

**TRƯỜNG** **ĐẠI** **HỌC** **KINH TẾ - ĐẠI HỌC ĐÀ NẴNG**

**KHOA** **THỐNG KÊ TIN HỌC**

**---------\*\*\*---------**

\



**BÁO CÁO BÀI TẬP NHÓM**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU BẰNG PYTHON**

**ĐỀ TÀI: DỰ BÁO GIÁ VÀNG VIỆT NAM**

**BẰNG MÔ HÌNH ARIMA**

**Group 1 – 48K21.2**

Giảng viên hướng dẫn: Lê Diên Tuấn

Thành viên: Phần trăm đóng góp:

Nguyễn Thị Cẩm Nhi (0354004759) 100%

Lê Thị Kim Cương 99%

Hồ Thị Thu Hà 100%

Lê Thị Mỹ Trâm 99%

Vũ Kiều Tố Như 99%

***Đà Nẵng, 12/5/2025***

# MỤC LỤC

[MỤC LỤC 2](#_Toc197968263)

[DANH MỤC CÁC BẢNG 3](#_Toc197968264)

[DANH MỤC HÌNH ẢNH 4](#_Toc197968265)

[LỜI MỞ ĐẦU 5](#_Toc197968266)

[CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN 6](#_Toc197968267)

[1. Lí do lựa chọn đề tài 6](#_Toc197968268)

[2. Mục tiêu 6](#_Toc197968269)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 7](#_Toc197968270)

[1. Quy trình nghiên cứu (Framework) 7](#_Toc197968271)

[2. Dãy số thời gian (Time series) 8](#_Toc197968272)

[2.1 Dãy số thời gian 8](#_Toc197968273)

[2.2 Các loại dãy số thời gian 9](#_Toc197968274)

[3. Tổng quan về mô hình dự báo chuỗi thời gian (Time Series Forecasting) 10](#_Toc197968275)

[3.1 Dự báo là gì? 10](#_Toc197968276)

[3.2 Dự báo chuỗi thời gian (Time Series Forecasting) 10](#_Toc197968277)

[3.3 Dự báo chuỗi thời gian bằng mô hình hồi quy tuyến tính (Linear Regression) 11](#_Toc197968278)

[3.4 Dự báo chuỗi thời gian bằng mô hình ARIMA 13](#_Toc197968279)

[CHƯƠNG 3: TRIỂN KHAI DỰ ÁN BẰNG PYTHON 14](#_Toc197968280)

[1. Tổng quan và mô tả dữ liệu 14](#_Toc197968281)

[1.1 Tổng quan dữ liệu 14](#_Toc197968282)

[1.2 Mô tả dữ liệu 15](#_Toc197968283)

[2. Kiểm tra và tiền xử lý dữ liệu 16](#_Toc197968284)

[2.1 Kiểm tra dữ liệu 16](#_Toc197968285)

[2.2 Tiền xử lý dữ liệu 17](#_Toc197968286)

[3. Trực quan hóa dữ liệu 18](#_Toc197968287)

[4. Xây dựng mô hình và đánh giá 21](#_Toc197968288)

[4.1 Mô hình ARIMA 21](#_Toc197968289)

[4.2 Mô hình Hồi quy tuyến tính 27](#_Toc197968290)

[5. So sánh và lựa chọn mô hình tốt nhất 30](#_Toc197968291)

[6. Dự báo kết quả 31](#_Toc197968292)

[CHƯƠNG 4: KẾT LUẬN 32](#_Toc197968293)

[TÀI LIỆU THAM KHẢO 33](#_Toc197968294)

[VIDEO 33](#_Toc197968295)

# 

# DANH MỤC CÁC BẢNG

[Bảng 1: Ví dụ về dãy số thời gian 8](#_Toc197981917)

[Bảng 2: Ví dụ về dãy số thời kỳ 9](#_Toc197981918)

[Bảng 3: Ví dụ về dãy số thời điểm 10](#_Toc197981919)

[Bảng 4: Tổng quan dữ liệu 14](#_Toc197981920)

[Bảng 5: Kết quả đánh giá mô hình ARIMA 26](#_Toc197981921)

[Bảng 6: So sánh đánh giá 2 mô hình 30](#_Toc197981922)

# DANH MỤC HÌNH ẢNH

[Hình 1: Quy trình nghiên cứu 8](#_Toc197965447)

[Hình 2: Tổng quan dữ liệu 14](#_Toc197965448)

[Hình 3: Mô tả dữ liệu 15](#_Toc197965449)

[Hình 4: Kết quả kiểm tra dữ liệu bị thiếu 16](#_Toc197965450)

[Hình 5: Boxplot để phát hiện điểm ngoại lai 16](#_Toc197965451)

[Hình 6: Chuyển kiểu dữ liệu cột Date sang datetime 17](#_Toc197965452)

[Hình 7: Kết quả sau khi bị xử lý giá trị ngoại lai 17](#_Toc197965453)

[Hình 8: Biểu đồ đường: Xu hướng giá theo thời gian 18](#_Toc197965454)

[Hình 9: Histogram phân phối dữ liệu 19](#_Toc197965455)

[Hình 10: Biểu đồ hộp (Boxplot) - Biến động giá theo năm 20](#_Toc197965456)

[Hình 11: Kết quả kiểm định tính dừng 20](#_Toc197965457)

[Hình 12: Biểu diễn kết quả sau sai phân bậc 1 21](#_Toc197965458)

[Hình 13: Kết quả sau khi sai phân bậc 1 21](#_Toc197965459)

[Hình 14: Biểu đồ phân tích thành phần mùa vụ 21](#_Toc197965460)

[Hình 15: Xây dựng hàm tự động và kết quả 22](#_Toc197965461)

[Hình 16: Xây dựng mô hình ARIMA cho giá mua và bảng tóm tắt mô hình 24](#_Toc197965462)

[Hình 17: Xây dựng mô hình ARIMA cho giá bán và bảng tóm tắt mô hình 24](#_Toc197965463)

[Hình 18: Biểu đồ giá trị thực tế và giá trị dự báo 25](#_Toc197965464)

[Hình 19: Tính toán các chỉ số đánh giá mô hình 25](#_Toc197965465)

[Hình 20: Tính trực quan hóa xu hướng giá vàng theo thời gian 26](#_Toc197965466)

[Hình 21: Hiển thị bảng sau khi thêm cột time 27](#_Toc197965467)

[Hình 22: Các đặc trưng và mục tiêu huấn luyện 27](#_Toc197965468)

[Hình 23: Kết quả xây dựng mô hình hồi quy 28](#_Toc197965469)

[Hình 24: Kết quả đánh giá mô hình 29](#_Toc197965470)

[Hình 25: Kết quả dự báo 30](#_Toc197965471)

# LỜI MỞ ĐẦU

Nghiên cứu này chủ yếu sử dụng mô hình ARIMA để dự đoán giá vàng tại Việt Nam dựa trên chuỗi dữ liệu lịch sử. Dữ liệu được thu thập, tiền xử lý và triển khai mô hình trên nền tảng Python. Ngoài ra, mô hình hồi quy tuyến tính (Linear Regression - LR) cũng được sử dụng nhằm mục đích so sánh hiệu quả dự báo với mô hình ARIMA. Các chỉ số đánh giá như MAD (Mean Absolute Deviation) và SSE (Sum of Squared Errors) cho thấy mô hình ARIMA cho kết quả chính xác hơn so với mô hình LR. Điều này cho thấy ARIMA phù hợp hơn trong việc mô hình hóa và dự đoán chuỗi thời gian có đặc điểm biến động như giá vàng. Nghiên cứu cũng đóng vai trò tham khảo cho các nhà đầu tư hoặc nhà phân tích tài chính trong việc lựa chọn phương pháp dự báo phù hợp.

# CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN

## Lí do lựa chọn đề tài

Trong bối cảnh nền kinh tế toàn cầu thường xuyên biến động và chịu ảnh hưởng mạnh từ các yếu tố như lạm phát, khủng hoảng tài chính và chính sách tiền tệ, vàng từ lâu đã được xem là một tài sản đầu tư an toàn và có tính phòng ngừa rủi ro cao. Việc dự báo giá vàng ngày càng trở thành một nhiệm vụ chiến lược trong lĩnh vực tài chính – đầu tư. Khả năng nhận diện xu hướng biến động của giá vàng không chỉ giúp nhà đầu tư tối ưu hóa danh mục, mà còn hỗ trợ các tổ chức tài chính và cơ quan quản lý trong việc xây dựng các chính sách điều hành phù hợp.

Đề tài **“Dự báo giá vàng tại Việt Nam bằng mô hình ARIMA”** tập trung vào việc khai thác các phương pháp phân tích chuỗi thời gian để dự đoán xu hướng giá vàng trong tương lai. Cốt lõi của đề tài là vận dụng mô hình ARIMA – một phương pháp phân tích dữ liệu hiệu quả, đặc biệt phổ biến trong lĩnh vực dự báo kinh tế và tài chính. Đồng thời, mô hình hồi quy tuyến tính cũng được triển khai song song nhằm so sánh độ chính xác và hiệu quả giữa các phương pháp truyền thống và hiện đại trong xử lý dữ liệu thời gian.

Đề tài có mối liên hệ chặt chẽ với học phần *Phân tích dữ liệu*, khi toàn bộ quy trình nghiên cứu – từ thu thập, xử lý, phân tích dữ liệu đến đánh giá mô hình – đều vận dụng các kiến thức và kỹ thuật cốt lõi của môn học, như: phân tích chuỗi thời gian, kiểm định tính dừng, lựa chọn mô hình tối ưu, và đánh giá sai số dự báo bằng các chỉ số thống kê như MAD, SSE. Qua đó, đề tài không chỉ thể hiện khả năng áp dụng lý thuyết vào thực tiễn mà còn góp phần phát triển tư duy phân tích dữ liệu một cách logic và có hệ thống.

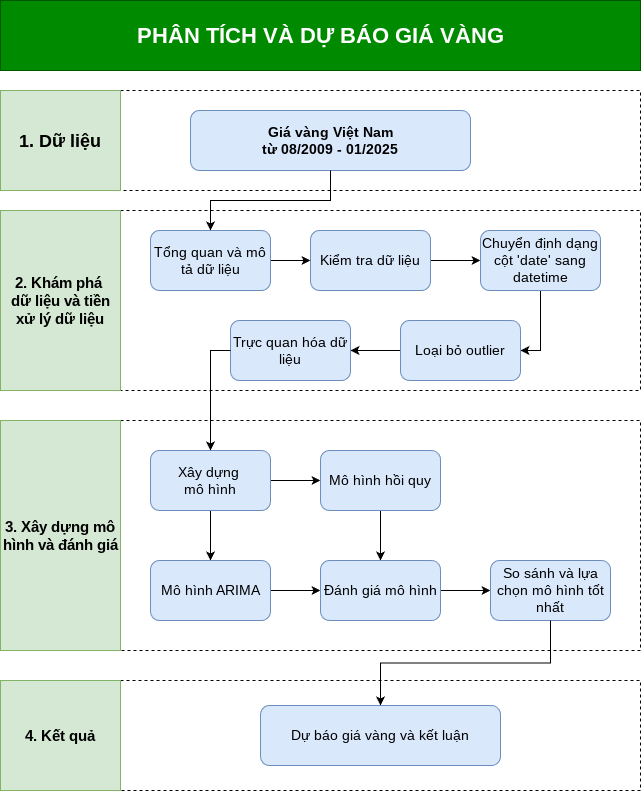
## Mục tiêu

* Mục tiêu tổng quát:
  + Nghiên cứu và dự báo giá vàng trong tương lai bằng cách sử dụng mô hình chuỗi thời gian ARIMA. Nghiên cứu nhằm phân tích xu hướng biến động giá vàng trong quá khứ, từ đó xây dựng mô hình dự báo có độ chính xác cao, hỗ trợ các nhà đầu tư và tổ chức tài chính ra quyết định hiệu quả trong bối cảnh thị trường nhiều biến động.
* Mục tiêu cụ thể:
  + Thu thập và xử lý dữ liệu giá nhà theo thời gian để đảm bảo tính đầy đủ, chính xác và phù hợp cho quá trình phân tích.
  + Phân tích đặc điểm chuỗi thời gian của giá nhà như xu hướng, mùa vụ và độ biến động nhằm hiểu rõ hành vi thay đổi của thị trường bất động sản.
  + Áp dụng mô hình chuỗi thời gian ARIMA để dự báo giá nhà trong tương lai dựa trên dữ liệu lịch sử đã xử lý.
  + Đánh giá hiệu suất dự báo của mô hình thông qua các chỉ số thống kê như MAD, SSE để đảm bảo tính chính xác và độ tin cậy của mô hình.
  + Đề xuất khả năng ứng dụng kết quả dự báo vào thực tiễn như hỗ trợ ra quyết định đầu tư, xây dựng chính sách phát triển nhà ở, hoặc điều chỉnh chiến lược kinh doanh bất động sản.
  + Mở rộng hướng ứng dụng mô hình ARIMA trong lĩnh vực tài chính – đầu tư, đồng thời làm cơ sở cho các nghiên cứu tiếp theo về dự báo tài chính sử dụng chuỗi thời gian.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Quy trình nghiên cứu (Framework)

Trong nghiên cứu này, mô hình dự đoán giá vàng trong tương lai được thực hiện theo quy trình sau:



Hình 1: Quy trình nghiên cứu

## Dãy số thời gian (Time series)

### Dãy số thời gian

* **Khái niệm**: Dãy số thời gian là dãy các trị số của chỉ tiêu thống kê được sắp xếp theo một thứ tự thời gian. (1)
* **Ví dụ**: Giá cả hàng ngày của một cổ phiếu nào đó ở thị trường chứng khoán, dãy số liệu về GDP của một quốc gia qua các năm.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Năm | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
| GDP(tỷ USD) | 101 | 112 | 134 | 155 | 170 | 191 | 201 |

Bảng 1: Ví dụ về dãy số thời gian

* **Kết cấu:** Dãy số thời gian gồm 2 phần:
  + Thời gian chỉ rõ biểu hiện của biến số là thời gian nào (tuần, tháng, quý, năm,...). Độ dài giữa hai thời gian liền nhau gọi là khoảng cách thời gian.
  + Chỉ tiêu về hiện tượng nghiên cứu như: Chỉ tiêu sản lượng, giá trị sản xuất,...Trị số của chỉ tiêu nghiên cứu gọi là các mức độ của dãy số thời gian được biểu hiện bằng số tuyệt đối hoặc số tương đối hoặc số bình quân của thời kỳ đó.
* **Ý nghĩa của dãy số thời gian:** 
  + Nghiên cứu các đặc điểm về sự biến động, tình hình phát triển của hiện tượng theo thời gian.
  + Vạch rõ xu hướng và tính quy luật của sự biến động.
  + Dự đoán các mức độ của hiện tượng trong tương lai.

### Các loại dãy số thời gian

* **Dãy số thời kỳ:** Là dãy số mà mỗi mức độ của nó biểu hiện quy mô, khối lượng của hiện tượng trong từng khoảng thời gian nhất định.
  + **Đặc điểm:** 
    - Mỗi mức độ của dãy số phản ánh quy mô hiện tượng trong thời kỳ nào đó.
    - Có thể cộng dồn các mức độ của dãy số để phản ánh quy mô hiện tượng trong khoảng thời gian dài hơn.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Năm | 2009 | 2010 | 2011 | 2012 | 2017 | 2018 | 2019 |
| Giá trị xuất khẩu   (Triệu USD) | 12,0 | 13,4 | 14,5 | 15,6 | 20,5 | 21,2 | 22,1 |

Bảng 2: Ví dụ về dãy số thời kỳ

* **Dãy số thời điểm:** Là dãy số mà mỗi mức độ của nó biểu hiện quy mô,khối lượng của hiện tượng tại một thời điểm nhất định.
  + **Đặc điểm:**
* Các trị số của dãy số thời điểm chỉ phản ánh mặt lượng của hiện tượng tại những thời điểm nhất định.
* Mức độ của các hiện tượng ở thời điểm sau thường bao gồm toàn bộ hoặc một bộ phận mức độ của hiện tượng tại thời điểm đó. Vì vậy việc cộng các trị số của chỉ tiêu không phản ánh quy mô của hiện tượng.
  + **Ví dụ**: Số liệu về giá trị hàng hóa tồn kho của DN B vào thời điểm đầu tháng của quý I trong năm 2022 như sau (kiểm kê vào 1 ngày hàng tháng)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thời điểm (Ngày) | 1/1/22 | 1/2/22 | 1/3/22 | 31/3/22 |
| Giá trị hàng hóa tồn kho (tr.đồng) | 420 | 470 | 352 | 364 |

Bảng 3: Ví dụ về dãy số thời điểm

## Tổng quan về mô hình dự báo chuỗi thời gian (Time Series Forecasting)

### Dự báo là gì?

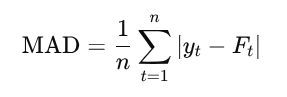
Dự báo là sự tiên đoán những vấn đề sẽ xảy ra trong tương lai dựa trên một cơ sở nào đó. Đây là một vấn đề luôn nhận được sự quan tâm của nhiều nhà khoa học, nhà quản lý bởi vì nó có một vai trò rất quan trọng trong thực tế. Tuy nhiên cho đến nay, dự báo vẫn là bài toán chưa có lời giải cuối cùng (Abbasov & Mamedova, 2003; Tai, 2019). Trong thống kê, dựa trên dữ liệu quá khứ, mô hình để dự báo cho tương lai được thiết lập. Đối với dữ liệu dạng chuỗi, hồi quy và chuỗi thời gian là hai mô hình được áp dụng phổ biến ngày nay.

### Dự báo chuỗi thời gian (Time Series Forecasting)

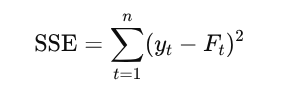
**Dự báo chuỗi thời gian** là một kỹ thuật dự đoán các sự kiện thông qua một chuỗi thời gian. Nó dự đoán các sự kiện trong tương lai bằng cách phân tích các xu hướng trong quá khứ, với giả định rằng các xu hướng trong tương lai sẽ tương tự như vậy.

**Dự báo chuỗi thời gian** cũng là một lĩnh vực quan trọng trong học máy và có thể được coi là vấn đề học tập có giám sát. Các phương pháp học máy như: Regression, Neural Network, Support Vector Machines, Random Forests,...có thể được áp dụng trong trường hợp này.

* **Các mô hình dự báo chuỗi thời gian phổ biến:**
* **Moving Average** (Trung bình trượt): Phương pháp Moving Average làm mượt dữ liệu bằng cách tính trung bình các điểm gần nhau trong quá khứ, giúp giảm nhiễu và làm rõ xu hướng ngắn hạn. Tuy nhiên, nó không hiệu quả với dữ liệu có xu hướng dài hạn hoặc mùa vụ rõ rệt.
* **Exponential Smoothing** (San bằng mũ): là kỹ thuật dự báo ưu tiên giá trị gần hiện tại, phản ánh rõ xu hướng mới. Phù hợp với dữ liệu có xu hướng hoặc mùa vụ, nhưng kém hiệu quả khi dữ liệu biến động đột ngột hoặc quá phức tạp.
* **Regression** (Hồi quy): Hồi quy chuỗi thời gian sử dụng các biến độc lập như thời gian, giá trị quá khứ hoặc yếu tố bên ngoài để dự đoán tương lai. Phù hợp với dữ liệu có mối liên hệ rõ giữa các biến, nhưng có thể không phản ánh tốt đặc điểm nội tại của chuỗi nếu xử lý chưa phù hợp.
* **ARIMA**: là mô hình phổ biến cho chuỗi thời gian đơn biến, kết hợp ba thành phần: tự hồi quy (AR), sai phân (I), và trung bình trượt (MA). Thích hợp với dữ liệu có xu hướng nhưng không có mùa vụ rõ rệt. Cần làm ổn định chuỗi trước khi áp dụng và chọn đúng tham số (p, d, q); nếu có mùa vụ, nên dùng SARIMA. (2)
* **Các chỉ số đánh giá độ chính xác của dự báo:**
* **MAD (Mean Absolute Deviation)**: Là trung bình của các sai số tuyệt đối giữa giá trị thực tế và giá trị dự báo.



* **SSE (Sum of Squared Errors):** Là tổng các bình phương sai số giữa giá trị thực tế và dự báo.



### Dự báo chuỗi thời gian bằng mô hình hồi quy tuyến tính

* **Khái niệm:** 
  + Là một thuật toán học có giám sát (supervised learning) trong [Machine Learning](https://trituenhantao.io/tu-dien-thuat-ngu/machine-learning/), nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target). Linear Regression giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến output từ các giá trị của các biến đầu vào.
  + Mô hình hồi quy thường được sử dụng để dự báo chuỗi thời gian, đặc biệt phù hợp với các loại dữ liệu có các đặc điểm sau:
    - **Tính dao động ngẫu nhiên lớn:** nghĩa là dữ liệu có sự biến động mạnh, không đều và không theo quy luật rõ ràng trong ngắn hạn. Những dao động này thường do yếu tố ngẫu nhiên, không dự đoán được gây ra. Ví dụ: Giá vàng, giá cổ phiếu,…
    - **Có tính xu thế**: Là dữ liệu có chiều hướng tăng lên hoặc giảm xuống ổn định theo thời gian, dù có thể vẫn dao động trong ngắn hạn.
* **Phân loại:** 
  + **Simple Linear Regression**: Mô hình này chỉ có một biến độc lập (input [feature](https://trituenhantao.io/tu-dien-thuat-ngu/feature/)) mô tả mối quan hệ tuyến tính giữa biến phụ thuộc (output target) và biến độc lập. Phương trinh của Simple Linear Regression có dạng:

******

* Trong đó:
  + - a: Hệ số chặn
    - b1, b2,…,bn: Hệ số góc tương ứng với từng biến đầu vào
    - ϵ: Sai số
* **Multiple Linear Regression:** Mô hình này có nhiều hơn một biến độc lập, biểu diễn mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc.
* Phương trình của Multiple Linear Regression có dạng:



* Trong đó:
  + - a: Hệ số chặn
    - b1, b2,…,bn: Hệ số góc tương ứng với từng biến đầu vào
    - ϵ: Sai số
* **Mục tiêu:** Mục tiêu của Linear Regression là tìm ra hệ số góc và điểm giao với trục tung sao cho hàm dự đoán tuyến tính đạt được sai số nhỏ nhất. Một trong những cách phổ biến để ước lượng các hệ số là sử dụng phương pháp Ordinary Least Squares (OLS), trong đó chúng ta cần tối thiểu hóa tổng bình phương sai số (sum of squared error).

### Dự báo chuỗi thời gian bằng mô hình ARIMA

* Mô hình ARIMA thường được sử dụng để dự báo chuỗi thời gian, đặc biệt phù hợp với các loại dữ liệu có các đặc điểm sau:
* Dãy số dao động ngẫu nhiên lớn:
  + Nếu có dữ liệu của chuỗi thời gian có sự dao động ngẫu nhiên mạnh, không có mô hình dễ nhận biết từ các yếu tố như tăng/ giảm đều đặn hay chu kỳ rõ ràng, mô hình ARIMA vẫn có thể xử lý tốt nhờ khả năng kết hợp các phần tự hồi quy (AR), trung bình động (MA), và tích phân (I).
* Không có xu thế rõ ràng:
  + ARIMA phù hợp với các chuỗi không có xu thế (trend) hoặc đã được làm mịn xu thế (detrending). Thành phần tích phân (I) của ARIMA giúp loại bỏ xu thế khi cần thiết. (3)
* Mô hình ARIMA sử dụng các tín hiệu quá khứ của chuỗi để dự báo giá trị tương lai, bao gồm hai thành phần chính: Auto Regression (AR) và Moving Average (MA).
  + Auto Regression (AR)
    - Khái niệm: Thành phần AR giả định rằng giá trị hiện tại của chuỗi thời gian (yₜ) phụ thuộc tuyến tính vào các giá trị trong quá khứ của chính nó, cùng với một sai số ngẫu nhiên (εₜ).
    - Ý nghĩa: AR tận dụng mối liên hệ giữa các giá trị trước đó để dự đoán giá trị hiện tại. Bậc của mô hình (p) quyết định số lượng giá trị quá khứ được sử dụng, cho phép mô hình phản ánh các mối quan hệ tuyến tính phức tạp hơn.
  + Moving Average (MA)
    - Khái niệm: Thành phần MA giả định rằng giá trị hiện tại của chuỗi (yₜ) được xác định từ tổ hợp tuyến tính của các sai số ngẫu nhiên (ε) tại các thời điểm trước đó.
    - Ý nghĩa: MA sử dụng thông tin từ các nhiễu động trước đó để điều chỉnh và cải thiện độ chính xác của dự báo.

# CHƯƠNG 3: TRIỂN KHAI DỰ ÁN BẰNG PYTHON

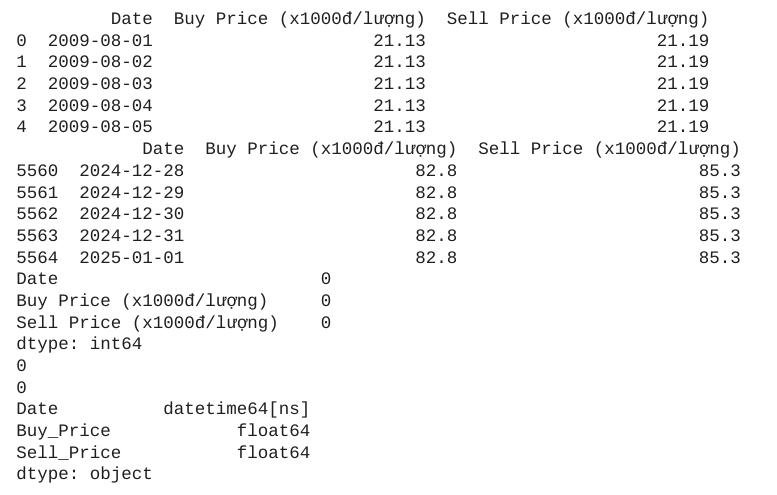
## Tổng quan và mô tả dữ liệu

### Tổng quan dữ liệu

Bộ dữ liệu này ghi nhận giá mua và bán vàng tại Việt Nam, phản ánh sự biến động của giá vàng trong giai đoạn từ ngày 01/08/2009 đến ngày 01/01/2025, phục vụ cho việc phân tích xu hướng biến động giá theo thời gian. Dữ liệu được thu thập theo đơn vị x1000 đồng/lượng, gồm 5.565 dòng và ba cột chính: ngày ghi nhận (Date), giá mua (Buy Price) và giá bán (Sell Price).

|  |  |  |
| --- | --- | --- |
| **Cột dữ liệu** | **Kiểu dữ liệu** | **Ý nghĩa** |
| **Date** | object | Ngày ghi nhận giá mua và bán của vàng. Cột này cho biết thời gian cụ thể của mỗi lần cập nhật dữ liệu giá. |
| **Buy Price** | float64 | Giá mua vào của vàng được ghi nhận tại thời điểm tương ứng, đơn vị tính là nghìn đồng trên một lượng vàng. |
| **Sell Price** | float64 | Giá bán ra của vàng tại thời điểm tương ứng, đơn vị tính là nghìn đồng trên một lượng vàng. |

Bảng 4: Tổng quan dữ liệu

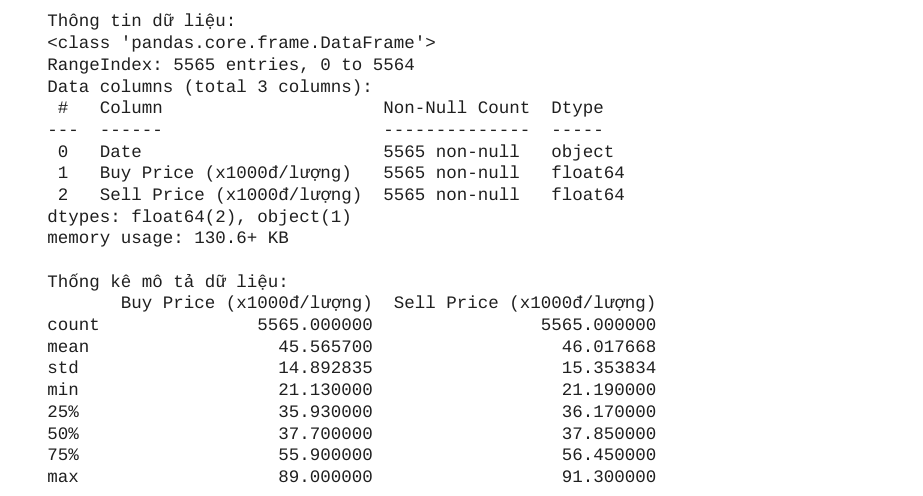


Hình 2: Tổng quan dữ liệu

(Nguồn:<https://www.kaggle.com/datasets/hoangnam729/gi-vng-ti-vit-nam-01082009-31122024>)

### Mô tả dữ liệu

* **Giá mua vàng (Buy Price)**
* Giámua dao động từ mức thấp nhất là 21.13 nghìn đồng/lượng đến mức cao nhất là 89 nghìn đồng/lượng.
* Trung bình giá mua là 45.57 nghìn đồng/lượng, với độ lệch chuẩn 14.89, phản ánh sự biến động khá mạnh trong suốt khoảng thời gian quan sát.
* Phân vị cho thấy 50% giá trị nằm trong khoảng từ 35.93 đến 55.90 nghìn đồng/lượng, tập trung quanh giá trị trung vị là 37.70 nghìn đồng/lượng.
* **Giá bán vàng (Sell Price)**
* Giábán dao động từ 21.10 nghìn đồng/lượng đến 91.30 nghìn đồng/lượng.
* Trung bình giá bán là 46.02 nghìn đồng/lượng, cao hơn giá mua một khoảng nhỏ (~0.45 nghìn đồng/lượng).
* Độ lệch chuẩn của giá bán là 15.35, tương tự giá mua, cho thấy sự biến động đồng nhất giữa hai loại giá.
* **Nhận xét:** 
  + **Mối quan hệ giữa giá mua và giá bán**: Giá bán luôn cao hơn giá mua, cho thấy khoảng chênh lệch lợi nhuận ổn định.
  + **Biến động giá**: Giá vàng có sự biến động lớn với biên độ rộng (21.13 - 89 nghìn đồng/lượng), đặc biệt là trong các giai đoạn giá tăng mạnh.
  + **Phân bổ dữ liệu**: Phần lớn giá trị tập trung trong khoảng 37-56 nghìn đồng/lượng, cho thấy sự ổn định tương đối của giá vàng trong một khoảng thời gian nhất định.
* Bộ dữ liệu này cung cấp thông tin quan trọng để phân tích xu hướng và dự báo giá vàng trong tương lai. Việc kiểm tra và mô tả dữ liệu chi tiết giúp đảm bảo độ tin cậy trước khi tiến hành các bước phân tích tiếp theo.

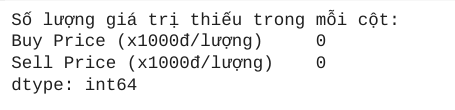


Hình 3: Mô tả dữ liệu

## Kiểm tra và tiền xử lý dữ liệu

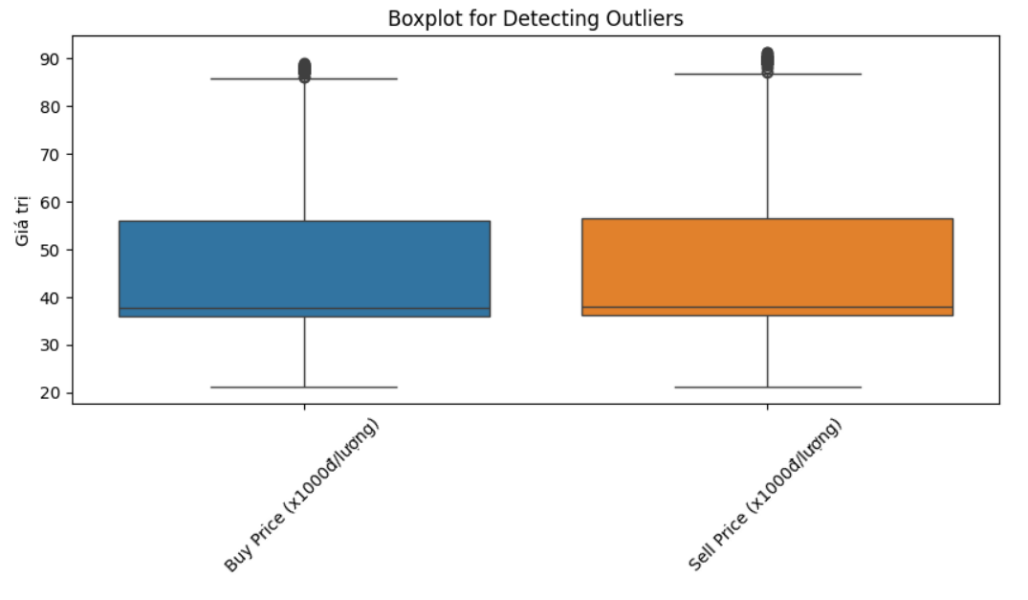
### Kiểm tra dữ liệu

* **Kiểm tra dữ liệu bị thiếu**

****

Hình 4: Kết quả kiểm tra dữ liệu bị thiếu

* **Dữ liệu đầy đủ:** Không có giá trị nào bị thiếu trong các thuộc tính, điều này đảm bảo rằng tất cả các biến đều được sử dụng trong mô hình mà không cần phải loại bỏ hoặc thay thế bất kỳ giá trị nào.
* **Kiểm tra giá trị ngoại lai**



Hình 5: Boxplot để phát hiện điểm ngoại lai

* **Nhận xét:**

Biểu đồ boxplot cho thấy cả "Buy Price" và "Sell Price" đều xuất hiện một số giá trị ngoại lai (outliers) ở phía trên, điều này phản ánh sự tồn tại của những mức giá cao bất thường trong dữ liệu. Những giá trị này có thể là kết quả của các biến động mạnh hoặc các sự kiện tác động đến thị trường vàng trong giai đoạn quan sát.

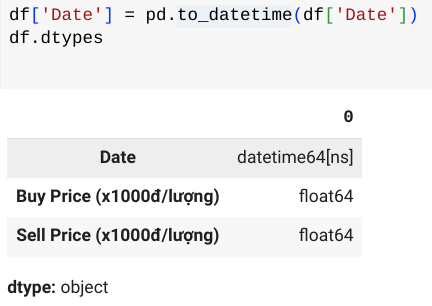
Tuy nhiên, phần lớn các giá trị của cả hai loại giá đều phân bố khá tập trung quanh khoảng giá trung vị, cho thấy giá vàng thường ổn định và không có sự thay đổi quá lớn trong một phần lớn thời gian.

Hình dạng phân bố của "Buy Price" và "Sell Price" tương đối giống nhau, điều này cho thấy sự chênh lệch giữa giá mua và giá bán là không quá lớn, duy trì ở mức ổn định. Sự tương đồng này có thể phản ánh chính sách giá của các cửa hàng vàng, nơi giá mua và bán chỉ chênh lệch một khoảng nhỏ, giúp đảm bảo lợi nhuận ổn định.

Tóm lại, mặc dù có sự biến động và các giá trị ngoại lai xuất hiện, bộ dữ liệu vẫn cho thấy một bức tranh tổng thể khá ổn định về giá vàng trong giai đoạn quan sát.

### Tiền xử lý dữ liệu

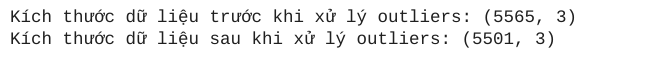
* **Chuyển kiểu dữ liệu cột Date sang datetime**



Hình 6: Chuyển kiểu dữ liệu cột Date sang datetime

Nhằm mục đích khai thác thông tin theo chuỗi thời gian và thực hiện các phép toán liên quan đến ngày tháng, cột 'Date' đã được chuyển đổi sang định dạng datetime64[ns]. Việc chuyển đổi này là bước tiền xử lý quan trọng để chuẩn bị cho các bước phân tích chuyên sâu hơn.

#### **Xử lý giá trị ngoại lai**



Hình 7: Kết quả sau khi bị xử lý giá trị ngoại lai

**Nhận xét:** Việc xử lý outliers đã dẫn đến việc giảm số lượng hàng trong tập dữ liệu từ 5565 xuống còn 5502. Điều này cho thấy có 5565 - 5502 = 63 hàng đã bị loại bỏ hoặc được điều chỉnh trong quá trình xử lý outliers.

## Trực quan hóa dữ liệu

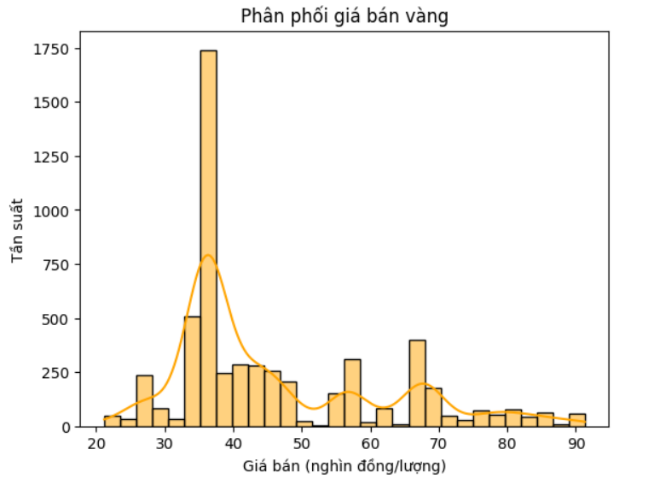
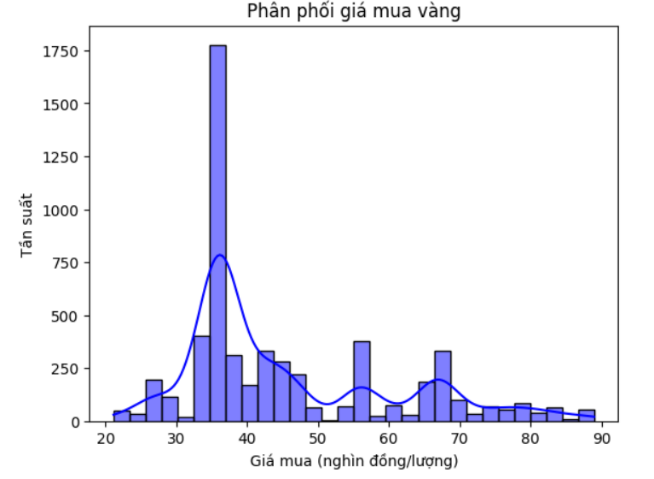
* **Biểu đồ đường: Xu hướng giá theo thời gian** 
  + Đầu tiên, chúng tôi trực quan hóa bằng biểu đồ đường vì đây là công cụ trực quan hiệu quả nhất để thể hiện dữ liệu chuỗi thời gian như giá vàng theo ngày. Nó giúp người xem dễ dàng theo dõi diễn biến của giá theo từng mốc thời gian cụ thể, nhận diện xu hướng dài hạn và ngắn hạn. Quan trọng hơn, biểu đồ này cho phép phát hiện các điểm bất thường, các đợt biến động mạnh và các thời kỳ ổn định – từ đó hỗ trợ việc phân tích nguyên nhân, tác động từ chính sách kinh tế, hay các sự kiện toàn cầu đến thị trường vàng. Việc sử dụng biểu đồ đường là rất cần thiết khi mục tiêu là hiểu rõ xu hướng và sự biến động theo thời gian.



Hình 8: Biểu đồ đường: Xu hướng giá theo thời gian

Nhận xét các đặc trưng nổi bật và ý nghĩa của phân bố dữ liệu

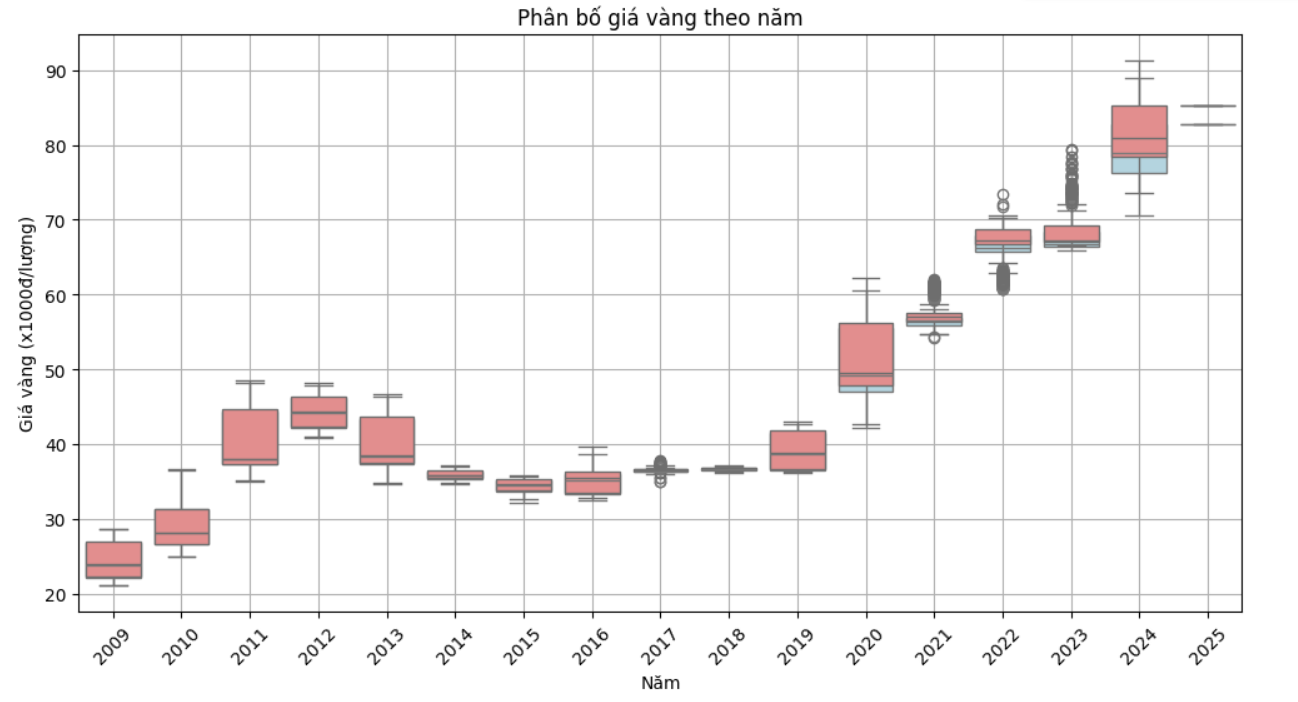
* Giá mua (Xanh dương): Phân bố theo thời gian có xu hướng tăng dài hạn, đặc biệt sau năm 2020, cho thấy giá vàng liên tục tăng mạnh trong những năm gần đây.
* Giá bán (Đỏ): Luôn cao hơn giá mua, nhưng khoảng cách giữa hai mức giá không quá lớn, phản ánh tính thanh khoản tốt của thị trường vàng.
* Xu hướng chung: Từ 2010 - 2018, giá vàng dao động trong khoảng 30-50 triệu đồng/lượng, sau đó có sự tăng trưởng mạnh từ 2019 đến 2024, thể hiện sự ảnh hưởng của các yếu tố kinh tế vĩ mô.
* Độ biến động: Trước năm 2018, giá vàng có nhiều giai đoạn biến động mạnh nhưng không theo xu hướng tăng rõ rệt. Từ 2019 trở đi, biên độ dao động lớn hơn với nhiều đợt điều chỉnh, cho thấy tác động từ thị trường toàn cầu.
* Chu kỳ tăng giảm: Có sự xuất hiện của các đợt tăng mạnh, sau đó là giai đoạn điều chỉnh trước khi tiếp tục xu hướng đi lên, phản ánh tính chu kỳ của giá vàng.
* Giai đoạn tăng đột biến: Xuất hiện các thời điểm giá vàng tăng nhanh, đáng chú ý nhất là giai đoạn 2020 - 2024, có thể do ảnh hưởng của đại dịch và bất ổn kinh tế.
* **Histogram: Phân phối dữ liệu** 
  + Biểu đồ histogram được sử dụng nhằm kiểm tra mức độ tập trung và phân tán của giá vàng trong toàn bộ tập dữ liệu. Biểu đồ này giúp xác định khoảng giá nào xuất hiện nhiều nhất, đồng thời phản ánh tính phân phối của dữ liệu — liệu có gần chuẩn không, hay bị lệch, có nhiều đỉnh hoặc các cụm dữ liệu. Điều này quan trọng trong việc hiểu thị trường có xu hướng neo giá tại mức nào, và liệu có sự xuất hiện của các mức giá bất thường, từ đó giúp đánh giá rủi ro hoặc hành vi giá trong quá khứ.



Hình 9: Histogram phân phối dữ liệu

Nhận xét biểu đồ Histogram – Phân bố giá vàng:

* Giá vàng phân bố không đều, tập trung nhiều ở một số mức giá nhất định, đặc biệt khoảng 38 triệu đồng/lượng.
* Nhiều cụm giá xuất hiện ở 30 triệu, 50 triệu, 60 triệu, 70 triệu đồng/lượng, phản ánh các giai đoạn biến động mạnh.
* Giá mua và giá bán có xu hướng gần nhau, nhưng giá bán luôn cao hơn, thể hiện mức chênh lệch lợi nhuận.
* Ít dữ liệu ở mức giá cao (trên 80 triệu đồng/lượng), cho thấy đây là mức giá hiếm gặp.
* Biểu đồ giúp hiểu rõ sự thay đổi và phân bố giá vàng qua thời gian, hỗ trợ dự báo xu hướng thị trường.
* **Biểu đồ hộp (Boxplot) – Biến động giá theo năm** 
  + Biểu đồ hộp giúp thể hiện mức độ phân tán của giá vàng theo từng năm. Nó giúp đánh giá mức độ biến động của giá vàng, xác định các năm có biến động mạnh hoặc ổn định, và phát hiện các giá trị ngoại lai (outliers) - những thời điểm giá thay đổi bất thường.



Hình 10: Biểu đồ hộp (Boxplot) - Biến động giá theo năm

Nhận xét:

* Xu hướng tăng mạnh: Giá vàng có xu hướng tăng rõ rệt từ năm 2019 đến 2025, đặc biệt tăng đột biến sau năm 2020.
* Biến động lớn hơn theo thời gian: Giai đoạn trước 2018, giá vàng khá ổn định với hộp nhỏ hơn, nhưng từ 2020 trở đi, biên độ dao động ngày càng lớn.
* Xuất hiện nhiều outlier (điểm bất thường): Đặc biệt từ 2020, giá vàng có nhiều biến động mạnh với các điểm dữ liệu vượt xa phạm vi hộp.
* Giai đoạn 2013 - 2018 tương đối ổn định: Giá vàng dao động quanh một mức cố định, ít biến động.
* Biểu đồ giúp xác định sự thay đổi của giá vàng theo thời gian, mức độ dao động và xu hướng chung, hỗ trợ phân tích rủi ro và dự báo thị trường.

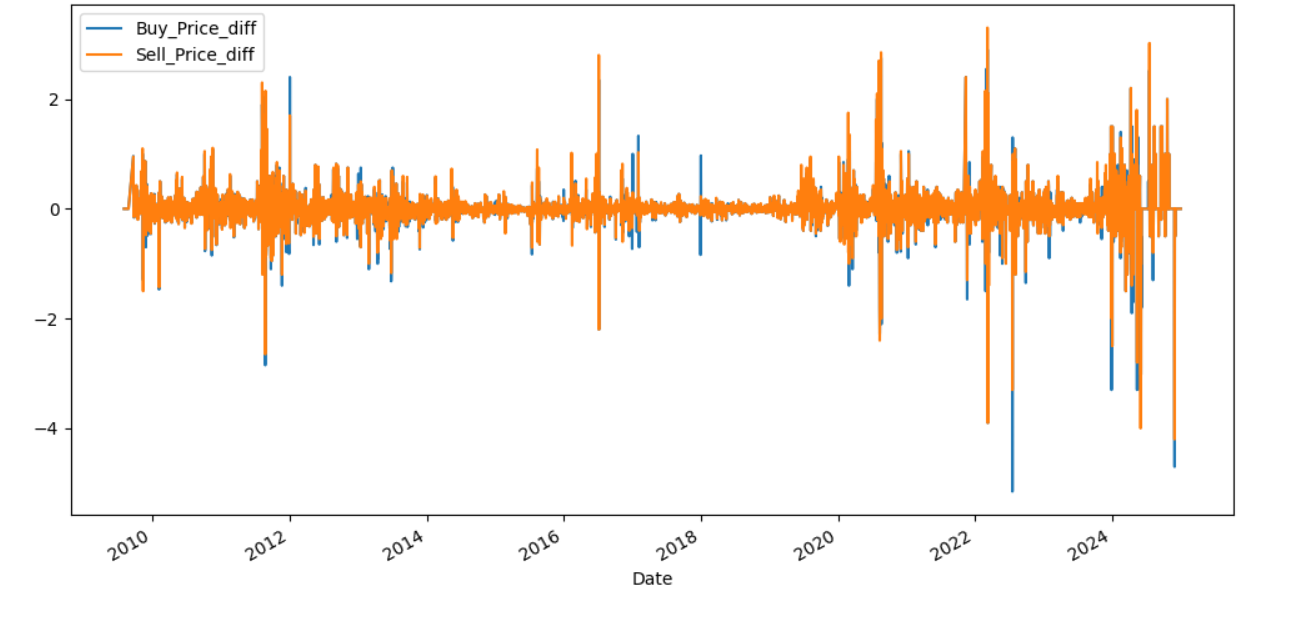
## Xây dựng mô hình và đánh giá

### Mô hình ARIMA

* **Bước 1: Kiểm định tính dừng**

Hình 11: Kết quả kiểm định tính dừng

* + Kết quả kiểm định ADF cho thấy cả "Giá Mua" (p-value: 0.9868) và "Giá Bán" (p-value: 0.9893) đều có p-value rất lớn (> 0.05). Điều này cho thấy cả hai chuỗi thời gian "Giá Mua" và "Giá Bán" đều **không dừng**. Do đó ta cần lấy sai phân trước khi áp dụng các mô hình ARIMA hoặc các phương pháp phân tích chuỗi thời gian dừng khác. (4)
  + **Sau khi lấy sai phân bậc 1:**



Hình 12: Biểu diễn kết quả sau sai phân bậc 1



Hình 13: Kết quả sau khi sai phân bậc 1

* + Sau khi lấy sai phân bậc 1, cả chuỗi thời gian "Buy Price\_diff" (Giá Mua sau sai phân) và "Sell Price\_diff" (Giá Bán sau sai phân) đều trở nên **dừng**. Việc chuyển đổi này là cần thiết để có thể áp dụng các mô hình chuỗi thời gian dừng như ARMA hoặc các mô hình ARIMA (với bậc sai phân d=1).
* **Bước 2: Phân tích thành phần mùa vụ**

|  |  |
| --- | --- |
|  |  |

Hình 14: Biểu đồ phân tích thành phần mùa vụ

* + **Nhận xét:** Sau khi áp dụng phép sai phân bậc một, cả chuỗi thời gian giá mua và giá bán đều thể hiện tính dừng, đáp ứng yêu cầu cho việc xây dựng mô hình ARIMA. Một quan sát đáng chú ý là thành phần mùa vụ, được phân tách trên dữ liệu đã dừng, cho thấy biên độ dao động không đáng kể ở cả hai chuỗi. Điều này gợi ý rằng yếu tố mùa vụ không đóng vai trò chủ đạo trong việc giải thích các biến động sau khi đã loại bỏ xu hướng. Phần dư sau phân tách vẫn cho thấy sự biến động, đòi hỏi phân tích sâu hơn để đảm bảo tính ngẫu nhiên.
  + Sau khi xác nhận chuỗi thời gian đã đạt được tính dừng và yếu tố mùa vụ không có ảnh hưởng đáng kể, bước tiếp theo trong quá trình xây dựng mô hình là xác định các bậc của thành phần tự hồi quy (p) và trung bình trượt (q). Thay vì sử dụng trực tiếp các đồ thị hàm tự tương quan (ACF) và hàm tương quan tự phần (PACF), việc lựa chọn sẽ được thực hiện thông qua một hàm tự động, dựa trên tiêu chí tối ưu như AIC hoặc BIC. Cụ thể, hàm sẽ thử nghiệm các tổ hợp giá trị p và q trong một phạm vi xác định, trên cơ sở mô hình ARIMA với bậc sai phân d = 1 và các thành phần mùa vụ (P, D, Q) bằng 0. Mô hình cho kết quả tối ưu nhất theo tiêu chí đã chọn sẽ được sử dụng để mô hình hóa và dự báo biến động giá một cách hiệu quả hơn.
* **Bước 3: Xác định các tham số ARIMA (p,d,q)** 
  + Lựa chọn tham số p, q bằng cách xây dựng hàm tự động

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |
| **Kết quả:** | **Kết quả:** |

Hình 15: Xây dựng hàm tự động và kết quả

* Nhận xét:
* Cả hai chuỗi thời gian (Buy Price và Sell Price) đều được mô hình ARIMA lựa chọn cấu trúc giống nhau: ARIMA(0,1,0)(0,0,0)[0] intercept

→ Điều này cho thấy **chuỗi chỉ cần sai phân bậc 1 (d = 1)** để trở nên ổn định, **không cần thành phần tự hồi quy (AR = 0)** hay trung bình trượt (MA = 0).

* **Không có yếu tố chu kỳ** (seasonal = False) là phù hợp vì chuỗi giá vàng thường không có tính chu kỳ rõ rệt trong ngắn hạn.
* **Thời gian huấn luyện nhanh** (chỉ vài giây) → mô hình đơn giản và dữ liệu đã được xử lý tốt (không còn outliers, đầy đủ).
* **Kết quả giống nhau** giữa hai chuỗi có thể gợi ý rằng: Hai chuỗi có mối tương quan chặt chẽ và cùng tuân theo một dạng biến động đơn giản: "tăng/giảm tuyến tính ngẫu nhiên" sau sai phân.
* **Kết luận:**
* ARIMA(0,1,0) là một mô hình random walk – tức biến động chủ yếu do yếu tố ngẫu nhiên sau khi loại bỏ xu hướng.
* Đây là lựa chọn hợp lý khi chuỗi không có tính mùa vụ và không có cấu trúc hồi quy rõ ràng.
* Mô hình này phù hợp làm baseline để so sánh với các mô hình phức tạp hơn như ARIMA có AR/MA, SARIMA, hoặc mô hình học sâu (LSTM) nếu cần cải thiện độ chính xác.
* **Bước 4: Xây dựng mô hình**
* **Buy Price**

|  |
| --- |
|  |
|  |

Hình 16: Xây dựng mô hình ARIMA cho giá mua và bảng tóm tắt mô hình

* **Sell Price**

|  |
| --- |
|  |
|  |

Hình 17: Xây dựng mô hình ARIMA cho giá bán và bảng tóm tắt mô hình

* **Nhận xét:**
* Mô hình ARIMA(0,1,0) được xây dựng cho cả giá mua và giá bán là một mô hình đơn giản nhưng phù hợp với đặc điểm của chuỗi dữ liệu, vốn có xu hướng nhưng không thể hiện tính chu kỳ rõ rệt. Đây là dạng mô hình sai phân bậc nhất (random walk), thường được sử dụng để mô tả các chuỗi thời gian mang tính biến động ngẫu nhiên quanh xu hướng. Kết quả ước lượng cho thấy phương sai phần dư (sigma²) rất nhỏ với p-value = 0.000, khẳng định các thông số của mô hình có ý nghĩa thống kê cao.. Nhìn chung, ARIMA(0,1,0) là một mô hình khởi đầu hợp lý, giúp phản ánh xu hướng dữ liệu và có thể được sử dụng làm cơ sở để thử nghiệm các mô hình nâng cao hơn.

|  |  |
| --- | --- |
|  |  |

Hình 18: Biểu đồ giá trị thực tế và giá trị dự báo

* **Bước 5: Đánh giá mô hình**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |

Hình 19: Tính toán các chỉ số đánh giá mô hình

**Kết quả đánh giá mô hình ARIMA**

|  |  |  |
| --- | --- | --- |
| **Buy Price** | MAD = 0.1298 | SSE = 471.8629 |
| **Sell Price** | MAD = 0.1282 | SSE = 445.3161 |

Bảng 5: Kết quả đánh giá mô hình ARIMA

* **Nhận xét:**
* Cả hai chỉ số Mean Absolute Deviation (MAD) cho giá mua và giá bán đều nhỏ (≈ 0.13), cho thấy sai số trung bình giữa giá thực tế và giá dự báo là rất thấp, chứng tỏ mô hình phù hợp tốt với dữ liệu.
* Tổng sai số bình phương (**SSE**) của hai biến số đều ở mức thấp (< 500), cho thấy mô hình không bị sai lệch lớn với dữ liệu thực tế.
* Đường fitted trên biểu đồ trước đó cũng khớp khá sát với dữ liệu gốc, củng cố độ tin cậy của mô hình.
* Cho thấy mô hình ARIMA cho kết quả dự báo rất khả quan với sai số thấp cả về trung bình tuyệt đối và tổng bình phương, cho thấy tiềm năng ứng dụng trong dự báo giá vàng.(5)

### Mô hình Hồi quy tuyến tính

* + **Bước 1: Thêm cột Time**

****  
Hình 21: Hiển thị bảng sau khi thêm cột time

* Sử dụng thư viện numpy để tạo một cột mới gọi là Time trong một DataFrame (df). Cột này được điền bằng một dãy số nguyên từ 0 đến độ dài của chỉ mục DataFrame. Phương thức head() sau đó được sử dụng để hiển thị vài hàng đầu tiên của DataFrame, bao gồm các cột Date, Buy Price, Sell Price, và Time.
  + **Bước 2: Xác định các đặc trưng và mục tiêu từ dữ liệu huấn luyện**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |

Hình 22: Các đặc trưng và mục tiêu huấn luyện

* Sử dụng dữ liệu thời gian (Time) làm đặc trưng đầu vào để huấn luyện mô hình dự đoán. Mục tiêu đầu ra là hai biến giá gồm Buy Price (giá mua) và Sell Price (giá bán). Cụ thể, khi dự đoán giá mua, biến mục tiêu là Buy Price; tương tự, khi dự đoán giá bán, biến mục tiêu là Sell Price. Việc lựa chọn đặc trưng và mục tiêu như vậy giúp mô hình học được mối quan hệ giữa thời gian và giá, từ đó đưa ra các dự đoán phù hợp cho từng trường hợp cụ thể.
  + **Bước 3: Huấn luyện mô hình**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |
| **Kết quả:** | **Kết quả:** |

Hình 23: Kết quả xây dựng mô hình hồi quy

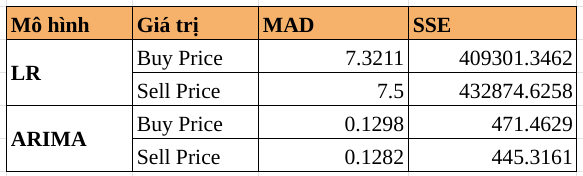
* **Nhận xét:** 
  + Mô hình hồi quy tuyến tính theo thời gian (Time-series forecasting with regression) được áp dụng cho cả chuỗi giá mua và giá bán cho thấy xu hướng tăng rõ rệt. Đường hồi quy màu cam trong cả hai biểu đồ thể hiện một xu hướng tuyến tính tăng dần, trong khi dữ liệu thực tế dao động quanh đường này với biên độ ngày càng lớn.
  + Cụ thể, Ở giai đoạn đầu (từ điểm 0 đến khoảng 1500), cả giá mua và giá bán đều dao động quanh đường hồi quy với biên độ vừa phải, cho thấy mô hình có thể phản ánh tương đối xu hướng trong giai đoạn này. Tuy nhiên, từ khoảng điểm 4000 trở đi, dữ liệu thực tế thể hiện các biến động mạnh, với các đợt tăng giá đột ngột (đỉnh giá vượt 80), trong khi đường hồi quy vẫn giữ nguyên độ dốc tuyến tính, không bắt kịp được sự thay đổi nhanh của chuỗi dữ liệu. Tại điểm cuối cùng trên biểu đồ (~5500), giá mua và giá bán thực tế đều đạt trên 85, trong khi mô hình chỉ dự đoán khoảng 60, chênh lệch khoảng 25 đơn vị – phản ánh mô hình hồi quy tuyến tính đang đánh giá thấp các biến động mạnh mẽ gần đây.
  + Vì vậy, mô hình này không phản ánh được các dao động phi tuyến và các cú sốc đột biến trong dữ liệu.
* **Đánh giá mô hình**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |
| **Kết quả:** | **Kết quả:** |

Hình 24: Kết quả đánh giá mô hình

* **Mô hình dự đoán giá mua (Buy\_Price):** Trên tập dữ liệu, mô hình hồi quy tuyến tính cho thấy sai số dự đoán ở mức trung bình, với độ lệch tuyệt đối trung bình (MAD) là **7.32** và tổng bình phương sai số (SSE) là **409301.35**. Điều này có nghĩa là, trung bình mỗi dự đoán của mô hình lệch khoảng 7.32 đơn vị so với giá mua thực tế. Mặc dù xu hướng chung được mô hình mô tả khá tốt qua đường hồi quy, nhưng biểu đồ cho thấy ở giai đoạn sau (sau điểm 4000), mô hình dự báo thấp hơn nhiều so với thực tế khi giá tăng vọt, làm tăng tổng sai số (SSE) đáng kể.
* **Mô hình dự đoán giá bán (Sell\_Price):** Tương tự, mô hình hồi quy tuyến tính áp dụng cho giá bán có MAD là **7.50** và SSE là **432874.63**, cao hơn một chút so với mô hình giá mua. Như biểu đồ minh họa, sự khác biệt giữa giá thực tế và đường dự đoán ngày càng tăng về cuối chuỗi, đặc biệt khi giá bán thực tế vượt mốc 85 trong khi mô hình chỉ dự đoán ở mức khoảng 60. Điều này dẫn đến sai số tổng thể lớn, cho thấy mô hình chưa theo kịp các biến động phi tuyến mạnh mẽ trong dữ liệu.
* **Tổng kết:** Cả hai mô hình đều thể hiện khả năng nhận diện xu hướng tuyến tính của chuỗi thời gian, nhưng chưa phản ánh tốt các biến động lớn và đột ngột. Giá trị MAD ở mức khoảng 7.3–7.5 và SSE trên 400,000

## So sánh và lựa chọn mô hình tốt nhất



Bảng 6: So sánh đánh giá 2 mô hình

* Kết quả so sánh cho thấy mô hình **ARIMA** vượt trội rõ rệt so với **Linear Regression (LR)** trong việc dự báo cả hai chuỗi **Buy Price** và **Sell Price**.
  + Với **MAD** (Mean Absolute Deviation), ARIMA chỉ dao động quanh **0.13**, trong khi LR ghi nhận sai số trên **7**, chênh lệch gần **60 lần**.
  + Về **SSE** (Sum of Squared Errors), ARIMA chỉ khoảng **450–470**, thấp hơn hàng nghìn lần so với LR (trên **400,000**).(7)
* Sự khác biệt này cho thấy LR không phù hợp trong bối cảnh dữ liệu chuỗi thời gian biến động phức tạp như giá vàng. Ngược lại, ARIMA tận dụng tốt thông tin thời gian và độ trễ, giúp giảm thiểu sai số dự báo. Vì vậy, ARIMA là lựa chọn phù hợp và hiệu quả hơn rõ rệt trong phân tích và dự báo giá vàng ngắn hạn.

## Dự báo kết quả

|  |  |
| --- | --- |
|  |  |

Hình 25: Kết quả dự báo

* Kết quả dự báo 10 ngày tới cho thấy cả giá mua và giá bán đều giữ mức ổn định:
  + **Giá Mua** dao động quanh 82.8
  + **Giá Bán** dao động cố định ở 85.3
* Dự báo này phản ánh mô hình ARIMA không phát hiện xu hướng biến động rõ rệt trong ngắn hạn, đồng thời cho thấy thị trường đang ở trạng thái tương đối cân bằng sau khi loại bỏ yếu tố không ổn định nhờ sai phân. Đây là đặc điểm thường gặp trong chuỗi thời gian ngẫu nhiên khi không có cú sốc hoặc yếu tố chu kỳ ảnh hưởng.

# CHƯƠNG 4: KẾT LUẬN

Trong nghiên cứu này, nhóm chúng em  đã áp dụng hai mô hình dự báo phổ biến trong phân tích chuỗi thời gian, đó là ARIMA (Autoregressive Integrated Moving Average) và Hồi quy tuyến tính (Linear Regression - LR) để dự báo giá vàng tại thị trường Việt Nam. Qua quá trình xây dựng, huấn luyện và đánh giá các mô hình, kết quả cho thấy mô hình ARIMA đạt độ chính xác cao hơn so với hồi quy tuyến tính, thể hiện qua các chỉ số lỗi như MAD, SSE. Những kết quả này không  những chỉ ra sự vượt trội của ARIMA trong việc xử lý và dự báo chuỗi thời gian có tính biến động mạnh, mà còn khẳng định khả năng mô hình này trong việc phản ánh các yếu tố như xu hướng, mùa vụ và các biến động ngắn hạn của giá vàng  điều mà mô hình hồi quy tuyến tính khó có thể bắt kịp.

Mô hình ARIMA đã chứng minh hiệu quả trong dự báo giá vàng, đặc biệt trong bối cảnh thị trường tài chính luôn thay đổi và chịu ảnh hưởng từ nhiều yếu tố kinh tế, chính trị, và xã hội. Điều này làm nổi bật tầm quan trọng của việc lựa chọn phương pháp phân tích phù hợp trong các tình huống thực tế, đặc biệt đối với các lĩnh vực như tài chính và đầu tư. Nhóm chúng em  cũng chỉ ra rằng việc áp dụng mô hình dự báo chính xác có thể hỗ trợ các nhà đầu tư, các tổ chức tài chính, cũng như các cơ quan quản lý trong việc đưa ra quyết định đầu tư, điều chỉnh chiến lược kinh doanh và giảm thiểu rủi ro tài chính.

Từ góc độ học thuật, nghiên cứu này không chỉ cung cấp thêm bằng chứng thực tiễn về ứng dụng phương pháp phân tích chuỗi thời gian trong lĩnh vực tài chính, mà còn tạo ra cơ hội để phát triển thêm các mô hình dự báo tiên tiến hơn. Thông qua việc ứng dụng các công cụ và kỹ thuật phân tích dữ liệu trong Python, nghiên cứu này không chỉ khẳng định tầm quan trọng của việc lựa chọn mô hình phù hợp, mà còn giúp người học hiểu rõ hơn về cách sử dụng các phương pháp thống kê và học máy để giải quyết các bài toán thực tế trong lĩnh vực tài chính và đầu tư.

**Hướng phát triển trong tương lai**

Mặc dù nghiên cứu này đã đạt được những kết quả khả quan với hai mô hình ARIMA và Hồi quy tuyến tính, nhưng còn rất nhiều cơ hội để cải thiện và mở rộng nghiên cứu trong tương lai. Dưới đây là một số hướng phát triển mà nhóm nghiên cứu sẽ tiếp tục khai thác:

* **Áp dụng các mô hình học máy tiên tiến**: Nhóm có thể thử nghiệm với các mô hình học máy phức tạp hơn như LSTM (Long Short-Term Memory) hoặc các mạng nơ-ron hồi tiếp (Recurrent Neural Networks - RNN), các mô hình này đã được chứng minh là rất hiệu quả trong việc xử lý dữ liệu chuỗi thời gian, đặc biệt là những chuỗi dữ liệu có tính biến động mạnh và phức tạp như giá vàng.
* **Kết hợp nhiều mô hình**: Một hướng phát triển tiềm năng khác là kết hợp các mô hình khác nhau (hybrid models). Ví dụ, có thể kết hợp ARIMA với các mô hình học sâu (deep learning) để tận dụng được những ưu điểm của cả hai phương pháp, giúp cải thiện độ chính xác và khả năng dự báo trong các tình huống thị trường phức tạp.
* **Kiểm tra và đánh giá mô hình trên các tập dữ liệu dài hạn hơn**: Để cải thiện độ chính xác, nhóm nghiên cứu có thể thử nghiệm các mô hình trên các tập dữ liệu dài hạn hơn và đa dạng hơn, từ đó có thể đánh giá được sự ổn định của mô hình khi áp dụng vào các chu kỳ kinh tế khác nhau.

**Ứng dụng vào thực tế**: Mô hình dự báo giá vàng ARIMA không chỉ giúp nhà đầu tư nhận diện xu hướng mà còn cung cấp đầy đủ thông tin để quyết định thời điểm mối giao dịch mua và bán. Điều này đặc biệt quan trọng đối với thị trường có biến động cao như thị trường vàng.

# TÀI LIỆU THAM KHẢO

1. *Time Series Analysis.* Được truy lục từ Tomorrow Marketers: https://blog.tomorrowmarketers.org/time-series-analysis/

2. *Chuỗi thời gian là gì? (What is time series?).* Được truy lục từ Viblo: https://viblo.asia/p/chuoi-thoi-gian-la-gi-what-is-time-series-pgjLNd6w432

3. *Mô hình ARIMA trong time series.* Được truy lục từ Pham Dinh Khanh: https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html

4. *Joshua Noble.* Được truy lục từ What are ARIMA model?: https://www.ibm.com/think/topics/arima-model

5. *Autoregressive Integrated Moving Average (ARIMA) Prediction Model.* Được truy lục từ Adam Hayes: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

6. *How to Create an ARIMA Model for Time Series Forecasting in Python.* Được truy lục từ Jason Brownlee: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

7. *ARIMA for Time Series Forecasting: A Complete Guide.* Được truy lục từ Zaina Saadeddin: https://www.datacamp.com/tutorial/arima

# VIDEO

Link youtube: [**https://youtu.be/tTv-NxIeLGc**](https://youtu.be/tTv-NxIeLGc)