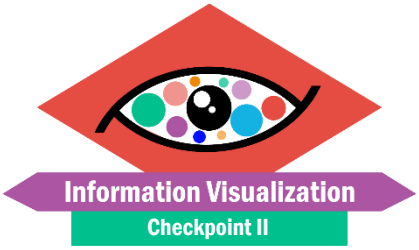

Checkpoint II: Data Cleaning & Processing

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.

For all three datasets used, each row/item contains data about an US state in a particular year. In the first dataset, each column contains information about its revenue, expenditure and NAEP grades for mathematics and reading for both fourth grade and eight grade students, separated by ethnicity. In the second dataset, each item has information regarding the party that governed that state in that year. In the third one, each item features that state's population in that year.

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.

The Attribute **STATE** is Nominal; Key; String and represents the US state the item is from. The attribute **YEAR** is Continuous; Diverging; Key; Integer; ranges from 1992-2016 and represents the year the item is from. The Attribute **RULING_PARTY** is Nominal; String; Democratic or Republican and represents the the ruling political party. The attribute **POPULATION** is Ratio; Sequential; Integer and represents the population numbers. The attribute **{T}_REVENUE(_PC)** is Ratio; Sequential; Integer and represents the revenue in dollars. T is one of TOTAL, FEDERAL, STATE or LOCAL and PC stands for per capita. The attribute **{T}_EXPENDITURE(_PC)** is Ratio; Sequential; Integer and represents the expenditure in dollars. T is one of TOTAL, INSTRUCTION, SUPPORT_SERVICES or CAPITAL_OUTLAY and PC stands for per capita. The attribute **G0{N}_{E}_A** is Ratio; Sequential; Integer and represents the 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) or TR (Two or More Races). The attribute **G0{N}_{E}_A_{S}** is Ratio; Sequential; Integer; ranges from 0-500 and represents the average grade on the NAEP exam. N and E mean the same as in **G0{N}_{E}_A** and 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)

1. How did changes in investment in education impact grades and student enrolment?

```
(from "data.csv")
STATE; YEAR; TOTAL_EXPENDITURE_PC; G0{N}_{E}_A; G0{N}_{E}_A_{S}
ALABAMA; 2015; 1.530; 55808; 217
```

2. Is state revenue more impactful for academic success than state investment in education?

```
(from "data.csv")
STATE; YEAR; TOTAL_REVENUE_PC; TOTAL_EXPENDITURE_PC;  G0{N}_{E}_A_{S}
ALABAMA; 2015; 1.501; 1.530; 217
```

3. Which type of investment (instruction, support services, capital outlay) has the greatest impact in enrolment and academic success?

```
(from "data.csv")
STATE;  YEAR;  INSTRUCTION_EXPENDITURE_PC;  SUPPORT_SERVICES_EXPENDITURE_PC;
CAPITAL_OUTLAY_EXPENDITURE_PC; G0{N}_{E}_A ; G0{N}_{E}_A_{S}
ALABAMA; 2015; 0.778; 0.492; 0.104; 7360222; 7501799; 217
```

4. Is there any correlation between the party that governs a state and its investment in education?

```
(from "data.csv")
STATE; YEAR; RULING_PARTY; TOTAL_EXPENDITURE_PC
ALABAMA; 2015; Republican; 1.530
```

5. Is there any trend regarding student ethnicity and their academic success?

```
(from "data.csv")
STATE;   YEAR;   G08_WH_A_READING;   G08_BL_A_READING;   G08_HI_A_READING;
G08_AS_A_READING
ALABAMA; 2015; 267; 243; 252; 0
```