Phase4: Decisions and Cleaning Steps

Date: November 23, 2020
Members: Qinyu Wu, Richard Zhou
DataGuideline: Glossary

Decisions:

- There are many SP values in our dataset. According to the dataset guideline, SP aims at protecting privacy of schools which have less than 50 enrollment. This means that as long as a school has less than 50 enrollment, the values on their attribute, except for the basic information, are SP. Since SP data only shows a small group of students and our dataset has been big enough, we decide to remove all of them.

- We treat all the N/D, N/R values as NULL and delete all of them except for the secondary achievement. This is because the schools ,which lose data in one attribute, always lose data in all other attributes. So it doesn't make sense to leave a blank row with all NULLs. However, in secondary Achievement, many schools don't have data only in applied Math but have all other data. So it is better to leave the applied Math null and keep other data.

- We split achievement table into elementaryAchievement and secondaryAchievement to aviod NULLs (It doesn't make sense for an elementary school to have ossaltgrades)

- only one school's school number is not int, so we delete the entire row of that school to reinforce the type of shool number follows the int

- Some schools miss the phone number which is the only way to contact them in our schema. So we don't reinforce a not null constraint on it in case that these schools want to update later

- ~~One difficulty we encounter is that "69%" 's type is text rather than int in our dataset and it is extremely difficult to convert it into int. So we keep it as text in our schema but we may change that later in order to do the comparison in queries~~
  we convert the text "69%" into float

- we add two domains: schooltype and schoollevel in order to reinforce the school types and levels we want

Cleaning Steps:

- **Step1:** Using Pandas library in Python to split the spreadsheet and saving each part into csv file

- **Step2:** Using Pandas to delete all "N/R, N/D, NULL, SP"

- **Step3:** Deleting extremely strange data manually from Excel. (like school number mentioned above)

- **Step4:** Using postgresql to reinforce the constraint (remove constraints first, and then add them back), and then convert relations into csv files.