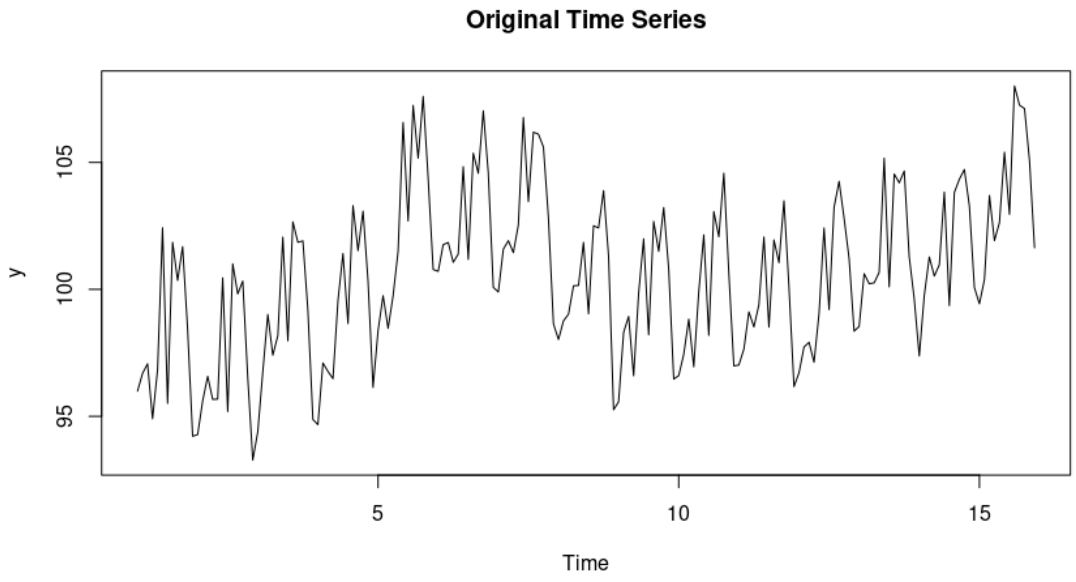
Introduction

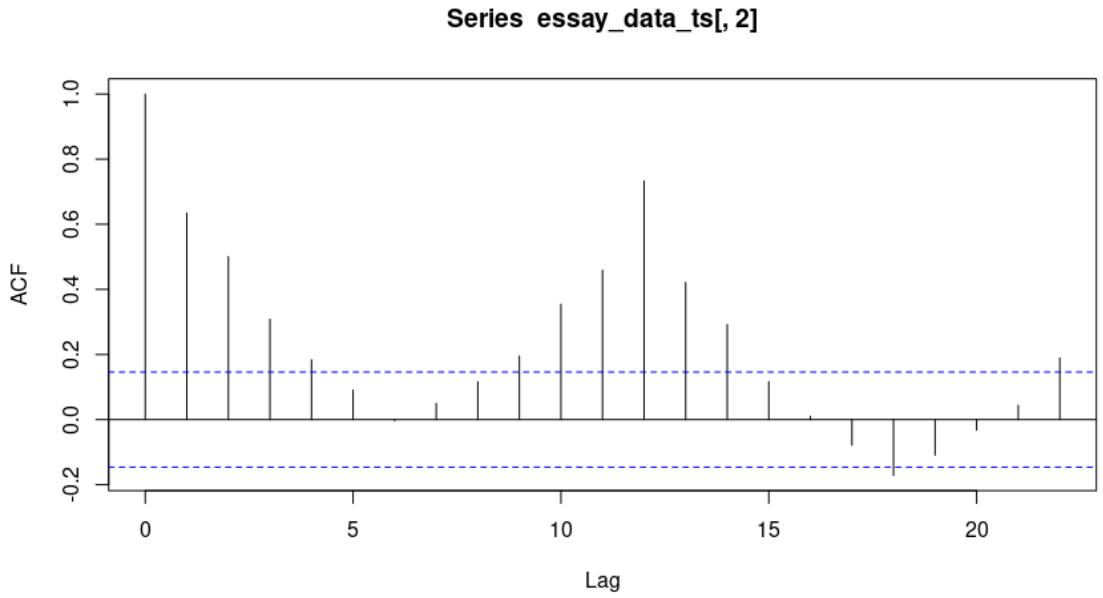
Forecasting data is an integral part of many jobs, such as those in consulting or research. As such, it is important to look at a set of data and be able to understand what patterns may appear in data, decide which model best works in capturing said patterns, estimate these models and properly forecast with these models. In this paper, we will analyze a time series dataset to be able develop the proper skills needed to create forecasts, applying our knowledge so far in ECO374. Using our knowledge, we will identify the correct ARIMA model to be used and use that to develop a 12-period forecast.

Cleaning the Data and Verifying Assumptions

The first step in this process is to clean and manipulate the data, then attempt to identify any patterns within the data, to do so, we can analyze a few things such as the plot of the time series itself.



Looking at the time series plot, we can see that the data may have a trend, either deterministic or stochastic is to be determined. However, since one cannot just eyeball it to determine what kind of trend it is we have to look at the ACF and Augmented Dickey Fuller (ADF) Test to see if the data may be non-stationary.



We can see that in the ACF, the data does not die down immediately as it rebounds later on, as such this suggests a non-stationary set of data. Moreover, there seems to potentially be some evidence of seasonality in the data as it looks cyclical near the latter half of the plot, so we may also want to correct that as well when we choose our ARIMA model. We can also look at the ADF to determine if it is stationary, the results are shown below.

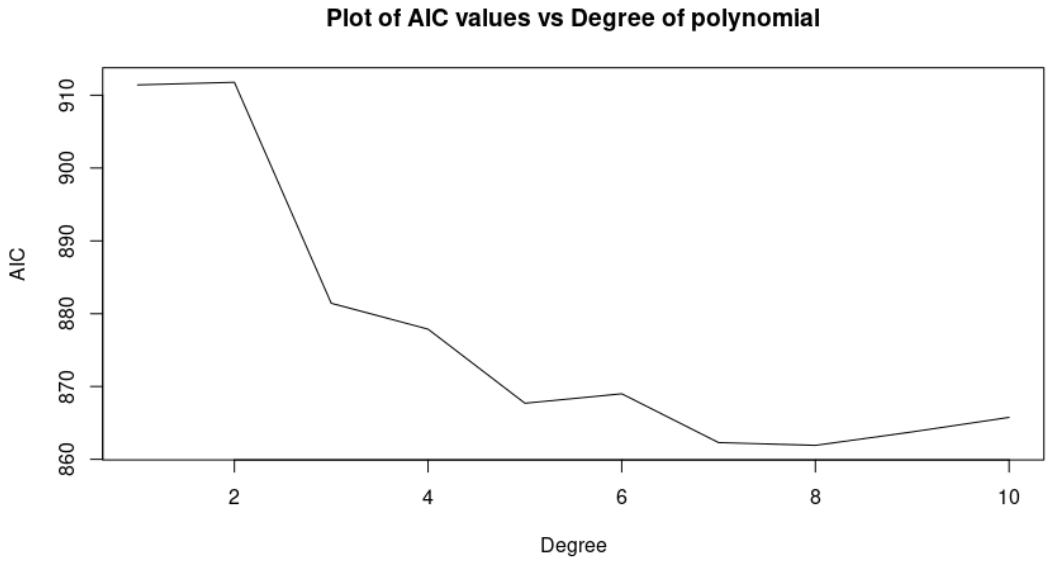
```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag     ADF p.value
[1,]   0 -0.0274   0.636
[2,]   1  0.0986   0.672
[3,]   2  0.0975   0.671
[4,]   3  0.2076   0.703
[5,]   4  0.2439   0.713

Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -6.34    0.01
[2,]   1 -4.82    0.01
[3,]   2 -4.98    0.01
[4,]   3 -4.95    0.01
[5,]   4 -4.64    0.01

Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -6.76    0.01
[2,]   1 -5.16    0.01
[3,]   2 -5.38    0.01
[4,]   3 -5.34    0.01
[5,]   4 -5.06    0.01
```

So based on the ADF outputs, we cannot reject the null hypothesis of non-stationarity and hence it suggests the time series is not stationary, since the p-values are not less than 0.05 significance for all the types, namely the values in the Type 1 section. As such, we will likely have to correct these discrepancies before moving forward. To do so, we can regress on a time constant to capture any deterministic trends in our data, which would leave us with a residual series. However, this series looks that it may be polynomial, and as a result, we will have to look at regressions using different numbers of polynomials

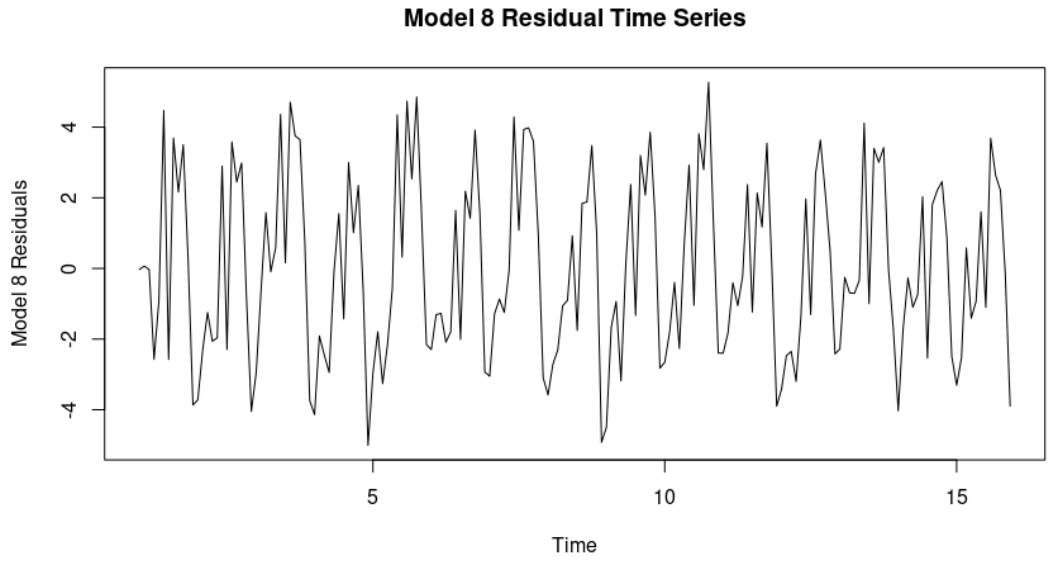
and choose the best model there to get our residual plots. Going up to 10 different models, we get the following list of AICs

```
[1] 911.4144 911.7727 881.4313 877.8733 867.7051 869.0018 862.2994 861.9199 863.7720 865.7667
```

If we continue to increase the number of degrees in our polynomial regression, the AIC will likely continue to get smaller, as seen in the following plots of the AICs



However, because these decreases are getting exponentially smaller, we will cap it off at only 10 models and choose the minimized AIC from this selection of 10. We can see that the smallest AIC comes from the 8^{th} model, and thus we will be using this model as our regression to capture the deterministic components of the data. This series yields the following time series plot.



The regression results of model 8 can be seen below and we can see the for the most part every estimate is statistically significant to at least the 10% significance level, except for the t^8 variable.

```
Call:
lm(formula = y ~ t + I(t^2) + I(t^3) + I(t^4) + I(t^5) + I(t^6) +
    I(t^7) + I(t^8), data = essay_data_ts)

Residuals:
   Min     1Q Median     3Q    Max
-5.009 -2.104 -0.301  2.141  5.270

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.528e+01  1.937e+00  49.178   <2e-16 ***
t            8.179e-01  4.818e-01   1.698   0.0914 .
I(t^2)      -7.902e-02  3.893e-02  -2.030   0.0439 *
I(t^3)       3.181e-03  1.439e-03   2.211   0.0284 *
I(t^4)      -6.192e-05  2.831e-05  -2.187   0.0301 *
I(t^5)       6.482e-07  3.158e-07   2.052   0.0417 *
I(t^6)      -3.753e-09  2.003e-09  -1.874   0.0626 .
I(t^7)       1.134e-11  6.722e-12   1.687   0.0933 .
I(t^8)      -1.398e-14  9.267e-15  -1.508   0.1333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 171 degrees of freedom
Multiple R-squared:  0.3808,    Adjusted R-squared:  0.3519
F-statistic: 13.15 on 8 and 171 DF,  p-value: 1.033e-14
```

As we can see the time series data looks much better, so it seems we have corrected any potential deterministic trends. We can confirm this by looking at the ADF test of the new detrended residual series to see if we need to do anything else further, such as first differencing, the output is shown below.

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -8.48    0.01
[2,]   1 -6.73    0.01
[3,]   2 -7.44    0.01
[4,]   3 -7.79    0.01
[5,]   4 -8.02    0.01

Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -8.45    0.01
[2,]   1 -6.71    0.01
[3,]   2 -7.41    0.01
[4,]   3 -7.76    0.01
[5,]   4 -8.00    0.01

Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -8.42    0.01
[2,]   1 -6.69    0.01
[3,]   2 -7.39    0.01
[4,]   3 -7.74    0.01
[5,]   4 -7.98    0.01
```

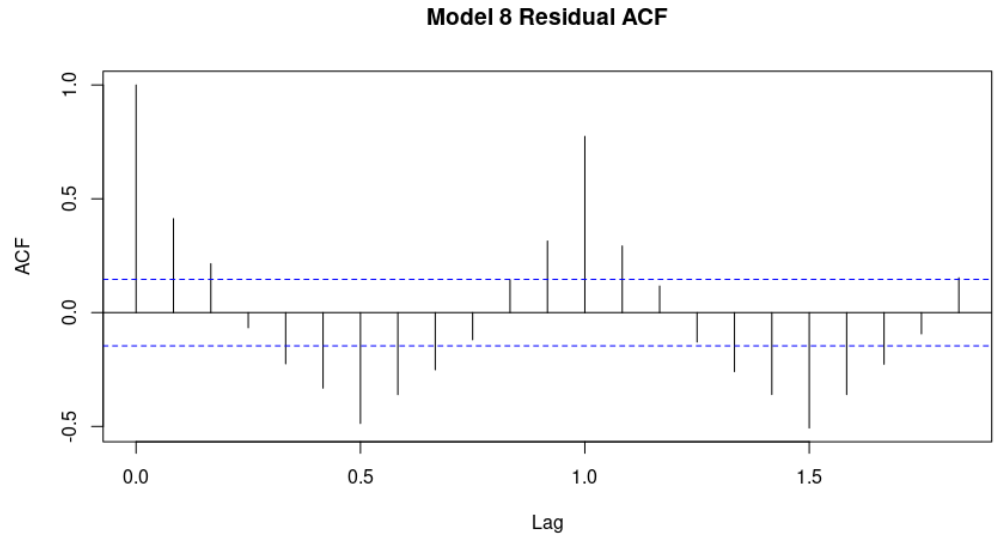
We can see that all p-values are less than 0.05 significance and hence we would reject the null hypothesis of the test and this means the detrended series is now stationary, as the alternative is the series is stationary. Thus, we do not need to do any further detrending, so as to not over difference our series. Now, we want to see if there is any autocorrelation in our data, which we will see using a Ljung-Box test.

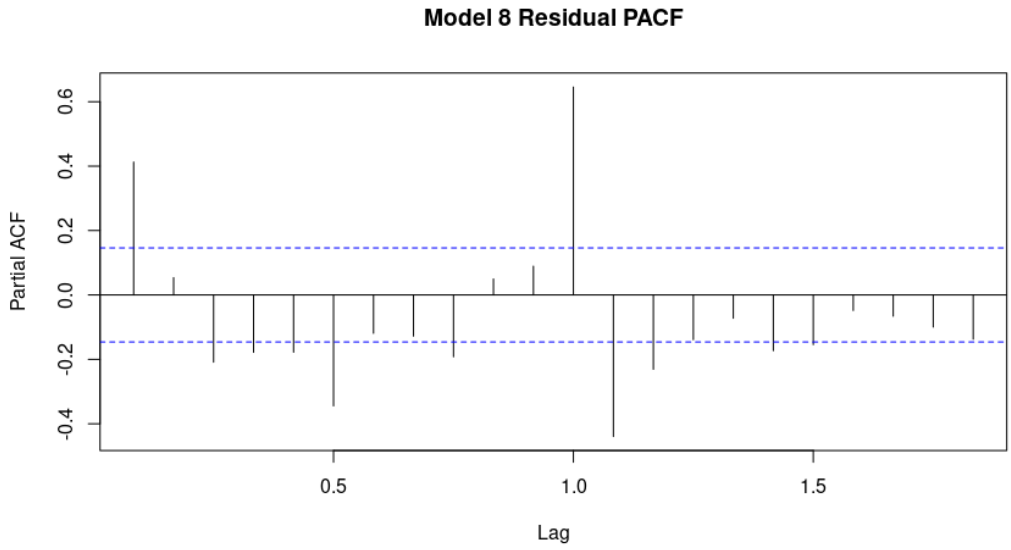
```
        Box-Ljung test

data:  model8_res
X-squared = 31.165, df = 1, p-value = 2.37e-08


        Box-Ljung test

data:  model8_res
X-squared = 39.659, df = 2, p-value = 2.444e-09


        Box-Ljung test

data:  model8_res
X-squared = 40.48, df = 3, p-value = 8.431e-09
```

Since the p-values of the test are quite small, this suggests we will reject the null hypothesis of this test. As a result, we would conclude that there is autocorrelation in the series, as the test suggests a rejection of the null that there is no autocorrelation. Thus, this tells us that there is a reason for us to model the residual series we have created. Thus, now that we have corrected for non-stationarity and confirmed that there is autocorrelation in our residual series, we can start our analysis of the data and create a model.

Determining Our Model

Now to determine what kind of model we want; we need to look at the ACF and PACF of the residual series we have.





Since it may be difficult to determine the correct model to use based on these plots, as it can be subjective, we will use the *auto.arima* function in the *forecast* package to help with determining the correct model. Moreover, there looks to be some degree of seasonality in our model, and the *auto.arima* function will help with this as well in choosing our model. In the ACF, the largest lag with statistical

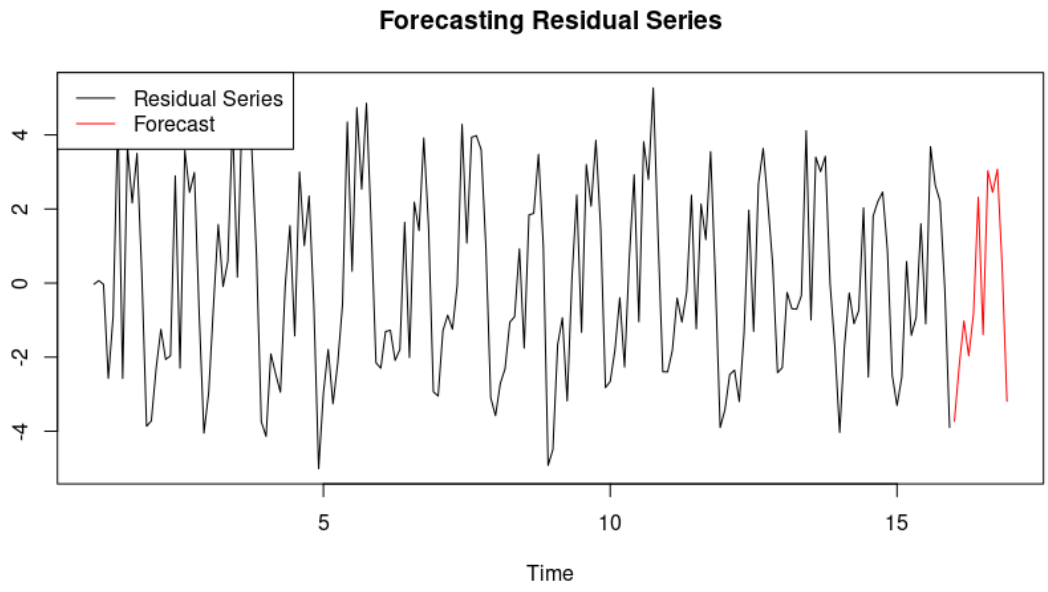
significance is the 21st lag and looking at the PACF, the largest statistically significant lag is lag 17. These values suggest that we will have a *max.q* = 21 and *max.p* = 17, the output of this function is shown below

```
Series: model8_res
ARIMA(1,0,1)(1,1,2)[12]

Coefficients:
         ar1      ma1     sar1     sma1     sma2
      0.7918  -0.3862  -0.0279  -0.7463  -0.0883
s.e.  0.0795   0.1096   0.7590   0.7637   0.6148

sigma^2 estimated as 0.8367:  log likelihood=-227.98
AIC=467.97   AICc=468.49   BIC=486.71

Training set error measures:
                      ME      RMSE       MAE      MPE     MAPE      MASE         ACF1
Training set -0.04815465 0.8704608 0.6754971 11.09856 105.3126 0.5594615 -0.04346215
```

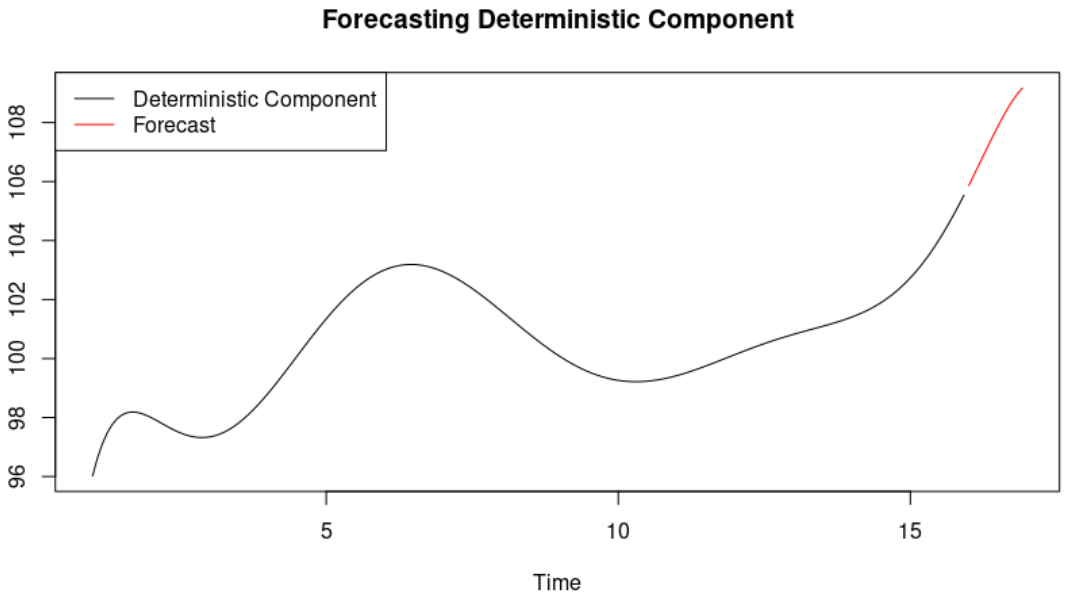
From this function, the best model to use is the ARIMA (1,0,1) model, i.e. a ARMA (1,1) model, with a seasonal ARIMA (1,1,2) term, and hence we will be using this model going forward.

Making Our Forecasts

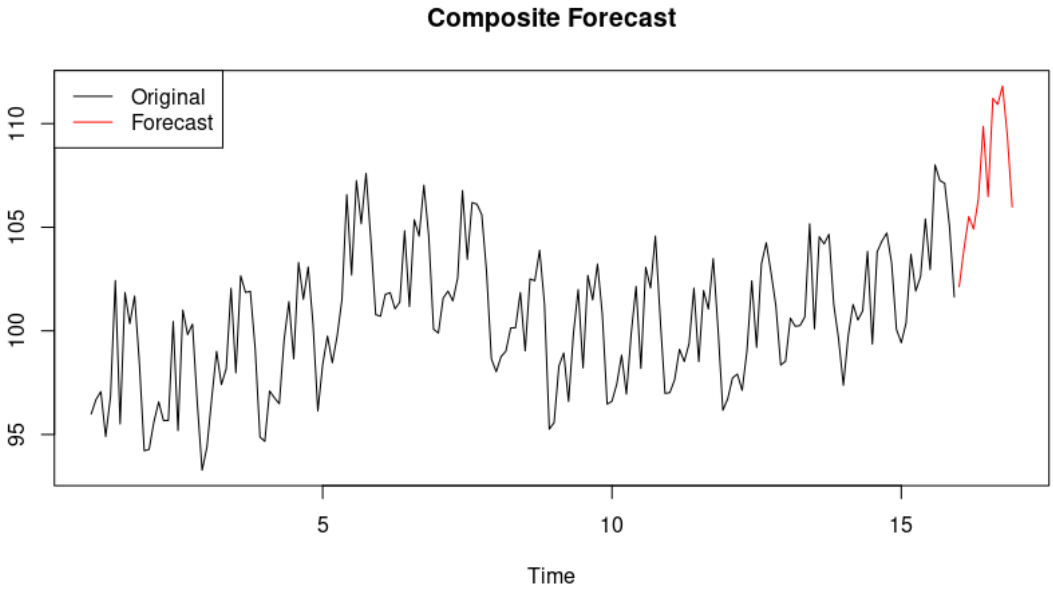
Now that we have the ARIMA model we want to use, we will now be able to forecast using it. The forecast for the residual series is shown below



The deterministic component forecast is shown below



Combining these two forecasts, the residual and deterministic component forecasts, yields the final composite forecast that we want for 12 periods. This plot is shown below.



Thus, we have now created the composite forecast we set out to complete using this set of data and the exact point forecasted values are as follows

```
     Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
102.1320 103.9419 105.5139 104.9149 106.3839 109.8684 106.4829 111.2170 110.9263 111.8160 109.4096 105.9784
```

We can also identify an interval of potential forecasted values for the 12 periods which can be found in the table below

Month	**Point Forecast**	**Lower Interval Forecast**	**Upper Interval Forecast**
January	102.1320002	95.77299277	108.4910077
February	103.9418911	96.98234699	110.9014353
March	105.5138645	97.75002723	113.2777018
April	104.9148845	96.11731433	113.7124548
May	106.3839353	96.3015154	116.4663553
June	109.868431	98.23063977	121.5062222
July	106.4829028	93.00019861	119.9656069
August	111.2169622	95.57971101	126.8542134
September	110.926319	92.80281239	129.0498257
October	111.8160275	90.85001775	132.7820372
November	109.4096358	85.21765031	133.6016213
December	105.9783829	78.14695124	133.8098145

Conclusion

In conclusion, using the knowledge that we have learned throughout the course, we were able to identify any trends within the data and rectify any violation of assumptions needed for the ARIMA model, such as stationarity, identify the best model to forecast with, a ARIMA (1,0,1) with seasonal ARIMA (1,1,2) terms, and forecasted for 12 periods using the chosen ARIMA model. This allowed for us to find the interval and point forecast values of the data.