**Data Mining Project**

Analyzing Housing Price Index of United states and other economic features

**As a part of Data Mining course this project is done. This project is done by using Python programming language. The main aim of our project is finding out which is at best chance in predicting Housing Price Index of United states (Correlated State or Correlated Economic Factor) based on knowledge discovery in data bases for the relation between Housing price index of United states with mostly correlated state and mostly correlated Economic index. We selected the state which is most correlated to United states and most correlated economic factor in predicting the Housing Price**

Maddulety Swamy Yettikadi

(700656805)

*Sumanth Vandavasi Venkata*

*(700654072)*

**Project Details**

**Programming Language: Python (3.5.2)**

**Libraries: MatPlotLib, Pandas, Sci- Kit Learn, Numpy, Simplejson, Quandl**

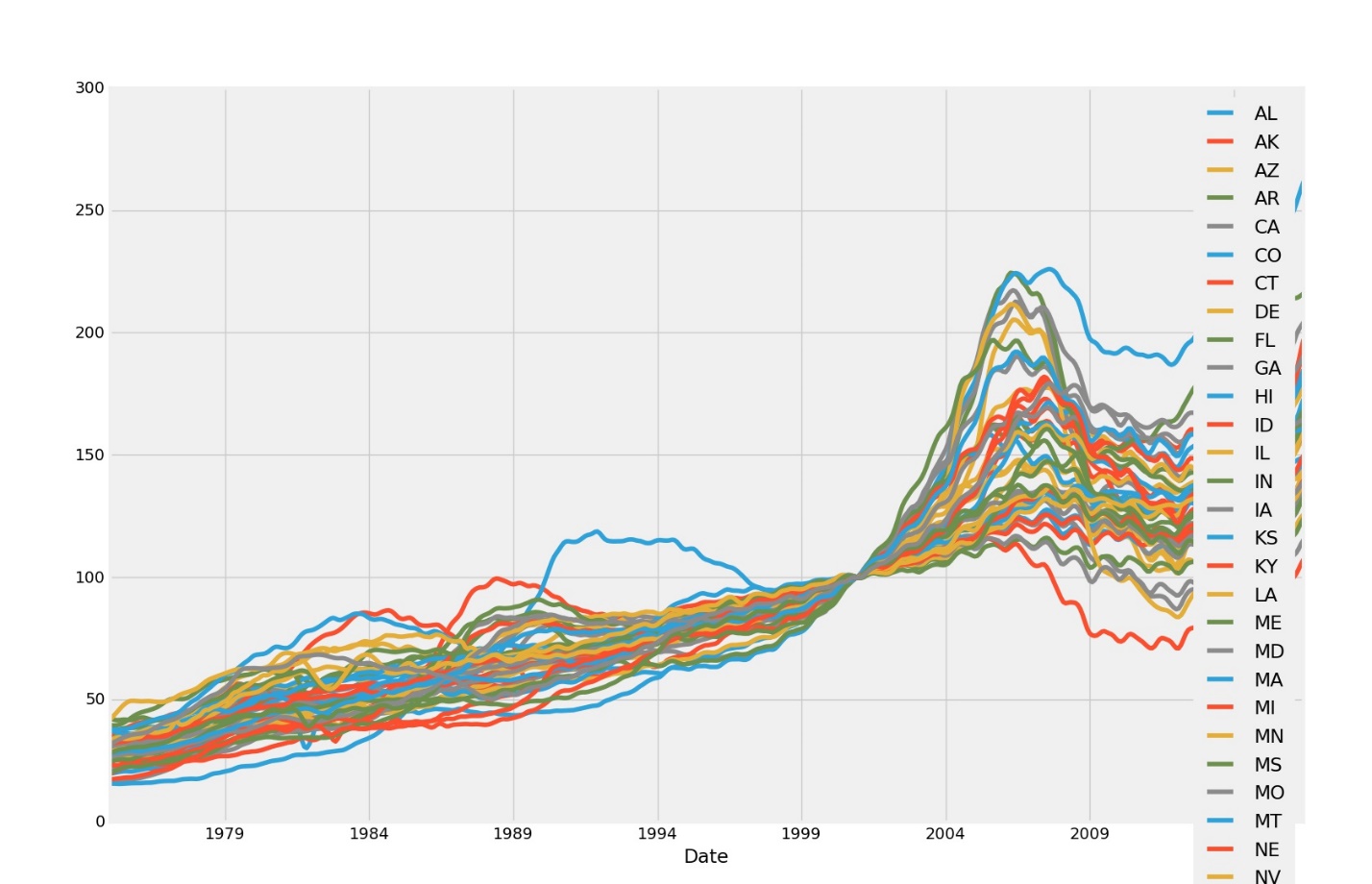
**IDE: Anaconda (4.1.1)**

**Domain : Financial Data**

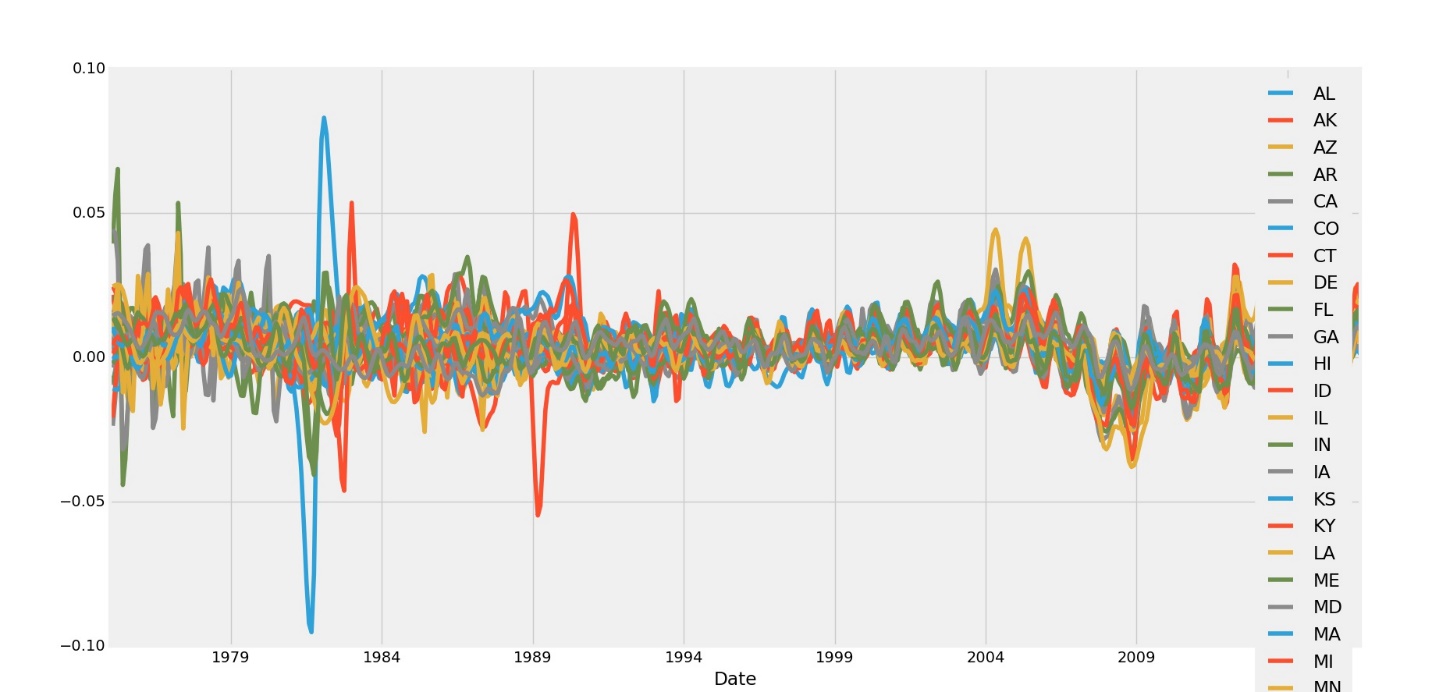
**Data Source: WEB (Quandl, Wikipedia)**

**Knowledge Discovery from Databases:**

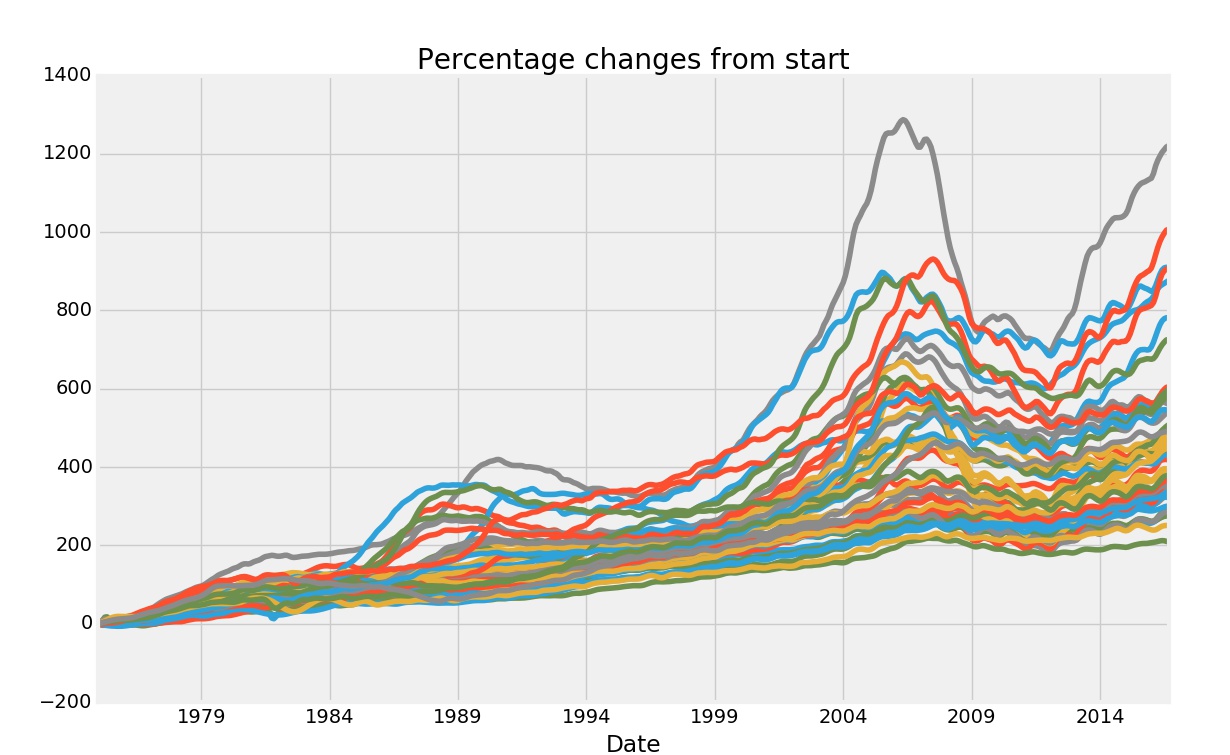
* **Data Collection: The data is collected mostly from Quandl, a financial database which provides required data through api, the data of each state need to be collected. Quandl provided data for each state separately so it would be hectic to hardcode for the 50 states so instead what we did is used a table from Wikipedia and used the data of states for extracting the Quandl data of states. Also, Quandl has all the economic indicators data and choosing the ones that best match to our needs and syncs with all other data is a challenging task. After collecting data using python we saved all the data in CSV format and PICKLE format so that they can be directly used instead of extracting from web each time.** All the data files have been included in the bundle.
* **Data Cleaning Plan: The data has been cleaned at many phases, initially for extracting the data the Wikipedia source had many tables which are absolutely of no use to use so only the states table is extracted and from the states table we needed only the names of states so all other information is ignored. While calculating the correlation between SP500 and Housing price index of United there are some null values that has been handled. Also, while calculating correlation between Housing Price Index of United states and Unemployment of United states, we had information of Unemployment till 2011 so the dates after that are removed to handle the null values. All other data sources are chosen in the required format that fits our data.**
* **Data Integration Plan: Housing Price Index of each state forms one dataset all the states information along with United states data has been integrated in one dataset (This could be found at main\_data\_with\_us.csv or main\_data\_with\_us.pickle ). The economic indicators used for comparison are GDP, Unemployment Rate, S&P500, Mortgage Values. All the indicators data is integrated to Housing price index db of United states to find the relation. (This could be found at df\_economicdata.csv or df\_economicdata.pickle). All the Individuals data can be found in their respective files in the bundle.**
* **Data Selection Plan: To select the attribute that are required to perform regression, we used the correlation operation. After performing correlation matrix for both United states with states and United states with economic indicators, we selected state with highest correlation and economic indicator with highest correlation. The correlation charts can be found in Correlation\_with\_Us\_States.csv and df\_economy\_correlation.csv. So the state with highest correlation Virginia ‘Va’ and highest economic indicator ‘S&pP500’ are selected to perform further data mining techniques.**
* **Data Mining Plan: After the parameters required are selected the linear regression with a single variable is used to predict the US housing prices. The regression algorithm for python is used using scikit-learn library( the regression code is used and modified accordingly from** <http://scikit-learn.org/>**). The results of the linear regression are with Virginia as attribute Coefficient(w= 0.**842) and Mean Squared error is 39.66 and variance score is 0.69. The result when the best economic indicator ‘S&p500 ’ when passed are Coefficients(w= 7.18) and Mean squared error 1559 variance score -2.83. After running many times with varying conditions these are the best results we came up with. So using states data we can predict the housing price index of United states much better than any economic indicators.
* **Pattern Evaluation Plan:** We used visualization techniques to understand the flow and the process and we captured all the process in graphs. The below graphs explain in detail all the process and results.

****

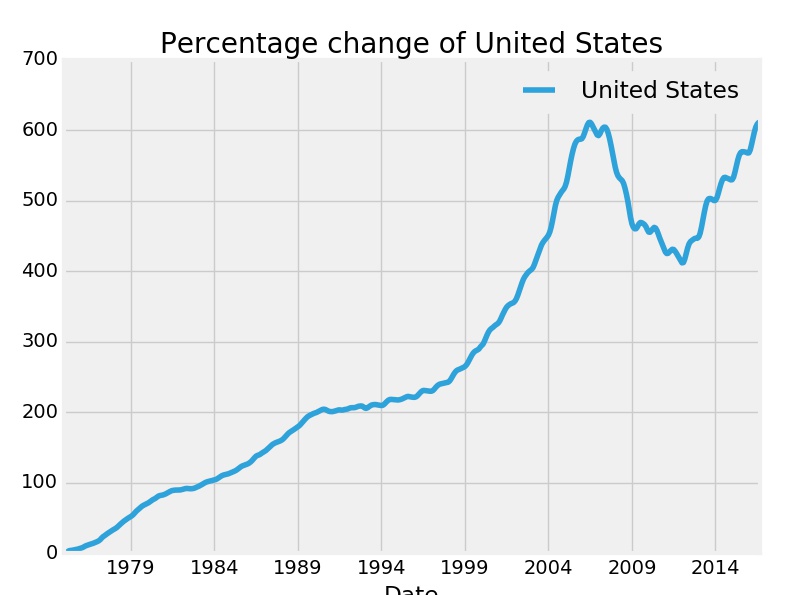
**This graph indicates data after clubbing together and it shows the percentage change with 2000 as base.**

****

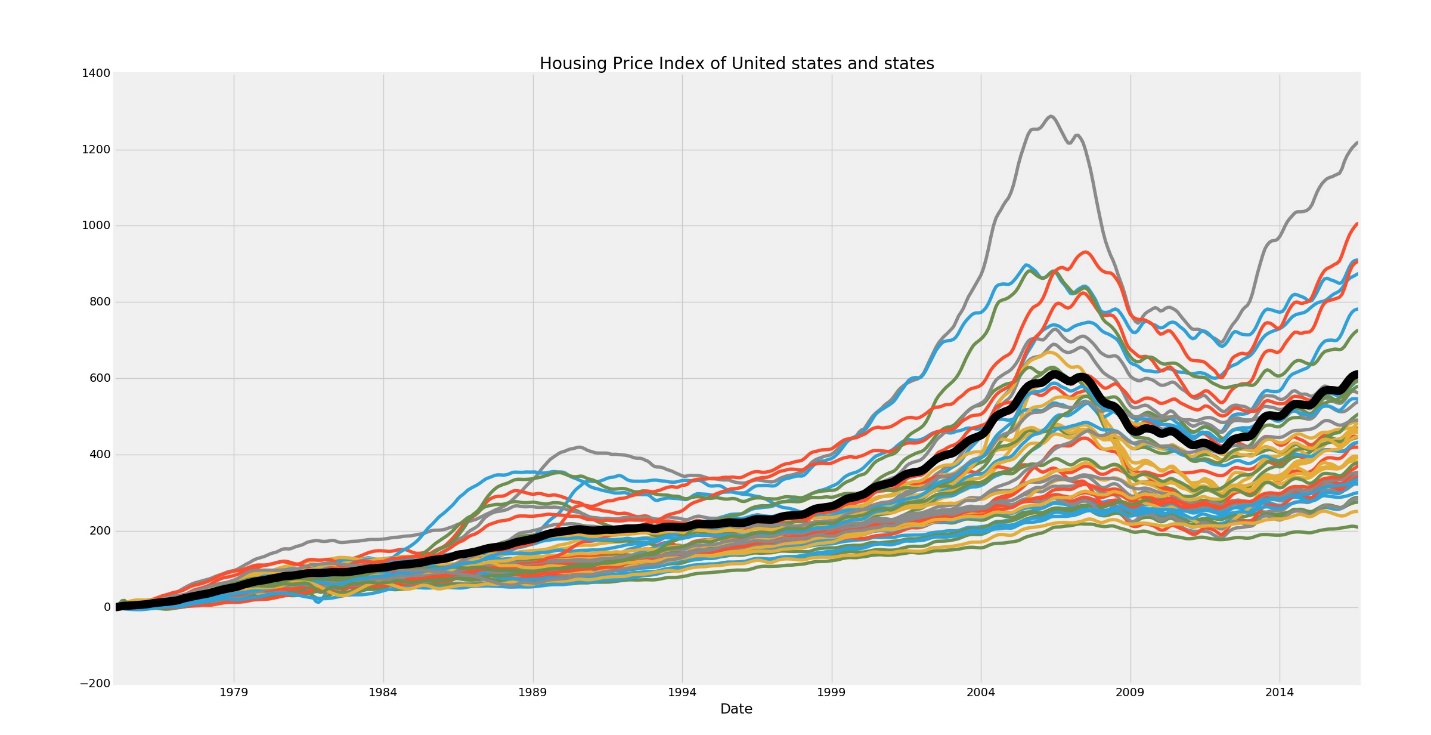
**The above graph shows percentage change of all the states when compared to the previous year.**

****

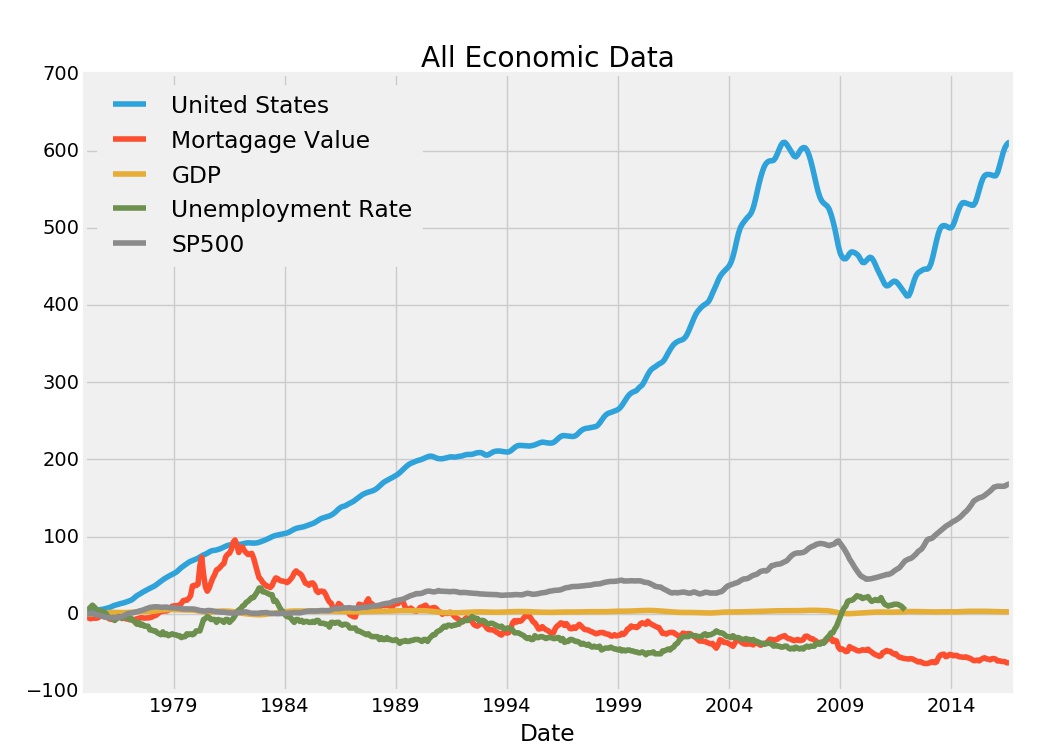
**The data is modified in such a way that we compare the percentage change from the start so that it is easy to analyze the pattern of the data.**

****

**The above graph shows the housing price index of the united states. This is percentage change from the start.**

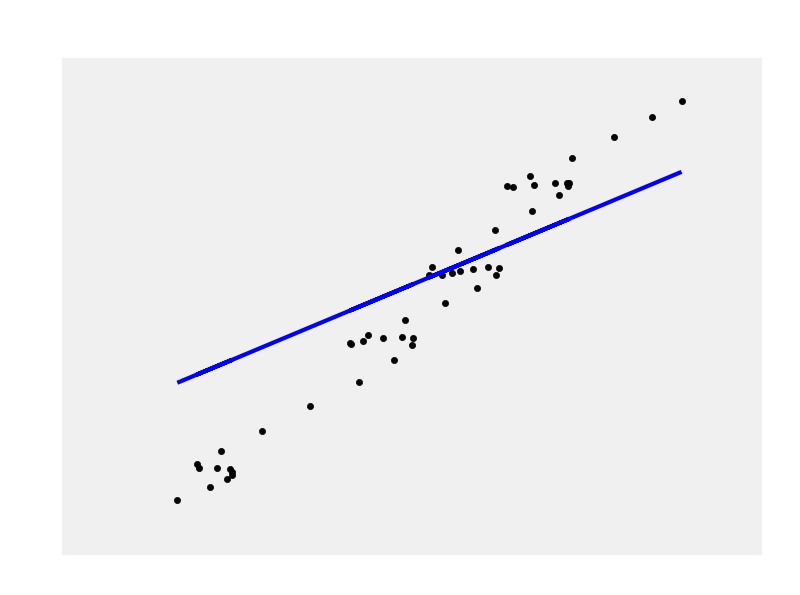
****

**The above graph shows the housing index prices changes along with United states. The solid black line is the United states.**

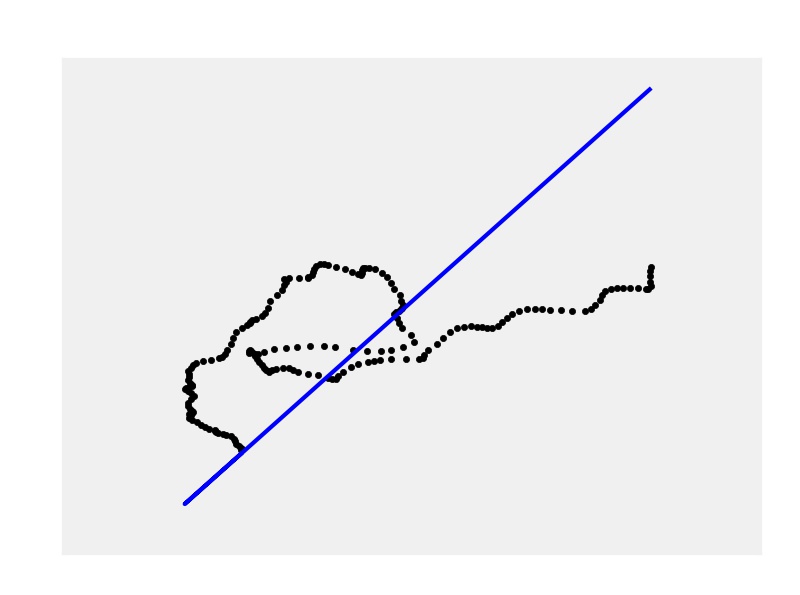
****

**The above graph is percentage change of all economic data including housing price index of United states. The individual graphs can be found in the bundle. These graph clearly shows ‘S&P500’ follows housing price index. Giving GDP as attribute would be under fitting and giving the other three attributes will be over fitting. So we used ‘S&P500’ as it suits or requirement to perform linear regression with the given data.**

**The resampling of data is done as the data needs to be consistent among parameters. The attribute values are resampled in between 1 and -1. The resulting regression results are as follows.**



After Training the set the following regression graph can be visualized with Virginia state as attribute and the blue line as predicting model.



The following regression is visualized after running the S&P500 to find the housing price index of united states. As we already saw in the graph earlier the attributes are not that easy to train to predict the values.

* Future developments : These project can be further extended to predict the housing prices for a particular region or any region. People sentiment is not taken into consideration these can be included taking them as attributes. The news of area or climate of particular areas can be used in deciding how much influence it has on the prices and can be used in better predicting the housing prices.

(Note: The source code is in project.py and we used each block of coed to do particular task that is explained in brief in the code. Instead of downloading data for each run we saved the data in pickle formats and used in later parts of code so that it is easier and faster to run the code. )

Thanks for giving an opportunity to learn and apply the concepts...!