Module Assignment
Module 7
QMB-6304 Foundations of Business Statistics

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

Preprocessing

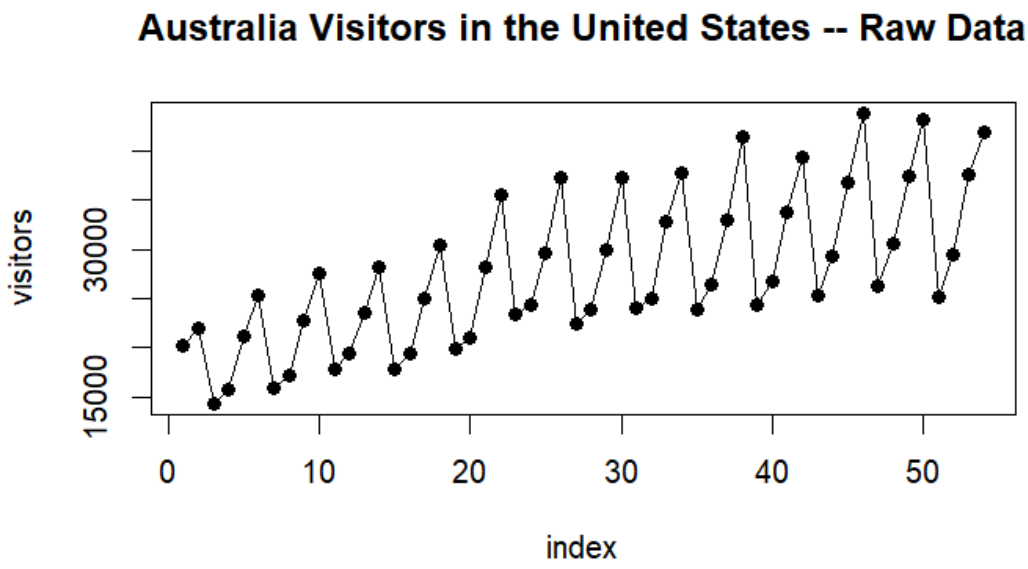
1. Load the file "6304 Module 7 Data.xlsx" into R. This data shows the number of visitors to the United States from the Commonwealth of Australia on a quarterly basis from years 1998 to 2012. The data shown is scaled in thousands of people.
2. Create a new "index" variable in the data frame which will be an identifying sequential numbering of rows from 1 to the number of rows in the data frame. This will be the only data set used for your analysis.

```
#Carolina Aldana Yabur
#U25124553

#Preprocessing
rm(list=ls())
setwd("C:/Users/calda/Desktop")
library(rio)
library(car)
australia=import("6304 Module 7 Assignment Data.xlsx")
colnames(australia)=tolower(make.names(colnames(australia)))
colnames(australia)[3]="visitors"
names(australia)
australia$index=seq(1:nrow(australia))
names(australia)
attach(australia)
```

Analysis

1. Show a plot of the data using the number of visitors as the "y" variable in the plot.

```
#Analysis
#Part 1
plot(index,visitors,type="o",pch=19,
     main="Australia Visitors in the United States -- Raw Data")
```

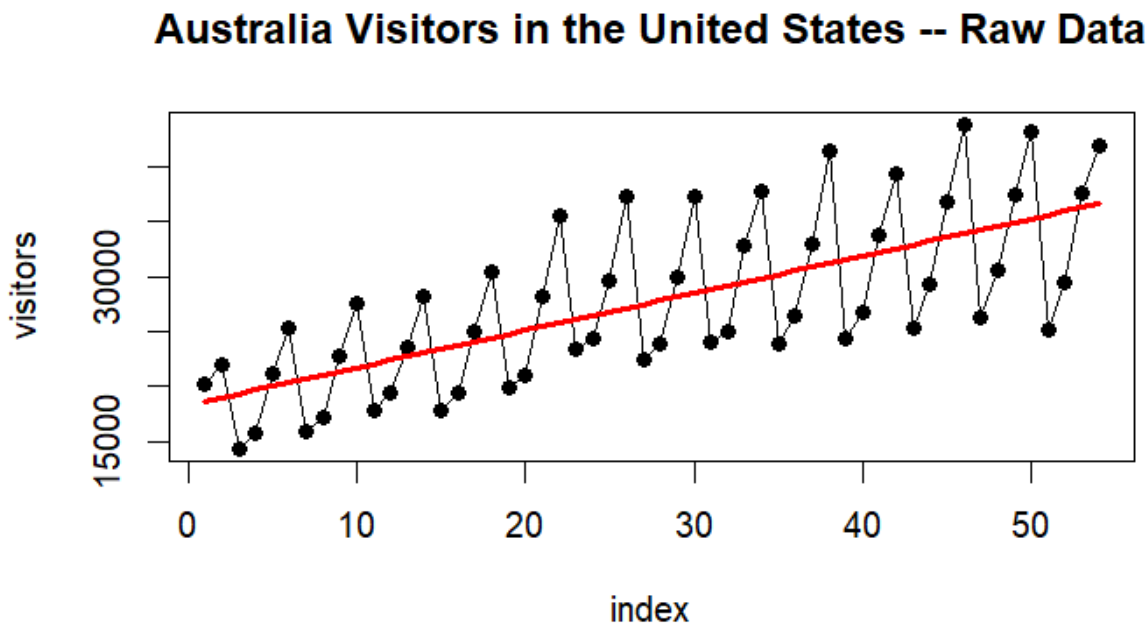


2. Using all the data parameterize a base time series simple regression model using "index" as the independent variable. Show the summary of your regression output.

```
> #Part 2
> base.out=lm(visitors~index,data=australia)
> summary(base.out)

Call:
lm(formula = visitors ~ index, data = australia)

Residuals:
     Min       1Q   Median       3Q      Max
-10405.0  -4299.8    875.2   3320.7  10194.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18314.62    1528.44  11.983  < 2e-16 ***
index         339.28      48.35   7.017 4.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5538 on 52 degrees of freedom
Multiple R-squared:  0.4863,  Adjusted R-squared:  0.4765
F-statistic: 49.23 on 1 and 52 DF,  p-value: 4.654e-09
```

3. Drawing on Analysis Part 1 above, show a properly titled plot of the time series data with the simple regression line layered on the graph in a contrasting color.

```
#Part 3
points(base.out$fitted.values,type="l",lwd=3,col="red")
```



4. Execute and interpret a Durbin-Watson test on your model results.

```
> #Part 4
> durbin.out=durbinWatsonTest(base.out)
> durbin.out
 lag Autocorrelation D-W Statistic p-value
   1     -0.08170513      2.143781   0.706
 Alternative hypothesis: rho != 0
```

The D-W statistic is 2.143781. Since it is close to 2 and falls within the normal range, then it indicates that there is no autocorrelation.

The null hypothesis for the Durbin-Watson test is that there is no autocorrelation in the residuals. Since the p-value is 0.706, which is greater than the 0.05 standard significance level, we do not have enough evidence to reject the null hypothesis that there is no autocorrelation.

5. Note the original data appears to have a pronounced cyclical pattern. Assuming the complete cycles are four quarters long, construct a set of seasonal indices which describe

the typical annual fluctuations in visitors. Use these indices to deseasonalize the visitors data. Store this deseasonalized data in a column in the original data frame.

```
#Part 5
#Making Seasonal Indices

indices=data.frame(month=1:4,average=0,index=0)
for(i in 1:4) {
  count=0
  for(j in 1:nrow(australia)) {
    if(i==australia$quarter[j]) {

indices$average[i]=indices$average[i]+australia$visitors[j]
      count=count+1
    }
  }
  indices$average[i]=indices$average[i]/count
  indices$index[i]=indices$average[i]/mean(australia$visitors)
}

#Deseasonalizing the original data

for(i in 1:4){
  for(j in 1:nrow(australia)){
    if(i==australia$quarter[j]){

australia$deseason.visitors[j]=australia$visitors[j]/indices$ind
ex[i]
    }
  }
}
```

6. Using the deseasonalized data parameterize four different regression models. A simple regression model will be the base case to be followed by second order, third order, and fourth order polynomial models which attempt to describe the longer-term secular fluctuations in the deseasonalized data.

```
#Part 6
#Conducting the deseasonalized regression
#1

> desreg.out=lm(deseason.visitors~index,data=australia)
> summary(desreg.out)

Call:
```

```
lm(formula = deseason.visitors ~ index, data = australia)

Residuals:
    Min      1Q  Median      3Q     Max
-3339.0  -729.9  -119.1   867.5  3742.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18535.3      357.0   51.91   <2e-16 ***
index          331.3       11.3   29.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1294 on 52 degrees of freedom
Multiple R-squared:  0.943,   Adjusted R-squared:  0.9419
F-statistic: 860.1 on 1 and 52 DF,  p-value: < 2.2e-16


> #2
>
secondreg.out=lm(deseason.visitors~index+I(index^2),data=austral
ia)
> summary(secondreg.out)

Call:
lm(formula = deseason.visitors ~ index + I(index^2), data =
australia)

Residuals:
     Min       1Q   Median       3Q      Max
-2089.62  -788.97    23.67   639.59  2970.65

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16755.219    445.746  37.589  < 2e-16 ***
index         521.979     37.391  13.960  < 2e-16 ***
I(index^2)     -3.468      0.659  -5.262 2.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1052 on 51 degrees of freedom
Multiple R-squared:  0.963,   Adjusted R-squared:  0.9616
F-statistic: 664.6 on 2 and 51 DF,  p-value: < 2.2e-16


> #3
```

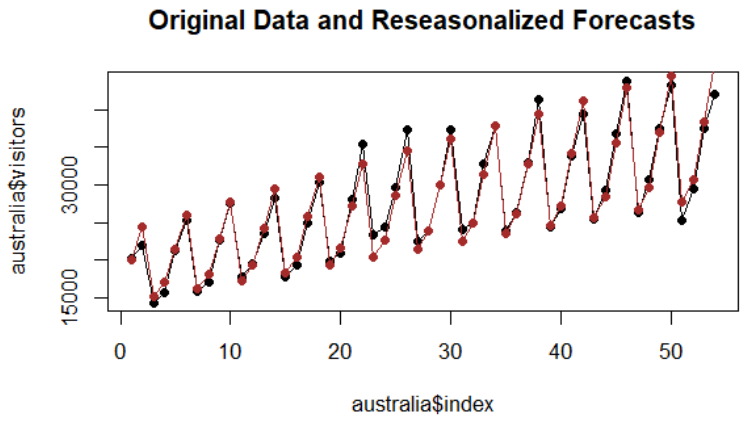
```
>
thirdreg.out=lm(deseason.visitors~index+I(index^2)+I(index^3),da
ta=australia)
> summary(thirdreg.out)

Call:
lm(formula = deseason.visitors ~ index + I(index^2) +
I(index^3),
    data = australia)

Residuals:
     Min       1Q   Median       3Q      Max
-1895.30  -813.00    13.51   712.29  3105.28

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.739e+04  6.071e+02  28.637  < 2e-16 ***
index        3.905e+02  9.472e+01   4.123 0.000141 ***
I(index^2)   2.454e+00  3.982e+00   0.616 0.540492
I(index^3)  -7.178e-02  4.762e-02  -1.507 0.138001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1039 on 50 degrees of freedom
Multiple R-squared:  0.9647,  Adjusted R-squared:  0.9625
F-statistic: 454.9 on 3 and 50 DF,  p-value: < 2.2e-16

> #4
>
fourthreg.out=lm(deseason.visitors~index+I(index^2)+I(index^3)+I
(index^4),data=australia)
> summary(fourthreg.out)

Call:
lm(formula = deseason.visitors ~ index + I(index^2) + I(index^3)
+
    I(index^4), data = australia)

Residuals:
     Min       1Q   Median       3Q      Max
-1903.72  -829.00    39.59   729.63  3057.26

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.760e+04  7.999e+02  21.998   <2e-16 ***
index        3.198e+02  1.976e+02   1.618    0.112
I(index^2)   8.120e+00  1.442e+01   0.563    0.576
```

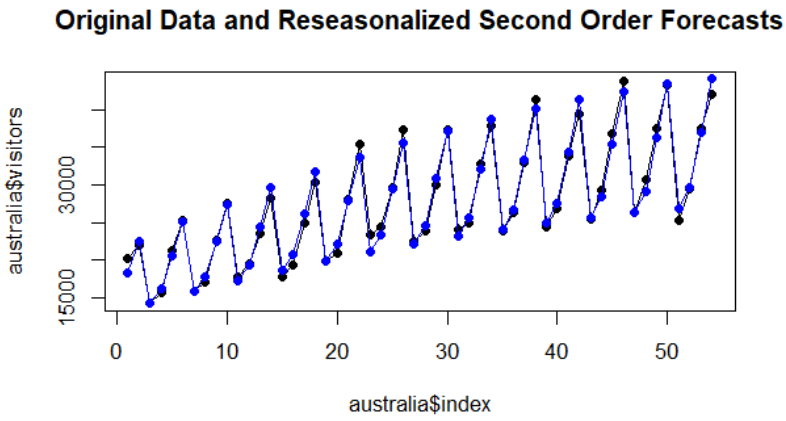
```
I(index^3)  -2.310e-01  3.922e-01  -0.589    0.559
I(index^4)   1.448e-03  3.539e-03   0.409    0.684
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1048 on 49 degrees of freedom
Multiple R-squared:  0.9648,  Adjusted R-squared:  0.9619
F-statistic: 335.5 on 4 and 49 DF,  p-value: < 2.2e-16
```

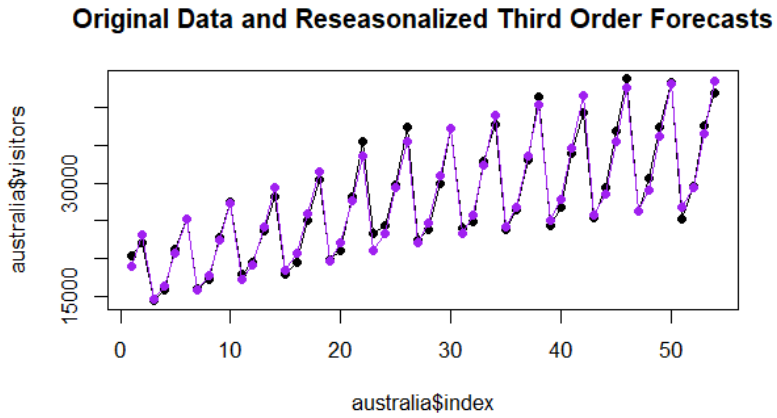
7. Reseasonalize the fitted values for each of the four models, storing the reseasonalized values in separate columns in the original data frame. Drawing on Analysis Part 3 above, construct a plot showing the original data and the fitted values for each of the four regression models. Show the four sets of fitted values plots in contrasting colors and title the graph appropriately.

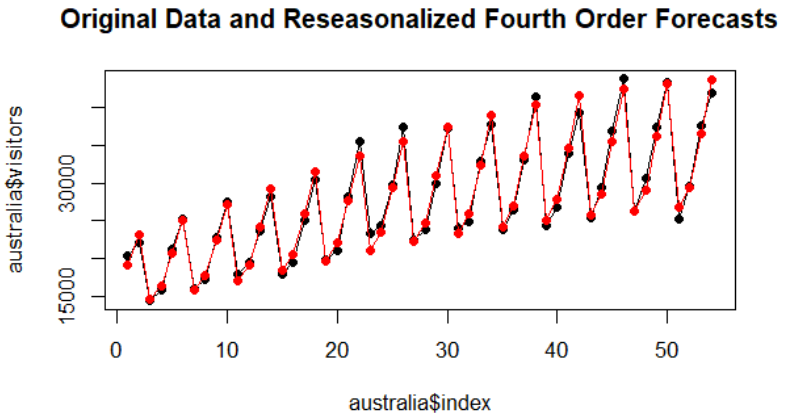
```
#Part 7
#Reseasonalizing Forecasts

australia$deseason.forecast=desreg.out$fitted.values
australia$deseasonsecond.forecast=secondreg.out$fitted.values
australia$deseasonthird.forecast=thirdreg.out$fitted.values
australia$deseasonfourth.forecast=fourthreg.out$fitted.values

for(i in 1:4){
  for(j in 1:nrow(australia)){
    if(i==australia$quarter[j]){

australia$reseason.forecast[j]=australia$deseason.forecast[j]*
        indices$index[i]

australia$reseasonsecond.forecast[j]=australia$deseasonsecond.fo
recast[j]*
        indices$index[i]

australia$reseasonthird.forecast[j]=australia$deseasonthird.fore
cast[j]*
        indices$index[i]

australia$reseasonfourth.forecast[j]=australia$deseasonfourth.fo
recas[j]*
        indices$index[i]
    }
  }
}
```

```
par(mfrow=c(2,2))
plot(australia$index,australia$visitors,type="o",pch=19,
     main="Original Data and Reseasonalized Forecasts")
points(australia$index,australia$reseason.forecast,
       type="o",pch=19,col="brown")
plot(australia$index,australia$visitors,type="o",pch=19,
     main="Original Data and Reseasonalized Second Order
Forecasts")
points(australia$index,australia$reseasonsecond.forecast,
       type="o",pch=19,col="blue")
plot(australia$index,australia$visitors,type="o",pch=19,
     main="Original Data and Reseasonalized Third Order
Forecasts")
points(australia$index,australia$reseasonthird.forecast,
       type="o",pch=19,col="purple")
plot(australia$index,australia$visitors,type="o",pch=19,
     main="Original Data and Reseasonalized Fourth Order
Forecasts")
points(australia$index,australia$reseasonfourth.forecast,
       type="o",pch=19,col="red")

par(mfrow=c(2,2))
```



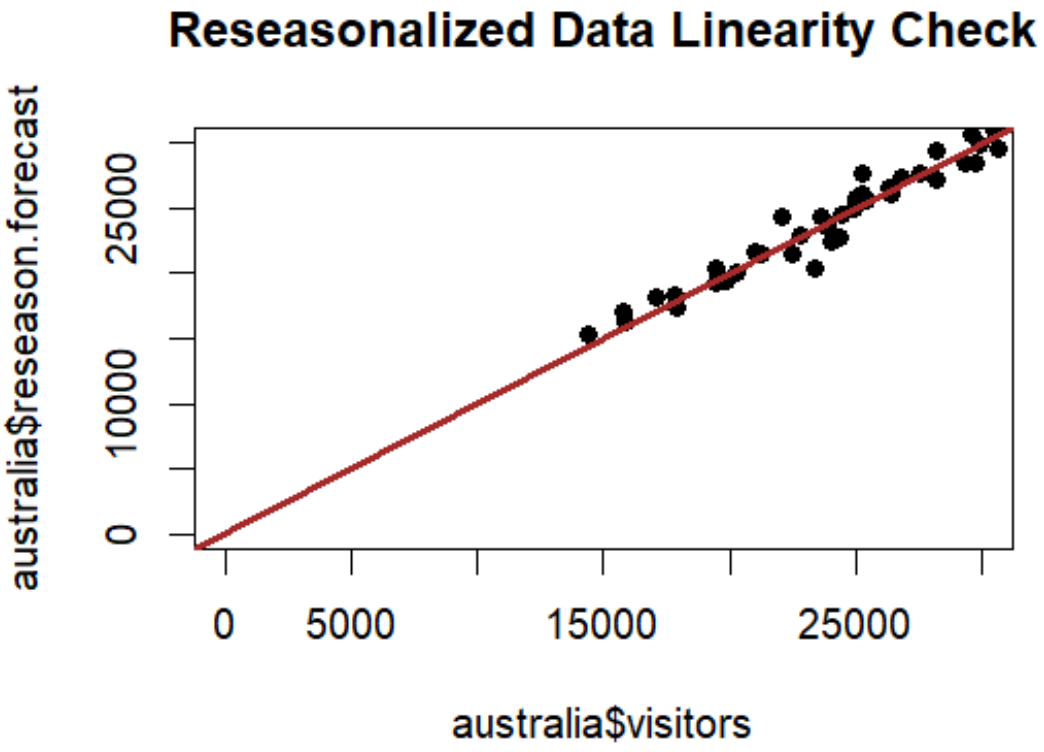




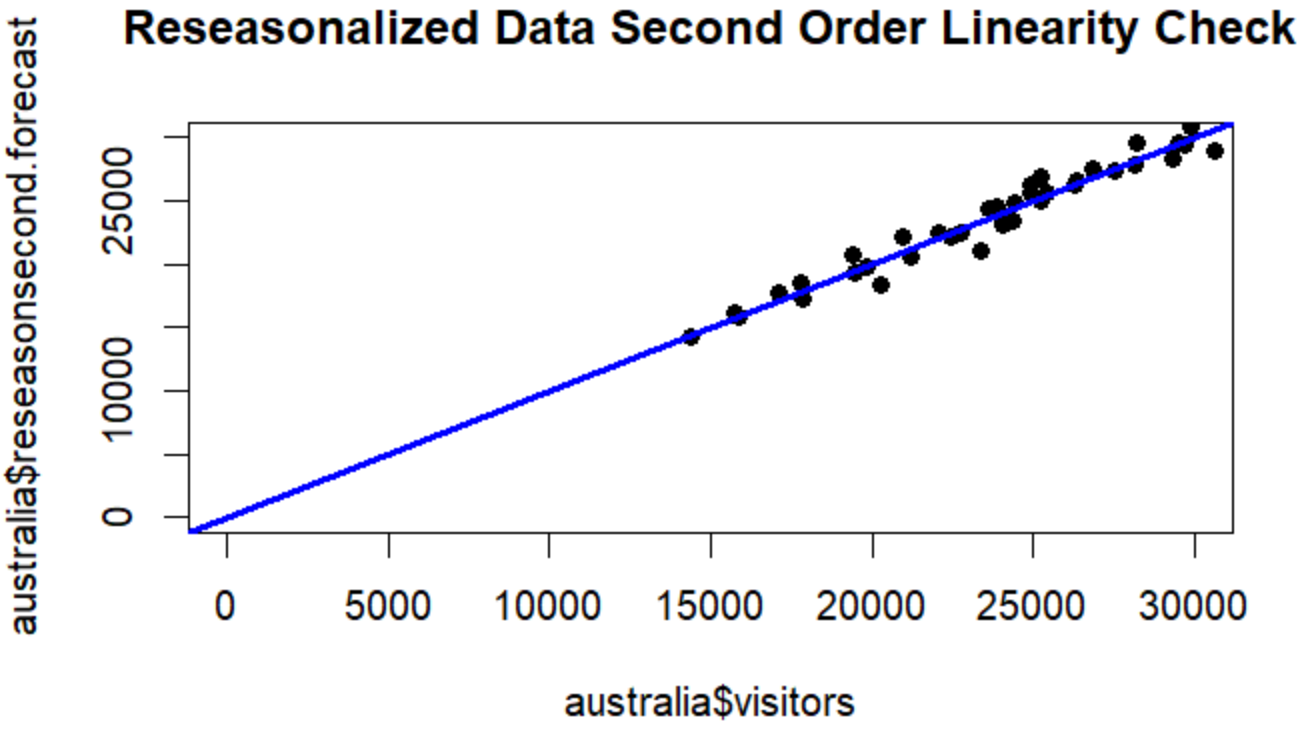


8. Select the model which in your view is the best fit to the deseasonalized data. Give a brief justification as to why you believe your selection is the best fit.

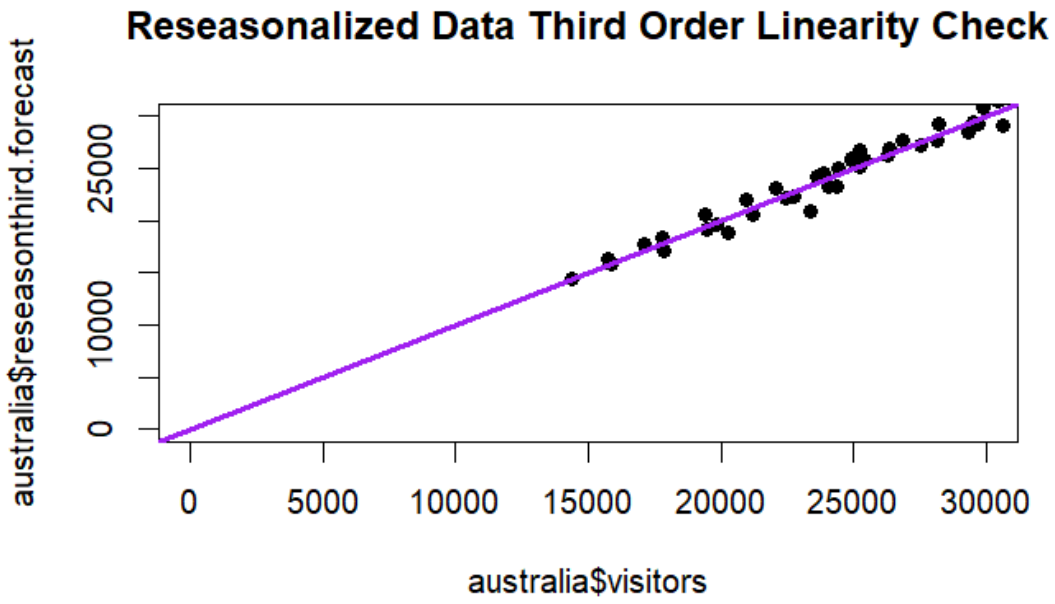
```
#Part 8
#Linearity First Model
plot(australia$visitors,australia$reseason.forecast,pch=19,
     xlim=c(0,30000),ylim=c(0,30000),
     main="Reseasonalized Data Linearity Check")
abline(0,1,lwd=3,col="brown")
> cor(australia$visitors,australia$reseason.forecast)
[1] 0.9852278
```



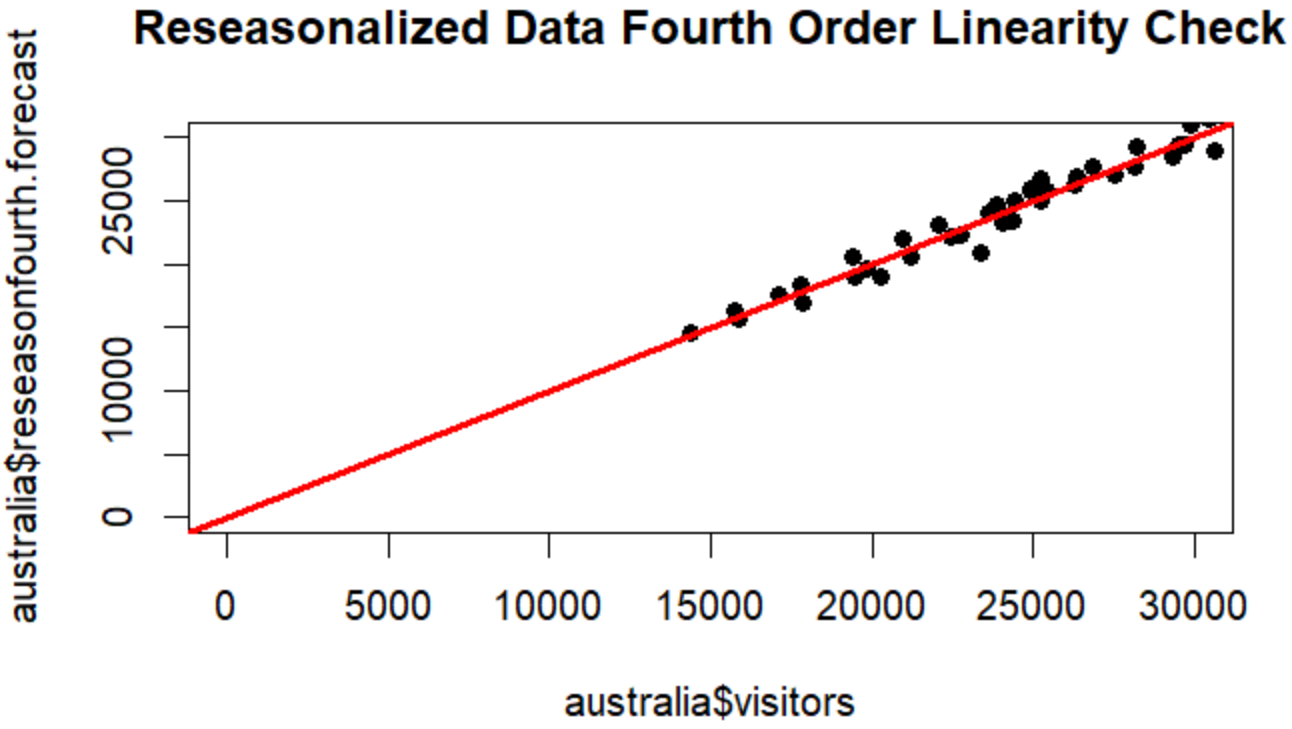

```
#Linearity Second Model
plot(australia$visitors,australia$reseasonsecond.forecast,pch=19
,
     xlim=c(0,30000),ylim=c(0,30000),
     main="Reseasonalized Data Second Order Linearity Check")
abline(0,1,lwd=3,col="blue")
> cor(australia$visitors,australia$reseasonsecond.forecast)
[1] 0.9908193
```




```
#Linearity Third Model
plot(australia$visitors,australia$reseasonthird.forecast,pch=19,
     xlim=c(0,30000),ylim=c(0,30000),
     main="Reseasonalized Data Third Order Linearity Check")
abline(0,1,lwd=3,col="purple")
> cor(australia$visitors,australia$reseasonthird.forecast)
[1] 0.9913627
```




```
#Linearity Fourth Model
plot(australia$visitors,australia$reseasonfourth.forecast,pch=19
,
     xlim=c(0,30000),ylim=c(0,30000),
     main="Reseasonalized Data Fourth Order Linearity Check")
abline(0,1,lwd=3,col="red")
> cor(australia$visitors,australia$reseasonfourth.forecast)
[1] 0.9913673
```



First Model: The R-squared is **0.943** and the correlation is **0.9852278**.
Second Model: The R-square is **0.963** and the correlation is **0.9908193.**
Third Model: The R-square is **0.9647** and the correlation is **0.9913627.**
Fourth Model: The R-square is **0.9648** and the correlation is **0.9913673**.

The fourth model seems to be the best fit for the deseasonalized data. It has the highest R-square value (0.9648) among the models, so it explains the most variance in the number of visitors to the United States from Australia (y variable). Although all the models have a very strong positive correlation, the fourth model also has the highest correlation among them (0.9913673).

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above preprocessing and analysis instructions and 2) the results of those instructions and needed written interpretations. The first line of your script file should be a "#" comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**