CORRELATION AND REGRESSION

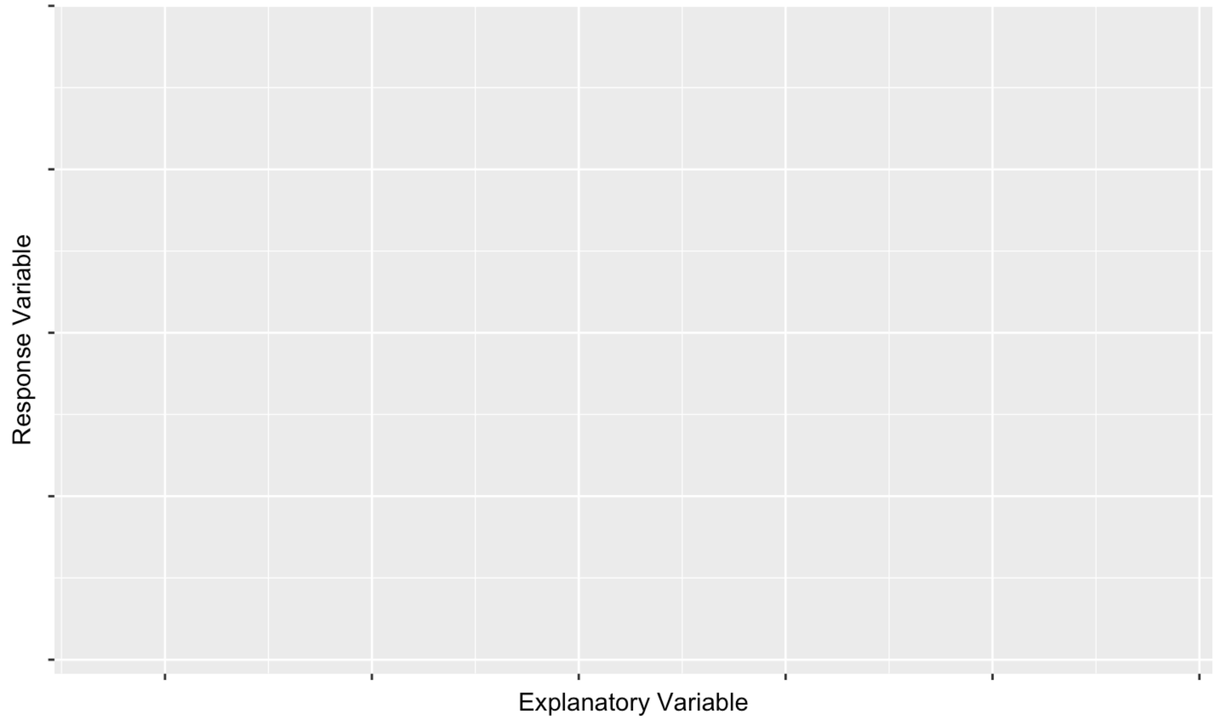
Modeling bivariate relationships

Bivariate relationships

- Both variables are numerical
- Response variable
  - a.k.a. *y*, dependent
- Explanatory variable
  - Something you think might be related to the response
  - a.k.a. *x*, independent, predictor

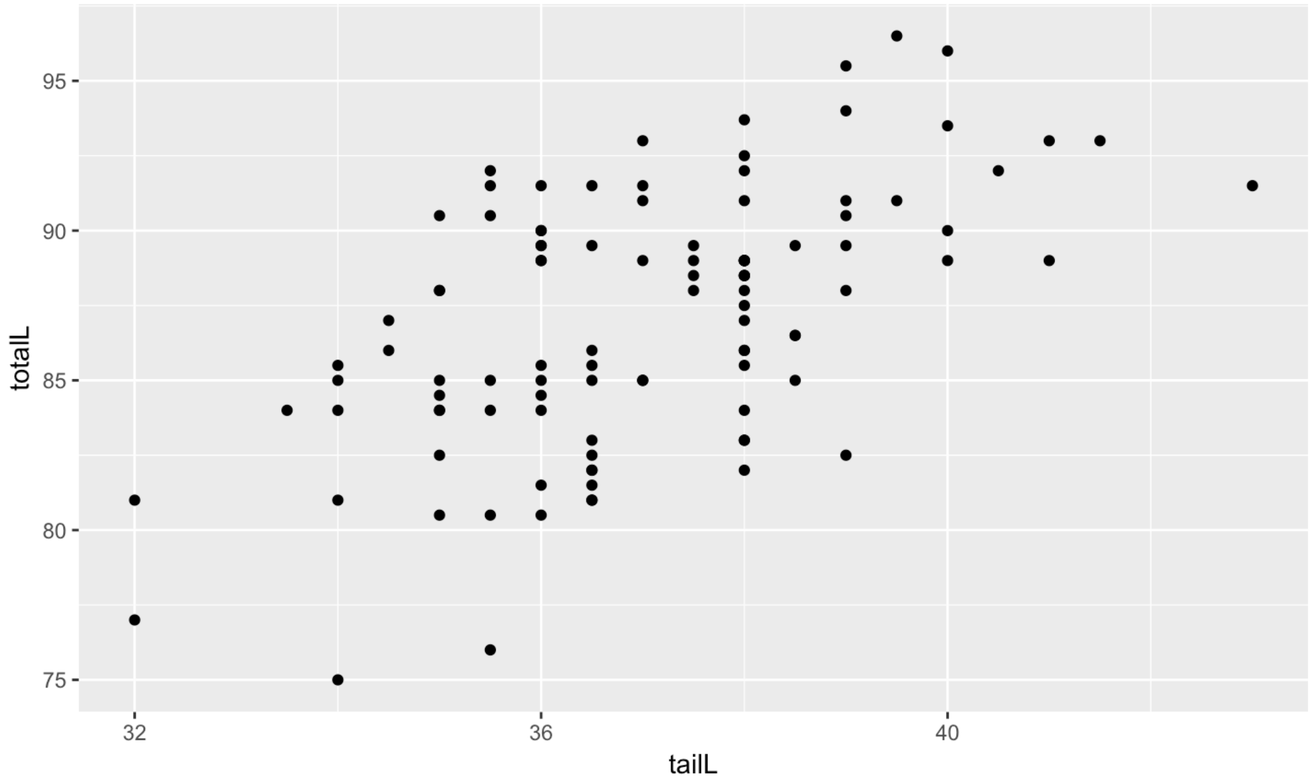
Graphical representations

- Put response on vertical axis
- Put explanatory on horizontal axis



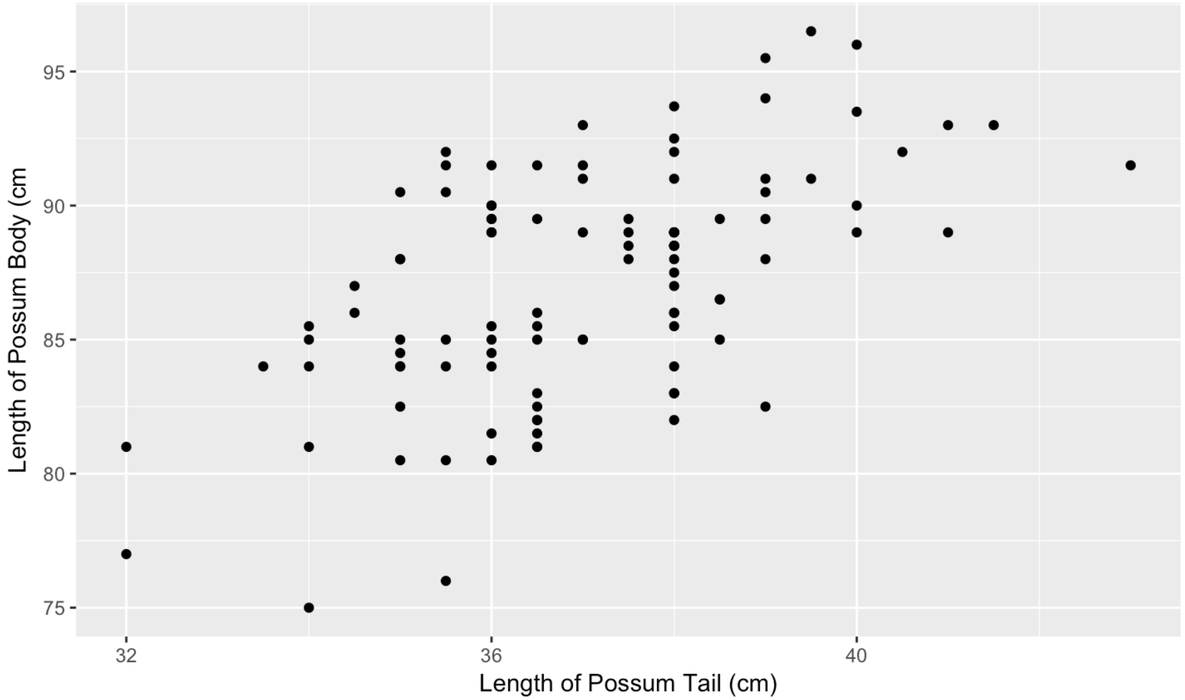
Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point()
```



Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() +
    scale_x_continuous("Length of Possum Tail (cm)") +
    scale_y_continuous("Length of Possum Body (cm)")
```

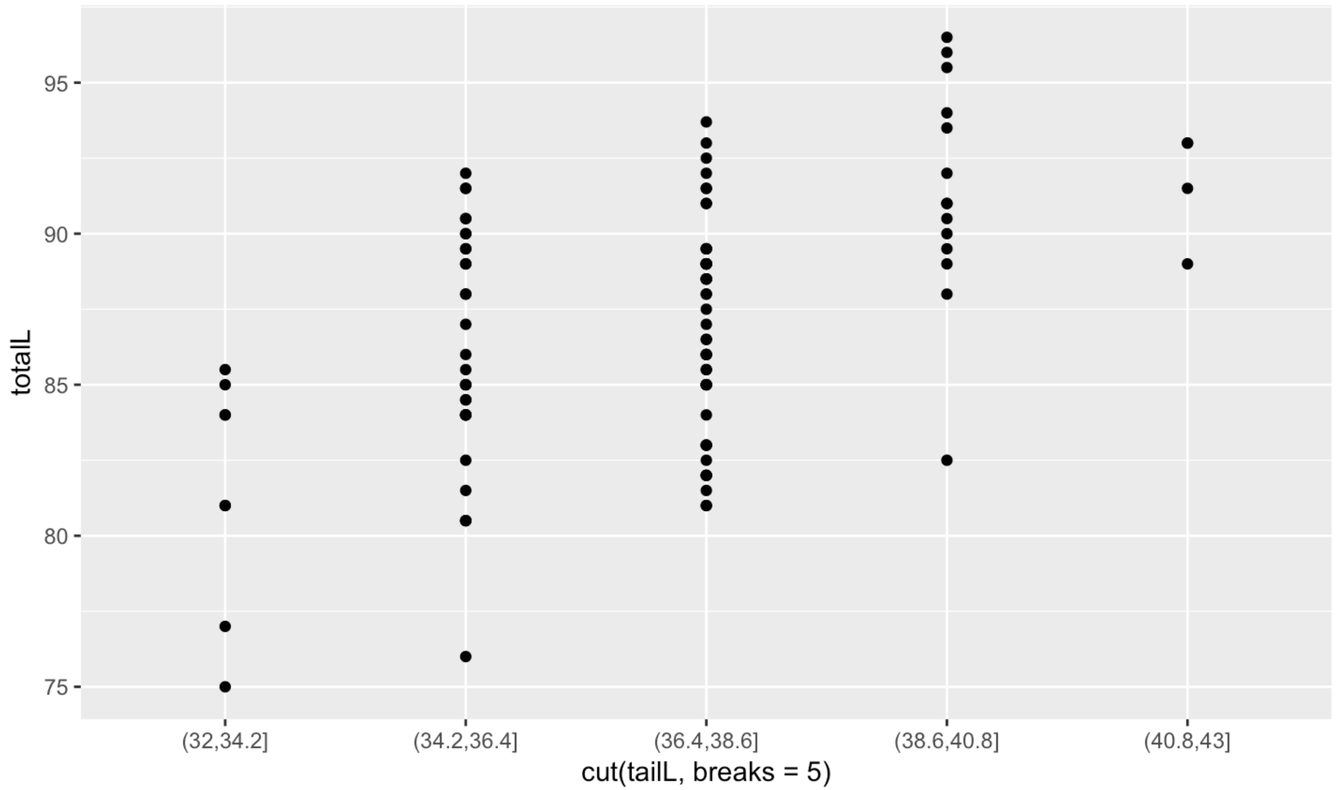


Bivariate relationships

- Can think of boxplots as scatterplots...
  - ...but with discretized explanatory variable
- `cut()` function discretizes
  - Choose appropriate number of "boxes"

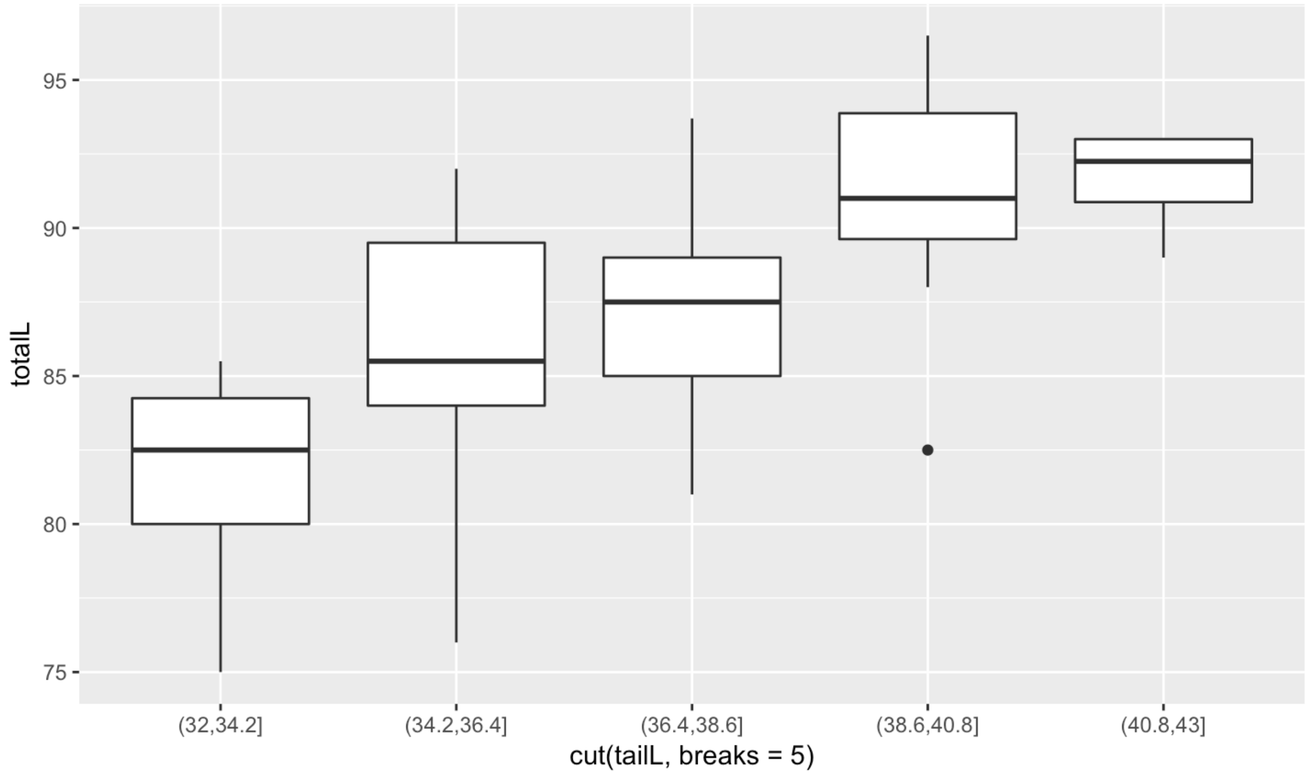
Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = cut(tailL, breaks = 5))) +
    geom_point()
```



Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = cut(tailL, breaks = 5))) +
    geom_boxplot()
```



CORRELATION AND REGRESSION

Let’s practice!

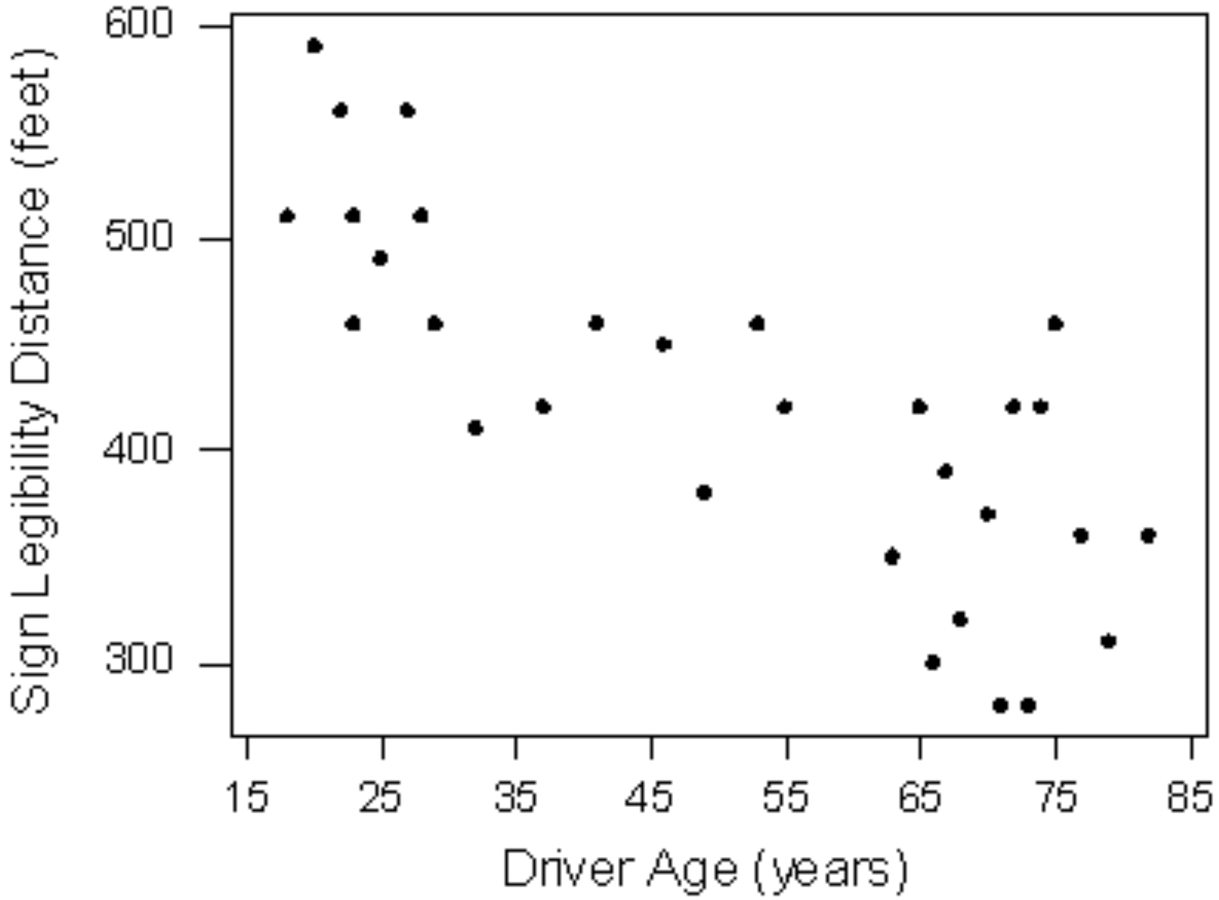
CORRELATION AND REGRESSION

Characterizing bivariate relationships

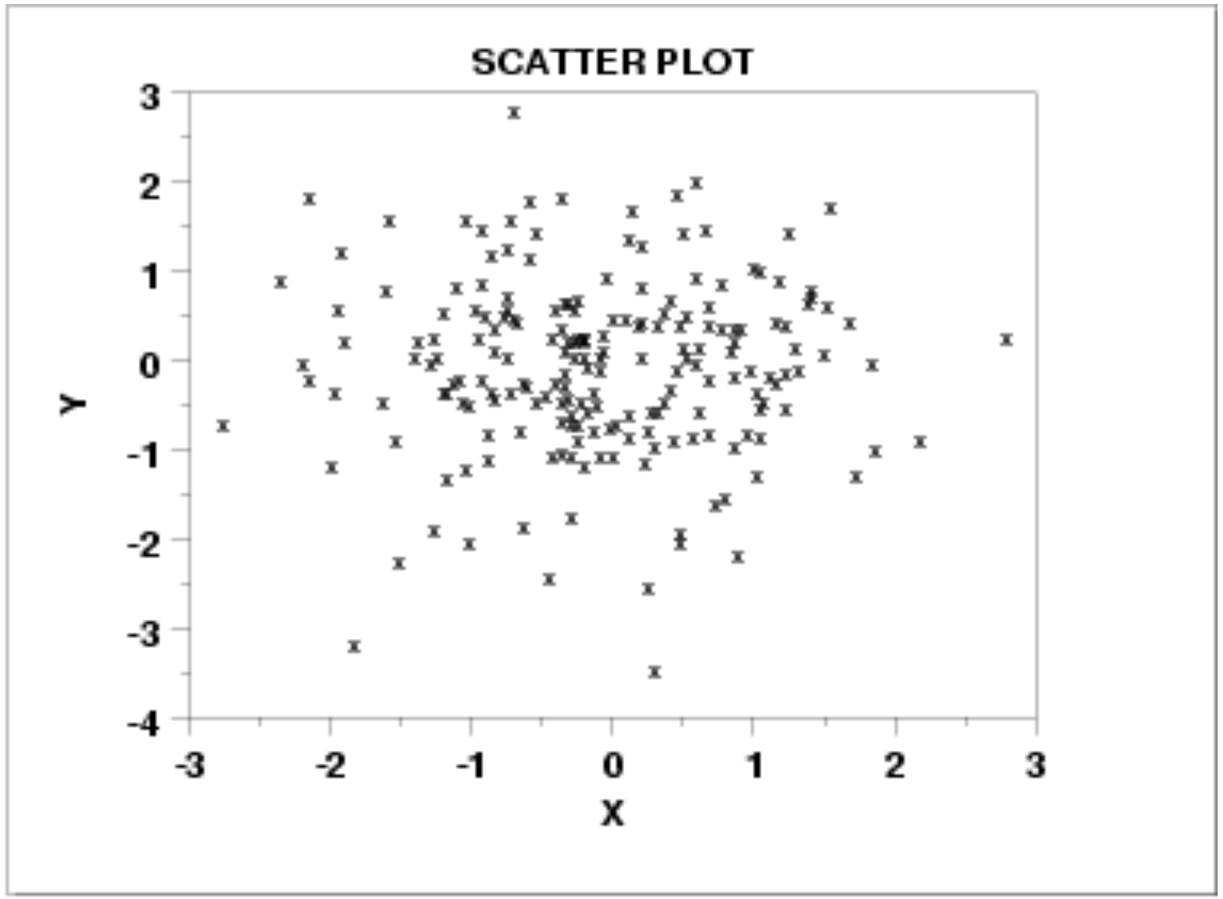
Characterizing bivariate relationships

- Form (e.g. linear, quadratic, non-linear)
- Direction (e.g. positive, negative)
- Strength (how much scatter/noise?)
- Outliers

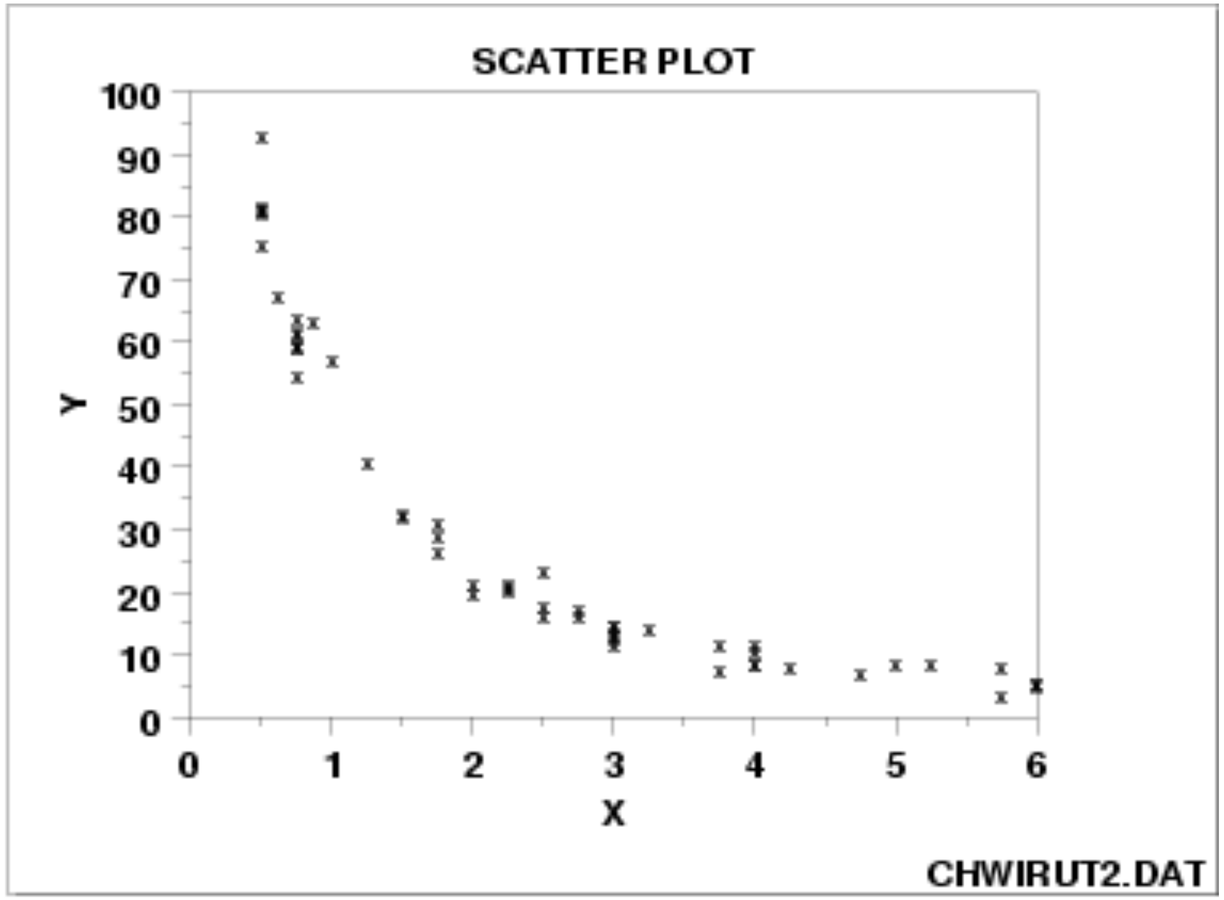
Sign legibility



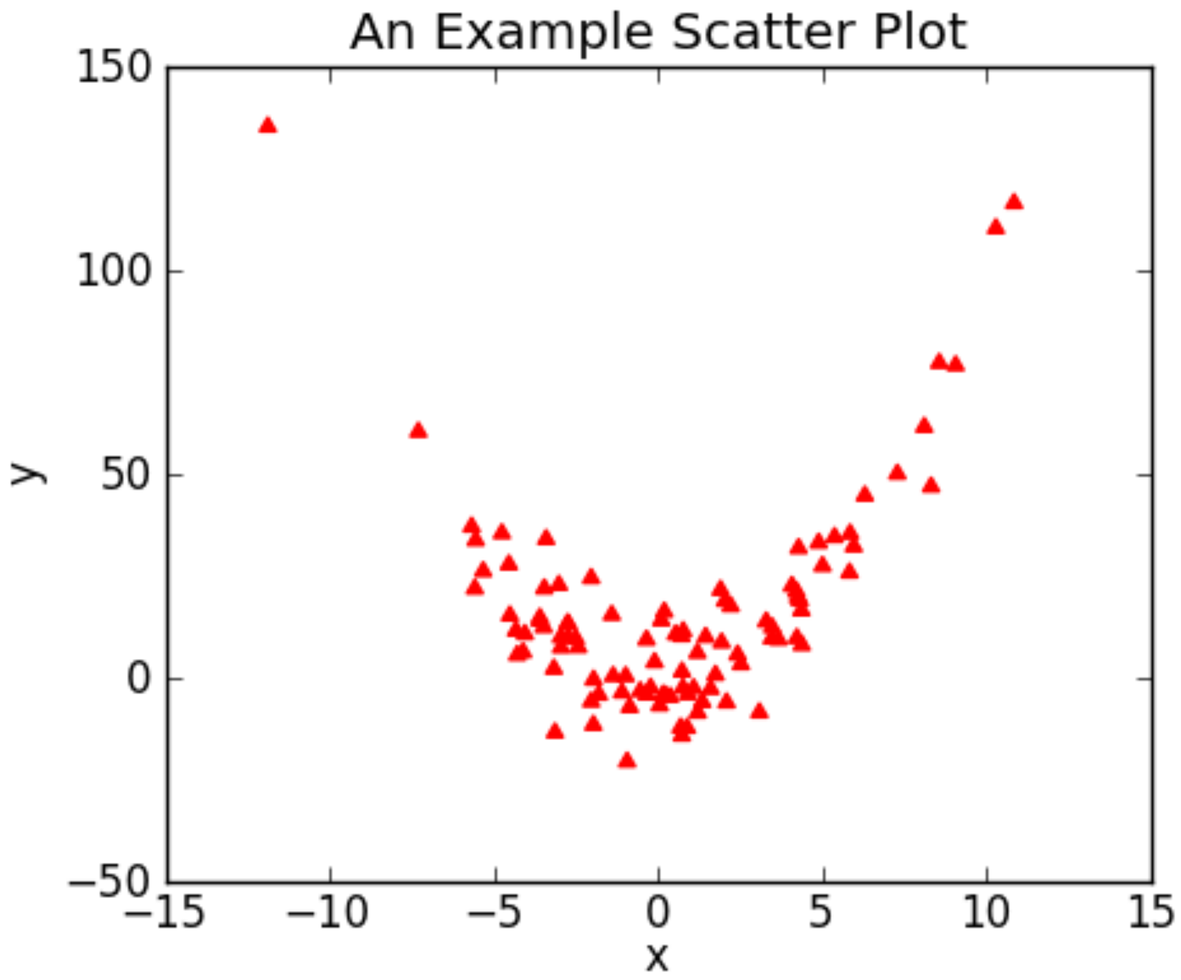
NIST



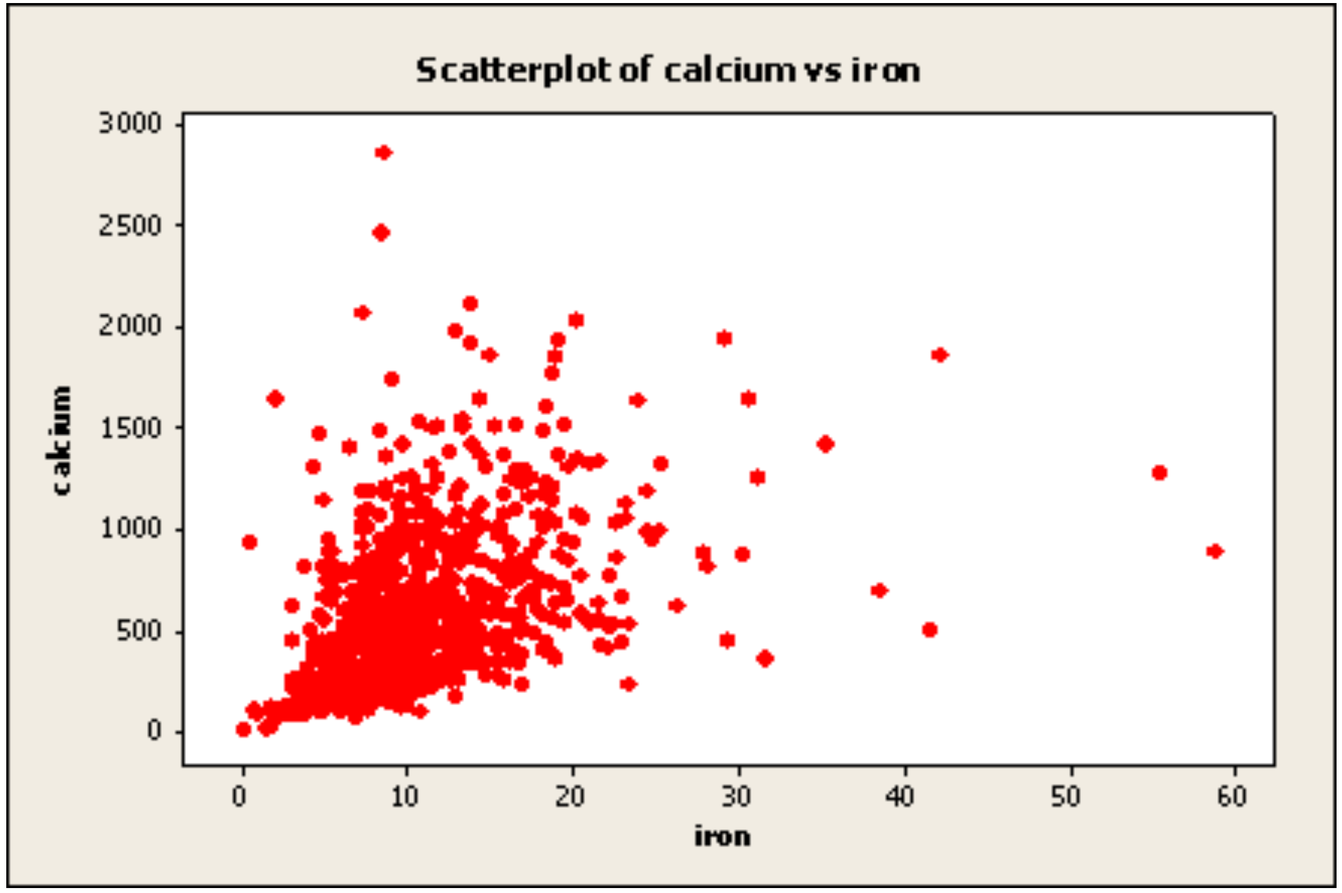
NIST 2


SCATTER PLOT
100
90
80
70
60
50
40
30
20
10
0
Y
0
1
2
3
4
5
6
X
CHWIRUT2.DAT

Non-linear



Fan shape


Scatterplot of calcium vs iron
calcium
iron
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60

CORRELATION AND REGRESSION

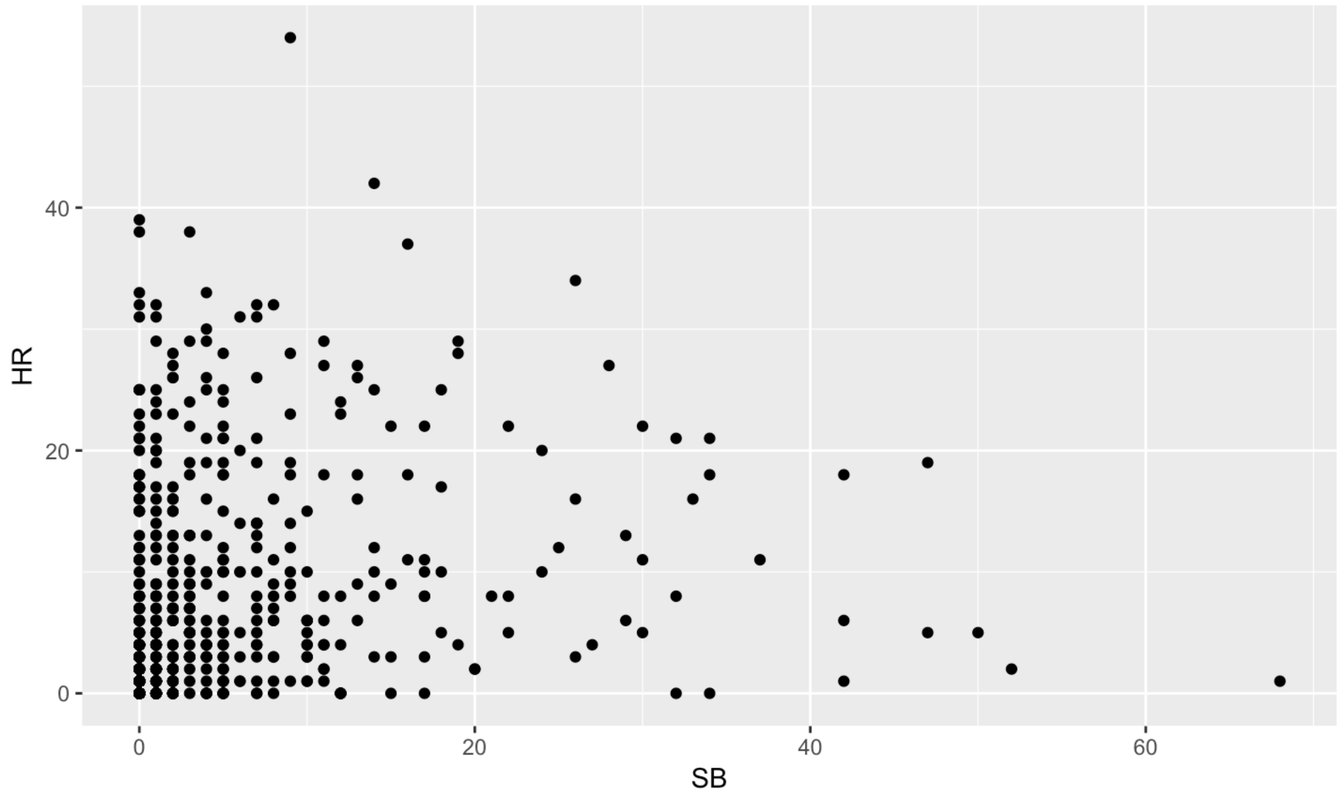
Let’s practice!

CORRELATION AND REGRESSION

Outliers

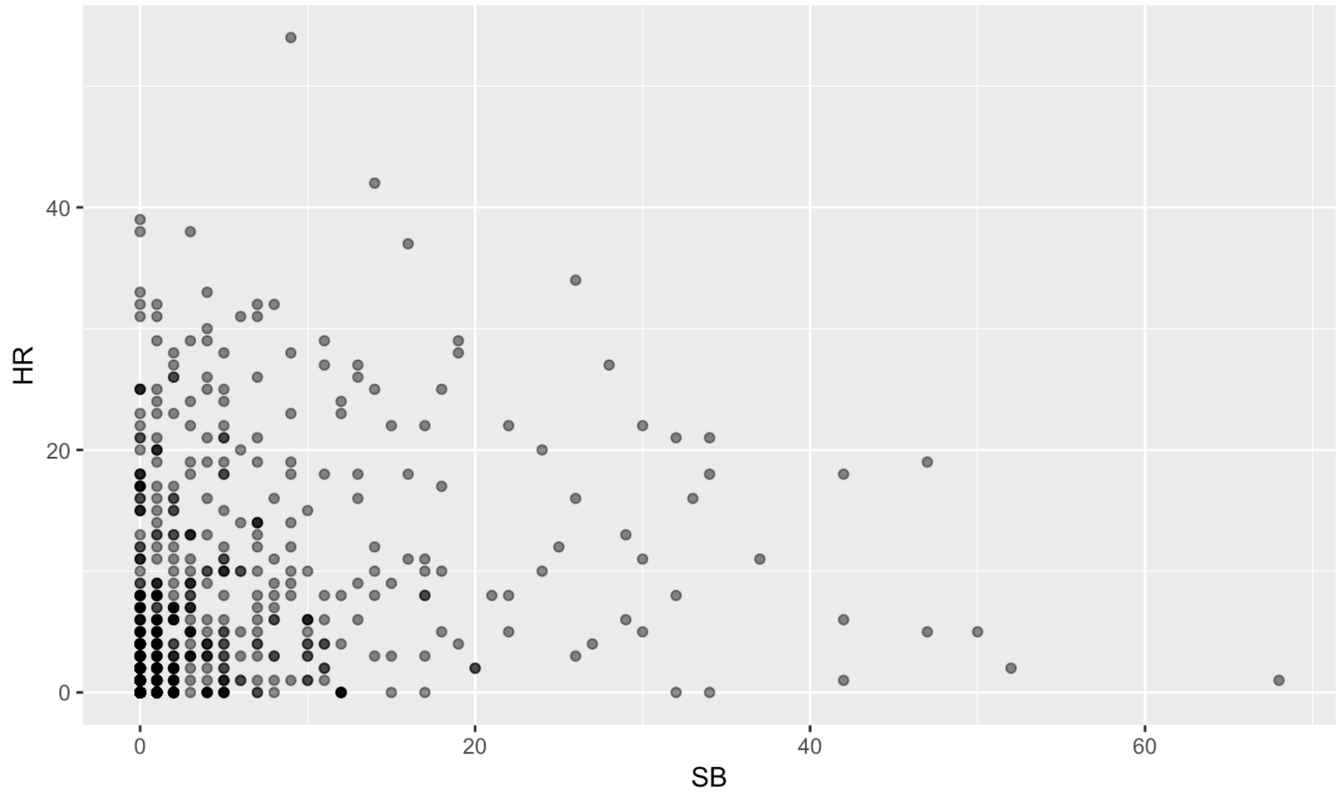
Outliers

```
> ggplot(data = mlbBat10, aes(x = SB, y = HR)) +
    geom_point()
```



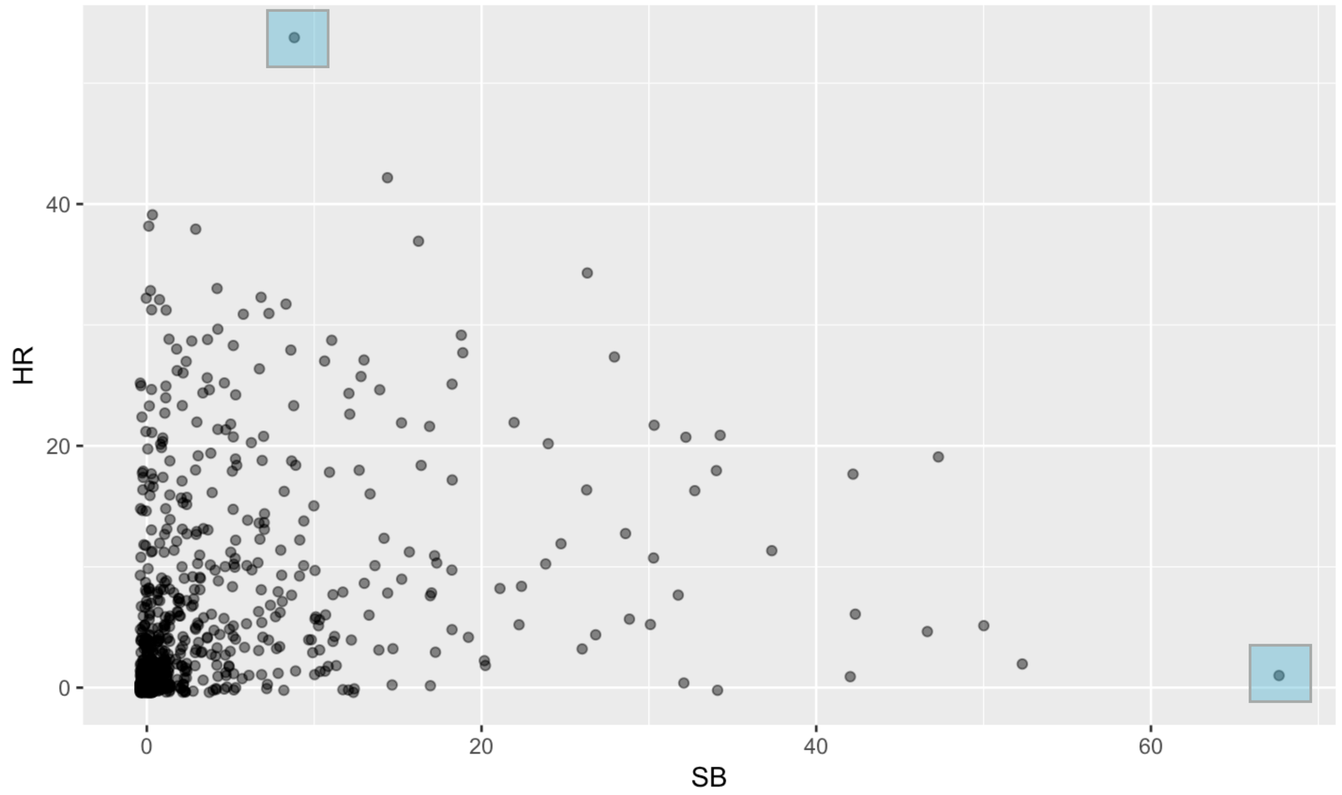
Add transparency

```
> ggplot(data = mlbBat10, aes(x = SB, y = HR)) +
    geom_point(alpha = 0.5)
```



Add some jitter

```
> ggplot(data = mlbBat10, aes(x = SB, y = HR)) +
    geom_point(alpha = 0.5, position = "jitter")
```



Identify the outliers

```
> mlbBat10 %>%
    filter(SB > 60 | HR > 50) %>%
    select(name, team, position, SB, HR)

  ##          name team position SB HR
  ## 1   J Pierre  CWS       OF 68  1
  ## 2 J Bautista  TOR       OF  9 54
```

CORRELATION AND REGRESSION

Let’s practice!