SI 649: Communication Visualization Project
Interactive Blog Post

Mingliang Duanmu
duanmuml@umich.edu

November 12, 2021

1 Learning Objectives

The viewer will be able to:

- **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, and across different states in the US.

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 a combination of Altair and Streamlit for my interactive visualizations. Streamlit is an easy-to-use, lightweight tool to deploy blog sites using Python. The interaction widgets like select boxes, sliders, checkboxes and radio buttons can be well combined with parameters of Altair graphs to provide convenient interactions with readers. As a person used to programming-based tools, Altair has more freedom of customizing all the parameters of a graph, together with versatility to add interface interactions and organize multiple subplots. I give up using Tableau in this interactive due to high learning cost.

Visualizations

Line plot for unemployment rate

Since the author claims that Red Lake County's unemployment rate is well below national level, I decide to use a line plot to compare it with the national data. Unluckily the unemployment rate I find only applies to each state since county-level data are missing for many states.

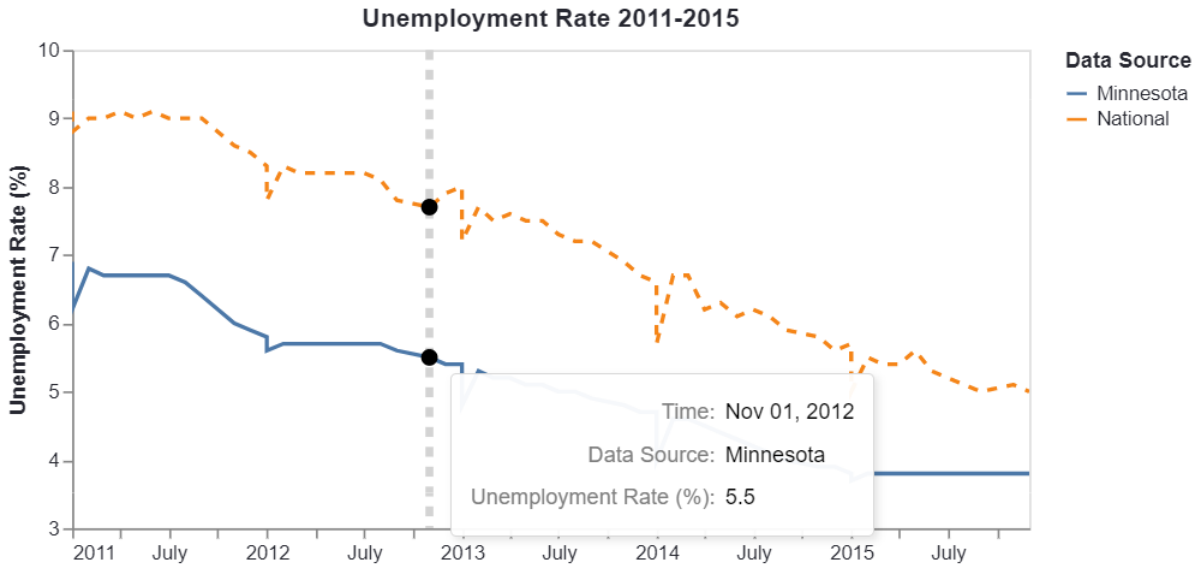


Figure 1: Line Plot for unemployment rate

Interactions:

- I add zooming to both x and y axis to allow users to view local values and trends easily.
- Similar as homework, I add a grey vertical line and two black dots when mouse is hovering on the data points, so that reader can easily compare the value of national and Red Lake's unemployment rate at the same time point.

What I dislike about this plot is that the information is not rich, which is due to the limitation of data from the source. The graph is already expressive so it does not require extra interaction to make it more efficient.

Dual scatter plots for average household income, and poverty rate

Since the author mentions that Red Lake County does not rank low on economic metrics like household income and poverty rate, I decide to use scatter plots to display such two metrics in both national scale and state scale.
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 dual plots in Figure 2.

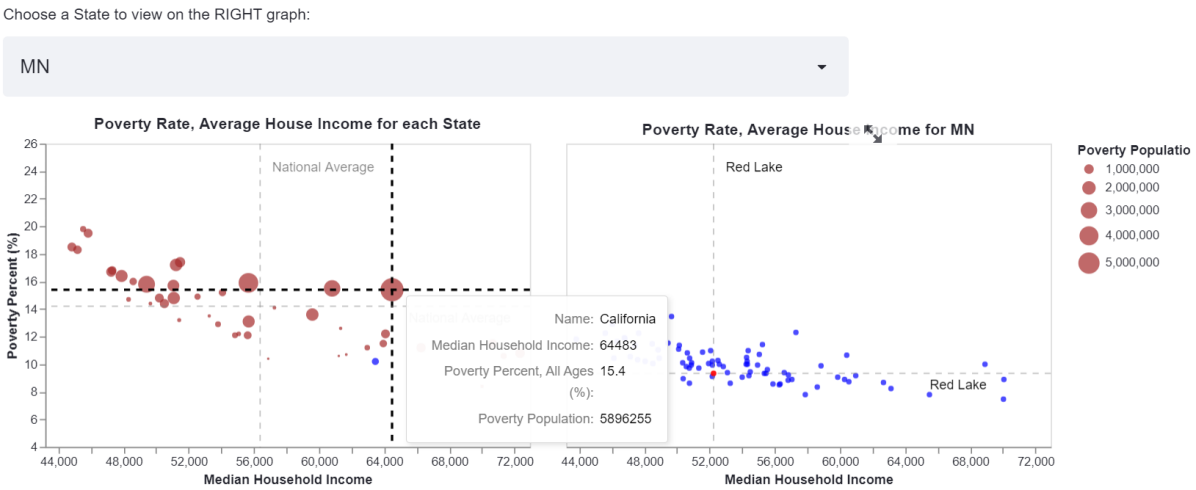


Figure 2: Dual Scatter Plot for poverty rate, poverty population, household income

The position of Minnesota is marked with blue color on the left plot. The grey 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.
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.
Interactions:

- I add zooming to x axis to allow users to view local values among a series of scatter points easily.
- When clicking on the points on the left plot, a vertical and a horizontal black dash line help readers read the exact values on both axises.
- The tooltip displays the state name and all three features on mouse hovering.
- Readers can use the selection menu to view data of different states in the right plot, at the same time comparing it to national data. The default of selection is MN.
- The point indicating Red Lake County is marked red and always displayed so readers can even know the position of Red Lake County in other states across the country.

Bar and scatter plot for 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.

Figure 3: Bar chart and dot chart for educational resources

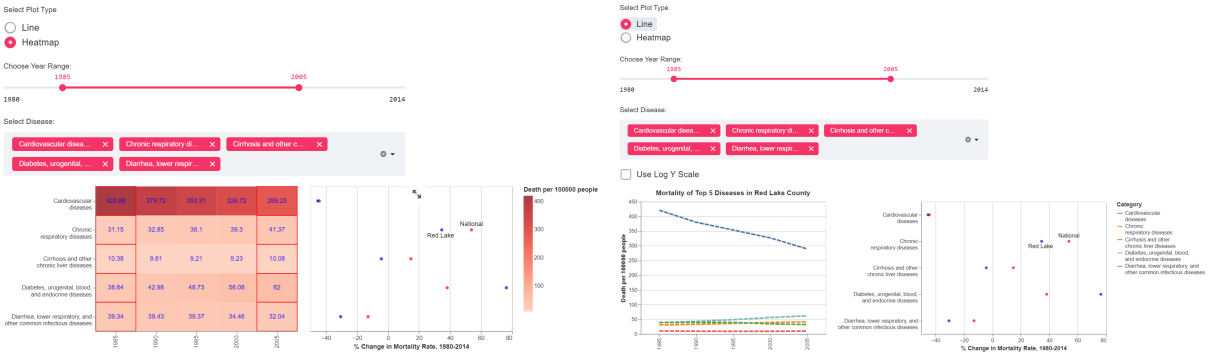
Interactions:

- Readers can use the selection menu to explore educational resources regarding the two aspects for different states across the country. The default of selection is MN.
- Using the checkboxes, readers are able to switch between top/bottom values for both metrics, as well as changing the display according to ascending/descending order.
- The sliders allow readers to customize the number of top/bottom values they want to display on both plots. A range between 10 and 20 is available for choice.
- The tooltip displays the exact numerical value on mouse hovering.
- The value of Red Lake County is marked using a horizontal dash line and always displayed so readers can even know the position of Red Lake County in other states across the country.

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, but I found there is no necessity and does not add to expressiveness.

Line/heatmap and scatter plot for 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. On the heatmap, the mortality rates of the most and least recent time are highlighted to inform readers the change during the period. 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.



(a) Heatmap

(b) Line

Figure 4: Heatmap/Line plot for top 5 mortal diseases in Red Lake

Interactions:

- Readers can choose either the heatmap or line plot to display the trend of diseases in the past 35 years using the radio button.
- For the line plot, readers can use the checkbox to change y values to log scale for comparisons between diseases.
- Readers can customize the range of timeline for both the heatmap and line plot using a double ended slider.
- Readers can choose different subsets of diseases to display on both plots using the multiple selection box. All five diseases are shown by default.

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 high educational resources. In contrast, red color represents negative data like mortality and bad enrollment rate. For example, lighter red represents less fatality for a disease (Figure 4). 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 2).
  Cognition:
  In the scatter plot, I create a pop-out effect using a contrast color to identify the point we want to observe (Figure 2). For multiple lines crossing each other in one chart (Figure 4), 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 1, 4) and categorical variable axis (Figure 4), 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 2), 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 2), heatmap cells (Figure 4), and customized color encodings (Figure 4), 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**
  Filter:
  I use filters on time, variables and items to display on the last two plots. A double ended slider on time interval allows user to choose data within any time period to display, and a multiple selection box makes readers focus only on the categories they are interested in.
  Explore:
  For the second and third plot, I offer a dropdown menu for readers to choose different states' data to display. Although the topic is about Red Lake County, the readers can explore the same metrics on different places across the nation, which gives them a broader picture of national level and makes a fairer context for Red Lake. Also, the slider and checkboxes on the number of top/bottom counties to display satisfies readers' curiosity in the counties doing good/bad on certain metric, as well as the large gap between top and bottom performances.
  Encoding:
  For the mortality plot, I provide two encoding options on the same data for readers to choose from, line plot and heatmap. Each encoding has its benefit in conveying information according to reader preference: line plot focuses on the trend of mortality of each disease along the timeline and provides parallel comparison between different diseases, while heatmap shows the severity of disease of any time point in a straightforward color encoding, and the text labels and highlights make the detail values clearer.

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:

- 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? Does Red Lake County defeat half of the counties in California in terms of poverty rate? Can you give an example of state with a lower average income and higher poverty rate than Red Lake County?
- What is good about Red Lake County and what is the problem we still need to solve in terms of public education? How big is the difference between top and bottom counties within a certain state in terms of education resources?

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.