Project -6

Time Series Forecasting

By Saleesh Satheeshchandran
PGP-BABI 2019-20 (G-6)

Objective

This project is intended to analyse timeseries data called Australian monthly gas production. The following are the parts of this analysis.

- Read the data as a time series object in R. Plot the data.
- What do you observe? Which components of the time series are present in this dataset?
- What is the periodicity of dataset?
- Is the time series Stationary? Inspect visually as well as conduct an ADF test? Write down the null and alternate hypothesis for the stationarity test? De-seasonalise the series if seasonality is present?
- Develop an ARIMA Model to forecast for next 12 periods. Use both manual and auto.arima (Show & explain all the steps).
- Report the accuracy of the model.

Assumptions

There are no particular assumptions.

Tool used for the analysis

RStudio Version 1.2.1335
R Version 3.6.0

Input Data

The data is available in a data frame within the R-package “forecast”

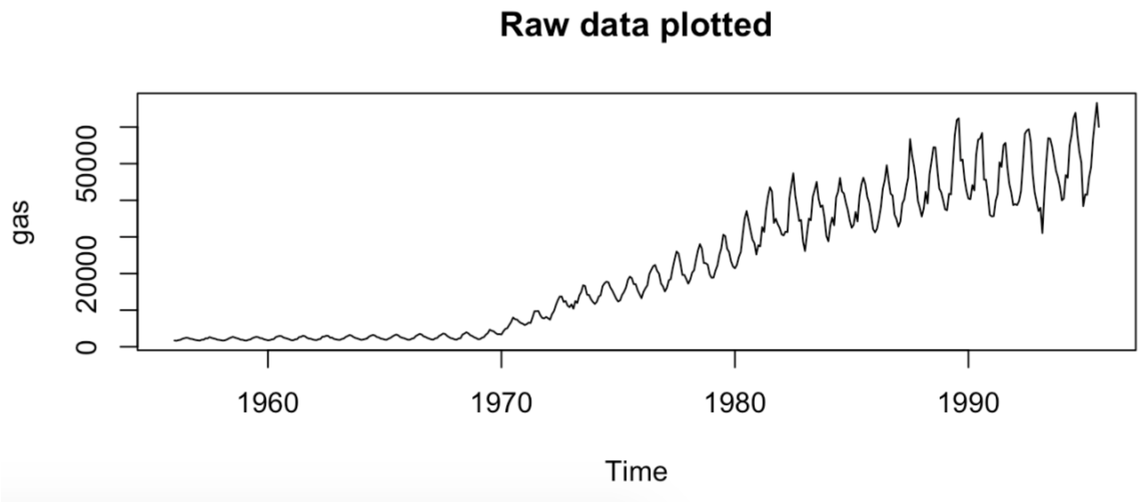
Loaded Libraries

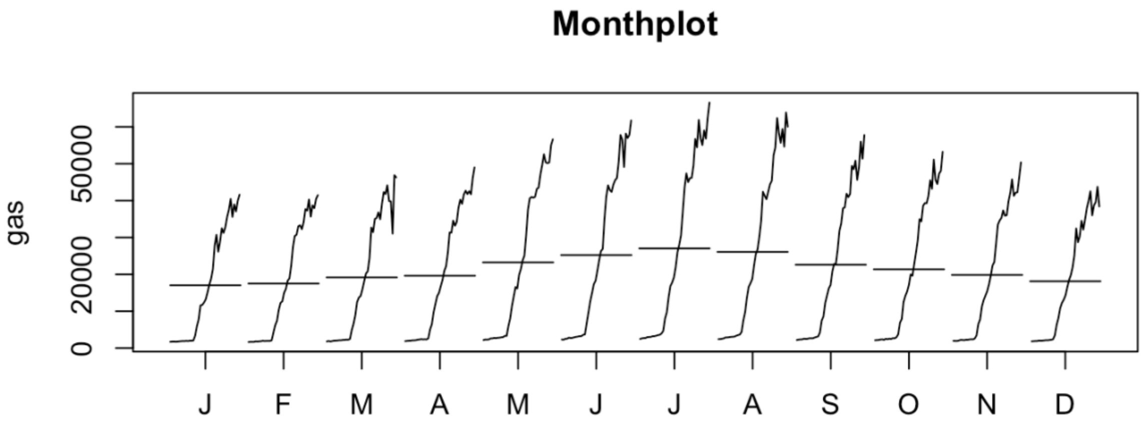
```
library(forecast)
library(fpp2)
library(tseries)
library(MLmetrics)
library(ggplot2)
library(stats)

```

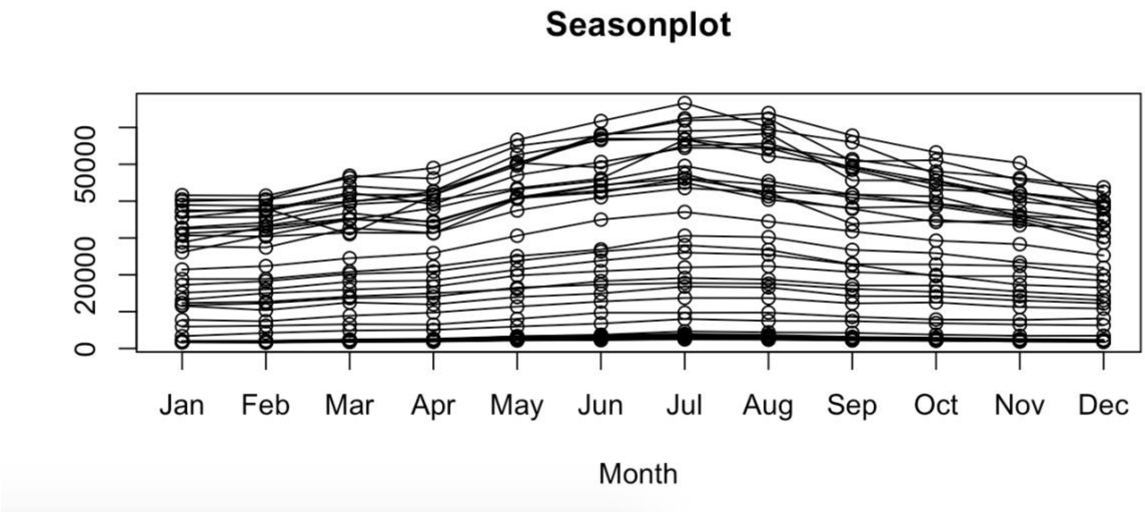
Exploratory Data Analysis

- The data given has single variable.
- There are 476 observations
- The starting year is 1956 and ends in 1996
- The observations are taken monthly
-

```
> head(gas)
      Jan  Feb  Mar  Apr  May  Jun
1956 1709 1646 1794 1878 2173 2321
> str(gas)
 Time-Series [1:476] from 1956 to 1996: 1709 1646 1794 1878 2173 ...
> frequency(gas)
[1] 12
```







Cleaning the data and making it time series.

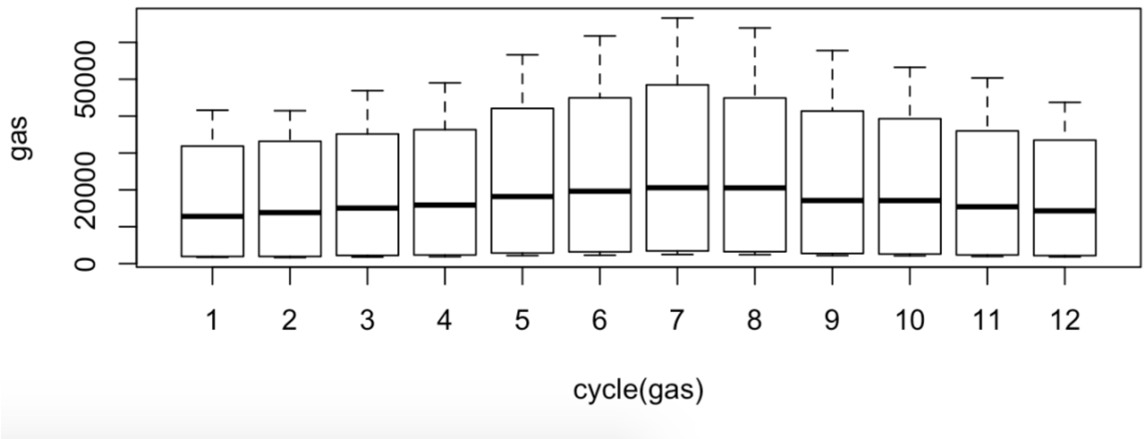
```
#Cleaning the data
TSgasdata=tsclean(gas)

#Saving as Time series data
TSGas <-ts(gas, frequency=12, start=c(1956,1))
summary(TSGas)
```

```
> summary(TSGas)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1646    2675   16788   21415   38628   66600
>
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1956	1709	1646	1794	1878	2173	2321	2468	2416	2184	2121	1962	1825
1957	1751	1688	1920	1941	2311	2279	2638	2448	2279	2163	1941	1878
1958	1773	1688	1783	1984	2290	2511	2712	2522	2342	2195	1931	1910
1959	1730	1688	1899	1994	2342	2553	2712	2627	2363	2311	2026	1910
1960	1762	1815	2005	2089	2617	2828	2965	2891	2532	2363	2216	2026
1961	1804	1773	2015	2089	2627	2712	3007	2880	2490	2237	2205	1984
1962	1868	1815	2047	2142	2743	2775	3028	2965	2501	2501	2131	2015
1963	1910	1868	2121	2268	2690	2933	3218	3028	2659	2406	2258	2057
1964	1889	1984	2110	2311	2785	3039	3229	3070	2659	2543	2237	2142
1965	1962	1910	2216	2437	2817	3123	3345	3112	2659	2469	2332	2110
1966	1910	1941	2216	2342	2923	3229	3513	3355	2849	2680	2395	2205
1967	1994	1952	2290	2395	2965	3239	3608	3524	3018	2648	2363	2247
1968	1994	1941	2258	2332	3323	3608	3957	3672	3155	2933	2585	2384
1969	2057	2100	2458	2638	3292	3724	4652	4379	4231	3756	3429	3461
1970	3345	4220	4874	5064	5951	6774	7997	7523	7438	6879	6489	6288

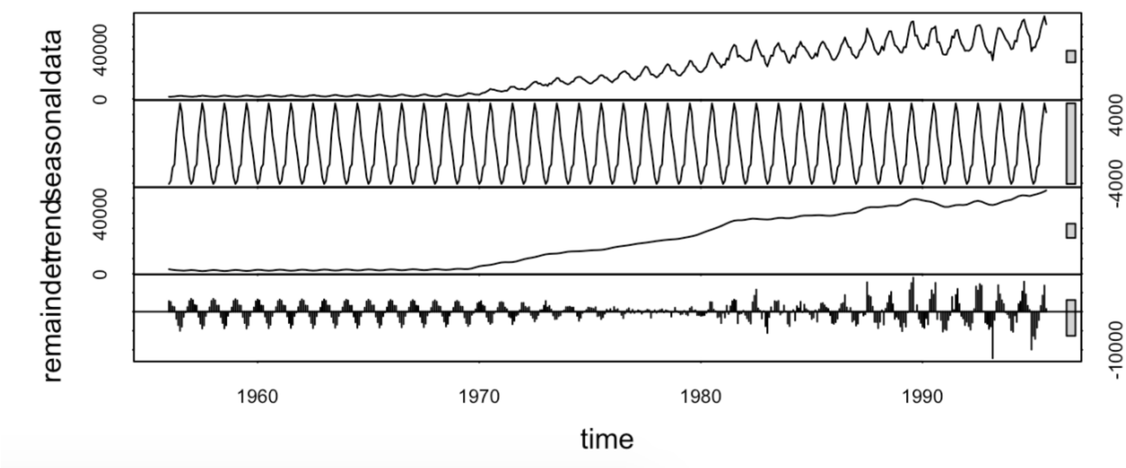
Checking for outliers

There are no outliers identified in Monthly boxplot.



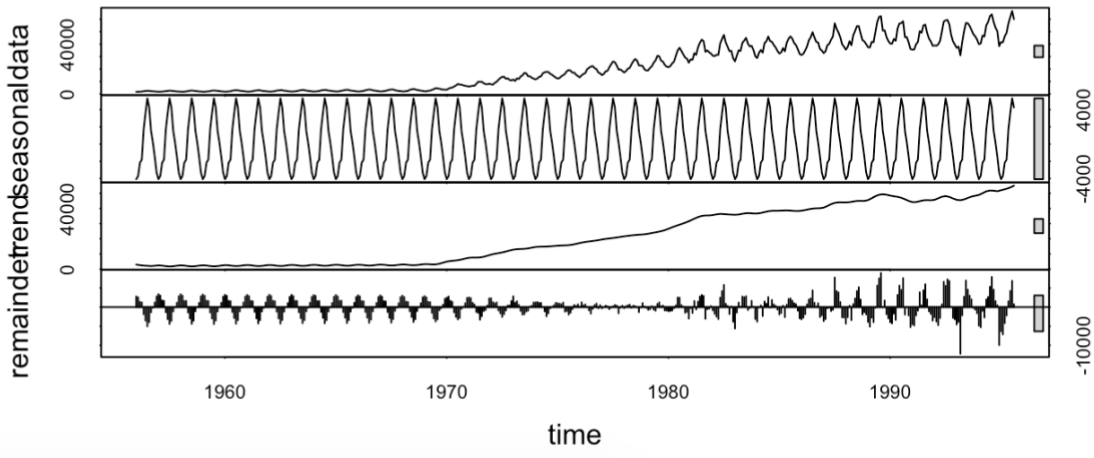
Decomposing the data

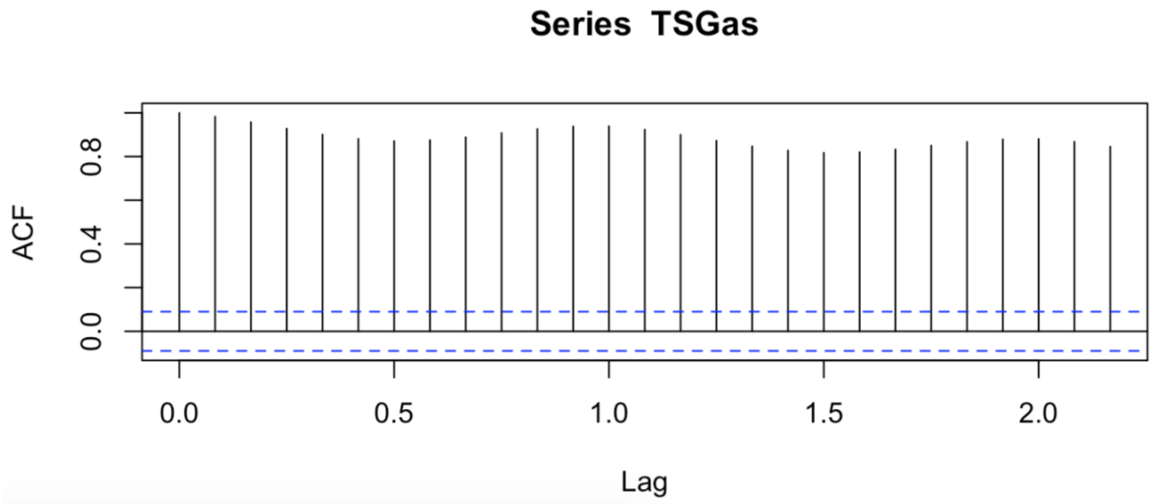
```
#Decompose the data
GasDec<-stl(TSGas, s.window='p')
plot(GasDec)
GasDec$time.series
```

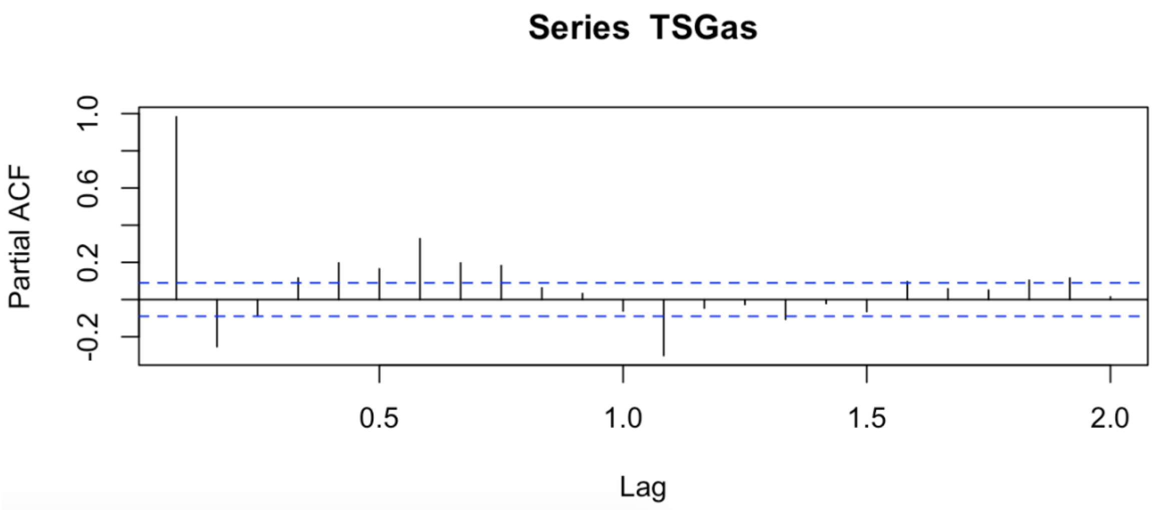


The decomposition of the data shows
- There is an upward trend starting from the year 1970.
- There is a clear seasonality
- The random factor is not a white noise.

Checking periodicity of the data.



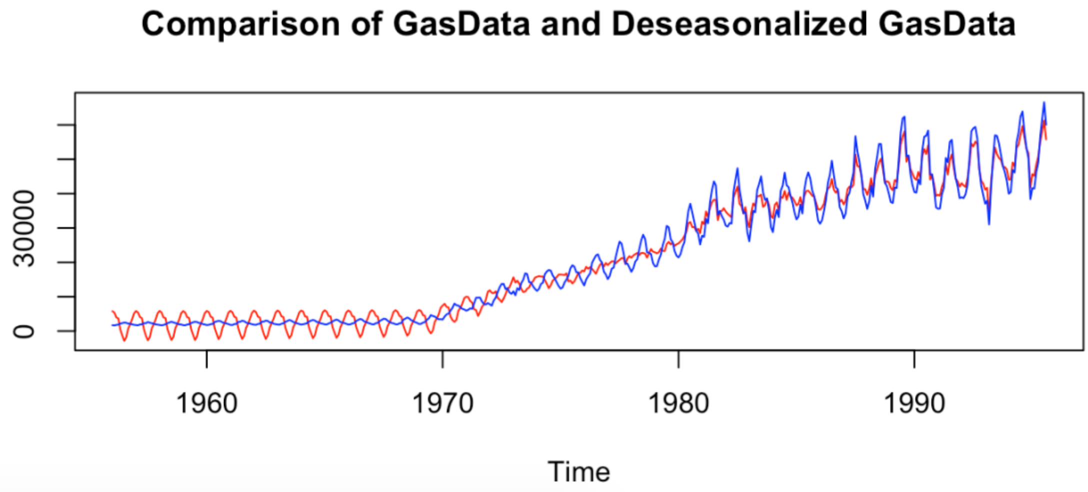




The periodicity is annual in nature

De-seasonalizing the data

```
#Deseasonalize the data
DeseasonGas <- (GasDec$time.series[,2]+GasDec$time.series[,3])
ts.plot(DeseasonGas, TSGas, col=c("red", "blue"), main="Comparison of GasData and Deseasonalized GasData")
deseasonal_gas=seasadj(GasDec)
```



Checking if the data is stationary

Null Hypothesis – The Data is not stationary
Alternate Hypothesis – The Data is stationary

```
        Augmented Dickey-Fuller Test

data:  TSGas
Dickey-Fuller = -2.7131, Lag order = 7, p-value = 0.2764
alternative hypothesis: stationary
```

The data is not stationary shown by high p value . In order to make it stationary differencing is done.

```
        Augmented Dickey-Fuller Test

data:  count_diff1
Dickey-Fuller = -18.14, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(count_diff1, alternative = "stationary") :
  p-value smaller than printed p-value
```

The data now is stationary

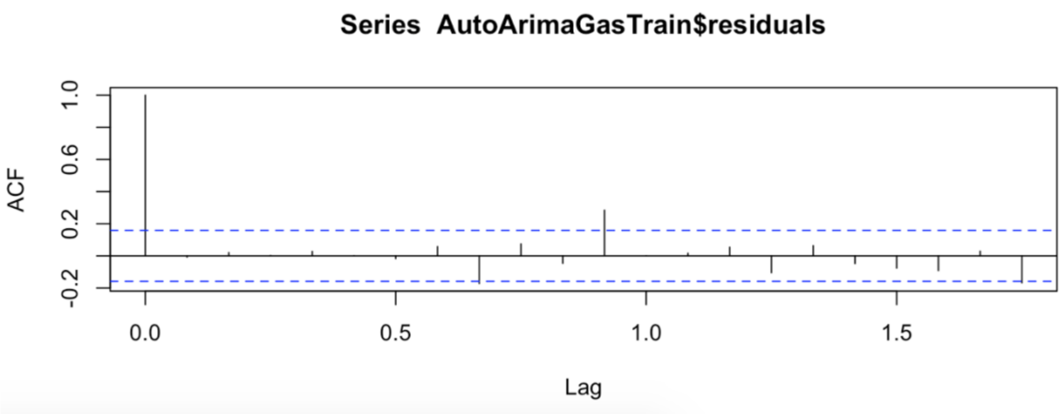
```
#Checking if the data is stationary
adf.test(TSGas, alternative = "stationary")

#Differencing the time series data
count_diff1 = diff(deseasonal_gas, differences = 1)
plot(count_diff1)
adf.test(count_diff1, alternative = "stationary")
```

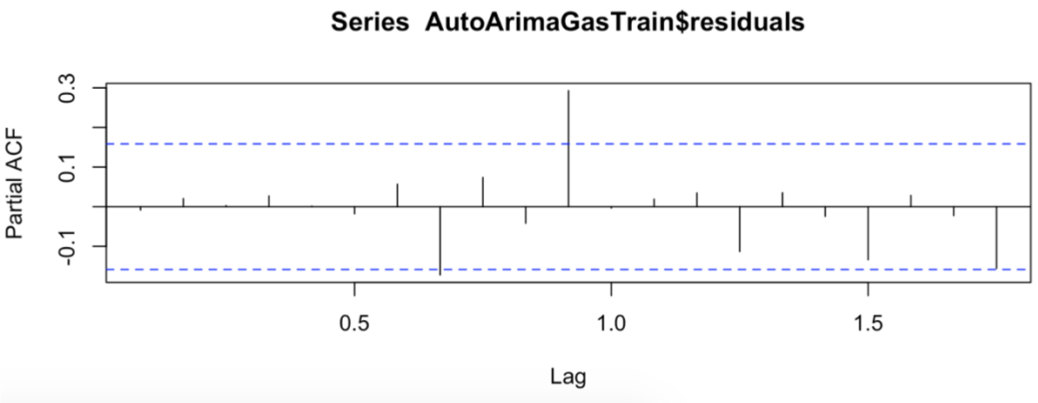
Splitting data to train and test

```
#Splitting into training and test sets

GasdataTrain <- window(count_diff1, start=c(1970,1), end=c(1982,9), frequency=12)
GasdataTest <- window(count_diff1, start=c(1982,10), frequency=12)
```

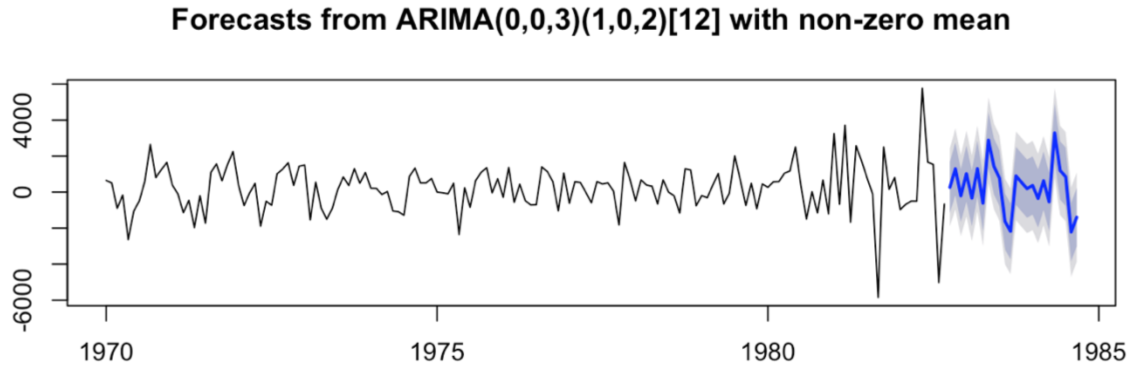
Auto ARIMA

```
> AutoArimaGasTrain=auto.arima(GasdataTrain,seasonal=TRUE)
> AutoArimaGasTrain
Series: GasdataTrain
ARIMA(0,0,3)(1,0,2)[12] with non-zero mean

Coefficients:
          ma1      ma2      ma3    sar1     sma1    sma2      mean
      -0.2808  -0.0092  -0.3611  0.7713  -0.4747  0.2229  228.8640
s.e.   0.0821   0.0771   0.0778  0.0961   0.1345  0.1215   82.4874

sigma^2 estimated as 1214780:  log likelihood=-1289.37
AIC=2594.75   AICc=2595.75   BIC=2618.99
> MAPE(AutoArimaGasTrain$fitted,AutoArimaGasTrain$x)
[1] 7.980251
```

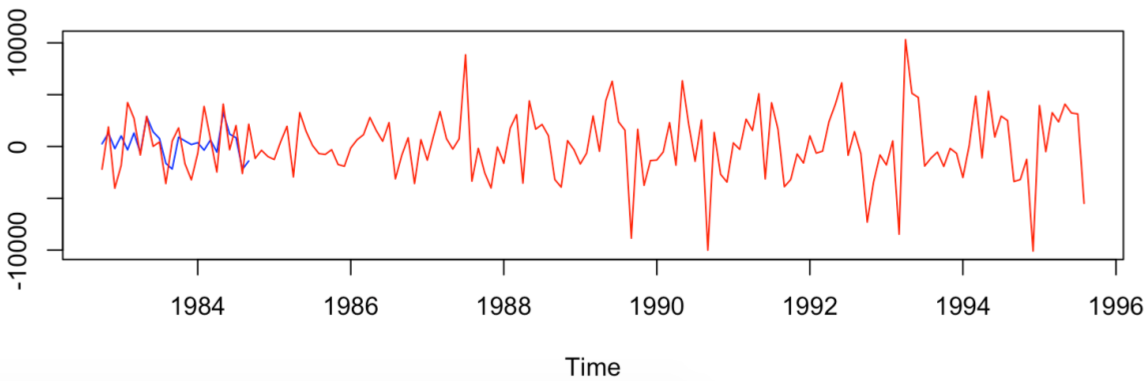




```
> acf(AutoArimaGasTrain$residuals)
> pacf(AutoArimaGasTrain$residuals)
> Box.test(AutoArimaGasTrain$residuals, lag = 30, type = "Ljung-Box")

        Box-Ljung test

data:  AutoArimaGasTrain$residuals
X-squared = 35.749, df = 30, p-value = 0.2164
```



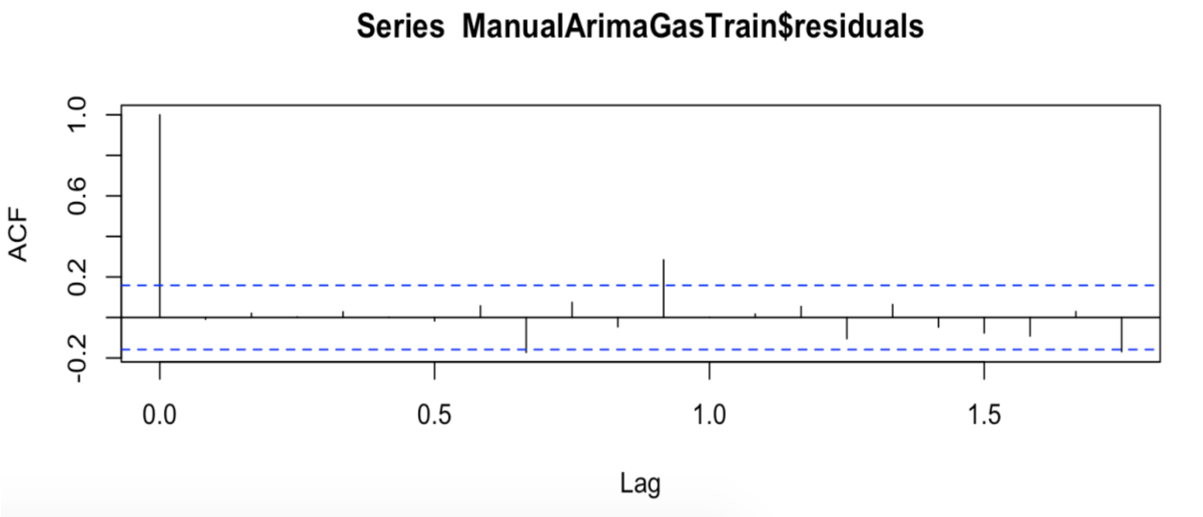

```
> GasAutoArimaForecast=forecast(AutoArimaGasTrain)
> plot(GasAutoArimaForecast)
> accuracy(GasAutoArimaForecast)
                    ME     RMSE      MAE       MPE     MAPE      MASE
Training set -7.995632 1076.662 787.9134 -483.9265 798.0251 0.7891952
                    ACF1
Training set -0.00831619
```

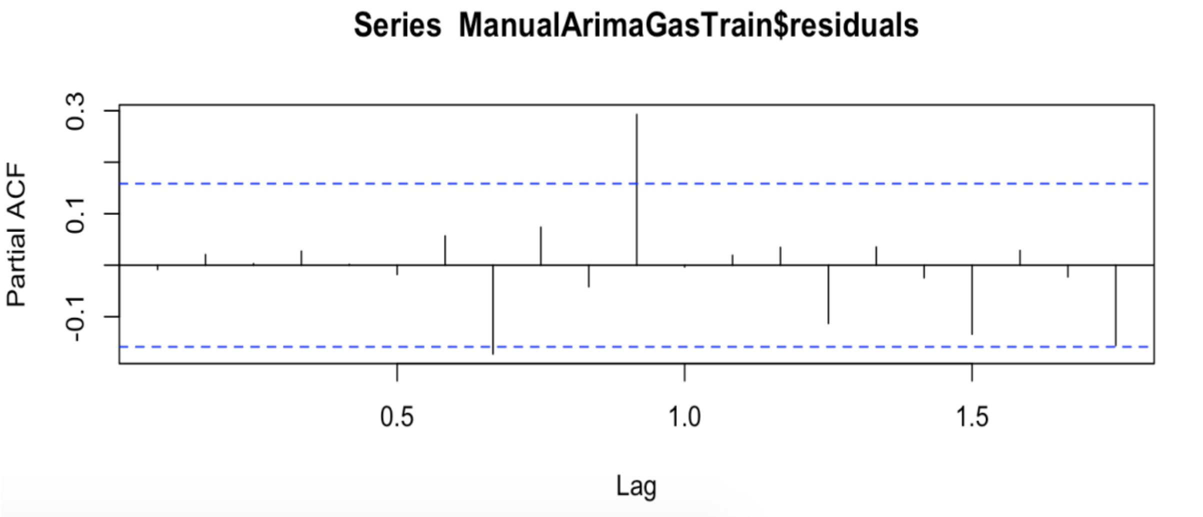



```
> vec.autoarima=cbind(GasAutoArimaForecast$mean,GasdataTest)
> ts.plot(vec.autoarima,col=c("blue", "red"))
```

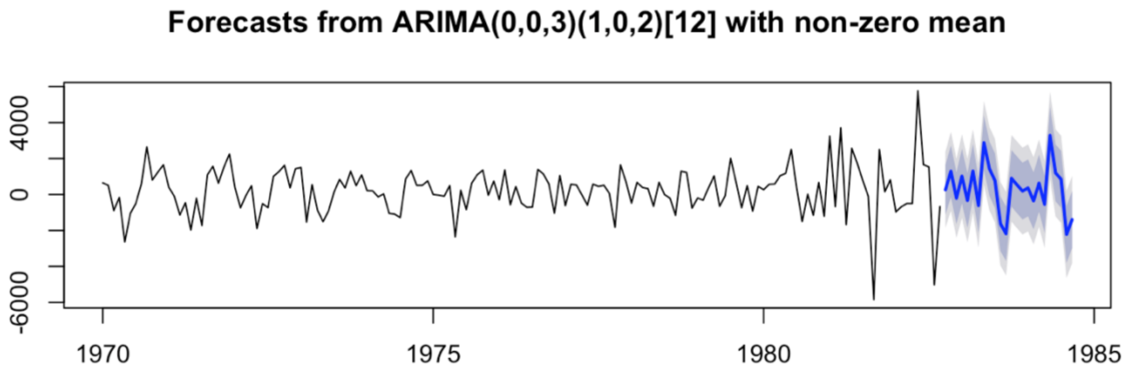
Auto ARIMA

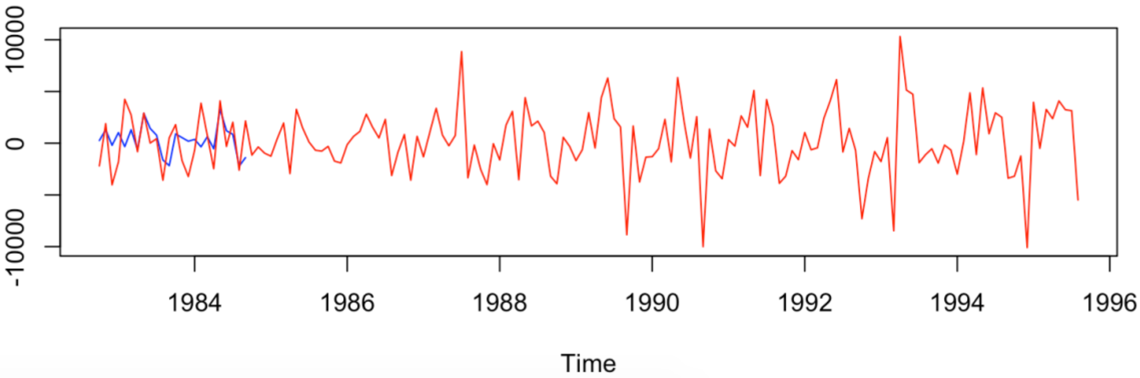
```
> ManualArimaGasTrain<-arima(GasdataTrain, order = c(0,0,3), season=list(order = c(1,
0,2), period=12))
> ManualArimaGasTrain

Call:
arima(x = GasdataTrain, order = c(0, 0, 3), seasonal = list(order = c(1, 0,
    2), period = 12))

Coefficients:
          ma1      ma2      ma3    sar1     sma1    sma2  intercept
      -0.2808  -0.0092  -0.3611  0.7713  -0.4747  0.2229   228.8640
s.e.   0.0821   0.0771   0.0778  0.0961   0.1345  0.1215    82.4874

sigma^2 estimated as 1159201:  log likelihood = -1289.37,  aic = 2594.75
>
```





```
        Box-Ljung test

data:  ManualArimaGasTrain$residuals
X-squared = 35.749, df = 30, p-value = 0.2164
```





Accuracy

```
> GasManualArimaForecast=forecast(ManualArimaGasTrain)
> plot(GasManualArimaForecast)
> accuracy(GasManualArimaForecast)
                    ME     RMSE      MAE       MPE     MAPE      MASE
Training set -7.995632 1076.662 787.9134 -483.9265 798.0251 0.7891952
                    ACF1
Training set -0.00831619
```

MAPE – 798.0251

Actionable Insights

From the various analysis and modelling done on this project, Auto ARIMA is identified to be the best model.
The model built is able to give a 12 year projection of the data.