Bài tập lập trình 1: Hiểu dữ liệu

Bài tập 1 của môn Khai phá Dữ liệu nhằm mục đích giúp sinh viên làm quen với việc hiểu dữ liệu và khai thác thông tin cần thiết thông qua các phương pháp thống kê cơ bản, đồng thời biết kết hợp kiến thức chuyên ngành (domain knowledge) để ứng dụng vào bài toán thực tế, trong trường hợp này là dữ liệu tài chính về thị trường chứng khoán.

Dữ liệu sử dụng bao gồm giá cổ phiếu của một số công ty niêm yết tại Việt Nam và các chỉ số tài chính theo quý của các công ty đó. Sinh viên sẽ thực hiện các phân tích như tính toán thống kê mô tả (trung bình, độ lệch chuẩn, phương sai, trung vị, phân vị), trực quan hóa dữ liệu để phát hiện xu hướng, và kiểm tra mối quan hệ giữa các biến số tài chính. Mục tiêu của bài tập là giúp sinh viên phát triển kỹ năng tiền xử lý và phân tích dữ liệu, từ đó có cái nhìn sâu hơn về cách dữ liệu tài chính phản ánh hoạt động kinh doanh và biến động thị trường.

Sinh viên có thể sử dụng các kiến thức về thống kê, cũng như kỹ năng lập trình xử lý dữ liệu với Python để làm bài tập này, bao gồm kỹ năng sử dụng các thư viện chuyên biệt trong Python.

Gợi ý một số thư viện có thể dùng:

- Tính toán thống kê: Pandas, NumPy, SciPy, Statsmodels (thống kê nâng cao)
- Trực quan hóa dữ liệu (dataviz): Matplotlib, Seaborn
- Ngoài ra, để mô hình hóa dữ liệu: scikit-learn (học máy và phân tích hồi quy), PyMC3, PyMC4 (thống kê Bayes)

- **Dữ liệu:**

Cho 2 bảng dữ liệu sau :

- Simplize_IMP_FinancialIdicatior_….. : Báo cáo về Chỉ số tài chính của một công ty (theo quý)
- Simplize_IMP_PriceHistory_… : Báo cáo về Lịch sử giá của một loại cổ phiếu của công ty

Thời gian: 5 năm trở lại đây (theo ngày)

Công ty:

STT	Tên doanh nghiệp	Mã cổ phiếu
1.	Công ty Cổ phần Dược phẩm Imexpharm	IMP
2.	Công ty Cổ phần FPT	FPT
3.	Tổng Công ty Khí Việt Nam - Công ty Cổ phần	GAS
4.	Công ty Cổ phần Tập đoàn Hòa Phát	HPG
5.	Công ty Cổ phần Tập đoàn MaSan	MSN
6.	Công ty Cổ phần Sữa Việt Nam	VNM

7.	Công ty Cổ phần Đầu tư Thế Giới Di Động	MWG

Lưu ý: Mỗi nhóm chọn 01 mã cổ phiếu để làm

- **Yêu cầu đề bài:**

  Lần lượt thực hiện các bước sau đây để phân tích dữ liệu về giá cổ phiếu

1. **Quan sát để hiểu doanh nghiệp và dữ liệu**

- Thực hiện thống kê miêu tả (descriptive statistics) hai bảng dữ liệu được cho: five-number summary (tính giá trị nhỏ nhất – **min**, tứ phân vị thứ nhất – **Q1**, trung vị – **median** hay tứ phân vị thứ hai – **Q2**, tứ phân vị thứ ba – **Q3**, Giá trị lớn nhất – **max**), trung bình (**average**), trung vị (**mean**), độ lệch chuẩn (**standard deviation**), phương sai (**variance**).
- Lấy giá đóng cửa mỗi ngày trong bảng Lịch sử giá làm giá cổ phiếu ngày hôm đó, hãy vẽ biểu đồ theo dõi sự biến động giá cổ phiếu.

  (Gợi ý: Có thể tạo ra một dataframe nhỏ các ngày cuối quý các năm từ 2021 đến 2025, ví dụ các ngày 31/3, 30/6, 30/9, 31/12)
- Quan sát lịch sử giá cổ phiếu xem có những đợt tăng hay giảm giá cổ phiếu nào (bất thường nếu có), tìm hiểu background (bối cảnh lịch sử) của công ty và thị trường xem có những yếu tố nào có thể là nguyên nhân tác động tới giá cổ phiếu của công ty (chưa cần thực hiện phép tính toán).

2. **Tiền xử lý dữ liệu**

◆ Tạo một bảng dữ liệu mới từ hai bảng đã cho.

➢ Các cột dữ liệu trong bảng bao gồm các thông tin quan trọng cho việc tính toán và mô hình hóa thống kê từ hai bảng đã cho:

➔ Dữ liệu lấy từ bảng Lịch sử giá (cần phải tính lại cho phù hợp nội dung của bảng Chỉ số tài chính): Mốc thời gian (quý/năm), Giá đóng cửa, Thay đổi giá, % thay đổi

➔ Dữ liệu lấy từ bảng Chỉ số tài chính của công ty, bao gồm các chỉ số có khả năng ảnh hưởng tới giá cổ phiếu: Biên lợi nhuận gộp, Biên lợi nhuận ròng, P/E (tỷ lệ giá trên lợi nhuận Price to Earnings Ratio), EPS (lợi nhuận trên mỗi cổ phiếu Earnings Per Share), Tăng trưởng EPS, ROE (tỷ suất sinh lời trên vốn chủ sở hữu), tỷ lệ Nợ phải trả/Vốn chủ sở hữu, Khả năng thanh toán tổng quát, Vòng quay tài sản (Asset turnover ratio), Giá trị sổ sách (Book Value Per Share).

➢ Mỗi hàng là các chỉ số trong một quý.

➢ Cách lấy số liệu cho mỗi hàng:

➔ Dữ liệu theo của quý đó trong Báo cáo Chỉ số tài chính của công ty.

➔ Đối với dữ liệu về Lịch sử giá thì tính trung bình dữ liệu của ngày cuối cùng của quý đó và 14 ngày trước và sau ngày đó.

- ➔ VD: Ngày cuối cùng của quý 4/2024 là ngày 31/12/2024, vậy để tính giá cổ phiếu của quý 4/2024 ta sẽ tính trung bình giá cổ phiếu từ ngày 17/12/2024 đến ngày 14/01/2025.
- ➔ Như vậy thì các cột Thay đổi giá và % thay đổi cũng phải tính lại theo giá trị mới của quý.

3. **Lựa chọn các yếu tố nguy cơ (risk factors) tiềm năng cho mô hình tài chính**

Ở bước này, sinh viên sẽ sàng lọc ra một số yếu tố nguy cơ (risk factors) « tiềm năng » có thể gây ra biến động về giá cổ phiếu. Các factors này có thể đến từ trong chính công ty đó – các yếu tố nội tại (internal) – thể hiện bằng các chỉ số trong bảng báo cáo tài chính của công ty, hoặc từ thị trường bên ngoài tác động vào – ví dụ như các biến đổi trên thị trường chứng khoán thế giới, các ngành nghề liên quan, các yếu tố văn hóa chính trị, dịch bệnh,… khiến nền kinh tế bị trì trệ, v.v.

Trong bài tập này chúng ta chỉ xét tới các yếu tố nội tại của chính công ty phát hành cổ phiếu. Để sàng lọc thì chúng ta cần tính độ tương quan Pearson của các chỉ số tài chính với giá cổ phiếu (dữ liệu được tạo ra ở mục 2).

Từ đó đưa ra kết luận:

- ➔ Giá cổ phiếu của công ty này có thể có mối quan hệ phụ thuộc vào những chỉ số tài chính nào?

4. **Viết báo cáo trình bày lại những nội dung đã thực hiện.**

Mục đích của báo cáo là để luyện tập kỹ năng phân tích dữ liệu (data analysis).

Yêu cầu:

Báo cáo phải trình bày code, in ra 5 dòng đầu tiên sau khi tạo ra bộ dữ liệu mới ở bước 2, biểu đồ có tên và chú thích, phân tích thống kê và nhận xét riêng những gì quan sát được qua dữ liệu, biểu đồ và tính tương quan của các chỉ số trong bộ dữ liệu. Ở mục 1 phải trình bày những gì quan sát và tìm hiểu được về công ty, loại cổ phiếu và dữ liệu.

Format: ipynb hoặc pdf.