

BÁO CÁO CASE 02 – PHÂN TÍCH DỮ LIỆU THỊ TRƯỜNG CHỨNG KHOÁN

I. Phương pháp xử lý và phân tích dữ liệu

1. Mô tả tổng quan dữ liệu

Dữ liệu đầu vào của bài toán được cung cấp dưới dạng các tệp CSV, mô phỏng dữ liệu thị trường chứng khoán. Bộ dữ liệu gồm ba bảng chính:

- **stock_info**: Chứa thông tin định danh cổ phiếu như mã cổ phiếu (ticker), sàn giao dịch (exchange) và ngành nghề (sector).
- **stock_price**: Chứa dữ liệu giá cổ phiếu theo thời gian, bao gồm giá đóng cửa theo từng ngày giao dịch.
- **transactions**: Chứa thông tin giao dịch của từng mã cổ phiếu, bao gồm khối lượng giao dịch và loại giao dịch.

2. Quy trình xử lý dữ liệu

Nhóm tiến hành xử lý dữ liệu theo một quy trình tuần tự, đảm bảo dữ liệu đầu vào được chuẩn hóa và sẵn sàng cho phân tích.

Bước 1 – Làm sạch dữ liệu (Task 01)

- Kiểm tra dữ liệu trùng lặp và loại bỏ các bản ghi không hợp lệ.
- Xử lý giá trị thiếu (NA) bằng cách loại bỏ hoặc thay thế phù hợp.
- Chuẩn hóa tên cột giữa các bảng để đảm bảo tính nhất quán khi xử lý.
- Kiểm tra và chuyển đổi kiểu dữ liệu, đặc biệt là các cột số và cột ngày.

Bước 2 – Khám phá dữ liệu (Task 02)

- Phân tích cấu trúc dữ liệu: số lượng dòng, số lượng cột và kiểu dữ liệu.
- Thực hiện thống kê mô tả (mean, min, max, std) cho giá đóng cửa và khối lượng giao dịch.
- Phân tích phân bố cổ phiếu theo sàn giao dịch và theo ngành nhằm nắm được đặc điểm tổng quan của dữ liệu.

Bước 3 – Gộp dữ liệu (Task 04)

- Chuẩn hóa cột thời gian về cùng một định dạng (`date`) nhằm tránh lỗi khi merge dữ liệu.
- Thực hiện merge bảng **stock_price** với **stock_info** theo khóa `ticker` để bổ sung thông tin ngành và sàn cho dữ liệu giá.
- Merge thêm bảng **transactions** theo khóa `ticker` nhằm hoàn thiện bảng dữ liệu tổng hợp phục vụ phân tích.

Bước 4 – Phân tích tổng hợp (Task 05)

- Sử dụng Pivot Table để tổng hợp dữ liệu theo nhiều chiều như ngành, sàn và thời gian.

- Thực hiện chuyển đổi dữ liệu bằng stack và unstack để hiểu cấu trúc MultiIndex trong pandas.
- Phân tích xu hướng tăng trưởng của các ngành dựa trên sự biến động giá theo thời gian.

II. Các bảng kết quả chính

Bảng 1. Giá đóng cửa trung bình theo ngành × tháng

Ngành	2024-01	2024-02
Bán Lẻ	51,358.27	53,155.17
Công Nghệ	54,769.23	55,000.00
Tài Chính	52,571.43	52,000.00
Tài Chính	52,184.78	51,873.68

Bảng 1 phản ánh giá đóng cửa trung bình của các ngành theo từng tháng. Ngành Công Nghệ duy trì mức giá cao nhất và tăng nhẹ trong tháng 02/2024. Ngành Bán Lẻ ghi nhận mức tăng rõ rệt, trong khi hai nhóm Tài Chính/Tài Chính có xu hướng giảm nhẹ.

Bảng 2. Giá đóng cửa trung bình theo sàn giao dịch

Sàn giao dịch	Giá đóng cửa trung bình
HOSE	53,214.56
HNX	51,987.34

Bảng 2 phản ánh sự khác biệt về mức giá cổ phiếu trung bình giữa các sàn giao dịch, cho thấy HOSE có mặt bằng giá cao hơn.

Bảng 2. Khối lượng giao dịch theo sàn × ngành

Sàn / Ngành	Bán Lẻ	Công Nghệ	Tài Chính	Tài Chính
HNX	2,625,000	0	0	0
HOSE	3,780,000	0	4,000,000	0
UPCOM	5,060,000	2,520,000	13,860,000	4,500,000

Bảng 2 cho thấy khối lượng giao dịch tập trung chủ yếu trên sàn UPCOM, đặc biệt ở nhóm ngành Tài Chính. Sàn HOSE có thanh khoản đáng kể ở ngành Bán Lẻ và Tài Chính, trong khi HNX chỉ ghi nhận giao dịch ở ngành Bán Lẻ.

Bảng 3. Phân tích tăng trưởng ngành

Ngành	Mức tăng giá (2024-02 – 2024-01)
Bán Lẻ	+1,796.90

Ngành	Mức tăng giá (2024-02 – 2024-01)
Công Nghệ	+230.77
Tài Chính	-311.10
Tài Chính	-571.43

Bảng 3 cho thấy ngành Bán Lẻ có mức tăng trưởng mạnh nhất trong giai đoạn phân tích. Ngành Công Nghệ tăng nhẹ, trong khi hai nhóm ngành Tài Chính/Tài Chính có xu hướng giảm giá.

III. Nhận xét dựa trên kết quả phân tích

- Kết quả phân tích cho thấy giá cổ phiếu có sự khác biệt rõ rệt giữa các ngành, phản ánh đặc thù hoạt động và mức độ phát triển không đồng đều của các lĩnh vực kinh tế.
- Một số ngành có xu hướng tăng trưởng ổn định qua các tháng, cho thấy tiềm năng phát triển tích cực và thu hút sự quan tâm của nhà đầu tư.
- Khối lượng giao dịch tập trung chủ yếu ở một số ngành và sản nhất định, cho thấy dòng tiền trên thị trường chưa phân bổ đồng đều.
- Việc sử dụng pivot table kết hợp với stack/unstack giúp nhóm phân tích dữ liệu theo nhiều chiều một cách linh hoạt và hiệu quả hơn.

IV. Kết luận

Qua Case 02, nhóm đã thực hiện đầy đủ quy trình xử lý và phân tích dữ liệu thị trường chứng khoán từ dữ liệu thô đến dữ liệu phân tích. Bắt đầu từ khâu làm sạch và chuẩn hóa dữ liệu, nhóm đã đảm bảo tính nhất quán giữa các bảng dữ liệu về giá cổ phiếu, thông tin doanh nghiệp và giao dịch.

Thông qua việc gộp dữ liệu và áp dụng các kỹ thuật phân tích như thống kê mô tả, pivot table và stack/unstack, nhóm đã khai thác được thông tin theo nhiều chiều khác nhau, bao gồm ngành, sản giao dịch và thời gian. Các bảng kết quả cho thấy sự khác biệt rõ rệt về giá cổ phiếu và khối lượng giao dịch giữa các ngành, đồng thời giúp xác định những ngành có xu hướng tăng trưởng tích cực trong giai đoạn phân tích.

Kết quả của bài làm không chỉ giúp hiểu rõ hơn về đặc điểm và xu hướng của thị trường chứng khoán mà còn rèn luyện kỹ năng xử lý, phân tích và trình bày dữ liệu bằng Python và Pandas. Đây là nền tảng quan trọng cho việc áp dụng phân tích dữ liệu vào các bài toán thực tế và các học phần nâng cao sau này.

V. Nhật ký nhóm

STT	Họ và tên thành viên	Nhiệm vụ đảm nhận
1	Mai Văn Duy	Đọc & làm sạch dữ liệu (task 01)
2	Phan Thị Thanh Huyền	Truy vấn & thống kê mô tả (task 02)
3	Nguyễn Thu Hương	GroupBy & Tổng hợp (task 03)

STT	Họ và tên thành viên	Nhiệm vụ đảm nhận
4	Nguyễn Đăng Long	Merge dữ liệu (task 04)
5	Nguyễn Ngọc Anh	Pivot Table + Stack/Unstack (task 05) ,tạo repo,viết báo cáo.