TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẠI HỌC QUỐC GIA – THÀNH PHỐ HỒ CHÍ MINH

-----o0o----

LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

REPORT

FINAL PROJECT

Nhóm sinh viên thực hiện:

20120041 – Trần Kim Bảo

20120071 – Nguyễn Thị Bích Hà

20120369 – Nguyễn Thanh Tân

20120521 – Trần Thị Phương Linh

Thành phố Hồ Chí Minh, tháng 01 năm 2023

MỤC LỤC

I. BẢNG THÀNH VIÊN NHÓM

MSSV	Họ và tên
20120041	Trần Kim Bảo
20120071	Nguyễn Thị Bích Hà
20120369	Nguyễn Thanh Tân
20120521	Trần Thị Phương Linh

II. BẢNG PHÂN CÔNG CÔNG VIỆC

Về phần tìm tập dữ liệu và viết báo cáo, tất cả các thành viên đều làm chung.

MSSV	Nội dung	Mức độ hoàn thành
20120041	• Với mỗi cột có kiểu dữ liệu dạng phân loại (categorical), các giá trị được phân bố như thế nào? • Số-lượng/tỉ-lệ các giá trị thiếu? • Số lượng các giá trị khác nhau? Show một vài giá trị → Có gì bất thường không? • Đặt và trả lời câu hỏi số 4.	100%
20120071	• Với mỗi cột có kiểu dữ liệu dạng số (numerical), các giá trị được phân bố như thế nào? • Số-lượng/tỉ-lệ các giá trị thiếu? • Min? max? → Có gì bất thường không? • Đặt và trả lời câu hỏi số 1.	100%
20120369	• Mỗi cột có ý nghĩa gì? • Mỗi cột hiện đang có kiểu dữ liệu gì? • Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp không? • Đặt và trả lời câu hỏi số 2.	100%

20120521	• Dữ liệu có bao nhiêu dòng và bao nhiêu cột? • Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không? • Dữ liệu có các dòng bị lặp không? • Đặt và trả lời câu hỏi số 3.	100%

III. BÁO CÁO CHI TIẾT

1. Thu thập dữ liệu

- Nhóm sử dụng dữ liệu ***population_en_v2.csv*** tại: https://data.opendevelopmentcambodia.net/en/dataset/vietnam-population-status-2011-2016
- Tập dữ liệu mô tả tình trạng dân số của các tỉnh Việt Nam từ năm 2011 đến năm 2020.
- LICENSE của dữ liệu là https://creativecommons.org/licenses/by-sa/4.0/
- Người dùng được phép sử dụng với mục đích phi thương mại và phải credit cho chủ sở hữu.
- Người ta thu thập dữ liệu tại website Tổng cục Thống kê (https://www.gso.gov.vn).

2. Khám phá dữ liệu

2.1. Số dòng và số cột

Dữ liệu gồm có tổng cộng: 630 dòng và 8 cột.

2.2. Ý nghĩa của mỗi dòng

- Mỗi dòng chứa thông tin về dân số của các tỉnh Việt Nam qua các năm từ năm 2011 đến năm 2020 và không xuất hiện vấn đề các dòng có ý nghĩa khác nhau.
- Không có dòng nào trong dữ liệu bị trùng lặp.

2.3. Ý nghĩa của mỗi cột

Bảng mô tả ý nghĩa các cột

STT	Tên cột dữ liệu	Mô tả
1	Provinces/city	Tên tỉnh thành
2	Population density	Mật độ dân số
3	Average population	Dân số trung bình
4	Sex ratio	Tỷ lệ giới tính
5	Population grow ratio	Tỷ lệ gia tăng dân số
6	15+ labor	Lực lượng lao động từ 15 tuổi trở lên
7	Region	Vùng
8	Year	Năm

- Các cột hầu như có kiểu dữ liệu phù hợp, chẳng hạn như Provinces/city và Region là tên nên có kiểu dữ liệu object.
- Cột Year để dễ xử lý cho các phần sau nên không đổi thành kiểu DateTime.
- Các cột còn lại là số nên có kiểu số thực.

2.4. Các cột có kiểu dữ liệu dạng số (numerical) và sự phân bố giá trị

- Ở dữ liệu này, hiện có 6 cột thuộc nhóm numeric: Population density, Average population, Sex ratio, Population grow ratio, 15+ labor, Year.
- Với mỗi cột numeric ta sẽ tính tỉ lệ % giá trị thiếu (từ 0 đến 100), min, max. Và lưu kết quả vào Dataframe nume_col_profiles_df.
- Dataframe này gồm có:
  - 3 dòng là missing_ratio, min, max lần lượt là tỉ lệ các giá trị thiếu mỗi cột, min và max.
  - 5 cột là Population density, Average population, Sex ratio, Population grow ratio, 15+ labor, Year.

	Population density	Average population	Sex ratio	Population grow ratio	15+ labor	Year
missing_ratio	0.0	0.0	0.0	0.00	0.0	0.0
min	42.9	300.4	91.1	-1.22	199.6	2011.0
max	4476.0	9227.6	114.1	5.30	4826.0	2020.0

2.5. Các cột có kiểu dữ liệu dạng phân loại (categorical) và sự phân bố giá trị

- Ở dữ liệu này, hiện có 2 cột categorical là Provinces/city, Region.
- Với mỗi cột categorical, ta tính tỉ lệ % giá trị thiếu (từ 0 đến 100), số lượng giá trị khác nhau (không xét giá trị thiếu), list/array các giá trị khác nhau (không xét giá trị thiếu).
- Kết quả được lưu vào dataframe cate_col_profiles_df. Dataframe này có 3 dòng là missing_ratio, num_diff_vals, ratio_diff_vals; và có 2 cột là Provinces/city, Region.

	Provinces/city	Region
missing_ratio	0.0	0.0
num_diff_vals	63	6
ratio_diff_vals	{'Binh Dinh': 0.02, 'Ha Nam': 0.02, 'Cao Bang': 0.02, 'Vinh Phuc': 0.02, 'Tuyen Quang': 0.02, 'Dak Nong': 0.02, 'Lai Chau': 0.02, 'Ha Noi': 0.02, 'Hoa Binh': 0.02, 'Thai Binh': 0.02, 'Quang Tri': ...	{'North Central and Central Coast': 0.22, 'Midlands and northern mountains': 0.22, 'Mekong Delta': 0.21, 'Hong river Delta': 0.17, 'South East': 0.1, 'Highlands': 0.08}

3. Đặt và trả lời câu hỏi

3.1. Tìm hiểu về sự phân bố lực lượng nhân công trên 15 tuổi (15+ labor) của các tỉnh thành và từng giai đoạn. Có nhận xét gì về xu hướng trong các giai đoạn trên và xu hướng của các tỉnh thành có lực lượng lao động cao vượt trội hơn so với những tỉnh thành còn lại.

Ý nghĩa khi trả lời câu hỏi:

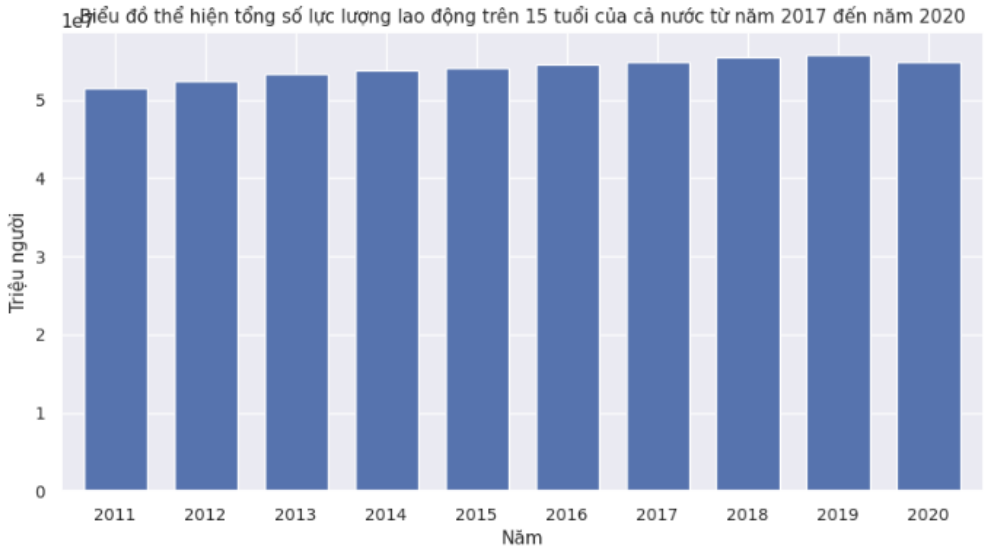
- Ta sẽ có được cái nhìn tổng quan về sự phân bố nhân lực trên 15 tuổi của Việt Nam theo từng vùng miền.
- Ngoài ra, ta sẽ tìm hiểu thêm những thành phố/ tỉnh thành có lực lượng nhân công dưới 15 đông nhất và lí do tại sao những nơi đó lại có sự phân bố như vậy.

Các bước cần làm để để trả lời câu hỏi:

- Đầu tiên ta sẽ tiến hành thống kê lực lượng lao động trên 15 tuổi theo từng năm. Sau đó ta lần lượt dùng Bar chart và Choropleth Map để trực quan hóa lực lượng lao động mỗi năm và đưa ra nhận xét tổng quan.
- Tiếp theo, ta chọn ra các tỉnh/ thành phố có lực lượng lao động trên 15 tuổi nhiều nhất ở mỗi miền để tiến hành phân tích chi tiết hơn. Đồng thời dùng Line chart để quan sát và so sánh xu hướng phân bổ lực lượng lao động trên 15 tuổi giữa mỗi tỉnh/ thành đang quan sát.
- Cuối cùng, dựa vào phân tích và trực quan, ta sẽ đưa ra kết luận.

Chi tiết thực hiện:

Bước 1: Ta sẽ tiến hành lấy các cột cần thiết từ dataframe, bao gồm 15+ labor, Provinces/city, Year để xử lý. Đồng thời, ta tiền xử lý cột 15+ labor bằng cách nhân 1000 cho mỗi giá trị. Vì đơn vị gốc của cột này chính là triệu người.
Bước 2: Ta sẽ tổng hợp lại tổng lực lượng lao động trên 15 tuổi mỗi năm và dùng Bar chart để trực quan hóa. Từ đó có được cái nhìn tổng quan về xu hướng trong các giai đoạn này.



Nhận xét:

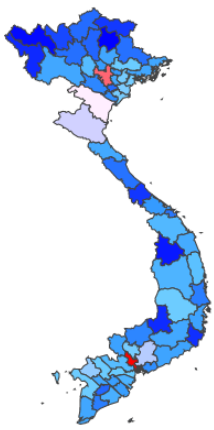
- Nhìn chung, lực lượng lao động tăng đều theo mỗi năm từ năm 2011 đến năm 2019. Cao nhất là năm 2019.
- Tuy nhiên, lại có sự sụt giảm đáng báo động vào năm 2020. Có thể lí giải đây là năm dịch Covid-19 hoành hành, gây nhiều khó khăn cho đời sống kinh tế, cơ hội việc làm nên mới gây ra hiện tượng giảm sút này.

Bước 3: Tiếp theo, để đi vào chi tiết hơn của xu hướng, ta sẽ tiến hành tiến hành tách dữ liệu theo từng năm và lưu trữ vào từng Dataframe riêng. Tổng cộng ta có tất cả 10 Dataframe riêng của 10 năm từ năm 2011 đến năm 2020.

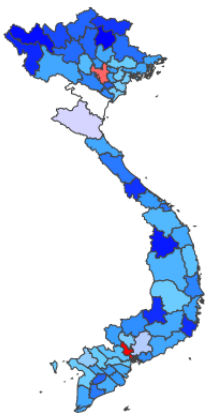
Bước 4: Tiếp theo, ta sẽ thực hiện visualize bằng Choropleth Map để dễ dàng quan sát và nhận xét. Ở đây ta sẽ dùng thêm file vietnam.geojson để phục vụ cho việc vẽ Choropleth Map. Đây là file dùng để trao đổi dữ liệu không gian địa lí.

- File vietnam.geojson được lấy từ Github với đường dẫn: https://github.com/Vizzuality/growasia_calculator
- Để thực hiện hóa việc này, ta cần phải tiền xử lý dữ liệu được đọc từ file vietnam.geojson, cụ thể lên tên các tỉnh thành từ tên tiếng việt sang tên tiếng anh (không dấu) và cách gọi tên tỉnh thành:
    - Ví dụ:
        - **Hồ Chí Minh city** sang **Ho Chi Minh**
        - **Thừa Thiên - Huế** sang **Thua Thien Hue**

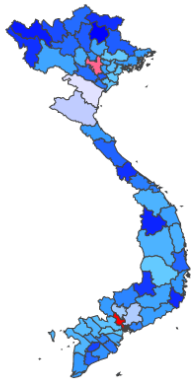
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2011



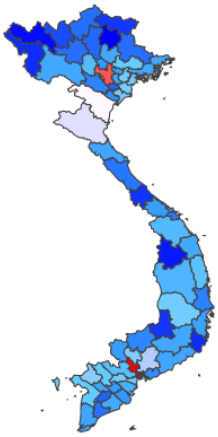
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2012



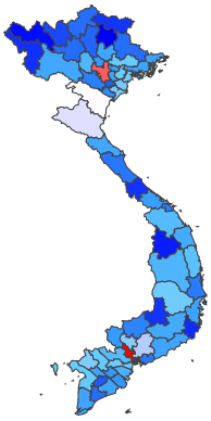
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2020



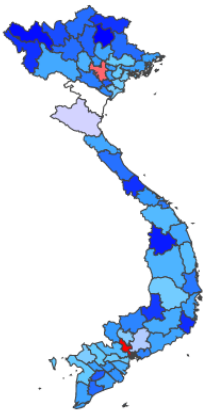
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2013



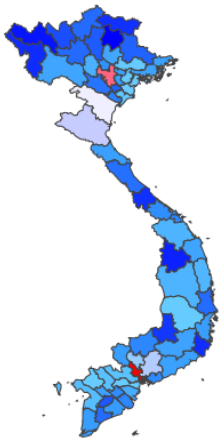
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2014



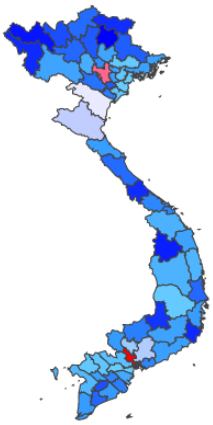
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2015



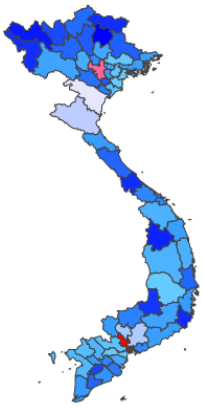
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2017



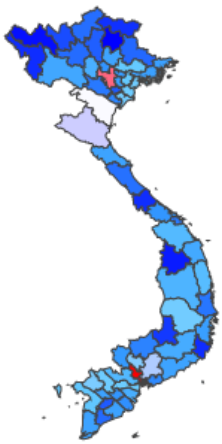
Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2018



Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2019



Bản đồ thể hiện phân bố lao động trên 15 tuổi năm 2016



Nhận xét:

Tổng quan:

- Đa phần ở 10 bản đồ, ta thấy màu xanh đậm và xanh nhạt chiếm gần hết cả nước từ Bắc vào Nam, tức là số lượng dân lao động trên 15 tuổi sẽ dao động từ khoảng 500.000 đến 1.500.000 người ở đại đa số các tỉnh thành/ thành phố.
- Ở miền Nam và nửa cuối miền Trung, nếu để ý, ta sẽ thấy đa số đều được phủ màu xanh nhạt trái ngược với phần đỉnh của bản đồ, có màu xanh đậm (tức là ở miền Bắc). Ta nhận thấy lực lượng lao động trên 15 tuổi tập trung chủ yếu ở 3 vùng là Đồng Bằng Sông Hồng, Bắc Trung Bộ và Duyên Hải Miền Trung và Đồng Bằng Sông Cửu Long.

Chi tiết:

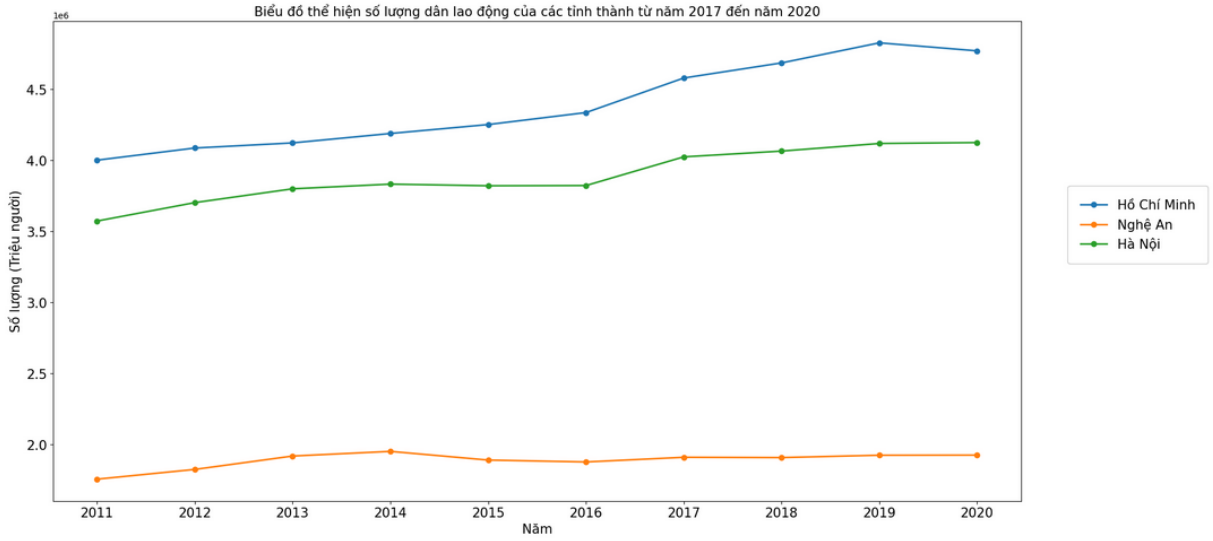
- Ở miền Bắc, ta thấy Hà nội có màu cam đỏ nếu quy đổi ra số liệu là trên 3.000.000 người (ta sẽ đặc biệt chú ý đến thành phố này) và Thanh Hóa (màu tím nhạt) với số liệu quy đổi (trên 2.000.000 người).
- Ở miền Trung, ta thấy nổi bật nhất là tỉnh Nghệ An (màu tím đậm) với lượng dân số lao động trên 15 trên 1.500.000 người.
- Ở miền Nam, ta sẽ thấy TP Hồ Chí Minh (màu đỏ) với lượng dân lao động trên 4 triệu người. Kế đến là Đồng Nai (màu tím nhạt) với số lượng trên 1.500.000 người. Nhưng nếu để ý, từ năm 2017 trở đi, tỉnh Bình Dương có xu hướng màu ngã sang tím nhạt, tức số lượng lao động ở Bình Dương từ năm 2017 có xu hướng tăng và xấp xỉ với Đồng Nai.

Ta sẽ tìm hiểu kĩ hơn về xu hướng và lí do của các tỉnh/ thành phố ở mỗi miền có số lượng dân lao động trên 15 tuổi cao nhất từ năm 2011 đến năm 2020. Ở đây dựa vào nhận xét phía trên, ta chọn ra lần lượt 3 tỉnh/ thành phố ứng với 3 miền

- Miền Bắc: TP Hà Nội
- Miền Trung: Tỉnh Nghệ An
- Miền Nam: TP Hồ Chí Minh

Bước 5: Ta tiến hành thống kê số lượng dân lao động trêm 15 tuổi từ năm 2017 đến năm 2020 của các tỉnh/ thành phố nêu trên, lần lượt là TP Hà Nội, TP Hồ Chí Minh, tỉnh Nghệ An.

Bước 6: Ở bước này ta sẽ dùng Line chart để quan sát về xu hướng của từng tỉnh thành từ năm 2011 đến năm 2020.



Nhận xét:

- Số lượng lực lượng lao động trên 15 tuổi có xu hướng tăng dần ở TP Hồ Chí Minh. Từ năm 2016, số lượng tăng vọt lên và đỉnh điểm là năm 2019, nhưng đến năm 2020, số lượng lại giảm xuống. Lí giải cho điều này:
    - Thành phố Hồ Chí Minh là đô thị đặc biệt; trung tâm lớn về kinh tế, văn hóa, giáo dục - đào tạo, khoa học - công nghệ; đầu mối giao lưu và hội nhập quốc tế; đầu tàu, động lực, có sức hút và sức lan tỏa lớn của Vùng. Kinh tế thành phố tăng trưởng khá và ổn định qua các năm, GRDP tăng bình quân đạt 8,3%/năm, quy mô GRDP của Thành phố năm 2020 ước chiếm 22,8% GDP cả nước và khoảng 48,4% GRDP của Vùng. Do đó nhận thức được điều này, rất nhiều người đã đổ về thành phố Hồ Chí Minh để đi tìm việc làm, cơ hội phát triển, nâng cao đời sống.

    - Theo tìm hiểu, nhu cầu tuyển dụng 06 tháng đầu năm 2016 tại thành phố tăng 2,53% so với cùng kỳ năm 2015. Có sự gia tăng về nhu cầu tuyển dụng lao động đã qua đào tạo, kinh nghiệm, trình độ và tính chuyên nghiệp luôn là sự quan tâm của doanh nghiệp khi tuyển dụng nhân sự. Nhu cầu tuyển dụng của doanh

nghiệp trong 06 tháng cuối năm 2016 tăng 13,15% so với 06 tháng đầu năm 2016. Điều này tác động tích cực đến thị trường lao động và một phần lí giải được vì sao lực lượng lao động lại tăng vọt từ năm 2016.

- Tính đến tháng 12 năm 2020, cả nước có 32,1 triệu người từ 15 tuổi trở lên bị ảnh hưởng tiêu cực bởi dịch Covid-19 bao gồm người bị mất việc làm, phải nghỉ giãn việc/nghỉ luân phiên, giảm giờ làm, giảm thu nhập,… Điều đó làm cho lực lượng lao động ở thành phố Hồ Chí Minh sụt giảm đáng kể.

- Tương tư như ở TP Hồ Chí Minh, ở Hà Nội, từ năm 2016, số lượng dân lao động trên 15 tuổi cũng tăng lên, nhưng tăng đều (trên 3.500.000 triệu người). Nhưng ở Hà Nội khác Hồ Chí Minh ở chỗ tổng lực lượng lao động trên 15 không giảm dần từ năm 2019 mà vẫn giữ vững là do từ năm 2019 đến năm 2020, mặc dù vẫn tồn tại các ca nhiễm covid nhưng Hà Nội không phải là tâm dịch như Hồ Chí Minh nên mọi hoạt động, việc làm vẫn ổn định, không xảy ra tình trạng trì trệ.
- Riêng ở tỉnh Nghệ An, số lượng dân lao động không có xu hướng tăng từ năm 2016 trở lên như ở 2 thành phố trên. Số lượng cao nhất rơi vào năm 2014 và có xu hướng giảm nhẹ đến năm 2016.
- TP Hồ Chí Minh có nguồn lực lượng lao động trên 15 tuổi hùng hậu và phát triển nhất, kế đến là thủ đô Hà Nội và sau cùng là tỉnh Nghệ An.

3.2. Dự đoán dân số theo các tỉnh thành trong vòng 5 năm kế tiếp (2021 - 2025)

Ý nghĩa khi trả lời câu hỏi:

- Đưa ra góc nhìn ngắn hạn để điều chỉnh chế độ kế hoạch hóa gia đình
- Dự trù cho nền kinh tế, dự đoán được các biến động xã hội có thể xảy ra khi dân số tăng vượt ngưỡng nào đó, từ đó đưa ra kế hoạch điều chỉnh kinh tế, xã hội (như quy hoạch hóa để xây dựng chung cư ở các thành phố lớn)

Các bước cần làm để để trả lời câu hỏi:

- Đối với mỗi tỉnh thành:
    1. Sử dụng mô hình để dự đoán
    2. Trích xuất đặc trưng
    3. Đưa ra thông tin dự đoán
    - Như vậy, mỗi tỉnh thành sẽ có một mô hình để dự đoán riêng.
    - Ở đây ta đưa ra 2 mô hình là Hồi quy tuyến tính và Cây quyết định.

Chi tiết thực hiện

❖ **HỒI QUY TUYẾN TÍNH**

a. Trích xuất thông tin cho từng tỉnh thành (ví dụ đối với thành phố Hồ Chí Minh)

- Chọn ra "Ho Chi Minh" trong cột "Provinces/city"
- Chọn đặc trưng:
    - Do mô hình hồi quy tuyến tính khá đơn giản nên chỉ chọn "Year" làm đặc trưng.
- Dữ liệu của mỗi tỉnh chỉ có 10 dòng nên chọn 8 dòng đầu làm tập train và 2 dòng sau làm tập test.

b. Huấn luyện và đưa ra đánh giá

- Sau khi huấn luyện và chạy trên tập test của thành phố Hồ Chí Minh, kết quả độ đo MSE cho ra vào cỡ 84467
- Đối với huấn luyện toàn bộ tập dữ liệu, do mỗi tỉnh thành sẽ có một mô hình riêng nên kết quả tổng cộng sẽ lấy trung bình độ đo MSE của từng tỉnh. Kết quả cho ra cỡ 12315.

c. Kết quả

Dự đoán của mô hình cho rằng hầu hết các tỉnh thành đều có dân số tăng nhẹ trong 5 năm tới. Chỉ có một vài tỉnh thành có dân số tăng nhanh.

d. Nhận xét

Điểm mạnh:

- Mô hình đơn giản và dễ sử dụng

Hạn chế:

- Do quá đơn giản nên có thể thiếu các đặc trưng khác
  - Tỉ lệ gia tăng dân số của 1 tỉnh nào đó cao đột biến hoặc thấp đột biến
  - Mật độ dân số ảnh hưởng đến sự di cư
  - Các chính sách được thực hiện trong năm để kìm hãm sự tăng dân số

Những nguyên nhân trên dẫn đến việc phải có một mô hình nào đó đơn giản mà có thêm các đặc trưng khác để đánh giá tốt hơn. Ta chọn Cây quyết định.

❖ CÂY QUYẾT ĐỊNH

a. Chọn đặc trưng

Các đặc trưng được sử dụng cho cây quyết định là:

- Population density
- Population grow ratio
- Sex ratio
- Region
- Year

Phân bổ tập train-test cũng là 8:2 như mô hình Hồi quy tuyến tính.

b. Tiền xử lý

Do có một thuộc tính Categorical là Region nên phải xử lý chuyển qua Numerical.

Ý tưởng của việc tiền xử lý là:

1. Xác định xem Region có bao nhiêu lớp
2. Gán số bắt đầu từ 0 cho từng lớp
3. Áp dụng cho dữ liệu của cột Region

Thư viện sklearn cung cấp LabelEncoder thực hiện tác vụ trên với bảng ánh xạ như sau:

Region	Label	Region	Label
Highlands	0	Midlands and northern mountains	3
Hong river Delta	1	North Central and Central Coast	4
Mekong Delta	2	South East	5

c. Đánh giá mô hình sau khi huấn luyện

Tương tự như đánh giá cho mô hình Hồi quy tuyến tính. Điểm số trung bình của các tỉnh cho ra là 19510, không tốt hơn mô hình hồi quy tuyến tính.

Vì vậy cần tinh chỉnh các tham số của mô hình cho phù hợp.

d. Tinh chỉnh tham số của mô hình

Các tham số của mô hình có thể tinh chỉnh là:

- **splitter:** chiến lược sử dụng để chọn nhánh ở mỗi node. Có hai lựa chọn là “best" - chọn nhánh tốt nhất và “random” - chọn nhánh ngẫu nhiên tốt nhất
- **max_depth:** chiều sâu tối đa của cây. Nếu là None thì sẽ mở rộng đến khi tất cả các lá là pure hoặc số lá nhỏ hơn “min_sample_leaf”, ở đây em chọn các giá trị là 1, 3, 5, 7 để kiểm tra
- **min_samples_leaf**: số lá tối thiểu của nhánh, ở đây em chọn 1, 2, 3, 4, 5 để kiểm tra
- **min_weight_fraction_leaf:** trọng số của nút lá/trọng số của đầu vào. Ở đây em chọn là 0.1; 0.2; 0.3; 0.4 (có thể mở rộng tới 1)
- **max_features:** số lượng đặc trưng được xem xét mỗi phần chọn nhánh
- **max_leaf_nodes**: số nút lá tối đa

Kết quả cho ra như sau:

```
{'max_depth': 7,
 'max_features': 'log2',
 'max_leaf_nodes': None,
 'min_samples_leaf': 1,
 'min_weight_fraction_leaf': 0.1,
 'splitter': 'best'}
```

Áp dụng mô hình với tham số được tinh chỉnh cho ra kết quả:

```
np.array(scores).mean()
```

```
17716.8211238995
```

Rõ ràng, kết quả đã tốt hơn trước nhưng không tốt bằng mô hình hồi quy tuyến tính.

e. Nhận xét

Điểm mạnh:

- Có khả năng đưa ra kết quả tốt hơn Hồi quy tuyến tính vì đưa ra phụ thuộc nhiều biến hơn.

Hạn chế:

- Không hiệu quả với tập dữ liệu này vì các yếu tố chủ quan và khách quan.
- Khách quan:
    - Dữ liệu tuân theo đường tuyến tính tốt hơn.
- Chủ quan:
    - Quá trình tinh chỉnh chưa tốt vì chưa bao quát hết các trường hợp có thể tinh chỉnh do sức mạnh của phần cứng không cho phép.
    - Lỗi lập trình:
        - Em gặp khó khăn vì mỗi lần chạy cây quyết định lại cho một kết quả khác nhau.
        - Em không thể tạo ra bộ test cho 5 năm sau đó.

Nếu có thể làm thêm, em sẽ tìm cách sử dụng 1 mô hình cho toàn bộ tỉnh thành thay vì mỗi tỉnh thành dùng 1 mô hình như trên.

3.3. Tỷ lệ lao động sẵn sàng của từng khu vực, các tác nhân nào có thể ảnh hưởng đến xu hướng thay đổi tỷ lệ lao động của từng khu vực?

Ý nghĩa khi trả lời câu hỏi:

- Có cái nhìn tổng quan về tỷ lệ người lao động của từng khu vực, xu hướng thay đổi về lực lượng lao động qua từng năm.
- Cố gắng tìm ra các tác nhân có thể ảnh hưởng đến xu hướng đó.

Các bước cần làm để trả lời câu hỏi:

- Tính tỷ lệ lao động trên 15 tuổi của từng khu vực qua từng năm.
- Trực quan hóa bằng biểu đồ heatmap và biểu đồ đường rồi rút ra nhận xét.
- Phân tích các tác nhân có khả năng sẽ liên quan đến xu hướng thay đổi tỷ lệ lao động và rút ra nhận xét.

Chi tiết thực hiện

Bước 1: Tính tỷ lệ lao động trên 15 tuổi của từng tỉnh

Ta tiến hành tính tỷ lệ lao động của mỗi tỉnh bằng cách lấy số lượng lao động chia cho dân số của tỉnh đó và lưu vào cột Labor_ratio.

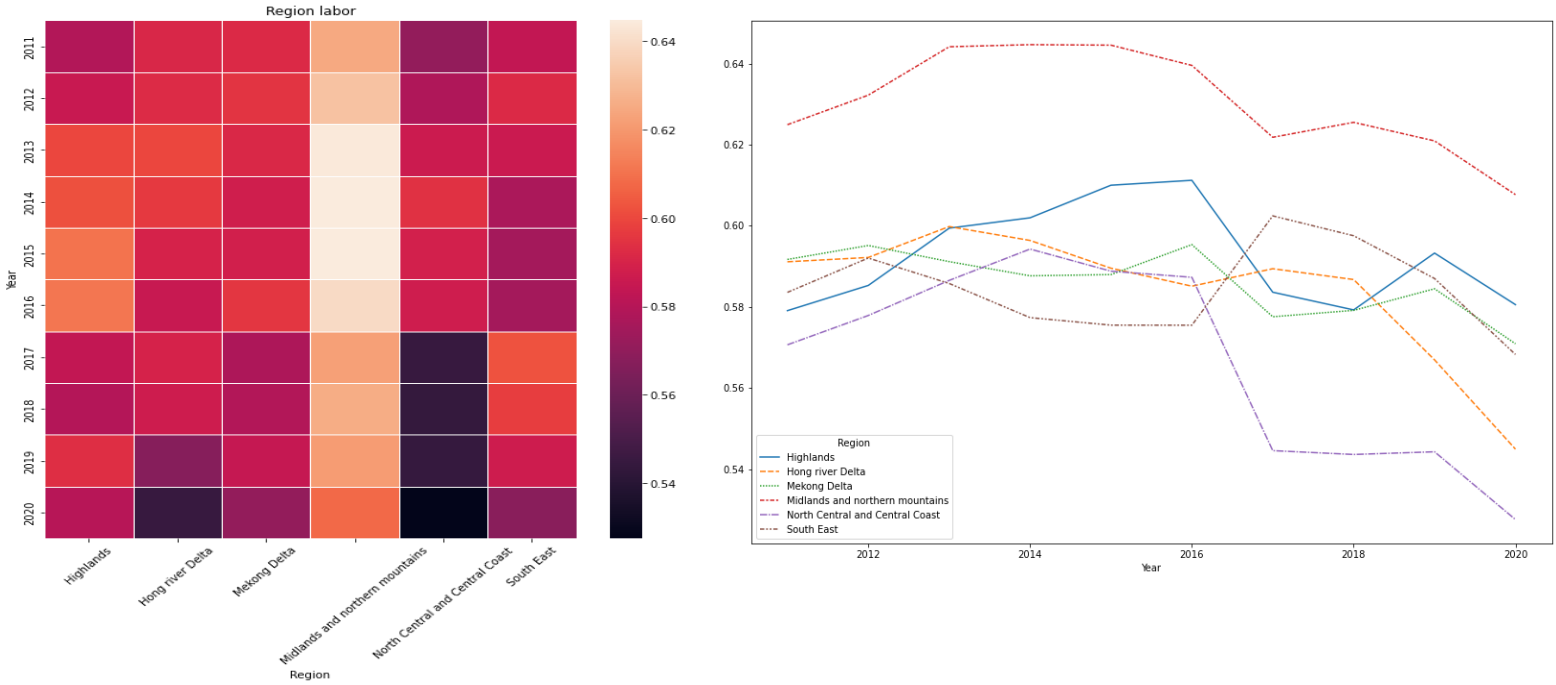
	Provinces/city	Average population	Population grow ratio	15+ labor	Region	Year	Labor_ratio
0	Ha Noi	6761.3	1.93	3572.90	Hong river Delta	2011	0.528434
1	Vinh Phuc	1011.4	0.38	608.30	Hong river Delta	2011	0.601444
2	Bac Ninh	1063.4	1.84	593.50	Hong river Delta	2011	0.558115
3	Quang Ninh	1168.0	0.93	675.00	Hong river Delta	2011	0.577911
4	Hai Duong	1729.8	0.78	1071.00	Hong river Delta	2011	0.619147
...	...	...	...	...	...	...	...
625	Can Tho	1240.7	0.39	716.78	Mekong Delta	2020	0.577722
626	Hau Giang	729.8	-0.33	402.33	Mekong Delta	2020	0.551288
627	Soc Trang	1195.7	-0.32	641.91	Mekong Delta	2020	0.536849
628	Bac Lieu	913.5	0.58	507.76	Mekong Delta	2020	0.555840
629	Ca Mau	1193.9	-0.03	669.77	Mekong Delta	2020	0.560993

Bước 2: Tính tỷ lệ lao động trung bình của từng khu vực qua từng năm

Tính tỷ lệ lao động trung bình của từng vùng qua các năm.

Region	Highlands	Hong river Delta	Mekong Delta	Midlands and northern mountains	North Central and Central Coast	South East
Year						
2011	0.579075	0.591130	0.591707	0.624949	0.570646	0.583586
2012	0.585297	0.592200	0.595151	0.632242	0.577886	0.592055
2013	0.599424	0.599824	0.591169	0.644178	0.586540	0.585780
2014	0.601983	0.596392	0.587670	0.644695	0.594271	0.577353
2015	0.610021	0.589546	0.587976	0.644571	0.588814	0.575504
2016	0.611232	0.585111	0.595370	0.639579	0.587302	0.575487
2017	0.583627	0.589419	0.577572	0.621847	0.544542	0.602459
2018	0.579262	0.586751	0.579144	0.625534	0.543597	0.597554
2019	0.593276	0.566933	0.584474	0.620953	0.544261	0.587004
2020	0.580561	0.544911	0.570912	0.607692	0.527541	0.568269

Bước 3: Trực quan hóa bằng Heat Map và Line Chart.



Bước 4: Từ hai biểu đồ trên, rút ra nhận xét

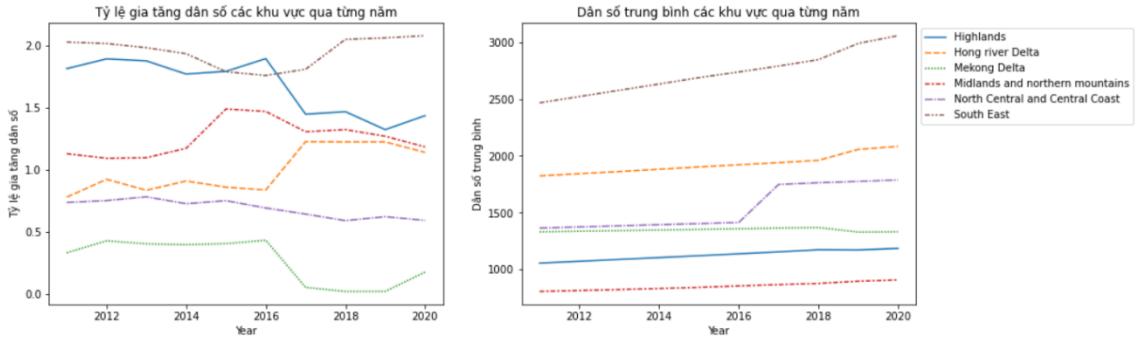
Nhận xét:

- Nhìn chung tỷ lệ người lao động trên 15 tuổi giữa các vùng không có sự chênh lệch nhiều lắm. Quan sát tone màu cùng với biểu đồ đường, ta có thể thấy rằng tỉ lệ người lao động trên 15 tuổi ở khu vực Trung du và miền núi Bắc Bộ là nhiều nhất. Việc này là dễ hiểu khi đa số người dân ở khu vực này chưa có trình độ học vấn cao nên họ thường có xu hướng học nghề và đi làm sau khi tốt nghiệp cấp 2. Việc tỉ lệ lao động các vùng đồng đều nhau như thế này cũng dẫn đến một tình trạng đó chính là lượng người lao động

ở vùng miền núi, nông thôn lại quá thừa trong khi nhu cầu nguồn nhân lực ở các vùng này lại không cao và các vùng thành thị vẫn đang trong tình trạng thiếu nhân lực lao động. Để giải quyết vấn đề trên, chính phủ cần phải có các chính sách để tạo cơ hội giúp người lao động ở vùng miền núi nông thôn có thể tìm việc làm ở thành thị, giúp đáp ứng thêm nhu cầu nhân lực ở đây.

- Ngoài ra ta cũng có thể thấy tỷ lệ lao động khoảng năm 2019-2020 có xu hướng giảm xuống, việc này có thể giải thích là do ảnh hưởng của trận đại dịch covid và việc giãn cách xã hội đã khiến nhiều người mất đi việc làm.

Bước 6: Xét các yếu tố về dân số để xem thử chúng có liên quan đến xu hướng tỷ lệ lao động hay không
Ta xét đến các yếu tố về dân số trung bình và tỷ lệ gia tăng dân số



Nhận xét:
Ta thấy dân số trung bình của các khu vực tăng đều qua từng năm nhưng tỷ lệ gia tăng dân số ở một vài nơi lại có nhiều lúc giảm mạnh, ví dụ như Đồng Bằng Sông Cửu Long dấu hiệu giảm từ 0.5 xuống ngưỡng 0 kể từ năm 2016, hay Bắc Trung Bộ và Duyên Hải miền Trung giảm nhẹ theo từng năm, hay Tây Nguyên cũng giảm dần từ 2016. Các thông số có vẻ chưa thực sự liên quan lắm đến tỷ lệ lao động. Để làm rõ điều này, ta chuyển tới bước tiếp theo là tính độ tương quan giữa các thông số.

Bước 7: Tính hệ số tương quan giữa các thông số về dân số so với tỷ lệ lao động của từng khu vực.

```
Correlation in Highlands region
Population grow ratio    0.486502
Average population      -0.108744
dtype: float64

Correlation in Hong river Delta region
Population grow ratio   -0.526665
Average population      -0.890253
dtype: float64

Correlation in Mekong Delta region
Population grow ratio    0.757882
Average population      -0.130824
dtype: float64

Correlation in Midlands and northern mountains region
Population grow ratio    0.186036
Average population      -0.653095
dtype: float64

Correlation in North Central and Central Coast region
Population grow ratio    0.888132
Average population      -0.934347
dtype: float64

Correlation in South East region
Population grow ratio    0.059276
Average population      -0.138370
dtype: float64
```

Nhận xét:

Ngoại trừ Bắc Trung Bộ và Duyên hải miền Trung thì các khu vực còn lại đều có độ tương quan rất thấp. Chứng tỏ hai tác nhân trên vẫn chưa đủ để ảnh hưởng mạnh đến tỷ lệ lao động, một tác nhân khác mà ta có thể nghĩ đến chính là việc dân số nhập/di cư từ khu vực này sang khu vực khác. Do tập dữ liệu này không có thông số khác cho việc nhập/di cư nên ta chỉ có thể lờ mờ đoán được từ các thông số về dân số trung bình mà tác nhân chủ yếu đến từ tỷ lệ sinh/tử và tỷ lệ nhập/di cư. Thông số liên quan đến tỷ lệ gia tăng dân số không thực sự có liên quan trực tiếp đến tỷ lệ lao động vì số lượng ca sinh hoàn toàn không liên quan tới lượng người lao động vì những đứa trẻ sinh vào khoảng thời gian đó cần phải 15 năm sau để có thể tham gia vào lực lượng lao động, tương tự thì ca tử phần lớn là những người già đã về hưu và không liên quan đến lực lượng lao động. Tuy vậy, nhờ nó mà ta có thể suy đoán được xu hướng nhập/di cư của người lao động ở các khu vực này và từ đó suy ra phần nào sự thay đổi về tỷ lệ người lao động. Ví dụ như tỉ lệ gia tăng dân số giảm ở 1 số khu vực nhưng dân số trung

bình lại tăng, chứng tỏ đã có 1 lượng người di cư tới đây, ta cũng có thể suy luận ngược lại với người nhập cư.

3.4. Vẽ biểu đồ và nhận xét về tỉ lệ tăng dân số với dân số trung bình? Có hiện tượng dân số trung bình tăng nhưng độ tăng dân số giảm hay không? Vì sao?

Ý nghĩa khi trả lời câu hỏi:

- Thấy được những thông tin ẩn đằng sau của tỷ lệ gia tăng dân số và dân số trung bình, khi chúng ta biểu diễn thành biểu đồ.
- Biết thêm được vì sao có hiện tượng dân số trung bình tăng, nhưng tỷ lệ gia tăng dân số giảm.

Các bước cần làm để để trả lời câu hỏi:

Bước 1: Tính trung bình cho các cột Population grow ratio và Average population theo khu vực.

- Gom cụm cột Region, Year: tính giá trị trung bình của cột Average population.
- Gom cụm cột Region, Year: tính giá trị trung bình của cột Population grow ratio.
- Chuyển các dataframe được tạo ở 2 ý trên thành bảng dữ liệu pivot(cột là tên các khu vực, dòng là các năm).

Region	Highlands	Hong river Delta	Mekong Delta	Midlands and northern mountains	North Central and Central Coast	South East
Year						
2011	1.812	0.780000	0.333077	1.128571	0.737143	2.025000
2012	1.890	0.922727	0.429231	1.092143	0.752143	2.013333
2013	1.874	0.835455	0.404615	1.096429	0.782143	1.980000
2014	1.768	0.909091	0.398462	1.172143	0.727143	1.931667
2015	1.790	0.860000	0.406154	1.487857	0.751429	1.786667
2016	1.892	0.837273	0.433846	1.467857	0.692857	1.756667
2017	1.446	1.225455	0.055385	1.305000	0.643571	1.806667
2018	1.466	1.223636	0.022308	1.322143	0.591429	2.046667
2019	1.322	1.223636	0.023077	1.268571	0.622857	2.058333
2020	1.434	1.139091	0.177692	1.183571	0.593571	2.076667

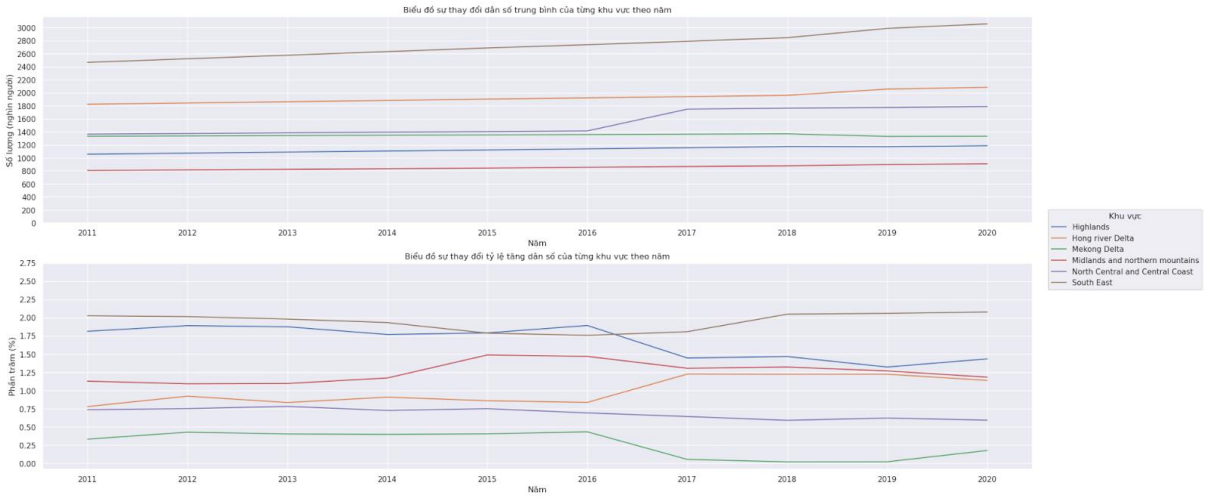
Tỷ lệ gia tăng dân số trung bình theo khu vực

Region Year	Highlands	Hong river Delta	Mekong Delta	Midlands and northern mountains	North Central and Central Coast	South East
2011	1056.42	1824.190909	1331.284615	807.185714	1364.621429	2466.583333
2012	1072.66	1843.163636	1336.892308	815.535714	1374.528571	2521.783333
2013	1089.14	1861.990909	1342.215385	824.035714	1384.828571	2576.550000
2014	1105.18	1882.290909	1347.515385	833.400000	1394.478571	2631.716667
2015	1121.58	1902.327273	1353.100000	843.121429	1404.135714	2687.966667
2016	1138.64	1921.254545	1358.515385	856.021429	1414.200000	2737.383333
2017	1155.70	1940.190909	1364.461538	867.778571	1748.042857	2789.933333
2018	1174.20	1960.581818	1369.592308	878.050000	1763.478571	2845.716667
2019	1172.26	2056.381818	1329.423077	897.807143	1774.957143	2988.383333
2020	1186.40	2083.654545	1332.192308	908.992857	1788.278571	3057.150000

Dân số trung bình theo khu vực

Bước 2: Vẽ biểu đồ theo từng khu vực.

- Chia không gian vẽ thành 2 biểu đồ: biểu đồ trên thể hiện sự thay đổi dân số trung bình. Biểu đồ dưới thể hiện sự thay đổi tỷ lệ gia tăng dân số. Để dễ so sánh và phát hiện các điểm bất thường.



Biểu diễn biểu đồ

Nhận xét:

Dân số trung bình:

- Thứ tự dân số ở các khu vực không có thay đổi, thứ tự tăng dần dân số: ĐB Sông Hồng, Đông Nam Bộ, ĐB Sông Cửu Long, Bắc Trung Bộ và duyên hải miền Trung, Trung du miền núi phía Bắc, Tây Nguyên.
- Khu vực Đồng bằng Sông Hồng, Đông Nam Bộ có dân số trung bình tăng theo năm, riêng ở khu vực Đồng bằng Sông Cửu Long có sự giảm

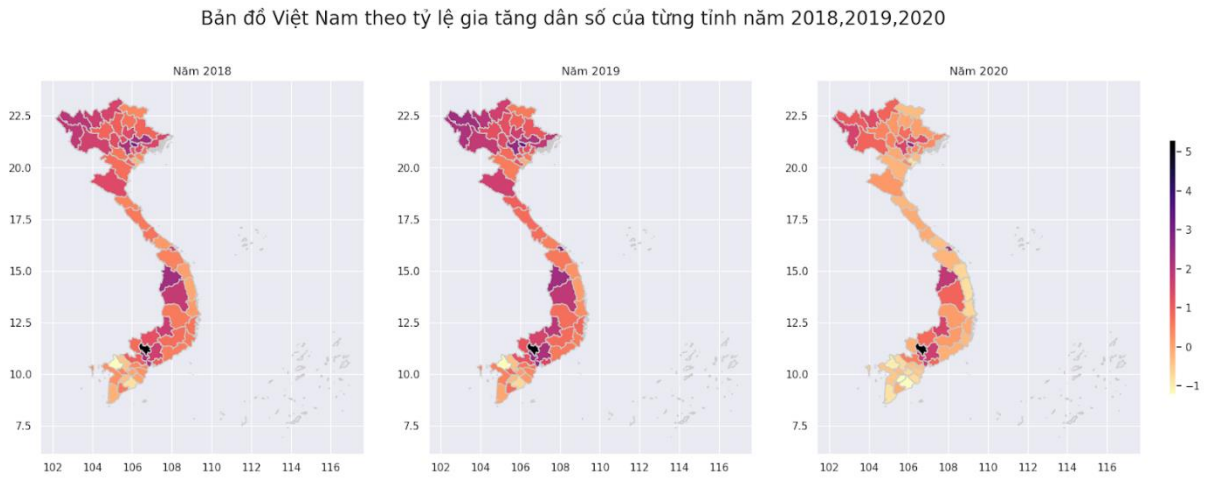
dân số ở năm 2018-2020. Các khu vực khác không có sự thay đổi nhiều.

Tỉ lệ tăng dân số:

- Các khu vực có sự tăng giảm tỉ lệ dân số theo từng năm. Riêng ở khu vực Đồng bằng Sông Cửu Long ta thấy sự bất thường ở từ năm 2017-2019 tỉ lệ xấp xỉ bằng 0.
- Năm 2016, 2017 hầu hết khu vực đều có sự thay đổi rõ ràng.

Bước 3: Vẽ chi tiết biểu đồ tỷ lệ tăng dân số giữa các tỉnh trong 3 năm: 2018, 2019, 2020.

- Đọc file geojson bản đồ Việt Nam vào vietnam_map.
- Xử lý dữ liệu tên các tỉnh cho đồng bộ (['Ho Chi Minh' 'Thua Thien Hue'], ['TP. Ho Chi Minh' 'Thua Thien - Hue']).
- Thêm các cột thuộc tính dân số trung bình và tỷ lệ gia tăng dân số theo từng năm cho dataframe vietnam_map.
- Lọc dữ liệu các năm cần vẽ: 2018, 2019, 2020 rồi thực hiện vẽ biểu đồ Choropleth Map.



Nhận xét:

- Có sự giảm tone màu của năm 2020 so với 2 năm 2018 và 2019.

- Các tỉnh: An Giang, Cà Mau, Đồng Tháp, Hậu Giang, Nam Định, Sóc Trăng, Vĩnh Long có nhiều hơn 1 lần có tỉ lệ gia tăng dân số âm.

Bước 4: Trả lời câu hỏi: Có hiện tượng dân số trung bình tăng nhưng độ tăng dân số giảm hay không và giải thích?

Trả lời:

- Tồn tại hiện tượng dân số trung bình tăng nhưng tỷ lệ tăng dân số giảm.

Giải thích:

- Tỷ lệ gia tăng dân số là tỉ lệ thể hiện sự chênh lệch giữa tỷ lệ sinh thô và tỷ lệ tử vong thô trong một dân số nhất định. Nên khi tỷ lệ dân số giảm có nghĩa là dân số đang già hóa hơn, tỷ lệ tử vong nhiều hơn tỷ lệ sinh.
- Tỷ lệ tăng dân số giảm, nhưng vẫn dương.
- Dân số trẻ và đang trong độ tuổi sinh nở nhiều, dẫn đến tỷ lệ sinh thô sẽ cao hơn tỷ lệ tử vong.