ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN THỰC HÀNH

|Đề tài|

THỰC HIỆN QUY TRÌNH KHOA HỌC DỮ LIỆU VÀ MÔ HÌNH HÓA DỮ LIỆU

|Trợ giảng phụ trách đồ án|

Trần Đại Chí
Nguyễn Bảo Long
Lê Nhựt Nam
Nguyễn Thái Vũ

Môn học: Nhập môn Khoa học dữ liệu

Thành phố Hồ Chí Minh - 2022

THÔNG TIN THÀNH VIÊN

MSSV	Họ tên
20120357	Nguyễn Đức Minh Quân
20120402	Nguyễn Hoàng Việt
20120120	Nguyễn Việt Khoa
20120565	Nguyễn Tấn Sơn

MỤC LỤC

Nội dung

CHI TIẾT ĐỒ ÁN

1. Giới thiệu sơ bộ về đồ án

Đồ án lần này được thực hiện để nâng cao khả năng thực hiện một quy trình khoa học dữ liệu. Đồ án tổng hợp lại tất cả kiến thức và kĩ năng đã được học ở môn Nhập môn Khoa học dữ liệu giúp cho sinh viên có thể ôn tập và cũng cố kiến thức cũng như nâng cao khả năng của bản thân.

2. Giới thiệu về thu thập dữ liệu

Tập dữ liệu của đồ án lần này được lấy từ trang web

https://www.worldometers.info/world-population/population-by-country/, vì là tập dữ liệu thu thập về dân số được công khai bởi UNESCO nên tập dữ liệu này đã được cấp phép, vì vậy chúng em sẽ sử dụng tập dữ liệu này cho đồ án.

Tập dữ liệu được thu thập bằng cách sử dụng thư viện scrapy để cào dữ liệu tại trang web trên sau đó lưu vào một file csv.

3. Khám phá dữ liệu và tiền xử lí dữ liệu

a. Khám phá dữ liệu

Dữ liệu có 4230 dòng và 13 cột

```
print(df.shape)
df.dtypes
```

```
(4230, 13)
```

Mỗi dòng ở đây thể hiện cho các đặc điểm liên quan đến dân số của một quốc gia ở một năm cụ thể.

Các cột thể hiện những mô tả về đặc điểm liên quan đến dân số:

- Country: tên quốc gia
- Year: năm thực hiện thống kê
- Population: dân số
- Yearly%Change: tỉ lệ gia tăng dân số (dương: tăng, âm: giảm)
- YearlyChange: số dân thay đổi hằng năm
- Migrants(net): số người di cư đến
- MedianAge: độ tuổi trung vị

- FertilityRate: tỉ suất sinh
- UrbanPop%: tỉ lệ dân thành thị
- UrbanPopulation: số dân thành thị
- %OfWorldPop: tỉ lệ % so với dân số thế giới
- GlobalRank: xếp hạng về dân số
- Continent: châu lục mà quốc gia đó trực thuộc

Các kiểu dữ liệu trong tập này vừa có dạng số vừa có dạng object (thực chất là dạng chuỗi), một số chỗ là NaN nên sẽ được gán là kiểu float trong pandas.

```
Year                  int64
Population            int64
Yearly%Change       float64
YearlyChange          int64
Migrants(net)       float64
MedianAge           float64
FertilityRate       float64
UrbanPop%           float64
UrbanPopulation     float64
Country%OfWorldPop  float64
GlobalRank            int64
Country              object
Continent            object
dtype: object
```

b. Tiền xử lí dữ liệu

Chúng ta có 2 vấn đề cần xử lí với kiểu dữ liệu này:

- Tìm cách thay thế các NaN hoặc giá trị thiếu ở các cột dạng số.
- Bỏ đi một số cột có thể suy ra được từ dữ liệu của các cột khác.

Xem xét trên df thấy cột UrbanPopulation, UrbanPop%, ta thấy các cột này chỉ thiếu vài giá trị, trong khi đó các quốc gia có giá trị thiếu ở các cột FertilityRate, MedianAge, Migrants(net) đều giống nhau (đều thiếu các giá trị ở cột đó và các giá trị thiếu cũng trùng năm với nhau).

Giải pháp: Dùng hồi quy để thay thế NaN, dựa trên cột UrbanPop% (vì những dữ liệu còn lại của FertilityRate, MedianAge, Migrants(net) phân bố khá đều và độ chênh lệch không cao, do đó thay thế bằng giá trị trung vị hoặc trung bình đều không hợp lí)

```
def regressionFillUrban(df,Country):
    df_=df[df['Country']==Country]
    r=df_[df_['UrbanPopulation']!=df_['UrbanPopulation']].index
    df_=df_[df_['UrbanPopulation']==df_['UrbanPopulation']]
    def estimate_coef(x, y):
        n = np.size(x)
        m_x = np.mean(x)
        m_y = np.mean(y)
        SS_xy = np.sum(y*x) - n*m_y*m_x
        SS_xx = np.sum(x*x) - n*m_x*m_x
        b_1 = SS_xy / SS_xx
        b_0 = m_y - b_1*m_x
        return (b_0, b_1)
    def regression(b,x):
        return b[0] + b[1]*x
    b=estimate_coef(df_['Year'],df_['UrbanPop%'])
    for i in r :
        df.at[i,'UrbanPop%']=min(regression(b,df.loc[i]['Year']),100)
        df.at[i,'UrbanPopulation']=round(df.loc[i]['UrbanPop%']*df.loc[i]['Population']/100,0)
for i in missing_priority:
    regressionFillUrban(df,i)
```

Về logic nếu có diện tích, dân số, ta có thể tính được mật độ dân số nên ta hoàn toàn có thể bỏ đi cột Density(P/Km2). Sau đó lưu dữ liệu đã được xử lí lại vào file csv.

4. Đặt câu hỏi và trả lời

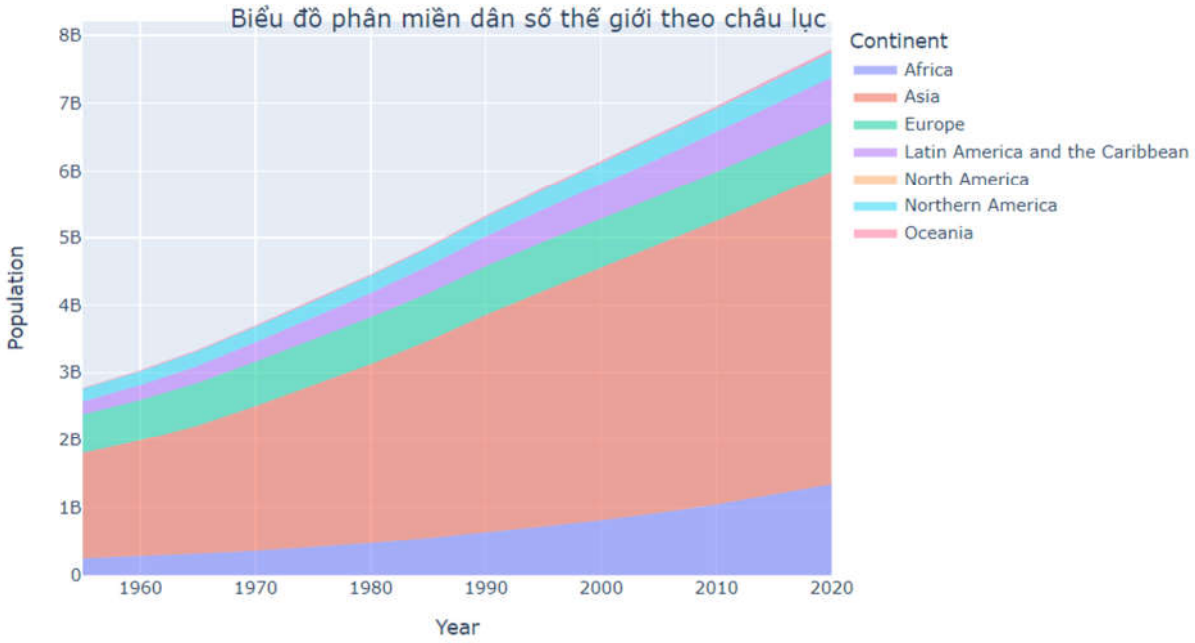
Dựa vào dữ liệu đã khám phá, chúng ta có thể đặt ra một số câu hỏi như sau:

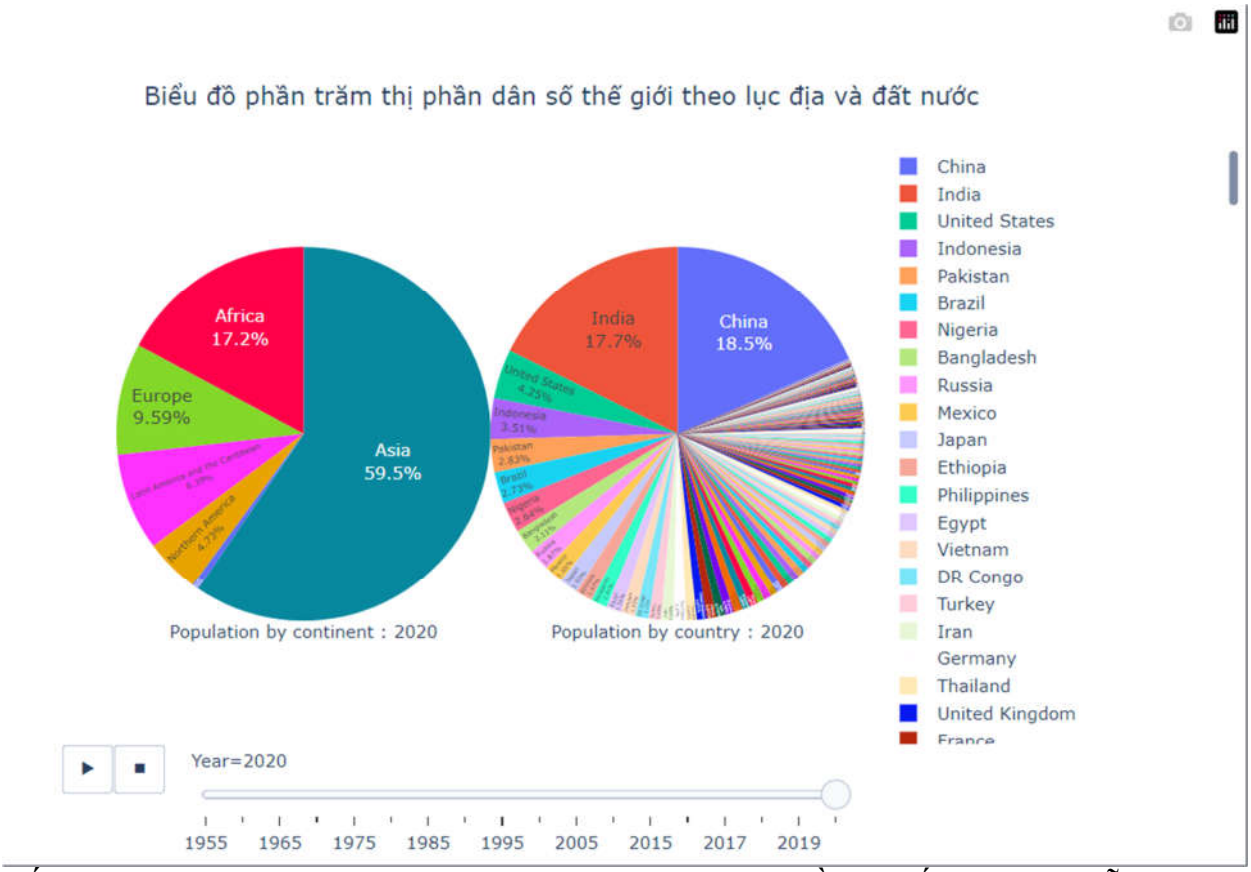
a. **Câu hỏi 1:** Làm thế nào để dễ dàng tra cứu thông tin dân số theo từng nước/lục địa qua các năm, tỉ lệ phần trăm dân số của từng nước/lục địa so với thế giới (qua các năm)?

*Tiền xử lý để trả lời câu hỏi: Do thông tin cần trình bày có rất nhiều đặc điểm cần thể hiện (dân số của các nước/lục địa, sự thay đổi theo thời gian,…) nên nhóm đã trực quan dữ liệu bằng interactive plot trong thư viện plotly (biểu đồ tương tác, có thể thay đổi dữ liệu hiển thị theo yêu cầu của người dùng).

*Trả lời: Bằng cách vẽ các biểu đồ:


Biểu đồ thống kê dân số các nước trên thế giới
Continent
Asia
Europe
Africa
Oceania
Latin America and the Caribbean
Northern America
North America
Population
1B
600M
200M
100M
40M
20M
7M
3M
1M
500k
200k
90k
40k
20k
7000
3000
1000
500
200
80
30
10
Afghanistan
Cambodia
Iran
Kyrgyzstan
Myanmar
Saudi Arabia
Tajikistan
Vietnam
Bosnia and Herzegovina
Faeroe Islands
Hungary
Lithuania
North Macedonia
Serbia
United Kingdom
Cabo Verde
DR Congo
Gabon
Liberia
Mayotte
Réunion
South Africa
Western Sahara
French Polynesia
New Zealand
Tokelau
Argentina
British Virgin Islands
Curaçao
Grenada
Martinique
Puerto Rico
Trinidad and Tobago
Greenland
Year=2020
Country
1955
1965
1975
1985
1995
2005
2015
2017
2019



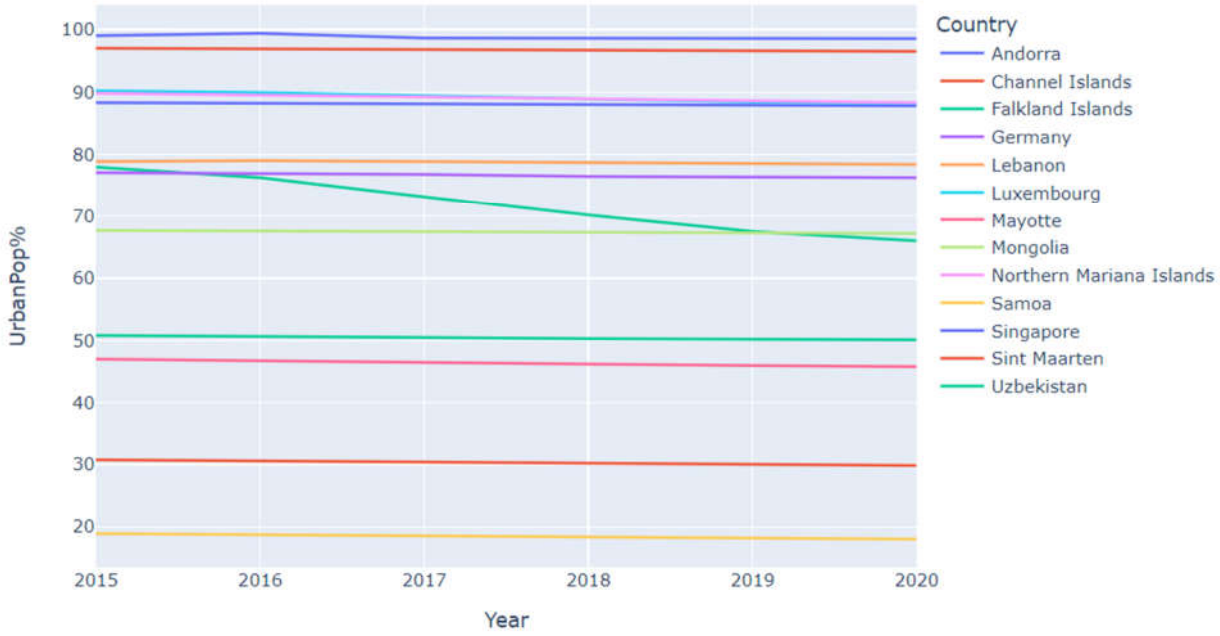


*Ý nghĩa của câu hỏi: Giúp việc tra cứu các thông tin về dân số một cách dễ dàng hơn.

b. <u>Câu hỏi 2:</u> Có nước nào có tỉ lệ dân thành thị giảm gần đây không? Điều đó mang ý nghĩa như thế nào?

<u>*Tiền xử lý để trả lời câu hỏi:</u> Chọn ra dữ liệu của 3 năm mới nhất trong df để so sánh giá trị cột UrbanPop%, nếu quốc gia nào có giá trị UrbanPop% giảm dần từ 2018 đến 2020 thì đó là nước có tỉ lệ dân thành thị giảm gần đây.

<u>*Trả lời:</u> Vẽ biểu đồ, các quốc gia được hiển thị trên biểu đồ là đáp án của câu hỏi này.



<u>*Ý nghĩa của việc trả lời câu hỏi:</u> Từ thông tin này ta có thể tìm hiểu thêm về cách thức và mục đích của việc giảm tỉ lệ dân thành thị ở các quốc gia đó. Có thể việc giảm tỉ lệ dân thành thị cũng liên quan đến việc giảm ùn tắc giao thông, tái cơ cấu nền kinh tế ở một số vùng miền để giảm sự chênh lệch về kinh tế giữa các vùng trong nước.

c. <u>Câu hỏi 3:</u> Mối tương quan giữa độ tuổi trung vị với tỉ lệ tăng dân số?

<u>*Trả lời:</u> Vẽ biểu đồ, các quốc gia được hiển thị trên biểu đồ là đáp án của câu hỏi này.

Giải thích: Mỗi điểm thể hiện cho một quốc gia, độ lớn của điểm thể hiện tổng dân số, cột x và y thể hiện lần lượt tỉ lệ gia tăng dân số và độ tuổi trung vị của quốc gia đó.

<u>*Ý nghĩa của việc trả lời câu hỏi:</u> Từ thông tin này ta có thể biết thêm mối tương quan giữa độ tuổi trung vị với tỉ lệ tăng dân số: độ tuổi trung vị càng thấp (tức dân số càng trẻ) thì tỉ lệ gia tăng dân số càng cao. Trong khi đó, các nước dân số già lại đang có xu hướng thực hiện lối sống "hưởng thụ", không quá áp đặt việc sinh con để có thể sống tự do hơn, điều đó khiến dân số của các nước này có khả năng ngày càng "già" hơn nữa.

5. Mô hình hóa dữ liệu

Nhóm sẽ sử dụng mô hình hồi quy tuyến tính (Linear Regression) để đánh giá và đưa ra một số dự đoán cho dữ liệu.

a. Đánh giá BestToLive (quốc gia đáng sống)

Ban đầu nhóm request html để lấy thông tin về tên của các quốc gia đáng sống (top 50 quốc gia đáng sống trên thế giới, theo thống kê từ tờ báo US News với đường dẫn https://www.usnews.com/news/best-countries/rankings/quality-of-life). Sau đó tiến hành gán nhãn cho các quốc gia trong dataframe df.

	Country	Population	Yearly%Change	MedianAge	UrbanPop%	Migrants(net)	FertilityRate	BestToLive
0	Afghanistan	38928346	2.33	18.4	25.4	-62920.0	4.56	False
1	Albania	2877797	-0.11	36.4	63.5	-14000.0	1.62	False
2	Algeria	43851044	1.85	28.5	72.9	-10000.0	3.05	False
3	American Samoa	55191	-0.22	0.0	88.1	0.0	0.00	False
4	Andorra	77265	-0.19	0.0	87.8	0.0	0.00	False

Chọn ra dữ liệu thuộc năm 2020 để phân lớp. Nhóm lựa chọn việc phân lớp theo hai thuộc tính.

- Phân lớp dựa vào UrbanPop% và Migrants(net):

```
Binary cross entropy loss là :  6.466844936886641
Độ chính xác của mô hình :  0.8127659574468085
```

- Phân lớp dựa vào FertilityRate và Yearly%Change (bỏ đi các giá trị bằng 0 trong FertilityRate)

```
Binary cross entropy loss là :  8.763605932921974
Độ chính xác của mô hình :  0.746268656716418
```

b. Dự đoán dân số trong tương lai

Tạo ra 2 hàm linear_regression_1 và linear_regression_2 để biểu diễn hồi quy tuyến tính (linear bậc 1) và hồi quy đa thức bậc 2 (linear bậc 2) để so sánh sự khác biệt. Ngoài ra, phải xử lý trường hợp mô hình dự đoán số dân của một số nước giảm về 0:

```
#Có những nước với dự đoán GIẢM có nguy cơ diệt vong (0 dân)
dict_pred['Population']=[i if i>0 else 1000 for i in dict_pred.pop('Population')]
dict_pred = [dict(zip(dict_pred,t)) for t in zip(*dict_pred.values())]
for row in dict_pred :
    yield row
```

Kết quả linear bậc 1:

	Year	YearlyChange	Migrants(net)	MedianAge	FertilityRate	UrbanPopulation	Yearly%Change	UrbanPop%	Country	Continent	LossFunctCheck	Population
0	2025	1.015350e+06	-44993.213325	16.582543	5.409869	9.517517e+06	2.686857	25.185605	Afghanistan	Asia	1.482467	3.778951e+07
1	2030	1.091289e+06	-42795.797468	16.486920	5.210928	1.027058e+07	2.708648	25.492221	Afghanistan	Asia	1.482467	4.028908e+07
2	2035	1.167229e+06	-40598.381611	16.391296	5.011988	1.102364e+07	2.727894	25.763014	Afghanistan	Asia	1.482467	4.278864e+07
3	2040	1.243168e+06	-38400.965754	16.295673	4.813048	1.177671e+07	2.745014	26.003915	Afghanistan	Asia	1.482467	4.528821e+07
4	2050	1.395047e+06	-34006.134040	16.104425	4.415167	1.328284e+07	2.774150	26.413873	Afghanistan	Asia	1.482467	5.028734e+07
...	...	...	...	...	...	...	...	...	...	...	...	...
1170	2025	2.038575e+05	-134017.823668	18.685330	3.276979	6.055059e+06	1.284438	38.150902	Zimbabwe	Africa	0.278630	1.587134e+07
1171	2030	2.075294e+05	-145549.966245	18.885654	2.957300	6.500572e+06	1.236126	38.719944	Zimbabwe	Africa	0.278630	1.678869e+07
1172	2035	2.112013e+05	-157082.108822	19.085978	2.637620	6.946084e+06	1.192821	39.230023	Zimbabwe	Africa	0.278630	1.770604e+07
1173	2040	2.148732e+05	-168614.251398	19.286302	2.317941	7.391596e+06	1.153782	39.689850	Zimbabwe	Africa	0.278630	1.862339e+07
1174	2050	2.222171e+05	-191678.536552	19.686949	1.678583	8.282621e+06	1.086206	40.485791	Zimbabwe	Africa	0.278630	2.045809e+07

Kết quả linear bậc 2:

	Year	YearlyChange	Migrants(net)	MedianAge	FertilityRate	UrbanPopulation	Yearly%Change	UrbanPop%	Country	Continent	LossFunctCheck	Population
0	2025	1.015350e+06	-44993.213325	16.582543	5.409869	9.517517e+06	2.686857	25.185605	Afghanistan	Asia	1.482467	3.778951e+07
1	2030	1.091289e+06	-42795.797468	16.486920	5.210928	1.027058e+07	2.708648	25.492221	Afghanistan	Asia	1.482467	4.028908e+07
2	2035	1.167229e+06	-40598.381611	16.391296	5.011988	1.102364e+07	2.727894	25.763014	Afghanistan	Asia	1.482467	4.278864e+07
3	2040	1.243168e+06	-38400.965754	16.295673	4.813048	1.177671e+07	2.745014	26.003915	Afghanistan	Asia	1.482467	4.528821e+07
4	2050	1.395047e+06	-34006.134040	16.104425	4.415167	1.328284e+07	2.774150	26.413873	Afghanistan	Asia	1.482467	5.028734e+07
...	...	...	...	...	...	...	...	...	...	...	...	...
1170	2025	2.038575e+05	-134017.823668	18.685330	3.276979	6.055059e+06	1.284438	38.150902	Zimbabwe	Africa	0.278630	1.587134e+07
1171	2030	2.075294e+05	-145549.966245	18.885654	2.957300	6.500572e+06	1.236126	38.719944	Zimbabwe	Africa	0.278630	1.678869e+07
1172	2035	2.112013e+05	-157082.108822	19.085978	2.637620	6.946084e+06	1.192821	39.230023	Zimbabwe	Africa	0.278630	1.770604e+07
1173	2040	2.148732e+05	-168614.251398	19.286302	2.317941	7.391596e+06	1.153782	39.689850	Zimbabwe	Africa	0.278630	1.862339e+07
1174	2050	2.222171e+05	-191678.536552	19.686949	1.678583	8.282621e+06	1.086206	40.485791	Zimbabwe	Africa	0.278630	2.045809e+07

***Nhận xét:**

- Linear bậc 1: biên độ bị lệch do ảnh hưởng từ các số liệu tăng đột biến. Ví dụ giả sử Afghanistan có dân số luôn tăng theo thời gian thì dân số năm 2025 được dự đoán sẽ nhỏ hơn dân số năm 2020.
- Linear bậc 2: Overfit, không phù hợp để chuẩn đoán các mốc thời gian xa hơn, tuy nhiên rất tốt để chuẩn đoán các giá trị liền kề.

Vì vậy, ta đặt mốc ổn định của thế giới là năm 1990 (tình hình thế giới có thể được coi là ít xung độ vũ trang) để cải tiến dự đoán của LD1:

	Year	YearlyChange	Migrants(net)	MedianAge	FertilityRate	UrbanPopulation	Yearly%Change	UrbanPop%	Country	Continent	LossFunctCheck	Population
0	2025	1.043706e+06	-70941.417972	17.952350	4.538854	1.089753e+07	2.420311	25.270914	Afghanistan	Asia	0.438159	4.312282e+07
1	2030	1.116381e+06	-89952.464963	18.336933	4.027601	1.212830e+07	2.351515	25.546706	Afghanistan	Asia	0.438159	4.747499e+07
2	2035	1.189056e+06	-108963.511954	18.721517	3.516348	1.335906e+07	2.294273	25.776180	Afghanistan	Asia	0.438159	5.182715e+07
3	2040	1.261731e+06	-127974.558945	19.106101	3.005095	1.458982e+07	2.245900	25.970099	Afghanistan	Asia	0.438159	5.617932e+07
4	2050	1.407081e+06	-165996.652927	19.875268	1.982589	1.705135e+07	2.168622	26.279893	Afghanistan	Asia	0.438159	6.488365e+07
...	...	...	...	...	...	...	...	...	...	...	...	...
1170	2025	1.870439e+05	-148825.326463	18.942374	3.607345	6.033273e+06	1.221004	39.384611	Zimbabwe	Africa	3.104178	1.531886e+07
1171	2030	1.876839e+05	-164660.079143	19.167848	3.451781	6.451871e+06	1.171257	40.263437	Zimbabwe	Africa	3.104178	1.602414e+07
1172	2035	1.883239e+05	-180494.831822	19.393322	3.296216	6.870469e+06	1.125704	41.068164	Zimbabwe	Africa	3.104178	1.672943e+07
1173	2040	1.889639e+05	-196329.584501	19.618796	3.140651	7.289066e+06	1.083837	41.807785	Zimbabwe	Africa	3.104178	1.743471e+07
1174	2050	1.902439e+05	-227999.089860	20.069744	2.829522	8.126262e+06	1.009504	43.120943	Zimbabwe	Africa	3.104178	1.884528e+07

TÀI LIỆU THAM KHẢO

[1] Tài liệu lý thuyết môn học **Nhập môn Khoa học dữ liệu 20_21**.

[2] https://population.un.org/wpp/Download/Standard/MostUsed/

[3] https://dash.plotly.com/interactive-graphing

[4] https://rpubs.com/lengockhanhi/445130