TÓM TẮT ÔN THI

PHÂN TÍCH DỮ LIỆU KINH DOANH

Business Data Analytics

Phân tích kinh doanh dựa trên công nghệ và dữ liệu

TS. Hà Thúc Huân — Viện Quản trị & Công nghệ FSB

6 module · Tài liệu ôn tập cô đọng

Mục lục

Cách dùng tài liệu này

Mỗi module gồm ba phần: **khái niệm cốt lõi** (định nghĩa phải thuộc), **công thức & ngưỡng** (con số phải nhớ) và **khi nào dùng** (nhận diện dạng bài). Hai phụ lục cuối tài liệu dùng để ôn nhanh trước khi thi: **Phụ lục A** giúp nhận diện nên dùng kỹ thuật nào cho từng dạng bài, **Phụ lục B** gom các con số, ngưỡng và những bẫy khái niệm hay bị nhầm.

Bản đồ toàn môn

Module	Trả lời câu hỏi	Kỹ thuật chính
1. Giới thiệu	Phân tích dữ liệu là gì, đặt câu hỏi thế nào?	DDDM, MECE, 5 Whys
2. Mô tả (Descriptive)	Chuyện gì đã và đang xảy ra?	Trực quan hóa, mean/median/mode, độ lệch chuẩn, CV
3. Suy luận thống kê	Kết quả này là thật hay do ngẫu nhiên?	CLT, khoảng tin cậy, kiểm định giả thuyết, p-value, A/B test
4. Dự báo (Predictive)	Điều gì sẽ xảy ra tiếp theo?	Hồi quy đơn & đa biến, phần dư, chuỗi thời gian, Delphi
5. Data Mining	Có nhóm/mẫu hình ẩn nào trong dữ liệu?	Phân cụm, phân loại (k-NN), luật kết hợp
6. Đề xuất (Prescriptive)	Nên làm gì là tốt nhất?	Tối ưu hóa: hàm mục tiêu, biến quyết định, ràng buộc

Module 1 — Giới thiệu về phân tích dữ liệu trong kinh doanh

1.1. Định nghĩa cốt lõi

> **Định nghĩa:** Phân tích dữ liệu trong kinh doanh là việc sử dụng dữ liệu, công nghệ thông tin, phân tích thống kê, phương pháp định lượng và mô hình toán học/mô hình dựa trên máy tính để giúp nhà quản lý có cái nhìn sâu sắc hơn về hoạt động kinh doanh và ra quyết định tốt hơn dựa trên dữ liệu thực tế.

> Chuỗi giá trị: Dữ liệu thô → Insight (sự thật ngầm hiểu) → Quyết định kinh doanh tốt hơn

- Phân tích dữ liệu **không thay thế** trực giác và kinh nghiệm lãnh đạo — nó là "trợ lý thông minh" giúp **kiểm chứng** và **làm sắc bén** trực giác.
- Kinh nghiệm là **con dao hai lưỡi**: vô giá ở giai đoạn đầu, nhưng khi hành vi tiêu dùng thay đổi, đối thủ mới xuất hiện và công nghệ phát triển thì kinh nghiệm dễ trở thành "lối mòn".
- **DDDM (Data-Driven Decision Making)** — ra quyết định dựa trên dữ liệu — không còn là lựa chọn mà là yêu cầu bắt buộc để tồn tại và phát triển.

1.2. Ba cấp độ phân tích (phải thuộc)

Cấp độ	Định nghĩa	Câu hỏi
Mô tả (Descriptive)	Dùng dữ liệu để hiểu hiệu quả kinh doanh trong quá khứ và hiện tại, từ đó ra quyết định sáng suốt.	Chuyện gì đã xảy ra?
Dự báo (Predictive)	Dự báo tương lai bằng cách kiểm tra dữ liệu lịch sử, phát hiện xu hướng, chu kỳ lặp lại hoặc mối tương quan giữa các biến, rồi ngoại suy cho tương lai.	Chuyện gì sẽ xảy ra?
Đề xuất (Prescriptive)	Xác định các phương án lựa chọn tốt nhất để giảm thiểu hoặc tối đa hóa một mục tiêu cụ thể.	Nên làm gì?

1.3. Hai khung tư duy đặt câu hỏi

MECE — Mutually Exclusive, Collectively Exhaustive

- "**Không trùng lặp, không bỏ sót**". Do **McKinsey & Company** phổ biến.
- **ME**: các yếu tố trong cùng một nhóm phải hoàn toàn riêng biệt, không chồng chéo.
- **CE**: gộp tất cả các nhóm lại phải bao quát mọi khả năng.
- Công dụng: chia nhỏ vấn đề mơ hồ để **khoanh vùng khu vực có vấn đề**.

5 Whys

- Phát triển bởi **Sakichi Toyoda**, một phần của Hệ thống sản xuất Toyota.
- Hỏi "vì sao" nhiều lần liên tiếp, không dừng ở câu trả lời bề mặt, đào đến **quy trình/hệ thống bị lỗi gốc rễ**.

> **Ví dụ yêu cầu tôi cần tránh:** "Doanh thu đang giảm, hãy tìm hiểu đi!" — không MECE, không có giả thuyết, không biết bắt đầu từ đâu.

1.4. Nguồn dữ liệu — "mỏ vàng" 3 tầng

Tầng	Trạng thái	Nội dung	Trả lời
Dữ liệu vận hành	Sẵn có	Dữ liệu máy móc (tốc độ, hiệu suất, thời gian dừng, tỷ lệ lỗi), ERP (tồn kho, đơn hàng), năng suất nhân công theo ca/chuyền	Chúng ta vận hành hiệu quả đến đâu?

Tầng	Trạng thái	Nội dung	Trả lời
Dữ liệu kinh doanh	Cần xây dựng	Bán hàng (sản phẩm, kênh, thời gian), khách hàng (nhân khẩu học, lịch sử mua), marketing (hiệu quả chiến dịch)	CÁI GÌ đang diễn ra trên thị trường?
Dữ liệu bên ngoài	Cần khai thác	Xu hướng ngành, báo cáo thị trường, sản phẩm/giá/khuyến mãi đối thủ, review – bình luận khách hàng	TẠI SAO thị trường diễn ra như vậy?

1.5. Dữ liệu lớn (Big Data) — 3V

> **Định nghĩa:** Các tập hợp dữ liệu cực kỳ lớn, phức tạp và được tạo ra liên tục đến mức các công cụ xử lý truyền thống (như Excel) không thể thu thập, quản lý và xử lý một cách hiệu quả.

> 3V: Volume (khối lượng lớn) · Velocity (tốc độ cao) · Variety (đa dạng loại hình)

1.6. Độ tin cậy vs. Tính hợp lệ (rất hay ra đề)

- **Độ tin cậy (Reliability)**: dữ liệu chính xác và **nhất quán**.
- **Tính hợp lệ (Validity)**: dữ liệu đo lường **đúng thứ cần đo**.

Tình huống	Tin cậy?	Hợp lệ?
Thiết bị đo áp suất lốp luôn hiển thị thấp hơn thực tế vài bar	Không (sai lệch hệ thống)	Có (vẫn đang đo áp suất lốp)
Đếm chính xác số cuộc gọi đến CSKH mỗi ngày, dùng để đo mức độ **không hài lòng**	Có	Không (nhiều cuộc gọi chỉ là hỏi thông tin)
Câu hỏi khảo sát về chất lượng thức ăn, dùng để đo **sự hài lòng tổng thể**	Không (khách có quan điểm trái ngược)	Không (hài lòng còn gồm cả dịch vụ)

1.7. Lợi ích, thách thức và rào cản

- **Lợi ích**: giảm chi phí · quản lý rủi ro tốt hơn · ra quyết định nhanh hơn · năng suất cao hơn · cải thiện lợi nhuận và sự hài lòng khách hàng.
- **Thách thức**: thiếu hiểu biết cách dùng phân tích · ưu tiên kinh doanh mâu thuẫn · kỹ năng phân tích không đủ · khó thu thập dữ liệu tốt và chia sẻ thông tin · không hiểu rõ lợi ích so với chi phí.
- **5 lý do doanh nghiệp "ngại" dữ liệu**: (1) tư duy chủ nghĩa kinh nghiệm/linh cảm; (2) thiếu nhân lực phân tích và diễn giải; (3) chi phí đầu tư công nghệ – hạ tầng; (4) chất lượng dữ liệu thấp, phân mảnh, không nhất quán; (5) tâm lý "làm nhanh thắng nhanh".

Module 2 — Phân tích mô tả (Descriptive Analytics)

2.1. Trực quan hóa dữ liệu

> **Con số phải nhớ:** Bộ não con người xử lý hình ảnh nhanh hơn văn bản tới **60.000 lần**.

5 mục đích của trực quan hóa

1. **Dễ hiểu** — biến phức tạp thành đơn giản, giảm tải nhận thức (biến danh sách chỉ đường thành tấm bản đồ).
2. **Cải thiện quyết định** — nắm bắt tình hình nhanh, so sánh lựa chọn dễ, ra quyết định dựa trên bằng chứng.
3. **Tăng mức độ tương tác** — hình ảnh thu hút, dễ chia sẻ, khơi gợi thảo luận trong cuộc họp.
4. **Xác định mô hình & xu hướng** — 4 thứ mắt người phát hiện tốt: Xu hướng · Tính thời vụ · Mối tương quan · Điểm bất thường (outlier).
5. **Kể chuyện bằng dữ liệu** — 3 yếu tố: **Dữ liệu + Hình ảnh + Bối cảnh & Lời kể**.

3 câu hỏi bắt buộc trước khi vẽ biểu đồ

- **Ai** là người đọc và diễn giải? Họ có đủ kiến thức hiểu thuật ngữ không?
- **Kỳ vọng** của người đọc là gì, kiểu dữ liệu nào hữu dụng nhất với họ?
- **Mục đích** của trực quan hóa là gì, người đọc sẽ hành động thế nào sau đó?

2.2. Bảng chọn biểu đồ — mỗi loại kể một câu chuyện

Biểu đồ	Kể câu chuyện về	Dùng khi
Bảng số liệu	Tra cứu chi tiết	Nhiều người đọc, mỗi người quan tâm một hàng; nhiều thước đo khác đơn vị. **Không dùng khi thuyết trình** (người ta ngừng nghe để đọc)
Đường (line)	**XU HƯỚNG**	Dữ liệu liên tục theo thời gian; xem tăng/giảm/đi ngang, có mùa vụ không
Cột (column)	**SO SÁNH**	Sản phẩm nào là "ngôi sao", sản phẩm nào yếu thế, chênh lệch có lớn không
Thanh (bar)	So sánh — **ưu tiên dùng**	Tên biến dài; đúng thứ tự mắt đọc (tên trước, dữ liệu sau)
Tròn (pie)	**TỶ TRỌNG**	Kênh nào đóng góp nhiều nhất, thị phần. **Không dùng để so sánh nhiều giá trị; tránh pie 3D**
Điểm phân tán	**QUAN HỆ giữa 2 biến**	Xác định 4 thứ: Hướng · Hình dạng · Độ mạnh · Giá trị ngoại lai
Bản đồ nhiệt	**SỰ TẬP TRUNG**	Thị trường trọng điểm ở đâu, khu vực tiềm năng nào đang bỏ lỡ
Thanh xếp chồng	So sánh + cơ cấu	Thông thường: so tổng và giá trị tuyệt đối từng phần. **100%**: chỉ so tỷ trọng, bỏ qua khác biệt tổng
Thác nước	Thay đổi lũy kế	Kể chi tiết "những gì đã xảy ra ở giữa": phân tích lợi nhuận, biến động nhân sự, biến động doanh thu
Dốc (slope)	Thay đổi giữa 2 mốc	So sánh nhiều hạng mục giữa hai thời điểm; so thứ hạng; phát hiện xu hướng trái ngược

5 sai lầm phổ biến

- **Quá tải thông tin** — nhồi quá nhiều KPI lên một dashboard.

- **Thiếu ngữ cảnh** — không có so sánh lịch sử hay mục tiêu.
- **Chọn biểu đồ kém** — dùng pie để so sánh nhiều giá trị, dùng biểu đồ 3D.
- **Hình ảnh gây hiểu lầm** — cắt/chỉnh trục Y, thang đo không nhất quán, màu sắc làm biến dạng dữ liệu.
- **Không tinh chỉnh cho đối tượng người đọc.**

2.3. Mô tả bằng số — 3 đặc điểm cần mô tả

> ① Trung tâm (dữ liệu tập trung ở đâu?) ② Độ biến thiên (phân tán bao nhiêu?) ③ Hình dạng (đối xứng hay lệch?)

Thước đo xu hướng trung tâm

Thước đo	Công thức	Đặc điểm / khi nào dùng
Trung bình cộng (Mean)	Tổng thể: $\mu = (\Sigma x_i) / N$ Mẫu: $\bar{x} = (\Sigma x_i) / n$	Chịu ảnh hưởng của **tất cả** phần tử → nhạy với ngoại lai. Là **điểm cân bằng** của phân phối
Trung vị (Median)	Phân vị thứ 50 n lẻ → giá trị giữa; n chẵn → TB hai giá trị giữa	Chia đôi dữ liệu. **Không nhạy với ngoại lai** → dùng khi dữ liệu lệch (thu nhập, giá nhà)
Mode	Giá trị xuất hiện nhiều nhất	Có thể có nhiều mode hoặc không có. Hữu ích nhất với dữ liệu **rời rạc / phân loại**; hiếm khi hữu ích với dữ liệu liên tục
Trung bình nhân (Geometric)	$G = \sqrt[n]{(x_1 \cdot x_2 \cdot \ldots \cdot x_n)}$	Dùng cho dữ liệu **theo tỷ lệ**: tốc độ tăng trưởng, tỷ suất lợi nhuận. Nhược: không xử lý được số 0 và số âm
Khoảng giữa (Midrange)	(x_max − x_min) / 2	Dễ tính nhưng **kém tin cậy** vì chỉ dựa trên hai giá trị cực trị

Hình dạng phân phối (skewness) — quy tắc so sánh

Hình dạng	Đuôi dài về	Quan hệ	Ví dụ trên slide
Lệch trái (âm)	Bên trái	**Mean < Median**	GPA của 158 sinh viên
Đối xứng	Cân đối hai bên	**Mean = Median = Mode**	Phân phối chuẩn
Lệch phải (dương)	Bên phải	**Mean > Median**	Chi tiêu tại 74 quán mì

Thước đo độ biến thiên

Thước đo	Công thức	Ghi nhớ
Phạm vi (Range)	R = x_max − x_min	Dễ tính, nhưng chỉ dùng 2 giá trị cực trị
Phương sai (Variance)	Tổng thể: $\sigma^2 = (1/N)\,\Sigma(x_i - \mu)^2$ Mẫu: $s^2 = \Sigma(x_i - \bar{x})^2 / (n - 1)$	Trung bình của bình phương độ lệch. Đơn vị bị **bình phương** nên khó diễn giải
Độ lệch chuẩn (Std. Dev.)	$\sigma = \sqrt{\sigma^2}$ · $s = \sqrt{s^2}$	**Cùng đơn vị với biến** → thước đo độ phân tán dùng nhiều nhất
Hệ số biến thiên (CV)	$CV = 100 \times (s / \bar{x})$ %	So sánh độ phân tán giữa các tập **khác đơn vị** (USD vs VNĐ) hoặc **khác trung bình**. CV có thể **> 100%** khi $s > \bar{x}$ (thường ở dữ liệu lệch, có ngoại lai)

> **Lưu ý khi thi:** Slide viết phương sai mẫu với mẫu số N; công thức chuẩn thống kê là **n − 1** (bậc tự do). Nếu đề yêu cầu tính phương sai mẫu, dùng n − 1 — đây cũng là cách Excel tính với hàm VAR.S / STDEV.S.

Hệ số tương quan r

> **Định nghĩa:** Đại lượng thống kê mô tả mức độ **tuyến tính** giữa các cặp quan sát trên hai biến định lượng X và Y.

> $-1 \leq r \leq +1$ · $r \approx 0$: gần như không có quan hệ tuyến tính · $r \approx +1$: đồng biến chặt · $r \approx -1$: nghịch biến chặt

- **Tương quan ≠ nhân quả.** r chỉ đo quan hệ **tuyến tính** — hai biến có quan hệ cong chặt chẽ vẫn có thể cho r ≈ 0.

Module 3 — Suy luận thống kê (Statistical Inference)

3.1. Từ mẫu đến tổng thể

- **Tổng thể (Population)**: toàn bộ đối tượng quan tâm (ví dụ: tất cả khách hàng của công ty).
- **Mẫu (Sample)**: một phần nhỏ được chọn ra để kiểm tra (ví dụ: 100 khách hàng được khảo sát).

> **Bài toán gốc:** Khảo sát 100 khách cho tỷ lệ nhận diện thương hiệu 30%. Liệu có thể tự tin nói 30% toàn bộ thị trường mục tiêu đã nhận diện, hay con số thật có thể là 20% hoặc 40%? Suy luận thống kê sinh ra để trả lời câu hỏi này.

3.2. Phân phối chuẩn và quy tắc 68–95–99,7

- Ba tính chất: (1) **đối xứng**; (2) tập trung quanh trung bình μ — đỉnh chuông đồng thời là **mean = median = mode**; (3) **độ lệch chuẩn σ quyết định độ "béo/gầy"** của chuông (σ nhỏ → cao và gầy; σ lớn → thấp và bè).

> $\approx 68\%$ dữ liệu nằm trong $\mu \pm 1\sigma$
> $\approx 95\%$ dữ liệu nằm trong $\mu \pm 2\sigma$
> $\approx 99{,}7\%$ dữ liệu nằm trong $\mu \pm 3\sigma$

> **Ví dụ ứng dụng:** Thời gian đi làm trung bình 40 phút, σ = 10 phút → **95% khả năng** một nhân viên đến văn phòng trong khoảng **20 đến 60 phút** → dùng để sắp xếp lịch họp buổi sáng.

3.3. Định lý giới hạn trung tâm (CLT)

> **Định lý quan trọng nhất trong thống kê ứng dụng:** Phân phối của **trung bình mẫu** sẽ luôn là **phân phối chuẩn**, miễn là cỡ mẫu đủ lớn — bất kể phân phối gốc của tổng thể có hình dạng kỳ quặc thế nào.

Minh họa (chuỗi cà phê, thang điểm 1–10, tổng thể 1.000 khách)

1. Lấy ngẫu nhiên 30 khách, tính trung bình mẫu = 7,2.
2. Lặp lại 1.000 lần với 1.000 mẫu khác nhau → dãy trung bình mẫu: 7,2 · 7,5 · 7,1 · 7,3 · …
3. Vẽ biểu đồ 1.000 trung bình mẫu đó → **luôn có dạng hình chuông**.

Hai hệ quả phải nhớ

- Trung bình của phân phối các trung bình mẫu **bằng** trung bình tổng thể.
- Độ lệch chuẩn của phân phối này gọi là **sai số chuẩn (Standard Error)** và **nhỏ hơn** độ lệch chuẩn tổng thể — các "trung bình mẫu" ít biến động hơn các "điểm cá nhân".

3.4. Khoảng tin cậy (Confidence Interval)

- Là khoảng giá trị xây dựng quanh trung bình mẫu, kèm **biên độ sai số** hợp lý, nhằm chứa giá trị thật của tổng thể.

> Ví dụ: khảo sát 100 khách, chi tiêu TB = 500.000đ → KTC 95% = [480.000đ ; 520.000đ]

- **Diễn giải chuẩn (dùng khi thi)**: "Nếu lặp lại quy trình lấy mẫu này 100 lần thì khoảng 95 trong 100 lần, khoảng tin cậy tạo ra sẽ chứa giá trị trung bình thật của tổng thể."
- **Cách nói đơn giản (chấp nhận trong kinh doanh)**: "Với mức độ tự tin 95%, chi tiêu trung bình của tất cả khách hàng nằm trong khoảng 480.000–520.000đ."
- **Nguyên lý**: dựa trên CLT — 95% các trung bình mẫu nằm trong **2 sai số chuẩn** so với trung bình thật; khoảng tin cậy chính là đảo ngược logic này.

3.5. Kiểm định giả thuyết — quy trình 4 bước

> **Tinh thần:** "Vô tội cho đến khi được chứng minh là có tội." H_0 là giả định vô tội; ta chỉ bác bỏ khi có bằng chứng đủ mạnh.

Ví dụ: nhà cung cấp quảng cáo cam kết CTR 3%; chạy thử 10.000 lượt hiển thị chỉ thu 250 click (CTR = 2,5%)

1. **Đặt giả thuyết** — H_0: tỷ lệ click thật là 3%. H_1: tỷ lệ click thật nhỏ hơn 3%.
2. **Thu thập bằng chứng** — dữ liệu mẫu cho CTR = 2,5%.
3. **Câu hỏi cốt lõi** — "Nếu H_0 đúng, xác suất quan sát được kết quả kém như 2,5% hoặc tệ hơn chỉ do ngẫu nhiên là bao nhiêu?"
4. **Xác suất đó chính là p-value** → so với ngưỡng α để kết luận.

> p-value < 0,05 → bác bỏ H_0 → "có bằng chứng thống kê đủ mạnh"
> p-value ≥ 0,05 → không đủ cơ sở bác bỏ H_0 → "chênh lệch có thể chỉ do ngẫu nhiên"

- Ví dụ p = 0,01 (1%): kết quả rất khó xảy ra nếu CTR thật là 3% → bác bỏ H_0, kết luận chiến dịch kém hiệu quả hơn cam kết.
- Ví dụ p = 0,20 (20%): hoàn toàn có thể do biến động ngẫu nhiên → không đủ bằng chứng bác bỏ H_0.

3.6. Kiểm định hai mẫu — A/B Testing

- **Định nghĩa**: phương pháp so sánh hai phiên bản một cách khoa học — A là bản đối chứng/bản gốc, B là biến thể đã thay đổi **một yếu tố duy nhất**.

> **Nguyên tắc vàng:** Tại một thời điểm chỉ thay đổi MỘT yếu tố. Đúng: đổi màu sắc nút. Sai: vừa đổi màu sắc vừa đổi câu chữ — vì không biết yếu tố nào tạo ra khác biệt.

- **Ví dụ**: 1.000 người xem A → chuyển đổi 4%; 1.000 người xem B (nút "Mua Ngay" màu cam) → 5,5%; chênh lệch **1,5%**. Có nên đầu tư hàng trăm triệu để đổi toàn bộ website?
- **Nền tảng CLT**: tỷ lệ chuyển đổi mỗi nhóm là biến ngẫu nhiên; **hiệu số của hai biến ngẫu nhiên cũng là biến ngẫu nhiên**, và phân phối của hiệu hai trung bình mẫu cũng có dạng chuẩn.
- **p-value ở đây**: "Giả sử A và B không khác gì nhau, xác suất tình cờ thấy chênh lệch ≥ 1,5% chỉ do may mắn là bao nhiêu?"

3.7. "Có ý nghĩa thống kê" ≠ "Quan trọng"

	Ý nghĩa thống kê	Ý nghĩa thực tiễn
Trả lời câu hỏi	Sự khác biệt này có thật không, hay chỉ do ngẫu nhiên?	Sự khác biệt có đủ lớn để tạo tác động kinh doanh và đáng để đầu tư không?
Căn cứ	p-value < 0,05	Phán quyết của nhà quản lý (chi phí – lợi ích)
Không cho biết	Khác biệt lớn hay nhỏ, có đáng hành động không	—

> **Ví dụ kinh điển:** Khảo sát 1 triệu người: B = 5,01% vs A = 5,00%. p-value rất nhỏ → **có ý nghĩa thống kê cao**, nhưng chênh lệch 0,01% **không đáng** để chi 1 tỷ đồng thay đổi hệ thống.
>
> **Câu chốt module:** Suy luận thống kê giúp chuyển từ "chúng tôi nghĩ rằng…" sang "dữ liệu cho thấy rằng…".

Module 4 — Phân tích dự báo (Predictive Analytics)

4.1. Hồi quy — khái niệm nền

> **Định nghĩa:** Phân tích hồi quy là công cụ xây dựng mô hình toán học và thống kê mô tả mối quan hệ giữa **một biến phụ thuộc Y** và **một hoặc nhiều biến độc lập (biến giải thích) X**, trong đó giá trị của tất cả các biến đều ở dạng số.

- **Hồi quy tuyến tính đơn**: chỉ 1 biến độc lập. **Hồi quy đa biến**: từ 2 biến độc lập trở lên.
- **Bước đầu tiên luôn là vẽ biểu đồ phân tán** để kiểm tra xu hướng: (a) tuyến tính → dùng hồi quy tuyến tính; (b) phi tuyến → thêm hạng tử bậc hai hoặc đổi phương pháp; (c) không có quan hệ → không dùng được.

4.2. Đường bình phương tối thiểu và phần dư

> Phần dư (residual) = Giá trị Y **thực tế** − Giá trị Y **dự báo**
> Đường hồi quy tốt nhất là đường **tối thiểu hóa tổng bình phương của phần dư (SSE)**

- **Mô hình tổng thể**: $Y = \beta_0 + \beta_1 X + \varepsilon$ · **Mô hình ước lượng từ mẫu**: $\hat{Y} = b_0 + b_1 X$

Case study xuyên suốt: giá trị thị trường 42 ngôi nhà

> Giá trị thị trường = 32.673 + 35,036 × diện tích (foot vuông) · $R^2 \approx 0{,}53$

- **Dự báo**: nhà 2.200 ft² → 32.673 + 35,036 × 2.200 = **$109.752**.
- **Diễn giải hệ số góc**: "Các yếu tố khác không đổi, diện tích tăng 1 ft² làm giá nhà tăng **$35,03**."

4.3. Công cụ Excel

Công cụ	Cách gọi	Dùng để
Hàm	=SLOPE(dữ_liệu_Y, dữ_liệu_X) =INTERCEPT(dữ_liệu_Y, dữ_liệu_X)	Lấy nhanh hệ số góc b_1 và hệ số chặn b_0
Đường xu hướng	Chuột phải vào chuỗi dữ liệu trên biểu đồ phân tán → Thêm đường xu hướng → tick **Hiển thị phương trình** và **Hiển thị giá trị R²**	Xem nhanh phương trình + R^2 trực tiếp trên biểu đồ
Regression	Data → **Data Analysis** → Regression (khai báo vùng Y, vùng X, vị trí xuất kết quả)	**Đầy đủ nhất**: R^2, Adjusted R^2, p-value, khoảng tin cậy, bảng & biểu đồ phần dư
Moving Average / Exponential Smoothing	Data → Data Analysis (khai báo dữ liệu thực tế, k hoặc hệ số giảm 1 − α)	Dự báo chuỗi thời gian ngắn hạn

> **Lưu ý kỹ thuật:** Trong hồi quy đa biến, Excel yêu cầu **các biến độc lập phải nằm ở các cột liền kề nhau** — có thể phải di chuyển cột thủ công trước khi chạy.

4.4. Đọc kết quả hồi quy — các con số quan trọng

Chỉ số	Ý nghĩa	Cách đọc / ngưỡng
Hệ số góc b_1	Tác động biên của X lên Y	"X tăng 1 đơn vị ⇒ Y thay đổi b_1, các yếu tố khác không đổi"
R^2 (hệ số xác định)	% biến động của Y được mô hình giải thích	Từ 0 (không phù hợp) đến 1 (hoàn toàn phù hợp). R^2 = 0,53 → "**53% biến động giá nhà được giải thích bởi diện tích**"

Chỉ số	Ý nghĩa	Cách đọc / ngưỡng
R^2 hiệu chỉnh	R^2 đã hiệu chỉnh theo cỡ mẫu và **số biến độc lập**	Dùng để **so sánh các mô hình khác số biến**; có thể tăng hoặc giảm khi thêm biến
p-value của hệ số	Biến X có ý nghĩa thống kê không	**p < 0,05** ⇒ bác bỏ H_0 với độ tin cậy 95% ⇒ biến X **có giá trị giải thích**
Khoảng tin cậy 95%	Giới hạn dưới / trên của hệ số thật	Nếu KTC **không chứa 0** ⇒ biến có ý nghĩa. Có thể dùng thay cho kiểm định p-value
Sai số chuẩn của ước lượng	Độ biến động giữa Y quan sát và Y dự báo	Sai số điển hình của dự báo, tính bằng đơn vị của Y
Multiple R	r — hệ số tương quan mẫu	Từ −1 đến +1; mang dấu âm nếu hệ số góc âm

> **Ví dụ về rủi ro ước lượng:** Với nhà 1.750 ft^2, nếu tham số thật nằm ở cực trị dưới của khoảng tin cậy thì ước tính có thể thấp tới 14.823 + 24,59 × 1.750 = **$57.856** — khác rất xa dự báo điểm.

4.5. Phân tích phần dư và kiểm tra giả định

> Phần dư chuẩn hóa = Phần dư ÷ Độ lệch chuẩn của phần dư

- **Ngưỡng ngoại lai: phần dư chuẩn nằm ngoài ±2 độ lệch chuẩn** là ngoại lai tiềm năng, cần xem lại (slide có điểm với phần dư chuẩn = 4,53).
- **Ví dụ tính**: giá trị thực $90.000, dự báo $96.159,13 → phần dư = 90.000 − 96.159,13 = **−$6.159,13**.

Đọc biểu đồ phần dư

Hình dạng biểu đồ phần dư	Chẩn đoán	Xử lý
Rải **ngẫu nhiên quanh 0**	Mô hình tốt, giả định thỏa mãn	Giữ nguyên
Hình **chữ U** (hoặc U ngược)	Quan hệ **phi tuyến**, mô hình tuyến tính không phù hợp	Thêm hạng tử bậc hai X^2
Hình **phễu** (loe dần)	Phương sai sai số **không đều**	Biến đổi biến (log), hoặc dùng mô hình khác
Điểm nằm xa (chuẩn > 2)	Ngoại lai	Kiểm tra lỗi nhập liệu / sự kiện bất thường

4.6. Hồi quy đa biến — 3 cái bẫy

> $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$

Bẫy 1 — R^2 luôn tăng khi thêm biến

- "Việc thêm một biến độc lập vào mô hình sẽ **luôn luôn** dẫn đến R^2 bằng hoặc lớn hơn R^2 của mô hình ban đầu." → **So sánh mô hình bằng R^2 hiệu chỉnh**, không phải R^2.

Quy trình xây dựng mô hình (loại lùi) — 3 bước

1. Xây dựng mô hình với **tất cả** các biến độc lập có sẵn.
2. Xác định biến có **p-value lớn nhất vượt quá mức ý nghĩa α**.
3. Loại biến đó rồi đánh giá R^2 hiệu chỉnh: **tăng ⇒ loại hẳn biến**; **giảm ⇒ giữ lại** biến dù không có ý nghĩa thống kê.

> **Nguyên tắc:** Chỉ loại **từng biến một**, không loại đồng loạt mọi biến có p > α. Mô hình tốt là mô hình đơn giản nhất có thể — **nguyên tắc tối giản (parsimony)**.

Bẫy 2 — Đa cộng tuyến (Multicollinearity)

- Xảy ra khi các **biến độc lập tương quan chặt với nhau** và dự báo lẫn nhau tốt hơn dự báo biến phụ thuộc.
- **Ngưỡng kiểm tra: |r| giữa hai biến X vượt quá 0,7.**
- Hậu quả: khó tách biệt tác động của từng biến · **dấu của hệ số có thể trái ngược lý thuyết** · hệ số khó diễn giải · **p-value bị thổi phồng**.

Bẫy 3 — Quá khớp (Overfitting)

- "Mô hình quá khớp với dữ liệu mẫu đến mức không còn tổng quát hóa được cho quần thể mà bạn thực sự quan tâm."
- Nguyên nhân: thêm quá nhiều biến. Giảm thiểu bằng **logic, trực giác, lý thuyết và tính tối giản**.

4.7. Biến phân loại, phi tuyến và tương tác

Biến giả (dummy variable)

- Hồi quy yêu cầu dữ liệu số → dữ liệu phân loại phải mã hóa.
- **2 phân loại** → 1 biến nhị phân 0/1. **k phân loại (k > 2)** → tạo **k − 1 biến nhị phân**; tạo đủ k biến sẽ gây **đa cộng tuyến**.
- Ví dụ: Y = giá nhà; X_1 = diện tích; X_2 = có mặt tiền (0/1) → β_2 là chênh lệch giá của nhà có mặt tiền so với nhà không có mặt tiền.

Hạng tử phi tuyến

> Đa thức bậc hai: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ (β_1 = tác động tuyến tính, β_2 = tác động cong)

- Dấu hiệu nhận biết: **biểu đồ phần dư có dạng chữ U**.
- Mô hình này **vẫn tuyến tính theo tham số β** nên vẫn chạy được bằng công cụ hồi quy tuyến tính của Excel.
- **Ví dụ**: Doanh số = 142.850 − 3.643,17 × nhiệt độ + 23,3 × nhiệt độ².

Tương tác (Interaction)

- "Tương tác xảy ra khi **tác động của một biến phụ thuộc vào một biến khác**."
- Kiểm tra bằng cách tạo biến mới $\mathbf{X_3 = X_1 \times X_2}$ và xét ý nghĩa thống kê của nó.

> **Quy tắc bắt buộc:** Nếu X_3 có ý nghĩa thống kê thì **phải luôn giữ cả X_1 và X_2** trong mô hình, dù hai biến này không có ý nghĩa thống kê.

4.8. Các kỹ thuật dự báo

- Hai nhóm phương pháp: **(1) định tính & phán đoán** và **(2) mô hình thống kê chuỗi thời gian**.

Dự báo định tính

- Dựa vào kinh nghiệm và trực giác; **cần thiết khi dữ liệu lịch sử không có sẵn** hoặc khi dự báo cho **tương lai xa**.
- **Phương pháp tương đồng quá khứ**: dự báo bằng cách so sánh với các tình huống trước đó.

> **Case study giá dầu 1988:** Đầu 1988 giá ~22 USD/thùng, giữa 1988 giảm còn 11 USD/thùng do cung vượt cầu. Trong quá khứ OPEC thường tăng giá → phương pháp tương đồng sẽ dự báo giá cao hơn. Nhưng

> giá tiếp tục giảm ngay cả khi OPEC cắt giảm sản lượng → **tương đồng quá khứ không phải lúc nào cũng giải thích được thực tế hiện tại**.

- **Hạn chế**: (1) chủ quan và có **thiên kiến** — thiên kiến lạc quan/bi quan, **thiên kiến mỏ neo**, **tư duy nhóm (groupthink)**; (2) không nhất quán, khó lặp lại, không kiểm chứng khoa học được; (3) tốn thời gian, khó mở rộng.

Phương pháp Delphi — 3 nguyên lý

1. **Tính ẩn danh để loại bỏ thiên kiến** (quan trọng nhất) — triệt tiêu hiệu ứng **HiPPO** (Highest Paid Person's Opinion) và groupthink; mọi ý kiến có **trọng số như nhau ở vòng đầu**.
2. **Phản hồi có kiểm soát** — sau mỗi vòng chỉ cung cấp **bản tóm tắt thống kê** (trung bình, khoảng dao động), không để tranh luận tự do.
3. **Cơ chế tự điều chỉnh** — người có dự báo ngoại lai tự soát lại lập luận và điều chỉnh ở vòng sau.

- **Hạn chế Delphi**: chậm, không hợp quyết định cần nhanh · phụ thuộc chất lượng chuyên gia · nguy cơ **đồng thuận giả tạo** · thiếu tranh luận sâu do ẩn danh.

Chuỗi thời gian (Time Series)

- **Định nghĩa**: một luồng dữ liệu lịch sử, chẳng hạn doanh số bán hàng theo tuần (T = số kỳ).

Thành phần	Mô tả	Ví dụ
Hành vi ngẫu nhiên	Biến động không quy luật. **Chuỗi thời gian dừng chỉ có thành phần này**	Nhiễu quanh mức trung bình
Xu hướng (Trend)	Hướng đi lên hoặc xuống trong dài hạn	Tổng sản lượng & tiêu thụ năng lượng
Hiệu ứng theo mùa	Lặp lại theo khoảng thời gian **cố định**: năm, tháng, tuần hoặc ngày	Lượng sử dụng gas theo mùa
Hiệu ứng chu kỳ	Lên xuống trong khung thời gian **dài hơn nhiều, vài năm**	Lãi suất liên bang

Hai mô hình dự báo cơ bản

- **Trung bình trượt (Moving Average)** — "phương pháp làm trơn dựa trên ý tưởng lấy trung bình các biến động ngẫu nhiên để xác định hướng cơ bản của chuỗi". Dự báo kỳ kế tiếp = trung bình của **k quan sát gần nhất**.

> Quy tắc: k càng lớn ⇒ mô hình dự báo càng trơn, vì giá trị cực trị ít tác động hơn

- **Làm trơn hàm mũ (Exponential Smoothing) đơn giản**:

> $F(t+1) = \alpha \cdot A(t) + (1-\alpha) \cdot F(t)$
> $F(t+1)$: dự báo kỳ t+1 · $F(t)$: dự báo kỳ t · $A(t)$: giá trị thực tế quan sát ở kỳ t
> α: hằng số làm trơn, $0 < \alpha < 1$ (α lớn ⇒ phản ứng nhanh với dữ liệu mới)

Đo sai số dự báo (4 chỉ số)

Chỉ số	Tên đầy đủ	Đặc điểm
MAD	Mean Absolute Deviation — Độ lệch tuyệt đối trung bình	Trung bình sai số ; cùng đơn vị với dữ liệu, dễ hiểu
MSE	Mean Squared Error — Sai số bình phương trung bình	Bình phương sai số → **phạt nặng sai số lớn**; đơn vị bị bình phương
RMSE	Root Mean Squared Error	Căn của MSE → đưa về cùng đơn vị dữ liệu

Chỉ số	Tên đầy đủ	Đặc điểm
MAPE	Mean Absolute Percentage Error — Sai số phần trăm tuyệt đối trung bình	Tính theo **%** → **so sánh được giữa các chuỗi khác đơn vị/quy mô**

Module 5 — Giới thiệu Data Mining

> **Định nghĩa:** Khai thác dữ liệu (data mining) là việc hiểu rõ hơn về các đặc điểm và mẫu hình giữa các biến trong **cơ sở dữ liệu lớn** bằng nhiều công cụ thống kê và phân tích.

Kỹ thuật	Định nghĩa	Trả lời câu hỏi
Phân tích cụm (Clustering)	Nhận dạng các nhóm mà trong đó các phần tử có điểm tương đồng về mặt nào đó	Khách hàng của tôi chia thành mấy nhóm?
Phân loại (Classification)	Phân tích dữ liệu để dự đoán cách phân loại một phần tử dữ liệu **mới**	Khách hàng mới này thuộc nhóm nào?
Luật kết hợp (Association Rules)	Xác định mối liên hệ tự nhiên giữa các biến để xây dựng quy tắc tiếp thị mục tiêu hoặc đề xuất mua hàng	Mua A thì thường mua kèm gì?

5.1. Phân tích cụm (Cluster Analysis)

- Còn gọi là **phân đoạn dữ liệu (segmentation)**.
- **Mục đích**: nhóm một tập đối tượng thành các cụm sao cho **các đối tượng trong cùng cụm có quan hệ chặt chẽ hơn** so với đối tượng ở cụm khác — tương đồng cao trong cụm, khác biệt cao giữa các cụm.

> Phép đo khoảng cách: **Khoảng cách Euclide** — khoảng cách đường thẳng giữa hai điểm
> $d = \sqrt{[(x_1 - x_2)^2 + (y_1 - y_2)^2]}$

- **Hai phương pháp**: phân cụm **chia tách (divisive)** — từ một cụm lớn tách dần; phân cụm **hợp nhất (agglomerative)** — từ từng điểm gộp dần lên.
- **Không có nhãn cho trước** (học không giám sát) — khác với phân loại.

5.2. Phân loại (Classification)

> **Định nghĩa:** Phân loại một quan sát/bản ghi vào một trong nhiều nhãn (label) **đã xác định trước** dựa trên các thuộc tính dữ liệu.

k-láng giềng gần nhất (k-NN)

- Tìm các điểm dữ liệu trong CSDL có giá trị bằng số **tương tự** với bản ghi mới; "tương đồng" = **độ gần**, đo bằng **khoảng cách Euclide chuẩn hóa**.
- **Quy tắc**: tìm k hàng xóm gần nhất, nhãn nào **chiếm đa số** thì gán nhãn đó.
- **1-NN**: gán cùng nhãn với bản ghi gần nhất. **3-NN**: gán nhãn có **ít nhất 2** trong 3 điểm lân cận.
- **Chọn k**: thử nhiều giá trị k khác nhau và kiểm tra kết quả để xác định **k tối ưu**.

Phân tích phân biệt (Discriminant Analysis)

- Sử dụng các lớp xác định trước dựa trên tập các **hàm phân biệt tuyến tính** của các biến dự báo.

> **Quy tắc hay ra đề:** Nếu chỉ có **hai nhãn** → có thể áp dụng phân tích hồi quy. Khi có **nhiều hơn hai nhãn** → không thể dùng hồi quy tuyến tính, phải dùng phần mềm chuyên biệt.

5.3. Khai phá luật kết hợp (Association Rule Mining)

- Còn gọi là **phân tích mối quan hệ**; khám phá các liên kết/tương quan trong tập dữ liệu lớn — xác định các thuộc tính **thường xuất hiện cùng nhau**.
- Ứng dụng điển hình: **phân tích giỏ hàng thị trường (market basket analysis)**.
- **Luật kết hợp có dạng câu lệnh nếu–thì**: tiền đề (Nếu…) → kết quả (Thì…).

Ba thước đo mức độ kết hợp (phải thuộc)

Thước đo	Cách tính	Trả lời câu hỏi	Ngưỡng
Độ hỗ trợ (Support)	% giao dịch chứa **tất cả** sản phẩm trong luật	Sự kết hợp này có **phổ biến** không?	Càng cao càng đáng chú ý
Độ tin cậy (Confidence)	Trong các giao dịch có vế "Nếu", bao nhiêu % cũng có vế "Thì"	Nếu khách mua A thì **khả năng** mua B là bao nhiêu?	Càng cao luật càng chắc
Độ tăng cường (Lift)	Độ tin cậy ÷ mức độ phổ biến của sản phẩm ở vế "Thì"	Mua A làm **tăng bao nhiêu lần** khả năng mua B so với bình thường?	**Lift > 1 ⇒ luật có giá trị**; càng cao luật càng mạnh

Case study trên slide: Dữ liệu mua máy tính cá nhân tùy chỉnh với 4 thuộc tính (Bộ vi xử lý · Kích thước màn hình · Bộ nhớ · Ổ cứng). Tương quan cao nhất là giữa **Intel Core i7 và Ổ cứng 750 GB — 8/67 giao dịch** có cả hai linh kiện.

Module 6 — Phân tích đề xuất & Tối ưu hóa (Prescriptive)

> **Định nghĩa:** Tối ưu hóa là quá trình lựa chọn các giá trị của **các biến quyết định** nhằm **tối thiểu hóa hoặc tối đa hóa** một đại lượng quan tâm.

- Ứng dụng rộng: vận hành, chuỗi cung ứng, tài chính, marketing…

6.1. Ba thành phần của mọi bài toán tối ưu hóa

Thành phần	Là gì	Câu hỏi để tìm ra
Biến quyết định	Những con số mà nhà quản lý cần tìm ra để ra quyết định	Tôi đang phải quyết định **cái gì**? (nhớ ghi rõ **đơn vị đo lường**)
Hàm mục tiêu	Đại lượng cần tối đa hóa hoặc tối thiểu hóa	Mục tiêu cuối cùng là gì? Tối đa lợi nhuận hay tối thiểu chi phí?
Ràng buộc	Giới hạn, quy tắc, điều kiện ngăn cản việc đạt mục tiêu vô hạn	Cái gì đang **giới hạn** tôi? (nguồn lực, công suất, thị trường, hợp đồng)

6.2. Case study: Công ty trượt tuyết SSC

- SSC sản xuất hai mẫu ván trượt: **Jordanelle** và **Deercrest**. Cần chọn số lượng sản xuất mỗi mẫu để **tối đa hóa lợi nhuận**. Quy trình qua hai giai đoạn: **chế tạo** và **hoàn thiện**.

Dữ kiện	Jordanelle	Deercrest	Năng lực
Giờ **chế tạo** / đôi	3,5 giờ	4 giờ	12 công nhân × 7 giờ = **84 giờ/ngày**
Giờ **hoàn thiện** / đôi	1 giờ	1,5 giờ	3 công nhân × 7 giờ = **21 giờ/ngày**
Lợi nhuận ròng / đôi	50 USD	65 USD	—
Ràng buộc thị trường	—	—	Số Deercrest bán được **ít nhất gấp đôi** Jordanelle

Bước 1 — Biến quyết định

> J = số đôi ván Jordanelle sản xuất mỗi ngày
> D = số đôi ván Deercrest sản xuất mỗi ngày

Bước 2 — Hàm mục tiêu

> Tối đa hóa: Lợi nhuận ròng = $50 \cdot J + 65 \cdot D$

Bước 3 — Hệ ràng buộc

> $3{,}5 \cdot J + 4 \cdot D \leq 84$ (năng lực chế tạo)
> $1 \cdot J + 1{,}5 \cdot D \leq 21$ (năng lực hoàn thiện)
> $D \geq 2 \cdot J$ (ràng buộc thị trường)
> $J \geq 0,\ D \geq 0$ (ràng buộc không âm)

> **Ghi nhớ khi làm bài:** Luôn viết đủ **ràng buộc không âm**. Bài toán hai biến giải được bằng **phương pháp đồ thị** (vẽ miền khả thi, kiểm tra các đỉnh); bài toán nhiều biến giải bằng **Excel Solver**.

Phụ lục A — Bảng tra nhanh: gặp bài toán này thì dùng kỹ thuật nào

Tình huống / câu hỏi kinh doanh	Kỹ thuật	Module
Mô tả tình hình bán hàng quý vừa qua cho ban giám đốc	Trực quan hóa + thống kê mô tả (mean, median, độ lệch chuẩn)	2
So sánh mức độ ổn định doanh thu giữa hai chi nhánh có quy mô rất khác nhau	**Hệ số biến thiên (CV)**	2
Dữ liệu có vài giá trị cực lớn làm lệch trung bình	Dùng **trung vị** thay cho trung bình	2
Tính tốc độ tăng trưởng trung bình qua nhiều năm	**Trung bình nhân**	2
Khảo sát 200 khách, muốn suy ra cả tệp khách hàng	**Khoảng tin cậy**	3
Đổi thiết kế landing page, tỷ lệ chuyển đổi tăng — thật hay ngẫu nhiên?	**A/B test + p-value** (ngưỡng 0,05)	3
Nhà cung cấp cam kết một chỉ số, thực tế thấp hơn — có bằng chứng không?	**Kiểm định giả thuyết một mẫu**	3
Chi bao nhiêu cho quảng cáo thì ra bao nhiêu doanh thu?	**Hồi quy tuyến tính**	4
Nhiều yếu tố cùng ảnh hưởng đến kết quả, yếu tố nào thực sự quan trọng?	**Hồi quy đa biến** + đọc p-value từng biến	4
Biến đầu vào là loại hình/nhóm (có/không, 3 kênh…)	**Biến giả**, k phân loại → k−1 biến	4
Quan hệ giữa X và Y có dạng cong	Thêm **hạng tử bậc hai X^2**	4
Dự báo doanh số tuần tới từ dữ liệu lịch sử	**Chuỗi thời gian**: trung bình trượt / làm trơn hàm mũ	4
Sản phẩm hoàn toàn mới, không có dữ liệu lịch sử	**Dự báo định tính / Delphi**	4
So sánh độ chính xác của các mô hình dự báo khác đơn vị	**MAPE**	4
Chia khách hàng thành các nhóm để chăm sóc khác nhau	**Phân cụm**	5
Dự đoán khách hàng mới thuộc nhóm nào	**Phân loại / k-NN**	5
Xây dựng gợi ý “mua kèm”, combo sản phẩm	**Luật kết hợp** (support, confidence, lift)	5
Nguồn lực có hạn, phải phân bổ sao cho lợi nhuận cao nhất	**Tối ưu hóa / quy hoạch tuyến tính**	6

Phụ lục B — Những con số và ngưỡng phải thuộc

Con số	Ý nghĩa
60.000 lần	Tốc độ não xử lý hình ảnh so với văn bản
$-1 \le r \le +1$	Phạm vi hệ số tương quan
68 – 95 – 99,7%	Tỷ lệ dữ liệu trong $\mu \pm 1\sigma, \pm 2\sigma, \pm 3\sigma$ (phân phối chuẩn)
$p < 0,05$	Ngưỡng bác bỏ H_0 với độ tin cậy 95%
KTC 95% không chứa 0	Hệ số hồi quy có ý nghĩa thống kê
$0 \le R^2 \le 1$	0 = không phù hợp, 1 = hoàn toàn phù hợp
±2 độ lệch chuẩn	Ngưỡng phần dư chuẩn hóa để xác định ngoại lai
$\lvert r \rvert > 0,7$	Ngưỡng cảnh báo đa cộng tuyến giữa hai biến độc lập
$k - 1$	Số biến giả cần tạo cho biến phân loại có k nhóm
Lift > 1	Luật kết hợp có giá trị
$0 < \alpha < 1$	Phạm vi hằng số làm trơn trong exponential smoothing
3V	Volume – Velocity – Variety (Big Data)

Bẫy khái niệm hay bị nhầm

- **Tương quan ≠ nhân quả** — r cao không chứng minh X gây ra Y.
- **Có ý nghĩa thống kê ≠ quan trọng về mặt kinh doanh** — cỡ mẫu lớn làm mọi khác biệt tí hon đều "có ý nghĩa".
- **$R^2 \neq R^2$ hiệu chỉnh** — chỉ R^2 hiệu chỉnh mới dùng được để so sánh mô hình khác số biến.
- **Độ tin cậy ≠ tính hợp lệ** — chính xác nhất quán không có nghĩa là đo đúng thứ cần đo.
- **Phân cụm ≠ phân loại** — phân cụm không có nhãn trước, phân loại có nhãn xác định trước.
- **Support ≠ Confidence** — support là % trên **toàn bộ** giao dịch, confidence là % trên **các giao dịch có vế Nếu**.