Zillow House Median Sold Price Predictions

Group Members:

1. Andy Cheon
2. Kevin Wong
3. Lin Meng
4. Roja Immanni

Table of contents:

1. Data description and Research question

The Zillow dataset consists of four variables:

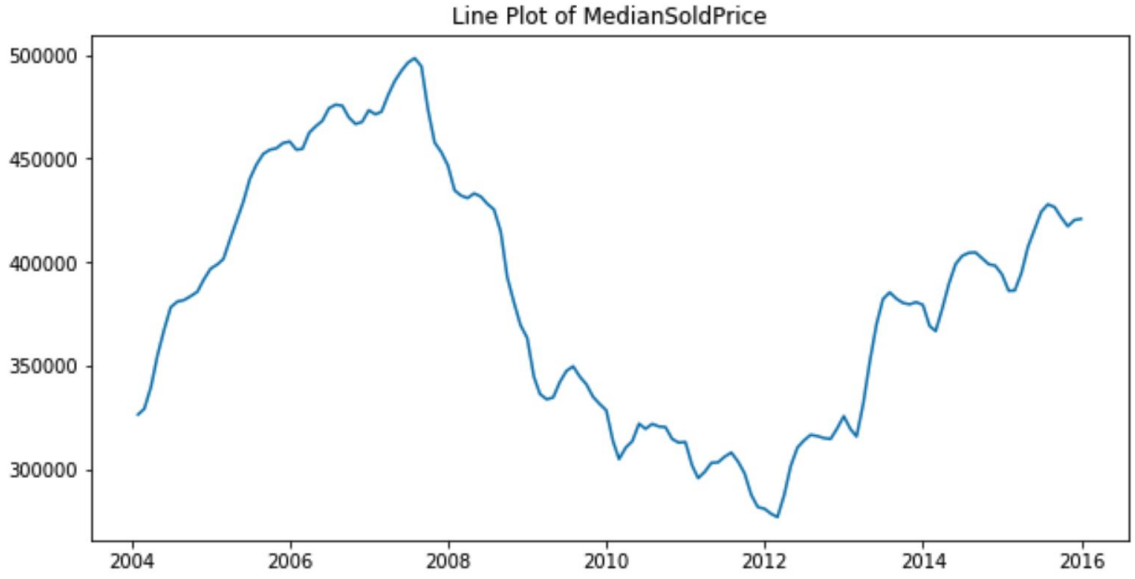
- Median rental price of all houses in California (predictor)
- Median mortgage rate (predictor)
- Unemployment rate (predictor)
- Median sold price of all houses in California (*target*)

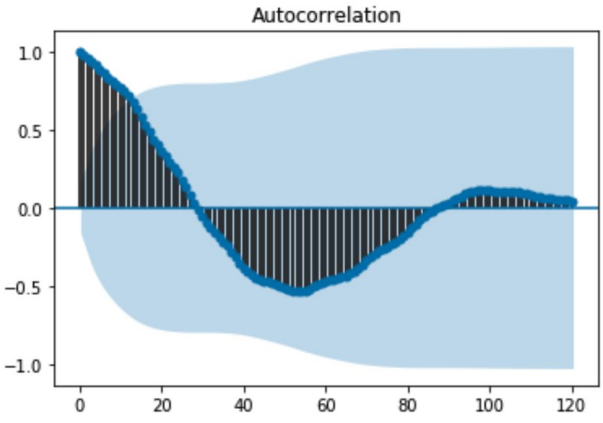
spanning January 2004 - August 2017. Data were collected at the end of each month, totaling 164 observations. Our goal is to forecast the median sold price of all homes in California from January 2016 through August 2017 (20 forecasted values).

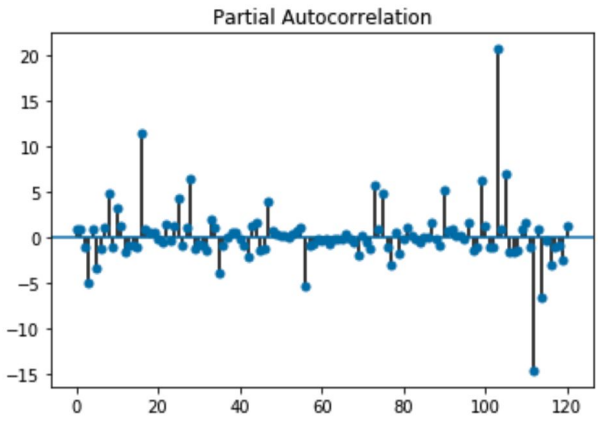
We first investigate the relationship between predictor variables and the target, as well as relationships among the predictor variables themselves. The intuition and domain knowledge gained from this step allows us to test out various time series models trained on different sets of predictors. We then identify which predictors, if any, contribute significantly to the success of our candidate models. Lastly, we choose a final model based on its predictive ability on our validation set and use this model to forecast median selling price.

2. Exploratory Data Analysis:

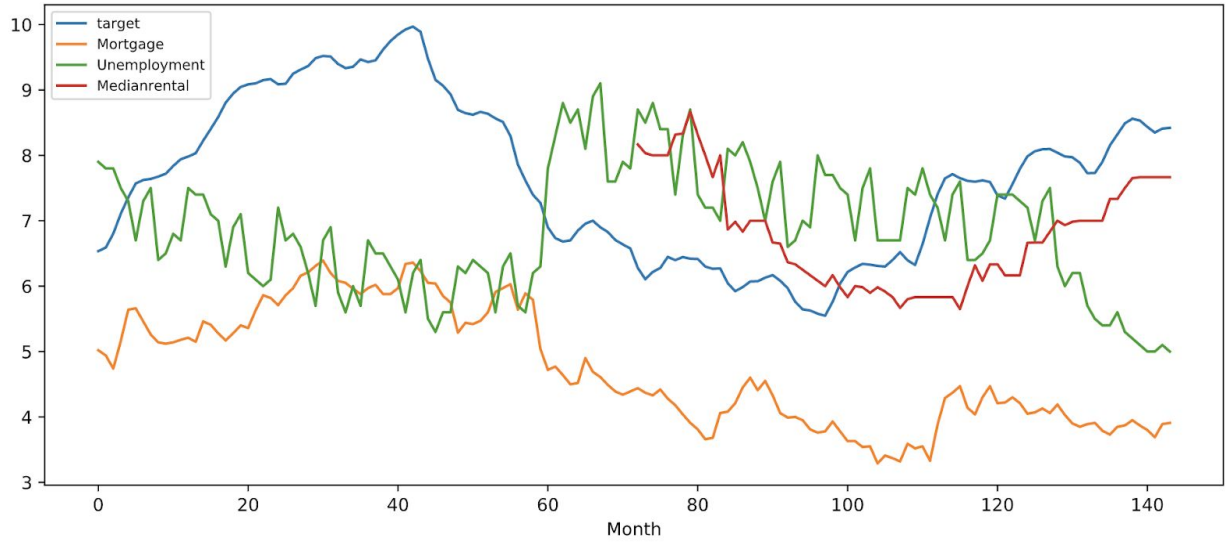
We first examine the line plot of the Median Sold Price from 2004 to 2016, which paints an overall picture of our target variable. Some interesting observations are the 2008 recession, the market rebound starting in 2012, and the three dips in 2014, 2015, and 2016 -- most likely market corrections. There are obvious trends (upward from 2004-2008, downward from 2008-2012, upward from 2012 onwards), and no obvious seasonality (although we show later in differencing plots that there is plausible annual seasonality). The pattern of the ACF plot is also a clear indicator of the trend in our original data.



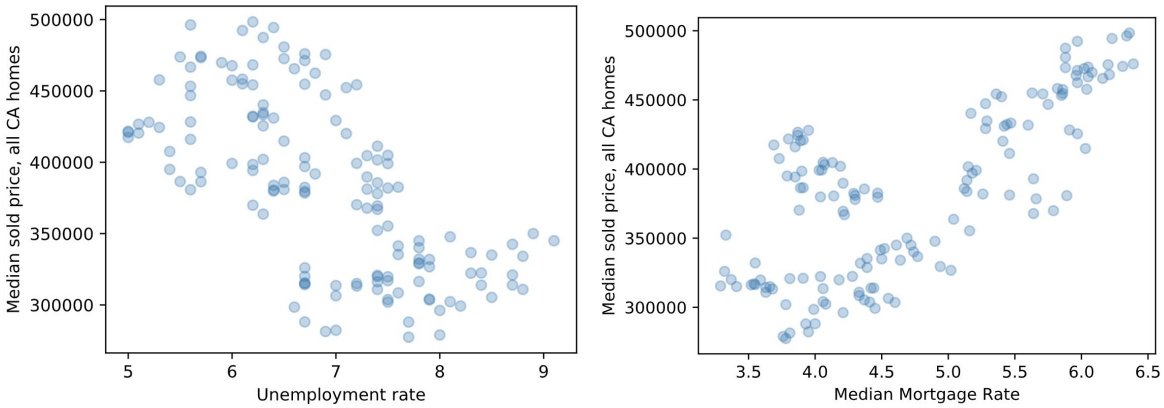




We then performed exploratory data analysis on our entire dataset to learn more about our given variables. From the line plot below, we observe that Median Mortgage Rate and Unemployment is closely associated with the target.

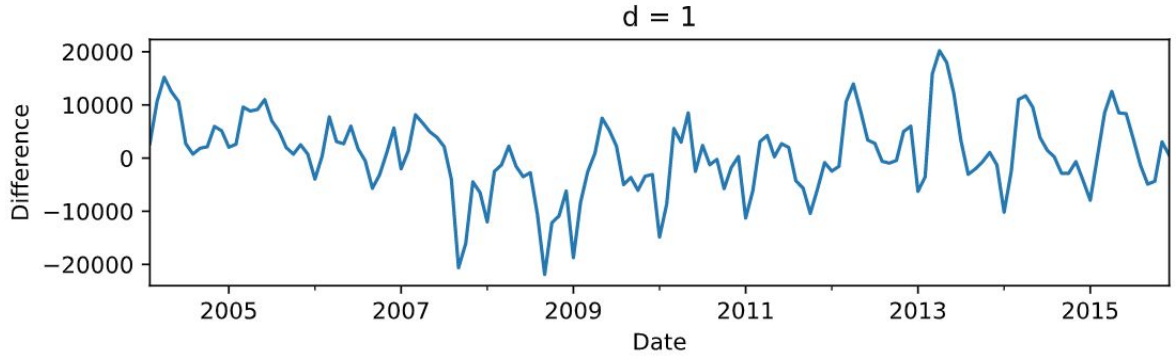


The scatterplots below help verify our beliefs. These may serve as useful variables in our multivariate models later on. We scaled the series so all show in the same range.



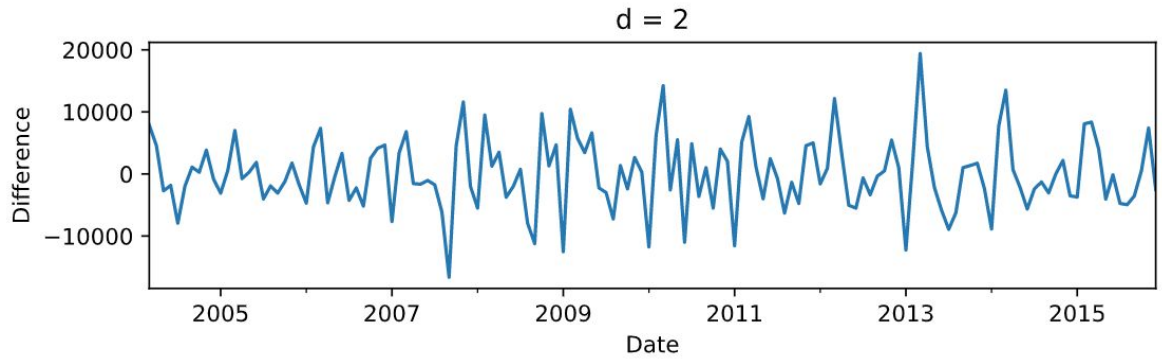
Further investigating relationships, scatterplots of unemployment and mortgage rates with the target -- Median Sold Price -- show decreasing and increasing correlations. These are likely useful to include in our latter multivariate models.

Trend & Seasonality: Time Differencing



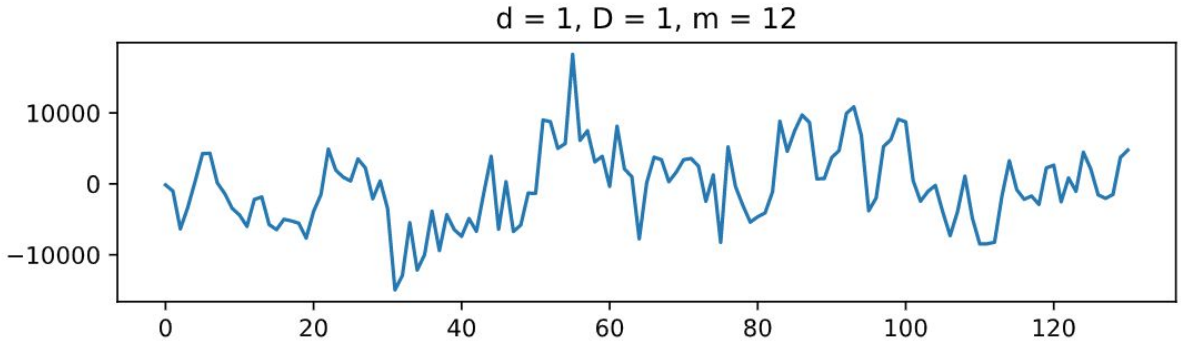

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.687305
p-value                          0.437665
#Lags Used                      12.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

After the 1-time trend differencing, a significant amount of trend has been removed from the data. To the naked eye, the data seems centered around 0, however, seasonality is still apparent (as expected) and the p-value for the stationary condition is not significant.



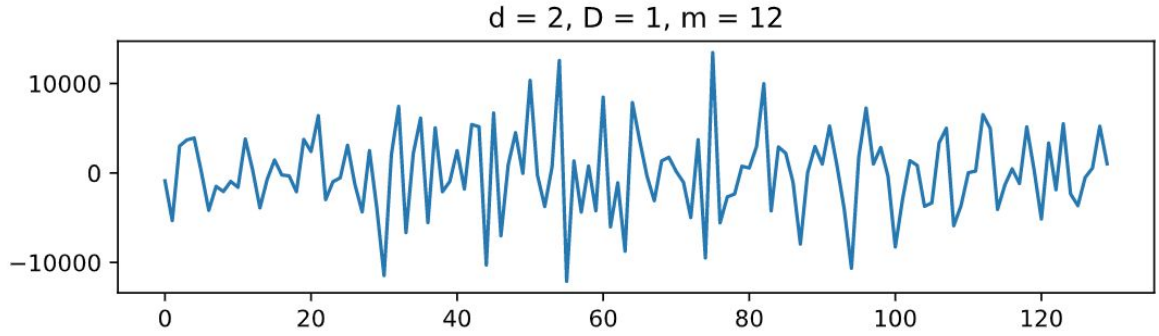

```
Results of Dickey-Fuller Test:
Test Statistic                  -4.918846
p-value                          0.000032
#Lags Used                      11.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

After the 2-time trend differencing, we see good results. The series appears even better centered at 0 and the p-value is clearly significant for the stationary condition.



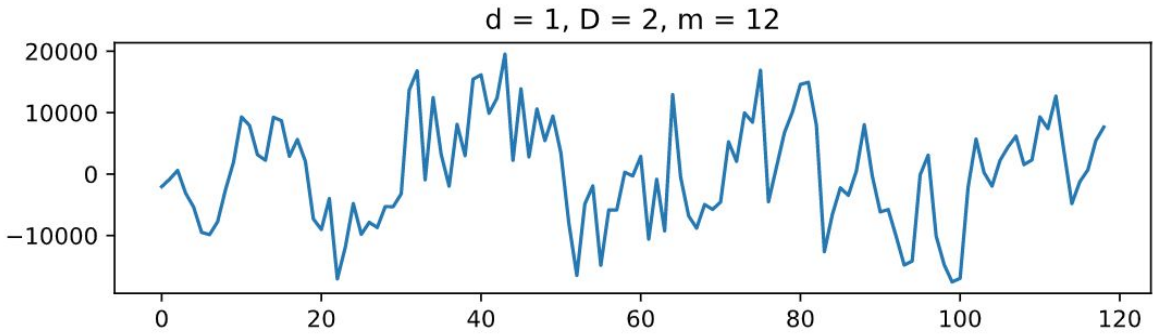

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.121323
p-value                          0.236028
#Lags Used                      13.000000
Number of Observations Used    117.000000
Critical Value (1%)             -3.487517
Critical Value (5%)             -2.886578
Critical Value (10%)            -2.580124
dtype: float64
```

Next, we focus on removing seasonality by choosing the correct D and m. We anticipate D = 1, but we try D = 2 for completeness. The plot above shows 1-time trend differencing with 1-time seasonality differencing with a seasonal lag of 12. The p-value is not, but approaching significance. We show later that the log-transformed target predictions perform best with these parameters.




```
Results of Dickey-Fuller Test:
Test Statistic                  -5.149454
p-value                          0.000011
#Lags Used                      13.000000
Number of Observations Used    116.000000
Critical Value (1%)             -3.488022
Critical Value (5%)             -2.886797
Critical Value (10%)            -2.580241
dtype: float64
```

With a 2-time trend differencing 1-time seasonality differencing, the p-value is significant.




```
Results of Dickey-Fuller Test:
Test Statistic                  -2.621665
p-value                          0.088602
#Lags Used                      13.000000
Number of Observations Used    105.000000
Critical Value (1%)             -3.494220
Critical Value (5%)             -2.889485
Critical Value (10%)            -2.581676
dtype: float64
```

We find later on that our best SARIMA model is, in fact, using d = 1, and D = 2. We show the trend-differenced plot and ADF test results above. Visually and statistically, we would not conclude to use these parameters, but based on the test data, this has the best results.

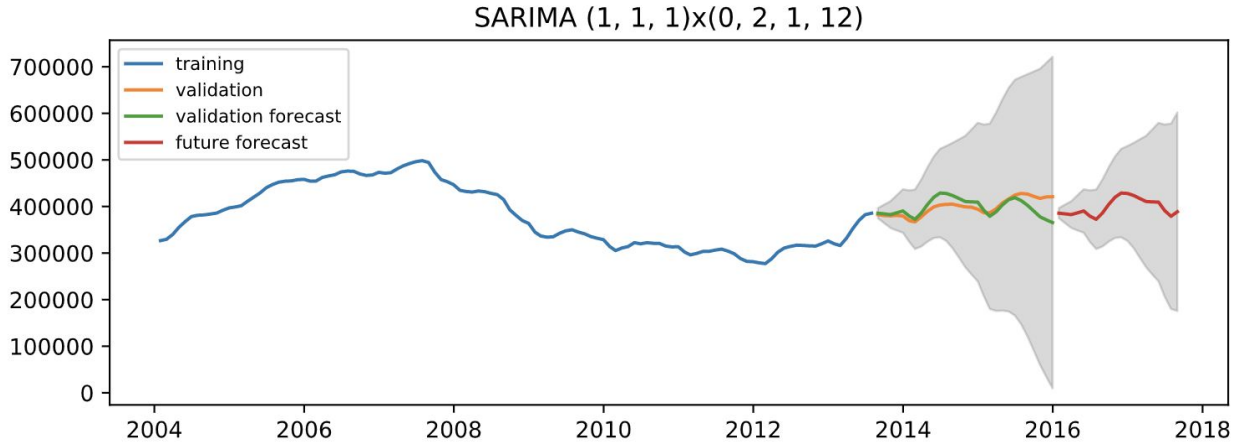
3. **Model Selection process:**
   To choose the best model in each category, several models are evaluated within each category using a range of category-specific parameters:

   - For SARIMA class models, we try different parameters of d and D, varying raw and log-transformed target data.
   - For exponential smoothing models, we focus on Triple Exponential Smoothing and evaluate combinations of additive and multiplicative trends and seasonality.
   - For SARIMAX, we try different combinations of exogenous variables and use auto-arima selection to see which model works best.
   - For VAR, to find the best VAR model, we tried every combination of the X variables and parameters p to choose the best model

   To evaluate and select the best models, we split data using the train-test split into training and validation sets on an 80/20 split. We calculate our models' predictions and RMSE on the validation set and choose the model with the lowest RMSE.

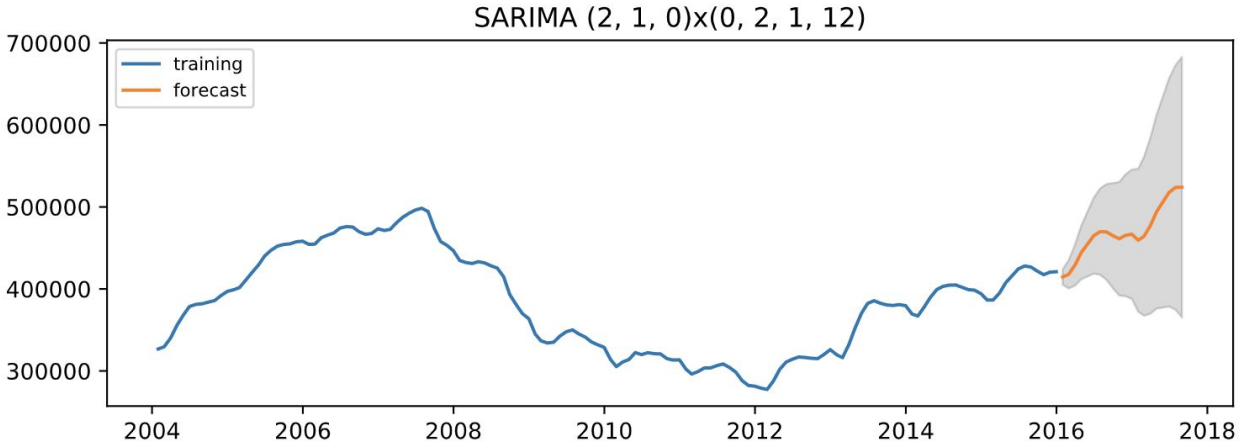
Best SARIMA Model

For SARIMA models, we iterate through d = 1 to d = 3, and D = 1 to D = 2. We set the seasonality m to 12 as we use seasonal time differencing. To our surprise, the best performing SARIMA model.



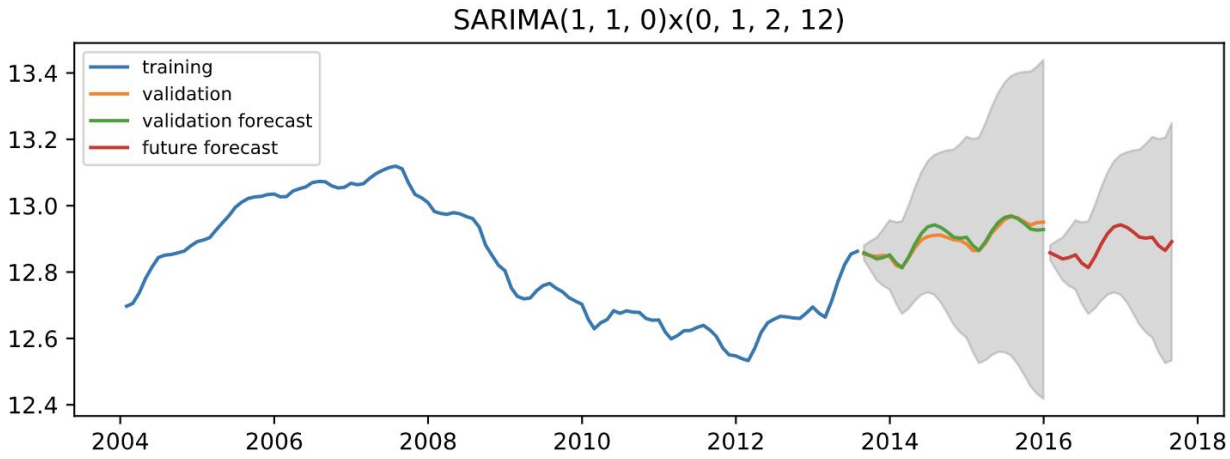
Validation RMSE: 20606.54

Forecast Visualization:



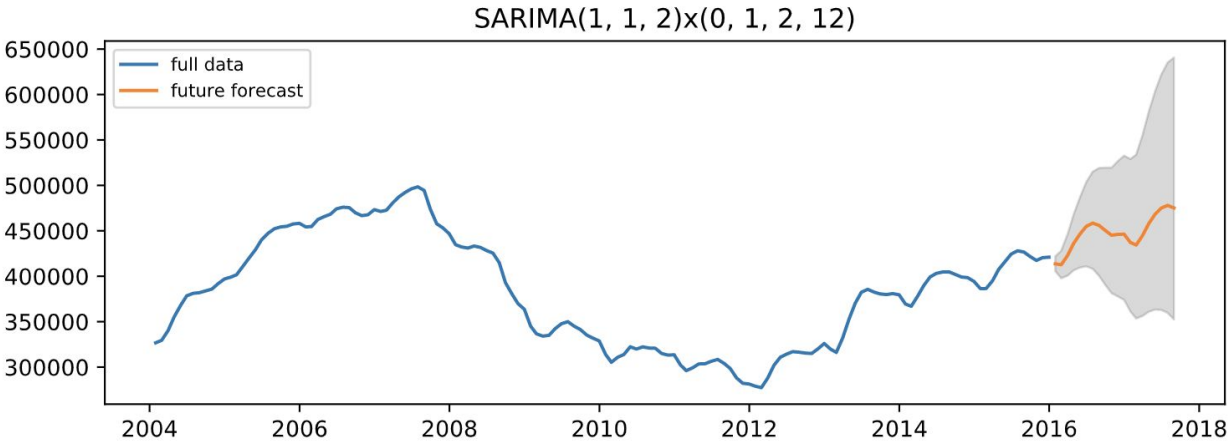
Best Log-Transformed SARIMA Model

For each of the above, we also train and forecast models using log-transformed target data. In this case, we find d = 1 and D = 1 gives the lowest error on validation data.



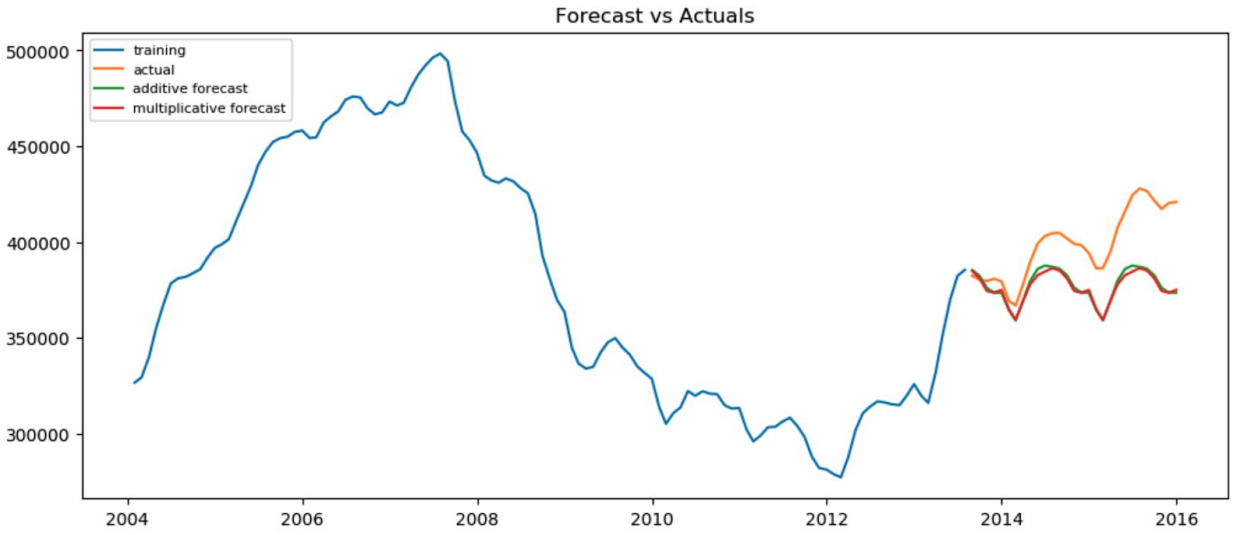
Validation RMSE: (transformed-back): 5640.62

Forecast Visualization (transformed-back):



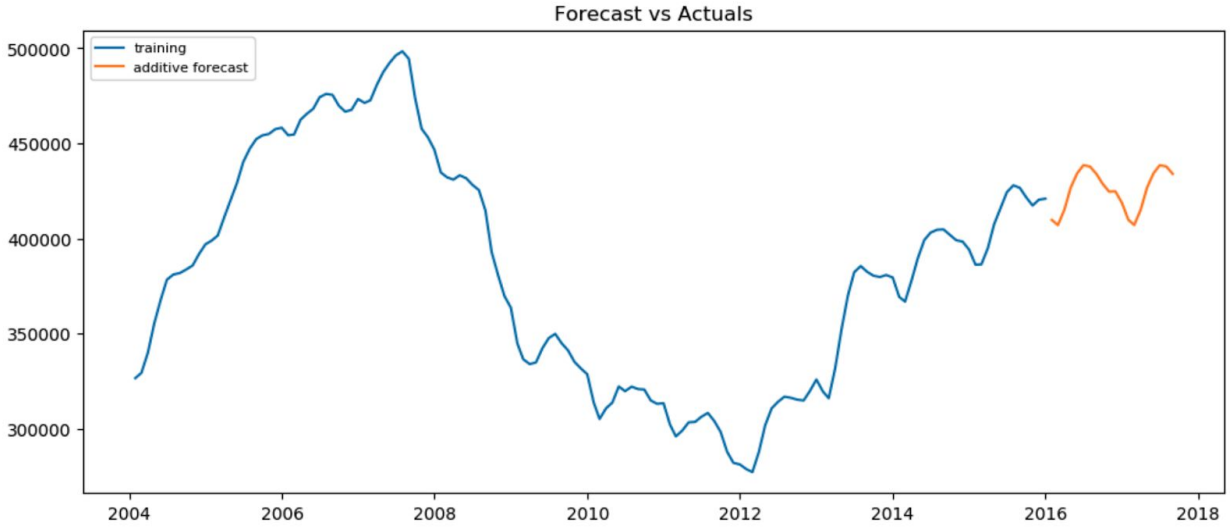
Best Exponential Smoothing

We also tried Exponential Smoothing using both the additive and multiplicative models. They had similar performances based on RMSE and prediction. In the end, the additive-additive model performed slightly better based on RMSE.



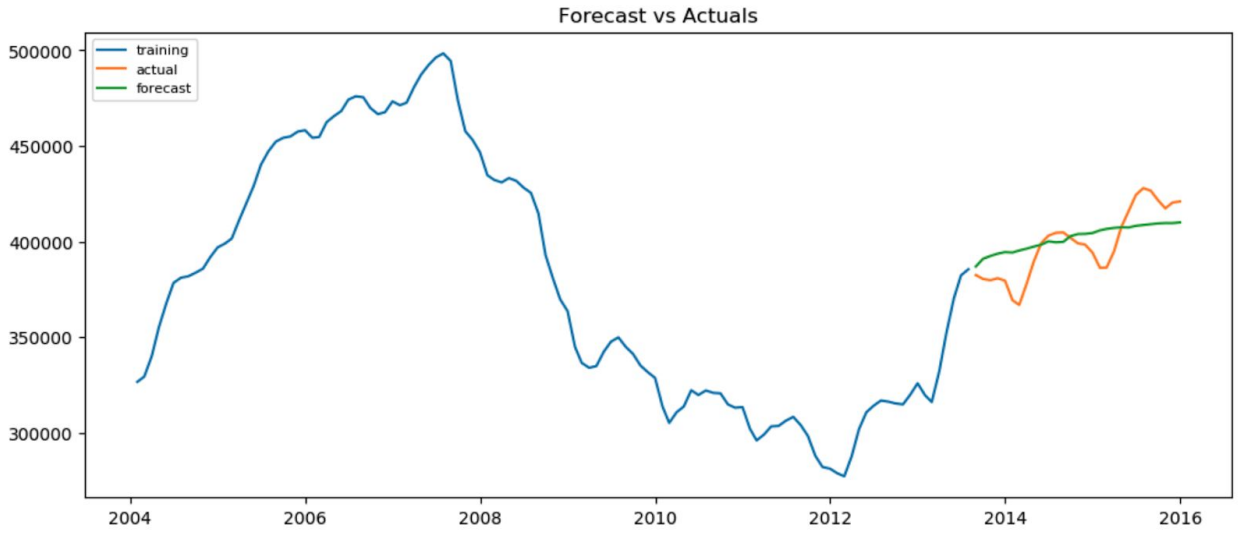
Validation RMSE: 25716.92

Forecast Visualization:



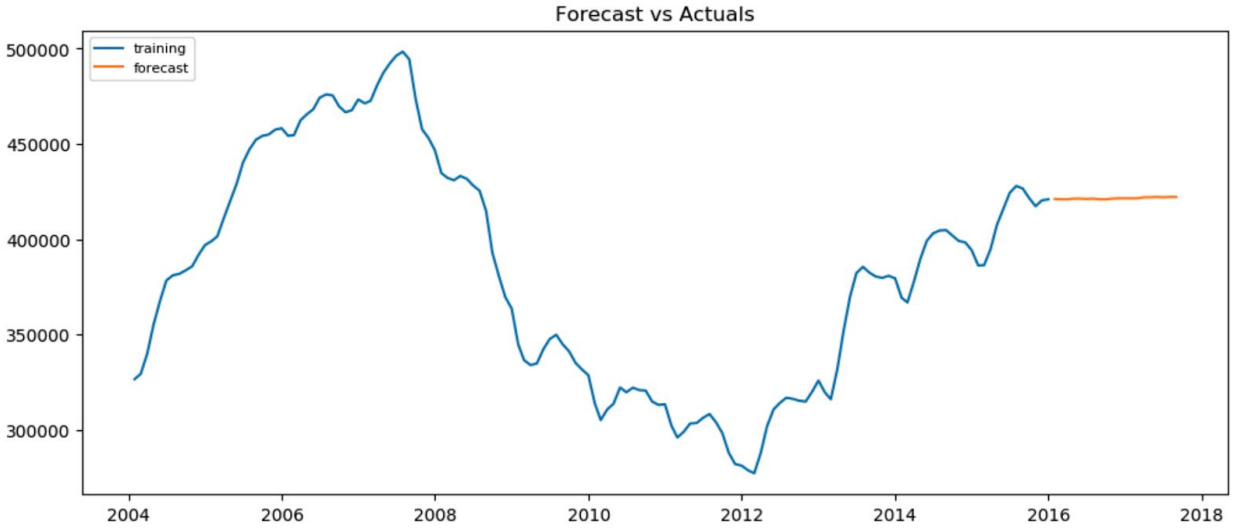
Best SARIMAX

Taking into consideration the “unemployment” rate and “mortgage” rate, we fit the SARIMAX model since the variables have some correlation with each other. It might help using other columns to predict our target. It turns out using only “unemployment” as X did a decent job. The model achieved validation RMSE at 13287.87.



Validation RMSE: 13287.87

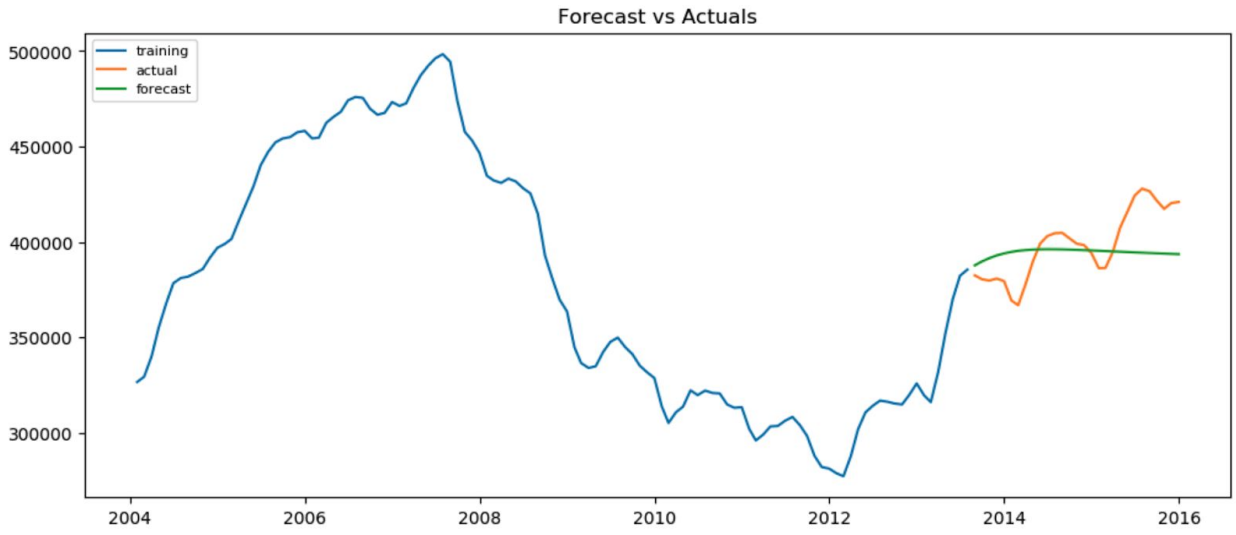
Forecast Visualization:



Best VAR

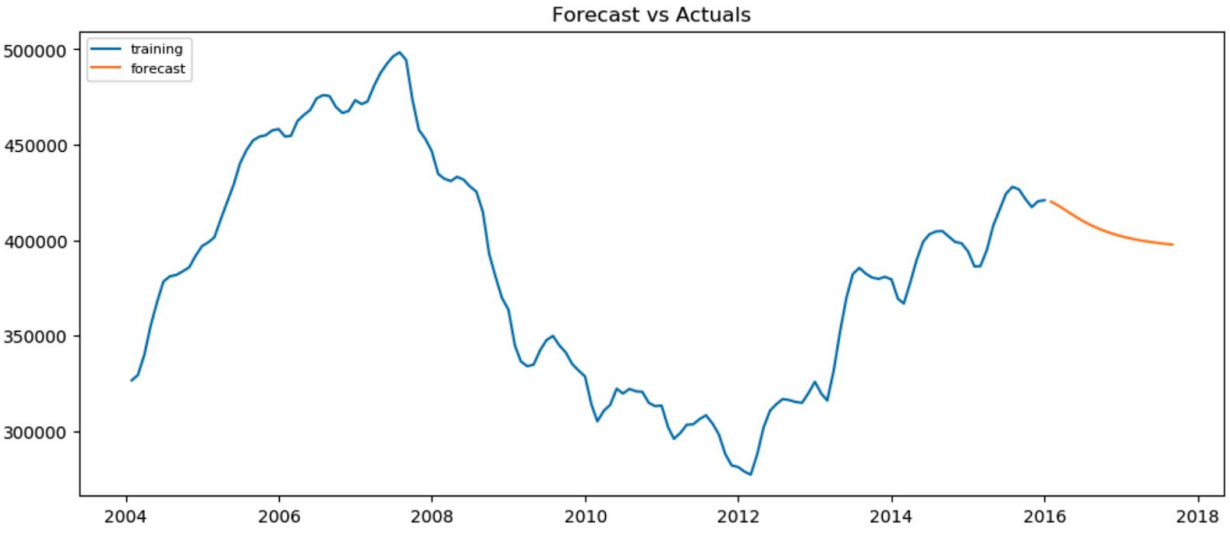
Initially, we tried using all variables that are available to predict the MedianSoldPrice.

Median Rental data has missing data, so to impute the missing data, we fit a VAR model using all the other variables except the target variable. Then using that data, we fit another full model to predict for our target using the additional imputed data. We found that using this imputed data, in fact, gave We found that the best model within VAR is using the parameters mortgage rate and the unemployment rate with p=2.



Validation RMSE: 17890.82

Forecast Visualization:



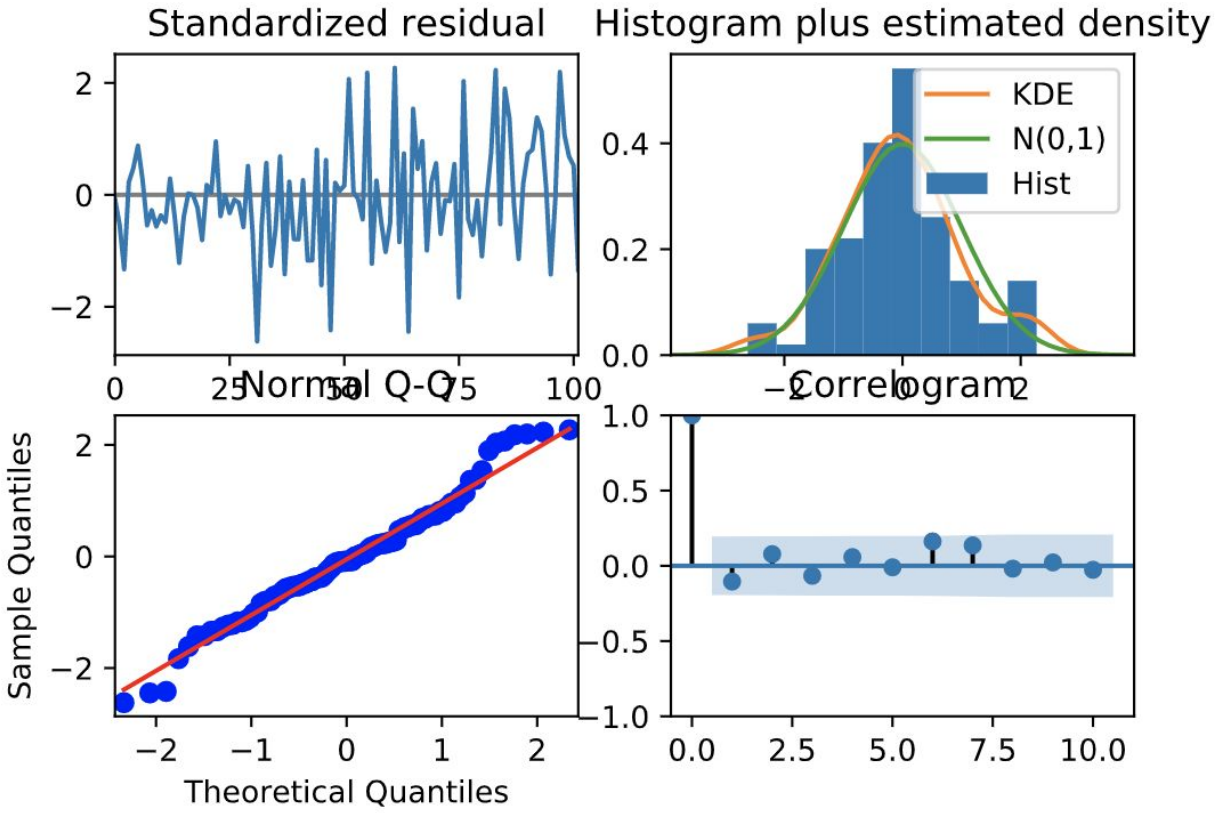
4. Final Model Choice and Explanation

Based on validation RMSE we choose the **SARIMA model (1, 1, 2) x (0, 1, 2, 12) fit on the full log-transformed target data** as our best model. Visually, the forecast of the validation set overlaid with true validation set looks good, and we have prior knowledge that median prices in California have been on the rise in recent years. More and more people have been moving into the state causing increased prices. The forecasted data predict an upward trend in January 2016 onwards which is what we know is true. This validates our prior assumptions.

Moreover, the model selects d = 1 and D = 1 as the time-differencing parameters, confirming what we learned in class that seasonal time-differencing should be of order 1. Trend time-differencing is 1 as well, which while not stationary according to ADF -- is approaching stationary. It selects AR order 1 and MA order 2 for the trend component, and no AR order but an MA order of 2 for the seasonal component.

5. Model Validation:

Final Model Diagnostics:



I. White Noise Behaviors

- The mean is consistently around 0
- Variance is smaller at the start, and larger but constant in the latter section
- Correlation between terms is 0 from Correlogram

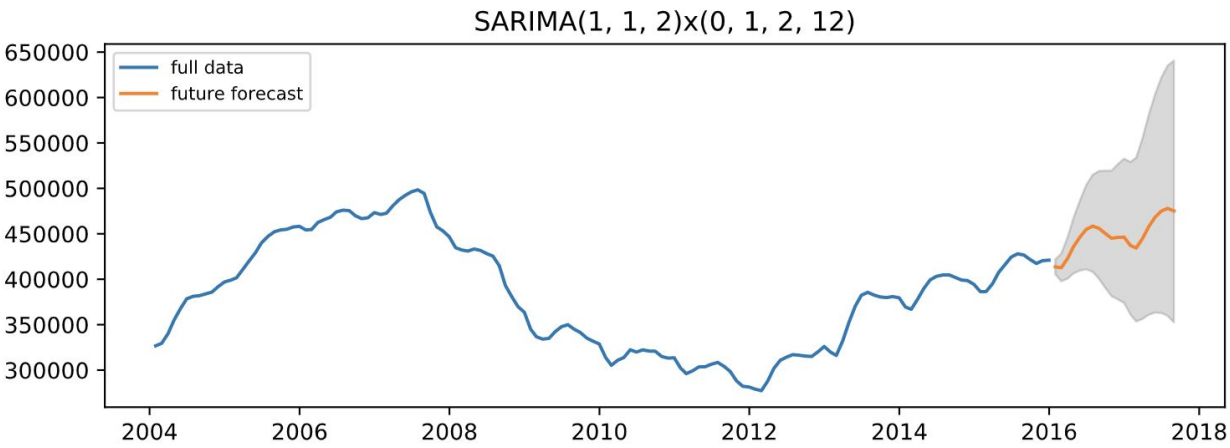
ii. Normality

- Histogram and Kernel Density Estimates show a normal distribution
- QQ-Plot shows close to the theoretical normal distribution

6. Forecast:

Using our chosen model, SARIMA model (1, 1, 2) x (0, 1, 2, 12) fit on log-transformed data, we forecast the missing sale price values:

2016-01-31	413651.8984
2016-02-29	412685.4501
2016-03-31	422938.9574
2016-04-30	436394.178
2016-05-31	446720.955
2016-06-30	454935.6407
2016-07-31	458494.6106
2016-08-31	455975.1019
2016-09-30	450488.9427
2016-10-31	445109.5476
2016-11-30	446055.2059
2016-12-31	446336.7706
2017-01-31	437112.2927
2017-02-28	434373.6639
2017-03-31	444843.1121
2017-04-30	457782.3259
2017-05-31	468224.4085
2017-06-30	474989.865
2017-07-31	477952.7139
2017-08-31	475179.0874


SARIMA(1, 1, 2)x(0, 1, 2, 12)
full data
future forecast
650000
600000
550000
500000
450000
400000
350000
300000
2004
2006
2008
2010
2012
2014
2016
2018

Team member contributions

Member	Proportion	Contributions
Andy Cheon	25%	Discussed problem statement, approach EDA on predictor relationships Ran SARIMA and SARIMAX on different orders; cross-verify with team members Integrated team members' notebooks into final deliverable Write report
Kevin Wong	25%	Discussed problem statement, approach Ran SARIMA and SARIMAX on different orders; cross-verify with team members. Focused on log-transformations of data Write report
Lin Meng	25%	Discussed problem statement, approach Ran TES, SARIMAX, VAR on different orders; cross-verify with team members Cleaned notebook Write report
Roja Immanni	25%	Discussed problem statement, approach EDA on predictor relationships Initial modeling to guide our group's modeling approach/direction Ran SARIMA and SARIMAX on different orders; cross-verify with team members Focus on VAR modeling Write report