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

1. Download dữ liệu từ trang web 3

2. Thực hiện xử lý file có đuôi .csv bằng Excel 4

II. PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU TRÊN PHẦN MỀM SPSS 5

1. Cách import dữ liệu vào SPSS 5

2. Làm sạch dữ liệu bằng SPSS: 7

3. Mô tả dữ liệu và phân tích kết quả 9

3.1 Mô tả dữ liệu bằng Bảng trong SPSS 9

3.2 Mô tả dữ liệu bằng Đồ thị 11

3.3 Mô tả đồ thị bằng đại lượng thống kê 14

4. Kiểm định 15

a. Kiểm định trung bình tổng thể 15

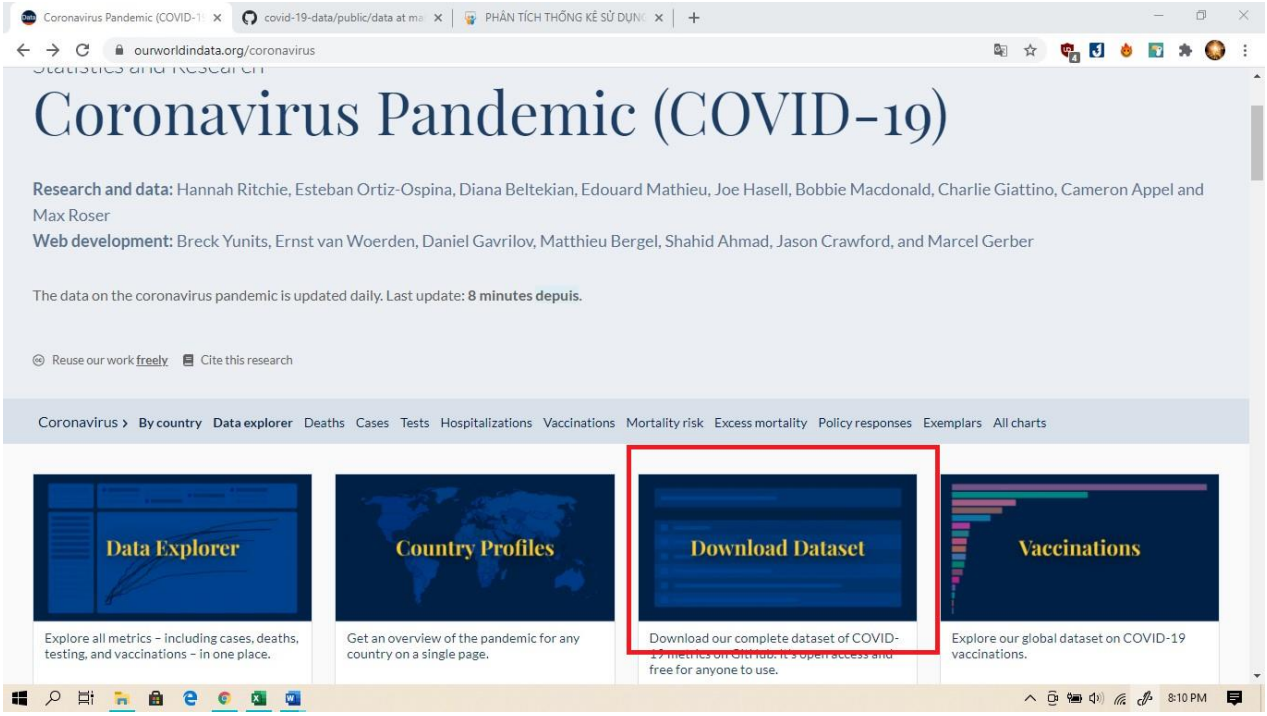
b. Kiểm định 1 mẫu 16

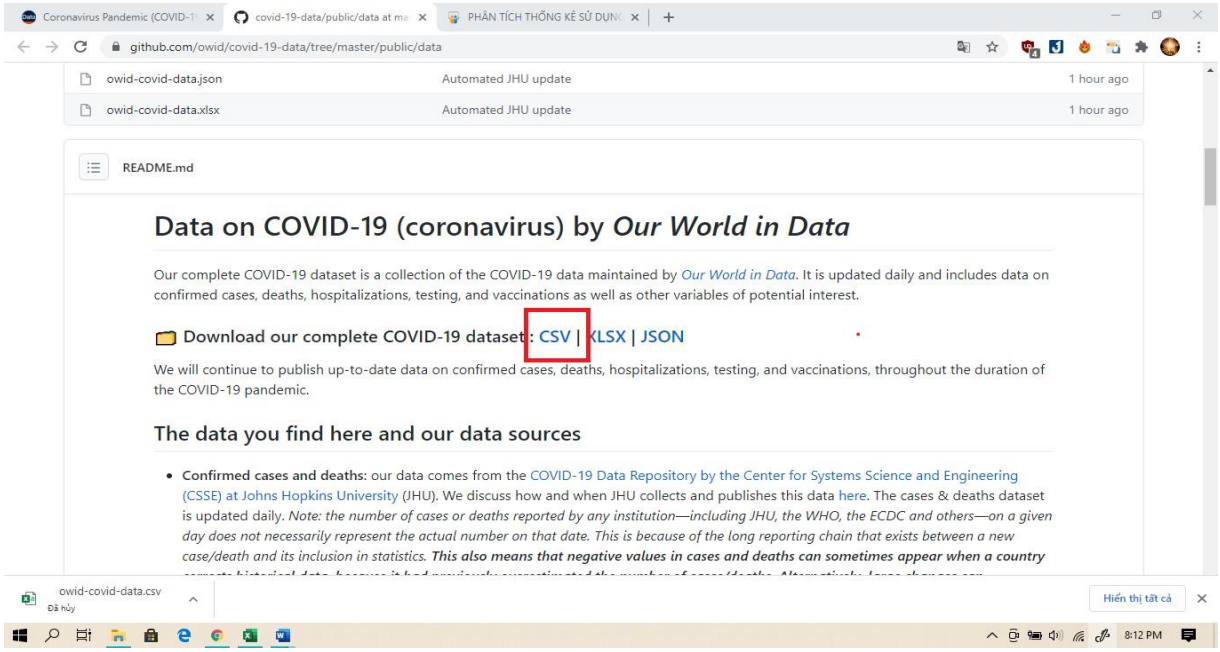
c. Kiểm định mẫu cặp 17

I. PHƯƠNG PHÁP DOWNLOAD VÀ LƯU TRỮ FILE DỮ LIỆU

1. 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 hợp tác với Đại học Oxford (UK)

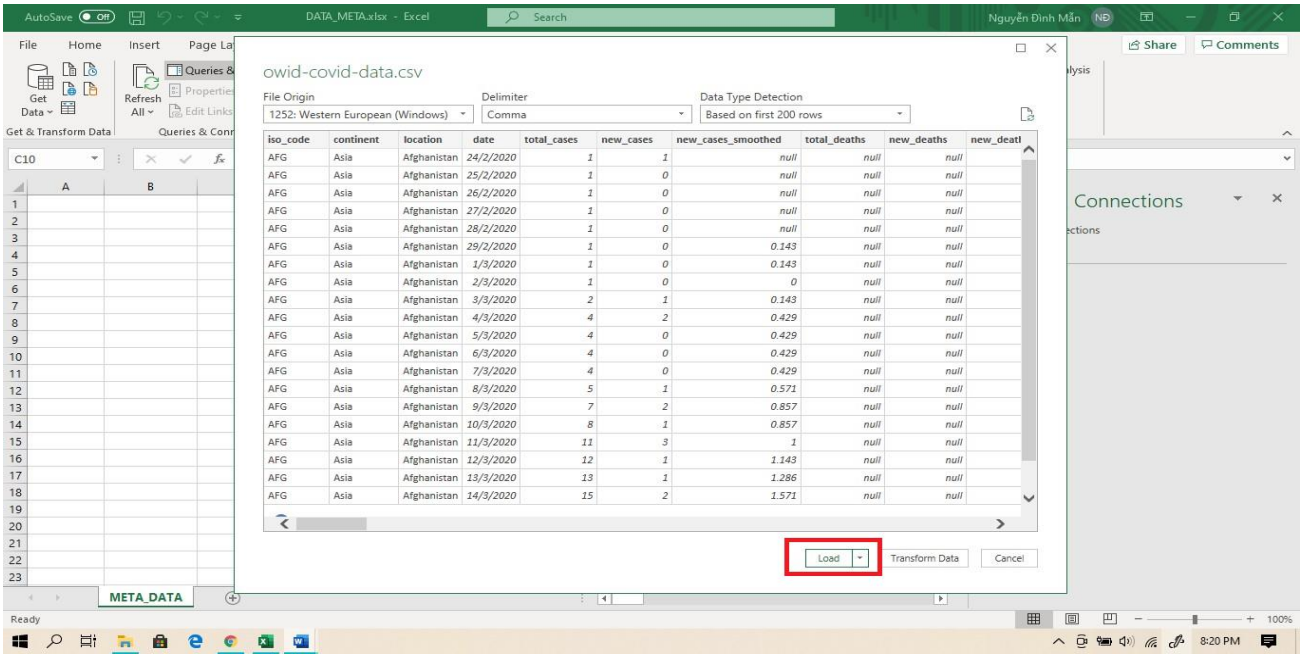




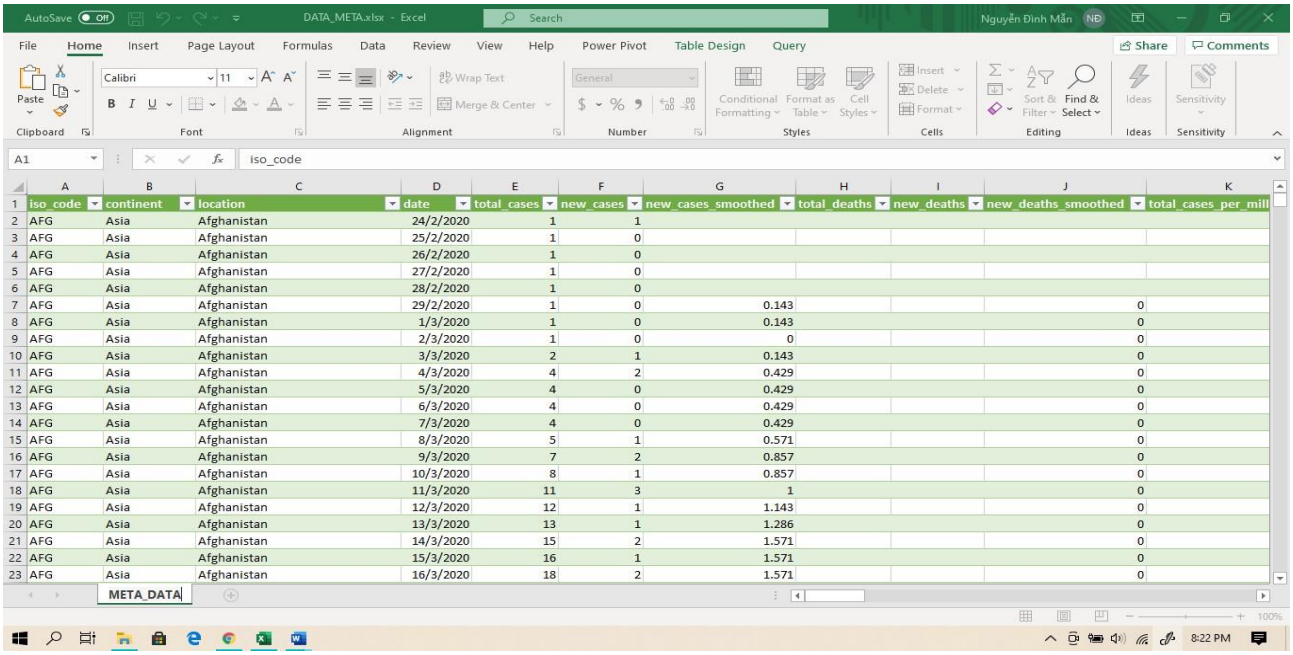
2. **Thực hiện xử lý file có đuôi .csv bằng Excel**
   (Việc xử lý được thực hiện trên Excel - Office 365)

   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:

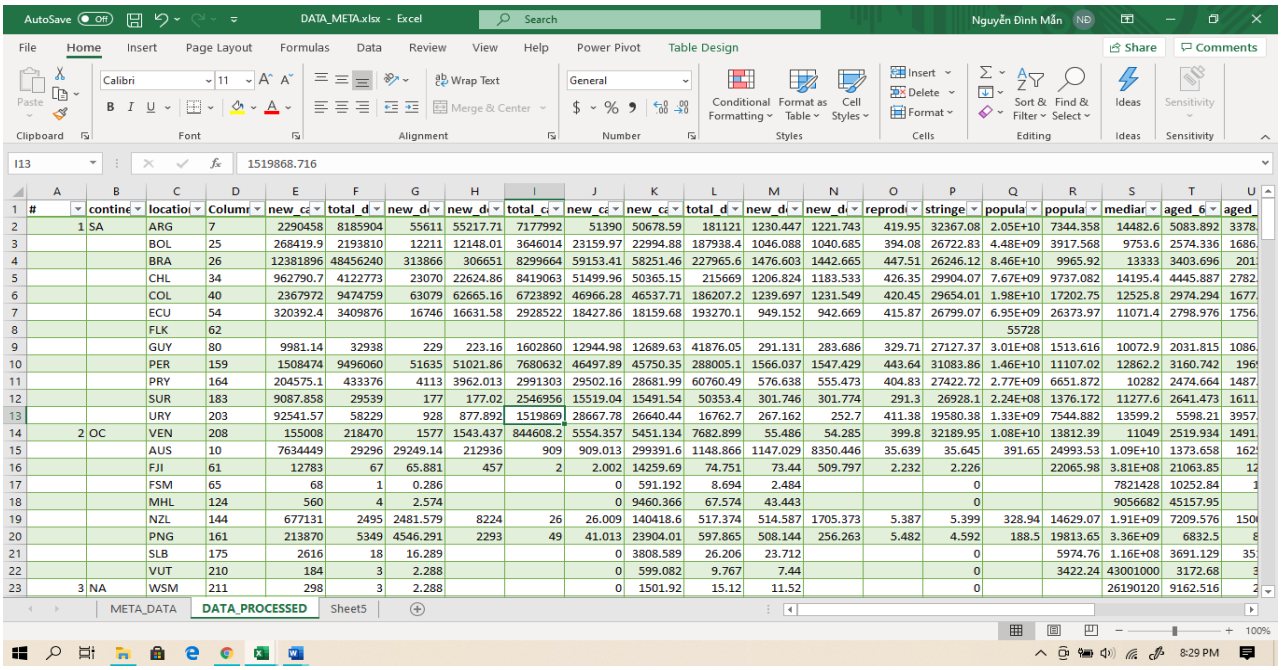


- Dữ liệu sau khi Import từ file *.csv



Bước 4: Thực hiện tổ chức lại dữ liệu

Dữ liệu sau khi được tổ chức lại:



#	continent	location	Column	new_cases	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_smoothed	new_cases_smoothed_per_million	total_deaths_per_million	new_deaths_per_million	new_deaths_smoothed_per_million	reproduction_rate	stringency_index	population	population_density	median_age	aged_65_older	aged
2	1	SA	ARG	7	2290458	8185904	55611	55217.71	7177992	51390	50678.59	181121	1230.447	1221.743	419.95	32367.08	2.05E+10	7344.358	14482.6	5083.892	3378
3		BOL	25	268419.9	2193810	12211	12148.01	3646014	23159.97	22994.88	187938.4	1046.088	1040.685	394.08	26722.83	4.48E+09	3917.568	9753.6	2574.336	1686
4		BRA	26	12381896	48456240	313866	306651	8299664	59153.41	58251.46	227965.6	1476.603	1442.665	447.51	26246.12	8.46E+10	9965.92	13333	3403.696	201
5		CHL	34	962790.7	4122773	23070	22624.86	8419063	51499.96	50365.15	215669	1206.824	1183.533	426.35	29904.07	7.67E+09	9737.082	14195.4	4445.887	2782
6		COL	40	2367972	9474759	63079	62665.16	6723892	46966.28	46537.71	186207.2	1239.697	1231.549	420.45	29654.01	1.98E+10	17202.75	12525.8	2974.294	1677
7		ECU	54	320392.4	3409876	16746	16631.58	2928522	18427.86	18159.68	193270.1	949.152	942.669	415.87	26799.07	6.95E+09	26373.97	11071.4	2798.976	1756
8		FLK	62													55728				
9		GUY	80	9981.14	32938	229	223.16	1602860	12944.98	12689.63	41876.05	291.131	283.686	329.71	27127.37	3.01E+08	1513.616	10072.9	2031.815	1086
10		PER	159	1508474	9496060	51635	51021.86	7680632	46497.89	45750.35	288005.1	1566.037	1547.429	443.64	31083.86	1.46E+10	11107.02	12862.2	3160.742	196
11		PRY	164	204575.1	433376	4113	3962.013	2991303	29502.16	28681.99	60760.49	576.638	555.473	404.83	27422.72	2.77E+09	6651.872	10282	2474.664	1487
12		SUR	183	9087.858	29539	177	177.02	2546956	15519.04	15491.54	50353.4	301.746	301.774	291.3	26928.1	2.24E+08	1376.172	11277.6	2641.473	1611
13		URY	203	92541.57	58229	928	877.892	1519869	28667.78	26640.44	16762.7	267.162	252.7	411.38	19580.38	1.33E+09	7544.882	13599.2	5598.21	3957
14	2	OC	VEN	208	155008	218470	1577	1543.437	844608.2	5554.357	5451.134	7682.899	55.486	54.285	399.8	32189.95	1.08E+10	13812.39	11049	2519.934	1491
15		AUS	10	7634449	29296	29249.14	212936	909	909.013	299391.6	1148.866	1147.029	8350.446	35.639	35.645	391.65	24993.53	1.09E+10	1373.658	162
16		FJI	61	12783	67	65.881	457	2	2.002	14259.69	74.751	73.44	509.797	2.232	2.226		22065.98	3.81E+08	21063.85	12
17		FSM	65	68	1	0.286			0	591.192	8.694	2.484			0			7821428	10252.84	1
18		MHL	124	560	4	2.574			0	9460.366	67.574	43.443			0			9056682	45157.95	
19		NZL	144	677131	2495	2481.579	8224	26	26.009	140418.6	517.374	514.587	1705.373	5.387	5.399	328.94	14629.07	1.91E+09	7209.576	150
20		PNG	161	213870	5349	4546.291	2293	49	41.013	23904.01	597.865	508.144	256.263	5.482	4.592	188.5	19813.65	3.36E+09	6832.5	8
21		SLB	175	2616	18	16.289			0	3808.589	26.206	23.712			0		5974.76	1.16E+08	3691.129	35
22		VUT	210	184	3	2.288			0	599.082	9.767	7.44			0		3422.24	43001000	3172.68	3
23	3	NA	WSM	211	298	3	2.288			0	1501.92	15.12	11.52			0			26190120	9162.516	2

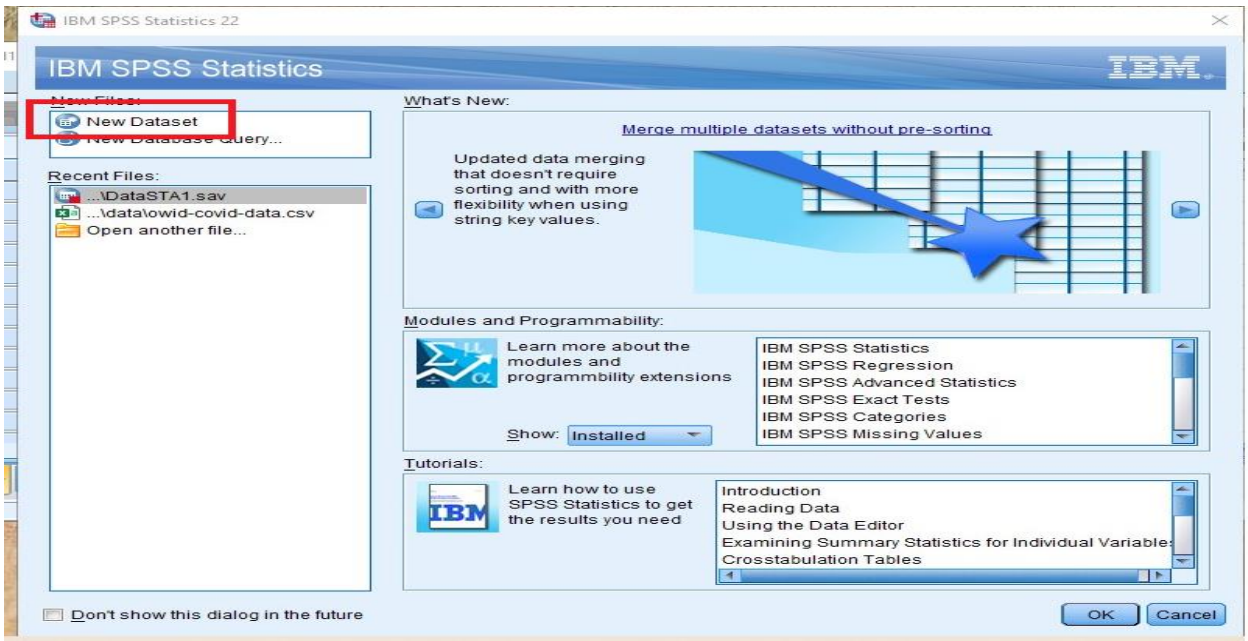
Bước 5: Lưu file với tên META_DATE.xlxx

II. PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU TRÊN PHẦN MỀM SPSS

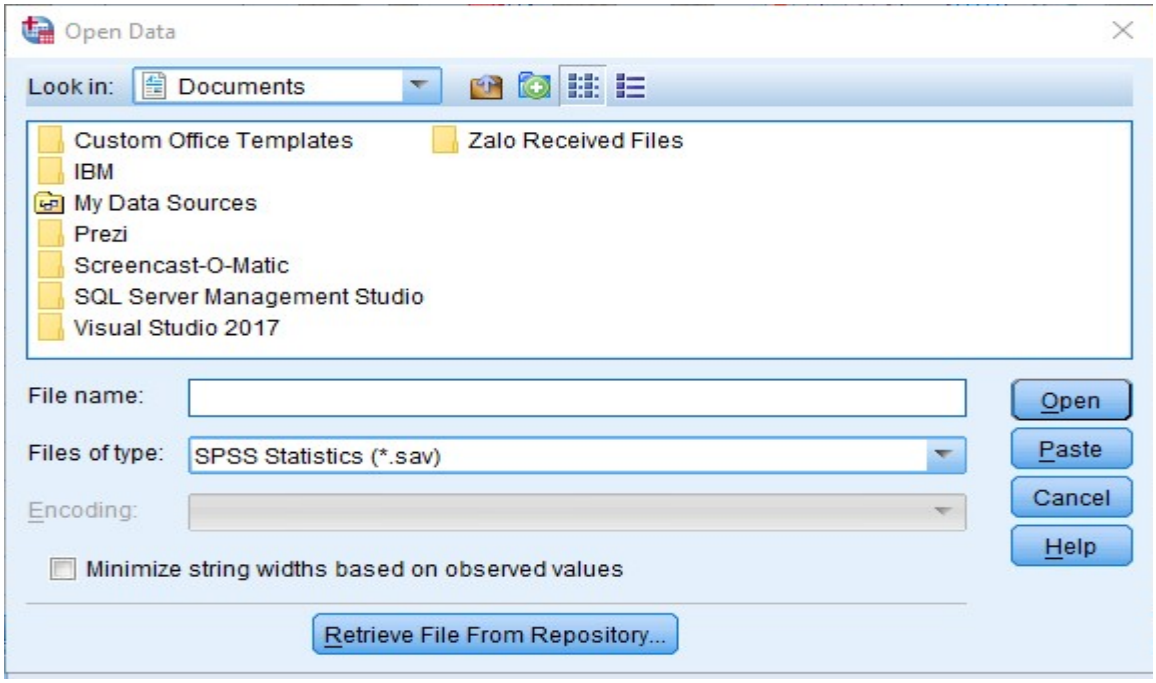
1. Cách import dữ liệu vào SPSS

Bước 1: Mở phần mềm IBM SPSS Statistic 22

Bước 2: Chọn New Dataset



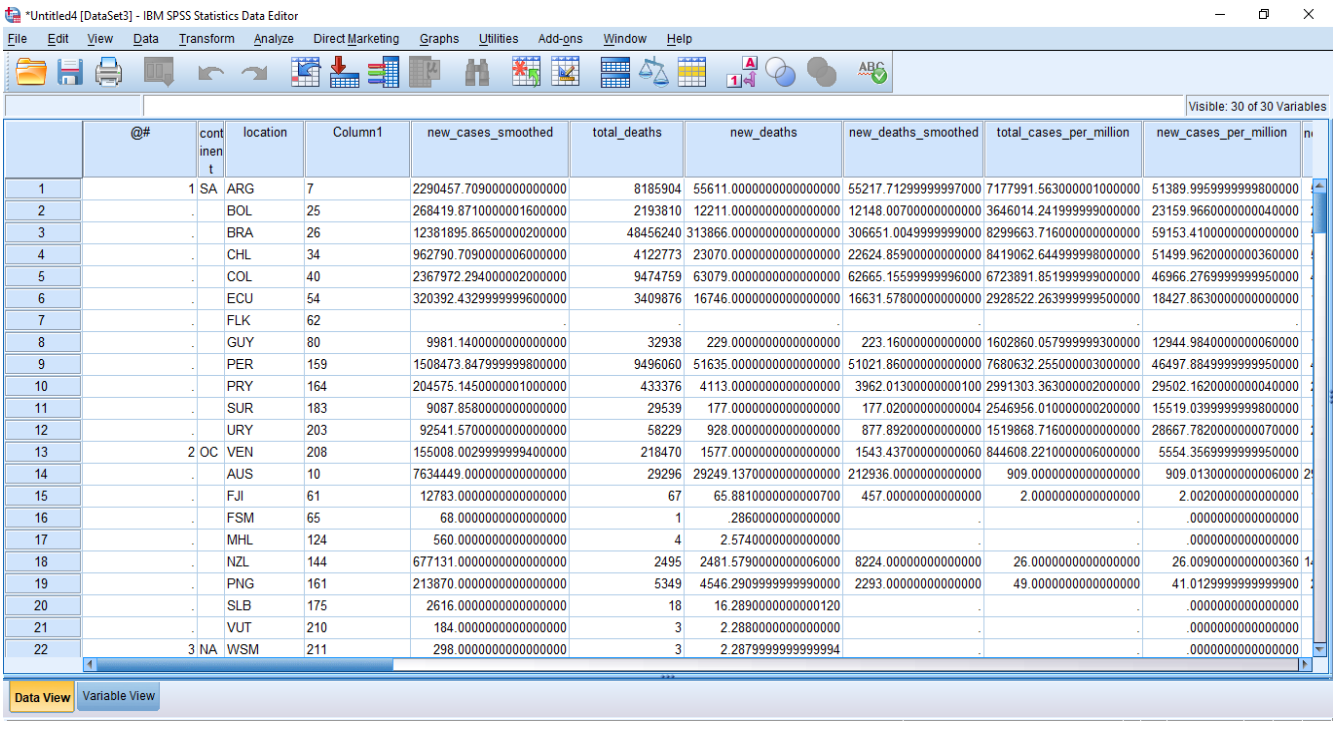
Bước 3: Chọn **File > Open > Data**, xuất hiện hộp thoại **Open Data**



Bước 4:

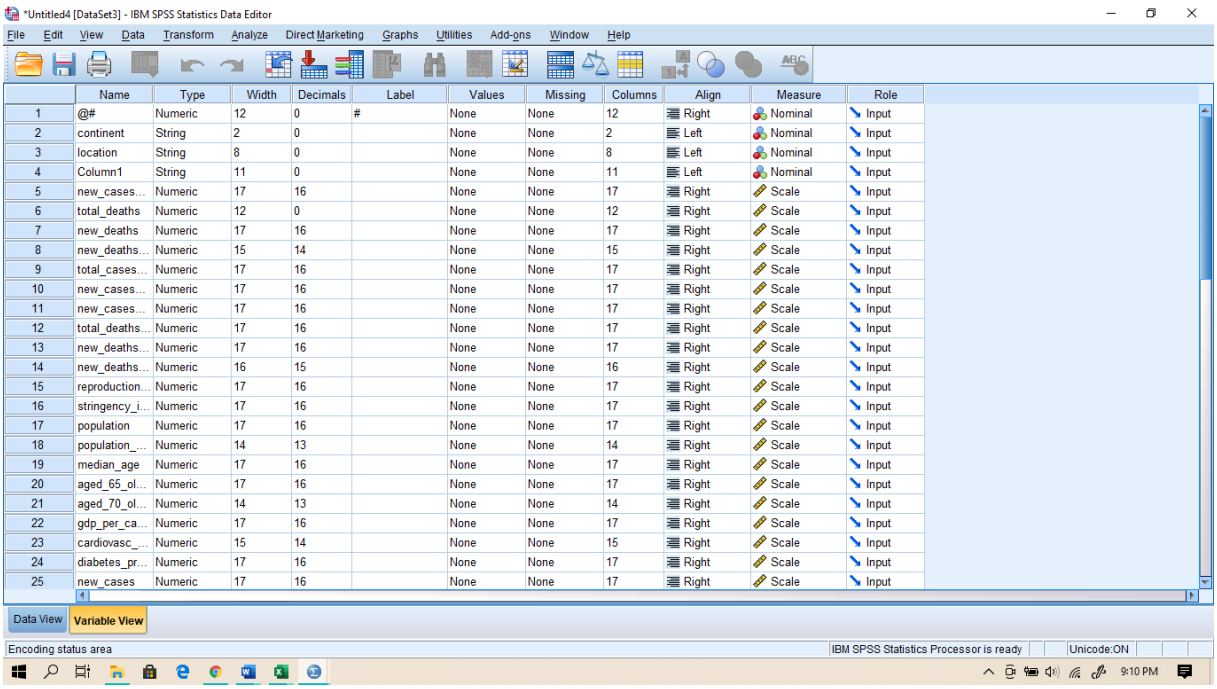
- Chọn thư mục lưu file dữ liệu
- Điều chỉnh **File of type** thành định dạng **Excel**
- Chọn file META_DATA.xlsx. Sau đó nhấn **Open**

- File dữ liệu sẽ được hiển thị như sau:



	@#	continent	location	Column1	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million
1	1	SA	ARG	7	2290457.709000000000000	8185904	55611.000000000000000	55217.712999999997000	7177991.563000001000000	51389.995999999980000
2	.		BOL	25	268419.871000001600000	2193810	12211.000000000000000	12148.007000000000000	3646014.241999999000000	23159.966000000040000
3	.		BRA	26	12381895.865000002000000	48456240	313866.000000000000000	306651.004999999900000	8299663.716000000000000	59153.410000000000000
4	.		CHL	34	962790.709000006000000	4122773	23070.000000000000000	22624.859000000000000	8419062.644999998000000	51499.962000000360000
5	.		COL	40	2367972.294000002000000	9474759	63079.000000000000000	62665.155999999996000	6723891.851999999000000	46966.276999999950000
6	.		ECU	54	320392.432999999600000	3409876	16746.000000000000000	16631.578000000000000	2928522.263999999500000	18427.863000000000000
7	.		FLK	62	.	.	.	.	.	.
8	.		GUY	80	9981.140000000000000	32938	229.000000000000000	223.160000000000000	1602860.057999999300000	12944.984000000060000
9	.		PER	159	1508473.847999999800000	9496060	51635.000000000000000	51021.860000000000000	7680632.255000003000000	46497.884999999950000
10	.		PRY	164	204575.145000001000000	433376	4113.000000000000000	3962.013000000000100	2991303.363000002000000	29502.162000000040000
11	.		SUR	183	9087.858000000000000	29539	177.000000000000000	177.020000000000004	2546956.010000000200000	15519.039999999800000
12	.		URY	203	92541.570000000000000	58229	928.000000000000000	877.892000000000000	1519868.716000000000000	28667.782000000070000
13	2	OC	VEN	208	155008.002999999900000	218470	1577.000000000000000	1543.437000000000060	844608.221000000600000	5554.356999999950000
14	.		AUS	10	7634449.000000000000000	29296	29249.137000000000000	212936.000000000000000	909.000000000000000	909.013000000006000
15	.		FJI	61	12783.000000000000000	67	65.881000000000700	457.000000000000000	2.000000000000000	2.002000000000000
16	.		FSM	65	68.000000000000000	1	.286000000000000	.	.	.000000000000000
17	.		MHL	124	560.000000000000000	4	2.574000000000000	.	.	.000000000000000
18	.		NZL	144	677131.000000000000000	2495	2481.579000000000600	8224.000000000000000	26.000000000000000	26.009000000000360
19	.		PNG	161	213870.000000000000000	5349	4546.290999999999900	2293.000000000000000	49.000000000000000	41.012999999999900
20	.		SLB	175	2616.000000000000000	18	16.289000000000120	.	.	.000000000000000
21	.		VUT	210	184.000000000000000	3	2.288000000000000	.	.	.000000000000000
22	3	NA	WSM	211	298.000000000000000	3	2.287999999999994	.	.	.000000000000000

- 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

2. 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,…

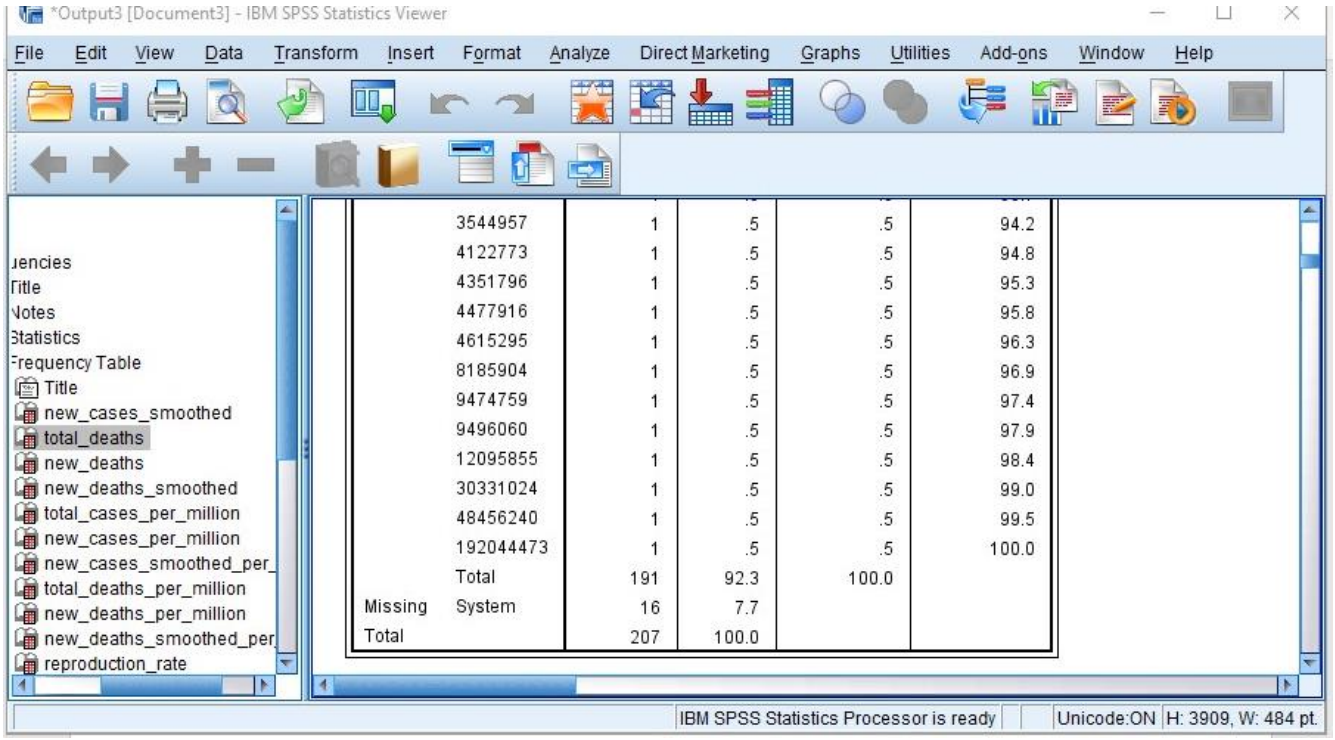
Do vậy trước khi dữ liệu được phân tích, cần làm sạch để hạn chết sai số trong quá trình phân tích số liệu.

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:

Statistics

continent

N	Valid	207
	Missing	0

continent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		201	97.1	97.1	97.1
	AF	1	.5	.5	97.6
	AS	1	.5	.5	98.1
	EU	1	.5	.5	98.6
	NA	1	.5	.5	99.0
	OC	1	.5	.5	99.5
	SA	1	.5	.5	100.0
	Total	207	100.0	100.0	

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.

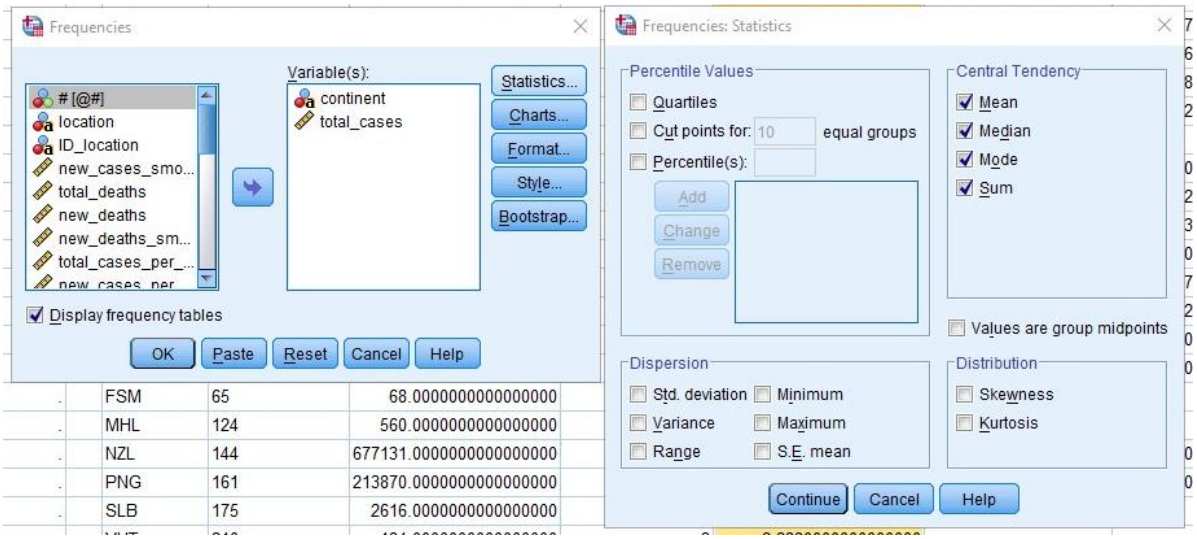
3. Mô tả dữ liệu và phân tích kết quả

3.1 Mô tả dữ liệu bằng Bảng trong SPSS

a. 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:

Frequencies

Statistics

		continent	total_cases
N	Valid	207	191
	Missing	0	16
Mean			31340772.16
Median			2801.340000
Mode			895.400000[a]
Sum			5986087482

a. Multiple modes exist. The smallest value is shown

Frequency Table

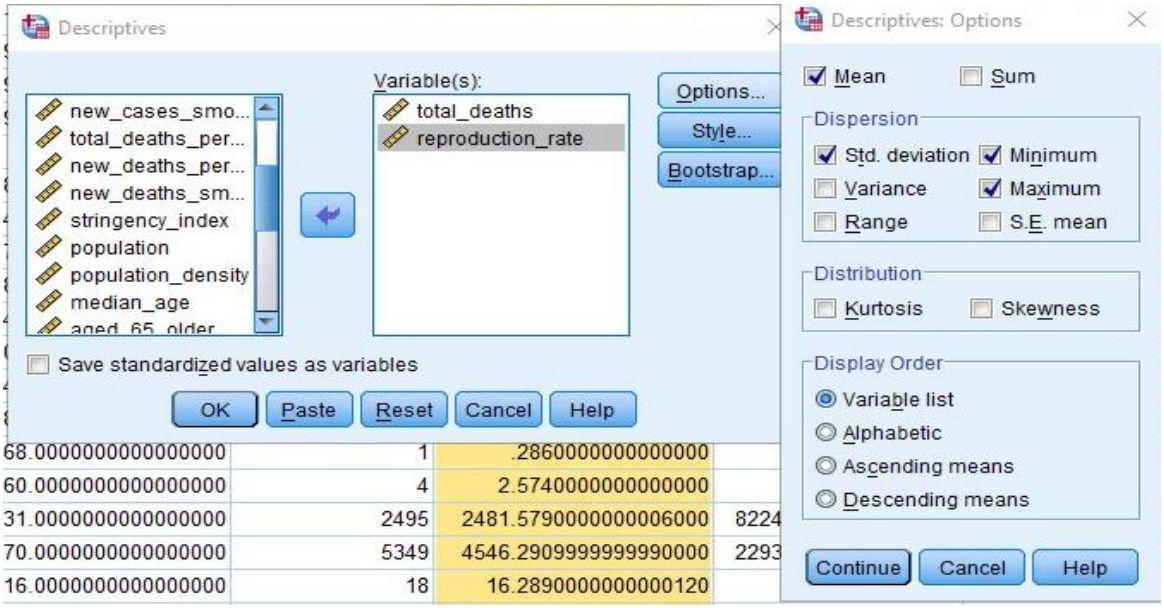
continent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		201	97.1	97.1	97.1
	AF	1	.5	.5	97.6
	AS	1	.5	.5	98.1
	EU	1	.5	.5	98.6
	NA	1	.5	.5	99.0
	OC	1	.5	.5	99.5
	SA	1	.5	.5	100.0
	Total	207	100.0	100.0	

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

Descriptive Statistics

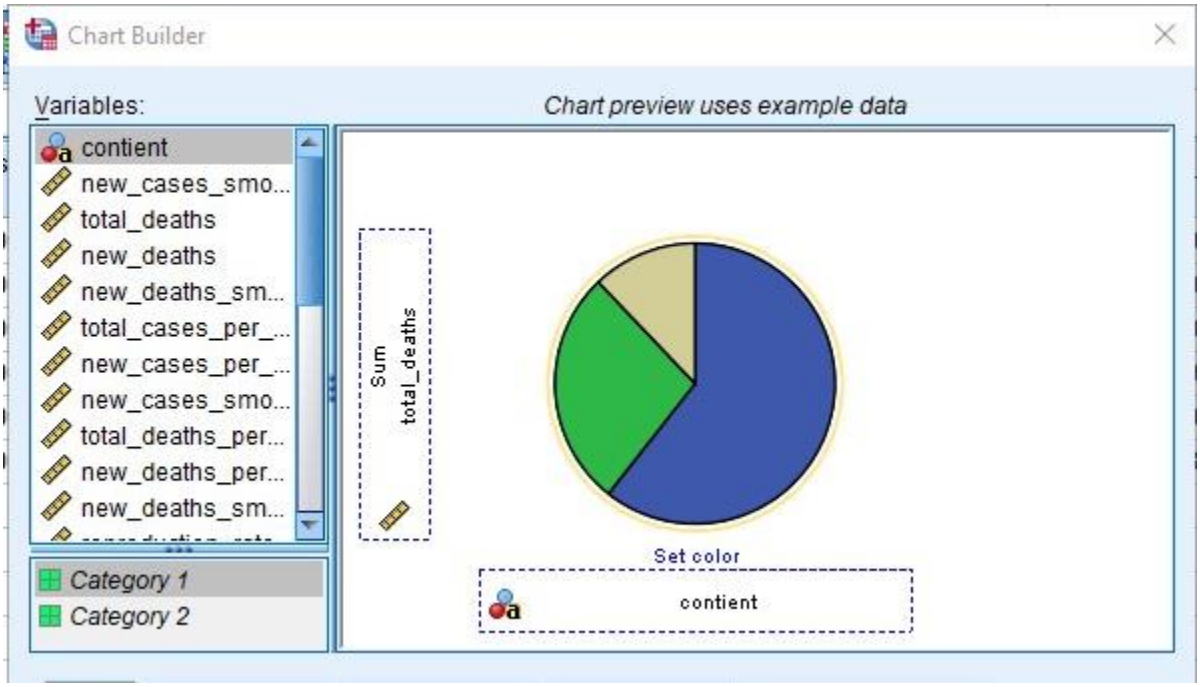
	N	Minimum	Maximum	Mean	Std. Deviation
new_cases_smoothed	191	68.00000000	1.41337E+10	147996406.6	1072579454
total_deaths	191	1	192044473	2010936.89	14498178.28
new_deaths	191	.2860000000	105176151.9	1101320.963	7950851.273
new_deaths_smoothed	181	71.00000000	345373335.7	3816279.952	26686499.89
total_cases_per_million	181	1.000000000	56629851.90	625744.2199	4393747.446
new_cases_per_million	191	.0000000000	2620323.245	27437.93974	194391.4896
new_cases_smoothed_per_million	191	591.1920000	561585494.4	5880476.381	40603955.80
total_deaths_per_million	191	6.728000000	6331683.960	66300.35560	458226.4863
new_deaths_per_million	191	2.484000000	4614989.362	48324.49594	333698.6269
new_deaths_smoothed_per_million	181	4.247000000	10580257.84	116908.9264	786895.1781
reproduction_rate	181	.3510000000	84719.72400	936.1295470	6290.175764
stringency_index	191	.0000000000	414946.9750	4344.994503	30626.34567
Valid N (listwise)	181				

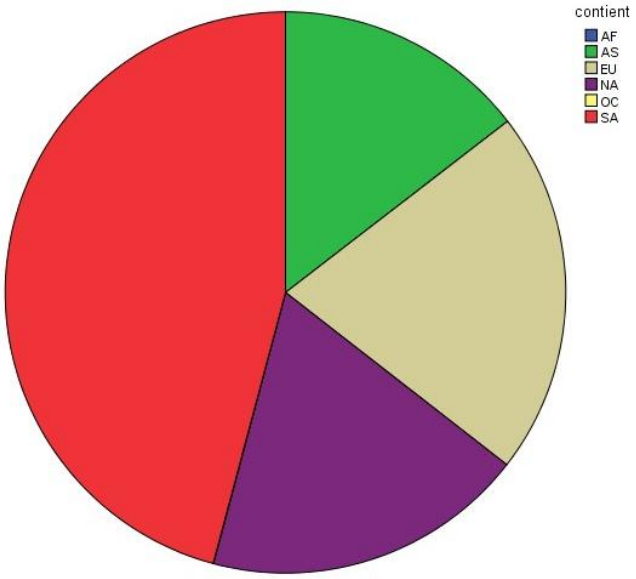
Bảng thể hiện các giá trị trung bình của các biến được chọn.

3.2 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

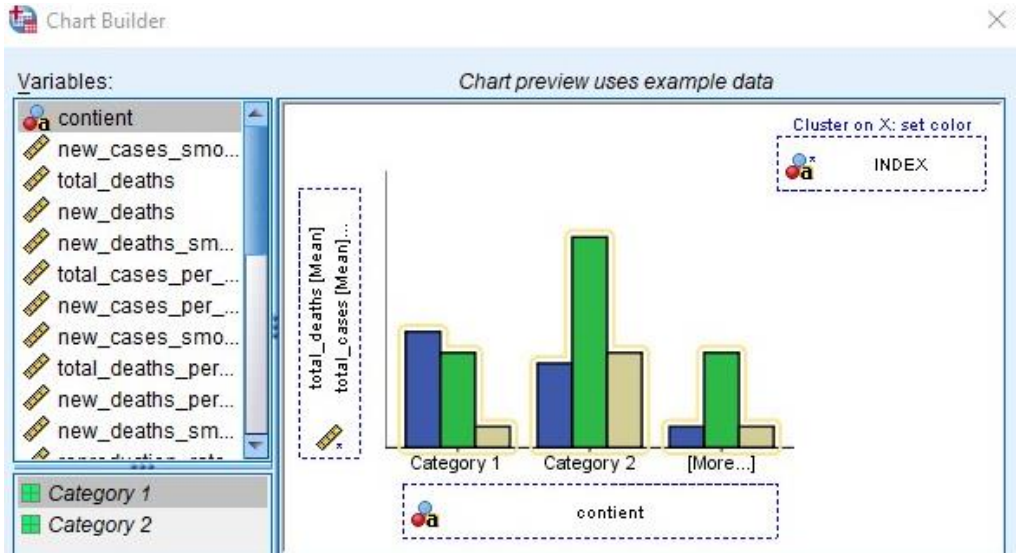


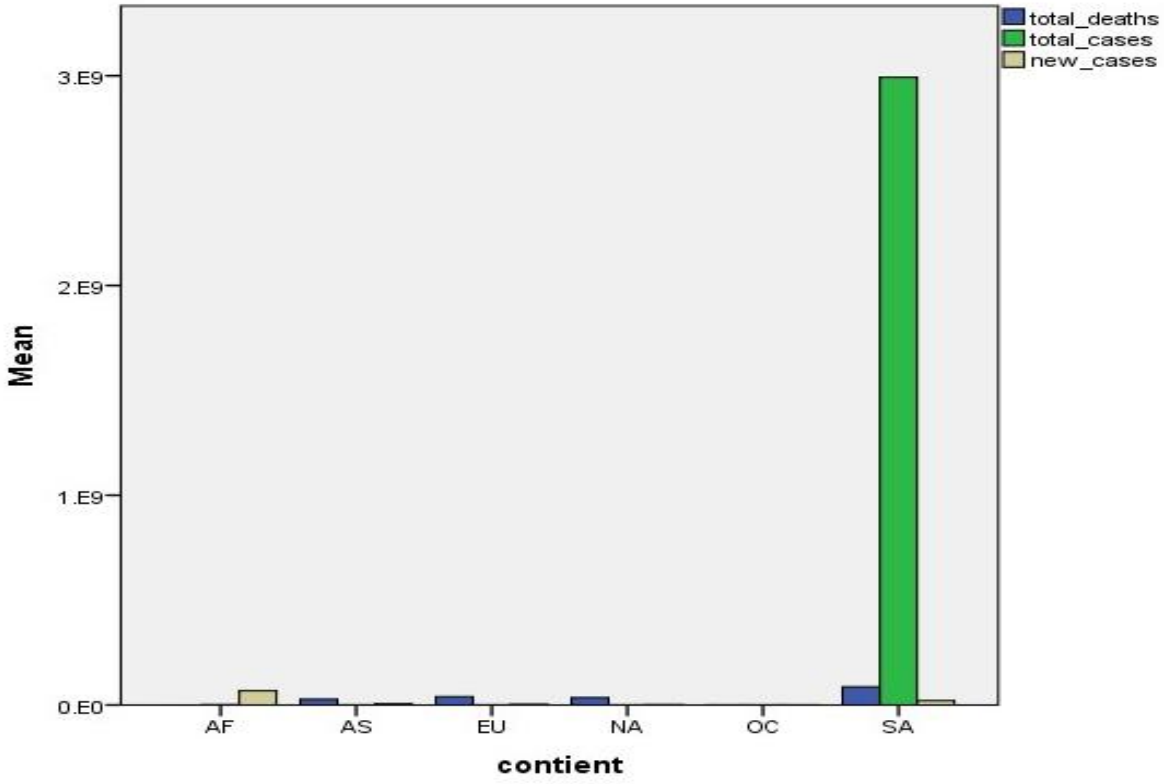


Nhận xét:

- 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

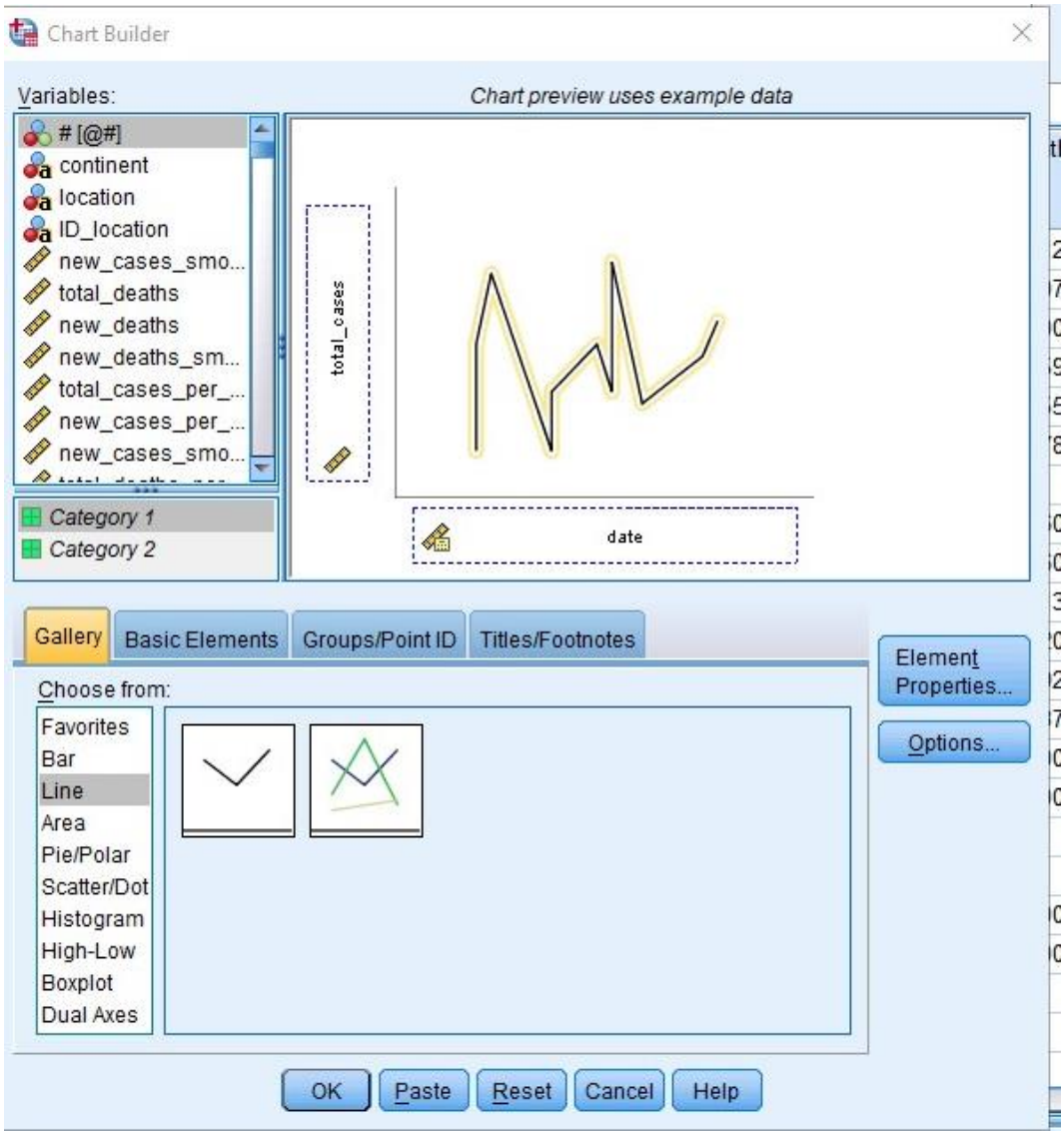


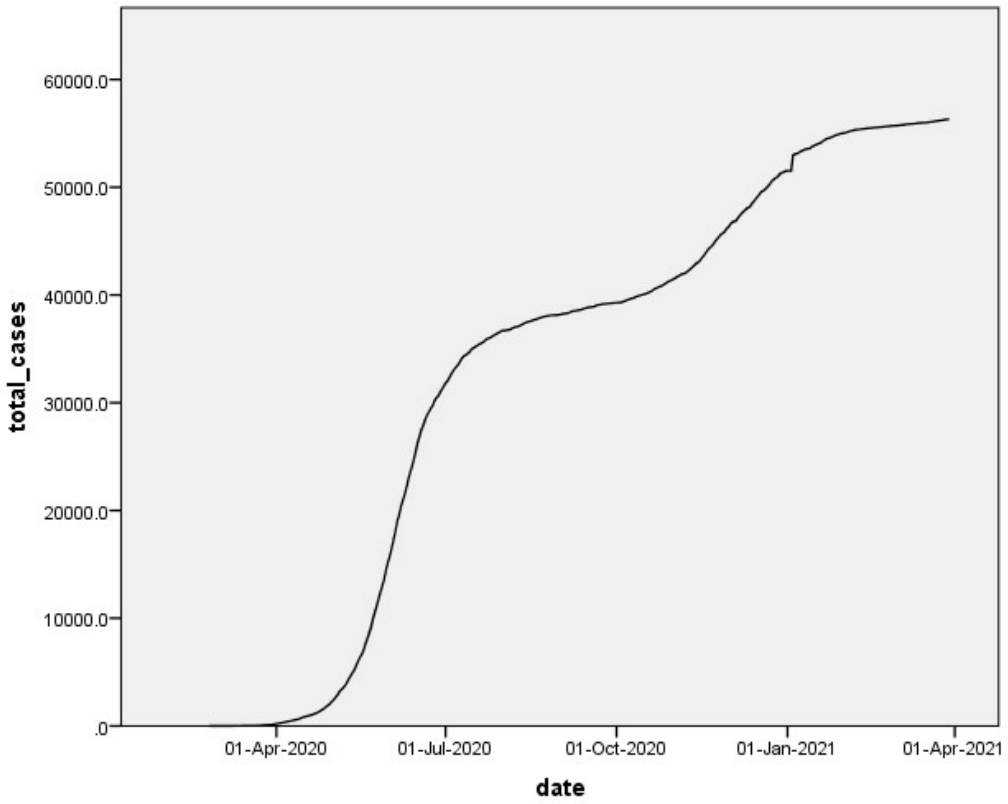


Nhận xét:

- 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.

3.3 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

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
total_cases	6	29695.60000	2992518115	499090694.7	1221525129
Valid N (listwise)	6				

Phân tích:

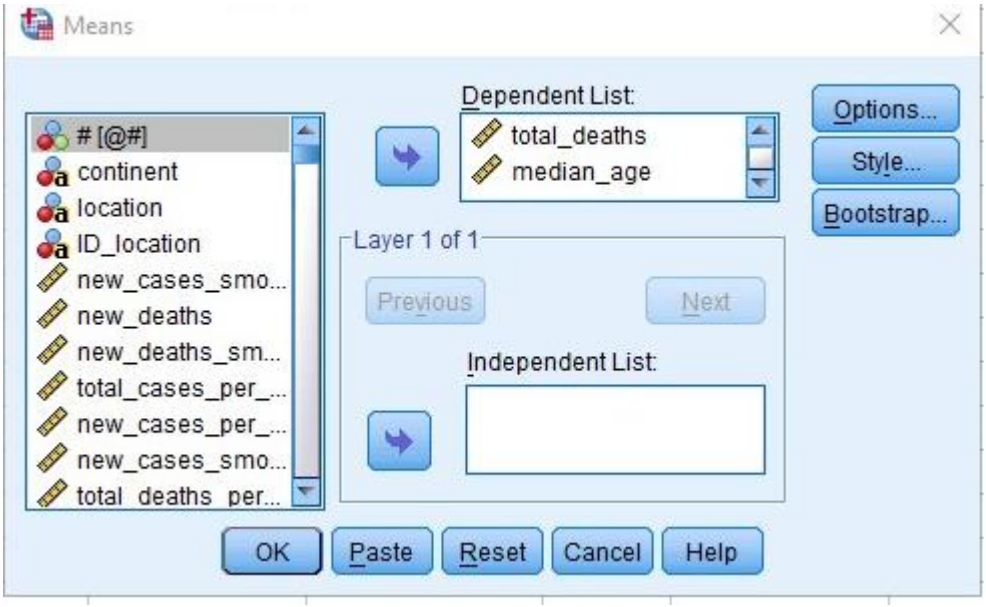
- Tổng số ca mắc ít nhất trong 6 khu vực trong khoảng thời gian thu thập số liệu là 29695 ca.
- Số ca mắc nhiều nhất được ghi nhận trên 6 khu vực trong khoảng thời gian thu thập là 499090694 ca.
- Độ lệch chuẩn của dữ liệu sau khi xử lý là 1221525129 ca.

4. Kiểm định

a. 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:

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
total_deaths	191	92.3%	16	7.7%	207	100.0%
median_age	205	99.0%	2	1.0%	207	100.0%

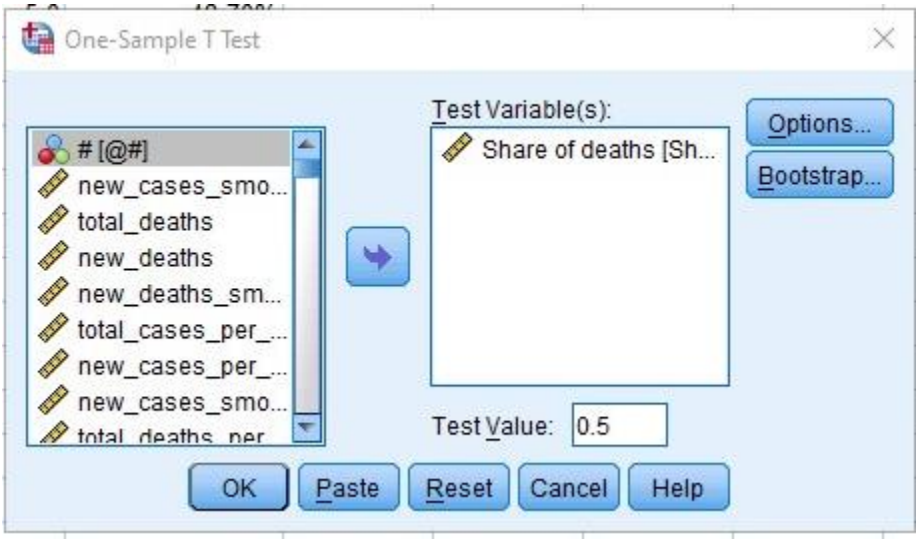
Report

	total_deaths	median_age
Mean	2010936.89	2.95920E+10
N	191	205
Std. Deviation	14498178.28	2.19605E+11

b. 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

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Share of deaths	6	33.3267%	36.99912%	15.10483%

One-Sample Test

	Test Value = 0.5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Share of deaths	2.173	5	.082	32.82667%	-6.0015%	71.6549%

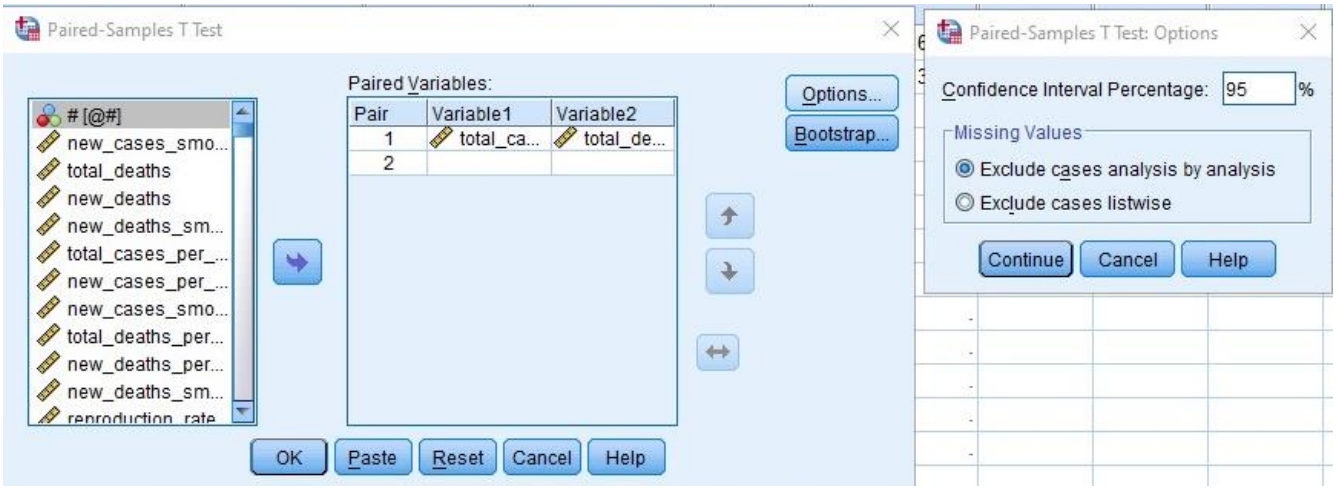
Phân tích: Quan sát trong bảng One-Sample Test, ta thấy giá trị Sig.(2-tailed) bằng 0.082, lớn hơn 0.05 . Vậy ta ***chưa đủ cơ sở để bác bỏ giả thuyết*** ban đầu, Khả năng mắc bệnh Covid ở các độ tuổi là 50%

c. 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

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	total_cases	2636.035909	44	784.9102138	118.3296670
	total_deaths	893160.727	44	1306648.799	196984.7182

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	total_cases & total_deaths	44	-.103	.507

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	total_cases - total_deaths	-890524.691	1306729.600	196996.8993	-1287806.80	-493242.581	-4.521	43	.000

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**