DATA
VISUALIZATION

DATA RELATIONSHIP

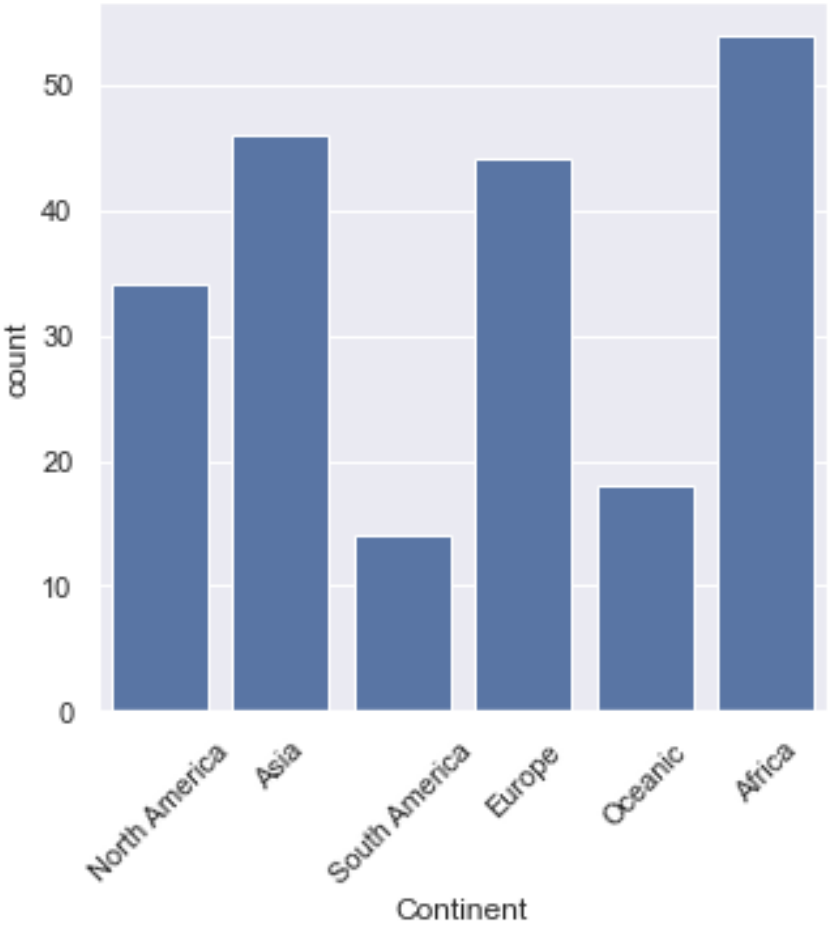
HUỲNH QUỐC DUY
19120494
LỚP 19CTT3

PHÂN TÍCH DỮ LIỆU

ĐƠN BIẾN

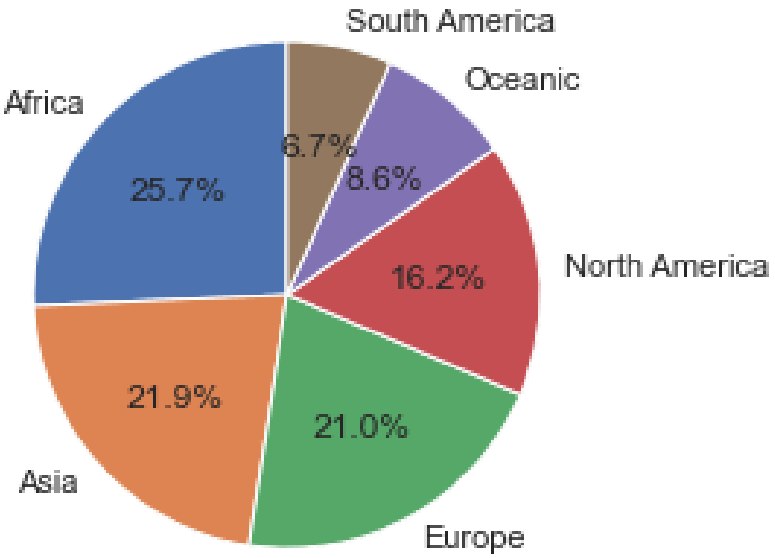
DATA VISUALIZATION

Cột Continent

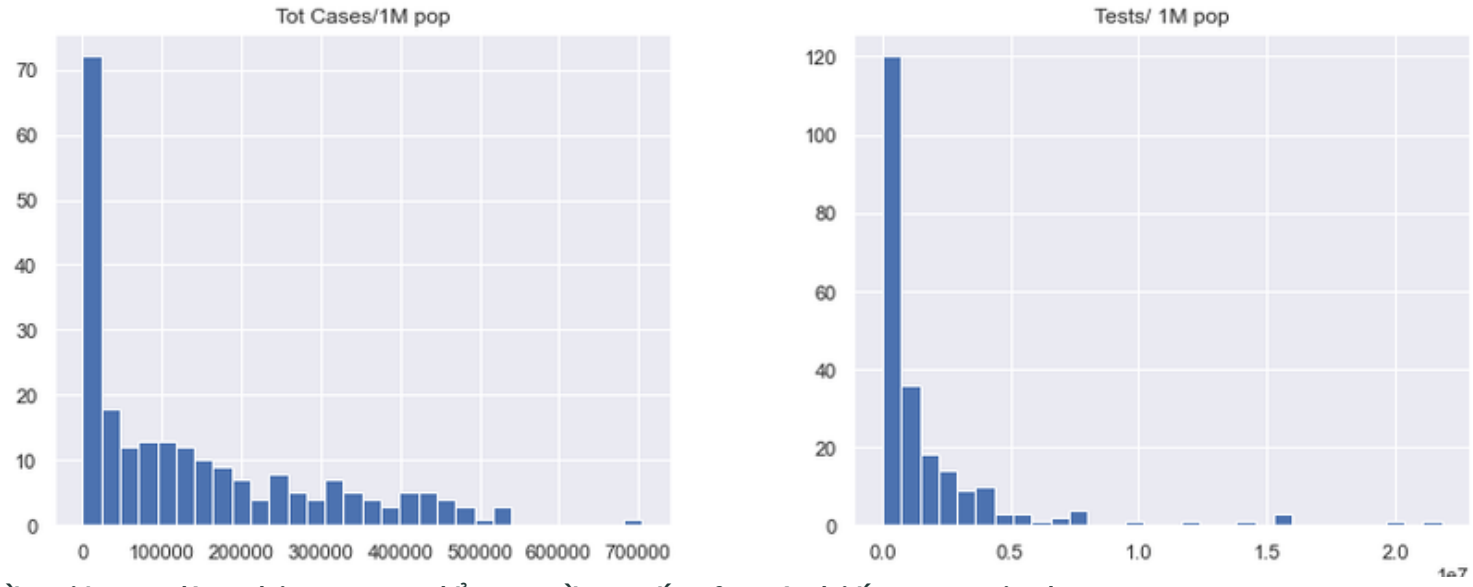


- Count plot để xem số lượng các quốc gia ở mỗi châu lục
- Ta thấy châu Phi có nhiều quốc gia nhất, châu Đại Dương có nhiều quốc gia hơn cả Nam Mỹ, có thể là vì tính cả các quần đảo

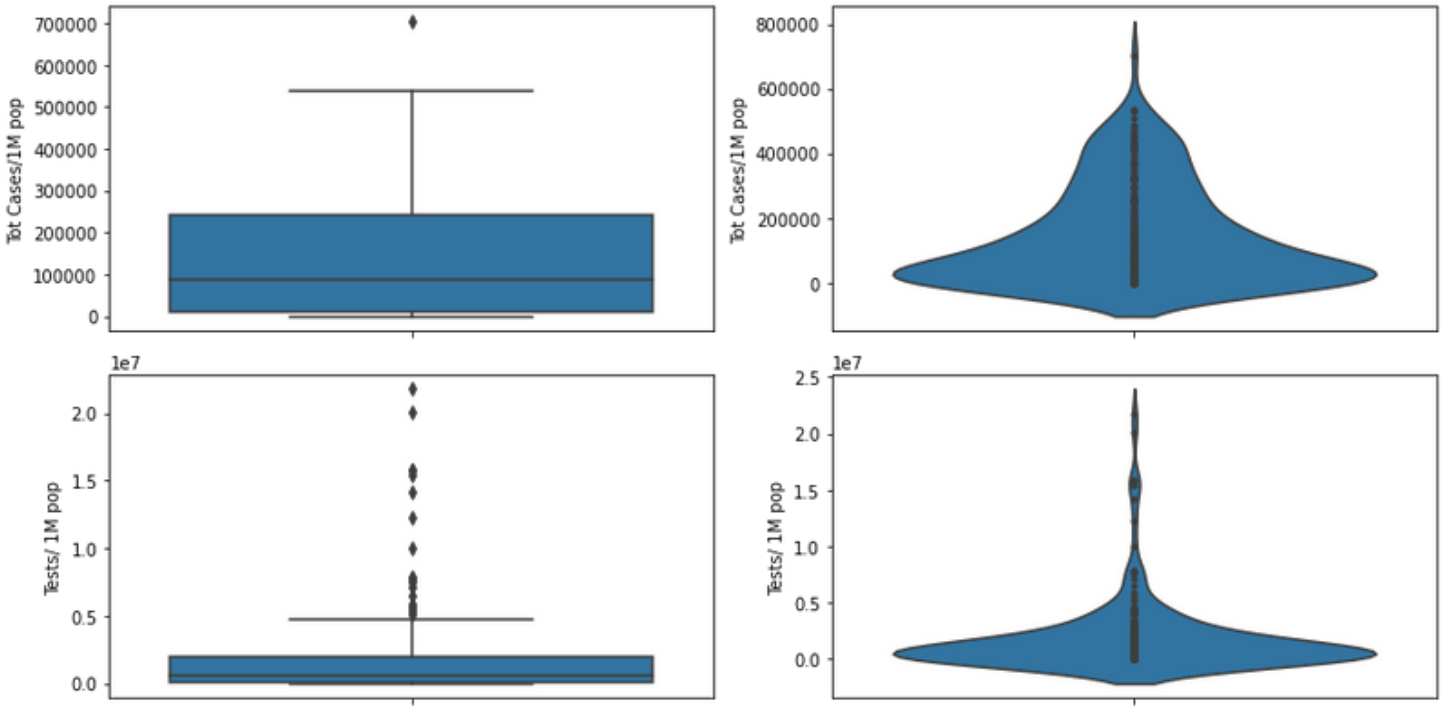
- Pie chart để xem phân bố số lượng quốc gia thuộc mỗi châu lục
- Châu Phi, châu Á và châu Âu chiếm gần 3/4 số lượng quốc gia trên thế giới. Riêng châu Phi có số lượng quốc gia bằng 1/4 toàn bộ quốc gia.



Cột Tot Cases/1M pop, cột Tests/ 1M pop



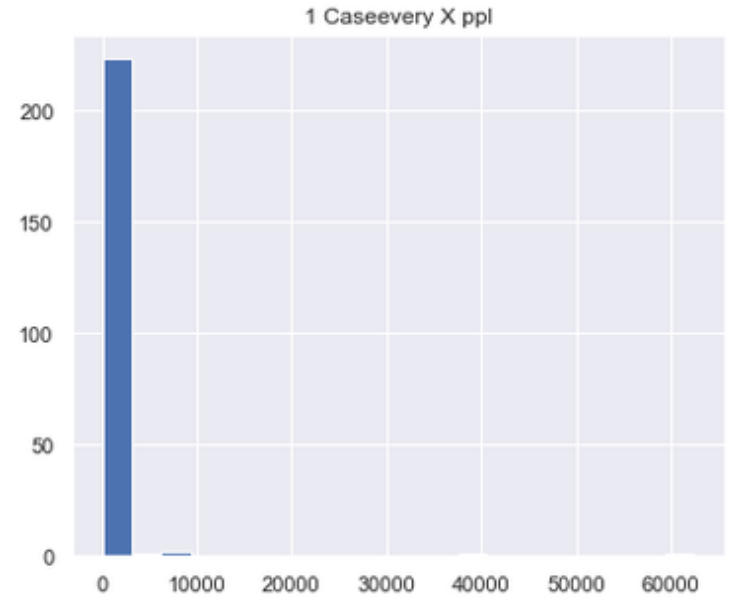
- Đầu tiên ta dùng histogram để xem tần suất của các biến numerical
- Có thể thấy phần lớn các nước đều có tỉ lê tổng số ca trên 1 triệu dân vào khoảng từ 0 đến 20000.
- Tương tự, phần lớn các nước đều có tỉ lệ tổng số lần test trên 1 triệu dân vào khoảng từ 0 đến 5 triệu lần test. Bên cạnh đó, các nước có số lần test trên 1 triệu dân dưới 100,000 là rất nhiều (120 nước)

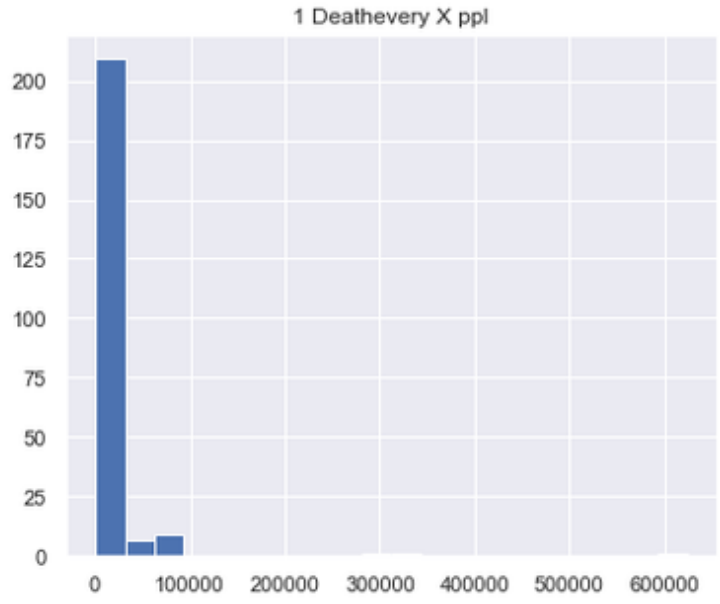


- Tiếp theo ta dùng box plot và violin plot để xem phân bố của các thuộc tính
- Đối với thuộc tính đầu tiên, ta có thể khẳng định kết luận từ biểu đồ histogram là đúng khi tỉ lệ tổng số ca trên 1 triệu dân nằm hầu hết từ 0 đến 2 triệu dân. Có 120 nước có thấp hơn 100,000 ca trên 1 triệu dân
- Biểu đồ của thuộc tính thứ 2 cũng hầu như tập trung ở dưới, ngoài ra cũng có nhiều quốc gia là outlier xuất hiện với số lượng test lớn hơn vượt trội. Số lượng các quốc gia có con số lớn hơn giá trị lớn nhất trong box plot là tận 124 nước

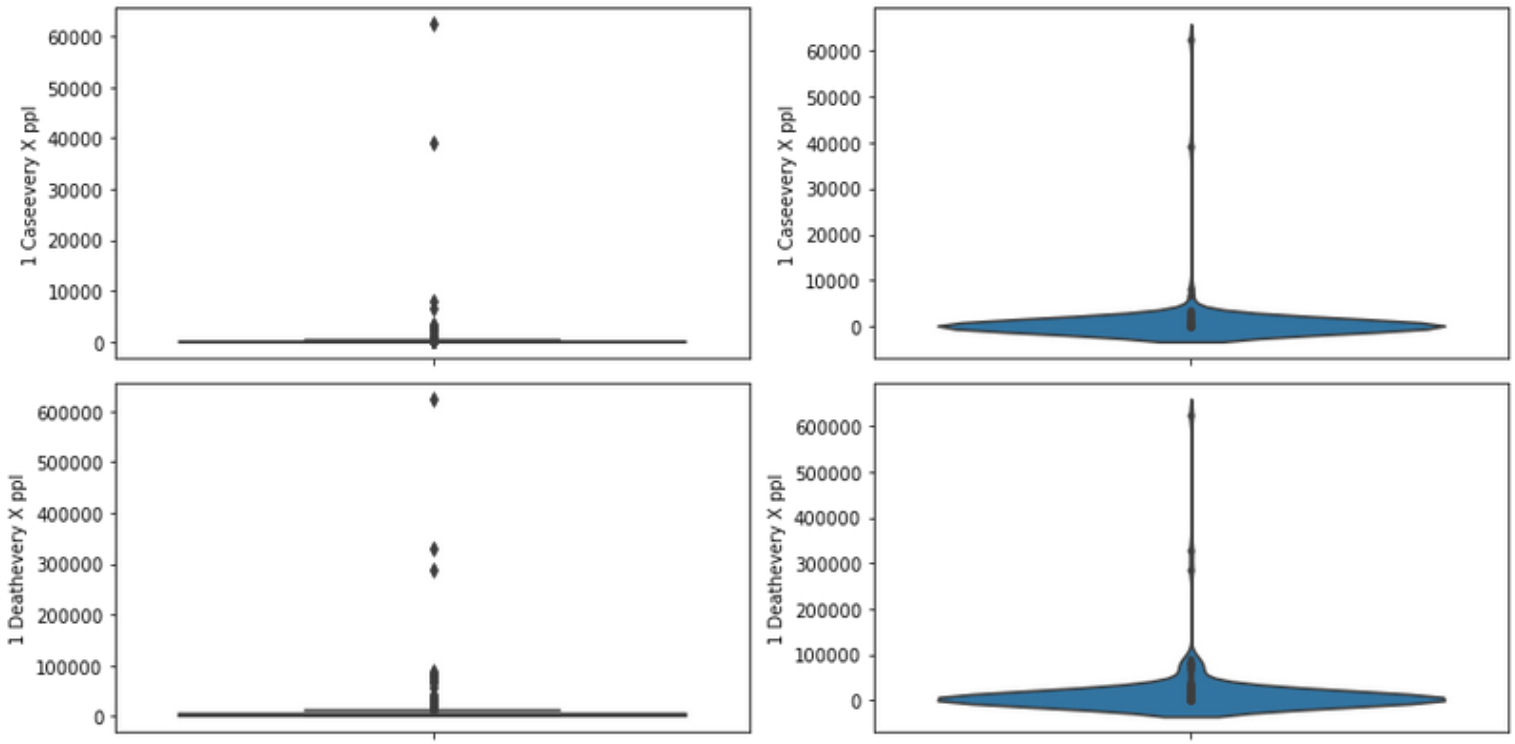
Cột 1 Caseevery X ppl, cột 1 Deathevery X ppl

Tương tự, ta dùng histogram để xem tần suất phân bố của các thuộc tính





- Hầu như phân bố của các thuộc tính được xét đều lệch về phía bên trái của biểu đồ (lệch về phía của 0). Các phân bố này không đều vì có sự xuất hiện của các outlier với giá trị vượt trội



- Có thể thấy phân bố của 2 thuộc tính này còn lệch hơn nhiều so với 2 thuộc tính ta xét ở trên. Ở biểu đồ boxplot ta còn không thấy được khoảng các giữa các phân vị

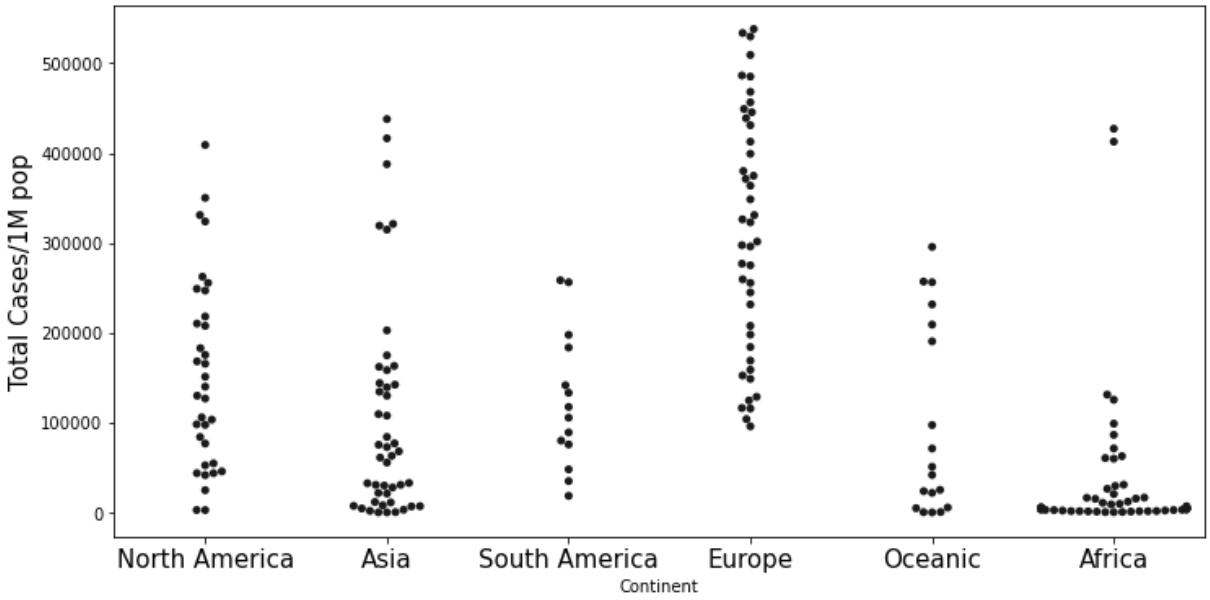
PHÂN TÍCH DỮ LIỆU

QUAN HỆ ĐA BIẾN

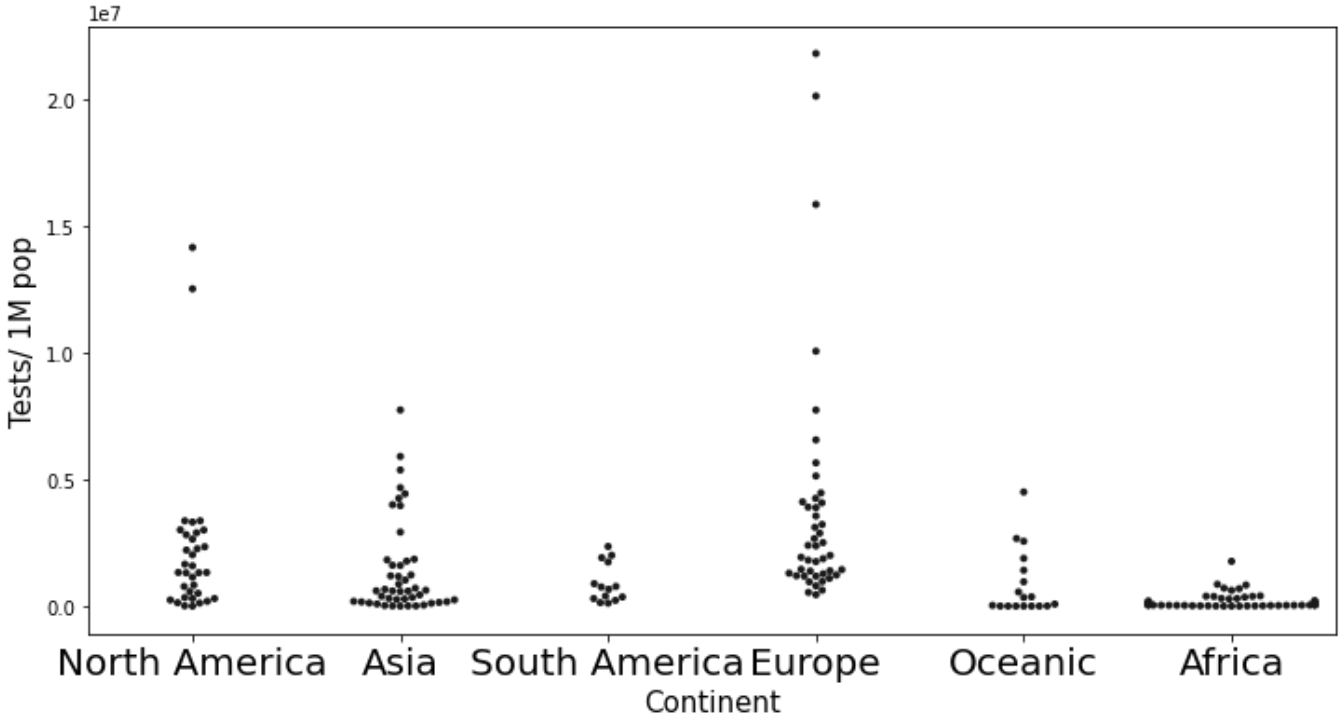
DATA RELATIONSHIP

DATA VISUALIZATION

Biến continent và các biến khác

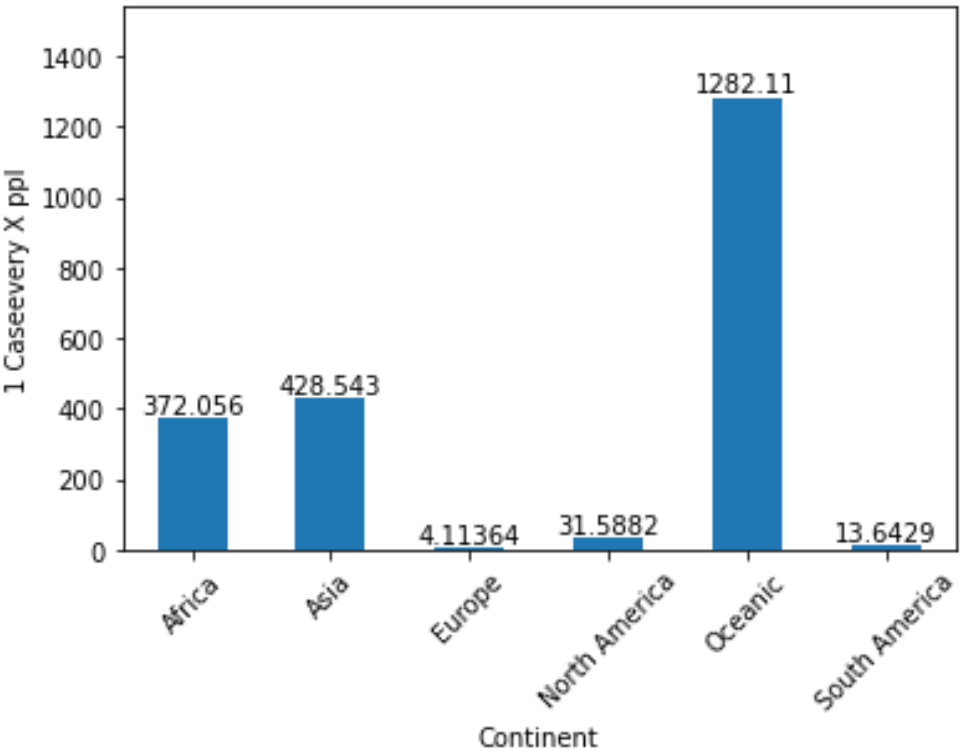


- Ta dùng swarm plot để xem sự phân bố của Tot Cases/1M pop theo từng châu lục
- Dễ thấy rằng ở châu Âu, số ca mắc Covid là vượt trội so với phần còn lại, ngay cả quốc gia có tỉ lệ số ca mắc trên 1 triệu người dân thấp nhất ở châu Âu cũng cao hơn hầu hết các nước ở châu Phi
-Từ biểu đồ ta thấy được các nước ở châu Phi có tỉ lệ số ca mắc trên 1 triệu người dân là vô cùng thấp, các giá trị đều tập trung ở sát đáy trục

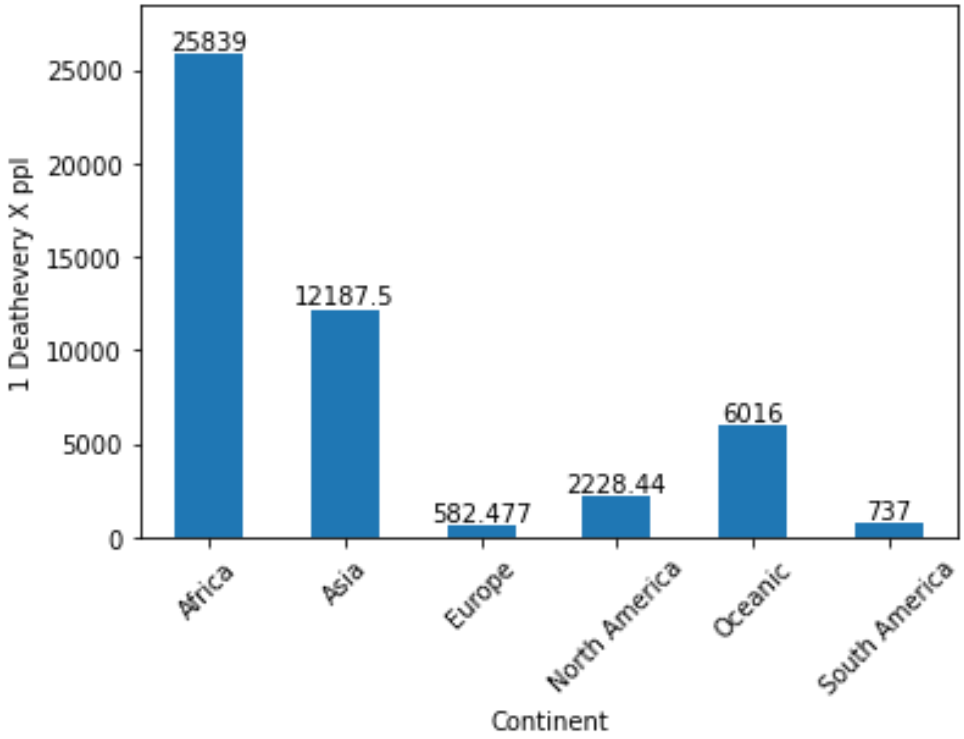


- Về số lần test thì ngoài một vài quốc gia có số lần test trên 1 triệu dân cao vượt trội thì hầu như phân bố của châu Âu cũng không quá chênh lệch với Bắc Mỹ và châu Á
- Ngoài ra để biết được phần nào tình hình hiện giờ thì ta cũng xem sự phân bố số ca đang bị nhiễm Covid trên 1 triệu dân ở các châu lục

Với các biến *1 Caseevery X ppl* và *1 Deathevery X ppl* thì ta không thể cộng dồn rồi so sánh chúng giữa các châu nên ta sẽ sử dụng giá trị trung bình của từng châu

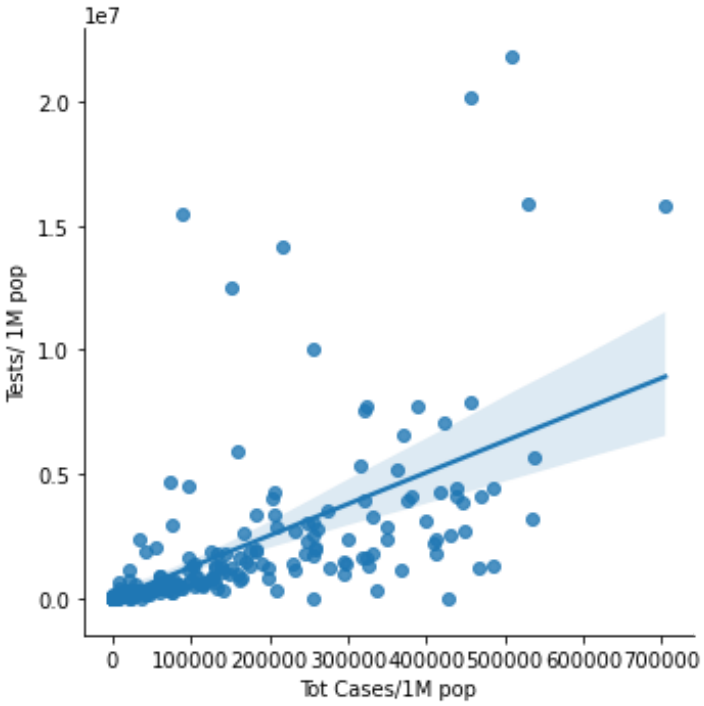


- Châu Đại Dương có tỉ lệ 1 Caseevery X ppl cao nhất, tức là trung bình cứ khoảng 2500 người mới có một người mắc
- Châu Âu, Bắc và Nam Mỹ có tỉ lệ trên rất thấp. Điều này có nghĩa là ở các châu này, cứ khoảng từ 10 đến 30 người là có 1 người mắc phải covid



- Ở châu Phi cứ 25832 người mới có một người chết vì Covid, điều này là rất đáng mừng. Tuy nhiên, các con số này vẫn có khả năng không chính xác vì nền khoa học ở châu Phi nhiều nơi còn lạc hậu
- Ở Bắc Mỹ và Nam Mỹ, con số là rất báo động . Do đó ở những nơi này cần phải đề cao phòng bị

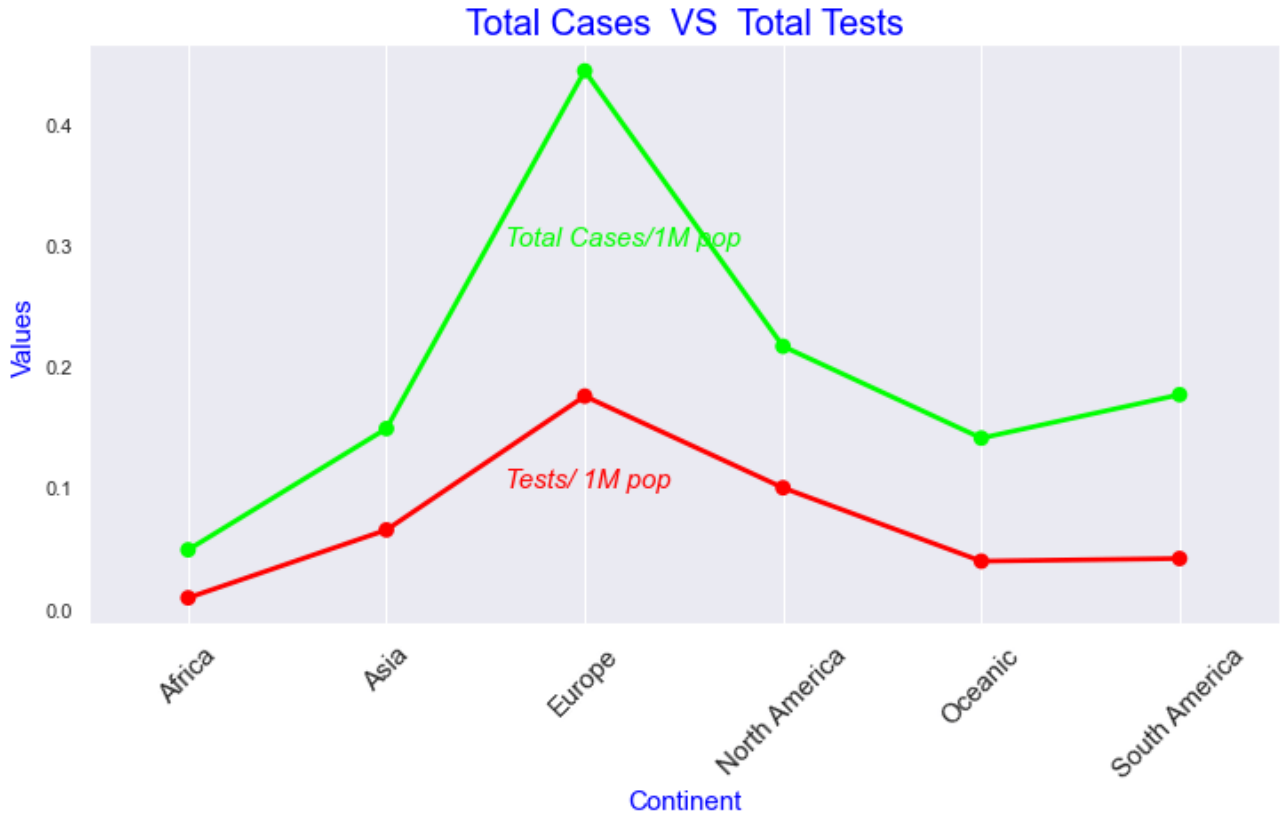
(Tot Cases/1M pop - Tests/ 1M pop)



Để tìm mối tương quan giữa Tot Cases/1M pop và Tests/ 1M pop, đầu tiên ta vẽ scatter plot của 2 thuộc tính cùng với một đường hồi quy

Từ heatmap ở trên ta biết được mối quan hệ giữa 2 thuộc tính là 0.61. Đường hồi quy vẽ từ các điểm trong biểu đồ cho thấy mối tương quan giữa 2 thuộc tính phần nào đó là tuyến tính

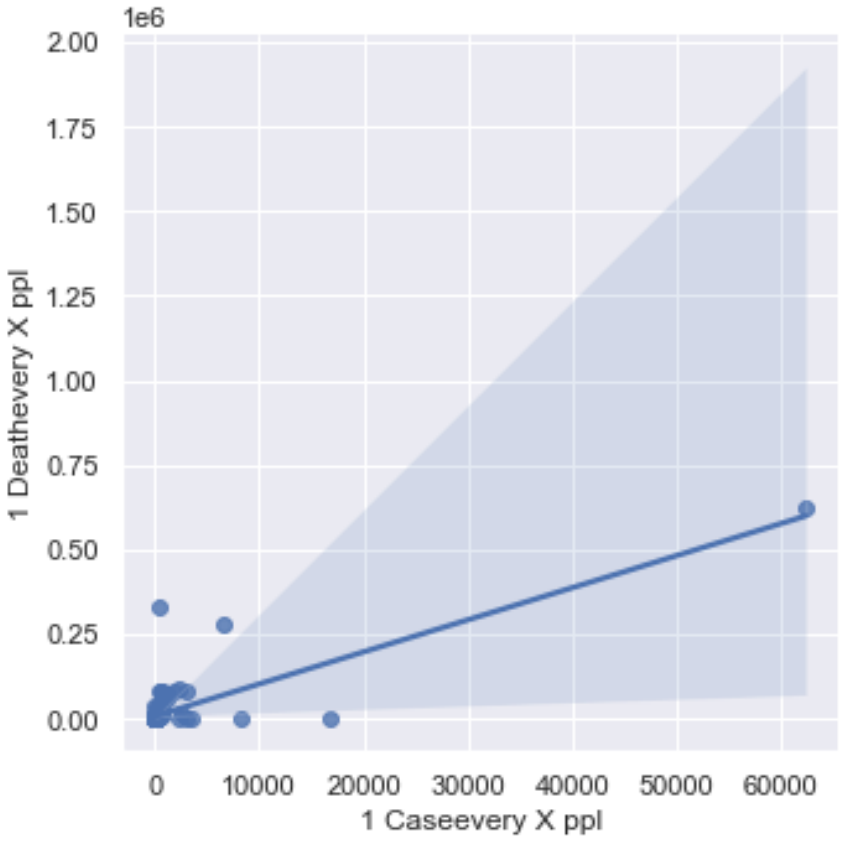
Tiếp theo ta chuẩn hóa MinMax cho 2 thuộc tính rồi biểu diễn bằng point plot để xem được sự thay đổi của chúng theo từng châu lục



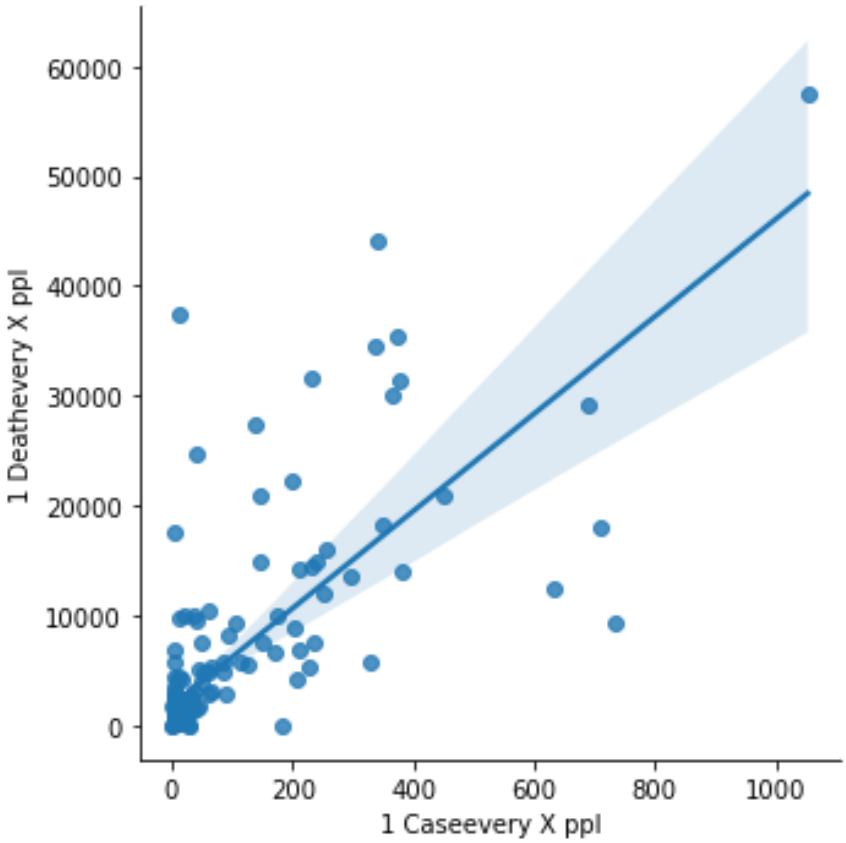
- Ở các châu lục có số lượng test cao thì tổng sô ca nhiễm cũng cao (châu Âu, Bắc Mỹ). Vì đây là tỉ lệ trên 1 triệu dân nên số liệu này khá khách quan (không bị ảnh hưởng bởi dân số từng vùng).
- Các châu lục có lượng lượt test thấp thì tổng số ca nhiễm cũng không cao (châu Phi, châu Đại Dương). Do đó những nơi này có số ca nhiễm thấp không hẳn là vì đã kiểm soát được dịch bệnh mà có thể là do người dân không thực hiện test covid

(1 Caseevery X ppl - 1 Deathevery X ppl)

Để tìm mối tương quan giữa Tot Cases/1M pop và Tests/ 1M pop, đầu tiên ta vẽ scatter plot của 2 thuộc tính cùng với một đường hồi quy.



Tuy nhiên dữ liệu của 2 thuộc tính chứa quá nhiều outlier nên ta sẽ loại bỏ các giá trị lớn hơn phân vị 95% rồi mới thực hiện vẽ laij scatter plot



Có thể thấy lúc này mối quan hệ giữa 2 thuộc tính đã là tuyến tính khi đường hồi quy khá đẹp. Giá trị tương quan lúc này đã tăng từ 0.69 lên 0.74

NGUỒN THAM KHẢO

SLIDE

Slide bài giảng GV. Bùi Tiến Lên

BOOK

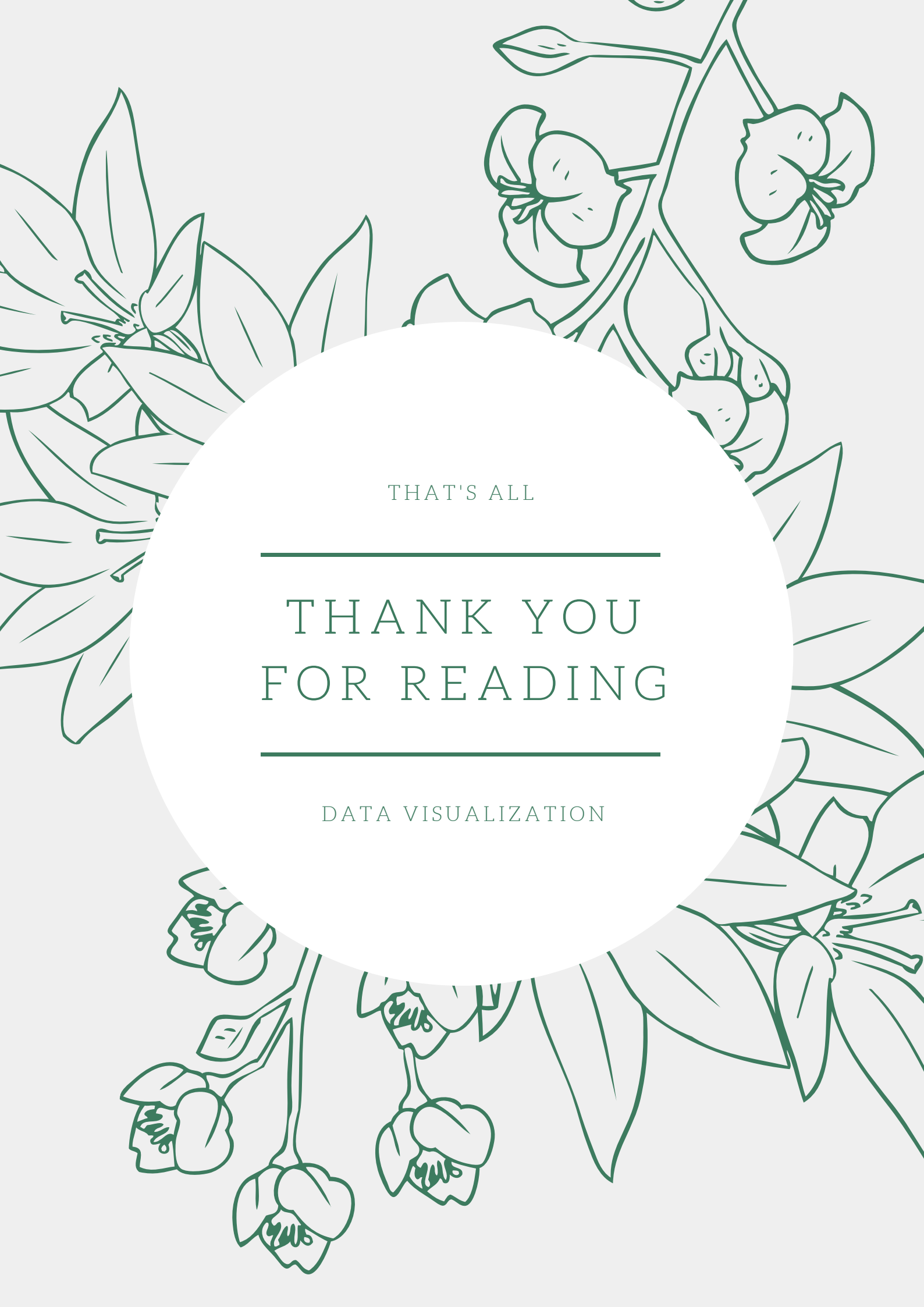
Data analysis của thầy Bùi Tiến Lên

KAGGLE

EDA with Seaborn

BOOK

Python for Data Science handbook


THAT'S ALL
THANK YOU
FOR READING
DATA VISUALIZATION