The statistical questions for the project are:

1. Health Outcomes: Everyone wants to experience long and healthy lives, yet places have different resources and opportunities. To understand the health outcomes in a community, we measure both length and quality of life by county within Texas.

2. Is there a significant statistical difference between rural and urban single parent households? Do Texas rural counties have a higher proportion of single parent households than urban counties?

3. Are high school graduation rates along with reading and writing scores better in urban schools? There is a myth that urban schools, or more specifically affluent suburban schools, are better than rural schools. Unfortunately, the dataset does not include this category only Urban and Rural.

4. Do country people have less health concerns and live longer than city dwellers? When we think of the country, we think of healthy living and fresh clean air.

The outcome of my EDA is that the distribution of single parents is similar in both samples. Although the rural population is only approximately 10 percent of the total population, the distributions are very similar for both samples. It would be interesting to see if the distribution is similar in other States. As far as the education goes, knowing that students are pushed through the system so the school districts can make “the grade”, I can see the reason for the skewness in the KDE and the histograms. Because of this, the math and reading schools are not a helpful indicators of graduation rates. The State of Texas has standardized testing and teachers simply teach towards this test. I know I am injecting my biases but even the results show an awfully contrived skewness towards the right. The distribution I would say is far from normal. The results of the quality of life factors are quite surprising. It shows a lack of sleep as being a factor which has a close correlation to poor health. We’ve heard about the lack of sleep and how it affects motor skills, concentration, etc. But it is surprising to see how statistically poor sleep and poor health go hand-in-hand. Other factors with positive correlations to poor health, as I suspected, are availability of Physicians, lack of health insurance, lack of availability of food, and housing cost burden. Housing cost burden I can see as a reason for poor health when a family is faced with putting food on the table and a roof over their heads, they will more than likely forgo health insurance and trips to the doctor.

While the quality of life factors have very close correlations with poor health and to each other, these factors have a small or non-linear relationship to life expectancy. In fact, in the Rural sample several factors have a small negative correlation to life expectancy. Surprising, even poor health had a small correlation to life expectancy. So statistically speaking, barring any non-natural fatalities, the mean for Life expectancy is 78 for city dwellers and 77 for rural denizen.

The source dataset contains a large amount of data. I chose to use only the non-ranked measures to view the statistical outcomes. For a more thorough analysis and probably more accurate results, I would have liked to see additional grade-level measures. The dataset does have the category of some college. I perhaps could have used this variable to assess the truthfulness of the high school graduation rate. But then again there are other factors to consider. The only other variable I did not use is the population counts for rural areas within urban counties. I did not use this figure in the single parent issue because the percent single figure did not show the percentage for these rural areas. Only by county. The same reason I didn’t use the rural counts for the quality of life question. The dataset only identifies the counts for these factors by county. I created the two samples by splitting the dataset by the population. Rural counties have 50000 or less inhabitants and the rest are considered urban.

The only challenges I had was interpreting the CDF. I still struggle with that. The other challenge was trying to use the difference of the means code from thinkstats2. I kept getting an error so I went with what I thought were equivalent or similar hypothesis tests. It’s still “Greek to me”. But going through this project has started to bring all the concepts together.