**ETL Project** **Group Members:**

**Brian Klovert, Yamini Sasidhar, Youqing Ye, Fardi Yueng**

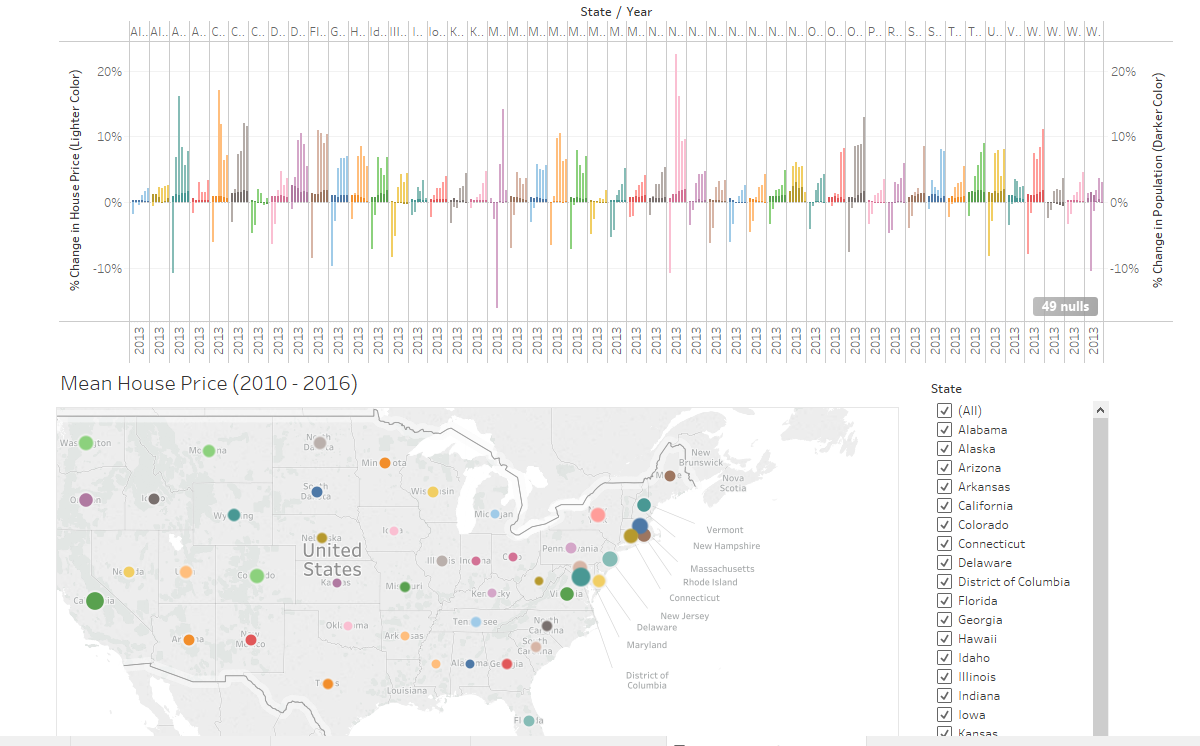
For the ETL project our group decided to extract, transform, and load datasets for use while analyzing housing trends in America. In the interest of time, we decided to keep our analysis domestically rather than compare globally. When deciding what date range to go off from when analyzing the housing 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 reacted from 2010-2016.

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, price of property by state, crime rates by state, average SAT scores by state, and overall population by state. We started extracting the data from multiple sites including, dataworld, collegeboard (SAT scores), 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.

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. In 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 to show where the majority of citizens reside in America for each year in our analysis.

The biggest stop gap we ran into during the course of the project was transforming the collegeboard SAT scores 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 average scores were stored for each state. The scores within the dataset were separated by “Male and Female”, where we had to join the genders into one score as our analysis is based around housing trends and has nothing to do with gender. We decided to drop the Writing scores of the SAT since that portion of the exam weighed our results down a bit and was not as statistically significant as the other two.

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



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 importing created engines from SQLAlchemy. We created the engine by connecting to mySQL which resulted in all files being pushed to the server. Once we pushed to mySQL we were able to see all data files 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.