPLYR

Hall of Fame

```
> library(Lahman)
> as.tibble(HallOfFame)
# A tibble: 4,088 × 9
    playerID yearID votedBy ballots needed votes inducted
       <chr>  <int>   <chr>   <int>  <int> <int>   <fctr>
1   cobbty01   1936   BBWAA     226    170   222        Y
2   ruthba01   1936   BBWAA     226    170   215        Y
3  wagneho01   1936   BBWAA     226    170   215        Y
4  mathech01   1936   BBWAA     226    170   205        Y
5  johnswa01   1936   BBWAA     226    170   189        Y
6  lajoina01   1936   BBWAA     226    170   146        N
7  speaktr01   1936   BBWAA     226    170   133        N
8  youngcy01   1936   BBWAA     226    170   111        N
9  hornsro01   1936   BBWAA     226    170   105        N
10 cochrmi01   1936   BBWAA     226    170    80        N
# ... with 4,078 more rows, and 2 more variables:
#   category <fctr>, needed_note <chr>
```


NATIONAL
BASEBALL
HALL OF FAME

JOINING DATA IN R WITH DPLYR

Let’s practice!

JOINING DATA IN R WITH DPLYR

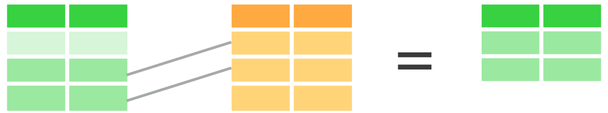
Congratulations!

Mutating joins

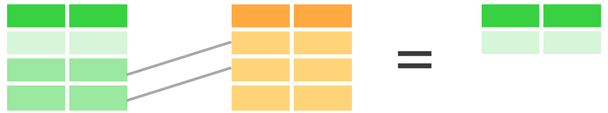
- `left_join()` 
- `right_join()` 
- `inner_join()` 
- `full_join()` 

Filtering joins

- `semi_join()`

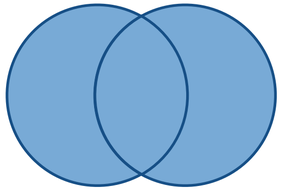


- `anti_join()`

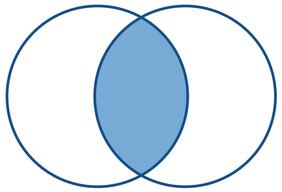


Set operations

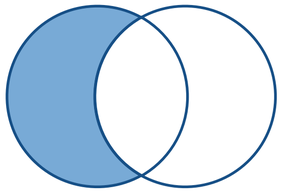
- `union()`



- `intersect()`

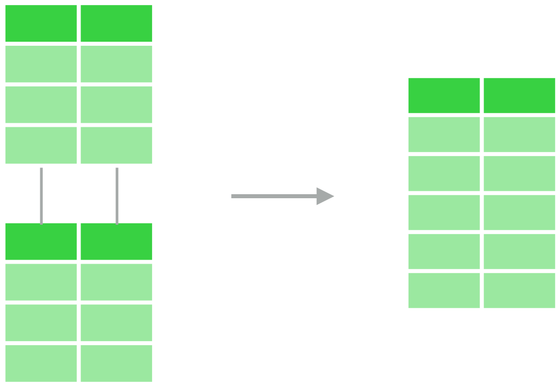


- `setdiff()`



Binds

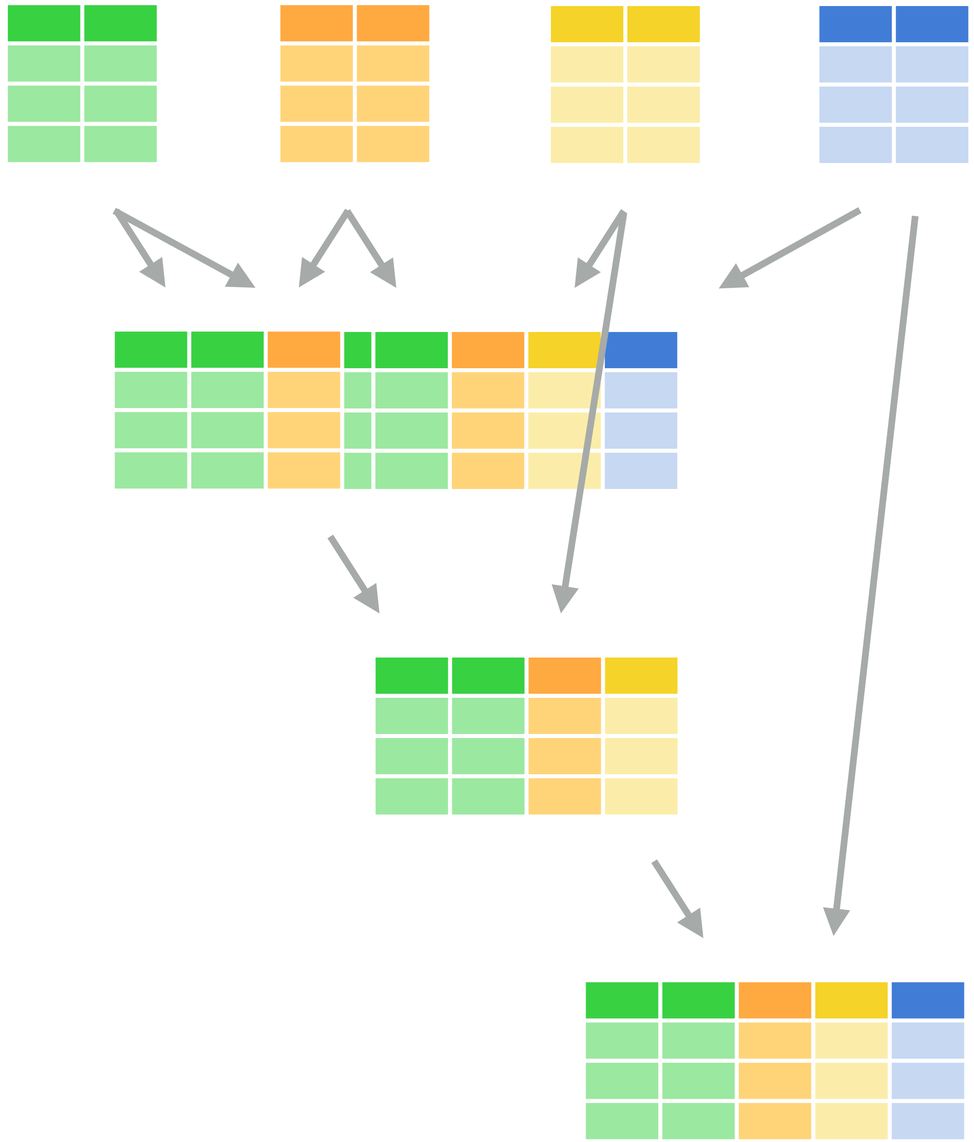
- `bind_rows()`



- `bind_cols()`

Data frames

- `data_frame()`
- `as_data_frame()`



`purrr::reduce()`

Advanced joins

	Missing	Duplicate
key values	Missing key values	Duplicate key columns
key columns	Missing key columns	Duplicate key columns

JOINING DATA IN R WITH DPLYR

Thanks!