Lab 4A - If the line fits ...

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

## How to make predictions

* Anyone can make predictions.
  + Data scientists use data to inform their predictions by using the information learned from the sample to make predictions for the whole population.
* In this lab, we'll learn how to make predictions by finding the *line of best-fit*.
  + You will also learn how to use the information from one variable to make predictions about another variable.

## Predicting heights

* Use the data() function to load the arm\_span data.
* This data comes from a sample of 90 people in the Los Angeles area.
  + The measurements of height and armspan are in inches.
  + A person's armspan is the maximum distance between their fingertips when they spread their arms out wide.
* Make a plot of the height variable.
  + **If you had to predict the height of someone in the LA area, what single height would you choose and why?**
  + **Would you describe this as a *good* guess? What might you try and improve your predictions?**

## Predicting heights knowing arm spans

* Create two subsets of our arm\_span data:
  + One for armspan >= 61 & armspan <= 63.
  + A second for armspan >= 64 & armspan <= 66.
* Create a histogram for the height of people in each subset. Answer the following based on the data:
  + **What height would you predict if you knew a person had an armspan around 62 inches?**
  + **What height would you predict if you knew a person had an armspan around 65 inches?**
  + **Does knowing someone's armspan helps you predict their height. Why or why not?**

## 

## Fitting lines

* Notice that there is a trend that people with a larger armspan also tend to have a larger mean height.
  + One way of describing this sort of trend is with a line.
* Data scientists often *fit* lines to their data to make predictions.
  + What we mean by *fit* is to come up with a line that's close to as many of the data points as possible.
* Create an scatterplot for height and armspan. Then run the following code. Draw a line by clicking twice on the *Plot* pane.

add\_line()

## Predicting with lines

* Draw a line that you think is a good *fit* and write down its equation. Using this equation:
  + **Predict how tall a person with a 62 and a person with a 65 inch armspan would be.**
* Using a line to make predictions also lets us make predictions for armspans that aren't in our data.
  + **How tall would you predict a person with a 63.5 inch armspan to be?**
* **Compare your answers with a neighbor's. Did both of you come up with the same equation for a line? If not, can you tell which line fits the data best?**

## Regression lines

* If you were to go around your class, each student would have created a different line that they feel *fit* the data best.
  + Which is a problem because everyone's line will make slightly different predictions.
* To avoid this variation in predictions, data scientists will use *regression lines*.
  + These line connects the mean height of people with similar arm\_spans.
  + Fill in the blanks below to create the a *regression line* using an lm, or *linear model*:

lm(\_\_\_\_ ~ \_\_\_\_, data = arm\_span)

## Predicting with regression lines

* Use the output of the code from the previous slide to write down the equation of the *regression line* in the form
* y = a + bx.
* Add this line to a scatterplot by filling in the blanks below:

add\_line(intercept = \_\_\_\_, slope = \_\_\_\_)

* Predict the height of a person with a 63.5 inch armspan and compare it with a neighbor. Ensure you both arrive at the same predicted value.
* **Measure your armspan and use the regression line to predict your height. How close was the prediction?**