

Copyright Notice

Tài liệu học "**Mastering Excel for Data Analysis**" này được phân phối **miễn phí** và **chỉ được sử dụng cho mục đích học tập cá nhân**. Không sử dụng trong các khóa học thương mại hoặc đào tạo doanh nghiệp.

Tất cả nội dung bao gồm slides, bài tập thực hành trên Excel được cấp phép theo **Giấy phép Attribution-NonCommercial 4.0 International**, cho phép sử dụng phi thương mại, không được phái sinh và phải ghi nguồn đầy đủ khi sử dụng hoặc trích dẫn.

© 2025 Nguyễn Thái Hà. Vui lòng không sao chép, phân phối lại, biên tập hoặc sử dụng cho mục đích thương mại mà không có sự cho phép bằng văn bản. Để biết thêm chi tiết về giấy phép, vui lòng tham khảo: [Attribution-NonCommercial 4.0 International](#)

Mastering Excel for Data Analysis

Tác giả: Nguyễn Thái Hà (Ph.D)

Mục tiêu của bộ tài liệu



Học Excel Từ Cơ Bản Đến Nâng Cao

Nắm vững các tính năng và công cụ của Excel từ căn bản đến chuyên sâu cho phân tích dữ liệu



Nắm Vững Kiến Thức Phân Tích Dữ Liệu Cơ Bản

Thành thạo quy trình phân tích dữ liệu sử dụng Excel



Sử Dụng Excel để Tìm Hiểu Xu Hướng Dữ Liệu

Vận dụng các phương pháp thống kê để nhận diện, phân tích và diễn giải chính xác các xu hướng ẩn trong dữ liệu doanh nghiệp



Sử Dụng Excel Trực Quan Hóa Dữ Liệu

Thành thạo sử dụng công cụ trực quan hóa dữ liệu, tạo ra các biểu đồ và dashboard thuyết phục cho việc ra quyết định



Phân Tích Nâng Cao & Tối Ưu Hóa

Thực hiện kiểm định giả thuyết, xây dựng mô hình hồi quy đa biến và áp dụng công cụ Excel Solver để tối ưu hóa quy trình kinh doanh

Tổng Quan Nội Dung Tài Liệu

Phần 1: Hiểu Tổng Quan Về Phân Tích Dữ Liệu

Khái niệm cơ bản và ứng dụng Excel trong phân tích dữ liệu.

Phần 2: Hiểu Xu Hướng Dữ Liệu Qua Thống Kê Cơ Bản

Phương pháp thống kê phát hiện xu hướng trong dữ liệu.

Phần 3: Trực Quan Hóa Dữ Liệu

Tạo biểu đồ và dashboard hiệu quả từ dữ liệu.

Phần 4: Kiểm Định Giả Thuyết (Hypothesis Testing)

Phương pháp đưa ra kết luận dựa trên kiểm định thống kê.

Phần 5: Tiền Xử Lý Dữ Liệu

Kỹ thuật làm sạch và chuẩn bị dữ liệu.

Phần 6: Sử dụng các mô hình hồi quy đơn biến và đa biến

Xây dựng mô hình dự báo từ dữ liệu kinh doanh.

Phần 7: Tối ưu hoá

Sử dụng Solver để tối ưu hoá quyết định kinh doanh.

Phần 1: Hiểu Tổng Quan Về Phân Tích Dữ Liệu

Dữ Liệu là gì?

Dữ liệu (data) là các sự kiện thô, các con số, hoặc văn bản chưa được xử lý và tổ chức. Dữ liệu tự thân không mang nhiều ý nghĩa cho đến khi được phân tích và diễn giải.

Dữ liệu (Data)

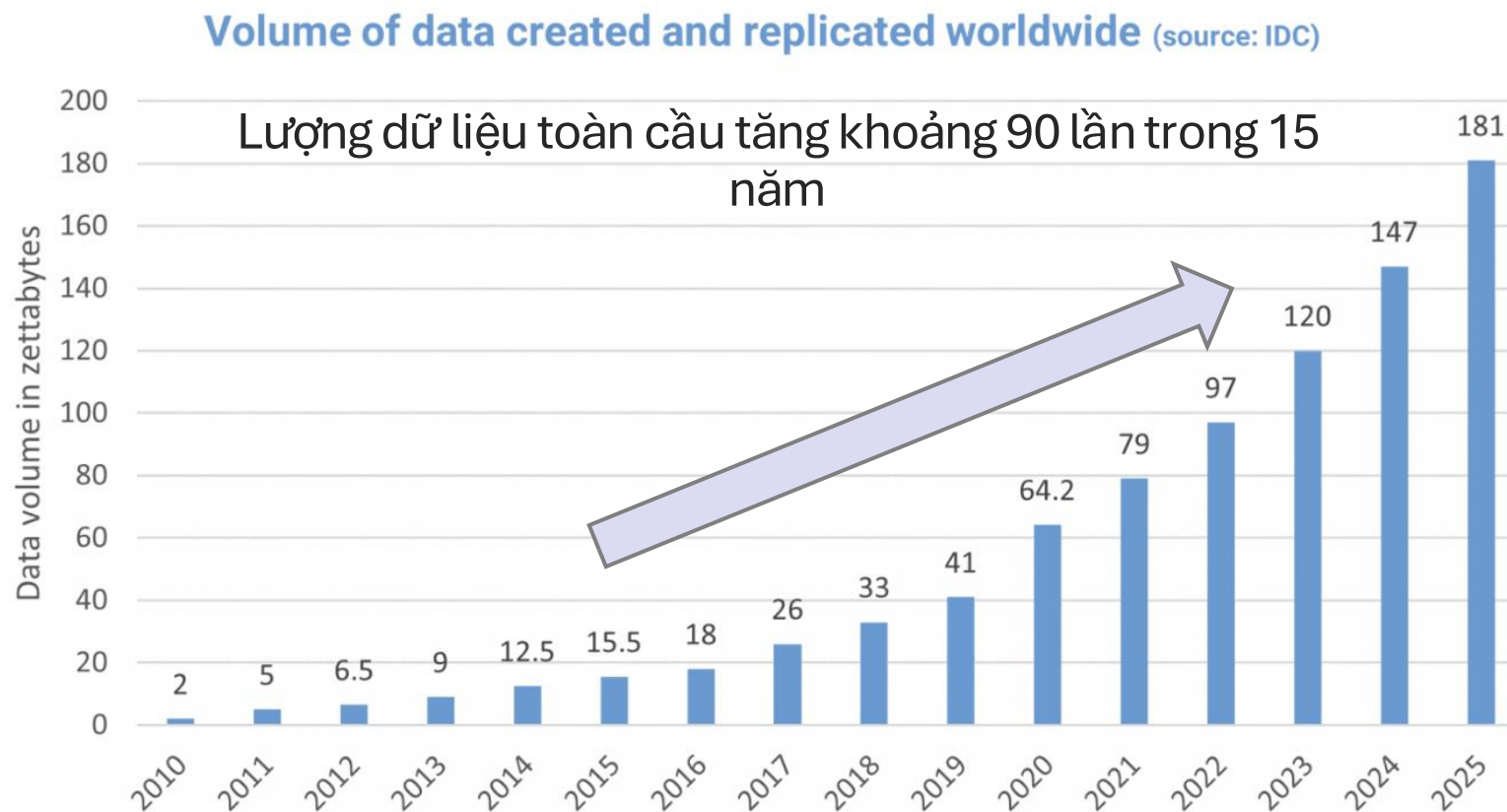
- Dạng thô, chưa qua xử lý
- Thường ở dạng số, ký tự, hình ảnh
- Không có ngữ cảnh hoặc ý nghĩa rõ ràng
- Ví dụ: 15.000.000, 42%, "TPHCM"
- Thường được lưu trữ trong cơ sở dữ liệu

Thông tin (Information)

- Dữ liệu đã được xử lý và tổ chức
- Có ngữ cảnh và ý nghĩa rõ ràng
- Hỗ trợ việc đưa ra quyết định
- Ví dụ: "Doanh thu tháng 6 là 15 triệu đồng, tăng 42% so với TPHCM"
- Được trình bày dưới dạng báo cáo, dashboard

Sự bùng nổ của dữ liệu toàn cầu

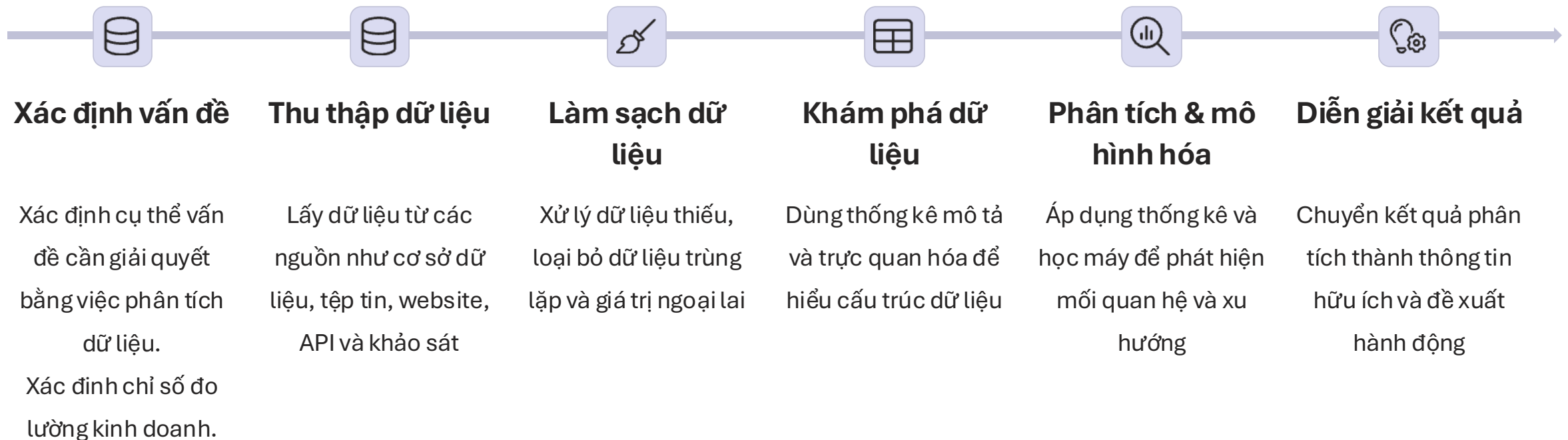
Trong hơn một thập kỷ qua, khối lượng dữ liệu toàn cầu đã tăng trưởng theo cấp số nhân, từ chỉ 2 zettabytes năm 2010 lên đến 181 zettabytes dự kiến vào năm 2025. Phân tích dữ liệu là chìa khoá để khai thác dữ liệu một cách hiệu quả.



Phân Tích Dữ Liệu là gì?

Phân tích dữ liệu là quá trình kiểm tra, biến đổi và phân tích dữ liệu nhằm khám phá thông tin giá trị, rút ra kết luận và hỗ trợ đưa ra quyết định chính xác dựa trên dữ liệu, thay vì cảm tính.

Quy Trình Phân Tích Dữ Liệu



Lợi ích của Phân Tích Dữ Liệu trong Doanh Nghiệp

Phân tích dữ liệu chuyển biến dữ liệu thành thông tin có ích giúp ra quyết định chính xác dựa trên bằng chứng xác thực, nhận diện xu hướng và tối ưu hiệu quả kinh doanh



Ra quyết định dựa trên bằng chứng

Thay vì dựa vào trực giác, các tổ chức sử dụng phân tích dữ liệu để đưa ra quyết định đầu tư và chiến lược dựa trên bằng chứng cụ thể.



Phát hiện xu hướng mới

Phân tích mạng xã hội giúp nhận biết sớm sự thay đổi trong sở thích người tiêu dùng.



Dự đoán nhu cầu thị trường

Sử dụng dữ liệu quá khứ để dự báo nhu cầu và tối ưu hàng tồn kho theo khu vực.



Tối ưu quy trình sản xuất

Áp dụng phân tích dữ liệu để giảm thiểu lãng phí trong sản xuất.



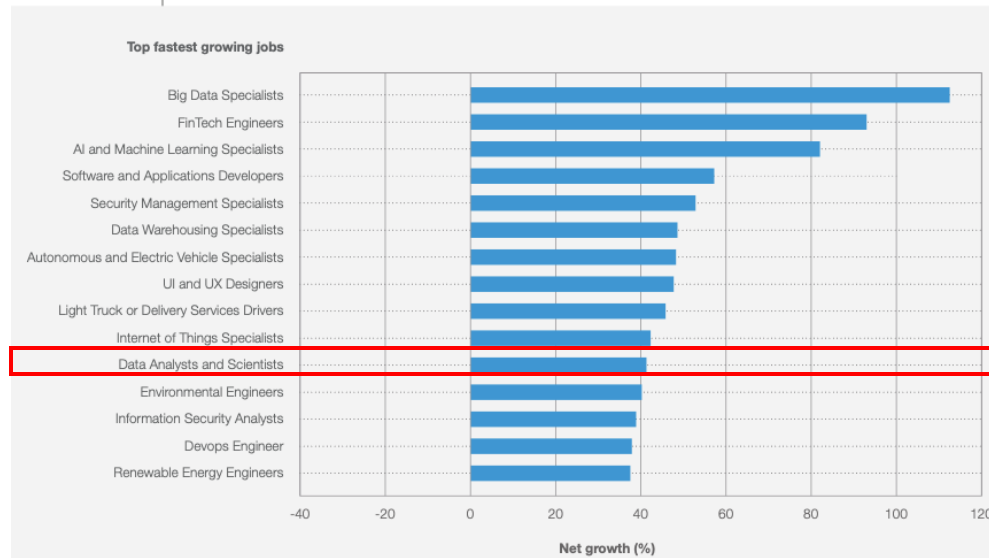
Cá nhân hóa trải nghiệm khách hàng

Phân tích hành vi người dùng giúp đề xuất nội dung phù hợp, nâng tỷ lệ giữ chân khách hàng.

Lý do nên học Phân Tích Dữ Liệu

Phân tích dữ liệu mở ra cơ hội nghề nghiệp và phát triển tư duy phân tích - kỹ năng cốt lõi được nhiều nhà tuyển dụng đánh giá cao

FIGURE 2.2 Fastest-growing and fastest-declining jobs, 2025-2030
Top jobs by fastest net growth and net decline, projected by surveyed employers

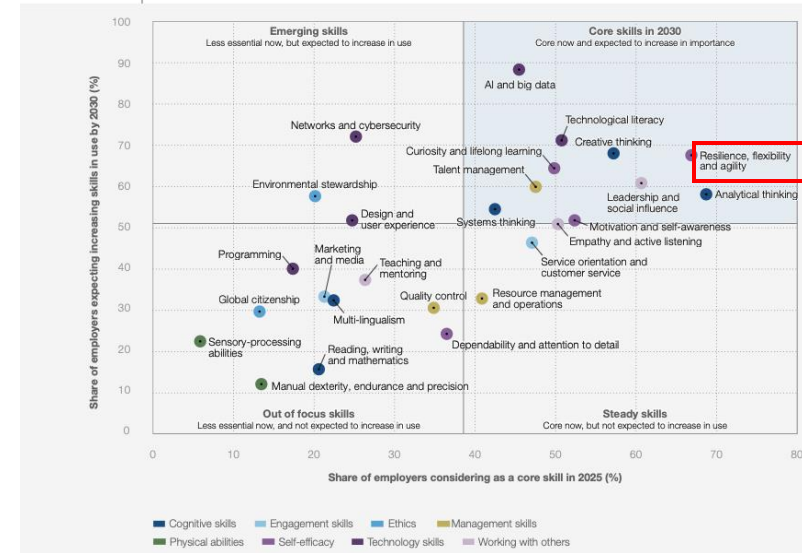


Cơ hội nghề nghiệp rộng mở

"Các nhà phân tích và khoa học dữ liệu thuộc nhóm nghề phát triển nhanh nhất" - *Future of Jobs Report*, tr.19

"Tại Nhật Bản, Chuyên gia An ninh thông tin và Phân tích dữ liệu được dự đoán là nghề tăng trưởng hàng đầu" - *Future of Jobs Report*, tr.65

FIGURE 3.6 Core skills in 2030
Share of employers considering skills to be a core skill in 2025 and share of employers expecting skills to increase in importance by 2030.






Tư duy phân tích là kỹ năng được săn đón nhất

"Tư duy phân tích vẫn là kỹ năng cốt lõi được các nhà tuyển dụng săn đón nhất" - *Future of Jobs Report*, tr.6

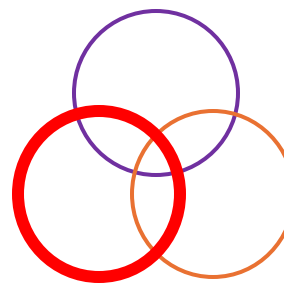
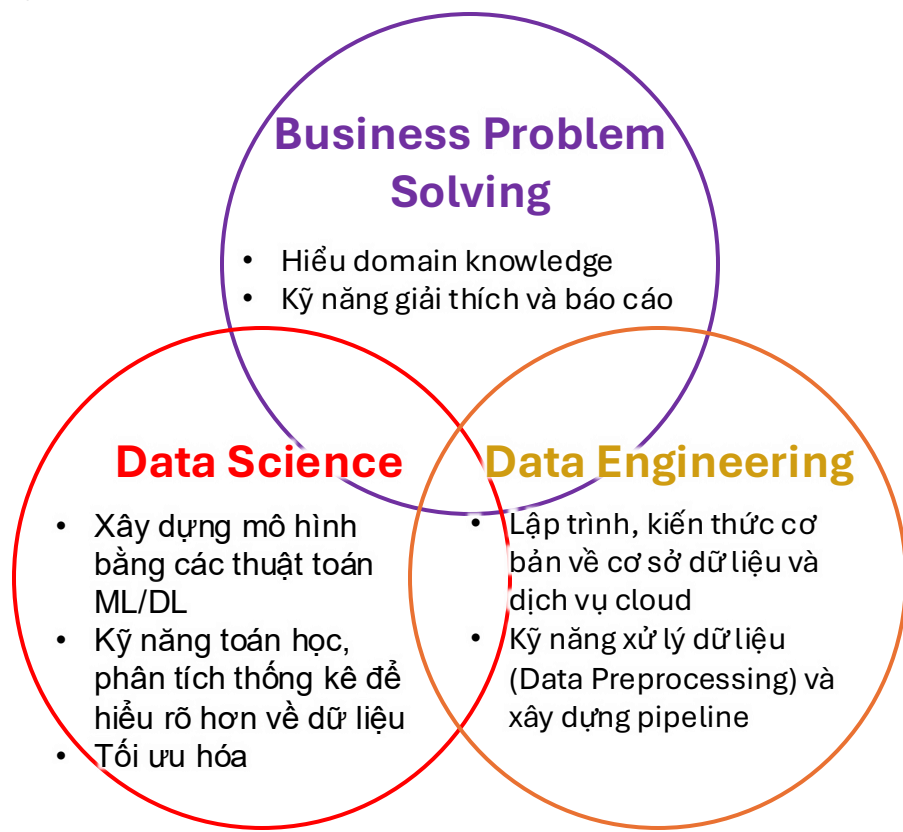
Công việc trong ngành phân tích dữ liệu

Ngành phân tích dữ liệu thường chia thành ba vai trò chính, mỗi vai trò có thế mạnh riêng và cùng nhau tạo nên một hệ sinh thái dữ liệu toàn diện.

Vai trò	Vai trò	Điểm mạnh	Công cụ phổ biến
Data scientist 	<ul style="list-style-type: none">Phân tích dữ liệu nâng cao để xây dựng mô hình và hỗ trợ đưa ra các quyết định kinh doanh	<ul style="list-style-type: none">Tư duy về dữ liệu tốt, khả năng làm việc với nhiều loại dữ liệuCó kỹ năng xây dựng mô hình và tối ưu mô hình	<ul style="list-style-type: none">Excel, Python, SQL, R, SparkCloud (AWS, GCP, Azure, Databricks)
Data Analyst BI Analyst 	<ul style="list-style-type: none">Phân tích dữ liệu để tạo ra các báo cáoCung cấp các insight về dữ liệu để hỗ trợ quyết định kinh doanh	<ul style="list-style-type: none">Khả năng sử dụng đa dạng các công cụ phân tích như Excel, SQL, Power BI và TableauKỹ năng mềm tốtAm hiểu lĩnh vực kinh doanh	<ul style="list-style-type: none">Excel, SQL, Tableau, Power BI, PythonPower Point, Presentation Tools
Data Engineer 	<ul style="list-style-type: none">Xây dựng pipeline dữ liệuXử lý, chuyển đổi dữ liệu (ETL)Quản lý hệ thống dữ liệu	<ul style="list-style-type: none">Kỹ năng lập trình tốt, am hiểu hệ thống cơ sở dữ liệu, có khả năng xử lý dữ liệu lớnKhả năng xây dựng kiến trúc dữ liệu và đảm bảo dữ liệu được ổn định	<ul style="list-style-type: none">Python, SQL, Spark, AirflowCloud (AWS, GCP, Azure, Databricks)

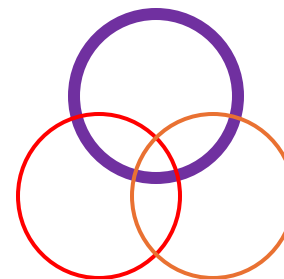
Bộ kỹ năng cần thiết tương ứng với từng vị trí

Công việc trong ngành phân tích dữ liệu thường yêu cầu 3 kỹ năng chính: Am hiểu về lĩnh vực chuyên môn, kỹ thuật xử lý dữ liệu, và kỹ năng về khoa học dữ liệu. Yêu cầu và phân bổ sẽ thay đổi dựa vào vị trí.



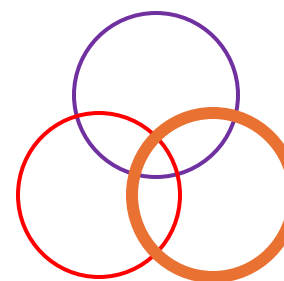
Data Scientist

Trọng tâm: Thiên về kỹ thuật – Xây dựng mô hình hoá – Phân tích thống kê sâu để hiểu dữ liệu - Tối ưu hoá



Data Analyst/Business Intelligence Analyst

Trọng tâm: Hiểu và phân tích bài toán kinh doanh – Truy vấn dữ liệu (SQL) - Trực quan hóa - Báo cáo – Giao tiếp và truyền đạt kết quả với bộ phận kinh doanh



Data Engineer

Trọng tâm: Xây dựng pipeline xử lý dữ liệu – Làm việc với cloud (AWS, GCP, Azure, Databricks) – Tối ưu hoá cơ sở dữ liệu và truy xuất dữ liệu

Đối tượng nên học phân tích dữ liệu

Phân tích dữ liệu là kỹ năng thiết yếu trong thời đại số, không chỉ giới hạn cho chuyên gia hay những người có chuyên môn về dữ liệu



Nhân viên văn phòng

Tự động hóa quy trình, nâng cao giá trị chuyên môn và tạo lợi thế cạnh tranh trong công việc



Chủ doanh nghiệp nhỏ

Tối ưu chi phí, hiểu hành vi khách hàng và xây dựng chiến lược kinh doanh dựa trên dữ liệu thực tế



Sinh viên và học sinh

Phát triển kỹ năng được nhà tuyển dụng săn đón, nâng cao khả năng nghiên cứu và chuẩn bị cho kỷ nguyên số hoá



Tất cả mọi người

Đưa ra quyết định dựa trên dữ liệu thay vì cảm tính, phát triển tư duy phản biện và giải quyết vấn đề hiệu quả

Các trường hợp ứng dụng phân tích dữ liệu

Phân tích dữ liệu được ứng dụng trong nhiều lĩnh vực với mục tiêu cụ thể:

Ngành giải trí

- Netflix phân tích thói quen xem để đề xuất nội dung
- Spotify tạo danh sách phát cá nhân hóa
- Disney tối ưu trải nghiệm công viên từ phản hồi

Thể thao

- NBA sử dụng dữ liệu cho chiến thuật và tuyển dụng
- CLB bóng đá theo dõi hiệu suất qua dữ liệu GPS
- Phân tích thống kê tối ưu chiến thuật thi đấu

Thương mại

- Amazon đề xuất sản phẩm từ lịch sử mua hàng
- Walmart tối ưu hàng tồn kho từ dữ liệu bán hàng
- Starbucks cá nhân hóa chương trình khuyến mãi

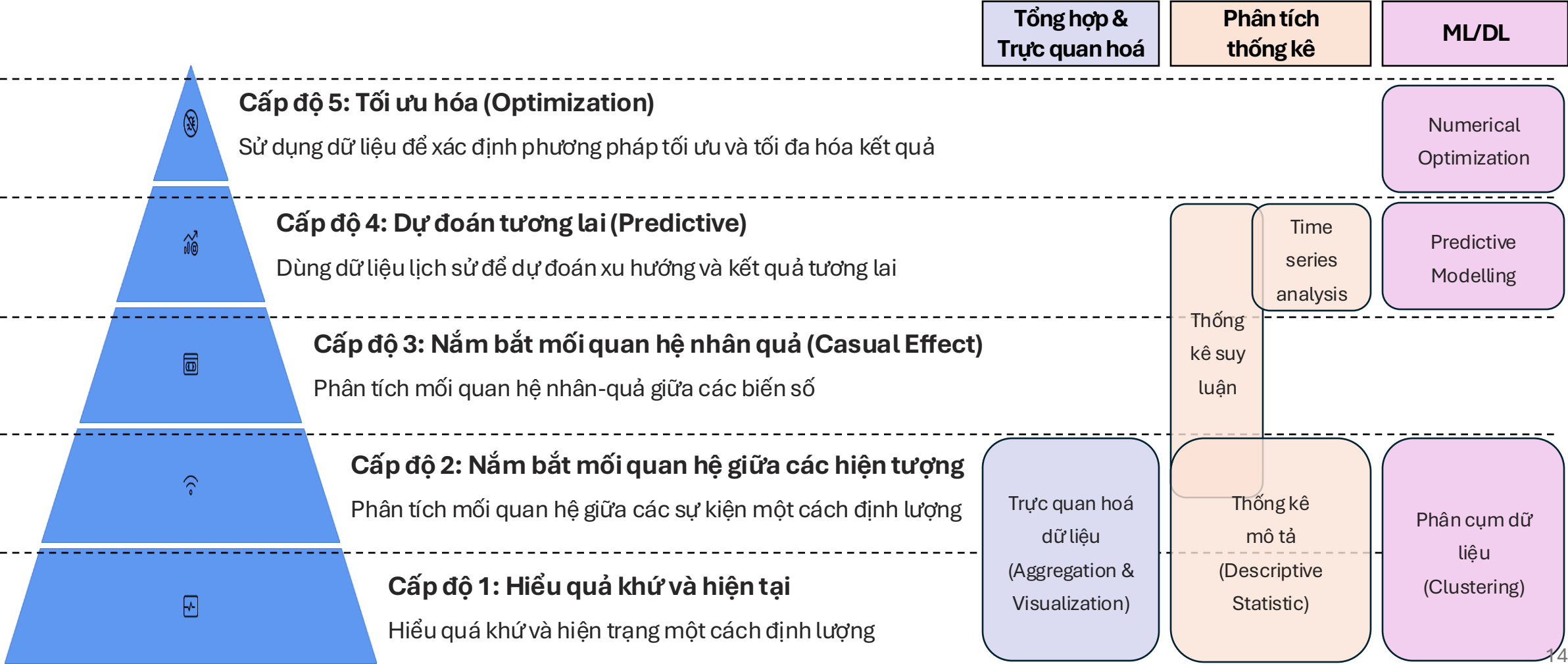
Y tế và giáo dục

- Bệnh viện dự đoán tỷ lệ tái nhập viện
- Trường học theo dõi tiến bộ để can thiệp sớm
- Cơ quan chính phủ phân bổ nguồn lực hiệu quả

Tham Khảo: Coursera, Data Analytics Foundations – Module 1

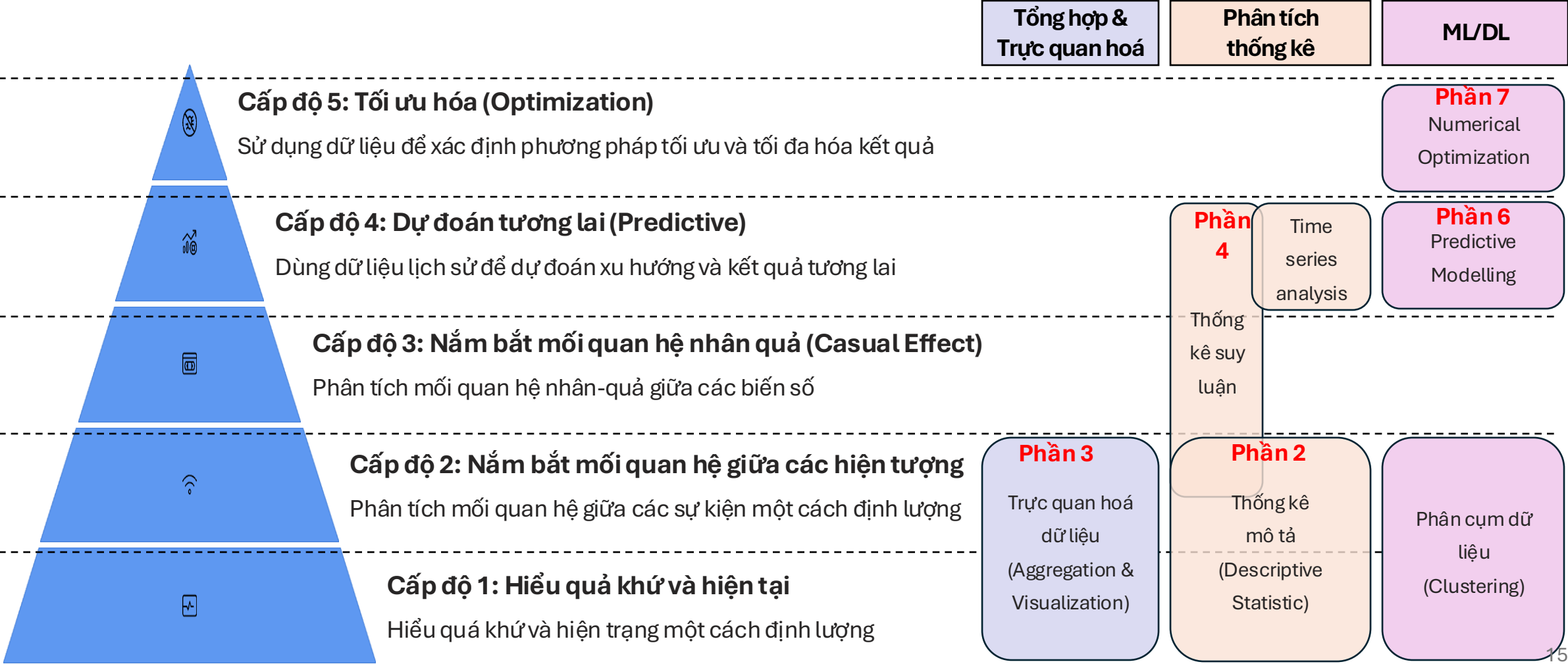
5 Cấp Độ Của Việc Phân Tích Dữ Liệu

Phân tích dữ liệu phát triển qua 5 cấp độ tăng dần về độ phức tạp: từ theo dõi cơ bản, nhận diện mối quan hệ, phân tích nhân quả, đến dự báo tương lai và tối ưu hóa.



Liên hệ với nội dung tài liệu theo từng phần

Phân tích dữ liệu phát triển qua 5 cấp độ tăng dần về độ phức tạp: từ theo dõi cơ bản, nhận diện mối quan hệ, phân tích nhân quả, đến dự báo tương lai và tối ưu hóa.



Lý do sử dụng Excel trong Phân Tích Dữ Liệu

Excel là công cụ cơ bản để phân tích dữ liệu, tạo hình ảnh trực quan và là kỹ năng quan trọng cho nhiều công việc



Công Cụ Tính Toán Phân Tích Thông Dụng

Excel được sử dụng rộng rãi, giúp bạn phân tích dữ liệu nhanh chóng.



Tạo Biểu Đồ Từ Dữ Liệu

Dễ dàng tạo biểu đồ giúp hiểu và trình bày thông tin rõ ràng hơn.



Mô Phỏng Đơn Giản

Giúp bạn hiểu các khái niệm xác suất và thống kê qua thực hành.



Tự Động Hóa Công Việc

Dùng macro để làm nhanh các công việc lặp đi lặp lại, tiết kiệm thời gian.

Excel vs Ngôn Ngữ Lập Trình Cho Phân Tích Dữ Liệu

Trong thực tế, nhiều người phân tích dữ liệu dùng Excel cho phân tích nhanh, đơn giản và ngôn ngữ như Python, R hoặc SQL cho việc phức tạp hơn.

Nên Dùng Excel Khi

- Dữ liệu nhỏ hoặc vừa (dưới 1 triệu dòng)
- Cần phân tích nhanh và tạo biểu đồ đơn giản
- Làm việc với người không chuyên môn về kỹ thuật
- Làm các phép tính thống kê cơ bản
- Cần tạo báo cáo nhanh và dễ chia sẻ
- Không cần tự động hóa phức tạp

Nên Dùng Ngôn Ngữ Lập Trình Khi

- Dữ liệu lớn (hàng triệu dòng trở lên)
- Cần tạo mô hình phức tạp (machine learning, AI)
- Cần tự động hóa công việc thường xuyên
- Cần kết nối với hệ thống và API khác
- Phân tích phức tạp với nhiều nguồn dữ liệu
- Muốn tạo ứng dụng hoặc bảng thông tin tương tác

Chuẩn Bị Công Cụ Excel Cho Phân Tích Dữ Liệu



Cài đặt công cụ phân tích

Cài đặt và kích hoạt "Data Analysis Tool" và "Solver" để mở rộng khả năng phân tích dữ liệu của Excel.



Thiết lập môi trường làm việc

Tùy chỉnh thanh công cụ và trang tính với các công thức thường xuyên sử dụng để tăng hiệu quả làm việc.



Tổ chức dữ liệu

Cấu trúc dữ liệu thành bảng có định dạng rõ ràng với cột tiêu đề và loại dữ liệu phù hợp.



Add-ins bổ sung

Cài đặt thêm các tiện ích mở rộng như Power Query và Power Pivot để xử lý dữ liệu phức tạp.

Thất Bại Khi Dùng Công Cụ Không Phù Hợp

Chọn đúng công cụ phù hợp với quy mô dữ liệu và thời gian. Với dữ liệu nhỏ và thời gian hạn chế, Excel thường là lựa chọn hiệu quả hơn so với viết code Python.



Yêu Cầu Ban Đầu

Sếp yêu cầu phân tích bộ dữ liệu nhỏ (50.000 dòng) về doanh số trong vòng 2 giờ để chuẩn bị cho cuộc họp



Lựa Chọn Sai Công Cụ

Quyết định sử dụng Python thay vì Excel dù dữ liệu đơn giản và thời gian hạn chế



Gặp Lỗi Trong Code

Code xử lý dữ liệu bị lỗi cú pháp và import thư viện, mất thời gian để debug



Kết Quả

Trễ deadline, không có kết quả phân tích để trình bày, trong khi Excel có thể hoàn thành trong 30 phút

Lưu ý khi phân tích dữ liệu

Phân tích dữ liệu cần dựa trên sự thật, phân biệt rõ giữa dữ liệu và diễn giải, bắt đầu bằng câu hỏi đúng, và luôn xác định trước kết quả mong muốn.



Luôn đưa ra sự thật (facts)

Báo cáo kết quả phân tích dữ liệu phải dựa trên dữ liệu thực tế, được kiểm chứng và có thể xác minh.



Phân biệt rõ ràng giữa sự thật và suy luận

Luôn chỉ rõ đâu là dữ liệu thực tế và đâu là kết luận dựa trên sự diễn giải của bạn để người đọc không bị nhầm lẫn.



Bắt đầu bằng câu hỏi "Tại sao"

Trước khi phân tích, hãy đặt câu hỏi "Tại sao" chúng ta cần phân tích dữ liệu này và chúng ta muốn tìm hiểu điều gì.



Phác thảo kết quả mong muốn

Xác định trước các kết quả dự kiến để định hướng quá trình phân tích và đảm bảo đạt được mục tiêu đề ra.

Luôn đưa ra sự thật (Fact)

Phân tích dữ liệu đáng tin cậy phải dựa trên sự thật rõ ràng, chính xác từ nguồn đáng tin cậy. Cần phân biệt rõ giữa dữ liệu thực tế và diễn giải trong báo cáo.

- **Sự thật (Fact):** Thông tin khách quan, có thể kiểm chứng.
- **Suy luận (Inference):** Kết luận dựa trên diễn giải dữ liệu, thường mang tính chủ quan.

Ví dụ trong ngành đầu tư tài chính

- **Sự thật (Fact):** “Giá cổ phiếu Công ty B đã tăng từ 40.000 VNĐ lên 52.000 VNĐ trong vòng 1 tháng, tương đương mức tăng 30%.”
- **Suy luận (Interpretation):** “Giá cổ phiếu tăng mạnh chứng tỏ Công ty B sắp được một quỹ đầu tư lớn rót vốn.”

Hậu quả có thể xảy ra

- **Gây hiểu lầm cho nhà đầu tư:** Nhà đầu tư không phân biệt đâu là dữ liệu, đâu là nhận định, có thể ra quyết định mua cổ phiếu chỉ vì một giả định chưa được xác thực.
- **Dẫn đến thua lỗ tài chính:** Nếu suy luận sai (ví dụ: không có quỹ nào rót vốn cả), cổ phiếu giảm mạnh sau đó → nhà đầu tư lỗ nặng vì quyết định dựa trên "sự thật tưởng tượng".

Bắt Đầu Phân Tích Dữ Liệu Bằng Việc Đặt Câu Hỏi

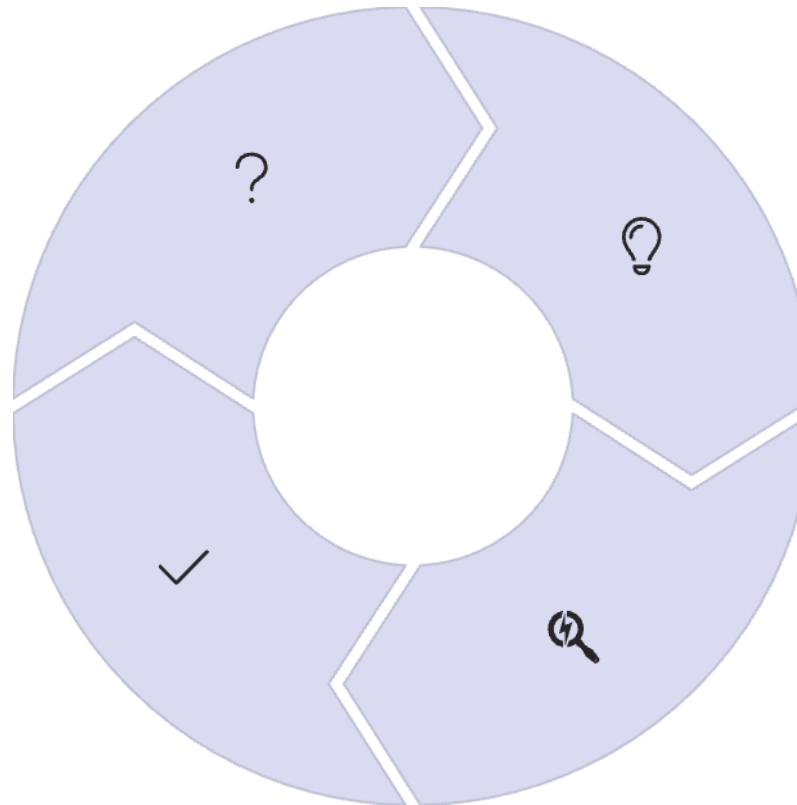
Phân tích dữ liệu hiệu quả bắt đầu từ câu hỏi đúng, giả thuyết rõ ràng và kiểm chứng có phương pháp để ra quyết định chính xác.

Đặt Câu Hỏi

Xác định câu hỏi đúng để định hướng phân tích và xác định thông tin cần thiết.

Kiểm Chứng

Phân tích dữ liệu để kiểm chứng giả thuyết và ra quyết định dựa trên bằng chứng.



Đặt Giả Thuyết

Xây dựng giả thuyết từ quan sát ban đầu để tạo hướng đi cụ thể.

Xác Định Cụ Thể

Thu hẹp phạm vi phân tích để tìm kiếm câu trả lời chính xác.

Bắt Đầu Phân Tích Dữ Liệu Bằng Việc Đặt Câu Hỏi

Nếu không đặt câu hỏi đầu tiên, cửa hàng có thể mất thời gian xem xét quá nhiều dữ liệu không cần thiết. Với câu hỏi rõ ràng, họ tập trung đúng vấn đề và nhanh chóng điều chỉnh giá hoặc chiến lược bán hàng.



1. Đặt Câu Hỏi

Xác định rõ vấn đề cần tìm hiểu: "Tại sao doanh số cửa hàng A giảm 15% trong quý vừa qua?"



2. Đặt Giả Thuyết

Các lý do có thể giải thích vấn đề: "Có thể do đối thủ mới, giá thay đổi, hoặc khách hàng không hài lòng."



3. Xác Định Cụ Thể

Phân tích dữ liệu: "Cần xem dữ liệu bán hàng theo sản phẩm, thời gian, đánh giá khách hàng, và giá đối thủ."



4. Kiểm Chứng

"Phân tích cho thấy doanh số giảm ở một loại sản phẩm, đúng lúc đối thủ giảm giá 20%."

Phác thảo kết quả mong muốn trước khi phân tích

Xác định rõ mục tiêu trước khi phân tích giúp tiết kiệm thời gian, nâng cao hiệu quả và định hướng đúng cho quá trình làm việc. Điều này đảm bảo thu thập đúng dữ liệu và áp dụng phương pháp phù hợp.



Phác hoạ kết quả

Xác định đầu ra cần thiết: biểu đồ, bảng tổng hợp hoặc báo cáo định hướng.



Thu thập dữ liệu

Tập hợp, làm sạch và chuẩn hóa dữ liệu với công cụ phù hợp.



Trực quan và insight

Tạo biểu đồ, bảng tổng hợp để hiển thị kết quả và rút ra thông tin cho quyết định.

Tình Huống Ứng Dụng Excel Trong Doanh Nghiệp



Phân tích doanh số bán hàng

Tổng hợp và phân tích doanh số theo sản phẩm, khu vực, và thời gian để đưa ra chiến lược kinh doanh hiệu quả.

Công cụ sử dụng:

- Sử dụng **SUMIFS** để tính tổng doanh số theo nhiều điều kiện
- Áp dụng **Pivot Table** để phân tích xu hướng theo quý/tháng
- Tạo **Dashboard** với biểu đồ so sánh hiệu suất bán hàng



Quản lý ngân sách và dự báo tài chính

Theo dõi chi phí, dự báo ngân sách, và tính toán chỉ số tài chính để kiểm soát tình hình tài chính doanh nghiệp.

Công cụ sử dụng:

- Dùng công thức **IF** và **VLOOKUP** để tự động phân loại chi phí
- Áp dụng **FORECAST** để dự báo doanh thu trong tương lai
- Sử dụng **Conditional Formatting** để đánh dấu các khoản chi vượt ngân sách

Thiết Kế Phân Tích Dữ Liệu Framework 5W2H

Framework 5W2H Trong Excel: Framework 5W2H giúp cấu trúc quá trình phân tích dữ liệu một cách logic và toàn diện.

Framework 5W2H

- **What (Cái gì):** Xác định vấn đề và mục tiêu cần phân tích
- **Why (Tại sao):** Lý do thực hiện phân tích, giá trị mang lại
- **Who (Ai):** Đối tượng liên quan và người dùng kết quả
- **When (Khi nào):** Khung thời gian phân tích và deadline
- **Where (Ở đâu):** Nguồn dữ liệu và phạm vi phân tích
- **How (Làm thế nào):** Phương pháp và công cụ Excel sử dụng
- **How much (Bao nhiêu):** Chi phí và nguồn lực cần thiết

Ví dụ cụ thể: Phân tích doanh số cửa hàng bán lẻ

- **What:** Phân tích nguyên nhân doanh số giảm 15% trong quý gần nhất
- **Why:** Để xác định chiến lược cải thiện doanh số và điều chỉnh danh mục sản phẩm
- **Who:** Phòng kinh doanh và quản lý cấp cao sẽ sử dụng kết quả
- **When:** Dữ liệu 3 quý gần nhất, cần kết quả trong 1 tuần
- **Where:** Dữ liệu từ hệ thống POS và báo cáo bán hàng (Excel files)
- **How:** Sử dụng Pivot Tables, VLOOKUP, và biểu đồ so sánh trong Excel
- **How much:** Cần 1 chuyên viên phân tích làm việc toàn thời gian trong 1 tuần

Phần 2: Hiểu Xu Hướng Dữ Liệu Qua Thống Kê Cơ Bản

Tại sao phải hiểu xu hướng dữ liệu?

Hiểu xu hướng dữ liệu không chỉ là nền tảng cho quyết định kinh doanh đúng đắn mà còn giúp doanh nghiệp dự đoán thay đổi, phát hiện bất thường và tối ưu hóa hiệu suất hoạt động.

Case study: Hai cửa hàng điện tử A và B cùng có doanh số trung bình 50 triệu/tháng trong Q1/2025

Cửa hàng A

- Doanh số ổn định: 48-52 triệu đồng/tháng (biến thiên $\pm 4\%$)
- Biến động nhỏ theo tuần (độ lệch chuẩn: 1.2 triệu)
- Khách hàng thân thiết: 70% doanh số (trung bình 35 triệu/tháng)
- Lợi nhuận biên: 22% (cao hơn mức trung bình ngành 3%)

⇒ Cần chiến lược chăm sóc khách hàng hiện tại và chương trình loyalty với mục tiêu tăng giá trị đơn hàng trung bình thêm 15%

Cửa hàng B

- Doanh số dao động mạnh: 30-70 triệu đồng/tháng (biến thiên $\pm 40\%$)
- Tăng 60% vào cuối tuần, giảm 35% đầu tuần (mẫu hình tuần rõ rệt)
- Khách hàng mới: 60% doanh số (trung bình 30 triệu/tháng)
- Lợi nhuận biên: 18% (thấp hơn do chi phí marketing cao)

⇒ Cần điều chỉnh nhân sự theo giờ cao điểm, tối ưu chiến dịch marketing theo ngày, và phát triển chiến lược chuyển đổi khách hàng mới thành khách hàng thường xuyên

Không Chỉ Là Giá Trị Trung Bình

Các chỉ số thống kê cơ bản cho ta nhiều góc nhìn khác nhau về dữ liệu. Chỉ dùng mỗi giá trị trung bình khi phân tích có thể gây hiểu nhầm.



Trung bình (Mean)

Là nền tảng quan trọng của phân tích dữ liệu nhưng không nên chỉ dựa vào chỉ số này.

Trong Excel: Sử dụng hàm **AVERAGE(range)** hoặc **=SUM(range)/COUNT(range)** để tính giá trị trung bình.



Trung vị (Median)

Giúp bạn thấy rõ hơn về phân bố dữ liệu, đặc biệt khi có giá trị ngoại lệ.

Trong Excel: Áp dụng hàm **MEDIAN(range)** để xác định giá trị nằm ở vị trí trung tâm của tập dữ liệu.



Phương sai và độ lệch chuẩn

Cho biết mức độ phân tán của dữ liệu, giúp tránh những kết luận sai lầm.

Trong Excel: Dùng **VAR.P(range)** cho phương sai và **STDEV.P(range)** cho độ lệch chuẩn của toàn bộ dữ liệu (hoặc VAR.S/STDEV.S cho mẫu).

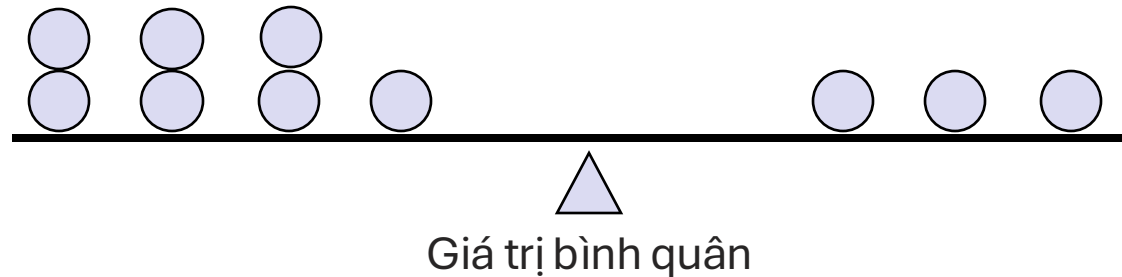
Hình ảnh giá trị trung bình (Mean)

Giá trị trung bình (Mean) là tổng các số chia cho số lượng phần tử. Khi giá trị bình quân được phác họa bằng hình ảnh ta có thể hình dung giá trị trung bình chính là điểm cân bằng.

🕒 Công thức toán học

$$\text{Mean} = \frac{\text{Tổng dữ liệu}}{\text{Số lượng dữ liệu (n)}}$$

🕒 Hình ảnh tưởng tượng



Hạn chế khi chỉ nhìn vào giá trị trung bình

- **Dễ bị mất thông tin quan trọng:** Nếu chỉ nhìn vào giá trị trung bình (ví dụ 50 triệu), ta sẽ không biết được giá trị thấp nhất hay cao nhất là bao nhiêu, hay doanh số có ổn định hay dao động mạnh

Dễ bị hiểu nhầm

- Nó không phải lúc nào cũng là điểm cân bằng (số lượng bên trái và bên phải không nhất thiết phải bằng nhau)
- Dữ liệu không nhất thiết tập trung nhiều quanh giá trị trung bình
- Và giá trị trung bình không luôn đại diện chính xác cho toàn bộ dữ liệu

Độ Phân Tán (Phương Sai, Độ Lệch Chuẩn)

Các chỉ số phân tán giúp hiểu rõ mức độ biến động của dữ liệu - yếu tố quan trọng khi đưa ra quyết định dựa trên dữ liệu

Phương sai (Variance)

- Đo lường mức độ phân tán của dữ liệu xung quanh giá trị trung bình.
- **Trong Excel:** Sử dụng hàm **VAR.S(range)** cho mẫu hoặc **VAR.P(range)** cho toàn bộ dữ liệu.
- **Ví dụ:** Doanh thu hàng ngày (triệu VND): 50, 52, 49, 51, 48
 - Trung bình: 50 triệu
 - Phương sai: 2.5 → Biến động thấp

Độ lệch chuẩn (Standard Deviation)

- Là căn bậc hai của phương sai, có cùng đơn vị với dữ liệu gốc nên dễ hiểu hơn.
- **Trong Excel:** Sử dụng hàm **STDEV.S(range)** hoặc **STDEV.P(range)**
- **Ví dụ:** Với phương sai 2.5, độ lệch chuẩn = $\sqrt{2.5} \approx 1.58$
 - Khoảng 68% dữ liệu nằm trong khoảng 50 ± 1.58 triệu (48.42 - 51.58 triệu)

Case study: So sánh 2 nhóm sản phẩm

- Nhóm A: 100, 105, 95, 102, 98 triệu → Độ lệch chuẩn: 3.8
- Nhóm B: 80, 120, 60, 140, 100 triệu → Độ lệch chuẩn: 31.6
- Dù cùng trung bình 100 triệu, Nhóm B có độ rủi ro cao hơn vì biến động lớn hơn nhiều.

Giá Trị Cực Đại – Cực Tiểu



Hàm MIN và MAX

Dùng MIN, MAX để xác định giá trị nhỏ nhất và lớn nhất trong tập dữ liệu, giúp bạn nhanh chóng nắm bắt phạm vi của dữ liệu.

Ví dụ: Với doanh số 5 cửa hàng: 120, 85, 160, 95, 110 (triệu VND)

=MIN(A1:A5) → 85 triệu |

=MAX(A1:A5) → 160 triệu

→ Nhanh chóng thấy cửa hàng hiệu quả nhất và kém nhất



Hàm LARGE và SMALL

Sử dụng LARGE, SMALL để tìm giá trị lớn thứ k hoặc nhỏ thứ k, hữu ích khi bạn muốn xác định top 5, top 10 giá trị cao nhất hoặc thấp nhất.

Ví dụ: 20 sản phẩm có doanh thu trong tháng:

=LARGE(B1:B20,3) → Top 3 sản phẩm bán chạy nhất

=SMALL(B1:B20,5) → 5 sản phẩm kém nhất cần cải thiện



Thiết lập cảnh báo

Thiết lập cảnh báo tự động khi giá trị vượt ngưỡng, giúp phát hiện sớm các bất thường trong dữ liệu kinh doanh.

Ví dụ: Định dạng có điều kiện → Quy tắc mới → Sử dụng công thức:

=OR(D2<\$G\$1, D2>\$G\$2) → Tự động tô đỏ các ô có giá trị nằm ngoài ngưỡng an toàn

→ Nhanh chóng phát hiện chỉ số bất thường để xử lý kịp thời

Các Hàm Thống Kê Cơ Bản Trong Excel

Chỉ Số Thống Kê	Hàm Excel	Cách sử dụng	Ví dụ cụ thể
Mean (Giá trị trung bình)	AVERAGE()	=AVERAGE(dãy_số)	=AVERAGE(B1:B10) → Tính giá trị trung bình điểm số của 10 học sinh
Median (Trung vị)	MEDIAN()	=MEDIAN(dãy_số)	=MEDIAN(D1:D15) → Xác định mức lương trung vị của 15 nhân viên
Mode	MODE()	=MODE(dãy_số)	=MODE(E1:E50) → Tìm kích cỡ giày phổ biến nhất trong 50 khách hàng
Giá trị lớn nhất - Giá trị nhỏ nhất	MAX()/MIN()	=MAX(dãy_số) hoặc =MIN(dãy_số)	=MAX(J1:J40) → Tìm doanh thu cao nhất trong 40 cửa hàng
Tổng	SUM()	=SUM(dãy_số)	=SUM(K1:K12) → Tính tổng chi phí hoạt động trong 12 tháng
Số lượng dữ liệu	COUNT()/ COUNTA()	=COUNT(dãy_số) hoặc =COUNTA(dãy_ô)	=COUNT(L1:L200) → Đếm số lượng giao dịch đã ghi nhận trong bảng dữ liệu
Variance (Phương sai)	VAR()	=VAR(dãy_số)	=VAR(G1:G25) → Tính độ phân tán của lợi nhuận 25 sản phẩm
Standard Deviation (Độ lệch chuẩn)	STDEV()	=STDEV(dãy_số)	=STDEV(F1:F30) → Đo mức độ biến động của doanh số trong 30 ngày
Standard Error (Sai số chuẩn)	STDEV()/SQRT(COUNT())	=STDEV(dãy_số)/SQRT(COUNT(dãy_số))	=STDEV(C1:C20)/SQRT(COUNT(C1:C20)) → Tính sai số chuẩn cho 20 mẫu đo lường
Kurtosis (Độ nhọn)	KURT()	=KURT(dãy_số)	=KURT(H1:H100) → Phân tích mức độ tập trung của giá trị trong 100 mẫu
Skewness (Độ lệch)	SKEW()	=SKEW(dãy_số)	=SKEW(I1:I80) → Kiểm tra tính đối xứng của phân phối thu nhập của 80 hộ gia đình

<Tham Khảo>

Các hàm Excel quan trọng cho phân tích dữ liệu

Phương thức	Hàm Excel	Cách sử dụng	Ví dụ
Kiểm tra điều kiện	IF()	=IF(điều_kiện, giá_trị_nếu_đúng, giá_trị_nếu_sai)	=IF(A1>100,"Cao","Thấp") → Nếu A1 lớn hơn 100 trả về "Cao", ngược lại trả về "Thấp"
Lồng nhiều điều kiện	IFS()	=IFS(điều_kiện1, giá_trị1, điều_kiện2, giá_trị2...)	=IFS(A1<50,"Thấp",A1<100,"Trung bình",TRUE,"Cao") → Phân loại giá trị dựa trên các mức
Tổng có điều kiện	SUMIF()	=SUMIF(phạm_vi, tiêu_chí, phạm_vi_tổng)	=SUMIF(B1:B10,"Hà Nội",C1:C10) → Tổng doanh số của các cửa hàng ở Hà Nội
Tổng nhiều điều kiện	SUMIFS()	=SUMIFS(phạm_vi_tổng, phạm_vi1, tiêu_chí1, phạm_vi2, tiêu_chí2...)	=SUMIFS(D1:D20,B1:B20,"Hà Nội",C1:C20,"Q1") → Tổng doanh số ở Hà Nội trong quý 1
Đếm có điều kiện	COUNTIF()	=COUNTIF(phạm_vi, tiêu_chí)	=COUNTIF(B1:B50,">100") → Đếm số sản phẩm có doanh số lớn hơn 100
Đếm nhiều điều kiện	COUNTIFS()	=COUNTIFS(phạm_vi1, tiêu_chí1, phạm_vi2, tiêu_chí2...)	=COUNTIFS(B1:B20,"Nam",C1:C20,">30") → Đếm số khách hàng nam trên 30 tuổi
Trung bình có điều kiện	AVERAGEIF()	=AVERAGEIF(phạm_vi, tiêu_chí, phạm_vi_trung_bình)	=AVERAGEIF(B1:B10,"Laptop",C1:C10) → Tính giá trung bình các mặt hàng laptop
Trung bình nhiều điều kiện	AVERAGEIFS()	=AVERAGEIFS(phạm_vi_trung_bình, phạm_vi1, tiêu_chí1...)	=AVERAGEIFS(D1:D10,B1:B10,"Laptop",C1:C10,">5 000000") → Giá trung bình laptop trên 5 triệu
Tìm kiếm theo hàng	HLOOKUP()	=HLOOKUP(giá_trị_tìm, bảng_tìm, chỉ_số_hàng, [chính_xác])	=HLOOKUP("Q1",A1:E5,3,FALSE) → Tìm giá trị ở hàng 3 dưới cột "Q1"
Tìm kiếm theo cột	VLOOKUP()	=VLOOKUP(giá_trị_tìm, bảng_tìm, chỉ_số_cột, [chính_xác])	=VLOOKUP("SP001",A1:F20,3,FALSE) → Tìm giá trị ở cột 3 của sản phẩm "SP001"
Truy xuất dữ liệu theo vị trí	INDEX()	=INDEX(mảng, số_hàng, [số_cột])	=INDEX(A1:D10,3,2) → Trả về giá trị ở hàng 3, cột 2 trong phạm vi A1:D10
Tìm vị trí của dữ liệu	MATCH()	=MATCH(giá_trị_tìm, phạm_vi_tìm, [kiểu_đối_chiếu])	=MATCH("SP005",A1:A20,0) → Trả về vị trí hàng của "SP005" trong phạm vi A1:A20

Sử dụng "Data Analysis Tool"

Với Data Analysis Tool > Descriptive Statistics, bạn nhận được đầy đủ các chỉ số như trung bình, mode, độ lệch chuẩn một cách nhanh chóng mà không cần nhập nhiều công thức.



Kích hoạt Data Analysis Tool

Bật công cụ này trong Excel



Chọn Descriptive Statistics

Dùng tính năng này để có kết quả nhanh



Diễn giải kết quả

Hiểu ý nghĩa các chỉ số thu được

Dùng Pivot Table

Pivot Table là công cụ mạnh mẽ giúp tổng hợp và phân tích lượng lớn dữ liệu một cách trực quan thông qua việc tái cấu trúc dữ liệu từ dạng bảng thành báo cáo có ý nghĩa.

1 Tạo Pivot Table cơ bản

Chọn dữ liệu → Insert → PivotTable
→ Kéo thả các trường vào 4 vùng:
Filters, Columns, Rows và Values

2 Lọc và nhóm dữ liệu

Sử dụng Slicers để lọc → Nhóm
theo thời gian → Tạo Calculated
Fields → Hiển thị dữ liệu dưới dạng
% với Show Values As

3 Tạo báo cáo trực quan

Chuyển sang PivotChart → Kết hợp
nhiều Pivot Table trong
Dashboard → Tự động cập nhật khi
nguồn thay đổi → Tạo từ nhiều
nguồn với Data Model

Nguyên Lý Hoạt Động Của Pivot Table

Pivot Table sắp xếp lại dữ liệu bằng cách gom nhóm và tính toán dựa trên các thành phần được đặt vào 4 khu vực chính:



Filters (Bộ lọc)

Chọn xem dữ liệu nào được hiển thị. Ví dụ: Chỉ xem số liệu của "Quý 1" hoặc "Khu vực miền Nam".



Columns (Cột)

Tạo các cột trong báo cáo. Ví dụ: Đặt "Tháng" vào đây sẽ tạo một cột cho mỗi tháng, giúp xem số liệu theo thời gian.




Rows (Hàng)

Tạo các hàng trong báo cáo. Ví dụ: Đặt "Sản phẩm" vào đây sẽ hiển thị mỗi sản phẩm trên một hàng, giúp so sánh giữa các sản phẩm.



Values (Giá trị)

Tính toán kết quả (tổng, trung bình, đếm...). Ví dụ: Kéo "Doanh thu" vào đây và chọn SUM sẽ hiển thị tổng doanh thu.

 Ưu điểm của Pivot Table là khả năng tự động tính toán lại khi bạn kéo thả các mục khác nhau vào các khu vực này, giúp xem dữ liệu từ nhiều góc độ khác nhau.

Các Phím Tắt hữu ích trong Excel

Phím Tắt Windows	Phím Tắt Mac	Chức Năng
Ctrl + C	Command (⌘) + C	Sao chép dữ liệu
Ctrl + V	Command (⌘) + V	Dán dữ liệu
Ctrl + Z	Command (⌘) + Z	Hoàn tác thao tác
Ctrl + E	Command (⌘) + E	Flash Fill (Tự động điền dữ liệu theo mẫu)
Ctrl + Y	Command (⌘) + Y	Làm lại thao tác
Ctrl + S	Command (⌘) + S	Lưu tập tin
F4	Command (⌘) + T	Lặp lại thao tác cuối cùng / Cố định tham chiếu trong công thức
F2	Control + U	Chỉnh sửa ô
Ctrl + Home	Command (⌘) + Home	Di chuyển đến ô A1
Ctrl + End	Command (⌘) + End	Di chuyển đến ô cuối có dữ liệu
Ctrl + →	Command (⌘) + →	Di chuyển đến cột cuối có dữ liệu
Ctrl + ↓	Command (⌘) + ↓	Di chuyển đến hàng cuối có dữ liệu
Alt + =	Command (⌘) + Shift + T	Tự động tính tổng
Ctrl + Shift + L	Command (⌘) + Shift + F	Bật/tắt bộ lọc
Ctrl + Space	Control + Space	Chọn toàn bộ cột
Shift + Space	Shift + Space	Chọn toàn bộ hàng
Ctrl + F / Ctrl + H	Command (⌘) + F / Command (⌘) + Shift + H	Tìm kiếm / Tìm và thay thế
Alt → H → O → I/or A	Home → Format → AutoFit Column Width/Row Height	Tự động điều chỉnh chiều rộng cột, hàng
Alt → I → R or C	Command (⌘) + Shift + +	Chèn hàng (R) hoặc cột (C)
Alt → H → H	unknown	Định dạng ô
Alt → H → F → C	unknown	Định dạng có điều kiện

Doanh Thu Iphone 15 Pro Max

Sử dụng các Hàm cơ bản trong Excel và công cụ Data Analysis để tính toán các thông số thống kê cơ bản miêu tả doanh thu của Iphone 15 Pro Max

Dữ liệu gốc		
Country	Product	Gross Sales
USA	iPhone 15 Pro Max	1082697
USA	iPhone 15 Pro Max	569525
USA	iPhone 15 Pro Max	767360
USA	iPhone 15 Pro Max	869275
USA	iPhone 15 Pro Max	199034
USA	iPhone 15 Pro Max	691823
USA	iPhone 15 Pro Max	272173
USA	iPhone 15 Pro Max	880066
USA	iPhone 15 Pro Max	131890
USA	iPhone 15 Pro Max	961598
USA	iPhone 15 Pro Max	737385
USA	iPhone 15 Pro Max	612689
USA	iPhone 15 Pro Max	605495
USA	iPhone 15 Pro Max	799733
USA	iPhone 15 Pro Max	1035936
USA	iPhone 15 Pro Max	326128
USA	iPhone 15 Pro Max	189442
USA	iPhone 15 Pro Max	535953
USA	iPhone 15 Pro Max	1152239
USA	iPhone 15 Pro Max	161865
USA	iPhone 15 Pro Max	1091090
USA	iPhone 15 Pro Max	152273
USA	iPhone 15 Pro Max	550341
USA	iPhone 15 Pro Max	694221
USA	iPhone 15 Pro Max	960399
USA	iPhone 15 Pro Max	364496
USA	iPhone 15 Pro Max	188243
USA	iPhone 15 Pro Max	140283
USA	iPhone 15 Pro Max	381282
USA	iPhone 15 Pro Max	972389
USA	iPhone 15 Pro Max	350108

Sử dụng hàm cơ bản trong Excel			
Gross Sales of iPhone 15 Pro Max			
Mean	\$	563,718.61	AVERAGE(range)
Standard Error	\$	35,079.85	STDEV.S(range)/SQRT(COUNT(range))
Median	\$	605,495.00	MEDIAN(range)
Mode	\$	605,495.00	MODE.SNGL(range)
Standard Deviation	\$	330,942.61	STDEV.S(range)
Sample Variance		109,523,008,099	VAR.S(range)
Kurtosis		-1	KURT(range)
Skewness		0	SKEW(range)
Range	\$	1,163,030.00	MAX(range) - MIN(range)
Minimum	\$	11,990.00	MIN(range)
Maximum	\$	1,175,020.00	MAX(range)
Sum	\$	50,170,956.00	SUM(range)
Count		89	COUNT(range)
Largest(1)	\$	1,175,020.00	LARGE(range, 1)
Smallest(1)	\$	11,990.00	SMALL(range, 1)
Confidence Level(95.0%)		68,755	CONFIDENCE.T(0.05, STDEV.S(range), COUNT(range))

Chi tiết: part2_descriptive statistic.xlsx

Sử dụng công cụ Data Analysis			
Gross Sales of iPhone 15 Pro Max			
Mean	\$	563,718.61	
Standard Error	\$	35,079.85	
Median	\$	605,495.00	
Mode	\$	605,495.00	
Standard Deviation	\$	330,942.61	
Sample Variance		109,523,008,099	
Kurtosis		-1	
Skewness		0	
Range	\$	1,163,030.00	
Minimum	\$	11,990.00	
Maximum	\$	1,175,020.00	
Sum	\$	50,170,956.00	
Count		89	
Largest(1)	\$	1,175,020.00	
Smallest(1)	\$	11,990.00	
Confidence Level(95.0%)		69,714	

Phần 3: Trực Quan Hóa Dữ Liệu

Vì Sao Cần Trực Quan Dữ Liệu?

Trực quan hóa giúp nhận diện xu hướng nhanh, truyền đạt thông tin hiệu quả và hỗ trợ ra quyết định dựa trên bằng chứng trực quan.

1 Nhận biết xu hướng nhanh chóng

Biểu đồ Excel chuyển số liệu thành hình ảnh, giúp phát hiện ngay xu hướng doanh số và mối tương quan mà bảng số không thể hiện rõ.

2 Truyền đạt thông điệp hiệu quả

Dashboard với biểu đồ trực quan giúp trình bày kết quả phân tích trong 2-3 phút thay vì 15 phút giải thích bảng số.

3 Hỗ trợ ra quyết định

Biểu đồ phân tán và nhiệt thể hiện tương quan dữ liệu, giúp đưa ra quyết định chính xác về phân bổ ngân sách và nguồn lực.

Doanh thu mặt hàng quần áo

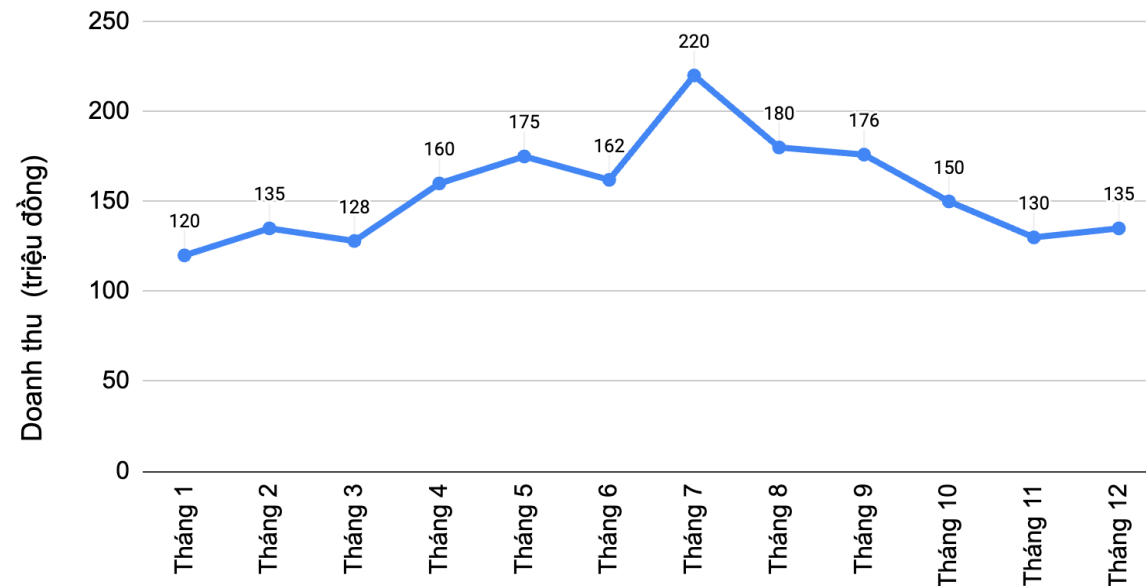
Trực quan hóa dữ liệu giúp chuyển đổi bảng số phức tạp thành biểu đồ dễ hiểu, cho phép nhận biết nhanh xu hướng và hỗ trợ ra quyết định hiệu quả.

✗ Khó nắm bắt xu hướng

Tháng	Doanh thu (triệu đồng)
Tháng 1	120
Tháng 2	135
Tháng 3	128
Tháng 4	160
Tháng 5	175
Tháng 6	162
Tháng 7	220
Tháng 8	180
Tháng 9	176
Tháng 10	150
Tháng 11	130
Tháng 12	135

○ Dễ dàng nắm bắt xu hướng thay đổi doanh số theo mùa

Doanh thu có xu hướng tăng vào mùa hè và bắt đầu giảm khi chuyển sang thu và đông



Các Loại Biểu Đồ Thường Được Sử Dụng



Biểu đồ cột (Bar Chart)

So sánh giá trị giữa các nhóm



Biểu đồ đường (Line Chart)

Hiển thị xu hướng theo thời gian



Biểu đồ phân tán (Scatter Plot)

Thể hiện mối liên hệ giữa hai biến số



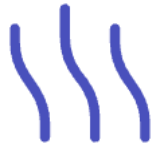
Biểu đồ tròn (Pie Chart)

Thể hiện tỷ lệ các phần trong tổng thể



Biểu đồ hộp (Box Plot)

Cho thấy phân phối dữ liệu và phát hiện giá trị bất thường



Heatmap

Biểu diễn dữ liệu bằng màu sắc, thể hiện mức độ quan hệ



Biểu đồ Phân Phối (Histogram)

Thể hiện tần suất xuất hiện của dữ liệu

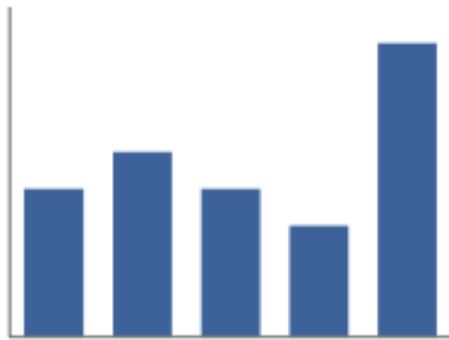


Biểu đồ Water Flow

Thể hiện sự thay đổi giá trị qua các giai đoạn

Biểu Đồ Cột (Bar Plot)

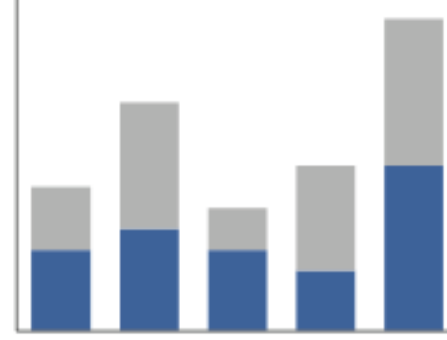
Có 4 loại biểu đồ cột được sử dụng thông dụng: vertical bar, horizontal bar, stacked vertical bar, stacked horizontal bar. Biểu đồ cột được sử dụng để so sánh giá trị giữa các danh mục, với chiều cao cột tương ứng với giá trị.



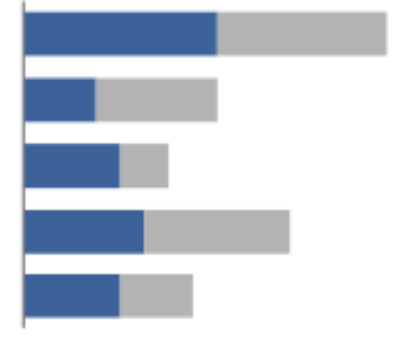
Vertical bar



Horizontal bar



Stacked vertical bar



Stacked horizontal bar

Image source: *Storytelling with Data: A Data Visualization Guide for Business Professionals*, Cole Nussbaumer Knaflic



Dễ đọc

Trực quan cho mọi đối tượng



So sánh hiệu quả

Phân biệt rõ giữa các nhóm



Linh hoạt

Có thể so sánh nhiều biến cùng lúc

4 Tips Để Tạo Biểu Đồ Cột Chuyên Nghiệp

❌ Bad Example

Survey results: summer learning program on science

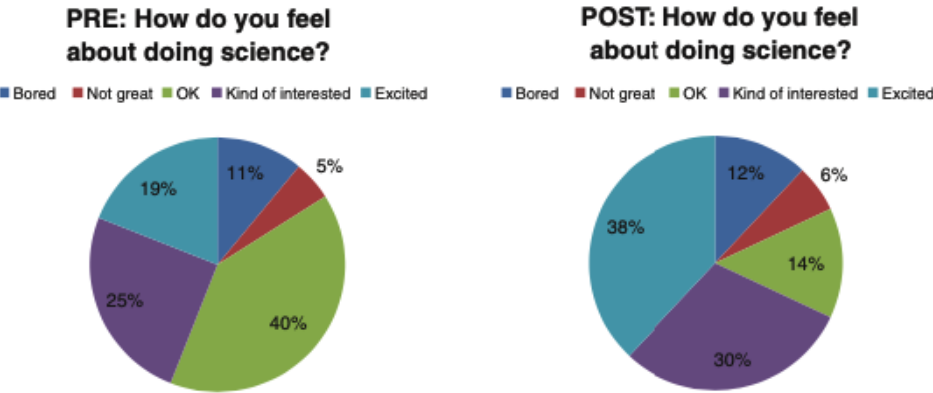
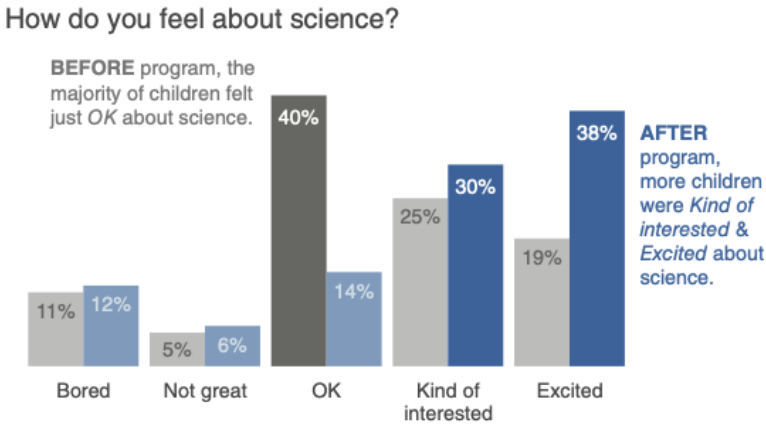


FIGURE 9.28 Original visual

🟡 Good Example

Pilot program was a success



Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

FIGURE 9.30 Simple bar graph

Image source: *Storytelling with Data: A Data Visualization Guide for Business Professionals*, Cole Nussbaumer Knaflic

1. Chọn đúng loại biểu đồ

- Chọn đúng biểu đồ giúp truyền đạt đúng thông tin, tránh rắc rối, gây hiểu nhầm cho người xem

2. Đơn Giản Hóa Nội Dung

- Hạn chế nhãn và thông tin trên biểu đồ.
- Sắp xếp dữ liệu theo thứ tự logic để người xem dễ hiểu thông tin.

3. Gắn Nhãn Rõ Ràng

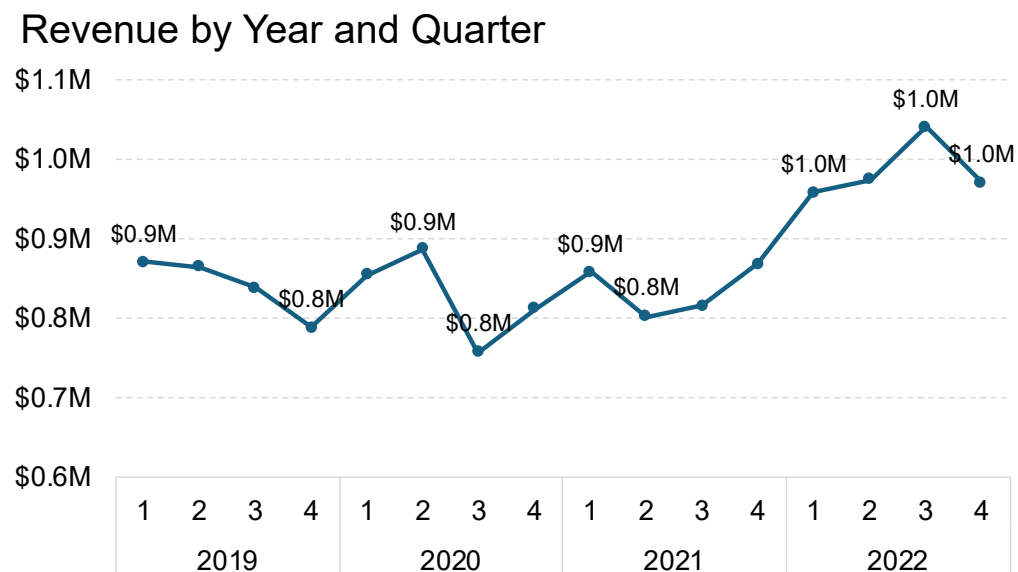
- Đặt tiêu đề ngắn gọn truyền đạt thông điệp chính
- Thêm chú thích khi cần để người đọc hiểu đúng thông điệp.

4. Sử Dụng Màu Sắc Phù Hợp

- Chọn màu phù hợp với thương hiệu công ty.
- Dùng màu tương phản để nhấn mạnh thông tin quan trọng.
- Giữ nhất quán màu sắc trong toàn bộ báo cáo.

Biểu Đồ Đường (Line Chart)

Biểu đồ đường giúp xem sự thay đổi của dữ liệu theo thời gian một cách rõ ràng.



Xem xu hướng qua thời gian

Thích hợp để hiển thị dữ liệu theo thời gian như doanh số theo tháng



So sánh nhiều loại dữ liệu

Dễ xem xu hướng của nhiều đối tượng trên cùng một biểu đồ

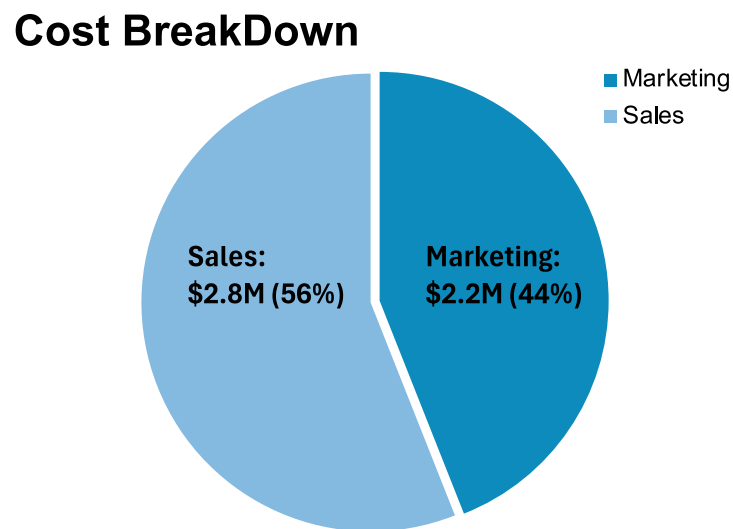


Các kiểu biểu đồ đường trong Excel

Có nhiều loại: biểu đồ đường đơn giản, biểu đồ có điểm đánh dấu, biểu đồ 2D/3D

Biểu Đồ Tròn (Pie Chart)

Biểu đồ tròn thể hiện tỷ lệ các phần trong một tổng thể. Hiệu quả khi so sánh doanh thu, thị phần hoặc phân bổ ngân sách.



Hiển thị tỷ lệ phần trăm

Giúp người xem nhanh chóng nắm bắt tỷ lệ của từng phần



Tối ưu với dưới 7 phần

Hiệu quả nhất khi hiển thị không quá 7 phân khúc

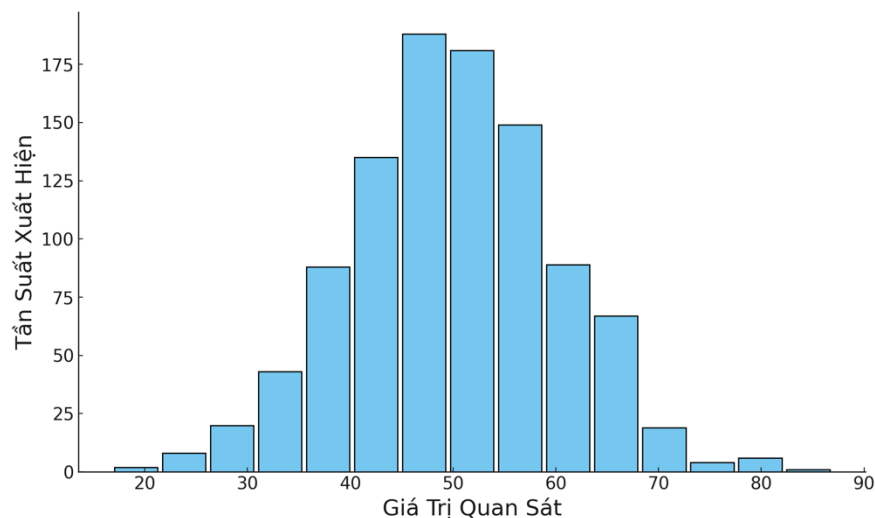


Các biến thể

Gồm biểu đồ tròn tiêu chuẩn, biểu đồ phân tách và biểu đồ hình khuyên

Biểu đồ Phân Phối (Histogram)

Biểu đồ histogram hiển thị phân phối tần suất của dữ liệu số bằng cách chia thành các khoảng và đếm số lượng điểm dữ liệu trong mỗi khoảng.



Đánh giá hình dạng phân phối

Nhận biết dữ liệu đối xứng, lệch phải hoặc lệch trái



Phát hiện giá trị ngoại lai

Xác định outliers không phù hợp với xu hướng chung



Kiểm tra tính chuẩn

Đánh giá sự phù hợp với phân phối chuẩn



Xác định mô hình thống kê

Chọn mô hình phù hợp dựa trên dạng phân phối

Heatmap

Heatmap là công cụ trực quan hóa dữ liệu bằng màu sắc, giúp nhận biết nhanh các mẫu và xu hướng thông qua thang màu đậm nhạt.

Tháng	Doanh thu	Lợi nhuận
Tháng 1	120	-12
Tháng 2	135	-7
Tháng 3	128	13
Tháng 4	160	16
Tháng 5	175	18
Tháng 6	162	16
Tháng 7	220	33
Tháng 8	180	18
Tháng 9	176	18
Tháng 10	150	15
Tháng 11	130	13
Tháng 12	135	14

※ Đơn vị: Triệu đồng

Tạo heatmap bằng Conditional Formatting

- Sử dụng Color Scales trong Conditional Formatting để tạo heatmap, giúp nhanh chóng nhận biết xu hướng dữ liệu gradient màu sắc.

Heatmap giúp phân tích hiệu suất nhanh chóng

- So sánh hiệu suất giữa khu vực, sản phẩm hoặc thời gian thông qua màu sắc đậm nhạt, nhanh chóng xác định điểm mạnh và yếu.

Biểu Đồ Water Flow

Biểu đồ Water Flow (hay còn gọi là biểu đồ thác nước) là dạng biểu đồ trực quan hóa thể hiện sự thay đổi giá trị lũy kế theo từng giai đoạn, giúp người xem hiểu được các yếu tố đóng góp tích cực hoặc tiêu cực vào kết quả cuối cùng.

2014 Headcount math

Though more employees transferred out of the team than transferred in, aggressive hiring means overall headcount (HC) increased 16% over the course of the year.

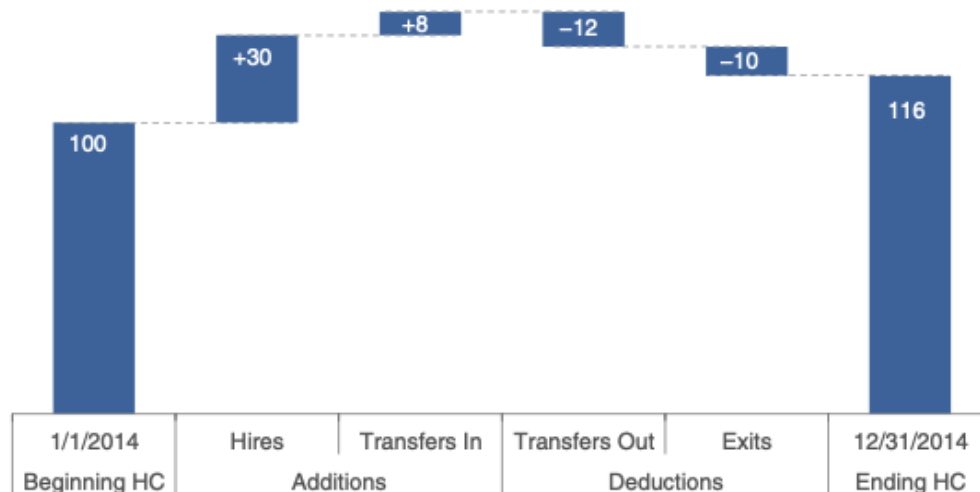


FIGURE 2.17 Waterfall chart

Image source: *Storytelling with Data: A Data Visualization Guide for Business Professionals*, Cole Nussbaumer Knafl

Đặc điểm chính

- Hiện thị sự biến động giữa hai điểm dữ liệu - điểm bắt đầu và kết thúc
- Các thanh tăng thường hiển thị màu xanh lá, các thanh giảm thường màu đỏ
- Thanh cuối cùng hiển thị giá trị tổng sau tất cả biến động

Khi nào nên sử dụng

- Phân tích tài chính: theo dõi sự thay đổi doanh thu, chi phí, lợi nhuận
- Phân tích KPI: so sánh chỉ số hiệu suất giữa các giai đoạn
- Phân tích nguyên nhân: xác định các yếu tố tác động tích cực/tiêu cực
- Nghiên cứu thị trường: theo dõi biến động thị phần, khách hàng

Tạo Biểu Đồ Water Flow Trong Excel

Biểu đồ Water Flow (hay biểu đồ thác nước) giúp trực quan hóa sự thay đổi giá trị lũy kế qua các giai đoạn, thể hiện các yếu tố tăng/giảm ảnh hưởng đến kết quả cuối cùng.

1

Chuẩn bị dữ liệu

Tổ chức dữ liệu thành các cột: giá trị ban đầu, các thay đổi (tăng/giảm), và giá trị cuối. Mỗi thay đổi được đặt trong một cột riêng biệt.

2

Tạo biểu đồ cột chồng

Chọn dữ liệu → Insert → Column Chart → Stacked Column. Đây là nền tảng cho biểu đồ Water Flow.

3

Chỉnh sửa hiển thị

Chuyển các giá trị âm thành màu đỏ, giá trị dương thành màu xanh lá. Thêm đường kết nối giữa các cột để tạo hiệu ứng "dòng chảy".

4

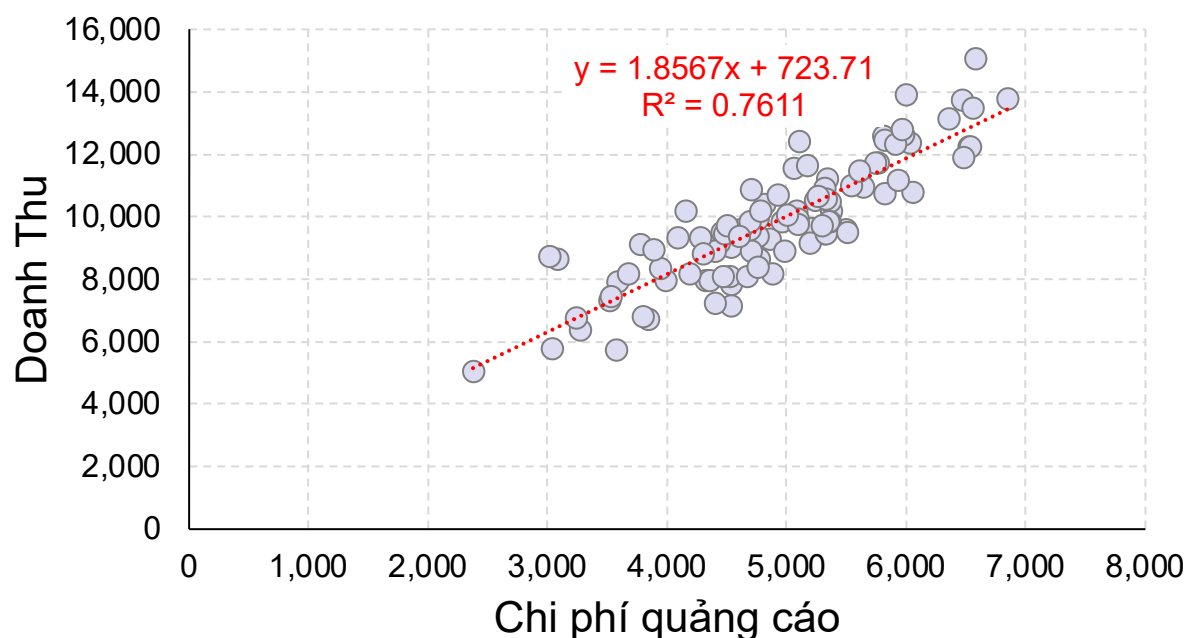
Tinh chỉnh định dạng

Thêm nhãn dữ liệu, điều chỉnh màu sắc và căn chỉnh để biểu đồ dễ đọc. Giá trị cuối cùng thường được đánh dấu bằng màu khác biệt.

Scatter plot

Biểu đồ phân tán hiển thị mối quan hệ giữa hai biến số, giúp phát hiện mẫu và xu hướng trong dữ liệu.

Biểu đồ tương quan giữa doanh thu và chi phí quảng cáo



Tạo scatter plot

- Mỗi điểm trên biểu đồ đại diện cho một cặp giá trị (x,y), giúp nhận diện mẫu và xu hướng.
- Trong Excel: chọn dữ liệu, sử dụng tùy chọn Scatter, thêm nhãn cho các điểm quan trọng.

Thêm trendline và phân tích

- Đường xu hướng làm rõ mối quan hệ giữa các biến. Excel hỗ trợ nhiều dạng: tuyến tính, đa thức, logarithm.
- Hiển thị phương trình và R^2 để đo lường độ mạnh của mối quan hệ và dự đoán giá trị.

3 Tips Để Tạo Scatter Plot Đẹp

1

Cài đặt độ trong suốt khi quá nhiều điểm dữ liệu

- Giảm độ đậm của điểm dữ liệu (opacity) khi visualize dataset lớn giúp nhìn rõ các khu vực tập trung cao và tránh hiện tượng chồng chéo.
- Trong Excel, điều chỉnh này được thực hiện qua Format Data Series.

2

Đặt giới hạn trên và dưới cho dữ liệu

- Thiết lập giới hạn trục x và y phù hợp giúp loại bỏ các giá trị ngoại lệ và tập trung vào phạm vi dữ liệu quan trọng, làm nổi bật xu hướng chính và tăng độ chính xác trong phân tích.

3

Hiển thị đường fitting line

- Thêm đường xu hướng (trendline) để làm rõ mối quan hệ giữa các biến.
- Hiển thị thêm phương trình và giá trị R^2 để đánh giá độ mạnh của mối quan hệ và khả năng dự đoán của mô hình.

Hệ Số Tương Quan

Hệ số tương quan đo lường mối quan hệ tuyến tính giữa hai biến, dao động từ -1 đến +1. Giá trị dương thể hiện quan hệ thuận, âm thể hiện quan hệ nghịch, và gần 0 cho thấy không có tương quan.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Công thức tính hệ số Pearson, với x_i, y_i là giá trị của hai biến, \bar{x}, \bar{y} là giá trị trung bình

+1

Tương quan dương hoàn hảo

Hai biến cùng tăng hoặc cùng giảm

0

Không có tương quan

Không có liên hệ tuyến tính

-1

Tương quan âm hoàn hảo

Một biến tăng, biến kia giảm

Dùng hàm CORREL hoặc PEARSON trong Excel để đo lường mối liên hệ tuyến tính giữa hai biến

Ma Trận Tương Quan

Ma trận tương quan phân tích mối quan hệ giữa nhiều biến cùng lúc, phát hiện đa cộng tuyến, và xác định các nhóm biến liên quan. Đây là công cụ thiết yếu trong phân tích dữ liệu đa biến.

Tạo ma trận tương quan

- Ma trận tương quan phân tích mối quan hệ giữa nhiều biến đồng thời. Trong Excel, bạn có thể tạo ma trận này bằng công thức CORREL hoặc VBA để tự động hóa quá trình.
- Ma trận tương quan hoàn chỉnh là ma trận đối xứng, với giá trị 1 trên đường chéo chính.

Phân tích ma trận tương quan

- Ma trận tương quan giúp phát hiện đa cộng tuyến và xác định nhóm biến có liên quan chặt chẽ.
- Việc phát hiện đa cộng tuyến rất quan trọng trong phân tích hồi quy, ảnh hưởng đến độ chính xác của mô hình. Nếu hai biến độc lập có tương quan cao (>0.7), nên cân nhắc loại bỏ một trong hai.

Sử Dụng Excel Template

Tạo Excel Template giúp tiết kiệm thời gian và đảm bảo nhất quán trong báo cáo doanh nghiệp



Bắt đầu từ template phù hợp

Sử dụng mẫu Dashboard, KPI Tracker hoặc Sales Report. Tổ chức dữ liệu thành Tables và dùng Named Ranges để kết nối dữ liệu với biểu đồ



Giữ dữ liệu – biểu đồ tách biệt

Tách dữ liệu thô, phân tích và biểu đồ vào các sheet riêng. Sử dụng INDIRECT hoặc OFFSET tạo liên kết động để tự động cập nhật biểu đồ



Tùy chỉnh màu sắc và phong cách

Tạo bộ màu nhất quán 4-5 màu. Dùng định dạng có điều kiện để tự thay đổi màu theo giá trị. Loại bỏ đường lưới, sử dụng font Sans-serif để tăng khả năng đọc



Lưu template biểu đồ tùy chỉnh

Chuột phải vào biểu đồ → "Lưu làm template" → lưu trong thư mục Chart Templates.

Tạo dashboard phân tích doanh số

Mục tiêu: Tạo một Dashboard trực quan giúp theo dõi doanh thu, chi phí, lợi nhuận, và các nguồn bán hàng chính.

Các bước tạo Dashboard trong Excel:

Bước 1 – Chuẩn bị dữ liệu (sheet: sales_data)

Bước 2 – Tạo Pivot Table & Pivot Chart (sheet: profit_by_year_month, revenue_trend, sales_by_source, cost_break_down)

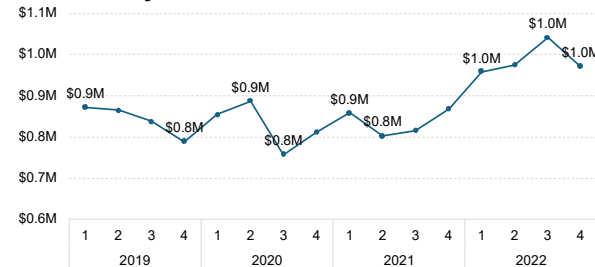
Bước 3 – Thiết kế Dashboard (sheet: dashboard)

Bước 4 – Sắp xếp biểu đồ và hoàn thiện Dashboard

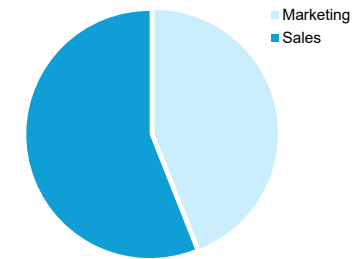
Chi tiết: part3_sales_analysis_dashboad.xlsx

Sales Analysis Dashboard			
Number of Sales	Profit	Net Revenue	Cost
1,613	\$8.3M	\$14.0M	\$5.0M

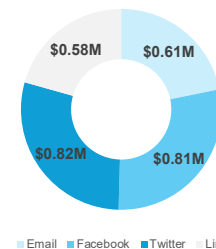
Revenue by Year and Quarter



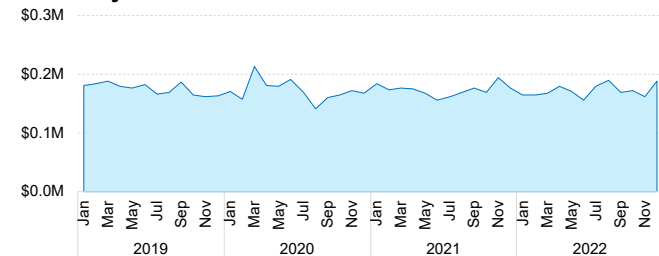
Cost BreakDown



Sales by Source



Profit by Year and Month



Sales Rep	Profit	Revenue	Profit Margin
Tokyo Office	\$1.70M	\$2.81M	60.57%
Osaka Office	\$1.68M	\$2.83M	59.40%
Kyoto Office	\$1.67M	\$2.81M	59.31%
Hokaido Office	\$1.65M	\$2.78M	59.25%
Kanagawa Office	\$1.62M	\$2.74M	59.14%

Phần 4: Kiểm Định Giả Thuyết (Hypothesis Testing)

Tại Sao Phải Kiểm Định Giả Thuyết?

Kiểm định giả thuyết là phương pháp thống kê giúp xác định tính đúng đắn của một giả thuyết dựa trên dữ liệu, cho phép đưa ra kết luận về tổng thể từ mẫu nghiên cứu.

1 Đưa ra quyết định khách quan

Thay thế quyết định cảm tính bằng bằng chứng thống kê, giảm thiểu sai lầm chủ quan.

2 Xác định ý nghĩa thống kê

Phân biệt kết quả ngẫu nhiên với hiệu ứng có ý nghĩa thực sự.

3 Đánh giá mối quan hệ và sự khác biệt

Xác định mối liên hệ giữa các biến và phát hiện khác biệt đáng kể giữa các nhóm.

4 Xác minh hiệu quả can thiệp

Đánh giá liệu phương pháp mới có tạo ra khác biệt có ý nghĩa so với phương pháp hiện tại.

Kiểm Định Giả Thuyết Là Gì?

Là quy trình thống kê để kết luận về giả thuyết từ dữ liệu. Phương pháp kiểm tra xem giả thuyết về tổng thể có được hỗ trợ bởi dữ liệu mẫu hay không.

Đặt giả thuyết

Xác định giả thuyết không (H_0) và thay thế (H_1)

- H_0 : Không có sự khác biệt/ảnh hưởng
- H_1 : Có sự khác biệt/ảnh hưởng

Tính toán giá trị thống kê

Sử dụng Excel để tính:

- t-value (T.TEST, T.INV.2T)
- z-value (NORM.S.DIST, NORM.S.INV)
- Chi-square (CHISQ.TEST, CHISQ.INV)

Xác định p-value

So sánh p-value với mức ý nghĩa (α)

- $p\text{-value} < \alpha$: Bác bỏ H_0
- $p\text{-value} \geq \alpha$: Không đủ cơ sở bác bỏ H_0

Thực hiện trong Excel

Các công cụ sẵn có:

- Data Analysis ToolPak (t-Test, ANOVA)
- Hàm thống kê (T.TEST, F.TEST, CHISQ.TEST)
- Tính toán thủ công với hàm xác suất

Population vs Sample

Tổng Thể (Population)

Tập hợp đầy đủ tất cả các đối tượng mà ta muốn nghiên cứu.

- Thường quá lớn hoặc không thực tế để kiểm tra toàn bộ
- Có các thông số đặc trưng gọi là tham số (parameters)
- Ký hiệu: μ (giá trị trung bình), σ (độ lệch chuẩn)

Mẫu (Sample)

Một phần nhỏ được chọn từ tổng thể để nghiên cứu.

- Dùng để suy luận về tổng thể
- Có các đặc điểm được gọi là thống kê mẫu (statistics)
- Ký hiệu: \bar{x} (giá trị trung bình mẫu), s (độ lệch chuẩn mẫu)

Mối Liên Hệ Với Kiểm Định Giả Thuyết



Suy Luận Từ Mẫu Về Tổng Thể

Kiểm định giả thuyết sử dụng dữ liệu mẫu để đưa ra kết luận về tổng thể, giúp kiểm tra tính đúng đắn của giả thuyết H_0 .



Mức Độ Tin Cậy

Mẫu càng đại diện tốt cho tổng thể, kết quả kiểm định càng đáng tin cậy, nhờ đó giảm thiểu sai số loại I và loại II.



Lấy Mẫu Ngẫu Nhiên

Để kết quả kiểm định giả thuyết có giá trị, mẫu phải được chọn một cách ngẫu nhiên và đủ đại diện cho tổng thể.

Các Khái Niệm Quan Trọng 1/2

Kiểm định giả thuyết dựa trên bốn khái niệm cơ bản: mức ý nghĩa (alpha), p-value, t-value và cặp giả thuyết H_0/H_1 .

Mức ý nghĩa (alpha)

Ngưỡng quyết định (thường là 0.05) để đánh giá kết quả thống kê. Nếu $p\text{-value} < \alpha$, bác bỏ H_0 . $\alpha = 0.05$ nghĩa là chấp nhận 5% khả năng sai lầm loại I.

p-value

Xác suất quan sát được kết quả cực đoan như dữ liệu mẫu, giả định H_0 đúng. p-value càng nhỏ, bằng chứng chống lại H_0 càng mạnh. Khi $p\text{-value} < \alpha$: bác bỏ H_0 .

t-value

Đo lường khoảng cách giữa trung bình mẫu và giá trị giả định trong H_0 . $|t|$ càng lớn, sự khác biệt càng đáng kể. Công thức: $t = (\bar{x} - \mu) / (s/\sqrt{n})$.

Giả thuyết H_0 và H_1

H_0 (giả thuyết gốc): Không có sự khác biệt/tác động ($\mu_1 = \mu_2$). H_1 (giả thuyết thay thế): Có sự khác biệt/tác động ($\mu_1 \neq \mu_2$). Kiểm định nhằm xác định có đủ bằng chứng bác bỏ H_0 không.

Các Khái Niệm Quan Trọng 2/2

Phân phối xác suất

Mô tả toán học về khả năng xuất hiện của các giá trị của một biến ngẫu nhiên. Phân phối xác suất cho biết mỗi giá trị có thể xảy ra với xác suất bao nhiêu.

Phân phối chuẩn(Normal Distribution)

Phân phối hình chuông đối xứng, được mô tả bởi giá trị trung bình (μ) và độ lệch chuẩn (σ). Nhiều hiện tượng tự nhiên và dữ liệu thực tế tuân theo phân phối này, làm cơ sở cho nhiều kiểm định thống kê.

Định lý giới hạn trung tâm (Central Limit Theorem – CLT)

Khi kích thước mẫu đủ lớn ($n \geq 30$), phân phối của trung bình mẫu sẽ xấp xỉ phân phối chuẩn, bất kể phân phối ban đầu của dữ liệu. Đây là cơ sở cho nhiều phương pháp suy luận thống kê.

Sai lầm loại I, II (Type I, II Error)

Sai lầm loại I (α): Bác bỏ H_0 khi nó đúng (phát hiện dương tính giả).
Sai lầm loại II (β): Không bác bỏ H_0 khi nó sai (phát hiện âm tính giả). Mỗi loại sai lầm có hậu quả khác nhau tùy vào bối cảnh nghiên cứu.

Phân Phối Xác Suất Là Gì?

Phân phối xác suất mô tả khả năng xảy ra của các kết quả trong một thử nghiệm.

Hiểu đơn giản về phân phối xác suất

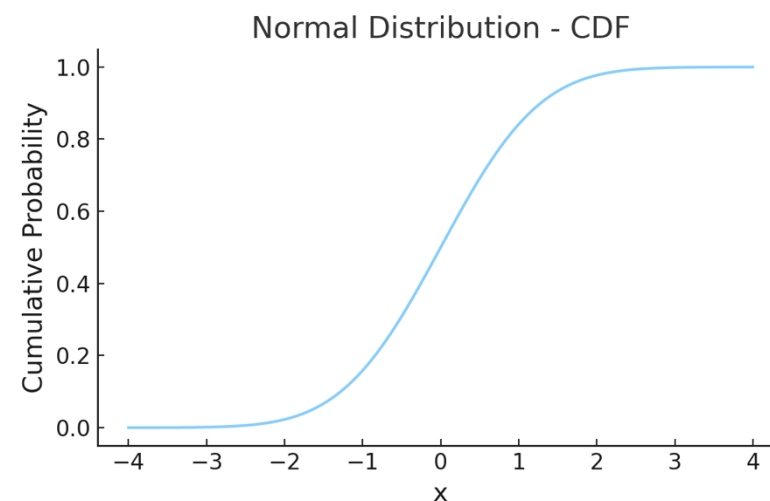
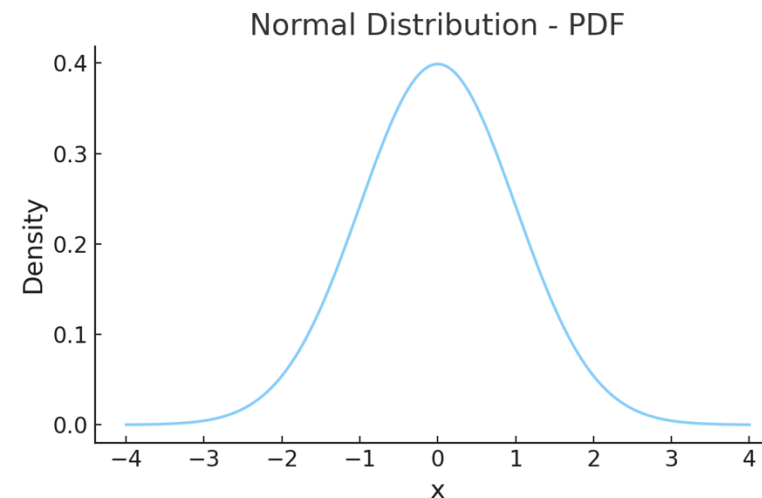
Khi tung đồng xu, xác suất ra mặt ngửa hoặc sấp là 50%. Nhiều sự kiện tự nhiên khi ghi lại (như chiều cao) tạo nên các "hình dạng" đặc trưng trên biểu đồ.

Hàm mật độ xác suất (PDF)

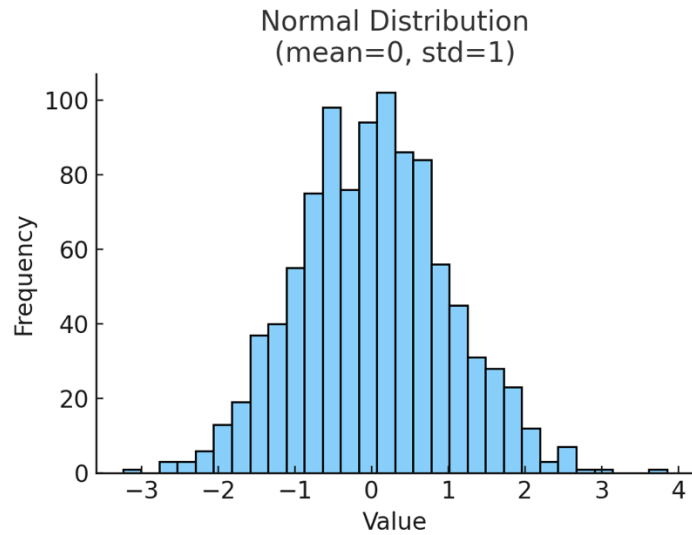
PDF biểu thị khả năng xuất hiện của các giá trị. Diện tích dưới đường cong bằng 1, giúp tính xác suất giá trị trong một khoảng.

Hàm phân phối tích lũy (CDF)

CDF cho biết xác suất một giá trị nhỏ hơn hoặc bằng một mức. CDF tăng dần từ 0 đến 1.



3 Loại Phân Phối Thường Gặp

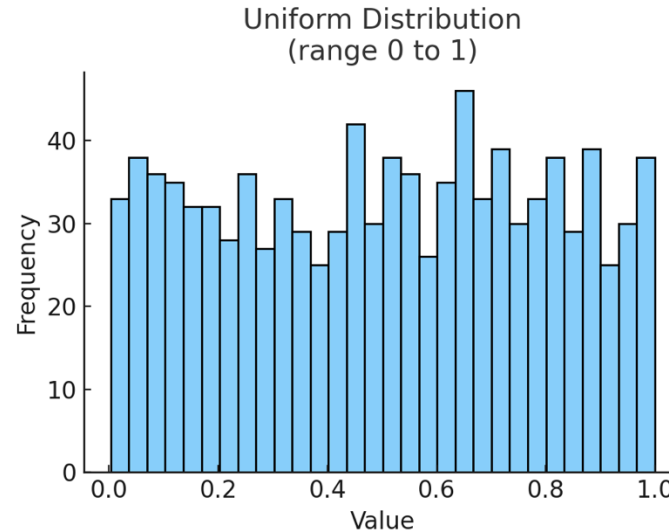


Phân Phối Chuẩn (Normal)

Dạng hình chuông đặc trưng, thường xuất hiện trong tự nhiên.

- Chiều cao của người trưởng thành
- Điểm thi của học sinh trong lớp đông
- Sai số đo lường trong khoa học

Trong Excel: Sử dụng hàm `NORM.DIST()` để tính xác suất

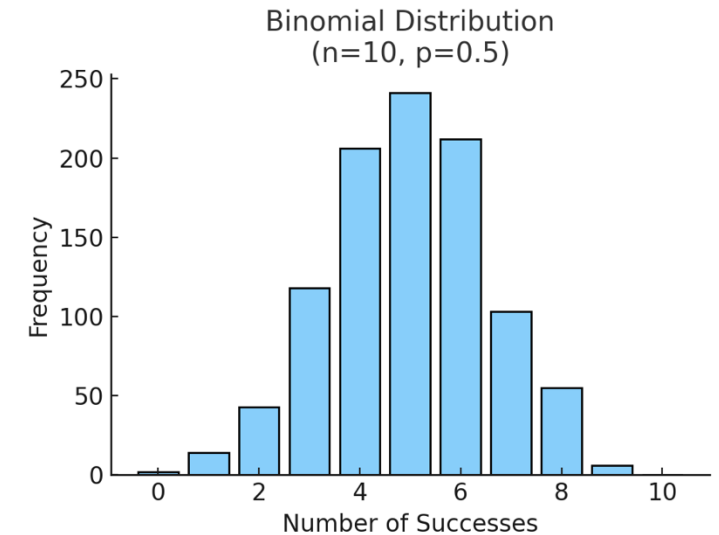


Phân Phối Đồng Đều (Uniform)

Mọi giá trị có xác suất xuất hiện như nhau.

- Số hiển thị khi tung xúc xắc
- Thời gian chờ đợi giữa 2 khách hàng
- Vị trí ngẫu nhiên của điểm trên đoạn thẳng

Trong Excel: Dùng hàm `RAND()` để tạo số ngẫu nhiên đồng đều



Phân Phối Nhị Thức (Binomial)

Mô tả số lần thành công trong n thử nghiệm độc lập.

- Số lần tung đồng xu được mặt ngửa
- Số sản phẩm lỗi trong dây chuyền sản xuất
- Số khách hàng chấp nhận mua sản phẩm

Trong Excel: Sử dụng hàm `BINOM.DIST()` để tính xác suất

Central Limit Theorem

Định lý giới hạn trung tâm khẳng định rằng với mẫu đủ lớn từ bất kỳ quần thể nào, phân phối của giá trị trung bình mẫu sẽ xấp xỉ phân phối chuẩn.

Tại sao định lý này đúng?

- **Ý tưởng cơ bản:** Khi lấy trung bình nhiều biến ngẫu nhiên, các lệch lạc cá nhân sẽ triệt tiêu, tạo ra biến mới **ổn định và cân đối hơn** (Điều kiện: Mẫu cần đủ lớn ($n \geq 30$) và các quan sát phải độc lập).
- **Hệ quả:** Phân phối mẫu có trung bình bằng μ và độ lệch chuẩn bằng σ/\sqrt{n} .

Ứng dụng của định lý

- **Xây dựng khoảng tin cậy (Confidence Interval):** Từ mẫu nhỏ, **ước lượng toàn bộ tổng thể** bằng cách dùng công thức chuẩn xác định khoảng giá trị chứa trung bình thực.
- **Kiểm định giả thuyết:** Các kiểm định t, z, chi-square dựa vào **phân phối trung bình mẫu là chuẩn** để đưa ra kết luận về tổng thể.
- **Ứng dụng thực tế:** Dự đoán hành vi hệ thống lớn từ mẫu nhỏ. Kiểm soát quy trình sản xuất bằng giám sát mẫu tính ngẫu nhiên.

Mức Ý Nghĩa và Khoảng Tin Cậy

Mức Ý Nghĩa (Significance Level)

Là ngưỡng xác suất để quyết định bác bỏ giả thuyết không (H_0).

- Thường được ký hiệu là α (alpha), giá trị phổ biến: 0.05, 0.01, 0.1
- $p\text{-value} < \alpha$: Bác bỏ H_0 , kết quả có ý nghĩa thống kê
- $p\text{-value} \geq \alpha$: Không đủ bằng chứng để bác bỏ H_0

Mức ý nghĩa 5% ($\alpha = 0.05$) nghĩa là chấp nhận rủi ro 5% kết luận sai khi bác bỏ H_0 .

Mối Quan Hệ Giữa Mức Ý Nghĩa và Khoảng Tin Cậy

Mức ý nghĩa $\alpha = 0.05$ tương ứng với khoảng tin cậy 95%. Hai cách tiếp cận bổ sung cho nhau: kiểm định dùng p-value, ước lượng dùng khoảng tin cậy.

Khoảng Tin Cậy (Confidence Interval)

Là khoảng giá trị ước lượng chứa tham số thực của tổng thể với mức độ tin cậy nhất định.

- Công thức: Ước lượng \pm Margin of Error
- CI 95%: Có 95% khả năng khoảng chứa giá trị thực của tham số
- Khoảng hẹp = độ chính xác cao, khoảng rộng = độ chính xác thấp

t-value & p-value

t-value (Giá trị t)

Là thước đo sự khác biệt giữa trung bình mẫu và giá trị giả định của tổng thể, tính theo đơn vị độ lệch chuẩn.

Công thức tính:

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$

Trong đó:

- \bar{x} : Giá trị trung bình của mẫu
- μ : Giá trị giả định của tổng thể (từ H_0)
- s : Độ lệch chuẩn của mẫu
- n : Kích thước mẫu

Giá trị t càng lớn, bằng chứng càng mạnh để bác bỏ giả thuyết không (H_0).

p-value (Giá trị p)

Là xác suất quan sát được kết quả cực đoan như dữ liệu mẫu (hoặc cực đoan hơn) với giả định H_0 là đúng.

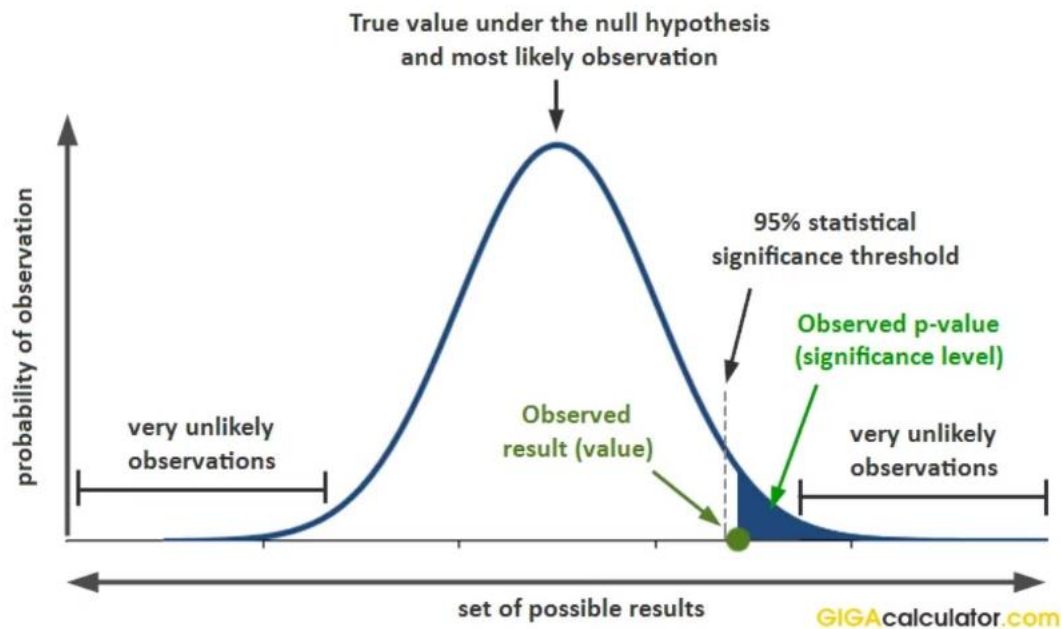
Cách xác định:

1. Tính t-value từ dữ liệu mẫu
2. Dựa vào phân phối t với bậc tự do phù hợp ($df = n-1$) để tìm xác suất
3. Trong Excel: =T.DIST.2T(|t-value|, df) cho kiểm định hai phía

P-value nhỏ (<0.05) nghĩa là có đủ bằng chứng để bác bỏ H_0 với mức ý nghĩa 5%.

t-value & p-value trong phân phối chuẩn

Probability & Statistical Significance Explained



Trong phân phối chuẩn:

- t-value: Khoảng cách từ giá trị trung bình theo độ lệch chuẩn
- Với phân phối chuẩn, t-value > 1.96 hoặc < -1.96 cho p-value < 0.05 (kiểm định hai phía)
- Vùng bác bỏ H_0 nằm ở hai đuôi của đường cong

Image Source: <https://medium.datadriveninvestor.com/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e>

So sánh giữa t-value và z-value

t-value

- Dùng khi không biết phương sai tổng thể
- Ước lượng từ phương sai mẫu
- Phụ thuộc vào bậc tự do (df)
- Chịu ảnh hưởng bởi mẫu nhỏ

z-value

- Dùng khi biết phương sai tổng thể
- Độc lập với kích thước mẫu
- Dựa trên phân phối chuẩn
- Phù hợp với mẫu lớn ($n > 30$)

Khi mẫu đủ lớn ($n > 30$), phân phối t xấp xỉ phân phối chuẩn, khiến t-value và z-value trở nên tương đương.



Kích thước mẫu tăng

Khi n tăng, bậc tự do ($df = n - 1$) cũng tăng, làm phân phối t gần với phân phối chuẩn.



Tiệm cận đến phân phối chuẩn

Với $n \rightarrow \infty$, phân phối t tiệm cận phân phối chuẩn, làm t-value gần giống z-value.



Ứng dụng thực tế

Nhiều nhà thống kê coi hai giá trị tương đương khi $n > 30$, cho phép dùng bảng z thay bảng t.

Kiểm định bằng Excel

Quy trình kiểm định giả thuyết thống kê trên Excel:

1

Xác định giả thuyết

Thiết lập giả thuyết gốc (H_0) và giả thuyết thay thế (H_1) dựa trên vấn đề cần kiểm định

2

Chọn kiểm định phù hợp

Lựa chọn kiểm định thích hợp như t-test, z-test, ANOVA, hoặc kiểm định F tùy thuộc vào dữ liệu và mục tiêu phân tích

3

Thực hiện kiểm định

Sử dụng Data Analysis Tool trên Excel để tiến hành kiểm định đã chọn với dữ liệu của bạn

4

Phân tích kết quả

So sánh p-value với mức ý nghĩa alpha (thường là 0.05). Nếu p-value < alpha, bác bỏ giả thuyết gốc H_0

5

Đưa ra kết luận

Diễn giải kết quả trong ngữ cảnh thực tế và đưa ra quyết định dựa trên kết quả kiểm định

Các Loại Kiểm Định Phổ Biến

Lựa chọn kiểm định phù hợp dựa trên loại dữ liệu và mục tiêu phân tích

t-Test

Mục đích: So sánh trung bình giữa hai nhóm

Ứng dụng: So sánh hiệu quả hai phương pháp, đánh giá trước-sau can thiệp

- One-sample t-test: So sánh với giá trị chuẩn
- Paired t-test: So sánh các cặp dữ liệu liên quan
- Independent t-test: So sánh hai nhóm độc lập

ANOVA

Mục đích: So sánh trung bình giữa nhiều nhóm (>2)

Ứng dụng: So sánh hiệu quả của nhiều phương pháp khác nhau

- One-way ANOVA: Một biến phân loại
- Two-way ANOVA: Hai biến phân loại
- MANOVA: Nhiều biến phụ thuộc

Kiểm Định Phi Tham Số

Mục đích: Dùng khi dữ liệu không tuân theo phân phối chuẩn

Ứng dụng: Phân tích dữ liệu thứ bậc, dữ liệu nhỏ, phân phối lệch

- Mann-Whitney U: Thay thế cho t-test độc lập
- Wilcoxon: Thay thế cho paired t-test
- Kruskal-Wallis: Thay thế cho ANOVA
- Chi-square: Kiểm định tính độc lập giữa các biến phân loại

Kiểm định hiệu quả chiến dịch marketing mới

1

1. Xác định vấn đề

- Công ty ABC muốn biết liệu chiến dịch quảng cáo mới có làm tăng doanh số bán hàng không

2

2. Thu thập dữ liệu

- Doanh số trước chiến dịch ($n=30$): trung bình $\mu_1=520$ triệu đồng/tháng
- Doanh số sau chiến dịch ($n=30$): trung bình $\mu_2=580$ triệu đồng/tháng

3

3. Thiết lập giả thuyết

- $H_0: \mu_1 = \mu_2$ (Doanh số trung bình không thay đổi)
- $H_1: \mu_1 < \mu_2$ (Doanh số trung bình tăng lên)

4

4. Thực hiện kiểm định

- Sử dụng t-Test: Paired Two Sample for Means từ Data Analysis ToolPak
- Kết quả:** p-value = 0.023, t-stat = 2.458

5

5. Đưa ra kết luận

- p-value = 0.023 < $\alpha = 0.05$, nên bác bỏ H_0

Kết luận: Với mức ý nghĩa 5%, có đủ bằng chứng thống kê để kết luận rằng chiến dịch marketing mới đã thực sự làm tăng doanh số bán hàng. Dữ liệu cho thấy doanh số đã tăng trung bình 60 triệu đồng mỗi tháng sau khi triển khai chiến dịch.

One-sample t-test

Kiểm chứng xem giá bán trung bình của sản phẩm điện tử trên thị trường có phải là 990USD hay không?

Case study	Kiểm định giá bán trung bình sản phẩm điện tử	
	Mục đích: Kiểm chứng xem giá bán trung bình của sản phẩm điện tử trên thị trường có thật sự là 990USD hay không	
Bước 1	Xác định giả thuyết	Giá trị
	Giả thuyết gốc (H_0)	Giá bán trung bình của sản phẩm điện tử = 990
	Giả thuyết thay thế (H_1)	Giá bán trung bình của sản phẩm điện tử \neq 990
	Mức ý nghĩa (α)	0.05
Bước 2	Lựa chọn kiểm định: t-test và thực hiện tính toán thống số liên quan	
	Phương pháp kiểm định	One-sample t-test
	Tính giá trị trung bình (Sample Mean \bar{x})	997.71
	Tính Độ Lệch Chuẩn (Sample Standard Deviation -s)	49.57
	Tính số lượng mẫu (n)	200
	Giá trị trung bình của Population (μ)	990
	Tính tử số hiệu của ($\bar{x} - \mu$)	7.71
	Tính mẫu số (s / \sqrt{n})	3.51
	Giá trị t (t-statistic) = $(\bar{x} - \mu) / (s / \sqrt{n})$	2.20
	Bậc tự do df = Số lượng mẫu n - 1	199
	Giá trị p (p-value)	0.0291
Bước 3	Đưa ra kết luận	
	So sánh p-value với α	$p = 0.0291 < 0.05$
	Kết luận thống kê	Có đủ bằng chứng để bác bỏ giả thuyết H_0 ở mức ý nghĩa 5%.
	Diễn giải kinh doanh	Giá bán trung bình của sản phẩm điện tử trên thị trường không thực sự gần với 990 USD. Mức giá trung bình thực tế là 997.71, cao hơn một cách có ý nghĩa thống kê.

Product Category	Company ID	Price per Unit (\$)
Electronics	ID_1	973
Electronics	ID_2	890
Electronics	ID_3	1038
Electronics	ID_4	1036
Electronics	ID_5	971
Electronics	ID_6	965
Electronics	ID_7	1037
Electronics	ID_8	1035
Electronics	ID_9	1000
Electronics	ID_10	943
Electronics	ID_11	935
Electronics	ID_12	953
Electronics	ID_13	989
Electronics	ID_14	1012
Electronics	ID_15	970
Electronics	ID_16	926
Electronics	ID_17	1033
Electronics	ID_18	1092
Electronics	ID_19	1017
Electronics	ID_20	1107
Electronics	ID_21	1109
Electronics	ID_22	967
Electronics	ID_23	985
Electronics	ID_24	1007
Electronics	ID_195	1054
Electronics	ID_196	1035
Electronics	ID_197	1078
Electronics	ID_198	914
Electronics	ID_199	996
Electronics	ID_200	972

Chi tiết:
part4_hypothesis_testing_examples.xlsx > one_sample_t_test

t-Test: Two-Sample Assuming Unequal Variances

Xác định có sự khác biệt đáng kể về giá bán trung bình giữa sản phẩm của Company A và Company B không?

Dữ liệu

- 32 sản phẩm của Company A
- 37 sản phẩm của Company B
- **Giả thuyết H_0 :** Không có sự khác biệt về giá bán trung bình giữa hai công ty.
- **Giả thuyết H_1 :** Có sự khác biệt về giá bán trung bình giữa hai công ty.

Kết quả kiểm định

t-Test: Two-Sample Assuming Unequal Variances		
	Price_Company A	Price_Company B
Mean	372.1875	421.4594595
Variance	21209.57661	11924.08859
Observations	32	37
Hypothesized Mean Difference	0	
df	57	
t Stat	-1.569877847	
P(T<=t) one-tail	0.060989334	
t Critical one-tail	1.672028888	
P(T<=t) two-tail	0.121978669	
t Critical two-tail	2.002465459	

Diễn giải kết quả

p-value = 0.122 > 0.05, nên không đủ bằng chứng để bác bỏ H_0 ở mức ý nghĩa 5%.

Nói cách khác: Chưa thể kết luận có sự khác biệt có ý nghĩa thống kê giữa giá bán trung bình của hai công ty.

Chi tiết: part4_hypothesis_testing_examples.xlsx > two_samples_t_test

Các Hàm và Công Cụ Thống Kê Trong Excel

Tổng hợp các hàm và công cụ Excel thường dùng cho kiểm định thống kê:

Loại kiểm định	Hàm/Công cụ Excel	Mô tả
Kiểm định t (t-Test)	Data Analysis > t-Test	So sánh trung bình của 2 tập dữ liệu (3 loại: paired, equal variance, unequal variance)
Kiểm định z (z-Test)	Data Analysis > z-Test	So sánh trung bình với mẫu lớn hoặc khi biết phương sai tổng thể
ANOVA	Data Analysis > ANOVA	So sánh trung bình của nhiều nhóm (single factor, two-factor)
Tương quan	CORREL(), Data Analysis > Correlation	Đo lường mối quan hệ tuyến tính giữa các biến
Hồi quy tuyến tính	Data Analysis > Regression	Phân tích mối quan hệ giữa biến phụ thuộc và biến độc lập
Thống kê mô tả	Data Analysis > Descriptive Statistics	Cung cấp các thống kê cơ bản (mean, median, mode, standard deviation, etc.)
Hàm phân phối chuẩn	NORM.DIST(), NORM.INV(), NORM.S.DIST()	Tính xác suất và giá trị của phân phối chuẩn
Hàm phân phối t	T.DIST(), T.INV(), T.TEST()	Tính xác suất và giá trị của phân phối t-Student
Hàm phân phối F	F.DIST(), F.INV(), F.TEST()	Tính xác suất và giá trị của phân phối F (dùng trong ANOVA)
Hàm phân phối Chi bình phương	CHISQ.DIST(), CHISQ.INV(), CHISQ.TEST()	Kiểm định tính độc lập giữa các biến phân loại
Phân tích p-value	Data Analysis (kết quả có sẵn)	So sánh p-value với mức ý nghĩa alpha để đưa ra kết luận

ⓘ Lưu ý: Để sử dụng Data Analysis, cần kích hoạt Add-in "Analysis ToolPak" trong Excel (File > Options > Add-ins).

Phần 5: Tiền Xử Lý Dữ Liệu

Tại sao cần tiền xử lý dữ liệu?

Tiền xử lý dữ liệu giúp làm sạch dữ liệu thô, nâng cao độ chính xác trong phân tích và tạo nền tảng cho việc phân tích chính xác



Làm sạch dữ liệu thô

Dữ liệu thô thường chứa giá trị thiếu, sai lệch và định dạng không đồng nhất cần được xử lý.



Đảm bảo độ chính xác

Chuẩn hóa và biến đổi dữ liệu thành định dạng phù hợp cho việc phân tích hiệu quả.



Hỗ trợ quyết định chính xác

Loại bỏ giá trị ngoại lệ và xử lý dữ liệu thiếu giúp ngăn chặn sai lệch và tối ưu quyết định kinh doanh.

Các phương pháp xử lý dữ liệu phổ biến

Excel cung cấp công cụ xử lý dữ liệu thiếu, loại bỏ giá trị bất thường, tạo biến giả và kiểm tra tính chính xác

1 Xử lý giá trị thiếu (Null values)

Sử dụng IF, ISBLANK và IFERROR để phát hiện ô trống, thay thế bằng giá trị trung bình, trung vị hoặc phổ biến nhất.

2 Loại bỏ giá trị ngoại lệ (Outliers)

Áp dụng phương pháp IQR với QUARTILE hoặc Z-score (STANDARDIZE) để xác định và xử lý giá trị nằm ngoài khoảng tin cậy.

3 Tạo biến giả (Dummy) cho biến phân loại

Chuyển biến phân loại thành biến nhị phân (0/1) bằng IF, SWITCH hoặc IFS để cải thiện phân tích hồi quy.

4 Kiểm tra và đánh giá sau xử lý

Sử dụng biểu đồ phân phối, PivotTable và thống kê mô tả để xác nhận dữ liệu đã được xử lý đúng.

Xử lý giá trị thiếu (Null values)

Xử lý đúng cách các giá trị thiếu là yếu tố quan trọng ảnh hưởng trực tiếp đến kết quả phân tích dữ liệu.

Phát hiện giá trị thiếu

Sử dụng ISBLANK(), ISNA(), IFERROR() để xác định các ô trống trong dữ liệu.

Thay thế bằng giá trị thống kê


Áp dụng AVERAGE, MEDIAN, hoặc MODE.SNGL để thay thế mà không gây sai lệch phân tích.

Lọc hoặc loại bỏ

Với dữ liệu có ít giá trị thiếu, dùng Filter hoặc Advanced Filter để lọc hoặc loại bỏ chúng.

Dự đoán giá trị thiếu

Dùng FORECAST hoặc phân tích hồi quy để ước tính giá trị dựa trên mối quan hệ với các biến khác.

 Lựa chọn phương pháp xử lý phụ thuộc vào bản chất dữ liệu, mục đích phân tích và tỷ lệ giá trị thiếu. Luôn ghi chú phương pháp đã sử dụng để đảm bảo tính minh bạch.

Phát Hiện Giá Trị Ngoại Lệ Trong Excel

Giá trị ngoại lệ (outliers) là những quan sát có giá trị cực kỳ cao hoặc thấp, có thể làm sai lệch kết quả phân tích và dự báo.

Nhận diện giá trị ngoại lệ


- Sử dụng Box Plot, Histogram hoặc Scatter Plot để trực quan hóa outliers.
- Áp dụng nguyên tắc IQR: Giá trị ngoài khoảng $(Q1 - 1.5 \cdot IQR)$ và $(Q3 + 1.5 \cdot IQR)$ là ngoại lệ.

Phương pháp thống kê phát hiện outliers

- Dùng QUARTILE.EXC để tính Q1 và Q3, sau đó áp dụng công thức IQR.
- Z-score: Sử dụng hàm STANDARDIZE, giá trị có $|z| > 3$ thường là ngoại lệ.

Kỹ thuật xử lý outliers

- Loại bỏ: Dùng Filter để loại trừ giá trị ngoại lệ nếu chắc chắn là sai sót.
- Thay thế: Áp dụng MEDIAN hoặc phương pháp Winsorization.

 Xử lý outliers cần được thực hiện cẩn thận, có cơ sở khoa học và phù hợp với ngữ cảnh phân tích.

Tạo biến giả (Dummy) cho biến phân loại

Kỹ thuật chuyển đổi biến phân loại thành dạng số học phù hợp cho phân tích định lượng trong Excel.

Khái niệm biến giả (Dummy)

Biến nhị phân (0/1) đại diện cho giá trị của biến phân loại, giúp đưa dữ liệu định tính vào mô hình định lượng.

Quy tắc tạo biến giả

Với n giá trị phân loại, tạo $(n-1)$ biến giả để tránh đa cộng tuyến. Giá trị 1 đại diện "có", 0 đại diện "không".

Tạo biến giả trong Excel

Sử dụng IF, IFS hoặc SWITCH kết hợp công thức mảng và Power Query để tự động hóa tạo biến giả cho dữ liệu lớn.

Kỹ thuật này đóng vai trò quan trọng trong phân tích hồi quy, phân tích phương sai và dự báo khi làm việc với dữ liệu phi số học.

Kiểm tra và đánh giá sau xử lý

Kiểm tra kết quả sau tiền xử lý là bước cần thiết để đảm bảo dữ liệu sẵn sàng cho phân tích.

Kiểm tra tính nhất quán

Kiểm tra các mâu thuẫn trong dữ liệu sau xử lý bằng COUNTIF, AVERAGEIF và các hàm điều kiện.

Đánh giá phân phối

Xem xét phân phối sau chuẩn hóa qua Histogram và QQ-plot. Sử dụng Data Analysis ToolPak để kiểm tra tính chuẩn.

So sánh trước-sau

Tạo biểu đồ so sánh để đánh giá hiệu quả tiền xử lý. Dùng bar charts và scatter plots để trực quan hóa thay đổi.

Lập lại quy trình

Tối ưu quy trình dựa trên đánh giá. Sử dụng macro và Power Query để tự động hóa, đảm bảo nhất quán và tiết kiệm thời gian.

Đánh giá kỹ lưỡng giúp phát hiện sớm vấn đề tiềm ẩn, cho phép điều chỉnh phương pháp xử lý, từ đó nâng cao độ chính xác của phân tích và dự báo.

<Ví dụ>

Tiền xử lý dữ liệu doanh số bán hàng

Dữ liệu thô (trước xử lý)

Tháng	Doanh số (triệu VND)	Khu vực	Ghi chú
T1/2023	45.2	Miền Bắc	Đầy đủ
T2/2023	NULL	Miền Nam	Thiếu dữ liệu
T3/2023	52.7	Miền Bắc	Đầy đủ
T4/2023	198.5	Miền Bắc	Nghi ngờ sai sót
T5/2023	54.8	MB	Đầy đủ
T6/2023	-12.3	Miền Trung	Giá trị âm

Dữ liệu sau khi xử lý

Tháng	Doanh số (triệu VND)	Khu vực	Miền_Bắc
T1/2023	45.2	Miền Bắc	1
T2/2023	48.95	Miền Nam	2
T3/2023	52.7	Miền Bắc	1
T4/2023	54.8	Miền Bắc	1
T5/2023	54.8	Miền Bắc	1
T6/2023	53.7	Miền Trung	3

- **Điền giá trị thiếu:** T2/2023 được điền bằng giá trị trung bình của T1 và T3 (48.95)
- **Xử lý giá trị ngoại lệ:** T4/2023 (198.5) được thay thế bằng trung vị của dữ liệu hợp lệ (54.8)
- **Chuẩn hóa tên:** "MB" được thống nhất thành "Miền Bắc"
- **Xử lý giá trị âm:** T6/2023 (-12.3) được thay bằng giá trị dự đoán từ xu hướng dữ liệu (53.7)
- **Tạo biến giả:** Thêm cột cho 3 miền với giá trị 1,2,3 đại diện cho khu vực Miền Bắc

Hàm Excel Phổ Biến Trong Tiền Xử Lý Dữ Liệu

Bảng dưới đây tổng hợp các phương pháp tiền xử lý dữ liệu phổ biến và các hàm Excel tương ứng để thực hiện chúng.

Phương pháp tiền xử lý	Hàm Excel/Công cụ	Mô tả
Xử lý giá trị thiếu (Null)	IFERROR(), IF(ISBLANK()), AVERAGE(), MEDIAN()	Phát hiện và điền giá trị thiếu bằng giá trị trung bình, trung vị hoặc giá trị dự đoán
Loại bỏ giá trị trùng lặp	Remove Duplicates, UNIQUE()	Loại bỏ các bản ghi trùng lặp trong tập dữ liệu
Phát hiện giá trị ngoại lệ	QUARTILE(), STDEV(), IF(), AVERAGEIF()	Xác định và xử lý các giá trị nằm ngoài khoảng bình thường
Chuẩn hóa dữ liệu	UPPER(), LOWER(), PROPER(), TRIM()	Thống nhất định dạng văn bản, loại bỏ khoảng trắng thừa
Tạo biến giả (Dummy)	IF(), IFS(), VLOOKUP()	Chuyển đổi biến phân loại thành biến nhị phân (0/1) hoặc mã hóa
Biến đổi dữ liệu	LN(), SQRT(), POWER(), LOG10()	Chuyển đổi phân phối dữ liệu (logarit, căn bậc hai, bình phương...)
Điều chỉnh thang đo	STANDARDIZE(), MIN(), MAX()	Chuẩn hóa dữ liệu về cùng thang đo (ví dụ: 0-1 hoặc z-score)
Tách cột dữ liệu	Text to Columns, LEFT(), RIGHT(), MID()	Phân tách dữ liệu từ một cột thành nhiều cột
Gộp dữ liệu	CONCATENATE(), &, TEXTJOIN()	Kết hợp dữ liệu từ nhiều cột thành một
Chuyển đổi định dạng thời gian	DATE(), DATEVALUE(), TEXT()	Chuẩn hóa các định dạng ngày tháng khác nhau

Phương Pháp Kết Hợp Bảng Dữ Liệu Trong Excel

Dưới đây là các phương pháp phổ biến để kết hợp nhiều bảng dữ liệu thành một bảng trong Excel:

Phương Pháp	Công Cụ/Hàm Excel	Mô Tả	Ưu Điểm	Hạn Chế
Tra cứu dữ liệu	VLOOKUP(), HLOOKUP()	Tìm kiếm giá trị dựa trên cột khóa và trả về giá trị tương ứng từ bảng khác	Dễ sử dụng, phổ biến	Chỉ tra cứu từ trái sang phải, không linh hoạt với dữ liệu thay đổi
Tra cứu đa chiều	INDEX() + MATCH()	Kết hợp để tìm giá trị từ bảng khác dựa trên dòng và cột	Linh hoạt hơn VLOOKUP, tìm kiếm theo mọi hướng	Cú pháp phức tạp hơn, khó học hơn
Tra cứu nâng cao	XLOOKUP()	Hàm tra cứu hiện đại thay thế VLOOKUP và INDEX-MATCH	Linh hoạt, hỗ trợ tìm kiếm theo nhiều hướng, có giá trị mặc định	Chỉ có trong Excel 365 và các phiên bản mới hơn
Nối dữ liệu	Power Query (Get & Transform)	Kết nối và biến đổi dữ liệu từ nhiều nguồn	Tự động làm mới, xử lý được khối lượng dữ liệu lớn	Yêu cầu học thêm về cách sử dụng Power Query
Tạo bảng động	Pivot Table	Tổng hợp và phân tích dữ liệu từ nhiều bảng	Phân tích linh hoạt, tính toán tự động	Chủ yếu dùng để tổng hợp, không phải kết hợp dữ liệu chi tiết
Hợp nhất dữ liệu	Consolidate	Kết hợp dữ liệu từ nhiều vùng hay sheet	Dễ sử dụng với dữ liệu có cùng cấu trúc	Ít linh hoạt, khó xử lý dữ liệu phức tạp
Tạo quan hệ	Data Model	Thiết lập quan hệ giữa các bảng trong mô hình dữ liệu Excel	Mạnh mẽ, xử lý được khối lượng dữ liệu lớn, quan hệ phức tạp	Yêu cầu hiểu biết về mô hình dữ liệu, phức tạp hơn
Công thức mảng	FILTER(), UNIQUE(), SORT()	Sử dụng các hàm mảng động để kết hợp và lọc dữ liệu	Mạnh mẽ, xử lý được các điều kiện phức tạp	Chỉ có trong Excel 365, đòi hỏi hiểu biết về công thức mảng

Phần 6: Sử dụng các mô hình hồi quy đơn biến và đa biến

Tại sao phải Phân Tích Hồi Quy?

Phân tích hồi quy là gì?

Kỹ thuật thống kê xác định mối quan hệ giữa biến phụ thuộc (Y) và biến độc lập (X), giúp hiểu cách một biến thay đổi khi biến khác biến động.

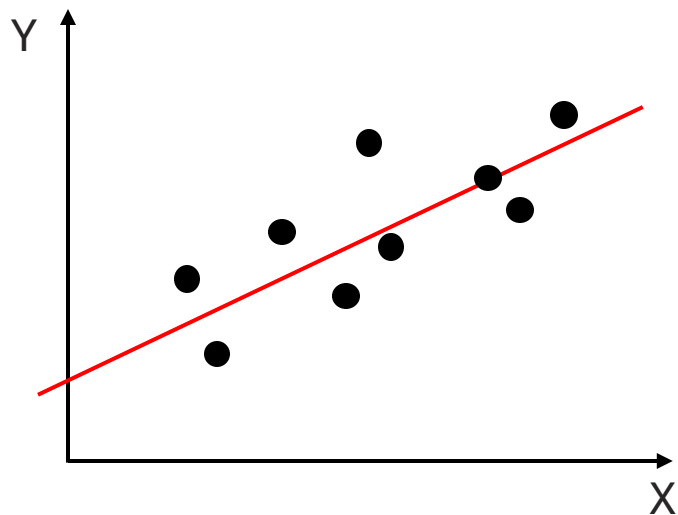
- **Hồi quy đơn biến:** Phân tích quan hệ giữa 1 biến X và 1 biến Y (ví dụ: giá ảnh hưởng đến doanh thu)
- **Hồi quy đa biến:** Phân tích quan hệ giữa nhiều biến X và 1 biến Y (ví dụ: giá, quảng cáo, thời tiết, v.v.)

Tại sao sử dụng phân tích hồi quy?

- **Dự báo:** Ước tính giá trị tương lai từ dữ liệu quá khứ
- **Phân tích nhân quả:** Hiểu mối quan hệ giữa các yếu tố với kết quả
- **Kiểm định giả thuyết:** Xác minh giả thuyết về quan hệ giữa các biến
- **Tối ưu hóa:** Xác định giá trị tối ưu của biến đầu vào

Mô Hình Hồi Quy Tuyến Tính

Mô hình hồi quy tuyến tính là mô hình dùng để dự đoán giá trị của một biến (Y) dựa trên mối quan hệ với một hoặc nhiều biến độc lập (X)



Phương trình tuyến tính cơ bản:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

1

Giải thích:

Y: biến phụ thuộc (doanh thu)

X: biến độc lập (giá, khuyến mãi,...)

β : hệ số hồi quy (mức ảnh hưởng)

ε : phần dư (sai số)

2

Đánh giá mô hình

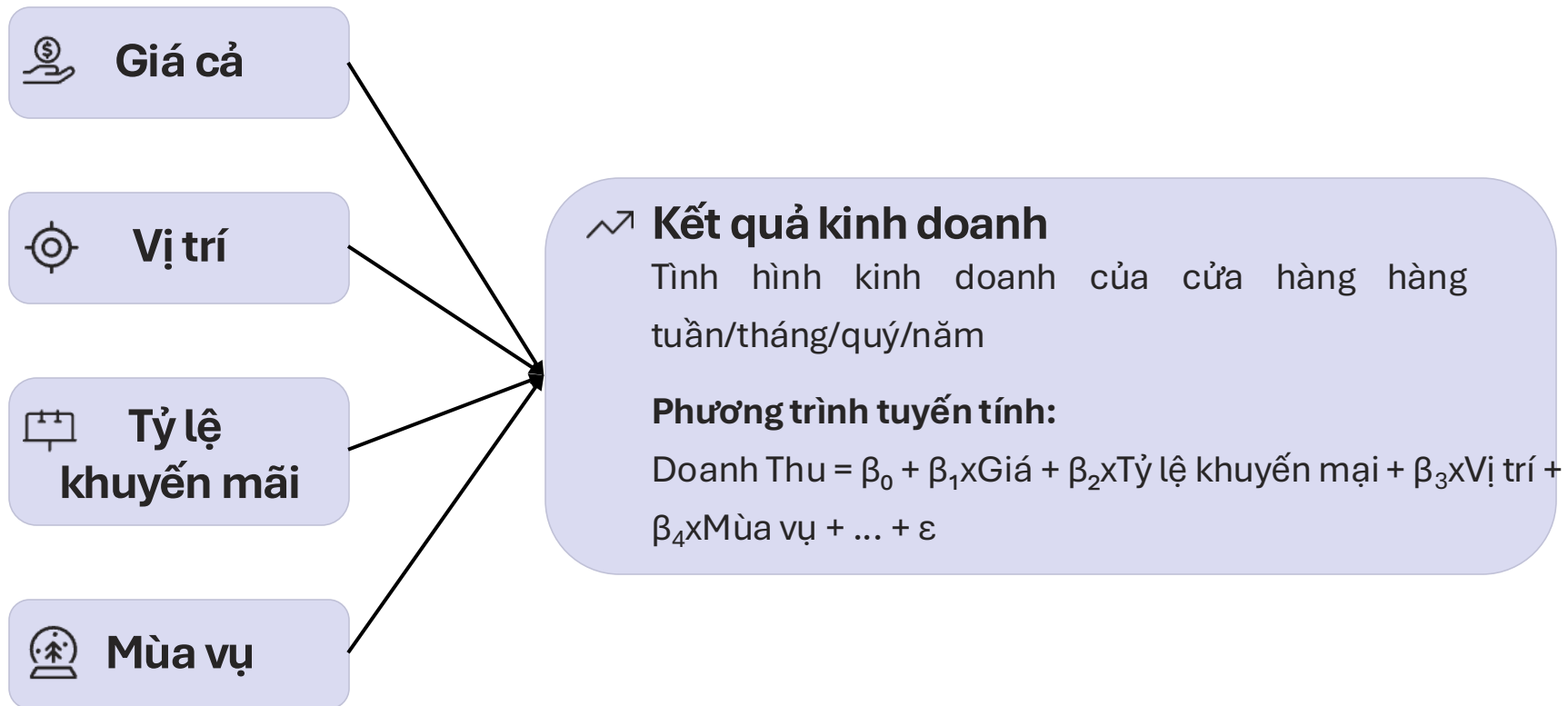
R^2 : cho biết mô hình giải thích bao nhiêu % biến động của Y

p-value: Đánh giá ý nghĩa của từng biến X

Phân tích phần dư: xem mô hình có vi phạm giả định không

Yếu Tố Ảnh Hưởng Đến Doanh Thu Cửa Hàng Bán Lẻ

Phân tích hồi quy giúp xác định mức độ ảnh hưởng của các yếu tố như giá cả, khuyến mãi, vị trí địa lý và mùa vụ đến hiệu quả kinh doanh. Giúp dự đoán doanh thu khi điều chỉnh các yếu tố như giá, khuyến mãi...



Thực Hiện Phân Tích Hồi Quy Trong Excel

Quy trình phân tích hồi quy trong Excel bao gồm 4 bước chính:



Bước 1: Bật Add-in Data Analysis Toolpak

Đảm bảo rằng bạn đã kích hoạt Data Analysis Toolpak trong Excel.



Bước 2: Chọn Regression

Trong Data Analysis, chọn Regression để bắt đầu phân tích.



Bước 3: Thiết lập vùng dữ liệu Y (doanh thu) và X (biến ảnh hưởng)

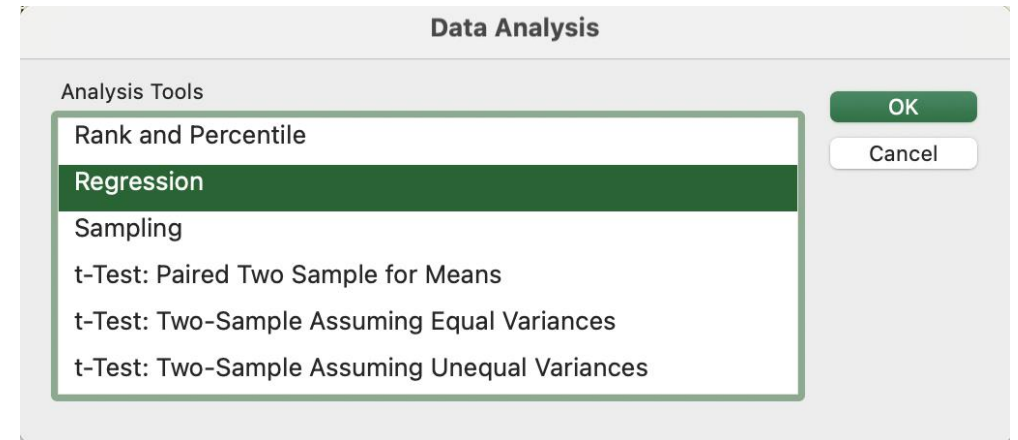
Xác định phạm vi dữ liệu cho biến phụ thuộc (Y) và biến độc lập (X).

4

Bước 4: Chạy mô hình và diễn giải kết quả

Phân tích các chỉ số quan trọng như R^2 , hệ số, p-value và sai số chuẩn để hiểu ý nghĩa thống kê và mức độ ảnh hưởng của mô hình.

Công cụ Data>Data Analysis trong Excel



Hồi quy đơn biến

Bài toán: Phân tích hồi quy xác định mức độ ảnh hưởng của chi phí đầu tư cho marketing (Marketing Spend) đến hiệu quả kinh doanh (Revenue)

Dữ liệu kinh doanh

Product_ID	Marketing_Spend	Revenue
SP001	726	42468
SP002	357	17112
SP003	304	27029
SP004	596	37162
SP005	747	32008
SP006	480	24585
SP007	982	63939
SP008	716	50748
SP009	532	24638
SP010	452	31361
SP011	408	36740
SP012	756	51311
SP013	494	34534
SP014	153	6412
SP015	458	30078
SP016	764	41215
SP017	264	24979
SP018	257	23011
SP019	578	35024
SP020	578	34504
SP021	670	37226
SP022	864	57819
SP091	734	45339
SP092	995	50182
SP093	420	34957
SP094	786	42521
SP095	633	44209
SP096	722	35365
SP097	236	16841
SP098	458	13429
SP099	316	13349
SP100	409	28757

Kết quả phân tích hồi quy đơn biến

Regression Result

SUMMARY OUTPUT

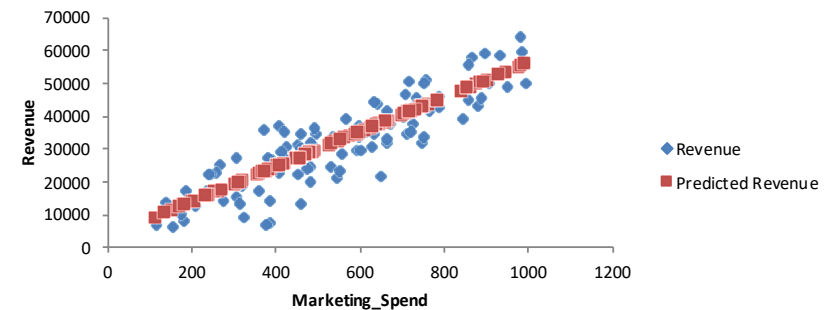
Regression Statistics	
Multiple R	0.878319475
R Square	0.771445101
Adjusted R Square	0.769112908
Standard Error	6430.57083
Observations	100

ANOVA

	df	SS	MS	F	Significance F
Regression	1	13678536012	13678536012	330.7810077	3.55488E-33
Residual	98	4052519637	41352241.2		
Total	99	17731055649			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2747.8894	1727.944011	1.590265299	0.114995928	-681.1592532	6176.938053	-681.1592532	6176.938053
Marketing_Spend	52.96027561	2.911923448	18.18738595	3.55488E-33	47.1816583	58.73889292	47.1816583	58.73889292

Marketing_Spend Line Fit Plot



Chi tiết: part6_regression.xlsx > univariate_regression

Hồi quy đơn biến – Lý giải kết quả hồi quy

Lý giải ý nghĩa của các giá trị

1. R Square = 0.77

- Khoảng 77% biến động trong doanh thu có thể được giải thích bởi biến Marketing Spend
- Đây là mức vừa phải – đầu tư vào marketing có ảnh hưởng, nhưng còn các yếu tố khác chưa được đưa vào mô hình

2. Coefficient (Marketing Spend) = 52.9

- Với mỗi 1 đơn vị tăng trong chi phí đầu tư, doanh thu tăng trung bình khoảng 52.9 đơn vị tiền tệ, giữ các yếu tố khác không đổi
- Vì P-value của Marketing Spend rất nhỏ ($\approx 3.6e-33$) → kết quả có ý nghĩa thống kê, tức là mối quan hệ này là thật, không phải do ngẫu nhiên

3. Intercept = 2747.9

- Nếu chi phí đầu tư bằng 0, mô hình dự đoán doanh thu sẽ là 2747.9 (đây là giá trị lý thuyết và không có ý nghĩa thực tế trong kinh doanh – vì giả định là chi phí đầu tư > 0)

4. Significance F = 3.55e-33

- Mô hình tổng thể có ý nghĩa thống kê → đủ tin cậy để sử dụng

Góc nhìn kinh doanh:

- Chi phí đầu tư vào Marketing có ảnh hưởng rõ ràng đến doanh thu: Tăng chi phí có thể dẫn đến doanh thu tăng, nhưng khi đầu tư quá nhiều ta phải cân nhắc đến lợi nhuận và ROI.
- Mô hình này chưa hoàn hảo, vì nó chưa xem xét các yếu tố khác như khuyến mãi, vị trí cửa hàng, gần ga tàu hay không, v.v.

Hồi quy đa biến

Bài toán: Xây dựng mô hình hồi quy đa biến có hiệu quả tốt hơn cho việc dự báo và đánh giá hiệu quả kinh doanh (Revenue)

Dữ liệu gốc

Product ID	Marketing_Spend	Discount_Ratio	Is_Near_Station	Store_Location	Day_of_Week	Store_Area	Num_Employees	Revenue
SP001	726	0.18	1	Quận 1	Thu	123	8	42468
SP002	357	0.22	1	Quận 3	Thu	198	8	17112
SP003	304	0.08	0	Quận 4	Tue	165	8	27029
SP004	596	0.08	1	Quận 2	Wed	84	9	37163
SP005	747	0.13	0	Quận 1	Fri	131	8	32008
SP006	480	0.22	1	Quận 3	Wed	152	7	24585
SP007	982	0.26	1	Quận 4	Fri	162	8	63839
SP008	716	0.19	1	Quận 4	Fri	145	8	50748
SP009	532	0.28	0	Quận 2	Fri	183	2	24638
SP010	452	0.15	1	Quận 4	Tue	170	4	31381
SP011	408	0.13	1	Quận 4	Thu	134	3	36740
SP012	756	0.14	1	Quận 4	Mon	181	5	51311
SP013	484	0.09	1	Quận 1	Thu	112	8	34534
SP014	153	0.26	1	Quận 3	Wed	61	9	6412
SP015	458	0.13	0	Quận 4	Fri	72	7	30078
SP016	764	0.19	0	Quận 3	Wed	200	4	41215
SP017	264	0.19	1	Quận 4	Wed	80	4	24979
SP018	257	0.18	1	Quận 2	Thu	184	4	23011
SP019	578	0.05	1	Quận 3	Mon	148	7	35024
SP020	578	0.30	1	Quận 4	Fri	55	9	34504
SP021	670	0.28	1	Quận 1	Mon	146	7	37226
SP022	864	0.10	1	Quận 4	Mon	182	8	57818
SP023	421	0.05	1	Quận 3	Mon	150	8	27244
SP024	139	0.19	1	Quận 3	Wed	84	9	13862
SP025	374	0.28	0	Quận 3	Tue	120	2	7184
SP026	458	0.20	1	Quận 4	Fri	144	9	34586
SP027	734	0.10	1	Quận 4	Thu	75	2	45339
SP028	995	0.23	0	Quận 2	Fri	190	8	50182
SP029	420	0.14	1	Quận 4	Fri	148	2	34957
SP030	786	0.22	0	Quận 2	Tue	135	8	42521
SP031	833	0.06	0	Quận 2	Thu	131	5	44209
SP032	722	0.21	0	Quận 4	Wed	161	3	35383
SP033	236	0.08	1	Quận 3	Fri	192	8	18841
SP034	458	0.24	0	Quận 3	Wed	91	8	13429
SP035	316	0.17	0	Quận 3	Wed	58	8	13349
SP036	409	0.08	0	Quận 4	Thu	76	6	28757

Dữ liệu đã qua tiền xử lý

Product ID	Marketing_Spend	Discount_Ratio	Is_Near_Station	Day_of_Week	Store_Area	Num_Employees	Quận 1_Quận 2	Quận 1_Quận 3	Quận 1_Quận 4	Revenue
SP001	726	0.18	1	4	123	8	0	0	0	42468
SP002	357	0.22	1	4	198	8	0	1	0	17112
SP003	304	0.08	0	2	165	8	0	0	1	27029
SP004	596	0.08	1	3	84	9	1	0	0	37163
SP005	747	0.13	0	5	131	8	0	0	0	32008
SP006	480	0.22	1	3	152	7	0	1	0	24585
SP007	982	0.26	1	5	162	8	0	0	1	63839
SP008	716	0.19	1	5	145	8	0	0	1	50748
SP009	532	0.28	0	5	183	2	1	0	0	24638
SP010	452	0.15	1	2	170	4	0	0	0	31381
SP011	408	0.13	1	4	134	3	0	0	1	36740
SP012	756	0.14	1	1	181	5	0	0	1	51311
SP013	484	0.09	1	4	112	8	0	0	0	34534
SP014	153	0.26	1	3	61	9	0	1	0	6412
SP015	458	0.13	0	5	72	7	0	0	1	30078
SP016	764	0.19	0	3	200	4	0	1	0	41215
SP017	264	0.19	1	3	80	4	0	0	1	24979
SP018	257	0.18	1	4	184	4	1	0	0	23011
SP019	578	0.05	1	1	148	7	0	1	0	35024
SP020	578	0.30	1	5	55	9	0	0	1	34504
SP021	670	0.28	1	1	146	7	0	0	0	37226
SP022	864	0.10	1	1	182	8	0	0	0	57818
SP023	421	0.05	1	3	150	8	1	0	0	27244
SP024	139	0.19	1	3	84	9	0	1	0	13862
SP025	374	0.28	0	2	120	2	0	1	0	7184
SP026	458	0.20	1	5	144	9	0	0	1	34586
SP027	734	0.10	0	4	75	2	0	0	1	45339
SP028	995	0.23	0	5	190	8	1	0	0	50182
SP029	420	0.14	1	5	148	2	0	0	1	34957
SP030	786	0.22	0	2	135	8	1	0	0	42521
SP031	833	0.06	0	4	131	5	1	0	0	44209
SP032	722	0.21	0	3	161	3	0	0	1	35383
SP033	236	0.08	1	5	192	8	0	1	0	18841
SP034	458	0.24	0	3	91	8	0	1	0	13429
SP035	316	0.17	0	3	58	8	0	0	1	13349
SP036	409	0.08	0	4	76	6	0	0	1	28757

Kết quả phân tích hồi quy đa biến

SUMMARY OUTPUT

Regression Statistics

Multiple R

0.975675

R Square

0.951943

Adjusted R Square

0.947137

Standard Error

3076.991

Observations

100

ANOVA

df

SS

MS

F

Significance F

Regression

9

1.7E+10

1875438536

198.0844024

2.5146E-55

Residual

90

8.5E+08

9467875.878

Total

99

1.8E+10

Coefficients

andard Err

t Stat

P-value

Lower 95%

Upper 95%

Lower 95.0%

Upper 95.0%

Intercept

2096.855

1877.35

1.116925033

0.266999321

-1632.8213

5826.53203

-1632.8213

5826.53203

Marketing_Spend

50.93728

1.42945

35.63429006

7.45842E-55

48.0974344

53.7771206

48.0974344

53.7771206

Discount_Ratio

-36378.91

4398.69

-8.270403963

1.12703E-12

-45117.669

-27640.154

-45117.669

-27640.154

Is_Near_Station

6974.563

632.495

11.0270685

2.15543E-18

5718.00161

8231.12399

5718.00161

8231.12399

Day_of_Week

22.99946

230.706

0.099691702

0.920810815

-435.33802

481.336946

-435.33802

481.336946

Store_Area

6.884375

7.0745

0.973125068

0.333099374

-7.170358

20.939109

-7.170358

20.939109

Num_Employees

90.35045

140.489

0.643114251

0.521785173

-188.75537

369.456265

-188.75537

369.456265

Quận 1_Quận 2

5390.506

892.08

6.042626524

3.36726E-08

3618.23317

7162.77796

3618.23317

7162.77796

Quận 1_Quận 3

-2682.171

975.191

-2.750405413

0.007195105

-4619.5587

-744.78362

-4619.5587

-744.78362

Quận 1_Quận 4

7118.014

843.126

8.442405329

4.96189E-13

5442.99608

8793.03098

5442.99608

8793.03098

Chi tiết: part6_regression.xlsx > multivariate_regression

Hồi quy đa biến – Lý giải kết quả hồi quy

Tổng quan mô hình

- **R Square = 0.9519**: Gần **95.2% biến động của doanh thu** được giải thích bởi các biến đầu vào → mô hình có tính giải thích cao
- **Adjusted R Square = 0.9471**: Sau khi điều chỉnh cho số lượng biến, mô hình vẫn rất mạnh
- **Significance F = 2.51e-55**: Mô hình tổng thể có ý nghĩa thống kê rất cao → đáng tin cậy

Góc nhìn kinh doanh

Những yếu tố doanh nghiệp nên quan tâm:

- Chi phí đầu tư vào Marketing là yếu tố tăng doanh thu rất rõ ràng
- Khuyến mãi mạnh làm giảm doanh thu, có thể do làm giảm giá trị đơn hàng
- Vị trí gần ga tàu là lợi thế rõ rệt về doanh thu
- Cửa hàng ở Quận 2 và Quận 4 hoạt động hiệu quả hơn đáng kể so với Quận 1
- Quận 3 nên xem xét lại chiến lược – doanh thu thấp hơn đáng kể

Những yếu tố có thể bỏ qua:

- Ngày trong tuần, diện tích cửa hàng, và số nhân viên không có mối liên hệ đáng kể với doanh thu → có thể là nhiễu hoặc ảnh hưởng gián tiếp

Ý nghĩa từng biến

Biến	Hệ số	P-value	Ý nghĩa thống kê?	Giải thích
Intercept	2096.86	0.267	Không	Không đáng kể – không cần quan tâm quá
Marketing_Spend	50.94	7.46E-55	Có	Mỗi đơn vị tăng giá → doanh thu tăng ~51
Discount_Ratio	-36,378.91	1.13E-12	Có	Giảm giá mạnh → doanh thu giảm nhiều
Is_Near_Station	6974.56	2.15E-18	Có	Gần ga tàu giúp tăng doanh thu gần 7,000
Day_of_Week	23	0.92	Không	Không có ảnh hưởng đáng kể (nhiều)
Store_Area	6.88	0.33	Không	Diện tích không rõ ảnh hưởng
Num_Employees	90.35	0.52	Không	Số nhân viên không có ý nghĩa
Quận 2 (so với Quận 1)	5390.51	3.37E-08	Có	Quận 2 cao hơn Quận 1 ~5390
Quận 3 (so với Quận 1)	-2682.17	0.007	Có	Quận 3 doanh thu thấp hơn Quận 1 ~2682
Quận 4 (so với Quận 1)	7118.01	4.96E-13	Có	Quận 4 vượt trội hơn Quận 1 ~7118

Cải Thiện Độ Chính Xác

Phương pháp nâng cao hiệu quả mô hình hồi quy trong Excel:

So sánh mô hình bằng R^2 hiệu chỉnh

R^2 thông thường tăng khi thêm biến mới, kể cả khi biến không có ý nghĩa. R^2 hiệu chỉnh (Adjusted R^2) khắc phục vấn đề này bằng cách tính đến số lượng biến độc lập.

- Xem giá trị "Adjusted R Square" trong kết quả phân tích
- Chọn mô hình có R^2 hiệu chỉnh cao nhất
- Cân bằng giữa độ phức tạp và độ chính xác

Kiểm tra đa cộng tuyến

Đa cộng tuyến xảy ra khi các biến độc lập có tương quan cao, làm giảm độ tin cậy của mô hình.

- Dùng Data Analysis → Correlation tạo ma trận tương quan
- Phát hiện biến có tương quan cao (>0.7)
- Áp dụng VIF qua hàm tùy chỉnh
- Loại bỏ hoặc kết hợp biến tương quan cao

Biến đổi dữ liệu trong Excel

Biến đổi dữ liệu cải thiện độ chính xác bằng cách làm cho mối quan hệ giữa các biến tuyến tính hơn.

- Logarithm: `=LN(cell)` hoặc `=LOG10(cell)`
- Căn bậc hai: `=SQRT(cell)`
- Nghịch đảo: `=1/cell`
- Kiểm tra dư bằng biểu đồ phân tán sau biến đổi

Xử Lý Ngoại Lệ & Đa Cộng Tuyến

1 Xác định điểm ngoại lệ (outliers)

Sử dụng biểu đồ phần dư để phát hiện các điểm dữ liệu nằm xa khỏi phần còn lại, có thể ảnh hưởng đến mô hình.

3 Xem xét loại bỏ hoặc biến đổi biến

Nếu phát hiện biến gây nhiễu, hãy cân nhắc loại bỏ hoặc biến đổi biến để cải thiện mô hình.

2 Phát hiện mô hình sai lệch

Biểu đồ phần dư có thể cho thấy các mẫu hình như quan hệ phi tuyến, phương sai không đồng nhất, hoặc sự phụ thuộc giữa các phần dư, gợi ý rằng mô hình không phù hợp.

4 Thêm biến tương tác hoặc dùng mô hình nâng cao hơn

Ngoài việc loại bỏ biến, bạn cũng có thể thử thêm biến tương tác hoặc sử dụng các mô hình hồi quy nâng cao hơn như phân tích thành phần chính (PCA).

Phần 7: Tối ưu hoá

Tại sao cần Tối Ưu Hóa?

Tối ưu hóa là quá trình tìm kiếm giải pháp tốt nhất cho một vấn đề trong điều kiện nhất định, giúp đạt được hiệu quả cao nhất với nguồn lực có hạn.



Tại sao cần tối ưu hóa?

Tối ưu hóa giúp doanh nghiệp tăng hiệu quả sử dụng nguồn lực, giảm chi phí sản xuất, và cải thiện chất lượng quyết định dựa trên dữ liệu.



Tối ưu hóa trong Excel

Excel cung cấp các công cụ hỗ trợ tối ưu như Solver, Goal Seek và Scenario Manager để giải quyết các bài toán tối ưu hóa phức tạp.



Ứng dụng trong thực tế

Tối ưu phân bổ nguồn lực, tối ưu giá bán, lên kế hoạch sản xuất, dự báo nhu cầu và tối ưu hóa chuỗi cung ứng.



Lợi ích chính

Tìm được giải pháp tối ưu trong thời gian ngắn, xử lý được các bài toán phức tạp với nhiều ràng buộc, và mô phỏng các kịch bản khác nhau.

Hiểu Ý Nghĩa Tối Ưu Hóa



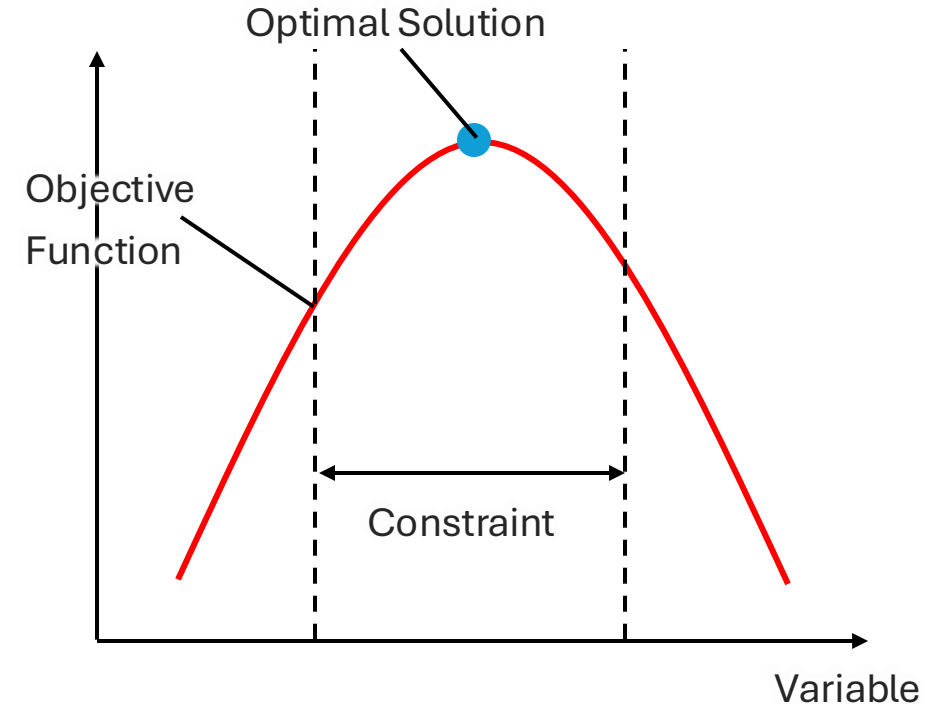
Bản chất của tối ưu hóa

- Tối ưu hóa là cách tìm giải pháp tốt nhất cho một vấn đề trong những điều kiện nhất định.
- Trong kinh doanh, tối ưu hóa thường là việc tối đa lợi nhuận cao nhất, giảm chi phí thấp nhất, hoặc cân bằng giữa nhiều mục tiêu khác nhau.



Tư duy tối ưu hóa

- Tư duy tối ưu hóa cần khả năng nhìn vấn đề dưới góc nhìn toán học.
- Xác định mục tiêu cần đạt (hàm mục tiêu cần tối ưu – objective function), các yếu tố có thể thay đổi (biến quyết định – variable), và những giới hạn phải tuân theo (constraint).



Tối ưu (Tối đa hoá/Tối thiểu hoá) chỉ số gì?

【Objective Function】
Hàm Mục Tiêu

Có thể thay đổi biến số gì?

【Variable】
Biến quyết định

Điều kiện bắt buộc phải tuân theo là gì?

【Constraint】
Ràng buộc

Trình Tự Tối Ưu Hoá Trên Excel

Excel cung cấp công cụ mạnh mẽ để giải quyết các bài toán tối ưu hóa trong kinh doanh và phân tích dữ liệu.

1

Định nghĩa mô hình tối ưu

Sử dụng công cụ Solver, xác định ô mục tiêu (tối đa hóa/tối thiểu hóa), các ô biến quyết định và các ràng buộc kèm theo.

2

Thiết lập trong Solver

Chọn "Set Objective" (ô mục tiêu), "To" (Max/Min/Value), "By Changing Variable Cells" (các ô biến đổi) và "Subject to the Constraints" (các ràng buộc cần tuân thủ).

3

Chọn phương pháp giải

Lựa chọn phương pháp phù hợp: GRG Nonlinear (cho bài toán không tuyến tính), Simplex LP (cho bài toán tuyến tính) hoặc Evolutionary (cho bài toán phức tạp).

4

Phân tích kết quả

Xem xét các báo cáo Answer, Sensitivity và Limits để hiểu sâu về giải pháp tối ưu và tác động của các thông số.

i Công cụ Solver trong Excel giúp giải quyết nhiều bài toán thực tế như: phân bổ ngân sách, tối ưu danh mục đầu tư, lập kế hoạch sản xuất, và tối ưu chuỗi cung ứng.

Tối Ưu Giá Gói Cước Điện Thoại – Đặt Vấn Đề

Tình huống: Bạn là PIC (Person In Charge) đang làm việc trong một công ty viễn thông. Nhiệm vụ của bạn là thiết lập mức giá tối ưu cho 3 gói cước điện thoại công ty đang cân nhắc sao đạt được lợi nhuận tốt nhất

1

Vấn đề đặt ra

- Bạn là PIC của công ty viễn thông, nhiệm vụ của bạn là phải thiết lập giá 3 gói data thế nào để kiếm được lợi nhuận nhiều nhất cho công ty
- 3 gói công ty bạn đang cân nhắc:
 - Gói Cơ Bản – Basic Plan (Giới hạn sử dụng 10GB)
 - Gói Nâng Cao – Advanced Plan (20GB)
 - Gói Cước Không Giới Hạn Dung Lượng

2

Dữ liệu có

- Công ty bạn đã thực hiện cuộc điều tra về nhu cầu khách hàng và có được kết quả sau đây:
 - Lượng data của khách sử dụng hàng tháng
 - Số tiền khách hàng sẵn sàng bỏ ra (Willingness-To-Pay WTP) để sử dụng gói cước đó
 - Nhóm hành vi – Segment (Để dễ phân loại khách hàng)
- File csv 500 khách hàng

3

Tối ưu doanh thu bằng solver trong Excel

- Solver là công cụ mạnh mẽ trong Excel giúp giải quyết các bài toán tối ưu hóa phức tạp.
- Để sử dụng Solver, bạn cần xác định:
 - Objective: Ô mục tiêu cần để tối ưu (Doanh thu)
 - Variable: Giá cước cho 3 gói cước
 - Constraint: Giá cước phải tuân theo quy chuẩn được công ty đề ra (Nằm trong giới hạn cho phép)
- Solver: GRG Nonlinear

Tối Ưu Giá Gói Cước Điện Thoại – Dữ Liệu Hiện Trạng

Biến mục tiêu cần tối ưu là tổng doanh thu (Total Revenue). Đây là KPI chính để đo lường hiệu quả của giải pháp giá được đề xuất. Biến quyết định là giá 3 gói cước Basic, Advanced và Unlimited.

Dữ liệu 500 khách hàng hiện có

CustomerID	Monthly_Data_GB	Segment	Max_Willingness_to_Pay_VN
KH_001	1.35	Low	114000
KH_002	4.06	Low	97000
KH_003	20	High	203000
KH_004	22.38	High	210000
KH_005	10.86	Moderate	131000
KH_006	14.28	Moderate	182000
KH_007	15.62	High	161000
KH_008	24.3	High	221000
KH_009	5.2	Low	112000
KH_010	3.36	Low	73000
KH_011	8.33	Moderate	153000
KH_012	9.53	Moderate	144000
KH_013	7.58	Moderate	128000
KH_014	3.99	Low	118000
KH_015	11.55	Moderate	130000
KH_485	9.02	Moderate	154000
KH_486	32.58	High	297000
KH_487	107.72	Heavy	607000
KH_488	11.27	Moderate	162000
KH_489	4.95	Low	98000
KH_490	16.01	High	186000
KH_491	2.94	Low	118000
KH_492	14.87	Moderate	158000
KH_493	5.97	Low	140000
KH_494	10.55	Moderate	140000
KH_495	86.44	Heavy	531000
KH_496	33.12	High	286000
KH_497	3.29	Low	88000
KH_498	11.65	Moderate	145000
KH_499	9.37	Moderate	150000
KH_500	0.33	Low	55000

Giá gói cước hiện tại công ty bạn đang sử dụng

Parameter	Giá gốc	Dung lượng tối đa (GB)
Basic Plan Price (VND)	125,000	10
Advanced Price (VND)	225,000	20
Unlimited Plan Price (VND)	600,000	1,000,000,000
Overage fee per GB	9,000	

Biến quyết định

Doanh Thu Hiện Tại

Chỉ số KPI	Giá trị	Đơn vị
Tổng doanh thu (Total Revenue)	61,656,710	VND
Chỉ số khách hàng rời đi (Churn Rate)	14.60%	%

Hàm mục tiêu

- Với giá gốc hiện nay của 3 gói dịch vụ, doanh thu hàng tháng của 500 khách hàng bạn quản lý là khoảng 62 triệu VND
- Tỷ lệ khách hàng rời đi Churn Rate là 14.6% (73 khách không sử dụng dịch vụ của công ty bạn cung cấp do giá gói cước vượt quá số tiền khách hàng sẵn sàng chi cho cước viễn thông)

Tối Ưu Giá Gói Cước Điện Thoại – Tính Toán Chi Phí

Trước khi tối ưu ta phải tính toán chi phí cho từng gói cước và chọn plan tương ứng phù hợp với chi phí của từng gói. Nếu chi phí này vượt quá WTP khách hàng sẽ không sử dụng dịch vụ.

Dữ liệu 500 khách hàng sau khi tính giá cho từng gói cước

CustomerID	Usage_GB	WTP_VND	Cost_Basic	Cost_Advanced	Cost_Unlimited	MinCost	ChosenPlan	ChurnFlag	Revenue
KH_001	1.35	164,000	110,000	165,239	534,281	110,000	Basic	0	110,000
KH_002	4.06	147,000	110,000	165,239	534,281	109999.94	Basic	0	0
KH_003	20	253,000	200,000	165,239	534,281	165238.72	Advanced	0	165,239
KH_004	22.38	260,000	221,420	186,659	534,281	186658.72	Advanced	0	186,659
KH_005	10.86	181,000	117,740	165,239	534,281	117739.94	Basic	0	117,740
KH_006	14.28	232,000	148,520	165,239	534,281	148519.94	Basic	0	148,520
KH_007	15.62	211,000	160,580	165,239	534,281	160579.94	Basic	0	160,580
KH_008	24.3	271,000	238,700	203,939	534,281	203938.72	Advanced	0	203,939
KH_009	5.2	162,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_010	3.36	123,000	110,000	165,239	534,281	109999.94	Basic	0	0
KH_011	8.33	203,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_012	9.53	194,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_013	7.58	178,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_014	3.99	168,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_015	11.55	180,000	123,950	165,239	534,281	123949.94	Basic	0	123,950
KH_016	8.52	189,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_017	6.27	154,000	110,000	165,239	534,281	109999.94	Basic	0	0
KH_018	6.81	154,000	110,000	165,239	534,281	109999.94	Basic	0	0
KH_474	3.84	163,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_475	13.42	217,000	140,780	165,239	534,281	140779.94	Basic	0	140,780
KH_476	2.81	154,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_477	11.33	219,000	121,970	165,239	534,281	121969.94	Basic	0	121,970
KH_478	3.01	159,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_479	1.87	133,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_480	4.06	154,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_481	7.01	164,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_482	9.17	171,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_483	10.89	208,000	118,010	165,239	534,281	118009.94	Basic	0	118,010
KH_484	1.29	145,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_485	9.02	204,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_486	32.58	347,000	313,220	278,459	534,281	278458.72	Advanced	0	278,459
KH_487	107.72	657,000	989,480	954,719	534,281	534281.11	Unlimited	0	534,281
KH_488	11.27	212,000	121,430	165,239	534,281	121429.94	Basic	0	121,430
KH_489	4.95	148,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_490	16.01	236,000	164,090	165,239	534,281	164089.94	Basic	0	164,090
KH_491	2.94	168,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_492	14.87	208,000	153,830	165,239	534,281	153829.94	Basic	0	153,830
KH_493	5.97	190,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_494	10.55	190,000	114,950	165,239	534,281	114949.94	Basic	0	114,950
KH_495	86.44	581,000	797,960	763,199	534,281	534281.11	Unlimited	0	534,281
KH_496	33.12	336,000	318,080	283,319	534,281	283318.72	Advanced	0	283,319
KH_497	3.29	138,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_498	11.65	195,000	124,850	165,239	534,281	124849.94	Basic	0	124,850
KH_499	9.37	200,000	110,000	165,239	534,281	109999.94	Basic	0	110,000
KH_500	0.33	105,000	110,000	165,239	534,281	109999.94	Basic	1	110,000

Tính toán chi phí cho từng gói cước sử dụng

$Cost_i = \text{Giá gói}_i + \text{Phí vượt quá hạn mức} \times \text{Dung lượng vượt hạn mức}$

- $i: 1$ trong 3 gói cước

Ví dụ:

- Khách sử dụng hàng tháng 18GB
- WTP (Chi phí sẵn sàng chi trả hàng tháng) : 250,000 VND
- Trường hợp sử dụng gói Cơ Bản:
 - Dung lượng tối đa 10GB, giá gốc 125,000VND
 - $Cost_Basic = 125,000 + 9,000 \times (18-10) = 197,000$

Các cột tiếp theo:

- MinCost** = MIN(range) Tìm chi phí thấp nhất trong 3 gói
- ChosenPlan**: Gói tương ứng với lượng sử dụng dữ liệu của khách hàng
- Revenue**: Nếu chi phí \leq WTP thì khách sẽ sử dụng dịch vụ **ChurnFlag=0**, ngược lại khách không sử dụng dịch vụ và **ChurnFlag = 1**

Chi tiết tính toán: part7_optimization.xlsx > NonOptimizedSolution

Tối Ưu Giá Gói Cước Điện Thoại – Tối Ưu Giá

Solver là công cụ tối ưu hóa có sẵn trong Excel, dùng để tìm giá trị tốt nhất của hàm mục tiêu bằng cách thay đổi các biến trong phạm vi giới hạn

Solver Parameters

Set Objective: **\$M\$3**

To: **Max** (Hàm mục tiêu)

By Changing Variable Cells: **\$M\$7:\$M\$9** (Biến thay đổi)

Subject to the Constraints:

- \$M\$7 <= \$S\$4
- \$M\$7 >= \$R\$4
- \$M\$8 <= \$S\$5
- \$M\$8 >= \$R\$5
- \$M\$9 <= \$S\$6
- \$M\$9 >= \$R\$6

Ràng buộc

☒ Make Unconstrained Variables Non-Negative

Select a Solving Method: **GRG Nonlinear**

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Close Solve

Chỉ số KPI	Giá trị	Đơn vị	Notes
Tổng doanh thu (Total Revenue)	75,095,421	VND	← Tối đa doanh thu
Chỉ số khách hàng rời đi (Churn Rate)	5.00%	%	

Parameter	Giá tối ưu	Dung lượng tối đa (GB)	Notes
Basic Plan Price (VND)	110,000	10	• Biến cần phải tối ưu
Advanced Price (VND)	165,239	20	• Biến cần phải tối ưu
Unlimited Plan Price (VND)	534,281	1,000,000,000	• Biến cần phải tối ưu • Không có giới hạn nên có thể đặt là 1 số rất lớn gần với ∞
Overage fee per GB	9,000		

Thông tin cơ bản của gói cước sử dụng (Giới hạn)		
	Giới hạn dưới (VND)	Giới hạn trên (VND)
GB_limit_Basic	80,000	150,000
GB_limit_Advanced	150,000	500,000
GB_limit_Unlimited	500,000	1,000,000

※ Giới hạn giá này phản ánh mức giá thị trường chấp nhận được và được quyết định bởi công ty trước

Tối Ưu Giá Gói Cước Điện Thoại – So Sánh Kết Quả

Kết quả tối ưu cho thấy giá gói cước hiện tại đang được thiết lập cao hơn giá trị lý tưởng. Sau khi tối ưu giá tổng doanh thu tăng từ khoảng 62 triệu VND/1 tháng lên khoảng 75 triệu VND/1 tháng (Tăng ~13 triệu VND).

	Trước khi tối ưu	Sau khi tối ưu		
Chỉ số KPI	Giá trị (VND)	Giá trị (VND)	Chênh lệch	Đơn vị
Tổng doanh thu (Total Revenue)	61,656,710	75,095,421	13,438,711	VND
Chỉ số khách hàng rời đi (Churn Rate)	14.60%	5.00%	-9.60%	%
Parameter	Giá gốc	Giá tối ưu	Chênh lệch	Dung lượng tối đa (GB)
Basic Plan Price (VND)	125,000	110,000	-15,000	10
Advanced Price (VND)	225,000	165,239	-59,761	20
Unlimited Plan Price (VND)	600,000	534,281	-65,719	1,000,000,000

※ Giá trị rất lớn – không giới hạn

Kết Quả:

- ✓ Tăng tổng doanh thu từ khoảng 62 triệu VND lên 75 triệu VND / tháng (Tăng ~13 triệu VND)
- ✓ Tỷ lệ khách hàng rời giảm mạnh từ 14.6% còn 5%
- ✓ Việc điều chỉnh giá gói cước không chỉ tăng doanh thu mà còn giữ chân khách hàng hiệu quả hơn – từ đó giúp doanh nghiệp ổn định dài hạn

Tối Ưu Dưới Ràng Buộc

Trong quá trình tối ưu hóa với Excel, việc hiểu và thiết lập ràng buộc đúng cách cùng với phương pháp giải phù hợp sẽ quyết định hiệu quả của giải pháp.

Thêm ràng buộc thực tế

Trong các bài toán tối ưu hóa thực tế, luôn có các ràng buộc cần xem xét. Ví dụ, khi tối ưu hóa giá, bạn có thể có ràng buộc về giá sàn (giá không được thấp hơn chi phí sản xuất), giá trần (giá không được cao hơn mức khách hàng chấp nhận), hoặc giới hạn về hàng tồn kho.

Solver cho phép bạn thêm nhiều loại ràng buộc khác nhau, bao gồm ràng buộc về giá trị (\leq , $=$, \geq), ràng buộc về tính nguyên (biến phải là số nguyên), và ràng buộc về tính nhị phân (biến chỉ có thể là 0 hoặc 1).

Các loại solver trong Excel

Excel cung cấp ba phương pháp giải khác nhau trong Solver:

- GRG Nonlinear: phù hợp cho các bài toán tối ưu hóa phi tuyến mượt mà.
- Simplex LP: hiệu quả cho các bài toán quy hoạch tuyến tính.
- Evolutionary: hữu ích cho các bài toán phức tạp, không mượt mà hoặc không liên tục.

Việc lựa chọn phương pháp giải phù hợp phụ thuộc vào đặc điểm của bài toán tối ưu hóa của bạn.

Tổng Kết

Hành trình học phân tích dữ liệu với Excel đã hoàn thành. Dưới đây là những kiến thức chính:

- 1 Hiểu Tổng Quan Về Phân Tích Dữ Liệu**
Khái niệm cơ bản, vai trò trong quyết định dựa trên bằng chứng, cách đặt câu hỏi phù hợp và lý do chọn Excel làm công cụ phân tích.
- 2 Hiểu Xu Hướng Dữ Liệu Qua Thống Kê Cơ Bản**
Sử dụng hàm thống kê, Data Analysis Tool và Pivot Table để phân tích xu hướng dữ liệu hiệu quả.
- 3 Trực Quan Hóa Dữ Liệu**
Các loại biểu đồ (cột, đường, tròn, histogram, heatmap, scatter plot) và cách truyền tải thông điệp dữ liệu.
- 4 Kiểm Định Giả Thuyết**
Thiết lập và kiểm định giả thuyết thống kê, hiểu về p-value, t-value và ứng dụng trong Excel.
- 5 Tiền Xử Lý Dữ Liệu**
Phương pháp xử lý dữ liệu thiếu, giá trị ngoại lệ và tạo biến giả để chuẩn bị dữ liệu.
- 6 Sử Dụng Các Mô Hình Hồi Quy**
Xây dựng và phân tích mô hình hồi quy đơn biến và đa biến để xác định yếu tố ảnh hưởng.
- 7 Tối Ưu Hoá**
Ý nghĩa, quy trình và ứng dụng tối ưu hóa vào bài toán thực tế như tối ưu giá gói cước điện thoại

Tài liệu tham khảo

1. Báo cáo The Future of Jobs: <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>
2. Sách Storytelling with Data: A Data Visualization Guide for Business Professionals, Cole Nussbaumer Knaflitz
3. Sách tiếng Nhật: 統計学の基礎から学ぶExcelデータ分析の全知識、三好大悟（著）/ 榎田洋資（監修）
4. Các hàm trong Google Sheet: <https://support.google.com/docs/table/25273>
5. Keyboard Shortcuts trong Google Sheet:
<https://support.google.com/docs/answer/181110?sjid=3200468911459127971-NC>
6. Coursera, Data Analytics Foundations – Module 1