USP 634 Assignment 1: Katie Conlon 4/19

1.) a.) Determine the type for all variables (if total number of variables in your dataset is less than 6; otherwise select 6 variables including both continuous and categorical variables) in your dataset and select 2 continuous and 2 categorical variables for the following exercises.

From the GAIN index I have picked out the variables of 'food,' 'water' and 'habitat.' These are all continuous variables that are in an index form where 1=highest and 0=lowest. From the CITI database, I have picked DOMMOV (freedom of domestic movement), FORMOV (Freedom of foreign movement), and SPEECH (freedom of speech). These are all categorical variables that are ratings on a country basis.

b.) For each of the 2 continuous variables, calculate these summary statistics:

a. mean; b. mode; c. median; d. range; e. interquartile range; f. variance; g. standard deviation. And then show with appropriate graphs and describe the distribution of each variable. Which of these two variables resembles the normal distribution more closely?

	Mean	Mode	Median	Rng	IQ Rg	Var	SD
Food	.5315	0.66741	.5315	.8042	.287	.034	.18486
Water	.4611	0.67785	.4713	.9892	.0913	0.03134	0.17709

c.) For the variable more normally distributed in 2), calculate the Z-scores for all the observations in your dataset. Next, choose two observations and identify the Z-scores for them. Assuming that this variable is normally distributed (even if it isn't), what proportion of observations would be predicted to lie between these two Z-score values? How does this prediction vary from the actual number of observations, and why?

Z-scores are a way to compare results from a test to a "normal" population. A z-score is the number of standard deviations from the mean value of the reference population

Formula: $z = x - \mu / \sigma$

So for instance, with the water data a Z-score of 1 would be one SD above the mean (-1 would be below the mean). So a Z score of 1/-1 and 2/-2 for water would be: .638/.284 & .815/.1069
For food the 1/-1 and 2/-2 Z scores would be: .7164/.3467 and .901/.16184

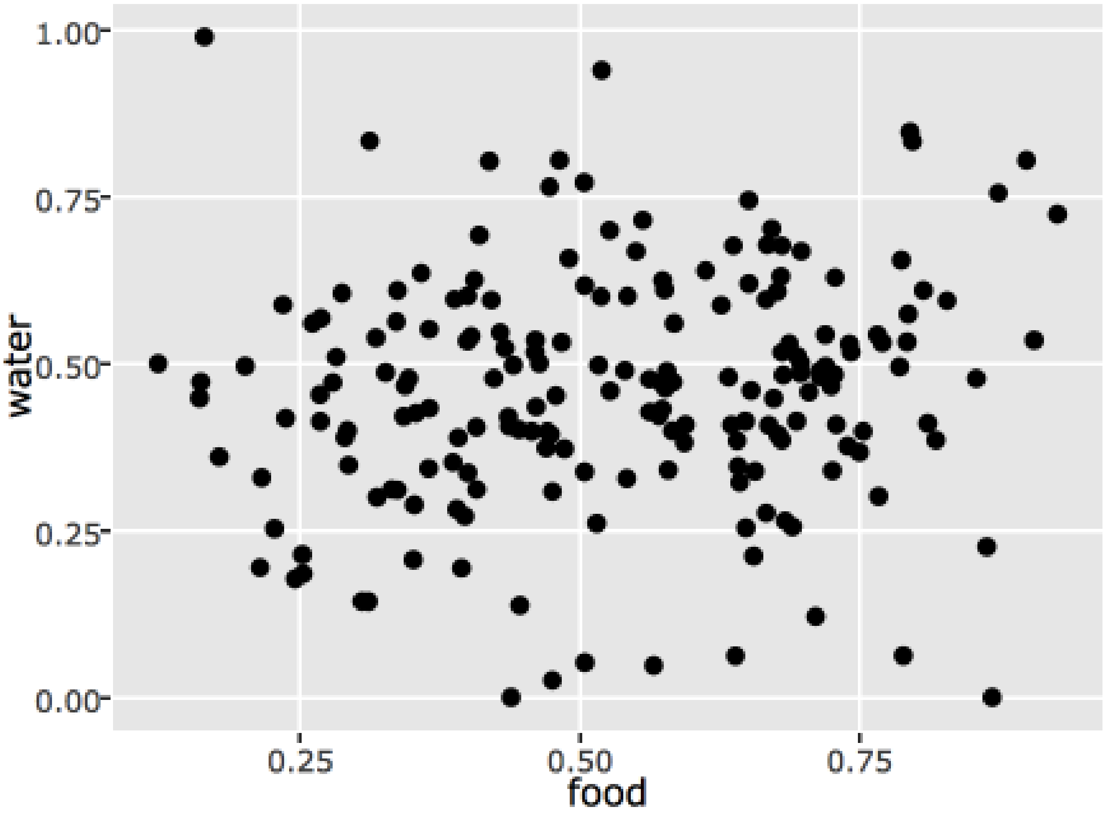
Sample Observations: Food - .7702761 Water - 0.5601717

Food: .7702761- .5315/.18486 = 1.292=
0.901821 probability or 90%
This observation is 1.2 standard deviations above the mean. 90% of the data fall below this number (this represents very high food security in relation to the other data).

Water: 0.5601717 - .4611/0.17709 = .55944 =
0.712069probability or 71%
This observation is .55 standard deviations above the mean. 71% of the data fall below this number, which means the majority of countries have less than 50% rating for water stability and security.

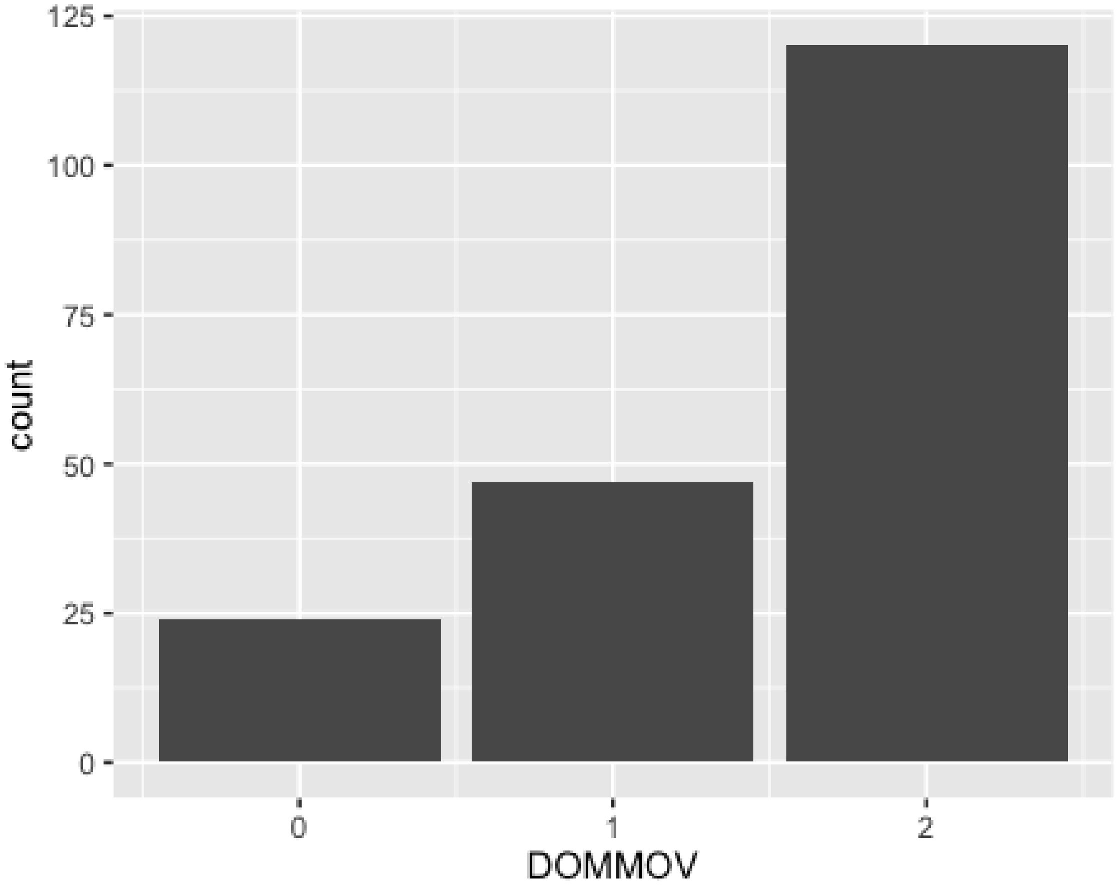
d.) Show with appropriate graphs and describe the relationship between the two continuous variables. Are they dependent? If so, positively or negatively?

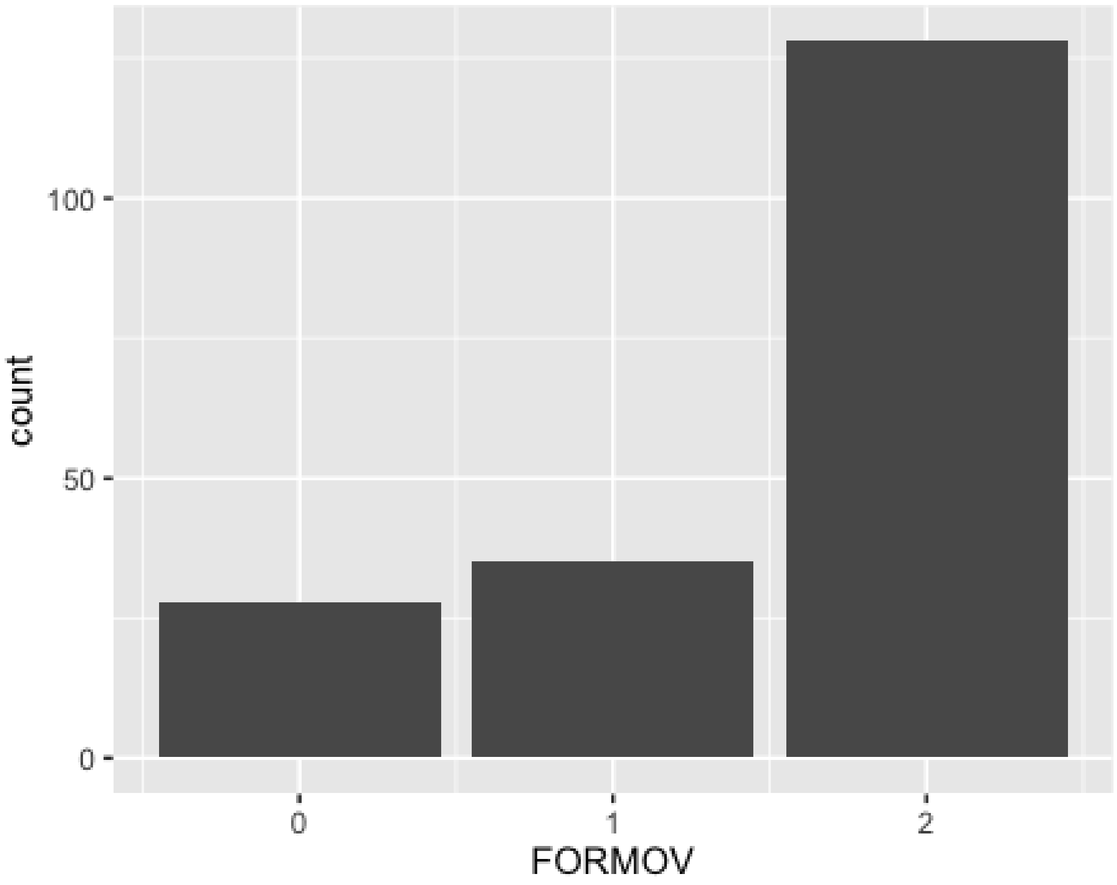
Water and Food are independent continuous variables. The relationship shows very little correlation, however, one might hypothesize that the greater the food sensitivity the greater the water sensitivity and visa versa. It would be very hard to draw a line through these points. It would be interesting to see what countries are in the bottom left square (the most vulnerable).

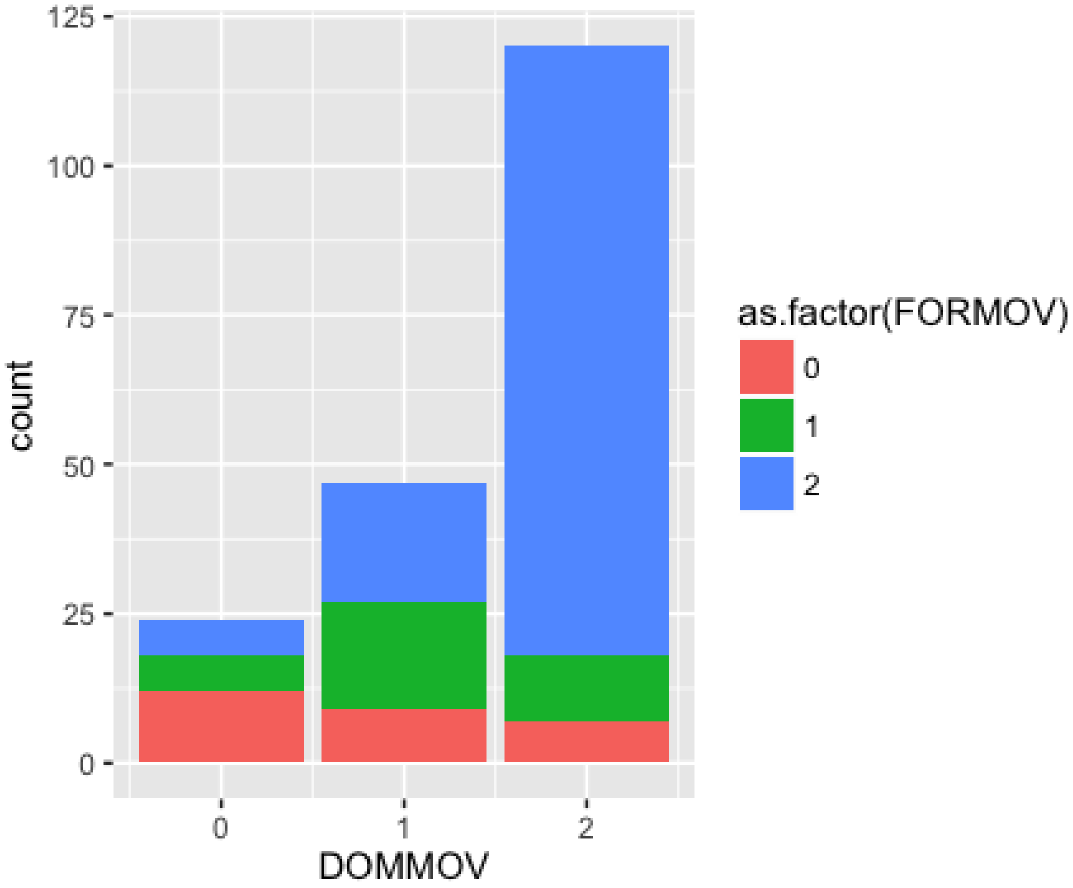


e.) For each of the 2 categorical variables, show with appropriate graphs and tables and describe its distribution. Show with appropriate graphs and describe the relationship between the 2 categorical variables. Are they dependent?

The 2 categorical variables are not dependent, but they are similar. In the DOMMOV, we can see that around 120 countries in the world have basically no rights (as defined further in the CITI coding doc) to domestic movement. Only about 25 countries are labeled as 'free' (this would not even include all of Europe, the US, Canada and AU – I will be curious to see which countries are excluded). The FORMOV data is interesting in the fact that more countries allow complete freedom of FORMOV over DOMMOV. However, both data sets show that globally, movement is highly restrictive both internally and externally for the majority of citizens.

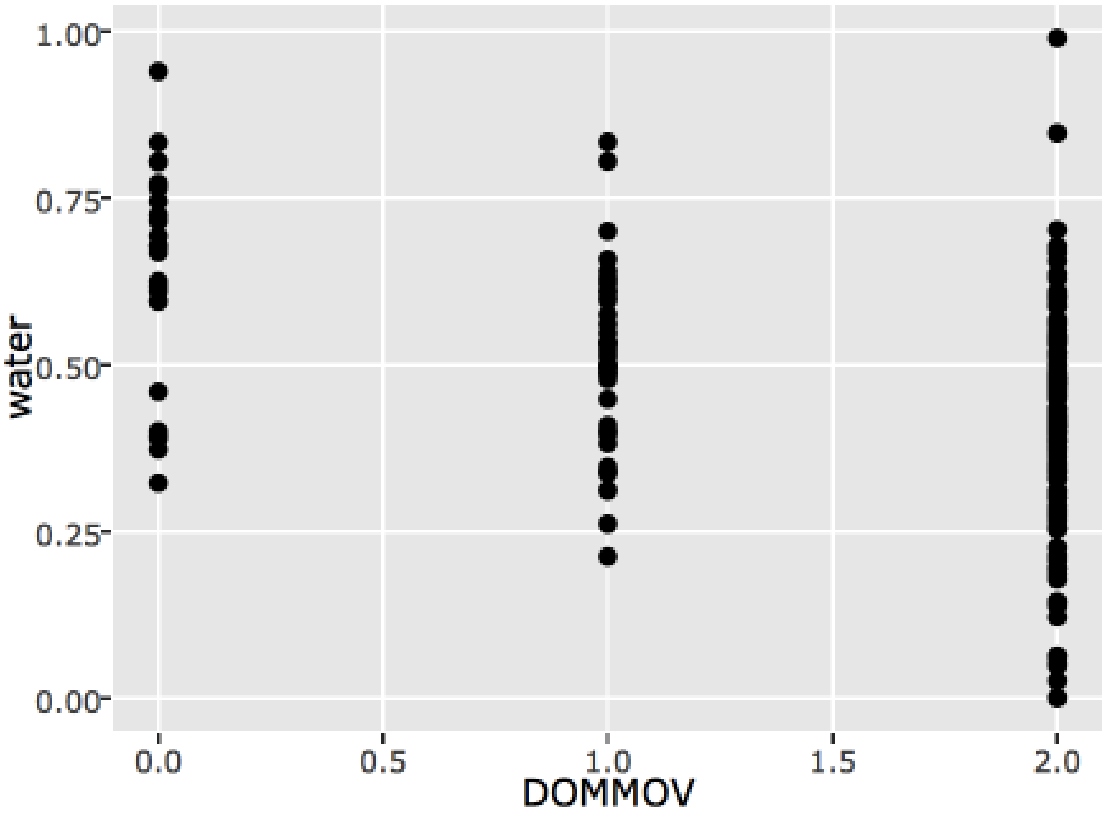



100
50
0
count
0
1
2
FORMOV



f.) Select one continuous variable and one categorical variable, show with an appropriate graph and describe the relationship between the two variables.

This scatterplot compares water vulnerability to DOMMOV (freedom of domestic movement). The graph shows that countries with greater water vulnerability are also countries that restrict internal movement. In the case of climate change and water shortages, for people living in vulnerable areas, this information suggests that vulnerable people will be put in more dire situations with little recourse for movement.



2. What is the probability of rolling two 5's with two fair dice? What is the probability of rolling snake eyes (two ones) twice in a row, followed by a four and a six, followed by a score adding to 10?

Two 5's: 1/6 x 1/6 = .02777 or 2.8% chance

(independent probability of event/outcome)

snake eyes etc :

1/6x1/6x 1/6 x1/6 x (2/36) = [.00077160493x.0555555555]= .00004286694

=.004% chance

(1/6 x 1/6 = the probability of getting 2 specific numbers on the roll, there are 2 ways to get numbers rolled to 10, so 2/36)

3. In 2005, the average annual ozone levels in Smogsville were normally distributed with a daily mean of 100 ppb (parts per billion) and a standard deviation of 25 ppb.

How many days in 2005 were smog levels either above 75 ppb (their air quality standard) or below 50 ppb?

34.1% +50% = 84.1%

.84 x365 = 307 days in 2005 above 75 ppb

[With Normal Distribution on a Bell Curve:
68% of the distribution lies within one standard deviation of the mean.
95% of the distribution lies within two standard deviations of the mean.
99.7% of the distribution lies within three standard deviations of the mean.]

4.) Inferring the direction and existence of causal relationships from observational data is plagued by selection bias, and reverse causality, and confounding variables (a third variable, or a number of other variables influence both explanatory and response variables). The following empirical patterns have been cited in press reports as potential evidence of causal relationships.

• Oakland is considering a Fresh Food Financing program that incentivizes grocery stores to locate in East Oakland. This program is based on studies showing that residents of neighborhoods without stores selling fresh foods have an unhealthy diet.

Response: This is faulty reasoning because there are other confounding factors at play and it is not correct to assume that lack of fresh foods causes an unhealthy diet. For instance, the price of fruits and vegetables might be a factor so merely having them available wouldn't make a difference if people didn't have the money to buy them. Or, maybe people like to go to a local market over a chain/non-cultural market. So, giving financing to grocery stores 'from outside' to locate n E Oakland might not create greater patronage, it might aggravate already tense racial situations. Preference might also play a part in the decision-making process of what to buy. Maybe people in E. Oakland don't like fresh foods or don't think to purchase them because they are not accustomed to it. There could also be an educational factor at play also, if people do not realize what kinds of foods are best to create a balanced nutrition. Etc. etc. Essentially, there are many other cultural and socio-economic factors at play besides access to more food options which show alternative interpretations of the data.

• Two percent of residents in Fresno, CA bike to work while eight percent bike in Berkeley. Berkeley has 50 more miles of bike lanes on their roads than Fresno. Therefore, if Fresno were to add more bike lanes its bike ridership would increase.

Response: Similar to the Oakland case, there could be other confounding socio-economic and cultural factors at play. Berkeley is a college town, with a high percentage of educated people and young people (more inclined to ride bikes) vs.

Fresno is a predominantly farm area where people might favor driving trucks etc for work reasons. Also, this model assumes that bike lanes came first and then the riders. However, maybe there was a reverse causality in Berkeley where there were riders first, who were highly politically engaged (it is Berkeley after all), and then they demanded bike lanes. So, if there is not a ridership in Fresno to call for more bike lanes, putting them in doesn't necessarily mean they will come. Also, the temperature in Berkeley is more temperate than Fresno, which might make some more inclined to ride there vs. in the dust and heat of Fresno. Thus, these show altered interpretations of the data.

- A recent study in Minneapolis found that people who live in neighborhoods where the majority of houses have porches are more likely to talk to their neighbors at least once a week in comparison with people who live in neighborhoods where there are few porches. To encourage social cohesion in neighborhoods, Minneapolis is therefore considering a new grant program to help people add porches to their houses.

Response: This could be a case of reverse causation. Maybe people who like to talk to their neighbors move to neighborhoods with more porches. This could also be a case of selection bias, where the researcher only went to homes with porches to see if they were more conversant with their neighbors. It would have been similar if they choose something like households with dogs are more likely to use their local parks. Thus, it would be better to create a random sample of all the neighborhoods in Minneapolis and that way get a better distribution of households that talk to their neighbors, regardless of if they have a porch or not. This would give a different interpretation of what kind of people talk to their neighbors.