Báo cáo bài tập 2

Nhập Môn Khoa Học Dữ Liệu

GV: Lê Ngọc Thành

MSSV:19127326

Họ và tên: Vương Thành An

1. Tổng kết dữ liệu thu thập được
   1. Dữ liệu gốc

* 15 cột dữ liệu:
* Country,Other: Tên các quốc gia trên thế giới
* TotalCases: Tổng số ca mắc covid
* NewCases: Số ca mắc covid mới trong ngày
* TotalDeaths: Tổng số ca chết vì covid
* NewDeaths: Số ca chết vì covid trong ngày
* TotalRecovered: Tổng số ca bị covid đã được chữa trị hết bệnh
* NewRecovered: Số ca đã được chữa hết bệnh trong ngày
* ActiveCases: Tổng số ca còn đang bị mắc covid
* Serious,Critical: Tổng số ca đang mắc covid và đang trong tình trạng nguy kịch
* TotalCases/1M pop: Tổng số ca mắc covid trên 1 triệu dân số
* Deaths/1M pop: Tổng số ca chết vì covid trên 1 triệu dân số
* TotalTests: Tổng lần xét nghiệm covid cho một người
* Tests/1M pop: Tổng số lần xét nghiệm covid
* Population: Tổng dân số ở quốc gia đó
* Continent: Châu lục của quốc gia đó

1. Tiền xử lý dữ liệu

1.1 Tổng quát toàn bộ dữ liệu

* #Load dữ liệu các file dữ liệu của các ngày đã thu thập được lên corona-{day}-12.csv ví dụ corona-3-12.csv
* Xóa hết dấu “+” và dấu “,” trong số liệu và chuyển thành dạng numeric
* Đổi dữ liệu bị N/A và None thành 0,riêng ở cột Continent những giá trị đó sẽ đổi thành “The remaining areas”
* Tính toán dữ liệu cho giá trị còn thiếu của các cột và tính thêm giá trị cho các cột mới
* Tính giá trị ở cột Population,TotalCases/1M pop,Deaths/1M pop,TotalTests,Tests/1M pop cho các mẫu có giá trị ở cột Country, Other là Total: (Vì đây là số liệu của các châu lục)
* Tính thêm 3 cột mới(NewCases/1M pop,NewDeaths/1M pop,ActiveCases/1M pop) ở mỗi mẫu từ cột NewCases,NewDeaths,ActiveCases,Population
* Sau đó duyệt lại lần nữa 3 cột dữ liệu mới của các mẫu để chuyển các giá trị None thành 0
* Cuối cùng lưu vào 1 file csv mới gắn thêm đuôi -pre ví dụ corona-3-12-pre.csv