Intro To DataScience

Semester I – 22KHDL

Mrs. Nguyen Ngoc Thao
Mr. Le Nhut Nam

Nội Dung

01 Thông tin

02 Giới thiệu

03 Giai đoạn

04 Quá trình thực hiện

05 Thực nghiệm

06 Kết luận

01

Thông tin

Thông tin

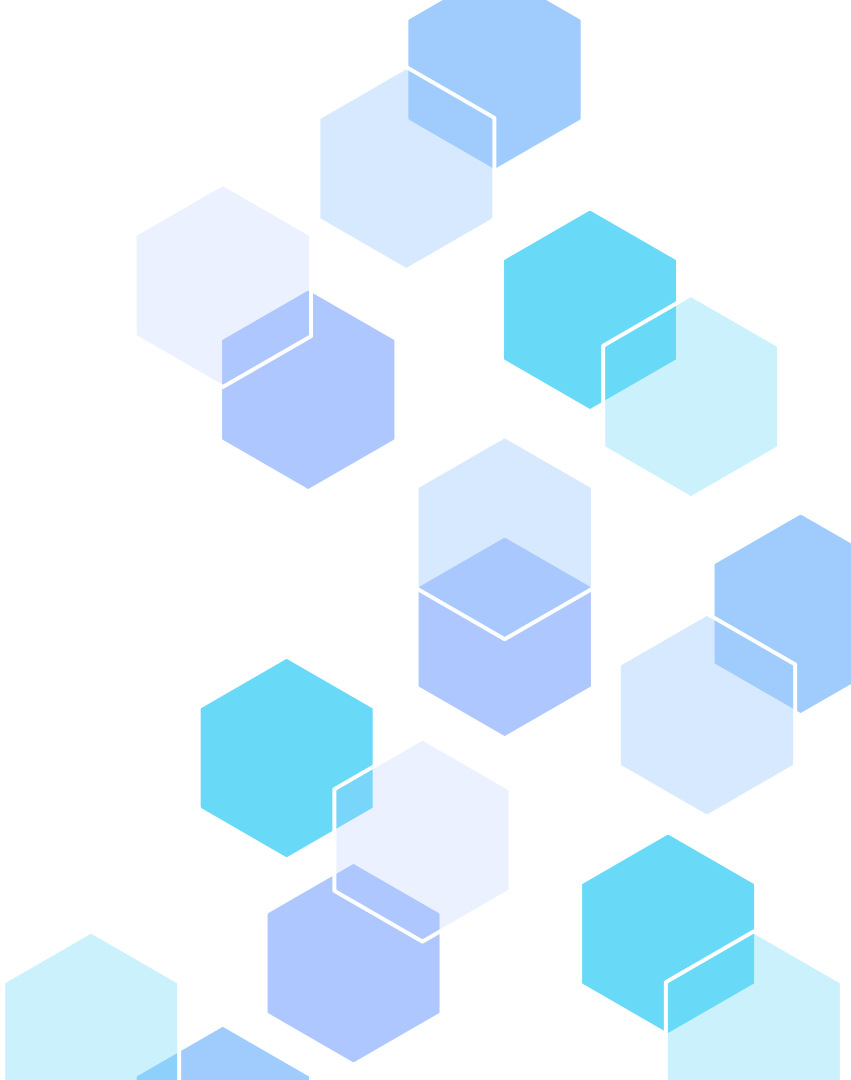
Đồ án được đưa ra nhằm mục đích áp dụng các kiến thức, kỹ thuật được dạy trong lĩnh vực data science để giải quyết bài toán thực tế với dữ liệu trên internet.

Thành viên

19127607 – Trần Nguyên Trung

02

Giới thiệu



Giới thiệu

Đề tài: Khảo sát dân số Việt Nam ở các 63 tỉnh thành và các vùng.
Lý do lựa chọn đề tài:

- Là vấn đề thực tế
- Có nguồn dữ liệu đáng tin cậy lấy từ Tổng Cục Thống Kê Việt Nam
- Hiểu rõ hơn về quy mô dân số Việt Nam

03

Giai Đoạn

Thời Gian Thực Hiện

Thời gian	Công Việc	Thực Hiện
14/10 – 29/10	Tìm kiếm, lựa chọn, thu thập dữ liệu	Trung
30/10 – 06/11	Mô tả, khám phá dữ liệu	Trung
30/11 – 06/11	Tiền xử lý dữ liệu	Trung
06/11 – 15/11	Đặt câu hỏi, trực quan hoá dữ liệu, rút ra ý nghĩa từ dữ liệu	Trung
15/11 – 19/11	Giải thích và kết luận cho các câu hỏi	Trung
20/11 – 04/12	Xây dựng mô hình để dự đoán	Trung
06/11 – 30/12	Kiểm tra, đánh giá đồ án	Trung
14/10 – 30/12	Viết báo cáo, slide và các tài liệu liên quan	Trung

04

Quá Trình Thực Hiện

Tổ Chức Và Quản Lý

Git Flow:

- Toàn bộ đồ án(gồm dữ liệu, code crawl data, notebook,…) được lưu trữ ở github.
- Các thay đổi đều được commit vào các sub-branch sau đó tạo pull request và review kỹ càng trước khi được merge vào nhánh main.

Môi Trường:

- Crawl Data: Mã nguồn được viết bằng python, sử dụng công cụ Selenium.
- Notebook: Môi trường thực thi notebook là Google Colab.

Bộ Dữ Liệu

- Dữ liệu được thu thập từ trang web của Tổng cục Thống kê Việt Nam https://www.gso.gov.vn/
- Gồm 2 bộ dữ liệu chính gồm:
  - Diện tích, dân số và mật độ dân số phân theo địa phương.
  - Tỷ suất sinh thô, tỷ suất chết thô và tỷ lệ tăng tự nhiên của dân số phân theo địa phương.
- Dữ liệu ghi nhận sự thay đổi về sự gia tăng dân số theo 63 tỉnh, thành phố của Việt Nam, được quan sát từ năm 2011 đến năm 2023.

Trường Dữ Liệu

	Địa phương	Năm	Diện tích(Km2)	Dân số trung bình (Nghìn người)	Mật độ dân số (Người/km2)	Diện tích(Km2) (*)	Diện tích (Km2) (*)	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13
0	An Giang	2011	3536,7	2097,5	593,1	..	..	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	An Giang	2012	3536,7	2077,9	587,5	..	..	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	An Giang	2013	3536,7	2051,6	580,1	..	..	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	An Giang	2014	3536,7	2024,6	572,5	..	..	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	An Giang	2015	3536,7	2000,9	565,8	..	..	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Diện tích, dân số và mật độ dân số phân theo địa phương.

	Địa phương	Phân tổ	2005	2007	2008	2009	2010	2011	2012	2013	...	Unnamed: 30	Unnamed: 31	Unnamed: 32	Unnamed: 33	Unnamed: 34	Unnamed: 35	Unnamed: 36	Unnamed: 37	Unnamed: 38	Unnamed: 39
0	An Giang	Tỷ suất sinh thô	18,40	16,90	16,90	17,60	17,30	16,40	17,50	16,30	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	An Giang	Tỷ suất chết thô	5,20	5,10	5,00	8,00	7,90	7,30	7,80	7,80	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	An Giang	Tỷ lệ tăng tự nhiên	13,20	11,80	11,90	9,70	9,30	9,00	9,60	8,50	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Bà Rịa - Vũng Tàu	Tỷ suất sinh thô	18,50	16,90	17,10	17,70	15,60	15,20	14,90	16,20	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Bà Rịa - Vũng Tàu	Tỷ suất chết thô	4,40	4,40	4,10	6,60	6,50	6,30	7,10	7,50	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Tỷ suất sinh thô, tỷ suất chết thô và tỷ lệ tăng tự nhiên của dân số phân theo địa phương.

Xử lý dữ liệu

Dữ liệu bị thiếu:

- Bộ dữ liệu 1:
  - Cột “Diện tích(Km2)” có một số dữ liệu bị thiếu, được bổ sung ở các cột “Diện tích phụ lục 1(Km2)” và “Diện tích phụ lục 2(Km2)”, do đó tiến hành điền giá trị còn thiếu bằng 2 cột phụ lục và loại bỏ 2 cột phụ lục.
- Bộ dữ liệu 2: Không có giá trị bị thiếu.

Chuẩn hoá dữ liệu:

- Bộ dữ liệu 2 có cấu trúc khác bộ dữ liệu 1, do đó tiến hành **melt** để làm phẳng dữ liệu và dùng **pivot_table** để đưa về cùng cấu trúc với bộ dữ liệu 1.
- Loại bỏ các cột không có ý nghĩa(UnNamed)
- Ở 2 bộ dữ liệu, cột “Địa phương” có kiểu dữ liệu object do đó tiến hành đưa về string.

```
Data columns (total 8 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Địa phương                      910 non-null    object
 1   Năm                             910 non-null    int64
 2   Diện tích(Km2)                  910 non-null    float64
 3   Dân số trung bình (Nghìn người) 910 non-null    float64
 4   Mật độ dân số (Người/km2)       910 non-null    float64
 5   Tỷ lệ tăng tự nhiên             910 non-null    float64
 6   Tỷ suất chết thô                910 non-null    float64
 7   Tỷ suất sinh thô                910 non-null    float64
dtypes: float64(6), int64(1), object(1)
memory usage: 57.0+ KB
```

```
set(df['Địa phương'].unique()) - set(data_pivot['Địa phương'].unique())
```

```
{'Hoà Bình', 'Khánh Hoà', 'Thanh Hoá', 'Thừa Thiên Huế', 'Tp.Hồ Chí Minh'}
```

```
set(data_pivot['Địa phương'].unique()) - set(df['Địa phương'].unique())
```

```
{'Hà Tây',
 'Hòa Bình',
 'Khánh Hòa',
 'Thanh Hóa',
 'Thừa Thiên - Huế',
 'Tp. Hồ Chí Minh'}
```

- Có một số tỉnh ở 2 bộ dữ liệu bị sai do đánh máy do đó tiến hành đổi tên lại cho phù hợp.
- Dùng **get_dummies** để đưa cột “Địa phương” về vector one-hot để tiến hành đưa vào mô hình.

Dữ Liệu Tổng Hợp

	Địa phương	Năm	Diện tích(Km2)	Dân số trung bình (Nghìn người)	Mật độ dân số (Người/km2)	Tỷ lệ tăng tự nhiên	Tỷ suất chết thô	Tỷ suất sinh thô
0	An Giang	2011	3536.7	2097.5	593.1	9.00	7.30	16.40
1	An Giang	2012	3536.7	2077.9	587.5	9.60	7.80	17.50
2	An Giang	2013	3536.7	2051.6	580.1	8.50	7.80	16.30
3	An Giang	2014	3536.7	2024.6	572.5	7.90	8.80	16.70
4	An Giang	2015	3536.7	2000.9	565.8	4.90	7.50	12.50
...	...	...	...	...	...	...	...	...
905	Đồng Tháp	2019	3383.8	1598.8	472.0	5.00	6.90	11.90
906	Đồng Tháp	2020	3382.3	1600.0	473.0	6.30	6.66	12.95
907	Đồng Tháp	2021	3382.3	1601.3	473.0	4.72	7.60	12.40
908	Đồng Tháp	2022	3382.3	1600.2	473.0	1.70	8.70	10.40
909	Đồng Tháp	2023	3382.3	1600.2	473.1	5.89	6.22	12.10

910 rows × 8 columns

Dữ liệu và các trường dữ liệu sau khi được tổng hợp lại

- **Địa phương**: tên 63 tỉnh/thành phố của Việt Nam
- **Năm**: năm ghi nhận số liệu (format: yyyy)
- **Diện tích**: diện tích của thành phố (km2)
- **Dân số trung bình**: theo từng thành phố (nghìn người)
- **Mật độ dân số**: Mật độ dân số là số người sinh sống trên một đơn vị diện tích, lấy theo giá trị trung bình. Từ giá trị này bạn có thể suy ra lượng tài nguyên mà một khu vực cần có, và dựa vào đó so sánh các khu vực khác nhau. Công thức tính:

  Mật độ dân số = (Dân số trung bình / Diện tích) * 1000 (đơn vị: Người/km2)
- **Tỉ suất sinh thô**: Tỷ suất sinh thô (CBR – Crude Birth Rate): được sử dụng rộng rãi trong dân số học, đó là tỷ số giữa số trẻ em được sinh ra trong năm so với số dân trung bình ở cùng thời gian ấy với đơn vị tính bằng phần nghìn (đơn vị ‰). Tỷ suất sinh thô được tính theo công thức :

  CBR = (Số trẻ em sinh ra trong năm / Tổng số dân trung bình của năm) * 1000
- **Tỷ suất chết thô** (CDR - Crude Death Rate) là chỉ tiêu đơn giản nhưng phổ biến nhất trong việc đánh giá mức tử vong của dân số. Nó được xác định bằng số người chết trong năm tính bình quân cho 1.000 dân số năm đó (đơn vị ‰)

  CDR = (Số người chết trong năm / Tổng số dân trung bình của năm) * 1000
- **Tỷ lệ tăng tự nhiên** cho biết cứ 1.000 dân số trung bình của một năm, thì có bao nhiêu người tăng lên trong năm do hậu quả của 2 yếu tố sinh ra và chết đi. Tỷ lệ gia tăng tự nhiên hay tỉ suất gia tăng tự nhiên là đến sự khác biệt hay sự chênh lệch giữa tỷ lệ sinh thô và tỷ lệ tử vong thô của một dân số nhất định (đơn vị ‰). Công thức xác định tỷ lệ gia tăng tự nhiên cụ thể như sau:

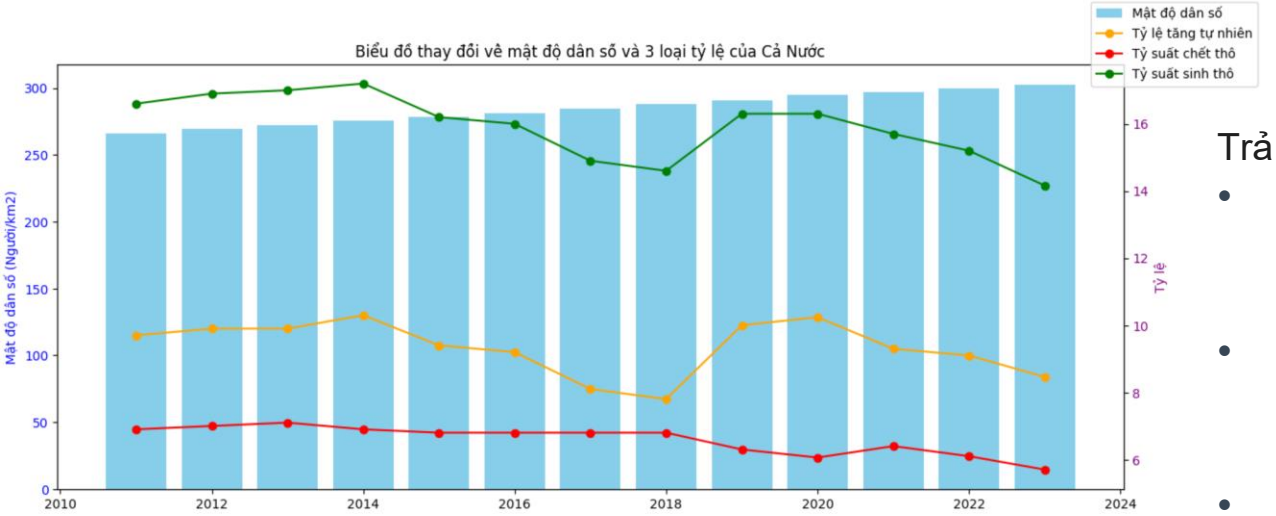
  Tỷ lệ tăng tự nhiên = Tỷ suất sinh thô – tỷ suất chết thô

Ý Nghĩa Dữ Liệu

- Đánh giá sự gia tăng dân số tự nhiên được hiểu cơ bản là quá trình tái sản xuất dân cư, thế hệ già được thay thế bằng thế hệ trẻ
- Tổ chức Y tế Thế giới (WHO) đã có được tất cả các giá trị được tính toán cho mỗi quốc gia trên toàn thế giới để từ đó có thể lên kế hoạch về hỗ trợ từng quốc gia.
- Tổ chức Y tế Thế giới cũng đã sử dụng các giá trị của tốc độ gia tăng tự nhiên để nhằm mục đích có thể đánh giá tiền tệ, nguồn nhân lực và sự hỗ trợ về mặt kĩ thuật đã cung cấp cho từng quốc gia.
- Để đưa ra các giải pháp thay đổi cơ cấu dân số, đồng thời một số vấn đề khác liên quan đến kinh tế và môi trường.

Câu Hỏi 1

- Câu Hỏi: Tổng quát về dân số Việt Nam qua các năm?
- Mục Tiêu: Khảo sát dân số của Cả Nước để có cái nhìn tổng quát.
- Thực Hiện: Vẽ biểu đồ Barchart(Mật độ dân số) kết hợp với biểu đồ Đường(3 loại tỷ suất).

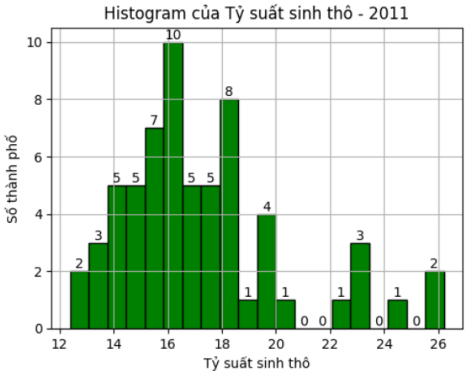


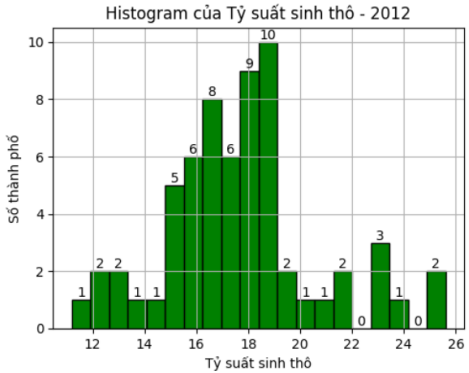
Trả lời:

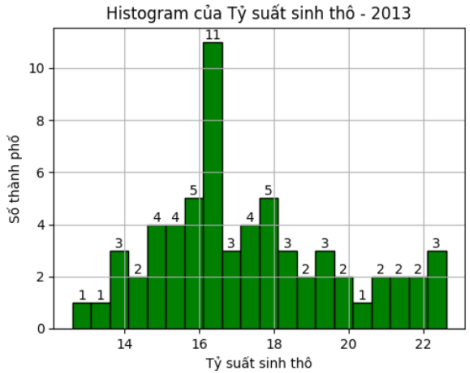
- Từ biểu đồ có thể thấy được mật độ dân số liên tục tăng dần đều theo các năm.
- Tỷ lệ sinh thô cao hơn nhiều so với tỷ lệ chết thô. Cả 3 tỷ lệ đều có biến động, đặc biệt là năm 2017 - 2020.
- Tỷ suất chết thô giảm nhẹ theo từng năm tuy nhiên tăng trở lại vào năm 2021 có thể do đại dịch Covid 19 ở Việt Nam.

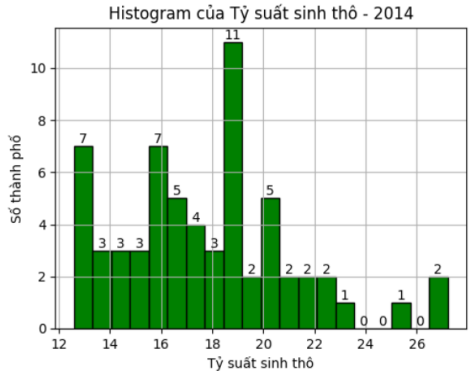
Câu Hỏi 2

- Câu Hỏi: Tỷ suất sinh thô, tỷ suất chết thô, tỷ lệ gia tăng tự nhiên của các thành phố qua các năm có sự biến đổi như thế nào?
- Mục Tiêu: Nhận xét phân phối về 3 loại tỷ suất từ năm 2011 - 2023
- Thực hiện: Dùng Histogram để biểu diễn sự phân phối của 3 trường dữ liệu Tỷ suất trên các thành phố theo các năm. Trong đó cột y là số thành phố, cột x là tỷ lệ. Số lượng thành phố ở một mức tỷ lệ nhất định sẽ được map vào độ dài tương ứng của từng cột.

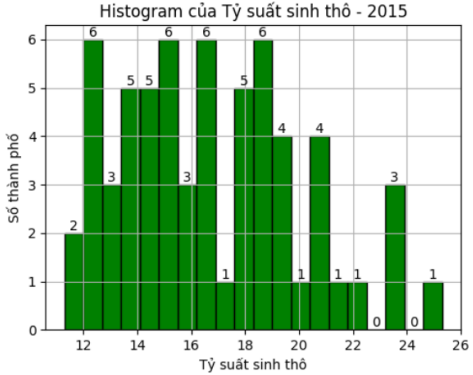


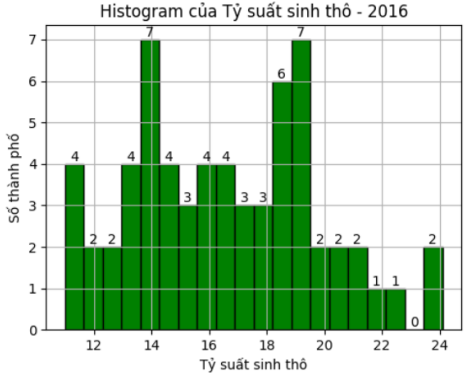


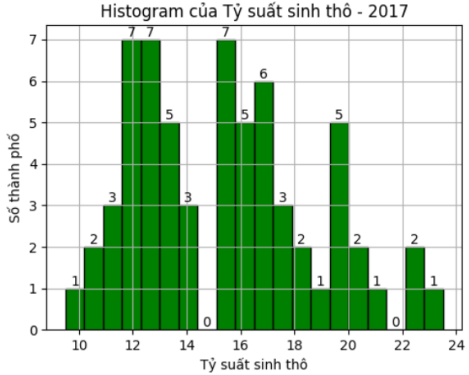


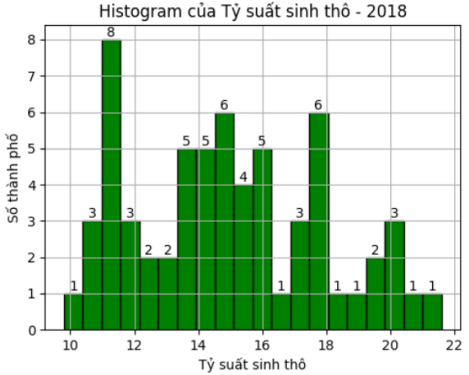


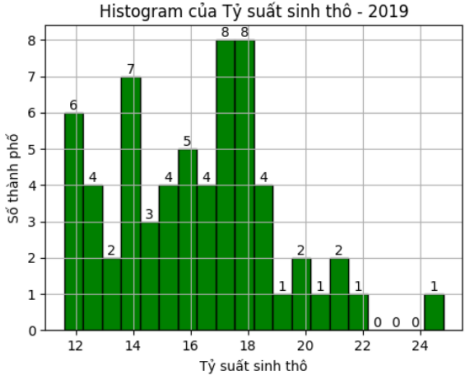
Câu Hỏi 2

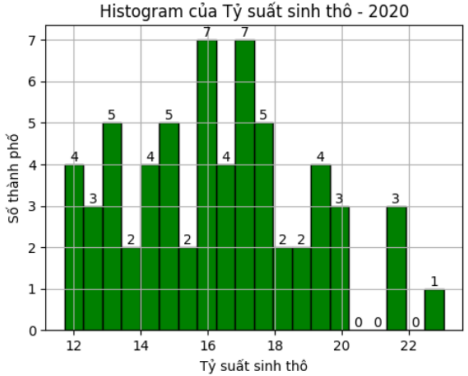

Histogram của Tỷ suất sinh thô - 2015
Số thành phố
Tỷ suất sinh thô

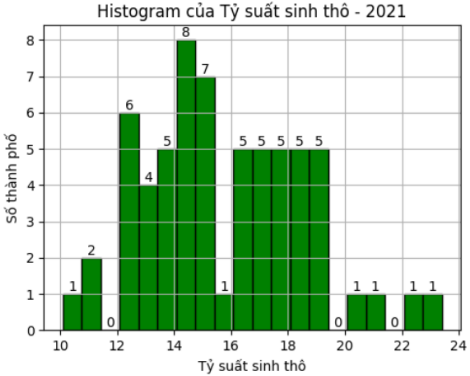

Histogram của Tỷ suất sinh thô - 2016
Số thành phố
Tỷ suất sinh thô

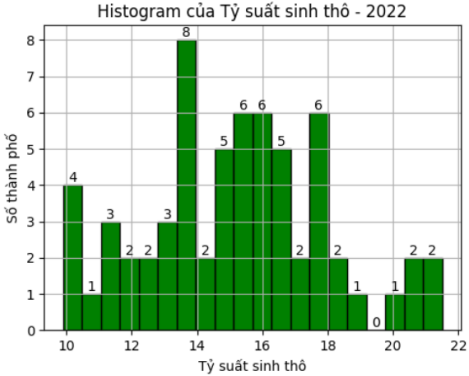

Histogram của Tỷ suất sinh thô - 2017
Số thành phố
Tỷ suất sinh thô


Histogram của Tỷ suất sinh thô - 2018
Số thành phố
Tỷ suất sinh thô

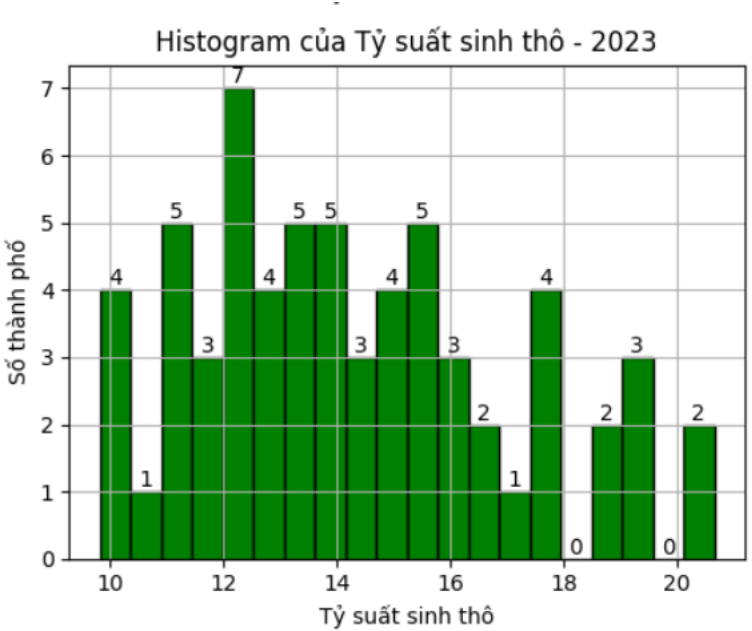

Histogram của Tỷ suất sinh thô - 2019
Số thành phố
Tỷ suất sinh thô


Histogram của Tỷ suất sinh thô - 2020
Số thành phố
Tỷ suất sinh thô


Histogram của Tỷ suất sinh thô - 2021
Số thành phố
Tỷ suất sinh thô


Histogram của Tỷ suất sinh thô - 2022
Số thành phố
Tỷ suất sinh thô

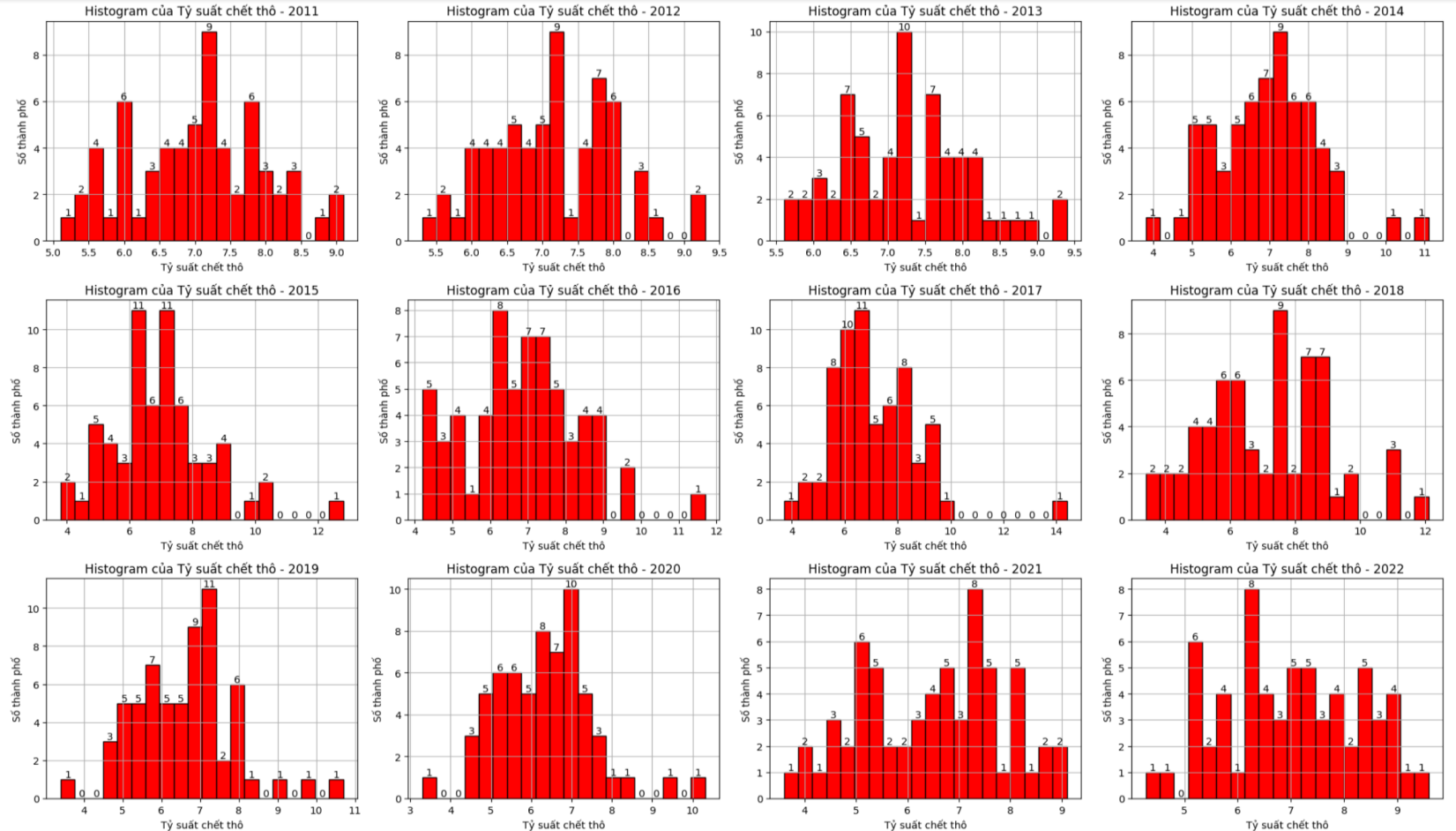
Câu Hỏi 2



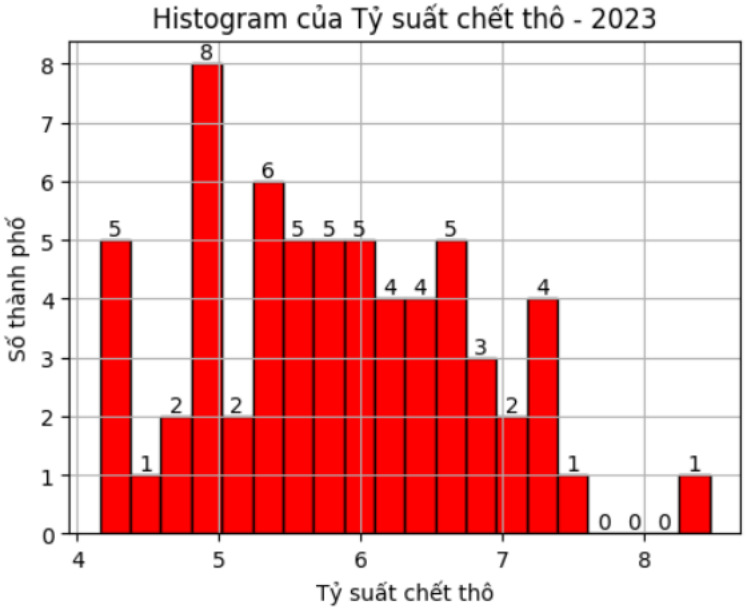
Trả lời:

- Quan sát biểu đồ biến đổi theo từng năm, có thể thấy được tỷ suất sinh thô của các thành phố có xu hướng giảm dần (histogram di chuyển từ phải sang trái).

Câu Hỏi 2


Histogram của Tỷ suất chết thô - 2011
Histogram của Tỷ suất chết thô - 2012
Histogram của Tỷ suất chết thô - 2013
Histogram của Tỷ suất chết thô - 2014
Histogram của Tỷ suất chết thô - 2015
Histogram của Tỷ suất chết thô - 2016
Histogram của Tỷ suất chết thô - 2017
Histogram của Tỷ suất chết thô - 2018
Histogram của Tỷ suất chết thô - 2019
Histogram của Tỷ suất chết thô - 2020
Histogram của Tỷ suất chết thô - 2021
Histogram của Tỷ suất chết thô - 2022
Số thành phố
Tỷ suất chết thô

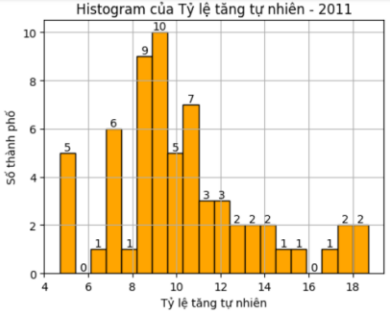
Câu Hỏi 2

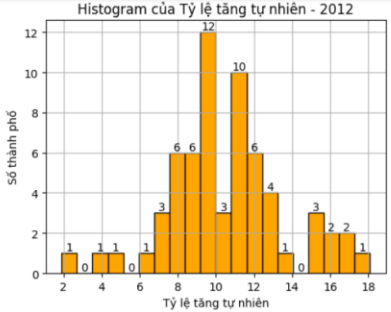


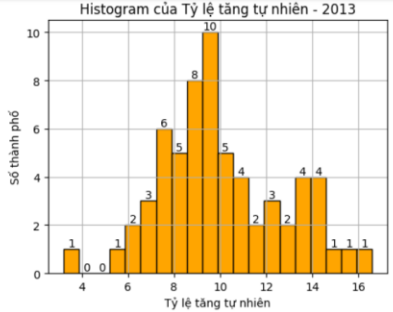
Trả lời:

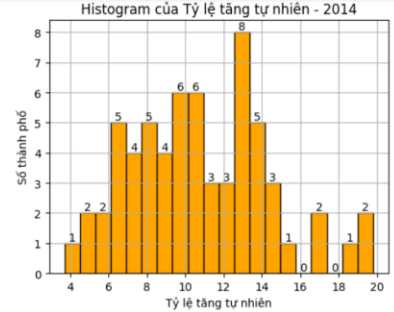
- Quan sát biểu đồ biến đổi theo từng năm, có thể thấy được tỷ suất chết thô thay đổi ít và cũng có xu hướng giảm dần.

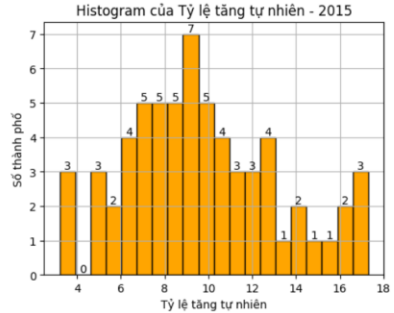
Câu Hỏi 2

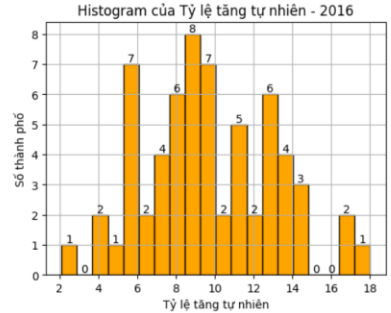

Histogram của Tỷ lệ tăng tự nhiên - 2011
Số thành phố
Tỷ lệ tăng tự nhiên

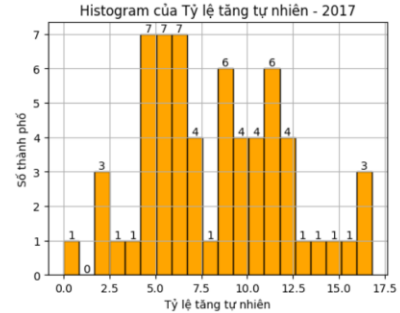

Histogram của Tỷ lệ tăng tự nhiên - 2012
Số thành phố
Tỷ lệ tăng tự nhiên

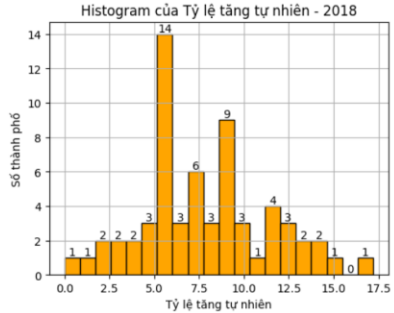

Histogram của Tỷ lệ tăng tự nhiên - 2013
Số thành phố
Tỷ lệ tăng tự nhiên

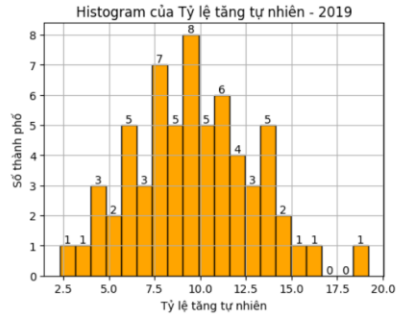

Histogram của Tỷ lệ tăng tự nhiên - 2014
Số thành phố
Tỷ lệ tăng tự nhiên

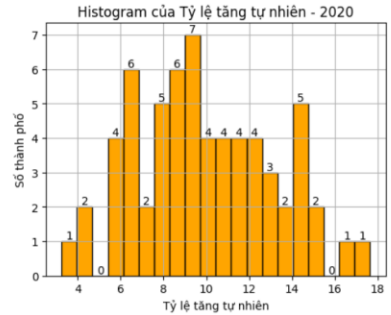

Histogram của Tỷ lệ tăng tự nhiên - 2015
Số thành phố
Tỷ lệ tăng tự nhiên

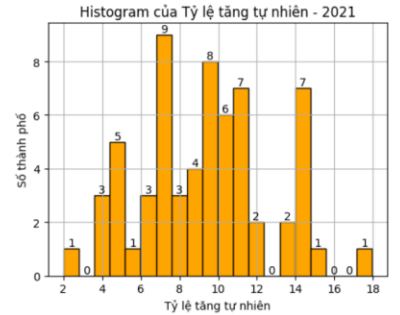

Histogram của Tỷ lệ tăng tự nhiên - 2016
Số thành phố
Tỷ lệ tăng tự nhiên

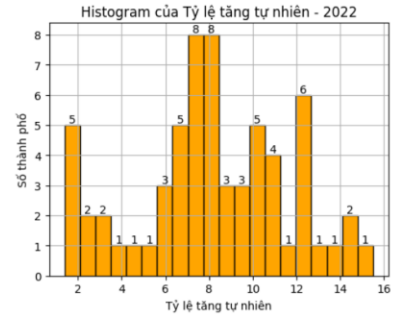

Histogram của Tỷ lệ tăng tự nhiên - 2017
Số thành phố
Tỷ lệ tăng tự nhiên


Histogram của Tỷ lệ tăng tự nhiên - 2018
Số thành phố
Tỷ lệ tăng tự nhiên

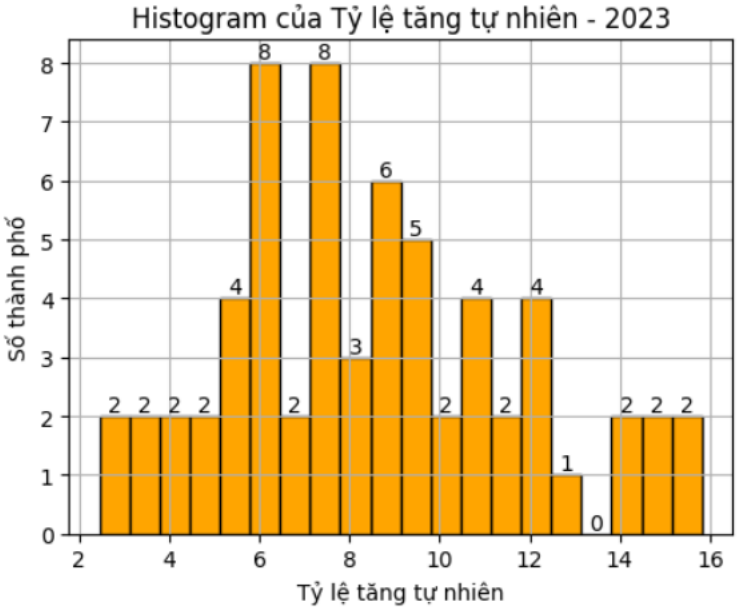

Histogram của Tỷ lệ tăng tự nhiên - 2019
Số thành phố
Tỷ lệ tăng tự nhiên


Histogram của Tỷ lệ tăng tự nhiên - 2020
Số thành phố
Tỷ lệ tăng tự nhiên


Histogram của Tỷ lệ tăng tự nhiên - 2021
Số thành phố
Tỷ lệ tăng tự nhiên


Histogram của Tỷ lệ tăng tự nhiên - 2022
Số thành phố
Tỷ lệ tăng tự nhiên

Câu Hỏi 2



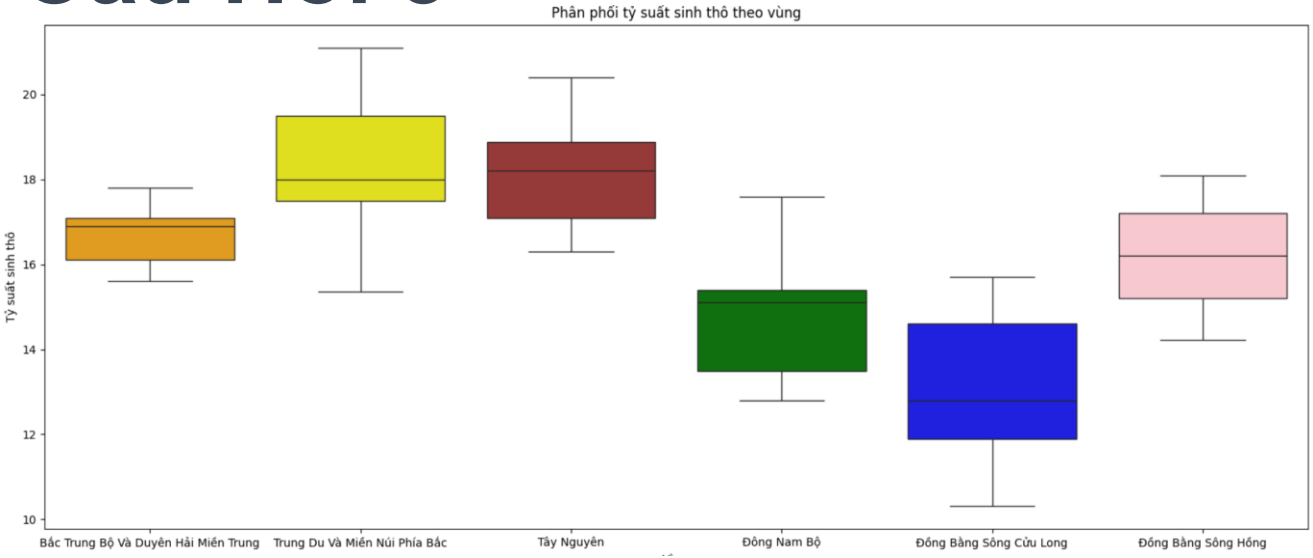
Trả lời:

- Tỷ lệ gia tăng tự nhiên biến động theo từng năm, có thể thấy vào năm 2014 nhiều thành phố có tỷ lệ tăng tự nhiên tập trung ở mức từ 10 đến 14.
- Tỷ lệ tăng tự nhiên thấp nhất vào năm 2017 và 2018, sau đó lại bắt đầu tăng cho đến năm 2021 và bắt đầu giảm từ năm 2022 đến 2023.

Câu Hỏi 3

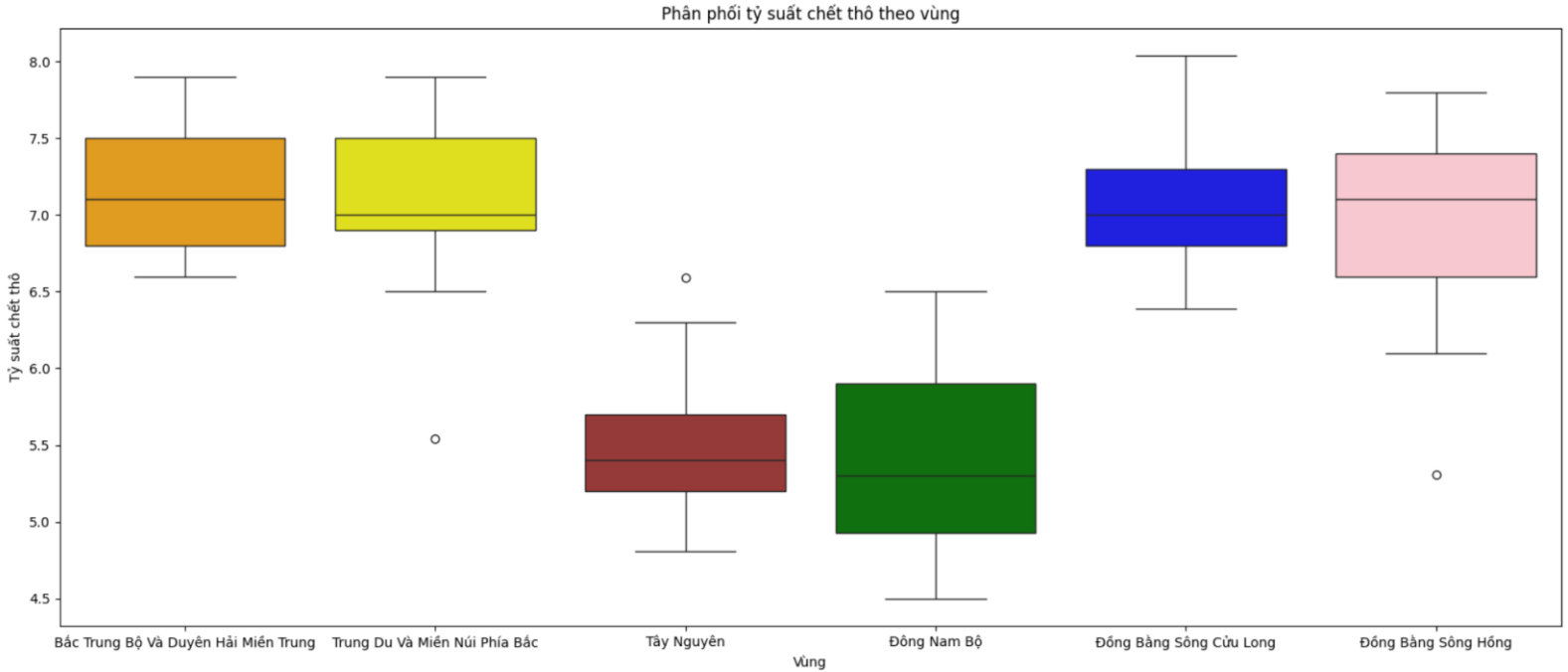
- Câu Hỏi: So sánh về phân bổ về tỷ suất sinh thô, tỷ suất chết thô và tỷ lệ gia tăng tự nhiên giữa các vùng ở Việt Nam?
- Mục Tiêu: Nhận xét về xu hướng dân số của các Vùng
- Thực hiện: Dùng Boxplot và Linechart để vẽ biểu đồ của các vùng:
    - Bắc Trung Bộ Và Duyên Hải Miền Trung
    - Trung Du Và Miền Núi Phía Bắc
    - Đông Nam Bộ
    - Đồng Bằng Sông Cửu Long
    - Đồng Bằng Sông Hồng
    - Tây Nguyên.

Câu Hỏi 3



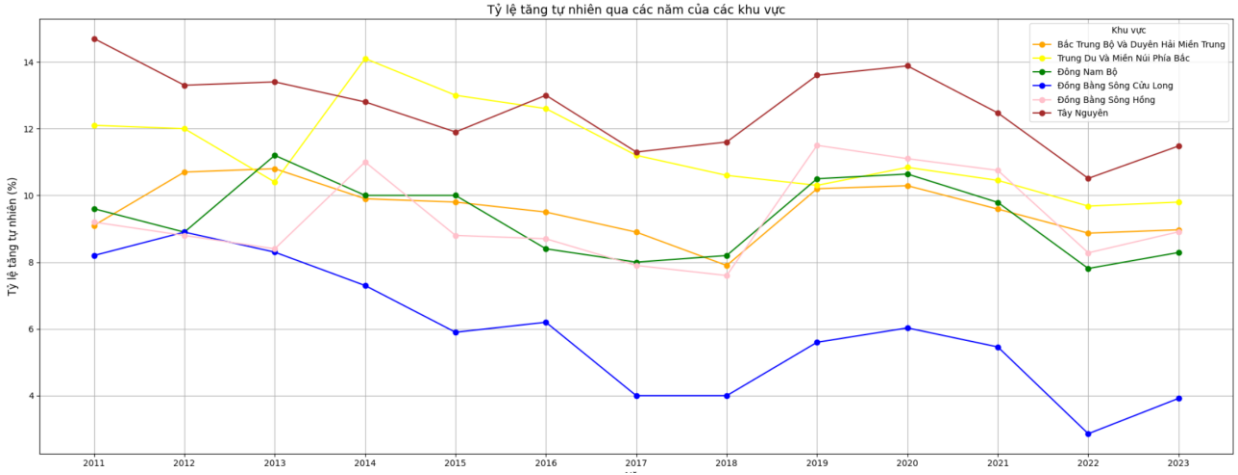
- Tây Nguyên và Trung Du và Miền Núi Phía Bắc có tỷ suất sinh thô cao hơn, có thể do đặc điểm kinh tế - xã hội hoặc văn hóa.
- Đông Nam Bộ và Đồng Bằng Sông Cửu Long có tỷ suất sinh thô thấp, điều này có thể liên quan đến mức độ đô thị hóa và sự phát triển kinh tế.
- Đồng Bằng Sông Cửu Long và Trung Du và Miền Núi Phía Bắc có khoảng phân tán lớn nhất, thể hiện sự chênh lệch đáng kể qua các năm của các địa phương trong vùng.
- Bắc Trung Bộ và Duyên Hải Miền Trung có khoảng phân tán nhỏ nhất, cho thấy sự đồng đều hơn trong tỷ suất sinh thô qua các năm của các địa phương trong vùng.

Câu Hỏi 3



- Bắc Trung Bộ và Duyên Hải Miền Trung và Trung Du và Miền Núi Phía Bắc có tỷ suất chết cao hơn, có thể liên quan đến các yếu tố như điều kiện tự nhiên khắc nghiệt hoặc chất lượng dịch vụ y tế.
- Đông Nam Bộ có độ phân tán lớn nhất, cho thấy sự biến động giữa các tỉnh trong vùng qua các năm.
- Tây Nguyên và Đông Nam Bộ có tỷ suất chết thấp hơn, phản ánh điều kiện y tế tốt hơn hoặc cấu trúc dân số trẻ hơn.

Câu Hỏi 3



Xu hướng chung:

- Tỷ lệ tăng tự nhiên ở các khu vực có xu hướng giảm dần qua các năm.
- Một số khu vực có sự biến động lớn ở một số thời điểm, nhưng xu hướng dài hạn vẫn là giảm.

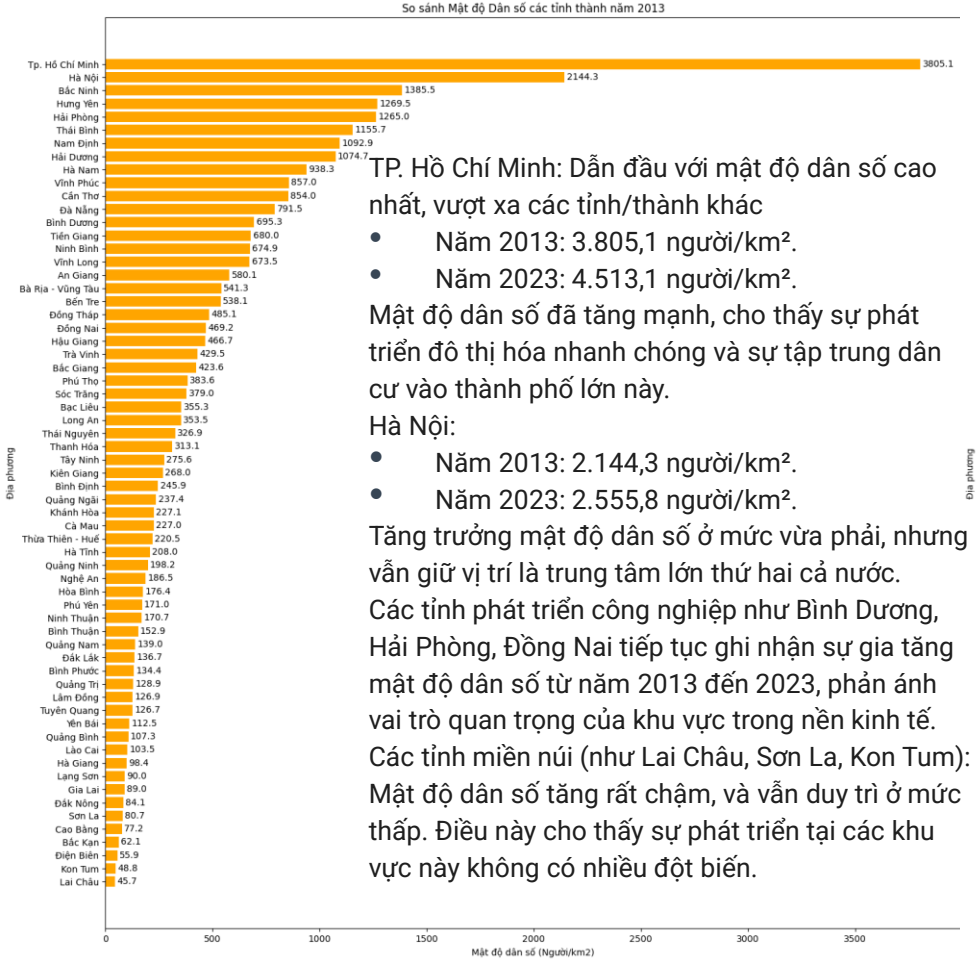
Phân tích từng khu vực:

- Tây Nguyên: Khu vực này có tỷ lệ tăng tự nhiên cao nhất so với các khu vực khác trong hầu hết các năm. Tuy nhiên, tỷ lệ này cũng giảm rõ rệt sau năm 2020.
- Đồng bằng Sông Cửu Long: Đây là khu vực có tỷ lệ tăng tự nhiên thấp nhất, với sự giảm đều từ 2012-2018 và tăng nhẹ từ 2018-2020 sau đó tiếp tục giảm mạnh.
- Đồng bằng Sông Hồng, Bắc Trung Bộ và Duyên Hải Miền Trung, Đông Nam Bộ: Biến động tương đối ổn định, nhưng xu hướng giảm nhẹ.
- Trung du và miền núi phía Bắc: Tỷ lệ tăng tự nhiên tăng mạnh từ 2013-2014 sau đó giảm dần qua từng năm.

Câu Hỏi 4

- Câu Hỏi: So sánh mật độ dân số và các tỷ lệ của 63 tỉnh thành giữa năm 2013 và 2023(1 thập kỷ)?
- Mục Tiêu: Nhận xét về thay đổi dân số như thế nào sau 1 thập kỷ của 63 tỉnh thành.
- Thực hiện:
  - Dùng Barchart để biểu diễn mật độ dân số của năm 2013 và năm 2023
  - Dùng Groupbar để biểu diễn các tỷ lệ giữa năm 2013 và năm 2023
  - Sử dụng Scatterplot để tìm hệ số tương quan của các trường ở năm 2013 và so kiểm tra độ tương quan đó ở năm 2023 có còn tương quan không
  - Dùng Heatmap để xem tổng quan về độ tương quan giữa các cột dữ liệu.

Câu Hỏi 4



TP. Hồ Chí Minh: Dẫn đầu với mật độ dân số cao nhất, vượt xa các tỉnh/thành khác

- Năm 2013: 3.805,1 người/km².
- Năm 2023: 4.513,1 người/km².

Mật độ dân số đã tăng mạnh, cho thấy sự phát triển đô thị hóa nhanh chóng và sự tập trung dân cư vào thành phố lớn này.

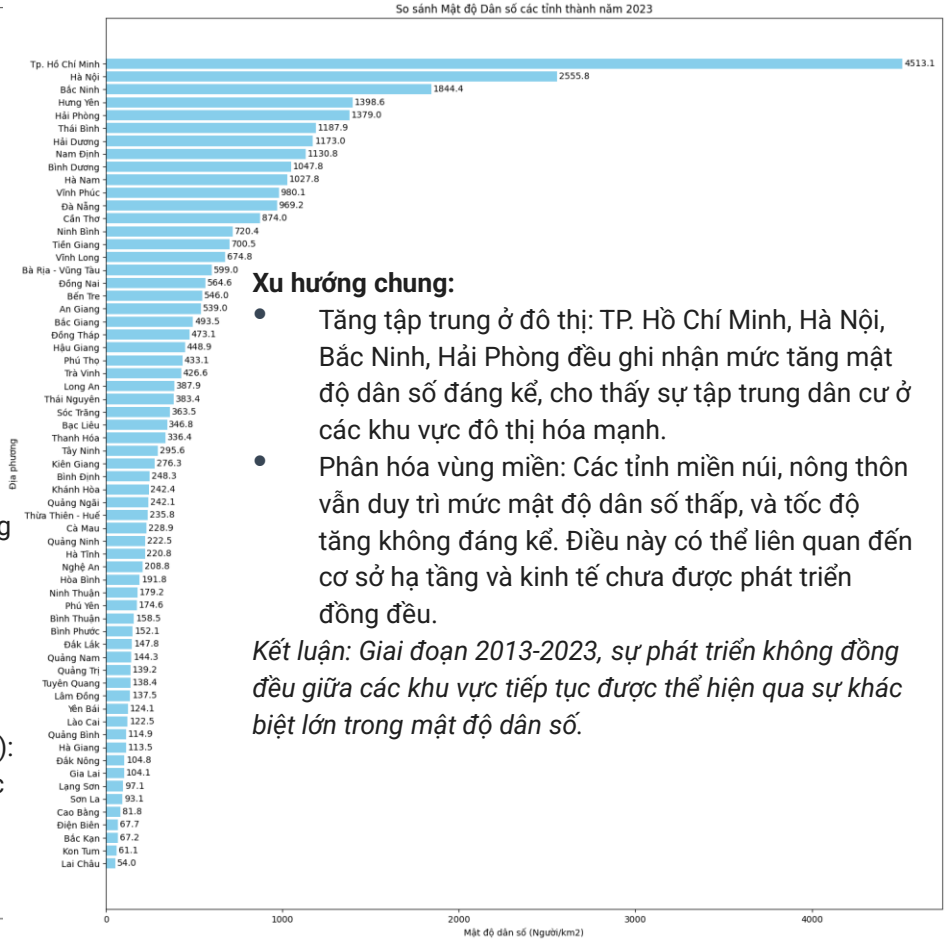
Hà Nội:

- Năm 2013: 2.144,3 người/km².
- Năm 2023: 2.555,8 người/km².

Tăng trưởng mật độ dân số ở mức vừa phải, nhưng vẫn giữ vị trí là trung tâm lớn thứ hai cả nước.

Các tỉnh phát triển công nghiệp như Bình Dương, Hải Phòng, Đồng Nai tiếp tục ghi nhận sự gia tăng mật độ dân số từ năm 2013 đến 2023, phản ánh vai trò quan trọng của khu vực trong nền kinh tế.

Các tỉnh miền núi (như Lai Châu, Sơn La, Kon Tum): Mật độ dân số tăng rất chậm, và vẫn duy trì ở mức thấp. Điều này cho thấy sự phát triển tại các khu vực này không có nhiều đột biến.



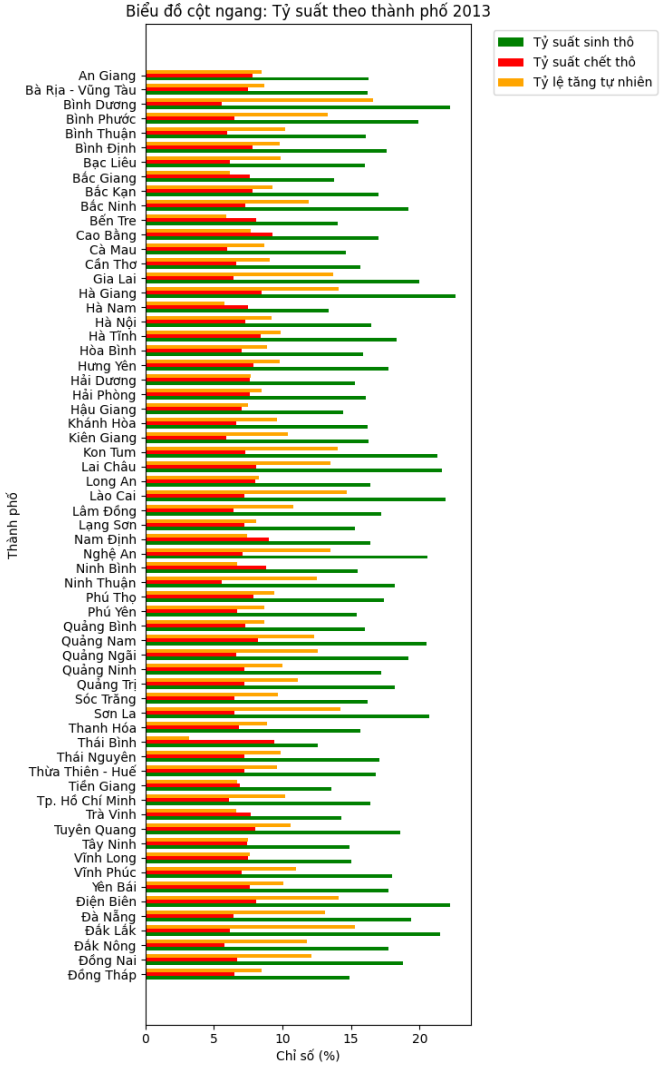
Xu hướng chung:

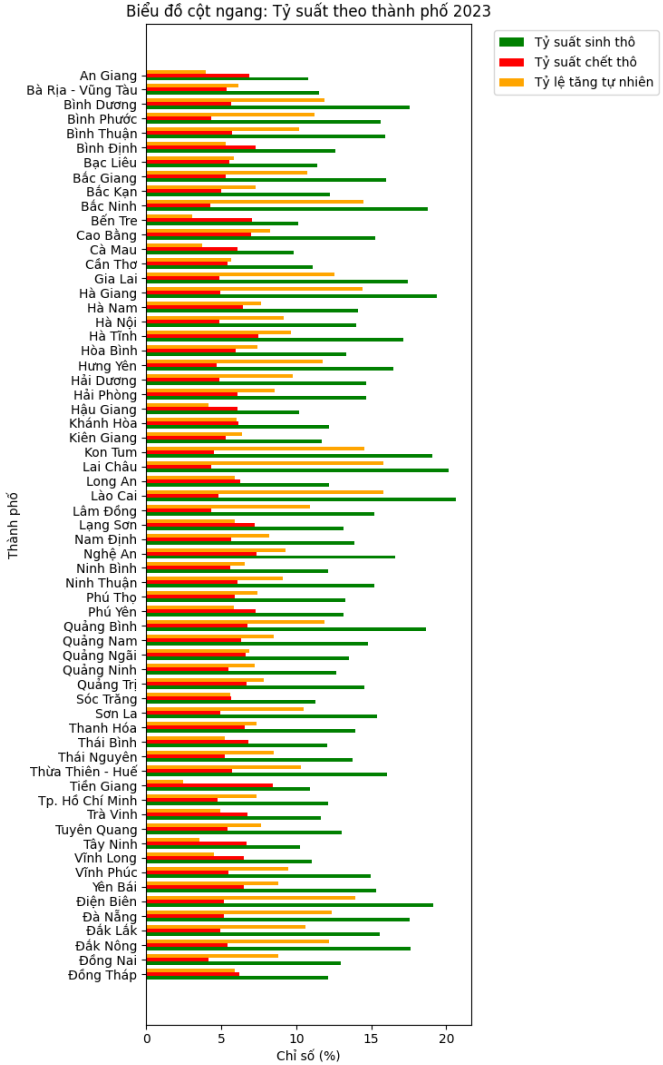
- Tăng tập trung ở đô thị: TP. Hồ Chí Minh, Hà Nội, Bắc Ninh, Hải Phòng đều ghi nhận mức tăng mật độ dân số đáng kể, cho thấy sự tập trung dân cư ở các khu vực đô thị hóa mạnh.
- Phân hóa vùng miền: Các tỉnh miền núi, nông thôn vẫn duy trì mức mật độ dân số thấp, và tốc độ tăng không đáng kể. Điều này có thể liên quan đến cơ sở hạ tầng và kinh tế chưa được phát triển đồng đều.

Kết luận: Giai đoạn 2013-2023, sự phát triển không đồng đều giữa các khu vực tiếp tục được thể hiện qua sự khác biệt lớn trong mật độ dân số.

Câu Hỏi 4

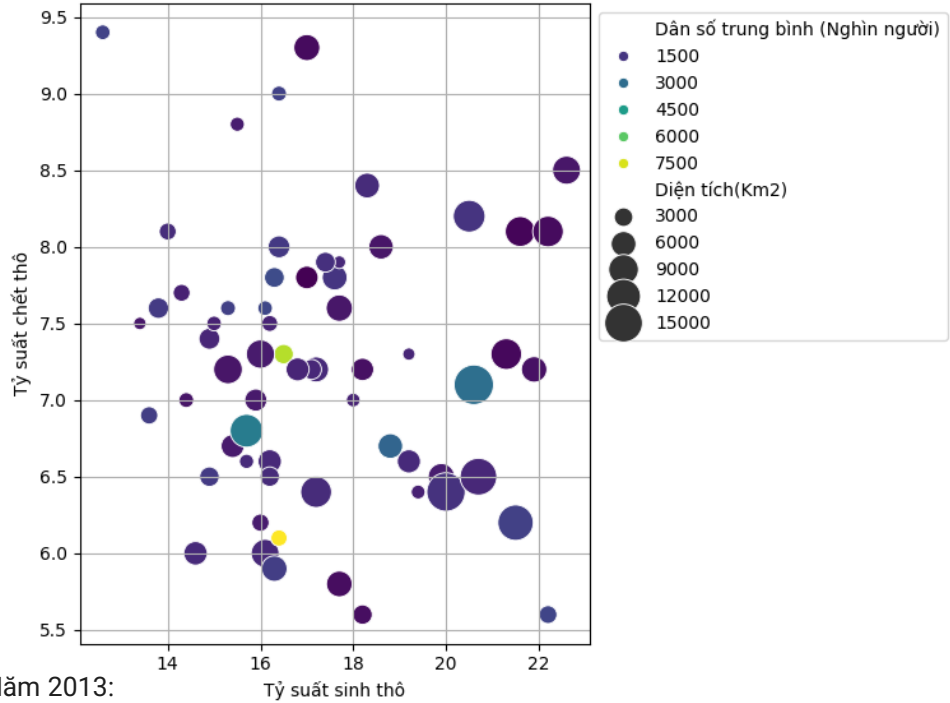
- Hầu hết các tỉnh có tỷ suất sinh thô giảm. Điều này có thể phản ánh sự thay đổi trong mô hình sinh sản, với xu hướng giảm sinh ở các gia đình.
- Tỷ suất chết thô ở hầu hết các tỉnh thành phố: không thay đổi nhiều, thậm chí có xu hướng giảm nhẹ ở một số địa phương, nhờ cải thiện hệ thống y tế, tuổi thọ gia tăng và giảm tỷ lệ tử vong do bệnh tật.Các tỉnh có điều kiện kinh tế kém phát triển hoặc dân cư lớn tuổi có thể có tỷ suất chết thô cao hơn.
- Tỷ lệ gia tăng tự nhiên giảm, điều này do sự giảm mạnh tỷ suất sinh thô mà không có sự thay đổi đáng kể trong tỷ suất chết thô. Một số tỉnh, đặc biệt là khu vực đô thị hóa, có thể đã bước vào giai đoạn tăng trưởng dân số thấp hoặc thậm chí dân số suy giảm tự nhiên (tỷ lệ gia tăng tự nhiên âm).



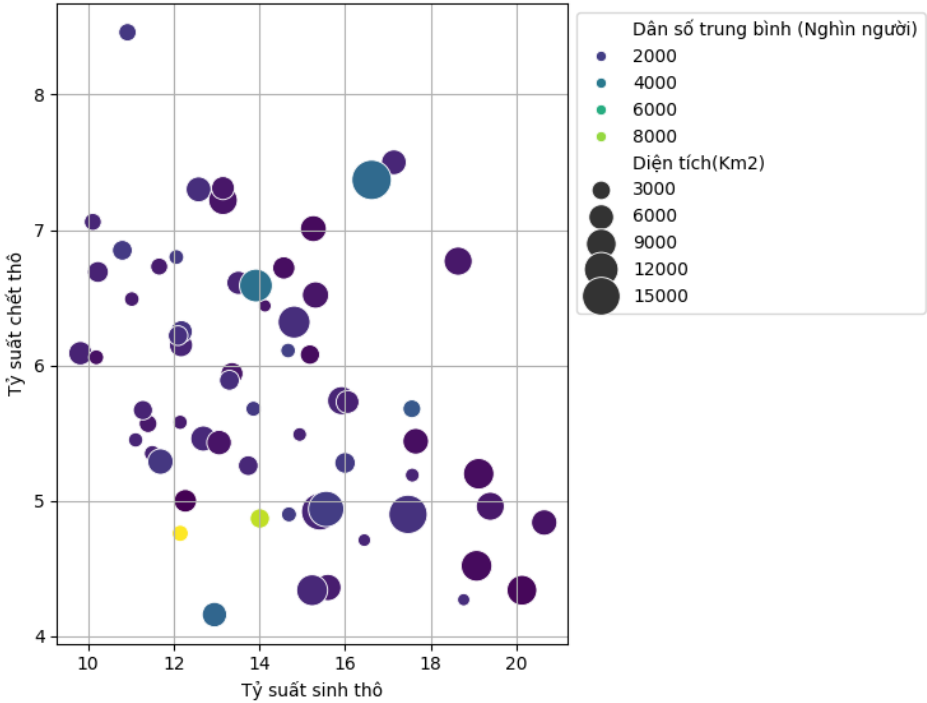


Câu Hỏi 4

Scatterplot giữa Tỷ suất sinh thô và Tỷ suất chết thô (Năm 2013)



Scatterplot giữa Tỷ suất sinh thô và Tỷ suất chết thô (Năm 2023)



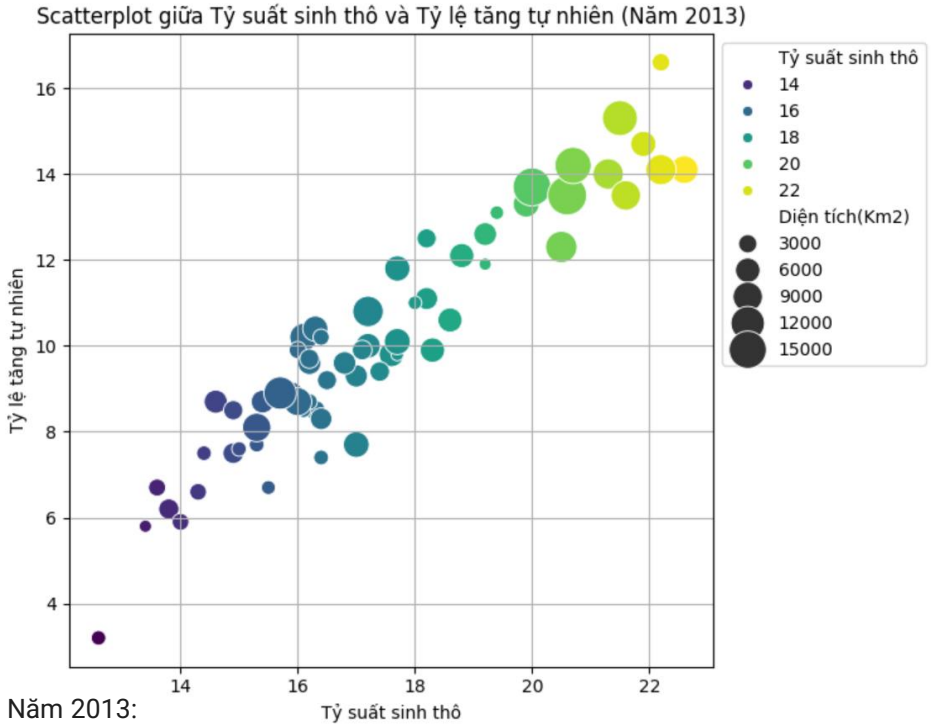
Năm 2013:

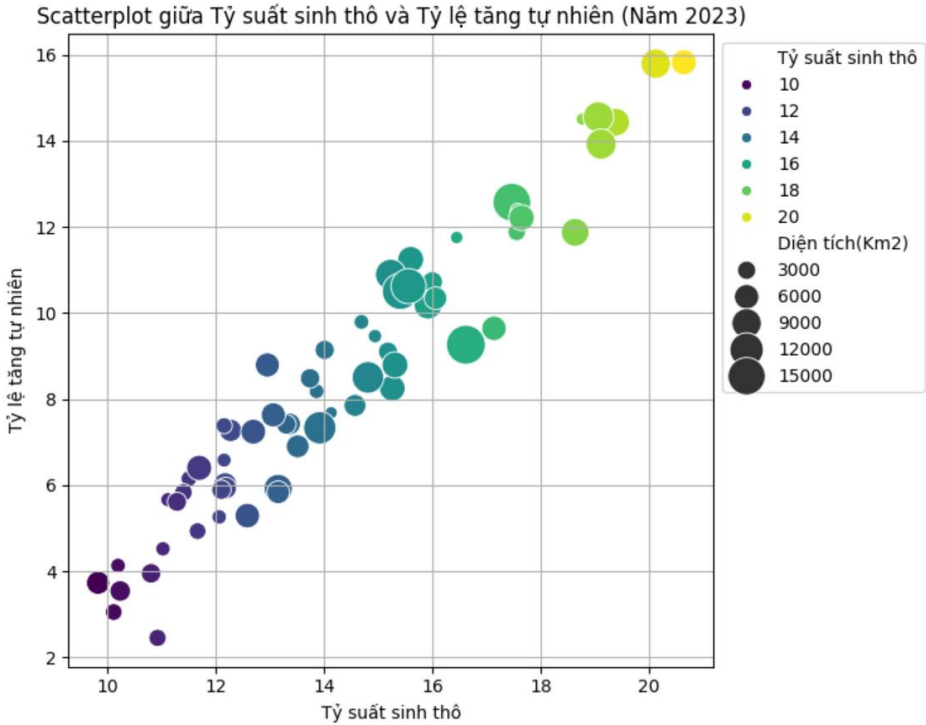
- Tương quan giữa tỷ suất sinh thô và tỷ suất chết thô có vẻ không rõ ràng, thể hiện qua sự phân tán các điểm dữ liệu. Các khu vực có tỷ suất sinh thô cao không nhất thiết đi kèm với tỷ suất chết thô cao và ngược lại.

Năm 2023:

- Tỷ suất sinh thô và tỷ suất chết thô có vẻ tương quan âm nhẹ, tức là ở các tỉnh/thành có tỷ suất sinh thô cao hơn, tỷ suất chết thô cũng có xu hướng thấp hơn. Điều này có thể phản ánh các khu vực có tỷ suất sinh thô cao thì điều kiện y tế đã được cải thiện rõ rệt dẫn đến tỷ suất chết thô thấp. Xu hướng này có thể phản ánh sự cải thiện trong y tế và chất lượng sống tại một số khu vực.

Câu Hỏi 4





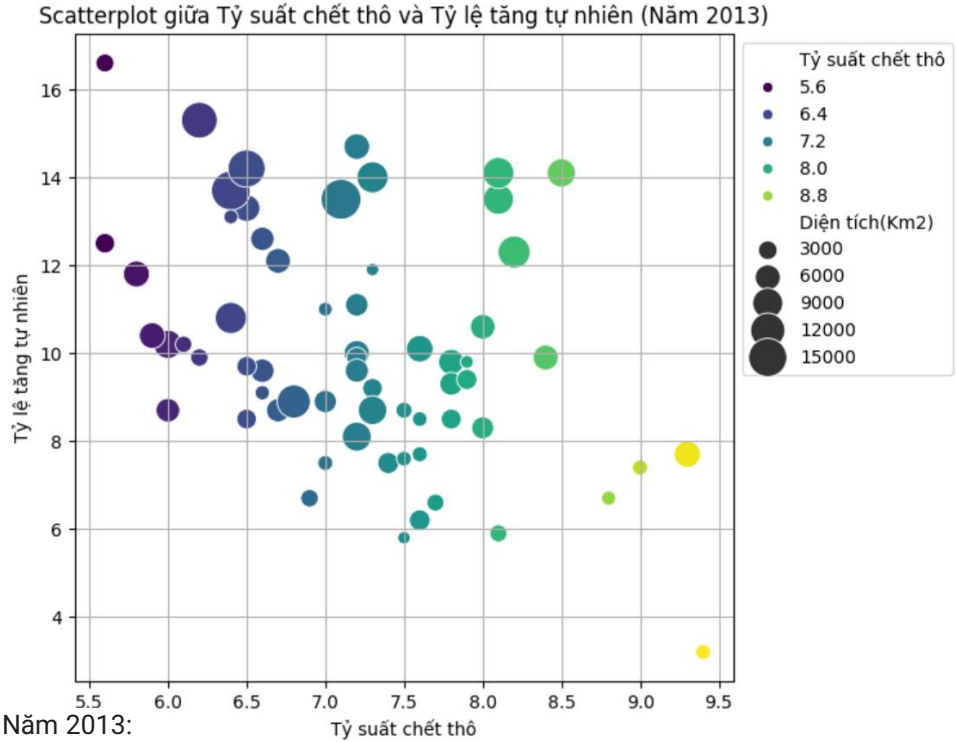
Năm 2013:

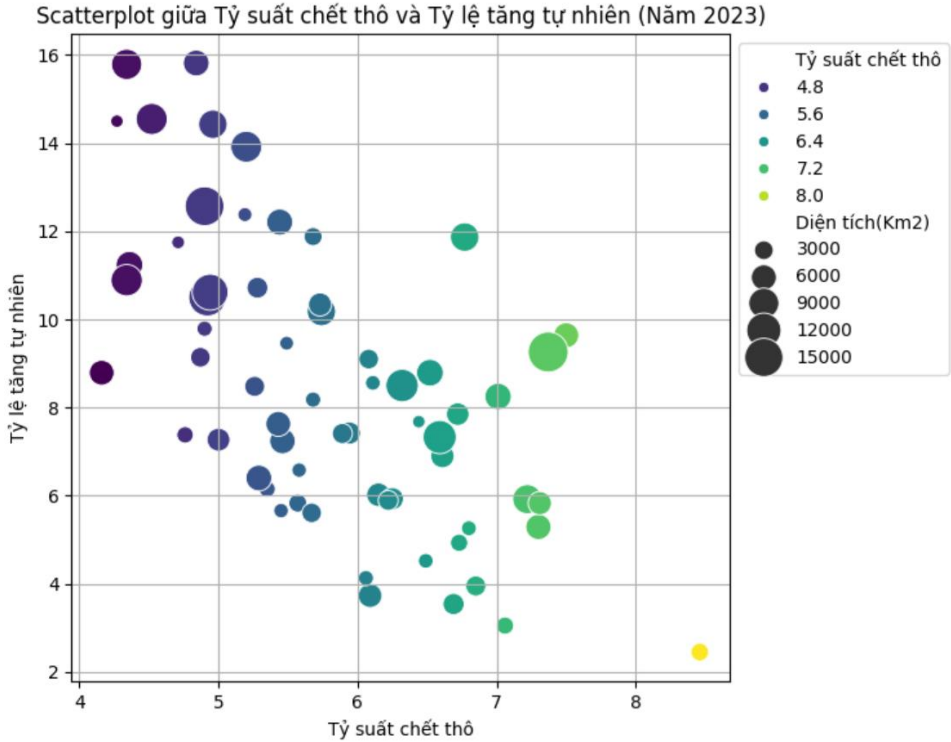
- Có một tương quan tuyến tính mạnh mẽ giữa tỷ suất sinh thô và tỷ lệ tăng tự nhiên. Các khu vực có tỷ suất sinh thô cao thường có tỷ lệ tăng tự nhiên cao hơn, Điểm dữ liệu tập trung trong khoảng tỷ suất sinh từ 14-22 và tỷ lệ tăng tự nhiên từ 6-16.

Năm 2023:

- Mối quan hệ giữa hai chỉ số này vẫn duy trì tương quan tuyến tính, nhưng phân bố dữ liệu đã thay đổi. Tỷ suất sinh thô đã giảm đáng kể ở hầu hết các khu vực, khiến tỷ lệ tăng tự nhiên cũng giảm tương ứng. Dữ liệu tập trung trong khoảng tỷ suất sinh từ 10-18 và tỷ lệ tăng tự nhiên từ 2-12, cho thấy xu hướng giảm sinh ảnh hưởng rõ rệt đến tăng tự nhiên.

Câu Hỏi 4





Năm 2013:

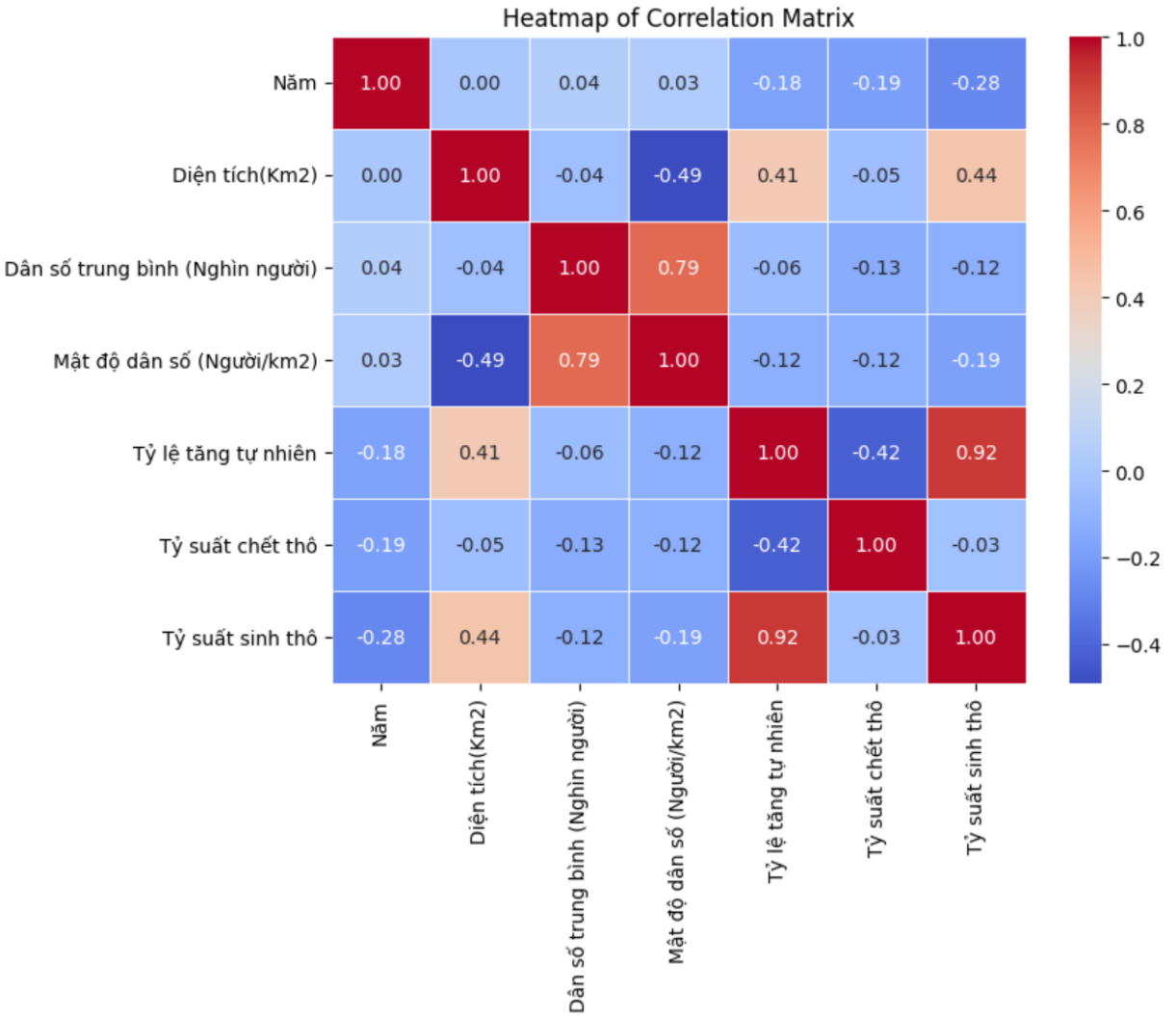
- Mối quan hệ giữa tỷ suất chết thô và tỷ lệ tăng tự nhiên có xu hướng nghịch biến: các khu vực có tỷ suất chết thô cao thường có tỷ lệ tăng tự nhiên thấp hơn. Các điểm dữ liệu tập trung trong khoảng tỷ suất chết thô từ 6-9 và tỷ lệ tăng tự nhiên từ 6-16. Một số khu vực có tỷ lệ tăng tự nhiên cao bất chấp tỷ suất chết thô cao, điều này có thể do tỷ suất sinh thô tại các khu vực này đủ lớn để bù đắp tỷ suất chết.

Năm 2023:

- Xu hướng nghịch biến vẫn được duy trì. Tỷ suất chết thô giảm ở nhiều khu vực, phổ biến trong khoảng 4-7. Tỷ lệ tăng tự nhiên giảm, tập trung trong khoảng 2-12.

Câu Hỏi 4

- Tương quan mạnh:
  - Tỷ lệ tăng tự nhiên có mối tương quan mạnh với tỷ suất sinh thô (hệ số tương quan 0.92). Điều này cho thấy tỷ lệ tăng tự nhiên phụ thuộc chặt chẽ vào tỷ suất sinh thô.
  - Mật độ dân số có mối tương quan cao với dân số trung bình (hệ số tương quan 0.79). Điều này hợp lý vì dân số trung bình càng cao, mật độ dân số cũng tăng.



Câu Hỏi 5

- Câu Hỏi: Phân tích về xu hướng dân số của 63 tỉnh thành.
- Mục Tiêu: Tìm ra về xu hướng dân số cả cả nước, về mật độ dân số tập trung nhiều ở khu vực nào thông qua bản đồ địa lý Việt Nam.
- Thực hiện:
  - Dùng Choropleth để map giá trị mật độ dân số lên bản đồ địa lý Việt Nam

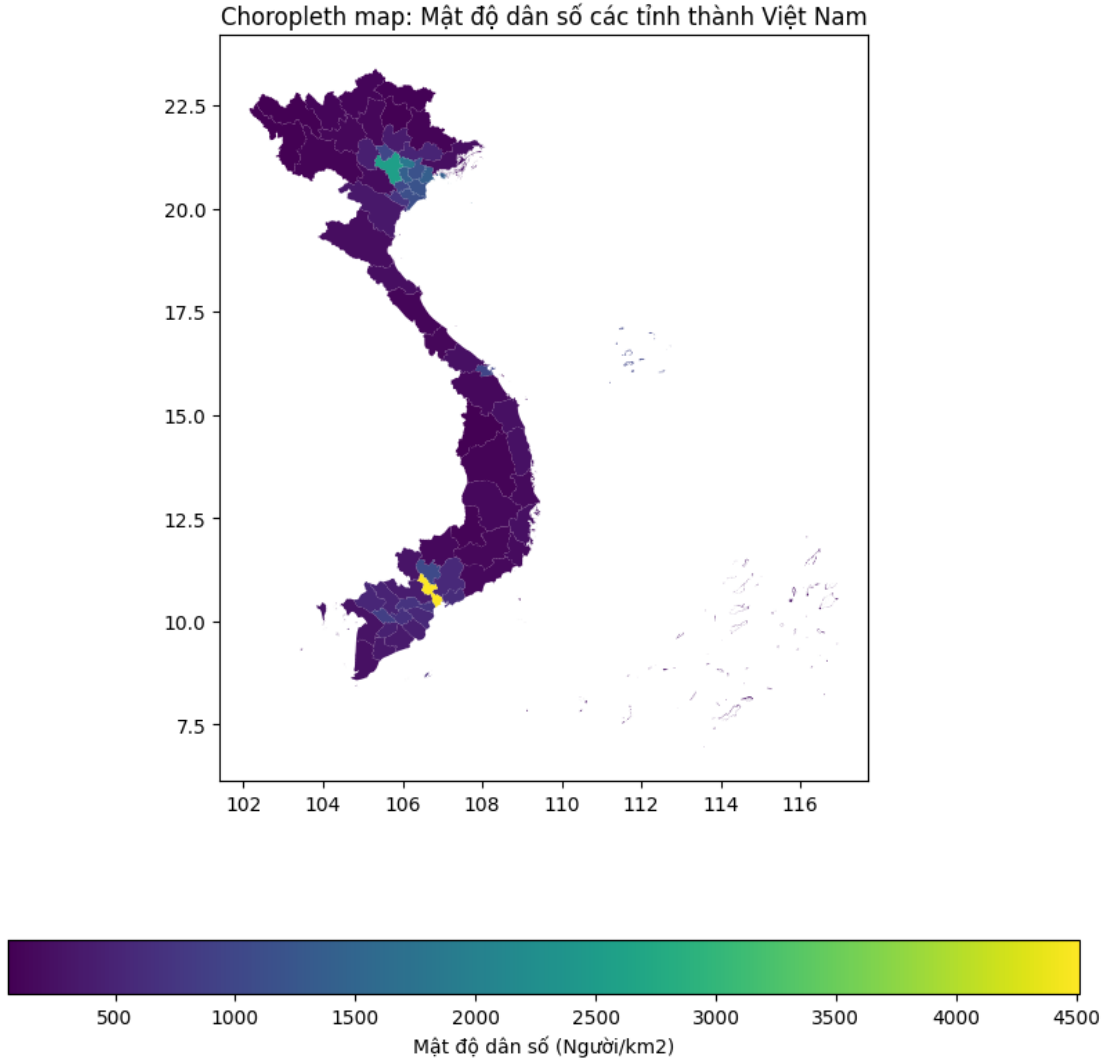
Câu Hỏi 5

Ta thấy rằng mật độ dân số của TP Hồ Chí Minh rất cao(>4500). Khu vực xung quanh TP Hồ Chí Minh cũng có mật độ dân số cao hơn cho thấy dân số có xu hướng tập trung vào các đô thị lớn.

Ở phía Bắc, TP Hà Nội cũng có mật độ dân số cao nhất khu vực(>2500). Các tỉnh thành phố xung quanh TP Hà Nội cũng có mật độ dân số khá cao, điều này khẳng định rõ việc dân số đang tập trung vào các vùng đô thị trọng điểm của cả nước.

Kết luận:

- Quy hoạch đô thị: Các khu vực có mật độ dân số cao cần được tập trung phát triển cơ sở hạ tầng, dịch vụ công cộng, và giải pháp giảm tải dân số (như phát triển vùng vệ tinh).
- Phát triển vùng: Các khu vực có mật độ thấp có tiềm năng phát triển đất đai và mở rộng dân số thông qua các chính sách khuyến khích di cư hoặc đầu tư vào hạ tầng.



Áp Dụng Mô Hình

Yêu Cầu: Dự đoán Tỷ Lệ Gia Tăng Tự Nhiên trong tương lai(2024-2030).

Mục tiêu:

- Áp dụng các loại mô hình khác nhau để so sánh độ hiệu quả giữa chúng.
- Tìm được xu hướng của dân số trong tương lai
- Tìm ra lý do để giải thích về xu hướng dân số qua đó có các giải pháp phù hợp với xu hướng này.

Áp Dụng Mô Hình

Hồi Quy Tuyến Tính OLS:

Chọn ra biến độc lập gồm ‘Năm’ và ‘Tỷ suất chết thô’, huấn luyện bằng tập dữ liệu của cả nước và kỳ vọng ở biến ‘Tỷ lệ gia tăng tự nhiên’

Dữ liệu trong tương lai gồm

- Năm: Từ 2024-2030
- Tỷ suất chết thô: Giả sử mỗi năm tỷ suất chết thô đều giảm(do y tế phát triển và điều kiện lý tưởng là không có dịch bệnh, thiên tai,…)

Kết quả dự đoán:

- Tỷ lệ gia tăng tự nhiên vẫn dương nhưng có xu hướng giảm dần theo hàm tuyến tính
- Y = 557.4623 - 0.2056 * Năm – 1.8931 * Tỷ suất chết thô

Đánh giá mô hình:

- Với biến “Năm” có p=0.047 thấy rằng có ý nghĩa thống kê 5%. Phản ánh xu hướng giảm dần
- Với R^2 = 0.352, mô hình giải thích được 35.2% biến phụ thuộc.

Kết luận:

- Vì số lượng dữ liệu hạn chế cũng như tỷ lệ tăng tự nhiên còn phụ thuộc vào nhiều yếu tố khác

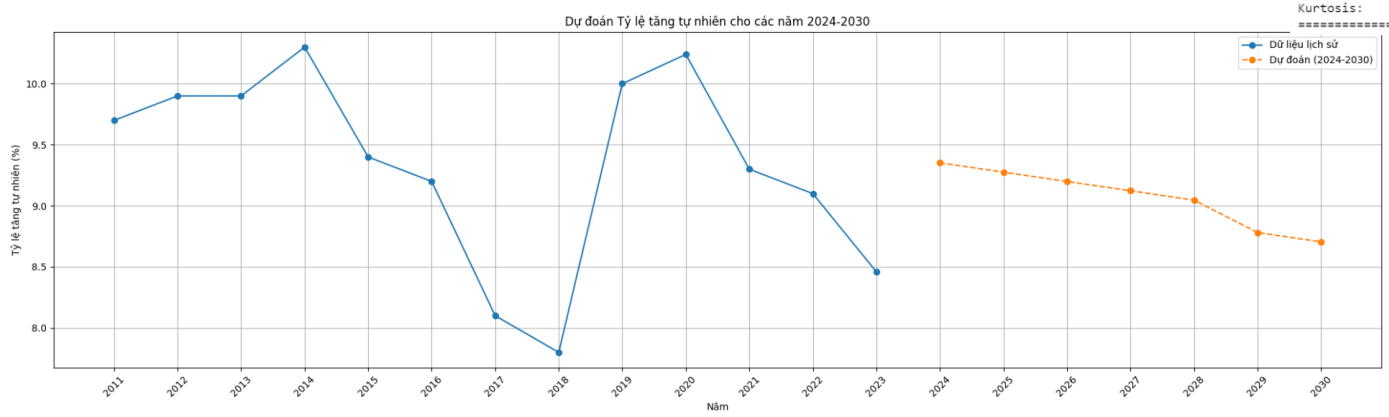
do đó không thể sử dụng hàm tuyến tính để dự đoán tỷ lệ tăng tự nhiên trong tương lai được

OLS Regression Results

Dep. Variable:	Tỷ lệ tăng tự nhiên	R-squared:	0.352
Model:	OLS	Adj. R-squared:	0.223
Method:	Least Squares	F-statistic:	2.719
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.114
Time:	08:18:51	Log-Likelihood:	-12.210
No. Observations:	13	AIC:	30.42
Df Residuals:	10	BIC:	32.12
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	557.4623	242.220	2.301	0.044	17.761	1097.163
Năm	-0.2656	0.117	-2.270	0.047	-0.526	-0.005
Tỷ suất chết thô	-1.8931	1.057	-1.791	0.104	-4.248	0.462

Omnibus:	3.084	Durbin-Watson:	1.193
Prob(Omnibus):	0.214	Jarque-Bera (JB):	1.249
Skew:	-0.324	Prob(JB):	0.536
Kurtosis:	1.627	Cond. No.	2.50e+06



Năm	Tỷ suất chết thô	Tỷ lệ gia tăng tự nhiên(Dự đoán)
2024	5.6	9.352182
2025	5.5	9.275922
2026	5.4	9.199662
2027	5.3	9.123403
2028	5.2	9.047143
2029	5.2	8.781576
2030	5.1	8.705316

Áp Dụng Mô Hình

Gradient Boosting Regressor:

Là mô hình dựa trên cây quyết định để giải các bài toán hồi quy, huấn luyện và kiểm tra bằng tập dữ liệu chung(bao gồm tất cả các tỉnh, vùng và cả nước) với tỷ lệ (8:2).

Tham số sử dụng

- learning_rate: 0.3 max_depth: 3 n_estimators: 200

Dữ liệu trong tương lai gồm

- Năm: Từ 2024-2030
- Tỷ suất chết thô: Giả sử mỗi năm tỷ suất chết thô đều giảm(do y tế phát triển và điều kiện lý tưởng là không có dịch bệnh, thiên tai,…)

Kết quả dự đoán:

- Tỷ lệ tăng tự nhiên có xu hướng tăng giảm nhẹ ổn định từ 2024-2027 sau đó tăng mạnh vào 2028 sau đó ổn định và giảm vào 2030.

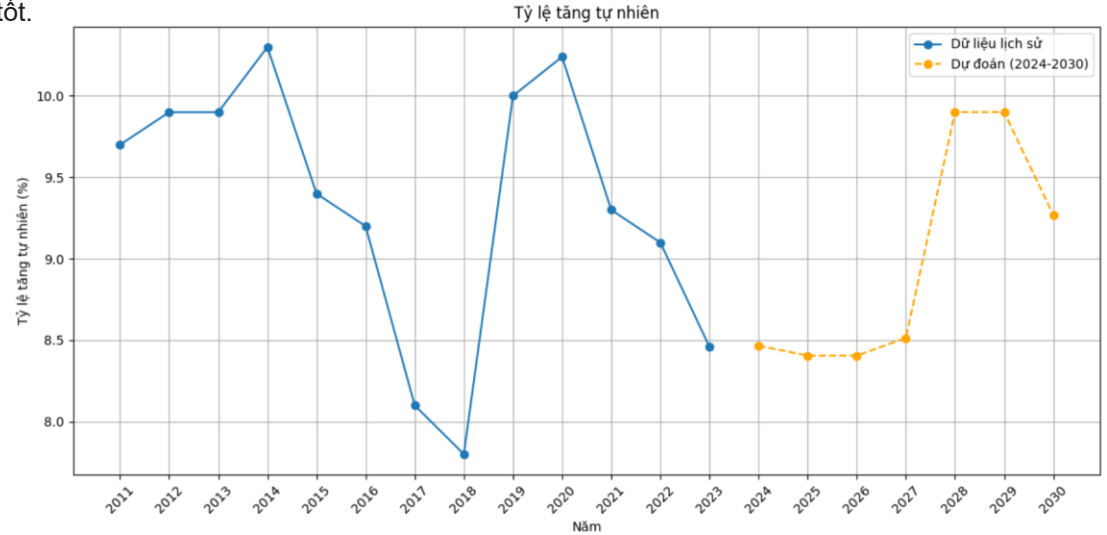
Đánh giá mô hình:

- MSE: 2.14, sai số thấp cho thấy mô hình hoạt động tương đối tốt.
- Với R^2 = 0.79, mô hình giải thích được 79% biến phụ thuộc.

Kết luận:

- Mô hình hoạt động tương đối hiệu quả.

	Năm	Tỷ suất chết thô	Tỷ lệ gia tăng tự nhiên(Dự đoán)
0	2024	5.6	8.466303
1	2025	5.5	8.405252
2	2026	5.4	8.405252
3	2027	5.3	8.513837
4	2028	5.2	9.900669
5	2029	5.2	9.900669
6	2030	5.1	9.269272



05

Thực Nghiệm

Tiến Độ Thực Hiện

STT	Công Việc	Tiến Độ
1	Thu thập dữ liệu từ trang https://www.gso.gov.vn/	Hoàn Thành
2	Tiền xử lý dữ liệu	Hoàn Thành
3	Khám phá dữ liệu	Hoàn Thành
3.1	Mô tả dữ liệu	Hoàn Thành
3.2	Đặt câu hỏi cho dữ liệu	Hoàn Thành
3.3	Trực quan hóa dữ liệu	Hoàn Thành
3.4	Rút ra ý nghĩa để trả lời câu hỏi	Hoàn Thành
4	Phân tích dữ liệu theo mô hình	Hoàn Thành
5	Slide thuyết trình và các tài liệu liên quan	Hoàn Thành

06

Kết Luận

Khó khăn gặp phải trong quá trình thực hiện đồ án?

Trong quá trình thực hiện đồ án không tránh khỏi những khó khăn, trong đó việc thực hiện đồ án một mình là thách thức lớn nhất. Đặc biệt là trong việc quản lý thời gian và phân bổ công việc hợp lý. Vì đồ án yêu cầu sự kết hợp giữa lý thuyết và thực hành, việc tự nghiên cứu và tìm ra giải pháp cho các vấn đề gặp phải là điều không hề dễ dàng. Ngoài ra, đôi khi gặp phải những vấn đề kỹ thuật phức tạp, chẳng hạn như việc thiếu dữ liệu, tối ưu mã nguồn, xử lý lỗi,… cũng đòi hỏi nhiều thời gian và công sức để giải quyết.

Những gì đã học được trong lúc làm đồ án?

Trong quá trình thực hiện đồ án, em đã học được rất nhiều điều quý giá. Đầu tiên, em hiểu rõ hơn về các quy trình trong lĩnh vực Khoa Học Dữ Liệu, từ thu thập dữ liệu, tiền xử lý dữ liệu, trực quan hoá dữ liệu, xây dựng mô hình,… Qua đó có cái nhìn tổng quát hơn về lĩnh vực này.

Ngoài ra, đồ án này giúp em học được cách làm việc từ nghiên cứu đến tìm kiếm giải pháp, cách quản lý thời gian và rèn luyện tư duy logic và khả năng giải quyết vấn đề từ đó giúp em tự tin hơn trong việc đối mặt với những thử thách trong công việc.

Nếu có thêm thời gian, nhóm sẽ làm gì?

Trong quá trình thực hiện đồ án, nhóm nhận thấy các điểm có thể cải thiện gồm:

- Bộ dữ liệu: Có thể mở rộng thêm từ những bộ dữ liệu liên quan trên trang Tổng Cục Thống Kê, từ đó có thêm nhiều trường dữ liệu để tìm hiểu và xây dựng mô hình hiệu quả hơn.
- Nhóm có tìm hiểu và thử xây dựng các mô hình ARIMA và KNN để dự đoán tuy nhiên vì không đạt được hiệu quả như mong đợi nên đã bỏ qua. Nếu có thêm thời gian, nhóm mong muốn có thể tìm hiểu thêm các mô hình khác để so sánh và tìm ra mô hình phù hợp nhất để giải quyết bài toán dân số trong tương lai.

Thank You!