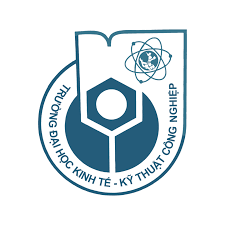
**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC KINH TẾ - KỸ THUẬT CÔNG NGHIỆP**

**KHOA KHOA HỌC ỨNG DỤNG**

**-----🕮-----**

****

**BÁO CÁO ĐỒ ÁN 1**

(NHÓM 1.3)

**Thăm dò và trực quan hóa dữ liệu của 3 mã cổ phiếu nhóm công nghệ thông tin FPT, VGI và VTC.**

Nhóm sinh viên thực hiện: Mã sinh viên:

Nguyễn Bá Duy Anh 22174600072

Nguyễn Văn Duy 22174600011

Ninh Hữu Duy 22174600042

Đồng Khắc Kiên 22174600074

Lê Văn Mạnh 22174600044

*(GV hướng dẫn: ThS. Trần Chí Lê)*

**Hà Nội – 2024**

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC KINH TẾ - KỸ THUẬT CÔNG NGHIỆP**

**KHOA KHOA HỌC ỨNG DỤNG**

**-----🕮-----**

**BÁO CÁO ĐỒ ÁN 1**

(NHÓM 1.3)

**Thăm dò và trực quan hóa dữ liệu của 3 mã cổ phiếu nhóm công nghệ thông tin FPT, VGI và VTC.**

Nhóm sinh viên thực hiện: Mã sinh viên:

Nguyễn Bá Duy Anh 22174600072

Nguyễn Văn Duy 22174600011

Ninh Hữu Duy 22174600042

Đồng Khắc Kiên 22174600074

Lê Văn Mạnh 22174600044

*(GV hướng dẫn: ThS. Trần Chí Lê)*

**Hà Nội – 2024**

# LỜI CẢM ƠN

Để hoàn thành báo cáo đồ án 1 này trước hết nhóm chúng em xin gửi lời cảm ơn chân thành đến quý Thầy/Cô trong khoa Khoa học ứng dụng trường Đại học Kinh Tế - Kĩ Thuật Công Nghiệp đã trang bị cho các thành viên trong nhóm những kiến thức nền tảng và tạo điều kiện thuận lợi cho nhóm chúng em trong suốt quá trình học tập và nghiên cứu tại trường. Đặc biệt, nhóm chúng em xin gửi đến thầy Trần Chí Lê - giảng viên bộ môn Đồ án 1, người đã tận tình hướng dẫn, giúp đỡ các thành viên trong nhóm hoàn thành báo cáo đồ án 1 này lời cảm ơn sâu sắc nhất. Sự chỉ bảo tận tình và những góp ý quý báu của thầy đã giúp các thành viên trong nhóm hoàn thiện đồ án 1 và mở rộng kiến thức của mình. Giúp chúng em hiểu rõ hơn về các kiến thức chuyên môn, cũng như cách thức tiếp cận và giải quyết vấn đề trong đề tài của mình.

Cuối cùng, các thành viên trong nhóm chúng em xin gửi lời chúc sức khỏe và thành công đến Thầy/Cô trong khoa và mong rằng sẽ tiếp tục nhận được sự hướng dẫn và hỗ trợ từ Thầy/Cô trong những học kỳ tiếp theo.

Một lần nữa, chúng em xin chân thành cảm ơn !

Hà Nội, ngày 9 tháng 12 năm 2024

Nhóm thực hiện đồ án

# TÊN ĐỒ ÁN VÀ CÁC YÊU CẦU CHI TIẾT

1. **THÔNG TIN THÀNH VIÊN NHÓM**

|  |  |  |
| --- | --- | --- |
| Họ và tên thành viên nhóm | Mã sinh viên | Điện thoại |
| Nguyễn Bá Duy Anh | 22174600072 | 0976343384 |
| Nguyễn Văn Duy | 22174600011 | 0822990072 |
| Ninh Hữu Duy | 22174600042 | 0914540465 |
| Đồng Khắc Kiên | 22174600074 | 0369579338 |
| Lê Văn Mạnh | 22174600044 |  |

1. **THÔNG TIN ĐỒ ÁN**

Tên đồ án: Thăm dò và trực quan hóa dữ liệu của 3 mã cổ phiếu nhóm Công nghệ thông tin FPT, VGI và VTC.

Mục đích của đồ án: Nhóm sẽ đi sâu vào kiểm tra và tìm hiểu 3 mã cổ phiếu trên để xác định nhiều điều cần chú ý từ đó so sánh hiệu suất đầu tư, phân tích rủi ro cơ hội và hỗ trợ đưa ra quyết định với một cách truyền đạt thông tin dễ hiểu nhất.

Đồ án môn học được thực hiện tại: Bộ Môn Toán & Khoa Học Dữ Liệu, Khoa Khoa Học Ứng Dụng, Trường Đại Học Kinh Tế - Kỹ Thuật Công Nghiệp.

Thời gian thực hiện: Từ ngày 22/10/2024 đến 23/12/2024.

1. **Nhiệm vụ của đề tài**

* Thu thập dữ liệu.
* Thống kê mô tả.
* Phân tích mối liên hệ giữa các biến.
* Trực quan hóa dữ liệu.
* Một số chủ đề nâng cao: Xây dựng mô hình dự đoán.
* Viết báo cáo.

# MỤC LỤC

[LỜI CẢM ƠN 2](#_Toc185708226)

[TÊN ĐỒ ÁN VÀ CÁC YÊU CẦU CHI TIẾT 3](#_Toc185708227)

[MỤC LỤC 4](#_Toc185708228)

[MỤC LỤC HÌNH VẼ 7](#_Toc185708229)

[MỤC LỤC BẢNG 9](#_Toc185708230)

[LỜI GIỚI THIỆU 10](#_Toc185708231)

[CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN CÁC KỸ THUẬT THĂM DÒ VÀ TRỰC QUAN HÓA DỮ LIỆU TRONG CHỨNG KHOÁN 11](#_Toc185708232)

[1.1 GIỚI THIỆU 11](#_Toc185708233)

[1.1.1. Tầm quan trọng của phân tích dữ liệu giá chứng khoán: 11](#_Toc185708234)

[1.1.2. Mục tiêu của việc phân tích dữ liệu giá chứng khoán: 11](#_Toc185708235)

[1.2 CÁC KỸ THUẬT PHỔ BIẾN 12](#_Toc185708236)

[1.2.1. Thống kê mô tả (Descriptive Statistics): 12](#_Toc185708237)

[1.2.2. Biểu đồ chuỗi thời gian: 13](#_Toc185708238)

[1.2.3. Phân phối giá cổ phiếu (Price Distribution): 13](#_Toc185708239)

[1.2.4. Phân tích khối lượng giao dịch (volume analysis): 14](#_Toc185708240)

[1.2.5. Phân tích tương quan: 14](#_Toc185708241)

[1.2.6. Phân tích xu hướng (Trend Analysis): 16](#_Toc185708242)

[1.2.7. Phát hiện bất thường (Anomaly Detection): 16](#_Toc185708243)

[1.2.8. Phân cụm dữ liệu: 16](#_Toc185708244)

[1.3 TRỰC QUAN HÓA DỮ LIỆU TRONG LĨNH VỰC CHỨNG KHOÁN 16](#_Toc185708245)

[1.3.1. Vai trò của trực quan hóa dữ liệu trong chứng khoán: 16](#_Toc185708246)

[1.3.2. Kết Hợp EDA Với Trực Quan Hóa Dữ Liệu Trong Chứng Khoán 17](#_Toc185708247)

[1.3.3. Ví dụ minh họa với dữ liệu chứng khoán 17](#_Toc185708248)

[CHƯƠNG 2: THU THẬP DỮ LIỆU 19](#_Toc185708249)

[2.1. NGUỒN DỮ LIỆU 19](#_Toc185708250)

[2.2. KỸ THUẬT DỮ LIỆU 19](#_Toc185708251)

[2.3. LÀM SẠCH DỮ LIỆU 21](#_Toc185708252)

[CHƯƠNG 3: THỐNG KÊ MÔ TẢ 23](#_Toc185708253)

[3.1. CÁC ĐẠI LƯỢNG, CHỈ SỐ THỐNG KÊ CƠ BẢN 23](#_Toc185708254)

[3.2. BIỂU ĐỒ TÓM TẮT DỮ LIỆU 27](#_Toc185708255)

[3.2.1. Biểu đồ thể hiện giá đóng cửa FPT, VGI, VTC 27](#_Toc185708256)

[3.3. SỬ DỤNG ƯỚC LƯỢNG VÀ KIỂM ĐỊNH ĐỂ PHÂN TÍCH MỐI LIÊN HỆ GIỮA GIÁ CHỨNG KHOÁN VÀ CÁC YẾU TỐ KHÁC 28](#_Toc185708257)

[3.3.1. Kiểm tra dữ liệu có theo phân phối chuẩn không để so sánh dữ liệu. 28](#_Toc185708258)

[3.3.2.Sử dụng các thống kê không tham số. 32](#_Toc185708259)

[3.3.3. Biểu đồ trực quan hóa. 34](#_Toc185708260)

[3.3.4**.** Phân tích sâu sự khác biệt giữa các cổ phiếu. 37](#_Toc185708261)

[3.3.5. Phân tích tương quan. 39](#_Toc185708262)

[3.3.6. Kiểm định giả thuyết về tương quan (Spearman). 42](#_Toc185708263)

[3.3.7.Phân tích mô hình biến động giá. 45](#_Toc185708264)

[3.3.8.Hồi quy tuyến tính. 50](#_Toc185708265)

[CHƯƠNG 4: PHÂN TÍCH MỐI QUAN HỆ GIỮA CÁC BIẾN 53](#_Toc185708266)

[4.1. BIỂU ĐỒ 53](#_Toc185708267)

[4.2. HỆ SỐ TƯƠNG QUAN VÀ ĐỘ ĐỒNG BIẾN 62](#_Toc185708268)

[CHƯƠNG 5: TRỰC QUAN HÓA DỮ LIỆU 67](#_Toc185708269)

[5.1. GIỚI THIỆU VỀ TRỰC QUAN HÓA DỮ LIỆU 67](#_Toc185708270)

[5.1.1. Trực quan hóa dữ liệu là gì ? 67](#_Toc185708271)

[5.1.2.Mục đích của trực quan hóa dữ liệu 67](#_Toc185708272)

[5.1.3. Các loại biểu đồ trực quan hóa: 68](#_Toc185708273)

[5.2. Trực quan hóa dữ liệu 70](#_Toc185708274)

[CHƯƠNG 6: CHỦ ĐỀ NÂNG CAO 77](#_Toc185708275)

[6.1. GIỚI THIỆU 77](#_Toc185708276)

[6.2. CÁC THUẬT TOÁN PHÂN TÍCH CỤM 77](#_Toc185708277)

[6.3. ƯU VÀ NHƯỢC ĐIỂM CỦA PHÂN TÍCH CỤM 77](#_Toc185708278)

[6.4. ỨNG DỤNG CỦA PHÂN TÍCH CỤM VÀO BỘ DỮ LIỆU ĐỂ PHÂN LOẠI CÁC CỔ PHIẾU 78](#_Toc185708279)

[6.4.1. Tiền xử lý dữ liệu: 79](#_Toc185708280)

[6.4.2. Phân cụm K-Means: 79](#_Toc185708281)

[6.4.3. Trực quan hóa: 80](#_Toc185708282)

[6.4.4 Phân tích từng cụm: 81](#_Toc185708283)

[6.4.5. Hạn chế 82](#_Toc185708284)

[6.4.6. Kết luận và Kiến nghị 82](#_Toc185708285)

[CHƯƠNG 7: TỔNG KẾT BÁO CÁO 83](#_Toc185708286)

[7.1. MỤC TIÊU VÀ PHẠM VI CỦA ĐỒ ÁN 83](#_Toc185708287)

[7.2. QUY TRÌNH VÀ PHƯƠNG PHÁP SỬ DỤNG 83](#_Toc185708288)

[7.3. KẾT QUẢ CHÍNH VÀ PHÁT HIỆN 83](#_Toc185708289)

[7.4. Ý NGHĨA VÀ ỨNG DỤNG KẾT QUẢ 84](#_Toc185708290)

[7.5. HẠN CHẾ VÀ ĐỀ XUẤT HƯỚNG PHÁT TRIỂN 84](#_Toc185708291)

[7.6. KẾT LUẬN CUỐI CÙNG 84](#_Toc185708292)

[TÀI LIỆU THAM KHẢO 86](#_Toc185708293)

# MỤC LỤC HÌNH VẼ

[Hình 1.1: Biểu đồ phân phối giá 14](#_Toc185708528)

[Hình 1.2: Ví dụ ma trận tương quan giữa các cổ phiếu 15](#_Toc185708529)

[Hình 1.3: Ví dụ về xu hướng giá cổ phiếu 18](#_Toc185708530)

[Hình 3.1: Giá đóng cửa của FPT, VGI và VTC 27](#_Toc185708531)

[Hình 3.2: Biểu đồ tần suất và Đường mật độ hạt nhân của mã cổ phiếu FPT 29](#_Toc185708532)

[Hình 3.3: Biểu đồ định lượng - định lượng cho biến close của mã cổ phiếu FPT 29](#_Toc185708533)

[Hình 3.4: Biểu đồ tần suất và Đường mật độ hạt nhân của mã cổ phiếu VGI 30](#_Toc185708534)

[Hình 3.5: Biểu đồ định lượng - định lượng cho biến close của mã cổ phiếu VGI 30](#_Toc185708535)

[Hình 3.6: Biểu đồ tần suất và Đường mật độ hạt nhân của mã cổ phiếu VTC 31](#_Toc185708536)

[Hình 3.7: Biểu đồ định lượng - định lượng cho biến close của mã cổ phiếu VTC 31](#_Toc185708537)

[Hình3.8: Biểu đồ boxplot so sánh giá đóng cửa của các cổ phiếu 35](#_Toc185708538)

[Hình 3.9: Biểu đồ violin thể hiện Phân phối giá đóng cửa của các cổ phiếu 39](#_Toc185708539)

[Hình 3.10: Biểu đồ Heatmap thể hiện tương quan giữa giá đóng cửa của các cổ phiếu 40](#_Toc185708540)

[Hình 3.11: Biểu đồ phân tán giá đóng cửa của 3 mã cổ phiếu FPT, VGI, VTC 42](#_Toc185708541)

[Hình 3.12: Biểu đồ phân tích xu hướng giá cổ phiếu FPT với SMA 7 ngày 45](#_Toc185708542)

[Hình 3.13: Biểu đồ phân tích mô hình biến động giá 49](#_Toc185708543)

[Hình 3.14: Biểu đồ Hồi quy tuyến tính đơn biến. 51](#_Toc185708544)

[Hình 4.1: Biểu đồ phân tán giá đóng cửa và khối lượng giao dịch 54](#_Toc185708545)

[Hình4.2: Biểu đồ phân tán giữa giá đóng cửa và khối lượng giao dịch 56](#_Toc185708546)

[Hình 4.3: Biểu đồ phân tán giữa giá mở cửa và giá đóng cửa 57](#_Toc185708547)

[Hình 4.4: Biểu đồ ma trận tương quan 59](#_Toc185708548)

[Hình 4.5: Biểu đồ tính ma trận tương quan 61](#_Toc185708549)

[Hình 4.6: Biểu đồ ma trận hệ số tương quan 64](#_Toc185708550)

[Hình 4.7: Biểu đồ ma trận độ đồng biến 65](#_Toc185708551)

[Hình 5.1. biểu đồ ma trận phân tán 70](#_Toc185708552)

[Hình 5.2. biểu đồ phân phối giá trị ngân hàng của FPT 71](#_Toc185708553)

[Hình 5.3. biểu đồ giá đóng cửa của các cổ phiếu FPT, VGI và VTC 71](#_Toc185708554)

[Hình 5.4: Biểu đồ giá đóng cửa và đường trung bình động MA30 của cổ phiếu FPT 72](#_Toc185708555)

[Hình 5.5: Biểu đồ ma trận tương quan 72](#_Toc185708556)

[Hình 5.6: Biểu đồ ma trận tương quan với dendrogram 73](#_Toc185708557)

[Hình 5.7 biểu đồ giá đóng cửa 73](#_Toc185708558)

[Hình 5.8. Biểu đồ giá đóng cửa và các đường trung bình động (MA) 74](#_Toc185708559)

[Hình 5.9: Biểu đồ giá đóng cửa và dải Bollinger Bands 74](#_Toc185708560)

[Hình 5.10: Biểu đồ Nến Nhật Mã Cổ phiếu FPT 75](#_Toc185708561)

[Hình 5.11. Biểu đồ Nến Nhật Mã Cổ phiếu VGI 75](#_Toc185708562)

[Hình 5.12. Biểu đồ Nến Nhật Mã Cổ phiếu VGI 76](#_Toc185708563)

[Hình 5.13: Biểu đồ Nến Nhật và Dải Bollinger Bands của Mã Cổ phiếu FPT 76](#_Toc185708564)

[Hình 6.1 Biểu đồ cặp các đặc trưng theo cụm 80](#_Toc185708565)

[Hình 6.2 Biểu đồ phân bố theo cụm 81](#_Toc185708566)

# MỤC LỤC BẢNG

[Bảng 1.1: Kết quả ví dụ 13](#_Toc185708679)

[Bảng 2.1: Mã cổ phiếu và thông tin cổ phiếu 21](#_Toc185708680)

[Bảng 2.3: Số lượng giá trị thiếu 22](#_Toc185708681)

[Bảng 3.1: Các đại lượng thống kê mã FPT 23](#_Toc185708682)

[Bảng 3.2: Các đại lượng thống kê mã VGI 24](#_Toc185708683)

[Bảng 3.3: Các đại lượng thống kê mã VTC 26](#_Toc185708684)

[Bảng 3.4: DataFrame của 3 mã cổ phiếu FPT, VGI, VTC. 35](#_Toc185708685)

[Bảng 3.5: Mô hình Garch 47](#_Toc185708686)

[Bảng 6.1: Bảng dữ liệu chứng khoán 79](#_Toc185708687)

# LỜI GIỚI THIỆU

Trực quan hóa dữ liệu là một phương thức biểu diễn trực quan của dữ liệu. Trực quan hóa dữ liệu có thể cho biết tổng quan về dữ liệu nghiên cứu, các xu hướng, các giá trị ngoại lệ và mối tương quan trong dữ liệu có thể cho biết tổng quan về dữ liệu. Trực quan hóa dữ liệu giúp những người phân tích dễ dàng hiểu được ý nghĩa của dữ liệu, từ đó đưa ra các quyết định phù hợp trong lĩnh vực mà mình đang phân tích, quản lý. Phương thức phổ biến nhất làm cho dữ liệu trở nên trực quan là thông qua phân tích thống kê và phân tích biểu đồ.

Trong bối cảnh toàn cầu hóa và sự bùng nổ của công nghệ, thị trường chứng khoán trở thành một chỉ dấu quan trọng phản ánh sức khỏe kinh tế và tiềm năng phát triển của các doanh nghiệp. Đặc biệt, nhóm ngành công nghệ thông tin đang là tâm điểm của sự chú ý, khi các doanh nghiệp trong lĩnh vực này không chỉ dẫn đầu về đổi mới sáng tạo mà còn có khả năng mở rộng quy mô và tác động mạnh mẽ đến nhiều lĩnh vực khác. Ba mã cổ phiếu tiêu biểu của nhóm ngành này tại Việt Nam – FPT, VGI, và VTC – đại diện cho những doanh nghiệp hàng đầu, đóng vai trò quan trọng trong hệ sinh thái kinh tế số của đất nước.

Việc phân tích dữ liệu tài chính của các mã cổ phiếu này không chỉ giúp nhà đầu tư hiểu rõ hơn về sự biến động của giá trị thị trường mà còn cung cấp những góc nhìn sâu sắc về xu hướng phát triển, hiệu quả kinh doanh và tiềm năng dài hạn. Trong đó, các kỹ thuật thăm dò và trực quan hóa dữ liệu trở thành công cụ không thể thiếu, hỗ trợ việc khám phá và trình bày dữ liệu một cách trực quan, dễ hiểu. Các biểu đồ, hình ảnh được tạo ra từ dữ liệu không chỉ giúp nhận diện các mẫu hình, xu hướng quan trọng mà còn làm sáng tỏ những yếu tố tiềm ẩn mà dữ liệu thô khó có thể thể hiện.

Bài báo cáo này tập trung vào việc áp dụng các kỹ thuật thăm dò và trực quan hóa dữ liệu để phân tích sâu ba mã cổ phiếu tiêu biểu trong ngành công nghệ thông tin là FPT, VGI, và VTC. Sử dụng các công cụ mạnh mẽ của Python như Pandas, Matplotlib, Seaborn, và vnstock, chúng tôi sẽ khám phá đặc điểm dữ liệu, phân tích biến động giá cổ phiếu, và trình bày các kết quả phân tích một cách trực quan. Mục tiêu của báo cáo không chỉ là cung cấp thông tin chi tiết, đáng tin cậy mà còn hỗ trợ nhà đầu tư và các bên liên quan trong việc đưa ra các quyết định chiến lược, từ đó tối ưu hóa lợi nhuận và giảm thiểu rủi ro trên thị trường tài chính đầy biến động.

# CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN CÁC KỸ THUẬT THĂM DÒ VÀ TRỰC QUAN HÓA DỮ LIỆU TRONG CHỨNG KHOÁN

## 1.1 GIỚI THIỆU

### 1.1.1. Tầm quan trọng của phân tích dữ liệu giá chứng khoán:

**Xác định xu hướng thị trường**: Phân tích dữ liệu giá giúp nhà đầu tư hiểu và dự đoán xu hướng tăng hay giảm của cổ phiếu hoặc thị trường chứng khoán. Việc này rất quan trọng trong việc xác định điểm vào và ra hợp lý.

**Đánh giá sức khỏe tài chính của công ty**: Dữ liệu giá cổ phiếu phản ánh tình hình tài chính của công ty, hiệu quả hoạt động, và niềm tin của nhà đầu tư. Phân tích sẽ giúp nhà đầu tư đánh giá liệu công ty có tiềm năng tăng trưởng hay không.

**Tối ưu hóa chiến lược đầu tư**: Phân tích giá chứng khoán giúp nhà đầu tư tối ưu hóa danh mục đầu tư, quyết định cổ phiếu nào cần nắm giữ, mua thêm, hoặc bán ra, từ đó tối ưu hóa lợi nhuận và giảm thiểu rủi ro.

**Phát hiện các biến động ngắn hạn**: Bên cạnh việc xác định xu hướng dài hạn, phân tích giá cổ phiếu còn giúp phát hiện các biến động ngắn hạn, như các đợt sóng ngắn hoặc những cơ hội đầu tư nhanh.

**Xác định điểm mua và bán hợp lý**: Phân tích kỹ thuật dựa trên dữ liệu giá (chẳng hạn như các chỉ số kỹ thuật, mô hình biểu đồ) giúp nhà đầu tư tìm điểm vào và ra tối ưu, từ đó cải thiện hiệu quả giao dịch.

**Quản lý rủi ro**: Qua phân tích dữ liệu giá, nhà đầu tư có thể dự báo các rủi ro tiềm ẩn và quản lý rủi ro tốt hơn. Việc theo dõi diễn biến giá cổ phiếu cũng giúp nhà đầu tư điều chỉnh chiến lược nhanh chóng trước khi thị trường chuyển biến xấu.

### 1.1.2. Mục tiêu của việc phân tích dữ liệu giá chứng khoán:

**Xác định xu hướng thị trường**: Phân tích dữ liệu giá giúp nhà đầu tư hiểu và dự đoán xu hướng tăng hay giảm của cổ phiếu hoặc thị trường chứng khoán. Việc này rất quan trọng trong việc xác định điểm vào và ra hợp lý.

**Đánh giá sức khỏe tài chính của công ty**: Dữ liệu giá cổ phiếu phản ánh tình hình tài chính của công ty, hiệu quả hoạt động, và niềm tin của nhà đầu tư. Phân tích sẽ giúp nhà đầu tư đánh giá liệu công ty có tiềm năng tăng trưởng hay không.

**Tối ưu hóa chiến lược đầu tư**: Phân tích giá chứng khoán giúp nhà đầu tư tối ưu hóa danh mục đầu tư, quyết định cổ phiếu nào cần nắm giữ, mua thêm, hoặc bán ra, từ đó tối ưu hóa lợi nhuận và giảm thiểu rủi ro.

**Phát hiện các biến động ngắn hạn**: Bên cạnh việc xác định xu hướng dài hạn, phân tích giá cổ phiếu còn giúp phát hiện các biến động ngắn hạn, như các đợt sóng ngắn hoặc những cơ hội đầu tư nhanh.

**Xác định điểm mua và bán hợp lý**: Phân tích kỹ thuật dựa trên dữ liệu giá (chẳng hạn như các chỉ số kỹ thuật, mô hình biểu đồ) giúp nhà đầu tư tìm điểm vào và ra tối ưu, từ đó cải thiện hiệu quả giao dịch.

**Quản lý rủi ro**: Qua phân tích dữ liệu giá, nhà đầu tư có thể dự báo các rủi ro tiềm ẩn và quản lý rủi ro tốt hơn. Việc theo dõi diễn biến giá cổ phiếu cũng giúp nhà đầu tư điều chỉnh chiến lược nhanh chóng trước khi thị trường chuyển biến xấu.

Mục tiêu và tầm quan trọng của đề tài này là áp dụng các phương pháp khoa học dữ liệu để tối ưu hóa danh mục đầu tư và dự đoán giá thị trường chứng khoán, nhằm cung cấp các giải pháp đầu tư hiệu quả và thông minh. Cụ thể, nghiên cứu nhằm giảm thiểu rủi ro và tối đa hóa lợi nhuận thông qua các mô hình phân tích và dự đoán chính xác.

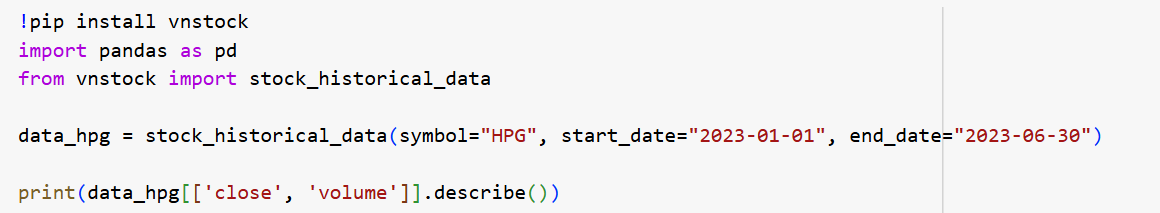
## 1.2 CÁC KỸ THUẬT PHỔ BIẾN

### 1.2.1. Thống kê mô tả (Descriptive Statistics):

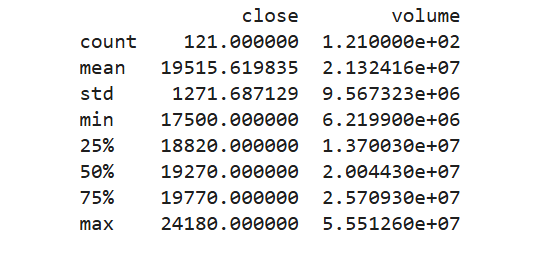
Thống kê mô tả cung cấp một cái nhìn tổng quan về dữ liệu, bao gồm:

* + - Các chỉ số trung tâm: Giá trị trung bình, trung vị, mode.
    - Sự phân tán: Độ lệch chuẩn, phương sai.
    - Khoảng giá trị: Giá cao nhất, giá thấp nhất, khoảng giá trị.

Ví dụ 1.1:



Kết quả hiển thị:



Bảng 1.1: Kết quả ví dụ

Dựa trên kết quả thống kê mô tả, chúng ta có thể thấy rằng giá đóng cửa trung bình của cổ phiếu đó trong giai đoạn phân tích là XX (đơn vị), với độ biến động là YY. Thông tin này cung cấp cái nhìn tổng quan về phạm vi dao động giá và khối lượng giao dịch, giúp nhà đầu tư đánh giá mức độ rủi ro của cổ phiếu.

### 1.2.2. Biểu đồ chuỗi thời gian:

Biểu đồ chuỗi thời gian là công cụ quan trọng để theo dõi biến động giá cổ phiếu trong một giai đoạn cụ thể, từ đó nhận diện các xu hướng tăng hoặc giảm.

* + - Vẽ biểu đồ giá cổ phiếu qua thời gian để phân tích xu hướng.
    - Kết hợp phân tích các chỉ số như đường trung bình trượt (MA - Moving Average) để xác định xu hướng dài hạn.

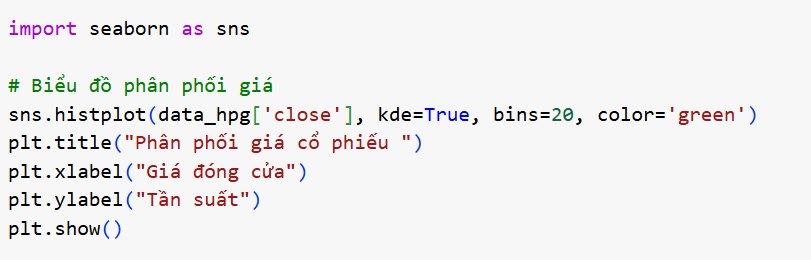
### 1.2.3. Phân phối giá cổ phiếu (Price Distribution):

Biểu đồ phân phối giúp người dùng hiểu cách giá cổ phiếu tập trung ở các mức cụ thể và phát hiện sự bất đối xứng hoặc cực đoan trong dữ liệu.

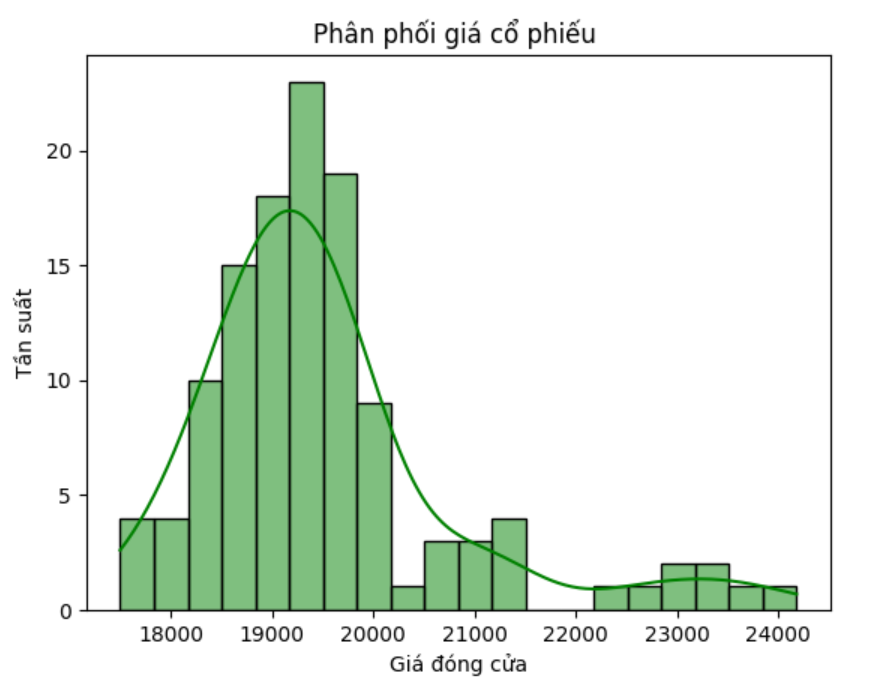
Phân tích phân phối giá giúp hiểu được hành vi của giá cổ phiếu:

* Xác định độ lệch, tính đối xứng.
* Sử dụng biểu đồ Histogram hoặc KDE để hiển thị.

Ví dụ 1.2:



Kết quả hiển thị:



Hình 1.1: Biểu đồ phân phối giá

Biểu đồ phân phối giá cổ phiếu cho thấy phần lớn giá đóng cửa tập trung trong khoảng XX-YY (đơn vị). Đỉnh của phân phối thể hiện mức giá phổ biến nhất, trong khi phần đuôi dài ở bên phải/bên trái có thể gợi ý các phiên giao dịch bất thường.

### 1.2.4. Phân tích khối lượng giao dịch (volume analysis):

Phân tích khối lượng giao dịch cùng với giá đóng cửa giúp nhà đầu tư hiểu rõ hơn về mối liên hệ giữa cung, cầu và sự biến động giá của cổ phiếu.

* Đánh giá mối quan hệ giữa giá cổ phiếu và khối lượng giao dịch.
* Xác định các giai đoạn "volume spike" (giao dịch tăng đột biến).

Khối lượng giao dịch của cổ phiếu tăng đột biến hoặc giảm đột biến tại một số thời điểm trùng với những giai đoạn giá có biến động mạnh. Điều này cho thấy sự quan tâm đặc biệt của nhà đầu tư trong các phiên giao dịch này, có thể liên quan đến tin tức quan trọng hoặc sự kiện thị trường.

### 1.2.5. Phân tích tương quan:

Phân tích tương quan giữa các cổ phiếu giúp xác định mối quan hệ động lực giữa chúng, hỗ trợ trong việc phân bổ tài sản và quản lý rủi ro.

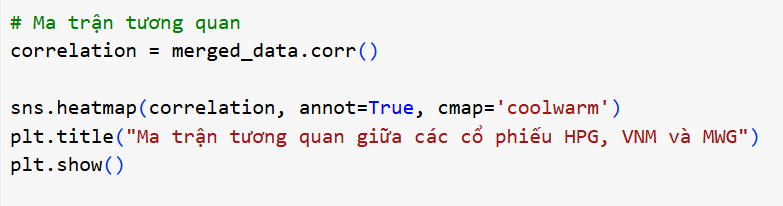
Sử dụng Heatmap để hiển thị trực quan.

Ví dụ 1.3:

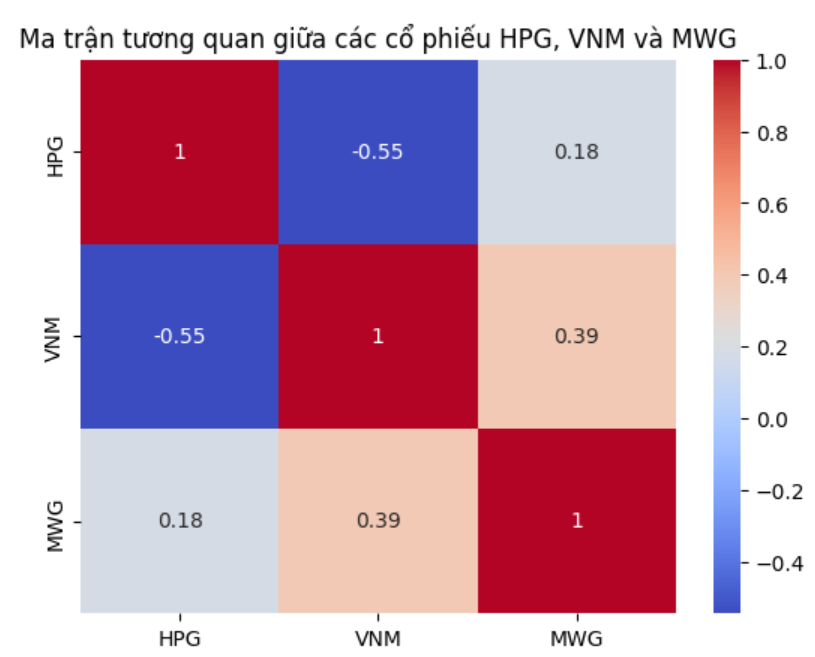
Từ bộ dữ liệu của vnstock lấy 3 mã cổ phiếu bất kì



rồi tạo ra ma trận tương quan giữa các mã:



Kết quả hiển thị:



Hình 1.2: Ví dụ ma trận tương quan giữa các cổ phiếu

* Ma trận tương quan cho thấy cổ phiếu HPG và VNM có mối quan hệ tương quan thấp, trong khi HPG và MWG có mối quan hệ tương quan trung bình. Điều này cho thấy các cổ phiếu thuộc ngành khác nhau thường không bị ảnh hưởng bởi những yếu tố chung, giúp đa dạng hóa danh mục đầu tư.

### 1.2.6. Phân tích xu hướng (Trend Analysis):

* + - Là một kỹ thuật quan trọng trong thăm dò dữ liệu, đặc biệt trong lĩnh vực tài chính và chứng khoán. Mục tiêu của phân tích xu hướng là nhận diện các mô hình và hành vi trong dữ liệu giá cổ phiếu theo thời gian. Kỹ thuật này giúp nhà đầu tư dự báo tiềm năng tương lai của cổ phiếu và đưa ra các quyết định giao dịch.
    - Công cụ phổ biến là sử dụng đường trung bình cộng (Moving Average – MA), một công cụ là mượt dữ liệu giá, giúp loại bỏ nhiễu và làm rõ xu hướng của cổ phiếu.
      * Sử dụng Moving Average (MA), Bollinger Bands để xác định xu hướng và biến động.
      * Phân tích các điểm giao cắt MA để phát hiện tín hiệu mua hoặc bán.

### 1.2.7. Phát hiện bất thường (Anomaly Detection):

Phát hiện bất thường giúp nhận diện các giá trị ngoại lệ trong dữ liệu, hỗ trợ nhà đầu tư trong việc đánh giá các yếu tố gây ra sự kiện bất thường:

* Sử dụng Boxplot để phát hiện các điểm giá bất thường (outliers).
* Phân tích các phiên giao dịch có khối lượng bất thường.

### 1.2.8. Phân cụm dữ liệu:

Phân cụm dữ liệu là kỹ thuật phân nhóm cổ phiếu dựa trên đặc điểm giống nhau, giúp phát hiện các nhóm cổ phiếu có xu hướng giao động tương đồng.

## 1.3 TRỰC QUAN HÓA DỮ LIỆU TRONG LĨNH VỰC CHỨNG KHOÁN

Trực quan hóa dữ liệu trong lĩnh vực chứng khoán không chỉ giúp xử lý lượng thông tin khổng lồ từ các thị trường tài chính mà còn hỗ trợ nhà đầu tư hiểu rõ hơn về xu hướng và mối quan hệ trong dữ liệu. Dưới đây là vai trò và ứng dụng quan trọng.

### 1.3.1. Vai trò của trực quan hóa dữ liệu trong chứng khoán:

Trực quan hóa dữ liệu là quá trình biến dữ liệu thô thành các biểu đồ, hình ảnh giúp người dùng dễ dàng hiểu và phân tích dữ liệu. Nó đóng vai trò quan trọng trong việc:

* Biểu diễn thông tin phức tạp:

Thị trường chứng khoán thường có nhiều biến động, với hàng loạt dữ liệu về giá cổ phiếu, khối lượng giao dịch, chỉ số ngành, và các chỉ báo tài chính khác. Trực quan hóa dữ liệu giúp làm rõ các xu hướng và mẫu hình này thông qua các biểu đồ như biểu đồ đường (line chart) để phân tích xu hướng giá, hoặc biểu đồ hộp (box plot) để so sánh biến động của cổ phiếu theo thời gian.

* So sánh trực tiếp giữa các yếu tố:

Nhà đầu tư cần so sánh hiệu suất của nhiều cổ phiếu hoặc chỉ số trong cùng một khoảng thời gian. Các biểu đồ cột (bar chart) hoặc biểu đồ tròn (pie chart) giúp minh họa rõ ràng tỷ lệ và sự khác biệt giữa các yếu tố.

* Dự đoán và phát hiện mẫu hình:

Với các công cụ như biểu đồ phân tán (scatter plot) hoặc biểu đồ nhiệt (heatmap), nhà phân tích có thể nhận ra các mối quan hệ tiềm ẩn giữa giá cổ phiếu và các yếu tố như khối lượng giao dịch hoặc chỉ báo kinh tế.

### 1.3.2. Kết Hợp EDA Với Trực Quan Hóa Dữ Liệu Trong Chứng Khoán

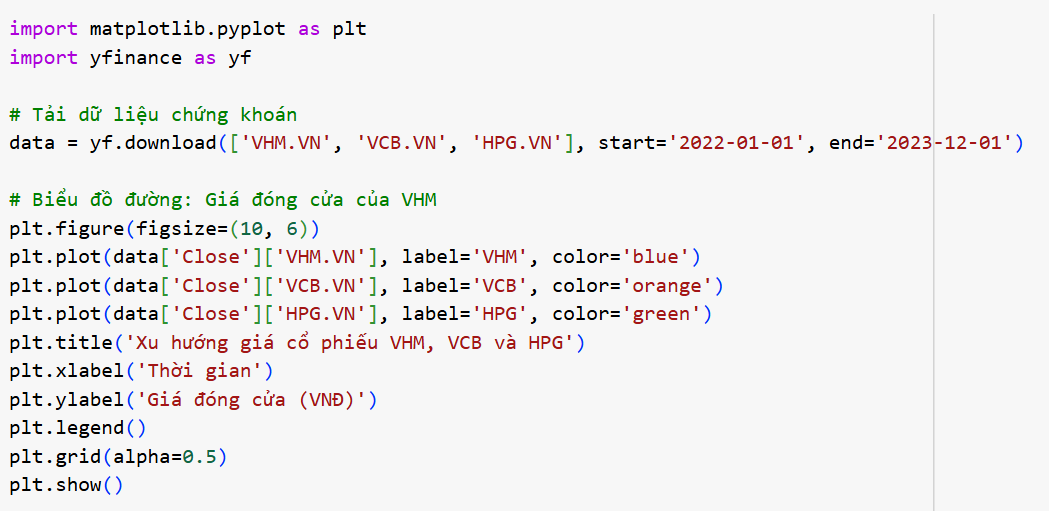
Phương pháp phân tích thăm dò dữ liệu (Exploratory Data Analysis - EDA) trong chứng khoán tập trung vào việc khám phá và hiểu dữ liệu trước khi ra quyết định:

* Kiểm tra dữ liệu ban đầu: Xem xét dữ liệu giá, khối lượng, và chỉ số liên quan để phát hiện các giá trị bị thiếu hoặc ngoại lai.
* Phân tích phân phối dữ liệu: Sử dụng histogram hoặc box plot để hiểu cách giá cổ phiếu phân phối qua thời gian.
* Khám phá mối quan hệ: Xác định tương quan giữa các biến số (ví dụ: giá cổ phiếu và khối lượng giao dịch) bằng biểu đồ phân tán hoặc heatmap.

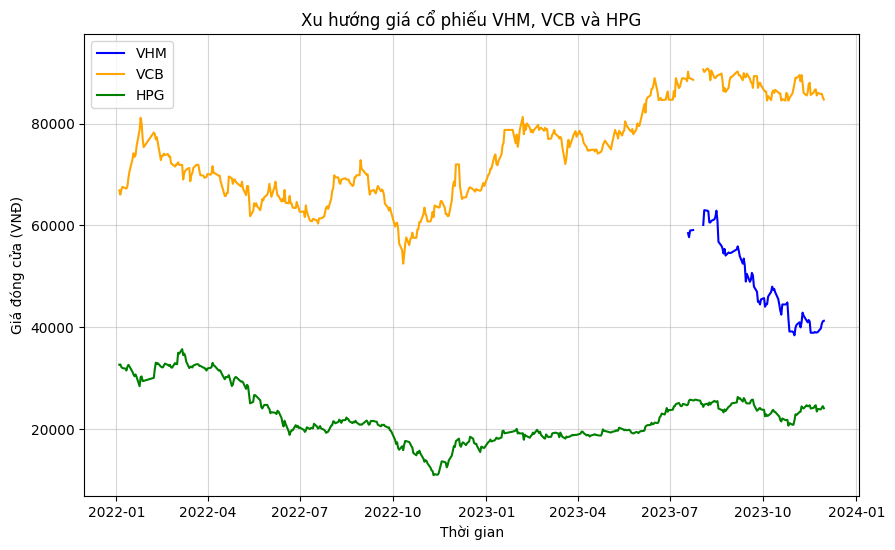
### 1.3.3. Ví dụ minh họa với dữ liệu chứng khoán

Ví dụ 1.4

Lấy dữ liệu cổ phiếu và khối lượng giao dịch:



Kết quả hiển thị:



Hình 1.3: Ví dụ về xu hướng giá cổ phiếu

Từ biểu đồ, nhà phân tích có thể quan sát xu hướng giá cổ phiếu của ba mã chứng khoán trong năm vừa qua. Các biến động giá rõ ràng thể hiện sự phản ứng của thị trường đối với các sự kiện kinh tế hoặc chính trị.

Từ đó thấy được tầm quan trọng của trực quan hóa dữ liệu, dữ liệu chứng khoán có thể rất phức tạp và khó hiểu dưới dạng bảng số liệu. Việc sử dụng các kỹ thuật trực quan hóa dữ liệu không chỉ làm cho dữ liệu dễ hiểu mà còn hỗ trợ nhà đầu tư ra quyết định nhanh chóng và chính xác hơn, tối ưu hóa chiến lược đầu tư và quản trị rủi ro.

# CHƯƠNG 2: THU THẬP DỮ LIỆU

## 2.1. NGUỒN DỮ LIỆU

  Sử dụng nguồn dữ liệu API: vnstock

Web: https://vnstocks.com/docs/category/t%C3%A0i-li%E1%BB%87u

VNStock là một nền tảng cung cấp thông tin tài chính và dữ liệu chứng khoán tại Việt Nam. Đây là công cụ hữu ích cho nhà đầu tư, nhà nghiên cứu hoặc các tổ chức quan tâm đến thị trường chứng khoán Việt Nam. Một số đặc điểm nổi bật của VNStock bao gồm:

Dữ liệu thị trường đa dạng:

VNStock cung cấp thông tin chi tiết về các cổ phiếu niêm yết trên các sàn giao dịch lớn tại Việt Nam, như HOSE (Sở Giao dịch Chứng khoán TP.HCM), HNX (Sở Giao dịch Chứng khoán Hà Nội) và UPCoM (Thị trường Công ty Đại chúng chưa niêm yết).

Phân tích tài chính:

Cung cấp các báo cáo tài chính, số liệu thống kê quan trọng (P/E, EPS, ROE, ROA, ...) và các chỉ số tài chính của các công ty niêm yết.

Cập nhật thời gian thực:

Thông tin giá cổ phiếu, khối lượng giao dịch và các chỉ số thị trường (VN-Index, HNX-Index, UPCoM-Index) được cập nhật liên tục theo thời gian thực.

Biểu đồ và công cụ phân tích kỹ thuật:

Hỗ trợ biểu đồ tương tác, công cụ phân tích kỹ thuật như đường trung bình (MA), chỉ số RSI, MACD,... để người dùng theo dõi xu hướng và đưa ra quyết định đầu tư.

Tin tức thị trường:

Tích hợp các tin tức liên quan đến tài chính, kinh tế, doanh nghiệp và các sự kiện có thể ảnh hưởng đến thị trường chứng khoán.

Cộng đồng và học liệu:

Một số nền tảng như VNStock cũng cung cấp tài liệu, khóa học, và diễn đàn để hỗ trợ nhà đầu tư mới học cách tham gia và phân tích thị trường.

## 2.2. KỸ THUẬT DỮ LIỆU

Để sử dụng Vnstock, cần cài đặt thư viện trong ứng dụng Terminal trong môi trường Python hoặc 1 ô chứa mã Python trong giao diện Google Colab. Copy các dòng lệnh lần lượt như sau:

!pip install vnstock3

Nạp thư viện để sử dụng:

from vnstock3 import Vnstock # Nạp thư viện

import pandas as pd

import numpy as np

Định nghĩa biến vnstock lưu thông tin mã chứng khoán & nguồn dữ liệu bạn sử dụng với 3 mã cổ phiếu:

FPT: Công ty Cổ phần FPT (FPT Corporation) - Tập đoàn công nghệ và viễn thông hàng đầu tại Việt Nam, hoạt động trong các lĩnh vực như công nghệ thông tin, viễn thông và giáo dục.

VGI: Tổng Công ty Cổ phần Đầu tư Quốc tế Viettel (Viettel Global Investment JSC) - Công ty con thuộc Tập đoàn Viettel, chuyên kinh doanh trong lĩnh vực đầu tư và phát triển viễn thông quốc tế.

VTC: Tổng Công ty Truyền thông Đa phương tiện VTC (Vietnam Multimedia Corporation) - Một doanh nghiệp Nhà nước hoạt động trong lĩnh vực truyền thông, công nghệ thông tin và dịch vụ nội dung số.

# Định nghĩa biến vnstock lưu thông tin mã chứng khoán & nguồn dữ liệu bạn sử dụng

stock\_fpt = Vnstock().stock(symbol='FPT', source='VCI')

stock\_vgi = Vnstock().stock(symbol='VGI', source='VCI')

stock\_vtc = Vnstock().stock(symbol='VTC', source='VCI')

Thiết lập thời gian tải dữ liệu và khung thời gian tra cứu là 1 ngày:

# Thiết lập thời gian tải dữ liệu và khung thời gian tra cứu là 1 ngày

df\_fpt = stock\_fpt.quote.history(start='2024-01-01', end='2024-12-31', interval='1D')

df\_vgi = stock\_vgi.quote.history(start='2024-01-01', end='2024-12-31', interval='1D')

df\_vtc = stock\_vtc.quote.history(start='2024-01-01', end='2024-12-31', interval='1D')

Tạo một list\_banks với tên của các mã chứng khoán ở trên

list\_banks=["FPT","VGI","VTC"]

Ghép dữ liệu của 3 mã cổ phiếu FPT, VGI, VTC lại với nhau

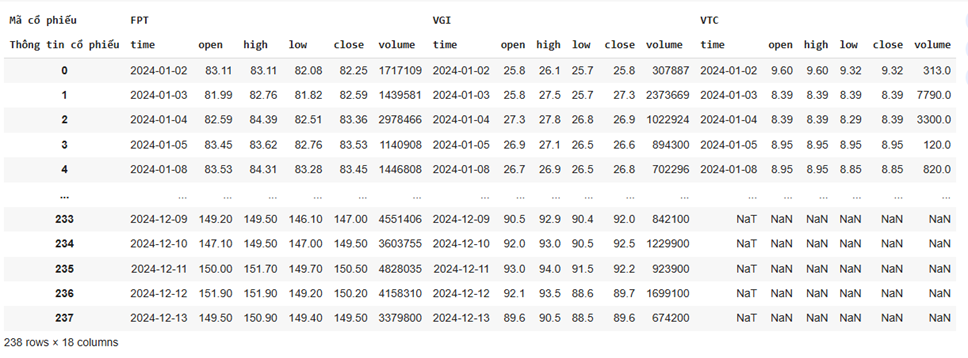
# ghép dữ liệu của 3 mã cổ phiếu FPT, VGI, VTC lại với nhau thành một DataFrame lớn hơn

bank\_stocks=pd.concat([df\_fpt,df\_vgi,df\_vtc],axis=1,keys=list\_banks)

bank\_stocks.columns.names = ['Mã cổ phiếu', 'Thông tin cổ phiếu']

bank\_stocks

# kết quả hiển thị



Bảng 2.1: Mã cổ phiếu và thông tin cổ phiếu

Sau khi thu thập thấy tập dữ liệu gồm các biến, cột

time: Thời gian (năm-tháng-ngày)

open (Giá mở cửa):

* Giá đầu tiên mà cổ phiếu được giao dịch khi thị trường bắt đầu mở cửa vào một ngày giao dịch.

high (Giá cao):

* Giá cao nhất mà cổ phiếu đạt được trong một ngày giao dịch.

low (Giá thấp):

* Giá thấp nhất mà cổ phiếu đạt được trong một ngày giao dịch.

close (Giá đóng cửa):

* Giá cuối cùng mà cổ phiếu được giao dịch trước khi thị trường đóng cửa vào cuối ngày.

volume (Khối lượng giao dịch):

* Tổng số lượng cổ phiếu được mua bán trong ngày giao dịch.

## 2.3. LÀM SẠCH DỮ LIỆU

Làm sạch dữ liệu và xử lý các giá trị thiếu:

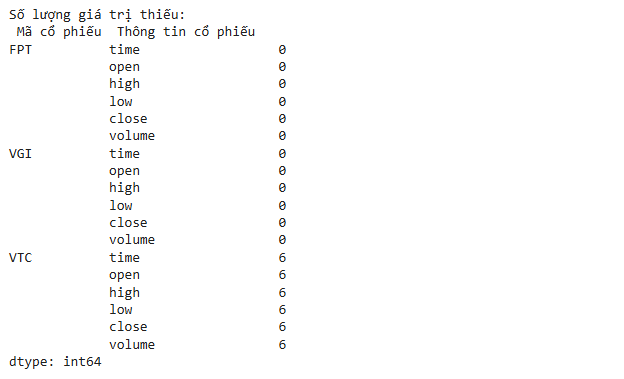
# Kiểm tra các giá trị thiếu trong dữ liệu

# Đếm số lượng giá trị thiếu ở từng cột

missing\_values = bank\_stocks.isnull().sum()

print("Số lượng giá trị thiếu:\n", missing\_values)

# kết quả hiển thị



Bảng 2.3: Số lượng giá trị thiếu

mã VTC ở tất cả các cột đều chứa 6 giá trị thiếu

Loại bỏ các hàng chứa giá trị thiếu:

bank\_stocks = bank\_stocks.dropna()

print("Số dòng sau khi loại bỏ dữ liệu thiếu:", bank\_stocks.shape[0])

# kết quả hiển thị



# CHƯƠNG 3: THỐNG KÊ MÔ TẢ

## 3.1. CÁC ĐẠI LƯỢNG, CHỈ SỐ THỐNG KÊ CƠ BẢN

Các đại lượng, chỉ số thống kê cơ bản:

Đối với chỉ số giá của chứng khoán như giá mở, giá cao, giá thấp, giá đóng cửa, chúng ta có thể tính các đại lượng thống kê cơ bản như:

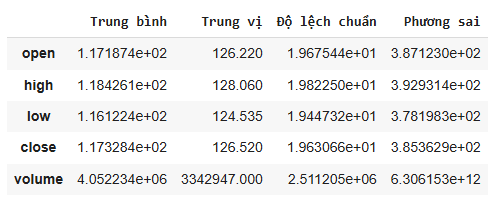
- Trung bình (Mean): Giá trung bình của từng loại chỉ số trong một khoảng thời gian.

- Trung vị (Median): Giá trị ở giữa tập dữ liệu, phản ánh giá trị trung tâm khi sắp xếp theo thứ tự tăng dần.

- Độ lệch chuẩn (Standard Deviation): Đo lường mức độ phân tán của giá trị xung quanh giá trị trung bình. Độ lệch chuẩn cao cho thấy sự biến động giá mạnh.

|  |
| --- |
| name=[ 'open','high','low','close','volume'] |

|  |
| --- |
| # Các đại lượng thống kê cơ bản : trung bình, trung vị, độ lệch chuẩn,phương sai  # Mã FPT  mean\_fpt=pd.DataFrame(df\_fpt[name].mean(),columns=["Trung bình"])  med\_fpt=pd.DataFrame(df\_fpt[name].median(),columns=["Trung vị"])  std\_fpt=pd.DataFrame(df\_fpt[name].std(),columns=["Độ lệch chuẩn"])  var\_fpt=pd.DataFrame(df\_fpt[name].var(),columns=["Phương sai"])  fpt=pd.concat([mean\_fpt,med\_fpt,std\_fpt,var\_fpt],axis=1)  fpt.columns=["Trung bình","Trung vị","Độ lệch chuẩn","Phương sai"]  fpt  # kết quả hiển thị |



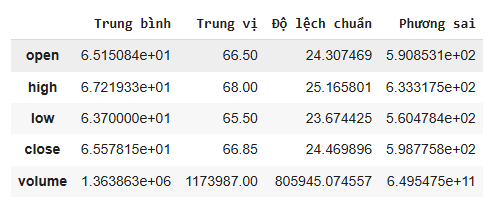
Bảng 3.1: Các đại lượng thống kê mã FPT

Nhận xét: Cổ phiếu FPT Dựa trên bảng số liệu thống kê của mã cổ phiếu FPT (năm 2024), các chỉ số trung bình, trung vị, độ lệch chuẩn và phương sai của giá mở, giá cao, giá thấp, giá đóng và khối lượng giao dịch được cung cấp. Sau đây là một số nhận xét:

1. Tính ổn định của giá cổ phiếu:
   * Giá mở, giá cao, giá thấp, và giá đóng đều có mức trung bình và trung vị khá sát nhau, cho thấy giá cổ phiếu FPT không biến động quá mạnh trong năm.
   * Độ lệch chuẩn và phương sai của giá các mức giá này khá thấp, phản ánh sự ổn định của mã cổ phiếu này.
2. Khối lượng giao dịch:
   * Khối lượng giao dịch trung bình là 3.93 triệu cổ phiếu, cao hơn rất nhiều so với trung vị 3.16 triệu cổ phiếu, cho thấy có những ngày giao dịch với khối lượng lớn vượt trội so với thông thường.
   * Độ lệch chuẩn của khối lượng giao dịch (2.49 triệu) cũng khá lớn, điều này chứng tỏ sự dao động của thanh khoản cổ phiếu FPT theo thời gian.
3. So sánh giá mở và giá đóng:
   * Giá mở trung bình (115.85) và giá đóng trung bình (115.95) rất gần nhau, điều này có thể chỉ ra rằng mã FPT thường không có sự thay đổi lớn giữa thời điểm mở và đóng cửa giao dịch.

Nhận định tổng quan: Cổ phiếu FPT có sự ổn định tốt về giá, phù hợp với các nhà đầu tư dài hạn tìm kiếm sự an toàn. Tuy nhiên, khối lượng giao dịch biến động lớn có thể tạo cơ hội cho nhà đầu tư ngắn hạn tận dụng những phiên giao dịch có thanh khoản cao.

|  |
| --- |
| # Các đại lượng thống kê cơ bản : trung bình, trung vị, độ lệch chuẩn,phương sai  # Mã VGI  mean\_vgi=pd.DataFrame(df\_vgi[name].mean(),columns=["Trung bình"])  med\_vgi=pd.DataFrame(df\_vgi[name].median(),columns=["Trung vị"])  std\_vgi=pd.DataFrame(df\_vgi[name].std(),columns=["Độ lệch chuẩn"])  var\_vgi=pd.DataFrame(df\_vgi[name].var(),columns=["Phương sai"])  vgi=pd.concat([mean\_vgi,med\_vgi,std\_vgi,var\_vgi],axis=1)  vgi.columns=["Trung bình","Trung vị","Độ lệch chuẩn","Phương sai"]  vgi  # kết quả hiển thị |



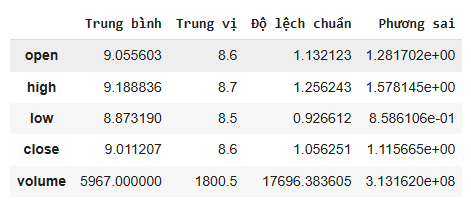
Bảng 3.2: Các đại lượng thống kê mã VGI

Nhận xét: Cổ phiếu VGI Dựa trên bảng thống kê dữ liệu của mã cổ phiếu VGI năm 2024, có thể đưa ra một số nhận xét như sau:

1. Tính ổn định của giá cổ phiếu:
   * Giá mở (63.13), giá cao (65.19), giá thấp (61.74) và giá đóng (63.58) có mức trung bình khá đồng đều. Trung vị của các giá này cũng gần nhau, phản ánh sự ổn định trong biến động giá của mã cổ phiếu VGI.
   * Tuy nhiên, độ lệch chuẩn của các mức giá (khoảng từ 24.38 đến 25.31) khá cao, cho thấy có sự dao động giá đáng kể trong suốt năm.
2. Khối lượng giao dịch:
   * Khối lượng giao dịch trung bình là 1.35 triệu cổ phiếu, lớn hơn trung vị 1.15 triệu cổ phiếu, cho thấy vẫn có những ngày giao dịch với khối lượng cao hơn mức thường lệ.
   * Độ lệch chuẩn của khối lượng giao dịch (818,371) cũng cho thấy thanh khoản của cổ phiếu này không đều, với một số phiên có giao dịch lớn hơn bình thường.
3. So sánh giá mở và giá đóng:
   * Giá đóng trung bình (63.58) nhỉnh hơn giá mở trung bình (63.13), điều này có thể cho thấy xu hướng tăng nhẹ vào cuối các phiên giao dịch.

Nhận định tổng quan: Cổ phiếu VGI có sự ổn định tương đối về giá, nhưng biến động vẫn xuất hiện với biên độ lớn hơn so với mã FPT. Điều này có thể thu hút các nhà đầu tư ưa thích chiến lược giao dịch ngắn hạn để tận dụng dao động giá. Tuy nhiên, nhà đầu tư cần lưu ý đến sự không đồng đều về thanh khoản để tối ưu hóa giao dịch.

|  |
| --- |
| # Các đại lượng thống kê cơ bản : trung bình, trung vị, độ lệch chuẩn,phương sai  # Mã VTC  mean\_vtc=pd.DataFrame(df\_vtc[name].mean(),columns=["Trung bình"])  med\_vtc=pd.DataFrame(df\_vtc[name].median(),columns=["Trung vị"])  std\_vtc=pd.DataFrame(df\_vtc[name].std(),columns=["Độ lệch chuẩn"])  var\_vtc=pd.DataFrame(df\_vtc[name].var(),columns=["Phương sai"])  vtc=pd.concat([mean\_vtc,med\_vtc,std\_vtc,var\_vtc],axis=1)  vtc.columns=["Trung bình","Trung vị","Độ lệch chuẩn","Phương sai"]  vtc  # kết quả hiển thị |



Bảng 3.3: Các đại lượng thống kê mã VTC

Nhận xét: Cổ phiếu VTC Dựa trên bảng thống kê dữ liệu của mã cổ phiếu VTC năm 2024, có thể đưa ra một số nhận xét như sau:

1.Tính ổn định của giá cổ phiếu:

* Các giá mở (9.12), giá cao (9.26), giá thấp (8.92) và giá đóng (9.07) có mức trung bình khá gần nhau, cho thấy biến động giá của cổ phiếu này không quá lớn.
* Trung vị của các giá cũng gần với giá trung bình (dao động trong khoảng 8.6 - 8.8), khẳng định thêm tính ổn định trong suốt năm.
* Độ lệch chuẩn thấp (từ 0.94 đến 1.28) cho thấy biên độ biến động của giá cổ phiếu này nhỏ hơn so với các mã như FPT hay VGI.

2. Khối lượng giao dịch:

* Khối lượng giao dịch trung bình là 6.33 triệu cổ phiếu, nhưng trung vị lại chỉ là 1.83 triệu cổ phiếu, cho thấy sự không đồng đều trong giao dịch. Một số phiên có khối lượng giao dịch đột biến rất lớn.
* Độ lệch chuẩn (18417.70) cao hơn trung vị rất nhiều, phản ánh sự bất ổn về thanh khoản.

3. So sánh giá mở và giá đóng:

* Giá mỡ trung bình (9.12) và giá đóng trung bình (9.07) rất gần nhau, điều này cho thấy cổ phiếu VTC không có sự thay đổi lớn giữa thời điểm bắt đầu và kết thúc mỗi phiên giao dịch.

Nhận định tổng quan: Cổ phiếu VTC có sự ổn định cao về giá, phù hợp với các nhà đầu tư dài hạn không muốn đối mặt với rủi ro lớn từ biến động giá. Tuy nhiên, khối lượng giao dịch có sự dao động lớn, nên các nhà đầu tư cần xem xét thanh khoản trong từng giai đoạn cụ thể để đảm bảo hiệu quả giao dịch.

## 3.2. BIỂU ĐỒ TÓM TẮT DỮ LIỆU

### 3.2.1. Biểu đồ thể hiện giá đóng cửa FPT, VGI, VTC

|  |
| --- |
| import matplotlib.pyplot as plt  # Đảm bảo cột 'time' có mặt trong dữ liệu nếu cần  df\_fpt['time'] = df\_fpt.index  df\_vgi['time'] = df\_vgi.index  df\_vtc['time'] = df\_vtc.index  # Danh sách dữ liệu và tên cổ phiếu  stocks = [  ('FPT', df\_fpt, '#1f77b4'), # Màu xanh lam  ('VGI', df\_vgi, '#ff7f0e'), # Màu cam  ('VTC', df\_vtc, '#2ca02c') # Màu xanh lá  ]  # Khởi tạo biểu đồ  plt.figure(figsize=(12, 6))  # Lặp qua từng cổ phiếu để thêm vào cùng một biểu đồ  for stock\_name, stock\_data, color in stocks:  plt.plot(stock\_data['time'], stock\_data['close'], label=stock\_name, color=color)  # Thêm tiêu đề, nhãn và chú thích  plt.title('Giá đóng cửa của FPT, VGI và VTC (2024)', fontsize=16)  plt.xlabel('Ngày', fontsize=12)  plt.ylabel('Giá đóng cửa (NGHÌN VND)', fontsize=12)  plt.grid(alpha=0.3)  plt.legend(title='Mã cổ phiếu', fontsize=10)  plt.tight\_layout()  # Hiển thị biểu đồ  plt.show()  # kết quả hiển thị |

Hình 3.1: Giá đóng cửa của FPT, VGI và VTC

## 3.3. SỬ DỤNG ƯỚC LƯỢNG VÀ KIỂM ĐỊNH ĐỂ PHÂN TÍCH MỐI LIÊN HỆ GIỮA GIÁ CHỨNG KHOÁN VÀ CÁC YẾU TỐ KHÁC

### 3.3.1. Kiểm tra dữ liệu có theo phân phối chuẩn không để so sánh dữ liệu.

-Phân phối chuẩn (Normal distribution) là một dạng phân phối xác suất phổ biến trong thống kê, có dạng hình chuông đối xứng. Việc xác định dữ liệu có tuân theo phân phối chuẩn hay không là rất quan trọng trong phân tích thống kê, vì nhiều phương pháp phân tích yêu cầu dữ liệu có phân phối chuẩn.

* Shapiro-Wilk Test: Kiểm định này dùng để kiểm tra giả thuyết gốc H0H\_0H0​ rằng dữ liệu tuân theo phân phối chuẩn.
* Nếu p-value > 0.05, không bác bỏ H0: dữ liệu có thể tuân theo phân phối chuẩn.
* Nếu p-value ≤ 0.05p, bác bỏ H0​: dữ liệu không tuân theo phân phối chuẩn.

import scipy.stats as stats

import seaborn as sns

import matplotlib.pyplot as plt

# Hàm kiểm tra phân phối chuẩn

def check\_normality(data, column\_name):

    # Histogram + KDE

    plt.figure(figsize=(12, 6))

    sns.distplot(data[column\_name], kde=True, color="blue", bins=30)

    plt.title(f"Histogram và KDE - {column\_name}")

    plt.xlabel(column\_name)

    plt.ylabel("Tần suất")

    plt.show()

    # Q-Q Plot

    plt.figure(figsize=(6, 6))

    stats.probplot(data[column\_name], dist="norm", plot=plt)

    plt.title(f"Q-Q Plot - {column\_name}")

    plt.show()

    # Kiểm định Shapiro-Wilk

    shapiro\_test = stats.shapiro(data[column\_name])

    print(f"Shapiro-Wilk Test: W = {shapiro\_test.statistic}, p-value = {shapiro\_test.pvalue}")

    # Kết luận từ p-value

    if shapiro\_test.pvalue > 0.05:

        print(f"'{column\_name}' có thể tuân theo phân phối chuẩn (không bác bỏ H0).\n")

    else:

        print(f"'{column\_name}' không tuân theo phân phối chuẩn (bác bỏ H0).\n")

# Kiểm tra từng mã cổ phiếu

dfs = {"FPT": df\_fpt, "VGI": df\_vgi, "VTC": df\_vtc}

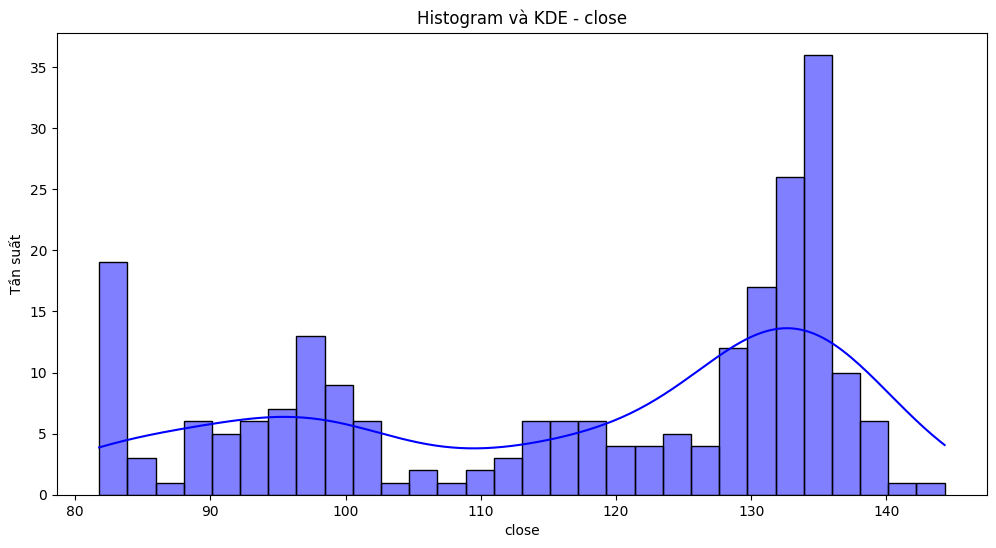
for name, df in dfs.items():

    print(f"--- Phân tích cho mã cổ phiếu {name} ---")

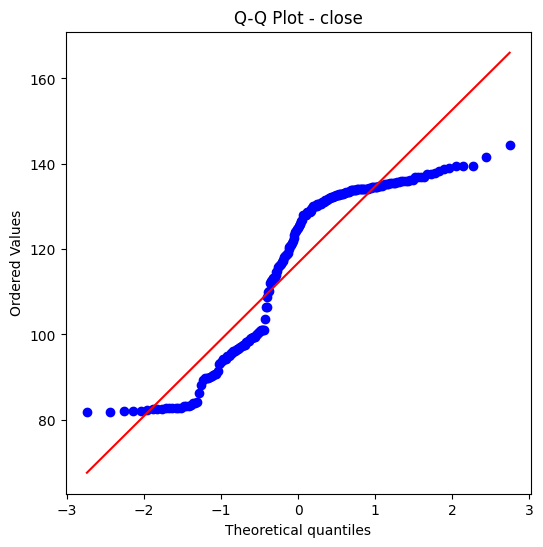
    check\_normality(df, "close")

Output:

--- Phân tích cho mã cổ phiếu VGI ---



Hình 3.2: Biểu đồ tần suất và Đường mật độ hạt nhân của mã cổ phiếu FPT

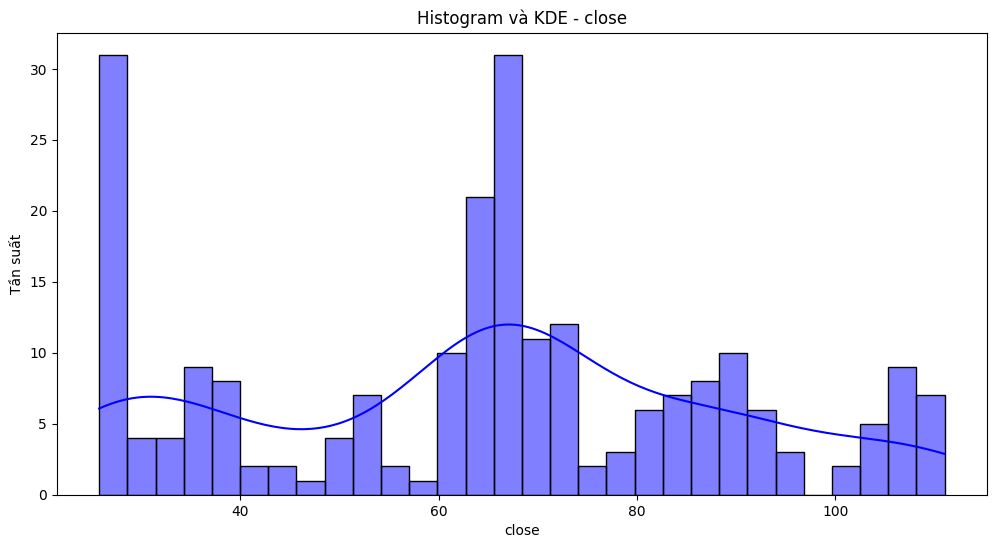


Hình 3.3: Biểu đồ định lượng - định lượng cho biến close của mã cổ phiếu FPT

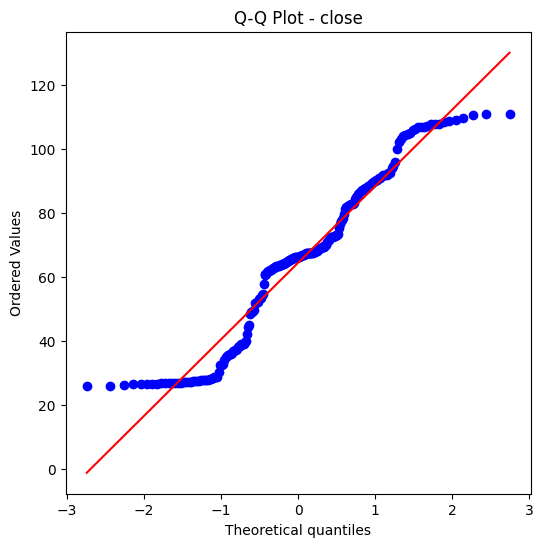
Shapiro-Wilk Test: W = 0.8630242452414278, p-value = 2.0374993055055205e-13

'close' không tuân theo phân phối chuẩn (bác bỏ H0).

--- Phân tích cho mã cổ phiếu VGI —



Hình 3.4: Biểu đồ tần suất và Đường mật độ hạt nhân của mã cổ phiếu VGI

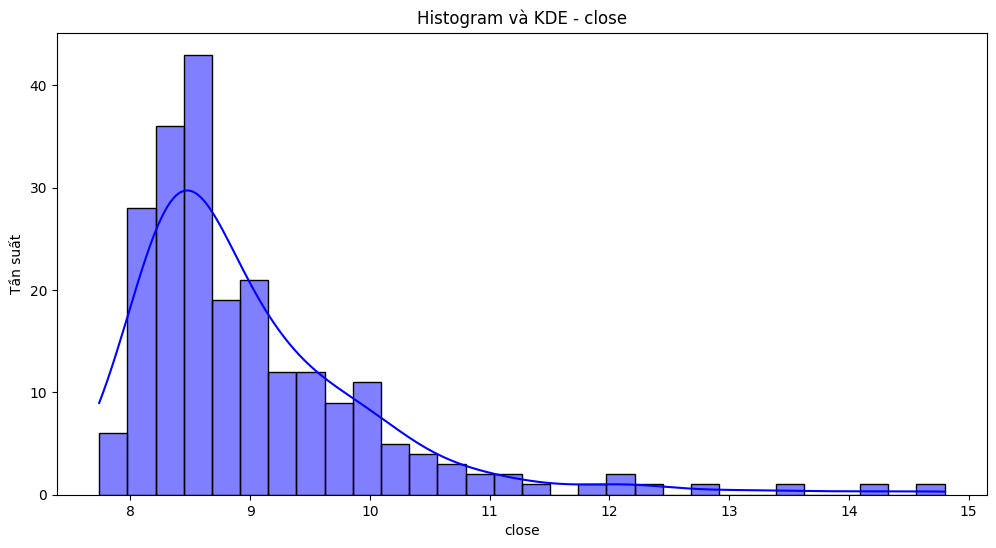


Hình 3.5: Biểu đồ định lượng - định lượng cho biến close của mã cổ phiếu VGI

Shapiro-Wilk Test: W = 0.9445826211291252, p-value = 1.2530188917177904e-07

'close' không tuân theo phân phối chuẩn (bác bỏ H0).

--- Phân tích cho mã cổ phiếu VTC —



Hình 3.6: Biểu đồ tần suất và Đường mật độ hạt nhân của mã cổ phiếu VTC



Hình 3.7: Biểu đồ định lượng - định lượng cho biến close của mã cổ phiếu VTC

Shapiro-Wilk Test: W = 0.7819530644747976, p-value = 7.001828753222947e-17

'close' không tuân theo phân phối chuẩn (bác bỏ H0).

* **Nhận xét:** Kết quả từ kiểm định Shapiro-Wilk cho thấy cột close không tuân theo phân phối chuẩn (p-value rất nhỏ, gần bằng 0). Vì vậy, cần điều chỉnh cách phân tích và sử dụng các phương pháp không giả định dữ liệu có phân phối chuẩn.

### 3.3.2.Sử dụng các thống kê không tham số.

* Thống kê không tham số (non-parametric statistics) là một nhánh của thống kê, tập trung vào việc phân tích dữ liệu mà không cần giả định cụ thể về phân phối xác suất của tổng thể (như phân phối chuẩn).

Đặc điểm:

* Không yêu cầu giả định về tham số của tổng thể (ví dụ: trung bình, phương sai).
* Thích hợp khi:
  + Dữ liệu không tuân theo phân phối chuẩn.
  + Cỡ mẫu nhỏ, khó xác định phân phối của dữ liệu.
  + Dữ liệu dạng thứ bậc, danh mục hoặc không thể đo lường liên tục.
* Thường dựa trên xếp hạng hoặc thứ tự thay vì giá trị thực tế.

Ưu điểm:

* Không yêu cầu dữ liệu phải có phân phối cụ thể.
* Có thể áp dụng cho nhiều loại dữ liệu khác nhau.
* Thường ít bị ảnh hưởng bởi giá trị ngoại lai.

Hạn chế:

* Có thể mất một số thông tin trong dữ liệu (do sử dụng xếp hạng thay vì giá trị thực tế).
* Kém mạnh mẽ hơn so với các phương pháp tham số nếu dữ liệu thực sự tuân theo phân phối chuẩn.

3.3.2.1. Phân tích trung vị và phạm vi liên phần tư (Median and Interquartile Range Analysis)

a. Trung vị (Median)

Khái niệm:

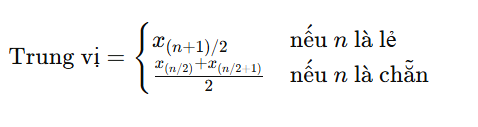
* Trung vị là giá trị nằm ở vị trí chính giữa của dữ liệu khi sắp xếp theo thứ tự tăng dần.
* Trong trường hợp số lượng quan sát là lẻ, trung vị là giá trị chính giữa. Với số lượng quan sát chẵn, trung vị là trung bình của hai giá trị ở giữa.

Đặc điểm:

* Trung vị là một đại lượng không bị ảnh hưởng bởi giá trị ngoại lai, khác với trung bình (mean).
* Được sử dụng phổ biến để biểu diễn dữ liệu phân phối không chuẩn hoặc dữ liệu có ngoại lệ lớn.

Công thức:

* Với n là số lượng phần tử:



b. Phạm vi liên phần tư (IQR - Interquartile Range)

Khái niệm:

* IQR là khoảng cách giữa phần tư thứ 3 (Q3 - quartile 75%) và phần tư thứ 1 (Q1 - quartile 25%) của dữ liệu.
* IQR đo lường mức độ phân tán của dữ liệu xung quanh trung vị, giúp loại bỏ ảnh hưởng của các giá trị ngoại lai.

Công thức: IQR = Q3 - Q1

Trong đó:

* Q1Q1Q1: Phần tư thứ nhất, giá trị tại 25% dữ liệu khi sắp xếp tăng dần.
* Q3Q3Q3: Phần tư thứ ba, giá trị tại 75% dữ liệu.

Đặc điểm:

* Giúp đánh giá mức độ biến thiên của dữ liệu.
* Thường được sử dụng thay cho phương sai hoặc độ lệch chuẩn khi dữ liệu không tuân theo phân phối chuẩn.

Input:

# Tính các thống kê không tham số

non\_parametric\_stats = pd.DataFrame({

    "Trung vị": df\_fpt['close'].median(),

    "IQR (Phạm vi liên phần tư)": df\_fpt['close'].quantile(0.75) - df\_fpt['close'].quantile(0.25)

}, index=["close"])

print("Thống kê không tham số cho cột 'close':\n", non\_parametric\_stats)

Output:

Thống kê không tham số cho cột 'close':

        Trung vị  IQR (Phạm vi liên phần tư)

close    125.25                     35.4675

* **Nhận xét:** Trung vị giá đóng cửa (close) là 125.25, và IQR (Phạm vi liên phần tư) là 35.4675, chỉ ra rằng phần lớn dữ liệu giá đóng cửa nằm trong khoảng hẹp.

3.3.2.2. So sánh giữa các cổ phiếu (Phương pháp không tham số)

Input:

# Kiểm định Kruskal-Wallis cho cột close giữa 3 cổ phiếu

kruskal\_test = stats.kruskal(df\_fpt['close'], df\_vgi['close'], df\_vtc['close'])

print(f"Kruskal-Wallis Test: H = {kruskal\_test.statistic}, p-value = {kruskal\_test.pvalue}")

# Kết luận

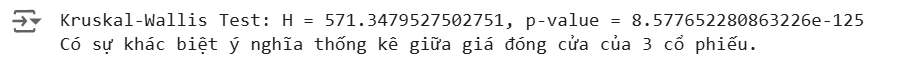
if kruskal\_test.pvalue > 0.05:

    print("Không có sự khác biệt ý nghĩa thống kê giữa giá đóng cửa của 3 cổ phiếu.")

else:

    print("Có sự khác biệt ý nghĩa thống kê giữa giá đóng cửa của 3 cổ phiếu.")

Output:



**Nhận xét:** Với p-value rất nhỏ (≈ 0), có sự khác biệt ý nghĩa thống kê giữa giá đóng cửa của các cổ phiếu FPT, VGI, và VTC.

### 3.3.3. Biểu đồ trực quan hóa.

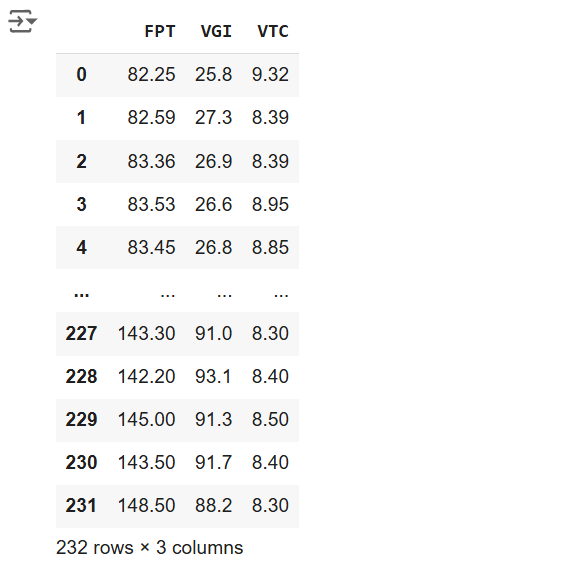
Kết hợp dữ liệu của 3 mã cổ phiếu để trực quan hóa.

df = pd.concat([df\_fpt['close'], df\_vgi['close'], df\_vtc['close']], axis=1)

df.columns = ['FPT', 'VGI', 'VTC']

df.dropna(inplace=True)

df



Bảng 3.4: DataFrame của 3 mã cổ phiếu FPT, VGI, VTC.

Input:

plt.figure(figsize=(12, 6))

sns.boxplot(df,

            notch=True, palette="Set2")

plt.xticks(["FPT", "VGI", "VTC"])

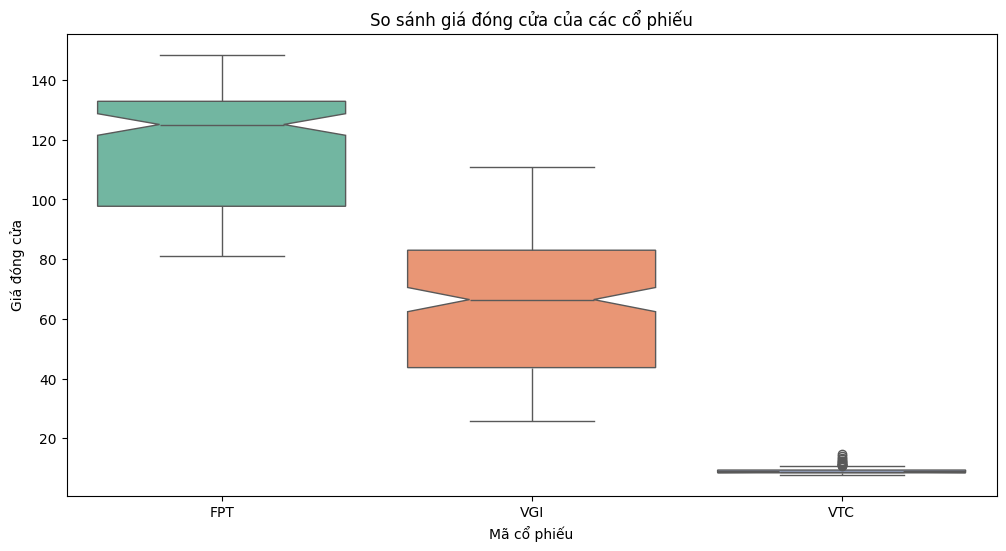
plt.title("So sánh giá đóng cửa của các cổ phiếu")

plt.ylabel("Giá đóng cửa")

plt.xlabel("Mã cổ phiếu")

plt.show()

Output:



Hình3.8: Biểu đồ boxplot so sánh giá đóng cửa của các cổ phiếu

Đoạn code trên sử dụng thư viện Matplotlib và Seaborn để tạo biểu đồ boxplot nhằm so sánh giá đóng cửa của ba cổ phiếu: FPT, VGI, và VTC.

* + figsize=(12, 6): thiết lập kích thước của biểu đồ với chiều rộng 12 và chiều cao 6 (đơn vị inch).
  + notch=True: Thêm khoảng khuyết (notch) vào hộp để làm nổi bật trung vị và đánh giá khoảng tin cậy của trung vị.
  + palette="Set2": Sử dụng bảng màu Set2 để tăng tính thẩm mỹ cho biểu đồ.
  + plt.title: Tiêu đề của biểu đồ.
  + plt.ylabel: Nhãn của trục Y thể hiện giá trị giá đóng cửa.
  + plt.xlabel: Nhãn của trục X thể hiện mã cổ phiếu.

Biểu đồ này giúp so sánh phân bố giá đóng cửa của ba cổ phiếu khác nhau (FPT, VGI, VTC) thông qua các đặc trưng như: trung vị, độ phân tán, và các giá trị ngoại lệ.

1. Cổ phiếu FPT:
   * Giá đóng cửa của cổ phiếu FPT nằm trong khoảng từ 80 đến 150.
   * Trung vị (median) nằm ở khoảng 130, cho thấy giá chủ yếu tập trung ở mức cao.
   * Phân phối của giá khá cân đối với độ rộng hộp lớn, biểu hiện sự biến động vừa phải của giá cổ phiếu.
2. Cổ phiếu VGI:
   * Giá đóng cửa của cổ phiếu VGI nằm trong khoảng từ 25 đến 110.
   * Trung vị nằm khoảng 65, cho thấy mức giá tập trung ở khoảng trung bình.
   * Phân phối khá lệch về phía dưới (right skewed), với một vài giá trị cao làm kéo dài phần trên của hộp.
3. Cổ phiếu VTC:
   * Giá đóng cửa của cổ phiếu VTC rất thấp, nằm trong khoảng 5 đến 15.
   * Trung vị nằm ở khoảng 10.
   * Phân phối rất hẹp và tập trung, cho thấy giá cổ phiếu VTC ít biến động.

Tổng quan:

* Cổ phiếu FPT có giá trị cao và ít biến động hơn so với VGI.
* Cổ phiếu VGI có sự biến động lớn hơn, với khoảng giá rộng và lệch về phía dưới.
* Cổ phiếu VTC có giá trị rất thấp và ổn định hơn so với hai mã cổ phiếu còn lại.

Nhìn chung, biểu đồ thể hiện sự khác biệt rõ rệt giữa ba mã cổ phiếu về giá trị và độ biến động.

### 3.3.4**.** Phân tích sâu sự khác biệt giữa các cổ phiếu.

Input:

from scipy.stats import mannwhitneyu

# So sánh từng cặp cổ phiếu

pairs = [('FPT', df\_fpt['close']), ('VGI', df\_vgi['close']), ('VTC', df\_vtc['close'])]

results = []

for i in range(len(pairs)):

    for j in range(i + 1, len(pairs)):

        stock1, data1 = pairs[i]

        stock2, data2 = pairs[j]

        test\_stat, p\_value = mannwhitneyu(data1, data2, alternative='two-sided')

        results.append((f"{stock1} vs {stock2}", test\_stat, p\_value))

# Hiển thị kết quả

for pair, stat, p\_val in results:

    print(f"Kiểm định Mann-Whitney U {pair}: U = {stat}, p-value = {p\_val}")

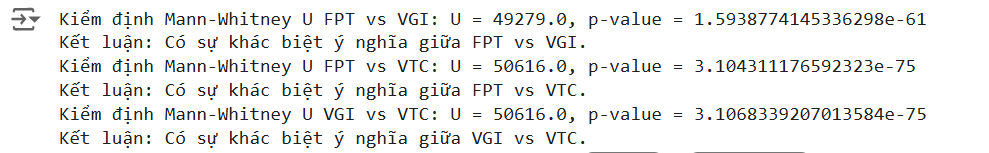
    if p\_val < 0.05:

        print(f"Kết luận: Có sự khác biệt ý nghĩa giữa {pair}.")

    else:

        print(f"Kết luận: Không có sự khác biệt ý nghĩa giữa {pair}.")

Output:



x, y: Hai tập dữ liệu cần so sánh.

alternative: Xác định giả thuyết thay thế. Có 3 tùy chọn:

* 'two-sided': Kiểm tra xem hai phân phối có khác biệt tổng quát không.
* 'less': Kiểm tra xem phân phối của x có nhỏ hơn y.
* 'greater': Kiểm tra xem phân phối của x có lớn hơn y.
* test\_stat: Giá trị thống kê U.
* p\_value: Giá trị p, thể hiện mức ý nghĩa của kiểm định.
* mannwhitneyu(data1, data2, alternative='two-sided'): So sánh hai tập dữ liệu data1 và data2.
* Giá trị p-value rất nhỏ (< 0.05):
  + Trong tất cả các cặp cổ phiếu được so sánh, giá trị p-value đều rất nhỏ (gần như bằng 0), cho thấy sự khác biệt giữa các cặp cổ phiếu là có ý nghĩa thống kê.
  + Điều này nghĩa là phân phối giá đóng cửa của mỗi cặp cổ phiếu không giống nhau một cách rõ ràng.
* Giá trị U:
  + Giá trị U càng lớn càng chỉ ra sự khác biệt giữa hai tập dữ liệu.
  + Các giá trị U của các cặp đều khá cao, điều này càng củng cố kết luận rằng phân phối giá đóng cửa giữa các cổ phiếu là khác nhau.
* **Nhận xét:** Kết quả kiểm định cho thấy sự khác biệt đáng kể về phân phối giá đóng cửa giữa tất cả các cặp cổ phiếu: FPT, VGI, và VTC. Điều này có thể phản ánh sự khác biệt trong tính chất tài chính, độ biến động, hoặc xu hướng thị trường của các cổ phiếu.

Trực quan hóa thêm:

plt.figure(figsize=(12, 6))

sns.violinplot(df,

               palette="muted", split=True)

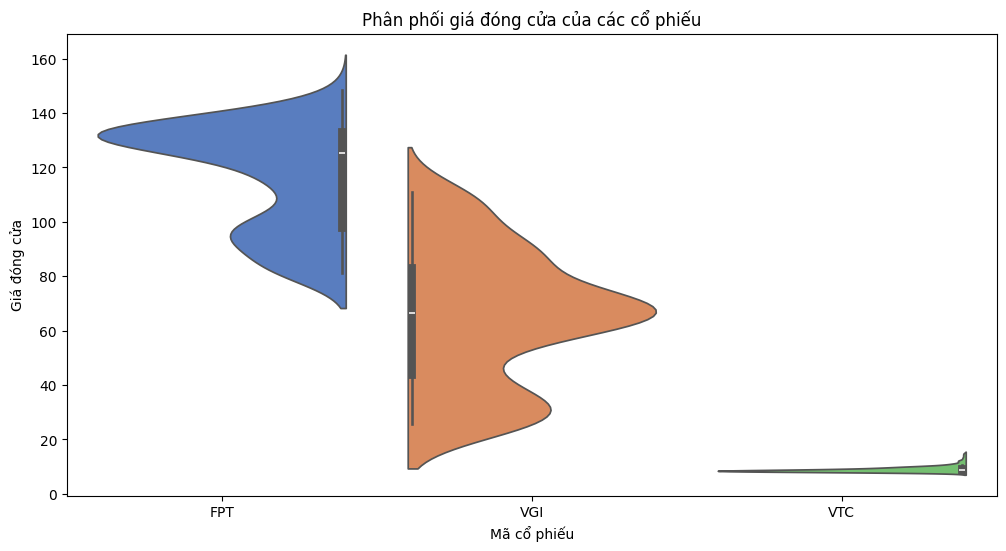
plt.xticks(["FPT", "VGI", "VTC"])

plt.title("Phân phối giá đóng cửa của các cổ phiếu")

plt.ylabel("Giá đóng cửa")

plt.xlabel("Mã cổ phiếu")

plt.show()  
#Hiển thị kết quả:



Hình 3.9: Biểu đồ violin thể hiện Phân phối giá đóng cửa của các cổ phiếu

plt.figure(figsize=(12, 6)):

* Tạo một khung hình (figure) với kích thước rộng 12 inch và cao 6 inch để chứa biểu đồ.
* Điều này giúp biểu đồ rõ ràng hơn và phù hợp khi làm việc với nhiều dữ liệu.

sns.violinplot(...):

* Hàm violinplot từ thư viện Seaborn được sử dụng để vẽ biểu đồ violin, biểu diễn phân phối của dữ liệu.
* palette="muted": Lựa chọn bảng màu nhẹ (muted) cho biểu đồ để dễ nhìn và chuyên nghiệp hơn.
* split=True (không áp dụng trong biểu đồ này do không có nhóm đối chứng): Thường được sử dụng khi phân phối có thêm một biến phân loại (nhóm khác nhau). Nếu có, hai phần của biểu đồ violin sẽ được tách ra để so sánh.

### 3.3.5. Phân tích tương quan.

3.3.5.1. Hệ số tương quan.

Hệ số tương quan là một đại lượng số học được sử dụng để đo lường mức độ mối quan hệ giữa hai biến số. Nó cho biết đến mức độ nào các biến số di chuyển cùng nhau trong một tập dữ liệu. Hệ số tương quan có giá trị từ -1 đến 1, trong đó:

* Nếu hệ số tương quan gần bằng 1, có một mối tương quan tuyến tính mạnh dương, tức là khi một biến tăng thì biến kia cũng tăng theo cùng hướng.
* Nếu hệ số tương quan gần bằng -1, có một mối tương quan tuyến tính mạnh âm, tức là khi một biến tăng thì biến kia giảm theo cùng hướng.
* Nếu hệ số tương quan gần bằng 0, không có mối tương quan tuyến tính giữa hai biến.

Input:

# Tính ma trận tương quan

close\_prices = pd.DataFrame({

    "FPT": df\_fpt["close"],

    "VGI": df\_vgi["close"],

    "VTC": df\_vtc["close"]

})

correlation\_matrix = close\_prices.corr()

print("Ma trận tương quan:")

print(correlation\_matrix)

# Vẽ biểu đồ heatmap cho ma trận tương quan

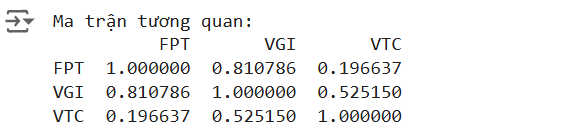
plt.figure(figsize=(8, 6))

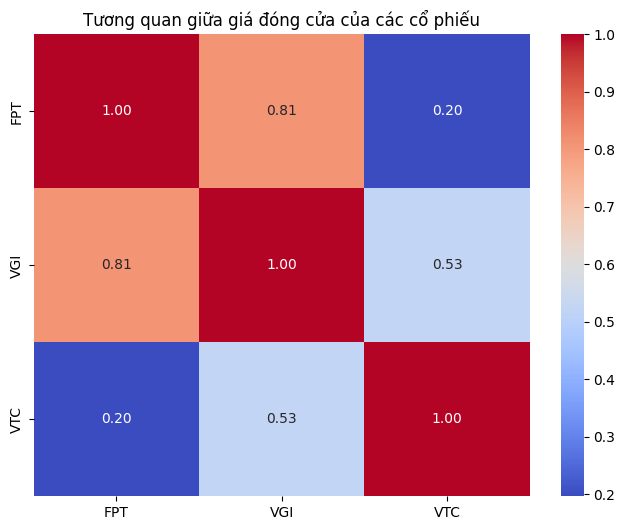
sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm", fmt=".2f", cbar=True)

plt.title("Tương quan giữa giá đóng cửa của các cổ phiếu")

plt.show()

Output:





Hình 3.10: Biểu đồ Heatmap thể hiện tương quan giữa giá đóng cửa của các cổ phiếu

**Nhận xét:**

FPT và VGI: Hệ số tương quan là 0.81 → Mối tương quan dương mạnh (cùng chiều).

FPT và VTC: Hệ số tương quan là 0.20 → Mối tương quan yếu.

VGI và VTC: Hệ số tương quan là 0.53 → Mối tương quan trung bình.

Biểu đồ heatmap trực quan hóa mối tương quan giữa các cổ phiếu và giúp nhận diện nhanh chóng mối quan hệ giữa chúng. Trong trường hợp này, FPT và VGI có mối tương quan mạnh, trong khi VTC ít tương quan hơn với hai cổ phiếu còn lại.

3.3.5.2.Biểu đồ phân tán (scatter plot).

# Vẽ biểu đồ scatter plot

plt.figure(figsize=(14, 6))

# FPT và VGI

plt.subplot(1, 3, 1)

sns.scatterplot(x=close\_prices["FPT"], y=close\_prices["VGI"], color="blue", alpha=0.6)

plt.title("FPT vs VGI")

plt.xlabel("FPT - Giá đóng cửa")

plt.ylabel("VGI - Giá đóng cửa")

# FPT và VTC

plt.subplot(1, 3, 2)

sns.scatterplot(x=close\_prices["FPT"], y=close\_prices["VTC"], color="green", alpha=0.6)

plt.title("FPT vs VTC")

plt.xlabel("FPT - Giá đóng cửa")

plt.ylabel("VTC - Giá đóng cửa")

# VGI và VTC

plt.subplot(1, 3, 3)

sns.scatterplot(x=close\_prices["VGI"], y=close\_prices["VTC"], color="orange", alpha=0.6)

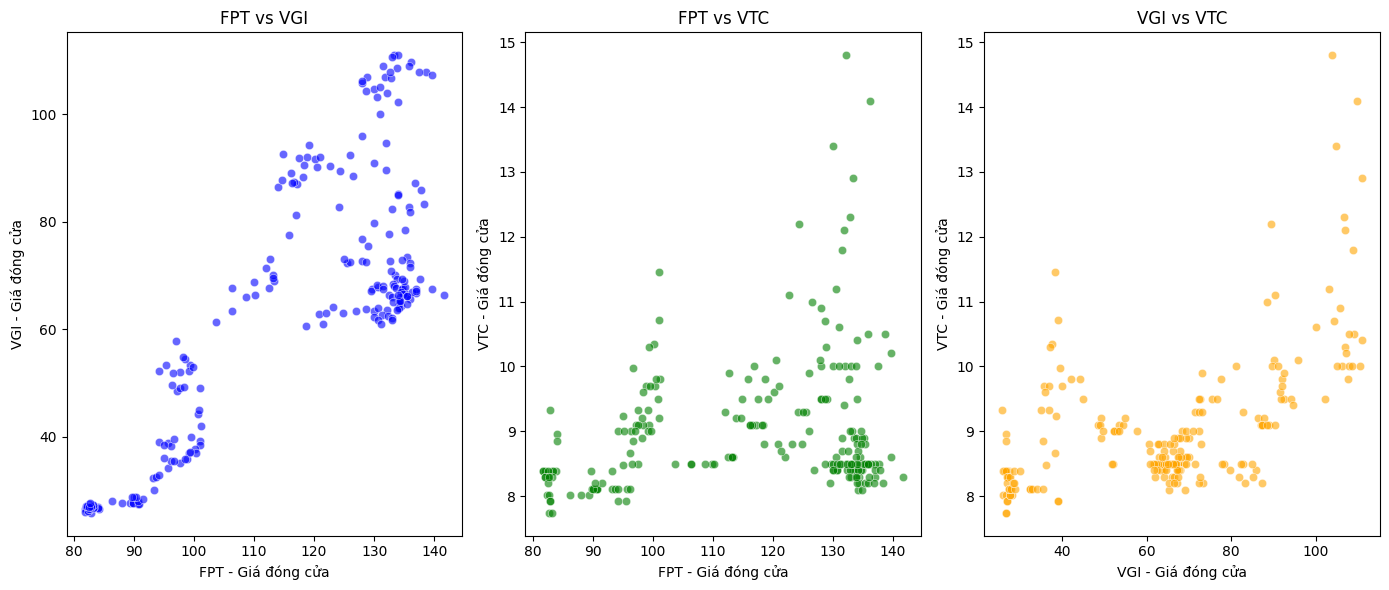
plt.title("VGI vs VTC")

plt.xlabel("VGI - Giá đóng cửa")

plt.ylabel("VTC - Giá đóng cửa")

plt.tight\_layout()

plt.show()

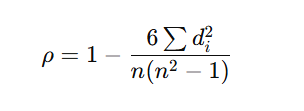


Hình 3.11: Biểu đồ phân tán giá đóng cửa của 3 mã cổ phiếu FPT, VGI, VTC

### 3.3.6. Kiểm định giả thuyết về tương quan (Spearman).

Kiểm định tương quan Spearman là một phương pháp thống kê được sử dụng để đo lường và kiểm định mối tương quan đơn điệu (monotonic relationship) giữa hai biến, thông qua thứ hạng của các giá trị thay vì giá trị thực.

Hệ số tương quan Spearman đo lường mức độ và chiều hướng của mối quan hệ giữa hai biến. Công thức tính như sau:



Trong đó:

* di: Hiệu giữa thứ hạng của cặp giá trị tương ứng (thứ hạng của xi và yi​).
* n: Kích thước mẫu (số cặp quan sát).
* Hệ số Spearman (ρ) nằm trong khoảng [−1,1]:
  + ρ>0: Mối tương quan thuận (cùng chiều).
  + ρ<0: Mối tương quan nghịch (ngược chiều).
  + ρ=0: Không có mối tương quan đơn điệu giữa hai biến.

Kiểm định giả thuyết Spearman

1. Giả thuyết thống kê

* H0 (Giả thuyết gốc): Không có mối tương quan đơn điệu giữa hai biến.
* H1 (Giả thuyết thay thế): Có mối tương quan đơn điệu giữa hai biến.

2. Ý nghĩa của kiểm định

* Kiểm định giúp xác định mối tương quan có ý nghĩa thống kê hay không.
* p-value được sử dụng để ra quyết định:
  + p≤α (thường là 0.05): Bác bỏ H0 → Có mối tương quan có ý nghĩa.
  + p>α: Không đủ bằng chứng để bác bỏ H0.

from scipy.stats import pearsonr, spearmanr

import numpy as np

import pandas as pd

# Lấy giá đóng cửa của các cổ phiếu

close\_prices = pd.DataFrame({

    "FPT": df\_fpt["close"],

    "VGI": df\_vgi["close"],

    "VTC": df\_vtc["close"]

})

# Loại bỏ các giá trị NaN và inf

close\_prices = close\_prices.dropna()  # Loại bỏ NaN

close\_prices = close\_prices[~close\_prices.isin([np.inf, -np.inf]).any(axis=1)]  # Loại bỏ inf

# Kiểm định tương quan Spearman

print("\n--- Kiểm định Spearman ---")

for col1 in close\_prices.columns:

    for col2 in close\_prices.columns:

        if col1 != col2:

            corr, p\_value = spearmanr(close\_prices[col1], close\_prices[col2])

            print(f"Tương quan Spearman giữa {col1} và {col2}: {corr:.2f}, p-value = {p\_value:.4f}")



Phân tích chéo các khoảng thời gian:

quarters = [fpt\_quy1, fpt\_quy2, fpt\_quy3, fpt\_quy4]

quarter\_names = ["Quý 1", "Quý 2", "Quý 3", "Quý 4"]

for i, quarter in enumerate(quarters):

    quarter\_prices = pd.DataFrame({

        "FPT": quarter["close"],

        "VGI": vgi\_quy1["close"].reindex(quarter.index),  # Đồng bộ index

        "VTC": vtc\_quy1["close"].reindex(quarter.index)

    })

    print(f"\n--- Tương quan Spearman giá đóng cửa trong {quarter\_names[i]} ---")

    print(quarter\_prices.corr(method="spearman"))



quarters: Chứa dữ liệu giá đóng cửa của cổ phiếu FPT cho từng quý.

quarter\_names: Tên các quý tương ứng.

Đồng bộ index: Dữ liệu của VGI và VTC được đồng bộ về cùng index (thời gian) với FPT bằng .reindex().

Tạo DataFrame quarter\_prices chứa giá đóng cửa của 3 cổ phiếu (FPT, VGI, VTC).

method="spearman": Tính toán hệ số tương quan Spearman.

**Nhận xét:**

* + Tương quan Spearman giữa các cổ phiếu FPT và VGI rất mạnh trong Quý 1 và Quý 2 nhưng giảm mạnh trong Quý 3 và Quý 4. Điều này có thể do sự thay đổi trong mối quan hệ đơn điệu giữa giá đóng cửa của 2 cổ phiếu trong nửa cuối năm.
  + Sự suy giảm tương quan trong Quý 3 và Quý 4 có thể do các yếu tố thị trường hoặc các sự kiện kinh tế, chính trị làm cho biến động giá giữa các cổ phiếu không còn đồng nhất như trước.

*\*Tìm kiếm xu hướng:* Phân tích trung bình động (SMA)

# Thêm SMA 7 ngày vào dữ liệu

df\_fpt['SMA\_7'] = df\_fpt['close'].rolling(window=7).mean()

plt.figure(figsize=(12, 6))

plt.plot(df\_fpt['time'], df\_fpt['close'], label='Giá đóng cửa', alpha=0.6)

plt.plot(df\_fpt['time'], df\_fpt['SMA\_7'], label='SMA 7 ngày', color='red')

plt.title("Xu hướng giá cổ phiếu FPT với SMA 7 ngày")

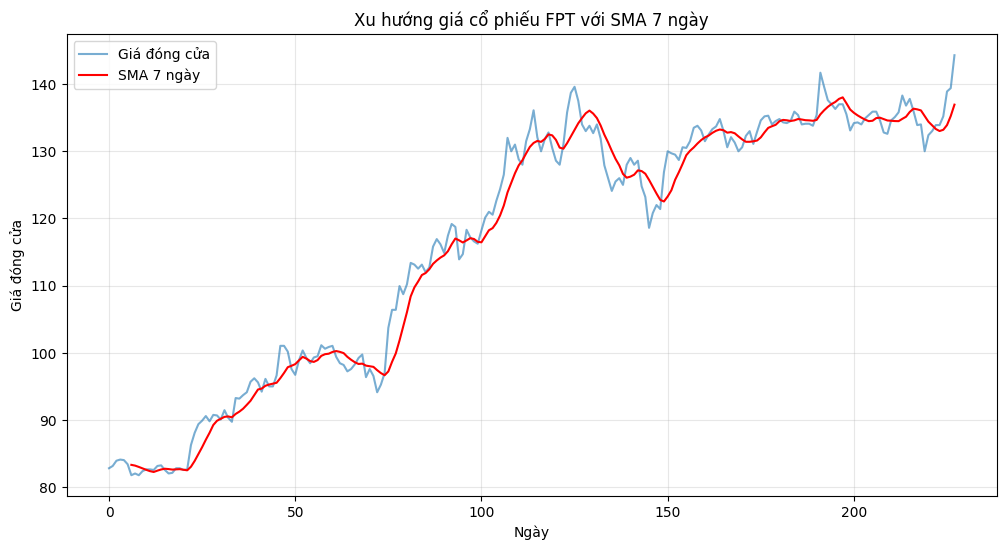
plt.xlabel("Ngày")

plt.ylabel("Giá đóng cửa")

plt.legend()

plt.grid(alpha=0.3)

plt.show()



Hình 3.12: Biểu đồ phân tích xu hướng giá cổ phiếu FPT với SMA 7 ngày

### 3.3.7.Phân tích mô hình biến động giá.

* Biến động giá: Được tính toán thông qua độ lệch chuẩn của tỷ suất sinh lợi hàng ngày (Daily Return), giúp đánh giá mức độ rủi ro khi đầu tư vào cổ phiếu đó.
* Biến động cao phản ánh sự không chắc chắn lớn trong thị trường, có thể làm tăng rủi ro nhưng cũng tạo cơ hội lớn cho nhà đầu tư.
* Nếu biến động nhỏ, cổ phiếu ổn định và phù hợp cho các nhà đầu tư ngại rủi ro.

# Tính tỷ suất sinh lợi hàng ngày

df\_fpt['daily\_return'] = df\_fpt['close'].pct\_change()

volatility = df\_fpt['daily\_return'].std()

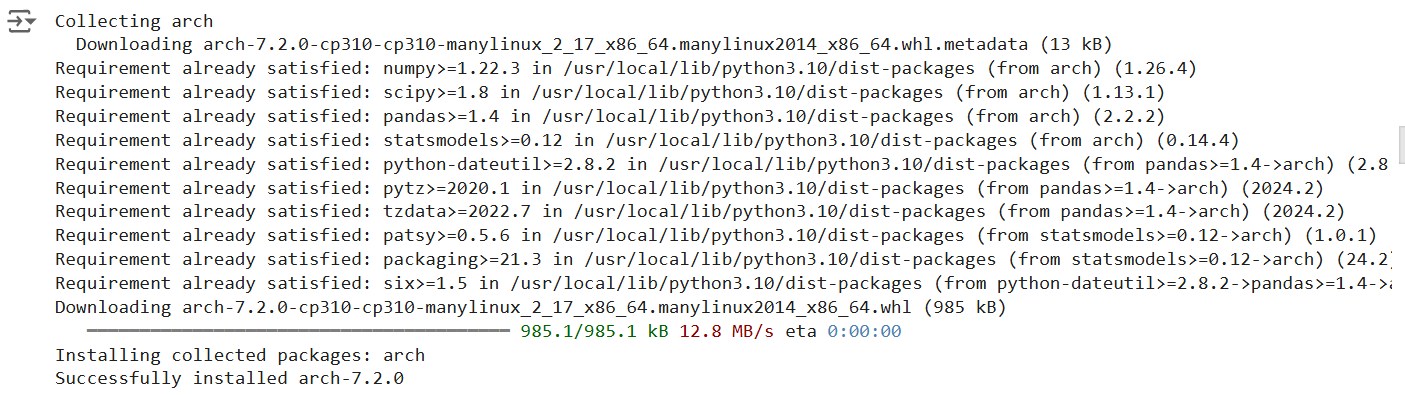
print(f"Độ lệch chuẩn của tỷ suất sinh lợi (FPT): {volatility:.4f}")



* ARCH/GARCH models: Các mô hình ARCH và GARCH có thể được áp dụng để ước lượng và dự báo mức độ biến động của giá cổ phiếu qua thời gian. Những mô hình này rất hữu ích khi phân tích tác động của tin tức thị trường đến sự biến động giá cổ phiếu, đặc biệt là khi dữ liệu có dấu hiệu biến động không đều đặn (heteroscedasticity).
* Mô hình GARCH cung cấp dự báo chính xác hơn về biến động giá, đặc biệt trong các thị trường có sự thay đổi không đồng đều.
* Nếu các tham số trong mô hình GARCH (như 𝛼 α và 𝛽 β) cao, điều này cho thấy biến động trong quá khứ ảnh hưởng lớn đến biến động hiện tại.
* Ý nghĩa: Dự báo mức độ biến động trong tương lai dựa trên biến động lịch sử.

Cài đặt thư viện Python **arch:**

!pip install arch



from arch import arch\_model

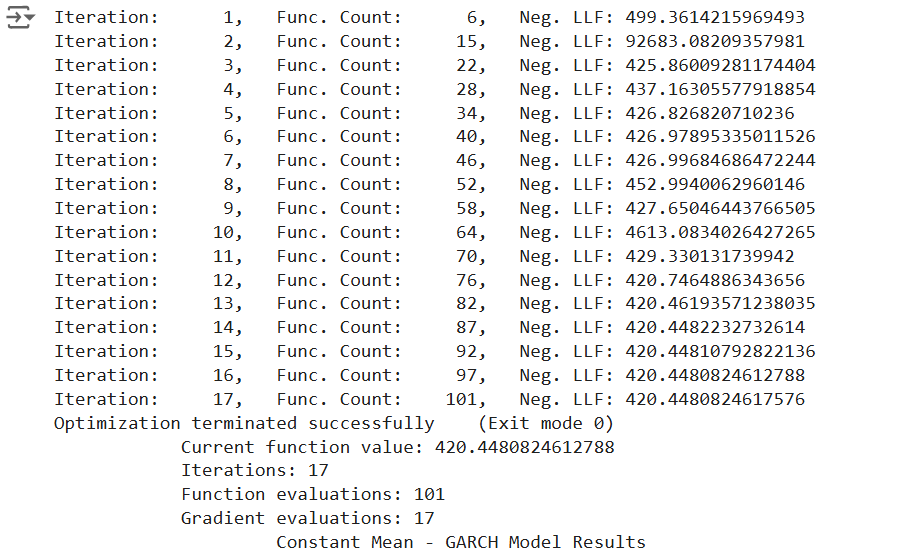
# Mô hình GARCH(1,1) cho tỷ suất sinh lợi

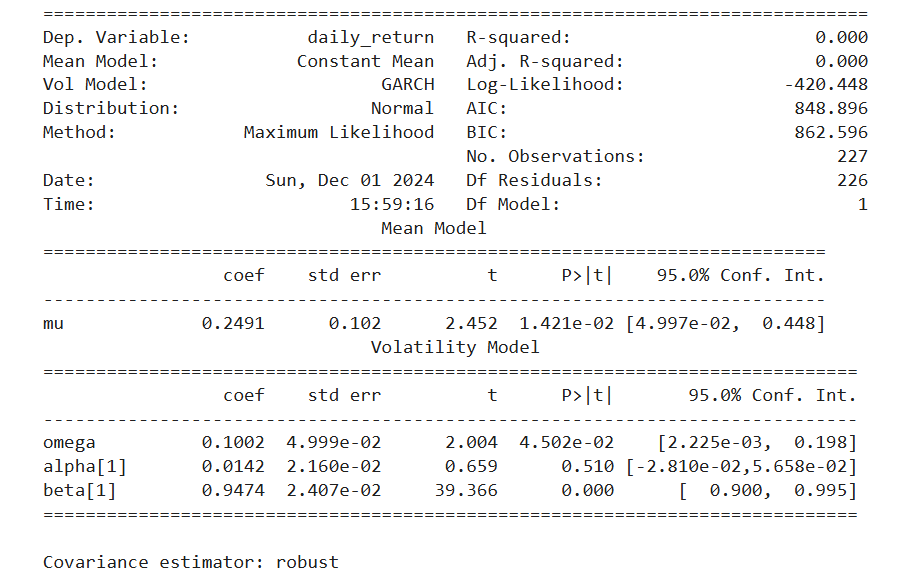
returns = df\_fpt['daily\_return'].dropna() \* 100  # Chuyển đổi về %

garch\_model = arch\_model(returns, vol='Garch', p=1, q=1)

garch\_result = garch\_model.fit()

print(garch\_result.summary())





Bảng 3.5: Mô hình Garch

**Nhận xét:**

1. Mức độ phù hợp mô hình:
   * Mô hình GARCH(1,1) có beta[1] rất lớn (0.9474) và alpha[1] rất nhỏ (0.0142).  
     → Phương sai có điều kiện chủ yếu phụ thuộc vào thành phần GARCH (độ trễ phương sai), trong khi thành phần ARCH (độ trễ sai số bình phương) không đáng kể.
2. Ý nghĩa kinh tế:
   * Tỷ suất sinh lợi hàng ngày có phương sai thay đổi theo thời gian.
   * Sự biến động của tỷ suất sinh lợi trong quá khứ có ảnh hưởng lớn đến phương sai hiện tại (thành phần GARCH chiếm ưu thế).
3. Ứng dụng:
   * Kết quả này hữu ích trong việc dự báo biến động giá và quản lý rủi ro, đặc biệt là trong lĩnh vực tài chính và đầu tư.

Biểu đồ phân tích mô hình biến động giá:

import matplotlib.pyplot as plt

# Lấy giá trị độ biến động dự đoán (Conditional Volatility)

conditional\_volatility = garch\_result.conditional\_volatility

# Vẽ biểu đồ

plt.figure(figsize=(14, 7))

# Biểu đồ Tỷ suất sinh lợi thực tế

plt.plot(returns.index, returns, label="Tỷ suất sinh lợi thực tế (%)", color='blue', alpha=0.6)

# Biểu đồ Độ biến động dự đoán

plt.plot(returns.index, conditional\_volatility, label="Độ biến động dự đoán (GARCH)", color='red', linewidth=2)

# Tùy chỉnh biểu đồ

plt.title("Phân tích mô hình biến động giá - GARCH(1,1)", fontsize=16)

plt.xlabel("Ngày", fontsize=12)

plt.ylabel("Tỷ suất sinh lợi (%) / Độ biến động", fontsize=12)

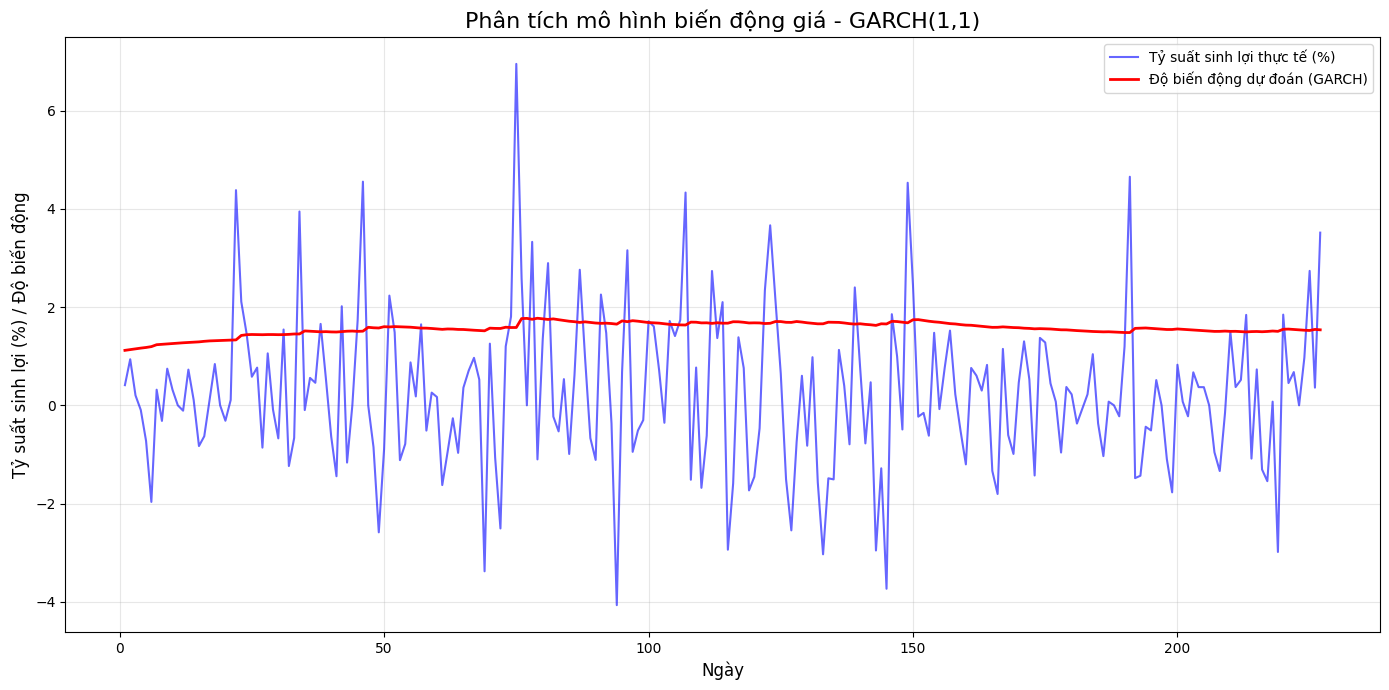
plt.legend()

plt.grid(alpha=0.3)

plt.tight\_layout()

# Hiển thị biểu đồ

plt.show()



Hình 3.13: Biểu đồ phân tích mô hình biến động giá

garch\_result.conditional\_volatility: Đây là chuỗi phương sai có điều kiện đã được ước lượng từ mô hình GARCH(1,1). Nó đại diện cho độ biến động được dự đoán.

**Nhận xét:**

Tỷ suất sinh lợi thực tế (đường màu xanh):

* Đường màu xanh thể hiện sự biến động mạnh mẽ và ngẫu nhiên của tỷ suất sinh lợi thực tế theo thời gian. Có những giai đoạn có biên độ dao động lớn, phản ánh sự bất ổn của thị trường.

Độ biến động dự đoán từ GARCH (đường màu đỏ):

* Đường màu đỏ khá mượt và thể hiện xu hướng của độ biến động theo thời gian.
* Độ biến động dự đoán (GARCH) tăng nhẹ tại một số giai đoạn có sự tăng vọt mạnh mẽ của tỷ suất sinh lợi thực tế (ví dụ, ở giữa hoặc cuối biểu đồ). Điều này cho thấy mô hình GARCH phản ứng với sự thay đổi lớn trong biến động giá.
* GARCH giúp dự đoán trước những giai đoạn có độ biến động cao hơn thông qua sự gia tăng của đường đỏ.

Sự tương quan giữa hai đường:

* Khi tỷ suất sinh lợi thực tế có những biến động lớn (đỉnh và đáy cao), độ biến động dự đoán từ GARCH có xu hướng tăng nhẹ, mặc dù không phản ứng ngay lập tức. Điều này là do mô hình GARCH(1,1) dựa vào phương sai có điều kiện và lịch sử biến động để điều chỉnh dự đoán.

### 3.3.8.Hồi quy tuyến tính.

* Phân tích hồi quy đơn biến: Dùng để ước lượng mối quan hệ giữa giá đóng cửa của cổ phiếu và khối lượng giao dịch. Đây là một phương pháp để đánh giá liệu có sự phụ thuộc giữa giá và khối lượng hay không.
* Ý nghĩa: Đánh giá mối quan hệ giữa giá đóng cửa và khối lượng giao dịch.

import pandas as pd

import statsmodels.api as sm

import matplotlib.pyplot as plt

# Giả sử df là DataFrame chứa dữ liệu

# Lấy dữ liệu giá đóng cửa (Y) và khối lượng giao dịch (X)

Y = df["close"]  # Giá đóng cửa

X = df["volume"]  # Khối lượng giao dịch

# Loại bỏ NaN

data = pd.DataFrame({"close": Y, "volume": X}).dropna()

# Thêm hệ số chặn vào X để phục vụ hồi quy

X = sm.add\_constant(data["volume"])  # Thêm cột "1" cho hệ số chặn

# Xây dựng mô hình hồi quy

model = sm.OLS(data["close"], X).fit()

# Hiển thị kết quả

print(model.summary())

# Vẽ đồ thị

plt.scatter(data["volume"], data["close"], color='blue', alpha=0.5, label='Dữ liệu thực tế')

plt.plot(data["volume"], model.predict(X), color='red', label='Đường hồi quy')

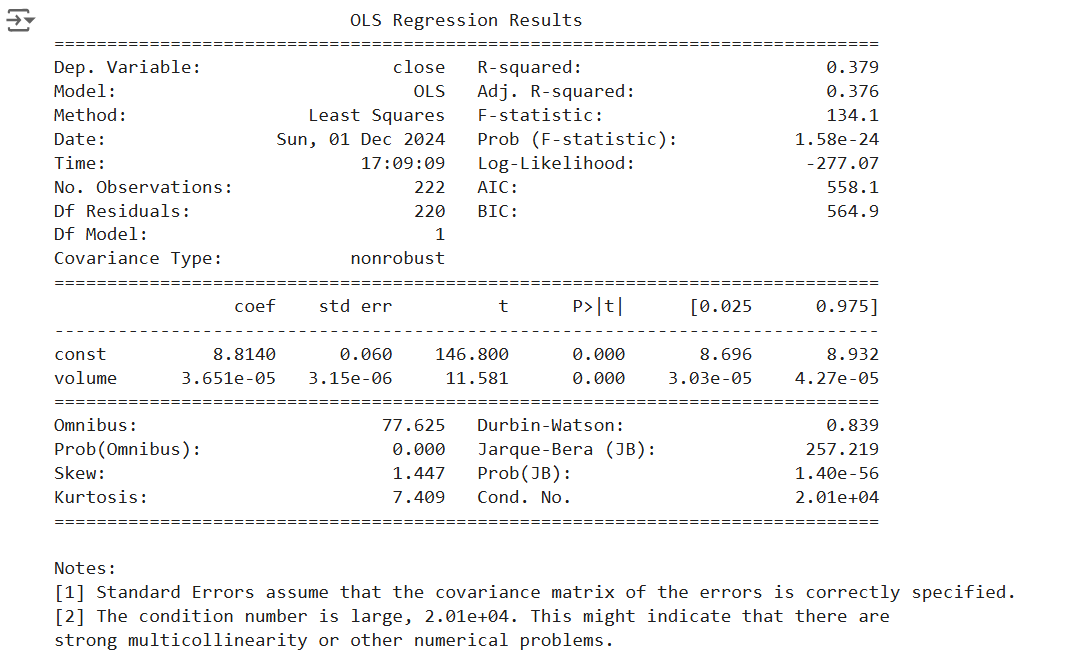
plt.xlabel("Khối lượng giao dịch")

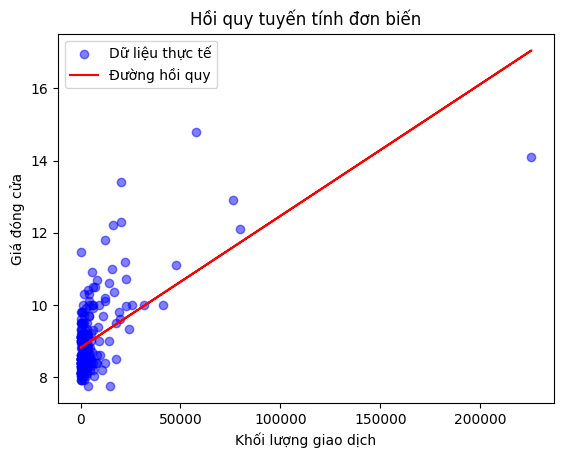
plt.ylabel("Giá đóng cửa")

plt.title("Hồi quy tuyến tính đơn biến")

plt.legend()

plt.show()  
# Kết quả hiển thị:





Hình 3.14: Biểu đồ Hồi quy tuyến tính đơn biến.

**Nhận xét:**

Mô hình hồi quy tuyến tính đơn biến chỉ ra mối quan hệ dương giữa khối lượng giao dịch và giá đóng cửa.

Biểu đồ tán xạ (Scatter Plot):

* + Các điểm dữ liệu (màu xanh) thể hiện mối quan hệ giữa khối lượng giao dịch và giá đóng cửa.
  + Phần lớn các điểm dữ liệu tập trung ở mức khối lượng giao dịch thấp, và giá đóng cửa cũng chủ yếu nằm trong khoảng 8-10.

Đường hồi quy (màu đỏ):

* + Đường hồi quy cho thấy mối quan hệ tuyến tính dương giữa khối lượng giao dịch và giá đóng cửa.
  + Tuy nhiên, có một số điểm dữ liệu nằm xa đường hồi quy, thể hiện các ngoại lệ và độ nhiễu trong dữ liệu.

Vấn đề tiềm ẩn:

* + Với R-squared thấp và phần dư phân phối không chuẩn, mô hình này có thể chưa thực sự tốt trong việc giải thích toàn bộ sự biến động của giá đóng cửa. Các yếu tố khác ngoài khối lượng giao dịch có thể ảnh hưởng đáng kể đến giá đóng cửa.

# CHƯƠNG 4: PHÂN TÍCH MỐI QUAN HỆ GIỮA CÁC BIẾN

## 4.1. BIỂU ĐỒ

Scatter Plot (biểu đồ phân tán) là công cụ trực quan hóa để phân tích mối liên hệ giữa hai biến định lượng:

* Giá đóng cửa (Closing price) và khối lượng giao dịch (Volume).
* Giá mở cửa (Opening price) và giá đóng cửa (Closing price).

Scatter Plot có thể được sử dụng để phân tích mối quan hệ giữa các cặp biến như:

* Giá cổ phiếu (Price) và khối lượng giao dịch (Volume).
* Tỷ suất sinh lời hàng ngày (Daily Return) và khối lượng giao dịch.
* Tỷ suất sinh lời và tin tức thị trường (nếu dữ liệu cảm xúc thị trường được định lượng).

Nếu hai biến có mối quan hệ tuyến tính (tức là khi giá trị của một biến tăng, giá trị của biến kia cũng tăng), biểu đồ phân tán sẽ thể hiện một đường xu hướng rõ ràng. Nếu không có mối quan hệ, các điểm sẽ phân tán ngẫu nhiên trên biểu đồ.

Biểu đồ (scatter plot) cũng dùng để vẽ đường biên hiệu quả phân tích mỗi quan hệ giữa các biến độ biến động và lợi nhuận kỳ vọng. Màu sắc của các điểm trên biểu đồ đại diện cho tỷ lệ Sharpe đo lường hiệu suất điều chỉnh theo rủi ro của mỗi danh mục đầu tư.

Ma trận tương quan (Correlation matrix):

Ma trận tương quan được sử dụng để hiển thị hệ số tương quan giữa nhiều biến số với nhau trong cùng một bảng.

Ma trận tương quan có thể bao gồm các biến:

* Giá mở cửa (Open Price).
* Giá cao nhất (High Price).
* Giá thấp nhất (Low Price).
* Giá đóng cửa (Close Price).
* Khối lượng giao dịch (Volume).
* Tỷ suất sinh lời hàng ngày (Daily Return). Kết quả xác định các biến có mối quan hệ tương quan mạnh hoặc yếu với nhau

import matplotlib.pyplot as plt

import seaborn as sns

# Biểu đồ phân tán giữa giá đóng cửa và khối lượng giao dịch

plt.figure(figsize=(12,6))

sns.scatterplot(data=df\_fpt, color='blue', alpha=0.6)

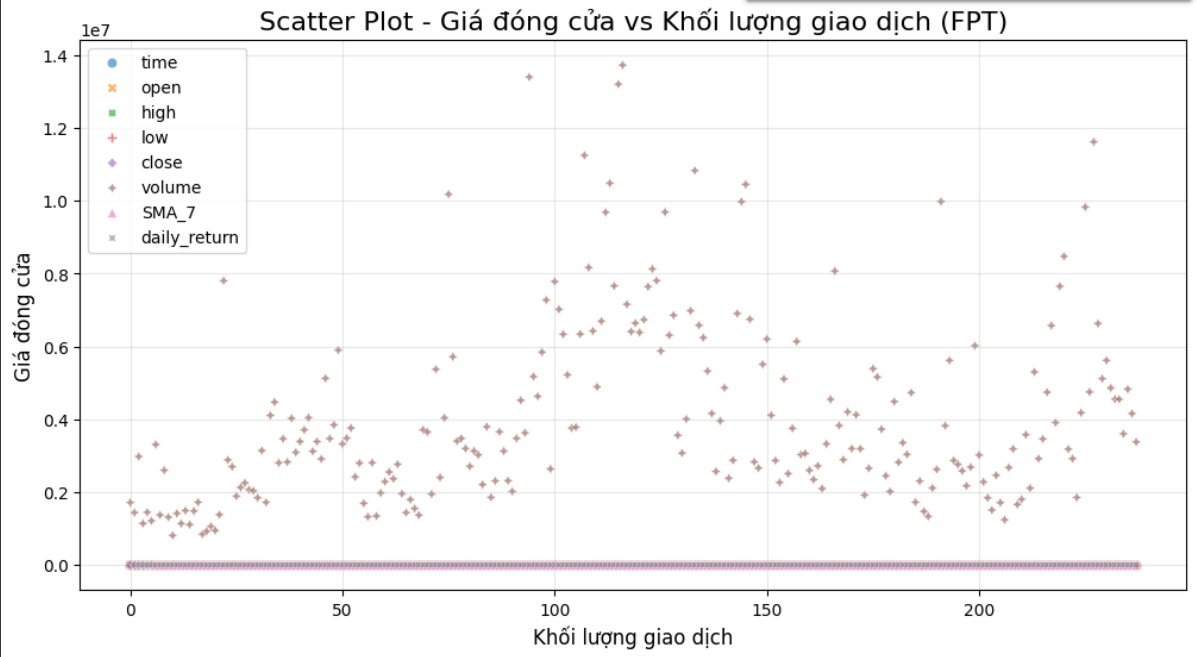
plt.title("Scatter Plot - Giá đóng cửa vs Khối lượng giao dịch (FPT)", fontsize=16)

plt.xlabel("Khối lượng giao dịch", fontsize=12)

plt.ylabel("Giá đóng cửa", fontsize=12)

plt.grid(alpha=0.3)

plt.show()



Hình 4.1: Biểu đồ phân tán giá đóng cửa và khối lượng giao dịch

Biểu đồ trên là biểu đồ tán xạ thể hiện mối quan hệ giữa Giá đóng cửa (trục tung) và Khối lượng giao dịch (trục hoành) của cổ phiếu FPT. Dưới đây là một số nhận xét:

* 1. Phân bổ dữ liệu:

Các điểm dữ liệu nằm rải rác, không có xu hướng rõ ràng. • Phần lớn giá trị khối lượng giao dịch tập trung ở khoảng từ 0 đến 250.

Giá đóng cửa dao động chủ yếu trong khoảng từ 0 đến 0.4 x 10^7.

2. Xu hướng:

Không có xu hướng rõ ràng giữa giá đóng cửa và khối lượng giao dịch. Điều này có thể cho thấy mối tương quan giữa hai yếu tố này là yếu hoặc không đáng kể.

Một số điểm dữ liệu có giá trị khối lượng giao dịch cao (>150) nhưng không tương ứng với giá đóng cửa cao.

1. Giá trị ngoại lệ:

Có một số điểm nổi bật với giá đóng cửa khá cao (1.0 x 10^7 đến 1.4 x 10^7) trong khi khối lượng giao dịch không lớn.

Các điểm này có thể là giá trị ngoại lệ hoặc biểu thị những phiên giao dịch đặc biệt.

4. Đường SMA\_7:

Đường màu tím nằm ngang ở phần dưới cùng của biểu đồ, biểu diễn trung bình động 7 ngày (SMA\_7), có vẻ như không biến động lớn.

Kết luận: Biểu đồ này cho thấy khối lượng giao dịch và giá đóng cửa của cổ phiếu FPT không có mối tương quan mạnh mẽ. Cần có thêm phân tích để làm rõ hơn tác động của các yếu tố khác đối với biến động giá và khối lượng.

import matplotlib.pyplot as plt

import seaborn as sns

# Biểu đồ phân tán giữa giá đóng cửa và khối lượng giao dịch

plt.figure(figsize=(12, 6))

# Check if the columns exist before plotting

if 'volume' in df\_fpt.columns and 'close' in df\_fpt.columns:

    sns.scatterplot(data=df\_fpt, x='volume', y='close', color='blue', alpha=0.6)

    plt.title("Scatter Plot - Giá đóng cửa vs Khối lượng giao dịch (FPT)", fontsize=16)

    plt.xlabel("Khối lượng giao dịch", fontsize=12)

    plt.ylabel("Giá đóng cửa", fontsize=12)

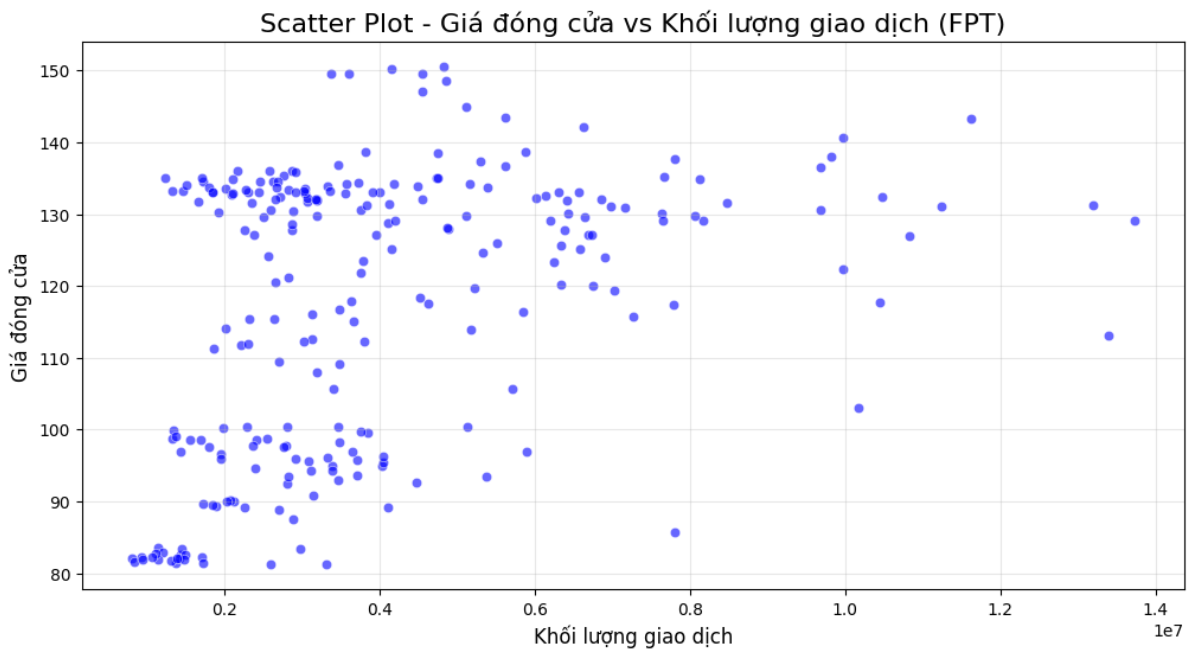
    plt.grid(alpha=0.3)

    plt.show()

else:

    print("Error: 'volume' or 'close' column not found in the DataFrame.")

    print("Available columns:", df\_fpt.columns)



Hình4.2: Biểu đồ phân tán giữa giá đóng cửa và khối lượng giao dịch

Dưới đây là nhận xét về biểu đồ Scatter Plot - Giá đóng cửa vs Khối lượng giao dịch (FPT):

1. Phân bố dữ liệu:

Dữ liệu phân tán rộng và không có xu hướng rõ ràng giữa giá đóng cửa và khối lượng giao dịch.

Các điểm dữ liệu tập trung dày đặc ở mức khối lượng giao dịch nhỏ hơn 6 triệu và giá đóng cửa nằm trong khoảng từ 80 đến 140.

1. Mối quan hệ giữa hai biến:

Không có mối tương quan tuyến tính rõ rệt giữa khối lượng giao dịch và giá đóng cửa.

Khối lượng giao dịch cao hơn không nhất thiết đi kèm với giá đóng cửa cao hay thấp.

1. Cụm dữ liệu:

Có một số cụm dữ liệu rõ rệt:

* Ở mức giá đóng cửa từ 130 đến 140, khối lượng giao dịch tập trung nhiều ở dưới 4 triệu.
* Ở mức giá đóng cửa thấp hơn (từ 80 đến 100), khối lượng giao dịch cũng tương đối thấp.

1. Giá trị cực trị:

Một số điểm có khối lượng giao dịch lớn (hơn 10 triệu) và nằm ở vùng giá đóng cửa từ khoảng 130 đến 140, nhưng đây là các điểm ngoại lệ.

Kết luận: Biểu đồ cho thấy không có mối quan hệ rõ ràng giữa giá đóng cửa và khối lượng giao dịch của cổ phiếu FPT. Dữ liệu phân tán và tập trung chủ yếu ở vùng khối lượng giao dịch thấp, giá đóng cửa trong khoảng 80 đến 140.

print(df\_fpt.columns)

Index(['time', 'open', 'high', 'low', 'close', 'volume', 'SMA\_7',

'daily\_return'],

dtype='object')

# Biểu đồ phân tán giữa giá mở cửa và giá đóng cửa

plt.figure(figsize=(12, 6))

sns.scatterplot(data=df\_fpt, x='open', y='close', color='green', alpha=0.6)

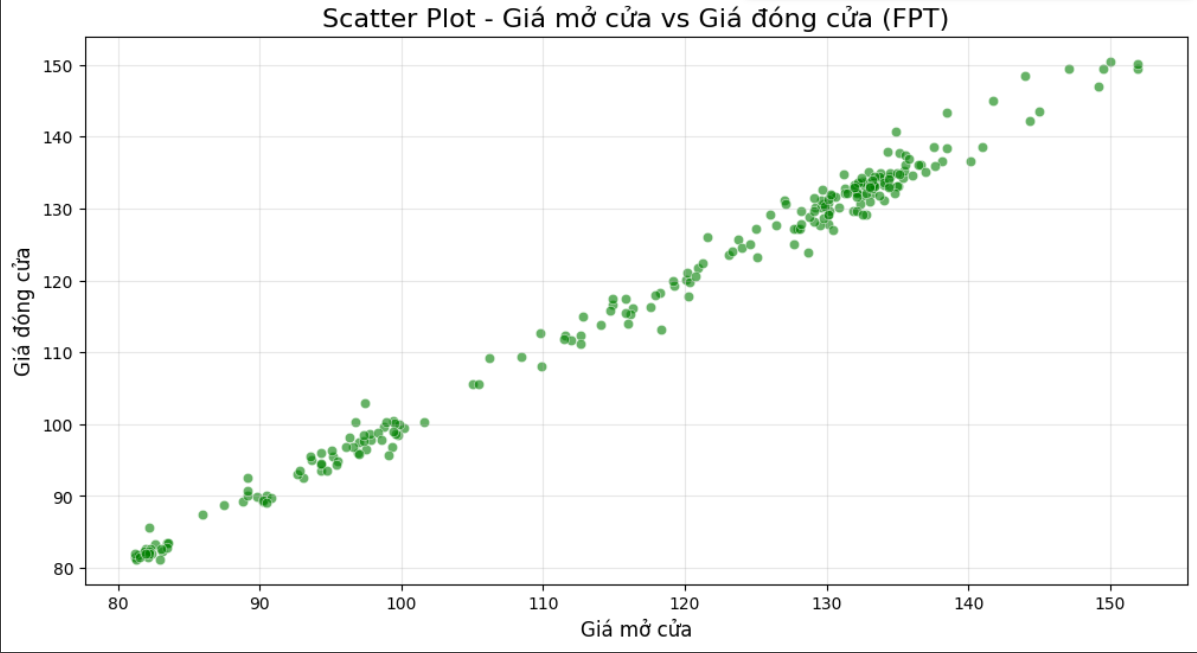
plt.title("Scatter Plot - Giá mở cửa vs Giá đóng cửa (FPT)", fontsize=16)

plt.xlabel("Giá mở cửa", fontsize=12)

plt.ylabel("Giá đóng cửa", fontsize=12)

plt.grid(alpha=0.3)

plt.show()



Hình 4.3: Biểu đồ phân tán giữa giá mở cửa và giá đóng cửa

Biểu đồ phân tán này thể hiện mối quan hệ giữa giá mở cửa và giá đóng cửa của cổ phiếu FPT. Dưới đây là một số nhận xét:

1. Xu hướng tương quan dương:

Có thể thấy các điểm dữ liệu tập trung thành một đường chéo đi lên từ trái sang phải. Điều này cho thấy mối tương quan thuận giữa giá mở cửa và giá đóng cửa, nghĩa là khi giá mở cửa tăng thì giá đóng cửa cũng có xu hướng tăng.

1. Mức độ tập trung của dữ liệu:

Các điểm dữ liệu khá tập trung gần một đường thẳng, cho thấy mối quan hệ tuyến tính mạnh mẽ giữa hai biến này.

1. Biến động nhỏ:

Có một số điểm nằm hơi xa so với đường xu hướng, nhưng nhìn chung các điểm dữ liệu không có sự biến động quá lớn. Điều này chứng tỏ giá mở cửa và giá đóng cửa không có sự chênh lệch đáng kể trong ngày.

1. Phạm vi giá trị:

Giá mở cửa dao động trong khoảng từ 80 đến 150.

Giá đóng cửa dao động trong khoảng từ 80 đến 150.

Kết luận: Biểu đồ này chỉ ra rằng giá mở cửa và giá đóng cửa có mối tương quan dương rất rõ ràng. Điều này có nghĩa là xu hướng giá trong ngày thường ổn định và không có biến động lớn giữa giá mở và giá đóng cửa.

Tính ma trận tương quan:

correlation\_matrix = df\_fpt[['open', 'high', 'low', 'close', 'volume']].corr()

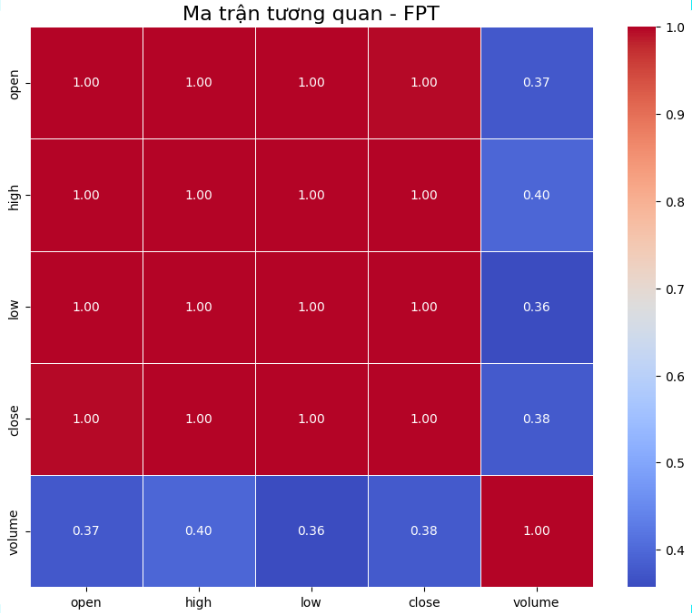
# Vẽ heatmap để trực quan hóa ma trận tương quan

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, fmt=".2f", cmap="coolwarm", linewidths=0.5)

plt.title("Ma trận tương quan - FPT", fontsize=16)

plt.show()



Hình 4.4: Biểu đồ ma trận tương quan

Nhận xét về Ma trận tương quan - FPT:

1. Mối tương quan giữa các biến:

Các biến open (giá mở cửa), high (giá cao nhất), low (giá thấp nhất), và close (giá đóng cửa) đều có hệ số tương quan gần bằng 1.00.

Điều này cho thấy mối quan hệ tuyến tính hoàn hảo giữa các biến này. Khi một biến tăng hoặc giảm, các biến còn lại cũng thay đổi tương tự.

1. Tương quan giữa khối lượng giao dịch (volume) và các biến còn lại:

Khối lượng giao dịch (volume) có tương quan thấp với các biến giá:

* + Với open: 0.37
  + Với high: 0.40
  + Với low: 0.36
  + Với close: 0.38

Điều này cho thấy khối lượng giao dịch và các biến giá không có mối quan hệ mạnh, tức là khối lượng giao dịch không ảnh hưởng đáng kể đến giá cổ phiếu và ngược lại.

1. Tổng quan:

Biểu đồ này cho thấy các biến giá (open, high, low, close) có mối tương quan cực kỳ mạnh với nhau (gần 1.00). Điều này có thể giải thích do các giá trị này thường biến động theo cùng một xu hướng trong ngày.

Tuy nhiên, volume chỉ có mối tương quan yếu với các biến giá, phản ánh rằng khối lượng giao dịch có thể bị ảnh hưởng bởi nhiều yếu tố khác ngoài giá cổ phiếu.

Kết luận:

* Giá mở cửa, giá cao nhất, giá thấp nhất và giá đóng cửa có mối quan hệ tuyến tính rất chặt chẽ.
* Khối lượng giao dịch và giá cổ phiếu có mối tương quan thấp, chứng tỏ khối lượng giao dịch không phải là yếu tố quyết định chính đến giá cổ phiếu FPT.

# Tạo dataframe chứa giá đóng cửa của 3 cổ phiếu

close\_prices = bank\_stocks.xs(key='close', axis=1, level='Thông tin cổ phiếu')

# Tính ma trận tương quan

correlation\_matrix\_all = close\_prices.corr()

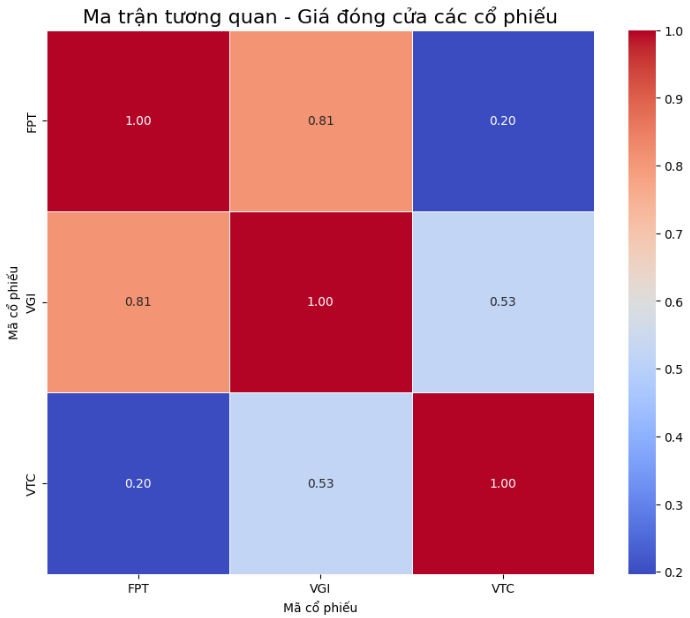
# Vẽ heatmap

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix\_all, annot=True, fmt=".2f", cmap="coolwarm", linewidths=0.5)

plt.title("Ma trận tương quan - Giá đóng cửa các cổ phiếu", fontsize=16)

plt.show()



Hình 4.5: Biểu đồ tính ma trận tương quan

Biểu đồ trên là ma trận tương quan thể hiện mối quan hệ giữa giá đóng cửa của các mã cổ phiếu FPT, VGI và VTC. Dưới đây là một số nhận xét chi tiết:

1. Ý nghĩa các hệ số tương quan:

Hệ số tương quan nằm trong khoảng từ -1 đến 1:

1.0: Tương quan thuận hoàn hảo.

0.0: Không có mối tương quan.

-1.0: Tương quan nghịch hoàn hảo.

Giá trị tương quan càng gần 1.0 hoặc -1.0, mối quan hệ càng chặt chẽ.

1. Nhận xét chi tiết:

FPT và VGI: Hệ số tương quan là 0.81

→ Hai cổ phiếu này có mối tương quan thuận mạnh, tức là khi giá đóng cửa của FPT tăng, giá đóng cửa của VGI cũng có xu hướng tăng.

FPT và VTC: Hệ số tương quan là 0.20

→ Mối tương quan giữa hai cổ phiếu này là rất yếu, gần như không có mối quan hệ đáng kể.

VGI và VTC: Hệ số tương quan là 0.53

→ Có mối tương quan thuận trung bình, tức là có sự ảnh hưởng nhất định giữa giá đóng cửa của hai cổ phiếu này.

1. Tổng kết mối quan hệ:

FPT và VGI có mối quan hệ chặt chẽ nhất trong ma trận tương quan.

FPT và VTC gần như độc lập với nhau.

VGI và VTC có sự liên hệ ở mức trung bình.

Kết luận: Biểu đồ này giúp xác định mức độ tương quan giữa giá đóng cửa của các cổ phiếu. Nhà đầu tư có thể dựa vào thông tin này để đánh giá rủi ro hoặc xây dựng danh mục đầu tư tối ưu dựa trên các mối tương quan.

## 4.2. HỆ SỐ TƯƠNG QUAN VÀ ĐỘ ĐỒNG BIẾN

Hệ số tương quan (Correlation Coefficient):

* Hệ số tương quan Pearson là một giá trị từ -1 đến 1, thể hiện mức độ mối liên hệ tuyến tính giữa hai biến:
* r = 1: Tương quan dương hoàn hảo (hai biến cùng tăng hoặc cùng giảm).
* r = -1: Tương quan âm hoàn hảo (một biến tăng thì biến kia giảm).
* r = 0: Không có mối liên hệ tuyến tính giữa hai biến.

Kết quả:

* Giá trị lớn hơn 0: Cho thấy mối tương quan dương giữa hai biến.
* Giá trị nhỏ hơn 0: Cho thấy mối tương quan âm giữa hai biến.

-Giá trị gần 0: Mối liên hệ tuyến tính giữa hai biến không mạnh.

Độ đồng biến (Covariance):

* Covariance đo lường mức độ hai biến cùng biến động với nhau. Kết quả:
* Cov > 0: Hai biến có xu hướng cùng tăng hoặc cùng giảm (đồng biến).
* Cov < 0: Khi một biến tăng thì biến kia giảm (ngược biến).
* Cov = 0: Không có sự liên quan giữa hai biến.

# Tính hệ số tương quan giữa giá đóng cửa và khối lượng giao dịch

correlation = df\_fpt['close'].corr(df\_fpt['volume'])

print(f"Hệ số tương quan Pearson giữa giá đóng cửa và khối lượng giao dịch: {correlation:.2f}")

Hệ số tương quan Pearson giữa giá đóng cửa và khối lượng giao dịch: 0.38

# Tính độ đồng biến giữa giá đóng cửa và khối lượng giao dịch

covariance = df\_fpt['close'].cov(df\_fpt['volume'])

print(f"Độ đồng biến giữa giá đóng cửa và khối lượng giao dịch: {covariance:.2f}")

Độ đồng biến giữa giá đóng cửa và khối lượng giao dịch: 18712145.08

# Ma trận hệ số tương quan

correlation\_matrix = df\_fpt[['open', 'high', 'low', 'close', 'volume']].corr()

print("Ma trận hệ số tương quan:")

print(correlation\_matrix)

Ma trận hệ số tương quan:

open high low close volume

open 1.000000 0.998025 0.998659 0.996362 0.374795

high 0.998025 1.000000 0.998152 0.998562 0.396848

low 0.998659 0.998152 1.000000 0.998249 0.357091

close 0.996362 0.998562 0.998249 1.000000 0.379583

volume 0.374795 0.396848 0.357091 0.379583 1.000000

# Ma trận độ đồng biến

covariance\_matrix = df\_fpt[['open', 'high', 'low', 'close', 'volume']].cov()

print("\nMa trận độ đồng biến:")

print(covariance\_matrix)

Ma trận độ đồng biến:

open high low close volume

open 3.871230e+02 3.892461e+02 3.821214e+02 3.848367e+02 1.851829e+07

high 3.892461e+02 3.929314e+02 3.847820e+02 3.885690e+02 1.975445e+07

low 3.821214e+02 3.847820e+02 3.781983e+02 3.810952e+02 1.743898e+07

close 3.848367e+02 3.885690e+02 3.810952e+02 3.853629e+02 1.871215e+07

volume 1.851829e+07 1.975445e+07 1.743898e+07 1.871215e+07 6.306153e+12

import seaborn as sns

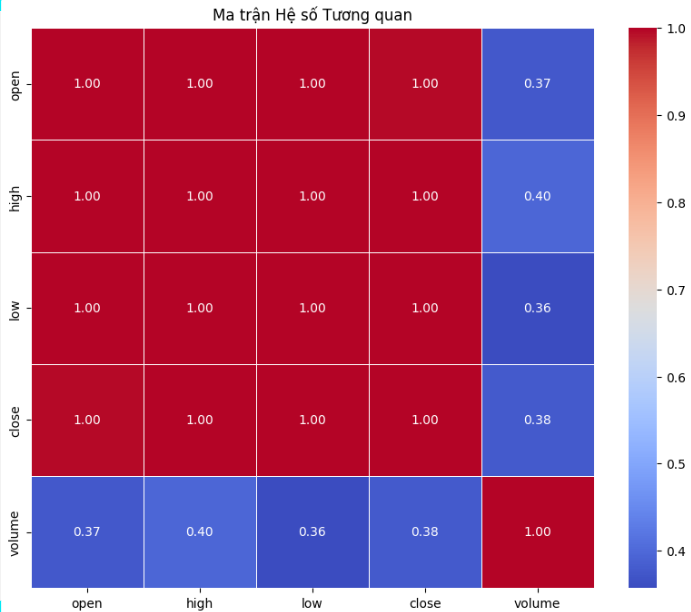
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, fmt=".2f", cmap="coolwarm", linewidths=0.5)

plt.title("Ma trận Hệ số Tương quan")

plt.show()



Hình 4.6: Biểu đồ ma trận hệ số tương quan

Biểu đồ trên là ma trận hệ số tương quan thể hiện mối quan hệ giữa các biến số: open (giá mở cửa), high (giá cao nhất), low (giá thấp nhất), close (giá đóng cửa), và volume (khối lượng giao dịch). Dưới đây là một số nhận xét chi tiết:

* 1. Tương quan giữa giá mở cửa, cao nhất, thấp nhất và đóng cửa:

Các hệ số tương quan giữa open, high, low, close đều bằng 1.00.

→ Điều này cho thấy 4 biến này có tương quan hoàn hảo với nhau. Khi một biến thay đổi, các biến còn lại cũng thay đổi theo cùng một xu hướng.

* 1. Tương quan giữa khối lượng giao dịch (volume) và các biến còn lại:

Volume và open: Hệ số tương quan là 0.37.

Volume và high: Hệ số tương quan là 0.40.

Volume và low: Hệ số tương quan là 0.36.

Volume và close: Hệ số tương quan là 0.38.

→ Các hệ số này đều nằm trong khoảng từ 0.36 đến 0.40, thể hiện mối tương quan yếu giữa khối lượng giao dịch và các biến giá. Mặc dù có sự liên hệ, nhưng khối lượng giao dịch không ảnh hưởng mạnh đến các mức giá mở cửa, cao nhất, thấp nhất hay đóng cửa.

* 1. Tổng kết:

Giá mở cửa, cao nhất, thấp nhất và đóng cửa có mối tương quan hoàn hảo (1.00), cho thấy các mức giá này di chuyển cùng chiều.

Khối lượng giao dịch có mối tương quan yếu với các biến giá, cho thấy khối lượng không có ảnh hưởng lớn đến sự thay đổi của các mức giá.

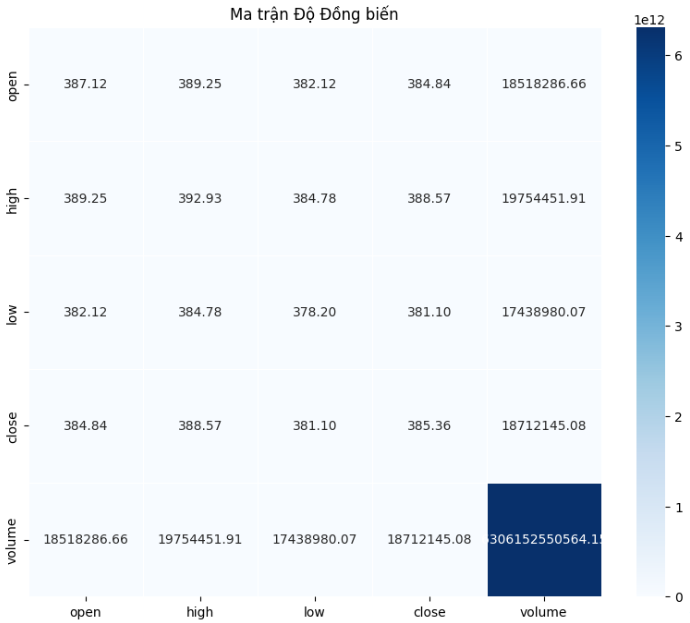
Kết luận: Biểu đồ cho thấy các mức giá có mối quan hệ chặt chẽ với nhau, trong khi khối lượng giao dịch không đóng vai trò quyết định trong biến động của giá. Nhà đầu tư có thể dựa vào thông tin này để đánh giá mối liên hệ giữa khối lượng và giá trong các quyết định giao dịch.

plt.figure(figsize=(10, 8))

sns.heatmap(covariance\_matrix, annot=True, fmt=".2f", cmap="Blues", linewidths=0.5)

plt.title("Ma trận Độ Đồng biến")

plt.show()



Hình 4.7: Biểu đồ ma trận độ đồng biến

Biểu đồ trên là ma trận độ đồng biến (ma trận hiệp phương sai) của các chỉ số liên quan đến giá chứng khoán (open, high, low, close) và khối lượng giao dịch (volume).

Nhận xét:

1. Các chỉ số giá (open, high, low, close):

Có giá trị đồng biến khá cao giữa các chỉ số giá với nhau.

Ví dụ:

• Open với High: 389.25

• Open với Close: 384.84

• Điều này phản ánh mối quan hệ tích cực giữa các mức giá trong phiên giao dịch.

1. Khối lượng giao dịch (volume):

Khối lượng giao dịch có giá trị đồng biến rất lớn (tính theo lũy thừa 10^12) khi so sánh với các chỉ số giá, ví dụ:

• Volume với Open: 1.85e+12

• Volume với High: 1.97e+12

• Điều này cho thấy khối lượng giao dịch không có mối tương quan mạnh như các chỉ số giá, nhưng vẫn đóng vai trò quan trọng trong dữ liệu.

1. Tổng thể:

• Ma trận này chủ yếu cho thấy mối tương quan cao giữa các giá trị giá chứng khoán (mở cửa, cao nhất, thấp nhất và đóng cửa).

• Khối lượng giao dịch (volume) lại có độ lớn rất khác biệt, làm nổi bật lên trong ma trận (màu xanh đậm).

Tóm lại, giá chứng khoán trong phiên có mối quan hệ chặt chẽ, trong khi khối lượng giao dịch mang tính chất khác biệt và có độ lớn vượt trội.

# CHƯƠNG 5: TRỰC QUAN HÓA DỮ LIỆU

## 5.1. GIỚI THIỆU VỀ TRỰC QUAN HÓA DỮ LIỆU

### 5.1.1. Trực quan hóa dữ liệu là gì ?

Điều gì có thể diễn đạt những trải nghiệm, ý tưởng và đặc biệt những con số “biết nói” tốt hơn ngôn từ đó chính là những hình ảnh bức tranh bởi hình ảnh có thể mang lại sự phong phú, sâu sắc và phản ánh được một phần của thế giới và cảm xúc của con người một cách không thể diễn tả hoàn toàn bằng từ ngữ. Trong một số trường hợp, hình ảnh có thể tạo ra một ấn tượng sâu sắc và không thể nào được truyền đạt bằng từ ngữ. Cũng như trích dẫn của Ludwig Wittgenstein, một nhà triết học Áo từng phát biểu rằng *"Một bức tranh, một bức ảnh có thể nói lên những điều không thể diễn đạt được bằng ngôn từ".*

Hiểu được tầm quan trọng của việc diễn đạt những con số đã được tổng hợp từ nhiều nguồn dữ liệu khác nhau là chìa khóa để có thể truyền tải thông tin một cách rõ ràng và hiệu quả. Những thế hệ đi trước đã dần dần sáng tạo ra khái niệm biểu đồ, đồ thị, bản đồ để từ đó giúp chúng ta giải được bài toán về sự giới hạn của ngôn từ đối với những con số của thống kê. Biểu đồ, đồ thị, bản đồ có thể đưa tới cho người xem, người đọc hiểu rõ hơn về bộ dữ liệu đó phải làm thế nào để họ hiểu, họ nắm được xem bộ dữ liệu đó nói về cái gì, miêu tả cái gì và rút ra điều gì từ bộ dữ liệu đó. Chữ viết hoàn toàn có thể sử dụng cho những bộ dữ liệu đơn giản nhưng không thể thực hiện được việc truyền tải nội dung về xu hướng và dao động của dữ liệu. Khi đó các công cụ về trực quan hóa dữ liệu trở nên cần thiết hơn bao giờ hết.

Vậy trực quan hóa dữ liệu là gì? Trực quan hóa dữ liệu là quá trình biểu diễn thông tin và dữ liệu dưới dạng hình ảnh hoặc đồ họa để giúp người xem hiểu và suy luận dữ liệu một cách dễ dàng và nhanh chóng hơn. Bằng cách sử dụng các biểu đồ, bản đồ, biểu đồ tia, biểu đồ cột và các công cụ trực quan hóa khác, trực quan dữ liệu giúp chuyển đổi dữ liệu phức tạp thành các hình ảnh đơn giản và dễ hiểu. Kết hợp với là một ngôn ngữ lập trình mạnh mẽ và phổ biến trong phân tích dữ liệu và trực quan hóa. Nó cung cấp nhiều gói và hàm mạnh mẽ để tạo ra các biểu đồ và đồ thị chất lượng cao.

### 5.1.2.Mục đích của trực quan hóa dữ liệu

Trực quan hóa dữ liệu đóng vai trò quan trọng trong quá trình phân tích và truyền tải thông tin. Một số mục đích chính của trực quan hóa dữ liệu đó là: hiểu dữ liệu, trình bày kết quả, ra quyết định, giao tiếp và thảo luận, tăng tương tác, giáo dục và đào tạo.

*Thứ nhất*, trực quan hóa dữ liệu giúp chúng ta hiểu sâu hơn về các mẫu, xu hướng và mối quan hệ tiềm ẩn trong dữ liệu. Khi xem xét các biểu đồ và đồ thị, chúng ta có thể nhận diện ngay lập tức các đặc điểm nổi bật mà có thể không rõ ràng khi chỉ xem xét các bảng số liệu đơn thuần. Việc khám phá này cho phép chúng ta phát hiện các ngoại lệ hoặc bất thường trong dữ liệu, kiểm tra các giả thuyết và thu thập thông tin chi tiết để làm rõ các hiện tượng phức tạp.

*Thứ hai,* một trong những mục đích chính của trực quan hóa dữ liệu là truyền tải thông tin một cách rõ ràng và ngắn gọn. Các biểu đồ và đồ thị giúp chúng ta trình bày kết quả phân tích một cách dễ hiểu, đặc biệt là khi đối tượng khán giả không có nền tảng kỹ thuật. Hình ảnh trực quan hấp dẫn không chỉ gây ấn tượng mạnh mẽ mà còn giúp thông tin dễ dàng được nhớ lâu hơn so với việc trình bày bằng các bảng số liệu khô khan.

*Thứ ba,* trực quan hóa dữ liệu đóng vai trò quan trọng trong việc hỗ trợ ra quyết định. Các biểu đồ và đồ thị cung cấp cái nhìn tổng quan và chi tiết về dữ liệu, giúp người quản lý và nhà phân tích đưa ra quyết định dựa trên dữ liệu thực tế. Khi các thông tin được trình bày một cách rõ ràng và dễ hiểu, nguy cơ sai sót trong quá trình ra quyết định được giảm thiểu, đồng thời tăng cường độ chính xác và hiệu quả của các quyết định chiến lược.

*Thứ tư,* trực quan hóa dữ liệu giúp giao tiếp thông tin một cách hiệu quả và dễ hiểu, tạo điều kiện cho các cuộc thảo luận và trao đổi thông tin. Các biểu đồ và đồ thị minh họa rõ ràng các điểm chính, giúp các bên liên quan dễ dàng hiểu và đồng thuận với các thông tin được trình bày. Điều này không chỉ giúp tăng cường sự minh bạch mà còn thúc đẩy quá trình hợp tác và thảo luận sâu hơn.

*Thứ năm,* với các công cụ trực quan hóa dữ liệu tương tác, người dùng có thể khám phá dữ liệu theo nhiều cách khác nhau, chẳng hạn như phóng to, thu nhỏ, và lọc dữ liệu. Khả năng tương tác này cho phép người dùng thực hiện các phân tích động, xem xét dữ liệu từ nhiều góc độ và tìm hiểu các chi tiết cụ thể. Trải nghiệm tương tác này không chỉ làm cho quá trình khám phá dữ liệu trở nên thú vị hơn mà còn tăng cường khả năng phát hiện những thông tin giá trị.

*Thứ sáu,* trực quan hóa dữ liệu là một công cụ mạnh mẽ trong giáo dục và đào tạo, giúp học sinh và sinh viên hiểu rõ các khái niệm phức tạp và ứng dụng thực tế của dữ liệu. Các biểu đồ và đồ thị giúp minh họa các quy trình và hệ thống phức tạp, làm cho việc giảng dạy và học tập trở nên trực quan và sinh động hơn. Điều này không chỉ giúp nâng cao hiệu quả học tập mà còn khuyến khích sự hứng thú và đam mê trong việc khám phá dữ liệu.

### 5.1.3. Các loại biểu đồ trực quan hóa:

Biểu đồ đường (Line Chart):

* Dữ liệu sử dụng: Giá cổ phiếu theo thời gian (giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất) và khối lượng giao dịch.
* Kết quả dự kiến: Hiển thị xu hướng tổng thể của các cổ phiếu, với các điểm nhấn như các giai đoạn giá cổ phiếu tăng mạnh hoặc giảm đột ngột, đồng thời so sánh được diễn biến giữa các mã cổ phiếu.

Biểu đồ cột (Bar Chart):

* Dữ liệu sử dụng: Khối lượng giao dịch hàng ngày hoặc theo tuần, tháng.
* Kết quả dự kiến: Hiển thị rõ ràng các phiên giao dịch với khối lượng lớn hoặc nhỏ hơn trung bình, giúp nhận diện những ngày có biến động giao dịch bất thường.

Biểu đồ phân tán (Scatter Plot):

* Dữ liệu sử dụng: Mối quan hệ giữa khối lượng giao dịch và giá cổ phiếu, hoặc giữa các chỉ số giá khác nhau như giá đóng cửa và giá mở cửa.
* Kết quả dự kiến: Biểu đồ này sẽ hiển thị rõ các cặp biến có mối liên hệ mật thiết, chẳng hạn như khi khối lượng giao dịch lớn có thể đi kèm với sự thay đổi mạnh về giá.

Biểu đồ nến Nhật (Candlestick Chart):

* Dữ liệu sử dụng: Giá cổ phiếu (mở cửa, đóng cửa, cao nhất, thấp nhất) trong một ngày hoặc một khoảng thời gian ngắn.
* Kết quả dự kiến: Biểu đồ nến cho phép quan sát các điểm đảo chiều, xu hướng tăng hoặc giảm của cổ phiếu, từ đó hỗ trợ việc dự đoán sự thay đổi giá trong tương lai.

Biểu đồ hình tròn (Pie Chart):

* Dữ liệu sử dụng: Tỷ trọng giao dịch hoặc vốn hóa của các cổ phiếu trong danh mục đầu tư.
* Kết quả dự kiến: Biểu đồ này sẽ hiển thị rõ tỷ trọng phân bổ của các cổ phiếu trong danh mục, giúp người xem có cái nhìn tổng quan về cơ cấu và sự phân bổ nguồn lực.

Biểu đồ vùng (Area Chart):

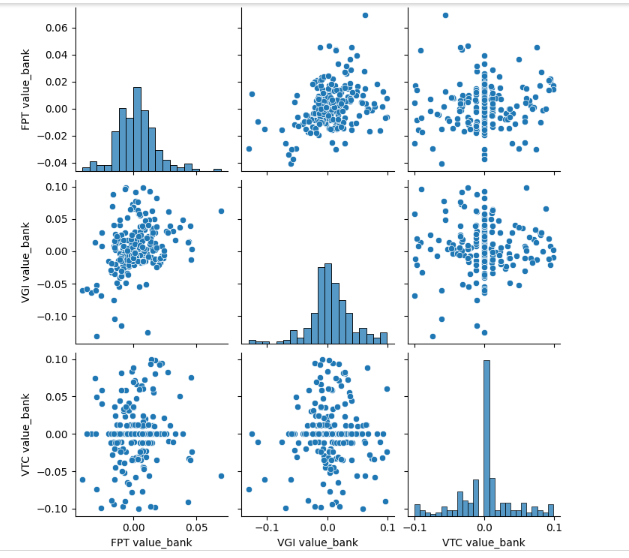
* Dữ liệu sử dụng: Giá cổ phiếu hoặc khối lượng giao dịch theo thời gian.

-Kết quả dự kiến: Biểu đồ vùng sẽ cho thấy sự biến động về khối lượng giao dịch hoặc giá cổ phiếu theo thời gian, giúp đánh giá sự thay đổi liên tục và tích lũy của các biến số này.

## 5.2. Trực quan hóa dữ liệu

import seaborn as sns

sns.pairplot(value\_banks)



Hình 5.1. biểu đồ ma trận phân tán

import matplotlib.pyplot as plt

import seaborn as sns

# Thiết lập phong cách của seaborn

sns.set\_theme(style="whitegrid") # Sử dụng whitegrid cho nền lưới màu xám

plt.figure(figsize=(15, 5)) # Tùy chỉnh kích thước biểu đồ

# Vẽ biểu đồ với viền cột màu trắng và cột màu xanh lá

sns.histplot(value\_banks['FPT value\_bank'],

color='green',

edgecolor='white', # Viền cột màu trắng

bins=100)

# Tùy chỉnh thêm

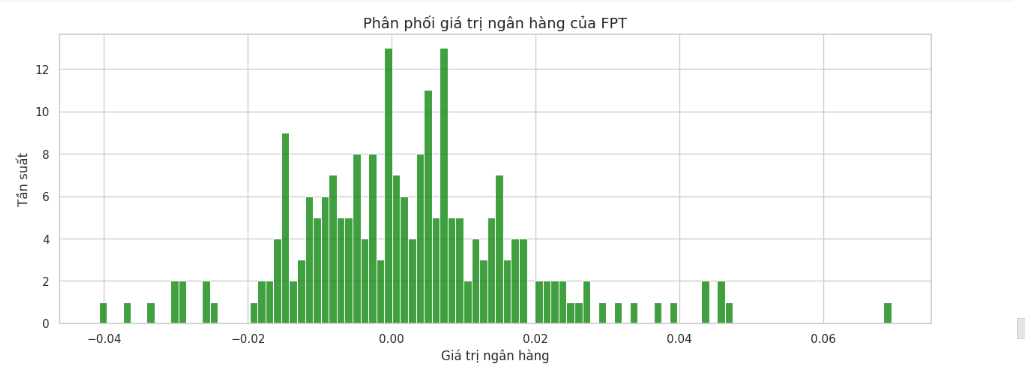
plt.title("Phân phối giá trị ngân hàng của FPT", fontsize=14)

plt.xlabel("Giá trị ngân hàng", fontsize=12)

plt.ylabel("Tần suất", fontsize=12)

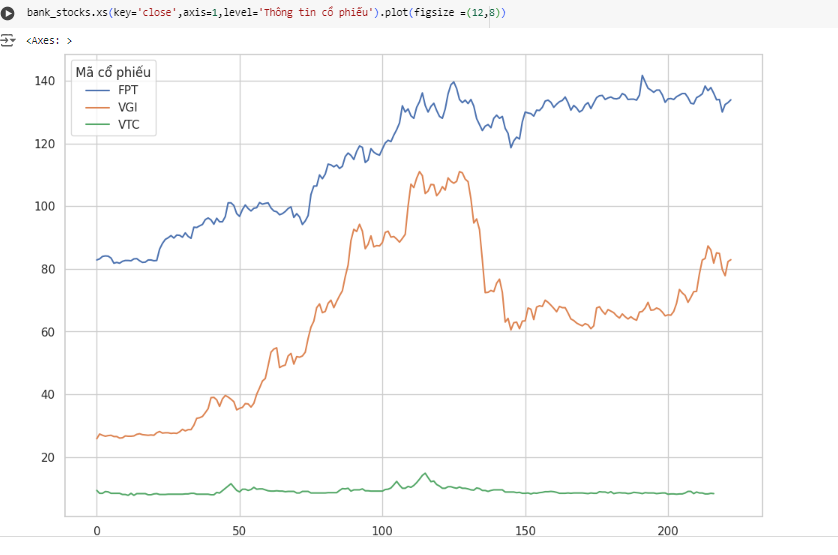
# Hiển thị biểu đồ

plt.show()



Hình 5.2. biểu đồ phân phối giá trị ngân hàng của FPT

bank\_stocks.xs(key='close',axis=1,level='Thông tin cổ phiếu').plot(figsize =(12,8))



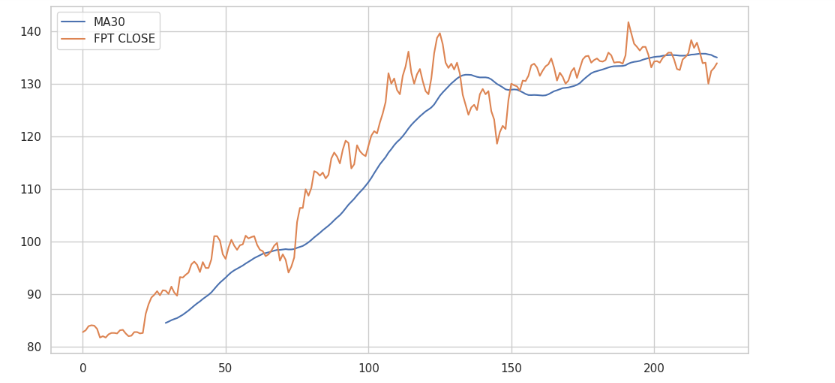
Hình 5.3. biểu đồ giá đóng cửa của các cổ phiếu FPT, VGI và VTC

plt.figure(figsize=(12,6))

df\_fpt['close'].rolling(window=30).mean().plot(label='MA30')

df\_fpt['close'].plot(label='FPT CLOSE')

plt.legend()



Hình 5.4: Biểu đồ giá đóng cửa và đường trung bình động MA30 của cổ phiếu FPT

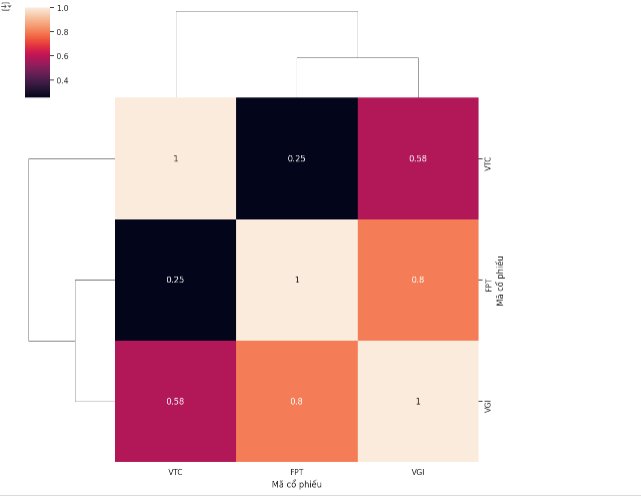
plt.figure(figsize=(12,5))

sns.heatmap(bank\_stocks.xs(key='close',axis=1,level='Thông tin cổ phiếu').corr(),annot=True)



Hình 5.5: Biểu đồ ma trận tương quan

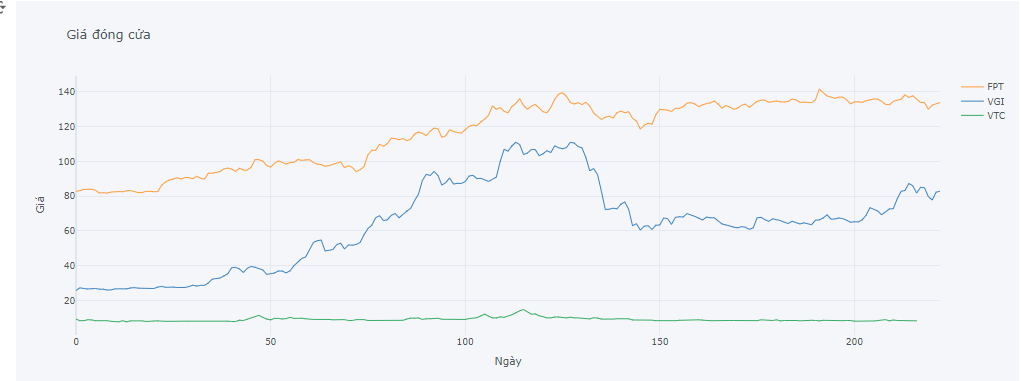
sns.clustermap(bank\_stocks.xs(key='close',axis=1,level='Thông tin cổ phiếu').corr(),annot=True)



Hình 5.6: Biểu đồ ma trận tương quan với dendrogram

bank\_stocks.xs(key='close',axis=1,level='Thông tin cổ phiếu').iplot(title = "Giá đóng cửa",

xTitle = "Ngày", yTitle = "Giá")



Hình 5.7 biểu đồ giá đóng cửa

stock\_fpt = Vnstock().stock(symbol='FPT', source='VCI')

stock\_vgi = Vnstock().stock(symbol='VGI', source='VCI')

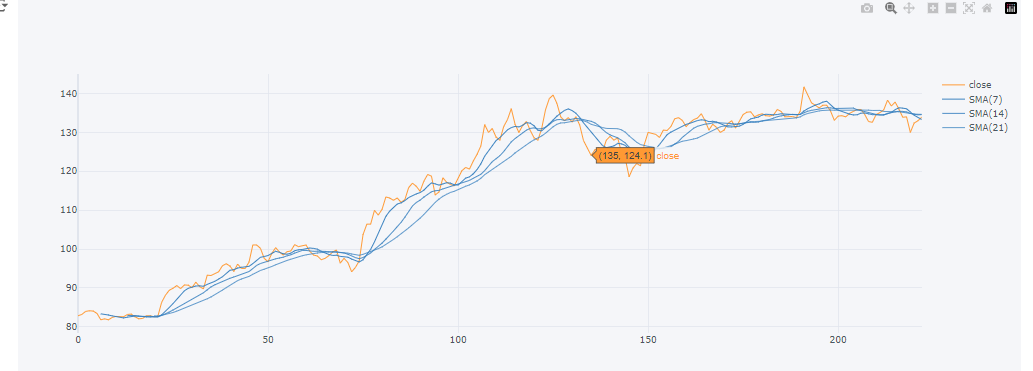
stock\_vtc = Vnstock().stock(symbol='VTC', source='VCI')

df\_fpt = stock\_fpt.quote.history(start='2024-01-01', end='2024-12-31', interval='1D')

df\_vgi = stock\_vgi.quote.history(start='2024-01-01', end='2024-12-31', interval='1D')

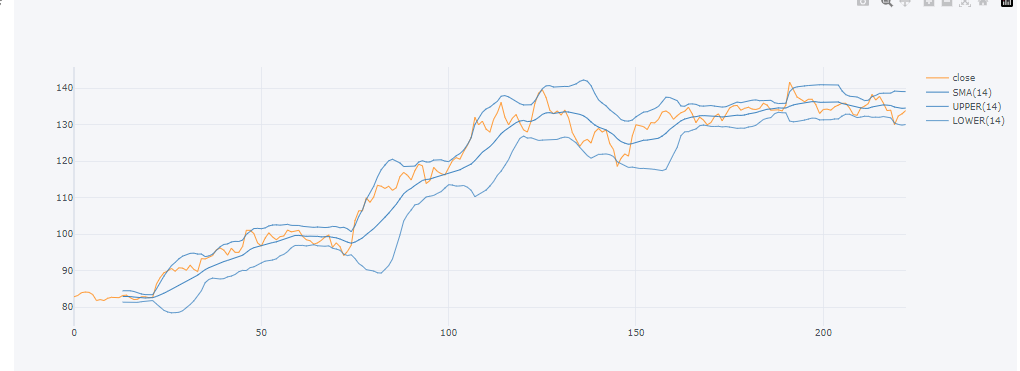
df\_vtc = stock\_vtc.quote.history(start='2024-01-01', end='2024-12-31', interval='1D')

df\_fpt['close'].ta\_plot(study='sma',periods=[7,14,21],title='SMA FPT')



Hình 5.8. Biểu đồ giá đóng cửa và các đường trung bình động (MA)

df\_fpt['close'].ta\_plot(study='boll')



Hình 5.9: Biểu đồ giá đóng cửa và dải Bollinger Bands

df\_fpt[['open', 'high', 'low', 'close']].iplot(kind='candle',title='Biểu đồ hình nến mã FPT')



Hình 5.10: Biểu đồ Nến Nhật Mã Cổ phiếu FPT

df\_vgi[['open', 'high', 'low', 'close']].iplot(kind='candle',title='Biểu đồ hình nến mã VGI')



Hình 5.11. Biểu đồ Nến Nhật Mã Cổ phiếu VGI

df\_vtc = df\_vtc.rename(columns={'Giá mở': 'open', 'Giá cao': 'high', 'Giá thấp': 'low', 'Giá đóng': 'close'})

df\_vtc[['open', 'high', 'low', 'close']].iplot(kind='candle',title='Biểu đồ hình nến mã VTC')



Hình 5.12. Biểu đồ Nến Nhật Mã Cổ phiếu VGI

cf.datagen.ohlc()

fpt\_plot = cf.QuantFig(df\_fpt,title = 'Biểu đồ hình nến FPT',legend = "Top", name = "Nến")

fpt\_plot.add\_bollinger\_bands(colors =["red", 'green'])

fpt\_plot.iplot()



Hình 5.13: Biểu đồ Nến Nhật và Dải Bollinger Bands của Mã Cổ phiếu FPT

# CHƯƠNG 6: CHỦ ĐỀ NÂNG CAO

Phân tích cụm (Cluster Analysis): sử dụng phân tích nhóm để phân loại các cổ phiếu công nghê thông tin dựa trên các đặc điểm tương đồng như mức độ biến động giá, khối lượng giao dịch, và chỉ số tài chính.

## 6.1. GIỚI THIỆU

Phân tích cụm (Cluster Analysis) là một kỹ thuật học máy không giám sát được sử dụng để nhóm các điểm dữ liệu thành các cụm dựa trên sự giống nhau của chúng. Nói cách khác, nó là một quá trình nhóm các đối tượng thành các nhóm sao cho các đối tượng trong cùng một nhóm (được gọi là cụm) giống nhau hơn (theo một nghĩa nào đó) so với các đối tượng trong các nhóm (cụm) khác. Nó được sử dụng rộng rãi để phân tích dữ liệu khám phá trong nhiều lĩnh vực, bao gồm học máy, nhận dạng mẫu, phân tích hình ảnh, truy xuất thông tin, sinh tin học, nén dữ liệu và đồ họa máy tính.

## 6.2. CÁC THUẬT TOÁN PHÂN TÍCH CỤM

Có nhiều thuật toán phân cụm khác nhau, nhưng một số thuật toán phổ biến nhất bao gồm:

* **Phân cụm K-means:** Đây là một trong những thuật toán phân cụm đơn giản và được sử dụng rộng rãi nhất. Nó hoạt động bằng cách lặp đi lặp lại gán các điểm dữ liệu cho cụm gần nhất dựa trên khoảng cách của chúng với tâm của cụm.
* **Phân cụm phân cấp:** Thuật toán này xây dựng một hệ thống phân cấp các cụm bằng cách bắt đầu với mỗi điểm dữ liệu trong cụm riêng của nó, sau đó hợp nhất các cặp cụm gần nhất cho đến khi chỉ còn lại một cụm.
* **DBSCAN:** Viết tắt của Phân cụm không gian dựa trên mật độ của các ứng dụng có nhiễu. Đây là một thuật toán phân cụm dựa trên mật độ nhóm các điểm gần nhau (các điểm có nhiều lân cận gần) và đánh dấu là các điểm ngoại lai các điểm nằm trong các vùng có mật độ thấp (có các lân cận gần nhất quá xa).
* **Phân cụm Gauss hỗn hợp (GMM):** Đây là một phương pháp tiếp cận xác suất để phân cụm giả định rằng các điểm dữ liệu được tạo ra từ một hỗn hợp các phân phối Gauss. Nó linh hoạt hơn so với k-means vì nó xem xét không chỉ khoảng cách của các điểm đến trung tâm mà còn xem xét phương sai của dữ liệu (cách dữ liệu được trải rộng).

## 6.3. ƯU VÀ NHƯỢC ĐIỂM CỦA PHÂN TÍCH CỤM

* Ưu điểm của phân tích cụm

Có một số ưu điểm khi sử dụng phân tích cụm, bao gồm:

* Nó có thể được sử dụng để khám phá cấu trúc ẩn trong dữ liệu. Phân tích cụm có thể giúp khám phá các mẫu và mối quan hệ trong dữ liệu không rõ ràng ngay lập tức.
* Nó có thể được sử dụng để giảm độ phức tạp của dữ liệu. Phân tích cụm có thể giúp đơn giản hóa các tập dữ liệu phức tạp bằng cách nhóm các điểm dữ liệu tương tự lại với nhau.
* Nó có thể được sử dụng để cải thiện hiệu suất của các thuật toán học máy khác. Phân tích cụm có thể được sử dụng làm bước tiền xử lý cho các thuật toán học máy khác, chẳng hạn như phân loại và hồi quy.
* Nhược điểm của phân tích cụm

Cũng có một số nhược điểm khi sử dụng phân tích cụm, bao gồm:

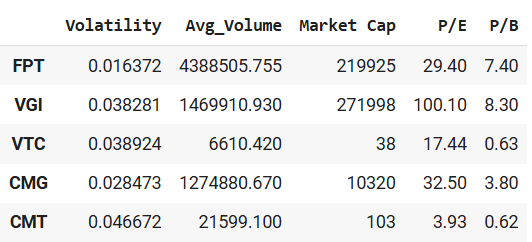
* Nó có thể chủ quan. Việc lựa chọn thuật toán phân cụm và số lượng cụm có thể chủ quan và có thể ảnh hưởng đến kết quả phân tích.
* Nó có thể tốn kém về mặt tính toán. Phân tích cụm có thể tốn kém về mặt tính toán, đặc biệt là đối với các tập dữ liệu lớn.
* Có thể khó giải thích. Có thể khó giải thích kết quả của phân tích cụm, đặc biệt là khi số lượng cụm lớn.

## 6.4. ỨNG DỤNG CỦA PHÂN TÍCH CỤM VÀO BỘ DỮ LIỆU ĐỂ PHÂN LOẠI CÁC CỔ PHIẾU

Phân tích cụm là một công cụ có giá trị có thể được sử dụng để cải thiện các quyết định đầu tư. Bằng cách hiểu các mối quan hệ giữa các cổ phiếu khác nhau, các nhà đầu tư có thể đưa ra các quyết định sáng suốt hơn về cổ phiếu nào nên mua và bán. Em xin trình bày kết quả phân tích cụm cho một nhóm gồm 5 cổ phiếu: FPT, VGI, VTC, CMG, và CMT. Mục tiêu của phân tích là nhóm các cổ phiếu này thành các cụm riêng biệt dựa trên sự tương đồng của chúng về các đặc trưng tài chính, từ đó rút ra những nhận định hữu ích cho việc đầu tư.

Dữ liệu sử dụng trong phân tích bao gồm 5 đặc trưng cho mỗi cổ phiếu:

* Volatility (Biến động giá): Thể hiện mức độ biến động của giá cổ phiếu.
* Average Volume (Khối lượng giao dịch trung bình): Thể hiện mức độ thanh khoản của cổ phiếu.
* Market Cap (Vốn hóa thị trường): Thể hiện quy mô của công ty.
* P/E (Price-to-Earnings Ratio): Tỷ lệ giá trên thu nhập, thể hiện mức định giá của cổ phiếu so với thu nhập của công ty.
* P/B (Price-to-Book Ratio): Tỷ lệ giá trên giá trị sổ sách, thể hiện mức định giá của cổ phiếu so với tài sản ròng của công ty



Bảng 6.1: Bảng dữ liệu chứng khoán

Có nhiều cách khác nhau để thực hiện phân tích cụm, nhưng phương pháp phổ biến nhất là sử dụng thuật toán phân cụm k-means. Thuật toán này hoạt động bằng cách nhóm các điểm dữ liệu thành k cụm, trong đó k là một số do người dùng chỉ định. Các điểm dữ liệu được gán cho cụm mà chúng gần nhất.

### 6.4.1. Tiền xử lý dữ liệu:

Dữ liệu được chuẩn hóa bằng phương pháp StandardScaler trong thư viện scikit-learn của Python. StandardScaler sẽ chuyển đổi từng đặc trưng sao cho chúng có trung bình bằng 0 và độ lệch chuẩn bằng 1. Việc chuẩn hóa này giúp loại bỏ ảnh hưởng của sự khác biệt về thang đo giữa các đặc trưng, đảm bảo rằng tất cả các đặc trưng đều có đóng góp như nhau vào quá trình phân cụm.

# Đọc thư viện

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

# Chuẩn hóa dữ liệu

scaler = StandardScaler()

scaled\_features = scaler.fit\_transform(features\_df\_all)

### 6.4.2. Phân cụm K-Means:

  Thuật toán K-Means được sử dụng để phân cụm các cổ phiếu. K-Means là một thuật toán phân cụm phổ biến, hoạt động bằng cách chia dữ liệu thành k cụm sao cho tổng bình phương khoảng cách từ các điểm dữ liệu đến tâm cụm của chúng là nhỏ nhất. Trong phân tích này, số cụm được chọn là k = 3.

# Phân cụm K-Means

n\_clusters = 3  # Thử nghiệm với 3 cụm

kmeans = KMeans(n\_clusters=n\_clusters, random\_state=0)

kmeans.fit(scaled\_features)

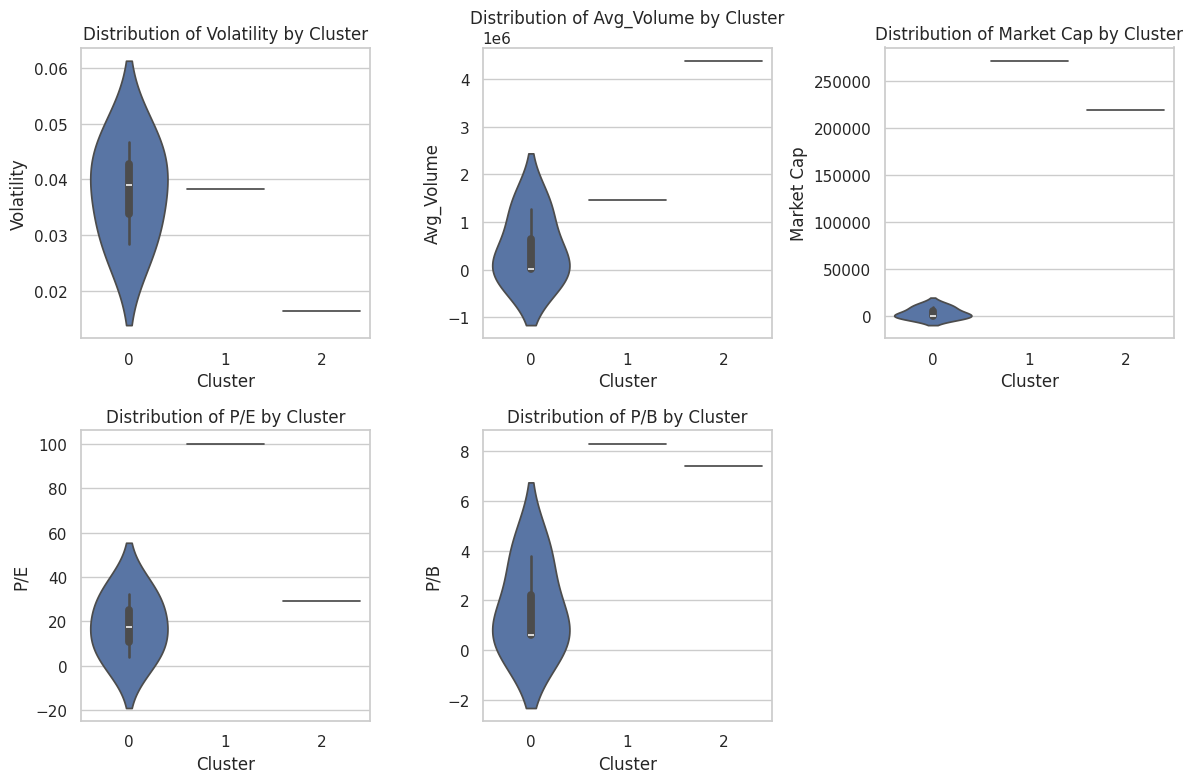
labels = kmeans.labels\_

### 6.4.3. Trực quan hóa:

  Kết quả phân cụm được trực quan hóa bằng các biểu đồ pairplot và violin từ thư viện seaborn của Python. Biểu đồ pairplot hiển thị mối quan hệ giữa từng cặp đặc trưng, với màu sắc thể hiện cụm của mỗi cổ phiếu. Biểu đồ violin hiển thị phân phối của từng đặc trưng cho mỗi cụm, giúp dễ dàng so sánh các cụm với nhau.



Hình 6.1 Biểu đồ cặp các đặc trưng theo cụm



Hình 6.2 Biểu đồ phân bố theo cụm

Kết quả phân tích cụm với n\_clusters = 3 cho thấy các cổ phiếu được phân chia thành các cụm như sau:

* **Cụm 0:** VTC, CMG, CMT
* **Cụm 1:** VGI
* **Cụm 2:** FPT

### 6.4.4 Phân tích từng cụm:

* **Cụm 0 (VTC, CMG, CMT):** Nhóm này bao gồm các cổ phiếu có đặc điểm chung là vốn hóa thị trường nhỏ, khối lượng giao dịch trung bình thấp (thể hiện tính thanh khoản thấp hơn), và tỷ lệ P/B thấp. Điều này cho thấy các cổ phiếu trong nhóm này có thể đang bị định giá thấp so với giá trị sổ sách của chúng. Tuy nhiên, các cổ phiếu này cũng có mức độ biến động giá (Volatility) từ trung bình đến cao. Đây có thể là nhóm cổ phiếu phù hợp với những nhà đầu tư chấp nhận rủi ro cao và tìm kiếm cơ hội từ các cổ phiếu có tiềm năng tăng trưởng trong tương lai.
* **Cụm 1 (VGI):** VGI được xếp vào một cụm riêng biệt do có sự khác biệt đáng kể so với các cổ phiếu khác. VGI có vốn hóa thị trường và khối lượng giao dịch lớn thứ hai trong nhóm, cho thấy tính thanh khoản tương đối cao. Tuy nhiên, VGI lại có tỷ lệ P/E rất cao (100.10), điều này có thể cho thấy thị trường đang kỳ vọng rất cao vào sự tăng trưởng của VGI trong tương lai hoặc cổ phiếu đang được định giá quá cao so với thu nhập hiện tại. VGI cũng có P/B cao và mức độ biến động giá cao.
* **Cụm 2 (FPT):** FPT được xếp vào một cụm riêng do có những đặc điểm nổi bật. FPT là cổ phiếu có vốn hóa thị trường lớn nhất và khối lượng giao dịch cao nhất trong nhóm, thể hiện tính thanh khoản rất cao. FPT cũng có mức độ biến động giá thấp nhất, cho thấy sự ổn định hơn so với các cổ phiếu khác. Tỷ lệ P/E và P/B của FPT ở mức vừa phải, cho thấy mức định giá hợp lý hơn so với VGI. FPT có thể được coi là một cổ phiếu blue-chip, phù hợp với các nhà đầu tư tìm kiếm sự ổn định và tăng trưởng bền vững.

### 6.4.5. Hạn chế

Phân tích này có một số hạn chế cần được lưu ý:

* Số lượng mẫu nhỏ: Phân tích chỉ dựa trên 5 cổ phiếu, do đó kết quả có thể không đại diện cho toàn bộ thị trường chứng khoán Việt Nam.
* Hạn chế về số lượng đặc trưng: Phân tích chỉ sử dụng 5 đặc trưng tài chính. Việc bổ sung thêm các đặc trưng khác, chẳng hạn như các chỉ số kỹ thuật, thông tin vĩ mô, hoặc các yếu tố định tính khác, có thể cải thiện kết quả phân cụm và cung cấp cái nhìn toàn diện hơn.
* Dữ liệu tĩnh: Dữ liệu được sử dụng trong phân tích là dữ liệu tĩnh tại một thời điểm. Việc phân tích dựa trên dữ liệu chuỗi thời gian có thể giúp theo dõi sự thay đổi của các cụm theo thời gian và cung cấp thông tin hữu ích hơn cho việc ra quyết định đầu tư.

### 6.4.6. Kết luận và Kiến nghị

Phân tích cụm đã giúp nhóm 5 cổ phiếu FPT, VGI, VTC, CMG, và CMT thành 3 nhóm riêng biệt dựa trên sự tương đồng về các đặc trưng tài chính. Kết quả phân tích đã cung cấp những thông tin hữu ích về đặc điểm của từng nhóm cổ phiếu, từ đó hỗ trợ nhà đầu tư trong việc lựa chọn cổ phiếu phù hợp với khẩu vị rủi ro và mục tiêu đầu tư của mình.

Để cải thiện phân tích, em kiến nghị:

* Mở rộng phạm vi phân tích với nhiều cổ phiếu hơn và nhiều đặc trưng hơn.
* Sử dụng dữ liệu chuỗi thời gian để theo dõi sự thay đổi của các cụm theo thời gian.
* Kết hợp phân tích cụm với các phương pháp phân tích cơ bản và kỹ thuật khác để đưa ra quyết định đầu tư toàn diện hơn.

# CHƯƠNG 7: TỔNG KẾT BÁO CÁO

## 7.1. MỤC TIÊU VÀ PHẠM VI CỦA ĐỒ ÁN

Mục tiêu của báo cáo này là tập trung vào việc thăm dò và trực quan hóa dữ liệu tài chính của ba mã cổ phiếu thuộc nhóm ngành công nghệ thông tin: FPT, VGI, và VTC. Thông qua phân tích, báo cáo hướng đến việc khám phá các đặc điểm chính của dữ liệu, bao gồm giá mở cửa, giá đóng cửa, khối lượng giao dịch, và các chỉ số biến động giá. Đồng thời, báo cáo cũng so sánh hiệu suất của ba cổ phiếu để nhận diện xu hướng và các yếu tố ảnh hưởng đến giá trị thị trường.

Phạm vi nghiên cứu bao gồm việc thu thập dữ liệu từ nguồn đáng tin cậy là thư viện Python vnstock, sử dụng các kỹ thuật thống kê và trực quan hóa để trình bày dữ liệu, đồng thời áp dụng các phân tích nâng cao nhằm cung cấp thông tin hữu ích cho việc ra quyết định đầu tư.

## 7.2. QUY TRÌNH VÀ PHƯƠNG PHÁP SỬ DỤNG

Đồ án được thực hiện qua một quy trình rõ ràng và khoa học, bắt đầu từ việc thu thập dữ liệu, xử lý và làm sạch dữ liệu, sau đó tiến hành các bước phân tích cơ bản và nâng cao. Dữ liệu được thu thập từ thư viện vnstock, bao gồm giá cổ phiếu, khối lượng giao dịch và các thông tin liên quan của ba mã cổ phiếu trong khoảng thời gian từ năm 2024. Sau khi làm sạch dữ liệu, loại bỏ các giá trị thiếu và ngoại lai, dữ liệu được tổ chức lại dưới dạng bảng dữ liệu Pandas để dễ dàng phân tích.

Các kỹ thuật thống kê mô tả được sử dụng để tóm tắt dữ liệu, kiểm định thống kê được thực hiện để đánh giá mối quan hệ giữa các biến, và các biểu đồ được tạo bằng Matplotlib và Seaborn để trực quan hóa kết quả. Ngoài ra, báo cáo còn sử dụng các phương pháp nâng cao như phân tích cụm để nhóm cổ phiếu có đặc điểm tương đồng và hồi quy tuyến tính để xác định mối quan hệ giữa giá cổ phiếu và khối lượng giao dịch.

## 7.3. KẾT QUẢ CHÍNH VÀ PHÁT HIỆN

Phân tích dữ liệu đã chỉ ra nhiều kết quả quan trọng, trong đó nổi bật là các đặc điểm khác nhau của ba mã cổ phiếu FPT, VGI, và VTC. Giá cổ phiếu FPT cho thấy sự ổn định cao và biên độ dao động thấp, phù hợp với chiến lược đầu tư dài hạn. Ngược lại, VGI có biên độ dao động lớn hơn, là mã cổ phiếu hấp dẫn cho các chiến lược giao dịch ngắn hạn. VTC có giá trị thấp nhất trong nhóm nhưng duy trì được sự ổn định, phù hợp với các nhà đầu tư bảo toàn vốn. Mối tương quan giữa các mã cổ phiếu cũng được phân tích, cho thấy FPT và VGI có tương quan dương mạnh, phản ánh xu hướng biến động đồng pha, trong khi VTC có tương quan thấp hơn với hai mã còn lại, giúp đa dạng hóa danh mục đầu tư. Phân tích cụm đã nhóm ba cổ phiếu thành các cụm riêng biệt, cho thấy sự khác biệt rõ rệt về tính chất và hiệu suất của từng mã cổ phiếu.

## 7.4. Ý NGHĨA VÀ ỨNG DỤNG KẾT QUẢ

Kết quả phân tích từ báo cáo mang lại nhiều ý nghĩa thực tiễn quan trọng, không chỉ đối với nhà đầu tư mà còn đối với các nhà phân tích thị trường và hoạch định chiến lược. Đối với nhà đầu tư, báo cáo cung cấp thông tin chi tiết về xu hướng, hiệu suất và đặc điểm của từng mã cổ phiếu, từ đó hỗ trợ việc ra quyết định đầu tư dựa trên dữ liệu. Các biểu đồ và phân tích trực quan hóa giúp dễ dàng nhận diện các xu hướng và mẫu hình, tạo nền tảng cho các chiến lược mua, giữ hoặc bán cổ phiếu. Đối với thị trường, mối tương quan giữa các mã cổ phiếu giúp xác định mức độ đồng bộ trong nhóm ngành, từ đó đánh giá được rủi ro chung.

## 7.5. HẠN CHẾ VÀ ĐỀ XUẤT HƯỚNG PHÁT TRIỂN

Báo cáo vẫn tồn tại một số hạn chế cần được khắc phục trong các nghiên cứu tiếp theo để nâng cao tính toàn diện và ứng dụng của kết quả. Phạm vi nghiên cứu chỉ giới hạn ở ba mã cổ phiếu, do đó chưa đủ bao quát để đại diện toàn bộ nhóm ngành công nghệ thông tin. Dữ liệu phân tích tập trung trong một năm, dẫn đến khó khăn trong việc đánh giá các xu hướng dài hạn và sự ảnh hưởng của các yếu tố kinh tế vĩ mô. Một số kỹ thuật nâng cao như dự báo chuỗi thời gian hoặc mô hình học sâu chưa được áp dụng trong báo cáo này.

Để cải thiện, các nghiên cứu tiếp theo nên mở rộng phạm vi dữ liệu bao gồm nhiều mã cổ phiếu hơn và thời gian dài hơn, đồng thời tích hợp các mô hình dự báo tiên tiến như LSTM hoặc ARIMA để phân tích xu hướng dài hạn. Bên cạnh đó, việc so sánh hiệu suất cổ phiếu giữa các nhóm ngành khác nhau cũng là một hướng nghiên cứu đáng được xem xét.

## 7.6. KẾT LUẬN CUỐI CÙNG

Báo cáo đã đạt được mục tiêu đặt ra khi cung cấp một bức tranh rõ ràng và chi tiết về ba mã cổ phiếu FPT, VGI, và VTC thông qua các kỹ thuật thăm dò và trực quan hóa dữ liệu hiện đại. Kết quả phân tích không chỉ giúp nhận diện xu hướng, mối tương quan, và đặc điểm nổi bật của từng mã cổ phiếu mà còn cung cấp những thông tin hữu ích hỗ trợ ra quyết định đầu tư chiến lược. Các biểu đồ và kết quả trực quan hóa không chỉ làm sáng tỏ dữ liệu thô mà còn mang lại giá trị thực tiễn cao. Để đạt được sự toàn diện hơn, cần tiếp tục mở rộng nghiên cứu, áp dụng các kỹ thuật nâng cao và phân tích sâu rộng hơn. Báo cáo này là một bước khởi đầu hữu ích, không chỉ giúp hiểu rõ hơn về thị trường chứng khoán mà còn mở ra nhiều tiềm năng phát triển cho các nghiên cứu và ứng dụng trong tương lai.

Kết luận lại, báo cáo này không chỉ đạt được mục tiêu ban đầu mà còn mở ra nhiều hướng nghiên cứu và ứng dụng mới. Thị trường chứng khoán không ngừng biến động, và sự kết hợp giữa phân tích dữ liệu hiện đại và các yếu tố vĩ mô sẽ giúp cải thiện khả năng dự báo và ra quyết định. Đây không chỉ là một bài học trong phân tích dữ liệu mà còn là một công cụ mạnh mẽ hỗ trợ các nhà đầu tư và nhà hoạch định chiến lược trong việc tối ưu hóa hiệu quả đầu tư và quản trị rủi ro. Bằng việc tiếp tục phát triển các kỹ thuật phân tích, chúng ta có thể tiến gần hơn đến việc hiểu rõ hơn về thị trường tài chính, tạo ra giá trị.

# TÀI LIỆU THAM KHẢO

[1]. Tài liệu học tập Đồ án 1: Trực quan hóa dữ liệu bằng R, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[2]. Tài liệu học tập: Nhập môn Trí tuệ Nhân tạo, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[3]. Tài liệu học tập: Trực quan hóa dữ liệu, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[4]. Link nguồn thư viện tham khảo: <https://github.com/thinh-vu/vnstock>

[5]. Link: https://www.youtube.com/playlist?list=PLv6GftO355AuLcf4lFigEFzRBTb4yM2vX

[6]. Hỗ trợ của Chat GPT 3.5. Link: https://chatgpt.com/