Decision Tree and Time Series Models for Predicting Housing Prices in the Alameda County

Pardeep Kumar

Capstone Project Report

Introduction to Data Sciene

Springboard Online

SEPTEMBER 2018

ABSTRACT

The main objective of this project was to make relatively accurate housing price predictions for the Alameda County (CA) using simple time series and decision tree based methods. The input data was divided into the training and the test set. Decision tree methods: bagging and random forest were the basis for developing the prediction models using 11 features, housing prices of the neighbouring counties, population, time, consumer price index, monthly inventory and the turnover rate. Relative importance of each feature was key in improving the dynamic regression model developed at the end. For time series models, the median monthly housing prices and the error terms were the only variable used to develop a prediction model. Time series model were relatively accurate than the decision tree models. Finally using the insight from both regression and time series methods a dynamic regression model was developed with improved accuracy. R-programming language was used as the base package for the data and the model analysis.

CONTENTS

[ABSTRACT ii](#_Toc526353158)

[CONTENTS iii](#_Toc526353159)

[1 Introduction 1](#_Toc526353160)

[1.1 Objective 1](#_Toc526353161)

[1.2 Scope of Study 1](#_Toc526353162)

[2 Exploratory Data Analysis 2](#_Toc526353163)

[2.1 Introduction 2](#_Toc526353164)

[2.2 Housing Price Data 3](#_Toc526353165)

[2.2.1 Time Series Plots 3](#_Toc526353166)

[2.2.2 Correlation Plots 4](#_Toc526353167)

[2.3 Other Predictors For Decision Tree Methods 5](#_Toc526353168)

[2.3.1 Population 5](#_Toc526353169)

[2.3.2 Housing Inventory 6](#_Toc526353170)

[2.3.3 Consumer Price Index 7](#_Toc526353171)

[2.3.4 Monthly Turnover 8](#_Toc526353172)

[3 Predictive Modeling Using Decision Trees 9](#_Toc526353173)

[3.1 Introduction 9](#_Toc526353174)

[3.1.1 Bagging 9](#_Toc526353175)

[3.1.2 Random Forest 9](#_Toc526353176)

[3.2 Bagging for Predictive Analysis 10](#_Toc526353177)

[3.3 Random Forest for Predictive Analysis 12](#_Toc526353178)

[3.4 Comparison of Bagging and Random Forest Models 15](#_Toc526353179)

[4 Forecasting with Time Series Models 17](#_Toc526353180)

[4.1 Time Series Data 17](#_Toc526353181)

[4.1.1 Time Series Decomposition 18](#_Toc526353182)

[4.2 Forecasting with Exponential Smoothing 20](#_Toc526353183)

[4.2.1 Holt’s Trend Methods 20](#_Toc526353184)

[4.3 ARIMA Model 22](#_Toc526353185)

[4.3.1 Stationarity 22](#_Toc526353186)

[4.3.2 Non-Seasonal ARIMA Model 26](#_Toc526353187)

[4.4 ARIMA(0,2,3) 26](#_Toc526353188)

[4.5 ARIMA(5,2,5) 28](#_Toc526353189)

[4.6 Breakpoint Analysis 30](#_Toc526353190)

[4.6.1 Intercept Only Model 30](#_Toc526353191)

[4.6.2 Linear Model 32](#_Toc526353192)

[4.7 ARIMA Model with 3 partitions 34](#_Toc526353193)

[4.8 ARIMA Model with 6 partitions 37](#_Toc526353194)

[5 Dynamic Regression Model 39](#_Toc526353195)

[5.1 ARIMA Model with Regression 39](#_Toc526353196)

[5.2 ARIMA Regression Model with 3 Partitions 41](#_Toc526353197)

[6 Summary 43](#_Toc526353198)

[6.1 Conclusion 43](#_Toc526353199)

[6.2 Future Research 44](#_Toc526353200)

[Bibliography 45](#_Toc526353201)

# Introduction

Since the recovery of US real estate market after 2008 financial crisis, the housing prices in San Francisco Bay Area have seen a steep rise. The housing prices has an indirect impact on the economic policies, as higher housing prices and sales implies increase in wealth, which leads firms to raise markups(Stroebel, 2015). Higher real estate prices are often related to low supply and strong economic growth. Numerous factors can weigh heavily on the demand and thus influencing the real estate cost in an area. An accurate housing prediction can help everyone involved in the real estate market, from sellers/ buyers to insurers, in better decision making.

## Objective

The objective of this project was to gain insight into the housing price trend of Alameda County and be able to forecast the future prices with fair accuracy. Six counties in closest proximity to Alameda County were selected for this study (including San Francisco County). Median monthly housing price data of the counties was obtained from the Zillow Research Website (Zillow, 2018). Exploratory data analysis on the time series revealed a high correlation between the housing prices in the 7 counties. Insights from both decision tree and time series analysis were used to develop a final dynamic regression model. Accuracy of all models was tested on the test set sample using the mean squared errors. Statistical programming language R was used for this study.

## Scope of Study

This report is divided into 5 chapters as follows:

Chapter 1. Introduction

Chapter 2. Exploratory data analysis

Chapter 3. Predictive modeling using decision trees

Chapter 4. Forecasting using time series models

Chapter 5. Dynamic regression models

Chapter 5. Summary

# Exploratory Data Analysis

## Introduction

Figure 2.1 shows the map of San Francisco Bay Area (Bay Area Census, 2018). Seven counties were selected for this study: Alameda, Contra Costa, Marin, San Francisco, San Mateo, Santa Clara and Solano. Selection of counties was based on their relative proximity to the Alameda County. The median-monthly housing price data for each county was obtained from the Zillow research website (Zillow, 2018) which spanned from April 1996 to July 2018. The available data was seasonally adjusted housing prices of the single-family homes, condominiums and co-operative houses combined. Any data (housing prices or other predictors) collected before January 2018 was considered training set and test set comprised the monthly data sets for year 2018.



Figure 2.1 Map of the San Francisco Bay Area.

## Housing Price Data

Median-monthly housing prices of the neighboring counties were included as features for decision tree models. The time series are discussed in the following sections.

### Time Series Plots

Figure 2.2 shows the time series plot of the selected counties. Few characteristics of the plotted time series were as follows:

1. Housing prices of all counties had similar trends. The housing prices increased linearly from 1996 to about 2006. During this time, housing prices in the Marin County were highest and the housing prices in San Francisco and San Mateo County were almost the same.
2. Due to the real estate recession, the housing prices fell from 2006 and did not properly recover until mid-2011.
3. 2012 onwards, the monthly rate of housing price increase was higher in the San Francisco and San Mateo County but remained almost same as before the period of real estate crisis, for the Solano and Contra Costa County.

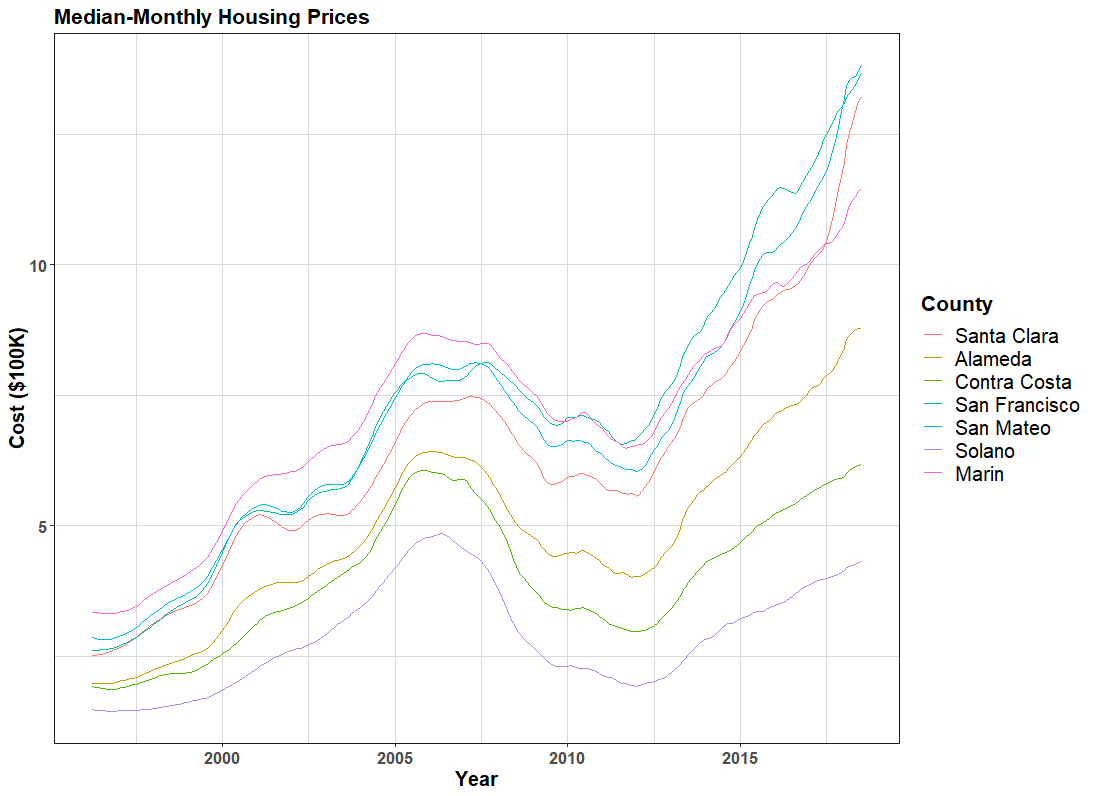


Figure 2.2 Time series plot of median-monthly hosing prices (in $100,000).

### Correlation Plots

Figure 2.3 shows the correlation plots and Table 2.1 lists the correlation matrix entries for the seven counties. As it was evident from the close trend of time series plots, the housing price of Alameda County were highly correlated to the other six counties, highest correlation with Marin County (0.99) and lowest with Solano (0.86).

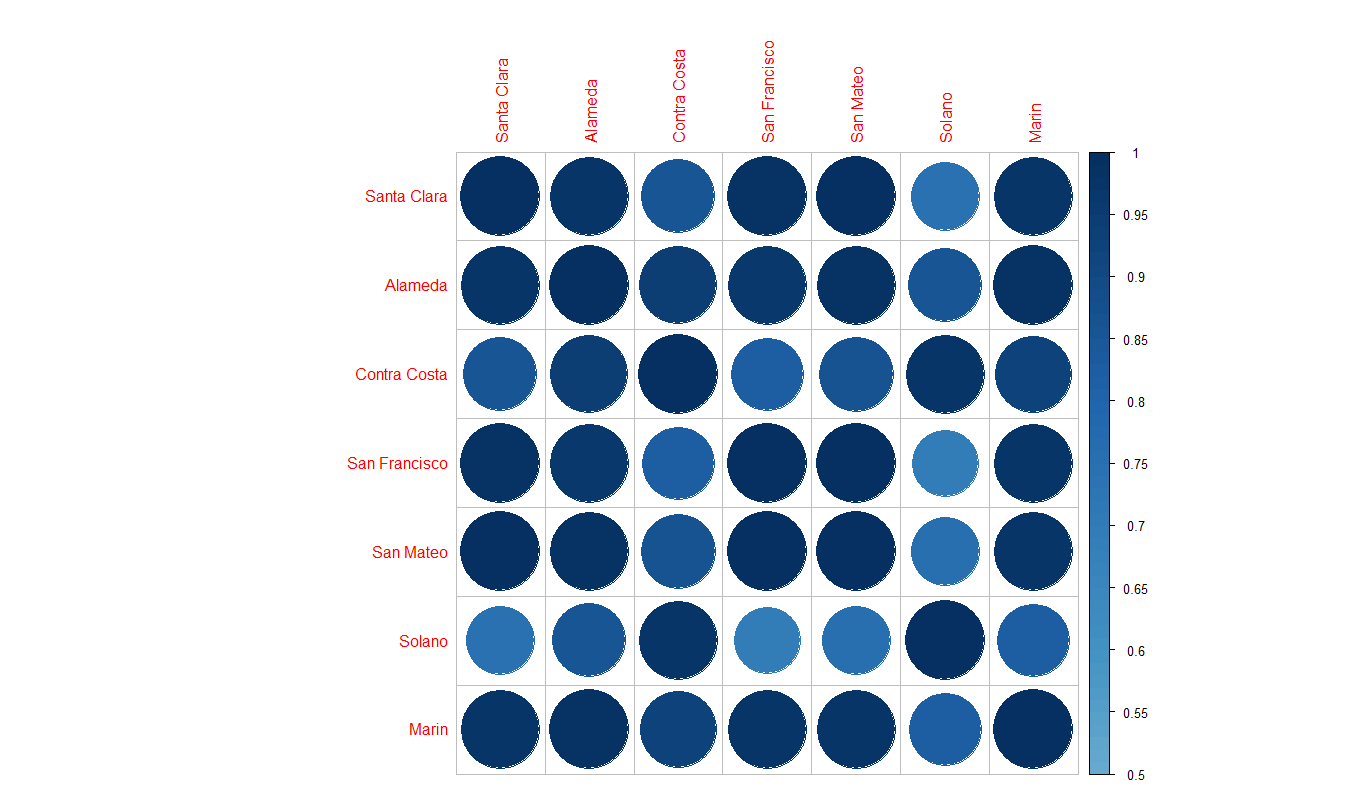


Figure 2.3 Correlation plot of the seven counties.

Table 2.1 Coefficient matrix of the seven counties (R output).

|  |
| --- |
| Santa Clara Alameda Contra Costa San Francisco San Mateo Solano Marin  Santa Clara 1.0000000 0.9776199 0.8587038 0.9888947 0.9978823 0.7417814 0.9709462  Alameda 0.9776199 1.0000000 0.9428272 0.9659201 0.9823701 0.8556051 0.9893881  Contra Costa 0.8587038 0.9428272 1.0000000 0.8286117 0.8690276 0.9776691 0.9210560  San Francisco 0.9888947 0.9659201 0.8286117 1.0000000 0.9915703 0.6969285 0.9708248  San Mateo 0.9978823 0.9823701 0.8690276 0.9915703 1.0000000 0.7542549 0.9756265  Solano 0.7417814 0.8556051 0.9776691 0.6969285 0.7542549 1.0000000 0.8241027  Marin 0.9709462 0.9893881 0.9210560 0.9708248 0.9756265 0.8241027 1.0000000 |

## Other Predictors For Decision Tree Methods

For the decision tree models, 5 more predictors (other than housing prices of other counties) were added to improve the model accuracy. Data for few predictors was available for only last eight years, so the housing data used for the decision tree methods was reduced to time series from January 2010 to June 2018. Data from January 2010 to December 2017 was assigned as the training set data total 96 observations per feature. Test set comprised of 6 observations (per feature) from January 2018 to June 2018. Time was one of the predictors added to capture any linear trends in the data. Other 4 predictors are discussed in the following sections.

### Population

Annual population estimates were obtained from State of California’s Department of Finance website (Department of Finance, 2018). The annual trend of estimated population was almost linear from 2010 to 2018 as plotted in Figure 2.4. Since the housing prices were observed on the monthly basis, monthly population estimates were obtained by linearly interpolating between the annual estimated.

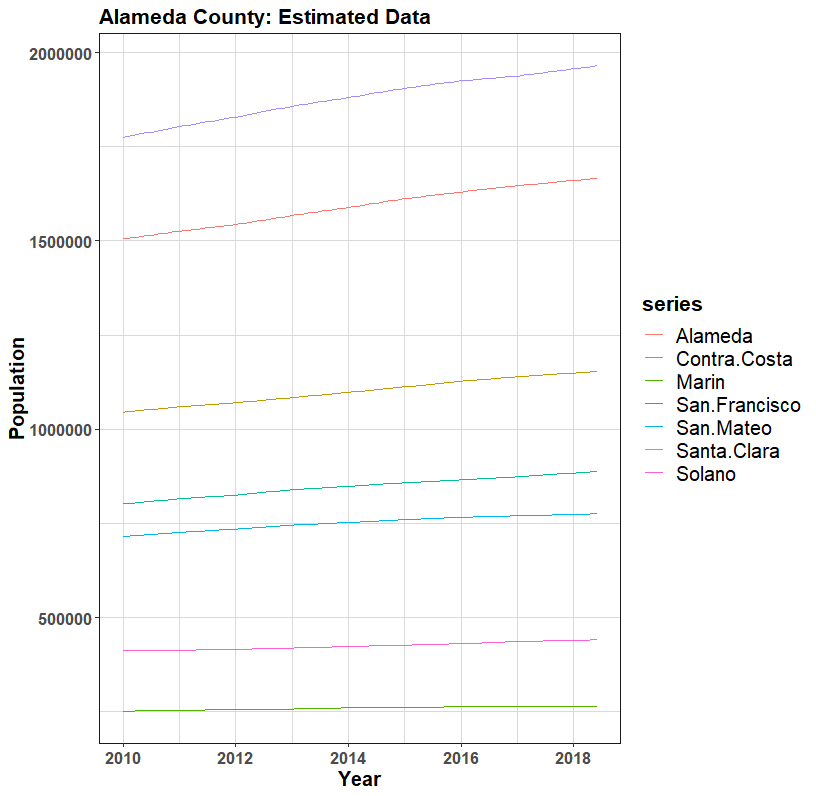


Figure 2.4 Population data from January 2010 to June 2018.

### Housing Inventory

Housing Inventory is the number of houses available for sale in a region. Figure 2.5 plots the monthly housing inventory of the 7 counties (Zillow, 2018). The inventory data for the San Francisco and Solano County was not available. Trend of housing inventory was inverse of the housing price time series from 2010 to 2018. After the real-estate financial crisis, there was sudden increase in the housing demand due to lower prices, which lead to sharp drop in the monthly inventory. Similar to the housing price data, the Alameda County inventory seems to follow the mean trend of all other counties.

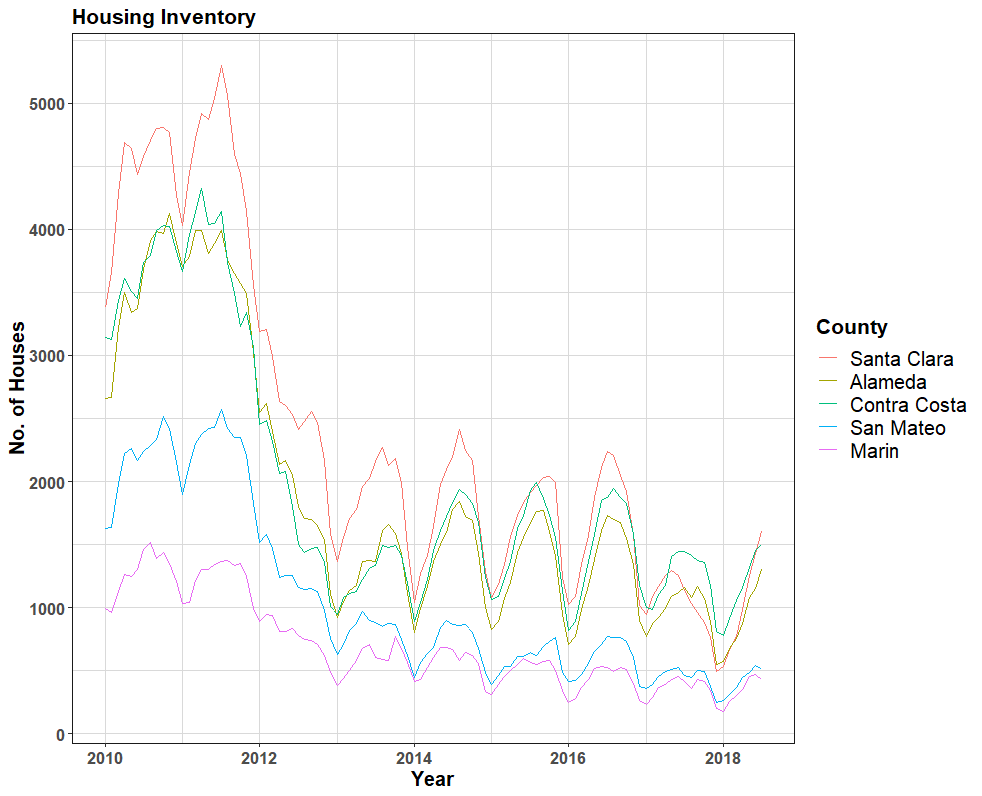


Figure 2.5 Monthly housing inventory from January 2010 to June 2018.

### Consumer Price Index

Consumer Price Index (CPI), as defined by the Bureau of Labor Statistics, is the measure of the average change over time in the prices paid by the urban consumer for a market basket of consumer goods and services (Bureau of Labor Statistics, 2018). Bureau of Labor Statistics provides this information on national and regional level. CPI thus measures the cost of living of an average individual in a region. CPI for the geographical region of San Francisco-Oakland-Hayward (All Urban Consumers) was obtained from the Federal Reserve Bank of St. Louis webpage (Federal Reserve Bank of St. Louis, 2018). Data was available on bi-monthly basis and the missing month values were obtained by averaging the first neighbor data. Figure 2.6 shows the seasonally adjusted, time series plot of the CPI for all consumer goods.

Annual inflation rate is computed as the annual rate of change in the CPI index and is an important factor in controlling the mortgage interest rates for the real estate. Thus, CPI has impact on not just the housing price but also the housing inventory. High interest rates reduce the buyers’ willingness to invest in the real estate.

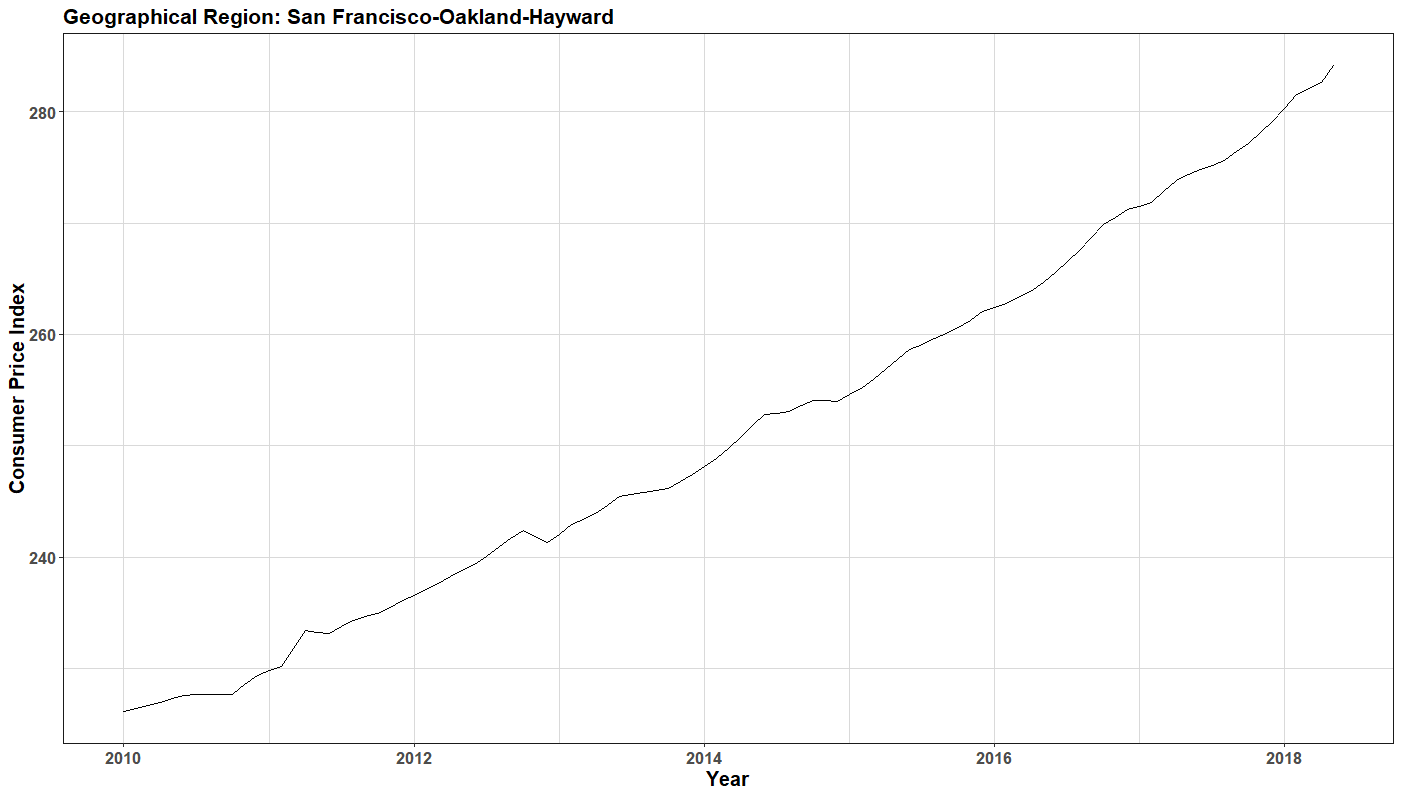


Figure 2.6 Consumer Price Index of San Francisco-Oakland-Hayward.

### Monthly Turnover

Monthly turnover rate in this study was defined as the percentage of houses sold in the previous month relative to total number of houses in the county. Figure 2.7 shows the monthly turnover time series (Zillow, 2018). Turnover rate of the Contra Costa County was higher relative to other counties. This could be due to high inventory and relatively less demand. The turnover rate is considered a measure of buyer/seller confidence and actual behind-the-scene strength of the real estate market. The high housing prices and demand in a region may seem like economy is booming but if the turnover rate is below the national average, it indicates a weak real-estate market. In the plot below, data for the Solano County was no available.

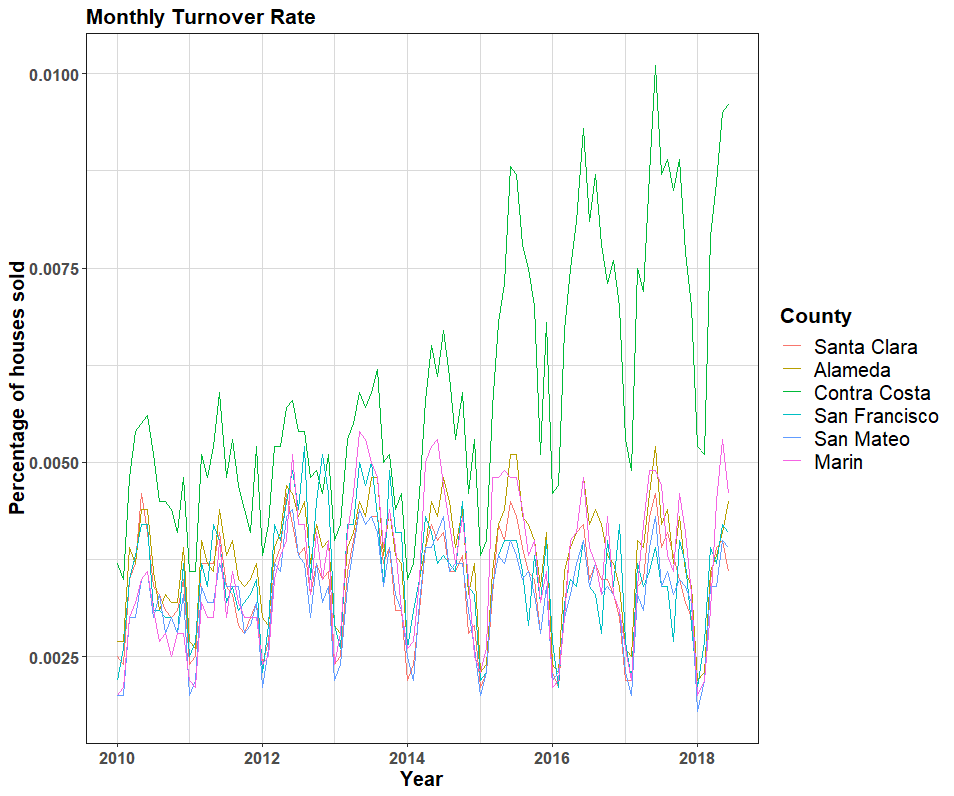


Figure 2.7 Monthly Turnover Rate.

# Predictive Modeling Using Decision Trees

## Introduction

Decision tree methods have applications in both regression and classification analysis. Decision tree models are hierarchical methods that partition the space until the algorithm reaches a certain threshold. For this study, two decision tree methods, bagging and random forest, were used to train the predictive models.

Bagging and random forest make multiple passes on the data and results are aggregated from all the trees. Averaging reduces the issue of overfitting by canceling the uncommon features of data, yet producing a flexible fitting model. This often results in break-off with the bias-variance tradeoff. Before discussing these methods in details, we first give a brief introduction to bootstrap method.

Bootstrapping is used where the access to data is limited or sample size is too small to get reliable output. Bootstrap method randomly selects the sample data points with replacement to create a new subset of same size as the parent sample. Bootstrap method induces distortion in the data, which is fed as a new observation set to the machine-learning algorithm. (James, Witten, Hastie, & Tibshirani, 2014).

### Bagging

Bagging uses bootstrapped random data samples for growing the decision trees. All predictors are considered for the branch splitting and the terminal node observations are collected over multiple trees. The number of trees to grow are usually user defined based on how well the model performs on the cross validation or test data. Finally the terminal node observations from all trees are aggregated to produce the final output. Averaging over multiple trees reduces overfitting and the variance error. The one drawback of bagging is if few predictive features outperform the other, these features will be included in most of the trees. Thus the final averaging will be on correlated trees which does not reduce variance effectively despite using more trees. Random forests technique is used to improve on this limitation.

### Random Forest

Often times few predictors dominate at the branch split, because on average they perform better than their competitors. These “weak” predictors can be useful for local data feature but suppressed and rarely used. This results in most of the trees being correlated. Random forest adds one more step in the bagging algorithm. At the time of branch splitting, random forests only uses a subset of the predictor set. This is done over multiple trees, thus including almost all the features in the final results. This limits the error in bias and error in variance.

## Bagging for Predictive Analysis

Final output of the bagging model averaging over 500 trees is shown in Table 3.1. The mean of squared residuals is also called the Out-of-Bag (OOB) error. OOB error is computed by using only 2/3rd of the training data for training and other 1/3rd for validating the tree. Figure 3.1 plots the OOB against the number of trees. As the number of trees increased, the OOB error reduced almost exponentially. The rate of error reduction was really high in the beginning (from 1 to 40 trees). After 200 trees the magnitude of OOB error was almost constant, 0.13%.

Table 3.1 also lists the predictor importance summary and Figure 3.2 shows the corresponding plot. The term **%IncMSE,** is based upon the mean drop in prediction accuracy on the out-of-bag samples when the given predictor is not included in the model. **IncNodePurity** is the average decrease in the node impurity when split is over that particular variable. Among the 6 predictor time series, San Francisco County housing prices had relatively higher influence on Alameda County housing prices. From the set of other 5 predictors, time had relatively more influence followed by population.

Table 3.2 compares the predictions and the test data. These magnitude of errors increased into the future months, indicating that bagging was more reliable for short term predictions (next one month or so). The test MSE was 29.51%.

Table 3.1 Bagging Model (R-output).

|  |
| --- |
| Bagging:: Model Summary |
| |  | | --- | | Call:  randomForest(formula = Alameda ~ .,data = AlamedaData,mtry = 11,importance = T,subset = train)  Type of random forest: regression  Number of trees: 500  No. of variables tried at each split: 11  Mean of squared residuals: 0.001347724  % Var explained: 99.93 | |
| Predictor Importance Summary |
| %IncMSE IncNodePurity  SantaClara 9.477318 20.95054501  ContraCosta 9.349447 20.77210016  SanFrancisco 13.076102 35.51751667  SanMateo 8.564825 15.81219437  Solano 8.072550 12.86226551  Marin 9.331034 15.11414058  time 10.058072 20.52681368  Pop 9.848822 18.35977401  CPI 9.430175 19.16544540  Inven 5.895297 0.09179841  TurnOver 4.591539 0.05488782 |

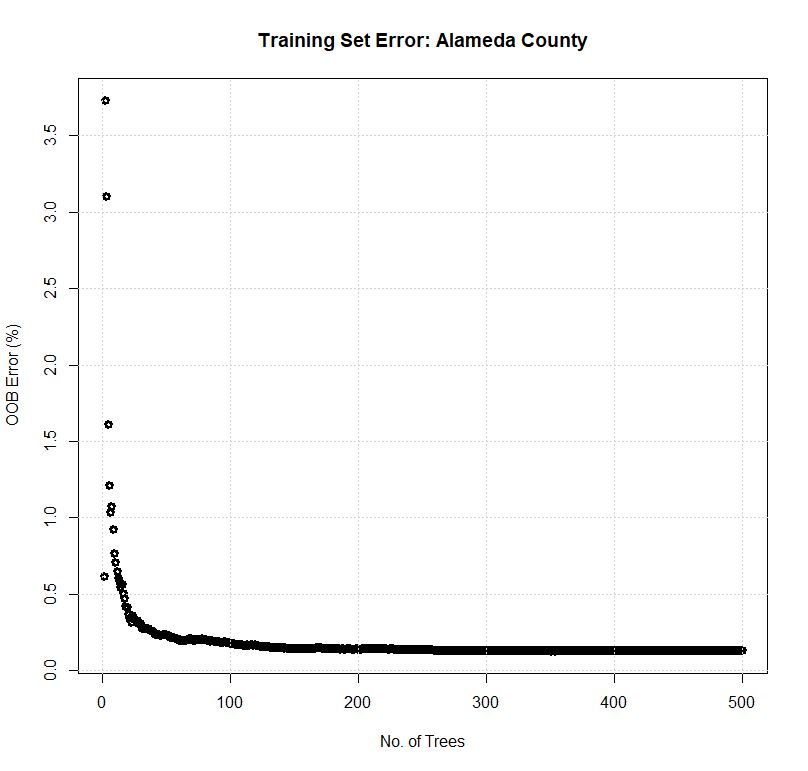


Figure 3.1 Bagging OOB error v/s no. of trees.

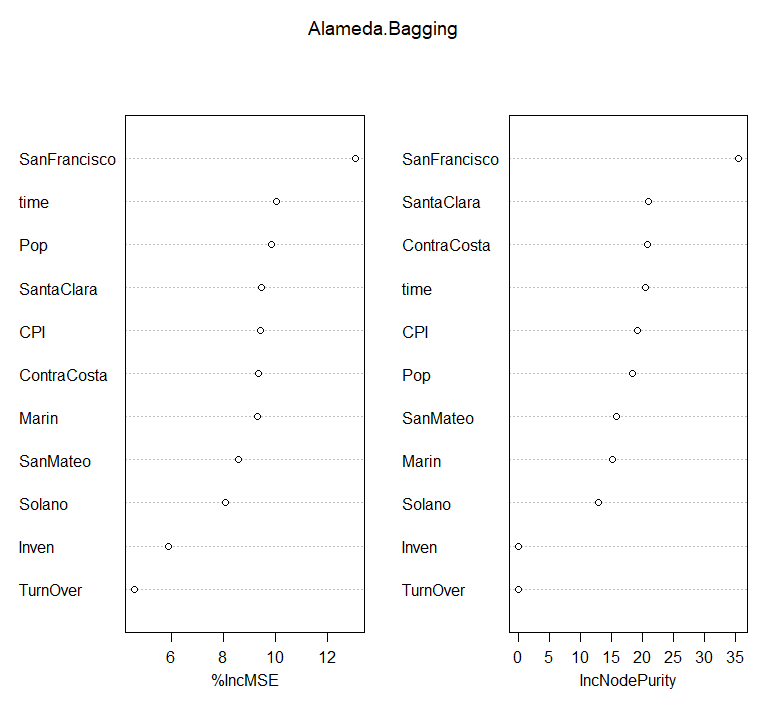


Figure 3.2 Predictor importance measures for bagging.

Table 3.2 Bagging Predictions and Test MSE

|  |
| --- |
| Test data, prediction and residuals |
| Test Data Model Predictions Residuals Residual^2  2018.01 8.423 8.137822 0.2851780 0.08132649  2018.02 8.569 8.135845 0.4331548 0.18762308  2018.03 8.654 8.132978 0.5210222 0.27146417  2018.04 8.698 8.126163 0.5718365 0.32699698  2018.05 8.755 8.100485 0.6545146 0.42838941  2018.06 8.786 8.096860 0.6891396 0.47491339 |
| Mean Test Error: **29.51%** |

## Random Forest for Predictive Analysis

Main difference between random forest and bagging is decorrelation of several trees on random bootstrapped samples. The idea is to reduce the variance by averaging the trees and eventually avoid overfitting by reducing the correlation between the random trees. Random forest uses a subset of predictors (without replacement) and thus multi-collinearity is not a major issue as all subsets are randomly selected. Table 3.3 gives the model summary of random forest. At each split a subset of 3 randomly selected predictors was used. The predictor importance summary suggested San Francisco County among the housing price predictors and population of the Alameda County from other 5 predictors to be relatively important, same as the bagging model (Table 3.1).

The OOB error of random forest and bagging was almost same (0.13%), Figure 4.3 compares the two. Random forest OOB error reduced faster than bagging during the initial increment in number of trees. After about 20 trees the random forest OOB error was half the magnitude of bagging OOB error. Thus on same training data set random forest performed better than bagging initially.

Table 3.4 compares the test data with the predictions from the random forest model. The mean test MSE of random forest model was 27.36% about 2% less than bagging model. Like bagging, random forest model predictions were more accurate for short term and the residuals increased further into the future.

Table 3.3 Random Forest Model (R-output).

|  |
| --- |
| Random Forest:: Model Summary |
| |  | | --- | | Call:  randomForest(formula = Alameda ~ ., data = AlamedaData, importance = T, subset = train)  Type of random forest: regression  Number of trees: 500  No. of variables tried at each split: 3  Mean of squared residuals: 0.001307916  % Var explained: 99.93 | |
| Predictor Importance Summary |
| %IncMSE IncNodePurity  SantaClara 11.316503 24.59108819  ContraCosta 9.639555 15.18653132  SanFrancisco 12.072966 26.77084846  SanMateo 10.171515 19.95339782  Solano 9.473884 15.83409830  Marin 10.251450 20.00878551  time 9.970050 19.15122676  Pop 9.896872 19.14841774  CPI 9.549034 18.73496533  Inven 3.625210 0.08620661  TurnOver 4.783974 0.04576179 |

Table 3.4 Random Forest Predictions and Test MSE

|  |
| --- |
| Test data, prediction and residuals |
| Test Data Model Predictions Residuals Residual^2  2018.01 8.423 8.146811 0.2761892 0.07628047  2018.02 8.569 8.146011 0.4229892 0.17891986  2018.03 8.654 8.144488 0.5095120 0.25960248  2018.04 8.698 8.138035 0.5599650 0.31356080  2018.05 8.755 8.133280 0.6217195 0.38653518  2018.06 8.786 8.132836 0.6531635 0.42662260 |
| Mean Test Error: **27.36%** |

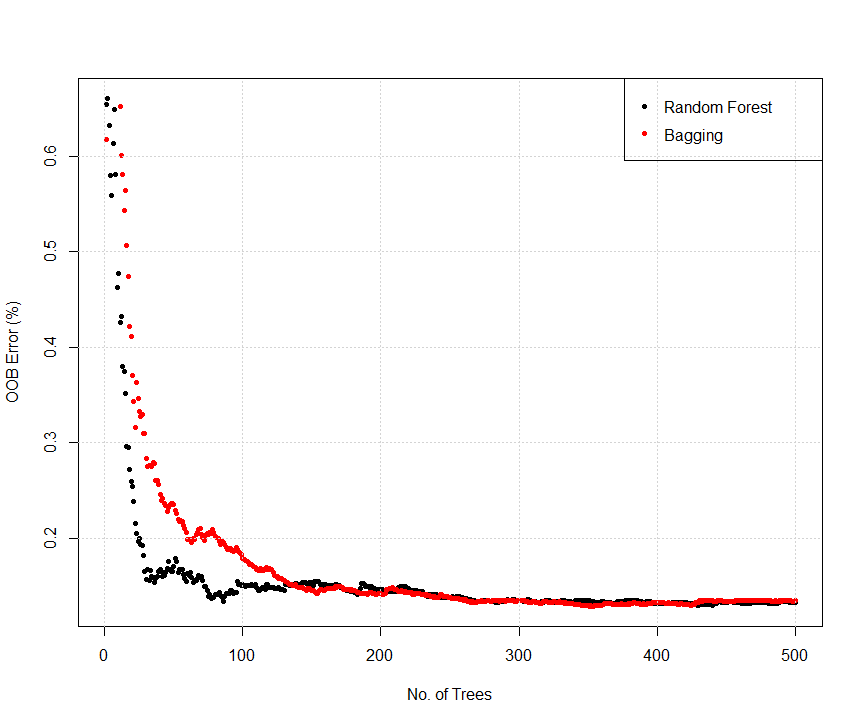


Figure 3.3 Comparison of Random Forest and Bagging OOB Error.

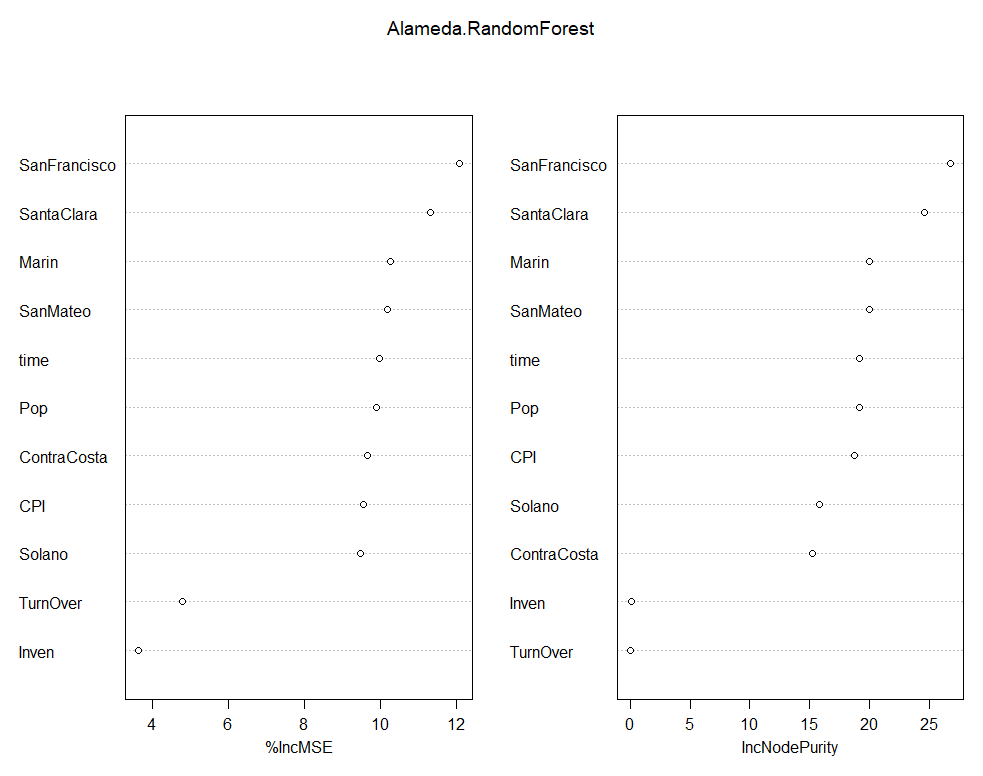


Figure 3.4 Predictor importance measures for random forest.

## Comparison of Bagging and Random Forest Models

Bagging and random forest models provided a good insight into the relative importance of predictor variables. San Francisco County had highest importance for Alameda County housing price models. Inventory and turnover percentage had least influence in the decision tree models. Performance of both models on training set was almost as the number of tree increased above 200, with OOB error of 0.13% after 500 trees. Both models suggested that San Francisco County was important predictor for Alameda County housing prices. The test MSE of random forest model (27.36%) was about 2% lower than bagging test MSE (29.51%). Thus the number of predictors chosen at the split did effect the test MSE. Figure 3.5 show the test and OOB errors plotted against the predictor subset size. Note that OOB error was scaled 100 times to compare its pattern with test MSE. The OOB error reduced significantly as the number of predictors increased from 1 to 2. At predictor subset size of 3, both OOB and test errors were at the minimum magnitude. After subset size of 7, the errors increased linearly. The last point on both curves corresponds to the bagging model with all 11 predictors. Figure 3.6 shows the squared-residual plots for both models plotted against each month. Random Forest performed slightly better than bagging in the months of May and June.

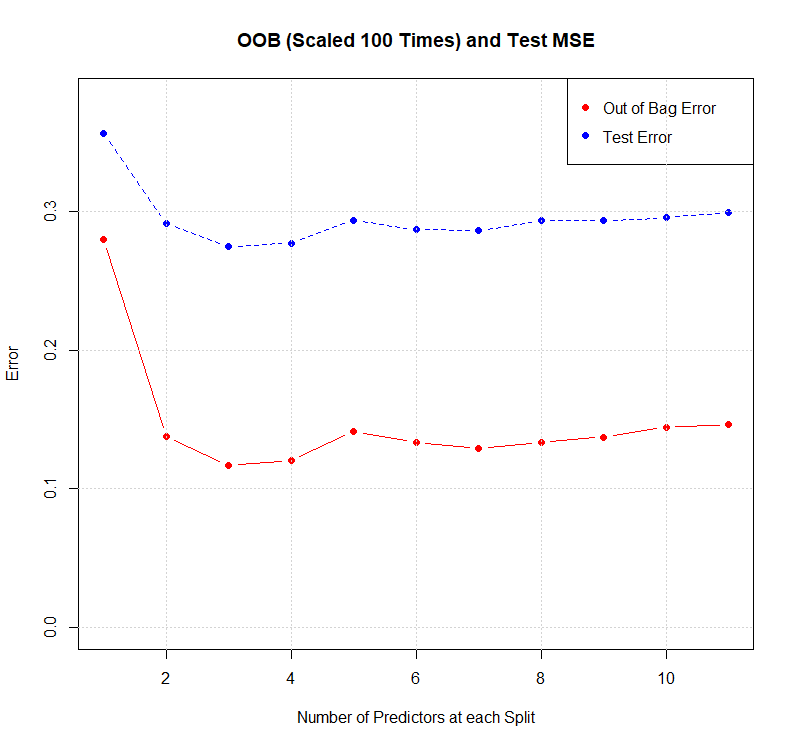


Figure 3.5 OOB comparison of Bagging and Random Forest.

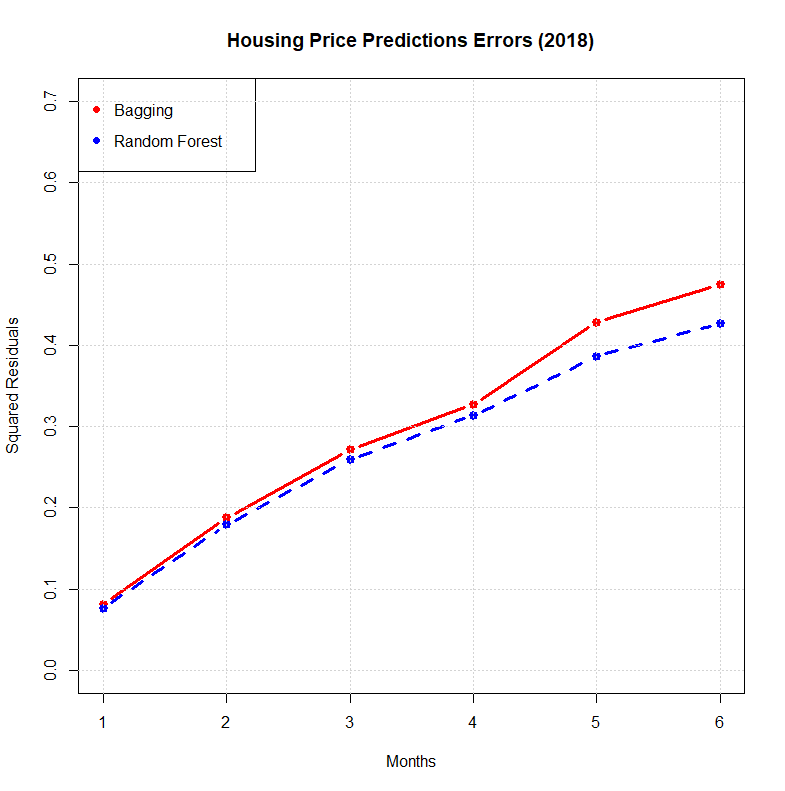


Figure 3.5 Squared residuals of housing Price prediction errors.

# Forecasting with Time Series Models

Sequential observation of the quantity of interest over time is defined as a time series. Observations can be recorded (or measured) on hourly, daily, weekly monthly, quarterly or annually basis. The basic idea behind the time series forecasting is to estimate how the past trend continues into the future. Most of the models used in this chapter can be mathematically written as,

(4.1)

The time series was only dependent on the past values and no external variable, which may affect the system. The *Error Term* allowed for the random variation and effects not included in the model. The training set for time series models was housing price data of the Alameda County from April 1996 to December 2017. Monthly median housing prices from June 2018 to July 2018 were used as the test set. Performance of each model on the test set was compared using the mean-squared errors (MSE).

Time series models in R often give the three error measures, Akaike’s Information Criterion (AIC), Akaike’s Information Criterion Correct (AICc) and Bayesian Information Criterion (BIC). AICc and BIC are modifications over AIC, and all three measures depends on the number of predictors used in the model (Vrieze, 2012). The model is penalized relative to the number of predictors used favoring simple models over the complex ones, as long as they are reasonably accurate. AICc was used as the error criterion for this study. Focus was on change in the value of AICc rather than the actual magnitude. Most of the error values were negative and the model with most negative values was chosen as long as it satisfied the other selection criteria of insignificant auto correlation between the residuals.

## Time Series Data

Time series plots for the housing price predictions were discussed in Chapter 2. Figure 4.1 plots the Alameda County housing price time series. Few features noted from the plot were as follows,

* The time series plot had two regions of positive trend, from 1996-2007 and 2012-2018.
* There was a negative trend from 2007- 2012, indicator of the real-estate financial crisis.
* Just looking at data it was hard to conclude if any seasonality was present in the time series. Seasonality was tested separately as it could affect the model selection (seasonal vs non-seasonal models).

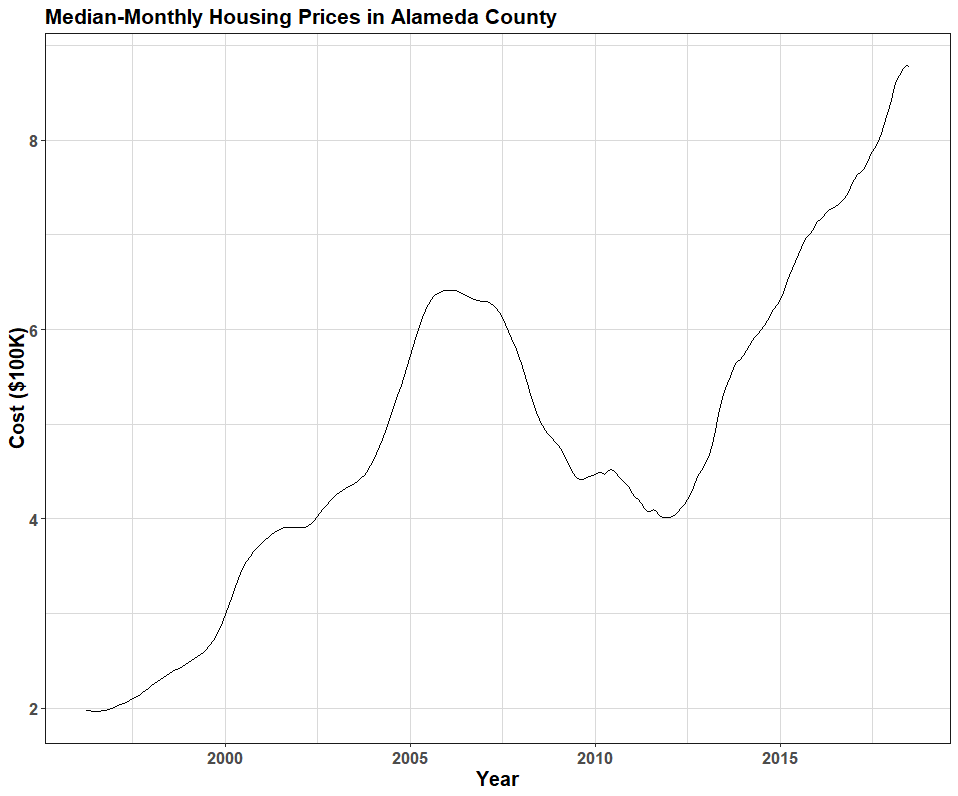


Figure 4.1 Median-monthly housing price (in $100,000).

### Time Series Decomposition

Pattern exhibited by the time series was crucial to understand the type of forecasting model to use. The time series was decomposed into three components: trend, seasonal and reminder, usin additive decomposition.

(4.2)

Depending on which component dominated the time series, forecasting methods were selected accordingly. Figure 4.2 show the decomposition of time series using STL method in R (Cleveland, Cleveland, McRae, & Terpening, 1990). A time series decomposition can be additive (equation 4.2) or multiplicative,

(4.3)

Multiplicative decomposition is used when the seasonal component is stronger than trend component, while additive decomposition is more appropriate for trend dominant time series. In order to justify using additive model, the strength of trend component in the time series was computed using trend strength measure (Wang , Smith, & Handyman, 2006). In case there is a strong trend, the seasonally adjusted data will have more variation than the remainder component but for weak trend data, two variations should be almost same. The trend strength measure varies from zero to one, with value closer to one suggesting stronger trend,

(4.4)

For the Alameda County time series,. supporting an additive model.

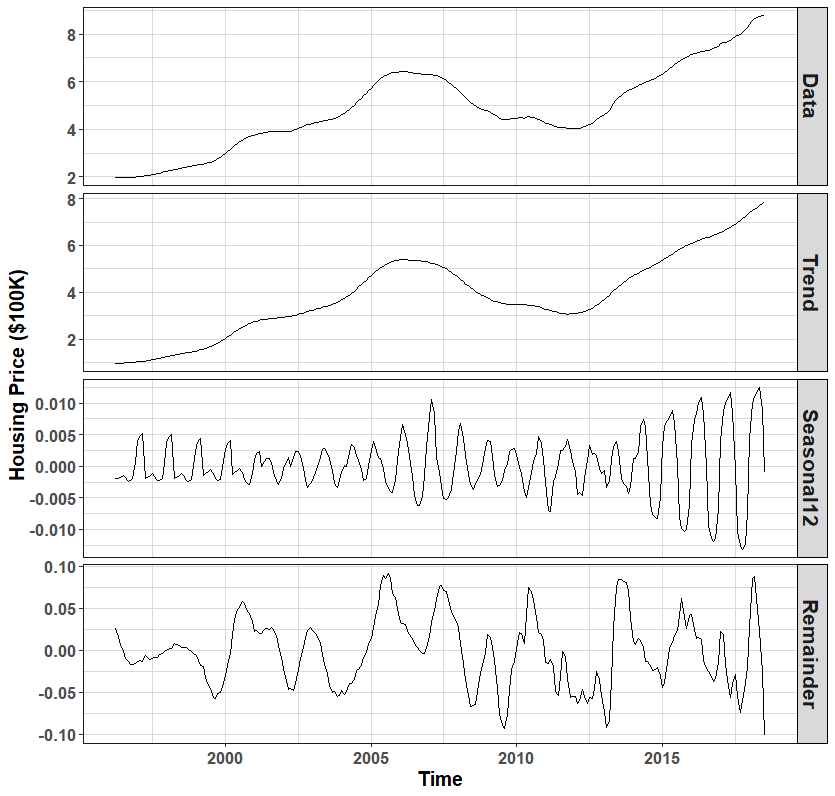


Figure 4.2 Decomposition of the additive time series.

## Forecasting with Exponential Smoothing

In the exponential smoothing methods, weights assigned to the lagged and error terms decay exponentially with higher weights assigned to most recent observation.

### Holt’s Trend Methods

Holt’s method involves a forecast equation and two smoothing equations (Holt, 1957). This method is good for time series with linear trends and low seasonality.

(4.5)

(4.6)

(4.7)

where = Predictions of the time series at time (t+h), = Estimate of the level of the time series at time t, = Estimate of the trend of the time series at time t, = Smoothing parameter for the level, and = Smoothing parameter for the trend.

The level equation is sum of the weighted average of current observation and the one-step training forecast (). Smoothing parameter for level () is always between zero and one, and penalizes the lag values exponentially as one goes back in time. Same concept and range is valid for the trend smoothing parameter (). Trend equation is sum of the trend at time t based on () and, the previous estimate of the training data slope.

Table 4.1 gives the summary of the Holt’s trend model on the train set data. The smoothing parameters were almost unity. So Equation 4.5 for the Alameda County Holt’s model was,

(4.6)

which is equations of a straight line with intercept and slope. Equation 4.6 implies that the future housing prices of Alameda County were only dependent on the current month’s housing prices and its rate. The estimate was on average good if there were no significant events disrupting the trend of the time series. With high volatility in financial sector and housing market, the resultswere not too reliable. Figure 4.3 shows prediction intervals, mean predictions and the test data. The mean cross-validation error on the training was 1%, and the test MSE was 1.27%. The magnitude of residuals was increased into the future linearly. The data in the table and in the rest of the report was in $100,000. So for first month, Holt’s method underestimated the median monthly price by $3,400 approximately, which is just 0.4% off.

Table 4.1 Holt’s Model (R output).

|  |
| --- |
| Model Summary |
| Holt's method  Call:  holt(y = AlamedaTS[1:261], h = 7)  Smoothing parameters:  alpha = 0.9999, beta = 0.9999  Initial states:  l = 1.9769, b = -0.0011  sigma: 0.0125  AIC AICc BIC  -828.5011 -828.2658 -810.6785 |
| Forecast Summary |
| Time Test Data Forecast Residuals Residual^2  2018-01-01 8.423 8.389000 -0.033999899 1.155993e-03  2018-02-01 8.569 8.495001 -0.073999398 5.475911e-03  2018-03-01 8.654 8.601001 -0.052998897 2.808883e-03  2018-04-01 8.698 8.707002 0.009001604 8.102888e-05  2018-05-01 8.755 8.813002 0.058002105 3.364244e-03  2018-06-01 8.786 8.919003 0.133002606 1.768969e-02  2018-07-01 8.783 9.025003 0.242003107 5.856550e-02 |
| Cross-validation error for Holt’s Model = 1.01 % |
| Test MSE for Holt’s Model = 1.27 % |

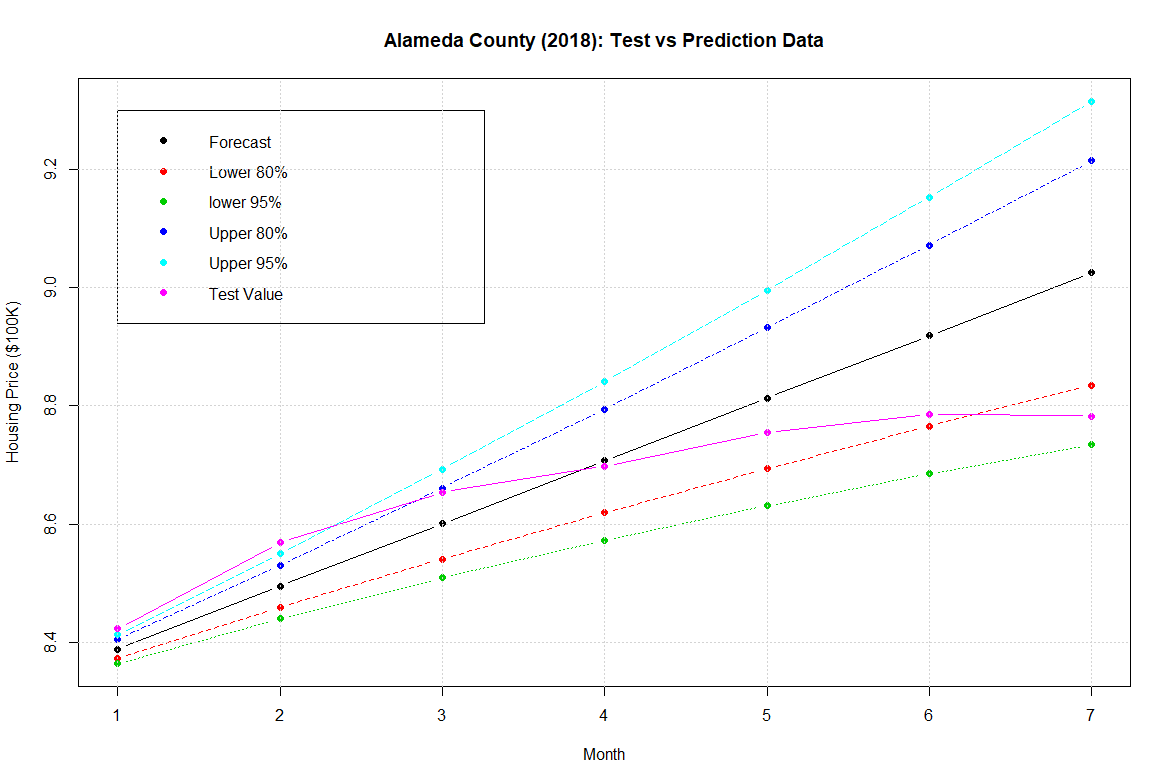


Figure 4.3 Comparison of test data with predictions.

## ARIMA Model

ARIMA model comprises of three main components:

1. Auto Regressive (AR) component, which relates the present values to the past,
2. Differential (Integrated) component, to make a non-stationary series at least weakly stationary or trend stationary, if necessary and
3. Moving Average (MA) component, which relates current value of time series to the error term (current and past).

(4.8)

Note that, if a time series has drift and trend then its mean is not constant but if the trend hold, it is classified as trend-stationary. Final performance of ARIMA models on the training data was measured based on relative magnitude of the AICc error and the ACF plots (4.3.1).

The AICc error is given as,

The AICc error penalizes the model based on the number of parameters used for the model fit.

The ACF of the series gives correlation between and, where lag *h* can be 1, 2, 3, etc.

Significant bound of ACF plots represents the 95% limit. Results were acceptable as long as 95% of the plotted residuals were within the boundary lines.

### Stationarity

The basic assumption for ARIMA model is that time series is at least weakly stationary. A weakly stationary time series satisfies following properties:

* The mean of time series is same for all time *t*.
* The variance of timer series is same for all time *t*.
* The correlation between and is same for all time *t*.

The last property of stationarity implies that the time series has same structure forward and backward. Often time one can obtained a stationary time series by taking first and second difference of the values. If drift and linear trend is present, differencing once or twice can eliminate the drift and trend term as follows,

where is the backshift operators and is the difference operator. After differencing, most of the time series pass the Unit Root Test for stationarity. Unit Root tests are measured developed to check if differencing is required for stationarity. R package *urca* has a number of unit root tests available which often output conflicting answer due to the inherit assumptions. For the current study, ADF (Dickey & Fuller, 1979) and KPSS (Kwiatkowski, Phillips, Schmidt, & Shin, 1992) test were used to test if a unit root was present and to what degree of differencing was needed.

In ADF when a null hypothesis is true, the process contains a unit roots and series is non-stationary. ADF equation with linear trend is,

(4.9)

Table 4.2 shows the R code-output for the ADF test. In case of no differencing, the output was close to the 10% critical acceptable level but more than 5%. After first and second difference, the test statics was less than 5% and 1% critical values, respectively. Both first and second difference suggested that the time parameter and intercept were not significant. Based on ADF test results, one difference was acceptable for the time series to be stationary.

The null hypothesis interpretation in KPSS is reversed. If null hypothesis is true then the process is stationary. The test static in KPSS test suggested need of differencing twice, Table 4.3.

Table 4.2 ADF Unit Root Test using ur.df test (R output).

|  |
| --- |
| Original Data |
| ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression trend  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -0.053221 -0.005003 -0.000057 0.007324 0.042110  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.584e-03 2.632e-03 2.882 0.00428 \*\*  z.lag.1 -2.673e-03 8.501e-04 -3.144 0.00186 \*\*  tt 4.619e-05 1.856e-05 2.489 0.01343 \*  z.diff.lag 9.684e-01 1.678e-02 57.704 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.01327 on 262 degrees of freedom  Multiple R-squared: 0.9286, Adjusted R-squared: 0.9278  F-statistic: 1135 on 3 and 262 DF, p-value: < 2.2e-16  Value of test-statistic is: -3.1439 3.6719 4.9515  Critical values for test statistics:  1pct 5pct 10pct  tau3 -3.98 -3.42 -3.13  phi2 6.15 4.71 4.05  phi3 8.34 6.30 5.36  First Difference |
| Test regression trend  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -0.057067 -0.004264 0.000039 0.005548 0.040855  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.204e-03 1.575e-03 0.765 0.44523  z.lag.1 -5.148e-02 1.597e-02 -3.224 0.00143 \*\*  tt 5.072e-07 1.021e-05 0.050 0.96042  z.diff.lag 3.742e-01 5.820e-02 6.430 6.05e-10 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.01259 on 261 degrees of freedom  Multiple R-squared: 0.1531, Adjusted R-squared: 0.1434  F-statistic: 15.73 on 3 and 261 DF, p-value: 1.976e-09  Value of test-statistic is: -3.2237 3.521 5.2796  Critical values for test statistics:  1pct 5pct 10pct  tau3 -3.98 -3.42 -3.13  phi2 6.15 4.71 4.05  phi3 8.34 6.30 5.36 |
| Second Difference |
| Test regression trend  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -0.049575 -0.005231 -0.000284 0.005729 0.036633  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5.258e-04 1.462e-03 0.360 0.719  z.lag.1 -9.246e-01 6.631e-02 -13.944 < 2e-16 \*\*\*  tt -3.753e-06 9.510e-06 -0.395 0.693  z.diff.lag 4.080e-01 5.760e-02 7.083 1.32e-11 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.01177 on 260 degrees of freedom  Multiple R-squared: 0.43, Adjusted R-squared: 0.4234  F-statistic: 65.38 on 3 and 260 DF, p-value: < 2.2e-16  Value of test-statistic is: -13.9443 64.8597 97.2713  Critical values for test statistics:  1pct 5pct 10pct  tau3 -3.98 -3.42 -3.13  phi2 6.15 4.71 4.05  phi3 8.34 6.30 5.36 |

Table 4.3 KPSS Unit Root Test using ur.kpss test (R output).

|  |
| --- |
| No Difference |
| Test is of type: tau with 5 lags.  Value of test-statistic is: 0.4702  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216  critical values 0.119 0.146 0.176 0.216 |
| First Difference |
| Test is of type: tau with 5 lags.  Value of test-statistic is: 0.3887  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216  critical values 0.119 0.146 0.176 0.216 |
| Second Difference |
| Test is of type: tau with 5 lags.  Value of test-statistic is: 0.0452  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216 |

### Non-Seasonal ARIMA Model

ARIMA(*p,d,q*) model of the differenced series can be written as,

(4.10)

Rearranging the above equation and using back difference operator equation 4.10 can be reformulated as,

(4.11)

## ARIMA(0,2,3)

R package has an *auto.arima* function that automatically outputs the *p, d* and *q* values. Table 4.4 shows the output of the auto.arima model for the housin prices in the Alameda County. ARIMA(0,2,3) was selected by R based as the best fit model on the training data. Using the coefficient estimates in Table 4.4, the housing price time series of Alameda County was,

The error terms had extremely low variance and zero mean, thus error impact on present housing prices was almost negligible. The model was approximated as,

ARIMA(0,2,3) model suggested that the present value of housing price was dependent only on previous months’ housing price and change in the housing price. Holt’s methods had the similar results, minus the noise term (4.2.1). Thus *auto.arima* model also had an almost linear trend (Figure 4.5). Table 4.4 lists the prediction summary and the test MSE. Test MSE of ARIMA(0,2,3) was 0.72% lower than test MSE of Holt’s method (Table 4.1). Figure 4.4 shows the residual plots for ARIMA(0,2,3) model. Both ACF and PACF plots suggested that the residuals were correlated as their magnitudes exceeded the significant bound at multiple lags. Thus the model was not a reliable forecasting model for the Alameda County housing prices. For the 36 lags shown, a model was acceptable as long as no more than 2 consecutive lag values exceeded the bound.

Table 4.4 ARIMA(0,2,3) Model (R output).

|  |
| --- |
| Model Summary |
| ARIMA(0,2,3)  Coefficients:  ma1 ma2 ma3  0.9007 -0.1267 -0.5625  s.e. 0.0584 0.0617 0.0654  sigma^2 estimated as 9.991e-05: log likelihood=825.5  AIC=-1642.99 AICc=-1642.83 BIC=-1628.76  Training set error measures:  ME RMSE MAE MPE MAPE  Training set 0.0003006419 0.00989911 0.006937339 0.01359208 0.1424329  MASE ACF1  0.1577909 -0.1653788 |
| Forecast Summary |
| Time TestData Forecast Residuals Residual^2  2018-01-01 8.423 8.385408 0.03759224 1.413177e-03  2018-02-01 8.569 8.480777 0.08822271 7.783246e-03  2018-03-01 8.654 8.575230 0.07876959 6.204648e-03  2018-04-01 8.698 8.669684 0.02831646 8.018221e-04  2018-05-01 8.755 8.764137 -0.00913666 8.347856e-05  2018-06-01 8.786 8.858590 -0.07258978 5.269277e-03  2018-07-01 8.783 8.953043 -0.17004291 2.891459e-02 |
| Test MSE for ARIMA (0,2,3) = 0.72% |

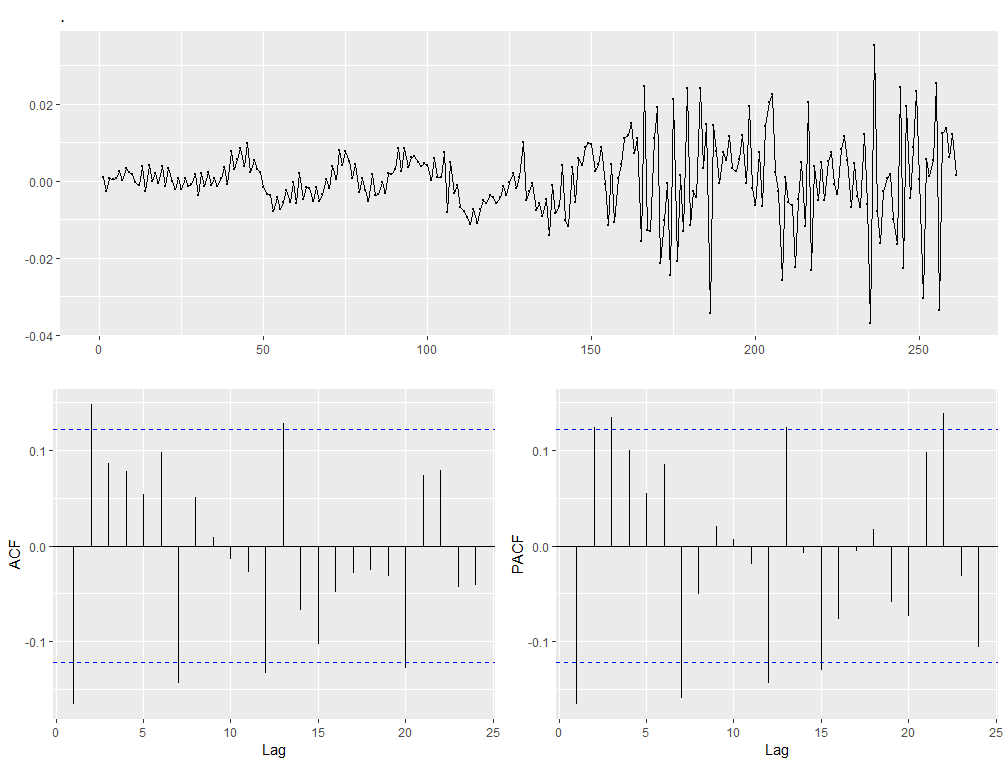


Figure 4.4 Residual data for the ARIMA(0,2,3) model.

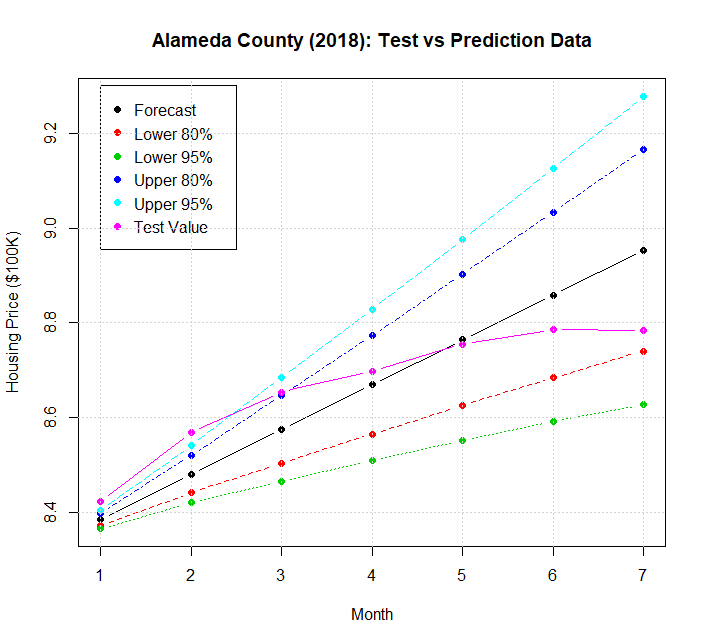


Figure 4.5 Test data and forecast of ARIMA(0,2,3).

## ARIMA(5,2,5)

Various ARIMA models were tried for improving the forecast accuracy while satisfying the set selection criterion, ARIMA (0,2,4), ARIMA (2,2,2), ARIMA (4,1,4), etc. The final model chosen with minimum AICc and acceptable ACF/PACF plots was ARIMA(5,2,5). Table 4.5 shows the output of the model and the ARIMA equation was,

AICc error of the ARIMA(5,2,5) was -1681.11 lower than the AICc error of ARIMA(0,2,3) thus satisfying the first criteria. Residual plots of ARIMA(5,2,5) also statisfied the second selection criteria and showing improvement over ARIMA(0,2,3), Figure 4.6. Test MSE of ARIMA(5,2,5) was 1.21% , higher than ARIMA(0,2,3) but because of low correlation between the residual this model was relatively more reliable when compared to ARIMA(0,2,3). Despite the highly complicated equation the forecasting trend of ARIMA(5,2,5) was almost linear, thus not adding much to the intuitive understanding to the housing price time series (Figure 4.7).

Table 4.5 ARIMA(5,2,5) Model (R output).

|  |
| --- |
| Model Summary |
| ARIMA(5,2,5)  Coefficients:  ar1 ar2 ar3 ar4 ar5 ma1 ma2 ma3  0.1361 -0.4602 0.3637 0.4009 0.4129 0.6241 0.3282 -0.3282  s.e. 0.0578 0.0546 0.0591 0.0555 0.0588 0.0269 0.0356 0.0372  ma4 ma5  -0.6241 -1.0000  0.0334 0.0326  sigma^2 estimated as 7.956e-05: log likelihood=852.09  AIC=-1682.18 AICc=-1681.11 BIC=-1643.05  Training set error measures:  ME RMSE MAE MPE MAPE  Training set 0.00039035 0.008712174 0.006138008 0.01390287 0.1250435  MASE ACF1  0.13961 -0.01636796 |
| Forecast Summary |
| Time Test Data Forecast Residuals Residual^2  2018-01-01 8.423 8.394737 0.02826331 0.0007988145  2018-02-01 8.569 8.509901 0.05909879 0.0034926674  2018-03-01 8.654 8.620950 0.03305031 0.0010923228  2018-04-01 8.698 8.718582 -0.02058186 0.0004236131  2018-05-01 8.755 8.815641 -0.06064088 0.0036773165  2018-06-01 8.786 8.921043 -0.13504260 0.0182365034  2018-07-01 8.783 9.022730 -0.23972992 0.0574704357 |
| Test MSE for ARIMA (5,2,5) = 1.22% |

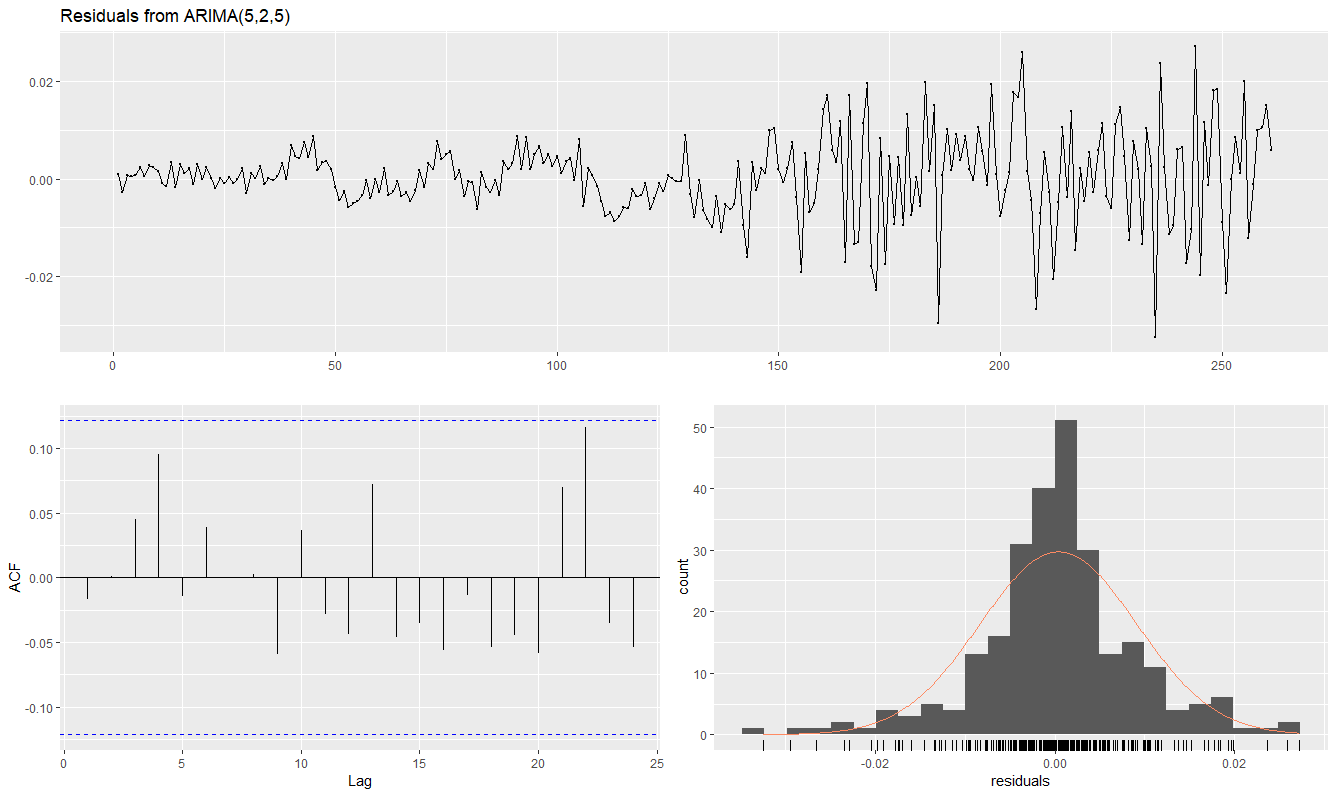


Figure 4.6 Residual data for the ARIMA(5,2,5) model.

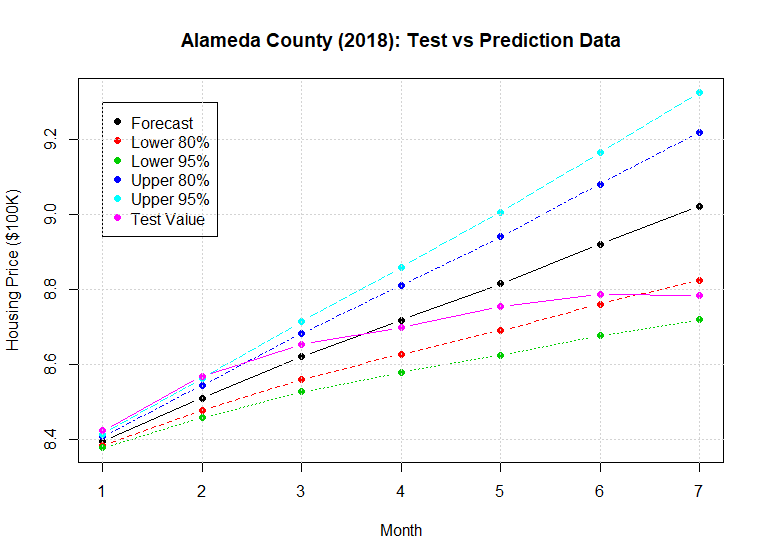


Figure 4.7 Test data and forecast of ARIMA(5,2,5).

## Breakpoint Analysis

The time series had three obvious trends as discussed earlier (section 4.1). In order to improve the prediction accuracy the impact of structural changes was incorporate in the model. The breakpoint model partitions the space into m+1 regions, where m is the number of breakpoints. Based on the type of regression model, each region is fitted with the polynomial function: constant fit, linear, quadratic, etc. In the current study, housing price in Alameda County were regressed on time. The forecasts based on the higher power are hard to interpret and their implementation is complicated when incorporating in other modes. Hence the discussion is only limited to constant and linear fit functions. Breakpoint analysis provides two error measures for the model fit: residual sum of square errors (RIC) and Bayesian Information Criterion (BIC). BIC was chosen as criteria to select the optimal number of partitions.

### Intercept Only Model

Intercept-only, breakpoint analysis finds the best constant fit for the partitioned regions. Table 4.6 gives the model summary detailing breakpoint index, date, and RIC and BIC error at each partition considered by the model. Figure 4.8 shows the residual errors for this model. Both residual sum of squares (RIC) and BIC errors were reduced as the number of time series partitions increased. Error magnitudes at 4th and 5th breakpoints was almost same. Figure 4.9 plots the observed and fitted time series data, along with the confidence intervals for the breakpoints.

Table 4.6 Breakpoint intercept-only model (R output).

|  |
| --- |
| Optimal (m+1)-segment partition:  Call:  breakpoints.formula(formula = AlamedaTS ~ 1)  Breakpoints at observation number:  m = 1 89  m = 2 57 224  m = 3 47 93 227  m = 4 48 99 144 221  m = 5 48 99 145 188 228  Corresponding to breakdates:  m = 1 2003(8)  m = 2 2000(12) 2014(11)  m = 3 2000(2) 2003(12) 2015(2)  m = 4 2000(3) 2004(6) 2008(3) 2014(8)  m = 5 2000(3) 2004(6) 2008(4) 2011(11) 2015(3)  Fit:  m 0 1 2 3 4 5  RSS 780.38 351.69 161.85 117.35 72.29 66.91  BIC 1058.17 855.74 658.95 583.97 465.28 455.76 |

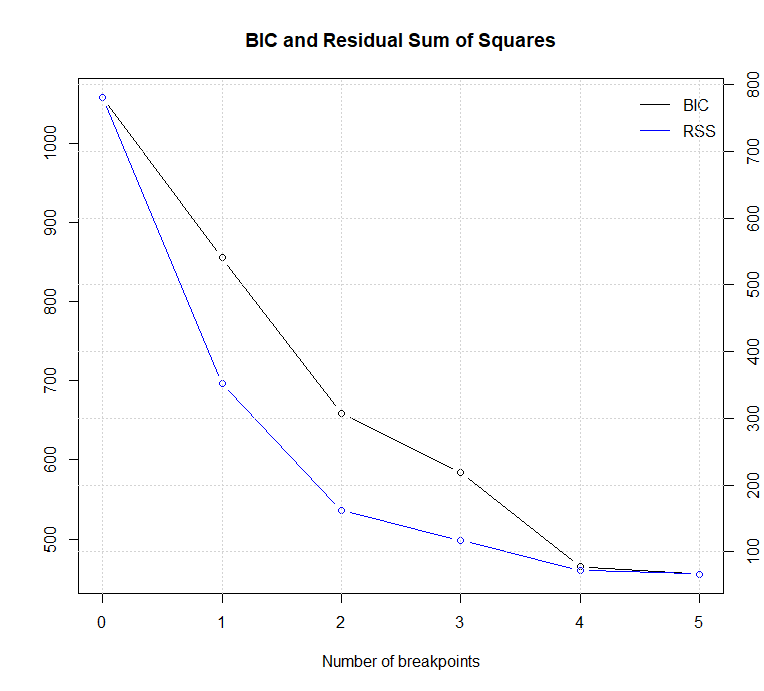


Figure 4.8 Intercept-only breakpoint analysis error measures.

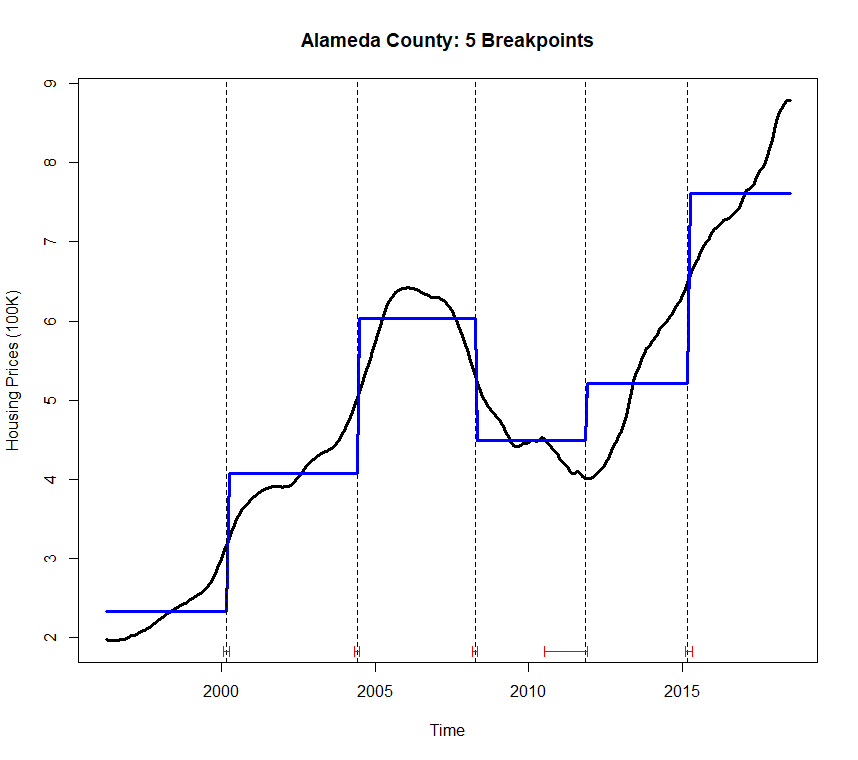


Figure 4.9 Constant fit with 5 breakpoints.

### Linear Model

To incorporate the time trend observed in the housing price time series, data was time fitted linearly. Figure 4.10 showed that the error measures decreased as the number of breakpoint increased. BIC error was minimum at 5 breakpoints. After 2 breakpoint the rate of error reduction was really low. Table 4.7 shows the summary of linear-fit breakpoint model. Figure 4.12 shows both five (suggested) and two breakpoint (chosen) plot of the original data with the linear fits. Since relative errors change from 3 partitions and 6 partition space were not significant, both plots were considered in the following sections for forecasting. Even though 5 breakpoint data fitted better, less number of breakpoints were easier to interpret when combined with other models.

Table 4.7 Breakpoint linear-fit model (R output).

|  |
| --- |
| breakpoints.formula(formula = AlamedaTS ~ timep)  Breakpoints at observation number:  m = 1 147  m = 2 108 186  m = 3 104 144 196  m = 4 81 121 161 201  m = 5 42 82 122 162 202  Corresponding to breakdates:  m = 1 2008(6)  m = 2 2005(3) 2011(9)  m = 3 2004(11) 2008(3) 2012(7)  m = 4 2002(12) 2006(4) 2009(8) 2012(12)  m = 5 1999(9) 2003(1) 2006(5) 2009(9) 2013(1)  Fit:  m 0 1 2 3 4 5  RSS 249.379 71.538 11.891 8.436 6.202 4.712  BIC 758.025 440.132 -24.007 -99.232 -164.921 -221.776 |

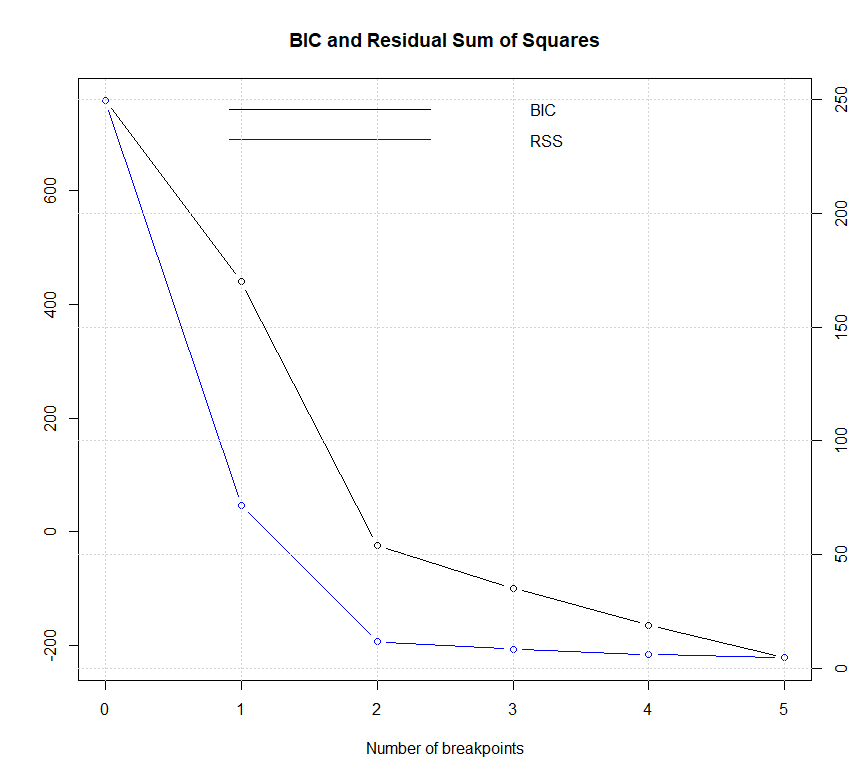


Figure 4.10 Linear breakpoint analysis error measures.

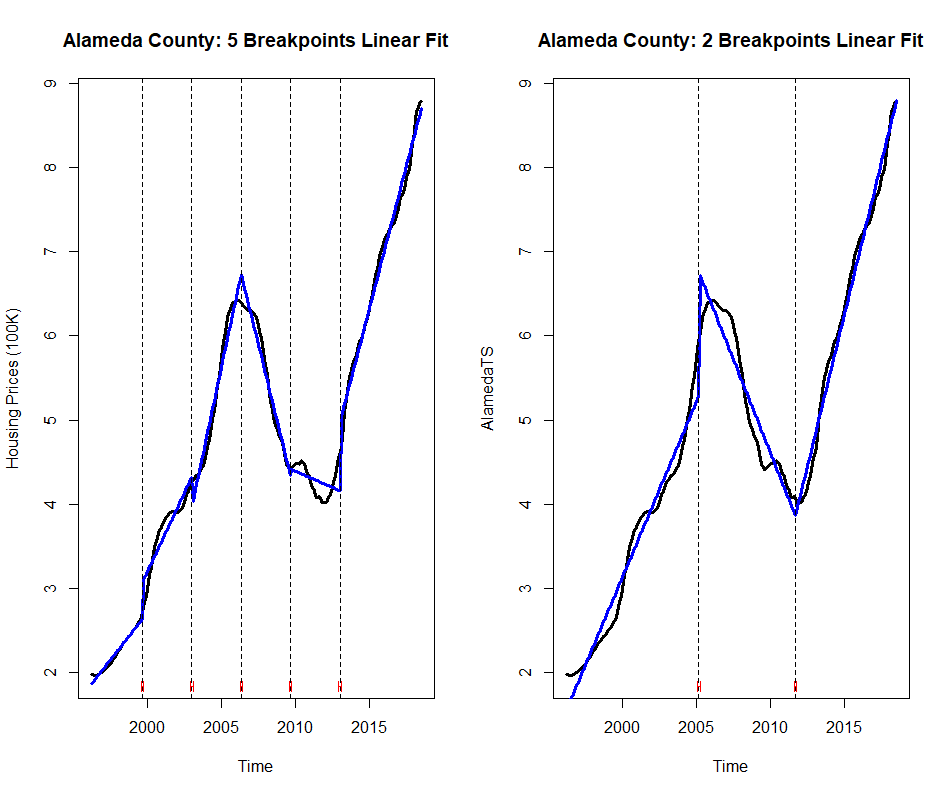


Figure 4.12 Breakpoint linear-fit with: (Left) 5 breakpoints, (Right) 2 breakpoints.

## ARIMA Model with 3 partitions

Since the trend of time series before and after the breakpoint was, complete different, only the data after the second breakpoint was considered for the housing price modeling. R function *auto.arima* was used for this reduced time series. The second breakpoint for the linear fit model was September 2011, Table 4.7. The time series was split with training set from September 2011 to December 2017 and test set from January 2018 to July 2018. Table 4.8 gives the summary of the suggested ARIMA(1,1,3) model written as,

Neglecting the really small error terms, the current value of housing prices were dependent on the past months housing price and about 88% of previoust months housing price change. A constant value of 0.0537 was also added in the fitted model which corresponds to the drift. Mathematically the model was represented as,

AICc error was -412.67, more than the previous two ARIMA models, Table 4.8. The previous ARIMA models were based on a different training set so comparing the AICc did not provide a reliable measure. However, the residual ACF and PACF plots are independent criteria as shown in Figure 4.13. Both residual functions were within the bound suggesting that errors were un-correlated and thus ARIMA with two structural breaks performed better that the previous two ARIMA models. Test MSE of ARIMA(1,1,3) was also lower, 0.43% (Table 4.8). Based on the forecast summery in Table 4.8, Figure 4.14 was plotted to compare the predicted values and test data. The mean predictions were not completely linear and the test values were relatively closer to the prediction both in short (1st and 2nd) and long (6th and the 7th) term.

Table 4.8 ARIMA(1,1,3) with 2 breakpoint (R-output).

|  |
| --- |
| Model Summary |
| ARIMA(1,1,3) with drift  Coefficients:  ar1 ma1 ma2 ma3 drift  0.8761 0.7764 -0.1662 -0.4276 0.0537  s.e. 0.0947 0.1674 0.1783 0.1764 0.0142  sigma^2 estimated as 0.0002053: log likelihood=212.95  AIC=-413.91 AICc=-412.67 BIC=-400  Training set error measures:  ME RMSE MAE MPE MAPE  Training set 0.00110619 0.01375076 0.0103362 0.01996238 0.1734625  MASE ACF1  0.179162 -0.1472886 |
| Forecast Summary |
| Time Test Data Forecast Residuals Residual^2  2018-01-01 8.423 8.385264 0.03773586 1.423995e-03  2018-02-01 8.569 8.475635 0.09336462 8.716952e-03  2018-03-01 8.654 8.556256 0.09774401 9.553892e-03  2018-04-01 8.698 8.633547 0.06445284 4.154169e-03  2018-05-01 8.755 8.707921 0.04707872 2.216406e-03  2018-06-01 8.786 8.779740 0.00626035 3.919199e-05  2018-07-01 8.783 8.849319 -0.06631883 4.398187e-03 |
| Test MSE for ARIMA (1,1,3) = 0.43% |

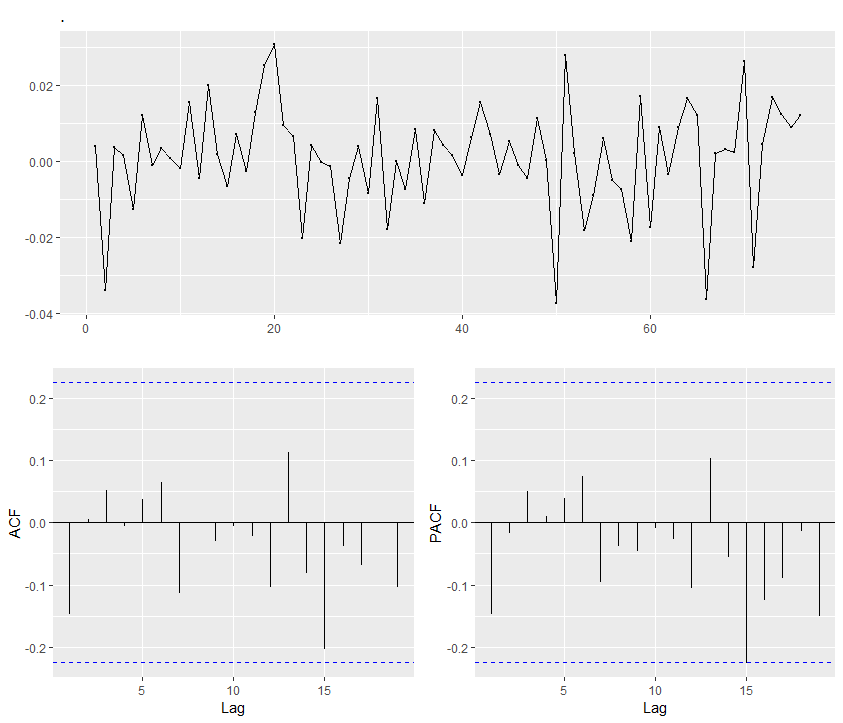


Figure 4.13 Residual, ACF and PACF plots for ARIMA(1,1,3).

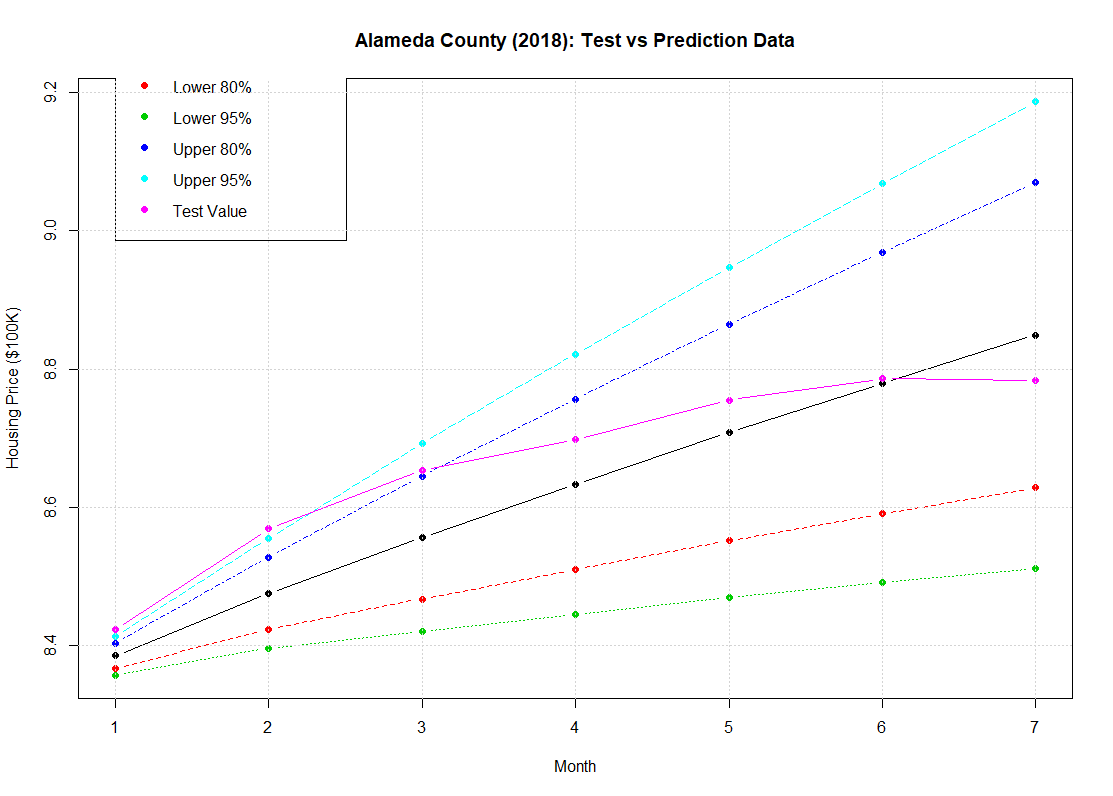


Figure 4.14 Test data and forecast of ARIMA(1,1,3).

## ARIMA Model with 6 partitions

Based on the BIC measures (Figure 4.10), 5 breakpoints provided the best linear fit to the time series. Table 4.9 shows the model output using *auto.arima* but only for data after the fifth breakpoint on January 2013 (Table 4.7). The final model equation was,

Thus current housing price in the Alameda County was dependent on the housing price of last month, and change in monthly housing price over previous three months. The AICc of ARIMA(2,2,0) was -313.99, higher than the ARIMA(1,1,3) model. The residual plot were satisfactory showing no correlation, Figure 4.14. The test mean square error of the ARIMA(2,2,0) was higher than ARIMA model with 3 partitions and ARIMA(0,2,3) obtained without considering structural breaks. This suggested that 5 breakpointsmodel was not able to learn a lot from the past data trends. Figure 4.15, shows that the final predictions were linear which was expected as the double differencing eliminates the constant term in the ARIMA equations just like previous ARIMA models without structural breaks.

Table 4.9 ARIMA(2,2,0) with 2 breakpoint (R-output).

|  |
| --- |
| Model Summary |
| ARIMA(2,2,0)  Coefficients:  ar1 ar2  0.4166 -0.5705  s.e. 0.1094 0.1114  sigma^2 estimated as 0.000239: log likelihood=160.22  AIC=-314.43 AICc=-313.99 BIC=-308.25  Training set error measures:  ME RMSE MAE MPE MAPE  Training set 0.0003015067 0.01493388 0.01220482 0.002553756 0.187845  MASE ACF1  0.1960479 0.05336483 |
| Forecast Summary |
| Time Test Data Forecast Residuals Residual^2  2018-01-01 8.423 8.385532 0.03746769 0.0014038276  2018-02-01 8.569 8.484338 0.08466209 0.0071676687  2018-03-01 8.654 8.583569 0.07043098 0.0049605233  2018-04-01 8.698 8.685103 0.01289660 0.0001663222  2018-05-01 8.755 8.787355 -0.03235471 0.0010468274  2018-06-01 8.786 8.888591 -0.10259076 0.0105248638  2018-07-01 8.783 8.988995 -0.20599481 0.0424338617 |
| Test MSE for ARIMA (1,1,3) = 0.97% |

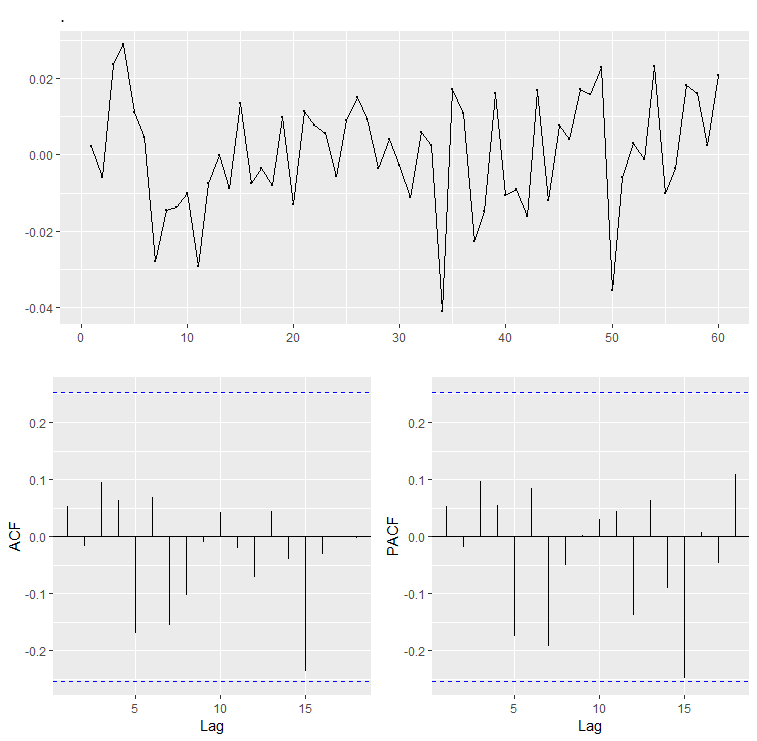


Figure 4.14 Residual, ACF and PACF plots for ARIMS(2,2,0).

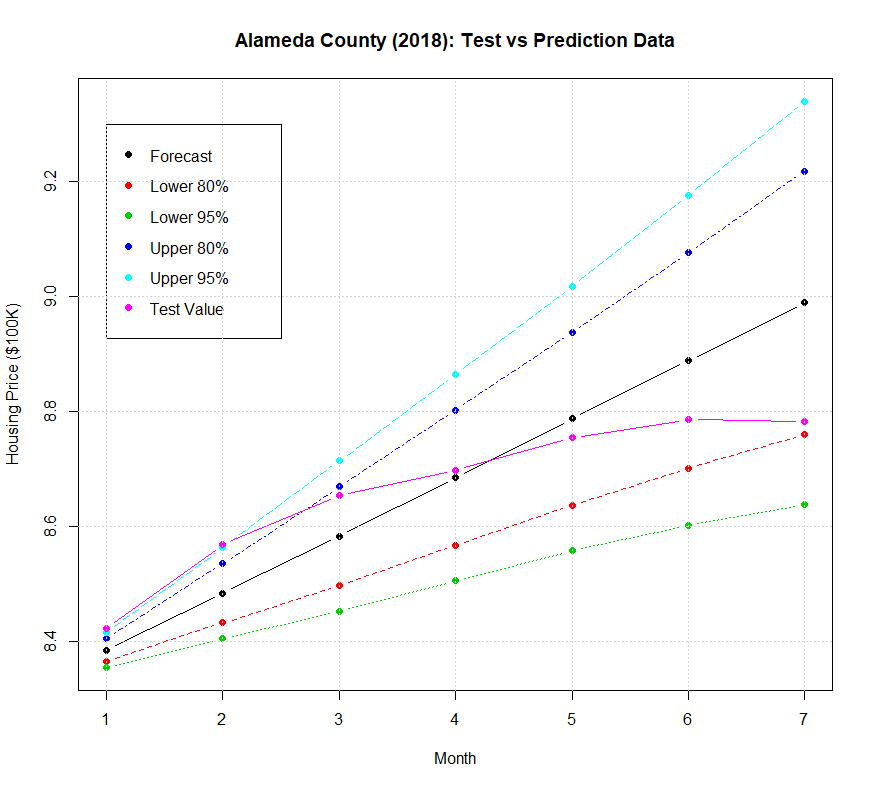


Figure 4.15 Test data and forecast of ARIMA(2,2,0).

# Dynamic Regression Model

Decision tree models allow inclusion of different predictors but they do not capture the effect of past structural changes in forecasting which can improve the model performance as discussed in the previous chapter. On the other hand, time series models only rely on the past data and error terms to predict the future values without including the effect of other predictors which can strongly impact the time series. Dynamic regression model combine both concepts to improve on the future predictions.

## ARIMA Model with Regression

ARIMA model can be written as,

(5.1)

This is modification of Equation 4.8 in the last chapter to including the *k* regression terms. The ACF for dynamic regression models, contain error from the regression thus was replaced by. Based on the results from decision tree, time and the housing price time series of San Francisco County were selected as the regressors for the model. Only data after the second structural break was used as it improved the model accuracy. ARIMA(3,0,2) was the best fit model.

Output of ARIMA(3,0,2) is shown in Table 5.1. Weights assigned to both regressions were approximately the same. No differencing was needed and the mathematical model was,

Figure 5.1 shows the residual plots from the regression and ARIMA model. Residual of ARIMA model was white noise as expected . ARIMA error was almost negligible compared to the regression error, which was close to zero near the prediction period. The AICc error for this model was -1659.19.

Figure 5.2 shows the test and predicted data for ARIMA(3,0,2) model. Due to the intercept term and zero differencing the predictions we curved in the direction of the test data. The test MSE for model was 0.33%, lower than any other model.

Table 5.1 ARIMA(3,0,2) with 2 breakpoint (R-output).

|  |
| --- |
| Model Summary |
| Regression with ARIMA(3,0,2) errors  Coefficients:  ar1 ar2 ar3 ma1 ma2 intercept tt  1.1662 0.4599 -0.6305 1.6267 0.8730 -366.2092 0.1842  s.e. 0.0603 0.1058 0.0590 0.0338 0.0332 112.0634 0.0559  SanFranciscoTS  0.1988  0.0580  sigma^2 estimated as 9.238e-05: log likelihood=838.95  AIC=-1659.91 AICc=-1659.19 BIC=-1627.83 |
| Forecast Summary |
| Time Test Data Forecast Residuals Residual^2  2018-01-01 8.423 8.385532 0.03746769 0.0014038276  2018-02-01 8.569 8.484338 0.08466209 0.0071676687  2018-03-01 8.654 8.583569 0.07043098 0.0049605233  2018-04-01 8.698 8.685103 0.01289660 0.0001663222  2018-05-01 8.755 8.787355 -0.03235471 0.0010468274  2018-06-01 8.786 8.888591 -0.10259076 0.0105248638  2018-07-01 8.783 8.988995 -0.20599481 0.0424338617 |
| Test MSE for ARIMA (1,1,3) = 0.33% |

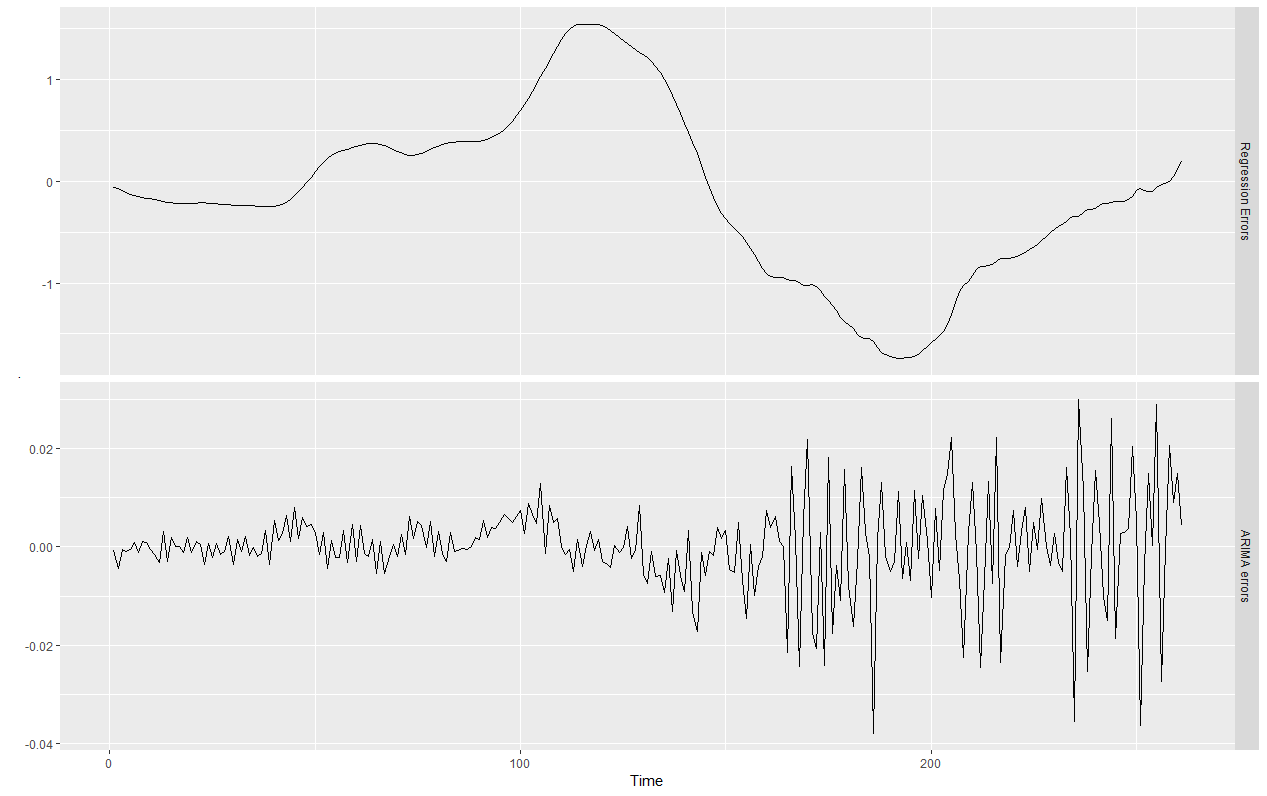


Figure 5.1 Residual for ARIMA(3,0,2).

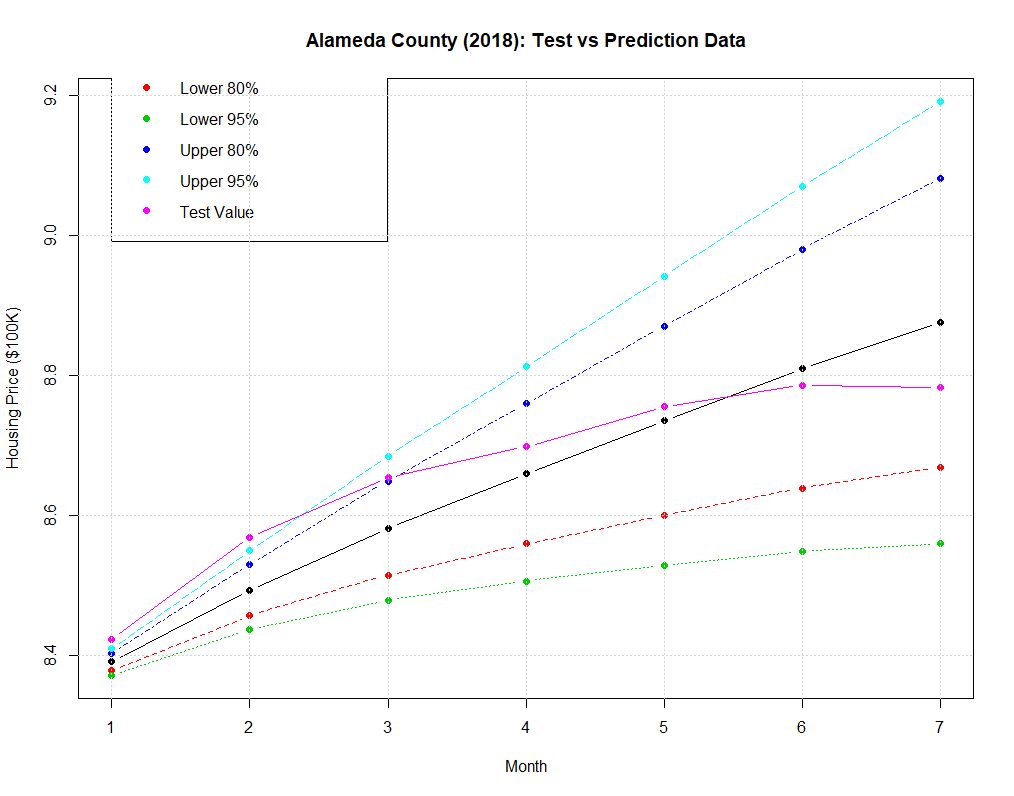


Figure 5.2 Test data and forecast of ARIMA(3,0,2).

# Summary

Total 8 predictive models were studied in this report. Two decision tree based models considered the impact of predictors like housing prices of other counties, consumer price index, etc., to predict the housing prices of the Alameda County. Time series model focused mainly on developing a forecast based entirely on the past data and error terms. Finally the observations from both models were combined to develop the finally dynamic regression model.

## Conclusion

Bagging and random forest provided consistent output in terms of importance variables. Both models suggested the housing prices in San Francisco County were relatively more important in predictive models for the Alameda County housing prices. Performance wise the random forest model was better than bagging model with smaller test error (Table 6.1).

Table 6.1 Test MSE of decision tree models.

|  |  |  |
| --- | --- | --- |
| MODEL | TEST MSE | No. of Predictors at each Splits |
| Bagging | 29.51% | 11 |
| Random Forest | 27.36% | 3 |

For time series model, consideration of structural changes played a crucial role in improve the predictability. All time series models, performed relatively better than decision tree model in term of the test MSE, Table 6.2. Holt’s exponential smoothing model has a linear forecast. ARIMA(0,2,3) and ARIMA(5,2,5) had linear forecast as well due to double differencing that eliminated the intercept term. Intercept term can provide a quadratic forecast model. This was clear when the ARIMA model was improved by including the structural partitions. Inclusion of 2 breakpoints in the model had better impact on the output compared to 5 breakpoints which was what the original analysis suggested. But that made the training set too small for the ARIMA model to recognize any pattern. All time series model suggested that the housing price and the rate of change in housing prices over the previous month were crucial for predicting the current values.

Knowledge from both, decision trees and the time series model, was combined in the dynamic regression model. The model considered the past housing prices and the two most prominent (un-correlated) predictors based on the decision tree models: time and the housing prices in the San Francisco County. The dynamic regression model outperformed all other models in terms of the test MSE (Table 6.2).

Based on the final model it was suggested that future housing prices of Alameda county can be computed with relative accuracy by considering the housing price data of past four months, the housing prices in San Francisco as long as there are no significant structural changes.

Table 6.2 Test MSE of ARIMA Models.

|  |  |  |
| --- | --- | --- |
| MODEL | TEST MSE | Notes |
| ARIMA(0,2,3) | 0.72% | auto.arima |
| ARIMA(5,2,5) | 1.22% | Random trials |
| ARIMA(1,1,3) | 0.43% | Time series with 2 breakpoints |
| ARIMA(2,2,0) | 0.97% | Time series with 5 breakpoints |
| ARIMA(3,0,2) | 0.33% | Dynamic Regression Model |

## Future Research

All models were studied for seasonally adjusted time series. Future studies can evaluate how simple models perform when effect of seasonality is included in the models. Breakpoint analysis with higher order functions might be able to provide a mode in-depth understanding. Finally a model considering more predictors might be able to improve on predictions.

# Bibliography

Bay Area Census. (2018). *Counties*. Retrieved from Bay Area Census: http://www.bayareacensus.ca.gov/counties/counties.htm

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpening, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3–73.

Department of Finance. (2018). *Estimates*. Retrieved from http://www.dof.ca.gov/Forecasting/Demographics/Estimates/

Dickey, D. A., & Fuller, W. A. (1979). Distributions of the Estimators For Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 75, 427-431.

Federal Reserve Bank of St. Louis. (2018, September). *FRED Economic Data*. Retrieved from fred.stlouisfed.org: https://fred.stlouisfed.org/series/CUURA422SA0?utm\_source=series\_page&utm\_medium=related\_content&utm\_term=other\_formats&utm\_campaign=other\_format

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer International Publishing.

Holt, C. E. (1957). *Forecasting seasonals and trends by exponentially weighted averages.* Pittsburgh: Carnegie Institute of Technology.

Hyndman, R. J., & Anthanasopoulos, G. (2018, 04). *Forecasting: Principles and Practice*. Retrieved from Otexts: https://otexts.org/fpp2/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications in R.* Springer International Publishing.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3), 159–178.

Shumway, R. H. (2017). *Time Series Analysis and its Applications with R Examples.* Springer International Publishing.

Statistics, B. o. (2018, September). *Consumer Price Index*. Retrieved from www.bls.gov: https://www.bls.gov/cpi/

Stroebel, J. (2015, January 28). *How do Rising house prices affect the local economy?* Retrieved from World Economic Forum: https://www.weforum.org/agenda/2015/01/how-do-rising-house-prices-affect-the-local-economy/

United States Department of Labor. (2018, August). *Bureau of Labor Statistics*. Retrieved from BLS Beta Labs: https://beta.bls.gov/dataViewer/view/timeseries/CUURS49BSA0

Vrieze, S. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Phsycological Methods*, 17, 228-243.

Wang , X., Smith, K. A., & Handyman, R. J. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3), 335–364.

Zillow. (2018, 09). *Zillow Research*. Retrieved from Zillow: https://www.zillow.com/research/data/