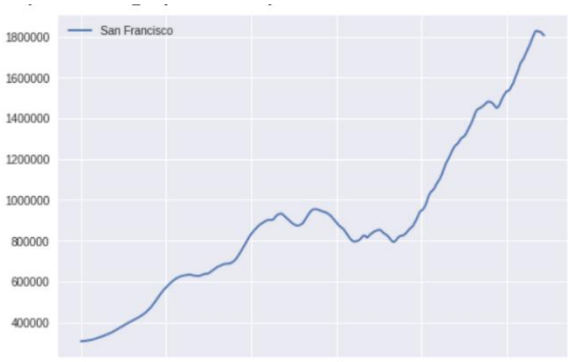
Time series model:
https://colab.research.google.com/drive/1B2y2_wwIThMOhnG1v9Ir4KPozPzVJJGc
Machine learning model:
https://colab.research.google.com/drive/1HOImytkTurMfISk71AU48y4zfZCrbfmE

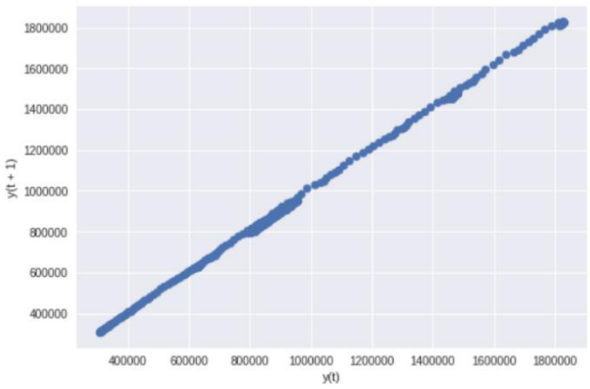
## Overview

In this assignment, I explore the house prices in San Francisco from 1996 to 2018. I use a time series model and a traditional machine learning model to make predictions. Specifically, I use the data from 1996 to 2016 to fit the models and predict the house prices in 2017 and 2018 with these two models.

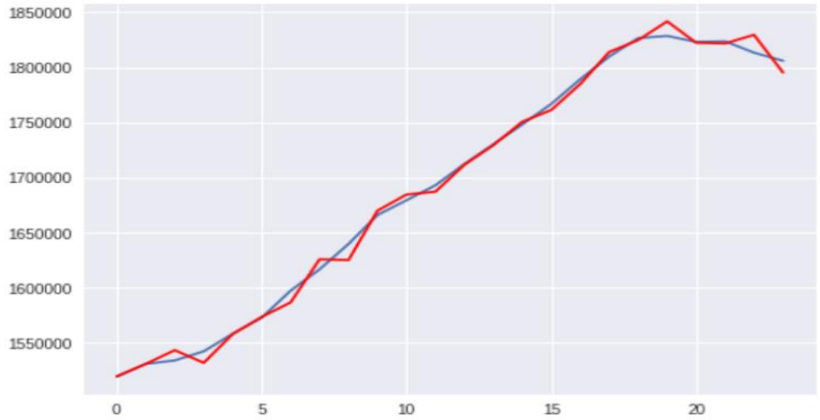
## Technique

My first model is an AR time series model. In data cleaning, I just keep the time and house prices in San Francisco and set the time as an index. To test whether the data is correlated with itself, I create a lag plot.



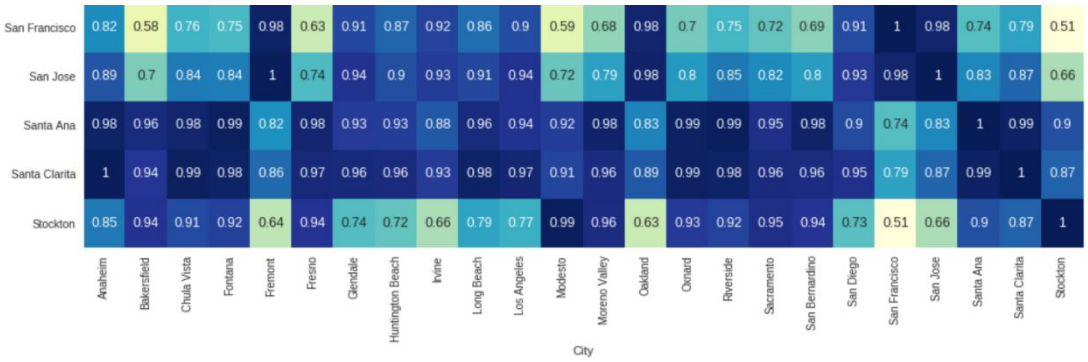




Comparing two plots, I can clearly see that there is a perfect correlation between the data of current month and the data of the last month, which also means an AR model is reasonable for the data. Then I split the whole data into two parts: the first part is the data before 2017. I use this part as train data. Another part is the data for 2017 and 2018, which are used as test data. After selection, I find a 15-lag data is best for this data and the final MSE is 54656832.853. The number looks pretty large, but that might because each house price is a large amount. To see how the model works, I visualize the comparison between predicted values and actual data.

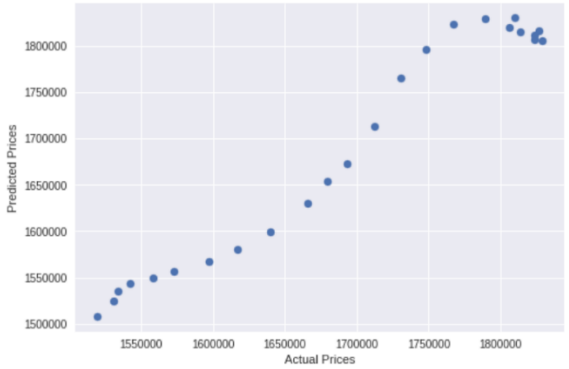


From this plot, I think this model makes a totally fine prediction because the overall trend is correctly predicted. Thus, this model is useful to some extent.

My second model is a traditional machine learning model. In this model, I select the house prices of 24 large cities in California from 1996 to 2016 as train data to predict the house prices in 2017 and 2018. First, I keep time, city name and prices data for all the 24 cities and San Francisco. Then, I group the data by city name and split the data into train data and test data. After transposing the data, I create a heatmap to visualize the correlation between San Francisco's house prices and each of the 24 big cities' prices in California.




From the heatmap, I find all these 24 cities' prices highly relates to San Francisco's house price, so I include all 24 cities into feature data. Then, I split the whole data into train and test and use a linear regression model to fit the data. The comparison between test data and predictions are shown in the below plot and the model score is 0.9416.




## Conclusion

Housing price data have obvious autocorrelation, and AR model can fit the data to a certain extent. The accuracy of prediction can be improved by improving the model or changing the form of data. For the machine learning model, the 24 largest cities in California show a high correlation with San Francisco home prices, and thus I get a high accuracy in this prediction.