Introduction to Time Series Analysis

DSLA COURSE

ROHIT PADEBETTU

Course Assignments

Programming Assignments

Reading Assignments

Presentation Assignments

Technical Skills Assignments

Writing Assignments

Technical Assignment

Install & Setup RStudio on AWS

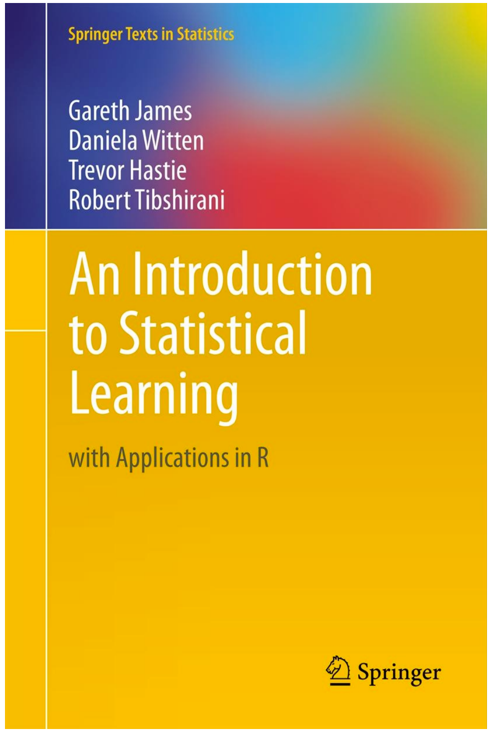
Programming Assignment

Install & Complete: Swirl – Exploratory Data Analysis

Install & Complete: Swirl – Statistical Inference

Reading Assignment

Read Chapter 5: Resampling Methods



Writing Assignment

Submit by Saturday

Written Report (not to exceed 15 pages) on Mushroom Classification Case

Presentation Assignment

By Saturday Submit

Your Presentations on Mushroom Classification Case

1. Technical Presentation
2. Business Presentation (Not to exceed 5 slides)

Types of Data Sets

There can be 3 types of data sets

- **Cross-Sectional**
    - All data collected at specific point in time
    - Does not vary over time
    - Example: survey of heights of 100 subjects
- **Time-series**
    - Data is collected over time
    - Changes with time
    - Example: weather, sales, market index, prices
- **Panel data**
    - Panel data consists of data that is both time series and cross sectional.

Where it is used ?

1. Agricultural crop yields forecasting
2. Unit product sales forecasting
3. Average price forecasting (gas prices, inflation, rental prices)
4. Unemployment rate forecasting (city, state, national)
5. Utilization demand forecasting (server, web traffic, industrial machines, utilities)
6. Forecasting birth rate or hospitalization
7. Forecasting Seizures, Heart attacks
8. Forecasting size of certain populations (rabbits, rodents, bacteria, humans)
9. Forecasting passengers in train station or traffic on certain roads
10. Stock Price Forecasting (do not try at home!)

Time Series Analysis

Time Series Description

- *To best capture or describe time series*
- *To understand the underlying causes*

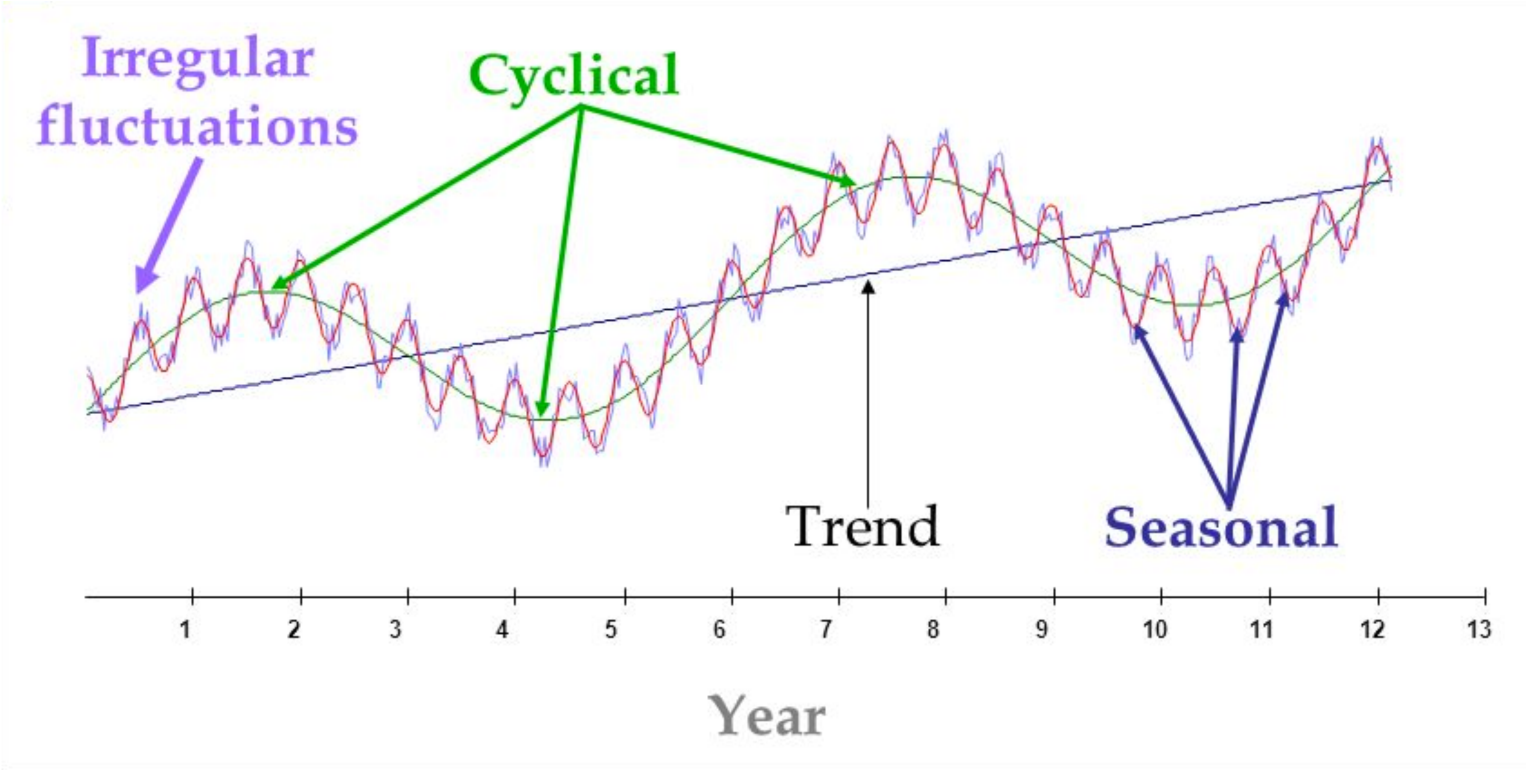
Time Series Forecasting

- *Predictions about the future within confidence intervals*
- *Using understanding and models built from the past*

How?

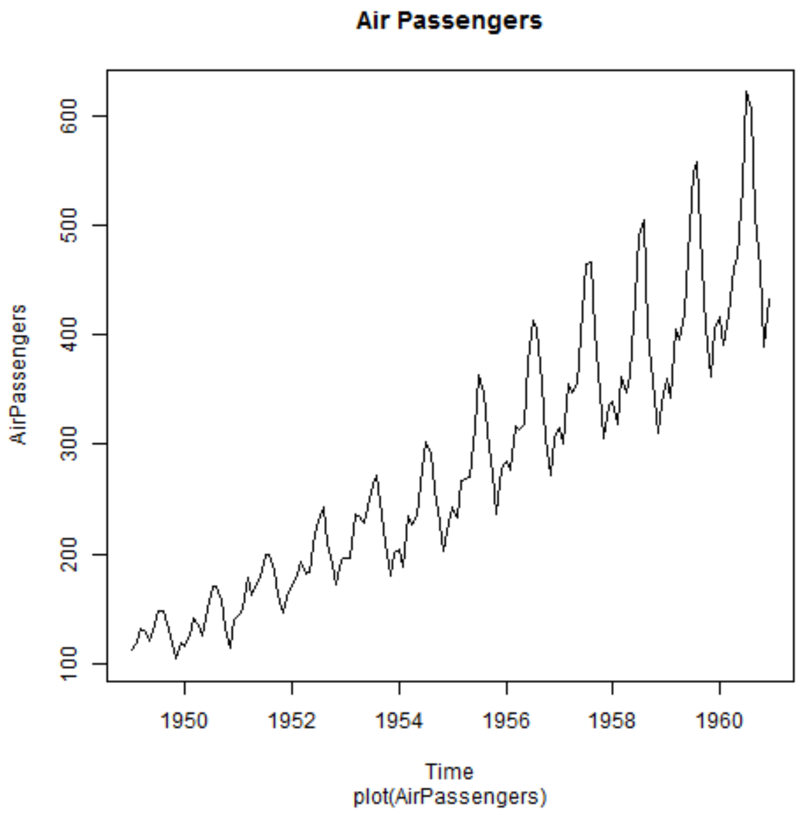
- *Decomposing a time series*
- *Modeling the parts*

Time Series Components

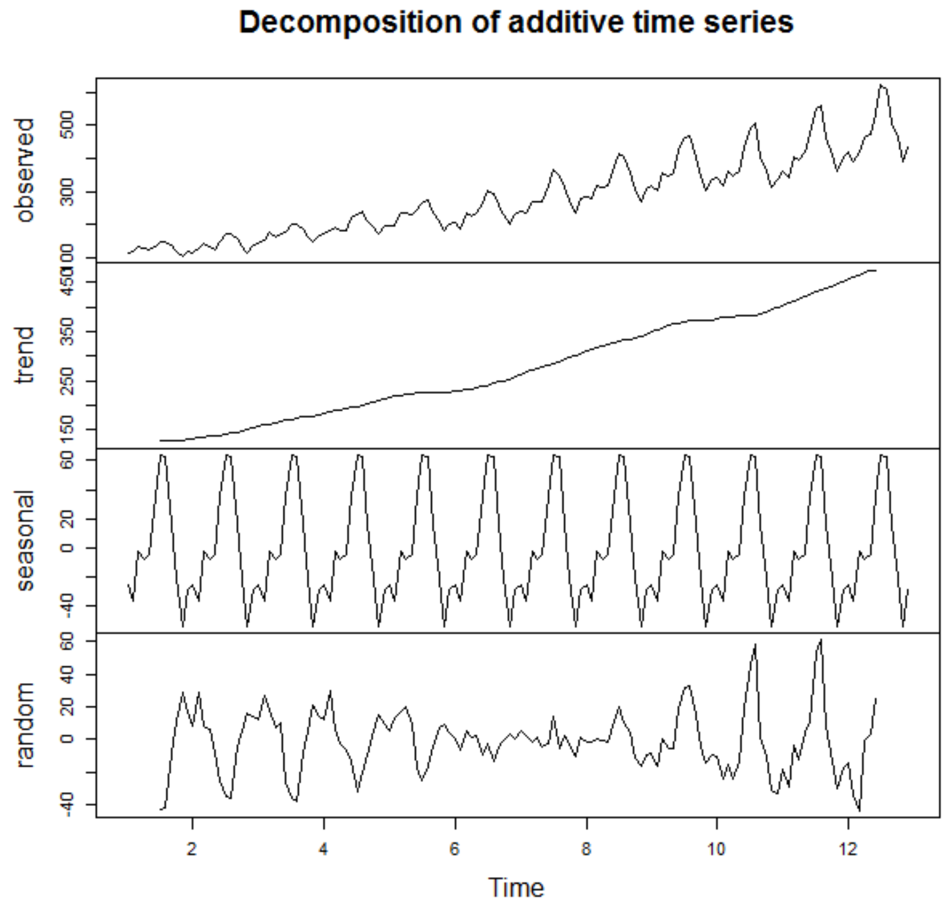


Trend	Overall Persistent Long Term movement
Seasonal	Regular periodic fluctuations, within 12 months
Cyclical	Repeated movements over 1 year
Irregular Random	Erratic or residual fluctuations

Time Series Decomposition

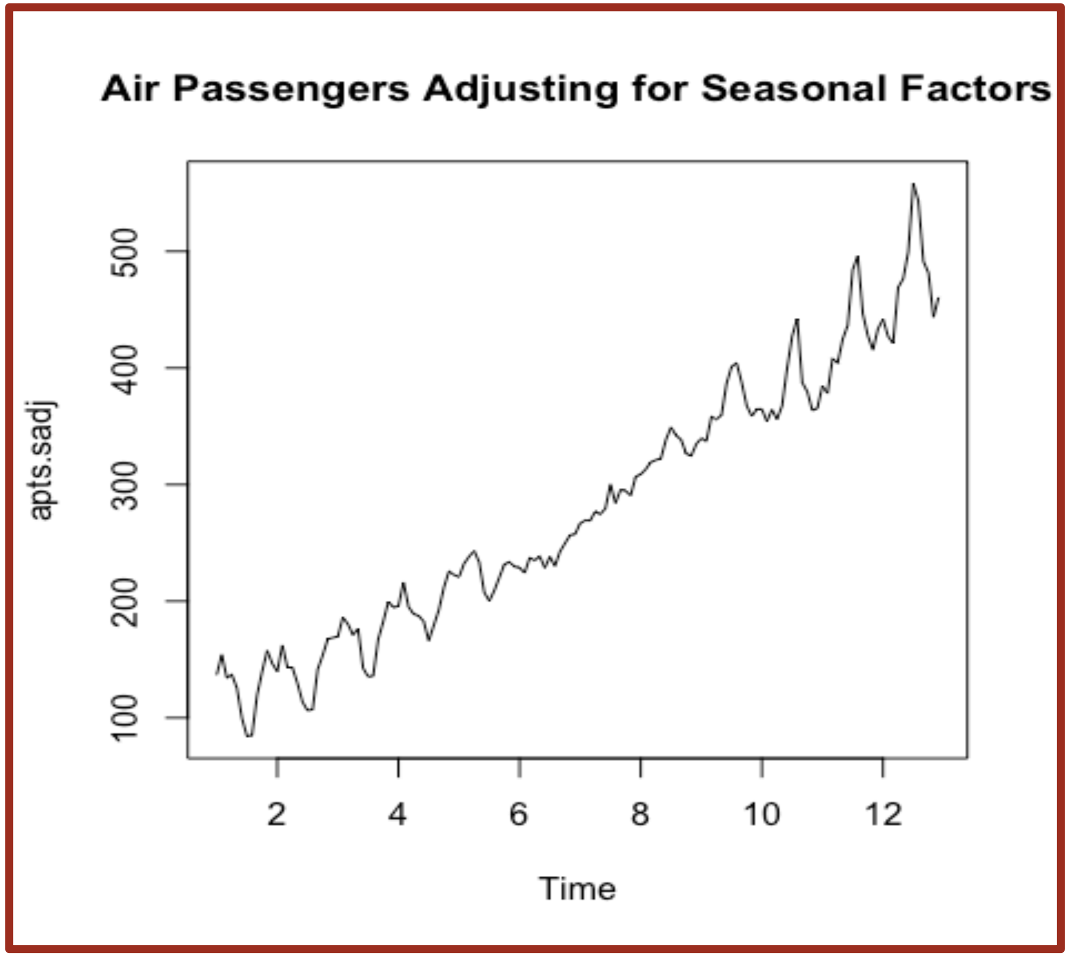


decompose()



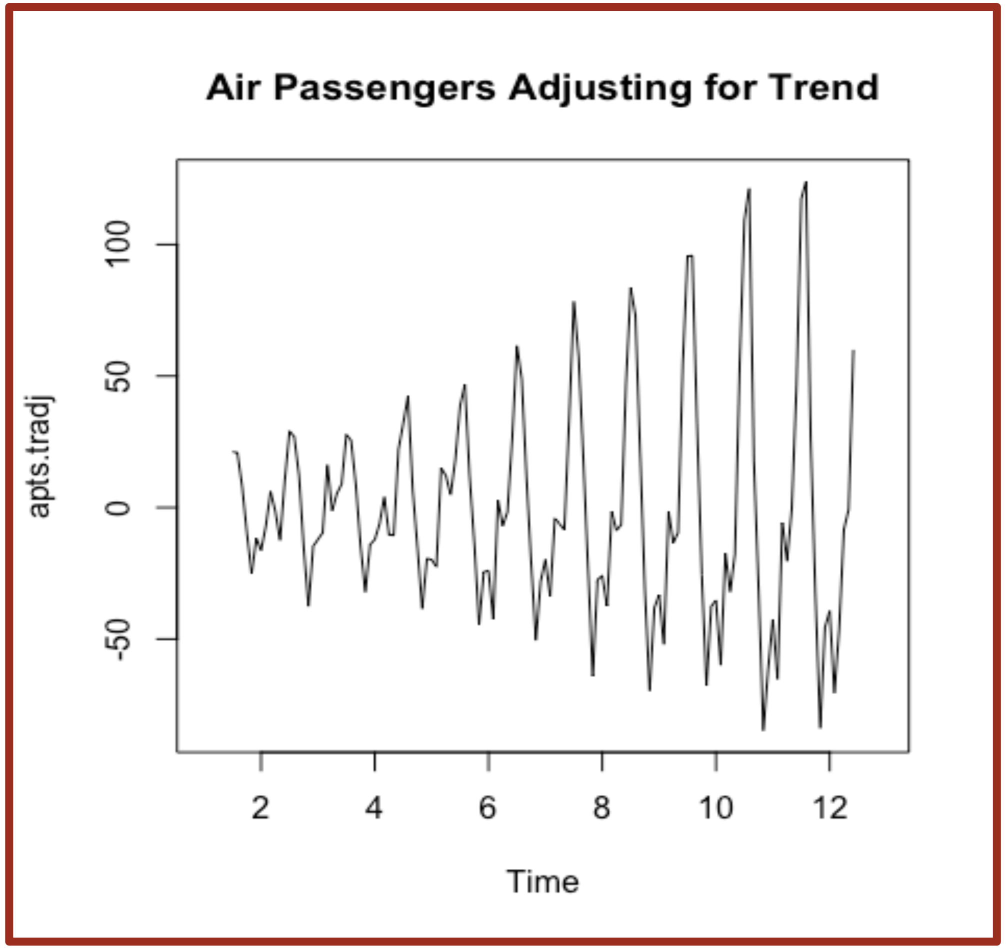
Time Series - Seasonal Adjustments

AirPassengers – AirPassengers$seasonal



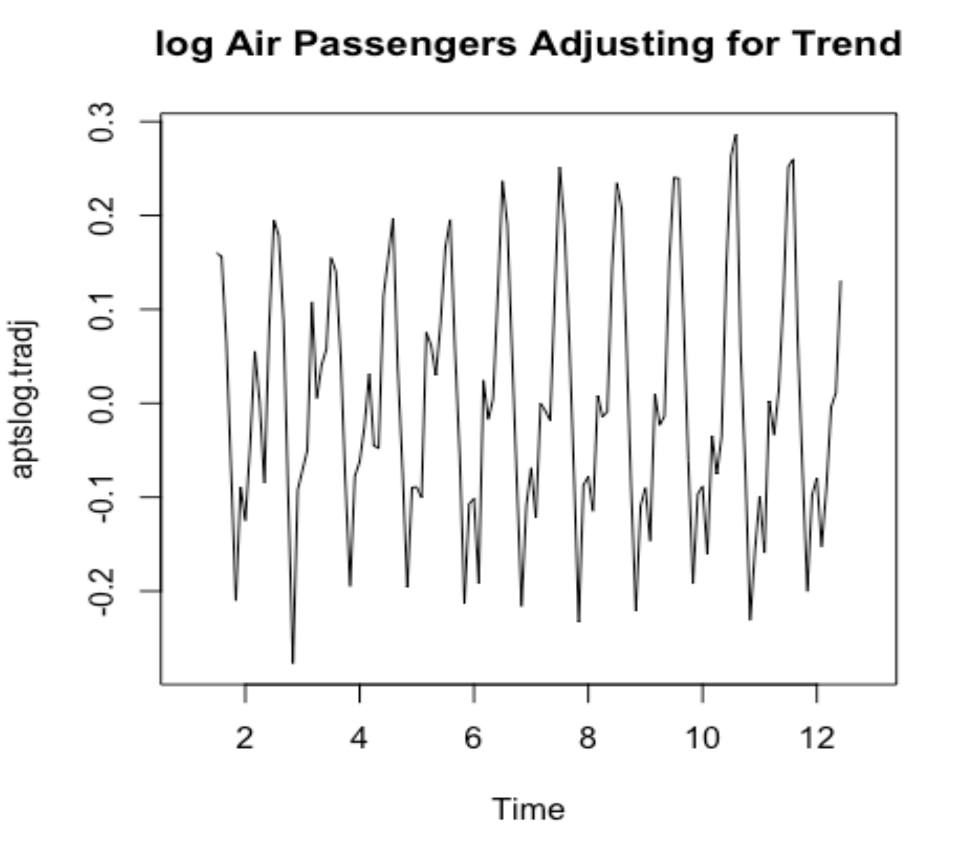
Time Series- Trend Adjusted

AirPassengers – AirPassengers$trend



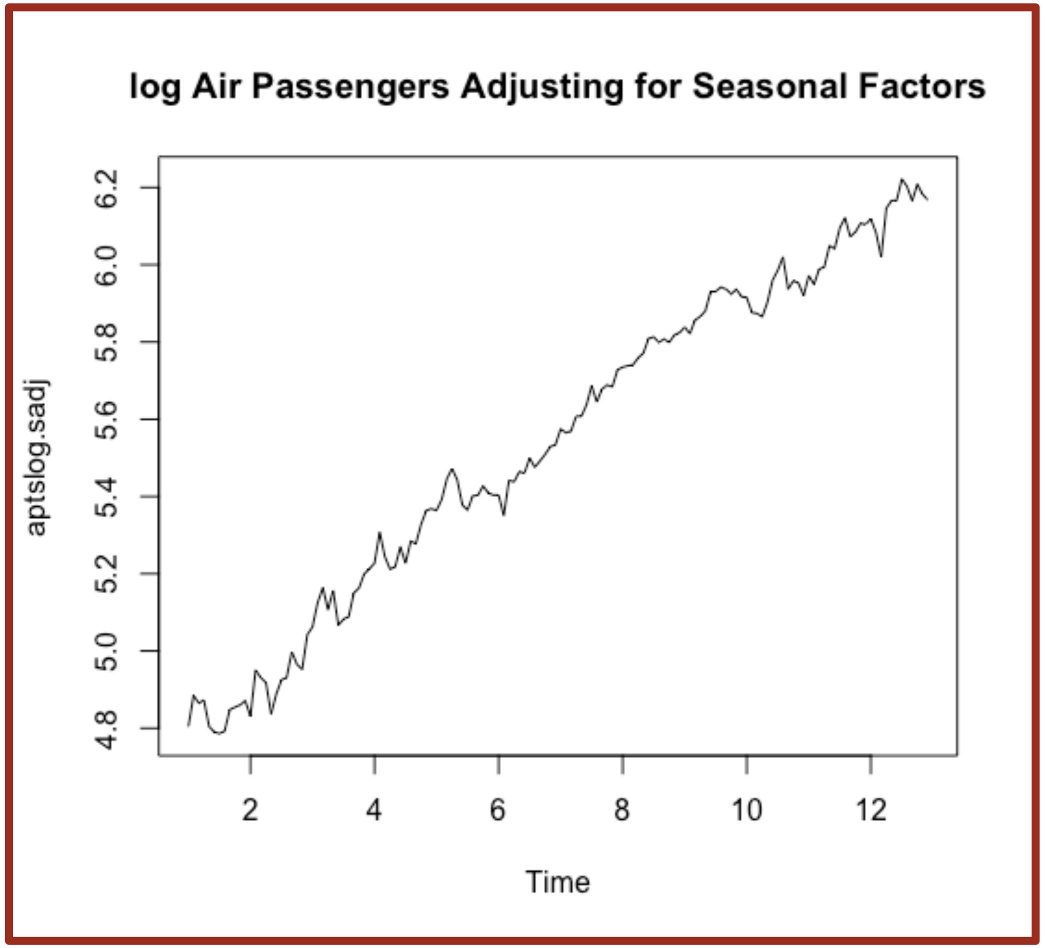
Time Series- Trend Adjusted (log)

log(AirPassengers) – log(AirPassengers)$trend



Time Series - Seasonal Adjustments (log)

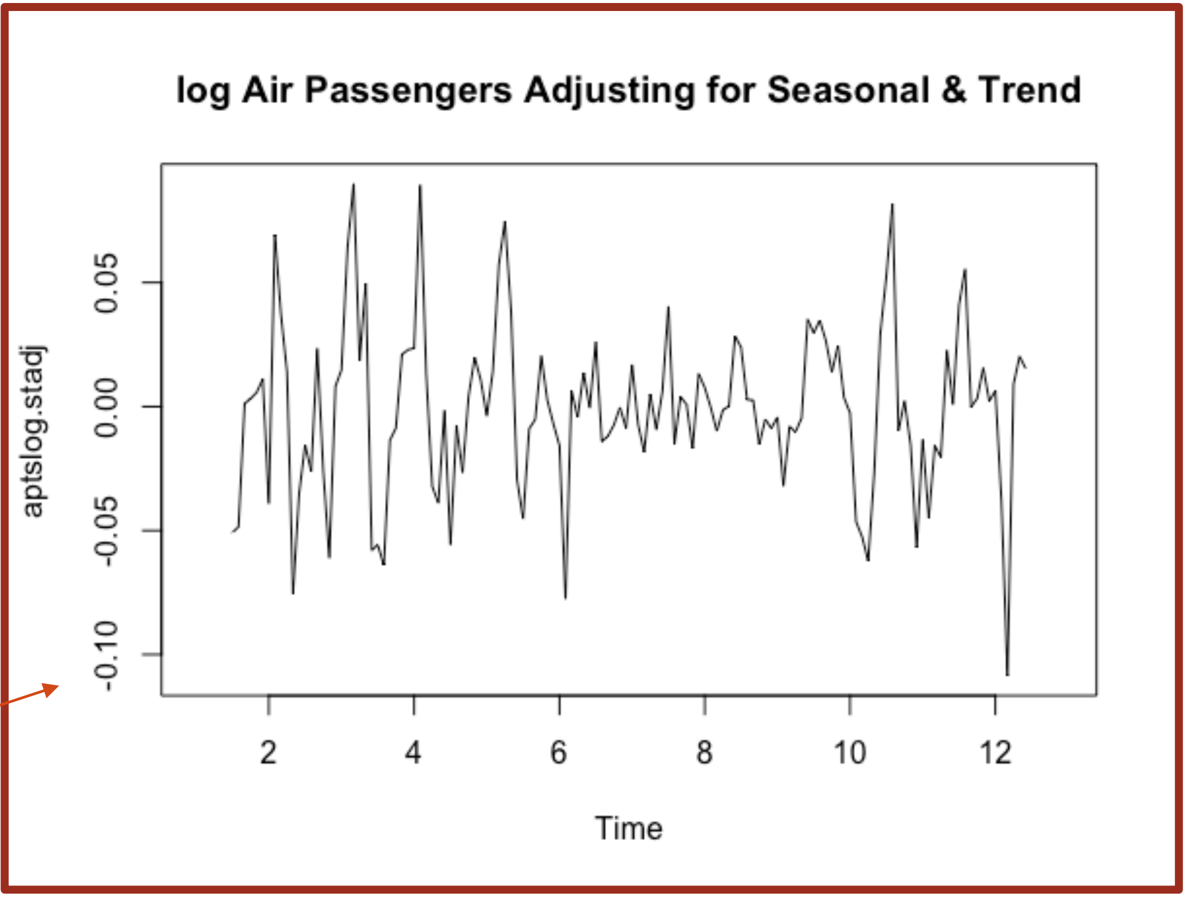
log(AirPassengers) – log(AirPassengers)$seasonal



Time Series – Random Fluctuations

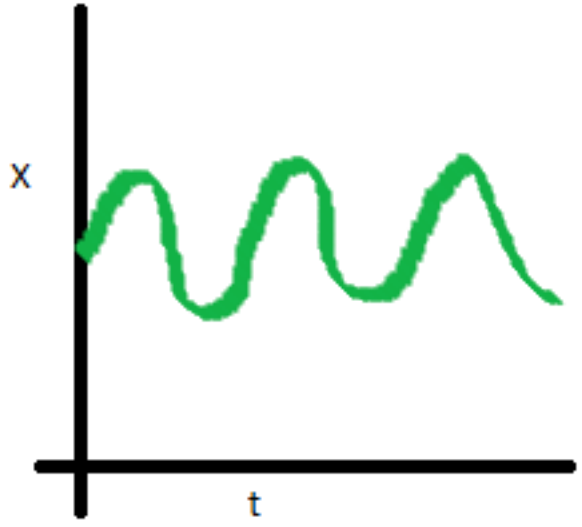
log(AirPassengers) –
log(AirPassengers)$seasonal
- log(AirPassengers)$trend

Stationary time series

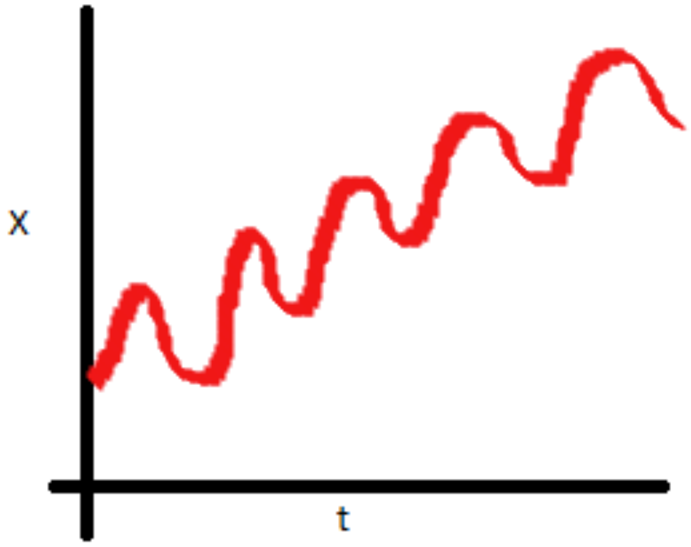


Stationary Time Series

Criteria 1: Mean of the series should not change with time



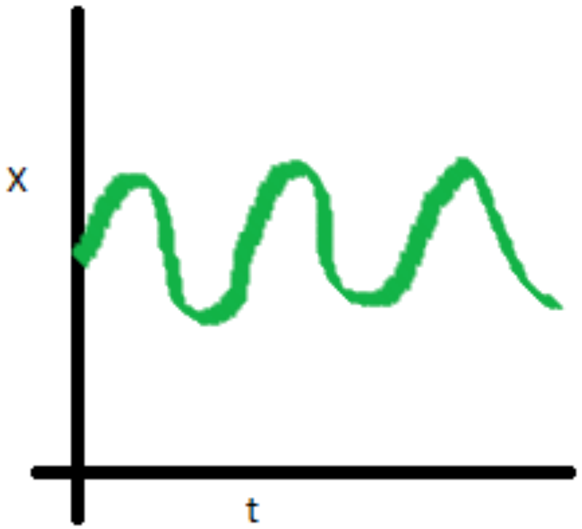
Stationary series



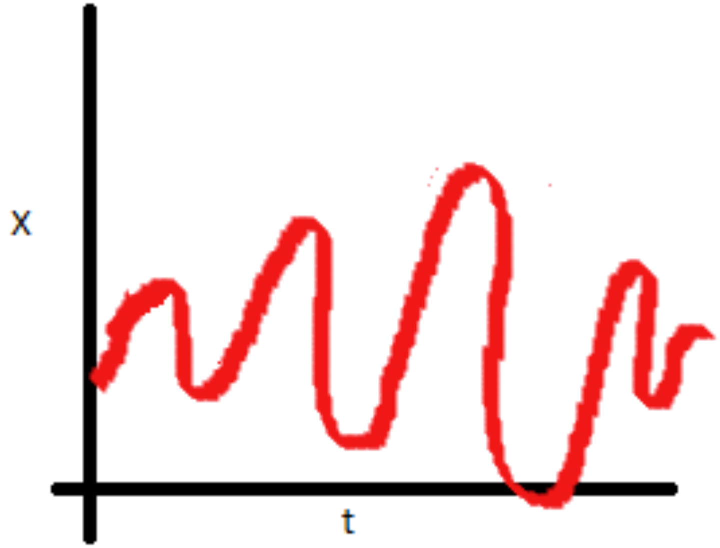
Non-Stationary series

Stationary Time Series

Criteria 2: Variance of the series should not change with time



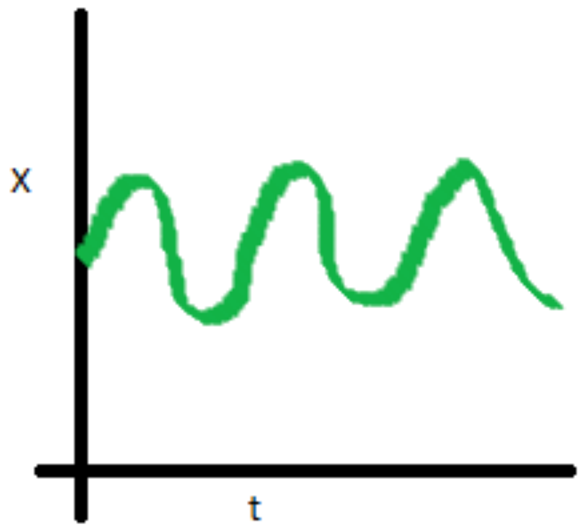
Stationary series



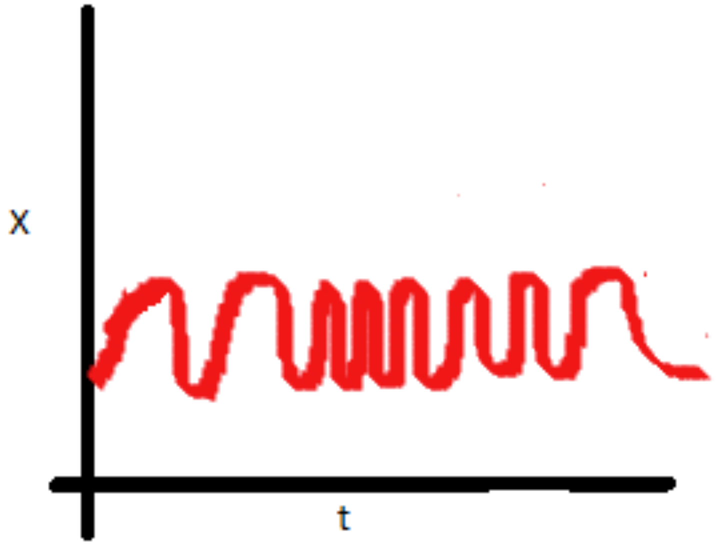
Non-Stationary series

Stationary Time Series

Criteria 3: Auto Covariance of series should not change with time



Stationary series



Non-Stationary series

Stationary Time Series

DSLA
DATA SCIENCE LEARNING ACADEMY

Why do we care ?

To model and forecast the time invariant component using statistical techniques and models.

Stationary series are easy to predict as statistical properties will stay same in future as they are in past!

How do we do it ?

Various methods such as:

- Box Cox Transformations
- De-trending
- Seasonal Adjustments
- Differencing

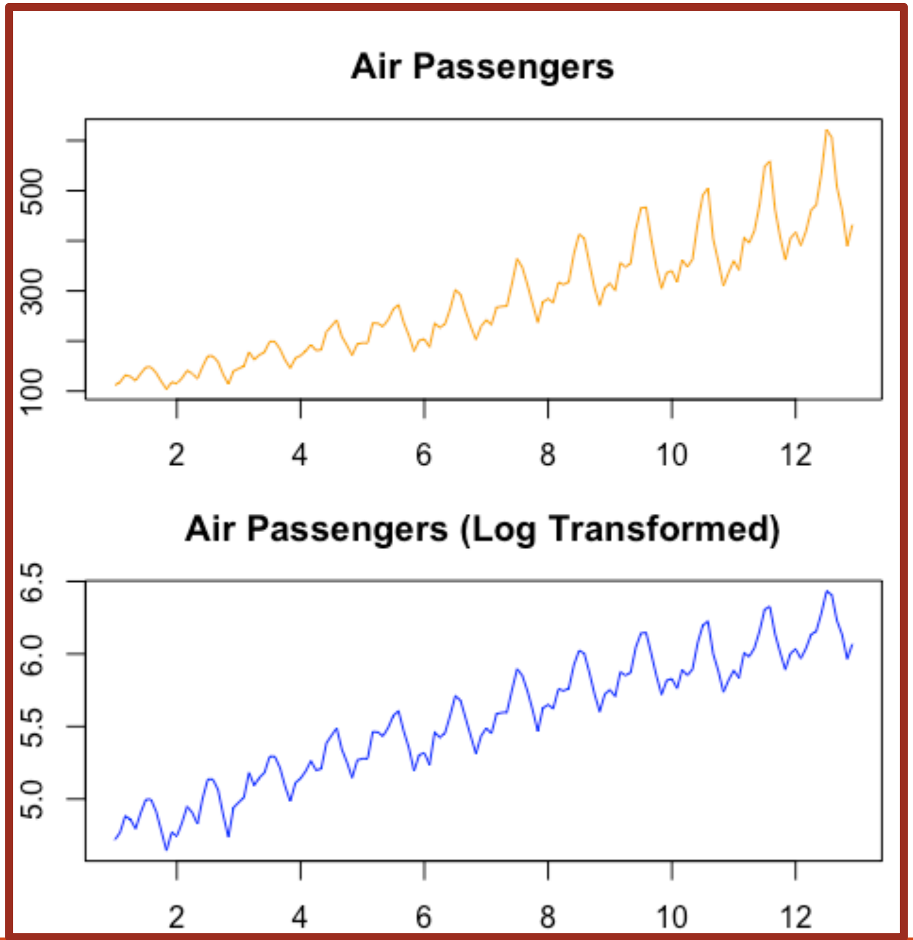
"I have seen the future and it is very much like the present, only longer."

Kehlog Albran
The Profit

Box Cox Transformations

Box Cox Transformations of time series are used to help stabilize the variance

Common Box-Cox Transformations	
Lambda	Suitable Transformation
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(Sqrt(Y))$
0	log(Y)
0.5	$Y^{0.5} = Sqrt(Y)$
1	$Y^1 = Y$
2	Y^2



DSLA
DATA SCIENCE LEARNING ACADEMY

Differencing

- Differencing time series is a method to help stabilize the mean and bring about stationarity
- Repeated Differencing can be used to remove identifiable trends and seasonal patterns in a time series
- The differenced series is a change between two equally spaced observations in a time series

Differencing on lag i

$$Y_t' = Y_t - Y_{t-i}$$

ILLUSTRATION OF SEASONAL DIFFERENCING AFTER INFLATION ADJUSTMENT

Consumer Price Index, 1990=1.0

Auto sales ($B)

Deflated auto sales:
16.13 = 4.79 / 0.297

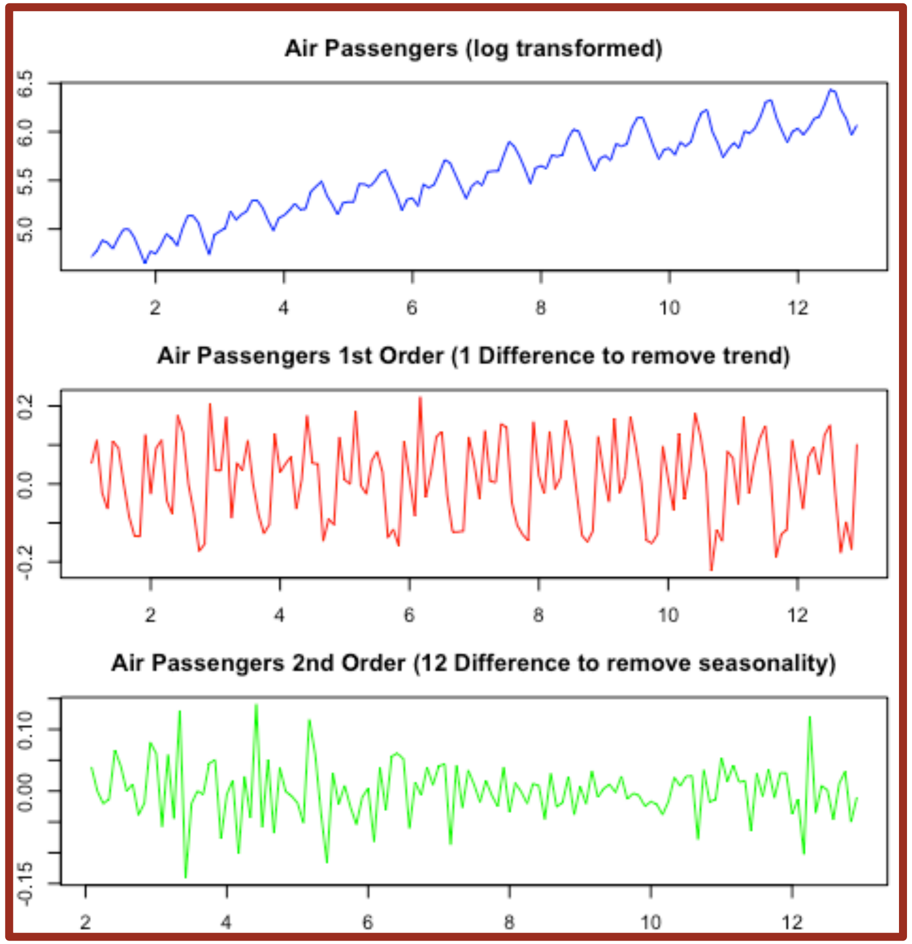
First difference of deflated auto sales:
0.51= 16.64 - 16.13, etc.

Seasonal difference of deflated auto sales:
0.44 = 16.57 - 16.13, etc.

First difference of seasonal difference:
0.72 = 1.16 - 0.44, etc.

DATE	AUTOSALE	CPI	AUTOSALE/CPI	DIFF(AUTOSALE/CPI)	SDIFF(AUTOSALE/CPI,12)	DIFF(SDIFF(AUTOSALE/CPI,12))
Jan-70	4.79	0.297	16.13			
Feb-70	4.96	0.298	16.64	0.51		
Mar-70	5.64	0.300	18.80	2.16		
Apr-70	5.98	0.302	19.80	1.00		
May-70	6.08	0.303	20.07	0.27		
Jun-70	6.55	0.305	21.48	1.41		
Jul-70	6.11	0.306	19.97	-1.51		
Aug-70	5.37	0.306	17.55	-2.42		
Sep-70	5.17	0.308	16.79	-0.76		
Oct-70	5.48	0.309	17.73	0.94		
Nov-70	4.49	0.311	14.44	-3.29		
Dec-70	4.65	0.312	14.90	0.46		
Jan-71	5.17	0.312	16.57	1.67	0.44	
Feb-71	5.57	0.313	17.80	1.23	1.16	0.72
Mar-71	6.92	0.314	22.04	4.24	3.24	2.08
Apr-71	7.10	0.315	22.54	0.50	2.74	-0.50
May-71	7.02	0.316	22.22	-0.32	2.15	-0.59
Jun-71	7.58	0.319	23.76	1.54	2.28	0.13
Jul-71	6.93	0.319	21.72	-2.04	1.75	-0.53

Differencing (Air Passengers Example)



Variance Stabilized time series still shows both Trend and Seasonality

First Order Differencing (1)

Trend is Removed, but Seasonality exists

Second Order Differencing (12)

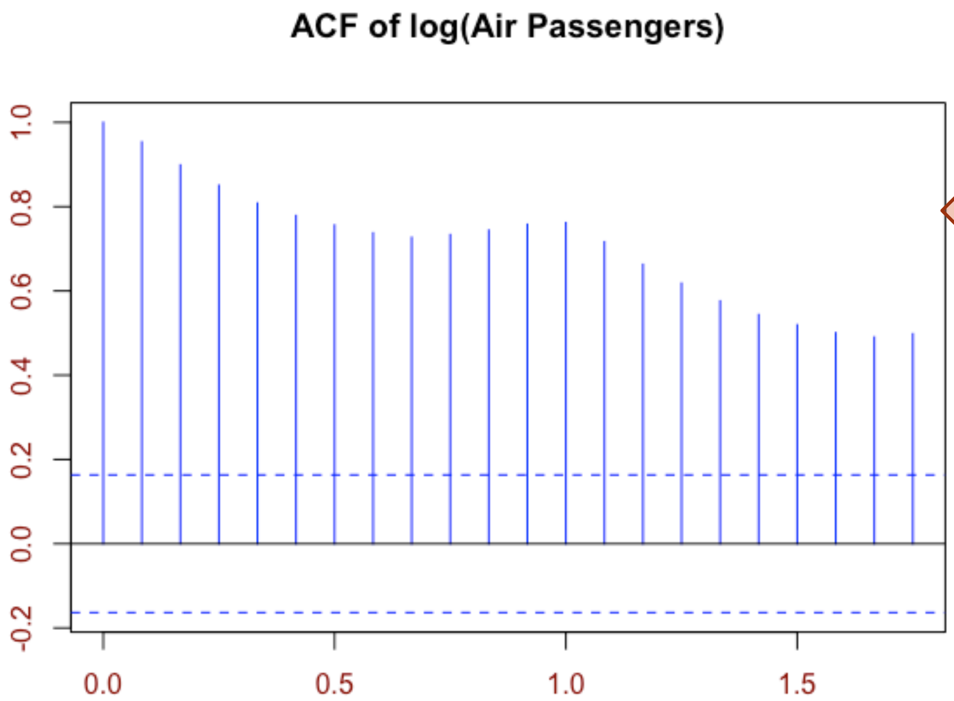
Seasonality also removed

Got Stationary Series!

Auto Correlation Function (ACF)

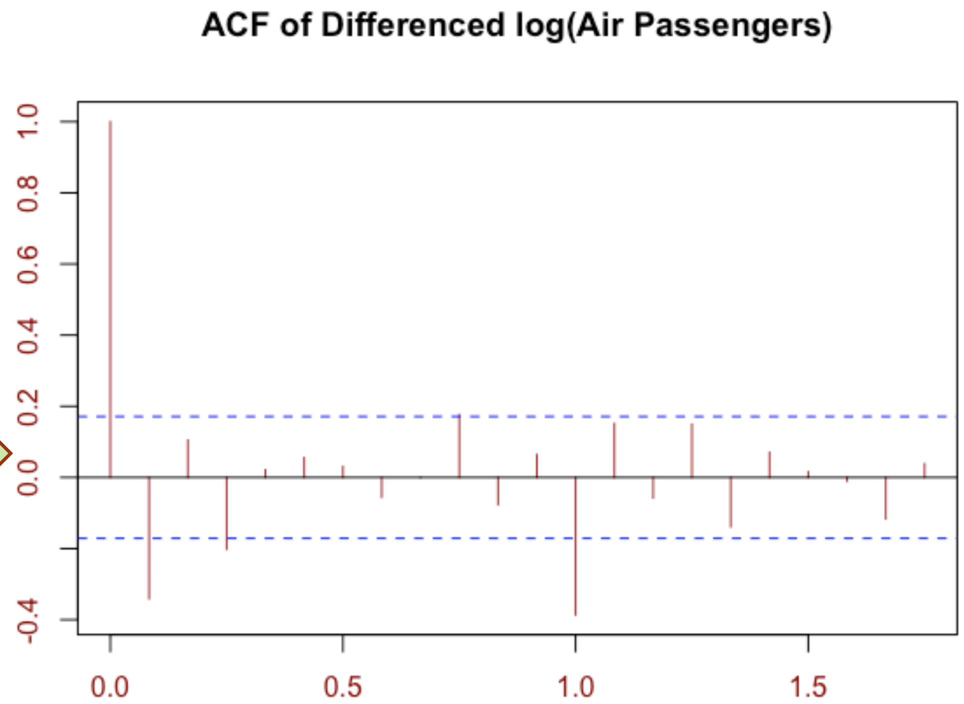
Autocorrelation is a correlation of a series with a delayed copy of itself.

It is similarities between observations as a function of time lag between them



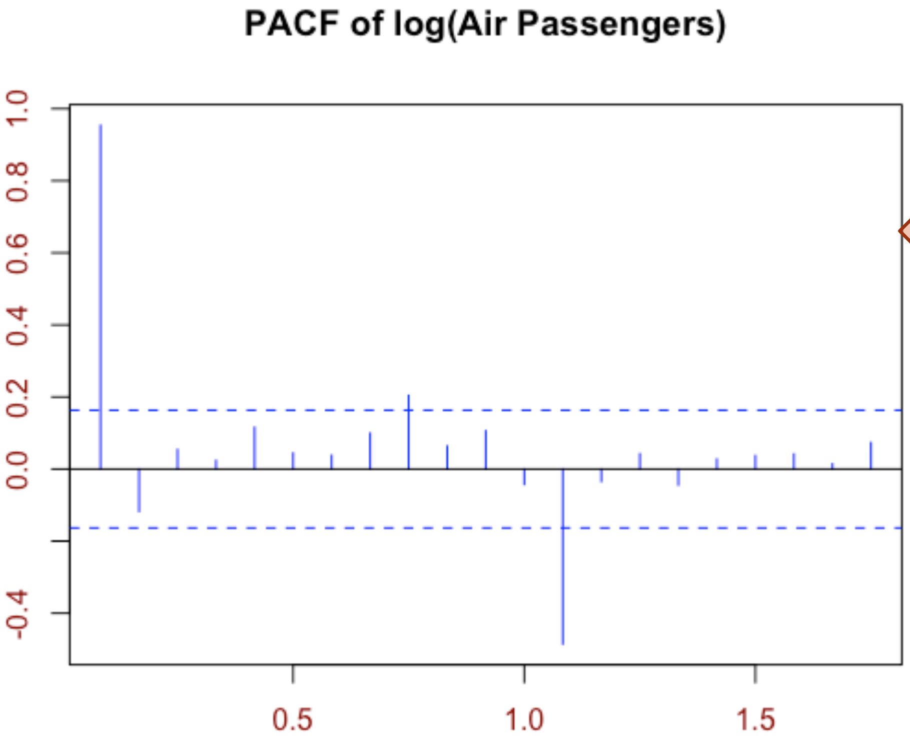
Gradual Decay
Non- Stationary

Typical of “almost” Stationary series



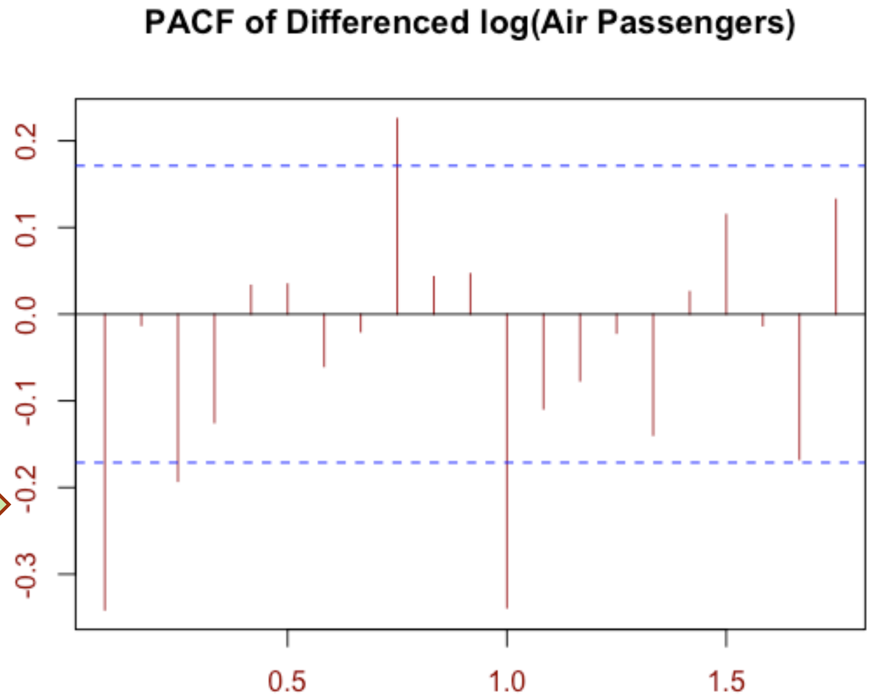
Partial Auto Correlation Function (PACF)

Partial correlation of a series with a delayed copy of itself removing effects of intermediate time lags



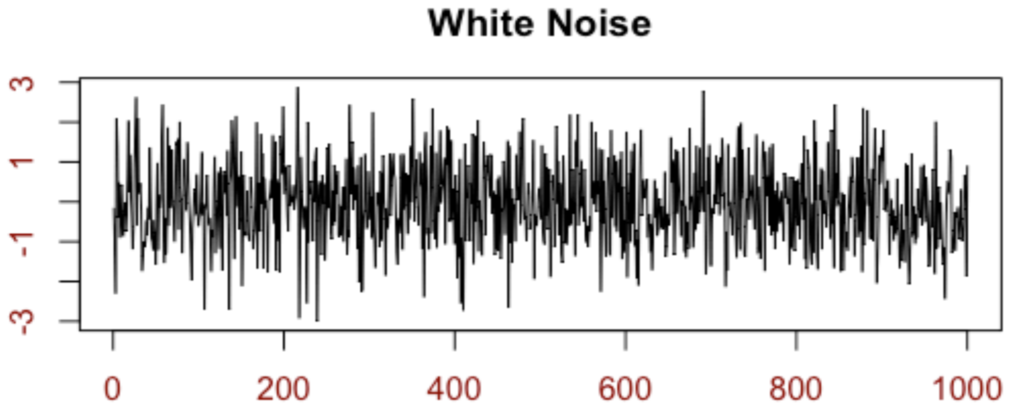
Sudden Cut off after 1 Lag

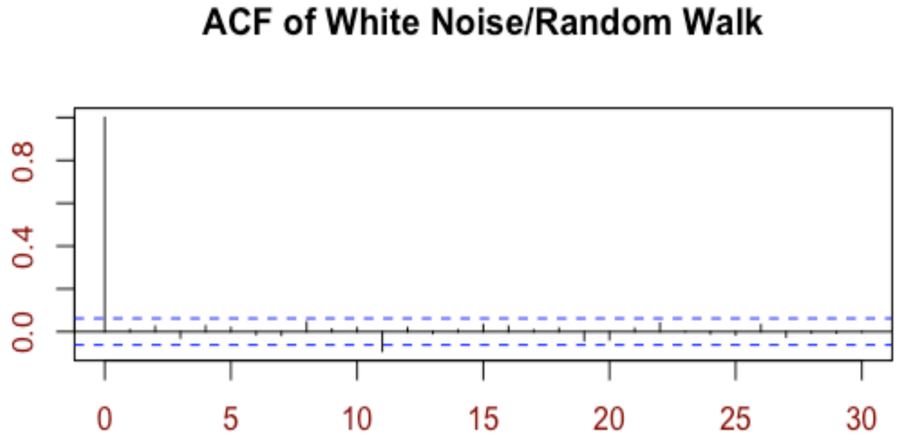
Sinusoidal Pattern

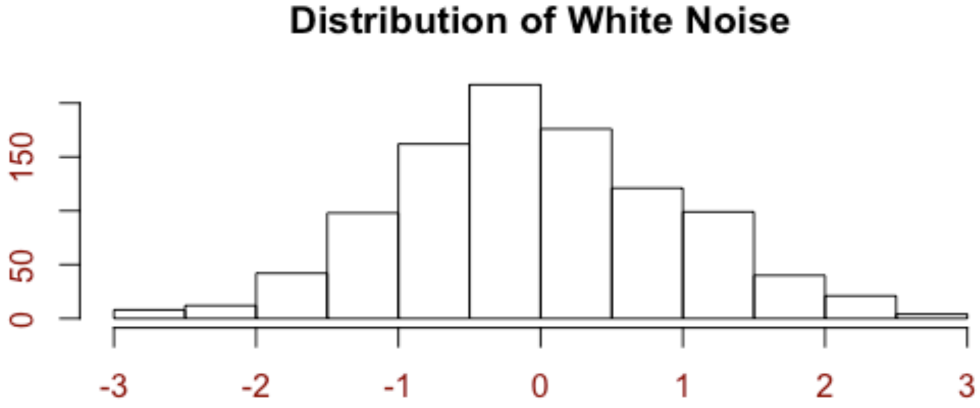


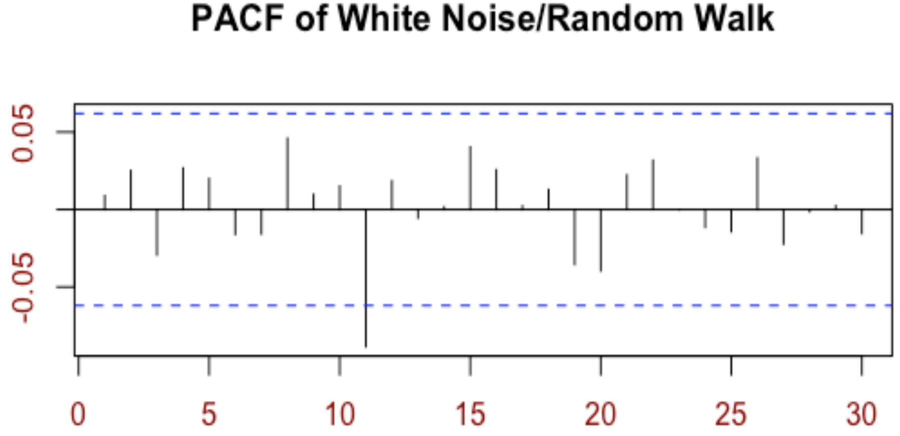
DSLA
DATA SCIENCE LEARNING ACADEMY

ACF & PACF for Stationary Series









Statistical tests for Stationarity

Ljung–Box Test
Box.test()

Examines whether there is significant evidence for non-zero correlations at lags 1-20.
Small p-values (i.e., less than 0.05) suggest that the series is stationary

Augmented Dickey-Fuller Test
adf.test()

Tests for presence of Unit Root using an Auto-Regressive Model to optimize information criteria across multiple lags
Small p-values suggest the data is stationary and doesn't need to be differenced for stationarity.

Kwiatkowski-Phillips-Schmidt-Shin Test
kpss.test()

Accepting the null hypothesis means that the series is stationarity.
Small p-values suggest that the series is not stationary and a differencing is required.

Auto Regressive Models (AR)

Once a Series is Stationary, we can use models to predict behavior

An autoregressive (AR) model **predicts future behavior based on past behavior.**

The process is basically a linear regression of the data in the current series against one or more past values in the same series.

AR(1) Model → $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \varepsilon_t$

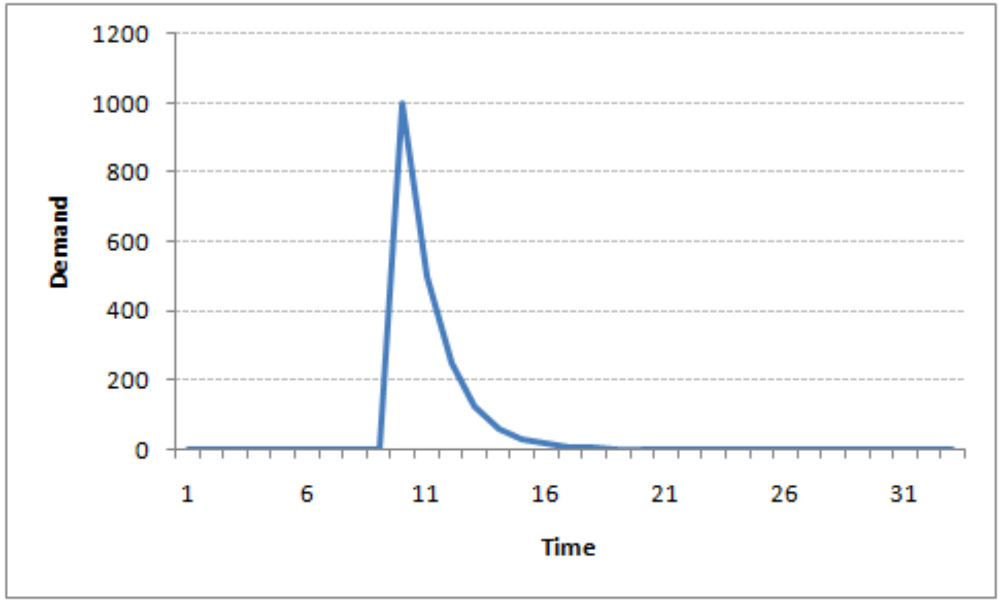
AR(2) Model → $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \varepsilon_t$

AR(p) Model → $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ \ldots + \alpha_p Y_{t-p} + \varepsilon_t$

Auto Regressive - Example

New Cell Phone Demand on Launch

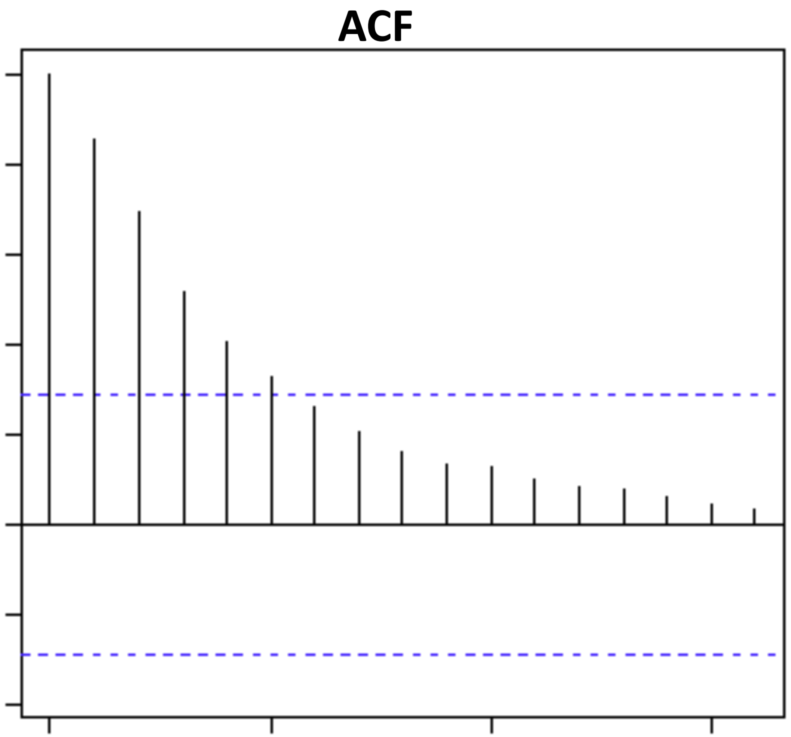
Gradual Decay of Demand post Launch day



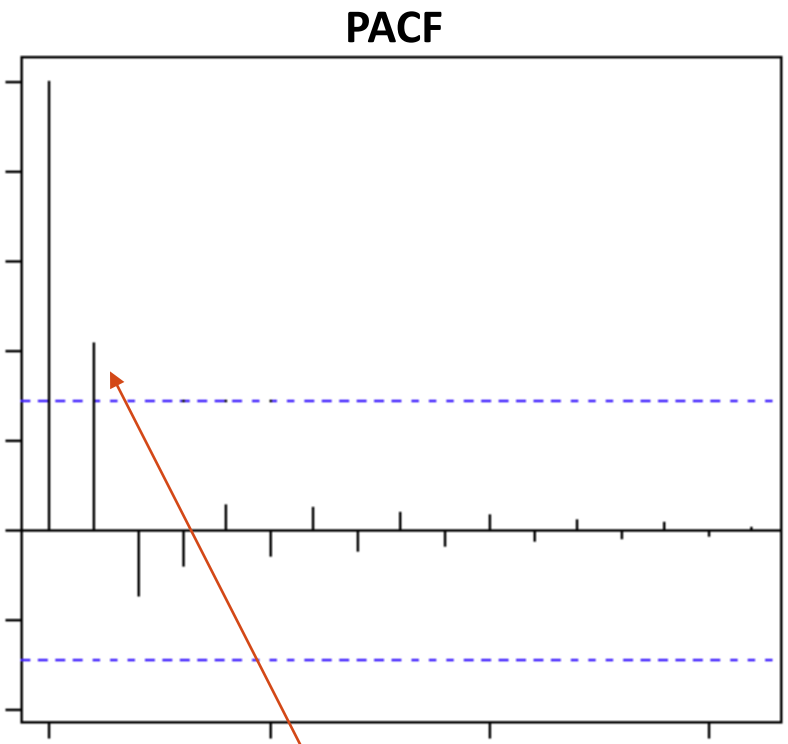
Auto Regressive Model can be guessed by plotting ACF and PACF

ACF Curve	*PACF Curve*	*Model*
Exponential or Oscillating Decay	Cut off at lag *'p'*	*AR(p) Model*

Auto Regressive Model - Selection



Exponential Decay



Cut off after lag 2, indicating AR(2) Model

Moving Average Models (MA)

Once a Series is Stationary, we can use models to predict behavior

An moving average (MA) model **predicts future behavior based on past errors**

The process is basically a linear regression of the data in the current series against past forecast errors

MA(1) Model → $Y_t = c + \theta_1 \varepsilon_{t-1} + \varepsilon_t$

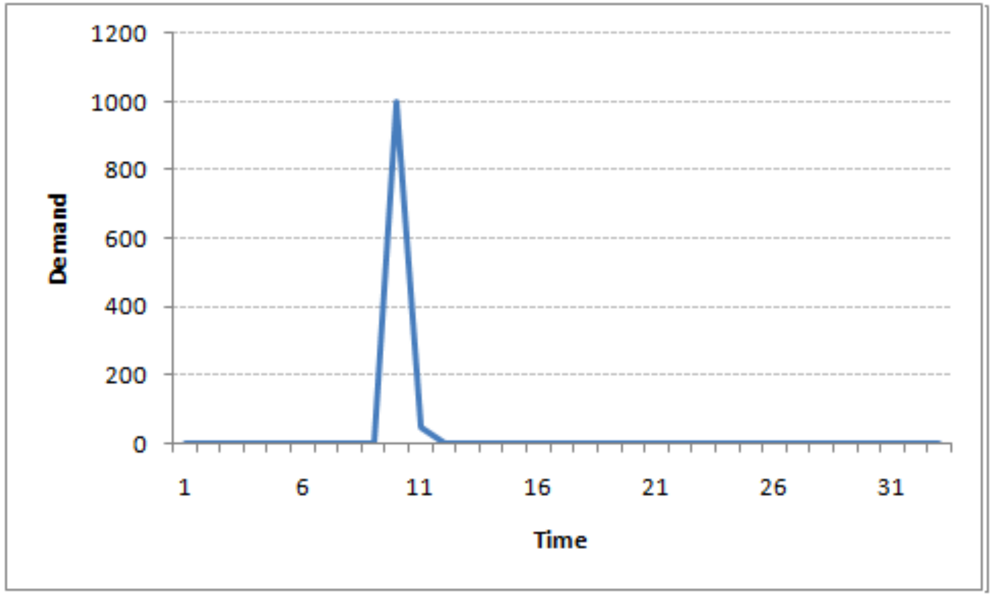
MA(2) Model → $Y_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$

MA(q) Model → $Y_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$

Moving Average - Example

Spike in Demand due to sudden drop in price erroneously

Shock is immediately decayed after price corrected

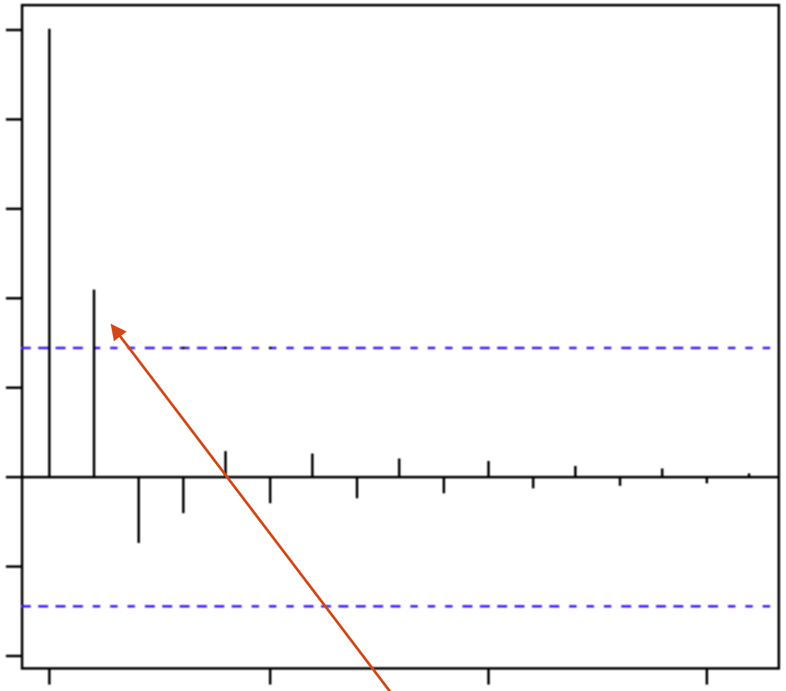


Moving Average Model can be guessed by plotting ACF and PACF

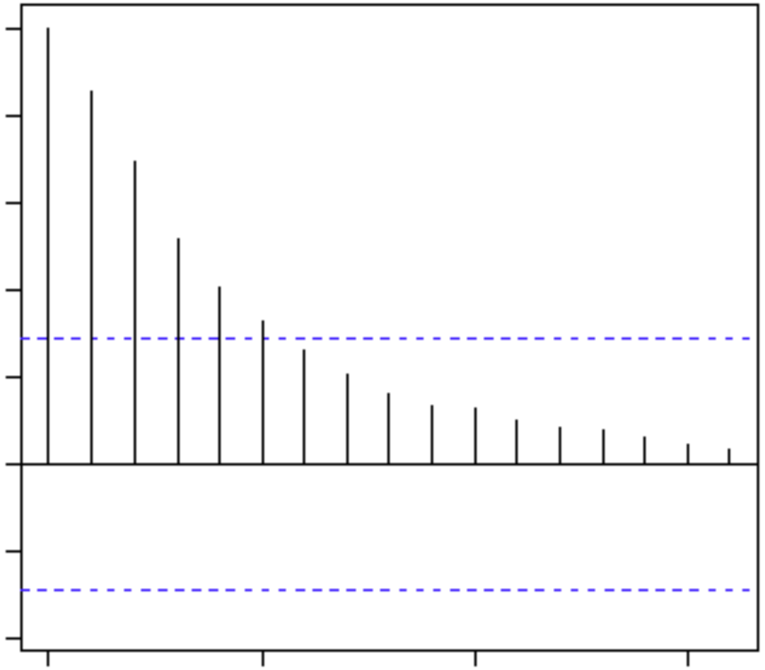
ACF Curve	*PACF Curve*	*Model*
Cut off at lag *'q'*	Exponential or Oscillating Decay	*MA(q) Model*

Moving Average Model - Selection

ACF



PACF



Cut off after lag 2, indicating MA(2) Model

Exponential Decay

ARIMA Models

Auto **R**egressive **I**ntegrated **M**oving **A**verage Model

The model is basically a filter which applies AR models, Differencing (I) models and MA Models to separate signal from noise. Signals are then extrapolated into future to make predictions

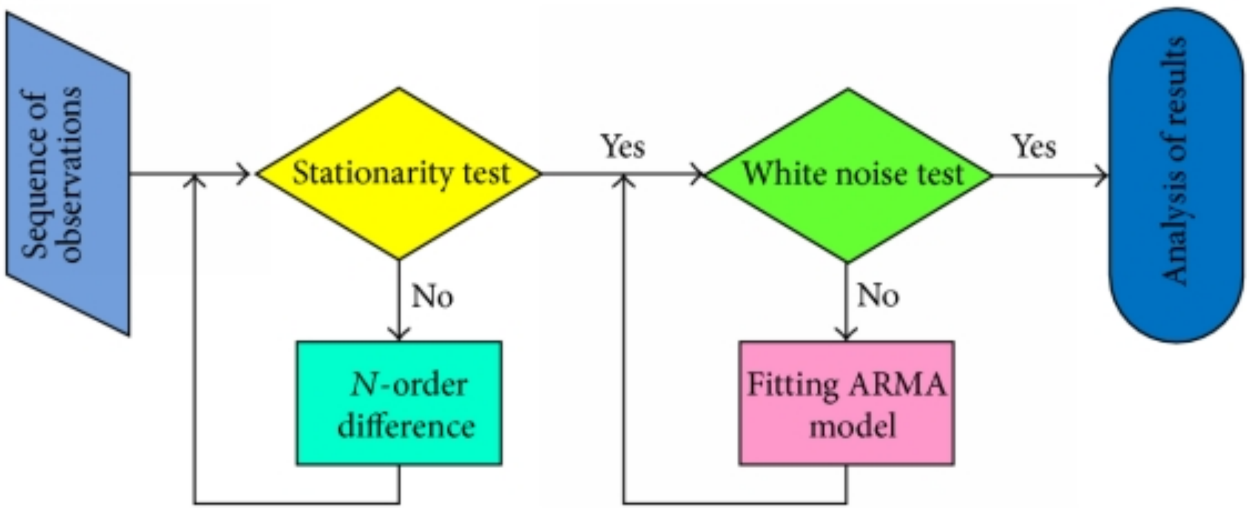
Order of Differencing

ARIMA(p, d, q)

Auto Regressive Terms

Moving Average Terms

ARIMA Model Selection



$$\text{ARIMA} \quad \underbrace{(p, d, q)}_{\substack{\uparrow \\ \left(\begin{array}{l}\text{Non-seasonal part} \\ \text{of the model}\end{array}\right)}} \quad \underbrace{(P, D, Q)_m}_{\substack{\uparrow \\ \left(\begin{array}{l}\text{Seasonal part} \\ \text{of the model}\end{array}\right)}}$$

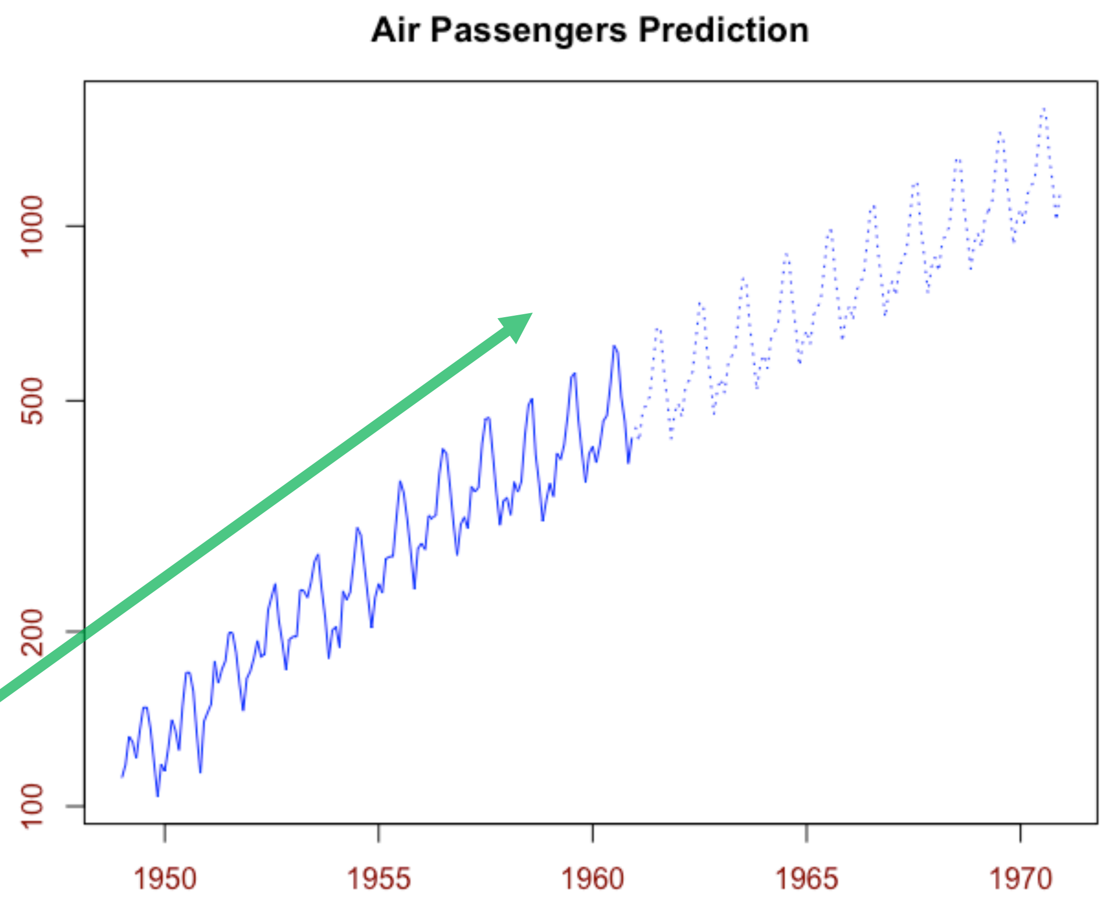
ARIMA Model Prediction

Log Transform the Air Passengers Series

Fit an ARIMA(1,1,0)$(1,1,0)_{12}$ Seasonal Model

Predict using the Model for 10 more years

Apply Exponential Transformation to Predictions



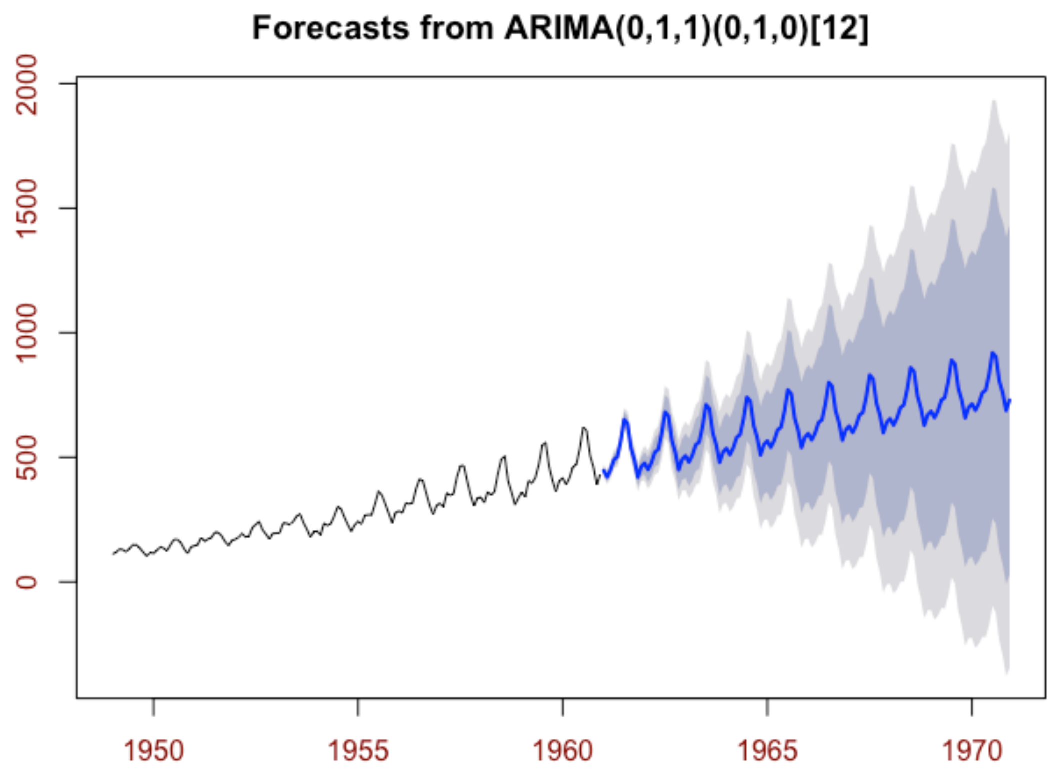
ARIMA Model Prediction

Use *auto.arima()* from *forecast* package to automatically select the best model

Fit the **best arima model** to *Air Passengers*

Predict using the Model using *forecast()* for 10 more years

Plot using *plot()* to get confidence intervals



Summary

1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

Time Series Analysis

Break Time

Demo

Forecasting Air Passengers