# Xử lí dữ liệu

### Tổng quan về giá của Bitcoin từ 17-9-2024 đến 8-12-2022

## 

Nhìn chung giá Bitcoin biết động rất mạnh , có sự tăng trưởng vược bậc trong giao đoạn 2017 và 2021.

## Xóa bỏ các trường không cần thiết

Ở đây trong bộ dữ liệu có giá mở cửa, giá đóng cửa, Giá cao nhất , Giá Thấp Nhất, Giá đóng cửa điều chỉnh và Ngày . Nhóm chỉ sử dụng **giá đóng cửa điều chỉnh** và **ngày** để thực hiện xây dựng mô hình dự báo.

## Chia bộ dữ liệu thành tập train và test

### Ở đây nhóm sẽ chia dự liệu thành 3 kịch bản cụ thể như sau

|  |  |  |
| --- | --- | --- |
| Kịch bản | Tập Train | Tập Test |
| 1 | 17-9-2014 tới 1-10-2022 (2937 data) | 2-10-2022 tới 8-10-2022 |
| 2 | 17-10-2014 tới 1-11-2022 (2938 data) | 2-11-2022 tới 8-11-2022 |
| 3 | 17-11-2014 tới 1-12-2022 (2937 data) | 2-12-2022 tới 8-12-2022 |

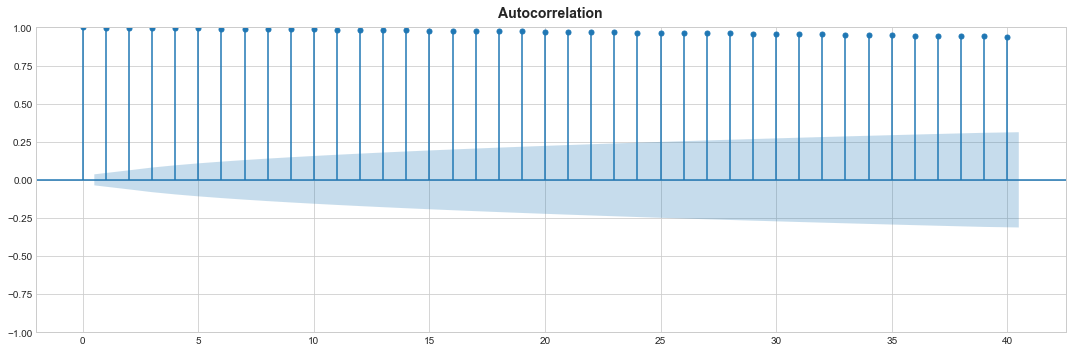
Mục tiêu sẽ thực hiện train model trong 3 kịch bản này thực hiện dự báo , nhầm đánh giá chọn ra model phù hợp cho dữ liệu.

## Kiểm tra tính dừng của chuỗi

Để có thể thể xây dựng được mô hình thì yêu cầu chuỗi dữ liệu phải dừng tức là

* Trung bình của phân phối không phụ thuộc vào thời gian.
* Phương sai của phân phối không phụ thuộc vào thời gian.
* Hiệp phương sai giữa các giá trị của nó tại hai thời điểm bất kỳ chỉ phụ thuộc vào khoảng cách giữa các điểm đó chứ không phụ thuộc vào thời gian.

Để có thể xác định chuỗi đã dừng hay chưa ta xem xét độ thì ACF và kiểm định Dickey–Fuller



Vẽ đồ thị ACF của giá đóng cửa điều chỉnh của Bitcoin ta được các độ trễ từ bậc 1 cho tới 40 điều dương và rất lớn sấp sỉ bằng 1 như vậy chuỗi không thể dừng.

Thực hiện kiểm định Dickey-Fuller với cặp giả thuyết như sau

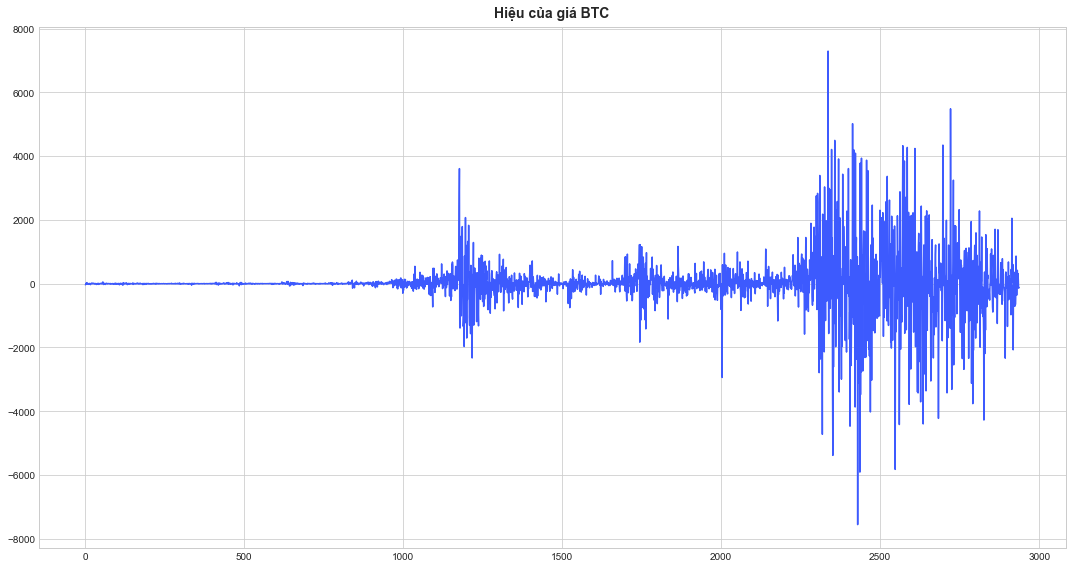
* H0: có nghiệm đơn vị (chuỗi không dừng)
* H1: không có nghiệm đơn vị (chuỗi dừng)

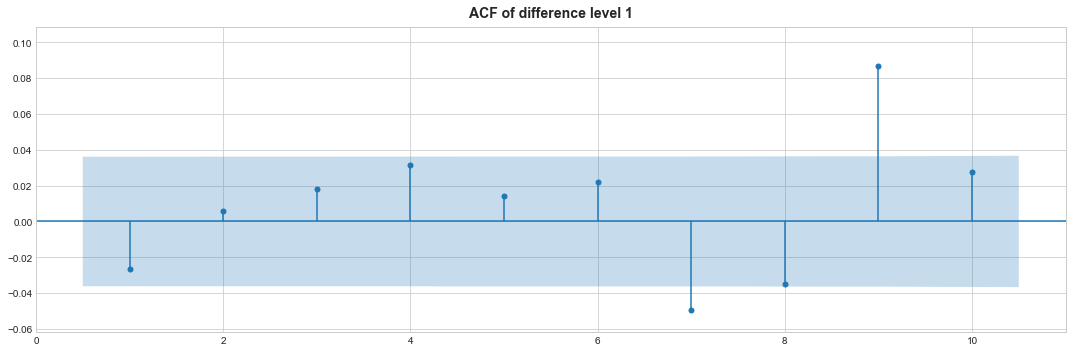
Với mức ý nghĩa 5%.

Ta đươc p\_value =0.445651 không bác bỏ H0 chuỗi không dừng.

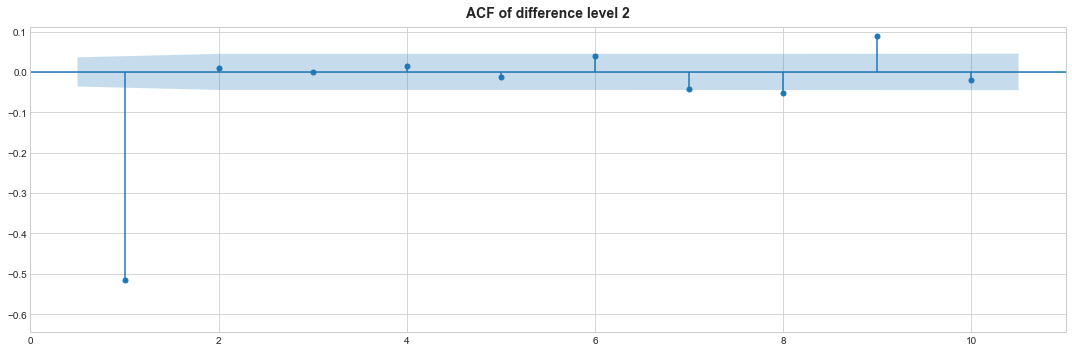
### Thực hiện lấy sai phân và xem lại biểu đồ ACF

train\_df['price\_diff\_1'] = train\_df['price'].diff()



Chuỗi đã bắt đầu giao đông quanh mức 0 hình dạng có khả năng là chuỗi sẽ dừng.

Ta xem lại biểu đồ ACF thì thấy các giá trị tự tương qua giảm mạnh gần về không việc lấy sai phân đã có hiệu quả . Nhưng để chắc chắn hơn thì chúng ta sẽ lấy thêm một lần sai phân nữa để kiểm tra đây có phải sai phân tất nhất hay chưa.



ACF của sai phân bậc 2 , ta nhận thấy hệ số tự tương của của độ trễ thứ nhất từ dương chuyển sang âm cho nên , dữ liệu đã bị overdifference quá mức cho nên hệ số sai phân (difference) băng 1 là hợp lí cho chuỗi dữ liệu dừng.

Thực hiện kiểm định Dickey-Fuller lại 1 lần nữa ta được Dickey–Fuller test: p=0.000 cho nên chuỗi đã dừng.

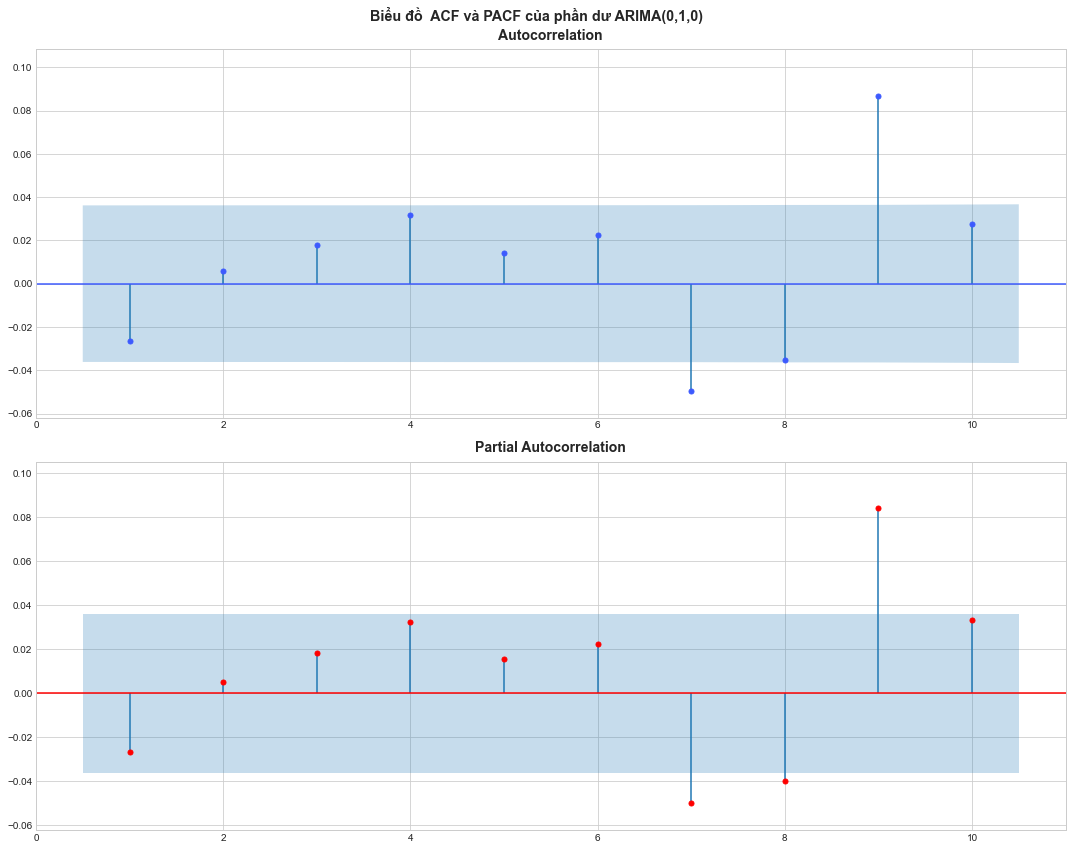
Kết luận hệ số sai phân (difference) 1 là tốt nhất để ổn định chuỗi dữ liệu trước khi ta thực hiện xây dựng mô hình ARIMA với tham số d=1.

Cách xác định được lấy từ nguồn tham khảo : [Identifying the order of differencing in an ARIMA model](https://people.duke.edu/~rnau/411arim2.htm)

# Chọn các mô hình ARIMA

#### Kiểm tra đồ thị ACF và PACF phần dư của mô hình ARIMA(0,1,0) tức d=1 , p=q=0 ở mức ý nghĩa 5%

Như chúng ta đã đã có sai phân bậc một (difference level 1) giá đóng cửa điều chỉnh của của Bitcoin cho nên nó chính là phần dư của mô hình ARIMA(0,1,0)



Các có các giá trị của ACF và PACF như sau

lag acf pacf  
 1 -0.026543 -0.026552  
 2 0.005949 0.005252  
 3 0.017885 0.018213  
 4 0.031502 0.032503  
 5 0.013977 0.015539  
 6 0.022258 0.0224556

7 -0.049511 -0.049885  
 8 -0.035045 -0.039894  
 9 0.086670 0.084177  
 10 0.027338 0.033237

Theo kết quả của đồ thị ACF thì ta có các lag 7,9 có ý nghĩa thống kê, ngược ra ngoài 0.035. Trong khi kết quả của biểu đồ PACF thì các lag 7,8,9 có ý nghĩa thống kê vược ra ngoài 0.035. Từ kết quả trên nhóm chọn được các tham số p và q để xây dựng mô hình ARIMA. Giá trị của d(difference) là một vì chuỗi dừng ở lần lấy sai phân bậc một. Từ các độ trễ của có nghĩa thống kê của ACF và PACF ta có q=[7,9] vì ACF thể hiện tính MA của chuỗi (Moving Average) và p=[7,8,9] vì PACF đại điện cho tính AR (Autoregressive). Kết hợp các tham số trên chúng ta sẽ có các mô hình ARIMA như sau: (7,1,7);(7,1,9);(8,1,7);(8,1,9);(9,1,7);(9,1,9) tổng cộng có 6 mô hình. Ta thực hiện xây dừng từng mô hình dựa dự báo trên 3 kịch bản của dữ liệu chọn mô hình nào cho MAPE bé nhất sẽ là mô hình ARIMA tốt nhất.