Project 6 - Time Series

Sanju Hyacinth C

23/10/2019

Contents

Project Objective:

To analyse and explore the time series dataset of **gas**, in-built in R, about the **Australian monthly gas production**. We are also asked to build an appropriate ARIMA model and report on the accuracy.

Required Packages:

```
#install.packages("forecast")
#install.packages("ggplot2")
#install.packages("caTools")
#install.packages("dplyr")
#install.packages("tseries")
```

Question 1

Data loading and exploration:

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.6.1

## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Registered S3 methods overwritten by 'forecast':
##   method             from
##   fitted.fracdiff    fracdiff
##   residuals.fracdiff fracdiff
```

```
# Loading the dataset from the forecsast package
forecast::gas
```

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov
## 1956  1709  1646  1794  1878  2173  2321  2468  2416  2184  2121  1962
## 1957  1751  1688  1920  1941  2311  2279  2638  2448  2279  2163  1941
## 1958  1773  1688  1783  1984  2290  2511  2712  2522  2342  2195  1931
## 1959  1730  1688  1899  1994  2342  2553  2712  2627  2363  2311  2026
## 1960  1762  1815  2005  2089  2617  2828  2965  2891  2532  2363  2216
## 1961  1804  1773  2015  2089  2627  2712  3007  2880  2490  2237  2205
## 1962  1868  1815  2047  2142  2743  2775  3028  2965  2501  2501  2131
## 1963  1910  1868  2121  2268  2690  2933  3218  3028  2659  2406  2258
## 1964  1889  1984  2110  2311  2785  3039  3229  3070  2659  2543  2237
## 1965  1962  1910  2216  2437  2817  3123  3345  3112  2659  2469  2332
## 1966  1910  1941  2216  2342  2923  3229  3513  3355  2849  2680  2395
## 1967  1994  1952  2290  2395  2965  3239  3608  3524  3018  2648  2363
## 1968  1994  1941  2258  2332  3323  3608  3957  3672  3155  2933  2585
## 1969  2057  2100  2458  2638  3292  3724  4652  4379  4231  3756  3429
## 1970  3345  4220  4874  5064  5951  6774  7997  7523  7438  6879  6489
## 1971  5919  6183  6594  6489  8040  9715  9714  9756  8595  7861  7753
## 1972  7778  7402  8903  9742 11372 12741 13733 13691 12239 12502 11241
## 1973 11569 10397 12493 11962 13974 14945 16805 16587 14225 14157 13016
## 1974 11704 12275 13695 14082 16555 17339 17777 17592 16194 15336 14208
## 1975 12354 12682 14141 14989 16159 18276 19157 18737 17109 17094 15418
## 1976 13260 14990 15975 16770 19819 20983 22001 22337 20750 19969 17293
## 1977 15117 16058 18137 18471 21398 23854 26025 25479 22804 19619 19627
## 1978 17243 18284 20226 20903 23768 26323 28038 26776 22886 22813 22404
## 1979 18839 18892 20823 22212 25076 26884 30611 30228 26762 25885 23328
## 1980 21433 22369 24503 25905 30605 34984 37060 34502 31793 29275 28305
## 1981 27730 27424 32684 31366 37459 41060 43558 42398 33827 34962 33480
## 1982 30715 30400 31451 31306 40592 44133 47387 41310 37913 34355 34607
## 1983 26138 30745 35018 34549 40980 42869 45022 40387 38180 38608 35308
## 1984 28801 33034 35294 33181 40797 42355 46098 42430 41851 39331 37328
## 1985 32494 33308 36805 34221 41020 44350 46173 44435 40943 39269 35901
## 1986 31239 32261 34951 38109 43168 45547 49568 45387 41805 41281 36068
```

```
## 1987 32791 34206 39128 40249 43519 46137 56709 52306 49397 45500 39857
## 1988 35567 37696 42319 39137 47062 50610 54457 54435 48516 43225 42155
## 1989 37541 37277 41778 41666 49616 57793 61884 62400 50820 51116 45731
## 1990 40459 40295 44147 42697 52561 56572 56858 58363 45627 45622 41304
## 1991 35592 35677 39864 41761 50380 49129 55066 55671 49058 44503 42145
## 1992 38963 38690 39792 42545 50145 58164 59035 59408 55988 47321 42269
## 1993 37059 37963 31043 41712 50366 56977 56807 54634 51367 48073 46251
## 1994 39975 40478 46895 46147 55011 57799 62450 63896 57784 53231 50354
## 1995 41600 41471 46287 49013 56624 61739 66600 60054
##        Dec
## 1956  1825
## 1957  1878
## 1958  1910
## 1959  1910
## 1960  2026
## 1961  1984
## 1962  2015
## 1963  2057
## 1964  2142
## 1965  2110
## 1966  2205
## 1967  2247
## 1968  2384
## 1969  3461
## 1970  6288
## 1971  8154
## 1972 10829
## 1973 12253
## 1974 13116
## 1975 14312
## 1976 16498
## 1977 18488
## 1978 19795
## 1979 21930
## 1980 25248
## 1981 32445
## 1982 28729
## 1983 30234
## 1984 34514
## 1985 32142
## 1986 34879
## 1987 37958
## 1988 39995
## 1989 42528
## 1990 36016
## 1991 38698
## 1992 39606
## 1993 43736
## 1994 38410
## 1995
```

```r
# Converting the gas data to a time series dataset
gts = ts(gas, start = c(1956,1), frequency = 12)
```

```r
# Printing the structure and class of data
str(gts)
```

```
##  Time-Series [1:476] from 1956 to 1996: 1709 1646 1794 1878 2173 ...
```

```r
# "ts" refers to time series data
class(gts)
```

```
## [1] "ts"
```

```r
# To print the start and end of the ts data
start(gts); end(gts)
```

```
## [1] 1956    1
```

```
## [1] 1995    8
```

```r
# Mode gives the data format
mode(gts)
```

```
## [1] "numeric"
```

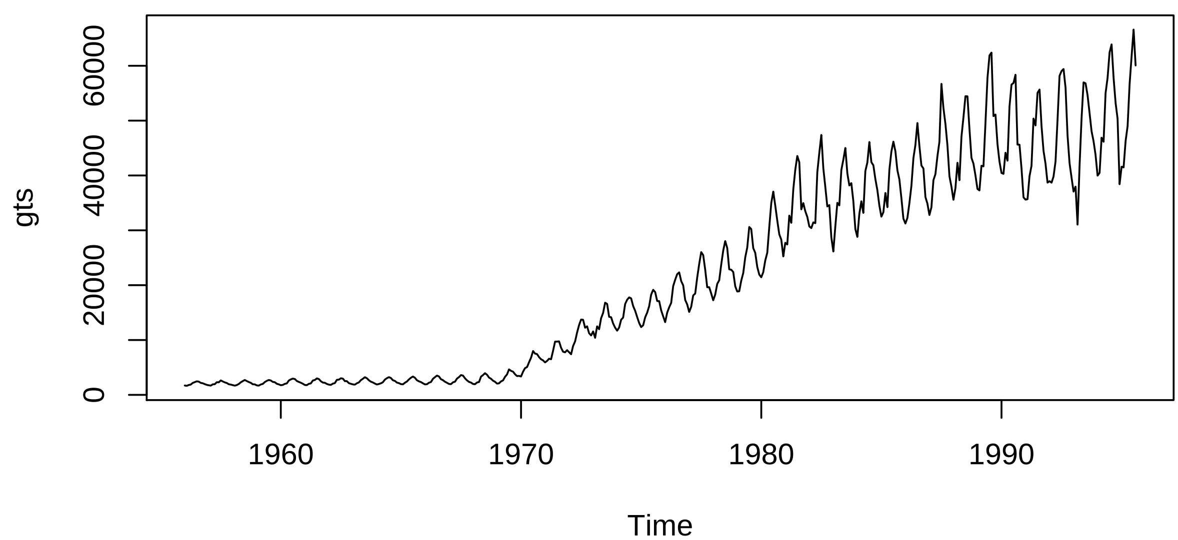
```r
# Print the summary of the ts data (usually not necessary for time series data)
summary(gts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1646    2675   16788   21415   38629   66600
```

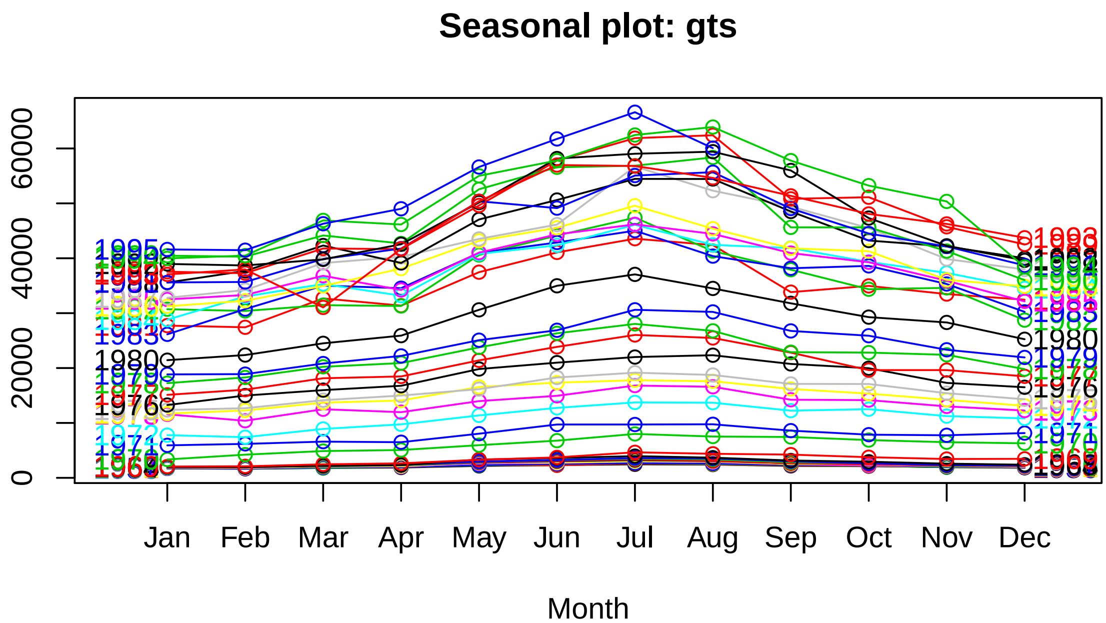
Data Plotting:

Let us look at the season plot and the month plot for the time series data.

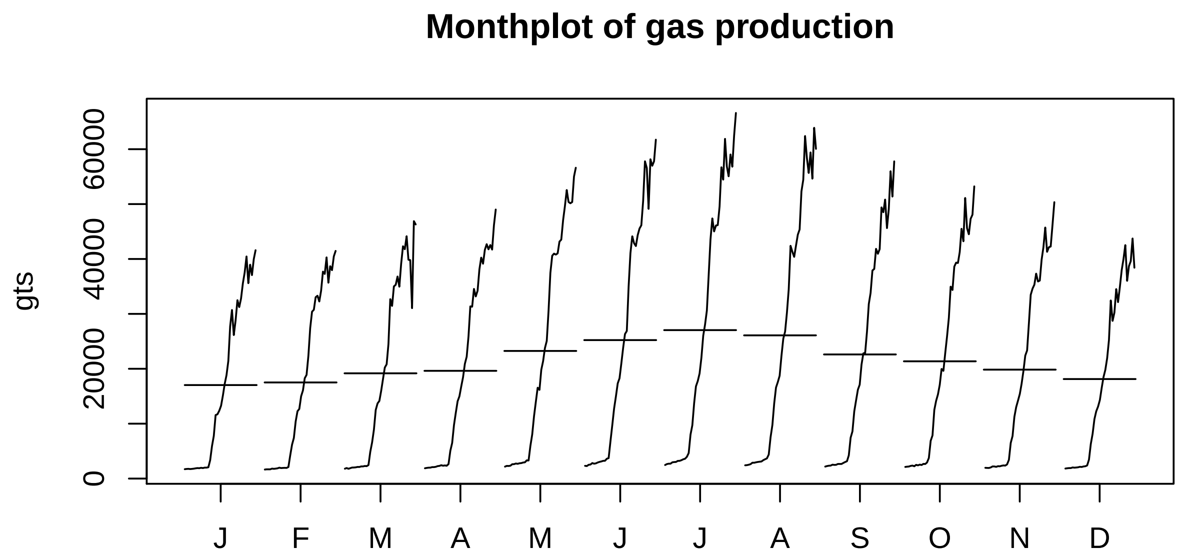
```r
plot(gts)
```



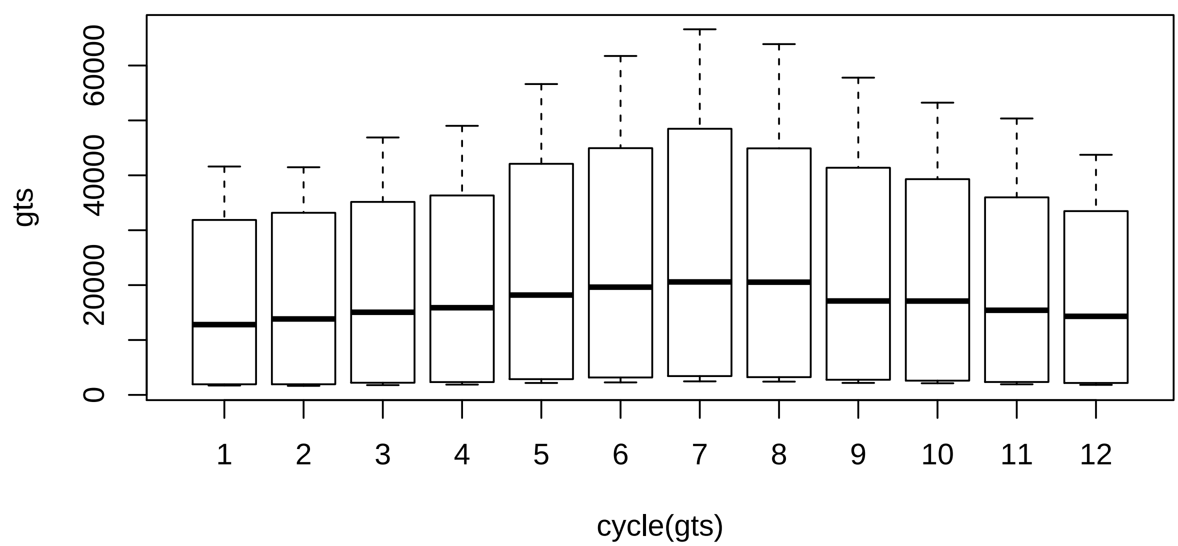
```r
# We see that until 1970, there has been a steady production of gas with no increase or decrease.
# But after 1970, we see a steady increase in production.
# Though there are dips here and there accounting for outliers, it pretty much shows a positive trend.
# We can ignore the years uptill 1970 as these were the starting years and do not show that much stabil

# Season plot:
seasonplot(gts, year.labels = TRUE, year.labels.left = TRUE, col = 1:12)
```




```r
# Month plot:
monthplot(gts, main = "Monthplot of gas production")
```

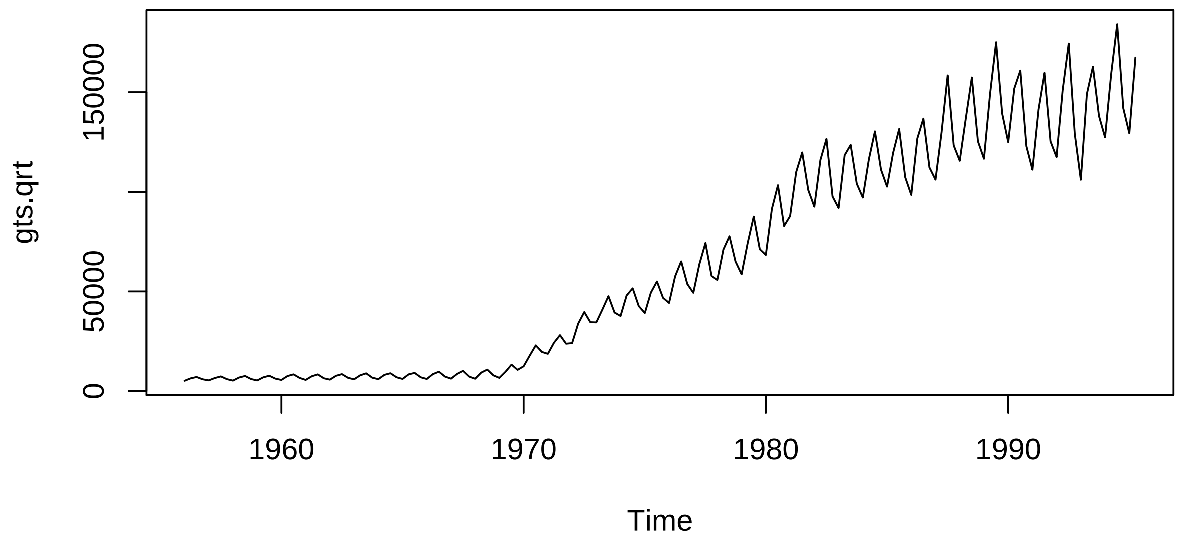



```r
# Boxplot:
boxplot(gts~ cycle(gts))
```

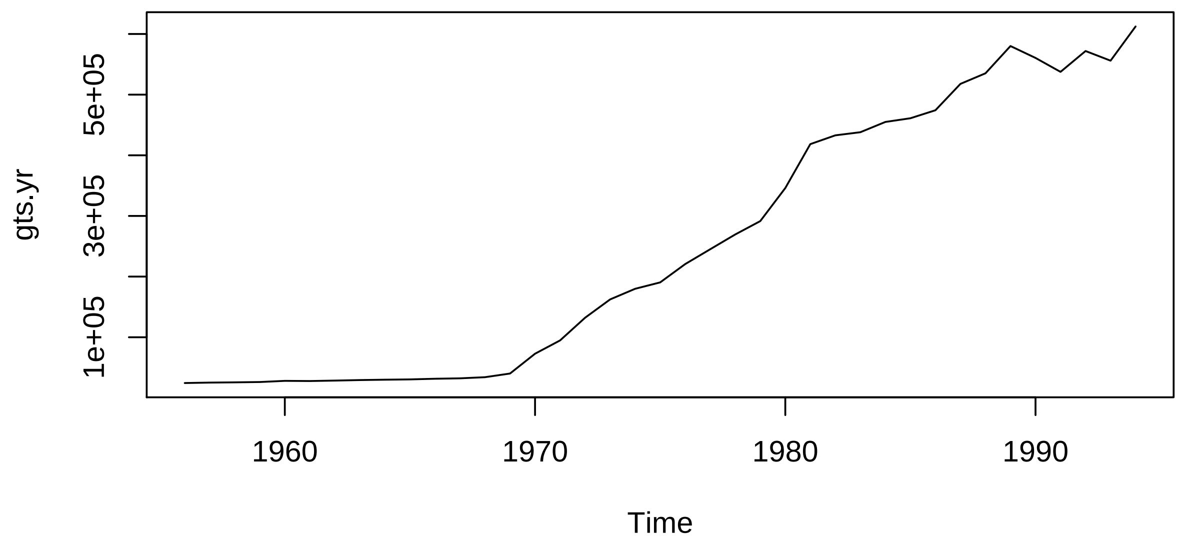



```r
# Aggregation at quarter and year level

gts.qrt = aggregate(gts, nfrequency = 4)
gts.yr = aggregate(gts, nfrequency = 1)
```

```r
plot(gts.qrt)
```




```r
plot(gts.yr)
```



Question 2

Observation and Inference from the plots:

Even in the season plot we see that uptil 1970, we do not find significantly growing production of gas. Except a little increase that started during the June of 1969 the rest of the months previous to 1969 show almost equal production within the years. The highest production recorded is during the July of 1995.

One more noteworthy point is, upon looking at the overall season plot, we see uptil the summer months there is an increased production and this goes down during the rainy and winter season. Most of the years show this change, which may contribute to seasonality.

From the monthplot and the boxplot we see that the production of gas follows a trend of increased production till July and gradually decreases throughout the year. We can, thereby assume it to follow some sort of a seasonlity pattern.

Components of the time series present in the data:

Upon continuing with the above inferences, we can assume the data to have both **trend and seasonality**. The data begins to be not showing any trend until 1970. After which, we see that the production of gas increased over the years, thus showing a **positive trend**.

The same can be applied for seasonlity but it still is quite tricky to conclude that the data has seasonlity. From our seasonplot, boxplot and monthplot, we can observe quite a pattern. Ultimately, the production of gas has followed an increasing pattern upto July, and is seen to be decreasing after July. With the lowest production during January and the highest being July.

Question 3

Periodicity:

The periodicity of the time series dataset is a pattern that occurs at regular time intervals.

The time series can be **seasonal or cyclical** based on the pattern repetition. We are sure that our data is a **monthly time series data** having one observation per month. But what can we say about the pattern or periodicity of the subject time series data.

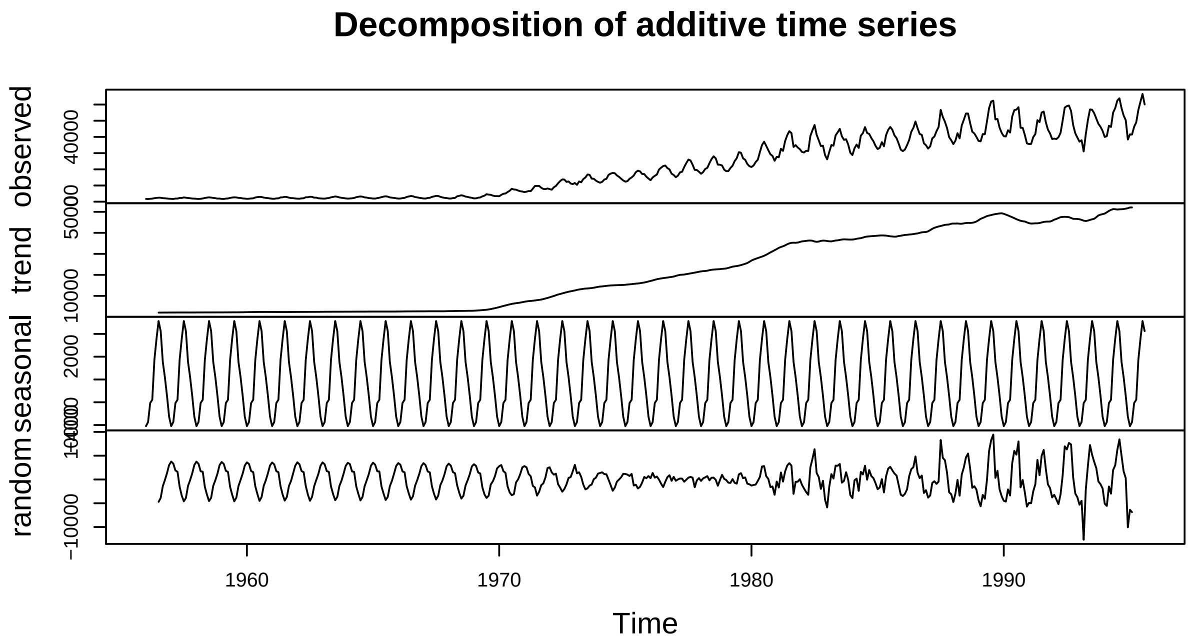
From the above graphs, we see that between the years **1974 - 1981** there seems to be a perfect seasonality and an increasing trend. But beyond that, we can see an **inconsistency** in both trend and seasonality. Though there is some of both seen till 1990; after 1990, we see there is neither a trend line nor a visibly stable seasonality. The few points were the value goes below the normal production cannot necessarily be considered outliers, but they seem inconsistent, which could be sampling error.

Also the fact that, the production was **constant** during the beginning years until 1970, a regular **increase** in production until 1990 and the **inconsistent or a mildly stagnant production** of gas after 1990 seems that this could might as well be a **long run cycle**. We might assume (to a very less degree) that the production of gas after the given years may **remain stagnant** for a few more years and **fall down or decrease**. This indicates that there are chances of identifying a **cyclic pattern** in this data.

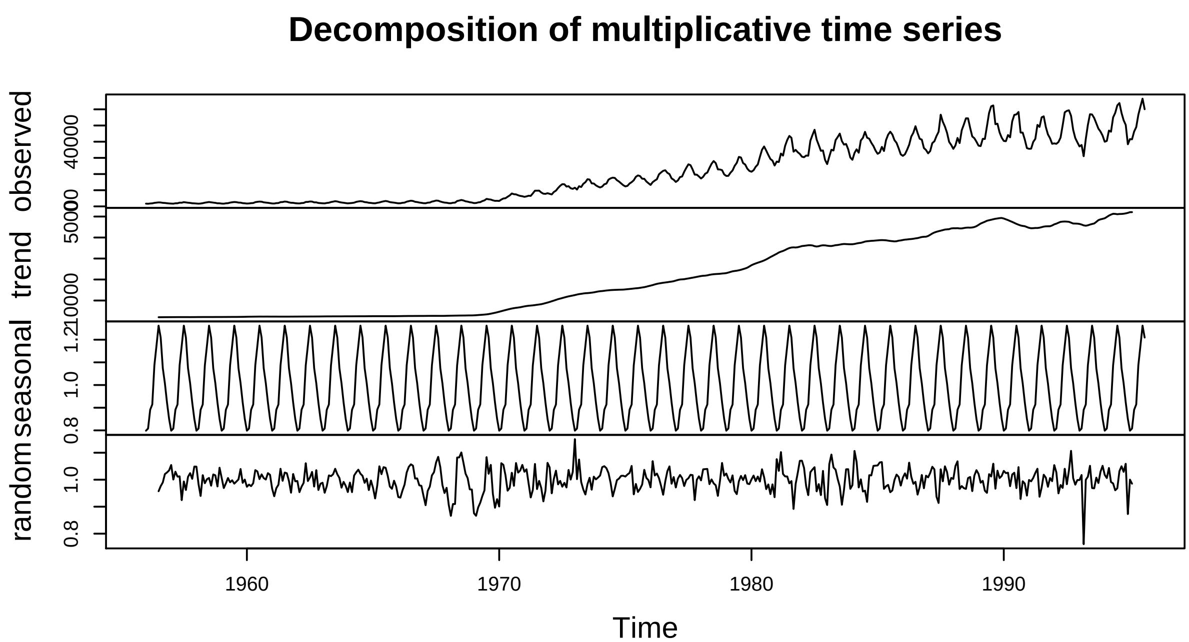
```r
frequency(gts)
```

```
## [1] 12
```

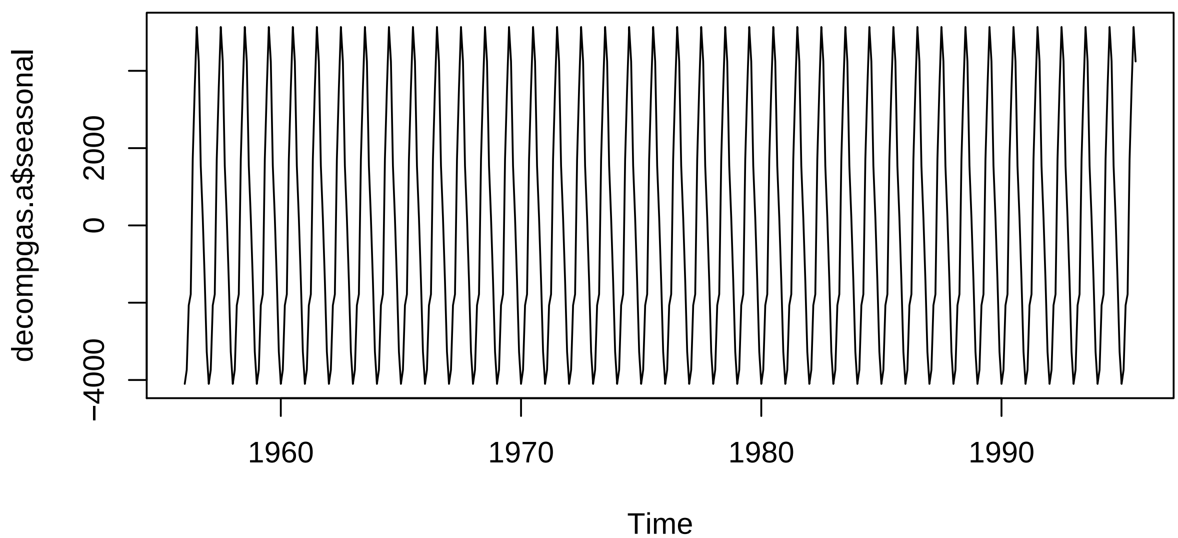
```r
decompgas.a = decompose(gts, type = "additive")
plot(decompgas.a)
```



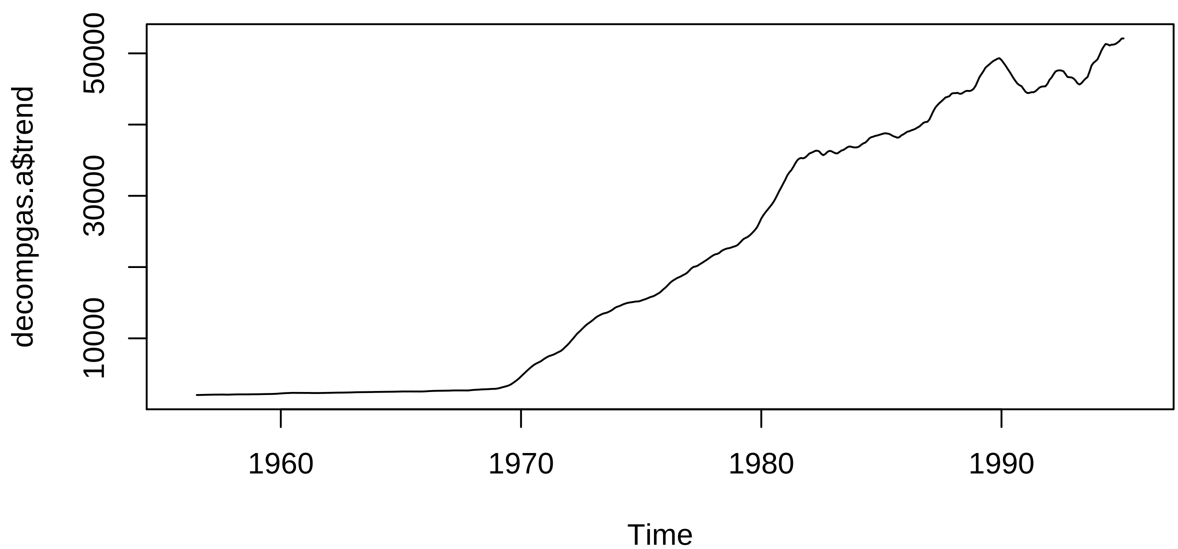

```r
decompgas.m = decompose(gts, type = "multiplicative")
plot(decompgas.m)
```



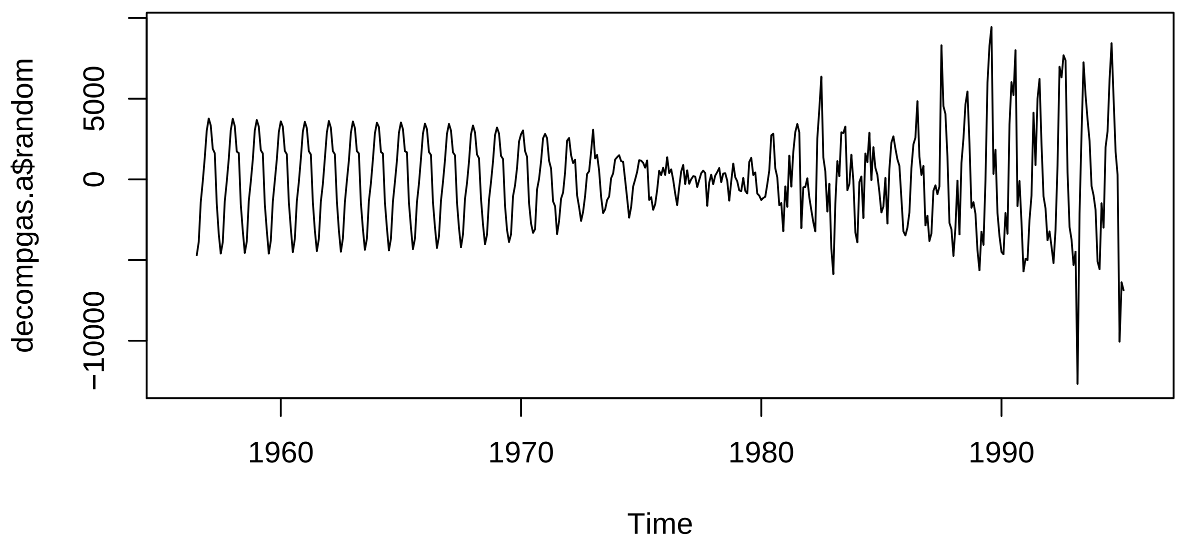

```r
# There isn't much change in the seasonality of both the graphs, hence we can consider the data as an a

# Individual components

plot(decompgas.a$seasonal)
```




```
plot(decompgas.a$trend)
```




```
plot(decompgas.a$random)
```




```r
# The trend seems to be significant
```

Question 4

Stationarity of the Time Series data:

A stationary time series is one whose properties do not depend on the timeat which the series is observed. The series will not have any predictable pattern. Another name for a stationary series is **White noise**.

We are aware that forecasting can be done only on a **stationary time series** data. If a time series data is not found to be stationary, we will first have to **stationarize the series**. A stationary time series has to have a **constant mean and variance**.

In our time series data, we find that the time series data definitely has a trend and a seasonality pattern. Usually time series with trend and seasonality is **non-stationary** as both trend and seasonality will affect the value of the time series at different times. Hence we may assume our data to be non stationary.

Test for Stationarity:

As per our visual assumption, we see that the time series data, as having a trend and seasonality, is non-stationary. Hence we will have to **stationarize** the data first to be able to do forecasting on it.

For this, we use the **Augmented Dickey-Fuller test**. This test is used to test whether a time series data is non stationary. There is a null and alternate hypothesis for the process. A lower p value will state that the time series is stationary.

Let us install and run the **tseries** library to access the **adf.test** function, which refers to the **augmented dickey-fuller test**.

```r
# Install and load the library tseries

library(tseries)
```

```
## Warning: package 'tseries' was built under R version 3.6.1
```

```r
# Augmented Dickey-Fuller test

adf.test(gts)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gts
## Dickey-Fuller = -2.7131, Lag order = 7, p-value = 0.2764
## alternative hypothesis: stationary
```

```r
# Dickey-Fuller = -2.7131
# p-value = 0.2764
# Lag order = 7
```

Hypothesis for ADF test:

Our above adf test on the gas.ts dataset has given a p-value of **0.2764**. As this is a test to stationarize the time series data, we have the below hypothesis made.

Null Hypothesis (H0) = Time series is not stationary

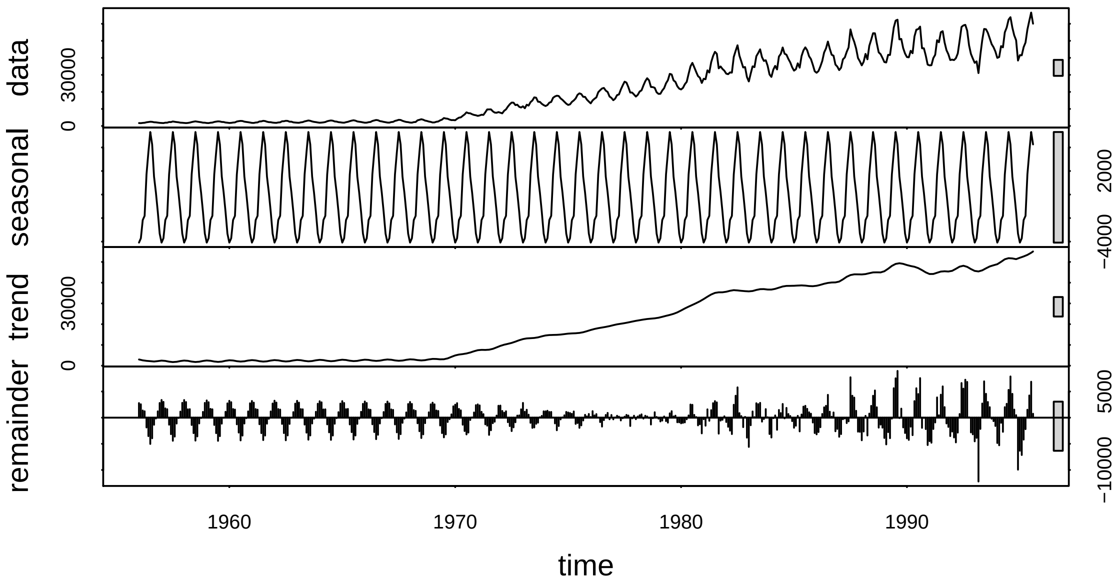
Alternate Hypothesis (Ha) = Time series is stationary

Only when the p-value is **less than or equal to 0.05** can we straight away reject the null hypothesis to approve the alternate hypothesis of the time series being stationary. In this case, as we have obtained a p-value not less than or equal to 0.05, we are **unable to reject the null hypothesis** and approve of the alternate hypothesis that the time series is stationary.

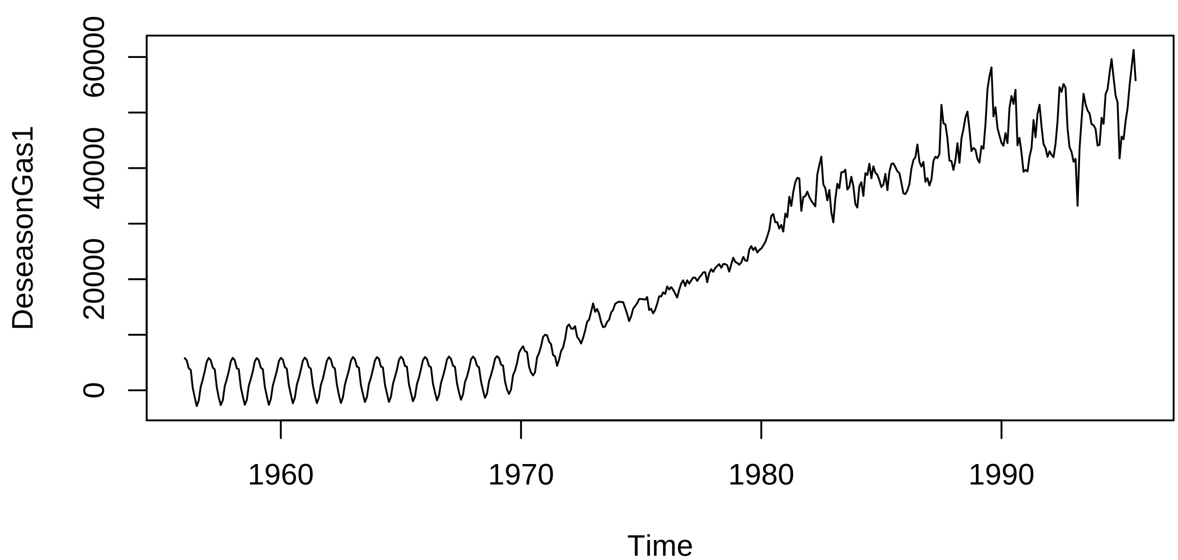
Since we have proved that our time series data is not stationary, we have create a difference series, ie. the difference of consecutive terms in a time series known as the difference series of order 1. This will help us to stationarize the time series data.

De-seasonalising the time series:

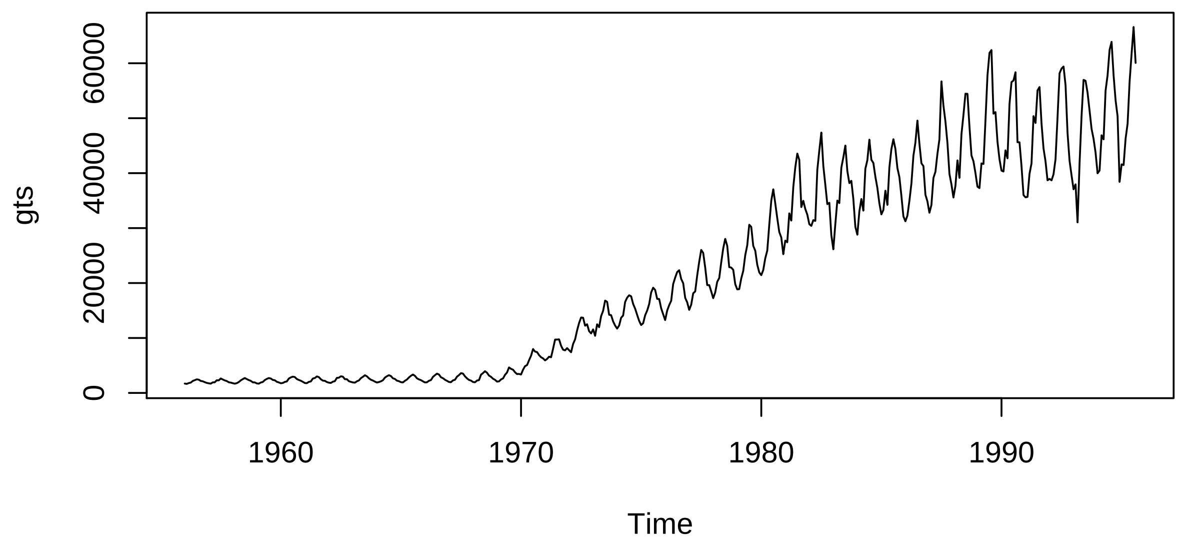
```r
# Decomposing using stl

des.gas = stl(gts, s.window = "p")
plot(des.gas)
```




```
# Deseasoning data

DeseasonGas1 = seasadj(des.gas)
plot(DeseasonGas1)
```




```
# Comparison
plot(gts)
```



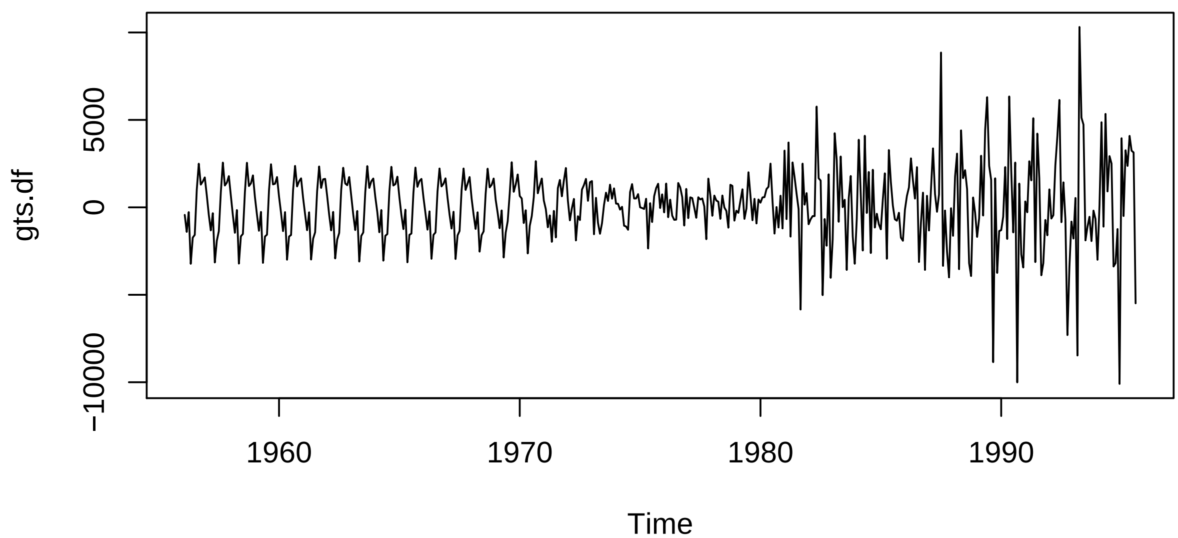

```
## We do not see that much seasonality upon decomposition
```

From the above **deseasonalised data**, we find that the effect of seasonality is now very less on the dataset. Deseasonlised data comprises of components exclusive of the seasonality factor. Hence while plotting these two together, we confirm that there has been a **presence of seasonality** and that now **the significance of seasonality is not so high on the data** after deseasonalisation.

Detrending the series:

Upon deseasonalising the data, we are now detrending the data inorder to make it a stationary series. And then we are about to take the Augmented Dickey-Fuller Test on the differenced series.

```
# Differencing the deseasonalised data

gts.df = diff(DeseasonGas1, differences = 1)
plot(gts.df)
```




```r
# We see a lot of sharp and extreme ponits beyond 1980 but they still lie close to the central line

# Augmented Dickey-Fuller Test on the differenced data

adf.test(gts.df)
```

```
## Warning in adf.test(gts.df): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  gts.df
## Dickey-Fuller = -18.14, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```r
# Dickey-Fuller = -18.14
# Lag order = 7
# p-value = 0.01
```

The result of Augmented Dickey-Fuller Test on the differenced data **gts.df** shows a very significant and less p-value of **0.01** (which is the minimum value to be printed showing that the p-value is less than the printed value of 0.01). By this we can **reject the null hypothesis** and approve of the alternate hypothesis that the time series is **stationary**.

This series is known as the **difference series of order one**

Question 5

Autocorrelations and Partial Autocorrelations:

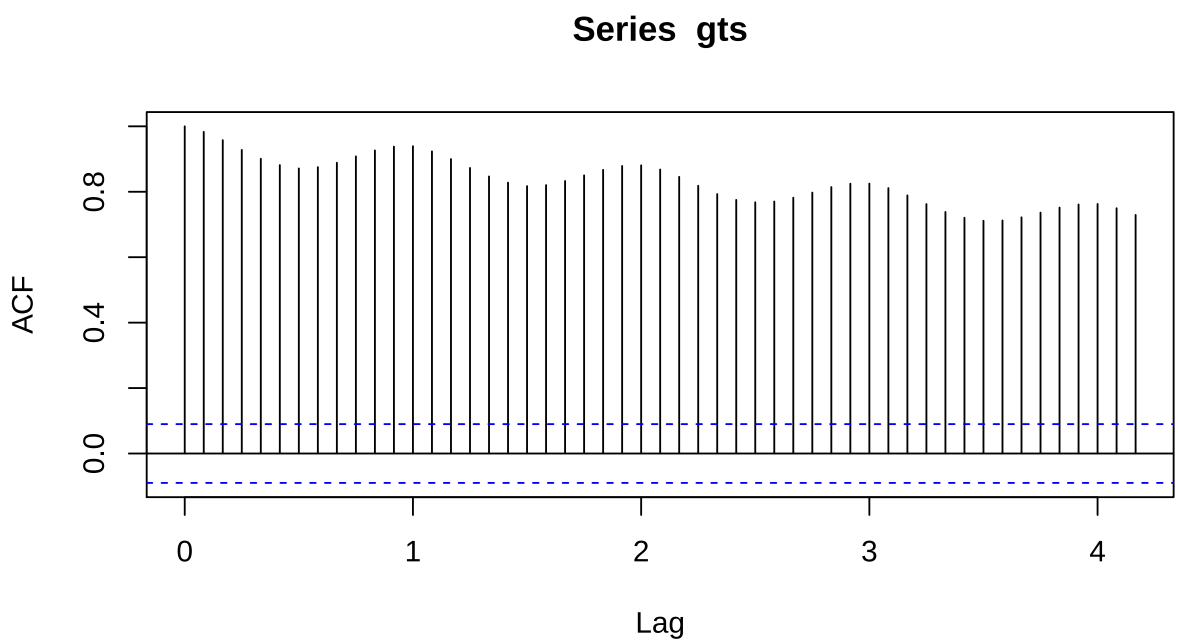
Though our original data is non stationary, we have our differenced data that is stationary. Hence, we can go about the next process of finding the **auto correlations and partial auto correlations** on the differenced data. Auto correlation can be done only on stationarised data, that does not have the effect of trend or seasonality.

Auto correlation is referred to as correlation with self. It consists of different **orders**. Auto corrrelation of different orders give inside information about the time series we are dealing with for analysis and forecasting. The auto corrrelation values range between **-1 and +1 only**. The values nearing -1 and +1 may correspond to a negative and positive correlation. And the values closer to 0 indicate no correlation.

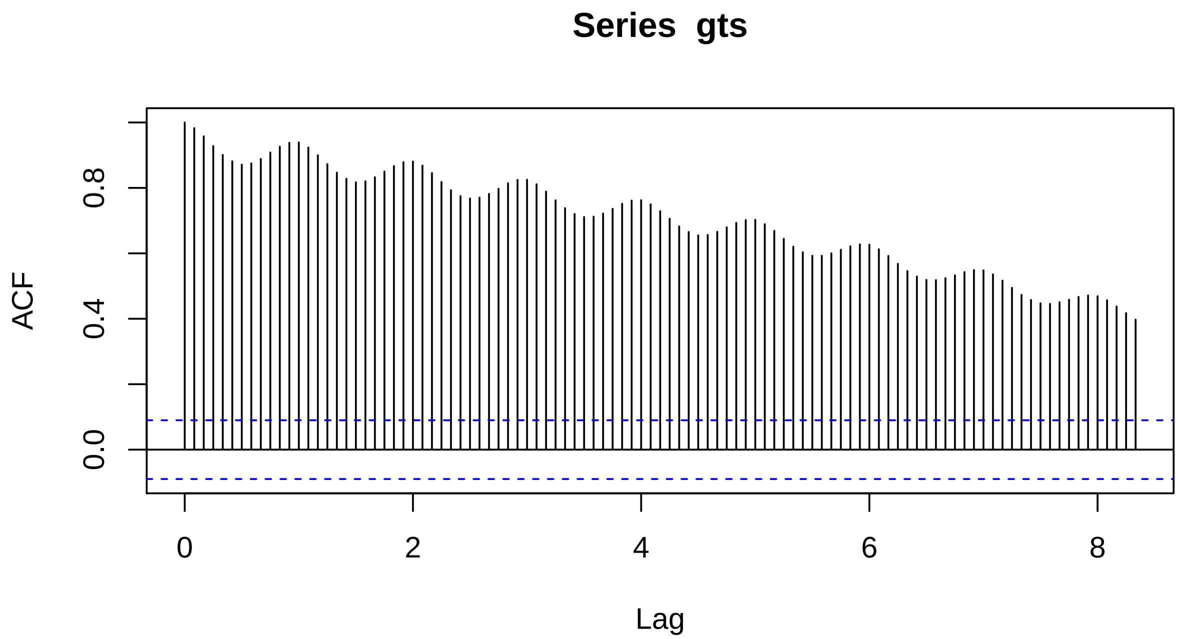
The auto corrrelation of order 0 will be 1 as all the values correlate to itself which will show a full or complete correlation. But we will also do correlations of different degrees or orders. The 1st order auto correlation will have the correlation between the original values with lag 1 values (shifting the values to the next corresponding place, like the first value moves to the second and so forth). There can be as many lags.

Let us look as the auto corrrelation for the data with lag upto 50.

```
# Auto correlation on the original data
acf(gts, lag.max = 50)
```




```
# Trying with a lag 100
acf(gts, lag.max = 100)
```



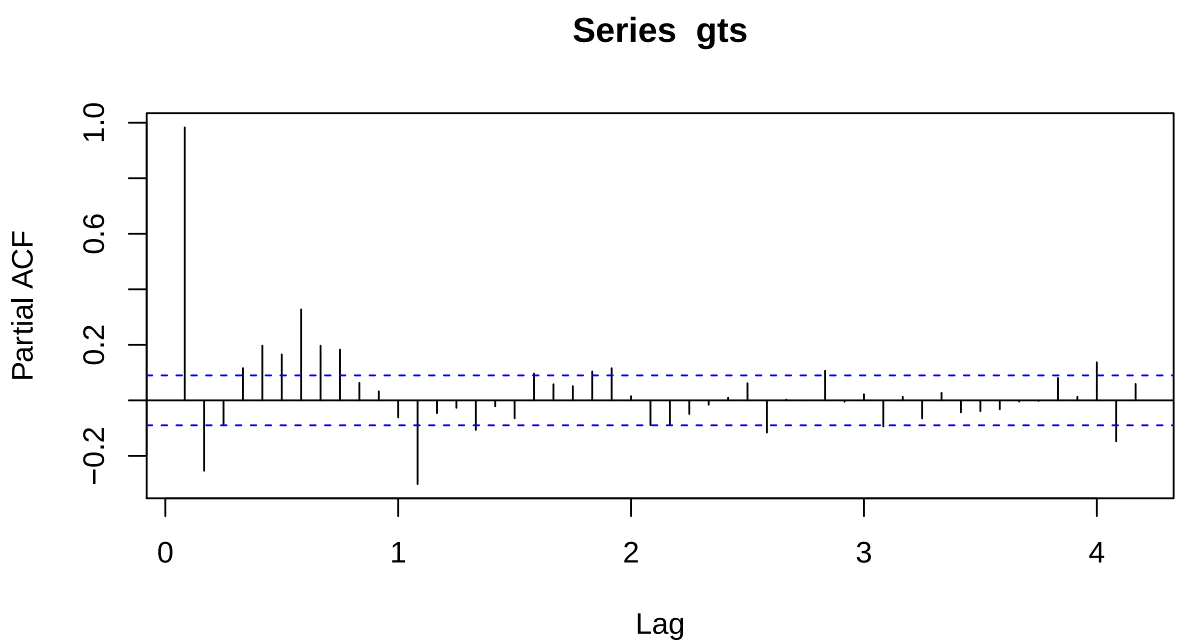

```
# The auto correlation of the original with lag 0 is always 100% or 1
```

The significance of the auto correlation is not much if the values are within the blue dotted bands. When they are outside of the bands, we can say that there is a dependency of the data on these auto-correlations. The original data is dependent on the so and so lag series. We find that none of these auto correlations lie within the blue dotted bands, hence we can say that all these of these are significant and remain close to 1 over many lag periods. Significant auto correlations imply that the observations of long past influences current observation. This also indicates that **the original series is non stationary**

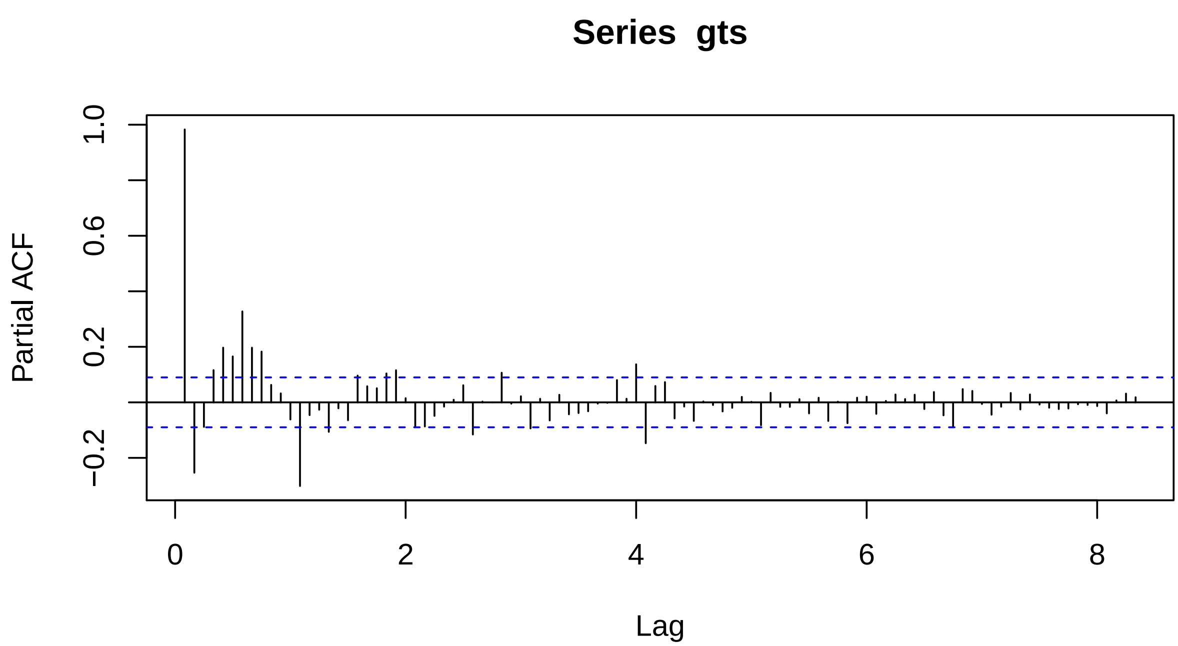
Partial auto correlation and auto correlation are actually the same, except for the fact that partial auto correlation excludes the effect of the intervening periods or lags while correlating.

For example, *PACF(1) = ACF(1)* as the correlation between original and lag1 will be the same for both, and there is no intervening periods in between. But PACF(2) is the correlation between the **original and lag2 series** after the effect or influence of lag1 series is eliminated. The same goes on for PACF(50) where the influence of lag1 upto lag49 is eliminated for the correlation between original series and the lag50 series. This is ideally the only difference between them.

```
# Partail Auto correlation on original data

pacf(gts, lag.max = 50)
```




```r
pacf(gts, lag.max = 100)
```




```r
# A mix of significant and insignificant correlations found
```

We see that upto lag 49, there is a mix of observations or correlations being significant and the vice. But beyond lag 50, we see that all of the correlations lie within the blue dotted region proving insignificance.There is a mix of both positive and negative correlation.

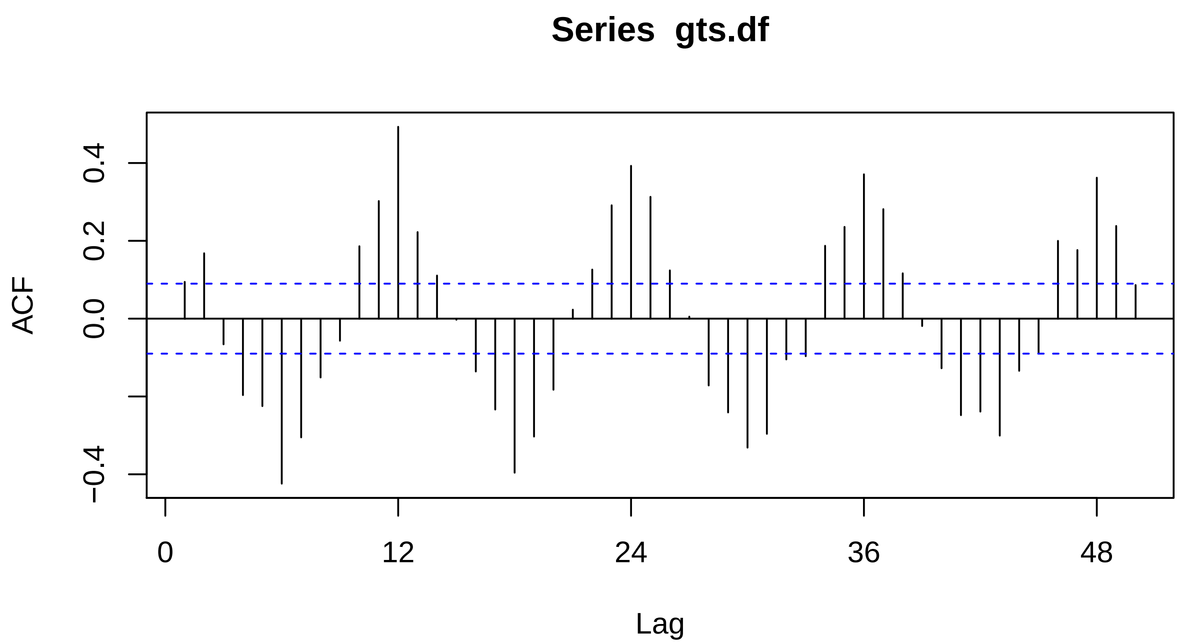
We see that the partial auto correlation of the original with lag 1 is close to 1, also when excluding influence of lag 1 for correlation between original and lag 2, we see that it is still significant though being negative. But when it comes to correlation between original and lag 3, the significance is not there without excluding the influence of lag 1 and lag 2. Likewise, there are certain correlations that even upon excluding the effect of the intervening period, remain significant.

This may tell that for a regression model, the response (current value) depends not only on the immediate previous value, as there are a few consecutive significant correlations and the data throughout the previous years maybe necessary for prediction.

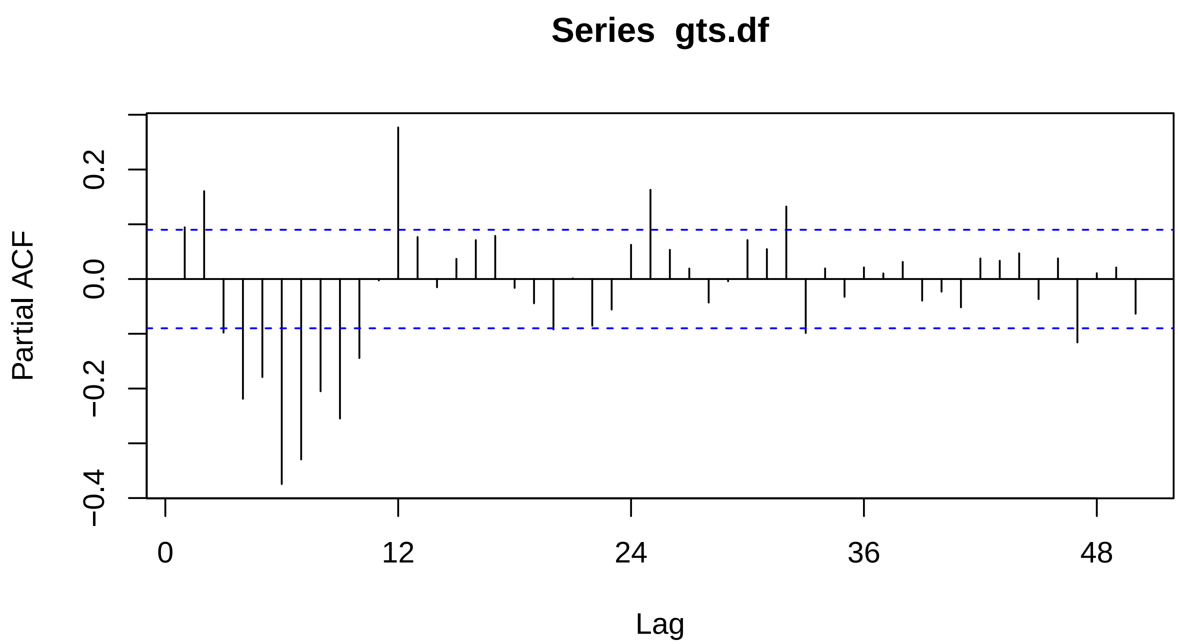
ACF and PACF on Differenced Series:

From the earlier acf and pacf we have found that all the correlations have given a value nearer to 1. This proves non stationarity of the series. Hence we are conducting the acf and pacf on the **differenced series**.

```
# ACF and PACF on differenced series
Acf(gts.df, lag.max = 50)
```




```
Pacf(gts.df, lag.max = 50)
```



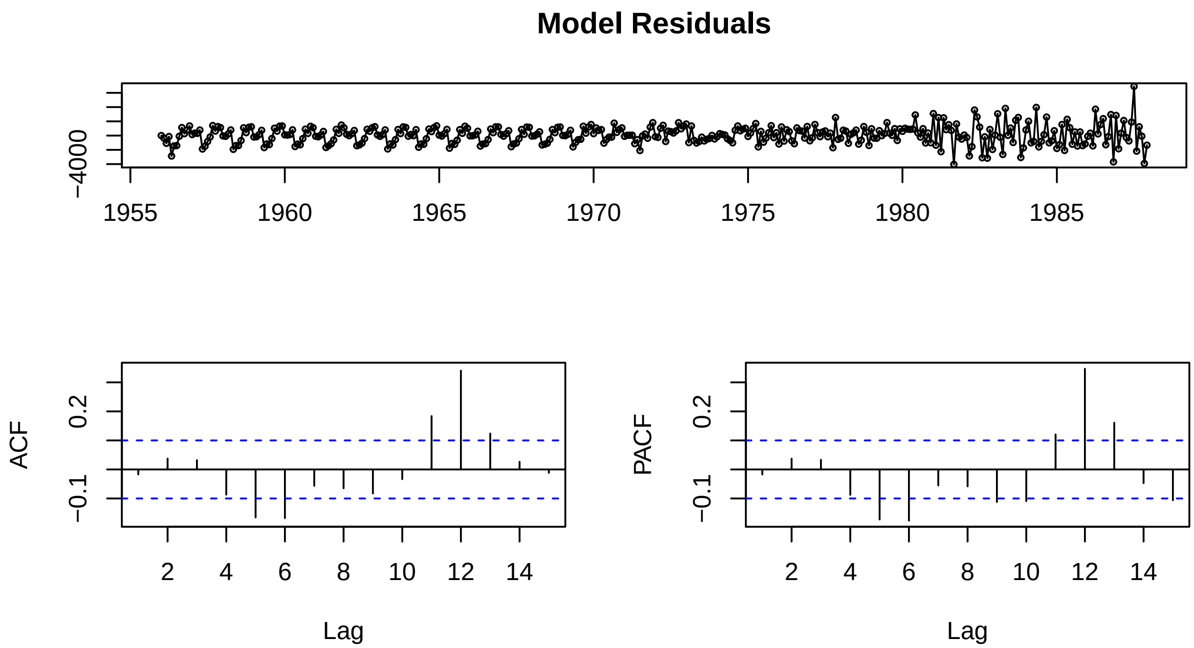

```r
# ACF cuts off after lag 1, so q=1
# PACF cuts off after 10. p=10
```

ARIMA Model:

```r
# Split data to train and test

gtstrain = window(DeseasonGas1, start = 1956, end = c(1987,12))
gtstest = window(DeseasonGas1, start = 1988, end = c(1995,8))

# Conducting the ARIMA model:

gtsARIMA = arima(gtstrain, order = c(2,1,10))
gtsARIMA
```

```
##
## Call:
## arima(x = gtstrain, order = c(2, 1, 10))
##
## Coefficients:
##           ar1     ar2     ma1      ma2      ma3      ma4     ma5      ma6
##       -0.5184  0.4462  0.5028  -0.6317  -0.3773  -0.1148  0.1792  -0.1734
## s.e.   0.0815  0.0827  0.0778   0.0760   0.0636   0.0748  0.0697   0.0591
##           ma7      ma8     ma9    ma10
##       -0.3681  -0.1092  0.5057  0.5688
## s.e.   0.0560   0.0730  0.0540  0.0591
##
## sigma^2 estimated as 1651318:  log likelihood = -3288.75,  aic = 6603.5
```

```r
tsdisplay(residuals(gtsARIMA), lag.max = 15, main = "Model Residuals")
```

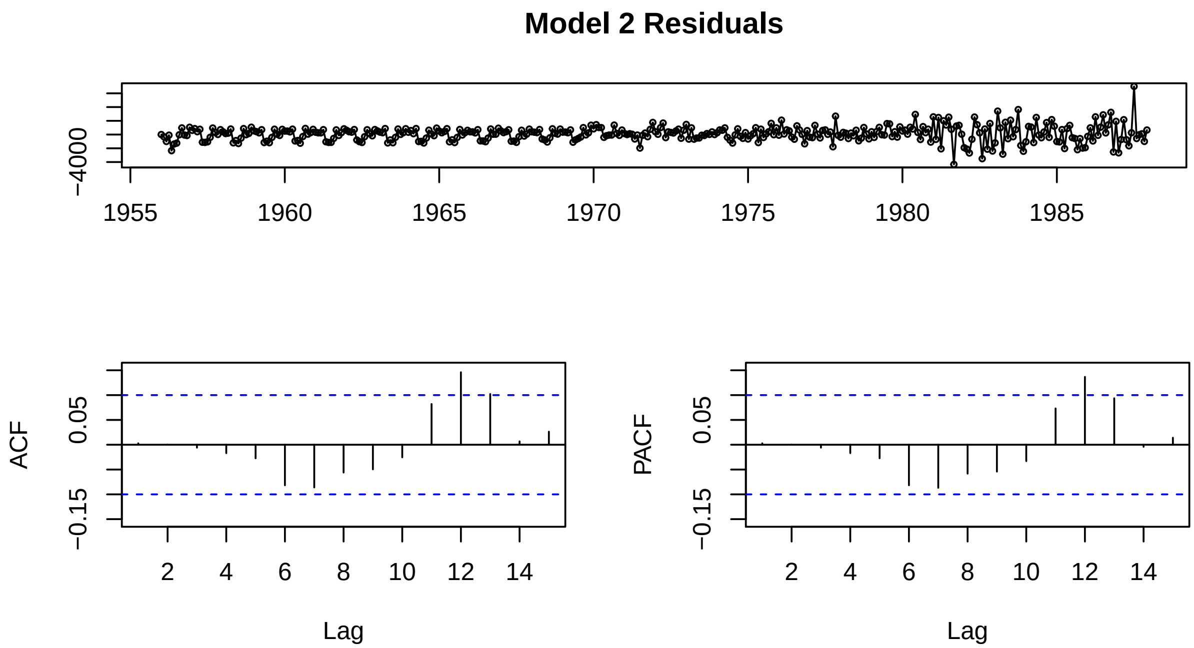


From this residual plot, we find that there is some amount of seasonality present in the plot. This is also evident from the ACF and PACF plots, which show some significant correlation present. Thus, this may not be the best model to predict on. Hence, we will have to build a better model

```
# ARIMA model 2

gtsARIMA2 = arima(gtstrain, order = c(2,1,20))
gtsARIMA2
```

```
##
## Call:
## arima(x = gtstrain, order = c(2, 1, 20))
##
## Coefficients:
##           ar1     ar2     ma1      ma2      ma3      ma4      ma5      ma6
##       -0.3403  0.5709  0.1691  -0.7147  -0.1472  -0.1801  -0.0723  -0.1573
## s.e.   0.1305  0.1480  0.1397   0.1493   0.0752   0.0825   0.0847   0.0800
##          ma7     ma8     ma9    ma10    ma11    ma12     ma13     ma14
##       0.1312  0.1625  0.3026  0.3091  0.1204  0.3545  -0.2961  -0.5267
## s.e.  0.0854  0.0792  0.0734  0.0984  0.0942  0.0814   0.0977   0.0838
##          ma15    ma16    ma17     ma18    ma19    ma20
##       -0.1945  0.0324  0.0876  -0.0806  0.1538  0.2724
## s.e.   0.0947  0.0858  0.0803   0.0847  0.1007  0.0744
##
## sigma^2 estimated as 1266436:  log likelihood = -3241.06,  aic = 6528.12
```

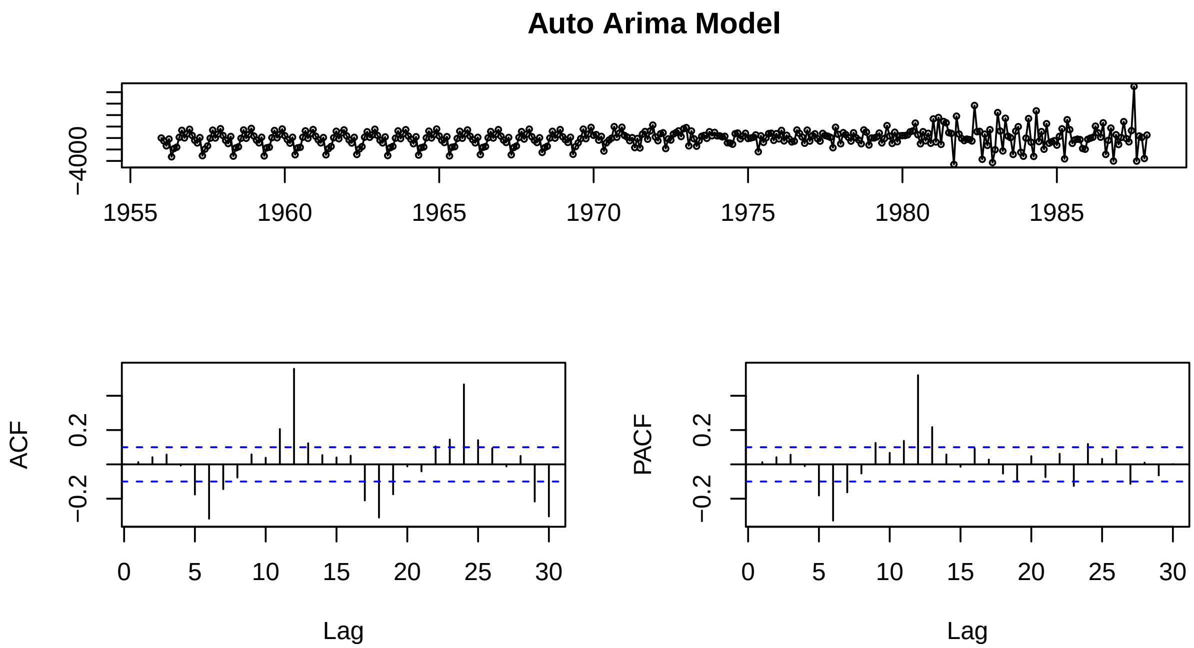
```
tsdisplay(residuals(gtsARIMA2), lag.max = 15, main = "Model 2 Residuals")
```



This model is much better than the previous one. And we see that there is a pattern in the beginning of the graph but transforms to a pattern less graph after 1970. Even the ACF and PACF has done better.

Fitting with Auto Arima

```
# let us use auto.arima

#auto.arima(gtstrain, ic = "aic", trace = TRUE)

fit = auto.arima(gtstrain, seasonal = FALSE)
tsdisplay(residuals(fit), lag.max = 30, main = "Auto Arima Model")
```




```r
# We see that there this is not a proper model
# Not a random pattern in the plot
# The acf and pacf plots identify some correlations
```

Diagnosis by Ljung box test:

- H0 - Residuals are independent
- Ha - Residuals are not independent

```r
# Diagnosis by Ljung box test:

Box.test(gtsARIMA$residuals)
```

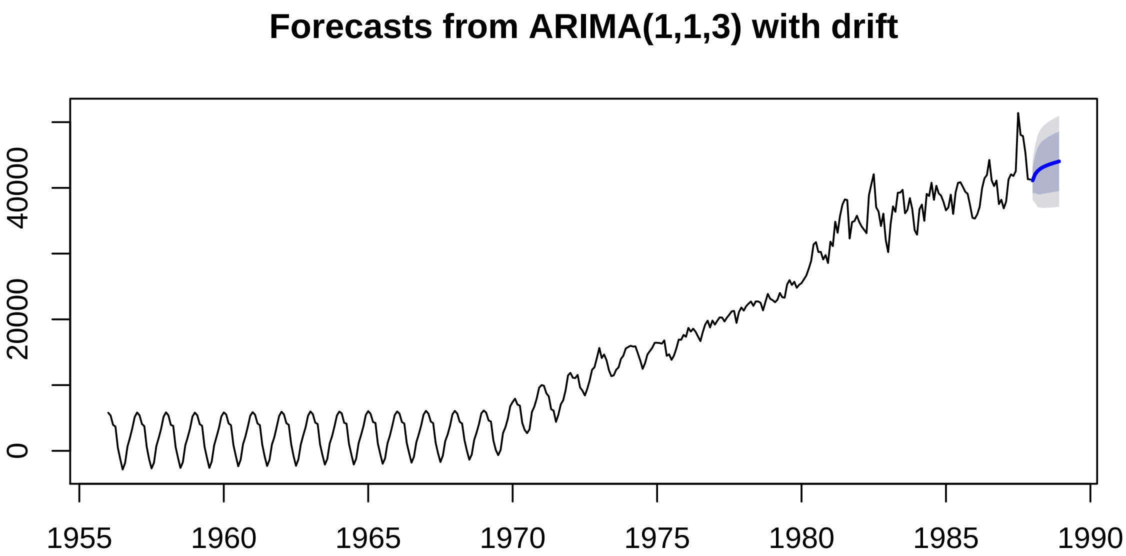
```
##
##  Box-Pierce test
##
## data:  gtsARIMA$residuals
## X-squared = 0.11342, df = 1, p-value = 0.7363
```

Since we have p value more than 0.05, we have not rejected the null hypothesis and confirm the **residuals are independent**

Forecast with ARIMA Model:

We are asked to forecast for the next 12 periods. Let us do that

```r
fcast12 = forecast(fit, h=12)

plot(fcast12)
```




```r
# This has not captured any seasonality as we have deseasonalised the series

# Accuracy:

accuracy(forecast(fit), gtstest)
```

```
##                        ME     RMSE      MAE        MPE      MAPE      MASE
## Training set  -15.61473 1506.906 1060.623 -26.661342 64.562973 0.7156339
## Test set     2547.19396 5036.507 3629.499   4.648702  7.254409 2.4489305
##                    ACF1 Theil's U
## Training set 0.01302978        NA
## Test set     0.70123694   1.45903
```

```r
# Though it is pretty less for the test, for the train, it is not a good model as it shows an error of
```

Report:

We find that the ARIMA model was okay for the test data set by considering the mape value. But the model still is not a great one. We have found that the auto arima has not performed well or given a good result