**CSE3506 Essentials of Data Analytics**

24/02/2022

# Name: V Maheysh

# Reg No: 19BCE1120

# Lab Exercise: 4 - Time Series Forecasting (Air Passengers)

### Objective: To apply time series forecasting on the AirPassengers dataset and visualize the decomposed datasset.

### Methods:

1. Store AirPassengers dataset as “x”.
2. Display the dataset.
3. Check is there is null value.
4. Look over the statistical summary of the dataset.
5. Plot no. of passengers year-wise.
6. Plot no. of passengers month-wise.
7. Decompose “x” and store in “ddata”.
8. Plot ddata / Trends/Seasons/Random separately.
9. Plot ACF and PACF
10. Build ARIMA model using x.

### Importing Libraries

#install.packages("forecast")  
library(ggfortify)

## Warning: package 'ggfortify' was built under R version 4.1.2

## Loading required package: ggplot2

library(tseries)

## Warning: package 'tseries' was built under R version 4.1.2

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library(forecast)

## Warning: package 'forecast' was built under R version 4.1.2

## Registered S3 methods overwritten by 'forecast':  
## method from   
## autoplot.Arima ggfortify  
## autoplot.acf ggfortify  
## autoplot.ar ggfortify  
## autoplot.bats ggfortify  
## autoplot.decomposed.ts ggfortify  
## autoplot.ets ggfortify  
## autoplot.forecast ggfortify  
## autoplot.stl ggfortify  
## autoplot.ts ggfortify  
## fitted.ar ggfortify  
## fortify.ts ggfortify  
## residuals.ar ggfortify

library(ggplot2)

### Importing Dataset

data("AirPassengers")  
data <- AirPassengers

### 1) Displaying dataset

data

## Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec  
## 1949 112 118 132 129 121 135 148 148 136 119 104 118  
## 1950 115 126 141 135 125 149 170 170 158 133 114 140  
## 1951 145 150 178 163 172 178 199 199 184 162 146 166  
## 1952 171 180 193 181 183 218 230 242 209 191 172 194  
## 1953 196 196 236 235 229 243 264 272 237 211 180 201  
## 1954 204 188 235 227 234 264 302 293 259 229 203 229  
## 1955 242 233 267 269 270 315 364 347 312 274 237 278  
## 1956 284 277 317 313 318 374 413 405 355 306 271 306  
## 1957 315 301 356 348 355 422 465 467 404 347 305 336  
## 1958 340 318 362 348 363 435 491 505 404 359 310 337  
## 1959 360 342 406 396 420 472 548 559 463 407 362 405  
## 1960 417 391 419 461 472 535 622 606 508 461 390 432

### 2) Check for unfilled data

sum(is.na(data))

## [1] 0

### Inference: There are no unfilled data in the dataset.

### 3) Display the statistical info of the dataset such as min, max, 1st quartile, 3rd quartile, mean and median.

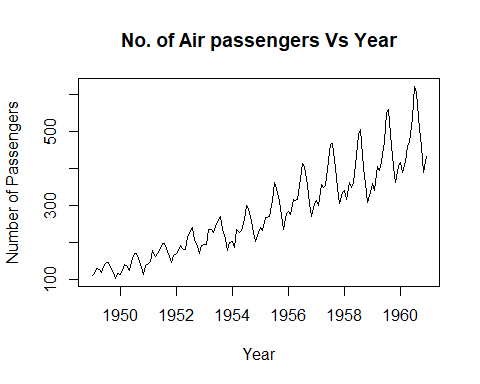
summary(data)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 104.0 180.0 265.5 280.3 360.5 622.0

**Inference**: Min value = 104 1st Quartile = 180.0 Median = 265.5 Mean = 280.3 3rd Quartile = 360.5 Max = 622

### 4) Plot ‘data’ (No. of Air passengers Vs Year)

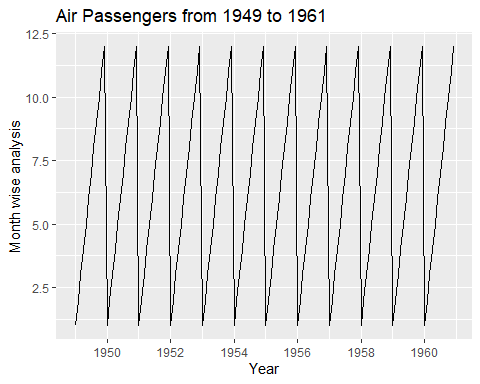
plot(data,xlab="Year", ylab = "Number of Passengers",main="No. of Air passengers Vs Year")



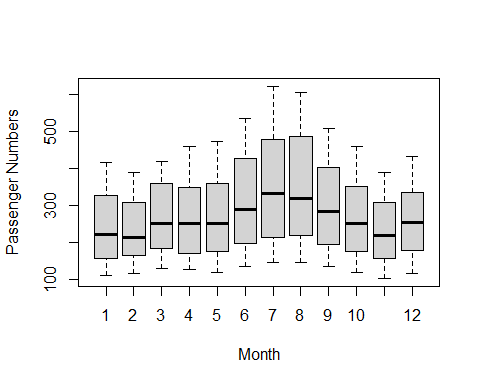
**Inference**: The number of passengers increase , as the year increases.

### 5) Plot as timeseries ‘data’ (monthwise)

autoplot(cycle(data)) + labs(x ="Year", y = "Month wise analysis", title="Air Passengers from 1949 to 1961")



boxplot(data~cycle(data),xlab="Month", ylab = "Passenger Numbers")



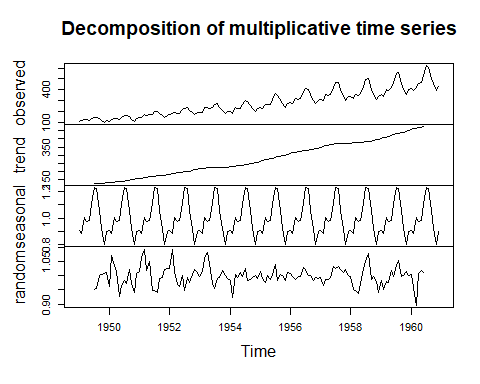
### Inference: It can be inferred from the boxplot that the total number of passengers in months 6 to 9 are higher and have greater mean and IQR range. Hence there is a seasonality variation in the 12-month cycle.

### 6) Decompose the data as multiplicative and store as ‘ddata’..

ddata <- decompose(data,"multiplicative")  
#ddata

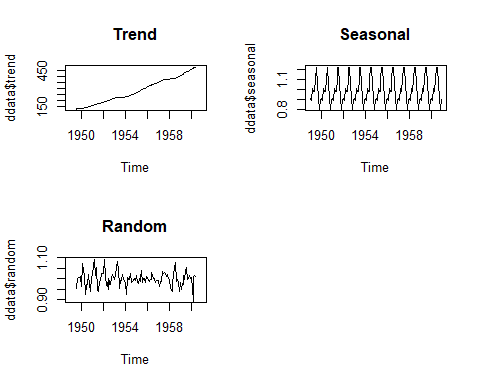
### 7) Plot dddata.

plot(ddata)



### 8) Plot the following: trend, seasonal and random separately.

par(mfrow=c(2,2))  
plot(ddata$trend, main="Trend")  
plot(ddata$seasonal, main="Seasonal")  
plot(ddata$random, main="Random")



**Inference:** Although there is a overall increase of passengers in the trend, there is also a seasonal increase and decrease.

### 9) Perform ADF test for stationarity

adf.test(data)

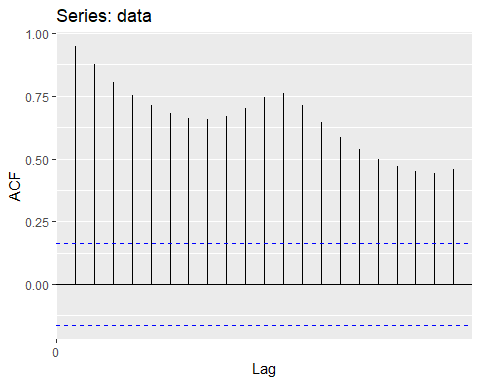
## Warning in adf.test(data): p-value smaller than printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: data  
## Dickey-Fuller = -7.3186, Lag order = 5, p-value = 0.01  
## alternative hypothesis: stationary

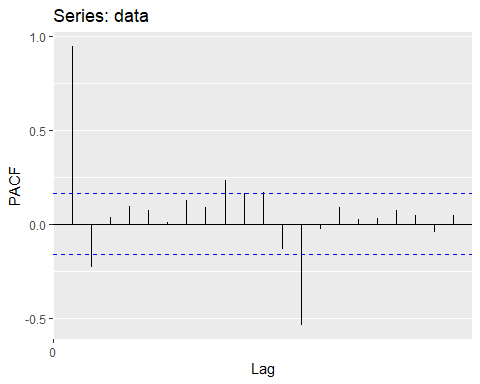
### Inference: p-value here is 0.01,which is less than 5%.This indicates a strong evidence against the null hypothesis, as there is less than a 5% probability the null hypothesis is correct (and the results are random). Hence the alternative hypothesis that the time series is stationaary is accepted.

### 10) Plot ACF and PACF.

#ACF  
autoplot(acf(data,plot = FALSE))



#PACF  
autoplot(pacf(data,plot=FALSE))



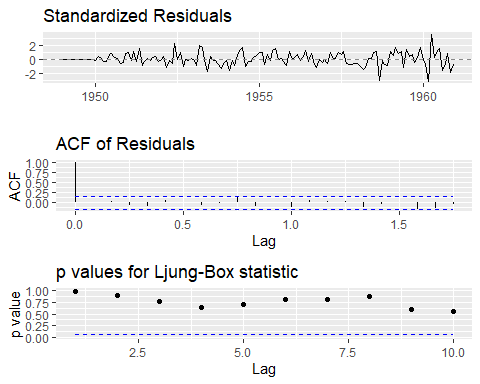
### Inference: There is a positive relationship within the 12 month cycle.

### 11) Model using ARIMA.

model <- auto.arima(data)  
model

## Series: data   
## ARIMA(2,1,1)(0,1,0)[12]   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.5960 0.2143 -0.9819  
## s.e. 0.0888 0.0880 0.0292  
##   
## sigma^2 = 132.3: log likelihood = -504.92  
## AIC=1017.85 AICc=1018.17 BIC=1029.35

ggtsdiag(model)



### Inference: The model has a first lag autoregressive term at model period 12 months. A smaller AIC value indicates a better model fit.

### Conclusion: No data cleaning is required, as there are no outliers or missing values. The AirPassengers dataset appears to be a multiplicative time series. This means that as the number of passengers grows, so does the seasonality pattern.The alternative hypothesis was accepted. The ARIMA model was finally created, with a period of 12 months.