

Lecture 12: Data wrangling

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 2022
- 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

Making a plan

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

Question: What steps could we take to answer this question?

- choose NY Mets
- choose seasons 1998 - 2018
- Look at $w \geq L$ for each season
- Add a column for GM \leftarrow mutate
- Calculate win Pct $\frac{w}{w+L}$ for each GM
- group-by, summarize

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:

```
1 Teams |> select columns to keep
2   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
12	1872	BRA	9	28
13	1872	RNA	39	8

|> means
"take THIS, then do
THAT"

Step 1: Focus on the Mets

== to check for equality

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   ... (franchID == "NYM")
```

What function do I use to choose only the rows corresponding to the Mets?

Step 1: Focus on the Mets

```
1 Teams |>  
2   select(yearID, franchID, W, L) |>  
3   filter(franchID == "NYM")
```

	yearID	franchID	W	L
1	1962	NYM	40	120
2	1963	NYM	51	111
3	1964	NYM	53	109
4	1965	NYM	50	112
5	1966	NYM	66	95
6	1967	NYM	61	101
7	1968	NYM	73	89
8	1969	NYM	100	62
9	1970	NYM	83	79
10	1971	NYM	83	79
11	1972	NYM	83	73
12	1973	NYM	82	79
13	1974	NYM	71	91

Step 2: Focus on the Mets between 1998 and 2018

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   filter(franchID == "NYM",
4           ...)
```

How do I specify the range of years I want?

$\text{yearID} \geq 1998$

$\text{yearID} \leq 2018$

or

$\text{yearID} \%in\% 1998:2018$

Step 2: Focus on the Mets between 1998 and 2018

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   filter(franchID == "NYM",
4          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
13	2010	NYM	79	83

↑ equivalent to
franchID == "NYM" &
yearID >= 1998 & yearID <= 2018

Step 3: 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?

- mutate

gm = ---

Logic: if year 1998 - 2003 then "Phillips"
if year 2004 then "Duquette"

⋮

Step 3: Who was the GM?

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   filter(franchID == "NYM",
4         yearID >= 1998, yearID <= 2018) |>
5   mutate(gm = case_when(
6     yearID <= 2003 ~ "Phillips",
7     yearID == 2004 ~ "Duquette",
8     yearID <= 2010 ~ "Minaya",
9     yearID <= 2018 ~ "Alderson"
10  ))
```

	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
13	2010	NYM	79	83	Minaya

Step 4: Summarize performance

How do I calculate performance for *each* GM?

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   filter(franchID == "NYM",
4           yearID >= 1998, yearID <= 2018) |>
5   mutate(gm = case_when(
6     yearID <= 2003 ~ "Phillips",
7     yearID == 2004 ~ "Duquette",
8     yearID <= 2010 ~ "Minaya",
9     yearID <= 2018 ~ "Alderson"
10  )) |>
11  summarize(wpct = sum(W)/sum(W + L))
```

```
      wpct
1 0.5019112
```

Step 4: Summarize performance

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   filter(franchID == "NYM",
4          yearID >= 1998, yearID <= 2018) |>
5   mutate(gm = case_when(
6     yearID <= 2003 ~ "Phillips",
7     yearID == 2004 ~ "Duquette",
8     yearID <= 2010 ~ "Minaya",
9     yearID <= 2018 ~ "Alderson"
10  )) |>
11  group_by(gm) |>
12  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

```
1 Teams |>
2   select(yearID, franchID, W, L) |>
3   filter(franchID == "NYM",
4         yearID >= 1998, yearID <= 2018) |>
5   mutate(gm = case_when(
6     yearID <= 2003 ~ "Phillips",
7     yearID == 2004 ~ "Duquette",
8     yearID <= 2010 ~ "Minaya",
9     yearID <= 2018 ~ "Alderson"
10  )) |>
11  group_by(gm) |>
12  summarize(wpct = sum(W)/sum(W + L)) |>
13  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

sorts the rows by one or more columns in the data

arrange(wpct) (lowest to highest wpct)

arrange(desc(wpct)) (highest to lowest wpct)

Class activity

<https://sta279->

[f23.github.io/class_activities/ca_lecture_12.html](https://sta279-f23.github.io/class_activities/ca_lecture_12.html)

