**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----o0o----

**BÁO CÁO HỌC PHẦN: KHAI PHÁ DỮ LIỆU**

**TÊN ĐỀ TÀI:** **TƯƠNG QUAN PEARSON**

**NHÓM: 15**

**Thành phố Hồ Chí Minh, tháng 4 năm 2025**

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----o0o----

**TÊN ĐỀ TÀI: TƯƠNG QUAN PEARSON**

|  |  |
| --- | --- |
| **Nhóm: 15**  Trưởng nhóm: Trần Công Minh  Thành viên:   1. Lê Đức Trung 2. Nguyễn Hữu Thắng | **Giảng viên hướng dẫn: Trần Như Ý** |

**Thành phố Hồ Chí Minh, tháng 4 năm 2025**

**BẢNG PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MSSV** | **Họ tên** | **Đóng góp tỷ lệ %** | **Nhiệm vụ được phân công** | **Mức độ hoàn thành công việc được phân công** |
| 2001230909 | Nguyễn Hữu Thắng | 90% | * Giả định và Hạn chế * Ứng dụng | Hoàn thành tốt, đúng hạn |
| 2001225676 | Lê Đức Trung | 90% | * Công thức * Ví dụ minh họa * So sánh | Hoàn thành tốt, đúng hạn |
| 2001222641 | Trần Công Minh | 100% | * Định nghĩa * Demo * Word * PowerPoint | Hoàn thành tốt, đúng hạn |

MỤC LỤC

[**PHẦN 1.** **NỘI DUNG** 2](#_Toc194596591)

[**1.** **Định nghĩa và công thức** 2](#_Toc194596592)

[**2.** **Diễn giải hệ số tương quan** 4](#_Toc194596593)

[**3.** **Giả định và hạn chế** 5](#_Toc194596594)

[**4.** **Ứng dụng trong khai phá dữ liệu** 7](#_Toc194596595)

[**5.** **So sánh với ANOVA và Chi-square** 7](#_Toc194596596)

[**6.** **Ví dụ minh họa** 8](#_Toc194596597)

[**7.** **Demo ứng dụng** 12](#_Toc194596598)

[**8.** **Kết luận** 14](#_Toc194596599)

[PHẦN 2. TÀI LIỆU THAM KHẢO 15](#_Toc194596600)

[Danh mục hình ảnh 15](#_Toc194596601)

**MỞ ĐẦU**

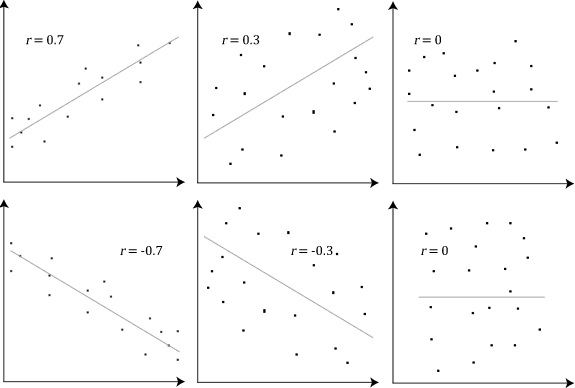
Tương quan Pearson là một công cụ thống kê quan trọng, được sử dụng rộng rãi để đo lường mối quan hệ tuyến tính giữa hai biến số liên tục. Được phát triển bởi Karl Pearson vào cuối thế kỷ 19, phương pháp này đóng vai trò nền tảng trong phân tích dữ liệu, đặc biệt trong khai phá dữ liệu, nơi cần hiểu rõ mối quan hệ giữa các đặc trưng. Báo cáo này sẽ trình bày định nghĩa, công thức, cách diễn giải, giả định, hạn chế và ứng dụng của tương quan Pearson trong khai phá dữ liệu, kèm theo các ví dụ minh họa thực tế.

1. **NỘI DUNG**
2. **Định nghĩa và công thức**

|  |  |
| --- | --- |
| Hệ số này được đưa ra bởi ***Karl Pearson***, đây chính là cha đẻ của hệ số Pearson. Karl Pearson (27 tháng 3 năm 1857 - 27 tháng 4 năm 1936) là một nhà toán học người Anh, được ghi nhận là người đã thành lập nên môn thống kê toán học. Năm 1911 ông thành lập khoa thống kê đầu tiên trên thế giới tại Đại học College London. Ông là một người theo thuyết ưu sinh và là người viết tiểu sử của Francis Galton. | Karl Pearson. |

Tương quan Pearson, hay hệ số tương quan sản phẩm-moment Pearson (PPMC), đo lường mức độ và hướng của mối quan hệ tuyến tính giữa hai biến x và y. Giá trị của nó nằm trong khoảng từ -1 đến 1:

* r = 1: Mối quan hệ tuyến tính dương hoàn hảo, tức là khi x tăng, y tăng theo tỉ lệ cố định
* r = -1: Mối quan hệ tuyến tính âm hoàn hảo, tức là khi x tăng, y giảm theo tỉ lệ cố định
* r = 0: Không có mối quan hệ tuyến tính giữa 2 biến

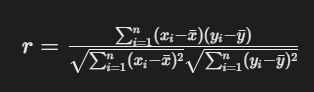


*Hình 1. Ảnh minh họa bằng biểu đồ Scartter Plot*

Mục đích của việc chạy tương quan Pearson là kiểm tra mối tương quan tuyến tính chặt chẽ giữa biến phụ thuộc với các biến độc lập. Bởi vì điều kiện để hồi quy là trước nhất phải tương quan. Tiếp đến đó là nhận diện vấn đề đa cộng tuyến khi các biến độc lập cũng có tương quan mạnh với nhau. Dấu hiệu đa cộng tuyến sẽ được xem xét khi phân tích hồi quy (Kiểm tra hệ số VIF)

Giữa hai biến định lượng có nhiều dạng liên hệ, có thể là tuyến tính hoặc phi tuyến hoặc không có bất kỳ một mối liên hệ nào. Thường các nhà nghiên cứu nhận diện sớm mối quan hệ một cặp biến thông qua đồ thị phân tán Scatter.

Công thức tính hệ số tương quan



Trong đó:

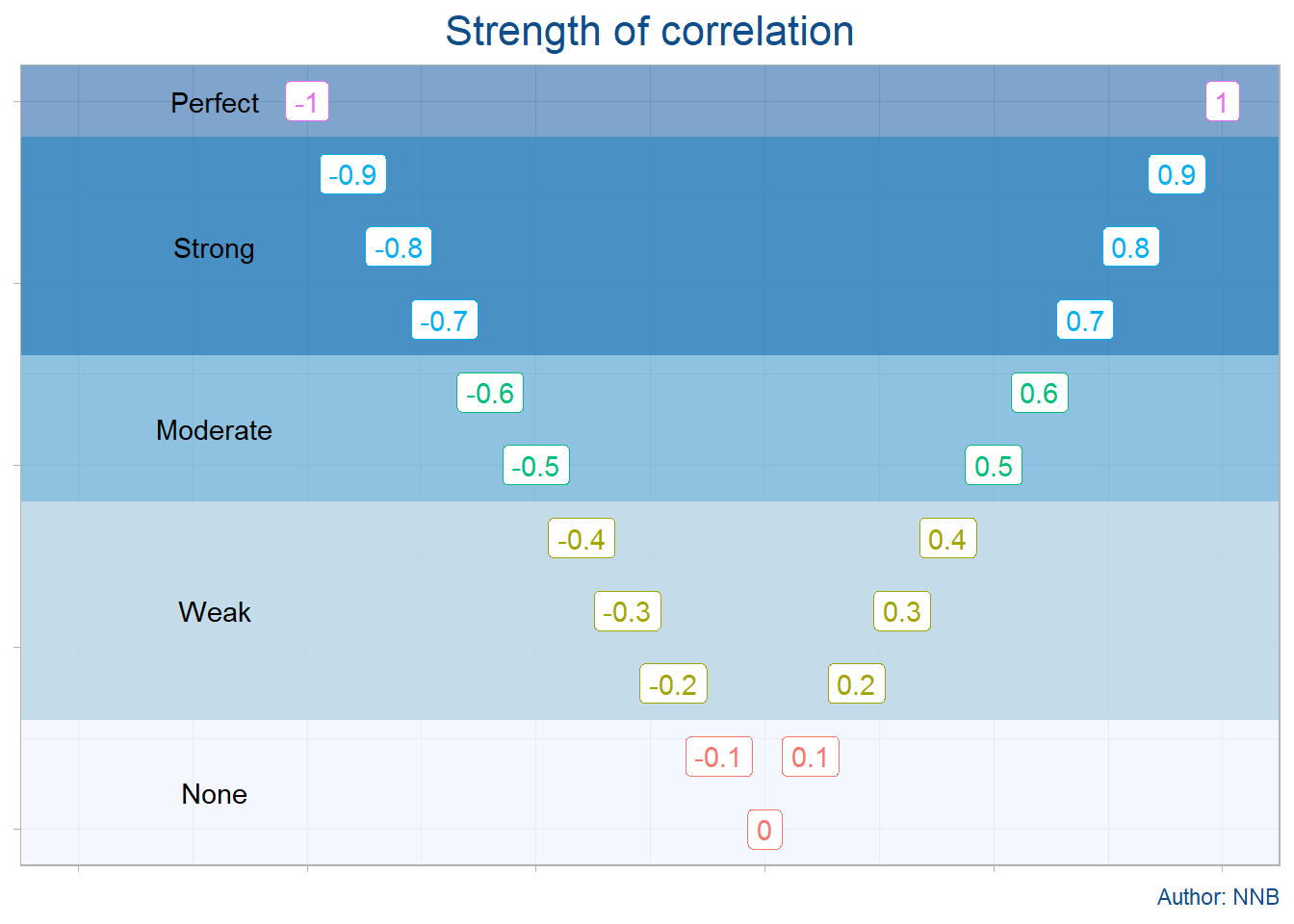
* n: Số lượng quan sát trong tập dữ liệu.
* xi: Giá trị của biến x tại quan sát thứ i.
* yi: Giá trị của biến y tại quan sát thứ i.
* : Giá trị trung bình của biến x, tính bằng
* : Giá trị trung bình của biến x, tính bằng

Công thức này thể hiện việc chuẩn hóa covariance (mức độ hai biến thay đổi cùng nhau) bằng tích của độ lệch chuẩn của *x* và *y* r luôn nằm trong khoảng [-1, 1].

1. **Diễn giải hệ số tương quan**

Hệ số tương quan Pearson r cung cấp thông tin về cả hướng và sức mạnh của mối quan hệ tuyến tính:

* Hướng:
* Nếu r > 0: Mối quan hệ dương, nghĩa là khi giá trị của x tăng, giá trị của y cũng tăng. Ví dụ, số giờ học và điểm thi thường có r > 0
* Nếu r < 0: Mối quan hệ âm, nghĩa là khi giá trị của x tăng, giá trị của y giảm. Ví dụ, nhiệt độ và lượng mưa có thể có r < 0.
* Sức mạnh:
* >= 0.7: Tương quan mạnh, hai biến có mối quan hệ chặt chẽ.
* 0.3 <= < 0.7: Tương quan trung bình, mối liên hệ rõ ràng nhưng không quá mạnh.
* < 0.3: Tương quan yếu, mối liên hệ rất mờ nhạt hoặc gần như không đáng kể.



*Hình 2. Sức mạnh của mối quan hệ tuyến tính*

Ví dụ cụ thể:

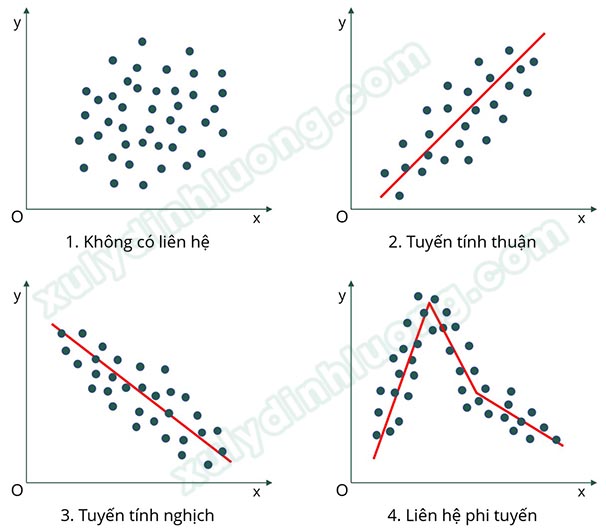
* Nếu r = 0.85 giữa số giờ học và điểm thi, điều này cho thấy mối quan hệ dương mạnh, nhưng không đảm bảo rằng học nhiều luôn dẫn đến điểm cao do có thể có các yếu tố khác (như chất lượng học).
* Nếu r = −0.6 giữa nhiệt độ và lượng mưa, điều này cho thấy mối quan hệ âm trung bình, tức là nhiệt độ cao thường đi kèm với lượng mưa thấp.

Lưu ý quan trọng: Tương quan không đồng nghĩa với nhân quả. Ví dụ, số lượng kem bán ra và số vụ cháy rừng có thể có r dương cao vào mùa hè, nhưng không có nghĩa kem gây ra cháy rừng – cả hai đều bị ảnh hưởng bởi nhiệt độ.

1. **Giả định và hạn chế**

Để áp dụng tương quan Pearson một cách chính xác, cần tuần thủ các giả định sau:

* Tính tuyến tính: Mối quan hệ giữa x và y phải là tuyến tính. Nếu mối quan hệ là phi tuyến tính (ví dụ, y = x2), r sẽ không phản ảnh đúng mức độ liên hệ.
* Phân phối gần chuẩn: Dữ liệu của x và y nên có phân phối gần giống phân phối chuẩn, đặc biệt khi sử dụng r để suy luận thống kê (tính giá trị p).
* Không có oultier nghiêm trọng: Các điểm dữ liệu bất thường (outlier) có thể làm sai lệnh r. Ví dụ, nếu một học sinh học 1 giờ nhưng đạt điểm 100 do thiên tài, r có thể giảm mạnh.
* Phương sai khác 0: Phương sai của x và y phải khác 0, và covariance giữa chúng phải hữu hạn.



*Hình 3. Hình minh họa các mối liên hệ*

1. **Ứng dụng trong khai phá dữ liệu**

Tương quan Pearson có nhiều ứng dụng thực tiễn trong khai phá dữ liệu trong khai phá dữ liệu:

* Chọn lọc đặc trưng (Feature Selection): Trong các mô hình học máy như hồi quy tuyến tính, nếu hai đặc trưng có r > 0.8 (ví dụ, diện tích nhà và số phòng ngủ), ta có thể loại bỏ một đặc trưng để tránh đa cộng tuyến, giúp mô hình đơn giản và hiệu quả hơn.
* Phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA): Trước khi xây dựng mô hình, tương quan Pearson giúp xác định mối quan hệ giữa các biến. Ví dụ, trong phân tích bán hàng, nếu chi phí quảng cáo và doanh thu có r = 0.75, ta có thể ưu tiên đầu tư vào quảng cáo.
* Đo lường sự tương đồng: Trong phân cụm (clustering), r được dùng để so sánh mức độ tương đồng giữa các đối tượng dữ liệu, hỗ trợ nhóm các điểm dữ liệu có đặc điểm giống nhau.

Ví dụ thực tế: Trong lĩnh vực y tế, tương quan Pearson có thể được dùng để phân tích mối quan hệ giữa liều lượng thuốc và mức độ giảm triệu chứng (ví dụ, 𝑟 = 0.65 cho thấy mối quan hệ dương trung bình).

1. **So sánh với ANOVA và Chi-square**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí** | **Tương quan Pearson** | **Phân tích phương sai (ANOVA)** | **Kiểm định Chi-square** |
| Mục đích | Đo lường mức độ và hướng của mối quan hệ tuyến tính giữa 2 biến liên tục. | So sánh trung bình của nhiều nhóm (≥3) để xem có sự khác biệt đáng kể không. | Kiểm tra mối quan hệ giữa 2 biến phân loại (categorical). |
| Loại dữ liệu | Biến liên tục (continuous) (ví dụ: điểm số, nhiệt độ). | Biến phụ thuộc liên tục, biến độc lập phân loại (ví dụ: điểm số theo nhóm tuổi). | Biến phân loại (ví dụ: giới tính, sở thích). |
| Kết quả | Hệ số r (từ -1 đến 1) và p-value. | Giá trị F và p-value. | Giá trị Chi-square và p-value. |
| Ví dụ | Số giờ học và điểm thi (r = 0.98). | Điểm thi của 3 nhóm học sinh khác nhau. | Mối quan hệ giữa giới tính và sở thích TV. |
| Ưu điểm | Đơn giản, định lượng mối quan hệ tuyến tính. | Phát hiện sự khác biệt giữa nhiều nhóm. | Dễ áp dụng cho dữ liệu phân loại. |
| Nhược điểm | Chỉ đo tuyến tính, nhạy với outlier. | Không cho biết mối quan hệ cụ thể giữa biến. | Không đo lường sức mạnh mối quan hệ. |
| Giả định | Dữ liệu gần chuẩn, tuyến tính, không outlier. | Dữ liệu chuẩn, phương sai đồng nhất. | Tần suất kỳ vọng đủ lớn (thường ≥ 5). |

1. **Ví dụ minh họa**

* **Ví dụ 1: Số giờ học và điểm thi**

Giả sử ta có dữ liệu của 5 học sinh:

* Số giờ học: 2, 4, 6, 8, 10
* Điểm thi: 40, 50, 65, 80, 95
* n = 5 (số quan sát)

Bước 1: Tính trung bình và

Bước 2: Tính các giá trị (), ̣(), và các tích liên quan

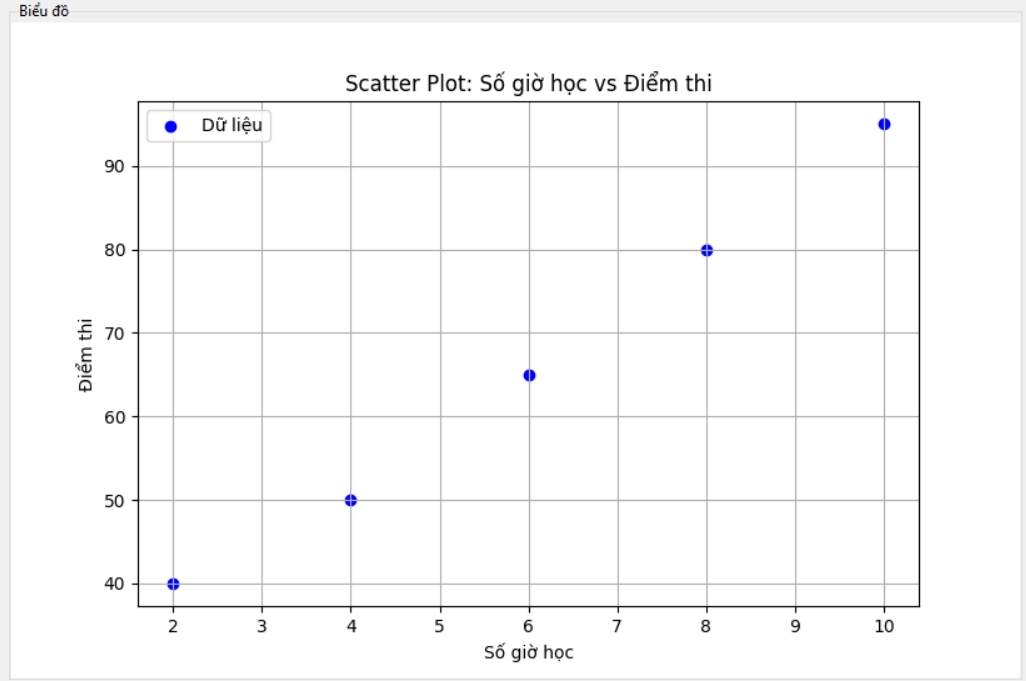
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| i |  |  |  |  | () ̣() | ()2 | ()2 |
| 1 | 2 | 40 | 2 - 6 = -4 | 40 – 66 = -26 | -4 x -26 = 104 | (-4)2 = 16 | (-26)2 = 676 |
| 2 | 4 | 50 | 4 – 6 = -2 | 50 – 66 = -16 | -2 x -16 = 32 | (-2)2 = 4 | (-16)2 = 256 |
| 3 | 6 | 65 | 6 – 6 = 0 | 65 – 66 = -1 | 0 x -1 = 0 | 02 = 0 | (-1)2 = 1 |
| 4 | 8 | 80 | 8 – 6 =2 | 80 – 66 = 14 | 2 x 14 = 28 | 22 = 4 | 142 = 196 |
| 5 | 10 | 95 | 10 – 6 = 4 | 95 – 66 = 29 | 4 x 29 = 116 | 42 = 16 | 292 = 841 |

Bước 3: Tính tổng các cột liên quan

Bước 4 : Thay vào công thức

* r =

Kết quả tính toán cho thấy r = 0.996, thể hiện mối quan hệ tuyến tính dương rất mạnh. Điều này cho thấy số giờ học tăng có liên quan chặt chẽ đến điểm thi cao, nhưng không đảm bảo nhân quả do có thể có yếu tố như tài năng tự nhiên.



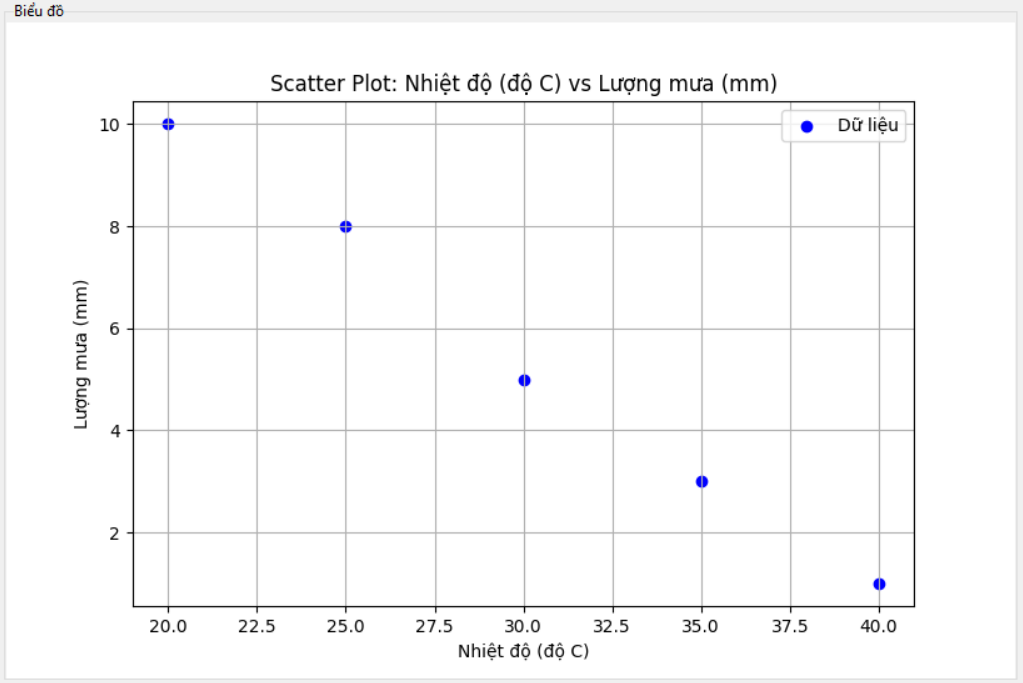
*Hình 4. Biểu đồ ví dụ 1*

* **Ví dụ 2: Nhiệt độ và lượng mưa**

Dữ liệu từ 5 ngày:

* Nhiệt độ(độ C): 20, 25, 30, 35, 40.
* Lượng mưa (mm): 10, 8, 5, 3, 1
* n = 5 (số quan sát)

Tương tự cách tính ở ví dụ 1, kết quả tính toán cho thấy r = - 0.97, thể hiện mối quan hệ tuyến tính âm rất mạnh. Nhiệt độ cao đi kèm với lượng mưa thấp, nhưng không có nghĩa nhiệt độ trực tiếp làm giảm mưa – cả hai có thể bị ảnh hưởng bởi mùa khô.



*Hình 5. Biểu đồ ví dụ 2*

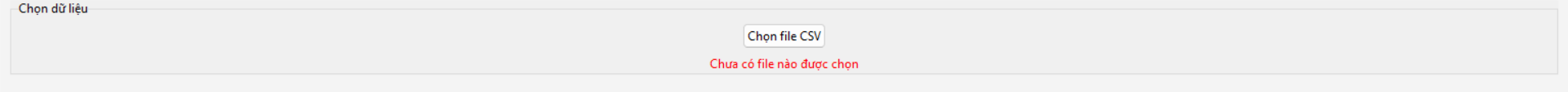
1. **Demo ứng dụng**

**Thư viện sử dụng:**

* tkinter: Xây dựng giao diện người dùng
* pandas: Đọc và xử lý dữ liệu CSV
* scipy.stats: Cung cấp hàm tính Pearson (pearsonr)
* numpy: Tính toán số học
* matplotlib: Vẽ đồ thị

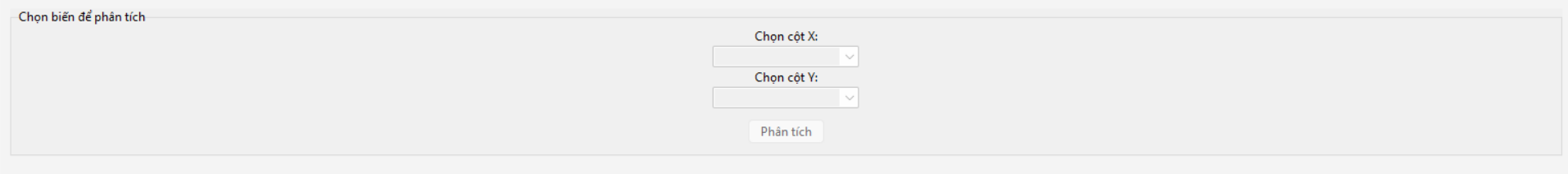
**Giao diện người dùng:**

* Cửa sổ chính (full màn hình)
* Ba phần chính:
* Phần chọn file



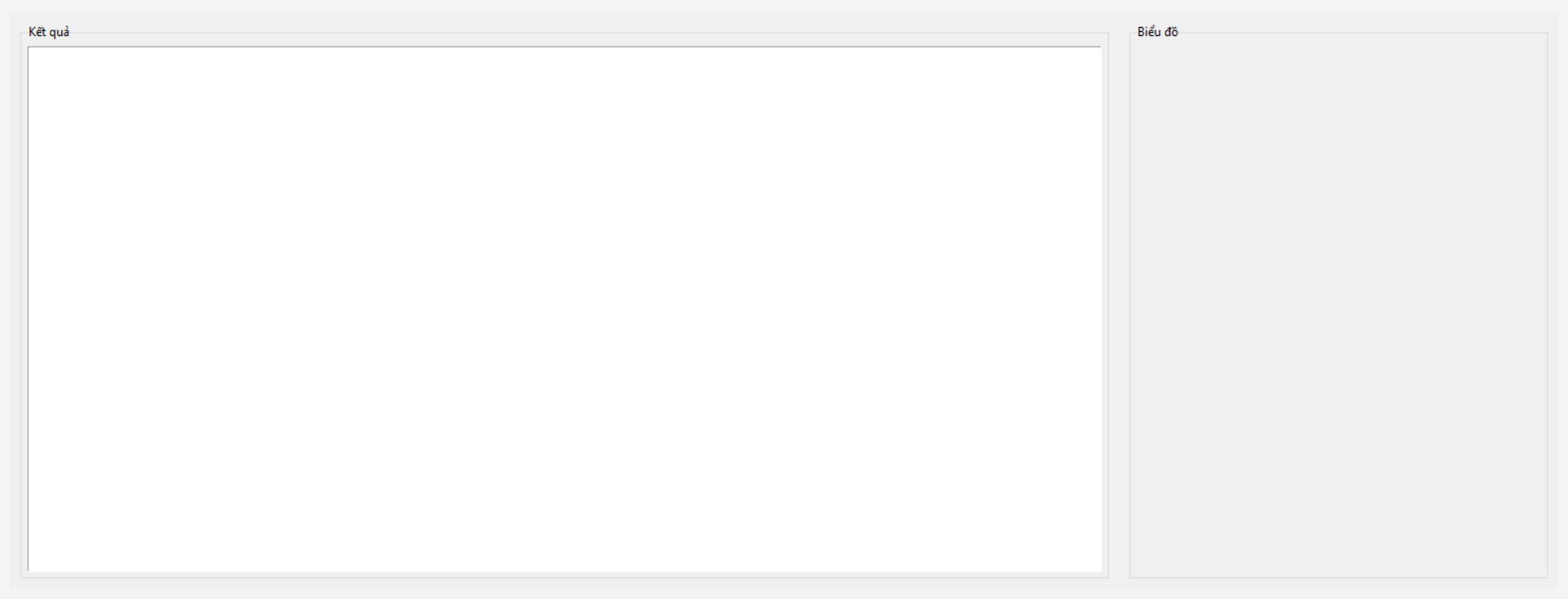
*Hình 6. Phần chọn file*

* Phần chọn biến phân tích (X và Y)



*Hình 7. Phần chọn biến phân tích (X và Y)*

* Phần hiển thị kết quả (kết quả tính toán và biểu đồ)



*Hình 8. Phần hiển thị kết quả*

**Luồng hoạt động chính**

**1. Đọc dữ liệu:**

* Khi người dùng click "Chọn file CSV", hàm *browse\_file()* mở hộp thoại chọn file
* Dữ liệu được đọc vào DataFrame pandas
* Danh sách các cột được nạp vào các combobox X và Y

**2. Phân tích dữ liệu:**

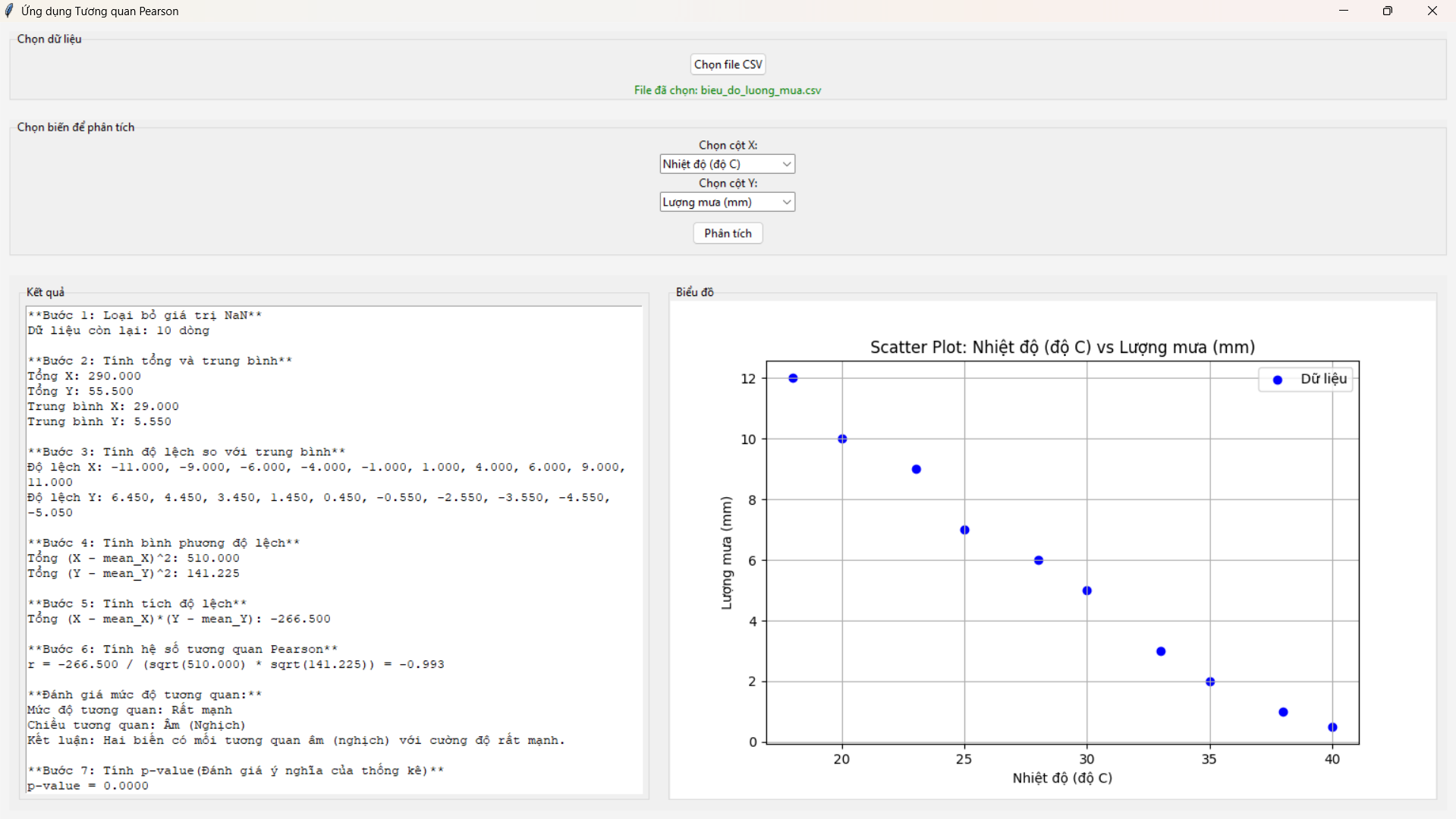
* Khi người dùng click "Phân tích", hàm *analyze()* thực hiện các bước:
* Kiểm tra tính hợp lệ của lựa chọn (đã chọn hai cột khác nhau)
* Loại bỏ các dòng có giá trị NaN
* Thực hiện các bước tính toán hệ số tương quan theo công thức:

**r = Σ[(X-mean\_X)(Y-mean\_Y)] / √[Σ(X-mean\_X)² × Σ(Y-mean\_Y)²]**

* Tính p-value để đánh giá ý nghĩa thống kê
* Tính khoảng tin cậy 95% sử dụng phép biến đổi Fisher z

**Hiển thị kết quả:**

* Hiển thị chi tiết từng bước tính toán trong khung kết quả:
* Số lượng dữ liệu sau khi loại bỏ NaN
* Tổng và trung bình của X, Y
* Độ lệch so với giá trị trung bình
* Tính bình phương độ lệch và tích độ lệch
* Hệ số tương quan r và p-value
* Khoảng tin cậy 95%
* Hiển thị biểu đồ phân tán (scatter plot) giữa biến X và Y



*Hình 9. Giao diện hiển thị kết quả*

1. **Kết luận**

Tương quan Pearson là một công cụ thống kê đơn giản nhưng hiệu quả để đo lường mối quan hệ tuyến tính giữa hai biến số. Trong khai phá dữ liệu, nó hỗ trợ từ việc phân tích khám phá, chọn lọc đặc trưng đến đo lường sự tương đồng, đóng vai trò nền tảng trước khi áp dụng các kỹ thuật phức tạp hơn. Tuy nhiên, để sử dụng đúng, cần chú ý đến các giả định như tính tuyến tính, không có outlier và không suy ra nhân quả từ tương quan. Với tính ứng dụng cao, tương quan Pearson là một bước không thể thiếu trong quá trình khai phá tri thức từ dữ liệu.

# PHẦN 2. TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Phạm Lộc Blog. (2015). *Phân tích tương quan Pearson trong SPSS*. từ <https://www.phamlocblog.com/2015/11/phan-tich-tuong-quan-pearson-trong-spss.html> |
| [2] | Trần Quang Quý. (2021). *Correlation Coefficient - RPubs*. từ <https://rpubs.com/tranquangquy_ictu/769561> |
| [3] | Nghiên cứu giáo dục (2025). *Phân tích tương quan Pearson r (Pearson’s Collection r)* từ  <https://nghiencuugiaoduc.com.vn/bai-4-phan-tich-tuong-quan-pearson-r-pearsons-correlation-r/> |
| [4] | Xử lí định lượng (2020). *Phân tích và đọc kết quả tương quan Pearson trên SPSS.* Từ  <https://xulydinhluong.com/phan-tich-tuong-quan-pearson-tren-spss/> |

# Danh mục hình ảnh

[*Hình 1. Ảnh minh họa bằng biểu đồ Scartter Plot* 3](#_Toc194596534)

[*Hình 2. Sức mạnh của mối quan hệ tuyến tính* 5](#_Toc194596535)

[*Hình 3. Hình minh họa các mối liên hệ* 6](#_Toc194596536)

[*Hình 4. Biểu đồ ví dụ 1* 10](#_Toc194596537)

[*Hình 5. Biểu đồ ví dụ 2* 11](#_Toc194596538)

[*Hình 6. Phần chọn file* 12](#_Toc194596539)

[*Hình 7. Phần chọn biến phân tích (X và Y)* 12](#_Toc194596540)

[*Hình 8. Phần hiển thị kết quả* 13](#_Toc194596541)

[*Hình 9. Giao diện hiển thị kết quả* 14](#_Toc194596542)