Information Visualization
Checkpoint II

Checkpoint II: Data Cleaning & Processing

Group: G14
Date: 2021/10/11

Initial Dataset

For this delivery we used three different datasets - the U.S Education Dataset from the Unification Project, the National Governors Association dataset, and the U.S. Census Bureau population dataset. The first dataset contains information aggregated by state and year about revenue, expenditure and NAEP grades for mathematics and reading for both fourth grade and eight grade students separated by student ethnicity. The second dataset includes information regarding the governing parties in each state throughout history. The third dataset contains every state's population for each decade.

Selected/Derived Data

From the investment and academic success dataset we discarded the key column since it encoded redundant information, all information regarding student gender and information regarding academic success from all academic levels but the fourth and eighth grade (only students from fourth and eighth grade take the NAEP reading and mathematics exam every two years). Finally, we selected data from the mainland states from 1992 to 2016 because they had the greatest amount of information.

For the US Governors dataset every column but the state, party and year was dropped. We then selected the data from the mainland states from 1992 to 2016.

For the population dataset we also dropped every column but the type of region (State, Nation, County, etc.), the year and the resident population. We then selected data regarding only the mainland states from 1992 to 2016.

Since different states will have different populations, the values for investment and revenue would be more appropriately visualized if they were *per capita*. To obtain that, we divided each of those columns by the population column and stored them in new columns.

Data Abstraction

The dataset used in this project contains time data organized by state. The data is stored in a table keyed by year and state, in csv format.

Attribute(s)	Description	Semantics
STATE	Nominal; Key; String;	The US state the record is from.
YEAR	Ratio; Key; Integer; 1992-2016;	The year the record is from.
RULING_PARTY	Nominal; String; Republican/Democrat;	The ruling political party.
POPULATION	Ratio; Integer;	The population numbers.

{T}_REVENUE(_PC)	Ratio; Integer;	Revenue in dollars. T is one of TOTAL, FEDERAL, STATE or LOCAL. PC stands for *per capita*.
{T}_EXPENDITURE(_PC)	Ratio; Integer;	Expenditure in dollars. T is one of TOTAL, INSTRUCTION, SUPPORT_SERVICES, CAPITAL_OUTLAY. PC stands for *per capita*.
G0{N}_{E}_A	Ratio; Integer;	Number of students enrolled. N (school year) is either 4 or 8. E (ethnicity) is one of AM (American Indian or Alaska Native), AS (Asian), HI (Hispanic/Latino), BL (Black or African American), WH (White), HP (Hawaiian Native/Pacific Islander), and TR (Two or More Races).
G0{N}_{E}_A_{S}	Ratio; Integer; 0-500;	Average grade on the NAEP exam. N and E are the same as students enrolled. S (subject) is either MATHEMATICS or READING.

Data Processing

The data processing was done using python with the pandas library.

For the investment and academic success dataset we first formatted the state values to the standard title format so we could later merge the various dataset on state and year. The next step was to fill the missing values with 0 (the original authors of the dataset left every 0 value with a NaN value). The last step was to aggregate the data separated by student genre regarding enrollment and grades in a single corresponding column.

While going through the U.S. Governors dataset we noticed some values were missing, more specifically, for the years between the last election and 2020, so we filled those with the last known value for each state.

The U.S. Census dataset only had one issue: because these censuses only happen every 10 years, we had to fill the missing values by interpolating between the existing ones.

We then removed from these last two datasets the states that were discarded from the main dataset due to lack of data. The last step was to merge the three datasets on state and year to produce the final dataset that is going to be used for the visualization. The original dataset had 1715 rows and 266 columns and after all the processing we were left with 951 rows and 69 columns.

Mapping (Data sample/Questions)

```
(from "data.csv")
STATE; YEAR; TOTAL_REVENUE; FEDERAL_REVENUE; STATE_REVENUE; LOCAL_REVENUE;
TOTAL_EXPENDITURE; INSTRUCTION_EXPENDITURE; SUPPORT_SERVICES_EXPENDITURE;
CAPITAL_OUTLAY_EXPENDITURE; ...; RULING_PARTY; POPULATION; TOTAL_REVENUE_PC;
FEDERAL_REVENUE_PC; STATE_REVENUE_PC; LOCAL_REVENUE_PC; TOTAL_EXPENDITURE_PC;
INSTRUCTION_EXPENDITURE_PC; SUPPORT_SERVICES_EXPENDITURE_PC;
CAPITAL_OUTLAY_EXPENDITURE_PC
Alabama; 1992; 2678885.0; 304177.0; 1659028.0; 715680.0; 2653798.0; 1481703.0;
735036.0; 174053.0; ...; Republican; 4121890; 0.649917; 0.073796; 0.402492;
0.173629; 0.643830; 0.359472; 0.178325; 0.042227
```