For Project 2 our research was aimed at comparing general housing sales data of Texas and the U.S. over a ten-year period. It has been apparent that this trend has increased in recent years, being marked by a meteoric climb in housing prices. The data we found and processed clearly illustrates this point. We relied heavily on Python Pandas, MongoDB, and CSV-formatted tables for the bulk of the project. We pulled the data from housing and government websites, filtered it, and loaded it for utilization.

Two website sources were utilized to pull the data were:

* <https://www.recenter.tamu.edu/data/building-permits#!/state/Texas>
* <https://www.census.gov/construction/bps/permitsbyusreg_cust.xls>

For the Texas data, the first URL was read into the program using browser.visit, and then Beautiful Soup was utilized to extract the data because the tables were formatted as xml documents. That data we then pushed to Pandas DataFrames. The extraction process produced two DataFrames because the tables on the site were not formatted as typical html tables, thus could not easily be read with Pandas. The two data frames (Year) and (Units) were concatenated to produce one data frame.

The national data was pulled from the census.gov site in a CSV file. Pandas was utilized to read in table and format it. Pandas was used to create two-column data frame to compare Year and Units like the web scraped data for Texas.

A MongoDB called ‘permits\_db’ was created to store the data in two collections; US and TX which can be queried from python. In order to import the dataframes to we had to convert them to a JSON format.