**HOME L.L.C ASSIGNMENT**

***To study the effect of national factors on house price rise in USA****.*

**Task -** Using publicly available data for the national factors that impact supply and demand of homes in US, build a model to study the effect of these variables on home prices.

**Approach -** The following variables are chosen for the study-

* Unemployment Rate
* Per Capita GDP
* Median Household Income
* Construction Prices
* CPI
* Interest Rates
* Number of new houses supplied
* Working Population
* Percentage of population above 65
* Housing subsidies
* Number of Households

As a proxy to the home prices, S&P CASE-SHILLER Index is used.

Most of the data is downloaded from [[https://fred.stlouisfed.org/]](https://fred.stlouisfed.org/%5D).

Data for all the variables is downloaded, pre-processed and combined to create a dataset. Data for different variables had different frequencies. So, to combine the data, necessary interpolations are made.

Linear Regression is used as most of the variables have high correlation with the target variable.

A few variables that could have been studied are below.

* Net-immigration (It is supposed to have a positive impact. No suitable data could be found)
* Marriage Rate (People tend to buy homes after they get married. So, it might have some effect. No data could be found)
* Average house size (The data was available only for the years after 2015. Though it is expected that an increase in the average house size would increase prices, it is found that the average home size has been consistently decreasing although the prica has been increasing)
* Land availability (Less land, higher prices. Perhaps this is why the prices has been rising even though the average house size has been decreasing. No relevant data could be found)
* Tax Rate (Too many brackets (7) and could not analyse due to time)
* Number of active listings (Data prior to 2017 was not found)

**Following sources were used to gather data -**

CASE-SCHILLER Home Price Index - <https://fred.stlouisfed.org/series/CSUSHPISA> Interest rates - <https://fred.stlouisfed.org/series/FEDFUNDS> Unemployment rate - <https://fred.stlouisfed.org/series/UNRATE> Income - <https://fred.stlouisfed.org/series/DSPIC96> Per Capita GDP - <https://fred.stlouisfed.org/series/A939RX0Q048SBEA> New Constructed units - <https://fred.stlouisfed.org/series/COMPUTSA> Construction price index - <https://fred.stlouisfed.org/series/WPUSI012011> percent urban population - <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?end=2021&locations=US&start=2001> Housing Subsidies (Federal) - <https://fred.stlouisfed.org/series/L312051A027NBEA> Total households - <https://fred.stlouisfed.org/series/TTLHH>

Pre-process of data was done all the 10 parameters. GDP data was quarterly. We will impute for other months using linear interpolation after we create the final data frame combining all the data.

The "Per\_Capita\_GDP" column had missing values because the data was quarterly. The missing values in the other columns is due to unavailability of fresh data. We will first fill the missing values in the "Per\_Capita\_GDP" column using linear interpolation. We will drop the rows having missing values in the other columns. We will use data from 2003 to 2023.

Our pre-processed dataset saved as "prepared\_dataset.csv" at the end of merging all files(parameters).

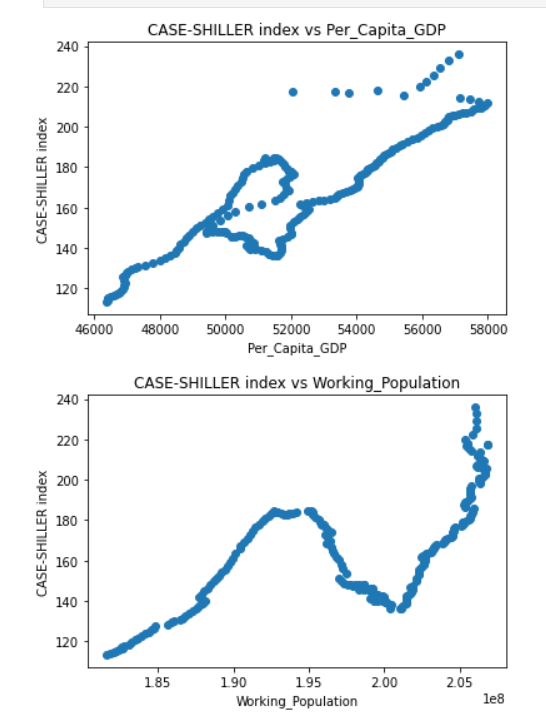
And we start the analysis part.

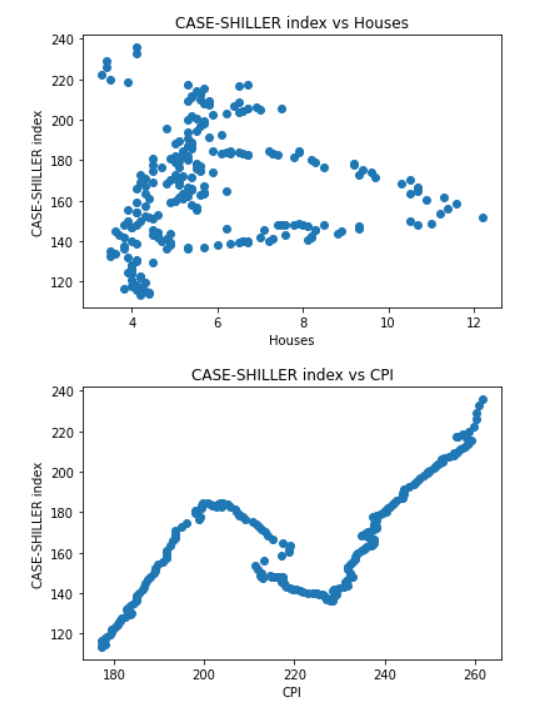
We don't need the month and year columns for our analysis. So, we dropped those columns.

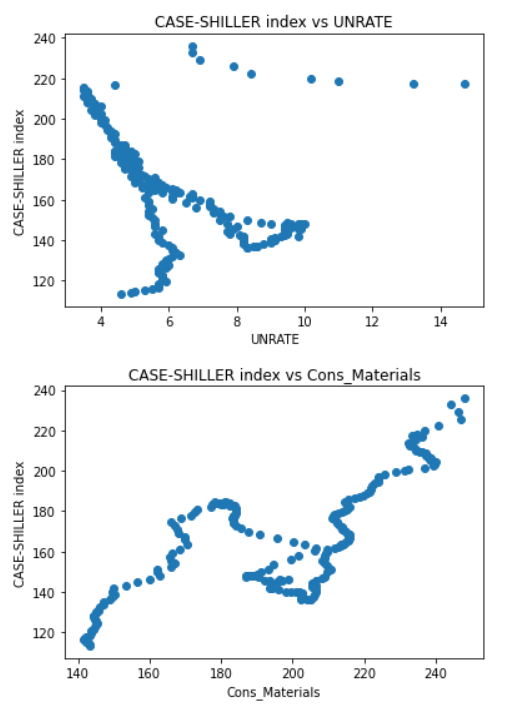
Let's look at the correlation among the variables.

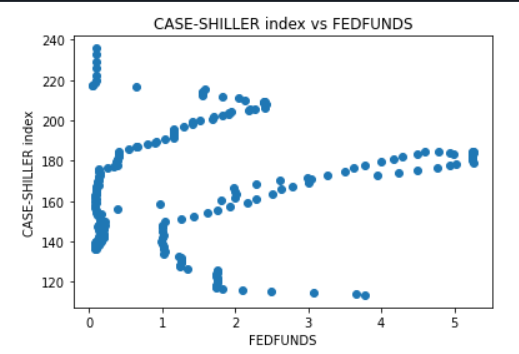


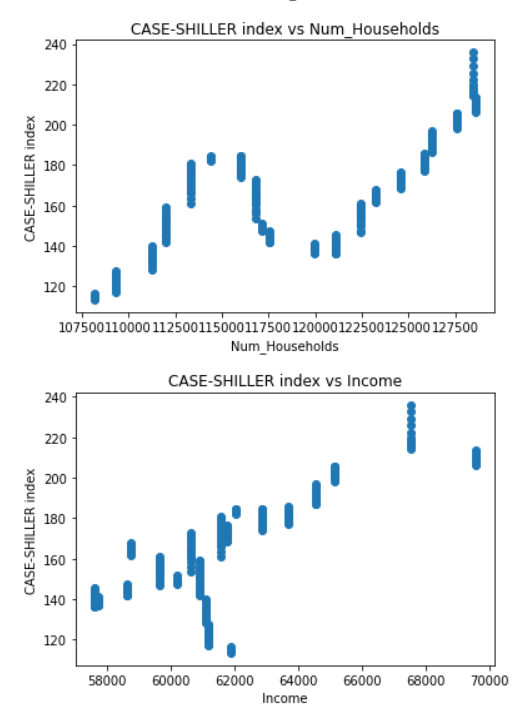
EDA

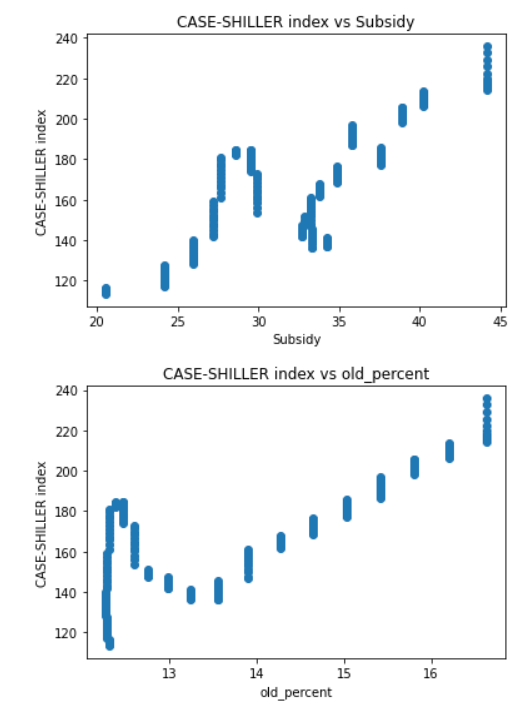










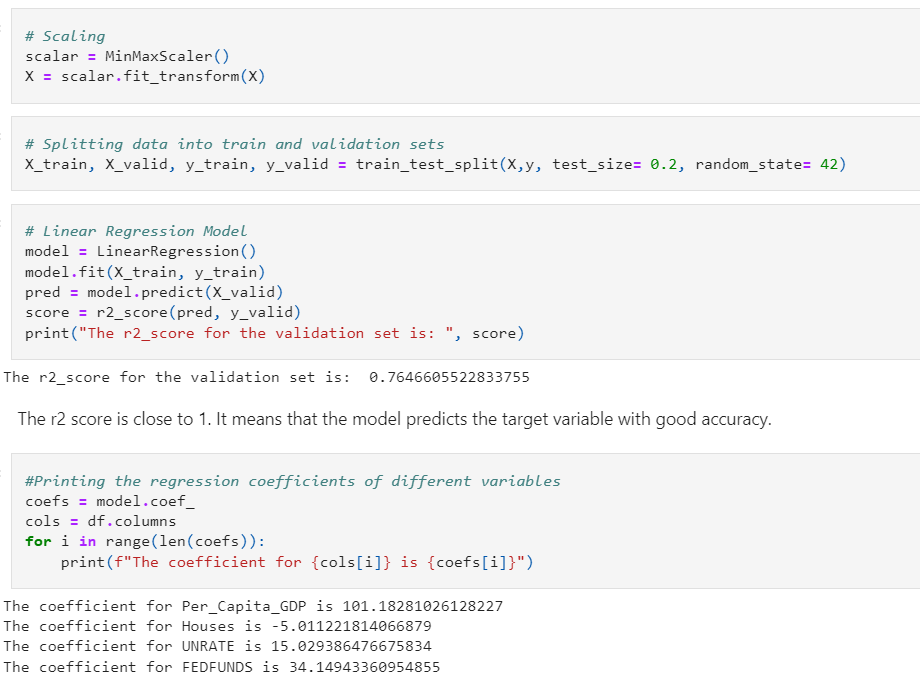


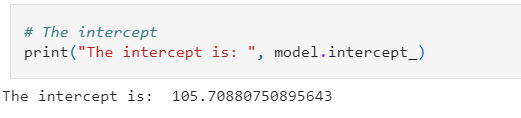
From the correlation matrix and the scatter plots, the following conclusions can be made-

* Unemployment rate is negatively correlated with the home prices. Which makes sense.
* Number of new houses has almost 0 correlation. This is unexpected as one would think that the number of houses supplied should also affect home prices. Maybe the number of new houses built are not enough to bring the prices down significantly.
* Impact of the great recession is reflected in almost all the plots.
* Interest rates have a slight negative impact on the home prices.
* All the other variables are positively correlated with the home prices.

From the correlation matrix above, it can be seen that several pairs of independent variables have high correlation. We drop one column from each such pair as they will cause multicollinearity. We will drop the column which has lower correlation with the target.

**Model Building**





**The coefficients of the variables are not in sync with their correlation coefficients with the target variable. For example, the coefficient of "UNRATE" (unemployment rate) is positive. It does not make sense as a rise in unemployment will surely lead to a decrease in the home prices. So, the question arises, why did this happen?**

I found the answer on [https://www.researchgate.net/post/What-does-it-indicating-If-there-is-positive-correlation-but-negative-regression-coefficient].

"... reason is what some call "real suppression". In this case suppression occurs because an important predictor variable, necessary in understanding the true relationship between the latent variables, suppresses the effect of another predictor variable."

As most variable increase with time. So, it is expected that they will have high correlation. So, instead of depending upon the linear model, we can directly study the impact of variables on the home price index using the scatter plots.

A few variables that could have been studied are below.

* Net-immigration (It is supposed to have a positive impact. No suitable data could be found)
* Marriage Rate (People tend to buy homes after they get married. So, it might have some effect. No data could be found)
* Average house size (The data was available only for the years after 2015. Though it is expected that an increase in the average house size would increase prices, it is found that the average home size has been consistently decreasing although the price has been increasing)
* Land availability (Less land, higher prices. Perhaps this is why the prices has been rising even though the average house size has been decreasing. No relevant data could be found)
* Tax Rate (Too many brackets (7) and could not analyse due to time)
* Number of active listings (Data prior to 2017 was not found)