Investment in Education in the United States of America

Martim Norte
Instituto Superior Técnico
Lisbon, Portugal
martim.norte@tecnico.ulisboa.pt

Miguel Neves
Instituto Superior Técnico
Lisbon, Portugal
miguelscfneves@tecnico.ulisboa.pt

ABSTRACT
Education is one of the most important aspects of a persons' life, it has as strong role in shaping their future and their personality. The education of a population is essential for the development of nations and society. However, when governments try to improve their school system, investing their funds in the right places is key, even if resources may not be the most important aspect for granting students a better education. In this project we aim to analyze the impacts of previous investments in education in the United States of America while also trying to find some interesting patterns and curiosities. This project was developed for the Information Visualization course from Instituto Superior Técnico, in Lisbon, Portugal during the academic year of 2021/22.

INTRODUCTION
The United States of America (USA) is the wealthiest nation of the world [Davies et al. 2021] but the results from the 2018 Programme for International Student Assessment (PISA) [5] show us the nation falling behind several less-wealthy countries.

Even though the USA are represented in the assessment as one nation, the states have different realities: they all have different levels of wealth and may invest their funds in different ways. Therefore, we may find different levels of academic success among the states.

In this project we developed an interactive visualization with data from the U.S. Census Bureau and the National Center for Education Statistics (NCES) that allows users to explore the relation between state-level investment in education and students' academic success and enrolment rates, allowing for a comparison between states and a deep analysis of the investment evolution throughout the years 1992 to 2015. The visualization also allows for a more fine-grained analysis of academic success by student ethnicity, in an effort to better understand if the social and economic differences between the different ethnic groups have any impact in their academic success.

We also present our own analysis, which focuses mainly on answering the following set of questions:

1. How did changes in investment in education impact grades and student enrolment?
2. Is state revenue more impactful than state investment in education for academic success?
3. Which type of investment (instruction, support services, capital outlay) has the greatest impact in enrolment and academic success?
4. Is there a nationwide trend regarding student ethnicity and their academic success?
5. Is there any correlation between geographical position (North/South, East/West, interior/littoral, etc.) and investment in education?

RELATED WORK
When it comes to education, there are several articles addressing its importance and its impact in society and economy. [Woessmann 2006] found that the school's resources are not as important as its environment for academic success. [Hanushek 1996] had already come to similar conclusions and suggested that a "serious restructuring of schools" in the USA was necessary to improve student performance.

THE DATA
The first dataset we found was the U.S. Education Dataset: Unification Project [2] which aggregated data from the U.S Census Bureau and the National Center for Education Statistics. This dataset contained information aggregated by state and year from 1992 to 2016 about revenue, expenditure on instruction, support services, capital outlay and NAEP grades for mathematics and reading for fourth and eighth grade students.

We also used data from the National Governors association that was aggregated by [Kaplan, J. 2021] which included information regarding the governing parties in each state throughout the history of the USA. We ended up not using this data on our final implementation as we found that it didn't really offer any insights of particular interest, allowing for a cleaner and simpler dashboard.

Finally, to obtain values per capita we used the dataset from the U.S. Census Bureau [6] that contains every state's population for every decade.

Data Processing
We merged the data from the three datasets using Python and the Pandas library, resulting in the final dataset used in the visualization.

For the U.S Census dataset we had to fill the missing population values - the censuses only happen every 10 years - by interpolating between the existing ones.

While going through the U.S. Governors dataset we noticed some values were missing, more specifically, for the years

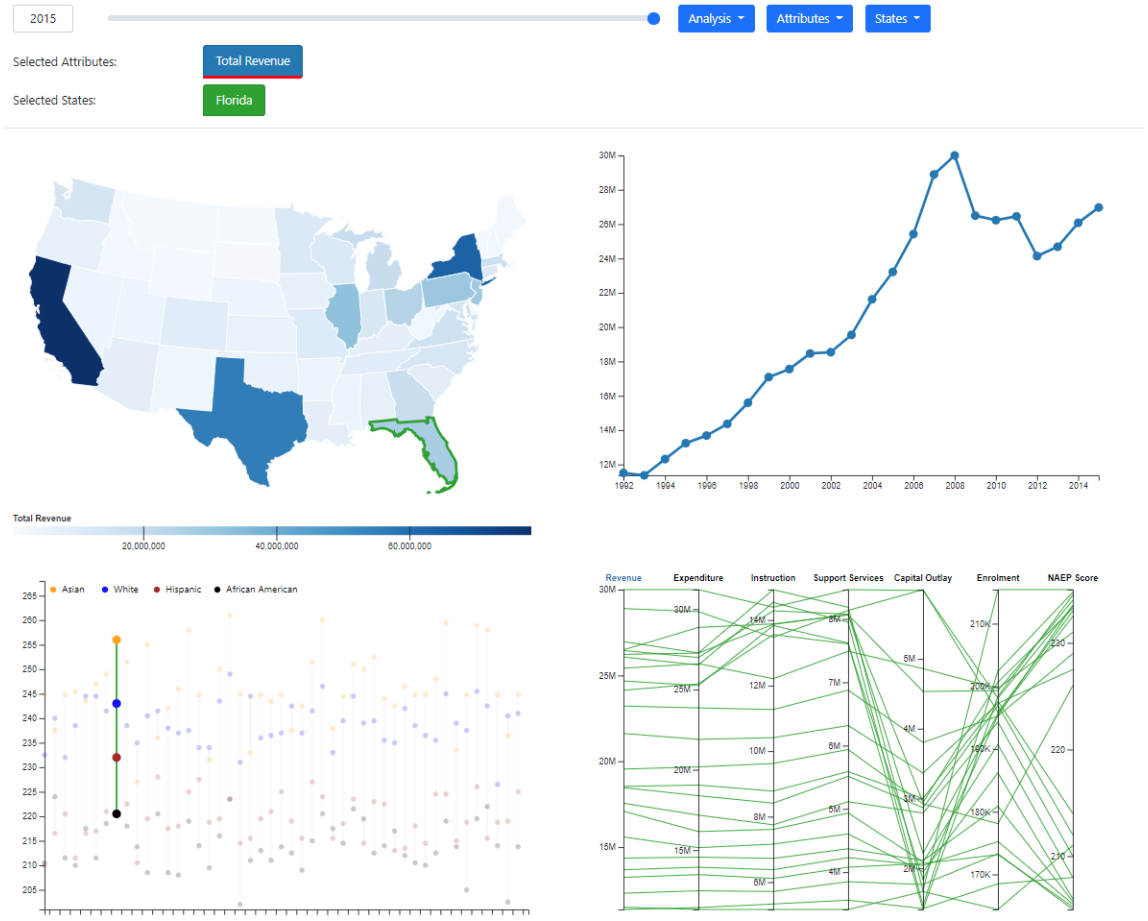

Figure 1. Layout of the visualization, with the default configuration.

between the last election and 2020, so we filled those with the last known value for each state.

The education dataset was much harder to clean. Most of the non-mainland states (Alaska; most of the isles) contained little to no data, or the data was severely outdated. Years before 1992 and after 2015 also suffered from this problem, containing no information regarding NAEP scores. Values for NAEP scores by ethnicity were only available after 2011, and for some states even later. Information regarding enrolment by ethnicity was very sparse, and for some states was simply not available. Due to these problems, we filtered data ranging from years 1992 to 2015, and filtered out every non-mainland state. For NAEP scores by ethnicity we applied the same method for filling missing values used for the average NAEP scores (described below). Information regarding enrolment by ethnicity was dropped, since it was too incomplete to get any type of meaningful analysis, even when interpolating. We also had to drop NAEP scores for two ethnic groups – Pacific Islanders and Native Americans – and for students with more than one ethnicity, as there was barely any data.

The education dataset contained NAEP scores separated by gender for a subset of the years. For this subset of years, only either male or female students took the exams. As we did not intend to take gender into account when visualizing the data, we merged this data into a single column.

The education dataset also contained NAEP scores for first, third, and tenth grades, but we decided to only keep values for the fourth and eighth grade as they contained way more data and were representative of the population. The NAEP scores' values from fourth and eighth grades were then aggregated by computing the average, to reduce the complexity of the final dataset.

After aggregating the values of the NAEP scores we noticed there were a lot of missing values - the exams were only taken every other year - which would hinder the visualization. For the states that had no NAEP scores for the base year, we calculated the average across every state, for the year 1992, and assigned it to the missing values. We then interpolated the values for the years with missing data, on a state basis.

Finally, we calculated for every attribute the percentual change from the previous year and the values per capita of every attribute but NAEP scores (as it does not make sense to calculate scores per capita) using the gross values and the population values for the corresponding year and state.

Figure 2. Layout of the visualization with Florida highlighted.

The final dataset contains information regarding year, state, revenue, the various investment types, average NAEP scores, NAEP scores per ethnic group, and enrolment. These values are presented in three different ways: gross, per capita, and percentual change regarding the previous year.

VISUALIZATION

Layout

Our visualization (Figure 1.) consists of a dashboard with two sections – a top section with a slider, three drop-down menus and two lists and a bottom section with four different idioms.

The bottom section is divided in two subsections, the left one shows data aggregated by year and features a choropleth map and a dot plot whereas the right section is aggregated by state and features a line chart and a parallel coordinates plot.

Selection Menu

The year slider allows the user to select the desired year for the idioms on the left side, by dragging the blue circle. The slider also features an input box that lets the user write the desired year, and reflects the year selected.

The selection menu allows the user to select/unselect what analysis, attributes and states appear on the visualization idioms. The user can select a maximum of 5 attributes/states and must have at least 1 attribute/state selected. Surpassing this limit will display a warning to the user (Figure 4.) on the top right of the screen. Changing analysis does not reset the attribute and state selection. The States list is sorted to facilitate state selection.

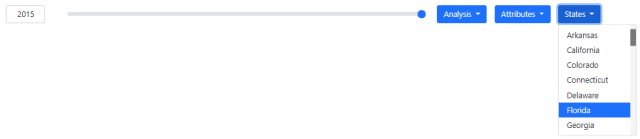

Figure 3. Selection menu with Florida selected.

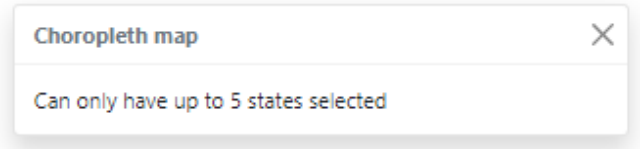

Figure 4. Warning displayed to the user after trying to select more than 5 states.

Attribute and State lists

The attribute and state lists reflect which attributes/states are present in the visualization. Each attribute/state is color encoded, and this color is used on every idiom to encode what attribute/state is being referred to. The user can select which attribute is displayed on the choropleth map by clicking on one of the attributes, that then gets underlined with red. Clicking on one the selected states removes it from the visualization.

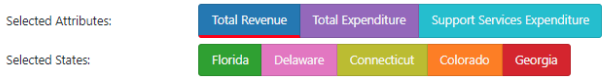

Figure 5 Attribute and State lists with Total Revenue, Total Expenditure, Support Services Expenditure and Florida, Delaware, Connecticut, Colorado, Georgia selected.

Choropleth Map

On the top left side of the bottom section, we have a choropleth map of the USA. In this map, each state's shade of blue represents the state's value for the attribute and year selected, according to the scale below it. When the user hovers a state, a tooltip shows up with its name and the corresponding value. This map also supports state selection: by clicking on a state, the user can select or unselect it, and the selected ones become highlighted with a colored outline. To facilitate this selection, the states get outlined with black when hovered, to better differentiate between them. The effects of this selection on the visualization are addressed in the other idioms' subsections. Finally, after hovering a state for 2 seconds, it gets highlighted on every idiom, by reducing the opacity of the other states (Figure 2.).

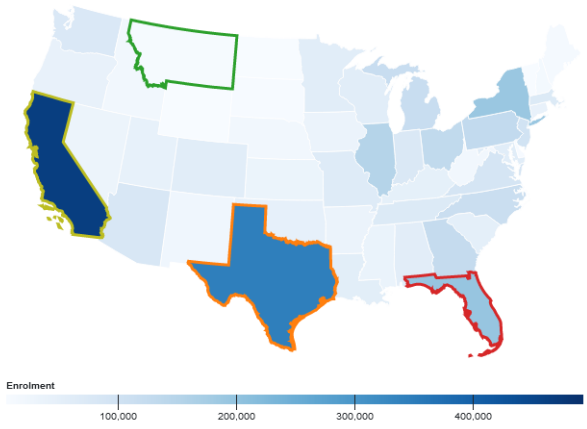

Figure 6. Choropleth Map representing gross values for Enrolment in 2007 with four states selected.

Dot Plot

The dot plot is located on the bottom left of the bottom section and, like the choropleth map above it, displays data from the selected year for every state. But, in this case, it shows the average results from students of different ethnicities, with its values represented by the dot's position on the y-axis. Each line corresponds to a different state and each dot to a different ethnicity. The different ethnicities are color encoded, and the legend is at the top right of the plot. The lines connecting the dots are colored with the color corresponding to the state it refers to (as in the state list). By hovering the mouse over a dot, a tooltip shows up, displaying the name of the state and the item's value. Whenever the user clicks on a dot, the corresponding state becomes selected or unselected for all the idioms (with a maximum limit of 5 states and minimum of 1). Finally, after hovering a dot corresponding to a specific state for 2 seconds, the state gets highlighted on every idiom, by reducing the opacity of the other states (Figure 2.)

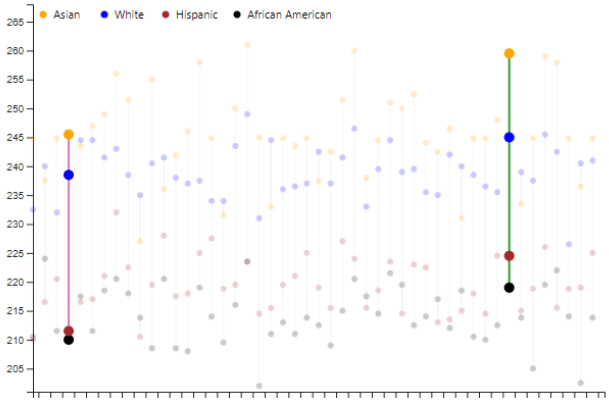

Figure 7. Dot plot representing the average NAEP results for different ethnicities in the year 2015 with two states selected.

Line Chart

The line chart on the top right of the bottom section represents the evolution through time of the selected attributes for the selected states. If only one state is selected, the chart will display the lines for all the selected attributes, with their color being defined by the attribute. With this mode, the user can select the attribute displayed on the choropleth map by clicking on one of the lines/dots. If multiple states are selected, the lines will correspond to the attribute selected on the map for the selected states, and their color will encode the corresponding state. This is easy to tell apart because the colors are consistent throughout the visualization i.e., the outer line on the choropleth map has the same color as the line in the line chart for that same state. If the user wants to highlight one specific line, they just need to hover the corresponding line for a couple of seconds (Figure 2.).

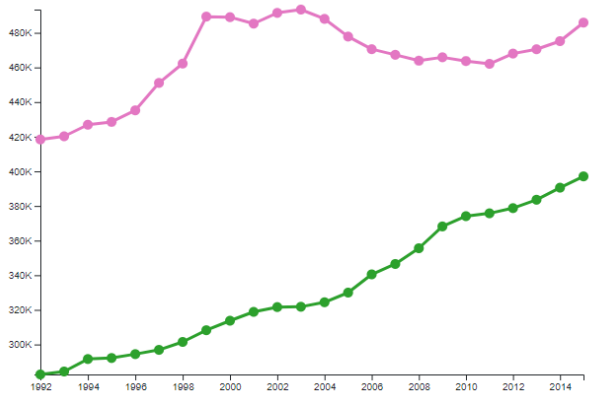

Figure 8. Line Chart showing the evolution of students enrolled from 1992 to 2016 in California (top, pink) and in Texas (bottom, green).

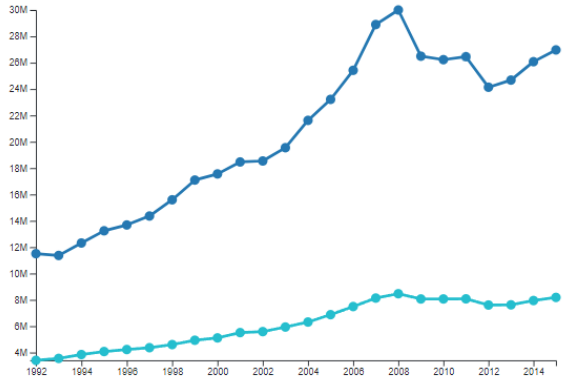

Figure 9. Line Chart showing the evolution of investment (top, dark blue) and expenditure (bottom, light blue) from 1992 to 2016 in Florida.

Parallel Coordinates

On the bottom right of the bottom section we find the parallel coordinates plot with a vertical axis for every attribute. Like the idioms mentioned before, the lines in this plot are also color coded according to the state they represent, and each one of them corresponds to a different year. To highlight a specific state, the user can hover any line for 2 seconds (Figure 2.). To facilitate the interpretation of the lines, we also allow the reordering of the attributes by dragging and dropping their axis into the desired position. The user can also click on the attributes to select/unselect them from the visualization. The font color of the axis reflects the coloring of the attribute (as in the attributes list).

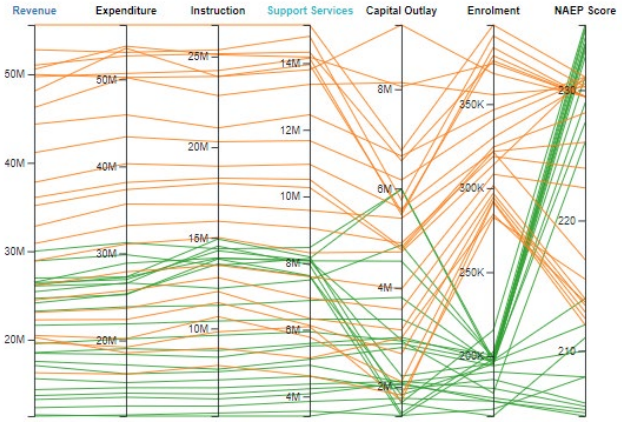

Figure 10. Parallel coordinates with Florida and Texas selected (green and orange lines, respectively). The attributes selected for the other idioms are colored with the respective color.

Rationale

When first addressing the ways we could visualize the data, we decided to offer two perspectives on the data, one based on year, and one based on state. With this idea in mind, we decided to split the visualization into two logical components, one for comparing states and the other for comparing years.

The easiest and most obvious idiom to pick was the choropleth map, as it can encode data for every state in the simplest way – with the geographical position - fits the problem well and would allow us to tell if there was any correlation between the values and the geographical position (North/South, East/West, interior/littoral, etc.).

When picking the rest of the idioms, we took into consideration the questions we wanted to answer with them. We first considered using a scatter plot to figure out whether there were any correlations between attributes but, because we wanted to compare multiple attributes at the same time, we opted by using a parallel coordinates plot instead, as it is more suitable for the job.

We also needed to find a way to visualize the student enrolment and performance data for different ethnicities at the same time. We wanted to have a nation-wide perspective of that and because the year didn't feel very relevant for this issue, we found the dot plot to be a good way to do it. We can represent every state at the same time in the horizontal axis and the values on the vertical axis, with each dot corresponding to a different ethnic group encoded by the dots' color.

Finally, we figured the line chart would be the best way to visualize the evolution of each attribute through time and would allow us to make comparisons both between attributes and between states.

Because of the size of the dataset, it is not possible to select multiple years at the same time, as it would require

computing aggregations on the fly, making the visualization extremely slow and unresponsive. Computing them *a priori* would make the dataset too large, since there is data from 23 years and an enormous number of combinations of year selections. With 50 states, having them all selected by default and allowing the user to filter the ones they are interested in would not be possible. Instead of the usual filter pattern, we went with the selection pattern – the visualization has a reduced default configuration, and the user can extend it to show information regarding states of interest. We limit the number of states/attributes to 5 since it is the maximum number of different colors an idiom can have, while keeping the colors contrast high enough to easily distinguish between them.

On the prototyping phase we planned on using another idiom – a Gantt chart – to encode what political party was governing the state at a point in time. After starting to implement the visualization, we found out that this chart did not improve on the design and only added additional complexity and clutter to the visualization. The conclusions taken from it were not relevant enough to keep it in the final version. Because of this, we replaced the question regarding political parties with a geography based one.

Initially, we did not have a way for the user to select the analysis and the states on the selection menu, and instead only allowed state selection on the choropleth map and every attribute was included on the attribute menu. We found out that requiring the user to know the geographical position of the states was bad design and added the state selection menu. We also found out that having over 50 attributes on one menu made it very hard to select a particular one, so we instead split them into different analysis categories and added an analysis selection menu, effectively reducing the size of the attribute list to one third.

Our first prototype did not allow for selecting multiple states. Because it is very important to be able to compare states across time, we improved the idioms to support this functionality.

Another problem with the first prototype was that it had a non-intuitive way of selecting what attribute is displayed on the map and on the line chart when multiple states are selected. Originally, this attribute would be the first attribute on the list, and the user would select it by reordering the list by dragging and dropping. Because of the lack of clarity with this implementation, we updated it, so the user instead simply clicks on the attribute, and it gets underlined with red.

While developing the dashboard we found out that displaying the states in a list improved the clarity of the visualization, so we added it to final version. This was not present in the original prototype.

Originally, the parallel coordinates plot did not allow for the reordering of the axis. Because finding correlations between different attributes is very important for the analysis presented in this work, we implemented the functionality to reorder the axis by dragging and dropping them into the desired position.

The warning messages were added to improve on the responsiveness and clarity of the visualization.

Finally, the state highlighting mechanism (Figure 2.) was implemented to simplify the visualization of only one state across the idioms when multiple states were selected.

Potential

With the developed dashboard we can easily answer the proposed questions. In this section we provide the means to answer two of them:

1. Is there a nationwide trend regarding student ethnicity and their academic success?

When it comes to analyzing trends regarding student ethnicity, if we select ‘Gross’ from the analysis menu and any year from 2011 onwards, we can then look at the dot plot and tell if some dots’ colors show up in higher positions more frequently than others. And the answer for the 5th question we proposed comes up clearly: we can conclude that students from Asian and White ethnicity, on average, have higher grades than those from Hispanic and African American ethnicity. We can also tell that between White and Asian, it is usually Asian students who have the best scores and between Hispanic and African American, it is usually the latter with the lowest grades.

If we change the type of analysis to ‘Percentage change’, the perspective we get is different and we can visualize whether some ethnicities seem to be improving more than others. The answer is not clear but there appears to be a bigger improvement in students with lower grades.

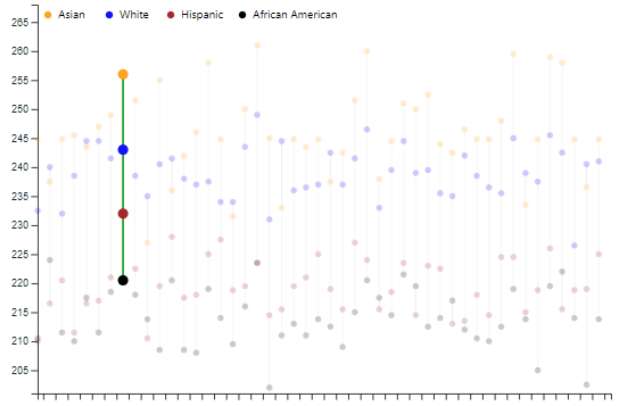

Figure 11. By looking at the plot, we can clearly see that the yellow dots are, on average higher than the rest. We can also see that the black dots are lower than the rest, on average.

2. Is there any correlation between geographical position (North/South, East/West, interior/littoral, etc.) and investment in education?

To figure out whether there is a correlation between geographical position and investment in education we have one of the best idioms for the purpose, the choropleth map. We first need to select 'Total Expenditure' from the attributes menu and make sure it's the one highlighted by clicking on it on the list of selected attributes and then change the analysis to 'Per Capita' because the population of a state has a strong influence on how much they need to invest, this way we will have a perception of the investment per habitant. Then, if we look at the map while dragging the slider to go through the years, we can have a sense of the investment's evolution through time, and it was clear that there was a drastic and nationwide improvement in that 23-year span, which can be confirmed, at least for some randomly selected states, by the line chart in Figure 12.

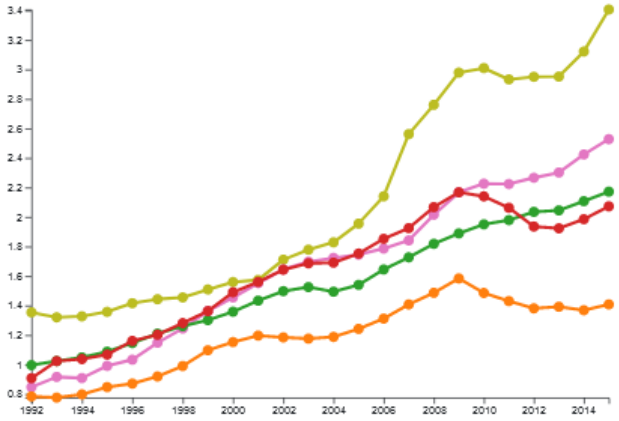

Figure 12. Line Chart showing the evolution of the total expenditure in education from 1992 to 2016 in Wyoming (yellowish green), Illinois (pink), Iowa (green), Texas (red) and North Carolina (orange).

Furthermore, and to finally answer the question, we notice slightly, yet significatively, darker color in states in the central area (between Wyoming and Illinois, both selected in Figure 13) and in the Northeast but especially Wyoming, New York, and New Jersey.

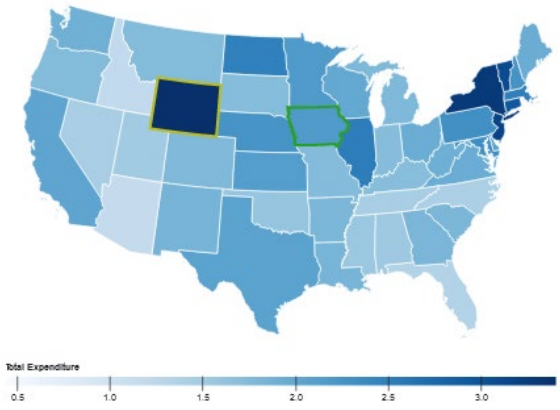

Figure 13. Choropleth map representing values per capita for Total Expenditure in 2015

We also found out when analyzing the data that the enrolment values per capita kept lowering through the years even though for most states the gross values increased. At first, we found that odd, but a possible justification would be that the adult population is increasing more rapidly than the younger one, possibly due to people living longer or due to immigration factors.

IMPLEMENTATION DETAILS

Every idiom, apart from the choropleth map, was developed from the ground up, as our use case was too unique for a general case solution to be good enough. This is due to the high levels of interactivity and customization of every idiom.

The colors used in the visualization are the 10 categorical colors of *d3.js* color scale collection. We chose the colors that had a higher contrast with blue for the states, as they would be used to outline the states in the choropleth map. The rest of the colors were used for the attributes.

Our implementation contains key global structures that store the application sate, of special importance:

1. *selectedYear*: an int with the selected year.
2. *selectedAnalysis*: a string with the selected analysis.
3. *selectedAttributes*: an array containing the selected attributes.
4. *selectedStates*: an array containing the selected states.
5. *mapAttribute*: a string with the attribute selected for the map.
6. *attributeColors*: an object containing the mapping between attribute and color.
7. *stateColors*: an object containing the mapping between attribute and color.
8. *yearData*: data filtered by the selected year. This exists for performance reasons.
9. *stateData*: an object containing the data for the selected states. Again, this exists for performance reasons.

With these structures, the idioms can get updated independently and transparently from one another. When a year selection occurs, *selectedYear* gets updated. Same goes for attribute selection and state selection, and their corresponding data structures. Changing analysis is more intricate – when changing analysis, we must swap every attribute to the correct one (which, even though harder, is far better than simply resetting the attribute selection), which encompasses changing the data binding in the attribute menu and attribute list and updating the internal structures. When a state change occurs, an *update* function gets called on every idiom, to update them based on the data in the global structures. Examples of updates would be changing the colored outline on the choropleth map, adding a line to the line chart, or more complex updates such as

changing the year the data is from or selecting/unselecting a state.

Reordering of the axis on the parallel coordinates works by listening to *dragStart*, *drag*, *dragEnd* events. When a *dragStart* event gets fired we store the x position of the axis being dragged. On *drag*, we sort an internal array containing the attributes by x position, update the x scale, and redraw the axis being dragged, and the lines connected to it. On *dragEnd* we set the position of the axis being dragged to its correct position, according to the x scale.

The state highlight functionality was implemented by developing two functions on every idiom, *highlightState* and *resetStateHighlight*, that implement this functionality for that specific idiom. *highlightState* essentially changes the opacity of the elements corresponding to the states not highlighted to 0.2 while keeping the highlighted state opacity at 1. *resetStateHighlight* changes the opacity of all elements back to 1.

CONCLUSION AND FUTURE WORK

In this project we created an interactive dashboard to visualize and analyze data related to education in the USA with focus on investment and results. The main takeaway from this project is the experience gained with data preprocessing, the increase in knowledge about different visualization idioms and their respective strengths, and the ability to implement interactivity in data visualization.

Even though we can answer every question with the aid of the dashboard, finding the answers ended up being tougher than expected, perhaps due to some questions being unnecessarily hard. If we were to start anew, we would have tried to focus on more specific aspects of the theme to provide a more in depth and worthy analysis. We would've also massively simplified the dataset, by reducing the dimensionality further.

If we had more time, we would have implemented brushing on the parallel coordinates, that would change the range of the data displayed on the parallel coordinates and the line chart. We would also have implemented a way to select a subset of years on the year slider, that would affect the line chart and the parallel coordinates.

REFERENCES

1. Davies, J., R. Lluberas and A. Shorrocks, Credit Suisse Global Wealth Databook 2021, Credit Suisse Research Institute, Zurich. https://www.credit-suisse.com/media/assets/corporate/docs/about-us/research/publications/global-wealth-databook-2021.pdf
2. Garrard, R. (2018). *U.S. Education Datasets: Unification Project* [Dataset]. https://www.kaggle.com/noriuk/us-education-datasets-unification-project
3. Hanushek, E. A. (1996). Measuring investment in education. *Journal of economic perspectives*, *10*(4), 9-30.
4. Kaplan, Jacob. United States Governors 1775-2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-01-16. https://doi.org/10.3886/E102000V3
5. OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en.
6. US Census Bureau. (2021, October 8). *Historical Population Change Data (1910–2020)* [Dataset]. https://www.census.gov/data/tables/time-series/dec/popchange-data-text.html
7. Woessmann, L. (2006). *Why Students in Some Countries Do Better*. Education Next. Retrieved November 14, 2021, from https://www.educationnext.org/whystudentsinsomecountriesdobetter/