Week 4.2

* Just because “negative correlation” does it mean “does not explain variation”?
  + gf\_point(FiveVegetables ~ Smokers, data = USStates, size = 4, color = "darkred", alpha = .5)
  + gf\_point(FiveVegetables ~ Obese, data = USStates, size = 4, color = "purple", alpha = .5)
* Does the scatterplot show that PhysicalActivity explains some of the variation in FiveVegetables? gf\_point(FiveVegetables~PhysicalActivity, data=USStates)
  + Yes,it can somewhat explain the variation because some people who are more physically active eat more healthy foods.
  + I don’t think knowing a state’s PhysicalActivity helps me make a better guess about a state’s FiveVegetables.
  + I’ve seen that people who work out or do a lot of physical activity are very likely to eat vegetables and fruit.
  + States that do a lot of physical activity probably have better programs that promote health so they would also have more people who eat healthy.
  + I think knowing PhysicalActivity helps me make a slightly better guess about a state’s FiveVegetables.
  + PhysicalActivity is a good explanatory variable because it shows that people with higher activity are the same ones eating more fruits and vegetables (at least more than people with low activity).
* Check out the data but be careful not to jump to conclusions. What’s your opinion about the USStates data? Do you think that Pres2008 explains some of the variation in FiveVegetables?
  + Let’s look at the first part of this data set. Consider both sides.
    - ShuffStates <- select(USStates, FiveVegetables, Pres2008)
    - head(ShuffStates, 8)
  + What are some reasons (from the data) you have for suspecting it does explain variation?
    - Causal stories (explanatory causes outcome)
      * People who vote for Obama think more about Obama and the Obamas are healthy so thinking more about them makes voters eat more healthy. (Pres2008 causes eating better.)
    - Non-causal stories
      * Backwards causation: People who eat better are more likely to vote Obama. (Eating better causes Pres2008.)
      * Third variable/confound: The type of states that vote for Obama have other qualities (like food traditions/access, income/education) that also cause them to eat more healthy (e.g., coastal states can ship in food from other places more cheaply; better education opportunities and better income)
  + What are some reasons (from the data) you have for not suspecting it explains variation?
  + Is it possible to have gotten this pattern of data by shuffling these groups randomly?
  + Remember Westvaco… could the DGP have been random?
    - Our current theory of the DGP is FiveVegetables = Pres2008 + other stuff
    - How would we simulate a random DGP
    - ShuffStates$ShuffPres1 <- shuffle(ShuffStates$Pres2008)
    - head(ShuffStates, 8)
    - http://bit.ly/cw4-2
  + Contrast to Westvaco activity
    - population/DGP
    - Sample (the real data)
    - Sampling distribution (random process)