Part 1:

1. According to the summary of Mod01, for every increase in the unit of poverty (x-value), the voting percentage decreases by 0.67 units. This means that as poverty increases, so too does voting. A p-value of 0.007, a number close to zero, tells us that we can reject the null hypothesis and confidently conclude that there **is** an association between poverty level and voting behavior.
2. There is a positive correlation between voting and percent of population over the age of 65. Given a p-value that is extremely close to zero (1.86e-07), we would reject the null hypothesis and conclude confidently that there is an association between the two variables. Compared to the association between voting and poverty, the poverty data is negatively correlated while the senior data is positively correlated. Based on the slopes for each dataset, the voting and senior data is stronger because the slope is bigger (|.96| > |.67|) and has a greater strength of effect. Additionally, looking at the r-squared values, the poverty-voting value is 0.1 and the seniors-voting value is 0.37. These values define the variation in y described by the variation in x. Given that the seniors-voting value is bigger, this also shows that this correlation is stronger.
3. Within each region, about half (~50%) of people in each of the four regions vote, with the range of mean values equating from 45% - 56%. The variation amongst the data is little but the difference in dispersion is large as the range of values for percent of votes across the regions has a broader standard deviation. The PCTOBM means have a wider range with a minimum of 29.2% in the Great Plain region, 42.8% along the Western Slope, 52.5% in the Mountain region, and 55.5% in the Front Range.
4. The F-value for the ANOVA test is 3.772. This means that the variation that we expecting between the means is greater than the variation that we would expect to see by chance. 3.772 is a high F-statistic and suggests a large variation between the means of the data. This means that the voting behavior in regions varies more than we’d expect it to if the population was truly equally randomly distributed. Other regional distinctions that might be worth examining might be income, demographic spread, or education.
5. When the k-squared value is less than 1.5, the data is not homoscedasticity. Given this Bartlett test, the k-squared value is 6.52, meaning that this data is not homoscedastic, but rather heteroscedastic. This means that the variance is different in all of the values that were tested and the assumption of homoscedasticity is not satisfied. We can use the ANOVA test because it considers this. Given the p-value of 0.09, we cannot confidently reject the null hypothesis, meaning that there is not a strong likelihood that there is a correlation between the two variables.
6. The p-value for the Kruskal-Wallis test is 0.09004, higher than the Bartlett test. We would still not be able to confidently reject the null hypothesis because the value is higher. In terms of achieving significance in a parametric or non-parametric test, the parametric test is going to be a bit more sensitive to the data because more of the conditions of validity have to be met in order for the data to fit the parametric structure. The parametric test is more likely to allow you to reject the null hypothesis because they have to meet more conditions of validity.
7. Looking at the ggplot, the graph is showing the spread of poverty and voting behavior in the four regions of Colorado. It appears that the range of poverty is the smallest along the Front Range and the largest in the Mountains. The 95% confidence interval extends from about 8 – 26 (~18) on the percent of poverty (x-axis), which is a large range compared to the other three regions that fall below this range number. All of the regions have negative regression lines, meaning that as poverty goes up, voting behavior goes down. The Western Slope has the highest amount of voting behavior while the Great Plains have the lowest.
8. The data is relatively well-behaved as it follows a negative L-curve with minimal outliers not following this pattern. Though it’s not a linear relationship, generally, as the elevation goes up there is less annual precipitation.
9. The values vary across the different variables, but generally the elevation can tell us a lot about the weather. For example, the higher the elevation, the lower the annual temperature; the annual temperature in January is lower as elevation goes up; as the elevation increases, the annual precipitation decreases; and annual precipitation generally decreases as the elevation increases.
10. According to the correlation matrix, there appear to be a positive correlation between latitude and precipitation but a negative correlation between precipitation and longitude and elevation. There doesn’t appear to be a pattern amongst the temperature fields, except that all of the temperature in January fields are negative with latitude, longitude, and elevation. I wouldn’t say that any of these three variables do a particularly good job of predicting temperature or precipitation. The one in this scenario that does the best job I would say is elevation at predicting temperature, as its values for tjan and tann are the closest to |1|.
11. The mean difference between double carbon levels and normal carbon levels is 4.09. The t-statistic for this data is 9.43. This provides good evidence that we could reject the null hypothesis that the two means are going to be equal, and accept the alternate hypothesis that they are different.
12. In the two-tailed test, the t-test is testing assuming that the null hypothesis is true—that the two means will be equal. The p-value for this is 0.06, meaning that there is a decent chance that we’ll be able to reject the null hypothesis. This means that we are pretty sure that we can expect the means to be different between Jan1x and Jan2x. In the one-tailed test, the t-test is testing the same information but against the added variable that the alternative hypothesis of the mean of Jan2x is greater than the mean of Jan1x. The p-value for this test is 0.03, which is an improved significance for us to reject the null hypothesis and have confidence in the expectation that the Jan2x mean will be greater than the Jan1x mean.

Part 4:

Research question: is there an association between rate of heart disease in American adults and location of food deserts across the country?

The primary spatial dataset that I will be working off of is from the National Vital Statistics System (NVSS) with the indicator of “rate of heart disease (diseases of the heart) mortality among US adults (18+)”. The data is in point form within each US state. The attribute table contains columns for the low and high ends of the confidence interval for the indicator of this data. I will need to figure out how to read and interpret this data in order to overlay it with food desert data and identify whether or not there is a statistical association between the two variables.

Centers for Disease Control and Prevention National Center for Health Statistics. “National Cardiovascular Disease Surveillance Data.” *Center for Disease Control and Prevention*, National Vital Statistics System, 29 June 2020, chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/National-Vital-Statistics-System-NVSS-National-Car/kztq-p2jf.