2022

Báo cáo thực hành

SCIKIT-LEARN VÀ PCA, LDA

BÙI NGỌC CHÍNH - 19127109

Mục lục

I. Thông tin về tập dữ liệu được chọn 2
1. China Population: 2
2. China Population Forecast: 2
II. Trực quan hóa dữ liệu 2
1. Bản đồ nhiệt về hệ số tương quan giữa các biến (correlations for multivariate data): 2
2. Bảng các thuộc tính có độ tương quan lớn nhất với nhau (most correlated): 3
3. Nên giữ lại bao nhiêu thành phần chính 3
4. Tổng dân số của TQ (gộp cả 2 tập dữ liệu): 4
5. Tỉ lệ thay đổi dân số theo năm (gộp cả 2 tập dữ liệu): 5
III. Phân tích thành phần chính (Principal Components Analysis) 5
1. Thư viện và hỗ trợ của thư viện: 5
2. Nhận xét 6
3. Hình ảnh kết quả 6
IV. Phân tích phân biệt tuyến tính (Linear Discriminant Analysis) 6
1. Thư viện và hỗ trợ thư viện 6
2. Nhận xét 6
3. Hình ảnh kết quả 7

I. Thông tin về tập dữ liệu được chọn

Link: https://www.kaggle.com/datasets/anandhuh/population-data-china

1. China Population:

- Dân số Trung Quốc từ năm 1955 đến năm 2020
- Thông tin tập dữ liệu:
  - Year – Năm từ 2020 tới 1955
  - Population - Dân số trong năm tương ứng
  - Yearly % Change - Phần trăm thay đổi dân số hàng năm
  - Yearly Change - Thay đổi dân số hàng năm
  - Migrants (net) - Tổng số người di cư
  - Median Age - Tuổi trung bình của dân số
  - Fertility Rate - Tỷ lệ sinh
  - Density (P/Km²)- Mật độ dân số (dân số trên km vuông)
  - Urban Pop %- Phần trăm dân số thành thị
  - Urban Population- Dân số đô thị
  - Country's Share of World Pop - Tỷ lệ dân số
  - World Population - Dân số Thế giới trong năm tương ứng
  - China Global Rank - Xếp hạng toàn cầu về dân số

2. China Population Forecast:

- Dự đoán dân số TQ trong tương lai.
- Thông tin tập dữ liệu giống China Population

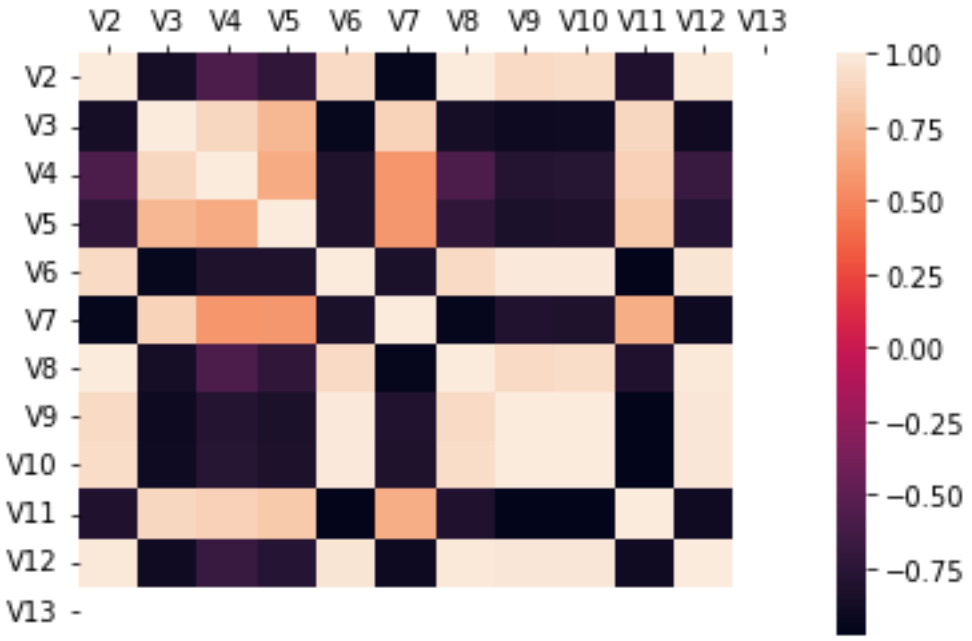
II. Trực quan hóa dữ liệu

1. Bản đồ nhiệt về hệ số tương quan giữa các biến (correlations for multivariate data):

a. Ý tưởng:

Dùng ma trận hệ số tương quan để vẽ lên thành 1 bản đồ nhiệt

b. Hình ảnh:



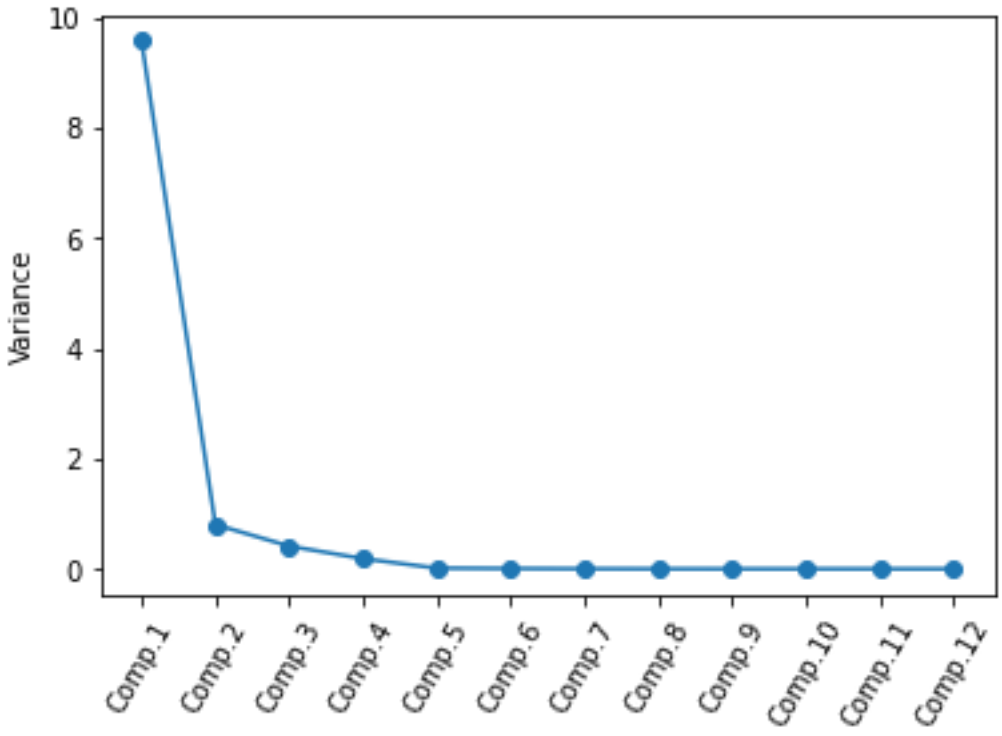
2. Bảng các thuộc tính có độ tương quan lớn nhất với nhau (most correlated):

	FirstVariable	SecondVariable	Correlation
0	V2	V8	0.999950
1	V9	V10	0.999377
2	V6	V10	0.988316
3	V6	V9	0.988199
4	V2	V12	0.984355
5	V8	V12	0.984047
6	V9	V11	-0.976744
7	V10	V12	0.976652
8	V9	V12	0.970542
9	V10	V11	-0.969559

3. Nên giữ lại bao nhiêu thành phần chính
   a. Tính độ lệch chuẩn:

	Standard deviation
PC1	9.577326
PC2	0.803574
PC3	0.412760
PC4	0.184926
PC5	0.012336
PC6	0.006262
PC7	0.002132
PC8	0.000627
PC9	0.000036
PC10	0.000016
PC11	0.000006
PC12	0.000000

b. Trực quan hóa (đổi từ độ lệch chuẩn sang phương sai):



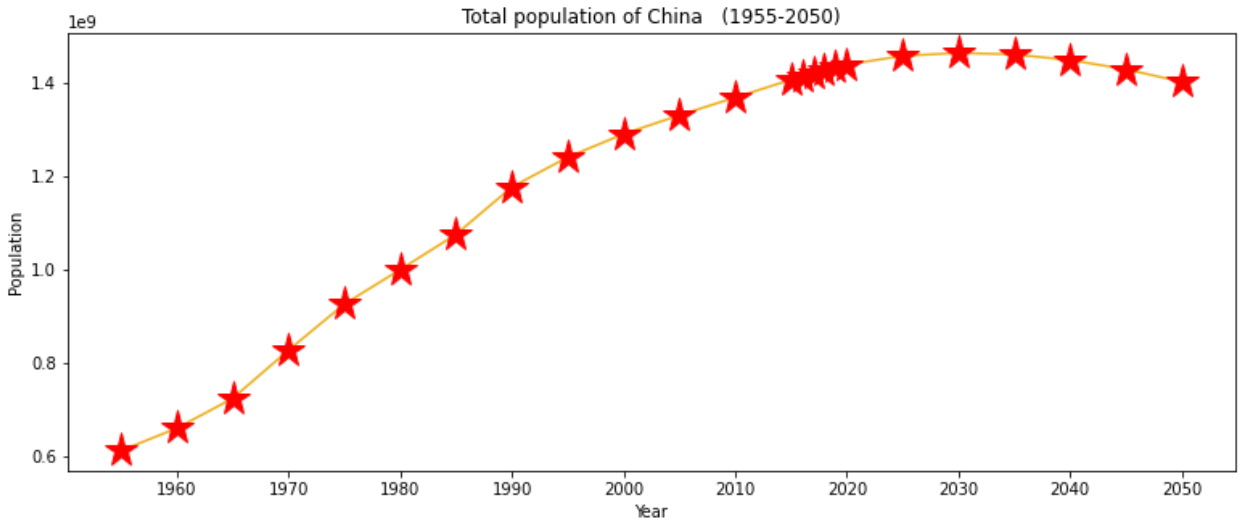
4. Tổng dân số của TQ (gộp cả 2 tập dữ liệu):

a. Ý tưởng:

Trục x: số năm

Trục y: dân số

b. Hình ảnh:



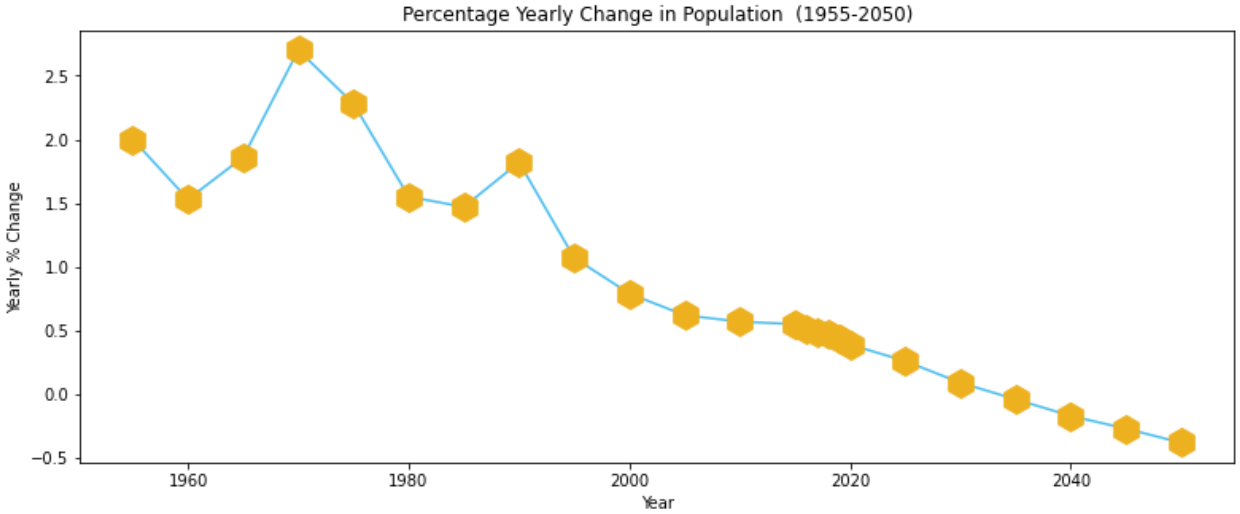
5. Tỉ lệ thay đổi dân số theo năm (gộp cả 2 tập dữ liệu):

a. Ý tưởng:

Trục x: số năm

Trục y: tỉ lệ tăng giảm

b. Hình ảnh kết quả:



III. Phân tích thành phần chính (Principal Components Analysis)

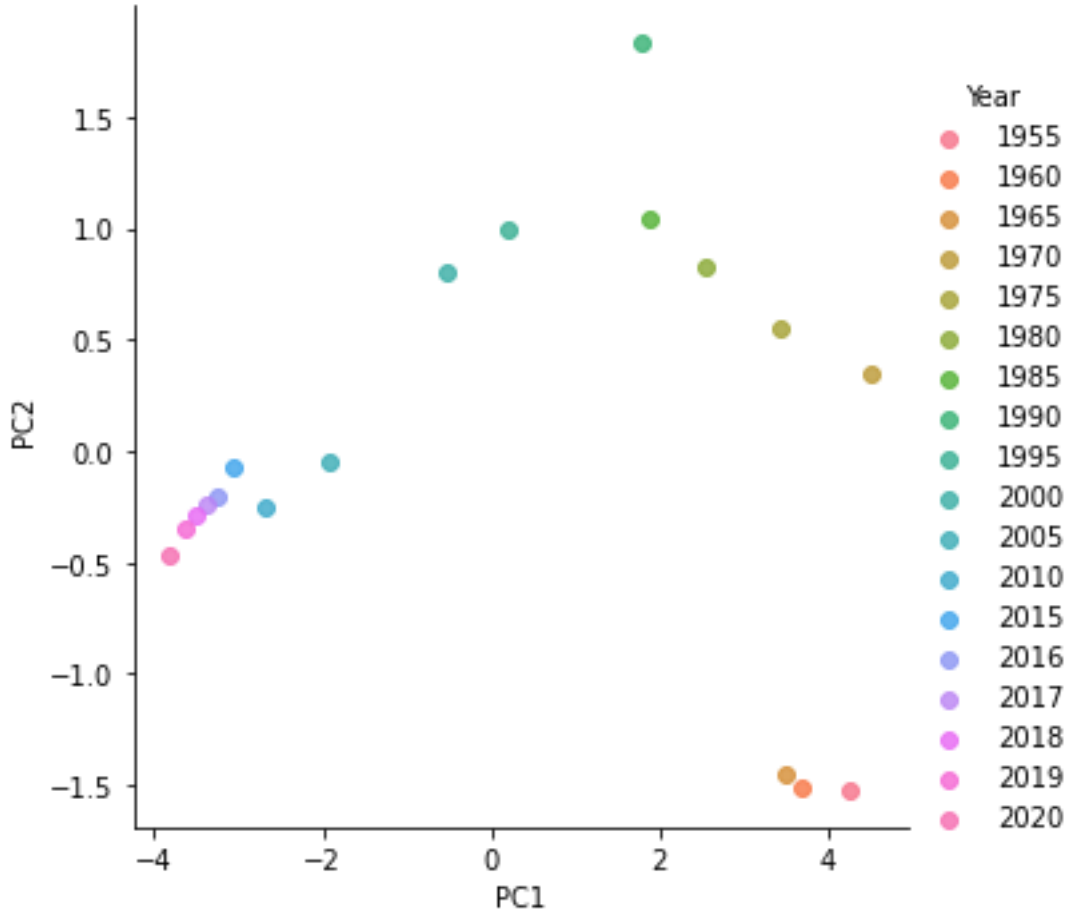
1. Thư viện và hỗ trợ của thư viện:

Theo hướng dẫn thực hành thì em sử dụng Scikit learn, trong thư viện này có hỗ trợ tính PCA nhanh gọn chỉ bằng cách gọi module PCA có sẵn của thư viện. Việc duy nhất chúng ta cần chuẩn bị là tập dữ liệu và chuẩn hóa tập dữ liệu đó bằng hàm scale() cũng được hỗ trợ bởi Scikit luôn.

2. Nhận xét

- Bài tập được thực hiện nhanh gọn bằng thư viện.
- Các module và hàm có sẵn đều có hướng dẫn trên trang chủ Scikit-Learn.

3. Hình ảnh kết quả



IV. Phân tích phân biệt tuyến tính (Linear Discriminant Analysis)

1. Thư viện và hỗ trợ thư viện

Cũng như PCA, LDA được hỗ trợ bởi Scikit-Learn, tập dữ liệu chuẩn bị cần chia thành X và Y. Với X là các thuộc tính cần xem xét phân tích, Y là mục tiêu hướng đến của phân tích đó. Ví dụ trong dataset China population trên, Y sẽ là xếp hạng dân số thế giới của TQ, trong khi đó, các thuộc tính còn lại sẽ là X.

2. Nhận xét

- Dễ sử dụng bằng thư viện.

3. Hình ảnh kết quả