**TIME SERIES FORECASTING**

**INDIVIDUAL ASSIGNMENT**

* **SINDHUJA HARIHARAN**

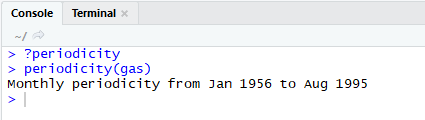
**Solution:**

**Q 1. *What is the periodicity of dataset?***

**Periodicity of any time series data can be defined as the frequency in which the data points have been collected. For the given dataset, the periodicity can be found using the following package in R**

* **Xts (eXtensible Time Series).**

**The above package has a function – “periodicity(<dataset>)”. This function takes in the dataset as an argument and returns the periodicity.**



**From the above screenshot we can infer that the data points have been recorded for every month since January 1956 until August 1995.**

**Q 2. *Which components of time series are present in this dataset?***

**Any time series data could had one or more of the following components within it.**

1. **Trend Component**
2. **Seasonal Component**
3. **Cyclic Component**
4. **Irregular Component**

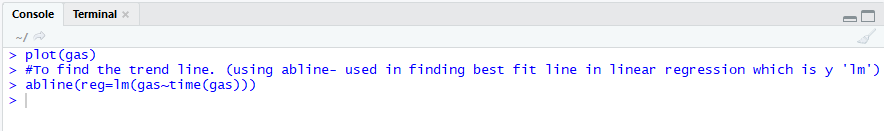
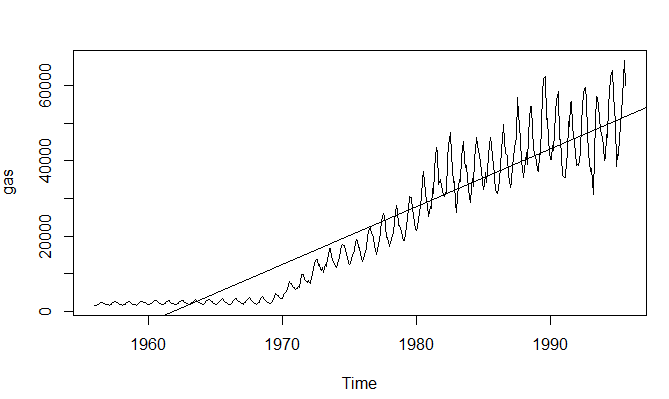
**Trend Component: When the time series is said to have trend component it means that there is long term variation (either linear or nonlinear). The trend can be increasing/decreasing (positive/negative in a sense). Trend also relates to the data being stationary. If there is no trend, the dataset is said to be stationary.**

**Seasonal Component: As the name suggests, seasonal Component relates to the change in one particular point in time. For example, the Sales of winter clothing would be at its peak just before winter and be a lot lesser during summer.**

**Cyclic Component: Cyclic change refers to the kind of variation which gets repeated after a certain period of time. Example, Road traffic – usually peaks in business hours – everyday.**

**Irregular Component: The variation component which remains even after removing all other components can be owed to Irregular component**

**For analyzing the components of a time series, we need to first plot the data set. The following screenshot explains about various components present in the chosen dataset.**

**In the above screenshot, we can clearly see that there is a definite “Trend” component. Although the initial ten to fifteen years don’t show a clear trend, the trend line shows that there is an increasing (positive) trend being observed after 1970 until 1990. (Long term variation)**

**There is “Seasonality” component too, since we can see a spike after every few months.**

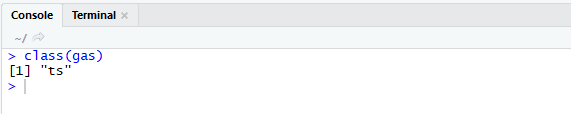
**It also looks like there could be some “Irregularity” components, since we can see a steep dip sometime after 1990.**

**Q 3. *Is the time series Stationary.***

**The trend line in the above graph has a positive effect over time, and hence shows that the chose dataset is not stationary.**

**Q 4. *Develop an ARIMA Model to forecast for next 12 periods. (Show all the steps)***

**Step1: To use any dataset for time series forecasting the dataset should belong to class ‘ts’.**



***Note:*** *If the dataset doesn’t belong to ‘ts’ class, it has to be converted to ‘ts’ type in order to be used in time series forecasting. The following command converts the dataset to a time series class. - ts(<dataset>,start=<starting period>,end==<starting period>,frequency = <how frequent the data points are collected>)*

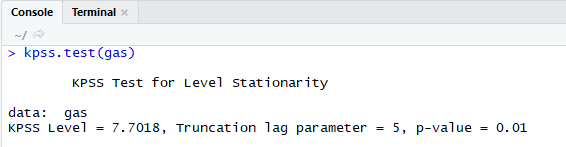
The above screenshot shows that the chosen dataset is already a part of ‘ts’ class.

**Step2:** For time series forecasting to be done, the dataset should be

* + - Should contain data points collected in regular intervals (periodicity)
    - Stationary

Periodicity: **As discussed in the above solution, the data points for the chosen data have been recorded for every month since January 1956 until August 1995.**

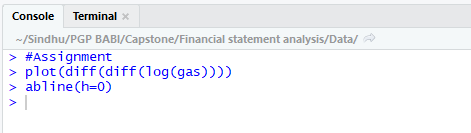
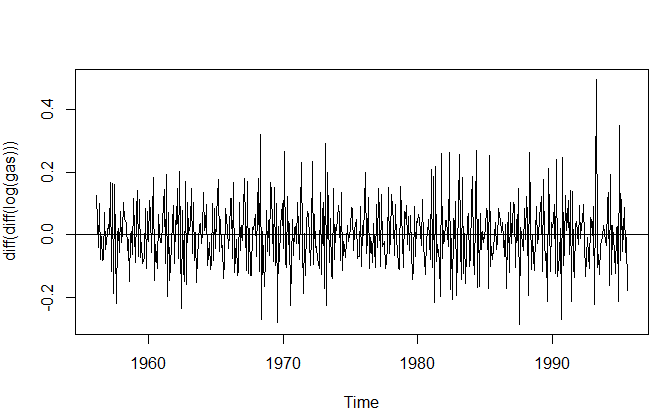
Stationarity: For performing time series forecasting, the dataset considered should be stationary. **A stationary time series is one where mean and variance are constant over time.** As discussed in the above solution, t**he trend line for this dataset when plotted has a positive effect over time, and hence shows that the chosen dataset is not stationary. This can be proved by conducting a simple test in R – KPSS test (Kwiatkowski–Phillips–Schmidt–Shin test)**

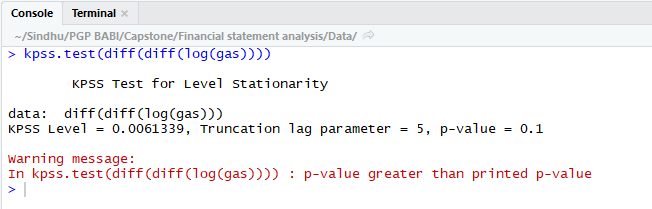


**The p-value is < 0.05 and hence it proves that the data set is not stationary.**

**To make the data stationary, the dataset is usually subjected to any sort of transformation techniques like log transformation/ taking differences between adjacent data points.**

**Let us now perform differencing and see if the kpss.test passes.**



After performing differentiation, the kpss.test gives a p-value > 0.05 and hence the Ho can be accepted. So, the dataset has become stationary.

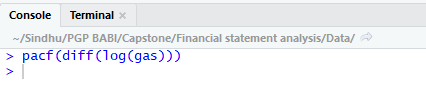
**Step3:** Modelling – We will build an ARIMA model to forecast gas consumption for the next 12 months. To perform ARIMA, we need to decide the following factors

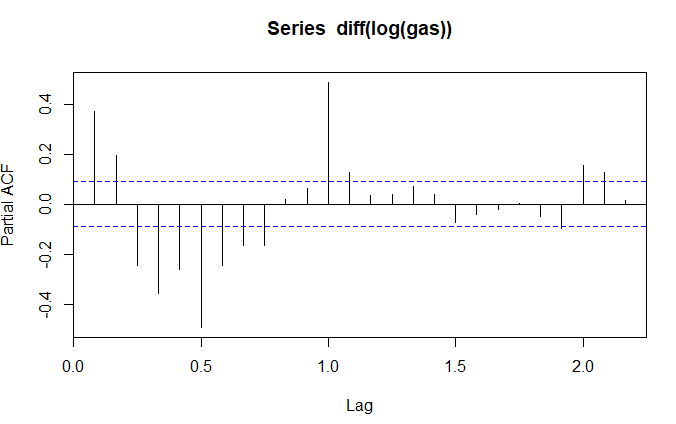
**AR -** Auto regressive (p) component

**I -** Integration (d) component

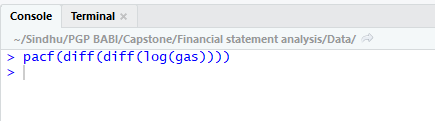
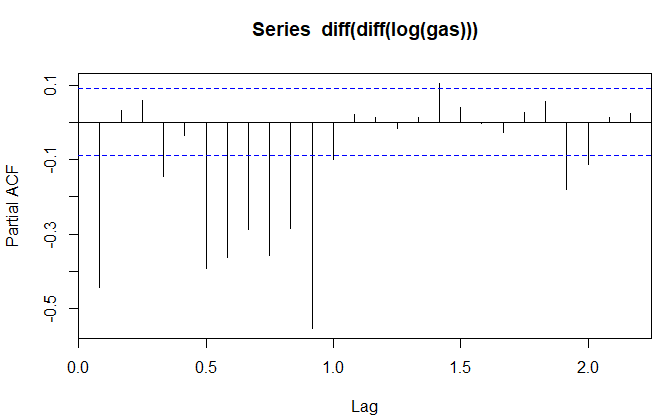
**MA –** Moving Average (q) component

**p -** This value can be found after plotting the ACF plot (Auto correlation function plot)





When the data was differenced once, the acceptable p value comes to 9. If the series has autocorrelations out to a high number of lags, then it probably needs a higher order of differencing. So in the above scenario, we will apply another order of differencing and see if the number of lags reduces.

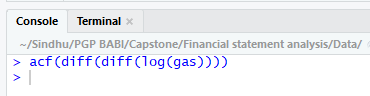
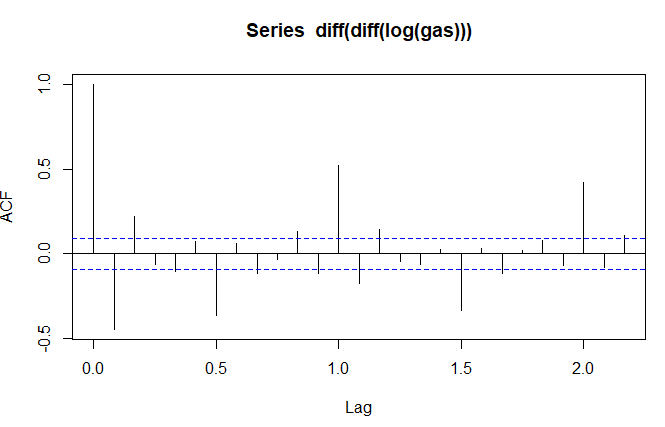
 

From the above PACF plot, we can conclude that the p value is – 1

**d -** Integration factor points to the order of differencing (how many times the dataset has been transformed/differentiated.

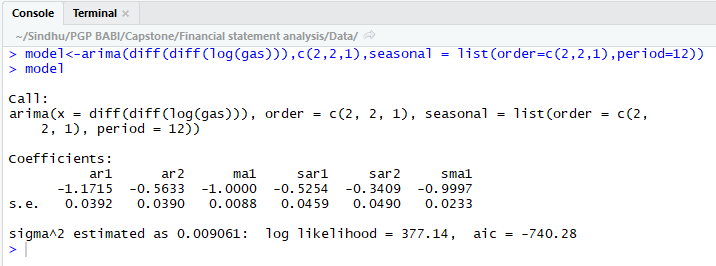
In the above scenario, d will be equal to 2, since the dataset has been differentiated twice.

**q -** Moving Average part can be ascertained by the ACF (complete Autocorrelation Function plot)

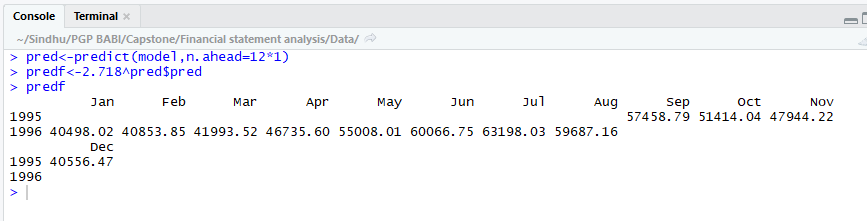
 

From the above PACF plot, we can conclude that the q value is – 2 (after omitting the ‘0’Th lag.

Running the model:



After running the above model, we can predict the gas consumption for the upcoming year until Aug 1996



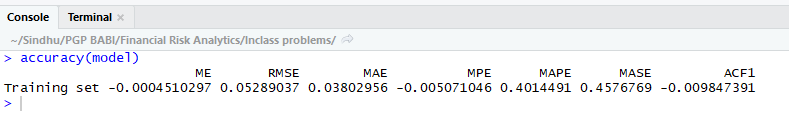
As we can see from the above screenshot, the values for the next 12 months have been predicted through ARIMA model for time series.

**Q 5. Report the accuracy of the model using all the four parameters.**

**Accuracy for the above model can be found using the following methods.**

* **ME: Mean Error**
* **RMSE: Root Mean Squared Error**
* **MAE: Mean Absolute Error**
* **MPE: Mean Percentage Error**
* **MAPE: Mean Absolute Percentage Error**
* **MASE: Mean Absolute Scaled Error**
* **ACF1: Autocorrelation of errors at lag 1.**

**The accuracy() function in timeseries package, gives a value for all the above parameters. For the above ARIMA model, the accuracy parameters came down to the values shown in the following screenshot.**



**ME: Mean Error:** **It is the average of all the forecast values. A mean error value other than zero suggests a tendency of the model to over forecast or under forecast (positive error).**

**In the above model, ME (Mean Error value) is -0.00045, which is almost equal to zero. And so the model can be taken as an ideal forecast model.**

**RMSE: Root Mean Squared Error: This parameter is a root of squared means. Hence, it is always positive. A ‘0’ for this parameter would mean that the model is a perfect fit model. The closer the value is to 0, the better the model is.**

**In the above model, the value is 0.05. So, the model fit is better.**

**MAE: Mean Absolute Error: This value is the same as Mean Error, except that, here the absolute values are taken instead of actual values.**

**In the above scenario, the value for MAE is 0.038, which is again close to 0. Hence, this parameter also shows that the model is a better fit.**

**MPE - Mean Percentage Error: The mean percentage error is the average of percentage errors by which forecasts of a model differ from actual values.**

**MAPE: Mean Absolute Percentage Error: Because the positive and negative errors may tend to cancel themselves, MPE is often replaced by the mean absolute percentage error (MAPE).**

**The closer MAPE approaches zero, the better the forecasting results. In the above scenario, MAPE is 0.4 and hence the model produces a better forecast.**

**MASE: Mean Absolute Scaled Error:** It**is a measure of the**[**accuracy**](https://en.wikipedia.org/wiki/Accuracy)**of**[**forecasts**](https://en.wikipedia.org/wiki/Forecasting)**. It is the mean absolute error of the forecast values, divided by the mean absolute error of the in-sample one-step naive forecast. The closer the number is to 0, the better the model.**

**Here the value is 0.4576, indicating that the model produces 45% bigger errors than a Naïve forecast would have produced.**

**ACF1: Autocorrelation of errors at lag 1: In the above model, it comes to -0.009847391, which means that the correlation is almost zero and hence a good fit.**