**STATISTIC ON COVID DATASET**

**NGUYEN NHAT NGUYEN**

**Giả thuyết thống kê cho dự án:** Số người mắc COVID-19 có mối quan hệ gì với số ca tử vong do COVID-19 hay không? Liệu số ca nhiễm càng nhiều có dẫn đến số người chết càng nhiều hay không?

**1. Thu thập dữ liệu**

Đầu tiên, dựa vào đề tài đã chọn ta có thể xác định được quần thể cần nghiên cứu là số người bị mắc COVID và số ca tử vong vì COVID trên thế giới.

Vì các chỉ số cần thống kê có số lượng rất lớn và hầu như không có cách nào để nghiên cứu toàn bộ, nên ta chọn dữ liệu về số người mắc và chết vì COVID-19 được tổng hợp bởi WHO để làm dữ liệu đại diện cho quần thể. Dữ liệu trên thống kê về số ca mắc và số ca chết lũy kế của hầu hết các quốc gia trên thế giới, đã cập nhập đến ngày 22/08/2021, nên có thể coi đây là dữ liệu được thu thập một cách khách quan và toàn diện, có thể đại diện cho quần thể.\

Source: <https://ourworldindata.org/covid-deaths>

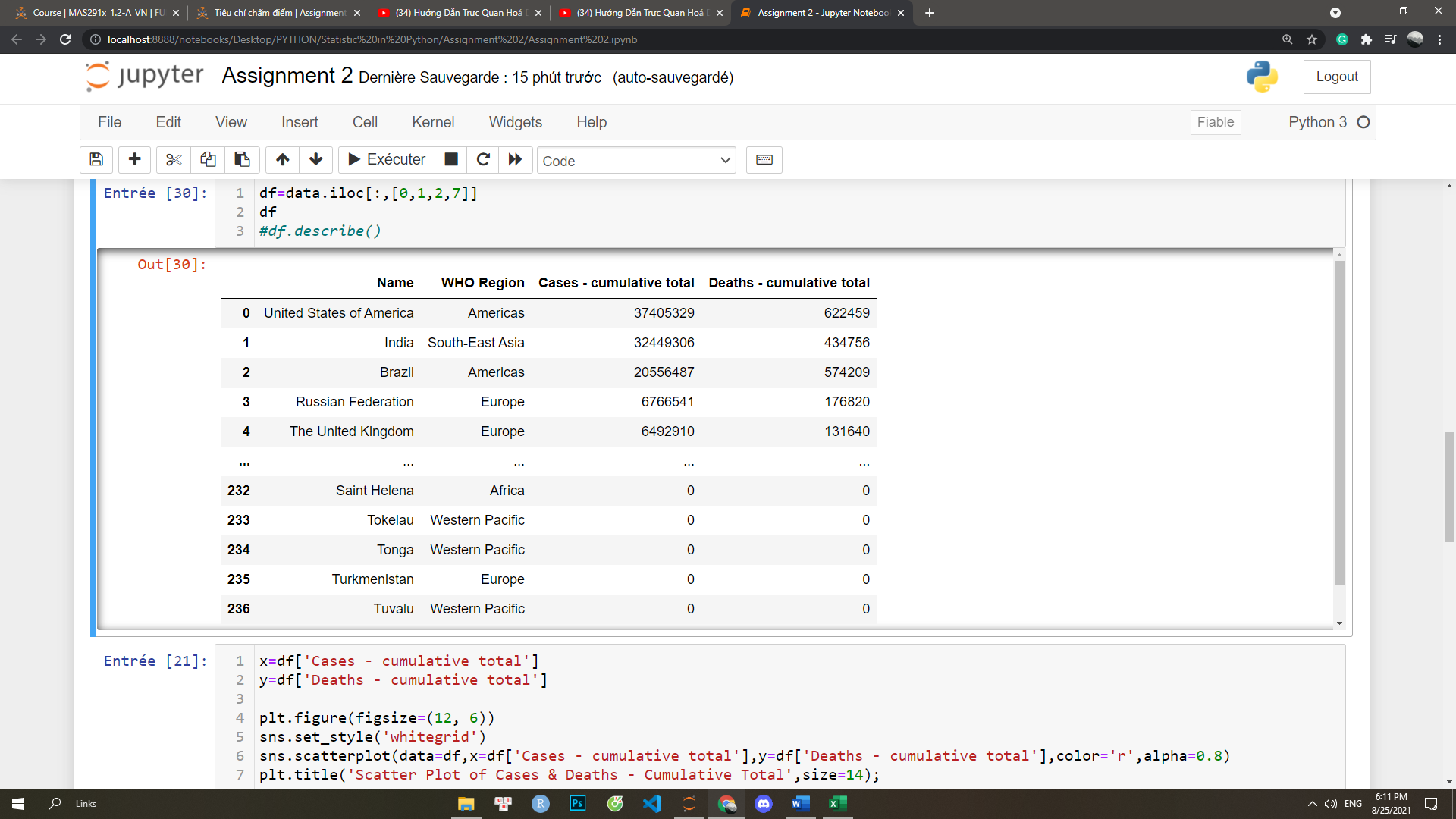
<https://covid19.who.int/table>

**Phân loại được dữ liệu:**

Dữ liệu cần thu thập sẽ cần có 2 biến liên tục chính (Continuous Variable) là tổng số ca mắc và số người chết thống kê theo từng quốc gia

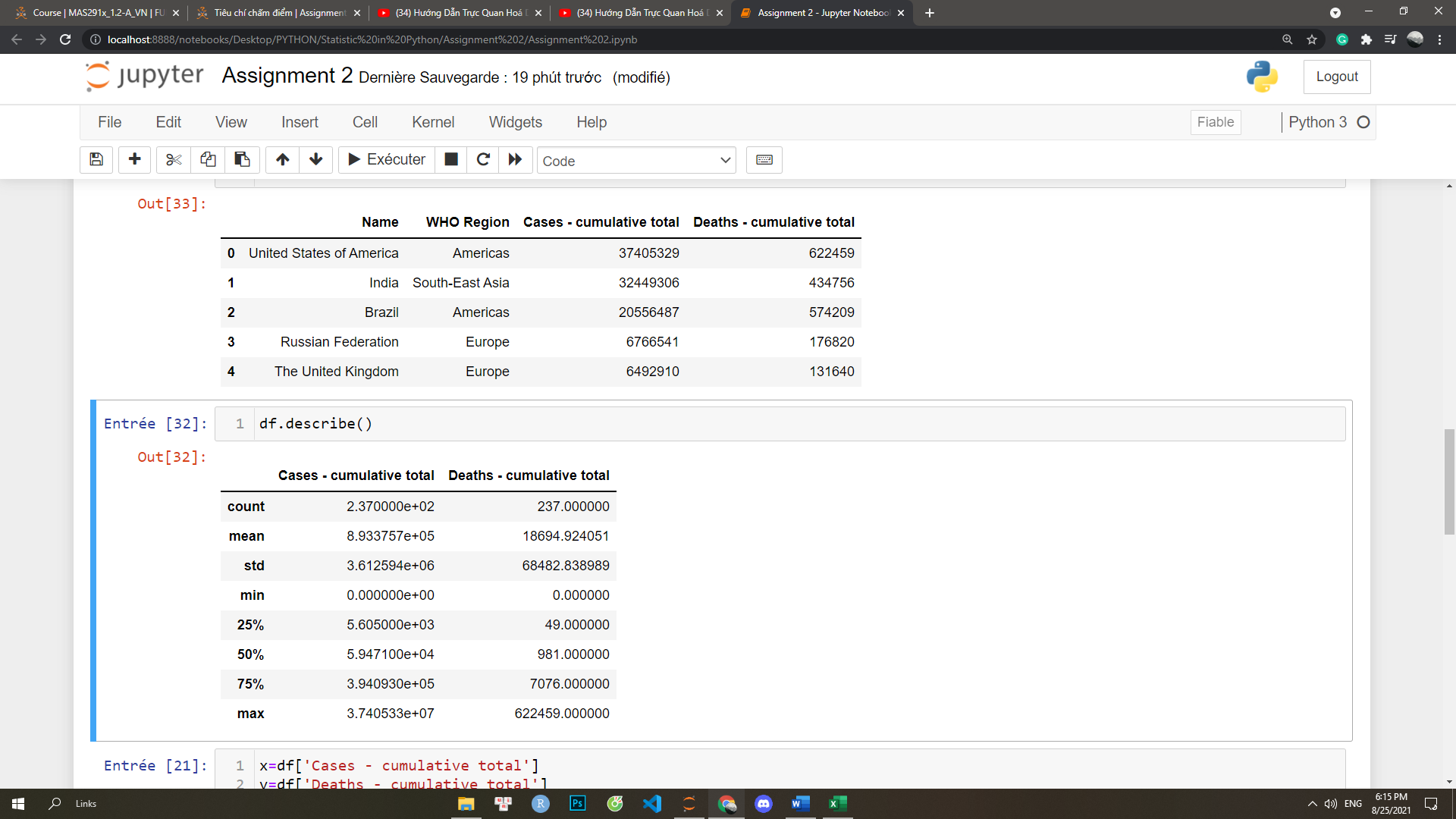
**Cách thu thập dữ liệu:** Nghiên cứu quan sát dựa trên số liệu thực tế

**2. Tổng hợp dữ liệu**

Bảng phân bố dữ liệu được đính kèm trong file. Sau khi loại bỏ những dữ liệu không liên quan ta có bảng tóm tắt như hình bên dưới

Các chỉ số cơ bản của dữ liệu cần thống kê:

( như Mean, Standard Deviation, Tứ phần vị, Max Min,..)



**3. Mô tả dữ liệu để loại bỏ các Outliers**

**3.1 Tìm các tham số trong dữ liệu**

Ta thử tìm hệ số tương quan và các tham số liên quan đến tương quan tuyến tính như r, slope- hệ số góc, và intercept-a hằng số của đường hồi quy tuyến tính

Ta có kết quả như sau:

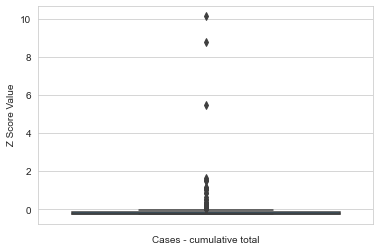
LinregressResult(slope=0.01765140764272694, intercept=2925.585831249611, rvalue=0.9311436786461393, pvalue=6.1412851231963546e-105, stderr=0.0004509287401127967)

Từ hệ số tương quan rvalue 0.93, ta thấy x và y có tương quan tuyến tính dương cực kì chặt chẽ. ( Mang tính chất tham khảo khi chưa thực hiện loại bỏ các outliers)

**3.2 Tìm các outliers trong dữ liệu:**

Có rất nhiều cách để tìm ra giá trị outliers tồn tại trong dữ liệu: dựa trên Z- Score , khoảng IQR, Boxplot, Scatter Plot, .v..v.

**Boxplot:**



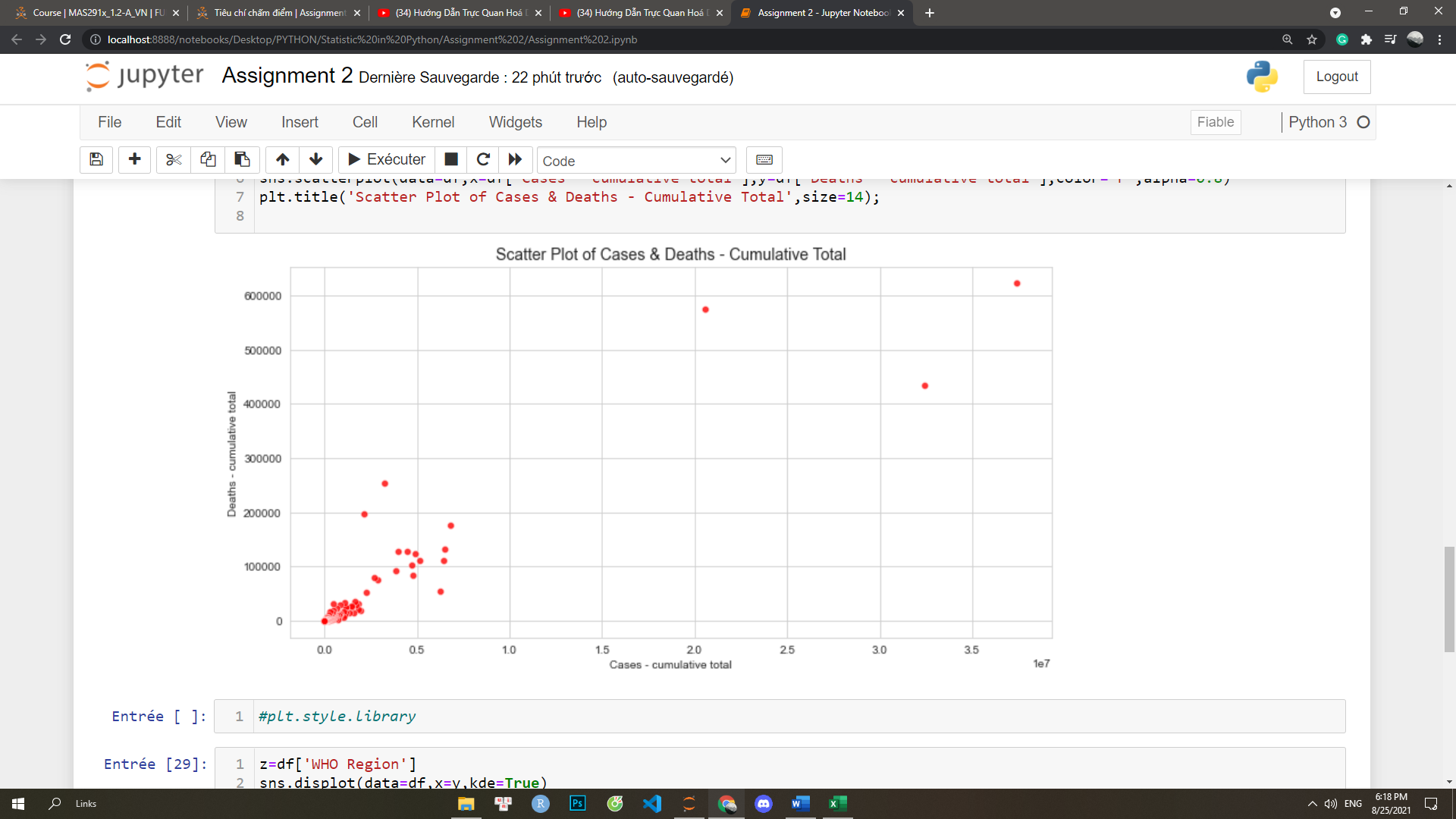
Bên trái là Boxplot vẽ trên giá trị số ca mắc Covid ( biến liên tục độc lập) dựa trên số ca nhiễm đã được chuẩn hóa dựa trên giá trị Z- Score.

Ta biết giá trị của Z-Score được xem là Outliers khi nằm ngoài khoảng [-3,3]( hay khoảng tin cậy 99.7%) .

Vì vậy nhìn vào Boxplot ta dễ dàng nhận ra có 3 Outliers lớn hơn giá trị 3 rất nhiều.Bên cạnh đó có quá nhiều giá trị 0- không có ý nghĩa thống kê và làm thay đổi đáng kể xu hướng của biểu đồ. Vì các giá trị này làm ảnh hưởng nặng nề tới hệ số tương quan và đường hồi quy tuyến tính nên ta cần loại bỏ các giá trị này trước khi đi tìm tương quan giữa x và y.

**Ta có thể tham khảo Scatter Plot để xem vị trí của các Outliers**

Theo quan sát, thực sự các Outliers rất lớn nhưng vẫn có xu hướng tuyến tính với phần dữ liệu còn lại. Tuy nhiên để tìm tương quan chính xác nhất cho quần thề, ta cần phải tiến hành loại bỏ các outliers để các chỉ số thể hiện khách quan và chính xác nhất.



**3.3 Xử lí các outliers**

Đầu tiên ta loại bỏ các giá trị không ca lây nhiễm ( x=0) ở trong tập dữ liệu. Sau đó ta chuẩn hóa các biến độc lập x sang giá trị Z-Score và loại bỏ các giá trị ngoài vùng tin cậy 99.7%. Có nghĩa là loại bỏ tất cả các giá trị có |z|>3.

Sau khi loại bỏ các giá trị Outliers khỏi tập dữ liệu, ta tiến hành kiểm định giả thuyết cho mô hình hồi quy tuyến tính.

**4. Kiểm định giả thuyết hồi quy tuyến tính**

**4.1 Xây dựng cặp giả thuyết**

Ta có, cặp giả thuyết Ho, Ha cho hệ số góc Beta:

Ho: Beta=0 ( Không có tương quan tuyến tính với y (Số ca nhiễm và Số người chết không có tương quan tuyến tính với nhau )

Ha: Beta khác 0 - Có tương quan tuyến tính giữa x và y

**4.2 Tìm được giá trị thống kê**

Sau khi tính toán ta thu được các giá trị thống kê như sau:

LinregressResult(slope=0.02315981717824699, intercept=-47.28889295770023, rvalue=0.842077431854981, pvalue=1.2187662332084403e-60, stderr=0.0010023908208546095)

Từ kết quả trên ta có thể tìm được đường hồi quy: **y=-47.29 + 0.023x**

**Hệ số tương quan r:** 0.842

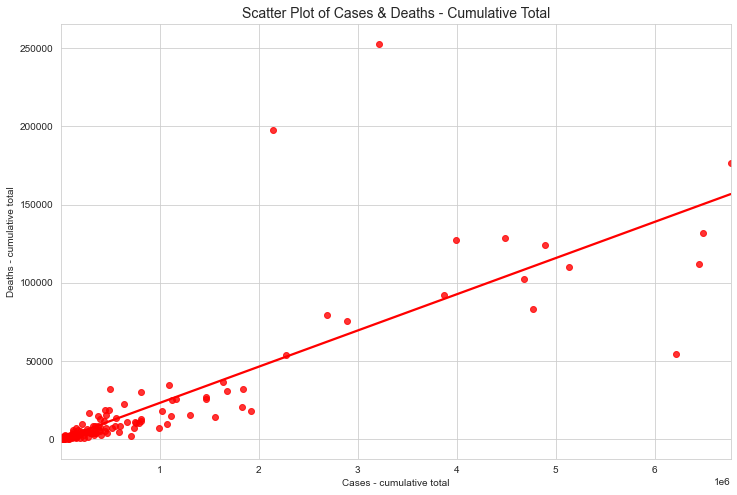
**Từ đó ta có thể tính được Hệ số xác định R- Bình phương:** 0.71

Hệ số xác định cho ta biết được phần tram biến động của y được giải thích bằng phương trình/ đường hồi quy. Nói cách khác, 71% các giá trị số ca tử vong được dự đoán đúng bởi giá trị số ca nhiễm cho trước.

**4.3 Tìm được miền bác bỏ giả thuyết gốc hoặc giá trị P tùy theo độ tin cậy bạn chọn**

Ta có p-value được tính toán ở trên có giá trị: pvalue=1.2187662332084403e-60 nhỏ hơn alpha=0.01 nên trong khoảng tin cậy CI= 99% , có thể bác bỏ giả thuyết Ho: x và y không có tương quan tuyến tính.

Ta có thể mô tả dữ liệu mẫu trên Scatter Plot để xem rõ hơn về tương quan giữa x và y thông qua đường hồi quy tuyến tính.



**4.4 Kết luận cho giả thuyết gốc và giả thuyết ban đầu**

Dựa vào kết quả đã tính toán ở trên, ta có thể bác bỏ giả thuyết Ho: x và y không có tương quan tuyến tính.

Từ đó, ta có thể khẳng định lại giả thuyết ban đầu về mối quan hệ giữa số ca nhiễm và số người chết vì Covid-19 hoàn toàn có tương quan tuyến tính chặt chẽ với nhau. Hơn nữa đó là tương quan đồng biến, nghĩa là số người nhiễm càng nhiều thì số ca chết vì Covid cũng càng cao.

**5. Dự đoán cho giá trị tương lai và ý kiến cá nhân về vấn đề đã nghiên cứu**

Ta có thể thấy, Covid-19 đang thật sự là một vấn đề nghiêm trọng của y tế và xã hội cộng đồng. Bằng cách phân tích dữ liệu mẫu được thu thập bởi WHO, ta có thể phần nào thấy được số ca nhiễm và số người chết trên thế giới.

Bên cạnh đó bằng cách phân tích tương quan giữa ca nhiễm và số người chết, ta có thể khẳng định 99% rằng số ca nhiễm và số người chết có tương quan dương với nhau. (CI=99%). Hơn nữa dựa vào hệ số xác định r2(0.71), ta có thể dự đoán 71% số người tử vong tương lai nếu biết được số ca nhiễm của mỗi khu vực.( thông qua phương trình hồi quy y=-47.29 + 0.023x) .

Tuy nhiên, đ ể tránh ngoại suy phi logic, chúng ta cũng cần phải xem xét về mức độ diễn biến một cách hợp lý trước khi tiên lượng các giá trị trong tương lai.