Australian GAS Production

- A Time-Series analysis

Pradeep Panicker 6/7/20 Time Series Forecasting

Table of Contents

PROJECT OBJECTIVE

Data Interpretation –

A dataset containing the Australian Monthly Gas Production for the period 1956 ~ 1995, is imported into the Global environment of R Studio – for further interpretation.

Project Objective –

1. Read the data as TS object in R and PLOT the Graph.
2. Define which all components are present in the Time Series.
3. What is the periodicity of the Data?
4. To check whether the time series is 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. To develop an initial forecast for next 20 periods. Check the same using the various metrics. After finalising the model, develop a final forecast for the 12 time periods. Use both manual and auto.arima .
6. Define the accuracy of the model.

WORKING ENVIORNMENT SETUP

1. R Studio is setup for programming.
2. Library ‘forecast’ is installed and attached to the R Environment.
3. Working file ‘gas’ is downloaded from the datafile library that exists in ‘Forecast’.
4. The data downloaded is a Time Series file and is now saved as a Time-Series object file named ‘gas’.
5. Manually checking is done for any MISSING VALUES – missing values will affect the continuity of the Time Series pattern.
6. Data covers the period JAN ‘1956 TO AUGUST 1995 = 476 DATA. THE SAME IS RECORDED AND VISIBLE IN THE Global Environment.

   Checking Missing Value –
   1956 ~ 1965 = 12 Months x 10 Years = 120 Data
   1966 ~ 1975 = 12 Months x 10 Years = 120 Data
   1976 ~ 1985 = 12 Months x 10 Years = 120 Data
   1986 ~ 1995 AUG = (12Months x 09 Years) + 08Months = 116 Data
   Total = 476 Data

   Data Structure in Global Environment states [1:476] Data.

   Hence NO MISSING VALUE IS REGISTERED BETWEEN THIS PERIOD.

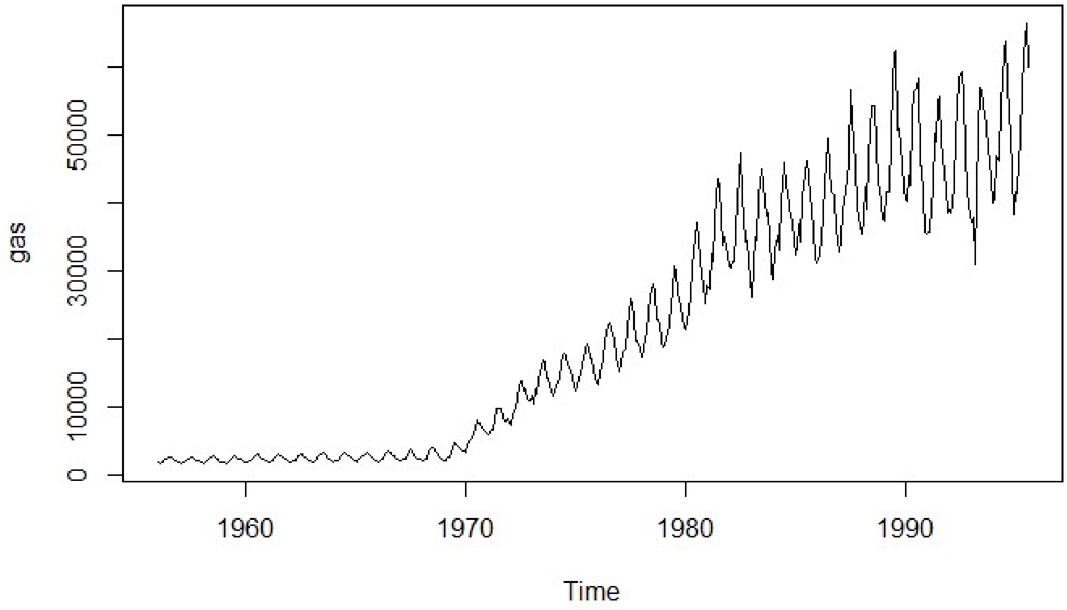
The TIME SERIES file ‘gas’ is now ready for Time-Series Analysis.

PLOTTING TIME-SERIES DATA

For generating a Time-Series Plot, command [plot(filename)] is used.

plot(gas) produced the below depicted graph –

GRAPH-1



TIME-SERIES DATA ➔ COMPONENT ANALYSIS

PERIOD 1956 ~ 1969 ➔

TREND ➔ From ‘Graph 1’, it is evident that the DATA does not have a SIZABLE Increasing nor Decreasing Trend. **The Production Output tend to increase very slightly** – Data in the year 1956 JAN starts @ 1709 Units and Data at 1969 December has reached ONLY 3461 Units.

SEASONALITY ➔ From ‘Graph 1’, **a Seasonality variation of data is seen present**. The volume of production peaks during a certain period of a Year.

From the data, it is observed that this peak is evident during the months JUNE ~ AUGUST of every year.

Also, **this Seasonality increases over every year** – which means during this peak period , the Demand during this period is increasing Year-On-Year.

The below data, as a sample lot from the main data, shows the peaking during the period.

Year	MAY	JUN	JUL	AUG	SEPT
1956	2173	2321	2468	2416	2184
1957	2311	2279	2638	2448	2279
1960	2617	2828	2965	2891	2532
1965	2817	3123	3345	3112	2659
1967	2965	3239	3608	3524	3018
1969	3292	3724	4652	4379	4231

PERIOD 1970 ~ 1983 ➔

TREND – An evident **presence of steep increasing trend** in the GAS Production is visible from the ‘Graph1’. The data reading also shows that the JANUARY 1970 production was @ 3345 Units whereas the DECEMBER 1983 production stood @ 30234 Units.

SEASONALITY – Seasonal peaking of production of GAS is present in every year. A slight spread in the seasonality period is seen in TWO years – 1972 & 1973, thereafter the seasonality period remains constant.
Also, every year, an increase is observed in the volume of production of Gas, Year-On-Year , during the seasonal period itself. i.e. The Gas PRODUCED during the Seasonal period in 1974 MAY ~ AUGUST IS SUBSTANTIALLY MORE than the Gas PRODUCED during the Seasonal period 1973 MAY ~ AUGUST.

PERIOD 1984 ~ 1995 ➔

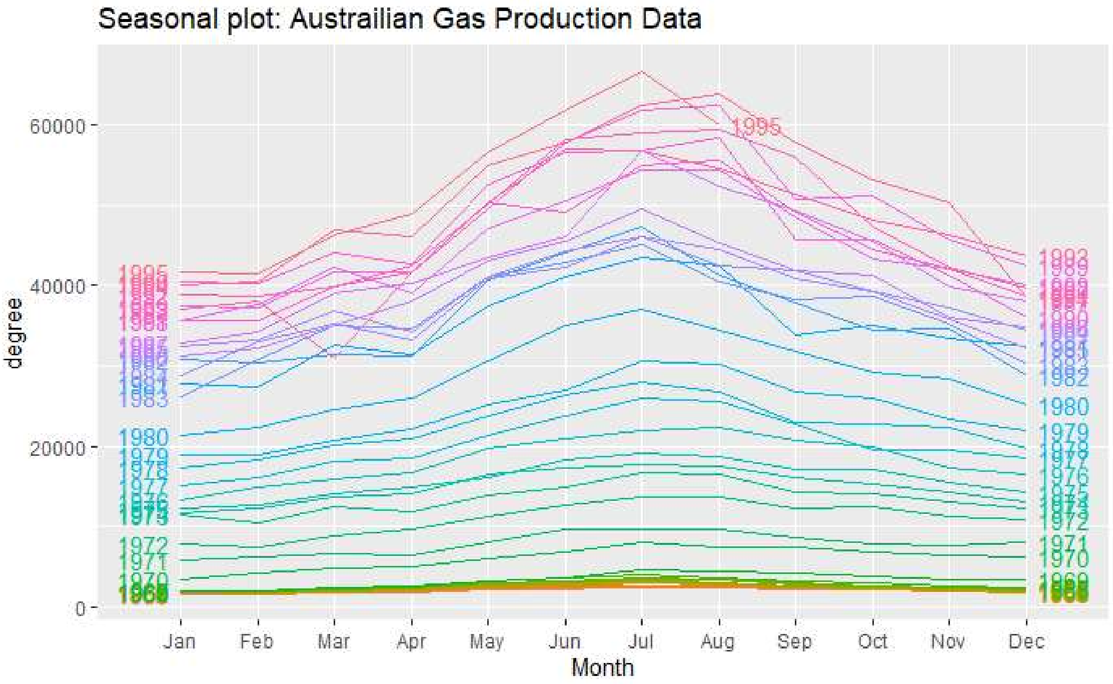
<u>TREND</u> – The data shows a continuous increase in production of Gas – an increasing TREND year-on-year. The steep increase as was there in the previous block of TEN YEARS is not present here, a slight dampening effect is there by which the steepness drops down as it reaches the year 1995.

<u>SEASONALITY</u> – Seasonality do exist throughout the period, but an increasing trend, Year-On-Year, during seasonal period is not witnessed. Off-shoot Spikes observed intermittently in some Season indicates HIGH PRODUCTION due to a certain non-repetitive factor. These spikes of Production are observed –

a) as increased production – during FOUR Seasons
b) as decreased production – during ONE Season.

Seasonality & Trend Graph –

<u>GRAPH 2-</u>



<u>From the above it is evident that the Seasonality is evident from the Year 1970 Onwards. The Seasonality is present during the months MAY ~ SEPTEMBER. Also the trend of Seasonality indicates it is a Multiplicative Seasonality.</u>

<u>Also the TREND of production is seen to increase over the years , starting from 1970.</u>

From 1956 to 1970, the series is somewhat STATIONARY.

PERIODICITY – Definition & Determination

PERIODICITY –

Periodicity of a Time Series data is defined as the frequency by which each data of the Time Series is placed.

In the dataset under consideration 'Gas' – the periodicity is MONTHLY.

This is since the data plotted shows the TREND & SEASONALITY pattern changes – every MONTH, and is recorded monthly , for FORTY years[less four months at the end of last year].

SPLITING DATASET – TRAIN & TEST.

Since a uniformity is found in the data with effect from the year 1970 onwards, A new sub-dataset is created which has data starting w.e.f. 1970 Onwards – new dataset known as 'Gas1'

gas1= window(gas, start= c(1970,1))

Dataset 'Gas1' is spilt into Train Data Set and Test Data Set.

gastrain = window(gas1, end= c(1993,12))
gastest= window(gas1, start= c(1994,1))

STATIONARY Dataset ; ADF TEST ; De-Seasonalisation of TS.

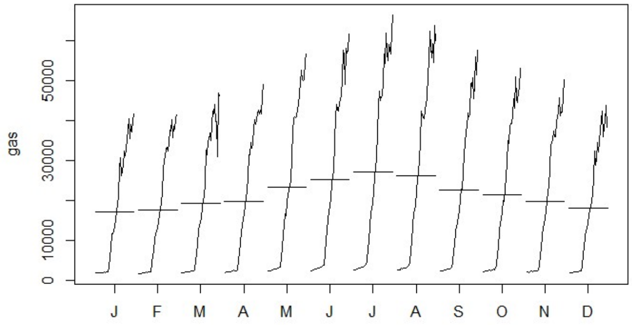
Stationary Time Series-

Three properties define a Time Series data to be Stationery –

a) The Mean & Variance of the data is to be constant over a period.

Upon visual inspection of Graph 1**, considering the dataset from the year 1970, Seasonality and Trend is seen in the dataset.**

Graph 3



Also, from the above MONTHPLOT, the Monthly Average varies .THEREBY THE DATASET IS BEING CATAGORISED AS NON-STATIONERY.

Augmented Dicky-Fuller [ADF] Test –

ADF test is conducted to check statistically whether the time series data is Stationary or Not.

A NULL Hypothesis & Alternate Hypothesis is generated, and the validity of NULL Hypothesis is checked with the support of ADF Test.

NULL Hypothesis H0 = The TS Data 'Gas' is Non-Stationery.

ALTERNATE Hypothesis H1 = The TS Data 'Gas' is Stationery.

If the 'p' value from the test results is higher than 5%, then the NULL Hypothesis will STAND GOOD, the TS Data will be proven Non-Stationery.

Results of the ADF Test –

```
Augmented Dickey-Fuller Test

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

The 'p' Value is found to be 27.64% , which is why the data is proven to be NON-STATIONERY.

De-Seasonalize the Time Series –

From Graph 2 & Graph 3 above, it is evident that Seasonality and Trend exists in the Time Series data – 'Gas'.

The Time Series data can be decomposed into its basic THREE Components –

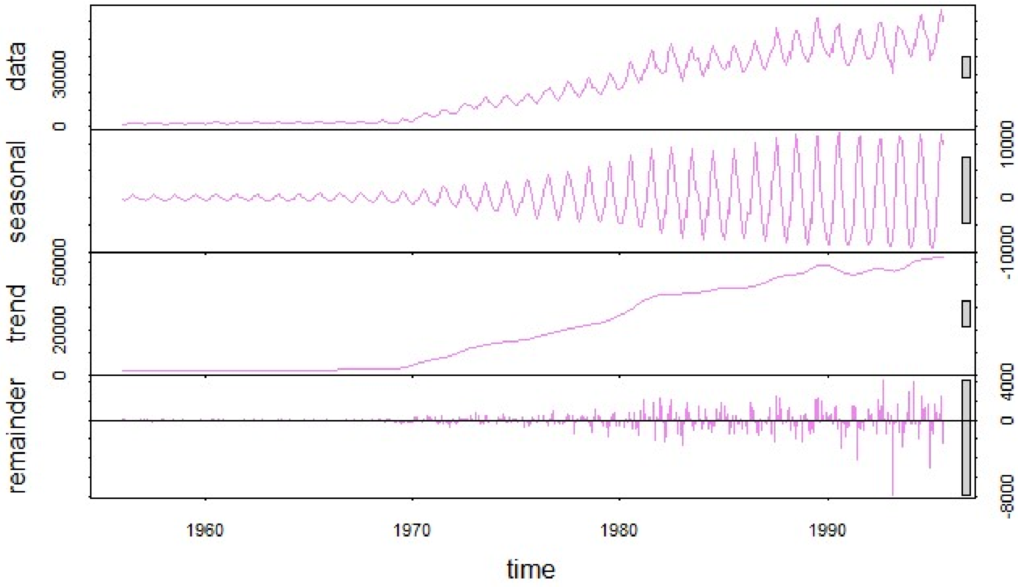
a) Seasonality
b) Trend &
c) Error / Random components.

The same is done using the 'stl' command in R, by which also, we can judge which component is significant in defining the characteristics of the data. The limitations of usage here is , for Multiplicative series, we need to take the 'Log' of data and conduct analysis as Additive series.

Hence, we will be using the following command ➔ gasdeslog= stl(log10(gas), s.window = 'p')

The graph plotted for the decomposed components is as below –

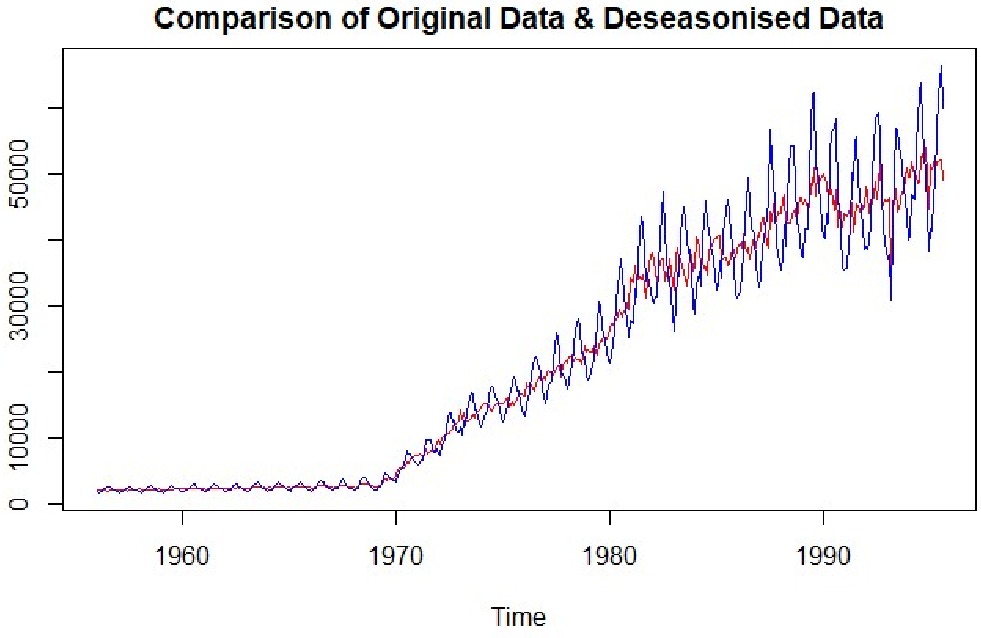
Graph 4



The rectangular bar at the RHS of each component box indicates the Scale of the data present in the respective component. The Lesser the length of the bar – THE HIGHER SIGNIFICANT IS THE COMPONENT.

Thereby, in the 'Gas' data – the TREND COMPONENT IS THE MOST SIGNIFICANT.

Graph 5



Since the ‘Seasonality’ factor is variable in nature, we need to execute a ‘Multiplicative’ De-Seasonalisation. ‘stl’ command is used after taking ‘log’ of data, this is since ‘Multiplicative’ model does not work good in ‘stl’ command, we are using ‘Additive’ model against logarithmic data.

Graph 5 is the result of De-seasonalisation.

The Blue line represents the Original Data whereas the RED Line represents the De-Seasonalized data. **The path of the RED Line superimposed on the path of the Blue Line – indicates that the variation of the data along the time series is EXPLAINED WHOLESOME BY THE TREND COMPONENT.**

FORCASTING

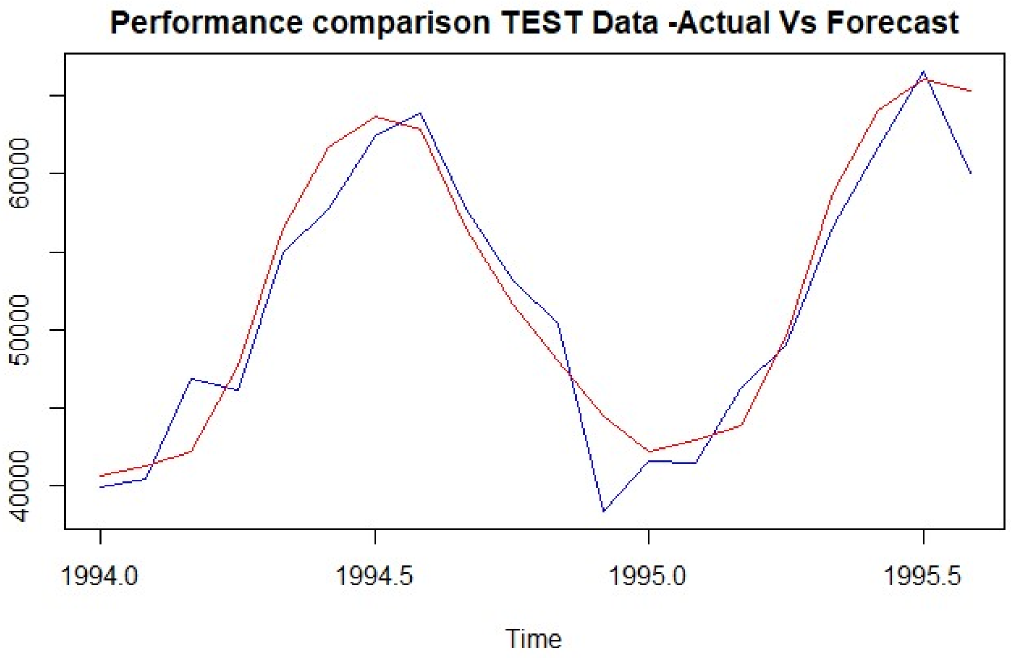
The data contains TREND & SEASONALITY components , and hence we will use the HOLT-WINTER Model for manual method of forecasting.

HOLT-Winter Model –

The Holt-Winter model is generated on the Train Data set, for 20 periods.

A performance plot is generated with the test data. Graph as below –

Graph 6



The Blue line represents the actual data while the RED line represents the FORECAST Performance of the model – both on the Test data.

The MAPE Value is 4.21%. The error is found to be Low suggesting a GOOD MODEL.

Interpreting the performance [Holt Winters Model]-

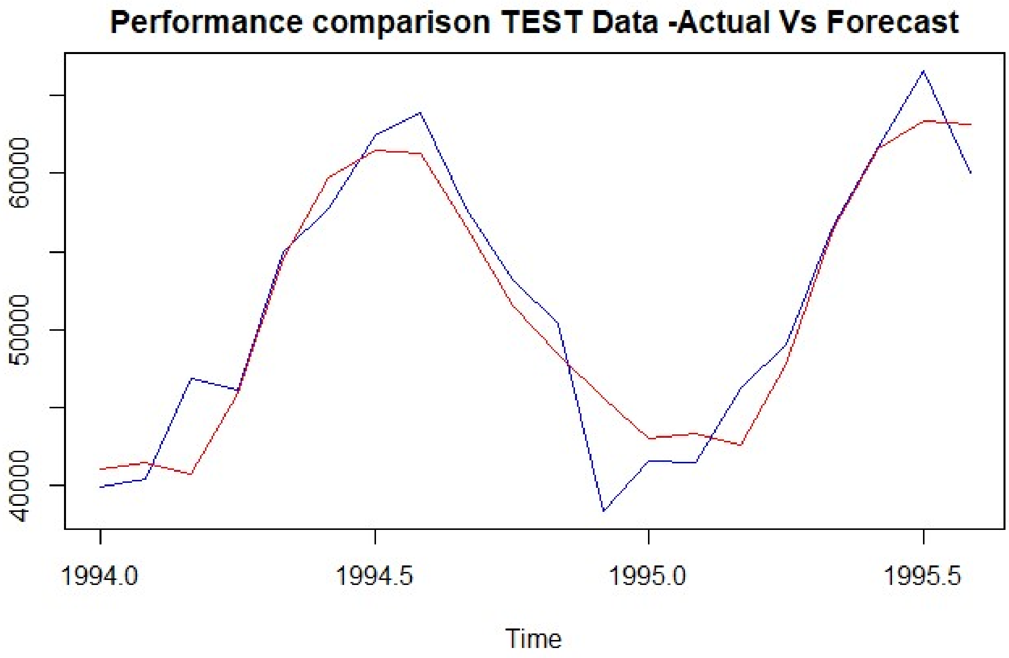
a) Wherever the production of Gas is LOW, the forecast model shows an error by forecasting higher values.
b) Wherever the production of Gas is at its PEAK / HIGHEST, the forecast model attains to reach the peak value SLIGHTLY EARLIER.
c) The TREND is maintained well.

Decomposition Model –

Here, the Train Dataset is decomposed using the Multiplicative model and is then used to forecast the data for 20 periods.

The forecast values and plotted against the Test Data set to compare the performance -

Graph 7



The Blue line represents the actual data while the RED line represents the FORECAST Performance of the model – both on the Test data.

The MAPE Value is 4.253%. The error is found to be Low suggesting a GOOD MODEL.

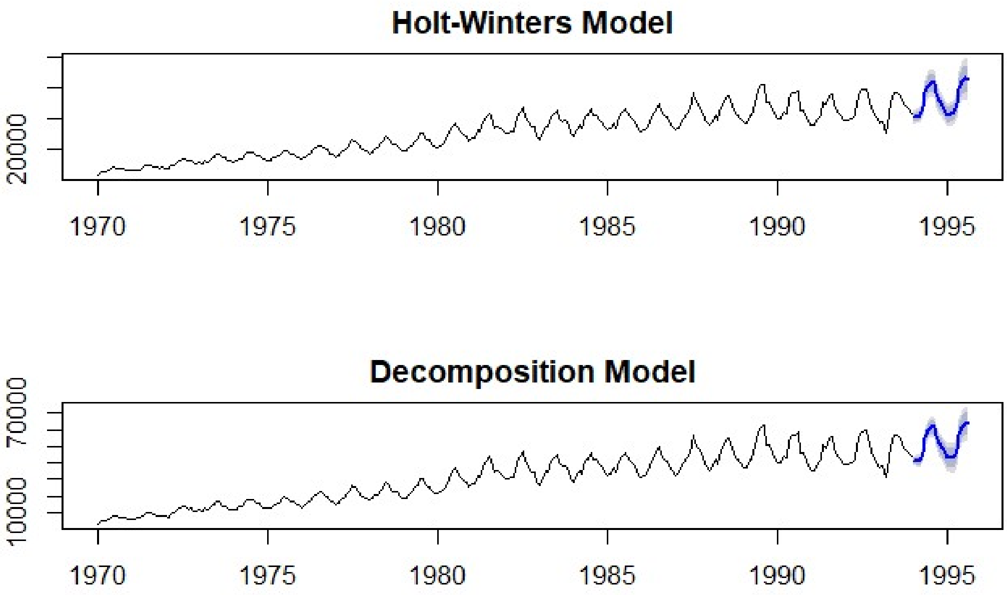
Interpreting the performance [Holt Winters Model]

The forecast misses to catch the actual data trend at the peak of production as well as the lowest range as well. Also, the forecast is on the Lower side as compared to the actual dataset during the seasonal changes.

In comparison to the HW Model performance, the HW Model is better in terms of MAPE as well as the forecast plotting.

Now, with the model built with two different methods, we can now Forecast on the Test Data and check the performance also –

Graph 8



The above visualization does not differ majorly, but it is evident that the 'Confidence Interval' bandwidth is high at the starting of the forecast. This is truly NOT good.

We will now try ARIMA Models and take a comparison of performance.

ARIMA MODELS –

Manual ARIMA Model -

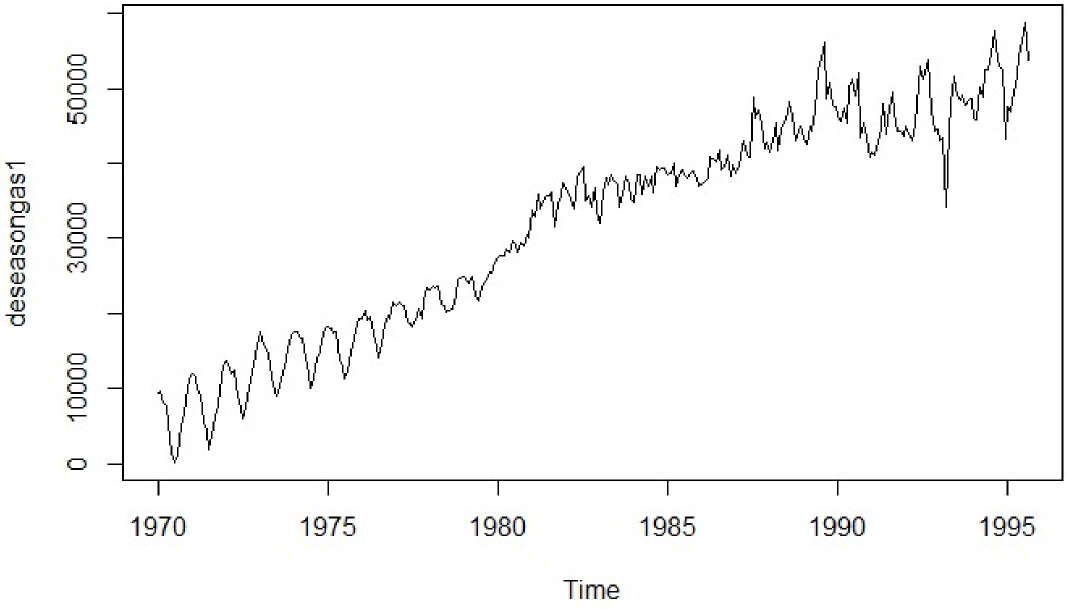
The process flow for generating an ARIMA Model is –

Data Ordering & Visual Check => Data Stationarize => Determine ACF[q] & PACF [p] => Differentiate Data [d] => Generate ARIMA Model with [p,d,q] values on Train Data => Run Model on TEST Data and Optimize Model with various iterations of [p,d,q] such that we choose a Model with Least AIC;BIC and is also a Parsimonious One.

Since we are using ARIMA and not SARIMA, we need to De-seasonalize the data.

Below depicted is the De-seasonalized data graph –

Graph 9



Visually, the dataset shows a trend but NO SEASONALITY.

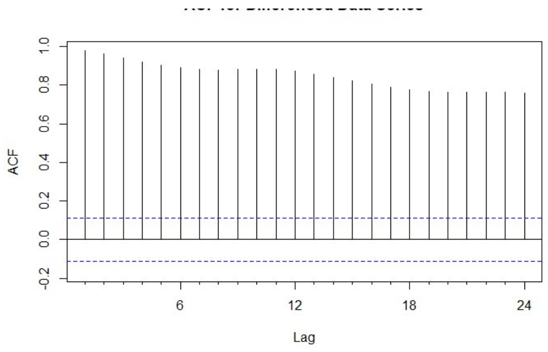
Checking whether the data has become Stationary , using the 'Augmented Dicky-Fuller' test, the results are –

```
data:  deseasongas1
Dickey-Fuller = -3.6706, Lag order = 6, p-value = 0.02665
alternative hypothesis: stationary
```

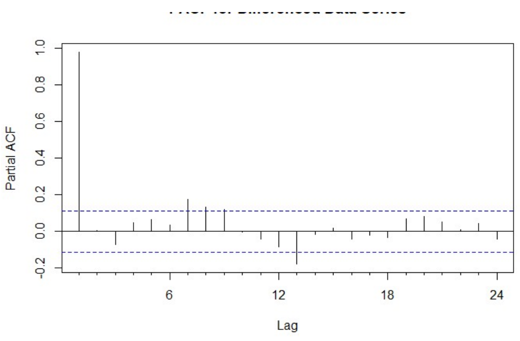
THE DE-SEASONALIZED DATA – 'deseasongas1' HAS BECOME STATIONARY NOW, since 'p-value' is less than 5% and the Null Hypothesis – 'Data is Not Stationary' condition IS REJECTED.

Determining ACF & PACF Values –

ACF Plot [Graph 10]–



PACF Plot[Graph 11]-



Considering the Parsimonious theory for parameter selection, we will now try Model buildings for [p,d,q] values as [4,1,6], [3,1,6], [4,1,10] & [4,1,8] on the De-Seasonalized dataset – 'deseasongas1'.

Train Data set was created as 'desgas1train= window(deseasongas1, start= c(1970,1), end= c(1993,12))'

Model Performance Parameters -

Model parameters were captured , as listed below –

MODEL NAME	[p,d,q] Values	AIC Value	AICc Value	BIC Value
gas1arima1	[3,1,6]	5150.26	5151.06	5186.86
gas1arima2	[4,1,10]	5138.14	5139.19	5193.03
gas1arima3	[4,1,6]	5139.49	5140.45	5179.74
gas1arima4	[4,1,8]	5136.99	5138.32	5184.46

'gas1arima4' gave the best 'AICc' values , claiming it to be the best model.

AUTO ARIMA -

The 'Auto ARIMA' model is generated on the original dataset, since all the parameters for De-Seasonlization as well as making the dataset Stationary will be taken care by the MODEL function internally itself.

The desired model is built using the command –

autoarimagas1= auto.arima(gastrain, seasonal = TRUE).

Model Performance Parameters-

The Model performance parameters were recorded as –

MODEL NAME	[p,d,q] Values	AIC Value	AICc Value	BIC Value
autoarimagas1	[2,1,1]	4939.33	4939.64	4961.03

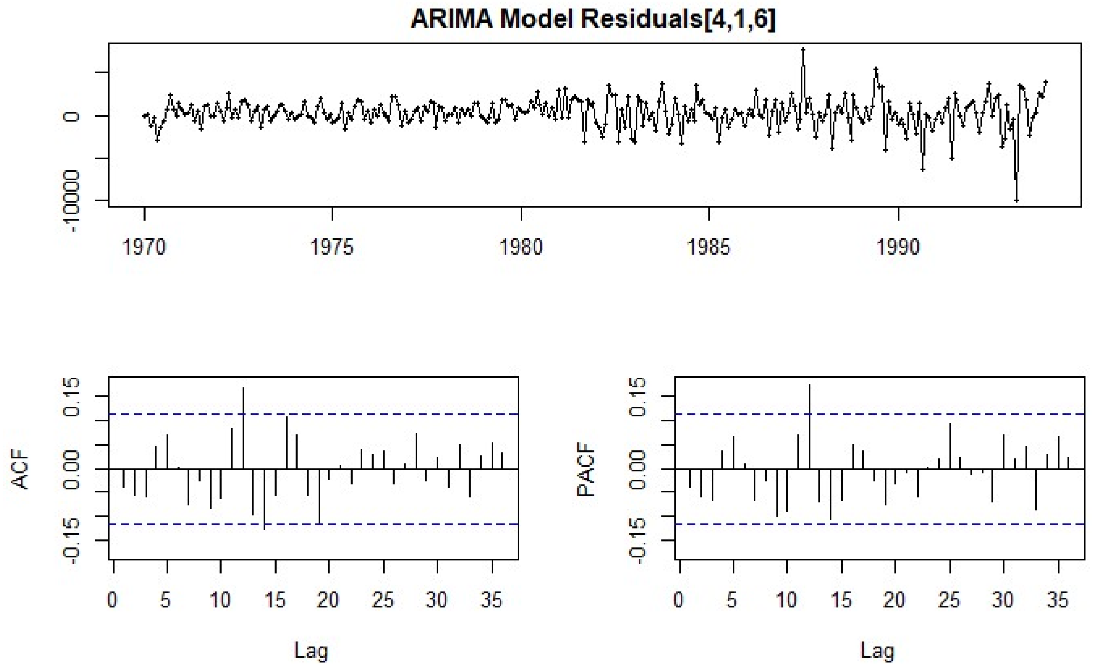
ARIMA Model Validation –

Two methods are used for ARIMA Model validations –

1) RESIDUALS behavior study – The Residuals SHOULD NOT CONTAIN ANY PATTERN.
2) LUENG-BOX Test – If the 'p' value is greater than 5%, the NULL Hypothesis that the RESIDUALS are INDEPENDENT is accepted.

The RESIDUALs behavior test proved that ARIMA Model with [p,d,q] values as [4,1,6] performs the best- since NO ACF & PACF Lags are seen Significant.

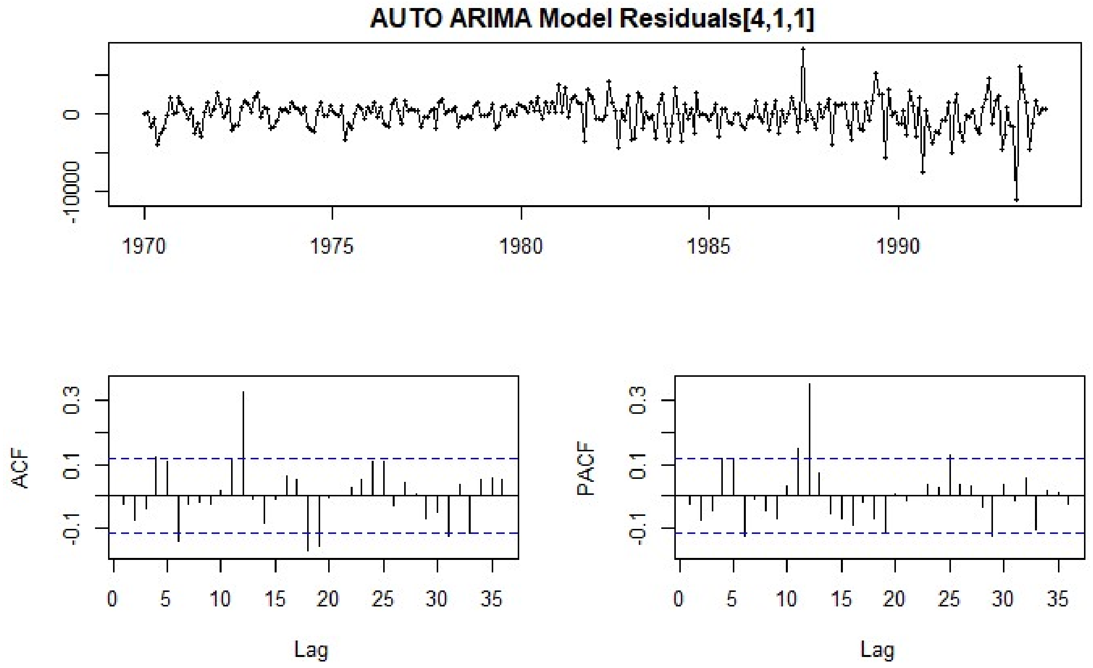
Graph 12-



The Lueng-Box test specifies 'p' value @ 51.83% by which it proves that the Residuals are INDEPENDENT.

Below graph depicts the status of RESIDUALs for the AUTO-ARIMA Model created 'autoarimagas1'

Graph 13-



The Lueng-Box test specifies 'p' value @ 65.63% by which it proves that the Residuals are INDEPENDENT.

Forecasting & Accuracy check - ARIMA Models -

We will now be FORECASTING using the train dataset with various models generated via Manual ARIMA as well as the AUTO-ARIMA method and then check the accuracy of the models. This forecasting is attempted for 20 periods.

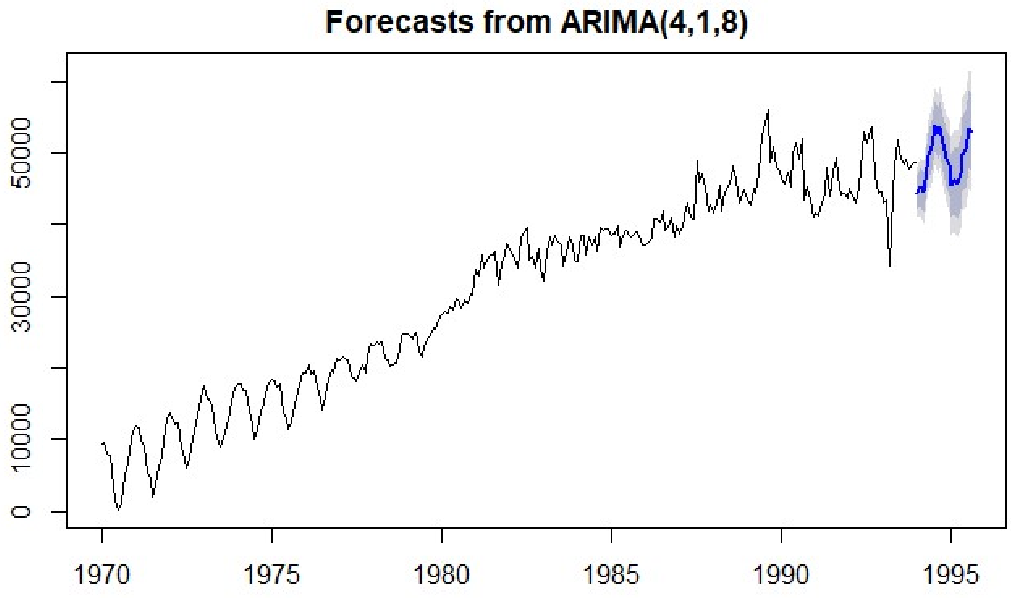
The results of accuracy – the MAPE factor , is tabulated as below –

MODEL NAME	[p,d,q] Values	AIC Value	AICc Value	BIC Value	MAPE - TEST DATA
fcastgas11	[3,1,6]	5150.26	5151.06	5186.86	5.517
fcastgas12	[4,1,10]	5138.14	5139.19	5193.03	5.627
fcastgas13	[4,1,6]	5139.49	5140.45	5179.74	5.728
fcastgas14	[4,1,8]	5136.99	5138.32	5184.46	5.609

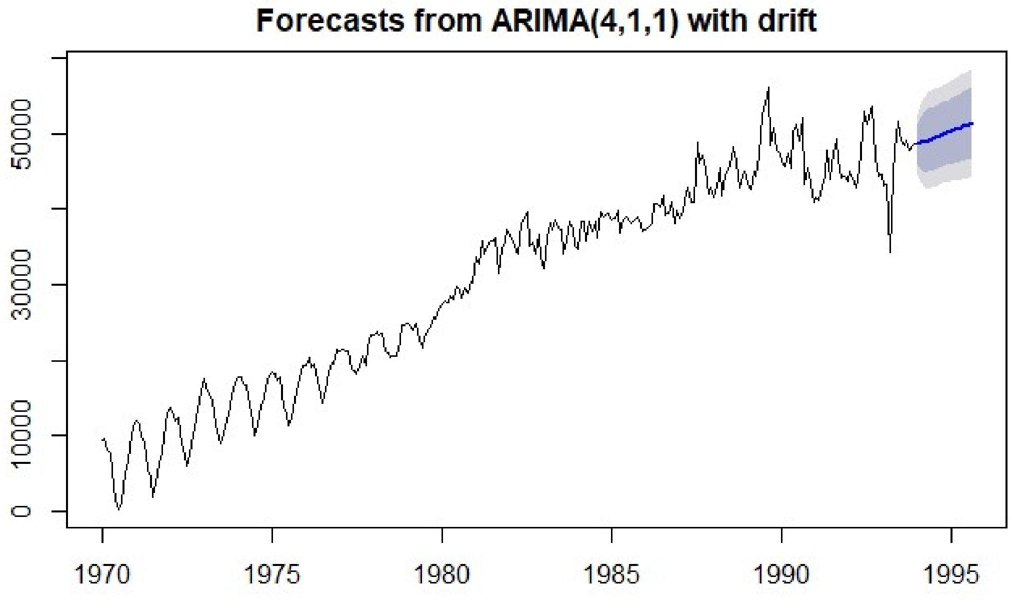
MODEL NAME	[p,d,q] Values	AIC Value	AICc Value	BIC Value	MAPE - TEST DATA
fcastautogas11	[2,1,1]	4939.33	4939.64	4961.03	6.894

We can consider 'fcastgas14' as the most accurate Model among the ARIMA Models generated.

Forecast with ARIMA [4,1,8] Graph13 –



Forecast with AUTO ARIMA [4,1,1] Graph 14 –



BEST MODEL SELECTION based on BEST ACCURACY –

We had generated various Models for forecasting and tested them on the respective test data. A consolidation of the accuracy parameters with respective Model names is tabulated below –

MODEL METHOD	MAPE -TEST DATA
HOLT WINTERS MODEL	4.21
DECOMPOSITION Model	4.253
MANUAL ARIMA[4,1,8]	5.609
fcastgas11 - ARIMA	5.517
fcastgas12 - ARIMA	5.627
fcastgas13 - ARIMA	5.728
fcastgas14 - ARIMA	5.609
fcastautogas11 -Auto ARIMA	6.894

HOLT-WINTER's Model stands out to be best model performed for the Gas Data.

Thereby, this Model will be now used to forecast the FUTURE TWELVE PERIODS .

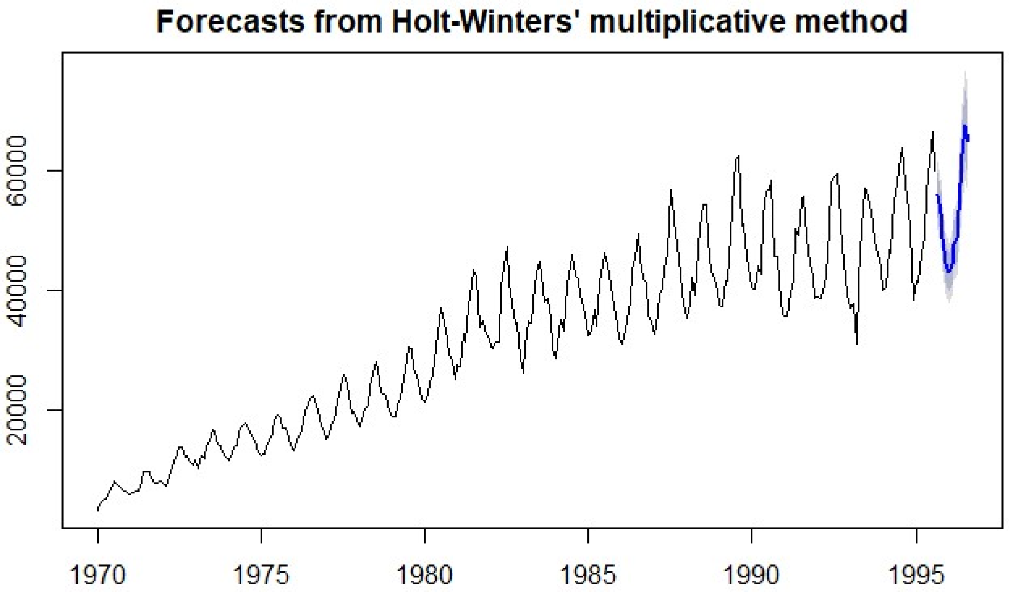
FORECASTING WITH BEST MODEL – Twelve Months

From the above tabulated accuracy parameters data, we choose the HOLT-WINTERs Model for future forecasting – since the MAPE value is the least.

We now take the complete dataset for the period from 1970 JANUARY to 1995 AUGUST for the forecast and apply the command function as –

```
finalfcast= hw(gas1, seasonal= 'm',h=12)
```

The graphical visualization is generated as Graph 14 –



The CONFIDENCE INTERVALS at the start of the forecast period is NARROW suggesting a robust forecasting.

GAS Production Forecast SEPT 1995 ~ AUG 1996 ➔

Sep-95	Oct-95	Nov-95	Dec-95	Jan-96	Feb-96	Mar-96	Apr-96
55919.31	53032.37	48789.96	44614.86	42962.29	44336.6	47698.69	48766.56

May-96	Jun-96	Jul-96	Aug-96
57967.31	62828.45	67670.21	64989.22

=========   THE END ========