**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

Sinh viên thực hiện: 18120408 - Trần Ngọc Lan Khanh

18120401 – Mai Khánh Huyền

18120379 – Võ Thị Hiếu

18120378 – Trần Văn Hiếu

GV phụ trách: Lê Ngọc Thành

  - TRỰC QUAN HÓA DỮ LIỆU

HỌC KỲ Ii – NĂM HỌC 2020-2021

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ tên** | **Email** | **Điện thoại** |
| 18120408 | Trần Ngọc Lan Khanh | 18120408@student.hcmus.edu.vn |  |
| 18120401 | Mai Khánh Huyền | 18120401@student.hcmus.edu.vn |  |
| 18120379 | Võ Thị Hiếu | [18120379@student.hcmus.edu.vn](mailto:18120379@student.hcmus.edu.vn) |  |
| 18120378 | Trần Văn Hiếu | [18120378@student.hcmus.edu.vn](mailto:18120378@student.hcmus.edu.vn) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Bảng phân công & đánh giá hoàn thành công việc** | | | |
| **Công việc thực hiện** | **Người thực hiện** | **Mức độ hoàn thành** | **Đánh giá của nhóm** |
| Công việc X | 1122333-Nguyễn Văn A | 80% | 7/10 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**YÊU CẦU ĐỒ ÁN- BÀI TẬP**

# **Yêu cầu của Đồ án/Bài tập**

Tìm hiểu mối quan hệ giữa các trường dữ liệu thực tế.

# **Kết quả**

1. Lấy dữ liệu:

Sử dụng thư viện BeautifulSoup và Requests để lấy dữ liệu từ trang……………….

1. Tiền xử lý dữ liệu:

Sử dụng thư viện Pandas, Numpy để đọc và sử lý dữ liệu

Chuyển dữ liệu từ Object thành numeric trừ cột Date và Country,Others

Để đảm bảo tính nhất quán và khách quan, không xử lý dữ liệu có giá trị NaN

1. Phân tích và trực quan dữ liệu:

Kết hợp thư viên Matplotlib.pyplot và Seaborn để trực quan hóa dữ liệu.

1. Hướng tiếp cận thứ nhất:

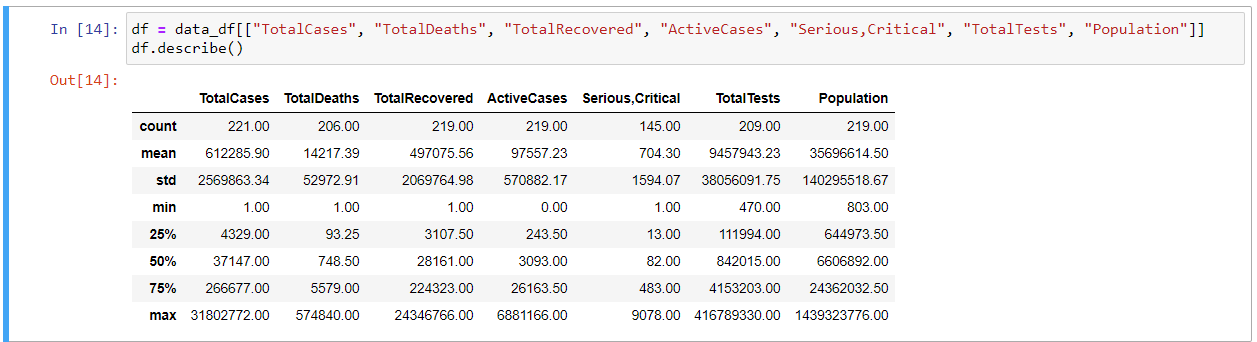
Tiếp cận dựa vào TotalCases

* 1. Phân tích một số trường dữ liệu

1. Các trường dữ liệu được chọn:

Chọn các thuộc tính TotalCases, TotalDeaths, TotalRecovered, ActiveCases, Serious,Critical, TotalTests, Population

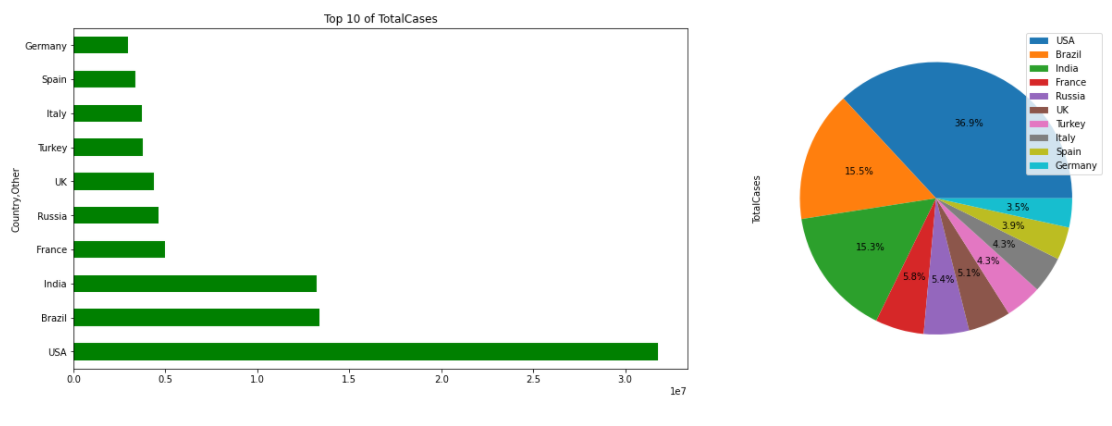
Tính chất các thuộc tính đã chọn:



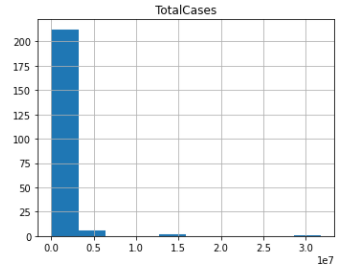
1. Phân tích các trường dữ liệu đã chọn

* TotalCases

Top 10 quốc gia có số ca nhiễm nhiều nhất:

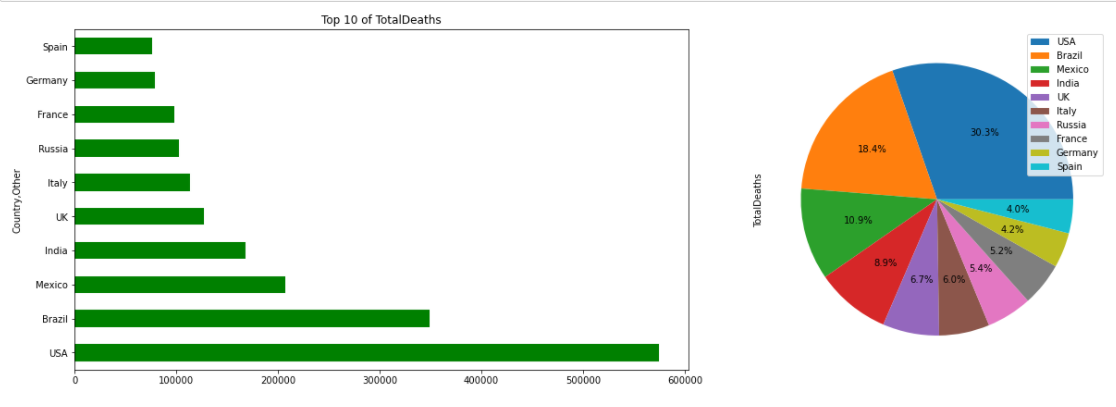


Biểu đồ thể hiện sự phân bổ về số ca nhiễm ở mỗi quốc gia:

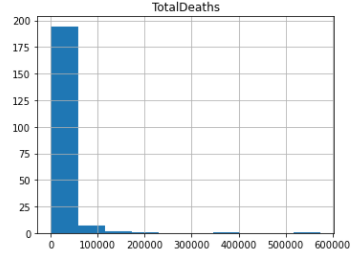


* Ta được các thông tin:
  + Hoa Kỳ là quốc gia có số ca nhiễm bệnh cao nhất
  + Phần lớn các quốc gia có số ca nhiễm ít hơn 5 triệu
  + 3 nước có số ca nhiễm lớn hơn 5 triệu là: Hoa Kỳ (hơn 30 triệu ca), Brazil (hơn 13 triệu ca), Ấn Độ (hơn 13 triệu ca)
  + Sự chênh lệch về số ca nhiễm của các quốc gia có sự khác biệt lớn theo từng nhóm ( Hoa Kỳ - Brazil, Ấn Độ - phần còn lại)
* TotalDeaths

Số ca tử vong ở 10 quốc gia có số ca nhiễm lớn nhất:

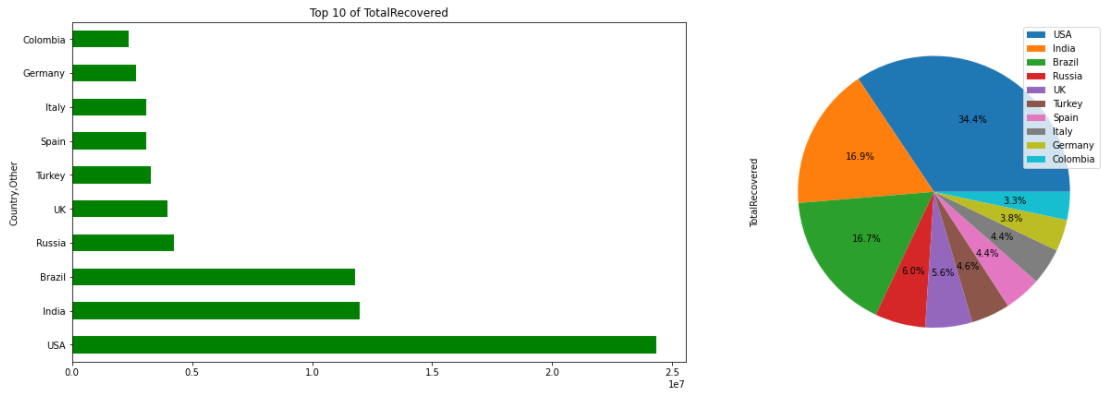


Số phân bố về số lượng quốc gia theo số ca tử vong:

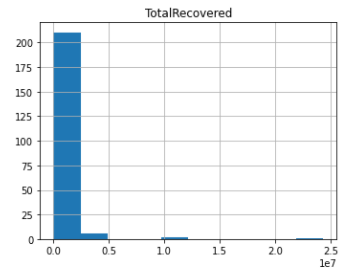


* Kết luận:
  + Hoa Kỳ dẫn đầu về số người tử vong (trong số 10 nước được xét)
  + Phần lớn các quốc gia có số ca tử vong dưới 50 000
  + Số ca tử vong của các quốc gia có sự chênh lệch lớn nhưng không phân chia thành các nhóm rõ ràng
* TotalRecovered

Tổng số ca hồi phục:

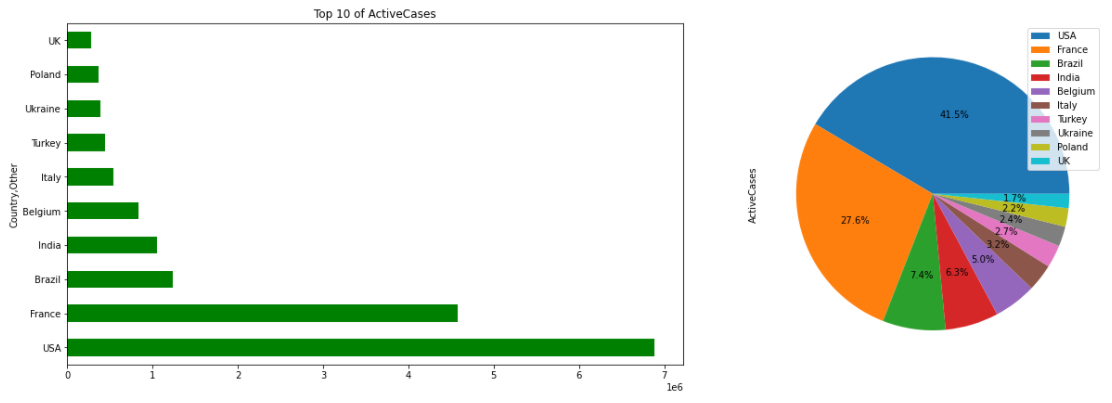


Sự phân bổ về số lượng quốc gia theo số ca hồi phục:

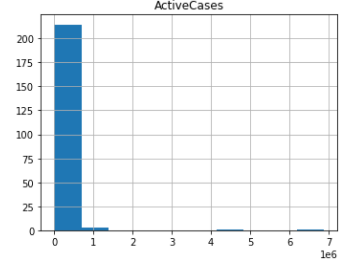


* Kết luận:
  + Hoa Kỳ có số ca hồi phục nhiều nhất
  + Phần lớn quốc gia có số ca hồi phục dưới 5 triệu
  + Số ca hồi phục có sự chênh lệch lớn và phân thành từng nhóm (Hoa Kỳ - Brazil, Ấn Độ - còn lại)
* ActiveCases

Số ca bệnh tại thời điểm phân tích

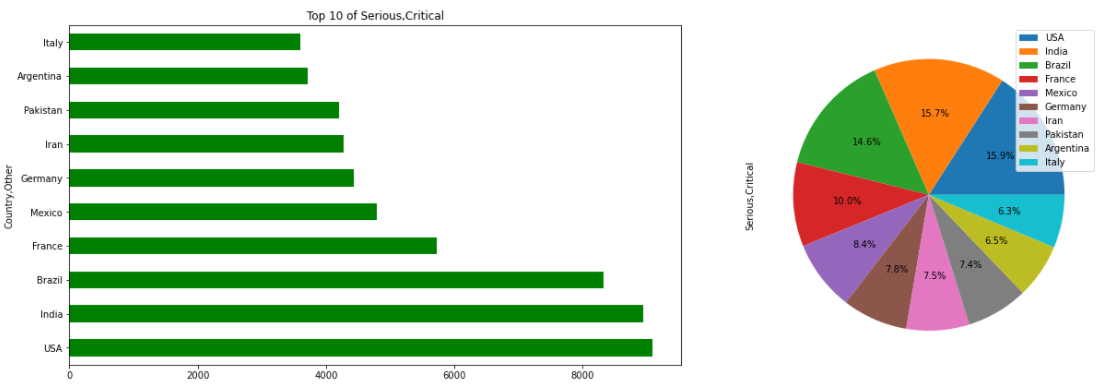


Sự phân bổ về số lượng quốc gia theo số ca dương tính:

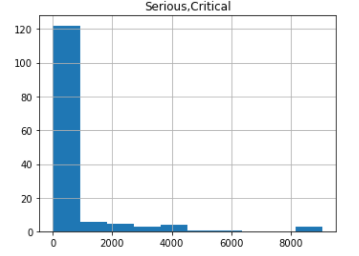


* Kết luận:
  + Hoa Kỳ có số ca dương tính nhiều nhất
  + Phần lớn quốc gia có số ca hồi phục dưới 1 triệu
  + Số ca dương tính có sự chênh lệch lớn giữa các quốc gia
* Serious,Critical

Số ca nghiêm trọng tại thời điểm phân tích

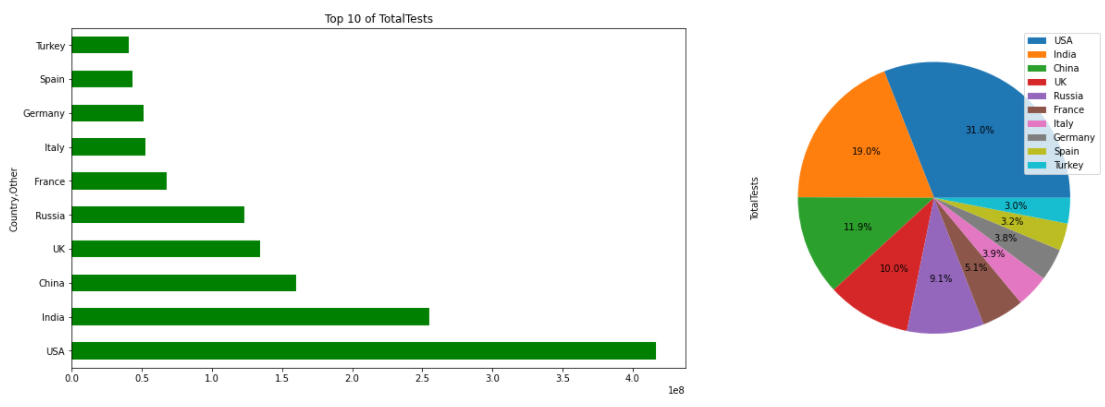


Sự phân bố về số lượng quốc gia theo số ca nghiêm trọng:

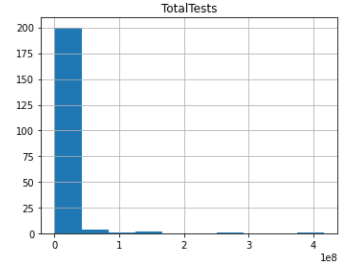


* Nhận xét:
  + Hoa kỳ có số ca nghiêm trọng nhiều nhất
  + Phần lớn quốc gia có số ca nghiêm trong ít hơn 1000
  + Số ca dương tính có sự chênh lệch không quá lớn giữa các quốc gia.
* TotalTests

Tổng số ca được xét nghiệm:

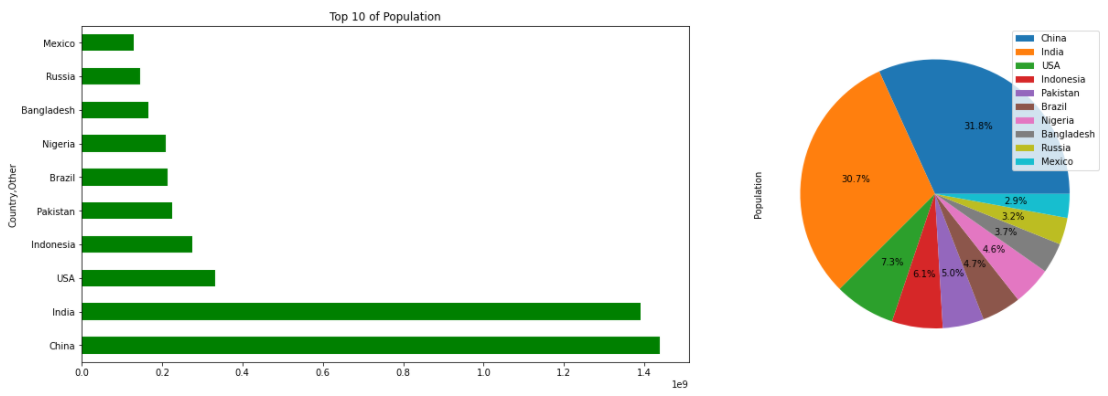


Sự phân bổ về số quốc gia theo số ca được xét nghiệm

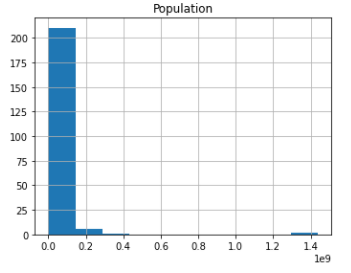


* Nhận xét:
  + Hoa kỳ có số ca xét nghiệm nhiều nhất
  + Phần lớn quốc gia có số ca xét nghiệm trong ít hơn 100 triệu
  + Số ca xét nghiệm của từng quốc gia có sự chênh lệch lớn
* Population

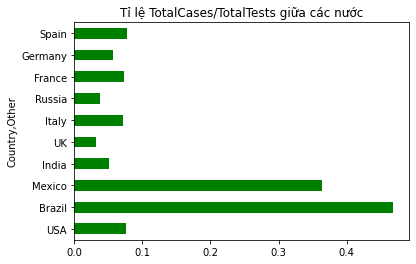
Dân số của các quốc gia:



Sự phân bổ về số quốc gia theo dân số:



* Nhận xét:
  + Trung Quốc và Ấn Độ có dân số lớn nhất
  + Phần lớn quốc gia có dân số ít hơn 200 triệu
  1. Phân tích quan hệ giữa các trường đã chọn:
* Tỉ lệ TotalCases/TotalTests giữa các nước:



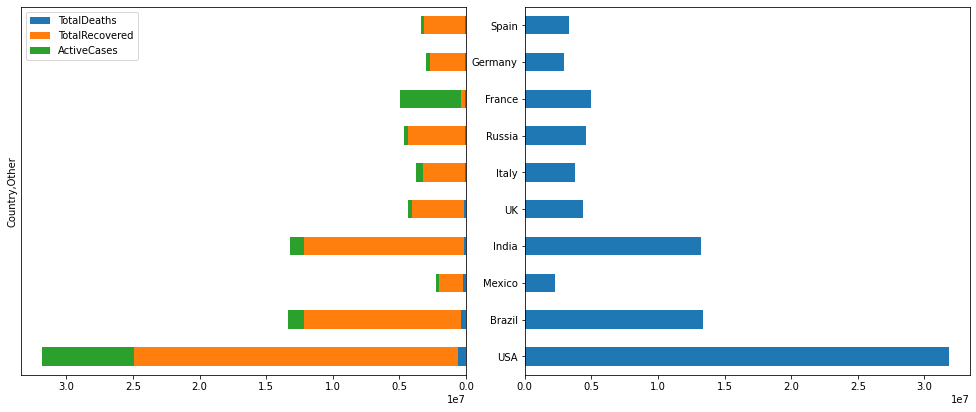
* Nhận xét:
* Tỉ lệ TotalCases/TotalTests của Mexico(>0.3) và Brazil(>0.4) là có cao vướt bậc so với các quốc gia khác
* Các quốc gia khác đều có tỉ lệ TotalCases/TotalTests gần như bằng nhau (0.05-0.1)
* Điều này cho thấy tỉ lệ dương tính trên số ca xét nghiệm COVID-19 ở Mexico và Brazil rất cao, điều này nghĩa là do chính phủ không thể nắm rõ và có đủ khả năng xử lý dịch bệnh của 2 nước Mỹ Latinh này.
* Tỉ lệ TotalRecovered/TotalCases giữa các nước:



* Nhận xét:
* Tỉ lệ TotalRecovered/TotalCases của Pháp là rất thấp với các nước còn lại (<0.1)
* Các quốc gia còn lại đều có TotalRecovered/TotalCases khá là cao (~0.8)
* Điều này cho thấy việc kiểm soát địch bệnh của Pháp rất kém, bệnh nhân khi mắc Covid-19 ít có cơ hồi phục hồi.

Tuy nhiên có dấu hiện đáng mừng khi tỉ lệ phục hồi bệnh của các nước đều cao.

* Số lượng TotalDeaths,TotalRecovered, ActiveCases giữa các nước:



* Nhận xét:
* Mỹ là quốc gia có tổng số ca mắc Covid-19 rất cao, cao vượt bậc so với các quốc gia khác.
* Mỹ, Brazil, Mexico là các quốc gia có số ca tử vong cao nhất, 3 nước này cũng lần lượt 3 nước có số ca mắc Covid cao nhất thế giới.
* Số lượng ca đang mắc Covid-19 tại thời điểm hiện tại của Mỹ và Pháp là cao nhất. Tuy nhiên, điều đáng lo ngại là Pháp là nước có tỉ lệ phục hồi thấp.

1. Hướng tiếp cận thứ 2