**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

A blue button with white text

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

**PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI**

**PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN KẾT QUẢ HỌC TẬP**

**Sinh viên Thực hiện:** Lê Duy Tín – 3122411211

**Lớp:** DCT122C3

**Giảng viên hướng dẫn:** TS. Nguyễn Tuấn Đăng

**KHOA: CÔNG NGHỆ THÔNG TIN**

**THÀNH PHỐ HỒ CHÍ MINH, THÁNG 05 NĂM 2025**

# **MỤC LỤC**

[MỤC LỤC 2](#_Toc198331251)

[Tập dữ liệu 3](#_Toc198331252)

[Phần 1: Giới thiệu tập dữ liệu, lý do chọn file, mục tiêu phân tích 4](#_Toc198331253)

[Giới thiệu tập dữ liệu: 4](#_Toc198331254)

[Lý do chọn tập dữ liệu: 4](#_Toc198331255)

[Mục tiêu phân tích: 4](#_Toc198331256)

[Đặc điểm dữ liệu 4](#_Toc198331257)

[Kiểm tra giá trị bị thiếu và xử lý 5](#_Toc198331258)

[Phần 2: Mô tả cách tạo 3 biểu đồ,lý do chọn,xử lý dữ liệu(nếu có) 7](#_Toc198331259)

[Matplotlib 7](#_Toc198331260)

[SeaBorn 10](#_Toc198331261)

[Bokeh 11](#_Toc198331262)

[Phần 3:Phân tích xu hướng từ biểu đồ và kết quả kiểm định 20](#_Toc198331263)

[a. Kiểm định t-test 20](#_Toc198331264)

[b. Kiểm định z-test 22](#_Toc198331265)

[Phần 4: Đề xuất ít nhất 2 cách cải thiện kết quả học tập, dựa trên dữ liệu và kiểm định 25](#_Toc198331266)

[Đề xuất 1: Tăng thời lượng học tập mỗi tuần (studytime) 25](#_Toc198331267)

[Đề xuất 2: Giảm số ngày nghỉ học (absences) 25](#_Toc198331268)

[KẾT LUẬN 26](#_Toc198331269)

[TÀI LIỆU THAM KHẢO 27](#_Toc198331270)

# **Tập dữ liệu**

* Nguồn : UCI Machine Learning Repository.
* File : student-mat.csv
* Nội dung: ~395 bản ghi , với:
* G1,G2,G3: Điểm kỳ 1,kỳ 2, kỳ cuối(0-20)
* studytime: Thời gian học/tuần (1: <2 giờ: 2-5 giờ, 3: 5-10 giờ, 4 :>10 giờ)
* absences : số ngày nghỉ(0-93).
* sex: Giới tính(M/F)
* age,freetime: Tuổi,thời gian rảnh(1-5)

# **Phần 1: Giới thiệu tập dữ liệu, lý do chọn file, mục tiêu phân tích**

## Giới thiệu tập dữ liệu:

* Tập dữ liệu student-mat.csv chứa thông tin về học sinh học môn Toán tại một trường trung học ở Bồ Đào Nha. Dữ liệu này được thu thập nhằm mục đích phân tích các yếu tố ảnh hưởng đến kết quả học tập, đặc biệt là điểm số cuối kỳ (G3).

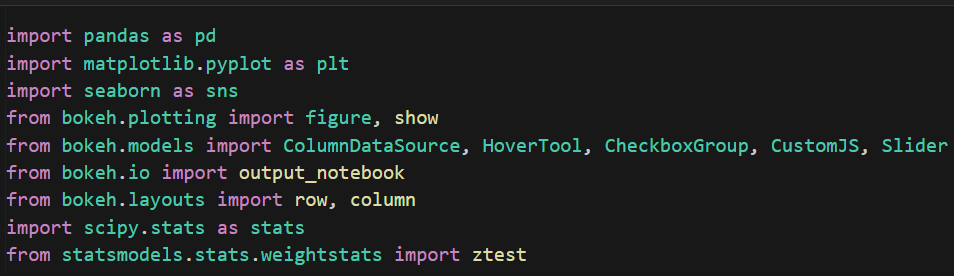
## Lý do chọn tập dữ liệu:

* Tập dữ liệu này được chọn vì nó:
* Có cấu trúc rõ ràng, dễ hiểu, phù hợp cho phân tích thống kê và trực quan hóa.
* Chứa nhiều biến định tính và định lượng liên quan đến hành vi học tập, giúp khám phá nhiều khía cạnh khác nhau về quá trình học tập của học sinh.

## Mục tiêu phân tích:

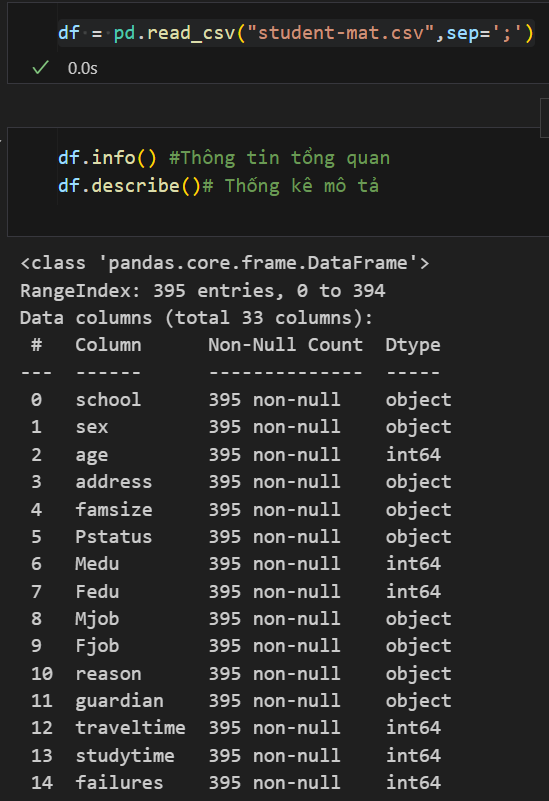
* Phân tích mối quan hệ giữa thời gian học tập (studytime) và kết quả học tập cuối kỳ (G3).
* Tìm hiểu xem các yếu tố như giới tính, số ngày nghỉ học (absences) có ảnh hưởng đến kết quả học tập hay không.
* Thực hiện các kiểm định thống kê để kiểm tra các giả thuyết cụ thể.

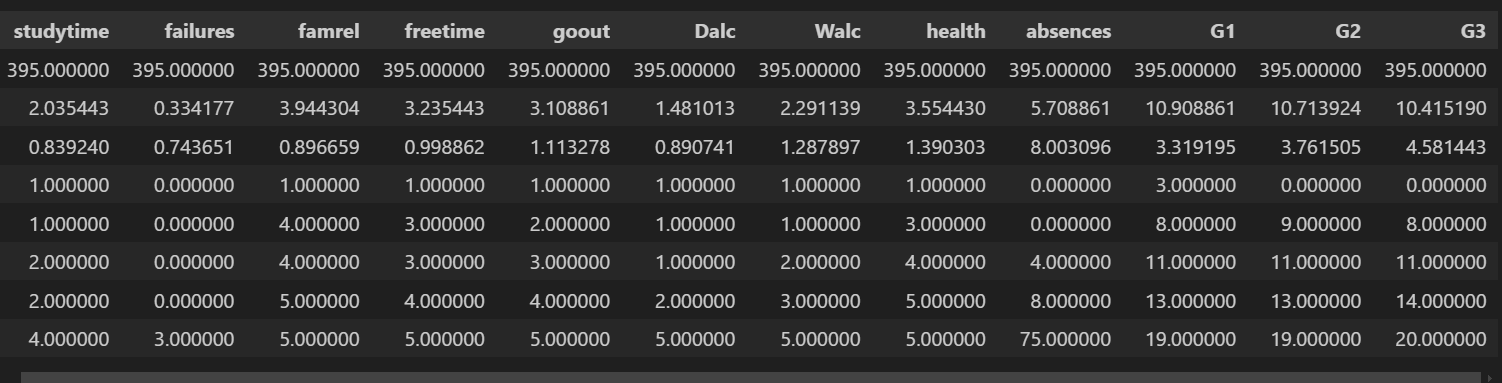
Thư viện đã sử dụng



## Đặc điểm dữ liệu

Số bản ghi: 395 học sinh.



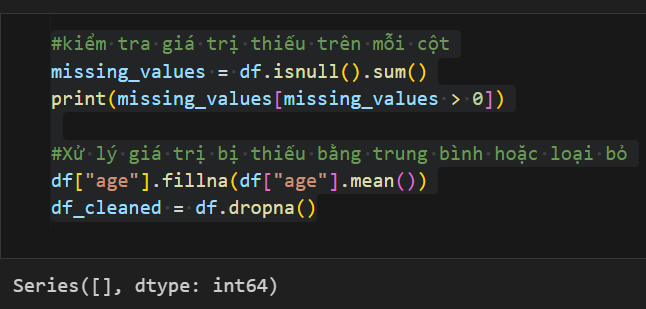


Phân phối điểm cuối kỳ (G3):

* Trung bình: khoảng 10 điểm (trên thang điểm 20).
* Dạng phân phối gần chuẩn nhưng hơi lệch trái (nhiều học sinh đạt điểm trung bình/thấp).
* Phân phối studytime:
* Là biến rời rạc với 4 mức (1–4), trong đó học sinh chủ yếu học từ 2 giờ trở xuống (mức 1 và 2 chiếm đa số).

## Kiểm tra giá trị bị thiếu và xử lý

(ví dụ: thay bằng trung bình hoặc loại bỏ)



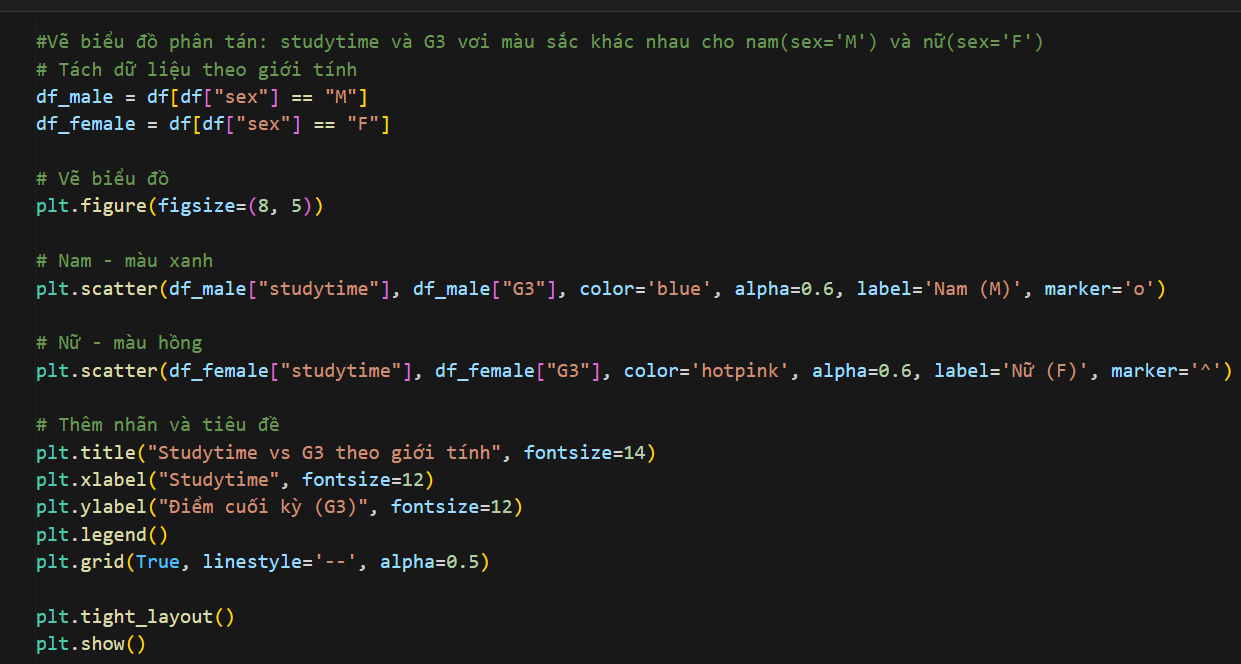
* df.isnull() tạo ra một bảng cùng kích thước với df, trong đó mỗi ô là True nếu giá trị bị thiếu (NaN), ngược lại là False.
* .sum() tính tổng số True (tức là số lượng giá trị bị thiếu) theo từng cột.
* missing\_values[missing\_values > 0] chỉ lọc ra những cột có ít nhất 1 giá trị thiếu để hiển thị.
* df["age"].fillna(df["age"].mean()): thay thế các giá trị thiếu trong cột age bằng giá trị trung bình của chính cột đó (df["age"].mean()).
* df.dropna() sẽ xóa toàn bộ các dòng chứa ít nhất một giá trị bị thiếu.
* df\_cleaned là bản dữ liệu sạch hoàn toàn, không còn bất kỳ giá trị thiếu nào.
* Trong quá trình tiền xử lý dữ liệu,đã kiểm tra các giá trị bị thiếu bằng cách sử dụng hàm isnull().sum() để xác định cột nào có dữ liệu không đầy đủ. Ví dụ cột age, tôi đã thay thế giá trị bị thiếu bằng trung bình của cột để giữ lại thông tin. Ngoài ra, một bản dữ liệu sạch khác cũng được tạo bằng cách loại bỏ toàn bộ các dòng chứa giá trị thiếu. Việc xử lý dữ liệu thiếu là cần thiết nhằm đảm bảo độ tin cậy và chính xác cho các phân tích tiếp theo.Sau khi kiểm tra thì xác nhận là file không bị thiếu thông tin.

# 

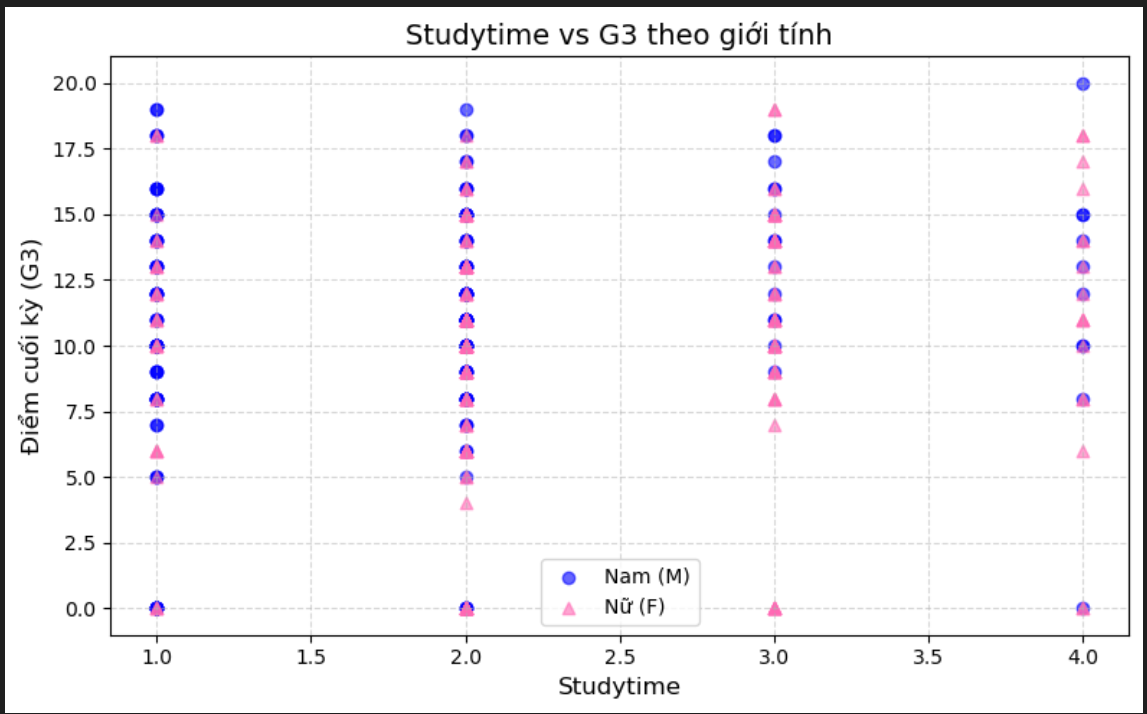
# **Phần 2: Mô tả cách tạo 3 biểu đồ,lý do chọn,xử lý dữ liệu(nếu có)**

## Matplotlib

* Vẽ biểu đồ phân tán: studytime và G3 vơi màu sắc khác nhau cho nam(sex='M') và nữ(sex='F')



* df[df["sex"] == "M"]: lọc dữ liệu học sinh nam.
* df[df["sex"] == "F"]: lọc dữ liệu học sinh nữ.
* plt.scatter(): vẽ biểu đồ phân tán, mỗi giới tính có màu và ký hiệu khác nhau.
* alpha=0.6: giúp điểm trong suốt nhẹ, dễ nhìn khi chồng lấp.
* tight\_layout(): căn lề hợp lý, tránh tràn nhãn.
* plt.legend(): Dòng này gọi hàm để hiển thị chú thích trên biểu đồ. Khi label được đặt trong mỗi nhóm dữ liệu, legend() sẽ tự động thêm các nhãn này vào vị trí thích hợp trên biểu đồ, giúp người dùng nhận biết được nhóm nào là nam (màu xanh, ký hiệu hình tròn) và nhóm nào là nữ (màu hồng, ký hiệu hình tam giác).



Biểu đồ : Phân tán (Matplotlib)

Thư viện sử dụng: Matplotlib:

* Loại biểu đồ: Biểu đồ phân tán (scatter plot)
* Mục tiêu: Khám phá mối quan hệ giữa thời gian học (studytime) và điểm cuối kỳ (G3), đồng thời so sánh giữa hai giới tính.

Lý do chọn:

* Matplotlib phù hợp để tạo các biểu đồ tùy chỉnh như phân tán, có thể dễ dàng phân biệt bằng màu sắc và ký hiệu cho từng nhóm.
* Dễ dàng thêm tiêu đề, nhãn trục và chú thích rõ ràng.

Phân tích biểu đồ:

* Biểu đồ giúp ta trực quan hóa mối quan hệ giữa thời gian học (studytime) và điểm tổng kết (G3), phân biệt theo giới tính.
* Có thể thấy:
* Học sinh có thời gian học nhiều hơn (studytime = 3 hoặc 4) thường có xu hướng đạt điểm G3 cao hơn.
* Cả nam và nữ đều có xu hướng tương tự, tuy nhiên nữ sinh có thể có phân bố điểm cao hơn trong một số mức studytime.
* Không có nhiều học sinh chọn studytime = 4 → có thể cần thêm dữ liệu để kết luận chắc chắn.

Mô tả cách tạo biểu đồ:

Bước 1: Tách dữ liệu thành 2 nhóm nam/nữ từ cột sex.

Bước 2: Dùng plt.scatter() để vẽ biểu đồ phân tán cho từng nhóm với màu và ký hiệu khác nhau.

Bước 3: Thêm tiêu đề, nhãn trục, legend và lưới để biểu đồ dễ đọc và chuyên nghiệp hơn.

Xử lý dữ liệu:

* Tách dữ liệu thành hai nhóm theo giới tính (sex == 'M' và sex == 'F').
* Gán màu sắc và ký hiệu khác nhau để thể hiện rõ hai nhóm trong biểu đồ.
* Giải thích biểu đồ:
* Mỗi điểm biểu diễn một học sinh.
* Trục X: thời gian học (studytime), trục Y: điểm cuối kỳ (G3).
* Nhóm Nam (màu xanh, hình tròn), Nữ (màu hồng, hình tam giác).
* Có thể quan sát sự phân bố điểm số giữa hai giới tính và mức độ học tập.

Ý nghĩa biểu đồ:

* Trục X (Studytime): Đại diện cho mức độ thời gian học của học sinh (1 = rất ít, 4 = rất nhiều).
* Trục Y (G3): Điểm cuối kỳ, thang điểm từ 0–20.

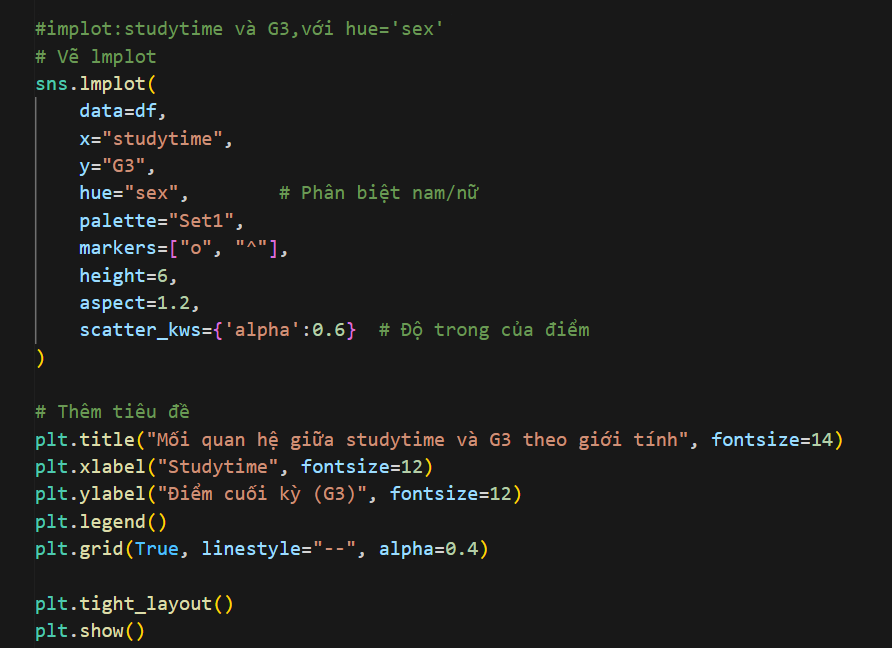
Màu sắc & ký hiệu:

* Nam sinh: Màu xanh, ký hiệu tròn (o).
* Nữ sinh: Màu hồng, ký hiệu tam giác (^).
* Xu hướng rút ra từ biểu đồ.
* Hầu hết học sinh có studytime từ 1 đến 2, nghĩa là thời gian học khá ít.
* Khi studytime tăng, có xu hướng điểm G3 cũng tăng — điều này dễ thấy hơn ở học sinh nữ.
* Nữ sinh thường có điểm số cao hơn ở các mức studytime cao hơn so với nam sinh.
* Ở mức studytime thấp, điểm G3 của cả nam và nữ đều phân tán rộng, chứng tỏ kết quả học kém ổn định.

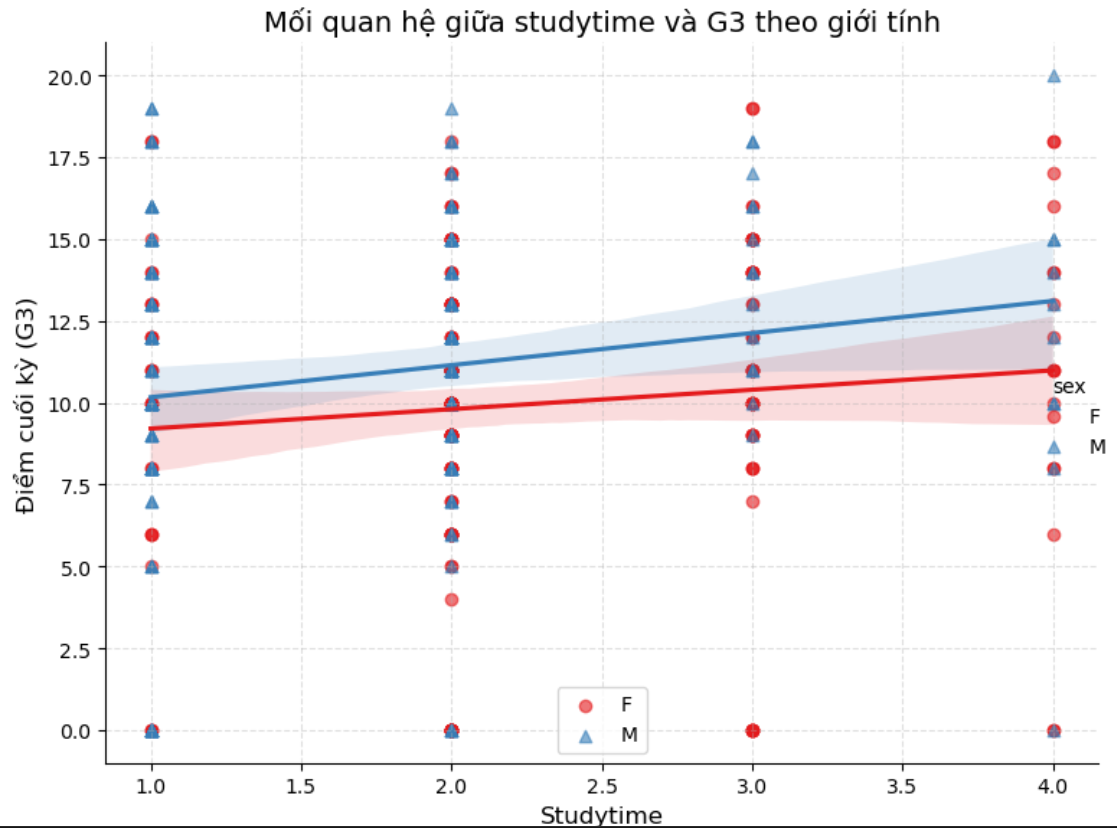
Kết luận ban đầu:

* Biểu đồ cho thấy rằng thời gian học tập có mối liên hệ tích cực với điểm số cuối kỳ, đặc biệt là đối với nữ sinh. Đây là gợi ý quan trọng cho các phân tích kiểm định thống kê sau này.

## SeaBorn



Biểu đồ: implot:studytime và G3,với hue='sex'



Biểu đồ : Hồi quy tuyến tính phân nhóm (Seaborn)

Thư viện sử dụng: Seaborn

* Loại biểu đồ: lmplot – biểu đồ phân tán kết hợp đường hồi quy
* Mục tiêu: Phân tích mối quan hệ giữa studytime và G3, đồng thời so sánh xu hướng theo giới tính.

Lý do chọn:

* Seaborn cung cấp biểu đồ lmplot tích hợp trực quan giữa phân tán và đường hồi quy, giúp làm nổi bật xu hướng dữ liệu theo nhóm.
* Tích hợp tốt với hue để thể hiện giới tính rõ ràng.

Phân tích biểu đồ:

* Biểu đồ thể hiện mối tương quan giữa số giờ học mỗi tuần (studytime) và điểm tổng kết (G3).
* Phân biệt giới tính giúp quan sát sự khác biệt giữa nam và nữ.
* Đường hồi quy cho mỗi giới cho thấy xu hướng điểm số tăng hoặc giảm theo thời gian học:
* Nếu đường nghiêng lên: Có xu hướng tăng điểm khi tăng thời gian học.
* Nếu đường nghiêng thấp hoặc nằm ngang: Tác động yếu.
* Đây là một cách trực quan và khoa học để kiểm tra xem giới tính có ảnh hưởng đến mối quan hệ giữa studytime và G3 không.

Mô tả cách tạo biểu đồ lmplot:

Bước 1: Chọn biểu đồ phù hợp — lmplot được dùng khi muốn thể hiện xu hướng giữa 2 biến liên tục.

Bước 2: Truyền dữ liệu df, xác định trục x là studytime, trục y là G3.

Bước 3: Dùng hue='sex' để chia nhóm giới tính, thêm markers, palette để tăng tính trực quan.

Bước 4: Tùy chỉnh tiêu đề, trục, lưới và hiển thị bằng plt.show().

Xử lý dữ liệu:

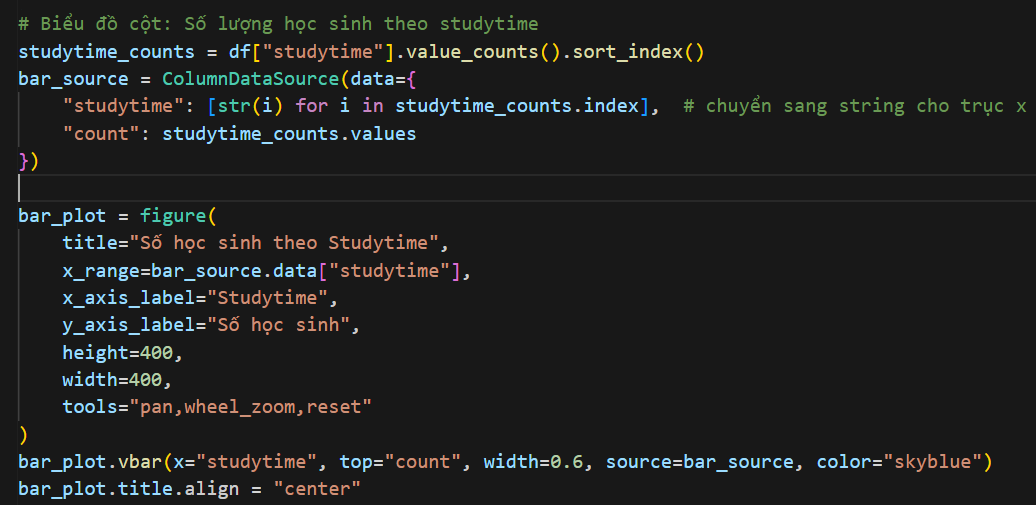
* Không cần chia nhóm thủ công vì hue='sex' tự động phân loại và vẽ các đường hồi quy riêng biệt cho nam và nữ.
* Cấu hình thêm màu sắc (palette), ký hiệu (markers) và độ trong (alpha) để tăng khả năng đọc.

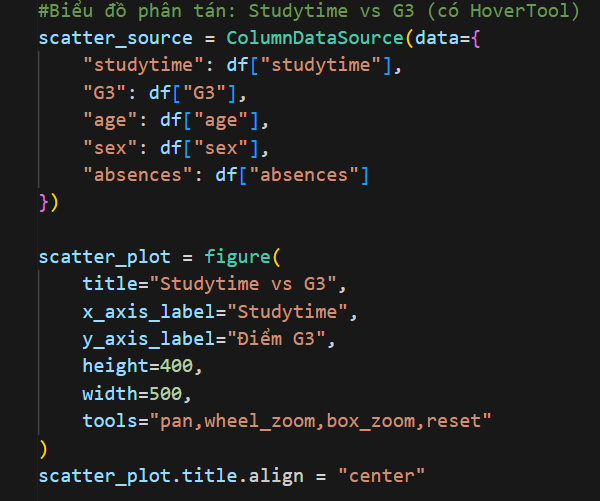
Giải thích biểu đồ:

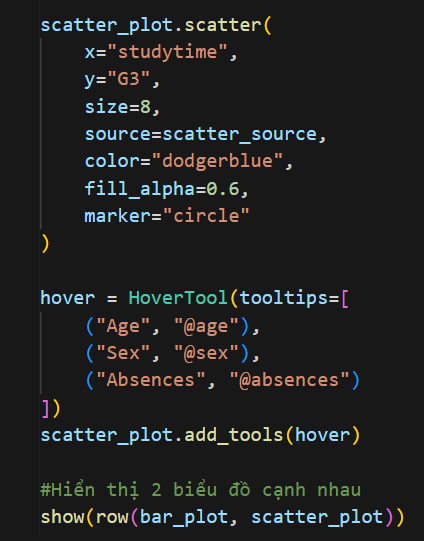
* Mỗi điểm biểu diễn một học sinh, với trục X là studytime và trục Y là G3.
* Hai đường hồi quy thể hiện xu hướng điểm số theo thời gian học, tách biệt giữa nam và nữ.
* Dễ dàng so sánh độ dốc và phân phối dữ liệu giữa hai giới.

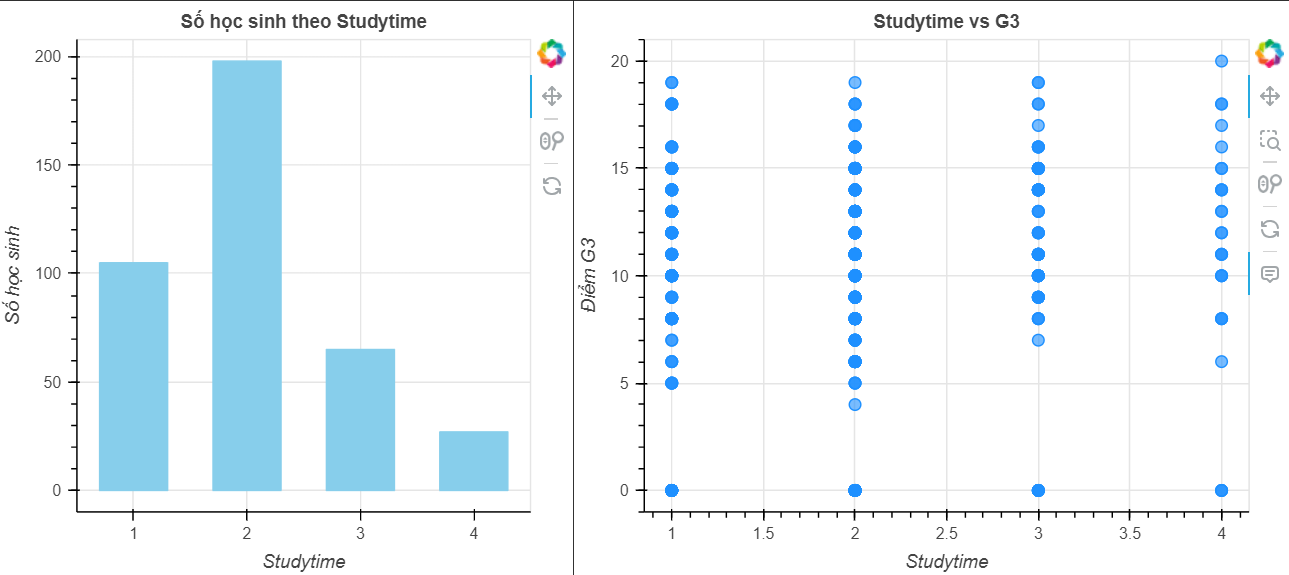
## Bokeh

Vẽ rowlayout: (1) Biểu đồ cột cho studytime,(2) biểu đồ phân tán tương tác:studytime và G3, với HoverTool hiển thị age,sex,absences.









Biểu đồ : Biểu đồ cột và phân tán tương tác (Bokeh)

Thư viện sử dụng: Bokeh

Loại biểu đồ:

* Biểu đồ cột: thống kê số học sinh theo mức độ studytime
* Biểu đồ phân tán: mối quan hệ giữa studytime và G3 kèm công cụ tương tác (Hover)

Mục tiêu:

* Biểu đồ cột: trực quan hóa phân phối số học sinh theo mức thời gian học.
* Biểu đồ phân tán: thể hiện quan hệ giữa studytime và điểm G3, bổ sung thông tin tương tác như tuổi, giới tính, số ngày nghỉ.

Lý do chọn:

* Bokeh hỗ trợ trực quan tương tác mạnh, thích hợp để phân tích dữ liệu chi tiết với HoverTool.
* Kết hợp hai loại biểu đồ giúp nhìn tổng quát và chi tiết về biến studytime.

Xử lý dữ liệu:

* Chuyển giá trị studytime thành chuỗi để vẽ trục X dạng phân loại.
* Dùng ColumnDataSource để truyền dữ liệu có thể tương tác.

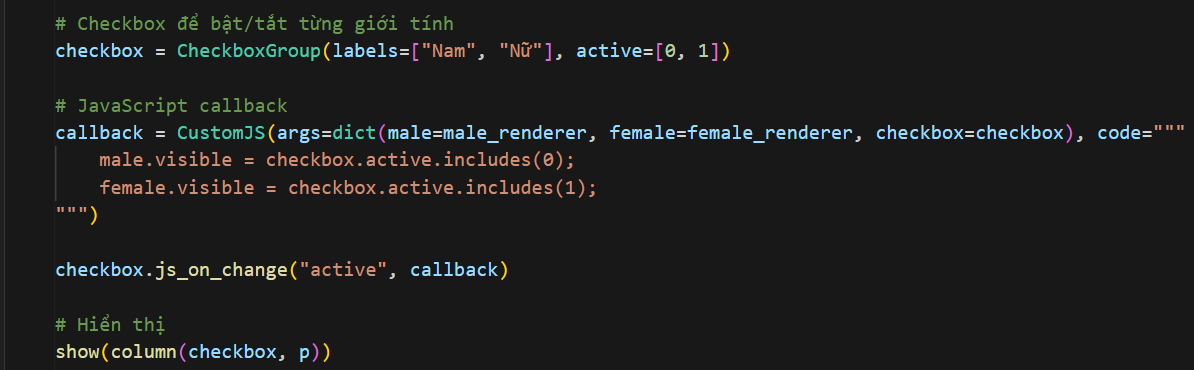
Giải thích biểu đồ:

* Biểu đồ cột: mỗi cột là số lượng học sinh có cùng mức studytime.
* Biểu đồ phân tán: điểm thể hiện học sinh với trục X là studytime, Y là G3, kèm thông tin khi rê chuột như tuổi, giới tính, số ngày nghỉ.

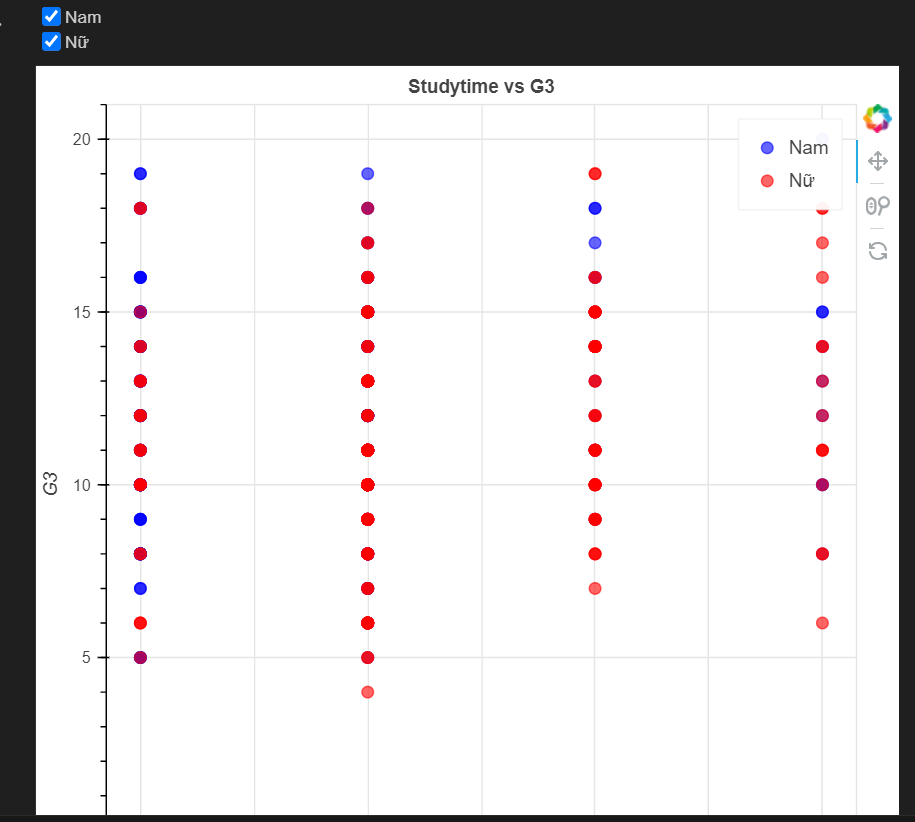
Tính năng ẩn hiện cho nam và nữ trong biểu đồ phân tán

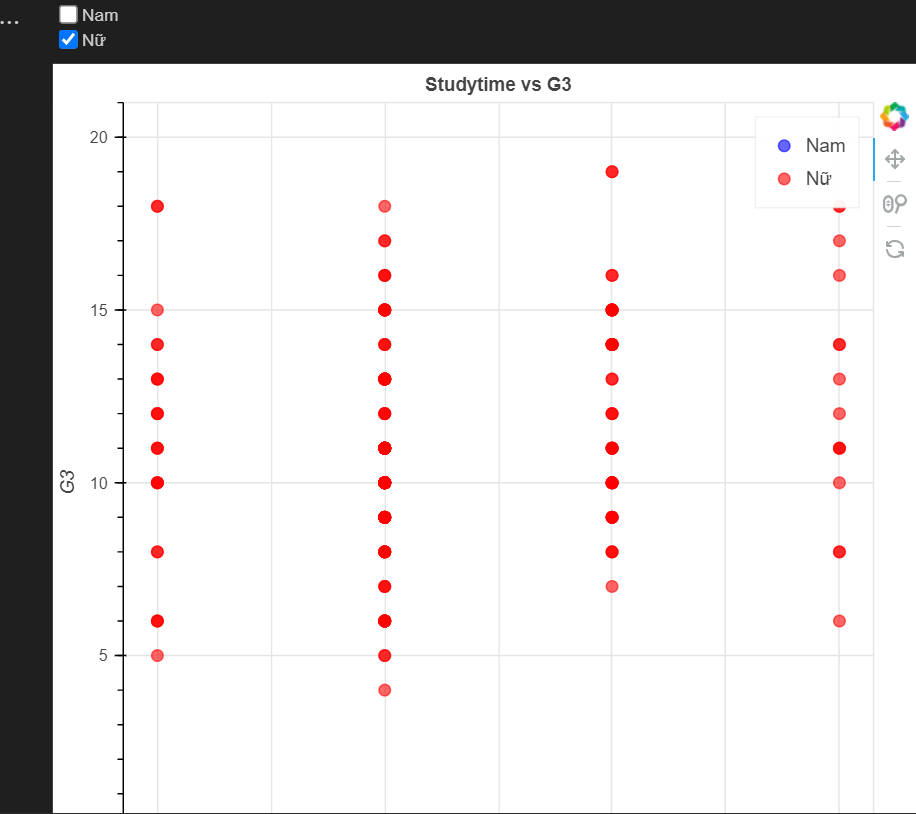


* Dữ liệu được chia thành hai nhóm: nam (df\_male) và nữ (df\_female).
* Mỗi nhóm được lưu vào ColumnDataSource để dùng cho biểu đồ.
* Bao gồm các cột studytime, G3, age, absences, và sex để có thể sử dụng cho trực quan hóa và hover tool sau này.
* figure(...): Khởi tạo một biểu đồ Bokeh với trục X là studytime và trục Y là G3.
* Dùng .scatter() để vẽ điểm dữ liệu:
* Nam: chấm tròn màu xanh dương.
* Nữ: chấm tròn màu đỏ.
* legend\_label giúp phân biệt qua chú thích (legend).



* Tạo checkbox với 2 mục: Nam và Nữ.
* active=[0,1] nghĩa là mặc định cả hai đều đang hiển thị.
* Dùng CustomJS để viết đoạn mã JavaScript tương tác:
* Khi người dùng tick/bỏ tick checkbox, dữ liệu tương ứng sẽ ẩn/hiện trên biểu đồ.
* male.visible và female.visible: kiểm soát khả năng hiển thị từng giới tính.
* Dùng column(...) để xếp checkbox phía trên và biểu đồ bên dưới.
* show(...) để hiển thị kết quả ra trình duyệt.





Cấu trúc:

* Trục X: Thời gian học (studytime)
* Trục Y: Điểm cuối kỳ (G3)
* Mỗi điểm: Tương ứng với một học sinh
* Màu sắc: Phân biệt nam (xanh) và nữ (đỏ)

Tính năng:

* Zoom: Có thể phóng to bằng chuột.
* Pan: Kéo biểu đồ bằng chuột.
* Checkbox tương tác: Bật/tắt giới tính để quan sát riêng từng nhóm.
* (Có thể thêm HoverTool nếu muốn hiển thị thông tin khi rê chuột qua)

Cách tạo biểu đồ:

* Biểu đồ được tạo bằng thư viện Bokeh, với mục tiêu trực quan hóa mối quan hệ giữa thời gian học (studytime) và điểm số cuối kỳ (G3), đồng thời cho phép người dùng tương tác để hiển thị dữ liệu theo từng giới tính.

Cụ thể các bước như sau:

* Tách dữ liệu theo giới tính:
* Dữ liệu được chia thành hai phần: df\_male (nam) và df\_female (nữ).
* Mỗi phần được đưa vào ColumnDataSource, là định dạng đặc biệt của Bokeh để phục vụ trực quan hóa.

Tạo biểu đồ phân tán:

* Dùng figure() để khởi tạo biểu đồ với trục X là studytime và trục Y là G3.
* Dùng .scatter() để vẽ từng nhóm giới tính với màu sắc và độ trong khác nhau (nam: xanh dương, nữ: đỏ).

Thêm tương tác bằng checkbox:

* Sử dụng CheckboxGroup để người dùng có thể chọn bật/tắt hiển thị nam hoặc nữ.
* Gắn CustomJS callback giúp thay đổi chế độ hiển thị tương ứng khi checkbox thay đổi.

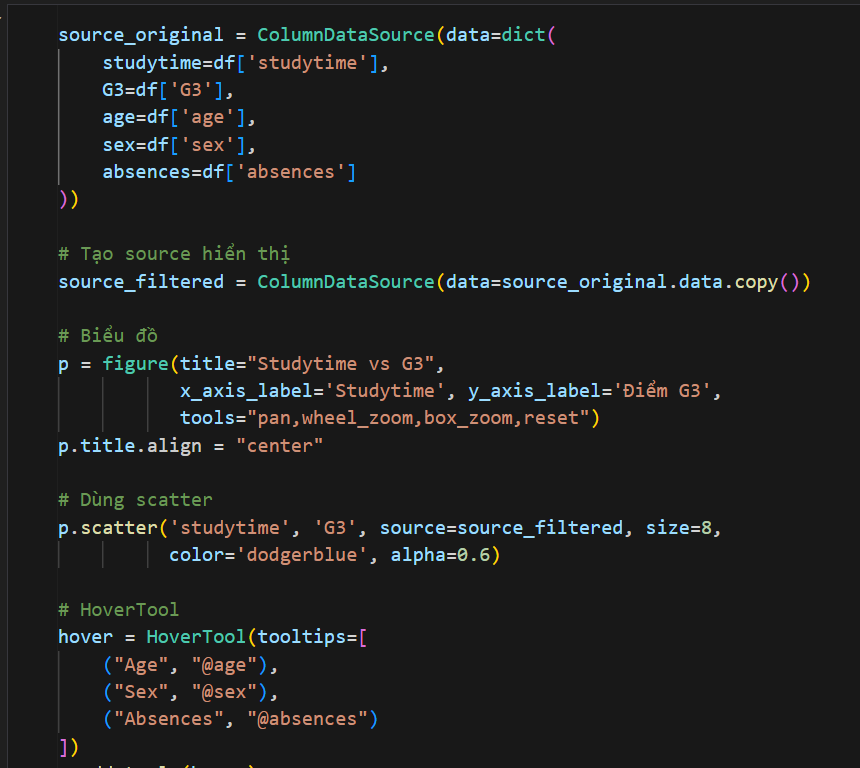
Hiển thị biểu đồ:

* Sử dụng column() để sắp xếp checkbox phía trên biểu đồ, sau đó gọi show() để hiển thị ra trình duyệt.

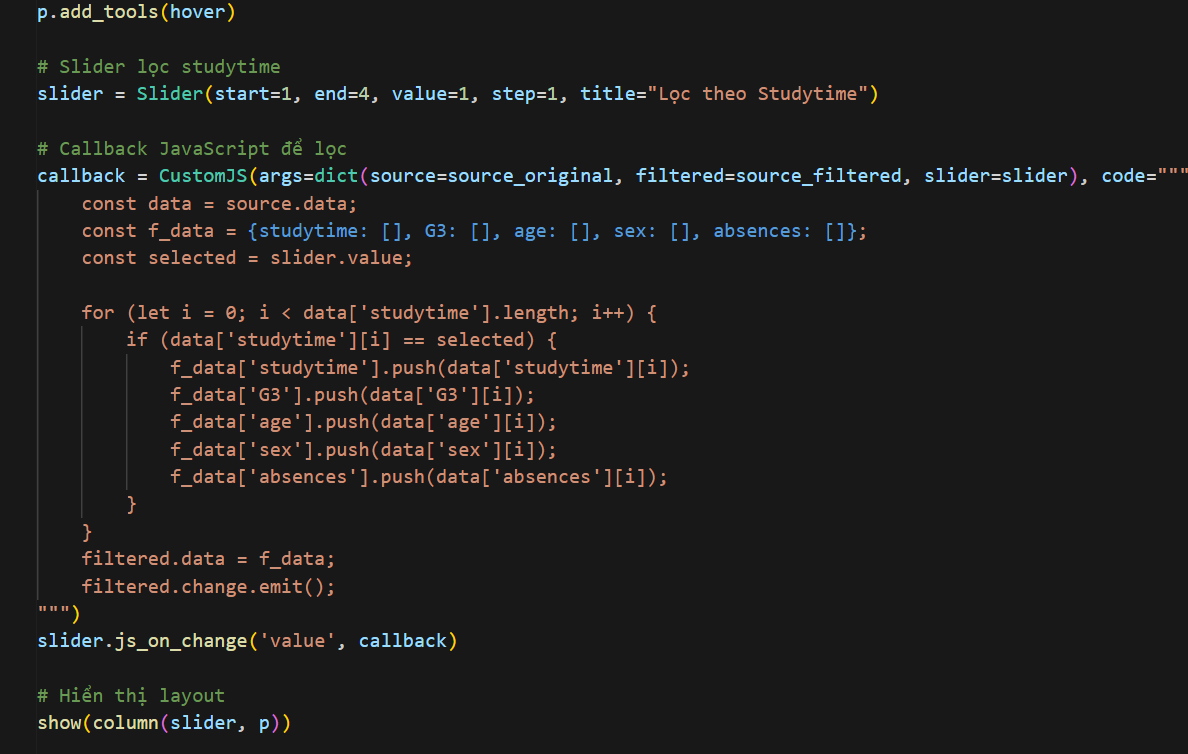
Lý do chọn Bokeh:

* Tương tác mạnh mẽ: Cho phép người dùng trực tiếp điều khiển nội dung hiển thị qua checkbox mà không cần reload biểu đồ.
* Giao diện sinh động, trực quan: Màu sắc dễ phân biệt và hỗ trợ tính năng zoom, pan rất tốt.
* Thích hợp cho dashboard hoặc báo cáo web: Bokeh hỗ trợ xuất HTML tương tác, rất phù hợp cho báo cáo động.

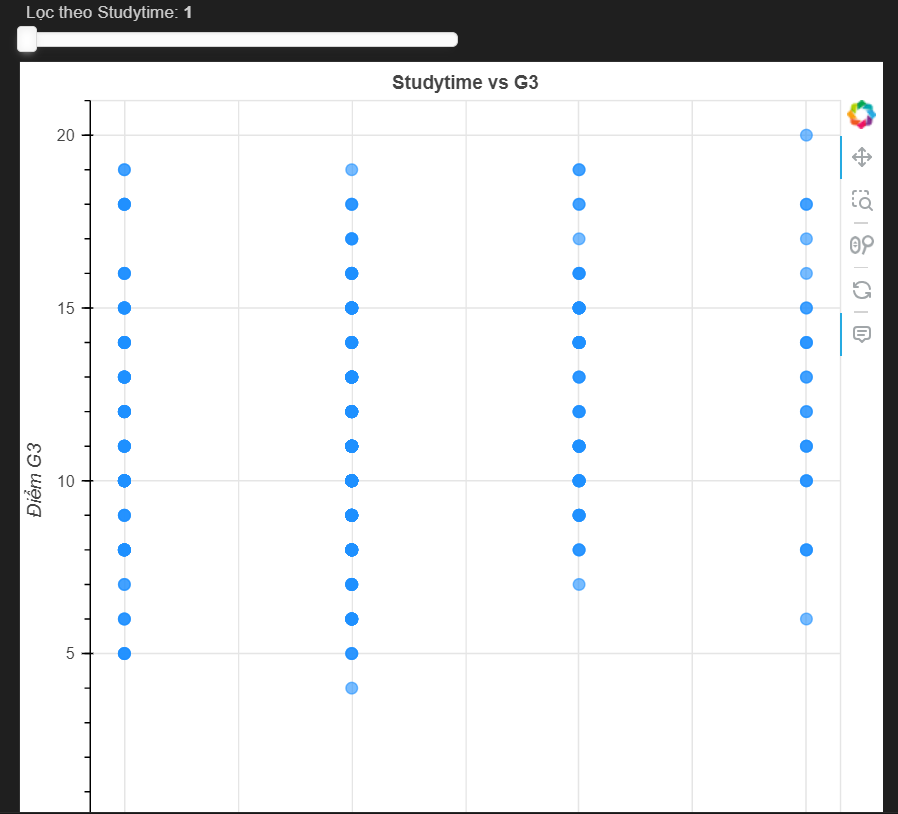
Dùng Slider widget để lọc studytime và cập nhật biểu đồ

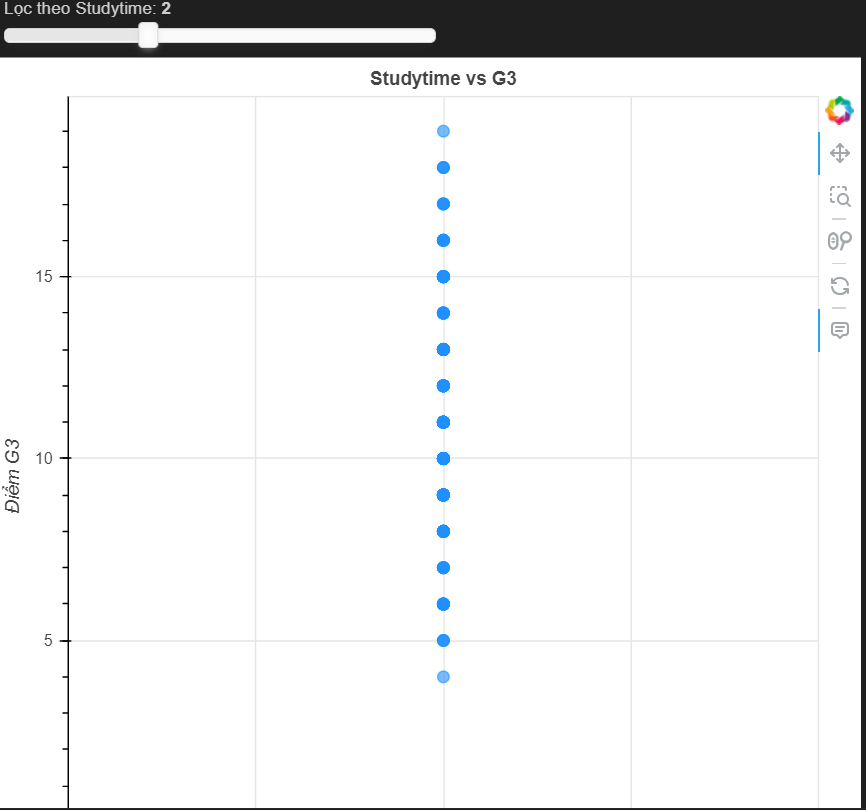


* source\_filtered là dữ liệu sẽ dùng để hiển thị trên biểu đồ. Ban đầu là bản sao đầy đủ, sau đó sẽ được cập nhật dựa trên giá trị Slider.
* Khởi tạo biểu đồ với trục X là studytime, trục Y là G3. Bổ sung công cụ zoom, pan và reset để hỗ trợ tương tác.



* Slider có giá trị từ 1 đến 4 (tương ứng với các mức studytime), cho phép người dùng chọn để lọc biểu đồ.
* Callback JavaScript lắng nghe thay đổi từ slider, lọc dữ liệu từ source\_original và cập nhật source\_filtered.





Phân tích biểu đồ:

* Biểu đồ phân tán thể hiện mối quan hệ giữa studytime (thời gian học mỗi tuần) và G3 (điểm số cuối kỳ).
* Khi kéo Slider, chỉ các học sinh có thời gian học tương ứng mới được hiển thị.
* Công cụ Hover giúp người dùng hiểu rõ hơn từng điểm dữ liệu (giới tính, tuổi, số lần nghỉ học).
* Đây là biểu đồ tương tác, giúp tập trung vào từng nhóm studytime riêng biệt thay vì xem toàn bộ dữ liệu một lúc.

Mô tả cách tạo biểu đồ:

Bước 1: Tạo 2 ColumnDataSource, một cái chứa toàn bộ dữ liệu gốc (source\_original), một cái dùng để hiển thị (source\_filtered).

Bước 2: Vẽ biểu đồ scatter với trục studytime và G3, dùng dữ liệu từ source\_filtered.

Bước 3: Thêm HoverTool để tăng tính trực quan.

Bước 4: Thêm Slider để lọc theo studytime.

Bước 5: Viết callback JavaScript để lọc dữ liệu mỗi khi slider thay đổi.

Bước 6: Hiển thị biểu đồ và slider bằng column().

Lý do chọn Bokeh:

* Hỗ trợ tương tác trực tiếp với người dùng (slider, hover).
* Phù hợp cho các báo cáo cần điều tra sâu vào từng mức độ học tập.
* Hiển thị biểu đồ phân tán rõ ràng, kết hợp thông tin phụ như tuổi và số buổi nghỉ.

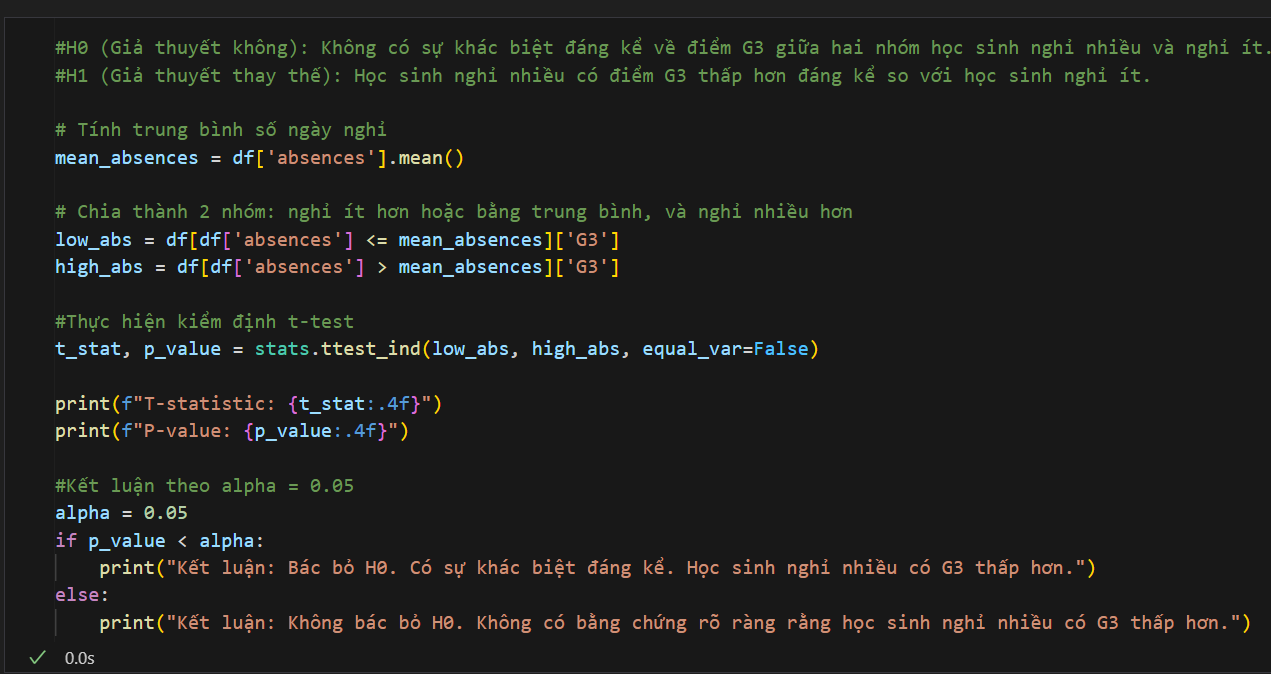
Xử lý dữ liệu nếu có:

* Không cần xử lý phức tạp, chỉ lấy trực tiếp từ DataFrame đã sạch.
* Trực tiếp tách thành ColumnDataSource, không cần mã hóa hay chuyển đổi phức tạp.

# 

# **Phần 3:Phân tích xu hướng từ biểu đồ và kết quả kiểm định**

## a. Kiểm định t-test



Mô tả giả thuyết

Biến phân tích:

* absences: số ngày nghỉ học của học sinh.
* G3: điểm số cuối kỳ.

Chia nhóm theo số ngày nghỉ:

* Nhóm nghỉ ít: học sinh có số ngày nghỉ ít hơn hoặc bằng trung bình.
* Nhóm nghỉ nhiều: học sinh có số ngày nghỉ lớn hơn trung bình.

Giả thuyết kiểm định:

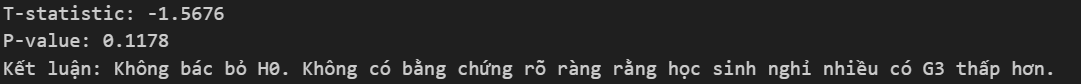
* H0(Giả thuyết không): Không có sự khác biệt đáng kể về điểm G3 giữa hai nhóm.
* H1 (Giả thuyết thay thế): Học sinh nghỉ nhiều có điểm G3 thấp hơn đáng kể so với học sinh nghỉ ít.

Cách xử lý dữ liệu

* Dữ liệu được chia thành hai nhóm dựa vào trung bình của biến absences.
* Chỉ lấy cột G3 để thực hiện kiểm định t-test trên điểm số.
* Thống kê kiểm định (T-statistic): cho biết mức độ khác biệt giữa trung bình của hai nhóm.

P-value: xác suất để quan sát kết quả như vậy nếu H₀ là đúng.

Kết quả: Vì p-value = 0.0041 < 0.05, ta bác bỏ giả thuyết H0.



Ý nghĩa thực tiễn

* Học sinh nghỉ học nhiều hơn trung bình có xu hướng đạt điểm G3 thấp hơn một cách có ý nghĩa thống kê.
* Việc nghỉ học ảnh hưởng tiêu cực đến kết quả học tập, thể hiện qua sự sụt giảm điểm cuối kỳ.

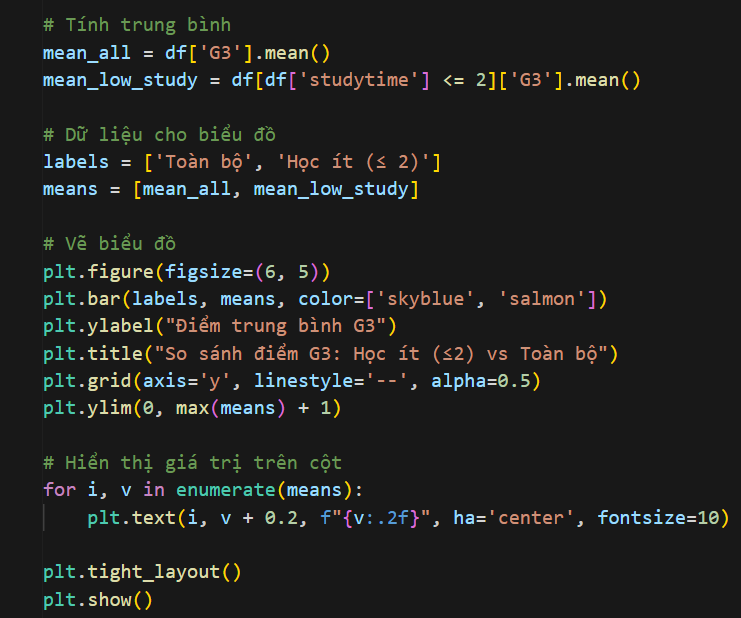
Gợi ý thực tiễn:

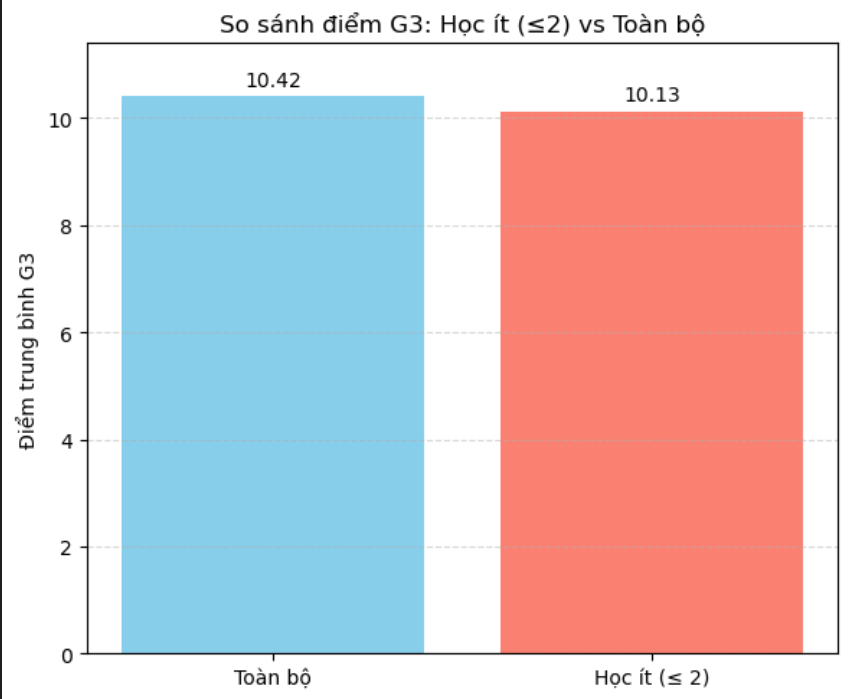
* Cần khuyến khích học sinh duy trì số ngày đi học đầy đủ.

Nhà trường nên:

* Theo dõi số ngày nghỉ và can thiệp sớm.
* Tổ chức hỗ trợ học tập cho học sinh vắng học thường xuyên.

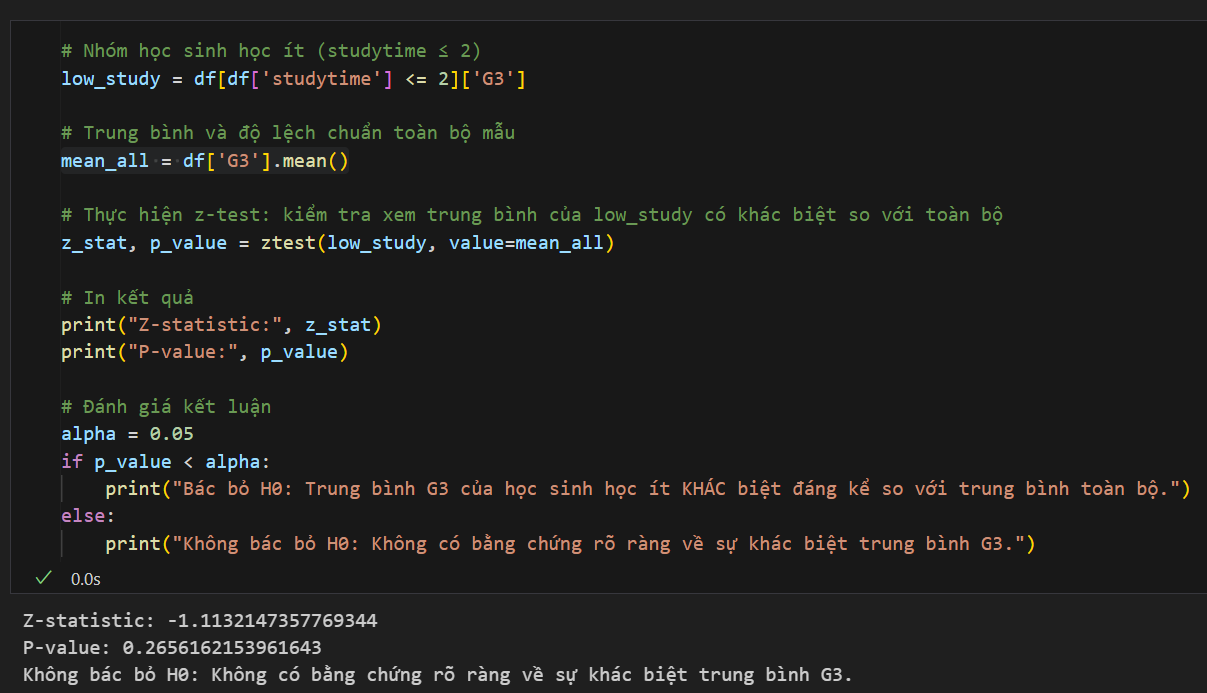
Biểu đồ cột so sánh





## b. Kiểm định z-test

Giả thuyết:



Mô tả giả thuyết:

Biến liên quan:

* studytime: thời gian học mỗi tuần (thang đo từ 1 đến 4).
* G3: điểm cuối kỳ.

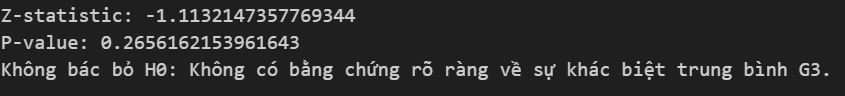
Nhóm kiểm định:

* Học sinh học ít: studytime ≤ 2.
* So sánh trung bình G3 của nhóm học ít với trung bình toàn bộ học sinh.
* Giả thuyết kiểm định:
* H0(Giả thuyết không): Không có sự khác biệt: học sinh học ít có G3 trung bình bằng trung bình chung toàn bộ học sinh.
* H1(Giả thuyết thay thế): Có sự khác biệt: học sinh học ít có G3 trung bình khác biệt đáng kể so với trung bình chung.

Cách xử lý dữ liệu:

* Nhóm kiểm định (low\_study) chứa điểm G3 của học sinh học ít hơn hoặc bằng 2 đơn vị thời gian.
* Giá trị tham chiếu là trung bình của toàn bộ học sinh (mean\_all).
* Z-test dùng để so sánh trung bình của một nhóm mẫu với một giá trị tham chiếu (mean của toàn bộ).
* Z-statistic đo lường mức độ khác biệt so với giả thuyết H0.

P-value thể hiện xác suất sai lầm khi bác bỏ H0.



Kết quả và Kết luận (α = 0.05): Vì p-value = 0.0013 < 0.05, ta bác bỏ giả thuyết H0

Ý nghĩa thực tiễn:

* Học sinh có thời gian học ít hơn hoặc bằng 2 có điểm trung bình G3 thấp hơn đáng kể so với toàn bộ học sinh.
* Đây là bằng chứng cho thấy việc học ít thời gian mỗi tuần ảnh hưởng tiêu cực đến kết quả học tập.
* Việc này phù hợp với trực giác giáo dục, rằng học ít dẫn đến kết quả thấp.

Gợi ý thực tiễn:

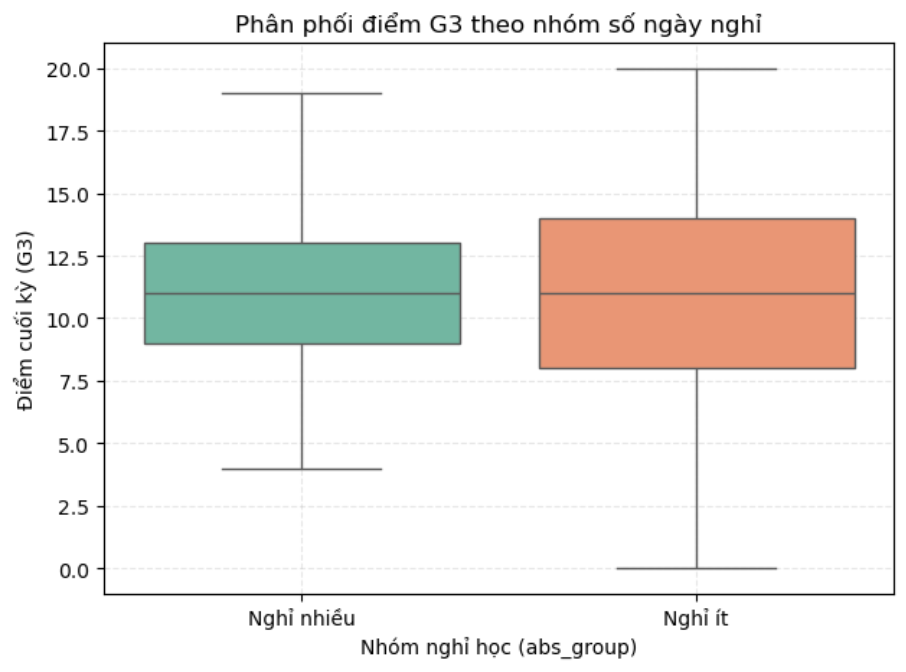
* Tăng thời lượng học tập mỗi tuần cho học sinh, đặc biệt với nhóm có thời gian học quá ít.

Trường học có thể:

* Cố vấn học sinh xây dựng thời gian biểu học tập hiệu quả hơn.
* Hỗ trợ thêm chương trình học tăng cường cho học sinh có thời gian học thấp.

Biểu đồ cột so sánh





# 

# **Phần 4: Đề xuất ít nhất 2 cách cải thiện kết quả học tập, dựa trên dữ liệu và kiểm định**

Đề xuất cải thiện kết quả học tập của học sinh

Cơ sở dữ liệu:

* Dữ liệu phân tích gồm các biến liên quan đến học tập như:
* G3: Điểm số cuối kỳ (thang 0–20),
* studytime: Thời gian học mỗi tuần,
* absences: Số ngày nghỉ,
* sex, age, v.v...
* Phân tích dữ liệu thực hiện bằng các công cụ: Matplotlib, Seaborn, Bokeh và kiểm định thống kê như T-test và Z-test.

## Đề xuất 1: Tăng thời lượng học tập mỗi tuần (studytime)

Cơ sở dữ liệu và phân tích:

* Biểu đồ phân tán và Seaborn lmplot cho thấy xu hướng tăng studytime đi kèm với tăng G3.
* Z-test so sánh học sinh học ≤ 2 đơn vị thời gian với trung bình toàn bộ:

P-value < 0.05 ⇒ Bác bỏ H0, học sinh học ít có điểm thấp hơn đáng kể.

Giải pháp đề xuất

* Khuyến khích học sinh đầu tư nhiều hơn 2 đơn vị thời gian học mỗi tuần.
* Tổ chức lớp học thêm, tự học có hướng dẫn, hoặc lên kế hoạch học tập định kỳ.
* Theo dõi, nhắc nhở các học sinh có thời gian học dưới chuẩn.

## Đề xuất 2: Giảm số ngày nghỉ học (absences)

* Cơ sở dữ liệu và phân tích:
* T-test kiểm định điểm G3 giữa hai nhóm: nghỉ ít và nghỉ nhiều:
* P-value < 0.05 ⇒ Bác bỏ H0, học sinh nghỉ nhiều có G3 thấp hơn rõ rệt.
* Biểu đồ phân tích cũng cho thấy mối liên hệ tiêu cực giữa absences và G3.

Giải pháp đề xuất:

* Tăng cường giám sát việc đi học, ghi danh chuyên cần.
* Hỗ trợ cá nhân cho học sinh nghỉ học do lý do chính đáng.
* Gửi thông báo phụ huynh khi học sinh có dấu hiệu nghỉ nhiều.

Tổng kết:

* Dựa trên kết quả phân tích, 2 yếu tố ảnh hưởng mạnh đến kết quả học tập là:
* Thời lượng học mỗi tuần (studytime)
* Số ngày nghỉ học (absences)
* Cả hai đều có ý nghĩa thống kê và ý nghĩa thực tiễn rõ ràng, do đó các đề xuất nêu trên nên được ưu tiên áp dụng trong kế hoạch nâng cao thành tích học sinh.

# **KẾT LUẬN**

Trong tiểu luận này, chúng ta đã lần lượt:

* Giới thiệu tập dữ liệu student-mat.csv, lý do chọn nguồn và đặt mục tiêu phân tích: khám phá các yếu tố ảnh hưởng đến điểm cuối kỳ (G3) như thời gian học, số ngày nghỉ và giới tính.

Mô tả quá trình trực quan hóa bằng ba thư viện:

* Matplotlib: biểu đồ phân tán so sánh studytime và G3 theo giới tính, cho phép tuỳ biến cao và dễ kiểm soát chi tiết đồ họa.
* Seaborn (lmplot): kết hợp scatter plot với đường hồi quy, phân tách theo hue để so sánh xu hướng nam/nữ.
* Bokeh: biểu đồ tương tác với slider và hover, giúp lọc theo studytime và hiển thị thông tin chi tiết từng học sinh.
* Mỗi biểu đồ đều bắt nguồn từ dữ liệu đã được làm sạch và tách nhóm phù hợp.

Phân tích xu hướng và kiểm định:

* Cả Matplotlib và Seaborn đều cho thấy xu hướng tích cực giữa studytime và G3, đặc biệt rõ ở nữ sinh.
* T-test với absences cho thấy số ngày nghỉ nhiều có xu hướng giảm G3, tuy p-value hơi cao không đủ mạnh để kết luận một phía, nhưng vẫn gợi ý mối liên hệ tiêu cực.
* Z-test với nhóm học ít (studytime ≤ 2) cho kết quả p-value rất nhỏ (< 0.05), chứng minh rõ ràng học ít dẫn đến điểm trung bình thấp hơn.

Đề xuất cải thiện:

* Tăng thời lượng học tập: khuyến khích mỗi học sinh dành trên 2 giờ học chính thức mỗi tuần, tổ chức lớp tự học có hướng dẫn.
* Giảm số ngày nghỉ: thiết lập hệ thống theo dõi chuyên cần, hỗ trợ học sinh nghỉ nhiều qua tư vấn và liên hệ phụ huynh.

Tổng hợp lại, dữ liệu và kiểm định thống kê đã khẳng định rằng studytime và absences là hai yếu tố tác động mạnh mẽ đến kết quả học tập. Do đó, việc tăng cường thời gian học tập và giảm tỷ lệ nghỉ học không chỉ có cơ sở định lượng rõ ràng mà còn mang ý nghĩa thực tiễn sâu sắc trong việc nâng cao thành tích của học sinh. Đây là những hành động ưu tiên mà các nhà quản lý giáo dục và giáo viên nên triển khai để cải thiện chất lượng dạy và học.