Rhw8

SihyuanHan

Identifying Table Keys in the NASA Weather Dataset

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

```r
library(help = "nasaweather")
library("nasaweather")
```

```
## Warning: package 'nasaweather' was built under R version 4.0.3
```

```
##
## Attaching package: 'nasaweather'
```

```
## The following object is masked from 'package:dplyr':
##
##     storms
```

```r
data(package = "nasaweather")
```

- 1-1 What are the data frames in this data set? atmos, borders, elev, glaciers, storms
- 1-2,3 What are the keys in each data frame?

```r
data("atmos")
head(atmos)
```

```
## # A tibble: 6 x 11
##     lat  long  year month surftemp  temp pressure ozone cloudlow cloudmid
##   <dbl> <dbl> <int> <int>    <dbl> <dbl>    <dbl> <dbl>    <dbl>    <dbl>
## 1  36.2 -114.  1995     1     273.  272.      835   304      7.5     34.5
## 2  33.7 -114.  1995     1     280.  282.      940   304     11.5     32.5
## 3  31.2 -114.  1995     1     285.  285.      960   298     16.5     26
## 4  28.7 -114.  1995     1     289.  291.      990   276     20.5     14.5
## 5  26.2 -114.  1995     1     292.  293.     1000   274     26       10.5
## 6  23.7 -114.  1995     1     294.  294.     1000   264     30        9.5
## # ... with 1 more variable: cloudhigh <dbl>
```

```r
atmos %>%
  group_by(lat,long,year,month) %>%
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 5
## # Groups:   lat, long, year, month [0]
## # ... with 5 variables: lat <dbl>, long <dbl>, year <int>, month <int>, n <int>
```

```r
data("borders")

borders %>%
  ungroup() %>%
  head()
```

```
## # A tibble: 6 x 4
##   country  long   lat group
##   <chr>   <dbl> <dbl> <int>
## 1 AG      -61.7  17.0     1
## 2 AG      -61.7  17.0     1
## 3 AG      -61.9  17.0     1
## 4 AG      -61.9  17.1     1
## 5 AG      -61.9  17.1     1
## 6 AG      -61.8  17.2     1
```

```r
borders %>%
  ungroup() %>%
  summarize(dist = nrow(distinct(.)))
```

```
## # A tibble: 1 x 1
##    dist
##   <int>
## 1  7778
```

```r
nrow(borders)
```

```
## [1] 7932
```

```r
borders %>%
  ungroup() %>%
  distinct(.keep_all = TRUE) %>%
  group_by(country,long,lat) %>%
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 4
## # Groups:   country, long, lat [0]
## # ... with 4 variables: country <chr>, long <dbl>, lat <dbl>, n <int>
```

```r
data("elev")
head(elev)
```

```
## # A tibble: 6 x 3
##    long    lat  elev
##   <dbl>  <dbl> <dbl>
## 1 -114. -21.2      0
## 2 -114. -18.7      0
## 3 -114. -16.2      0
## 4 -114. -13.7      0
## 5 -114. -11.2      0
## 6 -114.  -8.72     0
```

```r
elev %>%
  group_by(long,lat) %>%
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 3
## # Groups:   long, lat [0]
## # ... with 3 variables: long <dbl>, lat <dbl>, n <int>
```

```r
data("glaciers")
head(glaciers)
```

```
## # A tibble: 6 x 6
##   id           name           lat  long area   country
##   <chr>        <chr>        <dbl> <dbl> <chr>  <chr>
## 1 CO1A0101001 RAMIREZ E 4   10.8 -73.6 " NA"   CO
## 2 CO1A0101002 RAMIREZ E 3   10.8 -73.6 " NA"   CO
## 3 CO1A0101003 RAMIREZ E 2   10.8 -73.6 " NA"   CO
## 4 CO1A0101004 RAMIREZ E 1   10.8 -73.6 "0.03" CO
## 5 CO1A0101005 RAMIREZ 5 N   10.8 -73.6 "0.1"   CO
## 6 CO1A0101007 RAMIREZ 3 N   10.8 -73.6 "0.03" CO
```

```r
glaciers %>%
  group_by(id) %>%
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 2
## # Groups:   id [0]
## # ... with 2 variables: id <chr>, n <int>
```

```r
data("storms")
head(storms)
```

```
## # A tibble: 6 x 11
##   name    year month   day  hour   lat  long pressure  wind type          seasday
##   <chr>  <int> <int> <int> <int> <dbl> <dbl>    <int> <int> <chr>           <int>
## 1 Allis~  1995     6     3     0  17.4 -84.3     1005    30 Tropical De~       3
## 2 Allis~  1995     6     3     6  18.3 -84.9     1004    30 Tropical De~       3
## 3 Allis~  1995     6     3    12  19.3 -85.7     1003    35 Tropical St~       3
## 4 Allis~  1995     6     3    18  20.6 -85.8     1001    40 Tropical St~       3
## 5 Allis~  1995     6     4     0  22   -86        997    50 Tropical St~       4
## 6 Allis~  1995     6     4     6  23.3 -86.3      995    60 Tropical St~       4
```

```r
storms %>%
  group_by(name,year,month,day,hour,lat) %>%
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 7
## # Groups:   name, year, month, day, hour, lat [0]
## # ... with 7 variables: name <chr>, year <int>, month <int>, day <int>,
## #   hour <int>, lat <dbl>, n <int>
```

Lahman's Baseball Dataset

```r
library(Lahman)
help("Lahman-package")
```

```
## starting httpd help server ... done
```

- 2-1

```r
data("Master")
data("Batting")
data("Pitching")
data("Fielding")
data("Teams")
data("Salaries")
```

- 2-2

```r
# identify primary key
Teams %>%
  group_by(yearID,teamID) %>% # primary key
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 3
## # Groups:   yearID, teamID [0]
## # ... with 3 variables: yearID <int>, teamID <fct>, n <int>
```

```
Master %>%
  group_by(playerID) %>% # primary key
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 2
## # Groups:   playerID [0]
## # ... with 2 variables: playerID <chr>, n <int>
```

```
Fielding %>%
  group_by(playerID,yearID,stint,POS) %>% # primary key
  count() %>%
  filter(n>1)
```

```
## # A tibble: 0 x 5
## # Groups:   playerID, yearID, stint, POS [0]
## # ... with 5 variables: playerID <chr>, yearID <int>, stint <int>, POS <chr>,
## #   n <int>
```

```
Teams %>%
  filter(yearID >= 1903) %>%
  filter(LgWin == "Y") %>%
  filter(!is.na(WSWin)) %>% # not played each year
  filter(teamID == "BOS") %>%
  select(yearID,teamID,LgWin) ->
  team_bos_Lgwin

team_bos_Lgwin %>%
  left_join(Fielding, by = c("yearID","teamID")) %>%
  left_join(Master, by = "playerID") %>%
  filter(stint >= 1) %>%
  select(nameFirst,nameLast,yearID) %>%
  distinct() %>%
  arrange(nameLast) %>%
  head(n=10)
```

```
##    nameFirst nameLast yearID
## 1    Alfredo   Aceves   2013
## 2      Jerry    Adair   1967
## 3      Terry    Adams   2004
## 4        Sam    Agnew   1916
## 5        Sam    Agnew   1918
## 6       Nick  Altrock   1903
## 7        Abe  Alvarez   2004
## 8      Jimmy Anderson   2004
## 9      Ernie   Andres   1946
## 10       Kim   Andrew   1975
```

- 2-3-a

```r
# head(Salaries)
Salaries %>%
  group_by(yearID,playerID) %>%
  summarize(salary_total = sum(salary, na.rm = TRUE)) ->
  Salaries_3_a
```

```
## `summarise()` regrouping output by 'yearID' (override with `.groups` argument)
```

```r
Salaries_3_a
```

```
## # A tibble: 26,323 x 3
## # Groups:   yearID [32]
##    yearID playerID  salary_total
##     <int> <chr>            <int>
##  1   1985 ackerji01       170000
##  2   1985 agostju01       147500
##  3   1985 aguaylu01       237000
##  4   1985 alexado01       875000
##  5   1985 allenne01       750000
##  6   1985 almonbi01       255000
##  7   1985 anderal02        62500
##  8   1985 anderla02       250500
##  9   1985 andujjo01      1030000
## 10   1985 armasto01       915000
## # ... with 26,313 more rows
```

- 2-3-b

```r
Batting %>%
  left_join(Master, by = "playerID") %>%
  select(AB,H,playerID,yearID) %>%
  group_by(yearID,playerID) %>%
  summarize(sum_bats = sum(AB), sum_hits = sum(H)) ->
  Batting_3_b
```

```
## `summarise()` regrouping output by 'yearID' (override with `.groups` argument)
```

```r
Batting_3_b
```

```
## # A tibble: 99,402 x 4
## # Groups:   yearID [149]
##    yearID playerID  sum_bats sum_hits
##     <int> <chr>        <int>    <int>
##  1   1871 abercda01        4        0
##  2   1871 addybo01       118       32
##  3   1871 allisar01      137       40
##  4   1871 allisdo01      133       44
##  5   1871 ansonca01      120       39
##  6   1871 armstbo01       49       11
##  7   1871 barkeal01        4        1
##  8   1871 barnero01      157       63
```

```
## 9    1871 barrebi01        5        1
## 10   1871 barrofr01       86       13
## # ... with 99,392 more rows
```

- 2-4-a

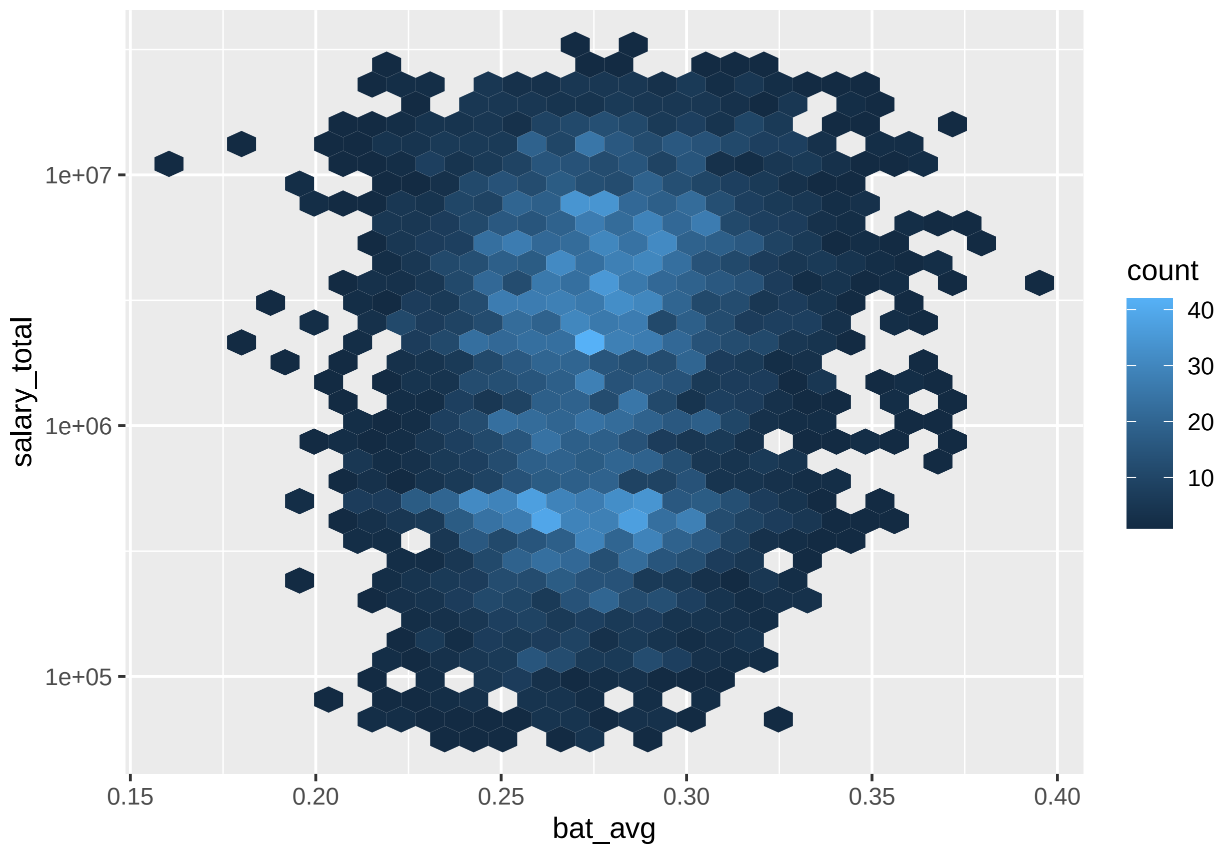
```
Batting_3_b %>%
  left_join(Salaries_3_a, by = c("yearID","playerID")) %>%
  mutate(bat_avg = sum_hits/sum_bats) %>%
  filter(sum_bats >= 400) %>%
  filter(yearID >= 1985) ->
  Batting_4_a
Batting_4_a
```

```
## # A tibble: 6,016 x 6
## # Groups:   yearID [35]
##    yearID playerID  sum_bats sum_hits salary_total bat_avg
##     <int> <chr>        <int>    <int>        <int>   <dbl>
##  1   1985 backmwa01      520      142       200000   0.273
##  2   1985 baineha01      640      198       675000   0.309
##  3   1985 balbost01      600      146       205000   0.243
##  4   1985 barfije01      539      156       325000   0.289
##  5   1985 barrema02      534      142       272500   0.266
##  6   1985 basske01       539      145       155000   0.269
##  7   1985 baylodo01      477      110       810000   0.231
##  8   1985 bellbu01       560      128       751297   0.229
##  9   1985 bellge02       607      167       335000   0.275
## 10   1985 beniqju01      411      125       365000   0.304
## # ... with 6,006 more rows
```

- 2-4-b hexplot
- Based on the hexplot, we can see batting average between 0.25-0.3 has lower salary than others.

```
Batting_4_a %>%
  ggplot(aes(bat_avg, salary_total))+
  geom_hex()+
  scale_y_log10()
```

```
## Warning: Removed 671 rows containing non-finite values (stat_binhex).
```

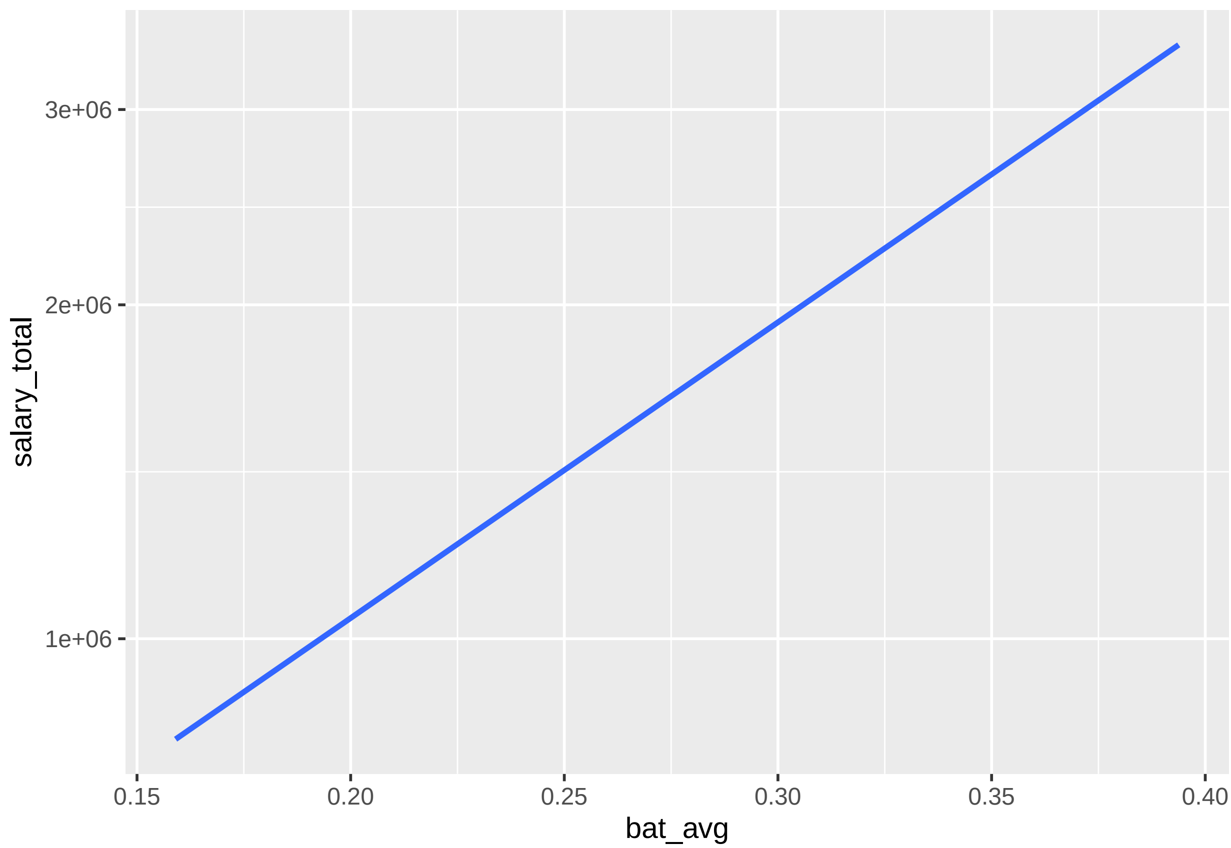


- 2-4-c
- We can learn from the plot(OLS) that the relationship between batting average and salary is positive, so when batting average is high, the salary is high.

```
Batting_4_a %>%
  ggplot(aes(bat_avg, salary_total))+
  scale_y_log10()+
  geom_smooth(se = FALSE, method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 671 rows containing non-finite values (stat_smooth).
```



- 2-4-d
- The pairwise complete correlation between batting average and log of the total salary by year has negative coefficients. As year pass, the correlation is decreasing.

```
Batting_4_a %>%
  group_by(yearID) %>%
  summarize(pc_cor = cor(bat_avg, log(salary_total), use="pairwise")) ->
  Batting_pc_cor
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Batting_pc_cor
```

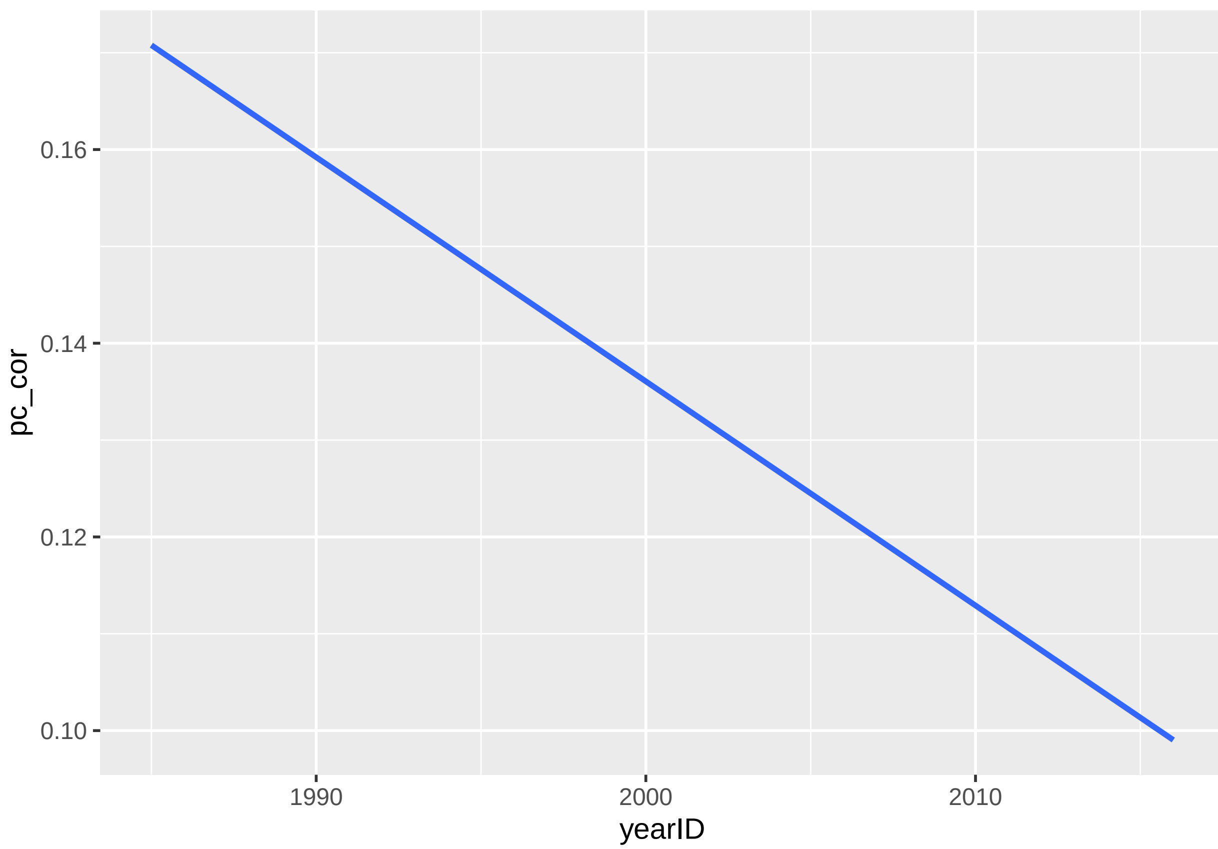
```
## # A tibble: 35 x 2
##    yearID   pc_cor
##     <int>    <dbl>
##  1   1985  0.196
##  2   1986  0.280
##  3   1987  0.0783
##  4   1988  0.204
##  5   1989  0.0813
##  6   1990 -0.00234
##  7   1991  0.0325
##  8   1992  0.153
```

```
## 9    1993  0.142
## 10   1994  0.138
## # ... with 25 more rows
```

```
Batting_pc_cor %>%
  ggplot(aes(yearID, pc_cor))+
  geom_smooth(se = FALSE, method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```



- 2-5

```
head(Master)
```

```
##    playerID birthYear birthMonth birthDay birthCountry birthState  birthCity
## 1 aardsda01      1981         12       27          USA         CO     Denver
## 2 aaronha01      1934          2        5          USA         AL     Mobile
## 3 aaronto01      1939          8        5          USA         AL     Mobile
## 4  aasedo01      1954          9        8          USA         CA     Orange
## 5  abadan01      1972          8       25          USA         FL Palm Beach
## 6  abadfe01      1985         12       17         D.R.  La Romana  La Romana
```

```
##   deathYear deathMonth deathDay deathCountry deathState deathCity nameFirst
## 1        NA         NA       NA         <NA>       <NA>      <NA>     David
## 2        NA         NA       NA         <NA>       <NA>      <NA>      Hank
## 3      1984          8       16          USA         GA   Atlanta    Tommie
## 4        NA         NA       NA         <NA>       <NA>      <NA>       Don
## 5        NA         NA       NA         <NA>       <NA>      <NA>      Andy
## 6        NA         NA       NA         <NA>       <NA>      <NA>  Fernando
##   nameLast        nameGiven weight height bats throws      debut  finalGame
## 1  Aardsma      David Allan    215     75    R      R 2004-04-06 2015-08-23
## 2    Aaron      Henry Louis    180     72    R      R 1954-04-13 1976-10-03
## 3    Aaron       Tommie Lee    190     75    R      R 1962-04-10 1971-09-26
## 4     Aase   Donald William    190     75    R      R 1977-07-26 1990-10-03
## 5     Abad    Fausto Andres    184     73    L      L 2001-09-10 2006-04-13
## 6     Abad Fernando Antonio    220     73    L      L 2010-07-28 2019-09-28
##    retroID   bbrefID  deathDate  birthDate
## 1 aardd001 aardsda01       <NA> 1981-12-27
## 2 aaroh101 aaronha01       <NA> 1934-02-05
## 3 aarot101 aaronto01 1984-08-16 1939-08-05
## 4 aased001  aasedo01       <NA> 1954-09-08
## 5 abada001  abadan01       <NA> 1972-08-25
## 6 abadf001  abadfe01       <NA> 1985-12-17
```

```
Salaries_3_a %>%
  left_join(Master, by = "playerID") %>%
  filter(nameFirst == "John") %>%
  filter(yearID %% 2 == 0) %>%
  arrange(desc(salary_total), n = 10) %>%
  select(yearID, nameFirst, nameLast, salary_total)
```

```
## # A tibble: 236 x 4
## # Groups:   yearID [16]
##    yearID nameFirst nameLast salary_total
##     <int> <chr>     <chr>           <int>
##  1   2010 John      Lackey       18700000
##  2   2016 John      Lackey       16000000
##  3   2012 John      Lackey       15950000
##  4   2016 John      Danks        15750000
##  5   2014 John      Lackey       15250000
##  6   2014 John      Danks        14250000
##  7   2008 John      Smoltz       14000000
##  8   2004 John      Smoltz       11666667
##  9   2006 John      Smoltz       11000000
## 10   2000 John      Smoltz        8500000
## # ... with 226 more rows
```

NYC Flights

- 3-1

```
library(dbplyr)
```

```
##
## Attaching package: 'dbplyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```r
library(RSQLite,lib.loc = "C:/Users/Stephanie/Documents/R/win-library/4.0")

conn <- dbConnect(drv = SQLite(), dbname = "../R_data/nycflights13.sqlite")
```

- 3-2

```r
dbListTables(conn)
```

```
## [1] "airlines"     "airports"     "flights"      "planes"       "sqlite_stat1"
## [6] "sqlite_stat4" "weather"
```

- 3-3

```r
airlines_db <- tbl(conn,"airlines")
airports_db <- tbl(conn,"airports")
flights_db <- tbl(conn,"flights")
planes_db <- tbl(conn,"planes")
weather_db <- tbl(conn,"weather")
```

- 3-4 in-memory data frame, only for flights that actually departed

```r
head(airports_db)
```

```
## # Source:   lazy query [?? x 8]
## # Database: sqlite 3.33.0
## #   [C:\Users\Stephanie\Documents\stat_612(R)\R_data\nycflights13.sqlite]
##   faa   name                           lat   lon   alt    tz dst   tzone
##   <chr> <chr>                        <dbl> <dbl> <dbl> <dbl> <chr> <chr>
## 1 04G   Lansdowne Airport             41.1 -80.6  1044    -5 A     America/New_Y~
## 2 06A   Moton Field Municipal Airp~  32.5 -85.7   264    -6 A     America/Chica~
## 3 06C   Schaumburg Regional           42.0 -88.1   801    -6 A     America/Chica~
## 4 06N   Randall Airport               41.4 -74.4   523    -5 A     America/New_Y~
## 5 09J   Jekyll Island Airport         31.1 -81.4    11    -5 A     America/New_Y~
## 6 0A9   Elizabethton Municipal Air~  36.4 -82.2  1593    -5 A     America/New_Y~
```

```r
head(flights_db)
```

```
## # Source:   lazy query [?? x 19]
## # Database: sqlite 3.33.0
## #   [C:\Users\Stephanie\Documents\stat_612(R)\R_data\nycflights13.sqlite]
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
```

```
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dbl>
```

```
flights_db %>%
  collect() ->
  flights

airports_db %>%
  collect() ->
  airports

flights %>%
  summarize(across(everything(),~sum(is.na(.))))
```

```
## # A tibble: 1 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <int>    <int>          <int>
## 1     0     0     0     8255              0      8255     8713              0
## # ... with 11 more variables: arr_delay <int>, carrier <int>, flight <int>,
## #   tailnum <int>, origin <int>, dest <int>, air_time <int>, distance <int>,
## #   hour <int>, minute <int>, time_hour <int>
```

```
flights %>%
  filter(!is.na(dep_time)) ->
  flights_check_dep
head(flights_check_dep)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dbl>
```

- 3-5 in-memory data frame, airports served by NYC airports, 104 rows

```
airports %>%
  semi_join(flights, by = c("faa" = "dest")) ->
  airports_dest

airports %>%
  semi_join(flights, by = c("faa" = "origin")) ->
  airports_origin
```

```r
rbind(airports_dest, airports_origin) ->
  airports_by_NYC

airports_by_NYC
```

```
## # A tibble: 104 x 8
##    faa   name                          lat    lon   alt    tz dst   tzone
##    <chr> <chr>                       <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
##  1 ABQ   Albuquerque International~  35.0 -107.   5355    -7 A     America/Denv~
##  2 ACK   Nantucket Mem               41.3  -70.1    48    -5 A     America/New_~
##  3 ALB   Albany Intl                 42.7  -73.8   285    -5 A     America/New_~
##  4 ANC   Ted Stevens Anchorage Intl  61.2 -150.    152    -9 A     America/Anch~
##  5 ATL   Hartsfield Jackson Atlant~  33.6  -84.4  1026    -5 A     America/New_~
##  6 AUS   Austin Bergstrom Intl       30.2  -97.7   542    -6 A     America/Chic~
##  7 AVL   Asheville Regional Airport  35.4  -82.5  2165    -5 A     America/New_~
##  8 BDL   Bradley Intl                41.9  -72.7   173    -5 A     America/New_~
##  9 BGR   Bangor Intl                 44.8  -68.8   192    -5 A     America/New_~
## 10 BHM   Birmingham Intl             33.6  -86.8   644    -6 A     America/Chic~
## # ... with 94 more rows
```

- 3-6

```r
flights_check_dep %>%
  anti_join(airports_by_NYC, by = c("dest" = "faa")) %>%
  group_by(dest) %>%
  summarize(total_flights = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 2
##   dest  total_flights
##   <chr>         <int>
## 1 BQN             891
## 2 PSE             361
## 3 SJU            5791
## 4 STT             518
```

- 3-7

```r
flights_check_dep %>%
  inner_join(airports, by = c("dest" = "faa")) %>%
  group_by(name,dest) %>%
  summarize(sum = n()) %>%
  arrange(desc(sum)) %>%
  head(n = 10) ->
  flights_dest_top10
```

```
## `summarise()` regrouping output by 'name' (override with `.groups` argument)
```

```
flights_dest_top10
```

```
## # A tibble: 10 x 3
## # Groups:   name [10]
##    name                                dest    sum
##    <chr>                               <chr> <int>
##  1 Hartsfield Jackson Atlanta Intl     ATL   16898
##  2 Chicago Ohare Intl                  ORD   16642
##  3 Los Angeles Intl                    LAX   16076
##  4 General Edward Lawrence Logan Intl BOS   15049
##  5 Orlando Intl                        MCO   13982
##  6 Charlotte Douglas Intl              CLT   13698
##  7 San Francisco Intl                  SFO   13230
##  8 Fort Lauderdale Hollywood Intl      FLL   11934
##  9 Miami Intl                          MIA   11633
## 10 Ronald Reagan Washington Natl       DCA    9157
```

- 3-8-a

```
head(airlines_db)
```

```
## # Source:   lazy query [?? x 2]
## # Database: sqlite 3.33.0
## #   [C:\Users\Stephanie\Documents\stat_612(R)\R_data\nycflights13.sqlite]
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
```

```
airlines_db %>%
  collect() ->
  airlines

airlines %>%
  semi_join(flights, by = "carrier") ->
  airlines_nyc
airlines_nyc
```

```
## # A tibble: 16 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
##  6 EV      ExpressJet Airlines Inc.
```

```
##   7 F9      Frontier Airlines Inc.
##   8 FL      AirTran Airways Corporation
##   9 HA      Hawaiian Airlines Inc.
##  10 MQ      Envoy Air
##  11 OO      SkyWest Airlines Inc.
##  12 UA      United Air Lines Inc.
##  13 US      US Airways Inc.
##  14 VX      Virgin America
##  15 WN      Southwest Airlines Co.
##  16 YV      Mesa Airlines Inc.
```

- 3-8-b

```
flights_dest_top10 %>%
  left_join(flights, by = "dest") %>%
  group_by(name, dest, carrier) %>%
  summarize(median_arr_delay = median(arr_delay, na.rm = TRUE),
            flights = n()) %>%
  group_by(dest) %>%
  arrange((median_arr_delay)) %>%
  slice(1:2) %>%
  select(-carrier)
```

```
## `summarise()` regrouping output by 'name', 'dest' (override with `.groups` argument)
```

```
## # A tibble: 20 x 4
## # Groups:   dest [10]
##    name                              dest  median_arr_delay flights
##    <chr>                             <chr>            <dbl>   <int>
##  1 Hartsfield Jackson Atlanta Intl    ATL                 -6     103
##  2 Hartsfield Jackson Atlanta Intl    ATL                 -4   10571
##  3 General Edward Lawrence Logan Intl BOS                -13     972
##  4 General Edward Lawrence Logan Intl BOS                -10     159
##  5 Charlotte Douglas Intl             CLT                 -9     282
##  6 Charlotte Douglas Intl             CLT                 -5    8632
##  7 Ronald Reagan Washington Natl      DCA                -14    1074
##  8 Ronald Reagan Washington Natl      DCA                 -8       2
##  9 Fort Lauderdale Hollywood Intl     FLL                 -7     182
## 10 Fort Lauderdale Hollywood Intl     FLL                 -7    2903
## 11 Los Angeles Intl                   LAX                -10    3582
## 12 Los Angeles Intl                   LAX                 -9    2501
## 13 Orlando Intl                       MCO                 -9    3663
## 14 Orlando Intl                       MCO                 -8    3217
## 15 Miami Intl                         MIA                -10    7234
## 16 Miami Intl                         MIA                 -9    2929
## 17 Chicago Ohare Intl                 ORD                -12    6059
## 18 Chicago Ohare Intl                 ORD                 -7    6984
## 19 San Francisco Intl                 SFO                -13    1858
## 20 San Francisco Intl                 SFO                -12    2197
```

- 3-8-c

```r
flights_dest_top10 %>%
  left_join(flights, by = "dest") %>%
  group_by(name, dest, carrier) %>%
  summarize(median_arr_delay = median(arr_delay, na.rm = TRUE),
            flights = n()) %>%
  arrange(desc(median_arr_delay)) %>%
  head(n = 10) %>%
  select(-carrier)
```

```
## `summarise()` regrouping output by 'name', 'dest' (override with `.groups` argument)
```

```
## # A tibble: 10 x 4
## # Groups:   name, dest [4]
##    name                          dest  median_arr_delay flights
##    <chr>                         <chr>            <dbl>   <int>
##  1 Chicago Ohare Intl            ORD              107         1
##  2 Chicago Ohare Intl            ORD               17.5       2
##  3 Charlotte Douglas Intl        CLT               14.5       2
##  4 Hartsfield Jackson Atlanta Intl ATL              6      2337
##  5 Ronald Reagan Washington Natl DCA                5      1717
##  6 Hartsfield Jackson Atlanta Intl ATL              4.5      59
##  7 Hartsfield Jackson Atlanta Intl ATL              4      1764
##  8 Hartsfield Jackson Atlanta Intl ATL              4      2322
##  9 Charlotte Douglas Intl        CLT                2      2508
## 10 Charlotte Douglas Intl        CLT                2      1620
```