

ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN



TIÊU LUẬN PHÂN TÍCH DỮ LIỆU

NGÀNH: CÔNG NGHỆ THÔNG TIN

SINH VIÊN THỰC HIỆN: TRẦN HẢO ĐIỀN

MSSV: 3122411042

LỚP: DCT122C3

GVHD: PGS TS. NGUYỄN TUẤN ĐĂNG

TP. HỒ CHÍ MINH, THÁNG ... NĂM

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình báo cáo của cá nhân tôi. Các nội dung nghiên cứu trong báo cáo “Tiêu Luận Phân Tích Dữ Liệu” của tôi là trung thực. Nếu có phát hiện bất kỳ sự gian lận nào thì tôi xin hoàn toàn chịu trách nhiệm về bài báo cáo của tôi.

Trần Hảo Điện

LỜI CẢM ƠN

Lời cảm ơn chân thành đến các đọc giả đã dành thời gian để quan tâm đến bài báo của tôi.

MỤC LỤC

MỤC LỤC.....	0
MỤC LỤC ANH.....	1
CHƯƠNG 1. GIỚI THIỆU	3
1.1. Tổng quan về tập dữ liệu.....	3
a) Giới thiệu tập dữ liệu	3
b) Lý do chọn file student-mat.csv cho việc phân tích môn Toán	4
c) Mục tiêu phân tích	5
1.2. Các công việc phân tích	5
a) Tải và đọc dữ liệu bằng Pandas	5
b) Phân tích cơ bản.....	7
c) Báo cáo ngắn về đặc điểm dữ liệu	12
Chương 2. TRỰC QUAN HÓA DỮ LIỆU	17
2.1. Yêu cầu	17
2.2. Kết quả	17
a) Matplotlib.....	17
b) Seaborn	22
c) Bokeh	29
CHƯƠNG 3. PHÂN TÍCH XU HƯỚNG VÀ KIỂM ĐỊNH	39
3.1. Yêu cầu	39
3.2. Kết quả	39
a) T-test	39
b) Z-test.....	41
c) Chi-square	44
d) Phân tích xu hướng các biểu đồ.....	48
CHƯƠNG 4. ĐỀ XUẤT PHƯƠNG PHÁP CẢI THIỆN KẾT QUẢ	49
4.1. Phương pháp	49
a) Phương pháp 1	49
b) Phương pháp 2	49
CHƯƠNG 5. TỔNG KẾT	52
5.1. Tổng kết	52
a) Chương 1	52
b) Chương 2	52
c) Chương 3	53
d) Chương 4	54
e) Tổng kết bài	54
TÀI LIỆU THAM KHẢO	56

MỤC LỤC ẢNH

Hình 1.2.1. Các thư viện cần dùng trong quá trình nghiên cứu	6
Hình 1.2.2. Tải tệp và đọc dữ liệu trên hệ thống.....	6
Hình 1.2.3. Thông tin của tệp dữ liệu	7
Hình 1.2.6. Kiểm tra giá trị bị thiếu trong dữ liệu	7
Hình 1.2.7. Kết quả của kiểm tra	7
Hình 1.2.4. Phân tích cơ bản của các đặc trưng dữ liệu.....	12
Hình 1.2.5. Kiểu dữ liệu trong tệp dữ liệu student-mat.csv.....	14
Hình 2.2.1. Mã nguồn tạo biểu đồ cột theo "Thời gian học" và "Số lượng học sinh"	17
Hình 2.2.2. Biểu đồ cột theo mã nguồn	18
Hình 2.2.3. Mô tả biểu đồ	18
Hình 2.2.4. Mã nguồn tạo biểu đồ phân tán theo "Studytime" và "G3" với đa màu sắc	20
Hình 2.2.5. Mã nguồn tạo biểu đồ hồi quy tuyến tính theo giới tính.....	23
Hình 2.2.6. Biểu đồ hồi quy tuyến tính theo mã nguồn trên.....	23
Hình 2.2.7. Mô tả biểu đồ	24
Hình 2.2.8. Mã nguồn biểu đồ hộp phân phối điểm "G3" theo "Studytime".....	25
Hình 2.2.9. Biểu đồ hộp với mức điểm "G3" và "Studytime"	26
Hình 2.2.10. Mã nguồn tạo biểu đồ xây dựng mối quan hệ giữa "G1","G2","G3" và "Studytime"	27
Hình 2.2.11. Biểu đồ mối quan hệ giữa "G1", "G2", "G3" và "Studytime"	28
Hình 2.2.12. Mã nguồn biểu đồ phân tán tương tác theo "Studytime" và "G3".....	30
Hình 2.2.13. Biểu đồ phân tán tương tác với chức năng HoverTool	31
Hình 2.2.14. Biểu đồ phân tán tương tác với chức năng Click Policy (Tắt dữ liệu giới tính nữ F)	31
Hình 2.2.15. Biểu đồ phân tán tương tác với chức năng Click Policy (Tắt dữ liệu giới tính nam M).....	32
Hình 2.2.16 Biểu đồ phân tán tương tác với chức năng Click Policy (Tắt hết dữ liệu).....	32
Hình 2.2.17. Biểu đồ phân tán tương tác theo mã nguồn trên với studytime là 1	33
Hình 2.2.18. Biểu đồ phân tán tương tác với studytime là 2	33
Hình 2.2.19. Biểu đồ phân tán tương tác với studytime là 3	34
Hình 2.2.20. Biểu đồ phân tán tương tác với studytime là 4	34
Hình 2.2.21. Mô tả biểu đồ (1).....	35
Hình 2.2.22. Mô tả biểu đồ (2).....	35
Hình 2.2.23. Biểu đồ cột cho studytime và biểu đồ phân tán tương tác theo định dạng row layout.....	37
Hình 2.2.24. Biểu đồ cột với "Studytime" và biểu đồ phân tán tương tác giữa "G3" và "Studytime"	38
Hình 3.2.1. Mã nguồn kiểm định T-Test	39
Hình 3.2.2. Giải thích kí hiệu cần dùng trong tính toán T-Test	40
Hình 3.2.3. Kết quả kiểm định T-Test	41
Hình 3.2.4. Mã nguồn kiểm định Z-Test	42
Hình 3.2.5. Kí hiệu cần dùng trong tính toán Z-Test	42
Hình 3.2.6. Giải thích phép toán trong mã nguồn Z-Test	43
Hình 3.2.7. Kết quả của kiểm định Z-Test.....	44
Hình 3.2.8. Mã nguồn phương pháp kiểm định Chi-square.....	45
Hình 3.2.9. Tạo biến nhị phân.....	46
Hình 3.2.10. Giá trị mẫu từ các biến nhị phân	46
Hình 3.2.11. Tạo bảng tần số	46
Hình 3.2.12. Kết quả của bảng tần số	46
Hình 3.2.13. Tính toán kiểm định Chi-square	47

Hình 3.2.14. Đưa ra kết luận cuối cùng	47
Hình 3.2.15. Kết luận của kết quả kiểm định Chi-square tìm mối liên hệ giữa Studytime và G3	47
Hình 4.1.1. Biểu đồ lý giải cho đề xuất phương án 1	49
Hình 4.1.2. Mã nguồn cho đề xuất phương pháp 2.....	50
Hình 4.1.3. Biểu đồ thể hiện cho mức điểm trung bình giữa các nhóm nghỉ học	50

CHƯƠNG 1. GIỚI THIỆU

1.1. Tổng quan về tập dữ liệu

a) Giới thiệu tập dữ liệu

- Tập dữ liệu được sử dụng trong nghiên cứu này là **student-mat.csv**, được lấy từ **UCI Machine Learning Repository** – một nguồn dữ liệu uy tín được sử dụng rộng rãi trong cộng đồng nghiên cứu học máy và khoa học dữ liệu. Dữ liệu này tiếp cận thành tích của học sinh trong giáo dục trung học của hai trường Bồ Đào Nha. Các thuộc tính dữ liệu bao gồm điểm số của học sinh, nhân khẩu học, xã hội và các đặc điểm liên quan đến trường học và nó được thu thập bằng cách sử dụng các báo cáo và bảng câu hỏi của trường. Tập dữ liệu này chứa thông tin chi tiết về **395 học sinh** đang học môn Toán tại hai trường trung học ở Bồ Đào Nha. Dữ liệu được thu thập trong khuôn khổ một nghiên cứu nhằm phân tích các yếu tố ảnh hưởng đến kết quả học tập của học sinh, được lưu với định dạng file CSV.
- Mỗi dòng dữ liệu đại diện cho một học sinh, với **33 đặc trưng (cột)**:
 - + **G1, G2, G3**: Điểm số học kỳ 1, học kỳ 2, và học kỳ cuối (thang điểm 0–20).
 - + **studytime**: Thời gian học mỗi tuần
 - . 1: < 2 giờ
 - . 2: 2-5 giờ
 - . 3: 5-10 giờ
 - . 4: >10 giờ
 - + **absences**: Số ngày nghỉ học (từ 0 đến 93 ngày).
 - + **sex**: Giới tính học sinh (M: nam/F: nữ).
 - + **age**: Tuổi của học sinh (dao động từ 15 đến 22 tuổi)
 - + **freetime**: Thời gian rảnh rỗi (1-5 tiếng).
 - + **school**: Trường học sinh theo học:
 - . GP: Gabriel Pereira
 - . MS: Mousinho da Silvera
 - + **address**: Nơi ở của học sinh:
 - . U: “urban” là thành phố
 - . R: “rural” là nông thôn
 - + **famsize**: Số lượng thành viên trong gia đình:
 - . GT3: “**greater than 3**” có nghĩa là nhiều hơn hoặc bằng 3 thành viên
 - . LE3: “**less than 3**” có nghĩa là ít hơn hoặc bằng 3 thành viên
 - + **Pstatus**: Tình trạng sống cùng phụ huynh:
 - . A: Không sống cùng nhau
 - . T: Sống cùng nhau
 - + **Medu,Fedu**: Trình độ học vấn của phụ huynh (cha hoặc mẹ).
 - . 0: Không đi học, chỉ có hai trường hợp của hai học sinh khác nhau
 - . 1: Đi học hết cấp 1 (từ lớp 1 – lớp 5)
 - . 2: Đi học hết cấp 2 (từ lớp 6 – lớp 9)
 - . 3: Đi học hết cấp 3 (từ lớp 10 – lớp 12)
 - . 4: Đi học hết đại học (4 năm)
 - + **Mjob,Fjob**: Công việc của phụ huynh
 - . at_home: Công việc tại nhà
 - . teacher: Giáo viên
 - . services: Làm phục vụ, dịch vụ
 - . health: Làm y khoa, mảng sức khỏe

- . other: Các ngành nghề khác
- + **Reason:** Lý do chọn ngôi trường để theo học:
 - . course: Do chương trình, khóa học
 - . home: Lý do gần nhà
 - . reputation: Danh tiếng của ngôi trường
 - . other: Các lý do khách quan khác
- + **Guardian:** Người giám hộ của học sinh là cha hoặc mẹ (father/mother)
- + **Travelttime:** Thời gian đi đến trường của học sinh:
 - . 1: <15 phút
 - . 2: 15-30 phút
 - . 3: 30 phút – 1 tiếng
 - . 4: >1 tiếng
- + **Failures:** Số lần học sinh rớt lớp trước đó từ 0 – 3 lớp
- + **Schoolsup, famsup:** Hỗ trợ học tập từ gia đình và nhà trường (yes/no)
- + **Paid:** Lớp học thêm có cần phải trả thêm học phí không (yes/no)
- + **Activities:** Hoạt động ngoại khóa (yes/no)
- + **Nursery:** Có học trường mẫu giáo không (yes/no)
- + **Higher:** Dự định muốn học cao hơn (yes/no)
- + **Internet:** Có kết nối internet ở nhà (yes/no)
- + **Romantic:** Có mối quan hệ tình cảm (yes/no)
- + **Famrel:** Tình trạng mối quan hệ trong gia đình (từ 1 là rất tệ đến 5 là rất tốt)
- + **Goout:** Đi ra ngoài chơi với bạn bè (từ 1 là rất ít đến 5 là rất nhiều)
- + **Dalc:** Tiêu thụ rượu trong ngày làm việc (từ 1 là rất ít đến 5 là rất nhiều)
- + **Walc:** Tiêu thụ rượu vào cuối tuần (từ 1 là rất ít đến 5 là rất nhiều)
- + **Health:** Tình trạng sức khỏe của học sinh (từ 1 là rất kém đến 5 là rất khỏe)

b) Lý do chọn file student-mat.csv cho việc phân tích môn Toán

Trong bối cảnh giáo dục hiện đại ngày càng chú trọng đến việc cá nhân hóa quá trình học tập và đưa ra các quyết định dựa trên dữ liệu, việc lựa chọn một bộ dữ liệu phù hợp là bước đầu tiên và quan trọng nhất để đảm bảo chất lượng của quá trình phân tích. Dữ liệu **student-mat.csv** đã được tôi lựa chọn làm cơ sở nghiên cứu vì các lý do sau:

1. **Tập trung sâu vào một môn học cụ thể:** Không giống như những bộ dữ liệu tổng hợp về kết quả học tập nhiều môn học khác nhau, “**student-mat.csv**” tập trung chuyên biệt vào môn Toán – một môn học được xem là nền tảng cho tư duy logic, phân tích và giải quyết vấn đề. Việc chỉ phân tích một môn học giúp loại bỏ các yếu tố gây nhiễu từ các môn khác, từ đó cho phép đi sâu vào việc tìm hiểu nguyên nhân và tác động cụ thể đến hiệu quả học tập của học sinh trong môn Toán.
2. **Bức tranh toàn diện về học sinh:** Một điểm nổi bật khác của bộ dữ liệu là sự kết hợp giữa dữ liệu học tập và dữ liệu phi học thuật. Bên cạnh các điểm số học kỳ (G1, G2, G3), tập dữ liệu còn ghi nhận các yếu tố như tình trạng gia đình, nghề nghiệp của cha mẹ, mức độ tiêu thụ rượu, thói quen học tập, thời gian rảnh rỗi và tình trạng đi học chuyên cần. Điều này mang đến một góc nhìn đa chiều, giúp tôi có thể khám phá những mối liên hệ tiềm ẩn giữa yếu tố xã hội và kết quả học tập, từ đó có thể đưa ra các hướng can thiệp chính sách giáo dục toàn diện.
3. **Tính ứng dụng cao trong thực tiễn giáo dục:** Với thông tin phong phú và chân thực, tập dữ liệu student-mat.csv không chỉ phù hợp để phục vụ các mô hình dự đoán học máy mà còn rất hữu ích cho các nghiên cứu giáo dục ứng dụng, chẳng hạn như thiết kế chương trình hỗ trợ học sinh yếu, điều chỉnh phương pháp giảng dạy phù hợp với từng nhóm học sinh, hoặc tư vấn học đường dựa trên dữ liệu. Như vậy, các kết quả rút ra từ tập dữ liệu này không chỉ mang tính học thuật mà còn có thể chuyển hóa thành các chiến lược cải thiện học tập thực tế.

4. **Khả năng theo dõi tiến trình học tập:** Việc bộ dữ liệu bao gồm cả ba giai đoạn điểm số (G1 – đầu kỳ, G2 – giữa kỳ, G3 – cuối kỳ) cho phép xây dựng một phát họa cụ thể theo thời gian về sự tiến bộ hoặc thụt lùi trong học tập của từng học sinh. Đây là một đặc điểm cực kỳ quý giá khi nghiên cứu về quá trình học tập, vì nó cho phép kiểm tra không chỉ kết quả cuối cùng mà còn cả quá trình phát triển và động lực học tập của học sinh.
5. **Tính sẵn sàng và minh bạch:** Dữ liệu được cung cấp công khai, có nguồn gốc rõ ràng, đã được sử dụng trong nhiều nghiên cứu trước đó, tạo điều kiện cho việc so sánh, kiểm định và tái sử dụng mô hình, đồng thời đảm bảo tính minh bạch và đáng tin cậy trong toàn bộ quá trình nghiên cứu.

=> Chính nhờ những đặc điểm nổi bật kể trên, tập dữ liệu “student-mat.csv” trở thành một lựa chọn lý tưởng cho việc phân tích và dự đoán kết quả học tập môn Toán. Thông qua đó, chúng ta không chỉ có thể hiểu rõ hơn về các yếu tố ảnh hưởng đến thành tích học tập của học sinh, mà còn có thể đề xuất các giải pháp thiết thực nhằm nâng cao chất lượng giảng dạy và tối ưu hóa quá trình học tập trong nhà trường.

c) Mục tiêu phân tích

- Khám phá mối tương quan giữa thời gian học tập, số ngày nghỉ và thời gian rảnh với kết quả học tập môn Toán, nhằm xác định các yếu tố đó theo thời gian có ảnh hưởng tích cực hoặc tiêu cực đến thành tích học tập của học sinh.
- Phân tích tác động của các yếu tố phi học thuật, chẳng hạn như mức độ vắng mặt và thói quen sử dụng thời gian rảnh đến việc sử dụng thức uống có cồn,...để từ đó đánh giá vai trò của kỷ luật học tập và cân bằng sinh hoạt trong quá trình học.
- Đề xuất các biện pháp cải thiện kết quả học tập dựa trên dữ liệu thực tế và kiểm định thống kê, giúp nhà trường và giáo viên định hướng chiến lược giảng dạy hiệu quả hơn, đồng thời hỗ trợ học sinh phát triển thói quen học tập khoa học và hợp lý.

1.2. Các công việc phân tích

a) Tải và đọc dữ liệu bằng Pandas

- Các thư viện cần dùng trong quá trình nghiên cứu nhằm có thể khai thác triệt để thông tin và đúc kết những kinh nghiệm thông qua dữ liệu, gồm các thư viện chính như:

- + **Pandas:** dùng cho việc đọc dữ liệu, tải nội dung từ các file với nhiều định dạng khác nhau.
- + **SciPy.stats:** dùng trong quá trình thực hiện các kiểm định thống kê (t-test, z-test, chi square).
- + **Matplotlib và Seaborn:** thư viện được dùng cho việc tạo các mô hình và biểu đồ tĩnh.
- + **Bokeh:** thư viện được dùng cho việc tạo các mô hình và biểu đồ tương tác với người dùng.
- + **NumPy:** thư viện được sử dụng cho mục đích tính toán.

```
● ● ●  
1 import findspark  
2  
3 import matplotlib.pyplot as plt  
4 import matplotlib.patches as mpatches  
5 import seaborn as sns  
6  
7 from pyspark.sql import functions as F  
8 from functools import reduce  
9  
10 from bokeh.plotting import figure, show, output_notebook,  
    curdoc  
11 from bokeh.models import ColumnDataSource, HoverTool, Label  
    ISet, Slider, CustomJS  
12  
13 from bokeh.layouts import column, row  
14  
15 from bokeh.palettes import Bright3  
16  
17 import numpy as np  
18 from scipy.stats import ttest_ind  
19 from scipy.stats import norm  
20  
21 from pyspark.sql import SparkSession  
22 spark = SparkSession.builder.appName('test').getOrCreate()  
23
```

Hình 1.2.1. Các thư viện cần dùng trong quá trình nghiên cứu

```
● ● ●  
1 math = spark.read.csv(r'student-mat.csv', header=  
    True, inferSchema= True , sep=';')  
2  
3 math.toPandas()
```

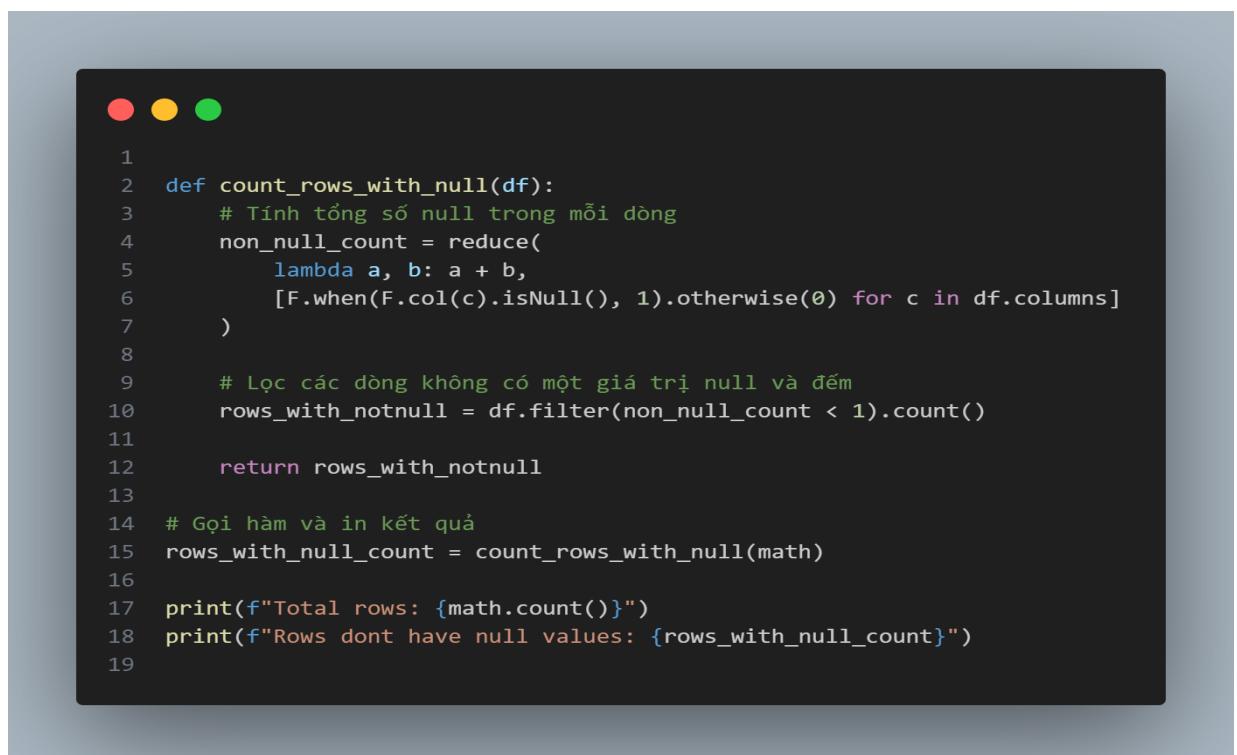
Hình 1.2.2. Tải tệp và đọc dữ liệu trên hệ thống

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10
...
390	MS	M	20	U	LE3	A	2	2	services	services	...	5	5	4	4	5	4	11	9	9	9
391	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	3	14	16	16
392	MS	M	21	R	GT3	T	1	1	other	other	...	5	5	3	3	3	3	3	10	8	7
393	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	0	11	12	10
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3	2	3	3	3	5	5	8	9	9

Hình 1.2.3. Thông tin của tệp dữ liệu

b) Phân tích cơ bản

- Kiểm tra tệp dữ liệu có các giá trị rỗng (NULL) hay không:



```

1
2 def count_rows_with_null(df):
3     # Tính tổng số null trong mỗi dòng
4     non_null_count = reduce(
5         lambda a, b: a + b,
6         [F.when(F.col(c).isNull(), 1).otherwise(0) for c in df.columns]
7     )
8
9     # Lọc các dòng không có một giá trị null và đếm
10    rows_with_notnull = df.filter(non_null_count < 1).count()
11
12    return rows_with_notnull
13
14 # Gọi hàm và in kết quả
15 rows_with_null_count = count_rows_with_null(math)
16
17 print(f"Total rows: {math.count()}")
18 print(f"Rows dont have null values: {rows_with_null_count}")
19

```

Hình 1.2.4. Kiểm tra giá trị bị thiếu trong dữ liệu

```

Total rows: 395
Rows dont have null values: 395

```

Hình 1.2.5. Kết quả của kiểm tra

- Các cột dữ liệu chứa giá trị ngoại lai có thể gây ảnh hưởng đến tổng thể dữ liệu:

+ Bảng giá trị trong cột **age**:

Age	count
16	104
17	98
15	82
18	82
19	24
20	3
22	1
21	1

=> Từ bảng sau ta có thể thấy được số tuổi dao động từ 15 đến 21 tuổi. Tuy nhiên vẫn có một giá trị “22” tuổi nằm trong dữ liệu trên. Nguyên nhân có thể đến từ việc nhập sai trong quá trình lưu trữ dữ liệu. Dù không quá lớn nhưng vẫn đem lại sự ảnh hưởng nhất định.

+ Bảng giá trị cho cột **Medu**:

Medu	count
4	131
2	103
3	99
1	59
0	3

=> Các giá trị được dao động từ khoảng 1 đến 4 (từ tiểu học đến hết đại học). Tuy nhiên vẫn có 3 giá trị ngoại lai với giá trị là 0. Điều này phản ánh rằng học sinh có mẹ không có trình độ học vấn có thể ảnh hưởng đến chất lượng điểm số của học sinh.

+ Bảng giá trị cho cột **Fedu**:

Fedu	count
2	115
3	100
4	96
1	82
0	2

=> Tương tự với **Medu** các giá trị được dao động từ khoảng 1 đến 4 (từ tiểu học đến hết đại học). Tuy nhiên vẫn có 2 giá trị ngoại lai với giá trị là 0. Điều này phản ánh rằng học sinh có cha không có trình độ học vấn có thể ảnh hưởng đến chất lượng điểm số của học sinh.

+ Bảng giá trị cho cột **Traveltime**:

Travel Time (1–4)	Số lượng
1 (≤ 15 phút)	257
2 (15–30 phút)	107
3 (30–60 phút)	23
4 (> 60 phút)	8

=> Thời gian đi đến trường của học sinh thường từ dưới 15 phút đến 1 tiếng. Bên cạnh có vẫn có 8 học sinh với thời gian đi học lâu hơn so với các bạn, cụ thể là trên 60 phút (1 tiếng). Điều này có thể ảnh hưởng đến số liệu của các cột điểm số hoặc số buổi học nghỉ nếu không đến đúng giờ điểm danh.

+ Bảng giá trị cho cột **Studytime**:

Study Time (1–4)	Số lượng
1 (< 2 giờ/tuần)	105
2 (2–5 giờ/tuần)	198
3 (5–10 giờ/tuần)	65
4 (> 10 giờ/tuần)	27

=> Với thời gian học của học sinh dao động trong tần suất từ 2 giờ đến 10 giờ trong tuần. Nhưng vẫn có những học sinh (27 học sinh) với sự tập trung và chăm chỉ hơn đã học với mức thời gian trên 10 tiếng. Chính những học sinh này sẽ đem lại sự ảnh hưởng lớn đến các bảng biểu dữ liệu sau này và thường các học sinh này đều là học sinh có điểm cuối kì cao trong lớp.

+ Bảng giá trị cho cột **Failures**:

Số môn rớt	Số lượng
0	312
1	50
2	17
3	16

=> Trong tổng thể 395 học sinh, phần lớn đều vượt qua các lớp (môn) nhưng bên cạnh đó vẫn có 83 học sinh tương ứng với 83 giá trị ngoại lai. Và rớt môn với số lượng đáng kể là 33 học sinh (2-3 môn rớt). Điều này thể hiện rằng học sinh gặp khó khăn trong việc học hoặc sự nỗ lực chưa thật sự đủ để có thể hoàn thành tốt quá trình học.

+ Bảng giá trị cho cột **Famrel**:

Điểm Famrel (1–5)	Số lượng
1	8
2	18
3	68
4	195
5	106

=> Ở bảng dữ liệu này ta có thể dễ dàng thấy được rằng tổng thể mặt bằng chung học sinh và gia đình có sự kết nối rất ổn định (từ bình thường đến rất tốt) đây là một biểu hiện tốt cho quá trình học tập của học sinh do có sự động viên và hỗ trợ từ người nhà. Tuy vậy vẫn có một nhiều trường hợp học sinh không có sự hợp tác từ phía gia đình (26 cá nhân) điều này phản ánh lên vấn đề kết nối giữa đôi bên và có thể dẫn đến tác động tiêu cực về mặt điểm số.

+ Bảng giá trị cho cột **Freetime**:

Điểm Freetime (1-5)	Số lượng
1	19
2	64
3	157
4	115
5	40

=> Thống kê được rằng có đến 19 giá trị ngoại lai với những học sinh có thời gian rảnh rất ít hoặc 40 trường hợp với thời gian rảnh quá nhiều. Điều này có thể chỉ ra sự thiếu đồng đều trong việc phân bổ thời gian cho các hoạt động ngoại khóa và thời gian học của học sinh.

+ Bảng giá trị cho cột **Absences**:

Số ngày vắng (absences)	Số lượng học sinh (count)
0	115
1	3
2	65
3	8
4	53
5	5
6	31
7	7
8	22
9	3
10	17
11	3
12	12
13	3
14	12
15	3
16	7
17	1
18	5
19	1
20	4
21	1

Số ngày vắng (absences)	Số lượng học sinh (count)
22	3
23	1
24	1
25	1
26	1
28	1
30	1
38	1
40	1
54	1
56	1
75	1

=> So với mức trung bình là 5.71 (~6 buổi) thì có các trường hợp được chỉ ra rằng số lượng học sinh vắng mặt ngoài phạm vi hợp lệ là rất nhiều. Dẫn đến xảy ra các giá trị ngoại lai như một trường hợp nghỉ đến 75 buổi.

+ Bảng giá trị cho cột G2:

G2 (Điểm)	Số lượng
9	50
10	46
12	41
13	37
11	35
15	34
8	32
14	23
7	21
5	15
6	14
16	13
0	13
18	12
17	5
19	3
4	1

=> Có 13 giá trị ngoại lai với giá trị thực tế là 0 điểm đã nằm ngoài phạm vi hợp lý của điểm học kỳ. Điều này có thể đến từ việc là do học sinh không tham gia kỳ thi hoặc do lỗi trong quá trình ghi nhận điểm số của nhà trường.

c) Báo cáo ngắn về đặc điểm dữ liệu

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	3.235443	3.108861	1.481013	2.291139	3.554430	5.708861	10.908861	10.713924	
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659	0.998862	1.113278	0.890741	1.287897	1.390303	8.003096	3.319195	3.761505	
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	3.000000	0.000000	
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000	1.000000	1.000000	3.000000	0.000000	8.000000	9.000000	
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	1.000000	2.000000	4.000000	4.000000	11.000000	11.000000	11.000000	
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000	2.000000	3.000000	5.000000	8.000000	13.000000	13.000000	
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	5.000000	5.000000	75.000000	19.000000	19.000000	20.000000	

Hình 1.2.6. Phân tích cơ bản của các đặc trưng dữ liệu

- Với các cột dữ liệu chứa đặc tính, đặc trưng cá nhân của các học sinh thì ta được một biểu đồ với các thống kê như sau:

1. age (Tuổi học sinh)

- Trung bình: 16 tuổi
- Dao động từ 15 đến 22 tuổi
- Phần lớn học sinh nằm trong độ tuổi từ 16 đến 18 (75% học sinh \leq 18 tuổi)
- Std: **1.28** => Hầu hết học sinh có độ tuổi khá gần nhau, không chênh lệch nhiều.

2. Medu (Trình độ học vấn của mẹ)

- Trung bình: 2.75 ~ 3 (tương đương với trình độ trung học phổ thông)
- Giá trị nhỏ nhất là 0 (không được đi học)
- Giá trị tối đa là 4 (giáo dục đại học)
- 50% học sinh có mẹ học đến trình độ trung học phổ thông trở lên
- Std: **1.09** => Mức độ phân tán vừa phải, có sự đa dạng trong trình độ học vấn của mẹ.

3. Fedu (Trình độ học vấn của cha)

- Trung bình: 2.52, tương đương với mức trung học cơ sở đến trung học phổ thông
- Phân bố tương tự như trình độ học vấn của mẹ
- Std: **1.09** => Mức phân tán gần như tương đương Medu.

4. traveltimes (Thời gian đi học)

- Trung bình: khoảng 1.45 (~<15 phút)
- Giá trị lớn nhất là 4 nghĩa là học sinh mất trên 1 tiếng đến trường
- Phần lớn học sinh chỉ mất 1–2 đơn vị thời gian (tức là dưới 30 phút) để đến trường
- Std: **0.70** => Mức độ chênh lệch thời gian đi học thấp, đa số gần trường.

5. studytime (Thời gian học mỗi tuần)

- Trung bình: 2.03 (~2–5 giờ mỗi tuần)
- Std: **0.84** => Có sự khác biệt tương đối giữa các học sinh về thời gian tự học.

6. failures (Số lần trượt môn trước đó)

- Trung bình: 0.33, phần lớn học sinh không bị trượt môn (75% có số lần trượt là 0)
- Tuy nhiên, có học sinh trượt đến 3 môn
- Std: **0.74** => Một vài học sinh có số lần trượt cao làm tăng độ phân tán.

7. famrel (Mối quan hệ trong gia đình)

- Trung bình: 3.94 (tốt)
- Giá trị dao động từ 1 (rất tệ) đến 5 (rất tốt)
- 75% học sinh đánh giá mối quan hệ gia đình từ mức 4 trở lên tức là từ tốt đến rất tốt
- **Std: 0.90 =>** Phân bố khá đồng đều, có sự khác biệt giữa học sinh về quan hệ gia đình.

8. freetime (Thời gian rảnh sau giờ học)

- Trung bình: 3.23 ~ 3 tiếng một ngày
- Đa số học sinh có thời gian rảnh ở mức trung bình (3–4 tiếng)
- **Std: 1.00 =>** Mức độ phân tán vừa phải, có học sinh ít rảnh, có người rảnh nhiều.

9. goout (Mức độ ra ngoài với bạn bè)

- Trung bình: 3.11
- Cho thấy hoạt động xã hội ở mức vừa phải
- Có học sinh rất ít khi ra ngoài (1) và có học sinh thường xuyên ra ngoài (5)
- **Std: 1.11 =>** Có sự đa dạng rõ rệt trong hoạt động xã hội giữa học sinh.

10. Dalc (Mức tiêu thụ rượu trong ngày thường)

- Trung bình: 1.48 (rất thấp)
- Phần lớn học sinh không sử dụng hoặc sử dụng rất ít rượu trong tuần
- Tuy vẫn có trường hợp học sinh sử dụng rất nhiều lên đến mức 5
- **Std: 0.89 =>** Mức độ biến động tương đối trong việc sử dụng rượu ngày thường.

11. Walc (Mức tiêu thụ rượu vào cuối tuần)

- Trung bình: 2.29
- Mức tiêu thụ cao hơn so với ngày thường, nhưng vẫn khá thấp nhìn chung
- **Std: 1.29 =>** Mức độ biến động cao hơn Dalc, phản ánh hành vi uống rượu vào cuối tuần đa dạng hơn.

12. health (Tình trạng sức khỏe hiện tại)

- Trung bình: 3.55
- Phần lớn học sinh tự đánh giá sức khỏe ở mức tốt (từ 3–4 trở lên)
- **Std: 1.39 =>** Có sự khác biệt rõ trong nhận thức và tình trạng sức khỏe cá nhân.

13. absences (Số buổi vắng mặt)

- Trung bình: 5.71 buổi
- Có học sinh vắng tới 75 buổi, cho thấy có giá trị ngoại lệ
- Đa số học sinh nghỉ dưới 8 buổi
- **Std: 8.00 =>** Độ lệch chuẩn rất cao, cần kiểm tra các ngoại lệ để tránh ảnh hưởng mô hình.

14. G1 (Điểm số kỳ 1)

- Trung bình: 10.91 (trên thang điểm 20)
- Phân bố đồng đều, điểm thấp nhất là 3 và cao nhất là 19
- **Std: 3.32 =>** Điểm số kỳ 1 có mức độ phân tán vừa phải.

15. G2 (Điểm số kỳ 2)

- Trung bình: 10.71
- Có xu hướng tương tự như G1
- Điểm thấp nhất là 0 và cao nhất là 19
- **Std: 3.76 =>** Mức phân tán điểm số cao hơn kỳ 1.

16. G3 (Điểm số cuối kỳ)

- Trung bình: 10.42
- Điểm thấp nhất là 0 và cao nhất là 20

- G3 phản ánh rõ kết quả học tập tổng thể và được dùng làm biến mục tiêu trong các mô hình dự đoán
- **Std: 4.58** => Mức độ biến động cao nhất trong 3 kỳ => thể hiện rõ sự khác biệt về kết quả học tập cuối cùng.

- Ngoài ra ta có thể thấy được kiểu dữ liệu của các cột đặc trưng được phân bố như sau:

	#	Column	Non-Null Count	Dtype
1		Data columns (total 33 columns):		
2		# Column	Non-Null Count	Dtype
3				
4	0	school	395 non-null	object
5	1	sex	395 non-null	object
6	2	age	395 non-null	int64
7	3	address	395 non-null	object
8	4	famsize	395 non-null	object
9	5	Pstatus	395 non-null	object
10	6	Medu	395 non-null	int64
11	7	Fedu	395 non-null	int64
12	8	Mjob	395 non-null	object
13	9	Fjob	395 non-null	object
14	10	reason	395 non-null	object
15	11	guardian	395 non-null	object
16	12	traveltime	395 non-null	int64
17	13	studytime	395 non-null	int64
18	14	failures	395 non-null	int64
19	15	schoolsup	395 non-null	object
20	16	famsup	395 non-null	object
21	17	paid	395 non-null	object
22	18	activities	395 non-null	object
23	19	nursery	395 non-null	object
24	20	higher	395 non-null	object
25	21	internet	395 non-null	object
26	22	romantic	395 non-null	object
27	23	famrel	395 non-null	int64
28	24	freetime	395 non-null	int64
29	25	goout	395 non-null	int64
30	26	Dalc	395 non-null	int64
31	27	Walc	395 non-null	int64
32	28	health	395 non-null	int64
33	29	absences	395 non-null	int64
34	30	G1	395 non-null	int64
35	31	G2	395 non-null	int64
36	32	G3	395 non-null	int64

Hình 1.2.7. Kiểu dữ liệu trong tệp dữ liệu student-mat.csv

1. Thông tin chung:

- + Tổng số bản ghi (số dòng): 395 học sinh
- + Tổng số thuộc tính (số cột): 33
- + Không có giá trị thiếu (null) ở bất kỳ cột nào – dữ liệu đầy đủ 100%
- + Dữ liệu chia làm 2 kiểu chính:
 - . 16 cột kiểu số nguyên (int64)
 - . 17 cột kiểu phân loại/ký tự (object)

2. Phân loại theo các nhóm biến bằng cách sử dụng phương thức **info()**:

1) Thông tin cá nhân và gia đình (Personal and Family Information)

Cột	Kiểu dữ liệu	Mô tả
school	object	Trường học: 'GP' (Gabriel Pereira) hoặc 'MS' (Mousinho da Silveira)
sex	object	Giới tính: 'F' (nữ), 'M' (nam)
age	int64	Tuổi của học sinh
address	object	Địa chỉ: 'U' (thành thị), 'R' (nông thôn)
famsize	object	Quy mô gia đình: 'LE3' (≤ 3), 'GT3' (> 3)
Pstatus	object	Tình trạng sống của cha mẹ: 'T' (sống cùng nhau), 'A' (đã ly thân)
Medu	int64	Trình độ học vấn của mẹ (0–4)
Fedu	int64	Trình độ học vấn của cha (0–4)
Mjob	object	Nghề nghiệp của mẹ
Fjob	object	Nghề nghiệp của cha
guardian	object	Người giám hộ: mẹ, cha hoặc người khác

2) Thông tin học tập (Academic Information):

Cột	Kiểu dữ liệu	Mô tả
reason	object	Lý do chọn trường
traveltime	int64	Thời gian đến trường (1–4)
studytime	int64	Thời gian tự học mỗi tuần (1–4)
failures	int64	Số lần trượt môn trong quá khứ
schoolsupsup	object	Có nhận hỗ trợ học tập ở trường không
famsup	object	Có nhận hỗ trợ học tập từ gia đình không
paid	object	Có tham gia lớp học thêm không
higher	object	Có ý định học lên cao không
internet	object	Có Internet ở nhà không
G1	int64	Điểm học kì đầu
G2	int64	Điểm học kì hai
G3	int64	Điểm học kì cuối

3) Hoạt động và hành vi xã hội (Social & Behavioral Factors):

Cột	Kiểu dữ liệu	Mô tả
activities	object	Có tham gia hoạt động ngoại khóa không
nursery	object	Có học mẫu giáo trước đây không
romantic	object	Có đang trong mối quan hệ tình cảm không

Cột	Kiểu dữ liệu	Mô tả
famrel	int64	Quan hệ gia đình (1–5)
freetime	int64	Thời gian rảnh sau giờ học (1–5)
goout	int64	Mức độ đi chơi với bạn (1–5)
Dalc	int64	Mức độ uống rượu trong ngày thường (1–5)
Walc	int64	Mức độ uống rượu vào cuối tuần (1–5)
health	int64	Đánh giá sức khỏe bản thân (1–5)
absences	int64	Số buổi nghỉ học

CHƯƠNG 2. TRỰC QUAN HÓA DỮ LIỆU

2.1. Yêu cầu

- Vẽ 3 biểu đồ từ 3 thư viện với tiêu đề, nhãn trục, màu sắc, chú thích. Thêm slider widget trong Bokeh để lọc studytime và cập nhật biểu đồ

2.2. Kết quả

a) Matplotlib

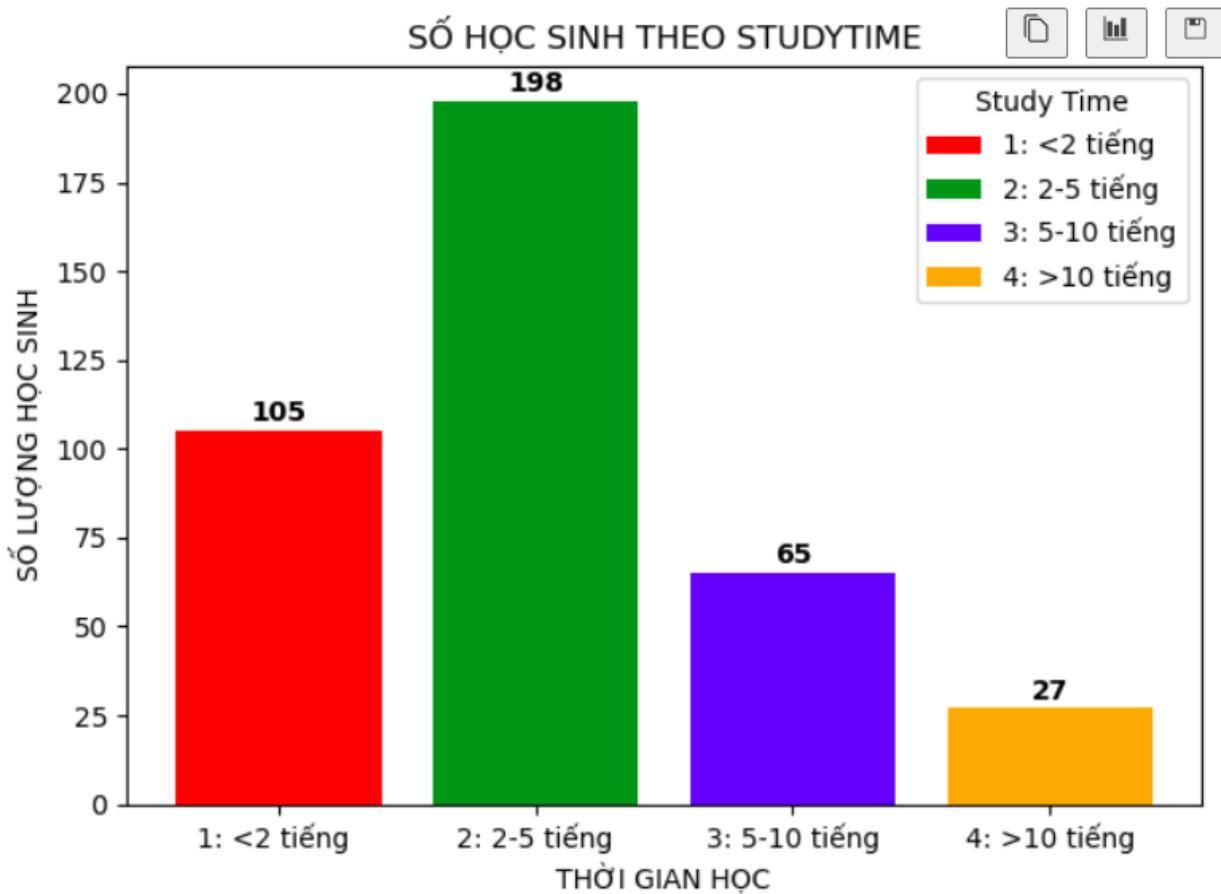
- Mã nguồn tạo biểu đồ cột:



```
1 import matplotlib.pyplot as plt
2
3 # Dữ liệu đầu vào
4 group1 = math.filter(math.studytime == 1).count()
5 group2 = math.filter(math.studytime == 2).count()
6 group3 = math.filter(math.studytime == 3).count()
7 group4 = math.filter(math.studytime == 4).count()
8
9 studytime_rate = ["1: <2 tiếng ", "2: 2-5 tiếng", "3: 5-10 tiếng", "4: >10 tiếng"]
10 student_amount = [group1, group2, group3, group4]
11 colors = ['red', 'green', 'blue', 'orange']
12
13 # Vẽ biểu đồ cột
14 bars = plt.bar(studytime_rate, student_amount, color=colors)
15
16 # Ghi số lượng trên đầu mỗi cột
17 for bar in bars:
18     height = bar.get_height()
19     plt.text(bar.get_x() + bar.get_width() / 2, height + 1, str(height),
20               ha='center', va='bottom', fontsize=10, fontweight='bold')
21
22 # Thêm tiêu đề và nhãn
23 plt.xlabel("THỜI GIAN HỌC")
24 plt.ylabel("SỐ LƯỢNG HỌC SINH")
25 plt.title("SỐ HỌC SINH THEO STUDYTIME")
26
27 # Ghi chú (legend)
28 plt.legend(bars, studytime_rate, title="Study Time")
29
30 # Hiển thị biểu đồ
31 plt.tight_layout()
32 plt.show()
```

Hình 2.2.1. Mã nguồn tạo biểu đồ cột theo "Thời gian học" và "Số lượng học sinh"

- Kết quả biểu đồ:



Hình 2.2.2. Biểu đồ cột theo mã nguồn

- Mô tả biểu đồ:

Biểu đồ cột - thư viện Matplotlib: Để tạo biểu đồ cột cho dữ liệu nhằm so sánh giữa số lượng học sinh theo 'studytime' cần làm các bước sau:

Bước	Mô Tả
1. Import thư viện	Import thư viện <code>matplotlib.pyplot</code> vào hệ thống.
2. Chia cột theo thời gian học	Theo thống kê, thời gian học dao động từ 1-4 giờ nên ta chia thành 4 cột theo từng thời gian.
3. Đếm số lượng học sinh	Đếm số lượng học sinh với từng điều kiện thời gian thông qua câu lệnh <code>filter</code> . Ví dụ: <code>group1 = math.filter(math.studytime == 1).count()</code> .
4. Gán giá trị cho trực	Gán các giá trị tìm được vào trực hoành và trực tung. Câu lệnh ví dụ: <code>studytime_rate = ["1 hour", "2 hours", "3 hours", "4 hours"]</code> <code>student_amount = [group1, group2, group3, group4]</code> .
5. Tạo biểu đồ	Sử dụng <code>plt.bar(studytime_rate, student_amount)</code> để tạo biểu đồ cột.
6. Đặt tên cho trực và tiêu đề	Đặt tên cho trực X, trực Y và tiêu đề cho biểu đồ.

Hình 2.2.3. Mô tả biểu đồ

- Mục đích sử dụng biểu đồ:

Biểu đồ cột được sử dụng nhằm trực quan hóa số lượng học sinh tương ứng với từng mức độ thời gian học trong tuần, từ đó hỗ trợ phân tích xu hướng và sự phân bố của học sinh theo mức độ đầu tư thời gian vào việc học.

- Lý do chọn biểu đồ cột:

- Dễ hiểu và trực quan: Biểu đồ cột thể hiện rõ ràng sự so sánh giữa các nhóm thời gian học khác nhau (từ <2 giờ đến >10 giờ mỗi tuần).
- Phù hợp với dữ liệu phân loại: "Studytime" là biến phân loại (categorical variable), nên biểu đồ cột là lựa chọn lý tưởng để trình bày sự phân bố của từng nhóm.
- Hiệu quả trong việc so sánh số lượng: Giúp người xem dễ dàng nhận biết nhóm nào có số lượng học sinh đông nhất, nhóm nào ít hơn, từ đó hỗ trợ việc đưa ra các nhận xét và phân tích tiếp theo.

- Kết luận từ biểu đồ:

- Nhóm học 2–5 tiếng mỗi tuần chiếm tỷ lệ lớn nhất, với 198 học sinh – gần một nửa tổng số mẫu khảo sát.
- Nhóm học dưới 2 tiếng vẫn khá lớn, gồm 105 học sinh, cho thấy một bộ phận đáng kể học sinh học với thời lượng rất thấp.
- Trong khi đó, nhóm học >10 tiếng rất ít, chỉ có 27 học sinh, cho thấy học quá nhiều không phải là xu hướng phổ biến.

⇒ Ý nghĩa: Biểu đồ này cung cấp cái nhìn tổng quan về thói quen học tập của học sinh, giúp xác định các nhóm học sinh chủ yếu và định hướng các phân tích sâu hơn, như việc liệu thời gian học có thực sự ảnh hưởng đến điểm số hay không.

- Mã nguồn tạo biểu đồ phân tán:

```

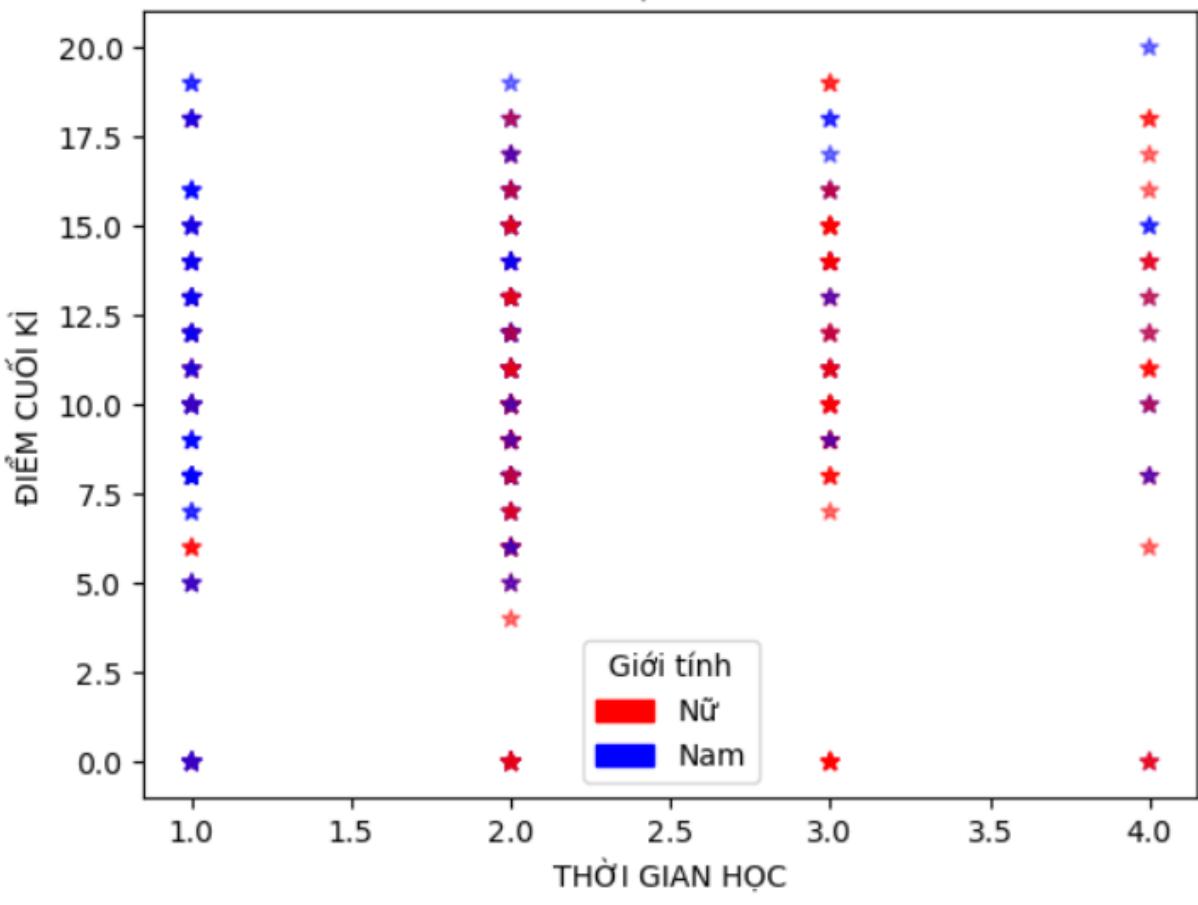
1 df_pandas = math.toPandas()
2
3 st = math.select('studytime').toPandas()
4 g3 = math.select('G3').toPandas()
5 sex = math.select('sex').toPandas()
6
7 df_pandas['sex_num'] = df_pandas['sex'].map({'F': 0, 'M': 1})
8
9 plt.scatter(x = st , y = g3 ,
10             c=df_pandas['sex_num'].map({0: 'red', 1: 'blue'}),
11             marker='*',alpha=0.5,label="Dữ liệu điểm số")
12
13 # Append the label on X-axis
14 plt.xlabel("THỜI GIAN HỌC")
15
16 # Append the label on X-axis
17 plt.ylabel("ĐIỂM CUỐI KÌ")
18
19 # Add the title to graph
20 plt.title("BIỂU ĐỒ PHÂN TÁN \n THỜI GIAN HỌC - ĐIỂM CUỐI KÌ ")
21
22
23 # Gắn chú thích đúng màu tương ứng
24 legend_labels = [
25     mpatches.Patch(color='red', label='Nữ'),    # C0 là màu của giá trị 0
26     mpatches.Patch(color='blue', label='Nam')    # C1 là màu của giá trị 1
27 ]
28 plt.legend(handles=legend_labels, title="Giới tính")
29
30 # Display the chart
31 plt.show()

```

Hình 2.2.4. Mã nguồn tạo biểu đồ phân tán theo "Studytime" và "G3" với da màu sắc

- Kết quả biểu đồ:

**BIỂU ĐỒ PHÂN TÁN
THỜI GIAN HỌC - ĐIỂM CUỐI KÌ**



- Mô tả biểu đồ:

1) **df_pandas = math.toPandas()**: Chuyển toàn bộ dữ liệu từ DataFrame PySpark (math) sang Pandas DataFrame (df_pandas) để dễ dàng thao tác và trực quan hóa bằng thư viện matplotlib.

2) **st = math.select('studytime').toPandas()**: Trích xuất cột studytime (thời gian học) từ DataFrame PySpark và chuyển thành Pandas Series.

3) **g3 = math.select('G3').toPandas()**: Trích xuất cột G3 (điểm cuối kỳ) từ DataFrame PySpark và chuyển thành Pandas Series.

4) **sex = math.select('sex').toPandas()**: Trích xuất cột sex (giới tính) từ DataFrame PySpark và chuyển thành Pandas Series.

5) **df_pandas['sex_num'] = df_pandas['sex'].map({'F': 0, 'M': 1})**: Mã hóa giới tính: gán giá trị 0 cho nữ (F) và 1 cho nam (M) để phục vụ việc phân loại màu sắc trong biểu đồ.

6) **plt.scatter(x = st , y = g3 , c=df_pandas['sex_num'].map({0: 'red', 1: 'blue'}), marker='*', alpha=0.5, label="Dữ liệu điểm số")**:

Vẽ biểu đồ phân tán:

- Trục hoành là thời gian học (studytime).
- Trục tung là điểm cuối kỳ (G3).
- Mỗi điểm dữ liệu được tô màu theo giới tính (đỏ cho nữ, xanh cho nam).
- Dùng biểu tượng hình sao (*) với độ trong suốt (alpha = 0.5) để dễ nhìn khi các điểm bị chồng lên nhau.

7) plt.xlabel("THỜI GIAN HỌC"): Đặt nhãn cho trục hoành là "THỜI GIAN HỌC".

8) plt.ylabel("ĐIỂM CUỐI KÌ"): Đặt nhãn cho trục tung là "ĐIỂM CUỐI KÌ".

9) plt.title("BIỂU ĐỒ PHÂN TÁN \n THỜI GIAN HỌC - ĐIỂM CUỐI KÌ "): Thêm tiêu đề cho biểu đồ, chia làm 2 dòng bằng \n để rõ ràng và dễ quan sát.

10) legend_labels = [mpatches.Patch(color='red', label='Nữ')mpatches.Patch(color='blue', label='Nam')]: Tạo thủ công phần chú thích cho biểu đồ để hiển thị đúng màu tương ứng với giới tính.

11) plt.legend(handles=legend_labels, title="Giới tính"): Hiển thị chú thích (legend) trên biểu đồ với tiêu đề "Giới tính", giúp người xem dễ phân biệt dữ liệu nam và nữ.

12) plt.show(): Hiển thị biểu đồ phân tán đã tạo ra.

- Mục đích:

Biểu đồ phân tán này được xây dựng nhằm mục đích phân tích mối quan hệ giữa thời gian học (studytime) và điểm cuối kỳ (G3) của học sinh, đồng thời so sánh sự khác biệt giữa nam và nữ trong việc học tập.

- Lý do chọn biểu đồ phân tán:

- + Nó thể hiện rõ mối quan hệ giữa hai biến liên tục: thời gian học và điểm cuối kì.
- + Cho phép ta nhìn thấy sự phân bố dữ liệu và xu hướng tổng thể.
- + Việc mã hóa giới tính bằng màu sắc (đỏ cho nữ, xanh cho nam) giúp ta so sánh trực quan giữa các nhóm giới tính trong từng mức thời gian học.
- + Sử dụng ký hiệu * giúp phân biệt dữ liệu rõ ràng hơn trên biểu đồ.

- Kết luận rút ra từ biểu đồ:

=> Nhìn chung, thời gian học tăng lên có xu hướng liên quan đến điểm số cao hơn, nhưng mối quan hệ này không hoàn toàn tuyến tính. Cả nam và nữ đều có khả năng đạt điểm cao nếu có thời gian học từ 2 trở lên, nhưng sự phân bố điểm ở từng mức thời gian học lại không giống nhau.

=> Không có sự khác biệt rõ rệt giữa nam và nữ về xu hướng điểm số, nhưng biểu đồ cho thấy:

- + Ở mức thời gian học thấp (1-2), điểm số rất phân tán, bao gồm cả điểm 0.
- + Ở mức thời gian học cao hơn (3-4), số lượng học sinh đạt điểm cao nhiều hơn.

=> Biểu đồ giúp giáo viên hoặc nhà phân tích dễ dàng xác định mức thời gian học nào nên được khuyến khích để cải thiện kết quả học tập.

b) Seaborn

- Mã nguồn biểu đồ Implot:

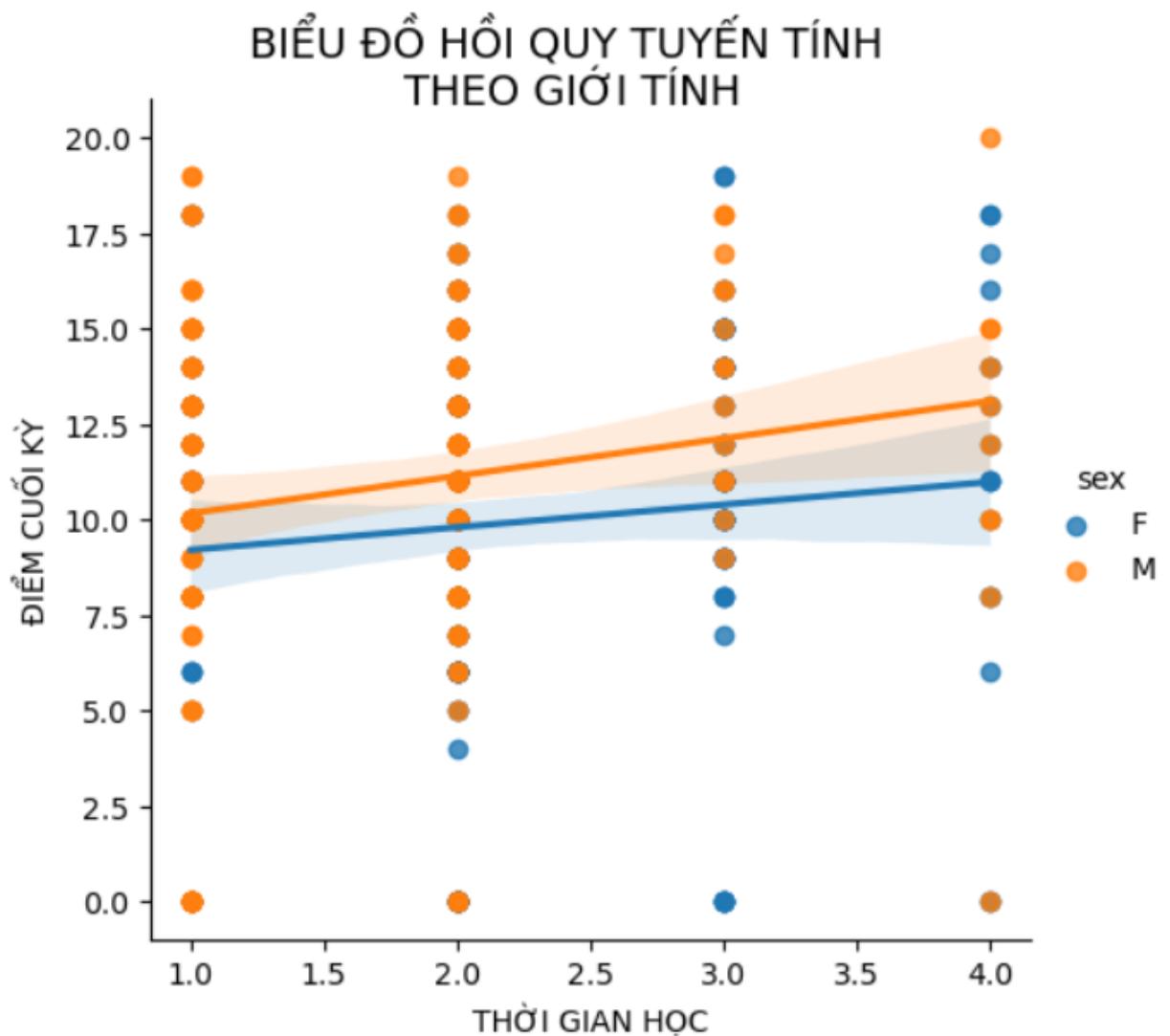
```

● ● ●
1 plot = sns.lmplot(x = 'studytime', y = 'G3', data = df_pandas, hue = 'sex')
2
3 # Ghi chú trực X và Y
4 plot.set_axis_labels("THỜI GIAN HỌC", "ĐIỂM CUỐI KỲ")
5
6 # Tiêu đề biểu đồ
7 plot.fig.suptitle("BIỂU ĐỒ HỒI QUY TUYẾN TÍNH \nTHEO GIỚI TÍNH", fontsize=14)
8
9 # Điều chỉnh vị trí tiêu đề cho không bị che
10 plot.fig.subplots_adjust(top=0.9)
11
12 #Show figure
13 plt.show()

```

Hình 2.2.5. Mã nguồn tạo biểu đồ hồi quy tuyến tính theo giới tính

- Kết quả biểu đồ:



Hình 2.2.6. Biểu đồ hồi quy tuyến tính theo mã nguồn trên

- Mô tả biểu đồ:

Biểu đồ implot - thư viện Seaborn: Để tạo dữ liệu biểu đồ hồi quy tuyến tính theo giới tính và phân bố của thời gian học (studytime) và điểm cuối kì (G3)

Bước	Mô Tả
1. Import thư viện	Import thư viện <code>seaborn</code> vào hệ thống.
2. Xác định trục và phân biệt giới tính	Trục hoành là <code>studytime</code> , trục tung là <code>G3</code> , và phân biệt các điểm dữ liệu theo giới tính (sex) sử dụng <code>hue='sex'</code> .
3. Đặt tên cho trục	Đặt tên cho trục X và trục Y. Ví dụ: <code>plot.set_axis_labels("THỜI GIAN HỌC", "ĐIỂM CUỐI KỲ")</code> .
4. Đặt tiêu đề cho biểu đồ	Đặt tiêu đề cho biểu đồ. Ví dụ: <code>plot.fig.suptitle("BIỂU ĐỒ HỒI QUY TUYẾN TÍNH \nTHEO GIỚI TÍNH", fontsize=14)</code> .
5. Điều chỉnh vị trí tiêu đề	Để tránh tiêu đề bị đè lên biểu đồ, sử dụng <code>plot.fig.subplots_adjust(top=0.9)</code> để nâng dòng chữ lên.

Hình 2.2.7. Mô tả biểu đồ

- Mục đích:

Biểu đồ hồi quy tuyến tính theo giới tính được sử dụng nhằm phân tích mối quan hệ giữa thời gian học và điểm cuối kỳ (G3), đồng thời đánh giá xem yếu tố giới tính có ảnh hưởng như thế nào đến mối quan hệ này.

- Lý do chọn:

- + **Biểu đồ phân tán** giúp trực quan hóa sự phân bố của dữ liệu giữa hai biến số liên tục (studytime và G3).
 - + Việc phân biệt theo màu sắc qua `hue = 'sex'` giúp dễ dàng so sánh dữ liệu giữa hai nhóm giới tính.
 - + Đường **hồi quy tuyến tính** được vẽ cho từng nhóm giúp thể hiện xu hướng thay đổi điểm số theo thời gian học của từng giới tính.
 - + Biểu đồ này hỗ trợ đánh giá **mức độ tương quan** và **sự khác biệt xu hướng học tập** giữa nam và nữ sinh.

- Kết luận rút ra từ biểu đồ:

- + Biểu đồ cho thấy xu hướng **điểm cuối kỳ tăng dần theo thời gian học** ở cả hai giới tính.
- + Tuy nhiên, **đường hồi quy của nam (M)** có độ dốc cao hơn, cho thấy thời gian học có ảnh hưởng tích cực mạnh hơn đến điểm số của nam so với nữ.
 - + Nữ sinh cũng có xu hướng điểm tăng theo thời gian học nhưng ở mức độ thấp hơn.
 - + Từ đó, có thể rút ra rằng **giới tính có ảnh hưởng nhất định đến mối quan hệ giữa thời gian học và kết quả học tập**, điều này có thể được khai thác để xây dựng các mô hình hỗ trợ học tập phù hợp với từng nhóm đối tượng.

- Mã nguồn tạo biểu đồ hộp (Boxplot):

```

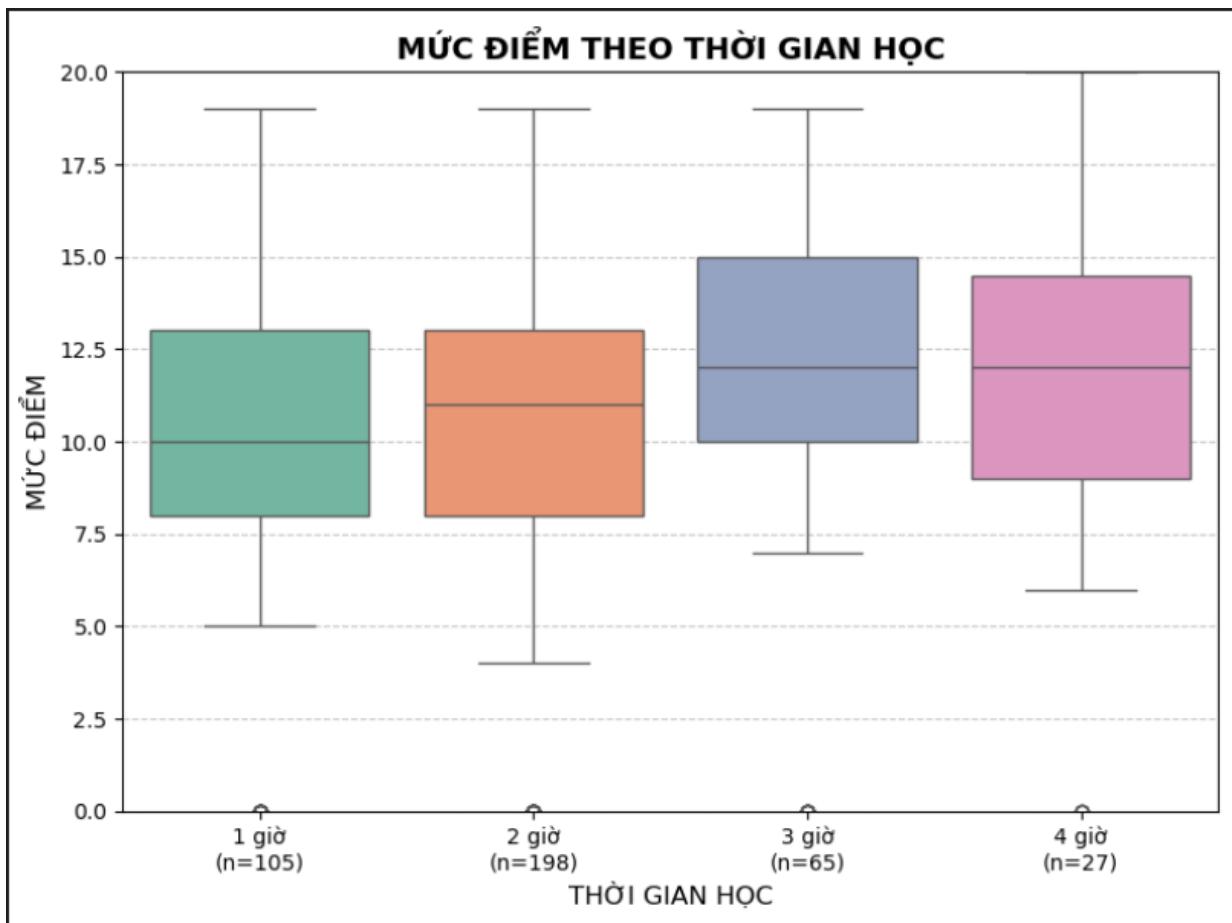
● ● ●

1 # Đếm số lượng học sinh cho từng mức studytime
2 counts = df_pandas['studytime'].value_counts().sort_index()
3
4 # Tạo nhãn mới với số lượng
5 labels = [f'{time} giờ\nn(n={counts[time]})' for time in sorted
6 (df_pandas['studytime'].unique())]
7
8 # Tạo boxplot với màu tùy chỉnh
9 plt.figure(figsize=(8, 6))
10 sns.boxplot(x='studytime', y='G3', data=df_pandas, palette='Set
11 2')
12
13 # Gắn lại nhãn trục x với số lượng học sinh
14 plt.xticks(ticks=range(len(labels)), labels=labels)
15
16 # Cài đặt nhãn và tiêu đề
17 plt.title('MỨC ĐIỂM THEO THỜI GIAN HỌC', fontsize=14, fontweight
18 = 'bold')
19 plt.xlabel('THỜI GIAN HỌC', fontsize=12)
20 plt.ylabel('MỨC ĐIỂM', fontsize=12)
21 plt.ylim(0, 20)
22
23 # Thêm grid nhẹ
24 plt.grid(axis='y', linestyle='--', alpha=0.7)
25 plt.show()

```

Hình 2.2.8. Mã nguồn biểu đồ hộp phân phối điểm "G3" theo "Studytime"

- Kết quả biểu đồ:



Hình 2.2.9. Biểu đồ hộp với mức điểm "G3" và "Studytime"

- Mô tả biểu đồ:

- + `counts = df_pandas['studytime'].value_counts().sort_index()`: Đếm số học sinh tương ứng với mỗi mức thời gian học (1 giờ, 2 giờ, 3 giờ, 4 giờ).
- + `labels = [f'{time} giờ\n(n={counts[time]})' for time in sorted (df_pandas['studytime'].unique())]`: Tạo nhãn trực x bao gồm cả thời gian học và số lượng học sinh trong mỗi nhóm.
- + `plt.figure(figsize=(8, 6)), sns.boxplot(x='studytime', y='G3', data=df_pandas, palette='Set2')`: Khởi tạo biểu đồ boxplot:
 - . Trục X là thời gian học (studytime)
 - . Trục Y là điểm cuối kỳ (G3)
 - . Sử dụng bảng màu Set2 để phân biệt các cột.
- + `plt.xticks(ticks=range(len(labels)), labels=labels)`: Gán các nhãn mới cho trục X kèm số học sinh theo từng nhóm.
- + `plt.title('MỨC ĐIỂM THEO THỜI GIAN HỌC', fontsize=14, fontweight='bold')`
- + `plt.xlabel('THỜI GIAN HỌC', fontsize=12)`
- + `plt.ylabel('MỨC ĐIỂM', fontsize=12)`
- + `plt.ylim(0, 20)`: Đặt tiêu đề và nhãn trục. Giới hạn trục Y từ 0 đến 20 (giá trị điểm tối đa).
- + `plt.grid(axis='y', linestyle='--', alpha=0.7)`: Thêm đường lưới theo trục Y để dễ đọc mức điểm.
- + `plt.tight_layout(), plt.show()`: Tự động điều chỉnh khoảng cách và hiển thị biểu đồ.

- Mục đích: Biểu đồ boxplot được sử dụng nhằm phân tích sự phân phối điểm số cuối kỳ (G3) của học sinh theo **thời gian học mỗi tuần** (từ 2 đến trên 10 giờ). Qua đó, ta có thể đánh giá mối liên hệ giữa thời gian học và mức điểm đạt được.

- Lý do chọn:

+ Độ phân tán của dữ liệu.

+ Giá trị trung vị (median) của điểm số trong từng nhóm thời gian học.

+ Các giá trị ngoại lai nếu có.

+ Việc bổ sung số lượng học sinh trên mỗi cột giúp tăng tính trực quan, cho biết nhóm nào có nhiều học sinh hơn.

+ Dễ dàng nhận biết liệu thời gian học có thực sự ảnh hưởng đến điểm số không, và mức độ ảnh hưởng như thế nào.

- Kết luận rút ra từ biểu đồ:

+ Điểm số trung vị tăng dần theo thời gian học, đặc biệt từ mức 2 trở lên.

+ Nhóm học 3 và 4 có **mức điểm trung bình cao hơn** và **phân phối điểm chặt hơn** (ít phân tán).

+ Mặc dù nhóm 4 ít học sinh ($n=27$), họ có xu hướng đạt điểm cao hơn so với các nhóm còn lại.

+ Kết luận: **Thời gian học nhiều hơn có thể giúp học sinh đạt điểm cao hơn**, tuy nhiên cần nghiên cứu sâu thêm về các yếu tố khác như chất lượng học, phương pháp học, hoặc động lực cá nhân.

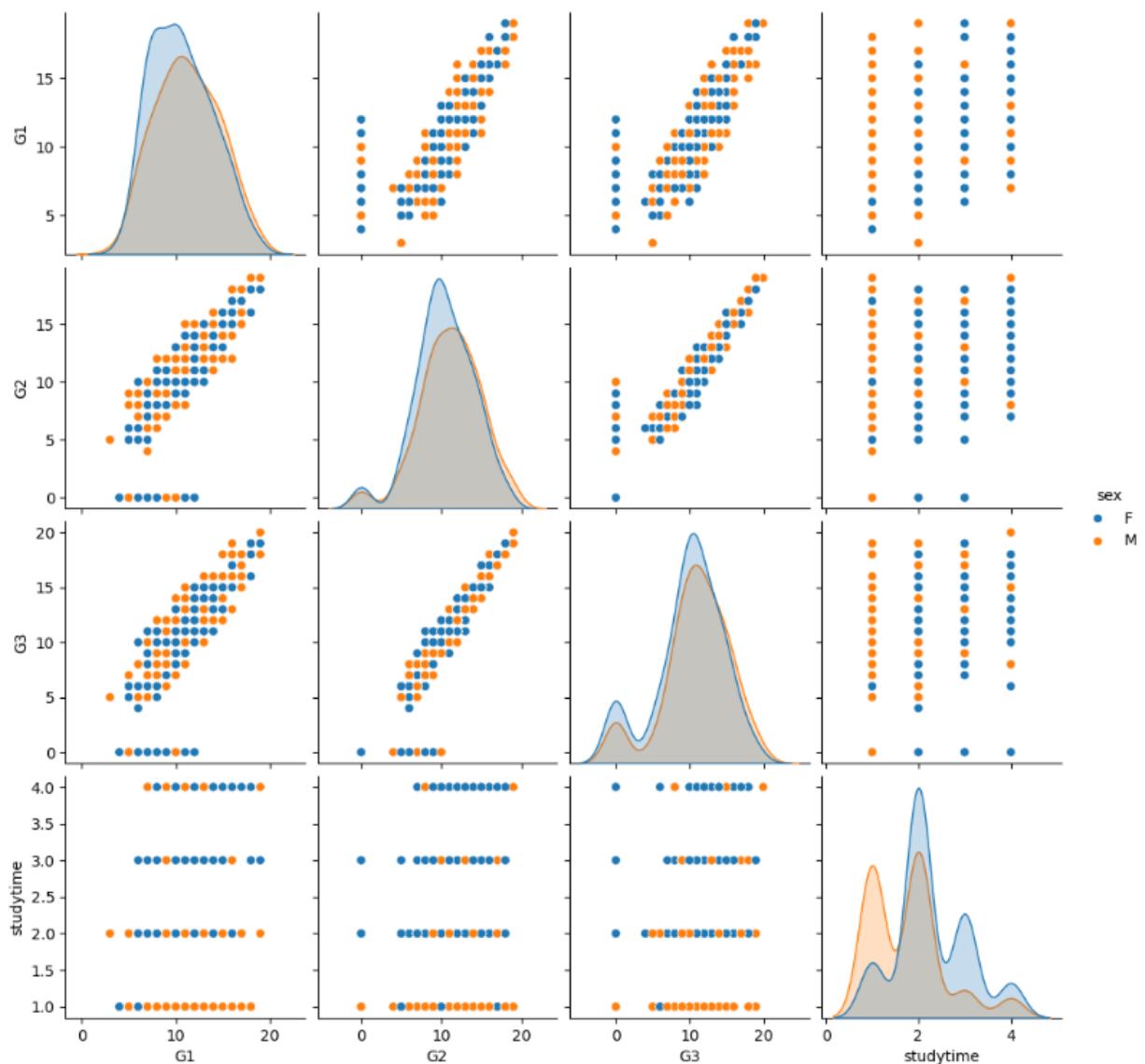
- Mã nguồn tạo biểu đồ Pairplot:

```
● ● ●  
1 # Tạo pairplot  
2 sns.pairplot(df_pandas[['G1', 'G2', 'G3', 'studytime', 'sex']],  
    hue='sex')  
3  
4 # Thêm tiêu đề tổng cho toàn bộ biểu đồ  
5 plt.suptitle('MỐI QUAN HỆ GIỮA G1, G2, G3 VÀ THỜI GIAN HỌC', y=  
    1.02)  
6  
7 # Hiển thị biểu đồ  
8 plt.show()
```

Hình 2.2.10. Mã nguồn tạo biểu đồ xây dựng mối quan hệ giữa "G1","G2","G3" và "Studytime"

- Kết quả biểu đồ:

MỐI QUAN HỆ GIỮA G1, G2, G3 VÀ THỜI GIAN HỌC



Hình 2.2.11. Biểu đồ mối quan hệ giữa "G1", "G2", "G3" và "Studytime"

- Mô tả biểu đồ:

+ `sns.pairplot(df_pandas[['G1', 'G2', 'G3', 'studytime', 'sex']], hue='sex')`: Tạo biểu đồ cặp từ các biến đã chọn: G1, G2, G3, studytime và sex, hue='sex' giúp phân biệt điểm dữ liệu nam và nữ bằng màu sắc.

+ `plt.suptitle('MỐI QUAN HỆ GIỮA G1, G2, G3 VÀ THỜI GIAN HỌC', y=1.02)`: Đặt tiêu đề tổng cho biểu đồ, y=1.02 dùng để điều chỉnh vị trí tiêu đề nằm cao hơn biểu đồ để không bị che.

+ `plt.show()`: Hiển thị toàn bộ biểu đồ lên màn hình.

- Mục đích: Biểu đồ **pairplot** được sử dụng để **khám phá mối quan hệ giữa các cặp biến định lượng** gồm: điểm số các kỳ (G1, G2, G3) và thời gian học (studytime), đồng thời phân tích sự khác biệt giữa hai giới tính (sex) thông qua phân biệt màu sắc.

- Lý do chọn:

- + Hiển thị mối quan hệ tuyến tính (hoặc phi tuyến) giữa các biến theo từng cặp.
- + Giúp xác định được liệu các điểm số G1, G2 có dự đoán tốt cho G3 không.

+ Cho phép so sánh trực quan theo giới tính để xem liệu nam và nữ có xu hướng điểm số hay thời gian học khác nhau không.

- Kết luận rút ra từ biểu đồ:

+ Có **mối tương quan tuyến tính mạnh giữa G1, G2 và G3**: điểm số ở các kỳ học trước là chỉ báo khá tốt cho điểm G3.

+ Nhóm học sinh có thời gian học cao hơn thường có G3 tốt hơn, nhưng mối liên hệ giữa studytime và điểm số không chặt như G1/G2 với G3.

+ Phân biệt giới tính qua màu sắc cho thấy **không có sự khác biệt rõ ràng về điểm số hay thời gian học giữa nam và nữ**.

+ Kết luận: G1 và G2 có thể sử dụng làm **biến dự báo** cho G3 trong các mô hình hồi quy hoặc phân loại điểm số.

c) Bokeh

- Biểu đồ phản ứng tương tác với 3 chức năng (hovertool, click policy, slider widget) nhằm cập nhật biểu đồ:

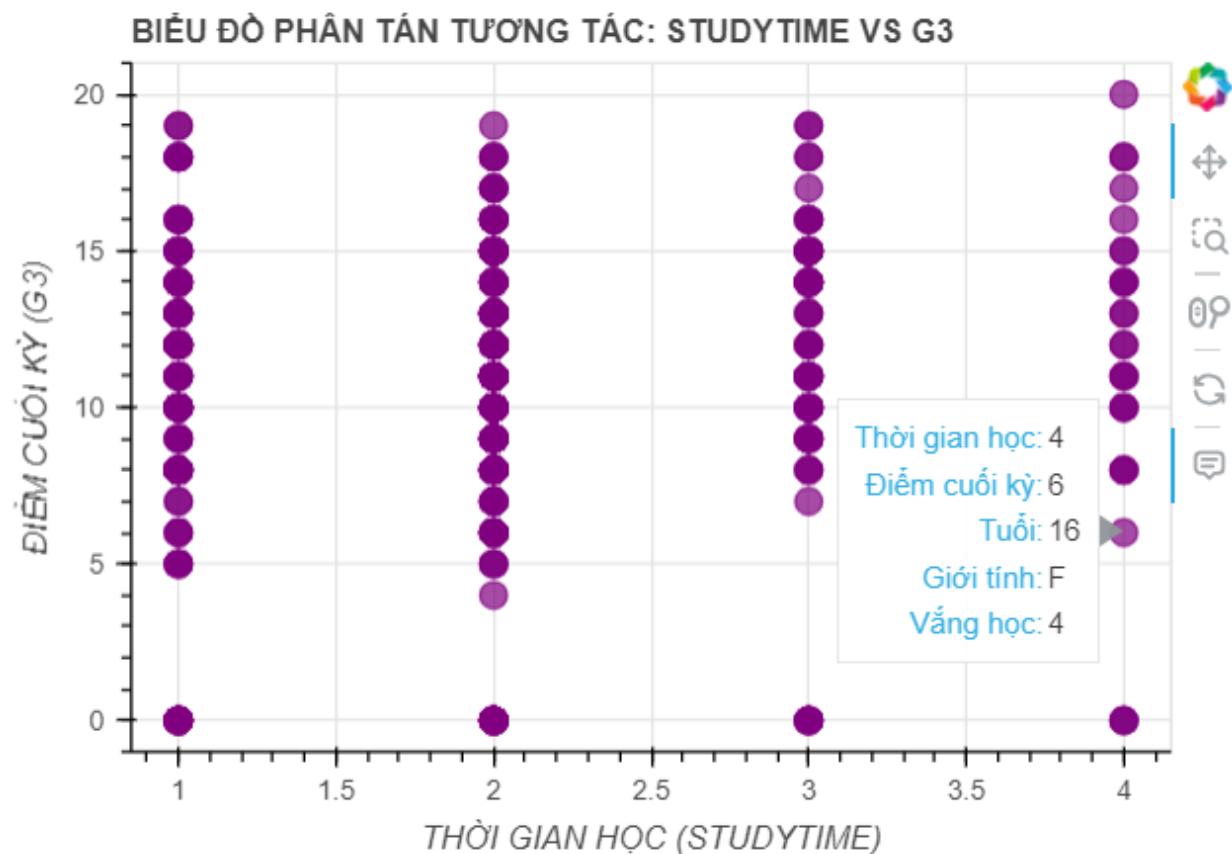
```

1  output_notebook()
2
3  # Giả sử df_pandas đã được chuẩn bị từ trước
4  # Chọn các cột cần thiết
5  df = df_pandas[['studytime', 'G3', 'age', 'sex', 'absences']]
6
7  # Tạo nguồn dữ liệu cho Bokeh
8
9  sources = []
10 original_sources = []
11 colors = {'F': 'blue', 'M': 'red'}
12
13 # Tạo biểu đồ
14 p = figure(width=500, height=350,
15             title="Biểu đồ phân tán tương tác: Studytime vs G3",
16             x_axis_label='Thời gian học (studytime)',
17             y_axis_label='Điểm cuối kỳ (G3)',
18             tools="pan,wheel_zoom,box_zoom,reset")
19
20 # Thêm các điểm cho mỗi giới tính với màu sắc khác nhau
21 for sex, color in colors.items():
22     subset = df[df["sex"] == sex]
23     source = ColumnDataSource(subset)
24     original_source = ColumnDataSource(subset)
25
26
27     original_sources.append(original_source)
28     sources.append(source)
29
30     p.scatter(x="studytime", y="G3",
31                 size=10, color=color, alpha=0.7,
32                 source=source, legend_label=f"Giới tính {sex}")
33
34
35 # Slider lọc theo studytime
36 callback = CustomJS(args=dict(sources=sources, original_sources=original_sources), code="""
37     var threshold = cb_obj.value;
38
39     for (var i = 0; i < sources.length; i++) {
40         var source = sources[i];
41         var original_data = original_sources[i].data;
42         var new_data = {studytime: [], G3: [], age: [], sex: [], absence
s: []};
43
44         for (var j = 0; j < original_data['studytime'].length; j++) {
45             if (original_data['studytime'][j] >= threshold) {
46                 new_data['studytime'].push(original_data['studytime']
[j]);
47                 new_data['G3'].push(original_data['G3'][j]);
48                 new_data['age'].push(original_data['age'][j]);
49                 new_data['sex'].push(original_data['sex'][j]);
50                 new_data['absences'].push(original_data['absences'][j]);
51             }
52         }
53
54         source.data = new_data;
55         source.change.emit();
56     }
57     """
58
59 slider = Slider(start=int(df["studytime"].min()),
60                  end=int(df["studytime"].max()),
61                  value=int(df["studytime"].min()),
62                  step=1,
63                  title="Lọc theo Thời gian học (studytime)")
64 slider.js_on_change("value", callback)
65
66 # Thêm HoverTool để hiển thị thông tin
67 hover = HoverTool(tooltips=[
68     ("Thời gian học", "@studytime"),
69     ("Điểm cuối kỳ", "@G3"),
70     ("Tuổi", "@age"),
71     ("Giới tính", "@sex"),
72     ("Vắng học", "@absences")
73 ])
74 p.add_tools(hover)
75
76 # Thiết lập click policy cho Legend để ẩn/hiện các nhóm
77 p.legend.click_policy = "hide"
78
79 # Hiển thị biểu đồ
80 show(column(p, slider))
81

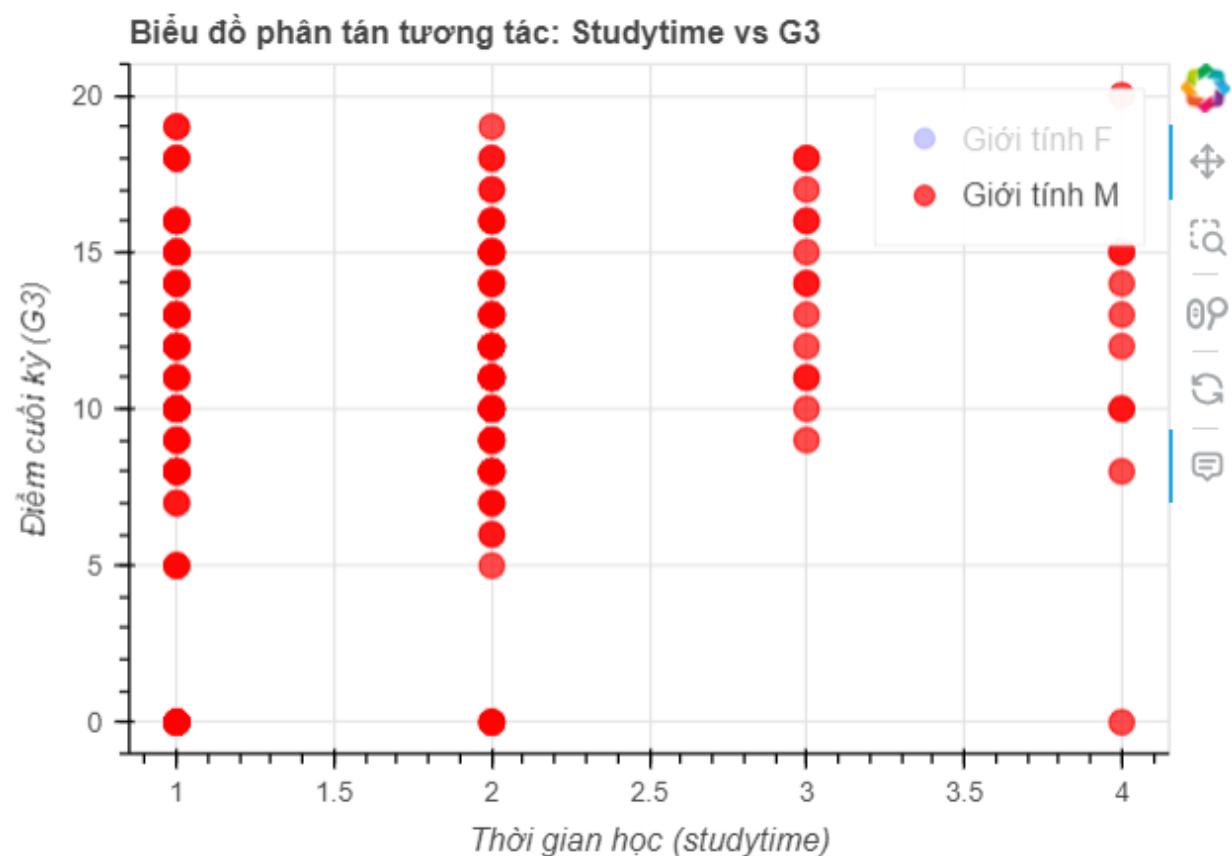
```

Hình 2.2.12. Mã nguồn biểu đồ phân tán tương tác theo “Studytime” và “G3”

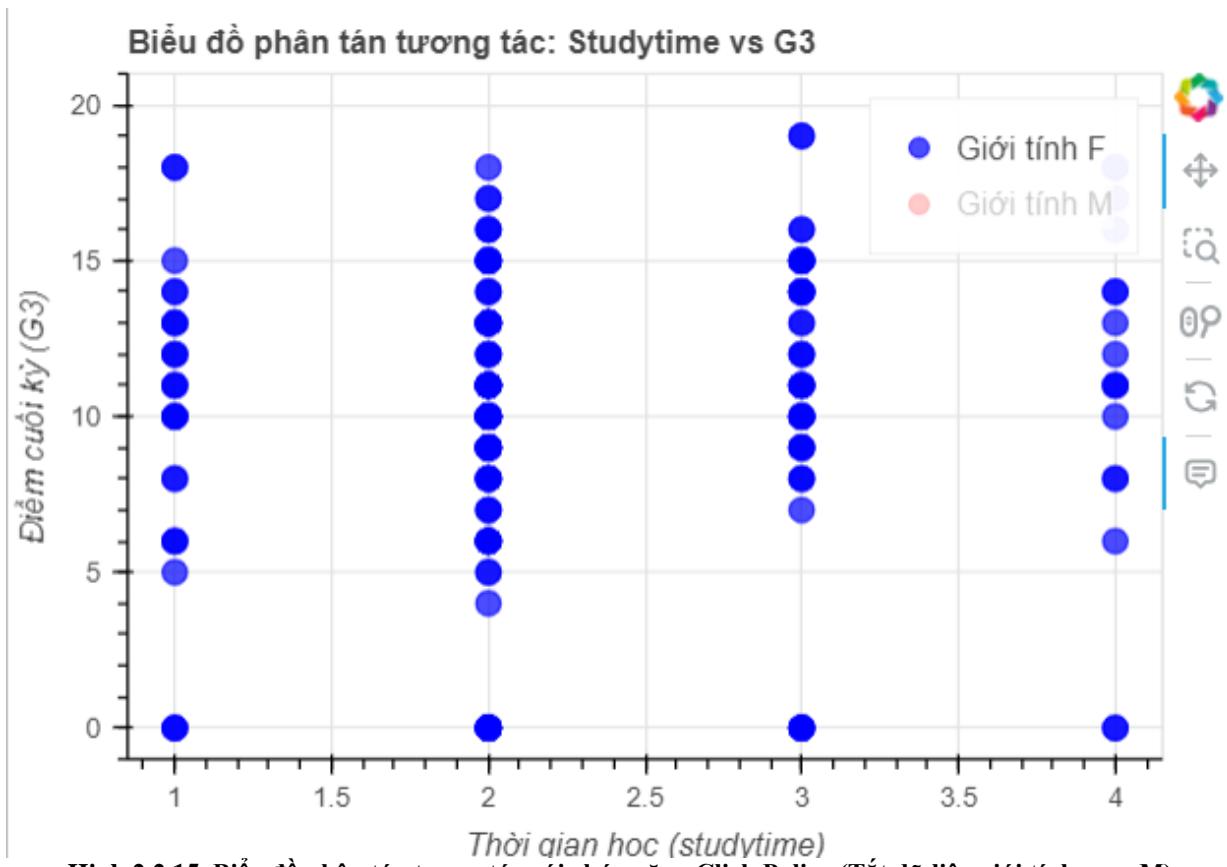
- Kết quả biểu đồ:



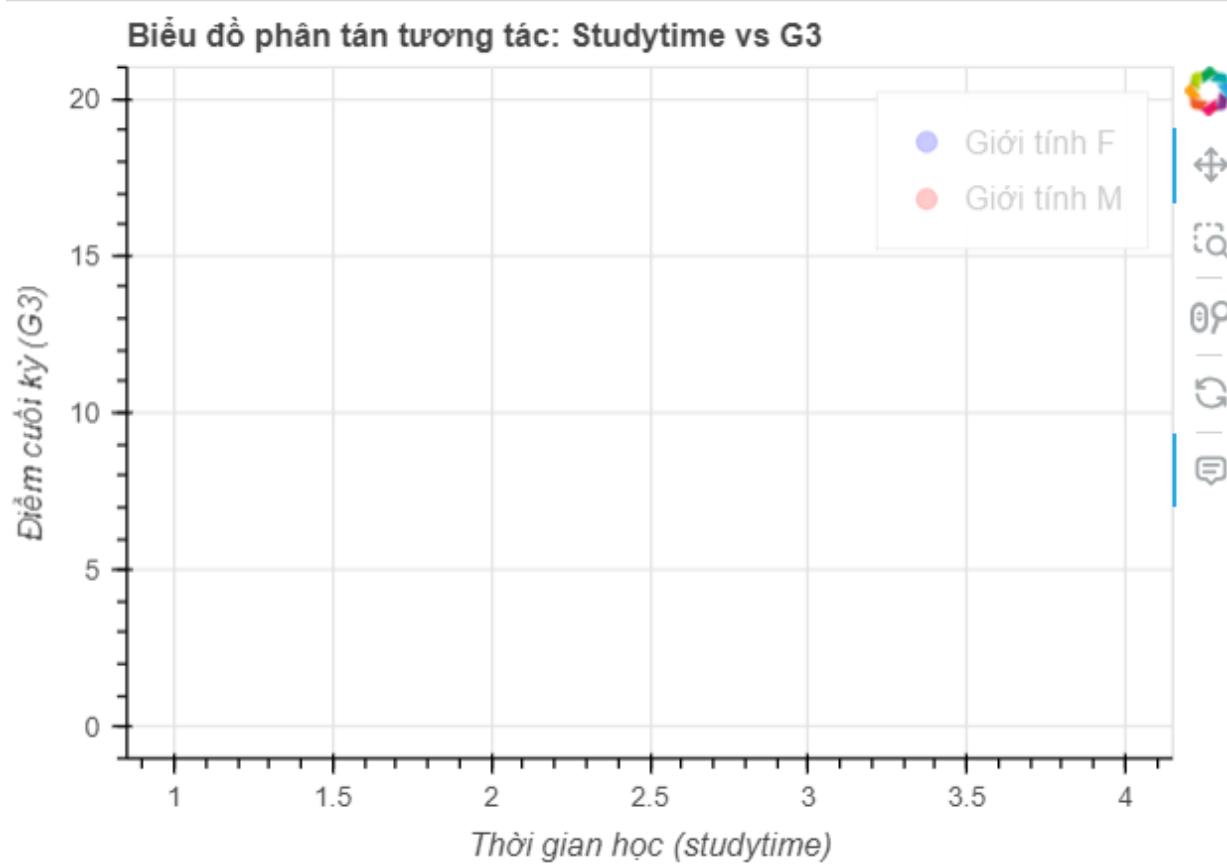
Hình 2.2.13. Biểu đồ phân tán tương tác với chức năng HoverTool



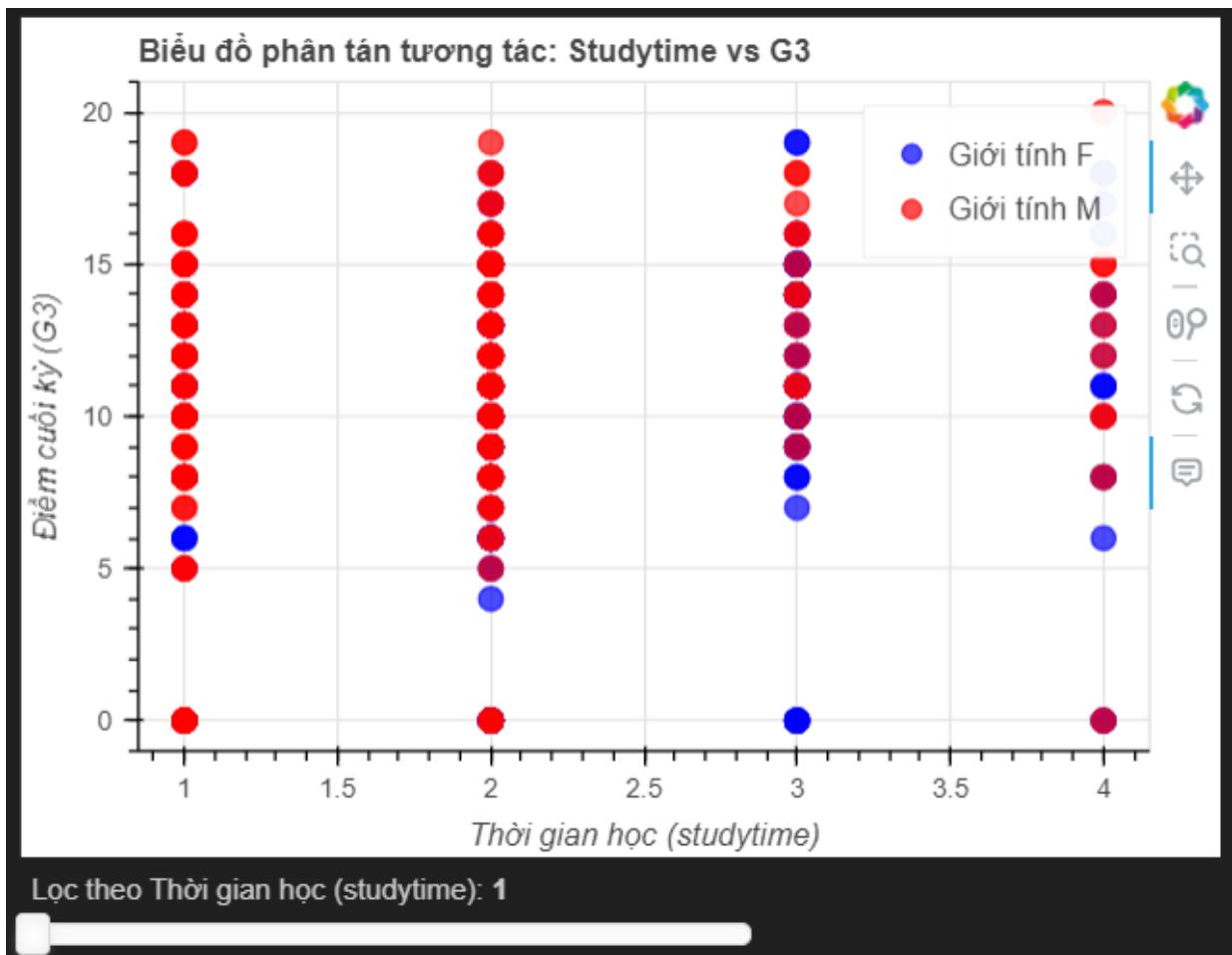
Hình 2.2.14. Biểu đồ phân tán tương tác với chức năng Click Policy (Tắt dữ liệu giới tính nữ F)



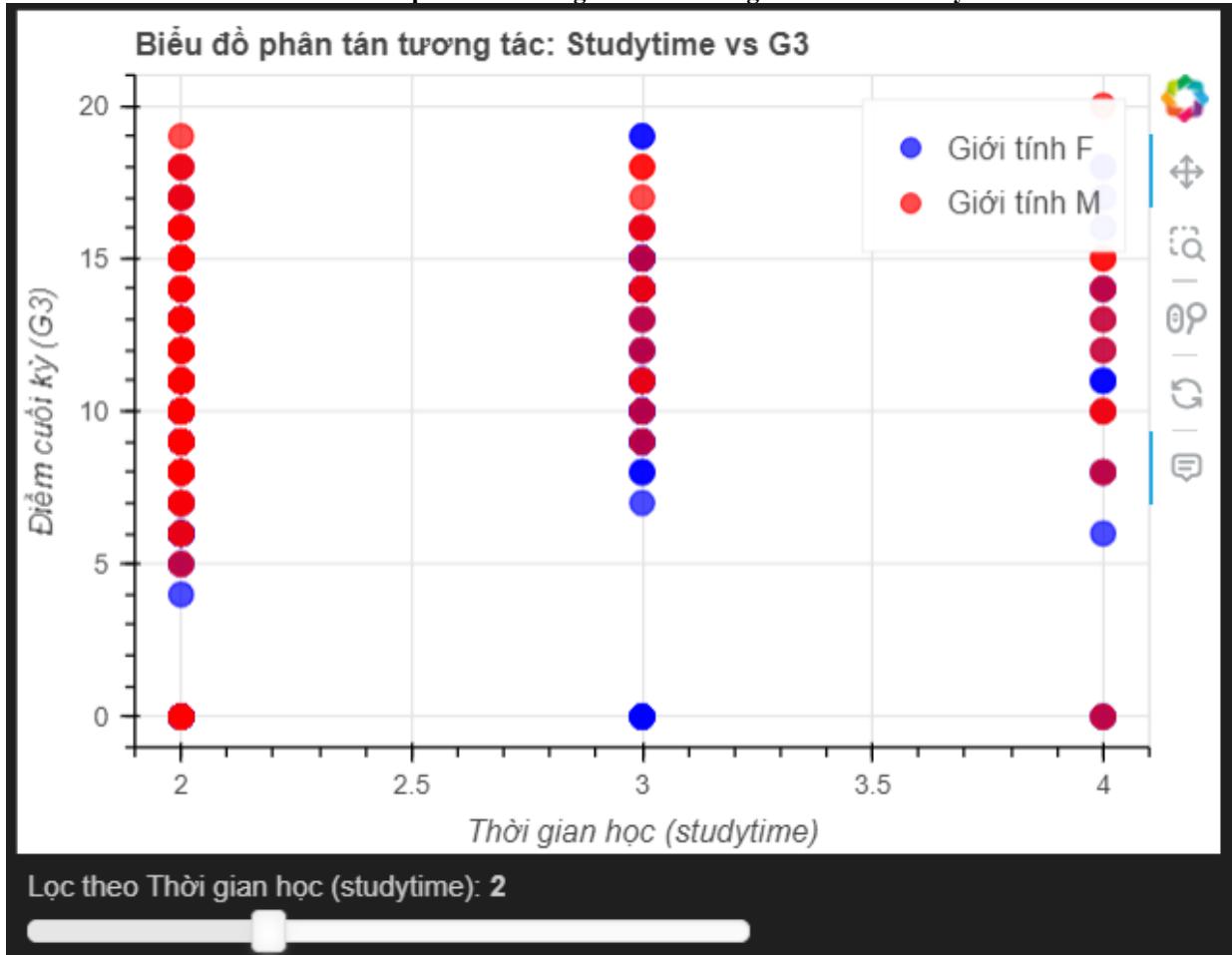
Hình 2.2.15. Biểu đồ phân tán tương tác với chức năng Click Policy (Tắt dữ liệu giới tính nam M)



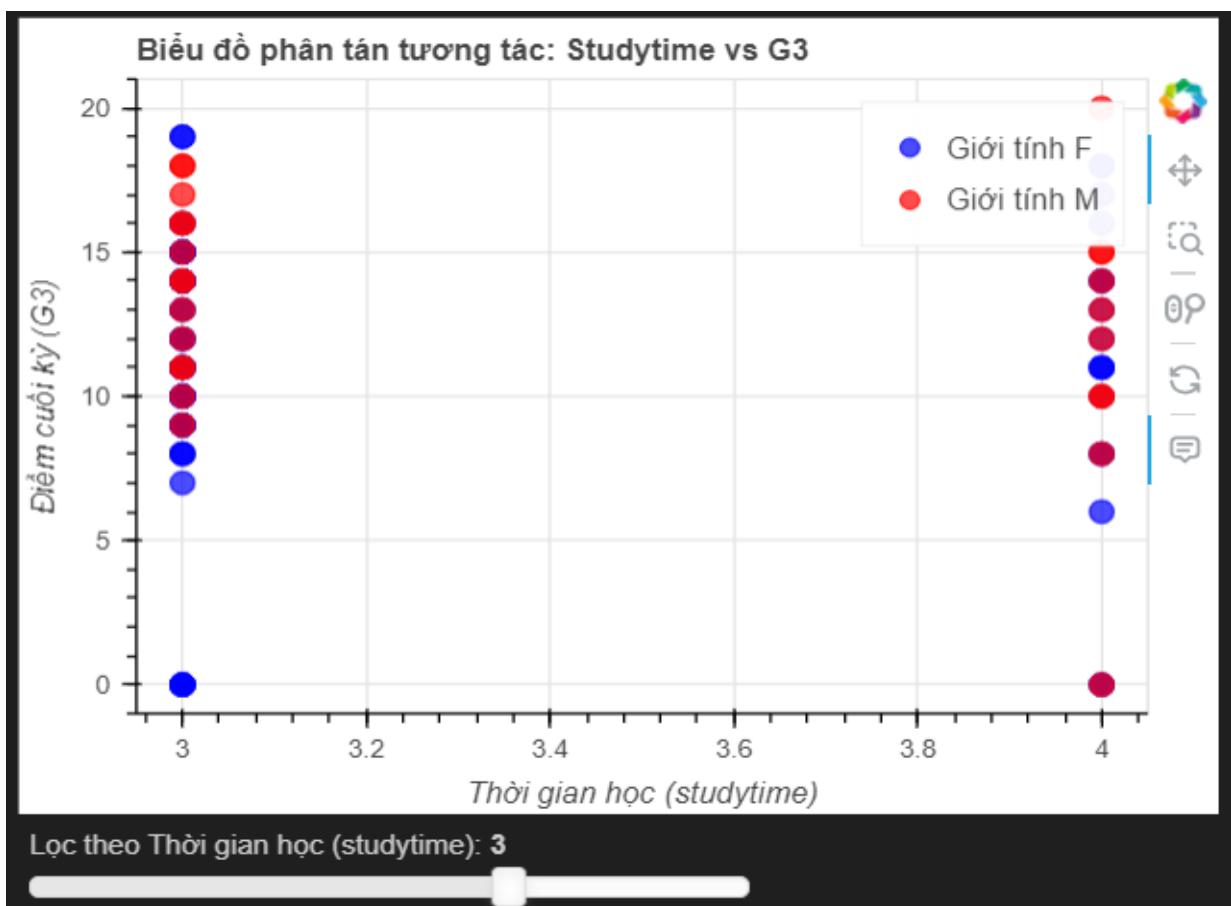
Hình 2.2.16 Biểu đồ phân tán tương tác với chức năng Click Policy (Tắt hết dữ liệu)



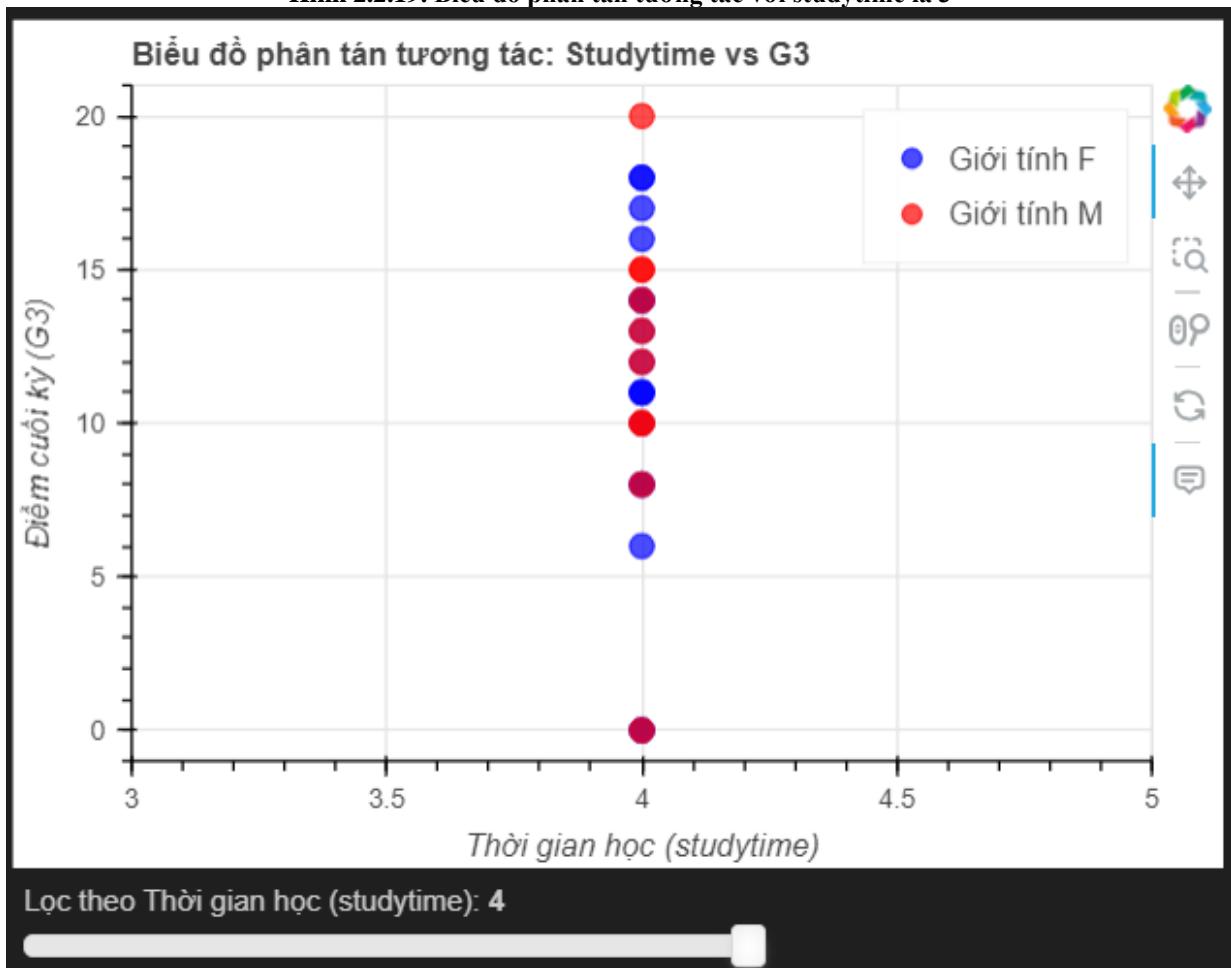
Hình 2.2.17. Biểu đồ phân tán tương tác theo mã nguồn trên với studytime là 1



Hình 2.2.18. Biểu đồ phân tán tương tác với studytime là 2



Hình 2.2.19. Biểu đồ phân tán tương tác với studytime là 3



Hình 2.2.20. Biểu đồ phân tán tương tác với studytime là 4

- Miêu tả biểu đồ:

Biểu đồ phân tán tương tác - thư viện Bokeh: Để tạo dữ liệu với biểu đồ phân tán tương tác thể hiện dữ liệu giữa: *studytme* và *G3* với *HoverTool* hiển thị *age*, *sex*, *absences*.

Bước	Mô Tả
1. Import thư viện	Import thư viện <code>bokeh</code> vào hệ thống cùng với các thành phần cần thiết như <code>figure</code> , <code>output_notebook</code> , <code>HoverTool</code> .
2. Sử dụng <code>output_notebook()</code>	Sử dụng hàm <code>output_notebook()</code> để hiển thị biểu đồ trực tiếp trong Jupyter Notebook. Đảm bảo biểu đồ hiển thị đúng trên notebook thay vì file.
3. Chuẩn bị dữ liệu	Tập dữ liệu lấy từ các cột: 'studytime', 'G3', 'age', 'sex', 'absences' và chuyển thành dữ liệu pandas. Tạo <code>ColumnDataSource</code> từ dữ liệu này.
4. Tạo khung biểu đồ	Tạo khung biểu đồ với các tham số sau: <ul style="list-style-type: none"> - <code>width=500</code>: Đặt chiều rộng của biểu đồ là 500 pixels. - <code>height=350</code>: Đặt chiều cao của biểu đồ là 350 pixels. - <code>title="BIỂU ĐỒ PHÂN TÁN TƯƠNG TÁC: STUDYTIME VS G3"</code>: Đặt tiêu đề cho biểu đồ. - <code>x_axis_label='THỜI GIAN HỌC (STUDYTIME)'</code>: Đặt tên cho trục X là "THỜI GIAN HỌC (STUDYTIME)". - <code>y_axis_label='ĐIỂM CUỐI KỲ (G3)'</code>: Đặt tên cho trục Y là "ĐIỂM CUỐI KỲ (G3)". - <code>tools="pan,wheel_zoom,box_zoom,reset"</code>: Định nghĩa các công cụ tương tác.

Hình 2.2.21. Mô tả biểu đồ (1)

5. Tạo biểu đồ phân tán	Tạo biểu đồ phân tán với các tham số sau: <ul style="list-style-type: none"> - <code>x="studytime"</code>: Xác định dữ liệu trục X là <code>studytime</code>. - <code>y="G3"</code>: Xác định dữ liệu trục Y là <code>G3</code>. - <code>size=11</code>: Đặt kích thước của các điểm. - <code>color="green"</code>: Đặt màu của các điểm là màu xanh lá cây. - <code>alpha=0.7</code>: Đặt độ trong suốt của các điểm (0 là trong suốt hoàn toàn, 1 là không trong suốt). - <code>source=source</code>: Cung cấp dữ liệu cho biểu đồ từ <code>ColumnDataSource</code>.
6. Thêm công cụ Hover	Thêm công cụ Hover để hiển thị thông tin khi di chuột qua các điểm dữ liệu: <ul style="list-style-type: none"> - <code>tooltips</code>: Danh sách các thông tin hiển thị: <ul style="list-style-type: none"> - "Thời gian học", "@studytime": Hiển thị thời gian học. - "Điểm cuối kỳ", "@G3": Hiển thị điểm cuối kỳ. - "Tuổi", "@age": Hiển thị tuổi học sinh. - "Giới tính", "@sex": Hiển thị giới tính học sinh. - "Vắng học", "@absences": Hiển thị số lần vắng học. - <code>p.add_tools(hover)</code>: Thêm công cụ Hover vào biểu đồ để hiển thị thông tin khi di chuột qua điểm dữ liệu.

Hình 2.2.22. Mô tả biểu đồ (2)

- Mục đích: Biểu đồ phân tán tương tác (interactive scatter plot) được lựa chọn nhằm phân tích mối quan hệ giữa thời gian học (*studytme*) và điểm cuối kỳ (*G3*), đồng thời so sánh giữa hai giới tính (nam và nữ). Việc tích hợp thanh trượt (slider) giúp người dùng lọc và khám phá dữ liệu theo thời gian học một cách linh hoạt, hỗ trợ việc trực quan hóa và phân tích sâu sắc hơn.

- Lý do chọn:

. Tương tác trực tiếp: Sử dụng Bokeh cho phép người dùng tương tác với biểu đồ, giúp họ khám phá thêm thông tin chi tiết về từng điểm dữ liệu (như tuổi, giới tính, số lần vắng học) khi di chuột qua các điểm.

. Hiển thị nhiều thông tin: Thông qua công cụ Hover, người dùng có thể dễ dàng truy cập các thông tin bổ sung như tuổi, giới tính, và số lần vắng học, từ đó cung cấp cái nhìn sâu sắc về mối quan hệ giữa thời gian học và điểm số.

. Thích hợp cho phân tích phức tạp: Khi có nhiều yếu tố ảnh hưởng đến kết quả học tập, như số lần vắng học hay tuổi tác, biểu đồ phân tán tương tác sẽ giúp làm rõ các mối liên hệ này và giúp người dùng đưa ra quyết định dễ dàng hơn.

- Kết luận rút ra từ biểu đồ:

+ Có một **xu hướng tích cực nhẹ** giữa thời gian học và điểm G3, đặc biệt rõ rệt hơn ở những học sinh dành nhiều thời gian học hơn ($\text{studytime} \geq 3$). Điều này gợi ý rằng **tăng thời gian học có thể góp phần nâng cao kết quả học tập**, mặc dù mối tương quan không hoàn toàn tuyến tính.

+ **Sự khác biệt giữa hai giới tính (nam và nữ)** không quá rõ rệt về xu hướng phân bố. Cả hai giới đều có sự phân tán điểm số tương đối đồng đều theo từng mức thời gian học, cho thấy **giới tính không phải là yếu tố quyết định rõ ràng đến hiệu quả học tập trong mối liên hệ với thời gian học**.

- Mã nguồn tạo biểu đồ theo “Row layout”:

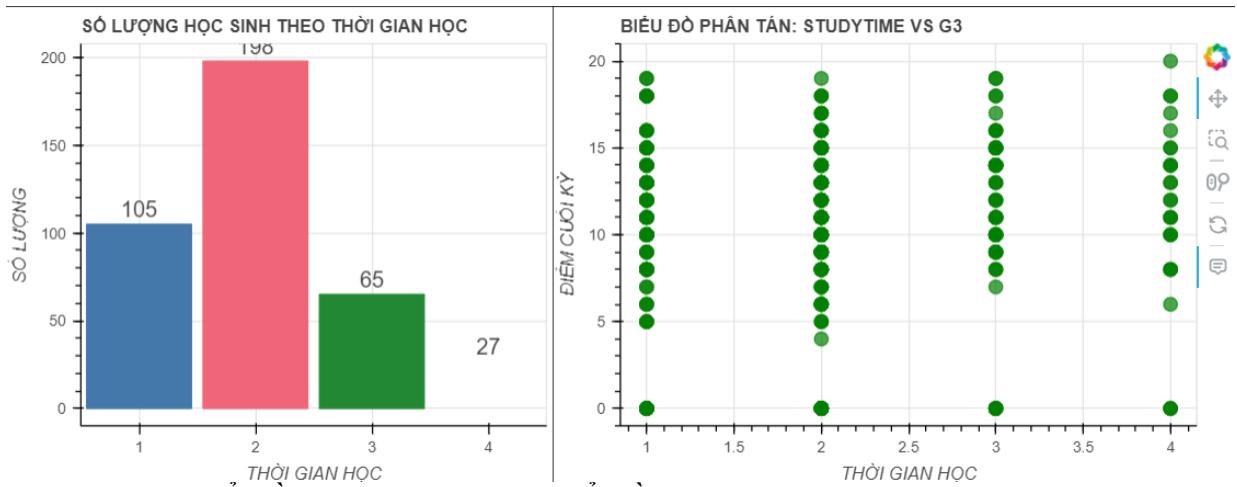
```

1  output_notebook()
2
3  # Tính số lượng theo studytime
4  count_study = df_pandas.groupby('studytime').size().reset_index(name='count')
5
6  # Tạo dữ liệu cho biểu đồ cột
7  source_bar = ColumnDataSource(data=dict(
8      studytime=[str(x) for x in count_study['studytime']],
9      count=count_study['count'],
10     color=Bright3[:len(count_study)]
11    ))
12
13  # Biểu đồ cột
14  bar_plot = figure(x_range=source_bar.data['studytime'], width=400, height=350,
15                     title="SỐ LƯỢNG HỌC SINH THEO THỜI GIAN HỌC",
16                     x_axis_label="THỜI GIAN HỌC", y_axis_label="SỐ LƯỢNG",
17                     toolbar_location=None)
18
19  bar_plot.vbar(x='studytime', top='count', width=0.9, color='color', source=source_bar)
20
21  # Thêm nhãn số lượng
22  labels = LabelSet(x='studytime', y='count', text='count', y_offset=5,
23                     text_align='center', source=source_bar)
24  bar_plot.add_layout(labels)
25
26  # Biểu đồ phân tán
27  df = df_pandas[['studytime', 'G3', 'age', 'sex', 'absences']]
28  source_scatter = ColumnDataSource(df)
29
30  scatter_plot = figure(width=500, height=350,
31                        title="BIỂU ĐỒ PHÂN TÁN: STUDYTIME VS G3",
32                        x_axis_label='THỜI GIAN HỌC', y_axis_label='ĐIỂM CUỐI KỲ',
33                        tools="pan,wheel_zoom,box_zoom,reset")
34
35  scatter_plot.scatter(x="studytime", y="G3", size=10, color="green", alpha=0.7, source=source_scatter)
36
37  hover = HoverTool(tooltips=[
38      ("Thời gian học", "@studytime"),
39      ("Điểm cuối kỳ", "@G3"),
40      ("Tuổi", "@age"),
41      ("Giới tính", "@sex"),
42      ("Vắng học", "@absences")
43  ])
44  scatter_plot.add_tools(hover)
45
46  # Xếp hai biểu đồ theo hàng ngang
47  layout = row(bar_plot, scatter_plot)
48
49  # Hiển thị
50  show(layout)

```

Hình 2.2.23. Biểu đồ cột cho studytime và biểu đồ phân tán tương tác theo định dạng row layout

- Kết quả biểu đồ:



Hình 2.2.24. Biểu đồ cột với "Studytime" và biểu đồ phân tán tương tác giữa "G3" và "Studytime"

- Mục đích: Biểu đồ này được thiết kế nhằm mục tiêu so sánh trực quan giữa phân bố số lượng học sinh theo thời gian học (studytime) và mối quan hệ giữa thời gian học với điểm số cuối kỳ (G3). Việc đặt hai biểu đồ cạnh nhau giúp người đọc vừa nắm được tần suất, vừa quan sát xu hướng kết quả học tập theo mức độ đầu tư thời gian.

- Lý do chọn:

+ **Biểu đồ cột** giúp mô tả phân bố số lượng học sinh theo từng mức thời gian học một cách trực quan, dễ hiểu, phù hợp với biến phân loại (studytime).

+ **Biểu đồ phân tán** cho phép theo dõi mối tương quan giữa hai biến số liên tục (studytime và G3), đồng thời cho phép khai thác thêm thông tin như giới tính, tuổi, số buổi nghỉ học thông qua HoverTool.

+ Việc kết hợp cả hai biểu đồ trong cùng một layout **tăng hiệu quả so sánh và phân tích**, cho phép xác định nhanh những nhóm học sinh nổi bật hoặc có đặc điểm bất thường.

+ **Tính tương tác cao** (zoom, hover, pan...) từ thư viện Bokeh giúp người dùng **khám phá dữ liệu sâu hơn** một cách linh hoạt và trực quan.

CHƯƠNG 3. PHÂN TÍCH XU HƯỚNG VÀ KIỂM ĐỊNH

3.1. Yêu cầu

- Phân tích xu hướng từ biểu đồ và kết quả kiểm định từ hai phương pháp:
 - + T-test
 - + Z-test
 - + Chi – square

3.2. Kết quả

a) T-test

- Giả thuyết: "Học sinh có số ngày nghỉ cao (absences > trung bình) thì có trung bình điểm G3 thấp hơn hơn đáng kể so với học sinh có số ngày nghỉ thấp (absences <= trung bình)
- Yêu cầu: Tạo hai nhóm dựa trên trung bình absences, thực hiện t-test, báo cáo p-value và kết luận (alpha = 0.05)
 - Kết quả kiểm định:
 - + Mô tả giả thuyết:
 - .Ta đặt: H_0 : Không có sự khác biệt điểm G3 giữa hai nhóm học sinh: nghỉ nhiều và nghỉ ít.
 - . Ta đặt: H_1 : Điểm trung bình G3 của nhóm có số ngày nghỉ cao hơn mức trung bình thấp hơn đáng kể so với học sinh có số ngày nghỉ thấp.

```
1 df = df_pandas[['absences', 'G3']]
2
3 absences = df["absences"]
4 G3 = df["G3"]
5
6 mean_absences = np.mean(absences)
7
8 group_low = G3[absences <= mean_absences] # nghỉ ít
9 group_high = G3[absences > mean_absences] # nghỉ nhiều
10
11
12 # T-test (2 mẫu độc lập)
13 t_stat, p_value = ttest_ind(group_high, group_low, equal_var=False)
14
15 print(f"T-statistic: {t_stat:.4f}")
16 print(f"P-value: {p_value:.4f}")
17
18 # Kết luận
19 alpha = 0.05
20 if p_value < alpha:
21     print("⇒ Bác bỏ  $H_0$ : Có sự khác biệt G3 giữa học sinh nghỉ nhiều và nghỉ ít.")
22 else:
23     print("⇒ Không bác bỏ  $H_0$ : Không có bằng chứng cho thấy G3 khác biệt rõ.")
```

Hình 3.2.1. Mã nguồn kiểm định T-Test

+ Các ký hiệu cần dùng trong việc tính toán của T-test:

Ký hiệu	Ý nghĩa	Biến trong code
\bar{X}_1	Trung bình mẫu của nhóm 1 (nhóm học sinh nghỉ ít)	<code>group_low.mean()</code> (Nhóm học sinh có số buổi nghỉ ≤ trung bình)
\bar{X}_2	Trung bình mẫu của nhóm 2 (nhóm học sinh nghỉ nhiều)	<code>group_high.mean()</code> (Nhóm học sinh có số buổi nghỉ > trung bình)
s_1^2	Phương sai mẫu của nhóm 1 (nhóm học sinh nghỉ ít)	<code>group_low.var(ddof=1)</code> (Phương sai mẫu của nhóm học sinh nghỉ ít)
s_2^2	Phương sai mẫu của nhóm 2 (nhóm học sinh nghỉ nhiều)	<code>group_high.var(ddof=1)</code> (Phương sai mẫu của nhóm học sinh nghỉ nhiều)
n_1	Cỡ mẫu (số học sinh trong nhóm 1)	<code>len(group_low)</code> (Số học sinh có số buổi nghỉ ≤ trung bình)
n_2	Cỡ mẫu (số học sinh trong nhóm 2)	<code>len(group_high)</code> (Số học sinh có số buổi nghỉ > trung bình)
t	Giá trị t-statistic (kiểm định T)	Tính toán trong <code>ttest_ind(group_high, group_low, equal_var=False)</code> (Kiểm định T giữa hai nhóm)
df	Độ tự do (degrees of freedom)	Tính toán tự động trong <code>ttest_ind()</code> (Độ tự do tính toán từ cỡ mẫu và phương sai mẫu)
p	P-value (xác suất)	Tính toán trong <code>ttest_ind(group_high, group_low, equal_var=False)</code> (Giá trị P từ kiểm định T)

Hình 3.2.2. Giải thích kí hiệu cần dùng trong tính toán T-Test

+ Cách xử lý dữ liệu:

1. Chuyển đổi dữ liệu thành dataframe Pandas của hai cột ‘absences’ và ‘G3’:

+ `df = df_pandas[['absences', 'G3']]`: chuyển đổi dữ liệu thành dạng pandas

+ `absences = df["absences"]`: biến absences chứa dữ liệu cột ‘absences’

+ `G3 = df["G3"]`: biến G3 chứa dữ liệu cột ‘G3’

2. Tính mức nghỉ trung bình của tất cả học sinh:

+ `mean_absences = np.mean(absences)`: dùng hàm ‘mean’ của thư viện numpy (np) nhằm tính mức nghỉ trung bình để làm mức so sánh với các nhóm học sinh.

3. Lọc nhóm học sinh:

+ `group_low = G3[absences <= mean_absences]`: Lọc nhóm học sinh có số buổi nghỉ ít hơn hoặc bằng trung bình (mean_absences).

+ Biến: ‘group_low’ là nhóm học sinh có ít buổi nghỉ.

+ `group_high = G3[absences > mean_absences]`: Lọc nhóm học sinh có số buổi nghỉ nhiều hơn trung bình (mean_absences).

+ Biến: ‘group_high’ là nhóm học sinh có nhiều buổi nghỉ.

4. Thực hiện kiểm định T hai mẫu độc lập:

+ `t_stat, p_value = ttest_ind(group_high, group_low, equal_var=False)`: Thực hiện kiểm định T hai mẫu độc lập giữa nhóm học sinh nghỉ nhiều và nhóm học sinh nghỉ ít.

+ ‘equal_var=False’: Cho phép kiểm định không giả định phương sai của hai nhóm bằng nhau.

+ Biến: ‘t_stat’ là giá trị t-statistic và ‘p_value’ là giá trị p từ kiểm định T.

+ Kết quả:

Số lượng buổi nghỉ trung bình: 5.708860759493671

Số lượng nhóm nghỉ ít: 249 và số lượng nhóm nghỉ nhiều: 146

Trung bình điểm nhóm nghỉ ít: 10.168674698795181

Trung bình điểm nhóm nghỉ nhiều: 10.835616438356164

T-statistic: 1.5676

P-value: 0.1178

⇒ Không bác bỏ H_0 : Không có bằng chứng cho thấy G3 khác biệt rõ.

Hình 3.2.3. Kết quả kiểm định T-Test

=> Ta bác bỏ giả thuyết H_1 rằng điểm trung bình của nhóm nghỉ nhiều thấp hơn đáng kể so với nhóm học sinh nghỉ ít. Chấp nhận giả thuyết H_0 rằng không có sự khác biệt điểm G3 giữa hai nhóm học sinh: nghỉ nhiều và nghỉ ít.

+ Ý nghĩa thực tiễn:

=> Ta có thể thấy được tuy trong tập dữ liệu này thì trung bình điểm của nhóm nghỉ nhiều và nhóm nghỉ ít có thể không có sự chênh lệch quá lớn. Nhưng vẫn chứng minh được rằng với số lượng ngày nghỉ tuy nhiều hay ít vẫn ảnh hưởng rất lớn đến số điểm trong quá trình học tập.

b) Z-test

- Giả thuyết: "Trung bình điểm G3 của học sinh học ít (studytime <= 2) thì khác biệt đáng kể so với trung bình điểm kỳ vọng của toàn bộ học sinh (giả định từ trung bình toàn mẫu.)"

- Yêu cầu: Tính trung bình G3 của nhóm studytime <= 2, so sánh với trung bình toàn bộ, thực hiện z-test, báo cáo p-value và kết luận

- Kết quả kiểm định:

+ Mô tả giả thuyết:

. Ta đặt: H_0 : Trung bình điểm của học sinh học ít bằng trung bình toàn bộ học sinh.

. Ta đặt: H_1 : Điểm trung bình G3 của nhóm học ít thì khác biệt đáng kể so với trung bình toàn học sinh.



```

1 df = df_pandas[['studytime','G3']]
2
3 group_low = df_pandas[df_pandas['studytime'] <= 2]['G3']
4
5 # Trung bình và độ lệch chuẩn của toàn bộ G3
6 mu = df_pandas['G3'].mean()
7 sigma = df_pandas['G3'].std(ddof=0) # độ lệch chuẩn toàn bộ, dùng ddof=0
8
9 # Trung bình nhóm học ít
10 x_bar = group_low.mean()
11 n = len(group_low)
12
13 # Tính z-value
14 z = (x_bar - mu) / (sigma / np.sqrt(n))
15
16 # Tính p-value (hai đầu)
17 p_value = 2 * (1 - norm.cdf(abs(z)))
18
19 # In kết quả
20 print(f"Trung bình toàn bộ: {mu:.4f}")
21 print(f"Trung bình nhóm học ít: {x_bar:.4f}")
22 print(f"Z-value: {z:.4f}")
23 print(f"P-value: {p_value:.4f}")
24
25 # Kết luận
26 alpha = 0.05
27 if p_value < alpha:
28     print("⇒ Bác bỏ H0: Trung bình G3 của học sinh học ít khác biệt đáng kể so với toàn bộ.")
29 else:
30     print("⇒ Không bác bỏ H0: Không có bằng chứng rõ ràng về sự khác biệt t.")

```

Hình 3.2.4. Mã nguồn kiểm định Z-Test

+ Các ký hiệu cần dùng cho việc tính toán Z-test:

Ký hiệu thống kê	Ý nghĩa	Biến trong code
\bar{x}	Trung bình mẫu (nhóm học ít)	x_bar = group_low.mean()
μ	Trung bình toàn bộ	mu = df_pandas['G3'].mean()
σ	Độ lệch chuẩn toàn bộ	sigma = df_pandas['G3'].std(ddof=0)
n	Cỡ mẫu (số học sinh học ít)	n = len(group_low)

Hình 3.2.5. Kí hiệu cần dùng trong tính toán Z-Test

+ Cách xử lý dữ liệu:

1. Chuyển đổi dữ liệu thành dataframe Pandas của hai cột ‘studytime’ và ‘G3’:

+ `df = df_pandas[['studytime','G3']]`: chuyển đổi dữ liệu thành dạng pandas

2. Lọc nhóm học sinh:

+ `group_low = df_pandas[df_pandas['studytime'] <= 2]['G3']`: biến “`group_low`” chứa dữ liệu điểm G3 của nhóm có **thời gian học ≤ 2**.

3. Tính trung bình điểm G3 của toàn bộ học sinh:

+ `mu = df_pandas['G3'].mean()`: biến “`mu`” chứa dữ liệu điểm trung bình cuối kì của toàn bộ học sinh

4. Tính độ lệch chuẩn của toàn bộ:

+ `sigma = df_pandas['G3'].std(ddof=0)`: biến “`sigma`” chứa dữ liệu tính toán của độ lệch chuẩn với “`ddof=0`” nghĩa là đang coi toàn bộ dữ liệu là tổng thể, không phải mẫu.

5. Tính trung bình điểm G3 của nhóm học ít:

+ `x_bar = group_low.mean()`: “`x_bar`” chứa dữ liệu của điểm trung bình cuối kì của nhóm học ít (thời gian ít hơn hoặc bằng 2 tiếng).

6. Tính tổng số lượng học sinh thuộc nhóm học ít:

+ `n = len(group_low)`: “`n`” là biến chứa số lượng học sinh của nhóm học ít.

7. Tính z-value:

+ `z = (x_bar - mu) / (sigma / np.sqrt(n))`: phép toán tính giá trị z

8. Tính p_value:

+ `p_value = 2 * (1 - norm.cdf(abs(z)))`: phép toán tính giá trị p

+ Giải thích phép toán:

Giải thích về phép tính:

1. Tính giá trị z:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Đây là công thức tính giá trị **z-value**, dùng để chuẩn hóa sự khác biệt giữa trung bình mẫu \bar{x} và trung bình toàn bộ μ so với độ lệch chuẩn σ và cỡ mẫu n .
- Cách tính này giúp đánh giá xem sự khác biệt giữa mẫu và tổng thể có ý nghĩa thống kê hay không, thông qua một chỉ số chuẩn hóa.

2. Tính p-value (hai đuôi):

$$p = 2 \times (1 - \text{norm.cdf}(|z|))$$

- `norm.cdf(abs(z))` tính giá trị hàm phân phối chuẩn (CDF) cho giá trị tuyệt đối của z.
- Sau đó, ta nhân với 2 để tính **p-value** cho kiểm định hai đuôi, với mục đích kiểm tra xem sự khác biệt có xảy ra ở cả hai phía của phân phối chuẩn hay không.

Hình 3.2.6. Giải thích phép toán trong mã nguồn Z-Test

+ Kết quả:

Trung bình toàn bộ: 10.4152
Trung bình nhóm học ít: 10.1287
Z-value: -1.0898
P-value: 0.2758
⇒ Không bác bỏ H0: Không có bằng chứng rõ ràng về sự khác biệt.

Hình 3.2.7. Kết quả của kiểm định Z-Test

=> Ta bác bỏ giả thuyết H1 rằng điểm trung bình G3 của nhóm học ít thì khác biệt đáng kể so với trung bình toàn học sinh. Chấp nhận giả thuyết H0 rằng trung bình điểm của học sinh học ít bằng xấp xỉ với trung bình toàn bộ học sinh.

+ Ý nghĩa thực tiễn:

=> Tập dữ liệu trên cho ta thấy được rằng điểm trung bình của toàn bộ học sinh tuy không cao hơn đáng kể với nhóm học sinh học ít nhưng một phần đã thể hiện được là việc học với thời gian ngắn có thể đem lại bất lợi trong việc cải thiện điểm số của học sinh. Và với mình chúng thực tế rằng với thời lượng học càng cao thì điểm số càng có phần được cải thiện hơn so với mặt bằng chung của học sinh, nhưng không mang ý nghĩa là học càng nhiều thì điểm sẽ càng cải thiện.

c) Chi-square

- Giả thuyết: Có mối quan hệ giữa mức thời gian học (studytime) và việc đạt điểm cao ($G3 \geq 12$) hay thấp ($G3 < 12$). Yêu cầu: Tạo biến nhị phân từ G3 (cao ≥ 12 , thấp < 12), xây dựng bảng tần số (contingency table) giữa studytime và biến mới, thực hiện chi-square, báo cáo p-value và kết luận.

- Kết quả kiểm định:

+ Mô tả giả thuyết:

. Ta đặt: H0 (Giả thuyết không): Không có mối quan hệ giữa thời gian học (studytime) và việc đạt điểm cao ($G3 \geq 12$) hay thấp ($G3 < 12$)

. Ta đặt: H1 (Giả thuyết đối): Có mối quan hệ giữa thời gian học và việc đạt điểm cao.

```

1 import pandas as pd
2 from scipy.stats import chi2_contingency
3
4 # Bước 1: Tạo biến nhị phân từ cột G3 (1 nếu G3 >= 12, 0 nếu < 12)
5 df['G3_binary'] = (df['G3'] >= 12).astype(int)
6
7 # Kiểm tra thử vài dòng đầu
8 print(df[['studytime', 'G3', 'G3_binary']].head(20))
9
10 print("\n")
11
12 # Bước 2: Tạo bảng tần số giữa studytime và G3_binary
13 contingency_table = pd.crosstab(df['studytime'], df['G3_binary'])
14
15 # In bảng tần số
16 print("Bảng tần số:")
17 print(contingency_table)
18
19 print("\n")
20
21 # Bước 3: Thực hiện kiểm định Chi-square
22 chi2, p_value, dof, expected = chi2_contingency(contingency_table)
23
24 print("\nKết quả kiểm định Chi-square:")
25 print(f"Chi-square statistic: {chi2:.4f}")
26 print(f"Degrees of freedom: {dof}")
27 print("Expected frequencies (tần số kỳ vọng nếu không có mối liên hệ):")
28 print(pd.DataFrame(expected, index=contingency_table.index, columns=contingency_table.columns))
29 print(f"p-value: {p_value:.4f}")
30
31 print("\n")
32
33 # Bước 4: Đưa ra kết luận
34 alpha = 0.05
35 print("Giả thuyết:")
36 print("H0: Không có mối quan hệ giữa thời gian học và việc đạt điểm cao (G3 ≥ 12).")
37 print("H1: Có mối quan hệ giữa thời gian học và việc đạt điểm cao (G3 ≥ 12).")
38
39 print("\n")
40
41 print("Kết luận kiểm định với mức ý nghĩa α = 0.05:")
42
43 if p_value < alpha:
44     print("Vì p-value < 0.05, bác bỏ H0.")
45     print("⇒ Kết luận: Có mối quan hệ giữa thời gian học và việc đạt điểm cao.")
46 else:
47     print("Vì p-value ≥ 0.05, không đủ bằng chứng để bác bỏ H0.")
48     print("⇒ Kết luận: Không tìm thấy mối quan hệ rõ ràng giữa thời gian học và việc đạt điểm cao.")
49

```

Hình 3.2.8. Mã nguồn phương pháp kiểm định Chi-square

+ Cách xử lý dữ liệu:

. `df['G3_binary'] = (df['G3'] >= 12).astype(int)`: Tạo biến nhị phân từ cột G3 với chỉ số là 1 nếu $G3 \geq 12$ điểm và 0 nếu $G3 < 12$

```
# Bước 1: Tạo biến nhị phân từ cột G3 (1 nếu G3 >= 12, 0 nếu < 12)
df['G3_binary'] = (df['G3'] >= 12).astype(int)
```

Hình 3.2.9. Tạo biến nhị phân

	studytime	G3	G3_binary
0		2 6	0
1		2 6	0
2		2 10	0
3		3 15	1
4		2 10	0
5		2 15	1
6		2 11	0
7		2 6	0
8		2 19	1
9		2 15	1
10		2 9	0

Hình 3.2.10. Giá trị mẫu từ các biến nhị phân

. `contingency_table = pd.crosstab(df['studytime'], df['G3_binary'])`: Tạo bảng tần số giữa studytime và G3_binary

```
# Bước 2: Tạo bảng tần số giữa studytime và G3_binary
contingency_table = pd.crosstab(df['studytime'], df['G3_binary'])
```

Hình 3.2.11. Tạo bảng tần số

Bảng tần số:		
G3_binary	0	1
studytime		
1	61	44
2	127	71
3	32	33
4	13	14

Hình 3.2.12. Kết quả của bảng tần số

. `chi2, p_value, dof, expected = chi2_contingency(contingency_table)`: Thực hiện việc kiểm định Chi-square

```
# Bước 3: Thực hiện kiểm định Chi-square
chi2, p_value, dof, expected = chi2_contingency(contingency_table)
```

Hình 3.2.13. Tính toán kiểm định Chi-square

- . **if p_value < alpha:** So sánh dữ liệu p_value và alpha = 0.05 để đưa ra quyết định cho kết luận cuối cùng

```
# Bước 4: Đưa ra kết luận
alpha = 0.05
print("Giả thuyết:")
print("H0: Không có mối quan hệ giữa thời gian học và việc đạt điểm cao (G3 ≥ 12).")
print("H1: Có mối quan hệ giữa thời gian học và việc đạt điểm cao (G3 ≥ 12).")

print("\n")

print("Kết luận kiểm định với mức ý nghĩa α = 0.05:")

if p_value < alpha:
    print("Vì p-value < 0.05, bác bỏ H0.")
    print("⇒ Kết luận: Có mối quan hệ giữa thời gian học và việc đạt điểm cao.")
else:
    print("Vì p-value ≥ 0.05, không đủ bằng chứng để bác bỏ H0.")
    print("⇒ Kết luận: Không tìm thấy mối quan hệ rõ ràng giữa thời gian học và việc đạt điểm cao.)
```

Hình 3.2.14. Đưa ra kết luận cuối cùng

- + Kết quả:

```
Giả thuyết:
H0: Không có mối quan hệ giữa thời gian học và việc đạt điểm cao (G3 ≥ 12).
H1: Có mối quan hệ giữa thời gian học và việc đạt điểm cao (G3 ≥ 12).

Kết luận kiểm định với mức ý nghĩa α = 0.05:
Vì p-value ≥ 0.05, không đủ bằng chứng để bác bỏ H0.
⇒ Kết luận: Không tìm thấy mối quan hệ rõ ràng giữa thời gian học và việc đạt điểm cao.
```

Hình 3.2.15. Kết luận của kết quả kiểm định Chi-square tìm mối liên hệ giữa Studytme và G3

=> Ta chấp nhận H0 rằng không có mối quan hệ rõ ràng giữa việc thời gian học cao sẽ tác động tích cực đến việc đạt điểm cao hoặc thấp. Và bác bỏ H1 rằng có sự tương tác và ảnh hưởng nhất định giữa hai đặc trưng này.

- + Ý nghĩa thực tiễn:

=> Kết quả kiểm định cho thấy không có đủ bằng chứng thống kê để khẳng định rằng thời gian học (studytme) có mối liên hệ rõ ràng với việc đạt điểm cao ($G3 \geq 12$) hay điểm thấp ($G3 < 12$).

Tuy nhiên, trên thực tế, khi quan sát dữ liệu:

- Những học sinh dành ít thời gian học (2-5 giờ) có xu hướng đạt điểm thấp hơn so với nhóm học sinh học nhiều hơn.
- Nhóm học nhiều giờ hơn (trên 5 giờ/tuần) nhìn chung có tỉ lệ đạt điểm cao nhỉnh hơn.

⇒ Điều này cho thấy:

Việc học với thời gian quá ngắn có thể là một bất lợi cho kết quả học tập, tuy nhiên học nhiều hơn không đồng nghĩa chắc chắn sẽ đạt điểm cao.

Tức là, tăng thời gian học có thể giúp cải thiện điểm số, nhưng không phải yếu tố quyết định duy nhất — chất lượng học, phương pháp học và các yếu tố khác cũng đóng vai trò quan trọng.

d) Phân tích xu hướng các biểu đồ

- **Biểu đồ cột:** Dựa vào biểu đồ cột đã trình bày phía trên, biểu đồ thể hiện rõ ràng về khả năng tập trung học của học sinh thông qua số lượng được thống kê theo từng mốc thời gian:

+ Tổng số lượng học sinh là 395 được phân chia thành các mốc thời gian học: Nhóm 1 (< 2 tiếng), nhóm 2 (2-5 tiếng), nhóm 3 (5-10 tiếng), 4 (>10 tiếng).

+ Số lượng học sinh tập trung ở mức 2 chiếm tỉ lệ cao nhất khoảng **50,13%** (198 học sinh). Cho thấy đây là khoảng thời gian học tập phổ biến nhất.

+ Mốc thời gian học có tỉ lệ thấp nhất nằm ở mức 4 chỉ chiếm một phần nhỏ là **6.84%** (27 học sinh).

+ Hai nhóm còn lại có tỷ lệ lần lượt là **26,58% (mức 1)** và **16,46% (mức 3)**, với sự chênh lệch đáng kể khoảng **10,12%** giữa hai nhóm này.

=> Kết quả này cho thấy phần lớn học sinh lựa chọn học trong khoảng thời gian vừa phải (mức 2), có thể là do đây là mốc thời gian cân bằng giữa hiệu quả và khả năng duy trì sự tập trung, đồng thời không gây cảm giác quá tải trong học tập.

- **Biểu đồ hồi quy tuyến tính:** Biểu đồ này tận dụng cột dữ liệu giới tính để phân tích mối quan hệ giữa điểm số cuối kì với thời lượng học tập của mỗi cá nhân như sau:

+ Biểu đồ này cho thấy mối quan hệ tích cực giữa thời gian học và điểm cuối kì. Dựa vào biểu đồ trên, ta có thể quan sát thấy rằng với số giờ học càng cao thì điểm số sẽ càng cao.

+ Đặc biệt, đường hồi quy **màu cam (nam)** luôn nằm phía trên so với đường **màu xanh dương (nữ)**, cho thấy trong cùng một thời lượng học tập, **nam sinh có xu hướng đạt điểm số cao hơn nữ sinh**.

+ Tuy nhiên, các điểm dữ liệu vẫn có sự **phân tán lớn**, biểu hiện bằng việc xuất hiện những học sinh học nhiều nhưng điểm số vẫn thấp, cho thấy thời gian học chỉ là một trong nhiều yếu tố ảnh hưởng đến kết quả học tập.

=> Kết luận được đưa ra rằng, với thời gian học càng cao thì điểm số của học sinh cũng được cải thiện đáng kể. Cùng với đó nam sinh có xu hướng đạt điểm cao hơn với nữ sinh ở cùng một mức học. Đây là hai xu hướng chung được thể hiện trong biểu đồ trên.

- **Biểu đồ phân tán tương tác:** Biểu đồ phân tán tương tác mô tả mối quan hệ giữa **thời gian học tập (studytime)** và **điểm số cuối kỳ (G3)**. Mỗi điểm tròn đại diện cho một học sinh.

+ Nhìn chung thì những học sinh có thời gian học cao hơn thì có xu hướng đạt điểm cuối kì cao hơn so với nhóm học ít.

+ Ở nhóm học sinh chỉ học mức **1-2**, điểm số có sự phân tán mạnh, trải rộng từ 0 đến 19 điểm. Điều này phản ánh **tính không ổn định** trong kết quả học tập của nhóm này.

+ Ngược lại, nhóm học mức **3-4** thể hiện sự tập trung điểm số cao hơn, ít xuất hiện học sinh điểm thấp. Các điểm dữ liệu tập trung hơn về phía trên, cho thấy **kết quả học ổn định và hiệu quả hơn**.

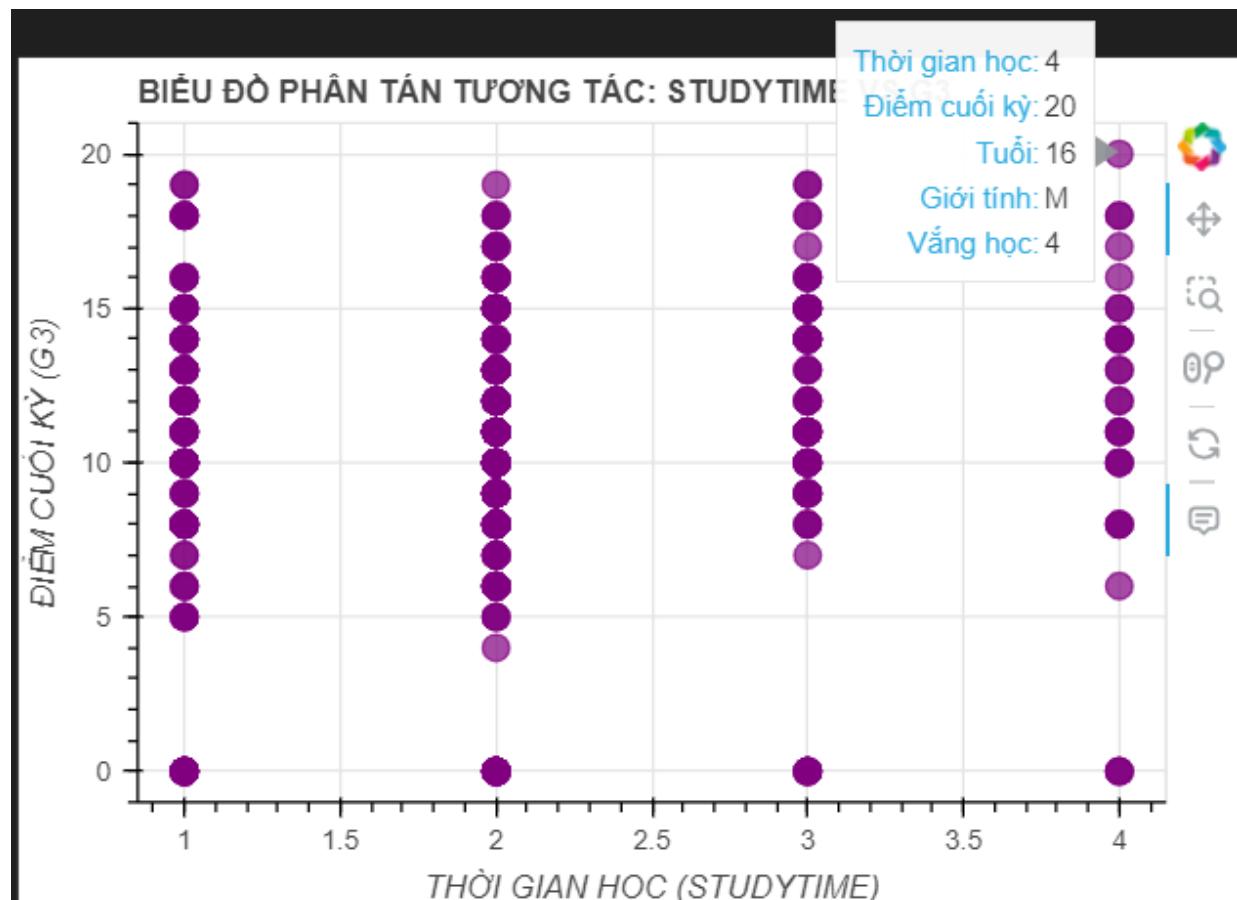
=> Biểu đồ này làm nổi bật xu hướng rằng, điểm số ở mức thời gian học ít (mức 1-2) thường được phân bố đều ở các mức điểm từ 0 đến 19 điểm, nhưng điểm cao vẫn chỉ chiếm số lượng nhỏ. Bên cạnh đó, với thời lượng học cao thì mức điểm thấp ít dần đi, mật độ điểm cao dày đặc hơn dao động từ 10 đến 20, tuy nhiên vẫn có các ngoại lệ.

CHƯƠNG 4. ĐỀ XUẤT PHƯƠNG PHÁP CẢI THIỆN KẾT QUẢ

4.1. Phương pháp

a) Phương pháp 1

- Tăng thời gian trung bình học lên từ dưới 2 tiếng đến 2-5 tiếng thành 5-10 tiếng hoặc trên 10 tiếng nhằm cải thiện điểm số. Vì theo thống kê qua các dữ liệu biểu đồ ta thấy được với mức thời gian học càng cao thì điểm cuối kì cũng được tăng lên đáng kể.



Hình 4.1.1. Biểu đồ lý giải cho đề xuất phương án 1

+ Giải thích:

. Ở mức thời gian học cao hơn thì mức dao động điểm G3 của học sinh cũng cao hơn đáng kể.

. Theo biểu đồ cho thấy: mức 2-5 đến 5-10 tiếng học thì dao động điểm trải dài từ 5-19 điểm nhưng ở mức trên 10 tiếng học điểm dao động được cải thiện hơn từ 10-20 điểm.

b) Phương pháp 2

- Với số điểm trung bình của nhóm học sinh có tổng buổi nghỉ trung bình dao động từ 1-6 buổi vẫn có sự chênh lệch đáng kể hơn với nhóm nghỉ nhiều trên 6 buổi. Đây là dữ liệu cho lý giải trên:

+ Mã nguồn:

```
df_pandas = math.toPandas()

Tabnine | Edit | Test | Explain | Document
def classify_absences(absences):
    if absences == 0:
        return 'Không vắng mặt'
    elif 1 <= absences <= 6:
        return 'Vắng trung bình'
    else:
        return 'Vắng nhiều'

df_pandas['Mức độ nghỉ'] = df_pandas['absences'].apply(classify_absences)

# Tính điểm trung bình G3 cho mỗi mức độ nghỉ
grouped = df_pandas.groupby('Mức độ nghỉ')['G3'].mean().reindex(['Không vắng mặt', 'Vắng trung bình', 'Vắng nhiều'])

# Vẽ biểu đồ cột
plt.figure(figsize=(8, 6))
grouped.plot(kind='bar', color=['green', 'orange', 'red'])

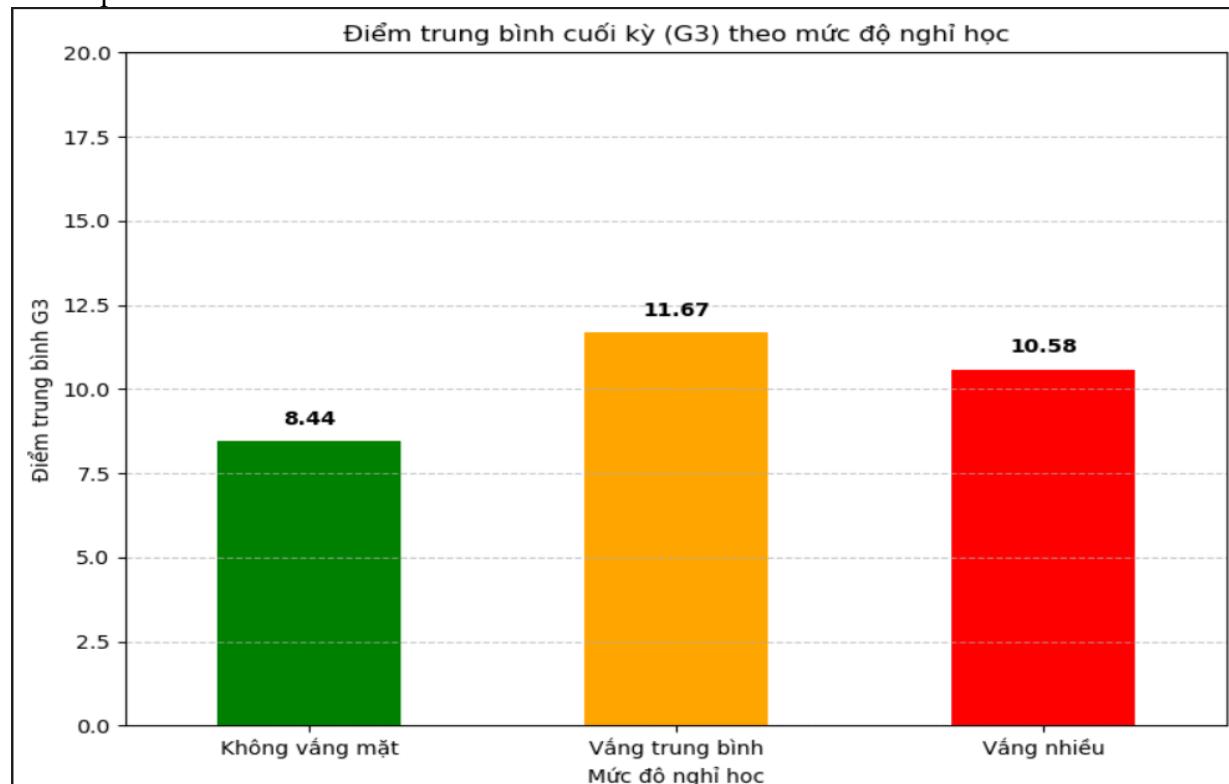
plt.title('Điểm trung bình cuối kỳ (G3) theo mức độ nghỉ học')
plt.xlabel('Mức độ nghỉ học')
plt.ylabel('Điểm trung bình G3')
plt.xticks(rotation=0)
plt.ylim(0, 20)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Hiển thị giá trị trên từng cột
for i, v in enumerate(grouped):
    plt.text(i, v + 0.5, f'{v:.2f}', ha='center', fontweight='bold')

plt.tight_layout()
plt.show()
```

Hình 4.1.2. Mã nguồn cho đề xuất phương pháp 2

+ Kết quả:



Hình 4.1.3. Biểu đồ thể hiện cho mức điểm trung bình giữa các nhóm nghỉ học

+ Giải thích:

- . Với 3 mức độ học sinh: không nghỉ, trung bình, nghỉ nhiều.
- . Mức điểm G3 trung bình của nhóm thứ 2 vẫn được duy trì cao hơn nhóm 3.
- . Không có ý nghĩa nếu tăng số lượng nghỉ lên của nhóm 1 thì điểm sẽ cao hơn.

+ Đề xuất: Nhóm nghỉ nhiều từ 6 buổi trở lên nên được nhà trường đưa ra phương án kiểm soát khả năng cúp, trốn học của học sinh nhằm tăng hệ số điểm cuối kì trở nên khả quan, tích cực hơn.

CHƯƠNG 5. TỔNG KẾT

5.1. Tổng kết

a) Chương 1

Chương 1 đóng vai trò **nền tảng cho toàn bộ quá trình phân tích**, với mục tiêu làm rõ cấu trúc, nội dung và chất lượng của tập dữ liệu student-mat.csv. Việc hiểu rõ đặc trưng của dữ liệu là bước đầu tiên và thiết yếu để đảm bảo rằng các phân tích và kết luận sau này được thực hiện trên một nền dữ liệu đầy đủ, hợp lệ và đáng tin cậy.

Bộ dữ liệu bao gồm **395 học sinh và 33 cột thông tin**, được chia thành ba nhóm lớn:

- **Thông tin cá nhân và gia đình** (tuổi, giới tính, nghề nghiệp và học vấn của phụ huynh, quy mô gia đình...),
- **Thông tin học tập** (điểm số G1, G2, G3, số lần rớt môn, thời gian học tập, sự hỗ trợ học tập, dự định học cao hơn...),
- **Hành vi xã hội và sinh hoạt cá nhân** (thời gian rảnh, ra ngoài, tiêu thụ rượu, sức khỏe, số buổi vắng mặt...).

Phân tích chương 1 tập trung vào các yếu tố sau:

- **Kiểm tra chất lượng dữ liệu:** Không phát hiện giá trị thiếu (null), dữ liệu hoàn chỉnh 100%.
- **Phát hiện ngoại lệ (outliers):** Một số cột như absences, age, G2 tồn tại giá trị ngoại lệ, tuy không phổ biến nhưng cần lưu ý khi mô hình hóa.
- **Thống kê mô tả các biến:** Giúp hiểu rõ sự phân bố, trung bình, độ lệch chuẩn và sự đa dạng của dữ liệu.

Mục tiêu quan trọng của chương này là:

- **Đánh giá mức độ tin cậy và sẵn sàng của dữ liệu** cho các phân tích nâng cao.
- **Hiểu rõ bản chất của từng biến**, để lựa chọn đúng kỹ thuật trực quan hóa, kiểm định và mô hình phù hợp trong các chương sau.
- **Đặt nền tảng cho việc phát hiện các mối quan hệ tiềm năng** giữa đặc trưng học tập và kết quả học tập (G3), từ đó định hướng phân tích thống kê và đề xuất cải thiện.

Việc phân tích chương 1 giúp đảm bảo rằng **mọi bước sau được thực hiện trên cơ sở dữ liệu đã được hiểu rõ, làm sạch, và xác định rõ ràng** — một yêu cầu quan trọng trong mọi nghiên cứu phân tích dữ liệu nghiêm túc.

b) Chương 2

Thư viện Matplotlib

Các biểu đồ sử dụng:

- **Biểu đồ cột:** Trực quan hóa phân bố số lượng học sinh theo mức độ studytime.
- **Biểu đồ phân tán:** Thể hiện mối quan hệ giữa studytime và G3, có phân biệt giới tính (qua màu sắc).

Kết luận từ biểu đồ:

- **Biểu đồ cột** cho thấy đa số học sinh học từ 2–5 giờ mỗi tuần, nhóm học nhiều hơn 10 giờ rất ít. Điều này phản ánh sự phổ biến của mức thời gian học trung bình và cho thấy các nhóm học sinh không phân bố đều.
- **Biểu đồ phân tán** chỉ ra rằng học sinh có thời gian học cao hơn có xu hướng đạt điểm G3 cao hơn. Tuy nhiên, xu hướng không hoàn toàn tuyến tính. Khi phân tích theo giới tính, biểu đồ không cho thấy sự khác biệt rõ ràng – cả nam và nữ đều có điểm cao nếu học đủ thời gian.

Xu hướng chung:

- => Matplotlib cung cấp cái nhìn định lượng sơ cấp về **sự phân bố thời gian học và mối liên hệ cơ bản với điểm số**.
- => Sự ảnh hưởng của giới tính trong thư viện này **chưa được thể hiện nổi bật**, chủ yếu nhấn mạnh rằng **thời gian học nhiều hơn có xu hướng dẫn đến điểm số cao hơn**, nhưng không phải yếu tố quyết định tuyệt đối.

Thư viện Seaborn

Các biểu đồ sử dụng:

- **Biểu đồ hồi quy tuyến tính (lmplot)**: Phân tích mối quan hệ giữa studytime và G3, phân theo giới tính.
- **Biểu đồ hộp (boxplot)**: Thể hiện sự phân phối điểm G3 theo từng mức studytime.
- **Biểu đồ cặp (pairplot)**: Khảo sát quan hệ giữa G1, G2, G3 và studytime, có phân biệt giới tính.

Kết luận từ biểu đồ:

- **Hồi quy tuyến tính** cho thấy studytime có ảnh hưởng tích cực đến điểm G3, với độ dốc lớn hơn ở nhóm nam → nam sinh có xu hướng hưởng lợi rõ rệt từ việc học nhiều hơn.
- **Boxplot** cho thấy điểm trung vị của G3 tăng theo studytime, đặc biệt ở nhóm học 3–4 đơn vị thời gian. Nhóm học nhiều có điểm ổn định và cao hơn.
- **Pairplot** chỉ ra mối quan hệ tuyến tính mạnh giữa G1, G2 và G3 (dự đoán được G3 tốt), nhưng studytime có quan hệ yếu hơn. Giới tính không tạo ra sự khác biệt rõ về phân phối điểm số.

Xu hướng chung:

- => Seaborn làm nổi bật được **mối quan hệ giữa thời gian học và kết quả học tập dưới dạng định lượng và xu hướng**.
- => Yếu tố giới tính thể hiện rõ hơn so với Matplotlib: **nam sinh có xu hướng cải thiện điểm tốt hơn khi tăng thời gian học**, nhưng vẫn cần xét đến độ phân tán.
- => Thư viện này nhấn mạnh rằng **mối liên hệ giữa thời gian học và điểm số là tích cực nhưng phụ thuộc vào giới tính và không hoàn toàn tuyến tính**.

Thư viện Bokeh

Các biểu đồ sử dụng:

- **Biểu đồ phân tán tương tác**: Giữa studytime và G3, có sử dụng công cụ tương tác như HoverTool, Click Policy, và Slider lọc theo studytime.
- **Biểu đồ dạng hàng (row layout)**: Kết hợp biểu đồ cột với phân tán tương tác.

Kết luận từ biểu đồ:

- Biểu đồ phân tán tương tác cho thấy: **khi người dùng điều chỉnh slider studytime, điểm số G3 có xu hướng tăng nhẹ ở mức studytime ≥ 3** .
- Khi tắt/mở dữ liệu theo giới tính, ta thấy **cả nam và nữ đều có phân bố điểm tương đối giống nhau**, nhưng các điểm cá biệt (outlier) rõ ràng hơn, giúp phân tích sâu hơn.
- Biểu đồ dạng hàng giúp **so sánh trực tiếp** giữa phân bố thời gian học và phân bố điểm số.

Xu hướng chung:

- => Bokeh cung cấp **trực quan tương tác mạnh mẽ**, giúp phát hiện xu hướng không thể hiện rõ trong biểu đồ tĩnh.
- => Yếu tố giới tính vẫn không quá nổi bật nhưng cho phép **khám phá sâu sắc các điểm cá biệt hoặc nhóm ẩn** khi người dùng thao tác trực tiếp.
- => Mối quan hệ giữa thời gian học và điểm số vẫn **mang tính tích cực nhẹ**, nhưng phụ thuộc vào nhiều yếu tố đi kèm như sự chuyên cần, động lực học tập hoặc phương pháp học.

c) Chương 3

Các kiểm định T-test, Z-test và Chi-square đã cung cấp hoặc bác bỏ một số giả thuyết:

- **T-test:** Không có sự khác biệt đáng kể về điểm G3 giữa nhóm nghỉ nhiều và nghỉ ít => nghỉ học chưa chắc ảnh hưởng mạnh đến điểm cuối kỳ nhưng vẫn mang lại một số ảnh hưởng tiêu cực.
- **Z-test:** Điểm G3 của nhóm học ít không khác biệt đáng kể so với toàn bộ => học ít không luôn dẫn đến điểm thấp, nhưng tiềm ẩn bất lợi.
- **Chi-square:** Không tìm thấy mối liên hệ thống kê giữa studytime và điểm G3 (≥ 12 hay < 12), tuy nhiên quan sát thực tế vẫn cho thấy học nhiều có thể hỗ trợ cải thiện kết quả.

=> Tổng thể, các kiểm định thống kê cho thấy **thời gian học và số ngày nghỉ không phải là yếu tố duy nhất ảnh hưởng điểm số**, mà cần xét đến nhiều biến tương tác khác.

d) Chương 4

Trong chương 4, hai phương pháp cải thiện điểm số cuối kỳ môn Toán đã được đề xuất, dựa trên các phân tích dữ liệu thực nghiệm. Cả hai phương án đều dựa trên những yếu tố hành vi và học tập cụ thể của học sinh, qua đó đưa ra các giải pháp có khả năng ứng dụng thực tiễn cao trong môi trường giáo dục phổ thông.

Phương pháp 1 – Tăng thời gian học mỗi tuần

Phương pháp này đề xuất nâng thời gian học trung bình của học sinh từ mức 1–2 lên mức 2–3 giờ hoặc mức 4 mỗi tuần. Cơ sở cho đề xuất đến từ quan sát thực nghiệm:

- Tại mức học **2–3**, điểm số G3 dao động từ 5–19 điểm.
- Trong khi đó, nhóm học mức 4 có phân bố điểm cao hơn, chủ yếu dao động từ **10–20 điểm**, đồng thời xuất hiện ít học sinh có điểm thấp.

Lợi ích nổi bật:

- Nâng cao **mức sàn điểm số**: Học sinh học nhiều hơn thường không bị rơi vào nhóm điểm thấp.
- Cải thiện **sự ổn định trong học tập**: Phân bố điểm chặt chẽ hơn cho thấy kết quả học tập ít dao động hơn ở nhóm học nhiều giờ.
- Gợi ý một **ngưỡng tối thiểu hiệu quả của thời gian học**, giúp giáo viên và học sinh điều chỉnh kế hoạch học tập phù hợp hơn với năng lực và mục tiêu cá nhân.

Phương pháp 2 – Quản lý nhóm học sinh có số buổi nghỉ cao

Phương pháp này tập trung vào mối liên hệ giữa số buổi nghỉ học và kết quả điểm G3. Dữ liệu cho thấy:

- Nhóm học sinh có **số buổi nghỉ từ 1–6** đạt điểm trung bình cao hơn đáng kể so với nhóm **nghỉ trên 6 buổi**.
- Nhóm "không nghỉ học" không nhất thiết có điểm cao vượt trội, nhưng **việc nghỉ quá nhiều** (trên 6 buổi) có **tác động tiêu cực rõ rệt đến điểm số**.

Lợi ích nổi bật:

- Cung cấp **chỉ báo cảnh báo sớm**: Số buổi nghỉ học là một chỉ số đơn giản nhưng hữu ích để phát hiện sớm học sinh có nguy cơ tụt dốc học tập.
- Giúp nhà trường thiết kế các **biện pháp can thiệp quản lý** như: tư vấn, liên lạc phụ huynh, cảnh báo học vụ nhằm hỗ trợ nhóm học sinh nghỉ học nhiều.
- Tăng tính **kỷ luật học đường**: Giảm thiểu nghỉ học không lý do giúp cải thiện hiệu quả học tập tổng thể

e) Tổng kết bài

Kết quả và đóng góp nổi bật

Bài tiểu luận đã triển khai một chuỗi quy trình phân tích dữ liệu khoa học, kết hợp giữa thống kê mô tả, trực quan hóa, kiểm định giả thuyết và đề xuất cải thiện hành vi học tập. Cụ thể:

- Thông qua việc sử dụng tập dữ liệu thực tế student-mat.csv, tiểu luận đã **khám phá mối quan hệ giữa thời gian học tập, số ngày nghỉ, các yếu tố xã hội và kết quả học tập môn Toán (G3)** của học sinh trung học.
- Việc vận dụng ba thư viện trực quan hóa (Matplotlib, Seaborn, Bokeh) đã cung cấp cái nhìn **đa chiều, rõ ràng và sinh động** về xu hướng trong dữ liệu, hỗ trợ ra quyết định chính xác hơn.
- Các kiểm định thống kê (T-test, Z-test, Chi-square) đóng vai trò xác minh giả thuyết, **bổ sung tính khách quan và độ tin cậy** cho các nhận định.
- Hai phương pháp cải thiện được đề xuất dựa trên bằng chứng dữ liệu, hướng tới **cá nhân hóa giáo dục và tăng hiệu quả học tập thông qua điều chỉnh hành vi và thời lượng học**.

Tổng thể, bài tiểu luận thể hiện một quy trình phân tích dữ liệu hoàn chỉnh – từ khám phá dữ liệu đến đề xuất giải pháp – với **tính thực tiễn cao và khả năng ứng dụng trong giáo dục hiện đại**.

Hạn chế của nghiên cứu

Mặc dù đạt được một số kết quả tích cực, nghiên cứu vẫn tồn tại một số hạn chế nhất định:

- **Phạm vi dữ liệu hẹp:** Tập trung vào một môn học (Toán) và một nhóm đối tượng cụ thể tại Bồ Đào Nha khiến cho các kết luận chưa thể khái quát hóa cho các môi trường giáo dục khác.
- **Thiếu yếu tố định tính:** Dữ liệu không bao gồm các biến mềm như động lực học tập, kỹ năng tự học, phương pháp học tập hoặc tương tác trong lớp học – những yếu tố có thể ảnh hưởng đáng kể đến kết quả học tập.
- **Giới hạn về thời gian và quy mô phân tích,** chưa áp dụng các mô hình dự đoán phác tệp hoặc học máy để mở rộng khả năng ứng dụng.

Định hướng phát triển trong tương lai

Để tiếp tục phát triển hướng nghiên cứu này và gia tăng giá trị ứng dụng, một số đề xuất như sau:

- **Mở rộng phạm vi nghiên cứu:** Tiến hành phân tích thêm các môn học khác (như Ngữ văn, Vật lý...) để so sánh **yếu tố ảnh hưởng giữa các môn**, từ đó nhận diện các đặc trưng học tập chung và riêng.
- **Ứng dụng học máy (Machine Learning):** Triển khai các mô hình dự đoán điểm số (như hồi quy, cây quyết định, random forest...) để **phát hiện sớm học sinh có nguy cơ rớt môn**, giúp giáo viên và nhà trường can thiệp kịp thời.
- **Kết hợp dữ liệu định tính:** Bổ sung khảo sát học sinh, phỏng vấn giáo viên để **hiểu rõ hơn động lực, cảm nhận và hoàn cảnh cá nhân**, từ đó nâng cao chiều sâu và tính nhân văn trong phân tích.

TÀI LIỆU THAM KHẢO

1. Chapter 5 (Matplotlib, Seaborn, Bokeh).
2. Scipy.stats cho kiểm định thống kê (SciPy Docs).
3. Tài liệu từ Matplotlib, Seaborn, Bokeh, UIC Machine Learning Repository.