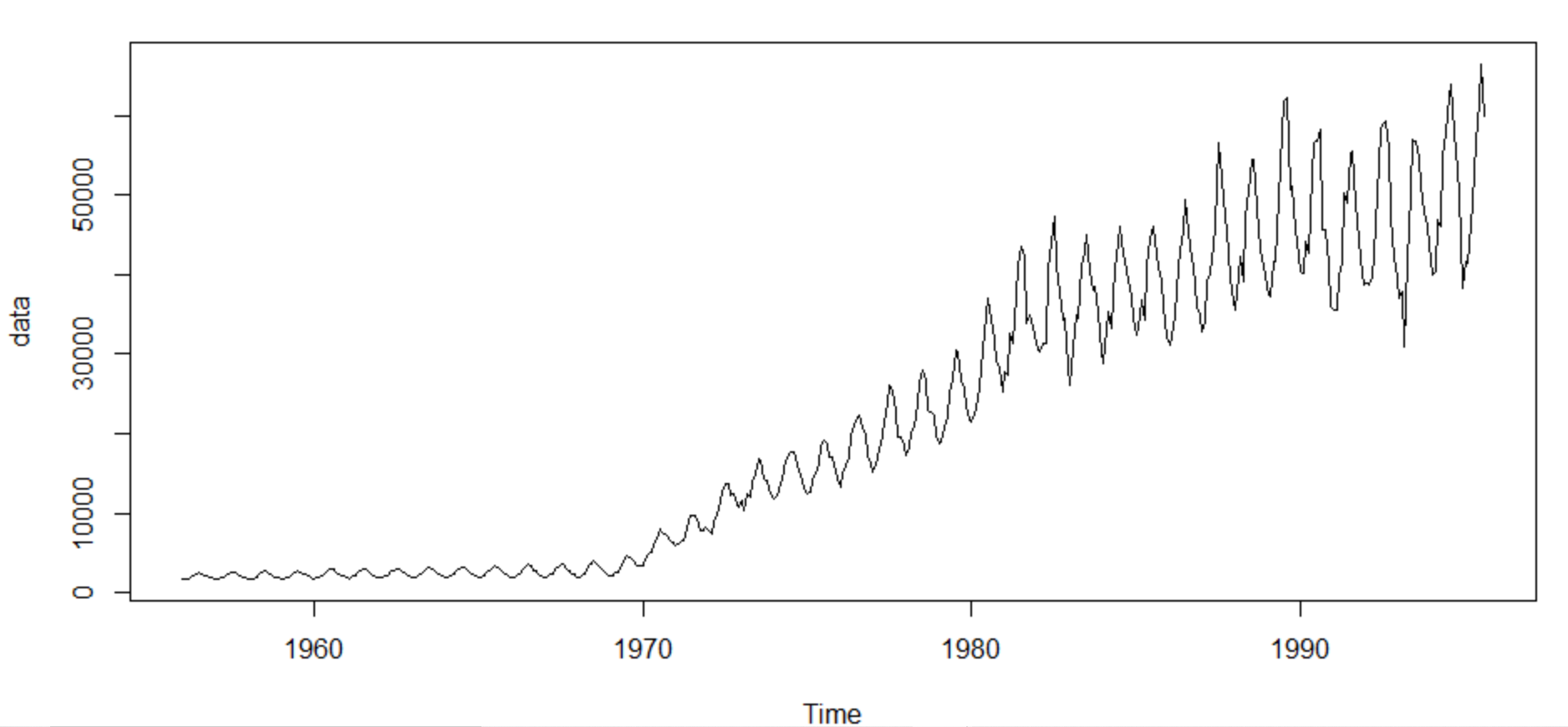
# Project Objective:

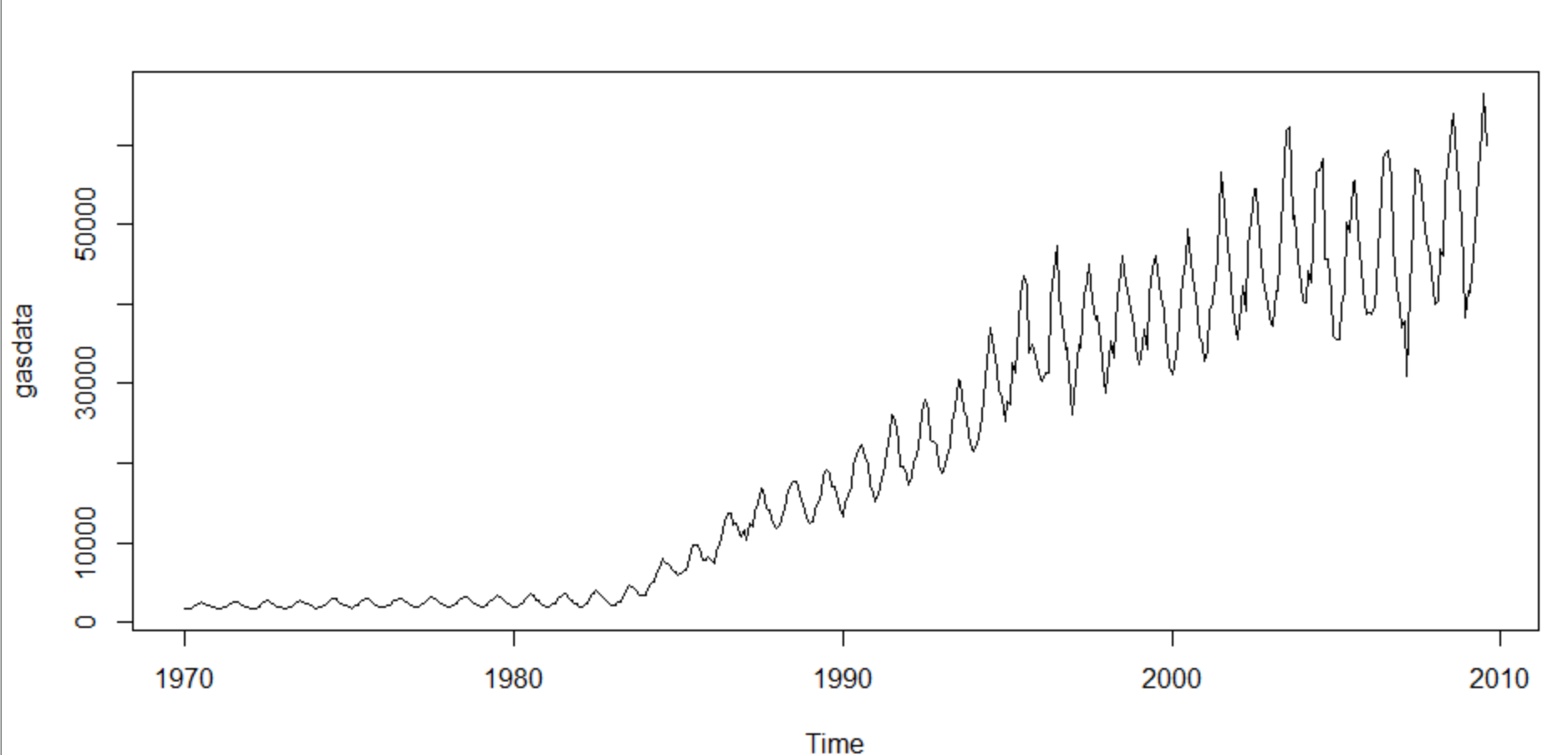
The objective of the project is to explore the **gas** (Australian monthly gas production) dataset in Forecast package to develop an initial forecast for next 20 periods and then check the same using the various metrics, after finalising the model, develop a final forecast for the 12 time periods

# Plot the data. What do you observe? Which components of the time series are present in this dataset?

The original data is present in forecast package under gas dataset. This dataset contains data from 1954.

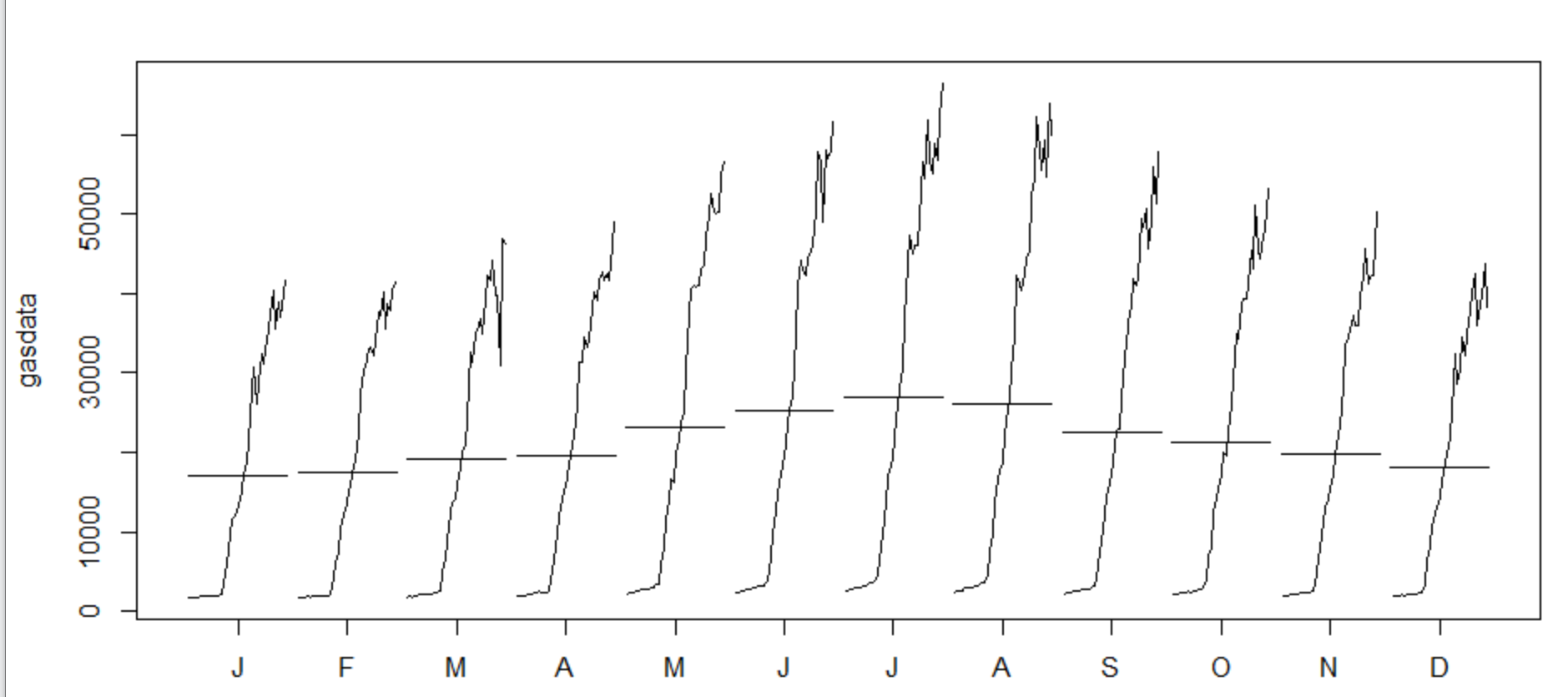


We will however, focus mainly on dataset starting from January 1970. So, our focused dataset looks like below:



Let us try to plot various components of the dataset:

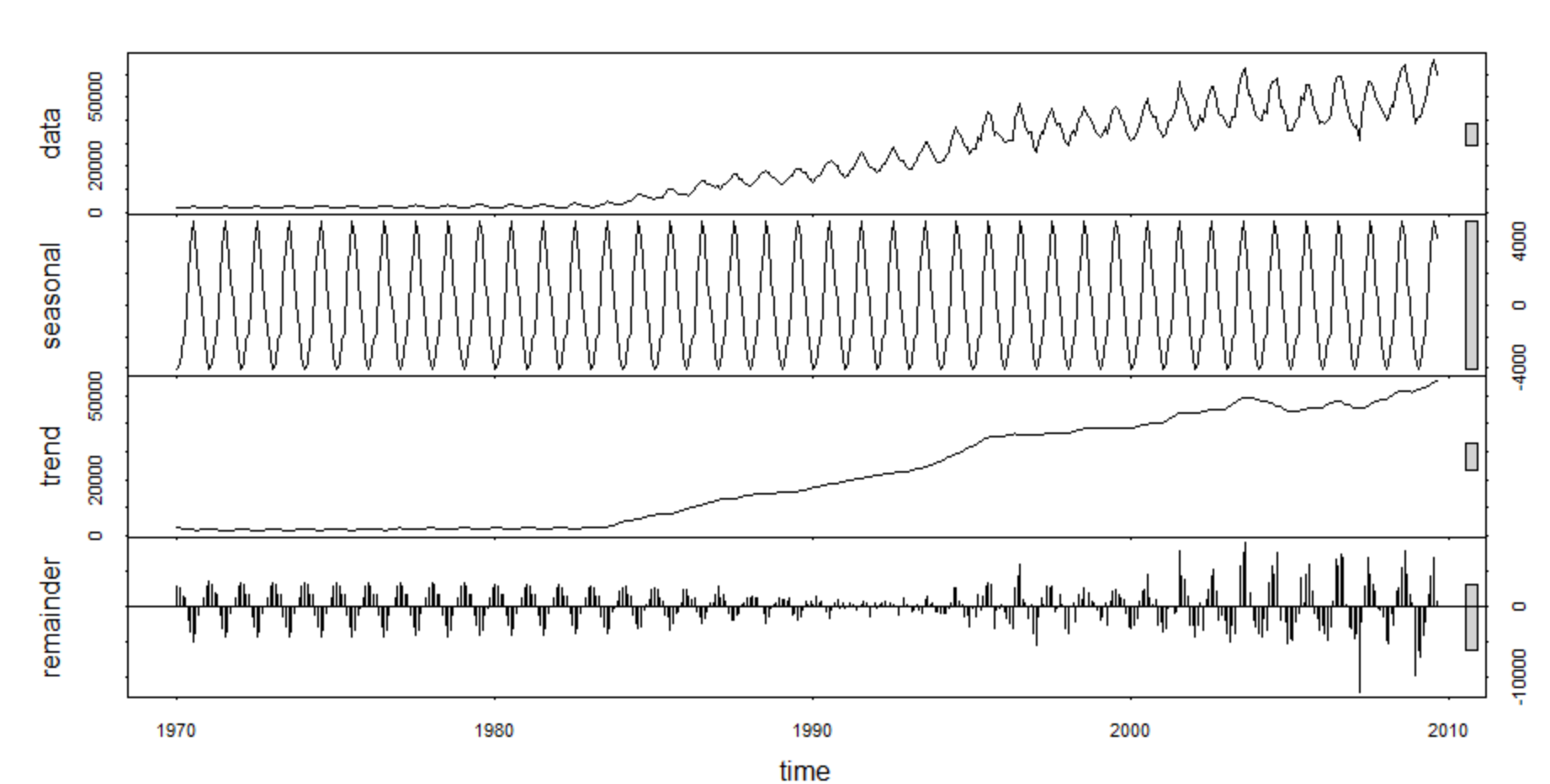
Month plot: This is used to give a clear picture of the seasonal component of the data. As we see here gas production has been increasing for every month of the year. However, there are some fluctuation during the last part of the series may be during 2008-2010



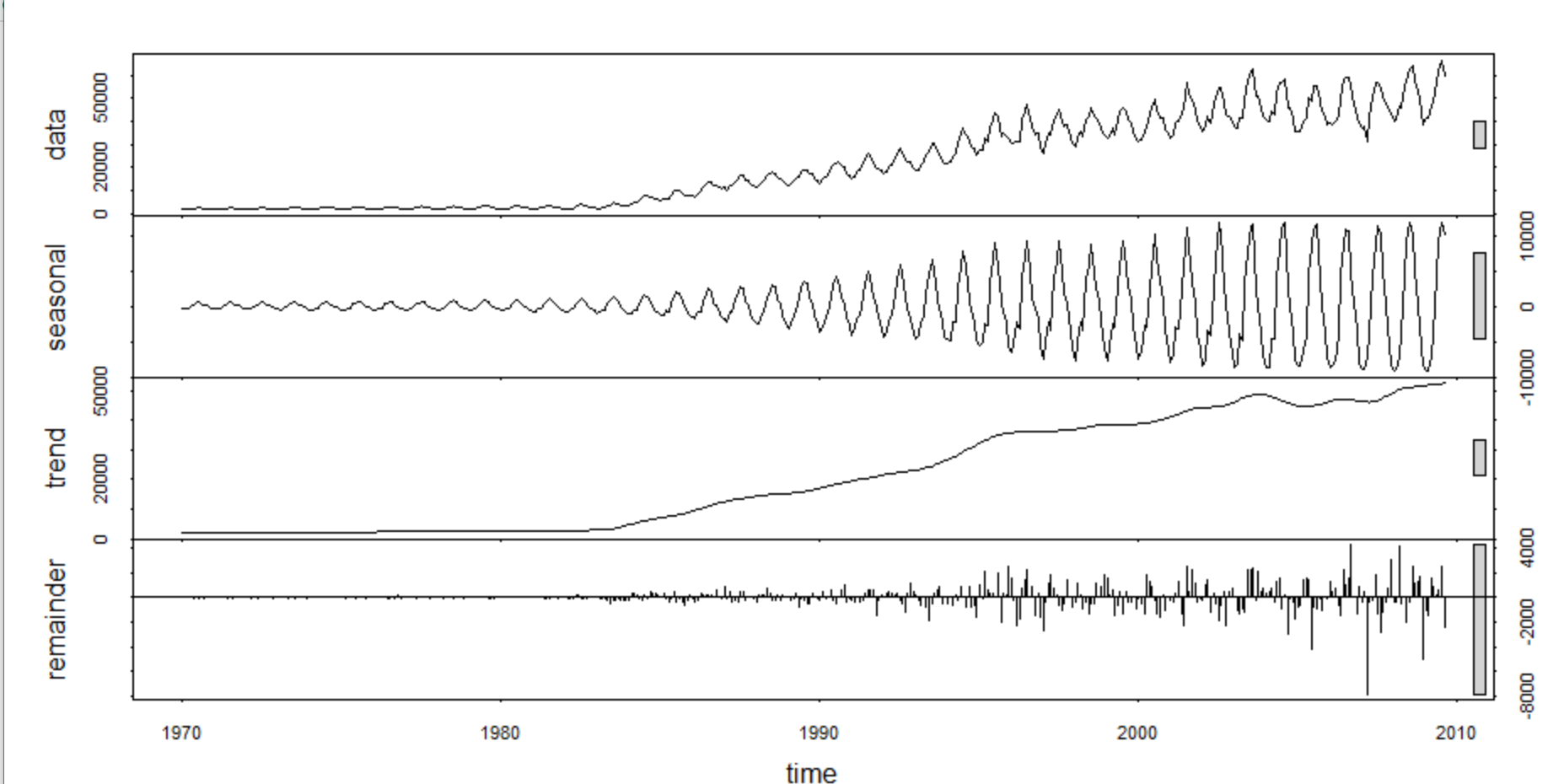
Decompose the dataset :

We use stl() to decompose the dataset to find out the trend , seasonality and error components of the data. This will help us to understand which component is contributing more in determining the future values of the series.

Let us first assume the seasonality is constant. If we decompose now, it is evident that Trend is much more significant that the seasonality in this data set .



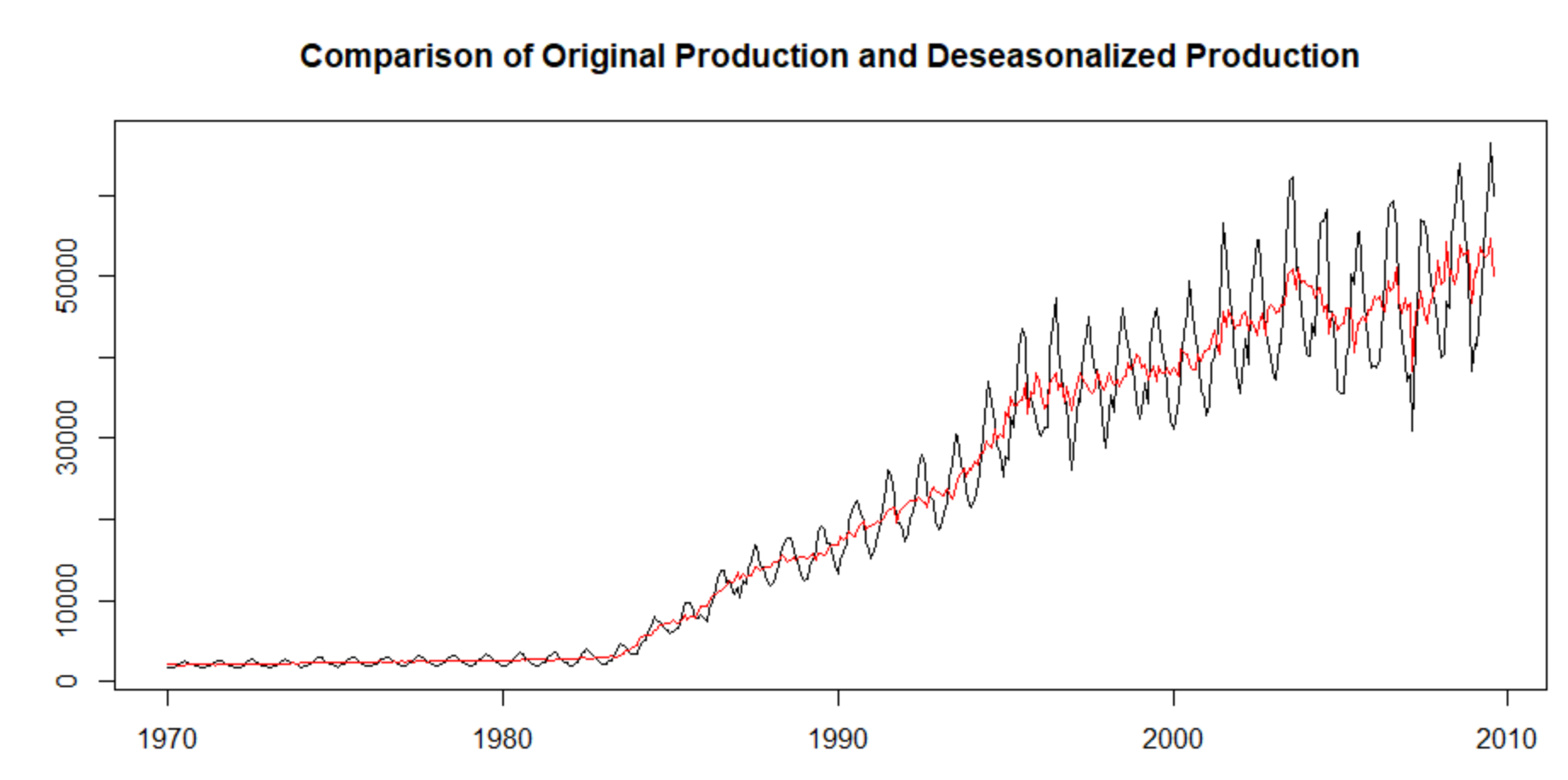
IF we decompose assuming the seasonality is changing, we get below plot .



In this decomposition too, the trend is more significant however we observer the seasonality increases over the time.

There is one big spike in the error part on 2008, mostly due to global recession. Looking at the increasing trend of seasonality, it looks like this data set is multiplicative

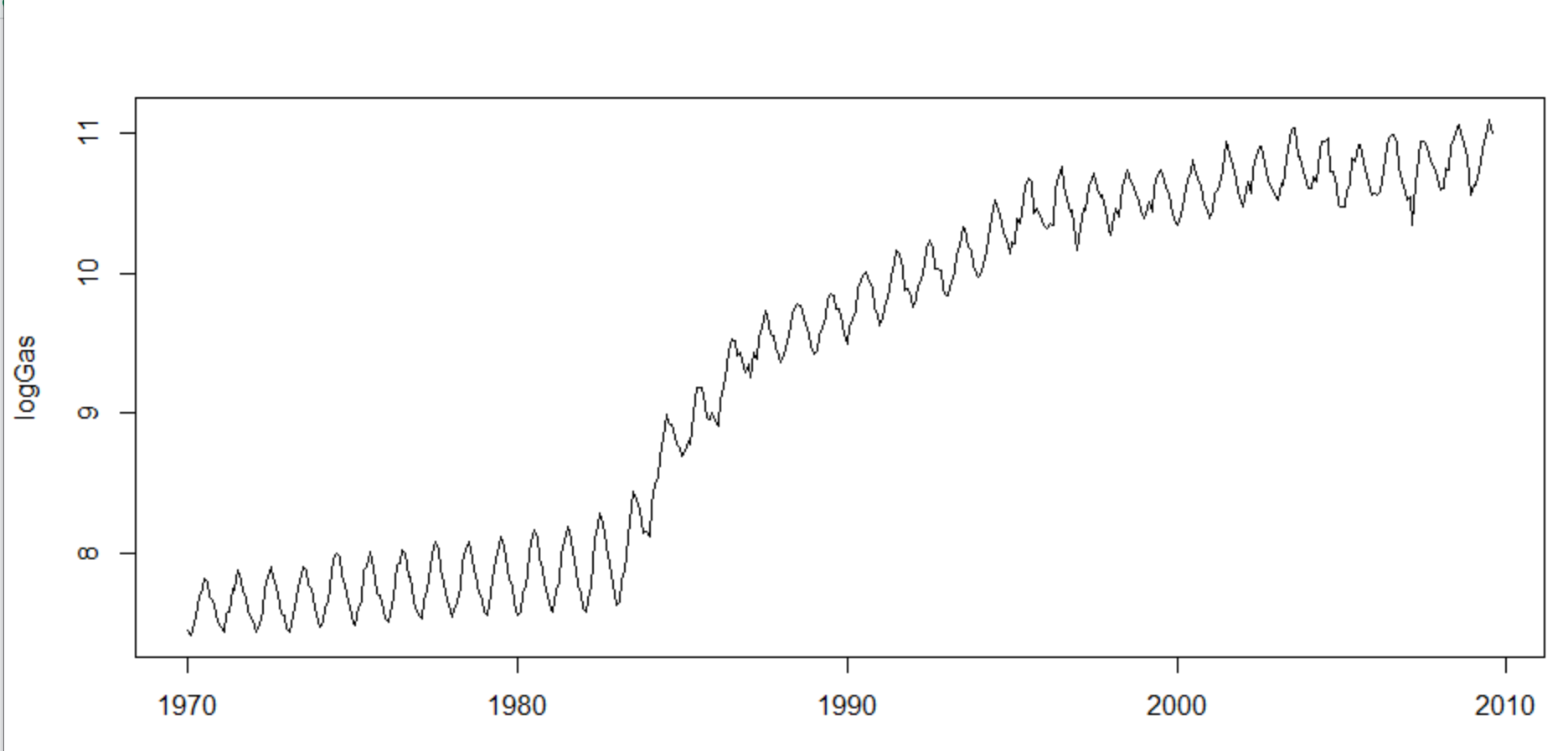
After decomposing, let’s build a dataset excluding the seasonality and plot it against the original dataset.



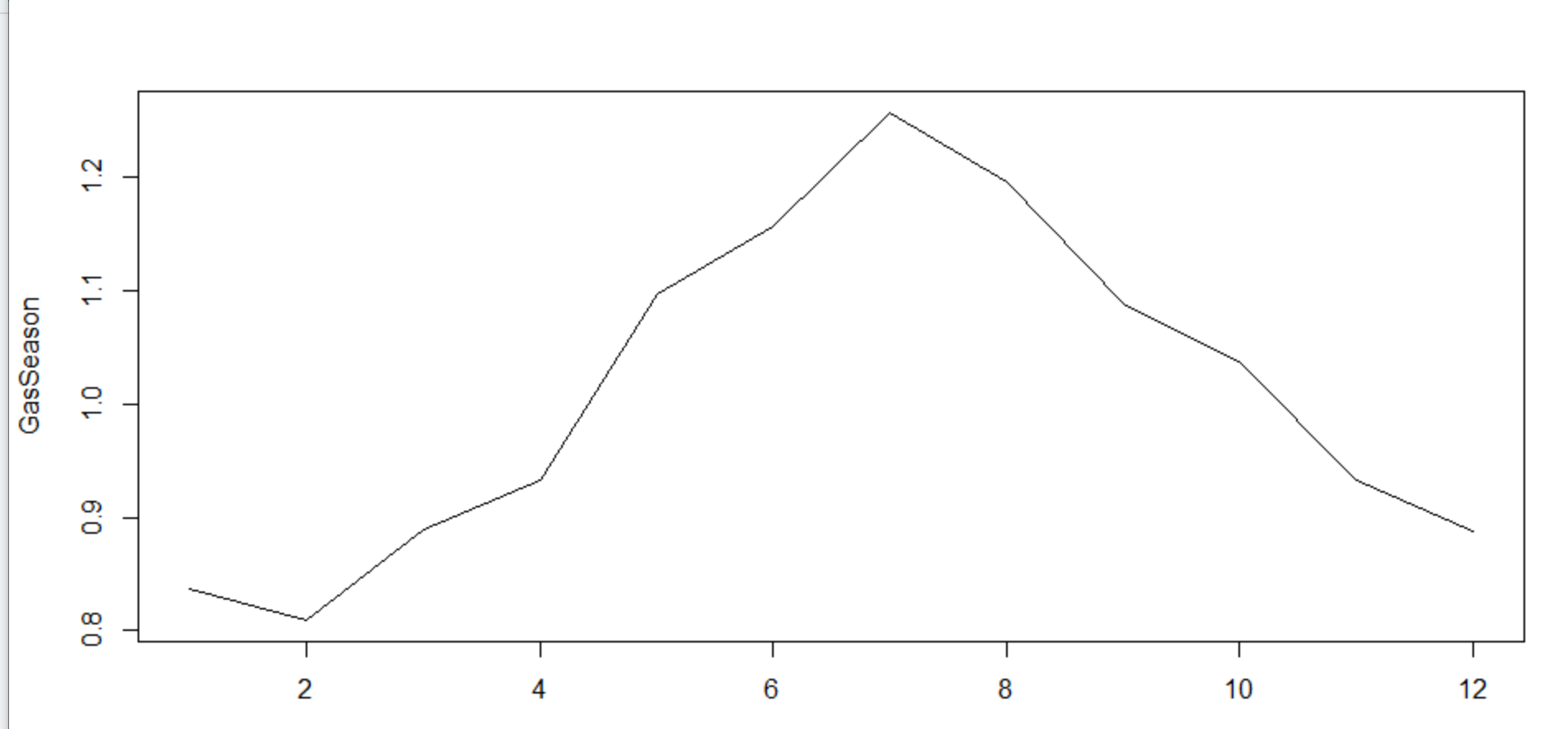
The red line shows the data set that contains only Trend and Error part. This comparison gives a clear indication of the seasonality of the dataset.

Since the series is multiplicative, we need to find the log of the original dataset and then we will find out the pattern of overall log series , seasonality and Trend.

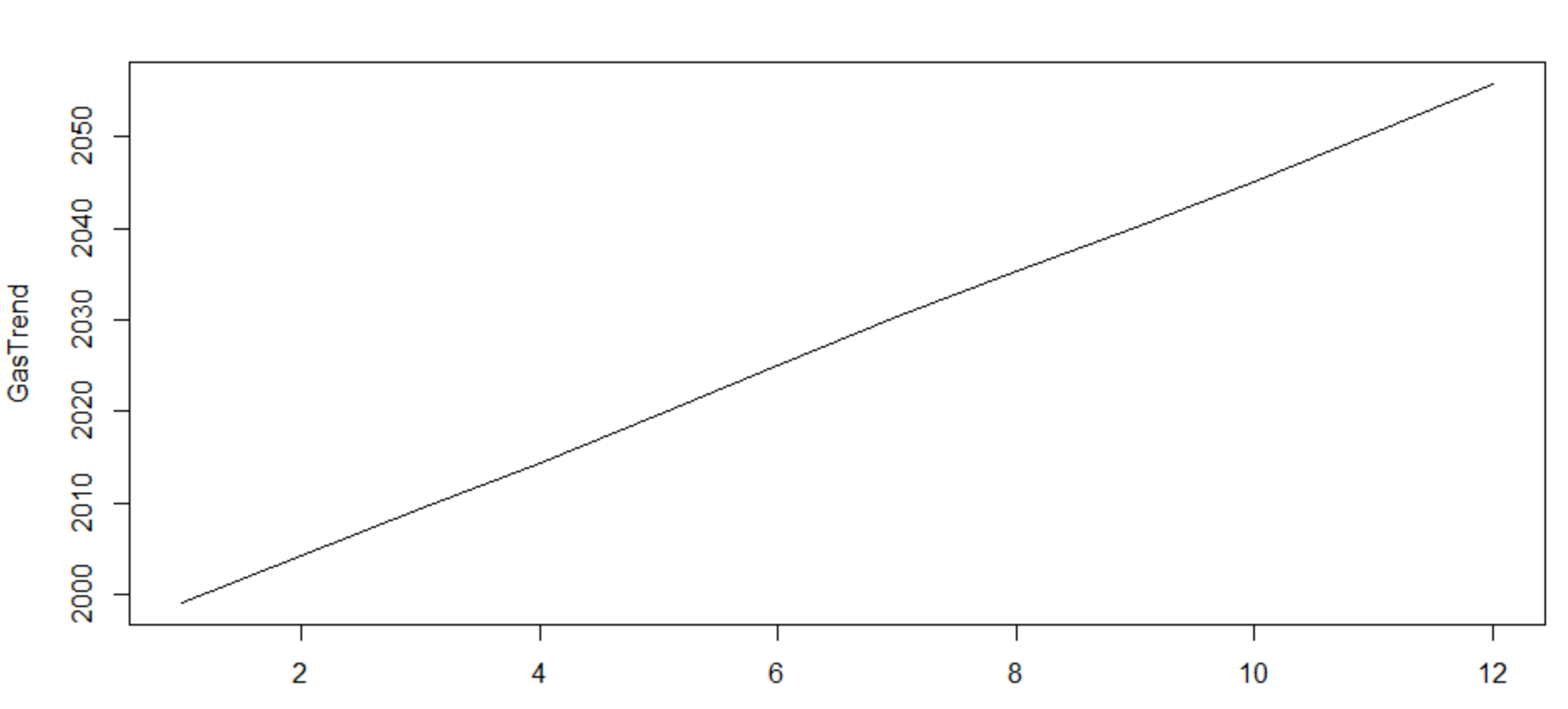
Log Series Plot: The series has now become an additive series with upward trend . From 1970 to 1985 , the trend is not much upward , however from 1985 onwards we can observe a huge upward trend in the series.



**Log plot of Seasonality:**



**Log Plot of Trend** :As expected, we get a very smooth upward trend in the dataset.

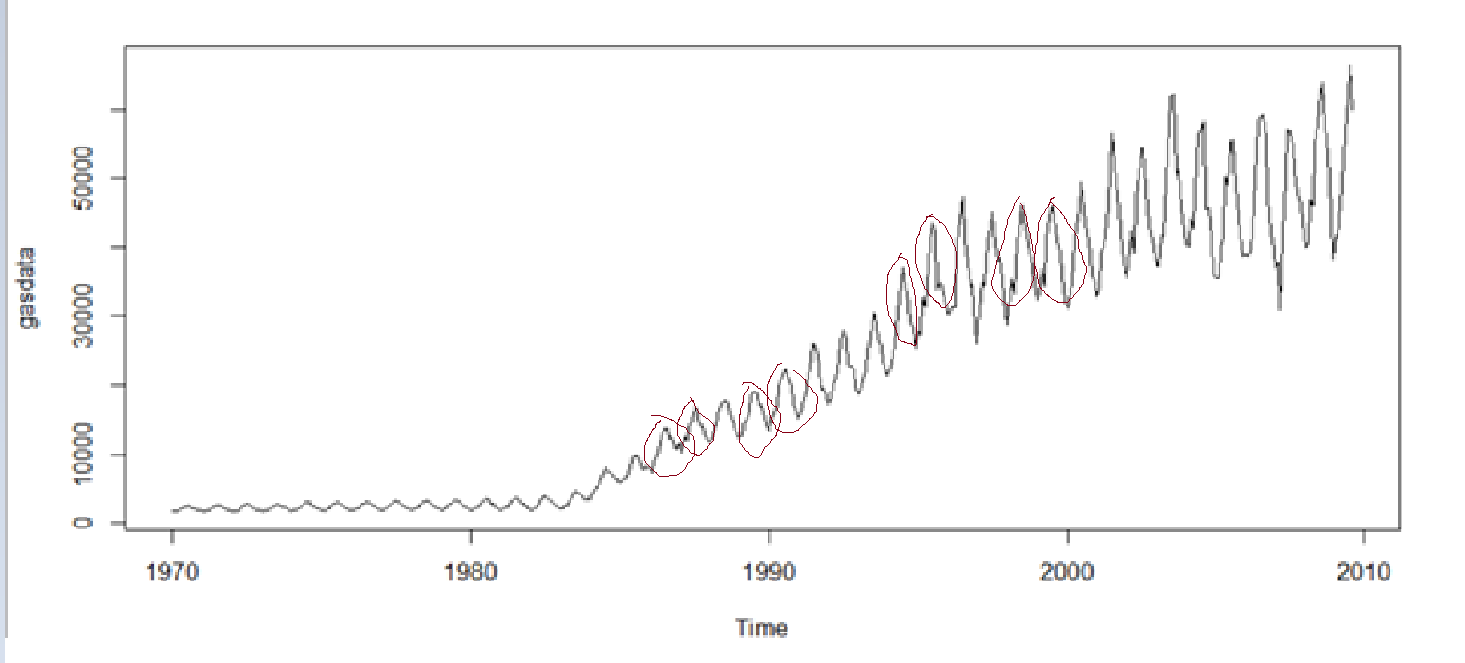


So we can conclude that, the data set has all the 3 components :

1. Trend which is most significant, this tells us the overall gas production has always been rising due to the increasing demand and improvement of technology in energy sector
2. Seasonality exists in the data set and it follows a pattern. From Jan to July, production rises and for the rest of the year production decreases slightly
3. There is also remainder or error component present in the series which is very significant during 2008 due to global recession.

# What is the periodicity of dataset?

Looking at the series plot it is evident that the series has a monthly periodicity. We can see every month the series attains a peak and this



We will also examine the series with a function called periodicity from xts package which returns the period value for a series.

periodicity(gasdata)

Monthly periodicity from Jan 1970 to Aug 2009

And hence we can conclude that the series has monthly periodicity.

# 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?

stationarity means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the *way* it changes does not itself change over time. The algebraic equivalent is thus a linear function, perhaps, and not a constant one; the value of a linear function changes as 𝒙 grows, but the *way* it changes remains constant — it has a constant slope; one value that captures that rate of change

* Stationary time series, mean of the time series will be a constant.
* Time series with trends, or with seasonality, are not stationary
* Trend and seasonality will affect the value of the time series at different times

As we have already established that our series has both high upward trend and seasonality, we can say that this series is not stationary.

A quick visualization of the series also confirms that average gas production volume is always increasing by every consecutive year.

Let’s do a statistical test called ADF (Augmented Dickey-Fuller Test)

* Null hypothesis H0: Time series non-stationary
* Alternative hypothesis Ha: Time series is stationary

We have performed ADF Test on both original Gas data set as well as Deseasonal Data set and in both scenario, we get a very high p value, so we fail to reject the null hypothesis and prove that the series is not stationary.

adf.test(gasdata)

Augmented Dickey-Fuller Test

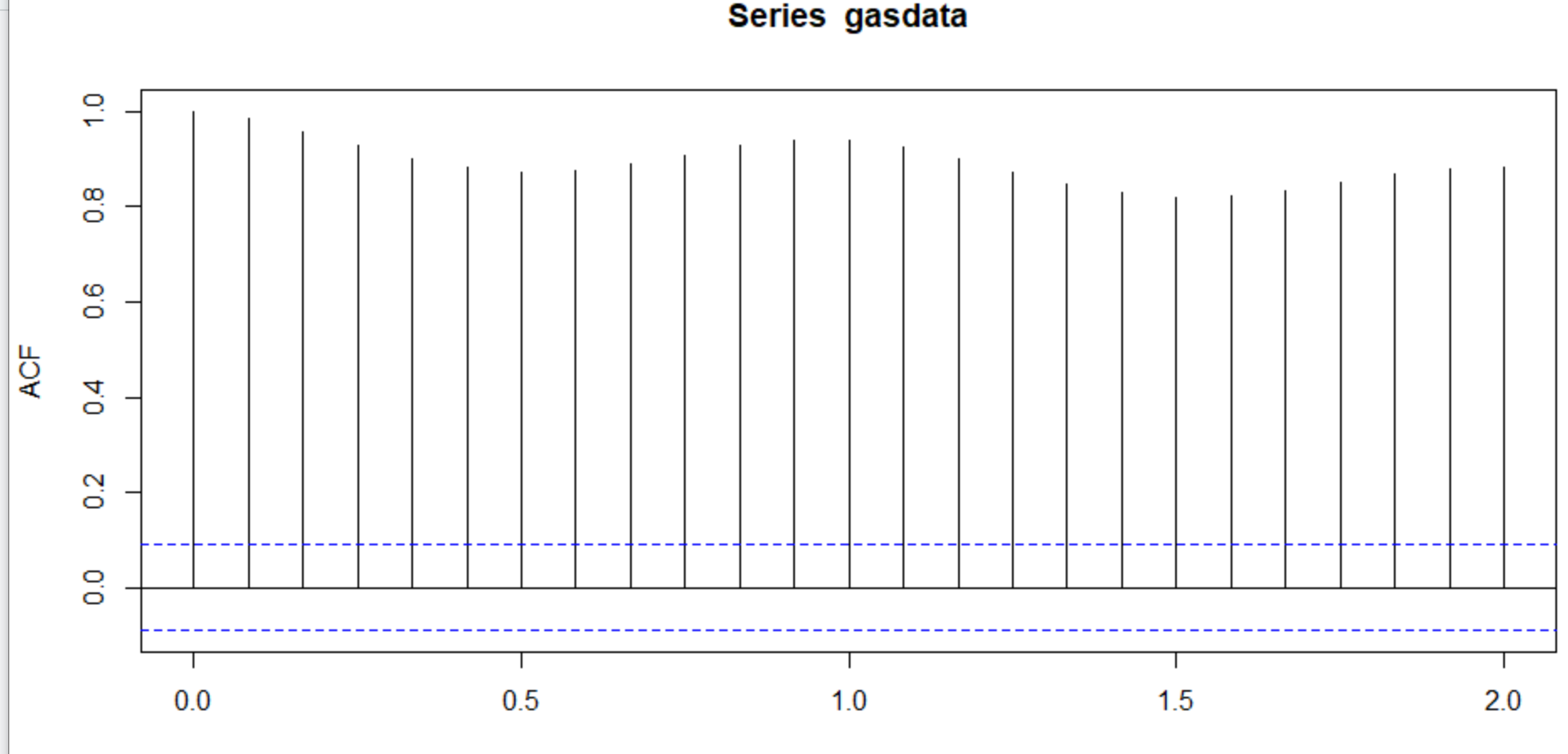
data: gasdata

Dickey-Fuller = -2.7131, Lag order = 7, p-value = 0.2764

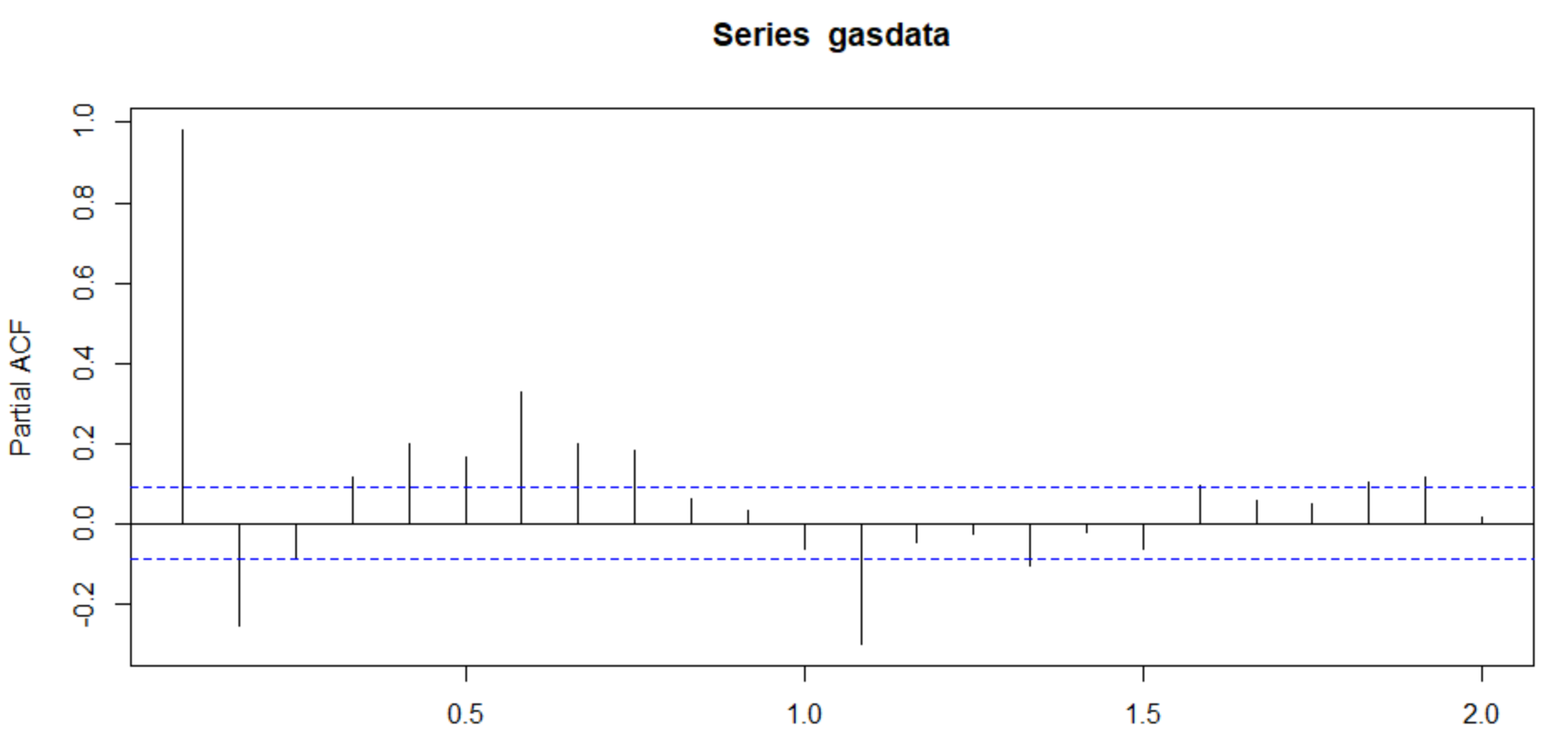
alternative hypothesis: stationary

We will use differencing method to remove.

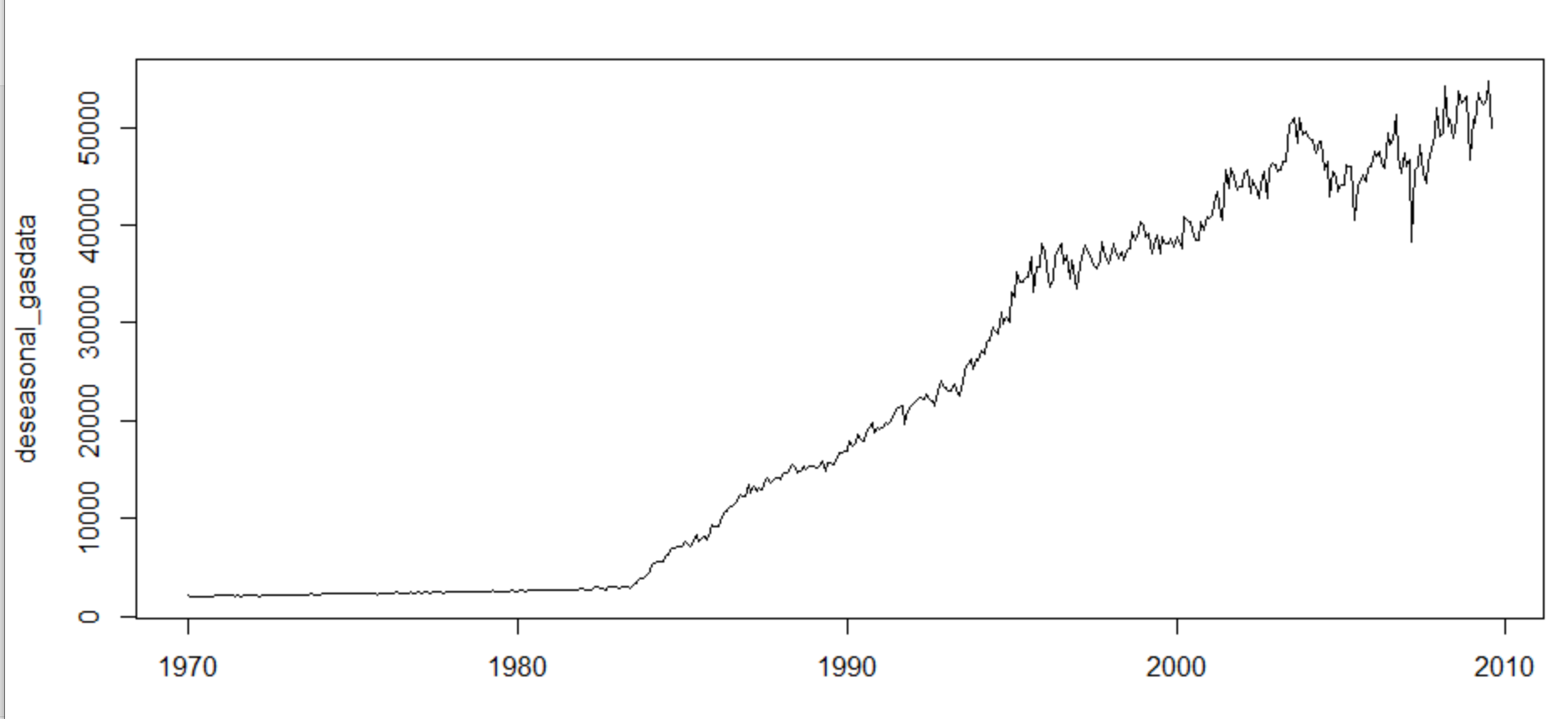
Let’s look at the acf plot for lag = 24. There are significant autocorrelations with many lags in our demand series, as shown by the ACF plot.



PACF Plot : PACF plot shows that there could be monthly seasonality since the plot peaks at intervals of 12

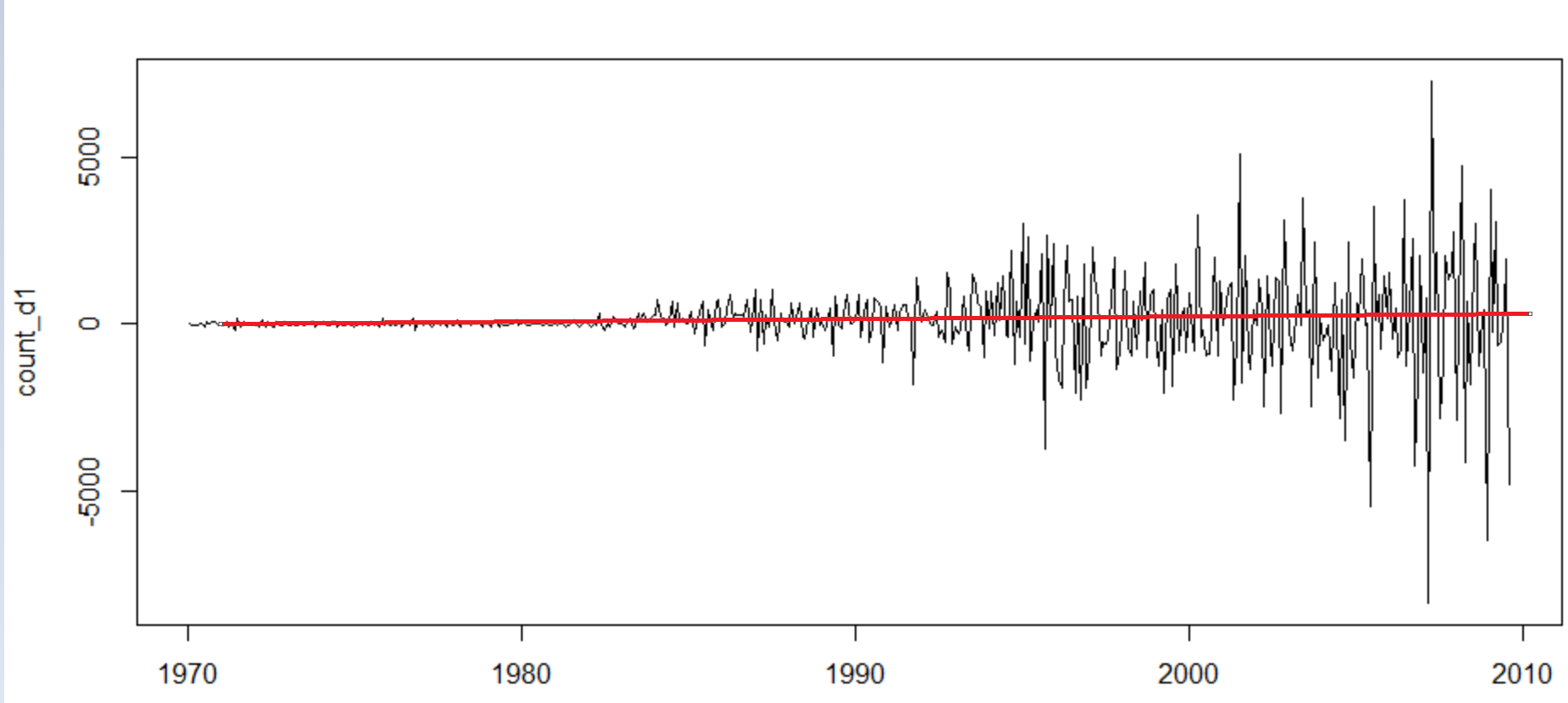


We have used seasadj() to deseasonalize the series and get the below plot .



Also we have used diff() for order 1 to remove the trend.

We plot the new series . We can see even though the variation is more towards the second half of the series , the overall mean is constant and overall variance is also constant . So by visualization method , it looks to be a stationary series . we will do a statistical ad test now.



Applying adf.test on the deseasonal set with removing trend , we get a p value of 0.01.

count\_d1 = diff(deseasonal\_gasdata, differences = 1)

plot(count\_d1)

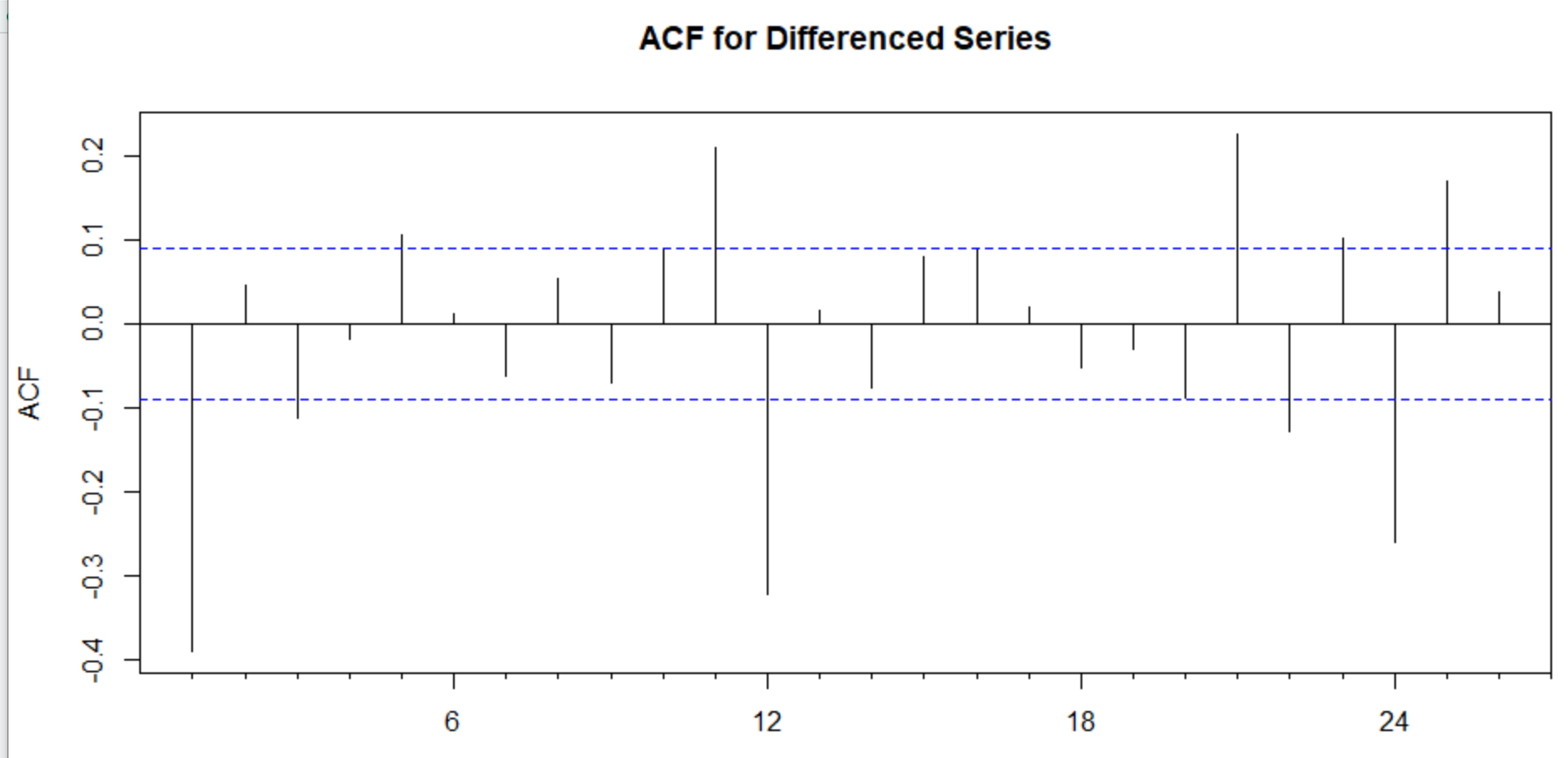
adf.test(count\_d1)

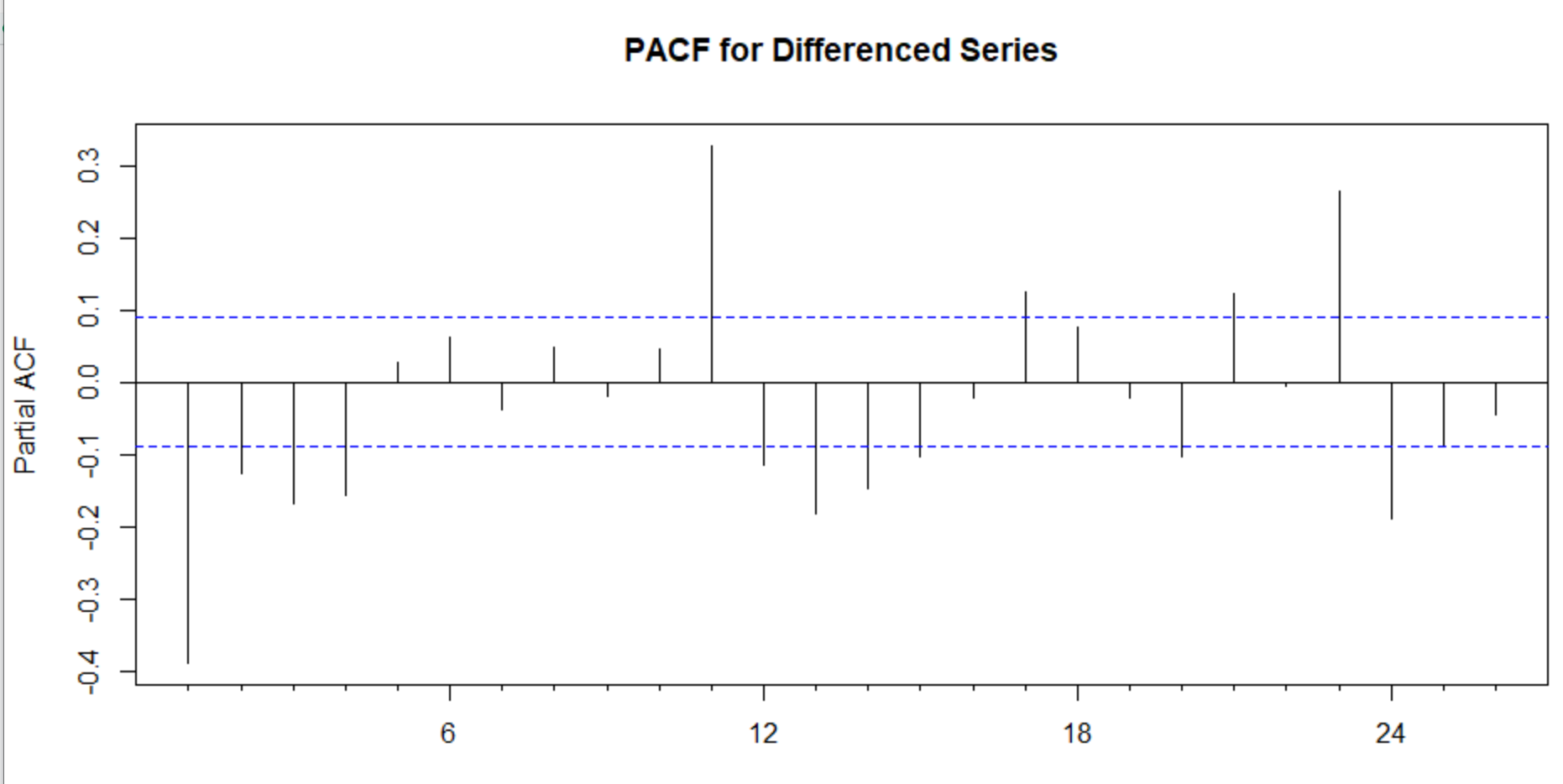
data: count\_d1

Dickey-Fuller = -8.1524, Lag order = 7, p-value = 0.01

So we reject the null hypothesis and accept the alternative hypothesis which means the series is stationary now. So we can use this series for further forecasting.

ACF plot for the stationary series:





The above 2 plots will help in Manual Arima modelling. From ACF plot we see , the first line ( ignore the first line in ACF ) is in the significant zone , so P value is 2 , Q value is 0.

# Develop an ARIMA Model to forecast for next 12 periods. Use both manual and auto.arima (Show & explain all the steps)

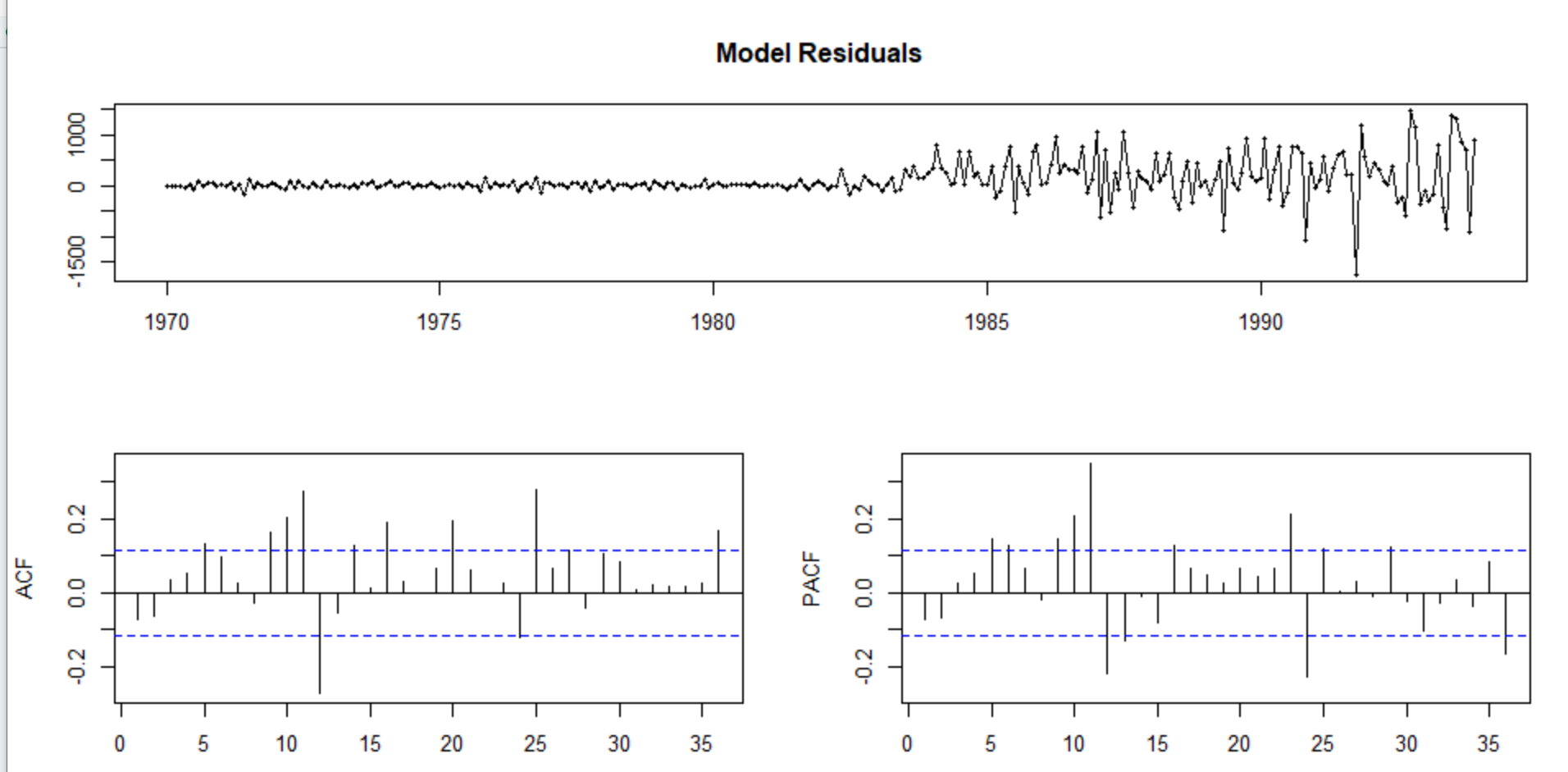
Split the series in train and test series.

* Train series runs from Jan 1970 to Dec 1993
* Test series runs from Jan 1994 to the end

We build manual ARIMA model with P = 2 . D = 1 and Q=0 on the training series

* MAPE = 2.47
* AIC = 4213.57

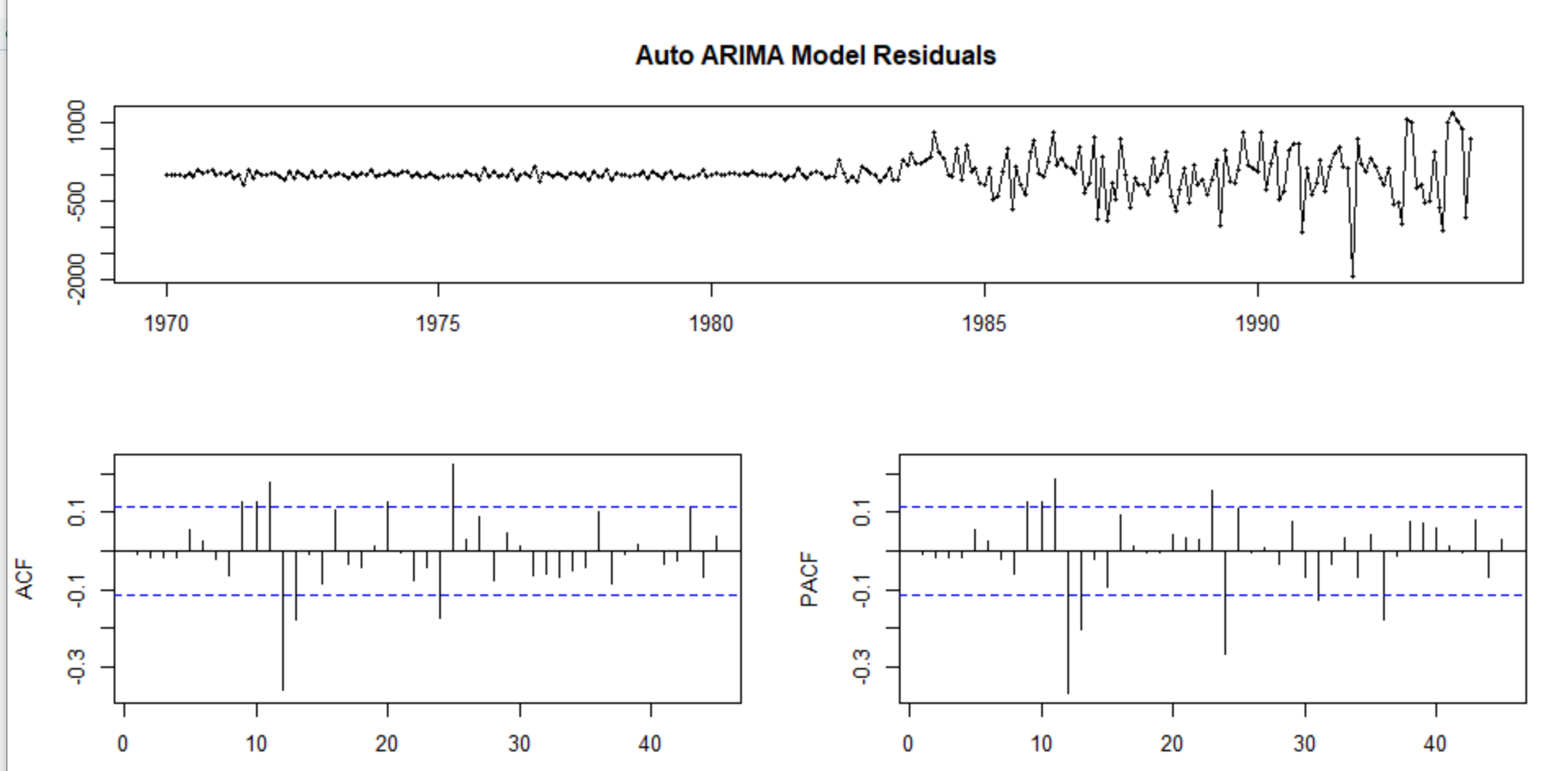
From the residual plot, we understand there are some observations in our model which are beyond significant zone. Residuals plots show few larger error range.



Auto Arima : We apply auto arima () . The function gives P = 0 , D = 2 and Q =3

* MAPE =2.370566
* AIC = 4155.15

So we observe slight improvement in both AIC and MAPE values.



The residual plot also shows improvement in terms of error range.

We will now perform Ljung box test to see the pattern of the residuals

* H0: Residuals are independent
* Ha: Residuals are not independent

The p value of the test is 0.2307, so we accept Ho ie. The residuals are

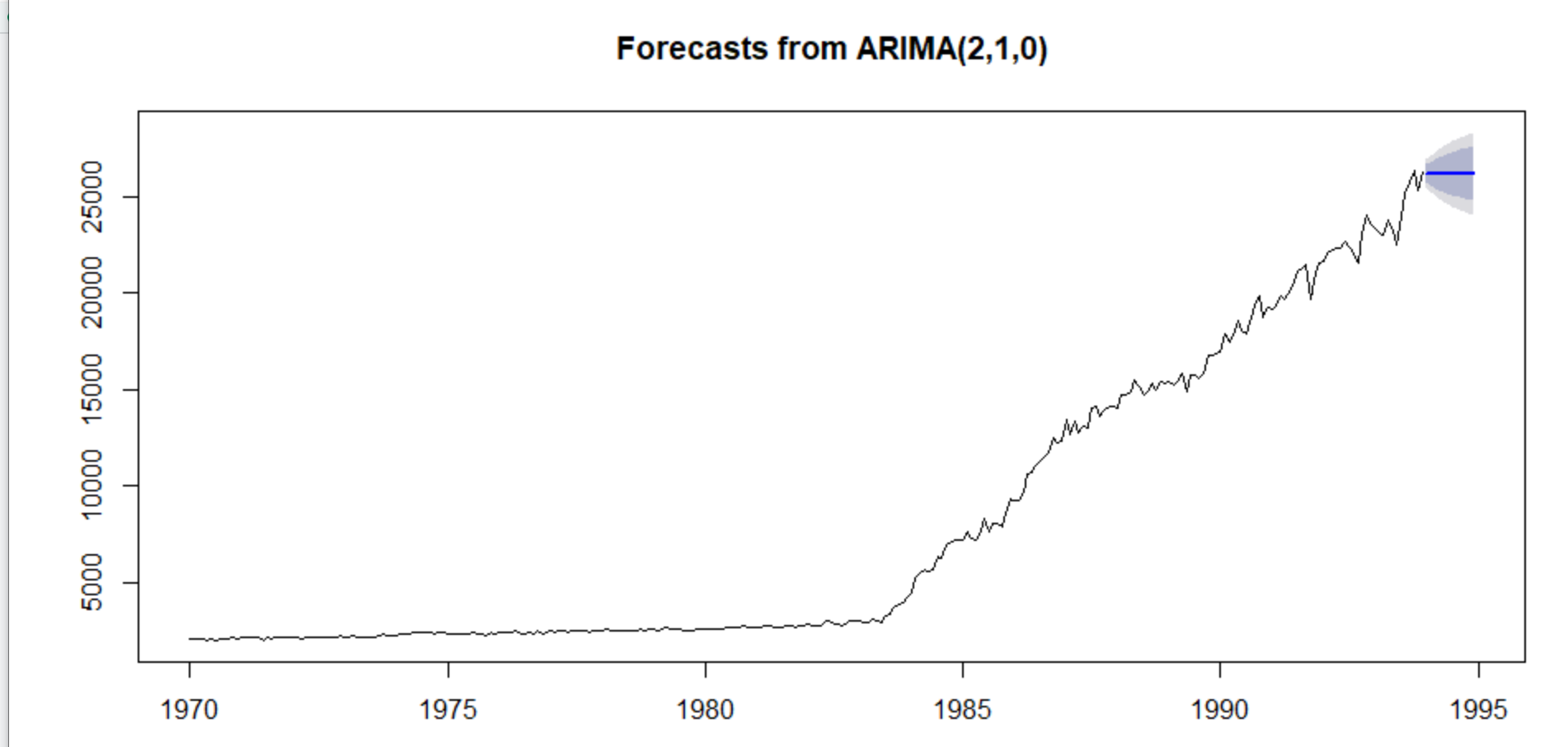
independent.

We will now forecast next 12 months value of the series by using both manual ARIMA model and Auto ARIMA model.

We will use the below process to forecast.

1. With our manual arima model , forecast based on train data .
2. Now run the model on test series to find the accuracy
3. Now run the model on whole series to forecast next 12 months .

Forecast plot for next 12 months based on manual arima model Train series



The overall accuracy using the manual arima :

> accuracy(f7, GasTStest)

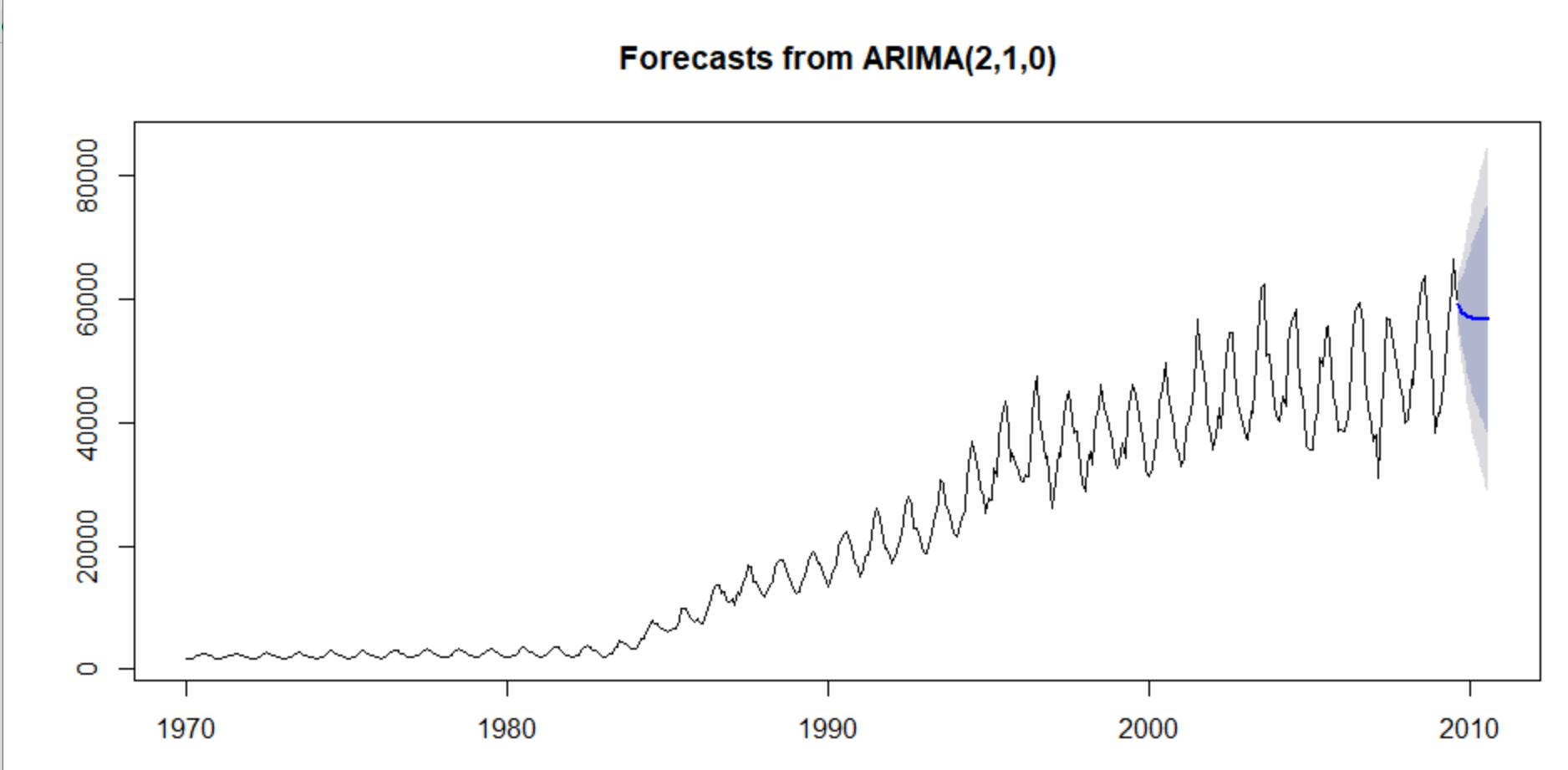
ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 98.44259 368.5403 212.6152 0.9674341 2.476283 0.215813 -0.07062703 NA

Test set 5647.45038 6576.7760 5652.9954 16.8053613 16.826572 5.738018 0.75670713 4.010643

In train series the MAPE is 2.47 , however in test series MAPE is 16.82.

Now we will forecast based on the complete series :



Below is the forecasted value with Manual Arima process.

|  |
| --- |
| Error measures:  ME RMSE MAE MPE MAPE MASE ACF1  Training set 67.59848 2648.252 1547.408 0.342698 7.108944 0.8294831 0.03276301  Forecasts:  Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  Sep 2009 59256.16 55858.72 62653.60 54060.22 64452.10  Oct 2009 57954.06 52522.88 63385.25 49647.78 66260.35  Nov 2009 57497.50 50102.74 64892.25 46188.20 68806.80  Dec 2009 57164.86 48035.78 66293.94 43203.14 71126.58  Jan 2010 57005.56 46312.50 67698.62 40651.93 73359.19  Feb 2010 56910.04 44803.21 69016.86 38394.24 75425.83  Mar 2010 56859.53 43460.58 70258.49 36367.60 77351.46  Apr 2010 56830.93 42240.46 71421.40 34516.73 79145.12  May 2010 56815.33 41116.51 72514.15 32806.06 80824.60  Jun 2010 56806.65 40069.09 73544.20 31208.76 82404.53  Jul 2010 56801.87 39084.47 74519.27 29705.44 83898.29  Aug 2010 56799.22 38152.39 75446.04 28281.36 85317.08  Sep 2010 56797.76 37265.16 76330.36 26925.22 86670.29  Oct 2010 56796.95 36416.75 77177.15 25628.13 87965.77  Nov 2010 56796.50 35602.44 77990.57 24382.97 89210.03  Dec 2010 56796.26 34818.36 78774.16 23183.96 90408.56  Jan 2011 56796.12 34061.35 79530.89 22026.29 91565.95  Feb 2011 56796.05 33328.78 80263.31 20905.97 92686.13  Mar 2011 56796.00 32618.42 80973.59 19819.58 93772.43  Apr 2011 56795.98 31928.35 81663.61 18764.22 94827.74  May 2011 56795.97 31256.92 82335.01 17737.37 95854.57  Jun 2011 56795.96 30602.70 82989.22 16736.83 96855.09  Jul 2011 56795.96 29964.43 83627.48 15760.68 97831.23  Aug 2011 56795.95 29341.00 84250.91 14807.23 98784.68 |
|  |
| |  | | --- | |  | |

We will follow the similar steps for Auto arima .

Error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set 73.0519 2606.1 1527.941 0.3105499 7.015954 0.8190477 0.004415276

Forecasts:

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Sep 2009 58640.96 55290.53 61991.39 53516.92 63764.99

Oct 2009 56168.61 50785.28 61551.94 47935.52 64401.70

Nov 2009 56168.61 48650.49 63686.73 44670.64 67666.58

Dec 2009 56168.61 46999.97 65337.25 42146.38 70190.83

Jan 2010 56168.61 45604.24 66732.97 40011.81 72325.41

Feb 2010 56168.61 44372.52 67964.70 38128.05 74209.17

Mar 2010 56168.61 43257.78 69079.44 36423.20 75914.02

Apr 2010 56168.61 42231.92 70105.30 34854.28 77482.94

May 2010 56168.61 41276.55 71060.66 33393.18 78944.04

Jun 2010 56168.61 40378.89 71958.33 32020.32 80316.90

Jul 2010 56168.61 39529.59 72807.63 30721.42 81615.80

Aug 2010 56168.61 38721.58 73615.64 29485.68 82851.54

## Report the accuracy of the model

So on overall series data both manual and auto arima performed similarly with MAPE around 7 which can be considered a good model .