Time Series
Final Project

Zillow Homes Median Sold Price

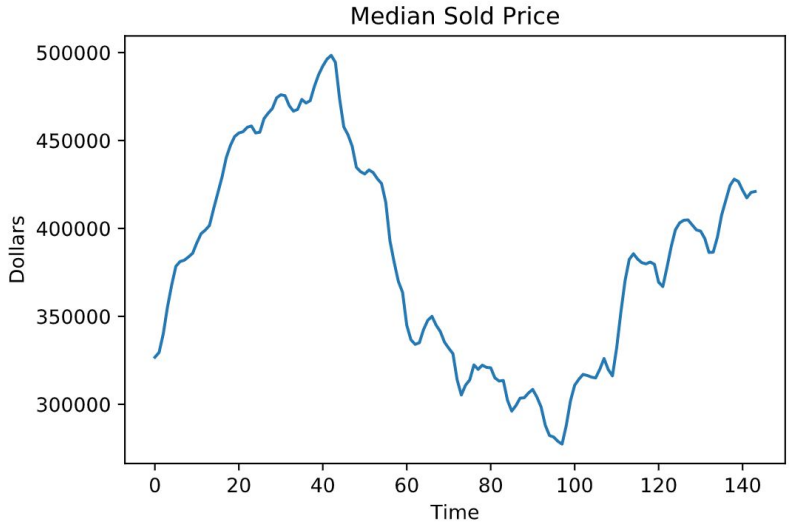
Group 1

Akanksha, Shreejaya Bharathan, Wendeng Hu, Ariana Moncada

DATA DESCRIPTION

The dataset used for our analysis is a Zillow dataset that contains information regarding monthly median sold price and related features for housing in California. The data is collected is monthly starting January 2004 till December 2015. The features present in the dataset are:

1. Date
2. Median Sold Price
3. Median Mortgage Rate
4. Unemployment Rate
5. Median Rental



The objective of this report is to forecast the monthly median sold price of California homes for the time period January 2016 through August 2017.

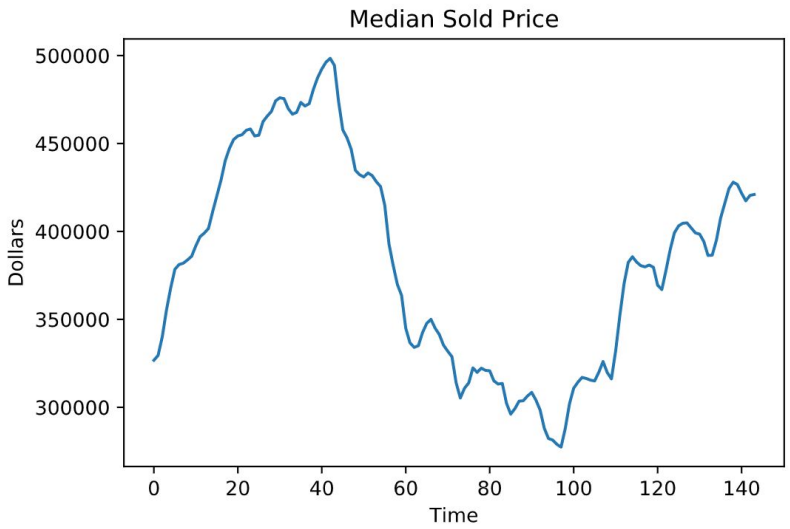
MODELING METHOD OVERVIEW

The following time series models were considered to predict median sold price:

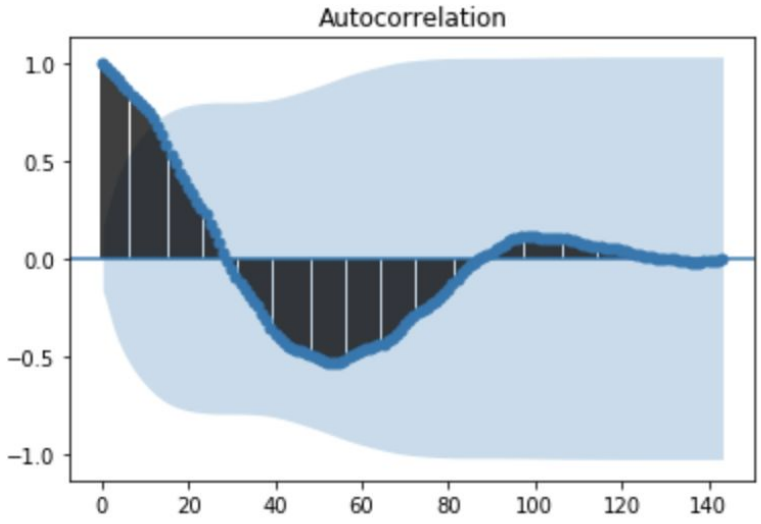
1. **Univariate Models - Median Sold Price**
    a. SARIMA Model
    b. TES (Exponential Smoothing) Model
2. **Multivariate Models - SARIMAX**
    a. Median Sold Price ~ Median Mortgage Rate
    b. Median Sold Price ~ Unemployment Rate
    c. Median Sold Price ~ Median Mortgage Rate + Unemployment Rate
3. **Multivariate Models - VAR**
    a. Median Sold Price ~ Median Mortgage Rate
    b. Median Sold Price ~ Median Mortgage Rate + Unemployment Rate
4. **Extra: Univariate Structural Break Median Sold Price**
    a. SARIMA Model
    b. TES (Exponential Smoothing) Model

INITIAL DATA EXPLORATION

Our first task before proceeding to modeling is to investigate the nature of the median sold price variable (the variable we are predicting for). Looking at the line plot for median sold price, we noticed that the median price fluctuates a lot throughout the 11 years. We see that it begins with an increasing trend then mid way in the series, experience a decrease in trend.



We can also confirm this by looking at the autocorrelation (ACF) plot below. The ACF plot decays slowly, which indicates the trend we noticed above in the line plot. Again, we can't determine the seasonality at this point.



This also tells us that the series is not a stationary process, indicating some work must be done to understand the data more before we proceed to modeling.

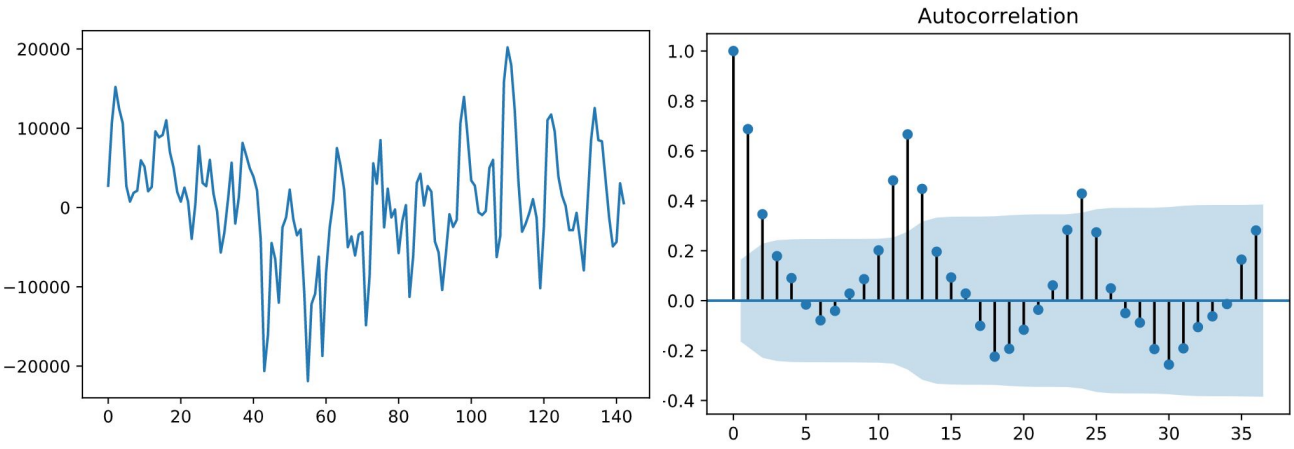
An Augmented Dickey-Fuller (ADF) test was performed to confirm the non-stationarity.

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.768391
p-value                          0.062933
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

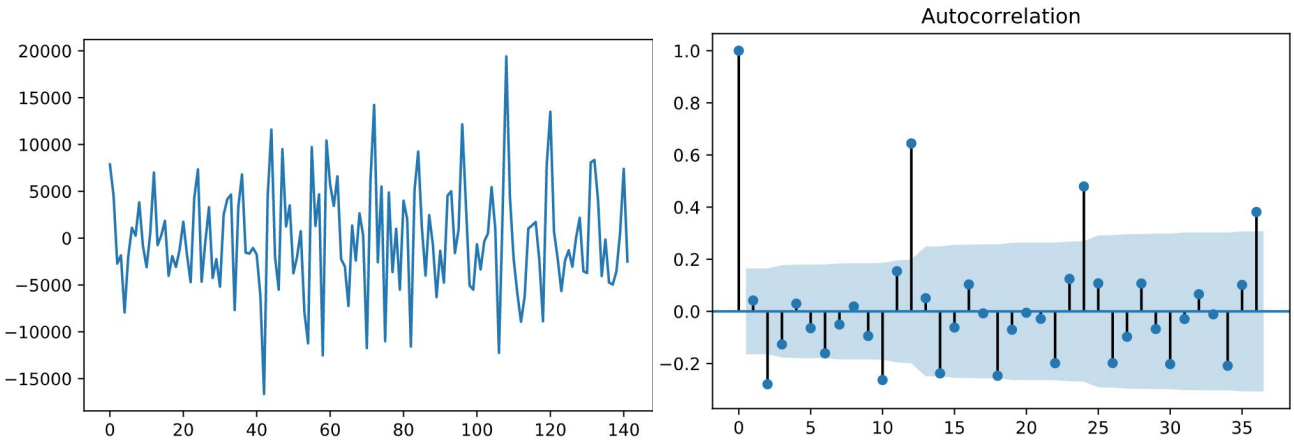
The null hypothesis for ADF test is that the process is not stationary. The alternative hypothesis for ADF test is that the process is stationary. With p-value 0.06293 which is higher than 0.05, confirming a non-stationary process.

DIFFERENCING

Next, we will utilize data differencing to eliminate trend in hopes to get a stationary process.



After one time differencing, we can see that there still exists some trend left in the data and as confirmed by the ADF test with a p-value of .43766, we still do not have a stationary process. Looking at the ACF plot, we can also notice at this point the obvious seasonality with lag = 12.



Looking at the plot above, we now see that after taking the second difference of the data, we have removed the trend. Notice, the seasonality still exists and we will proceed to perform seasonal differencing to remove seasonality. With a one time seasonal differencing with lag=12, we were able to achieve a stationary process.

Before we proceed to modeling with the knowledge acquired in this section we will split the data before January 2016, assigning 75% of the data for training purposes and leave 25% of the data for validation between models. Lastly, we will choose the model with the smallest validation error as our final model and then proceed to forecasting.

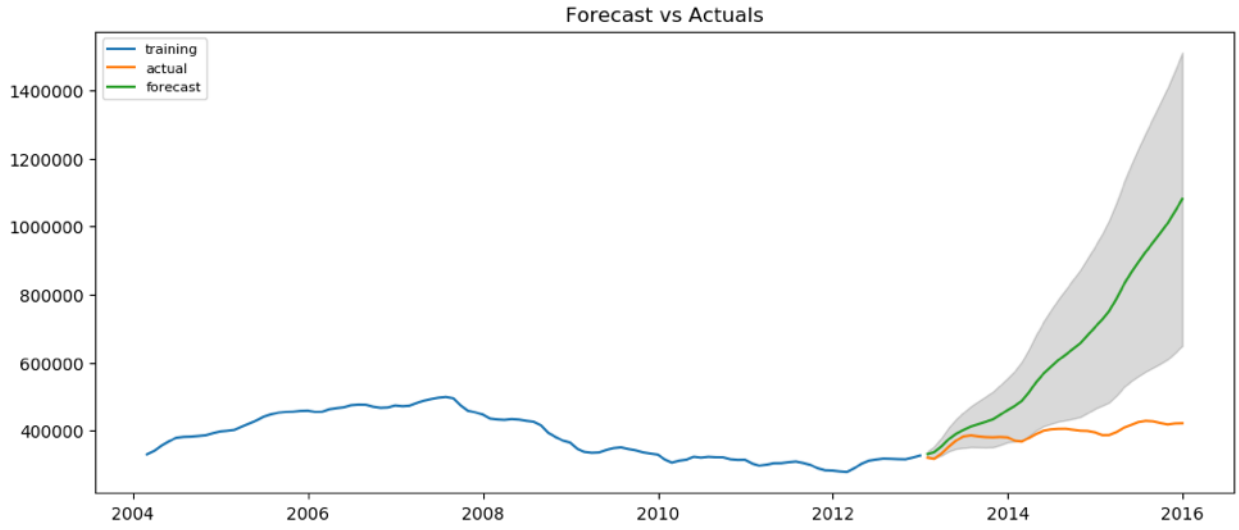
MOVING FORWARD

Now that we got a better sense of the trend and seasonality in our monthly median sold price of Californian homes, we first investigate how the univariate models like SARIMA and Exponential Smoothing perform on the data. Additionally, we will explore how external exogenous and endogenous variables like monthly *unemployment rate* and *median mortgage rate* predict median sold price.

UNIVARIATE MODELS

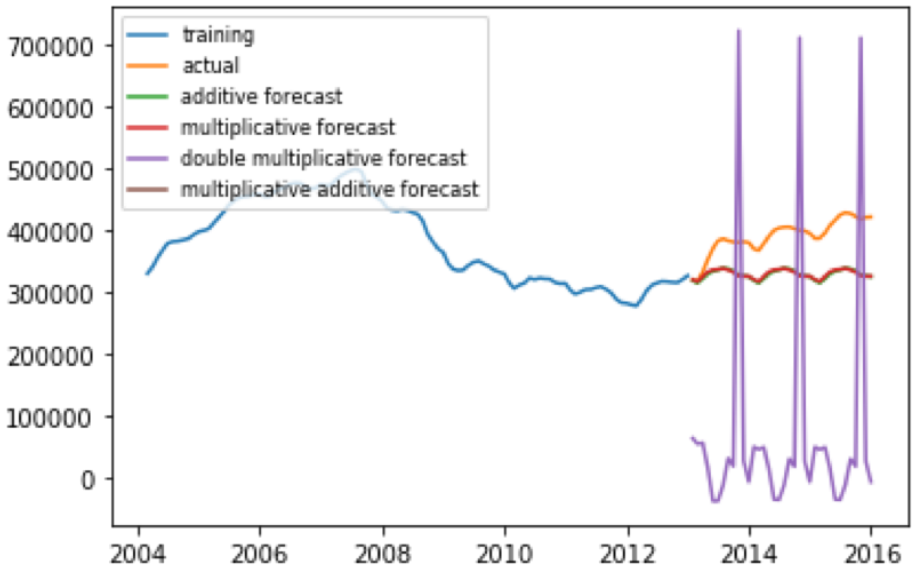
SARIMA

Based off of the differencing performed in the initial data exploration, we identified the correct differencing to remove trend was d=2 and to remove seasonality D=1. Using auto arima for model selection, the best model chosen based on BIC was a SARIMAX(2, 2, 1)x(0, 1, 0, 12) model. The model was evaluated using our validation set and received a Root Mean Squared Error (RMSE) value of **307111.42**. The prediction plot for the test data set was plotted as below.



TES

Since the line plot for median sold price experienced a drastic change in trend from positive to negative, it wasn't as obvious to identify whether the trend was additive or multiplicative. We suspected a multiplicative seasonality but decided to might as well test for both multiplicative and additive seasonality as well.



Four different TES models were evaluated. Model I is an additive trend and seasonality. Model II is an additive trend and multiplicative seasonality. Model III is a multiplicative trend and seasonality. Model IV is a multiplicative trend with an additive seasonality. The table below shows the respective RMSE values for the four models, highlighting the one with the lowest RMSE value.

TES Model	RMSE
I	66504.55
II	64851.94
III	NaN
IV	66092.52

Since the model with both multiplicative trend and seasonality was skewing the range, we plotted just models I and II with the actual values. With the nature of TES to weight the most recent previous values, it assumed the same pattern and did not extrapolate the exact magnitude of the jump in the actual data.

MULTIVARIATE MODELS

SARIMAX

In these models, we consider the external variables other than median sold price to be exogenous i.e. they influence the prediction of median sold price of a home but not the other way around.

*Predict Median Sold Price with respect to **Median Mortgage Rate:***

Based off of the differencing performed in the initial data exploration, we identified the correct differencing to remove trend was d=2 and to remove seasonality D=1.

Here, we consider *median mortgage rate* to be the exogenous variable and fit our model on the training data. Using auto arima for model selection, the best model chosen based on BIC was a SARIMAX(1,2,1)x(0,1,0,12) model. The model was evaluated using our validation set and received a RMSE value of **17690.74.**

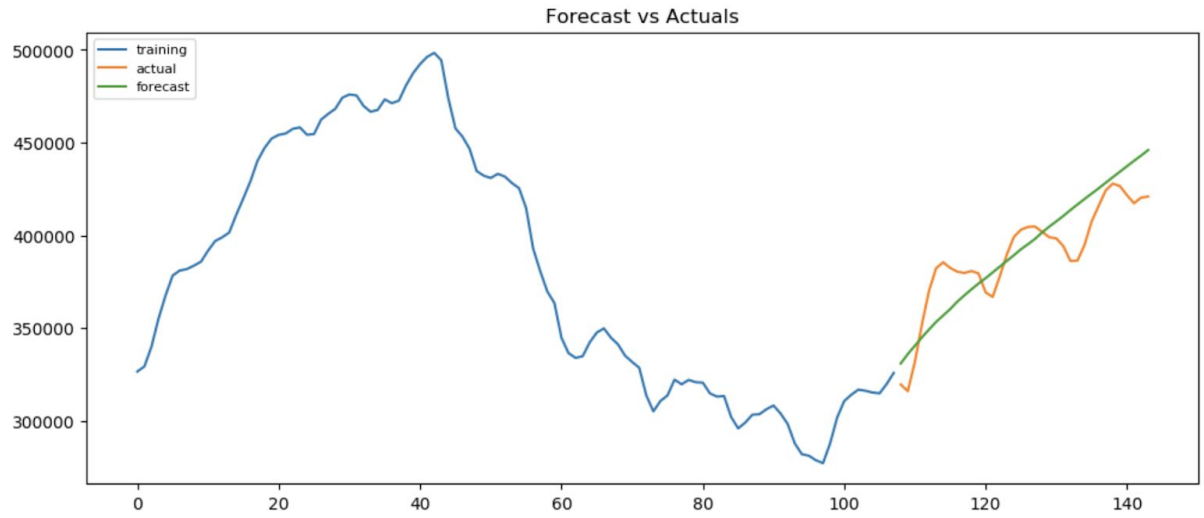
*Predict Median Sold Price with respect to **Median Unemployment Rate:***

Based off of the differencing performed in the initial data exploration, we identified the

correct differencing to remove trend was d=2 and to remove seasonality D=1

Here, we consider *median unemployment rate* to be the exogenous variable and fit our model on the training data. Using auto arima for model selection, the best model chosen based on BIC was a SARIMAX(1,2,1)x(0,1,0,12) model. The model was evaluated using our validation set and received a RMSE value of **16105.74**.

This model does better than the SARIMAX model using median mortgage rate, but still has a high RMSE. Below is a prediction plot with the actual and forecast for the validation set. We see that is was able to forward that trend very well.



Predict Median Sold Price with respect to ***Median Mortgage Rate and Unemployment:***

Seeing that both of the SARIMAX models above performed reasonably well separately, we wanted to see if both of their influence together performed any better than alone.

Again, using the trend differencing of d=2 and seasonality differencing of D=1, the two exogenous variables used are *median mortgage rate* and *unemployment rate* to predict median sold price on the training data. Using auto arima for model selection, the best model chosen based on BIC was a SARIMAX(1,2,1)x(0,1,0,12) model. The model was evaluated using our validation set and received a RMSE value of **17690.74.**

Vector AR (VAR)

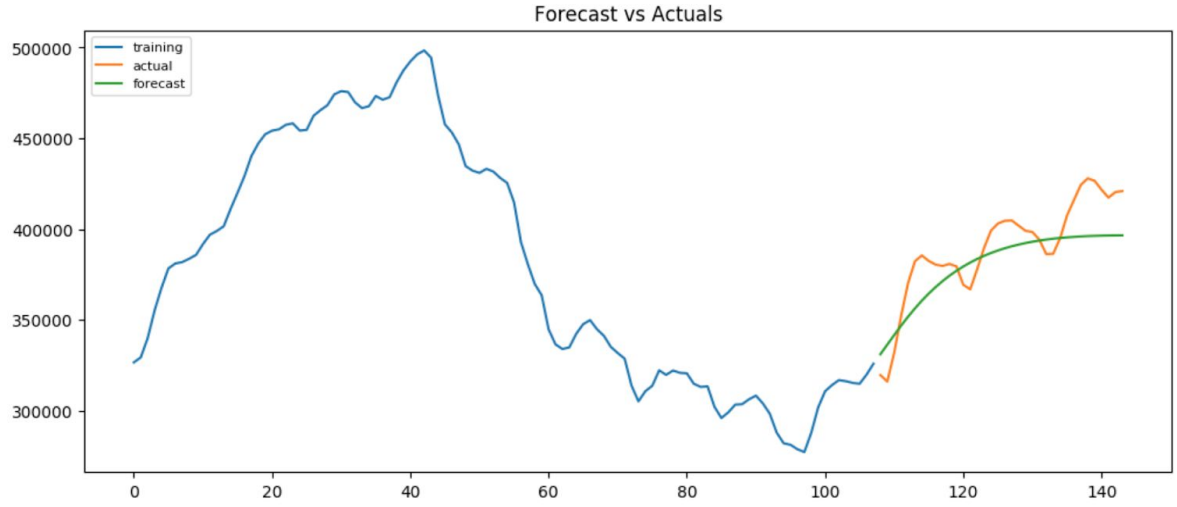
In these models, we consider the variables other than median sold price to be endogenous i.e. they are correlated with the median price of a home and have a two way influence.

Predict Median Sold Price with respect to ***Median Mortgage Rate:***
Our basis for choosing median mortgage rate to be correlated with median sold price was due to the idea that if a home has a high mortgage rate it may consequently have a high sold price. Additionally, if a price is sold for more (worth more) it could also influence the mortgage rate of a home.

First, we fit a VAR(1) model to the training data, and then used the fitted model to forecast the validation set. Computing the RMSE for the validation data we got a value of **32122.98**. Additionally, we want to test other orders of the VAR model and decided to check up to an order=3 VAR model.

VAR Model	RMSE
VAR(1)	32122.98
VAR(2)	14090.08
VAR(3)	21090.75

Below is the forecasted plot for the median sold price for the VAR order 2 model. Although, not too well at capturing the seasonality, it did predict the overall line shape well.



Predict Median Sold Price with respect to ***Median Unemployment Rate:***

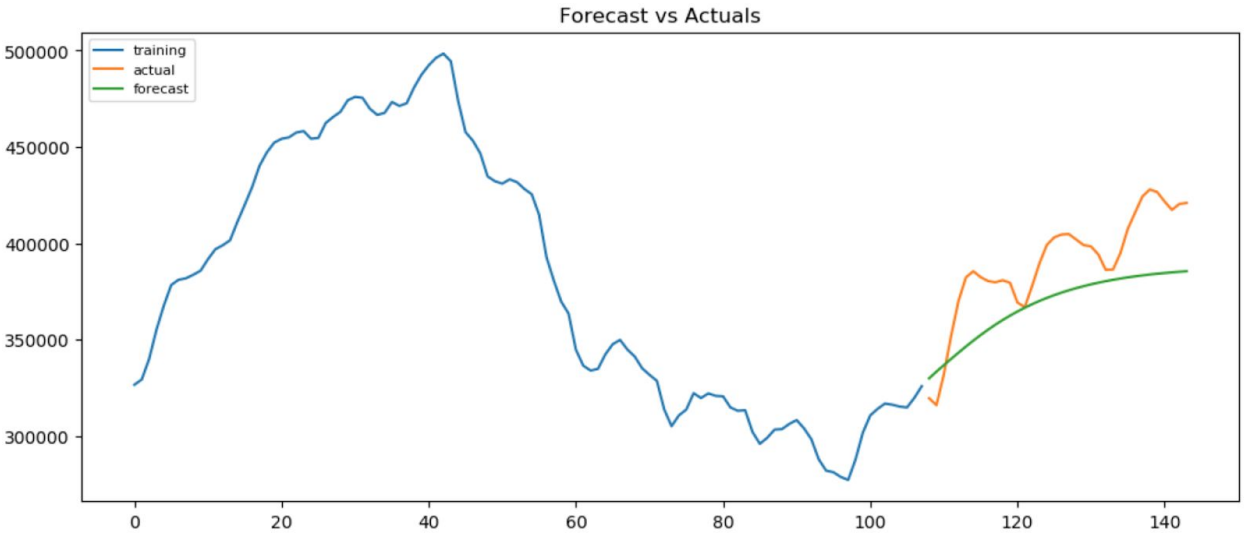
We did not fit a VAR model using unemployment rate to predict median sold price of home because it did not appear to us that a median sold price of homes will influence the median unemployment rate for that month.

Predict Median Sold Price with respect to ***Median Mortgage Rate and Unemployment:***

Although, we mentioned that median sold price cannot affect unemployment rate alone, we were curious to see how a VAR model with median mortgage rate and unemployment rate would perform. For this model we consider the predicting median sold price from the variables *median mortgage rate* and *unemployment rate* on the training data. We decided to try multiple AR orders=1, 2, and 3 for this combination of variables. Results are as follows:

VAR Model	RMSE
VAR(1)	51980.44
VAR(2)	26080.43
VAR(3)	47200.498

Amongst the VAR models we identified that a model of order 2 was best. Below is the plot for the VAR(2) model. As we saw earlier in the VAR model using just mortgage rate, this VAR model is only able to capture the overall increasing shape and not the seasonality.



UNIVARIATE STRUCTURAL BREAK MODELS

Structural Break Test

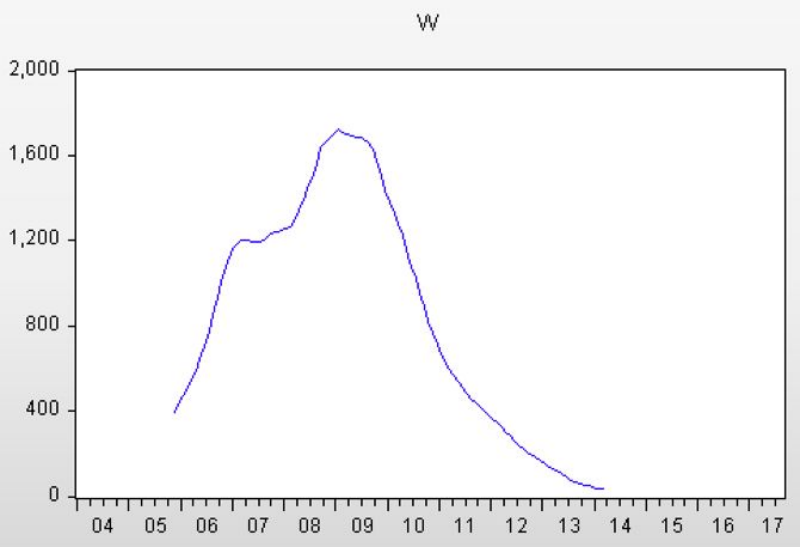
Looking at the time series plot for median house price above, it seems that structural breaks exist here. To identify the existence of a structural break we used the Quandt-Andrews test, which utilizes hypothesis testing to determine the proper structural break in the data.

Quandt-Andrews unknown breakpoint test
Null Hypothesis: No breakpoints within 15% trimmed data
Varying regressors: All equation variables
Equation Sample: 2004M01 2015M12
Test Sample: 2005M11 2014M03
Number of breaks compared: 101

Statistic	Value	Prob.
Maximum LR F-statistic (2009M01)	122.9733	0.0000
Maximum Wald F-statistic (2009M01)	1721.626	0.0000
Exp LR F-statistic	58.29053	0.0000
Exp Wald F-statistic	856.1981	0.0000
Ave LR F-statistic	60.05837	0.0000
Ave Wald F-statistic	840.8172	0.0000

Note: probabilities calculated using Hansen's (1997) method

The Maximum statistic is a statistic that can provide valuable information. The large value of maxF statistic were detected. Its associated p-value is close to zero. The null hypothesis that there is no structural break can be rejected. Hence, it is highly likely that there exists a structural break. What's more, the data in the parenthesis (2009M01) is the date where the F-statistic gets its maximum value among the 15% trimmed data. The F-statistic over these dates were plotted further.

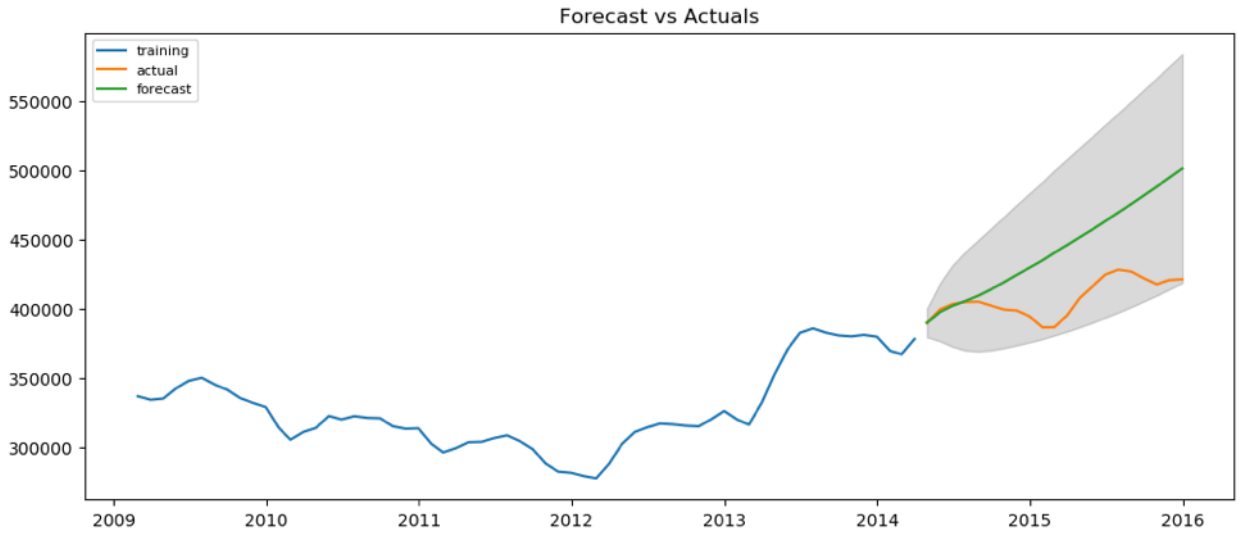


According to the F-statistic plot above, the F-statistic reaches the peak in 2009.

Based on the structural break test, I will use the data after 2009 (Month 1) to build my prediction model for forecasting.

SARIMA - *Structural Break*

Although we performed a SARIMA model on the structural break training dataset beginning after the year 2009, we did not see significant improvements between the above discussed models. Where the structural break dataset really shined was in the TES models (which we will discuss next). However, we were able to achieve a pretty good RMSE value of **43538.76** with a forecasted plot as seen below.

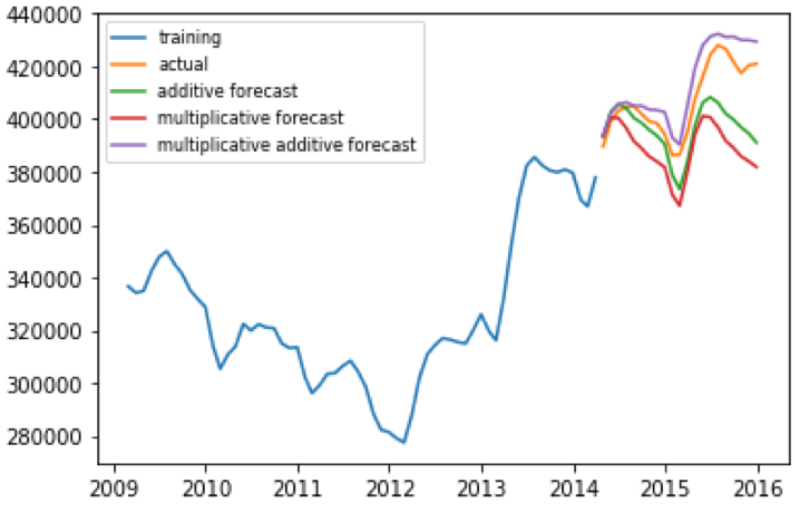


TES - *Structural Break*

Four different TES models were evaluated. Model I is an additive trend and seasonality. Model II is an additive trend and multiplicative seasonality. Model III is a multiplicative trend with an additive seasonality. The table below shows the respective RMSE values for the four models, highlighting the one with the lowest RMSE value.

TES Model	RMSE
I	14583.89
II	21674.04
III	7229.81

The prediction plot for models I, II, and III is plotted as below.



RESULTS

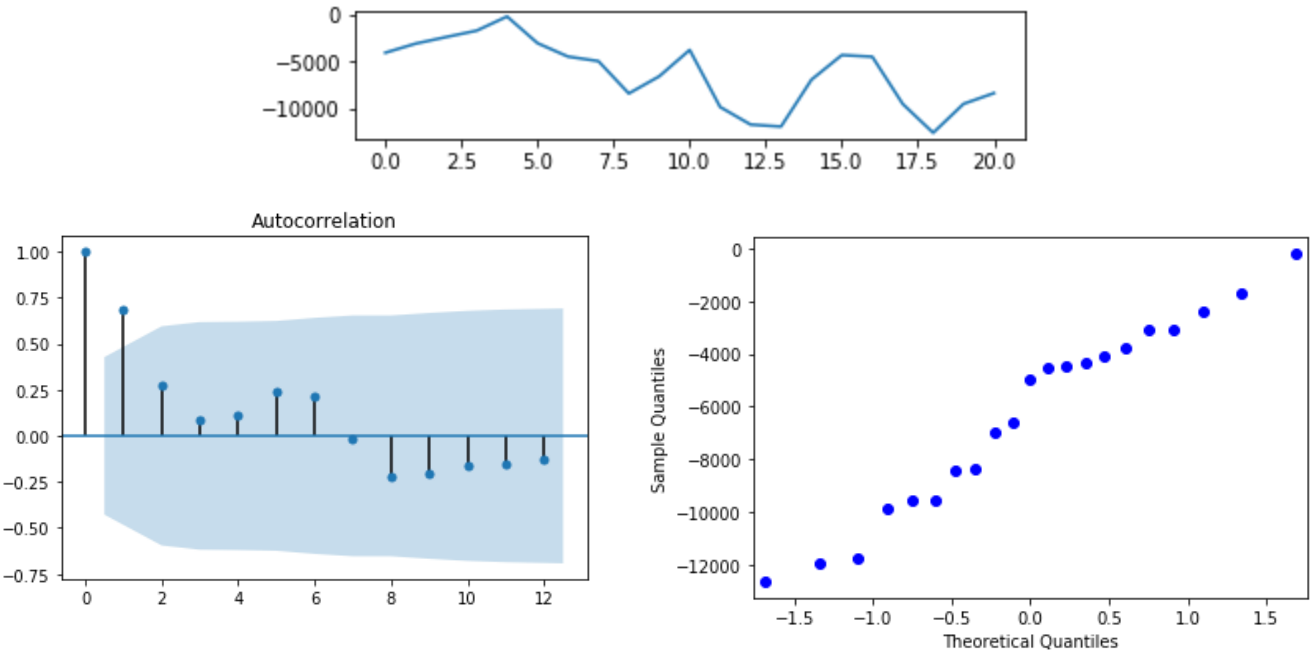
All models described above were evaluated using a validation set of 25% of the available data. The following table is the comparison Root Mean Squared Error (RMSE) scores for the validation set.

Models	RMSE
(Univariate of MedianSoldPrice)	
SARIMA	307111.42
TES I	66504.55
TES II	64851.94
TES IV	66092.52
SARIMAX *(Median MortgageRate)*	17690.74
SARIMAX *(Unemployment)*	16105.74
SARIMAX *(Median MortgageRate & Unemployment)*	17892.50
(MedianMortgage Rate)	
VAR(1)	32122.98
VAR(2)	14090.08
VAR(3)	21090.75
(MedianMortgageRate & Unemployment)	
VAR(1)	51980.44
VAR(2)	26080.43
VAR(3)	47200.49
(Structural Break)	
SARIMA	43522.61
TES I	14583.89
TES II	21366.22
TES III	7229.81

Based on the above table, the *TES Structural break* model was the model with the smallest validation RMSE score, so we will choose to use this model for forecasting median sold price for January 2016 through August 2017.

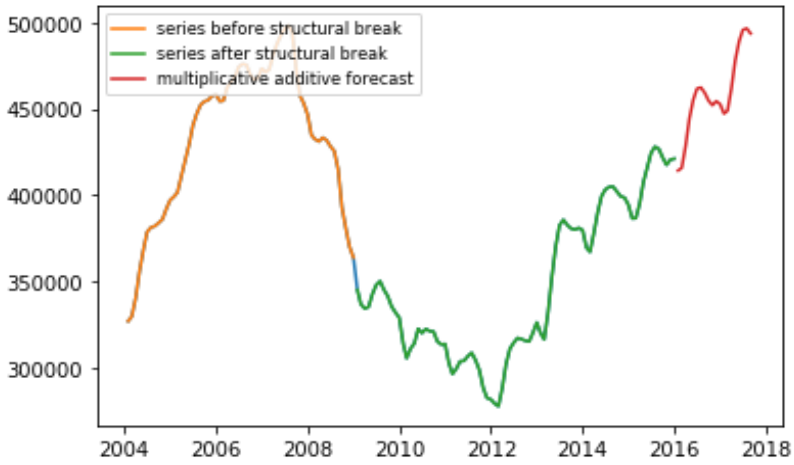
Residual Diagnosis for TES Structural Break Model

Residual diagnosis was conducted for the third TES structural break model.



Based on the residual line plot, the mean of the residual is close to a constant. The ACF plot shows that the residual is uncorrelated. The qq-plot shows that the residual follows the normality. So the residual is white noise with normality. This TES model follows the model assumption.

FORECASTING



CONCLUSION

We found that the structural break - TES model 3 (multiplicative trend and additive seasonality) had the lowest RMSE score. From a business perspective, we can see this as a shift in trends after the 2008 recession. Clearly, the models that consider the data after the structural break, perform better at predicting the future data.

CONTRIBUTION OF THE TEAM MEMBERS

Group members	Akanksha .	Wendeng Hu	Ariana Moncada	Shreejaya Bharathan
Proportion of work	25%	25%	25%	25%
List of work	- Discussion of problems and chosen methods - VAR models - Final discussion - Write the report	- Discussion of problems and chosen methods - SARIMA, TES - Structural break models - Final discussion -Write the report	- Discussion of problems and chosen methods - VAR model - SARIMAX model - Final discussion - Write the report	- Discussion of problems and chosen methods - SARIMAX models - Final discussion - Write the report