Prediction and correlation

Agenda

- Today:
  - Statistical communication exercise
  - prediction and correlation
- Other:
  - Project Part 1 released, due next Friday
  - Don't wait too late to try loading data into R!
    - See me if you have any issues

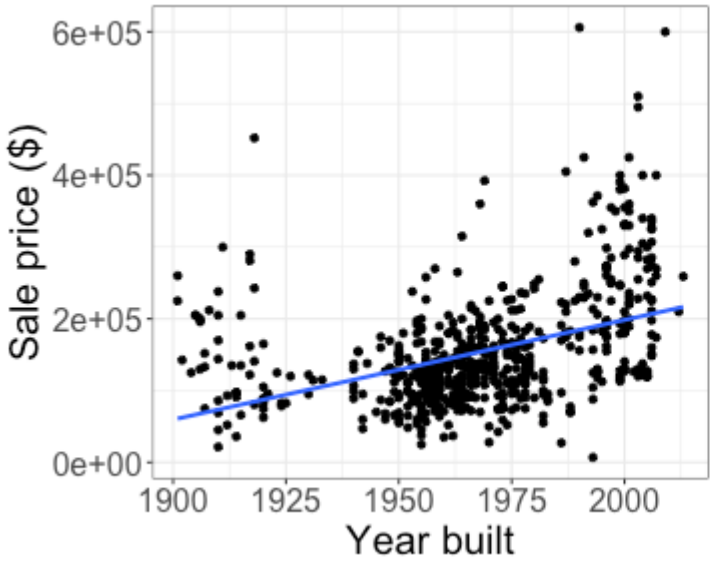
Statistical communication exercise

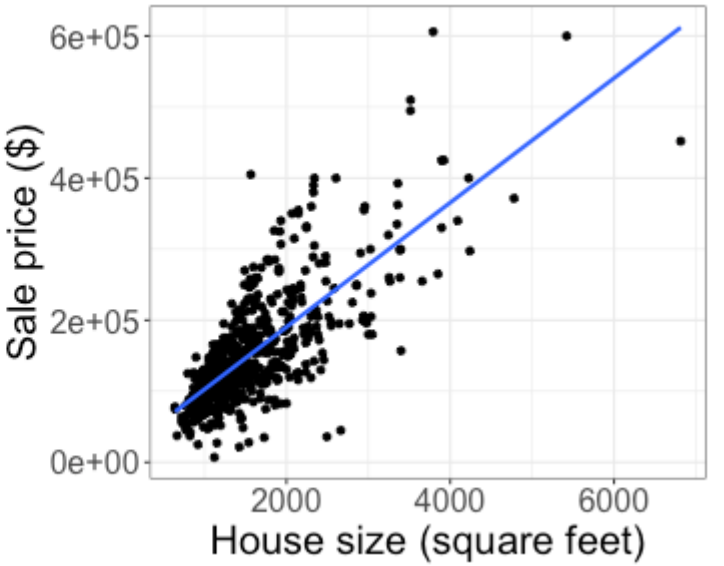
You are a statistical consultant, and a client reaches out to you to ask about some data they have collected. You fit a linear model, and the client has a question for you.

Spend ~5 minutes, individually, writing a brief email to the client to answer their question. Then we will discuss as a class.

Last time: SSE

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$





SSE: 3.41×10^{12}

SSE: 1.74×10^{12}

These numbers are very large!

Last time: SSE

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Depends on the scale of the response variable
- Depends on the *number* of observations
- Depends on the strength of the relationship

Alternative: RMSE

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

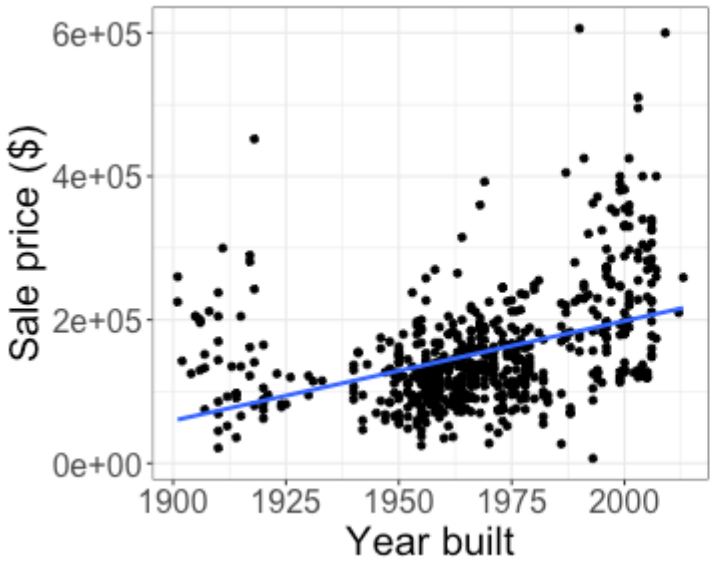
- Depends on the scale of the response variable
- Depends on the *number* of observations
- Depends on the strength of the relationship

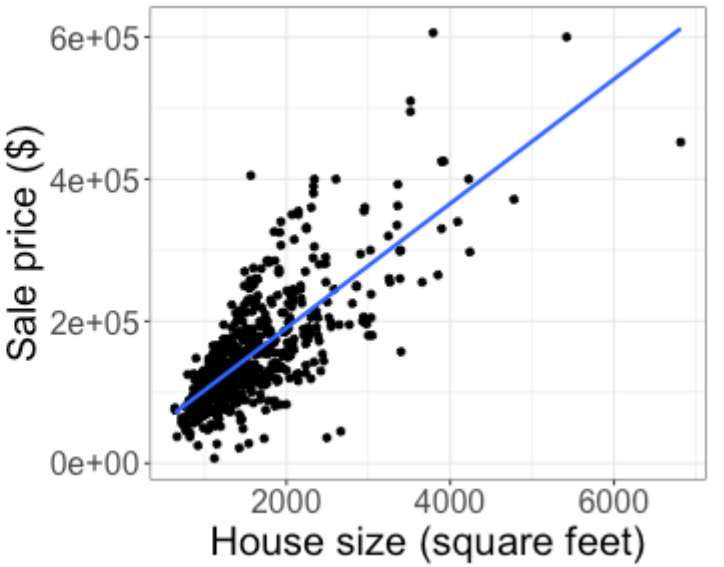
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Depends on the scale of the response variable
- Depends on the strength of the relationship
- Does *not* depend on the number of observations

Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

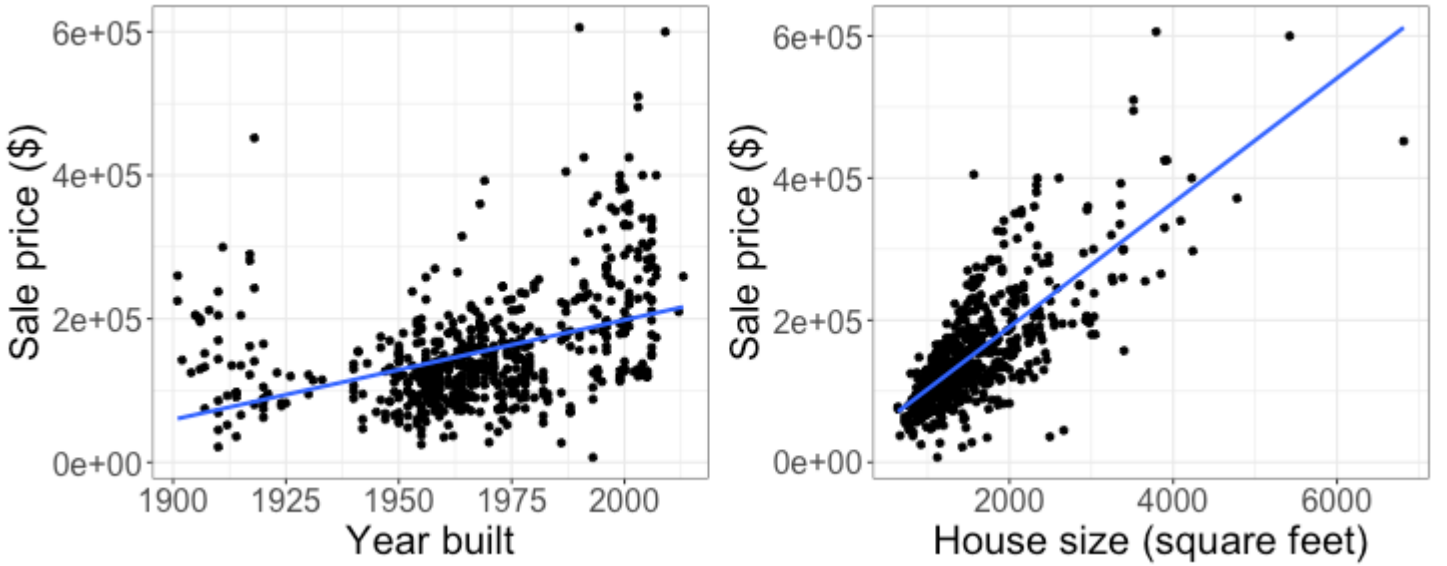




RMSE: 73121

RMSE: 52745

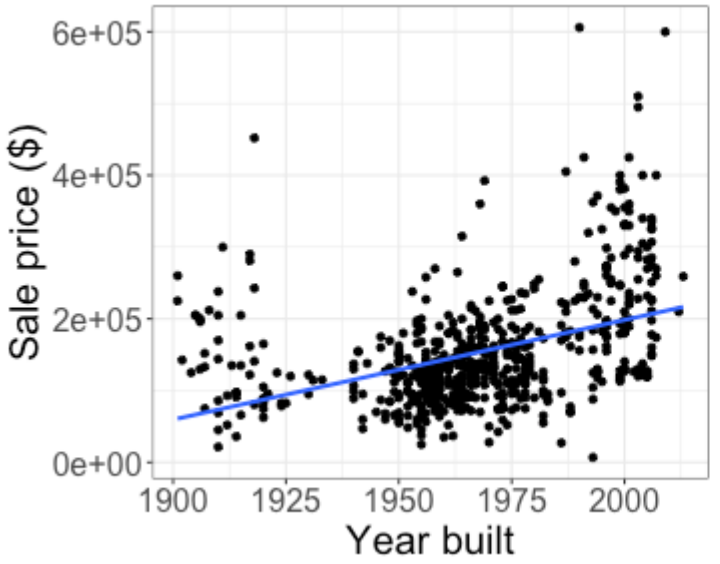
Root mean square error (RMSE)

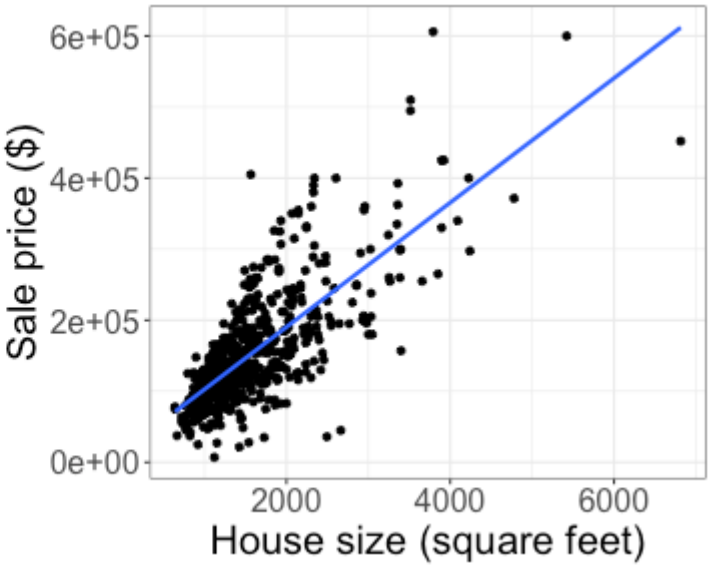


- Both SSE and RMSE are useful for summarizing our ability to make predictions
- Both depend on the scale of the response variable

Is there a measure of the strength of a linear relationship that does *not* depend on scale?

Correlation





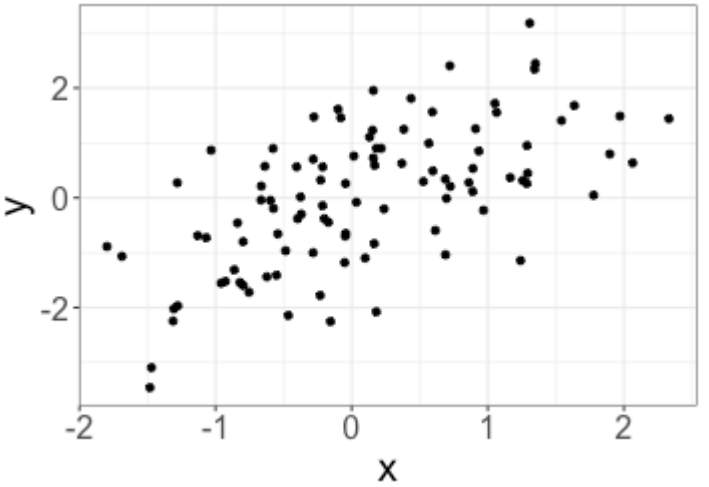
Correlation: 0.41

Correlation: 0.75

> What is the possible range of values for a correlation?

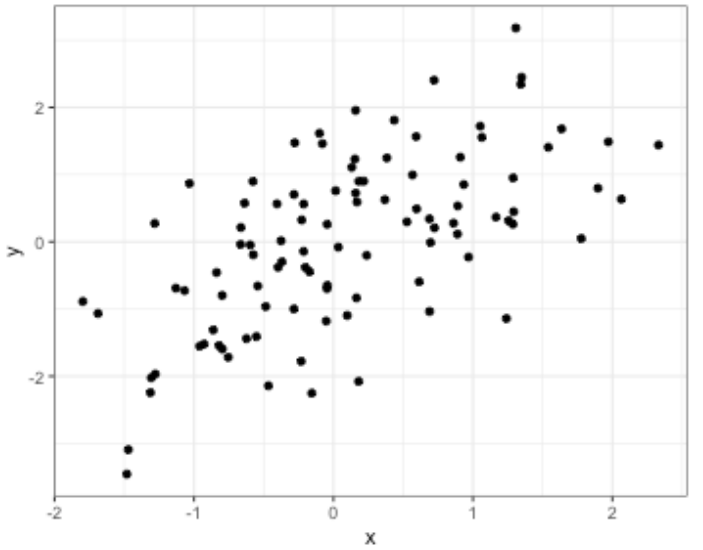
Correlation

Observe points $(x_1, y_1), \ldots, (x_n, y_n)$



$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Correlation



(A) 0.2

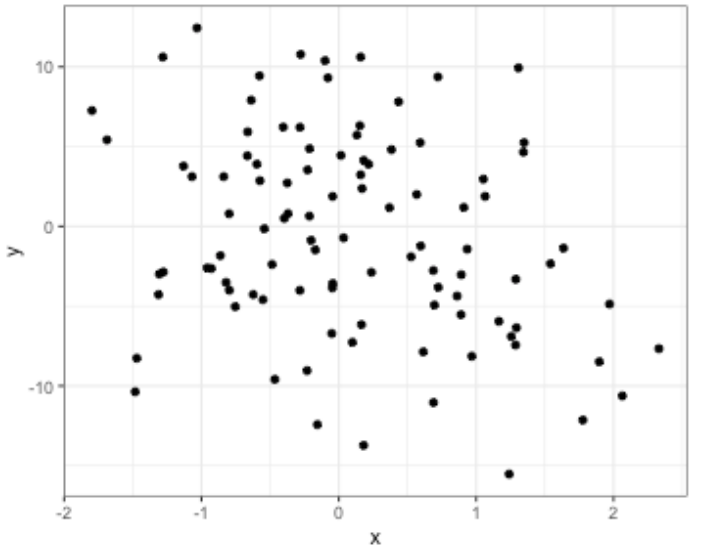
(B) -0.2

(C) 0.6

(D) 0.9

> What would you estimate as the correlation for this data?

Correlation



(A) 0

(B) -0.28

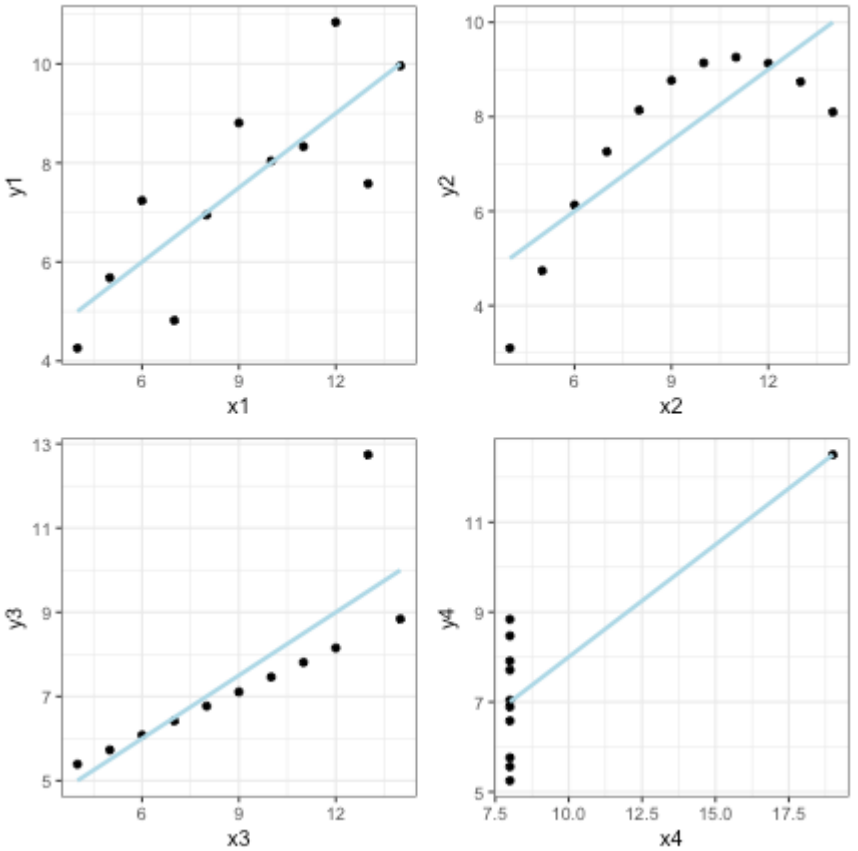
(C) 0.28

(D) -0.6

> What would you estimate as the correlation for this data?

Correlation

> Which of these plots shows the highest correlation?

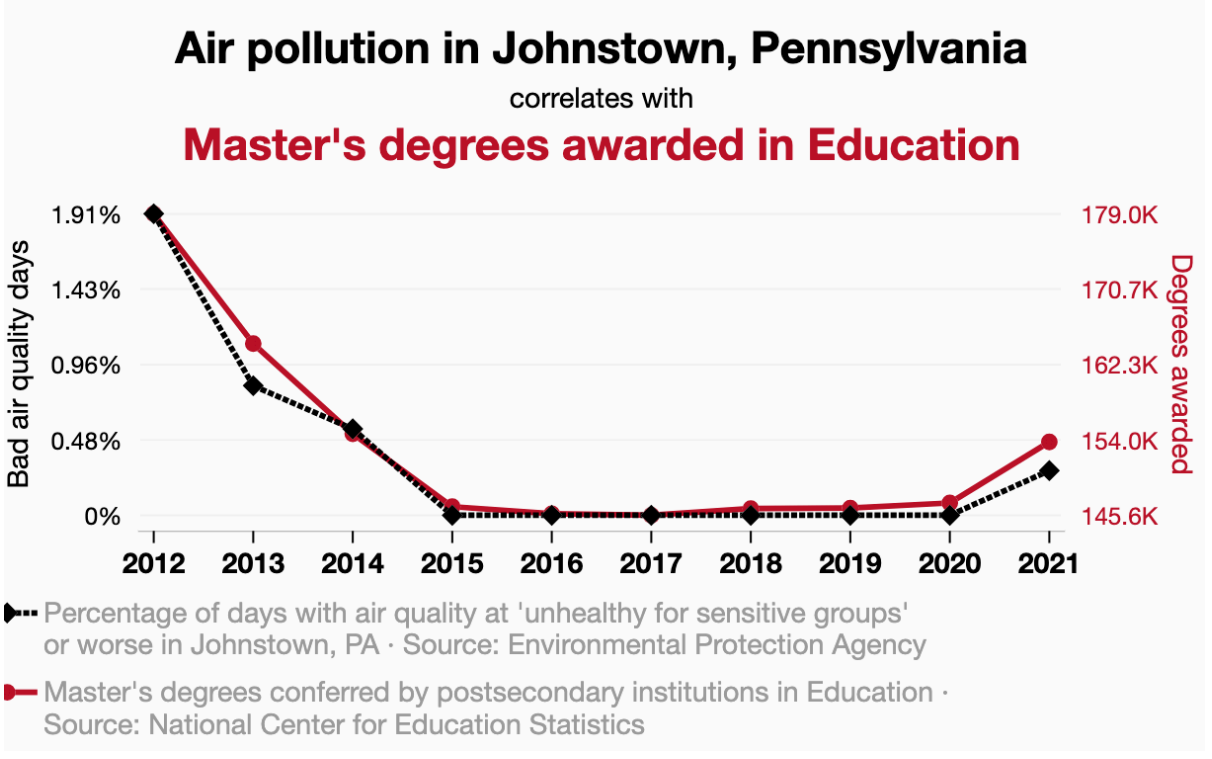


Prediction and correlation

- SSE:
  - depends on strength of the relationship
  - depends on number of observations
  - depends on scale of the response
- RMSE:
  - depends on strength of the relationship
  - depends on scale of the response
- Correlation:
  - depends on strength of the relationship
  - values always between [-1, 1]
  - measures strength of a **linear** relationship -- may not be suitable for nonlinear relationships
  - influenced by outliers and extreme points

Correlation?

> What is going on in this plot? Do you think this correlation is meaningful?



r = 0.989

Activity: spurious correlations

https://sta112-s26.github.io/class_activities/ca_07.html

- Play around with spurious correlations!
- No need to submit anything for this activity