Bài tập thực hành môn Tiền xử lý và xây dựng bộ dữ liệu

Câu a: Mô tả dữ liệu

Câu hỏi

Mô tả dữ liệu, số lượng dòng, cột và thông tin các cột.

Đáp án

Bộ dữ liệu COVID-19 được lưu trong file `data/covid_19_data.csv` có:

- Số cột trong bảng dữ liệu: 8
- Số dòng trong bảng dữ liệu: 306429
- Các cột trong bộ dữ liệu bao gồm: `SNo`, `ObservationDate`, `Province/State`, `Country/Region`, `Last Update`, `Confirmed`, `Deaths`, `Recovered`

Dữ liệu bao gồm thông tin về các ca nhiễm, tử vong và hồi phục của dịch COVID-19 theo từng quốc gia/vùng, được cập nhật theo ngày.

Nhận xét

Bộ dữ liệu có kích thước lớn với hơn 300 nghìn bản ghi, chứa thông tin đầy đủ về diễn biến dịch COVID-19 theo thời gian. Việc có dữ liệu phân chia theo quốc gia và vùng giúp phân tích chi tiết tình hình dịch bệnh ở các khu vực khác nhau. Dữ liệu được tổ chức tốt với các trường thông tin rõ ràng.

Câu b: Phân tích dữ liệu ở Việt Nam

Câu hỏi

Phân tích dữ liệu ở Việt Nam.

Đáp án

Sau khi lọc dữ liệu cho Việt Nam (Vietnam), kết quả phân tích cho thấy:

- Số ca nhiễm ghi nhận vào tháng 2/2021: 28 ca
- Tổng số bản ghi về ca nhiễm ở Việt Nam trong tháng 1 và tháng 2/2021: 59 bản ghi

So sánh với các quốc gia khác trong khu vực:

- Indonesia: 59 bản ghi về ca nhiễm trong tháng 1 và tháng 2/2021
- Philippines: Không có bản ghi nào trong tháng 1 và tháng 2/2021

Nhận xét

Việt Nam có số ca nhiễm thấp so với các quốc gia trong khu vực, cho thấy hiệu quả của công tác phòng chống dịch. Indonesia có số bản ghi tương đương Việt Nam, trong khi Philippines có thể thiếu dữ liệu hoặc

cách ghi nhận khác. Điều này phản ánh sự khác biệt trong cách các quốc gia ứng phó với dịch bệnh và hệ thống ghi nhận dữ liệu.

Câu c: Phân tích dữ liệu ở Trung Quốc

Câu hỏi

Phân tích dữ liệu ở Trung Quốc.

Đáp án

Phân tích dữ liệu của Trung Quốc (Mainland China) cho tháng 2/2021:

- Tổng số bản ghi theo tỉnh: 896 bản ghi (mỗi tỉnh có 28 bản ghi)
- Tổng số ca nhiễm tính đến ngày 28/2/2021: 2,513,697 ca
- Các tỉnh có số ca nhiễm cao nhất:
    - Hubei: 1,908,212 ca
    - Guangdong: 60,700 ca
    - Zhejiang: 60,182 ca

Nhận xét

Trung Quốc là quốc gia chịu ảnh hưởng nặng nề nhất trong giai đoạn đầu của đại dịch, với tỉnh Hubei (tâm dịch Vũ Hán) chiếm đa số ca nhiễm. Sự chênh lệch rất lớn giữa số ca nhiễm ở Hubei và các tỉnh khác cho thấy tính chất cục bộ của đợt bùng phát đầu tiên và hiệu quả của các biện pháp phong tỏa nghiêm ngặt được áp dụng.

Câu d: Số ca nhiễm ở Việt Nam tháng 1-2 năm 2021

Câu hỏi

In ra số dữ liệu về ca lây nhiễm nhiều nhất trong khoảng tháng 01 và 02 tại Việt Nam (Lấy năm 2021).

Đáp án

Kết quả phân tích cho thấy có 59 bản ghi về ca lây nhiễm COVID-19 ở Việt Nam trong khoảng tháng 01 và 02 năm 2021. Số liệu này phản ánh tình hình dịch bệnh ở Việt Nam trong giai đoạn đầu năm 2021.

Nhận xét

Số lượng bản ghi (59) tương đối thấp so với các quốc gia khác, điều này phản ánh công tác kiểm soát dịch bệnh hiệu quả của Việt Nam trong giai đoạn đầu năm 2021.

Câu e: So sánh với Indonesia và Philippines

Câu hỏi

Thực hiện tương tự câu d) cho Indonesia và Philippines.

Đáp án

Kết quả phân tích cho thấy:

- Indonesia: có 59 bản ghi về ca lây nhiễm COVID-19 trong khoảng tháng 01 và 02 năm 2021
- Philippines: không có bản ghi nào về ca lây nhiễm trong cùng khoảng thời gian (tháng 01 và 02 năm 2021)

Câu f: Thống kê số lượng record theo tỉnh của Trung Quốc

Câu hỏi

Thống kê số lượng record theo từng tỉnh của Trung Quốc trong tháng 02/2021.

Đáp án

Kết quả thống kê dữ liệu của Trung Quốc (Mainland China) trong tháng 02/2021 cho thấy:

- Mỗi tỉnh có chính xác 28 bản ghi (tương ứng với 28 ngày trong tháng 2)
- Tổng cộng có 32 tỉnh/vùng với dữ liệu được ghi nhận
- Tổng số record: 896 bản ghi (32 tỉnh/vùng × 28 ngày)

Nhận xét

Điều này cho thấy hệ thống báo cáo dịch tễ của Trung Quốc vận hành nhất quán và có tính hệ thống. Tính đồng nhất này giúp cho việc phân tích và so sánh dữ liệu giữa các vùng miền trở nên dễ dàng và chính xác hơn.

Câu g: Thống kê số ca nhiễm theo tỉnh của Trung Quốc

Câu hỏi

Đếm số lượng ca nhiễm mới theo từng tỉnh của Trung Quốc trong tháng 02/2021.

Đáp án

Tổng số ca nhiễm tích lũy theo tỉnh của Trung Quốc trong tháng 02/2021:

- Tổng số ca nhiễm trên toàn quốc: 2,513,697 ca
- Các tỉnh có số ca nhiễm cao nhất:
    - Hubei: 1,908,212 ca
    - Guangdong: 60,700 ca
    - Zhejiang: 60,182 ca
    - Shanghai: 49,269 ca
    - Henan: 36,499 ca
    - Zhejiang: 36,935 ca

Nhận xét

Phân tích số liệu cho thấy sự chênh lệch rất lớn giữa Hubei (tâm dịch Vũ Hán) với gần 2 triệu ca nhiễm và các tỉnh khác. Hubei chiếm khoảng 76% tổng số ca nhiễm của toàn Trung Quốc trong thời gian này. Các tỉnh có số ca nhiễm cao tiếp theo như Guangdong, Zhejiang, Shanghai là những trung tâm kinh tế lớn với mật độ dân

số cao và lưu lượng di chuyển nhiều. Trong khi đó, những vùng xa xôi và ít dân cư hơn như Tibet chỉ có 28 ca nhiễm.

Câu h: In ra dữ liệu tử vong ở Trung Quốc từ 01/02/2021 đến 15/02/2021

Câu hỏi

In ra dữ liệu về số ca tử vong ở Trung Quốc từ ngày 01/02/2021 đến ngày 15/02/2021.

Đáp án

Dữ liệu về số ca tử vong ở Trung Quốc từ 01/02/2021 đến 15/02/2021:

```
Dữ liệu về số ca tử vong ở Trung Quốc từ ngày 01/02/2021 đến ngày 15/02/2021:
       ObservationDate Province/State Country/Region  Deaths
283802      02/01/2021          Anhui   Mainland China      6
283803      02/01/2021        Beijing   Mainland China      9
283804      02/01/2021      Chongqing   Mainland China      6
...             ...            ...              ...    ...
284447      02/15/2021        Xinjiang   Mainland China      3
284448      02/15/2021         Yunnan   Mainland China      2
284449      02/15/2021       Zhejiang   Mainland China      1
```

Nhận xét

Trung Quốc đã kiểm soát được số ca tử vong trong giai đoạn đầu năm 2021, với số ca tử vong thấp ở hầu hết các tỉnh. So với giai đoạn đầu dịch (2020), số ca tử vong đã giảm đáng kể, phản ánh hiệu quả của các biện pháp y tế và kiểm soát dịch bệnh. Tuy nhiên, cần lưu ý rằng có thể có sự khác biệt trong cách ghi nhận và báo cáo số liệu.

Câu i: So sánh dữ liệu Nhật Bản

Câu hỏi

Thực hiện tương tự cho Nhật Bản (Japan).

Đáp án

Phân tích dữ liệu của Nhật Bản trong tháng 02/2021 cho thấy:

- Tổng số bản ghi: 1,372 bản ghi (49 tỉnh/vùng × 28 ngày)
- Mỗi tỉnh/vùng có 28 bản ghi (tương ứng với 28 ngày trong tháng 2)
- Tổng số ca nhiễm tích lũy: 11,627,561 ca
- Các tỉnh có số ca nhiễm cao nhất:
    - Tokyo: 2,978,730 ca
    - Osaka: 1,283,309 ca
    - Kanagawa: 1,210,638 ca

Nhận xét

Nhật Bản có số lượng bản ghi và tổng số ca nhiễm cao hơn nhiều so với Trung Quốc, với 49 tỉnh/vùng được ghi nhận dữ liệu. Tokyo là tâm dịch chính của Nhật Bản với gần 3 triệu ca nhiễm, chiếm khoảng 25.6% tổng số ca nhiễm của cả nước. Các vùng đô thị lớn khác như Osaka và Kanagawa cũng ghi nhận số ca nhiễm cao. Sự phân bố ca nhiễm ở Nhật Bản tương đối đồng đều hơn so với Trung Quốc, cho thấy dịch bệnh đã lan rộng khắp cả nước thay vì tập trung chủ yếu ở một vùng. Hệ thống báo cáo của Nhật Bản cũng rất chi tiết và đầy đủ, với dữ liệu được cập nhật đều đặn cho tất cả các tỉnh/vùng.

Câu j: So sánh số ca nhiễm mới tại Việt Nam

Câu hỏi

So sánh số ca nhiễm mới tại Việt Nam giữa tháng 05/2020 và tháng 05/2021.

Đáp án

Kết quả so sánh số ca nhiễm mới tại Việt Nam giữa hai thời điểm:

- Tổng số ca nhiễm mới tại Việt Nam trong tháng 5/2020: 58 ca
- Tổng số ca nhiễm mới tại Việt Nam trong tháng 5/2021: 3,966 ca
- Tỷ lệ tăng: 6,737.9%

Biểu đồ đường thể hiện số ca nhiễm mới trong 2 tháng trên cho thấy sự khác biệt rõ rệt, với đường biểu diễn tháng 5/2021 cao hơn nhiều so với tháng 5/2020.

Nhận xét

Số liệu cho thấy sự gia tăng đột biến về số ca nhiễm mới tại Việt Nam giữa tháng 5/2020 và tháng 5/2021. Trong khi tháng 5/2020 ghi nhận tổng cộng chỉ 58 ca nhiễm mới, con số này đã tăng lên 3,966 ca vào tháng 5/2021, tương đương mức tăng gần 6,738%. Sự gia tăng mạnh này phản ánh đợt bùng phát dịch mới vào năm 2021 nghiêm trọng hơn nhiều so với năm 2020. Tháng 5/2020, Việt Nam kiểm soát dịch bệnh rất tốt với rất ít ca nhiễm mới, nhưng tháng 5/2021 đánh dấu sự bùng phát đáng kể của làn sóng dịch mới tại Việt Nam.

Câu k: So sánh số ca nhiễm COVID-19 tại Việt Nam, Indonesia và Philippines

Câu hỏi

Vẽ biểu đồ so sánh số ca nhiễm COVID-19 tại Việt Nam, Indonesia và Philippines trong tháng 1 và tháng 2 năm 2021.

Đáp án

Kết quả phân tích và so sánh số ca nhiễm COVID-19 tại ba quốc gia:

- Vietnam: Số ca nhiễm cao nhất là 2,448 ca vào ngày 28/02/2021
- Indonesia: Số ca nhiễm cao nhất là 1,334,634 ca vào ngày 28/02/2021
- Philippines: Số ca nhiễm cao nhất là 576,352 ca vào ngày 28/02/2021

So sánh số ca nhiễm mới ở Việt Nam giữa tháng 5/2020 và tháng 5/2021:

- Tổng số ca nhiễm mới tháng 5/2020: 58

- Tổng số ca nhiễm mới tháng 5/2021: 3966
- Tăng: 6737.9%

Nhận xét

Dữ liệu cho thấy sự chênh lệch rất lớn về số ca nhiễm giữa ba quốc gia, với Indonesia có số ca nhiễm cao nhất, tiếp theo là Philippines, và Việt Nam có số ca nhiễm thấp nhất. Điều này phản ánh các chiến lược khác nhau trong việc kiểm soát dịch bệnh và quy mô dân số khác nhau.

Đặc biệt, Việt Nam đã chứng kiến sự tăng vọt về số ca nhiễm mới giữa tháng 5/2020 và tháng 5/2021, với mức tăng hơn 6700%. Điều này cho thấy làn sóng COVID-19 năm 2021 nghiêm trọng hơn nhiều so với năm 2020. Trong khi tháng 5/2020, Việt Nam kiểm soát dịch bệnh rất tốt với rất ít ca nhiễm mới, thì tháng 5/2021 đánh dấu sự bùng phát đáng kể với số ca nhiễm mới tăng lên gần 4000 ca.