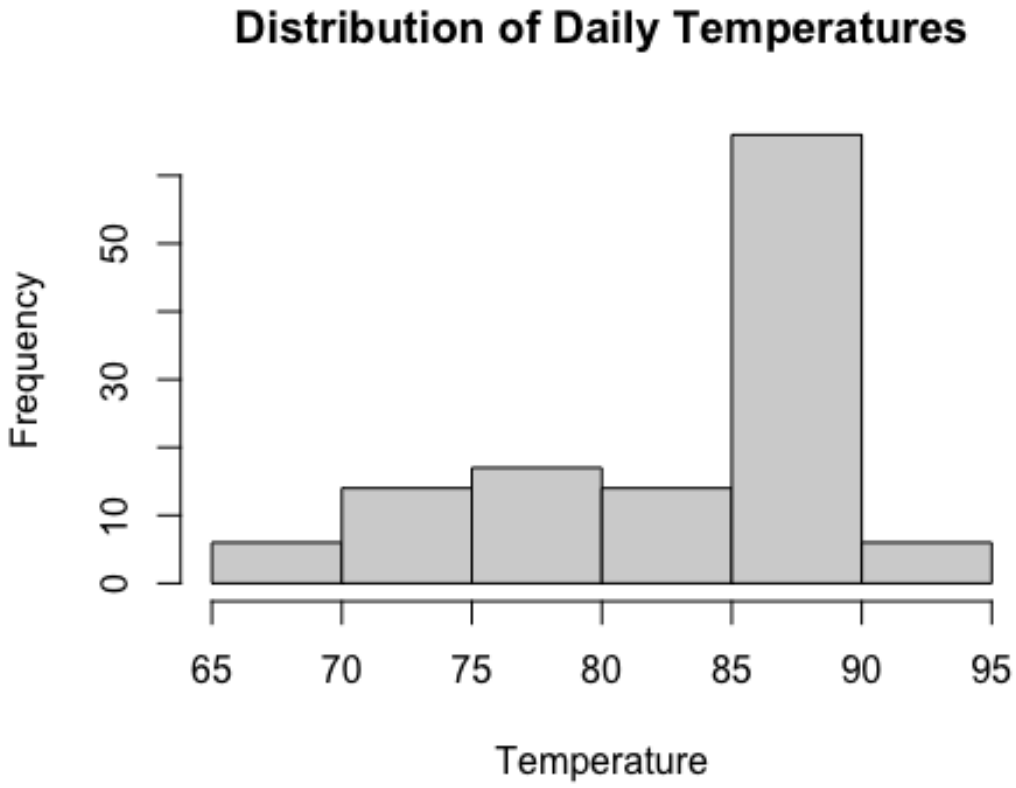
Time Series Modeling

The goal of this question was to determine whether the unofficial end of summer has gotten later over the past 20 years in Atlanta or not. To do so, a combination of approaches including time series models was used.

Data Pre-Processing

Before beginning to fit a time series model, the nature of the data needed to be analyzed. First, the distribution of temperatures between July and August from 1996 to 2015 was examined.

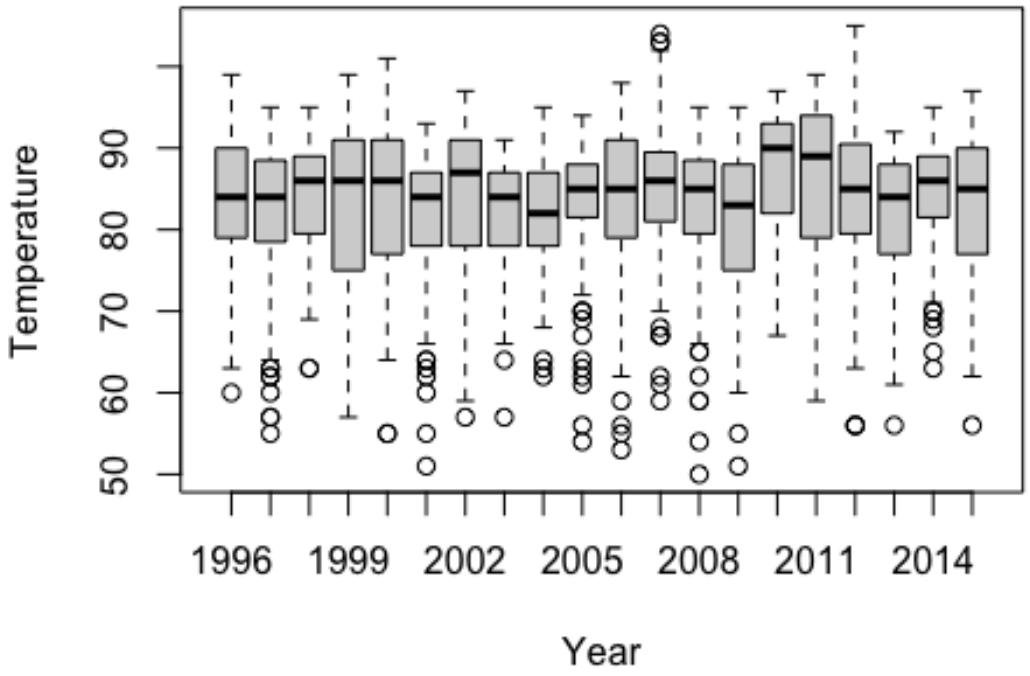


From this, it can be seen that the data is skewed to the left, however, an assumption of normality is not needed for fitting time series models. If the goal was to create forecasts with said models, however, then the data would need to be scaled. Since this is not the goal though, the data was not transformed to adjust for this left skew.

It can also be seen below that there are no missing data points in the data set. A collective outlier would be problematic here since we will be using the data to fit a time series model, and if any data points had been missing an imputation technique would have been needed.

```
sum(is.na(temps))

## [1] 0
```

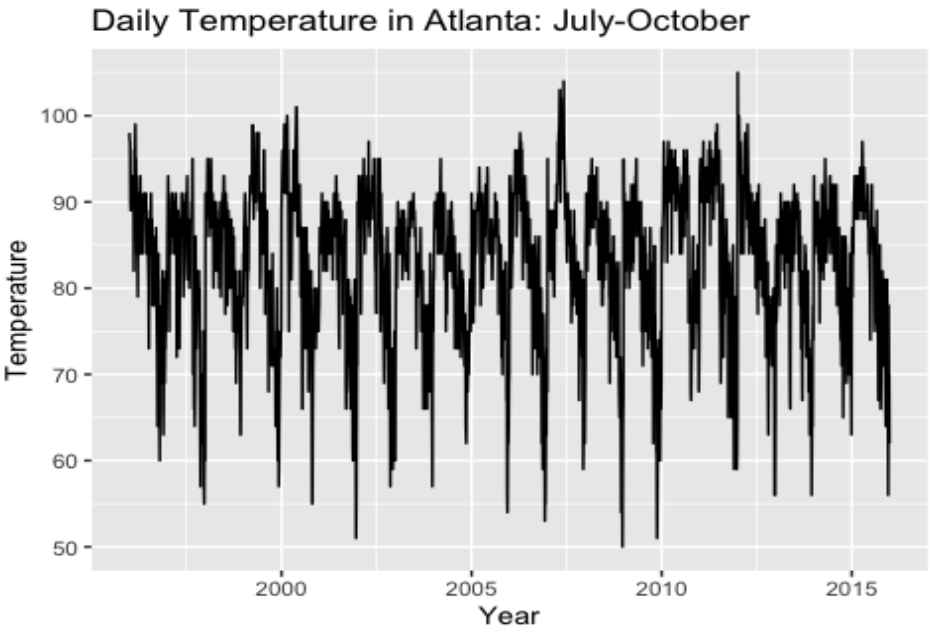
Lastly, a box plot of yearly temperatures was viewed to determine if there were any significant outliers in the dataset.



From this, it can be seen that there are a number of possible lower outliers in the data set. Given that there is some inherent randomness in temperature data and that there is no reason to believe that any data was entered incorrectly, however, no possible outliers were chosen to be imputed.

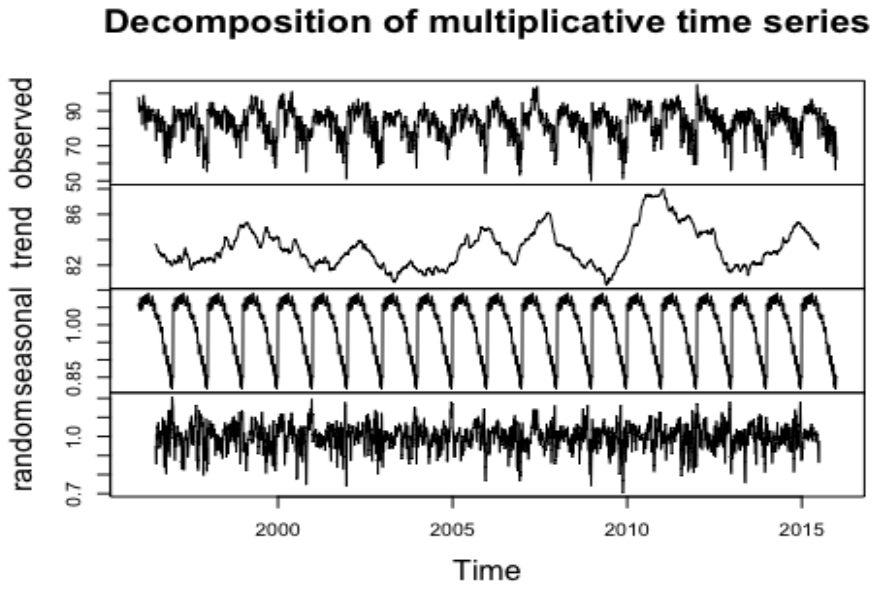
Initial Time-Series Plots

After the nature of the data had been examined, the next step was to view the time series of daily temperature data.



From this plot, a clear seasonal pattern can be seen. In addition, it does not appear that there is much of a trend across the data.

Both of these observations were affirmed when viewing the decomposed time series, which shows a clear seasonal pattern but no real clear trend in either direction. Both of these conclusions would be utilized when fitting time series models.



Fitting Time Series Models

Exponential Smoothing

First, a number of exponential smoothing models were fit on the temperature data. The whole data set was used when fitting models, and the squared sum of errors (SSE) was used to evaluate the performance of any models fit.

The first model fit was a Holt-Winters model with a multiplicative seasonal factor. This model returned an SSE of 68,904.57.

```
# Multiplicative Model
es.mult = HoltWinters(temps.ts,
                      seasonal = 'multiplicative')
## Multiplicative Model SSE
es.mult$SSE

## [1] 68904.57
```

To determine if an additive seasonal factor would be more appropriate for the data. A model using this parameter was fit next. Notably, this model yielded a better fit than the multiplicative seasonal factor, with a lower SSE of 66,244.25.

```
# Additive Model
es.add = HoltWinters(temps.ts,
                     seasonal = 'additive')
## Additive Model SSE
es.add$SSE

## [1] 66244.25
```

The alpha, beta, and gamma values for the additive triple exponential smoothing model can be seen below:

```
## Additive Model Parameters
es.add$alpha

##     alpha
## 0.6610618

es.add$beta

## beta
##    0

es.add$gamma

##     gamma
## 0.6248076
```

The alpha value showed that there was some randomness in the data but that differing values were mainly a real indicator of change, since the alpha value favored the observed data. Additionally, the best beta value for the model was 0, showing that there was no real trend in the data. This supported the conclusion drawn earlier when viewing the decomposed time series, where no clear trend was found. Furthermore, there was a clear seasonal pattern which was not mainly a result of randomness, since the gamma value favored the observed data.

ARIMA

Next, an ARIMA model was fit to determine if it resulted in a better fit for the data. This model was also analyzed using the sum of squared errors to compare it to the best exponential smoothing model found previously.

A generic ARIMA model for exponential smoothing was fit with an additional seasonal component, since it was already determined that there were clear seasonality in the data. The model used was ARIMA (0,1,1)(0,1,1). This model returned a SSE of 49,992.02, which was notably smaller than the best exponential smoothing model found previously.

```
# ARIMA (0,1,1)(0,1,1) Model
arima.fit = arima(temps.ts,
      order = c(0,1,1),
      seasonal = c(0,1,1)
      )
## ARIMA (0,1,1)(0,1,1) Model SSE
sum(arima.fit$residuals^2)

## [1] 49992.02
```

Model Combinations

To see if a combined approach would yield a more accurate model, the best exponential smoothing and ARIMA model were combined and their SSE was calculated.

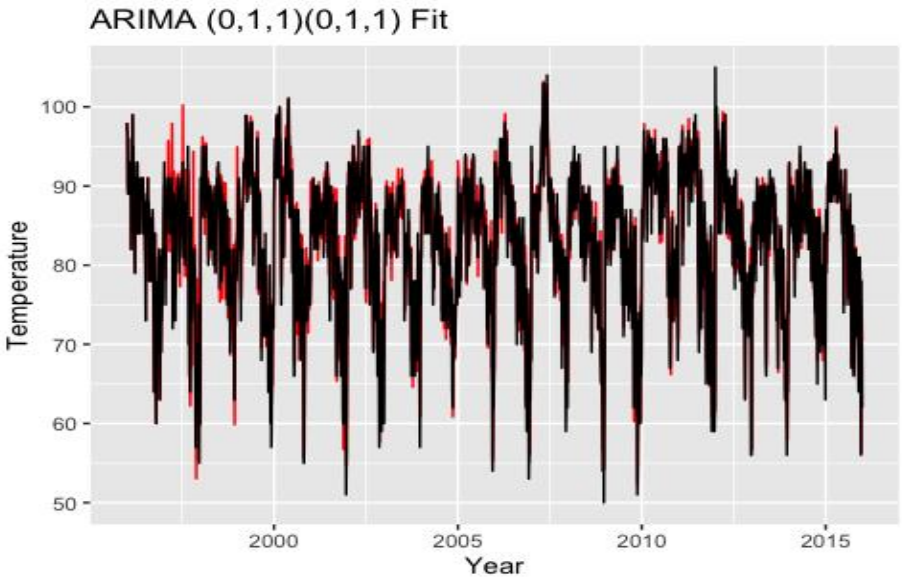
```
# Combined Model
arima.fitted = temps.ts - arima.fit$residuals
combined.fit = (es.add$fitted[,1] + arima.fitted) / 2
## Combined Model SSE
sum((combined.fit - temps.ts)^2)

## [1] 54847.37

## ARIMA (0,1,1)(0,1,1) Model SSE
sum(arima.fit$residuals^2)

## [1] 49992.02
```

This combined model returned a SSE of 54,847.37, which was not lower than the best ARIMA model.

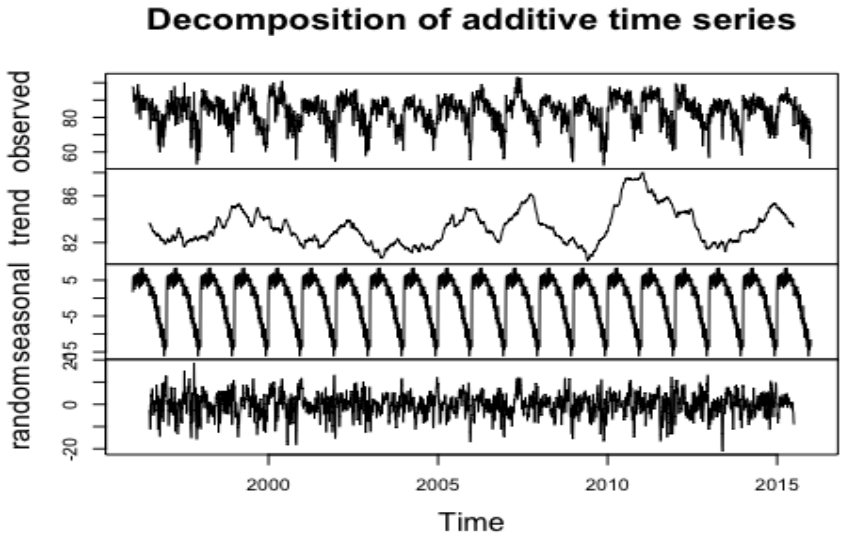
Analyzing Best Model: ARIMA (0,1,1)(0,1,1)

The fit for the best model found, the ARIMA (0,1,1)(0,1,1) model, can be seen below:

```
## Don't know how to automatically pick scale for object of type ts. Defaulti
ng to continuous.
```



When viewing the decomposition of the time series of fitted values for this model it can be see that there is no clear trend and a relatively normal white noise in the randomness, both of which affirmed the findings from the exponential smoothing model earlier. Additionally, it can be seen that the seasonal pattern remained consistent across the data. When examining whether or not the unofficial end of summer has gotten later over the past 20 years in Atlanta, this would not seem to provide evidence of this as there is no change in the seasonal pattern.



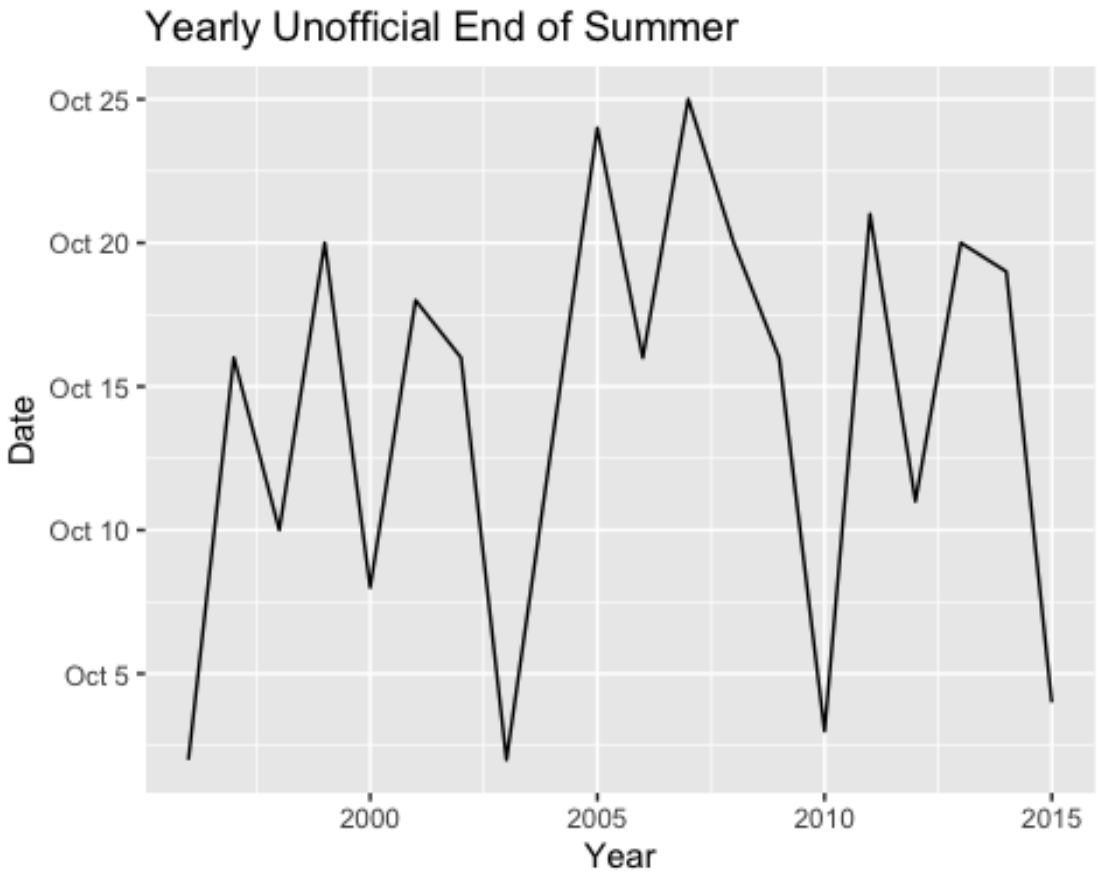
Affirming Conclusions: CUSUM

To affirm the conclusion that the unofficial end of summer has not gotten later over the past 20 years in Atlanta, the CUSUM approach for a Change Detection model was used as well.

To do so, end of the summer season was determined for each year. This was done by detecting a decrease in temperature for each year using the CUSUM approach. The critical value was set at one standard deviation, while the threshold was set at three standard deviations. The date for which the threshold was crossed for each year was then stored.

```
# df to store dates where threshold crossed for each year
yearly_dates = c()
# CUSUM Yearly
for(y in 2:21){
  # subset df
  yearly_temps = data.frame(Day = temps$Day,
                            Temp = temps[,y])
  # yearly mean
  yearly_mean = mean(yearly_temps$Temp)
  # yearly C
  yearly_c = sd(yearly_temps$Temp)
  # yearly T
  yearly_T = yearly_c*3
  # Generate St
  yearly_temps[1,"S.Value"] = 0
  for(i in 2:nrow(yearly_temps)){
    yearly_temps[i,"S.Value"] = max(0,yearly_temps[i-1,"S.Value"]+(yearly_mea
n-yearly_temps[i,"Temp"]-yearly_c))
  }
  yearly_dates[y-1] = yearly_temps[yearly_temps$S.Value > yearly_T,]$Day[1]
}
```

Below a plot can be seen of when the threshold was crossed for each year:



Notably, the date the the threshold was crossed seems to vary from year to year, but with no real trend in either direction.

By analyzing the yearly end of summer as a time series using the Holt Winters method (no seasonality), it can be seen that the value for beta was 0.4357513, meaning that the trend present was mainly a result of randomness.

```
end.fit = HoltWinters(end.ts,
            gamma = F)
end.fit$beta

##      beta
## 0.4357513
```

This would affirm the conclusion from the ARIMA model that there has been no change in the unofficial end of summer from the period of 1996 to 2015 in Atlanta.