Big Idea:

- When working with lots of variables, it is still very important to look at univariate and bivariate relationships and consider "preprocessing" your variables before you start building models.

Example 1: Predicting house prices

Kaggle has a large dataset on housing prices for over 20,000 homes sold in King County, Washington (Kaggle, 2018).

(a) Identify some potential sources of variation in housing prices (and conjecture whether you think the association will be positive or negative).

(b) Using the Two Quantitative Variables applet (test version), use the pull-down menus to select different explanatory variables (including the binary waterfront variable). Also check the Show correlation coefficient box. Are the associations in the direction(s) you expected? Which of these explanatory variables appears to explain the most variation in housing prices?

(c) Consider the *price* vs. *sq ft* relationship. Is it linear? Any problems with using this relationship to predict housing prices?

(d) Check the **Show transformations** box. Try different transformations, do any provide a more appropriate linear association? How are you deciding?

(e) Are there any homes you would consider removing from this dataset? Why or why not?

(f) What else would you consider doing before using these data with students?

Big Idea:

- When we cannot control a source of variation in an experiment (e.g., an observational study), we can still analytically "adjust" the association between the RV and EV of interest by other potential confounding variables.

Example 2: As we age, our medial temporal lobe (MTL) in our brain, particularly the total hippocampal volume, gets smaller, leading to impaired memory, Alzheimer's disease etc. Aerobic fitness is positively correlated with total hippocampal volume. Using MRI scan's, Siddarth et al. (2018) measured total MTL thickness for 35 non-demented middle-aged and older adults (25 women, 10 men; 45-75 years), measuring physical activity levels using the International Physical Activity Questionnaire (IPAQ). Participants meeting criteria for depressive or anxiety disorders (n = 9) were excluded.

(a) Create a sources of variation diagram, identify inclusion criteria as well as known sources of variation and the anticipated directions of those associations. What other sources of unexplained variation can you think of?

Observed Variation in:	Sources of explained variation	Sources of unexplained variation
Inclusion criteria: *Constant by Design:*		

From the article: "The IPAQ-E consists of 4 sets of questions assessing walking, moderate physical activities, vigorous physical activities, and average time spent sitting per day. Total physical activity is quantified by weighting each type of activity by its energy requirements defined in metabolic equivalent units (METs) to yield a score in MET-minutes per week. … Sitting data from the IPAQ-E was reported as average number of hours spent sitting per weekday over the past week. Weekend days were not included as prior work has shown lower accuracy of self-reported sitting for weekend days."

The data are in S1Table.xlsx from the article. Total MTL thickness is in column W. We have copied this column and a few other key variables into "Sheet 1." Load these data into the Multiple Variables applet (test version).

(b) Do lower levels of physical activity appear to be associated with less MTL thickness? How are you deciding?

(c) Does more time spent sitting appear to be associated with less MTL thickness? How are you deciding?

(d) Is it plausible that sedentary behaviors are independent from exercise and other physical activities? What are the implications of this statement?

(e) In the applet, using TOTAL as the response variable and sitting as the explanatory variable, move Age to the Explanatory box *above* sitting. The scatterplot color codes the dots by the values of this quantitative variable.

- What do you learn about the association between TOTAL (MTL thickness) and age from this graph?

- What do you learn about the association between Sitting and Age from this graph?

(Check the 2-variable graph to confirm your observations.)

So how do we adjust the relationship between MTL and sitting for age? The positive association between MTL and age says part of the reason for higher MTL values is lower ages and part of the reason for lower MTL values is larger ages. So to "subtract the age effect," we want to lower the MTL of the younger participants and raise the MTL of the older participants.

(f) So if I want to "adjust" the MTL response values by age, the white and blue dots will move _____ _____ and the darker dots will move _____

In an observational study, we also have to consider the association between the explanatory variables. In this case, there is a (weak) negative association between age and sitting. Age also explains some of the variation in sitting and we want to "separate" the variation in sitting that is not related to age. In other words, we can't adjust for age without also considering the relationship with sitting.

(g) So to adjust the sitting values for age as well, the white and blue dots will move _____ and the darker dots will move _____

Check your answers by checking the adjust y-values box and then the adjust x-values box. The resulting graph is often called an "added-variable graph." It demonstrates the leftover relationship between MTL thickness and sitting, after adjusting for age. This is equivalent to graphing the residuals of MTL on age against the residuals of sitting on age.

(h) Check the Show equation box. How has the coefficient of sitting changed?

(i) Check the Correlation box. How does the "partial correlation coefficient" change?

(j) Admittedly, the changes in these values are not large. Why not?

(k) Replace Age with Education and repeat the analysis. Is Education a confounding variable in this study? How are you deciding? What do you notice about the SS values in the pie chart if you reverse the order of the variables?

(l) Did you use *education* as a quantitative or categorical variable in your model? What are pros and cons of this modelling choice?

(m) Replace Education with Sex and repeat the analysis. Is Sex a confounding variable in this study? How are you deciding? Include an interpretation of the coefficient of *sitting* in this model.

(n) The above analysis assumed parallel lines between males and females, just possibly with different intercepts. Move Sex to the Subset By box to fit an interaction in this model. Based on the graph, describe the nature of the interaction.

(o) Fit a model that predicts MTL thickness from *sitting*, after adjusting for age, sex, education, is *sitting* still statistically significant?

(p) From Roback and Legler, a *New York Times* article was published discussing Siddarth et al. (2018) with the title "Standing Up at Your Desk Could Make You Smarter" (Friedman 2018). Do you agree with this headline choice? Why or why not?

In the article, the researchers also considered classifying the MET values as "lower" (below 1500 MET-minutes per week) and "higher" (above 1500).

(q) Analyze the relationship between this variable and the MTL thickness.

(r) What if use move IPAQgrp as the response variable? Is sitting a significant predictor?

Example 1 revisited

Let's consider a processed version of the Kaggle dataset: (n = 2000)
www.rossmanchance.com/data/KingCounty2.txt

(g) To create a regression model with more than one explanatory variable, move the data (or log transformed or scaled) to the Multiple Variables applet (test). Drag price to the Response box and sq ft to the Explanatory box to recreate out bivariate model. Check the **Statistical model** box. Is the association statistically significant? Does that mean increasing the size of your house by 1,000 sq ft will increase the price of your home by $310,000?

(h) Now move the Bedrooms variable to be the second variable in the Explanatory list, so that Bedrooms appears on the horizontal axis, color-coding the homes by square footage. Can you picture the relationship between price and bedrooms among the black dots? The red dots? The yellow dots? The blue dots?

(i) What happens when you Adjust the y values? What about the x values?

(j) Summarize the squarefootage-adjusted relationship between price and number of bedrooms. (How do you interpret the coefficient of bedrooms in this model?) Does this relationship make sense in context? Explain.

(k) How would you explore whether there was an interaction between square footage and number of bedrooms?

Example 3: You can also consider interactions between two quantitative variables. For example, the file harris.xls contains data on 93 employees of Harris Bank Chicago in 1977 (being investigated for discrimination). Variables include beginning salaries in dollars, years of schooling at time of hire, and number of months of previous work experience (Diehlman, 2001).

```
Coefficients:
             Estimate
(Intercept) 5455.2691
Exper          8.6726
Educ         450.1557
Exper:Educ    -1.2624
```

(a) Interpret the nature of the interaction in context.

(b) How would you explain/illustrate this interaction to a non-statistician?

Example 4: Memory quiz

From the main Rossman/Chance applets page, you can set up a memory quiz for your students. For now, use Memory.txt by typing this into the data window.

(a) Create a graph of the Memory Scores. Check out the new violin/boxplot feature. What information is still missing?

(b) Is *sequence* a significant predictor of memory score?

(c) Do we need to adjust for caffeine and/or sleep?