**Final Report**

**Extract**:

For this project, I chose to get housing cost data. The data provided was in 8 separate CSV files. I retrieved the data from the following url:

(<https://www.kaggle.com/paultimothymooney/zillow-house-price-data>)

All of the CSVs were loaded in as pandas data frames.

**Transform:**

Each of the CSVs were separated based on housing type. The first step I took was, for each CSV, I added a new column specifying the housing type (1 bedroom, 2 bedroom, 3 bedroom…etc). This was done so that when the dataframes were combined, they could be distinguished.

After the housing type was specified in each dataframe, all 8 dataframes were combined using “append.” This was done so they would all be added on as additional rows.

I only wanted to look at the past 5 years, so I used “.loc” to select the columns that were for the past 5 years.

The columns were month-by-month, but I wanted to see the data as year-by-year. I averaged all the monthly prices by year and created columns for the yearly averages.

I removed the month-by-month prices leaving only the yearly averages.

Lastly, I sorted the all the values based on the “RegionID” column.

**Load:**

I chose to use MongoDB for my database. I made this decision because it was the simpler option.

For the final collection, the following headers were used: 'RegionID', 'housing\_type', 'RegionName', 'RegionType', 'StateName', 'State', 'Metro', 'CountyName', '2016\_avg', '2017\_avg', '2018\_avg', '2019\_avg', and '2020\_avg'.

These were chosen because I only wanted to look at yearly averages for the past 5 years.

To begin the loading process, I needed to make sure that all the data was in string format. I used a loop to convert all data to strings, create a dictionary for each converted row, then insert the dictionary into MongoDB one by one.