Forecasting the Median Sale Price of Homes in Philadelphia, PA

Victoria Mullin

University of North Carolina Wilmington

I. Introduction

Zillow is an American online real estate database company that houses an abundance of useful data. Housing prices are notoriously difficult to forecast and typically involve trend and seasonality which need to be dealt with. For this reason, I thought using Zillow data would be a great learning experience for my final project to apply everything we have learned throughout this course. The dataset used in this analysis includes the raw median sale prices for all homes and is grouped by major cities in the U.S. Zillow defines "all homes" as single-family, condominium, and co-operative homes with a county record. The data is monthly and includes prices from February, 2008 to July, 2020. This project focuses on forecasting the median sale price for Philadelphia, PA since that is the city closest to where I grew up. Zillow provides smoothed and seasonally adjusted data, however, I thought using the raw values would allow me to apply decomposition myself. The goal of this project is to apply the models that I have learned throughout this course and determine which model produces the most accurate forecast.

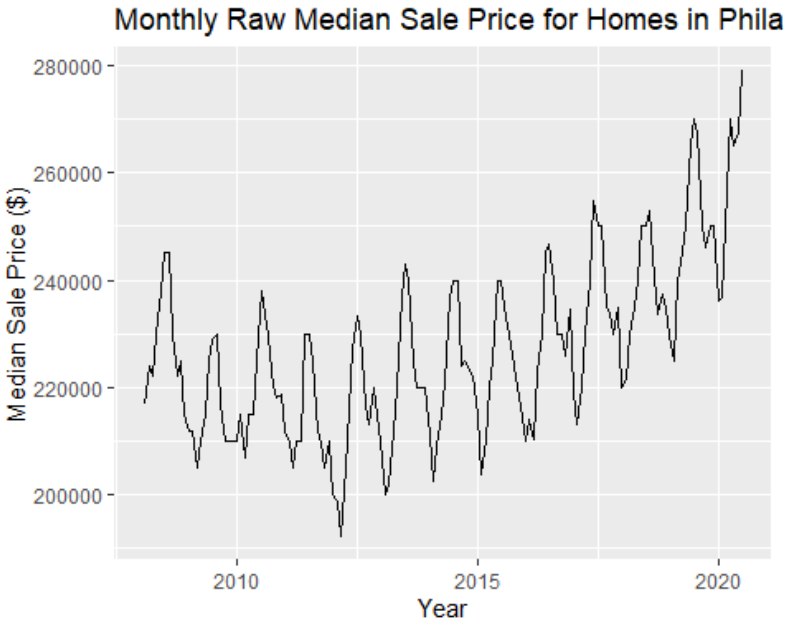
II. Methodology

Before we begin the analysis, it is necessary to create some exploratory plots in order to determine characteristics about the time series such as seasonality, trend, and cyclicity. These plots include a simple plot of the time series, a seasonal plot, a subseries plot, lag plot, and autocorrelation plot. In the first part of this project, the time series will be split into a train and test set so that the accuracy of the models can be evaluated. This will be done by assigning the first 80% of the time series to the training set and the remaining 20% to testing. From here, we will implement the four benchmark methods: naive, seasonal naive, drift, and average method. The models will be evaluated based on the RMSE and MAE values as well as with residual diagnostics. Next, we will perform STL decomposition to split the time series into seasonal, trend, and remainder components. Then we will explore several smoothing methods such as simple exponential smoothing (SES), Holt's trend methods, Holt-Winters' seasonal methods, as well as ETS models and determine which is best for this data. The final part of this analysis will be determining the best ARIMA models or any alternative models such as autoregressive only, moving average only, or ARMA models. These models will be evaluated by their AIC, RMSE, and MAE values as well as the behavior of the residuals. In the conclusion of this project, the various methods and models used throughout will be compared in order to determine which performed best.

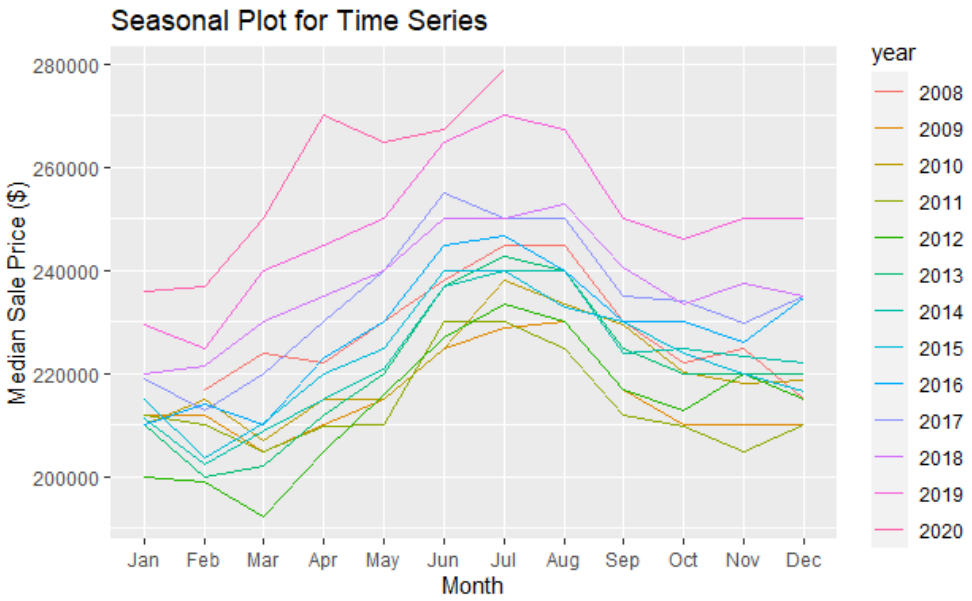
III. Analysis

a. Exploratory Plots

First, we will explore the time series by simply plotting it.

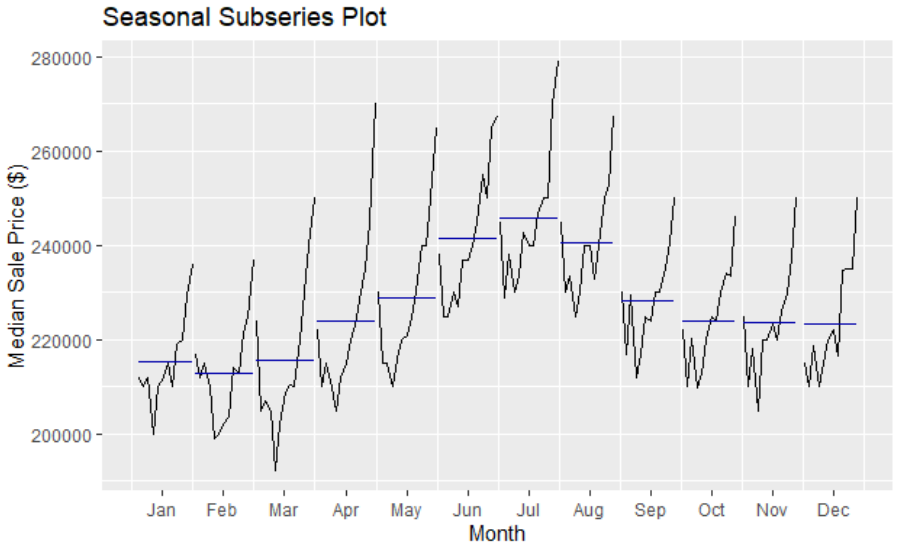


Just from looking at the plot, there appears to be yearly seasonality as well as a slight upward trend. It does not appear to have any cyclic behavior. To explore the seasonality more, we can look at the seasonal plot.

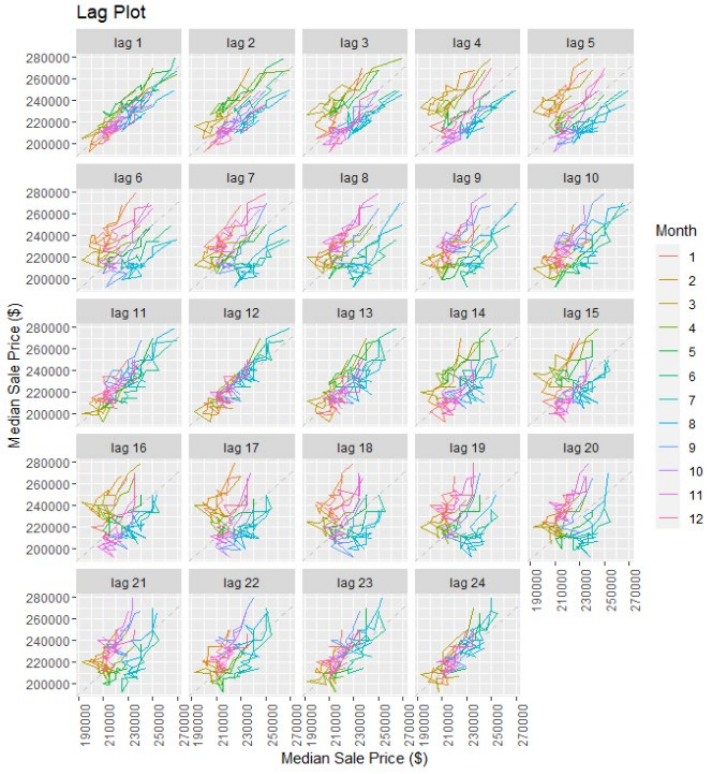


This confirms my suspicion since every year the median sale price increases up until around July and then decreases for the remainder of the year. The upward trend is also visible here since as the earlier years are at the bottom of the graph and as time progresses the more recent years are at the top of the graph.

In the subseries plot below, the blue lines represent the mean sale price for each month.

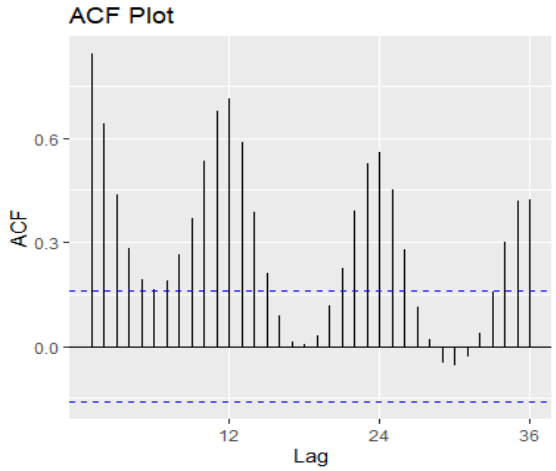


The mean value dips slightly in February, increases consistently, peaks in July, and then dips back down and levels off in the remaining months. Another useful plot to explore this time series is a lag plot in which the original time series is plotted against lagged versions of itself.



The relationship appears strongly positive at lags 12 and 24, indicating strong seasonality in the data.

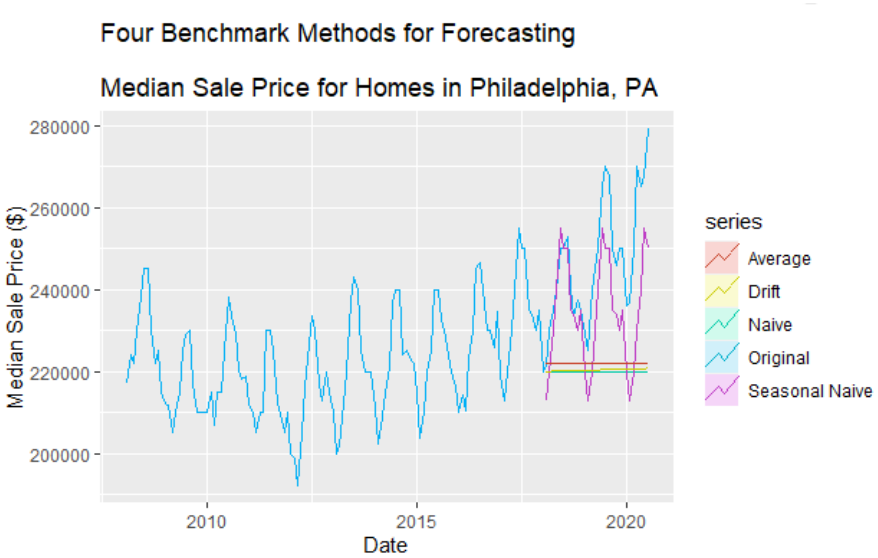
Autocorrelation measures the linear relationship between lagged values of a time series. This is another way of confirming seasonality in the data as well as trend.



From the ACF plot above we can see peaks at lag multiples of 12, indicative of the seasonal pattern in the data. The autocorrelations also have positive values that slowly decrease as the number of lags increase, indicative of a trended time series. The ACF plot can also tell us whether a series is white noise, however, we are more interested in whether the residuals of a model are white noise or not, which we will check later.

b. Four Benchmark Methods

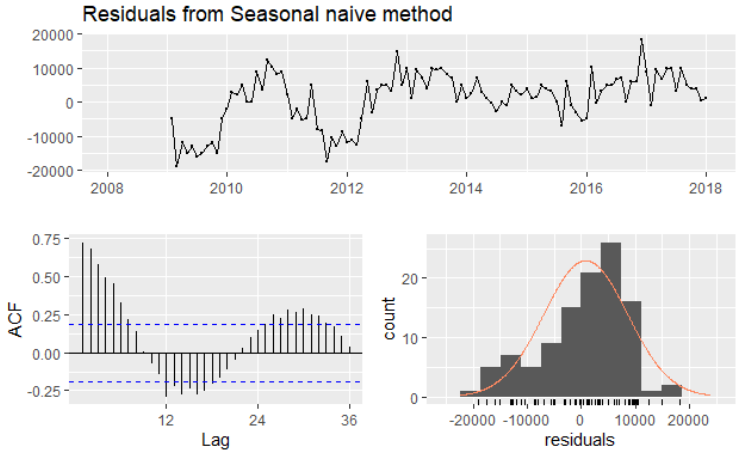
The length of the time series is 150, meaning the first 120 (80%) values will be assigned to the training set. The remainder will be the testing set which we can compare our forecast to. The first method we will use to forecast the median sale price of homes in Philadelphia is the naive method, which simply takes the last observed value and uses that for every future value. The next method, seasonal naive, takes the last observed value for each season and applies the respective values to every future season. The average method takes the average value of the time series and uses that for every future value. The final method, drift, basically draws a line between the first and last observations and extrapolates that into the future.



Looking at the plot of the four methods above, the average, naive, and drift methods all look somewhat similar and are all very poor forecasts. The seasonal naive method, on the other hand, looks somewhat realistic and is not extremely far off from the actual data.

	RMSE <dbl>	MAE <dbl>
Seasonal Naive	16370.28	13259.13
Average	29109.95	25266.68
Drift	30492.16	26881.71
Naive	30900.91	27272.47

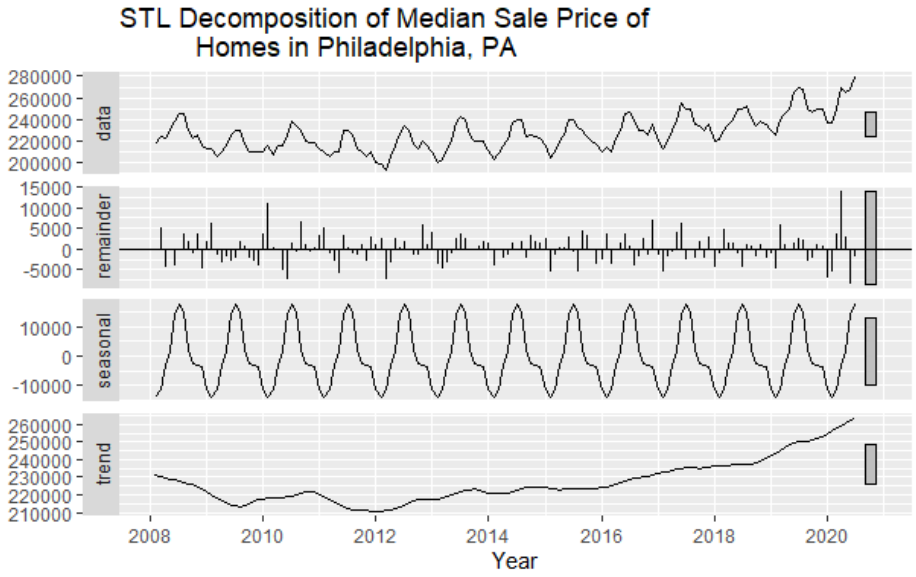
The RMSE and MAE for the seasonal naive is much lower than that of the other methods. For this reason, we will check the residuals for just the seasonal naive method since it is the best.



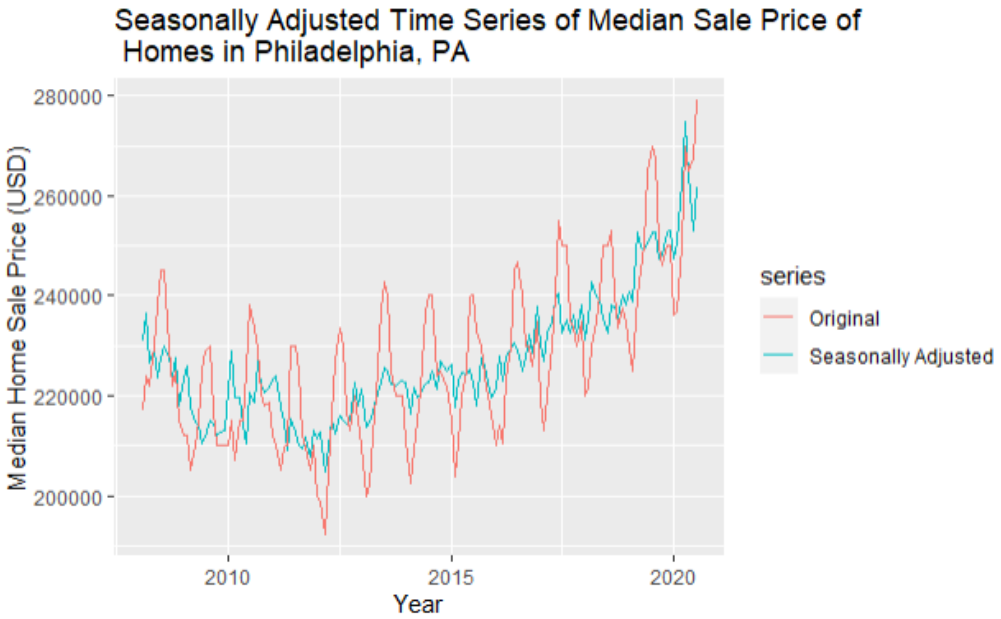
The ACF of the residuals of the seasonal naive method indicates that the residuals are *not* white noise since many of the autocorrelations go beyond the blue bands, indicating that they are significantly different from 0. This is confirmed with the results of the Ljung-Box test since the p-value is 2.2e-16. This means there is enough evidence to reject the null hypothesis that the model does not exhibit lack of fit. This is unfortunate because it means that there is still information in the residuals that is not being used in the model when it should be. Luckily there are plenty of other methods and models available which we will explore next.

c. Decomposition

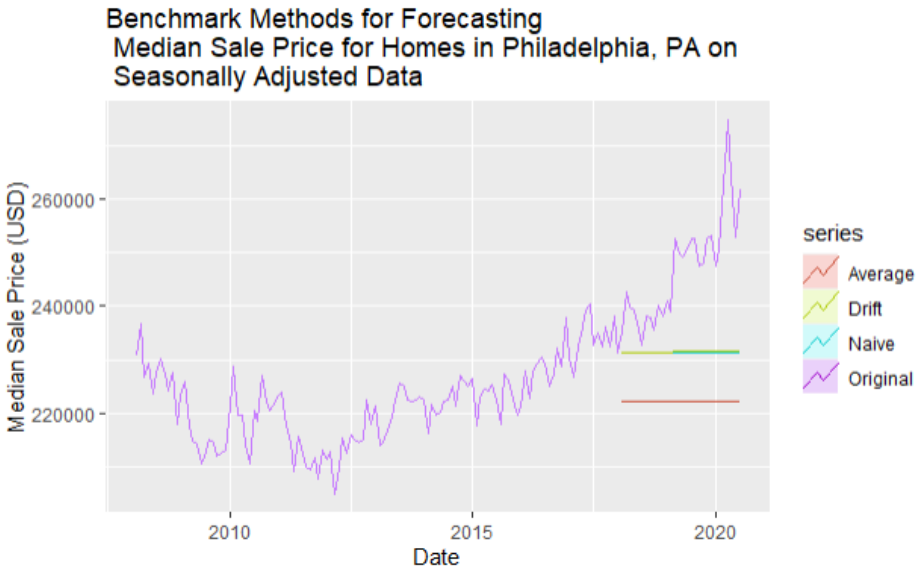
Since we detected seasonality and possible trend earlier, it may be useful to apply decomposition on this time series to further understand the data. There are several methods of decomposition and we will use the STL method which stands for "Seasonal and Trend decomposition using Loess," This method allows the seasonal component to change over time and can be defined, as well as allows the user to control the smoothness of the trend-cycle. I chose the seasonal window to be infinity, meaning the seasonal component is identical across years, which it appears to be in the season plot from earlier. The decomposition plots can be seen below.



We will also remove the seasonal component from the original data to obtain the seasonally adjusted time series, which may improve our models since the variation due to seasonality is not of primary interest. The plot of the seasonally adjusted and original data can be seen below.



Next, we can try using the benchmark methods on the seasonally adjusted data except for seasonal naive.

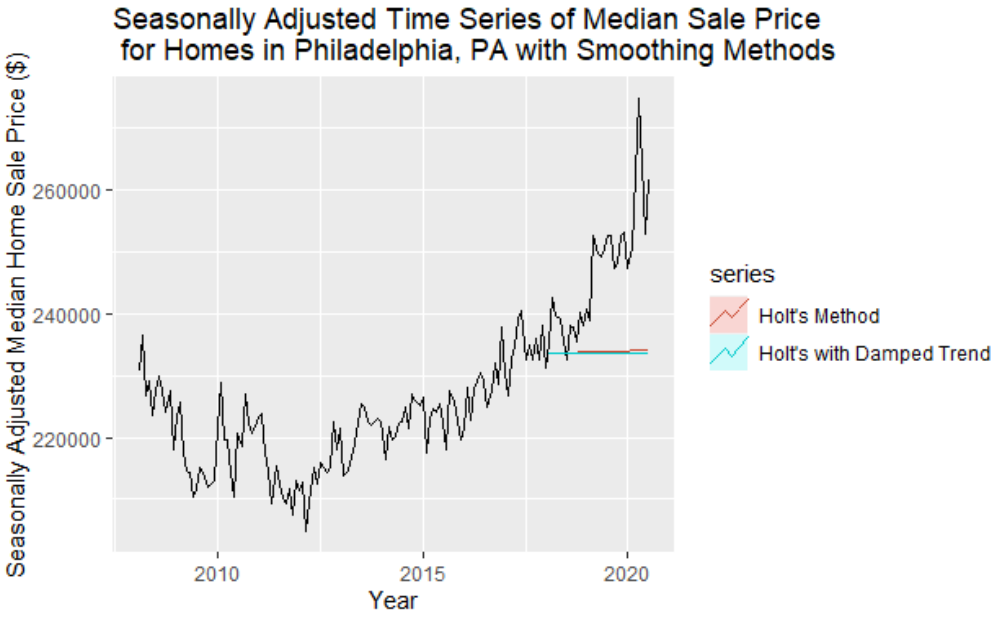


Looking at the plot above, the drift and naive methods on seasonally adjusted data appear almost the same since the forecasts overlap each other. Comparing all the RMSE and MAE scores below, taking out the seasonal component improved the benchmark models significantly, now with much lower RSME and MAE scores, however, the forecasts still are not ideal.

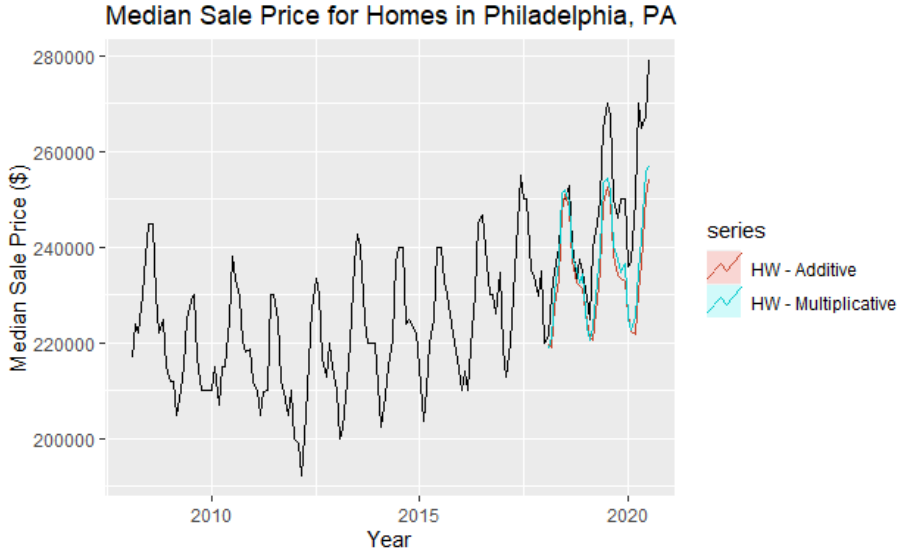
	RMSE <dbl>	MAE <dbl>
Seasonal Naive	16370.28	13259.13
Drift SA	18574.60	15788.54
Naive SA	18650.50	15857.63
Average SA	27020.60	25174.12
Average	29109.95	25266.68
Drift	30492.16	26881.71
Naive	30900.91	27272.47

d. Smoothing

Next, we will use exponential smoothing which gives the more recent observations a higher weight in the forecast model. Simple exponential smoothing is one method and is suited for time series that have no clear trend nor seasonal pattern. This is not the case for our time series and therefore SES is most likely not the best smoothing technique here. A second smoothing technique is Holt's linear trend method which allows trend but does not allow seasonality. This method involves a forecast equation and two smoothing equations: one for the level and one for the trend. Due to the no seasonality requirement, we cannot use this method on our original data, however, we can use it on the seasonally adjusted data. We can also try this method with damped trend which flattens the trend to a flat line sometime in the future. We can see in the plot below that these two methods are very similar in this case and appear to be very poor forecasting models for this data.

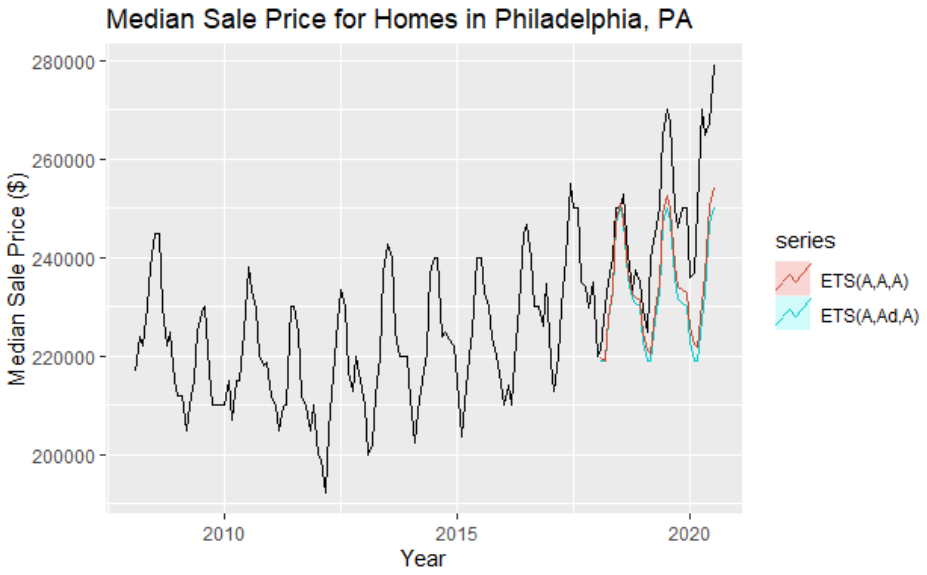


Holt-Winter's seasonal method is another form of smoothing that allows both trend and seasonality, meaning we can apply it to our original series. This method includes one forecast equation and three smoothing equations, one for level, one for trend, and one for the seasonal component, which correspond to smoothing parameters $\alpha, \beta,$ and γ. There are two types of this method in which the seasonal component is either additive or multiplicative. The additive method is recommended when the seasonal variations are somewhat constant throughout the series. We will try both methods and compare their results.



Looking at the plot of both Holt-Winters' methods above, they appear almost identical and may be the best forecast we have used thus far. Furthermore, the multiplicative method slightly outperformed the additive method according to the RMSE and MAE values.

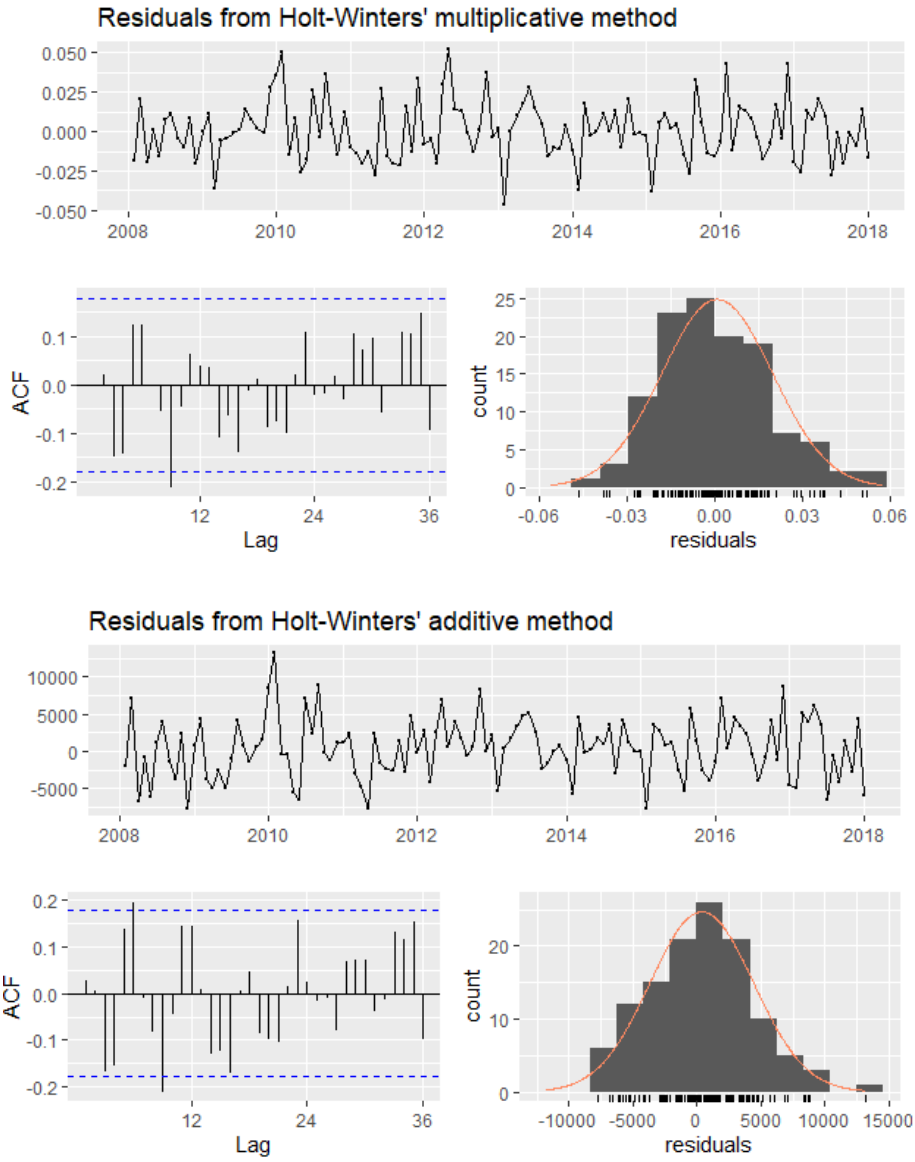
A final smoothing method that we can try is using ETS models, which stands for "Error, Trend, and Seasonality." The error can either be additive (A) or multiplicative (M), the trend can either be defined as none (N), linear (A), or damped (Ad), and the seasonality can either be none (N), additive (A), or multiplicative (M). We want to pick the best combination of these components such that AIC, AICc, and BIC are minimized. Using R's built-in function, it will choose the best ETS model for us, meaning the model with the lowest AICc by default. The model R chose uses additive error, damped trend, and additive seasonality. We can also try this model with a damped trend in order to see if that yields better results. We will use these models to forecast values and use its RMSE and MAE values to compare to our other models.



When plotting all four smoothing forecasts together, they all overlapped each other, making it difficult to distinguish between them. For this reason, I plotted the ETS models separately above, which we can see are also very close to each other.

	RMSE <dbl>	MAE <dbl>
HW-Multiplicative	13226.92	10515.77
HW - Additive	15578.79	12588.05
ETS(A,A,A)	15582.88	12591.55
Seasonal Naive	16370.28	13259.13
ETS(A,Ad,A)	17638.17	14488.32
Drift SA	18574.60	15788.54
Naive SA	18650.50	15857.63
Average SA	27020.60	25174.12
Average	29109.95	25266.68
Drift	30492.16	26881.71

Based on the scores above, the best model thus far is Holt-Winters' multiplicative method closely followed by the additive version. We can check the residuals of these models to see if they behave like white noise or not, giving us more insight into the quality of fit.



Looking at the ACF plot of the residuals for both models, it seems like the residuals are white noise since all but one or two of the autocorrelations are inside of the blue bands, indicating that they are not significantly different from zero. However, looking at the results of the Ljung-Box test for both models, the p-values are both small (0.0006477 for the multiplicative method and 6.227e-07 for the additive method) indicating the contrary - that the residuals are *not* white noise. This is not good because it means that there is still useful information, not just noise, leftover in the residuals which are not being used in the model. We will try more complex models such as ARIMA in the next section to find a better model.

e. ARIMA

Before we begin implementing ARIMA models, we need to determine whether this series is stationary and, if not, then what order differencing we will need to take. A series is stationary if its properties do not depend on the time at which the series is observed. This means that

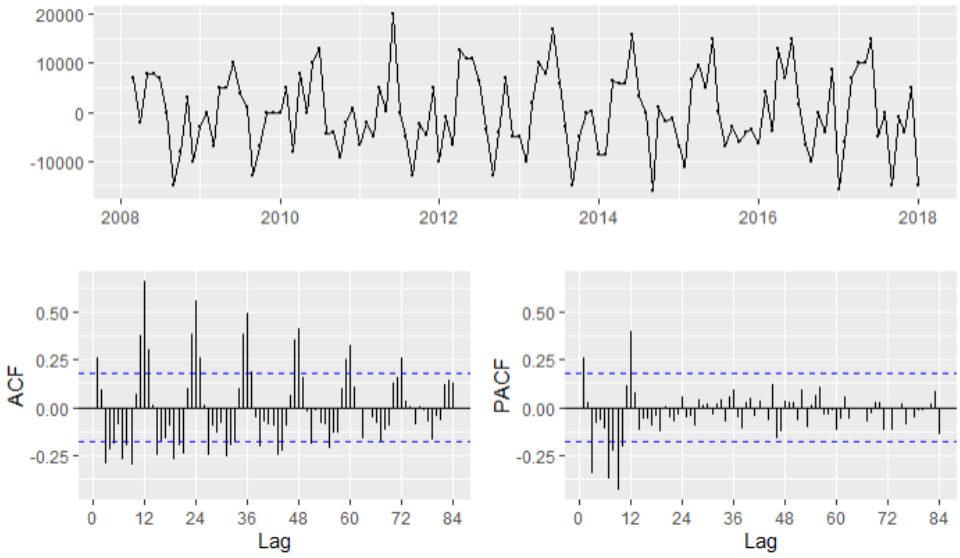
seasonality and trend make a series nonstationary. Although I am almost certain that this series is not stationary, we can use two unit root tests to confirm that assumption. The output from these tests can be seen below.

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 1.6429

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739

[1] 1
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  ts
Dickey-Fuller = -4.8357, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

The results of the tests are contradicting each other since the conclusion from KPSS test is that the series is not stationary and we need to take a difference since the test-statistic is large. Alternatively, the ADF test p-value is small (<0.05), meaning there is enough evidence to reject the null hypothesis that the series is not stationary. This means that it *is* stationary. We will go with the KPSS results and take a first difference. The "ndiffs" function returned a value of 1, meaning that we should take the first difference to achieve stationarity. We can re-run these tests again on the differenced data to determine if the differenced series is stationary.

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 0.0583

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739

p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  diff(ts)
Dickey-Fuller = -7.5802, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

After re-running both tests on the differenced series, the results both agree that the new series, after taking the difference, is stationary. To determine what models to try first, we can look at the ACF and PACF plots.



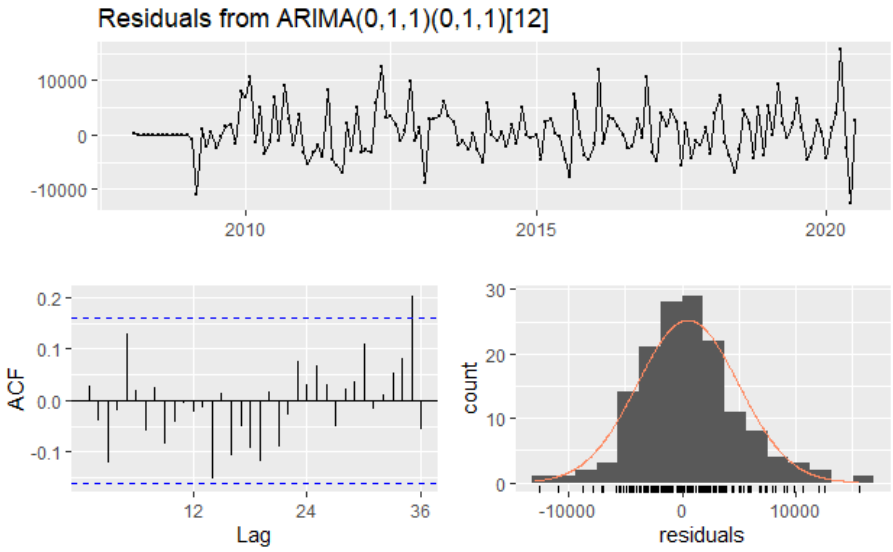
Looking at the PACF plot, significant spikes stop after lag 12. For this reason, we can try the AR(12) model first. This will be an ARIMA(12,1,0) model since we can specify the first difference as the value of d. Looking at the ACF plot, there are significant spikes at lags of multiples of 12 up until lag 84, indicating that a MA(6) model on the seasonal part may be worth trying. This will be an ARIMA(0,1,0)(0,0,6)[12] model indicating the first difference is taken and Q=6 for the moving average of the seasonality part.

R also has a built-in function called “auto.arima” that we can use which will select the best Arima model based on the lowest AIC value. The model R chose is ARIMA (0,1,1)(0,1,1)[12]. Based on this, we can also try an ARIMA(12,1,0)(0,1,1)[12] to see if incorporating the autoregressive model is beneficial.

The RMSE, MAE, and AIC values for each model attempted can be seen below. According to this, the ARIMA(0,1,1)(0,1,1)[12] model performed best, closely followed by the ARIMA(12,1,0)(0,1,1)[12] model.

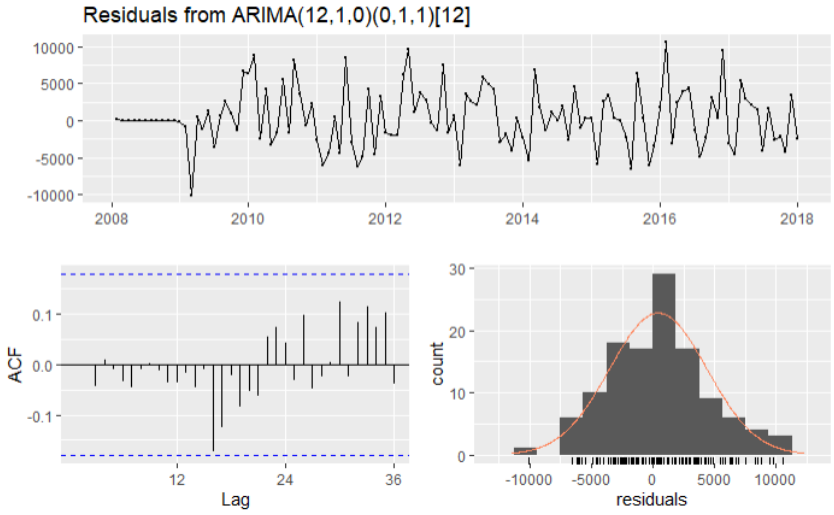
Model <chr>	AIC <dbl>
ARIMA(0,1,1)(0,1,1)[12]	2111.246
ARIMA(12,1,0)(0,1,1)[12]	2123.625
ARIMA(12,1,0)	2372.235
ARIMA(0,1,0)(0,0,6)[12]	2390.530

Since these two models are best based on the AIC, we can look at the residuals to evaluate them further. First, we will check the residuals of the model selected by auto-arima, ARIMA(0,1,1)(0,1,1)[12].



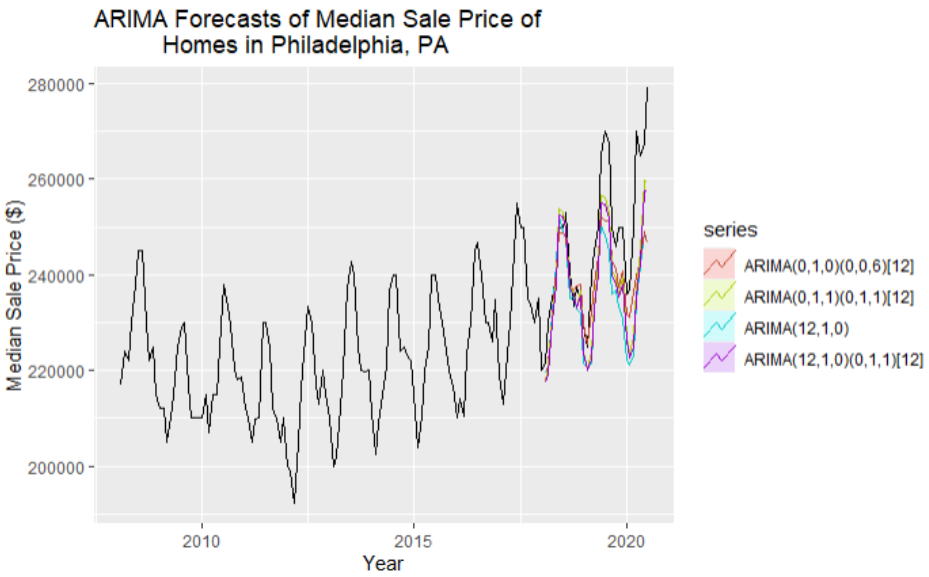
Both the ACF of the residuals of this model and the results of the Ljung-Box test indicate that the residuals of this model are white noise since the p-value is "big", specifically 0.5289. This is good because it means that there is no leftover information in the residuals and they are just noise, meaning it is OK to leave them out. Also, the residuals appear somewhat to be normal, which is also good.

Checking the residuals of the ARIMA(12,1,0)(0,1,1)[12] model, the ACF of the residuals can be seen below.



The residuals appear to be white noise based on the ACF plot since none of the lags go beyond the blue bands, indicating that they are not significantly different from 0. The Ljung-Box test also conclude that the residuals are white noise because the p-value is "big," specifically 0.4031. The distribution of the residuals appears somewhat normal as well.

The plot of the ARIMA forecasts can also be seen below.



Just from looking at the plot, the four ARIMA models tested appear very close together. According to the AIC values, the yellow line which represents the ARIMA(0,1,1)(0,1,1)[12] model is best. From the plot, these forecasts appear somewhat realistic, however, they do not predict the upward trend that actually occurs from the original series.

IV. Conclusion

After using various methods and models to forecast the median sale price of homes in Philadelphia, PA, we can compare all the RMSE and MAE values.

	RMSE <dbl>	MAE <dbl>	AIC <dbl>
ARIMA(0,1,1)(0,1,1)[12]	13048.99	10454.260	2111.246
ARIMA(12,1,0)(0,1,1)[12]	11996.95	9571.023	2123.625
ARIMA(12,1,0)	15537.38	12425.525	2372.235
ARIMA(0,1,0)(0,0,6)[12]	12140.74	8783.421	2390.530
ETS(A,Ad,A)	17638.17	14488.316	2599.797
ETS(A,A,A)	15582.88	12591.546	2601.673
Naive	30900.91	27272.467	*NA*
Seasonal Naive	16370.28	13259.133	*NA*
Average	29109.95	25266.677	*NA*
Drift	30492.16	26881.710	*NA*
Naive SA	18650.50	15857.634	*NA*
Average SA	27020.60	25174.118	*NA*
Drift SA	18574.60	15788.536	*NA*
HW - Additive	15578.79	12588.051	*NA*
HW-Multiplicative	13226.92	10515.768	*NA*

The best model according to AIC is the ARIMA(0,1,1)(0,1,1)[12] model. However, according to RMSE values, the ARIMA(12,1,0)(0,1,1)[12] model is best. As we discovered in the last section, the residuals of both models appear to be white noise. To determine which of these models is best, we can use cross validation as well as compare the size of the prediction

intervals. The results from cross validation can be seen below, comparing all the Arima and ETS models we used.

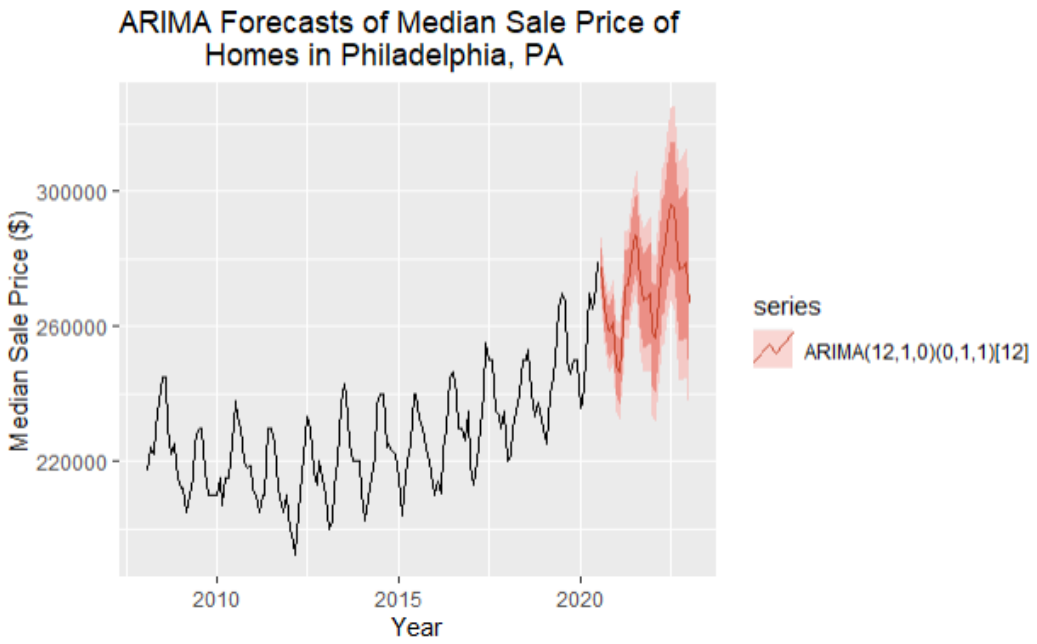
	Model <chr>	MSE <dbl>
2	ARIMA(12,1,0)(0,1,1)[12]	120029519
4	ARIMA(0,1,0)(0,0,6)[12]	178751444
3	ARIMA(12,1,0)	186123337
1	ARIMA(0,1,1)(0,1,1)[12]	337034113
6	ETS(A,A,A)	1416378871
5	ETS(A,Ad,A)	1447129942

Based on cross validation, the ARIMA(12,1,0)(0,1,1)[12] is best. To further confirm this, we can look at the range of the 80% prediction intervals of this model, the ARIMA(0,1,1)(0,1,1)[12] model, and the ARIMA(0,1,0)(0,0,6)[12] model since it is ranked the second best model by cross validation. The width the of 80% prediction intervals can be seen below followed by the column sums underneath. Since it is difficult to incorporate special characters in columns names in R, I did not include the specific ARIMA model names in the column names.

PI_Width_Model_1 <dbl>	PI_Width_Model_2 <dbl>	PI_Width_Model_3 <dbl>
11448.77	11450.71	12405.73
12797.89	12779.96	17510.55
14017.76	14496.55	21432.14
15139.65	15466.18	24739.72
16183.96	16196.32	27654.49
17164.86	17539.17	30290.07
18092.65	18985.83	32713.99
18975.13	20141.42	34970.31
19818.35	21251.52	37089.62
20627.13	21723.25	39094.21

```
PI_Width_Model_1 PI_Width_Model_2 PI_Width_Model_3
        831518.9         844123.2        1594528.7
```

Based on prediction intervals, the model with the smallest intervals overall is Model 1, which is the ARIMA(0,1,1)(0,1,1)[12] model, closely followed by Model 2, which is the ARIMA (12,1,0)(0,1,1)[12] model. Based on having the lowest RMSE value, the second to lowest AIC value, the lowest MSE value from cross validation, and the second smallest prediction intervals, we can conclude that the ARIMA(12,1,0)(0,1,1)[12] model is the best that we have found to predict the median sale price of homes in Philadelphia, PA. The plot of the forecast of this model with the prediction interval can be seen below.



This further demonstrates the that size of the prediction interval does not appear overly large and this model continues the upward trend at the end of the series. I will be interesting to compare this model to the actual values in a month. This project has given me insight into potentially what to expect if I end up moving back to Pennsylvania and start searching for a home.