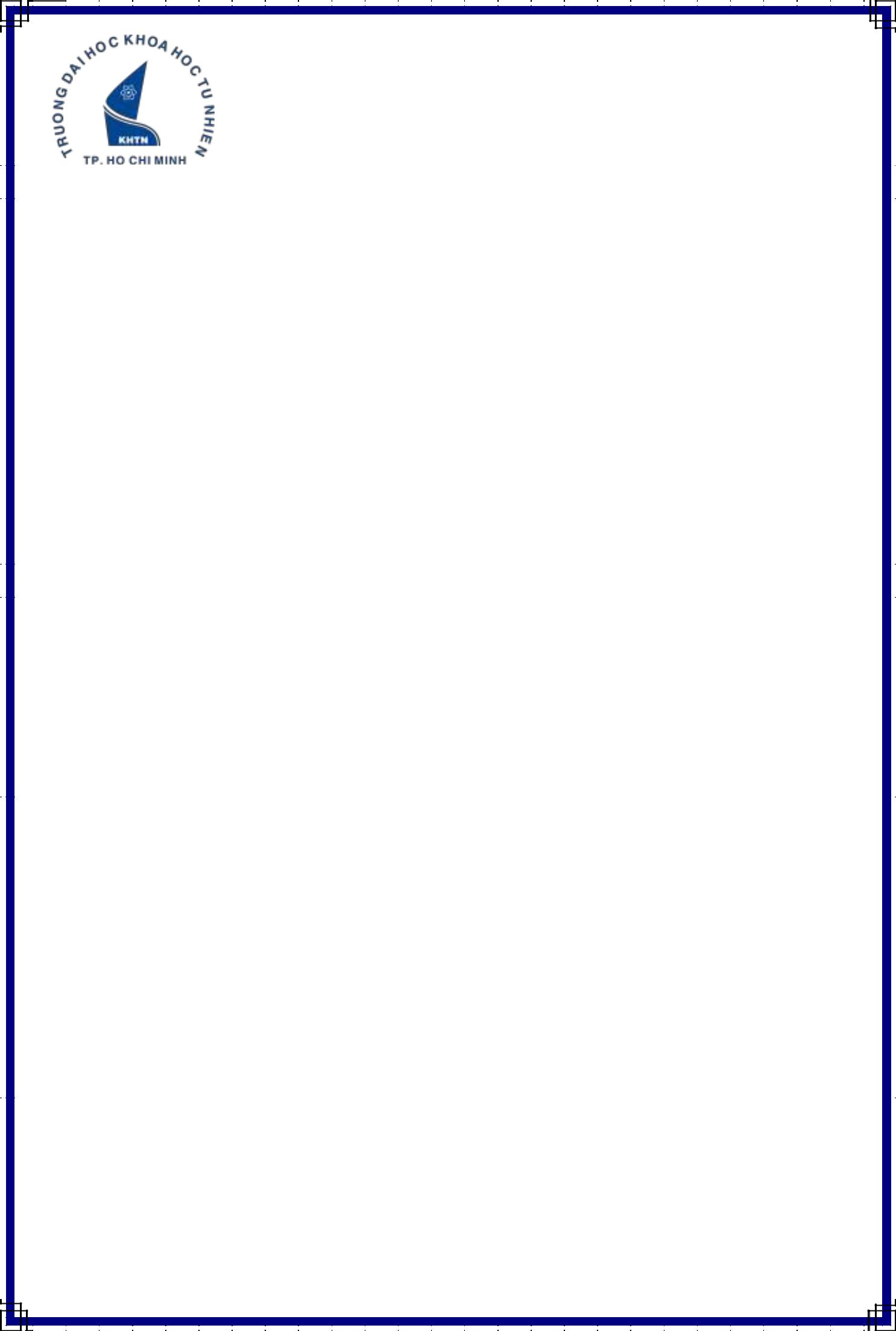
**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TPHCM  
KHOA CÔNG NGHỆ THÔNG TIN  
HỆ: CHÍNH QUI  
MÔN: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**ĐỒ ÁN 01**

**TRỰC QUAN DỮ LIỆU COVID-19**

**(Python)**

TP.HCM, ngày 25 tháng 10 năm 2019

MỤC LỤC

[**I.** **Thông tin nhóm** 3](#_Toc42610895)

[**II.** **Mức độ hoàn thành:** 3](#_Toc42610896)

[**III.** **Báo cáo đồ án:** 4](#_Toc42610897)

[**a.** **Cấu trúc chương trình:** 4](#_Toc42610898)

[**b.** **Tiền xử lý dữ liệu:** 4](#_Toc42610899)

[**c.** **Trực quan và nhận xét:** 5](#_Toc42610900)

[**i)** **Bar chart: top 15 nước có số ca nhiễm nhiềm nhất** 5](#_Toc42610901)

[**ii)** **Scatter: Số ca mắc/ca tử vong** 6](#_Toc42610902)

[**iii)** 8](#_Toc42610903)

[**iv)** **D** 8](#_Toc42610904)

[**v)** **D** 8](#_Toc42610905)

[**vi)** **D** 8](#_Toc42610906)

[**vii)** **D** 8](#_Toc42610907)

[**viii)** **D** 8](#_Toc42610908)

[**ix)** **D** 8](#_Toc42610909)

[**x)** **D** 8](#_Toc42610910)

[**xi)** **D** 8](#_Toc42610911)

[**xii)** **D** 8](#_Toc42610912)

[**xiii)** **D** 8](#_Toc42610913)

[**xiv)** **d** 8](#_Toc42610914)

[**d.** **Lưu ý về dữ liệu:** 8](#_Toc42610915)

[**IV.** **Nguồn/Tài liệu tham khảo:** 9](#_Toc42610916)

1. **Thông tin nhóm**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ tên** | **Vai trò** | **Nhiệm vụ** | **% hoàn thành** |
| 1 | 1712732 | Thái Bá Sơn | Nhóm trưởng |  |  |
| 2 | 1712739 | Lương Bội Sương | Thành viên |  |  |
| 3 | 1712773 | Võ Thế Sơn | Thành viên |  |  |

1. **Mức độ hoàn thành:**

|  |  |  |
| --- | --- | --- |
| **Phần** | **Tự đánh giá** | **Phần trăm hoàn thành** |
| Tiền xử lý | Thiếu phần định nghĩa phân cấp | 80% |
| Chạy biểu đồ | Hoàn thành | 100% |
| Đánh giá | Hoàn thành | 100% |
| Results | Hoàn thành | 100% |

1. **Báo cáo đồ án:** 
   1. **Cấu trúc chương trình:**

Chương trình gồm file menu.py, visualization.py, preprocess.py với chắc năng cụ thể:

* menu.py: Chứa menu để chọn các biểu đồ muốn trực quan, thực thi trên màn hình command.
* visualization.py: Thư viện chứa các hàm để trực quan hóa dữ liệu dưới dạng biểu đồ.
* preprocess.py: Thư viện chứa hàm tiền xử lý dữ liệu.

File input: Dữ liệu về covid-19 được thống kê vào ngày 03/06 (03thg6.csv) và dữ liệu đã được tiền xử lý nằm ở file 03thg6\_preprocessed.csv.

A picture containing holding, black, man, white

Description automatically generated

File hướng dẫn chạy chương trình: Readme.txt

* 1. **Tiền xử lý dữ liệu:**
* Dữ liệu input được lấy trực tiếp từ trang <https://www.worldometers.info/coronavirus/> bằng cách sao chép và dán vào trang Excel trống. Lưu lại dưới dạng .csv.

**A picture containing screen, television, monitor, holding

Description automatically generated**

* Lấy dữ liệu từ file .csv để đưa vào dataframe (trả về biến df) dựa trên thư viện pandas.

**A screen shot of a computer

Description automatically generated**

* Các bước tiến xử lý dữ liệu như:
  + Xóa dấu cách (space) trong dữ liệu.
  + Xóa cột chứa dữ liệu không dùng đến (cột # chứa số thứ tự).
  + Chuyển đổi các ô rỗng thành giá trị 0.
  + Đặt tên quốc gia làm chỉ mục (set\_index).
  + Đổi tên các cột thuộc tính thành biến dễ thao tác hơn. Vd: Country,Other -> Country.
  1. **Trực quan và nhận xét:**
     1. **Bar chart: top 15 nước có số ca nhiễm nhiềm nhất**

**A screen shot of a smart phone

Description automatically generated**

Sắp xếp top 15 quốc gia có số ca mắc nhiều nhất theo thuộc tính TotalCases, trực quan bằng biểu đồ Bar chart trong thư viện plot.

Ta được biểu đồ trực quan sau:

**A screenshot of a cell phone

Description automatically generated**

**Nhận xét, ý nghĩa:** Biểu đồ cho thấy số ca mắc ở Mỹ (USA) hiện 1.881.205 ca, bỏ xa các quốc gia xếp sau. Cho thấy tình hình dịch bệnh ở Mỹ đặc biệt nghiêm trọng, hơn cả quốc gia bùng phát dịch là Trung Quốc (nằm ngoài top 15).

* + 1. **Scatter: Số ca mắc/ca tử vong**

**A screenshot of a computer screen

Description automatically generated**

Trực quan tình hình các quốc gia bằng điểm trên biểu đồ với trục x là số ca tử vong và trục y là số ca mắc (chỉ hiển thị tên của điểm hiển thị đối với 15 quốc qua đầu bảng để dễ quan sát). Đồng thời hiện đường phân tách tuyến tính để đánh giá. Đường phân tách tuyến tính có dạng y = ax + b với a và b được rút ra từ hàm polyfit của thư viện numpy.

**A screenshot of a cell phone

Description automatically generated**

**Nhận xét, ý nghĩa:** Dựa vào biểu đồ, ta thấy được độ hiệu quả trong việc phòng chống dịch. Đánh giá dựa vào đường phân tách tuyến tính của biểu đồ, nếu điểm nằm phía trên thì hệ thống y tế quốc gia đó chưa tốt và ngược lại nếu xét điểm nằm dưới đường phân tách.

Tuy Mỹ(USA) có số ca mắc nhiều nhất nhưng được đánh giá tốt về hệ thống y tế (nằm phía trên đường phân tách).

Hệ thống y tế các quốc gia Spain, Italy, France bị đánh giá là chưa tốt vì số ca mắc nhiều nhưng số ca tử vòng cũng chiếm tỉ lệ cao. Với số ca mắc khá tương đồng thì Đức (Germany) đã làm tốt hơn các nước trên số ca tử vong thấp (điểm biểu diễn nằm trên đường phân tách).

Tương tự ta có thể đánh giá tốt các quốc gia Russia, Brazil, Chile, India…

A screenshot of a cell phone

Description automatically generated

Đặc biệt, nếu ta hiển thị tên tất cả các quốc gia và zoom trên biểu đồ thì có thể thấy nhiều quốc gia chống dịch rất tốt, tuy có số ca mắc nhiều nhưng vẫn chưa xuất hiện ca tử vong nào như Uganda, Viet Nam, Campuchia…

* + 1. **Scatter: Số ca mắc, tử vong / 1 triệu dân:**

**A screenshot of a cell phone

Description automatically generated**

Về phần code, dựa vào hàm của mục iii) phía trên, chỉ thay đổi cột thuộc tính mà ta cần trực quan là TotalCases/1M và TotalDeaths/1M. Ta được biểu đồ trực quan sau:

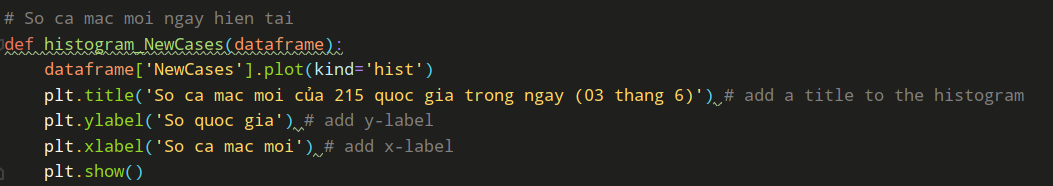
**A screenshot of a cell phone

Description automatically generated**

**Nhận xét, ý nghĩa:** Qua biểu đồ, ta thấy San Mario là đáng báo động nhất khi nằm khá xa phía trên đường phân tách, nghĩa là số người mắc và chết trên/ 1 triệu dân đều cao. Các quốc gia khác cũng nằm trong vùng tình hình xấu như Belgium, Andorra, Italy, Spain…

Ngược lại Qatar, Vatican, Bahrain… kiểm soát dịch tốt nếu tính số ca trên 1 triệu dân.

* + 1. **Histogram: Ca nhiễm mới trong ngày**



Trực quan bằng biểu đồ histogram của số ca mắc mới trong ngày đang xét (03thg6). Sử dụng histogram trong thư viện plot, trực quan thuộc tính NewCases trong dataframe.

**A screenshot of a cell phone

Description automatically generated**

**Nhận xét, ý nghĩa:** Theo biểu đồ trực quan, ta thấy tình hình dịch đang được đa số các nước kiểm soát tốt, hơn 200 quốc gia có ca mắc ít hơn 2.500 ca/ngày. Tuy nhiên vẫn có một vài quốc gia có số lượng mắc mới khá cao, trong khoảng 10.000 ca và trong khoảng 25.000 ca.

* + 1. **D**
    2. **D**
    3. **D**
    4. **D**
    5. **D**
    6. **D**
    7. **D**
    8. **D**
    9. **D**
    10. **d**
  1. **Lưu ý về dữ liệu:**
* Dữ liệu không hoàn toàn được liệt kê chính xác so với thực tế, cụ thể ở thống kê về số ca đang điều trị (Active Cases) và số ca hồi phục (Total Recovered) của các nước Spain và UK không hiển thị giá trị, gây khó khăn cho việc xử lý với dữ liệu. Trường hợp này không thể dùng phép tính suy ra để bổ sung giá trị nên không đưa vào quá trình tiền xử lý.
* Các nước trên toàn thế giới đều có ca nhiễm.
* Một vài nước không xét nghiệm ca nào (Total Test) nhưng vẫn có thống kê về ca mắc.

1. **Nguồn/Tài liệu tham khảo:**

<https://www.sciencedirect.com/topics/computer-science/concept-hierarchy>

<https://ongxuanhong.wordpress.com/2015/08/20/tien-xu-ly-du-lieu-horse-colic-dataset/>