

A group of people in a room

Description automatically generated

1. **Background & Method**

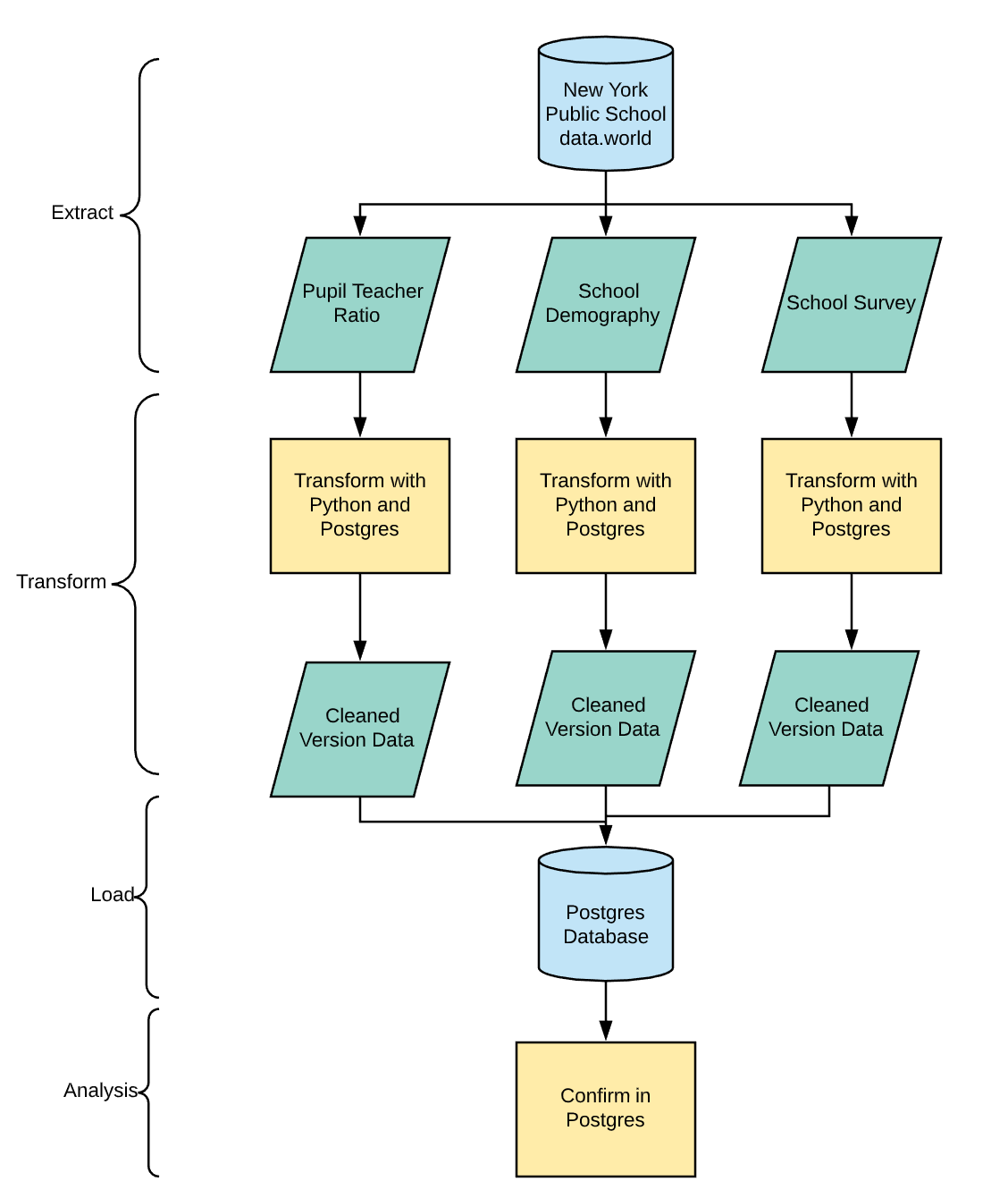
**Background on this project:** This project requires us to perform Extract, Transform, and Load process from a data source to a new database.

**Background why we chose the dataset:** We are interested to see a dataset which has correlation among them. We chose New York public school dataset because it has the same unique value (School ID and school name) among every data files. For every CSV file, it consists of several variables such pupil to teacher ratio, demography, and school survey result. Later on, we are interested if the school quality (reflected by school survey result) has an impact from gender demography and pupil to teacher ratio.

However, in this project, we only focus on the ETL part. The analysis and the visualization is not included in the project scope.

1. **Method**

These are the steps we did for the ETL project. We will explain in detail on each process.



A picture containing clock, drawing

Description automatically generated

A close up of a logo

Description automatically generated

A close up of a sign

Description automatically generated

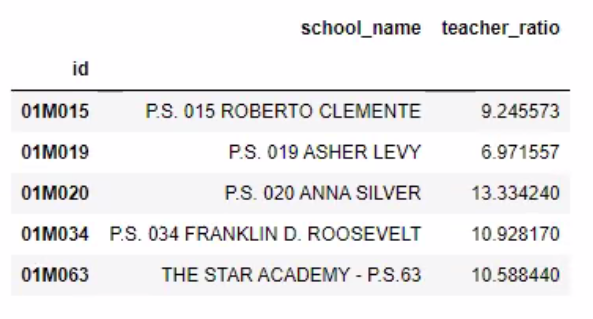
1. **Extract** 
   1. 3 CSV files from New York School Dataset in data.world. <https://data.world/city-of-ny/uchs-jqh4>
   2. We extracted our CSV files sources with Python and read it as Pandas Dataframe.
2. **Transform**

We did the transformation part in two environments. The first one is with Python and the second in Postgres.

* 1. Using Python, we did several data cleaning such as delete column, rename, and formatting.

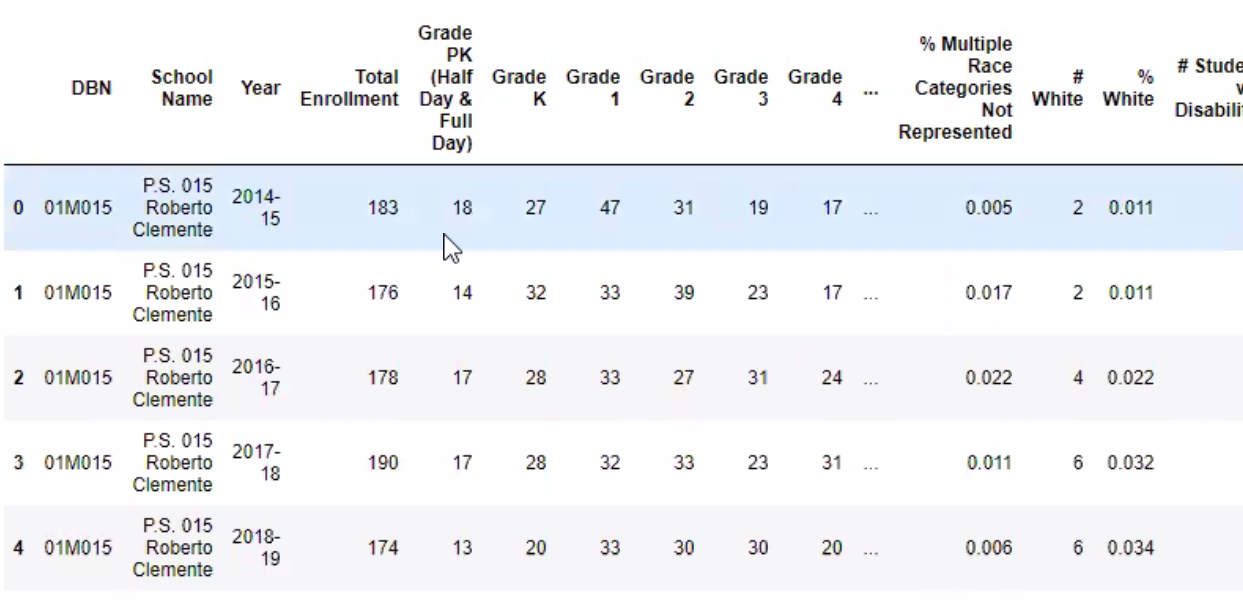
First table: Pupil\_teacher ratio

In this table, first we rename all columns in this table and we set the index into id.

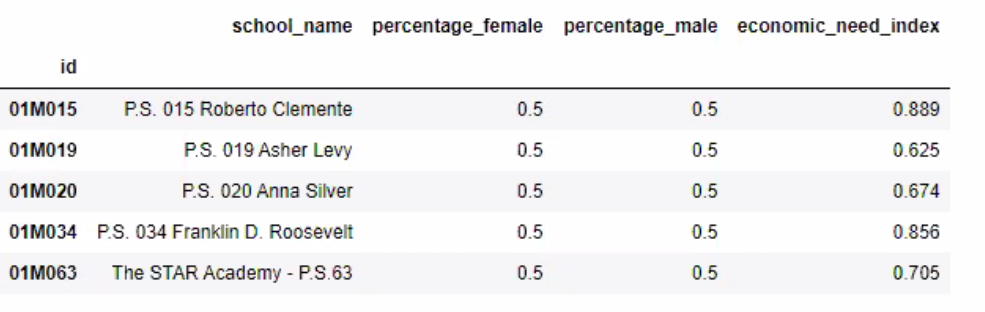


Second table: school\_demography

We only select columns that are relevant to our project. After that, we filter the data only from selected year (2014-2015). And the we rename the columns so that it has consistency. We also did the formatting for some number, and set the ID become the index.



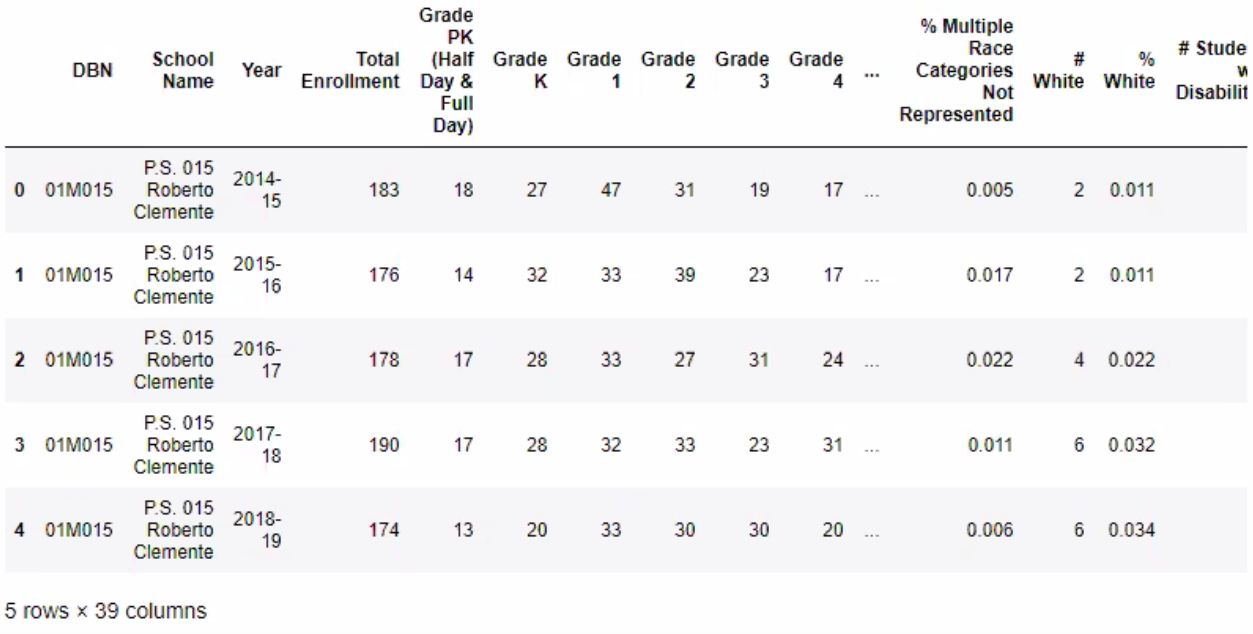
after



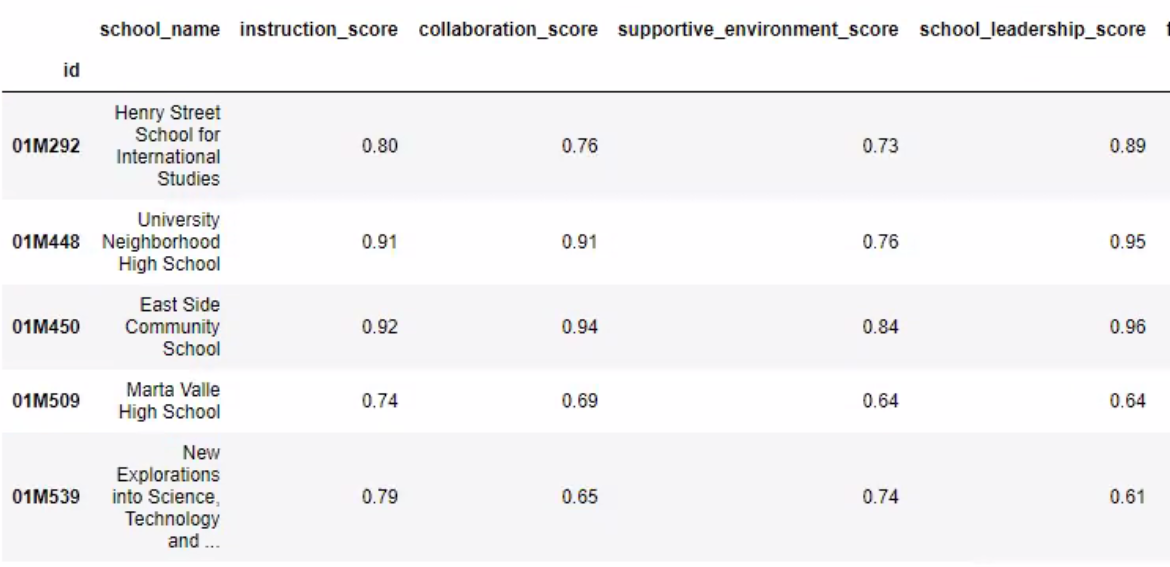
Third table: school\_survey

We selected columns that are relevant to our project. And the we rename the columns so that it has consistency with other tables. We also did the formatting for some number.

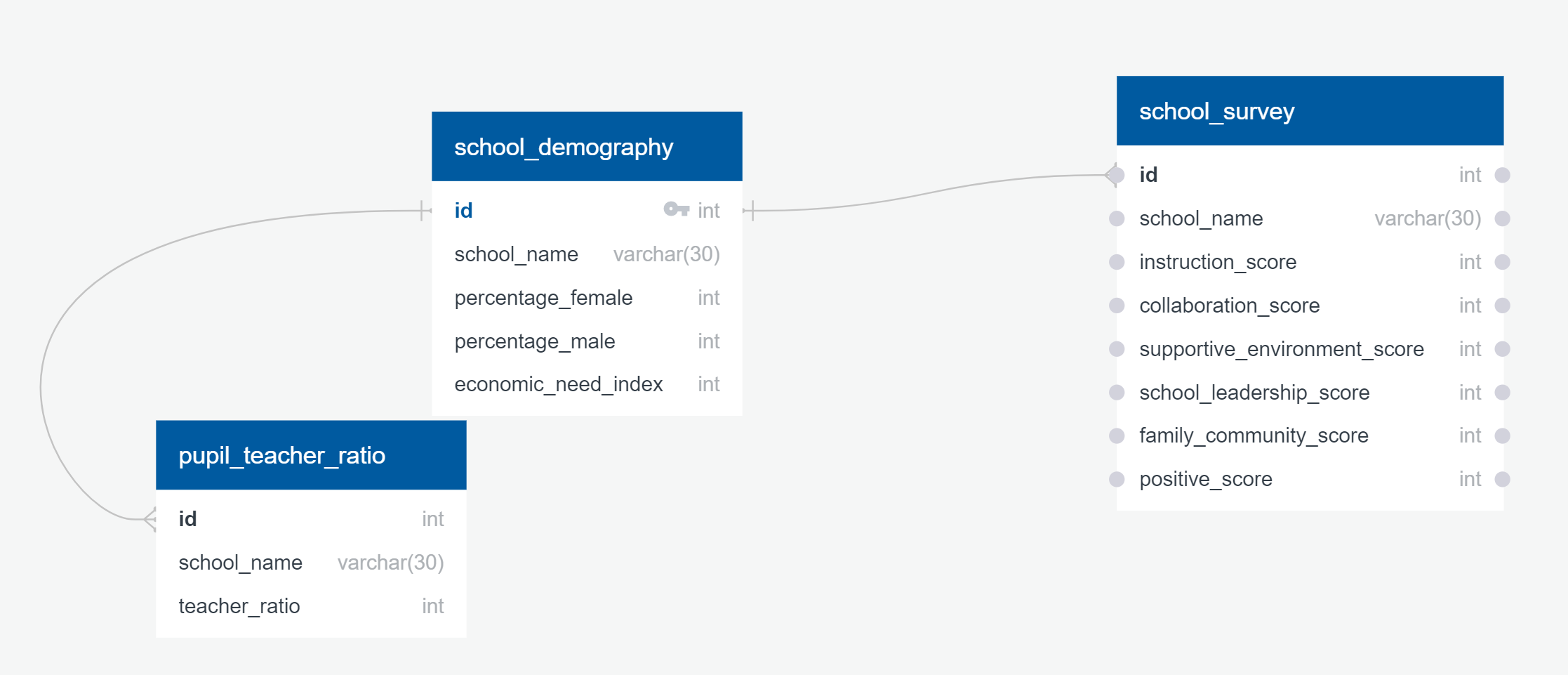
Before



After

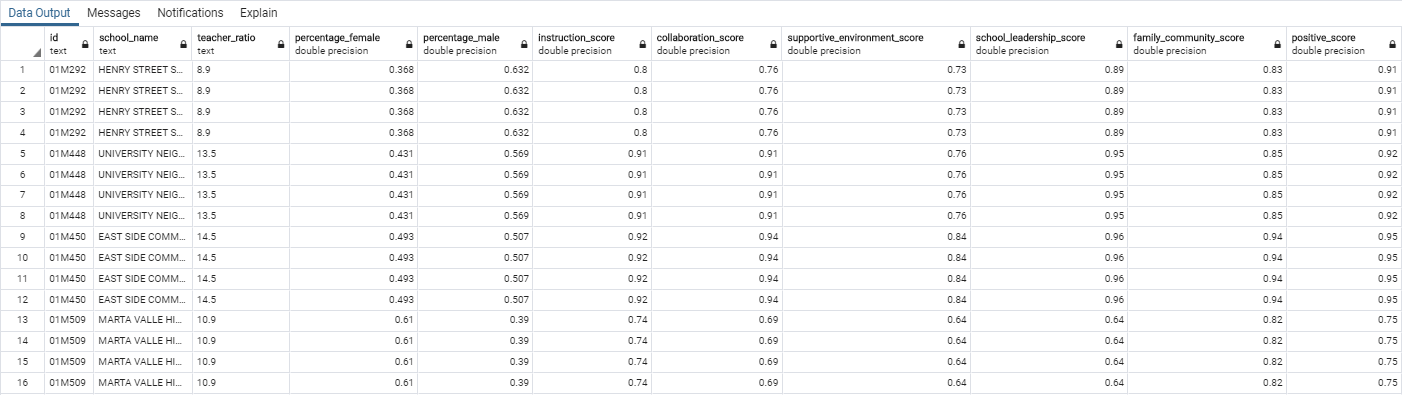


1. Load
   1. We created a new Postgres database. School id is the Primary Key and Foreign Key.



1. **Analysis**

After we load the database, we joined the tables to see the combined output with school id as the Primary key.



After we can combine all the tables, this dataset is ready to be analyzed. In the future, we can see the relationship and correlation among the variables: such as the impact of gender demography and pupil teacher ratio to selected school quality.