Introduction

Today I wanted to use the time series commands and complete a forecast in R. I wanted to use these packages and commands to explore the data and then create time series plots and forecasts to make predictions about various points in the data. I wanted to be more focused and use the Qtr data on page 2 of the downloaded dataset. I wanted to first get to know our data and then use the formatting tools to read it in as time series data. We then transform the dataset into something that we can use and make sure that we can use the times series commands on.

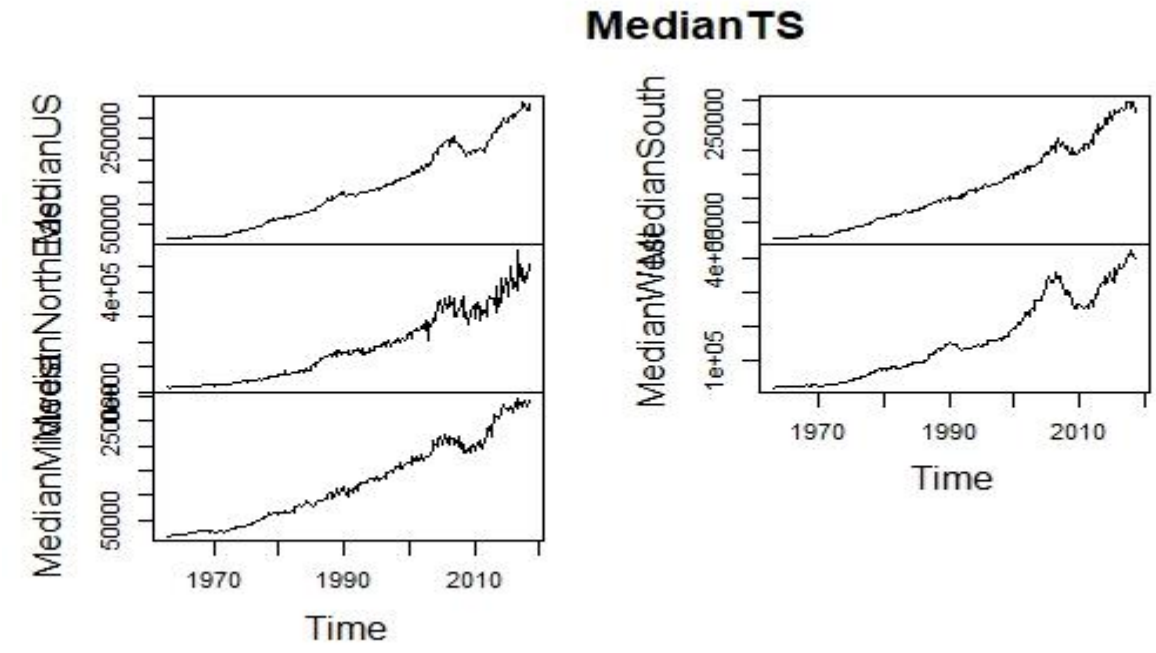
Methods and Results

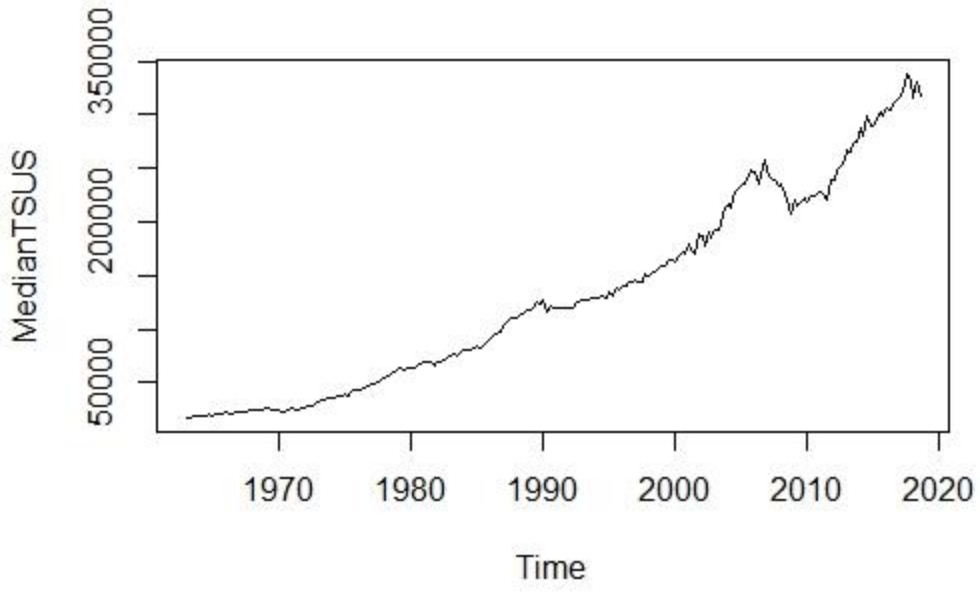
First, we were trying to see the structure of the data and describe what we saw along with some initial inferences. We read in the data first and then must make sure it is in data frame format. Once it is in this format we can complete transformations to ensure that the data is usable and in a time series data format for the assignment.

```
'data.frame':   223 obs. of  11 variables:
 $ YearQtr           : Factor w/ 223 levels "1963Q2","1963Q3",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ MedianUS          : num  18000 17900 18500 18500 18900 18900 19400 20200 19800 20200 ...
 $ MedianNorthEast   : num  20600 19600 20600 20300 19800 20200 21400 21000 21900 21200 ...
 $ MedianMidwest     : num  17700 17800 19100 18700 19800 18900 20800 21900 20800 22100 ...
 $ MedianSouth       : num  15800 15900 15800 16500 16800 16800 16700 17400 16400 17700 ...
 $ MedianWest        : num  18900 19000 19500 19600 20100 20600 21500 21600 22100 21500 ...
 $ AverageUS         : num  19400 19200 19600 19600 20200 20500 20900 21500 21000 21600 ...
 $ AvgerageNorthEast: Factor w/ 172 levels "(NA)","102000",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ AverageMidwest    : Factor w/ 173 levels "(NA)","102500",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ AverageSouth      : Factor w/ 170 levels "(NA)","102300",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ AverageWest       : Factor w/ 169 levels "(NA)","103400",..: 1 1 1 1 1 1 1 1 1 1 ...
```

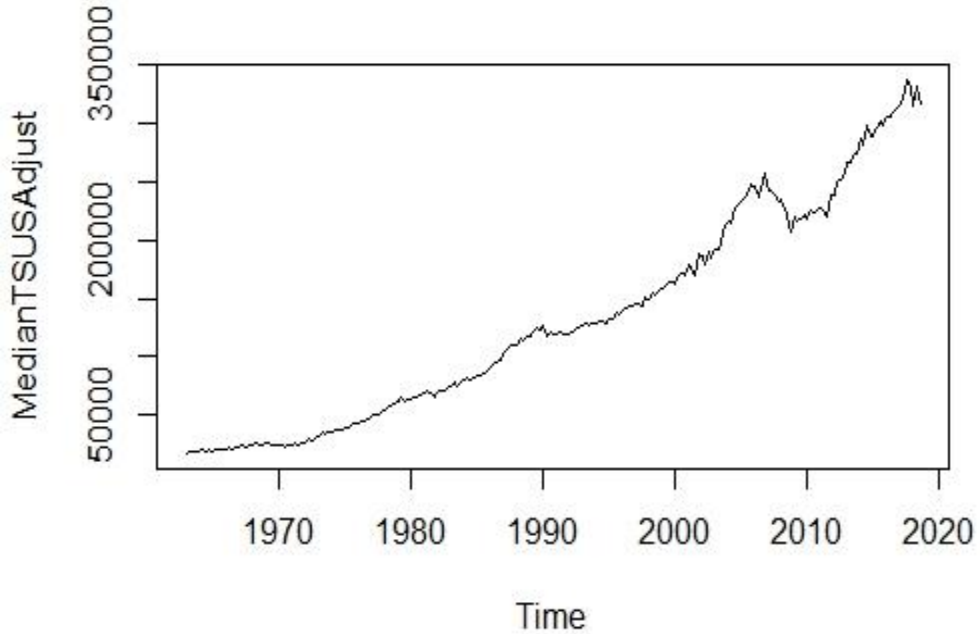
	YearQtr	MedianUS	MedianNorthEast	MedianMidwest	MedianSouth	MedianWest	AverageUS	AvgerageNorthEast
1	1963Q2	18000	20600	17700	15800	18900	19400	(NA)
2	1963Q3	17900	19600	17800	15900	19000	19200	(NA)
3	1963Q4	18500	20600	19100	15800	19500	19600	(NA)
4	1964Q1	18500	20300	18700	16500	19600	19600	(NA)
5	1964Q2	18900	19800	19800	16800	20100	20200	(NA)
6	1964Q3	18900	20200	18900	16800	20600	20500	(NA)

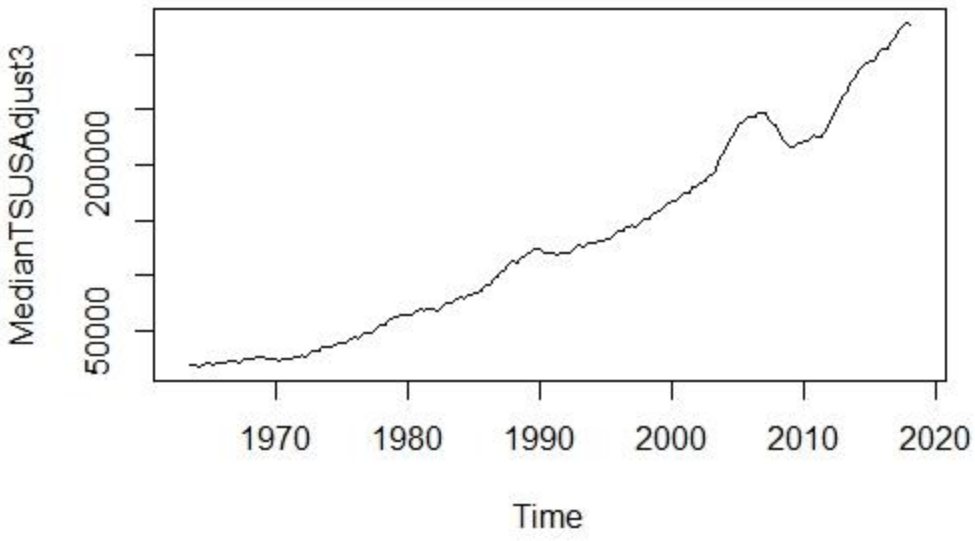
Once the data was converted into times series format I wanted to look at just the median as we were told for the Qtrs so that's what I pulled out and I was thinking about maybe looking at other variables later. For now, I wanted to start to look at the suggested commands with the median in order to start informing my forecast decision for the assignment. I also wanted to look at all the regions at the same time in order to be able to ensure that the data and models were reliable for the entire US before splitting out one of the regions for the assignment.

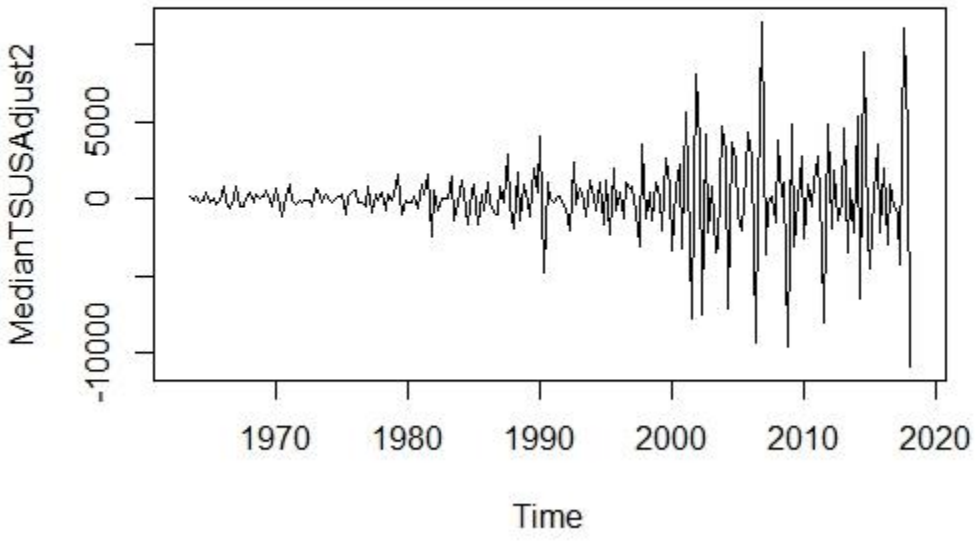


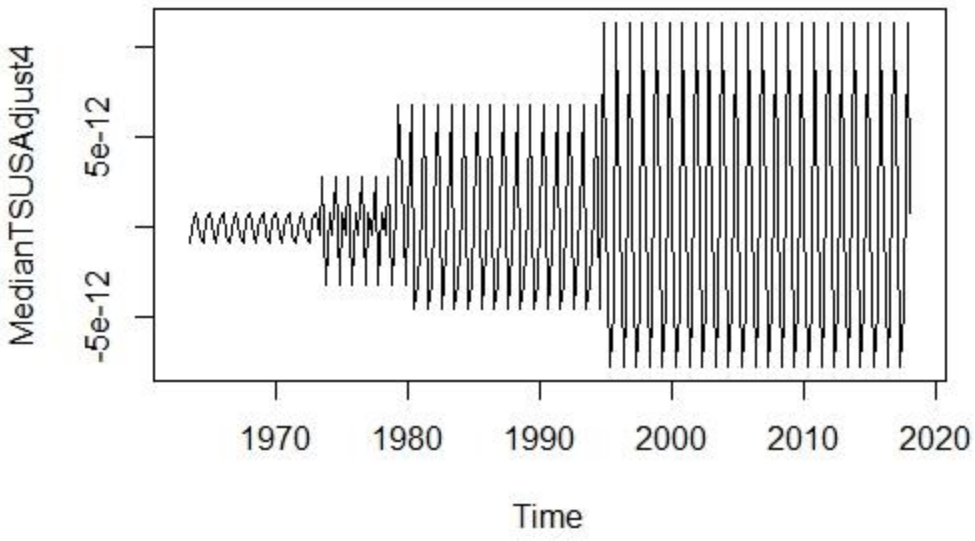


After the adjustments were made placing the seasonal, trends, random, and then all 4 into one I ensured that all the different.




MedianTSUSAdjust3
50000
200000
1970
1980
1990
2000
2010
2020
Time


MedianTSUSAdjust2
-10000
0
5000
1970
1980
1990
2000
2010
2020
Time

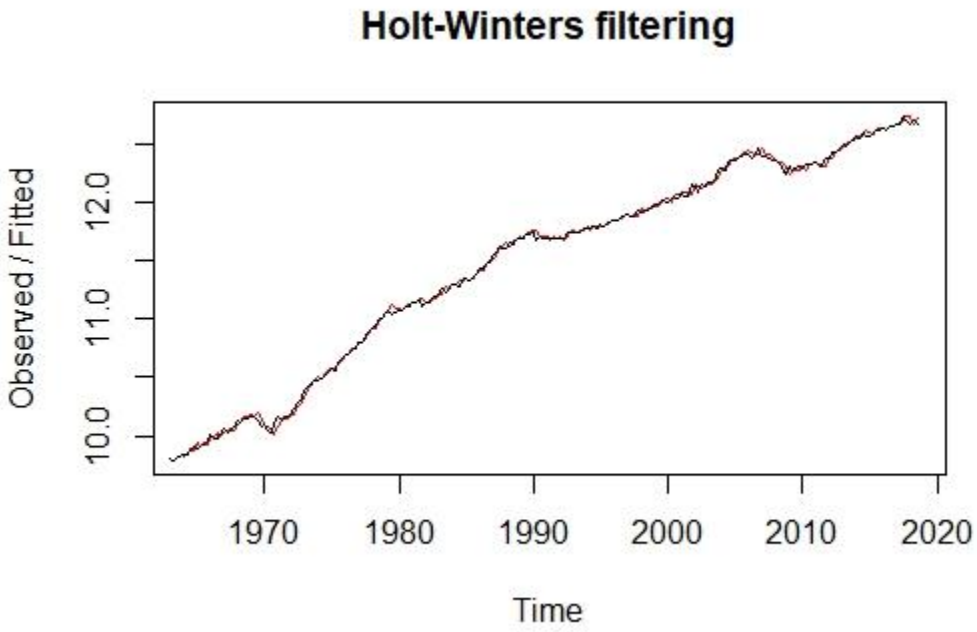


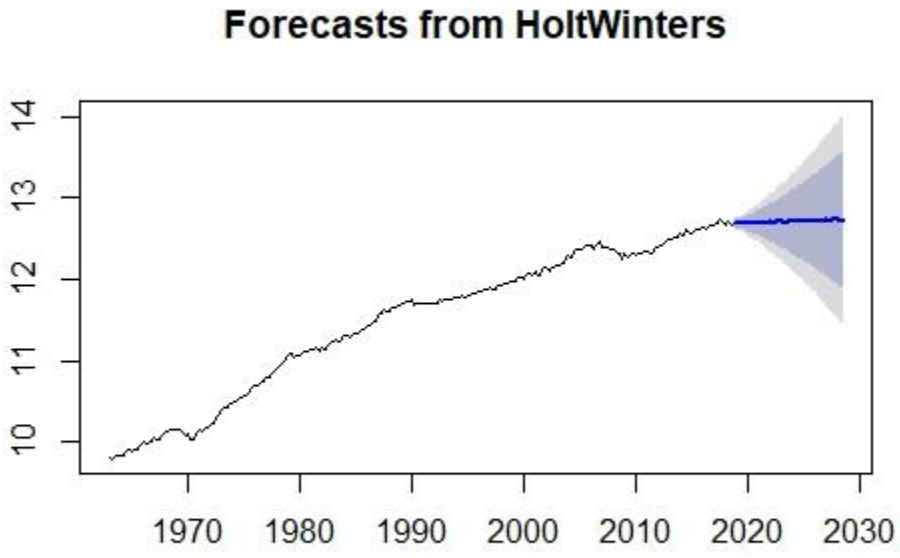
After running some of the commands to visualize the plots of the adjustments for seasonality and randomization, we can see that over time the variance is rising. This informs us that an additive model will not work because our time series data that we transformed is not being represented very well due to the multiplicative form it is currently in. So, for us to be able to use the additive models we will have to transform the data again so that we can use our additive models for the forecasting. Therefore, I will log transform the data and then use the Holt Winters model commands to evaluate the data and then forecast from. Basically, we are using smoothing methods that we learned about to use the additive models and account for the randomization and seasonality.

```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = MedianTSUSL)

Smoothing parameters:
 alpha: 0.6975487
 beta : 0.1925415
 gamma: 0.07524737

Coefficients:
            [,1]
a  12.6804267713
b   0.0011517708
s1  0.0005034694
s2  0.0040914964
s3 -0.0083802233
s4  0.0016014193
```

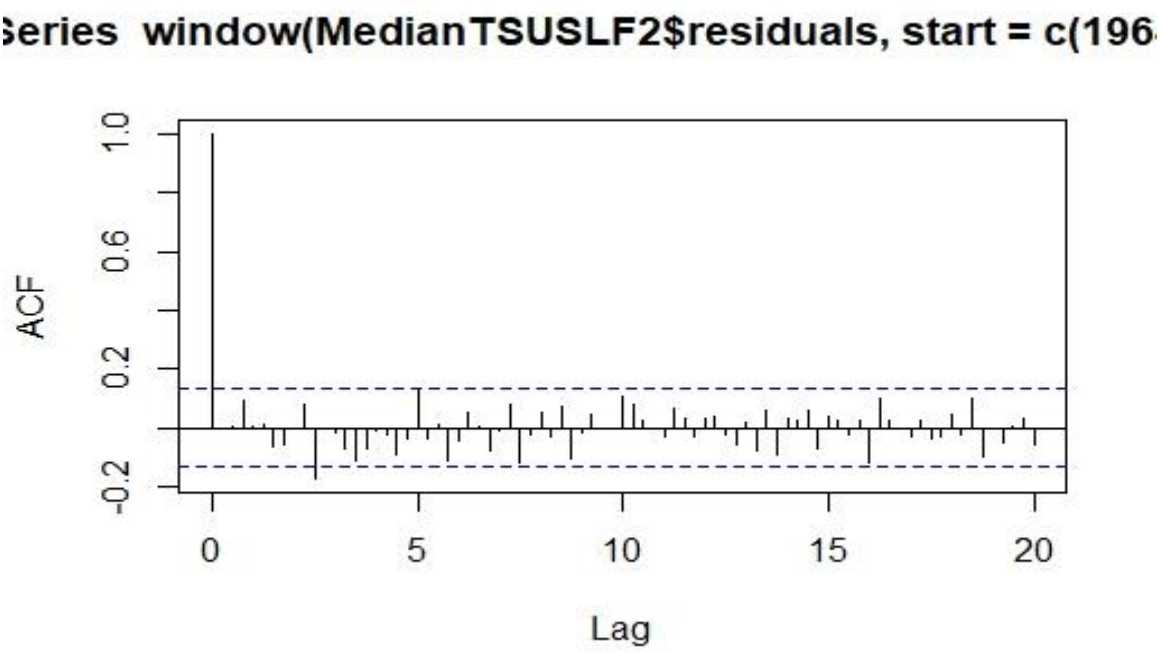
The smoothing metrics above show us how our data is being weighed. It also helps us to create our model to work from going forward. The alpha parameter tells us how much weight is put on our recent observations, the beta parameter tells us the slope, and the gamma parameter tells us how much weight and impact the seasonal components impact the model over time. Now that we can see how the US is split out, we can see how the model we created will be against actual numbers and

make the forecasts/predictions from it.





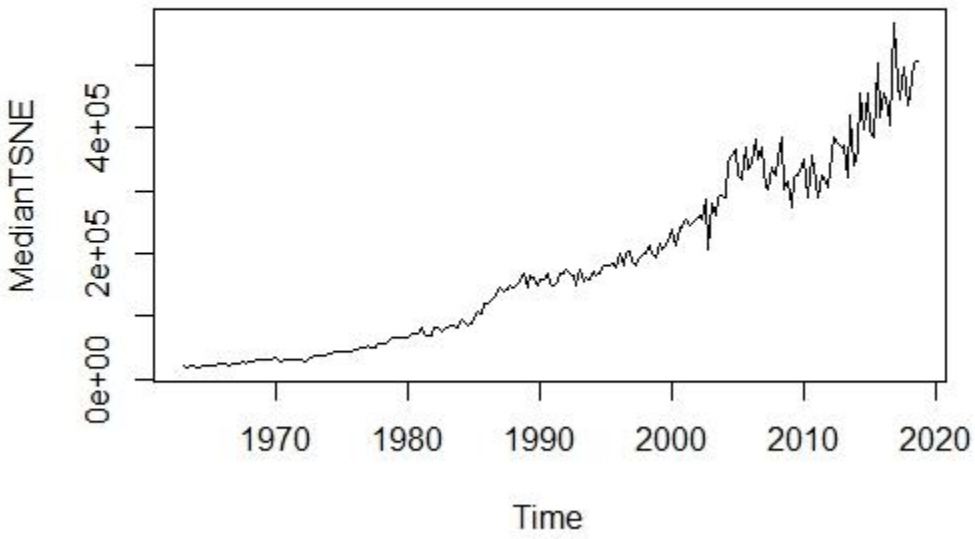
Now that we have a forecast we need to consider autocorrelations. The forecast can probably be improved if we account for the auto correlations. That will also give us a chance to cover the rest of the suggested commands. So, we will look at some commands to ensure that the autocorrelation is low because we have lagged values.

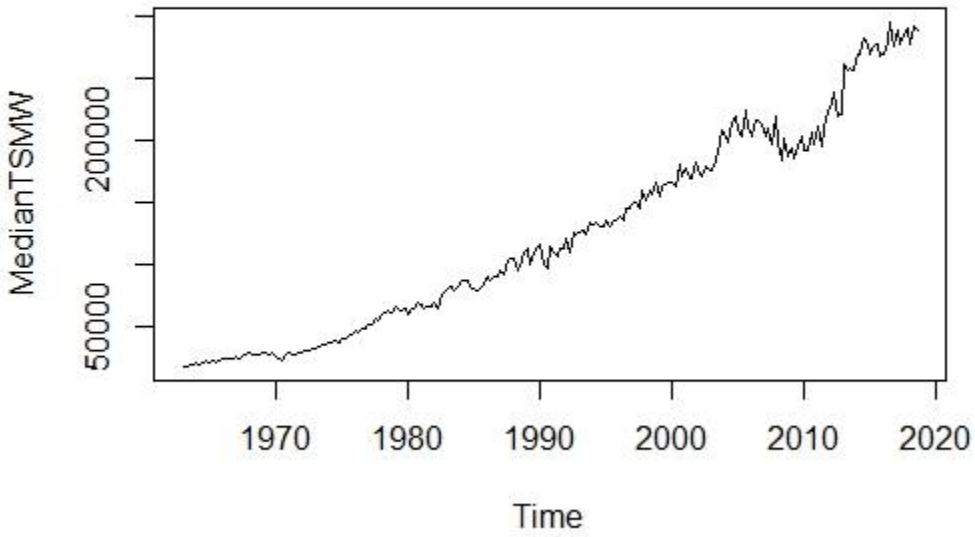


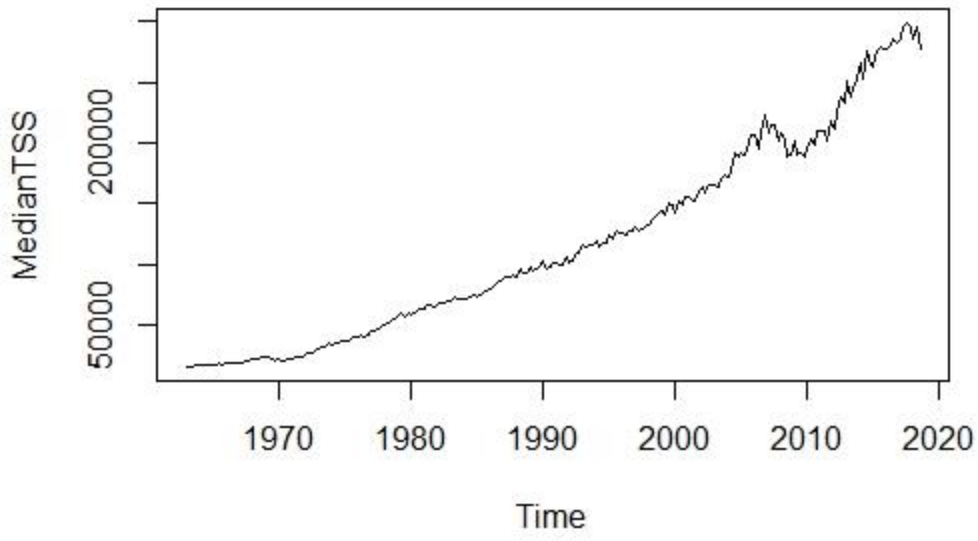

```
        Box-Pierce test

data:  window(MedianTSUSLF2$residuals, start = c(1964, 1))
X-squared = 63.919, df = 80, p-value = 0.9057
```

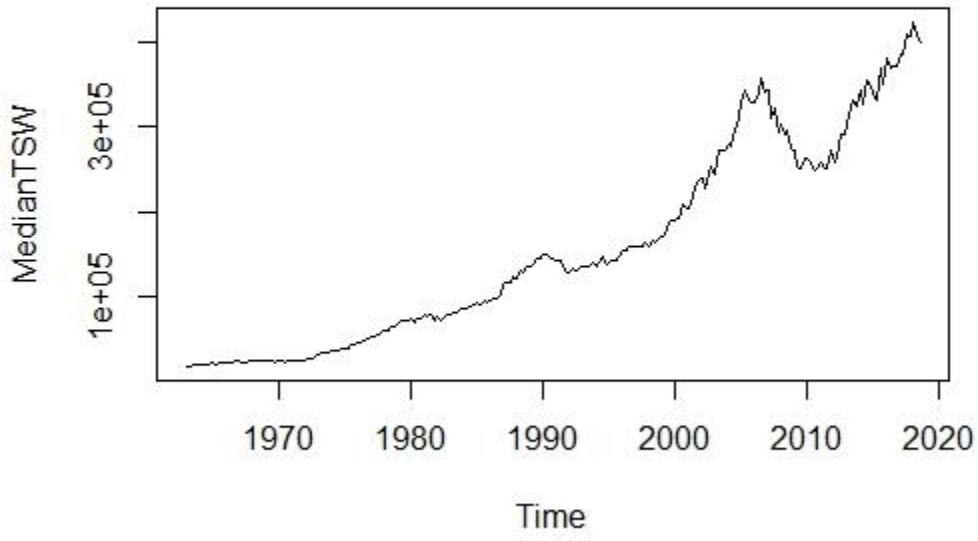
We then run the box-pierce test in order to see if the autocorrelation would be a factor in misrepresenting our data. However, we see that the p value is not significant and over .5 and closer to .9, so, it tells us we do not need to go one further evaluating the data for other misrepresentation factors or more. What this means overall is that the median is now covered, and we can see that the forecast for the next 10 years that we will use is reliable. So, we can now compare the regions and see how they also compare to this model. I understand we could choose one region for the

assignment, however, my curiosity hit me after cleaning the data and I wanted to see at least most of them.





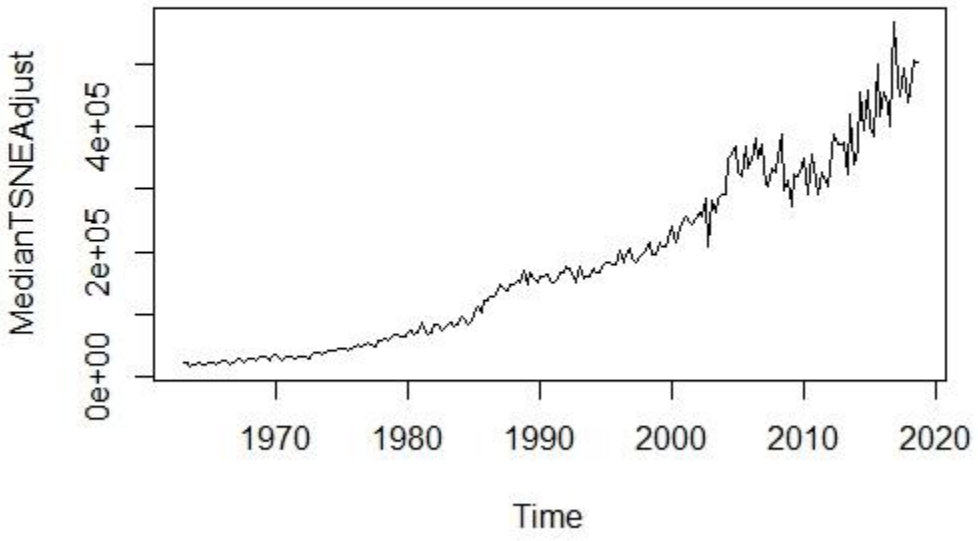




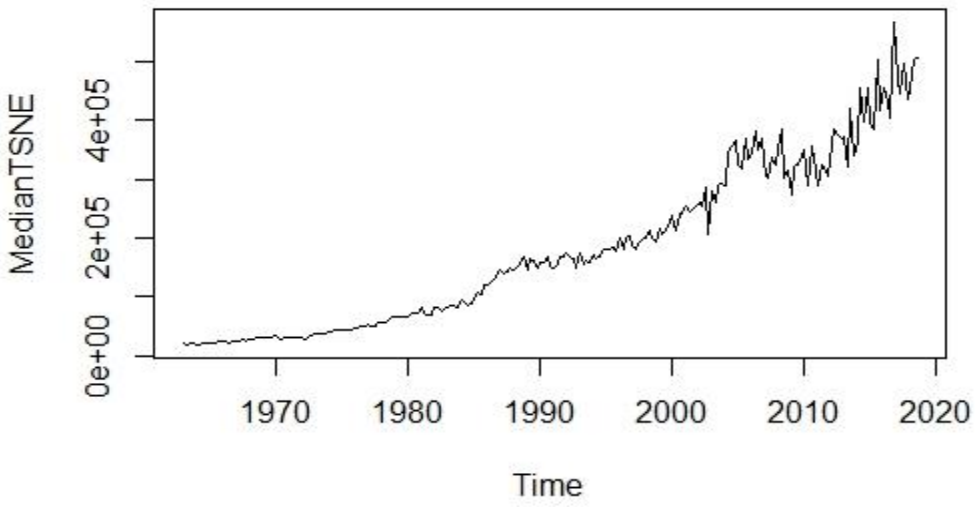
I can now see that each region above is different, however, I wanted to focus on the southwest because it had a very interesting augmentation to its trend around 2010. However, I thought it might be a good idea to compare all 4 of the major regions to the US to make a forecast. So, I did all 4

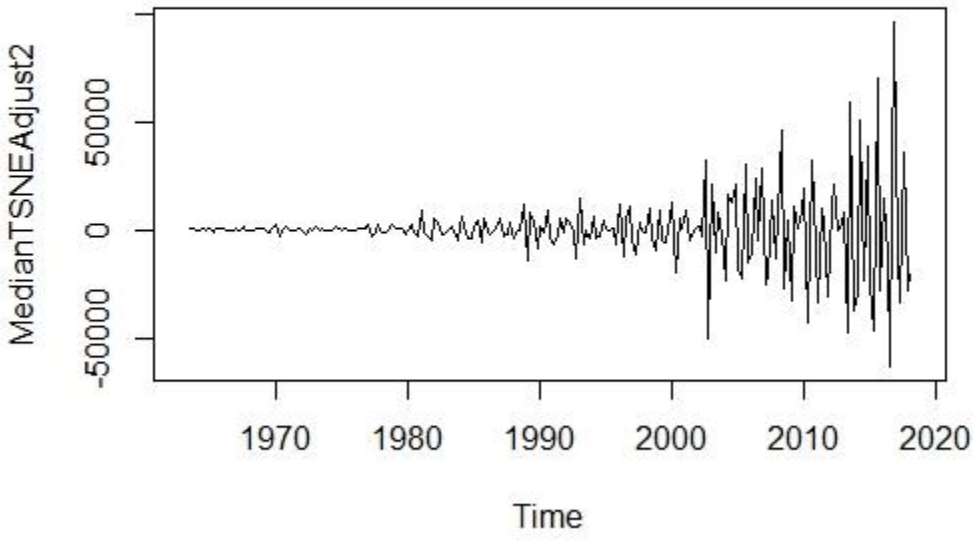
regions from above as well as the south in order to make a more comprehensive forecast. I started with the Northeast, then Midwest, then the South, then finally the Southwest. I completed the same steps as I did for the US in order to make my forecasts and compare them later.

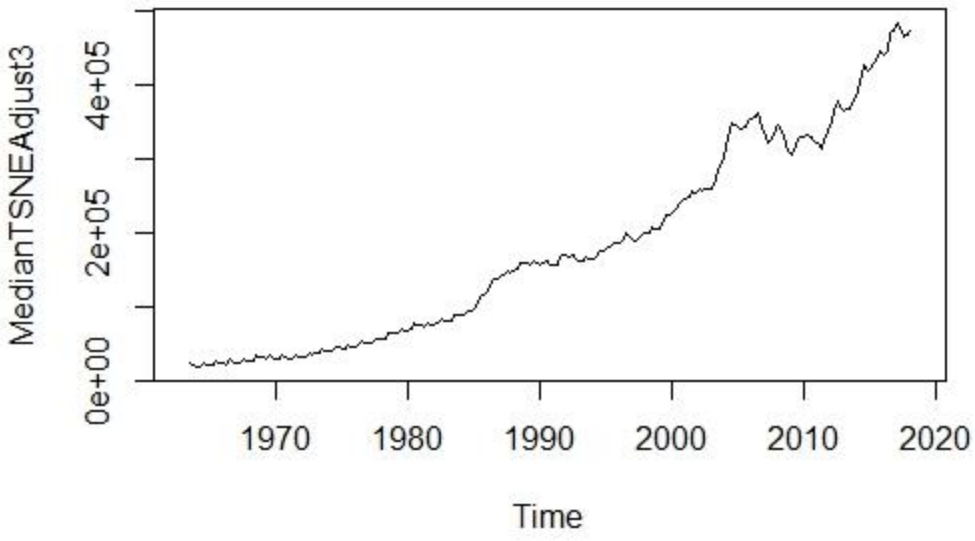
I decided to start with the Northeast

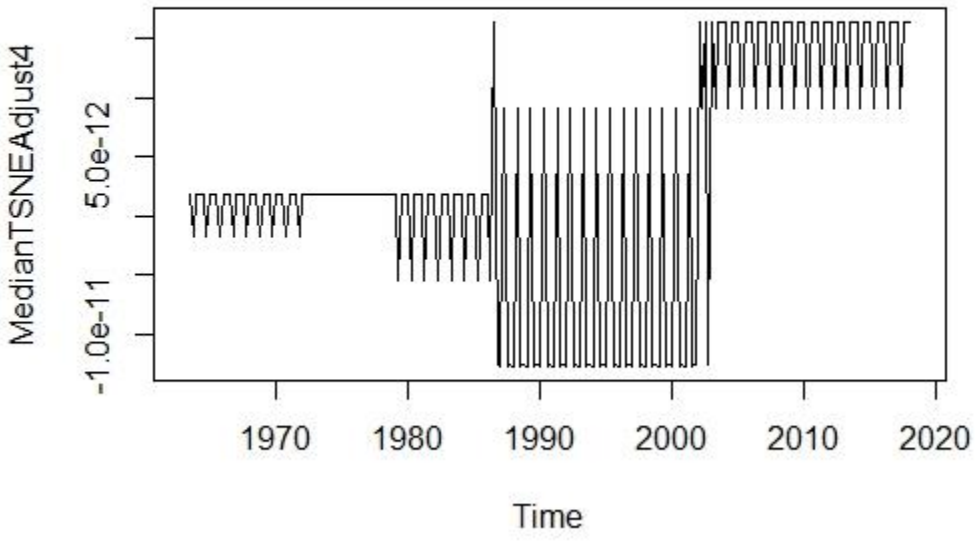


’s forecast.


MedianTSNE
0e+00
2e+05
4e+05
1970
1980
1990
2000
2010
2020
Time

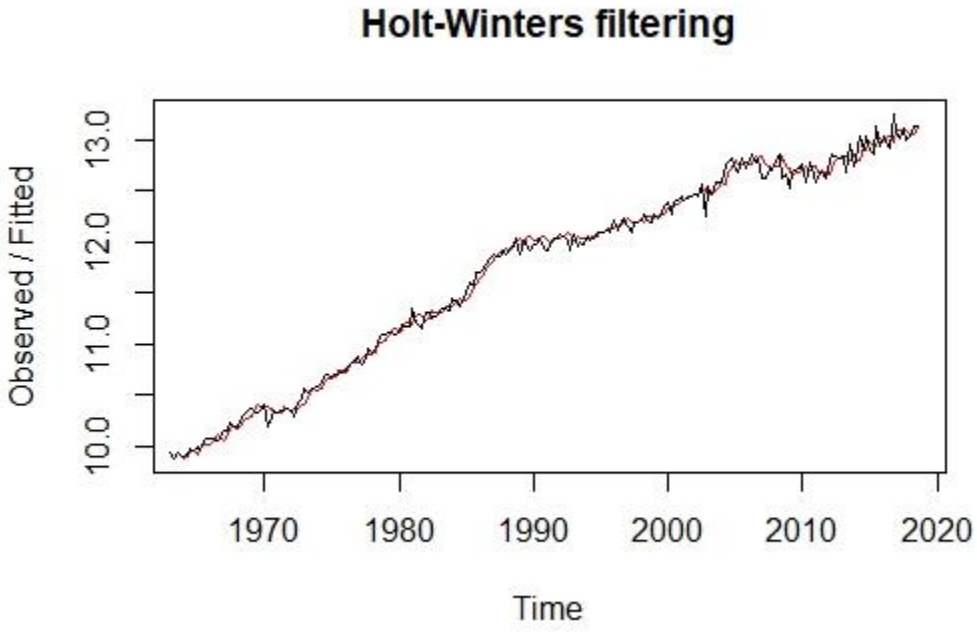

MedianTSNEAdjust2
-50000
0
50000
1970
1980
1990
2000
2010
2020
Time



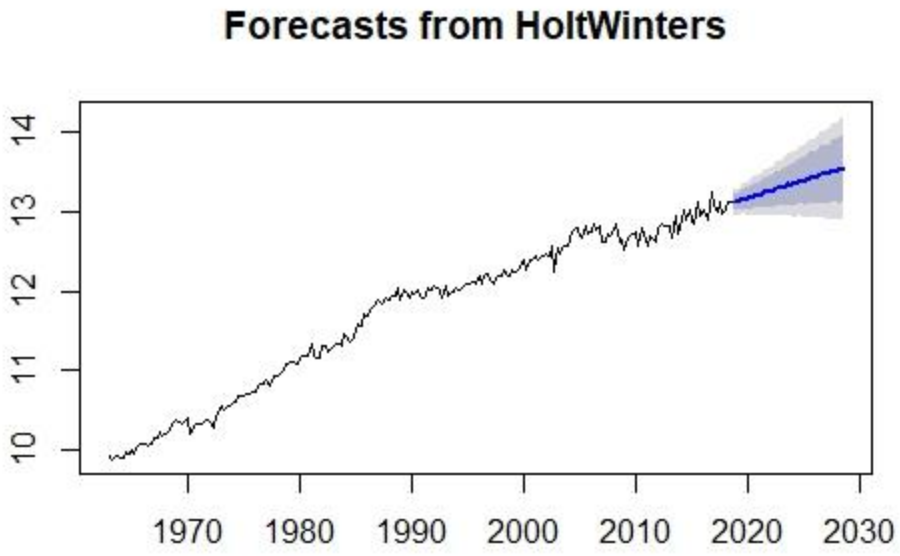


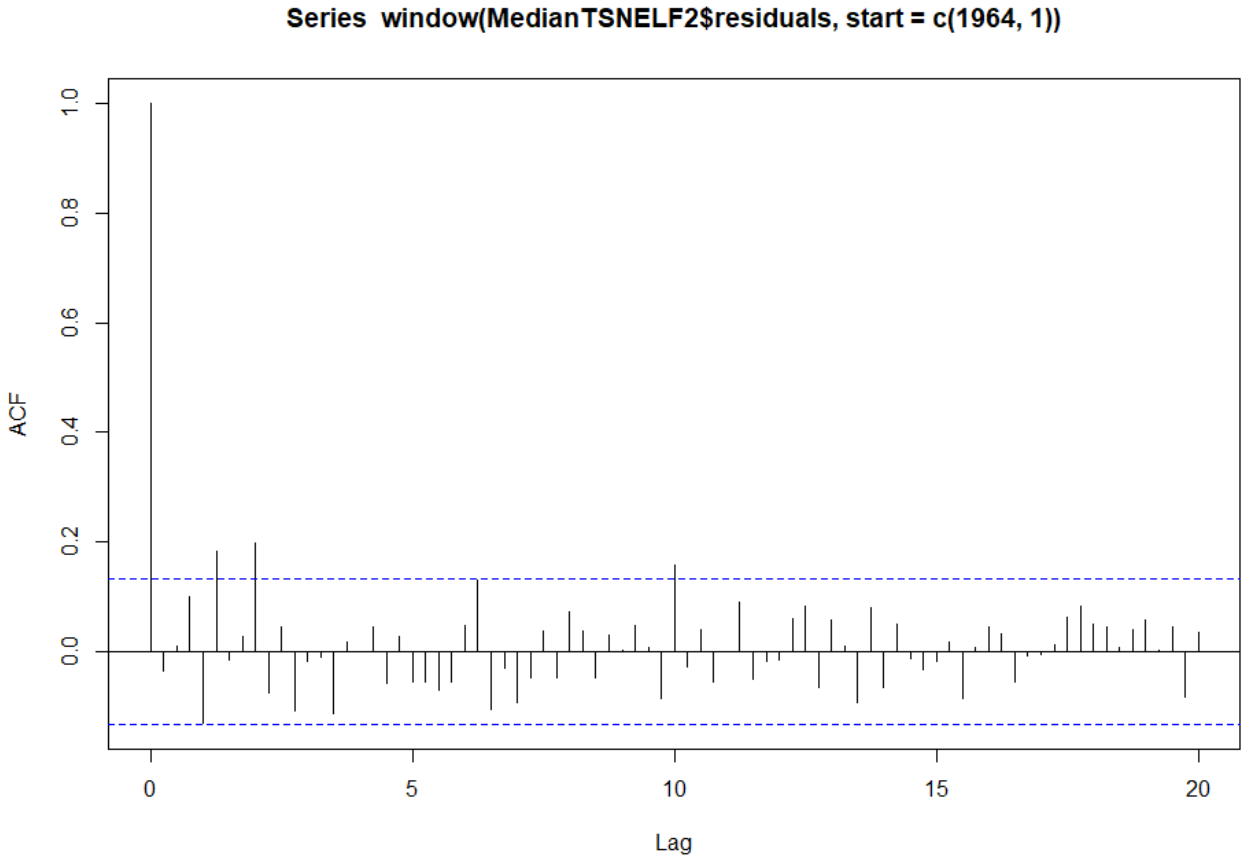
The variance here for the NE varies over time but also stays higher after 2003. There is a very large increase in seasonality in the late 80's to early 2000's. We will log transform the data again to deal with this since it is just as variable as the US data was.

```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = MedianTSNEL)

Smoothing parameters:
 alpha: 0.3840861
 beta : 0.03364569
 gamma: 0.04879961

Coefficients:
           [,1]
a  13.106701525
b   0.010929685
s1  0.001369425
s2 -0.002573501
s3 -0.006664032
s4  0.019896358
```



The forecast for the NE was then created and checked for autocorrelation.

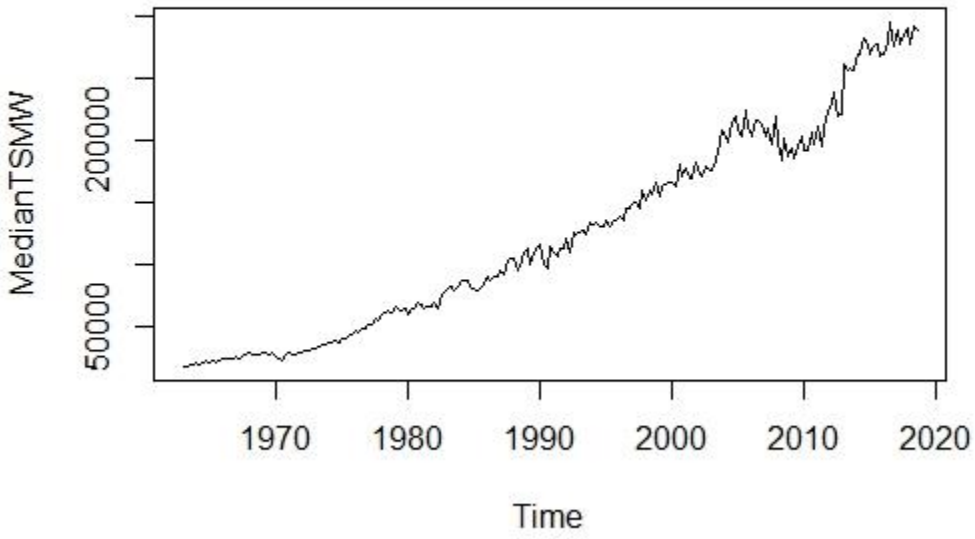

Forecasts from HoltWinters
14
13
12
11
10
1970
1980
1990
2000
2010
2020
2030

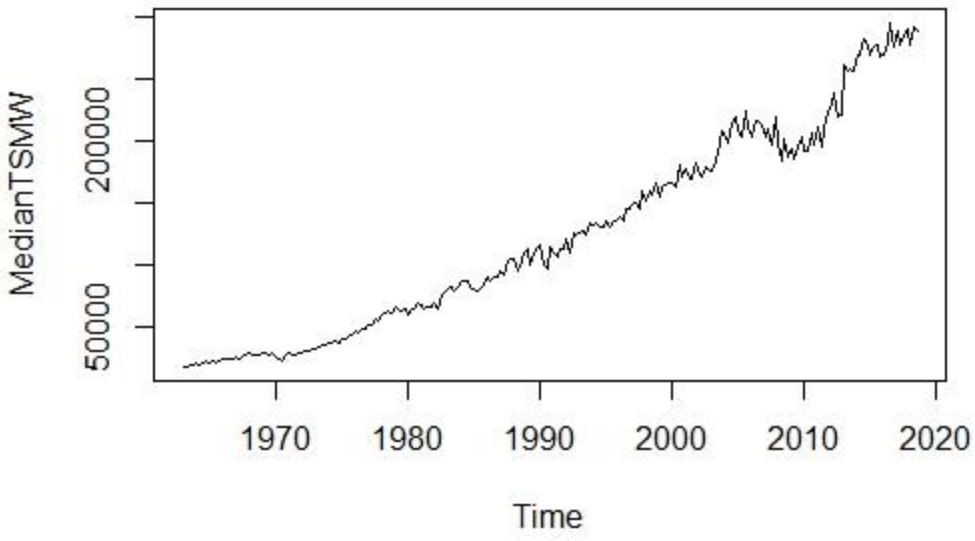

Series window(MedianTSNELF2$residuals, start = c(1964, 1))
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Lag

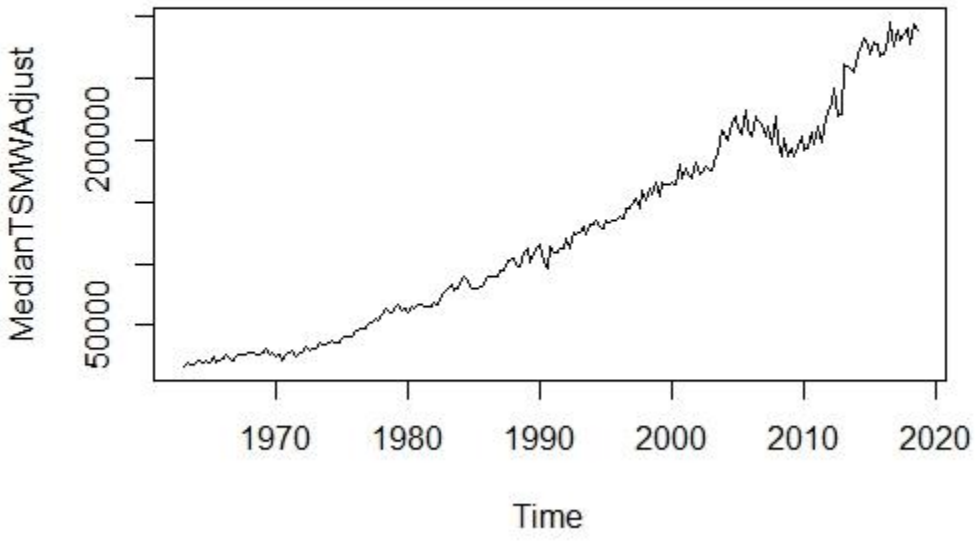
```
        Box-Pierce test

data:  window(MedianTSNELF2$residuals, start = c(1964, 1))
X-squared = 76.714, df = 80, p-value = 0.5834
```

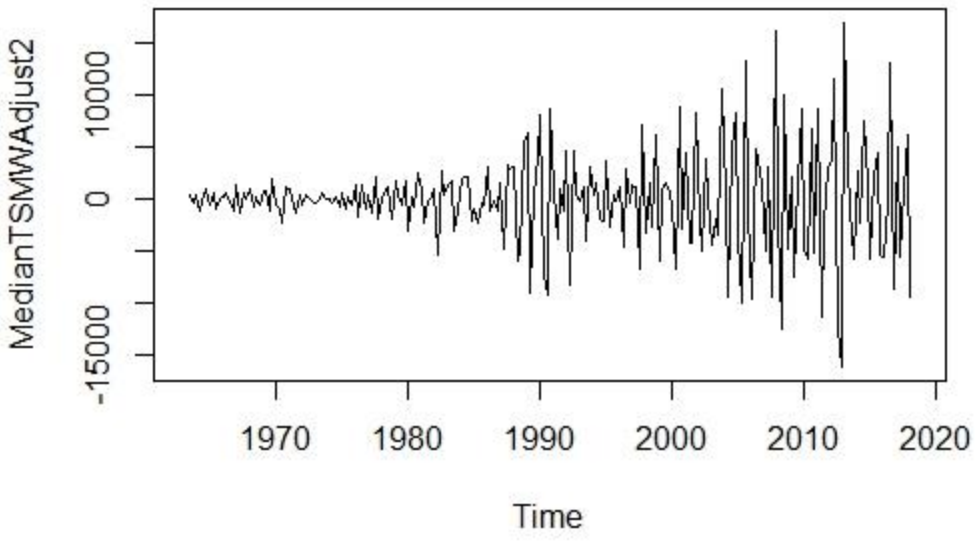
The box-pierce text and the forecast for the NE shows that there is no autocorrelation that is equal to 0. Especially since the p value is over .5

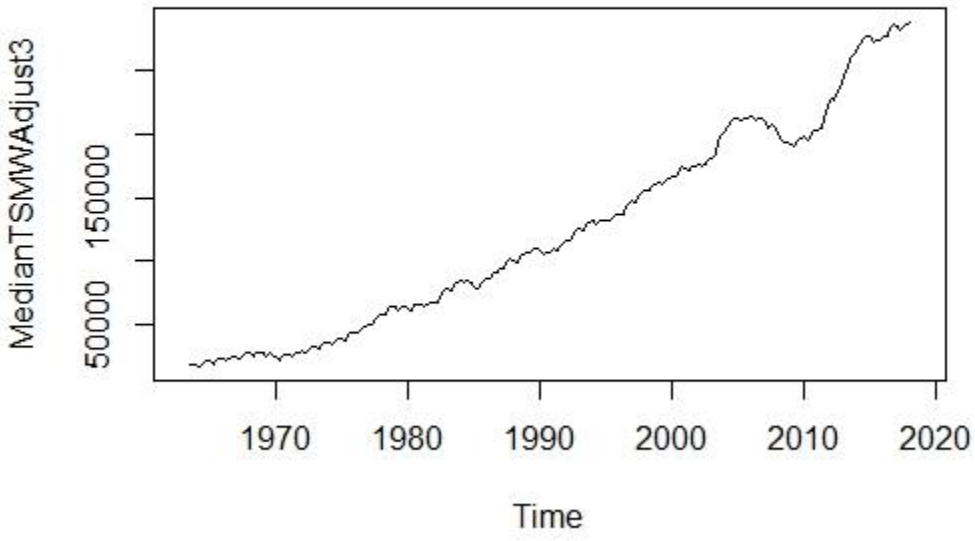
I then did the Midwest forecast to add to the comparisons.

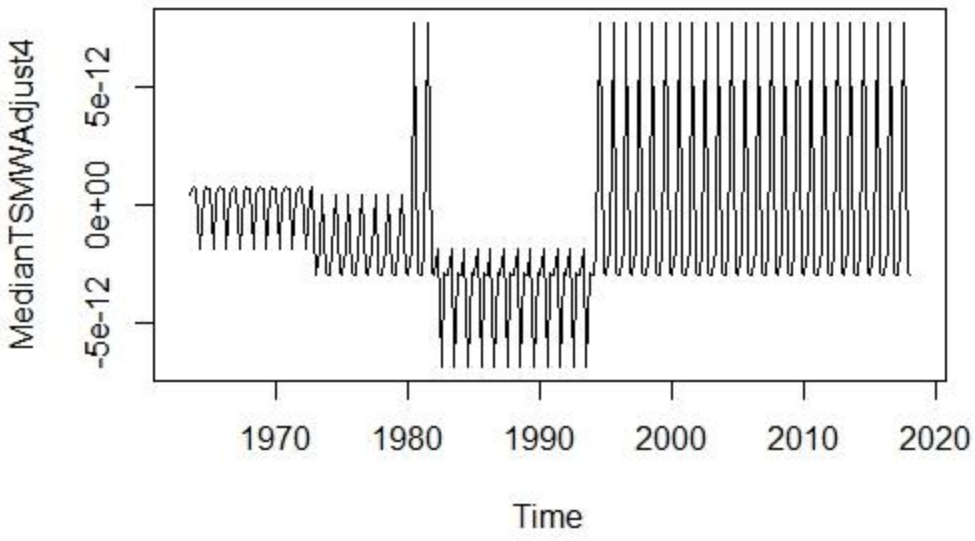



MedianTSMW
50000
200000
1970
1980
1990
2000
2010
2020
Time


MedianTSMWAdjust
50000
200000
1970
1980
1990
2000
2010
2020
Time


MedianTSMWAdjust2
10000
0
-15000
1970
1980
1990
2000
2010
2020
Time

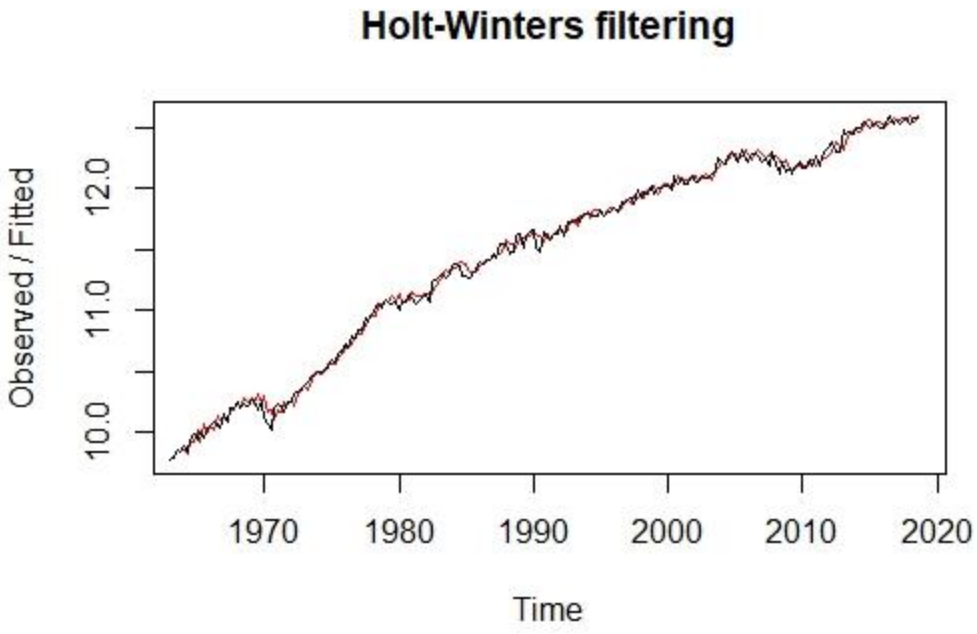

MedianTSMWAdjust3
150000
50000
1970
1980
1990
2000
2010
2020
Time



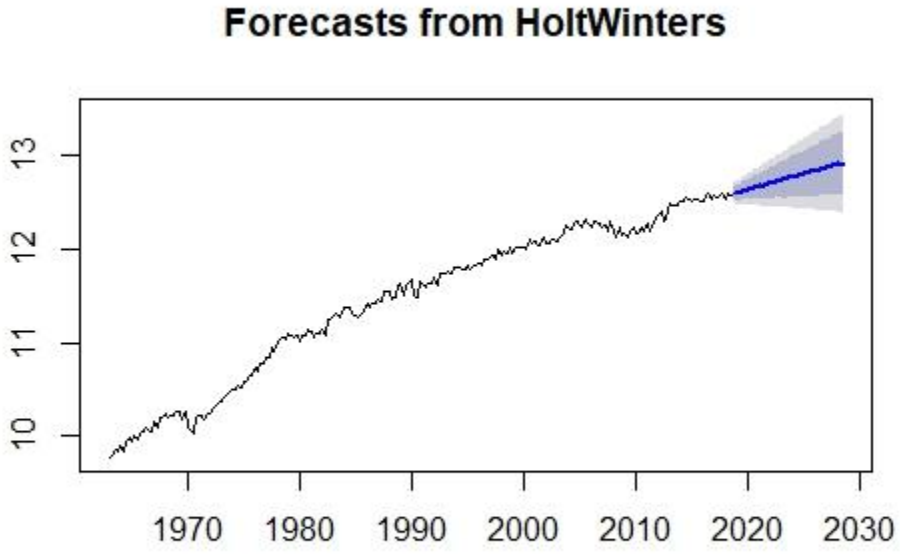
We again see a great change in the variance which increases over time. This again calls for a log transformation of the data to better assess autocorrelation.

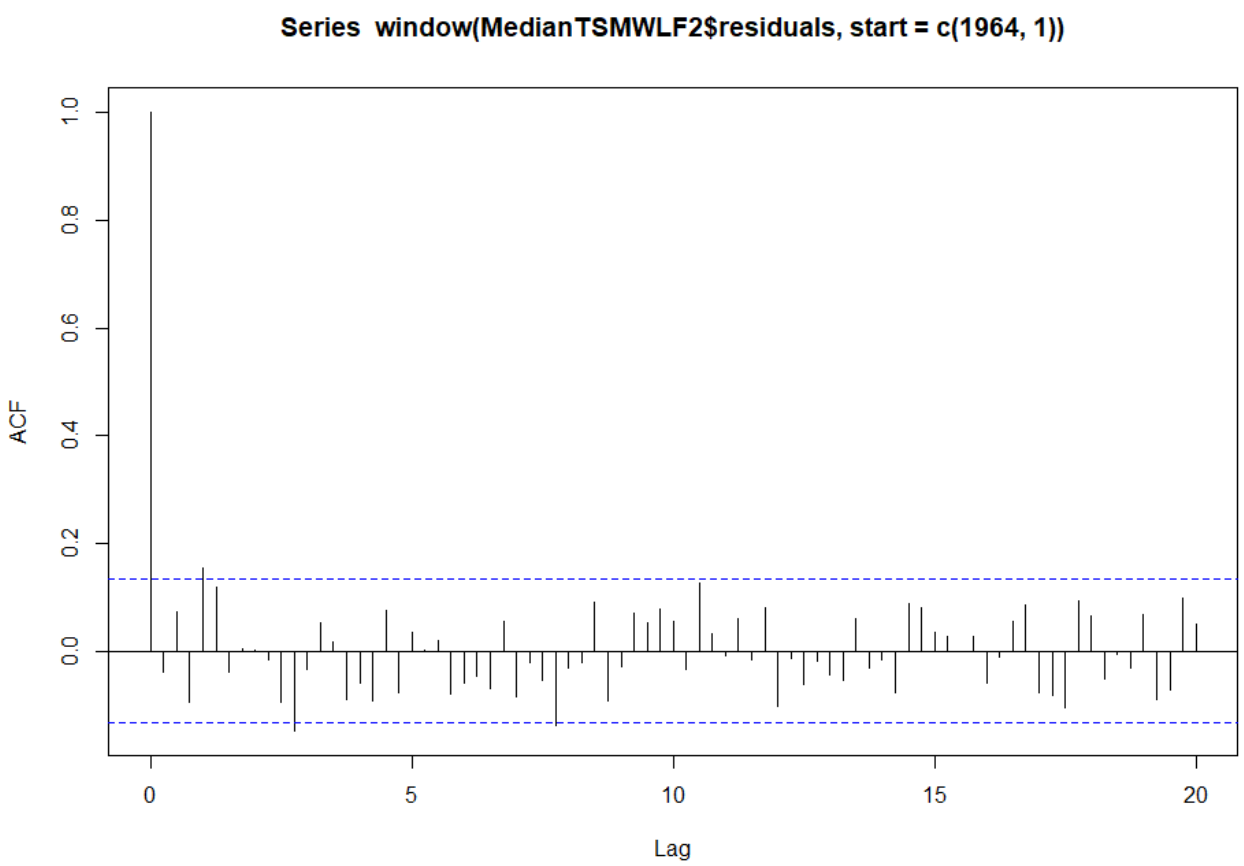
```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = MedianTSMWL)

Smoothing parameters:
 alpha: 0.5124846
 beta : 0.02604193
 gamma: 0.1638493

Coefficients:
           [,1]
a  12.599768385
b   0.008363989
s1 -0.021928713
s2 -0.018067890
s3 -0.029917276
s4 -0.016700952
```

We again see the weight distribution for Alpha, Beta, and Gamma.



I then went and created the forecast for the MW region and assessed the autocorrelation.

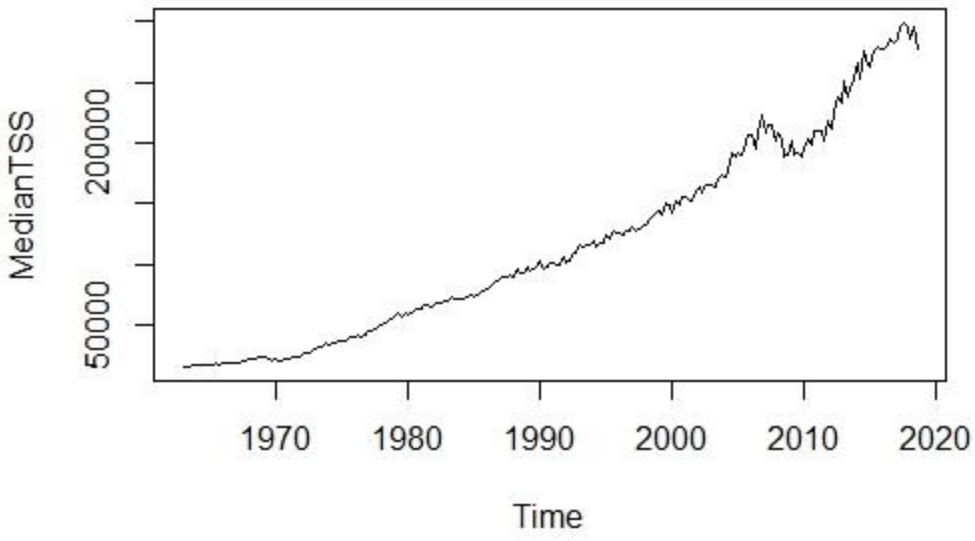


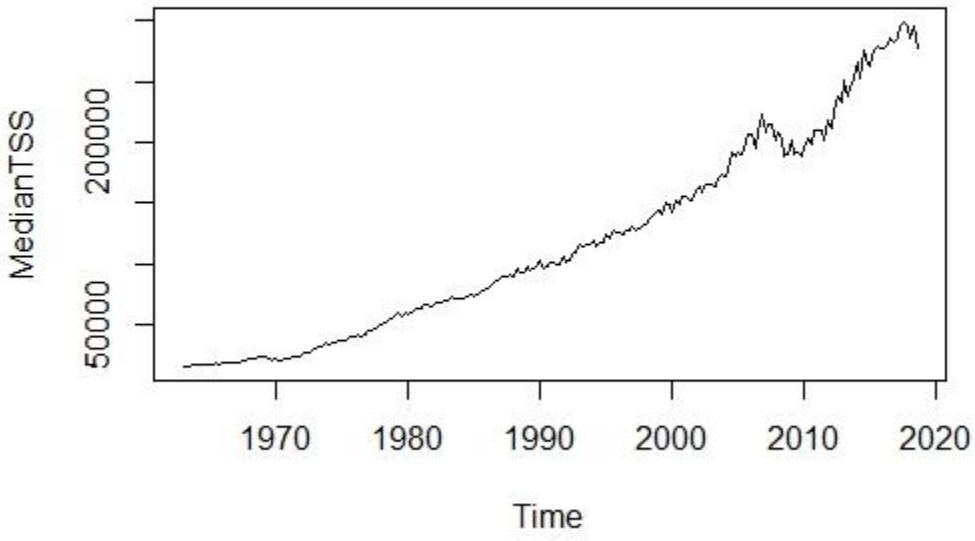


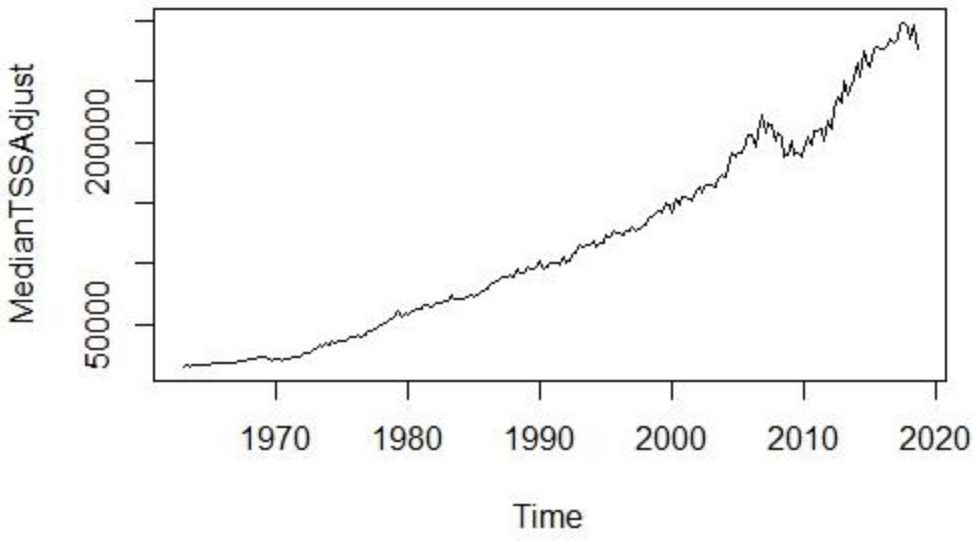

```
        Box-Pierce test

data:  window(MedianTSMWLF2$residuals, start = c(1964, 1))
X-squared = 81.124, df = 80, p-value = 0.4439
```

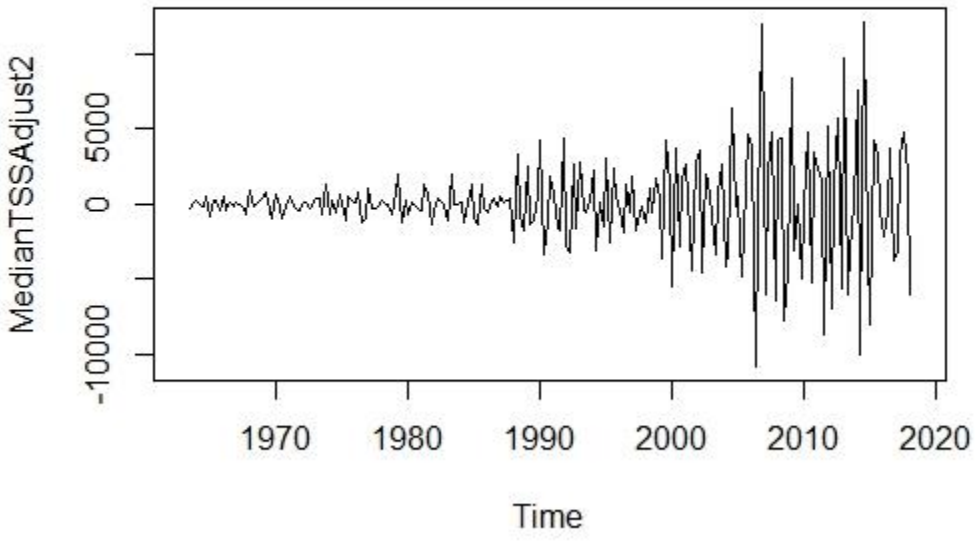
We again see here that the autocorrelation is low due to the p value not being significant as well as there not being too many points in the series window that go outside of the range.

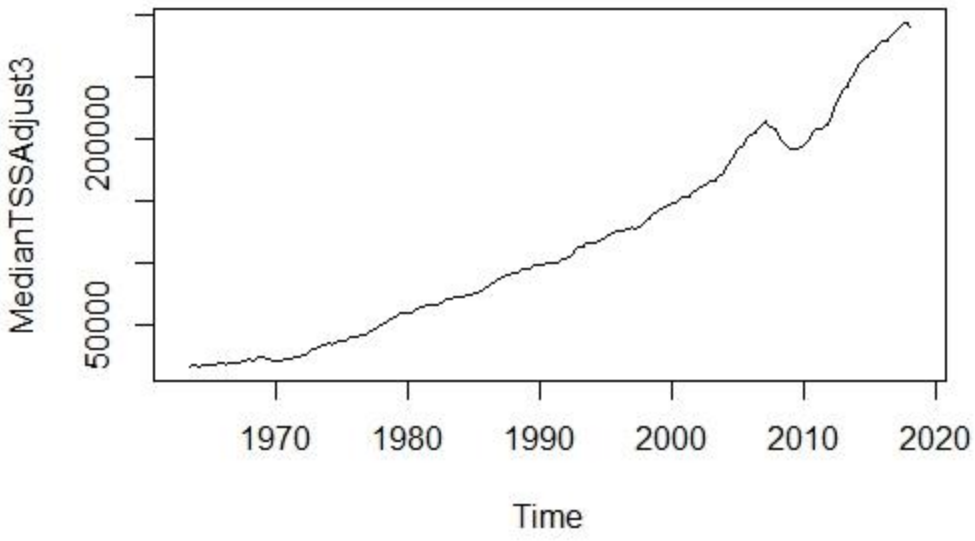
I then started the forecast for the south with the same process, placing adjustments for seasonality, assessing the seasonality, deciding and assessing the autocorrelation, and then the forecast.

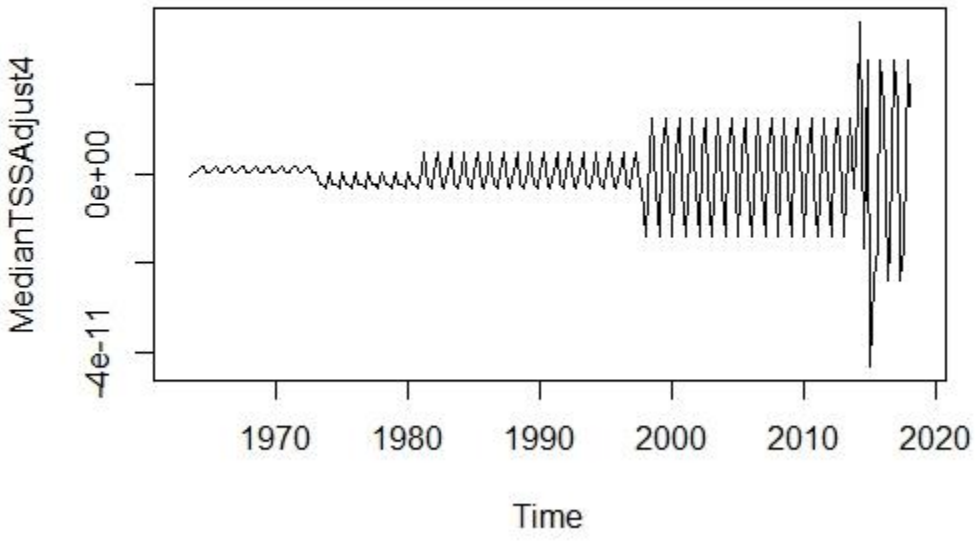

MedianTSS
50000
200000
1970
1980
1990
2000
2010
2020
Time


MedianTSS
50000
200000
1970
1980
1990
2000
2010
2020
Time


MedianTSSAdjust
50000
200000
1970
1980
1990
2000
2010
2020
Time


MedianTSSAdjust2
-10000
0
5000
1970
1980
1990
2000
2010
2020
Time

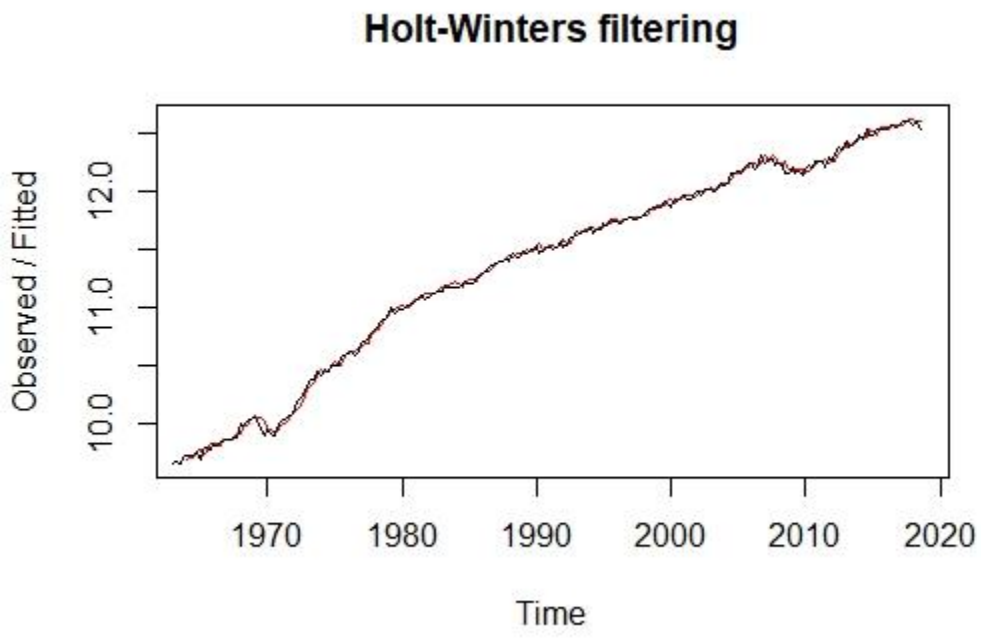




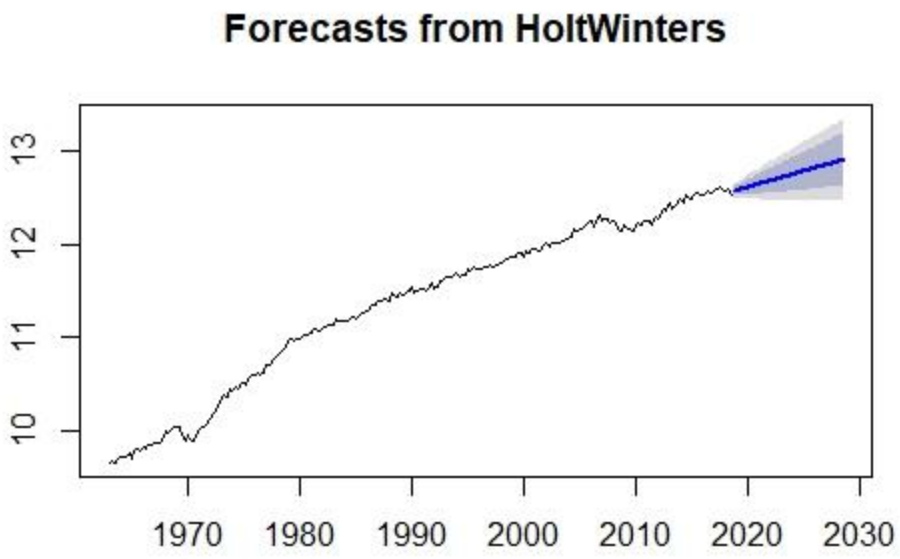
After the adjustments were in place I assessed the seasonality like the other regions. The seasonality again showed a variance that grew over time and required a log transformation in order to assess autocorrelation.

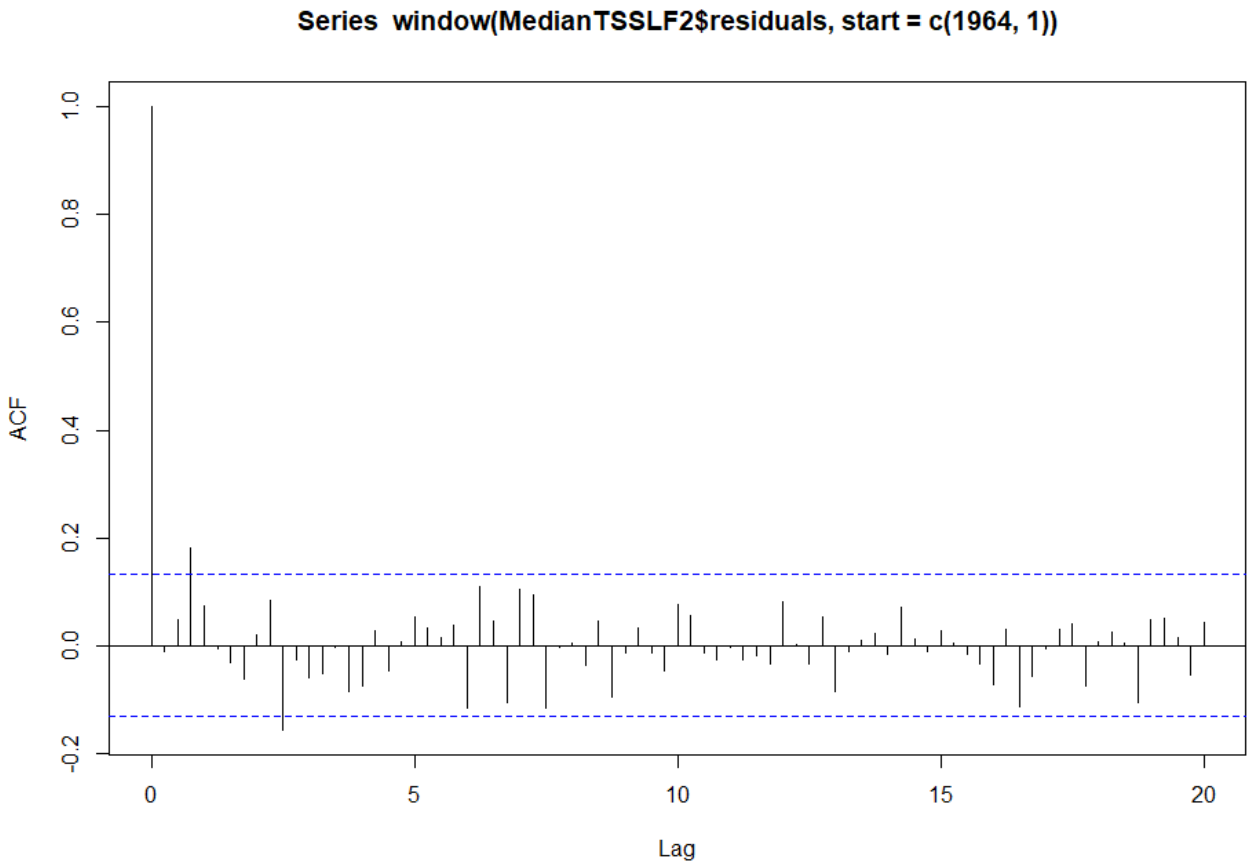
```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = MedianTSSL)

Smoothing parameters:
 alpha: 0.6645175
 beta : 0.02474877
 gamma: 0.1796906

Coefficients:
           [,1]
a  12.560236782
b   0.008637997
s1 -0.004618487
s2 -0.002840072
s3 -0.006425683
s4 -0.005616603
```

We then completed the smoothing and building of the model to assess the Alpha, Beta, and Gamma.



I then completed the forecast for comparison of the MW region to the other later on.

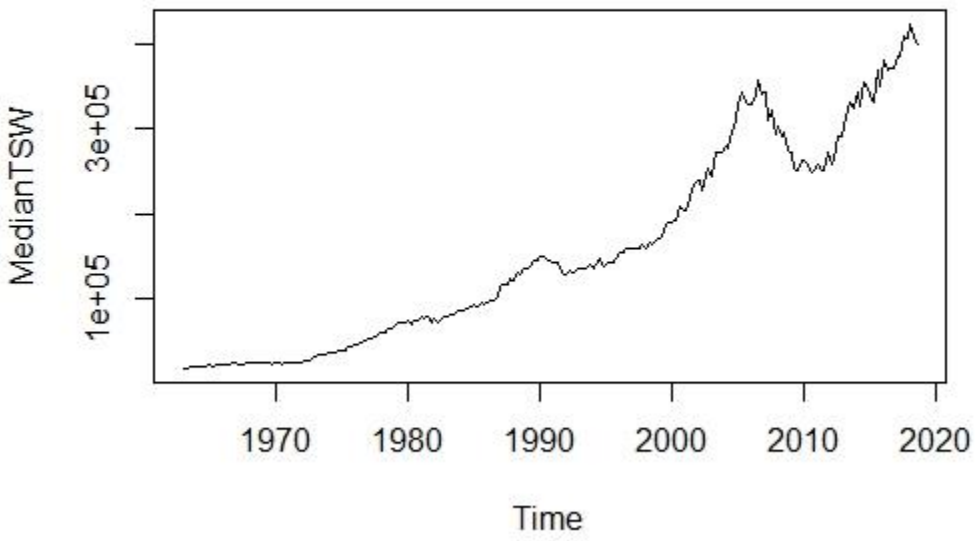

Forecasts from HoltWinters
13
12
11
10
1970
1980
1990
2000
2010
2020
2030

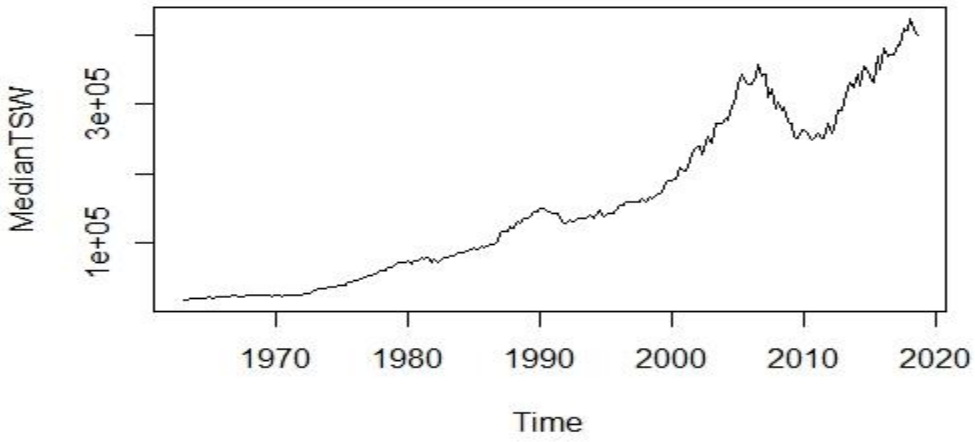

Series window(MedianTSSLF2$residuals, start = c(1964, 1))
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag

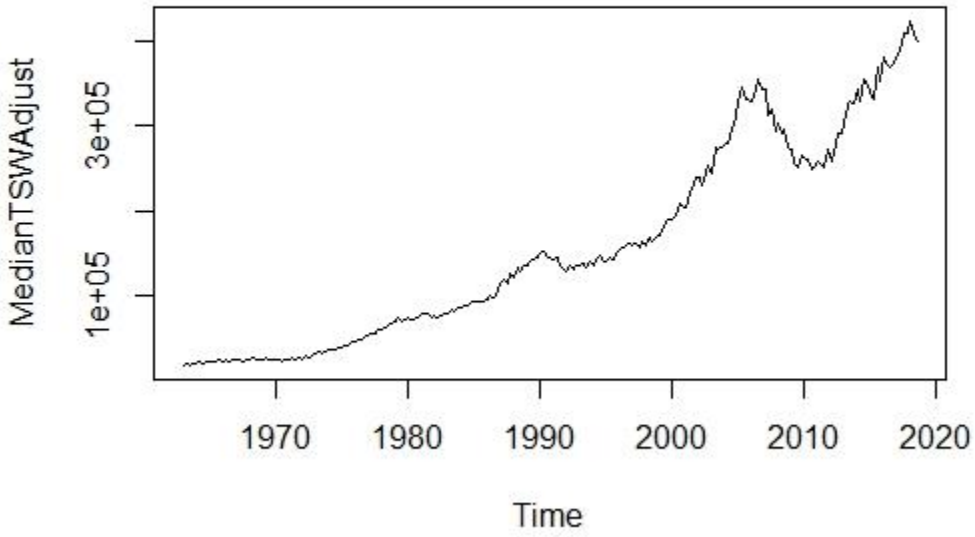
```
        Box-Pierce test

data:  window(MedianTSSLF2$residuals, start = c(1964, 1))
X-squared = 62.257, df = 80, p-value = 0.929
```

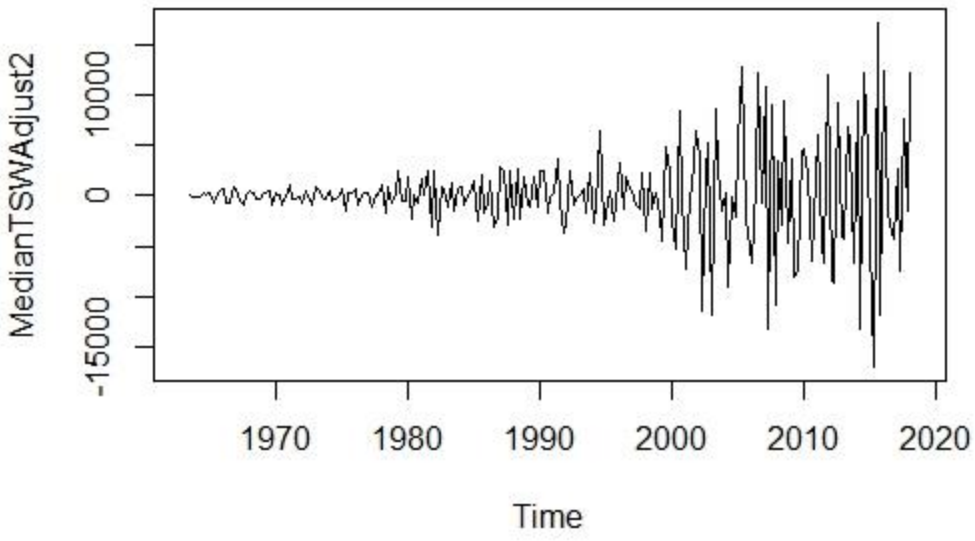
We again see that there is no autocorrelation that is not equal to 0 here due to such a high p value and very few marks on the series window that go out of range.

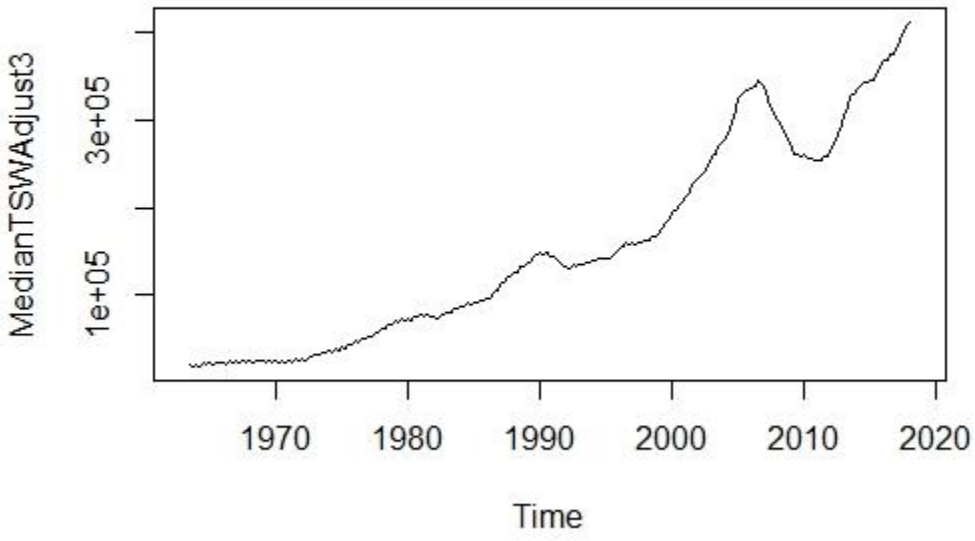
We then started the forecast for the Southwest with the same process, placing adjustments for seasonality, assessing the seasonality, deciding and assessing the autocorrelation, and then the forecast.

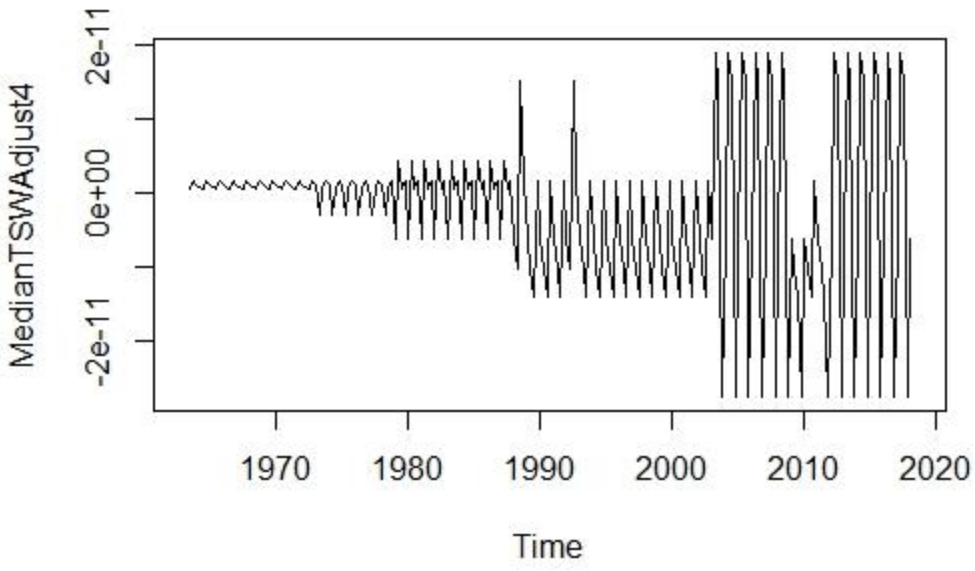



MedianTSW
1e+05
3e+05
1970
1980
1990
2000
2010
2020
Time


MedianTSWAdjust
1e+05
3e+05
1970
1980
1990
2000
2010
2020
Time


MedianTSWAdjust2
-15000
0
10000
1970
1980
1990
2000
2010
2020
Time

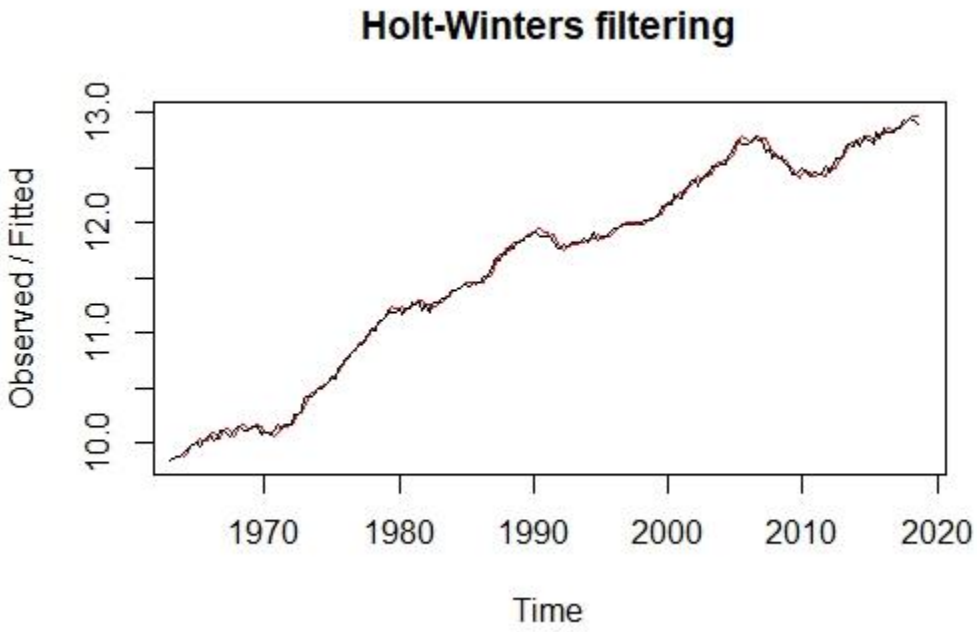

MedianTSWAdjust3
1e+05
3e+05
1970
1980
1990
2000
2010
2020
Time



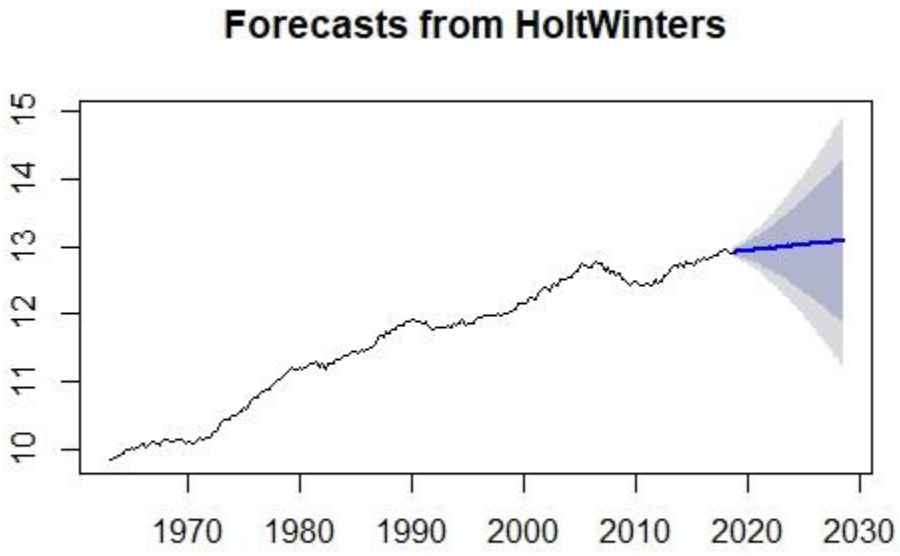
After the adjustments were in place I assessed the seasonality like the other regions. The seasonality again showed a variance that grew over time and required a log transformation in order to assess autocorrelation. This variance grew in the mid to late 2000's and on.

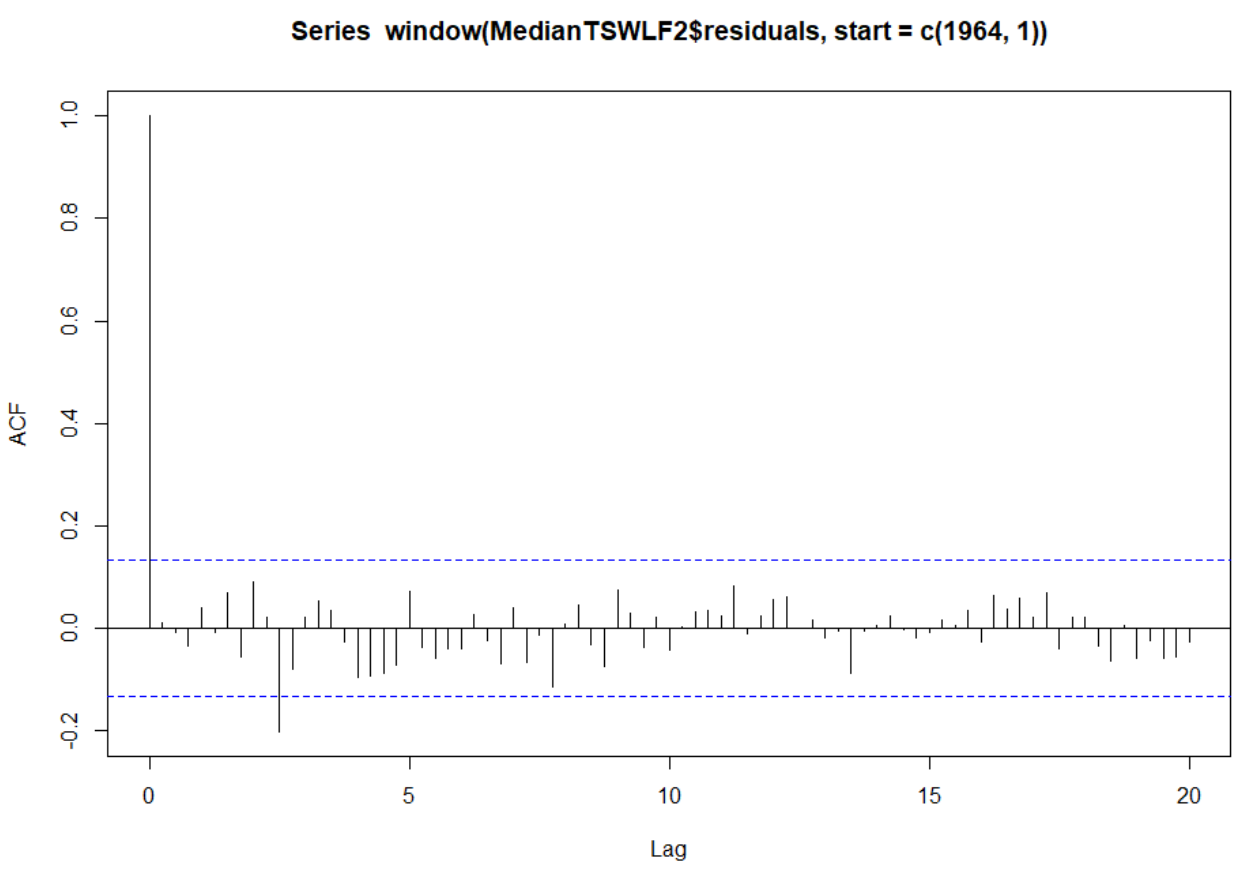
```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = MedianTSWL)

Smoothing parameters:
 alpha: 0.6584821
 beta : 0.2277654
 gamma: 0.03437146

Coefficients:
           [,1]
a  12.914038373
b   0.004301407
s1 -0.002841750
s2  0.001809947
s3 -0.006303842
s4  0.005984083
```

We then completed the smoothing and building of the model to assess the Alpha, Beta, and Gamma.



We then created the forecast for the SW region in order to help us compare the regions and the US.

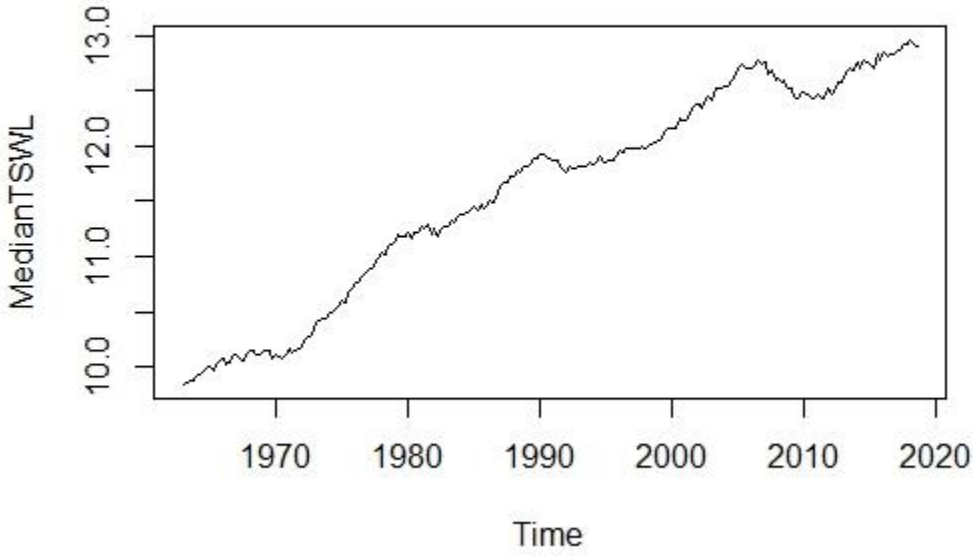


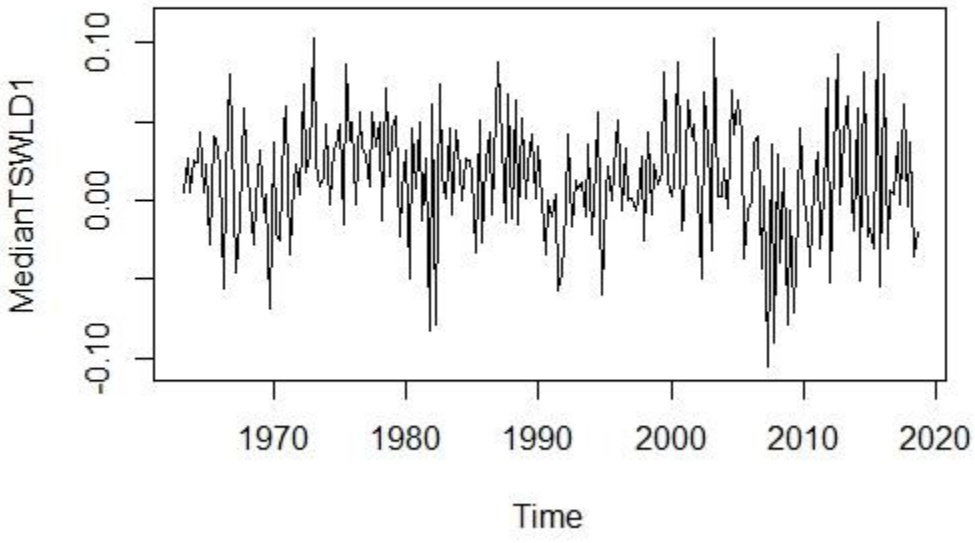



```
        Box-Pierce test

data:  window(MedianTSWLF2$residuals, start = c(1964, 1))
X-squared = 48.376, df = 80, p-value = 0.998
```

We again see that there is no autocorrelation that is not equal to 0 here due to a very high p value and very few marks on the series window that go out of range. After the forecast was created, I also wanted to make a different stationary model that was different from the above models for the Southwest. I did this since that is the region that peaked my interest due to the large dip in the market around 2010. So, I completed a ARIMA model for this region to see if it differed or confirmed the same forecast for the region.

First, I made my data stationary and then continued with building the rest of the ARIMA model.





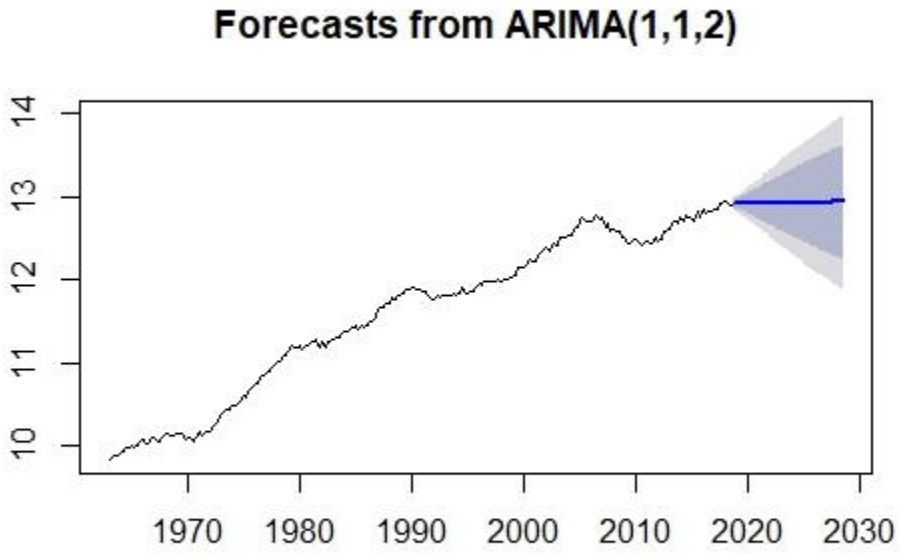
I looked at the above two plots and saw that they were drastically different in presentation. This led to me investigating more and wanting to find the other variables to create the ARIMA model by using the ARIMA function.

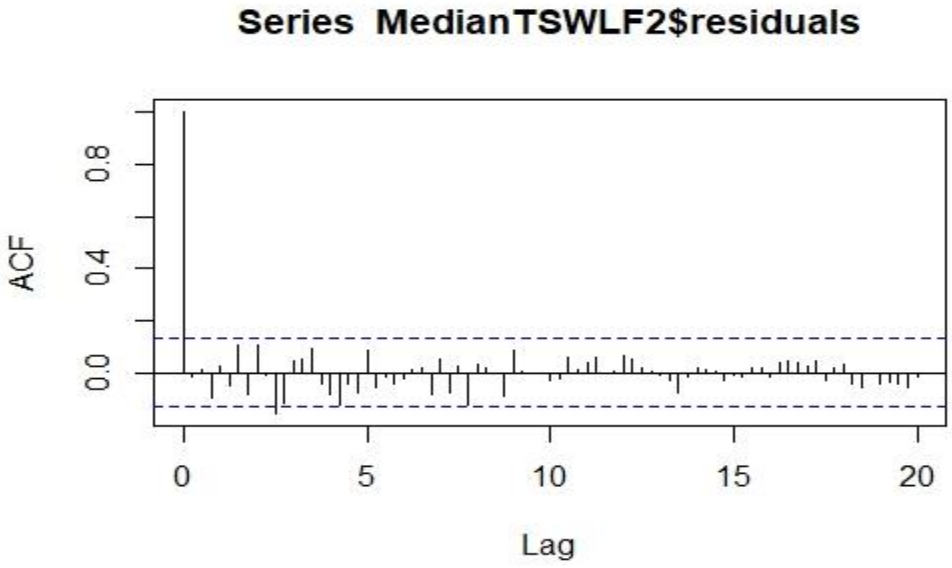
```
Series: MedianTSWL
ARIMA(1,1,2) with drift

Coefficients:
         ar1      ma1     ma2   drift
      0.8586  -1.1314  0.3652  0.0136
s.e.  0.0631   0.0873  0.0661  0.0040

sigma^2 estimated as 0.001341:  log likelihood=421.02
AIC=-832.05   AICc=-831.77   BIC=-815.03
```

```
Series: MedianTSWLD1
ARIMA(1,0,2) with non-zero mean

Coefficients:
         ar1      ma1     ma2    mean
      0.8586  -1.1314  0.3652  0.0136
s.e.  0.0631   0.0873  0.0661  0.0040

sigma^2 estimated as 0.001341:  log likelihood=421.02
AIC=-832.05   AICc=-831.77   BIC=-815.03
```

We see in the non-zero mean analysis that the middle number in the ARIMA variable is 0 which means that the difference is correct. It shows us that the auto ARIMA difference in the function is 0 through the number in the middle number for the non zero mean analysis while the original data has a 1 in the middle number. This means we now have the optimal variable for the ARIMA model and we can now forecast for the Southwest like we want.

```
Call:
arima(x = MedianTSWL, order = c(1, 1, 2))

Coefficients:
          ar1      ma1     ma2
      0.9370  -1.1779  0.3745
s.e.  0.0327   0.0703  0.0660

sigma^2 estimated as 0.001353:  log likelihood = 417.84,  aic = -827.69
```

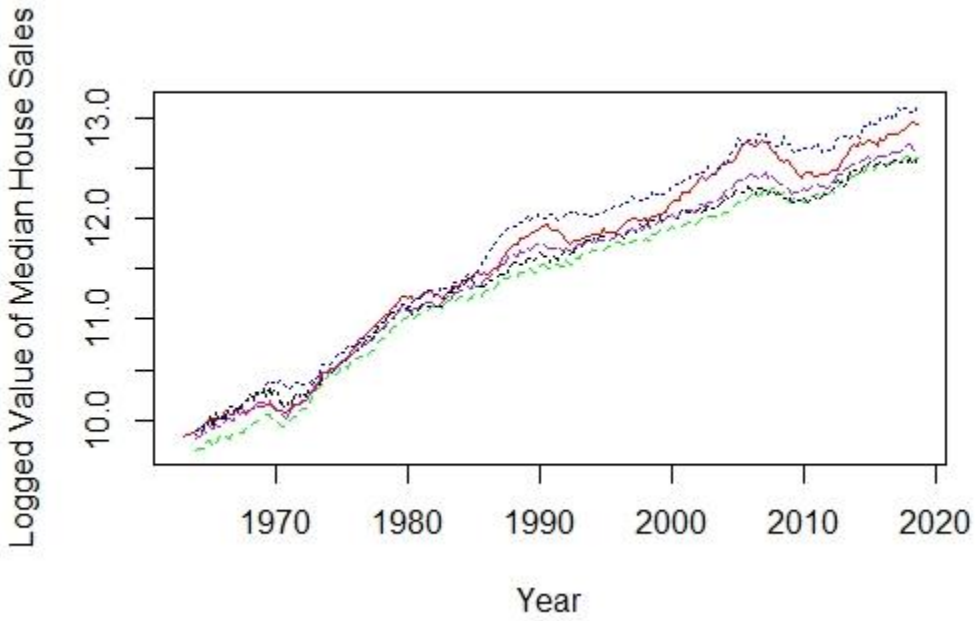




```
        Box-Pierce test

data:  window(MedianTSWLF2$residuals, start = c(1964, 1))
X-squared = 51.274, df = 80, p-value = 0.9948
```

We can see that with the ARIMA stationary forecast that it is almost identical to the forecast we did with the live data. We see that there is no autocorrelation that is not equal to 0 due to the extremely high p value which is identical to the previous one for the south, as well as there are not many marks going out of the range of the series window. The only difference is that the X-squared is slightly higher. But it still proves to be a good forecast.

Once we have all five models I wanted to compare them all on the same chart to see the differences and variations. I used color coding for the US and its different regions in this assignment.



Legend:

- Southwest: Red

- South: Green
- Northeast: Blue
- Midwest: Black
- US overall: Purple

Analysis and Interpretations

From the charts and forecasts all being compared together it looks like the Northeast and the Southwest moved ahead dramatically in the housing sales markets over a 20-year period. With the forecasts it looks like they both will keep climbing as compared to the rest of the country. However, through looking at the two forecasts for the Southwest there may be a slow down to the climb compared to the rest of the country and the regions. The stationary forecast shows a leveling off of the growth while the other forecast shows more of a slight climb compared to an incline like the rest of the markets. We can only speculate what that might mean for that region. Maybe building and regulations are different or maybe there have been some changes in buyers wants.

However, looking at all the forecasts it looks like median price is on the rise for the future no matter what region you may live in. Fore those who are looking to make an investment and gain value over time it seems like the northeast, Midwest, and the south have the steeper inclines for price going forward. I would recommend these markets to people looking to maybe flip a home or even buy one to gain value. But we have to be cautious because these forecasts do not tell us average prices or even the price at all for that matter. What it gives us are the trends of the values since we used the log transformations of our data. The final plot also only helps us to see how the markets are performing in relation to each other.

Conclusion

My suggestions for the continuation of the evaluation or even analysis of this data would be to look at the actual pricing or average pricing in the regions to get a better understanding of what the trends for pricing look like. This would have to be done after the forecasts are calculated and understood. This would be a great way to forecast for weather, and other finance related things in the future.

References:

1. Census.gov. New Residential Sales. Available at: https://www.census.gov/construction/nrs/historical_data/index.html
2. Stack Overflow. Forecast in R. Available at: https://stackoverflow.com/questions/45374807/r-forecast-holtwinters-in-forecast-package-not-found