Practical No.5

AIM: Practical of Time-series forecasting.

Theory:

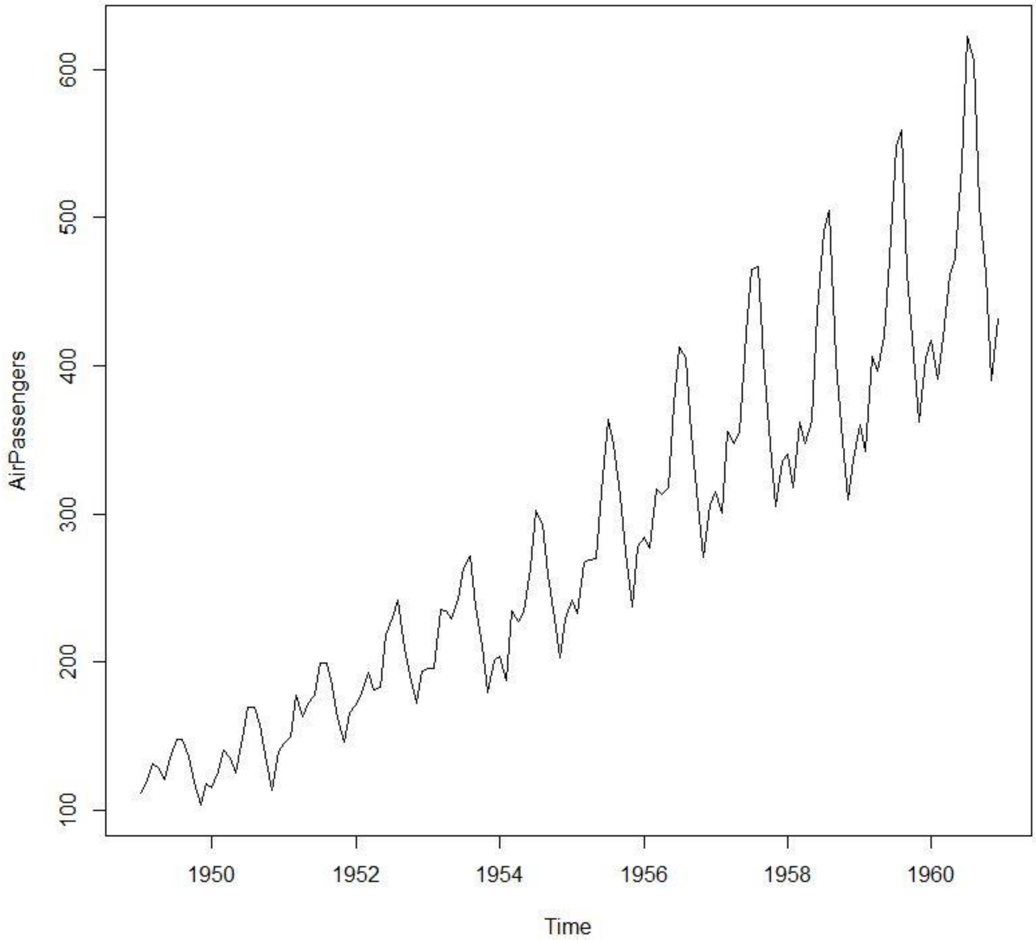
Making predictions about the future is called extrapolation in the classical statistical handling of time series data.
More modern fields focus on the topic and refer to it as time series forecasting.
Forecasting involves taking models fit on historical data and using them to predict future observations.
Descriptive models can borrow for the future (i.e. to smooth or remove noise), they only seek to best describe the data.
An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened.
The skill of a time series forecasting model is determined by its performance at predicting the future. This is often at the expense of being able to explain why a specific prediction was made, confidence intervals and even better understanding the underlying causes behind the problem.
Exploration of Time Series Data in R:
Here we’ll learn to handle time series data on R. Our scope will be restricted to data exploring in a time series type of data set and not go to building time series models.

I have used an inbuilt data set of R called AirPassengers. The dataset consists of monthly totals of international airline passengers, 1949 to 1960.
Loading the Data Set
Following is the code which will help you load the data set and spill out a few top level metrics.

```
> data(AirPassengers)

> class(AirPassengers)
[1] "ts"
#This tells you that the data series is in a time series format

> start(AirPassengers)
[1] 1949    1
#This is the start of the time series

> end(AirPassengers)
[1] 1960   12
#This is the end of the time series

> frequency(AirPassengers)
[1] 12
#The cycle of this time series is 12months in a year
> summary(AirPassengers)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  104.0   180.0   265.5   280.3   360.5   622.0 
```

#The number of passengers are distributed across the spectrum

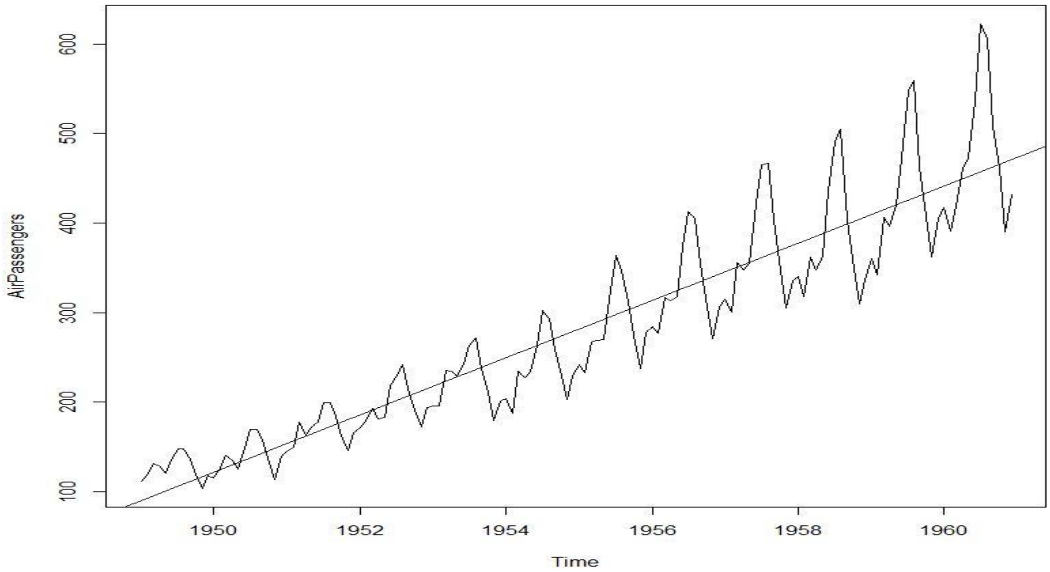
> plot(AirPassengers)



#This will plot the time series

> abline(reg=lm(AirPassengers~time(AirPassengers)))

This will fit in a line



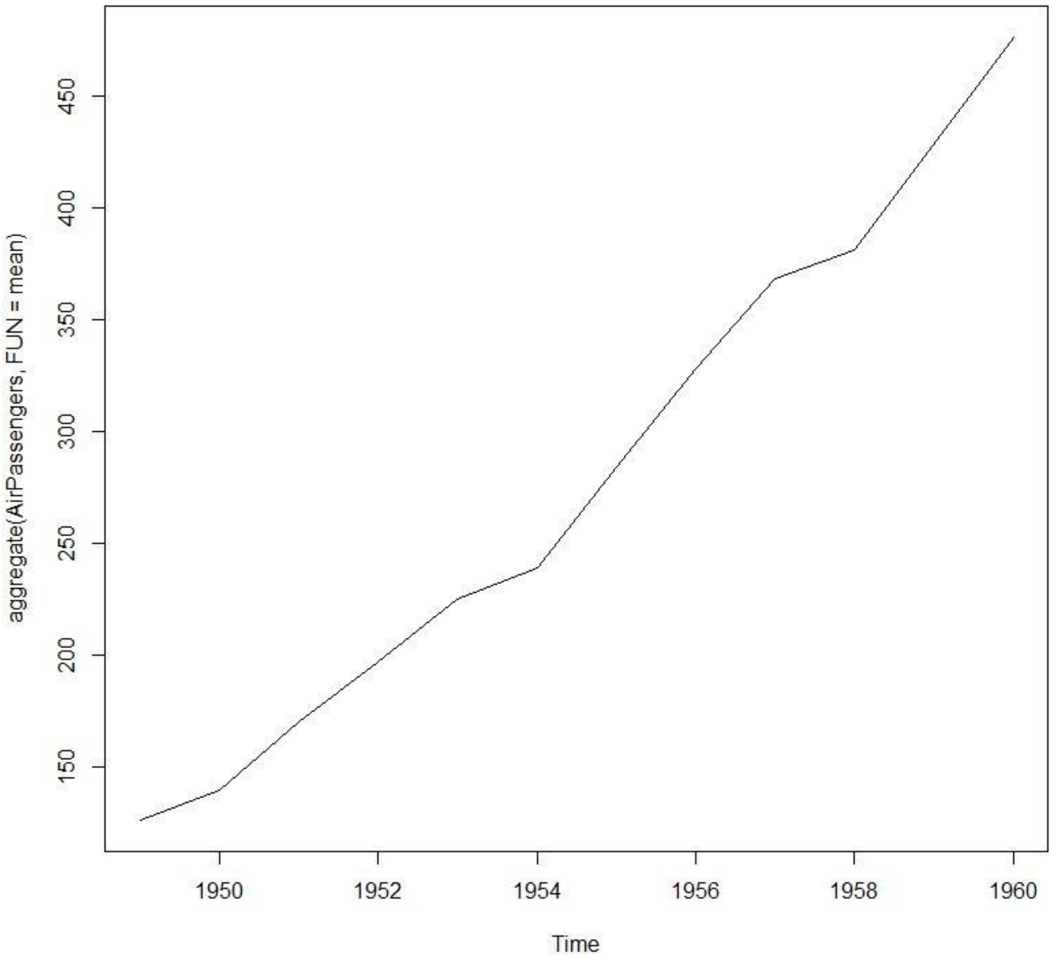
> cycle(AirPassengers)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	1	2	3	4	5	6	7	8	9	10	11	12
1950	1	2	3	4	5	6	7	8	9	10	11	12
1951	1	2	3	4	5	6	7	8	9	10	11	12
1952	1	2	3	4	5	6	7	8	9	10	11	12
1953	1	2	3	4	5	6	7	8	9	10	11	12
1954	1	2	3	4	5	6	7	8	9	10	11	12
1955	1	2	3	4	5	6	7	8	9	10	11	12
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
1958	1	2	3	4	5	6	7	8	9	10	11	12
1959	1	2	3	4	5	6	7	8	9	10	11	12
1960	1	2	3	4	5	6	7	8	9	10	11	12

#This will print the cycle across years.

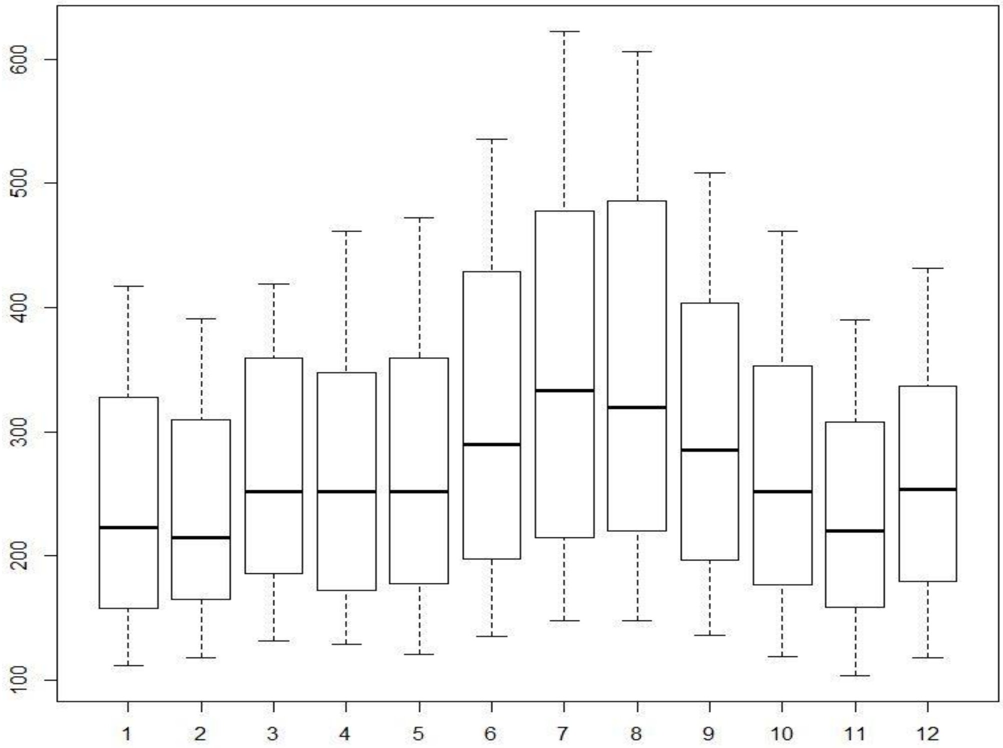
```
> plot(aggregate(AirPassengers,FUN=mean))
```

#This will aggregate the cycles and display a year on year trend



```
> boxplot(AirPassengers~cycle(AirPassengers))
```

#Box plot across months will give us a sense on seasonal effect



Important Inferences

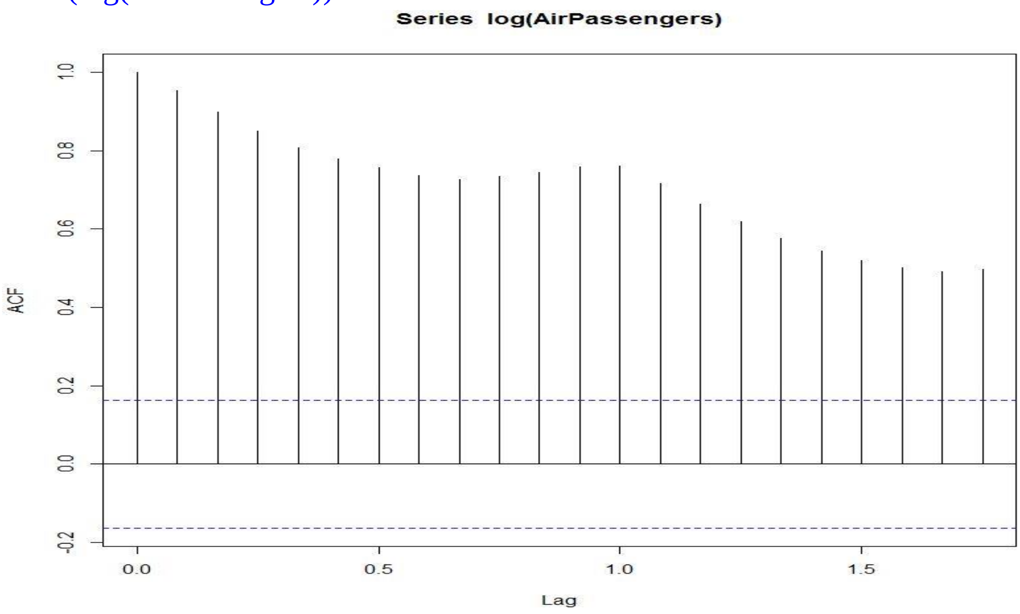
1. The year on year trend clearly shows that the #passengers have been increasing without fail.
2. The variance and the mean value in July and August is much higher than rest of the months.
3. Even though the mean value of each month is quite different their variance is small. Hence, we have strong seasonal effect with a cycle of 12 months or less.

Exploring data becomes most important in a time series model – without this exploration, you will not know whether a series is stationary or not. As in this case we already know many details about the kind of model we are looking out for.

Let's now take up a few time series models and their characteristics. We will also take this problem forward and make a few predictions.

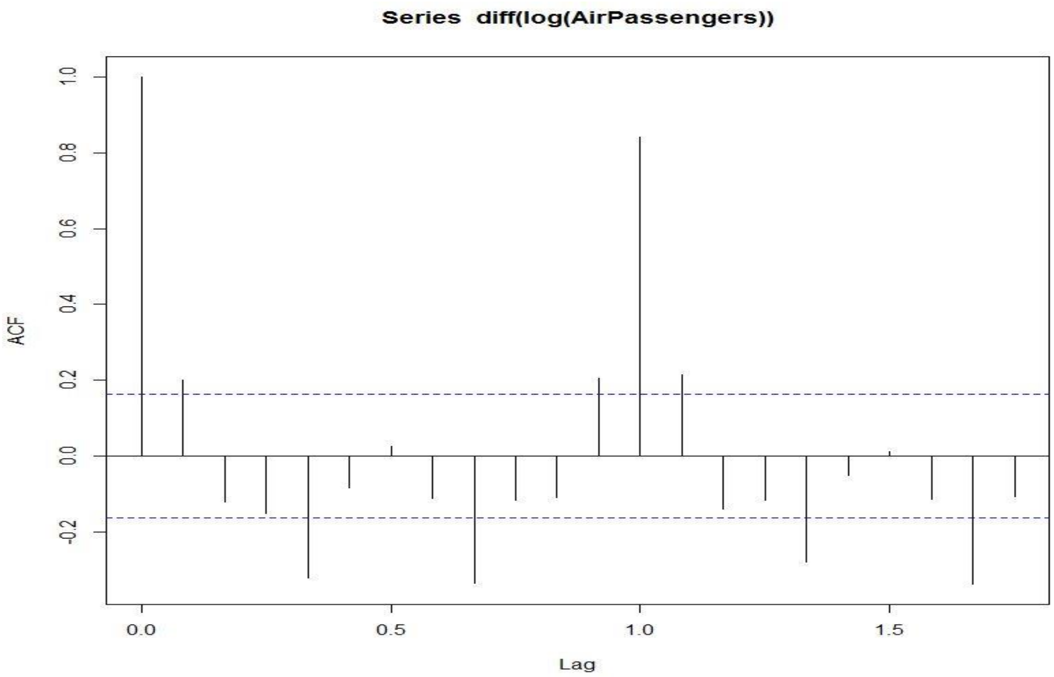
Auto – correlation Function(ACF): ACF is a plot of total correlation between different lag functions.
Following are the ACF plots for the series :

> acf(log(AirPassengers))

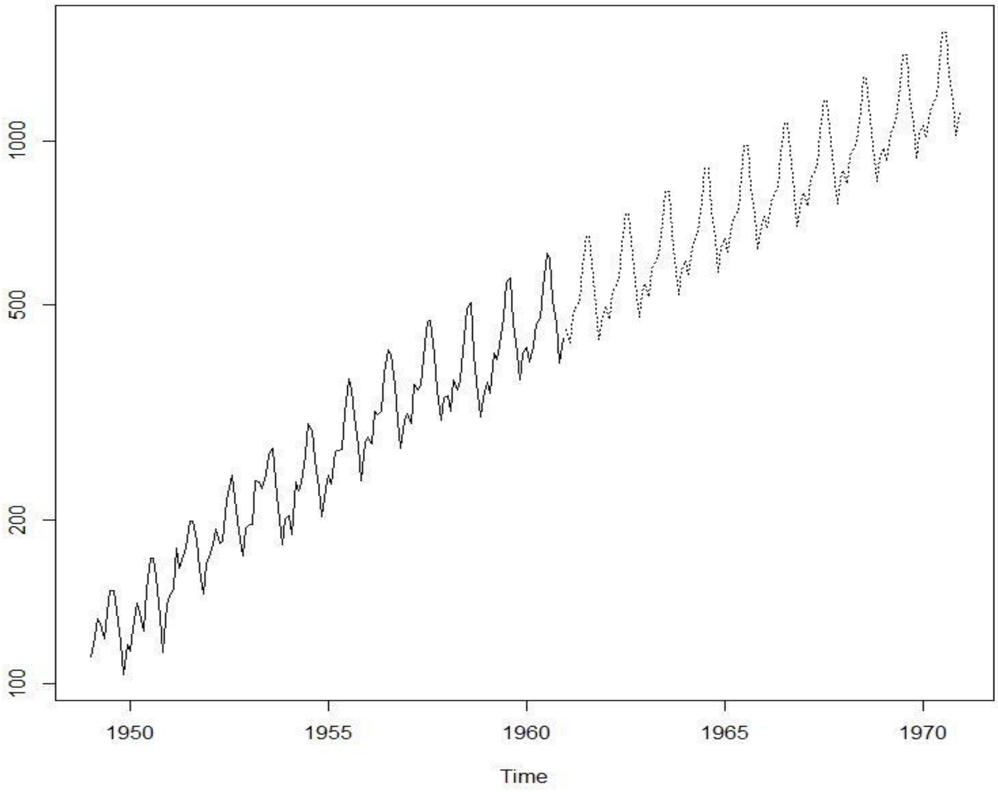


What do you see in the chart shown above?
Clearly, the decay of ACF chart is very slow, which means that the population is not stationary. We have already discussed above that we now intend to regress on the difference of logs rather than log directly. Let's see how ACF curve come out after regressing on the difference.

> acf(diff(log(AirPassengers)))



```
> (fit <- arima(log(AirPassengers), c(0, 1, 1),seasonal = list(order = c(0, 1, 1), period = 12)))

Call:
arima(x = log(AirPassengers), order = c(0, 1, 1), seasonal = list(order = c(0,
    1, 1), period = 12))

Coefficients:
         ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = -483.4

> pred <- predict(fit, n.ahead = 10*12)
> ts.plot(AirPassengers,2.718^pred$pred, log = "y", lty = c(1,3))
```



All Command:

```
data(AirPassengers)
class(AirPassengers)
start(AirPassengers)
end(AirPassengers)

frequency(AirPassengers)
summary(AirPassengers)
plot(AirPassengers)

abline(reg=lm(AirPassengers~time(AirPassengers)))

cycle(AirPassengers)

plot(aggregate(AirPassengers,FUN=mean))

boxplot(AirPassengers~cycle(AirPassengers))

acf(log(AirPassengers))

acf(diff(log(AirPassengers)))

(fit <- arima(log(AirPassengers), c(0, 1, 1),seasonal = list(order = c(0, 1, 1), period = 12)))

pred <- predict(fit, n.ahead = 10*12)

ts.plot(AirPassengers,2.718^pred$pred, log = "y", lty = c(1,3))
```