**COVID-19**

1. **Thông tin về dữ liệu**
2. **Ai cung cấp?**

* Bộ dữ liệu được NYT (The New York Times) đây là một tòa soạn phát hành bộ dữ liệu trên GitHub

1. **Bộ dữ liệu lấy từ đâu? Thu thập & độ tin cậy?**

* **Nguồn gốc:** NYT tổng hợp chuỗi thời gian ca mắc & tử vong từ các sở y tế bang, quận và lãnh thổ của Mỹ, rồi phát hành công khai trên GitHub. Họ dùng chính dữ liệu này để vận hành bản đồ & bài báo trên NYTimes.com ([Link](https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html))
* **Có 2 tệp:**
* Historical (final counts at the end of each day), cho nên dữ liệu sẽ được cập nhật vào cuối ngày => Ổn định, tránh sai sót.
* Live (updated throughout the day) là số cập nhật giữa ngày, có thể mới chỉ là một phần, mỗi nơi cập nhật không cùng giờ, nên dễ lệch và ít tin cậy hơn.
* **Trọng tâm:** chuỗi thời gian lũy kế theo ngày của ca và tử vong ở 3 cấp: toàn quốc (us.csv), bang (us-states.csv), quận (us-counties\*.csv) – phục vụ theo dõi xu hướng & bản đồ trên toàn nước Mỹ. [GitHub](https://github.com/nytimes/covid-19-data)

1. **Có bao nhiêu cột? Ý nghĩa & quan hệ giữa các cột**

* Bao gồm 6 cộtdate, county, state, fips, cases, deaths

**Trong đó:**

* **date:** ngày báo cáo đến nửa đêm giờ ET (Eastern Time - Giờ miền Đông): Đây là tên gọi chung cho múi giờ ở các bang miền Đông của Hoa Kỳ và một phần của Canada
* ET được chọn để có một tiêu chuẩn chốt ngày quốc gia, và vì các bang ở múi giờ muộn hơn thường vẫn công bố đủ sớm để được tính trước nửa đêm ET, giúp dữ liệu ngày-tháng ổn định và so sánh được giữa các vùng
* **county**: tên **quận** (có giá trị đặc biệt: **“New York City”** = gộp 5 quận của NYC (thường **không có FIPS**) hoặc “Unknown” khi chưa xác định nơi cư trú).
* Vì NYT **không liệt kê từng quận riêng lẻ** ở dòng “New York City” mà **gộp cả 5 quận thành một mục duy nhất**. → cột fips để trống cho hàng đó.
  + Bronx (Bronx County – FIPS 36005)
  + Brooklyn (Kings County – 36047)
  + Manhattan (New York County – 36061)
  + Queens (Queens County – 36081)
  + Staten Island (Richmond County – 36085)
* **state**: tên **bang/lãnh thổ** (VD: Washington, New York…).
* **fips: mã địa lý chuẩn (giữ dạng chuỗi 5 ký tự ở cấp county) để join dân số/bản đồ.**
* **cases, deaths: lũy kế (bao gồm confirmed + probable) đến cuối ngày đó; dùng diff theo nhóm (state, county) để suy ra ca/ngày & tử vong/ngày; nên dùng 7-day avg để làm mượt & giảm nhiễu do backlog/điều chỉnh**

**Quan hệ/Ảnh hưởng dữ liệu:**

* **cases/deaths là gốc lũy kế → suy ra chỉ số theo ngày (tăng/giảm/đỉnh-đáy, tốc độ tăng).**
* **fips quyết định join đúng với dân số → chuẩn hóa /100k để so sánh giữa nơi đông & thưa dân.**
* **date theo ngày công bố (không phải ngày khởi phát) → có thể xuất hiện nhảy bậc hoặc giảm lũy kế khi điều chỉnh; cần gắn cờ/loại dị thường khi tính xu hướng.**

**Mục tiêu của bộ dữ liệu NYT là gì?**

* **Thống nhất quốc gia**: gom và chuẩn hóa số **ca/tử vong lũy kế theo ngày** ở **toàn bộ bang/quận** để **theo dõi dịch** và dựng **bản đồ/bài báo** của NYT; đồng thời mở cho cộng đồng nghiên cứu dùng. [GitHub+1](https://github.com/nytimes/covid-19-data?utm_source=chatgpt.com)
* **Mức độ tổng hợp** (không vi mô): không phải line-list theo từng bệnh nhân; không chuẩn “ngày khởi phát”/“ngày lấy mẫu” mà là **ngày công bố**—chọn chuẩn này để ghép được 50+ hệ thống khác nhau.

**Nếu không nhằm “nhân-quả”, nó phân tích cái gì?**

Bộ này **rất mạnh cho mô tả (descriptive / surveillance)**:

* **Xu hướng thời gian**: ca/ngày, tử vong/ngày (từ lũy kế → diff), **7-day average**, đỉnh/đáy, tốc độ tăng/giảm. [GitHub](https://github.com/nytimes/covid-19-data?utm_source=chatgpt.com)
* **Không gian**: so sánh **bang/county**, bản đồ nóng, cụm bùng phát, đồng bộ/khác biệt theo vùng.
* **Đặc điểm sóng dịch**: nhận diện các “wave” theo giai đoạn, độ trễ ca→tử vong, mô tả mô hình báo cáo.
* **Chất lượng dữ liệu**: phát hiện **anomalies**, ngày âm/điều chỉnh, di dời ca giữa county (NYT có mục ghi chú).

1. Columns

**date** (YYYY-MM-DD):

EDA:

* Xu ly flips vi cac contry NYC khong duoc danh so va Unknown