Đề thi:

PYTHON FOR MACHINE LEARNING, DATA SCIENCE AND VISUALIZATION

Thời gian: 120 phút

Ngày thi : 08/05/2022

**** Học viên tạo 1 thư mục là **LDS2_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.
- Mỗi câu là 1 file viết trên Jupyter Notebook, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

1. Numpy Array (1.5 điểm)

- Yêu cầu: sử dụng thư viện Numpy thực hiện các yêu cầu sau :
  - Xây dựng hàm kiểm tra số nguyên tố **def kiem_tra_so_nguyen_to (so)** để kiểm tra số truyền vào có phải là số nguyên tố hay không (số nguyên tố là số lớn hơn 1, chỉ chia hết cho 1 và chính nó, ví dụ : 2,3,5,7,11,13….). Kết quả trả về True nếu là số nguyên tố, ngược lại trả về False. (0.5 điểm)
  - Phát sinh mảng 2 chiều có kích thước 4x4 với các phần tử có giá trị phát sinh ngẫu nhiên từ 1 đến 100 với np.random.seed(4). (0.5 điểm)
  - Kiểm tra và thay thế các phần tử trong mảng là số nguyên tố bằng giá trị phần tử xuất hiện nhiều nhất trong mảng. (0.5 điểm)
- Một số kết quả gợi ý :

```
Danh sách các phần tử được phát sinh ngẫu nhiên trong mảng:
[[47 56 70  2]
 [88 73 51 10]
 [59 95 56 56]
 [58 37 51 45]]

Phần tử xuất hiện nhiều nhất trong mảng là :  56
Mảng sau khi thay thế các pt là số nguyên tố :
[[56 56 70 56]
 [88 56 51 10]
 [56 95 56 56]
  [58 56 51 45]]
```

2. Reviews book (1.5 điểm)

Cho dữ liệu **Reviews_book.csv.csv** thực hiện các yêu cầu sau :

a. Đọc dữ liệu và tạo đoạn text từ cột **ReviewContent.** Sau đó thực hiện chuẩn hóa đoạn text bằng cách loại bỏ thêm các từ cho là không có ý nghĩa sau 'sea', 'girl', 'end', 'one', 'movie', 'reading', 'keep', 'put', 'end', 'time', 'author', 'wait', 'kept', 'book', 'character', 'story', 'rachel', 'twist', 'still', 'chapter', 'story', 'book', 'character', 'author', 'reading', 'gone', 'got', 'thing'. (0.5 điểm)

	ReviewContent
0	Good. It IS a page turner. You can read this b...
1	There are no words for how much I loathed this...
2	I think I would ordinarily cut this book more ...
3	Three disjointed characters for whom it's hard...
4	Was snookered into this novel as it was compar...

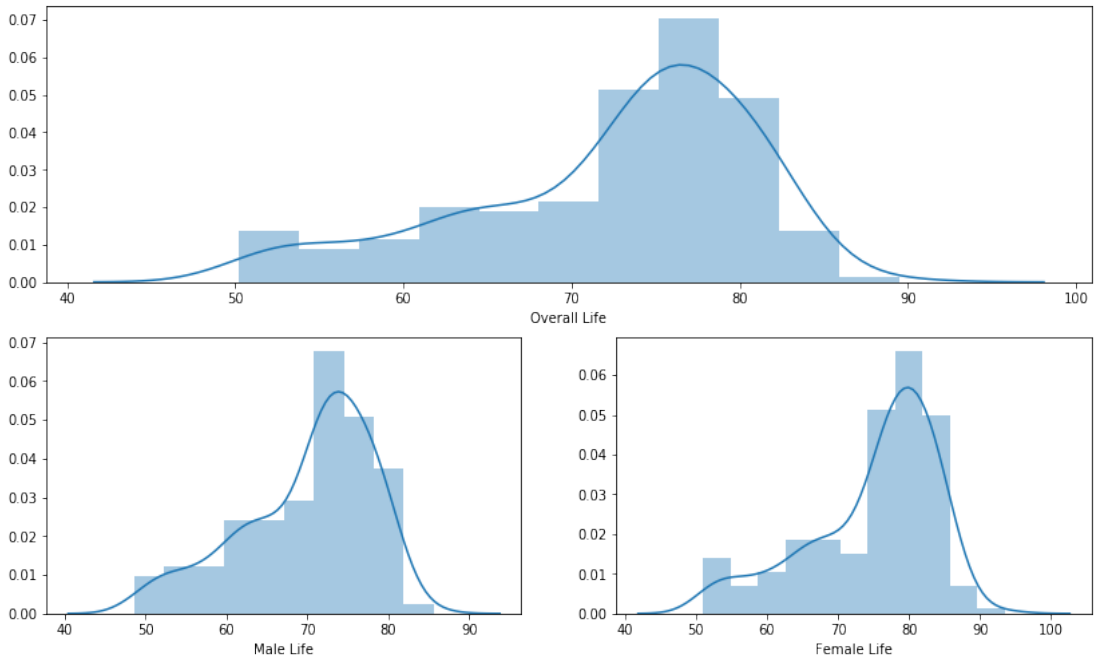
b. Tạo biểu đồ Wordcloud có kết quả gợi ý như sau : (0.5 điểm)

c. Cho tập tin hình ảnh **leaf1.png**, hãy tạo biểu đồ có kết quả gợi ý như hình sau : (0.5 điểm)

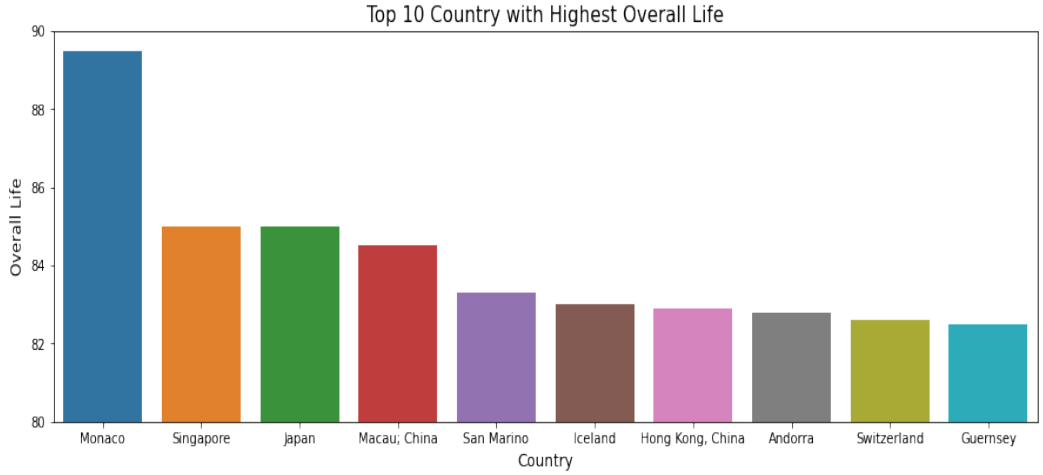
3. Life expectancy: (4 điểm)

- Cho dữ liệu **Life_expectancy_dataset.csv**, thực hiện các yêu cầu sau :

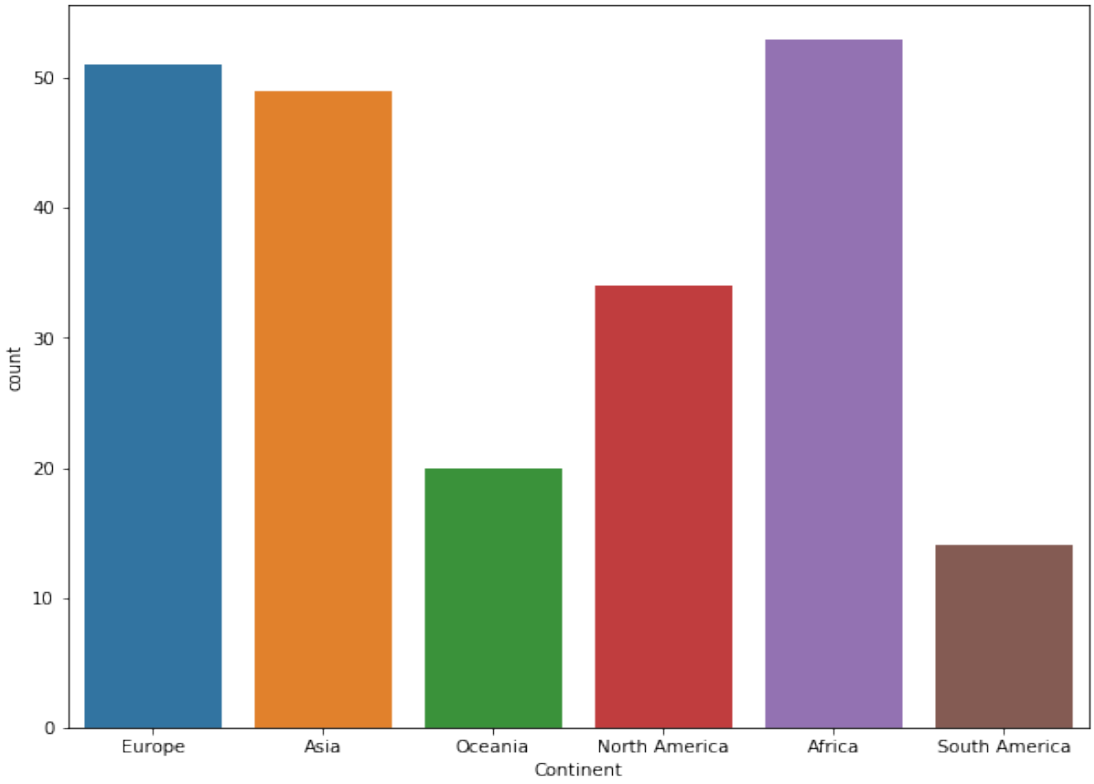
1. Đọc dữ liệu, hiển thị thông tin chung của dữ liệu : head, tail, info, describe. (0.25 điểm)
2. Kiểm tra xem dữ liệu có bị null ở cột nào hay không ? (0.25 điểm)
3. Thay giá trị nan bằng giá trị trung bình của cột đó. (0.25 điểm)
4. Vẽ biểu đồ thể hiện sự phân bố tuổi thọ của nam, nữ gợi ý như hình sau và nhận xét. (0.5 điểm)



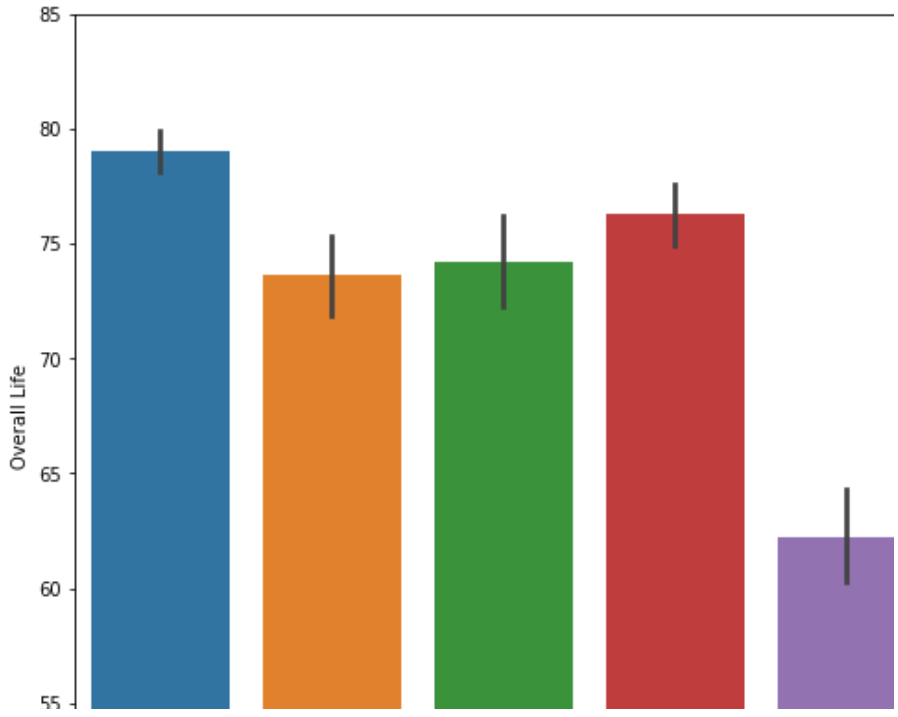
5. Cho biết top 10 quốc gia có tuổi thọ cao nhất. Vẽ biểu đồ như hình sau : (0.5 điểm)



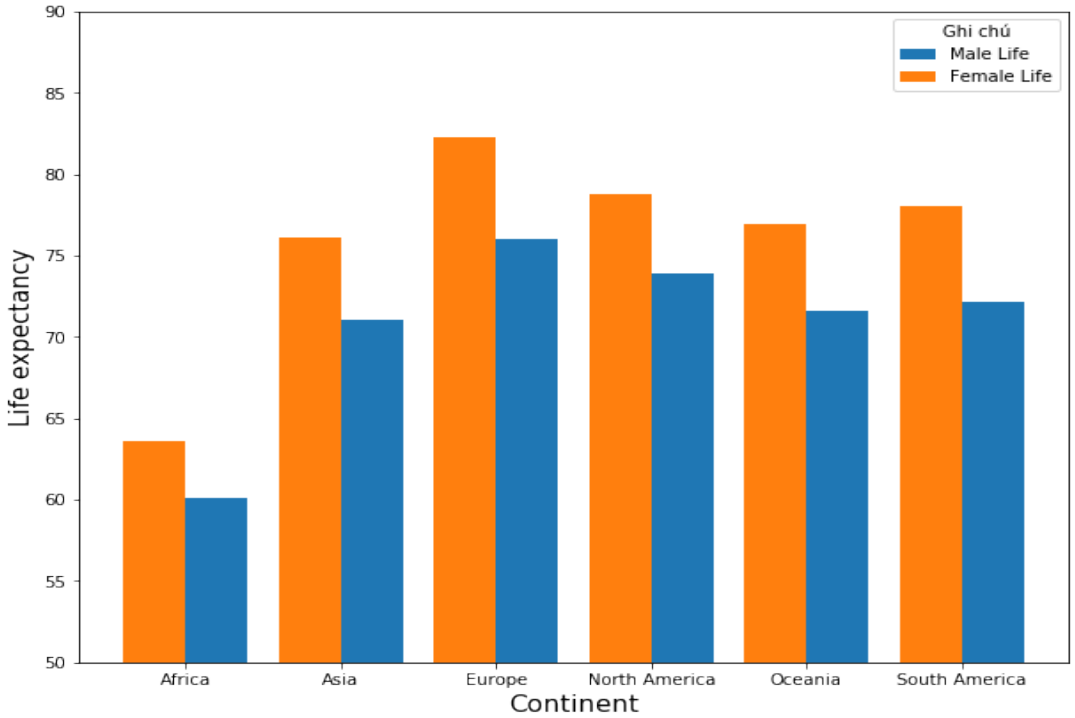
6. Vẽ biểu đồ cho biết số quốc gia trong mỗi châu lục như hình gợi ý. (0.25 điểm)



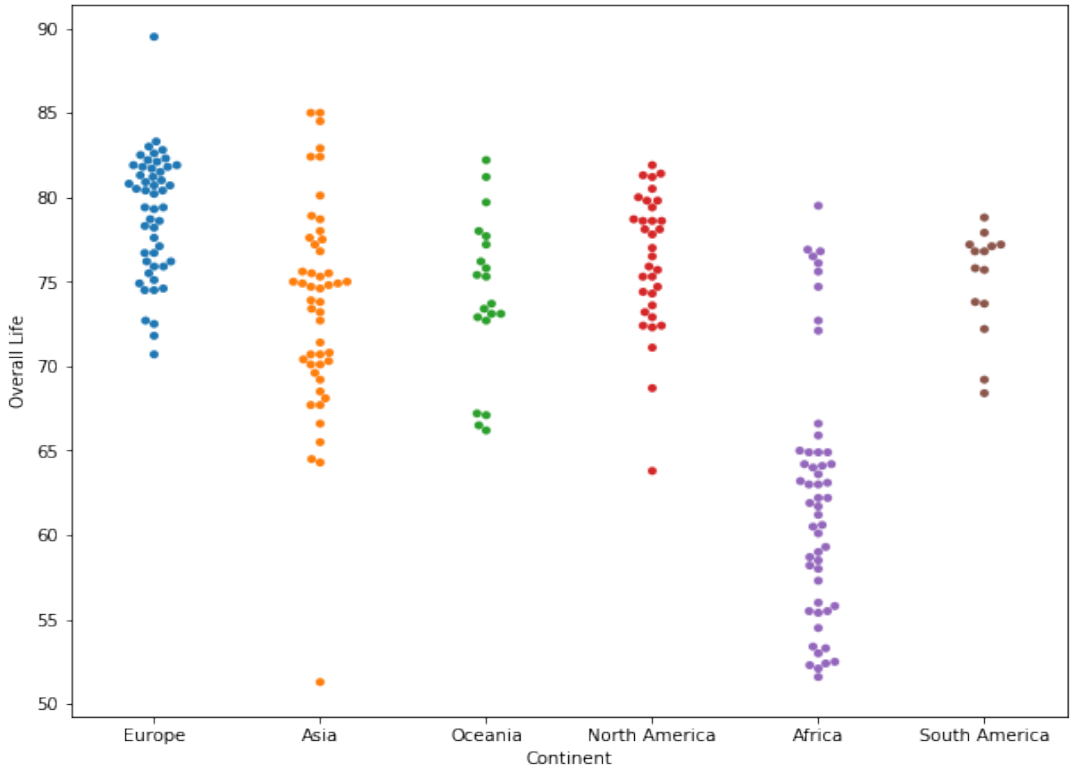
7. Vẽ biểu đồ thể hiện tỉ lệ tuổi thọ trung bình của mỗi châu lục như hình gợi ý và nhận xét : (0.5 điểm)



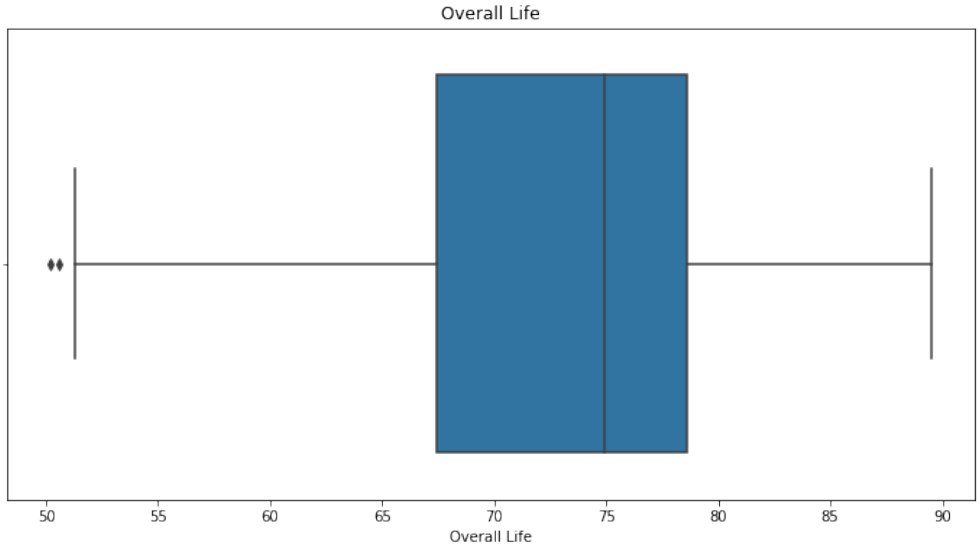
8. Vẽ biểu đồ so sánh tuổi thọ trung bình của nam và nữ ở mỗi châu lục. Bạn có nhận xét gì về biểu đồ này. (0.5 điểm)



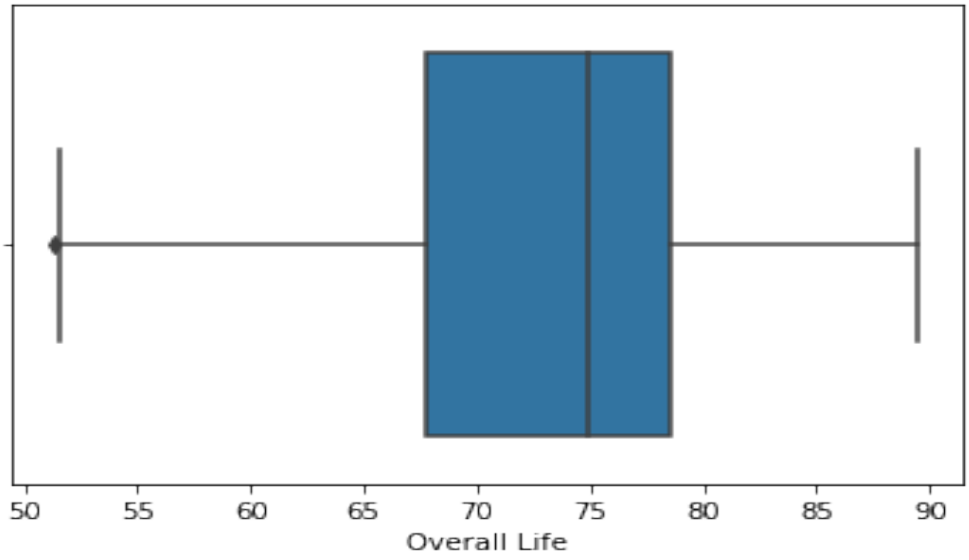
9. Vẽ biểu đồ thể hiện tuổi thọ chung ở các châu lục như gợi ý. Sau đó nhận xét về biểu đồ: (0.25 điểm)



10. Vẽ biểu đồ kiểm tra dữ liệu của cột 'Overall Life' gợi ý như hình sau : (0.25 điểm)



11. Dữ liệu của cột 'Overall Life' theo như hình trên có outliers hay không, nếu có thì loại bỏ tất cả các dòng trong data có outliers? (0.5 điểm)



4. Trực quan hóa dữ liệu bản đồ (3 điểm)

- Cho dữ liệu **World_Power_Consumption.csv** và **world-countries.json**, thực hiện các yêu cầu sau :

1. Đọc dữ liệu **World_Power_Consumption.csv**, hiển thị thông tin chung của dữ liệu bao gồm : head, tail, info, describe (0.75 điểm)

	Text
0	China 5,523,000,000,000
1	United 3,832,000,000,000
2	European 2,771,000,000,000
3	Russia 1,065,000,000,000
4	Japan 921,000,000,000

2. Tạo dataframe mới có 2 cột là **Country** và **Power_Consumption** được tách ra từ cột **Text** (0.5 điểm)

	Country	Power_Consumption
0	China	5,523,000,000,000
1	United	3,832,000,000,000
2	European	2,771,000,000,000
3	Russia	1,065,000,000,000
4	Japan	921,000,000,000

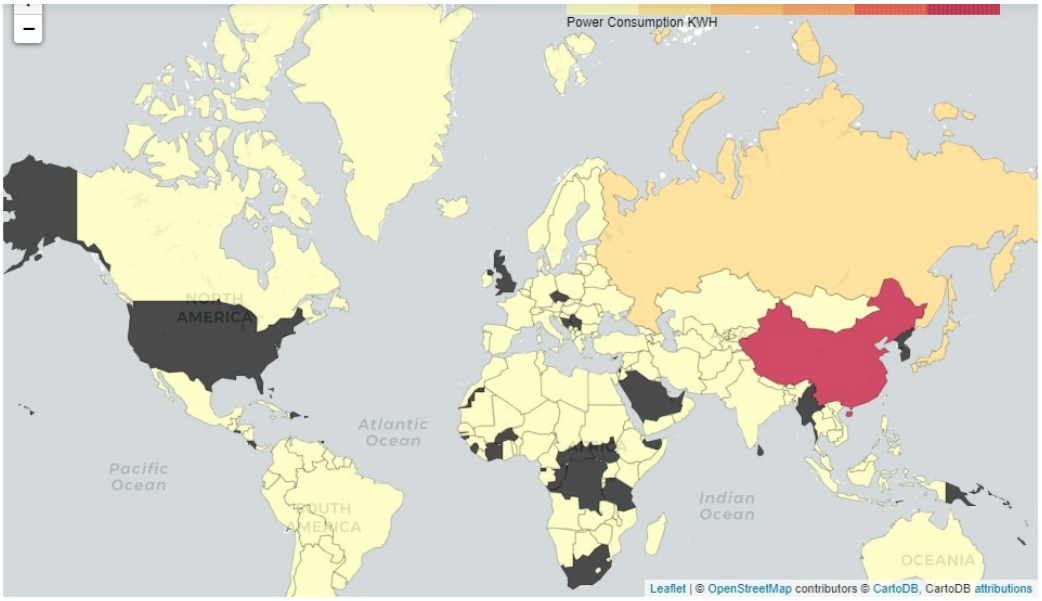
3. Chuyển đổi kiểu dữ liệu của cột **Power_Consumption** sang kiểu int64 (0.25 điểm)

```
1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 219 entries, 0 to 218
Data columns (total 2 columns):
Country              219 non-null object
Power_Consumption    219 non-null int64
dtypes: int64(1), object(1)
memory usage: 4.3+ KB
```

4. Tạo bản đồ có kiểu **cartodbpositron** với center (location=[0, 0]) và zoom level (zoom_start=3) gợi ý như hình sau : (0.75 điểm)

5. Tạo choropleth map theo **Power_Consumption** của từng quốc gia theo gợi ý như hình sau : (0.75 điểm)



--- Chúc các bạn làm bài tốt ☺ ---