Welcome to DATA 151

I'm so glad you're here!

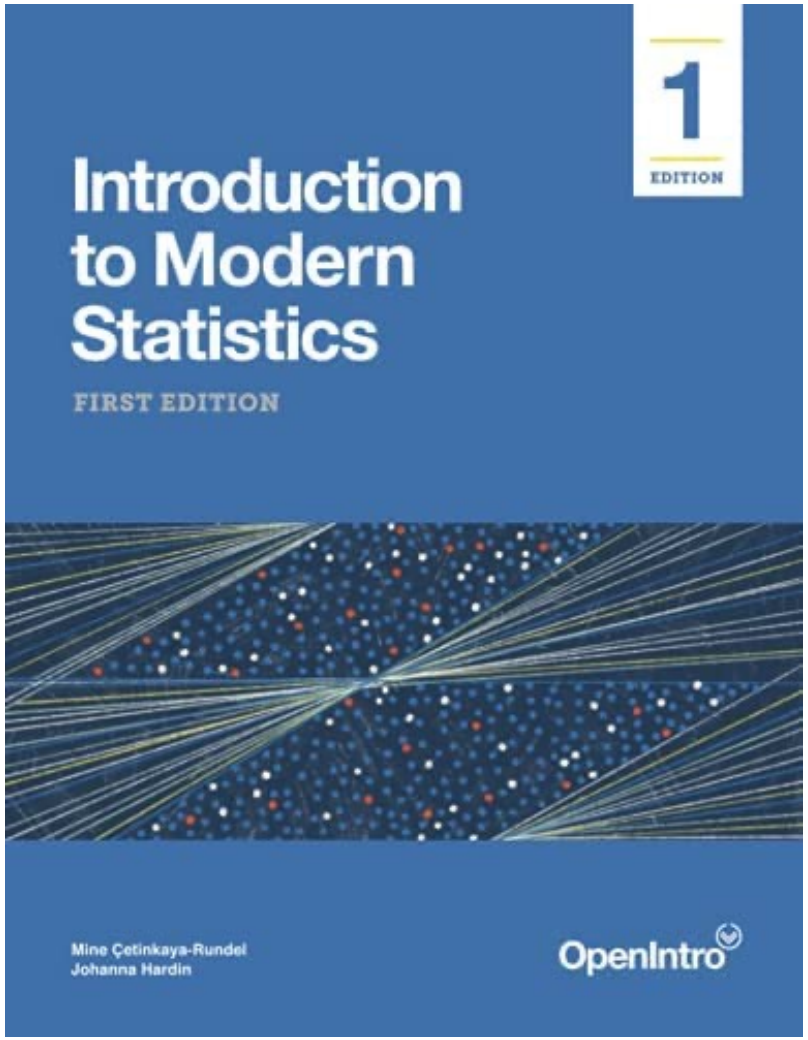
DATA 151: CLASS 11B
INTRODUCTION TO DATA SCIENCE (WITH R)

LINEAR REGRESSION

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

- Tuesday and Thursday:
  - Introduction to Modern Statistics
  - Ch 7: Relationships between two variables

HOMEWORK REMINDER

Due this week:

- ***NOTHING!***

HOMEWORK REMINDER

Due next week:

- ***DUE 11/17*** *Project Milestone #6*
  - Relationships between two numeric
- ***CANCELLED***
  - ~~***DUE 11/17*** *HW #10: DC Correlation and Regression*~~

EXAMPLE:

LINEAR REGRESSION MODEL CONDITIONS

***Example*: Climate Change and Fish Habitats**

As the climate grows warmer, we expect many animal species to move toward the poles in an attempt to maintain their preferred temperature range.

Do data on fish in the North Sea confirm this suspicion?

The data are 25 years of mean winter temperatures at the bottom of the North Sea (degrees Celsius) and the center of the distribution of anglerfish (sometimes called monkfish) in degrees of north latitude.

LINEAR REGRESSION MODEL CONDITIONS

https://www.oceaneyephoto.com/photo_5365532.html

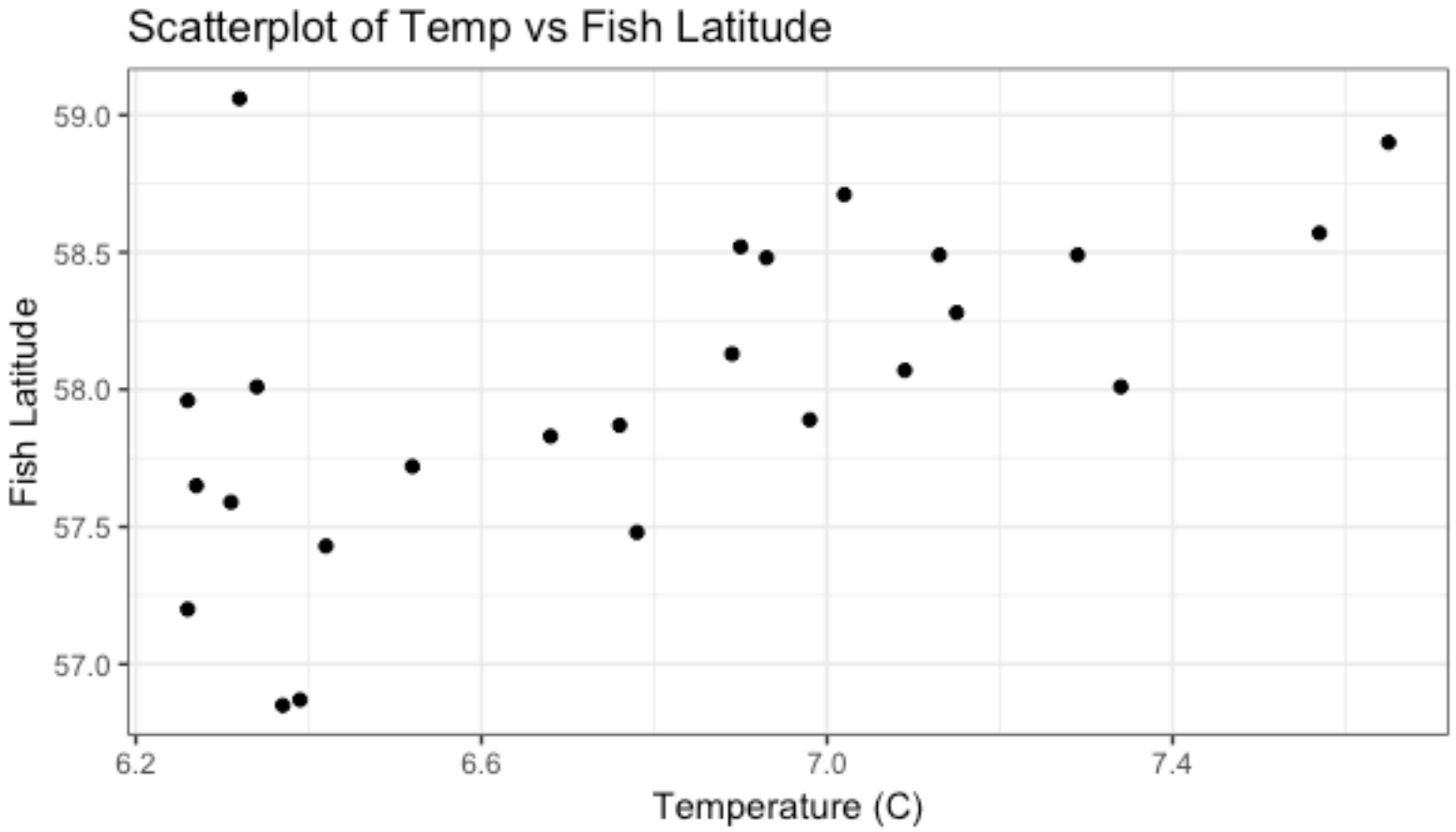
LINEAR REGRESSION MODEL CONDITIONS

Step 0: Load the data into R. Create a data frame.

```
# EXAMPLE: Climate Change and Fish Habitats
# Data on anglerfish distribution
# Explanatory: Temp in C (mean winter temperature at bottom of North Sea)
# Response: Latitude of center for distribution of anglerfish

year<-c(1977:2001)
temp<-c(6.26, 6.26, 6.27, 6.31, 6.34, 6.32, 6.37, 6.39, 6.42,
        6.52, 6.68, 6.76, 6.78, 6.89, 6.90, 6.93, 6.98,
        7.02, 7.09, 7.13, 7.15, 7.29, 7.34, 7.57, 7.65)
lat<-c(57.20, 57.96, 57.65, 57.59, 58.01, 59.06, 56.85, 56.87, 57.43,
       57.72, 57.83, 57.87, 57.48, 58.13, 58.52, 58.48, 57.89,
       58.71, 58.07, 58.49, 58.28, 58.49, 58.01, 58.57, 58.90)

fish<-data.frame(year, temp, lat)
```

LINEAR REGRESSION MODEL CONDITIONS

Step 1: Look at the data! Create a scatterplot!



Looks to have a moderately strong, positive, linear relationship.

They may be possible outlier in 1982.

LINEAR REGRESSION MODEL CONDITIONS

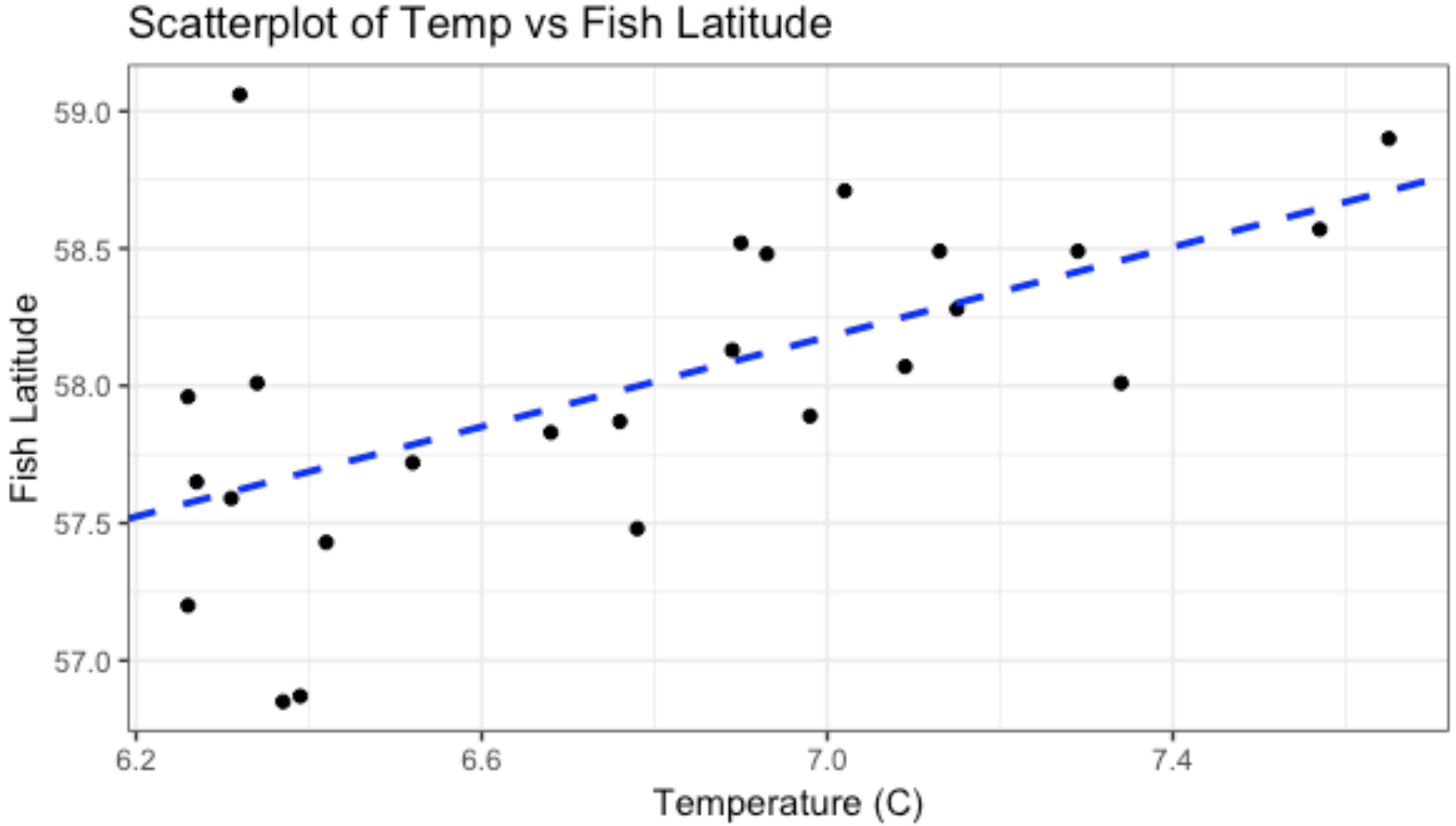
Step 2: Create a simple linear model

```
# Fit a simple linear model
mod2<-lm(lat~temp)
summary(mod2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81309 -0.27207 -0.02401  0.20523  1.43781

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.4524     1.5324  34.229  < 2e-16 ***
temp          0.8180     0.2254   3.629  0.00141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4734 on 23 degrees of freedom
Multiple R-squared:  0.3641,    Adjusted R-squared:  0.3364
F-statistic: 13.17 on 1 and 23 DF,  p-value: 0.001408
```

$$\hat{y} = 52.45 + 0.818x$$

LINEAR REGRESSION MODEL CONDITIONS

<u>Step 2</u>: Create a simple linear model… and add the fitted line to your scatter plot



LINEAR REGRESSION MODEL CONDITIONS

In our anglerfish example we found the fitted equation to be:

$$\hat{y} = 52.45 + 0.818x$$

Suppose that we want to **predict** the latitude of angler fish for this year when we estimate the temperature of the bottom of the North Sea to be 8.5°C?

OH NO!!! CAUTION!!! THIS IS AN EXAMPLE OF EXTRAPOLATION!!!

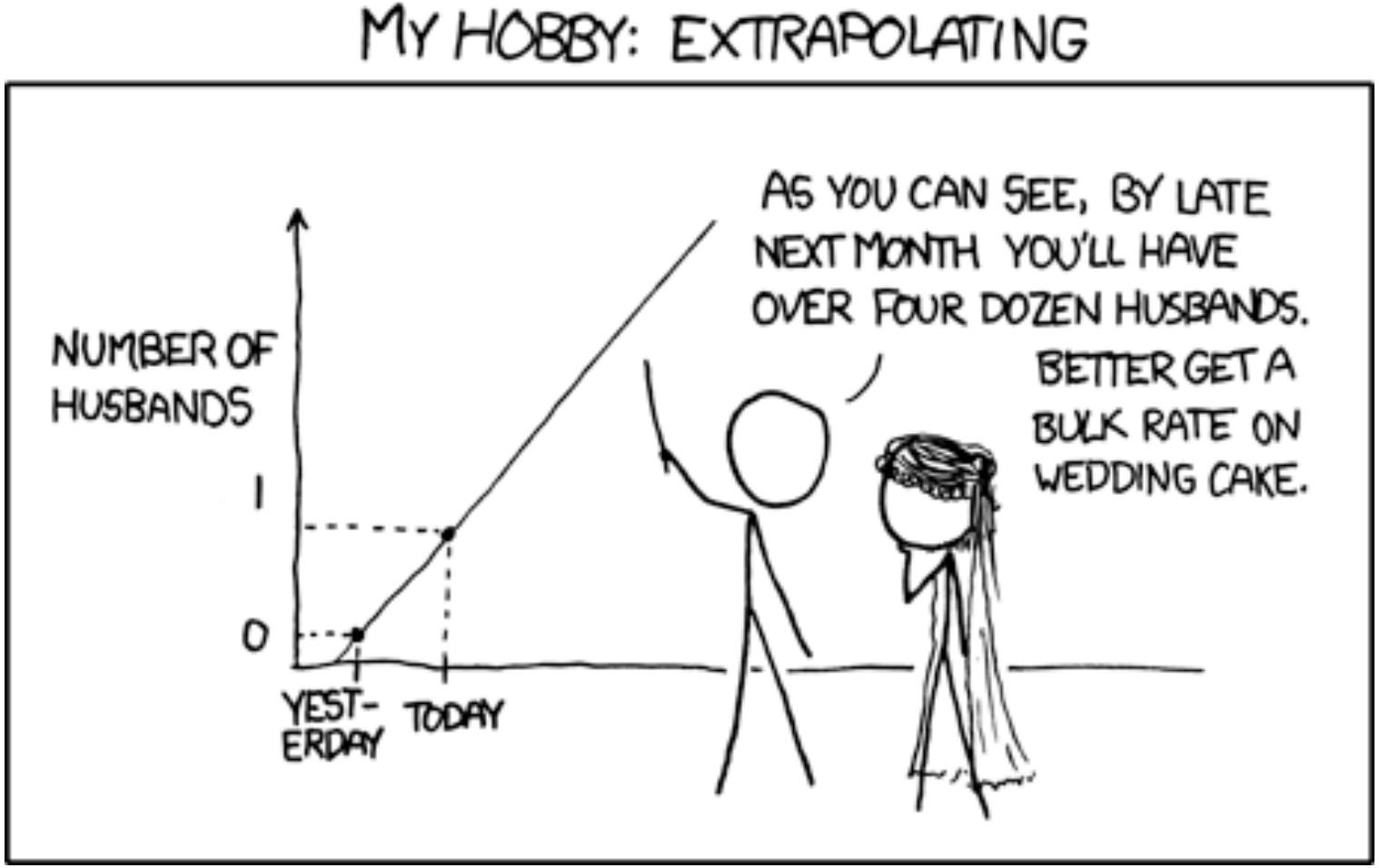
- Extrapolation is when we are trying to predict outside of the range of x in our data. The pattern may be different outside of this range.
- We only observed temperatures between 6.26 and 7.65 °C

ESTIMATING / PREDICTING

ESTIMATING / PREDICTING

- As stated previously, we can use the regression equation (or the line on the graph) to estimate or predict values of the response variable.
- Caution of EXTRAPOLATION!
  - Extrapolation is extending a trend outside of the range of observed values

ESTIMATING / PREDICTING

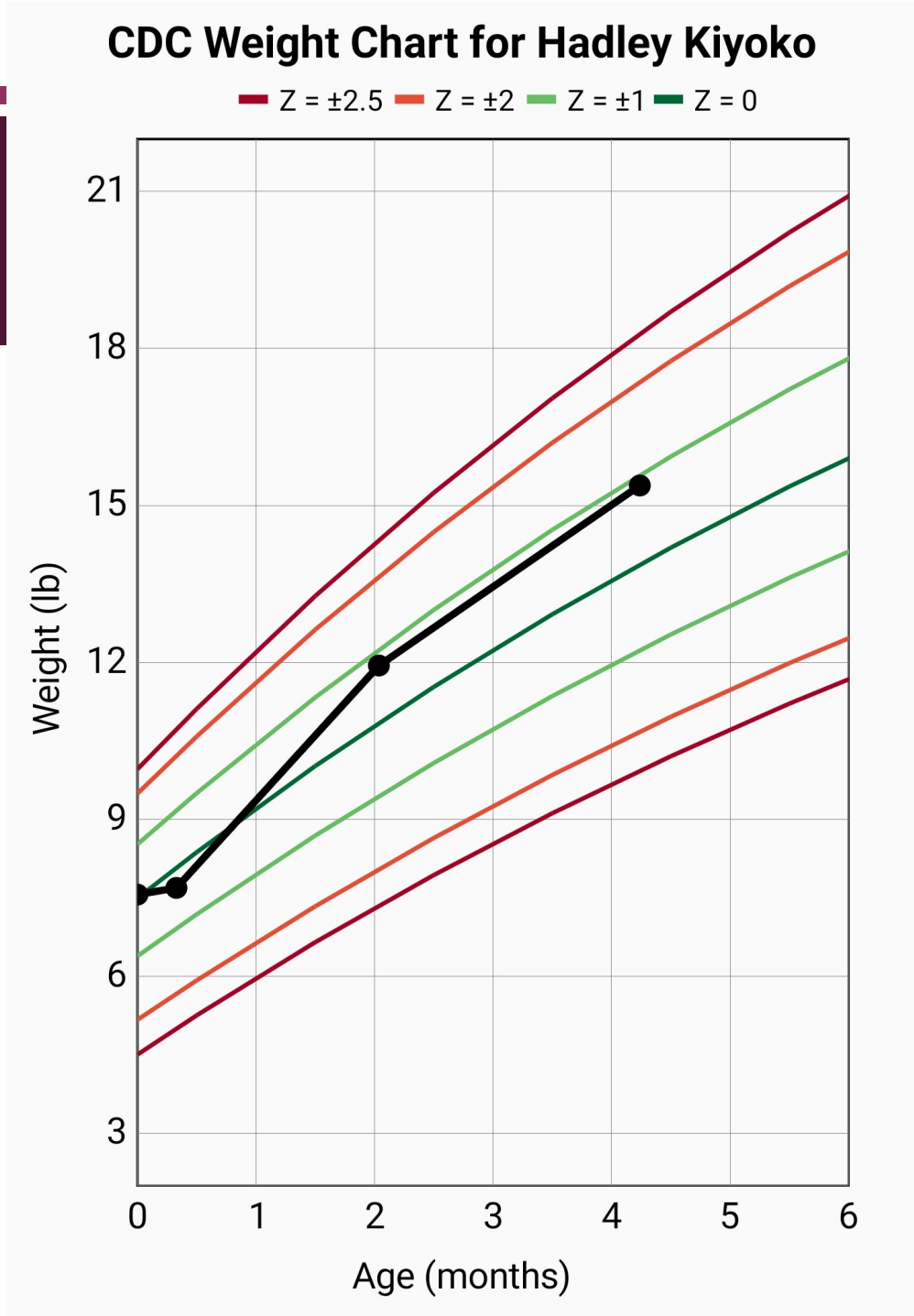


BEWARE OF EXTRAPOLATION!

EXTRAPOLATION

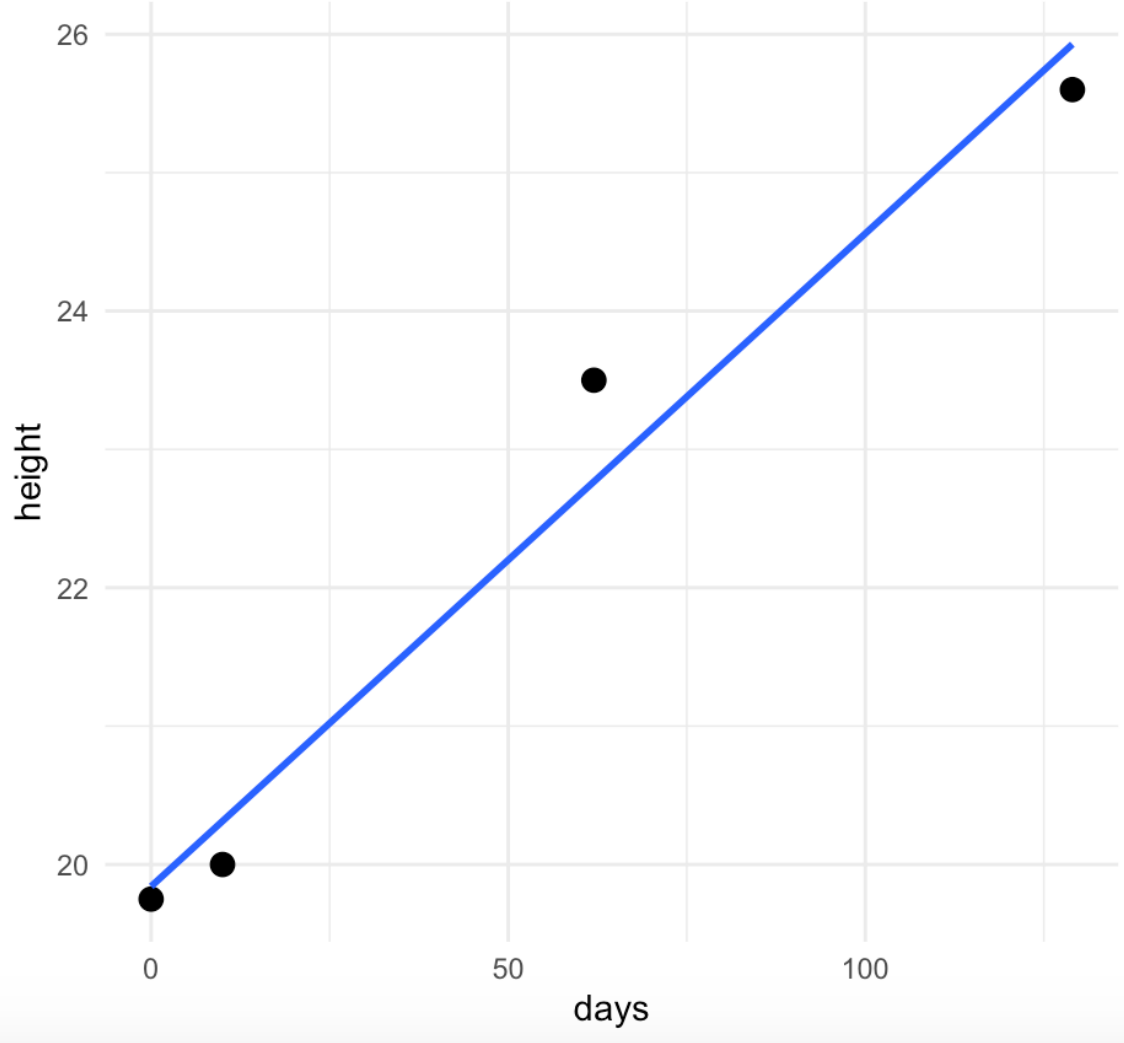
EXTRAPOLATION

- This is a growth chart for Hadley's baby wellness check-up
- Note that the z-scores are displayed



EXTRAPOLATION

```
growth<-data.frame(days=c(0, 10, 62, 129),
                   height=c(19.75, 20.0, 23.5, 25.6))

ggplot(growth, aes(days, height))+
  geom_point(size=3)+
  geom_smooth(method="lm", se=FALSE)+
  theme_minimal()
```



EXTRAPOLATION

```
> gMod<-lm(height~days, growth)
> summary(gMod)

Call:
lm(formula = height ~ days, data = growth)

Residuals:
       1        2        3        4
-0.09162 -0.31344  0.73312 -0.32805

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.841624   0.429515   46.20 0.000468 ***
days         0.047182   0.005987    7.88 0.015725 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6131 on 2 degrees of freedom
Multiple R-squared:  0.9688,    Adjusted R-squared:  0.9532
F-statistic:  62.1 on 1 and 2 DF,  p-value: 0.01572
```

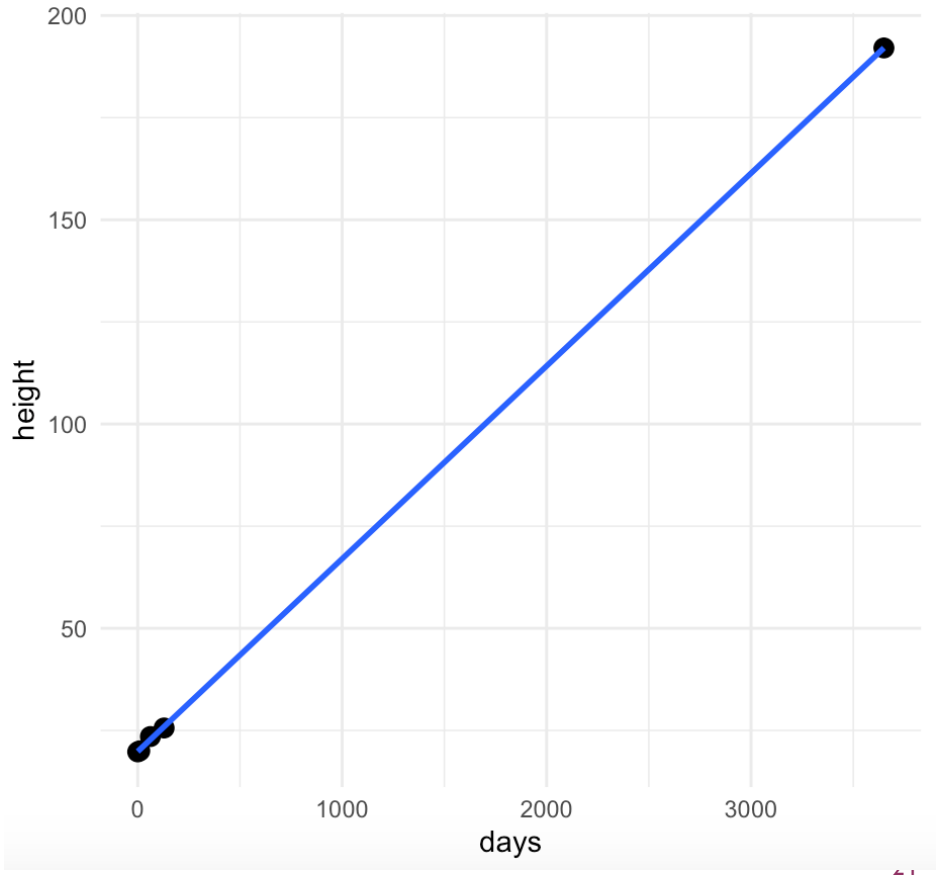
EXTRAPOLATION

Hadley's height (in inches) was plotted against her age (in days). The regression equation was found to be $\hat{y}$ = 19.84 + 0.047x

Can you predict her height at on her 10th birthday?

EXTRAPOLATION

Hadley's height (in inches) was plotted against her age (in days). The regression equation was found to be $\hat{y}$ = 19.84 + 0.047x



SOLUTION

Can you predict her height at on her 10th birthday?

19.84+0.047*3650 = 191.39 inches

191.39/12 = 15.95 feet

EXPLORING SUBGROUPS

INDICATORS AND INTERACTIONS

***Example*: Shipping Books**

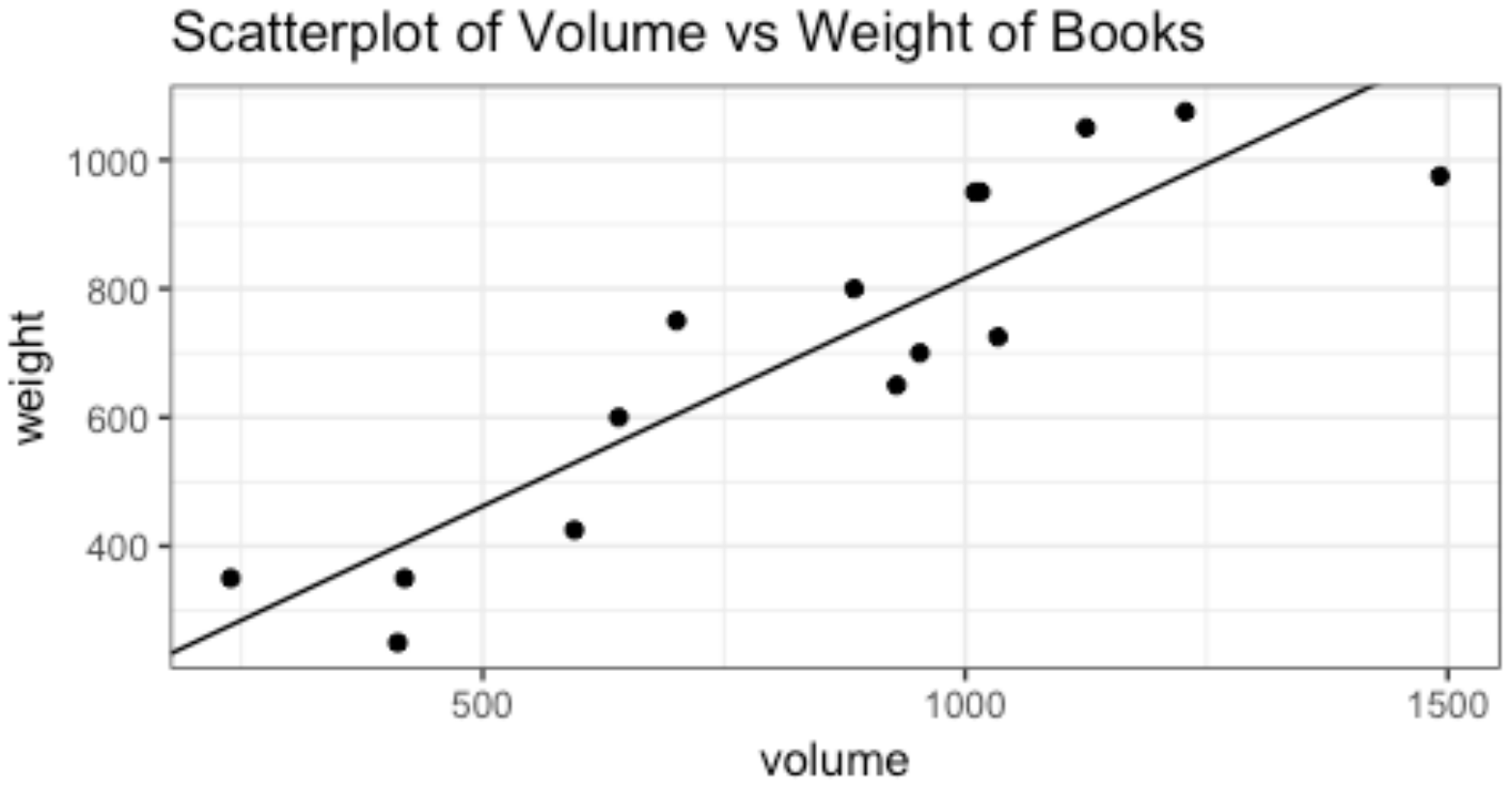
When you buy a book off Amazon, you get a quote for how much it costs to ship. This is based on the weight of the book. If you didn't know the weight of the book, what other characteristics of it could you measure to help predict the weight?

GROUP CODING

START WITH SLR

Example: **Shipping Books**

```
m3<-lm(weight~volume, data=books)
summary(m3)
```



START WITH SLR

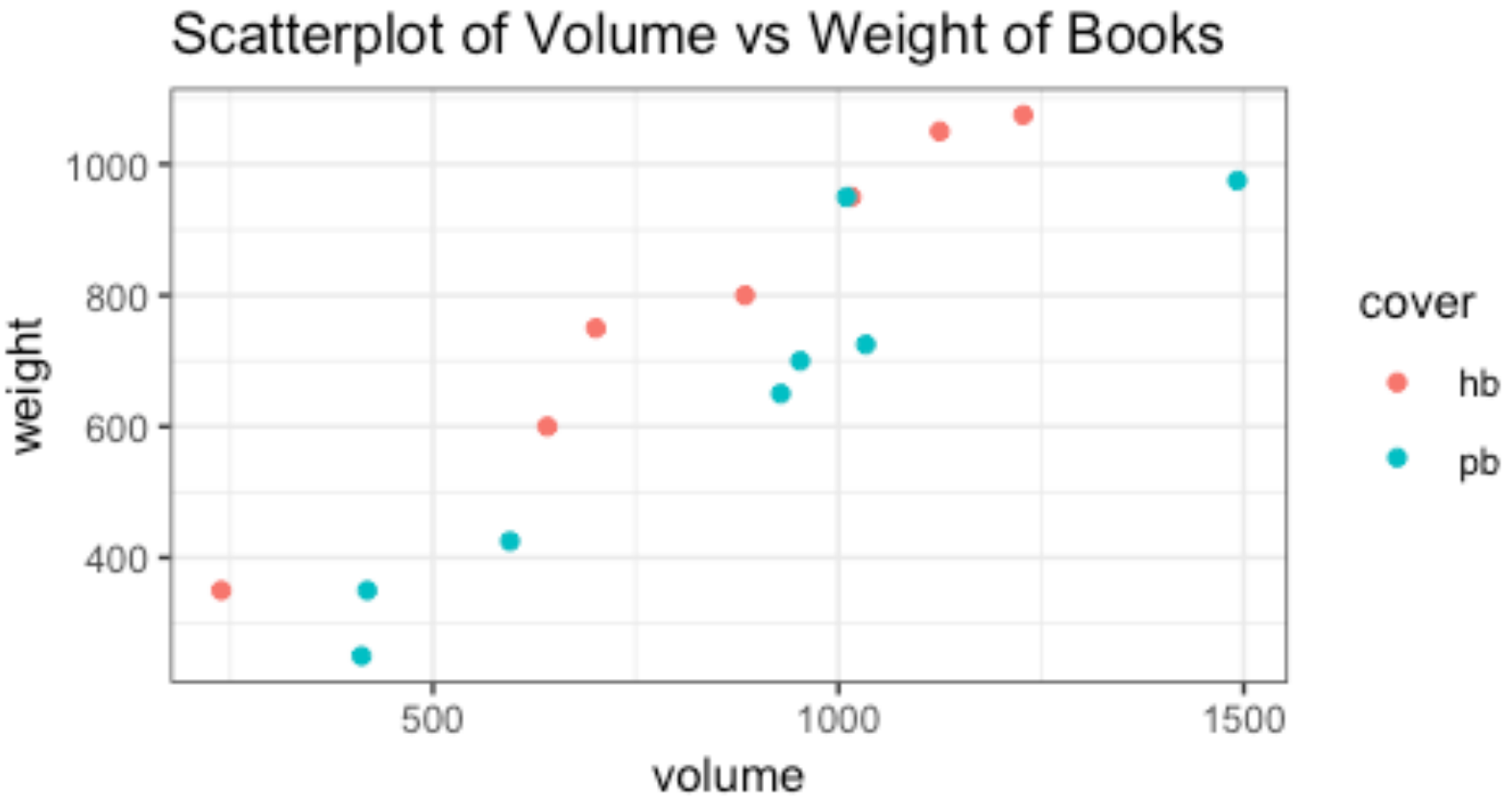
Example: Shipping Books (Model Output)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared:  0.8026,    Adjusted R-squared:  0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

$$\widehat{\text{weight}} = 107.68 + 0.71 \times volume$$

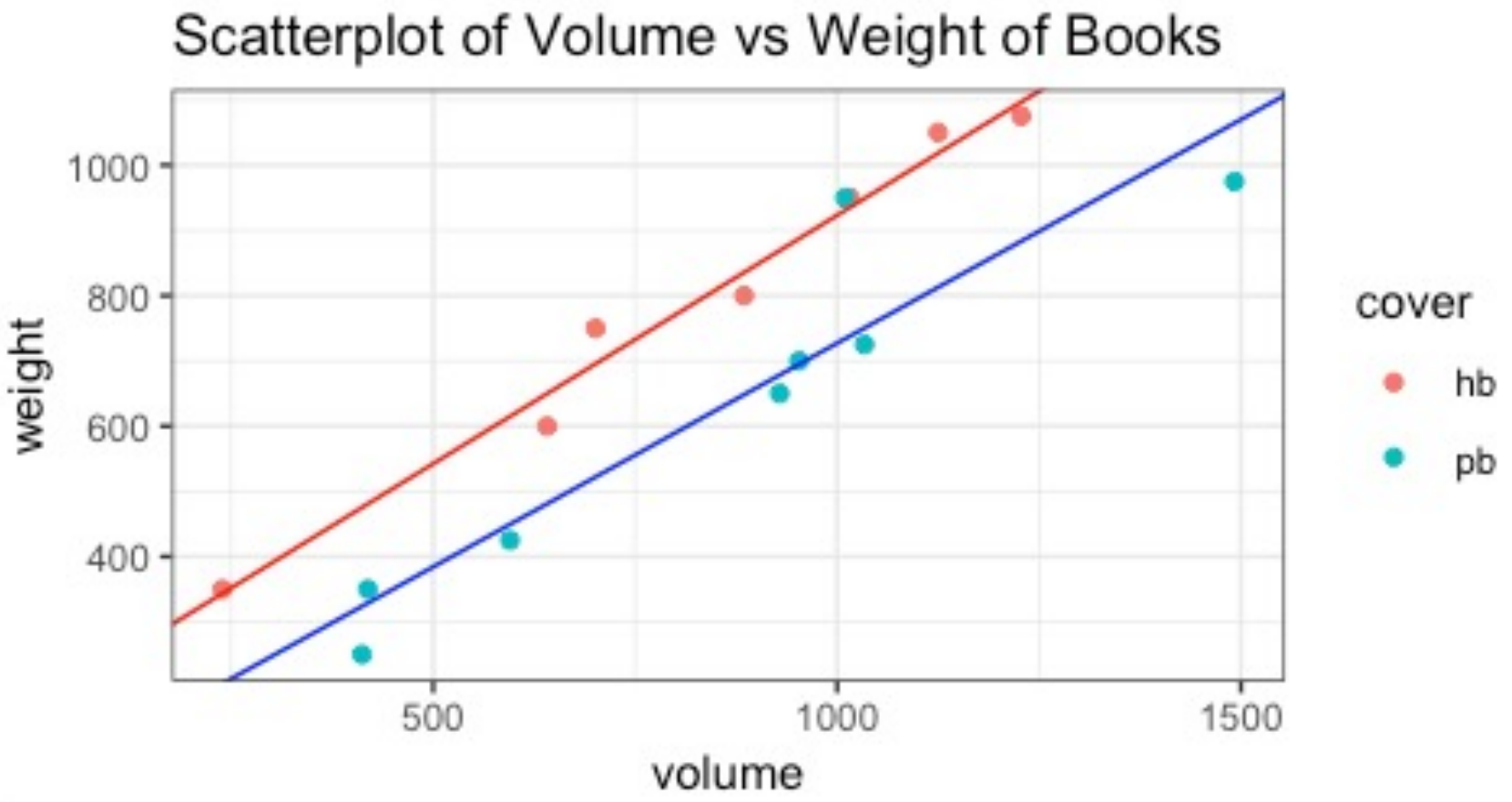
PARALLEL LINES

Would including cover type help out model explain more variation?



GROUP CODING

INTERACTIONS



The change in the slope was not significant, with a p-value of 0.566.

HOW DOES THIS WORK? WHAT ARE THESE LINES?

MORE DETAILS IN DATA 152 AND DATA 252

INTERACTIONS

- In R
  - "*" All possible subsets of interactions (and main effects)
  - ":" Only the specified interaction

- Test significance of interaction

- **Hierarchical principle**: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

INTERACTIONS

A shift in the intercept was significant, maybe we should also allow for different slopes.

```
# Include interaction to shift intercept and change slope
m5<-lm(weight~volume*cover, data=books)
summary(m5)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     161.58654   86.51918   1.868   0.0887 .
volume            0.76159    0.09718   7.837 7.94e-06 ***
coverpb        -120.21407  115.65899  -1.039   0.3209
volume:coverpb   -0.07573    0.12802  -0.592   0.5661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.41 on 11 degrees of freedom
Multiple R-squared:  0.9297,    Adjusted R-squared:  0.9105
F-statistic:  48.5 on 3 and 11 DF,  p-value: 1.245e-06
```

INTERACTIONS

`volume:cover` is an interaction term.

- It describes how the relationship between volume and weight may be different for the two cover type groups.

So we really have two different lines with different intercepts and slopes,

- Hardcover: $weight = 161.59 + 0.76 \times volume + (-120.21) \times 0 + (-0.08) \times volume \times 0$
  $\rightarrow weight = 161.59 + 0.76 \times volume$
- Paperback: $weight = 161.59 + 0.76 \times volume + (-120.21) \times 1 + (-0.08) \times volume \times 1$
  $\rightarrow weight = 41.38 + 0.68 \times volume$

INDICATORS AND INTERACTIONS

Take home messages:

- There is a statistically significant relationship between volume and weight.
- There is a statistically significant difference in weight between paperback and hardcover books, when controlling for volume.
- There is no strong evidence that the relationship between volume and weight differs between paperbacks and hardbacks.

FIVETHIRTYEIGHT ACTIVITY

READ THE ARTICLE

FiveThirtyEight

Politics **Sports** Science Podcasts Video

SEP. 29, 2017, AT 12:16 PM

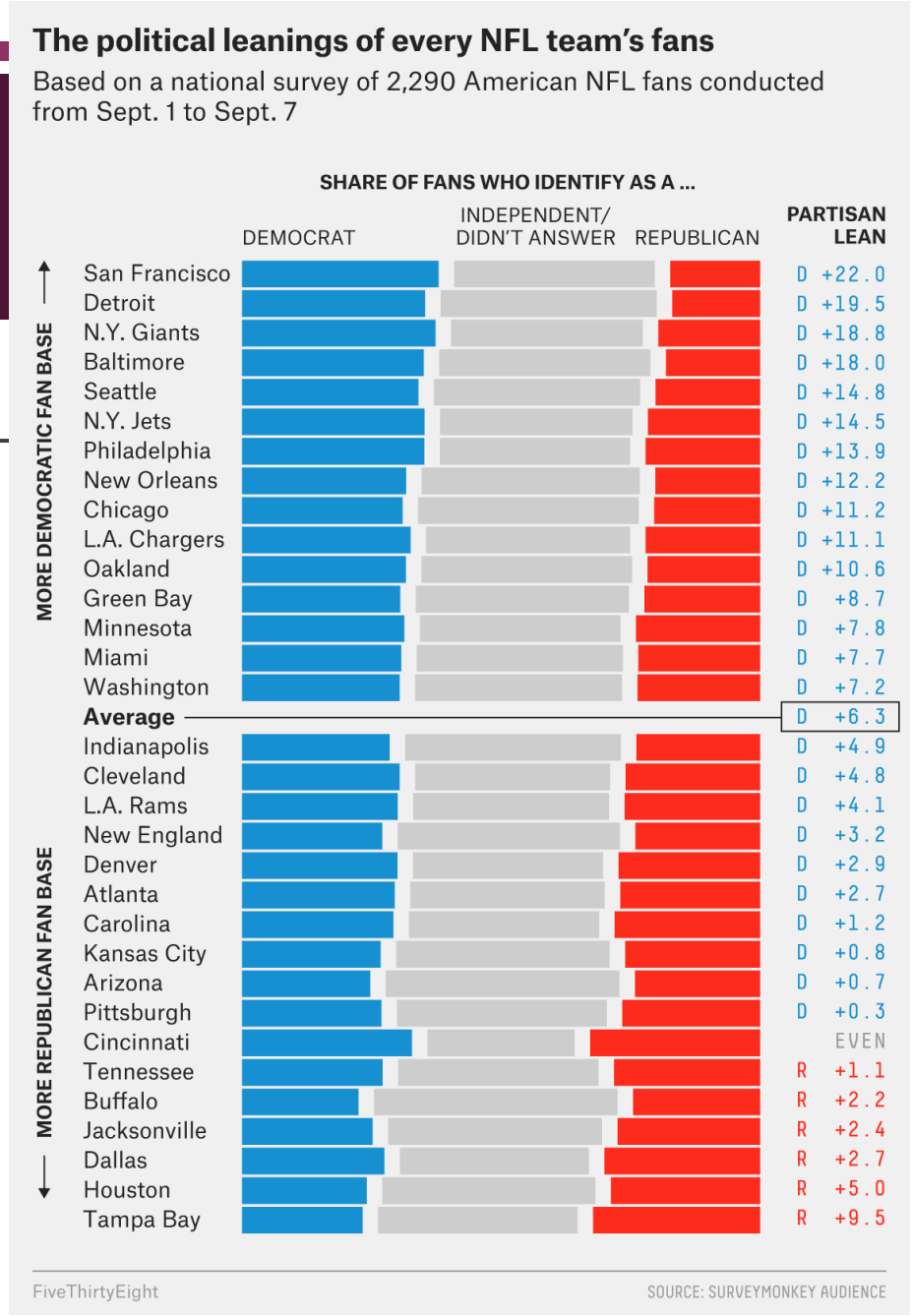
How Every NFL Team's Fans Lean Politically

By Neil Paine, Harry Enten and Andrea Jones-Rooy

Filed under NFL

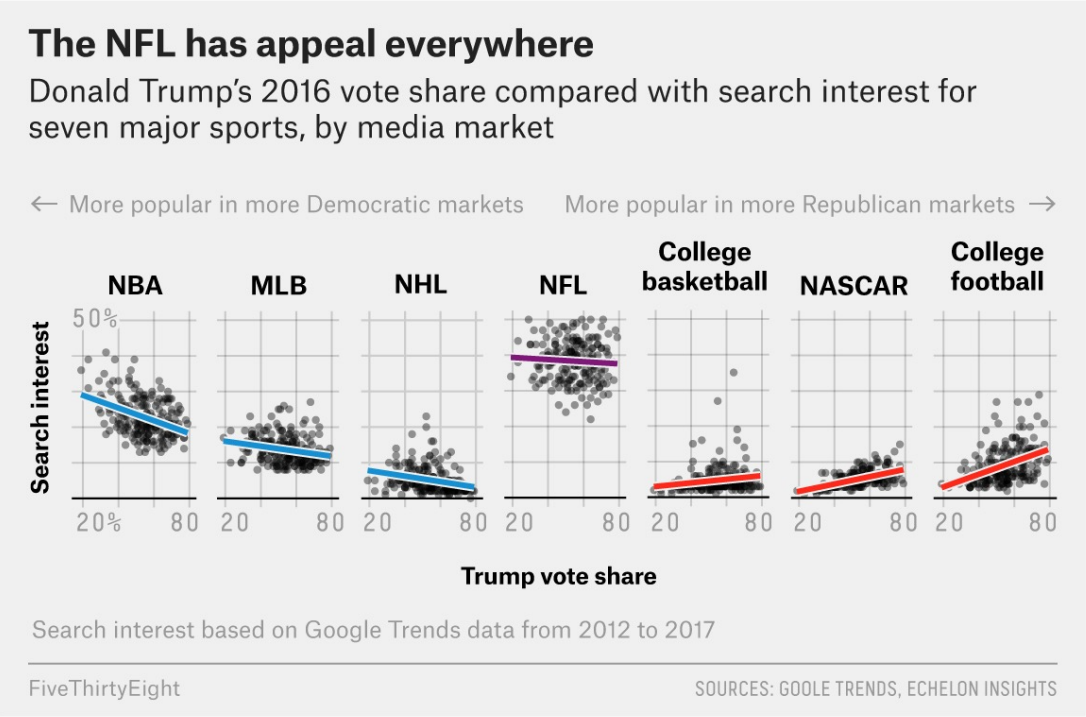
Get the data on GitHub

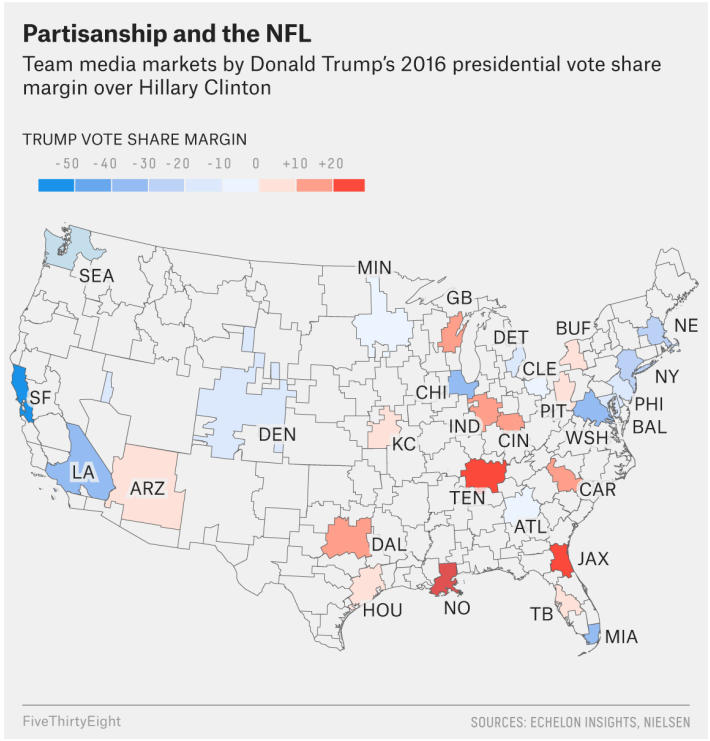
The showdown between President Trump and the NFL over some players' decision to kneel during the national anthem to protest racial injustice has raised all kinds of important issues. It's also put the most popular major sports league in the United States in a difficult position. The NFL's fan base is much more bipartisan than those of other major sports leagues, and it risks angering one side or the other if it mishandles the situation.

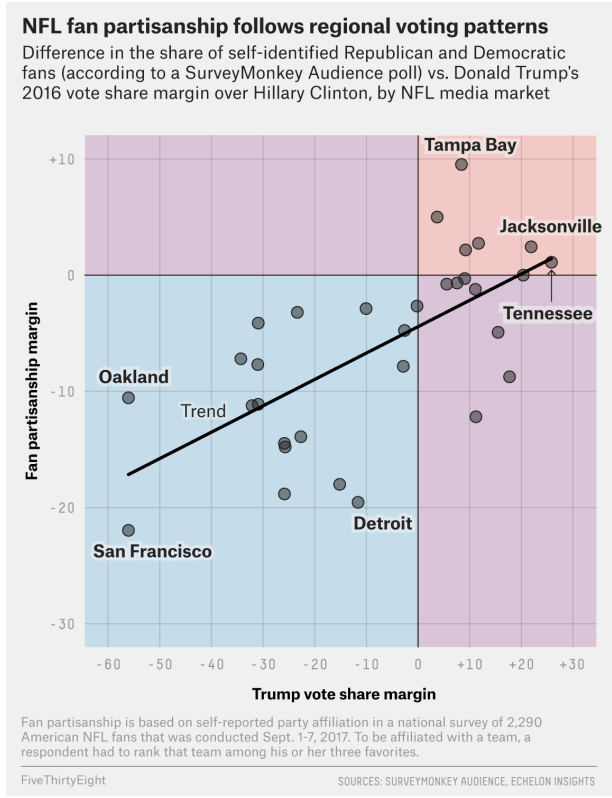


DISCUSS IN SMALL GROUPS

1. How are graphics used to tell the author's story?
2. What geometries are used?







WHAT DOES THE RAW DATA LOOK LIKE?

How to access the data:

```
# Load the tidyverse
library(tidyverse)

# Import data
sports<-read.csv("https://raw.githubusercontent.com/kitadasmalley/FA2020_DataViz/main/data/NFL_fandom_data.csv",
                 header=TRUE)
```

ARE WE GOING TO NEED TO TIDY THE DATA?

1. **Tidy the data:**

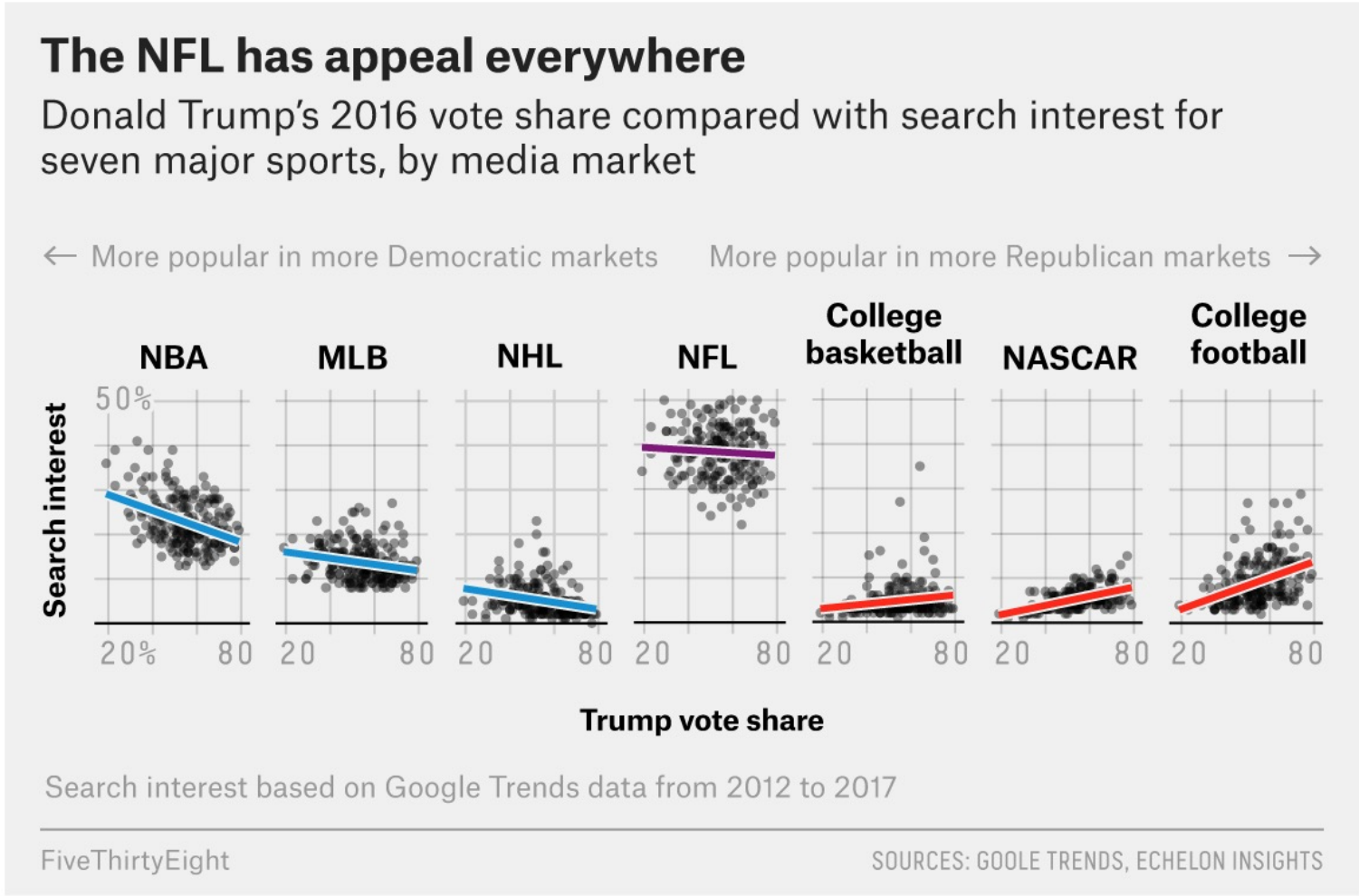
```
# Tidy the data
## Use gather to create:
### column for sport (categorical variable)
### Column for search interest (numeric - percent)

sportsT<-sports%>%
  gather("sport", "searchInterest",-c(DMA, PctTrumpVote))
```

WE MIGHT WANT TO RELEVEL THE SPORTS

2. **Relevel the data so that its in the right order:**

```
# Level the sport variable so that its in the right order
sportsT$sport<-factor(sportsT$sport,
                      level=c("NBA", "MLB", "NHL", "NFL", "CBB", "NASCAR", "CFB"))
```

RECREATE THIS GRAPH IN SMALL GROUPS



Task: Using the tools we have covered so far, recreate this graph.

Bonus Challenge: *Change the color of the lines.*