Justin Huang
Joaquin Alvarado

ETL Project Report

Synopsis:
Working in the field of education, Joaquin's school was presented with a COVID challenge in the form of end-of-the-year assessments. Typically, schools in CA must participate in state testing every year, this assessment gauges the school's mastery level per grade and compares it with the state. Deeper analysis of this data allows schools to measure growth in grade levels, subgroups, and priority standards. Joaquin also uses this data to initiate interventions and proper placement of student courses (i.e. honors courses). Due to COVID, state testing was cancelled, but Joaquin's school still needs to show growth in their subgroup populations (i.e. SpEd, ELL...). Furthermore, to compensate for the learning gaps created during Distance Learning, Joaquin will be leading 4 summer school course (2 math and 2 ELA). In lieu of state testing, his district encouraged assessing with IReady (an online platform assessment that diagnosis student mastery of grade level standards). Since students are still working on the assessment, IReady does not allow users to export current data via csv for further analysis.

Extract:
The data we wanted lived on the IReady platform and was organized in a table. Naturally, we thought of scrapping the webpage and then creating a pandas dataframe. After multiple attempts, we encountered a challenge with login credentials and thus our data scraping came to a halt. After some help we were successful in saving the webpages of interest into our desktop and using pandas to scrape and create dataframes for 6th, 7th, and 8th grade math and reading.

Knowing that student names might have to remain confidential when publishing data findings, we needed a way to identify students by student ID. Thus, Joaquin obtained a master list of all students, their grade level, and their student ID. This data was extracted from one of Joaquin's online platforms in the form of a CSV.

Transform:
Given that the IReady data was acquired from html, Justing worked diligently on cleaning our newfound dataframe. He dropped all columns that were not of interest, renamed columns, got rid of nulls, reset the index, and split grade level mastery by blank space to only keep grade level number (i.e. instead of showing "Grade 4", the dataframe shows "4"). He also merged all 3 math grade level data into one dataframe and also merged all 3 grade level data into one dataframe.

The csv proved to be a quicker process as we only needed a couple of transformations to ensure uniformity with our html dataframe. We first merged the last_name and first_name columns and then renamed them to ensure they matched our other dataframe. This would prove to help us when loading onto our final database.

Load:
Both data sets were ultimately loaded into an SQL database (PostgreSQL). By joining these two data sets in PostgreSQL, we can create the queries needed in Joaquin's analysis and keep student names anonymous (dropping name columns when publishing).