ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**



BÁO CÁO ĐỒ ÁN CUỐI KỲ

HỌC PHẦN: PHÂN TÍCH DỮ LIỆU

**ĐỀ TÀI: Phân tích ảnh hưởng của một số yếu tố đến kết quả học tập của sinh viên hệ CLC khóa 21 tại Đại học Sài Gòn**

THÀNH VIÊN:

Lê Minh Phúc – 3119411051-DCT119C1

Nguyễn Công Thọ – 3119411077-DCT119C1

Lê Công Minh - 3119411037-DCT119C2

GVHD: PGS.TS Nguyễn Tuấn Đăng

NĂM HỌC: 2023-2024/HK2

TP. HỒ CHÍ MINH, THÁNG 5 NĂM 2024

# MỤC LỤC

[MỤC LỤC i](#_Toc166444934)

[DANH MỤC BẢNG BIỂU iii](#_Toc166444935)

[DANH MỤC HÌNH ẢNH, SƠ ĐỒ iv](#_Toc166444936)

[DANH MỤC CHỮ VIẾT TẮT vi](#_Toc166444937)

[MỞ ĐẦU 1](#_Toc166444938)

[1. Lý do chọn đề tài 1](#_Toc166444939)

[2. Khái quát về đề tài 1](#_Toc166444940)

[3. Mục đích nghiên cứu 1](#_Toc166444941)

[4. Phạm vi nghiên cứu 2](#_Toc166444942)

[5. Phương pháp nghiên cứu 2](#_Toc166444943)

[6. Cấu trúc của đề tài 2](#_Toc166444944)

[CHƯƠNG 1: GIỚI THIỆU BỘ DỮ LIỆU 3](#_Toc166444945)

[1.1. GPA – tiêu chí đánh giá kết quả học tập 3](#_Toc166444946)

[1.2. Thu thập dữ liệu 4](#_Toc166444947)

[CHƯƠNG 2: XỬ LÝ DỮ LIỆU 6](#_Toc166444948)

[2.1. Công cụ và môi trường thực nghiệm 6](#_Toc166444949)

[2.1.1. Công cụ thực nghiệm 6](#_Toc166444950)

[2.1.2. Môi trường thực nghiệm 6](#_Toc166444951)

[2.2. Tiền xử lý dữ liệu 7](#_Toc166444952)

[2.2.1. Chỉnh sửa kiểu dữ liệu của bộ dữ liệu 7](#_Toc166444953)

[2.2.2. Xử lý dữ liệu khuyết thiếu (Handling missing data) 9](#_Toc166444954)

[2.2.3. Mã hoá dữ liệu cá nhân nhạy cảm 10](#_Toc166444955)

[2.2.4. Tính điểm trung bình GPA và đánh giá xếp loại sinh viên 11](#_Toc166444956)

[2.2.5. Sắp xếp một số trường thông tin số liệu 14](#_Toc166444957)

[2.3. Phân tích một số yếu tố ảnh hưởng đến kết quả học tập của sinh viên 18](#_Toc166444958)

[2.3.1. Phân tích sự ảnh hưởng của khoảng cách từ nhà đến trường đến thành tích học tập 18](#_Toc166444959)

[2.3.2. Phân tích sự ảnh hưởng của thu nhập do làm thêm đến thành tích học tập 23](#_Toc166444960)

[2.3.3. Phân tích sự ảnh hưởng của số lượng số người đồng hành (anh/chị/em trong gia đình hoặc bạn bè) đến thành tích học tập 29](#_Toc166444961)

[CHƯƠNG 3: NHẬN XÉT VÀ ĐÁNH GIÁ 38](#_Toc166444962)

[KẾT LUẬN 39](#_Toc166444963)

[TÀI LIỆU THAM KHẢO 40](#_Toc166444964)

[PHỤ LỤC 41](#_Toc166444965)

[Bộ mã nguồn chương trình 41](#_Toc166444966)

[Bảng phân công công việc 50](#_Toc166444967)

# DANH MỤC BẢNG BIỂU

[Bảng 1. Bảng số liệu ví dụ điểm số của một số môn học dùng để tính GPA của một sinh viên 3](#_Toc166444969)

[Bảng 2. Bảng số liệu quy đổi điểm số GPA thang điểm 10 sang thang điểm 4 theo quy định của Đại học Sài Gòn [1] 4](#_Toc166444970)

[Bảng 3. Bảng số liệu thành tích học tập của sinh viên 13](#_Toc166444971)

[Bảng 4. Dữ liệu thống kê số sinh viên theo nhóm khoảng cách từ nhà đến trường và thành tích học tập 22](#_Toc166444972)

[Bảng 5. Dữ liệu thống kê số sinh viên theo nhóm thu nhập và thành tích học tập 26](#_Toc166444973)

[Bảng 6. Dữ liệu thống kê số sinh viên theo thành tích học tập và nhóm bạn 32](#_Toc166444974)

[Bảng 7. Dữ liệu thống kê số sinh viên theo thành tích học tập và nhóm số anh chị em 36](#_Toc166444975)

# DANH MỤC HÌNH ẢNH, SƠ ĐỒ

[Hình 1. Hình ảnh ví dụ của một mẫu dữ liệu của bộ dữ liệu 7](#_Toc166445022)

[Hình 2. Dữ liệu được đọc bởi pyspark 8](#_Toc166445023)

[Hình 3. Kiểu dữ liệu ban đầu của các Schema trong bộ dữ liệu 9](#_Toc166445024)

[Hình 4. Hình ảnh các schema được chỉnh sửa kiểu dữ liệu 9](#_Toc166445025)

[Hình 5. Dữ liệu trước khi mã hóa 11](#_Toc166445026)

[Hình 6. Dữ liệu sau khi được mã hóa 11](#_Toc166445027)

[Hình 7. Điểm số GPA và xếp loại theo mức thành tích của một số sinh viên 12](#_Toc166445028)

[Hình 8. Hình ảnh minh hoạ số lượng sinh viên được xếp theo các mức thành tích 13](#_Toc166445029)

[Hình 9. Kết quả của mẫu dữ liệu sau khi sắp xếp điểm số GPA theo thứ tự giảm dần 14](#_Toc166445030)

[Hình 10. Kết quả của mẫu dữ liệu sau khi sắp xếp số lượng bạn bè của sinh theo thứ tự tăng dần cùng kết quả học tập của sinh viên 15](#_Toc166445031)

[Hình 11. Kết quả của mẫu dữ liệu sau khi sắp xếp số lượng anh/chị/em trong gia đình của sinh viên theo thứ tự tăng dần cùng kết quả học tập của sinh viên 17](#_Toc166445032)

[Hình 12. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và khoảng cách từ nhà đến trường 19](#_Toc166445033)

[Hình 13. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm khoảng cách từ nhà đến trường 21](#_Toc166445034)

[Hình 14. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và thu nhập từ việc làm thêm 24](#_Toc166445035)

[Hình 15. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm thu nhập (với M là million dong nghĩa là triệu đồng) 26](#_Toc166445036)

[Hình 16. Biểu đồ hộp mô tả việc thống kê số lượng sinh viên theo từng nhóm thu nhập 28](#_Toc166445037)

[Hình 17. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và số lượng bạn bè 30](#_Toc166445038)

[Hình 18. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm bạn 32](#_Toc166445039)

[Hình 19. Biểu đồ scatter mô tả mối quan hệ giữa số lượng anh/chị/em trong gia đình và thành tích học tập của sinh viên 34](#_Toc166445040)

[Hình 20. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm số lượng anh/chị/em 35](#_Toc166445041)

# DANH MỤC CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| CLC | : | Chất lượng cao |
| GPA | : | Grade Point Average |
| csv | : | Comma separated value |
| AES | : | Advanced Encryption Standard |

# MỞ ĐẦU

## Lý do chọn đề tài

Giáo dục đại học là một trong những vấn đề cấp thiết trong việc đào tạo nguồn nhân lực chất lượng cao nhằm phục vụ cho việc công nghiệp hoá hiện đại hoá đất nước trong thời đại Cách mạng công nghiệp 4.0 hiện nay. Trong đó, chương trình đào tạo hệ Chất lượng cao (CLC) là một trong những vấn đề cần được quan tâm vì các sinh viên hệ CLC nhận được những chế độ chính sách đặc biệt về phương pháp học tập, môi trường đào tạo và mục tiêu đào tạo. Vì thế, nhóm chúng em chọn đề tài này với lý do nhằm hiểu rõ những yếu tố ảnh hưởng đến kết quả học tập của sinh viên hệ CLC. Qua đó, có thể đưa ra các biện pháp hỗ trợ kịp thời và phù hợp giúp cho sinh viên đạt được kết quả tốt nhất.

## Khái quát về đề tài

Đề tài nghiên cứu tập trung vào kết quả học tập vào năm nhất và năm hai của một số sinh viên thuộc khóa 21 ngành Công nghệ thông tin hệ đào tạo Chất lượng cao tại trường Đại học Sài Gòn cùng một số yếu tố ảnh hưởng như sau:

* Quãng đường từ nhà đến trường
* Thu nhập do việc làm thêm
* Số lượng bạn bè
* Số lượng anh chị em trong nhà

## Mục đích nghiên cứu

Mục đích nghiên cứu của đề tài gồm:

* Xác định các yếu tố ảnh hưởng đến kết quả học tập của sinh viên hệ CLC khóa 21 tại Đại học Sài Gòn.
* Đánh giá mức độ ảnh hưởng của từng yếu tố.
* Đề xuất các biện pháp hỗ trợ phù hợp để giúp sinh viên hệ CLC nâng cao kết quả học tập.

## Phạm vi nghiên cứu

Đề tài nghiên cứu được thực hiện trên đối tượng là các sinh viên khoá 21 ngành Công nghệ thông tin thuộc hệ đào tạo CLC đang theo học tại Đại học Sài Gòn và một số yếu tố ảnh hưởng đến kết quả học tập được đề cập tại mục 2

## Phương pháp nghiên cứu

Một số phương pháp nghiên cứu được áp dụng trong đề tài:

* Phương pháp khảo sát và phỏng vấn được áp dụng trong bước thu thập và xây dựng bộ dữ liệu
* Phương pháp thống kê được áp dụng trong bước phân tích dữ liệu nhằm đưa ra kết quả phân tích cùng nhận xét đánh giá mức độ ảnh hưởng đến kết quả học tập của các yếu tố được nêu.

## Cấu trúc của đề tài

Báo cáo của đề tài gồm 5 phần:

* Phần mở đầu
* Chương 1: Giới thiệu bộ dữ liệu
* Chương 2: Xử lý dữ liệu
* Chương 3: Nhận xét đánh giá
* Phần kết luận

# CHƯƠNG 1: GIỚI THIỆU BỘ DỮ LIỆU

## GPA – tiêu chí đánh giá kết quả học tập

* GPA (Grade Point Average – điểm trung bình các môn học) là chỉ số được các trường đại học sử dụng một cách phổ biến trong việc đánh giá kết quả học tập của sinh viên trong suốt quá trình học tập tại trường. Chỉ số GPA có thể được đánh giá theo toàn khóa học, một học kỳ hay một năm học.
* Chỉ số GPA được xác định bằng cách tính trung bình cộng tích số điểm của các môn đã học nhân với hệ số tín chỉ của từng môn học, sau đó chia cho tổng số tín chỉ của các môn đã học (theo công thức (1))

(1)

Trong đó:

p: số điểm của môn học

t: số tín chỉ của môn học

* Để có thể hiểu rõ hơn về chỉ số GPA thì ta có ví dụ từ bảng số liệu từ bảng 1:

Bảng 1. Bảng số liệu ví dụ điểm số của một số môn học dùng để tính GPA của một sinh viên

|  |  |  |
| --- | --- | --- |
| Môn học | Số điểm | Số tín chỉ |
| A | 9 | 3 |
| B | 8 | 2 |
| C | 7 | 3 |
| D | 10 | 4 |
| E | 9 | 3 |

Với số liệu từ Bảng 1, ta có thể tính chỉ số GPA như sau:

* Mức điểm GPA thường được sử dụng ở thang điểm 4 (0.0 đến 4.0) nhằm để đánh giá thành tích học tập cụ thể:
  + 0.0: Mức kém (F)
  + 1.0: Mức yếu (D)
  + 2.0: Mức trung bình (C)
  + 3.0: Mức tốt (B)
  + 4.0: Mức xuất sắc (A)
* Theo quy định của [1]Đại học Sài Gòn thì kết quả học tập được tính theo 2 loại GPA là GPA thang điểm 10 và GPA thang điểm 4 (theo thang quy đổi bảng 2)

Bảng 2. Bảng số liệu quy đổi điểm số GPA thang điểm 10 sang thang điểm 4 theo quy định của Đại học Sài Gòn [1]

|  |  |  |
| --- | --- | --- |
| Thang điểm 10 | Thang điểm 4 | Mức điểm |
| < 4.0 | 0.0 | F |
| >=4,0 và <5.5 | 1.0 | D |
| >=5.5 và <7.0 | 2.0 | C |
| >=7.0 và <8.5 | 3.0 | B |
| >=8.5 | 4.0 | A |

## Thu thập dữ liệu

* Trong đề tài này, dữ liệu được thu thập trong phạm vi khóa 21 ngành Công nghệ thông tin hệ đào tạo chất lượng cao (CLC) đang học tập tại trường Đại học Sài Gòn. Có thể thu hẹp phạm vi nghiên cứu bằng cách thu thập dữ liệu của các sinh viên trong một lớp học nằm trong phạm vi nghiên cứu cụ thể là lớp DCT121C3
* Nhóm chúng em đã thu thập một số thông tin liên quan đến kết quả học tập trong khoảng thời gian năm nhất và năm hai của các sinh viên thuộc lớp DCT121C3 cùng với một số thông tin về các yếu tố ảnh hưởng đến kết quả học tập của họ, cụ thể:
  + Các thông tin cá nhân của sinh viên (gồm mã sinh viên, tên và lớp): Các thông tin này là các thông tin cá nhân nhạy cảm có thể từ chối cung cấp
  + Kết quả học tập trong năm nhất và năm hai của sinh viên: được tính theo GPA (thang điểm 10) tích lũy tương ứng năm nhất và năm 2
  + Một số thông tin liên quan đến các yếu tố ảnh hưởng đến kết quả học tập của sinh viên
    - Khoảng cách từ nhà đến trường (tính bằng km)
    - Thu nhập do làm thêm của sinh viên (tính theo triệu đồng)
    - Số anh chị em của sinh viên
    - Số bạn bè thân thiết của sinh viên
* Hình thức thu thập dữ liệu: Dữ liệu được thu thập dưới dạng bảng câu hỏi và thu thập bằng công cụ Google Form. Đồng thời kết quả thu thập được lưu trong một bảng tính Google Spreadsheet và có thể tải về dưới dạng file csv (comma separated value)

# CHƯƠNG 2: XỬ LÝ DỮ LIỆU

## Công cụ và môi trường thực nghiệm

### Công cụ thực nghiệm

* Công cụ lưu trữ dữ liệu: Bộ dữ liệu được lưu trong một tập tin csv (các giá trị trong tập tin được phân tách bằng dấu phẩy) vì tập tin csv có cấu trúc đơn giản dễ sử dụng được hỗ trợ bởi các công cụ phổ biến như pandas hay pyspark. Tuy nhiên, đối với dữ liệu phức tạp thì tập tin csv không được dùng để lưu trữ vì tập tin csv không có định dạng cụ thể nào. Vì vậy, với quy mô của một đồ án môn học, dữ liệu thường có cấu trúc tương đối đơn giản và có ít dữ liệu do đó tập tin csv là công cụ phù hợp để lưu trữ bộ dữ liệu này.
* Công cụ phân tích dữ liệu: pyspark là một interface của Apache Spark là một thư viện xử lý dữ liệu lớn mã nguồn mở mạnh mẽ. Nó cho phép bạn sử dụng ngôn ngữ lập trình Python để viết các chương trình Spark để xử lý và phân tích dữ liệu lớn trên cụm máy tính. Trong đề tài này, nhóm chúng em tận dụng một số tiện ích của pyspark như Spark DataFrame (trong việc xử lý dữ liệu dạng bảng) và Spark SQL (trong việc xử lý dữ liệu có cấu trúc) cùng với tính linh hoạt của nó trong việc hỗ trợ nhiều loại tập tin trong đó có dạng csv.
* Công cụ trực quan hoá dữ liệu: Thư viện Matplotlib của python là một công cụ trực quan hoá mã nguồn mổ tiêu chuẩn và phổ biến vì được sự hỗ trợ mạnh mẽ từ cộng đồng. Bên cạnh đó, Matplotlib còn có khả năng tùy chỉnh các thuộc tính của biểu đồ một cách linh hoạt.

### Môi trường thực nghiệm

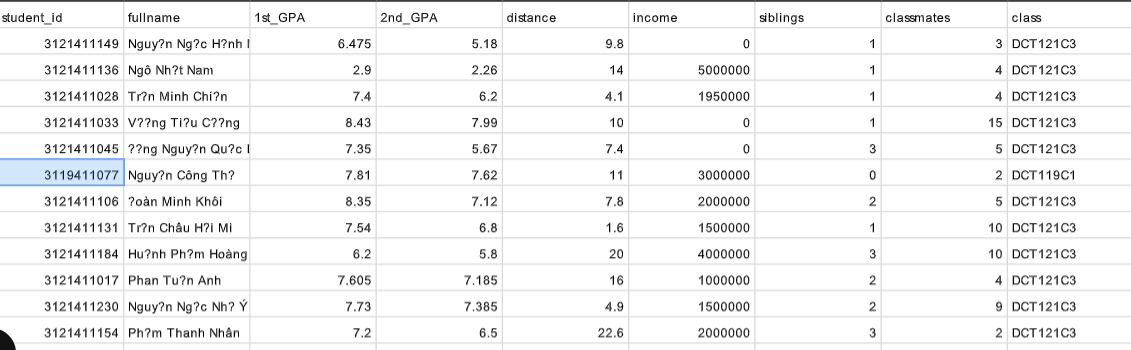
* Nhóm chúng em sử dụng Google Colab để tiến hành thực hiện đề tài này cùng với việc lưu trữ bộ mã nguồn và dữ liệu trên Google Drive vì có thể chia sẻ cho đội nhóm và dễ dàng quản lý tiến độ của các thành viên thông qua lịch sử hoạt động của các thành viên được ghi nhận trên thư mục của dự án
* Một tiện ích khác khi sử dụng Google Colab là việc môi trường lập trình được hệ thống máy chủ của Google cài đặt sẵn hầu hết mọi thư viện hỗ trợ cơ bản vả có giao diện như một Jupyter notebook giúp cho việc lập trình và quan sát kết quả dễ dàng hơn trên một ô (coding cell) cũng như khả năng chú thích các tài liệu quan trọng trước khi lập trình thông qua các markdown.
* Tuy nhiên, không phải thư viện nào của python cũng được Google cài đặt sẵn trong đó có thư viện pyspark vì vậy trước khi thực hiện chạy các ô code phía sau thì phải cài đặt pyspark bằng cú pháp

|  |
| --- |
| %pip install pyspark |

## Tiền xử lý dữ liệu

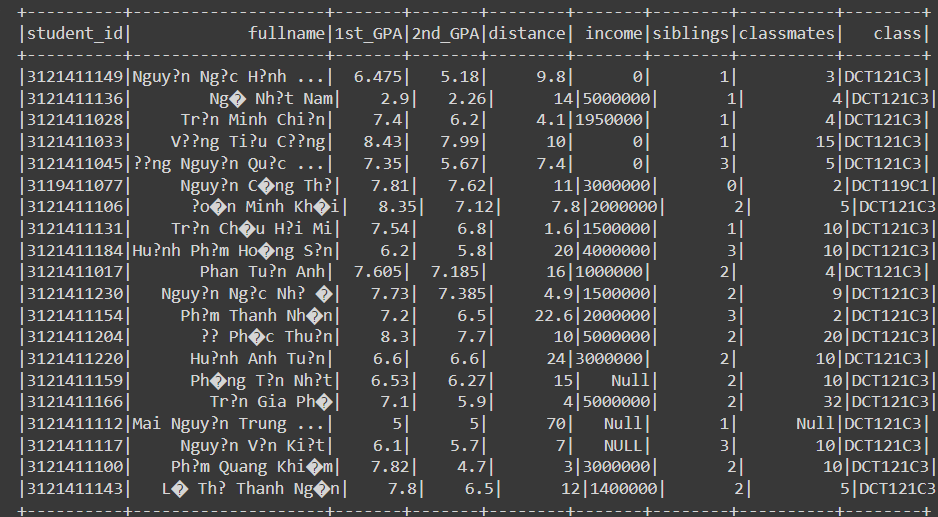
### Chỉnh sửa kiểu dữ liệu của bộ dữ liệu

* Như đã giới thiệu trong chương 1, bộ dữ liệu đã được thu thập là một tập tin csv gồm 9 trường dữ liệu (data fields) tượng trưng cho 9 cột và 31 thể hiện (data instance) tượng trưng cho 31 dòng dữ liệu



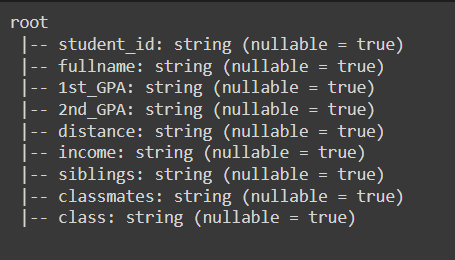
Hình 1. Hình ảnh ví dụ của một mẫu dữ liệu của bộ dữ liệu

* Theo hình ảnh minh hoạ từ Hình 1, bộ dữ liệu gồm các trường dữ liệu sau:
  + Student\_id: mã số sinh viên của sinh viên
  + Fullname: tên của sinh viên
  + 1st\_GPA: chỉ số GPA của sinh viên vào năm thứ nhất (tính theo thang điểm 10)
  + 2nd\_GPA: chỉ số GPA của sinh viên vào năm thứ hai (tính theo thang điểm 10)
  + Distance: khoảng cách từ nhà đến trường của mỗi sinh viên (tính theo km)
  + Income: thu nhập của sinh viên đến từ công việc làm thêm (tính theo VNĐ)
  + Siblings: Số anh chị em của sinh viên
  + Classmates: Số lượng bạn thân cùng lớp
  + Class: là lớp của sinh viên tham gia khảo sát



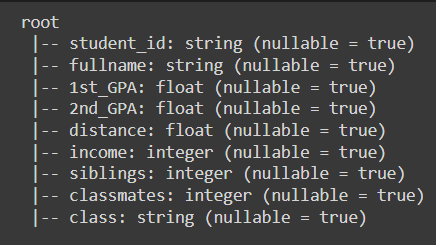
Hình 2. Dữ liệu được đọc bởi pyspark

* Một vấn đề gặp phải trong bộ dữ liệu (được thể hiện quan Hình 1 và Hình 2) là việc hiển thị tiếng Việt bị lỗi font chữ do lỗi mã hóa Unicode (UTF-8). Vì thế trường fullname có thể bị lược bỏ và các sinh viên có thể được quản lý nhờ vào mã số sinh viên và mã lớp của sinh viên.
* Tiếp theo là kiểu dữ liệu của các cột dữ liệu (schema) được ghi nhận nhờ phương thức printSchema() của pyspark



Hình 3. Kiểu dữ liệu ban đầu của các Schema trong bộ dữ liệu

* Theo hình 3, các schema được mặc định kiểu dữ liệu là kiểu chuỗi (string) mặc dù trong bảng dữ liệu (dataframe) còn có các cột mang kiểu dữ liệu là số nguyên hoặc số thực. Vì thế việc xử lý đổi kiểu dữ liệu là việc cần thiết nhằm phục vụ cho việc phân tích dữ liệu. Việc xử lý đổi kiểu dữ liệu được thể hiện qua Hình 4.



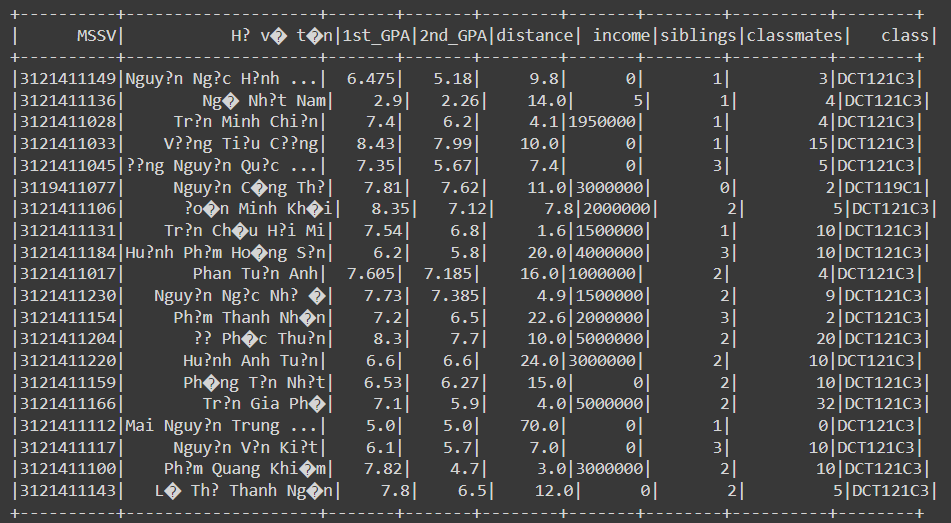
Hình 4. Hình ảnh các schema được chỉnh sửa kiểu dữ liệu

### Xử lý dữ liệu khuyết thiếu (Handling missing data)

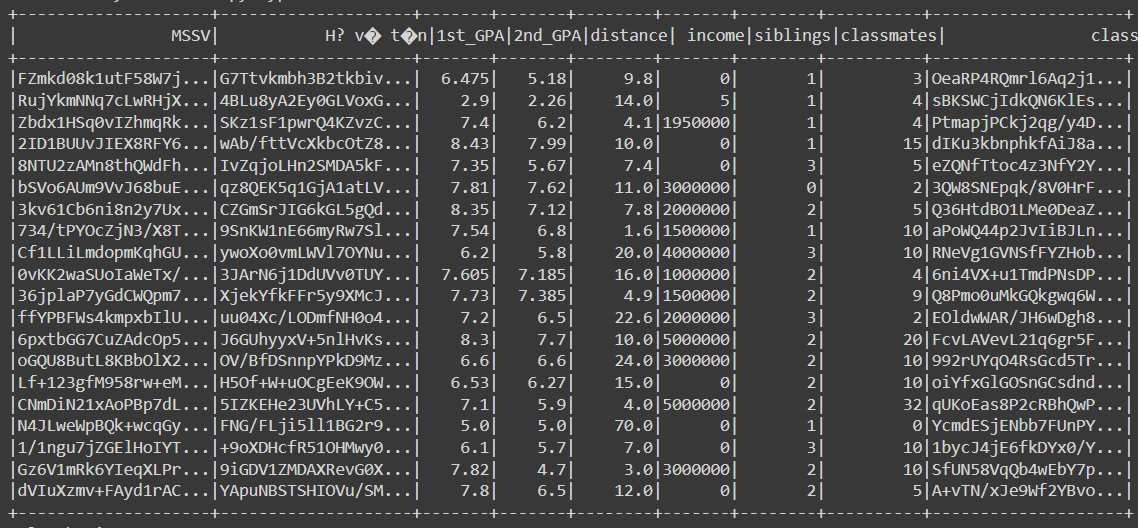
* Dữ liệu khuyết thiếu là một vấn đề phổ biến trong việc phân tích dữ liệu. Dữ liệu có thể bị khuyết thiếu vì một số nguyên nhân:
  + Lỗi thu thập dữ liệu
  + Lỗi nhập liệu
  + Người khảo sát từ chối cung cấp dữ liệu trong quá trình thu thập dữ liệu (có thể do dữ liệu cá nhân nhạy cảm nên từ chối cung cấp)
* Việc xử lý dữ liệu khuyết thiếu là rất quan trọng vì nó có thể ảnh hưởng đến độ chính xác và hiệu quả của các phân tích
* Có nhiều phương pháp khác nhau để xử lý dữ liệu khuyết thiếu. Phương pháp phù hợp nhất sẽ phụ thuộc vào loại dữ liệu khuyết thiếu, nguyên nhân gây ra dữ liệu khuyết thiếu và mục đích sử dụng dữ liệu. Một số phương pháp phổ biến bao gồm:
  + Xóa bỏ dòng dữ liệu khuyết thiếu
  + Gán giá trị
  + Dự đoán giá trị bằng máy học

### Mã hoá dữ liệu cá nhân nhạy cảm

* Mã hoá dữ liệu cá nhân nhạy cảm là quá trình chuyển đổi thông tin cá nhân có giá trị cao thành dạng không đọc được hoặc không thể hiểu được mà chỉ có những người được ủy quyền mới có khả năng giải mã. Mục đích chính của việc mã hoá dữ liệu cá nhân nhạy cảm là bảo vệ thông tin riêng tư của người dùng và ngăn chặn truy cập trái phép vào thông tin này.
* Chúng tôi sử dụng thuật toán [2]AES (Advanced Encryption Standard) là một tiêu chuẩn mã hóa được sử dụng rộng rãi để bảo mật dữ liệu. AES cung cấp bảo mật mạnh mẽ và được hỗ trợ bởi nhiều thư viện và công cụ mã hóa khác nhau.



Hình 5. Dữ liệu trước khi mã hóa



Hình 6. Dữ liệu sau khi được mã hóa

### Tính điểm trung bình GPA và đánh giá xếp loại sinh viên

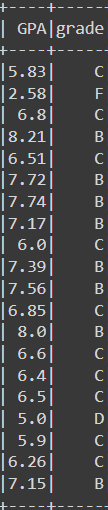
* Về việc đánh giá GPA của các sinh viên qua GPA của năm nhất và GPA của năm thứ hai thì nhóm chúng em sử dụng phương pháp tính giá trị trung bình của GPA trong hai năm học thông qua công thức (2)

Trong đó:

* GPA là giá trị GPA trung bình trong 2 năm học của sinh viên
* A là điểm GPA của sinh viên trong năm nhất
* B là GPA của sinh viên trong năm thứ hai

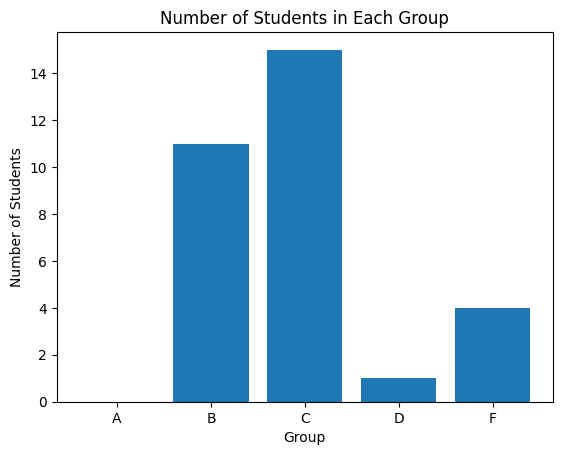
Lưu ý: Kết quả của việc tính điểm GPA được làm tròn 2 chữ số thập phân

* Khi đã tính được GPA của từng sinh viên, sau đó nhóm chúng em đã tiến hành đánh giá dựa trên mức thành tích (dựa theo thông tin của Bảng 2) thì nhóm chúng em có được thành tích của một số sinh viên tương ứng (theo hình 7)



Hình 7. Điểm số GPA và xếp loại theo mức thành tích của một số sinh viên

* Từ một phần bảng số liệu trích ra từ bộ dữ liệu (như hình 7), nhóm chúng em đã thống kê bao nhiêu sinh viên đạt hạng A, B, C, D và F tương ứng và đã trực quan hoá dưới dạng biểu đồ cột (như hình 8)



Hình 8. Hình ảnh minh hoạ số lượng sinh viên được xếp theo các mức thành tích

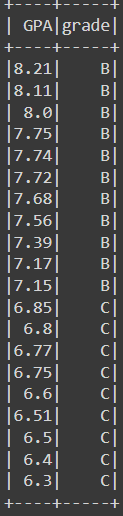
* Với biểu đồ cột như hình 6, nhóm chúng em đã thống kê được các nhóm sinh viên được chia theo thành tích học tập như bảng 3:

Bảng 3. Bảng số liệu thành tích học tập của sinh viên

|  |  |
| --- | --- |
| Thành tích học tập | Số lượng sinh viên |
| A | 0 |
| B | 11 |
| C | 15 |
| D | 1 |
| F | 4 |
| Tổng cộng | 31 |

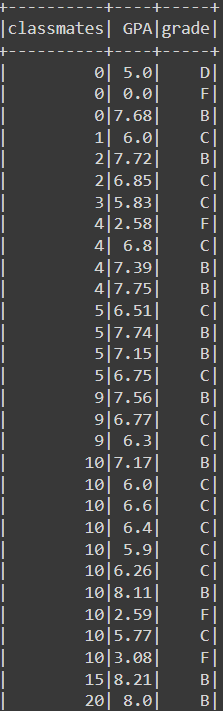
### Sắp xếp một số trường thông tin số liệu

* Để có thể dễ dàng quan sát thứ tự của các thông tin quan trọng thì việc sắp xếp lại các trường thông tin ấy là việc quan trọng nhằm giúp người phân tích có thể quan sát các quy luật của dữ liệu dễ dàng hơn so với việc quan sát và phân tích dữ liệu thô ban đầu
* Trong đề tài này, nhóm chúng em đã sắp xếp một số thông tin quan trọng như điểm số GPA, xếp loại thành tích, số lượng anh/chị/em, bạn bè của sinh viên,…



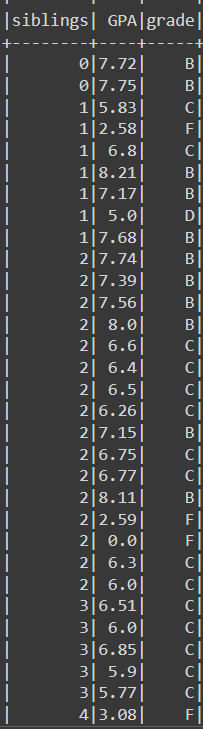
Hình 9. Kết quả của mẫu dữ liệu sau khi sắp xếp điểm số GPA theo thứ tự giảm dần

* Với hình 9, nhóm chúng em sắp xếp danh sách các sinh viên theo điểm số GPA và xếp loại thành tích của sinh viên. Qua đó có thể quan sát được một cách dễ dàng dữ liệu của sinh viên theo các nhóm xếp loại thành tích



Hình 10. Kết quả của mẫu dữ liệu sau khi sắp xếp số lượng bạn bè của sinh theo thứ tự tăng dần cùng kết quả học tập của sinh viên

* Trong khi đó, với hình 10, nhóm chúng tiến hành sắp xếp dữ liệu bạn bè của sinh viên đồng thời hiển thị cùng điểm số cùng xếp loại thành tích học tập của sinh viên. Qua đó, nhóm chúng em đã quan sát được sơ lược mối quan hệ giữa số lượng bạn bè và kết quả học tập của từng sinh viên.

**

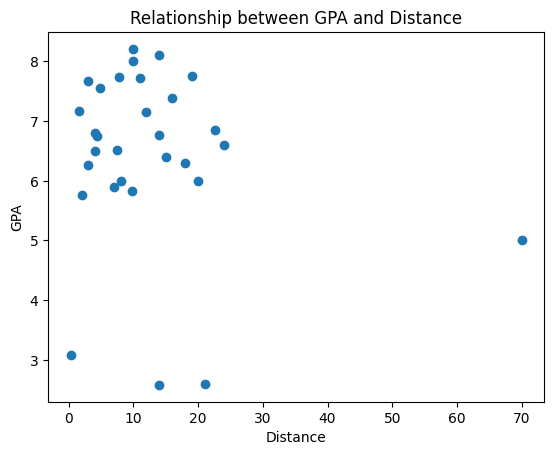
Hình 11. Kết quả của mẫu dữ liệu sau khi sắp xếp số lượng anh/chị/em trong gia đình của sinh viên theo thứ tự tăng dần cùng kết quả học tập của sinh viên

* Đồng thời, hình 11 mô tả việc nhóm chúng em đã tiến hành sắp xếp số lượng anh/chị/em trong gia đình cùng với thành tích học tập của sinh viên. Qua đó, nhóm chúng em có thể quan sát được sự ảnh hưởng của số lượng anh/chị/em trong gia đình đối với việc học tập của sinh viên.

## Phân tích một số yếu tố ảnh hưởng đến kết quả học tập của sinh viên

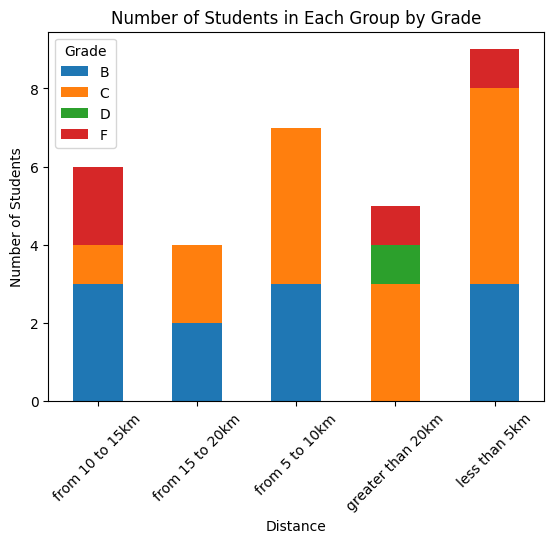
### Phân tích sự ảnh hưởng của khoảng cách từ nhà đến trường đến thành tích học tập

* Mối liên hệ giữa khoảng cách từ nhà đến trường và thành tích học tập của sinh viên là một mối quan hệ phức tạp với nhiều yếu tố cần suy xét. Trong đó có những giả thuyết về những tác động khác nhau liên quan đến quãng đường đến trường ảnh hưởng đến kết quả học tập của sinh viên:
  + Giả thuyết liên quan đến những tác động tiêu cực đến kết quả học tập của sinh viên: Khoảng cách từ nhà đến trường ảnh hưởng tiêu cực đến thành tích học tập của sinh viên. Quãng đường đi học của sinh viên càng xa thì thể lực và tinh thần của sinh viên sẽ suy giảm ảnh hưởng đến sự tập trung của chính sinh viên trong việc học tập và tham gia các hoạt động ngoại khoá
  + Giả thuyết liên quan đến những tác động tích cực đến kết quả học tập của sinh viên: Khoảng cách từ nhà đến trường tuy có tác động tiêu cực đến thành tích học tập của sinh viên nhưng không phải là tất cả. Một số sinh viên ở xa trường có thể rèn luyện cho mình một sự kỷ luật mạnh mẽ cùng với kỹ năng quản lý thời gian một cách chặt chẽ và khả năng tự lập của sinh viên. Đồng thời, việc ở xa trường sẽ cho sinh viên có động lực để thành công hơn vì bản thân họ đã nhận thức được rằng họ cần sự nỗ lực nhiều hơn để đạt được mục tiêu của mình.
* Thông qua bộ dữ liệu để thực hiện đề tài, nhóm chúng em sẽ phân tích thực trạng về sự ảnh hưởng của quãng đường từ nhà đến trường đến kết quả học tập của sinh viên bằng những số liệu đã được thu thập
* Đầu tiên thì chúng ta xét đến mối quan hệ giữa thành tích học tập và khoảng cách từ nhà đến trường thông qua biểu đồ scatter được thể hiện ở hình 12



Hình 12. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và khoảng cách từ nhà đến trường

* Theo hình 12 thì ta có xu hướng chủ yếu của sinh viên được khảo sát có khoảng cách từ nhà đến trường là khoảng 20km. Tuy nhiên vẫn có ngoại lệ là một sinh viên ở cách trường đến 70km
* Xét về mặt thành tích học tập của các sinh viên được khảo sát theo thông tin mà hình 7 cung cấp cho thấy thành tích của sinh viên thường tập trung ở mức 5,5 đến hơn mức 8 một chút. Đồng thời cũng phát hiện 4 sinh viên có thành tích học tập dưới mức 5
* Khi nhận xét kết hợp giữa thành tích học tập và khoảng cách từ nhà đến trường thì ta có thể nhận thấy các sinh viên ở nhà tương đối gần (mức dưới 15km) thì thành tích học tập tương đối cao thậm chí có sinh viên có mức điểm trên 8 và các sinh viên có khoảng cách nhà xa hơn thì có khoảng cách kém hơn (có thể ví dụ một sinh viên phải đi 70km để đi học thì có thành tích ở mức dưới 5). Tuy nhiên, chúng ta vẫn có một ngoại lệ khác là ba bạn sinh viên có thành tích dưới mức 5 nhưng khoảng cách từ nhà đến trường chỉ khoảng 20km.
* Để có thể hiểu rõ hơn mối quan hệ giữa thành tích học tập thì nhóm chúng em đã thống kê chia nhóm lại lại cho các sinh viên theo cách như sau:
  + Đầu tiên, nhóm chúng em sẽ chia khoảng cách từ nhà đến trường thành nhiều nhóm có quãng đường theo cột mốc cụ thể:
    - Nhóm nhà gần (dưới 5km)
    - Nhóm nhà tương đối gần (từ 5 đến 10km)
    - Nhóm nhà trung bình (từ 10 đến 15km)
    - Nhóm nhà tương đối xa (từ 15 đến 20km)
    - Nhóm nhà xa (nhiều hơn 20km)
  + Đồng thời, về thành tích học tập vẫn chia thành các mốc A, B, C, D và F dựa theo thông tin từ bảng 2 và bảng 3
  + Khi đã có dữ liệu thống kê thì nhóm chúng em sử dụng biểu đồ cột chồng (stacked bar plot) để vẽ biểu đồ thống kê có bao nhiêu sinh viên theo từng nhóm khoảng cách từ nhà đến trường và đếm cụ thể có bao nhiêu sinh viên A, B, C, D và F trong từng nhóm (được mô tả trong hình 13)



Hình 13. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm khoảng cách từ nhà đến trường

* Theo hình 13, biểu đồ hiển thị số lượng sinh viên được xếp theo các nhóm khoảng cách từ nhà đến trường. Đồng thời cũng đếm số sinh viên trong mỗi nhóm theo thành tích học tập (dữ liệu đã được thống kê theo bảng 4)

Bảng 4. Dữ liệu thống kê số sinh viên theo nhóm khoảng cách từ nhà đến trường và thành tích học tập

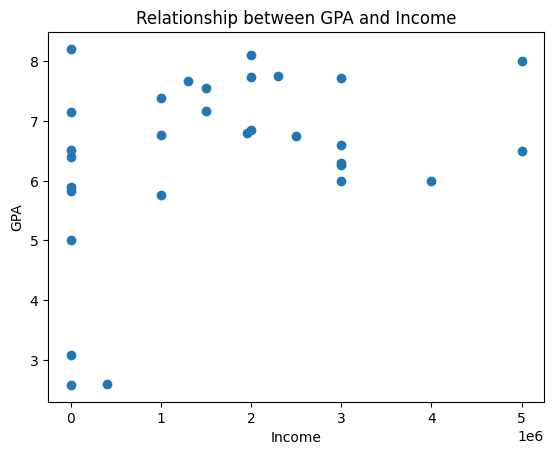
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Khoảng cách  Số sinh viên | Dưới 5km | Từ 5 đến 10km | Từ 10 đến 15km | Từ 15 đến 20km | Xa hơn 20km | Tổng cộng |
| Mức điểm A | 0 | 0 | 0 | 0 | 0 | 0 |
| Mức điểm B | 3 | 3 | 3 | 2 | 0 | 11 |
| Mức điểm C | 5 | 4 | 1 | 2 | 3 | 15 |
| Mức điểm D | 0 | 0 | 0 | 0 | 1 | 1 |
| Mức điểm F | 1 | 0 | 2 | 0 | 1 | 4 |
| Tổng cộng | 8 | 7 | 6 | 4 | 5 | 31 |

* Theo dữ liệu từ bảng 4, ta có 8 sinh viên đi học với khoảng cách dưới 5km, 7 sinh viên với quãng đường từ 5 đến 10km, 6 sinh viên có khoảng cách từ 15 đến 20km, 4 sinh viên có quãng đường đi học từ 15 đến 20km và 5 sinh viên có khoảng cách từ nhà đến trường xa hơn 20km.
* Về mặt thành tích học tập thì mức điểm cao nhất mà các sinh viên có thể đạt được là B (không có một sinh viên nào đạt mức điểm A). Đồng thời, ta có thể nhìn thấy được mức điểm B đều có mặt ở 4 nhóm quãng đường đi học gần hơn 20km mà không có một sinh viên nào đạt mức điểm ở nhóm nhà xa trường hơn 20km.
* Mặt khác, ta thấy ở mức điểm D thì không có sinh viên nào thuộc 4 nhóm có khoảng cách nhà gần trong khoảng 20km và có 1 thuộc nhóm nhà xa hơn 20km. Điều này có thể nói lên được rằng giả thuyết nhà càng ở xa thì thể lực và khả năng tập trung cho việc học sẽ kém hơn (có thể chứng minh thêm bằng một dẫn chứng khác là các nhóm nhà có khoảng cách trong khoảng 20km thì thành tích cao nhất là B còn nhóm nhà xa hơn 20km thì thành tích cao nhất là C)
* Khi xét đến nhóm mức điểm F thì có 4 sinh viên có mặt tại 3 nhóm:
  + Nhóm gần hơn 5km
  + Nhóm từ 10 đến 15km
  + Nhóm xa hơn 20km
* Như vậy thì chỉ có bạn sinh viên có thành tích học tập là F mà nhà xa hơn 20km thì có thể liên quan đến việc quãng đường đi học xa ảnh hưởng đến kết quả học tập do dành nhiều thời gian để đến trường, suy giảm thể lực và tinh thần khiến cho sự tập trung bị ảnh hưởng. Mặt khác đối với các bạn sinh viên đạt điểm F còn lại họ có quãng đường từ nhà đến trường ngắn hơn 15km thì khoảng cách từ nhà đến trường có thể không ảnh hưởng nhiều đến họ. Việc họ đạt được điểm F có thể do nhiều nguyên nhân khác trong đó có vấn đề ý thức kỷ luật của bản thân sinh viên.

### Phân tích sự ảnh hưởng của thu nhập do làm thêm đến thành tích học tập

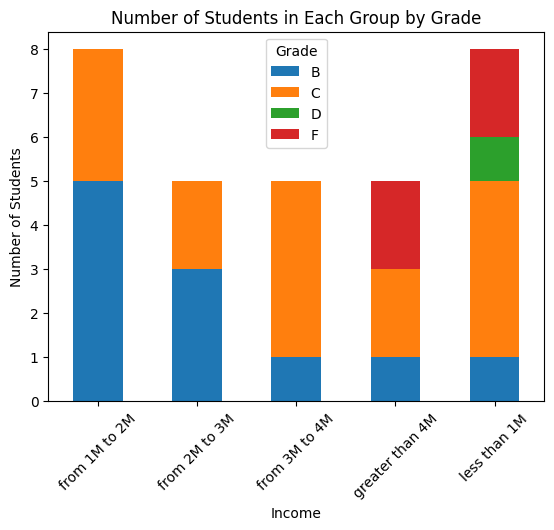
* Sinh viên khi đã bước vào môi trường mới tại các cơ sở giáo dục Đại học hoặc Cao đẳng thì chi phí duy trì cho việc học tập và sinh hoạt phí của các sinh viên sẽ là rất lớn đặc biệt là các bạn sinh viên ngoại tỉnh. Bởi vì không phải gia đình nào cũng đủ điều kiện vật chất để chu cấp cho con em mình đang học tại các trường Đại học, Cao đẳng nên việc đi làm thêm là một nhu cầu tất yếu giúp cho các bạn sinh viên có thể duy trì chi phí sinh hoạt của bản thân và hỗ trợ gia đình trong khoản được cho là “gánh nặng” này.
* Tuy nhiên, một số gia đình có điều kiện vật chất khá giả hơn thì khoản chi phí đầu tư cho con em học Đại học không quá lớn nhưng bản thân các em lại có mong muốn tự lập dần ít phụ thuộc vào gia đình hay tìm kiếm một trải nghiệm mới ngoài việc học trên giảng đường thì họ cũng có thể chọn làm thêm như một cách hiệu quả có thể thỏa mãn nhu cầu của họ.
* Nguyên nhân của việc sinh viên đi làm thêm thì có rất nhiều nhưng sự ảnh hưởng của nó lên kết quả học tập của sinh viên sẽ được nhóm chúng em phân tích thông qua các thông tin liên quan được thu thập trong bộ dữ liệu của đề tài.

- Đầu tiên thì chúng ta xét đến mối quan hệ giữa thành tích học tập và thu nhập từ việc làm thêm thông qua biểu đồ scatter được thể hiện ở hình 14



Hình 14. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và thu nhập từ việc làm thêm

* Biểu đồ này thể hiện mối quan hệ giữa điểm trung bình tích lũy (GPA) và thu nhập. Trục hoành thể hiện thu nhập, từ 0 đến 5tr, và trục tung thể hiện GPA, từ 0 đến 10. Các điểm dữ liệu trên biểu đồ thể hiện GPA và thu nhập của các cá nhân khác nhau.
* Nhìn chung, có một mối quan hệ tích cực giữa GPA và thu nhập. Có nghĩa là, những người có GPA cao hơn có xu hướng kiếm được nhiều tiền hơn những người có GPA thấp hơn. Điều này là do GPA cao thường được coi là dấu hiệu của sự thông minh, chăm chỉ và kỷ luật, tất cả đều là những phẩm chất quan trọng để thành công trong công việc.
* Tuy nhiên, mối quan hệ này không hoàn toàn hoàn hảo. Có một số người có GPA thấp nhưng vẫn kiếm được nhiều tiền, và cũng có một số người có GPA cao nhưng thu nhập lại thấp. Điều này là do có nhiều yếu tố khác cũng ảnh hưởng đến thu nhập, chẳng hạn như kinh nghiệm làm việc, kỹ năng, ngành nghề và may mắn..
* Để có thể hiểu rõ hơn mối quan hệ giữa thành tích học tập và thu nhập từ việc làm thêm thì nhóm chúng em đã thống kê chia nhóm lại lại cho các sinh viên theo cách như sau:
  + Đầu tiên, nhóm chúng em sẽ chia thu nhập từ việc làm thêm thành nhiều nhóm có thu nhập theo cột mốc cụ thể:
    - Nhóm thu nhập rất thấp (dưới 1tr)
    - Nhóm thu nhập thấp (từ 1tr đến 2tr)
    - Nhóm thu nhập trung bình (từ 2tr đến 3tr)
    - Nhóm thu nhập khá (từ 3tr đến 4tr)
    - Nhóm thu nhập cao (nhiều hơn 4tr)
  + Đồng thời, về thành tích học tập vẫn chia thành các mốc A, B, C, D và F dựa theo thông tin từ bảng 2 và bảng 3
  + Khi đã có dữ liệu thống kê thì nhóm chúng em sử dụng biểu đồ cột chồng (stacked bar plot) để vẽ biểu đồ thống kê có bao nhiêu sinh viên theo từng nhóm thu nhập và đếm cụ thể có bao nhiêu sinh viên A, B, C, D và F trong từng nhóm (được mô tả trong hình 15)



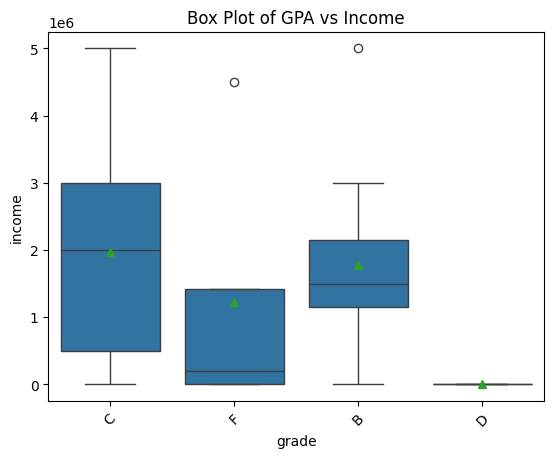
Hình 15. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm thu nhập (với M là million dong nghĩa là triệu đồng)

* Theo hình 15, biểu đồ hiển thị số lượng sinh viên được xếp theo các nhóm thu nhập. Đồng thời cũng đếm số sinh viên trong mỗi nhóm theo thành tích học tập (dữ liệu đã được thống kê theo bảng 5)

Bảng 5. Dữ liệu thống kê số sinh viên theo nhóm thu nhập và thành tích học tập

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thu nhập  Số sinh viên | Dưới 1tr | Từ 1tr đến 2tr | Từ 2tr đến 3tr | Từ 3tr đến 4tr | Nhiều hơn 4tr | Tổng cộng |
| Mức điểm A | 0 | 0 | 0 | 0 | 0 | 0 |
| Mức điểm B | 2 | 4 | 3 | 1 | 1 | 11 |
| Mức điểm C | 4 | 3 | 2 | 4 | 2 | 15 |
| Mức điểm D | 1 | 0 | 0 | 0 | 0 | 1 |
| Mức điểm F | 3 | 0 | 0 | 0 | 1 | 4 |
| Tổng cộng | 10 | 7 | 5 | 5 | 4 | 31 |

* Theo dữ liệu từ bảng 5, ta có 10 sinh viên có thu nhập từ việc làm thêm dưới 1tr7 sinh viên có thu nhập từ 1tr đến 2tr5 sinh viên có thu nhập từ 2tr đến 3tr5 sinh viên có thu nhập từ 3tr đến 4tr và 5 sinh viên có thu nhập từ 4tr trở lên.
* Mức điểm A: Không có sinh viên nào ở tất cả các nhóm thu nhập đạt được mức điểm A.
* Mức điểm B: Tỷ lệ sinh viên đạt mức điểm B cao nhất ở nhóm thu nhập từ 2 triệu đến 3 triệu đồng.
* Mức điểm C: Tỷ lệ sinh viên đạt mức điểm C cao nhất ở nhóm thu nhập từ 1 triệu đến 2 triệu đồng.
* Mức điểm D: Tỷ lệ sinh viên đạt mức điểm D thấp nhất ở tất cả các nhóm thu nhập.
* Mức điểm F: Tỷ lệ sinh viên đạt mức điểm F cao nhất ở nhóm thu nhập dưới 1 triệu đồng.
* Nhìn chung, có thể thấy rằng có mối quan hệ tích cực giữa thu nhập và thành tích học tập của sinh viên. Sinh viên ở các nhóm thu nhập cao hơn có xu hướng đạt điểm cao hơn sinh viên ở các nhóm thu nhập thấp hơn. Tuy nhiên, điều quan trọng cần lưu ý là đây chỉ là một xu hướng chung và không phải là một quy tắc cứng nhắc. Có nhiều yếu tố khác cũng ảnh hưởng đến thành tích học tập của sinh viên, chẳng hạn như năng lực học tập, sự chăm chỉ, môi trường học tập và sự hỗ trợ từ gia đình.

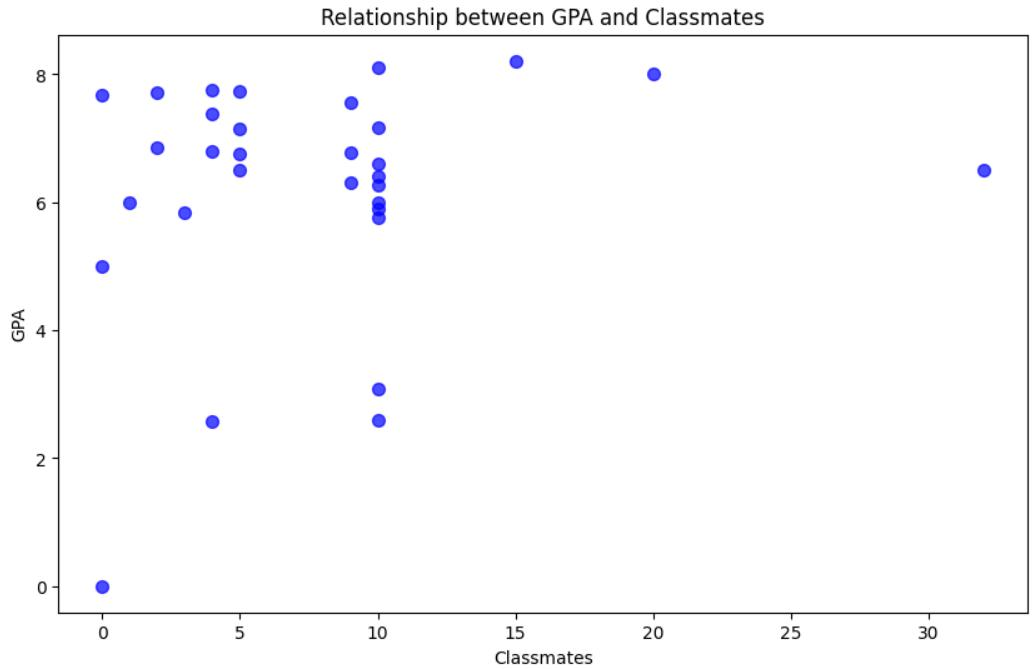


Hình 16. Biểu đồ hộp mô tả việc thống kê số lượng sinh viên theo từng nhóm thu nhập

* Dựa vào biểu đồ hình 16 ta khó mà khẳng định được mối quan hệ giữa thu nhập từ việc làm thêm và GPA:
* Sự nhiễu loạn: Mối quan hệ bị che khuất bởi sự nhiễu loạn do các yếu tố khác ảnh hưởng đến cả GPA và income. Ví dụ: khả năng học tập bẩm sinh, nỗ lực học tập, thời gian dành cho việc học, loại công việc làm thêm, v.v.
* Hiệu ứng lựa chọn: Những người có GPA cao có thể dành nhiều thời gian hơn cho việc học và do đó có ít thời gian hơn để làm thêm, dẫn đến thu nhập từ tiền làm thêm thấp hơn. Tuy nhiên, cũng có thể những người có năng lực học tập tốt hơn có thể tìm được công việc làm thêm với mức lương cao hơn. Tuy nhiên, có thể thấy được một xu hướng nhẹ là những sinh viên có GPA cao có thể có thu nhập từ tiền làm thêm cao hơn. Điều này được thể hiện bởi: Vị trí trung tuyến của hộp income cao hơn so với hộp GPA. Râu của hộp income cũng dài hơn so với hộp GPA, cho thấy sự biến động của income cao hơn
* Biểu đồ hộp cho thấy GPA của nhóm dữ liệu có phân bố không cân đối với giá trị trung bình thấp hơn mức trung tâm của phân bố. Sự biến động của GPA cao hơn ở mức GPA cao và có một số sinh viên có GPA rất thấp.Có thể có một xu hướng nhẹ là những sinh viên có GPA cao có thể có thu nhập từ tiền làm thêm cao hơn. Tuy nhiên, cần có thêm dữ liệu và phân tích để khẳng định mối quan hệ này một cách chính xác.
* Như vậy từ ba biểu đồ trên thì việc sinh viên học kém trong nhóm mức điểm F này có thể không liên quan đến việc thu nhập từ việc làm thêm mà phụ thuộc vào các yếu tố khác ảnh hưởng đến chất lượng học tập của sinh viên trong đó có vấn đề ý thức kỷ luật của chính bản thân sinh viên

### Phân tích sự ảnh hưởng của số lượng số người đồng hành (anh/chị/em trong gia đình hoặc bạn bè) đến thành tích học tập

* Việc có người đồng hành cùng sinh viên trong suốt quá trình học tập tại giảng đường (người đó có thể là anh/chị/em trong gia đình hoặc bạn bè) có thể mang lại nhiều tác động tích cực hoặc tiêu cực đến kết quả học tập của sinh viên. Cụ thể ta có các giả thuyết:
* Giả thuyết liên quan đến các tác động tích cực đến kết quả học tập của sinh viên: Người thân hoặc bạn bè là những người có khả năng động viên, thúc đẩy về mặt ý chí, tinh thần của sinh viên giúp cho sinh viên không còn cảm thấy sự cô đơn trống trải trong môi trường Đại học. Hơn thế nữa, với những anh/chị đi trước hoặc bạn bè cùng lớp cùng ngành còn có thể hỗ trợ nhau về mặt chuyên môn nghiệp vụ, chia sẻ kiến thức cho nhau. Qua đó, sinh viên có thể cải thiện về kết quả học tập.
* Ngược lại, về giả thuyết liên quan đến tác động tiêu cực đến kết quả học tập: Việc sinh viên có nhiều mối quan hệ bạn bè đặc biệt là các mối quan hệ không chất lượng (ở đây chỉ những người bạn chỉ biết cám dỗ khiến cho sinh viên buông lỏng ý chí quyết tâm để học tập để “sa lầy” vào các cuộc vui vô bổ) thì sẽ khiến cho kết quả học tập của sinh viên “trượt dốc không phanh” dần dần khiến cho sinh viên chán nản và muốn bỏ học
* Để có thể biết được thực trạng như thế nào thì thông qua đề tài này, nhóm chúng em sẽ phân tích những ảnh hưởng của số lượng bạn bè và số lượng anh/chị/em sẽ tác động như thế nào đến kết quả học tập của sinh viên bằng những số liệu mà chúng em đã thu thập được để thực hiện đề tài.
* Đầu tiên thì chúng ta hãy xem hình 17 để thấy được mối quan hệ giữa thành tích học tập với số lượng bạn bè của mỗi cá nhân sinh viên.

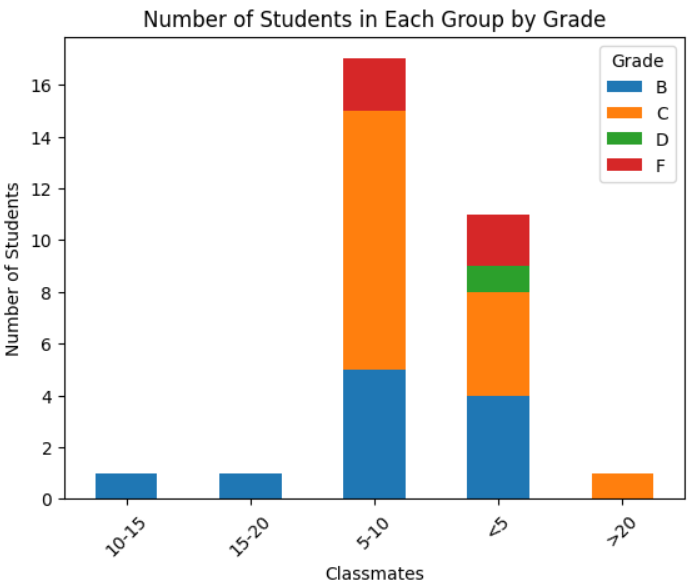


Hình 17. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và số lượng bạn bè

* Biểu đồ (thể hiện bởi hình 17) thể hiện sự tương quan giữa điểm tích lũy hệ 10 (GPA) và số lượng bạn bè. Trong đó, trục hoành thể hiện số bạn bè (classmates) của mỗi cá nhân, từ 0 đến 32, và trục tung thể hiện điểm GPA, từ 0 đến 10. Mỗi điểm dữ liệu trên biểu đồ thể hiện GPA và bạn bè của mỗi cá nhân sinh viên khác nhau.
* Xét về số lượng bạn bè, ta thấy được xu hướng chủ yếu của sinh viên là có từ 5 đến 10 người bạn, nhưng vẫn có một vài trường hợp ngoại lệ như một số sinh viên không có người bạn nào hoặc có quá nhiều bạn cụ ví dụ như có sinh viến có đến 32 người bạn. Xét về điểm số, ta có thấy rằng thành tích của các sinh viên nằm chủ yếu ở khoảng hơn 5.5 cho đến hơn 8 một chút, và cũng có một số sinh viên nằm ở dưới mức 5.
* Nếu phân tích kỹ hơn về sự tương quan giữa điểm GPA và số bạn bè, ta có thể thấy hầu hết các sinh viên có từ 5 đến 10 người bạn đều có số điểm khá cao, thậm chí có sinh viên có số điểm trên 8. Tuy nhiên, ta vẫn có một ngoại lệ khi có 2 sinh viên bị điểm dưới 5 trong khi có 10 người bạn.
* Nhằm làm rõ các mối quan hệ giữa điểm GPA và số bạn bè thì chúng em đã thông kê và chia nhóm các sinh viên thành các nhóm theo số lượng bạn:
* Nhóm có ít hơn 5
* Nhóm có từ 5 - 10
* Nhóm có từ 10 - 15
  + Nhóm có từ 15 - 20
  + Nhóm nhiều hơn 20

Đồng thời, em cũng chia thành tích học tập thành các mốc A, B, C, D và F dựa theo thông tin từ bảng 2 và bảng 3.

* Khi đã có dữ liệu thống kê thì nhóm chúng em sử dụng biểu đồ cột chồng (stacked bar plot) để vẽ biểu đồ thống kê số sinh viên theo từng nhóm đã chi ở trên và đếm cụ thể có bao nhiêu sinh viên nằm ở mốc điểm A, B, C, D và F trong từng nhóm (được mô tả trong hình 18)



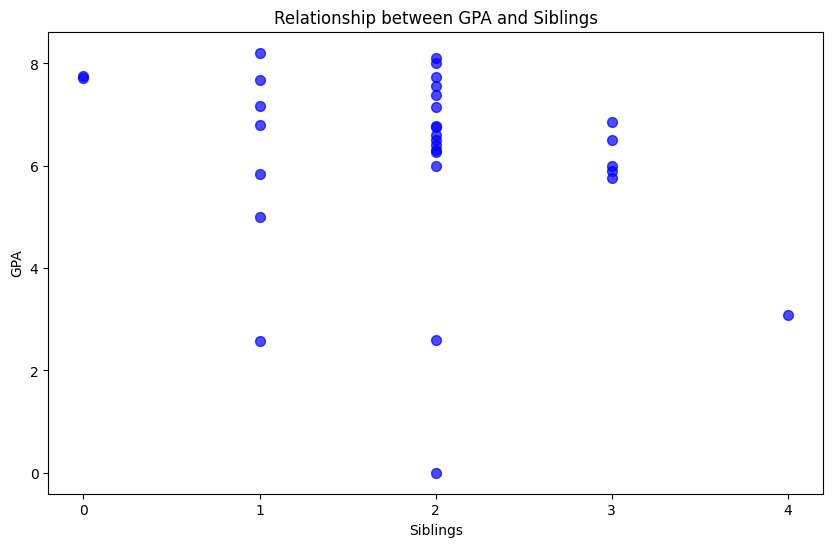
Hình 18. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm bạn

* Trên hình 18, biểu đồ hiển thị số lượng sinh viên được xếp theo các nhóm bạn. Đồng thời cũng đếm số sinh viên trong mỗi nhóm theo thành tích học tập (dữ liệu đã được thống kê theo bảng 6)

Bảng 6. Dữ liệu thống kê số sinh viên theo thành tích học tập và nhóm bạn

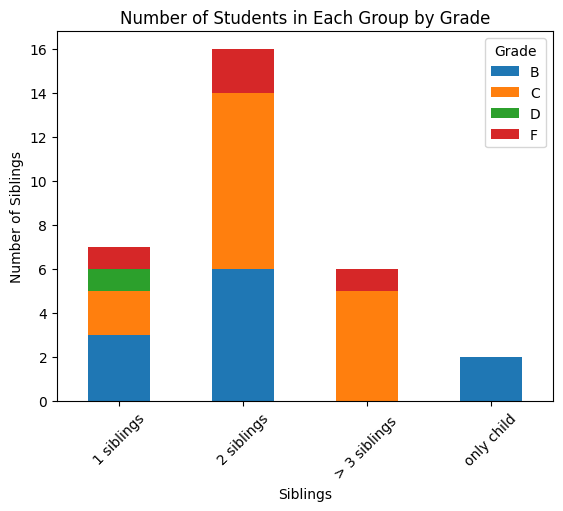
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Số bạn bè  Số sinh viên | <5 | 5-10 | 10-15 | 15-20 | >20 | Tổng cộng |
| Mức điểm A | 0 | 0 | 0 | 0 | 0 | 0 |
| Mức điểm B | 4 | 5 | 1 | 1 | 0 | 11 |
| Mức điểm C | 4 | 10 | 0 | 0 | 1 | 15 |
| Mức điểm D | 1 | 0 | 0 | 0 | 0 | 1 |
| Mức điểm F | 2 | 2 | 0 | 0 | 0 | 4 |
| Tổng cộng | 11 | 17 | 1 | 1 | 1 | 31 |

* Theo dữ liệu từ bảng 6, ta có 11 sinh viên có ít hơn 5 người bạn, 17 sinh viên có từ 5-10 bạn, 1 sinh viên ở mỗi nhóm 10-15, 15-20 và hơn 20 bạn.
  + Mức điểm A: Không có sinh viên nào ở tất cả các nhóm được mức điểm A.
  + Mức điểm B: Tỷ lệ sinh viên đạt mức điểm B cao nhất ở nhóm tử 5-10.
  + Mức điểm C: Tỷ lệ sinh viên đạt mức điểm C cao nhất ở nhóm từ 5-10.
  + Mức điểm D: Tỷ lệ sinh viên đạt mức điểm D thấp nhất ở nhóm ít hơn 5
  + Mức điểm F: Có 2 sinh viên đạt mức điểm F ở cả 2 nhóm ít hơn 5 và nhóm từ 5-10 trên tổng số 4 học sinh bị điểm F.
* Tóm lại, ta có thể thấy được mối quan hệ có chiều hướng tích cực giữa điểm GPA và số bạn bè. Sinh viên nằm ở nhóm có từ 5-10 người bạn có xu hướng có điểm cao hơn so với các nhóm còn lại, vì tỷ lệ sinh viên đạt mức điểm B và C đều được chiếm đa số bởi các sinh viên thuộc nhóm từ 5-10 bạn. Nhưng, vẫn có một vài trường hợp đặc biệt khi có 2 trên sinh viên thuộc vào từ 5-10 bạn nằm trong mức điểm F. Tuy nhiên, chúng ta cũng phải lưu ý rằng số lượng bạn bè chỉ là một trong các yếu tố nhỏ làm ảnh hưởng đến thành tích học tập của sinh viên mà còn có nhiều yếu tố khác như công việc, quãng đường đến trường,… Ngoài ra, đây có thể chỉ là xu hướng chung của bộ dữ liệu này do bộ dữ liệu còn chưa đủ lớn chứ không phải là quy tắc chung áp dụng cho bộ dữ liệu lớn hơn.
* Sau khi phân tích mối quan hệ giữa bạn bè và thành tích học tập, nhóm chúng em tiếp tục phân tích sự ảnh hưởng của số lượng anh/chị/em trong gia đình đến kết quả học tập của sinh viên.
* Ta có hình 17 là một biểu đồ scatter mô tả mối quan hệ giữa số lượng anh chị em và thành tích học tập của sinh viên:



Hình 19. Biểu đồ scatter mô tả mối quan hệ giữa số lượng anh/chị/em trong gia đình và thành tích học tập của sinh viên

* Dựa vào thông tin được cung cấp bởi hỉnh 19, đa số các sinh viên đều nằm ở nhóm có từ 1 đến 2 anh chị em trong gia đình và một số ít còn lại nằm ở nhóm con một (số anh/chị/em bằng 0) hoặc ở nhóm có nhiều hơn 2 anh/chị/em
* Khi xét về thành tích học tập thì một bạn sinh viên thuộc nhóm con một thì có bạn đã đạt thành tích tương đối cao (tiệm cận dưới của mức 8 điểm).
* Trong khi đó, khi xét đến nhóm chiếm đa số (nhóm từ 1 đến 2 anh/chị/em) thì phần lớn các bạn đạt được thành tích khá ấn tượng (đều ở mức trung bình khá - nhóm từ 5 điểm trở lên). Tuy nhiên, trong nhóm này vẫn có tồn tại một vài sinh viên có thành tích không mấy khả quan (mức dưới 5)
* Về nhóm có 3 anh/chị/em thì các bạn có thành tích khá tốt nhưng điều tương phản với nhóm này là một bạn sinh viên có 4 anh/chị/em và mang thành tích tương đối kém.
* Để có thể hiểu rõ hơn sự ảnh hưởng của số lượng anh/chị/em đôi với kết quả học tập của sinh viên thì nhóm chúng em đã chia số sinh viên theo các nhóm sau:
* Nhóm 1: nhóm con một
* Nhóm 2: nhóm có 1 anh/chị/em
* Nhóm 3: nhóm có 2 anh/chị/em
* Nhóm 4: nhóm có nhiều hơn 3 anh/chị/em
* Đồng thời, trong mỗi nhóm các sinh viên đều được chia theo các mức thành tích khác nhau. Để có thể trực quan hoá ý tưởng thì nhóm chúng em sử dụng biểu đồ cột chồng (được thể hiện ở hình 18)



Hình 20. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm số lượng anh/chị/em

* Theo những thông tin được hiển thị ở hình 20, số lượng các sinh viên được thống kê theo các nhóm anh/chị/em và theo các mức thành tích như bảng 7

Bảng 7. Dữ liệu thống kê số sinh viên theo thành tích học tập và nhóm số anh chị em

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Số lượng anh chị em  Mức điểm | Con một | 1 anh/chị/em | 2 anh/chị/em | Nhiều hơn 3 anh/chị/em | Tổng cộng |
| Mức điểm A | 0 | 0 | 0 | 0 | 0 |
| Mức điểm B | 2 | 3 | 6 | 0 | 11 |
| Mức điểm C | 0 | 2 | 8 | 5 | 15 |
| Mức điểm D | 0 | 1 | 0 | 0 | 1 |
| Mức điểm F | 0 | 1 | 2 | 1 | 4 |
| Tổng cộng | 2 | 8 | 16 | 3 | 31 |

* Theo dữ liệu của bảng 7, ta có 2 sinh viên là con 1, 8 sinh viên có 1 anh/chị/em, 16 sinh viên có 2 anh/chị/em và có 3 sinh viên có nhiều hơn 3 anh/chị/em.
* Về thành tích học tập thì không có sinh viên nào đạt mức điểm A, nhóm có đông anh/chị/em (nhóm nhiều hơn 3 anh/chị/em) thì lại có mức điểm cao nhất là C. Trong khi đó, các nhóm còn lại có mức điểm là B.
* Khi xét mức điểm F, nhóm con một thì không có sinh viên nào đạt mức F trong khi 3 nhóm còn lại thì có sinh viên đạt mức điểm này. Khi xét mức điểm D, thì chỉ có nhóm sinh viên có 1 anh/chị/em có một bạn sinh viên đạt mức điểm này.
* Như vậy, từ những dữ liệu được phân tích ở bảng 7 thì gia đình có ít hơn 2 anh/chị/em thì thành tích của họ cao, trong đó 6 sinh viên đạt điểm B và 8 sinh viên đạt điểm C trên tồng số 16 sinh viên. Tuy nhiên cũng có một số trường hợp ngoại lệ như nhóm 1 và 2 anh/chị/em thì có 3 sinh viên bị mức điểm F.
* Nhìn chung, ta có thể thấy có một xu hướng rõ ràng rằng sinh viên có ít hơn 2 anh/chị/em thường có thành tích học tập tốt hơn, với phần lớn đạt mức điểm B hoặc C. Điều này có thể cho thấy sự ổn định và hỗ trợ từ gia đình có thể có ảnh hưởng tích cực đến hiệu suất học tập của sinh viên. Ngược lại, nhóm có nhiều hơn 3 anh/chị/em thường có mức điểm trung bình thấp hơn, với phần lớn đạt mức điểm C. Điều này có thể cho thấy mặc dù có nhiều người trong gia đình, nhưng sự ổn định và hỗ trợ không được tốt như nhóm có ít hơn 2 anh/chị/em. Tóm lại, mặc dù thành tích học tập và số anh/chị/em có liên hệ với nhau, nhưng đây không phải là là một xu hướng luôn chính xác vì ngoài anh/chị/em ra thì vẫn còn nhiều yếu tố khác ảnh hưởng đến thành tích của sinh viên.

# CHƯƠNG 3: NHẬN XÉT VÀ ĐÁNH GIÁ

Từ dữ liệu về những yếu tố ảnh hưởng đến kết quả học tập của sinh viên khoá K21, hệ đào tạo chất lượng cao (CLC) tại Đại học Sài Gòn, nhóm chúng em đã tiến hành phân tích những yếu tố ảnh hưởng đến kết quả học tập của sinh viên. Qua những phân tích, nhóm chúng em đã đúc kết một số đánh giá như sau:

* Về ảnh hưởng của quãng đường từ nhà đến trường đến kết quả học tập của sinh viên thì quãng đường sinh viên đi học càng xa thì thành tích của sinh viên càng giảm. Điều này phù hợp với giả thuyết vì lý do quãng đường đi học xa nên sinh viên phải dành nhiều thời gian hơn để di chuyển. Đồng thời, trong quá trình di chuyển, thể lực và tinh thần không đảm bảo cho sinh viên có thể tập trung cũng như giành thời gian để tương tác với giảng viên cùng các bạn cùng lớp để xây dựng bài học. Tuy nhiên, tồn tại một số ngoại lệ, một số sinh viên nhà gần nhưng lại có thành tích học tập tương đối kém.
* Về ảnh hưởng của mức lương làm thêm đến kết quả học tập của sinh viên thì mức lương của sinh viên nằm trong mức vừa phải (từ 2 đến 3 triệu đồng) thì sinh viên có thành tích học tập cao. Có lẽ là với mức lương này sinh viên đã có thể đáp ứng được nhu cầu sinh hoạt cũng như cho sinh viên một số trải nghiệm thực tế. Tuy nhiên, có một số trường hợp ngoại lệ như nhóm sinh viên có mức lương quá nhiều hoặc quá ít thì thường lơ là trong việc học tập vì nhiều nguyên nhân khác nhau
* Về ảnh hưởng liên quan đến những người đồng hành với sinh viên thì nhóm chúng em sẽ tổng hợp những phân tích đánh giá liên quan đến bạn bè và anh/chị/em trong gia đình
* Về số lượng bạn bè ảnh hưởng đến kết quả học tập của sinh viên thì phần lớn sinh viên có ít bạn (ít hơn 10 người bạn) nhưng lại có thành tích học khá tốt. Tuy nhiên cũng trong nhóm này cũng có vài sinh viên đạt thành tích kém
* Về số lượng anh/chị/em trong gia đình thì đa phần sinh viên có ít hơn 2 anh/chị/em. Số lượng anh/chị/em càng ít thì thành tích học tập càng cao. Tuy nhiên, cũng có những trường hợp ngoại lệ là số anh/chị/em trong khoảng 1 đến 2 nhưng lại có thành tích là F.

Nhìn chung thì các yếu tố được nêu trong báo cáo chỉ có tác động một cách tương đối đối với kết quả học tập của sinh viên. Kết quả học tập của sinh viên còn bị ảnh hưởng từ nhiều yếu tố chủ quan và khách quan khác nhau.

# KẾT LUẬN

Từ những dữ liệu thô liên quan đến những yếu tố ảnh hưởng đến thành tích học tập của sinh viên, nhóm chúng em đã tiến hành phân tích, đánh giá mức độ tương quan giữa các yếu tố tác động đến kết quả học tập. Đồng thời nhóm chúng em đã thực hiện thêm một số công việc nhằm đảm bảo tính đúng đắn và tính bảo mật thông tin cá nhân của sinh viên. Cụ thể như sau:

Đầu tiên, nhóm chúng em đã cùng các sinh viên khoá 21 hệ đào tạo chất lượng cao ngành công nghệ thông tin thuộc trường Đại học Sài Gòn xây dựng nên bộ dữ liệu. Đây là bước đầu tiên và là bước quan trọng để thực hiện các bước tiếp theo

Thứ hai, khi đã xây dựng thành công bộ dữ liệu thì nhóm chúng em đã thực hiện các bước quan sát tổng quan bộ dữ liệu và thực hiện xử lý các dữ liệu khuyết thiếu do nhiều nguyên nhân khác nhau. Việc xử lý các dữ liệu khuyết thiếu nhằm đảm bảo độ chính xác của kết quả phân tích dữ liệu giúp cho việc phân tích dữ liệu đạt được hiệu quả cao hơn

Thứ ba, nhóm chúng em tiến hành mã hoá các thông tin cá nhân nhạy cảm của các sinh viên. Trong thực tế, việc bảo vệ thông tin cá nhân là việc rất quan trọng nhằm phòng tránh những rủi ro do rò rỉ hay lộ thông tin cá nhân.

Cuối cùng, nhóm chúng em đã tiến hành việc phân tích mối quan hệ tương quan của những yếu tố ảnh hưởng đến thành tích học tập của sinh viên một cách trực quan. Qua đó, nhóm chúng em đã đưa ra những nhận xét dựa trên dữ liệu mức độ ảnh hưởng của từng yếu tố và đi đến kết luận các yếu tố chỉ ảnh hưởng đến việc học của sinh viên chỉ ở mức tương đối. Việc học tập của sinh viên còn phụ thuộc vào nhiều yếu tố chủ quan và khách quan khác nhau cần được nghiên cứu và làm rõ.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Trường Đại học Sài Gòn, Sổ tay sinh viên năm học 2022-2023,  TPHCM: Trường Đai học Sài Gòn, 2022. |
| [2] | B. Q. Hoạt, "Tìm hiểu thuật toán mã hóa khóa đối xứng AES - viblo asia,"  viblo, 29 July 2017.  [Online]. Available: https://viblo.asia/p/tim-hieu-thuat-toan-ma-hoa-khoa-doi-xung-aes-gAm5yxOqldb?fbclid=IwZXh0bgNhZW0CMTAAAR0baazPnqVUOTAC-E\_hT3Kf4HYYd2RL7K4XLB9HpPMgqq2kRw1VS3818I\_  aem\_AZtq2uSEBFuN5HnvRQhNNTechrpwIVaqz0qMDYrYNnGwxIQaYF1u5JfDuSC1f4d-SNueWRm7k6Zi7owhI\_9dLV7c. [Accessed 12 May 2024]. |

# PHỤ LỤC

## Bộ mã nguồn chương trình

* Kết nối Google Drive

|  |
| --- |
| from google.colab import drive  drive.mount('/content/drive')  base\_path = '/content/drive/MyDrive/PTDL/data\_and\_code/final\_assignment' |

* Cài đặt pyspark

|  |
| --- |
| %pip install pyspark |

* Tạo Spark session

|  |
| --- |
| from pyspark.sql import SparkSession  def spark\_session(app\_name):  spark = SparkSession.builder \  .appName(app\_name)\  .getOrCreate()  # To make sure that we are working with the correct version of Spark  print("Running PySpark version: ", spark.version)  return spark  spark = spark\_session('assignment1') |

* Đọc dữ liệu từ file csv

|  |
| --- |
| dataset\_path='/content/drive/MyDrive/PTDL/data\_and\_code/final\_assignment /final\_assignment\_dataset.csv' #dataset path string  data = spark.read.csv(dataset\_path, header=True) |

* Đếm số dòng dữ liệu

|  |
| --- |
| # Count the number of rows in the DataFrame  row\_count = data.count()  # Print the number of rows  print("Number of rows in the DataFrame:", row\_count) |

* In lược đồ cấu trúc (schema)

|  |
| --- |
| data.printSchema() |

* Chỉnh sửa kiểu dữ liệu của các schema

|  |
| --- |
| from pyspark.sql.functions import col  data = data.withColumn("1st\_GPA", col("1st\_GPA").cast("float"))  data = data.withColumn("2nd\_GPA", col("2nd\_GPA").cast("float"))  data = data.withColumn("distance", col("distance").cast("float"))  data = data.withColumn("income", col("income").cast("integer"))  data = data.withColumn("siblings", col("siblings").cast("integer"))  data = data.withColumn("classmates", col("classmates").cast("integer")) |

* Gán các giá trị thuộc ô dữ liệu khuyết thiếu thành giá trị bằng 0

|  |
| --- |
| from pyspark.sql.functions import when  # Chuyển đổi giá trị null trong cột "income" thành 0  data = data.withColumn("income", when(data["income"].isNull(), 0).otherwise(data["income"]))  data = data.withColumn("classmates", when(data["classmates"].isNull(), 0).otherwise(data["classmates"]))  data.show() |

* Ứng dụng mã hoá AES mã hoá các thông tin cá nhân của sinh viên

|  |
| --- |
| %pip install pycryptodome  from Crypto.Cipher import AES  from Crypto.Util.Padding import pad, unpad  from Crypto.Random import get\_random\_bytes  import base64  from pyspark.sql.functions import udf  from pyspark.sql.types import StringType  # Khởi tạo khóa bí mật  key = b'n\x1ak\x06{\x12\xeb\xe3\x19=\x91d\xad\xa7\xaeY' # 128-bit (16 bytes) key  # Hàm để mã hóa dữ liệu  def encrypt\_data(data, key):  cipher = AES.new(key, AES.MODE\_CBC)  ct\_bytes = cipher.encrypt(pad(data.encode(), AES.block\_size))  iv = base64.b64encode(cipher.iv).decode('utf-8')  ct = base64.b64encode(ct\_bytes).decode('utf-8')  return iv + "|" + ct  # Hàm UDF để áp dụng mã hóa cho mỗi giá trị trong cột  encrypt\_udf = udf(lambda x: encrypt\_data(x, key), StringType())  # Áp dụng mã hóa cho các cột và thêm vào DataFrame  # Mã hóa các cột trên DataFrame hiện tại  df\_encrypted = data.withColumn("student\_id", encrypt\_udf(col("student\_id"))) \  .withColumn("fullname", encrypt\_udf(col("fullname"))) \  .withColumn("class", encrypt\_udf(col("class")))  # Hiển thị DataFrame sau khi mã hóa  df\_encrypted.show()  # Hàm để giải mã dữ liệu  def decrypt\_data(data, key):  iv, ct = data.split("|")  iv = base64.b64decode(iv)  ct = base64.b64decode(ct)  cipher = AES.new(key, AES.MODE\_CBC, iv)  pt = unpad(cipher.decrypt(ct), AES.block\_size)  return pt.decode('utf-8')  # Hàm UDF để áp dụng giải mã cho mỗi giá trị trong cột  decrypt\_udf = udf(lambda x: decrypt\_data(x, key), StringType())  # Áp dụng giải mã cho các cột đã mã hóa và thêm vào DataFrame  df\_decrypted = df\_encrypted.withColumn("student\_id", decrypt\_udf(df\_encrypted["student\_id"])) \  .withColumn("fullname", decrypt\_udf(df\_encrypted["fullname"])) \  .withColumn("class", decrypt\_udf(df\_encrypted["class"]))  # Hiển thị DataFrame sau khi giải mã  df\_decrypted.show()  # Lưu DataFrame đã được thay đổi thành tập tin CSV  output\_path = '/content/drive/MyDrive/PTDL/data\_and\_code/final\_assignment/final\_assignment\_dataset\_modified/'  df\_encrypted.write.csv(output\_path, header=True, mode="overwrite") |

* In thử mã số sinh viên của một sinh viên sau khi mã hoá

|  |
| --- |
| # In ra một dòng dữ liệu đã được mã hóa  df\_encrypted.select("student\_id").show(1, truncate=False)  # Giải mã dòng dữ liệu đó  print(decrypt\_data('HS9OhhjRbrRa2yaCbfrJVA==|DS0ki98W/Yap51SfBybtRA==',key)) |

* Tính điểm GPA của các sinh viên bằng cách áp dụng công thức (2) và làm tròn 2 số thập phân

|  |
| --- |
| from pyspark.sql.functions import round  GPA\_df = df\_encrypted.withColumn("GPA", round((col("1st\_GPA")+col("2nd\_GPA"))/2,2)) |

* Xếp loại sinh viên theo điều kiện của bảng 2

|  |
| --- |
| from pyspark.sql.functions import expr  GPA\_df = df\_encrypted.withColumn("GPA", round((col("1st\_GPA")+col("2nd\_GPA"))/2,2))  grade\_df = GPA\_df.withColumn("grade", expr("CASE WHEN GPA >= 8.5 THEN 'A' WHEN GPA < 8.5 AND GPA >=7 THEN 'B' WHEN GPA >= 5.5 AND GPA <7 THEN 'C' WHEN GPA >=4 AND GPA < 5.5 THEN 'D' ELSE 'F' END"))  grade\_df.show() |

* Sắp xếp dữ liệu theo điểm số GPA

|  |
| --- |
| from pyspark.sql.functions import col  grade\_df = grade\_df.orderBy(col("GPA").desc())  grade\_df.select("GPA", "grade").show(20) |

* Sắp xếp dữ liệu theo số lượng bạn bè

|  |
| --- |
| from pyspark.sql.functions import col  # Sắp xếp DataFrame theo cột "classmates"  grade\_classmates = grade\_df.orderBy(col("classmates"))  # Hiển thị 30 hàng đầu tiên  grade\_classmates.select("classmates", "GPA", "grade").show() |

* Sắp xếp dữ liệu theo số lượng anh/chị/em trong gia đình

|  |
| --- |
| from pyspark.sql.functions import asc, col  # Sắp xếp DataFrame theo cột "siblings"  grade\_siblings = grade\_df.orderBy(col("siblings").asc())  # Hiển thị 30 hàng đầu tiên  grade\_siblings.select("siblings", "GPA", "grade").show() |

* Vẽ biểu đồ cột mô tả số sinh viên được xếp theo các mức thành tích học tập

|  |
| --- |
| # prompt: visualize how many students in each group  # Failure group who has GPA under 5  # Satisfied group who has GPA from 5 to 7.9  # Good group who has GPA from 8 to 8.9  # Excellence group who has GPA greater or equal 9  # Count the number of students in each group  A\_count = grade\_df.filter(grade\_df["grade"] == 'A').count()  B\_count = grade\_df.filter(grade\_df["grade"] == 'B').count()  C\_count = grade\_df.filter(grade\_df["grade"] == 'C').count()  D\_count = grade\_df.filter(grade\_df["grade"] == 'D').count()  F\_count = grade\_df.filter(grade\_df["grade"] == 'F').count()  # Create a bar chart  labels = ["A","B","C","D","F"]  values = [A\_count, B\_count, C\_count, D\_count, F\_count]  plt.bar(labels, values)  # Set axis labels and title  plt.xlabel("Group")  plt.ylabel("Number of Students")  plt.title("Number of Students in Each Group")  # Display the bar chart  plt.show() |

* Vẽ đồ thị scatter mô tả mối quan hệ giữa thành tích học tập của sinh viên và khoảng cách từ nhà đến trường

|  |
| --- |
| import matplotlib.pyplot as plt  # Assuming 'df\_encrypted' is a Pandas DataFrame with columns 'GPA' and 'distance'  pdf = GPA\_df.toPandas()  # Extract GPA and distance values into separate lists  gpa\_values = pdf['GPA'].tolist()  distance\_values = pdf['distance'].tolist()  # Create a scatter plot  plt.scatter(distance\_values, gpa\_values)  # Set axis labels and title  plt.xlabel('Distance')  plt.ylabel('GPA')  plt.title('Relationship between GPA and Distance')  # Display the plot  plt.show() |

* Xếp các sinh viên thành các nhóm quãng đường từ nhà đến trường theo thứ tự tăng dần

|  |
| --- |
| distance\_groups =grade\_df.withColumn("distance\_group", expr("CASE WHEN distance <5 THEN 'less than 5km' WHEN distance >=5 AND distance <=10 THEN 'from 5 to 10km' WHEN distance > 10 AND distance <15 THEN 'from 10 to 15km' WHEN distance >= 15 AND distance < 20 THEN 'from 15 to 20km' ELSE 'greater than 20km' END")) |

* Vẽ biểu đồ cột chồng mô tả số sinh viên theo từng mức thành tích và quãng đường đi học của họ

|  |
| --- |
| # prompt: visualize stacked barplot to show how many students have grade A, B, C, D and F in each group distance less than 5km, distance from 5 to 10km, distance from 15 to 20 km and >20km  # Extract the required columns into Pandas DataFrame  pdf = distance\_groups.select('distance\_group', 'grade').toPandas()  # Group the data by distance and grade  grouped\_pdf = pdf.groupby(['distance\_group', 'grade']).size().unstack().fillna(0)  # Create stacked bar plot  grouped\_pdf.plot(kind='bar', stacked=True)  # Set axis labels and title  plt.xlabel('Distance')  plt.ylabel('Number of Students')  plt.title('Number of Students in Each Group by Grade')  plt.xticks(rotation = 45)  # Add legend  plt.legend(title='Grade')  # Display the plot  plt.show() |

* Vẽ biểu đồ cột mô tả mối quan hệ giữa thành tích học tập và thu nhập do làm thêm

|  |
| --- |
| import matplotlib.pyplot as plt  # Giả sử 'df\_encrypted' là một PySpark DataFrame với các cột 'GPA' và 'income'  pdf = GPA\_df.toPandas()  # Sắp xếp dữ liệu theo GPA  sorted\_pdf = pdf.sort\_values(by='GPA')  # Tạo biểu đồ cột  plt.figure(figsize=(10, 6))  # Vòng lặp để tạo từng cột  for index, row in sorted\_pdf.iterrows():  plt.bar(row['GPA'], row['income'], color='skyblue', edgecolor='black') # Tạo một cột cho mỗi cặp giá trị GPA và thu nhập  # Đặt nhãn và tiêu đề  plt.xlabel('GPA')  plt.ylabel('Income')  plt.title('Relationship between GPA and Income (Bar Plot)')  plt.xticks(rotation=45) # Xoay nhãn trục x để tránh chồng chéo  plt.tight\_layout()  plt.show() |

* Vẽ đồ thị scatter mô tả quan hệ giữa thành tích học tập và thu nhập do làm thêm

|  |
| --- |
| import matplotlib.pyplot as plt  # Assuming 'df\_encrypted' is a Pandas DataFrame with columns 'GPA' and 'income'  pdf = GPA\_df.toPandas()  # Extract GPA and income values into separate lists  gpa\_values = pdf['GPA'].tolist()  income\_values = pdf['income'].tolist()  # Create a scatter plot  plt.scatter(distance\_values, gpa\_values)  # Set axis labels and title  plt.xlabel('Income')  plt.ylabel('GPA')  plt.title('Relationship between GPA and Income')  # Display the plot  plt.show() |

* Vẽ biểu đồ scatter mô tả mối quan hệ tổng quát giữa số lượng bạn bè và thành tích học tập

|  |
| --- |
| import matplotlib.pyplot as plt  # Trích xuất dữ liệu từ DataFrame PySpark và chuyển đổi thành mảng dữ liệu thường  classmates = grade\_df.select('classmates').rdd.flatMap(lambda x: x).collect()  gpa = grade\_df.select('GPA').rdd.flatMap(lambda x: x).collect()  # Tạo biểu đồ scatter plot sử dụng Matplotlib cho số bạn bè  plt.figure(figsize=(10, 6))  plt.scatter(classmates, gpa, s=50, c='blue', alpha=0.7)  # Đặt tiêu đề và nhãn trục  plt.title('Mối quan hệ giữa điểm và số bạn bè')  plt.xlabel('Số bạn bè')  plt.ylabel('Điểm GPA')  # Hiển thị biểu đồ  plt.show() |

* Vẽ biểu đồ scatter mô tả mối quan hệ giữa số lượng anh/chị/em trong gia đình và kết quả học tập

|  |
| --- |
| import matplotlib.pyplot as plt  # Trích xuất dữ liệu từ DataFrame PySpark và chuyển đổi thành mảng dữ liệu thường  classmates = grade\_df.select('siblings').rdd.flatMap(lambda x: x).collect()  gpa = grade\_df.select('GPA').rdd.flatMap(lambda x: x).collect()  # Tạo biểu đồ scatter plot sử dụng Matplotlib cho số bạn bè  plt.figure(figsize=(10, 6))  plt.scatter(classmates, gpa, s=50, c='blue', alpha=0.7)  # Đặt tiêu đề và nhãn trục  plt.title('Mối quan hệ giữa điểm và số anh/chị/em')  plt.xlabel('Số anh chị em')  plt.ylabel('Điểm GPA')  # Hiển thị biểu đồ  plt.show() |

* Vẽ biểu đồ cột chồng mô tả số sinh viên được xếp theo các nhóm có số lượng bạn bè khác nhau. Trong đó các nhóm sinh viên được đếm số lượng sinh viên dựa trên thành tích

|  |
| --- |
| pdf = classmates\_groups.select('classmates\_group', 'grade').toPandas()  # Group the data by distance and grade  grouped\_pdf = pdf.groupby(['classmates\_group', 'grade']).size().unstack().fillna(0)  # Create stacked bar plot  grouped\_pdf.plot(kind='bar', stacked=True)  # Set axis labels and title  plt.xlabel('Classmates')  plt.ylabel('Number of Students')  plt.title('Number of Students in Each Group by Grade')  plt.xticks(rotation = 45)  # Add legend  plt.legend(title='Grade')  # Display the plot  plt.show() |

* Vẽ biểu đồ cột chồng mô tả số sinh viên được xếp theo các nhóm có số lượng bạn bè khác nhau. Trong đó các nhóm sinh viên được đếm số lượng sinh viên dựa trên thành tích

|  |
| --- |
| pdf = siblings\_groups.select('siblings\_group', 'grade').toPandas()  # Group the data by distance and grade  grouped\_pdf = pdf.groupby(['siblings\_group', 'grade']).size().unstack().fillna(0)  # Create stacked bar plot  grouped\_pdf.plot(kind='bar', stacked=True)  # Set axis labels and title  plt.xlabel('Siblings')  plt.ylabel('Number of Siblings')  plt.title('Number of Students in Each Group by Grade')  plt.xticks(rotation = 45)  # Add legend  plt.legend(title='Grade')  # Display the plot  plt.show() |

## Bảng phân công công việc

|  |  |  |
| --- | --- | --- |
| STT | Công việc | Người thực hiện |
| 1 | Liên hệ xây dựng bộ dữ liệu | Nguyễn Công Thọ |
| 2 | Tiền xử lý dữ liệu   * Chỉnh sửa kiểu dữ liệu * Xử lý dữ liệu khuyết thiếu * Mã hoá thông tin cá nhân nhạy cảm | Lê Công Minh |
| 3 | Phân tích sự ảnh hưởng của khoảng cách từ nhà đến trường đến kết quả học tập | Nguyễn Công Thọ |
| 4 | Phân tích sự ảnh hưởng của số lượng anh/chị/em đến kết quả học tập | Lê Minh Phúc |
| 5 | Phân tích sự ảnh hưởng của số lượng bạn bè đến kết quả học tập | Lê Minh Phúc |
| 6 | Phân tích sự ảnh hưởng của mức thu nhập do làm thêm đến kết quả học tập | Lê Công Minh |
| 7 | Viết báo cáo tổng kết | Tất cả thành viên |