ETL on Housing Price, Population, Crime, and SAT Score Data

Written by Brian Klovert, Yamini Sasidhar, Youqing Ye, and Fardi Yueng

Table of Contents:

Summary

For the ETL project our group decided to extract, transform, and load datasets for use in analyzing housing trends in America. In the interest of time, we decided to keep our analysis domestic versus comparing data globally. When deciding what date range to go off from when analyzing the housing market, we definitely wanted to pick a period of time when drastic trends (either good or bad) occurred. Since the housing bubble in 2008 cause the great economic downturn started in America, we wanted to see how the housing market responded in 2010 – 2016.

Data Overview

In order to make our analysis as clean as possible, we wanted to keep our variables straight forward and only focus on the factors we believe make the biggest impact when making a real estate purchase. The variables we decided to focus on were: overall population by state, price of property by state, crime rates by state, and SAT performance by state. We started extracting the data from multiple sites including *data.world* and *The College Board*. The process to extract the data from sites did not take us as long as some of the other phases of the project. We had a couple CSV files from the previously mentioned data sites, the cleanup was fairly simply as we dropped various years from our file since we are only analyzing 2010-2016 housing data. We utilized Pandas to rename columns from our crime rates CSV dataset as follows, “unnamed columns” to “2010”, “2011”, “2012” etc.

Data Cleanup

Extraction. Our extraction of the Housing Price CSV file has price by state, but prices are populated with year-month columns. Only price from 2010-01 to 2016-12 are included. In addition, there are regional data mingled in the same column where states are saved. We removed Regional rows from the dataset as our analysis is focused around state rather than region. The size rank column is also removed as there is no plan to use this data. For transformation, additional columns for each year are added from 2010 till 2016 which the average of 12 months for the corresponding year are loaded into these columns.

Eventually, we removed the column data of yyyy-mm and condensed the data into 6 separate years (2010-2016). In order to prepare proper data analysis in a clean fashion, we transformed the table into a pivot table with each state showcasing its yearly housing price. With regard to the Population by State CSV, no extraction was needed as the file contained proper population and state data. The only transformation that has been performed is to create a pivot table out of it so each state like the house price will have its own

population number by year. In the pivot table, we were able to create a chart (**Figure 1**) to show where the majority of citizens reside in America for each year in our analysis.

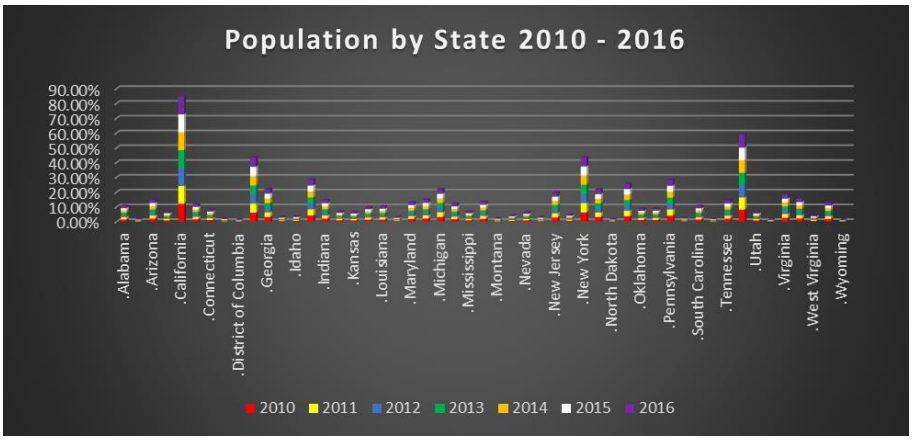

Figure 1: Population by State (2010–2016)

Transformation. The biggest stop gap we ran into during the course of the project was transforming SAT scores *The College Board* into what could be relevant for this analysis. We stumbled upon a great range of data on the site, but it was stored in a PDF which made it very difficult to scrape. We had to get around the scraping by going to Jupyter Notebook and writing a script that would import the PDF data and extract the portion of the PDF where Critical Reading and Mathematics performance was stored for the 75th percentile of each state. We decided to drop the Writing scores of the SAT since that portion of the exam is now optional for students. A sample table is provided in **Table 1**.

Table 1: SAT Percentiles per Section (Hawaii)

SAT	State		
Percentile	Critical Reading	Mathematics	Writing
75th	560	580	540
50th	490	510	470
25th	420	440	400

Load. Once all of our files were cleaned up and all the data we needed to conduct our proper analysis, we were ready to take on the final phase of the project. The loading process consisted of creating connections to relational databases using SQLAlchemy. The local database was created using MySQL Workbench, and loaded with five tables total, one for each dataset retrieved: population, housing_price, crime_rate, crit_scores, and math_scores. The CSV data and web-scraped SAT data were pushed into dataframes, and loaded into the MySQL tables using the database connection. Once we pushed to MySQL we were able to see all data in one condensed place. We can view every state's population sum, crime rates, housing prices, math and critical reading SAT scores all in one place.

Additional Data Analysis

We used Tableau to overlay the percentage change over the prior year for each state on Population and the mean of House price. Along with the US map with mean house price for the year from 2010 to 2016 on each state. Using dashboard, we tied the data together so users can interactively compare different state to see if there are any noticeable trends by filtering per state however they like (**Figure 2**).

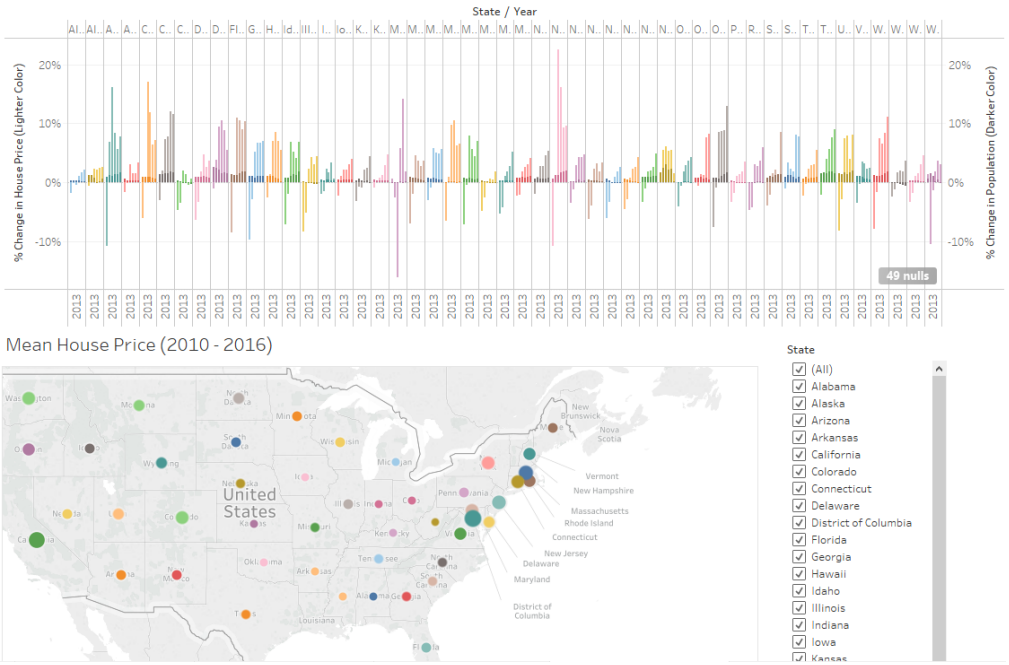

Figure 2: Tableau visualization of changes in housing price and population between 2010-2016.

For example, if you wanted to see the trends of the mean of housing prices between Alabama, California, and Connecticut, all the user would need to do is simply check off every state other than AL, CA, and CO. **Figure 3** gives a visualization of how each state's housing change from 2010-2016 is represented.

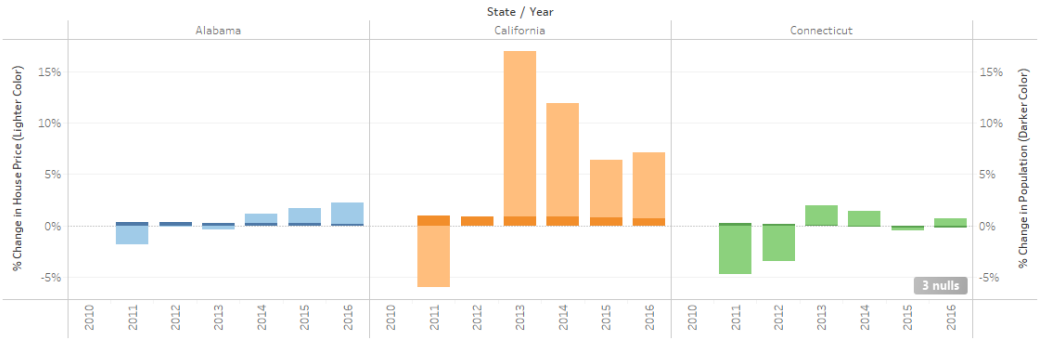

Figure 3: Tableau visualization of changes in housing price and population between 2010-2016.

Discussion

We performed this data analysis to see whether we could identify any relationships between housing prices, crime rates, and SAT performance per state as the datasets under study. The tools we used include: Pandas, Tableau, MySQL, SQLAlchemy, BeautifulSoup, Splinter, PyPDF2, and MS Excel PivotTables.

From our surface analysis, we succeeded in finding: (1) housing prices dropped across the board from 2010 to 2011 after the housing crash in 2008, and (2) housing prices changed substantially in most states after 2011. We found that the changes in population are much lower in percentage, and warrant further analysis for finding the statistical significance of these results. From this surface analysis, we find there is no noticeable relationship between population and housing prices.