URL to Colab:

https://colab.research.google.com/drive/1zMil6gVM1Qf8wS3H515PdYYgo5cI-YDP

Project Explanation:

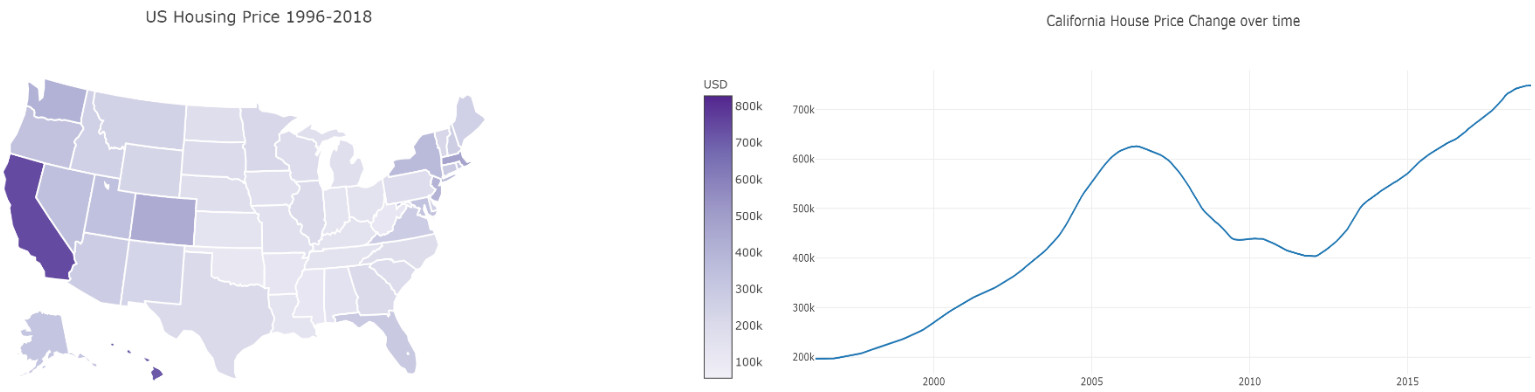
This Zillow House Price Project is to analyze US house price data from 1966 to 2018 and create a California house pricing prediction model (Time Series model) based on time and house size rank.

Data Processing

Since we need to conduct time series analysis, "Time" is the independent variable in the model. In this case, I utilized melt function in pandas to convert columns with time data to the new variable "Time".

Data Visualization

California/US House Pricing Change 1966 - 2018



From the above two graphs, we compare the house prices of different states in US and the house price change from 1966 to 2018 in California. Since the house price in California is extremely high compared with other states and increased dramatically from 200K USD to over 700K USD during the past 50 years, it is worthwhile to predict the house price in California.

Data Modeling

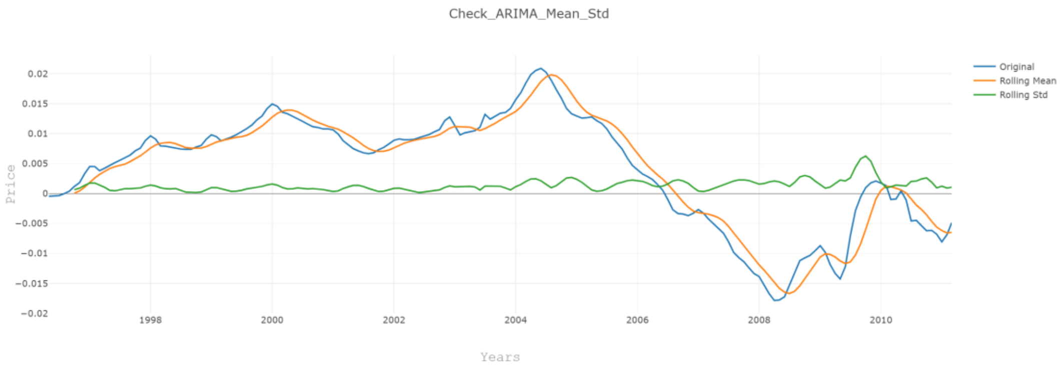
Time Series Model

- Check Stationarity

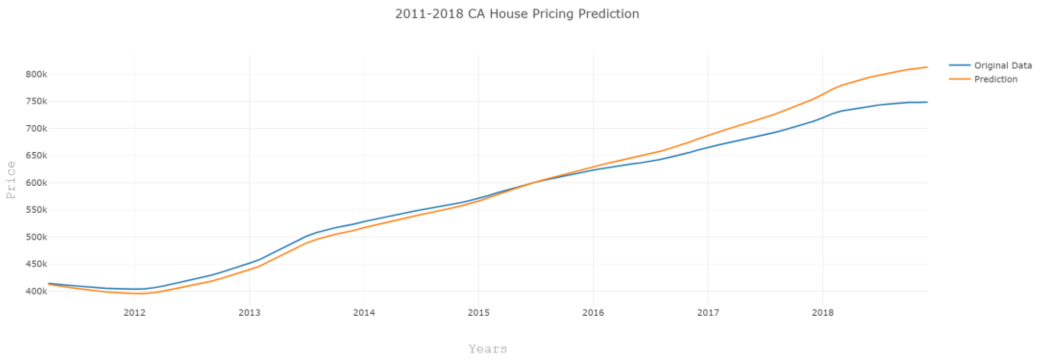
Three criteria are required to check the stationarity of current time series data. The first one for stationary is the constant mean. We fail in the constant mean test because the mean of housing price varies from 200K to 600K USD. The second one is the constant variance. In the current data set, the variance is around 0. The third one is that test statistic needs to be less than the critical value. In current model, test statistic is -2.06 and critical value is -2.58 at 90% confidence level. As a result, we are sure that our time series model is not stationary.

- Differencing Method to Eliminating Trend and Seasonality

In order to improve the stationarity of the model, we use differencing method to eliminate data trend and seasonality. In this technique, we take the difference of the observed house price data at a particular instant with that at the previous instant. From the below graph, we can see the three criteria of stationarity are all met.



- ARIMA Model to Forecast Price



With ARIMA model created, we are able to predict the California house price. From the above graph, the average house price in California 2018 is around 800K USD, which is 50K USD higher than the actual average price.

Model Comparison Conclusion:

I also created another Machine Learning Model – XGBoost Model by utilizing time series data and Size Rank data. The Root Mean Squared Error (RSME) of ARIMA Time Series Model is 24178.3941, while the RMSE of XGBoost Model is 25204.9083, which is slightly higher than the ARIMA Model. I will choose the model with smaller RSME, which is ARIMA Time Series Model. However, the time series model can be improved by adding house features to the data set, including house location, size, type and so on, so that we are able to predict the house price of specific individual house.

Recommendations to Management:

Based on the ARIMA Time Series forecasting model, we predict that in 2018, the average house price in California is around 800K USD, and there is a trend that price will rise steadily in the next few years.