SHWETA GUPTA

PGPBABI

Time Series Forecasting Project

Forecasting the gas PRODUCTION in australia

Background & Objectives

# BACKGROUND

* The Gas dataset from the forecast package has Australian monthly gas production data from year 1956–1995.
* This dataset is used to model the Gas production for Australia using various timeseries methods.
* Finally the best model is shortlisted basis the accuracy on key evaluation measures and is used to predict the future Gas Production

# OBJECTIVES

|  |  |
| --- | --- |
| 1. Read the data as a time series object in R. Plot the data |  |
| 2. Which components of the time series are present in this dataset? |  |
| 3. What is the periodicity of dataset? |  |
| 4. 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 |  |
| 5. Develop an ARIMA Model to forecast for next 12 periods. Use both manual and auto.arima (Show & explain all the steps) |  |
| 6. Report the accuracy of the model  Assumptions Of Time Series   * There is information about past * This information can be quantified in form of data * The pattern of past will continue in future   Model Building Flow |  |

Data Preparation and Exploratory Analysis

### Read the data as a time series object in R. Plot the data

summary(gas)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1646 2675 16788 21415 38628 66600

any(is.na(gas))

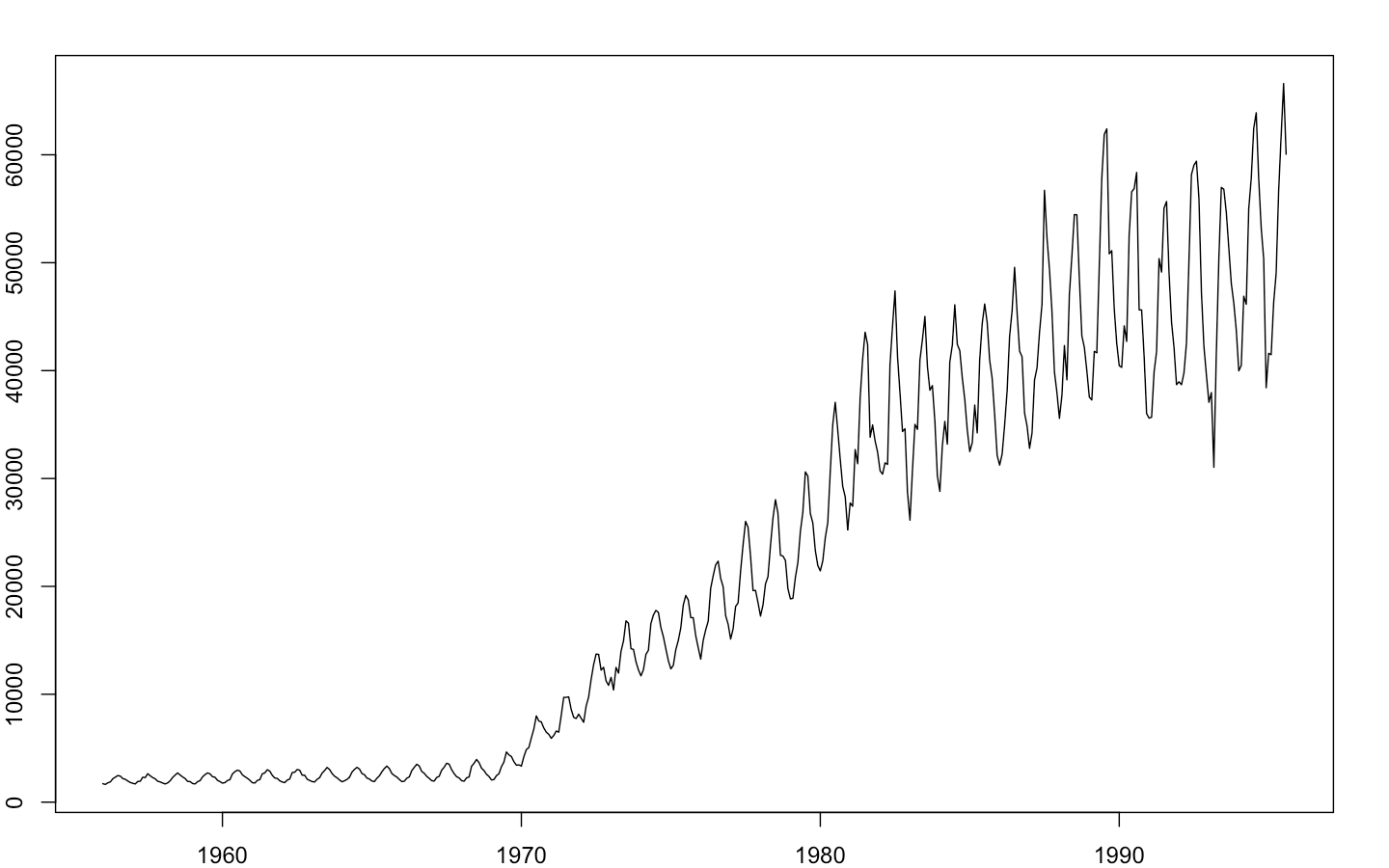
[1] FALSE

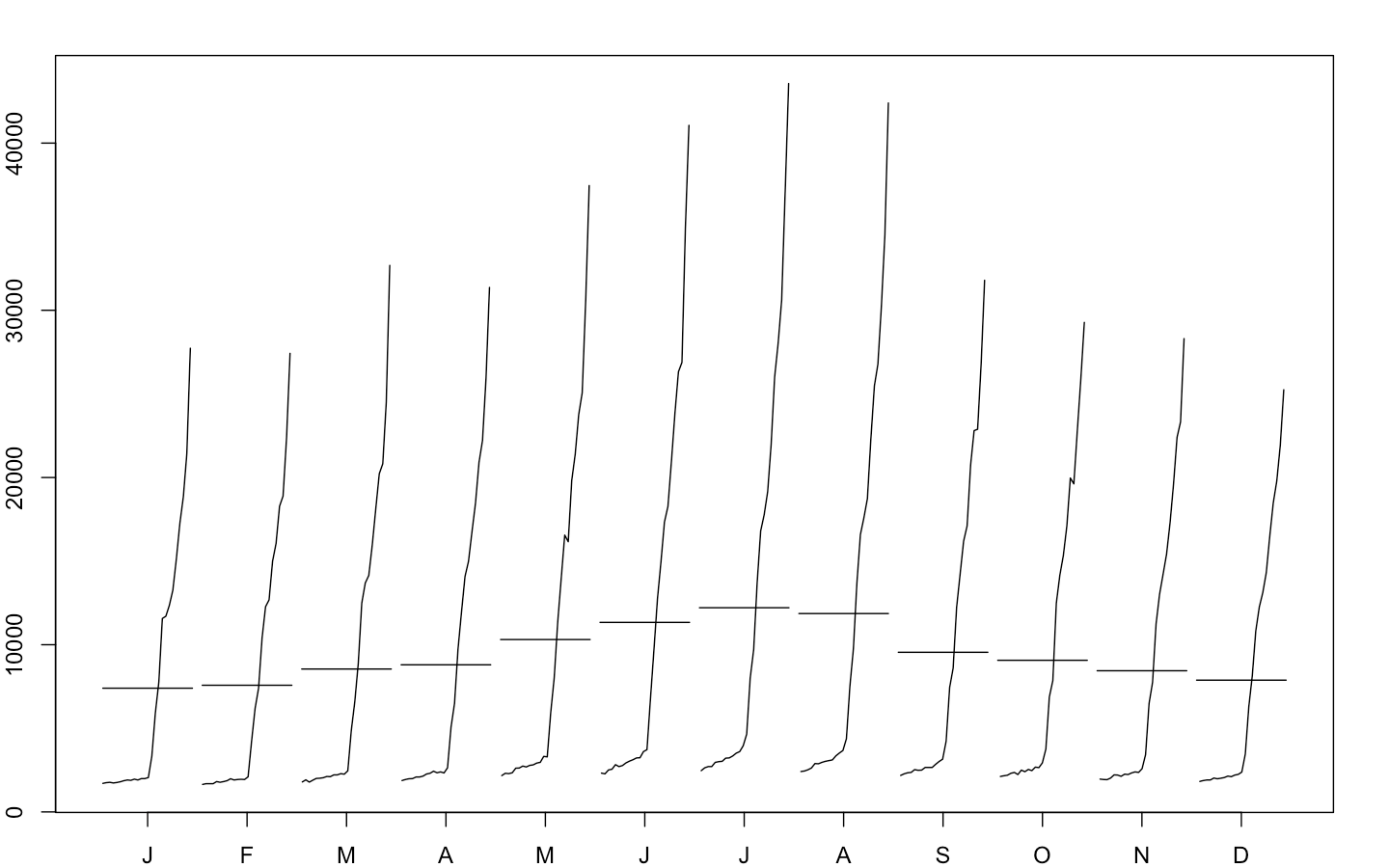
**Read the data as time series object**

gas\_ts = ts(data, start = c(1956,1), frequency = 12)

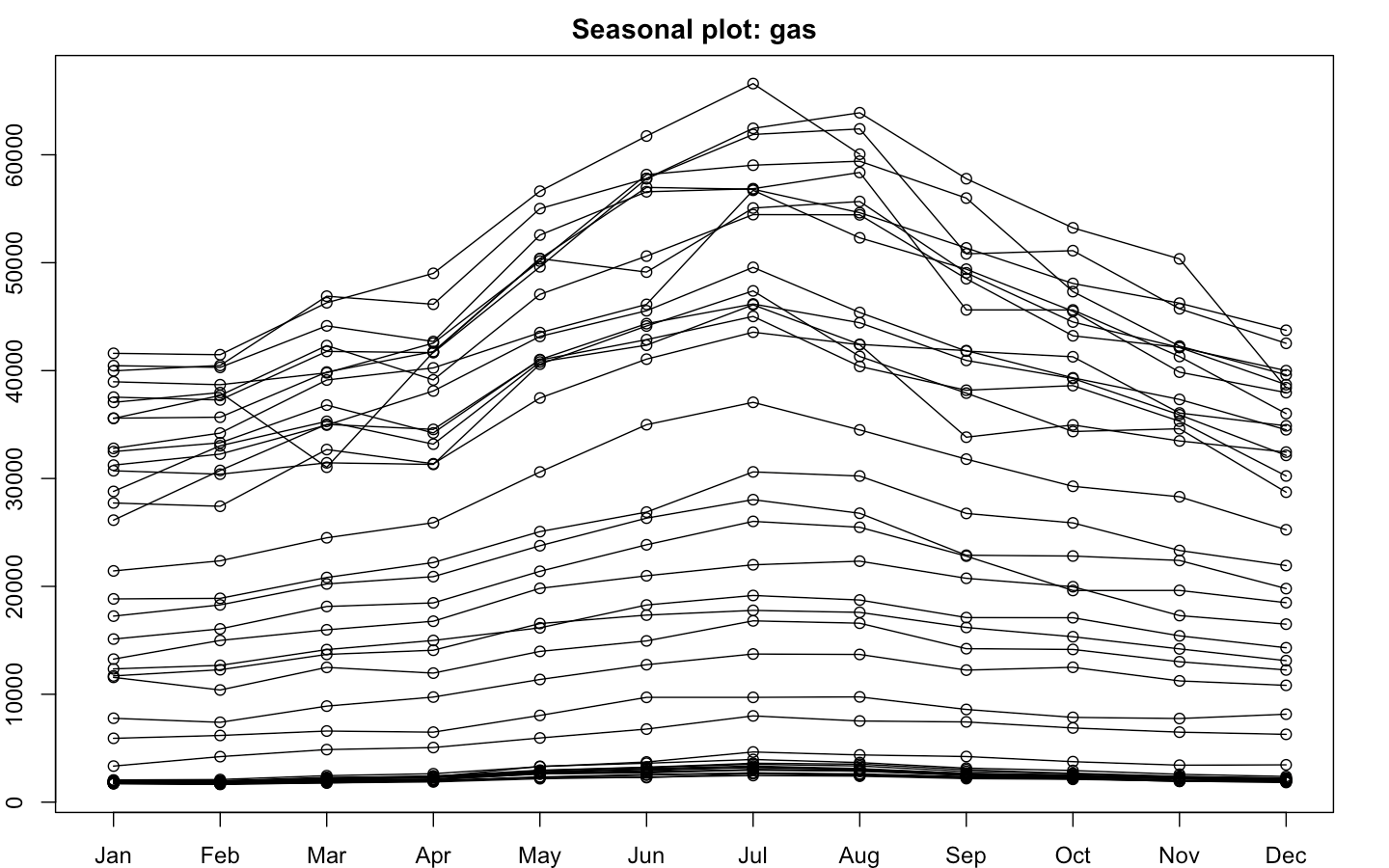
**Plot the dataset**

plot(gas\_ts)

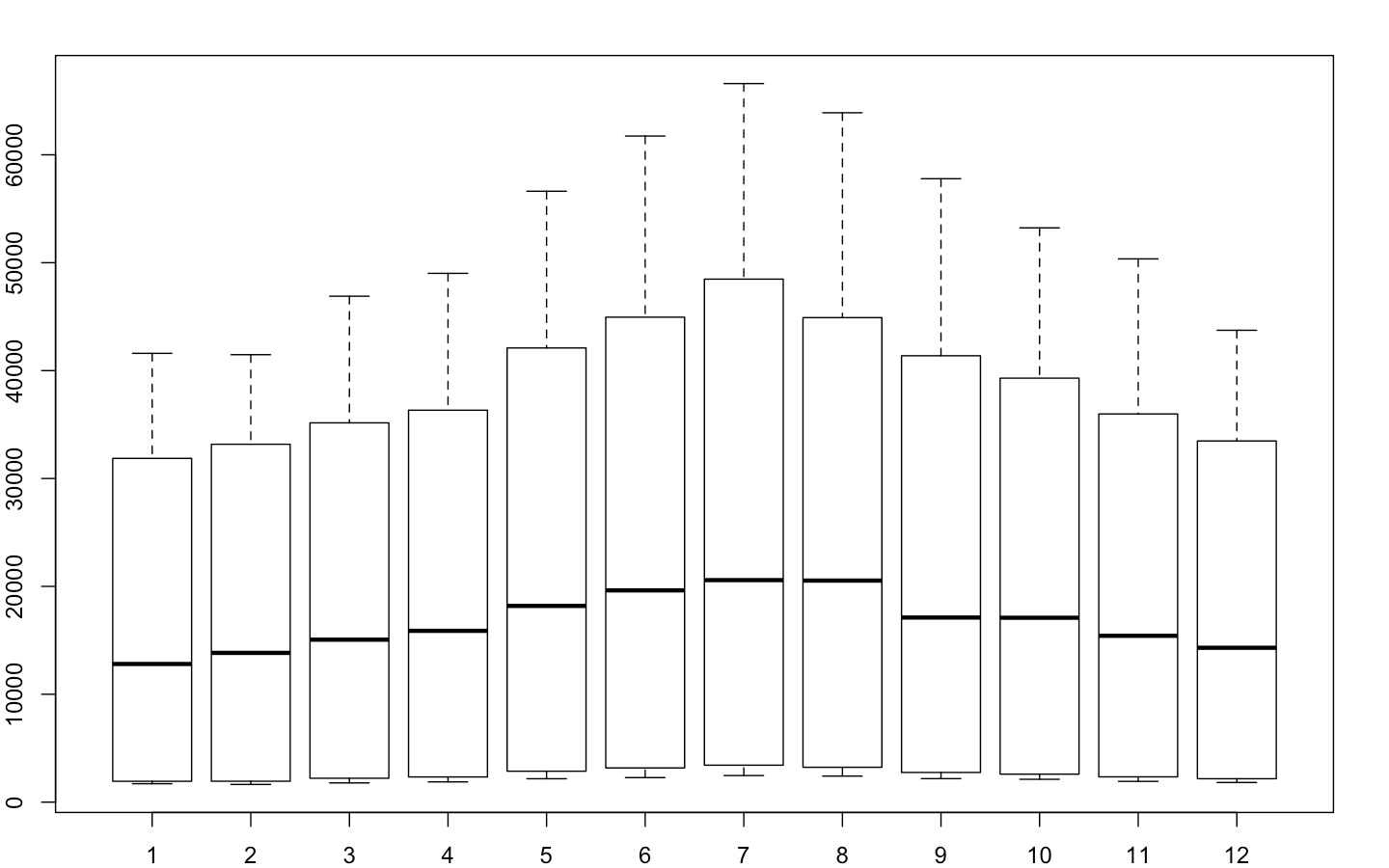
Timeseries - Gas

Monthplot - Gas

SeasonalPlot – Gas

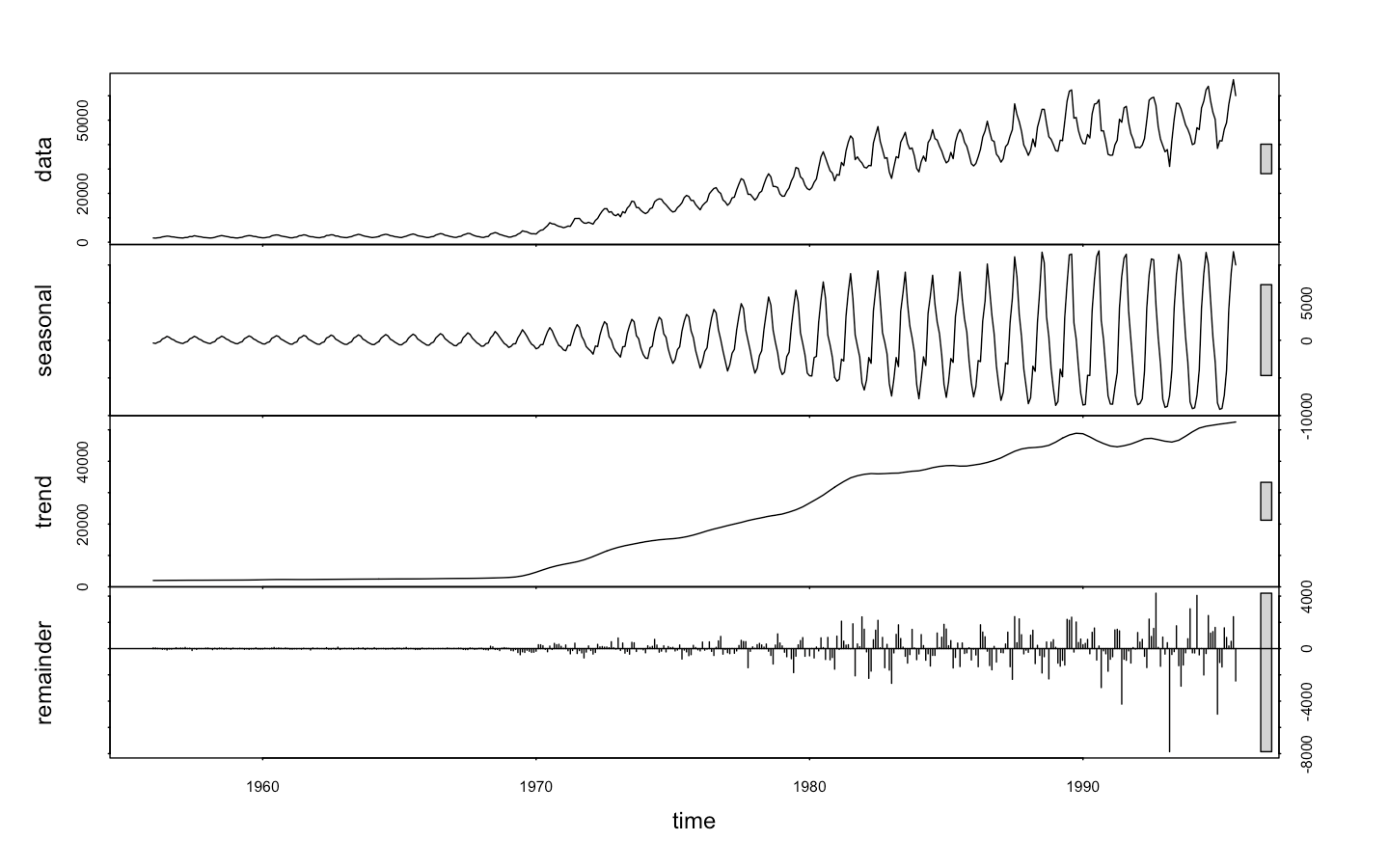


BoxPlot - Gas



* There are no missing values in the time series data
* From 1956 to 1970 there is not much movement in data and series seem flat
* There is an increasing linear trend in the series from 1970 onwards
* There is also seasonality present in the data, since the production increases during May, June, July and August during almost every year
* Seasonality is significantly varying with time, therefore the series has multiplicative seasonality
* The series is non-stationary because its mean and variance is not constant

### What components are present in the time series

Decomposition Plot - Gas

* Monthly Gas production is a combination of Trend, Seasonality and Irregular component
* There is a definite upward movement YOY
* Seasonal fluctuations increasing as total production increases
* Therefore, the Seasonality is multiplicative
* Hence the model is multiplicative as below
* However, Trend is contributing more as compared to Seasonality to explain the time series

Yt =Tt \*St \*It

Production = Trend \*Seasonality \* Error

### What is the periodicity of dataset

* The periodicity of data set is monthly and the data is present from January , 1956, to August, 1995

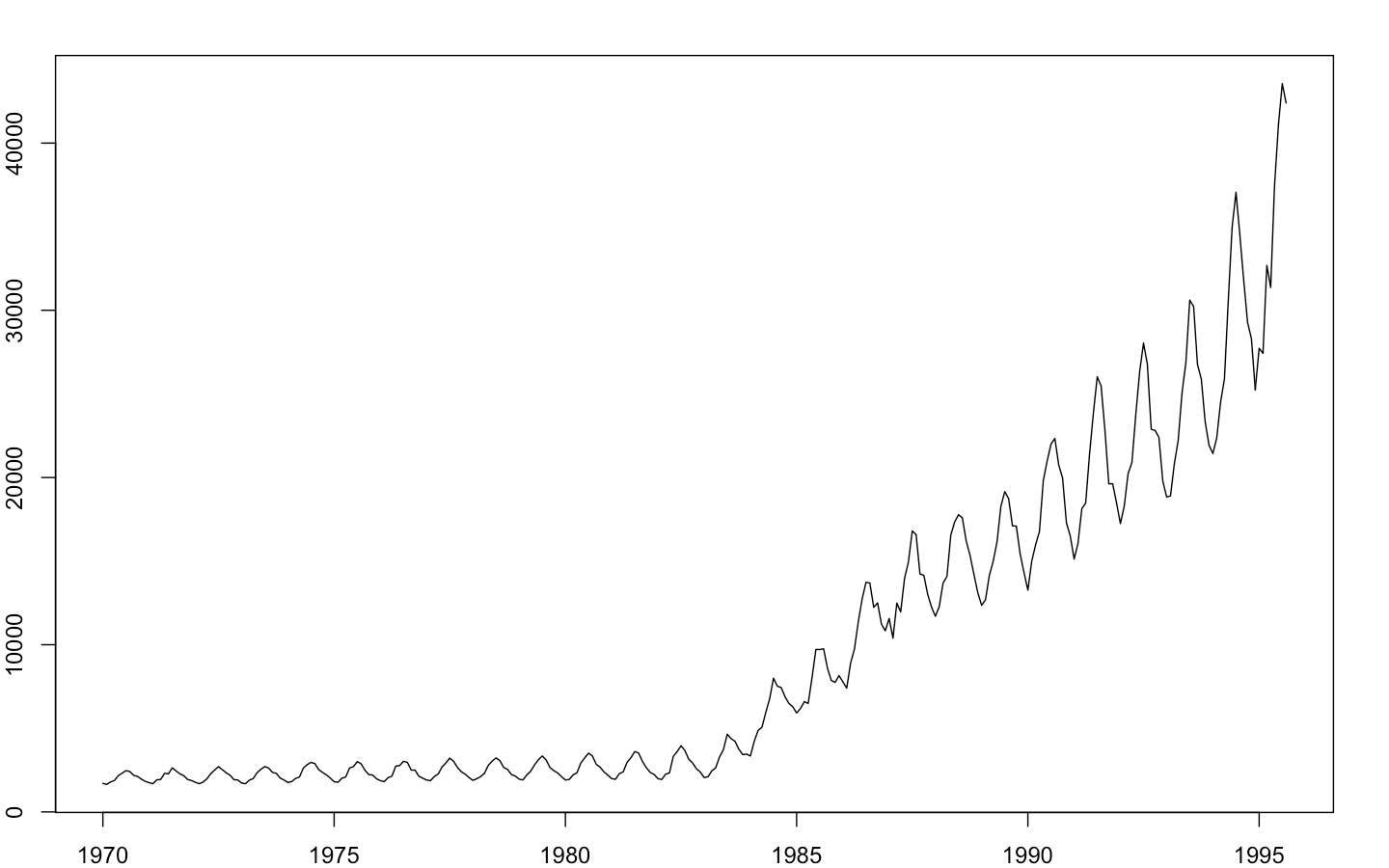
*1956 1 2 3 4 5 6 7 8 9 10 11 12*

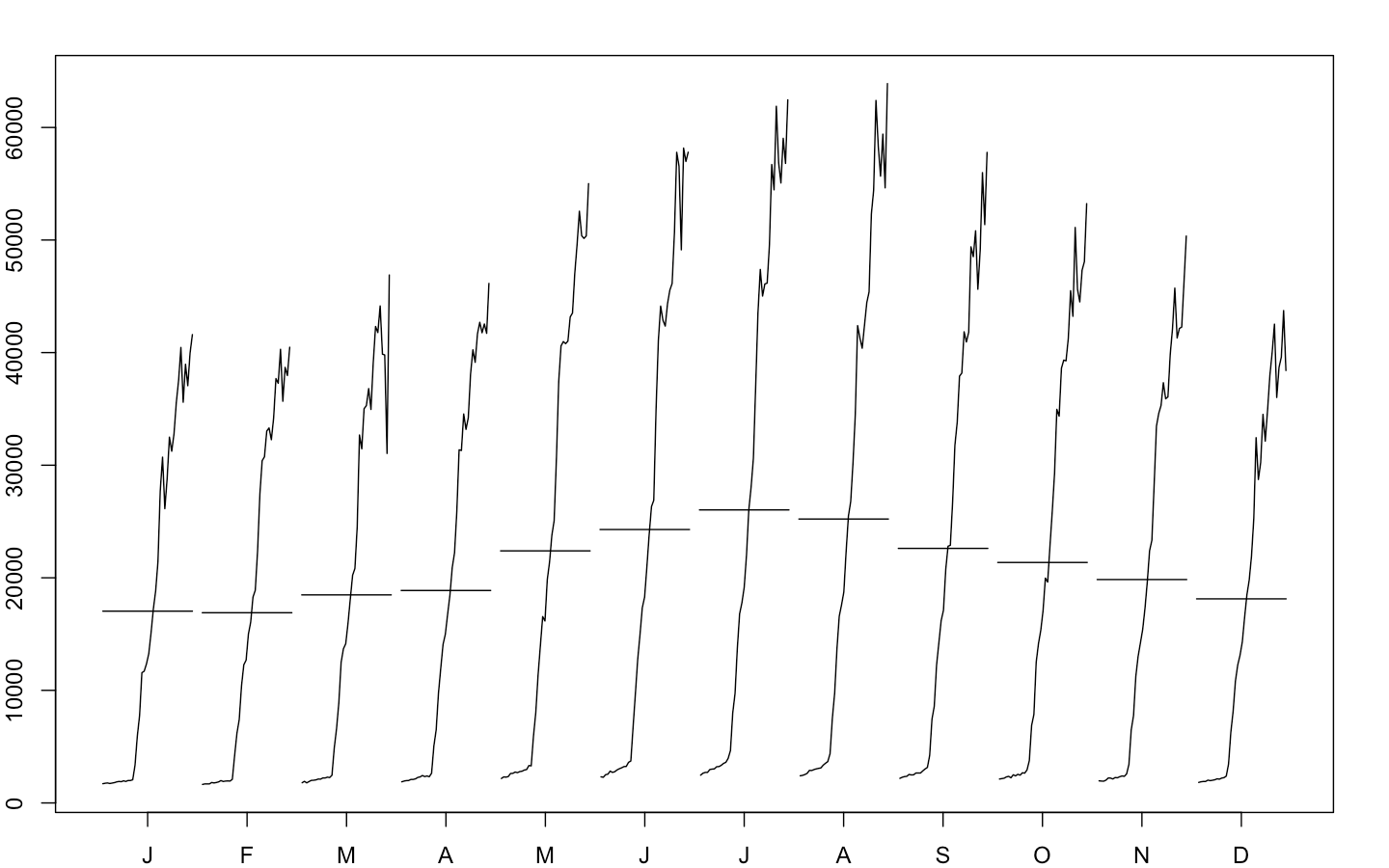
*1957 1 2 3 4 5 6 7 8 9 10 11 12*

*---------------------------------------------------------*

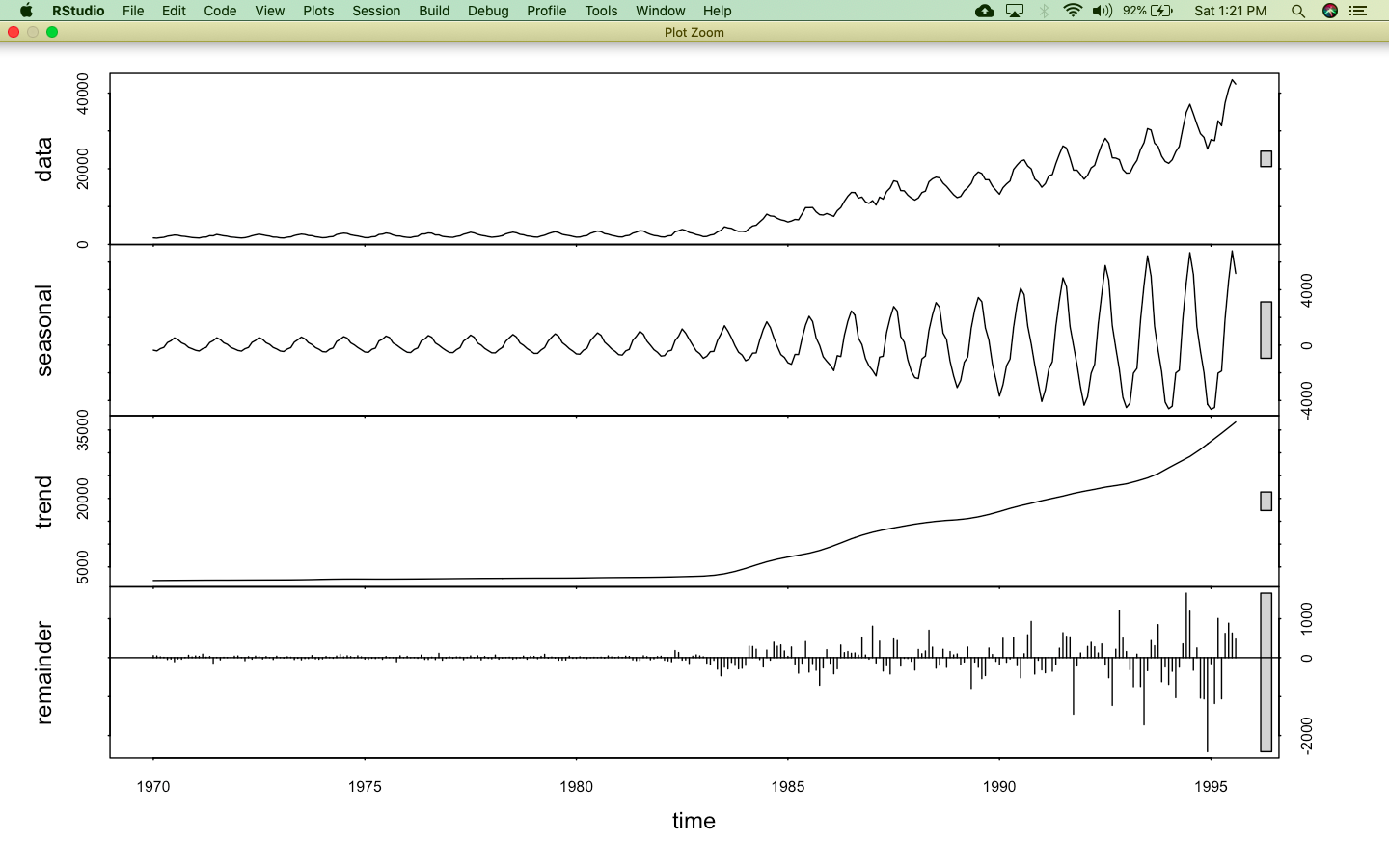
*1994 1 2 3 4 5 6 7 8 9 10 11 12*

*1995 1 2 3 4 5 6 7 8*

Timeseries – Gas : Year 1970 Onwards

Monthplot– Gas : Year 1970 Onwards

Decomposition Plot – Gas : Year 1970 Onwards



### Is the time series Stationary?

For time series to be stationary it should have below properties:-

* Mean of the time series should be 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.

* Variance of the time series should be constant
* The correlation between the t-th term in the series and the t+m-th term in the series should be constant for all time periods and for all m

#### Inspect Visually basis above graphs:

* It can be seen from the above time series plots that mean and variance are not constant
* Also there is trend and seasonality present in our data which affect the value of the series at different times
* The correlation between the tth term and t +mth term does not seem to be constant as the time series has a linear increasing trend and hence correlation will keep on increasing or decreasing depending upon the chosen time point
* A stationary time series is one whose properties do not depend on the time at which the series is observed. However, our timeseries depends on the time periods at which it is observed. For example, post 1985 there is an increasing trend whereas before 1985 the time series is flat
* Also a stationary time series does not have any predictable pattern. But our series does have a predictable increasing linear pattern and seasonality.
* Hence the time series in question is non-stationary.

#### Conduct an ADF test.Write down the null and alternate hypothesis for the stationarity test

**Augmented Dickey-Fuller Test**

* Tests whether a time series is NON-STATIONARY
  + Null hypothesis H0: Time series non-stationary
  + Alternative hypothesis Ha: Time series is stationary
* Rejection of null hypothesis implies that the series is stationary

Augmented Dickey-Fuller Test Results

data: gas\_ts2

Dickey-Fuller = 0.73972, Lag order = 6, p-value = 0.99

alternative hypothesis: stationary

* Since the p value is larger than cutoff i.e. 5% at 95% CL, we fail to reject the null hypothesis and hence conclude that the time series is non - stationery.

**Stationarize the time series**

gas\_diff = diff(gas\_ts2)

data: gas\_diff

Dickey-Fuller = -15.575, Lag order = 6, p-value = 0.01

alternative hypothesis: stationary

* Here we have stationarised the time series by differencing it at order 1 and conducted ADF test again and since p value is significantly lower then 5% at 95% CL, we can conclude that series is stationary.

### De-seasonalise the series if seasonality is present

1. **Decompose the time series**

* We decompose the time series in its systematic and irregular components using Stl command and argument s.window = 5 since seasonality is changing over time.

Call:

stl(x = gas\_ts2, s.window = 5)

Components

seasonal trend remainder

Jan 1970 -318.15778 2020.396 6.761473e+00

Feb 1970 -385.56584 2024.782 6.784331e+00

Mar 1970 -208.27565 2029.167 -2.689106e+01

Apr 1970 -153.98138 2033.552 -1.570518e+00

May 1970 166.67428 2037.937 -3.161138e+01

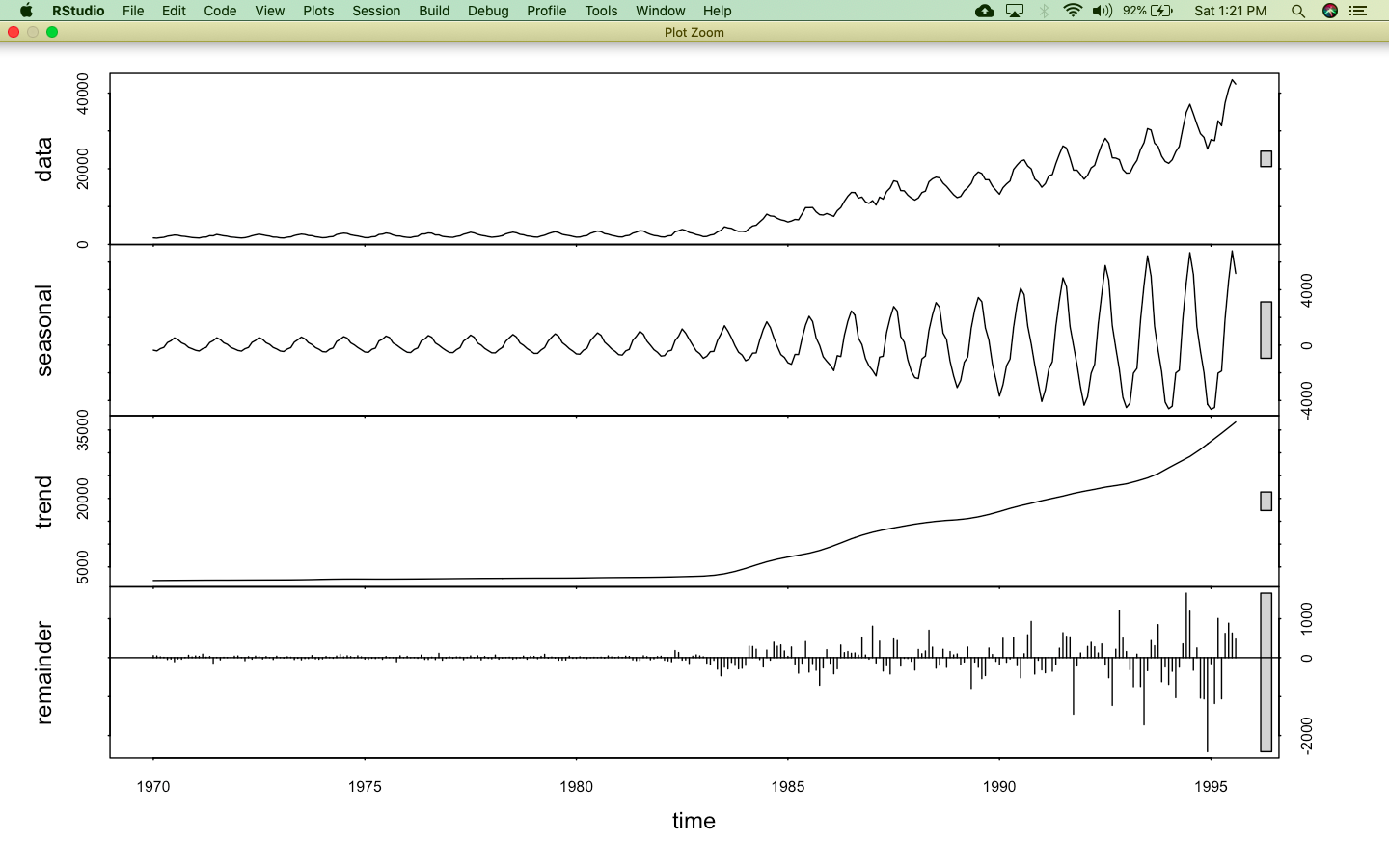
Jun 1970 238.07324 2042.205 4.072222e+01

------ continued

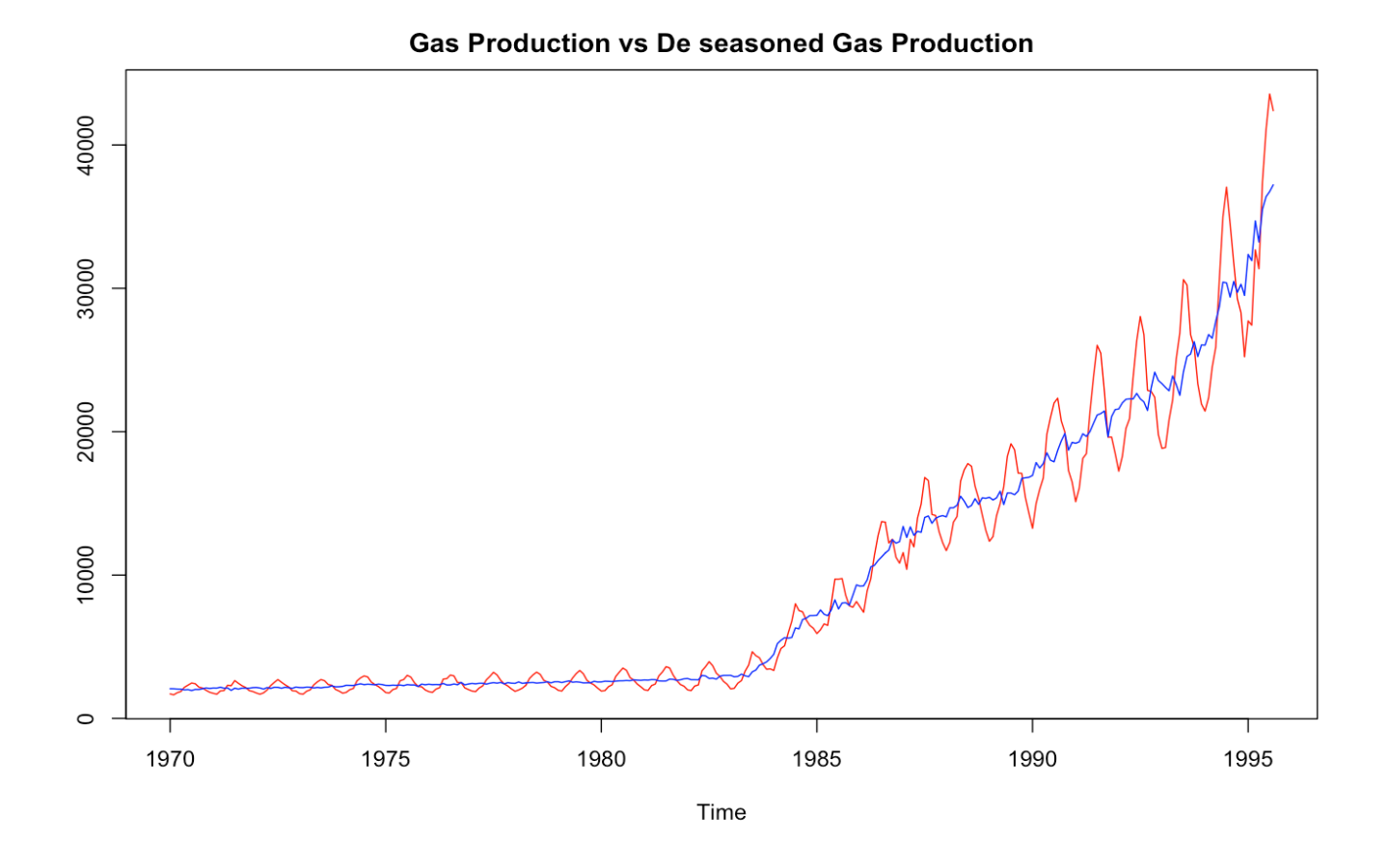
1. **Deseasonalise the time series**

* We now remove the seasonality component from the time series to remove the impact of seasonality on the series.

Deseason\_gas = (decomp2$time.series[,2] + decomp2$time.series[,3])

Decomposition Plot – Gas : Year 1970 Onwards

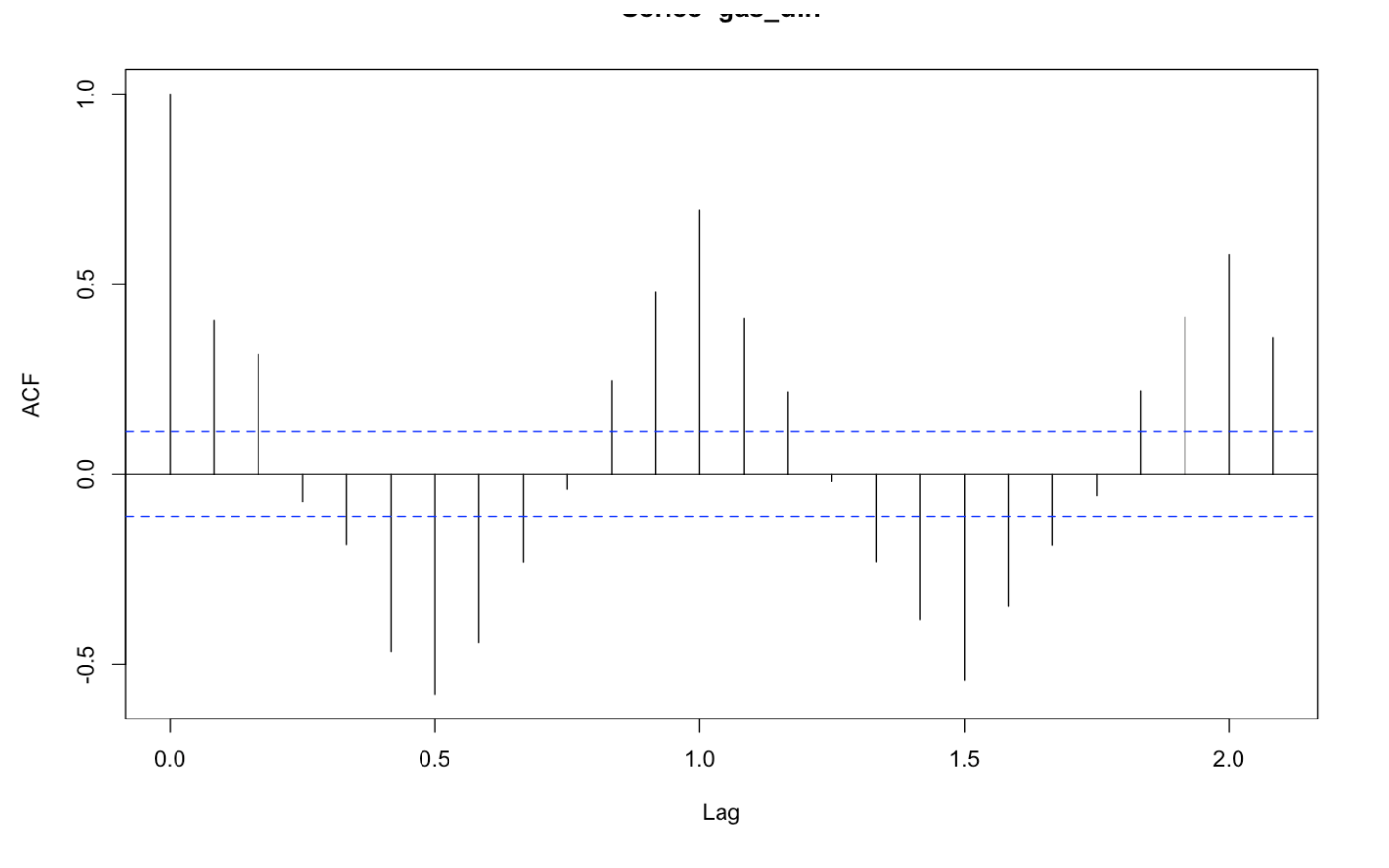
Original Series versus Deseasonalised Time series



* From the decomposition plot we can derive that Trend has a more defining role in our series as compared to Seasonality
* From the original vs deseasonalised plot we can notice a clear linearly increasing trend in the series, once seasonality is removed

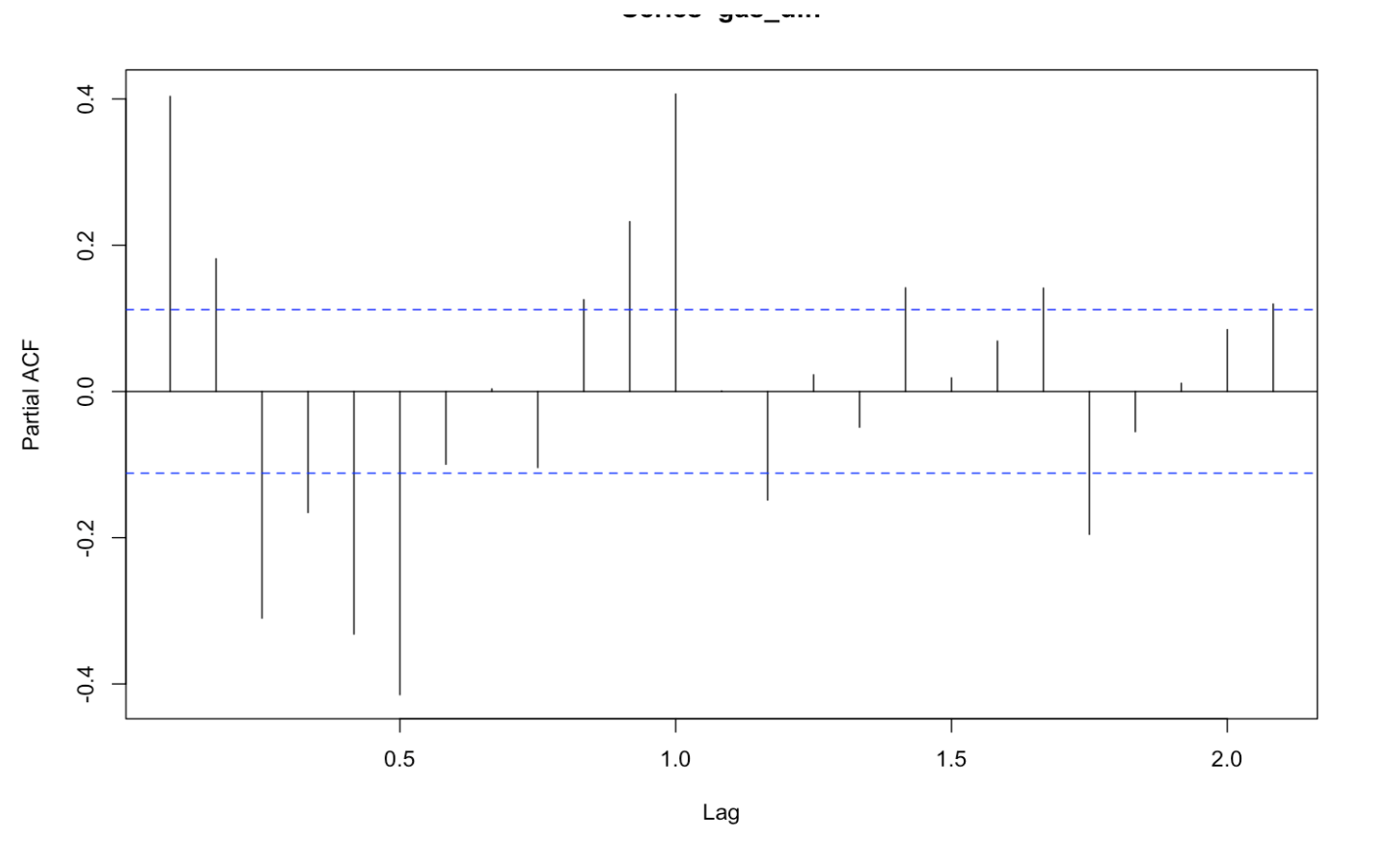
Auto Correlation And Partial Correlation Plots

* The ACF and PACF plots are observed on the differenced/ stationarised Time series
* General ARIMA model is based on below concept:
  + AR: Current observation is regressed on past observations – inputted in the model as a value for p , that is no of past significant partial autocorrelations to be considered in the ARIMA model
  + MA: Current observation is regressed on past forecast errors - inputted in the model as a value for q, no of past significant forecast errors to be considered in the model
  + I : Integrated or stationary time series – inputted in model by parameter d or order of differencing to stationarise the time series
* For ARIMA(p, d, q) identifies a non-seasonal model which needs to be differenced d times to make it stationary and contains p AR terms and q MA terms

ACF Plot 

From the ACF plot, there is a cut off after lag 2. This implies that q=2.

PACF Plot

PACF cuts off after lag 6, hence p = 6

Therefore our parameters for fitting an ARIMA Models are as below

P -Order of Auto regression – PACF = 6

D – Order of Differencing. = 2

Q – Order of moving average - ACF = 2

Data Partitioning

gasTStrain = window(Deseason\_gas, start=1970, end=c(1993,12))

gasTStest= window(Deseason\_gas, start=1994)

* Data in train – 1970 to 1993
* Data in test – 1994 onwards

Model Building

# ARIMA Model

**Steps**

* Identify p, d, q parameters
  + Test if the data is stationary, if it not then stationarize it by differencing the series
    - On the stationarised data check ACF and PACF plots to identify q and p parameters
    - Finalise p, d and q parameters from above steps
* Test if the original data has seasonality, if yes deseasonalize the data
* Partition the deseasonalized data as train and test data
* Create the model on train data using p, d and q parameters from the above steps
* Evaluate Model Validity by testing the residuals independence
* Evaluate the Model accuracy on Test data

### ARIMA Model Summary

*Call:*

*arima(x = gasTStrain, order = c(6, 2, 2))*

*Coefficients:*

*ar1 ar2 ar3 ar4 ar5 ar6 ma1*

*-0.1675 -0.2095 -0.0517 -0.0196 0.1194 0.0823 -1.0873*

*s.e. 0.4357 0.1590 0.1433 0.0931 0.0822 0.0790 0.4339*

*ma2*

*0.1374*

*s.e. 0.4001*

*sigma^2 estimated as 83931: log likelihood = -2028.73, aic = 4075.4*

* Akaike’s Information Criterion (AIC) for our ARIMA model is 4075.4.We will use this to choose between various models. The model with the lowest AIC will be shortlisted.

### Model Validity

#### Box-Ljung test:

*#H0: Residuals are Independent*

*#Ha: Residuals are not Independent*

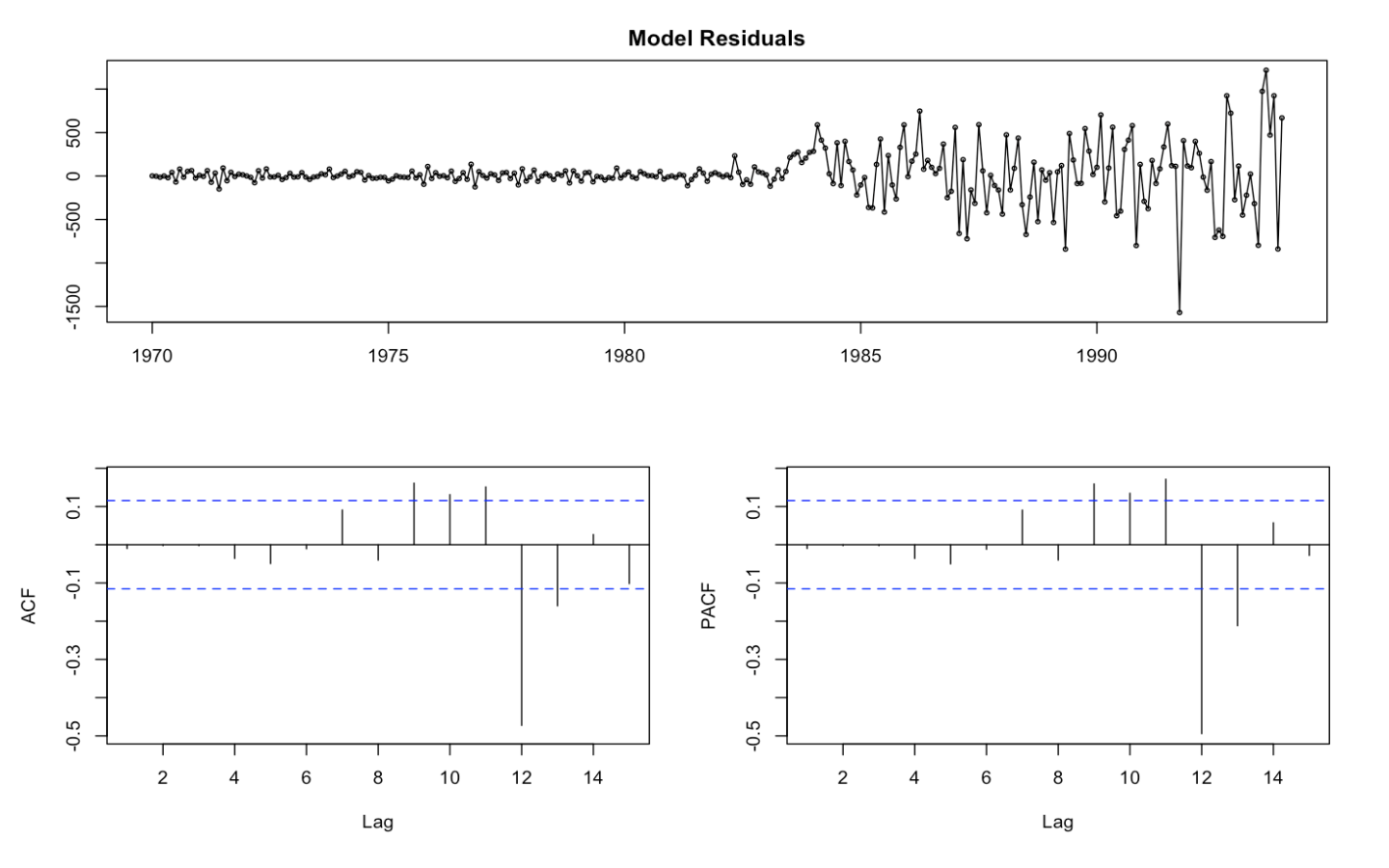
Box-Ljung test

data: gasARIMA$residuals

X-squared = 0.028413, df = 1, p-value = 0.8661

* The p-values for the Ljung-Box test all are well above 0.05, indicating non-significance and hence null hypothesis remains.

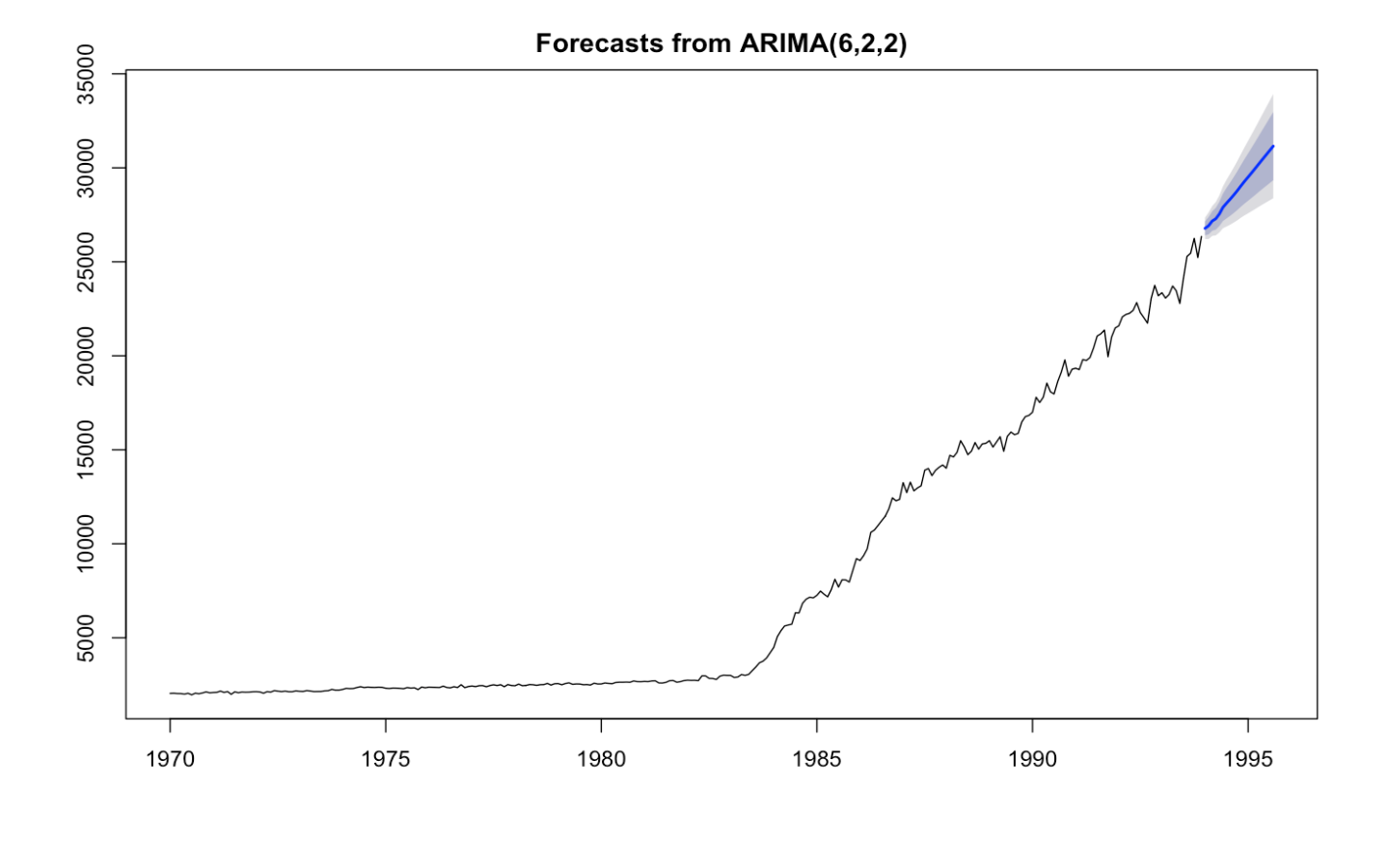
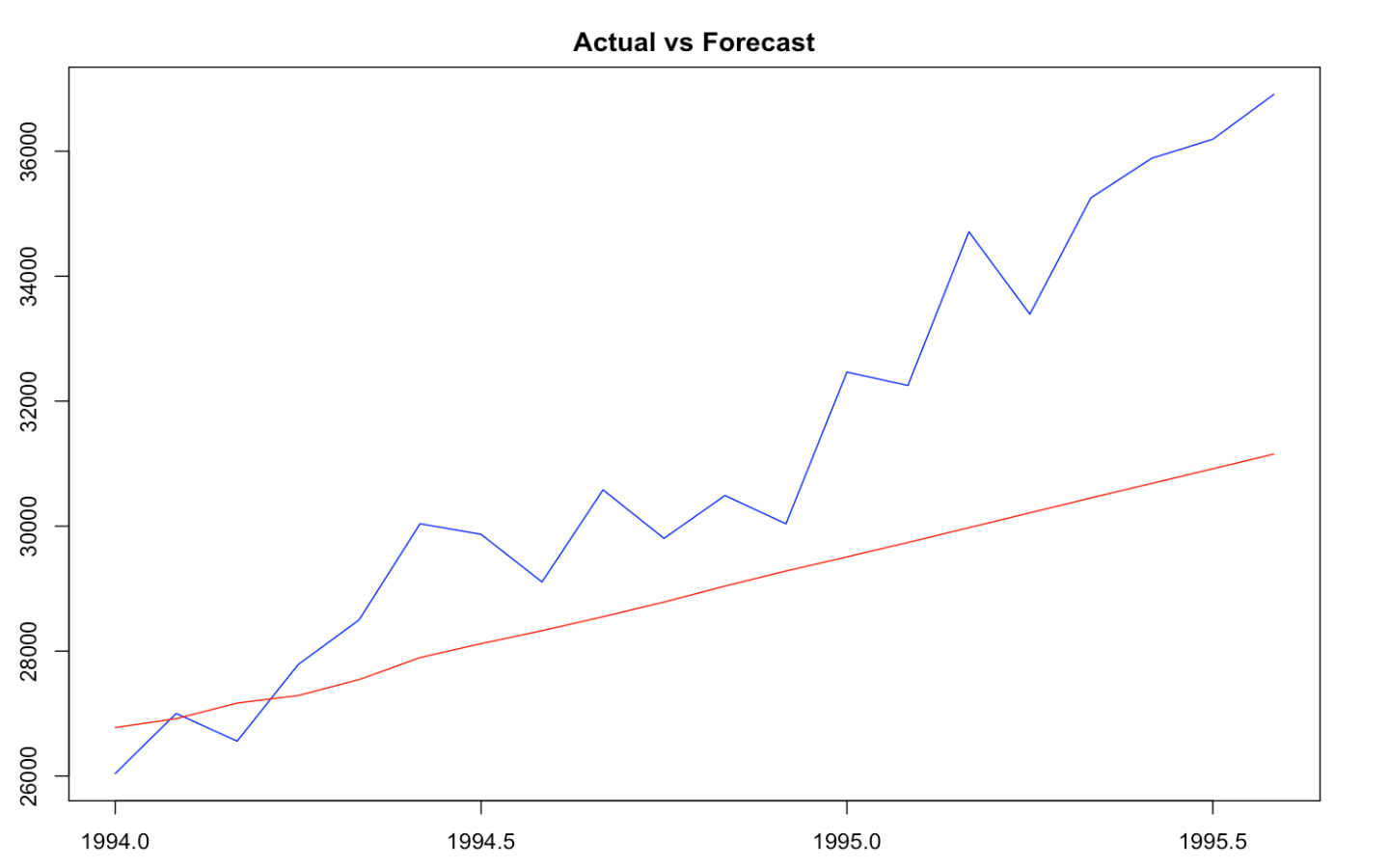
#### Check Model Residuals graphs for independence



* The residual values are normal as they rest on a line and aren’t all over the place. They don’t show any trend either.
* The ACF of the residuals shows no significant autocorrelations.
* As all the graphs are in support of the assumption that there is no pattern in the residuals, we can go ahead and calculate the forecast.

### Forecasting on Test Data

ARIMA Forecast Plots

**

* *The forecasted line is quite deviated from the actual line on the test data*

### Arima Model Accuracy

*ME RMSE MAE MPE MAPE*

*Training set 20.29908 288.7012 168.2114 0.317523 2.021709*

*Test set 2226.59855 2973.4311 2361.2744 6.583129 7.095742*

*MASE ACF1 Theil's U*

*Training set 0.1709974 -0.009932642 NA*

*Test set 2.4003826 0.710082936 2.288605*

* The Mape or mean absolute percentage error for Arima model is 7.09% on test data
* The Mean absolute error is 2361 on test data
* Root mean squared error is 2973 on test data

# AUTO ARIMA Model

**Steps**

* Test if the original data has seasonality, if yes deseasonalize the data
* Partition the deseasonalized data as train and test data
* Create the model on train data
* Evaluate Model Validity by testing the residuals
* Evaluate the Model accuracy on Test data

### AUTO ARIMA Model Summary

*Series: gasTStrain*

*ARIMA(0,2,3)*

*Coefficients:*

*ma1 ma2 ma3*

*-1.2505 0.1586 0.1419*

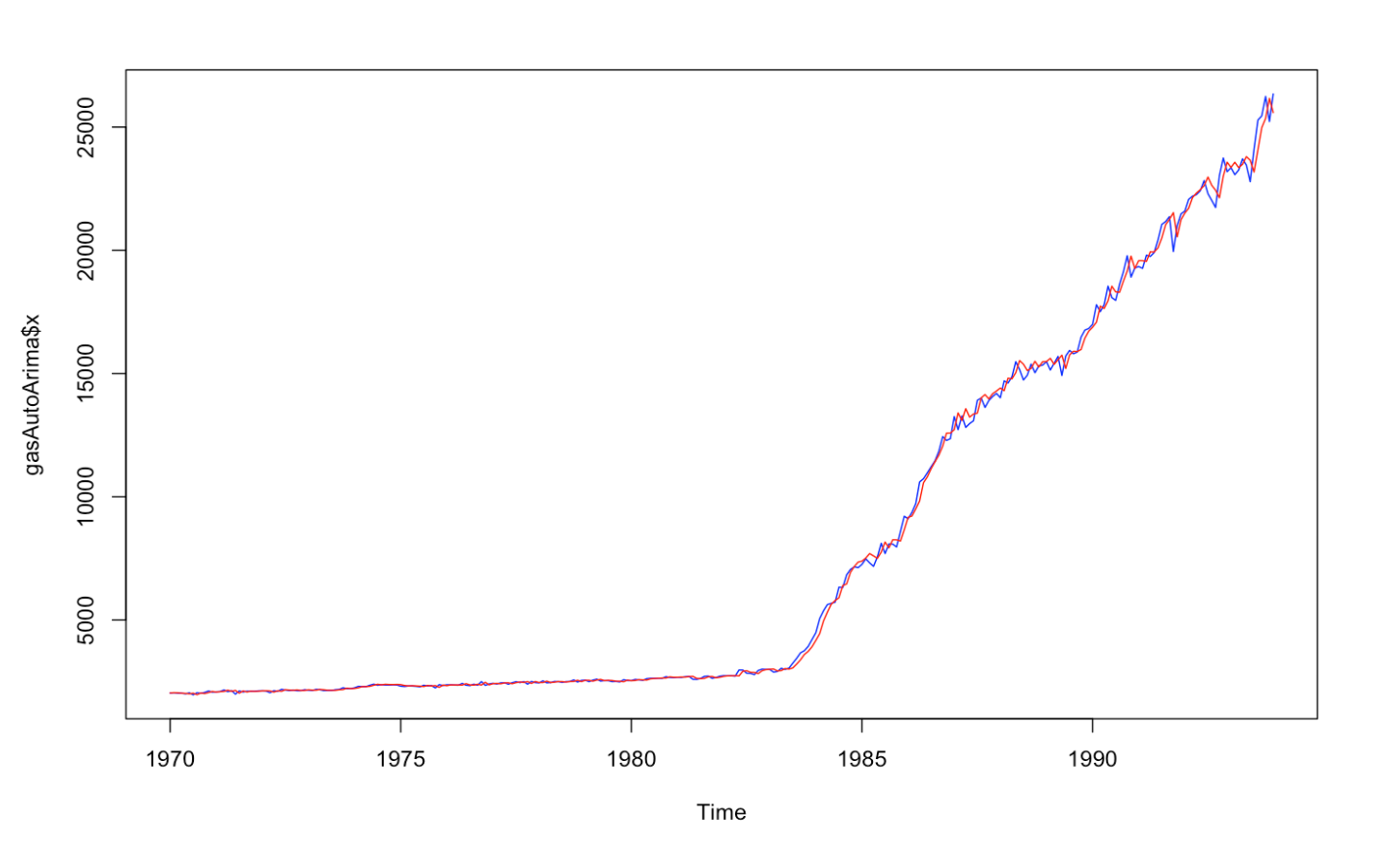
*s.e. 0.0607 0.0983 0.0581*

*sigma^2 estimated as 86437: log likelihood=-2031.34*

*AIC=4070.69 AICc=4070.83 BIC=4085.31*

* Akaike’s Information Criterion (AIC) for our AUTO ARIMA model is 4070.8.This is lower than manual ARIMA model.
* The p, d, q values identified by auto arima are 0,2,3

Actual vs Fitted values plot



* Actual and fitted values on train data are perfectly overlapping

### Model Validity

#### Box-Ljung test:

*#H0: Residuals are Independent*

*#Ha: Residuals are not Independent*

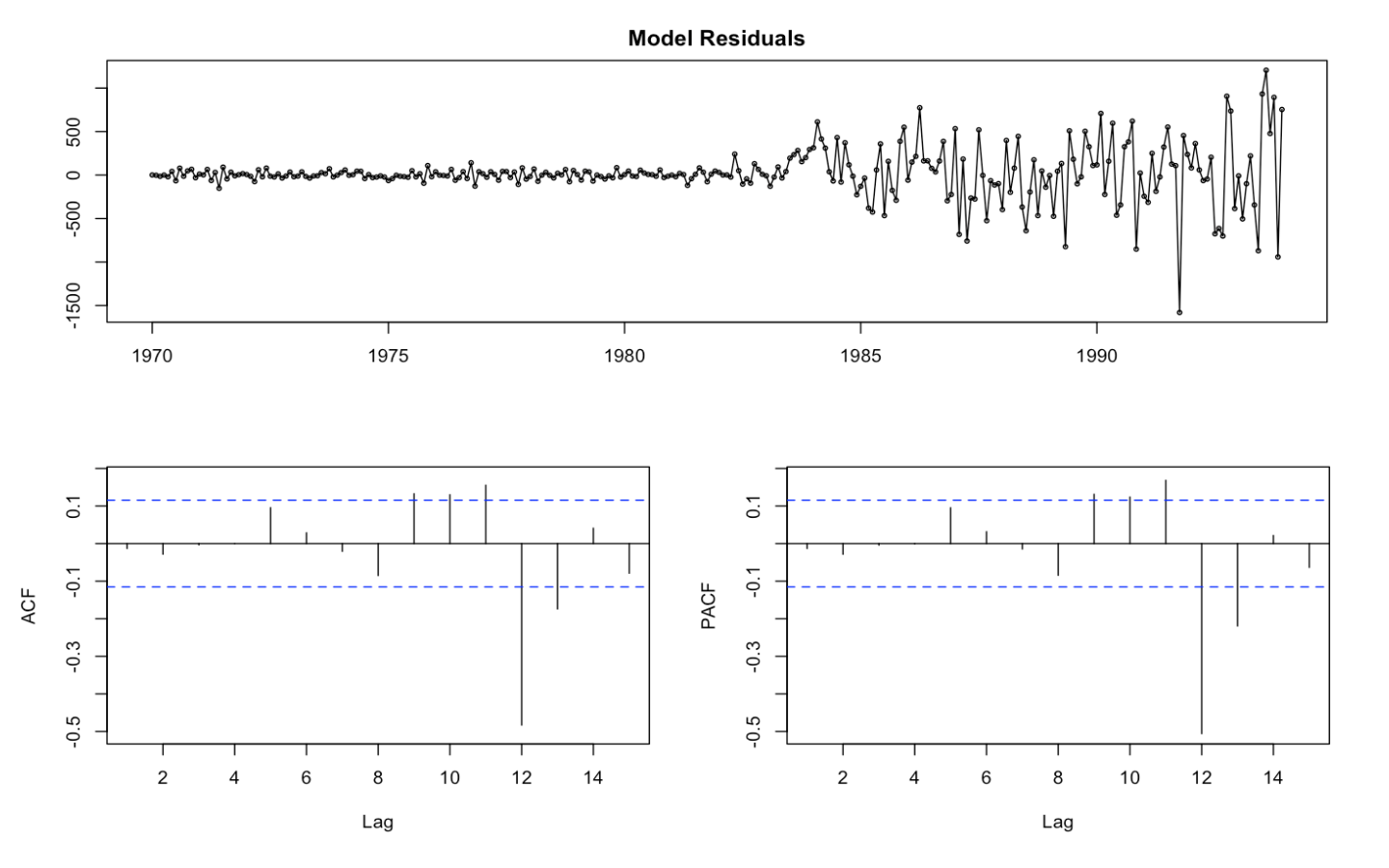
Box-Ljung test

data: gasAutoArima$residuals

X-squared = 0.049604, df = 1, p-value = 0.8238

* The p-values for the Ljung-Box test all are well above 0.05, indicating non-significance and hence null hypothesis remains.

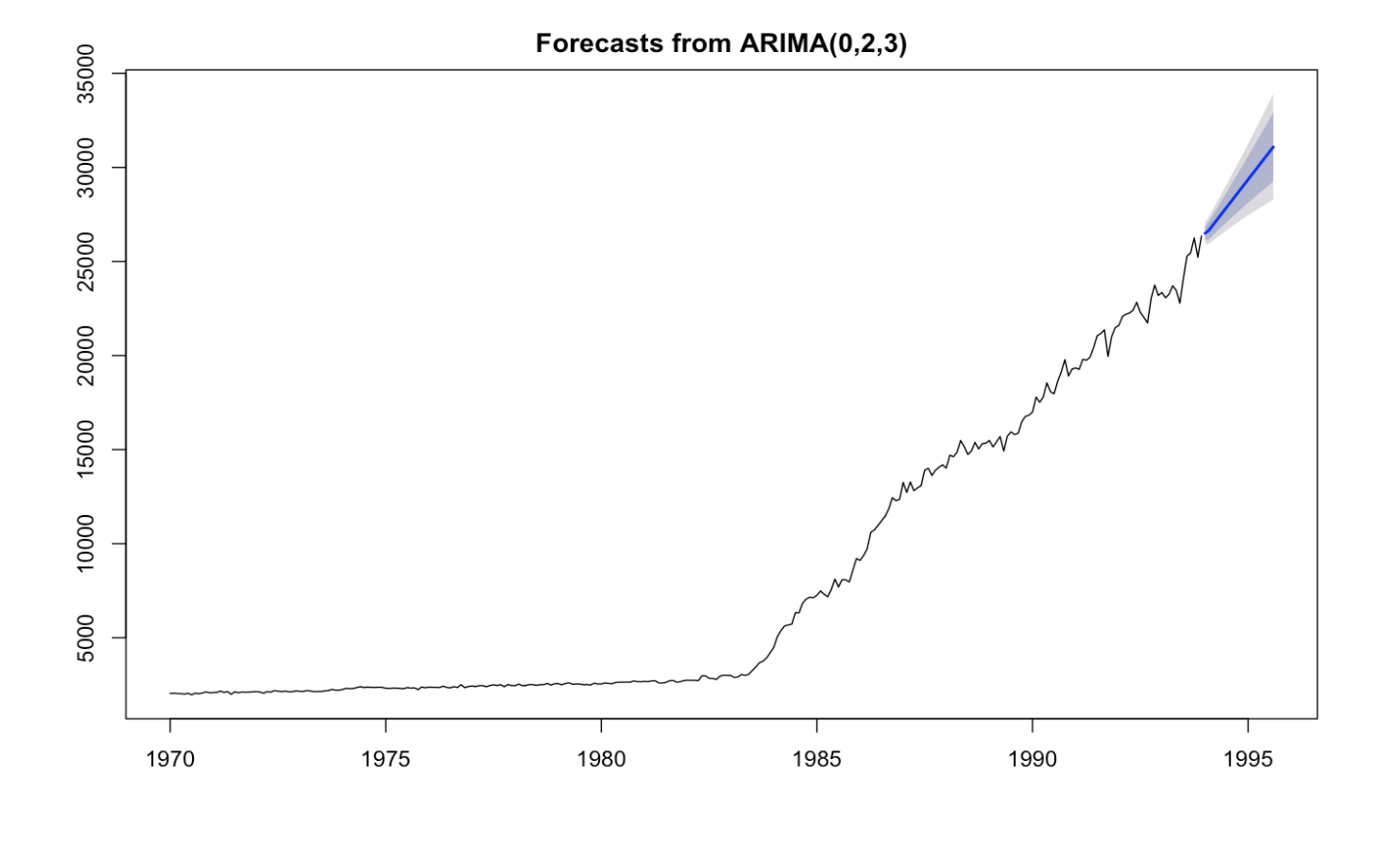
#### Check Model Residuals graphs for independence

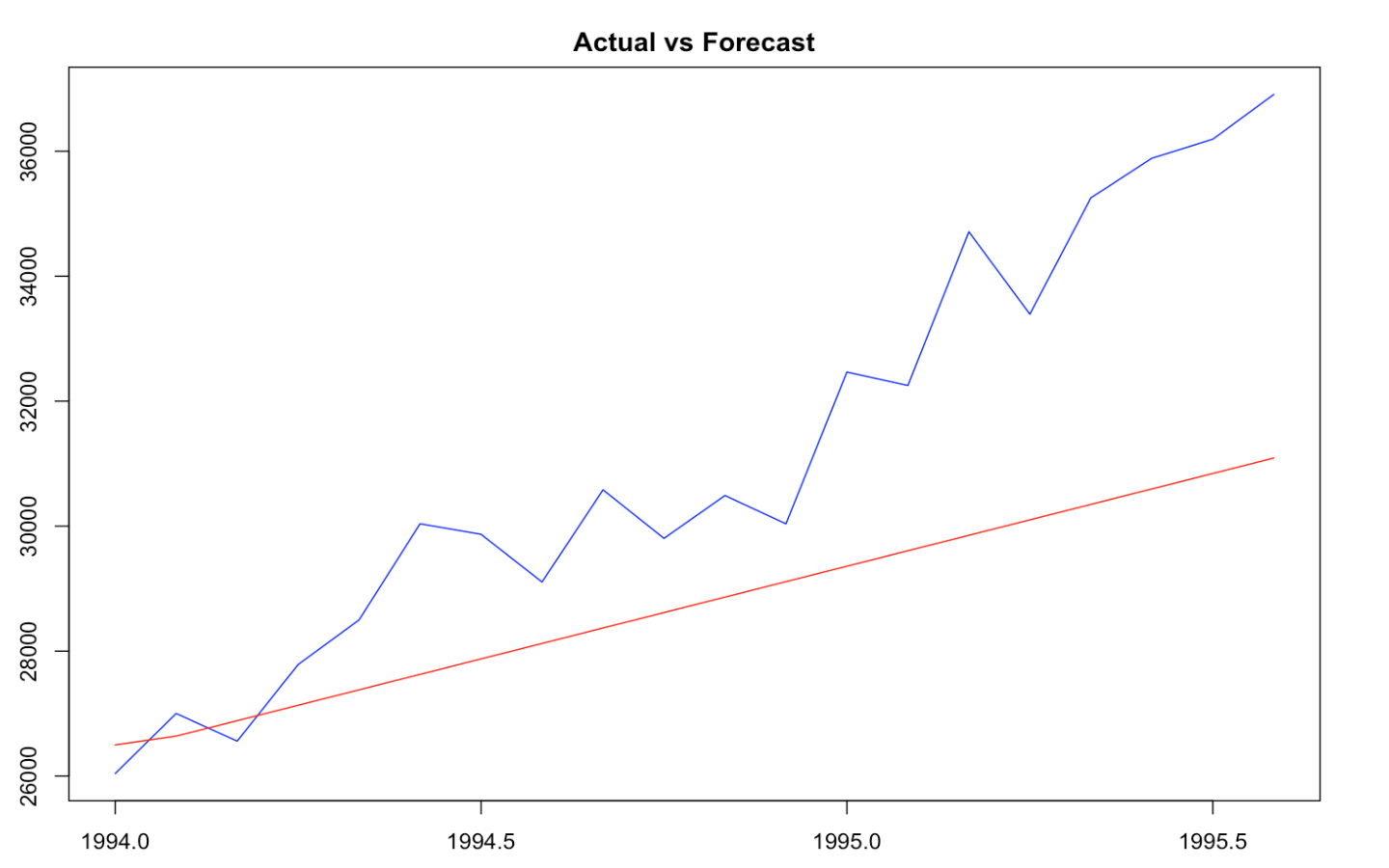


* The residual values are normal as they rest on a line and aren’t all over the place. They don’t show any trend either.
* The ACF of the residuals shows no significant autocorrelations.
* As all the graphs are in support of the assumption that there is no pattern in the residuals, we can go ahead and calculate the forecast.

### Forecasting on Test Data

AUTO ARIMA Forecast Plots





* *The forecasted line is quite deviated from the actual line on the test data*

### Auto Arima Model Accuracy

ME RMSE MAE MPE MAPE

Training set 17.15253 291.4387 169.5996 0.2831761 2.039798

Test set 2397.40487 3068.2542 2476.0176 7.1573524 7.456784

MASE ACF1 Theil's U

Training set 0.1724085 -0.01312389 NA

Test set 2.5170263 0.70277618 2.374159

* The Mape or mean absolute percentage error for Auto Arima model is 7.45% on test data, slightly higher then Manual Arima model
* The Mean absolute error is 2397 on test data
* Root mean squared error is 3068 on test data

# Holt - Winters Model

**Steps**

* Identify, if the data has both trend and seasonality, for holt winters to be applicable, both should be present in data.
* Partition the Original data as train and test data
* Identify whether additive or multiplicative model works
* Create the model on train data
* Evaluate Model Validity by testing the residuals independence
* Evaluate the Model accuracy on Test data

### Holt Winters Model Summary

Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:

HoltWinters(x = as.ts(gasTStrain2), seasonal = "additive")

Smoothing parameters:

alpha: 0.2813768

beta : 0.06945073

gamma: 1

Coefficients:

[,1]

a 25149.9117

b 194.4332

s1 -4190.3876

s2 -3925.9030

s3 -1835.8374

s4 -555.0107

s5 2201.5462

s6 4073.7517

s7 7212.2391

s8 6095.5329

s9 2007.4591

s10 723.9971

s11 -1648.5060

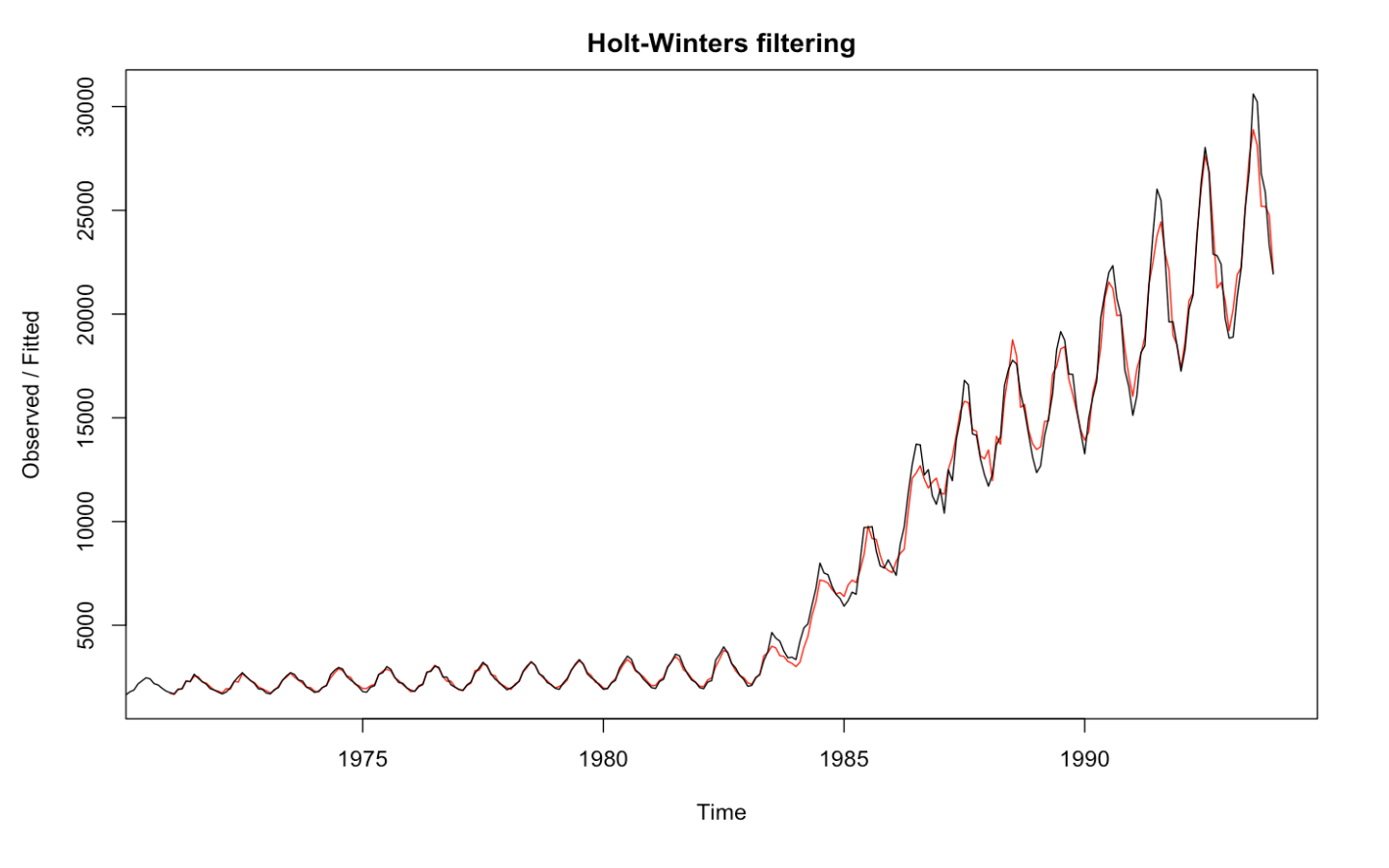
s12 -3219.9117

* Model is additive

▪ α: 0.28 - Smooths level;  
▪ β: 0.06 - Smooths trend;  
▪ γ : 1 - Smooths seasonality; - highest value for gamma indicates all fluctuations are due to seasonality

0 < α < 1 ; 0 < β < 1; 0 < γ < 1

Actual vs Fitted values plot



* Actual and fitted values on train data are overlapping quite well

### Model Validity

#### Box-Ljung test:

*#H0: Residuals are Independent*

*#Ha: Residuals are not Independent*

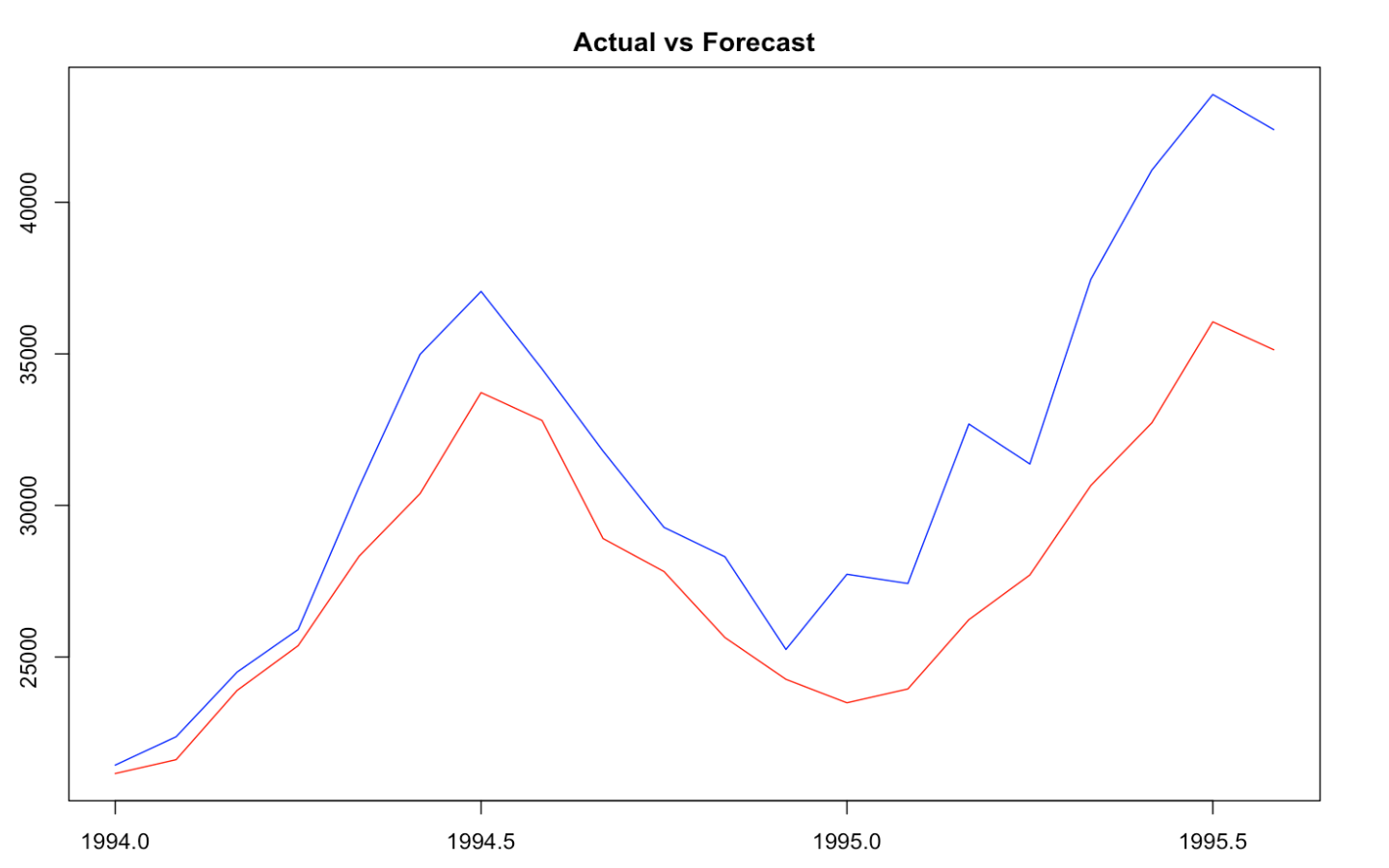
Box-Ljung test

data: hwBForecast$residuals

X-squared = 138.15, df = 20, p-value < 2.2e-16

* The p-values for the Ljung-Box test all are well below 0.05, indicating significance and hence null hypothesis is rejected, we conclude that residuals are not independent

### Forecasting on Test Data

HW Forecast Plots

* The trend and seasonality is captured better by holt winters model compared to the ARIMA models on Test data

### Holt winters Model Accuracy

ME RMSE MAE MPE MAPE

Training set 34.98522 549.0614 330.6199 0.2430721 3.965674

Test set 3490.84426 4301.7599 3490.8443 10.1655038 10.165504

MASE ACF1 Theil's U

Training set 0.3362248 0.3483775 NA

Test set 3.5500242 0.6596379 1.28634

* The Mape or mean absolute percentage error for HW model is 10.165% on test data, higher than both ARIMA model

Finalizing The Best Model And Developing The Future Forecast

**Steps**

* Compare the model validation measures for all three models
* Compare the model accuracy measures for all three models
* Basis the above metrics shortlist the final model
* Create the final model of the shortlisted approach using the full data
* Predict for future time periods

# Model Validation Measures - All Models

#### Box-Ljung test:

*#H0: Residuals are Independent*

*#Ha: Residuals are not Independent*

|  |  |  |  |
| --- | --- | --- | --- |
|  | MANUAL ARIMA | AUTO ARIMA | HOLT WINTERS |
| P VALUE | 0.8661 | 0.8238 | 2.2e-16 |

* As indicated by a lower than 5% p value of HW - Holt winters model’s residuals are not independent , concluding that the HW model is not valid

# Model Accuracy Measures - All Models

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | MANUAL ARIMA | AUTO ARIMA | HOLT WINTERS |
| AIC | *4075.4* | *AIC=4070.69* | - |
| MAPE | *7.095742* | 7.456784 | 10.165504 |

* The AIC is lowest for AUTO ARIMA Model
* Mape is lowest for manual ARIMA, however AUTO ARIMA MAPE is close to manual ARIMA MAPE
* Therefore, we will go ahead with AUTO ARIMA model as our final model based on AIC and MAPE values

# Final Model Building

* We build the final AUTO ARIMA Model on the full deseasonalized data:

gasAutoArimafinal<-auto.arima(Deseason\_gas, seasonal=FALSE)

gasAutoArimafinal

Series: Deseason\_gas

ARIMA(3,2,1)

Coefficients:

ar1 ar2 ar3 ma1

-0.4759 -0.1468 -0.2243 -0.8792

s.e. 0.0609 0.0697 0.0594 0.0312

sigma^2 estimated as 134786: log likelihood=-2240.79

AIC=4491.57 AICc=4491.77 BIC=4510.19

### Model Validity

#### Box-Ljung test:

*#H0: Residuals are Independent*

*#Ha: Residuals are not Independent*

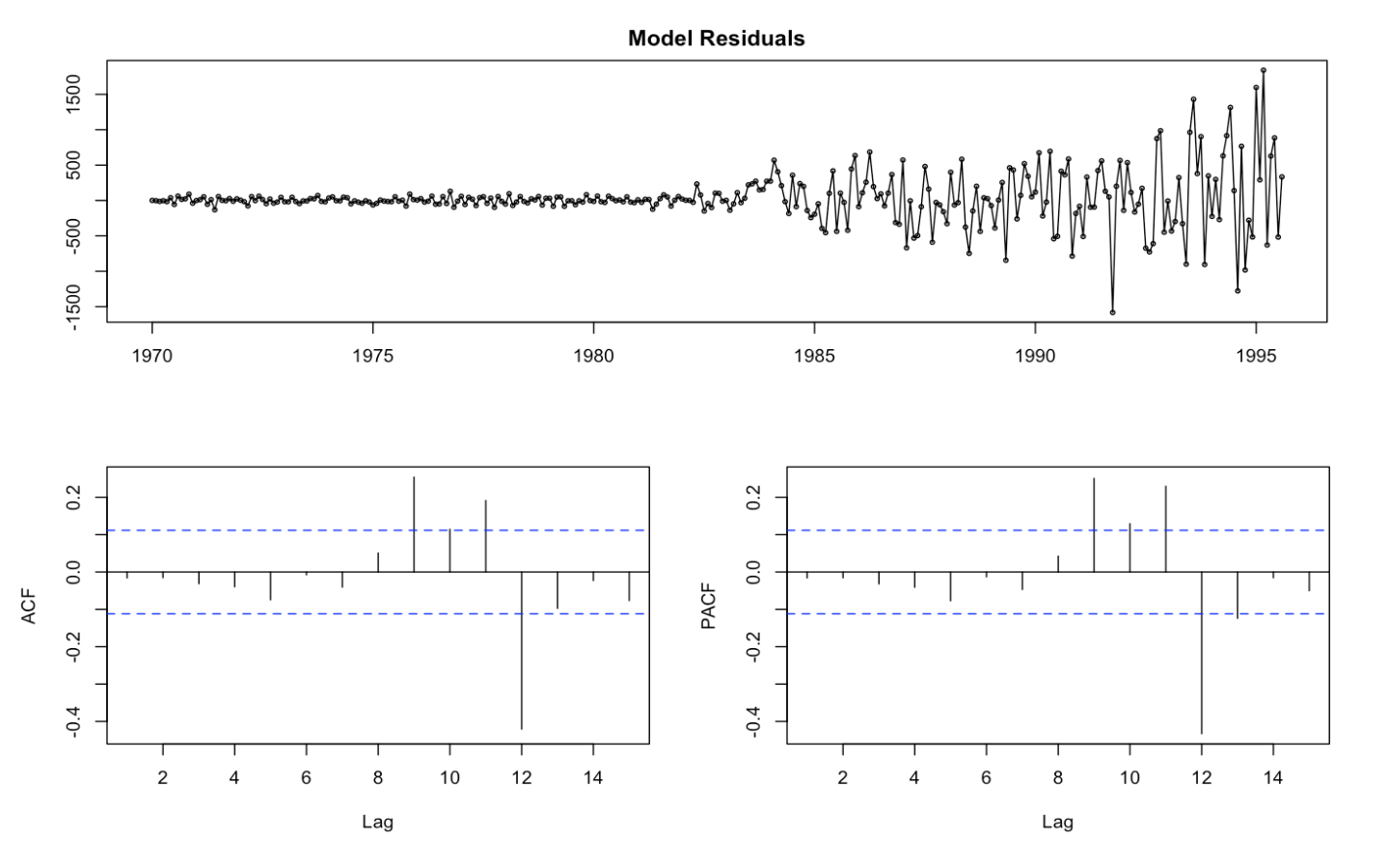
Box-Ljung test

data: gasAutoArimafinal$residuals

X-squared = 0.07563, df = 1, p-value = 0.7833

* Since p value is higher than 5%, our null hypothesis is vali

#### Check Model Residuals graphs for independence



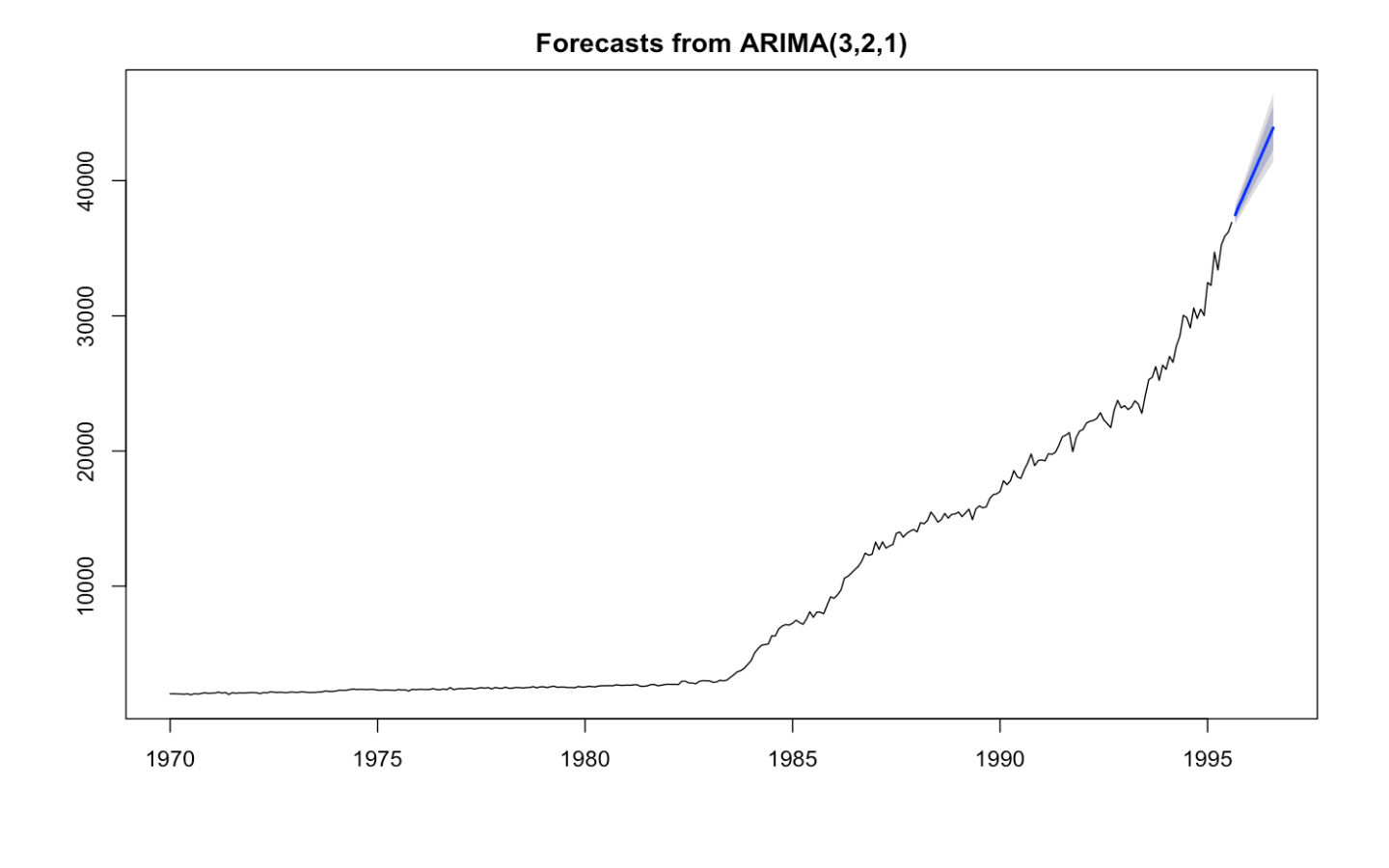
* As all the graphs are in support of the assumption that there is no pattern in the residuals, we can go ahead and calculate the forecast.

### Building Future Forecasts

* Future forecasts for next 12 months

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Point Forecast** | **Lo 80** | **Hi 80** | **Lo 95** | **Hi 95** |
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| Sep-95 | 37459.2 | 36988.7 | 37929.7 | 36739.6 | 38178.7 |
| Oct-95 | 38102.4 | 37542.6 | 38662.3 | 37246.2 | 38958.7 |
| Nov-95 | 38633.5 | 37962 | 39305 | 37606.5 | 39660.5 |
| Dec-95 | 39241.7 | 38500.1 | 39983.2 | 38107.6 | 40375.7 |
| Jan-96 | 39808.8 | 38953.7 | 40663.9 | 38501.1 | 41116.5 |
| Feb-96 | 40409.3 | 39455.4 | 41363.2 | 38950.4 | 41868.2 |
| Mar-96 | 40982.7 | 39916 | 42049.4 | 39351.3 | 42614.1 |
| Apr-96 | 41573.3 | 40402 | 42744.5 | 39781.9 | 43364.6 |
| May-96 | 42152.2 | 40866.4 | 43437.9 | 40185.8 | 44118.5 |
| Jun-96 | 42740.2 | 41341 | 44139.3 | 40600.4 | 44880 |
| Jul-96 | 43321.7 | 41803.2 | 44840.2 | 40999.3 | 45644.1 |
| Aug-96 | 43907.6 | 42269.2 | 45546 | 41401.9 | 46413.3 |

Future Forecasts – next 12 Months Graph



* The forecasts are shown as a blue line, with the 80% prediction intervals as a dark shaded area, and the 95% prediction intervals as a light shaded area.
* Prediction intervals are narrower, implying better reliability of the estimates

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_