Australian Monthly Gas Production

Table of contents

1 Project Objective

The primary objective of this report is to find the Australian monthly gas production .In given dataset , predictor is time and dependent variable is gas . Our aim is to forecast the next 12 periods in the gas productions

2 Exploratory data analysis

Generally, Exploratory data analysis are carried out to discover the patterns in the given dataset,missing values of the given dataset.

2.1 Environment set up and data Import

The working directory should be one , where the code and dataset are placed. The following R packages used for the analysis of Churn prediction dataset

- Ggplot2 package-To infer about the attributes of the loan prediction data using graphical plots
- Forecast-To import gas production dataset
- Tseries-To perform ARIMA modelling

2.2 Variable Identification:

Some basics R functions such as mean , sd ,round are used for the statistical calculation. Here are the following functions are getting used for the better understanding of data and to make further decision **.**

Structure(Str)- To get the structure of the employee transport mode such as class category,name and the count of the fields

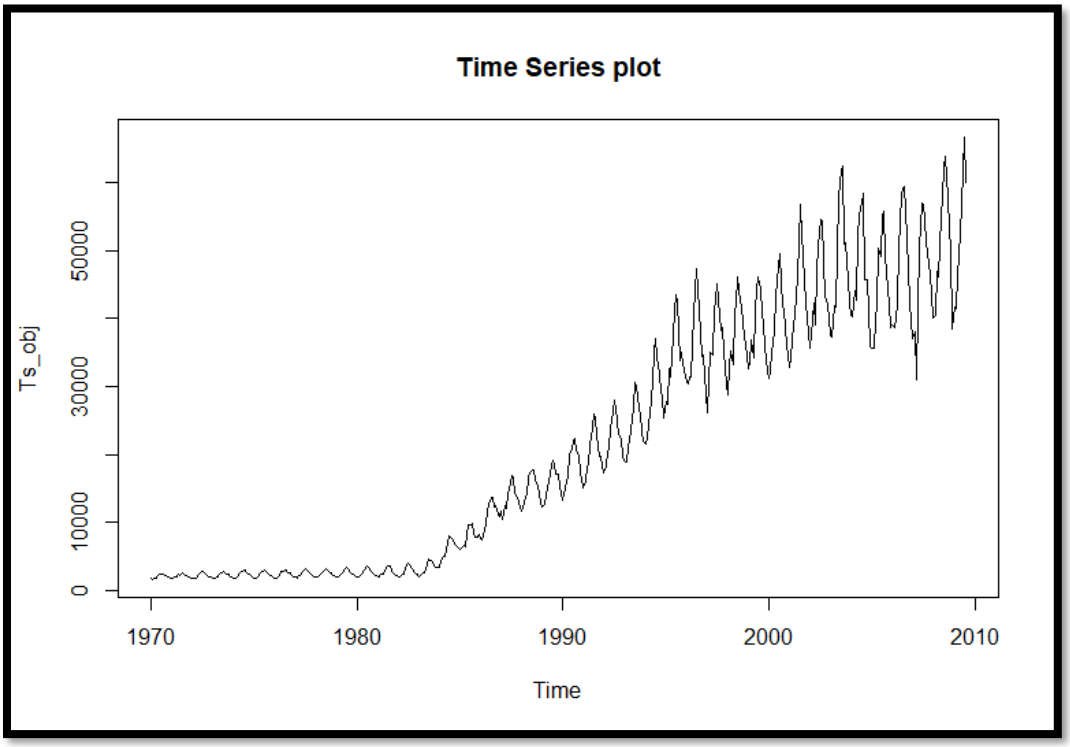
Summary –Generally , summary function will give the classification of attributes.If the field is categorical data, summary function give the count of each sub-category If the field is integer, then it will return the five stats notably minimum,25th percentile , median or 50th percentile , 75th percentile and maximum.

Summary statistics

```
> summary(Ts_obj)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1646    2675   16788   21415   38629   66600
> 
```

The summary of gas production dataset shows that it had max production of 66600 and min of 1646(in the earlier times of production).From the summary ,the data are in the upward trend.

Time series plot:



Features of Time series :

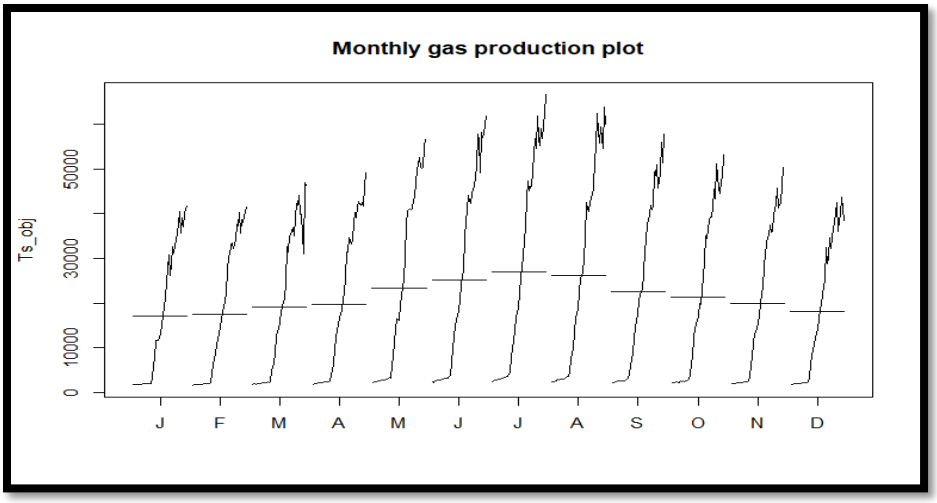
1)The given monthly gas production data are dependent and ordered.

2)There is no missing data

3)No outliers found

Data are dependent and ordered.As the time series plot showed that monthly gas production are stable between 1970 and 1985.After 1985 there was sudden increase in the production and become more significant till 2010 .Also noted that mean and variance are not constant throughout the gas production(due to increasing upward trend and partial seasonality) .

Monthly Plot:



The above plot gives the monthly gas production across all the months from 1970 -2010.Initially it seems , gradual increase in the production ,but july month of every year showed highest production when compared to rest of the months in most of the times .

3)Component of Time Series:

The component of time series involves

1) Trend- Long term increase or decrease in the given data

2) Seasonality-Short term stable fluctuations/variations within the year

3) Reminder- Unusual observations/Irregular components in the given data

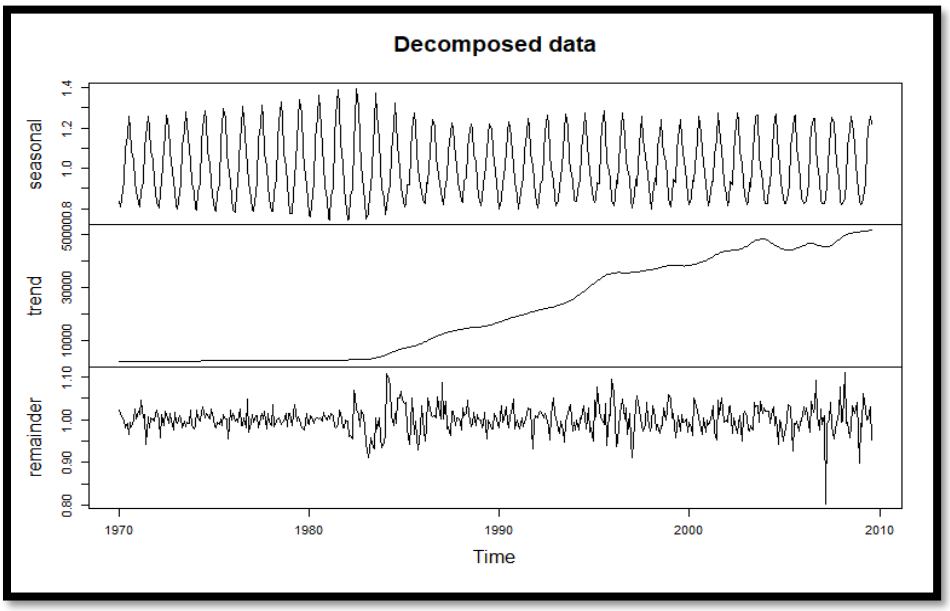
In the monthly gas production data, upward trend is significant and dominating the entire time series data. Generally decomposition model will describe the data appropriately.

Decomposition Model:

1) Additive Model

2) Multiplicative Model

Australian gas production shows high variance in the data(particularly from 1980 to 1990),so multiplicative model would be used to describe the gas production data. So, multiplicative model is the additive decomposition of logarithms.



Periodicity of Data:

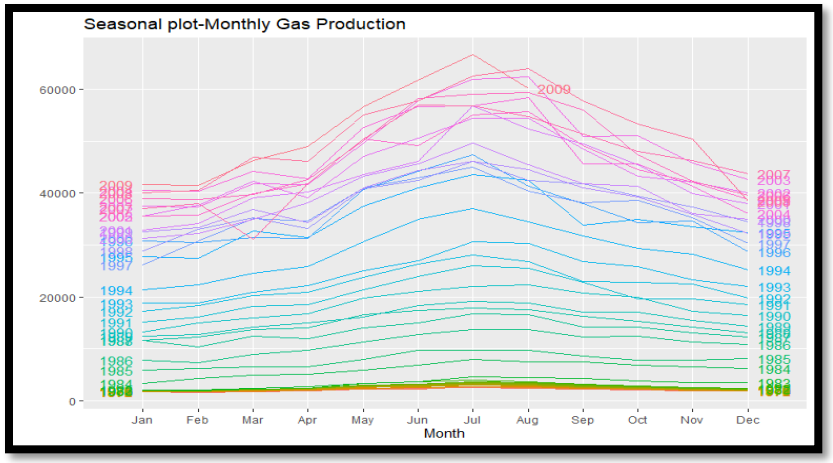
A fundamental characteristic of time series data is how frequently the observations are spaced in time. How often the observations of a time series occur is called the sampling frequency or the periodicity of the series. Time series with one observation each month has a monthly sampling frequency or monthly periodicity and so is called a monthly time series.

Periodicity/frequency of dataset=12

4)De-seasonalised data:

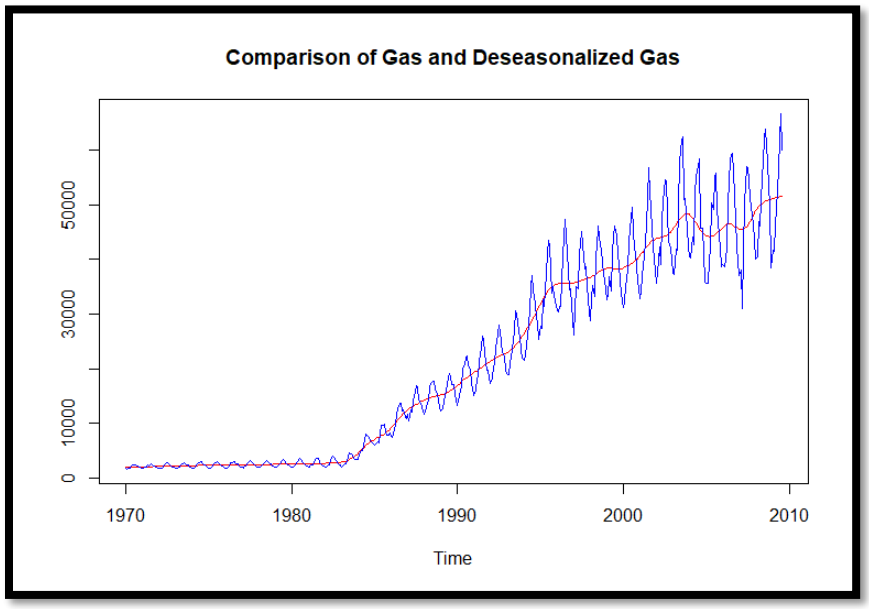
Seasonal Plot:

A seasonal plot is similar to a time plot except that the data are plotted against the individual "seasons" in which the data were observed. A seasonal plot allows the underlying seasonal pattern to be seen more clearly, and is especially useful in identifying years in which the pattern changes. In the seasonal plot, we observed that there is an partial seasonality(semi-annual seasonal) in the Australian gas production .



De-seasonalization removes the seasonality in the data and help us to understand the impact of other components of times series. As the data shows the high variance in the seasonality and it also proportional to the level of time series, multiplicative decomposition is more appropriate. From the comparison of given gas data and de-seasonalized gas ,it is concluded that Seasonal components also seems to be significant.

```
#Multiplicative model
Log_Ts_obj=log(Ts_obj)
Log_Ts_obj_dec=stl(Log_Ts_obj, s.window=7)
Decompose_obj=exp(Log_Ts_obj_dec$time.series)
plot(Decompose_obj,main="Decomposed data")
#de-seasonalise
deseason_obj=Decompose_obj[,2]+Decompose_obj[,3]
ts.plot(deseason_obj, Ts_obj, col=c("red", "blue"), main="Comparison of Gas and Deseasonalized Gas
```



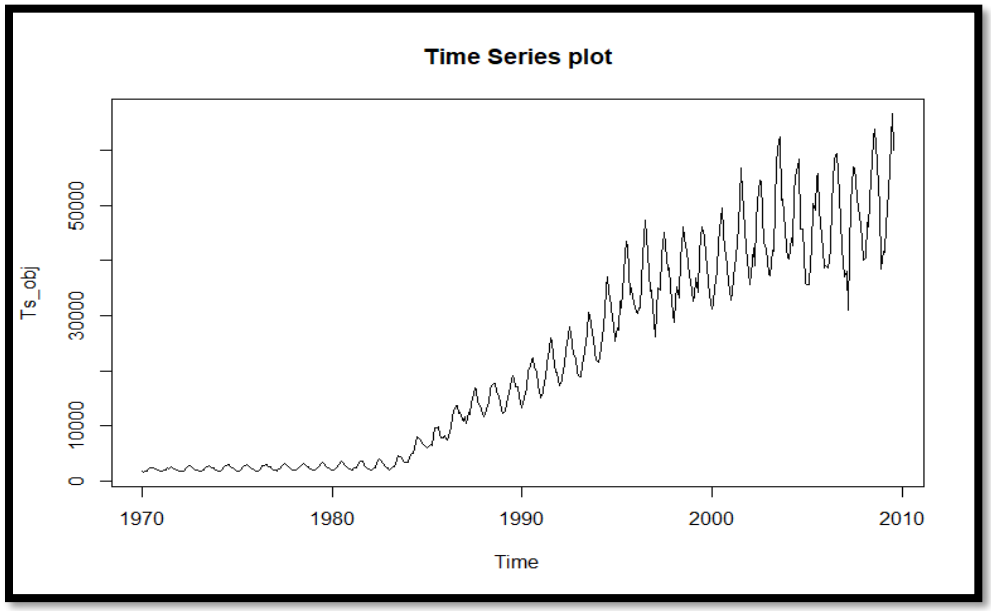
5)Stationary time series

Stationary time series is the one which neither have trend nor seasonality .It is also known as white noise. There are three conditions should be satisfied for being time series as stationary.

1) Mean of the time series should be constant.

2) Variance of time series should be constant.

3) The correlation between the t-th term in the series and the t+m-th term in the series is constant for all time periods and for all m



From the plot itself , we can easily say the Australian gas production dataset doesn't have constant mean and constant variance.

5.1)ADF(Augmented Dickey Fuller) Test:

ADF test is used to check whether the given time series data is non-stationary or not.To identify its status in R,the package named "tseries" is used.

Hypothesis Testing:

Null Hypothesis,H0:Time series is Non-Stationary

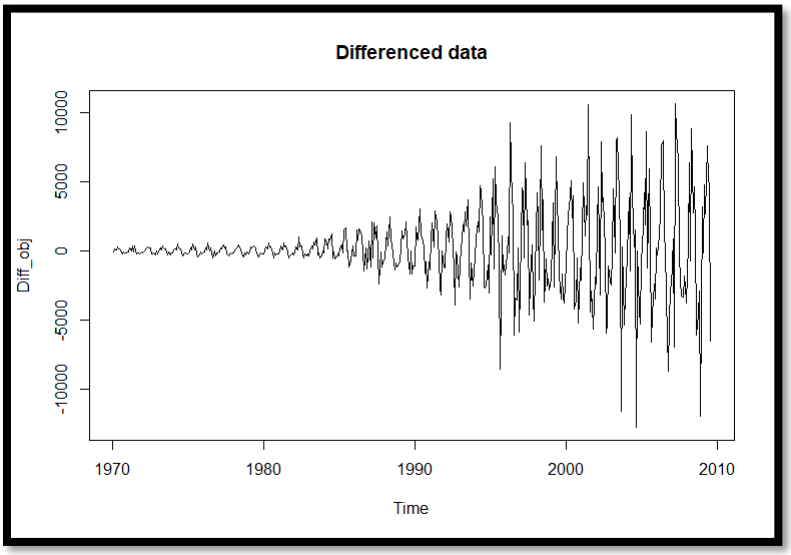
Alternate Hypothesis , H1:Time is Stationary

```
> ########################################
> library("tseries")
> adf.test(Ts_obj)

        Augmented Dickey-Fuller Test

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

From the ADF test, it is concluded that given gas production dataset is non-stationary series. So we are failed to reject Null hypothesis.

5.2)Differencing:

To make the gas production series to be stationary, we need to difference the time series data until the data become stationary.



The above plot is the first order differenced data. After differencing the time series data for the first time , again perform the adf test to test the stationarity of the series.

```
> #NUll Hypothesis:Time series Non-stationary
> #Alternate Hypothesis:Time series stationary
> adf.test(Diff_obj)#now it is stationary

        Augmented Dickey-Fuller Test

data:  Diff_obj
Dickey-Fuller = -19.321, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

After performing the adf test , differenced data shows that given time series is stationary.Here the p-value is 0.01 which less than 0.05, so we can reject null hypothesis and accept alternate hypothesis(ie given data is stationary) . Split the data into training and testing using window function.

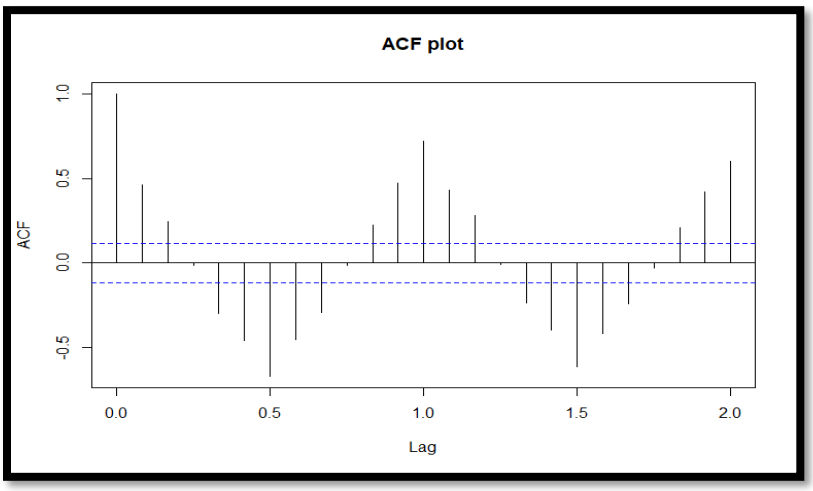
```
########################################
library("tseries")
#Data Split
Train_data=window(Ts_obj,start=c(1970,1),end=c(1993,12),frequency=12)
Test_data=window(Ts_obj,start=c(1994,1),frequency=12)
```

6)Forecating:

ARIMA(Auto Regression Integrated Moving Average)

To fit arima model, the given time series should be stationary. So, here the given gas production data is stationary after differencing. The parameters of arima model are p,d,q, to determine these three parameters , auto-correlations functions are used.

Auto-regression is value of a time series depends on its value at the previous time point .The order of the auto regression is determined by the ACF and PACF plot.

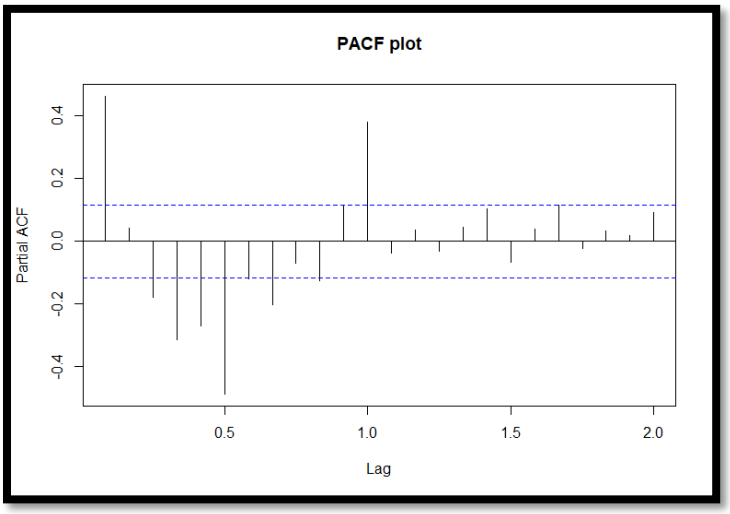


The auto correlation function is calculated by lag series. For the given gas production data , lag was taken as 50.The acf is significant only if it outside the blue line .In the above plot ,some points are positively correlated and negatively correlated. ACF considers all these components while finding correlations hence it's a 'complete auto-correlation plot'.

Disadvantage of ACF:

The future period of gas production depends mostly on the recent data points(ie previous month).The auto-correlation function concluded that data which are too previous are more significant but in the real time scenario, the gas production only depends on previous months(recent past observations)

PACF(Partial auto correlation function) adjust for the intervening periods. Basically instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence 'partial – auto correlation'.

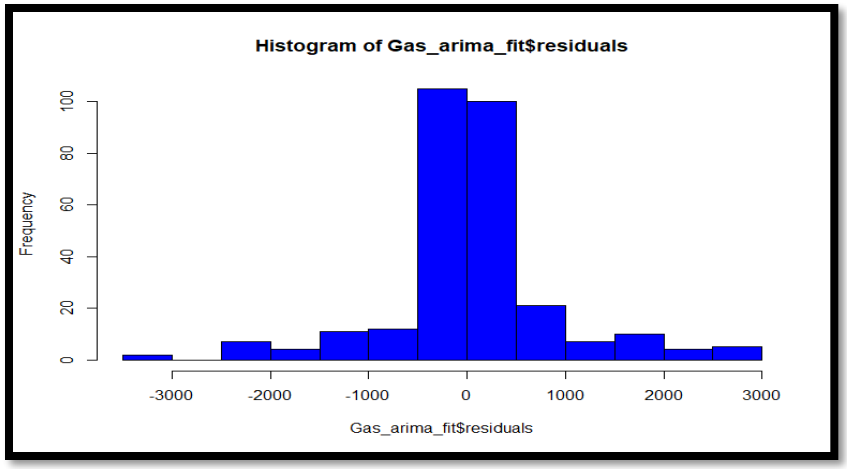


From the PACF plot, it is concluded that p value is 1, because rest of the lags are within blue band or negatively correlated.

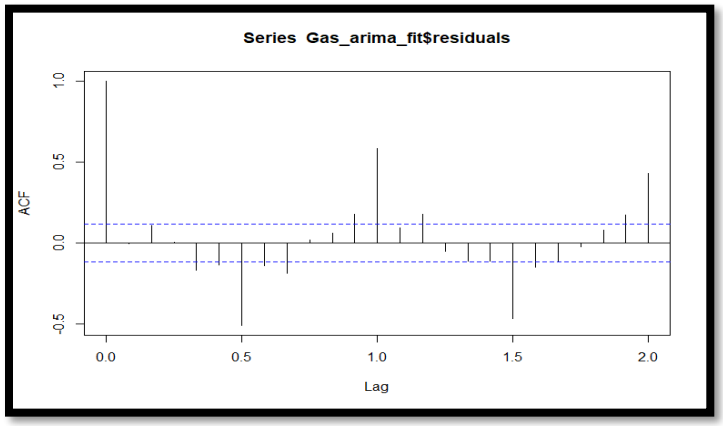
ARIMA Model:

To fit an arima model , the given time series should be stationary.The three parameters p,d,q values are c(1,1,1) respectively.

```
> Gas_arima_fit

Call:
arima(x = Train_data, order = c(1, 1, 1))

Coefficients:
          ar1      ma1
      0.5134  -0.0584
s.e.  0.0850   0.0893

sigma^2 estimated as 728709:  log likelihood = -2344.47,  aic = 4694.94
>
```



To check the adequacy of the model , we have to see the residual values. The histogram plot confirmed that the data are normally distributed. From this we can say the model is working well.



To check autocorrelation among the residuals, used ACF plot. Except the first lag all the nearest lags seems to be insignificant.

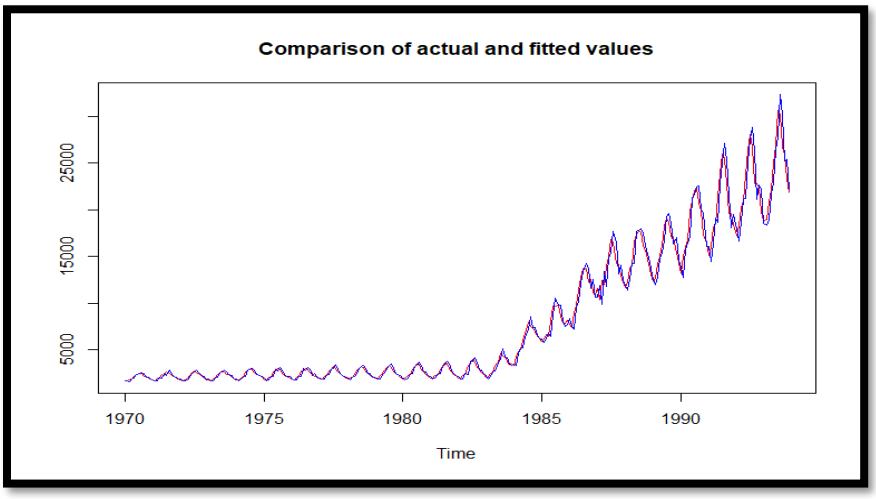
Portmanteau test-Ljung Box Test:

This test used to find whether the residuals are white noise (ie stationary)

H0: Residuals are independent Ha: Residuals are not independent

```
> #Ljung-Box test
> Box.test(Gas_arima_fit$residuals, lag=30, type="Ljung-Box")

        Box-Ljung test

data:  Gas_arima_fit$residuals
X-squared = 396.91, df = 30, p-value < 2.2e-16

>
```

From the ljung test it is concluded that we are failed to reject null hypothesis , because the p value is less than 0.05.The residuals of the gas production dataset are independent.



The above plot is the comparison of actual values and fitted values, more 90% values are coincided. So the arima model is a good model.

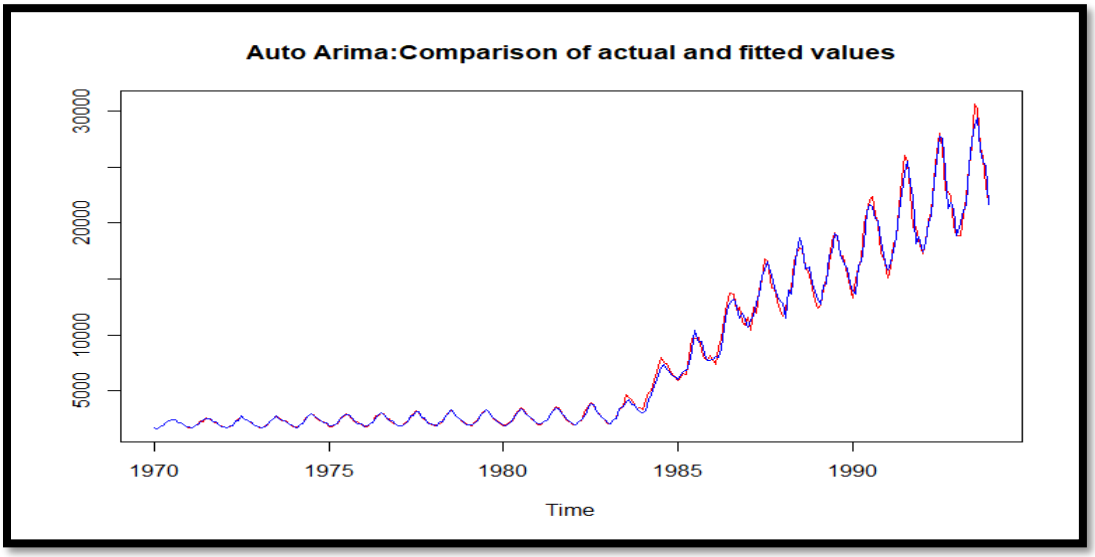
Auto Arima Model:

```
> Auto_arima_model=auto.arima(Train_data,ic="aic",trace=TRUE)

 Fitting models using approximations to speed things up...

 ARIMA(2,1,2)(1,1,1)[12]                    : 4067.452
 ARIMA(0,1,0)(0,1,0)[12]                    : 4121.304
 ARIMA(1,1,0)(1,1,0)[12]                    : 4079.053
 ARIMA(0,1,1)(0,1,1)[12]                    : 4055.018
 ARIMA(0,1,1)(0,1,0)[12]                    : 4093.278
 ARIMA(0,1,1)(1,1,1)[12]                    : 4068.115
 ARIMA(0,1,1)(0,1,2)[12]                    : 4055.294
 ARIMA(0,1,1)(1,1,0)[12]                    : 4070.799
 ARIMA(0,1,1)(1,1,2)[12]                    : Inf
 ARIMA(0,1,0)(0,1,1)[12]                    : 4065.58
 ARIMA(1,1,1)(0,1,1)[12]                    : 4048.483
 ARIMA(1,1,1)(0,1,0)[12]                    : 4088.126
 ARIMA(1,1,1)(1,1,1)[12]                    : 4062.33
 ARIMA(1,1,1)(0,1,2)[12]                    : 4050.001
 ARIMA(1,1,1)(1,1,0)[12]                    : 4066.641
 ARIMA(1,1,1)(1,1,2)[12]                    : Inf
 ARIMA(1,1,0)(0,1,1)[12]                    : 4059.087
 ARIMA(2,1,1)(0,1,1)[12]                    : 4051.279
 ARIMA(1,1,2)(0,1,1)[12]                    : 4050.221
 ARIMA(0,1,2)(0,1,1)[12]                    : 4052.607
 ARIMA(2,1,0)(0,1,1)[12]                    : 4058.076
 ARIMA(2,1,2)(0,1,1)[12]                    : 4053.47

 Now re-fitting the best model(s) without approximations...

 ARIMA(1,1,1)(0,1,1)[12]                    : 4215.794

 Best model: ARIMA(1,1,1)(0,1,1)[12]
```

The auto arima model has given the best model with p , d and q parameters.

ARIMA(1,1,1)(0,1,1)[12]

```
> #Ljung-Box test
> Box.test(Auto_arima_model$residuals, lag=30, type="Ljung-Box")

        Box-Ljung test

data:  Auto_arima_model$residuals
X-squared = 90.356, df = 30, p-value = 5.799e-08
```



The above plot is the comparison of actual values and fitted values, more 90% values are coincided. So the auto arima model is the best model.

Forecasting 20 periods:

```
#Forecast - 20 periods for arima model
Predicted_arima=forecast(Gas_arima_fit,h=20)
autoplot(Predicted_arima)
accuracy(Predicted_arima,Test_data[1:20])
Vec<- cbind(Test_data[1:20],Predicted_arima$mean)
ts.plot(Vec, col=c("blue", "red"), main="Gas Production: Actual vs Forecast")
#Auto arima model
Predicted_auto_arima=forecast(Auto_arima_model,h=20)
Vec1<- cbind(Test_data[1:20],Predicted_auto_arima$mean)
ts.plot(Vec1, col=c("blue", "red"), main="Gas Production: Actual vs Forecast")
accuracy(Predicted_auto_arima,Test_data[1:20])
```

Both arima and auto arima model is not good at forecasting 20 months from the present accuracy is decreasing as the year increases.

Forecasting 12 periods:

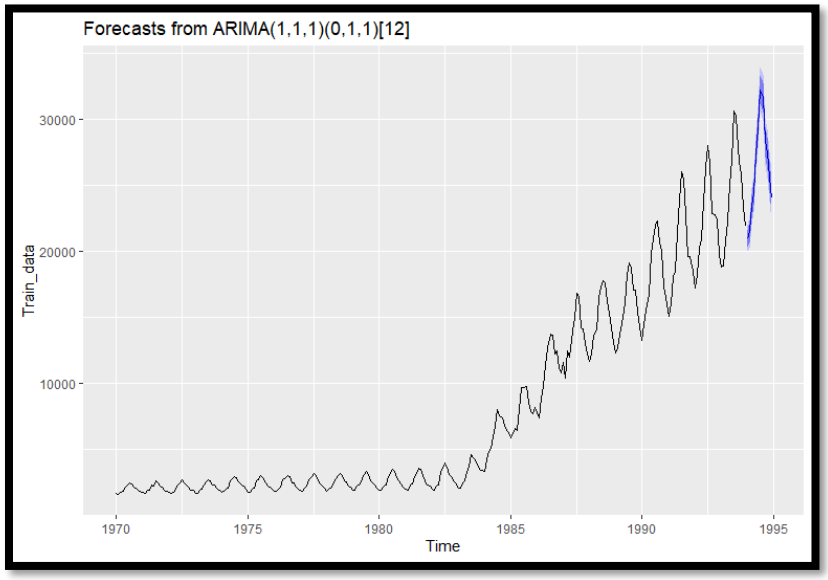
```
#Forecasting 12 periods
#Fit with seasonality-auto arima model
Fit_with_season=auto.arima(Train_data,seasonal = TRUE)
Predicted_arima_12=forecast(Fit_with_season,h=12)
Vec2<- cbind(Test_data[1:12],Predicted_arima_12$mean)
ts.plot(Vec2, col=c("blue", "red"), main="Gas Production with seasonality(Auto Arima): Actual vs F
accuracy(Predicted_arima_12,Test_data[1:12])

#Fit with seasonality- arima model
Fit_with_season_1=arima(Train_data,c(1,1,1),seasonal =list(order=c(1,1,1),period=12))
Predicted_arima_12_1=forecast(Fit_with_season_1,h=12)
Vec3<- cbind(Test_data[1:12],Predicted_arima_12_1$mean)
ts.plot(Vec3, col=c("blue", "red"), main="Gas Production with seasonality(Arima): Actual vs Foreca
accuracy(Predicted_arima_12_1,Test_data[1:12])
```

With the seasonal component, both arima and auto arima are good at predicting the future values.

Summary of auto arima model with seasonality:

```
> summary(Predicted_arima_12)

Forecast method: ARIMA(1,1,1)(0,1,1)[12]

Model Information:
Series: Train_data
ARIMA(1,1,1)(0,1,1)[12]

Coefficients:
         ar1      ma1     sma1
      0.5489  -0.8076  -0.4130
s.e.  0.1061   0.0698   0.0581

sigma^2 estimated as 259078:  log likelihood=-2103.9
AIC=4215.79   AICc=4215.94   BIC=4230.26

Error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 27.33581 494.6562 290.8485 0.3410585 3.494452 0.2957792 -0.01884895

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 1994       20964.82 20312.52 21617.13 19967.21 21962.44
Feb 1994       21472.27 20660.26 22284.27 20230.41 22714.12
Mar 1994       23397.32 22496.10 24298.55 22019.02 24775.63
Apr 1994       24475.53 23512.26 25438.80 23002.33 25948.72
May 1994       27325.34 26312.73 28337.94 25776.69 28873.98
Jun 1994       29352.03 28296.79 30407.26 27738.18 30965.87
Jul 1994       32242.76 31148.80 33336.73 30569.69 33915.83
Aug 1994       31663.37 30533.17 32793.57 29934.89 33391.86
Sep 1994       28306.14 27141.45 29470.83 26524.91 30087.38
Oct 1994       27416.17 26218.31 28614.04 25584.20 29248.15
Nov 1994       25664.80 24434.83 26894.77 23783.73 27545.88
Dec 1994       24035.73 22774.56 25296.89 22106.94 25964.52
```

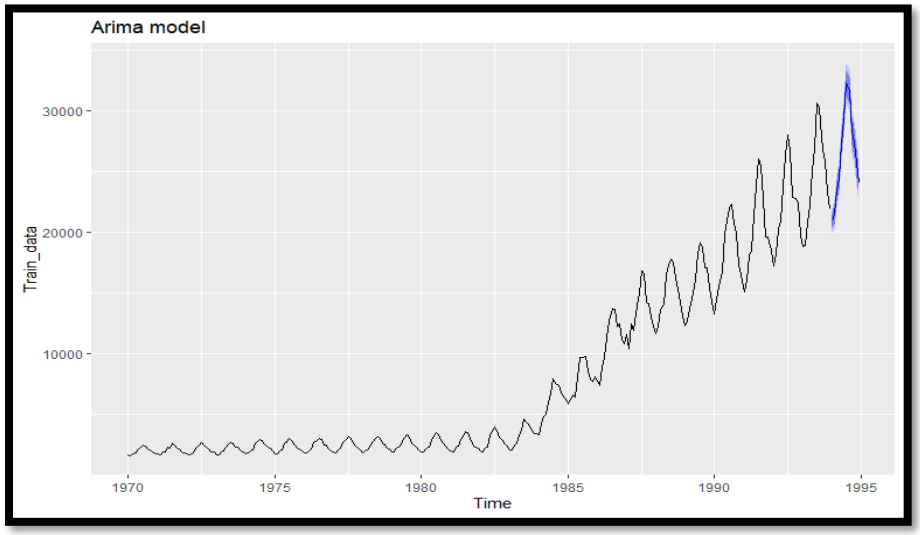
Autoplot of forecasted model from auto arima:



Summary of arima model with seasonality:

```
> summary(Predicted_arima_12_1)

Forecast method: ARIMA(1,1,1)(1,1,1)[12]

Model Information:

Call:
arima(x = Train_data, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1),
    period = 12))

Coefficients:
         ar1      ma1     sar1     sma1
      0.5332  -0.7987  -0.0643  -0.3622
s.e.  0.1136   0.0746   0.1324   0.1246

sigma^2 estimated as 255989:  log likelihood = -2103.77,  aic = 4217.55

Error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 26.53454 494.4026 290.9154 0.3326174 3.489578 0.2958472 -0.01900282

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 1994       20958.54 20310.14 21606.95 19966.89 21950.19
Feb 1994       21484.96 20680.43 22289.50 20254.53 22715.39
Mar 1994       23419.38 22527.69 24311.08 22055.65 24783.11
Apr 1994       24489.08 23536.33 25441.83 23031.98 25946.18
May 1994       27347.81 26346.06 28349.57 25815.77 28879.86
Jun 1994       29401.52 28357.08 30445.97 27804.18 30998.87
Jul 1994       32275.45 31191.98 33358.91 30618.43 33932.46
Aug 1994       31665.23 30545.11 32785.34 29952.15 33378.30
Sep 1994       28267.00 27111.91 29422.09 26500.44 30033.56
Oct 1994       27398.47 26209.70 28587.24 25580.40 29216.54
Nov 1994       25696.43 24475.04 26917.81 23828.48 27564.38
Dec 1994       24023.62 22770.55 25276.70 22107.21 25940.04
```

Autoplot of forecasted model from arima model:



7)Model Accuracy:

```
> ###Model Accuracy
> #Auto arima model
> accuracy(Predicted_arima_12,Test_data[1:12])
                    ME      RMSE       MAE       MPE     MAPE      MASE        ACF1
Training set   27.33581  494.6562  290.8485 0.3410585 3.494452 0.4736167 -0.01884895
Test set     2472.14320 2916.6446 2472.1432 7.9758770 7.975877 4.0256294          NA
> #Arima model
> accuracy(Predicted_arima_12_1,Test_data[1:12])
                    ME      RMSE       MAE       MPE     MAPE      MASE        ACF1
Training set   26.53454  494.4026  290.9154 0.3326174 3.489578 0.4737256 -0.01900282
Test set     2462.87508 2903.3892 2462.8751 7.9460017 7.946002 4.0105372          NA
>
```

The given gas production data is having increasing trend and partial seasonality,so with the seasonality component both arima and auto arima model is predicting the data 90% correctly.

As the MAPE(Mean absolute percentage error) is very less ,lesser than difference between the actual and forecasted values. In this scenario MAPE for both the model is very less and good in forecasting the future. Australian gas production data upward trend and semi-annual component are significant components of time series.