

Module 3 Tuần 1 - Excel for Data Analysis phần 1

Time-Series Team

Ngày 2 tháng 8 năm 2025

Buổi học thứ 7 (ngày 2/8/2025) được chia thành 3 phần chính nhằm giúp ta hiểu cách sử dụng Excel cho phân tích dữ liệu cơ bản, thống kê và trực quan hóa:

- **Phần 1: Hiểu Tổng Quan Về Phân Tích Dữ Liệu** – Khái niệm dữ liệu, quy trình và ứng dụng Excel.
- **Phần 2: Hiểu Xu Hướng Dữ Liệu Qua Thống Kê Cơ Bản** – Phương pháp thống kê và Pivot Table để phát hiện xu hướng.
- **Phần 3: Trực Quan Hóa Dữ Liệu** – Tạo biểu đồ, dashboard với tips hiệu quả.

Phần 1: Hiểu Tổng Quan Về Phân Tích Dữ Liệu

1.1 Định nghĩa Dữ liệu và Thông tin

Dữ liệu (Data) là các sự kiện thô, chưa qua xử lý, thường ở dạng số, ký tự, hình ảnh và chưa mang nhiều ý nghĩa. Thông tin (Information) là dữ liệu đã được xử lý và có ngữ cảnh rõ ràng, hỗ trợ ra quyết định.

- **Dữ liệu:** Thô, ví dụ: 15.000.000, 42%, "TPHCM".
- **Thông tin:** Ví dụ: "Doanh thu tháng 6 là 15 triệu đồng, tăng 42% so với TPHCM".

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

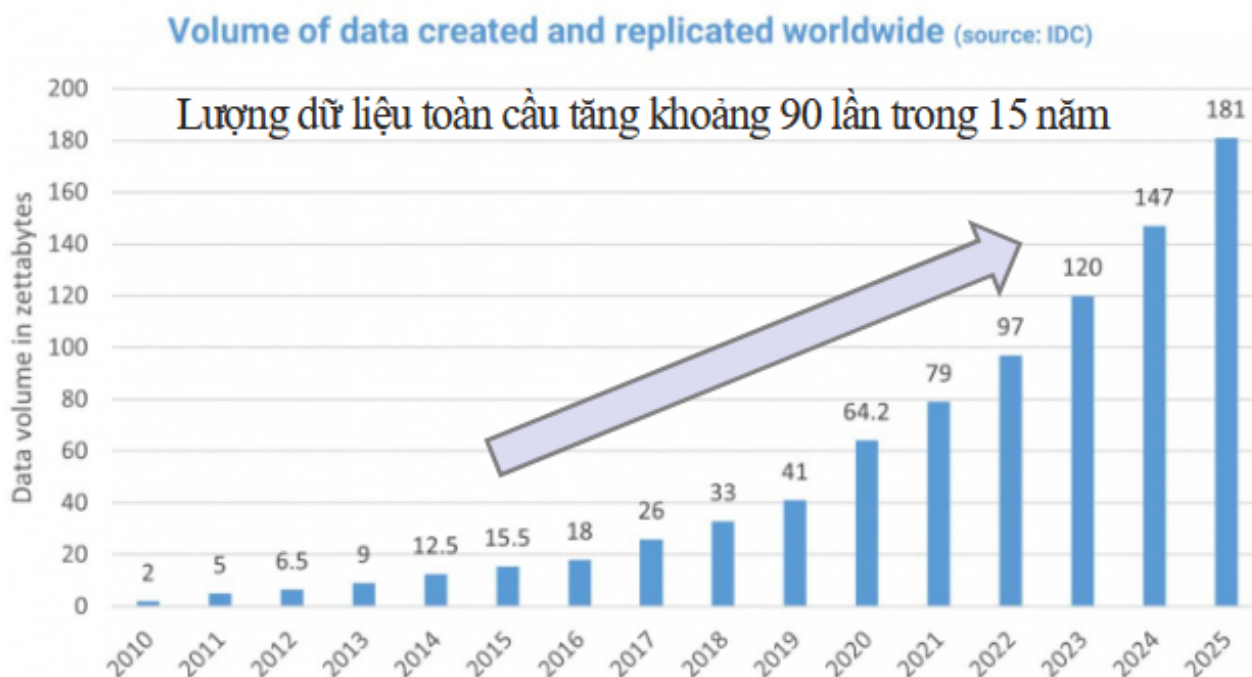
Thông tin (Information)

- 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

Hình 1: Data vs Information

1.2 Sự bùng nổ dữ liệu toàn cầu

Khối lượng dữ liệu toàn cầu đã tăng mạnh trong hơn một thập kỷ qua, từ 2 zettabyte năm 2010 lên tới hơn 180 zettabyte vào năm 2025 (dự kiến). Xu thế này cho thấy việc biết cách phân tích dữ liệu trở nên quan trọng để tìm ra giá trị từ khối lượng thông tin khổng lồ đó.



Hình 2: Data Explosion

1.3 Quy trình phân tích dữ liệu

1. **Xác định vấn đề:** Bắt đầu với câu hỏi đúng, xác định chỉ số kinh doanh cần phân tích. Ví dụ: *Shopee muốn biết vì sao tỷ lệ hủy đơn tăng 12% trong tháng 11 (tháng sale 11.11).*
2. **Thu thập dữ liệu:** Lấy dữ liệu từ cơ sở dữ liệu, file, API, khảo sát. Ví dụ: *Lấy dữ liệu đơn hàng từ hệ thống Shopee, dữ liệu phản hồi từ khách hàng, và dữ liệu vận chuyển từ đối tác giao hàng.*
3. **Làm sạch dữ liệu:** Xử lý dữ liệu thiếu, loại bỏ trùng lặp, ngoại lai. Ví dụ: *Loại bỏ các đơn hàng test của shop, xử lý các bản ghi thiếu thông tin địa chỉ.*
4. **Khám phá dữ liệu:** Thống kê mô tả, trực quan hoá để hiểu cấu trúc. Ví dụ: *Vẽ biểu đồ tỷ lệ hủy theo tỉnh/thành phố để phát hiện TP.HCM và Hà Nội có tỷ lệ cao bất thường.*
5. **Phân tích & mô hình hoá:** Áp dụng thống kê, học máy để tìm mối quan hệ. Ví dụ: *Xây dựng mô hình để phân tích mối quan hệ giữa thời gian giao hàng và khả năng khách hủy đơn.*
6. **Diễn giải & hành động:** Chuyển kết quả thành insight, hỗ trợ quyết định. Ví dụ: *Đề xuất Shopee hợp tác thêm đối tác vận chuyển trong giai đoạn flash sale để giảm tỷ lệ hủy do giao chậm.*

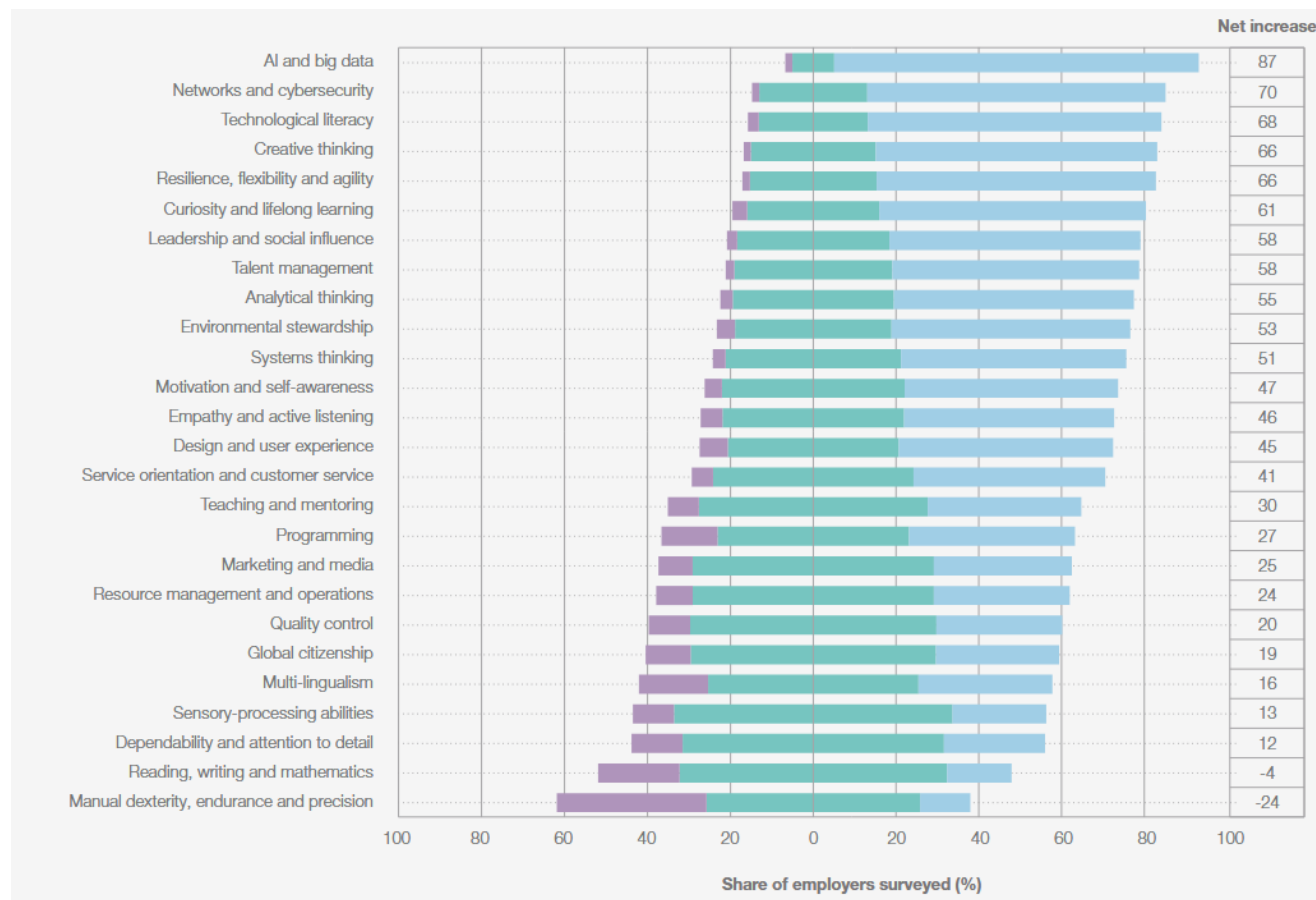
Trong bài ngày hôm nay, chúng ta sẽ đi từ tổng quan tới từng truy trình để phân tích dữ liệu với ví dụ cụ thể (i.e. use case) và vì sao mình nên dùng Excel cho công việc có dữ liệu < 1 triệu mẫu thay vì dùng Code để phân tích mất nhiều thời gian.

1.4 Lợi ích của phân tích dữ liệu

Phân tích dữ liệu giúp các quyết định bớt phụ thuộc vào cảm tính. Nó cho phép phát hiện sớm xu hướng, dự báo nhu cầu để quản lý tồn kho, tối ưu quy trình để giảm lãng phí, và cá nhân hóa trải nghiệm khách hàng. Những lợi ích này đều góp phần cải thiện hiệu quả hoạt động và tăng giá trị cho tổ chức.

1.5 Lý do nên học phân tích dữ liệu




Kỹ năng phân tích dữ liệu mở ra nhiều cơ hội nghề nghiệp như Data Analyst, Data Scientist hay Data Engineer. Ngoài ra, khả năng tư duy phân tích được đánh giá cao trong nhiều lĩnh vực và thường được nhắc đến trong các báo cáo về xu hướng việc làm toàn cầu.



Hình 3: Top những ngành lên trend nhiều nhất theo %

1.6 Các vai trò phổ biến trong lĩnh vực dữ liệu

Trong ngành này có 3 vai trò chính. Data Scientist thường tập trung vào mô hình nâng cao và học máy. Data Analyst hoặc BI Analyst chú trọng việc đặt câu hỏi kinh doanh, phân tích số liệu và xây dựng báo cáo. Data Engineer chịu trách nhiệm xây dựng hạ tầng dữ liệu và pipeline. Mỗi vai trò đều cần kỹ năng kỹ thuật kết hợp với hiểu biết về nghiệp vụ.

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 liệu Có kỹ năng xây dựng mô hình và tối ưu mô hình 	<ul style="list-style-type: none"> Excel, Python, SQL, R, Spark Cloud (AWS, GCP, Azure, Databricks)
Data Analyst BI Analyst 	<ul style="list-style-type: none"> Phân tích dữ liệu để tạo ra các báo cáo Cung 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 Tableau Kỹ năng mềm tốt Am hiểu lĩnh vực kinh doanh 	<ul style="list-style-type: none"> Excel, SQL, Tableau, Power BI, Python Power Point, Presentation Tools
Data Engineer 	<ul style="list-style-type: none"> Xây dựng pipeline dữ liệu Xử 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 lớn Khả năng xây dựng kiến trúc dữ liệu và đảm bảo dữ liệu được ổn định 	<ul style="list-style-type: none"> Python, SQL, Spark, Airflow Cloud (AWS, GCP, Azure, Databricks)

Hình 4: Một số công việc liên quan đến dữ liệu

1.7 Ai nên học phân tích dữ liệu

Hầu hết mọi người đều có thể hưởng lợi từ kỹ năng này. Nhân viên văn phòng có thể tự động hóa công việc. Người làm doanh nghiệp vừa hoặc nhỏ có thể theo dõi chi phí và khách hàng. Sinh viên có thể thích ứng tốt hơn với môi trường khi biết cách tự phân tích dữ liệu, và có thể phát triển góc nhìn trực quan hơn về thị trường với kiến thức phân tích dữ liệu.

1.8 Ứng dụng trong thực tế

Phân tích dữ liệu được áp dụng ở nhiều lĩnh vực. Netflix và Spotify dùng để gợi ý nội dung. Trong thể thao, dữ liệu được dùng để theo dõi hiệu suất. Amazon và Walmart ứng dụng trong thương mại và quản lý chuỗi cung ứng. Trong y tế và giáo dục, phân tích dữ liệu hỗ trợ dự báo và theo dõi tiến trình.

1.9 5 cấp độ của phân tích dữ liệu

Có năm cấp độ thường được nhắc đến: mô tả (descriptive), quan hệ (relational), nhân quả (causal), dự báo (predictive) và tối ưu hóa (optimization). Mỗi cấp độ giúp trả lời những loại câu hỏi khác nhau, từ việc mô tả dữ liệu quá khứ đến dự đoán tương lai và tìm cách cải thiện hiệu suất.

1. **Descriptive:** Hiểu quá khứ và hiện tại.
2. **Relational:** Phát hiện mối quan hệ giữa hiện tượng.
3. **Causal:** Tìm quan hệ nhân quả.
4. **Predictive:** Dự đoán xu hướng tương lai.
5. **Optimization:** Tối ưu hoá kết quả.

1.10 Lưu ý khi phân tích dữ liệu

Trong phân tích dữ liệu, Ta ĐẶT BIỆT cần PHÂN BIỆT rõ giữa **SỰ THẬT** (fact) và **SUY LUẬN** (inference). Việc phân tích nên bắt đầu từ những câu hỏi cụ thể và có mục tiêu rõ ràng. Điều này giúp tiết kiệm thời gian và tránh bị lệch hướng vì sự thật chỉ đúng khi dữ liệu theo sau.

- **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”.

Hình 5: Phân biệt Fact và Inference

1.11 Framework 5W2H trong phân tích dữ liệu

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

2.1 Tại sao cần hiểu xu hướng dữ liệu?

Phát hiện xu hướng trong dữ liệu giúp doanh nghiệp dự báo, nhận diện bất thường và tối ưu hiệu suất. Ví dụ, hai cửa hàng điện tử có cùng doanh thu trung bình, nhưng một cửa hàng có doanh thu ổn định, cửa hàng còn lại dao động mạnh. Nếu chỉ nhìn vào giá trị trung bình, ta dễ bỏ sót sự khác biệt quan trọng này.

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	Cửa hàng B
<ul style="list-style-type: none">• 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%) <p>⇒ 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%</p>	<ul style="list-style-type: none">• 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) <p>⇒ 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</p>

Hình 6: Case Study: So sánh xu hướng dữ liệu

2.2 Không chỉ dựa vào giá trị trung bình

Giá trị trung bình (Mean) là thước đo phổ biến để mô tả dữ liệu, nhưng có thể gây hiểu lầm nếu dữ liệu chứa ngoại lệ (outlier).

- **Mean (Trung bình):** Tính bằng tổng các giá trị chia cho số lượng phần tử. Phù hợp khi dữ liệu phân bố đều và ít outlier.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- **Median (Trung vị):** Là giá trị nằm giữa khi sắp xếp dữ liệu theo thứ tự. Ít bị ảnh hưởng bởi outlier, nên nó phản ánh tốt hơn vị trí trung tâm khi dữ liệu lệch hoặc có giá trị cực đoan.

$$Median = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ lẻ} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & n \text{ chẵn} \end{cases}$$

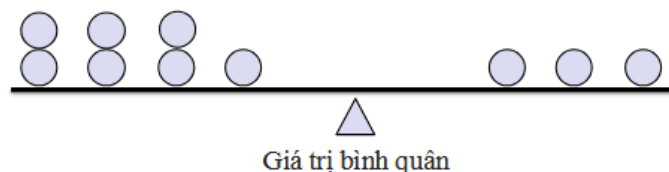
Ví dụ: Thu nhập của 5 nhân viên lần lượt là {10, 12, 14, 15, 100} triệu.

- Mean = (10+12+14+15+100)/5 = 30.2 triệu → bị kéo lên bởi giá trị 100 triệu.
- Median = 14 triệu → phản ánh mức điển hình hơn cho phần lớn nhân viên.

🧮 Công thức toán học

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

🖼️ Hình ảnh tương tượng



Hình 7: So sánh Mean và Median

2.3 Đo lường mức độ phân tán

Ngoài giá trị trung tâm, cần quan tâm đến mức độ dao động của dữ liệu. Hai chỉ số thường dùng là:

- **Variance (Phương sai):** Đo lường mức độ phân tán của dữ liệu quanh giá trị trung bình. Nên dùng trong bài toán hồi quy khi cần tính giá trị bình phương cho khoảng cách.

$$Var(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- **Standard Deviation (Độ lệch chuẩn):** Là căn bậc hai của phương sai, có cùng đơn vị với dữ liệu gốc nên dễ hiểu hơn. Dùng khi muốn giải thích mức độ biến động cùng đơn vị với dữ liệu gốc.

$$SD(X) = \sqrt{Var(X)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Ví dụ minh họa:

- Dữ liệu doanh thu (triệu VND): 50, 52, 49, 51, 48.
- Trung bình = 50.
- Phương sai = 2.5 (thấp) → Doanh thu dao động ít quanh 50.
- Độ lệch chuẩn = $\sqrt{2.5} \approx 1.58$ → Khoảng 68% doanh thu nằm trong khoảng [48.42, 51.58].

So sánh hai nhóm sản phẩm:

- Nhóm A: {100, 105, 95, 102, 98} → Độ lệch chuẩn = 3.8 (dao động nhỏ).
- Nhóm B: {80, 120, 60, 140, 100} → Độ lệch chuẩn = 31.6 (dao động rất lớn).
- Cả hai nhóm đều có trung bình 100, nhưng Nhóm B rủi ro cao hơn vì biến động mạnh.

Case study: so sánh 2 nhóm sản phẩm sử dụng Variance và Standard Deviation trong phân tích dữ liệu.

Phương sai (Variance)	Độ lệch chuẩn (Standard Deviation)	Case study: So sánh 2 nhóm sản phẩm
<ul style="list-style-type: none">• 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<ul style="list-style-type: none">• Trung bình: 50 triệu• Phương sai: 2.5 → Biến động thấp	<ul style="list-style-type: none">• 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 = $\sqrt{2.5} \approx 1.58$<ul style="list-style-type: none">• Khoảng 68% dữ liệu nằm trong khoảng 50 ± 1.58 triệu (48.42 - 51.58 triệu)	<ul style="list-style-type: none">• 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.

Hình 8: Ứng dụng Variance và Standard Deviation trong phân tích dữ liệu

2.4 Một số hàm thống kê cơ bản trong Excel

Excel cung cấp nhiều hàm thống kê hỗ trợ phân tích:

- **AVERAGE()** – Tính trung bình.
- **MEDIAN()** – Tính trung vị.
- **MODE()** – Giá trị xuất hiện nhiều nhất.
- **VAR.P()** – Phương sai toàn bộ.
- **STDEV.P()** – Độ lệch chuẩn toàn bộ.
- **SKEW()** – Độ lệch của phân phối.
- **KURT()** – Độ nhọn của phân phối.

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

Hình 9: Các hàm thống kê trong Excel

2.5 Hàm điều kiện và tổng hợp

Ngoài thống kê cơ bản, Excel còn hỗ trợ các hàm điều kiện giúp lọc và tính toán theo tiêu chí cụ thể:

- IF(), IFS() – Hàm điều kiện.
- SUMIF(), SUMIFS() – Tổng có điều kiện.
- COUNTIF(), COUNTIFS() – Đếm có điều kiện.
- AVERAGEIF(), AVERAGEIFS() – Trung bình có điều kiện.

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 > 100 thì 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 mức
Tổng có điều kiện	SUMIF()	=SUMIF(phạm_vi, tiêu_chí, phạm_vi_tổng)	=SUMIF(B1:B10,"Hà Nội",C1:C10) → Tính 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,">100") → 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 khách hàng nam trên 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 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,">5000000") → 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 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 của "SP005" trong phạm vi A1:A20

Hình 10: Hàm tổng hợp và điều kiện trong Excel

2.6 Pivot Table

Pivot Table là công cụ mạnh trong Excel để tóm tắt dữ liệu theo nhiều chiều. Người dùng có thể kéo thả trường dữ liệu vào các ô Rows, Columns, Values và Filters để nhanh chóng tạo báo cáo động. Đây là cách đơn giản nhưng hiệu quả để phân tích tập dữ liệu lớn.

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

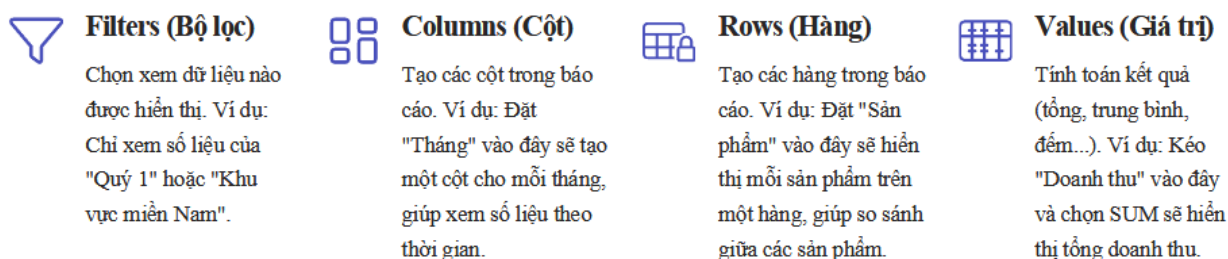
Sử dụng Slicers để lọc → Nhóm theo thời gian → Tạo Calculated Fields → Hiển thị dữ liệu dưới dạng % với Show Values As

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

Hình 11: Quy trình sử dụng 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:



Hình 12: Nguyên lý hoạt động của Pivot Table

Phần 3: Trực Quan Hóa Dữ Liệu

3.1 Vì sao cần trực quan hóa dữ liệu?

Trực quan hóa giúp con người dễ dàng nhận ra xu hướng và thông tin quan trọng từ dữ liệu thô. Thay vì đọc hàng trăm dòng số liệu, biểu đồ hoặc sơ đồ cho phép nắm bắt thông điệp chỉ trong vài giây. Đây là yếu tố quan trọng trong báo cáo, thuyết trình, và hỗ trợ ra quyết định dựa trên dữ liệu. Ví dụ, minh họa dữ liệu có tính biến động theo thời gian sẽ dễ hiểu hơn thay vì dùng bản với số đơn thuần.

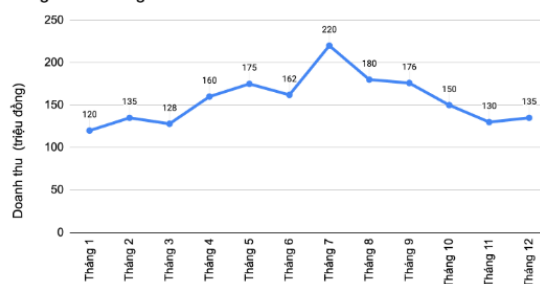
✗ **Khó nắm bắt xu hướng**

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



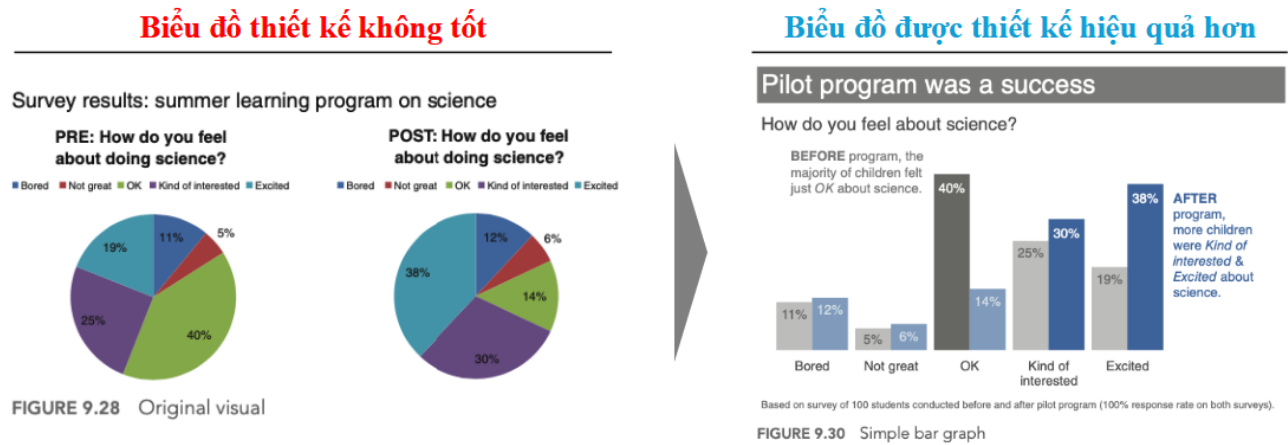
Hình 13: Tầm quan trọng của trực quan hóa dữ liệu

3.2 Cách sử dụng biểu đồ Hiệu quả

Một biểu đồ tốt không chỉ chính xác về mặt dữ liệu mà còn cần dễ hiểu và truyền tải thông điệp rõ ràng. Khi thiết kế biểu đồ, nên chú ý:

- **Chọn đúng loại biểu đồ:** Bar chart phù hợp để so sánh, line chart để thể hiện xu hướng, pie chart để hiển thị tỷ trọng.
- **Đơn giản và rõ ràng:** Tránh quá nhiều màu sắc, ký hiệu hoặc chi tiết dư thừa làm rối mắt người xem.
- **Sử dụng nhãn và tiêu đề rõ ràng:** Đặt tên trục, đơn vị đo, chú thích (legend) nếu cần.
- **Nhấn mạnh thông tin chính:** Dùng màu tương phản hoặc chú thích để làm nổi bật dữ liệu quan trọng.

- **Tránh lạm dụng 3D hoặc hiệu ứng:** Các yếu tố này có thể làm méo mó nhận thức về dữ liệu.



Hình 14: So sánh biểu đồ trực quan kém và hiệu quả

3.3 Các loại biểu đồ phổ biến

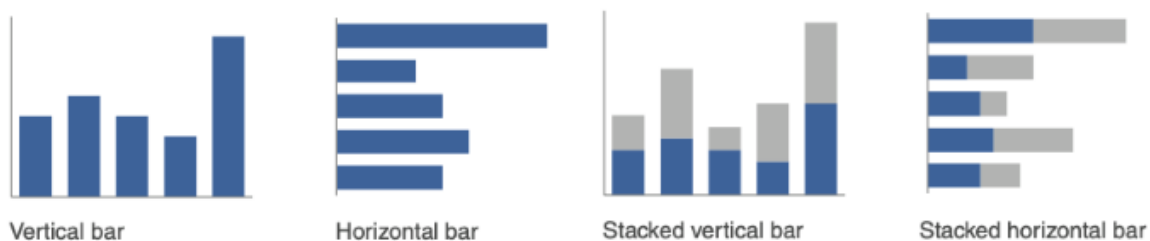
Có nhiều loại biểu đồ và sơ đồ, mỗi loại phù hợp cho một mục đích khác nhau: so sánh, theo dõi xu hướng, phân tích mối quan hệ, hoặc mô tả phân phối dữ liệu.



Hình 15: Các loại biểu đồ, plot và diagram

3.4 Biểu đồ cột (Bar Chart)

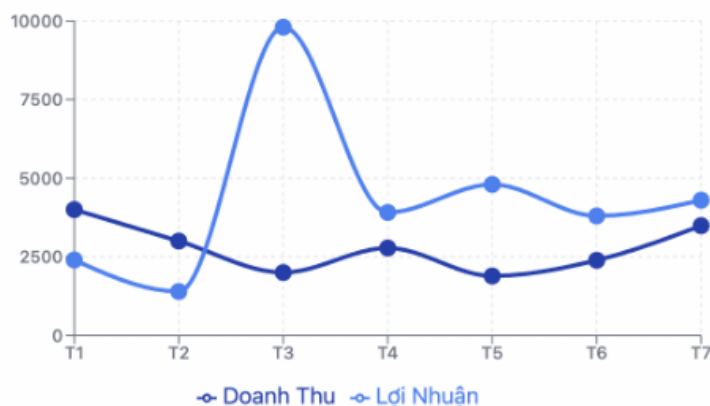
Dùng để so sánh giá trị giữa các nhóm hoặc hạng mục. **Ví dụ:** So sánh doanh số bán hàng của các chi nhánh trong cùng một tháng.



Hình 16: Ví dụ về Bar Chart

3.5 Biểu đồ đường (Line Chart)

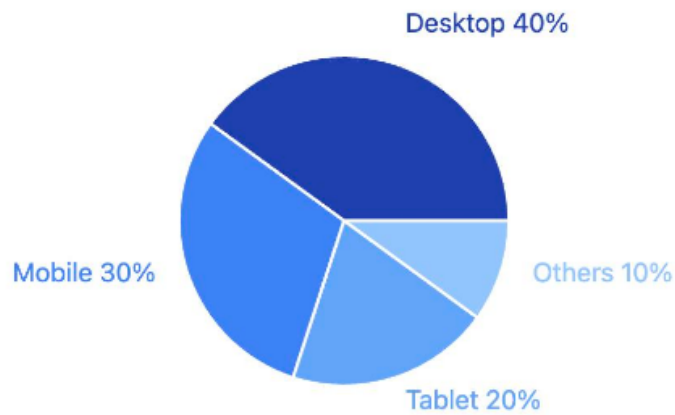
Phù hợp để thể hiện xu hướng theo thời gian. **Ví dụ:** Theo dõi doanh thu hàng tháng trong một năm, hoặc lượng truy cập website theo ngày.



Hình 17: Ví dụ về Line Chart

3.6 Biểu đồ tròn (Pie Chart)

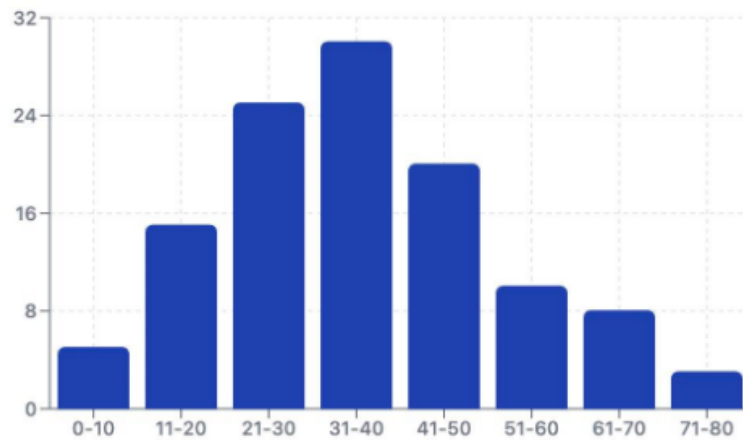
Hiển thị tỷ trọng các thành phần trong tổng thể. Chỉ nên dùng khi số nhóm ít (dưới 7). **Ví dụ:** Tỷ lệ khách hàng theo khu vực (Bắc, Trung, Nam), hoặc Tỷ lệ bán của Máy tính bàn, điện thoại, máy tính bảng và các sản phẩm khác.



Hình 18: Ví dụ về Pie Chart

3.7 Histogram

Dùng để mô tả phân phối dữ liệu và tần suất xuất hiện. **Ví dụ:** Phân phối điểm thi của sinh viên, hoặc độ tuổi khách hàng trong một cửa hàng.



Hình 19: Ví dụ về Histogram

3.8 Heatmap

Dùng màu sắc để biểu diễn cường độ hoặc giá trị, để nhận biết vùng mạnh – yếu. **Ví dụ:** Ma trận tương quan giữa các doanh thu và lợi nhuận theo từng tháng.

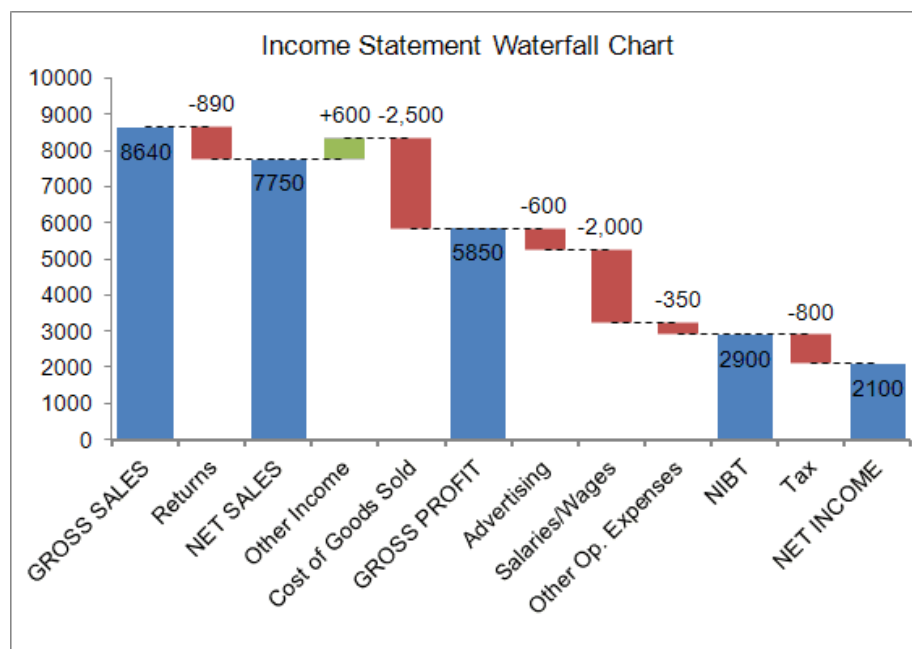
Tháng	Doanh thu	Lợi nhuận
Tháng 1	120	-12
Tháng 2	135	-7
Tháng 3	128	13
Tháng 4	160	16
Tháng 5	175	18
Tháng 6	162	16
Tháng 7	220	33
Tháng 8	180	18
Tháng 9	176	18
Tháng 10	150	15
Tháng 11	130	13
Tháng 12	135	14

※ Đơn vị: Triệu đồng

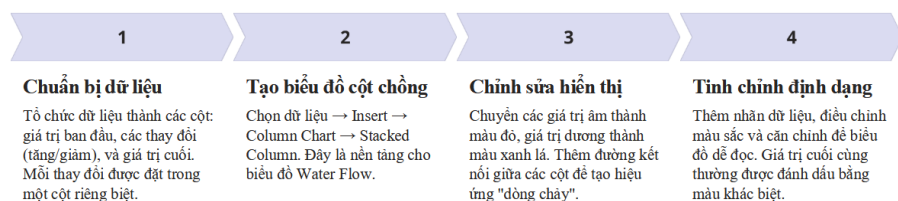
Hình 20: Ví dụ về Heatmap

3.9 Waterfall Chart

Minh họa sự thay đổi tích lũy của một chỉ số qua các yếu tố cộng/trừ. **Ví dụ:** Phân tích lợi nhuận ròng từ doanh thu, trừ chi phí và thuế để ra kết quả cuối cùng.



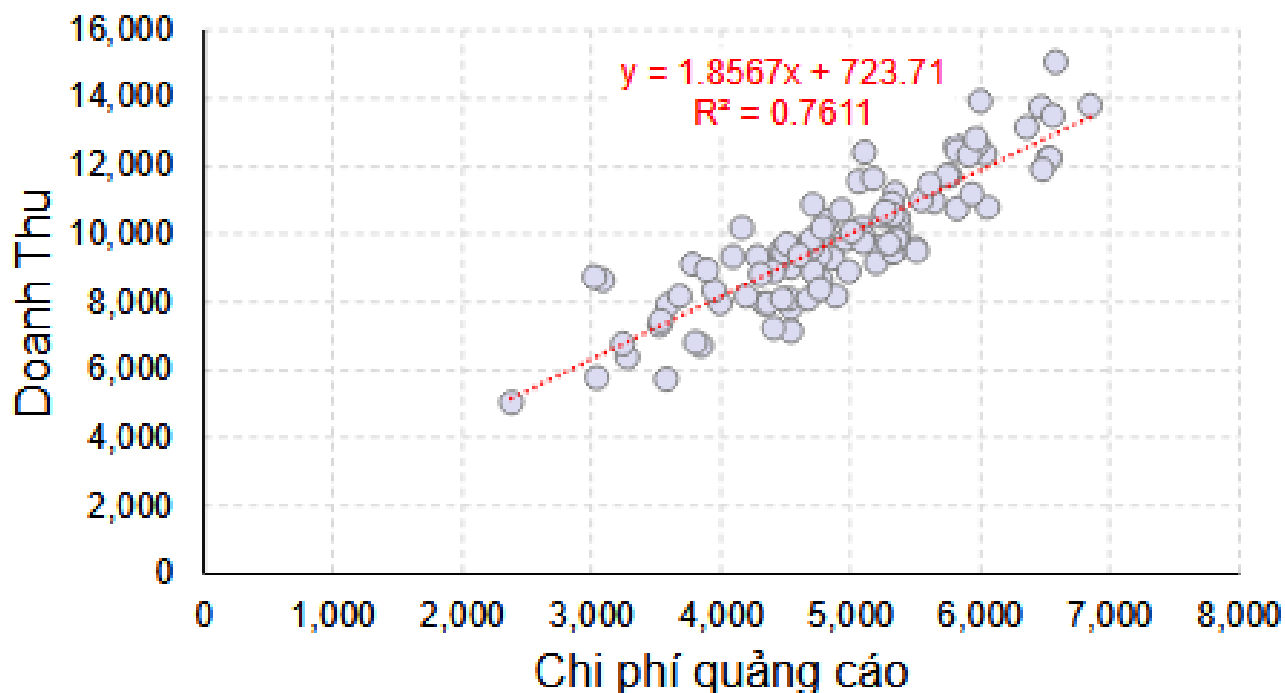
Hình 21: Ví dụ về Waterfall Chart



Hình 22: Nguyên lý xây dựng Waterfall Chart

3.10 Scatter Plot

Dùng để xem mối quan hệ giữa hai biến. Có thể thêm đường hồi quy để đánh giá xu hướng. **Ví dụ:** So sánh số giờ học và điểm thi của sinh viên, hoặc mức chi tiêu quảng cáo và doanh số bán hàng.



Hình 23: Ví dụ về Scatter Plot

3.11 Hệ số tương quan (Correlation Coefficient)

Hệ số tương quan cho ta biết mức độ và chiều hướng mối quan hệ tuyến tính giữa hai biến số. Giá trị này nằm trong khoảng từ -1 đến +1:

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

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	0	-1
Tương quan dương hoàn hảo	Không có tương quan	Tương quan âm hoàn hảo
Hai biến cùng tăng hoặc cùng giảm	Không có liên hệ tuyến tính	Một biến tăng, biến kia giảm

Hình 24: Hệ số tương quan

- Gần +1: Hai biến có mối quan hệ tuyến tính **cùng chiều mạnh** (biến A tăng thì biến B cũng tăng).
- Gần -1: Hai biến có mối quan hệ tuyến tính **ngược chiều mạnh** (biến A tăng thì biến B giảm).
- Gần 0: Không tồn tại mối quan hệ tuyến tính rõ ràng.

Cách phân tích:

- Giá trị lớn về độ lớn ($|r|$ gần 1) → Mối quan hệ đáng tin cậy hơn, có thể dự đoán biến này dựa trên biến kia.
- Giá trị nhỏ ($|r|$ gần 0) → Quan hệ yếu, cần kiểm tra thêm hoặc dùng mô hình phi tuyến.
- Lưu ý: Tương quan không có nghĩa là quan hệ nhân quả. Ví dụ: số kem bán ra và số vụ đuối nước cùng tăng vào mùa hè, nhưng không phải nguyên nhân trực tiếp.

Khi nào dùng CORREL hoặc PEARSON trong Excel:

- `=CORREL(array1, array2)`: Hàm ngắn gọn, trực tiếp trả về hệ số tương quan Pearson. Dùng khi chỉ cần kiểm tra nhanh mối liên hệ giữa hai biến.
- `=PEARSON(array1, array2)`: Cho cùng kết quả như CORREL, nhưng rõ ràng về mặt ý nghĩa thống kê. Nên dùng khi làm báo cáo hoặc tài liệu học thuật để nhấn mạnh đây là hệ số Pearson.