# More data Wrangling with joins and tidyR

Math 241, Week 4

```
# it's good practice to check that all the packages required are loaded and installed
libs <- c('tidyverse','knitr','viridis','mosaicData','babynames','mdsr','Lahman','nycflights13')
for(l in libs){
  if(!require(l,character.only = TRUE, quietly = TRUE)){
    message( sprintf('Did not have the required package << %s >> installed. Downloading now ... ',l))
    install.packages(l)
  }
  library(l, character.only = TRUE, quietly = TRUE)
}
```

### Goals of this in-class activity:

- Practice data wrangling and joins with tidyR

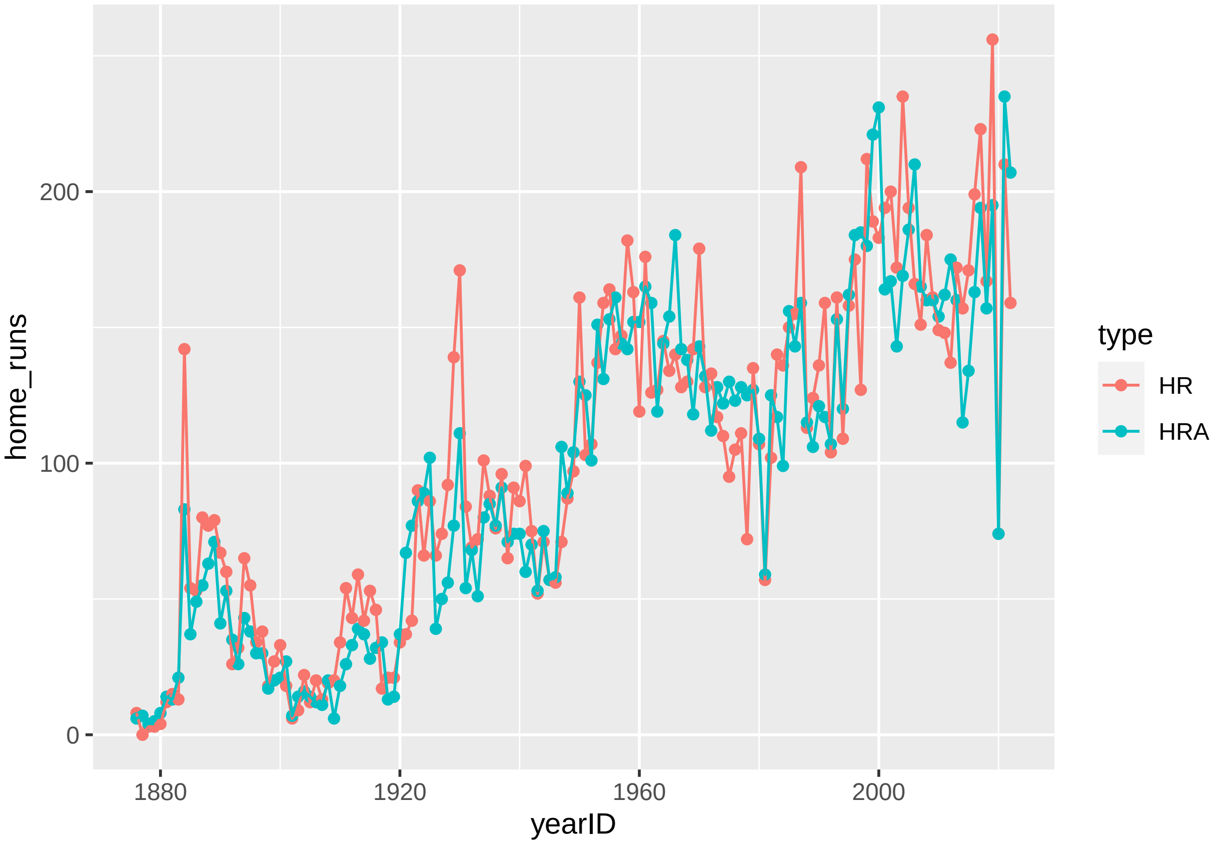
### Notes:

- Be prepared to ask for help from me, Tory, and your classmates!

### Problem 1 (Medium):

Consider the number of home runs hit (HR) and home runs allowed (HRA) for the Chicago Cubs (CHN) baseball team. Reshape the Teams data from the `Lahman` package into “long” format and plot a time series conditioned on whether the HRs that involved the Cubs were hit by them or allowed by them.

```
Teams %>%
  filter(teamID == "CHN") %>%
  select(yearID, HR, HRA) %>%
  pivot_longer(-yearID, names_to = "type", values_to = "home_runs") %>%
  ggplot(aes(x = yearID, y = home_runs, color = type)) +
  geom_point() +
  geom_line()
```



## Problem 2 (Medium):

Use the `nycflights13` package and the `flights` and `planes` tables to answer the following questions:

a. How many planes have a missing date of manufacture?

```
library(nycflights13)
planes2 <- select(planes, tailnum, year, manufacturer)
flights2 <- select(flights, tailnum)
nyc_flights <- left_join(planes2, flights2)
nyc_flights2 <- nyc_flights %>%
  filter(is.na(year)) %>%
  distinct(tailnum)
nrow(nyc_flights2)
```

```
## [1] 70
```

There are 70 airplanes with a missing date of manufacture.

b. What are the five most common manufacturers?

```
nyc_flights %>%
  select(manufacturer, tailnum, year) %>%
  unique() %>%
  group_by(manufacturer) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 35 x 2
##    manufacturer                  count
##    <chr>                         <int>
##  1 BOEING                         1630
##  2 AIRBUS INDUSTRIE                400
##  3 BOMBARDIER INC                  368
##  4 AIRBUS                          336
##  5 EMBRAER                         299
##  6 MCDONNELL DOUGLAS               120
##  7 MCDONNELL DOUGLAS AIRCRAFT CO   103
##  8 MCDONNELL DOUGLAS CORPORATION    14
##  9 CANADAIR                          9
## 10 CESSNA                            9
## # i 25 more rows
```

## Problem 3 (Medium):

Use the nycflights13 package and the flights and planes tables to answer the following questions:

a. What is the oldest plane (specified by the tailnum variable) that flew from New York City airports in 2013?

```
planes2 <- dplyr::select(planes, tailnum, year)
flights2 <- dplyr::select(flights, tailnum)
nyc_flights <- left_join(planes2, flights2)
head(nyc_flights)
```

```
## # A tibble: 6 x 2
##   tailnum  year
##   <chr>   <int>
## 1 N10156   2004
## 2 N10156   2004
## 3 N10156   2004
## 4 N10156   2004
## 5 N10156   2004
## 6 N10156   2004
```

N381AA, manufactured in 1956, is the oldest plane that flew from NYC in 2013.

b. How many airplanes that flew from New York City are included in the planes table?

```
nyc_flights2 <- distinct(nyc_flights)
nrow(nyc_flights2)
```

```
## [1] 3322
```

There are 3322 unique airplanes.

## Problem 4 (Medium):

The knitr package allows the analyst to display nicely formatted tables and results when outputting to pdf files. Use the following code chunk as an example to create a similar display for the penguins dataset, in the palmerpenguins package, instead (you can model penguins' body_mass_g as a function of their flipper_length_mm and sex):

```
mod <- broom::tidy(lm(cesd ~ mcs + sex, data = HELPrct))
knitr::kable(
  mod,
  digits = c(0, 2, 2, 2, 4),
  caption = "Regression model from HELP clinical trial.",
  longtable = TRUE
)
```

Table 1: Regression model from HELP clinical trial.

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | 55.79 | 1.31 | 42.62 | 0.0000 |
| mcs | -0.65 | 0.03 | -19.48 | 0.0000 |
| sexmale | -2.95 | 1.01 | -2.91 | 0.0038 |

```
library(palmerpenguins)
mod <- broom::tidy(lm(body_mass_g ~ flipper_length_mm + sex, data = penguins))
knitr::kable(mod, digits = c(0, 1, 1, 1, 4), longtable = TRUE)
```

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | -5410.3 | 285.8 | -18.9 | 0 |
| flipper_length_mm | 47.0 | 1.4 | 32.6 | 0 |
| sexmale | 347.9 | 40.3 | 8.6 | 0 |