**Steven Taylor**

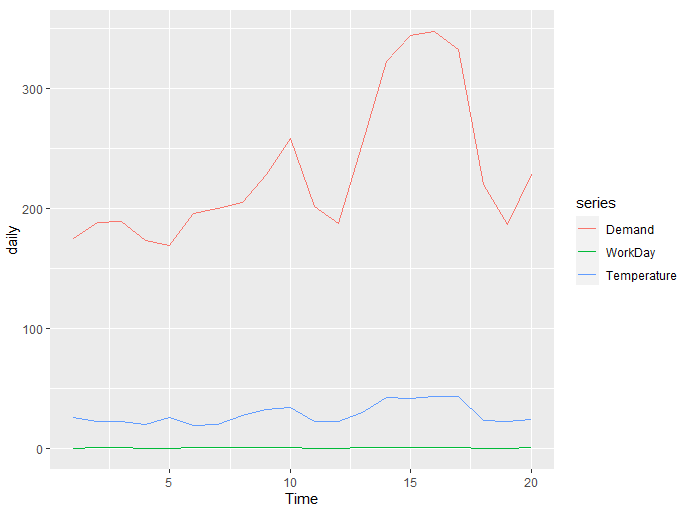
**Predictive Analytics & Forecastign**

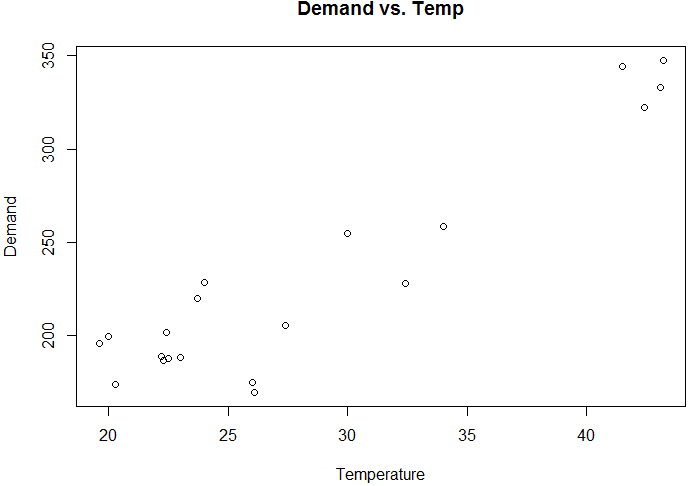
**Homework Assignment 1:**

**Chapter 5**

1. Daily electric demand
   1. Plot the data and find the regression model for Demand with temperature as an explanatory variable. Why is there a positive relationship?

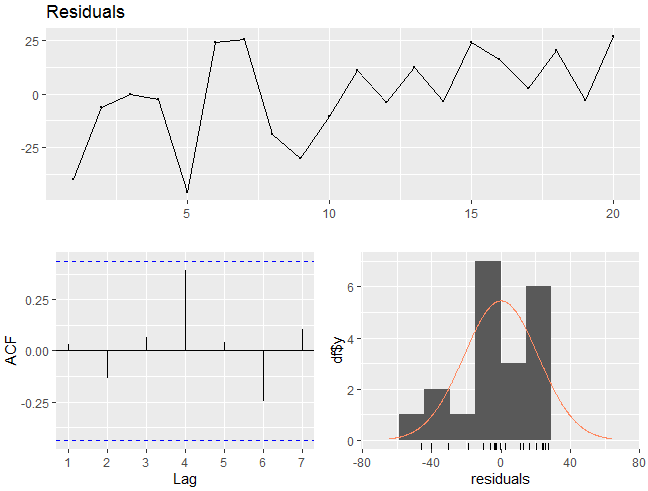
Demand and temperature are positively correlated. As temperature increases we see large increases in demand. This is evident when time (on the x axis) equals ~7.5 and when time equals 12.





* 1. Produce a residual plot. Is the model adequate? Are there any outliers or influential observations?

There is no clear correlation between the residuals.



* 1. Use the model to forecast the electricity demand that you would expect for the next day if the maximum temperature was 15∘ and compare it with the forecast if the with maximum temperature was 35∘. Do you believe these forecasts?

Temp = 15

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

21 140.5701 108.681 172.4591 90.21166 190.9285

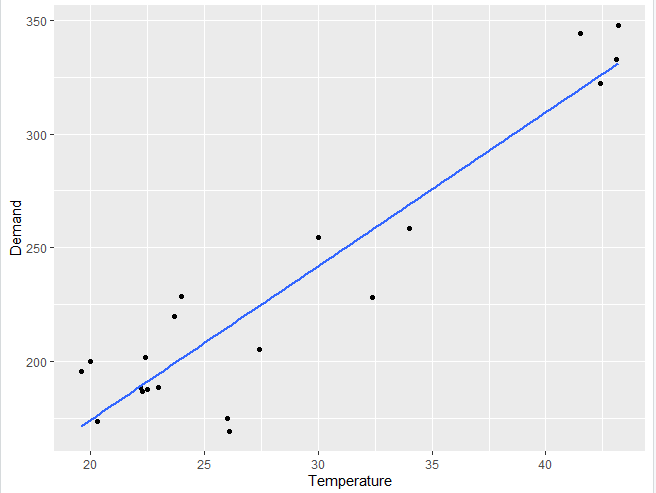
Temp = 35

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

21 275.7146 245.2278 306.2014 227.5706 323.8586

These forecasts seem reasonable as demand increases coincide with temperature increases.

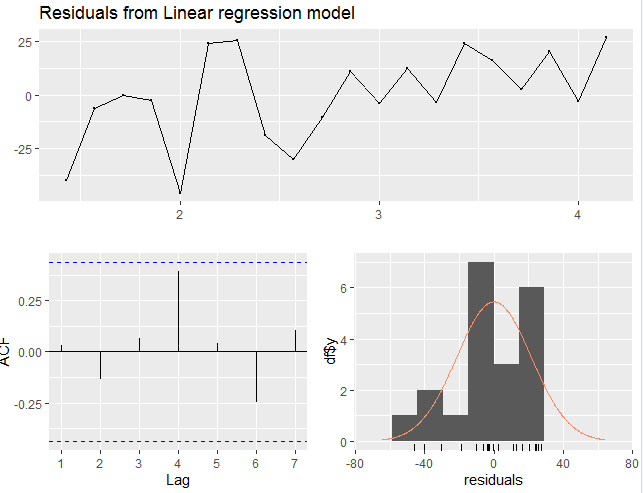
* 1. Give prediction intervals for your forecasts.



Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

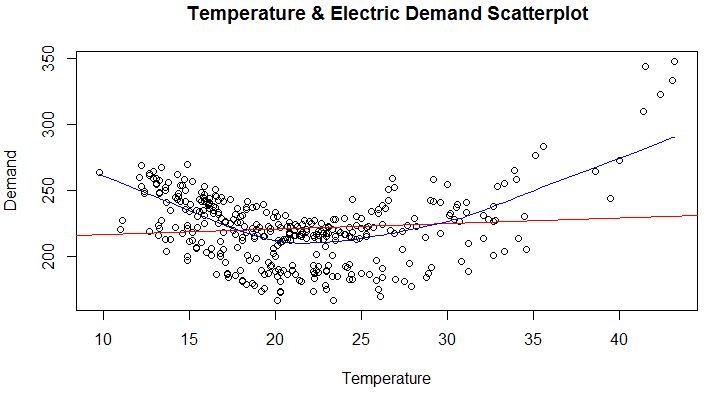
4.285714 140.5701 108.6810 172.4591 90.21166 190.9285

4.428571 275.7146 245.2278 306.2014 227.57056 323.8586



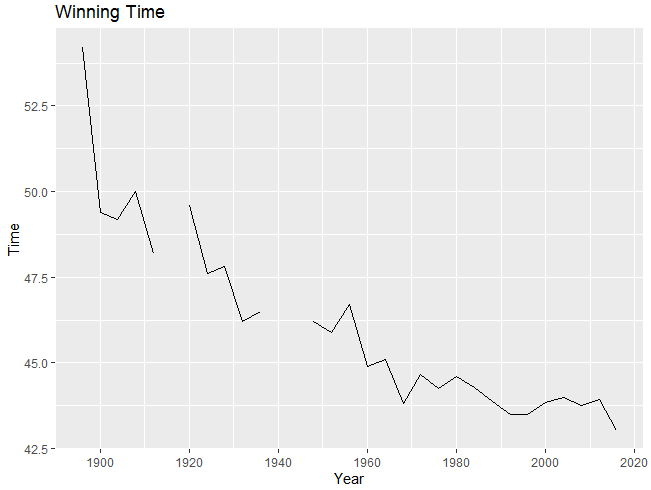
* 1. Plot Demand vs Temperature for all of the available data in elecdaily. What does this say about your model?

Electric demand and temperature have a non-linear relationship as energy demand fluctuates based on temperature. Energy demand is high when the temperature is low (12-15 degrees Celsius) and gradually decreases as temperature increases. The plot flattens when temperature reaches a comfortable 22 degrees Celsius. As temperature increases beyond the comfort level we’ll begin to see electric demand increase.



1. Mens400
   1. Plot the winning time against the year. Describe the main features of the plot.

The winning time for the Men’s 400 meter has consistently decreased since recordkeeping began in the late 1800s. By the mid-1960s the wining time had decreased to less than 45 seconds.



* 1. Fit a regression line to the data. Obviously the winning times have been decreasing, but at what average rate per year?

Winning time decrease at about 0.07 seconds per year.

Call:

tslm(formula = mens400 ~ mtime, data = mens400)

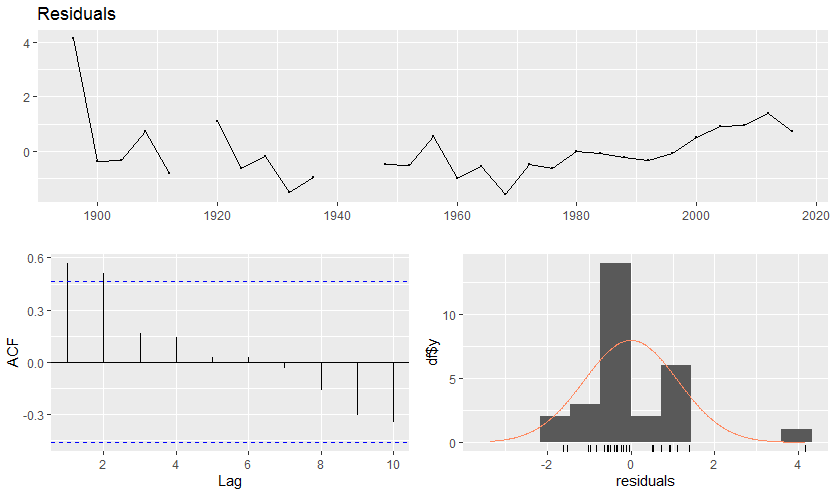
Coefficients:

(Intercept) mtime

172.48148 -0.06457

* 1. Plot the residuals against the year. What does this indicate about the suitability of the fitted line?

These residuals indicate that our model underestimated improvements in the beginning of our time series and overestimates improvement in the present time. This model expects the times to keep decreasing but our athletes seem to have hit a plateau around the 42 second mark.



* 1. Predict the winning time for the men’s 400 meters final in the 2020 Olympics. Give a prediction interval for your forecasts. What assumptions have you made in these calculations?

I forecast the 2020 Men’s 400-meter winning time to be 42.04 seconds. I’m assuming the improvement of 0.07 seconds per year will continue to hold true.

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

2020 42.04231 40.44975 43.63487 39.55286 44.53176

1. Type easter(ausbeer) and interpret what you see.

Qtr1 Qtr2 Qtr3 Qtr4

1956 0.67 0.33 0.00 0.00

1957 0.00 1.00 0.00 0.00

1958 0.00 1.00 0.00 0.00

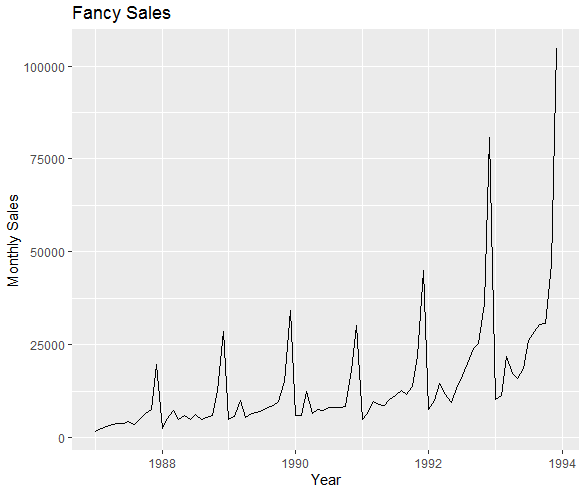
1959 1.00 0.00 0.00 0.00

1960 0.00 1.00 0.00 0.00

1961 0.33 0.67 0.00 0.00 …..

1. Unknown
2. ‘Fancy’ data
   1. Produce a time plot of the data and describe the patterns in the graph. Identify any unusual or unexpected fluctuations in the time series.

The ‘fancy’ dataset shows seasonal spikes in sales volume. There are steady seasonal highs of ~ 30,000 in the early stages of our time series. Then around 1992 our sales experience exponential growth surges. Fluctuation in the data are due to tourist events focused around holiday travel and surfing festivals. Our current monthly high is over 100,000.



* 1. Explain why it is necessary to take logarithms of these data before fitting a model.

Logs allow us more flexibility when dealing with data that has exponential growth periods, similar to our seasonal data above.

* 1. Use R to fit a regression model to the logarithms of these sales data with a linear trend, seasonal dummies and a “surfing festival” dummy variable.

Call:

tslm(formula = BoxCox(fancy, 0) ~ season + trend)

Coefficients:

(Intercept) season2 season3 season4 season5 season6

7.60586 0.25104 0.69521 0.38293 0.40799 0.44696

season7 season8 season9 season10 season11 season12

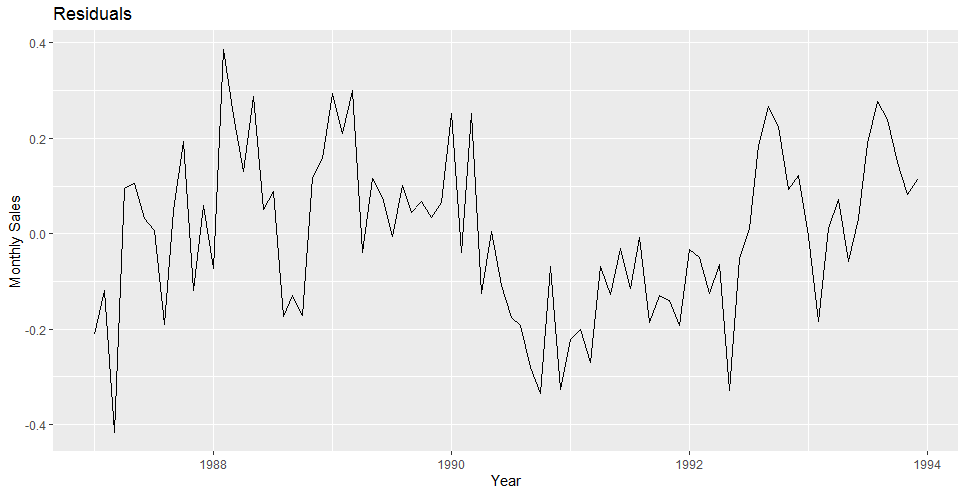
0.60822 0.58535 0.66634 0.74403 1.20302 1.95814

trend

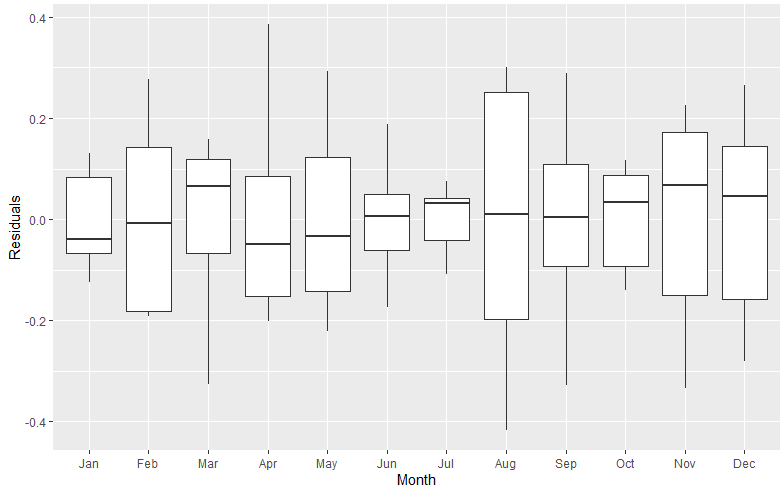
0.02239

* 1. Plot the residuals against time and against the fitted values. Do these plots reveal any problems with the model?

One potential flaw with the model is that our residuals appear to have some correlation.



* 1. Do boxplots of the residuals for each month. Does this reveal any problems with the model?



* 1. What do the values of the coefficients tell you about each variable?

Season 11 and Season 12 appear to be the seasons effected most by changes to our predictors.

(Intercept) season2 season3 season4 season5

7.60586042 0.25104367 0.69520658 0.38293405 0.40799437

season6 season7 season8 season9 season10

0.44696253 0.60821562 0.58535238 0.66634464 0.74403359

season11 season12 trend

1.20301639 1.95813657 0.02239298

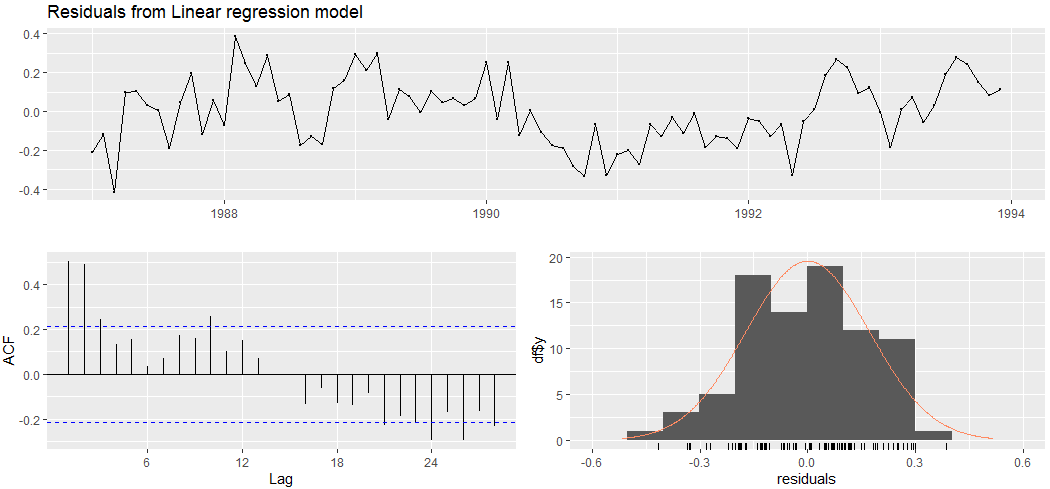
* 1. What does the Breusch-Godfrey test tell you about your model?

A p value less than .05 indicates residuals that are correlated.

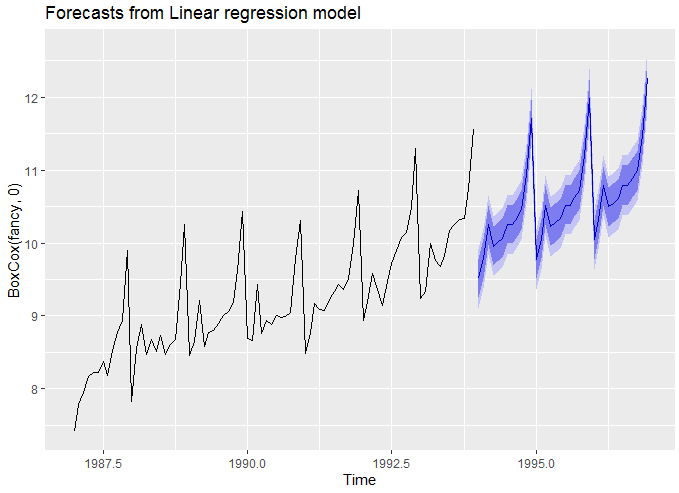
Breusch-Godfrey test for serial correlation of order up to 17

data: Residuals from Linear regression model

LM test = 37.152, df = 17, p-value = 0.003209



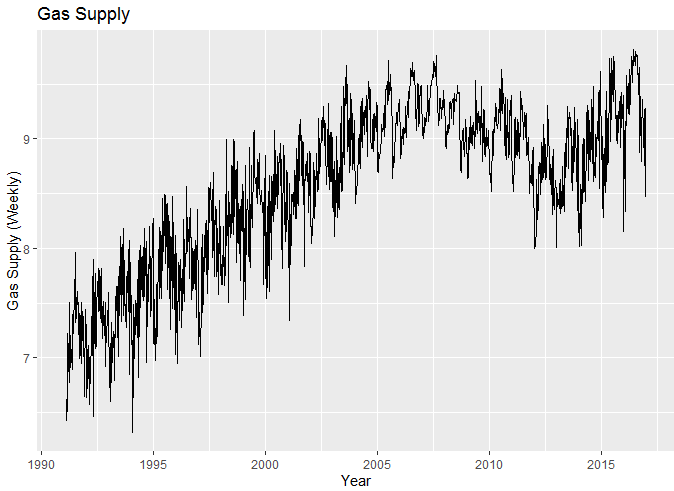
* 1. Regardless of your answers to the above questions, use your regression model to predict the monthly sales for 1994, 1995, and 1996. Produce prediction intervals for each of your forecasts.

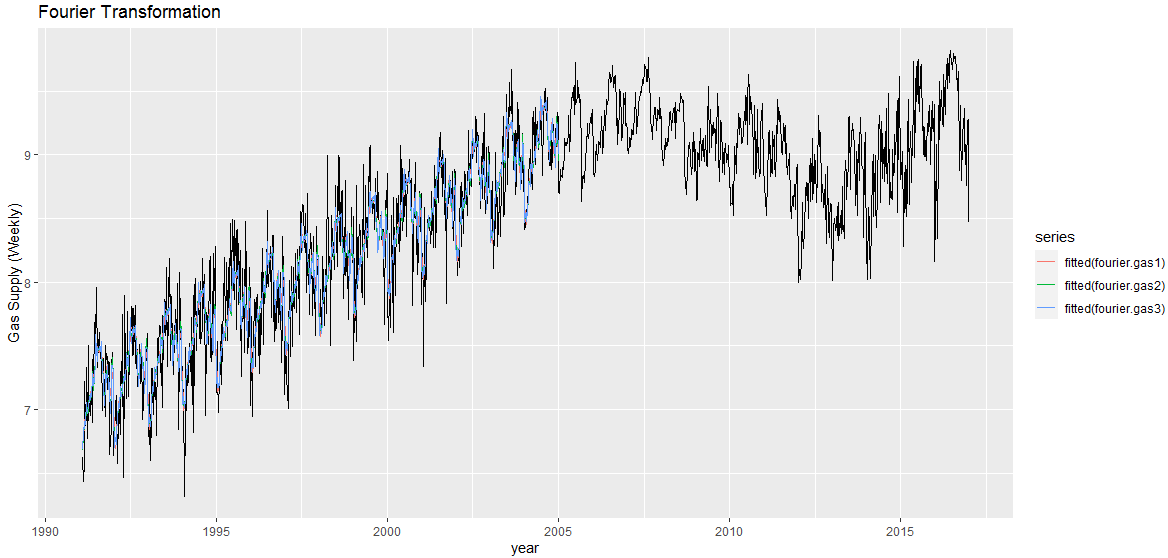


How could you improve these predictions by modifying the model?

One way we could improve the model would be to find a way to reduce correlation between predictors.

1. Gasoline
   1. Fit a harmonic regression with trend to the data. Experiment with changing the number Fourier terms. Plot the observed gasoline and fitted values and comment on what you see.





* 1. Select the appropriate number of Fourier terms to include by minimising the AICc or CV value

Model 2, where K=12 looks to be the best model. This model has the largest Adjusted R squared and a large AIC compared to the other models.

> CV(fourier.gas1)

CV AIC AICc BIC

0.1763075 -2344.7333753 -2344.2746047 -2256.1745910

AdjR2

0.6745614

> CV(fourier.gas2)

CV AIC AICc BIC

0.1770436 -2339.4200165 -2338.2780225 -2198.7678296

AdjR2

0.6756498

> CV(fourier.gas3)

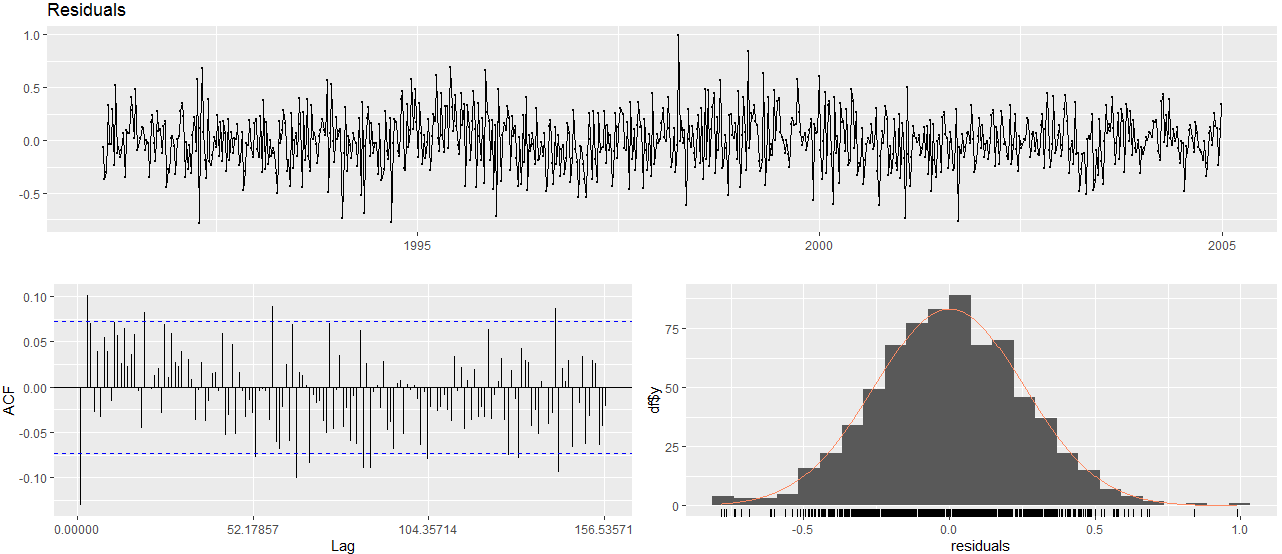
CV AIC AICc BIC

0.1803512 -2315.2407553 -2312.3477890 -2091.2391243

AdjR2

0.6735819

* 1. Check the residuals of the final model using the checkresiduals() function. Even though the residuals fail the correlation tests, the results are probably not severe enough to make much difference to the forecasts and prediction intervals. (Note that the correlations are relatively small, even though they are significant.)



* 1. To forecast using harmonic regression, you will need to generate the future values of the Fourier terms.

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

2017.011 9.116974 8.577060 9.656888 8.290904 9.943044

2017.030 9.092743 8.552807 9.632680 8.266639 9.918848

2017.049 9.124136 8.584203 9.664068 8.298038 9.950234

2017.069 9.123557 8.583635 9.663479 8.297475 9.949638

2017.088 9.137493 8.597636 9.677350 8.311511 9.963476

2017.107 9.224236 8.684420 9.764051 8.398316 10.050155

2017.126 9.328017 8.788202 9.867833 8.502098 10.153937

2017.145 9.365027 8.825220 9.904834 8.539121 10.190933

2017.164 9.359203 8.819401 9.899006 8.533304 10.185102

2017.184 9.390837 8.851031 9.930644 8.564933 10.216742

2017.203 9.455963 8.916155 9.995770 8.630056 10.281870

2017.222 9.482896 8.943091 10.022700 8.656993 10.308798

2017.241 9.466060 8.926257 10.005862 8.640161 10.291959

2017.260 9.472998 8.933193 10.012802 8.647096 10.298900

2017.279 9.516625 8.976818 10.056431 8.690720 10.342529

2017.298 9.532584 8.992779 10.072388 8.706681 10.358486

2017.318 9.511049 8.971247 10.050852 8.685150 10.336949

2017.337 9.539299 8.999495 10.079102 8.713397 10.365200

2017.356 9.652957 9.113151 10.192762 8.827053 10.478860

2017.375 9.744536 9.204732 10.284341 8.918634 10.570439

2017.394 9.713586 9.173783 10.253388 8.887686 10.539485

2017.413 9.635803 9.096000 10.175607 8.809903 10.461704

2017.433 9.665979 9.126174 10.205784 8.840076 10.491883

2017.452 9.808203 9.268398 10.348008 8.982300 10.634106

2017.471 9.908857 9.369054 10.448660 9.082957 10.734757

2017.490 9.881279 9.341476 10.421083 9.055379 10.707180

2017.509 9.814508 9.274703 10.354313 8.988605 10.640411

2017.528 9.818999 9.279195 10.358804 8.993097 10.644902

2017.548 9.872664 9.332861 10.412467 9.046764 10.698564

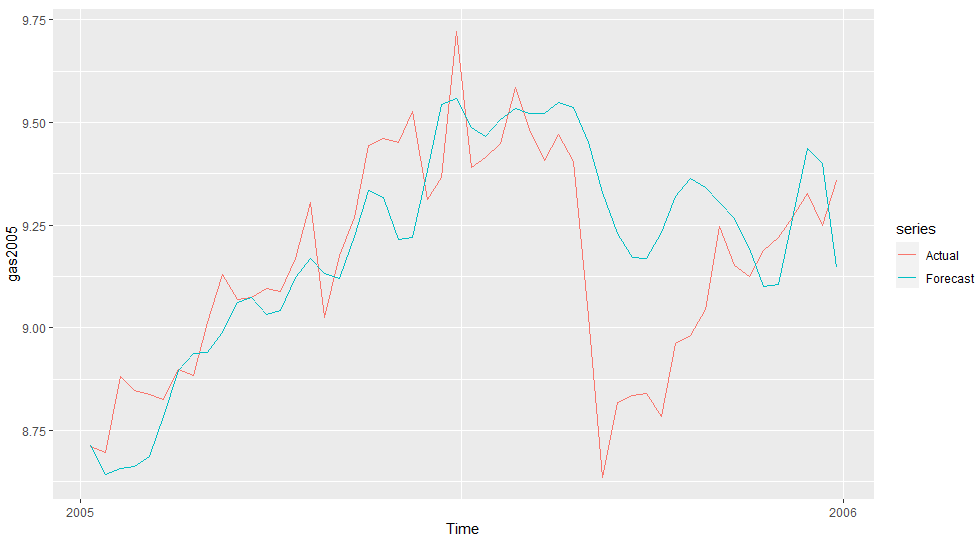
2017.567 9.892463 9.352660 10.432266 9.066563 10.718363

2017.586 9.876971 9.337166 10.416776 9.051069 10.702873

2017.605 9.882064 9.342259 10.421869 9.056161 10.707967

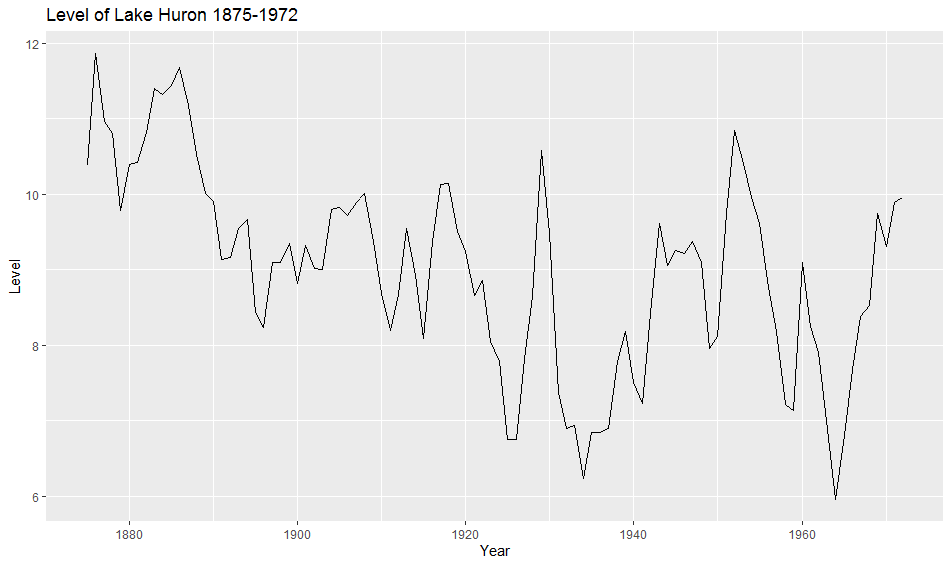
* 1. Plot the forecasts along with the actual data for 2005. What do you find?

Our model performs well through the first half of 2005. In the third quarter we experience a sharp decline in units which the model failed to predict.



1. Huron
   1. Plot the data and comment on its features.

Lake Huron water levels fluctuate between 6 and 12. Beginning in 1920 the fluctuations have become far more volatile from year to year.



* 1. Fit a linear regression and compare this to a piecewise linear trend model with a knot at 1915.

In comparing both models the first item I notice is that both p-values are too large. Looking beyond this, the piecewise model has a lower F-statistic and light adjusted R-squared.

**Linear**

Call:

tslm(formula = huron ~ trend)

Residuals:

Min 1Q Median 3Q Max

-2.50997 -0.72726 0.00083 0.74402 2.53565

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.202037 0.230111 44.335 < 2e-16 \*\*\*

trend -0.024201 0.004036 -5.996 3.55e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.13 on 96 degrees of freedom

Multiple R-squared: 0.2725, Adjusted R-squared: 0.2649

F-statistic: 35.95 on 1 and 96 DF, p-value: 3.545e-08

**Piecewise**

Call:

tslm(formula = huron ~ Time + t2)

Residuals:

Min 1Q Median 3Q Max

-2.49626 -0.66240 -0.07139 0.85163 2.39222

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 132.90870 19.97687 6.653 1.82e-09 \*\*\*

Time -0.06498 0.01051 -6.181 1.58e-08 \*\*\*

t2 0.06486 0.01563 4.150 7.26e-05 \*\*\*

---

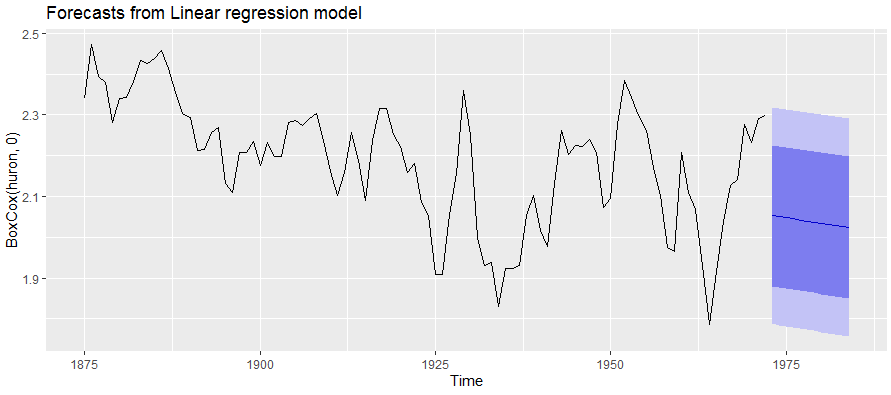
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 95 degrees of freedom

Multiple R-squared: 0.3841, Adjusted R-squared: 0.3711

F-statistic: 29.62 on 2 and 95 DF, p-value: 1.004e-10

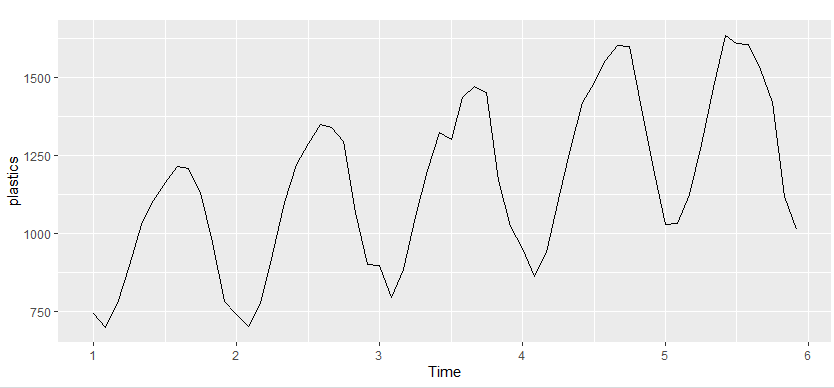
* 1. Generate forecasts from these two models for the period up to 1980 and comment on these.



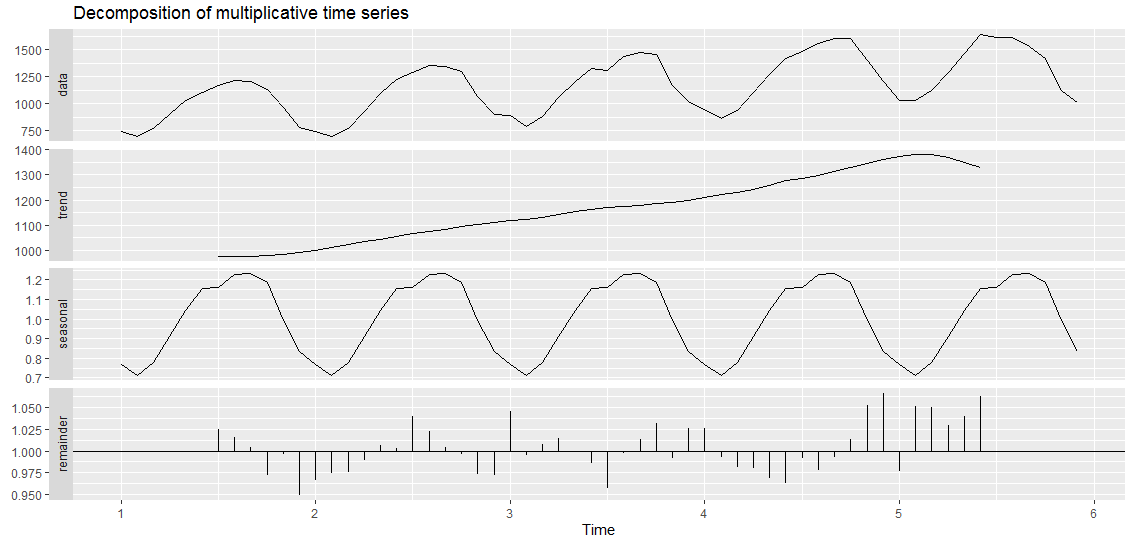
**Chapter 6**

1. N/A
2. Plastics
   1. Plot the time series of sales of product A. Can you identify seasonal fluctuations and/or a trend-cycle?

This plot shows that our data is highly seasonal.



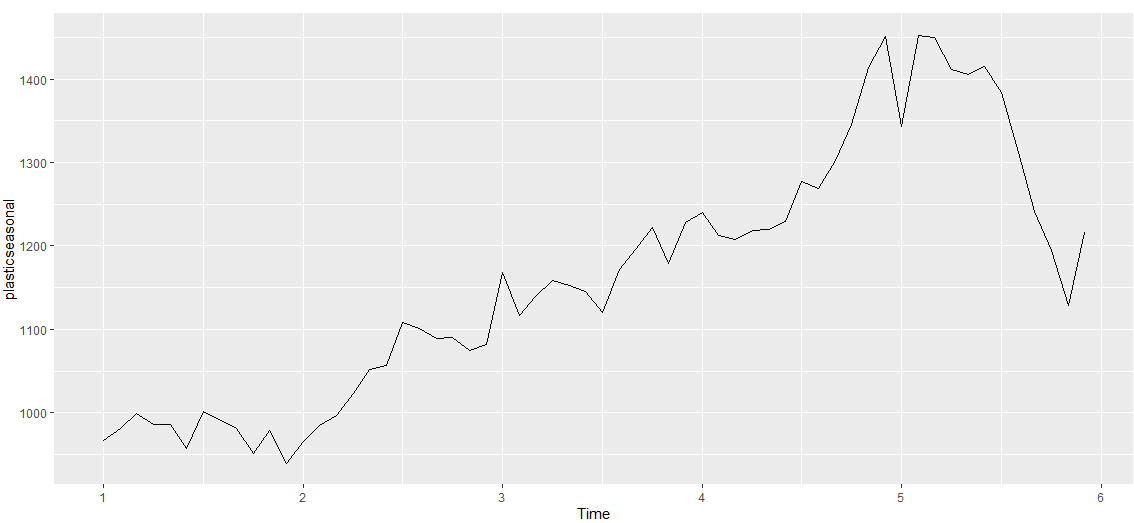
* 1. Use a classical multiplicative decomposition to calculate the trend-cycle and seasonal indices.



* 1. Do the results support the graphical interpretation from part A?

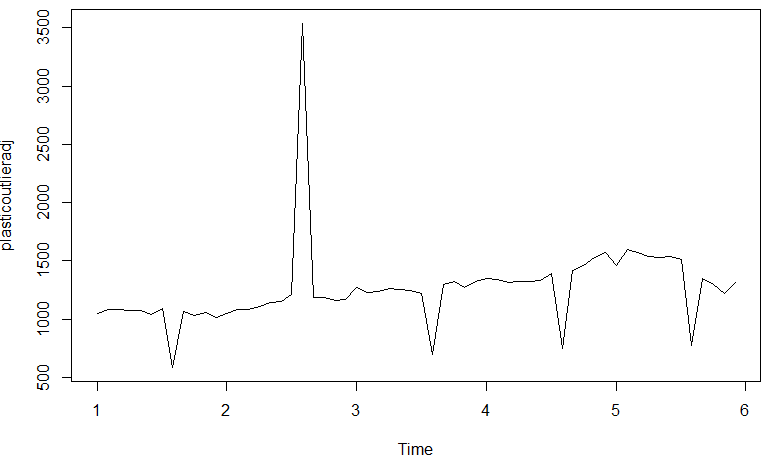
Yes, this graph does support the seasonality claim we made in part A.

* 1. Compute and plot the seasonally adjusted data.



* 1. Change one observation to be an outlier (e.g., add 500 to one observation), and recompute the seasonally adjusted data. What is the effect of the outlier?

Changing one variable to an extreme outlier completely disrupted this graph. I’ve tried it with 6,000 and 400. I’ve noticed the larger the outlier, the more un-usable this graph becomes.



* 1. Does it make any difference if the outlier is near the end rather than in the middle of the time series?

This should be irrelevant.

1. (from Exercise 3 in Section [2.10](https://otexts.com/fpp2/graphics-exercises.html#graphics-exercises)). Decompose the series using X11. Does it reveal any outliers, or unusual features that you had not noticed previously?

N/A

1. Decomposition
   1. Write about 3–5 sentences describing the results of the decomposition. Pay particular attention to the scales of the graphs in making your interpretation.

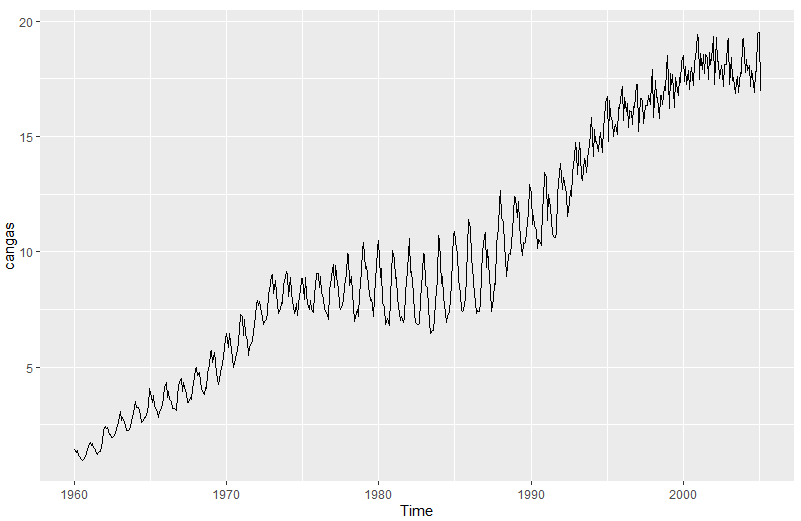
Figure 6.16 shows clear seasonality movement as the volatility is routine and predictable. There appears to be unusual movement between years 1991 and 1994.

* 1. Is the recession of 1991/1992 visible in the estimated components?

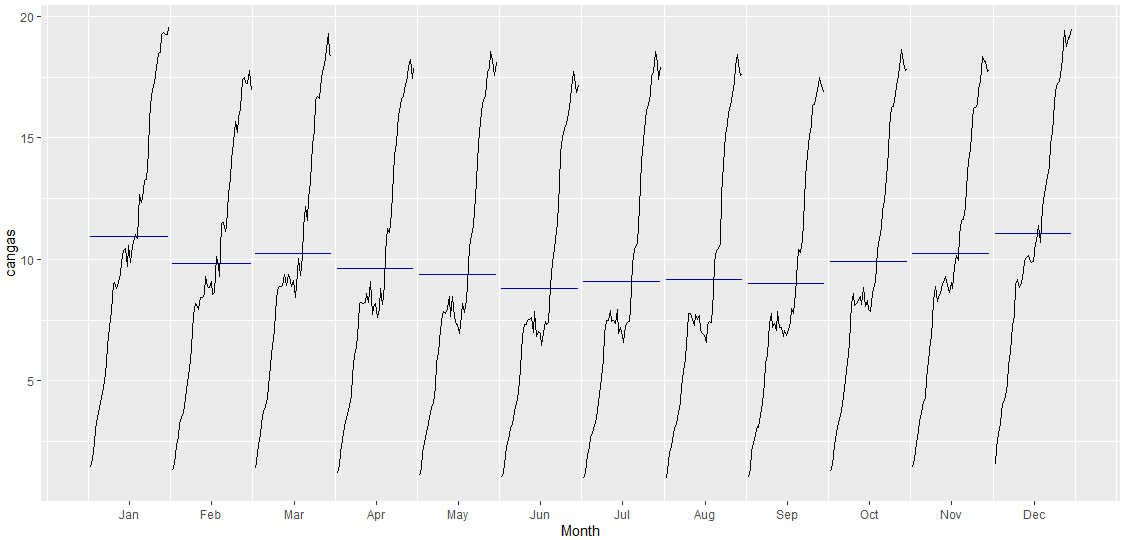
The recession is visible. The top ‘data’ graph has a clear divot at the 1991 mark. There is also a pronounced move on the ‘remainder’ graph during the 1991 recession.

1. Cangas
   1. Plot the data using autoplot(), ggsubseriesplot() and ggseasonplot() to look at the effect of the changing seasonality over time. What do you think is causing it to change so much?

Autoplot:

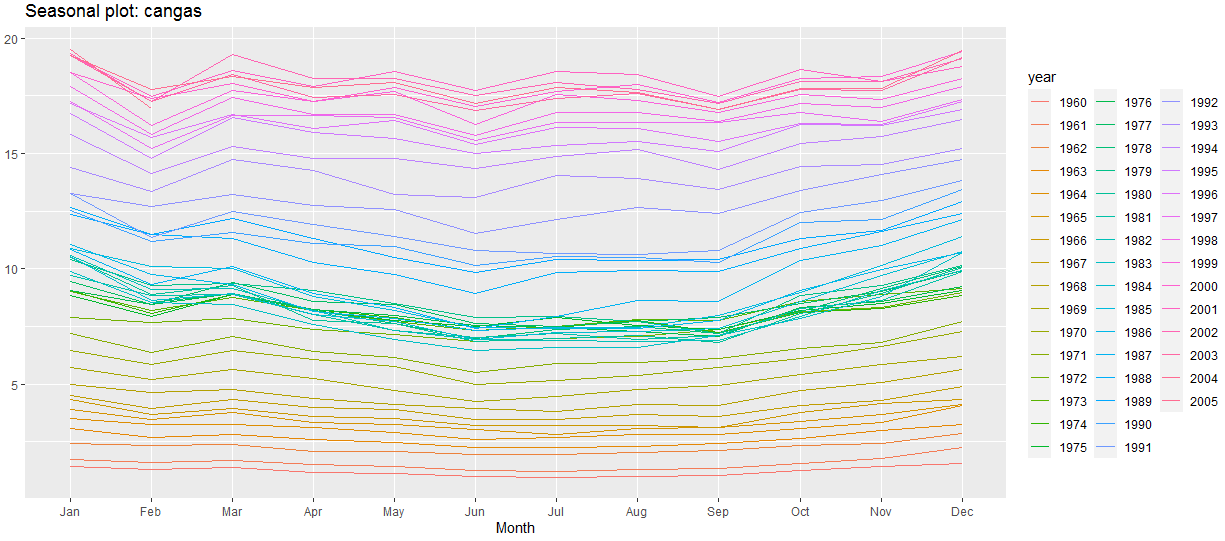


GG Sub Series Plot:

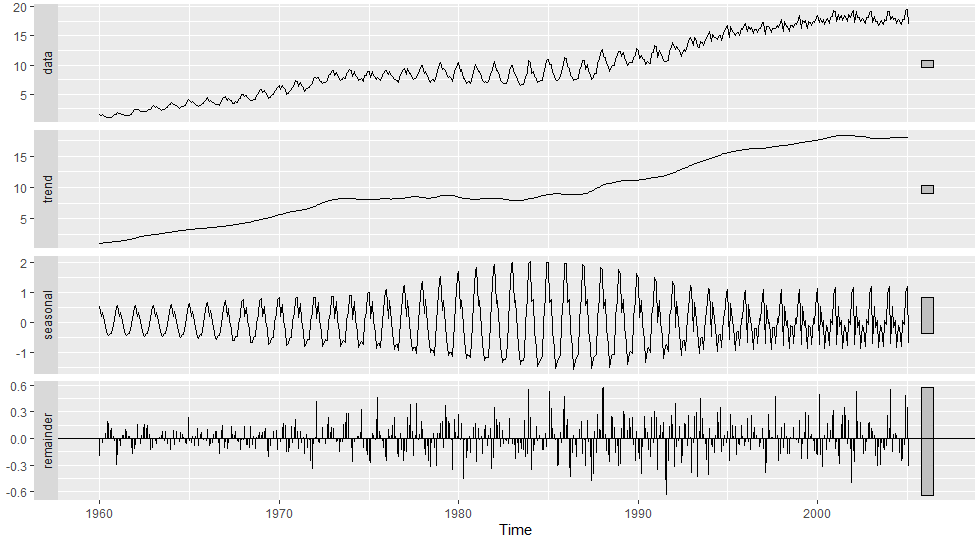


GG Season Plot:

Beginning in March there is a slight monthly decrease in production. This trend seems to bottom out in June before gradually working to its peak between December and January.



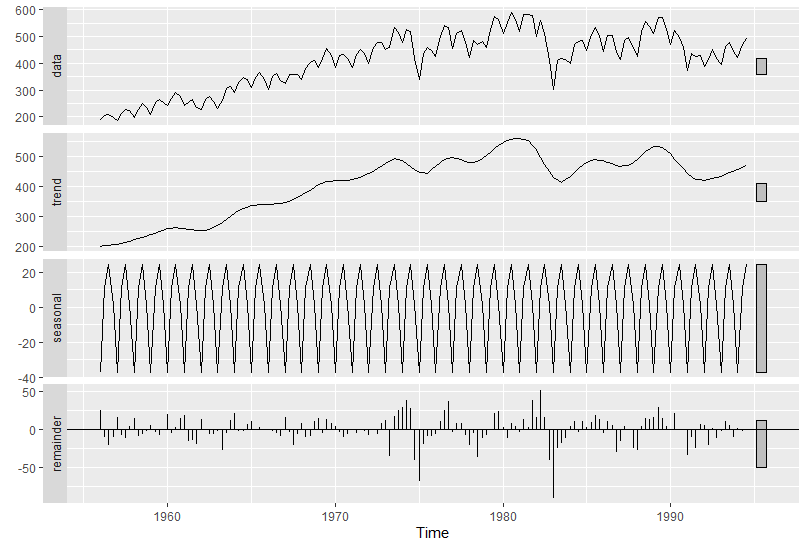
* 1. Do an STL decomposition of the data. You will need to choose s.window to allow for the changing shape of the seasonal component.



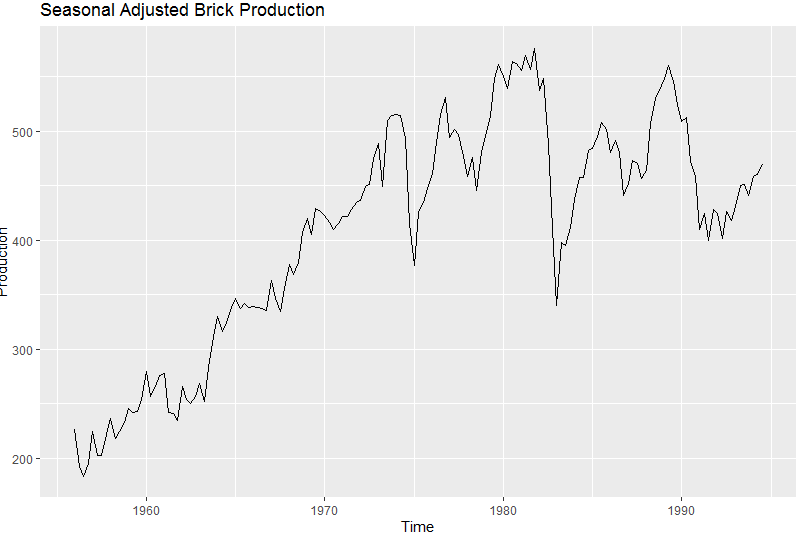
* 1. Compare the results with those obtained using SEATS and X11. How are they different?

Seasonality was at its worst around 1988. In reality, seasonality was high throughout the entire decade of the 80’s.

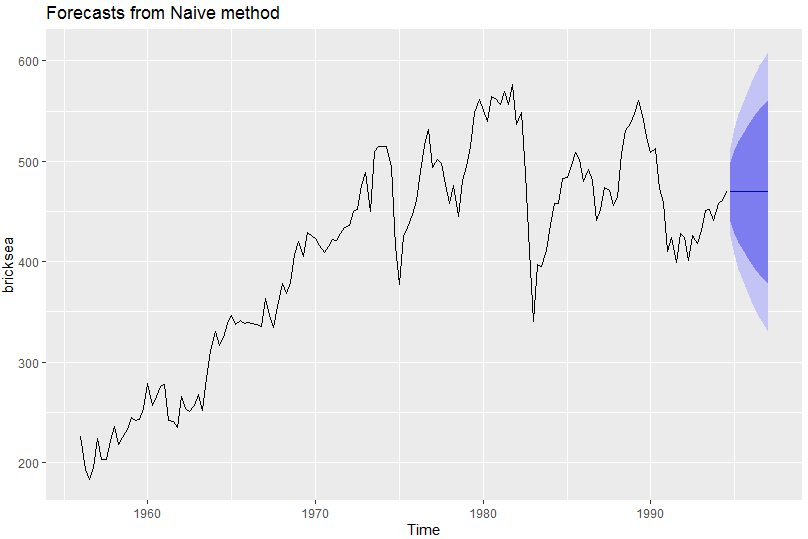
1. Bricksq
   1. Use an STL decomposition to calculate the trend-cycle and seasonal indices. (Experiment with having fixed or changing seasonality.)



* 1. Compute and plot the seasonally adjusted data.



* 1. Use a naïve method to produce forecasts of the seasonally adjusted data.



* 1. Use stlf() to reseasonalise the results, giving forecasts for the original data.

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

1994 Q4 467.0649 441.7635 492.3662 428.3698 505.7599

1995 Q1 426.0748 390.2933 461.8563 371.3517 480.7979

1995 Q2 484.0357 440.2125 527.8590 417.0139 551.0576

1995 Q3 494.0000 443.3973 544.6027 416.6098 571.3902

1995 Q4 467.0649 410.4893 523.6404 380.5400 553.5897

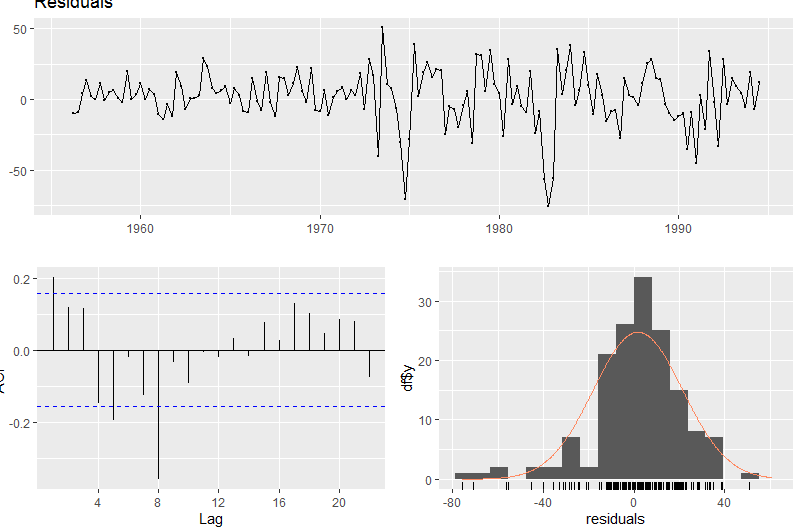
1996 Q1 426.0748 364.0994 488.0502 331.2916 520.8580

1996 Q2 484.0357 417.0946 550.9768 381.6582 586.4133

1996 Q3 494.0000 422.4370 565.5630 384.5538 603.4462

* 1. Do the residuals look uncorrelated?

This data does not appear to be correlated.



* 1. Repeat with a robust STL decomposition. Does it make much difference?

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

1994 Q4 471.7973 442.1625 501.4322 426.4748 517.1199

1995 Q1 471.7973 429.8874 513.7073 407.7015 535.8931

1995 Q2 471.7973 420.4683 523.1264 393.2963 550.2983

1995 Q3 471.7973 412.5277 531.0670 381.1522 562.4425

1995 Q4 471.7973 405.5318 538.0628 370.4530 573.1417

1996 Q1 471.7973 399.2071 544.3876 360.7802 582.8145

1996 Q2 471.7973 393.3909 550.2037 351.8851 591.7096

1996 Q3 471.7973 387.9774 555.6173 343.6058 599.9889

1996 Q4 471.7973 382.8928 560.7018 335.8296 607.7650

1997 Q1 471.7973 378.0838 565.5109 328.4748 615.1199

* 1. Compare forecasts from stlf() with those from snaive(), using a test set comprising the last 2 years of data. Which is better?