**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**□&□**

A picture containing logo

Description automatically generated

**Học phần:** Trực quan hóa dữ liệu **| Lớp:** 19\_21

**Giáo viên hướng dẫn:** Bùi Tiến Lên

**Sinh viên thực hiện:**

* Dương Thanh Hiệp - 19120505
* Lê Minh Hữu - 19120525
* Nguyễn Nhật Huy - 19120528
* Nguyễn Tường Khải - 19120538
* Hồ Công Lượng - 19120572

**ĐỒ ÁN CUỐI KỲ**

**PHÂN TÍCH DỮ LIỆU**

**VIETNAM POPULATION STATUS**

**Hồ Chí Minh, ngày 24 tháng 6 năm 2022**

**MỤC LỤC**

[**I.** **Thông tin chung** 3](#_Toc106974990)

[**II.** **Phân tích dữ liệu** 4](#_Toc106974991)

[**1.** **Thống kê mô tả** 4](#_Toc106974992)

[**2.** **Phân tích dữ liệu đơn giản** 4](#_Toc106974993)

[**3.** **Phân tích hồi quy, dự đoán** 6](#_Toc106974994)

[**III.** **Trực quan** 7](#_Toc106974995)

[**1.** **Biểu đồ dân số các vùng/tỉnh** 7](#_Toc106974996)

[**2.** **Biểu độ mật độ dân số các tỉnh** 8](#_Toc106974997)

[**3.** **Biểu đồ mật độ dân số các vùng** 8](#_Toc106974998)

[**4.** **Biểu đồ tỉ lệ gia tăng dân số của các vùng** 9](#_Toc106974999)

[**5.** **Biểu đồ tỉ lệ gia tăng dân số của các tỉnh thành** 10](#_Toc106975000)

[**6.** **Biểu đồ tỉ lệ giới tính trung bình của các vùng** 10](#_Toc106975001)

[**7.** **Dân số của từng vùng qua các năm** 11](#_Toc106975002)

[**8.** **Biểu đồ thể hiện sự thay đổi mật độ dân số các vùng qua các năm.** 12](#_Toc106975003)

[**9.** **Biểu đồ tỉ lệ giới tính ở các tỉnh thuộc khu vực Tây Nguyên qua từng năm.** 12](#_Toc106975004)

[**10.** **Tổng lực lượng lao động của các vùng theo năm.** 13](#_Toc106975005)

[**11.** **Tỉ lệ dân số và tỉ lệ lực lượng lao động của từng vùng** 14](#_Toc106975006)

[**12.** **Biểu đồ thể hiện tỉ lệ gia tăng dân số các tỉnh vùng Đông Nam Bộ qua các năm.** 15](#_Toc106975007)

[**13.** **Tình hình dân số, tỷ lệ lao động của thành phố Hồ Chí Minh qua các năm.** 15](#_Toc106975008)

[**14.** **Dân số và lực lượng lao động của từng vùng** 16](#_Toc106975009)

[**15.** **Biểu đồ tương quan giữa mật độ dân số và dân số trung bình của các tỉnh** 17](#_Toc106975010)

[**16.** **Biểu đồ mật độ dân số, tỷ lệ gia tăng dân số và dân số trung bình của các vùng trong năm 2020** 17](#_Toc106975011)

[**IV.** **Tài liệu tham khảo** 18](#_Toc106975012)

1. **Thông tin chung**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ và tên** | **Mức độ hoàn thành (%)** |
| 19120505 | Dương Thanh Hiệp | 100% |
| 19120525 | Lê Minh Hữu | 100% |
| 19120528 | Nguyễn Nhật Huy | 100% |
| 19120538 | Nguyễn Tường Khải | 100% |
| 19120572 | Hồ Công Lượng | 100% |

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Công Việc** | **Mức độ hoàn thành (%)** |
| 19120505 | Phân tích tổng quan, trực quan câu hỏi số 1, 2, 3 | 100% |
| 19120525 | Phân tích tổng quan, trực quan câu hỏi số 4, 5, 6 | 100% |
| 19120528 | Phân tích tổng quan, trực quan câu hỏi số 7, 8, 9 | 100% |
| 19120538 | Phân tích tổng quan, trực quan câu hỏi số 10, 11, 12 | 100% |
| 19120572 | Phân tích tổng quan, trực quan câu hỏi số 13, 14, 15, 16 | 100% |

1. **Phân tích dữ liệu**
2. **Thống kê mô tả**

Chart

Description automatically generated

A picture containing histogram

Description automatically generated

* **Nhận xét dựa vào biểu đồ boxplot và histogram:**
* Các thuộc tính đều bị lệch về bên phải, các giá trị tập trung hầu ở các giá trị nhỏ, riêng chỉ có vài tỉnh thành là ngoại lệ.
* Mật độ dân số phân bố chủ yếu ở mức 100 - 600 /km^2
* Dân số trung bình phân bố chủ yếu ở 800 - 1500 ngàn người
* Tỷ lệ gia tăng dân số thường ở mức 0.5 - 1 % mỗi năm
* Lực lượng lao động > 15 ở mức 300 - 1000 ngàn người
* Các tỉnh có ngoại lệ so với biểu đồ gồm: Hà Nội, Hồ Chí Minh. Điều nàytương đối dễ hiểu vì đây là 2 TP lớn của Việt Nam.
* Có nhiều dữ liệu ngoại lai ở cả 4 thuộc tính.

1. **Phân tích dữ liệu đơn giản**

Phân bố các tỉnh theo vùng kinh tế - xã hội Việt Nam

Icon

Description automatically generated

* **Nhận xét:**
* Có 6 loại vùng dựa theo vùng kinh tế  xã hội ở Việt Nam: Trung du và miền núi phía Bắc, Bắc Trung bộ và duyên hải miền Trung, Đồng bằng sông Cửu Long, Đồng bằng sông Hồng, Đông Nam Bộ, Tây Nguyên.Có 6 loại vùng dựa theo vùng kinh tếxã hội ở Việt Nam: Trung du và phía Bắc, Bắc Trung bộ và duyên hải miền Trung, Đồng bằng sông Cửu Long, Đồng bằng sông Hồng, Đông Nam Bộ, Tây Nguyên.
* Các vùng kinh tế có nhiều tỉnh thành nhất là: Trung du và miền núi phía Bắc, Bắc Trung bộ và duyên hải miền Trung và Đồng bằng sông Cửu Long.

Chart, pie chart

Description automatically generated

* **Nhận xét:**
* Lực lượng dân cư tập trung đông ở 3 vùng kinh tế trọng điểm của Bắc, Trung, Nam.
* Riêng vùng kinh tế phía Nam: Dân số tập trung đông gần gấp đôi 2 vùng kinh tế Bắc, Trung.
* Dân cư tập trung thưa thớt ở Tây Nguyên và Trung du miền núi phía Bắc.

1. **Phân tích hồi quy, dự đoán**

* **OLS:**
* Dựa vào tập dữ liệu, ta có thể tính được dân số trung bình của Việt Nam tại từng năm từ 2011 – 2020. Nhận thấy dân số Việt Nam tăng đều và ổn định, ta áp dụng Bình Phương Cực Tiểu (OLS) để tìm ra mô hình hồi quy cho dân số Việt Nam theo từng năm.

Chart, scatter chart

Description automatically generated

Mô hình kết quả: Y = **1739.4545** \* T – **3411623**

**Dự đoán:**Dân số Việt Nam năm 2022 sẽ là **105.554.000** người​.

* **Nhận xét:**
* Kết quả của mô hình cho ra hệ số tương quan cao ( mô hình đáng tin cậy), chúng ta có thể dùng mô hình này để dự đoán các khoảng thời gian ngắn.
* Chúng ta cần cân nhắc loại bỏ các outlier khi dân số Việt Nam tăng đột biến so với năm trước đó như năm 2016 – 2017.
* **ARIMA:**

Chart, line chart

Description automatically generated

* **Nhận xét:**
* Mô hình dễ thực thi nhưng đem lại kết quả dự đoán lại rất chênh lệch (MSE của ARIMA lớn hơn mô hình OLS nhiều lần)
* Mô hình không phù hợp với tập dữ liệu quá nhỏ (ARIMA yêu cầu tối thiểu 50 sample)

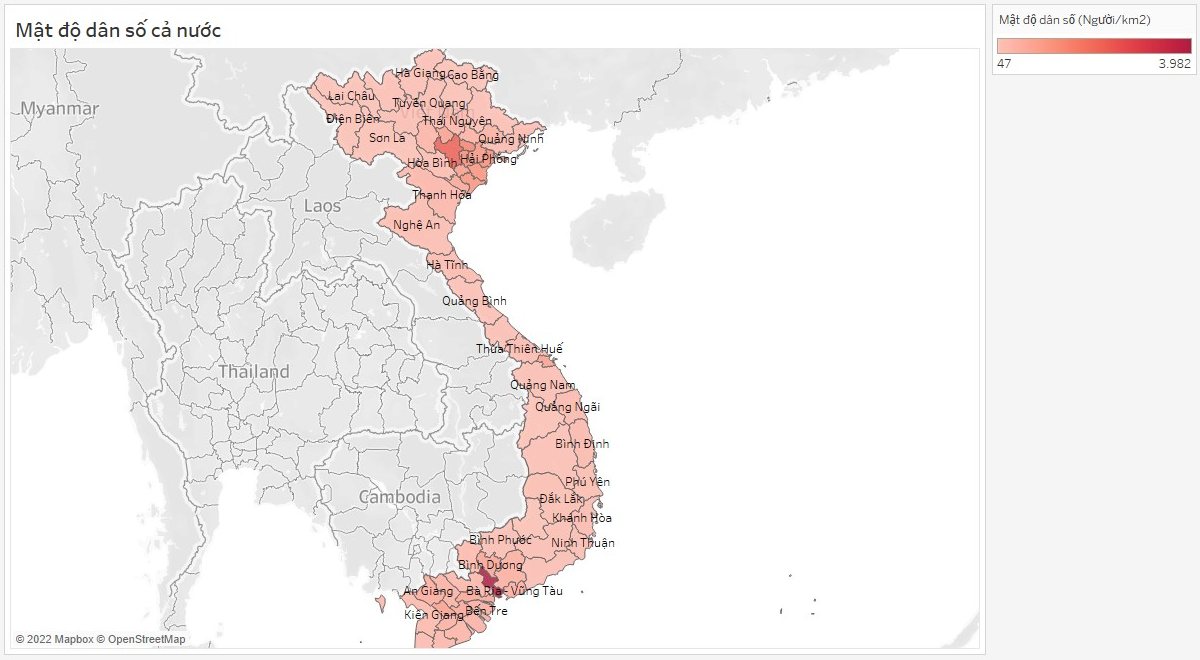
1. **Trực quan**
2. **Biểu đồ dân số các vùng/tỉnh**

Chart, treemap chart

Description automatically generated

* **Nhận xét:**
* Nhìn vào biểu đồ ta có thể thấy Thành phố Hồ Chí Minh và Hà Nội là hai tỉnh có dân số lớn nhất cả nước, và vượt trội so với phần còn lại. Đây là nơi tập trung dân cư đông đúc và là lá cờ đầu trong việc phát triển kinh tế của nước ta.
* Vùng Đồng bằng sông Hồng, Bắc Trung Bộ và duyên hải miền trung là nơi có đông dân cư nhất nước. Vùng Đông Nam Bộ có khá ít tỉnh thành nhưng dân cư ở đây cũng khá đông. Tây nguyên là vùng có ít dân nhất cả nước vì đặt thù về địa lý cũng như địa hình.

1. **Biểu độ mật độ dân số các tỉnh**



* **Nhận xét:**
* Thành phồ Hồ Chí Minh và Hà Nội là những nơi có mật độ dân số rất cao, nổi bật trên cả nước. Các khu vực kế tiếp là Hưng Yên, Bắc Ninh, Hải Phòng…
* Các khu vực còn lại thì có mật độ phân bố dân cư khá đồng đều nhau.
* Nhìn chung, dân cư tập trung chủ yếu ở đồng bằng sông Hồng và Đông Nam Bộ. Bởi vì nơi đây có nhiều khu công nghiệp, có điều kiện sản xuất và sinh sống rất thuận lợi. Điều này đã tạo ra một nguồn lao động dồi dào, thị trường tiêu thụ lớn và tiềm năng phát triển cho các tỉnh ở đây.

1. **Biểu đồ mật độ dân số các vùng**

Chart, bar chart

Description automatically generated

* **Nhận xét:**
* Đồng bằng sông Hồng, Đông Nam Bộ là 2 vùng có mật độ dân số cao nhất cả nước và cao hơn nhiều so với các vùng còn lại vì là nơi tập trung công nghiệp cao nhất cả nước.
* Tây Nguyên là nơi có mật độ dân số thấp nhất cả nước vì đây là vùng cao, dân cư thưa thớt, diện tích lại rất rộng.

1. **Biểu đồ tỉ lệ gia tăng dân số của các vùng**

Chart, bar chart

Description automatically generated

* **Nhận xét:**
* Tất cả các vùng đều có tỉ lệ gia tăng giân số > 0% và dân số gia tăng ở mọi vùng.
* Tỉ lệ giữa các vùng có sự khác biệt rõ rệt: Đông Nam Bộ, Tây Nguyên, Trung du và miền núi phía Bắc có tỉ lệ gia tăng dân số > 1% trong đó Đông Nam Bộ là lớn nhất với ~ 2%. Điều này bởi vì đây là những vùng phát triển kinh tế, thu hút đông dân đến làm việc và đa số là dân số trẻ, trong độ tuổi sinh đẻ.
* Đồng bằng sông Cửu Long là vùng có tỉ lệ gia tăng dân số ít nhất. Vì đa số là các tỉnh miền Tây gần trung tâm kinh tế của cả nước là TP.HCM, Bình Dương nên người dân trong độ tuổi lao động có xu hướng di dân để tìm việc làm.

1. **Biểu đồ tỉ lệ gia tăng dân số của các tỉnh thành**

Chart, bar chart, histogram

Description automatically generated

* **Nhận xét**
* Sự chênh lệch của tỉ lệ gia tăng dân số của các tỉnh trong 1 vùng là khá rõ.
* Các tỉnh có tỉ lệ cao nổi bật của từng vùng: Đà Nẵng, Bắc Ninh, Bình Dương, Đăk Nông. Đây đa số là trung tâm kinh tế của mỗi vùng nên dân các vùng xung quanh sẽ tập trung lại đây để làm việc à dân số tăng mạnh
* Các tỉnh có tỉ lệ rất thấp thậm chí là âm đa số là các tỉnh vùng Đồng bằng sông Cửu Long: Hậu Giang, Đồng Tháp, Sóc Trăng, An Giang. Do người dân ở những vùng này có xu hướng di cư đi các vùng kinh tế phát triển lân cận: Tp.HCM, Bình Dương để sinh sống và làm việc à dân số giảm

1. **Biểu đồ tỉ lệ giới tính trung bình của các vùng**

Chart, bubble chart

Description automatically generated

* **Nhận xét**
* Tỉ lệ nam/ 100 nữ của đa số các vùng là thấp hơn 100 tức là nam giới đang có số lượng ít hơn nữ giới. Tuy nhiên sự chênh lệch là không nhiều cho thấy ở nước ta sự cân bằng giới tính vẫn rất tốt.
* Tây Nguyên có tỉ lệ nam lớn hơn nữ rõ ràng nhất. Có thể là vì ở Tây Nguyên còn tình trạng trọng nam khinh nữ, người dân còn quan tâm nhiều đến giới tính của thai nhi.

1. **Dân số của từng vùng qua các năm**

Calendar

Description automatically generated

* **Nhận xét:**
  + Dân số của các vùng nhìn chung có sự tăng nhẹ qua các năm.
  + Đặc biệt từ năm 2016 – 2017 Bắc Trung bộ và duyên hải miền trung có sự gia tăng dân số vượt bậc so với các vùng.
  + Riêng Đồng bằng sông Cữu Long có sự sụt giảm dân số trong khoảng năm 2017 – 2018.

1. **Biểu đồ thể hiện sự thay đổi mật độ dân số các vùng qua các năm.**

A picture containing diagram

Description automatically generated

* **Nhận xét:**
* Mật độ dân số của Đồng bằng sông Cửu Long và Đông Nam Bộ có xu hướng tăng mạnh qua các năm. Lí do: đây là 2 vùng kinh tế trọng điểm của cả nước, người dân tập trung đến để sinh sống và làm việc càng ngày càng nhiều
* Các vùng còn lại có xu hướng cân bằng, không thay đổi qua các năm.

1. **Biểu đồ tỉ lệ giới tính ở các tỉnh thuộc khu vực Tây Nguyên qua từng năm.**

Line chart

Description automatically generated

* **Nhận xét:**
  + Tỉ lệ giới tính ở các tỉnh thành thuộc vùng Tây Nguyên tuy lớn ( nam > nữ) những những năm gần đây đã có xu hướng giảm về trạng thái cân bằng à các tỉnh có chính sách phù hợp trong việc sinh đẻ và cân bằng giới tính.

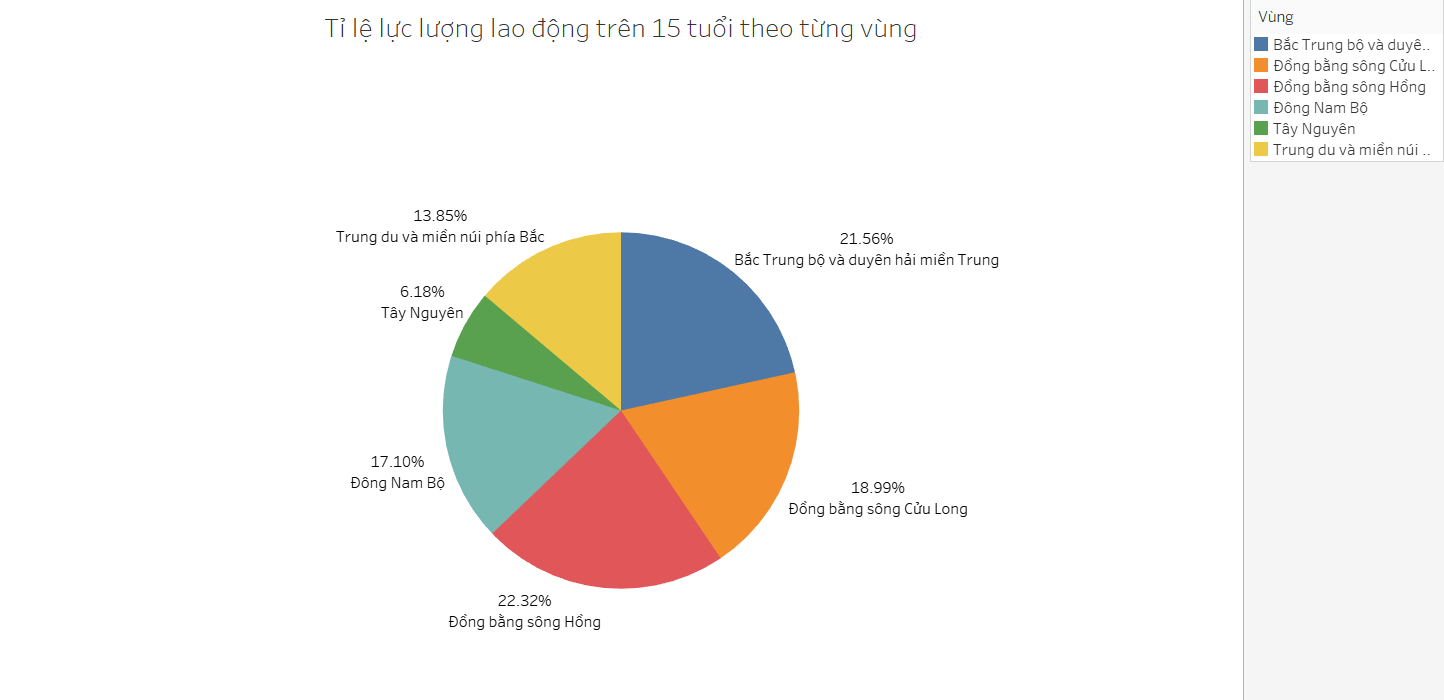
1. **Tổng lực lượng lao động của các vùng theo năm.**

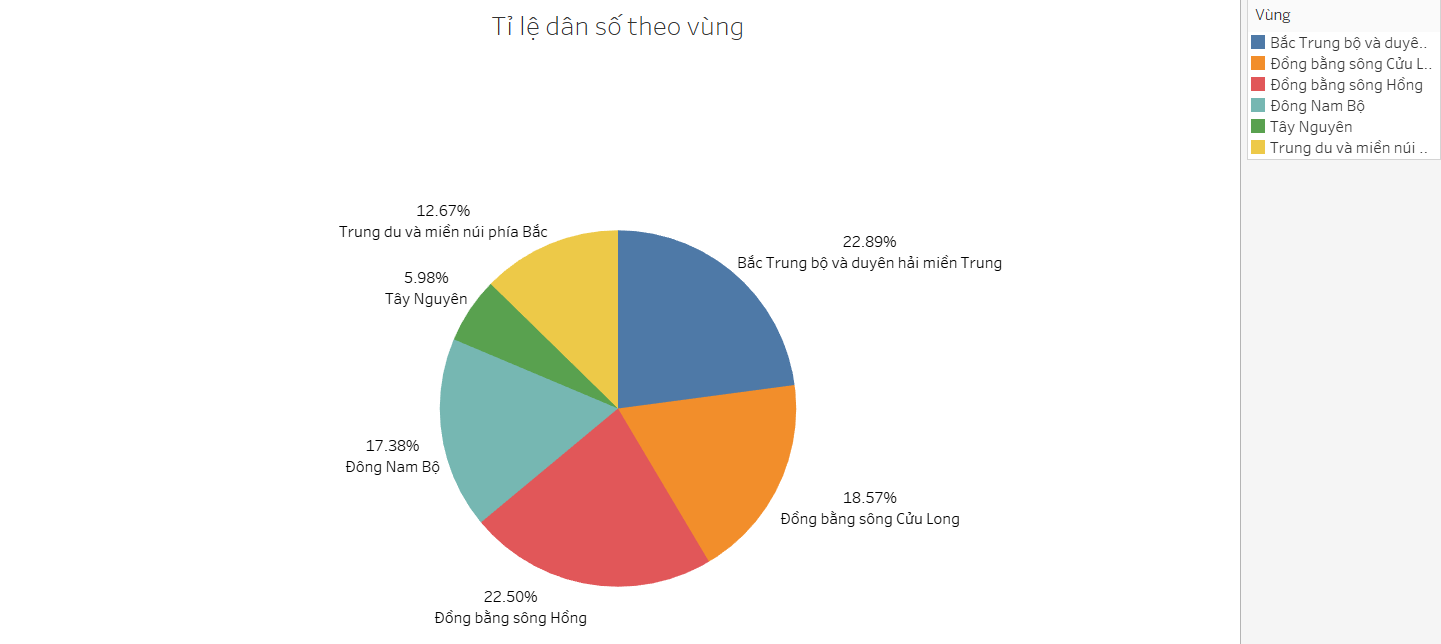
Diagram

Description automatically generated

* **Nhận xét:**
  + Nhìn chung tổng lực lượng lao động của các vùng qua các năm không thay đổi nhiều.
  + Lực lượng lao động ở vùng Tây Nguyên khá thấp so với các vùng còn lại. Điều này dễ lý giải vì vị trí địa lý cũng như địa hình ở Tây Nguyên khá khó khăn nên thu hút ít lực lượng lao động.
  + Lực lượng lao động ở vùng Trung du và miền núi Bắc Bộ cũng tương đối ít, điều kiện địa hình ở đây cũng khá giống với Tây Nguyên, tuy nhiên ở đây là nơi tập trung phần lớn khoáng sản ở Việt Nam nên thu hút được khá nhiều người lao động.
  + Lực lượng lao động các vùng còn lại rất đông và là những khu vực đóng góp nhiều nhất cho nền kinh tế Việt Nam.
  + Lực lượng lao động ở vùng Đồng bằng sông Cửu Long có xu hướng giảm trong vài năm gần đây do có sự can thiệt của Khoa học Công nghệ vào sản xuất nông nghiệp.

1. **Tỉ lệ dân số và tỉ lệ lực lượng lao động của từng vùng**





* **Nhận xét:**
* Ta có thể thấy tỉ lệ phần trăm dân số và tỉ lệ phần trăm lực lượng lao động là không chên lệch nhiều.
* Đồng bằng sông Hồng và Bắc Trung bộ và duyên hải miền Trung có tỉ lệ dân số lớn nhất và cũng là nơi có lực lượng lao động trên 15 tuổi lớn nhất.

1. **Biểu đồ thể hiện tỉ lệ gia tăng dân số các tỉnh vùng Đông Nam Bộ qua các năm.**

Chart, line chart

Description automatically generated

* **Nhận xét:**
* Tỉ lệ gia tăng dân số ở các tỉnh vùng Đông Nam Bộ hầu như có xu hướng tăng nhanh.
* Đặc biệt là tỉnh Bình Dương có tốc độ gia tăng dân số nhanh nhất, qua các năm đều tăng hơn 3.04, cao nhất là 5.3. Điều này là do sự đầu tư phát triển kinh tế được đầu tư mạnh mẻ những năm gần đây, thu hút người dân từ nhiều nơi về làm việc.
* Khu vực này có nguy cơ bùng nổ dân số trong những năm tới.
* Tỉnh Tây Ninh có tỉ lệ gia tăng dân số thấp nhất trong vùng, có có xu hướng giảm ở những năm gần đây.

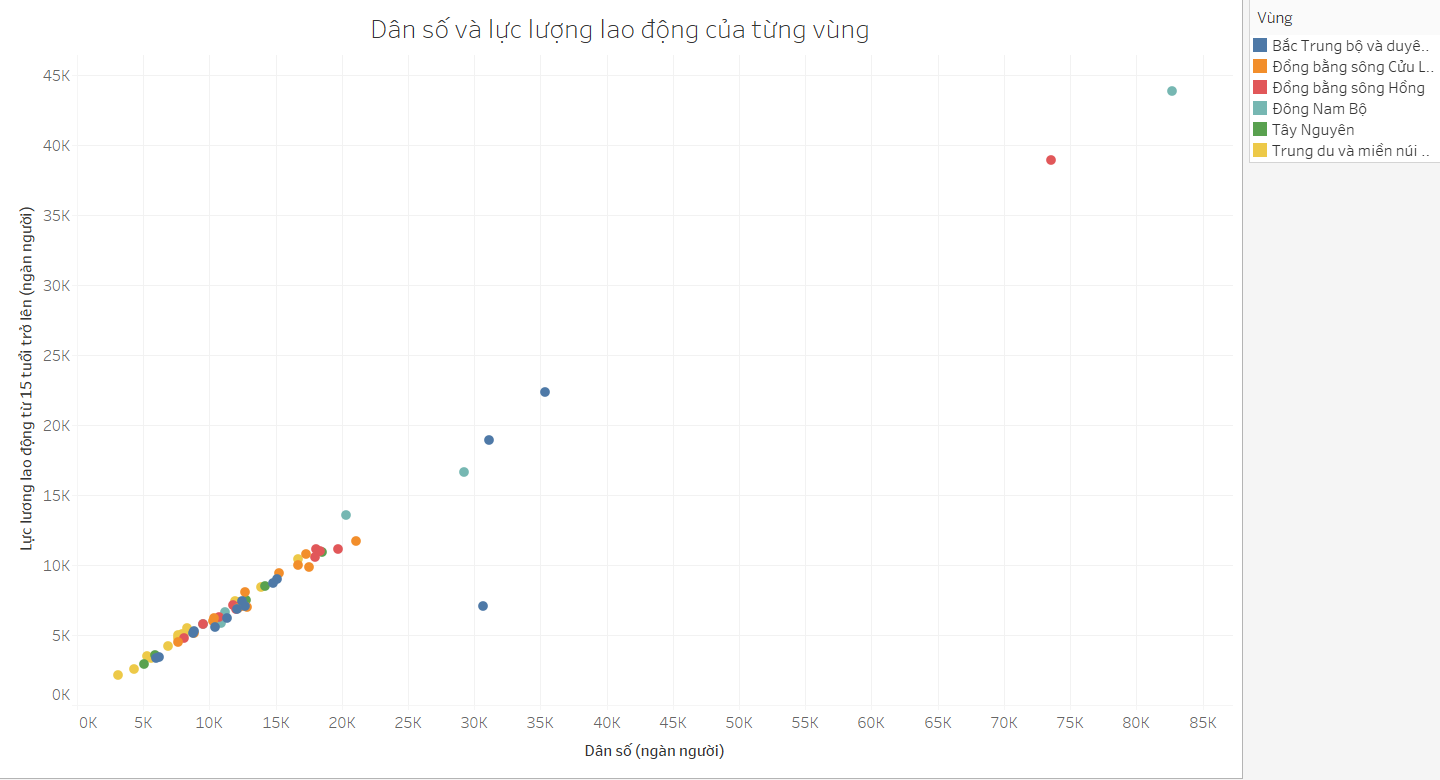
1. **Tình hình dân số, tỷ lệ lao động của thành phố Hồ Chí Minh qua các năm.**

Chart, bar chart

Description automatically generated

* **Nhận xét:**
* Dân số ở TPHCM có tỷ lệ gia tăng rất cao qua các năm.
* Bên cạnh đó, TPHCM cũng có nguồn lao động rất dồi dào, số người lao động luôn chiếm trên 50% tổng dân số ở thành phố này. Điều này rất dễ hiểu bởi vì TPHCM có cơ cấu dân số trẻ và nhiều nguồn lao động từ các tỉnh khác nhập cư.

1. **Dân số và lực lượng lao động của từng vùng**



* **Nhận xét:**
* Sự tương quan giữa dân số và lực lượng lao động trên 15 tuổi giữa các vùng là không có sự khác biệt gì lớn.
* Dân số càng cao thì lực lượng trong độ tuổi lao động càng cao và không có sự lệch về các vùng tập trung công nghiệp.

1. **Biểu đồ tương quan giữa mật độ dân số và dân số trung bình của các tỉnh**

Graphical user interface, chart, scatter chart

Description automatically generated with medium confidence

* **Nhận xét:**
* Nhìn vào biểu đồ ta có thể dễ dàng nhận thấy TP.HCM và Hà Nội là 2 tỉnh/thành phố nỗi bật về cả mật độ dân số và dân số so với các tỉnh còn lại, bởi vì đây là 2 vùng kinh tế phát triển nhất, tập trung rất đông lao động làm việc.
* Các tỉnh: Đồng Nai, Thanh Hóa có dân số cao hơn nhiều so với các tỉnh: Bắc Ninh, Hưng Yên, Hải Phòng tuy nhiên mật độ dân số lại thể hiện ngược lại. Lí do là diện tích của ĐN, TH rất lớn so với dân số.

1. **Biểu đồ mật độ dân số, tỷ lệ gia tăng dân số và dân số trung bình của các vùng trong năm 2020**

Chart, scatter chart, bubble chart

Description automatically generated

* **Nhận xét:**
* Đồng bằng sông Hồng và Đông Nam Bộ là hai vùng có mật độ dân số cao nhất toàn quốc. Đồng thời, đồng bằng sông Hồng cũng là nơi tập trung dân cư lớn nhất của cả nước
* Trung du và miền núi phía Bắc và Tây Nguyên là hai vùng có mật độ dân số thấp nhất. Trong đó, Tây Nguyên là nơi có ít dân cư sinh sống nhất.
* Ở giai đoạn này, Đông Nam Bộ có tỷ lệ tăng dân số bình quân cao nhất cả nước; Đồng bằng sông Cửu Long có tỷ lệ tăng dân số bình quân thấp nhất cả nước.
* Một số khu vực có tỷ lệ gia tăng dân số vượt mức trung bình cả nước như Đông Nam bộ, ĐB sông Cửu Long, Tây Nguyên..
* Ở Tây Nguyên, tuy mật độ dân số rất thấp nhưng tỷ lệ gia tăng dân số lại khá cao. Điều này xảy ra với nguyên nhân chính đó là sự di dân tự do đến các tỉnh ở vùng núi Tây nguyên.

1. **Tài liệu tham khảo**

* [1]. <https://medium.com/data-folks-indonesia/predicting-indonesian-population-using-arima-33af3c97bd11>
* [2]. <https://www.tableau.com/learn/articles/data-visualization>