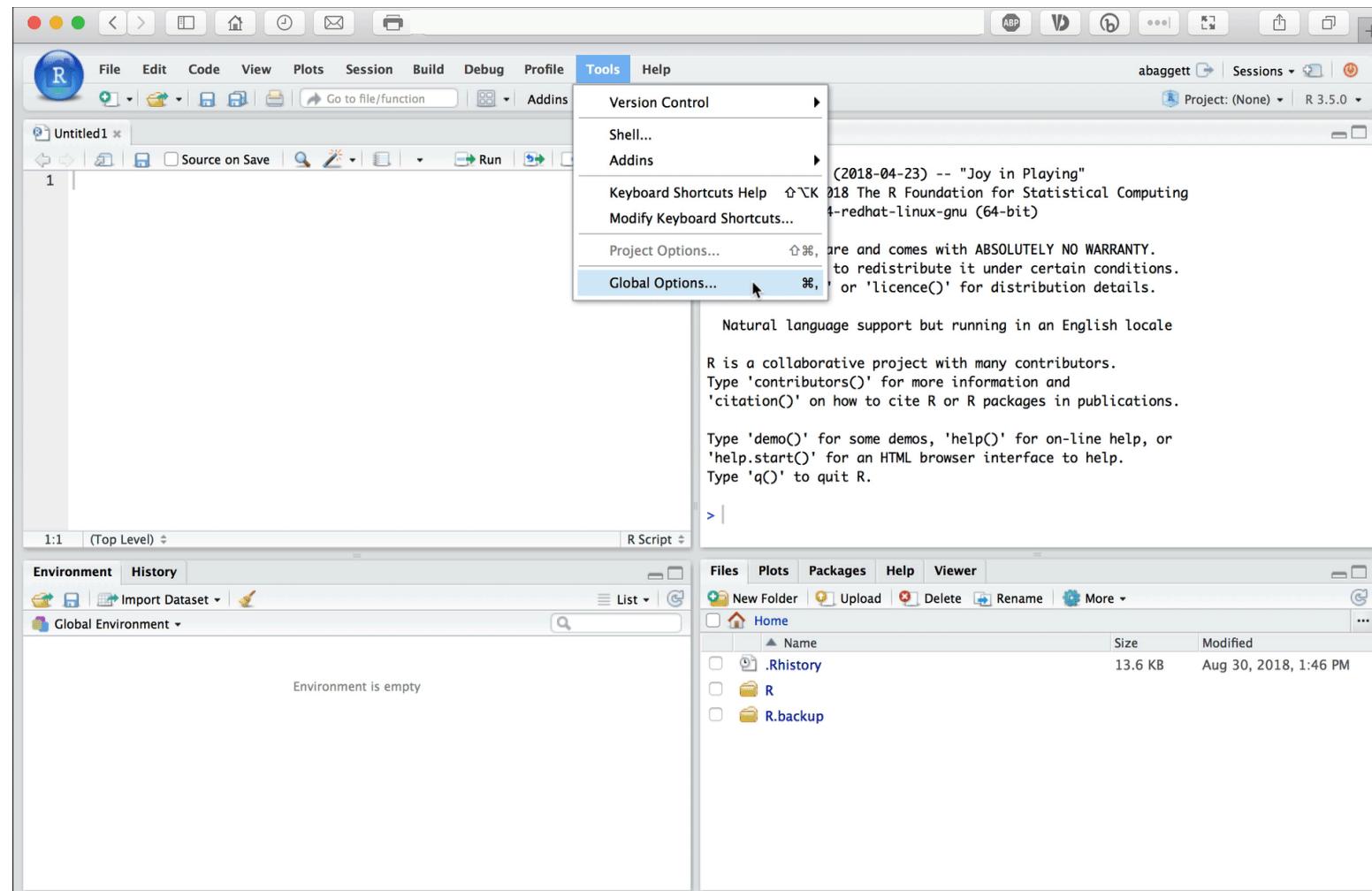


# WELCOME!

- Before we get started:
  1. Login to [r.umhb.edu](http://r.umhb.edu)
  2. File > New File > R Script
  3. Make yours look like this



Slightly  
↓

# A Deeper Dive into R

Introducing the Verbs of Data Manipulation

---

**Aaron R. Baggett, Ph.D.**

University of Mary Hardin-Baylor

March 6, 2019



# FOLLOW ALONG

[bit.ly/r\\_dive](https://bit.ly/r_dive)

# WHAT IS R?

- R is:
  - A powerful, flexible statistics and data software program
  - Free and open source
  - Surging in adoption worldwide
    - Ranks 14<sup>th</sup> among 50 of the most used programming languages in the world<sup>[1]</sup>
  - Like a set of tools



[1] TIOBE Programming Index, <http://bit.ly/2x4TQGh>, March 2, 2019

# R AS A COLLECTION OF TOOLS



VS.



*The tidyverse is a coherent system of packages for  
data manipulation, exploration, and visualization that  
share a common design philosophy.*



# tidyverse Verbs

- Verbs of data manipulation
  1. Mutate
  2. Filter
  3. Select
  4. Summarize



# LET'S GET STARTED!

# NFL 2018 DATA

- Let's look at some data from the 2018 NFL season
  1. What was each team's winning percentage?
  2. What was the mean number of points scored per game over the course of the season, across all teams?
  3. To what extent are points scored and winning percentage associated?



# NFL 2018 DATA

- Let's load the `tidyverse` package library and read in the NFL data

```
library(tidyverse)  
  
load(url("http://bit.ly/nfl_18"))
```

# NFL 2018 DATA

- Let's take a look at the NFL data

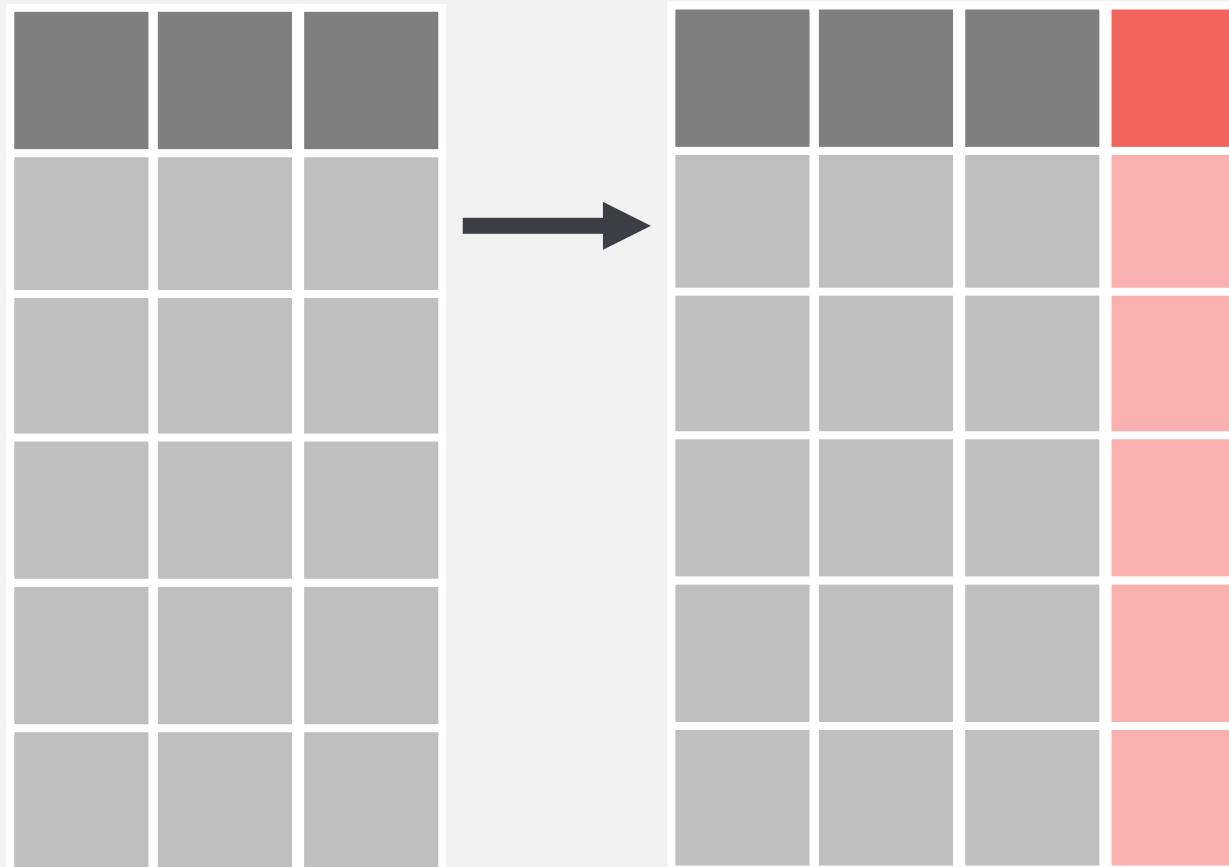
```
nfl
```

```
# A tibble: 32 x 12
  tm          w   l   t   ps   pa   pd   mov   sos   srs   osrs   dsrs
  <chr>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 New Orleans Saints    13     3     0   504   353   151    9.4    0.6   10.1    7.9     2.2
2 Los Angeles Rams      13     3     0   527   384   143    8.9   -0.4    8.5    9.5   -1.1
3 Kansas City Chiefs    12     4     0   565   421   144     9   -0.1    8.9   12.6   -3.8
4 Los Angeles Chargers   12     4     0   428   329    99    6.2   -0.2     6     3     2.9
5 Chicago Bears          12     4     0   421   283   138    8.6   -2.3    6.3    1.5     4.8
6 New England Patriots   11     5     0   436   325   111    6.9   -1.8    5.2    3.1     2.1
7 Houston Texans          11     5     0   402   316    86    5.4   -1.5    3.8    2.4     1.4
8 Baltimore Ravens        10     6     0   389   287   102    6.4    0.6     7    0.6     6.4
9 Indianapolis Colts      10     6     0   433   344    89    5.6   -2.2    3.4    3.9   -0.6
10 Dallas Cowboys         10     6     0   339   324    15    0.9    0.2    1.1   -1.9     2.9
# ... with 22 more rows
```

# 1. Mutate

Compute new column(s)

# 1. Mutate



# 1. Mutate

- Let's create a new column for teams' winning percentage

```
nfl
```

```
##   tm          w    l
## 1 New Orleans Saints 13  3
## 2 Los Angeles Rams   13  3
## 3 Kansas City Chiefs 12  4
## 4 Los Angeles Chargers 12  4
## 5 Chicago Bears      12  4
## . ...
## 32 Arizona Cardinals  3  13
```

→

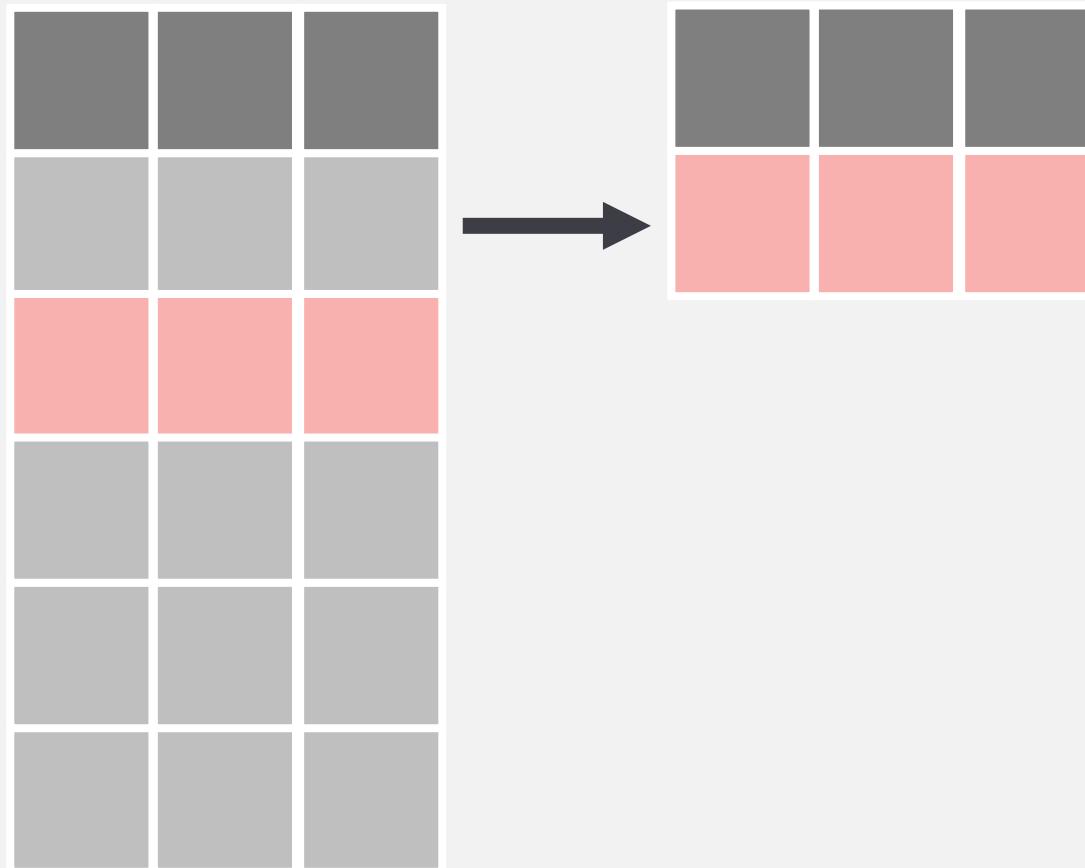
```
nfl <- nfl %>%
  mutate(wlpct = w / (w + l))
```

##	tm	w	l	wlpct
1	New Orleans Saints	13	3	0.812
2	Los Angeles Rams	13	3	0.812
3	Kansas City Chiefs	12	4	0.750
4	Los Angeles Chargers	12	4	0.750
5	Chicago Bears	12	4	0.750
.	...	.	.	.
32	Arizona Cardinals	3	13	0.188

# 2. Filter

Extract rows that meet logical criteria

# 2. Filter



# 2. Filter

- Let's filter out and keep only teams with winning percentages above 0.800

```
nfl_18
```

```
##   tm          w   l wlpct
## 1 New Orleans Saints 13  3 0.812
## 2 Los Angeles Rams   13  3 0.812
## 3 Kansas City Chiefs 12  4 0.750
## 4 Los Angeles Chargers 12  4 0.750
## 5 Chicago Bears      12  4 0.750
## ...
## 32 Arizona Cardinals  3   13 0.188
```



```
nfl_18 %>%
  filter(wlpct > 0.8))
```

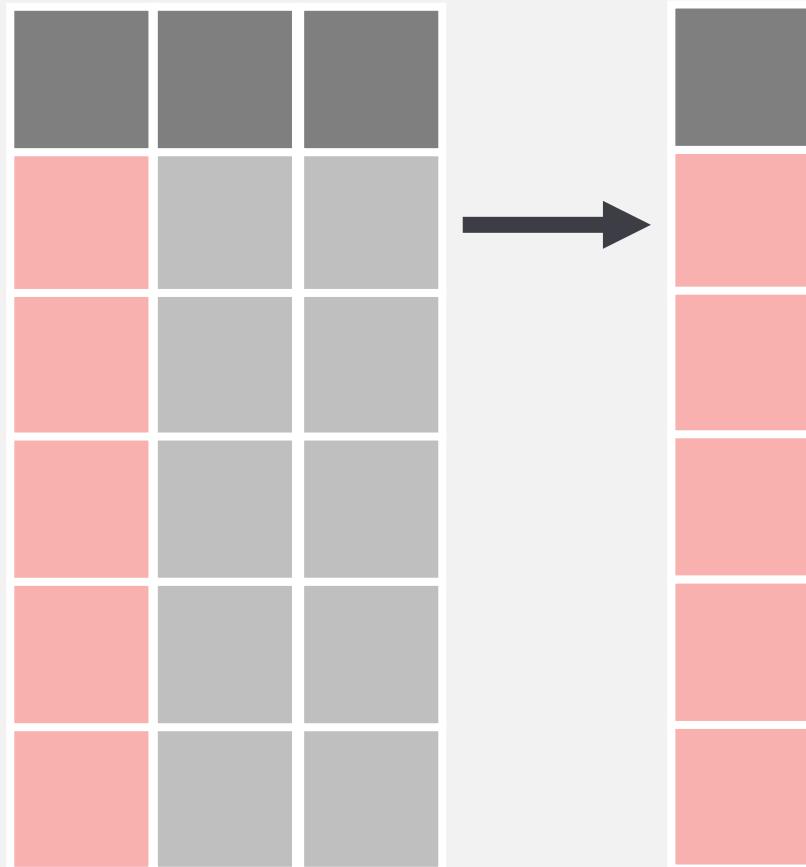
```
##   tm          w   l wlpct
## 1 New Orleans Saints 13  3 0.812
## 2 Los Angeles Rams   13  3 0.812
```

w l wlpct

# 3. Select

Extract columns as a table

# 3. Select



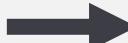
# 3. Select

- Let's get rid of w and l, keeping only wlpct

```
nfl_18
```

```
##      tm          w      l    wlpct
## 1 New Orleans Saints   13    3  0.812
## 2 Los Angeles Rams    13    3  0.812
## 3 Kansas City Chiefs   12    4  0.750
## 4 Los Angeles Chargers  12    4  0.750
## 5 Chicago Bears        12    4  0.750
## . ...
## 32 Arizona Cardinals    3   13  0.188
```

```
nfl_18 %>%
  select(wlpct)
```

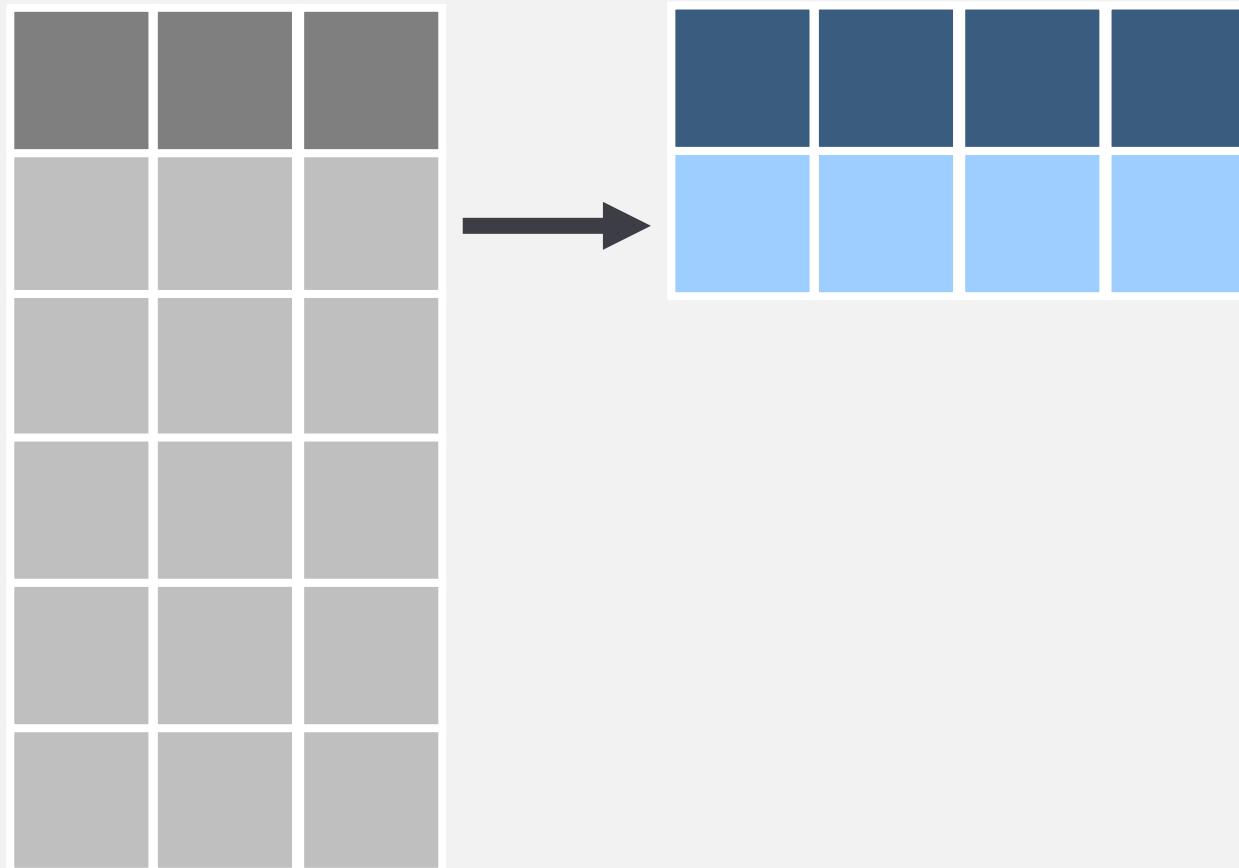


```
##      tm    wlpct
## 1 New Orleans Saints 0.812
## 2 Los Angeles Rams 0.812
## 3 Kansas City Chiefs 0.750
## 4 Los Angeles Chargers 0.750
## 5 Chicago Bears 0.750
## . ...
## 32 Arizona Cardinals 0.188
```

# 4. Summarize

Compute table of summaries

# 4. Summarize



# 4. Summarize

- Let's calculate the mean number of points scored per game, over the course of the season, across all teams

```
nfl_18
```

```
##   tm          w   l   ps
## 1 New Orleans Saints 13  3  504
## 2 Los Angeles Rams   13  3  527
## 3 Kansas City Chiefs 12  4  565
## 4 Los Angeles Chargers 12  4  428
## 5 Chicago Bears      12  4  421
## ...
## 32 Arizona Cardinals  3   13 225
```

```
nfl_18 %>%
  summarise(ppg = mean(ps) / 16)
```



```
## # A tibble: 1 x 1
##   ppg
##   1 23.3
```

# 4. Visualize

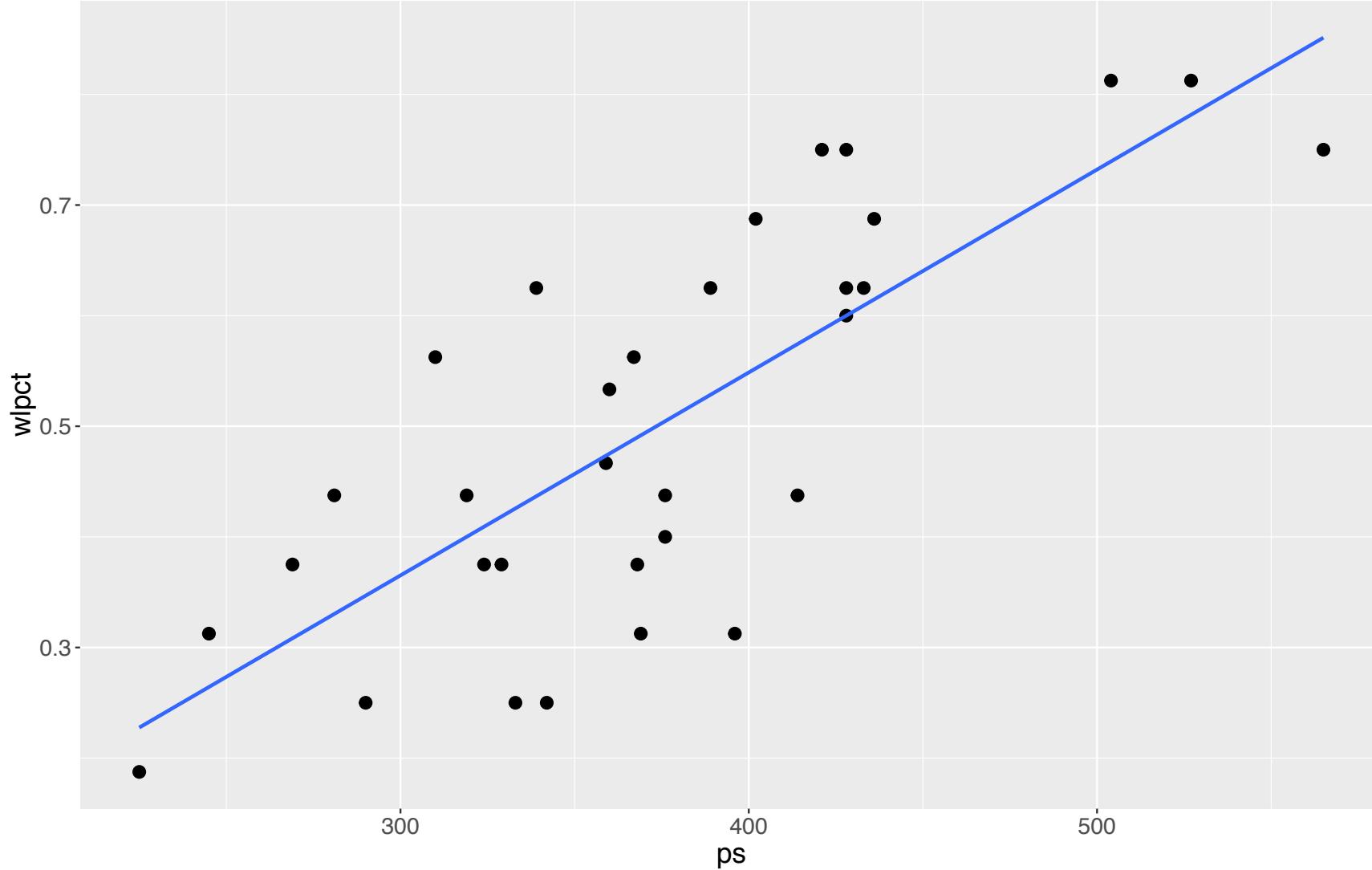
Plot distributions, relationships, and factors

# 5. Visualize

- Let's visualize the association between winning percentage and points scored over the season

```
ggplot(data = nfl,  
aes(x = ps, y = wlpct)) +  
geom_point() +  
geom_smooth(method = "lm", se = FALSE)
```

# 5. Visualize



# RECAP

# RECAP

- Data analysis often involves “tidying” the data before visualizing and modeling
- The verbs of the tidyverse are modern tools
- Applicable to any data analysis workflow



# QUESTIONS?

# GET IN TOUCH

**Aaron R. Baggett, Ph.D.**

Department of Psychology

[abaggett@umhb.edu](mailto:abaggett@umhb.edu)

Ext. 4553

# APPENDIX

# DATA SCIENCE WORKFLOW

- The following slide contains a basic workflow for tackling data analysis projects
- For each step in the workflow, the corresponding tidyverse packages are listed
- For more information about the tidyverse see:

<https://www.tidyverse.org/>

