Trường Đại học Khoa học tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

Ngành Công nghệ thông tin

Nhóm 4

Học phần: Trực quan hóa dữ liệu

Sinh viên:
Trần Nguyễn Nhật Cường (22127048)
Nguyễn Công Tuấn (22127436)
Trần Đăng Tuấn (22127438)

Giảng viên:
Bùi Tiến Lên
Võ Nhật Tân
Lê Ngọc Thành

Ngày 24 tháng 3 năm 2025

Mục lục

1 Thông tin chung

1.1 Thông tin nhóm

Họ tên	**MSSV**
Trần Nguyễn Nhật Cường	22127048
Nguyễn Công Tuấn	22127436
Trần Đăng Tuấn	22127438

Bảng 1: Thông tin các thành viên

1.2 Mức độ hoàn thành tổng thể của mỗi yêu cầu

Yêu cầu	**Mức độ hoàn thành (%)**
Thu thập dữ liệu	100%
Tiền xử lý dữ liệu	100%
Trực quan hóa dữ liệu	100%

Bảng 2: Mức độ hoàn thành tổng thể

1.3 Mức độ hoàn thành của từng thành viên

Công việc	**Thành viên phụ trách**	**Mức độ hoàn thành**
Thu thập dữ liệu	Trần Nguyễn Nhật Cường, Nguyễn Công Tuấn	100%
Tiền xử lý dữ liệu	Trần Nguyễn Nhật Cường, Nguyễn Công Tuấn	100%
Phân tích cơ bản về dữ liệu	Trần Đăng Tuấn	100%
Xác định mục tiêu phân tích và lựa chọn các trường dữ liệu	Tất cả thành viên	100%
Phân tích, nhận xét và đánh giá dữ liệu trên biểu đồ thu được	Tất cả thành viên	100%
Viết báo cáo	Trần Nguyễn Nhật Cường, Nguyễn Công Tuấn	100%

Bảng 3: Mức độ hoàn thành của từng thành viên

2 Giới thiệu đồ án

2.1 Các thư viện sử dụng trong đồ án

- **matplotlib** được dùng để vẽ biểu đồ trong Python. Có thể dùng để vẽ các loại biểu đồ như histogram, biểu đồ tán xạ (scatter plot), biểu đồ đường (line chart) và biểu đồ cột (bar chart), ...
- **pandas** được dùng để xử lý và phân tích dữ liệu dạng bảng và cung cấp cấu trúc dữ liệu như DataFrame và Series để dễ thao tác.
- **seaborn** dùng để vẽ biểu đồ dựa trên thư viện **matplotlib** nhưng biểu đồ sẽ trông đẹp hơn và cũng hỗ trợ các loại biểu đồ thống kê như violin plot, box plot, heatmap.
- **wbgapi** dùng để lấy/truy cập dữ liệu từ ngân hàng thế giới (World Bank).
- **sklearn** (scikit-learn) là thư viện học máy phổ biến trong Python, được sử dụng để thực hiện các tác vụ học máy và xử lý dữ liệu. Trong đồ án này, các thành phần sau của sklearn được sử dụng:
  - `sklearn.cluster.KMeans`: Cung cấp thuật toán KMeans để gom cụm dữ liệu thành các cụm dựa trên sự tương đồng, được sử dụng để phân loại các nền kinh tế thành các cụm.
  - `sklearn.metrics.silhouette_score`: Dùng để tính điểm Silhouette, một chỉ số đánh giá chất lượng của việc gom cụm, giúp xác định số lượng cụm tối ưu bằng cách đo lường mức độ tách biệt giữa các cụm.
  - `sklearn.preprocessing.StandardScaler`: Dùng để chuẩn hóa dữ liệu, đưa các giá trị về cùng thang đo (trung bình bằng 0 và độ lệch chuẩn bằng 1), đảm bảo các chỉ số kinh tế có trọng số công bằng khi áp dụng thuật toán gom cụm.

3 Nhiệm vụ đồ án

3.1 Thu thập dữ liệu

Dữ liệu được thu thập qua **World Developemt Indicators** từ **World Bank** bằng việc gọi **API** được hỗ trợ bởi package **wbgapi**.

Nhóm đã lấy được thông tin về những chỉ số phát triển theo chủ đề

1. Giáo dục (Education)
2. Biến đổi khí hậu (Climate change)
3. Phát triển kinh tế (Economic and Growth)

của ba quốc gia là Việt Nam (VNM), Hoa Kỳ (USA) và Nhật Bản (JPN) từ năm 2000 đến năm 2020 và lần lượt lưu vào các file

1. wdi_2000_2020_VNM.csv
2. wdi_2000_2020_USA.csv
3. wdi_2000_2020_JPN.csv

Bảng chỉ số phát triển của tập dữ liệu được thu thập bao gồm:

Các chỉ số Giáo dục

Mã chỉ số	**Tên chỉ số**	**Mô tả**
SE.PRM.UNER.FE	Trẻ em nữ không đi học ở cấp tiểu học	Số lượng trẻ em gái trong độ tuổi đi học tiểu học nhưng không đi học
SE.PRM.UNER.MA	Trẻ em nam không đi học, cấp tiểu học	Số lượng trẻ em trai trong độ tuổi đi học tiểu học nhưng không đi học
SE.XPD.TOTL.GD.ZS	Chi tiêu công cho giáo dục (% GDP)	Tổng chi tiêu của chính phủ cho giáo dục tính theo phần trăm GDP
SE.XPD.PRIM.PC.ZS	Chi tiêu công trên mỗi học sinh tiểu học (% GDP bình quân đầu người)	Tỷ lệ chi tiêu công cho mỗi học sinh cấp tiểu học so với GDP bình quân đầu người
SE.XPD.SECO.PC.ZS	Chi tiêu công trên mỗi học sinh trung học (% GDP bình quân đầu người)	Tỷ lệ chi tiêu công cho mỗi học sinh cấp trung học so với GDP bình quân đầu người
SE.XPD.TERT.PC.ZS	Chi tiêu công trên mỗi sinh viên đại học (% GDP bình quân đầu người)	Tỷ lệ chi tiêu công cho mỗi sinh viên đại học so với GDP bình quân đầu người
SE.PRM.GINT.FE.ZS	Tỷ lệ tuyển sinh lớp 1, nữ	Số lượng học sinh nữ nhập học lớp 1 trên tổng số trẻ em nữ trong độ tuổi tương ứng (tính theo phần trăm)
SE.PRM.GINT.MA.ZS	Tỷ lệ tuyển sinh lớp 1, nam	Số lượng học sinh nam nhập học lớp 1 trên tổng số trẻ em nam trong độ tuổi tương ứng (tính theo phần trăm)
SL.TLF.TOTL.FE.ZS	Lực lượng lao động nữ (% tổng lực lượng lao động)	Tỷ lệ lao động nữ trong tổng lực lượng lao động của quốc gia
SL.TLF.TOTL.IN	Tổng lực lượng lao động	Tổng số người tham gia vào lực lượng lao động của quốc gia, bao gồm cả người có việc làm và người thất nghiệp đang tìm việc

Mã chỉ số	**Tên chỉ số**	**Mô tả**
SE.ADT.LITR.FE.ZS	Tỷ lệ biết chữ ở người lớn, nữ	Tỷ lệ phần trăm phụ nữ từ 15 tuổi trở lên có thể đọc và viết
SE.ADT.LITR.MA.ZS	Tỷ lệ biết chữ ở người lớn, nam	Tỷ lệ phần trăm nam giới từ 15 tuổi trở lên có thể đọc và viết
SE.ADT.1524.LT.FE.ZS	Tỷ lệ biết chữ ở thanh niên, nữ	Tỷ lệ phần trăm nữ từ 15-24 tuổi có thể đọc và viết
SE.ADT.1524.LT.MA.ZS	Tỷ lệ biết chữ ở thanh niên, nam	Tỷ lệ phần trăm nam từ 15-24 tuổi có thể đọc và viết
SE.PRM.ENRL.TC.ZS	Tỷ lệ học sinh trên giáo viên cấp tiểu học	Số lượng học sinh trung bình trên một giáo viên cấp tiểu học
SE.PRM.REPT.FE.ZS	Tỷ lệ học sinh nữ lưu ban cấp tiểu học	Phần trăm học sinh nữ bị lưu ban tại cấp tiểu học
SE.PRM.REPT.MA.ZS	Tỷ lệ học sinh nam lưu ban cấp tiểu học	Phần trăm học sinh nam bị lưu ban tại cấp tiểu học
SE.PRE.ENRR	Tỷ lệ nhập học mầm non (% tổng số)	Tỷ lệ tổng số trẻ em nhập học mầm non so với tổng số trẻ em trong độ tuổi mầm non
SE.PRM.ENRR	Tỷ lệ nhập học tiểu học (% tổng số)	Tỷ lệ tổng số học sinh nhập học cấp tiểu học so với tổng số trẻ em trong độ tuổi tiểu học
SE.SEC.ENRR	Tỷ lệ nhập học trung học (% số lượng thực tế)	Tỷ lệ tổng số học sinh nhập học trung học so với số trẻ em trong độ tuổi tương ứng
SE.PRM.TCAQ.ZS	Tỷ lệ giáo viên tiểu học được đào tạo	Phần trăm giáo viên tiểu học có bằng cấp đào tạo chính thức
SL.UEM.TOTL.FE.ZS	Tỷ lệ thất nghiệp ở nữ	Phần trăm nữ giới trong lực lượng lao động không có việc làm nhưng đang tìm kiếm việc làm
SL.UEM.TOTL.ZS	Tổng tỷ lệ thất nghiệp	Phần trăm tổng trong lực lượng lao động không có việc làm nhưng đang tìm kiếm việc làm

Các chỉ số Biến đổi khí hậu

Mã chỉ số	Tên chỉ số	Mô tả
EG.ELC.ACCS.ZS	Dùng điện năng (% dân số)	Tỷ lệ phần trăm dân số có quyền truy cập vào điện
AG.LND.IRIG.AG.ZS	Đất nông nghiệp có tưới tiêu (% tổng diện tích đất nông nghiệp)	Diện tích đất nông nghiệp được tưới tiêu trên tổng diện tích đất nông nghiệp
AG.LND.AGRI.ZS	Đất nông nghiệp (% tổng diện tích đất)	Phần trăm diện tích đất được sử dụng cho mục đích nông nghiệp so với tổng diện tích đất
NV.AGR.TOTL.ZS	Giá trị gia tăng của nông, lâm, ngư nghiệp (% GDP)	Tỷ lệ đóng góp của nông nghiệp, lâm nghiệp và thủy sản vào tổng sản phẩm quốc nội (GDP)
ER.H2O.FWTL.K3	Tổng lượng nước ngọt dùng hàng năm (tỷ m^3)	Lượng nước ngọt được sử dụng hàng năm cho mục đích sinh hoạt, công nghiệp và nông nghiệp
AG.LND.ARBL.ZS	Đất canh tác (% tổng diện tích đất)	Tỷ lệ phần trăm đất có thể trồng trọt trên tổng diện tích đất
AG.YLD.CREL.KG	Sản lượng ngũ cốc (kg/ha)	Sản lượng ngũ cốc trung bình trên mỗi hecta đất trồng trọt
EG.USE.ELEC.KH.PC	Tiêu thụ điện bình quân đầu người (kWh/người)	Lượng điện năng tiêu thụ trung bình mỗi người trong một năm
EG.USE.PCAP.KG.OE	Tiêu thụ năng lượng bình quân đầu người (kg dầu tương đương/người)	Lượng năng lượng trung bình mỗi người tiêu thụ hàng năm, tính theo đơn vị kg dầu tương đương
AG.LND.FRST.ZS	Diện tích rừng (% tổng diện tích đất)	Tỷ lệ diện tích rừng so với tổng diện tích đất của quốc gia
AG.LND.FRST.K2	Tổng diện tích rừng (km^2)	Tổng diện tích đất được bao phủ bởi rừng, tính theo đơn vị km^2
AG.LND.EL5M.ZS	Đất thấp dưới 5 mét so với mực nước biển (% tổng diện tích đất)	Phần trăm tổng diện tích đất nằm ở độ cao dưới 5 mét so với mực nước biển

Mã chỉ số	**Tên chỉ số**	**Mô tả**
SP.POP.GROW	Tốc độ tăng dân số (% hàng năm)	Tốc độ thay đổi dân số theo tỷ lệ phần trăm hàng năm
EN.POP.EL5M.ZS	Dân số sống ở vùng thấp dưới 5 mét so với mực nước biển (% tổng dân số)	Tỷ lệ phần trăm dân số sinh sống ở khu vực có độ cao dưới 5 mét so với mực nước biển
SP.POP.TOTL	Tổng dân số	Tổng số người sinh sống trong một quốc gia hoặc vùng lãnh thổ tại một thời điểm nhất định
EG.ELC.RNEW.ZS	Sản lượng điện tái tạo (% tổng sản lượng điện)	Tỷ lệ điện sản xuất từ các nguồn năng lượng tái tạo (như gió, mặt trời, thủy điện) so với tổng lượng điện sản xuất
EG.FEC.RNEW.ZS	Tiêu thụ năng lượng tái tạo (% tổng tiêu thụ năng lượng cuối cùng)	Tỷ lệ năng lượng tái tạo trong tổng mức tiêu thụ năng lượng cuối cùng
ER.PTD.TOTL.ZS	Khu vực bảo tồn trên cạn và trên biển (% tổng diện tích lãnh thổ)	Phần trăm diện tích lãnh thổ được bảo vệ trên cạn và biển, bao gồm các khu bảo tồn thiên nhiên, vườn quốc gia
SP.URB.TOTL	Dân số đô thị	Tổng số người sinh sống tại khu vực đô thị
SP.URB.TOTL.IN.ZS	Dân số đô thị (% tổng dân số)	Phần trăm dân số sống tại khu vực đô thị so với tổng dân số của quốc gia

Các chỉ số phát triển kinh tế

Mã chỉ số	Tên chỉ số	Mô tả
NY.ADJ.SVNG.GN.ZS	Tiết kiệm ròng điều chỉnh, bao gồm thiệt hại do phát thải hạt (% GNI)	Chỉ số đo lường mức tiết kiệm ròng sau khi điều chỉnh cho suy giảm tài nguyên thiên nhiên và thiệt hại do ô nhiễm không khí
NV.AGR.TOTL.ZS	Giá trị gia tăng của nông, lâm, ngư nghiệp (% GDP)	Đóng góp của ngành nông nghiệp, lâm nghiệp và thủy sản vào GDP
GC.DOD.TOTL.GD.ZS	Nợ chính phủ trung ương (% GDP)	Tổng nợ của chính phủ trung ương so với GDP của quốc gia
BM.GSR.ROYL.CD	Thanh toán sử dụng quyền sở hữu trí tuệ (BoP, Tính theo tỷ giá đồng Dollar hiện tại)	Tổng số tiền một quốc gia thanh toán cho việc sử dụng bằng sáng chế, bản quyền, thương hiệu và các tài sản trí tuệ khác
BX.GSR.ROYL.CD	Thu nhập từ quyền sở hữu trí tuệ (BoP, Tính theo tỷ giá đồng Dollar hiện tại)	Tổng số tiền một quốc gia nhận được từ việc cấp phép sử dụng tài sản trí tuệ
BN.CAB.XOKA.CD	Cán cân tài khoản vãng lai (BoP, Tính theo tỷ giá đồng Dollar hiện tại)	Sự chênh lệch giữa xuất nhập khẩu hàng hóa, dịch vụ, thu nhập ròng và chuyển nhượng ròng
GC.XPN.TOTL.GD.ZS	Chi tiêu của chính phủ (% GDP)	Tổng chi tiêu của chính phủ so với GDP
NE.EXP.GNFS.ZS	Xuất khẩu hàng hóa và dịch vụ (% GDP)	Tổng giá trị xuất khẩu hàng hóa và dịch vụ so với GDP
DT.DOD.DECT.GN.ZS	Tổng nợ nước ngoài (% GNI)	Tổng giá trị nợ nước ngoài của một quốc gia so với tổng thu nhập quốc dân
DT.DOD.DECT.CD	Tổng nợ nước ngoài (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng giá trị nợ nước ngoài của quốc gia, tính bằng đô la Mỹ
BX.KLT.DINV.CD.WD	Đầu tư trực tiếp nước ngoài, dòng vốn vào ròng (BoP, Tính theo tỷ giá đồng Dollar hiện tại)	Tổng số vốn đầu tư trực tiếp từ nước ngoài vào quốc gia

Mã chỉ số	**Tên chỉ số**	**Mô tả**
NY.GDP.MKTP.CD	GDP (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng sản phẩm quốc nội (GDP) của một quốc gia theo giá trị thị trường hiện tại
NY.GDP.MKTP.KD.ZG	Tăng trưởng GDP (% hàng năm)	Tỷ lệ tăng trưởng hàng năm của GDP, điều chỉnh theo lạm phát
NY.GDP.PCAP.CD	GDP bình quân đầu người (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng GDP chia cho dân số của quốc gia
NY.GDP.PCAP.KD.ZG	Tăng trưởng GDP bình quân đầu người (% hàng năm)	Tốc độ tăng trưởng của GDP bình quân đầu người theo giá thực tế
NY.GDP.PCAP.PP.CD	GDP bình quân đầu người theo sức mua tương đương (PPP, USD quốc tế)	GDP bình quân đầu người được điều chỉnh theo ngang giá sức mua (PPP)
NY.GNP.PCAP.CD	GNI bình quân đầu người, phương pháp Atlas (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng thu nhập quốc dân bình quân đầu người theo phương pháp Atlas
NY.GNP.PCAP.PP.CD	GNI bình quân đầu người theo PPP (USD quốc tế)	Tổng thu nhập quốc dân bình quân đầu người điều chỉnh theo ngang giá sức mua
NY.GNP.ATLS.CD	GNI, phương pháp Atlas (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng thu nhập quốc dân của quốc gia theo phương pháp Atlas
NY.GNP.MKTP.PP.CD	GNI theo PPP (USD quốc tế)	Tổng thu nhập quốc dân được điều chỉnh theo ngang giá sức mua
BX.GRT.EXTA.CD.WD	Viện trợ không hoàn lại, trừ hợp tác kỹ thuật (BoP, Tính theo tỷ giá đồng Dollar hiện tại)	Tổng giá trị viện trợ không hoàn lại mà một quốc gia nhận được, không bao gồm hợp tác kỹ thuật
NE.GDI.TOTL.ZS	Tổng đầu tư vốn (% GDP)	Tổng chi tiêu cho đầu tư vốn so với GDP
NY.GNS.ICTR.ZS	Tiết kiệm gộp (% GDP)	Phần trăm GDP được tiết kiệm sau khi trừ đi tiêu dùng của chính phủ và hộ gia đình

Mã chỉ số	**Tên chỉ số**	**Mô tả**
NE.IMP.GNFS.ZS	Nhập khẩu hàng hóa và dịch vụ (% GDP)	Tổng giá trị nhập khẩu hàng hóa và dịch vụ so với GDP
NV.IND.TOTL.ZS	Giá trị gia tăng của ngành công nghiệp (% GDP)	Đóng góp của ngành công nghiệp, bao gồm xây dựng, vào GDP
NY.GDP.DEFL.KD.ZG	Lạm phát, giảm phát GDP (% hàng năm)	Mức tăng giá chung của nền kinh tế được đo lường bằng chỉ số giảm phát GDP
FP.CPI.TOTL.ZG	Lạm phát, chỉ số giá tiêu dùng (% hàng năm)	Tỷ lệ lạm phát hàng năm dựa trên chỉ số giá tiêu dùng (CPI)
NV.MNF.TECH.ZS.UN	Giá trị gia tăng của ngành sản xuất công nghệ vừa và cao (% tổng giá trị gia tăng của ngành sản xuất)	Phần trăm giá trị gia tăng của ngành công nghiệp chế tạo có công nghệ vừa và cao
DT.ODA.ODAT.GN.ZS	Hỗ trợ phát triển nhận được (% GNI)	Tổng viện trợ phát triển chính thức mà quốc gia nhận được so với GNI
DT.ODA.ODAT.PC.ZS	Hỗ trợ phát triển nhận được trên đầu người (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng viện trợ phát triển chính thức mà quốc gia nhận được chia cho dân số
DT.ODA.ODAT.CD	Hỗ trợ phát triển nhận được (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng giá trị viện trợ phát triển chính thức mà quốc gia nhận được
PA.NUS.PPP	Hệ số chuyển đổi PPP, GDP (LCU trên 1 USD quốc tế)	Hệ số chuyển đổi từ đơn vị tiền tệ nội địa (LCU) sang đô la quốc tế theo phương pháp ngang giá sức mua (PPP)
BX.TRF.PWKR.CD.DT	Kiều hối cá nhân nhận được (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng giá trị kiều hối cá nhân mà một quốc gia nhận được từ công dân làm việc ở nước ngoài
PA.NUS.PRVT.PP	Tỷ lệ mức giá của hệ số chuyển đổi PPP (GDP) so với tỷ giá hối đoái thị trường	Tỷ lệ giữa hệ số chuyển đổi PPP và tỷ giá hối đoái thị trường, đo lường mức giá nội địa so với giá quốc tế
GC.REV.XGRT.GD.ZS	Doanh thu, không bao gồm viện trợ (% GDP)	Tổng doanh thu của chính phủ không bao gồm viện trợ quốc tế, tính theo % GDP

Mã chỉ số	Tên chỉ số	Mô tả
DT.DOD.DSTC.ZS	Nợ ngắn hạn (% tổng dự trữ)	Tổng nợ ngắn hạn của một quốc gia so với tổng dự trữ ngoại hối và vàng
BX.GRT.TECH.CD.WD	Viện trợ hợp tác kỹ thuật (BoP, Tính theo tỷ giá đồng Dollar hiện tại)	Tổng giá trị viện trợ quốc tế dành cho hợp tác kỹ thuật mà một quốc gia nhận được
DT.TDS.DECT.EX.ZS	Nghĩa vụ trả nợ (% xuất khẩu hàng hóa, dịch vụ và thu nhập sơ cấp)	Tổng nghĩa vụ trả nợ (cả gốc và lãi) của một quốc gia so với tổng giá trị xuất khẩu hàng hóa, dịch vụ và thu nhập từ nước ngoài
FI.RES.TOTL.CD	Dự trữ ngoại hối và vàng (Tính theo tỷ giá đồng Dollar hiện tại)	Tổng giá trị dự trữ ngoại tệ và vàng của quốc gia

Tổng quan về bộ dữ liệu trước khi tiền xử lý
Ví dụ: Năm dòng đầu của dữ liệu Việt Nam [wdi_2000_2020_VNM.csv] như sau:

series	YR2000	YR2001	YR2002	...	YR2019	YR2020
AG.LND.AGRI.ZS	28.23	30.48	30.45	...	39.52	39.43
AG.LND.ARBL.ZS	19.93	21.37	21.25	...	21.64	21.65
AG.LND.EL5M.ZS	15.94	NaN	NaN	...	NaN	NaN
AG.LND.FRST.K2	117841.00	119444.96	121048.92	...	145671.90	146430.90
AG.LND.FRST.ZS	37.88	38.40	38.98	...	46.48	46.72

Bảng 7: Năm dòng đầu của dữ liệu Việt Nam

Nhận xét chung về cả ba bộ dữ liệu

- Cả ba bộ dữ liệu đều có 81 dòng và 22 cột (thuộc tính).
- Tất cả các thuộc tính đều có kiểu dữ liệu là float64, ngoại trừ thuộc tính series có kiểu object.
- Các thuộc tính có dạng YRabcd, trong đó abcd là năm dữ liệu được thu thập. Ví dụ, YR2015 của chỉ số A có giá trị B, nghĩa là chỉ số A tại năm 2015 ghi nhận giá trị B.
- Hầu hết các năm đều có dữ liệu bị thiếu, có thể do dữ liệu không được thu thập hoặc không có sẵn. Do đó, nhóm không thực hiện điền giá trị thiếu, vì việc này có thể làm mất tính chính xác và không có cơ sở hợp lý để bổ sung.

3.2 Tiền xử lý dữ liệu

- **Định dạng lại ba bộ dữ liệu**
  - Mục đích: Chuyển dữ liệu từ dạng rộng (wide format) sang dạng bảng chuẩn, giúp dễ dàng phân tích theo năm và quốc gia.
  - Mô tả:
    1. Đọc dữ liệu với mỗi chỉ số (series) là một hàng, các năm (YR2000–YR2020) là cột vào cột Year và giá trị vào cột Value.
    2. Chuyển sang long format bằng cách dùng `melt()` để đưa năm vào cột Year và giá trị vào cột Value.
    3. Chuẩn hóa dữ liệu bằng cách chuyển Year về dạng số, thay dấu chấm (.) trong tên chỉ số bằng gạch dưới (_).
       * Ví dụ: **AG.LND.ARBL.ZS → AG_LND_ARBL_ZS**
    4. Pivot lại dữ liệu bẳng cách chuyển mỗi chỉ số thành một cột riêng, với Year làm cột chính.
    5. Lưu kết quả sau khi đã áp dụng cho từng quốc gia và ghi đè vào file tương ứng.

Tổng quan về dữ liệu (sau khi thực hiện định dạng lại bộ dữ liệu)
Ví dụ: Năm dòng đầu của dữ liệu Việt Nam [wdi_2000_2020_VNM.csv] sau khi đã được định dạng lại như sau:

Year	AG_LND_AGRI_ZS	...	SP_URB_TOTL_IN_ZS
2000	28.23	...	24.37
2001	30.48	...	24.94
2002	30.45	...	25.51
2003	30.76	...	26.09
2004	31.59	...	26.68

Bảng 8: Năm dòng đầu của dữ liệu Việt Nam sau khi đã định dạng lại

Nhận xét chung về cả ba bộ dữ liệu sau khi đã định dạng lại

- Dữ liệu có 21 dòng và 82 thuộc tính (cột).
- Tất cả các thuộc tính đều có kiểu dữ liệu là float64 trừ thuộc tính Year có kiểu dữ liệu là int32.

Với bộ dữ liệu của Việt Nam, các cột (thuộc tính) không bao gồm dữ liệu gồm

- AG_LND_IRIG_AG_ZS
- BM_GSR_ROYL_CD
- BX_GSR_ROYL_CD

- GC_DOD_TOTL_GD_ZS
- GC_REV_XGRT_GD_ZS
- GC_XPN_TOTL_GD_ZS
- SE_PRM_UNER_FE
- SE_PRM_UNER_MA
- SE_XPD_SECO_PC_ZS

Với bộ dữ liệu của Hoa Kỳ, các cột (thuộc tính) không bao gồm dữ liệu gồm

- BX_GRT_EXTA_CD_WD
- BX_GRT_TECH_CD_WD
- DT_DOD_DECT_CD
- DT_DOD_DECT_GN_ZS
- DT_DOD_DSTC_ZS
- DT_ODA_ODAT_CD
- DT_ODA_ODAT_GN_ZS
- DT_ODA_ODAT_PC_ZS
- DT_TDS_DECT_EX_ZS
- SE_ADT_1524_LT_FE_ZS
- SE_ADT_1524_LT_MA_ZS
- SE_ADT_LITR_FE_ZS
- SE_ADT_LITR_MA_ZS

Với bộ dữ liệu của Nhật Bản, các cột (thuộc tính) không bao gồm dữ liệu gồm

- BX_GRT_EXTA_CD_WD
- BX_GRT_TECH_CD_WD
- DT_DOD_DECT_CD
- DT_DOD_DECT_GN_ZS
- DT_DOD_DSTC_ZS
- DT_ODA_ODAT_CD
- DT_ODA_ODAT_GN_ZS

- DT_ODA_ODAT_PC_ZS
- DT_TDS_DECT_EX_ZS
- GC_REV_XGRT_GD_ZS
- SE_ADT_1524_LT_FE_ZS
- SE_ADT_1524_LT_MA_ZS
- SE_ADT_LITR_FE_ZS
- SE_ADT_LITR_MA_ZS
- SE_PRM_GINT_FE_ZS
- SE_PRM_GINT_MA_ZS
- SE_PRM_TCAQ_ZS
- SE_PRM_UNER_FE
- SE_PRM_UNER_MA

3.3 Trực quan hóa dữ liệu

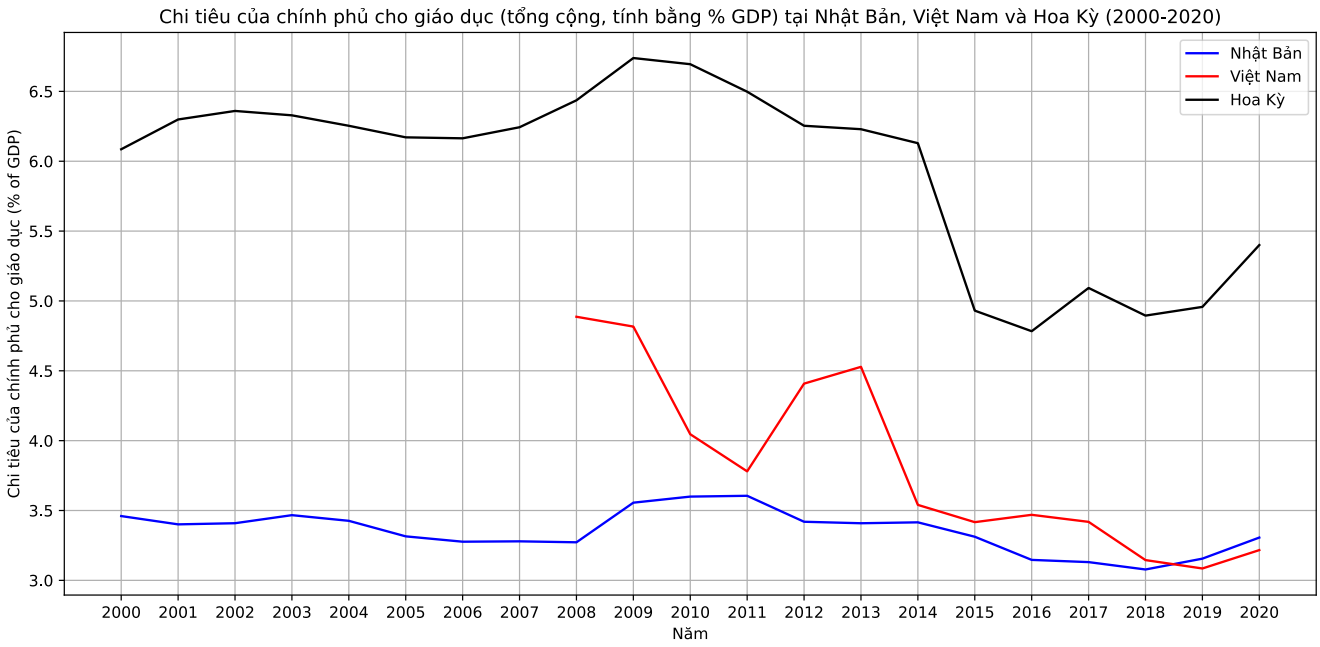
Phân tích cơ bản về dữ liệu

- **Giới thiệu về dữ liệu**: Ba tập dữ liệu chứa thông tin về các chỉ số phát triển của ba chủ đề là Giái dục, Biến đổi khí hậu và Phát triển Kinh tế trên World Development Indicators từ ngân hàng nhà nước (World Bank) của ba quốc gia là Việt Nam, Hoa Kỳ và Nhật Bản trong giai đoạn từ năm 2000 đến 2020.
- **Cỡ mẫu và cấu trúc**: Sau khi qua bước tiền xử lý dữ liệu để định dạng lại cả ba tập dữ liệu thì sẽ thu được ba bộ dữ liệu mới với **20** dữ liệu qua các năm của tất cả 82 chỉ số.

Các câu hỏi về chủ đề Giáo dục

Câu 1: Chi tiêu của chính phủ cho giáo dục (tổng cộng, tính bằng % GDP) tại Nhật Bản, Việt Nam và Hoa Kỳ đã thay đổi như thế nào từ năm 2000 đến năm 2020?

- **Mục tiêu phân tích**: So sánh mức chi tiêu giáo dục (% GDP) qua các năm của từng quốc gia và xác định xu hướng (tăng, giảm hoặc dao động) trong giai đoạn từ năm 2000 đến 2020.
- **Lựa chọn trường dữ liệu**: Sử dụng 2 trường dữ liệu là:
  - `SE_XPD_TOTL_GD_ZS` là tổng chi tiêu của chính phủ cho giáo dục (% GDP).
  - `Year` trong giai đoạn từ 2000 đến 2020.
- **Mối quan hệ giữa các trường**: `SE_XPD_TOTL_GD_ZS` (chi tiêu công cho giáo dục, % GDP) thay đổi theo `Year` (năm), với Year là biến thời gian và `SE_XPD_TOTL_GD_ZS` là biến phụ thuộc. Mối quan hệ này cho thấy xu hướng chi tiêu giáo dục tại Nhật Bản, Việt Nam và Hoa Kỳ từ 2000 đến 2020.
- **Loại biểu đồ sử dụng**: Biểu đồ đường (line chart)
- **Lý do chọn biểu đồ**:
  - Phù hợp để thể hiện xu hướng thay đổi theo thời gian.
  - Dễ dàng quan sát sự tăng giảm của mức chi tiêu giáo dục (% GDP) của từng quốc gia.
- **Các bước thực hiện**:
  1. Lọc dữ liệu (lấy thuộc tính `Year` và `SE_XPD_TOTL_GD_ZS`) từ bộ dữ liệu của Nhật Bản, Việt Nam và Hoa Kỳ.
  2. Vẽ biểu đồ đường để thể hiện sự thay đổi của chi tiêu giáo dục theo thời gian cho từng quốc gia.
  3. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ.
  4. Lưu biểu đồ dưới định dạng SVG và hiển thị kết quả
  5. Viết nhận xét.



Hình 1: Chi tiêu của chính phủ của các quốc gia cho giáo dục từ năm 2000 đến 2020

Nhận xét

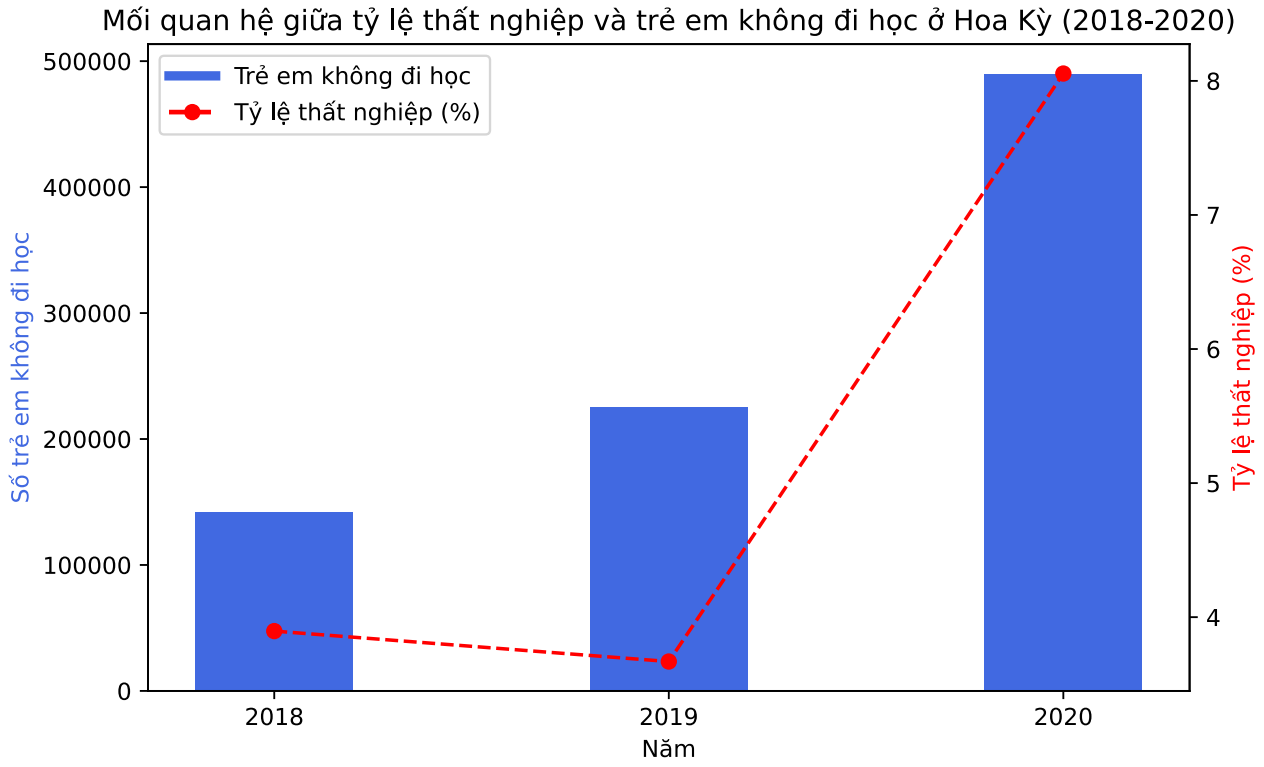
- Ở **Nhật Bản** (đường màu xanh), chi tiêu cho giáo dục của chính phủ dao động trong khoảng **3.1% - 3.6% GDP**, có xu hướng tổng thể ổn định nhưng hơi giảm nhẹ. Mức chi tiêu cao nhất là khoảng **3.6%** vào năm 2010, trong khi mức thấp nhất là khoảng **3.1%** vào năm 2018.

- Ở **Việt Nam** (đường màu đỏ), chi tiêu cho giáo dục dao động khá mạnh trong khoảng **3.0% - 4.9%** GDP, với xu hướng tổng thể giảm dần theo thời gian. Mức chi tiêu cao nhất đạt gần **4.9%** vào năm 2008, nhưng sau đó giảm dần và đạt mức thấp nhất khoảng **3.0%** vào năm 2019.

- Ở **Hoa Kỳ** (đường màu đen), chi tiêu cho giáo dục biến động trong khoảng **4.8% - 6.7% GDP**, có xu hướng tổng thể giảm dần. Mức cao nhất là khoảng **6.7%** vào năm 2009, sau đó giảm đáng kể xuống **4.8%** vào năm 2015, rồi phục hồi nhẹ vào năm 2020.

Câu 2: Tỷ lệ thất nghiệp tổng thể có mối liên hệ như thế nào với số lượng trẻ em không đi học ở cấp tiểu học tại Hoa Kỳ trong một khoảng thời gian 2018-2020? Liệu rằng có phải do sự khó khăn tài chính mà trẻ em không được đến trường?

- **Mục tiêu phân tích**: Xác thực xem liệu rằng có phải do sự khó khăn tài chính mà trẻ em không được đến trường ở Hoa Kỳ trong giai đoạn 2018-2020.
- **Lựa chọn trường dữ liệu**: Sử dụng 3 trường dữ liệu là:
  - `SE_PRM_UNER_FE` là số trẻ em nữ không đi học ở cấp tiểu học.
  - `SE_PRM_UNER_MA` là số trẻ em nam không đi học ở cấp tiểu học.
  - `SL_UEM_TOTL_ZS` là tỷ lệ thất nghiệp tổng thể trong lực lượng lao động (%).
- **Mối quan hệ giữa các trường**: `SL_UEM_TOTL_ZS` (tỷ lệ thất nghiệp tổng thể) được xem là biến độc lập, có thể ảnh hưởng đến tổng số trẻ em không đi học (`SE_PRM_UNER_FE` + `SE_PRM_UNER_MA`), là biến phụ thuộc. Mối quan hệ này kiểm tra xem sự thay đổi của tỷ lệ thất nghiệp có liên quan đến số trẻ em không đi học ở cấp tiểu học tại Hoa Kỳ từ 2018-2020, với giả thuyết rằng khó khăn tài chính do thất nghiệp có thể khiến trẻ em không được đến trường.
- **Loại biểu đồ sử dụng**: Biểu đồ kết hợp bao gồm biểu đồ cột (bar chart) và biểu đồ đường (line chart).
- **Lý do chọn biểu đồ**:
  - Biểu đồ cột thể hiện tổng số trẻ em không đi học giúp dễ dàng so sánh giá trị tuyệt đối qua các năm.
  - Biểu đồ đường thể hiện tỷ lệ thất nghiệp cho thấy xu hướng thay đổi liên tục, kết hợp với trục y kép để quan sát mối liên hệ giữa hai đại lượng một cách trực quan.
- **Các bước thực hiện:**
  1. Lấy dữ liệu từ bộ dữ liệu của Hoa Kỳ cho ba chỉ số trên trong khoảng thời gian 2018-2020.
  2. Tính tổng số trẻ em không đi học bằng cách cộng `SE_PRM_UNER_FE` và `SE_PRM_UNER_MA` theo từng năm.
  3. Tính hệ số tương quan (Pearson's r) để kiểm tra mức độ liên hệ giữa tỷ lệ thất nghiệp và tổng số trẻ em không đi học.
  4. Vẽ biểu đồ kết hợp bồm gồm biểu đồ cột thể hiện tổng số trẻ em không đi học theo từng năm và biểu đồ đường thể hiện tỷ lệ thất nghiệp để quan sát xu hướng biến đổi. Sử dụng trục y kép để dễ so sánh hai đại lượng.
  5. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ.
  6. Lưu và hiển thị biểu đồ.
  7. Viết nhận xét.



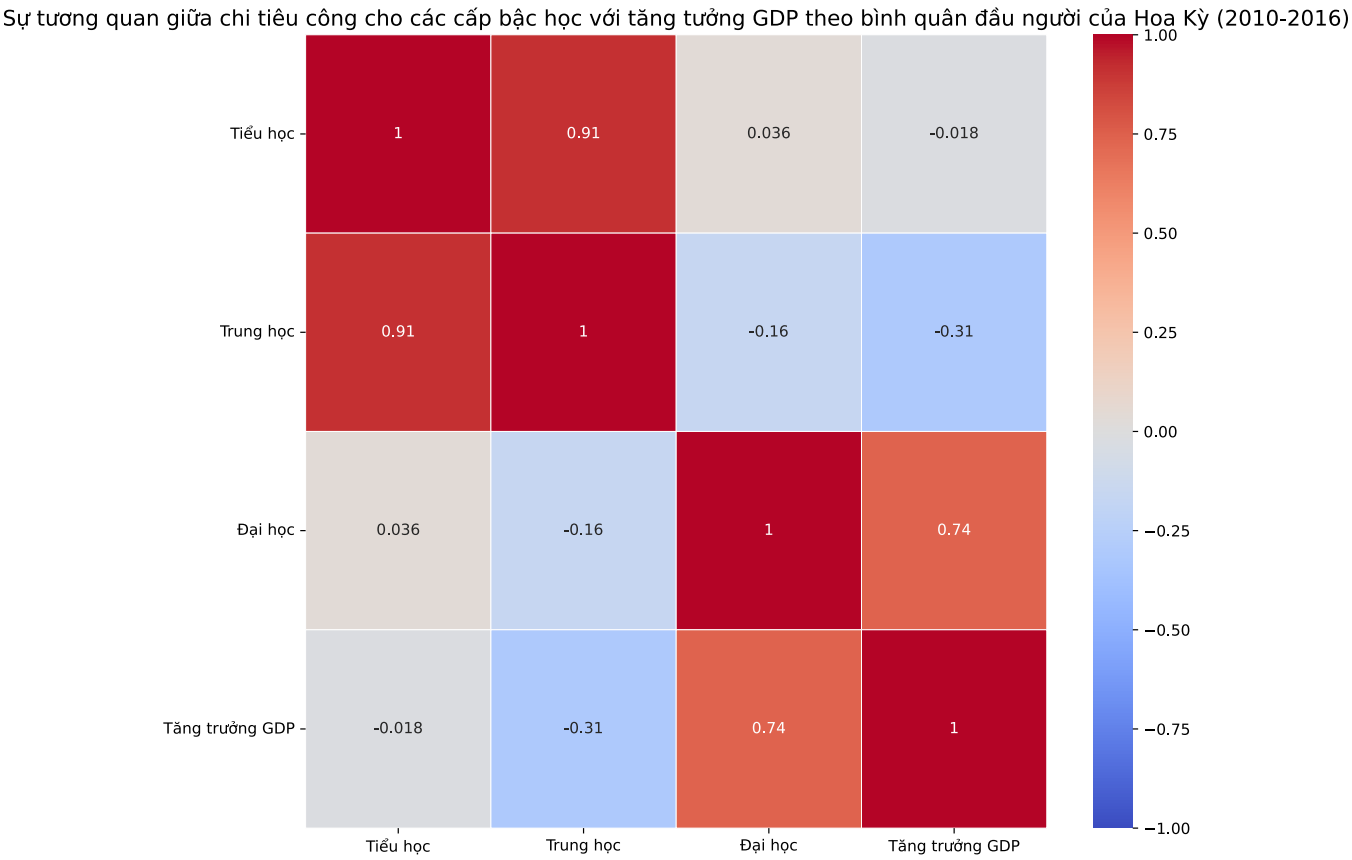
Hình 2: Phân bố thời lượng bài hát trên Spotify tại Việt Nam

Nhận xét

- Số trẻ em không đi học (cột màu xanh) ở Hoa Kỳ có xu hướng tăng dần qua các năm từ 2018 đến 2020 và tỷ lệ thất nghiệp (đường đỏ) ở Hoa Kỳ giảm nhẹ từ 2018 đến 2019, sau đó tăng vọt vào năm 2020.
- Vào năm 2018 - 2019, ta nhận thấy tỷ lệ thất nghiệp có xu hướng giảm nhẹ, nhưng số trẻ em không đi học tăng lên → Cho thấy không có mối quan hệ chặt chẽ giữa thất nghiệp và việc trẻ em bỏ học trong giai đoạn này.
- Nhưng vào năm 2020, tỷ lệ thất nghiệp tăng đột biến và số trẻ em không đi học cũng tăng vọt → Cho thấy một mối quan hệ rõ ràng giữa thất nghiệp và số trẻ em không đi học. Nguyên nhân có thể là do suy thoái kinh tế, gia đình gặp khó khăn tài chính, hoặc trường học đóng cửa do đại dịch.

Câu 3: Chi tiêu giáo dục ở các cấp học (tiểu học, trung học, đại học) có ảnh hưởng như thế nào đến tăng trưởng GDP ở Hoa Kỳ từ 2010-2016? Liệu đầu tư vào giáo dục có thực sự thúc đẩy tăng trưởng kinh tế?

- **Mục tiêu phân tích**: Xác định xem liệu đầu tư vào giáo dục có thực sự thúc đẩy tăng tưởng kinh tế ở Hoa Kỳ trong giai đoạn 2010-2016.
- **Lựa chọn trường dữ liệu**: Sử dụng 4 trường dữ liệu là
  - `SE_XPD_PRIM_PC_ZS` là chi tiêu của chính phủ cho mỗi học sinh tiểu học (% GDP bình quân đầu người).
  - `SE_XPD_SECO_PC_ZS` là chi tiêu của chính phủ cho mỗi học sinh trung học (% GDP bình quân đầu người).
  - `SE_XPD_TERT_PC_ZS` là chi tiêu của chính phủ cho mỗi sinh viên đại học (% GDP bình quân đầu người).
  - `NY_GDP_PCAP_KD_ZG` là tăng trưởng GDP bình quân đầu người hàng năm (%).
- **Mối quan hệ giữa các trường**: `SE_XPD_PRIM_PC_ZS`, `SE_XPD_SECO_PC_ZS` và `SE_XPD_TERT_PC_ZS` (chi tiêu công cho giáo dục ở các cấp tiểu học, trung học, và đại học) là các biến độc lập, có thể ảnh hưởng đến `NY_GDP_PCAP_KD_ZG` (tăng trưởng GDP bình quân đầu người), là biến phụ thuộc. Mối quan hệ này kiểm tra mức độ tác động của đầu tư giáo dục ở từng cấp học đến tăng trưởng kinh tế tại Hoa Kỳ từ 2010-2016, với giả thuyết rằng chi tiêu giáo dục cao hơn sẽ thúc đẩy tăng trưởng GDP.
- **Loại biểu đồ sử dụng**: Biểu đồ cột nhiệt (heatmap).
- **Lý do chọn biểu đồ**:
  - Biểu đồ nhiệt (heatmap) hiển thị ma trận tương quan, giúp dễ dàng so sánh mức độ liên hệ giữa chi tiêu giáo dục ở các cấp học và tăng trưởng GDP một cách trực quan.
  - Màu sắc trong heatmap thể hiện rõ ràng cường độ và hướng (âm/dương) của mối tương quan, hỗ trợ việc đánh giá nhanh tác động của từng biến.
- **Các bước thực hiện:**
  1. Lấy dữ liệu từ bộ dữ liệu của Hoa Kỳ cho bốn chỉ số trên trong khoảng thời gian 2010-2016.
  2. Tính ma trận tương quan Pearson để xác định mức độ liên hệ giữa chi tiêu giáo dục ở từng cấp học và tăng trưởng GDP.
  3. Vẽ biểu đồ nhiệt (heatmap) để trực quan hóa mối quan hệ giữa các biến sau đó đưa ra nhận xét về tác động của chi tiêu giáo dục đối với tăng trưởng kinh tế dựa trên kết quả phân tích.
  4. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ.
  5. Lưu và hiển thị biểu đồ.
  6. Viết nhận xét.



Hình 3: Sự tương quan giữa chi tiêu công cho các cấp bậc học với tăng tưởng GDP theo bình quân đầu người của Hoa Kỳ (2010-2016)

Nhận xét

- Giáo dục tiểu học **(-0.018)** và trung học **(-0.31)** có mối tương quan âm với tăng trưởng GDP bình quân đầu người, đặc biệt là trung học có giá trị âm đáng kể → Việc chi tiêu công cho các bậc học này không trực tiếp tác động tích cực đến tốc độ tăng trưởng kinh tế ngắn hạn.
- Giáo dục đại học **(0.74)** có mối tương quan dương khá mạnh với tăng trưởng GDP bình quân đầu người → Cho thấy đầu tư vào giáo dục bậc đại học có thể góp phần thúc đẩy tăng trưởng kinh tế, có thể do tác động của nghiên cứu, đổi mới công nghệ và nguồn nhân lực chất lượng cao.

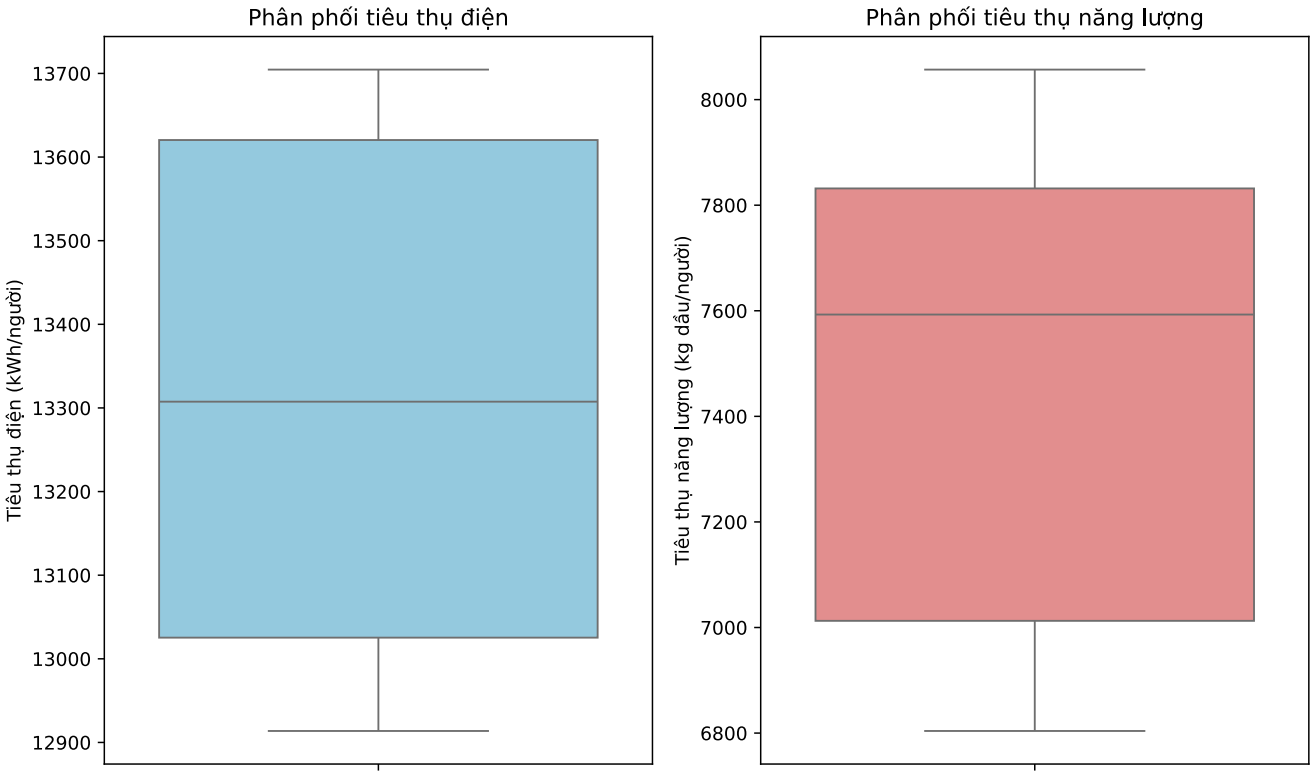
Kết luận

- Đầu tư vào giáo dục đại học có mối tương quan cao với tăng trưởng GDP bình quân đầu người, cho thấy rằng tăng cường đầu tư vào bậc đại học có thể có lợi cho nền kinh tế trong trung và dài hạn. Điều này phù hợp với quan điểm rằng nền kinh tế tri thức cần một lực lượng lao động có trình độ cao để thúc đẩy đổi mới và tăng trưởng.
- Dù giáo dục bậc thấp hơn không tác động ngay lập tức đến tăng trưởng kinh tế, nhưng nó đóng vai trò quan trọng trong việc xây dựng nền tảng tri thức và kỹ năng cho lực lượng lao động trong tương lai.

Các câu hỏi về chủ đề Biến đổi khí hậu

Câu 4: Biến động tiêu thụ năng lượng điện và năng lượng sơ cấp đầu người ở Hoa Kỳ từ 2000–2020 như thế nào? Có năm nào vượt trội hoặc bất thường không?

- **Mục tiêu phân tích**: Phân tích sự biến động của tiêu thụ năng lượng điện và năng lượng sơ cấp bình quân đầu người ở Hoa Kỳ trong giai đoạn 2000-2020, nhằm đánh giá xu hướng tiêu thụ năng lượng và xác định các năm có giá trị bất thường.
- **Lựa chọn trường dữ liệu**: Sử dụng 2 trường dữ liệu:
  - `EG_USE_ELEC_KH_PC` là sự tiêu thụ điện bình quân đầu người (kWh/người).
  - EG_USE_PCAP_KG_OE là sự tiêu thụ năng lượng bình quân đầu người (kg dầu tương đương/người).
- **Mối quan hệ giữa các trường**: `EG_USE_ELEC_KH_PC` và `EG_USE_PCAP_KG_OE` là hai biến độc lập, đại diện cho mức tiêu thụ năng lượng ở hai khía cạnh khác nhau (điện và tổng năng lượng sơ cấp). Mối quan hệ này giúp so sánh sự phân bố và biến động của hai loại tiêu thụ năng lượng qua thời gian, đồng thời xác định các điểm bất thường trong giai đoạn 2000-2020.
- **Loại biểu đồ sử dụng**: Biểu đồ hộp (box plot).
- **Lý do chọn biểu đồ**:
  - Biểu đồ hộp cho phép hiển thị rõ ràng sự phân bố, độ biến thiên, giá trị trung vị và các điểm ngoại lệ (outliers) của tiêu thụ năng lượng điện và năng lượng sơ cấp qua các năm, giúp dễ dàng nhận diện năm nào bất thường.
  - Việc sử dụng hai biểu đồ hộp song song hỗ trợ so sánh trực quan giữa hai chỉ số năng lượng, nhấn mạnh sự khác biệt trong xu hướng và mức độ ổn định của chúng.
- **Các bước thực hiện**:
  1. Trích xuất dữ liệu cho 2 chỉ số trên từ năm 2000–2020 từ bộ dữ liệu.
  2. Vẽ biểu đồ hộp (box plot) để trực quan hóa sự phân bố của từng chỉ số trong giai đoạn 2000–2020.
  3. Phân tích xem có năm nào là ngoại lệ (outlier) hoặc sự biến động có lớn không giữa các năm.
  4. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ.
  5. Lưu biểu đồ dưới định dạng SVG và hiển thị kết quả.
  6. Dưa ra nhận xét về mức độ ổn định trong tiêu thụ năng lượng và xu hướng thay đổi trong hai thập kỷ qua.



Hình 4: Phân phối tiêu thụ điện và phân phối tiêu thụ năng lượng

Nhận xét

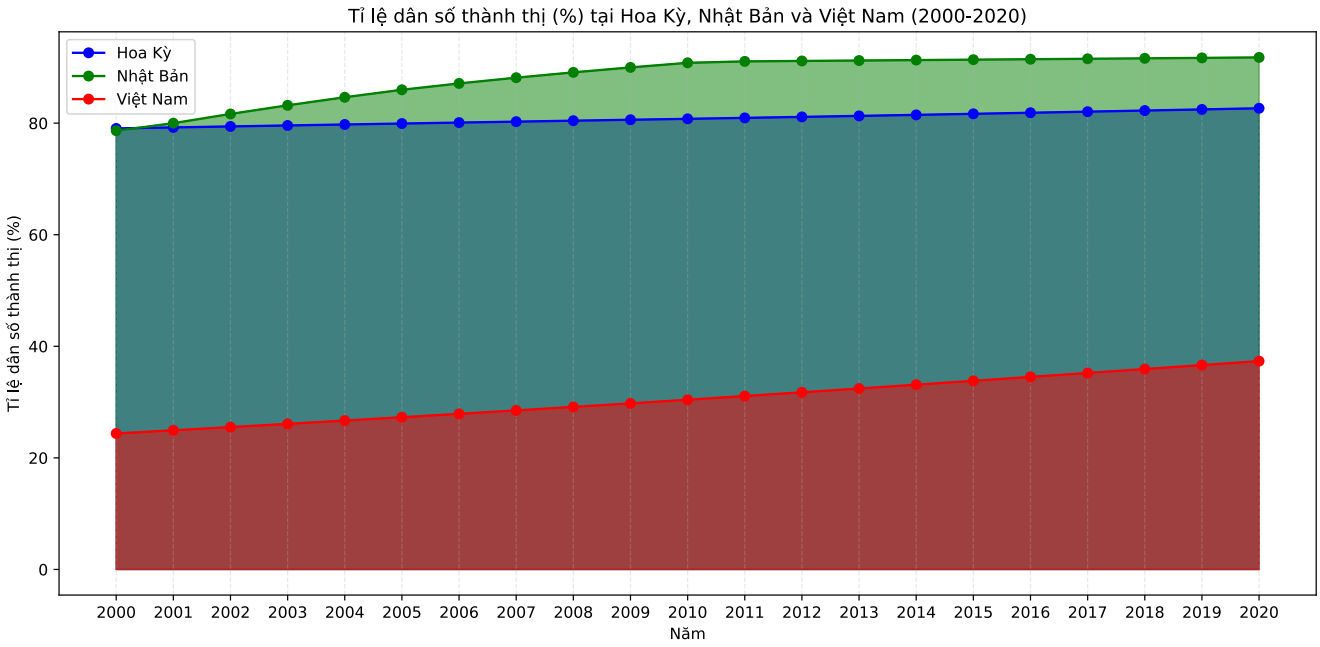
- Tiêu thụ điện (kWh/người) có phân phối khá hẹp, dao động chủ yếu trong khoảng 12900–13700 kWh/người, với trung vị khoảng 13300 kWh/người. Không có giá trị ngoại lai rõ ràng, cho thấy mức tiêu thụ điện tương đối ổn định trong giai đoạn này. Khoảng tứ phân vị (IQR) không quá rộng, chứng tỏ sự dao động giữa các năm là nhỏ.
- Tiêu thụ năng lượng sơ cấp (kg dầu/người) có phân phối rộng hơn một chút, với giá trị dao động trong khoảng 6800–8100 kg dầu/người. Trung vị rơi vào khoảng 7600 kg dầu/người. Có vẻ như phân phối hơi lệch về phía thấp khi phần dưới của hộp (Q1–Median) dài hơn phần trên. Điều này phản ánh mức tiêu thụ năng lượng có sự sụt giảm nhẹ trong một vài năm (có thể do khủng hoảng hoặc chuyển dịch sang năng lượng tái tạo).

Kết luận

- Mức tiêu thụ điện bình quân đầu người ở Hoa Kỳ khá ổn định trong hai thập kỷ qua, phản ánh sự ổn định trong lối sống, hạ tầng và nhu cầu dân cư.
- Trong khi đó, tiêu thụ năng lượng sơ cấp có xu hướng dao động mạnh hơn, điều này có thể liên quan đến sự thay đổi trong chính sách năng lượng, công nghệ tiết kiệm nhiên liệu, và xu hướng sử dụng năng lượng tái tạo.
- Không có giá trị ngoại lai rõ rệt trong cả hai biểu đồ, cho thấy dữ liệu tương đối 'sạch' và không có biến động đột ngột bất thường.

Câu 5: Tỷ lệ dân số sống ở khu vực đô thị của Hoa Kỳ, Nhật Bản và Việt Nam đã thay đổi như thế nào từ năm 2000 đến năm 2020?

- **Mục tiêu phân tích**: Phân tích xu hướng thay đổi tỷ lệ dân số sống ở khu vực đô thị của Hoa Kỳ, Nhật Bản và Việt Nam trong giai đoạn 2000-2020, nhằm so sánh mức độ đô thị hóa giữa ba quốc gia và nhận diện sự khác biệt trong quá trình phát triển đô thị.
- **Lựa chọn trường dữ liệu**: Sử dụng 1 trường dữ liệu là:
  - `SP_URB_TOTL_IN_ZS` là sự tiêu thụ điện bình quân đầu người (kWh/người)
- **Mối quan hệ giữa các trường**: `SP_URB_TOTL_IN_ZS` là biến phụ thuộc, được đo lường qua thời gian (năm) cho từng quốc gia. Mối quan hệ này thể hiện sự thay đổi tỷ lệ đô thị hóa theo thời gian ở mỗi quốc gia, cho phép so sánh xu hướng giữa Hoa Kỳ (đã đô thị hóa cao), Nhật Bản (đô thị hóa bão hòa) và Việt Nam (đang phát triển đô thị nhanh).
- **Loại biểu đồ sử dụng**: Biểu đồ miền (area plot) kết hợp đường (line plot).
- **Lý do chọn biểu đồ**:
  - Biểu đồ miền kết hợp đường giúp trực quan hóa xu hướng thay đổi tỷ lệ dân số thành thị theo thời gian một cách rõ ràng, với phần diện tích (area) nhấn mạnh sự tăng trưởng hoặc ổn định, còn đường (line) và điểm đánh dấu (marker) làm nổi bật giá trị cụ thể qua từng năm.
  - Sử dụng màu sắc khác nhau cho từng quốc gia giúp dễ dàng so sánh mức độ đô thị hóa giữa Hoa Kỳ, Nhật Bản và Việt Nam trong cùng một khung thời gian.
- **Các bước thực hiện**:
  1. Trích xuất dữ liệu cho chỉ số `SP_URB_TOTL_IN_ZS` của ba quốc gia Hoa Kỳ (US), Nhật Bản (JP), và Việt Nam (VN) từ năm 2000 đến 2020.
  2. Tạo DataFrame gồm các cột: 'Year', 'US', 'JP', 'VN' tương ứng với từng quốc gia.
  3. Sử dụng biểu đồ miền (area plot) kết hợp đường line để trực quan hóa xu hướng thay đổi theo thời gian.
  4. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ.
  5. Lưu biểu đồ dưới định dạng SVG và hiển thị kết quả.
  6. Nhận xét.



Hình 5: Tỉ lệ dân số thành thị (%) tại Hoa Kỳ, Nhật Bản và Việt Nam (2000-2020)

Nhận xét

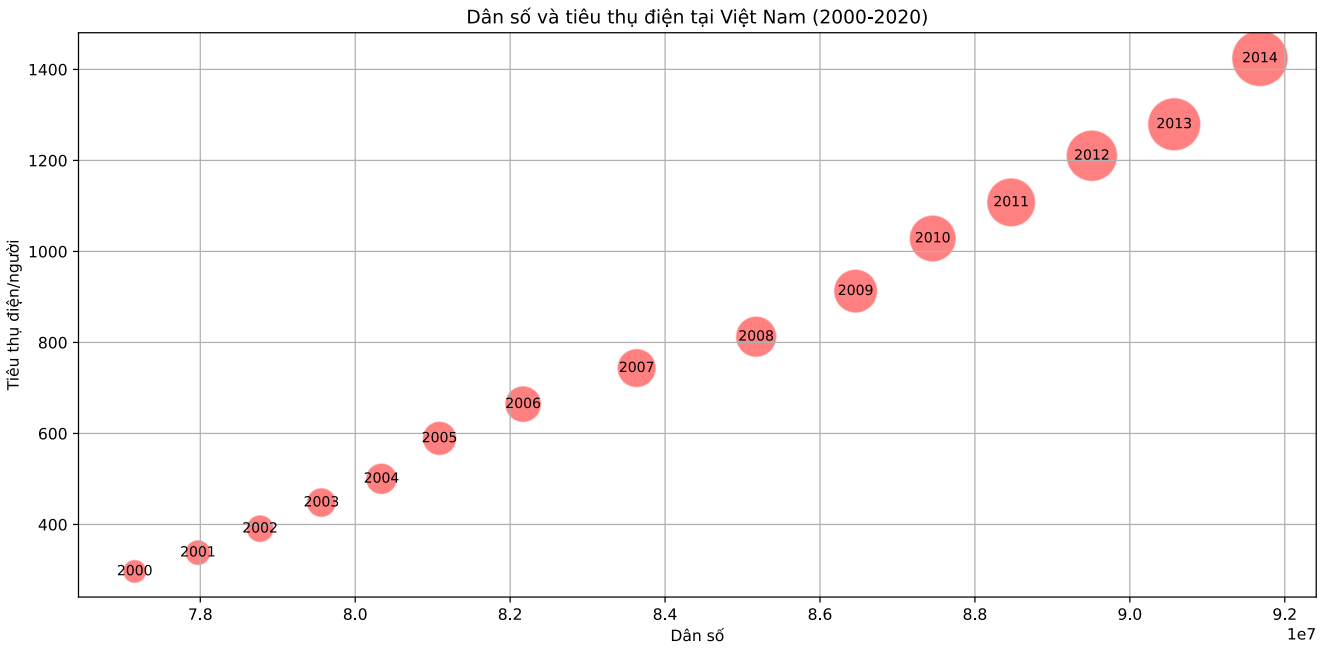
- Hoa Kỳ có tỷ lệ dân số thành thị khá ổn định, dao động trong khoảng 79% đến 83%. Đường biểu diễn cho thấy mức tăng nhẹ qua thời gian, phản ánh sự phát triển đều và ổn định về đô thị hóa.
- Nhật Bản là quốc gia có tỷ lệ đô thị hóa cao nhất trong ba nước, vượt mốc 90% từ sau năm 2010 và duy trì ở mức 91–92% đến năm 2020. Quá trình đô thị hóa của Nhật Bản diễn ra nhanh giai đoạn đầu và dần bão hòa.
- Việt Nam có tỷ lệ đô thị hóa thấp nhất nhưng lại tăng nhanh nhất, từ mức 25% năm 2000 lên gần 38% vào năm 2020. Biểu đồ cho thấy xu hướng tăng rõ rệt, đặc biệt sau năm 2010, phản ánh sự chuyển dịch mạnh mẽ từ nông thôn sang đô thị.

Kết luận

- Nhật Bản đã đạt mức đô thị hóa rất cao và gần như ổn định trong thập kỷ qua.
- Hoa Kỳ giữ mức độ đô thị hóa cao, với đà tăng chậm và đều.
- Việt Nam đang trong giai đoạn tăng tốc mạnh về đô thị hóa, thể hiện xu hướng phát triển kinh tế – xã hội nhanh chóng.

Câu 6: Mối quan hệ giữa quy mô dân số và tổng lượng điện năng tiêu thụ của Việt Nam từ năm 2000 đến 2020 thay đổi như thế nào? Những năm nào cho thấy sự tăng trưởng đột phá về nhu cầu điện tính trên đầu người?

- **Mục tiêu phân tích**: Phân tích mối quan hệ giữa quy mô dân số và tổng lượng điện năng tiêu thụ tại Việt Nam từ năm 2000 đến 2020, đồng thời xác định các năm có sự tăng trưởng đột phá về nhu cầu điện tính trên đầu người, nhằm đánh giá sự phát triển kinh tế và nhu cầu năng lượng.
- **Lựa chọn trường dữ liệu**: Sử dụng 2 trường dữ liệu là:
  - `EG_USE_ELEC_KH_PC` là sự tiêu thụ điện bình quân đầu người (kWh/người).
  - `SP_POP_TOTL` là tổng dân số.
- **Mối quan hệ giữa các trường**: `SP_POP_TOTL` (dân số) và `EG_USE_ELEC_KH_PC` (tiêu thụ điện bình quân đầu người) là hai biến chính, từ đó tính ra tổng lượng điện tiêu thụ (`TOTL_ELEC`) bằng cách nhân hai chỉ số này. Mối quan hệ này thể hiện cách dân số và mức tiêu thụ điện đầu người ảnh hưởng đến tổng nhu cầu điện, đồng thời cho phép đánh giá xu hướng tăng trưởng nhu cầu điện qua thời gian.
- **Loại biểu đồ sử dụng**: Biểu đồ bong bóng (bubble plot).
- **Lý do chọn biểu đồ**:
  - Biểu đồ bong bóng cho phép hiển thị ba chiều dữ liệu cùng lúc: dân số (trục x), tiêu thụ điện bình quân đầu người (trục y), và tổng lượng điện tiêu thụ (kích thước bong bóng), giúp trực quan hóa mối quan hệ giữa các biến một cách rõ ràng.
  - Nhãn năm trên từng bong bóng giúp nhận diện dễ dàng các mốc thời gian cụ thể, hỗ trợ việc xác định các năm có sự tăng trưởng đột phá về nhu cầu điện.
- **Các bước thực hiện**:
  1. Trích xuất dữ liệu của Việt Nam từ năm 2000 đến 2020 cho hai chỉ số: dân số (`SP_POP_TOTL`) và tiêu thụ điện bình quân đầu người (`EG_USE_ELEC_KH_PC`).
  2. Tính tổng lượng điện năng tiêu thụ bằng cách nhân hai chỉ số.
  3. Tạo biểu đồ Bubble Plot với
     - Trục X: Dân số
     - Trục Y: Tiêu thụ điện bình quân đầu người
     - Kích thước bong bóng: Tổng lượng điện tiêu thụ
     - Nhãn từng năm giúp nhận diện các mốc thời gian cụ th
  4. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ.
  5. Lưu biểu đồ dưới định dạng SVG và hiển thị kết quả.
  6. Nhận xét biểu đồ.



Hình 6: Dân số và tiêu thụ điện tại Việt Nam (2000–2020)

Nhận xét

- Giai đoạn 2000–2014: Dữ liệu cho thấy cả dân số và tiêu thụ điện/người đều tăng mạnh mẽ.
  - Dân số tăng từ khoảng 77 triệu (năm 2000) lên gần 92 triệu (năm 2014).
  - Tiêu thụ điện/người tăng từ khoảng 300 kWh/người lên hơn 1400 kWh/người.
  - Điều này cho thấy mức sống và nhu cầu sử dụng điện tăng rõ rệt, phản ánh quá trình công nghiệp hóa và hiện đại hóa đất nước.
- Tốc độ tăng tiêu thụ điện/người vượt xa tốc độ tăng dân số, cho thấy nhu cầu điện không chỉ đến từ dân cư mà còn từ hoạt động sản xuất, kinh doanh.
- Kích thước bong bóng (biểu thị tổng sản lượng điện tiêu thụ) cũng tăng theo thời gian, khẳng định mức tiêu thụ điện toàn quốc tăng mạnh.
- Xu hướng tăng đều và ổn định qua từng năm từ 2000 đến 2014, không có năm nào bị gián đoạn hoặc sụt giảm.

Lưu ý: Do thiếu dữ liệu về tiêu thụ điện đầu người từ năm 2015 trở đi, nên biểu đồ không hiển thị từ năm 2015 đến 2020.

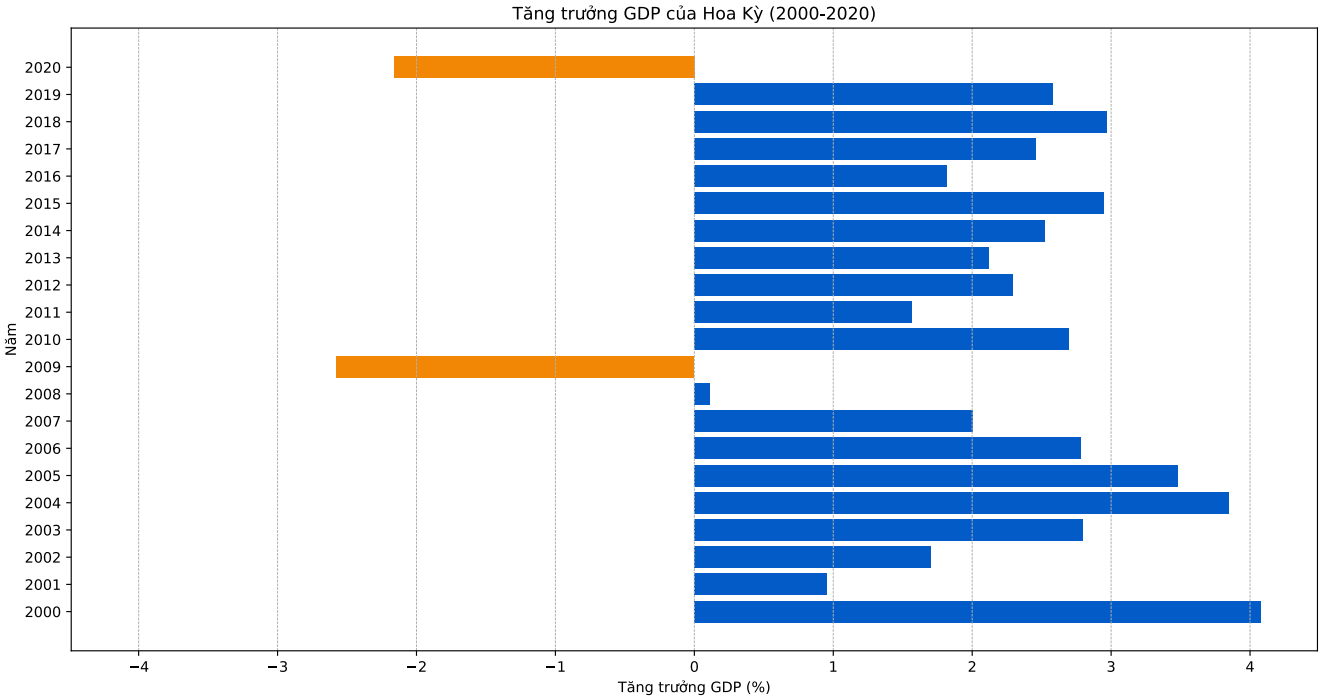
Kết luận

- Việt Nam đang trong quá trình tăng trưởng mạnh về nhu cầu sử dụng điện, phù hợp với đà phát triển kinh tế và đô thị hóa.
- Sự tăng nhanh về tiêu thụ điện/người là chỉ dấu tích cực về mức sống, công nghệ và công nghiệp hóa tại Việt Nam trong giai đoạn 2000–2014.

Các câu hỏi về chủ đề Kinh tế & Phát triển

Câu 7: Tình hình tăng trưởng GDP của Hoa Kỳ qua các năm (2000 - 2020)

- **Mục tiêu phân tích**: Phân tích xu hướng tăng trưởng GDP của Hoa Kỳ trong giai đoạn 2000-2020 để hiểu rõ hơn về biến động kinh tế qua các năm, bao gồm các giai đoạn tăng trưởng dương và suy thoái (tăng trưởng âm).
- **Lựa chọn trường dữ liệu**: Sử dụng 2 trường dữ liệu là:
  - `year` là năm từ 2000 đến 2020.
  - `NY_GDP_MKTP_KD_ZG` là tăng trưởng GDP hàng năm của Hoa Kỳ (% thay đổi so với năm trước).
- **Mối quan hệ giữa các trường**: `year` là biến độc lập (trục thời gian), trong khi `NY_GDP_MKTP_KD_ZG` là biến phụ thuộc, thể hiện mức tăng trưởng GDP tương ứng với từng năm. Mối quan hệ này giúp theo dõi sự thay đổi của tăng trưởng kinh tế theo thời gian và nhận diện các giai đoạn kinh tế quan trọng (ví dụ: khủng hoảng tài chính 2008-2009).
- **Loại biểu đồ sử dụng**: Biểu đồ cột ngang (horizontal bar chart).
- **Lý do chọn biểu đồ**:
  - Biểu đồ cột ngang cho phép hiển thị rõ ràng giá trị tăng trưởng GDP qua từng năm, với trục ngang thể hiện mức tăng trưởng (cả dương và âm) và trục dọc thể hiện các năm, giúp dễ dàng quan sát xu hướng theo thời gian.
  - Sử dụng màu sắc khác nhau (xanh cho tăng trưởng dương, cam cho tăng trưởng âm) để trực quan hóa nhanh các giai đoạn tăng trưởng và suy thoái kinh tế.
- **Các bước thực hiện**:
  1. Lấy dữ liệu tăng trưởng GDP hàng năm (`NY_GDP_MKTP_KD_ZG`) của Hoa Kỳ từ bộ dữ liệu cho giai đoạn 2000-2020.
  2. Tạo biểu đồ cột ngang bằng cách sử dụng `year` làm trục y và `NY_GDP_MKTP_KD_ZG` làm trục x.
  3. Tùy chỉnh biểu đồ: đặt tiêu đề, nhãn trục, giới hạn trục x dựa trên giá trị lớn nhất của tăng trưởng GDP, thêm lưới trục x, và sử dụng màu sắc có điều kiện để phân biệt tăng trưởng dương/âm.
  4. Điều chỉnh bố cục biểu đồ để đảm bảo các nhãn hiển thị đầy đủ và không bị cắt.
  5. Hiển thị biểu đồ và đưa ra nhận xét dựa trên xu hướng quan sát được.



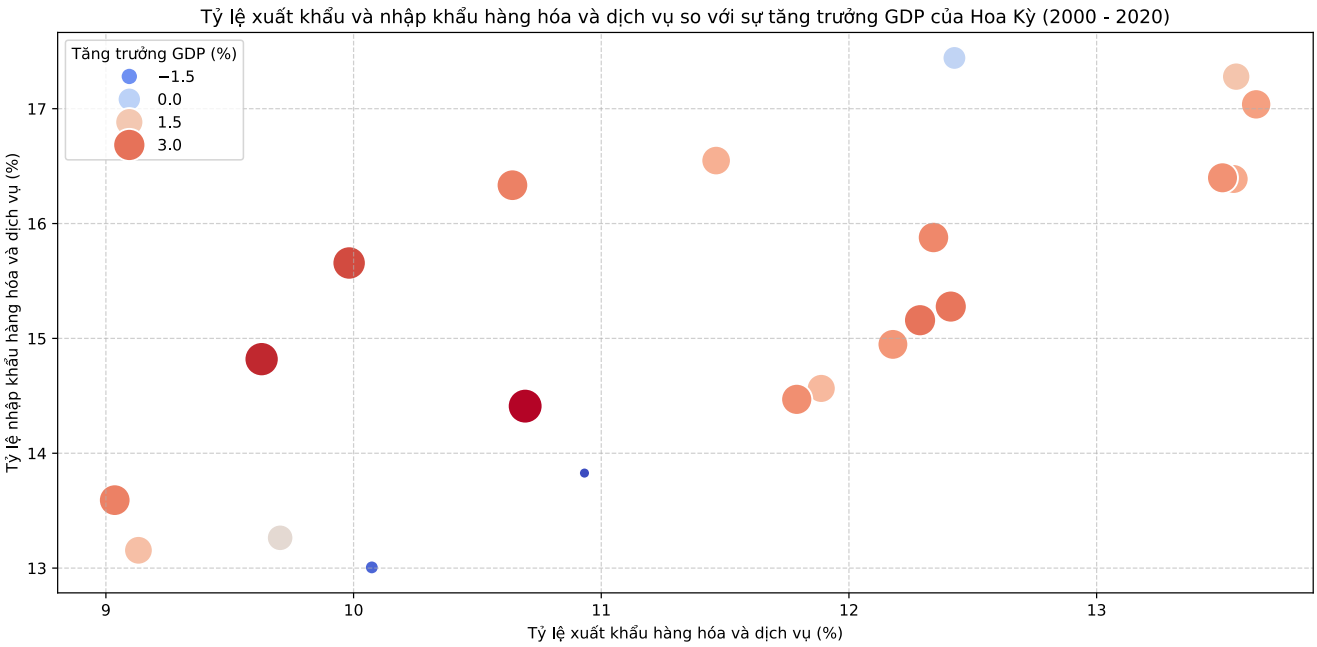
Hình 7: Tăng trưởng GDP của Hoa Kỳ (2000 - 2020)

Nhận xét

- Về tổng quan, nền kinh tế Hoa Kỳ phát triển mạnh mẽ khi hầu hết mức tăng trưởng GDP qua các năm đều dương cùng sự phục hồi mạnh mẽ khi mức GPD tăng vọt chỉ sau 1 năm suy thoái.
- Trong tất cả các năm từ 2000 đến 2020, chỉ có 2 năm mức tăng trưởng GDP của Mỹ đạt mức âm, đó là năm 2009 và năm 2020.
  - Năm 2009: Năm này hứng chịu mọi hậu quả của cuộc khủng hoảng tài chính toàn cầu, xuất phát từ việc Lehman Brothers phá sản - công tỳ dịch vụ tài chính toàn cầu.
  - Năm 2020: Đại dịch COVID-19 bùng phát dẫn đến phong tỏa diện rộng, tỷ lệ thất nghiệp tăng mạnh (14.7

Câu 8: Tỷ lệ xuất nhập khẩu của Hoa Kỳ ảnh hưởng đến sự tăng trưởng GDP như thế nào (2000 - 2020)?

- **Mục tiêu phân tích**: Xác định mối quan hệ giữa tỷ lệ xuất khẩu và nhập khẩu (tính theo phần trăm GDP) với tăng trưởng GDP của Hoa Kỳ trong giai đoạn 2000-2020, nhằm đánh giá mức độ ảnh hưởng của hoạt động thương mại quốc tế đến sự phát triển kinh tế.
- **Lựa chọn trường dữ liệu**: Sử dụng 3 trường dữ liệu là
  - `NE_EXP_GNFS_ZS`: Tỷ lệ xuất khẩu hàng hóa và dịch vụ (% GDP).
  - `NE_IMP_GNFS_ZS`: Tỷ lệ nhập khẩu hàng hóa và dịch vụ (% GDP).
  - `NY_GDP_MKTP_KD_ZG`: Tăng trưởng GDP hàng năm (% thay đổi so với năm trước).
- **Mối quan hệ giữa các trường**: `NE_EXP_GNFS_ZS` (tỷ lệ xuất khẩu) và `NE_IMP_GNFS_ZS` (tỷ lệ nhập khẩu) là các biến độc lập, có thể ảnh hưởng đến `NY_GDP_MKTP_KD_ZG` (tăng trưởng GDP), là biến phụ thuộc. Mối quan hệ này kiểm tra giả thuyết rằng thương mại quốc tế (xuất và nhập khẩu) có tác động đến tăng trưởng kinh tế, với tỷ lệ xuất nhập khẩu cao hơn hoặc thấp hơn có thể liên quan đến mức tăng trưởng GDP khác nhau.
- **Loại biểu đồ sử dụng**: Biểu đồ phân tán (scatter plot).
- **Lý do chọn biểu đồ**:
  - Biểu đồ phân tán cho phép thể hiện mối quan hệ giữa tỷ lệ xuất khẩu (trục x) và tỷ lệ nhập khẩu (trục y), đồng thời tích hợp tăng trưởng GDP qua kích thước và màu sắc của các điểm, giúp trực quan hóa sự tương quan giữa ba biến một cách đồng thời.
  - Màu sắc (theo bảng màu 'coolwarm') và kích thước điểm (từ nhỏ đến lớn) phản ánh rõ ràng mức tăng trưởng GDP, giúp dễ dàng nhận diện các năm có tăng trưởng cao/thấp và mối liên hệ với hoạt động xuất nhập khẩu.
- **Các bước thực hiện**:
  1. Lấy dữ liệu từ bộ dữ liệu của Hoa Kỳ cho ba chỉ số `NE_EXP_GNFS_ZS`, `NE_IMP_GNFS_ZS` và `NY_GDP_MKTP_KD_ZG` trong giai đoạn 2000-2020.
  2. Vẽ biểu đồ phân tán với `NE_EXP_GNFS_ZS` trên trục x, `NE_IMP_GNFS_ZS` trên trục y, và sử dụng `NY_GDP_MKTP_KD_ZG` để xác định kích thước và màu sắc của các điểm.
  3. Tùy chỉnh biểu đồ: thêm tiêu đề, nhãn trục, chú thích (legend) cho tăng trưởng GDP, và lưới nền để dễ quan sát.
  4. Điều chỉnh bố cục biểu đồ để đảm bảo các thành phần hiển thị rõ ràng và không bị cắt.
  5. Hiển thị biểu đồ và đưa ra nhận xét về mối quan hệ giữa tỷ lệ xuất nhập khẩu và tăng trưởng GDP dựa trên xu hướng quan sát được.



Hình 8: Tỷ lệ xuất khẩu và nhập khẩu hàng hóa và dịch vụ so với sự tăng trưởng GDP của Hoa Kỳ (2000 - 2020)

Nhận xét

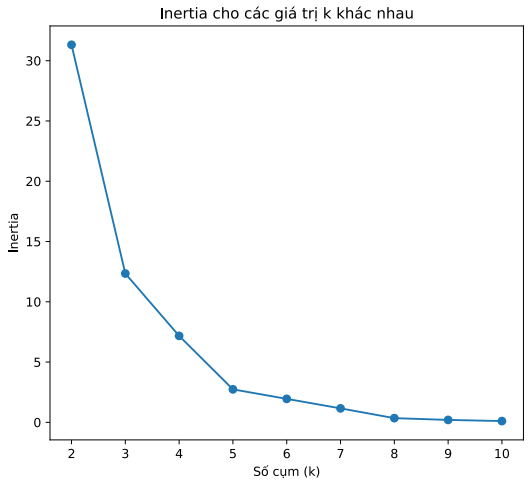
- Dựa vào biểu đồ trên, ta nhận thấy tỷ lệ xuất khẩu và tỷ lệ nhập khẩu có tương quan thuận, cùng tương quan thuận nhẹ với sự tăng trưởng GDP.
- Bên cạnh đó, Hoa Kỳ còn khá phụ thuộc vào nhập khẩu khi tỷ lệ nhập khẩu cao tác động thuận đến mức độ tăng trưởng GDP nhiều hơn tỷ lệ xuất khẩu.
- Đặc biệt, vào năm 2008, khi xảy ra khủng hoảng kinh tế, dù tỷ lệ xuất nhập khẩu khá cao, song vẫn chịu mức tăng trưởng GDP gần 0.

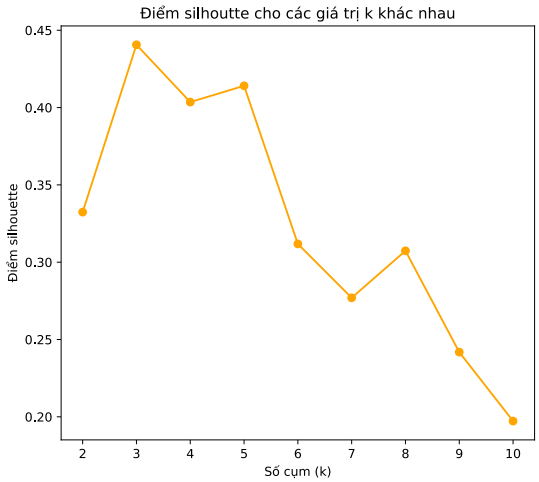
Câu 9: Làm thế nào để gom cụm các nền kinh tế trên thế giới và đưa gợi ý phát triển cho các cụm nền kinh tế (năm 2023)?

- **Mục tiêu phân tích**: Gom cụm các nền kinh tế trên thế giới dựa trên các chỉ số kinh tế quan trọng (GDP, tăng trưởng GDP, GDP bình quân đầu người, GNI) trong năm 2023, sử dụng thuật toán KMeans clustering, và đưa ra gợi ý phát triển phù hợp cho từng cụm nền kinh tế.
- **Lựa chọn trường dữ liệu**: Sử dụng 4 trường dữ liệu là:
  - `NY_GDP_MKTP_CD`: Tổng GDP (theo USD hiện hành).
  - `NY_GDP_MKTP_KD_ZG`: Tăng trưởng GDP hàng năm (% thay đổi so với năm trước).
  - `NY_GDP_PCAP_CD`: GDP bình quân đầu người (theo USD hiện hành).
  - `NY_GNP_ATLS_CD`: Tổng thu nhập quốc dân (GNI) theo phương pháp Atlas (theo USD hiện hành).

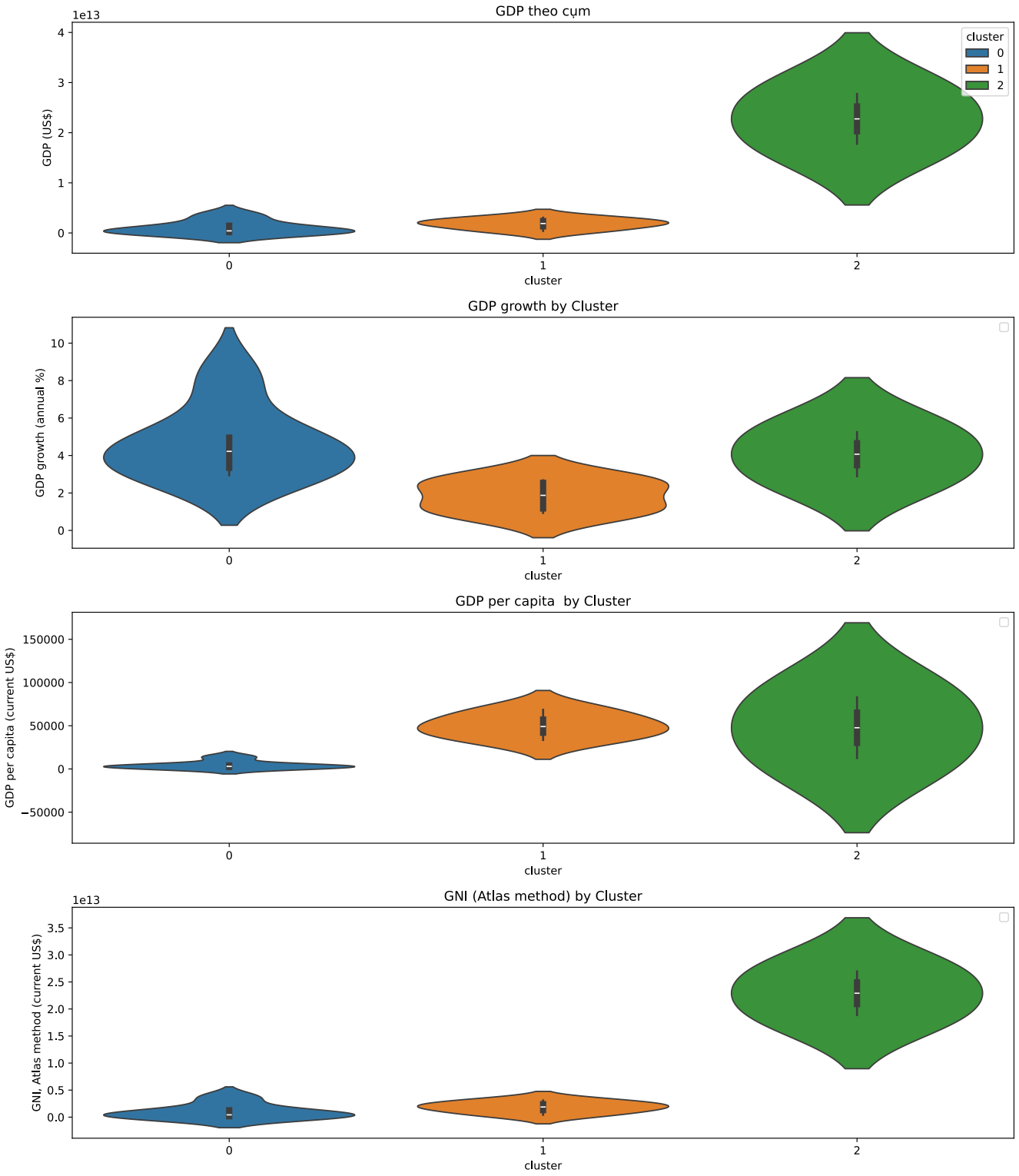
  Dữ liệu được lấy từ 12 quốc gia đại diện: Ai Cập (EGY), Liberia (LBR), Ghana (GHA), Trung Quốc (CHN), Ấn Độ (IND), Việt Nam (VNM), Đan Mạch (DNK), Pháp (FRA), Tây Ban Nha (ESP), Canada (CAN), Hoa Kỳ (USA), Mexico (MEX).
- **Mối quan hệ giữa các trường**: Các chỉ số `NY_GDP_MKTP_CD`, `NY_GDP_MKTP_KD_ZG`, `NY_GDP_PCAP_CD`, và `NY_GNP_ATLS_CD` được sử dụng để đo lường quy mô, tốc độ tăng trưởng, mức sống và thu nhập quốc dân của các nền kinh tế. Các chỉ số này có mối quan hệ tương hỗ, phản ánh mức độ phát triển kinh tế tổng thể, và được sử dụng làm cơ sở để gom cụm các quốc gia thành các cụm có đặc điểm kinh tế tương đồng.
- **Loại biểu đồ sử dụng**:
  - Biểu đồ đường (line plot) để đánh giá số lượng cụm tối ưu (dựa trên Inertia và Silhouette Score).
  - Biểu đồ violin (violin plot) để trực quan hóa phân phối các chỉ số kinh tế theo từng cụm.
- **Lý do chọn biểu đồ**:
  - Biểu đồ đường cho Inertia và Silhouette Score giúp xác định số lượng cụm tối ưu (k) một cách trực quan: Inertia giảm mạnh tại "elbow point" và Silhouette Score đạt giá trị cao nhất cho thấy sự phân cụm tốt.
  - Biểu đồ violin thể hiện rõ ràng sự phân phối và biến thiên của các chỉ số kinh tế (GDP, tăng trưởng GDP, GDP bình quân đầu người, GNI) trong từng cụm, giúp dễ dàng so sánh đặc điểm giữa các cụm nền kinh tế.
  - Sử dụng màu sắc khác nhau cho từng cụm trong biểu đồ violin giúp phân biệt rõ ràng các cụm và hỗ trợ việc nhận diện đặc điểm kinh tế của từng cụm.
- **Các bước thực hiện**:
  1. Trích xuất dữ liệu từ World Bank cho 12 quốc gia vào năm 2023, với 4 chỉ số: `NY_GDP_MKTP_CD`, `NY_GDP_MKTP_KD_ZG`, `NY_GDP_PCAP_CD`, và `NY_GNP_ATLS_CD`.

2. Chuẩn hóa dữ liệu bằng StandardScaler để đảm bảo các chỉ số có cùng thang đo, tránh ảnh hưởng của sự khác biệt về đơn vị.
3. Thử nghiệm thuật toán KMeans với số lượng cụm (k) từ 2 đến 10, tính toán Inertia và Silhouette Score cho từng giá trị k, sau đó vẽ biểu đồ đường để xác định số lượng cụm tối ưu.





4. Chọn k = 3 (dựa trên biểu đồ Inertia và Silhouette Score) và áp dụng KMeans để gom cụm các quốc gia thành 3 cụm.
5. Vẽ biểu đồ violin để trực quan hóa phân phối của từng chỉ số kinh tế theo cụm, tùy chỉnh nhãn, tiêu đề và màu sắc.
6. Phân tích đặc điểm của từng cụm và đưa ra gợi ý phát triển phù hợp.



Hình 9: GDP theo cụm

Nhận xét các nền kinh tế năm 2023

Sau khi gom cụm, ta chia được các nghệ sĩ vào 3 cụm

- Cụm 0 là cụm các nền kinh tế đang phát triển
  - **Đặc điểm**
    - GDP và GNI thấp nhất trong cả 3 cụm, thể hiện:
      - Sản xuất kém phát triển, quy mô nhỏ.
      - Cơ hội việc làm hạn chế, chất lượng lao động thấp.
      - Thiếu nguồn thu từ nước ngoài và đầu tư quốc tế.
    - Tăng trưởng GDP dao động mạnh, có nền kinh tế tăng trưởng nhanh nhưng cũng có nền kinh tế bị trì trệ.
    - GDP bình quân đầu người thấp phản ánh mức sống kém.
  - **Gợi ý phát triển**
    - Đầu tư phát triển cơ sở hạ tầng (giao thông, viễn thông,...).
    - Áp dụng các chính sách khuyến khích đầu tư nước ngoài.
- Cụm 1 là cụm các nền kinh tế mới nổi
  - **Đặc điểm**
    - GDP và GNI dù cao hơn cụm 0 nhưng vẫn khá thấp.
    - Tăng trưởng GDP ít dao động, thể hiện nền kinh tế ổn định nhưng thiếu sự đột phá.
    - GDP bình quân đầu người ở mức trung bình, vẫn còn tiềm năng để nâng cao mức sống.
  - **Gợi ý phát triển**
    - Đầu tư phát triển giáo dực và khả năng truy cập kỹ thuật số.
    - Đa dạng hóa nền kinh tế bằng cách chuyển từ nên kinh tế phát triển tập trung sang nền kinh tế dựa vào tri thức.
- Cụm 2 là cụm các nền kinh tế phát triển
  - **Đặc điểm**
    - GDP và GNI cao nhất trong 3 cụm, thể hiện:
      - Sản xuất mạnh mẽ với quy mô lớn.
      - Chất lượng sản xuất và chất lượng lao động tốt.
      - Vị thế trên thị trường quốc tế.
    - Tăng trưởng GDP ảm đạm hơn cụm 1, thể hiện nền kinh tế ổn định với nhiều hơn những đột phá.
    - GDP bình quân đầu người cao nhất, phản ánh mức sống cao cùng cơ sở hạ tầng hiện đại.
  - **Gợi ý phát triển**
    - Đầu tư phát triển công nghệ, AI.
    - Áp dụng các chính sách thúc đẩy tính bền vững, xu thế phát triển ngày nay.