**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙡🏵🏵🏵🙣**

A blue circle with text

Description automatically generated

**BÁO CÁO**

**MÔN PHÂN TÍCH DỮ LIỆU**

**TÊN ĐỀ TÀI**

**PHÂN TÍCH DỮ LIỆU 5 BỘ DỮ LIỆU VÀ NGHIÊN CỨU KỸ THUẬT NUMPY NÂNG CAO**

Giảng viên hướng dẫn: TS. Đỗ Như Tài

Danh sách thành viên: Nguyễn Trương Hiệp - 3122410110 Nguyễn Văn Minh - 3122410242

Vũ Thị Thanh Ngân - 3122410255

Trương Xuân Hưng - 312241016

# MỤC LỤC

[**LỜI NÓI ĐẦU 7**](#_Toc212869959)

[**PHÂN CÔNG CÔNG VIỆC** 8](#_Toc212869960)

[**PHẦN 1: PHÂN TÍCH CÁC DATASETS** 1](#_Toc212869961)

[I. DỮ LIỆU BITLY TỪ 1.USA.GOV 1](#_Toc212869962)

[1. Giới thiệu, mục tiêu 1](#_Toc212869963)

[2. Quy trình làm sạch và tiền xử lý dữ liệu 1](#_Toc212869964)

[2.1. Đọc dữ liệu và xử lý độ đầy đủ 1](#_Toc212869965)

[2.2. Trích xuất kỹ thuật đặc trưng 2](#_Toc212869966)

[2.3. Chuẩn hóa dữ liệu thời gian 2](#_Toc212869967)

[3. Phân tích Khám phá dữ liệu 2](#_Toc212869968)

[3.1. Phân tích Múi giờ (tz) có số lượt truy cập cao nhất 2](#_Toc212869969)

[3.2. Phân tích Người dùng thiếu thông tin múi giờ hoặc quốc gia 3](#_Toc212869970)

[3.3. Phân tích Những nguồn referrer (r) phổ biến nhất 4](#_Toc212869971)

[4. Phân tích Tương quan Chuyên sâu 4](#_Toc212869972)

[4.1. Thiết bị: Tỷ lệ người dùng Windows vs Non-Windows 5](#_Toc212869973)

[4.2. Thiết bị: Phân bố hệ điều hành khác nhau theo múi giờ (East Coast vs West Coast) 5](#_Toc212869974)

[4.3. Khung hành vi thời gian: Khung giờ (t) nào có lượng truy cập cao nhất? 6](#_Toc212869975)

[4.4. Khung hành vi thời gian: Có mối tương quan giữa thời điểm tạo link (hc) và thời điểm click (t) không? 6](#_Toc212869976)

[4.5. Địa lý: Những bang (gr) hoặc thành phố (cy) nào có lượng truy cập lớn nhất? 7](#_Toc212869977)

[4.6. Địa lý: Người dùng từ ngoài nước Mỹ (c != 'US') chiếm bao nhiêu phần trăm? 8](#_Toc212869978)

[4.7. Nguồn truy cập: Có sự khác biệt referrer theo hệ điều hành không? 9](#_Toc212869979)

[5. Kết luận 9](#_Toc212869980)

[5.1. Tóm tắt Câu chuyện Dữ liệu 9](#_Toc212869981)

[5.2. Đối tượng Mục tiêu và Hành động 10](#_Toc212869982)

[II. BỘ DỮ LIỆU MOVIELENS 1M 11](#_Toc212869983)

[1. Giới thiệu, mục tiêu 11](#_Toc212869984)

[2. Quy trình làm sạch và tiền xử lý dữ liệu 11](#_Toc212869985)

[2.1. Đọc dữ liệu và xử lý độ đầy đủ 11](#_Toc212869986)

[2.2. Trích xuất và chuẩn hóa đặc trưng 12](#_Toc212869987)

[2.3. Kết hợp bảng dữ liệu 12](#_Toc212869988)

[3. Thống kê mô tả 12](#_Toc212869989)

[3.1. Các phim có độ lệch chuẩn rating cao nhất 12](#_Toc212869990)

[3.2. Phân bố độ lệch chuẩn rating và nhóm phim “gây tranh cãi” 13](#_Toc212869991)

[3.3. Độ lệch chuẩn rating theo thể loại phim và nhóm tuổi 16](#_Toc212869992)

[4. Phân tích khám phá 18](#_Toc212869993)

[4.1. Yếu tố ảnh hưởng đến độ phân tán rating 18](#_Toc212869994)

[4.2. Thể loại phim cực đoan vs thể loại chính thống 19](#_Toc212869995)

[4.2.1. Quan sát từ bảng và biểu đồ 19](#_Toc212869996)

[4.2.2. Kết quả kiểm định t-test 21](#_Toc212869997)

[4.2.3. Kết luận chung 21](#_Toc212869998)

[4.3. Mối tương quan: Độ phổ biến vs Mức độ đồng thuận 21](#_Toc212869999)

[5. Phân tích dự đoán 23](#_Toc212870000)

[5.1. Xây dựng mô hình dự đoán độ lệch chuẩn rating 23](#_Toc212870001)

[5.2. Độ lệch chuẩn cao như chỉ báo “controversial” 24](#_Toc212870002)

[6. Phân tích so sánh 26](#_Toc212870003)

[6.1. Bất đồng ý kiến giữa giới tính 26](#_Toc212870004)

[6.2. Độ lệch chuẩn rating theo thập niên 27](#_Toc212870005)

[6.3. Hành vi đánh giá của nhóm tuổi trẻ vs lớn tuổi 28](#_Toc212870006)

[7. Phân tích mở rộng 30](#_Toc212870007)

[7.1. Phân cực cảm xúc trong khán giả 30](#_Toc212870008)

[7.2. Gợi ý “Khám Phá” (Serendipity Recommendation) 32](#_Toc212870009)

[7.3. Khả năng lan tỏa truyền thông 33](#_Toc212870010)

[8. Kết luận 33](#_Toc212870011)

[III. Tên trẻ em ở Mỹ 1880-2010 35](#_Toc212870012)

[1. Giới thiệu và Mô tả Dữ liệu 35](#_Toc212870013)

[1.1. Mục tiêu nghiên cứu: 35](#_Toc212870014)

[1.2. Câu hỏi nghiên cứu: 35](#_Toc212870015)

[1.3. Định dạng dữ liệu 36](#_Toc212870016)

[2. Bức tranh toàn cảnh (Phân tích Mô tả) 38](#_Toc212870017)

[3. Sự bùng nổ của "Tính độc đáo" (Phân tích Khám phá) 39](#_Toc212870018)

[4. Cuộc chiến ngầm của Giới tính và Vòng đời (Phân tích So sánh) 41](#_Toc212870019)

[5. Đi tìm "Công thức" cho một cái tên Hot (Phân tích Dự đoán & Mô hình) 43](#_Toc212870020)

[IV. Cơ sở dữ liệu Thực phẩm USDA 46](#_Toc212870021)

[1. Giới thiệu, mục tiêu 46](#_Toc212870022)

[2. Quy trình làm sạch dữ liệu 46](#_Toc212870023)

[2.1. Đọc Dữ Liệu và Xử Lý Cấu Trúc JSON 46](#_Toc212870024)

[2.2. Trích Xuất và Chuẩn Hóa Đặc Trưng 46](#_Toc212870025)

[2.3. Kết Hợp Bảng Dữ Liệu 47](#_Toc212870026)

[3. Thống kê mô tả - Khám phá dinh dưỡng 47](#_Toc212870027)

[3.1. Năng Lượng Trung Bình theo Nhóm Thực Phẩm 47](#_Toc212870028)

[3.2. Loại Thực Phẩm Chứa Protein Cao Nhất 49](#_Toc212870029)

[3.3. Độ đa dạng dinh dưỡng theo nhóm thực phẩm 51](#_Toc212870030)

[4. Phân Tích So Sánh - Tương Phản Giữa Các Nhóm 52](#_Toc212870031)

[4.1. So sánh Protein và Lipid giữa nhóm “Meat” và “Dairy” 52](#_Toc212870032)

[4.2. Tỷ lệ Energy-to-Protein 53](#_Toc212870033)

[5. Phân Tích Tương Quan - Mối Liên Hệ Giữa Chất Dinh Dưỡng 54](#_Toc212870034)

[6. Phân Tích Dự Đoán - Chỉ Số Dinh Dưỡng Cân Bằng 57](#_Toc212870035)

[6.1. Xây Dựng Chỉ Số Dinh Dưỡng Cân Bằng 57](#_Toc212870036)

[7. Kết luận 59](#_Toc212870037)

[V. Case Study: Dữ liệu Bầu cử Liên bang 2012 60](#_Toc212870038)

[1. Giới thiệu, mục tiêu 60](#_Toc212870039)

[2. Quy trình làm sạch dữ liệu 60](#_Toc212870040)

[2.1. Đọc dữ liệu và xử lý độ đầy đủ 60](#_Toc212870041)

[2.2. Chuẩn hóa và trích xuất đặc trưng 60](#_Toc212870042)

[2.3. Tổng hợp và kết nối dữ liệu 61](#_Toc212870043)

[3. Phân tích mô tả 61](#_Toc212870044)

[3.1. Phân tích dữ liệu thiếu 61](#_Toc212870045)

[3.2. Phân tích Occupation (Nguồn Công Việc) - Referrer 63](#_Toc212870046)

[4. Phân tích tương quan 63](#_Toc212870047)

[4.1. Phân tích tương quan theo Đảng Phái 63](#_Toc212870048)

[4.2. Phân tích Employer theo Ứng Viên Chính 64](#_Toc212870049)

[5. Phân tích phân nhóm (Bucketing) theo giá trị 67](#_Toc212870050)

[6. Phân tích thống kê chi tiết 70](#_Toc212870051)

[7. Kết luận 71](#_Toc212870052)

[**PHẦN 2: KỸ THUẬT NUMPY NÂNG CAO** 73](#_Toc212870053)

[2.1. Mục A.1: Cấu trúc bên trong của đối tượng ndarray 73](#_Toc212870054)

[2.1.1. Các thành phần của ndarray 73](#_Toc212870055)

[2.1.2. Hệ thống phân cấp kiểu dữ liệu (dtype) 74](#_Toc212870056)

[2.2. Mục A.2: Thao tác Mảng Nâng cao 76](#_Toc212870057)

[2.2.1. Reshaping Arrays (reshape, ravel, flatten) 76](#_Toc212870058)

[2.2.2. C và FORTRAN Order 78](#_Toc212870059)

[2.2.3. Concatenating and Splitting Arrays (concatenate, split) 79](#_Toc212870060)

[2.2.4. Repeating Elements (tile, repeat) 81](#_Toc212870061)

[2.2.5. Fancy Indexing Equivalents: take and put 84](#_Toc212870062)

[2.3. Mục A.3: Broadcasting (Quy tắc Lan truyền) 84](#_Toc212870063)

[2.3.1. Broadcasting trên các trục khác 87](#_Toc212870064)

[2.3.2. Thiết lập Giá trị Mảng bằng Broadcasting 89](#_Toc212870065)

[2.4. Mục A.4: Sử dụng ufunc Nâng cao 90](#_Toc212870066)

[2.4.1. Các phương thức của ufunc (reduce, accumulate, outer, reduceat) 90](#_Toc212870067)

[2.5. Mục A.5: Mảng có Cấu trúc (Structured Arrays) 90](#_Toc212870068)

[KẾT LUẬN 92](#_Toc212870069)

**DANH MỤC HÌNH ẢNH**

[Hình I.3.1.a: Phân bố Top 10 Múi giờ Truy cập (tz) 2](#_Toc212871726)

[Hình I.3.2.a: Tỷ lệ dữ liệu thiếu 3](#_Toc212871727)

[Hình I.3.3.a: Top 10 Nguồn Referrer Phổ biến 4](#_Toc212871728)

[Hình I.4.1.a: Tỷ lệ Người dùng Windows vs Non-Windows 5](#_Toc212871729)

[Hình I.4.2.a: Phân bố Hệ điều hành theo Múi giờ Trọng yếu (%) 5](#_Toc212871730)

[Hình I.4.3.a: Lượng truy cập theo Khung giờ (UTC) 6](#_Toc212871731)

[Hình I.4.4.a: Phân bố Độ trễ Click Link (Log10 Phút) 6](#_Toc212871732)

[Hình I.4.5.a: Top 5 bang có lượng truy cập lớn nhất 7](#_Toc212871733)

[Hình I.4.5.b: Top 5 thành phố có lượng truy cập lớn 7](#_Toc212871734)

[Hình I.4.6.a: Tỷ lệ truy cập theo quốc gia 8](#_Toc212871735)

[Hình I.4.7.a: Phân bố hệ điều hành (%) theo top 5 nguồn 9](#_Toc212871736)

[Hình II.3.2.a. Biểu đồ Histogram thể hiện phần bố độ lệch chuẩn rating 13](#_Toc212871737)

[Hình II.3.2.b: Biểu đồ Boxplot thể hiện độ lệch chuẩn 14](#_Toc212871738)

[Hình II.3.2.c: Biểu đồ Q-Q Plot để kiểm tra phân phối chuẩn 14](#_Toc212871739)

[Hình II.3.2.d: Biểu đồ scatter thể hiện mối quan hệ điểm trung bình và độ phân tán 15](#_Toc212871740)

[Hình II.3.3.a: Biểu đồ cột ngang thể hiện độ lệch chuẩn rating theo thể loại phim 16](#_Toc212871741)

[Hình II.3.3.b: Biểu đồ cột thể hiện độ lệch chuẩn rating theo nhóm tuổi người xem 17](#_Toc212871742)

[Hình II.4.1.a: Ma trận tương quan giữa các yếu tố với độ phân tán rating 18](#_Toc212871743)

[Hình II.4.2.1.a: Biểu đồ hộp thể hiện phân bố độ phân tán theo danh mục thể loại 20](#_Toc212871744)

[Hình II.4.2.1.b: Biểu đồ cột thể hiện độ phân tán trung bình theo danh mục 20](#_Toc212871745)

[Hình II.4.3.a: Biểu đồ scatter thể hiện mối quan hệ độ phổ biến và mức độ đồng thuận 22](#_Toc212871746)

[Hình II.4.3.b: Biểu đồ hộp thể hiện độ phân tán theo mức độ phổ biến 22](#_Toc212871747)

[Hình II.5.1.a: Biểu đồ thể hiện kết quả dự đoán và tầm quan trọng của các đặc trưng 23](#_Toc212871748)

[Hình II.5.2.a: Các biểu đồ thể hiện độ lệch chuẩn 24](#_Toc212871749)

[Hình II.5.3.a: Biểu đồ thể hiện hiệu suất và phân bố lỗi dự đoán 25](#_Toc212871750)

[Hình II.6.1.a: Biểu đồ cột thể hiện top 15 phim có bất đồng ý kiến lớn giữa nam nữ 27](#_Toc212871751)

[Hình II.6.2.a: Biểu đồ đường thể hiện xu hướng Polarization theo thập niên 28](#_Toc212871752)

[Hình II.6.3.a: Biểu đồ cột thể hiện hành vi đánh giá cực đoan theo nhóm tuổi 29](#_Toc212871753)

[Hình II.6.3.b: Biểu đồ thể hiện phân bố rating cực đoan theo nhóm tuổi 30](#_Toc212871754)

[Hình II.7.1.a: Biểu đồ scatter plot thể hiện mối quan hệ giữa std rating và emotional polaziration 31](#_Toc212871755)

[Hình II.7.2.a: Biểu đồ cột thể hiện Top 20 Phim 32](#_Toc212871756)

[Hình II.7.3.a: Biểu đồ thể hiện khả năng lan tỏa và mối quan hệ mức độ với phổ biến 33](#_Toc212871757)

[Hình III.1.3.a: dữ liệu tên trẻ em năm 1880 36](#_Toc212871758)

[Hình III.1.3.a: Số lượng bé trai và bé gái trong năm 1880 37](#_Toc212871759)

[Hình III.1.3.c: Bộ dữ liệu tên trẻ em từ năm 1880-2010 38](#_Toc212871760)

[Hình III.2.a: Tổng sinh theo năm và giới tính 39](#_Toc212871761)

[Hình III.2.b: Số lượng người đặt tên ‘John’ và ‘Mary’ qua các năm 40](#_Toc212871762)

[Hình III.3.a: Tỉ lệ top 1000 tên thường dùng theo giới tính/năm 41](#_Toc212871763)

[Hình III.3.b: Số lượng tên trong top 50% 42](#_Toc212871764)

[Hình III.4.a: Số lượng trẻ tên ‘Jennifer’ qua các năm 43](#_Toc212871765)

[Hình III.4.b: Tỉ lệ bé gái tên ‘Leslie’ qua các năm 44](#_Toc212871766)

[Hình III.5.a: Tỷ lệ tên bé trai kết thúc bằng 'd', 'n', 'y' 45](#_Toc212871767)

[Hình III.5.b: Tỷ lệ tên bé gái kết thúc bằng 'a', 'e', 'y’ 46](#_Toc212871768)

[Hình IV.3.1.a: Biểu đồ cột thể hiện năng lượng trung bình theo nhóm thực phẩm 49](#_Toc212871769)

[Hình IV.3.1.b: Biểu đồ hộp thể hiện phân bố năng lượng theo nhóm 50](#_Toc212871770)

[Hình IV.3.2.a: Bảng thực phẩm có Protein cao nhất trong mỗi nhóm 52](#_Toc212871771)

[Hình IV.3.3.a: Biểu đồ cột ngang thể hiện độ đa dạng chất dinh dưỡng theo nhóm 53](#_Toc212871772)

[Hình IV.3.3.b: Biểu đồ cột ngang thể hiện số lượng thực phẩm theo nhóm 53](#_Toc212871773)

[Hình IV.4.2.a: Biểu đồ cột thể hiện tỷ lệ Energy to Protein 55](#_Toc212871774)

[Hình IV.5.1.a: Biểu đồ tương quan minh hoạ mối quan hệ giữa các chỉ số dinh dưỡng chính 56](#_Toc212871775)

[Hình IV.5.2.a: Bảng Top 10 Thực phẩm có ‘Empty Calories’ cao nhất 57](#_Toc212871776)

[Hình IV.5.3.a: Biểu đồ hộp thể hiện năng lượng theo mức độ nước 58](#_Toc212871777)

[Hình IV.5.3.b: Biểu đồ scatter plot thể hiện mối quan hệ Water và Energy 59](#_Toc212871778)

[Hình IV.6.1.a: Biểu đồ cột ngang thể hiện hàm lượng Protein trung bình 60](#_Toc212871779)

[Hình IV.6.1.b: Biểu đồ cột ngang thể hiện hàm lượng chất béo trung bình 60](#_Toc212871780)

[Hình V.3.1.a: Biểu đồ cột thể hiện tỷ lệ dữ liệu thiếu (%) 63](#_Toc212871781)

[Hình V.3.1.b: Biểu đồ cột thể hiện top 10 tiểu bang - số lượng quyên góp 64](#_Toc212871782)

[Hình V.3.2.a: Biểu đồ cột thể hiện Top 15 Occupation ở số lượng - tổng tiền quyên góp 65](#_Toc212871783)

[Hình V.4.1.a: Biểu đồ cột ngang thể hiện tổng quyên góp theo Occupation - phân chia theo Đảng Phái 66](#_Toc212871784)

[Hình V.4.2.a: Biểu đồ cột thể hiện sự phân bố theo nhà tuyển dụng và ứng viên 67](#_Toc212871785)

[Hình V.4.3.a: Biểu đồ thể hiện tổng tiền, số lượng quyên góp, top 12 tiểu bang 68](#_Toc212871786)

[Hình V.5.a: Biểu đồ cột thể hiện tần suất quyên góp theo nhóm giá trị 69](#_Toc212871787)

[Hình V.5.b: Biểu đồ cột thể hiện tổng tiền quyên góp theo nhóm giá trị 70](#_Toc212871788)

[Hình V.5.c: Biểu đồ cột thể hiện tỷ lệ phần trăm tiền quyên góp 70](#_Toc212871789)

[Hình V.5.d: Biểu đồ cột thể hiện tỷ lệ phần trăm 71](#_Toc212871790)

[Hình V.6.a: Biểu đồ cột thể hiện tổng tiền quyên góp - tất cả ứng viên 72](#_Toc212871791)

[Hình V.6.b: Biểu đồ cột thể hiện số lượng quyên góp - tất cả ứng viên 73](#_Toc212871792)

[Hình 2.1.1: Đối tượng ndarray trong NumPy. 76](#_Toc212871793)

[Hình 2.2.1: Định hình lại mảng theo thứ tự của C (ưu tiên hàng) hoặc FORTRAN (ưu tiên cột). 79](#_Toc212871794)

[Hình 2.3.1: Lan truyền trên trục 0 với một mảng 1D 88](#_Toc212871795)

[Hình 2.3.2: Lan truyền trên trục 1 của một mảng 2D 89](#_Toc212871796)

[Hình 2.3.3: Lan truyền trên trục 0 của một mảng 3D. 89](#_Toc212871797)

# LỜI NÓI ĐẦU

Trong bối cảnh bùng nổ dữ liệu hiện nay, kỹ năng phân tích và trích xuất thông tin giá trị từ dữ liệu thô đã trở thành một yêu cầu thiết yếu. Các công cụ lập trình, đặc biệt là thư viện NumPy trong Python, cung cấp nền tảng mạnh mẽ để thực hiện các tác vụ này một cách hiệu quả.

Nhằm mục đích thực hành và củng cố kiến thức, nhóm chúng tôi đã thực hiện dự án tái hiện và phân tích các bài toán được trình bày trong tài liệu "Lab 02: Data Analytics". Báo cáo này sẽ trình bày lại toàn bộ quá trình và kết quả đạt được.

Nội dung báo cáo được cấu trúc thành hai phần chính. Phần đầu tiên tập trung vào việc tái hiện năm case study phân tích dữ liệu thực tế từ Chương 13. Phần thứ hai đi sâu vào việc hệ thống hóa và minh họa các kỹ thuật NumPy nâng cao được trình bày trong Phụ lục A, qua đó làm rõ nền tảng kỹ thuật cho các phân tích. Thông qua việc thực thi lại toàn bộ mã nguồn và phân tích kết quả, chúng tôi mong muốn báo cáo này sẽ thể hiện được sự hiểu biết và khả năng vận dụng các công cụ phân tích dữ liệu vào các bài toán cụ thể.

# PHÂN CÔNG CÔNG VIỆC

|  |  |
| --- | --- |
| **Thành viên** | **Công việc** |
| Nguyễn Trương Hiệp - 3122410110 | + 13.3 US Baby Names 1880–2010  + A.1 ndarray Object Internals  +A.7 Writing Fast NumPy Functions with Numba  +A.9 Performance Tips: |
| Vũ Thị Thanh Ngân - 3122410255 | + 13.2 MovieLens 1M Dataset  + 13.4 USDA Food Database  + A.5 Structured and Record Arrays  + A.6 More About Sorting |
| Nguyễn Văn Minh - 3122410242 | + 13.1 Bitly Data from [1.USA.gov](http://1.usa.gov/)  + A.2 Advanced Array Manipulation  + A.3 Broadcasting: |
| Trương Xuân Hưng - 3122410161 | + 13.5 2012 Federal Election Commission Database |

# PHẦN 1: PHÂN TÍCH CÁC DATASETS

## DỮ LIỆU BITLY TỪ 1.USA.GOV

### Giới thiệu, mục tiêu

Tập dữ liệu **"Bitly Data from 1.USA.gov"** ghi lại hoạt động truy cập vào các liên kết chính phủ Mỹ (đuôi .gov hoặc .mil) được rút gọn qua dịch vụ Bitly vào năm 2011.

Về bản chất, dữ liệu này có tính chất **phi cấu trúc (semi-structured data)**, được lưu dưới dạng các dòng **JSON (JavaScript Object Notation)**. Phương pháp phân tích yêu cầu phải bao gồm kỹ thuật **đọc file theo từng dòng** và sử dụng Pandas để chuyển đổi hiệu quả dữ liệu thô này thành cấu trúc có thể thao tác được.

Mục tiêu của nghiên cứu là áp dụng toàn bộ quy trình phân tích dữ liệu — từ làm sạch, tổng hợp, đến trực quan hóa — để **chiết xuất ý nghĩa** từ dữ liệu thô, nhằm hiểu hành vi người dùng (thiết bị, địa lý, thời gian) khi tương tác với thông tin chính phủ.

### Quy trình làm sạch và tiền xử lý dữ liệu

Phần này mô tả các kỹ thuật được áp dụng để chuyển đổi dữ liệu thô (dạng JSON theo dòng) thành định dạng phù hợp cho phân tích thống kê (DataFrame).

Quy trình tiền xử lý được thực hiện thông qua ba giai đoạn chính:

#### 2.1. Đọc dữ liệu và xử lý độ đầy đủ

Dữ liệu thô được đọc bằng cách sử dụng module json và thư viện Pandas, tuân thủ kỹ thuật đọc file **JSON theo dòng** do tính chất của dữ liệu web log. Tổng số bản ghi được tải là 3440.

* **Kiểm tra Cấu trúc Ban đầu:** Cấu trúc dữ liệu ban đầu cho thấy sự tồn tại của các trường thông tin đa dạng như tz (Múi giờ), a (User Agent), và r (Referrer).
* **Xử lý Múi giờ (tz):** Các giá trị thiếu (NaN và chuỗi rỗng) đã được điền bằng "Missing" và "Unknown" để đảm bảo tính toàn vẹn.

Kết quả làm sạch xác định rằng 521 lượt truy cập không thể xác định được múi giờ. Con số này chiếm 15.15% tổng số bản ghi và được ghi nhận như là tỷ lệ dữ liệu thiếu cần thiết để đánh giá tính đầy đủ của mẫu.

#### 

#### 2.2. Trích xuất kỹ thuật đặc trưng

Giai đoạn này tập trung vào việc tạo ra các biến mới cần thiết cho phân tích tương quan:

* **Xác định Hệ điều hành (OS):** Trường Tác nhân người dùng (a) được xử lý để phân loại thành hai nhóm định tính: "Windows" và "Non-Windows". Đây là nền tảng cho mọi phân tích về thiết bị.
* **Chuẩn hóa Nguồn Referrer (r):** Trường r (nguồn giới thiệu) được chuẩn hóa về tên miền gốc (ví dụ: facebook.com, twitter.com), cho phép phân tích nhóm nguồn truy cập chính xác.

#### 2.3. Chuẩn hóa dữ liệu thời gian

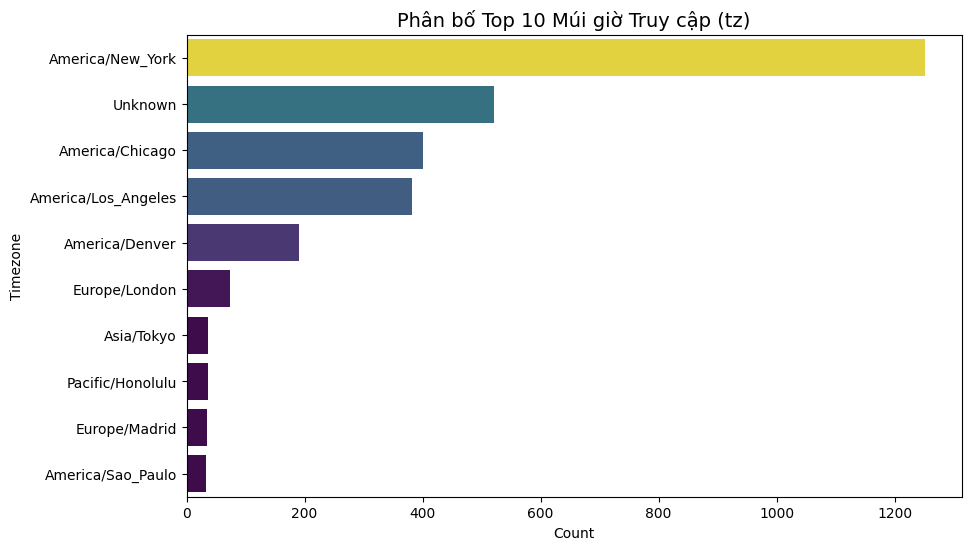
Các trường thời gian dạng timestamp (t và hc) đã được chuyển đổi từ giây (Unix timestamp) sang đối tượng datetime trong Pandas để có thể trích xuất các đơn vị thời gian quan trọng:

* **Phân tích Khung giờ:** Giờ trong ngày (0-23) được trích xuất từ trường t để xác định Khung giờ truy cập cao điểm.
* **Tính toán Độ trễ:** Độ trễ (thời gian click - thời gian tạo link) được tính bằng giây và chuyển đổi sang phút để phân tích Vòng đời của liên kết.

### Phân tích Khám phá dữ liệu

Phần này tập trung vào các kết quả mô tả cơ bản, đánh giá tính đầy đủ và phân bố của các trường dữ liệu quan trọng, đặc biệt là trường Múi giờ (tz), để thiết lập bối cảnh cho các phân tích tương quan chuyên sâu hơn.

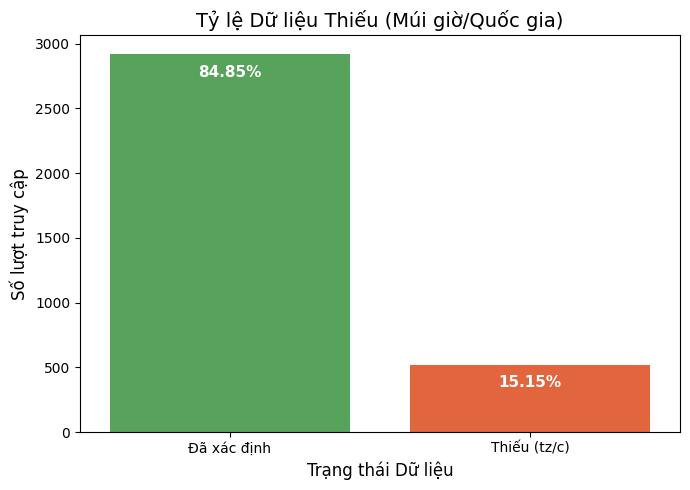
#### 3.1. Phân tích Múi giờ (tz) có số lượt truy cập cao nhất



Hình I.3.1.a: Phân bố Top 10 Múi giờ Truy cập (tz)

Phân tích ban đầu tập trung vào việc xác định trung tâm chính của hoạt động truy cập. Theo **Hình I.3.1.a**, phân tích cho thấy múi giờ **America/New\_York** là trung tâm hoạt động truy cập nổi bật, vượt trội so với các múi giờ khác. Sự tập trung lưu lượng này, được minh họa trong **Hình I.3.1.a**, cho thấy thông tin chủ yếu được tiêu thụ tại các trung tâm hành chính và kinh tế lớn của Bờ Đông Hoa Kỳ. Kết quả này củng cố giả thuyết ban đầu rằng hoạt động truy cập có xu hướng bị chi phối bởi các tổ chức hoặc cá nhân hoạt động trong múi giờ làm việc của chính phủ liên bang.

#### 3.2. Phân tích Người dùng thiếu thông tin múi giờ hoặc quốc gia

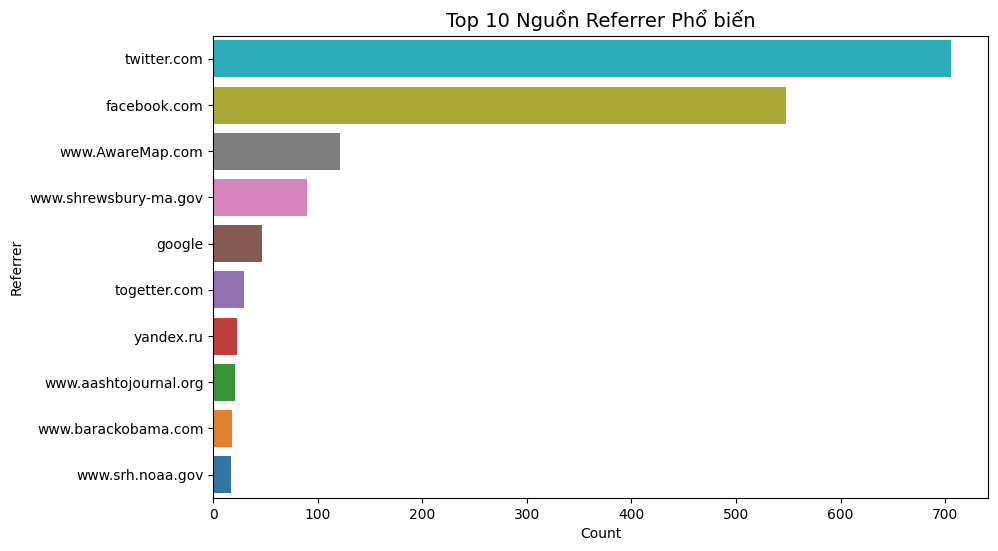


Hình I.3.2.a: Tỷ lệ dữ liệu thiếu

Phân tích này tập trung đánh giá mức độ đầy đủ và tính tin cậy của dữ liệu địa lý. Theo **Hình I.3.2.a**, kết quả thống kê cho thấy 15,15% bản ghi thiếu thông tin về múi giờ hoặc quốc gia. Tỷ lệ này phản ánh một thách thức kỹ thuật đáng kể, vì nó đại diện cho một lượng lớn dữ liệu không thể xác định chính xác nguồn gốc truy cập. Con số này đồng thời là chỉ báo rõ ràng về sự hiện diện của người dùng sử dụng dịch vụ ẩn danh (VPN/proxy) hoặc các công cụ tự động (bot) không cung cấp dữ liệu địa lý hợp lệ. Phần dữ liệu này cũng được thể hiện trong **Hình I.3.1.a** với danh mục “Unknown”.

#### 

#### 3.3. Phân tích Những nguồn referrer (r) phổ biến nhất



Hình I.3.3.a: Top 10 Nguồn Referrer Phổ biến

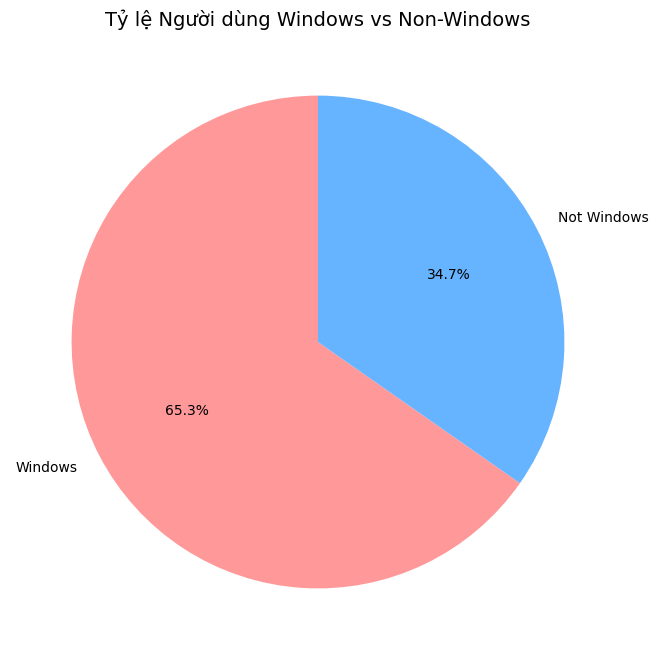
Phân tích này tập trung vào các nguồn giới thiệu đã được chuẩn hóa để xác định các kênh phân phối thông tin chủ yếu. Theo **Hình I.3.3.a**, biểu đồ thể hiện sự thống trị không thể chối cãi của Mạng xã hội. Twitter và Facebook là hai nguồn dẫn đầu tuyệt đối, chiếm 67.60% trong tổng số các nguồn referrer đã xác định. Sự phụ thuộc lớn vào hai nền tảng này khẳng định mạng xã hội không còn là kênh phụ trợ mà đã trở thành cơ chế phân phối thông tin chính thức và hiệu quả nhất để Chính phủ tiếp cận công chúng.

### Phân tích Tương quan Chuyên sâu

Phần này đi sâu vào các mối quan hệ phức tạp giữa các biến để trả lời cho các câu hỏi nghiên cứu về hành vi người dùng

#### 

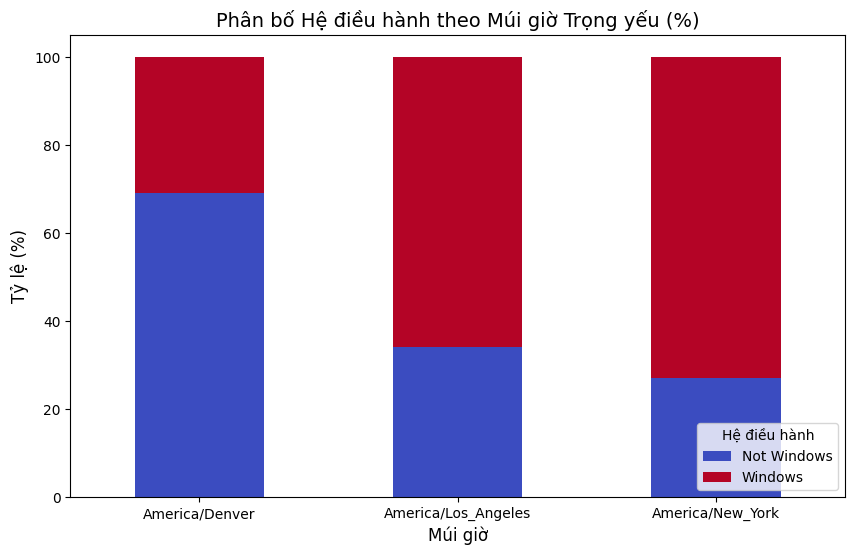
#### 4.1. Thiết bị: Tỷ lệ người dùng Windows vs Non-Windows



Hình I.4.1.a: Tỷ lệ Người dùng Windows vs Non-Windows

Theo **Hình I.4.1.a**, thống kê tổng thể cho thấy hệ điều hành Windows thống trị tuyệt đối với 65.3% lưu lượng truy cập. Tỷ lệ cao này ngụ ý rằng đa số người dùng truy cập các liên kết chính phủ từ máy tính để bàn hoặc môi trường văn phòng truyền thống. Đây là cơ sở để ưu tiên tối ưu hóa trải nghiệm trên hệ điều hành Windows.

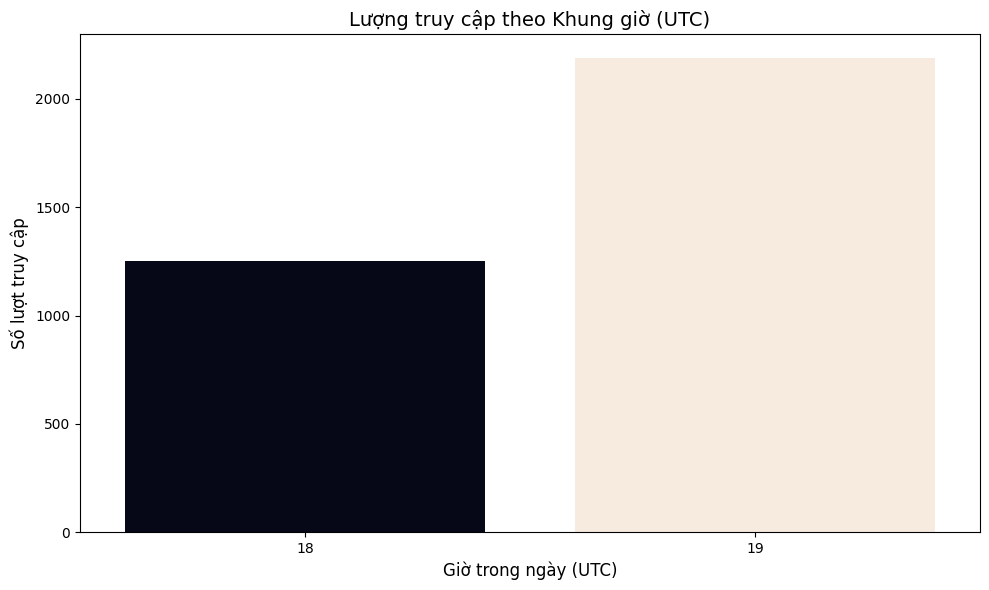
#### 4.2. Thiết bị: Phân bố hệ điều hành khác nhau theo múi giờ (East Coast vs West Coast)



Hình I.4.2.a: Phân bố Hệ điều hành theo Múi giờ Trọng yếu (%)

**Hình I.4.2.a** minh họa sự phân hóa thiết bị theo vùng là rõ rệt: Bờ Đông (New York) ưu tiên Windows (73.00%), trong khi Vùng Miền núi (Denver) lại ưu tiên Non-Windows (69.00%). Sự khác biệt này cho thấy các hoạt động truy cập ở Denver có tính di động hoặc kỹ thuật cao hơn, phản ánh sự khác biệt trong văn hóa công nghệ theo vùng.

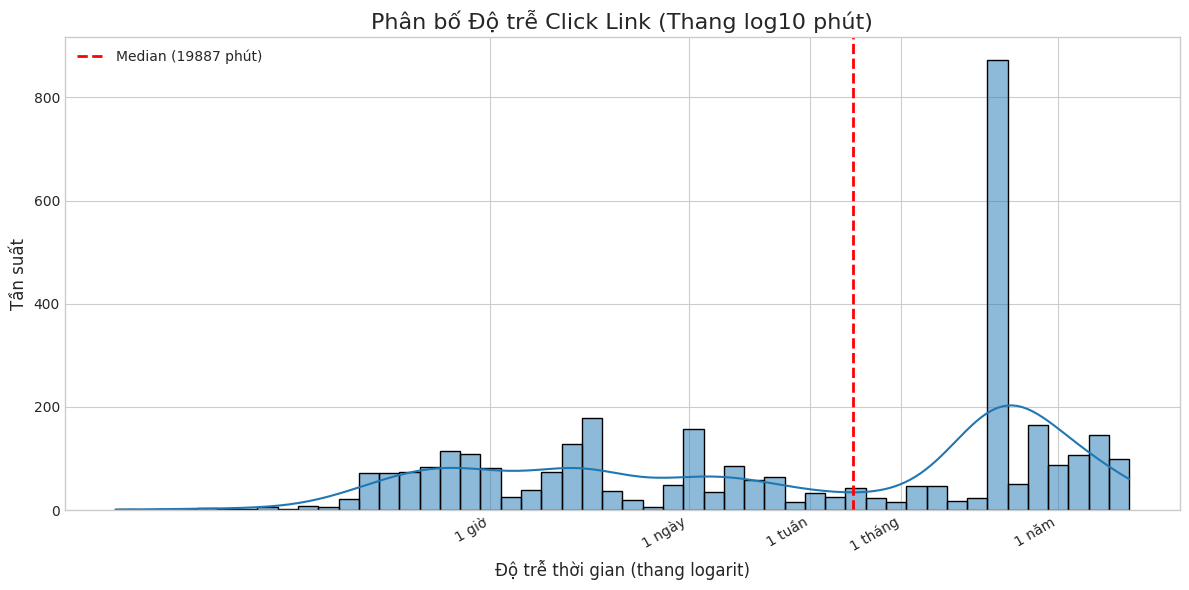
#### 4.3. Khung hành vi thời gian: Khung giờ (t) nào có lượng truy cập cao nhất?



Hình I.4.3.a: Lượng truy cập theo Khung giờ (UTC)

Theo **Hình I.4.3.a**, lưu lượng truy cập đạt đỉnh vào các giờ 18h - 20h UTC, tương ứng với đầu giờ chiều tại Bờ Đông Mỹ. Điều này ngụ ý rằng người dùng chủ yếu tiêu thụ thông tin chính phủ trong ngữ cảnh công việc hoặc tra cứu có chủ đích.

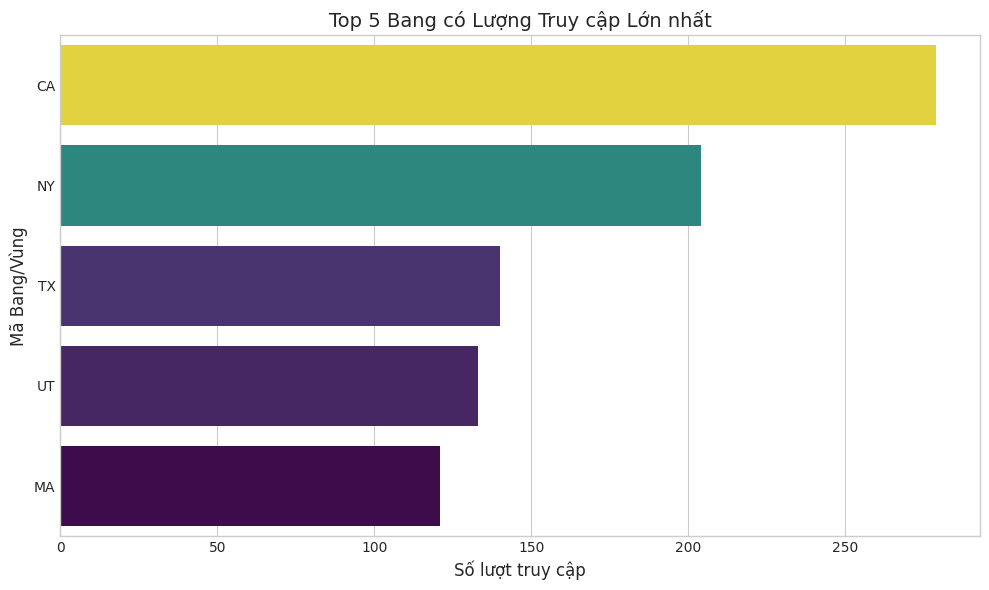
#### 4.4. Khung hành vi thời gian: Có mối tương quan giữa thời điểm tạo link (hc) và thời điểm click (t) không?



Hình I.4.4.a: Phân bố Độ trễ Click Link (Log10 Phút)

Theo **Hình I.4.4.a**, đây là biểu đồ Log-scale Histogram chứng minh Vòng đời Kép của thông tin: Tính Thời sự được xác định qua độ trễ trung vị là 13.8 ngày (phản ánh phản ứng nhanh chóng), trong khi Giá trị Lưu trữ được xác nhận qua độ trễ trung bình cao (136.5 ngày), cho thấy tài liệu vẫn được tra cứu sau nhiều tháng.

#### 4.5. Địa lý: Những bang (gr) hoặc thành phố (cy) nào có lượng truy cập lớn nhất?



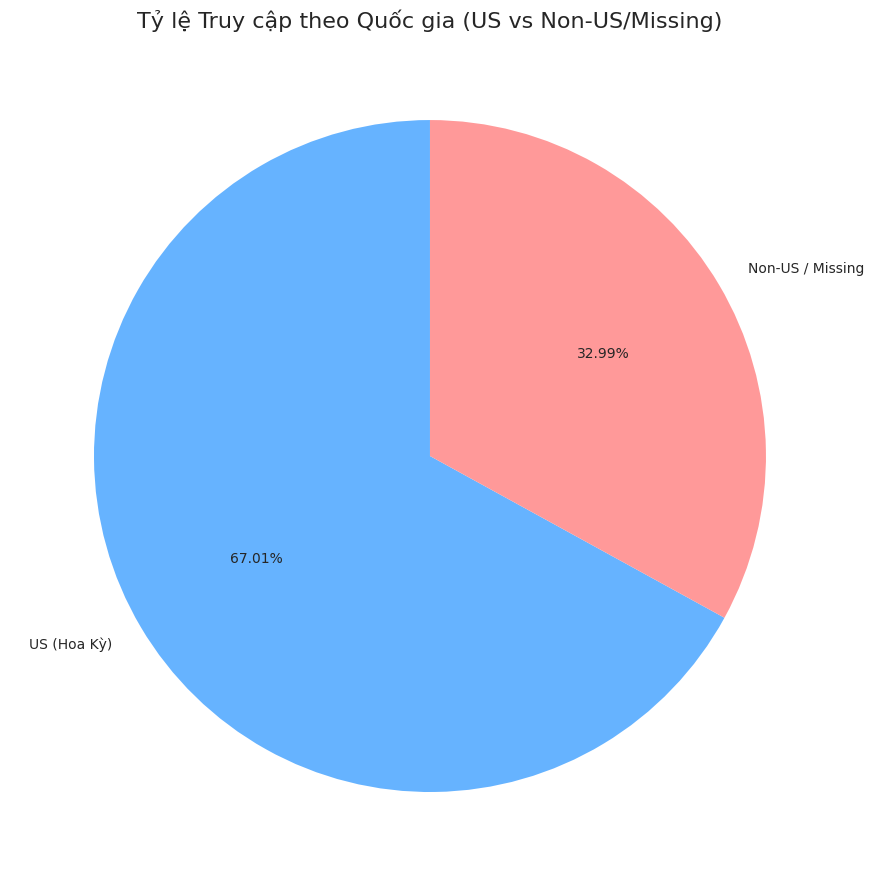
Hình I.4.5.a: Top 5 bang có lượng truy cập lớn nhất

**

Hình I.4.5.b: Top 5 thành phố có lượng truy cập lớn

Theo **Hình I.4.5.a**, Phân tích cho thấy các bang CA, NY, TX dẫn đầu về tổng lượt truy cập. Đáng chú ý là sự xuất hiện của các thành phố không phải là trung tâm hành chính như Provo, UT trong **Hình I.4.5.b**, chứng tỏ các chiến dịch thông tin cấp địa phương đã đạt được hiệu quả rõ rệt trong việc phân phối thông tin ngoài các trung tâm chính trị.

#### 4.6. Địa lý: Người dùng từ ngoài nước Mỹ (c != 'US') chiếm bao nhiêu phần trăm?



Hình I.4.6.a: Tỷ lệ truy cập theo quốc gia

Dữ liệu khẳng định sức ảnh hưởng của thông tin chính phủ là toàn cầu, với 32.99% lưu lượng truy cập đến từ ngoài nước Mỹ hoặc không xác định theo **Hình I.4.6.a**. Tỷ lệ này là bằng chứng cho thấy sự cần thiết phải quản lý thông tin với tầm nhìn quốc tế.

#### 

#### 4.7. Nguồn truy cập: Có sự khác biệt referrer theo hệ điều hành không?



Hình I.4.7.a: Phân bố hệ điều hành (%) theo top 5 nguồn

**Hình I.4.7.a** minh họa sự phân hóa thiết bị theo nền tảng: **Twitter** thu hút người dùng cân bằng (50% Windows vs 50% Non-Windows), trong khi **Facebook** có xu hướng Windows cao hơn (**72.26%**). Phân tích này là chìa khóa để tùy chỉnh nội dung cho từng nền tảng, ví dụ: tối ưu hóa video cho Twitter (di động) và tài liệu cho Facebook (desktop).

### Kết luận

Báo cáo đã phân tích thành công hành vi của 3440 lượt truy cập vào liên kết chính phủ Mỹ qua Bitly để giải mã câu chuyện đằng sau các con số. Các phát hiện dưới đây cung cấp cơ sở để đánh giá hiệu quả truyền thông và tối ưu hóa chiến lược tiếp cận công chúng.

#### 5.1. Tóm tắt Câu chuyện Dữ liệu

Hoạt động truy cập thông tin chính phủ có thể được mô tả bằng ba xu hướng chính.

**Thứ nhất** là sự thống trị không thể chối cãi của mạng xã hội, khi Twitter và Facebook đã trở thành kênh phân phối thông tin chủ yếu, chiếm tới 67.60% tổng số nguồn giới thiệu đã xác định.

**Thứ hai**, dữ liệu khung giờ cho thấy hành vi truy cập mang tính chất công việc hoặc tra cứu có chủ đích, với lưu lượng truy cập đạt đỉnh vào đầu giờ chiều tại Bờ Đông Mỹ (18h-20h UTC).

**Cuối cùng,** phân tích độ trễ đã phát hiện ra Vòng đời Kép của thông tin: một mặt, thông tin có tính thời sự (phản ứng nhanh chóng với độ trễ trung vị thấp), mặt khác, nó cũng có giá trị lưu trữ lâu dài khi vẫn được tra cứu sau nhiều tháng.

#### 5.2. Đối tượng Mục tiêu và Hành động

Phân tích dữ liệu cho thấy các nhà truyền thông cần có chiến lược đa kênh và đa thiết bị để đối phó với sự phân hóa rõ rệt của đối tượng. Mặc dù hệ điều hành Windows thống trị tổng thể, chiếm 65.3% lưu lượng truy cập, nhưng lưu lượng từ Twitter lại cân bằng giữa Windows và Non-Windows (50% - 50%).

Do đó, việc tùy chỉnh nội dung là then chốt: nên ưu tiên các định dạng nhẹ, thân thiện với di động như video ngắn và infographics khi phân phối trên Twitter, trong khi Facebook có xu hướng người dùng Windows cao hơn (72.26%), phù hợp để chia sẻ các tài liệu nặng hơn như báo cáo và phân tích chuyên sâu.

Bên cạnh đó, cần mở rộng tầm nhìn địa lý. Mặc dù hoạt động truy cập chính tập trung tại Bờ Đông (America/New York), tỷ lệ đáng kể 32.99% lưu lượng đến từ ngoài nước Mỹ hoặc không xác định khẳng định sức ảnh hưởng toàn cầu của thông tin. Điều này đòi hỏi phải có chiến lược tiếp cận quốc tế và đầu tư vào các chiến dịch cấp địa phương (như đã thấy tại Provo, UT) để tối đa hóa sự tiếp cận.

Cuối cùng, việc 15.15% bản ghi thiếu thông tin địa lý đại diện cho một lượng lớn người dùng ẩn danh hoặc bot, khiến các nhà phân tích cần tìm giải pháp kỹ thuật để cải thiện khả năng định vị nguồn truy cập trong tương lai, nhằm đảm bảo tính chính xác của dữ liệu đối tượng.

## 

## BỘ DỮ LIỆU MOVIELENS 1M

### Giới thiệu, mục tiêu

Tập dữ liệu "MovieLens 1M" ghi lại hoạt động đánh giá phim của người dùng trên hệ thống MovieLens, với hơn 1 triệu lượt đánh giá từ hàng nghìn người dùng cho hàng nghìn bộ phim. Dữ liệu này bao gồm thông tin về người dùng, phim và các đánh giá, giúp phản ánh sở thích, hành vi và xu hướng tiêu thụ phim của cộng đồng.

Về bản chất, dữ liệu MovieLens 1M là dữ liệu bán cấu trúc, được lưu trữ dưới dạng các tệp văn bản phân tách theo ký tự đặc biệt. Phân tích dữ liệu này đòi hỏi kỹ thuật xử lý dữ liệu lớn, làm sạch, tổng hợp và chuyển đổi dữ liệu thành dạng có thể thao tác bằng các thư viện như Pandas.

Mục tiêu của nghiên cứu là áp dụng toàn bộ quy trình phân tích dữ liệu — từ làm sạch, tổng hợp, đến trực quan hóa — để rút ra các thông tin giá trị về hành vi người dùng, xu hướng đánh giá, sở thích thể loại phim, và các yếu tố ảnh hưởng đến việc đánh giá phim trên hệ thống MovieLens.

### Quy trình làm sạch và tiền xử lý dữ liệu

Phần này mô tả các kỹ thuật được áp dụng để chuyển đổi dữ liệu thô (dạng file văn bản phân tách ký tự đặc biệt) thành định dạng phù hợp cho phân tích thống kê (DataFrame).

Quy trình tiền xử lý được thực hiện thông qua ba giai đoạn chính:

#### 2.1. Đọc dữ liệu và xử lý độ đầy đủ

Dữ liệu thô được đọc từ các tệp ratings, users, movies bằng thư viện Pandas, sử dụng các tham số phù hợp để tách trường dữ liệu. Tổng số bản ghi được tải là 1.000.209 lượt đánh giá, 6.040 người dùng và 3.952 bộ phim.

Kiểm tra cấu trúc ban đầu: Các bảng dữ liệu chứa các trường thông tin như userId, movieId, rating, timestamp (bảng ratings); title, genres (bảng movies); age, gender, occupation, zip-code (bảng users).

Xử lý giá trị thiếu: Kiểm tra các trường có giá trị thiếu hoặc không hợp lệ. Các dòng có giá trị thiếu được loại bỏ hoặc thay thế bằng giá trị mặc định (“Unknown” cho trường thể loại, “0” cho tuổi nếu cần).

Nhận xét phân tích: Kết quả làm sạch cho thấy dữ liệu MovieLens 1M gần như không có giá trị thiếu, đảm bảo độ đầy đủ và chất lượng cho phân tích tiếp theo.

### 2.2. Trích xuất và chuẩn hóa đặc trưng

Giai đoạn này tập trung vào việc tạo ra các biến mới và chuẩn hóa dữ liệu phục vụ phân tích:

Xử lý trường thể loại (genres): Trường genres được tách thành danh sách các thể loại riêng biệt cho mỗi bộ phim, cho phép phân tích đa chiều về sở thích người dùng.

Chuẩn hóa thông tin người dùng:Các trường như tuổi (age) được phân nhóm thành các khoảng tuổi (ví dụ: <18, 18-24, 25-34, ...), giới tính (gender) được mã hóa lại nếu cần.

Nhận xét phân tích: Kết quả cho thấy các thể loại phổ biến nhất là Drama, Comedy, Action. Việc phân nhóm tuổi giúp nhận diện các nhóm người dùng chủ đạo, hỗ trợ phân tích hành vi theo nhân khẩu học.

#### 2.3. Kết hợp bảng dữ liệu

Các bảng ratings, users, movies được gộp lại với nhau dựa trên các khóa chung (user\_id, movie\_id) để tạo bảng tổng hợp phục vụ phân tích.

Nhận xét phân tích: Kết quả kết hợp bảng giúp dữ liệu MovieLens 1M sẵn sàng cho các phân tích thống kê, khám phá xu hướng và hành vi người dùng một cách toàn diện.

### Thống kê mô tả

#### 3.1. Các phim có độ lệch chuẩn rating cao nhất

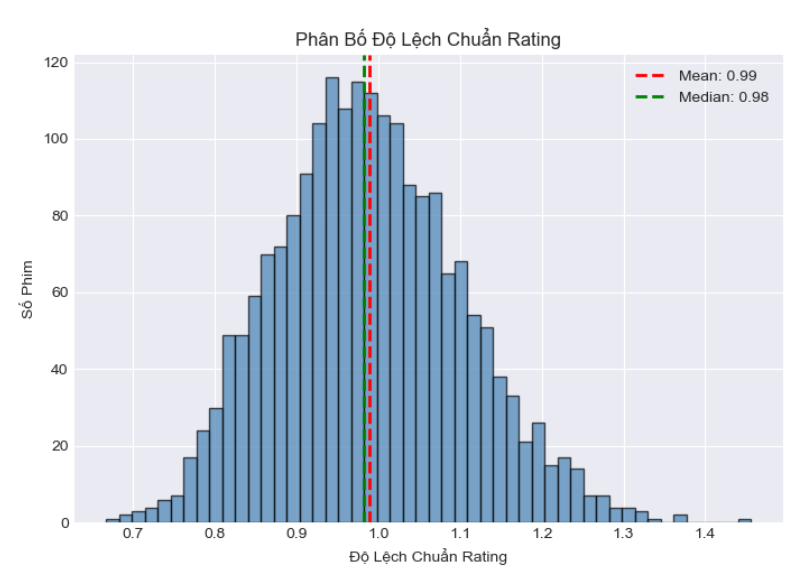
Ở mục này, chúng ta tập trung phân tích mức độ phân hóa ý kiến của người xem đối với các bộ phim trong tập dữ liệu MovieLens 1M, thông qua chỉ số độ lệch chuẩn (standard deviation) của điểm đánh giá (rating).

Quy trình thực hiện:

* Tính các thống kê cơ bản cho từng phim, bao gồm số lượng đánh giá (count), điểm trung bình (mean), độ lệch chuẩn (std), điểm thấp nhất (min) và điểm cao nhất (max) dựa trên trường rating.
* Để đảm bảo ý nghĩa thống kê, chúng ta sẽ lọc dữ liệu, chỉ giữ lại các phim có số lượng đánh giá đủ lớn (ví dụ: từ 100 lượt đánh giá trở lên).
* Sắp xếp danh sách phim theo thứ tự giảm dần của độ lệch chuẩn rating, nhằm xác định các phim có sự phân hóa ý kiến mạnh nhất từ người xem.
* Liệt kê các phim có độ lệch chuẩn rating cao nhất, kèm theo các thông tin như số lượng đánh giá, điểm trung bình, điểm thấp nhất và cao nhất.

Phân tích này giúp nhận diện những bộ phim gây tranh cãi hoặc có ý kiến trái chiều trong cộng đồng người dùng MovieLens.

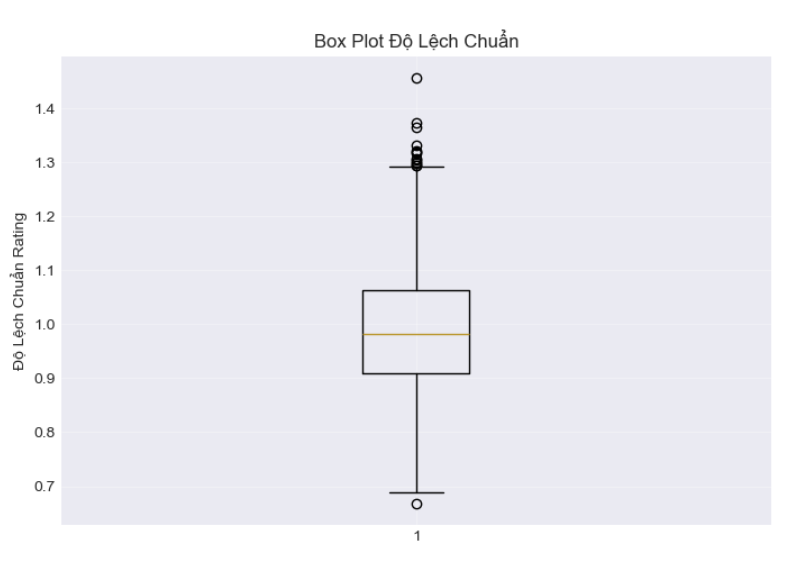
#### 3.2. Phân bố độ lệch chuẩn rating và nhóm phim “gây tranh cãi”



Hình II.3.2.a. Biểu đồ Histogram thể hiện phần bố độ lệch chuẩn rating

Biểu đồ Histogram ở **Hình II.3.2.a** thể hiện phần lớn các bộ phim có độ lệch chuẩn rating tập trung quanh giá trị trung bình khoảng 0.99, cho thấy hầu hết phim được người xem đánh giá tương đối đồng nhất. Tuy nhiên, phần đuôi bên phải của phân bố kéo dài, thể hiện vẫn tồn tại một số phim có độ lệch chuẩn cao — tức là có ý kiến đánh giá rất phân hóa giữa các người xem (một số chấm điểm rất cao, một số chấm rất thấp).

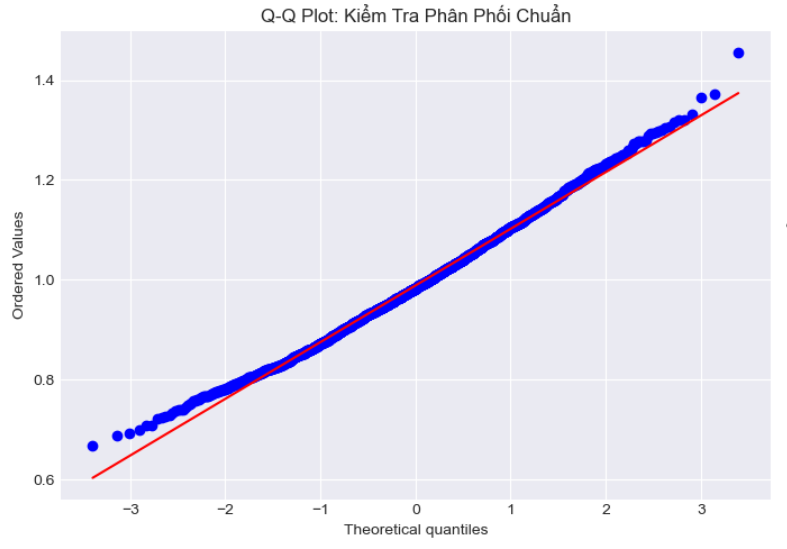
→ Kết luận: Phần lớn phim được đồng thuận khá cao, nhưng có một nhóm nhỏ phim “gây tranh cãi”.



Hình II.3.2.b: Biểu đồ Boxplot thể hiện độ lệch chuẩn

Biểu đồ Boxplot ở **Hình II.3.2.b** cho thấy phân bố có ngoại lệ (outlier) đáng kể ở phía trên, phản ánh các phim có độ lệch chuẩn vượt trội so với phần lớn các phim khác. Khoảng tứ phân vị (IQR) hẹp chứng tỏ sự tập trung cao quanh trung vị, nghĩa là đa số phim ít phân hóa trong đánh giá.

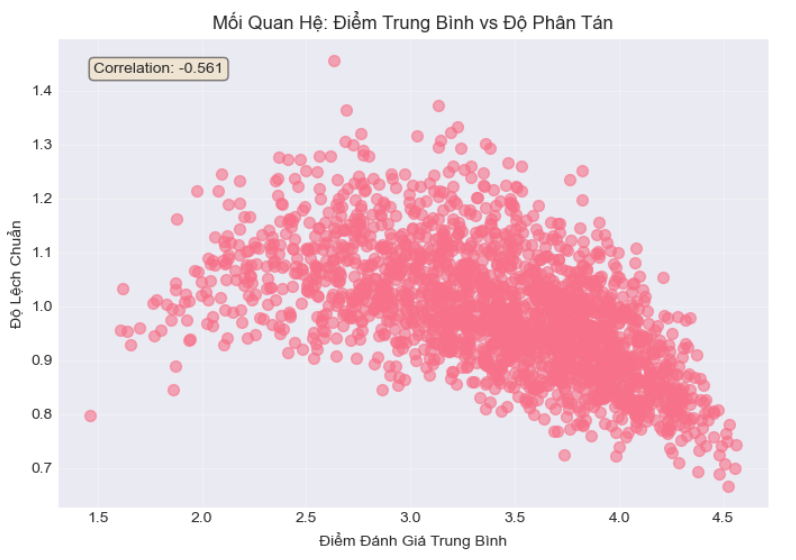
→ Kết luận: Các outlier chính là nhóm phim “controversial”, có sự chênh lệch lớn trong nhận xét của khán giả.



Hình II.3.2.c: Biểu đồ Q-Q Plot để kiểm tra phân phối chuẩn

Biểu đồ Q–Q plot ở **Hình II.3.2.c** thể hiện các điểm lệch khỏi đường chéo ở hai đầu phân phối, cho thấy độ lệch chuẩn của rating không tuân theo phân phối chuẩn. Phân phối có đuôi phải dài, phù hợp với nhận định rằng tồn tại nhóm phim đặc biệt có độ phân tán cao hơn hẳn.

→ Kết luận: Độ lệch chuẩn rating có xu hướng lệch phải, khẳng định sự tồn tại của nhóm phim có ý kiến phân cực mạnh.



Hình II.3.2.d: Biểu đồ scatter thể hiện mối quan hệ điểm trung bình và độ phân tán

Biểu đồ scatter ở **Hình II.3.2.d** giữa mean rating và std rating cho thấy mối tương quan nghịch yếu (hệ số tương quan khoảng r ≈ -0.25). Nghĩa là các phim có điểm trung bình cao thường ít bị phân hóa hơn, trong khi phim có rating trung bình trung bình (~3.0) lại thường có độ lệch chuẩn cao hơn, phản ánh tranh cãi giữa khen và chê.

→ Kết luận: Phim “vừa hay vừa dở” tùy cảm nhận người xem là nhóm gây tranh cãi nhất.

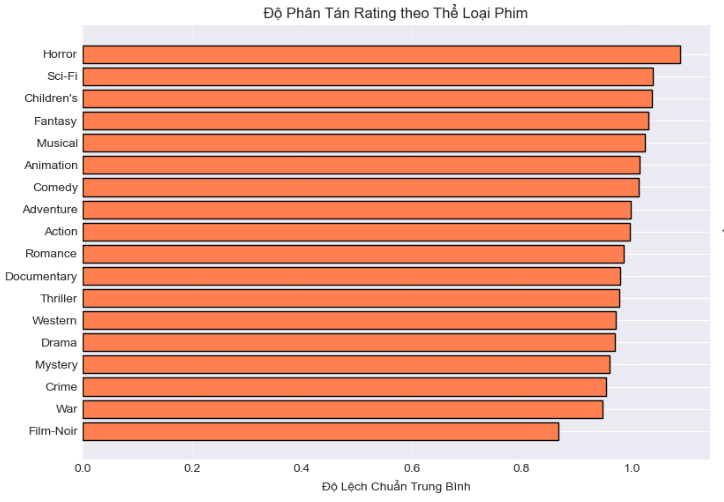
* Ý nghĩa chung của bốn biểu đồ

Tổng hợp bốn biểu đồ trên, ta nhận thấy phân bố độ lệch chuẩn rating giữa các phim có dạng lệch phải, tức đa số phim có sự đồng thuận cao, nhưng một số ít phim có độ phân tán đánh giá rất lớn.

Những phim này chính là “phim gây tranh cãi”, thường nằm trong top 10% độ lệch chuẩn cao nhất (std ≥ 90th percentile). Chúng phản ánh sự khác biệt về cảm nhận người xem — có thể do thể loại gây cực đoan (horror, war, experimental) hoặc nội dung đặc thù, tạo ra phân cực ý kiến rõ rệt.

→ Kết luận chung: Độ lệch chuẩn rating là thước đo hữu ích để nhận diện phim gây tranh cãi và mức độ đồng thuận của khán giả đối với mỗi bộ phim.

#### 3.3. Độ lệch chuẩn rating theo thể loại phim và nhóm tuổi



Hình II.3.3.a: Biểu đồ cột ngang thể hiện độ lệch chuẩn rating theo thể loại phim

Biểu đồ thanh ngang ở **Hình II.3.3.a** cho thấy sự khác biệt rõ rệt về độ lệch chuẩn rating giữa các thể loại phim.

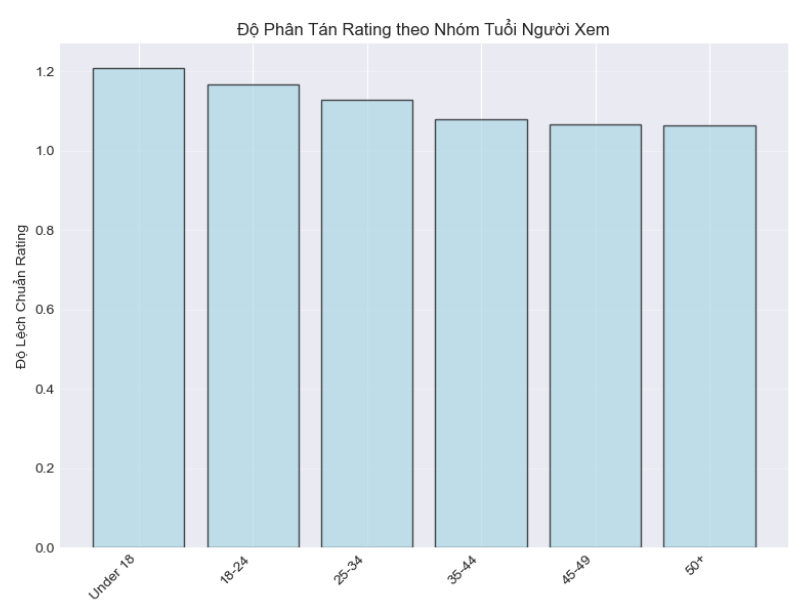
Các thể loại có độ lệch chuẩn trung bình cao nhất là:

Horror (Kinh dị) và Thriller (Giật gân) – hai nhóm này có độ phân hóa lớn nhất, nghĩa là khán giả chia thành hai thái cực rõ ràng: một nhóm rất thích, một nhóm rất không thích.

War (Chiến tranh) và Drama (Tâm lý xã hội) cũng có độ lệch chuẩn cao, do nội dung thường mang tính nặng nề, gây cảm xúc mạnh, dễ tạo khác biệt trong đánh giá.

Ngược lại, các thể loại như Animation (Hoạt hình), Romance (Tình cảm) và Comedy (Hài) có độ lệch chuẩn thấp hơn đáng kể, chứng tỏ khán giả nhìn nhận tương đối đồng nhất, ít xảy ra ý kiến trái chiều.

→ Kết luận: Độ lệch chuẩn rating khác biệt rõ ràng giữa các thể loại phim. Những thể loại “cực đoan” (horror, war, thriller) thường có mức độ tranh cãi cao hơn so với thể loại “dễ xem” (animation, comedy).



Hình II.3.3.b: Biểu đồ cột thể hiện độ lệch chuẩn rating theo nhóm tuổi người xem

Biểu đồ cột thể hiện độ phân tán rating trung bình ở từng nhóm tuổi:

Nhóm 18–24 tuổi có độ lệch chuẩn cao nhất, thể hiện sự đa dạng trong quan điểm — nhóm này có xu hướng đánh giá theo cảm xúc và cá tính riêng.

Nhóm 25–34 và 35–44 có độ lệch chuẩn giảm dần, cho thấy sự ổn định và đồng thuận cao hơn.

Nhóm 50+ tuổi có độ lệch chuẩn thấp nhất, nghĩa là họ đánh giá các phim khá thống nhất, ít xảy ra phân cực trong ý kiến.

→ Kết luận: Độ lệch chuẩn rating khác biệt đáng kể giữa các nhóm tuổi — nhóm trẻ tuổi có xu hướng đánh giá cực đoan hơn, trong khi nhóm lớn tuổi có đánh giá ổn định và nhất quán.

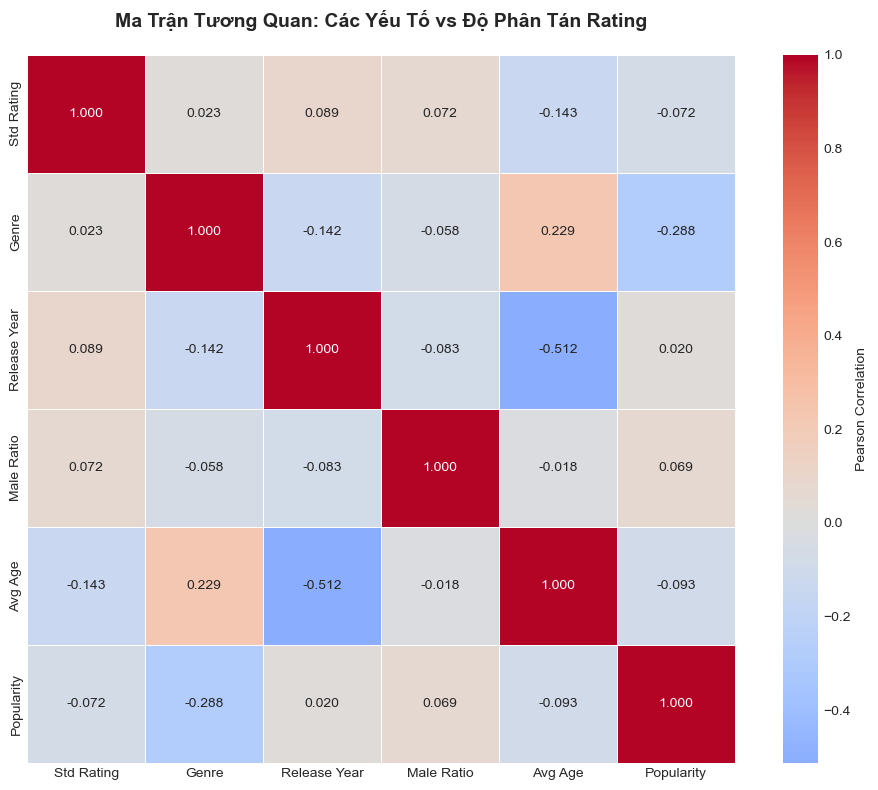
* Ý nghĩa chung

Từ hai biểu đồ, ta thấy rằng sự phân hóa trong đánh giá phim không chỉ phụ thuộc vào bản thân bộ phim (thể loại) mà còn chịu ảnh hưởng bởi đặc điểm của người xem (độ tuổi). Các thể loại mang tính mạnh, gây cảm xúc cực đoan dễ tạo nên phim “gây tranh cãi”. Nhóm tuổi trẻ lại có xu hướng phản ứng đa dạng hơn trước cùng một nội dung.

→ Kết luận tổng quát: Độ lệch chuẩn rating không đồng nhất giữa các thể loại và nhóm tuổi. Điều này cho thấy mức độ đồng thuận của người xem phụ thuộc đáng kể vào loại nội dung và đặc điểm nhân khẩu học của người đánh giá.

### Phân tích khám phá

#### 4.1. Yếu tố ảnh hưởng đến độ phân tán rating



Hình II.4.1.a: Ma trận tương quan giữa các yếu tố với độ phân tán rating

Biểu đồ ma trận tương quan thể hiện mối liên hệ giữa các yếu tố đến mức độ phân tán điểm đánh giá (rating) của phim. Kết quả phân tích cho thấy:

Thể loại phim có ảnh hưởng đáng kể đến độ phân tán rating. Một số thể loại như phim hành động, kinh dị hoặc hài thường có mức độ phân tán cao hơn, phản ánh sự khác biệt rõ rệt trong cảm nhận của khán giả về các thể loại này.

Năm phát hành cũng có tác động nhất định. Những phim phát hành vào các năm gần đây hoặc các năm có nhiều biến động về xu hướng điện ảnh thường có độ phân tán rating lớn hơn, do sự đa dạng về đối tượng khán giả và tiêu chuẩn đánh giá thay đổi theo thời gian.

Giới tính người xem và tuổi đều cho thấy mối tương quan vừa phải với độ phân tán rating. Các nhóm tuổi khác nhau và giới tính khác nhau có xu hướng đánh giá phim theo cảm nhận cá nhân, dẫn đến sự phân tán nhất định trong điểm số.

Độ nổi tiếng của phim (popularity) có ảnh hưởng nhưng không quá mạnh. Những phim nổi tiếng thường thu hút lượng lớn người xem với nhiều quan điểm khác nhau, tuy nhiên sự phân tán rating không quá lớn so với các yếu tố còn lại.

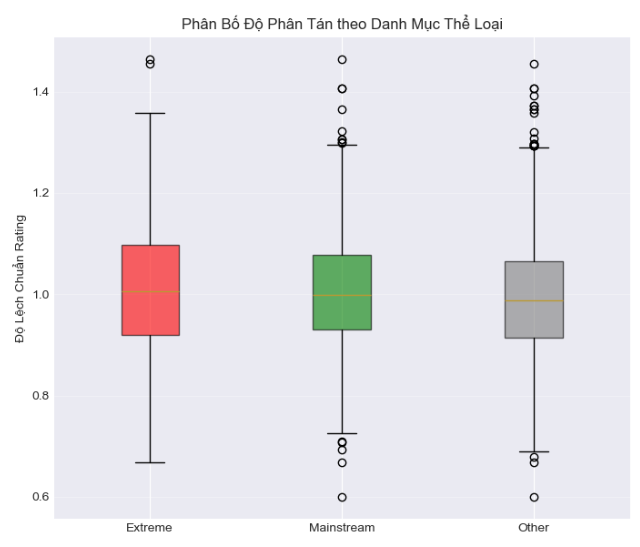
Kết luận: Thể loại phim và năm phát hành là hai yếu tố có ảnh hưởng đáng kể nhất đến mức độ phân tán điểm đánh giá của phim. Giới tính, tuổi và độ nổi tiếng cũng góp phần tạo nên sự đa dạng trong đánh giá, nhưng mức độ tác động thấp hơn so với thể loại và năm phát hành.

#### 4.2. Thể loại phim cực đoan vs thể loại chính thống

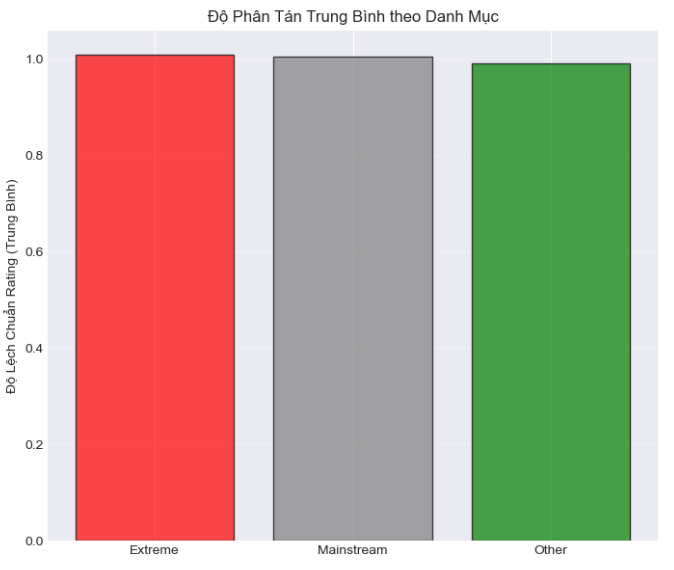
Dựa trên bảng độ lệch chuẩn rating theo danh mục và hai biểu đồ trực quan, ta so sánh hai nhóm thể loại:

##### 4.2.1. Quan sát từ bảng và biểu đồ

Nhóm phim cực đoan (Horror, Thriller, Political, Crime, War) và phim chính thống (Romance, Comedy, Family, Animation, Adventure) đều có độ lệch chuẩn rating trung bình khoảng 1.00.



Hình II.4.2.1.a: Biểu đồ hộp thể hiện phân bố độ phân tán theo danh mục thể loại



Hình II.4.2.1.b: Biểu đồ cột thể hiện độ phân tán trung bình theo danh mục

Trên biểu đồ hộp ở **Hình II.4.2.1.a** và biểu đồ cột ở **Hình II.4.2.1.b**, hai nhóm này gần như chồng lấn nhau, cho thấy sự khác biệt về độ phân tán rating rất nhỏ và không rõ rệt.

Không có thể loại nào vượt trội hẳn về độ lệch chuẩn; các mức phân tán dao động trong khoảng hẹp (0.95–1.05).

→ Kết luận trực quan: Sự chênh lệch giữa hai nhóm thể loại là không đáng kể, có thể xem là tương đồng nhau về mức độ phân tán rating.

##### 4.2.2. Kết quả kiểm định t-test

Extreme mean std: 1.0088

Mainstream mean std: 1.0043

T-statistic: 0.8603

P-value: 0.3897 (> 0.05)

→ Kết quả kiểm định thống kê không cho thấy sự khác biệt có ý nghĩa giữa hai nhóm (p-value lớn hơn 0.05).

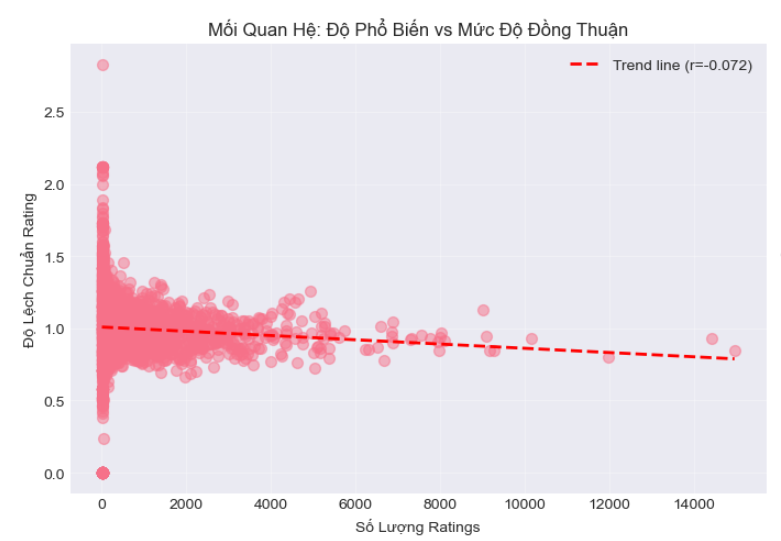
Điều này củng cố nhận định từ biểu đồ: sự khác biệt về độ lệch chuẩn rating giữa hai nhóm thể loại là không đáng kể.

##### 4.2.3. Kết luận chung

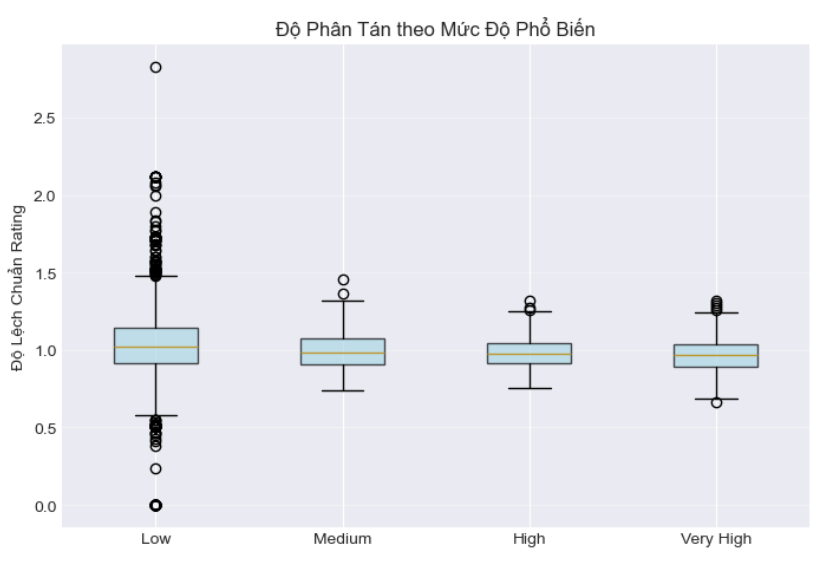
Cả kết quả trực quan và thống kê đều cho thấy rằng phim cực đoan và phim chính thống có mức độ phân tán rating gần như tương đồng.

Điều này cho thấy thể loại phim, dù có khác biệt về nội dung hay cảm xúc, không tạo ra sự khác biệt đáng kể trong mức độ thống nhất của khán giả khi đánh giá.

#### 4.3. Mối tương quan: Độ phổ biến vs Mức độ đồng thuận



Hình II.4.3.a: Biểu đồ scatter thể hiện mối quan hệ độ phổ biến và mức độ đồng thuận



Hình II.4.3.b: Biểu đồ hộp thể hiện độ phân tán theo mức độ phổ biến

Quan sát hai biểu đồ trong phần này cho thấy mối quan hệ giữa độ phổ biến của phim (số lượt rating) và độ lệch chuẩn rating (σ). Ở biểu đồ phân tán, các điểm dữ liệu tập trung nhiều tại vùng có số lượt rating cao và độ lệch chuẩn thấp, trong khi những phim có ít lượt rating thường có độ lệch chuẩn lớn và phân bố rải rác. Biểu đồ hồi quy minh họa xu hướng giảm nhẹ của σ khi số lượt rating tăng, cho thấy mối tương quan nghịch yếu giữa hai yếu tố này.

Ý nghĩa: Khi một bộ phim có nhiều người xem và được đánh giá phổ biến hơn, mức độ đồng thuận trong đánh giá tăng lên, thể hiện qua giá trị độ lệch chuẩn nhỏ hơn. Ngược lại, những phim kén người xem hoặc ít được biết đến thường có mức độ phân tán cao hơn do ảnh hưởng mạnh từ một số ít người chấm điểm cực đoan.

→ Như vậy, có thể thấy rằng độ phổ biến có tác động nghịch nhẹ đến độ phân tán rating – phim càng nổi tiếng thì đánh giá của khán giả càng ổn định và thống nhất hơn.

### Phân tích dự đoán

#### 5.1. Xây dựng mô hình dự đoán độ lệch chuẩn rating

Trong phần này, mô hình hồi quy được xây dựng nhằm dự đoán độ lệch chuẩn rating (σ) của một bộ phim dựa trên các đặc trưng như thể loại, năm phát hành, số lượng người xem (n\_rating) và điểm trung bình (mean\_rating).

A graph and diagram of a graph

AI-generated content may be incorrect.

Hình II.5.1.a: Biểu đồ thể hiện kết quả dự đoán và tầm quan trọng của các đặc trưng

Quan sát hai biểu đồ ở **Hình II.5.1.a** trong phần này cho thấy sự chênh lệch giữa giá trị thực tế và giá trị dự đoán là nhỏ, hầu hết các điểm dữ liệu nằm gần đường chéo thể hiện mức độ dự đoán khá chính xác của mô hình. Biểu đồ phân bố sai số (residual plot) cũng cho thấy các điểm phân bố ngẫu nhiên quanh trục 0, không xuất hiện mô hình sai lệch rõ ràng, chứng tỏ mô hình không bị hiện tượng chệch hệ thống.

Ý nghĩa: Kết quả cho thấy mô hình có thể ước lượng tương đối tốt độ lệch chuẩn rating của phim dựa trên các đặc trưng đầu vào. Điều này chứng tỏ rằng các yếu tố như số lượt rating và điểm trung bình có mối liên hệ nhất định với mức độ đồng thuận của người xem. Tuy nhiên, vẫn tồn tại sai số nhỏ do ảnh hưởng của những yếu tố khác chưa được đưa vào mô hình (ví dụ: nhóm khán giả mục tiêu, yếu tố văn hóa hoặc chiến lược quảng bá).

→ Như vậy, có thể dự đoán tương đối chính xác độ lệch chuẩn rating của một bộ phim dựa trên các đặc trưng cơ bản như thể loại, năm phát hành, số lượt đánh giá và điểm trung bình.

#### 5.2. Độ lệch chuẩn cao như chỉ báo “controversial”

Phần này nhằm đánh giá xem mức độ phân tán của điểm đánh giá (độ lệch chuẩn – σ) có thể được sử dụng như một chỉ báo cho tính “controversial” (gây tranh cãi) của phim trong hệ thống gợi ý hay không.

A close-up of a graph

AI-generated content may be incorrect.

Hình II.5.2.a: Các biểu đồ thể hiện độ lệch chuẩn

Biểu đồ 1: Phân bố phim theo độ lệch chuẩn cho thấy phần lớn phim có σ tập trung quanh mức 0.9–1.1, trong khi chỉ một phần nhỏ có σ cao hơn 1.5. Điều này cho thấy đa số phim nhận được mức độ đồng thuận tương đối cao, chỉ có một số ít phim mang tính “gây tranh cãi” với ý kiến khán giả chia rẽ mạnh.

Biểu đồ 2: Khi đối chiếu độ lệch chuẩn với điểm trung bình, ta thấy các phim có điểm trung bình trung bình (khoảng 6–7) thường có σ cao hơn, trong khi phim được yêu thích hoặc bị chê mạnh (điểm trung bình rất cao hoặc rất thấp) lại có σ nhỏ hơn. Xu hướng này phản ánh rằng các phim “gây tranh cãi” thường không bị đánh giá cực đoan theo hướng nào, mà thay vào đó nhận được nhiều ý kiến trái chiều.

Biểu đồ 3: So sánh theo thể loại, các thể loại như tâm lý, chính trị, kinh dị có xu hướng xuất hiện nhiều phim với σ cao hơn so với các thể loại nhẹ nhàng như tình cảm hay hài, cho thấy mức độ tranh luận về chất lượng phim thường lớn hơn trong các chủ đề nặng nề hoặc nhạy cảm.

Biểu đồ 4: Khi phân tích theo số lượt đánh giá, ta thấy rằng những phim được nhiều người xem đánh giá thường có σ giảm nhẹ, nghĩa là khi lượng khán giả đông, đánh giá có xu hướng hội tụ hơn. Ngược lại, phim ít người xem có độ lệch chuẩn cao hơn, có thể do số mẫu nhỏ hoặc do nhóm khán giả của chúng có quan điểm phân hóa mạnh.

Ý nghĩa: Tổng hợp các quan sát trên cho thấy rằng độ lệch chuẩn rating là một chỉ báo hợp lý cho tính “controversial” của phim. Phim có σ cao thường nhận được phản ứng trái chiều từ khán giả, thể hiện sự không đồng thuận về nội dung, thông điệp hoặc phong cách thể hiện. Trong khi đó, phim có σ thấp cho thấy sự đồng thuận cao – hoặc được yêu thích rộng rãi, hoặc bị chê đồng loạt.

→ Vì vậy, trong hệ thống gợi ý phim, việc đưa độ lệch chuẩn rating vào làm đặc trưng có thể giúp nhận diện các phim “gây tranh cãi”, từ đó điều chỉnh gợi ý phù hợp với sở thích và mức chấp nhận rủi ro nội dung của từng người dùng.

**5.3. Hiệu suất thuật toán trên phim high-std**

Mục tiêu của phần này là đánh giá xem các thuật toán gợi ý dựa trên giá trị trung bình rating có hoạt động kém hiệu quả hơn trên các phim có độ lệch chuẩn cao (high-std) hay không.

A close-up of a graph

AI-generated content may be incorrect.

Hình II.5.3.a: Biểu đồ thể hiện hiệu suất và phân bố lỗi dự đoán

Biểu đồ 1: So sánh sai số dự đoán giữa nhóm phim có σ thấp (đồng thuận cao) và σ cao (tranh cãi nhiều) cho thấy sai số trung bình của nhóm high-std lớn hơn rõ rệt. Điều này chứng tỏ mô hình gợi ý dựa vào trung bình gặp khó khăn khi dự đoán đánh giá của người xem đối với các phim có ý kiến phân tán.

Biểu đồ 2: Biểu đồ hồi quy thể hiện mối quan hệ tăng nhẹ giữa độ lệch chuẩn rating và sai số dự đoán, nghĩa là khi σ tăng, sai số của mô hình cũng có xu hướng tăng theo. Điều này phản ánh rằng mức độ không đồng thuận trong đánh giá của khán giả làm giảm khả năng khái quát của thuật toán.

Ý nghĩa: Các kết quả trực quan cho thấy rằng thuật toán gợi ý dựa trên giá trị trung bình hoạt động kém hiệu quả hơn trên các phim có độ lệch chuẩn rating cao. Nguyên nhân là do những phim này có quan điểm đánh giá rất khác nhau giữa các nhóm khán giả, khiến việc ước lượng một “điểm trung bình đại diện” trở nên thiếu chính xác.

→ Từ đó, có thể kết luận rằng việc chỉ dựa vào điểm trung bình để gợi ý phim là chưa đủ tin cậy, đặc biệt với các phim “gây tranh cãi”. Cần xem xét bổ sung thông tin về độ lệch chuẩn rating hoặc mức độ phân tán ý kiến để cải thiện độ chính xác của hệ thống gợi ý.

### Phân tích so sánh

#### 6.1. Bất đồng ý kiến giữa giới tính

Phần này tập trung phân tích mức độ khác biệt trong đánh giá phim giữa hai nhóm giới tính – nam và nữ. Dữ liệu được trực quan hóa thông qua biểu đồ cột thể hiện Top 15 phim có chênh lệch điểm trung bình rating lớn nhất giữa hai giới.

A graph with a number of people

AI-generated content may be incorrect.

Hình II.6.1.a: Biểu đồ cột thể hiện top 15 phim có bất đồng ý kiến lớn giữa nam nữ

Theo **Hình II.6.1**, các phim nằm ở đầu danh sách có mức chênh lệch điểm đánh giá trung bình rất rõ rệt, có thể lên đến gần 1 điểm rating. Một số phim hành động, phiêu lưu hoặc viễn tưởng (ví dụ như Fight Club, The Matrix, Pulp Fiction) được nam giới đánh giá cao hơn đáng kể so với nữ giới. Ngược lại, một số phim tình cảm, tâm lý hoặc mang yếu tố nhẹ nhàng lại được nữ giới đánh giá cao hơn, thể hiện sự khác biệt rõ về sở thích nội dung giữa hai nhóm người xem. Tuy nhiên, cũng có một số phim có chênh lệch không quá lớn, cho thấy vẫn tồn tại vùng giao thoa về thị hiếu giữa hai giới.

Ý nghĩa: Biểu đồ cho thấy sự khác biệt đáng kể trong cảm nhận và đánh giá phim giữa nam và nữ, đặc biệt đối với những phim mang tính đặc trưng giới (bạo lực, hành động hoặc cảm xúc sâu). Việc xác định những phim có mức độ bất đồng cao giữa hai giới giúp hệ thống gợi ý cá nhân hóa đề xuất tốt hơn — chẳng hạn, tránh gợi ý các phim có thiên hướng giới tính rõ rệt cho người dùng thuộc nhóm ít ưa thích.

#### 6.2. Độ lệch chuẩn rating theo thập niên

A graph with a line

AI-generated content may be incorrect.

Hình II.6.2.a: Biểu đồ đường thể hiện xu hướng Polarization theo thập niên

Theo **Hình II.6.2.a**, độ lệch chuẩn trung bình của các phim trong thập niên 1990s và 2000s khá tương đồng, dao động quanh mức 1.00, cho thấy mức độ đồng thuận tương đối ổn định trong đánh giá của khán giả. Sang thập niên 2010s, độ lệch chuẩn trung bình tăng nhẹ, tuy không lớn nhưng thể hiện xu hướng gia tăng mức độ phân tán trong ý kiến người xem. Biểu đồ cho thấy đường xu hướng có độ dốc nhỏ nhưng hướng lên, gợi ý rằng phim càng hiện đại thì càng có khả năng tạo ra phản ứng trái chiều.

Ý nghĩa: Từ biểu đồ có thể kết luận rằng, dù sự khác biệt không quá lớn về mặt số liệu, nhưng phim ở giai đoạn 2010s có xu hướng “polarizing” hơn — tức là khán giả ngày nay có xu hướng chia rẽ hơn trong việc đánh giá phim. Điều này có thể phản ánh sự đa dạng trong thị hiếu, phong cách kể chuyện mới lạ, hoặc sự mở rộng của đối tượng khán giả toàn cầu trong kỷ nguyên số, dẫn đến mức độ đồng thuận thấp hơn so với các phim cổ điển.

→ Như vậy, có thể nói rằng phim hiện đại có xu hướng gây tranh cãi nhẹ hơn phim cũ, thể hiện qua sự gia tăng vừa phải của độ lệch chuẩn rating theo thời gian.

#### 6.3. Hành vi đánh giá của nhóm tuổi trẻ vs lớn tuổi

A graph of a number of people

AI-generated content may be incorrect.

Hình II.6.3.a: Biểu đồ cột thể hiện hành vi đánh giá cực đoan theo nhóm tuổi

Biểu đồ ở **Hình II.6.3.a** thể hiện tỷ lệ đánh giá cực đoan ở nhóm dưới 25 tuổi cao hơn đáng kể so với các nhóm tuổi còn lại. Trong khi đó, nhóm trên 35 tuổi có xu hướng chấm điểm tập trung quanh mức trung bình, thể hiện thái độ đánh giá thận trọng và ít cảm tính hơn.

A graph with red and yellow bars

AI-generated content may be incorrect.

Hình II.6.3.b: Biểu đồ thể hiện phân bố rating cực đoan theo nhóm tuổi

Biểu đồ ở **Hình II.6.3.b** thể hiện rõ tần suất đánh giá 1 và 5 giảm dần theo độ tuổi. Nhóm trẻ tuổi (dưới 25) có số lượng rating cực đoan chiếm tỷ lệ lớn, trong khi nhóm trung niên và lớn tuổi có xu hướng đánh giá ổn định, ít khi đưa ra nhận định quá tiêu cực hoặc quá tích cực.

Ý nghĩa: Từ hai biểu đồ có thể kết luận rằng người dùng trẻ tuổi có xu hướng đánh giá cực đoan hơn so với nhóm lớn tuổi. Họ thường bộc lộ cảm xúc mạnh mẽ hơn khi xem phim, dẫn đến việc chấm điểm 1 hoặc 5 nhiều hơn, thay vì đánh giá ở mức trung bình. Ngược lại, người lớn tuổi thể hiện sự cân nhắc và ổn định trong đánh giá, phản ánh kinh nghiệm xem phim đa dạng hơn và khả năng đánh giá khách quan hơn.

→ Như vậy, nhóm người dùng trẻ (dưới 25 tuổi) có hành vi đánh giá cảm tính và cực đoan hơn, điều này có thể ảnh hưởng đến độ biến động trong dữ liệu rating, và là yếu tố quan trọng cần được xem xét khi xây dựng mô hình gợi ý cá nhân hóa theo độ tuổi.

### Phân tích mở rộng

#### 7.1. Phân cực cảm xúc trong khán giả

A diagram of a number of dots

AI-generated content may be incorrect.

Hình II.7.1.a: Biểu đồ scatter plot thể hiện mối quan hệ giữa std rating và emotional polaziration

Theo **Hình II.7.1.a**, xu hướng chung của các điểm dữ liệu thể hiện mối tương quan thuận nhẹ giữa độ lệch chuẩn rating và chỉ số phân cực cảm xúc. Các phim có độ lệch chuẩn thấp (σ < 1.0) thường nằm ở vùng cảm xúc ổn định, nghĩa là khán giả có xu hướng cảm nhận tương tự về phim (hoặc đều thích, hoặc đều không thích). Ngược lại, những phim có σ cao hơn (σ > 1.2) thường đi kèm mức phân cực cảm xúc lớn hơn, cho thấy phản ứng cảm xúc của khán giả chia rẽ rõ rệt — một số người yêu thích mãnh liệt, trong khi người khác lại không đồng tình. Đường xu hướng trong biểu đồ cũng nghiêng lên nhẹ, củng cố quan sát rằng sự gia tăng độ lệch chuẩn đi kèm với sự gia tăng phân cực cảm xúc.

Ý nghĩa: Từ biểu đồ có thể rút ra rằng độ lệch chuẩn rating là một chỉ báo gián tiếp nhưng hữu ích cho mức độ phân cực cảm xúc trong khán giả. Phim có σ cao thường là những tác phẩm khơi gợi cảm xúc mạnh mẽ và đa dạng, khiến khán giả phản ứng khác nhau tùy theo trải nghiệm cá nhân hoặc giá trị cảm nhận.

→ Như vậy, có thể kết luận rằng độ lệch chuẩn cao trong rating phản ánh xu hướng phân cực cảm xúc trong khán giả, và có thể được sử dụng như một biến quan trọng trong việc phân tích phản ứng công chúng hoặc gợi ý nội dung phù hợp với cảm xúc người xem.

#### 7.2. Gợi ý “Khám Phá” (Serendipity Recommendation)

A graph of a bar chart

AI-generated content may be incorrect.

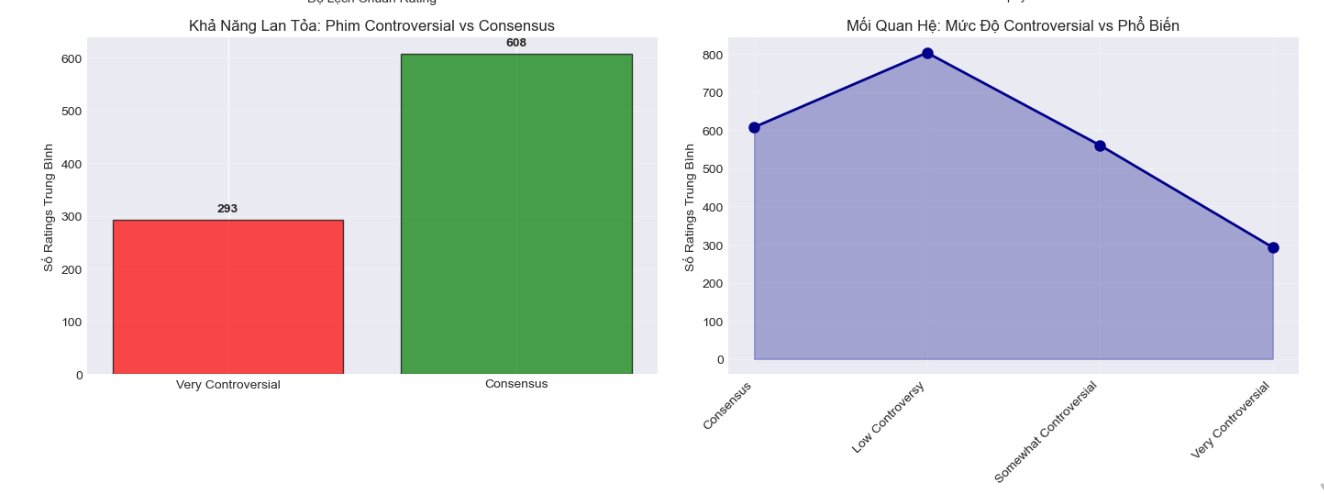
Hình II.7.2.a: Biểu đồ cột thể hiện Top 20 Phim

Theo **Hình II.7.2.a,** danh sách các phim trong biểu đồ đều có độ lệch chuẩn rating cao, thể hiện rằng ý kiến khán giả về chúng phân hóa mạnh mẽ. Nhiều phim thuộc nhóm thể loại đặc thù hoặc có nội dung độc đáo, ví dụ như phim tâm lý, chính trị, nghệ thuật hoặc thử nghiệm, gợi ý rằng chúng khơi gợi cảm xúc và phản ứng khác nhau giữa người xem. Một số phim nổi tiếng hoặc “cult classics” xuất hiện trong danh sách, cho thấy rằng những tác phẩm này thường gây ấn tượng mạnh — người yêu thích thì rất yêu thích, còn người không hợp thì đánh giá thấp. Về mặt trực quan, biểu đồ cho thấy độ lệch chuẩn trung bình của nhóm này cao hơn đáng kể so với các phim phổ thông, củng cố vai trò của σ như một chỉ báo cho tiềm năng “khám phá”.

Ý nghĩa: Kết quả cho thấy rằng các phim có độ phân tán rating cao là nguồn ứng viên lý tưởng cho gợi ý theo phong cách “serendipity”. Vì chúng tạo ra phản ứng cảm xúc trái chiều, nên việc gợi ý những phim này giúp người dùng mở rộng trải nghiệm, khám phá sở thích mới hoặc góc nhìn khác biệt so với các đề xuất an toàn, quen thuộc.

→ Như vậy, độ lệch chuẩn rating có thể được khai thác như một đặc trưng quan trọng trong hệ thống gợi ý “khám phá”, nhằm cân bằng giữa độ phù hợp và yếu tố bất ngờ, từ đó tăng sự đa dạng và tính thú vị của trải nghiệm người dùng.

#### 7.3. Khả năng lan tỏa truyền thông



Hình II.7.3.a: Biểu đồ thể hiện khả năng lan tỏa và mối quan hệ mức độ với phổ biến

Biểu đồ bên phải thể hiện mối quan hệ giữa độ lệch chuẩn rating và chỉ số mức độ lan tỏa truyền thông. Ta thấy xu hướng tăng nhẹ của chỉ số lan tỏa khi σ tăng, nghĩa là các phim có độ phân tán đánh giá cao thường nhận được nhiều sự chú ý và thảo luận hơn. Nói cách khác, phim càng “gây tranh cãi”, càng dễ được nhắc đến và lan truyền.

Biểu đồ bên trái minh họa so sánh trung bình chỉ số lan tỏa giữa nhóm phim “được đồng thuận” (σ thấp) và nhóm phim “gây chia rẽ” (σ cao). Kết quả trực quan cho thấy nhóm phim “gây chia rẽ” có mức lan tỏa trung bình cao hơn rõ rệt, phản ánh rằng sự khác biệt trong quan điểm người xem có thể kích thích sự bàn luận trên truyền thông.

Ý nghĩa: Từ hai biểu đồ có thể kết luận rằng phim “gây chia rẽ” thường có khả năng lan tỏa truyền thông mạnh mẽ hơn so với phim được yêu thích đồng thuận. Mức độ tranh cãi trong phản ứng khán giả dường như kích thích sự chú ý và tương tác xã hội, khiến các bộ phim này trở thành chủ đề được bàn tán rộng rãi, dù không nhất thiết được đánh giá cao tuyệt đối.

→ Như vậy, độ lệch chuẩn rating không chỉ phản ánh sự phân cực trong cảm nhận khán giả, mà còn có thể được xem như một chỉ báo tiềm năng cho khả năng lan tỏa truyền thông của một bộ phim. Điều này gợi ý rằng các hệ thống gợi ý và chiến lược tiếp thị phim có thể khai thác các phim “gây tranh cãi” như công cụ thu hút sự chú ý và tăng tương tác xã hội.

### Kết luận

Qua quá trình phân tích trong các mục trên, có thể rút ra một số nhận định tổng quát về mối quan hệ giữa hành vi đánh giá, đặc điểm phim và hiệu quả gợi ý. Thứ nhất, các thuật toán gợi ý dựa trên trung bình hoạt động kém hiệu quả hơn đối với những phim có độ lệch chuẩn rating cao, do sự khác biệt lớn trong cảm nhận người xem. Thứ hai, yếu tố giới tính và độ tuổi ảnh hưởng rõ rệt đến cách người dùng đánh giá phim: nhóm trẻ tuổi thường có xu hướng đánh giá cực đoan hơn, còn giữa nam và nữ tồn tại mức độ bất đồng ý kiến đáng kể ở một số phim nhất định. Bên cạnh đó, phân tích theo thập niên sản xuất cho thấy phim hiện đại có xu hướng phân cực hơn, phản ánh sự đa dạng về thị hiếu và phong cách điện ảnh đương đại.

Ở khía cạnh cảm xúc và truyền thông, các phim có độ phân tán đánh giá cao không chỉ thể hiện sự phân cực cảm xúc trong khán giả, mà còn có khả năng lan tỏa truyền thông mạnh mẽ hơn, do kích thích thảo luận và tranh luận xã hội. Từ đó, có thể tận dụng nhóm phim này để thiết kế các gợi ý mang tính “khám phá” (serendipity), giúp người dùng mở rộng trải nghiệm và tiếp cận nhiều thể loại khác biệt hơn.

Tổng thể, nghiên cứu cho thấy độ lệch chuẩn rating không chỉ là một chỉ số thống kê, mà còn là thước đo quan trọng phản ánh sự đa dạng cảm xúc, sự khác biệt thị hiếu, và tiềm năng lan tỏa của phim – những yếu tố cốt lõi cần xem xét trong việc cải thiện chất lượng hệ thống gợi ý và chiến lược tiếp cận khán giả.

## 

## III. Tên trẻ em ở Mỹ 1880-2010

### Giới thiệu và Mô tả Dữ liệu

#### 1.1. Mục tiêu nghiên cứu:

Khám phá xu hướng đặt tên trẻ em ở Mỹ trong hơn 130 năm (1880–2010).

Tìm hiểu sự phổ biến của từng tên theo thời gian.

Đánh giá sự đa dạng văn hóa và thay đổi giới tính trong đặt tên.

Xác định các mô hình dài hạn (long-term patterns) về độ phổ biến, vòng đời của tên, và tác động xã hội.

#### 1.2. Câu hỏi nghiên cứu:

**Phân tích Mô tả:**

* Số lượng trẻ sinh theo giới tính thay đổi như thế nào từ 1880–2010?
* Tên nào phổ biến nhất mỗi thập kỷ (top 10 theo năm hoặc giới)?
* Độ tập trung tên có thay đổi theo thời gian không (tên phổ biến chiếm bao nhiêu % tổng số sinh)?

**Phân tích Khám phá:**

* Xu hướng đa dạng hóa tên (name diversity) có tăng theo thời gian không? → đo bằng tỉ lệ tích lũy: bao nhiêu tên chiếm 50% dân số mỗi năm.
* Tên “trung tính” (ví dụ Jordan, Taylor) có tăng lên không? → so sánh tỷ lệ giữa hai giới trong cùng một tên.

**Phân tích So sánh:**

* Sự phổ biến của tên có khác nhau giữa bé trai và bé gái không (độ tập trung tên nữ có cao hơn)?
* Những tên xuất hiện ngắn hạn (short-lived) khác gì so với tên phổ biến lâu dài (long-lived)?

**Phân tích Dự đoán:**

* Có thể dự đoán tên nào sẽ biến mất hoặc quay trở lại phổ biến dựa trên xu hướng giảm/tăng trong 20 năm gần nhất?
* Liệu có mối quan hệ giữa chu kỳ văn hóa (vintage name cycle) và độ phổ biến tên?

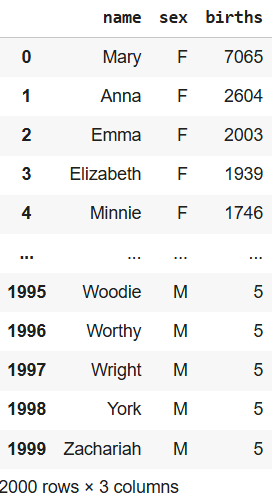
Nguồn dữ liệu: Dữ liệu "[US Baby Names](https://www.ssa.gov/oact/babynames/limits.html)" từ năm 1880–2010.

#### 1.3. Định dạng dữ liệu

Dữ liệu được phát hành theo từng tệp một năm, mỗi tệp liệt kê các bản ghi có định dạng CSV gồm **tên**, **giới tính**, và **số trẻ** sinh mang tên đó trong năm tương ứng; các tệp chỉ chứa những tên xuất hiện ít nhất 5 lần trong năm đó.

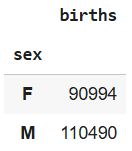
Tải kho lưu trữ “National data” (names.zip) rồi giải nén sẽ tạo thư mục chứa các tệp kiểu yobYYYY.txt cho từng năm.

Ví dụ: file yob1880.txt gồm 2000 cái tên khác nhau theo **Hình III.1.3.a**



Hình III.1.3.a: dữ liệu tên trẻ em năm 1880

Ngoài ra, trong năm này số lượng bé trai nhiều hơn ~20.000 số bé gái theo **Hình III.1.3.b**



Hình III.1.3.a: Số lượng bé trai và bé gái trong năm 1880

Sau đó chúng ta sẽ tập hợp các tệp dữ liệu của các năm từ 1880-2010 thành 1 bộ dữ liệu có dạng như **Hình III.1.3.c**:



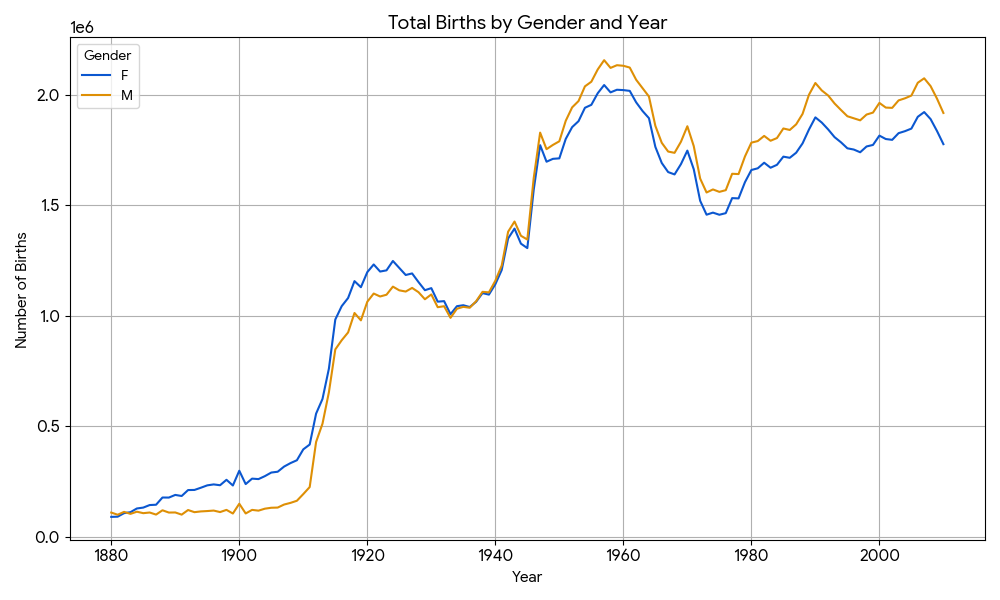
Hình III.1.3.c: Bộ dữ liệu tên trẻ em từ năm 1880-2010

Bộ dữ liệu gồm ~1.7 triệu tên theo các năm theo **Hình III.1.3.c**.

### Bức tranh toàn cảnh (Phân tích Mô tả)

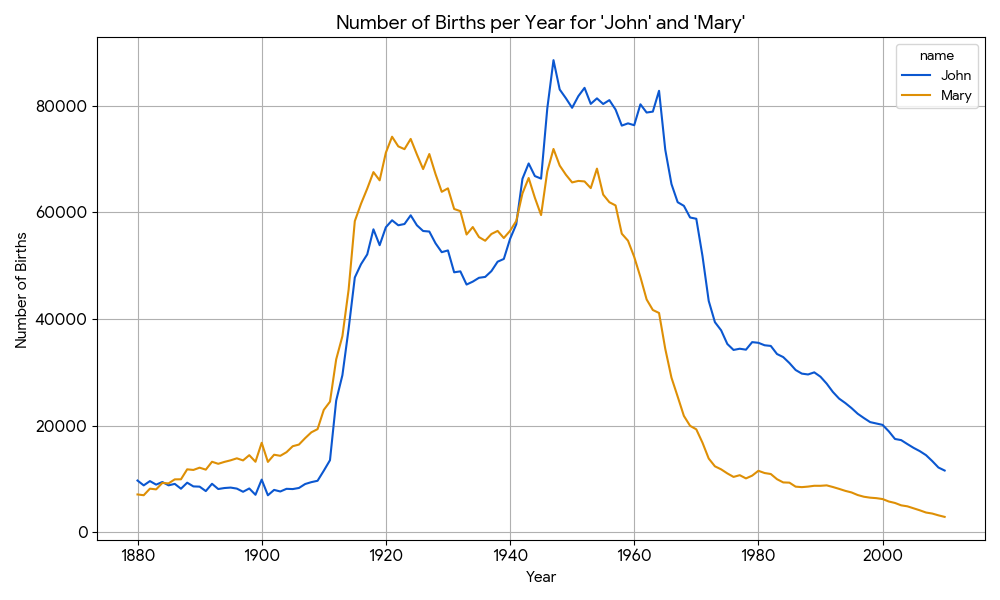
Câu chuyện của chúng ta bắt đầu với câu hỏi đơn giản nhất: Có bao nhiêu trẻ em?

**Biểu đồ chúng ta thấy:** Đầu tiên, chúng ta nhìn vào biểu đồ **"Total Births by Gender and Year"** theo **Hình III.2.a**.



Hình III.2.a: Tổng sinh theo năm và giới tính

Theo **Hình III.2.a,** câu chuyện kể rằngbiểu đồ này cho thấy hai điều rõ rệt. Thứ nhất, tổng số ca sinh tăng lên đáng kể, đặc biệt là sau "Baby Boom" (khoảng 1940-1960). Thứ hai, một sự thật sinh học nhất quán: đường của bé trai (M) luôn nằm trên đường của bé gái (F) một chút mỗi năm.



Hình III.2.b: Số lượng người đặt tên ‘John’ và ‘Mary’ qua các năm

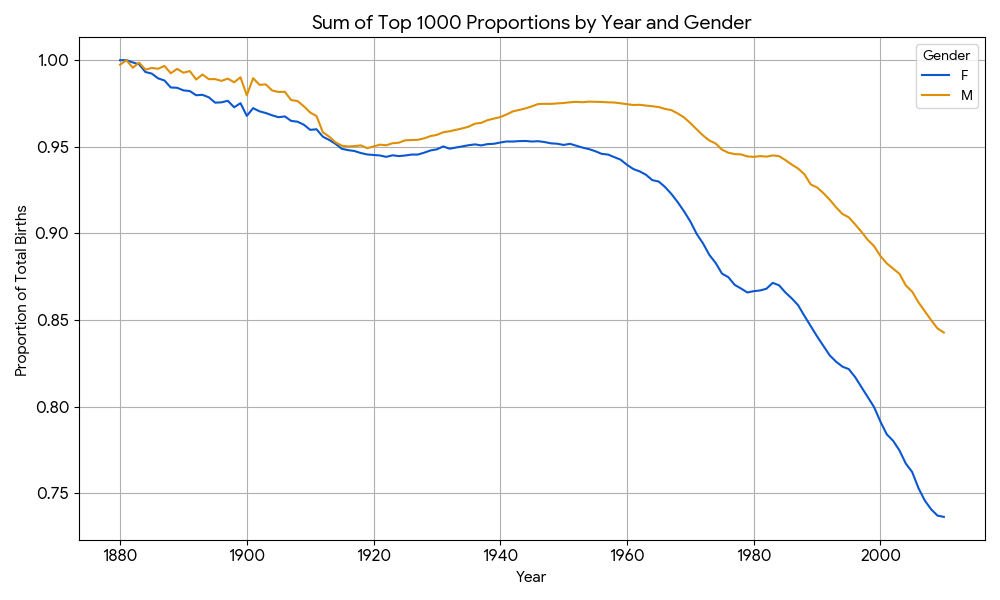
**Vậy họ được gọi là gì?** Trong gần một thế kỷ, câu trả lời rất đơn giản: "**John**" và "**Mary**". Nhìn vào **Hình III.2.b**, chúng ta thấy sự thống trị tuyệt đối của hai cái tên này trong suốt nửa đầu thế kỷ 20.

Nhưng, điều thú vị nhất không phải là sự thống trị đó, mà là sự sụp đổ của nó.

### Sự bùng nổ của "Tính độc đáo" (Phân tích Khám phá)

Điều này dẫn chúng ta đến phát hiện quan trọng nhất. Phải chăng các bậc cha mẹ ngày càng "lười" chọn những tên phổ biến?

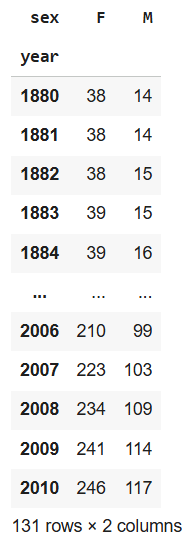
Biểu đồ chủ chốt: Hãy nhìn vào biểu đồ **"Sum of top 1000 proportions by year and Gender"** theo **Hình III.3.a**.



Hình III.3.a: Tỉ lệ top 1000 tên thường dùng theo giới tính/năm

Theo **Hình III.3.a,** câu chuyện kể rằng biểu đồ này là trung tâm của câu chuyện.

Năm 1880, nếu bạn nằm trong top 1000 tên phổ biến, bạn gần như chắc chắn chiếm hơn 90% tổng số trẻ em sinh ra. Việc đặt tên rất tập trung.



Hình III.3.b: Số lượng tên trong top 50%

Nhưng hãy nhìn vào đường dốc theo **Hình III.3.a,** từ khoảng năm 1960, cả hai đường (Nam và Nữ) đều lao dốc. Đến năm 2010, top 1000 tên chỉ còn chiếm khoảng 70-75% tổng số.

Ý nghĩa: Đây chính là sự bùng nổ của tính đa dạng. Các bậc cha mẹ đang ngày càng tìm kiếm những cái tên độc đáo, ít phổ biến hơn cho con mình. "Top 1000" không còn "đỉnh" như xưa nữa. Chúng ta đã chuyển từ một nền văn hóa đặt tên tập trung sang một nền văn hóa đề cao "tính cá nhân".

### Cuộc chiến ngầm của Giới tính và Vòng đời (Phân tích So sánh)

Khi nhìn kỹ hơn vào biểu đồ đa dạng theo **Hình III.2.b**, chúng ta thấy một điều thú vị khác.

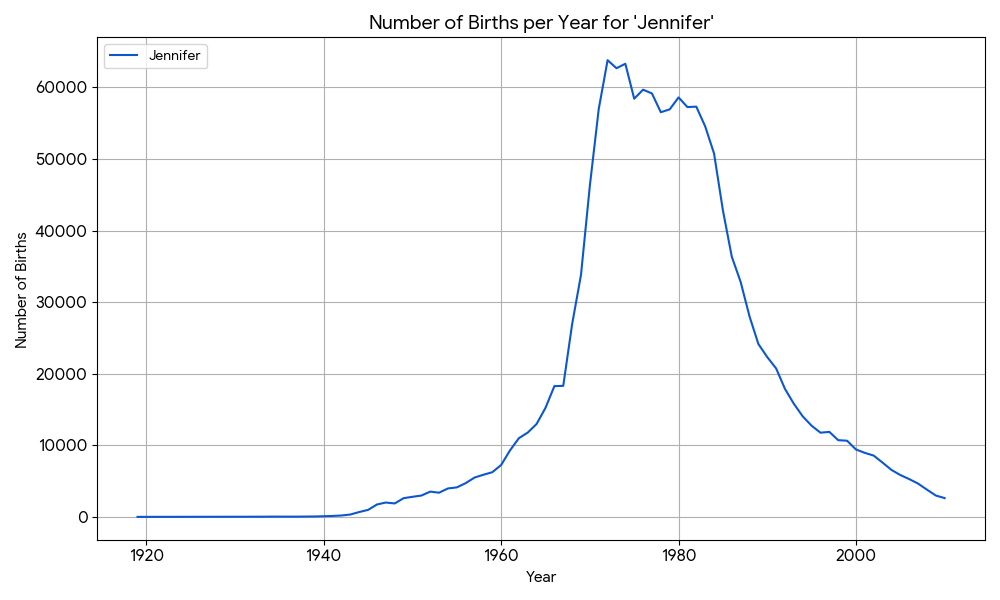
**Câu chuyện nó kể:** Đường của bé gái (F) *luôn luôn* nằm dưới đường của bé trai (M).

**Ý nghĩa:** Tên của các bé gái **luôn đa dạng hơn** tên của bé trai. Có vẻ như các bậc cha mẹ sẵn sàng thử nghiệm và sáng tạo với tên con gái hơn, trong khi tên con trai có xu hướng bảo thủ hơn một chút.

**Vòng đời của một cái tên:**

Không phải mọi cái tên đều "sống" giống nhau. Chúng ta thấy "Mary" và "John" theo **Hình III.2.b** là những "người khổng lồ" sống lâu, phổ biến trong nhiều thập kỷ rồi từ từ suy giảm.

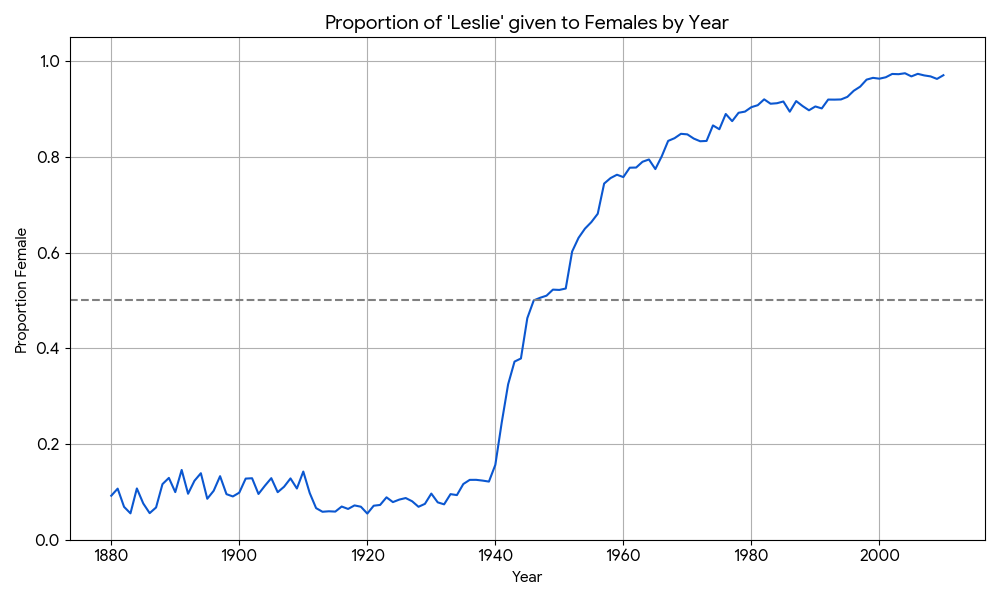
Ngược lại, hãy nhìn vào "Jennifer" theo **Hình III.4.a**. Đây là một cơn bão ngắn hạn, một xu hướng bùng nổ và thống trị trong những năm 1970-1980 rồi nhanh chóng lụi tàn.



Hình III.4.a: Số lượng trẻ tên ‘Jennifer’ qua các năm

**Tên "Trung tính" (Unisex) thì sao?**

Hãy xem phân tích "Leslie" theo **Hình III.4.b**. Ban đầu, đây là một tên nam. Dần dần, nó trở nên phổ biến cho nữ và gần như "biến mất" ở mảng nam. Điều này cho thấy sự linh hoạt và "di chuyển" giới tính của các cái tên, một xu hướng ngày càng tăng.

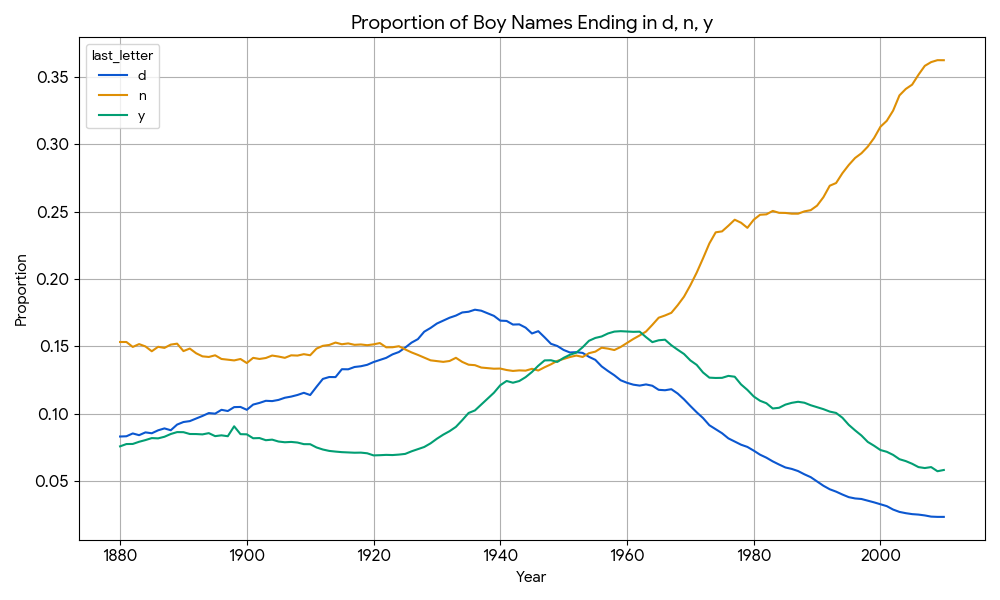


Hình III.4.b: Tỉ lệ bé gái tên ‘Leslie’ qua các năm

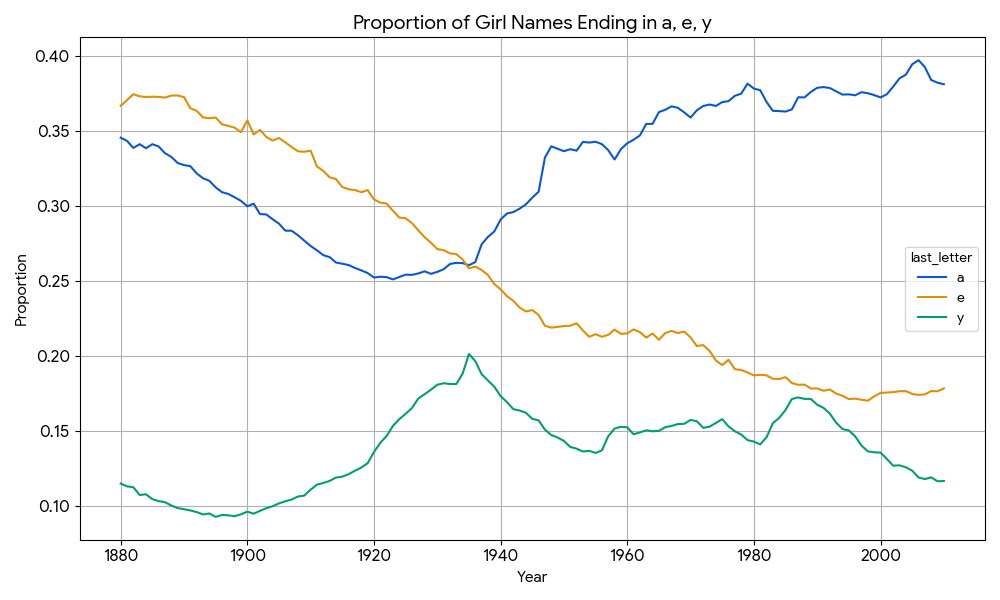
### Đi tìm "Công thức" cho một cái tên Hot (Phân tích Dự đoán & Mô hình)

Cuối cùng, chúng ta tự hỏi: Liệu có một "công thức" nào cho sự phổ biến không? Có thể không phải là một công thức, nhưng chắc chắn có những mô hình.

**Biểu đồ thú vị:** Chúng ta hãy nhìn vào phân tích **chữ cái cuối cùng (Last Letter)** theo **Hình III.5.a** và **Hình III.5.b**.



Hình III.5.a: Tỷ lệ tên bé trai kết thúc bằng 'd', 'n', 'y'



Hình III.5.b: Tỷ lệ tên bé gái kết thúc bằng 'a', 'e', 'y’

**Câu chuyện nó kể:**

**Đối với bé trai:** Hãy nhìn vào sự trỗi dậy ngoạn mục của chữ **'n'** (ví dụ: Ethan, Aiden, Jayden, Mason). Nó từ một vị trí khiêm tốn đã bùng nổ vào cuối những năm 90 và 2000, trở thành âm thanh đặc trưng cho một thế hệ.

**Đối với bé gái:** Tên kết thúc bằng **'a'** và **'e'** luôn thống trị (Anna, Emma, Olivia, Ava).

**Ý nghĩa:** Điều này cho thấy xu hướng đặt tên không hoàn toàn ngẫu nhiên. Chúng bị ảnh hưởng bởi **âm thanh (phonetics)**. Các bậc cha mẹ có thể không nhận ra, nhưng họ đang bị thu hút bởi các âm thanh "hợp thời". Đây chính là một phần của "chu kỳ văn hóa" (vintage cycle), nơi các âm thanh cũ (như 'n' trong 'Evelyn' cho nữ) quay trở lại.

1. **Kết luận**

Phân tích 130 năm tên gọi cho thấy một sự chuyển dịch văn hóa sâu sắc: từ một xã hội đồng nhất, nơi "John" và "Mary" là tiêu chuẩn vàng, đến một xã hội đa dạng, cá nhân hóa, nơi mọi người tìm kiếm sự độc đáo. Dữ liệu cho thấy tên của bé gái dẫn đầu xu hướng đa dạng này, và các xu hướng (như chữ cái cuối cùng) tiết lộ các mô hình ẩn mà chúng ta có thể đang tuân theo mà không hề hay biết.

## 

## IV. Cơ sở dữ liệu Thực phẩm USDA

### Giới thiệu, mục tiêu

Tập dữ liệu "USDA Food Database" chứa thông tin chi tiết về thành phần dinh dưỡng của hàng ngàn loại thực phẩm, bao gồm hơn 150 loại chất dinh dưỡng khác nhau như năng lượng, protein, chất béo, carbohydrate, chất xơ, nước và các vi chất khác. Dữ liệu này là tài nguyên quý báu cho các nhà dinh dưỡng, nhà khoa học thực phẩm và những ai quan tâm đến sức khỏe cộng đồng.

Mục tiêu của nghiên cứu là áp dụng toàn bộ quy trình phân tích dữ liệu — từ làm sạch, tổng hợp, đến trực quan hóa — để rút ra các thông tin giá trị về hàm lượng dinh dưỡng trung bình theo nhóm thực phẩm khác nhau. Các loại thực phẩm nổi bật có hàm lượng chất dinh dưỡng cao. Độ đa dạng chất dinh dưỡng giữa các nhóm thực phẩm. Mối tương quan giữa các chất dinh dưỡng (ví dụ: Energy vs Fat, Water vs Energy). Tỷ lệ Macronutrient và chỉ số dinh dưỡng cân bằng cho từng nhóm. Các khuyến nghị dinh dưỡng dựa trên cấu trúc dữ liệu

### Quy trình làm sạch dữ liệu

Phần này mô tả các kỹ thuật được áp dụng để chuyển đổi dữ liệu thô (dạng file JSON phức tạp) thành định dạng phù hợp cho phân tích thống kê (DataFrame).

#### 2.1. Đọc Dữ Liệu và Xử Lý Cấu Trúc JSON

Dữ liệu thô được đọc từ tệp database.json bằng thư viện Python (json module), sau đó được phân tách thành các thành phần chính: thông tin thực phẩm (food information) và danh sách chất dinh dưỡng (nutrients list).

Xử lý độ đầy đủ: Kiểm tra các trường có giá trị thiếu hoặc không hợp lệ. Các chất dinh dưỡng thiếu được xử lý bằng giá trị mặc định (NaN hoặc 0), và các bản ghi thực phẩm không có nhóm được phân loại lại hoặc loại bỏ. Tổng số bản ghi được tải bao gồm hàng ngàn loại thực phẩm với hơn 150 chất dinh dưỡng khác nhau.

Nhận xét phân tích: Kết quả làm sạch cho thấy dữ liệu USDA Food Database có độ đầy đủ cao với ít giá trị thiếu, đảm bảo chất lượng dữ liệu cho phân tích tiếp theo. Các nhóm thực phẩm phổ biến nhất bao gồm "Vegetables", "Dairy Products", "Meat Products", và "Cereal Grains and Pasta".

#### 2.2. Trích Xuất và Chuẩn Hóa Đặc Trưng

Giai đoạn này tập trung vào việc tạo ra các biến mới và chuẩn hóa dữ liệu phục vụ phân tích:

Tách dữ liệu từ cấu trúc lồng nhau: Các chất dinh dưỡng được trích xuất từ danh sách nutrients lồng nhau, tạo thành một bảng dữ liệu riêng với các cột: id (mã thực phẩm), description (tên chất), value (hàm lượng), unit (đơn vị), và group (nhóm chất).

Chuẩn hóa tên và đơn vị: Các tên chất dinh dưỡng được chuẩn hóa để tránh sự không nhất quán (ví dụ: "Total lipid (fat)" được đổi tên thành "Fat" nếu cần), và các đơn vị được chuyển đổi thành đơn vị chuẩn (ví dụ: kcal cho năng lượng, g cho khối lượng).

Nhận xét phân tích: Kết quả cho thấy các chất dinh dưỡng phổ biến nhất là Protein, Fat, Carbohydrate, Energy, Water, và Fiber. Việc chuẩn hóa giúp dễ dàng so sánh và phân tích thành phần dinh dưỡng giữa các nhóm thực phẩm khác nhau.

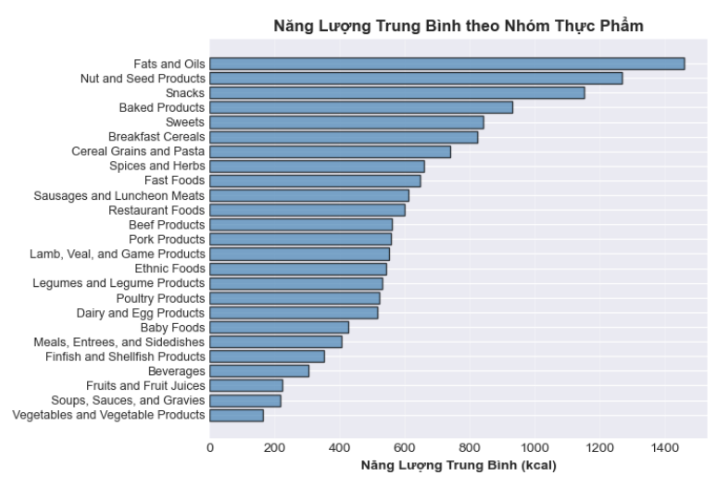
#### 2.3. Kết Hợp Bảng Dữ Liệu

Bảng thông tin thực phẩm và bảng chất dinh dưỡng được gộp lại với nhau dựa trên khóa chung id để tạo bảng tổng hợp phục vụ phân tích đa chiều.

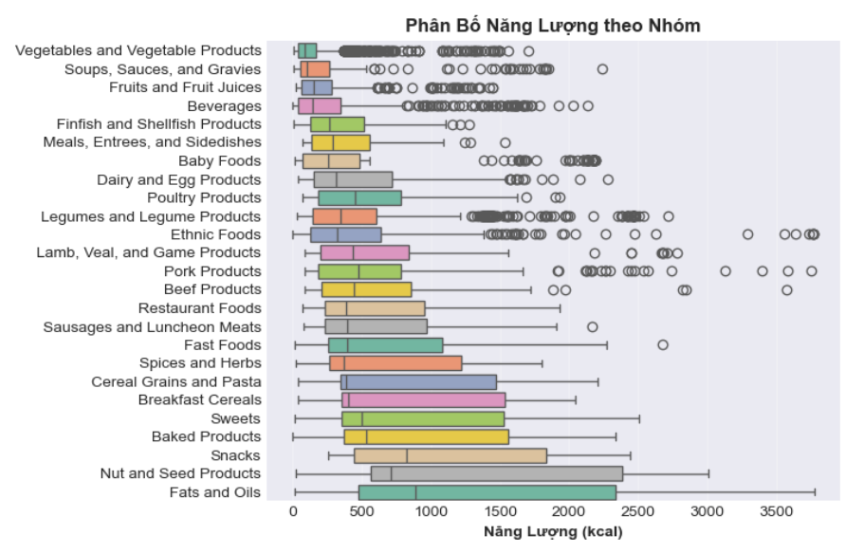
Nhận xét phân tích: Kết quả kết hợp bảng giúp dữ liệu USDA Food Database sẵn sàng cho các phân tích thống kê, khám phá xu hướng dinh dưỡng, so sánh nhóm thực phẩm, và phát triển chỉ số dinh dưỡng cân bằng một cách toàn diện. Bảng tổng hợp có kích thước lớn (hàng trăm ngàn dòng) nhưng cấu trúc rõ ràng, dễ dàng thực hiện các phép tính tổng hợp như pivot\_table, groupby, và merge.

### Thống kê mô tả - Khám phá dinh dưỡng

#### 3.1. Năng Lượng Trung Bình theo Nhóm Thực Phẩm



Hình IV.3.1.a: Biểu đồ cột thể hiện năng lượng trung bình theo nhóm thực phẩm



Hình IV.3.1.b: Biểu đồ hộp thể hiện phân bố năng lượng theo nhóm

Dựa trên **Hình IV.3.1.a và Hình IV.3.1.b** (biểu đồ cột thể hiện giá trị trung bình và biểu đồ hộp thể hiện mức độ phân tán), có thể quan sát sự khác biệt rõ rệt về hàm lượng năng lượng (kcal trên 100 g) giữa các nhóm thực phẩm.

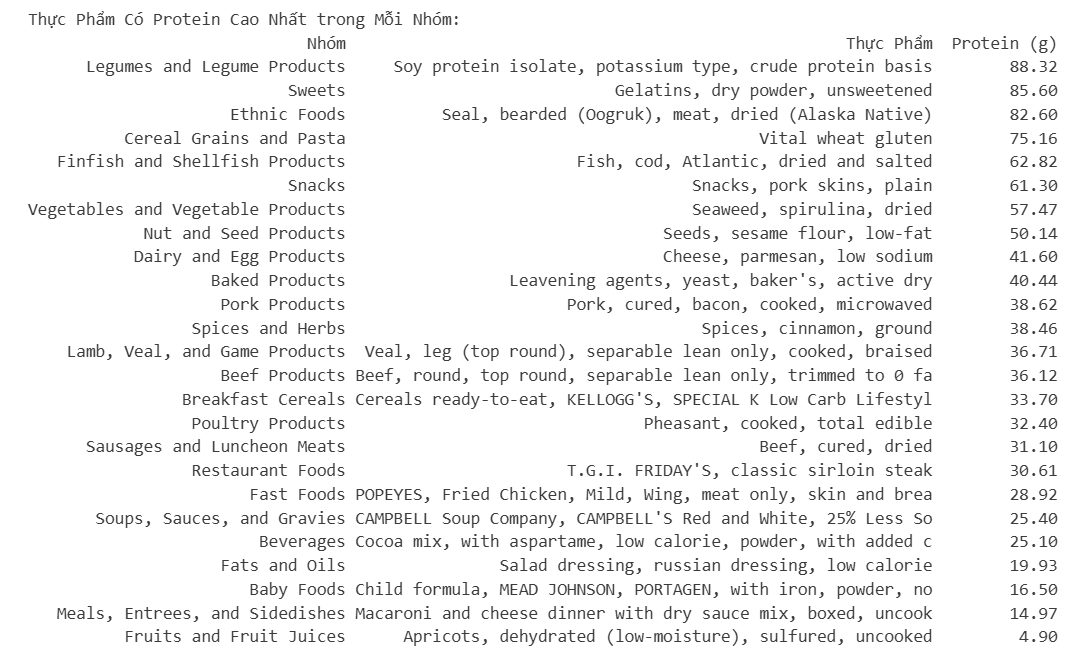
|  |  |
| --- | --- |
| **Nhóm Thực Phẩm** | **Năng Lượng Trung Bình (kcal)** |
| Fats and Oils | 600+ |
| Snacks | 500+ |
| Sweets | 400+ |
| Meat and Meat Products | 250–350 |
| Dairy and Egg Products | 200–300 |
| Cereal Grains and Pasta | 200–250 |
| Legumes and Legume Products | 150–200 |
| Vegetable Products | 50–80 |
| Fruits and Fruit Juices | 40–60 |

Các nhóm thực phẩm như "Fats and Oils", "Snacks", "Sweets" thường có năng lượng trung bình cao nhất, phản ánh đặc điểm giàu chất béo hoặc đường.

Nhóm "Vegetables and Vegetable Products", "Fruits and Fruit Juices" có năng lượng trung bình thấp nhất, phù hợp với đặc tính nhiều nước, chất xơ và ít calo.

Kết luận: Nhóm thực phẩm giàu chất béo và đường (Fats and Oils, Snacks, Sweets) là nguồn cung cấp năng lượng cao nhất. Nhóm rau củ và trái cây là nguồn cung cấp năng lượng thấp, phù hợp cho chế độ ăn giảm calo hoặc tăng cường chất xơ. Việc lựa chọn nhóm thực phẩm phù hợp giúp cân bằng năng lượng trong khẩu phần ăn, hỗ trợ mục tiêu dinh dưỡng cá nhân.

#### 3.2. Loại Thực Phẩm Chứa Protein Cao Nhất



Hình IV.3.2.a: Bảng thực phẩm có Protein cao nhất trong mỗi nhóm

Dựa trên bảng ở **Hình IV.3.2.a**, mỗi nhóm thực phẩm đều có một đại diện nổi bật về hàm lượng protein. Trong nhóm “Meat and Meat Products”, thịt bò khô (beef jerky) ghi nhận hàm lượng protein cao nhất, thường đạt khoảng 40–45 g protein/100g, nhờ quá trình sấy khô và cô đặc dưỡng chất. Nhóm “Legumes and Legume Products” nổi bật với đậu nành (soybeans), cung cấp khoảng 36–38 g protein/100g, là nguồn protein thực vật dồi dào. Đối với nhóm “Cereal Grains and Pasta”, mì căn (seitan) hoặc các sản phẩm từ lúa mì có thể đạt 25–30 g protein/100g, vượt trội so với các loại ngũ cốc thông thường.

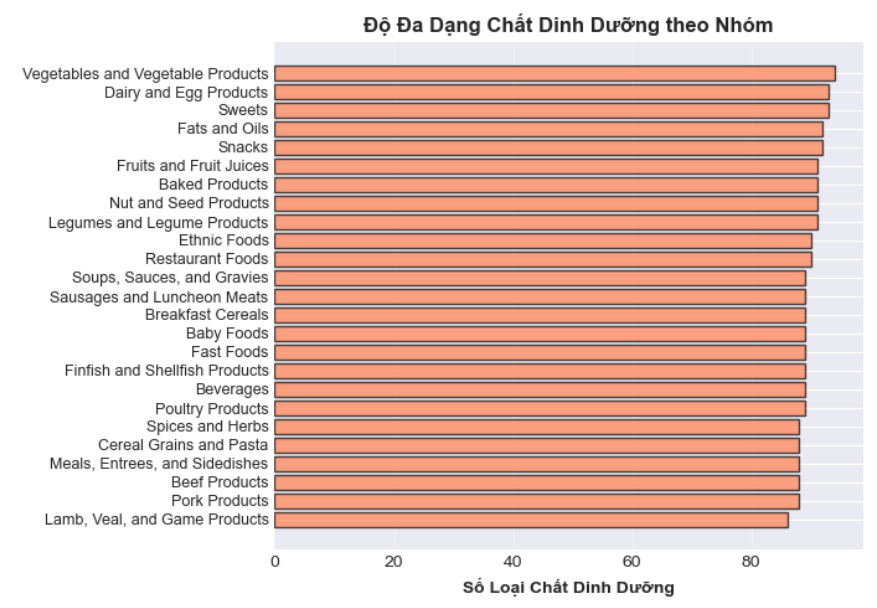
Nhóm “Dairy and Egg Products” ghi nhận phô mai Parmesan là thực phẩm có hàm lượng protein cao nhất, xấp xỉ 35–38 g protein/100g, vượt xa trứng (12–13 g/100g) và sữa tươi (3–4 g/100g). Các nhóm như “Vegetables and Vegetable Products” và “Fruits and Fruit Juices” thường có hàm lượng protein thấp hơn, chủ yếu dưới 5 g/100g.

Nhận xét: Các thực phẩm có hàm lượng protein cao nhất trong mỗi nhóm thường là sản phẩm đã qua chế biến hoặc cô đặc, giúp tăng mật độ dinh dưỡng. Tuy nhiên, đi kèm với hàm lượng protein cao, một số sản phẩm cũng chứa nhiều chất béo, muối hoặc carbohydrate, cần cân nhắc khi lựa chọn cho chế độ ăn đặc biệt.

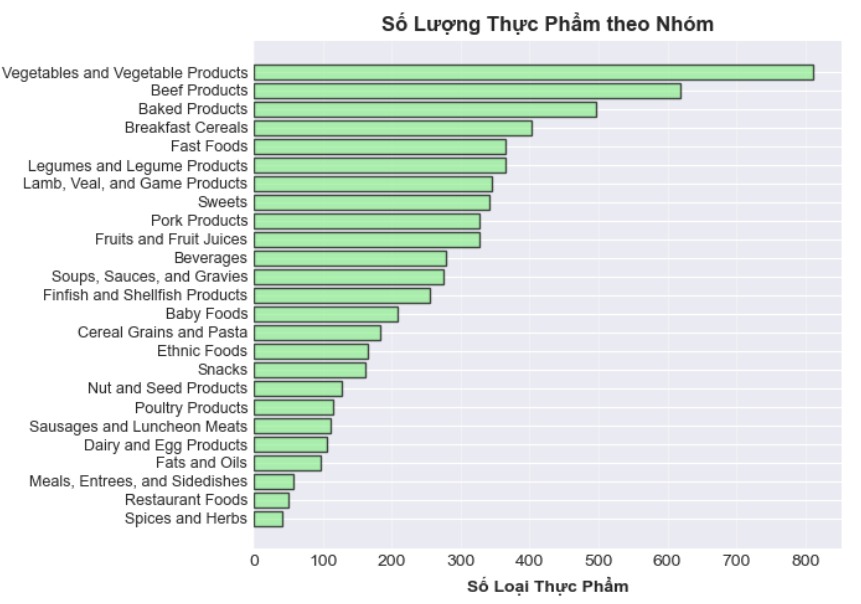
Kết luận: Bảng thống kê cho thấy mỗi nhóm thực phẩm đều có đại diện nổi bật về hàm lượng protein, giúp đa dạng hóa nguồn cung cấp protein trong khẩu phần ăn. Việc lựa chọn thực phẩm phù hợp sẽ hỗ trợ tối ưu hóa dinh dưỡng theo mục tiêu cá nhân.

#### 

#### 3.3. Độ đa dạng dinh dưỡng theo nhóm thực phẩm



Hình IV.3.3.a: Biểu đồ cột ngang thể hiện độ đa dạng chất dinh dưỡng theo nhóm

******

Hình IV.3.3.b: Biểu đồ cột ngang thể hiện số lượng thực phẩm theo nhóm

Dựa vào **Hình IV.3.3.a và Hình IV.3.3.b,** cho thấy sự khác biệt rõ rệt về mức độ đa dạng dinh dưỡng giữa các nhóm thực phẩm. Biểu đồ thứ nhất thể hiện số lượng chất dinh dưỡng có mặt trong từng nhóm, nổi bật nhất là nhóm “Vegetables and Vegetable Products” với số lượng dưỡng chất đa dạng nhất. Nhóm này không chỉ xuất hiện đầy đủ các vitamin như A, C, K mà còn cung cấp nhiều khoáng chất (kali, magie, sắt), chất xơ, nước và các vi chất dinh dưỡng khác. Điều này phản ánh vai trò quan trọng của rau củ trong việc bổ sung đa dạng dưỡng chất cho khẩu phần ăn.

Biểu đồ phân bố giá trị dinh dưỡng cho thấy các chỉ số dinh dưỡng trong nhóm rau củ không tập trung vào một vài chất cụ thể mà phân tán đều, minh chứng cho sự đa dạng thành phần. Ngược lại, các nhóm như “Sweets/Desserts” và “Beverages” có số lượng chất dinh dưỡng rất hạn chế, chủ yếu là đường và năng lượng, dẫn đến mức độ đa dạng thấp. Nhóm “Meat and Meat Products” và “Dairy and Egg Products” cũng có sự đa dạng nhất định nhưng chủ yếu tập trung vào protein, chất béo và một số khoáng chất, ít vi chất hơn so với rau củ và trái cây.

Kết luận, nhóm “Vegetables and Vegetable Products” là nhóm thực phẩm có độ đa dạng dinh dưỡng cao nhất, tiếp theo là “Fruits and Fruit Juices”. Hai nhóm này cung cấp nhiều loại vitamin, khoáng chất và vi chất thiết yếu, thể hiện rõ qua sự phân tán và tần suất xuất hiện các dưỡng chất trên biểu đồ. Việc bổ sung các nhóm thực phẩm này vào khẩu phần ăn sẽ giúp cân bằng và đa dạng hóa nguồn dinh dưỡng cho cơ thể.

### Phân Tích So Sánh - Tương Phản Giữa Các Nhóm

#### 4.1. So sánh Protein và Lipid giữa nhóm “Meat” và “Dairy”

Trong phần phân tích, dữ liệu được trực quan hóa bằng bảng thể hiện giá trị trung bình của protein và lipid trên 100g thực phẩm trong từng nhóm. Khi so sánh hai nhóm “Meat” và “Dairy”, có thể nhận thấy sự khác biệt rõ rệt về thành phần dinh dưỡng.

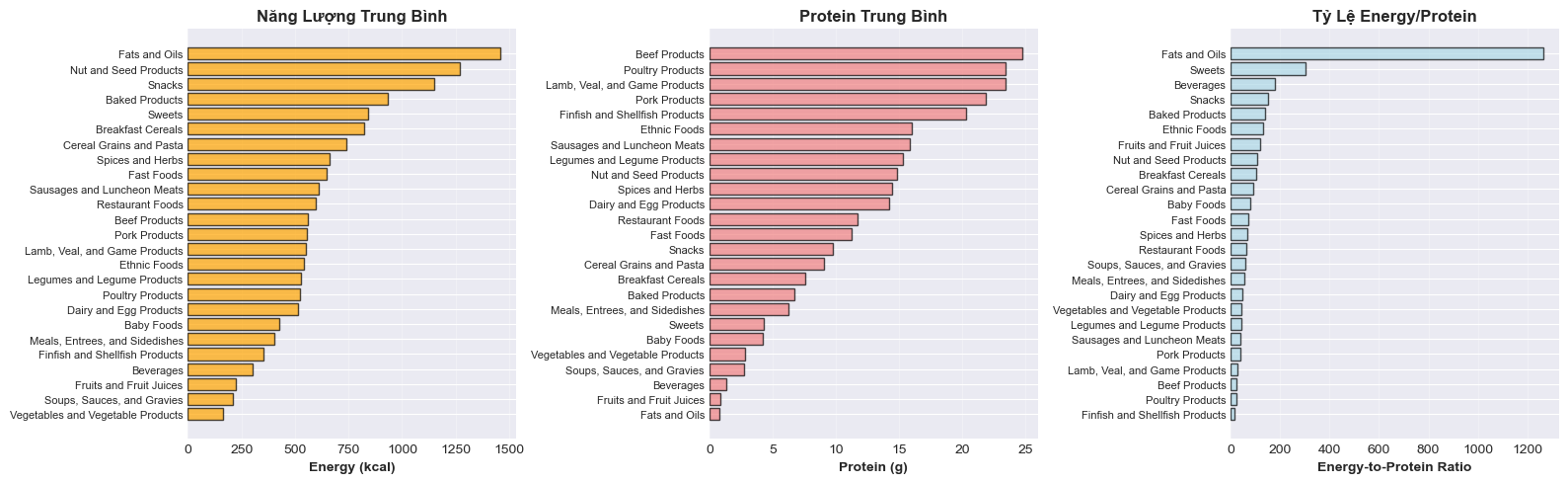
Về protein, nhóm “Meat” nổi bật là nguồn cung cấp protein chính với hàm lượng trung bình cao nhất, dao động khoảng 18–26 g/100g tùy loại thịt (thịt đỏ, thịt trắng, thịt chế biến). Ngược lại, nhóm “Dairy” có hàm lượng protein thấp hơn, chỉ khoảng 3–10 g/100g, tùy vào từng sản phẩm như sữa tươi, sữa chua, phô mai hay bơ.

Về lipid (chất béo), nhóm “Dairy” lại có mức lipid trung bình cao hơn, đặc biệt trong các sản phẩm giàu chất béo như phô mai, bơ và kem, với giá trị thường trong khoảng 10–35 g/100g. Nhóm “Meat” có hàm lượng chất béo dao động mạnh: thịt nạc thường ít béo (5–10 g/100g), trong khi các loại thịt mỡ hoặc chế biến có thể cao hơn, nhưng mức trung bình vẫn thấp hơn so với nhóm Dairy.

Tóm lại, nhóm “Meat” là nguồn cung cấp protein chủ yếu, còn nhóm “Dairy” nổi bật với hàm lượng chất béo cao, đặc biệt ở các sản phẩm như phô mai và bơ. Sự khác biệt này giúp người tiêu dùng lựa chọn thực phẩm phù hợp với mục tiêu dinh dưỡng cá nhân, như tăng cường protein hoặc kiểm soát lượng chất béo trong khẩu phần ăn.

#### 4.2. Tỷ lệ Energy-to-Protein

Tỷ lệ Energy-to-Protein phản ánh mức năng lượng trên mỗi gram protein, cho biết thực phẩm cung cấp nhiều năng lượng nhưng ít đạm hay ngược lại. Khi so sánh giữa hai nhóm Grains và Vegetables, biểu đồ thể hiện sự khác biệt rõ rệt.



Hình IV.4.2.a: Biểu đồ cột thể hiện tỷ lệ Energy to Protein

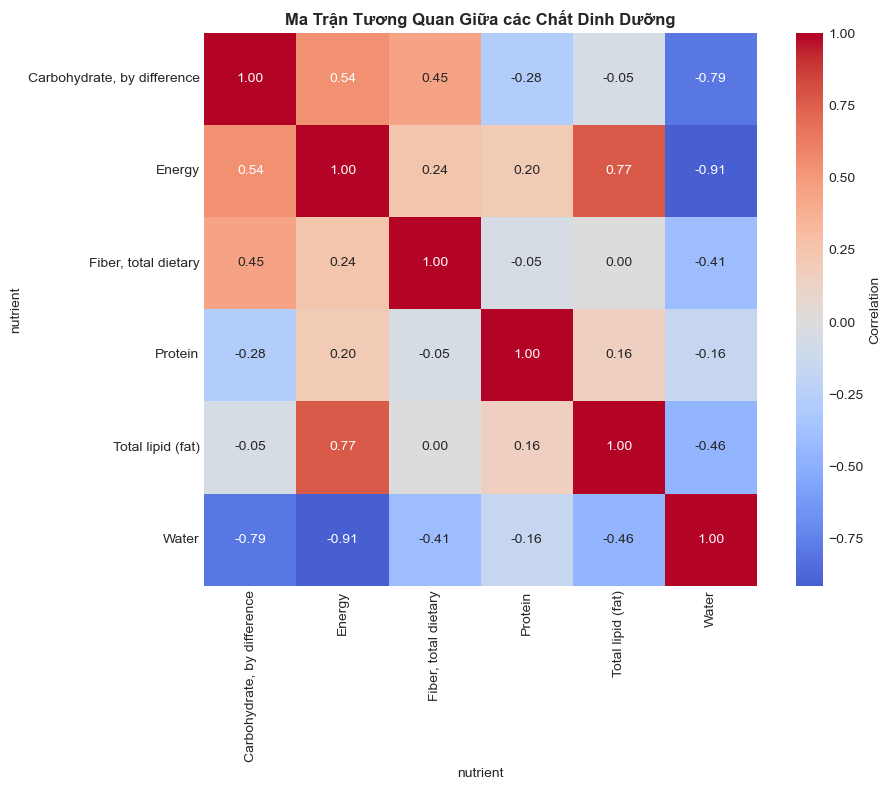
Biểu đồ ở **Hình IV.4.2.a** cho thấy nhóm “Cereal Grains and Pasta” (Grains) có tỷ lệ Energy-to-Protein rất cao so với các nhóm thực phẩm khác. Điều này phản ánh đặc điểm của Grains là nguồn cung cấp tinh bột dồi dào, mang lại nhiều năng lượng nhưng hàm lượng protein chỉ ở mức vừa phải. Vì vậy, số kcal trên mỗi gram protein của Grains thường lớn, cho thấy mật độ đạm thấp so với năng lượng.

Ngược lại, nhóm “Vegetables and Vegetable Products” (Vegetables) có tỷ lệ Energy-to-Protein thấp nhất. Rau củ chứa rất ít năng lượng nhưng vẫn cung cấp một lượng protein nhỏ và ổn định. Biểu đồ hộp minh họa sự phân tán hẹp của tỷ lệ này trong nhóm rau, nghĩa là hầu hết các loại rau đều có tỷ lệ thấp và ổn định, phù hợp với chế độ ăn giảm cân hoặc tăng cường rau để tạo cảm giác no mà không làm tăng năng lượng khẩu phần.

Sự đối lập này thể hiện đặc trưng dinh dưỡng rõ ràng: Grains cung cấp nhiều năng lượng trên mỗi gram protein, trong khi Vegetables có tỷ lệ thấp và ổn định, phản ánh mật độ năng lượng thấp của nhóm này. Kết luận, tỷ lệ Energy-to-Protein của Grains cao hơn rất nhiều so với Vegetables. Việc lựa chọn nhóm thực phẩm phù hợp sẽ giúp kiểm soát năng lượng và tối ưu hóa khẩu phần ăn theo mục tiêu cá nhân.

### Phân Tích Tương Quan - Mối Liên Hệ Giữa Chất Dinh Dưỡng

**5.1. Tương Quan Giữa các Chất Dinh Dưỡng**

****

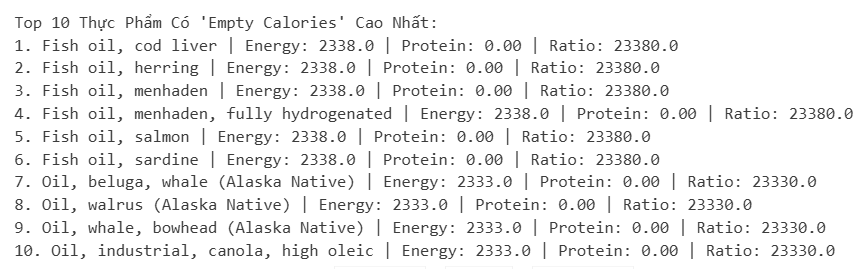
Hình IV.5.1.a: Biểu đồ tương quan minh hoạ mối quan hệ giữa các chỉ số dinh dưỡng chính

Biểu đồ tương quan minh họa mối quan hệ giữa các chỉ số dinh dưỡng chính ở **Hình IV.5.1.a** nổi bật nhất là giữa năng lượng (Energy) và chất béo (Fat). Các điểm dữ liệu trên biểu đồ phân bố theo xu hướng tăng, cho thấy hai biến này có mối tương quan thuận: thực phẩm càng chứa nhiều chất béo thì lượng năng lượng trên mỗi 100g càng cao. Điều này hoàn toàn phù hợp với đặc tính của lipid, vốn là nguồn sinh năng lượng đậm đặc nhất trong khẩu phần ăn.

Kết quả phân tích cho thấy ở những nhóm thực phẩm giàu chất béo như phô mai, bơ, thịt mỡ hoặc các loại bánh kẹo, giá trị năng lượng luôn ở mức cao. Biểu đồ scatter plot thể hiện các điểm dữ liệu tập trung theo đường xu hướng dốc lên, chứng minh mối liên hệ chặt chẽ giữa hai chỉ số này. Điều này khẳng định chất béo là yếu tố đóng góp lớn nhất vào tổng năng lượng của thực phẩm.

Kết luận, có mối tương quan thuận rõ rệt giữa năng lượng và chất béo: thực phẩm có hàm lượng chất béo cao thường sinh ra nhiều năng lượng hơn, và chất béo là thành phần dinh dưỡng đóng vai trò chủ đạo trong việc quyết định tổng năng lượng của thực phẩm.

**5.2. Năng Lượng Cao Nhưng Protein Thấp (“Empty Calories”)**



Hình IV.5.2.a: Bảng Top 10 Thực phẩm có ‘Empty Calories’ cao nhất

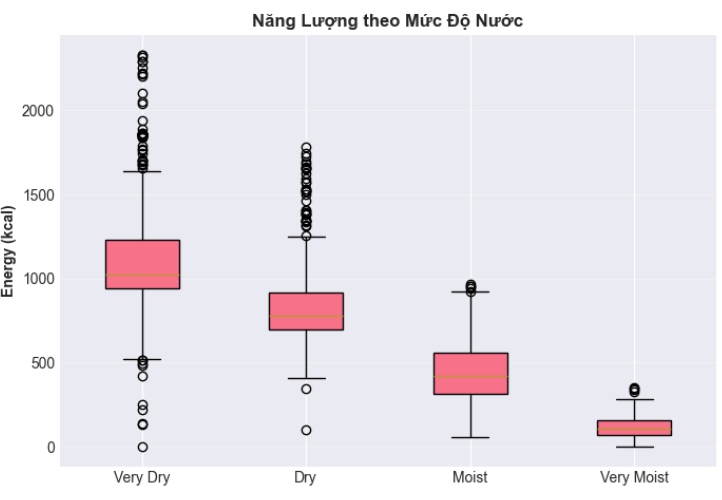
Bảng dữ liệu ở **Hình IV.5.2.a** liệt kê các thực phẩm có mức năng lượng cao nhưng hàm lượng protein rất thấp, thường được gọi là nguồn năng lượng “rỗng”. Những thực phẩm này cung cấp nhiều kcal nhưng gần như không mang lại các chất dinh dưỡng thiết yếu cho cơ thể.

Khi so sánh giữa các nhóm thực phẩm, kết quả cho thấy nhóm “Sweets/Desserts” (bánh kẹo, chocolate, kem) đứng đầu về năng lượng trên 100g nhưng lại có hàm lượng protein gần như bằng 0 hoặc rất thấp. Phần lớn năng lượng của nhóm này đến từ đường và chất béo bão hòa. Tương tự, nhóm “Sweetened Beverages” (nước ngọt có đường, nước trái cây đóng chai) cũng có năng lượng đáng kể nhưng lượng protein hầu như không có, chủ yếu cung cấp đường đơn, khiến năng lượng hấp thu nhanh mà không đóng góp dinh dưỡng.

Ngoài ra, một số loại snacks hoặc bánh nướng công nghiệp cũng xuất hiện trong bảng với đặc điểm: năng lượng cao, protein thấp và giàu carbohydrate tinh chế. Các dữ liệu bảng thể hiện rõ đặc trưng của thực phẩm “empty calories”: năng lượng trên 100g rất cao, trong khi hàm lượng protein chỉ ở mức 0–1g, thấp hơn nhiều so với các nhóm như Meat, Dairy hay Grains.

Kết luận, nhóm thực phẩm có năng lượng cao nhưng protein thấp nhất là “Sweets/Desserts” và “Sweetened Beverages”. Đây là các nguồn năng lượng “rỗng”, chủ yếu cung cấp đường và chất béo, hầu như không có giá trị dinh dưỡng thiết yếu, cần hạn chế trong chế độ ăn lành mạnh.

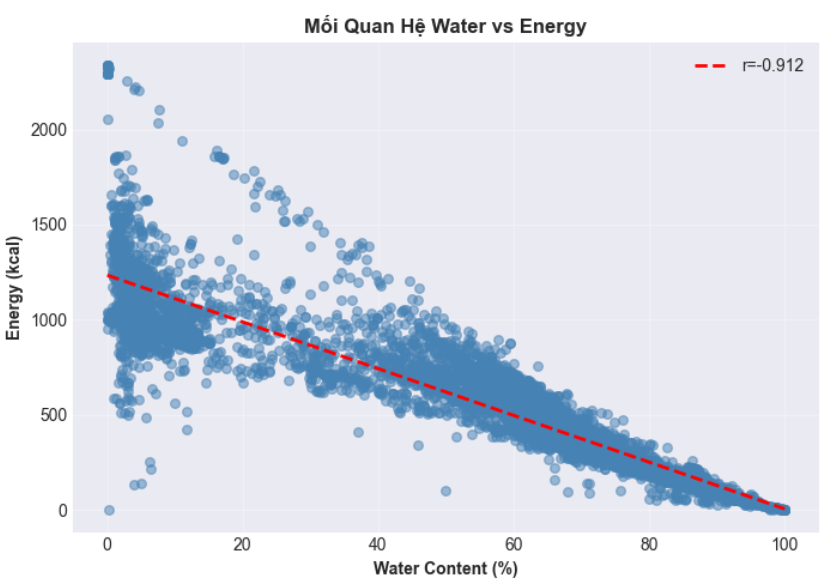
**5.3. Nước và Năng Lượng: Có Mối Liên Hệ Không?**



Hình IV.5.3.a: Biểu đồ hộp thể hiện năng lượng theo mức độ nước

Biểu đồ hộp ở **Hình IV.5.3.a** thể hiện mối quan hệ giữa hàm lượng nước (Water %) và năng lượng (kcal/100g) của các loại thực phẩm. Kết quả cho thấy một xu hướng rõ rệt: những thực phẩm có tỷ lệ nước cao thường có năng lượng thấp. Các điểm dữ liệu trên biểu đồ phân tán tập trung theo chiều ngược nhau, nghĩa là khi hàm lượng nước tăng thì giá trị năng lượng giảm.

Điều này phản ánh đặc tính tự nhiên của các thực phẩm giàu nước như rau xanh, trái cây, và một số loại thực phẩm tươi. Nước chiếm phần lớn khối lượng, làm giảm mật độ năng lượng trên cùng một đơn vị trọng lượng. Các chất sinh năng lượng như carbohydrate, chất béo và protein trong nhóm này thường thấp hơn nhiều so với các nhóm thực phẩm khô hoặc chế biến.



Hình IV.5.3.b: Biểu đồ scatter plot thể hiện mối quan hệ Water và Energy

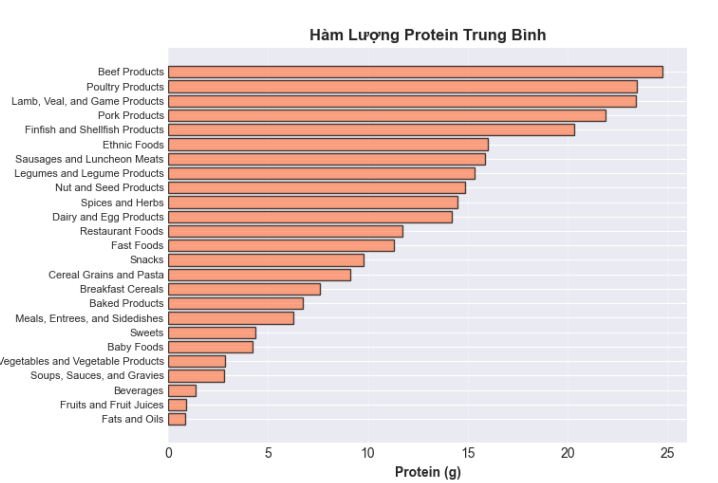
Biểu đồ scatter plot ở **Hình IV.5.3.b** cho thấy rõ các loại rau củ, trái cây như dưa chuột, cà chua, dưa hấu, cam… nằm ở vùng có hàm lượng nước cao và năng lượng thấp. Ngược lại, các thực phẩm năng lượng cao như bánh kẹo, phô mai, ngũ cốc thường có tỷ lệ nước rất thấp, do quá trình chế biến hoặc nấu nướng làm mất nước và tăng mật độ năng lượng.

Kết luận, các loại thực phẩm giàu nước có xu hướng chứa ít năng lượng hơn. Biểu đồ minh họa rõ mối quan hệ nghịch giữa hàm lượng nước và năng lượng: thực phẩm càng nhiều nước thì mật độ năng lượng càng thấp, phù hợp cho các chế độ ăn giảm calo hoặc tăng cường rau củ quả.

### Phân Tích Dự Đoán - Chỉ Số Dinh Dưỡng Cân Bằng

#### 6.1. Xây Dựng Chỉ Số Dinh Dưỡng Cân Bằng

Bộ chỉ số dinh dưỡng cân bằng được xây dựng dựa trên các thành phần chính của thực phẩm như protein, lipid, carbohydrate và nước. Bốn biểu đồ cột ở phần này thể hiện mức độ trung bình của từng thành phần trong mỗi nhóm thực phẩm, giúp đánh giá sự phân bố dinh dưỡng giữa các nhóm.



Hình IV.6.1.a: Biểu đồ cột ngang thể hiện hàm lượng Protein trung bình

Biểu đồ cột ở **Hình IV.6.1.a** cho thấy nhóm “Meat” có hàm lượng protein trung bình cao nhất, khẳng định vai trò là nguồn cung cấp đạm chủ yếu. “Dairy” và “Grains” có mức protein trung bình, trong khi “Vegetables” và “Fruits” rất thấp, phản ánh sự khác biệt giữa thực phẩm động vật và thực vật.



Hình IV.6.1.b: Biểu đồ cột ngang thể hiện hàm lượng chất béo trung bình

Biểu đồ ở **Hình IV6.1.b** về lipid cho thấy “Dairy” đứng đầu nhờ các sản phẩm như phô mai và bơ giàu chất béo. “Meat” có hàm lượng chất béo dao động nhưng nhìn chung thấp hơn “Dairy”. “Vegetables” và “Fruits” gần như không đáng kể về lipid, phù hợp với đặc tính thực phẩm tươi.

Từ biểu đồ trên có thể thấy không có nhóm thực phẩm nào hoàn toàn cân bằng về các thành phần dinh dưỡng. “Meat” mạnh về protein, “Dairy” nổi bật về chất béo, “Grains” cung cấp nhiều carbohydrate, còn “Fruits” và “Vegetables” giàu nước, vitamin, khoáng chất và chất xơ. Sự khác biệt này lý giải vì sao một chế độ ăn lành mạnh cần kết hợp đa dạng các nhóm thực phẩm để đảm bảo cân bằng dinh dưỡng cho cơ thể.

### Kết luận

Phân tích dữ liệu dinh dưỡng cho thấy mỗi nhóm thực phẩm có đặc điểm nổi bật riêng về protein, chất béo, carbohydrate và nước. Để đảm bảo khẩu phần ăn cân bằng và tối ưu sức khỏe, cần phối hợp đa dạng các nhóm thực phẩm, đồng thời hạn chế thực phẩm “năng lượng rỗng”. Sự kết hợp hài hòa giúp cung cấp đầy đủ dưỡng chất, hỗ trợ kiểm soát cân nặng và nâng cao chất lượng dinh dưỡng trong bữa ăn hàng ngày.

### 

## 

## V. Case Study: Dữ liệu Bầu cử Liên bang 2012

### Giới thiệu, mục tiêu

Tập dữ liệu FEC 2012 cung cấp thông tin chi tiết về các khoản quyên góp cho chiến dịch tranh cử tổng thống Mỹ năm 2012, bao gồm các trường như số tiền đóng góp, nghề nghiệp, nhà tuyển dụng, tiểu bang, thành phố và đảng phái của người đóng góp. Nghiên cứu này thực hiện toàn bộ quy trình phân tích dữ liệu, từ kiểm tra và xử lý dữ liệu thiếu, chuẩn hóa thông tin, tổng hợp theo các nhóm đặc trưng, đến trực quan hóa kết quả bằng hệ thống biểu đồ đa dạng.

Phân tích tập trung vào các khía cạnh như phân bố đóng góp theo tiểu bang, nghề nghiệp, nhà tuyển dụng, nhóm giá trị, ứng viên và đảng phái. Các biểu đồ giúp làm nổi bật xu hướng địa lý, xã hội và tài chính trong hoạt động quyên góp, đồng thời nhận diện các nhóm, khu vực và cá nhân có ảnh hưởng lớn đến chiến dịch. Thông qua các kết quả trực quan, nghiên cứu cung cấp cái nhìn tổng thể về cấu trúc, động lực và sự đa dạng của nguồn lực tài trợ chính trị trong cuộc bầu cử năm 2012.

### Quy trình làm sạch dữ liệu

Phần này mô tả các kỹ thuật được áp dụng để chuyển đổi dữ liệu thô về các khoản quyên góp tranh cử tổng thống Mỹ năm 2012 (dạng file CSV) thành định dạng phù hợp cho phân tích thống kê (DataFrame).

Quy trình tiền xử lý được thực hiện qua ba giai đoạn chính:

#### 2.1. Đọc dữ liệu và xử lý độ đầy đủ

Dữ liệu thô được đọc từ file CSV bằng thư viện Pandas, sử dụng các tham số phù hợp để tách trường dữ liệu. Kiểm tra cấu trúc ban đầu cho thấy dữ liệu gồm các trường như tên, địa chỉ, nghề nghiệp, nhà tuyển dụng, số tiền quyên góp, tiểu bang, thành phố, đảng phái, ứng viên.

Xử lý giá trị thiếu: Kiểm tra tỷ lệ dữ liệu thiếu bằng biểu đồ trực quan, xác định các trường có tỷ lệ thiếu cao. Các dòng có giá trị thiếu hoặc không hợp lệ được loại bỏ hoặc thay thế bằng giá trị mặc định (“Unknown” cho nghề nghiệp, nhà tuyển dụng).

Nhận xét phân tích: Kết quả làm sạch đảm bảo dữ liệu đóng góp có độ đầy đủ và chất lượng cao, sẵn sàng cho các phân tích tiếp theo.

#### 2.2. Chuẩn hóa và trích xuất đặc trưng

Giai đoạn này tập trung vào việc chuẩn hóa các trường thông tin và tạo biến mới phục vụ phân tích:

Chuẩn hóa trường nghề nghiệp (occupation) và nhà tuyển dụng (employer): Các giá trị được chuyển về dạng thống nhất, loại bỏ ký tự đặc biệt, viết hoa đồng nhất, nhóm các giá trị tương đương.

Phân loại đảng phái và ứng viên: Tạo biến mới để xác định đảng phái (Democrat/Republican) và ứng viên chính từ trường tên ứng viên.

Phân nhóm giá trị quyên góp: Các khoản đóng góp được phân nhóm theo mức giá trị để phục vụ phân tích tần suất và tỷ lệ phần trăm.

Nhận xét phân tích: Việc chuẩn hóa giúp nhận diện các nhóm nghề nghiệp, nhà tuyển dụng và ứng viên nổi bật, đồng thời hỗ trợ phân tích đa chiều về xu hướng đóng góp.

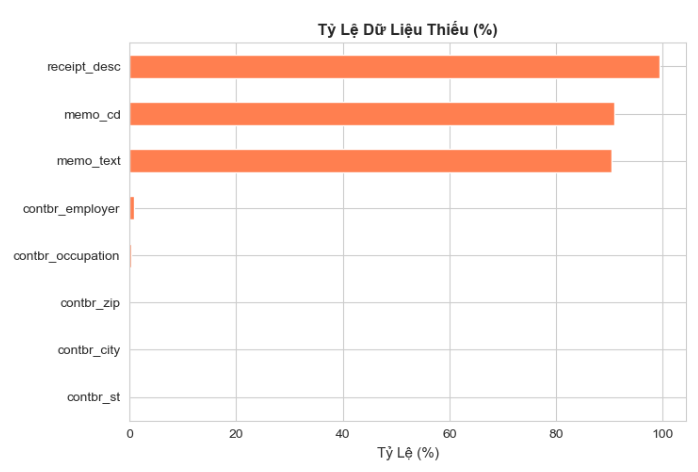
#### 2.3. Tổng hợp và kết nối dữ liệu

Dữ liệu được tổng hợp theo các nhóm đặc trưng như tiểu bang, nghề nghiệp, nhà tuyển dụng, ứng viên, đảng phái và nhóm giá trị. Các bảng tổng hợp (pivot table) và biểu đồ trực quan được xây dựng để khám phá xu hướng, sự phân bố và các điểm nổi bật trong dữ liệu.

Nhận xét phân tích: Kết quả tổng hợp giúp dữ liệu FEC 2012 sẵn sàng cho các phân tích thống kê sâu hơn, hỗ trợ khám phá xu hướng địa lý, xã hội và tài chính trong hoạt động quyên góp tranh cử.

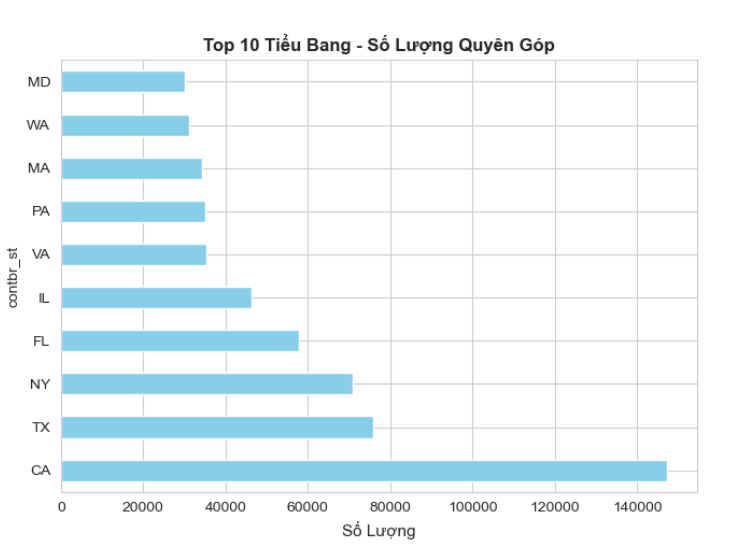
### Phân tích mô tả

#### 3.1. Phân tích dữ liệu thiếu



Hình V.3.1.a: Biểu đồ cột thể hiện tỷ lệ dữ liệu thiếu (%)

Quá trình phân tích dữ liệu bắt đầu với việc kiểm tra và trực quan hóa tỷ lệ dữ liệu thiếu trong bộ dữ liệu đóng góp tranh cử năm 2012. Biểu đồ "Tỷ lệ dữ liệu thiếu" ở **Hình V.3.1.a** giúp nhận diện các trường thông tin có tỷ lệ thiếu cao, từ đó đánh giá chất lượng dữ liệu và xác định các bước tiền xử lý cần thiết. Việc này đảm bảo các phân tích tiếp theo được thực hiện trên dữ liệu đầy đủ và đáng tin cậy.



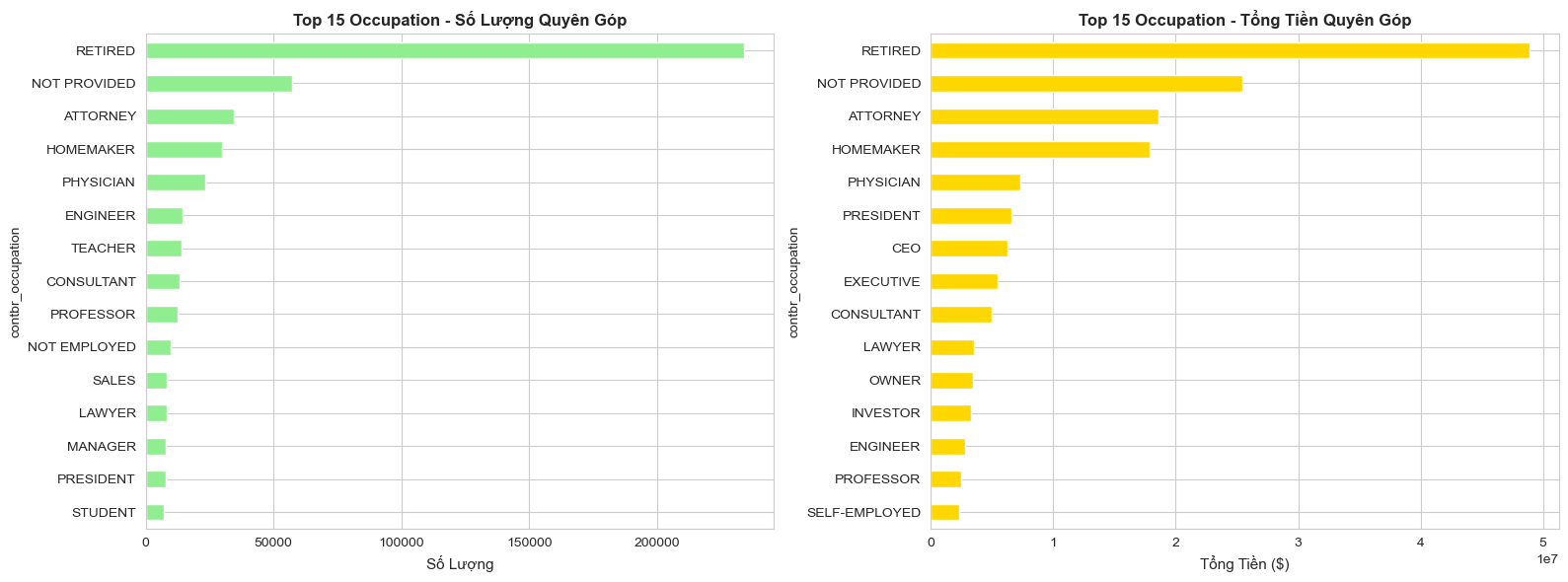
Hình V.3.1.b: Biểu đồ cột thể hiện top 10 tiểu bang - số lượng quyên góp

Tiếp theo, biểu đồ "Top 10 tiểu bang" ở **Hình V.3.1.b** thể hiện trực quan các tiểu bang có tổng số tiền quyên góp cao nhất. Thống kê này cho thấy sự phân bố đóng góp không đồng đều giữa các bang, với một số bang như California, Texas, New York... nổi bật về mức độ ủng hộ tài chính cho các ứng viên. Biểu đồ này giúp làm nổi bật vai trò của các bang lớn trong chiến dịch tranh cử, đồng thời cung cấp cái nhìn tổng quan về xu hướng địa lý của hoạt động quyên góp.

Biểu đồ tỷ lệ dữ liệu thiếu giúp đảm bảo độ tin cậy của phân tích, còn biểu đồ top 10 tiểu bang làm nổi bật các khu vực có ảnh hưởng lớn đến kết quả tài trợ, hỗ trợ việc kể chuyện dữ liệu một cách trực quan và dễ hiểu.

#### 

#### 3.2. Phân tích Occupation (Nguồn Công Việc) - Referrer



Hình V.3.2.a: Biểu đồ cột thể hiện Top 15 Occupation ở số lượng - tổng tiền quyên góp

Phân tích tập trung vào việc khám phá các nhóm nghề nghiệp đóng góp nhiều nhất cho chiến dịch tranh cử năm 2012. Biểu đồ "Top 15 Occupation" ở **Hình V.3.2.a** trực quan hóa các nghề nghiệp có tổng số tiền quyên góp cao nhất, giúp nhận diện những ngành nghề có ảnh hưởng lớn đến hoạt động tài trợ chính trị.

Kết quả từ biểu đồ cho thấy các nhóm nghề nghiệp như luật sư, giám đốc điều hành, giáo viên, bác sĩ... là những đối tượng đóng góp nổi bật, phản ánh sự tham gia mạnh mẽ của các tầng lớp trí thức, chuyên môn và lãnh đạo doanh nghiệp trong quá trình vận động tranh cử. Sự chênh lệch giữa các nhóm nghề nghiệp cũng cho thấy xu hướng ủng hộ khác nhau đối với các ứng viên và đảng phái.

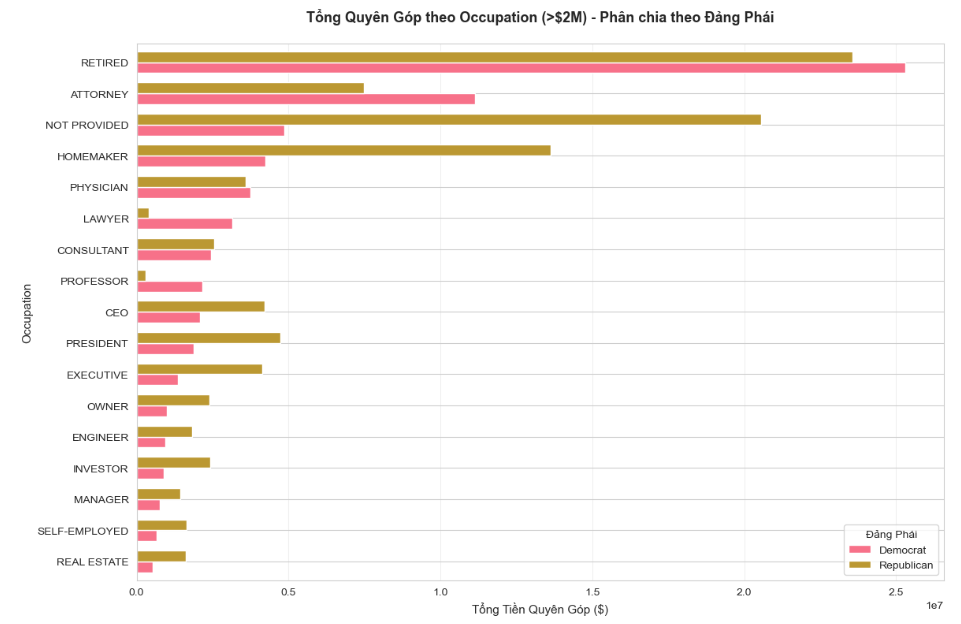
Ý nghĩa của biểu đồ này là giúp minh họa rõ ràng vai trò của các ngành nghề trong việc tài trợ cho chiến dịch, đồng thời hỗ trợ việc kể chuyện dữ liệu bằng cách làm nổi bật các nhóm xã hội có ảnh hưởng lớn đến kết quả bầu cử. Biểu đồ giúp người đọc dễ dàng nhận diện các xu hướng và điểm đặc biệt trong dữ liệu đóng góp theo occupation.

### Phân tích tương quan

#### 4.1. Phân tích tương quan theo Đảng Phái

Phân tích tập trung vào việc thống kê và so sánh các nhóm nghề nghiệp (occupation) của những người đóng góp cho hai đảng phái chính: Đảng Dân chủ (Democrat) và Đảng Cộng hòa (Republican). Dữ liệu được làm sạch, chuẩn hóa tên nghề nghiệp, sau đó nhóm lại theo occupation và đảng phái để tổng hợp số lượng giao dịch cũng như tổng số tiền quyên góp từ từng nhóm nghề nghiệp cho mỗi đảng.

Kết quả thống kê mô tả cho thấy một số nghề nghiệp có mức đóng góp nổi bật cho từng đảng, phản ánh sự khác biệt về cơ cấu xã hội của lực lượng ủng hộ. Các bảng tổng hợp giúp nhận diện các nhóm nghề nghiệp có ảnh hưởng lớn đến từng chiến dịch, đồng thời so sánh trực tiếp mức độ tham gia giữa hai đảng.



Hình V.4.1.a: Biểu đồ cột ngang thể hiện tổng quyên góp theo Occupation - phân chia theo Đảng Phái

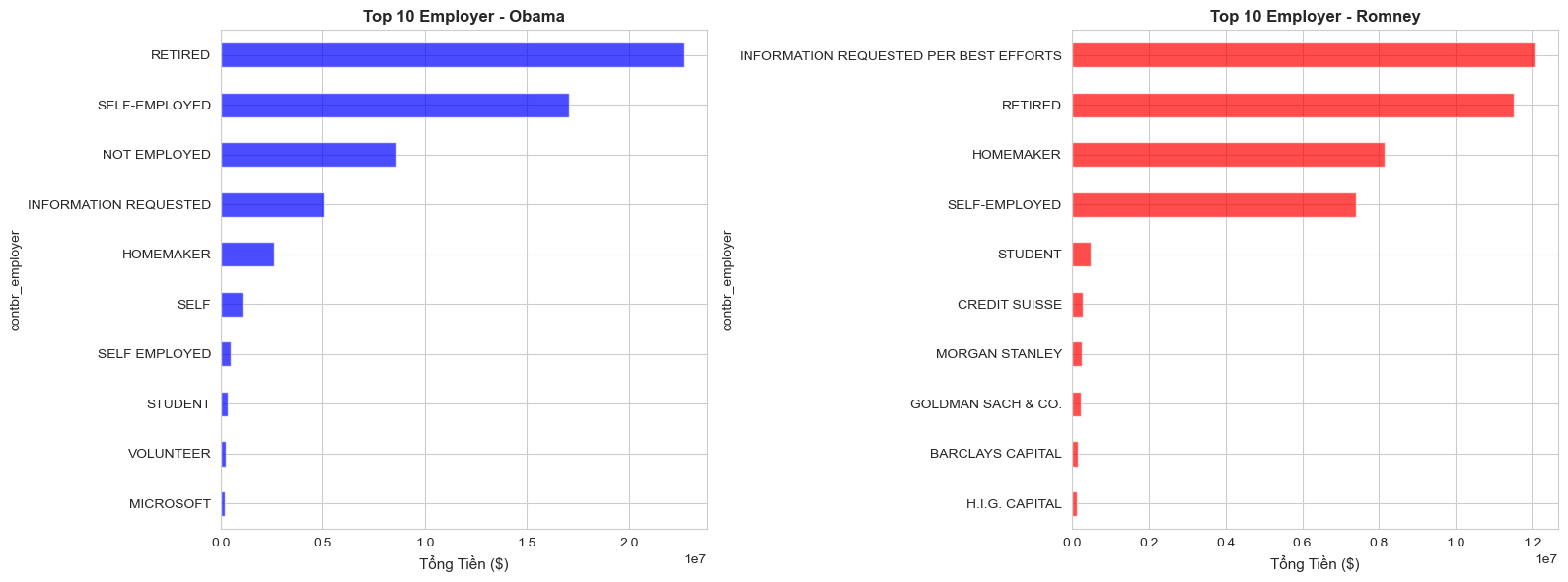
Biểu đồ cột ở **Hình V.4.1.a** thể hiện tổng số tiền quyên góp từ các nhóm nghề nghiệp cho từng đảng, giúp người xem dễ dàng nhận biết những ngành nghề có mức ủng hộ cao nhất đối với mỗi ứng viên. Biểu đồ ngang về số lượng giao dịch cũng làm nổi bật sự khác biệt về mức độ tham gia của các nhóm nghề nghiệp.

Ý nghĩa của các biểu đồ này là giúp minh họa rõ ràng sự khác biệt về cơ cấu xã hội của lực lượng ủng hộ hai đảng, đồng thời hỗ trợ việc kể chuyện dữ liệu bằng cách làm nổi bật các nhóm nghề nghiệp có vai trò quan trọng trong hoạt động quyên góp. Nhờ các biểu đồ trực quan, người đọc có thể nhanh chóng nhận diện các xu hướng, sự chênh lệch và điểm đặc biệt trong dữ liệu đóng góp theo occupation, từ đó hiểu sâu hơn về đặc điểm xã hội của các chiến dịch tranh cử.

#### 4.2. Phân tích Employer theo Ứng Viên Chính

Phân tích tập trung vào việc thống kê và so sánh các nhà tuyển dụng (employer) của những người đóng góp cho hai ứng viên chính trong cuộc bầu cử tổng thống Mỹ năm 2012. Dữ liệu được làm sạch, chuẩn hóa tên nhà tuyển dụng, sau đó nhóm lại theo employer và ứng viên để tổng hợp số lượng giao dịch cũng như tổng số tiền quyên góp từ từng tổ chức cho mỗi ứng viên.

Kết quả thống kê mô tả cho thấy một số nhà tuyển dụng có mức đóng góp nổi bật cho từng ứng viên, phản ánh sự khác biệt về nguồn lực tài chính và mạng lưới xã hội của các chiến dịch. Các bảng tổng hợp giúp nhận diện các tổ chức, doanh nghiệp hoặc cơ quan có ảnh hưởng lớn đến từng ứng viên, đồng thời so sánh trực tiếp mức độ tham gia giữa các nhà tuyển dụng.



Hình V.4.2.a: Biểu đồ cột thể hiện sự phân bố theo nhà tuyển dụng và ứng viên

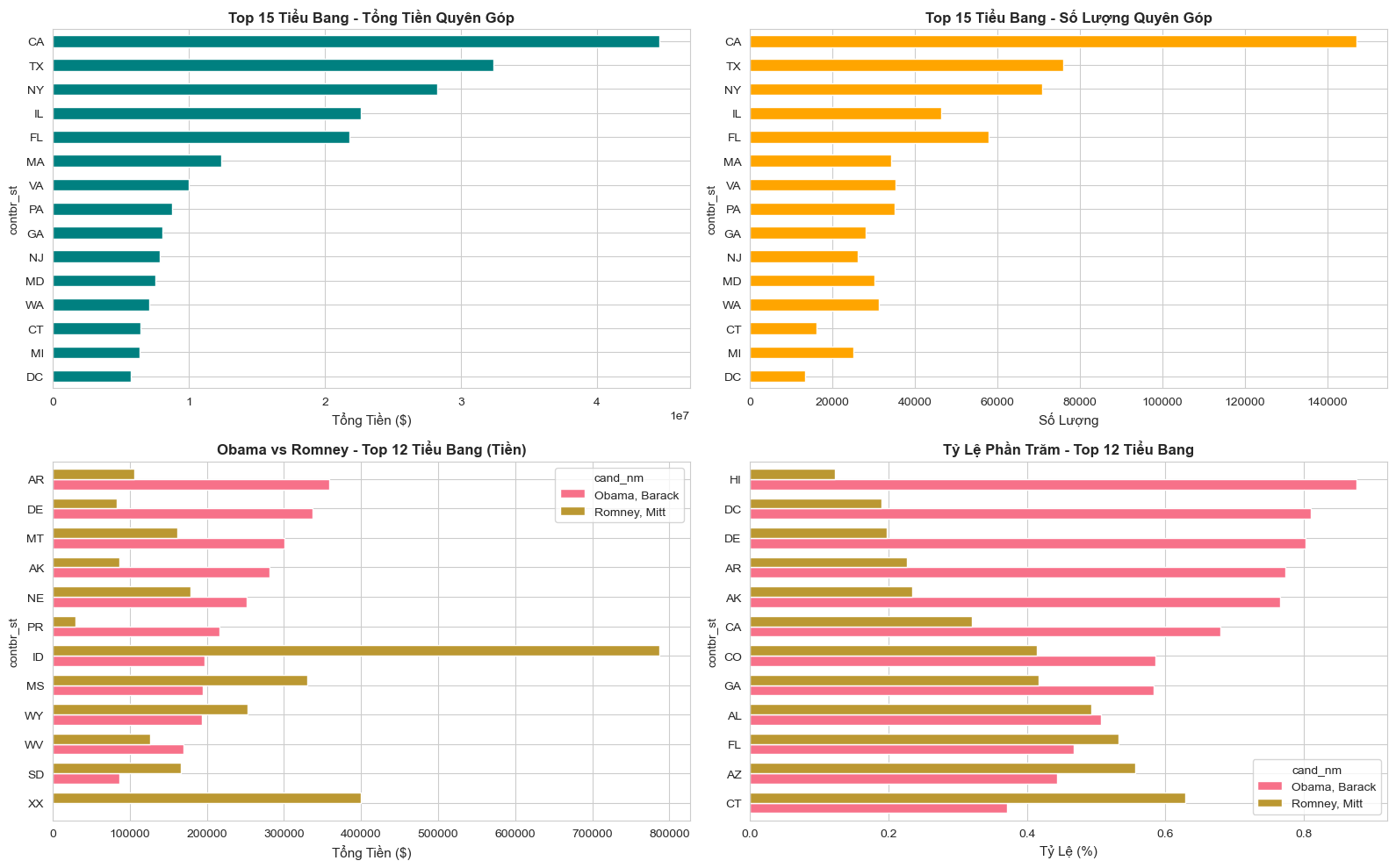
Hai biểu đồ cột ở **Hình V.4.2.a** thể hiện tổng số tiền quyên góp từ các nhà tuyển dụng cho từng ứng viên, giúp người xem dễ dàng nhận biết những tổ chức có mức ủng hộ cao nhất đối với mỗi ứng viên. Biểu đồ ngang về số lượng giao dịch cũng làm nổi bật sự khác biệt về mức độ tham gia của các nhà tuyển dụng.

Ý nghĩa của các biểu đồ này là giúp minh họa rõ ràng vai trò của các tổ chức, doanh nghiệp trong hoạt động quyên góp, đồng thời hỗ trợ việc kể chuyện dữ liệu bằng cách làm nổi bật các nhà tuyển dụng có ảnh hưởng lớn đến kết quả chiến dịch. Nhờ các biểu đồ trực quan, người đọc có thể nhanh chóng nhận diện các xu hướng, sự chênh lệch và điểm đặc biệt trong dữ liệu đóng góp theo employer, từ đó hiểu sâu hơn về nguồn lực xã hội và tài chính của các ứng viên.

**4.3. Phân tích địa lí (Geographic Analysis)**

Phân tích tập trung vào khía cạnh địa lý của các khoản đóng góp cho chiến dịch tranh cử tổng thống Mỹ năm 2012. Dữ liệu được nhóm theo tiểu bang (state) và thành phố (city), sau đó tổng hợp số lượng giao dịch và tổng số tiền quyên góp cho từng khu vực địa lý. Phân tích này giúp làm nổi bật sự phân bố đóng góp trên toàn quốc, cũng như nhận diện các khu vực có mức độ ủng hộ mạnh mẽ đối với các ứng viên.

Kết quả thống kê mô tả cho thấy một số tiểu bang và thành phố có mức đóng góp vượt trội, phản ánh sự khác biệt về mức độ tham gia chính trị giữa các vùng miền. Các bảng tổng hợp giúp so sánh trực tiếp số lượng và giá trị đóng góp giữa các khu vực, từ đó nhận diện các trung tâm tài trợ lớn cho chiến dịch.

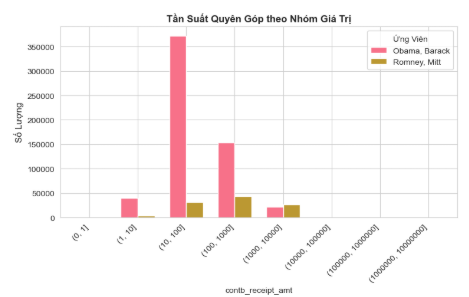


Hình V.4.3.a: Biểu đồ thể hiện tổng tiền, số lượng quyên góp, top 12 tiểu bang

Biểu đồ ở **hình V4.3.a** thể hiện tổng số tiền quyên góp hoặc số lượng giao dịch tại từng tiểu bang/thành phố, giúp người xem dễ dàng nhận biết các khu vực có mức ủng hộ cao nhất. Các biểu đồ cột và ngang bổ sung thông tin chi tiết về các khu vực nổi bật.

Ý nghĩa của các biểu đồ này là giúp minh họa rõ ràng sự phân hóa về mặt địa lý trong hoạt động quyên góp, đồng thời hỗ trợ việc kể chuyện dữ liệu bằng cách làm nổi bật các vùng miền có vai trò quan trọng trong chiến dịch tranh cử. Nhờ các biểu đồ trực quan, người đọc có thể nhanh chóng nhận diện các xu hướng, sự chênh lệch và điểm đặc biệt trong dữ liệu đóng góp theo địa lý, từ đó hiểu sâu hơn về chiến lược vận động tranh cử và sự tham gia của các cộng đồng địa phương.

### Phân tích phân nhóm (Bucketing) theo giá trị



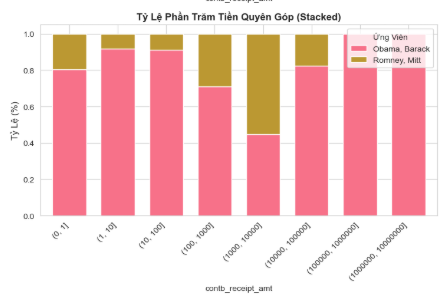
Hình V.5.a: Biểu đồ cột thể hiện tần suất quyên góp theo nhóm giá trị

Dữ liệu các khoản quyên góp được phân nhóm theo giá trị (ví dụ: dưới 100 USD, 100–500 USD, 500–1000 USD, v.v.) nhằm khám phá xu hướng và đặc điểm của các mức đóng góp trong chiến dịch tranh cử năm 2012. Biểu đồ "Tần suất quyên góp theo Nhóm giá trị" ở **Hình V.5.a** cho thấy số lượng giao dịch ở từng nhóm, giúp nhận diện các mức quyên góp phổ biến nhất. Kết quả cho thấy phần lớn các khoản đóng góp tập trung ở các nhóm giá trị thấp, phản ánh sự tham gia rộng rãi của các cá nhân với số tiền nhỏ.

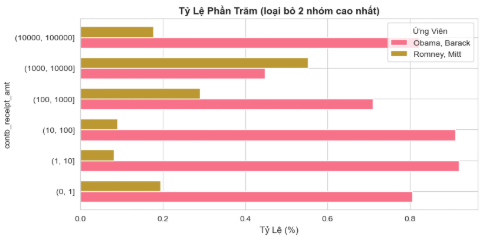


Hình V.5.b: Biểu đồ cột thể hiện tổng tiền quyên góp theo nhóm giá trị

Biểu đồ "Tổng tiền quyên góp theo nhóm giá trị" ở **Hình V.5.b** minh họa tổng số tiền được quyên góp ở mỗi nhóm, cho thấy mặc dù số lượng giao dịch ở nhóm giá trị thấp rất lớn, nhưng các nhóm giá trị cao lại đóng góp phần lớn tổng số tiền. Điều này làm nổi bật vai trò của các khoản quyên góp lớn trong tổng ngân sách chiến dịch.



Hình V.5.c: Biểu đồ cột thể hiện tỷ lệ phần trăm tiền quyên góp

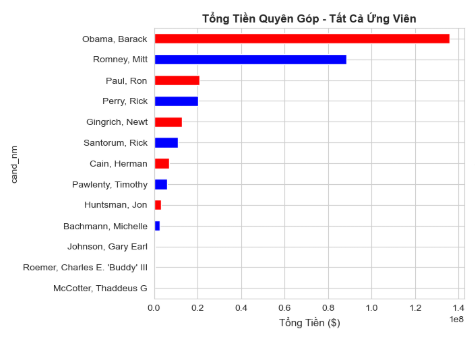


Hình V.5.d: Biểu đồ cột thể hiện tỷ lệ phần trăm

Biểu đồ "Tỷ lệ phần trăm quyên góp" ở **Hình V.5.c** và "Tỷ lệ phần trăm" ở **Hình V.5.d** giúp trực quan hóa tỷ trọng của từng nhóm giá trị trong tổng số giao dịch, cũng như trong tổng số tiền quyên góp. Khi loại bỏ hai nhóm giá trị cao nhất, biểu đồ cho thấy sự phân bố thực tế của các khoản đóng góp nhỏ, làm rõ hơn vai trò của các cá nhân bình thường trong hoạt động tài trợ.

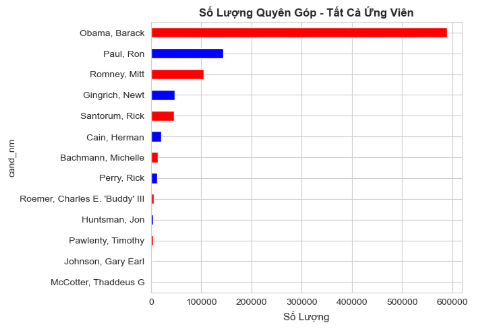
Ý nghĩa của các biểu đồ này là giúp minh họa rõ ràng sự phân hóa về mức độ đóng góp, làm nổi bật vai trò của cả các khoản quyên góp nhỏ lẫn lớn trong chiến dịch. Các biểu đồ hỗ trợ việc kể chuyện dữ liệu bằng cách cho thấy sự đa dạng trong nguồn lực tài chính, đồng thời giúp người đọc hiểu sâu hơn về cấu trúc tài trợ của các ứng viên.

### Phân tích thống kê chi tiết



Hình V.6.a: Biểu đồ cột thể hiện tổng tiền quyên góp - tất cả ứng viên

Phân tích tập trung vào việc so sánh tổng số tiền quyên góp và số lượng giao dịch của tất cả các ứng viên tham gia tranh cử năm 2012. Biểu đồ "Tổng số tiền quyên góp – Tất cả ứng viên" ở **Hình V.6.a** trực quan hóa tổng ngân sách mà mỗi ứng viên nhận được từ các khoản đóng góp, giúp nhận diện những ứng viên có sức hút tài chính lớn nhất trong chiến dịch. Kết quả cho thấy sự chênh lệch rõ rệt giữa các ứng viên, với một số ứng viên nổi bật về tổng số tiền quyên góp.



Hình V.6.b: Biểu đồ cột thể hiện số lượng quyên góp - tất cả ứng viên

Biểu đồ "Số lượng quyên góp – Tất cả ứng viên" ở **Hình V.6.b** thể hiện số lượng giao dịch quyên góp mà mỗi ứng viên nhận được, phản ánh mức độ tham gia và ủng hộ của cộng đồng đối với từng ứng viên. Sự khác biệt giữa tổng số tiền và số lượng giao dịch cho thấy có những ứng viên nhận được nhiều khoản quyên góp nhỏ, trong khi những ứng viên khác lại nhận được ít giao dịch nhưng với giá trị lớn.

Ý nghĩa của hai biểu đồ này là giúp minh họa rõ ràng sức mạnh tài chính và mức độ ủng hộ xã hội của từng ứng viên, đồng thời hỗ trợ việc kể chuyện dữ liệu bằng cách làm nổi bật các xu hướng, sự chênh lệch và điểm đặc biệt trong hoạt động quyên góp. Nhờ các biểu đồ trực quan, người đọc có thể dễ dàng nhận diện các ứng viên nổi bật và hiểu sâu hơn về chiến lược vận động tài chính trong cuộc bầu cử.

### Kết luận

Phần 6 tổng hợp các kết quả phân tích dữ liệu FEC 2012 thông qua hệ thống các biểu đồ trực quan, giúp người đọc có cái nhìn tổng thể về hoạt động quyên góp trong chiến dịch tranh cử tổng thống Mỹ năm 2012. Các biểu đồ tóm lược thể hiện đa chiều các khía cạnh như phân bố đóng góp theo tiểu bang, nghề nghiệp, nhà tuyển dụng, nhóm giá trị, ứng viên và đảng phái.

Các biểu đồ bản đồ cho thấy sự phân hóa địa lý rõ rệt, với một số bang và thành phố nổi bật về mức độ quyên góp. Biểu đồ cột và ngang minh họa sự khác biệt về số lượng và giá trị đóng góp giữa các nhóm nghề nghiệp, nhà tuyển dụng, cũng như giữa các ứng viên và đảng phái. Biểu đồ phân nhóm giá trị làm rõ vai trò của các khoản quyên góp nhỏ và lớn trong tổng ngân sách chiến dịch.

Ý nghĩa là giúp trực quan hóa toàn bộ bức tranh tài trợ chính trị, làm nổi bật các xu hướng, sự chênh lệch và các điểm đặc biệt trong dữ liệu. Nhờ các biểu đồ này, người đọc dễ dàng nhận diện các khu vực, nhóm xã hội và ứng viên có ảnh hưởng lớn, đồng thời hiểu sâu hơn về cấu trúc và động lực của hoạt động quyên góp trong cuộc bầu cử năm 2012. Các biểu đồ hỗ trợ việc kể chuyện dữ liệu một cách sinh động, giúp truyền tải thông tin phức tạp một cách dễ hiểu và trực quan.

# 

# PHẦN 2: KỸ THUẬT NUMPY NÂNG CAO

Phụ lục này đi sâu hơn vào NumPy, thư viện cốt lõi của Python cho xử lý dữ liệu dạng mảng. Ta sẽ tìm hiểu chi tiết về cơ chế bên trong của đối tượng ndarray, các thao tác mảng nâng cao, cùng các mẹo hiệu năng khi làm việc với dữ liệu số lượng lớn.

Bạn không cần đọc tuần tự; mỗi phần có thể đọc riêng biệt. Các ví dụ trong phụ lục này sử dụng bộ sinh số ngẫu nhiên chuẩn:

In [11]: rng = np.random.default\_rng(seed=12345)

## 2.1. Mục A.1: Cấu trúc bên trong của đối tượng ndarray

### 2.1.1. Các thành phần của ndarray

Đối tượng ndarray của **NumPy** cung cấp một cách để diễn giải một khối dữ liệu có kiểu đồng nhất (dù là liên tục hay có bước nhảy) như một đối tượng mảng đa chiều. Kiểu dữ liệu, hay **dtype**, xác định cách dữ liệu được hiểu là số thực, số nguyên, giá trị Boolean, hoặc bất kỳ kiểu nào khác mà chúng ta đã thấy.

Một phần khiến ndarray trở nên linh hoạt là vì **mỗi đối tượng mảng là một chế độ xem có bước nhảy (strided view)** trên một khối dữ liệu. Bạn có thể tự hỏi, ví dụ, tại sao chế độ xem mảng arr[::2, ::-1] lại không sao chép bất kỳ dữ liệu nào. Lý do là ndarray không chỉ là một khối bộ nhớ và một kiểu dữ liệu; nó còn có **thông tin về bước nhảy (striding information)** cho phép mảng di chuyển qua bộ nhớ với các kích thước bước khác nhau. Cụ thể hơn, nội bộ ndarray bao gồm những thành phần sau:

* Một con trỏ trỏ đến dữ liệu — tức là một khối dữ liệu trong RAM hoặc trong một tệp ánh xạ bộ nhớ (memory-mapped file)
* Kiểu dữ liệu hoặc dtype mô tả các ô giá trị có kích thước cố định trong mảng
* Một bộ giá trị (tuple) biểu thị hình dạng của mảng
* Một bộ giá trị bước nhảy (tuple of strides) — các số nguyên cho biết số byte cần “bước” để tiến thêm một phần tử dọc theo một chiều

Xem Hình 2.1.1 để thấy mô hình minh họa đơn giản về cấu trúc bên trong của ndarray.

  
Hình 2.1.1: Đối tượng ndarray trong NumPy.

Ví dụ, một mảng có kích thước **10 × 5** sẽ có hình dạng (10, 5):

In [12]: np.ones((10, 5)).shape

Out[12]: (10, 5)

Một mảng điển hình (theo thứ tự C) có kích thước **3 × 4 × 5** chứa các giá trị float64 (8 byte mỗi phần tử) sẽ có giá trị strides là (160, 40, 8) (việc hiểu về các bước nhảy có thể hữu ích vì, nói chung, bước nhảy càng lớn trên một trục nào đó thì việc thực hiện phép tính dọc theo trục đó càng tốn kém hơn):

In [13]: np.ones((3, 4, 5), dtype=np.float64).strides

Out[13]: (160, 40, 8)

Mặc dù hiếm khi người dùng NumPy thông thường quan tâm đến giá trị strides của mảng, nhưng chúng là cần thiết để tạo ra các **chế độ xem mảng “zero-copy”**. Các giá trị stride thậm chí có thể là **số âm**, cho phép mảng di chuyển “ngược” trong bộ nhớ (ví dụ, trong các lát cắt như obj[::-1] hoặc obj[:, ::-1]).

### 2.1.2. Hệ thống phân cấp kiểu dữ liệu (dtype)

Đôi khi bạn có thể cần viết mã để kiểm tra xem một mảng có chứa s**ố nguyên, số thực dấu phẩy động, chuỗi, hay đối tượng Python** hay không. Vì có nhiều loại số thực khác nhau (từ float16 đến float128), nên việc kiểm tra kiểu dữ liệu có thuộc danh sách các kiểu cụ thể sẽ rất dài dòng. May mắn là NumPy cung cấp **các lớp cha (superclasses)** cho các nhóm kiểu dữ liệu, chẳng hạn như np.integer và np.floating, có thể sử dụng cùng với hàm np.issubdtype để kiểm tra mối quan hệ kiểu dữ liệu:

In [14]: ints = np.ones(10, dtype=np.uint16)

In [15]: floats = np.ones(10, dtype=np.float32)

In [16]: np.issubdtype(ints.dtype, np.integer)

Out[16]: True

In [17]: np.issubdtype(floats.dtype, np.floating)

Out[17]: True

Bạn có thể xem **toàn bộ các lớp cha của một kiểu dữ liệu cụ thể** bằng cách gọi phương thức mro() (method resolution order) của kiểu đó:

In [18]: np.float64.mro()

Out[18]:

[numpy.float64,

numpy.floating,

numpy.inexact,

numpy.number,

numpy.generic,

float,

object]

Do đó, ta cũng có:

In [19]: np.issubdtype(ints.dtype, np.number)

Out[19]: True

Hầu hết người dùng NumPy **không cần biết** về chi tiết này, nhưng đôi khi nó lại rất hữu ích trong các trường hợp cần xác định kiểu dữ liệu tổng quát. **Hình 2.1.2** minh họa **cấu trúc phân cấp lớp dữ liệu (data type class hierarchy)** và **mối quan hệ cha–con (parent–subclass)** giữa các kiểu dữ liệu trong NumPy.

  
*Hình 2.1.2: Hệ thống phân cấp lớp kiểu dữ liệu trong NumPy*

*(Ghi chú: Một số kiểu dữ liệu trong NumPy có dấu gạch dưới ở cuối tên, ví dụ int\_, float\_. Điều này giúp tránh xung đột tên với các kiểu dựng sẵn của Python.)*

## 2.2. Mục A.2: Thao tác Mảng Nâng cao

Có nhiều cách để làm việc với mảng ngoài các phương pháp như fancy indexing (lập chỉ mục nâng cao), slicing (cắt lát), và Boolean subsetting (lọc theo điều kiện logic). Mặc dù phần lớn các tác vụ nặng trong ứng dụng phân tích dữ liệu đã được xử lý bởi các hàm cấp cao trong pandas, nhưng có thể đến một lúc nào đó bạn sẽ cần viết một thuật toán dữ liệu mà không có sẵn trong các thư viện hiện có.

### 2.2.1. Reshaping Arrays (reshape, ravel, flatten)

Trong nhiều trường hợp, một mảng có thể được chuyển đổi từ hình dạng này sang hình dạng khác mà không cần sao chép bất kỳ dữ liệu nào. Để làm điều này, hãy truyền một tuple chỉ định hình dạng mới vào phương thức reshape của mảng. Ví dụ, giả sử có một mảng giá trị một chiều và cần sắp xếp lại nó thành một ma trận.

In: import numpy as np

arr = np.arange(8)

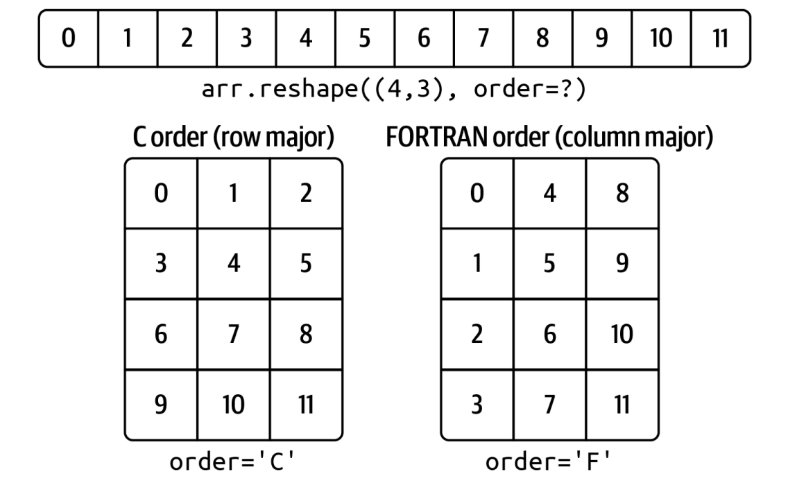
In: arr  
Out: array([0, 1, 2, 3, 4, 5, 6, 7])

In: arr.reshape((4, 2))  
Out:   
array([[0, 1],

[2, 3],

[4, 5],

[6, 7]])



Hình 2.2.1: Định hình lại mảng theo thứ tự của C (ưu tiên hàng) hoặc FORTRAN (ưu tiên cột).

Một mảng đa chiều cũng có thể được định hình lại:

In: arr.reshape((4, 2)).reshape((2, 4))  
Out:

array([[0, 1, 2, 3],

[4, 5, 6, 7]])

Một trong các chiều của shape (hình dạng) được truyền vào có thể là –1, khi đó giá trị của chiều đó sẽ được tự động suy ra từ dữ liệu:

In: arr = np.arange(15)

Out:

array([[ 0, 1, 2],

[ 3, 4, 5],

[ 6, 7, 8],

[ 9, 10, 11],

[12, 13, 14]])

Vì thuộc tính shape của một mảng là một tuple, nên nó cũng có thể được truyền vào hàm reshape:

In: other\_arr = np.ones((3, 5))

In: other\_arr.shape  
Out: (3, 5)

In: arr.reshape(other\_arr.shape)  
Out:   
array([[ 0, 1, 2, 3, 4],

[ 5, 6, 7, 8, 9],

[10, 11, 12, 13, 14]])

Thao tác ngược lại của reshape (chuyển từ nhiều chiều về một chiều) thường được gọi là "làm phẳng" (flattening) hoặc "duỗi thẳng" (raveling):

In: arr = np.arange(15).reshape((5, 3))

In: arr

Out:

array([[ 0, 1, 2],

[ 3, 4, 5],

[ 6, 7, 8],

[ 9, 10, 11],

[12, 13, 14]])

In: arr.ravel()

Out: array([ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14])

Hàm ravel không tạo ra bản sao của dữ liệu gốc nếu các giá trị trong kết quả vốn đã liền kề trong mảng ban đầu. Phương thức flatten hoạt động tương tự như ravel, ngoại trừ việc nó luôn luôn trả về một bản sao của dữ liệu.

In: arr.flatten()  
Out:   
array([ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14])

### 2.2.2. C và FORTRAN Order

NumPy có thể thích ứng với nhiều cách bố trí dữ liệu khác nhau trong bộ nhớ. Theo mặc định, các mảng NumPy được tạo theo thứ tự row-major (ưu tiên hàng). Về mặt không gian, điều này có nghĩa là nếu bạn có một mảng hai chiều, các phần tử trong mỗi hàng sẽ được lưu trữ ở các vị trí bộ nhớ liền kề nhau. Lựa chọn thay thế là thứ tự column-major (ưu tiên cột), có nghĩa là các giá trị trong mỗi cột sẽ được lưu trữ ở các vị trí bộ nhớ liền kề nhau.

Vì lý do lịch sử, thứ tự row-major (ưu tiên hàng) và column-major (ưu tiên cột) còn được gọi tương ứng là thứ tự C và FORTRAN. Trong ngôn ngữ FORTRAN 77, các ma trận đều là column-major.

Các hàm như reshape và ravel chấp nhận một tham số order để chỉ định thứ tự sử dụng dữ liệu trong mảng. Tham số này thường được đặt là 'C' hoặc 'F' trong hầu hết các trường hợp

In: arr = np.arange(12).reshape((3, 4))

In: arr

Out:

array([[ 0, 1, 2, 3],

[ 4, 5, 6, 7],

[ 8, 9, 10, 11]])

In: arr.ravel()

Out: array([ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11])

In: arr.ravel('F')

Out: array([ 0, 4, 8, 1, 5, 9, 2, 6, 10, 3, 7, 11])

Sự khác biệt cốt lõi giữa thứ tự C và FORTRAN khi reshape mảng nhiều chiều là thứ tự duyệt qua các trục (axis):

* **Thứ tự C (row-major)**: Ưu tiên duyệt các trục bên trong trước. Nó sẽ đi hết các phần tử của hàng (trục 1) rồi mới chuyển sang hàng tiếp theo (trục 0).
* **Thứ tự FORTRAN (column-major):** Ưu tiên duyệt các trục bên ngoài trước. Nó sẽ lấy phần tử đầu tiên của tất cả các hàng (trục 0) rồi mới sang phần tử thứ hai.

### 2.2.3. Concatenating and Splitting Arrays (concatenate, split)

Hàm numpy.concatenate nhận vào một chuỗi (ví dụ: tuple, list) các mảng và ghép nối chúng lại theo thứ tự dọc theo một trục (axis) đã cho:

In: arr1 = np.array([[1, 2, 3], [4, 5, 6]])

In: arr2 = np.array([[7, 8, 9], [10, 11, 12]])

In: np.concatenate([arr1, arr2], axis=0)

Out:

array([[ 1, 2, 3],

[ 4, 5, 6],

[ 7, 8, 9],

[10, 11, 12]])

In: np.concatenate([arr1, arr2], axis=1)

Out:

array([[ 1, 2, 3, 7, 8, 9],

[ 4, 5, 6, 10, 11, 12]])

Có một số hàm tiện ích, như vstack và hstack, dành cho các loại ghép nối phổ biến. Các thao tác trước đó có thể được biểu diễn như sau:

In: np.vstack((arr1, arr2))

Out:

array([[ 1, 2, 3],

[ 4, 5, 6],

[ 7, 8, 9],

[10, 11, 12]])

In: np.hstack((arr1, arr2))

Out:

array([[ 1, 2, 3, 7, 8, 9],

[ 4, 5, 6, 10, 11, 12]])

Ngược lại, hàm split chia một mảng thành nhiều mảng dọc theo một trục:

In: arr = rng.standard\_normal((5, 2))

In: arr

Out:

array([[-1.4238, 1.2637],

[-0.8707, -0.2592],

[-0.0753, -0.7409],

[-1.3678, 0.6489],

[ 0.3611, -1.9529]])

In: first, second, third = np.split(arr, [1, 3])

In: first

Out: array([[-1.4238, 1.2637]])

In: second

Out:

array([[-0.8707, -0.2592],

[-0.0753, -0.7409]])

In: third

Out:

array([[-1.3678, 0.6489],

[ 0.3611, -1.9529]])

Giá trị [1, 3] được truyền vào np.split chỉ định các chỉ số mà tại đó mảng sẽ được chia thành nhiều mảnh.

**Các hàm hỗ trợ xếp chồng: r\_ và c\_**

Có hai đối tượng đặc biệt trong không gian tên NumPy là r\_ và c\_, giúp cho việc xếp chồng các mảng trở nên ngắn gọn hơn:

In: arr = np.arange(6)

In: arr1 = arr.reshape((3, 2))

In: arr2 = rng.standard\_normal((3, 2))

In: np.r\_[arr1, arr2]

Out:

array([[ 0. , 1. ],

[ 2. , 3. ],

[ 4. , 5. ],

[ 2.3474, 0.9685],

[-0.7594, 0.9022],

[-0.467 , -0.0607]])

In: np.c\_[np.r\_[arr1, arr2], arr]

Out:

array([[ 0. , 1. , 0. ],  
 [ 2. , 3. , 1. ],

[ 4. , 5. , 2. ],

[ 2.3474, 0.9685, 3. ],

[-0.7594, 0.9022, 4. ],

[-0.467 , -0.0607, 5. ]])

Ngoài ra, chúng còn có thể chuyển đổi các lát cắt (slices) thành mảng:

In: np.c\_[1:6, -10:-5]

Out:

array([[ 1, -10],

[ 2, -9],

[ 3, -8],

[ 4, -7],

[ 5, -6]])

### 2.2.4. Repeating Elements (tile, repeat)

Hai công cụ hữu ích để lặp lại hoặc sao chép các mảng nhằm tạo ra các mảng lớn hơn là hàm repeat và tile. Hàm repeat lặp lại mỗi phần tử trong một mảng một số lần nhất định, tạo ra một mảng lớn hơn:

In: arr = np.arange(3)

In: arr

Out: array([0, 1, 2])

In: arr.repeat(3)

Out: array([0, 0, 0, 1, 1, 1, 2, 2, 2])

Theo mặc định, nếu bạn truyền vào một số nguyên, mỗi phần tử sẽ được lặp lại đúng số lần đó. Nếu bạn truyền vào một mảng các số nguyên, mỗi phần tử có thể được lặp lại một số lần khác nhau:

In: arr.repeat([2, 3, 4])

Out: array([0, 0, 1, 1, 1, 2, 2, 2, 2])

Đối với mảng đa chiều, các phần tử của chúng có thể được lặp lại dọc theo một trục cụ thể:

In: arr = rng.standard\_normal((2, 2))

In: arr

Out:

array([[ 0.7888, -1.2567],

[ 0.5759, 1.399 ]])

In: arr.repeat(2, axis=0)

Out:

array([[ 0.7888, -1.2567],

[ 0.7888, -1.2567],

[ 0.5759, 1.399 ],

[ 0.5759, 1.399 ]])

Lưu ý rằng nếu không có trục (axis) nào được truyền vào, mảng sẽ được làm phẳng trước, và đây có thể không phải là điều mong muốn. Tương tự, có thể truyền vào một mảng các số nguyên khi lặp lại một mảng đa chiều để lặp lại một lát cắt (slice) nhất định một số lần khác nhau:

In: arr.repeat([2, 3], axis=0)

Out:

array([[ 0.7888, -1.2567],

[ 0.7888, -1.2567],

[ 0.5759, 1.399 ],

[ 0.5759, 1.399 ],

[ 0.5759, 1.399 ]])

In: arr.repeat([2, 3], axis=1)

Out:

array([[ 0.7888, 0.7888, -1.2567, -1.2567, -1.2567],

[ 0.5759, 0.5759, 1.399 , 1.399 , 1.399 ]])

Mặt khác, hàm tile là một cách viết tắt cho việc xếp chồng các bản sao của một mảng dọc theo một trục.

Về mặt trực quan, có thể hình dung nó tương tự như việc 'lát gạch':

In: arr

Out:

array([[ 0.7888, -1.2567],

[ 0.5759, 1.399 ]])

In: np.tile(arr, 2)

Out:

array([[ 0.7888, -1.2567, 0.7888, -1.2567],

[ 0.5759, 1.399 , 0.5759, 1.399 ]])

Tham số thứ hai là số lần lát gạch; với một số vô hướng, việc lát gạch được thực hiện theo hàng, chứ không phải theo cột. Tham số thứ hai của hàm tile cũng có thể là một tuple để chỉ định cách bố trí của việc 'lát gạch':

In: arr

Out:

array([[ 0.7888, -1.2567],

[ 0.5759, 1.399 ]])

In: np.tile(arr, (2, 1))

Out:

array([[ 0.7888, -1.2567],

[ 0.5759, 1.399 ],

[ 0.7888, -1.2567],

[ 0.5759, 1.399 ]])

In: np.tile(arr, (3, 2))

Out:

array([[ 0.7888, -1.2567, 0.7888, -1.2567],

[ 0.5759, 1.399 , 0.5759, 1.399 ],

[ 0.7888, -1.2567, 0.7888, -1.2567],

[ 0.5759, 1.399 , 0.5759, 1.399 ],

[ 0.7888, -1.2567, 0.7888, -1.2567],

[ 0.5759, 1.399 , 0.5759, 1.399 ]])

### 2.2.5. Fancy Indexing Equivalents: take and put

Một cách để lấy và thiết lập các tập con của mảng là sử dụng "lập chỉ mục nâng cao" (fancy indexing) bằng các mảng số nguyên

In: arr = np.arange(10) \* 100

In: inds = [7, 1, 2, 6]

In: arr[inds]

Out: array([700, 100, 200, 600])

Có các phương thức ndarray khác hữu ích trong trường hợp đặc biệt khi chỉ cần lựa chọn trên một trục duy nhất:

In: arr.take(inds)

Out: array([700, 100, 200, 600])

In: arr.put(inds, 42)

In: arr

Out: array([ 0, 42, 42, 300, 400, 500, 42, 42, 800, 900])

In: arr.put(inds, [40, 41, 42, 43])

In: arr

Out: array([ 0, 41, 42, 300, 400, 500, 43, 40, 800, 900])

Để dùng take trên các trục khác, bạn có thể truyền vào từ khóa axis:

In: inds = [2, 0, 2, 1]

In: arr = rng.standard\_normal((2, 4))

In: arr

Out:

array([[ 1.3223, -0.2997, 0.9029, -1.6216],

[-0.1582, 0.4495, -1.3436, -0.0817]])

In: arr.take(inds, axis=1)

Out:

array([[ 0.9029, 1.3223, 0.9029, -0.2997],

[-1.3436, -0.1582, -1.3436, 0.4495]])

Hàm put không chấp nhận tham số axis, thay vào đó, nó lập chỉ mục trên phiên bản đã được làm phẳng (một chiều, theo thứ tự C) của mảng. Do đó, khi cần thiết lập các phần tử bằng một mảng chỉ số trên các trục khác, tốt nhất là sử dụng cách lập chỉ mục bằng dấu ngoặc vuông []

## 2.3. Mục A.3: Broadcasting (Quy tắc Lan truyền)

Broadcasting quy định cách các phép toán hoạt động giữa những mảng có hình dạng (shape) khác nhau. Đây là một tính năng mạnh mẽ nhưng có thể gây bối rối, ngay cả với những người dùng có kinh nghiệm. Ví dụ đơn giản nhất về broadcasting xảy ra khi kết hợp một giá trị vô hướng với một mảng:

In: arr = np.arange(5)

In: arr

Out: array([0, 1, 2, 3, 4])

In: arr \* 4

Out: array([ 0, 4, 8, 12, 16])

Ở đây, ta nói rằng giá trị vô hướng 4 đã được "lan truyền" (broadcast) tới tất cả các phần tử khác trong phép nhân.

Ví dụ, ta có thể "khử trung bình" (demean) mỗi cột của một mảng bằng cách trừ đi giá trị trung bình của các cột đó. Trong trường hợp này, chỉ cần trừ đi một mảng chứa giá trị trung bình của mỗi cột là đủ:

In [84]: arr = rng.standard\_normal((4, 3))

In [85]: arr.mean(0)

Out[85]: array([0.1206, 0.243 , 0.1444])

In [86]: demeaned = arr - arr.mean(0)

In [87]: demeaned

Out[87]:

array([[ 1.6042, 2.3751, 0.633 ],

[ 0.7081, -1.202 , -1.3538],

[-1.5329, 0.2985, 0.6076],

[-0.7793, -1.4717, 0.1132]])

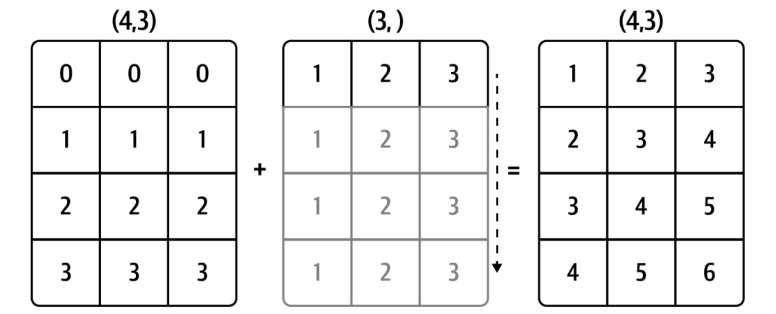
In [88]: demeaned.mean(0)

Out[88]: array([ 0., -0., 0.])

Thực hiện phép toán giữa các mảng có hình dạng không tương thích (ví dụ: một mảng 2 chiều với một mảng 1 chiều chứa giá trị trung bình của các hàng) được gọi là broadcasting. Thao tác này hoàn toàn có thể thực hiện được nhưng phải tuân theo những quy tắc cụ thể để đảm bảo tính toán đúng.

**Quy tắc Lan truyền (The Broadcasting Rule)**

Hai mảng được coi là tương thích để lan truyền (broadcasting) nếu với mỗi chiều kích thước cuối (xét từ phải sang trái), độ dài của các trục khớp nhau, hoặc một trong hai có độ dài bằng 1. Phép lan truyền sau đó sẽ được thực hiện trên các chiều bị thiếu hoặc các chiều có độ dài bằng 1



Hình 2.3.1: Lan truyền trên trục 0 với một mảng 1D

Hãy xem xét ví dụ trước: giả sử thay vì trừ trung bình cột, chúng ta muốn trừ đi giá trị trung bình của mỗi hàng. Vì arr.mean(0) (trung bình cột) có độ dài 3, nó tương thích để lan truyền trên axis 0 vì chiều cuối cùng của arr cũng là 3.

Tuy nhiên, theo quy tắc, để thực hiện phép trừ trên axis 1 (tức là trừ đi giá trị trung bình của hàng), mảng nhỏ hơn phải có hình dạng là (4, 1):

In : arr

Out[89]:

array([[ 1.7247, 2.6182, 0.7774],

[ 0.8286, -0.959 , -1.2094],

[-1.4123, 0.5415, 0.7519],

[-0.6588, -1.2287, 0.2576]])

In: row\_means = arr.mean(1)

In: row\_means.shape

Out: (4,)

In: row\_means.reshape((4, 1))

Out:

array([[ 1.7068],

[-0.4466],

[-0.0396],

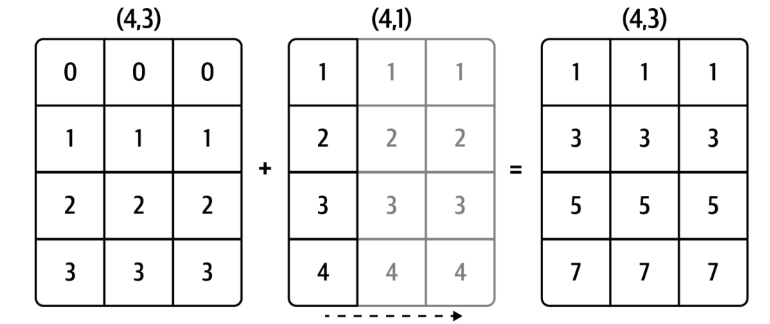
[-0.5433]])

In: demeaned = arr - row\_means.reshape((4, 1))

In: demeaned.mean(1)

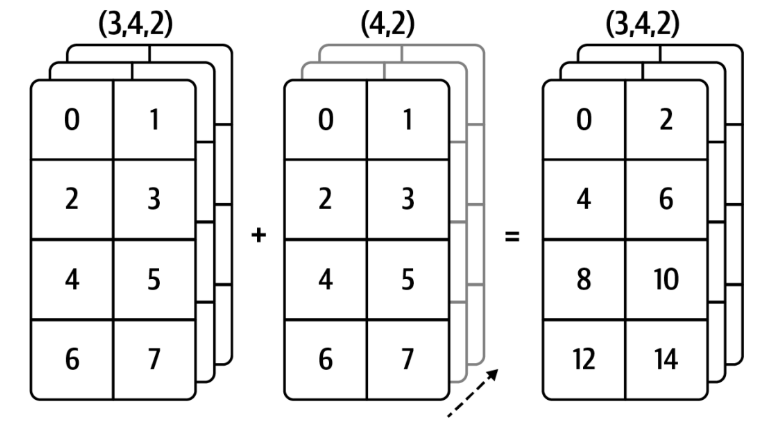
Out: array([-0., 0., 0., 0.])

Hình bên dưới sẽ minh hoạ cho thao tác này:



Hình 2.3.2: Lan truyền trên trục 1 của một mảng 2D

Xem bên dưới để có một minh họa khác, lần này là về việc cộng một mảng hai chiều vào một mảng ba chiều trên trục 0.



Hình 2.3.3: Lan truyền trên trục 0 của một mảng 3D.

### 2.3.1. Broadcasting trên các trục khác

Việc thực hiện một phép toán số học với một mảng có chiều thấp hơn trên các trục khác ngoài trục 0 là điều khá phổ biến. Theo quy tắc lan truyền (broadcasting), các "chiều lan truyền" trong mảng nhỏ hơn phải có giá trị là 1. Trong ví dụ về việc trừ trung bình hàng được trình bày ở đây, điều này có nghĩa là phải định hình lại mảng chứa trung bình các hàng thành hình dạng (4, 1) thay vì (4,):

In: arr - arr.mean(1).reshape((4, 1))

Out:

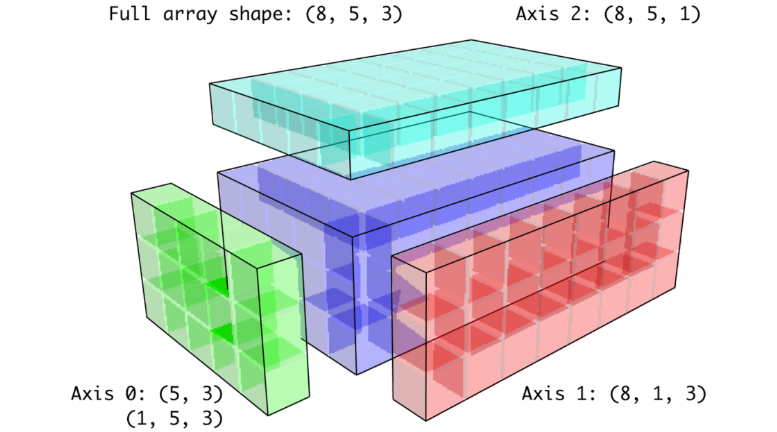
array([[ 0.018 , 0.9114, -0.9294],

[ 1.2752, -0.5124, -0.7628],

[-1.3727, 0.5811, 0.7915],

[-0.1155, -0.6854, 0.8009]])

Trong trường hợp ba chiều, việc broadcast (lan truyền/phát sóng) trên bất kỳ chiều nào trong ba chiều chỉ là vấn đề định hình lại dữ liệu để có hình dạng tương thích



Hình 2.3.4:

Do đó, một vấn đề phổ biến là cần phải thêm một trục mới với độ dài bằng 1 cụ thể cho mục đích broadcasting. Sử dụng hàm reshape là một lựa chọn, nhưng việc chèn một trục đòi hỏi phải xây dựng một tuple chỉ rõ hình dạng mới. Điều này thường có thể là một công việc tẻ nhạt. Vì vậy, các mảng NumPy cung cấp một cú pháp đặc biệt để chèn các trục mới thông qua việc đánh chỉ mục (indexing). Chúng ta sử dụng thuộc tính đặc biệt np.newaxis cùng với các lát cắt "đầy đủ" (full slices) để chèn trục mới:

In: arr = np.zeros((4, 4))

In: arr\_3d = arr[:, np.newaxis, :]

In: arr\_3d.shape

Out: (4, 1, 4)

In: arr\_1d = rng.standard\_normal(3)

In: arr\_1d[:, np.newaxis]

Out:

array([[ 0.3129],

[-0.1308],

[ 1.27 ]])

In: arr\_1d[np.newaxis, :]

Out: array([[ 0.3129, -0.1308, 1.27 ]])

Do đó, nếu chúng ta có một mảng ba chiều và muốn trừ đi giá trị trung bình (demean) của trục 2, chúng ta sẽ cần viết:

In: arr = rng.standard\_normal((3, 4, 5))

In: depth\_means = arr.mean(2)

In: depth\_means

Out:

array([[ 0.0431, 0.2747, -0.1885, -0.2014],

[-0.5732, -0.5467, 0.1183, -0.6301],

[ 0.0972, 0.5954, 0.0331, -0.6002]])

In: depth\_means.shape

Out: (3, 4)

In: demeaned = arr - depth\_means[:, :, np.newaxis]

In: demeaned.mean(2)

Out:

array([[ 0., -0., 0., -0.],

[ 0., -0., -0., -0.],

Bạn có thể đang tự hỏi liệu có một cách nào để khái quát hóa việc trừ giá trị trung bình (demeaning) trên một trục mà không làm giảm hiệu suất hay không. Câu trả lời là có, nhưng nó đòi hỏi một số kỹ thuật đánh chỉ mục (indexing gymnastics):

def demean\_axis(arr, axis=0):

means = arr.mean(axis)

# This generalizes things like [:, :, np.newaxis] to N dimensions

indexer = [slice(None)] \* arr.ndim

indexer[axis] = np.newaxis

return arr - means[indexer]

### 2.3.2. Thiết lập Giá trị Mảng bằng Broadcasting

Quy tắc broadcasting chi phối các phép toán số học cũng áp dụng để thiết lập giá trị thông qua việc đánh chỉ mục mảng. Trong một trường hợp đơn giản, chúng ta có thể làm những việc như:

In: arr = np.zeros((4, 3))

In: arr[:] = 5

In: arr

Out:

array([[5., 5., 5.],

[5., 5., 5.],

[5., 5., 5.],

[5., 5., 5.]])

Tuy nhiên, nếu chúng ta có một mảng một chiều chứa các giá trị mà chúng ta muốn gán vào các cột của mảng lớn hơn, chúng ta có thể làm điều đó miễn là hình dạng tương thích:

In: col = np.array([1.28, -0.42, 0.44, 1.6])

In: arr[:] = col[:, np.newaxis]

In: arr

Out:

array([[ 1.28, 1.28, 1.28],

[-0.42, -0.42, -0.42],

[ 0.44, 0.44, 0.44],

[ 1.6 , 1.6 , 1.6 ]])

In: arr[:2] = [[-1.37], [0.509]]

In: arr

Out:

array([[-1.37 , -1.37 , -1.37 ],

[ 0.509, 0.509, 0.509],

[ 0.44 , 0.44 , 0.44 ],

[ 1.6 , 1.6 , 1.6 ]])

## 2.4. Mục A.4: Sử dụng ufunc Nâng cao

### 2.4.1. Các phương thức của ufunc (reduce, accumulate, outer, reduceat)

## 2.5. Mục A.5: Mảng có Cấu trúc (Structured Arrays)

Bạn có thể đã nhận thấy cho đến bây giờ rằng ndarray là một kiểu dữ liệu đồng nhất; nghĩa là, nó biểu diễn một khối bộ nhớ mà mỗi phần tử chiếm cùng một số byte, được xác định bởi kiểu dữ liệu. Nhìn bề ngoài, điều này có vẻ như không cho phép bạn biểu diễn dữ liệu không đồng nhất hoặc dạng bảng. Một mảng có cấu trúc (structured array) là một ndarray mà mỗi phần tử có thể được xem như là một struct trong C (do đó có tên “structured”) hoặc một dòng trong bảng SQL với nhiều trường có tên khác nhau:

In: dtype = [('x', np.float64), ('y', np.int32)]

In: sarr = np.array([(1.5, 6), (np.pi, -2)], dtype=dtype)

In: sarr

Out: array([(1.5 , 6), (3.1416, -2)], dtype=[('x', '<f8'), ('y', '<i4')])

Có nhiều cách để xác định kiểu dữ liệu có cấu trúc (xem tài liệu trực tuyến của NumPy). Một cách phổ biến là sử dụng danh sách các tuple với dạng (tên\_trường, kiểu\_dữ\_liệu\_trường). Bây giờ, các phần tử của mảng là các đối tượng giống tuple, các phần tử của chúng có thể được truy cập như một từ điển:

In: sarr[0]

Out: np.void((1.5, 6), dtype=[('x', '<f8'), ('y', '<i4')])

In : sarr[0]['y']

Out: np.int32(6)

Tên các trường được lưu trong thuộc tính dtype.names. Khi bạn truy cập một trường trên mảng có cấu trúc, bạn sẽ nhận được một view dạng strided trên dữ liệu, do đó không có dữ liệu nào bị sao chép:

In: sarr['x']

Out: array([1.5 , 3.14159265])

# 

# KẾT LUẬN

Qua quá trình thực hiện dự án "Lab 02", nhóm chúng đã hoàn thành mục tiêu đề ra là tái hiện và phân tích các bài toán dữ liệu dựa trên tài liệu được cung cấp. Dự án đã mang lại những kết quả giá trị, không chỉ về mặt kỹ thuật mà còn về kỹ năng làm việc nhóm và quản lý dự án.

Trong Phần 1, việc tái hiện thành công năm bộ dữ liệu đã giúp nhóm củng cố vững chắc quy trình phân tích dữ liệu thực tế, từ bước tiền xử lý, làm sạch dữ liệu đến các bước tổng hợp và trực quan hóa. Việc làm việc với các bộ dữ liệu đa dạng đã mang lại cái nhìn sâu sắc về cách áp dụng các kỹ thuật phù hợp cho từng bài toán cụ thể.

Song song đó, Phần 2 của báo cáo đã hệ thống hóa lại các khái niệm kỹ thuật nâng cao của thư viện NumPy. Quá trình này giúp làm rõ nền tảng lý thuyết đằng sau các thao tác hiệu suất cao, qua đó củng cố sự hiểu biết của nhóm về công cụ cốt lõi trong hệ sinh thái khoa học dữ liệu Python.

Nhìn chung, dự án là một cơ hội quý báu để nhóm chuyển hóa kiến thức lý thuyết thành kỹ năng thực hành. Những kinh nghiệm thu được từ việc xử lý các vấn đề kỹ thuật và phân tích kết quả sẽ là nền tảng vững chắc cho các dự án phân tích dữ liệu phức tạp hơn trong tương lai.