Data wrangling fundamentals

Last time

- `filter`: choose certain rows
- `summarize`: calculate summary statistics
- `group_by`: group rows together
- `mutate`: create new columns

Data for today

- Data on professional baseball teams between 1871 and 2023
- 3015 rows and 48 columns
- Each row represents one year (season) for one team
- Variables include:
  - `yearID`: Year
  - `franchID`: Franchise
  - `W`: Wins
  - `L`: Losses

Data for today

- Variables include:
  - `yearID`: Year
  - `franchID`: Franchise
  - `W`: Wins
  - `L`: Losses

We want to know: which NY Mets general manager performed best between 1998 - 2018

Warmup activity

Work on the activity (handout) with a neighbor, then we will discuss as a class

Step 0: Make the columns more manageable

There are 48 columns in the initial data! Let’s only focus on the ones we care about:

```r
Teams |>
  select(yearID, franchID, W, L)
```

```
     yearID franchID   W   L
1      1871      BNA  20  10
2      1871      CNA  19   9
3      1871      CFC  10  19
4      1871      KEK   7  12
5      1871      NNA  16  17
6      1871      PNA  21   7
7      1871      ROK   4  21
8      1871      TRO  13  15
9      1871      OLY  15  15
10     1872      BLC  35  19
11     1872      ECK   3  26
```

Step 1: Focus on the Mets between 1998 and 2018

```
Teams |>
  select(yearID, franchID, W, L) |>
  filter(...)
```

Question: What goes in my `filter`?

Step 1: Focus on the Mets between 1998 and 2018

```r
Teams |>
  select(yearID, franchID, W, L) |>
  filter(franchID == "NYM",
         yearID >= 1998, yearID <= 2018)
```

```
   yearID franchID  W  L
1    1998      NYM 88 74
2    1999      NYM 97 66
3    2000      NYM 94 68
4    2001      NYM 82 80
5    2002      NYM 75 86
6    2003      NYM 66 95
7    2004      NYM 71 91
8    2005      NYM 83 79
9    2006      NYM 97 65
10   2007      NYM 88 74
11   2008      NYM 89 73
12   2009      NYM 70 92
```

Step 2: Who was the GM?

- 1998 - 2003: Steve Phillips
- 2004: Jim Duquette
- 2005 - 2010: Omar Minaya
- 2011 - 2018: Sandy Alderson

How should we add this information to the data?

Step 2: Who was the GM?

```
Teams |>
  select(yearID, franchID, W, L) |>
  filter(franchID == "NYM",
         yearID >= 1998, yearID <= 2018) |>
  mutate(gm = case_when(
    yearID <= 2003 ~ "Phillips",
    yearID == 2004 ~ "Duquette",
    yearID <= 2010 ~ "Minaya",
    yearID <= 2018 ~ "Alderson"
  ))
```

```
   yearID franchID  W  L       gm
1    1998      NYM 88 74 Phillips
2    1999      NYM 97 66 Phillips
3    2000      NYM 94 68 Phillips
4    2001      NYM 82 80 Phillips
5    2002      NYM 75 86 Phillips
6    2003      NYM 66 95 Phillips
7    2004      NYM 71 91 Duquette
8    2005      NYM 83 79   Minaya
9    2006      NYM 97 65   Minaya
```

```
10   2007      NYM 88 74   Minaya
11   2008      NYM 89 73   Minaya
12   2009      NYM 70 92   Minaya
```

Step 3: Summarize performance

```
   yearID franchID  W  L       gm
1    1998      NYM 88 74 Phillips
2    1999      NYM 97 66 Phillips
3    2000      NYM 94 68 Phillips
4    2001      NYM 82 80 Phillips
5    2002      NYM 75 86 Phillips
6    2003      NYM 66 95 Phillips
7    2004      NYM 71 91 Duquette
8    2005      NYM 83 79   Minaya
9    2006      NYM 97 65   Minaya
10   2007      NYM 88 74   Minaya
11   2008      NYM 89 73   Minaya
12   2009      NYM 70 92   Minaya
```

How would I calculate the win percentage for *each* GM?

Step 3: Summarize performance

```r
Teams |>
  select(yearID, franchID, W, L) |>
  filter(franchID == "NYM",
         yearID >= 1998, yearID <= 2018) |>
  mutate(gm = case_when(
    yearID <= 2003 ~ "Phillips",
    yearID == 2004 ~ "Duquette",
    yearID <= 2010 ~ "Minaya",
    yearID <= 2018 ~ "Alderson"
  )) |>
  group_by(gm) |>
  summarize(wpct = sum(W)/sum(W + L))
```

```
# A tibble: 4 × 2
  gm         wpct
  <chr>     <dbl>
1 Alderson 0.485
2 Duquette 0.438
3 Minaya   0.521
4 Phillips 0.517
```

Finally: arrange results

```r
Teams |>
  select(yearID, franchID, W, L) |>
  filter(franchID == "NYM",
         yearID >= 1998, yearID <= 2018) |>
  mutate(gm = case_when(
    yearID <= 2003 ~ "Phillips",
    yearID == 2004 ~ "Duquette",
    yearID <= 2010 ~ "Minaya",
    yearID <= 2018 ~ "Alderson"
  )) |>
  group_by(gm) |>
  summarize(wpct = sum(W)/sum(W + L)) |>
  arrange(desc(wpct))
```

```
# A tibble: 4 × 2
  gm        wpct
  <chr>    <dbl>
1 Minaya   0.521
2 Phillips 0.517
3 Alderson 0.485
4 Duquette 0.438
```

Class activity

https://sta279-f25.github.io/class_activities/ca_03.html

- Work with a neighbor on the class activity
- At the end of class, submit your work as an HTML file on Canvas (one per group, list all your names)

Monday's class will be reserved for getting Git and GitHub setup. We will use these tools for the rest of the semester.

- Work through the **Git and GitHub assignment instructions** on the course website
- If you successfully complete all steps, you do not need to come to class