ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HOC MÁY TÍNH

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG – CS313.N22

Bài tập Thực hành 1: *Tìm hiểu và phân tích dữ liệu*

Nhóm 3:

Bùi Nguyễn Anh Trung (NT)	20520332
Hồ Thanh Tịnh	20520813
Nguyễn Trần Minh Anh	20520394
Lê Nguyễn Bảo Hân	20520174
Nguyễn Văn Đức Ngọc	20521666

Hồ Chí Minh, 03 tháng 05 năm 2023

Nội dung

Bảng phân công

Thành viên / Công Việc	Bùi Nguyễn Anh Trung	Hồ Thanh Tịnh	Nguyễn Trần Minh Anh	Lê Nguyễn Bảo Hân	Nguyễn Văn Đức Ngọc
Phân công, quản lý công việc chung			✓		
Tìm hiểu dữ liệu	✓	✓	✓	✓	✓
Làm sạch dữ liệu			✓		
Phân tích dữ liệu		✓			✓
Giải thích kết quả				✓	
Tổng hợp nội dung và định dạng báo cáo				✓	
Mức độ hoàn thiện (%)	100%	100%	100%	100%	100%

1. Tìm hiểu dữ liệu

Bảng 1.1 Mô tả các bảng dữ liệu

STT	Tên bảng	Diễn giải	Chi tiết
1	SINHVIEN	Danh sách sinh viên	Gồm các thông tin cá nhân (như năm sinh, giới tính, nơi sinh,…) và thông tin ngành học, khóa học của sinh viên. Gồm 8295 điểm dữ liệu.
2	DIEM	Danh sách điểm các môn học của sinh viên	Gồm các thông tin về môn học (như số tín chỉ, môn học tiếp theo), thời điểm học, điểm của sinh viên và trạng thái môn học
3	CHUNGCHI	Danh sách chứng chỉ ngoại ngữ của sinh viên	Gồm thông tin về chứng chỉ ngoại ngữ của sinh viên (giấy xác nhận, điểm các kỹ năng) và kết quả so với chuẩn đầu ra. Có 3464 điểm dữ liệu.
4	XEPLOAIAV	Danh sách xếp loại lớp Anh Văn của sinh viên	Gồm thông tin điểm thi anh văn đầu vào (nghe + đọc) và kết quả xếp lớp. Có 6349 điểm dữ liệu.
5	THISINH	Danh sách thí sinh và kết quả thi tốt nghiệp	Gồm thông tin trường THPT của thí sinh, tỉnh, diện tốt nghiệp và điểm tốt nghiệp
6	GIAYXACNHAN	Danh sách các loại giấy xác nhận của sinh viên	Gồm thông tin sinh viên yêu cầu (loại giấy, ngày nộp, lý do) và các

			trạng thái xác nhận (đã ký, đóng dấu, phát,…)
8	XLHV	Danh sách sinh viên bị xử lý học vụ	Gồm thông tin loại cảnh cáo, lý do, thời điểm và số quyết định, ngày quyết định. Có 3452 điểm dữ liệu.
10	DIEMRL	Danh sách điểm rèn luyện của sinh viên	Gồm thông tin sinh viên (lớp sinh hoạt), học kỳ và điểm. Có 54058 điểm dữ liệu.
12	BAOLUU	Danh sách sinh viên bảo lưu	Gồm thông tin tình trạng, học kỳ, lý do bảo lưu và số quyết định, ngày quyết định. Có 1878 điểm dữ liệu.
14	TOTNGHIEP	Danh sách sinh viên tốt nghiệp	Gồm thông tin loại tốt nghiệp, số quyết định và ngày cấp văn bằng. Có 1845 điểm dữ liệu.

Mô tả từng bảng dữ liệu

Bảng 1.2 Mô tả bảng SINHVIEN và các thuộc tính

Bảng SINHVIEN				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	id	Số thứ tự	Số	
2	masv_tham so dau	Mã số sinh viên tham số đầu, là khóa của sinh viên	Chuỗi	Chuỗi 4 ký tự hoặc 2 ký tự. Gồm các khóa: 1252, 1352 và khóa 13 đến 19
3	masv	Mã sinh viên	Chuỗi	Chuỗi 40 ký tự

4	namsinh	Năm sinh	Số và chuỗi	Từ 1979 đến 2001, có rỗng ''
5	gioitinh	Giới tính	Số	• 0: Nữ • 1: Nam
6	noisinh	Nơi sinh	Chuỗi	Có rỗng ''
7	lopsh	Lớp sinh hoạt	Chuỗi	• Biểu diễn dưới dạng KHMT2016 hoặc KHMT2016.1 Với 4 ký tự đầu là ký hiệu khoa, 4 ký tự tiếp theo là năm nhập học và phân lớp '.1', '.2', '.3' (nếu có) • Một vài dữ liệu có phần năm nhập học là '0001' Ví dụ như KTPM0001
8	khoa	Khoa	Chuỗi	• Có 6 khoa: CNPM, HTTT, KHMT, KTTT, MMT&TT • Gồm 2 dạng biểu diễn CNPM và 'CNPM'
9	hedt	Hệ đào tạo	Chuỗi	• Có 5 hệ: CLC, CNTN, CQUI, CTTT, KSTN • Gồm 2 dạng biểu diễn là CLC và 'CLC'
10	khoahoc	Khóa học	Số	Từ 8 đến 14
11	chuyennganh2	Mã chuyên ngành 2	Chuỗi	Biểu diễn dưới dạng D480102 hoặc 'D480102'
12	tinhtrang	Tình trạng	Số	Gồm các tình trạng: • 1: • 2: Cảnh cáo • 3: • 4: • 5: Thôi học • 6: • 7: Gia hạn • 8: Tự do

13	diachi_tinhtp	Địa chỉ tỉnh, thành phố	Chuỗi	Tồn tại rỗng (‘ và ‘’) hoặc dữ liệu lệch sang cột tiếp theo.

Bảng 1.3 Mô tả bảng DIEM và các thuộc tính

Bảng DIEM				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	id	Số thứ tự	Số	
2	masv	Mã số sinh viên	Chuỗi	Chuỗi 40 ký tự
3	mamh	Mã môn học	Chuỗi	Biểu diễn dưới dạng ký hiệu và số như ‘IT001’, ’ADENG1’,…
4	malop	Mã lớp	Chuỗi	Gồm ký hiệu mã môn học, mã phân lớp và ký hiệu khoa (nếu có) Ví dụ: IT001.D11, IT001.E11.ANTT
5	sotc	Số tín chỉ	Số	• Từ 0 đến 5 • sotc = 0 là những môn không tính vào ĐTB và điểm tích lũy
6	namhoc	Năm học	Số	Từ năm 2012 đến 2016
7	hocky	Học kỳ	Số	Học kỳ 1, 2 và 3
8	diem	Điểm	Số thực	Từ 0 đến 10 hoặc NULL
9	trangthai	Trạng thái môn học	Số	Gồm các trạng thái: • 0: Hủy • 1: Bình Thường • 2: Trả Nợ • 3: Cải Thiện • 4: Miễn • 5: Hoãn

10	mamh_tt	Mã môn học tiếp theo	Chuỗi	Mã môn học tiếp theo hoặc NULL

Bảng 1.4 Mô tả bảng CHUNGCHI và các thuộc tính

Bảng CHUNGCHI				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	id	ID sinh viên	Số	
2	masv	Mã số sinh viên đã qua mã hóa	Chuỗi	Chuỗi 40 ký tự
3	ngaythi	Ngày sinh viên thi lấy chứng chỉ ngoại ngữ	Date	
4	url	Nguồn dẫn tới ảnh xác nhận chứng chỉ ngoại ngữ	Chuỗi	
5	loaixn	Loại giấy xác nhận mà sinh viên cung cấp	Chuỗi	
6	url	Điểm kỹ năng nghe	Số	Thuộc tính bị lặp lại Dữ liệu là của thuộc tính ***listening***
7	loaixn	Điểm kỹ năng nói	Số	Thuộc tính bị lặp lại Dữ liệu là của thuộc tính ***speaking***
8	listening	Điểm kỹ năng đọc	Số	Dữ liệu là của thuộc tính ***reading***
9	speaking	Điểm kỹ năng viết	Số	Dữ liệu là của thuộc tính ***writing***
10	reading	Điểm tổng	Số	Dữ liệu là của thuộc tính ***tongdiem***

11	writing	Chuẩn kỹ năng	Chuỗi	Dữ liệu là của thuộc tính ***lydo***
12	tongdiem	Trình độ ngoại ngữ của sinh viên so với chuẩn đầu ra	Chuỗi	Dữ liệu là của thuộc tính ***trangthai*** Thường là số tự nhiên Nếu khác số tự nhiên tức là bị dữ liệu của thuộc tính *writing* lấn sang một cột
13	lydo	Ngày sinh viên nhận xếp loại trình độ ngoại ngữ	Date	Dữ liệu là của thuộc tính ***ngayxl***
14	trangthai		Date	Thường là trống, nếu có dữ liệu thì tức là dữ liệu bị lấn qua từ thuộc tính *lydo*
15	ngayxl			Thuộc tính trống do hai thuộc tính thừa lấn dữ liệu

Bảng 1.5 Mô tả bảng XEPLOAIAV và các thuộc tính

Bảng XEPLOAIAV				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	id	ID sinh viên	Số	
2	masv	Mã số sinh viên đã qua mã hóa	Chuỗi	Chuỗi 40 ký tự
3	listening	Điểm kỹ năng nghe	Số	
4	reading	Điểm kỹ năng đọc	Số	
5	total	Điểm tổng	Số	

6	mamh	Mã trình độ tiếng anh có được từ điểm tổng	Chuỗi	Chuỗi 5 ký tự Xen lẫn cách tính điểm TOEIC
7	ghichu	Ghi chú thêm về điểm	Chuỗi	

Bảng 1.6 Mô tả bảng THISINH và các thuộc tính

Bảng THISINH				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	masv	Mã số sinh viên	Chuỗi	
2	dien_tt	Diện tốt nghiệp	Chuỗi	
3	diem_tt	Điểm thi tốt nghiệp	Số thực	
4	lop12_matinh	Mã tỉnh	Số	Số nguyên từ 1 đến 63
5	lop12_matruong	Mã trường	Số	
6	TEN_TRUONG	Tên trường	Chuỗi	

Bảng 1.7 Mô tả bảng GIAYXACNHAN và các thuộc tính

Bảng GIAYXACNHAN				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	giayxacnhan_id	Mã ID giấy xác nhận	Chuỗi	

2	maloaigiay	Mã loại giấy	Số nguyên thuộc miền [1,7]	Giấy xác nhận sinh viên Giấy làm lại thẻ sinh viên Giấy vay vốn ngân hàng Giấy miễn giảm học phí Giấy xác nhận điểm rèn luyện Giấy xác nhận Ưu đãi Giáo dục Giấy xác nhận Xác nhận Học bổng
3	ngaysubmit	Ngày tháng nộp giấy yêu cầu	Datatime	
4	masv	Mã sinh viên yêu cầu giấy xác nhận	Chuỗi	
5	lydoxacnhan	Lí do xác nhận	Chuỗi	
6	dain	Trạng thái đã in giấy xác nhận chưa	Số nguyên miền giá trị [0,1]	0: Chưa in 1: Đã in
7	baosai	Trạng thái có báo sai hay không	Số nguyên miền giá trị [0,1]	0: Báo đúng 1: Báo sai
8	lydocapthe	Lý do cấp thẻ	Chuỗi	Dùng khi mã xác nhận loại giấy là 1
9	hocky	Học kì đang yêu cầu giấy	Số nguyên miền giá trị [0,2]	
10	namhoc	Năm học	Chuỗi	
11	lydosai	Lí do báo sai giấy yêu cầu	Chuỗi	Dùng khi baosai = 1

12	daky	Đã ký giấy xác nhận	Số nguyên miền giá trị [0,1]	0: Chưa ký 1: Đã ký
13	dadongdau	Đã đống dấu giấy xác nhận	Số nguyên miền giá trị [0,1]	0: Chưa đóng dấu 1: Đã đóng dấu
14	daphat	Đã phát ra giấy xác nhận cho sinh viên	Số nguyên miền giá trị [0,1]	0: Chưa phát 1: Đã phát
15	trangthai	Không xác định	Số nguyên miền giá trị [-3,3]	
16	ngayphat	Ngày phát giấy	Datetime	

Bảng 1.8 Mô tả bảng XLHV và các thuộc tính

Bảng XLHV				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	id	Số thứ tự	Số	Theo form xxyyy, trong đó xx là số thứ tự từ 7-21, yyy là số thứ tự từ 1-999
2	masv	Mã sinh viên được mã hóa	Chuỗi	
3	tinhtrang	Mức độ cảnh cáo, các loại cảnh cáo	Số	Chỉ có 2,5,8

4	lydo	Các lý do ứng với mã của tinhtrang	Chuỗi	*(2): BỊ CẢNN CÁO vì: ĐTB HK 1<3, ĐTB 2 học kỳ liên tiếp <4, đóng học phí trễ || BUỘC THÔI HỌC: được hội đồng đang xem xét hạ mức *(5): BUỘC THÔI HỌC vì: ĐTB HK=0, hết hạn bảo lưu không nhập học lại, bị cảnh cáo 3 lần liên tiếp,'BTH ???c H? xem xét chuy?n T? xa' *(8): Sinh viên quá thời gian theo thiết kế của chương trình đạo tạo chuy?n t? do', BTH được hội đồng xem xét chuy?n t? do'
5	hocky	Học kỳ	Số và có nhiễu	Ngoài 1,2, còn có 3=0, 3<3, 3 < 3\n- Ä□Ã³ng há»□c phÃ- trá»…
6	namhoc	Năm học	Date hoặc Số	Từ 2016-2020, còn có 1 (do lỗi thụt lề)
7	soqd	Mã quyết định để phân biệt các đợt quyết định khác nhau	Chuỗi	Có form 'xxx/QÄ-ÄHCNTT' trong đó xxx là số {155, 180, 189, 223, 244, 245, 252, 633, 692, 713, 714, 748}
8	ngayqd	Ngày quyết định theo đợt	Date	Theo định dạng yyyy-mm-dd và 'yyyy-mm-dd' (có nhiễu)

Bảng 1.9 Mô tả bảng DIEMRL và các thuộc tính

Bảng DIEMRL				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	id	Số thứ tự	Số	Từ 7 - 136342, một vài dữ liệu ở cuối có lỗi chính tả. Ví dụ: '(136342'

2	masv	Mã sinh viên	Chuỗi	
3	lopsh	Mã lớp	Chuỗi	"xxxxyyyy" x là tên viết tắt của khoa y là năm học từ 2013-2019 Có nhiễu của mỗi khoa TMĐT → TMÄT
4	hocky	Học kỳ	Số	Chỉ có 1 và 2
5	namhoc	Năm học	Date, Số	2013-2020, không nhiễu
6	drl	Điểm rèn luyện	Số	Số có định dạng xxx hoặc 'xxx' Max = 100 Có số âm -23 -13 -13 -2
7	ghichu	ghi chú có thể dùng để phân biệt mã lớp .1.2	String, null	NULL chiếm đa số Ngoài ra có có các dữ liệu khác: ", NULL), ") 'MMCL2019.2') 'KHMT0001') 'KHDL2019')

Bảng 1.10 Mô tả bảng BAOLUU và các thuộc tính

Bảng BAOLUU				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	masv	Mã sinh viên	Chuỗi	Chuỗi 40 ký tự
2	tinhtrang	Tình trạng	Số	Tình trạng 3

3	lydo	Lý do	Chuỗi	Gồm 5 loại: TN *(năm 2016)* Tá»‘t nghiá»‡p *(năm 2020)* Chá»©ng chá»‰ Anh vÄfn khÃ´ng Ä‘áº¡t --> CNTN tá»‘t nghiá»‡p chuyá»fn sang há»‡ CQÄ□T *(năm 2017)* QÄ□ Ä‘iá»□u chá»‰nh ngÃ nh TN *(năm 2017)* Null
4	hocky	Học kỳ	Số	Có ba học kỳ: 1, 2 và 3
5	namhoc	Năm học	Số	Từ năm 2016 đến năm 2020
6	soqd	Số quyết định	Chuỗi	Có 2 dạng biểu diễn: Chỉ bao gồm Số QĐ: 236, 258, và 306 *(năm 2018)* Chuỗi: (Số QĐ)/QÄ-ÄHCNTT hoặc (Số QĐ)/QD-DHCNTT
7	ngayqd	Ngày quyết định	Date	Định dạng yyyy-mm-dd

Bảng 1.11 Mô tả bảng TOTNGHIEP và các thuộc tính

Bảng TOTNGHIEP				
STT	**Thuộc tính**	**Nội dung**	**Kiểu dữ liệu**	**Diễn giải**
1	masv	Mã sinh viên	Chuỗi	Chuỗi 40 ký tự
2	xeploai	Xếp loại	Chuỗi	Có 4 loại: Giỏi Khá TB Khá hoặc Trung bình khá Xuất sắc
3	soqd	Số quyết định	Chuỗi	Có dạng (Số QĐ)/QÄ-ÄHCNTT hoặc (Số QĐ)/QD-DHCNTT

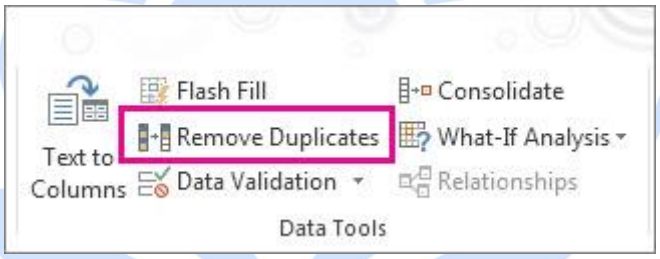
7	ngaycapvb	Ngày cấp văn bằng	Date	Định dạng dd/mm/yyyy hoặc dd-mm-yyyy

2. Làm sạch dữ liệu

2.1. Xóa dữ liệu trùng lặp và bất thường

Để xóa các dữ liệu trùng lặp trên Excel, có thể sử dụng tính năng "Remove Duplicates". Các bước thực hiện như sau:

1. Chọn dữ liệu cần xóa trùng lặp. Nếu dữ liệu của bạn có tiêu đề, hãy đảm bảo chọn cả tiêu đề.
2. Vào tab "Data" trên thanh menu và chọn "Remove Duplicates".
3. Trong hộp thoại "Remove Duplicates", chọn các cột dữ liệu mà ta muốn kiểm tra trùng lặp và bỏ chọn các cột muốn giữ nguyên.
4. Nhấn "OK" để xóa các giá trị trùng lặp.



Hình 2.1 Minh họa cho tính năng 'Remove Duplicates' của Microsoft Excel

2.2. Điều chỉnh sự không nhất quán

Sự không nhất quán trong dữ liệu (inconsistency in data) là hiện tượng dữ liệu có sự mâu thuẫn, không đồng nhất hoặc không nhất quán trong cách sắp xếp, mô tả hoặc phân loại các thông tin. Điều này có thể dẫn đến kết quả sai lệch hoặc khó khăn trong việc tìm ra mô hình hoặc kết luận chính xác từ dữ liệu.

Ví dụ: ở bảng 04.xeploaiav, ta có các cột '*id*', '*masv*', '*listening*', '*reading*', '*total*' và '*ghichu*'. Trong đó cột '*total*' được tính dựa trên tổng của các cột '*listening*' và '*reading*', nhưng trong đó cũng có các mẫu dữ liệu của cột '*total*' mà tại đó giá trị của chúng không phải là kết quả từ tổng của cột '*listening*' và '*reading*'.

3470	3509	C865FEF8X	31	62	505	ENG05	Qui Ä‘á»•i TOEIC
3471	3510	C0AE2077X	40	53	505	ENG05	Qui Ä‘á»•i TOEIC
3472	3511	3FD1462DX	37	56	505	Miá»…nEN	Qui Ä‘á»•i TOEIC
3473	3512	3D3E151DX	35	58	505	ENG05	Qui Ä‘á»•i TOEIC
3474	3513	918BCF59X	39	54	505	Miá»…nEN	Qui Ä‘á»•i TOEIC
3475	3514	21E88E8EX	36	57	505	Miá»…nEN	Qui Ä‘á»•i TOEIC
3476	3515	C4B4869CX	35	57	500	ENG05	Qui Ä‘á»•i TOEIC
3477	3516	B72FC904X	38	54	500	ENG05	Qui Ä‘á»•i TOEIC
3478	3517	16424E36X	41	51	500	ENG05	Qui Ä‘á»•i TOEIC

Hình 2.2 Minh họa cho sự thiếu nhất quán trong dữ liệu

2.3. Chuẩn hóa cấu trúc và định dạng dữ liệu

Trong khai thác dữ liệu, chuẩn hóa cấu trúc (data normalization) là quá trình biến đổi dữ liệu từ định dạng ban đầu thành định dạng chuẩn để phân tích và xử lý dữ liệu dễ dàng hơn. Quá trình này bao gồm việc điều chỉnh, biến đổi và tổng hợp dữ liệu để đưa chúng về dạng chuẩn, ví dụ như chuẩn hóa đơn vị đo lường hoặc chuẩn hóa định dạng số.

Mục đích của việc chuẩn hóa cấu trúc là loại bỏ sự khác biệt không cần thiết giữa các đặc trưng trong dữ liệu, giúp cho việc phân tích và xử lý dữ liệu trở nên dễ dàng và chính xác hơn. Việc chuẩn hóa cấu trúc cũng giúp tăng cường tính nhất quán và độ tin cậy của dữ liệu và giảm thiểu sự ảnh hưởng của dữ liệu nhiễu.

Ví dụ: cột '*lopsh*' trong file 01.sinhvien có các giá trị chuỗi mà tại đó chuỗi bắt đầu bằng khoảng trống ' '. Để chuẩn hóa cột dữ liệu này cần xóa các khoảng trắng đó.

KTPM0001

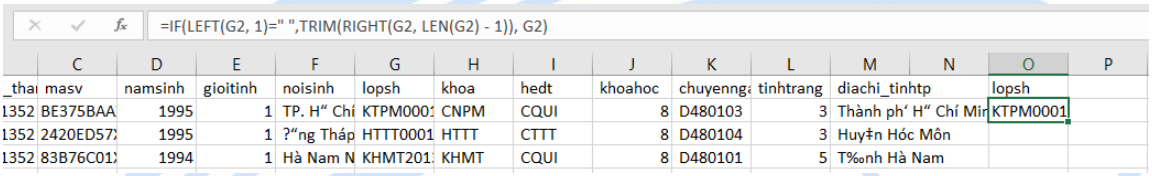
	C	D	E	F	G	H	I	J	K	L	M	N	O
_tha	masv	namsinh	gioitinh	noisinh	lopsh	khoa	hedt	khoahoc	chuyennga	tinhtrang	diachi_tinhtp		
352	BE375BAA	1995	1	TP. H“ Ch	KTPM0001	CNPM	CQUI	8	D480103	3	Thành ph‘ H“ Chí Minh		
352	2420ED57X	1995	1	?“ng Tháp	HTTT0001	HTTT	CTTT	8	D480104	3	Huy‡n Hóc Môn		
352	83B76C01X	1994	1	Hà Nam N	KHMT201	KHMT	CQUI	8	D480101	5	T‰nh Hà Nam		
352	91F785ABX	1995	1	TP. H“ Chí	HTTT0001	HTTT	CTTT	8	D480104	3	Thành ph‘ H“ Chí Minh		
352	007C275DX	1995	1	Thành ph‘	MMTT000	MMT&TT	CQUI	8	D480201	8	Thành ph‘ H“ Chí Minh		

Hình 2.3 Minh họa cho dữ liệu chưa được chuẩn hóa

Để xóa các khoảng trắng ở vị trí đầu tiên trong chuỗi dữ liệu trong Excel, bạn có thể sử dụng hàm TRIM kết hợp với hàm IF.

Với dữ liệu ta cần xóa nằm trong cột F, bắt đầu từ G2. Trong ô O2, nhập công thức sau: =IF(LEFT(G2, 1)=" ",TRIM(RIGHT(G2, LEN(G2) - 1)), G2) Sao chép công thức trong ô O2 và dán nó vào các ô phía dưới trong cột O để áp dụng công thức cho toàn bộ cột.

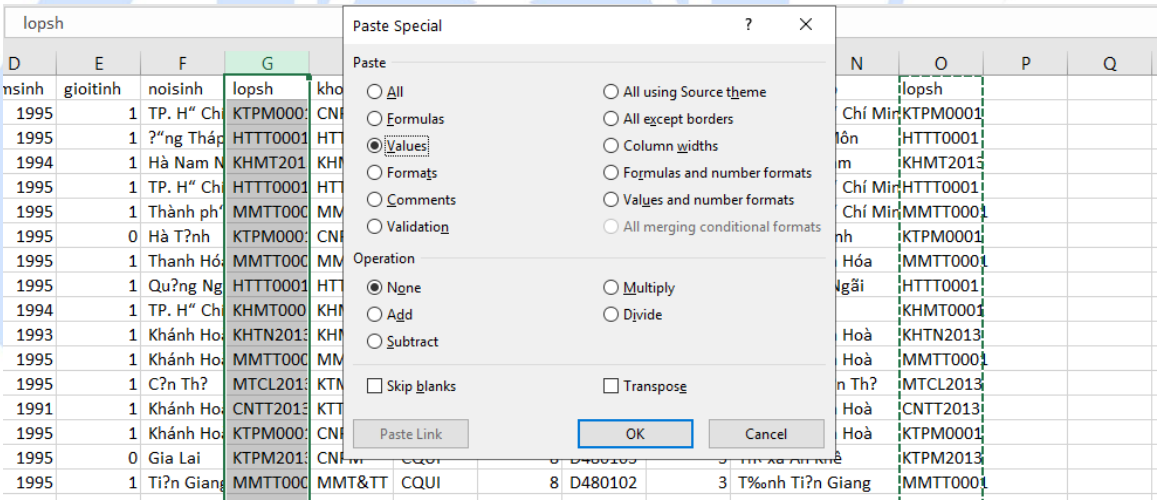
Công thức này kiểm tra xem ký tự đầu tiên của văn bản trong ô G2 có phải là một khoảng trắng hay không. Nếu là, nó sẽ xóa khoảng trắng bằng cách sử dụng hàm TRIM, và nếu không, nó sẽ trả về văn bản ban đầu. Các hàm LEFT, RIGHT và LEN được sử dụng để thao tác chuỗi văn bản.



_thai	masv	namsinh	gioitinh	noisinh	lopsh	khoa	hedt	khoahoc	chuyennga	tinhtrang	diachi_tinhtp	lopsh
1352	BE375BAA	1995	1	TP. H“ Chí	KTPM0001	CNPM	CQUI	8	D480103	3	Thành ph‘ H“ Chí Min	KTPM0001
1352	2420ED57	1995	1	?“ng Tháp	HTTT0001	HTTT	CTTT	8	D480104	3	Huy‡n Hóc Môn	
1352	83B76C01	1994	1	Hà Nam N	KHMT201	KHMT	CQUI	8	D480101	5	T‰nh Hà Nam	

Hình 2.4 Cách sử dụng hàm TRIM để chuẩn hóa dữ liệu

Kết quả ta được một mới với giá trị đã được chuẩn hóa từ cột *'lopsh'*. Để chuyển đổi giá trị từ cột O sang cột G, ta chỉ cần copy giá trị từ cột O, sau đó paste lên cột G bằng tính năng *Paste Special*, tích chọn *Values*.



Hình 2.5 Copy + Paste giá trị đã được chuẩn hóa về lại vị trí cũ bằng tính năng Paste Special

Kết quả ta có được cột giá trị *'lopsh'* đã được chuẩn hóa.

2.4. Xử lý khoảng trắng và các lỗi cú pháp khác

2.4.1. Xử lý khoảng trắng

Khi làm việc với dữ liệu trong Excel, thường xuyên chúng ta gặp phải các trường hợp có khoảng trắng không mong muốn, gây ra những khó khăn trong việc sử lý và phân tích dữ liệu. Sau đây là một số cách xử lý khoảng trắng trong Excel:

1. Xóa khoảng trắng đầu và cuối chuỗi dữ liệu: Sử dụng hàm TRIM. Ví dụ, nếu chuỗi dữ liệu cần xử lý là " Example ", ta sẽ nhập hàm "=TRIM(A1)" vào một ô khác, kết quả trả về sẽ là "Example".
2. Xóa các khoảng trắng trong chuỗi dữ liệu: Sử dụng hàm SUBSTITUTE. Ví dụ, nếu chuỗi dữ liệu cần xử lý là "Example data", ta sẽ nhập hàm "=SUBSTITUTE(A1," ","")" vào một ô khác, kết quả trả về sẽ là "Exampledata".
3. Xóa các khoảng trắng trong một cột dữ liệu: Sử dụng tính năng Find and Replace. Chọn cột dữ liệu cần xử lý, nhấn tổ hợp phím Ctrl + H, nhập " " vào ô Find what, nhập "" vào ô Replace with, nhấn Replace All.
4. Xóa các khoảng trắng ở đầu chuỗi trong một cột dữ liệu: Sử dụng hàm IF và LEFT. Ví dụ, nếu cột dữ liệu cần xử lý là A, ta sẽ nhập hàm "=IF(LEFT(A1,1)=" ", RIGHT(A1,LEN(A1)-1), A1)" vào ô B1, kết quả trả về sẽ là cột dữ liệu A đã được xử lý khoảng trắng đầu chuỗi.

Ngoài các khoảng trắng nằm bất định khiến cho việc xử lý dữ liệu trở nên khó khăn, bộ dữ liệu còn có các dòng trống. Để xử lý các dòng trống trong Excel, bạn có thể sử dụng các bước sau:

1. Chọn toàn bộ dữ liệu trong bảng tính Excel bằng cách nhấp vào nút góc trên cùng bên trái của bảng tính, bên cạnh tên cột A và số hàng 1.
2. Chọn tab "Home" trên thanh công cụ.
3. Chọn nút "Find & Select" trong nhóm "Editing".
4. Chọn "Go To Special".

5. Trong hộp thoại "Go To Special", chọn "Blanks" và nhấn "OK".
6. Những dòng trống trong bảng tính sẽ được chọn. Nhấn nút delete hoặc nhấn phím F5 và chọn "Special", sau đó chọn "Blanks" và nhấn "OK" để xóa các dòng trống đó.

2.4.2. Các lỗi cú pháp khác

Ngoài các lỗi đã kể trên, bộ dữ liệu còn có rất nhiều file có chuỗi bị mã hóa thành nhiều bộ kỳ tự.

Ví dụ: file 04.xeploaiav có các dòng đã bị mã hóa bằng bộ mã 1252: Western European.

	A	B	C	D	E	F	G	H
1	3456	47574797›	41	60	545	ENG05	Qui Ä‘á»•i	TOEIC
2	3457	E7D078A4‡	40	61	545	Miá»…nEN	Qui Ä‘á»•i	TOEIC
3	3458	B3990246›	40	60	540	Miá»…nEN	Qui Ä‘á»•i	TOEIC
4	3459	04CD33EB‡	43	57	540	ENG05	Qui Ä‘á»•i	TOEIC
5	3460	A28D2C72‡	36	64	540	Miá»…nEN	Qui Ä‘á»•i	TOEIC
6	3461	379C0BF8›	38	62	540	Miá»…nEN	Qui Ä‘á»•i	TOEIC
7	3462	B688B7AB‡	36	64	540	ENG05	Qui Ä‘á»•i	TOEIC
8	3463	1B9C9DEF‡	36	63	535	ENG05	Qui Ä‘á»•i	TOEIC
9	3464	43641E61›	40	59	535	ENG05	Qui Ä‘á»•i	TOEIC
10	3465	47E0992A›	42	57	535	Miá»…nEN	Qui Ä‘á»•i	TOEIC
11	3466	40F48AAD‡	38	61	535	ENG05	Qui Ä‘á»•i	TOEIC
12	3467	4C51140A‡	39	60	535	Miá»…nEN	Qui Ä‘á»•i	TOEIC
13	3468	AD2D3287	44	55	535	Miá»…nEN	Qui Ä‘á»•i	TOEIC
14	3469	E60244AA‡	41	58	535	Miá»…nEN	Qui Ä‘á»•i	TOEIC
15	3470	0E1B821F›	34	64	530	Miá»…nEN	Qui Ä‘á»•i	TOEIC

Hình 2.6 Minh họa cho bộ dữ liệu bị mã hóa

Để xử lý ta có thể sử dụng tính năng Load Data from Text/CSV của Microsoft Excel. Sau đó chọn bộ ký tự 65001: Unicode (UTF-8) và chọn Load. File mới của ta sẽ được load dưới bộ mã Unicode và dưới dạng file .xlsx (Excel Workbook).

3442	3481	BF736238XPvAibaEXe+KaiYHkmXYCbzr1mG14ddo	41	56	525	MiễnENG03	Qui đổi TOEIC
3443	3482	7DE869A0XPvAibaEXe8mAGU8TlDzsvC0KaxQL2ug	33	64	525	MiễnENG03	Qui đổi TOEIC
3444	3483	AE6F90F2XPvAibaEXe9Hf/fJ8gF3WoHWTRJwGHdY	35	62	525	MiễnENG03	Qui đổi TOEIC
3445	3484	B5453DB8XPvAibaEXe+jRlsyXvmhz3fdu48G7zN3	39	58	525	ENG05	Qui đổi TOEIC
3446	3485	4C5558ACXPvAibaEXe9urngX21O5XV6BAtnm2j34	38	58	520	ENG05	Qui đổi TOEIC
3447	3486	933A6450XPvAibaEXe8JDP83oALuETBHivu8sRom	37	59	520	ENG05	Qui đổi TOEIC
3448	3487	0330BA3CXPvAibaEXe8Z5hiNBvM0RW96I9r3cKzN	40	56	520	MiễnENG03	Qui đổi TOEIC
3449	3488	A67F245CXPvAibaEXe9yRt5bXcOfstXiEpKBa8kE	42	54	520	ENG05	Qui đổi TOEIC
3450	3489	8A03E471XPvAibaEXe9drMWeKYE4jEX0sTSaswVe	40	56	520	ENG05	Qui đổi TOEIC
3451	3490	E2FF3F25XPvAibaEXe8+b4756rik/nUx75UVX7mY	39	57	520	MiễnENG03	Qui đổi TOEIC
3452	3491	6CCFF9D3XPvAibaEXe/KP8Z+xE/sYxf47ovvLcr7	37	59	520	MiễnENG03	Qui đổi TOEIC

Hình 2.7 Bộ

Hình 2.7: Dữ liệu sau khi chuyển về bộ mã Unicode (UTF-8)

3. Phân tích dữ liệu

Đề tài: Dự đoán thời hạn và xếp loại tốt nghiệp của sinh viên

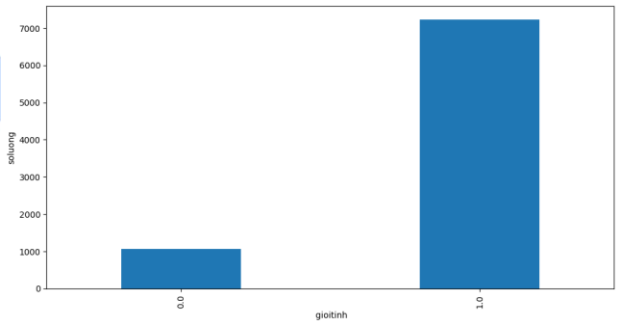
3.1. Câu hỏi nghiên cứu

(1) Các yếu tố trình độ đầu vào, điểm các môn học và điểm rèn luyện ảnh hưởng thế nào đến thời hạn và xếp loại tốt nghiệp?

(2) Việc bảo lưu và nộp học phí đúng hạn ảnh hưởng thế nào đến kết quả tốt nghiệp của sinh viên?

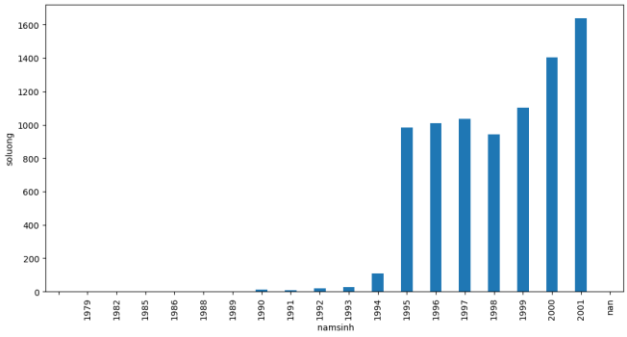
3.2. Phân tích dữ liệu

3.2.1. Bảng SINHVIEN



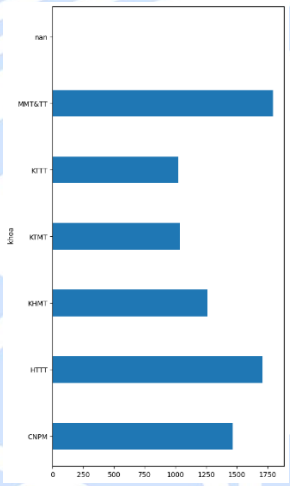
Hình 3.1 Thống kê giới tính sinh viên

Với giá trị 0 là nữ và 1 là nam, chênh lệch tỷ lệ giới tính của bộ dữ liệu là rất lớn.



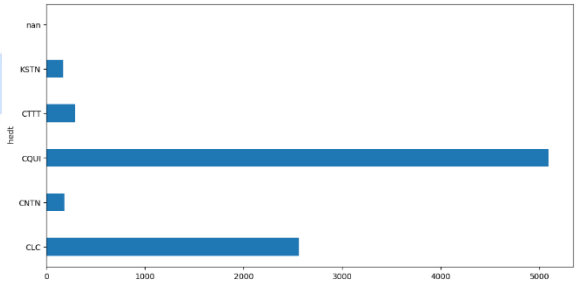
Hình 3.2 Thống kê số lượng sinh viên theo năm sinh

Phân bố chủ yếu ở sinh viên sinh năm 1995 đến 2001.



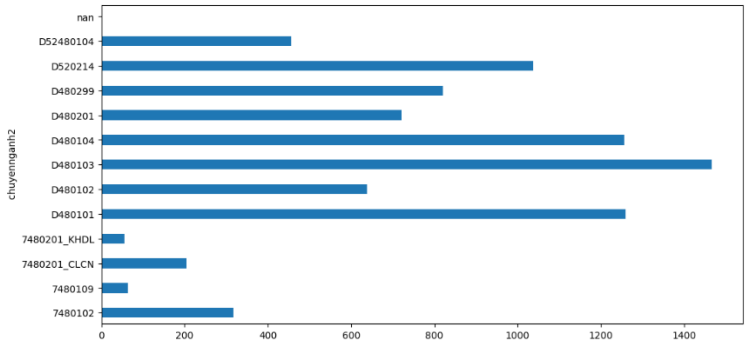
Hình 3.3 Thống kê số lượng sinh viên theo khoa

Số lượng sinh viên giữa các khoa nhìn chung không có sự chênh lệch quá lớn. Ba khoa có số lượng sinh viên nhiều nhất lần lượt là MMT&TT, HTTT và CNPM.

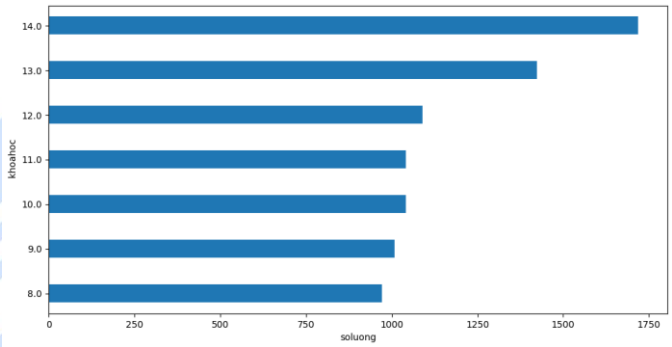


Hình 3.4 Thống kê số lượng sinh viên theo hệ đào tạo

Bộ dữ liệu có sinh viên chủ yếu thuộc hệ đào tạo CQUI và CLC.

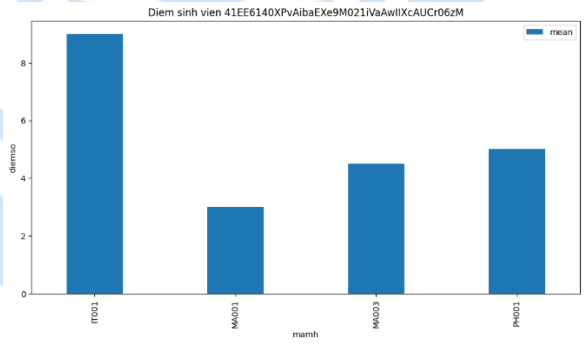


Hình 3.5 Thống kê số lượng sinh viên học chuyên ngành 2

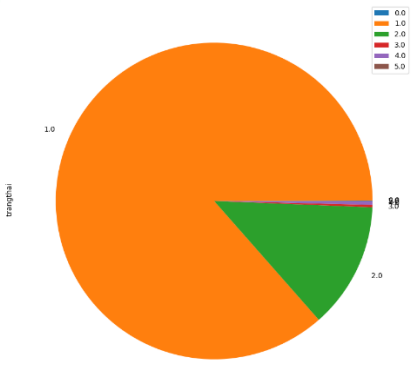


Hình 3.6 Thống kê số lượng sinh viên theo khóa học

3.2.2. Bảng DIEM

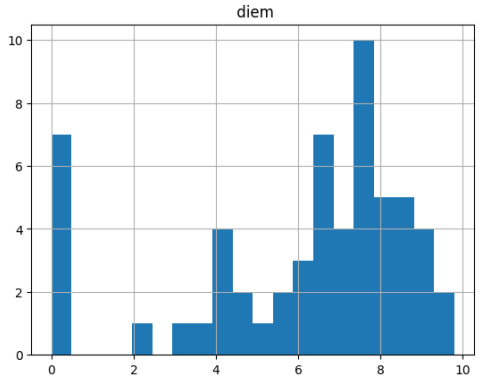


Hình 3.7 Điểm các môn học của một sinh viên



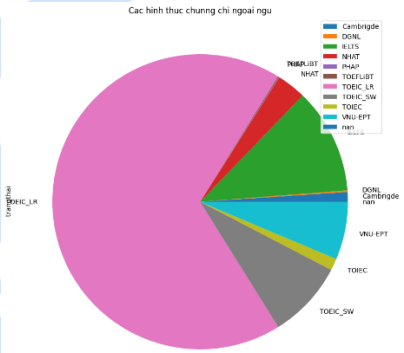
Hình 3.8 Thống kê trạng thái các môn học của sinh viên

Trạng thái môn học ‘Bình thường’ chiếm đa số, kế đến là trạng thái ‘Trả nợ’.



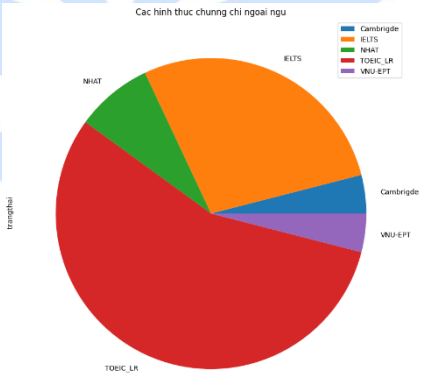
Hình 3.9 Phân phối điểm của môn IT001, năm 2015

3.2.3. Bảng CHUNGCHI



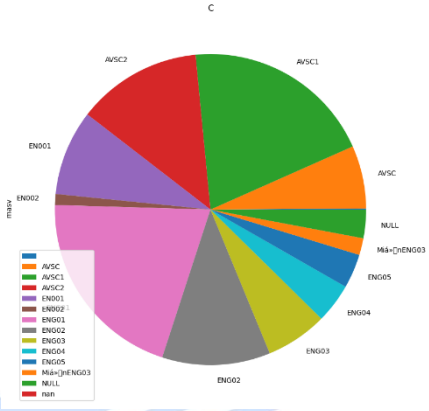
Hình 3.10 Thống kê các chứng chỉ ngoại ngữ của sinh viên

TOEIC_LR chiếm phần lớn tỷ lệ, kế đến là IELTS và TOEIC_SW.



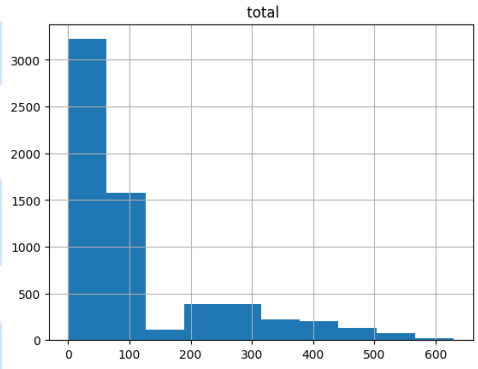
Hình 3.11 Thống kê chứng chỉ ngoại ngữ của sinh viên vào năm 2016

3.2.4. Bảng XEPLOAIAV



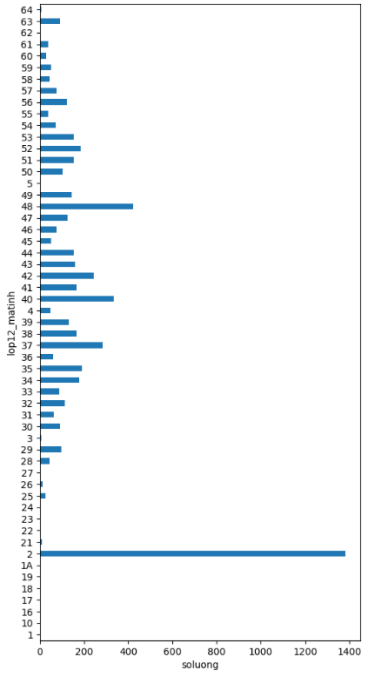
Hình 3.12 Thống kê phân lớp Anh Văn đầu vào của sinh viên

Phân lớp chiếm ưu thế là ENG01, AVSC1 và AVSC2.

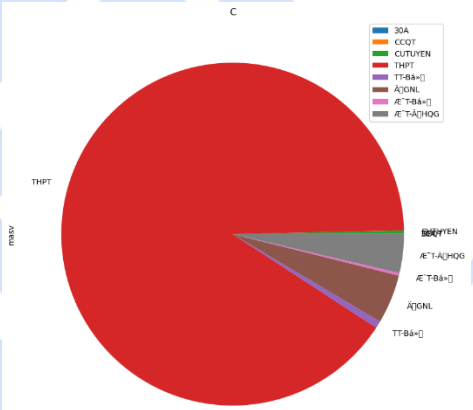


Hình 3.13 Phân bố điểm thi Anh Văn đầu vào

3.2.5. Bảng THISINH

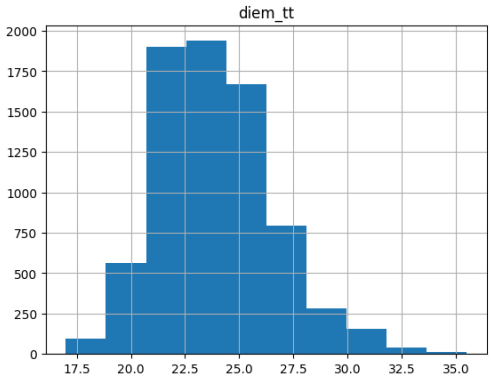


Hình 3.14 Phân bố thí sinh theo tỉnh

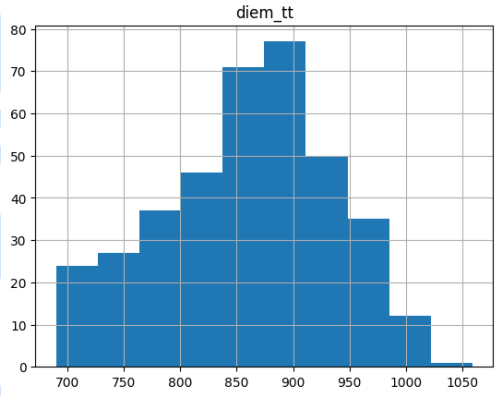


Hình 3.15 Thống kê diện tốt nghiệp của thí sinh

Thí sinh tốt nghiệp theo diện THPT chiếm đa số, hai diện khác có số lượng đáng kể là ĐGNL và ƯT-ĐHQG.

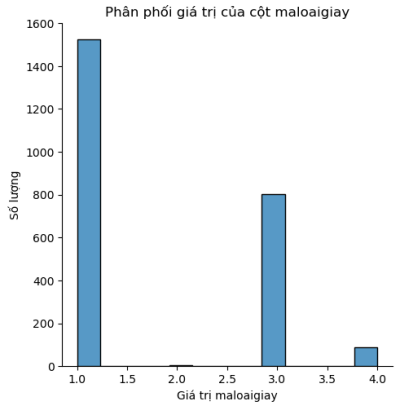


Hình 3.16 Phân phối điểm thi diện THPT



Hình 3.17 Phân phối điểm thi diện ĐGNL

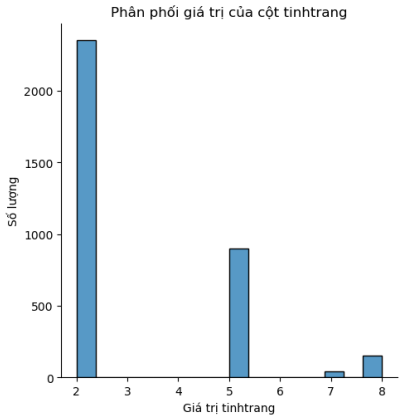
3.2.6. Bảng GIAYXACNHAN



Hình 3.18 Phân phối loại giấy

Loại giấy được yêu cầu xác nhận nhiều nhất là Giấy xác nhận sinh viên, tiếp theo là Giấy vay vốn ngân hàng và Giấy miễn giảm học phí.

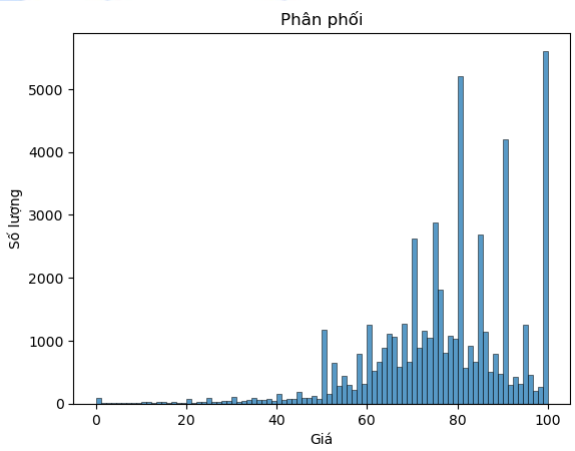
3.2.7. Bảng XLHV



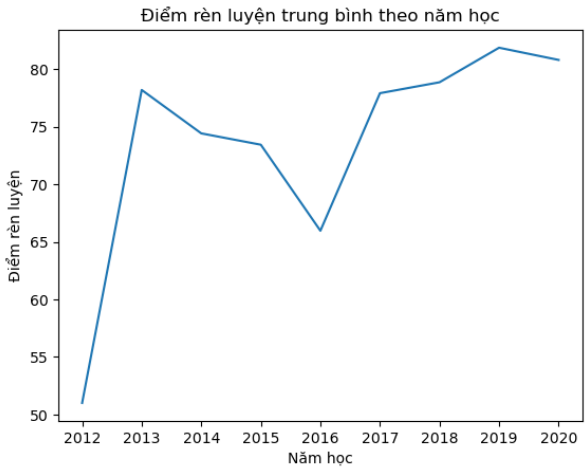
Hình 3.19 Phân phối tình trạng xử lý học vụ của sinh viên

Tình trạng xử lý học vụ nhiều nhất là Cảnh cáo, kế đến là Gia hạn.

3.2.8. Bảng DIEMRL

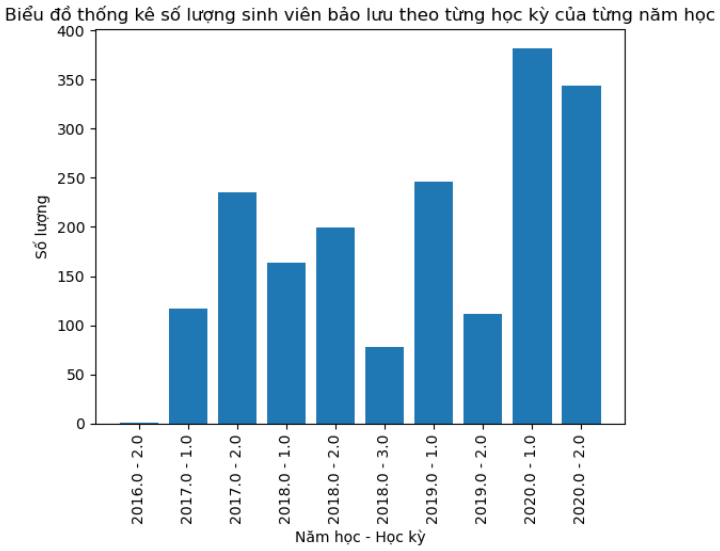


Hình 3.20 Phân phối điểm rèn luyện



Hình 3.21 Điểm rèn luyện trung bình theo năm học

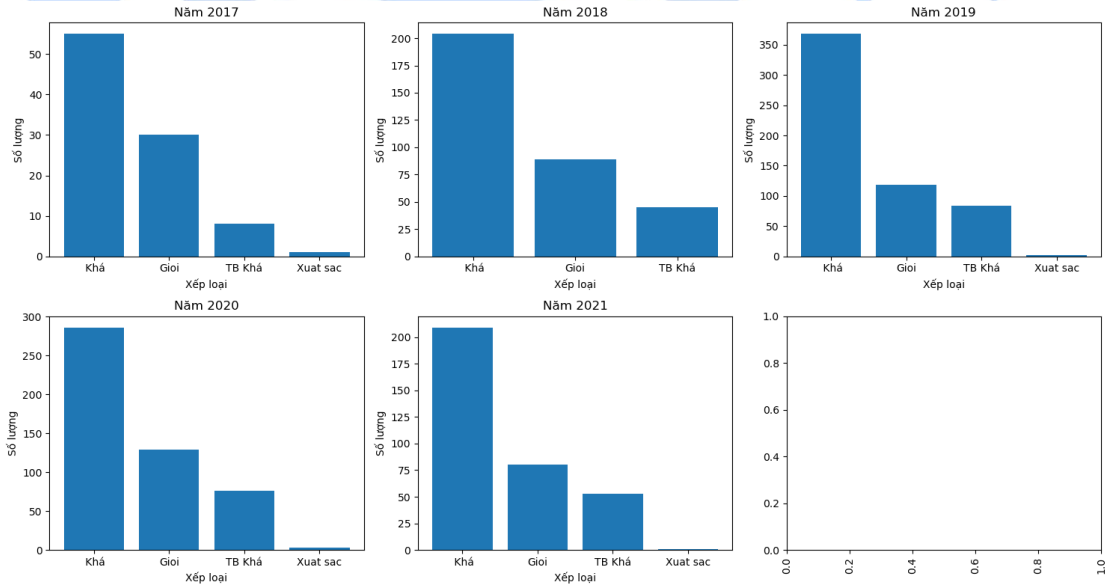
3.2.9. Bảng BAOLUU



Hình 3.22 Thống kê số lượng sinh viên bảo lưu theo từng học kỳ của từng năm học

Năm 2020 có số lượng sinh viên bảo lưu nhiều nhất.

3.2.10. Bảng TOTNGHIEP



Hình 3.23 Thống kê loại tốt nghiệp của sinh viên theo từng năm