Đề thi:

PYTHON FOR MACHINE LEARNING, DATA SCIENCE AND VISUALIZATION

Thời gian: 120 phút

Ngày thi : 25/09/2022

**** Học viên tạo 1 thư mục là **DS2_HoVaTen**, lưu tất cả bài làm vào để nộp chấm điểm ****

**** Học viên được sử dụng tài liệu ****

Chú ý, với mỗi câu:

- Học viên cần kiểm tra xem dữ liệu có bị thiếu (NaN, null, hoặc để trống) hay không, nếu có thì cần chuẩn hóa trước khi làm bài.
- Cần hiển thị thông tin chung của dữ liệu bằng cách dùng shape, head(), tail(), info()… để có cái nhìn ban đầu về dữ liệu.
- Lần lượt thực hiện các bước làm bài như đã được hướng dẫn làm bài tập trong lớp.
- Các câu viết trên 1file Jupyter Notebook, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

Câu 1. Numpy Array (1.5 điểm)

1. Yêu cầu: sử dụng thư viện Numpy thực hiện các yêu cầu sau:

– Phát sinh mảng 2 chiều có kích thước 5 x 3 với các phần tử có giá trị phát sinh ngẫu nhiên từ 1 đến 10 với np.random.seed(2). (0.5 điểm)

```
array([[9, 9, 7],
       [3, 9, 8],
       [3, 2, 6],
       [5, 5, 6],
       [8, 4, 7]])
```

– Cập nhật các phần tử nhỏ hơn 5 bằng phần tử xuất hiện nhiều nhất trong mảng, các giá trị còn lại giữ nguyên. (0. 5 điểm)

```
array([[9, 9, 7],
       [9, 9, 8],
       [9, 9, 6],
       [5, 5, 6],
       [8, 9, 7]])
```

– In ra các phần tử < 8 trong mảng (0.5 điểm)

```
array([7, 6, 5, 5, 6, 7])
```

Câu 2: Text (spam.csv)(1.5 điểm)

Cho dữ liệu **spam.csv** thực hiện các yêu cầu sau :
1. Đọc dữ liệu và tạo đoạn text từ cột content. Sau đó thực hiện chuẩn hóa đoạn text (loại bỏ các từ không quan trọng) (0.5 điểm)
Gợi ý: dùng stopwords

2. Tạo biểu đồ Wordcloud có kết quả gợi ý như sau: (0.5 điểm)

3. Cho tập tin hình ảnh **virus.png**, hãy tạo biểu đồ có kết quả gợi ý như hình sau: (0.5 điểm)

Câu 3. Xử lý dữ liệu và trực quan hóa: (5 điểm)

1. Tạo DataFrame df chứa danh sách các sinh viên khoa CN, khoa TO và khoa VL. Các danh sách này được đọc từ các tập tin sinh_vien_CN.csv, sinh_vien_TO.csv và sinh_vien_VL.csv (0.5 điểm)

2. Hiển thị thông tin chung của dữ liệu: head, tail, info, số dòng, số cột của dữ liệu (0.5 điểm)

	masv	ho	ten	gioitinh	ngaysinh	email	didong	cmnd	hocbong	makh
0	C0001	Khương Thảo	Loan	False	1999-01-04 00:00:00	ktloan@gmail.com	987314518	586900775484	2000000	CN
1	C0002	Đặng Bạch	Ngọc	True	1999-11-11 00:00:00	dbngoc@gmail.com	987587327	274387352269	0	CN
2	C0003	Phạm Văn Minh	Thiện	True	1999-11-23 00:00:00	pvmthien@gmail.com	987858734	835772714136	0	CN
3	C0004	Đinh Thị Thanh	Dung	False	1999-01-19 00:00:00	dttdung@gmail.com	987508413	373064334392	1000000	CN
4	C0005	Trần Mạnh	Thiện	True	1999-08-19 00:00:00	tmthien@gmail.com	987579776	233384596844	1000000	CN

	masv	ho	ten	gioitinh	ngaysinh	email	didong	cmnd	hocbong	makh
995	V0996	Trần Mạnh	Thăng	True	1999-06-21 00:00:00	tmthang@gmail.com	987250062	768656897616	2000000	VL
996	V0997	Trần Vĩnh	Thiện	True	1999-07-23 00:00:00	tvthien@gmail.com	987716502	468267857113	3000000	VL
997	V0998	Khương Thảo	Hạnh	False	1999-08-16 00:00:00	kthanh@gmail.com	913696608	702859343545	0	VL
998	V0999	Đinh Ngọc	Ngọc	True	1999-05-25 00:00:00	dnngoc@gmail.com	913689521	187831202224	3000000	VL
999	V1000	Lý Mạnh	Bảo	True	1999-06-03 00:00:00	lmbao@gmail.com	987656186	865173188581	0	VL

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   masv      1000 non-null   object
 1   ho        1000 non-null   object
 2   ten       1000 non-null   object
 3   gioitinh  1000 non-null   bool
 4   ngaysinh  1000 non-null   object
 5   email     1000 non-null   object
 6   didong    1000 non-null   int64
 7   cmnd      1000 non-null   int64
 8   hocbong   1000 non-null   int64
 9   makh      1000 non-null   object
dtypes: bool(1), int64(3), object(6)
memory usage: 71.4+ KB

Số dòng:  1000
Số cột:  10
```

3. Tạo DataFrame df_ket_qua từ tập tin ket_qua.csv (0.25 điểm)

	masv	mamh	diem
0	C0001	CSD1	7.0
1	C0001	CSD2	6.0
2	C0001	CTDL	9.0
3	C0001	KTLT	5.5
4	C0001	LQL1	7.0

4. Tạo DataFrame df_dtb bằng cách nhóm df_ket_qua theo masv và tính điểm trung bình cho mỗi sinh viên. (0.25 điểm)

	masv	diem
0	C0001	6.5
1	C0002	7.2
2	C0003	7.7
3	C0004	6.9
4	C0005	7.7

5. Trộn df và df_dtb theo masv và có kết quả của df như sau: (0.25 điểm)

	masv	ho	ten	gioitinh	ngaysinh	email	didong	cmnd	hocbong	makh	diem
0	C0001	Khương Thảo	Loan	False	1999-01-04 00:00:00	ktloan@gmail.com	987314518	586900775484	2000000	CN	6.5
1	C0002	Đặng Bạch	Ngọc	True	1999-11-11 00:00:00	dbngoc@gmail.com	987587327	274387352269	0	CN	7.2
2	C0003	Phạm Văn Minh	Thiện	True	1999-11-23 00:00:00	pvmthien@gmail.com	987858734	835772714136	0	CN	7.7
3	C0004	Đinh Thị Thanh	Dung	False	1999-01-19 00:00:00	dttdung@gmail.com	987508413	373064334392	1000000	CN	6.9
4	C0005	Trần Mạnh	Thiện	True	1999-08-19 00:00:00	tmthien@gmail.com	987579776	233384596844	1000000	CN	7.7

6. Cho biết **5 sinh viên có điểm lớn nhất của mỗi khoa**, sắp tăng theo mã khoa và giảm theo điểm(0.5điểm)

	masv	ho	ten	gioitinh	ngaysinh	email	didong	cmnd	hocbong	makh	diem
0	C0355	Văn Thị	Mai	False	1999-11-10 00:00:00	vtmai@gmail.com	987327541	855354458643	0	CN	8.75
1	C0387	Lê Bạch	Ngân	False	1999-04-29 00:00:00	lbngan@gmail.com	913158437	647119351103	0	CN	8.70
2	C0483	Lý Vĩnh	Thiệu	True	1999-06-21 00:00:00	lvthieu@gmail.com	913453657	432742077115	0	CN	8.70
3	C0154	Cao Văn Minh	Bảo	True	1999-05-11 00:00:00	cvmbao@gmail.com	913598140	530460808272	3000000	CN	8.60
4	C0052	Lê Thị Thảo	Ly	False	1999-02-14 00:00:00	lttly@gmail.com	168434476	874368763621	1000000	CN	8.55
5	T0508	Trần Thị Thảo	Nhung	False	1999-11-10 00:00:00	tttnhung@gmail.com	987228466	988514148472	1000000	TO	8.35
6	T0537	Văn Văn	Vỹ	True	1999-09-22 00:00:00	vvvy@gmail.com	913735282	876167216537	2000000	TO	8.25
7	T0584	Đàm Bạch	Ngân	True	1999-04-28 00:00:00	dbngan@gmail.com	168535534	165541702107	0	TO	8.25
8	T0513	Thái Văn Minh	Kiệt	True	1999-04-19 00:00:00	tvmkiet@gmail.com	913922643	753805518894	3000000	TO	8.20
9	T0557	Phạm Thị Thanh	Hoa	False	1999-07-18 00:00:00	ptthoa@gmail.com	913171650	396663865661	1000000	TO	8.15
10	V0846	Lê Văn	Lộc	True	1999-10-31 00:00:00	lvloc@gmail.com	913128685	658180634180	3000000	VL	8.65
11	V0799	Thái Nam	Tuấn	True	1999-03-07 00:00:00	tntuan@gmail.com	913950748	166625619025	1000000	VL	8.45
12	V0747	Mã Phú	Tâm	True	1999-08-25 00:00:00	mptam@gmail.com	168643676	335571036603	3000000	VL	8.45
13	V0700	Cao Thảo	Ly	False	1999-11-06 00:00:00	ctly@gmail.com	168435083	830725613394	3000000	VL	8.45
14	V0987	Trần Mạnh	Thiệu	True	1999-08-26 00:00:00	tmthieu@gmail.com	913149120	558722018236	2000000	VL	8.45

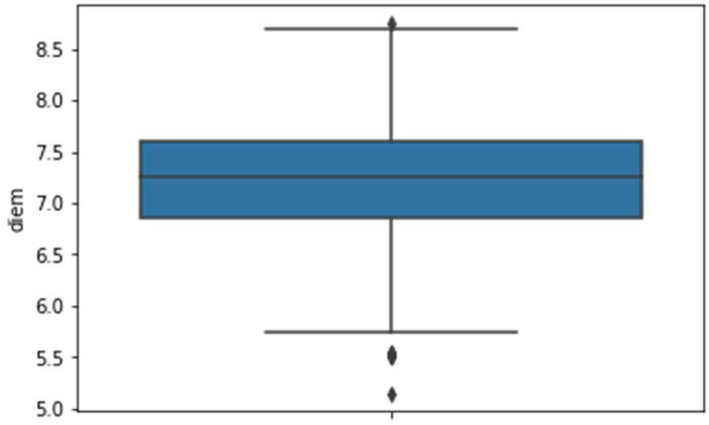
7. Trong DataFrame df tạo thêm cột ketqua với kết quả là giỏi nếu điểm>=9, kết quả là khá nếu điểm<9 và điểm >=6.5, kết quả là trung bình nếu điểm <6.5 và điểm >=5, kết quả là yếu nếu điểm<5 (0.25 điểm)

	masv	ho	ten	makh	diem	ketqua
0	C0001	Khương Thảo	Loan	CN	6.50	khá
1	C0002	Đặng Bạch	Ngọc	CN	7.20	khá
2	C0003	Phạm Văn Minh	Thiện	CN	7.70	khá
3	C0004	Đinh Thị Thanh	Dung	CN	6.90	khá
4	C0005	Trần Mạnh	Thiện	CN	7.70	khá
5	C0006	Trần Văn Minh	Hiệu	CN	7.80	khá
6	C0007	Lê Thị	Mai	CN	7.65	khá

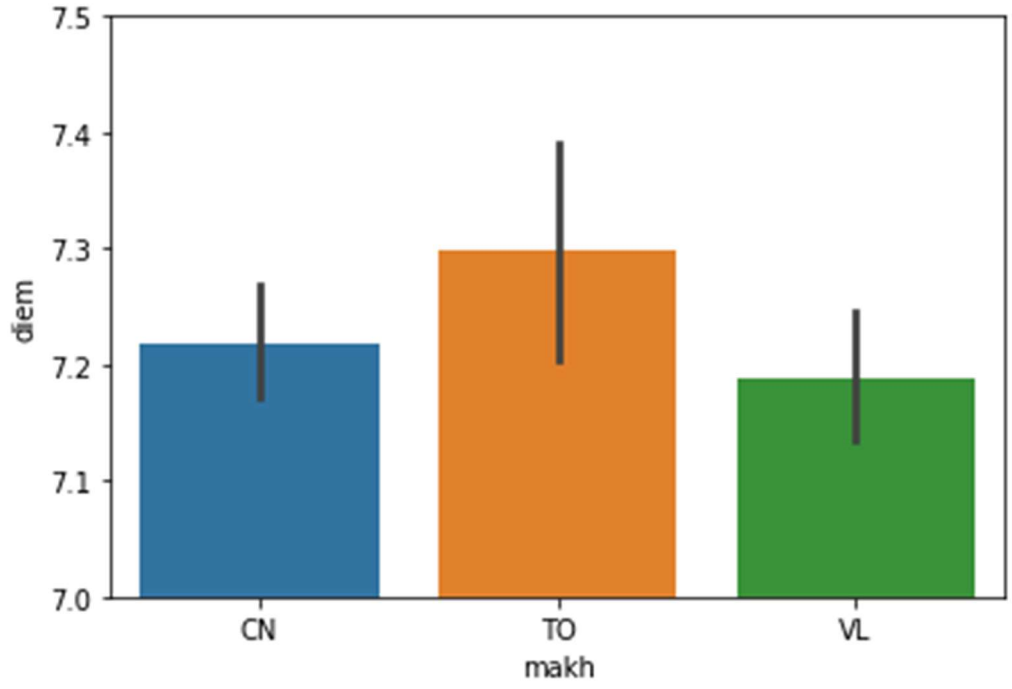
8. Tạo bảng tính điểm trung bình của mỗi khoa theo kết quả như mẫu: (0.25 điểm)

ketqua	khá	trung bình
makh		
CN	7.311874	6.176829
TO	7.360638	6.333333
VL	7.302095	6.210714

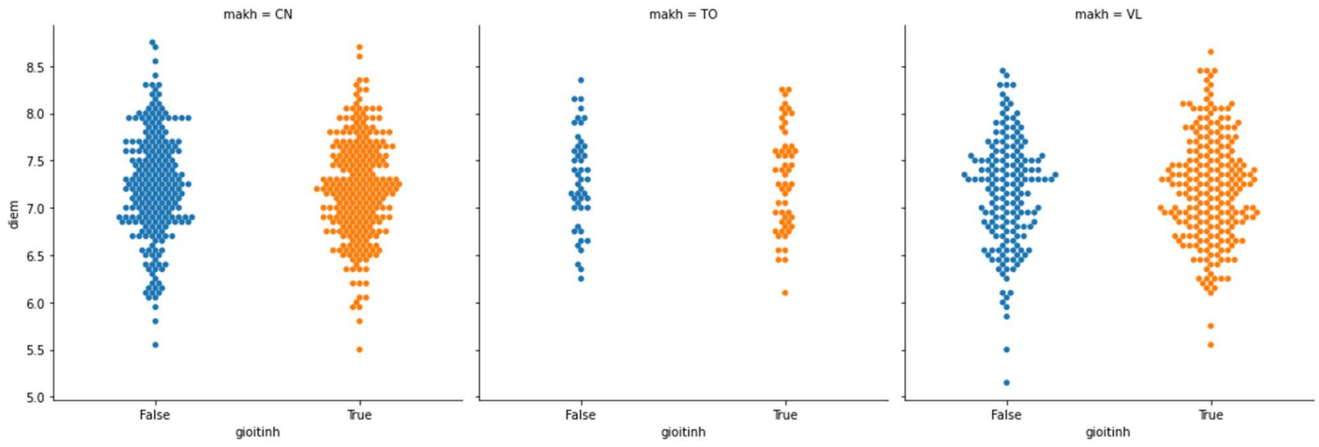
9. Vẽ biểu đồ boxplot với dữ liệu là DataFrame df. Chép các outlier ra df_outliers (0.5 điểm)



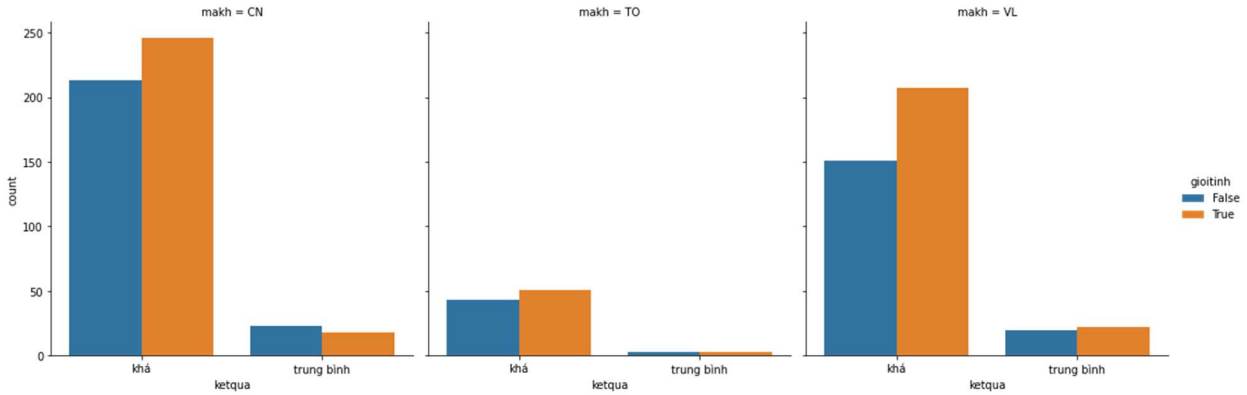
10. Vẽ biểu đồ barplot thể hiện điểm trung bình mỗi khoa theo mẫu và nhận xét (0.5 điểm)



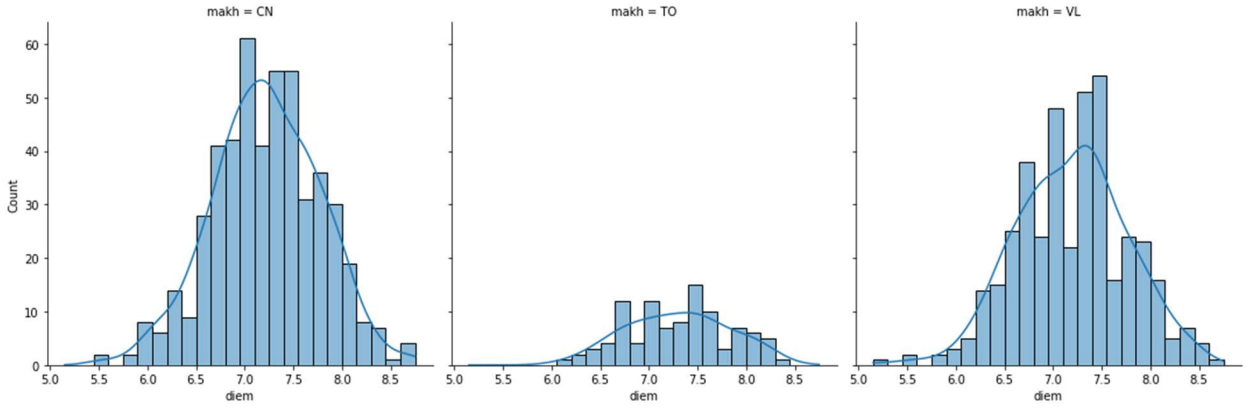
11. Vẽ biểu đồ thể hiện điểm số của sinh viên mỗi khoa theo giới tính như mẫu (0.25 điểm)



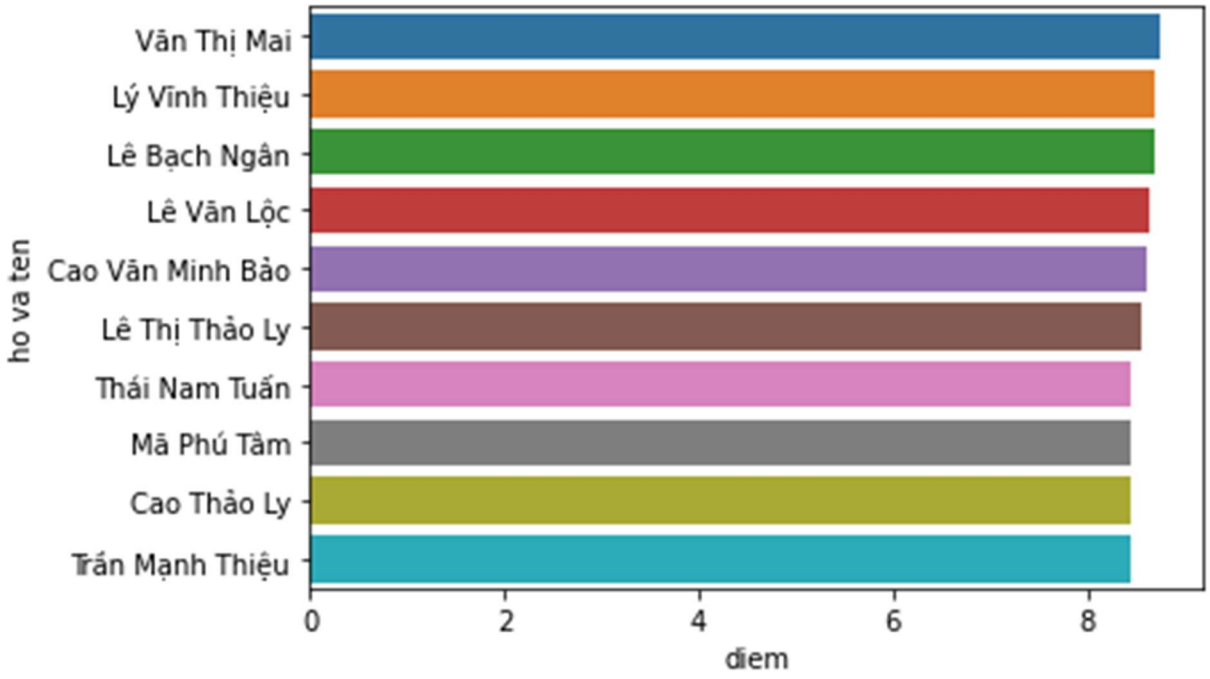
12. Vẽ biểu đồ đếm số sinh viên theo nhóm kết quả và giới tính **của mỗi khoa và nhận xét** (0.5 điểm)



13. Vẽ biểu đồ phân bố điểm của mỗi khoa theo mẫu sau và nhận xét: (0.25 điểm)

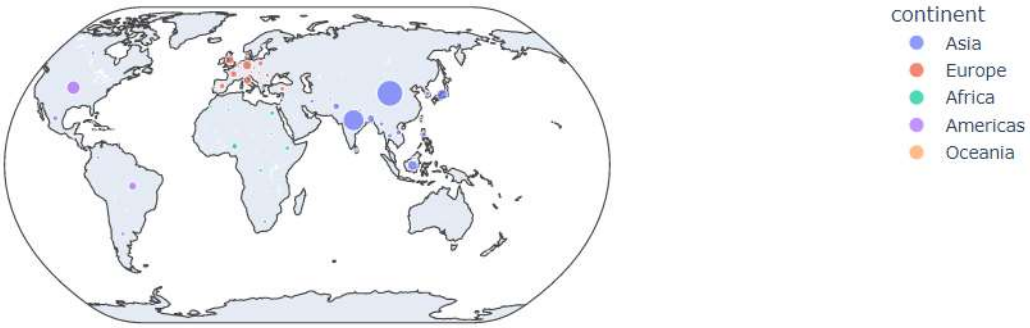


14. Lấy ngẫu nhiên 10 sinh viên của khoa CN (random_state = 0) rồi vẽ biểu đồ thể hiện sinh viên có điểm cao nhất, thấp nhất theo mẫu sau: (0.25 điểm)

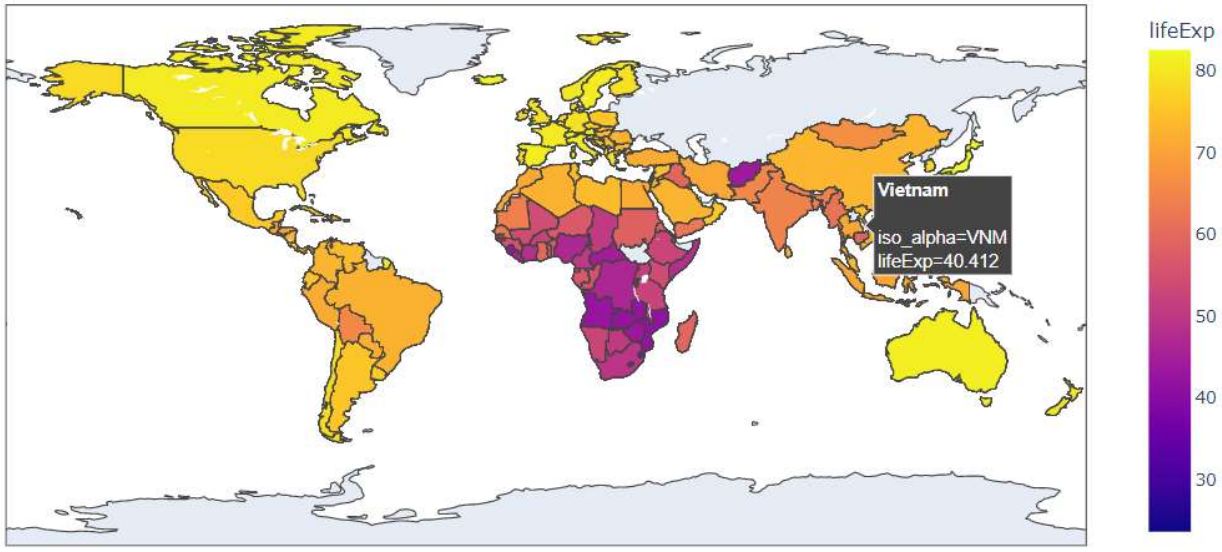


Câu 4: Trực quan hóa dữ liệu bản đồ (2 điểm)

1. Đọc dữ liệu **gapminder (có sẵn trên plotly)**, hiển thị thông tin chung của dữ liệu bao gồm: head, tail, info, describe (0.5 điểm)
2. Tạo scatter_geo map theo '**pop**' của từng country theo gợi ý như hình sau: (0.5 điểm)



3. Tạo Choropleth map theo lifeExp của từng country theo gợi ý như hình sau: (1.0 điểm)



--- Chúc các bạn làm bài tốt ☺---