Đại học Quốc gia TP. HCM
Trường Đại học Khoa học Tự nhiên

BÁO CÁO LAB 01
MỐI QUAN HỆ TRONG DỮ LIỆU

Trực quan hoá dữ liệu (CSC10108)

Nhóm 17

TP Hồ Chí Minh, ngày 29/04/2021

Contents

1 Thông tin nhóm

1. **Đường link GitHub**: `https://github.com/baolongnguyenmac/CinemaManagementSystem`
2. **Danh sách thành viên**

STT	MSSV	Họ tên	Email	SĐT
1	18120078	Ngô Phù Hữu Đại Sơn	18120078@student.hcmus.edu.vn	0919070940
2	18120201	Nguyễn Bảo Long	18120201@student.hcmus.edu.vn	0981850699
3	18120227	Phạm Văn Minh Phương	18120227@student.hcmus.edu.vn	0981850699
4	18120253	Mai Ngọc Tú	18120253@student.hcmus.edu.vn	0981850699
5	1712424	Hàn Văn Gia Hiên	1712424@student.hcmus.edu.vn	0911572108

Table 1: Bảng danh sách thành viên nhóm

2 Phân tích hoàn thiện yêu cầu

2.1 Tổng quan mức độ hoàn thành mỗi yêu cầu

STT	Yêu cầu	Công việc	Hoàn thành (%)
1	Thu thập dữ liệu	- Cài đặt chương trình thu thập dữ liệu - Tiền xử lý dữ liệu	100/100
2	Trực quan mối quan hệ	- Chọn trường dữ liệu cần trực quan - Trực quan mối quan hệ, giải thích ý nghĩa, nhận xét biểu đồ	90/100

Table 2: Bảng phân tích đóng góp cá nhân

2.2 Mức độ hoàn thành của thành viên nhóm

STT	Họ tên	Công việc tham gia	Hoàn thành (%)
1	Ngô Phù Hữu Đại Sơn	- Thu thập dữ liệu - Hồi quy tuyến tính cho cá quan hệ - Biểu diễn quan hệ 2 biến & 4 biến	20%
2	Nguyễn Bảo Long	- Biểu diễn quan hệ 3 biến & 4 biến - Giải thích lý do sử dụng biểu đồ đường - Nhận xét dữ liệu 3 biến & 4 biến	20%
5	Phạm Văn Minh Phương	- Giải thích biểu đồ stacked bar chart - Nhận xét dữ liệu stacked bar chart	20%
5	Mai Ngọc Tú	- Giải thích biểu đồ scatter - Nhận xét quan hệ dữ liệu 2 biến	20%
5	Hàn Văn Gia Hiên	-	20%

Table 3: Bảng phân tích đóng góp cá nhân

3 Thu thập dữ liệu

3.1 Thu thập số liệu thống kê từng ngày

- Nguồn dữ liệu: `https://www.worldometers.info/coronavirus/`

3.2 Tiền xử lý

- Vấn đề dữ liệu gặp phải:
- Hướng tiền xử lý:

3.3 Khám phá dữ liệu

4 Trực quan hoá mối quan hệ giữa các trường dữ liệu

4.1 Chọn trường dữ liệu

- Với mỗi biến X, Y của tập dữ liệu, ta tính hệ số tương quan

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}} = \frac{Cov(X,Y)}{\sigma(X) \times \sigma(Y)}$$

  Với $Cov(X,Y) = E[XY] - E[X]E[Y]$

- Từ đó, tính được ma trận hệ số tương quan giữa các cặp biến X, Y trong tập dữ liệu. Biểu diễn ma trận này bằng biểu đồ heatmap để xác định các trường dữ liệu trực quan.

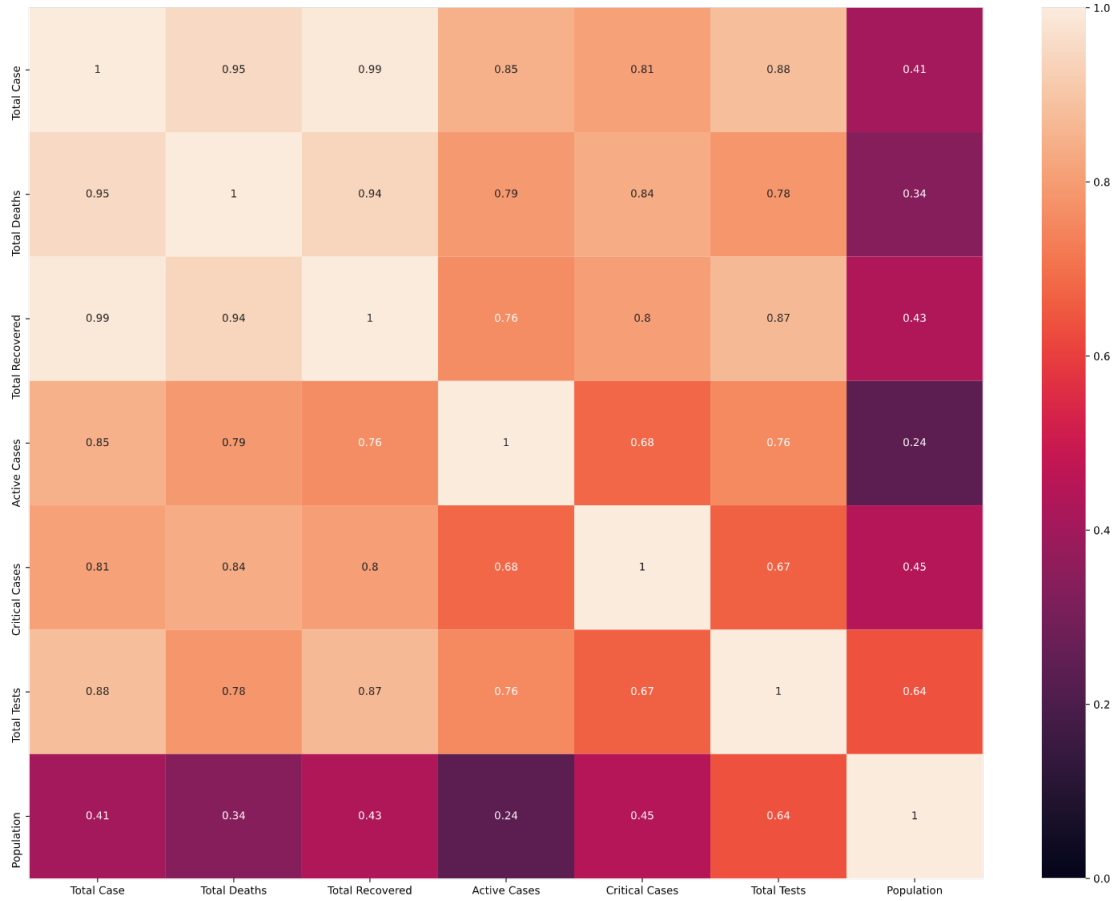

Figure 1: Biểu đồ thể hiện hệ số tương quan giữa các trường dữ liệu

- Từ biểu đồ trên, chọn ra các cặp thuộc tính có hệ số tương quan lớn hơn 0.7, ta được bảng các cặp thuộc tính sau

STT	Var1	Var2	Hệ số tương quan
1	Total Deaths	Total Cases	0.95
2	Total Recovered	Total Cases	0.99
3	Total Recovered	Total Deaths	0.94
4	Active Cases	Total Cases	0.85
5	Active Cases	Total Deaths	0.79
6	Active Cases	Total Recovered	0.76
7	Critical Cases	Total Cases	0.81
8	Critical Cases	Total Deaths	0.84
9	Critical Cases	Total Recovered	0.8
10	Total Test	Total Cases	0.88
11	Total Test	Total Deaths	0.78
12	Total Test	Total Recovered	0.87
13	Total Test	Active Cases	0.76

Table 4: Bảng các thuộc tính có hệ số tương quan trên 0.7

- Từ bảng trên, ta thấy có 6 thuộc tính có mức độ tương quan với nhau cao là **Total Case, Total Deaths, Total Recovered, Active Cases, Critical Cases, Total Tests**. Do đó, ta tập trung vào tìm hiểu mối quan hệ giữa các thuộc tính này.

4.2 Biểu diễn quan hệ giữa 2 trường dữ liệu

- Với mỗi 2 trường dữ liệu X, Y trong danh sách các trường dữ liệu xuất hiện trong Table 4, trực quan các điểm dữ liệu thuộc 2 trường đó bằng biểu đồ scatter.
- Tiếp đó, tìm một quy luật xấp xỉ phù hợp nhất với dữ liệu quan sát bằng một đường thẳng $\bar{y} = w_0 + xw_1$ với w_0, w_1 được tính theo công thức sau

$$w_1 = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - \overline{X}^2}$$

$$w_0 = \overline{Y} - w_1\overline{X}$$

- Lý do sử dụng biểu đồ scatter kết hợp biểu diễn quy luật bằng một đường tuyến tính: Sử dụng biểu đồ scatter để trực quan giúp người đọc dễ dàng nhận ra một sự xấp xỉ tuyến tính (nếu có) tồn tại giữa 2 biến dữ liệu. Cộng thêm việc biểu diễn quy luật bằng một đường tuyến tính, người đọc có thể thấy rõ được "chất lượng" của đường tuyến tính này trong việc giải thích mối quan hệ giữa 2 trường dữ liệu.
- Kết quả thu được $C_6^2 = 15$ biểu đồ như sau

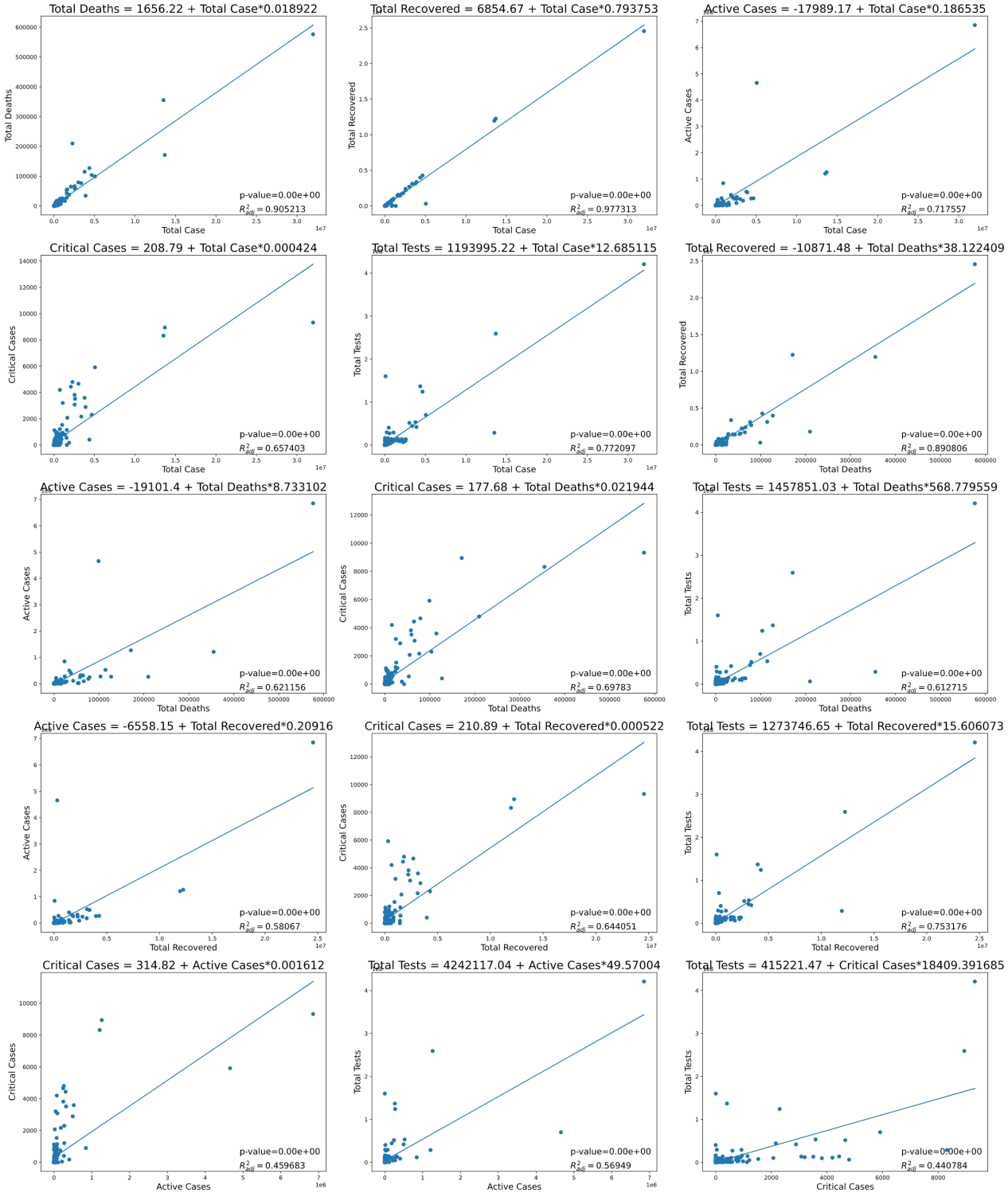

Figure 2: Biểu đồ thể hiện mỗi quan hệ giữa 2 biến dữ liệu

- Nhận xét dữ liệu
  - Với 5 cặp biến (Độc lập - Phụ thuộc) là **(Total Case, Total Deaths), (Total Case, Total Recovered), (Total Case, Total Tests), (Total Deaths, Total Recovered), (Total Recovered, Total Tests)**, ta có các kết luận sau
    - Chỉ số p-value của các cặp biến này rất nhỏ (trong khoảng $[10^{-180}, 10^{-68}]$). Do đó, các biến độc lập nêu trên có ý nghĩa về mặt thống kê.

* Mô hình phù hợp tốt với dữ liệu quan sát về mặt thống kê (p-value trong khoảng $[10^{-180}, 10^{-68}]$).
* Hệ số xác định hiệu chỉnh R^2_{adj} của mỗi bộ dữ liệu đều nằm trong khoảng $[75\%, 98\%]$. Do đó, các biến độc lập giải thích được từ 75% đến 98% sự thay đổi của các biến phụ thuộc.
* Cụ thể phương trình hồi của các cặp biến được ghi trên biểu đồ.

– Với các cặp biến còn lại, chỉ số $R^2_{adj} < 75\%$. Do đó, việc giải thích sự thay đổi của biến phụ thuộc dựa trên biến độc lập là không đủ tin cậy.

4.3 Biểu diễn quan hệ giữa 3 trường dữ liệu

4.4 Biểu diễn quan hệ giữa 4 trường dữ liệu