**ĐẠI HỌC ĐÀ NẴNG**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

**KHOA KHOA HỌC CÔNG NGHỆ TIÊN TIẾN**



BÁO CÁO MÔN HỌC

PHƯƠNG PHÁP THỐNG KÊ

VÀ

PHÂN TÍCH DỮ LIỆU

|  |  |
| --- | --- |
|  | GVHD: TS. Lê Dân  SVTH: Nguyễn Đình Mẫn  Lớp: 18PFIEV3  Nhóm: 18.87 |

Đà Nẵng, 4/2021

PHỤ LỤC

[**I.** **PHƯƠNG PHÁP DOWNLOAD VÀ LƯU TRỮ FILE DỮ LIỆU** 3](#_Toc68222123)

[**1.** **Download dữ liệu từ trang web** 3](#_Toc68222124)

[**2.** **Thực hiện xử lý file có đuôi .csv bằng Excel** 4](#_Toc68222125)

[**II.** **PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU TRÊN PHẦN MỀM SPSS** 5](#_Toc68222126)

[**1.** **Cách import dữ liệu vào SPSS** 5](#_Toc68222127)

[**2.** **Làm sạch dữ liệu bằng SPSS:** 7](#_Toc68222128)

[**3.** **Mô tả dữ liệu và phân tích kết quả** 9](#_Toc68222129)

[**3.1** **Mô tả dữ liệu bằng Bảng trong SPSS** 9](#_Toc68222130)

[**3.2** **Mô tả dữ liệu bằng Đồ thị** 11](#_Toc68222131)

[**3.3** **Mô tả đồ thị bằng đại lượng thống kê** 14](#_Toc68222132)

[**4.** **Kiểm định** 15](#_Toc68222133)

[**a.** **Kiểm định trung bình tổng thể** 15](#_Toc68222134)

[**b.** **Kiểm định 1 mẫu** 16](#_Toc68222137)

[**c.** **Kiểm định mẫu cặp** 17](#_Toc68222138)

1. **PHƯƠNG PHÁP DOWNLOAD VÀ LƯU TRỮ FILE DỮ LIỆU**
2. **Download dữ liệu từ trang web**

* Dữ liệu được lấy từ trang web: <https://ourworldindata.org/coronavirus>
* Chọn **Download Dataset**  và chọn file định dạng .csv để tải về
* Nguồn dữ liệu đã được kiểm định, có mức độ tin cậy cao: Đây là dự án của  [Global Change Data La](https://global-change-data-lab.org/) hợp tác với Đại học Oxford ( UK)

**Graphical user interface, application

Description automatically generated**

Graphical user interface, text, application, email

Description automatically generated

1. **Thực hiện xử lý file có đuôi .csv bằng Excel**

( Việc xử lý được thực hiện trên Excel - Office 365)

**Bước 1:** Mở trang Excel mới

**Bước 2:** Chọn **Data > Get Data > From File > From Text/Csv**

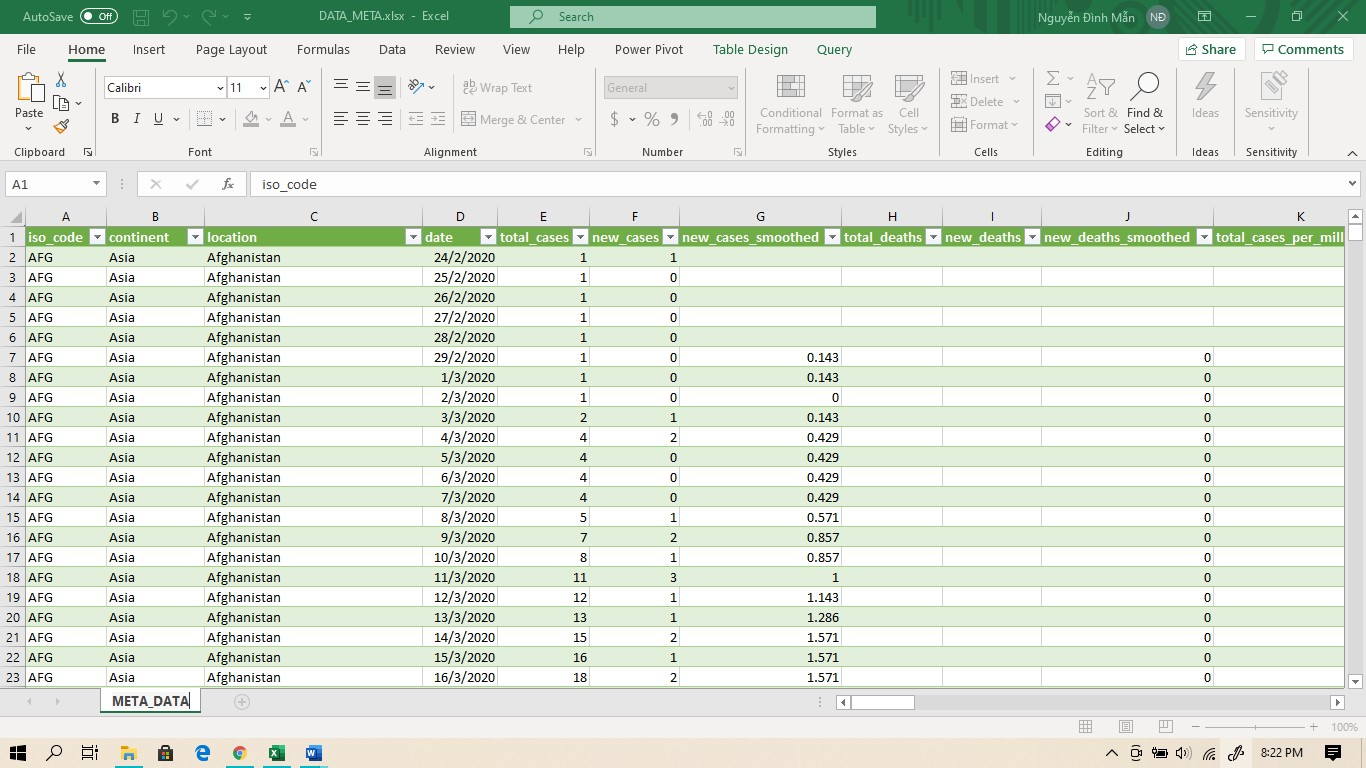
**Bước 3:** Chọn file có đuôi \*.csv muốn import vào Excel, sau đó nhấn **Load.**

Quá trình load file dữ liệu:

Graphical user interface, application, table, Excel

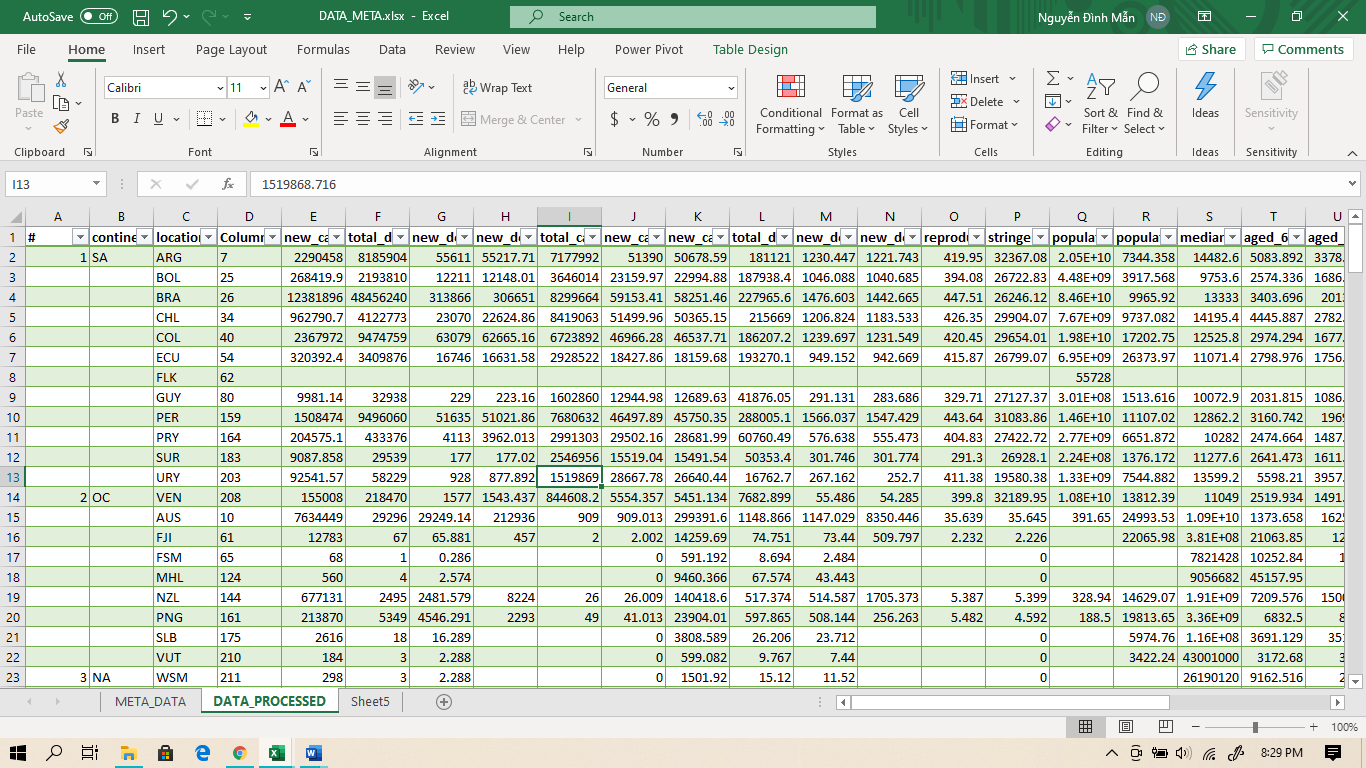
Description automatically generated

* Dữ liệu sau khi Import từ file \*.csv



**Bước 4:** Thực hiện tổ chức lại dữ liệu

Dữ liệu sau khi được tổ chức lại:



**Bước 5:** Lưu file với tên META\_DATE.xlxx

1. **PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU TRÊN PHẦN MỀM SPSS**
2. **Cách import dữ liệu vào SPSS**

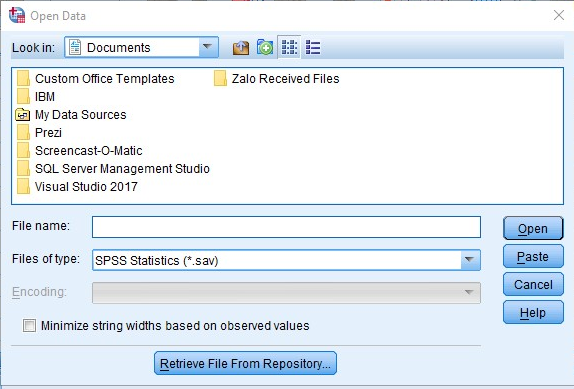
**Bước 1:** Mở phần mềm IBM SPSS Statistic 22

**Bước 2:**  Chọn New Dataset

Graphical user interface, application

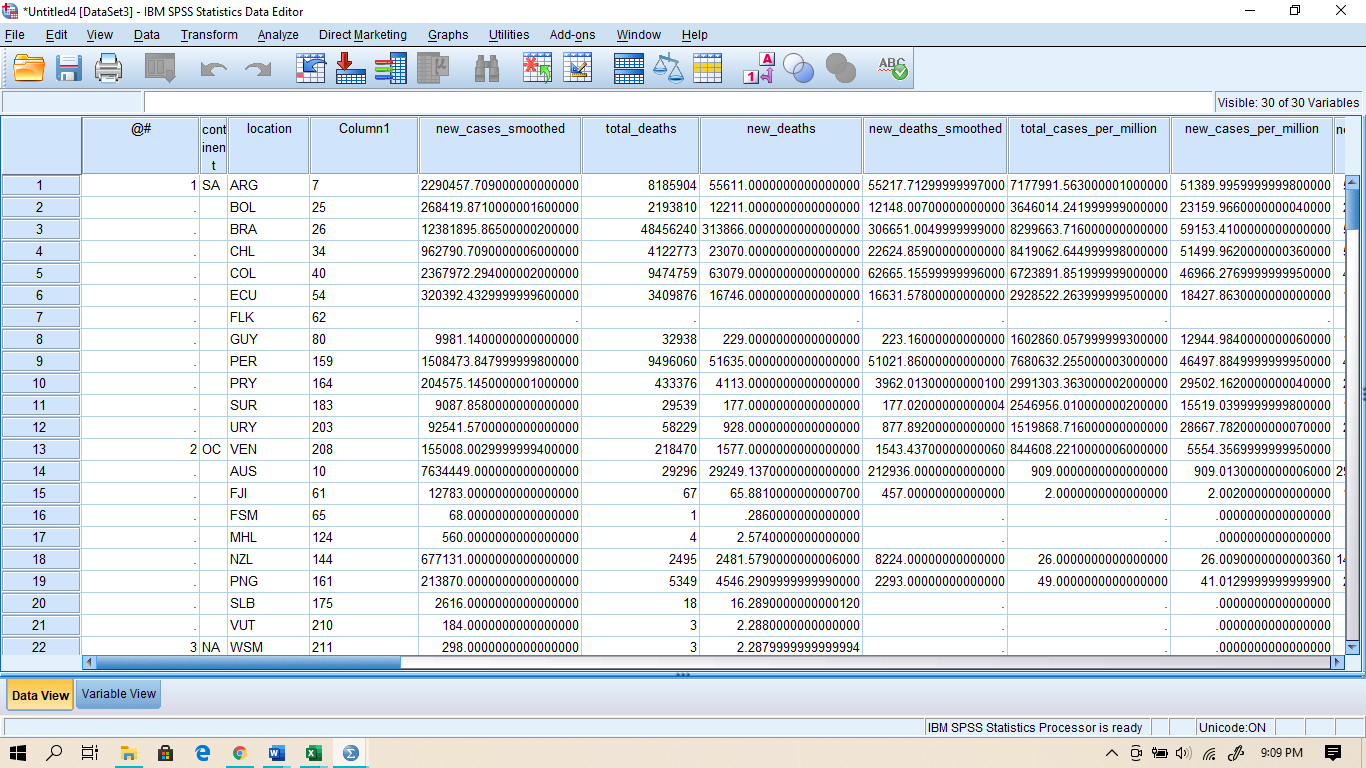
Description automatically generated

**Bước 3:** Chọn **File > Open > Data**, xuất hiện hộp thoại **Open Data**

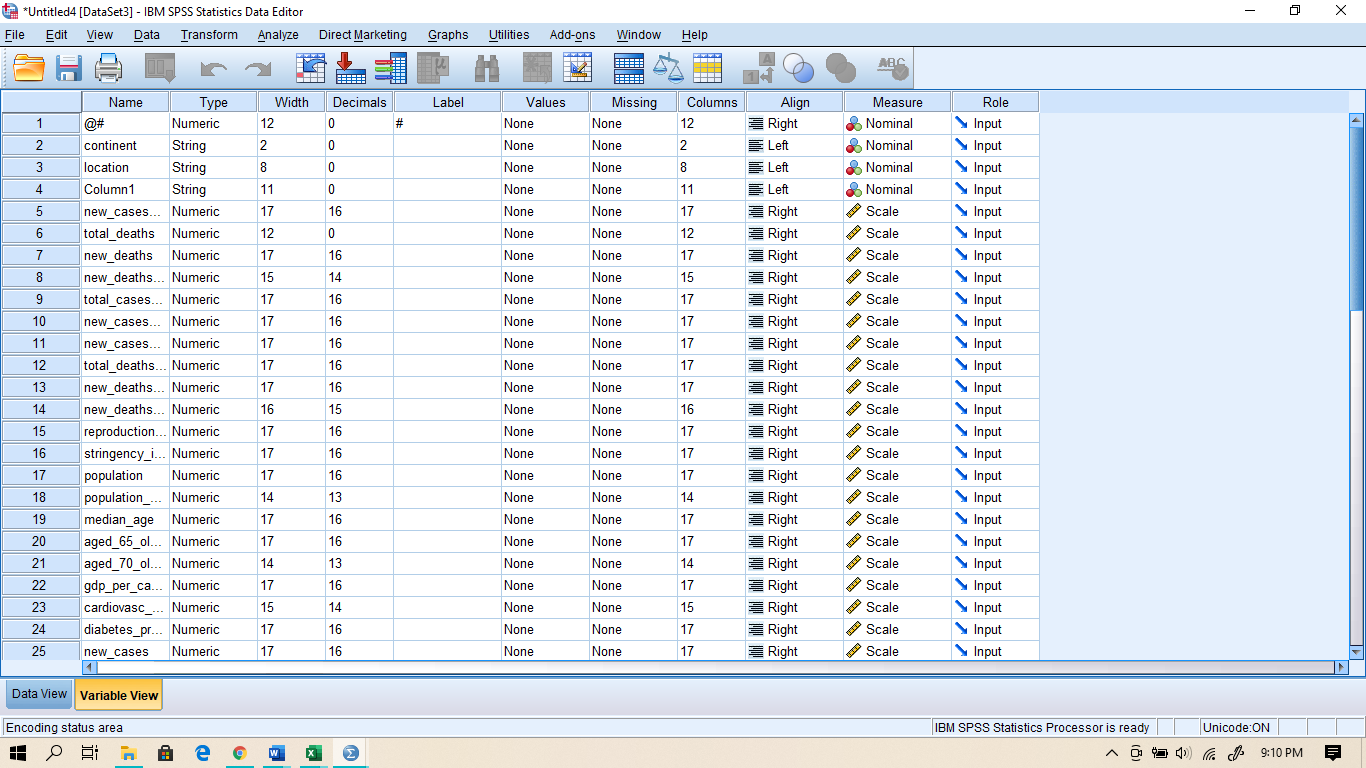
****

**Bước 4:**

* Chọn thư mục lưu file dữ liệu
* Điều chỉnh **File of type** thành định dạng **Excel**
* Chọn file META\_DATA.xlsx. Sau đó nhấn **Open**
* File dữ liệu sẽ được hiển thị như sau:



* Chuyển sang Variable View để xem thuộc tính của các biến và các tổ chức dữ liệu



Trong bảng Variable View cho thấy có 30 biến trong file dữ liệu được import vào.

**Bước 5:** Chọn **File > Save As** với tên DATA\_SPSS.sav

1. **Làm sạch dữ liệu bằng SPSS:**

Do trong quá trình thực hiện cập nhật dữ liệu, các số liệu có thể bị nhập sai hoặc tại thời điểm nhập dữ liệu, một số thông tin vẫn chưa có nên chưa thể nhập. Ví dụ như các biến vẫn chưa nhập dữ liệu, nhập số âm,…

Do vậy trước khi dữ liệu được phân tích, cần làm sạch để hạn chết sai số trong quá trình phân tích số liệu.

**Cách thực hiện:**

Chọn: Analyze > Descriptive Statistics > Frequencies

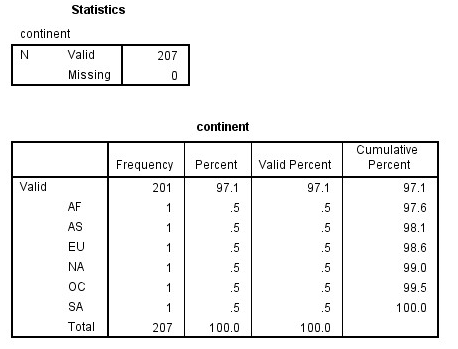
Chọn các biến muốn làm sạch, nhấn OK

Sau khi xử lý, bảng tần số xuất hiện như sau



Nhận xét: ta thấy có lỗi Missing, tức là lỗi hàng trống trong dữ liệu nhập vào.

* Hoặc đối với biến continent:



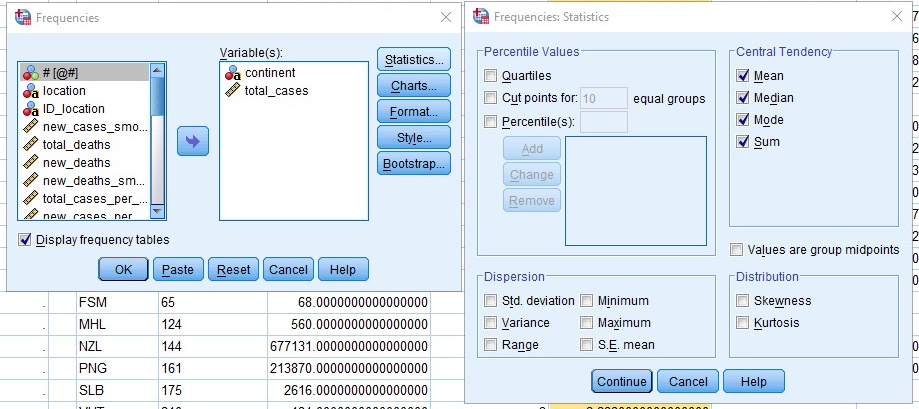
Nhận xét: ta thấy lỗi Missing = 0, điều này cho biết dữ liệu nhập vào không sai sót.

Ngoài cách lọc tần số nói trên, ta cũng có thể lọc dữ liệu bằng cách dùng bảng kết hợp hay xử lý trực tiếp trên excel.

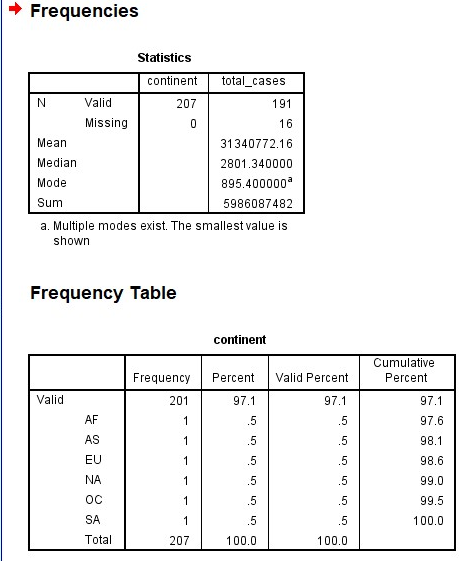
1. **Mô tả dữ liệu và phân tích kết quả** 
   1. **Mô tả dữ liệu bằng Bảng trong SPSS**
2. **Mô tả một biến**

**- Cách thực hiện mô tả theo tần số xuất hiện:**

* Chọn **Analyze > Descriptive Statistics > Frequencies**
* Chọn biến cần mô tả
* Nhấn vào nút Statistics để chọn các thông số
* Nhấn OK



* Sau khi thực hiện xong, của sổ Output sẽ xuất hiện:

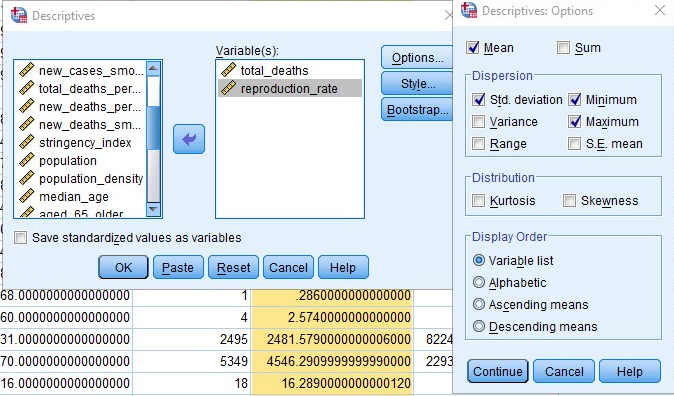


**Nhận xét:**

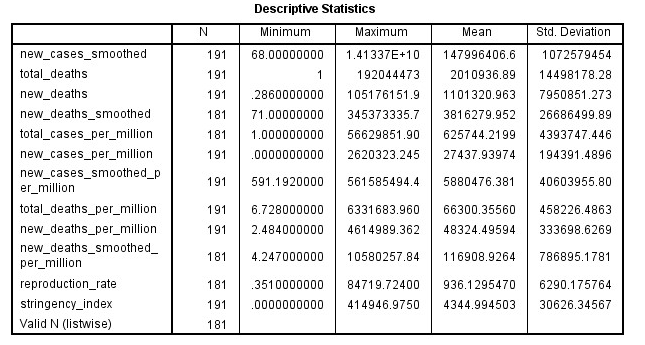
Bảng Frequencies, thể hiện thông số của các biến và các lỗi có thể có trong dữ liệu của biến.

Bảng Frequency Table thể hiện các giá trị tần số xuất hiện của các giá trị trong biến.

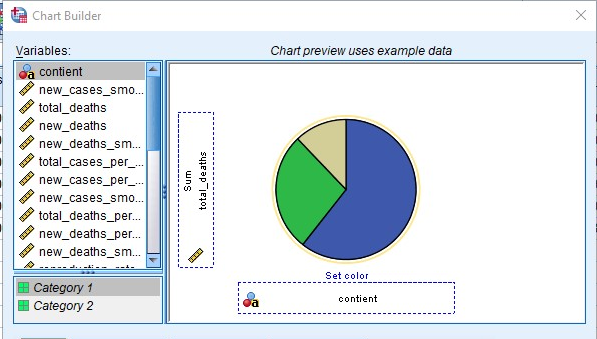
* **Cách thực hiện mô tả giá trị trung bình Mean**
* Chọn **Analyze > Descriptive Statistics > Descriptives**
* Chọn biến cần mô tả
* Nhấn vào nút Options để chọn các thông số
* Nhấn OK

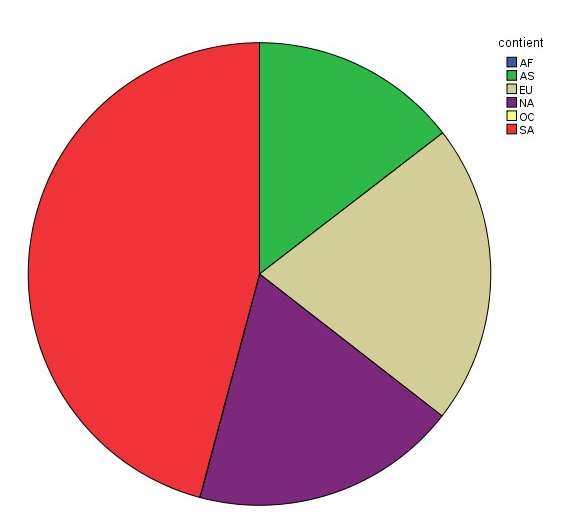


* Bảng Output

 Bảng thể hiện các giá trị trung bình của các biến được chọn.

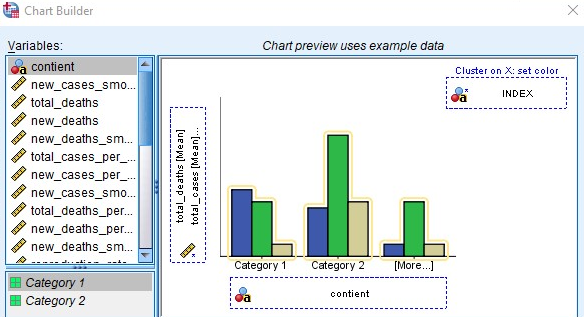
* 1. **Mô tả dữ liệu bằng Đồ thị**
* **Cách thực hiện:**
* Chọn **Graphs > Chart Builder**
* Chọn loại biểu đồ phù hợp với dữ liệu
* Chọn các biến cho biểu đồ
* Đồ thị thể hiện tổng số ca đã tử vong do covid-19 trên 6 lục địa





Nhận xét:

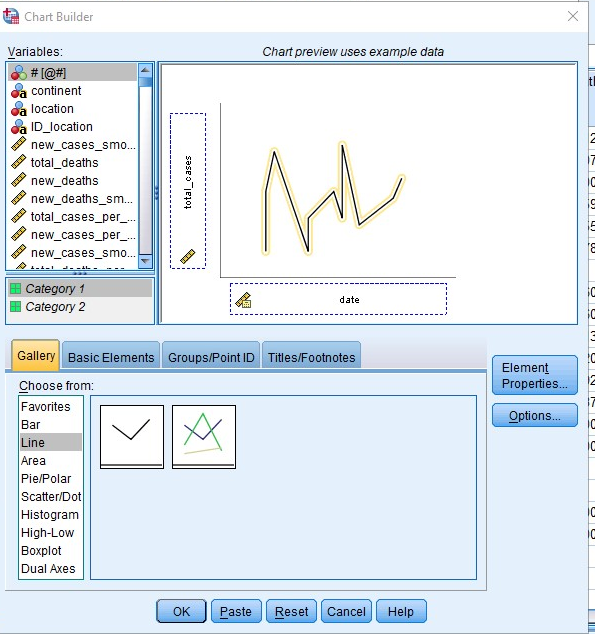
* Từ đồ thị trên ta thấy rằng, tổng số ca tử vong do covid 19 tập trung chủ yếu ở vùng SA( Bắc Mỹ)
* Hai vùng OC( Châu Đại Dương) và AF (Châu Phi) có số lượng người tử vong ít nhất.
* Đồ thị thể hiện số ca mắc,tử vong và ca mắc mới do Covid-19

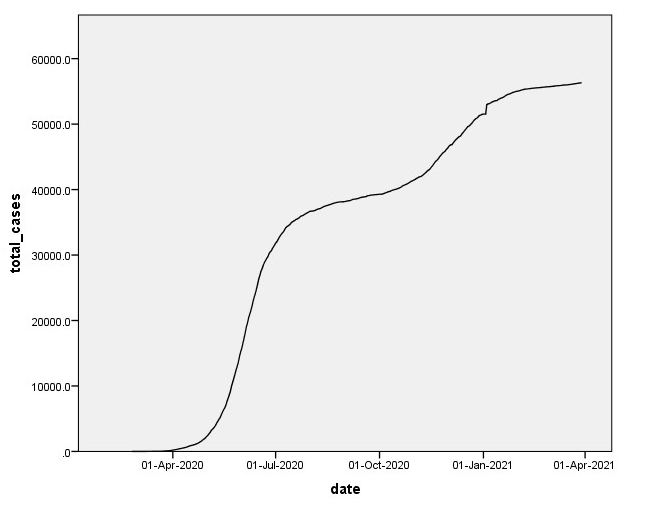




Nhận xét:

* Ta thấy tằng số ca mắc tập trung chủ yếu ở vùng SA( Nam Mỹ) Bên cạnh đó số ca tử vong do Covid-19 cũng nhiều nhất ở khu vực này.
* Số ca mắc mới nhiều nhất ở khu vực AF (Châu Phi).
* Đồ thị thể hiện số ca mắc theo ngày của Châu Á



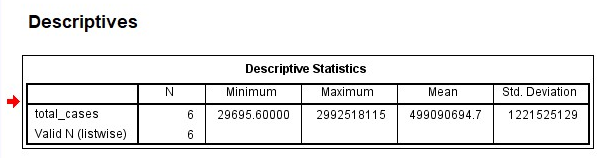


Nhận xét:

* Từ đồ thị ta thấy rằng, số ca mắc của Asia tăng trong khoảng thời gian từ tháng 2/2020 đến 4/2021
* Cả chỉ số tăng nhanh trong khoảng thời gian này.
  1. **Mô tả đồ thị bằng đại lượng thống kê**

**Dữ liệu sử dụng được thu thập từ ngày 7/2/2020-31/3/2021**

* Bài này thực hiện trên biến Total\_case (numeric) trong bảng Continents.
* Các đại lượng thống kê dùng cho bài này: Mean(), Std.deviation(), Minimum, Maximum.
* **Cách thực hiện:**
* Chọn **Analyze > Descriptive Statistics > Descriptives**
* Chọn Options, sau đó chọn các đại lường cần tính.
* Nhấn OK.
* Bảng Output



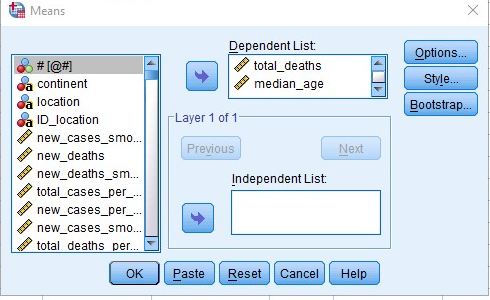
**Phân tích:**

* Tổng số ca mắc ít nhất trong 6 khu vực trong khoảng thời gian thu thập số liệu là 29695 ca.
* Số ca mắc nhiều nhất được ghi nhận trên 6 khu vực trong khoảng thời gian thu thập là 499090694 ca.
* Độ lệch chuẩn của dữ liệu sau khi xử lý là 1221525129 ca.

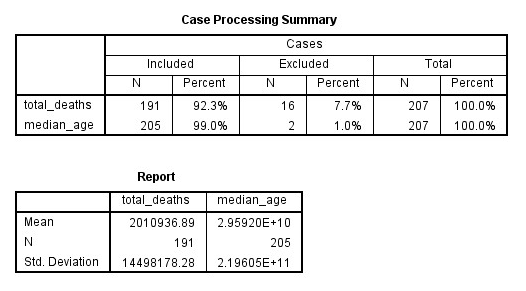
1. **Kiểm định**
2. **Kiểm định trung bình tổng thể**

**Cách thực hiện:**

* Chọn **Analyze > Compare Means > Means**
* Đưa biến Total\_deaths, median\_age vào Dependent List
* Nhấn OK



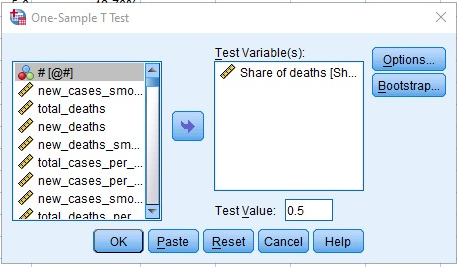
* Kết quả Output:

****

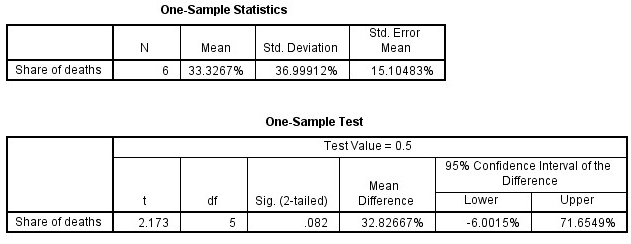
1. **Kiểm định 1 mẫu**

**Giả thuyết: Khả năng mắc bệnh Covid ở các độ tuổi là 50%**

* **Cách thực hiện:**
* Chọn **Analyze > Compare Means > One-Sample T Test**
* Đưa biến Share of death vào Test Variables
* Nhập Test Value là 50%



* Bảng Output

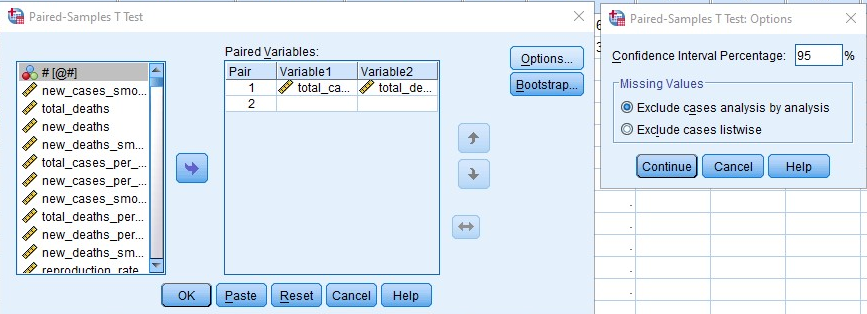


**Phân tích:** Quan sát trong bảng One-Sample Test, ta thấy giá trị Sig.(2-tailed) bằng 0.082, lớn hơn 0.05 . Vậy ta ***chưa đủ cơ sở để bác bỏ giả thuyết*** ban đầu, Khả năng mắc bệnh Covid ở các độ tuổi là 50%

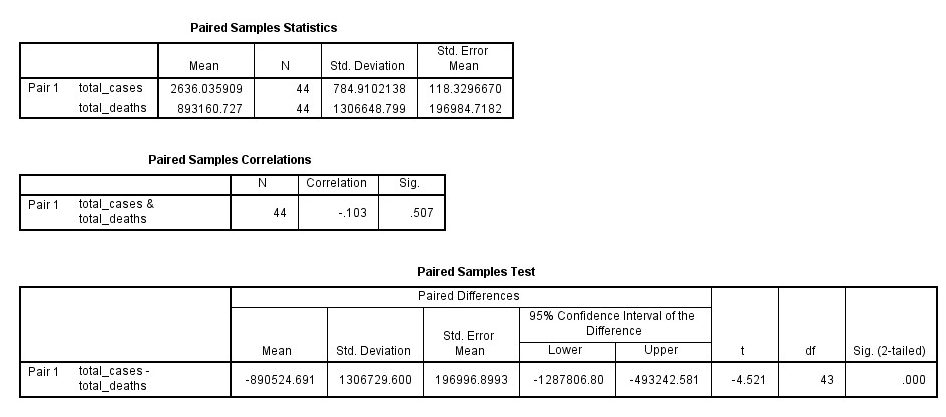
1. **Kiểm định mẫu cặp**

**Giả thuyết:** Không có sự khác nhau giữa số ca mắc mới và tổng số ca tử vong do Covid

* **Cách thực hiện:**
* Chọn **Analyze > Compare Means > Paired Sample T Test.**
* Đưa biến new\_case vào Pairs\_Variables- variable1
* Đưa biến total\_case vào Pairs\_Variables-variable2
* Thực hiện kiểm định với độ tin cậy 95%(mức ý nghĩa 5%)



* Kết quả Output



**Phân tích:** Từ bảng Pair Samples Test, ta thấy rằng chỉ số Sig.(2-tailed) = 0.00 chỉ số này bé hơn 0.05. Vì vậy ta có thể **bác bỏ giả thuyết trên**