Lab 2H - Eyeballing Normal

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

## What's normal?

* The *normal distribution* is a curve we often see in real data.
  + We see it in people's blood pressures and in measurement errors.
* When data appears to be *normally distributed*, we can use the *normal model* to:
* Simulate *normally distributed* data.
* Easily compute probabilities.
* In this lab, we'll look at some previous data sets to see if we can find data that are roughly normally distributed.

## The normal distribution

* The normal distribution is *symmetric about the mean*:
  + The mean is found in the very center of the distribution.
  + And the curve looks the same to the left of the mean as it does on the right.
* Use the following to draw a normal distribution:

plotDist('norm', mean = 0, sd = 1)

## The mean and sd of it

* To draw a normal curve, we need to know exactly 2 things:
  + The mean and sd.
* The sd, or *standard deviation*, is a measure of spread that's similar to the MAD.
* **Which part of the normal curve changes when the value of the mean changes?**
* **Which part of the normal curve changes when the value of the sd changes?**
* *Hint*: Try changing the mean and sd values in the plotDist function.

## Finding normal distributions

* Load the cdc data and use the histogram function to answer the following:
* **Based on what you know about these variables, which of the variables do you think have distributions that will look like the normal distribution?**
  + **Make histograms of these variables. Which ones look like the normal distribution?**
  + *Hint*: To help answer this question, try including the option fit = "normal" in the histogram function. You might also try faceting by gender.

## Using normal models

* Data scientists like using normal models because it often resembles real data.
  + *But not EVERYTHING is normally distributed.*
* As a data scientist in training, you must decide when a normal model seems appropriate.
  + No model is ever perfect 100% of the time.
  + If you choose a model, you should be able to justify why you chose it.

## On your own

* **For each of the following, determine which, if any, appear to be normally distributed. Explain your reasoning:**
  + **The weight of people in our cdc data, faceted by gender.**
  + **The difference in mean weights between Males and Females for 500 random shuffles.**
  + **The difference in median weights between Males and Females for 500 random shuffles.**