

# Visualizing bivariate relationships

CORRELATION AND REGRESSION IN R



**Ben Baumer**

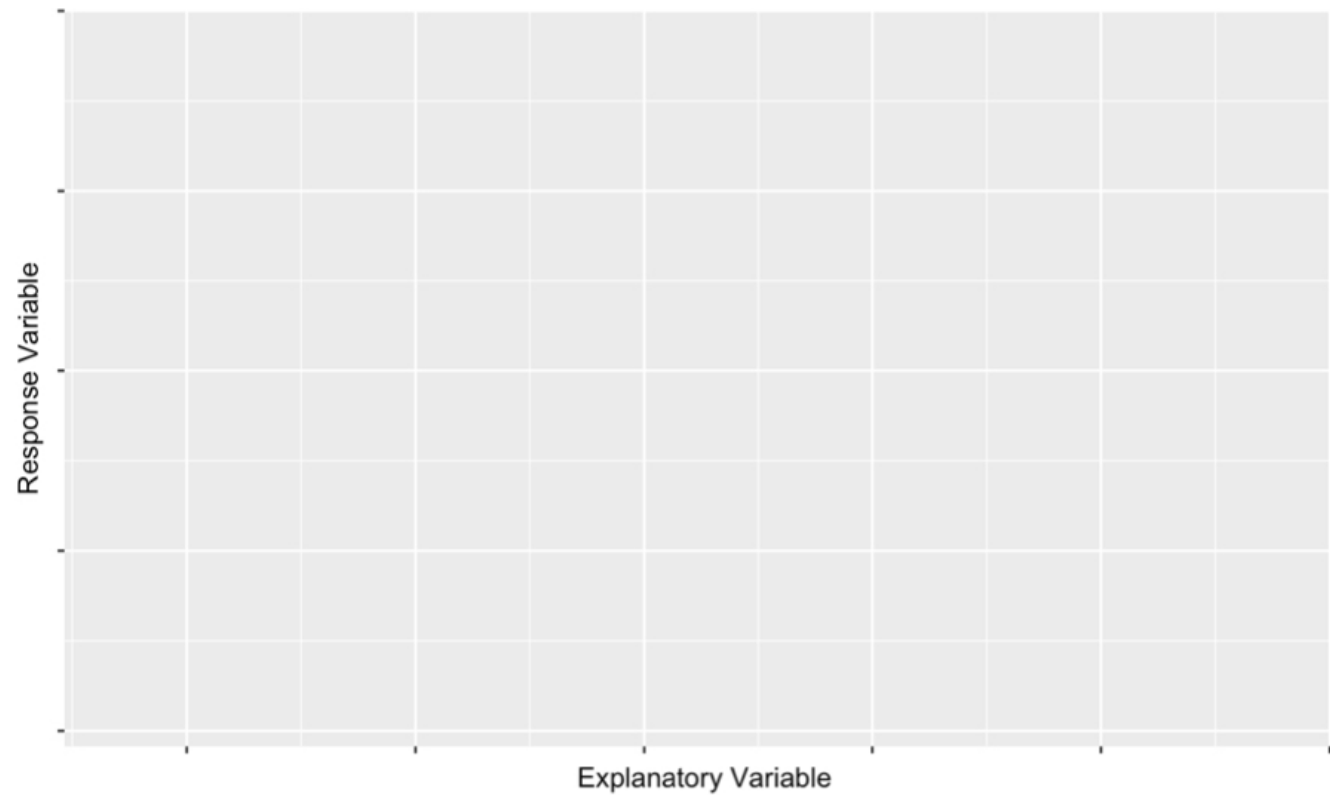
Assistant Professor at Smith College

# 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

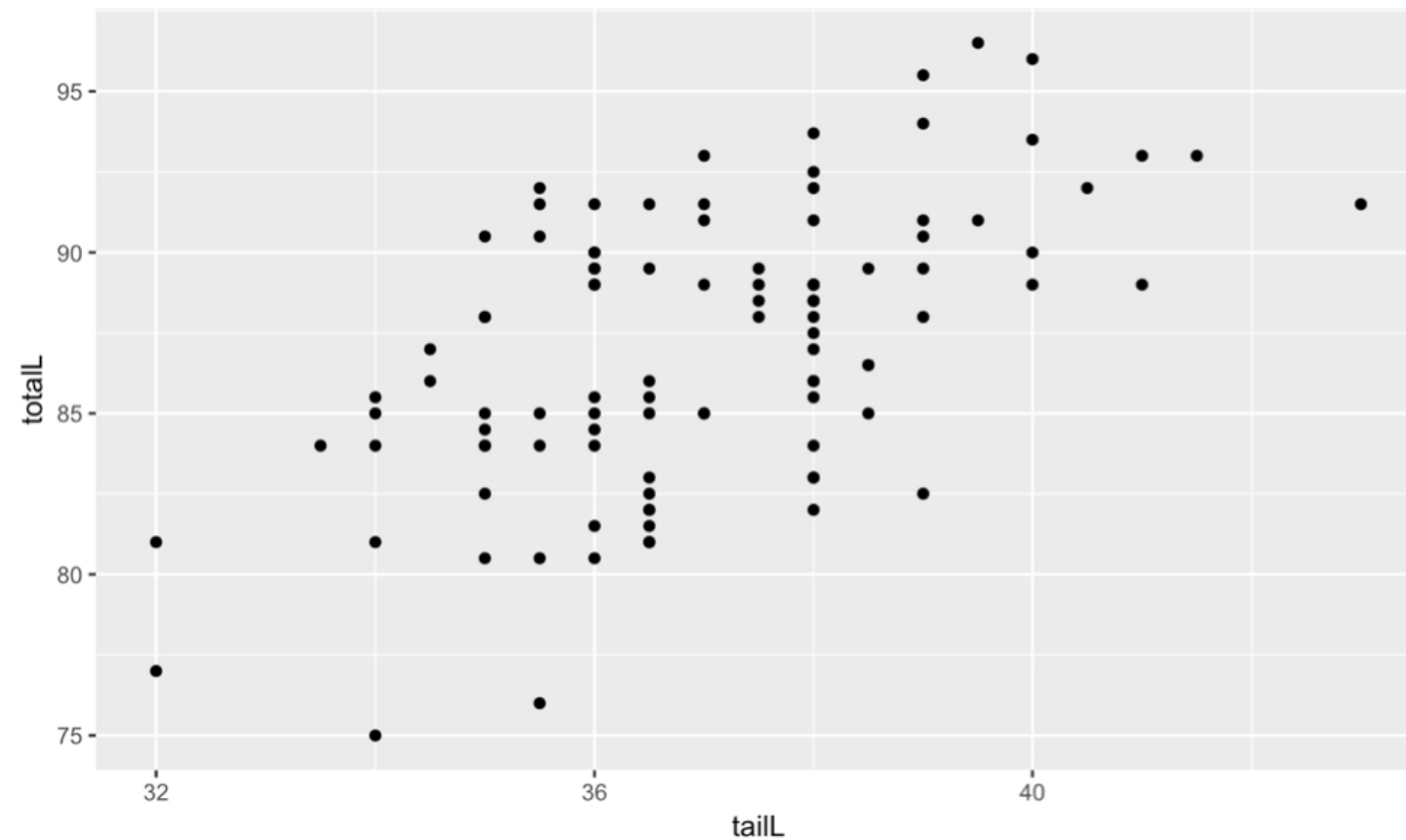
# Graphical representations

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



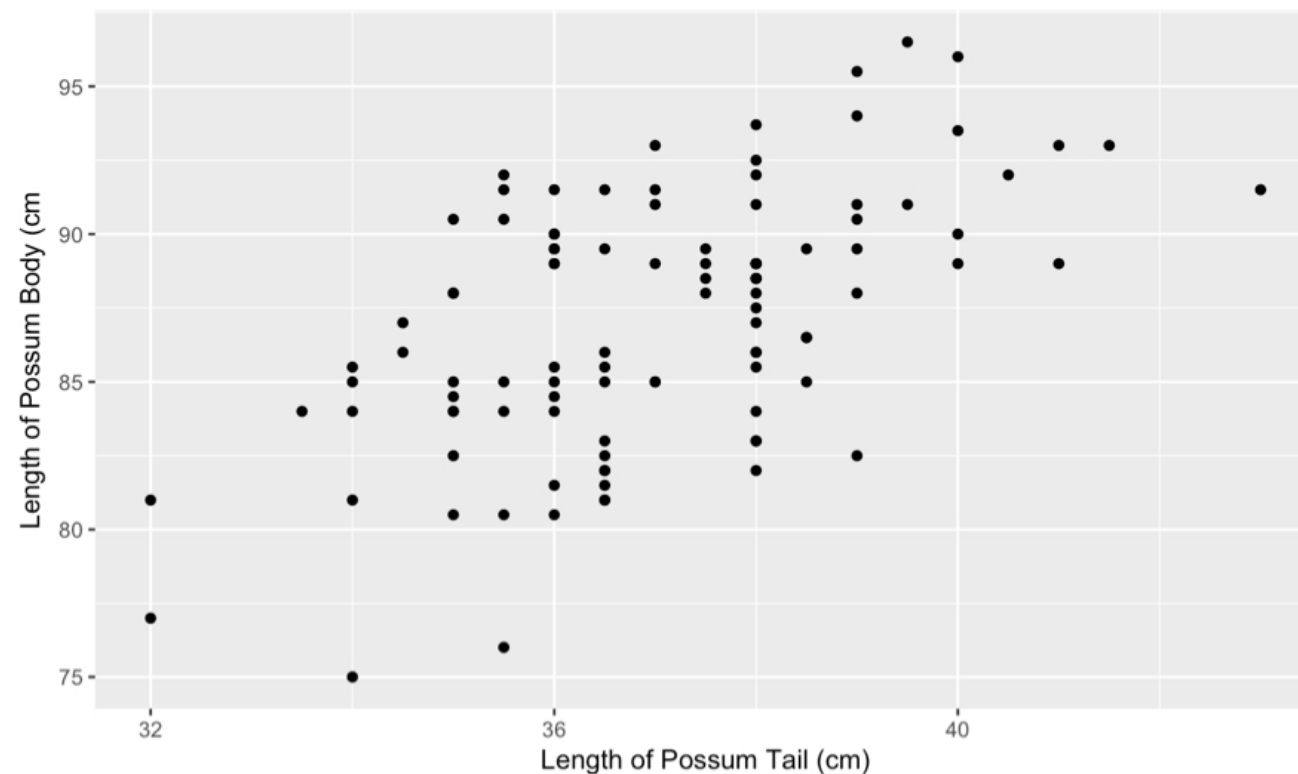
# Scatterplot

```
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)")
```

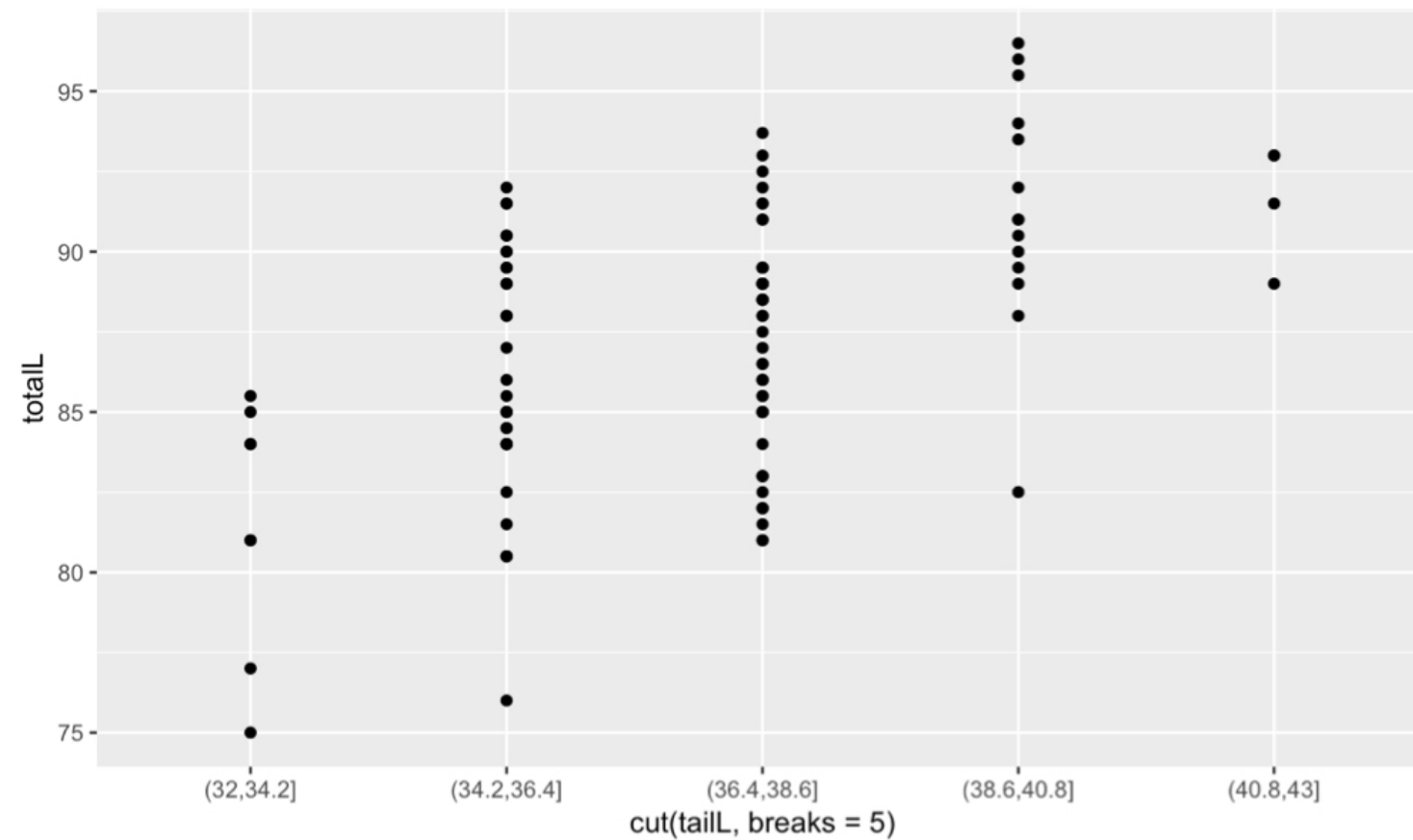


# Bivariate relationships

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

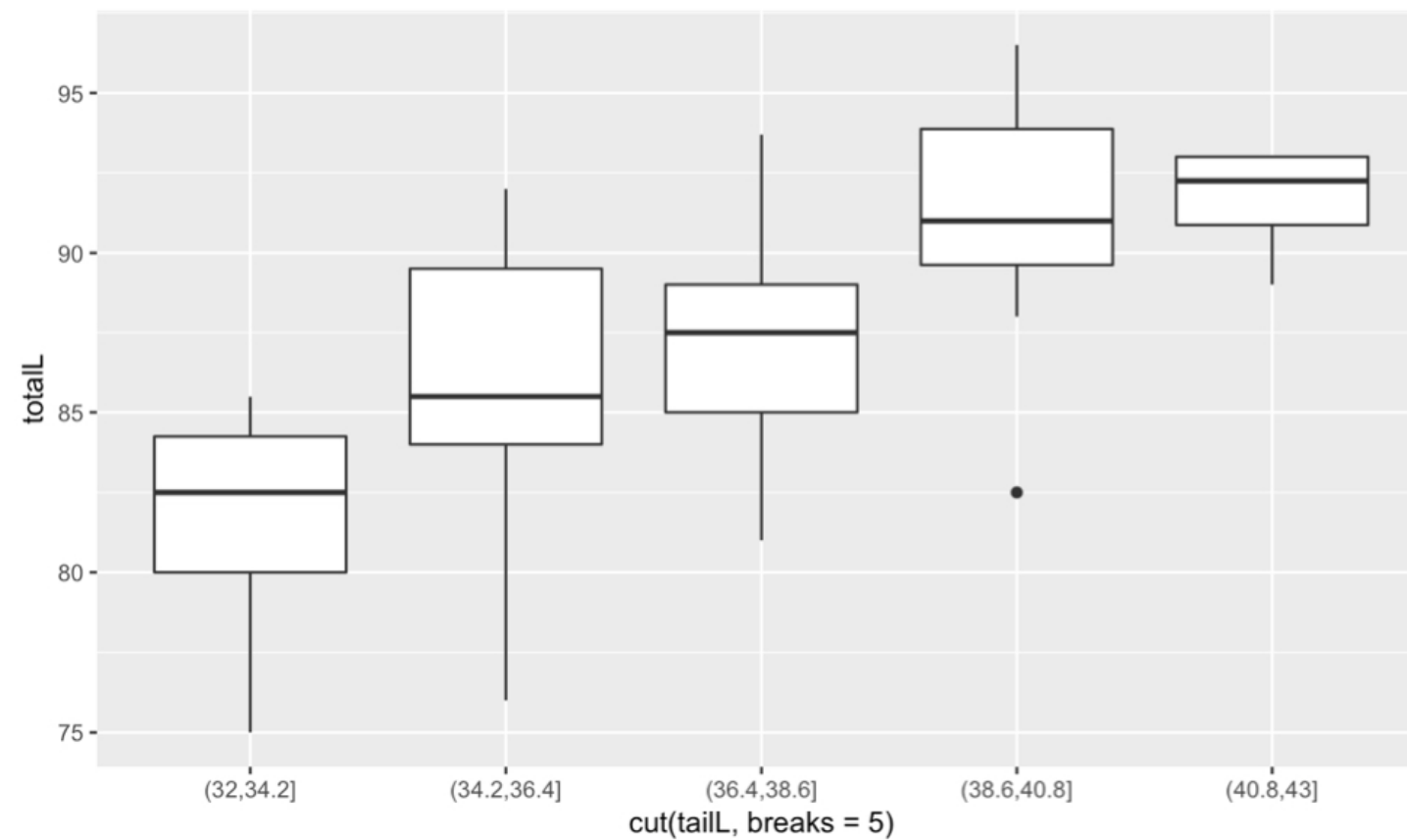
# Scatterplot

```
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()
```





# Let's practice!

CORRELATION AND REGRESSION IN R

# Characterizing bivariate relationships

CORRELATION AND REGRESSION IN R

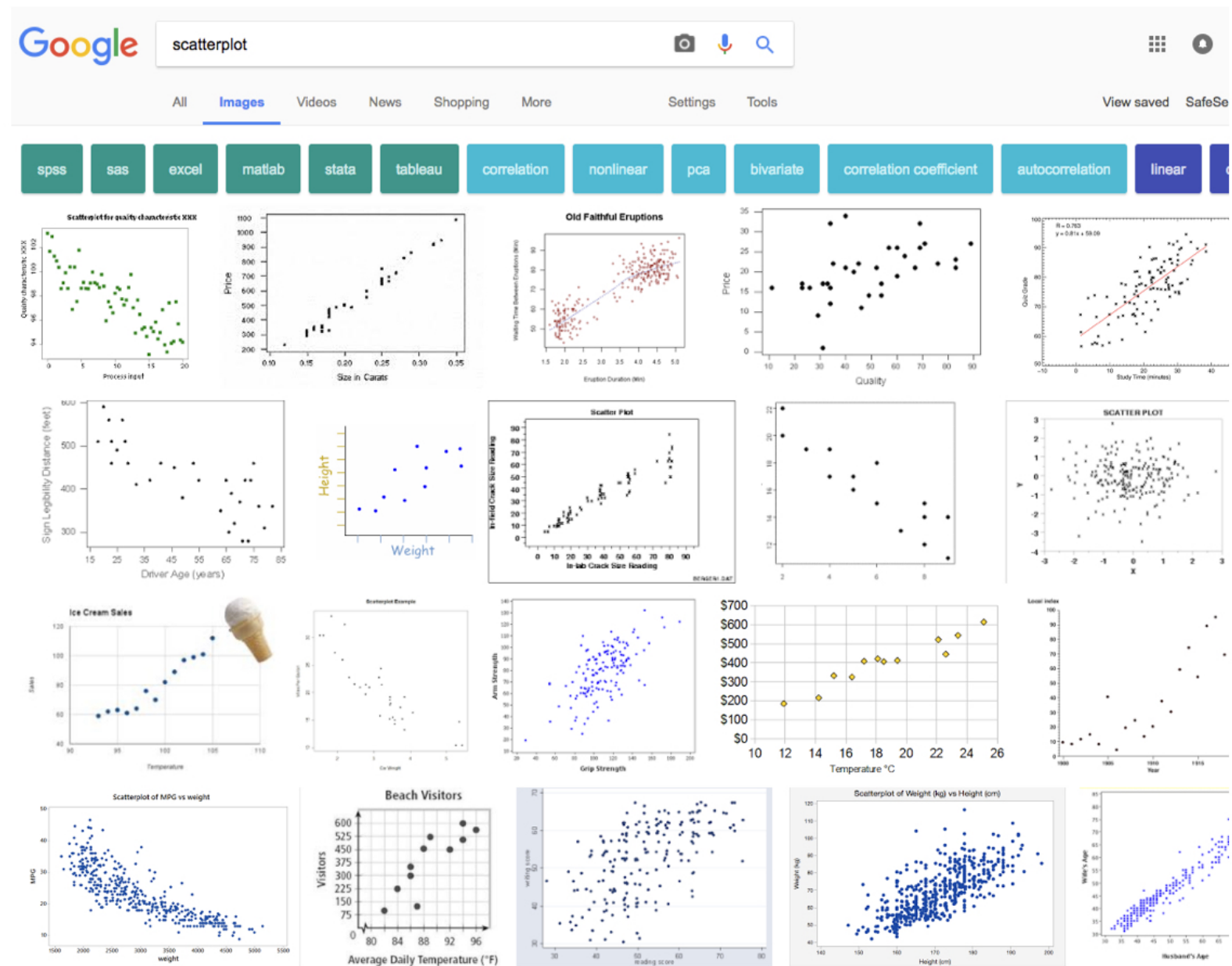


**Ben Baumer**

Assistant Professor at Smith College

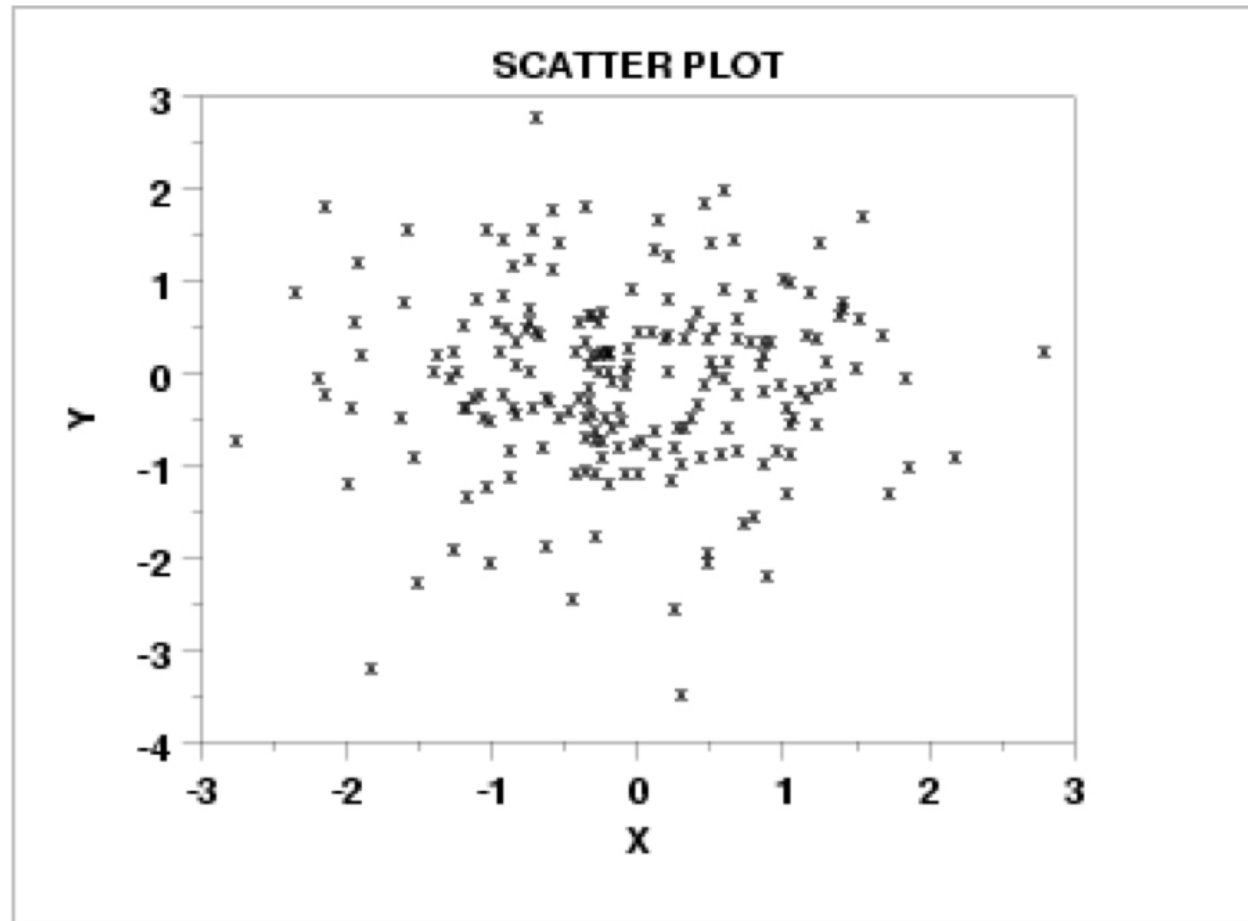
# Characterizing bivariate relationships

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

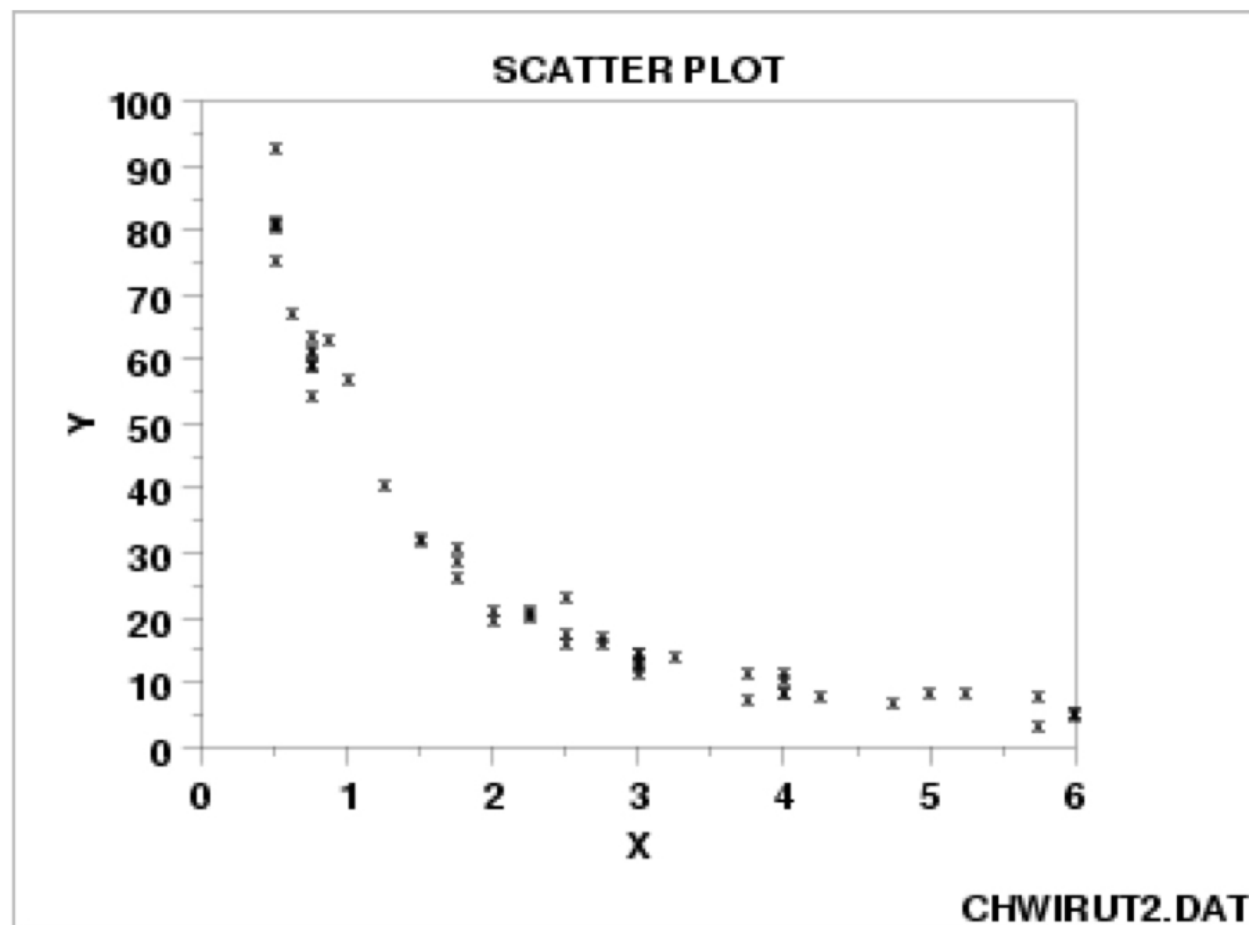


# Sign legibility

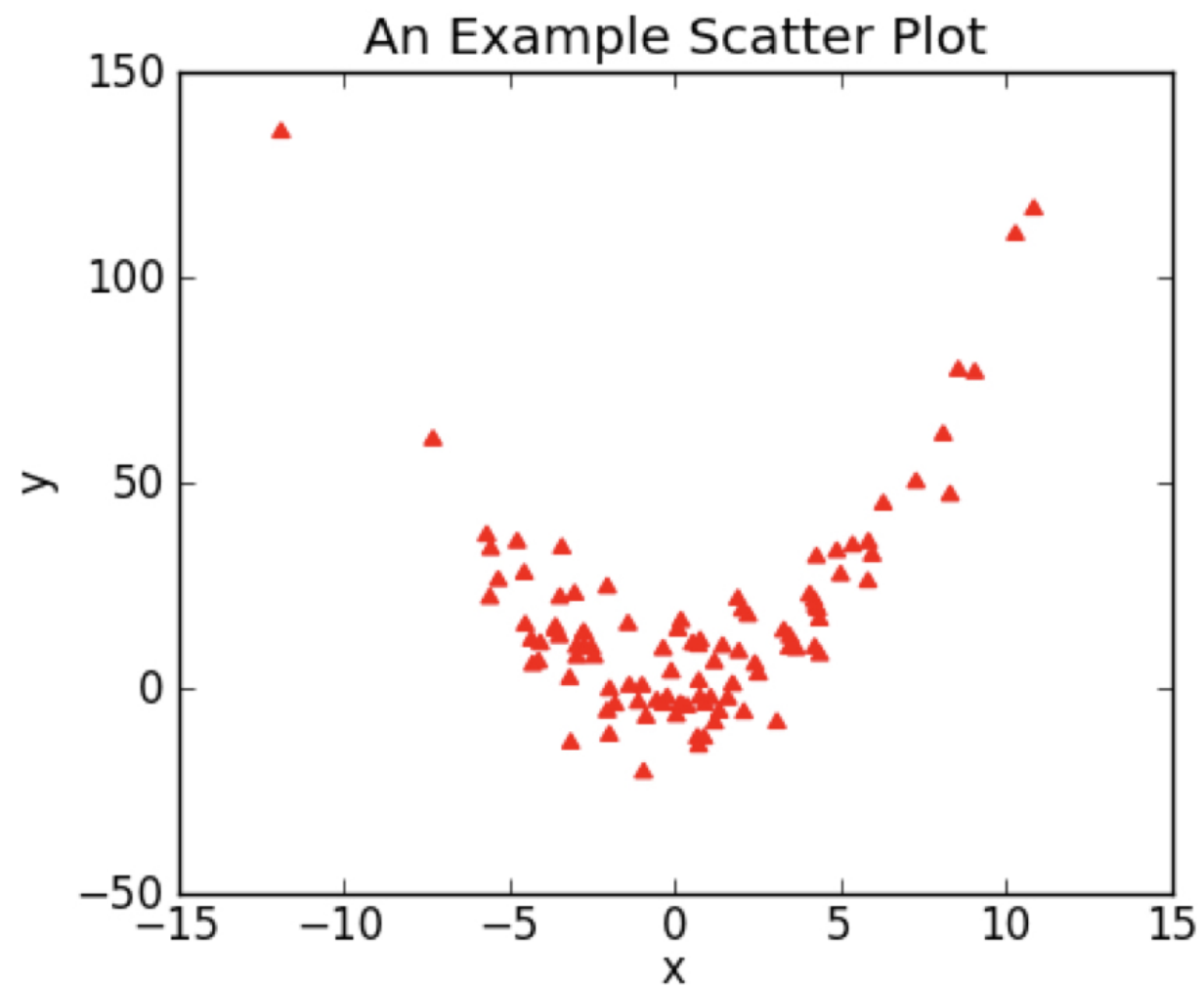




# NIST 2

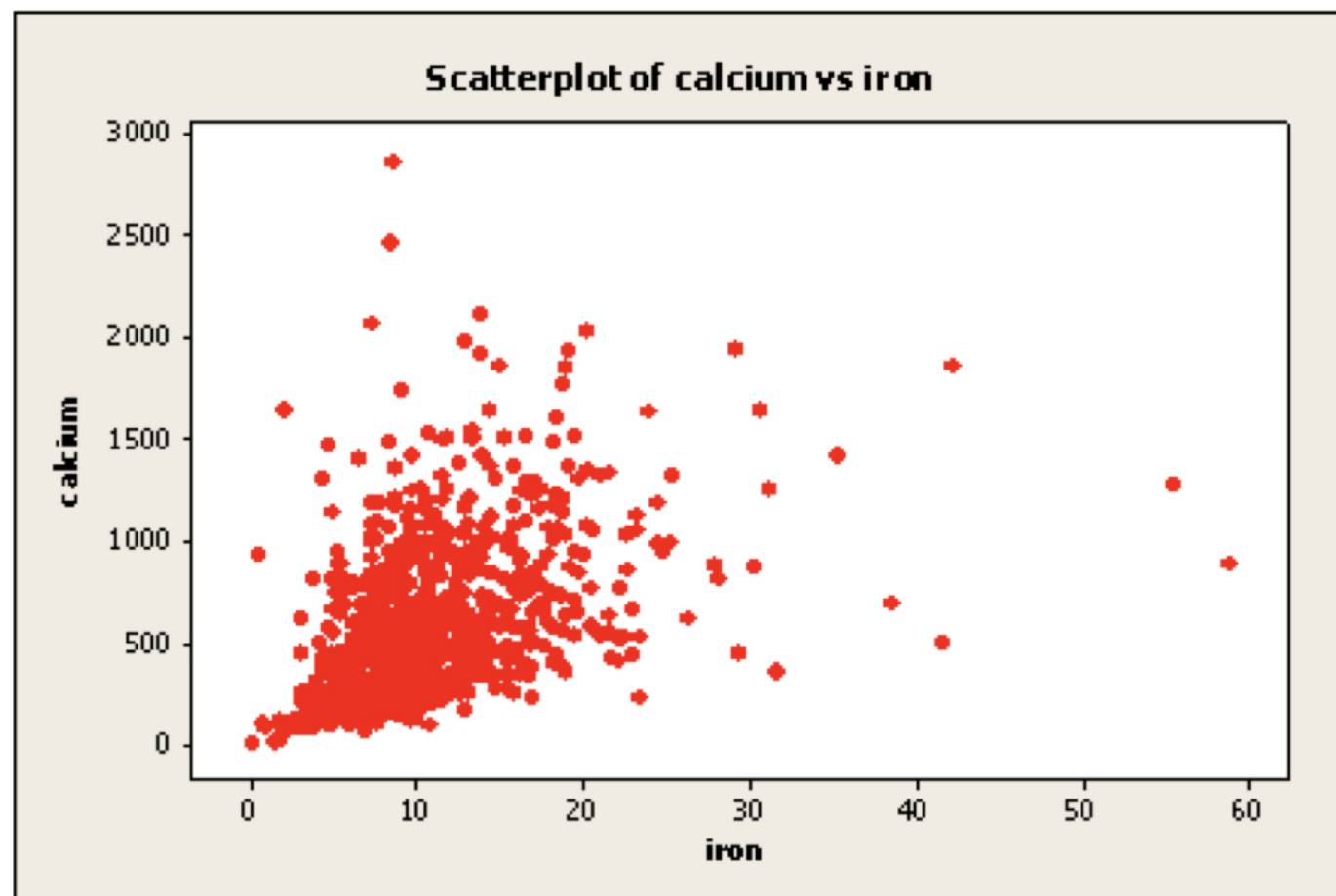


# Non-linear





# Fan shape



# Let's practice!

CORRELATION AND REGRESSION IN R

# Outliers

CORRELATION AND REGRESSION IN R

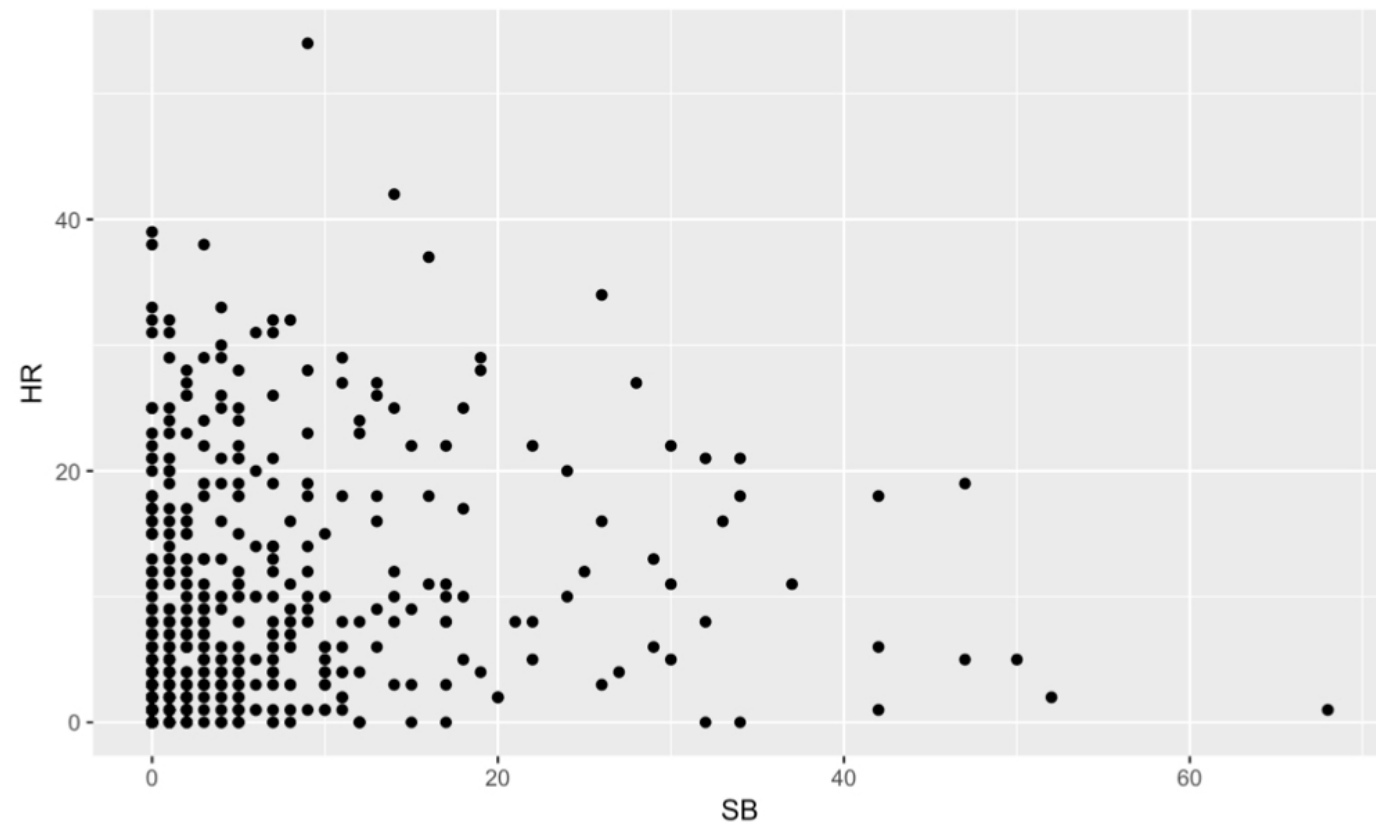


**Ben Baumer**

Assistant Professor at Smith College

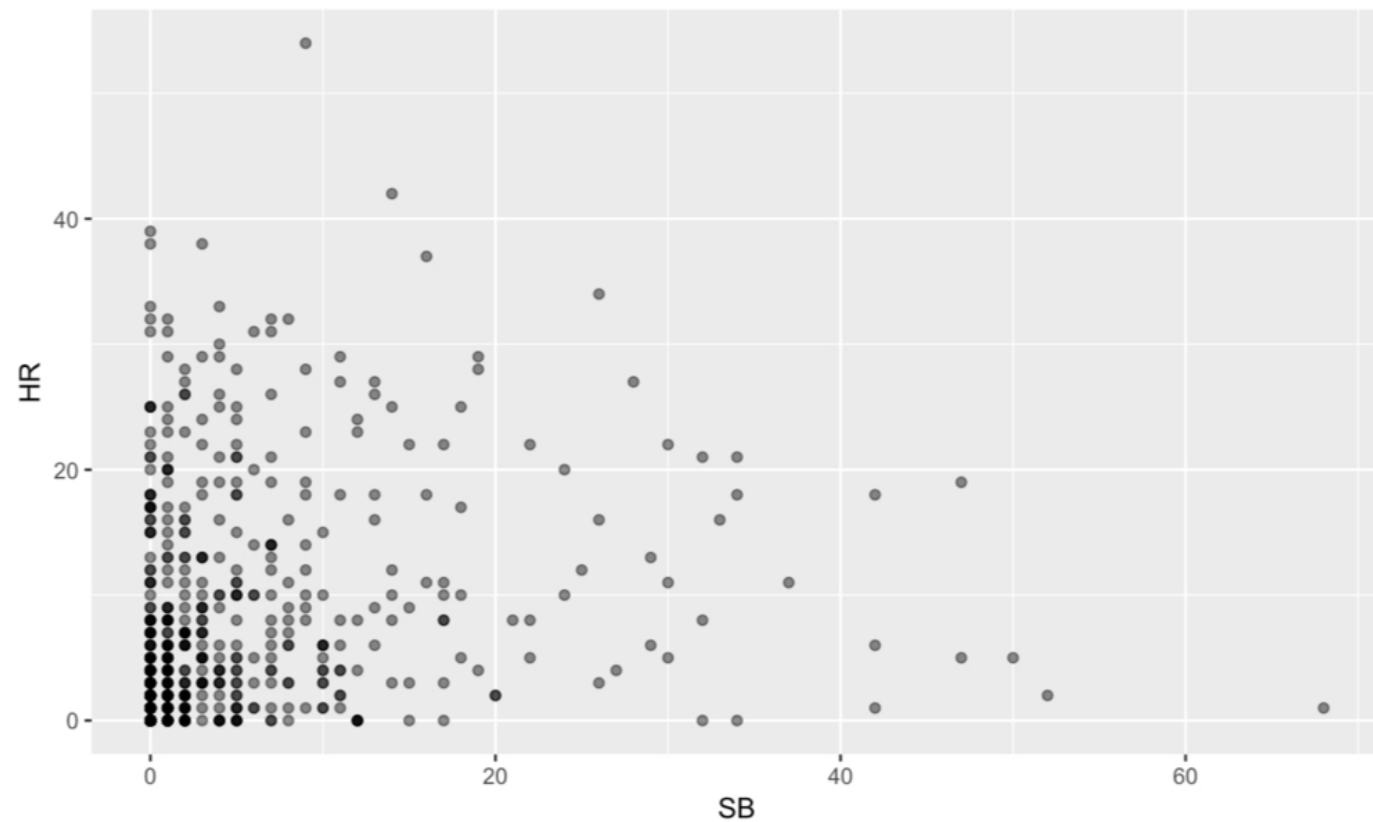
# Outliers

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



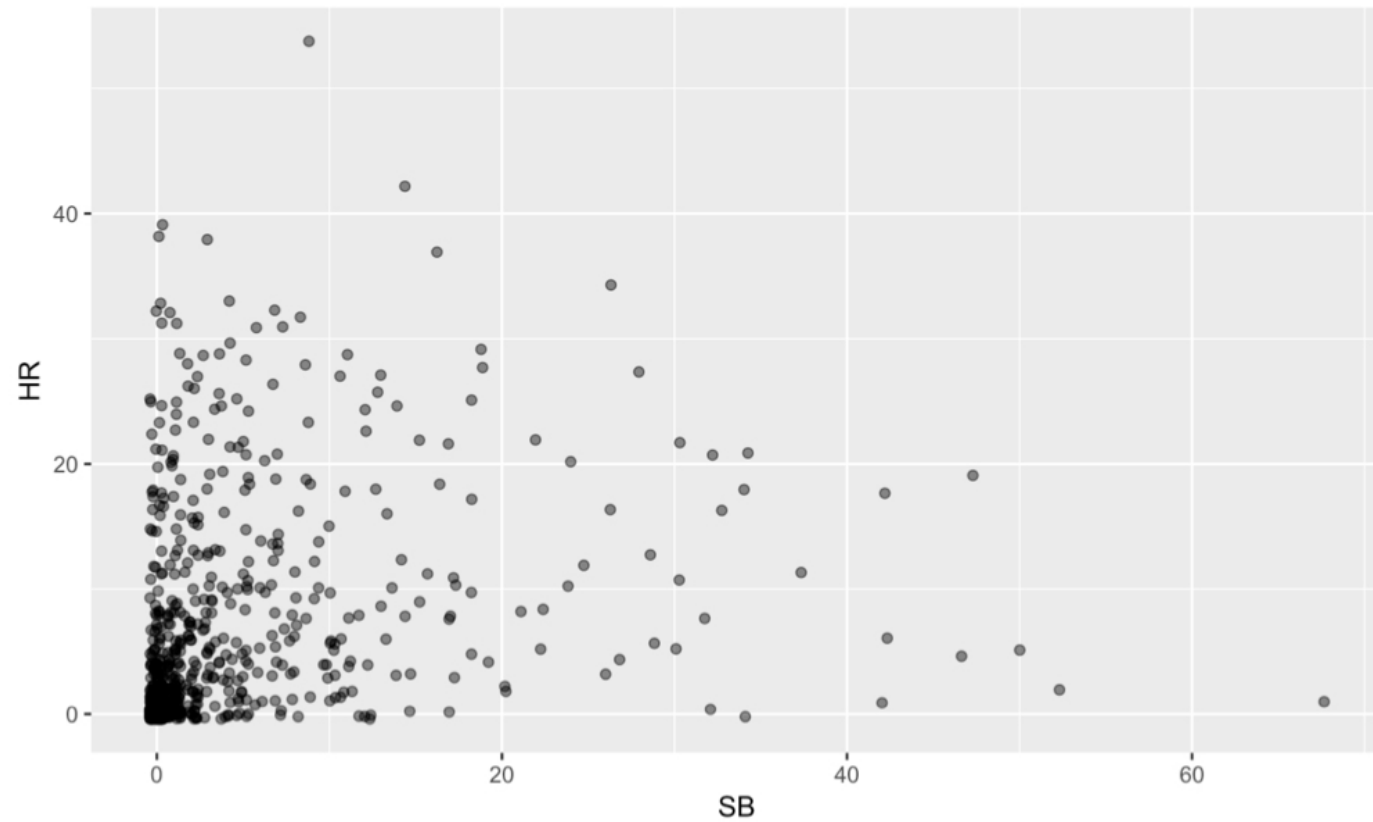
# Add transparency

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



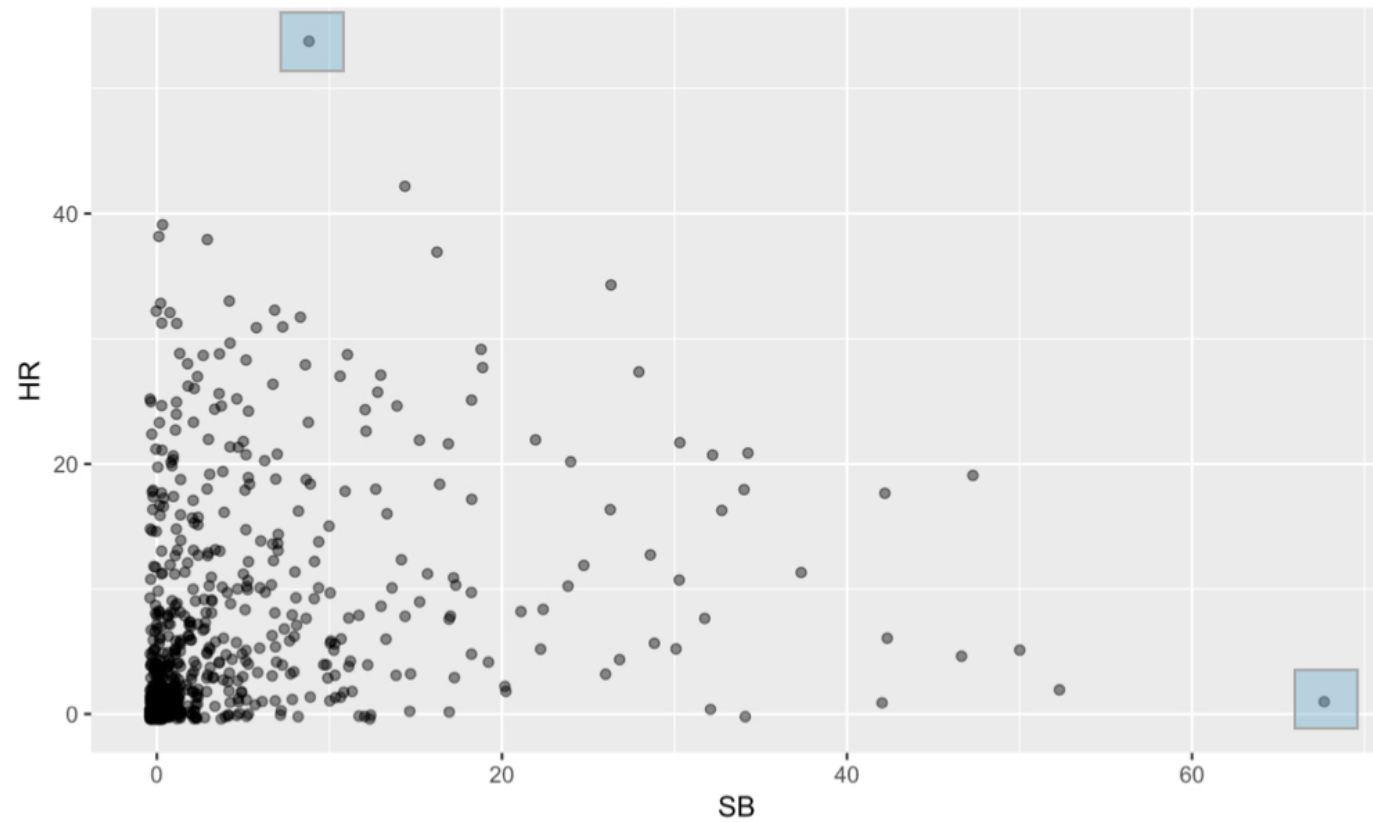
# Add some jitter

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



# 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



# Let's practice!

CORRELATION AND REGRESSION IN R