SI 649: Communication Visualization Project
Static Blog Post

Mingliang Duanmu
duanmuml@umich.edu

October 31, 2021

1 Learning Objectives

The viewer will be able to:

- **Compare** the natural amenity rating and geographic distribution across different states in the US.
- **Conclude** the diseases with highest mortality in Red Lake County and the trend of mortality during the past 30 years.
- **Contrast** the unemployment rate, average household income, and poverty rate of Red Lake County to the national and state-level data.
- **Observe** the difference between educational resources within the state of Minnesota.

2 Design Process

Idea

Most of my ideas of building the visualizations are based on the aspects mentioned by the author in the article "I called this place Americas worst place to live. Then I went there". The static version of my visualization focuses on comparisons of Red Lake County statistics against national average or counties within the same state. I apply various methods of encoding, representation and annotation to make it quick for readers to notice the contrast. I also go a little beyond to display extra aspects of Red Lake County that add to a more complete impression of the place for readers.

Data

I search extensively online to find useful and credible datasets for my visualizations. The data sources include Kaggle (the mortality data, public school data), USDA (natural amenity), U.S. Census Bureau (poverty and household income), U.S. Bureau of Labor Statistics (unemployment rate).
Based on the date of the article, all my visualizations are based on the data before 2015. I do some data cleaning and filtering on the raw data to kick out wrong values and useless information, as well as reorganizing structures of tables to make them more suitable to Altair grammar.

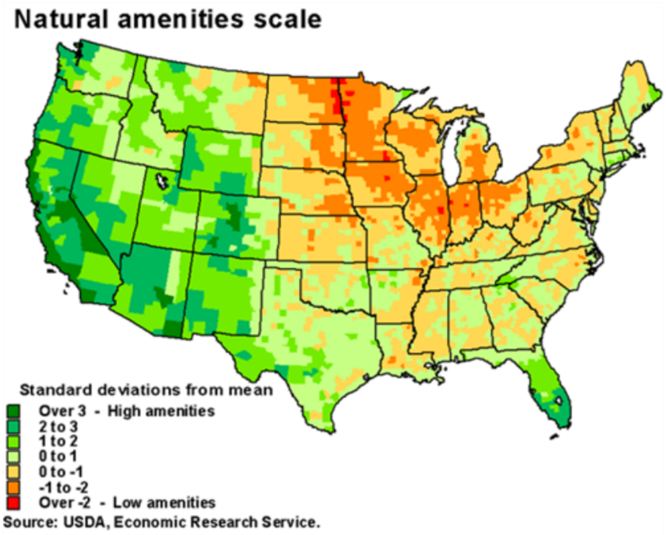
Tool

I use both Altair and Tableau for my static visualizations. Tableau is convenient for plotting map graphs while Altair has more freedom of customizing all the parameters of a graph, together with versatility to add layers and organize multiple subplots. Also, my interactive visualizations can be directly built upon the static Altair graphs. Personally, I prefer programming-based tools instead of software with rich GUI that requires more learning cost.

Visualizations

Map graph of national natural amenity distribution

First things first, I focus on the most important fact that leads to the conclusion of "Americas worst place to live", which is the natural amenity rating. I think using a map graph showing geographical difference and distribution of such an index is the most clear and straightforward way.



Source: USDA, Economic Research Service.

Figure 1: Colored choropleth map on USDA

At first I wanted to use Altair for plotting a colored choropleth map inspired by Figure 1, however due to the lack of documentation and tutorials, together with Altair's inability of making zoom interactions on the map graphs, I switched to Tableau for this plot in Figure 2.

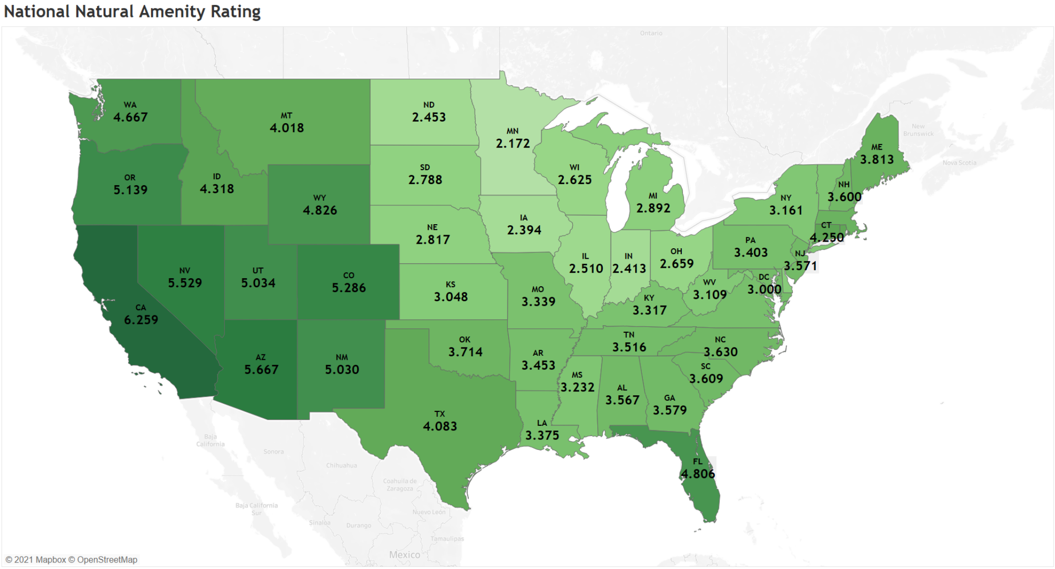


Figure 2: My colored choropleth map created by Tableau

Figure 3 is my failed attempt to encode values in color in Altair, I can only draw scatter plots using longitude and latitude of each county.

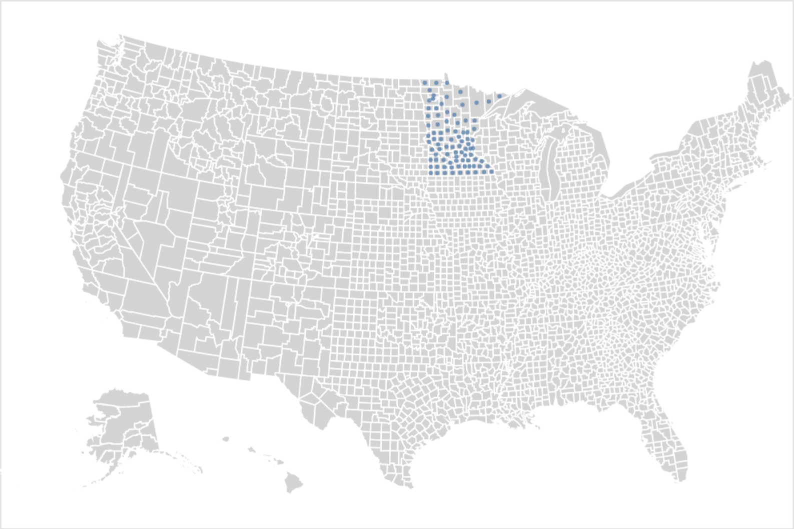


Figure 3: I cannot figure out how to plot data on map using Altair

Line & scatter plot of unemployment rate, average household income, and poverty rate

Since the author claims that Red Lake County does a good job by economic metrics regarding unemployment rate, average household income and poverty rate, I decide to use scatter plot to compare them to other counties within the state, as well as the national data. While the unemployment rate I find only applies to each state, I use a line plot to compare it with the national data.

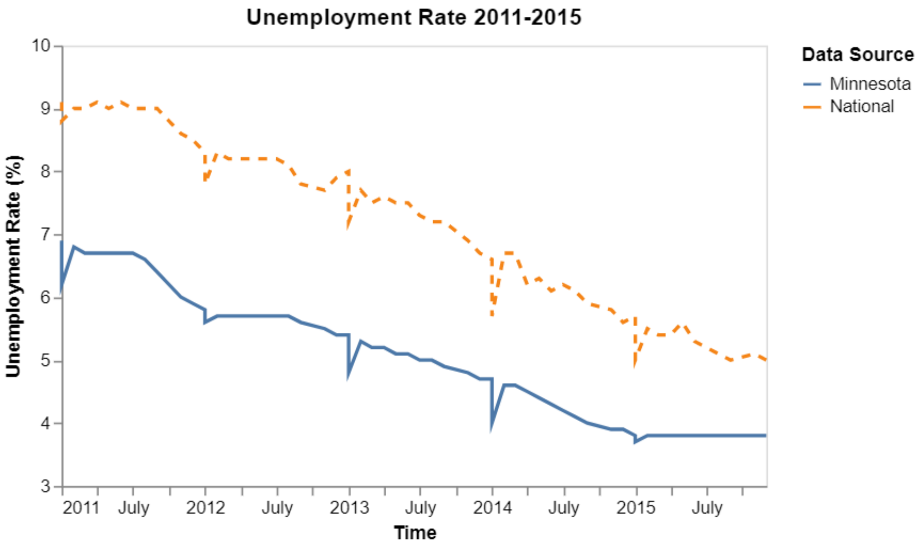


Figure 4: Line plot of unemployment rate from 2011 to 2015

For the scatter plot, I try different encodings of three features on both axises and size/color, to minimize influence on the size of graph by outliers. Finally I came up with the parallel plots in Figure 5.
The position of Minnesota is marked with blue color on the left plot. The dash lines indicate average value of each dimension on the left plot and the value of Red Lake County on the right plot. Both plots share the same y-axis.

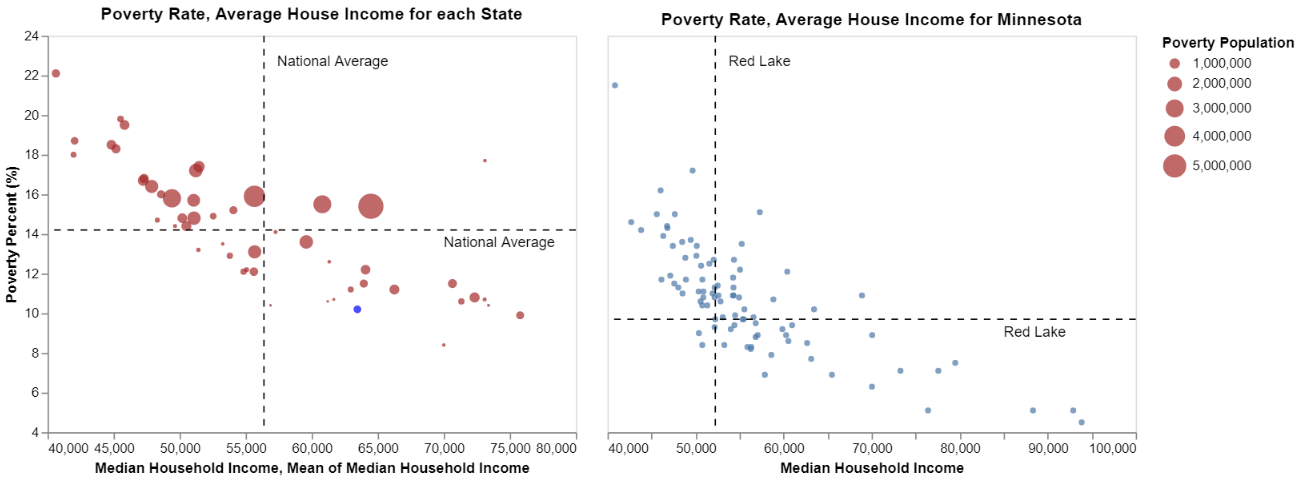


Figure 5: Poverty rate, poverty population and household income

As shown in Figure 6, I tried log scale on the x-axis but I think it prevents readers from knowing the real numbers and it makes deviation from average value look smaller for large data. Also, The distance between dots are narrower.

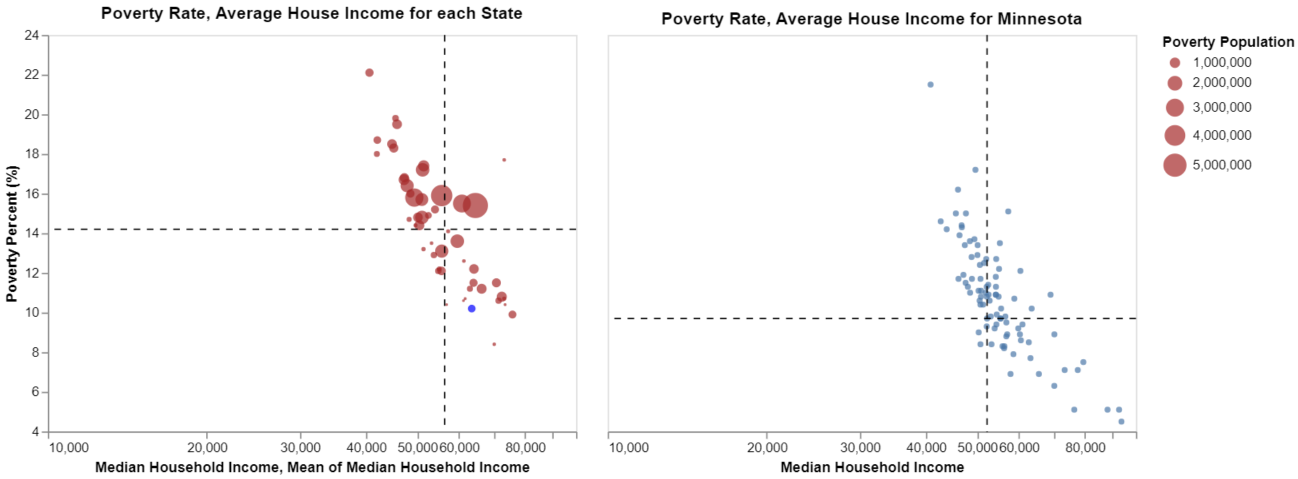


Figure 6: Poverty rate, poverty population and household income, log scale x-axis

Bar chart of public school statistics

As the author mentions education as an important thing in Red Lake County, I decide to contrast the level of educational resources of the county with other counties in Minnesota and the average domestic statistics. I only include data of public schools because there is no private ones in Red Lake.
I created two calculation fields, one is student per teacher and the other is enrollment rate. A small number of students assigned to each teacher means the educational resources is more abundant. The enrollment rate, which equals to enrollment number divided by the right-age population, is a good indicator of the willingness of pursuing education. The dash lines in Figure 7 stand for national average values.

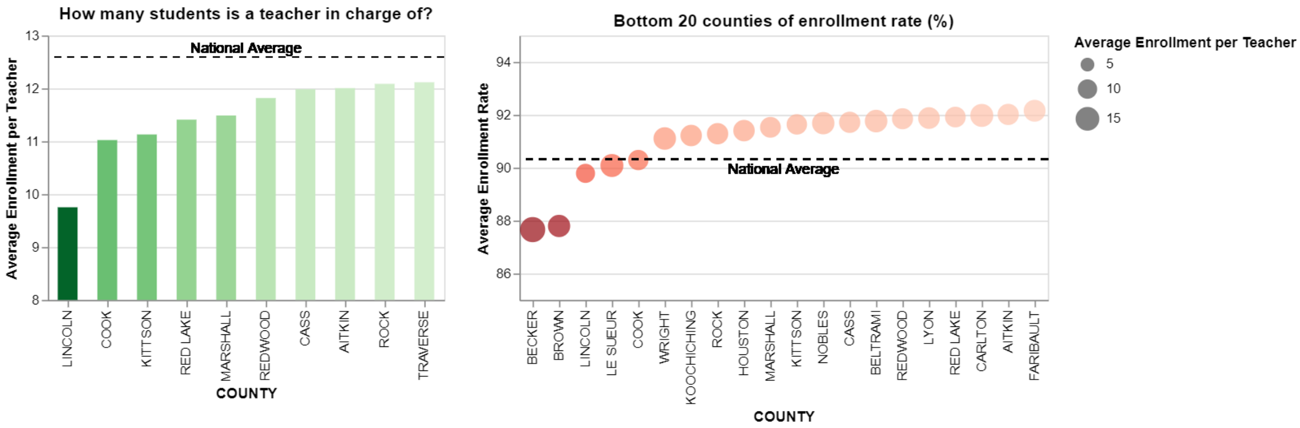


Figure 7: Least enrollment per teacher counties and bottom counties of enrollment rate

What I dislike about the graph is that the size encoding on the right graph is not distinct enough, since the range of average enrollment per teacher is narrow.
Previously I tried to include the top/bottom 5 values for the left plot in Tableau as Figure 8, but I found there is no necessity and does not add to expressiveness.

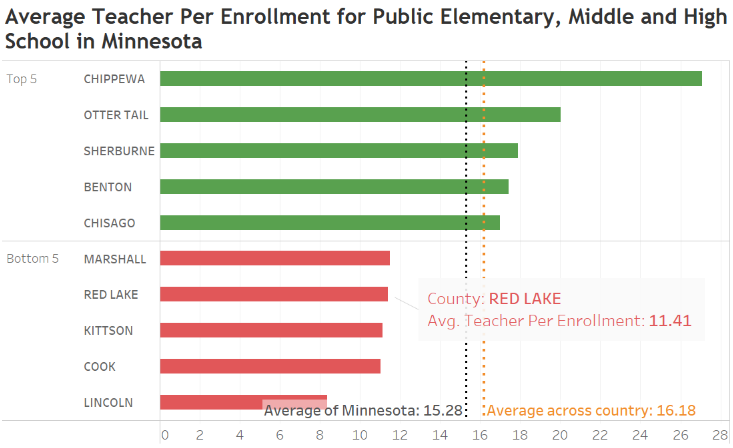


Figure 8: Top/bottom 5 counties for enrollment per teacher

Line & heatmap plot of mortality

In this graph, I go a little beyond the content of the article, but I think health is also an important evaluation of whether a place is suitable to live in. I want to see the major diseases causing death and if the health condition improves with time in such a "worst place to live".
I use a line chart, a scatter plot and a heatmap to display the top 5 mortal diseases in Red Lake and the change of mortality with time. I also make a comparison with national value on the top 5 diseases in Red Lake in terms of the percentage of change in mortality from 1980 to 2014. I like the composed plot as it shows rich information in trend, number, and comparison, together with a good visual effect.

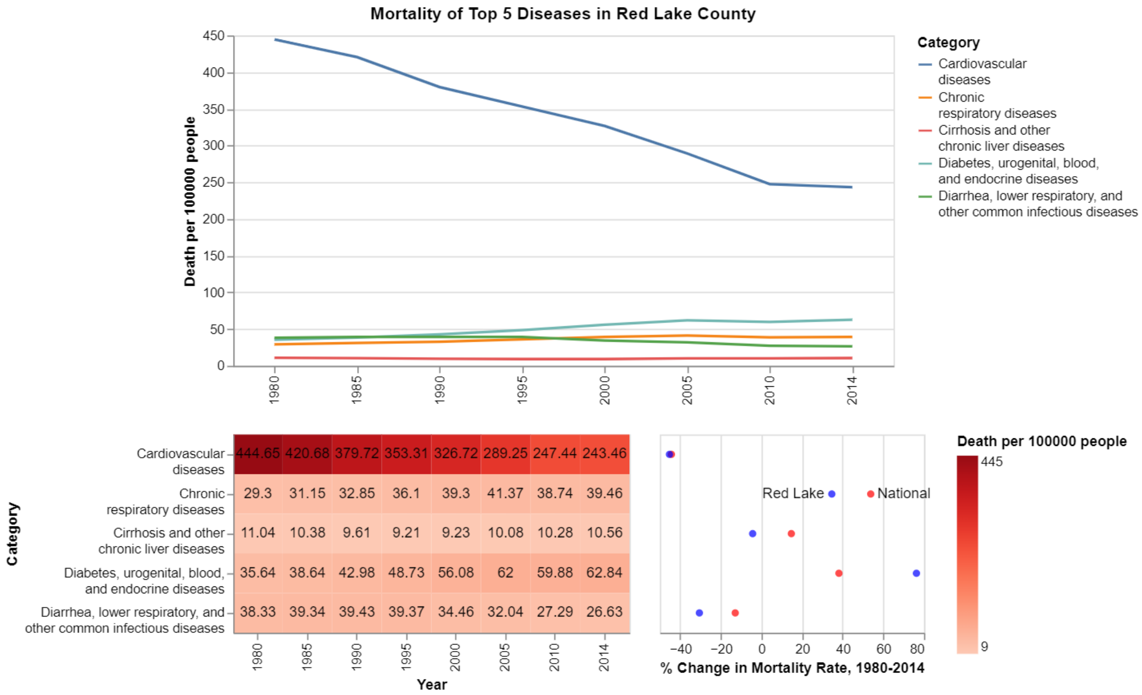


Figure 9: Composed plot for top 5 mortal diseases in Red Lake

As the dataset provides min/max values for each year's data, I planned to embed those boundaries into the line chart and the scatter plot, however it appeared too crowdy on the graphs and it is enough for readers to figure out the trends and features of data without knowing the specific ranges.

3 Final Design

- **Perception**
  Color: I use different color encodings in my visualizations that conform to reader's perception habits. I use green color to encode data containing positive aspects like natural amenity rating and high educational resources. In contrast, red color represents negative data like mortality and bad enrollment rate. For example, higher natural amenity rating leads to darker green (Figure 2), lighter red represents less fatality for a disease (Figure 9). In the scatter plot, there is a color consistency between the highlighted data point on the left graph and dots in the right graph (Figure 5).
  Cognition: In the scatter plot, I create a pop-out effect using a contrast color to identify the point we want to observe (Figure 5). For multiple lines crossing each other in one chart (Figure 9), color encoding is a good way to reinforce Gestalt Principle of continuity for readers.

- **Design**
  Data Ink: I pay careful attention to maximize the data ink ratio in my visualizations. For time-oriented axis (Figure 4, 9.1) and categorical variable axis (Figure 7, 9.3), I erase the grid lines in those dimensions and only leave the grid lines for numerical variables in the purpose of precise reference. Also for the scatter plot (Figure 5), I change the range of both axis to the data range to avoid leaving much blank and make data points sparse on the plot.
  Graphic Integrity: All the data in my visualizations are compared and drawn in the same scale, so the lie factors of all graphs are one. I avoid using a logarithm scale on population data to make deviations from mean value different for small and great values, as log scale makes distance between great values narrower. I make necessary annotations for average values (Figure 5, 7), heatmap cells (Figure 9.2), and customized color encodings (Figure 9.3), through bold black font and dash grey lines. All my plots are in 2D, carrying at most three dimensions of data in total, with the extra dimension encoded with color or size.

- **Interaction** (Not decided yet in static version)
  The Altair-based graphs are easy to implement interactions on, so I will continue adding features to

my graphs to provide efficient and meaningful interactions.

4 Evaluation Method

To ensure the readers can fulfill my learning object described in the first section, a simple way is to ask a few domain questions about the information we want them to take away, for example:

- Which general regions of US have the highest/lowest natural amenity rating?
- What diseases have a high death rate for Red Lake County and are there any trends of decrease/increase in mortality with the time? Which diseases did Red Lake County effectively protect against compared to national average after 35 years?
- Where is the position of Red Lake County's poverty situation and household income if we put it in a national scale?
- What is good about Red Lake County and what is the problem we still need to solve in terms of public education?

Also a usability test should be done to collect feedback from different users, depending on their social roles, ages, and degree of mastery on reading electronic visualizations. The usability test should be conducted using the control variate method, which means we may only change one factor each time, for example, color encoding, sequence of graph, or annotation font size, to figure out their actual influence on completing the learning objectives.