PHÂN TÍCH DỮ LIỆU KINH DOANH

LAB04. PHÂN TÍCH DỮ LIỆU CHUỖI THỜI GIAN

(Time Series Analysis)

CÔNG CỤ: R, PYTHON, EXCEL

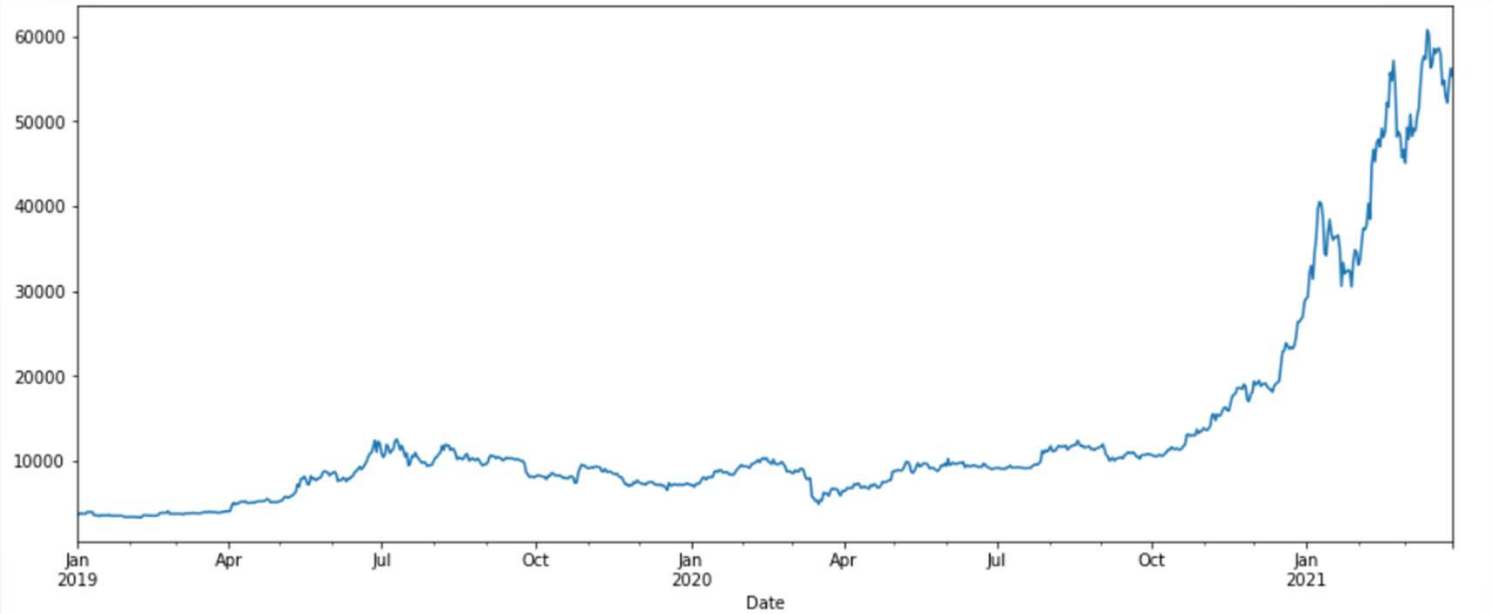
Trình bày: Nguyễn Minh Nhựt

SĐT: 0939013911 - 09851734105

4.1. Chuỗi thời gian là gì?

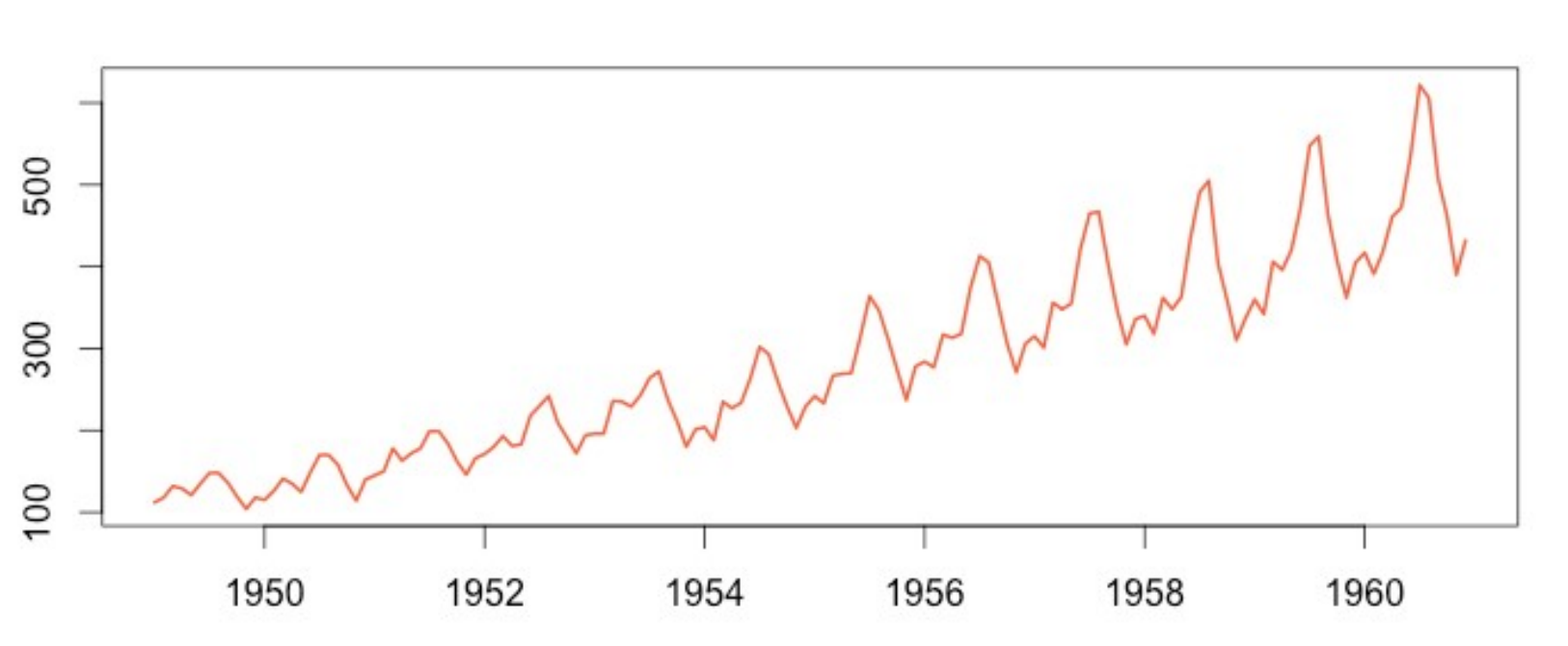
- ***Định nghĩa***

Time Series (Chuỗi thời gian) được xem là bao gồm các **phép đo một giá trị** theo **thời gian** ví dụ như giá BTC từ ngày 6/1/2019 – 6/1/2021 trong một khoảng thời gian. Phân tích chuỗi thời gian có mục đích **nhận dạng** và **tập hợp lại các yếu tố**, ảnh hưởng mà thời gian có ảnh hưởng đến **giá trị mà quan sát**.



4.1. Chuỗi thời gian là gì?

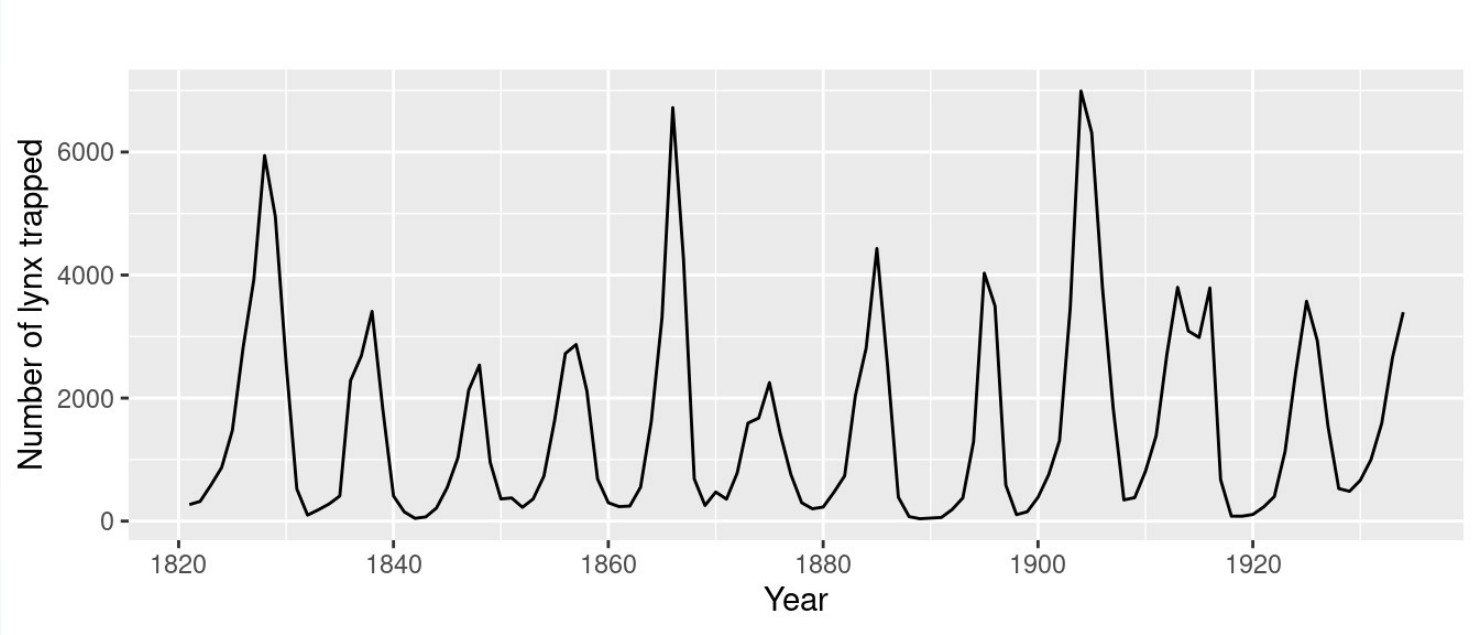
- ***Các thành phần của chuỗi thời gian***
  - **Chuỗi thời gian dạng xu hướng/xu thế (Trend)**



Dữ liệu có giá trị **theo xung hướng tăng/giảm** theo thời gian.

4.1. Chuỗi thời gian là gì?

- ***Các thành phần của chuỗi thời gian***
  - **Chuỗi thời gian dạng mùa vụ (Seasonal)**

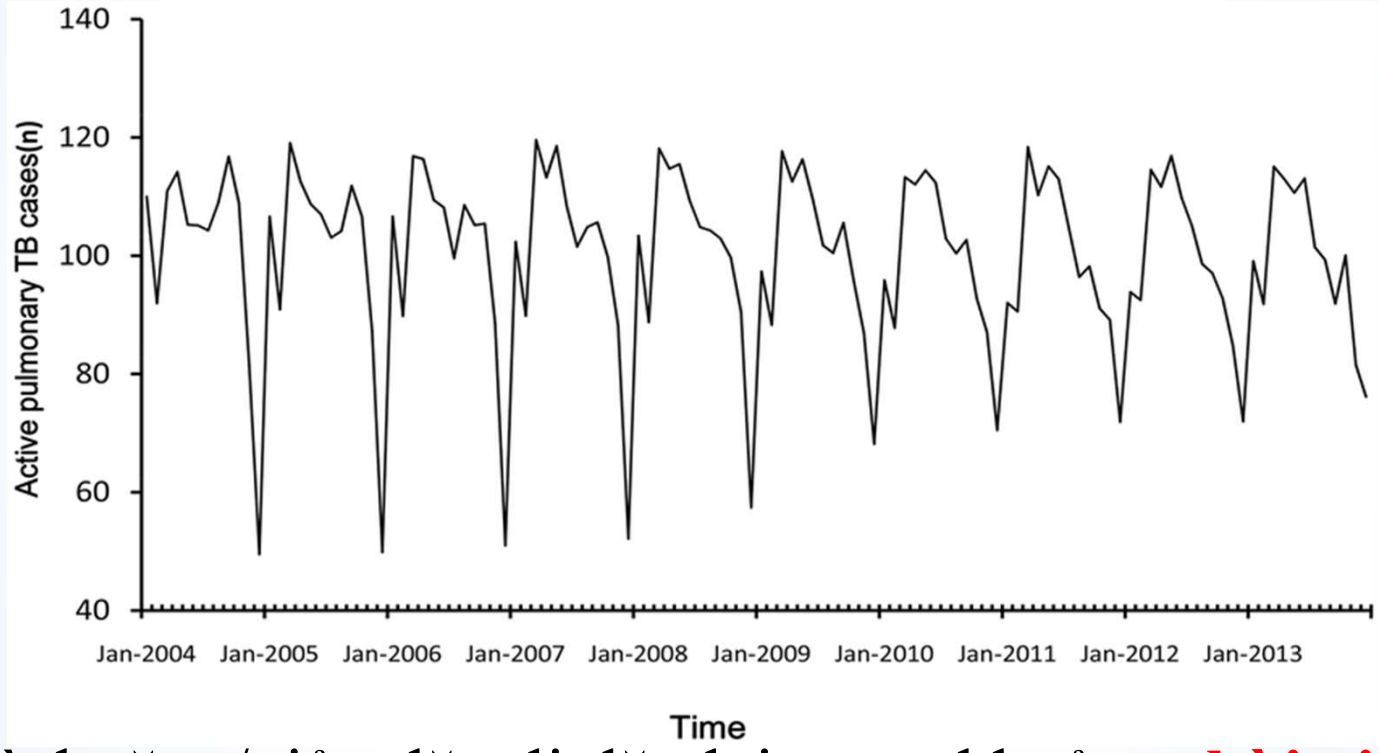


Quá trình tăng/giảm lặp đi, lặp lại trong khoảng **thời gian ngắn.**

Theo tháng, quí, mùa

4.1. Chuỗi thời gian là gì?

- ***Các thành phần của chuỗi thời gian***
  - **Chuỗi thời gian dạng chu kỳ (Cylic)**

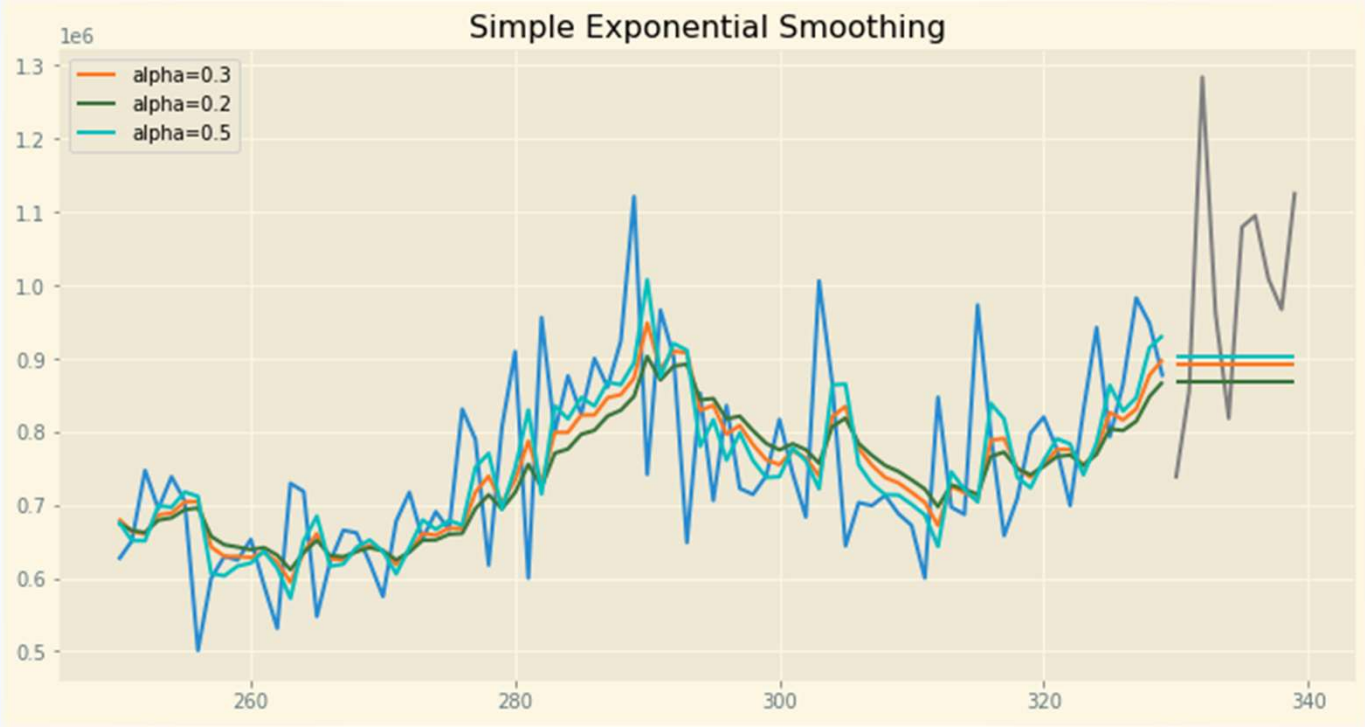


Quá trình tăng/giảm lặp đi, lặp lại trong khoảng **thời gian dài hơn.**

Theo Năm

4.1. Chuỗi thời gian là gì?

- ***Các thành phần của chuỗi thời gian***
  - **Chuỗi thời gian dạng ngẫu nhiên**



Dữ liệu thể hiện **KHÔNG** theo quy luật nào cả tăng ngẫu nhiên

biến động.

4.2 Xử lý dữ liệu ngày tháng trong R và Python

- **Trong ngôn ngữ R**
  - Tạo một ngày trong ngôn ngữ R

```
> date<-as.Date('05/08/2022',format='%m/%d/%Y')
> date
[1] "2022-05-08"
>
```

  - Ví dụ khác

```
> date2<-as.Date('05~08/2022',format='%m~%d/%Y')
> date2
[1] "2022-05-08"
>
```

  - Có nhiều cách format ngày giờ khác nhau trong ngôn ngữ R. Link tham khảo: https://www.stat.berkeley.edu/~s133/dates.html

4.2 Xử lý dữ liệu ngày tháng trong R và Python

- **Trong ngôn ngữ R**
  - Lấy ngày/tháng/năm trong một đối tượng ngày đã tạo sẵn

```
> format(date,format="%d")
[1] "08"
>
```

Lấy ngày trong R

  - Cho tập dữ liệu dưới đây liệt kê ra số ngày có trong tập dữ liệu, số tháng có trong tập dữ liệu, số năm kết quả trả về 1 mảng. https://drive.google.com/file/d/1CLJ1KvJZoL2Juqj5ii4tlYHbZzVnPEFy/view?usp=sharing
  - Ví dụ **[1] 01 02 03 05**

```
 [1] "01" "02" "03" "04" "05" "06" "07" "16" "17" "18" "19" "20" "21"
[14] "22" "23" "24" "25" "26" "27" "28" "29" "30" "31" "01"
>
```

4.2 Xử lý dữ liệu ngày tháng trong R và Python

- **Trong ngôn ngữ Python**

```
In [1]: from datetime import datetime

In [4]: year = 2022
        month = 5
        day = 9
        hour = 12
        minute = 8
        second = 59

In [5]: my_date = datetime(year,month,day,hour,minute,second)
        my_date
Out[5]: datetime.datetime(2022, 5, 9, 12, 8, 59)
```

Tạo đối tượng datetime bằng thư viện **`datetime`**

Thực hiện lại ví dụ giống như trên ngôn ngữ R?

Lấy tháng, năm trong `my_date`

```
In [6]: my_date.month
Out[6]: 5

In [8]: my_date.day
Out[8]: 9

In [9]: my_date.year
Out[9]: 2022
```

```
In [17]: import numpy as np

In [23]: type(df.date)
Out[23]: pandas.core.series.Series

In [35]: df['datetime_new'] = pd.to_datetime(df.date)

In [42]: type(df['datetime_new'])
Out[42]: pandas.core.series.Series

In [47]: df['datetime_year']=pd.DatetimeIndex(df['datetime_new']).year

In [49]: df['datetime_day']=pd.DatetimeIndex(df['datetime_new']).day

In [50]: df
```

4.2 Xử lý dữ liệu ngày tháng trong R và Python

- **Trong ngôn ngữ Python**

Đánh index trong Python

```
In [51]: df.index = df['datetime_new']

In [52]: df
```

Out[52]:

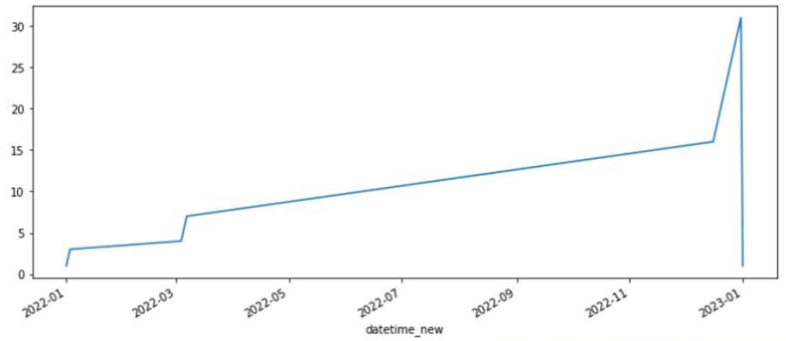
datetime_new	date	datetime_new	datetime_year	datetime_day
2022-01-01	1/1/2022	2022-01-01	2022	1
2022-01-02	1/2/2022	2022-01-02	2022	2
2022-01-03	1/3/2022	2022-01-03	2022	3
2022-03-04	3/4/2022	2022-03-04	2022	4
2022-03-05	3/5/2022	2022-03-05	2022	5
2022-03-06	3/6/2022	2022-03-06	2022	6
2022-03-07	3/7/2022	2022-03-07	2022	7
2022-12-16	12/16/2022	2022-12-16	2022	16
2022-12-17	12/17/2022	2022-12-17	2022	17
2022-12-18	12/18/2022	2022-12-18	2022	18
2022-12-19	12/19/2022	2022-12-19	2022	19
2022-12-20	12/20/2022	2022-12-20	2022	20
2022-12-21	12/21/2022	2022-12-21	2022	21
2022-12-22	12/22/2022	2022-12-22	2022	22
2022-12-23	12/23/2022	2022-12-23	2022	23
2022-12-24	12/24/2022	2022-12-24	2022	24
2022-12-25	12/25/2022	2022-12-25	2022	25
2022-12-26	12/26/2022	2022-12-26	2022	26
2022-12-27	12/27/2022	2022-12-27	2022	27
2022-12-28	12/28/2022	2022-12-28	2022	28
2022-12-29	12/29/2022	2022-12-29	2022	29
2022-12-30	12/30/2022	2022-12-30	2022	30
2022-12-31	12/31/2022	2022-12-31	2022	31

Mục tiêu của **đánh index** phục vụ cho việc xây dựng **mô hình chuỗi thời gian**

Vẽ hình chuỗi thời gian

```
In [53]: df['datetime_day'].plot(figsize=(12,5))
```

Out[53]: <AxesSubplot:xlabel='datetime_new'>



Code mẫu

```
df['datetime_day'].plot(figsize=(12
,5))
```

4.2 Xử lý dữ liệu ngày tháng trong R và Python

- **Trong ngôn ngữ Python**

Ví dụ bài tập

Lấy dữ liệu Bitcoin từ trang

https://finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD

Từ ngày 09/05/2017 – 09/05/2022

1. Đưa cột [Date] trở thành index của tập dữ liệu
2. Vẽ hình mối liên hệ giữa [Date] và giá đóng cửa [Close]
3. Vẽ hình mối liên hệ giữa [Date] và giá mở cửa [Open]
4. Vẽ hình mối liên hệ giữa [Date] và lượng giao dịch [volume]
5. Vẽ hình mối liên hệ giữa [Date] và giá đóng cửa [Close], [Date] và giá mở cửa [Open] trên cùng một biểu đồ. Chú thích biểu đồ.
6. Nhận xét các biểu đồ trên.

4.3 Mô hình ARIMA

- ***Định nghĩa mô hình ARIMA***

Mô hình ARIMA là viết tắt của **quá trình tự hồi quy** (Auto Regression -AR), **quá trình trung bình trượt** (Moving Average – MA) và **tích hợp sai phân Integrated** - I

Điểm quan trọng: Mô hình ARIMA không phải là mô hình dự báo hoàn hảo ứng với bất kỳ dữ liệu chuỗi thời gian nào nào.

Mô hình ARIMA chỉ hoạt động tốt nhất nếu dữ liệu phụ thuộc nhiều vào thời gian. Những dữ liệu dạng ngẫu nhiên thường ít hoạt động đối với mô hình ARIMA.

Mô hình ARIMA chỉ dự báo tốt tại dự báo dạng điểm thời gian.

4.3 Mô hình ARIMA

- ***Các loại mô hình ARIMA***

  Mô hình ARIMA không có tính mùa vụ

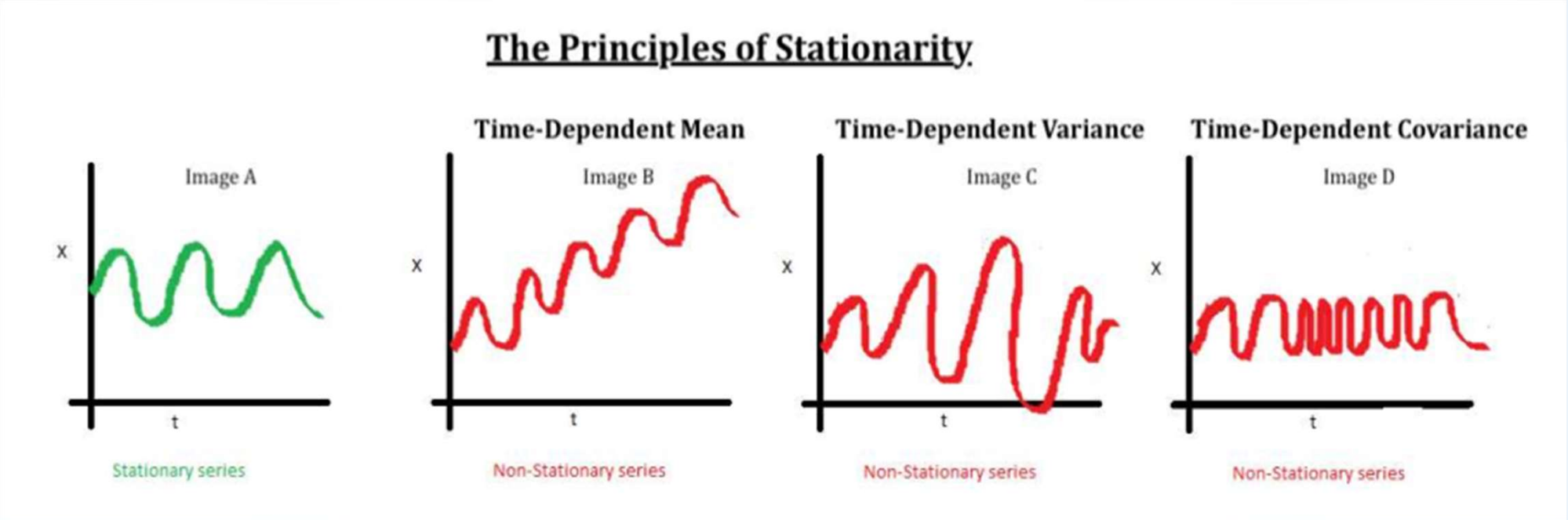
  Mô hình ARIMA có tính mùa vụ (Seasonal ARIMA – SARIMA)

- ***Chuỗi dừng***

Một chuỗi thời gian có **tính dừng** là một chuỗi các giá trị **mean, variance, autocorrelation** không thay đổi theo thời gian và nó không bao hàm yếu tố xu thế.Với hầu hết các phương pháp thống kê dự báo, đều phải đảm bảo tính dừng của **chuỗi dữ liệu** vì thế việc kiểm tra tính dừng là **rất quan trọng.**

4.3 Mô hình ARIMA

- ***Chuỗi dừng***



Để kiểm định tính dừng của dữ liệu ta có hai phương pháp kiểm định phổ biến: Kiểm định **Dickey Fuller3 (DF)** và **Dickey Fuller cải tiến (ADF4)**

4.3 Mô hình ARIMA

- ***Chuỗi dừng***

Sinh viên hoạt động **30 phút** thảo luận nhóm và trình phương pháp **ADF và DF** thực hiện lấy dữ liệu một chuỗi dữ liệu chuỗi thời gian thử để kiểm tra tính dừng của chuỗi đó?

Link tài liệu về kiểm định ADF, DF

1. https://www.statisticshowto.com/adf-augmented-dickey-fuller-test/
2. https://www.youtube.com/watch?v=T5BhGv742j4
3. https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

4.3 Mô hình ARIMA

- ***Mô hình ARIMA không có tính mùa vụ***

ARIMA (p,d,q)

Với p, d, q lần lượt là các **số không âm**

I(d): Integrated – So sánh sự khác nhau giữa d quan sát (Hiệu giữa giá trị hiện tại và **d** giá trị trước đó

Sai phân lần 1 I(1): $\Delta y_t = y_t - y_{t-1}$

Sai phân lần 2 I(2): $\Delta(\Delta y_t) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$

Sai phân lần d được ký hiệu là I(d)

Tài liệu tham khảo tính bằng excel:
https://www.youtube.com/watch?v=1DaTkEd_uHY

4.3 Mô hình ARIMA

- ***Mô hình ARIMA không có tính mùa vụ***

ARIMA (p,d,q)

Với p, d, q lần lượt là các **số không âm**

I(d): Integrated – So sánh sự khác nhau giữa d quan sát (Hiệu giữa giá trị hiện tại và **d** giá trị trước đó

Original Data

Time1	10
Time2	12
Time3	8
Time4	14
Time5	7

First Difference

Time1	NA
Time2	2
Time3	-4
Time4	6
Time5	-7

Second Difference

Time1	NA
Time2	NA
Time3	-6
Time4	10
Time5	-13

4.3 Mô hình ARIMA

- ***Mô hình ARIMA không có tính mùa vụ***

ARIMA (p,d,q)

Với p, d, q lần lượt là các **số không âm**

AR(p): Autoregression – là quá trình tìm mối quan hệ giữa dữ liệu hiện tại và **p** dữ liệu quá khứ trước đó. (Gọi là lag)

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + \varepsilon_t$$

Tài liệu tham khảo tính bằng excel:
https://www.youtube.com/watch?v=1DaTkEd_uHY

Điều kiện dừng của việc chọn ***p*** $\sum_{i=0}^{p} a_i < 1$

4.3 Mô hình ARIMA

- ***Mô hình ARIMA không có tính mùa vụ***

ARIMA (p,d,q)

Với p, d, q lần lượt là các **số không âm**

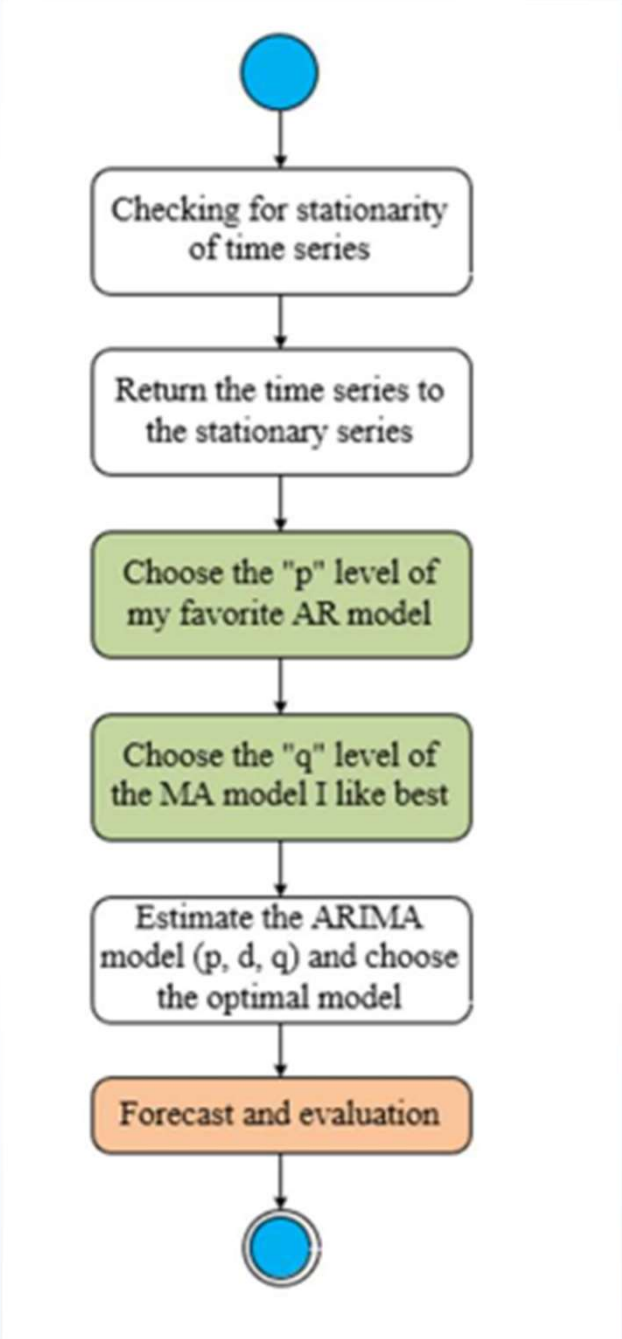
MA(q): Moving Average – là quá trình tìm mối quan hệ giữa dữ liệu hiện tại và **q** phần lỗi quá khứ trước đó.

$$y_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q} + \mu_t$$

Tài liệu tham khảo tính bằng excel:
https://www.youtube.com/watch?v=9DH9IhkT2wo

Điều kiện dừng của việc chọn ***q*** $\sum_{i=0}^{q} \beta_i < 1$

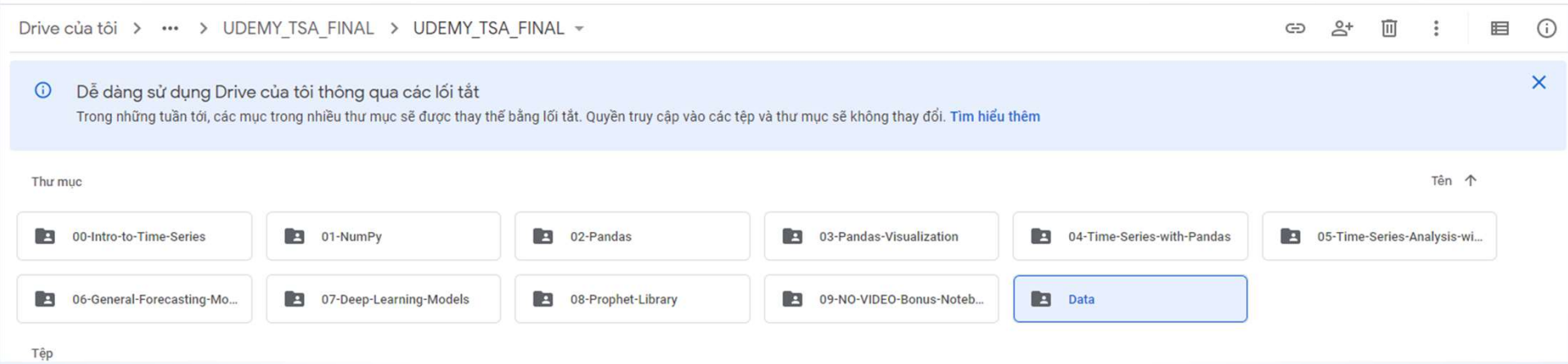
4.4 Mô hình Box-Jenkins



4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

Tập dữ liệu thực hiện bài tập trên ngôn ngữ R



Tập dữ liệu được lấy trong thư mục Data

Các tập dữ liệu dành cho thực hành

- `DailyTotalFemaleBirths.csv`
- `TradeInventories.csv`

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

```
Error in choose.file() : could not find function "choose.file"
> df<-read.csv(choose.files())
> df
          Date Births
1     1/1/1959     35
2     1/2/1959     32
3     1/3/1959     30
4     1/4/1959     31
5     1/5/1959     44
6     1/6/1959     29
7     1/7/1959     45
8     1/8/1959     43
9     1/9/1959     38
10   1/10/1959     27
11   1/11/1959     38
12   1/12/1959     33
13   1/13/1959     55
14   1/14/1959     47
15   1/15/1959     45
```

Đánh index trong R

```
> rownames(df)<-df$Date
> df
                Date Births
1/1/1959    1/1/1959     35
1/2/1959    1/2/1959     32
1/3/1959    1/3/1959     30
1/4/1959    1/4/1959     31
1/5/1959    1/5/1959     44
1/6/1959    1/6/1959     29
1/7/1959    1/7/1959     45
1/8/1959    1/8/1959     43
1/9/1959    1/9/1959     38
1/10/1959  1/10/1959     27
1/11/1959  1/11/1959     38
1/12/1959  1/12/1959     33
1/13/1959  1/13/1959     55
1/14/1959  1/14/1959     47
1/15/1959  1/15/1959     45
1/16/1959  1/16/1959     37
1/17/1959  1/17/1959     50
```

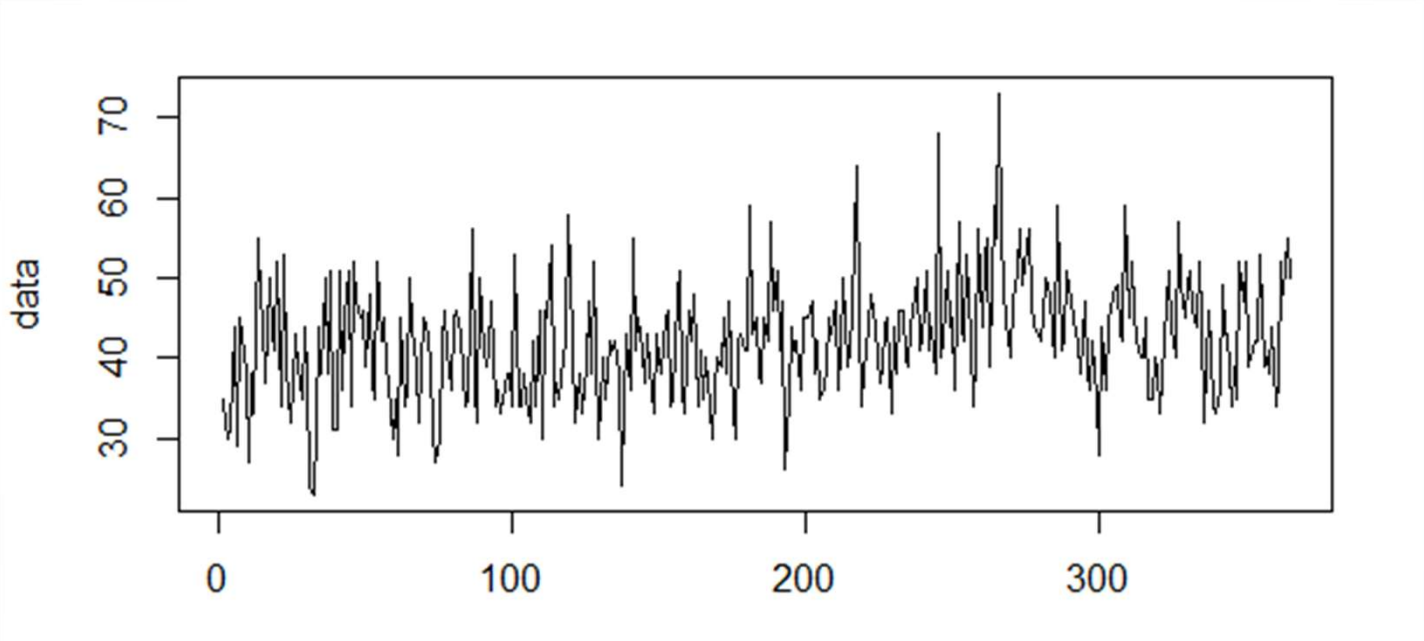
- Bước 1: Import file csv và đưa dữ liệu từ kiểu choỗi thành kiểu datetime.
- Bước 2: Đưa về ngày về dạng Date nếu chưa đưa về dạng Date.

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 3: Vẽ biểu đồ theo chuỗi thời gian

```
> data<-ts(df$Births,frequency = 1,start(1959,1,1))
> plot(data
+ )
> plot(data)
```



Nhìn vào biểu đồ ta thấy chuỗi thời gian dừng nhưng cần kiểm tra

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 4. Kiểm định ADF cho chuỗi thời gian. Đầu tiên thêm thư viện vào để kiểm định: `library(tseries)`

```
> adf.test(data)

        Augmented Dickey-Fuller Test

data:  data
Dickey-Fuller = -5.1042, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Dựa vào kết quả của kiểm định ADF ta có p-value = 0.01 < 0.05 nên chuỗi thời gian là **chuỗi dừng**

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 5. Để tiến hành dự báo và chọn mô hình ARIMA phù hợp: `library(forecast)`

```
Fitting models using approximations to speed things up...

ARIMA(2,1,2) with drift         : 2469.859
ARIMA(0,1,0) with drift         : 2646.36
ARIMA(1,1,0) with drift         : 2561.517
ARIMA(0,1,1) with drift         : 2459.039
ARIMA(0,1,0)                    : 2644.345
ARIMA(1,1,1) with drift         : 2462.563
ARIMA(0,1,2) with drift         : 2456.728
ARIMA(1,1,2) with drift         : 2464.245
ARIMA(0,1,3) with drift         : 2457.653
ARIMA(1,1,3) with drift         : Inf
ARIMA(0,1,2)                    : 2455.685
ARIMA(0,1,1)                    : 2457.773
ARIMA(1,1,2)                    : 2462.991
ARIMA(0,1,3)                    : 2456.743
ARIMA(1,1,1)                    : 2461.332
ARIMA(1,1,3)                    : Inf

Now re-fitting the best model(s) without approximations...

ARIMA(0,1,2)                    : 2459.637

Best model: ARIMA(0,1,2)
```

Kết quả mô hình tốt nhất ARIMA(0,1,2)

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 6. Chia dữ liệu theo train và test

```
> train<-head(df,90)
> train
                Date Births DateTimeNew
1/1/1959    1/1/1959     35    1/1/1959
1/2/1959    1/2/1959     32    1/2/1959
1/3/1959    1/3/1959     30    1/3/1959
1/4/1959    1/4/1959     31    1/4/1959
1/5/1959    1/5/1959     44    1/5/1959
1/6/1959    1/6/1959     29    1/6/1959
1/7/1959    1/7/1959     45    1/7/1959
1/8/1959    1/8/1959     43    1/8/1959
1/9/1959    1/9/1959     38    1/9/1959
1/10/1959  1/10/1959     27   1/10/1959
1/11/1959  1/11/1959     38   1/11/1959
1/12/1959  1/12/1959     33   1/12/1959
1/13/1959  1/13/1959     55   1/13/1959
1/14/1959  1/14/1959     47   1/14/1959
1/15/1959  1/15/1959     45   1/15/1959
1/16/1959  1/16/1959     37   1/16/1959
```

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 7. Train dữ liệu với mô hình ARIMA đã fit, Forecast test

```
> fitARIMAtrain<-arima(train['Births'],order=c(0,1,2))
> predict<-forecast(fitARIMAtrain,90)
> predict
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
276       48.50121 39.25964 57.74278 34.36745 62.63497
277       48.10055 38.79544 57.40566 33.86961 62.33149
278       48.10055 38.78274 57.41836 33.85019 62.35091
279       48.10055 38.77006 57.43104 33.83080 62.37030
280       48.10055 38.75740 57.44370 33.81144 62.38966
281       48.10055 38.74476 57.45634 33.79210 62.40900
282       48.10055 38.73213 57.46897 33.77279 62.42831
283       48.10055 38.71952 57.48158 33.75350 62.44760
```

Đưa predict về dạng dataframe
Tạo thêm cột predict trong dataframe
predict đưa về dạng số

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 7. Tính RMSE

Để tiến hành dự báo và chọn mô hình ARIMA phù hợp:

```
library(Metrics)
```

```
> rmse(test$Births,predict$predict)
[1] 7.734749
>
```

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 6. Chia dữ liệu theo train và test

```
> test<-tail(df,nrow(df)-90)
> test
                Date Births DateTimeNew
4/1/1959   4/1/1959     39    4/1/1959
4/2/1959   4/2/1959     41    4/2/1959
4/3/1959   4/3/1959     47    4/3/1959
4/4/1959   4/4/1959     34    4/4/1959
4/5/1959   4/5/1959     36    4/5/1959
4/6/1959   4/6/1959     33    4/6/1959
4/7/1959   4/7/1959     35    4/7/1959
4/8/1959   4/8/1959     38    4/8/1959
4/9/1959   4/9/1959     38    4/9/1959
```

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ R***

- Bước 8. Dự báo

```
> forecast(fitARIMA,10)
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
366       44.32892 35.31350 53.34434 30.54103 58.11682
367       43.76542 34.64618 52.88465 29.81874 57.71209
368       43.76542 34.63746 52.89338 29.80541 57.72542
369       43.76542 34.62874 52.90209 29.79208 57.73875
370       43.76542 34.62004 52.91079 29.77877 57.75206
371       43.76542 34.61134 52.91949 29.76547 57.76536
372       43.76542 34.60265 52.92818 29.75218 57.77865
373       43.76542 34.59397 52.93686 29.73891 57.79192
374       43.76542 34.58530 52.94553 29.72565 57.80519
375       43.76542 34.57664 52.95419 29.71240 57.81843
> plot(forecast(fitARIMA,10))
>
```

4.5 Thực hành dự báo chuỗi thời gian

- ***Phân tích mô hình ARIMA trên ngôn ngữ Python***

  - Xem trong file mẫu hướng dẫn

4.6 Link tài liệu hướng dẫn thực hành

https://drive.google.com/drive/folders/1N7gJ3vlKVP2fLoJJa3K89ofnbqNoJryy?usp=sharing