School Attendance in NYC 2015-2018

Extract: Originally I was going to use a csv file found on Kaggle, however, it only covered the 2011 school year, which felt limiting. After a brief Google search I came across the NYC Open Data site (<https://data.cityofnewyork.us/> ) with more up-to-date records and over a wider timespan. I was able to pull the data by importing the Socrata module (Socrata Open Data API) into my Python script.

Transform: This provided me with daily attendance reports for all NYC schools from 2015-2018. I began by grouping the data by ‘school id’ and ‘school year’ to find the average tally for present students and absent students. I dropped the ‘released column’ as the numbers seemed insignificant. After discussing with my team, I decided to further group the data only by school id, effectively getting the average tally over the the whole three year span, instead of on a year by year basis. From here I added a column where I calculated the percentage of absent students for each school, as this allowed for a quicker side-by-side comparison. Finally I indexed the table by school id, as this was the common factor for joining with my other team members’ data.

Load: We merged our separate Jupyter Notebooks together using nbmerge in Terminal. Once merged, we did some additional cleaning and unifying of column names. After generating the frameworks for the final tables in PgAdmin, we uploaded our transformed databases into them. In PgAdmin, we joined the tables into one using ‘school id’ as the common index. For this we used an inner join, as we were only interested in the schools that had data in all tables. Finally, we after more discussion we dropped some columns that felt less relevant to the broader picture.