**Name:**

**Classwork 6**

**US States: Democrats and Vegetables**

We’re going to look at a dataframe (you explored it a little in one of the practice quizzes from Ch. 3) called **USStates**. Here are the variables in this dataframe:

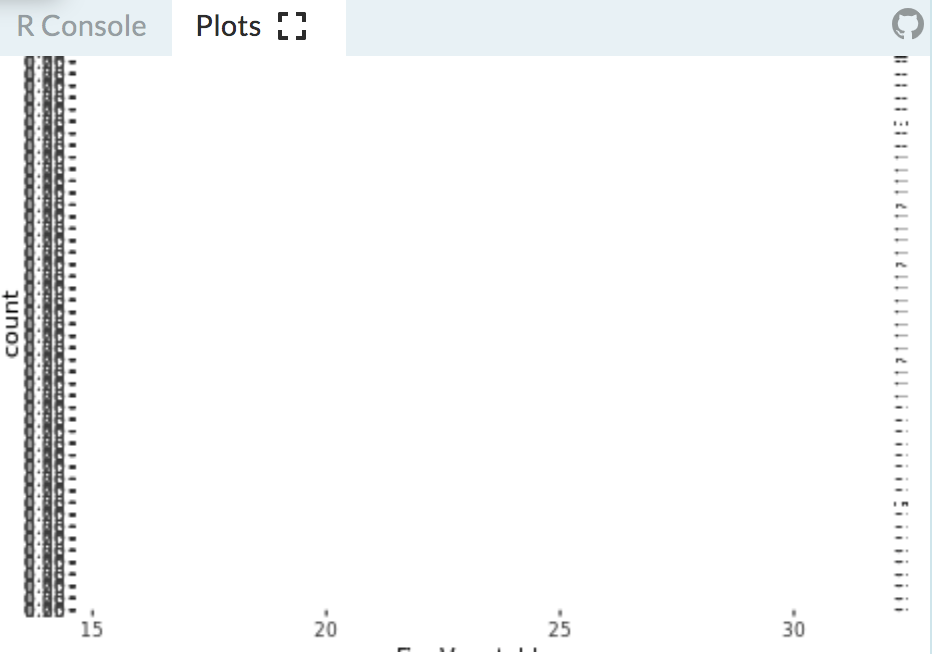
A data frame with 50 observations on the following 17 variables.

* State Name of state
* HouseholdIncome Mean household income (in dollars)
* IQ Mean IQ score of residents
* McCainVote Percentage of votes for John McCain in 2008 Presidential election
* Region Area of the country: MW=Midwest, NE=Northeast, S=South, or W=West
* ObamaMcCain Which 2008 Presidential candidate won state? M=McCain or O=Obama
* Pres2008 Which 2008 Presidential candidate won state? M=McCain or O=Obama
* Population Number of residents (in millions)
* EighthGradeMath Average score on standardized test administered to 8th graders
* HighSchool Percentage of high school graduates
* GSP Gross State Product (dollars per capita)
* FiveVegetables Percentage of residents who eat at least five servings of fruits/vegetables per day
* Smokers Percentage of residents who smoke
* PhysicalActivity Percentage of residents who have competed in a physical activity in past month
* Obese Percentage of residents classified as obese
* College Percentage of residents with college degrees
* NonWhite Percentage of residents who are not white
* HeavyDrinkers Percentage of residents who drink heavily

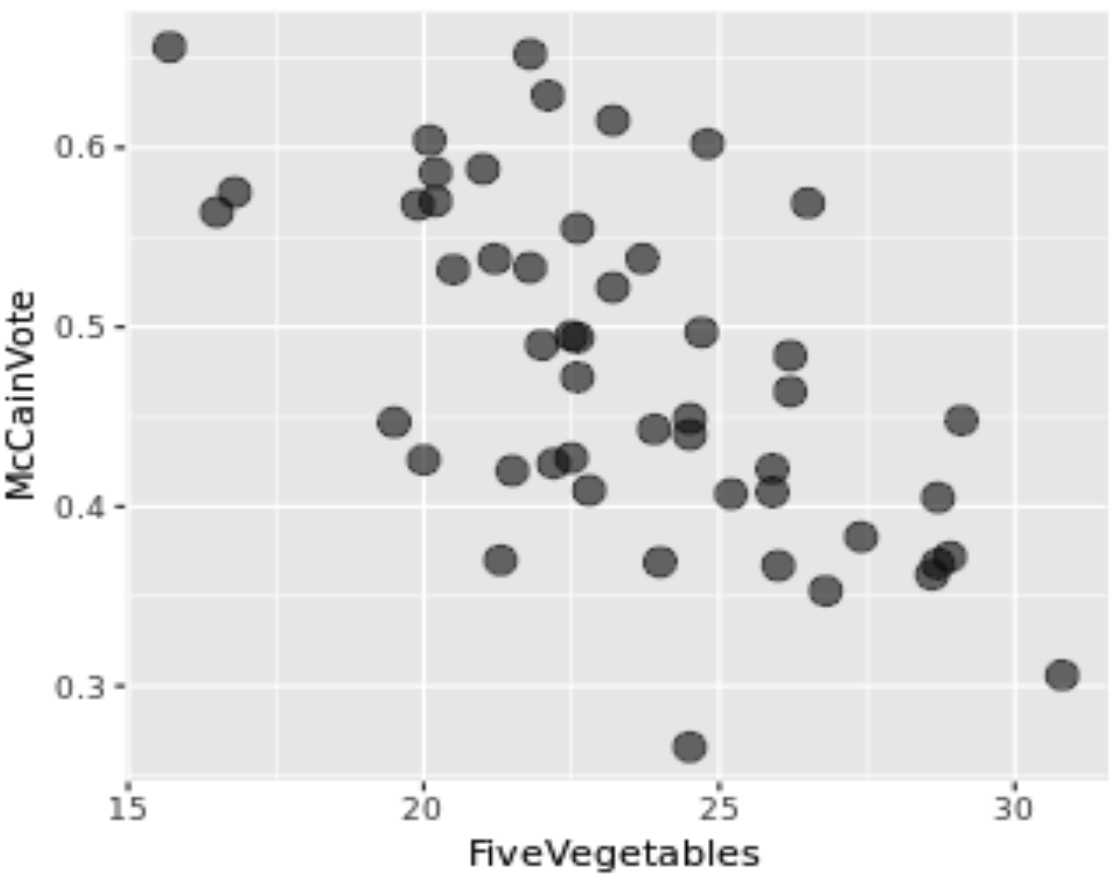
1. Does the **FiveVegetables** variable tell you how many vegetables a person ate? Why or why not?
2. I want to focus our attention on just two variables in this data set: **McCainVote** and **FiveVegetables**. How do we write R code to just print those two variables?
3. **McCainVote** and **FiveVegetables** are both percentages but one is out of 1.00 and the other is out of 100. What can we do to make them both consistent?
4. I want to look at whether some of the variation in the percentage of votes for John McCain (in 2008) is explained by **FiveVegetables**. If I run this code, I get a very unfortunately looking plot. Why?

gf\_histogram(~ McCainVote, data = USStates) %>%

gf\_facet\_grid(FiveVegetables ~ .)



1. I also tried making a boxplot (using gf\_boxplot) and a scatterplot (using gf\_point). Which one helps us explore the idea that some of the variation in **McCainVote** is explained by **FiveVegetables**? Why is that one better?

1. How would we write a word equation for the hypothesis that **FiveVegetables** explains some of the variation in **McCainVote**?
2. Why do we see this pattern of data? What are your ideas about the Data Generating Process (DGP)?
3. Is it possible for healthy eating to *cause* voting behavior? Why or why not?
4. Is it possible for voting behavior to *cause* healthy eating? Why or why not?
5. Are there any variables that might cause changes in both voting behavior and healthy eating?
6. Check out the data but be careful not to jump to conclusions.
   1. What are some reasons (from the data) for suspecting that **FiveVegetables** really does explain some of the variation in **McCainVote**?
   2. What are some reasons (from the data) for suspecting that **FiveVegetables** does not explain some of the variation in **McCainVote**?
   3. Is it possible to have gotten this pattern of data if we just shuffled these so that a value from **FiveVegetables** is randomly paired with a value from **McCainVote**?
7. Is this a well conducted experiment? What does that mean for us as data scientists?
8. Time to shuffle some papers! Grab 8 post-its. Write the values of the outcome variable on each post-it (try to write it on the bottom half of your post-it) and try make a scatterplot. What do you notice?

**Simulating a Random Process (Lab)**

1. Remember Mueller? How did we “simulate” a random process with paper? Which R function acts like that?
2. Now we have a slightly different situation – we have outcome values that are **not** in two groups. How can wemimic a random process with paper? How can we mimic that process with the shuffle() function?
3. How would we create a visualization of the shuffled data? Write the code here.
4. How is that different from the code for creating a visualization of the actual data?
5. Let’s go back to our “whole thing” diagram. If we create a bunch of scatterplots from shuffled data, where would we put that in this diagram?
6. If we shuffled the **FiveVegetable** values and randomly paired them with **McCainVote** values, would our actual sample look similar or different from those random shuffled?
7. One of the scatterplots below is the empirical sample. Can you tell which one it is? What are you looking for that is different in the empirical sample?

|  |  |  |
| --- | --- | --- |
| 1. | 2. | 3. |
| 4. | 5. | 6. |
| 7. | 8. | 9. |

1. Do you think the likelihood of getting a pattern of data like the empirical sample from a random process is high? Low? Medium? Explain your reasoning.
2. If the likelihood of getting a sample like the empirical sample from a random process is low, what theory of the DGP would that rule out? What theory would it support?