CS/INFO 3300 Project 1 Report

By
Dana Luong
Joe Mo
SM Mashuque

1. Description of the Data

We acquired the dataset from Kaggle, a popular site that contains many datasets. We selected a dataset that contains U.S. Education data. Some of the columns of this dataset include data such as on state total and instructional expenditure, average test scores in each state, and enrollment numbers in each state during the years 1992 through 2017.
Dataset: https://www.kaggle.com/noriuk/us-education-datasets-unification-project

To summarize, the variables that we have used in our visualization include:

- State name
- Year
- Total Expenditure in a year
- Average Exam Score of 8th Graders in a year

We decided to visualize the variables for state expenditure and average exam scores for 8th graders in each state to investigate whether or not there was a correlation between expenditure and students' performance on exams. We encountered some difficulties in the data because the dataset only had complete records for expenditure and test scores for certain years; consequently we limited the data for our scatterplot to a subset of years. We formatted the data for each graph in a dictionary, and filtered the invalid records (records with missing expenditure or exam score data).

Additional files used to create the visualization include *us-map-data/us.json* for TopoJSON data for U.S. and *us-map-data/us-state-names.tsv* for U.S. state names and IDs that match the TopoJSON file. These two files were taken from the course website. Since the U.S. Education dataset does not have an ID field for each state, one challenge was linking the U.S. Education dataset with the TopoJSON data. We linked the Education dataset with the map by using the lowercase state names with underscores to find the state ID by using the state names file, then the IDs were then mapped to the TopoJSON file.

2. Overview of Design Rationale

We thought we could best capture the data from the dataset by drawing multiple graphs.

2.1.1 Heat-map of Education Budget of the United States in 2016

This graph shows the education expenditure of all the states in America for the year 2016. We took the total expenditure amount for each state and then graphed this data into a U.S. map.

The marks used in this graph are the state lines of each state. The channel in this graph is the color hue saturation; lighter color present lower spending and darker colors represent higher spending for a state. Another channel is the spatial regions of each state in the map.

I made a design choice to label the map with state abbreviations because it made the graph more readable. Also, when designing the color scale labels, I made a design choice to format the numbers with the d3 format function and that made the legend more human readable.

2.1.2 Heat-map of Education Spending Efficiency

This graph is similar to the last graph, but it is mapping efficiency numbers to the U.S. map. The efficiency numbers for each state are calculated this way:

$$\frac{Avg\ math\ score\ in\ 2017 - Avg\ math\ score\ in\ 1992}{Avg\ Total\ Expenditure\ in\ 2016 - Avg\ Total\ Expenditure\ in\ 1992}$$

This number gives us math scores increased per dollar spent by each state. Finally, efficiency numbers are scaled between zero and one before graphing them on the map.

One challenge I faced after scaling the data between zero and one is that a lot of states had very small values close to zero and few states had values close to one. I made a design choice to multiply all the efficiency numbers by 20, and that gave the states with the smaller values some color to the graph.

2.2 Line-Chart Comparison

This graph maps the Average Total Expenditure for all states with the Average Grade 8 Math Score for years ranging from 1992 - 2017. A line chart was used as an attempt to visualize the relationship, if any, of math exam scores with total yearly average expenditure of all 50 states.

The marks in the graph are the points representing yearly average expenditure and average math scores for grade 8 math, and the lines connecting them to make trends, if any, more apparent. The channels in this graph are the x and y positioning of each point due to variation in average expenditure and math scores, and the different colors to clarify both y axes.

When displaying the data for total average expenditure per year, the result was divided by 1000 so that the labels on the left y axis could fit without minimizing the width of the svg too much. Because the numbers were fairly large, the labels on the left y axis are calculated in thousands. Furthermore, when designing the scales, three linear scales were used for all three axes because the data values for each respective axis were relatively consistent. In designing the graph, two contrasting colors (orange and green) were utilized to clearly indicate that two different sets of data were being represented. To explain this with the visualization, a legend was created at the bottom indicating which color represents which particular axis, as well as indicates that the left y axis should be interpreted in thousands. Moreover, path elements that connected circles for each y axis was incorporated to make the visualization of trends easier to locate.

2.3 Bar Chart Comparison

Our third graph shows the education expenditure and average math and reading test scores for each state in 2015. A combination bar-line chart was used to investigate trends between state expenditure per student and average math and reading exam scores in each state.

The marks in this graph are the vertical bars indicating state expenditure per student as well as the points and lines indicating the average math and reading exam scores for grade 8 students. The channels in this graph are the heights and positions of the bars according to the left y-axis, the positions of the points and lines according to the right y-axis, and the colors used to distinguish reading from math scores.

The expenditure per student was calculated and graphed because of the large differences in population, and as a result expenditure, in each state. Bars were chosen

for the expenditure values and line graphs for average exam score values because this allows users to more effectively observe differences in exam scores relative to differences in expenditure. Bars were used in addition to line plots to allow for easier interpretation of the dual y axis (which can often be difficult to interpret if the data plotted is not clearly distinguished). A legend is provided at the bottom of the chart to inform users of the color channel used to distinguish math from reading scores.

3. Story

We visualize U.S. Education data in our graphs. Here are the stories we are trying to tell for each graph.

3.1.1 Heat-map of Education Budget of the United States in 2016

With this graph, we are trying to convey how much money each state spends on public education relative to other states by using a color scale. We can see a trend where states with bigger population tends to spend more money. In this graph, we see that California, New York, and Texas the biggest spenders on education.

3.1.2 Heat-map of Education Spending Efficiency

The goal of this graph was to locate the states that were the most money efficient during the 24 year period between 1992 and 2017. This graph shows how effective each state was relative to others in increasing the budget and seeing improved math scores by 8th graders.

In this graph, Michigan, West Virginia, Indiana, Wyoming did well in terms of seeing an increase in math score with respect to budget increase between 1992 and 2016. If you compare this graph with the previous graph, you see that these states spent much less money than others in 2016. This graph shows that the states with smaller budgets can still increase their education proficiency.

3.2 Line-Chart Comparison

The aim of this visualization was an attempt to locate any trends or relationships between the total expenditure of states with grade 8 math exams scores. Before creating the graph, we assumed that the visualization would exhibit a positive relationship, in that a greater total average expenditure for a year would translate into higher average

math scores for that specific year. Initially, the visualization conveys that higher expenditure correlates with higher scores for the first several years. However, this proves to be false in that the last data point represents a score that is lower than the previous years' despite having a higher total average expenditure. Ultimately, there was no dramatic change in the range of test scores despite the increases in total average expenditure, which was quite ironic as well as surprising to us.

3.3 Graph Three

The purpose of this graph is to investigate whether or not there is a trend between state expenditure per student and average exam scores. For some states, higher expenditure was correlated with higher exam scores, but there was no strong relationship between the two variables. Initially, the visualization was implemented using total expenditure on the left y-axis, but we received feedback that it would be good to factor in enrollment numbers for each state, as states that had higher populations naturally had higher total expenditure. In future investigations on whether or not there is a relationship between per student expenditure and exam scores, it may be useful to aggregate data from different exams, as not all students in each state may take this specific exam. Also, it was not completely clear what the total expenditure per state included. In future studies, it would be beneficial to limit data to expenditures directly affecting quality of education.

4. Outline of Team Contributions to the Project

We all did an equal amount of work during the research phase and during the implementation phase.

SM Mashuque:

- Created the sketch ideas for the heat-map graph and line-chart graph.
- Implemented the heat-map graph with U.S. map and dataset.
- Submitted milestone 2
- Total time spent: ~15 Hours

Joe Mo:

- Created sketch ideas that were similar to the heat-map graph implemented as well as the line-chart graph except in the form of a scatter-plot (no lines)
- Implemented the line-graph map "Average Total Expenditure vs. Average Grade 8 Exam Scores Between 1992 and 2017
- Submitted milestone 3
- Total time spent: ~15 Hours

Dana Luong:

- Created similar sketch ideas for US heat-map, scatterplot, and final sketch of combination bar-line chart
- Implemented combination bar-line chart
- Submitted milestone 1
- Total time spent: ~15 Hours