Analysis and Forecasting of Median Price of single family homes in Sacramento

University of Miami | Xiaoli Han & Mingjun Li

2/15/2020

Part 1. Introduction

For a long time, real estate can be a worthy investment opportunity. Buying or selling real estate, for a majority of investors, is one of the most important decisions they will make. It could be risky if you invest in real estate by following the media, buying a property and waiting for its value to increase. Real estate investing requires research, which includes a plan of the right price and the right time. This report will give investors who regard the Sacramento housing market as a potential investment opportunity some guidance.

The data we have in hand is median prices for single family homes in Sacramento, an island of sanity in an overpriced, over-regulated and overheated West Coast housing market. Sacramento is the capital of California, whose real estate market contains around two million people. The data are collected from Apr 1996 to Dec 2019 from Zillow. In this report, we focus on the period starting from 2014 because there was a long housing boom until the great recession, which started in the late 2000s and lasted until 2013.

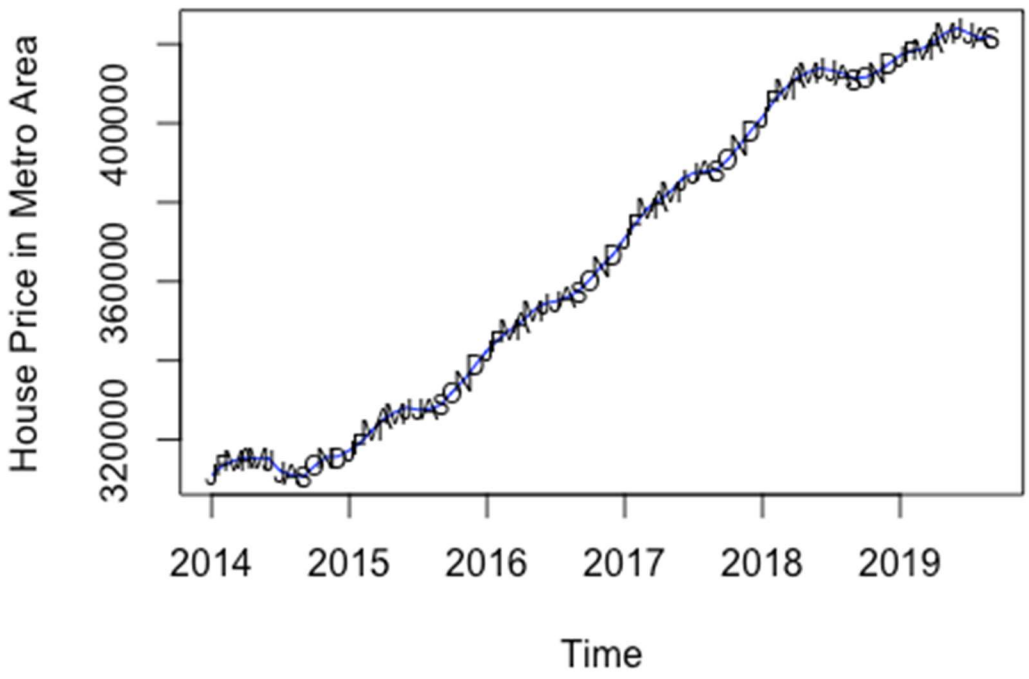
The median home prices of Sacramento risen by 3.03% over the last year. Looking forward in this year, the Sacramento real estate market forecast is that home prices will continue to increase by 5.4%. Since 2014, the median home price in Sacramento has increased from $311,111 to $429,831. In the latest quarter of 2019, the real estate appreciation rate in Sacramento has been around 1.5%. If it remains steady, the annualized appreciation rate would be around 5%, which could trigger a good amount of interest in the Sacramento real estate investment opportunities.

What are the Sacramento real estate market predictions for 2020? Is the Sacramento housing market 2020 shaping up to continue the trend of the last few years as one of the hottest markets in California? In this report, we shall model and predict median prices for single family homes in Sacramento. Please note that there are many variables that can potentially impact the value of a home in Sacramento (or any other market) and some of these variables are impossible to predict in advance.

Part 2. Model Specification

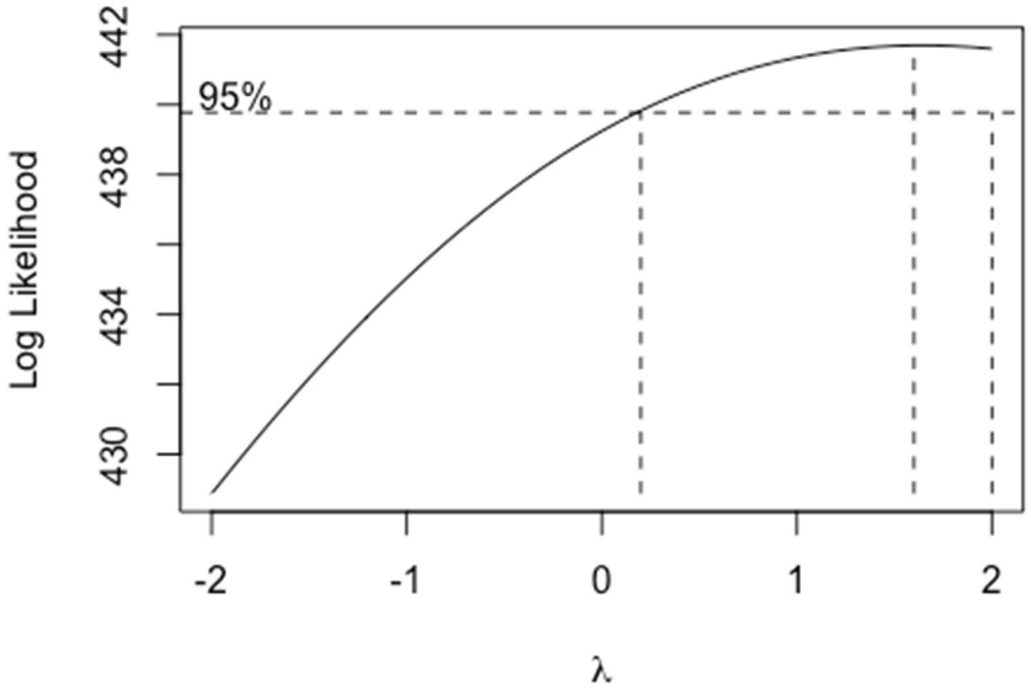
Data Overview

First, we get the data of median prices for single family homes in Sacramento from 2014 to 2019. In order to compare your forecasts to the actual values of the process, we withhold the data between 2019.10 and 2019.12. We plot the rest data where a clear ascending trend are shown.



BoxCox Transformation

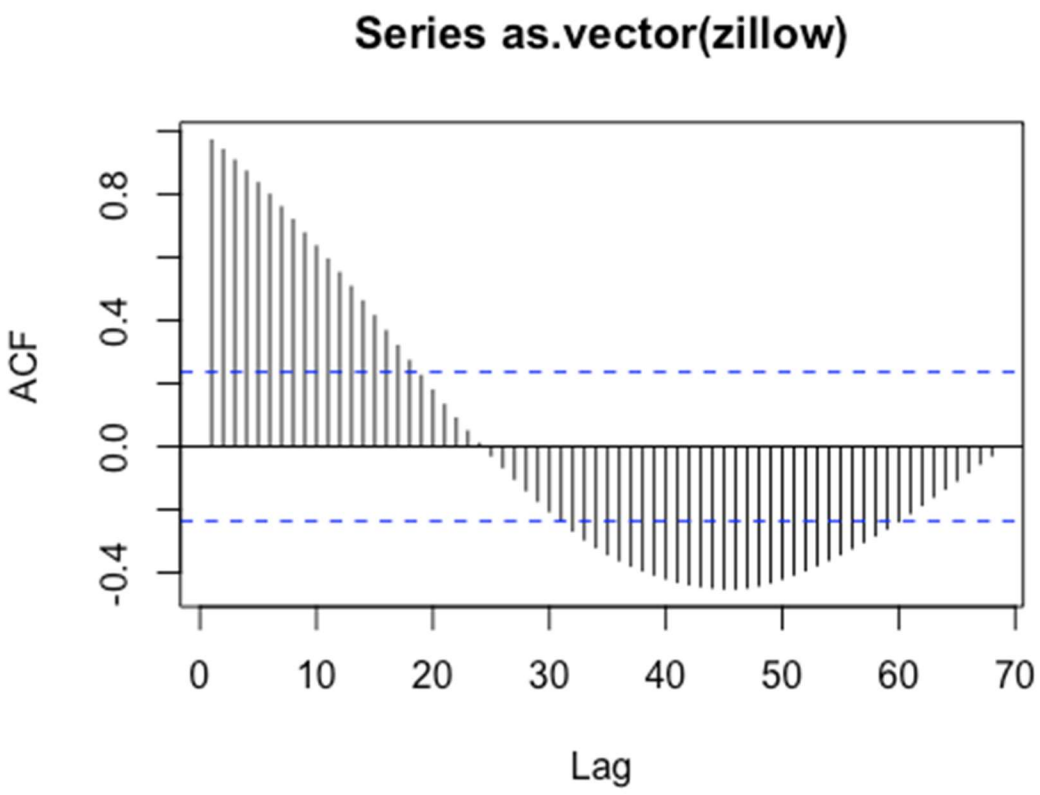
From the plot, we can see there is an obvious upward trend and a clear seasonal fluctuation. It's hard to say whether the data has a stable variance, so we'd better to run the Boxcox to decide if we should do the transformation or not.



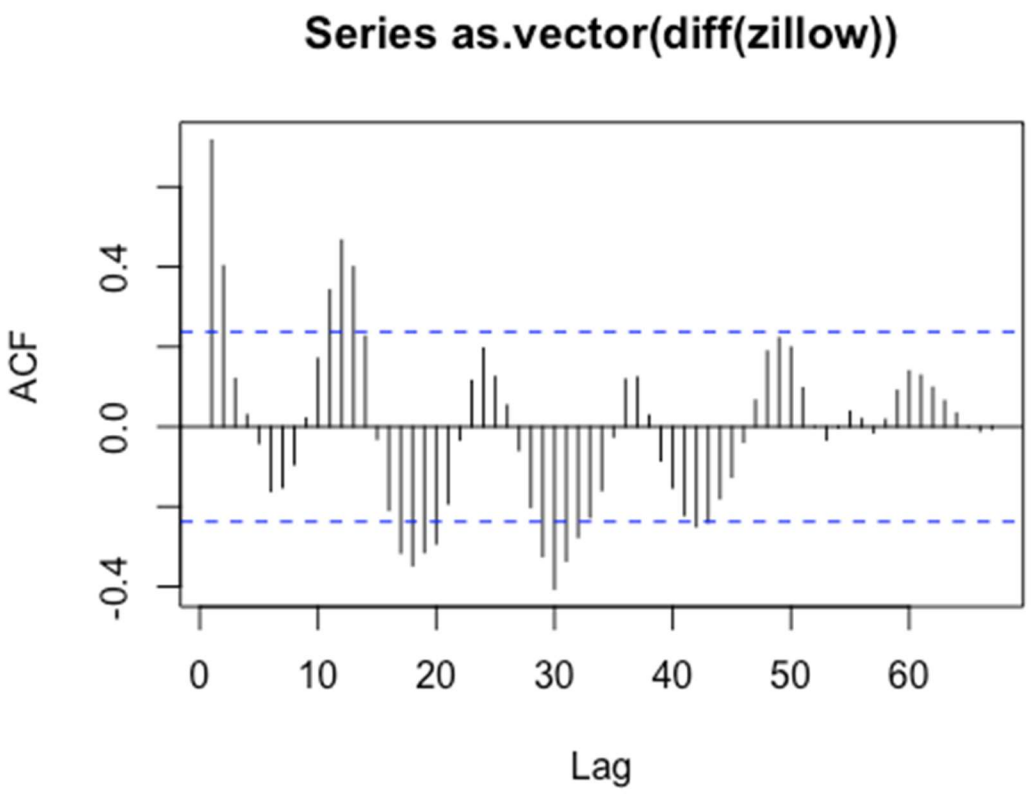
According to the plot of Boxcox, we can see that 1 is in the 95% confidence interval so that we choose the original data set to do modeling.

Remove trend and stickyness

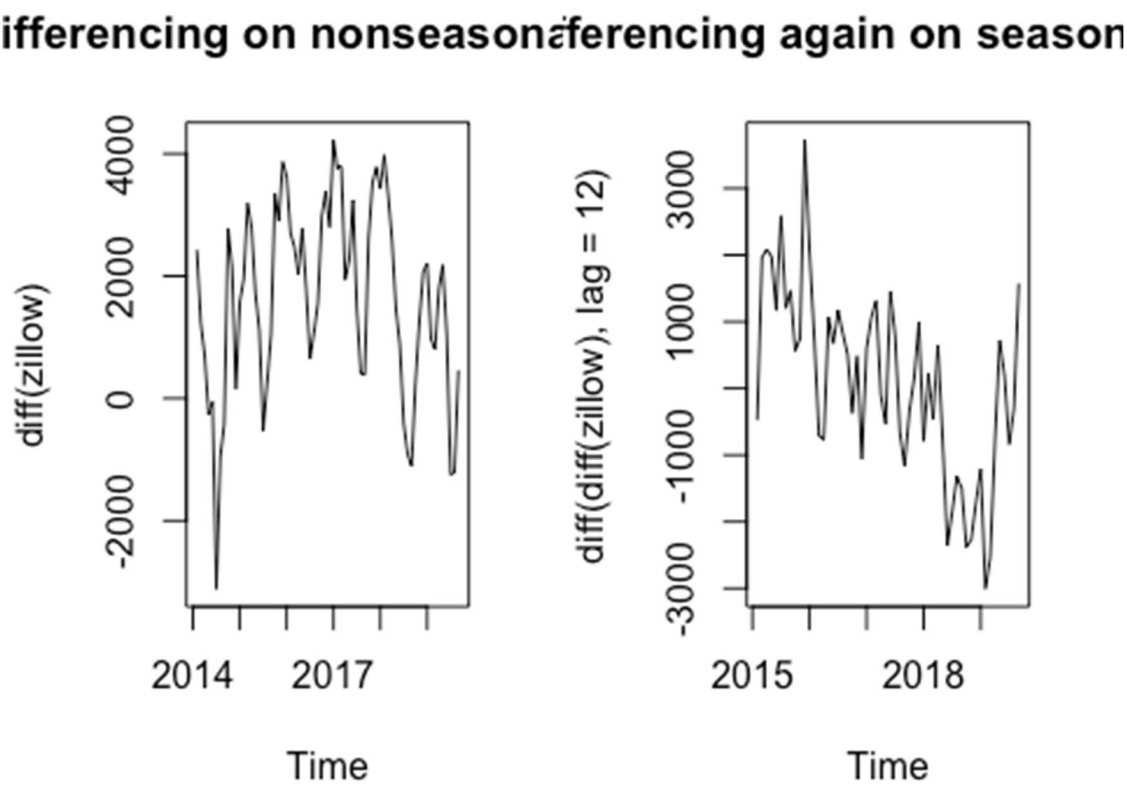
At the non-seasonal level, the ACF plot of the data set decays very slowly, indicating that the data has a very high autocorrelation. So that we need to do a difference to remove the trend.



At the seasonal level, it is hard to say if the ACF after differencing decays slowly on seasonal lag.

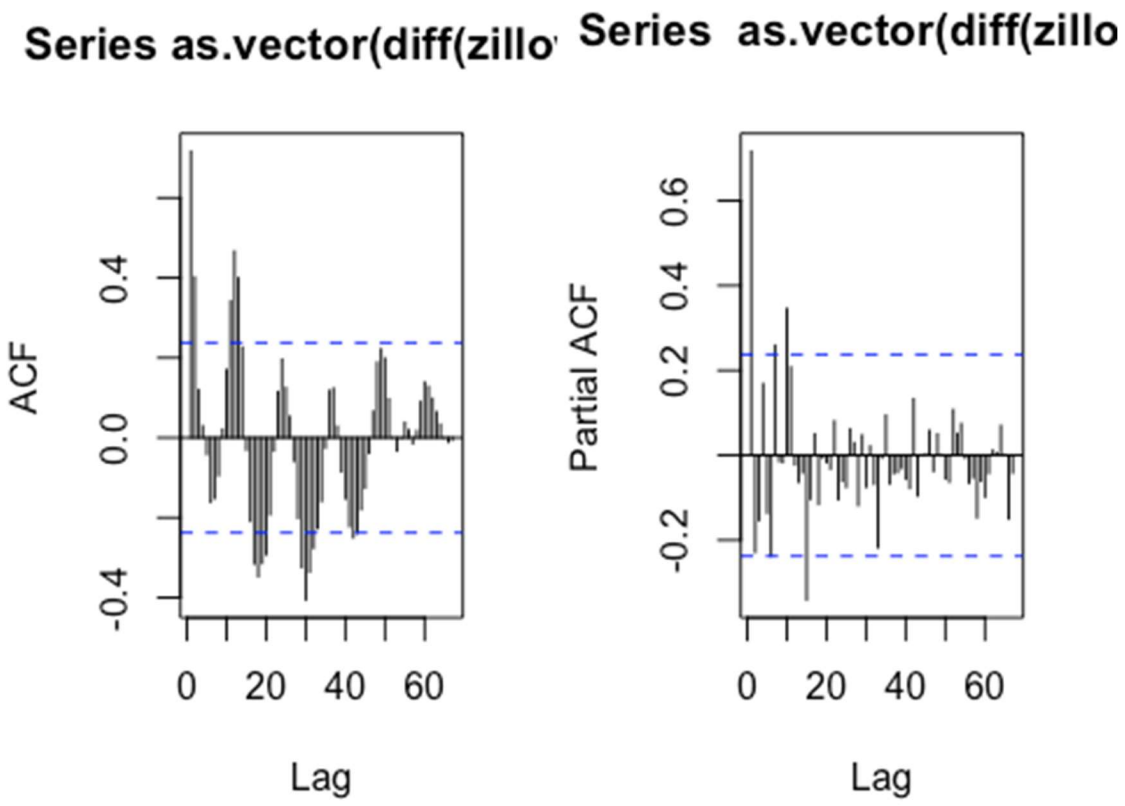


Therefore, we plot each time series plot after differencing. Differencing again on seasonal lags vanishes stationarity, hence we don't have to difference on seasonal lags.



Model Specification

After that, we try to plot the ACF, PACF and EACF of the differenced data.




```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x o o o o o o o o x  x  x  o
## 1 x x o o o x o o o o o  x  x  o
## 2 x x o o o x o o o o o  o  o  o
## 3 x o o o x o o o o o o  o  o  o
## 4 x o x o x o o o o o o  o  o  o
## 5 x x x o x o o o o o o  o  o  o
## 6 x x o x o o o o o o o  o  o  o
## 7 o o o o o o o o o o o  o  o  o
```

At seasonal level, the ACF tails off, the PACF has a small lag on the first seasonal lag which indicates either a weak seasonality or it is caused by sample variation. Therefore, we can regard PACF as cutting of at the first lag. We include seasonal ARIMAs(1,0,0) as a candidate model.

At nonseasonal level, ACF tails off or cuts off at lag 2, and PACF tails off or cut off at lag 1. Therefore, we include nonseasonal ARIMA(1,1,0), ARIMA(0,1,2) and ARIMA(1,1,1)

In conclude, the candidate models are ARIMA(1,1,0)*ARIMAs(1,0,0)* *ARIMA(0,1,2)*ARIMAs(1,0,0) ARIMA(1,1,1)*ARIMAs(1,0,0)

Model Selection

ARIMA(0,1,2)*ARIMAs(1,0,0) has lower AIC and much lower sigmar square and all the coefficients are significant, we choose it to do further diagnostics.

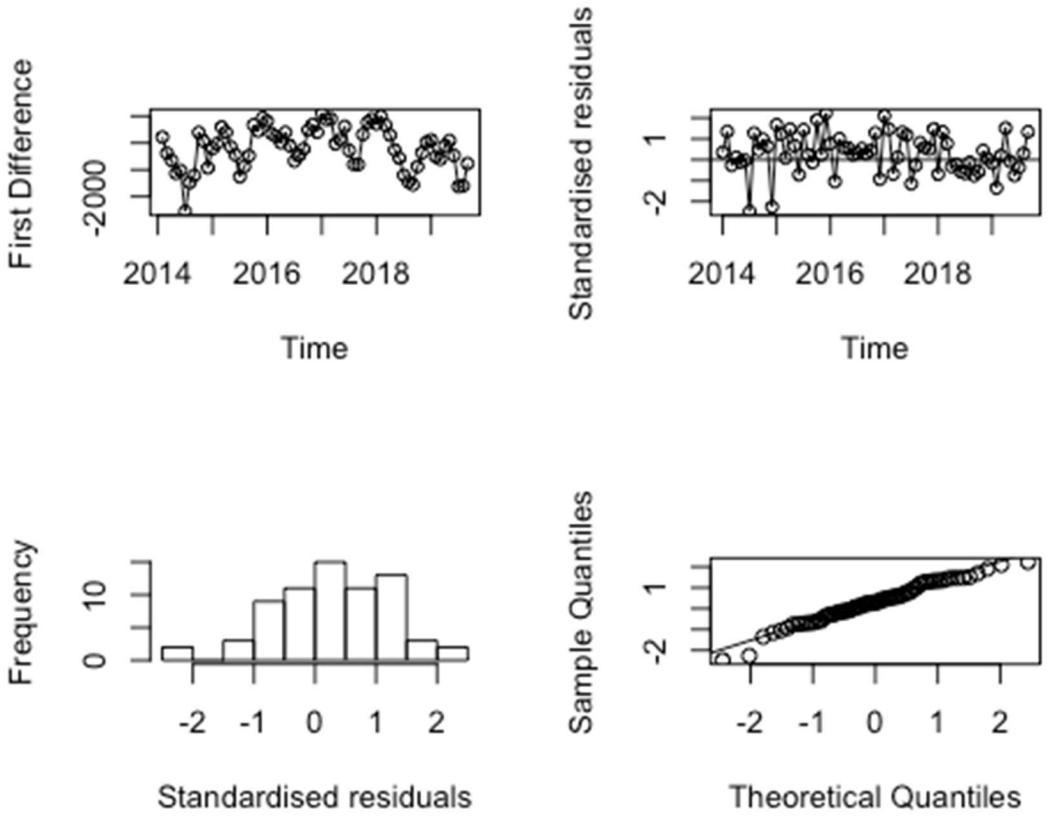
Part 3. Fitting and Diagnostics

Model Diagnostics

The residual plot has a little stickyness but the runs test doesn't reject the independency of residuals at 5% significance level. In addition, the ACF of residuals and Ljung-Box statistic also prove that the residuals are white noise.

Besides, the histogram of residuals and the QQ plot and Shapiro-Wilk test proved normality of residuals.

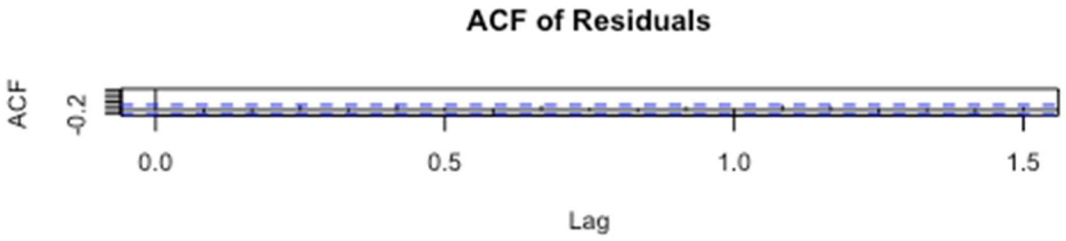
In conclusion, the residuals of ARIMA(0,1,2)*ARIMAs(1,0,0) meets all the assumptions of normality and white noise, so the model is adequate for further analysis.

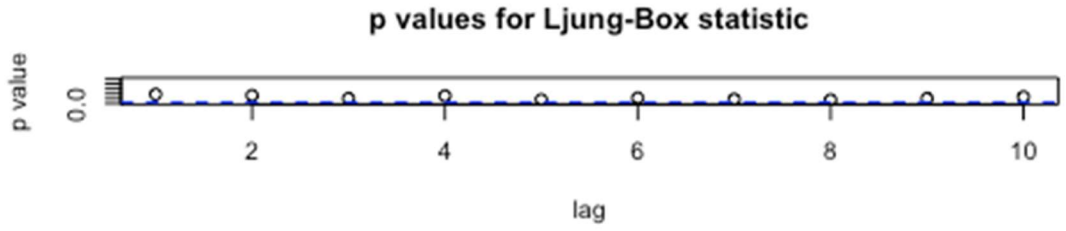



```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  rstandard(diag_model)
## W = 0.9788, p-value = 0.2901

## $pvalue
## [1] 0.16
##
## $observed.runs
## [1] 27
##
## $expected.runs
## [1] 32.88406
##
## $n1
## [1] 25
##
## $n2
## [1] 44
##
## $k
## [1] 0
```





Overfitting

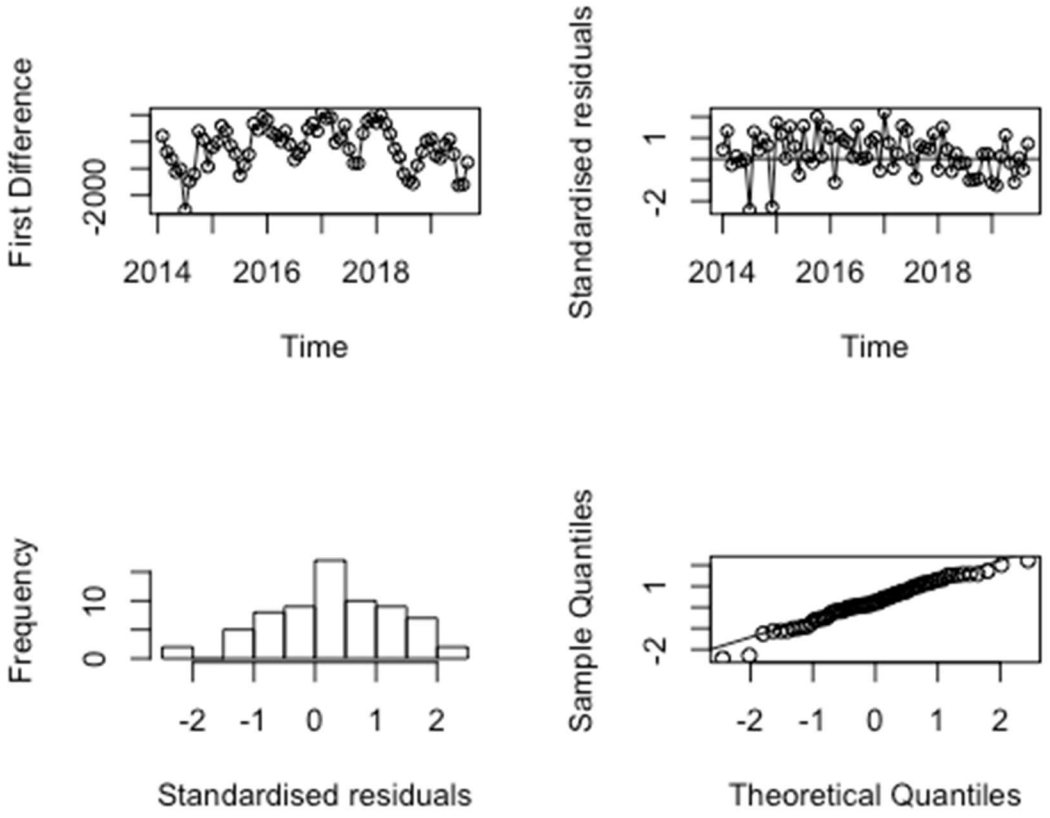
The coefficients of overfitting on nonseasonal ARIMA are unsignificant, they should be tossed out.

All coefficients of overfitting on seasonal ARIMA are significant, and the AIC and sigma square significantly decrease, so we can keep the overfitting model.

Besides, the ARIMA(0,1,2)*ARIMAs(1,0,1) performs better than ARIMA(0,1,2)*ARIMAs(2,0,0), so we select the former to do diagnostics again.

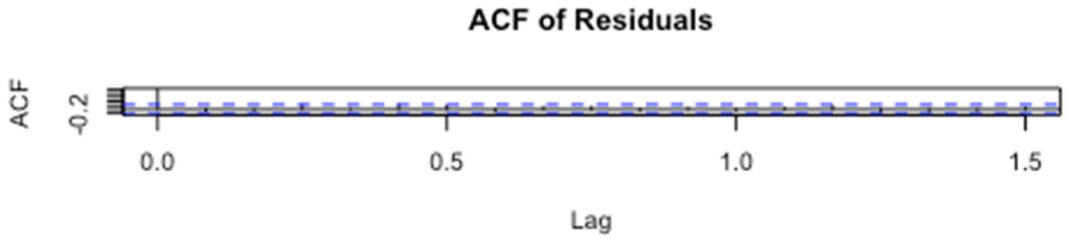
Diagnostics for Overfitting

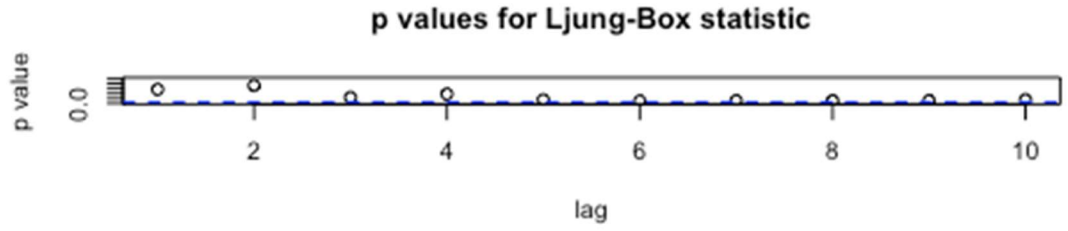
The residuals of ARIMA(0,1,2)*ARIMAs(1,0,1) meets all the assumptions of normality and white noise, so the overfitting model is a better model which will be used for prediction.




```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(diag_model)
## W = 0.97989, p-value = 0.3305
```

```
## $pvalue
## [1] 0.831
##
## $observed.runs
## [1] 31
##
## $expected.runs
## [1] 32.30435
##
## $n1
## [1] 24
##
## $n2
## [1] 45
##
## $k
## [1] 0
```

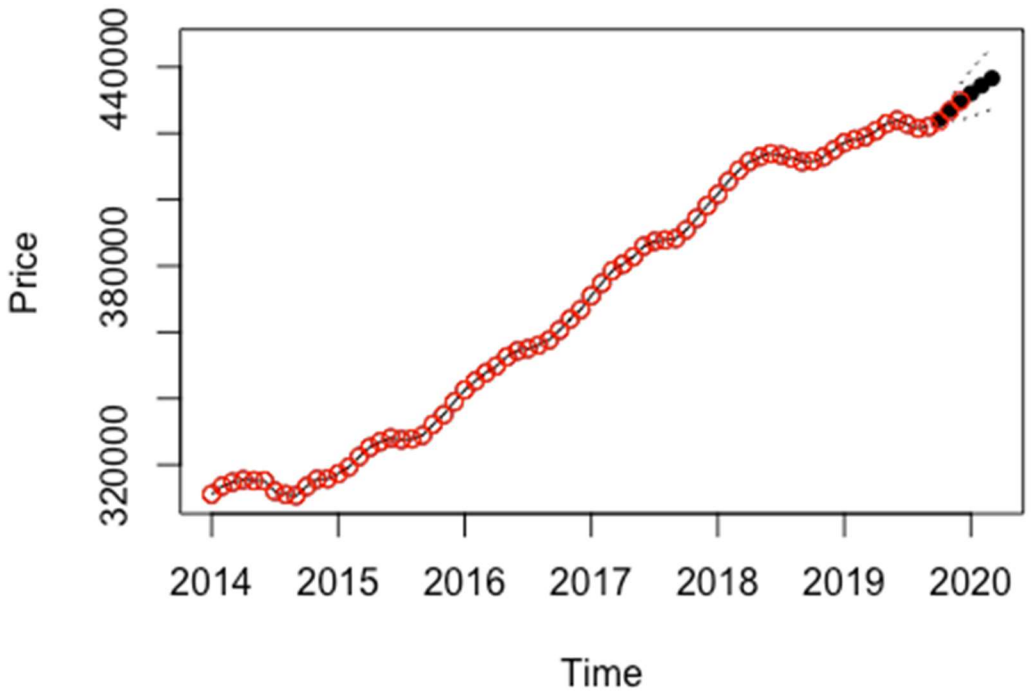




Part 4. Forecasting

Here is a short and crisp Sacramento housing market forecast for half a year ending with March of 2020. The Mean Absolute Percentage Error (MAPE) of the last three months in 2019 is 0.17% which is low enough for us to believe the final model is a good model.

If this price forecasting is correct, median prices for single family homes in Sacramento will be 4.2% higher in 2020 March than they were in 2019 March.



Part 5. Discussion

In general, we completed a thorough time series analysis of the median prices for single-family homes in Sacramento from 2014 to 2019 in this report. To preprocess the data, we ran the Boxcox and decide to use the original data. Then we did a difference at the non-seasonal level to remove the nonstationarity. To specify the model, we tried seasonal ARIMAs(1,0,0) as a candidate seasonal model and nonseasonal ARIMA(1,1,0), ARIMA(0,1,2) and ARIMA(1,1,1) as candidate nonseasonal models. In the end of data fitting, ARIMA(0,1,2)*ARIMAs(1,0,0) performed the best. During diagnosing the model, the residuals met the assumptions of normality and white noise. Then we move to model overfitting, the ARIMA(0,1,2)*ARIMAs(1,0,1) performs better than ARIMA(0,1,2)*ARIMAs(1,0,0). Since ARIMA(0,1,2)*ARIMAs(1,0,1) passes all the diagnosis

test, we choose it as our final model and use it to predict 6 months ahead. The Mean Absolute Percentage Error (MAPE) of the last three months in 2019 is 0.11% which is low enough for us to believe the final model is good.

However, there are still some problems in the process of study. For example, at the very beginning, the seasonal lags of the PACF of the data show very low values. we need to learn more to diagnose what it indicates and how to figure it out.

Part 6. Appendices

```
library(tidyr)
library(TSA)

#data cleaning
data<-read.csv("Metro_SingleFamily.csv",header=T)
n=row.names(data[grep('sacramento',data$RegionName,ignore.case = TRUE),])
zillow_all <- ts(as.numeric(data[n,-c(1,2,3)]), start=c(1996,04),frequency =
12)
zillow<-window(zillow_all,start=c(2014,1),end=c(2019,9))
test <- window(zillow_all,start=c(2019,10))

#show the ts
plot(zillow,ylab='House Price in Metro Area',type="l",col='blue')
points(y=zillow,x=time(zillow),pch=as.vector(season(zillow)),cex=.75)

BoxCox.ar1=function (y, order, lambda = seq(-2, 2, 0.01), plotit = TRUE,
          method = c("mle", "yule-walker", "burg", "ols", "yw"), ...)
{
  if (missing(method))
    method = "mle"
  y = as.vector(y/(max(abs(y)) + 1))
  if (any(y <= 0))
    stop("Data values must be positive")
  order = ar(log(y), method = method)$order
  nlngmy <- sum(log(y))
  if (!missing(lambda))
    xl <- lambda
  else xl <- seq(-2, 2, 0.1)
  loglik <- as.vector(xl)
  for (i in 1:length(xl)) if (abs(xl[i]) > 0) {
    if (missing(order))
      ar.result = ar((y^xl[i] - 1)/xl[i], method = method)
    else ar.result = ar((y^xl[i] - 1)/xl[i], method = method,
                        order.max = order)
    n = length(y) - ar.result$order
    ar.res = ar.result$resid
    n = length(y)
    loglik[i] <- -n/2 * log(ar.result$var.pred) + (xl[i] -
                                                     1) * nlngmy
  }
```

```
  else {
    if (missing(order))
      ar.result = ar(log(y), method = method)
    else ar.result = ar(log(y), method = method, order.max = order)
    n = length(y) - ar.result$order
    ar.res = ar.result$resid
    n = length(y)
    loglik[i] <- -n/2 * log(ar.result$var.pred) - nlngmy
  }
  if (plotit) {
    plot(xl, loglik, xlab = expression(lambda), ylab = "Log Likelihood",
         type = "l", ylim = c(min(loglik), max(loglik)))
    lambdahat <- loglik[loglik == max(loglik)]
    limit <- lambdahat - 0.5 * qchisq(0.95, 1)
    in.interval = xl[loglik >= limit]
    lower = in.interval[1]
    upper = rev(in.interval)[1]
    mle = (xl[loglik == max(loglik)])[1]
    lines(x = c(lower, lower), y = c(min(loglik), limit),
          lty = 2)
    lines(x = c(upper, upper), y = c(min(loglik), limit),
          lty = 2)
    lines(x = c(mle, mle), y = c(min(loglik), max(loglik)),
          lty = 2)
    abline(limit, 0, lty = 2)
    scal <- (par("usr")[4] - par("usr")[3])/par("pin")[2]
    text(c(xl[1]) + 0.1, limit + 0.08 * scal, " 95%")
  }
  invisible(list(lambda = xl, loglike = loglik, mle = mle,
                 ci = c(lower, upper)))
}

# observation: trend
#Box Cox test
BoxCox.ar1(zillow,method='burg')

#diff or not?
acf(as.vector(zillow),lag.max = 72)

acf(as.vector(diff(zillow)),lag.max = 72)

par(mfrow=c(1,2))
plot(diff(zillow),main='Differencing on nonseasonal lags')
plot(diff(diff(zillow),lag=12),main='Differencing again on seasonal lags')

par(mfrow=c(1,2))
acf(as.vector(diff(zillow)),lag.max = 72)
pacf(as.vector(diff(zillow)),lag.max = 72)

eacf(diff(zillow))
```

```r
#seasonal AR(1)

  #nonseasonal AR(1)
arima(zillow,order=c(1,1,0),seasonal = list(order=c(1,0,0)),method='ML')

  #nonseasonal MA(2)
arima(zillow,order=c(0,1,2),seasonal = list(order=c(1,0,0)),method='ML')

  #nonseasonal ARMA(1,1)
arima(zillow,order=c(1,1,1),seasonal = list(order=c(1,0,0)),method='ML')

par(mfrow=c(2,2))
diag_model = arima(zillow,order=c(0,1,2),seasonal = list(order=c(1,0,0)),meth
od='ML')
plot(diff(zillow),ylab='First Difference',type='o')
plot(rstandard(diag_model),xlab="Time",ylab="Standardised residuals",type='o'
)
abline(h=0)
hist(rstandard(diag_model),xlab="Standardised residuals",main="")
qqnorm(rstandard(diag_model),main="")
qqline(rstandard(diag_model))

shapiro.test(rstandard(diag_model))
runs(rstandard(diag_model))

tsdiag(diag_model)

# overfit ARIMA(0,1,2)*ARIMAs(1,0,0)
arima(zillow,order=c(1,1,2),seasonal = list(order=c(1,0,0)),method='CSS-ML')
arima(zillow,order=c(0,1,3),seasonal = list(order=c(1,0,0)),method='ML')
arima(zillow,order=c(0,1,2),seasonal = list(order=c(2,0,0)),method='ML')
arima(zillow,order=c(0,1,2),seasonal = list(order=c(1,0,1)),method='ML')

par(mfrow=c(2,2))
diag_model = arima(zillow,order=c(0,1,2),seasonal = list(order=c(1,0,1)),meth
od='ML')
plot(diff(zillow),ylab='First Difference',type='o')
plot(rstandard(diag_model),xlab="Time",ylab="Standardised residuals",type='o'
)
abline(h=0)
hist(rstandard(diag_model),xlab="Standardised residuals",main="")
qqnorm(rstandard(diag_model),main="")
qqline(rstandard(diag_model))

shapiro.test(rstandard(diag_model))
runs(rstandard(diag_model))

tsdiag(diag_model)
```

```r
pred <- predict(diag_model,n.ahead=6)
round(pred$pred)
round(pred$se)

# Create lower and upper prediction interval bounds
lower.pi<-pred$pred-qnorm(0.975,0,1)*pred$se
upper.pi<-pred$pred+qnorm(0.975,0,1)*pred$se

# Display prediction intervals
data.frame(Time=c(2019.10,2019.11,2019.12,2020.01,2020.02,2020.03),lower.pi,u
pper.pi)

z<- window(zillow_all,start=2014)
plot(diag_model,n.ahead=6,col='black',type='b',pch=16,n1=2014,ylab="Price",xl
ab="Time")
points(x=time(z),y=z,col='red',type='b')
x.temp=c(2019.10,2019.11,2019.12,2020.01,2020.02,2020.03)
#abline(h=coef(diag_model)[names(coef(diag_model))=='intercept'])

#install.packages('MLmetrics')
library(MLmetrics)
MAPE(pred$pred[1:3],test)
```