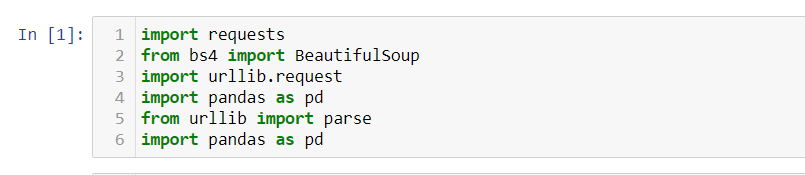
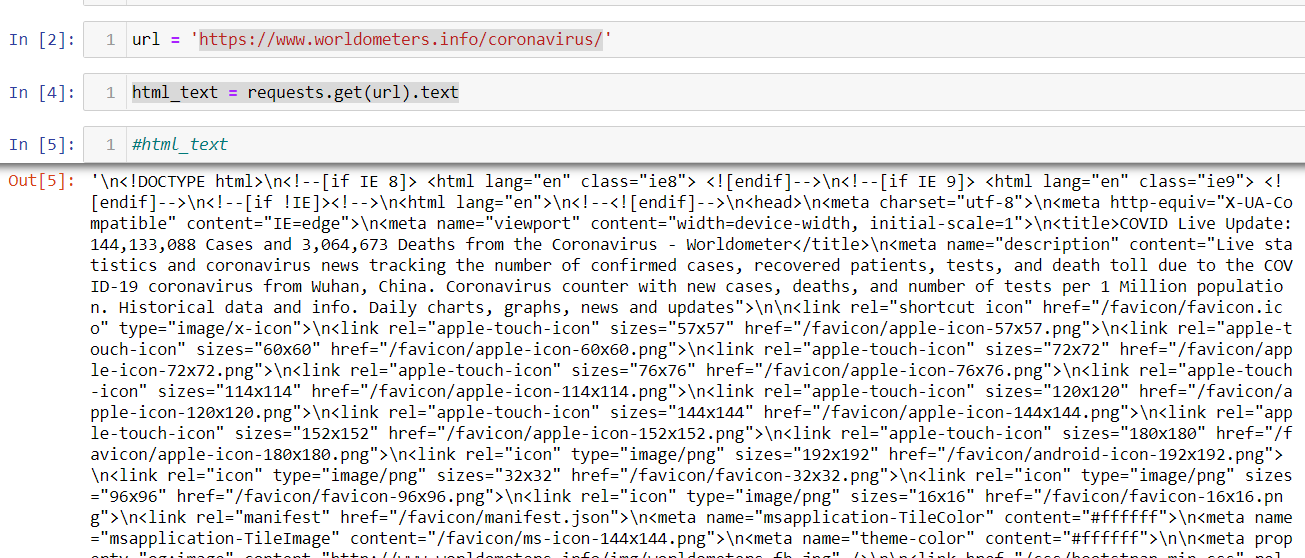
Báo cáo

1. Thu thập dữ liệu:

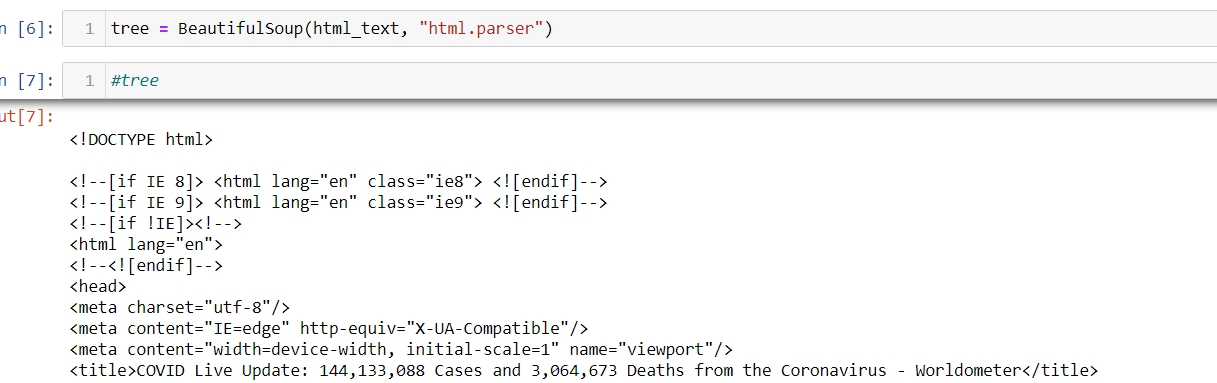
* Hình thức thu thập: parser html và BeautifulSoup.
* Các thư viện sử dụng:



* Các bước thu thập:
  + Bước 1: Gởi 1 requests đến url cần lấy dữ liệu, cụ thể url của bài này là: <https://www.worldometers.info/coronavirus/>.



* + Bước 2: Tạo một cây (tree) html bằng BeautifullSoup.



* + Bước 3: Lấy dữ liệu:

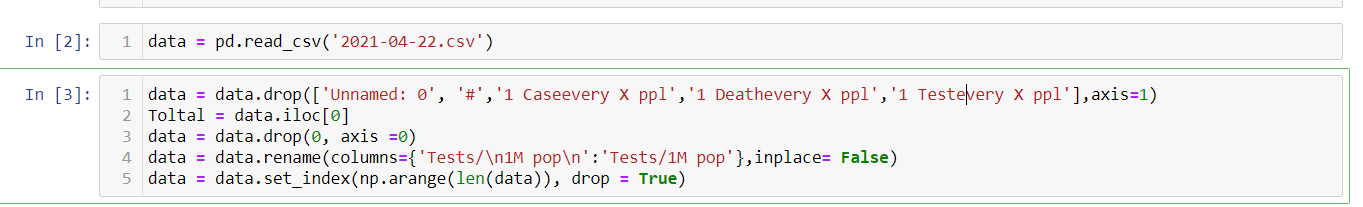


Lưu dữ liệu vào một DataFrame để lưu vào file:

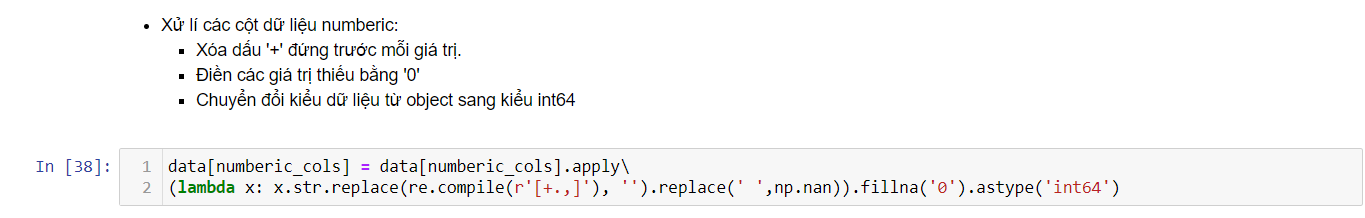


1. Tiền xử lí dữ liệu:

* Xóa một số thuộc tính không cần thiết và sửa tên của thuộc tính Tests/1M pop



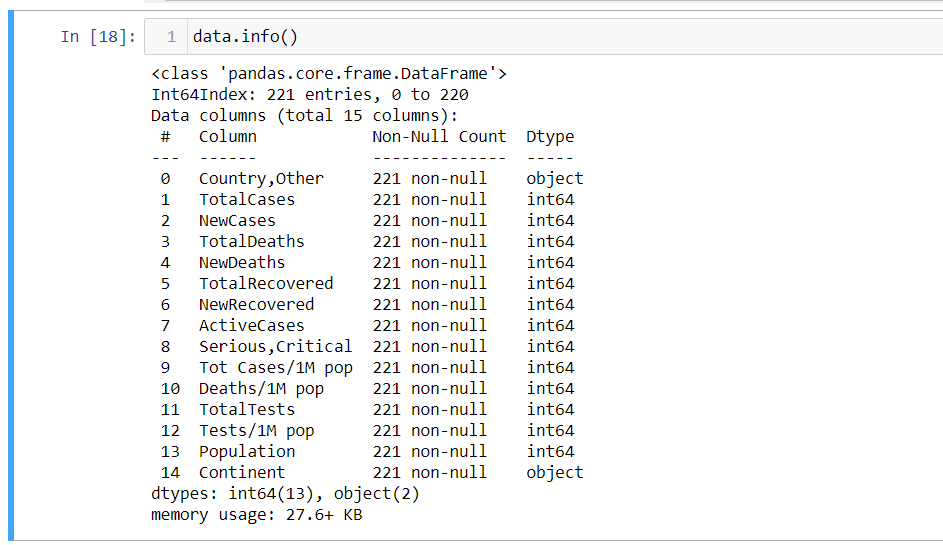
* Điền dữ liệu thiếu và chuyển đổi kiểu dữ liệu cho các thuộc tính số:



* Điền dữ liệu thiếu ở các thuộc tính categrical

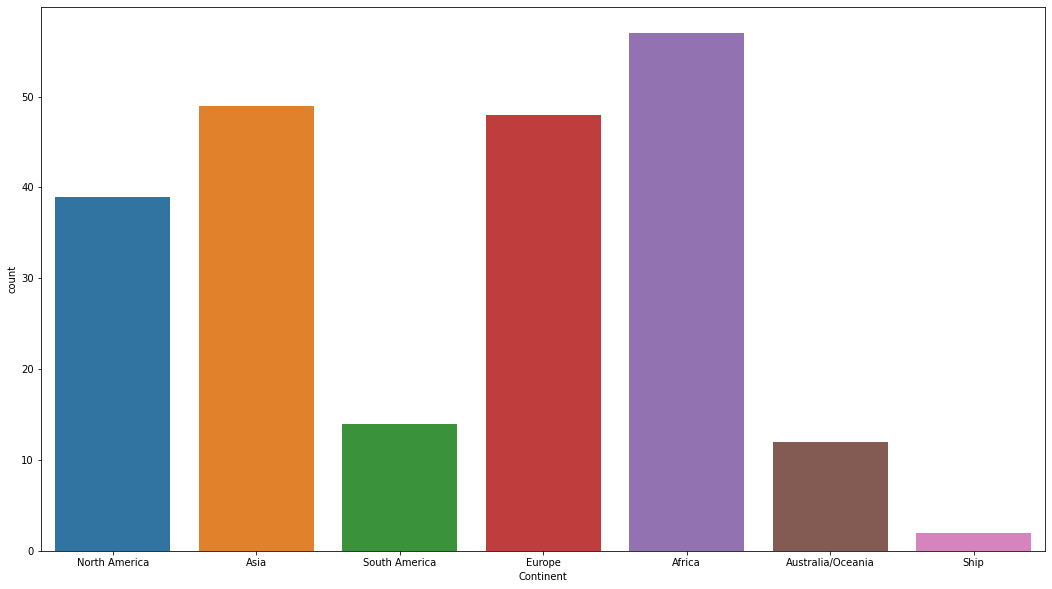


* Dữ liệu sau khi tiền xử lí:

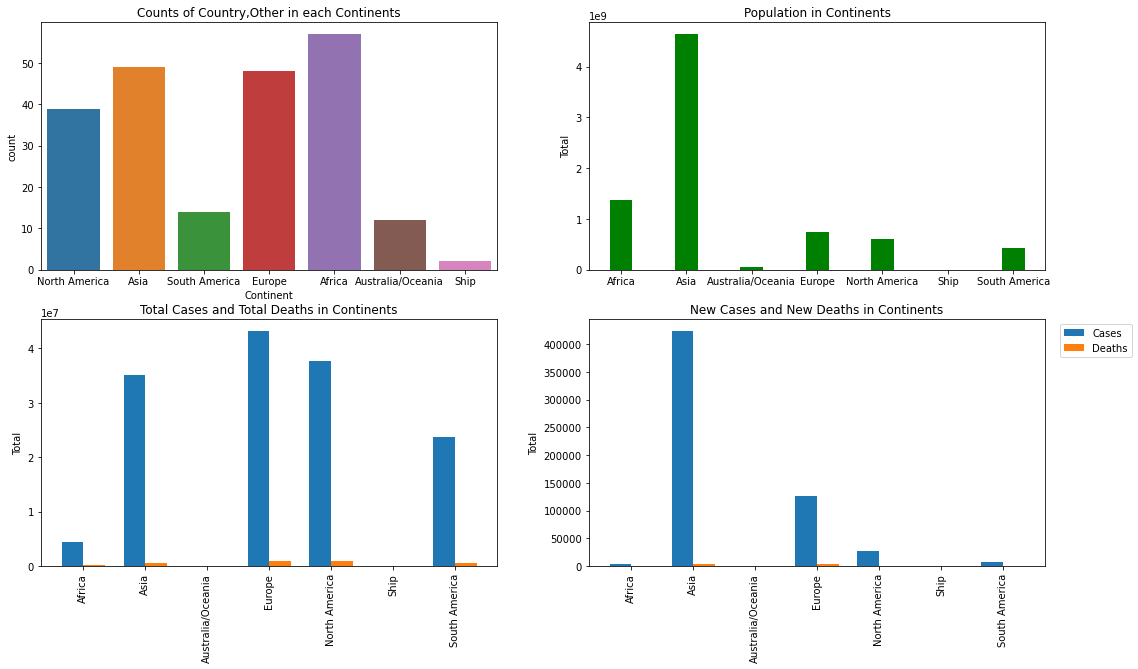


1. Mô phỏng:
   1. Vẽ biểu đồ cho thuộc tính Continents:

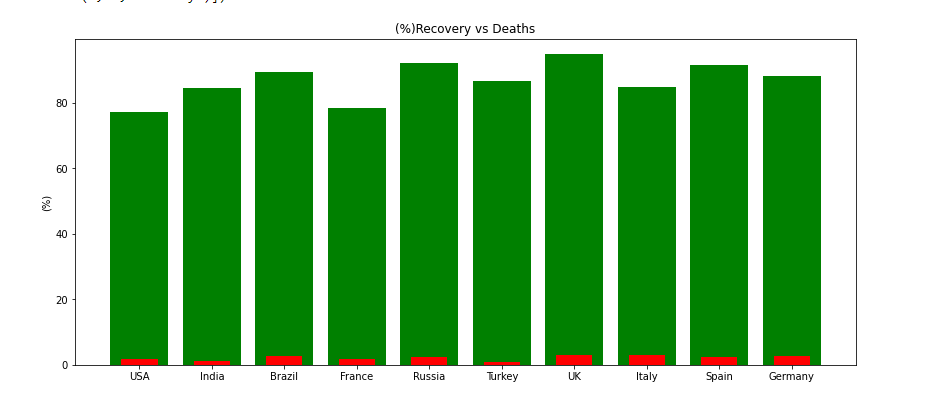
* Dùng biểu đồ countplot để đếm số lượng các quốc gia ở từng lục địa ở thuộc tính Continents.
* Hàm countplot của thư viện seaborn sẽ đếm số lượng các quan sát trong thuộc tính và vẽ biểu đồ.



* 1. Vẽ biểu đồ tương quan giữa Continent, NewCases, NewDeaths, Population.
* Kiểm tra sự tương quan giữa tổng số dân so với tổng số ca nhiễm và số ca tử vong ở từng khu vực.

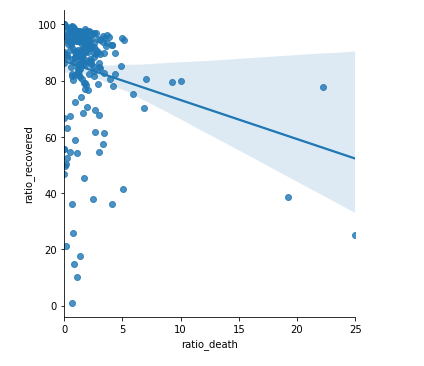


* Nhận xét:
  + Vào ngày 22/4/2021 số ca mắc ở Asia( châu Á) là cao nhất.
  + Asia( châu Á) là châu lục có dân số cao nhất thế giới, tuy nhiên tổng số ca nhiễm và tổng số ca tử vong lại thấp hơn Europe( châu Âu) và American ( châu Mĩ).
  1. Biểu đồ thể hiện % recovered và death của top 10 nước có TotalCases cao nhất



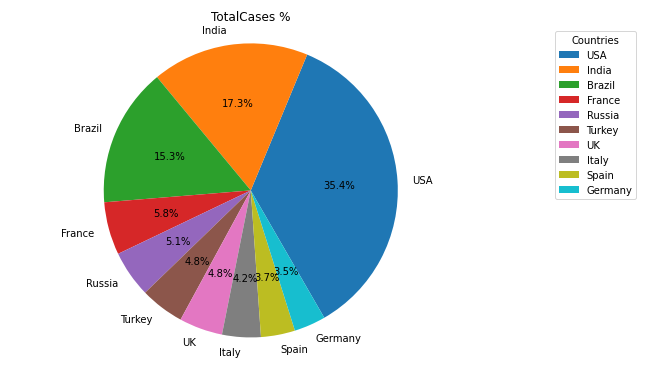
- UK có tỉ lệ Recovered cao nhất, và tỉ lệ Death cũng tương đối cao, điều này có vẻ mâu thuẫn

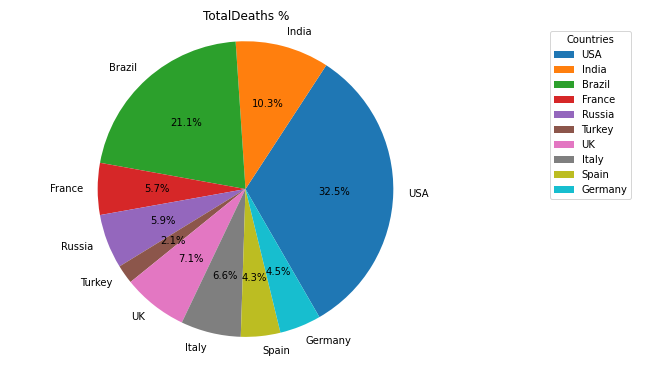
- France có tỉ lệ Recovered thấp nhất và tỉ lệ Death cũng ở tỉ lệ trung bình so với các nước khác

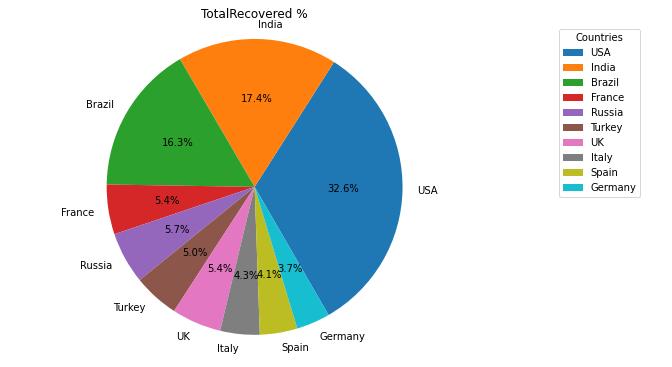


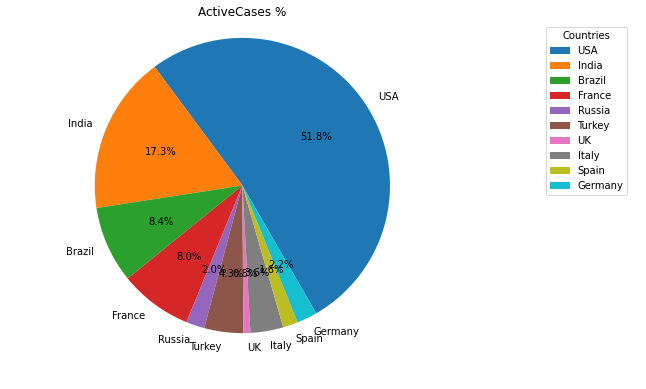
- Có vài nước có tỉ lệ Recovered và Death rất cao, có điều bất thường ở đây

* 1. Biểu đồ thể hiện tỉ lệ totalcases, totaldeath, totalrecovered, activecases của top 10 nước có totaldeath cao nhất



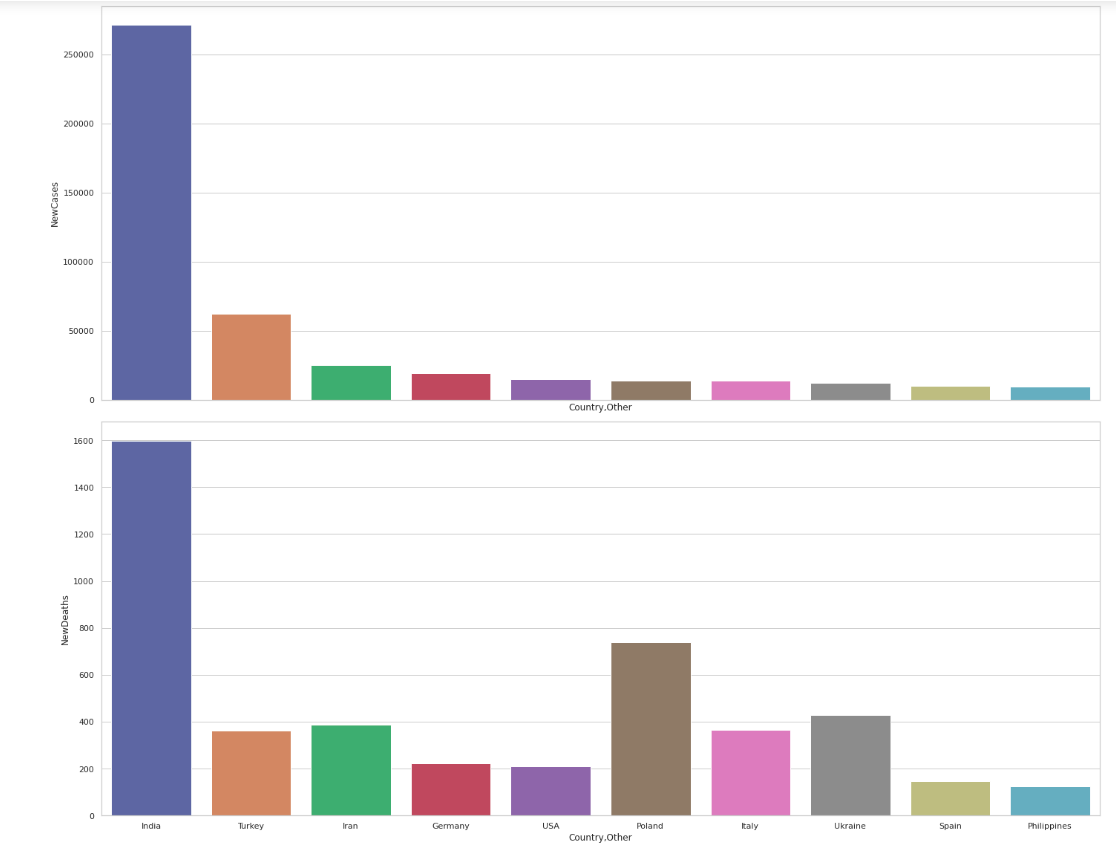


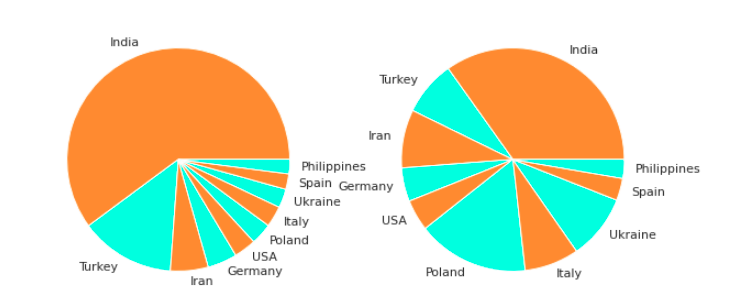




- Qua 4 biểu đồ tròn ở trên, ta thấy 3 nước USA, India, Brazil đều có tỉ lệ Recovered, Deaths, Activecases cao hơn nhiều so với các nước khác

- 3 thuộc tính Cases, Deaths và Recovered của 10 nước này có mối tương quan với nhau

6. Biểu đồ thể hiện số ca nhiễm mới và số ca chết mới ở top 10 nước có số ca nhiễm cao nhất

Dựa vào 2 biểu đồ ta có thể thấy không có mối quan hệ nhân quả (cause-effect) giữa số ca nhiễm mới và số ca chết dựa trên số liệu của top 10 nước có số ca nhiễm mới cao nhất.