**Lab 1: Project Kickoff & Initial Implementation**

# **Bước 1: Mô tả bài toán, mục tiêu dự án và người dùng cuối**

## **1.1 Bài toán**

* **Bài toán:** Dự đoán giá cổ phiếu của công ty Coca-Cola (mã KO) trong tương lai dựa trên dữ liệu lịch sử.
* **Mục tiêu dự án:**
  + Cung cấp dự đoán chính xác và đáng tin cậy về giá cổ phiếu trong tương lai bằng cách sử dụng các mô hình AI.
  + Giải thích các yếu tố ảnh hưởng đến giá chứng khoán, bao gồm việc xác định các ngày trong quá khứ có ảnh hưởng lớn nhất đến dự đoán giá, bằng cách sử dụng Explainable AI (XAI) để phân tích và làm rõ kết quả từ mô hình.
* **Người dùng cuối:**
  + **Nhà đầu tư cá nhân:** Những người muốn tối ưu hóa danh mục đầu tư của mình dựa trên dự đoán giá cổ phiếu.
  + **Quỹ đầu tư:** Các tổ chức tìm kiếm cơ hội đầu tư thông qua phân tích dữ liệu chuyên sâu.
  + **Nhà quản lý tài chính:** Những người cần thông tin chính xác để tư vấn cho khách hàng.
  + **Nhà phân tích thị trường chứng khoán:** Các chuyên gia muốn hiểu rõ hơn về xu hướng và các yếu tố ảnh hưởng đến giá cổ phiếu.

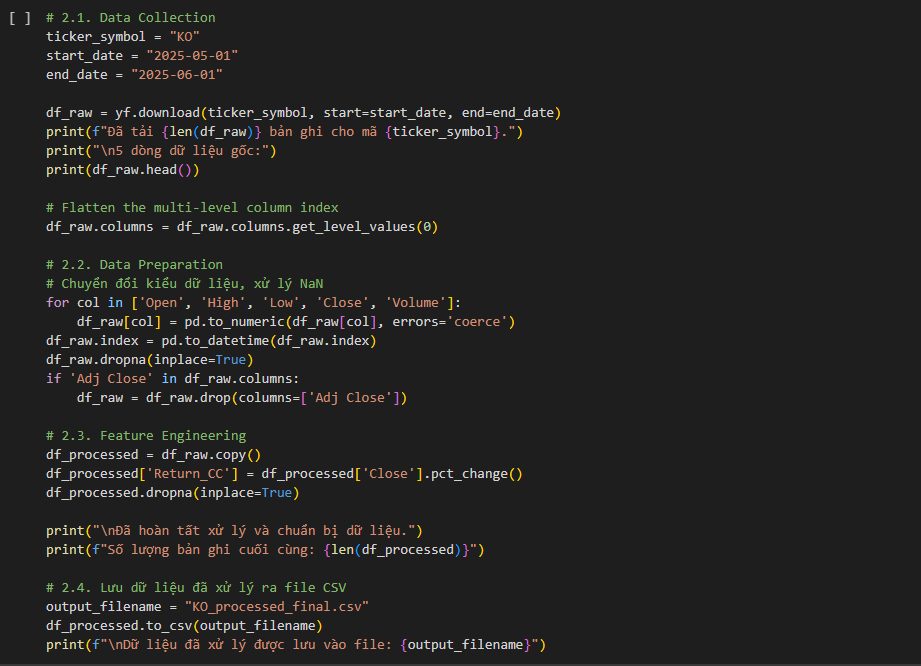
## **1. 2 Xác định các câu hỏi chính (research questions) và chỉ số đánh giá (KPIs)**

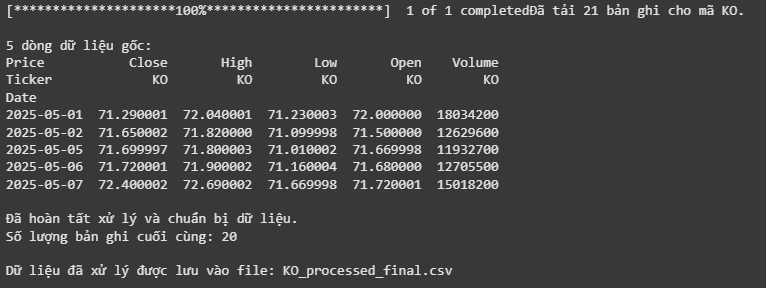
* **Câu hỏi chính:**
  + Có thể dự đoán chính xác giá cổ phiếu của Coca-Cola trong tương lai bằng các dự liêụ giá/khối lượng trong quá khứ không?
  + Những yếu tố nào (giá mở cửa, khối lượng giao dịch, giá của các ngày trong quá khứ, v.v.) ảnh hưởng đến giá cổ phiếu?
  + Giá của những ngày nào trong quá khứ có ảnh hưởng lớn nhất đến giá cổ phiếu dự đoán?
  + Có xu hướng hoặc mẫu hình nào trong dữ liệu lịch sử có thể hỗ trợ dự đoán giá cổ phiếu không?
* **Chỉ số đánh giá (KPIs):**
  + **Độ chính xác của các mô hình dự đoán:** MAE, MSE, RMSE, MAPE
  + **So sánh giữa các model dự đoán giá chứng khoán**
  + **Tỷ lệ dự đoán chính xác xu hướng giá:** Phần trăm dự đoán đúng giá tăng hoặc giảm.
  + **Các chỉ số đo lường trong chứng khoán:** SMA, EMA, RSI, MACD

### **Bước 2: Thu thập dữ liệu (Data Collection)**

**Mô tả kết quả:**

* **Thu thập:** Dữ liệu lịch sử giá cổ phiếu KO từ ngày 01/05/2025 đến 01/06/2025 đã được tải thành công bằng thư viện yfinance, thu được 21 bản ghi thô.
* **Chuẩn bị:** Dữ liệu đã được làm sạch bằng cách chuyển đổi kiểu dữ liệu cho các cột giá và khối lượng, xử lý các giá trị bị thiếu (NaN). Cột Adj Close không cần thiết đã được loại bỏ.
* **Kỹ thuật đặc trưng (Feature Engineering):** Một cột dữ liệu mới là Return\_CC (tỷ suất sinh lợi liên tục) đã được tạo ra để phục vụ cho các phân tích sau này.
* **Kết quả cuối cùng:** Sau khi xử lý, bộ dữ liệu cuối cùng còn lại 20 bản ghi, sạch và sẵn sàng cho các bước phân tích tiếp theo. Dữ liệu này được lưu vào tệp KO\_processed\_final.csv.

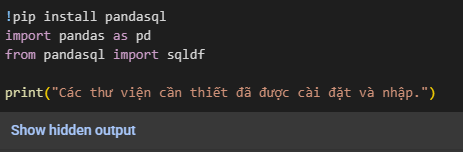


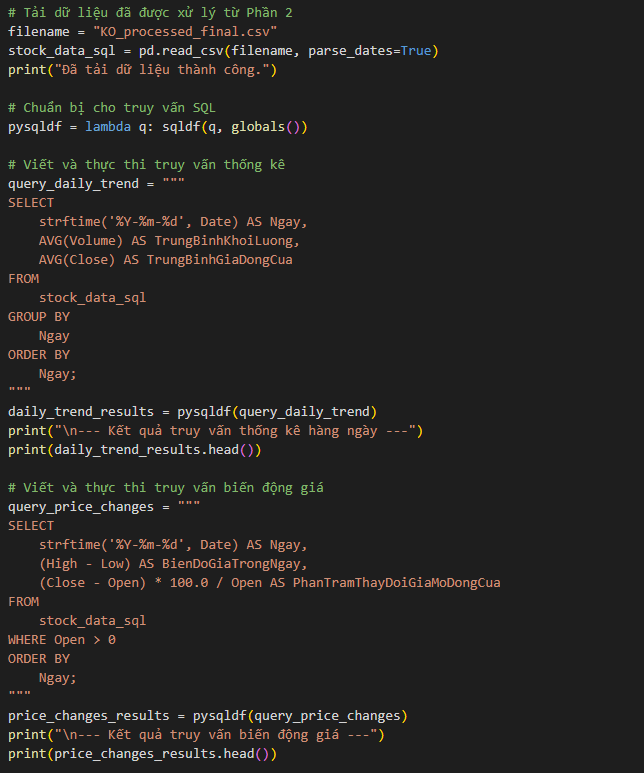


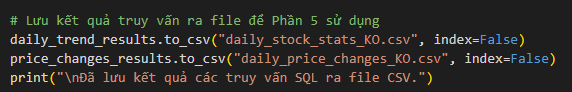
### **Bước 3: Phân tích Dữ liệu bằng SQL**

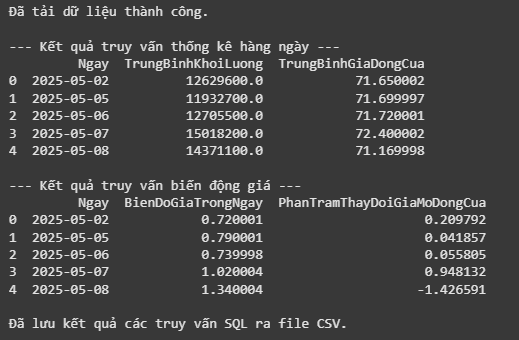
**Mô tả kết quả:**

* Sử dụng thư viện pandasql, các truy vấn SQL đã được thực thi trực tiếp trên DataFrame của Pandas để thực hiện các phân tích thống kê.
* **Truy vấn 1:** Tính toán khối lượng giao dịch trung bình (TrungBinhKhoiLuong) và giá đóng cửa trung bình (TrungBinhGiaDongCua) cho mỗi ngày.
* **Truy vấn 2:** Tính toán các chỉ số về biến động giá, bao gồm biên độ giá trong ngày (BienDoGiaTrongNgay) và phần trăm thay đổi giữa giá mở cửa và đóng cửa (PhanTramThayDoiGiaMoDongCua).
* **Kết quả:** Hai bảng kết quả đã được tạo ra và lưu thành các tệp CSV (daily\_stock\_stats\_KO.csv và daily\_price\_changes\_KO.csv) để sử dụng cho bước trực quan hóa dữ liệu.





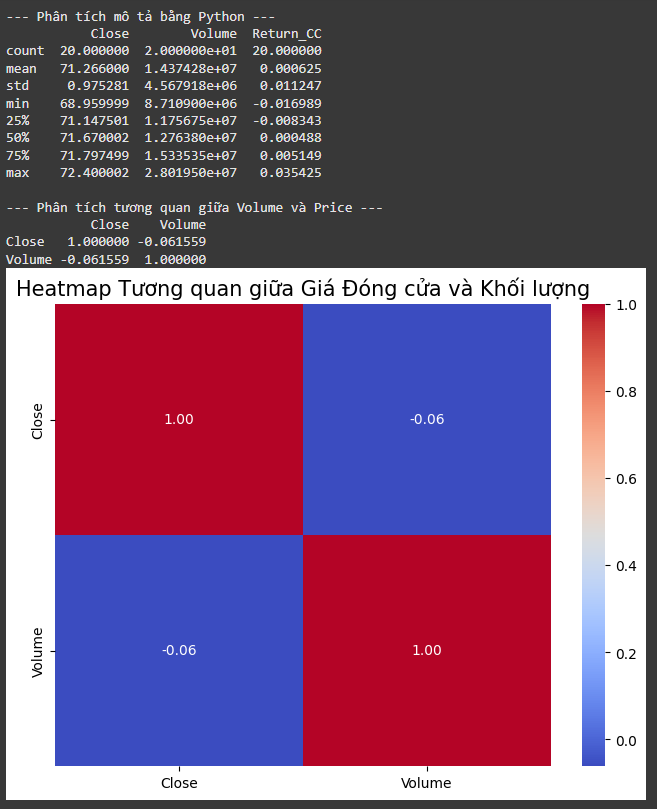
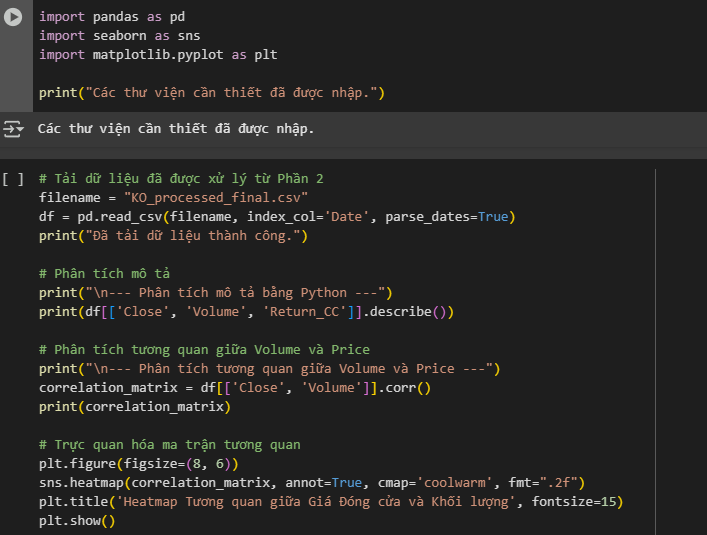




### **Bước 4: Phân tích Dữ liệu bằng Python**

**Mô tả kết quả:**

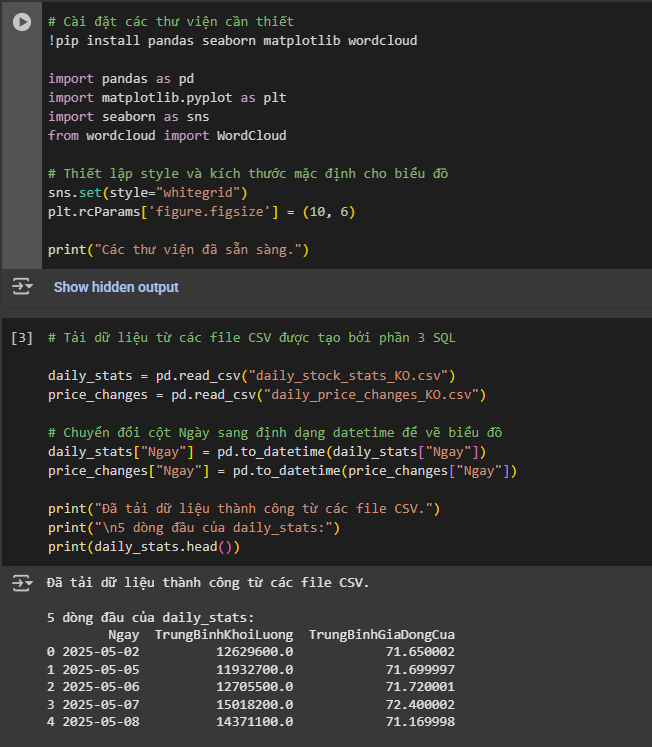
* **Phân tích mô tả:** Sử dụng hàm .describe() của Pandas, các chỉ số thống kê cơ bản (như giá trị trung bình, độ lệch chuẩn, min, max) cho các cột Close, Volume, và Return\_CC đã được tính toán. Kết quả cho thấy giá đóng cửa trung bình trong giai đoạn này là khoảng **71.27 USD** với khối lượng giao dịch trung bình là **14.37 triệu cổ phiếu/ngày**.
* **Phân tích tương quan:** Ma trận tương quan đã được tính toán để xem xét mối quan hệ giữa giá đóng cửa (Close) và khối lượng giao dịch (Volume). Hệ số tương quan là **-0.0616**, cho thấy một mối tương quan rất yếu và gần như không có giữa hai biến này trong khoảng thời gian phân tích.

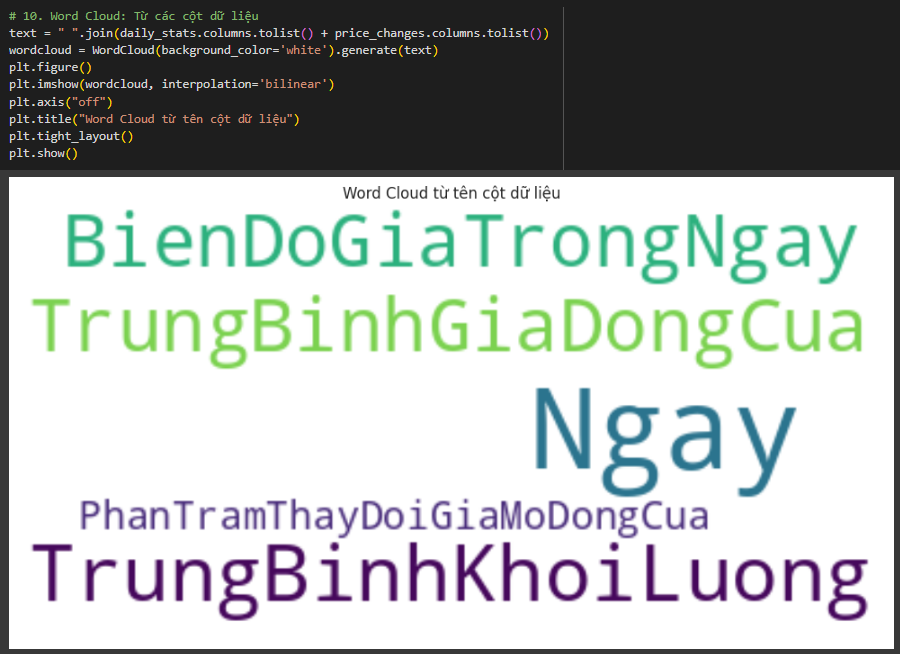
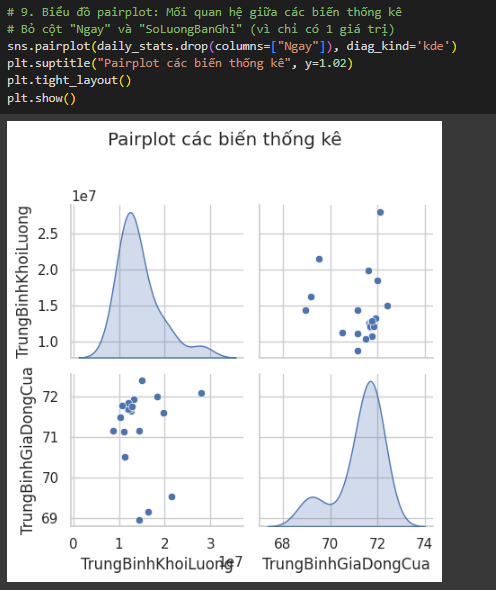
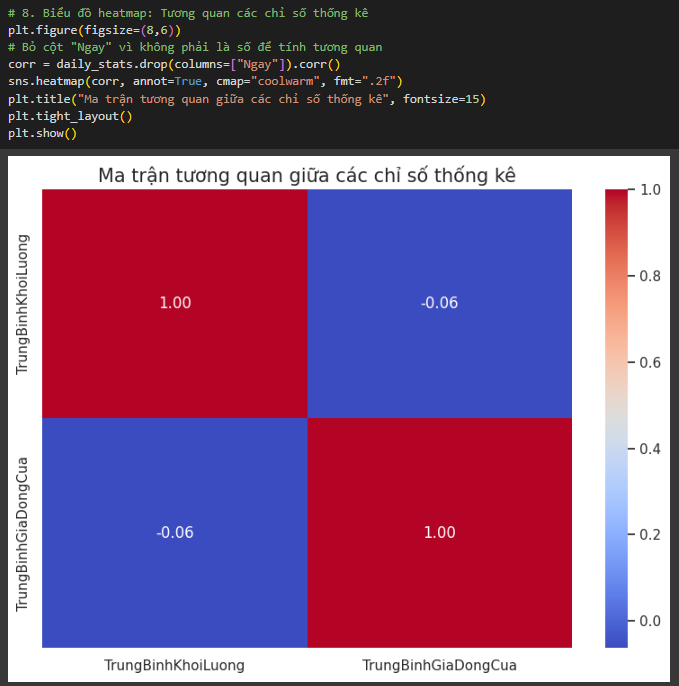
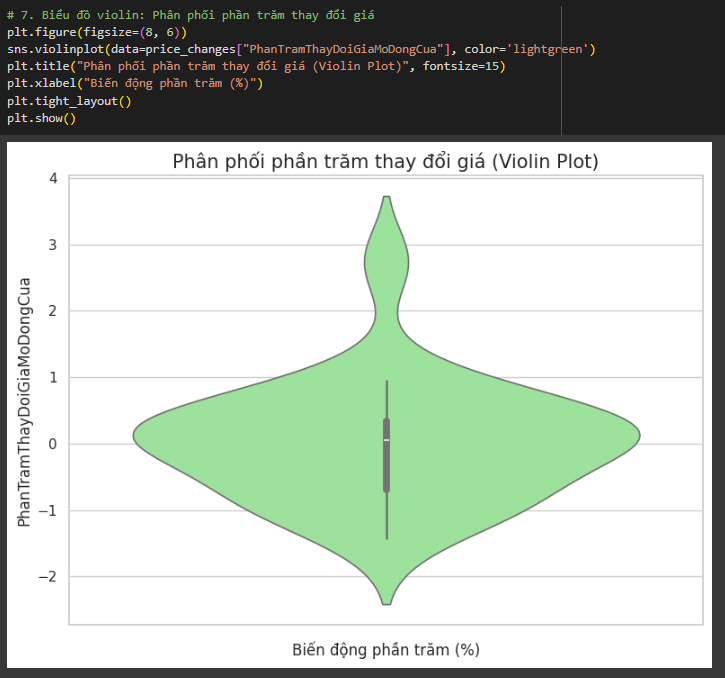
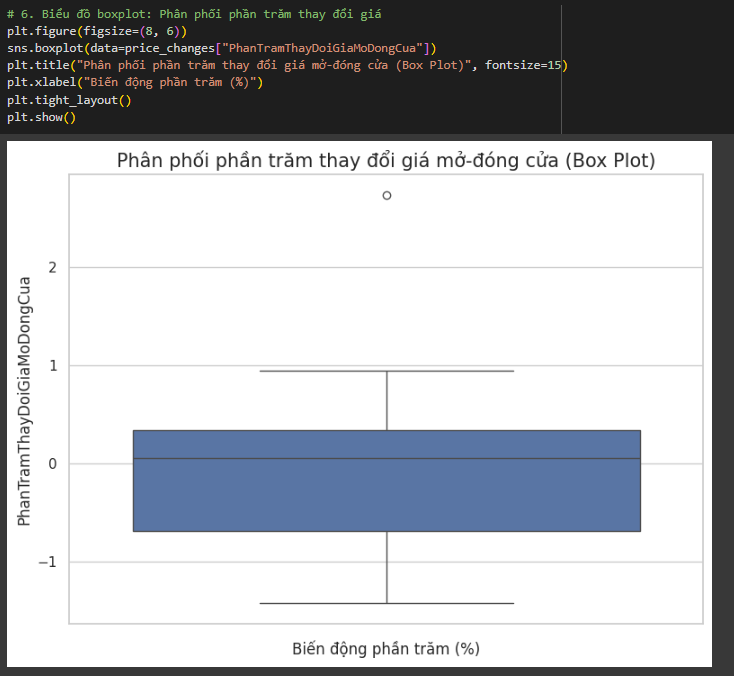
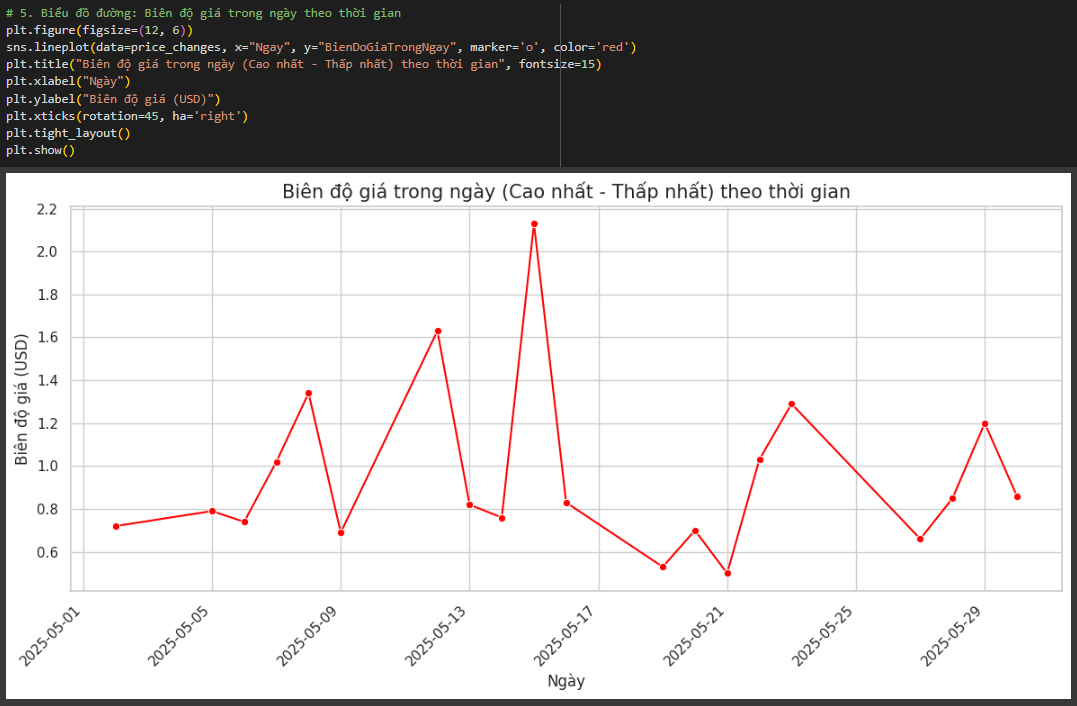
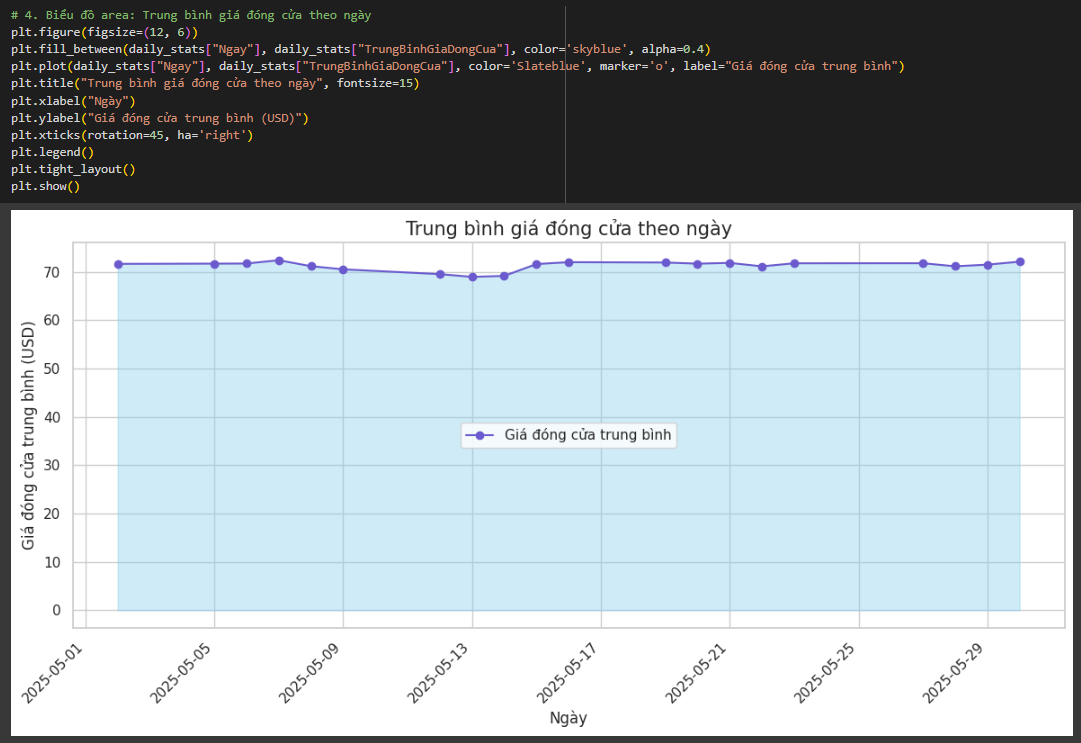
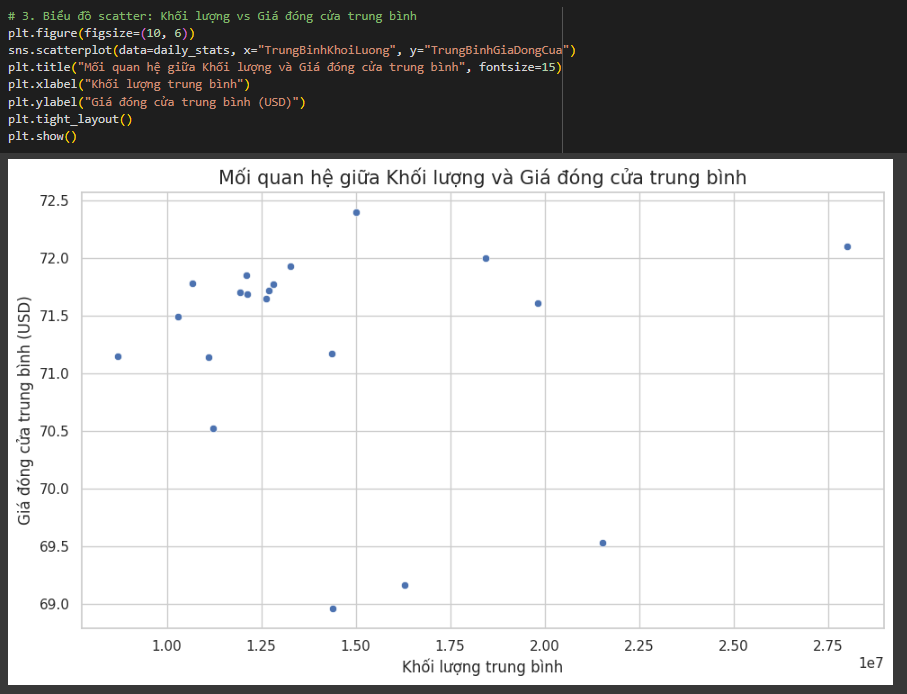
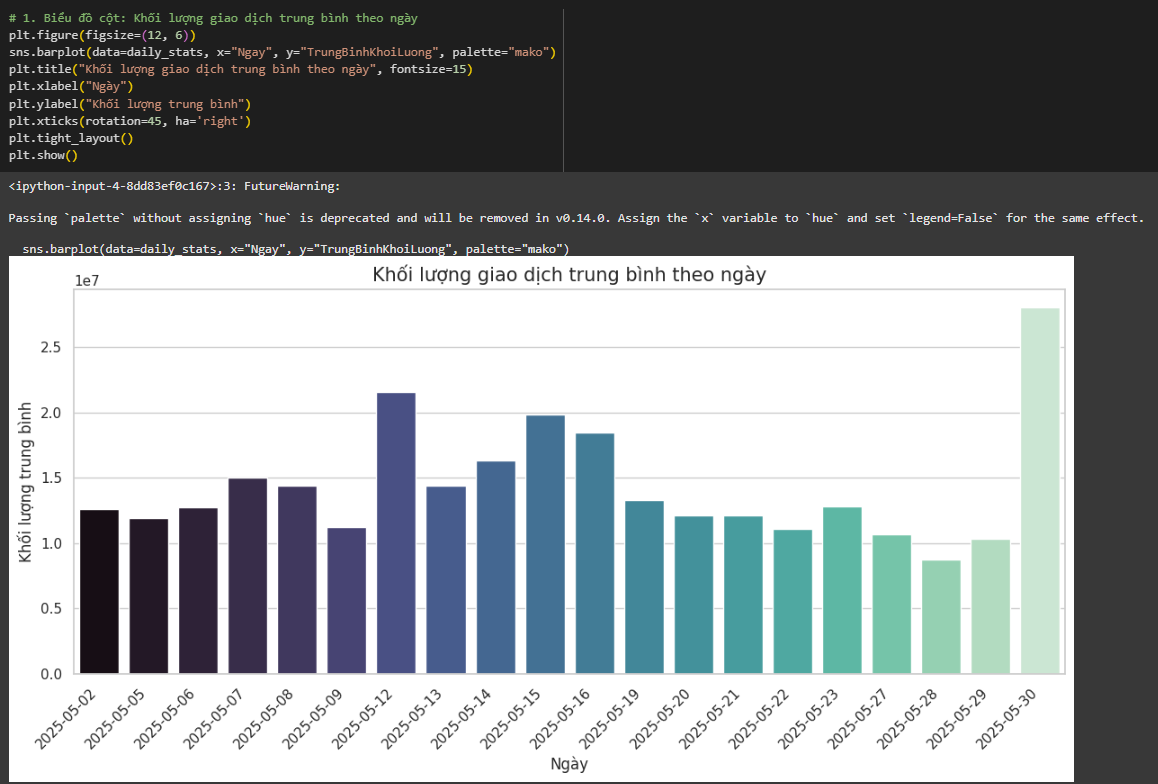


### **Bước 5: Trực quan hóa Dữ liệu**

**Mô tả kết quả:** Dữ liệu được phân tích từ các bước trước đã được trực quan hóa để có cái nhìn sâu sắc hơn:

* **Biểu đồ cột và Biểu đồ vùng:** Cho thấy xu hướng của khối lượng giao dịch và giá đóng cửa trung bình theo từng ngày trong tháng 5/2025.
* **Biểu đồ histogram:** Phân phối của phần trăm thay đổi giá đóng-mở cửa được thể hiện, cho thấy phần lớn các ngày, giá cổ phiếu có biến động nhỏ quanh mức 0%.
* **Biểu đồ scatter:** Minh họa mối quan hệ giữa khối lượng giao dịch và giá đóng cửa, xác nhận trực quan về sự tương quan yếu giữa hai yếu tố này.
* **Box plot và Violin plot:** Cung cấp một cái nhìn chi tiết hơn về sự phân bổ và các giá trị ngoại lai của biến động giá hàng ngày.
* **Heatmap tương quan:** Trực quan hóa ma trận tương quan, làm nổi bật mối quan hệ yếu giữa giá và khối lượng.
* **Word Cloud:** Một cách trực quan sáng tạo để hiển thị các thuật ngữ (tên cột) được sử dụng trong phân tích.





### **Bước 6: Xây dựng và Đánh giá Mô hình Dự báo**

**Mô tả kết quả:**

* **Xây dựng mô hình:** Năm mô hình khác nhau đã được xây dựng và huấn luyện để dự báo giá đóng cửa của cổ phiếu KO, bao gồm: ARIMA, RandomForest, LinearRegression, DecisionTree, và SVR.
* **Đánh giá hiệu suất:** Các mô hình được đánh giá dựa trên hai chỉ số chính là Sai số tuyệt đối trung bình (MAE) và Sai số phần trăm tuyệt đối trung bình (MAPE).
* **Kết quả so sánh:**
  + **ARIMA:** MAE = 0.3630, MAPE = 0.51%
  + **SVR:** MAE = 0.5707, MAPE = 0.80%
  + **RandomForest:** MAE = 0.5847, MAPE = 0.82%
  + **LinearRegression:** MAE = 0.6050, MAPE = 0.84%
  + **DecisionTree:** MAE = 1.1100, MAPE = 1.55%
* **Mô hình tốt nhất:** Dựa trên kết quả, mô hình **ARIMA** cho kết quả tốt nhất với chỉ số MAE và MAPE thấp nhất, chứng tỏ khả năng dự báo gần với giá thực tế nhất.
* **Trực quan hóa:** Kết quả dự báo từ tất cả các mô hình đã được vẽ trên cùng một biểu đồ so với giá thực tế, cho thấy đường dự báo của ARIMA bám sát nhất với đường giá thực tế.

