ETL Report

Allison, Carlos, and Curtis

For this project we extracted, transformed, and loaded multiple data sets to a non-relational database. We extracted the first dataset from Kaggle.com. It contained ~10,000 rows and 251 columns of data pulled from Zillow.com. The columns provided information regarding housing prices over a period of 12 years. We extracted the second dataset from the Tennessee Department of Education, and included standardized test scores for every school district in Tennessee. After saving these data into CSV files, we created data frames using Python’s Pandas library on Jupyter Lab.

After creating these two data frames, we removed the confusing and cluttered column headers and merged the 3rd and 4th rows to serve as new headers (on the data set from the Tennessee D.O.E). Next, we split the column with school district information. We did this so that the school district would stand alone in a column, and enable us to merge the two datasets on district/county (many of the school districts were simply named after the county they were in). After merging on the county/school district, we dropped unnecessary columns that contained outdated data (we were looking at only the past five years). Additionally, we dropped the subject specific test scores, so that we could focus on the composite scores. Finally, we performed a mathematical operation to find the percentage increase in housing value for each school district, which you can find in the ‘CAGR’ column of the clean\_df data frame.

In order to load the data frames into a non-relational database (MongoDB) we created two dictionaries by iterating through the data frames. The first has the metro and region name as keys. The second has the region name, school scores, and CAGR as keys. We inserted the two dictionaries into MongoDB as ‘Metro’ and ‘Cities’. We chose to use a non-relational database for its flexibility regarding inserting arrays as entries. See below for a visual example of the database mentioned above.

