Time Series Report

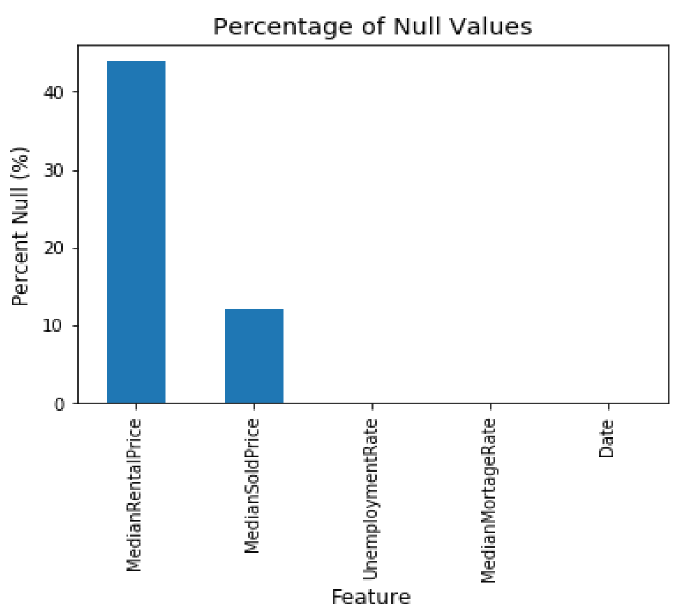
Geoffrey Hung, Jacob Goffin, Sean Tey, Sunny Kwong, and Andrew Eaton

Description of the problem

Our Zillow dataset contains 164 rows representing California housing data. We are forecasting the monthly median sold price across all homes in California. The other features are the median mortgage rate, unemployment rate, and median rental price. The dataset at a glance looks like:

Date	MedianSoldPrice_AllHomes.California	MedianMortageRate	UnemploymentRate	MedianRentalPrice_AllHomes.California
2004-01-31	326752.55	5.02	7.9	NaN
2004-02-29	329501.50	4.94	7.8	NaN
2004-03-31	340125.45	4.74	7.8	NaN
2004-04-30	355329.50	5.16	7.5	NaN
2004-05-31	367818.15	5.64	7.3	NaN
...	...	...	...	...
2017-04-30	NaN	3.91	4.4	2600.0
2017-05-31	NaN	3.83	4.3	2650.0
2017-06-30	NaN	3.88	4.4	2675.0
2017-07-31	NaN	3.88	4.3	2695.0
2017-08-31	NaN	3.74	4.3	2695.0

An interesting aspect of this dataset is that there is missing data for certain features. For example, there are no observations for the median monthly rental price until 2010, whereas the other features have observations dating back to 2004.

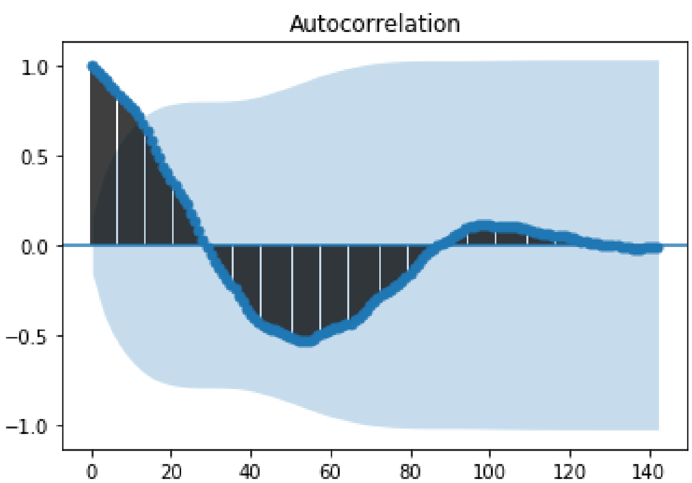


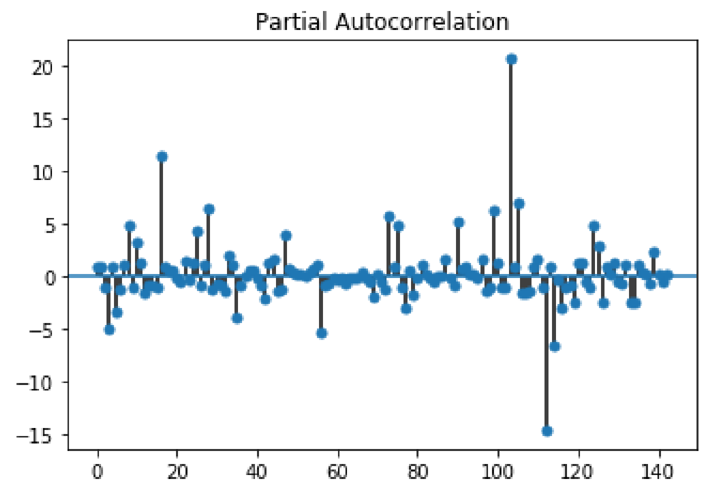
Using the data that we have, we will predict the Median Sold Price for all dates past December 31st, 2015.

Description of methods chosen

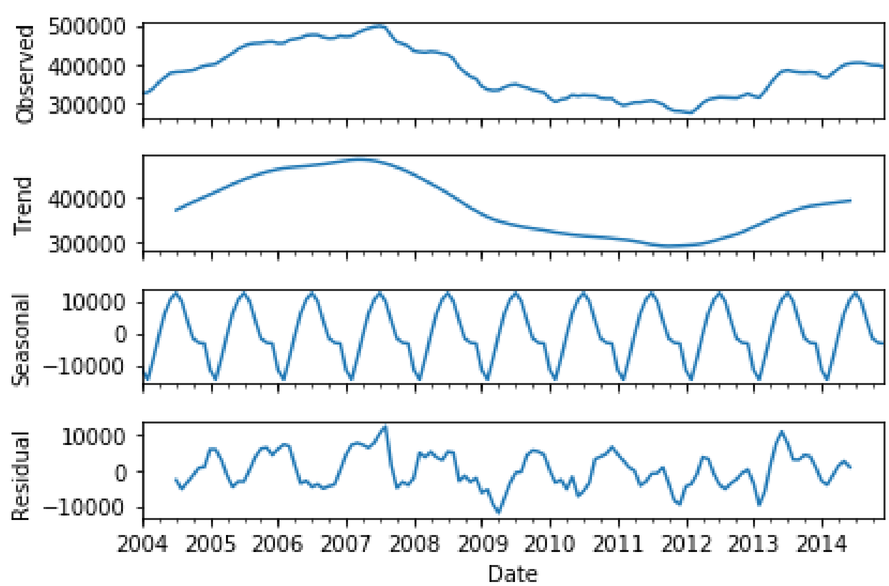
We investigated different time series model for forecasting, including ARIMA (or SARIMA) and Exponential Smoothing. We have multiple features, so we also wanted to explore multivariate time series models such as SARIMAX and VAR. Below is our decision making process in choosing the final models to try out.

We eliminated AR and MA as potential models, since from the PACF and ACF plots it was not clearly one or the other.





We also eliminated ARMA, ARIMA, Single Exponential Smoothing (SES), and Double Exponential Smoothing (DES) models. From seasonal decomposition, we notice that there is a clear trend and seasonal component to the Median sold price over time.



Therefore, the final 4 models we have chosen to fit is SARIMAX, TES, SARIMA and VAR. SARIMA and TES are both univariate time series models, and both support seasonal and trend components. The main difference is that TES uses exponentially decreasing weights for past observations. Since we are unsure whether there are exponentially decreasing weights or if we can weigh all past observations equally, we will experiment with both SARIMA and TES.
For TES, we also assume an additive trend and seasonality, as the amplitude of seasonal variation and trend doesn't seem to increase with time.

Since there are two other variables involved, we will experiment with SARIMAX and VAR as well. There is both an endogenous and exogenous variables to consider, as mortgage rate does have an impact on our target, but the unemployment rate does not. SARMAX is good for exogenous relationships, while VAR is good for endogenous relationships, therefore we will experiment with both models.

<u>Textual and visual report of findings</u>

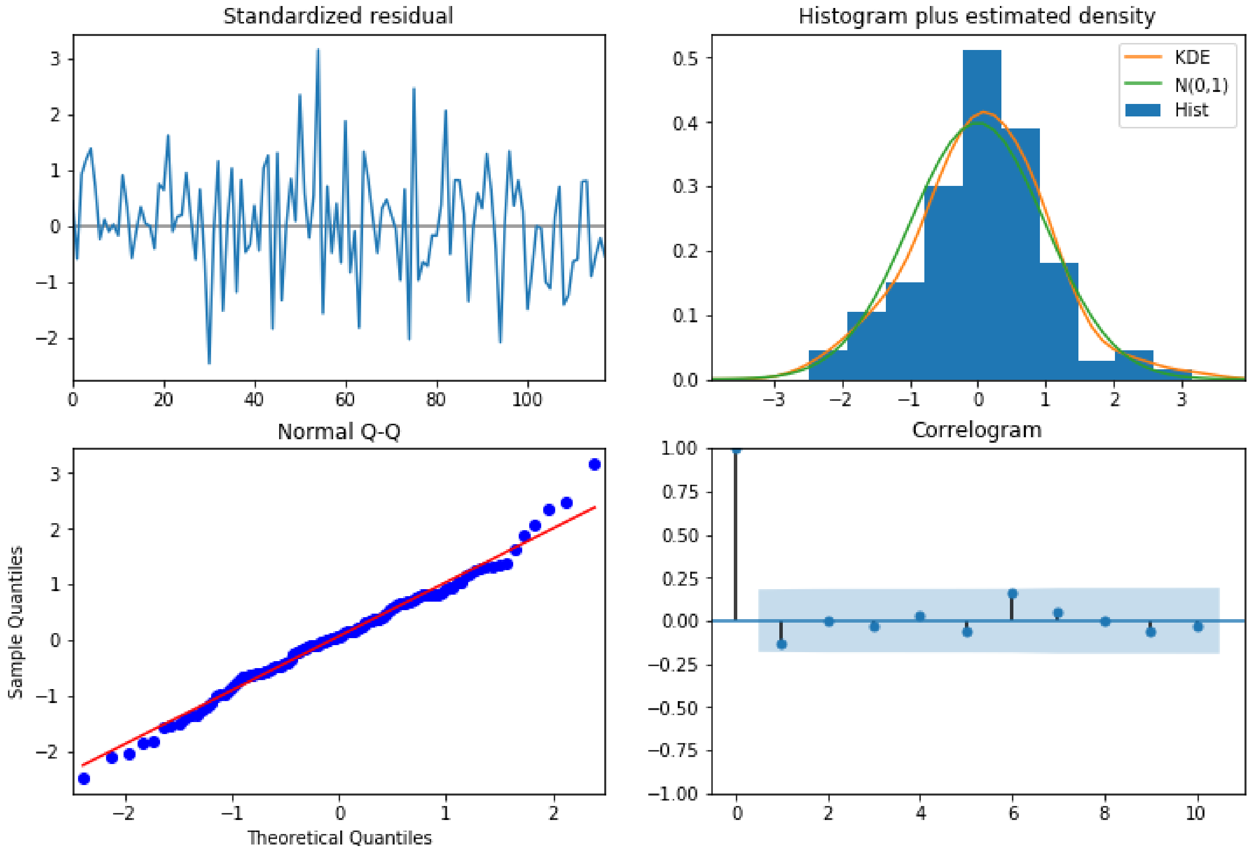
This section will go over the various models we fitted, including the parameters we have found and the Root Mean Square Error for each model.

<u>SARIMA</u>

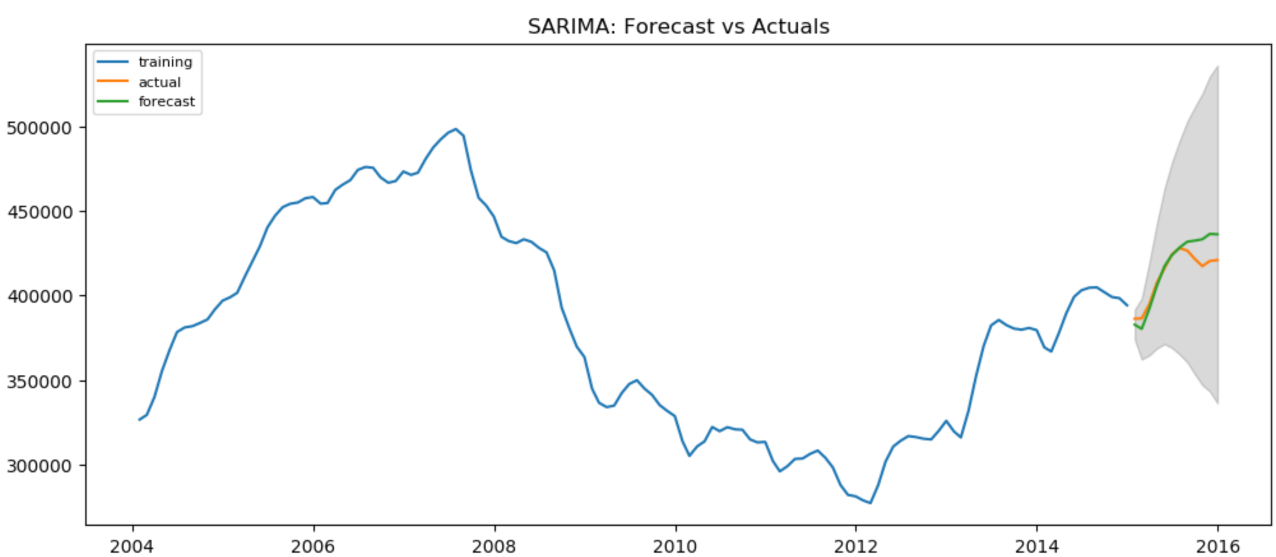
Using the differencing process, we found that the estimated d (power of the trend) would be around 2 and the estimated m (seasonality) would be around 12. Afterwards, we used Auto-ARIMA function to find out the other parameters, and the final model we ended up with is SARIMAX (1,2,1)X(0,1,0,12).

```
                                      Statespace Model Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 2, 1)x(0, 1, 0, 12)   Log Likelihood               -1158.927
Date:                            Sat, 07 Dec 2019   AIC                           2325.853
Time:                                    21:07:36   BIC                           2336.936
Sample:                                         0   HQIC                          2330.353
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     20.8805     13.147      1.588      0.112      -4.888      46.648
ar.L1          0.7441      0.063     11.720      0.000       0.620       0.869
ma.L1         -0.9999      0.120     -8.323      0.000      -1.235      -0.764
sigma2      2.068e+07   6.83e-07   3.03e+13      0.000    2.07e+07    2.07e+07
===================================================================================
Ljung-Box (Q):                       59.21   Jarque-Bera (JB):                 1.67
Prob(Q):                              0.03   Prob(JB):                         0.43
Heteroskedasticity (H):               1.18   Skew:                             0.10
Prob(H) (two-sided):                  0.62   Kurtosis:                         3.55
===================================================================================
```

Furthermore, we checked the assumptions, and we see that all of the assumptions are satisfied, although the QQ plot shows the distribution of errors is a little shaky around normal distribution.

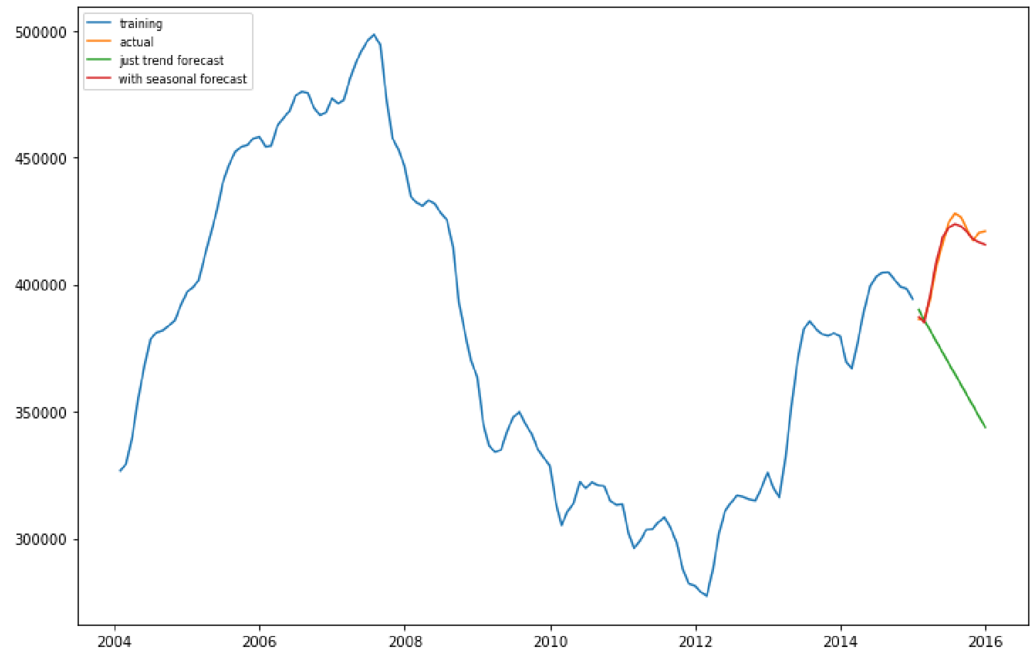


From the prediction plot here, we can see that initially the forecast seems to follow the actual results. However near the end of 2016, the forecast seems to start diverging from the actual results. The RMSE we found from this model is 8869.10.



<u>Exponential Smoothing</u>

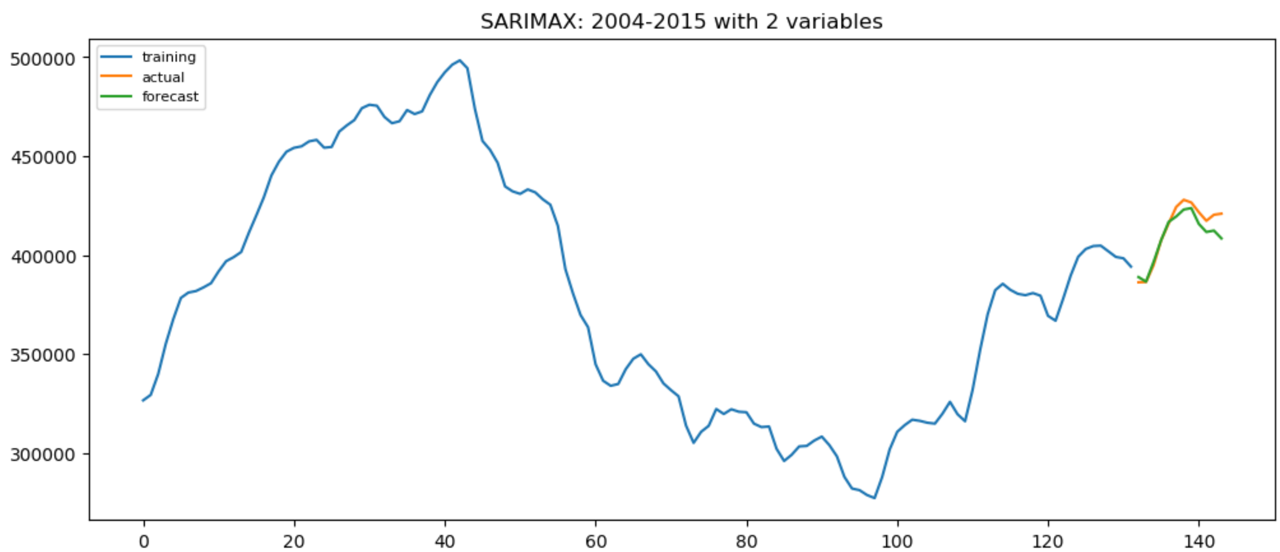
We used the seasonality results we found earlier and visual inspection of the Median sold price over time (m=12 and additive trend/seasonality) to fit this model into Python and we can see from this graph that the forecast seems the follow the actual data closely. The RMSE we found from this model is 2806.50.



SARIMAX

After fitting SARIMA model, we used the results from SARIMA to fit a SARIMAX model using the two variables Median Mortgage Rate and Unemployment Rate. As stated earlier, we stated three models, training years from 2004-2015 using the two non-missing variables, training from 2010-2015 using the two non-missing predictive variables, and training from 2010-2015 using all three predictive variables.

From the prediction plots below, for training from 2004-2015, we can see that the forecast does follow the path of the actual values, but it seems to underestimates the model near the end of 2016. The RMSE obtained from this model is 5420.14.

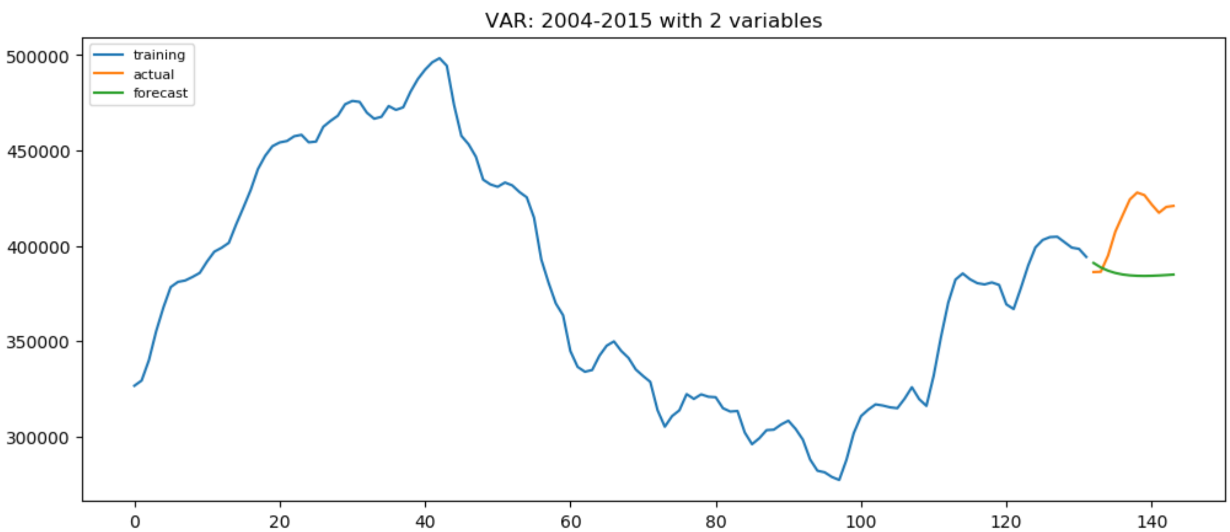


From the prediction plots below, for training from 2010-2015, using the two non-missing predictive variables, we can see that the forecast does follow the path of the actual values, but it seems to underestimates the model at all times. The RMSE obtained from this model is 5311.28.

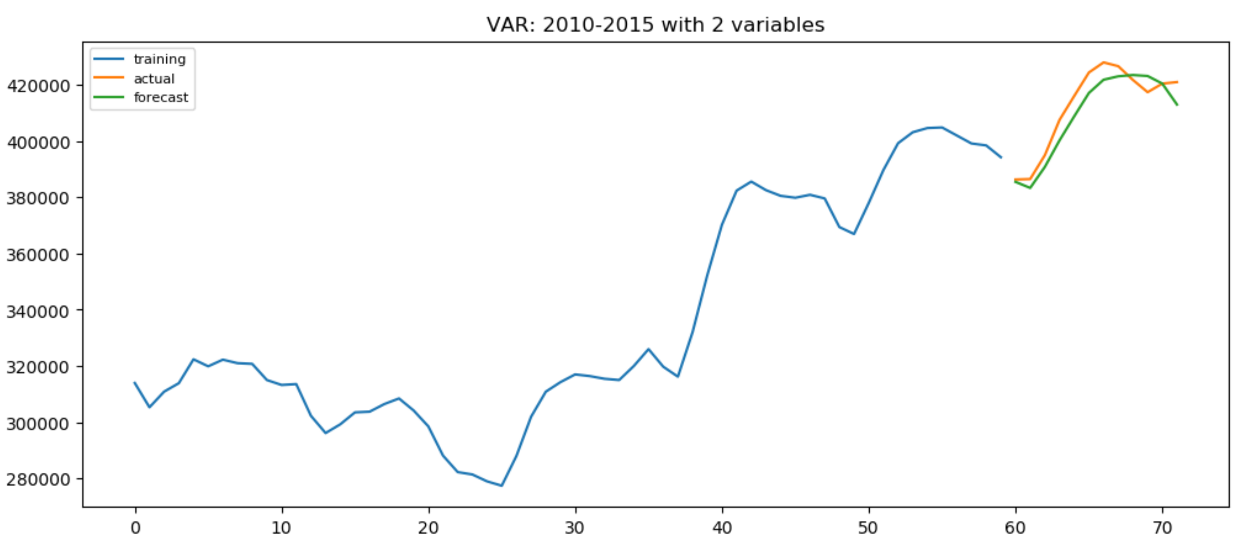
VAR

We also fit a VAR model for all versions of the training sets. As stated earlier, the versions of the training sets are training years from 2004-2015 using the two non-missing variables, training from 2010-2015 using the two non-missing predictive variables, and training from 2010-2015 using all three predictive variables.

From the prediction plots below, for training from 2004-2015, we can see that the forecast does doesn't follow the path at all, and under-estimates the actual values. The RMSE obtained from this model is 31423.47.



From the next set of prediction plots below, for training from 2010-2015, using the two non-missing predictive variables, we can see that the forecast does follow the path of the actual results, but under-estimates the actual in the beginning, but over-estimates it at the end. The RMSE obtained from this model is 5311.28.

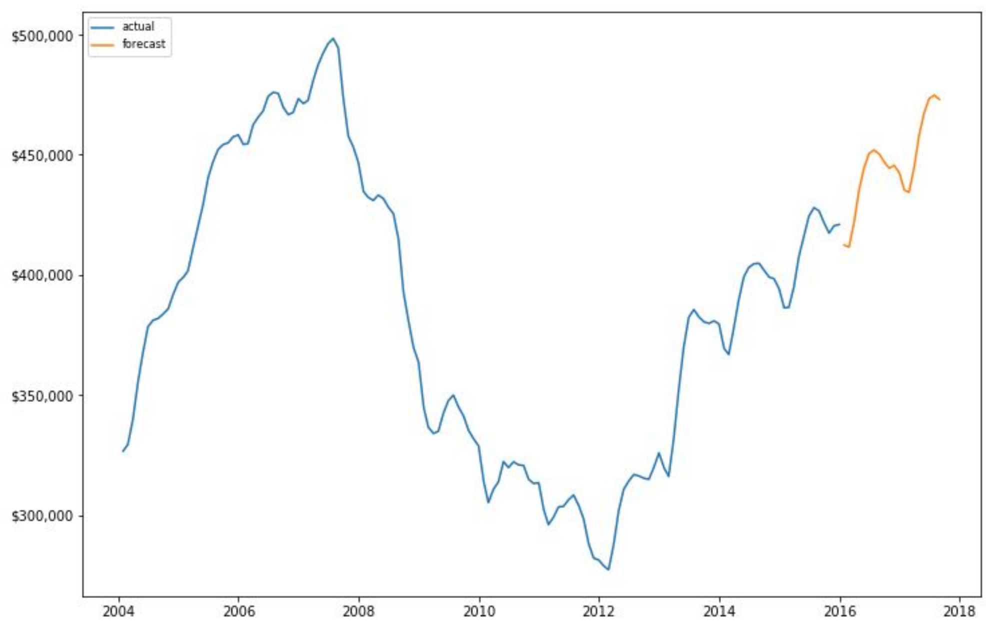


Conclusion

From these models we have fitted, we decided to go with exponential smoothing. This model has the smallest RMSE and the “forecast” line follows the “actual” line the closest. It seems to pick up on the seasonal behavior the best, whereas some of the other models just have straight lines for the forecast prediction.

Prediction

Using the model we selected (exponential smoothing) we predicted future values of the Median Sold Price. A summary and graph of the prediction of future values is shown below:



Date	2016-01	2016-02	2016-03	2016-04	2016-05	2016-06	2016-07
Prediction	$412,463	$411,603	$421,928	$435,306	$444,648	$450,536	$451,960
Date	2016-08	2016-09	2016-10	2016-11	2016-12	2017-01	2017-02
Prediction	$450,266	$447,050	$444,406	$445,593	$442,432	$435,301	$434,442
Date	2017-03	2017-04	2017-05	2017-06	2017-07	2017-08	
Prediction	$444,766	$458,145	$467,486	$473,374	$474,798	$473,104	

Proportion of work

Group Members	Sean	Sunny	Andrew	Geoffrey	Jacob
Proportion of Work	20%	20%	20%	20%	20%
List of Work	Discussion of Procedures and Methods EDA Fit Sarimax Model Fit VAR Model Prediction	Discussion of Procedures and Methods Write Report Code Review	Discussion of Procedures and Methods Write Report Code Review	Discussion of Procedures and Methods EDA Fit Sarimax Model Fit VAR Model	Discussion of Procedures and Methods EDA Fit Sarima Model Fit Exponential Smoothing Model