Stat 145 Problem Set 1

Anne Christine Amores

2025-08-27

1. Importing

```
# Importing the datasets

# TS1
ts1 <- read.csv("TS1.csv")
# TS2
ts2 <- read.csv("TS2.csv")

head(ts1)
```

```
##         date        y
## 1 2005-01-01 9.039826
## 2 2005-02-01 8.975322
## 3 2005-03-01 9.595398
## 4 2005-04-01 8.955137
## 5 2005-05-01 9.220265
## 6 2005-06-01 8.800123
```

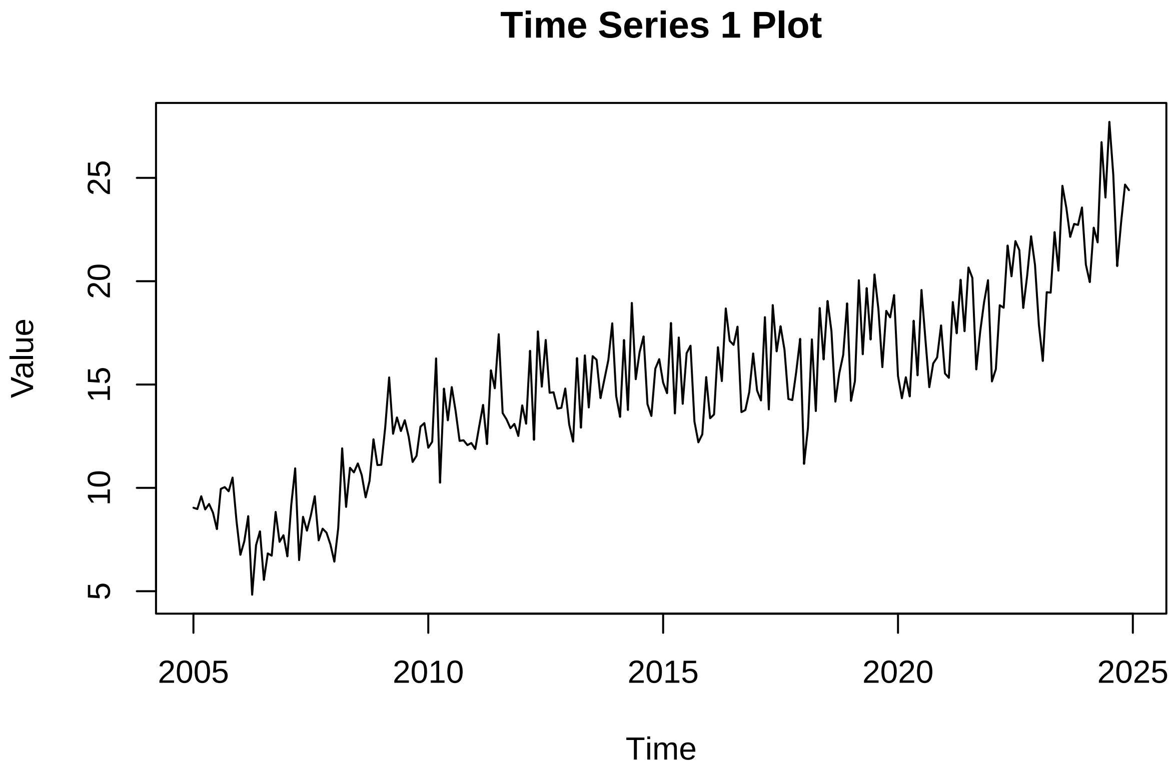
```
head(ts2)
```

```
##         date        y
## 1 2005-01-01 8.408048
## 2 2005-02-01 7.712259
## 3 2005-03-01 6.823825
## 4 2005-04-01 3.812830
## 5 2005-05-01 2.978984
## 6 2005-06-01 1.751672
```

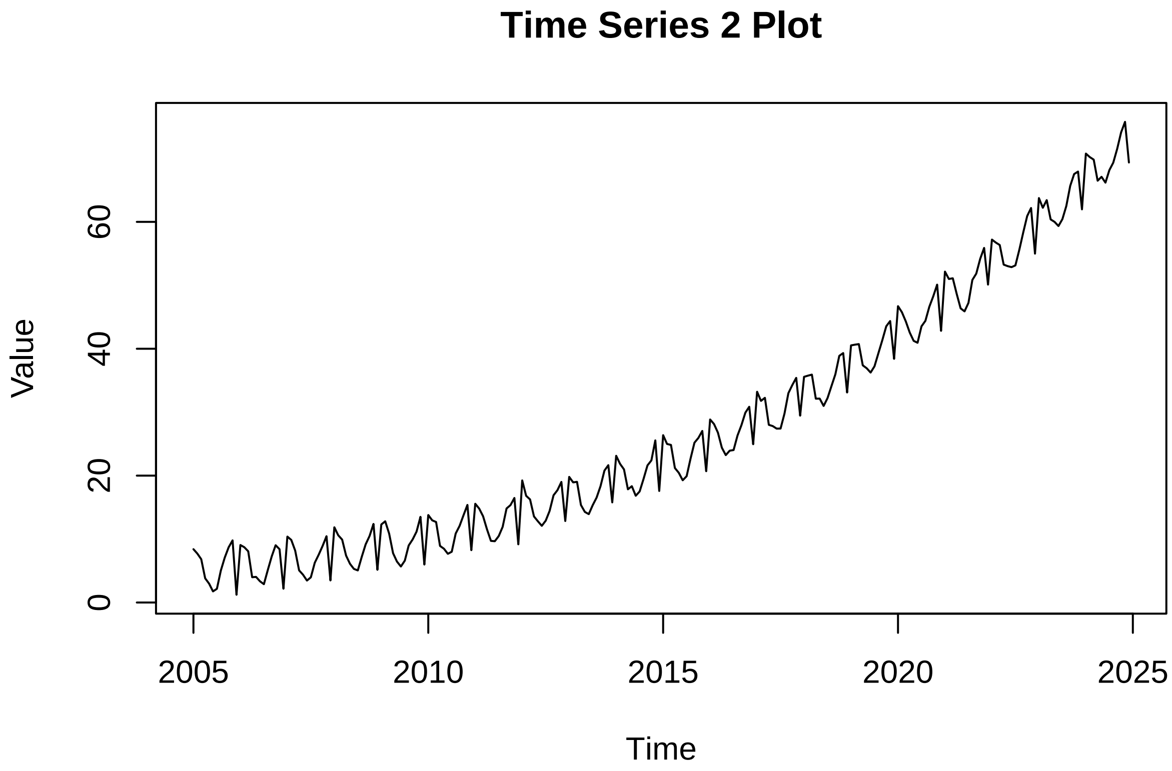
2. Plotting

```
# Plotting the time series for both datasets

# TS1
series1 <- ts(ts1$y, start=c(2005,1), frequency = 12)
plot.ts(series1, main = "Time Series 1 Plot", ylab = "Value")
```



```r
# TS2
series2 <- ts(ts2$y, start=c(2005,1), frequency = 12)
plot.ts(series2, main = "Time Series 2 Plot", ylab = "Value")
```



3. Trend

Based on visual inspection, there seems to be an upward trend for both series. To confirm, we conduct formal tests. Firstly, for the parametric test:

Linear regression:

Let

$$H_o : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Decision rule: Reject H_o if p-value < 0.05.

```
# Creating a time index

# TS1
ts1$date <- 1:nrow(ts1)
# TS2
ts2$date <- 1:nrow(ts2)

# Conducting formal testing - parametric
# Linear Regresion Model
```

```r
# TS1
trend1_par <- lm(y ~ date, ts1)
summary(trend1_par)
```

```
##
## Call:
## lm(formula = y ~ date, data = ts1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8271 -1.4695  0.1571  1.4582  6.3579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.203164   0.274997   29.83   <2e-16 ***
## date        0.055953   0.001978   28.28   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.123 on 238 degrees of freedom
## Multiple R-squared:  0.7707, Adjusted R-squared:  0.7697
## F-statistic: 799.8 on 1 and 238 DF,  p-value: < 2.2e-16
```

```r
# TS2
trend2_par <- lm(y ~ date, ts2)
summary (trend2_par)
```

```
##
## Call:
## lm(formula = y ~ date, data = ts2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.072  -4.593  -1.085   4.564  15.632
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.453560   0.770257   -7.08 1.61e-11 ***
## date         0.274416   0.005542   49.52  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.948 on 238 degrees of freedom
## Multiple R-squared:  0.9115, Adjusted R-squared:  0.9112
## F-statistic:  2452 on 1 and 238 DF,  p-value: < 2.2e-16
```

The null hypothesis is rejected for both models, indicating the presence of a linear trend for both series. Conducting a non-parametric test:

Mann-Kendall Test

Let

$$H_0 : \text{There is no monotonic trend.}$$

$$H_a : \text{There is a monotonic trend.}$$

Decision rule: Reject H_o if p-value < 0.05.

```r
# Conducting formal testing - nonparametric
# Mann-Kendall Test (from 'trend' package)
# Two-sided test

# TS1
mk.test (series1)
```

```
##
##  Mann-Kendall trend test
##
## data:  series1
## z = 16.201, n = 240, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.014200e+04 1.545533e+06 7.023013e-01
```

```r
# TS2
mk.test (series2)
```

```
##
##  Mann-Kendall trend test
##
## data:  series2
## z = 20.363, n = 240, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.531600e+04 1.545533e+06 8.827057e-01
```

The null hypothesis of the Mann-Kendall trend test is rejected for both series, indicating the presence of a monotonic trend.

Let

$$H_0 : \text{There is no monotonic trend.}$$

$$H_a : \text{There is a monotonic upward trend.}$$

Decision rule: Reject H_o if p-value < 0.05.

```
# One-sided test

# TS1
mk.test(series1, alternative = "greater")
```

```
##
##  Mann-Kendall trend test
##
## data:  series1
## z = 16.201, n = 240, p-value < 2.2e-16
## alternative hypothesis: true S is greater than 0
## sample estimates:
##            S         varS          tau
## 2.014200e+04 1.545533e+06 7.023013e-01
```

```
# TS2
mk.test(series2, alternative = "greater")
```

```
##
##  Mann-Kendall trend test
##
## data:  series2
## z = 20.363, n = 240, p-value < 2.2e-16
## alternative hypothesis: true S is greater than 0
## sample estimates:
##            S         varS          tau
## 2.531600e+04 1.545533e+06 8.827057e-01
```

More specifically, at $\alpha = 0.05$, there is sufficient evidence to conclude that an **upward trend** is present for both series, based on the rejection of the null hypothesis of a one-sided Mann-Kendall trend test (Ho: S = 0; Ha: S>0).

4. Stationarity

The **Augmented Dickey-Fuller test (ADF)** is a statistical test used to determine the stationarity of a time series. Its null hypothesis is that the series has a unit root, and is therefore non-stationary. On the other hand, its alternative hypothesis is that the series does not have a unit root, and is therefore stationary.

(https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/)

Let

$$H_o: \text{ The time series is non-stationary.}$$

$$H_a: \text{ The time series is stationary.}$$

Decision rule: Reject H_o if p-value < 0.05.

```
# Conducting ADF for both series

# TS1
adf.test(series1)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  series1
## Dickey-Fuller = -3.4816, Lag order = 6, p-value = 0.04521
## alternative hypothesis: stationary
```

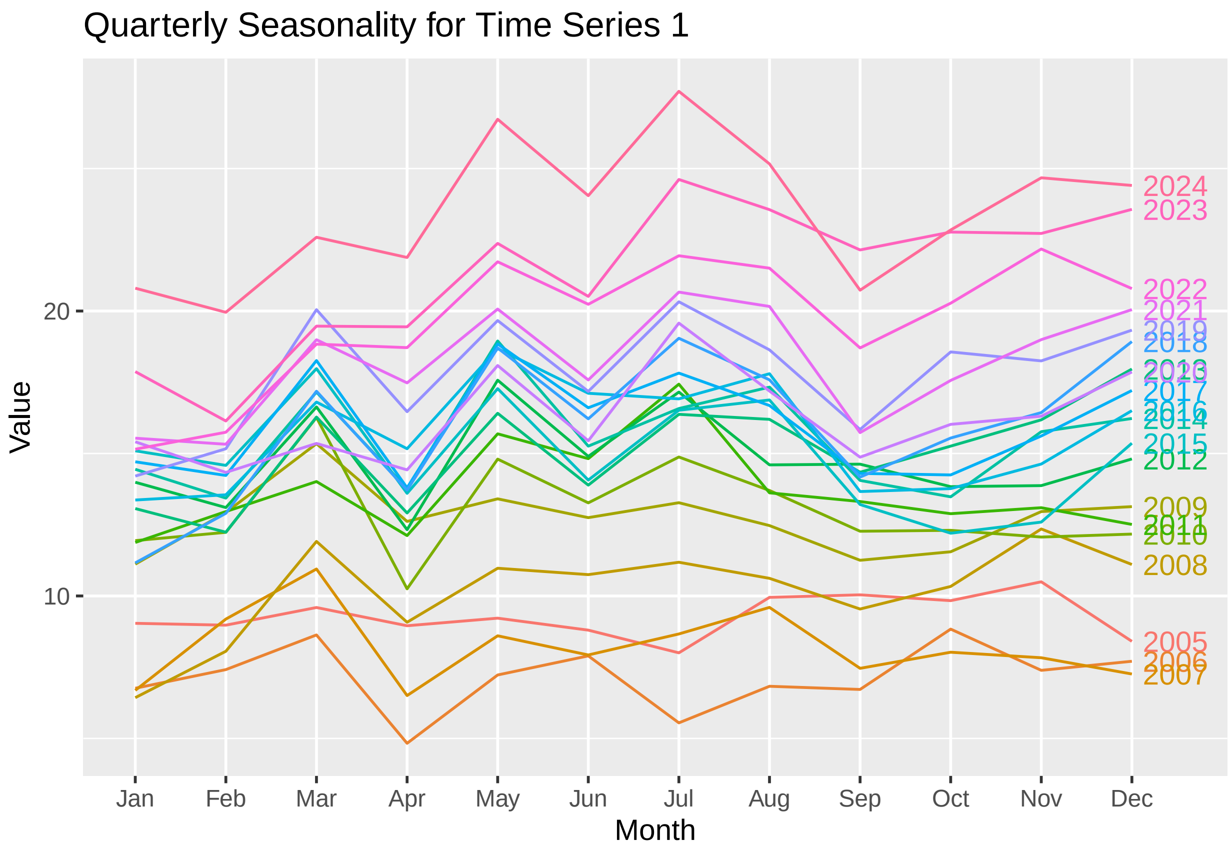
```
# TS2
adf.test(series2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  series2
## Dickey-Fuller = -1.1447, Lag order = 6, p-value = 0.9136
## alternative hypothesis: stationary
```

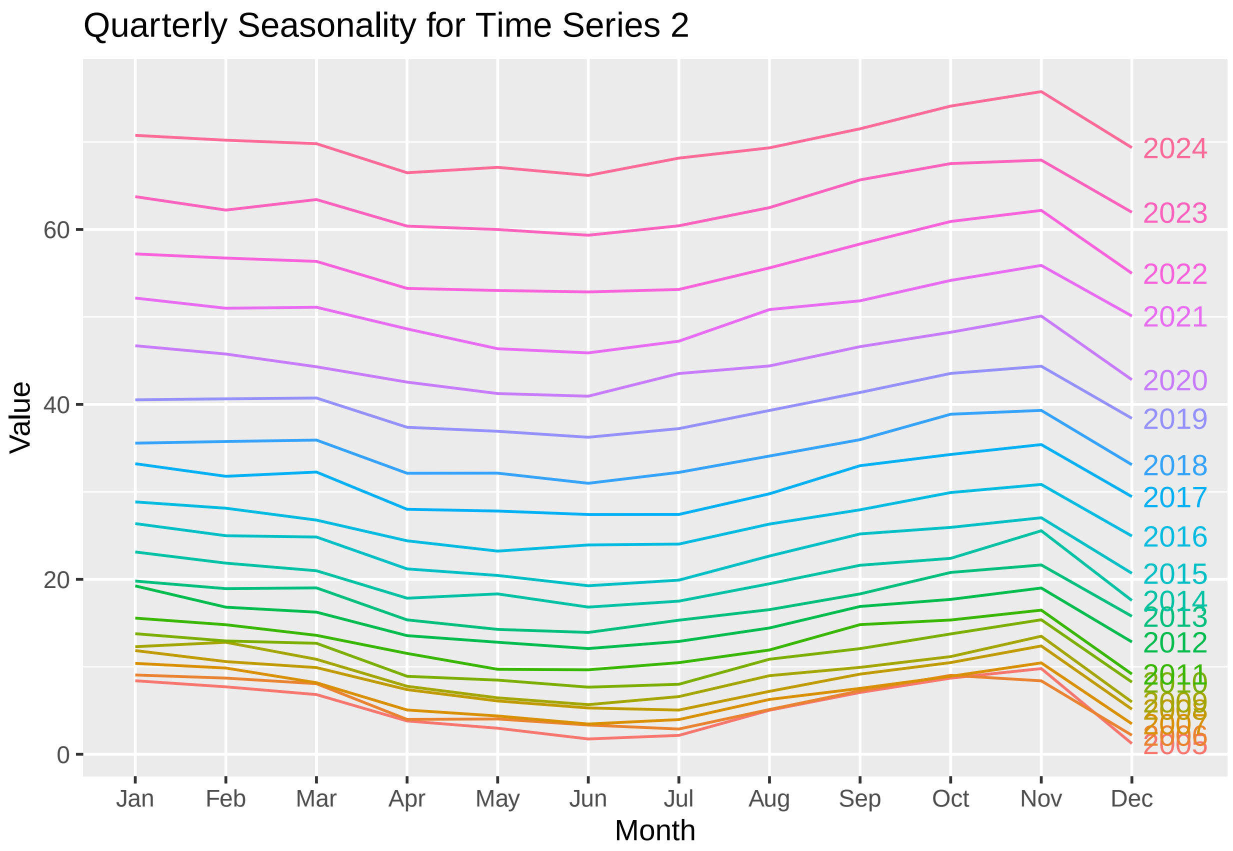
For time series 1, we have sufficient evidence to say, at $\alpha = 0.05$, that the time series is stationary. Meanwhile, for time series 2, there is insufficient evidence to say, at $\alpha = 0.05$, that the time series is stationary.

5. Seasonality

```
# Conducting
ggseasonplot(series1, year.labels = TRUE, main = "Quarterly Seasonality for Time Series 1", ylab = "Valu
```



```
ggseasonplot(series2, year.labels = TRUE, main = "Quarterly Seasonality for Time Series 2", ylab = "Valu
```



Based on visual inspection, for Time Series 1, the value rises for all years during March and May. Most years see a rise in value as well for the month of July. Meanwhile, there is a dip in the value for almost all years during the following months: April, June, and September. This suggests that seasonality is present in Time Series 1.

For Time Series 2, the value peaks in November for all years. Meanwhile, the value decreases for all years from April to July. This suggests that seasonality is also present in Time Series 2.

We now perform formal statistical tests to verify.

Seasonal Dummy Models:

Let

$$H_0 : \text{All } \beta_i = 0$$

$$H_a : \text{At least one } \beta_i \neq 0$$

Decision rule: Reject H_o if p-value < 0.05.

```
# Parametric test: regression on seasonal dummies

# TS1
```

```r
seasonality1 <- lm(series1 ~ seasonaldummy(series1)) #using 'forecast' package
summary(seasonality1)
```

```
##
## Call:
## lm(formula = series1 ~ seasonaldummy(series1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1650  -2.3250   0.3437   2.1988  10.9988
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                15.7653     0.9671  16.302   <2e-16 ***
## seasonaldummy(series1)Jan  -2.8320     1.3676  -2.071   0.0395 *
## seasonaldummy(series1)Feb  -2.6417     1.3676  -1.932   0.0547 .
## seasonaldummy(series1)Mar   0.3499     1.3676   0.256   0.7983
## seasonaldummy(series1)Apr  -2.3630     1.3676  -1.728   0.0854 .
## seasonaldummy(series1)May   0.9856     1.3676   0.721   0.4718
## seasonaldummy(series1)Jun  -0.8021     1.3676  -0.587   0.5581
## seasonaldummy(series1)Jul   0.9460     1.3676   0.692   0.4898
## seasonaldummy(series1)Aug   0.2385     1.3676   0.174   0.8617
## seasonaldummy(series1)Sep  -1.9156     1.3676  -1.401   0.1627
## seasonaldummy(series1)Oct  -1.2578     1.3676  -0.920   0.3587
## seasonaldummy(series1)Nov  -0.5444     1.3676  -0.398   0.6910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.325 on 228 degrees of freedom
## Multiple R-squared:  0.08871,    Adjusted R-squared:  0.04474
## F-statistic: 2.018 on 11 and 228 DF,  p-value: 0.02773
```

```r
# TS2
seasonality2 <- lm(series2 ~ seasonaldummy(series2))
summary(seasonality2)
```

```
##
## Call:
## lm(formula = series2 ~ seasonaldummy(series2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.151 -17.055  -5.742  13.701  43.979
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                25.3767     4.5308   5.601 6.11e-08 ***
## seasonaldummy(series2)Jan   4.5588     6.4075   0.711    0.478
## seasonaldummy(series2)Feb   3.7374     6.4075   0.583    0.560
## seasonaldummy(series2)Mar   3.2194     6.4075   0.502    0.616
## seasonaldummy(series2)Apr   0.1083     6.4075   0.017    0.987
## seasonaldummy(series2)May  -0.5839     6.4075  -0.091    0.927
## seasonaldummy(series2)Jun  -1.2379     6.4075  -0.193    0.847
## seasonaldummy(series2)Jul  -0.4665     6.4075  -0.073    0.942
## seasonaldummy(series2)Aug   1.6611     6.4075   0.259    0.796
## seasonaldummy(series2)Sep   3.7321     6.4075   0.582    0.561
## seasonaldummy(series2)Oct   5.4172     6.4075   0.845    0.399
## seasonaldummy(series2)Nov   6.6966     6.4075   1.045    0.297
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 20.26 on 228 degrees of freedom
## Multiple R-squared:  0.0164, Adjusted R-squared:  -0.03105
## F-statistic: 0.3457 on 11 and 228 DF,  p-value: 0.9742
```

The model for TS1 has a p-value < 0.05. Therefore, we reject the null hypothesis. That is, we have sufficient evidence to conclude that seasonality is present in the data.

Meanwhile, TS2 has a p-value > 0.05. Therefore, we do not reject the null hypothesis. That is, we do not have sufficient evidence to conclude that seasonality is present in the data.

Kruskal-Wallis Test

Let

$$H_o : \text{Time series is not seasonal.}$$

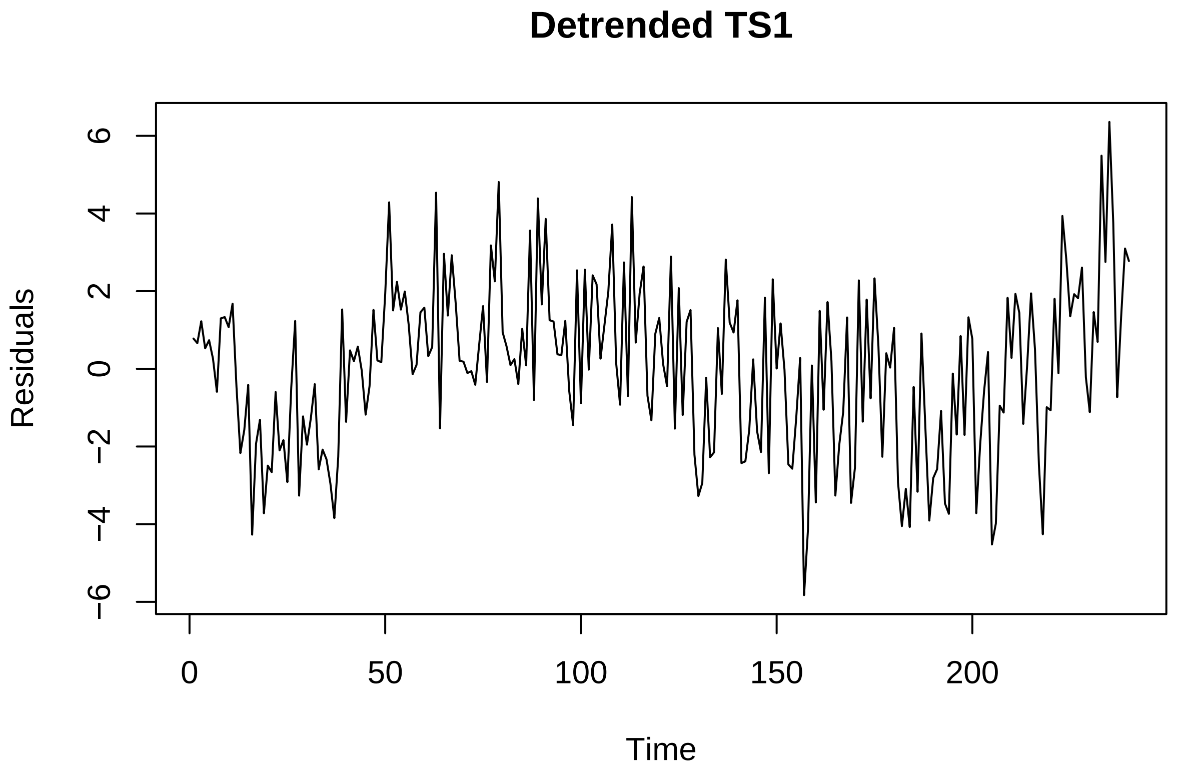
$$H_a : \text{Time series is seasonal.}$$

As it is, the Kruskal-Wallis test cannot be applied directly because both datasets exhibit a trend; **the trend has to be removed first** to properly test for seasonality.

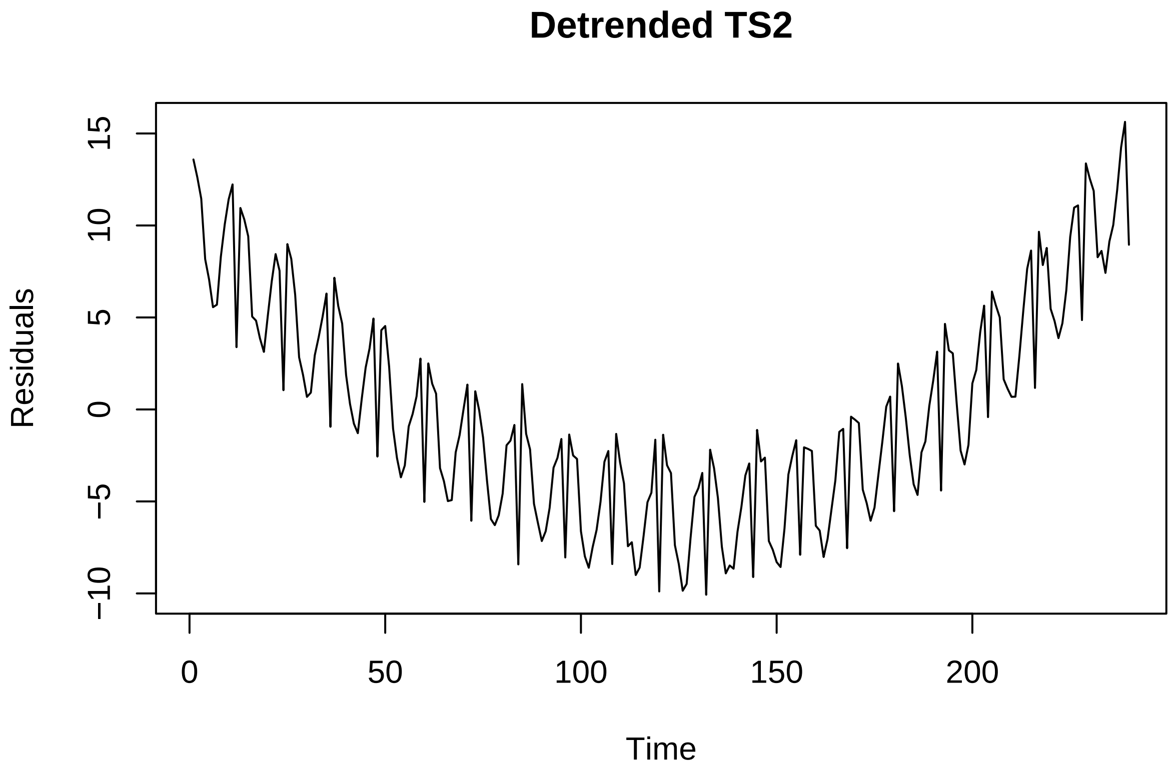
```r
# Creating data frames where each observation is labeled w/ its month for Kruskal-Wallis Test

# TS1
df1_2 <- data.frame(
  value = resid(trend1_par),            #resid() is used to remove trend
  month = factor(cycle(series1))
)
# TS2
df2_2 <- data.frame(
  value = resid(trend2_par),
  month = factor(cycle(series2))
)

# TS1 residuals
plot(resid(trend1_par), type = "l", main = "Detrended TS1", ylab = "Residuals", xlab = "Time")
```



```
# TS2 residuals
plot(resid(trend2_par), type = "l", main = "Detrended TS2", ylab = "Residuals", xlab = "Time")
```



```
# Kruskal-Wallis Test for Seasonality

# TS1
kruskal.test(value ~ month, df1_2)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by month
## Kruskal-Wallis chi-squared = 85.287, df = 11, p-value = 1.389e-13
```

```
# TS2
kruskal.test(value ~ month, df2_2)
```

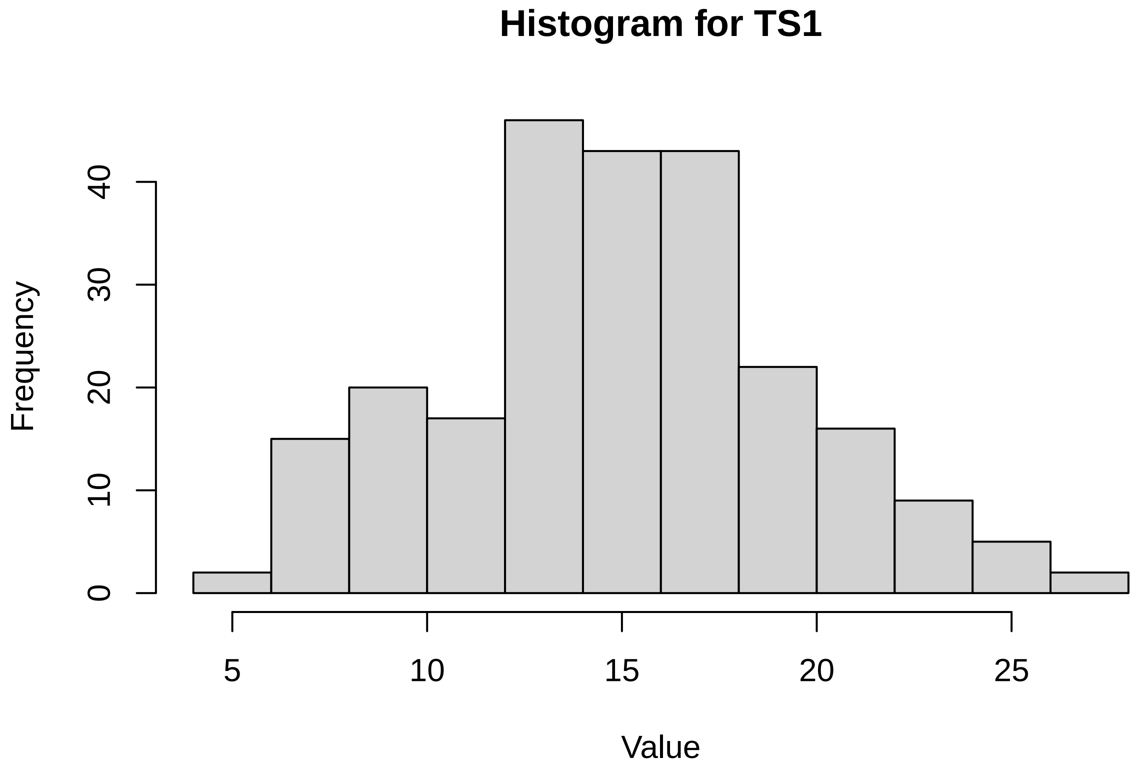
```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by month
## Kruskal-Wallis chi-squared = 50.019, df = 11, p-value = 6.211e-07
```

The residual plot of Time Series 1 is roughly flat, showing that trend has been removed to some extent. Meanwhile, for Time Series 2, the residual plot shows a convex (or upsidedown cave) shape. This means that a linear trend model is not able to fully remove the trend.

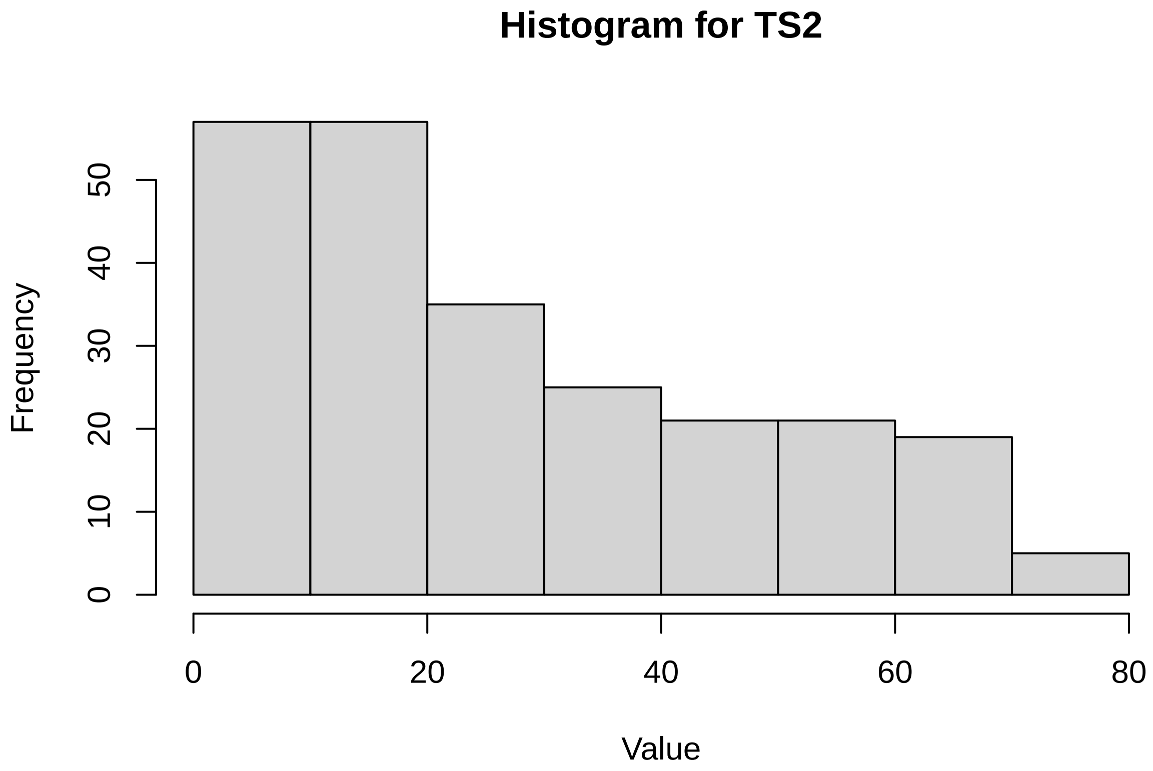
Nevertheless, we proceeded with the Kruskal-Wallis test for Time Series 1 and Time Series 2. This resulted to obtaining a p-value << 0.05 for both, implying rejection of their null hypotheses. Thus, we have sufficient evidence to conclude that **seasonality is present in both series.**

6. Descriptive Statistics

```
# Histogram for TS1
ts1_histogram <- hist(ts1$y, main = "Histogram for TS1", xlab = "Value")
```



```
# Histogram for TS2
ts2_histogram <- hist(ts2$y, main = "Histogram for TS2", xlab = "Value")
```



```
# Summary Measures

# TS1
summary(ts1$y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.829  12.263  14.845  14.946  17.638  27.710
```

```
# TS2
summary(ts2$y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.226  10.483  21.630  27.614  41.661  75.764
```

```
# Measures of Spread for TS1

# TS1
sd(ts1$y)
```

```
## [1] 4.425
```

```
var(ts1$y)
```

```
## [1] 19.58062
```

```
# Measures of Spread for TS2
sd(ts2$y)
```

```
## [1] 19.95479
```

```
var(ts2$y)
```

```
## [1] 398.1937
```

```
# Measures of Shape for TS1
skewness(ts1$y)
```

```
## [1] 0.1374598
```

```
kurtosis(ts1$y)
```

```
## [1] 2.807618
```

```
# Measures of Shape for TS2
skewness(ts2$y)
```

```
## [1] 0.6861874
```

```
kurtosis(ts2$y)
```

```
## [1] 2.294622
```