Multivariate Stats (PC & Discrimination) in the Analysis of Covid-19

Jorge Luis Romeu, Ph.D.
https://www.researchgate.net/profile/Jorge_Romeu
http://web.cortland.edu/romeu/
Email: romeu@cortland.edu
Copyright. May 15, 2020

Introduction

The present analysis uses metrics and data from New York State (NYS) Regions, released on V/13/20 by NYS State authorities, to illustrate the power of multivariate statistical analysis in the fight against the Coronavirus Pandemic. A Principal Components analysis will be implemented to explore the impact of the variables used. A Discrimination Analysis will then be used to assess which one of the variables utilized, does a better job in separating Regions into those that can safely open their economy, and those that should still wait.

The variables used are heterogeneous and have different units. In addition, *statistics* that use say Maximum values of a given period, are not as useful as *statistics* that use better values (e.g. units respect to total population), because they do not convey all available information (e.g. periods with different numbers of daily hospitalization or deaths, may have the same maximum value). Technically, such statistics are referred to as *not being sufficient statistics* as they do not provide all the information in the sample.

Other variables such as *availability of beds* in a hospital, or in an ICUs, are not proportional to a region's population. Variables *number of available beds in ICUs* and *number of available beds in the hospitals*, are not proportional to their respective regional populations, either.

Most interesting variables are, *new hospitalizations per 1000 population*, and the *ratio of existing versus required (according to CDC) diagnostic testing capacity*. This is because these variables reflect a relative impact with respect to their regional population.

We present, in Table 1, NYS data from https://forward.ny.gov/regional-monitoring-dashboard .

Table 1: Description of Metrics:

County	NewHosp	Deaths	Hosp/Res	ShareBeds	ShareICU	%Tested	Status	Number
Capital	18	6	1.32	0.33	0.51	1.105069	Yes	1
CNY	6	3	1.07	0.43	0.53	0.914839	Yes	1
FingerL	11	3	0.94	0.42	0.53	1.354115	Yes	1
LongIs	425	99	2.66	0.3	0.33	1.599155	No	-1
MidHud	132	69	2.25	0.33	0.49	1.725668	No	-1
Mohawk	4	2	0.62	0.51	0.64	1.101031	Yes	1
NorthC	3	1	0.16	0.52	0.68	1.083532	Yes	1
Southern	5	2	0.21	0.44	0.42	1.510269	Yes	1
Western	28	9	1.95	0.45	0.55	1.044895	Yes	1
NYC	820	502	2.64	0.28	0.24	1.610549	No	-1

We present, in Table 2, their data definitions according to the same State source.

Table 2: Description of the Metrics: Established based on guidance from Center for Disease Control and Prevention, World Health Organization, U.S. Department of State, and other public health experts.

Metric #1—Decline in Total Hospitalizations. Region must show a sustained decline in the three-day rolling average of total net hospitalizations (defined as total number of people in hospital on a given day) over the course of a 14-day period. Alternatively, regions satisfy this metric if daily net increase in total hospitalizations (measured on a 3-day rolling average basis) has never exceeded 15. The **first number** represents the number of **consecutive days of decline in** the three-day rolling average of **total net hospitalizations**; if this number is 14 or greater the region automatically satisfies this metric. The **second number** represents the **maximum daily net increase in total hospitalizations** measured on a three day rolling average the region has experienced; if this number is 15 or less region satisfies metric.

Metric #2—Decline in Deaths. Region must show a sustained decline in the three-day rolling average of daily hospital deaths over the course of a 14-day period. Alternatively, regions can satisfy this metric if the three-day rolling average of daily new hospital deaths has never exceeded 5. The **first number** in this cell represents number **of consecutive days of decline** in three-day rolling average **of daily hospital deaths**; if this number is 14 or greater the region automatically satisfies this metric. The **second number** represents **maximum daily increase** in three-day rolling average **of new hospital deaths** that region has experienced; if this number is 5 or less the region automatically satisfies this metric.

Metric #3—New Hospitalizations. Region must experience fewer than 2 **new hospitalizations per 100,000 residents**, measured on a three-day rolling average. New hospitalizations include both new admissions and prior admissions subsequently confirmed as positive COVID cases.

Metric #4—Hospital Bed Capacity. Regions must have at least 30% of their **hospital beds available**.

Metric #5—ICU Bed Capacity. Regions must have at least 30% of their **ICU beds available**

Metric #6—Diagnostic Testing Capacity. Average daily diagnostic testing over the past 7 days must be sufficient to **conduct 30 tests per 1,000 residents** per month. This entry reflects the ratio of Test per 1000 residents, over the required tests per 1000 residents, to fulfill said metric.

Principal Component Analysis Assess Relationship between Variables & Regions

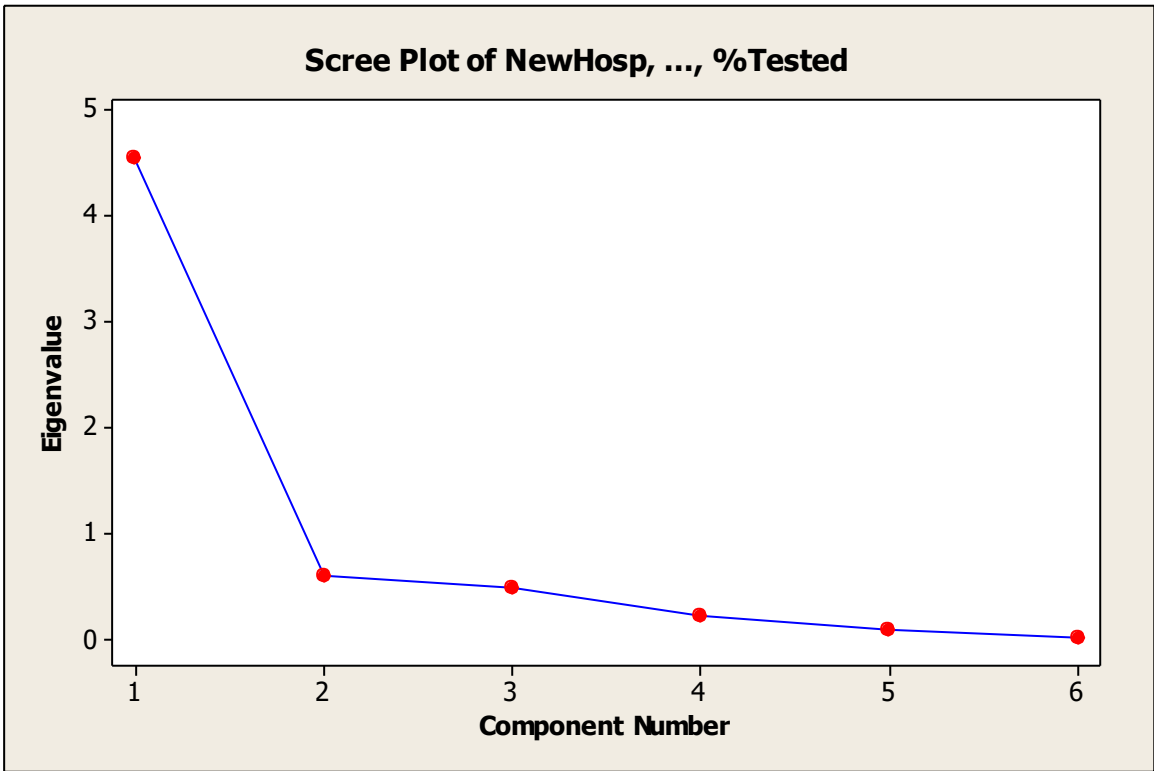
Variablese Analyzed: NewHosp, Deaths, Hosp/Res, ShareBeds, ShareICU, %Tests

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.5495	0.6110	0.4986	0.2205	0.0946	0.0259
Proportion	0.758	0.102	0.083	0.037	0.016	0.004
Cumulative	0.758	0.860	0.943	0.980	0.996	1.000

Variable	PC1
NewHosp	0.437
Deaths	0.404
Hosp/Res	0.394
ShareBeds	-0.420
ShareICU	-0.436
%Tested	0.353

Scree Plot of NewHosp, ..., %Tested



The Analysis for One Principal Components and all six variables shows how 75% of the Total explanation is given by the First Component, 10% by the second, 8% by the third, and the fourth and following components' contribution is minimal, corroborated by the above Scree Plot.

The First Component singles out, by being negative, variables Share of Hospital and Share of ICUs Beds, which is consistent with Covid-19 infection results. As number of hospitalizations and deaths in ICU Beds increase, because the infection rate increases, the number of available hospital beds and ICUs then decreases.

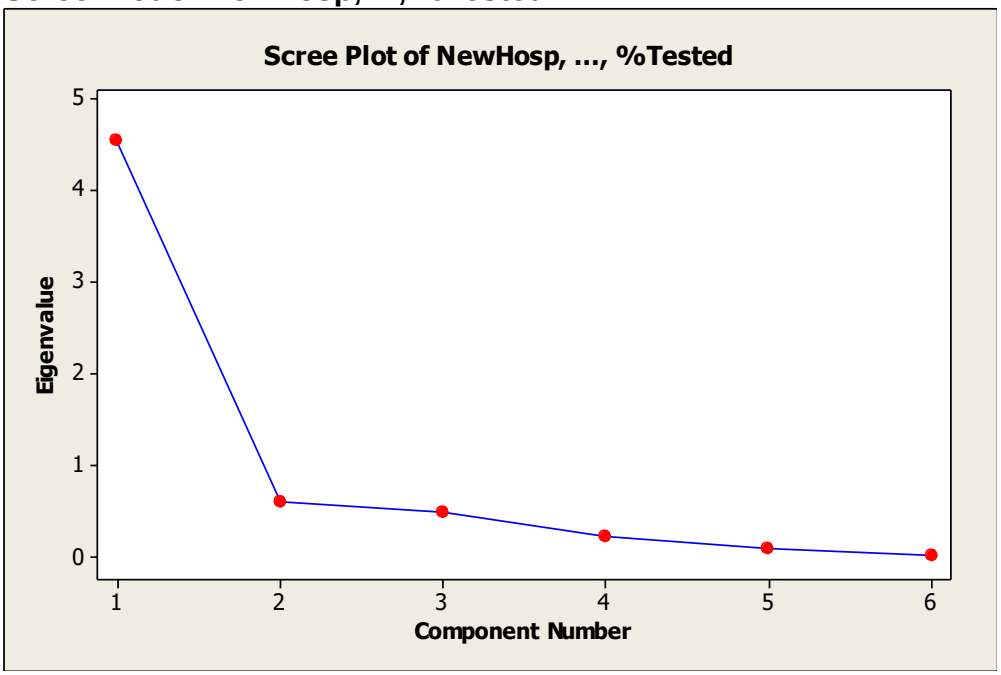
Principal Component Analysis: now considering Two Components

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.5495	0.6110	0.4986	0.2205	0.0946	0.0259
Proportion	0.758	0.102	0.083	0.037	0.016	0.004
Cumulative	0.758	0.860	0.943	0.980	0.996	1.000

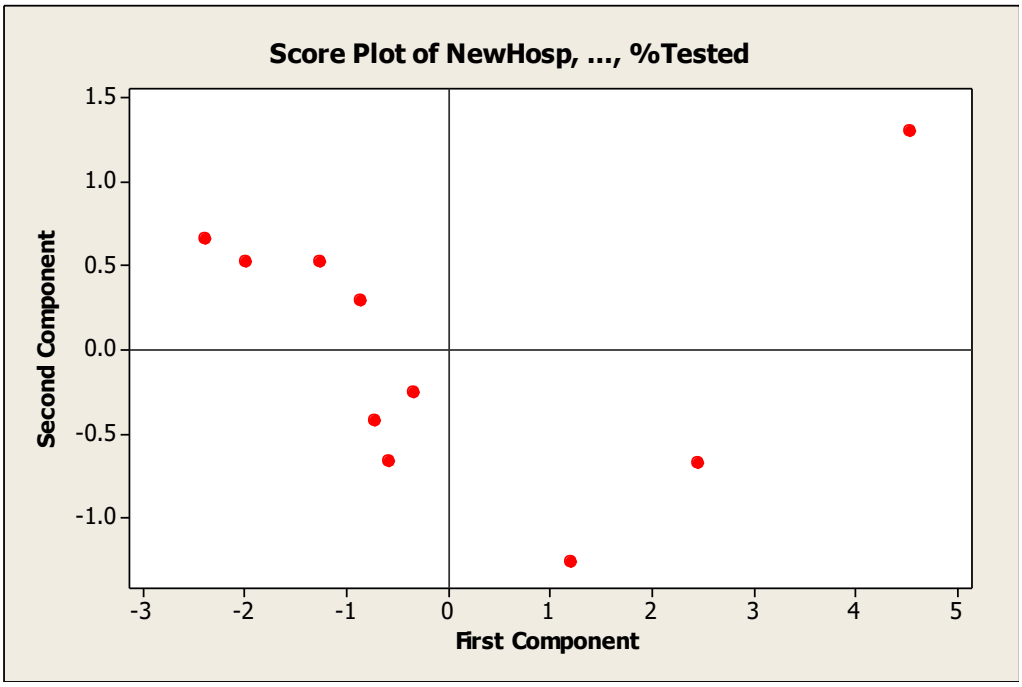
Variable	PC1	PC2
NewHosp	0.437	0.414
Deaths	0.404	0.586
Hosp/Res	0.394	-0.105
ShareBeds	-0.420	0.317
ShareICU	-0.436	0.065
%Tested	0.353	-0.608

Scree Plot of NewHosp, ..., %Tested



For Two Principal Components and all six variables: the analysis confirms how 75% of Total explanation is given by the First Component, 10% by the second, 8% by the third and the fourth and following components' contribution is minimal. However, this variant adds the possibility of plotting and grouping, using the two componets' coefficients, the NYS Regions according to their Covid-19 infection characteristics.

Score Plot of NewHosp, ..., %Tested



Principal Component Loadings (coefficients):

County	PC1	PC2
Capital	-0.35116	-0.2595
CNY	-1.27015	0.51828
FingerL	-0.73114	-0.4210
LongIs	2.44612	-0.6760
MidHud	1.20713	-1.2678
Mohawk	-1.98939	0.5191
NorthC	-2.38718	0.65836
Southern	-0.59404	-0.6611
Western	-0.85686	0.28644
NYC	4.52666	1.30348

Notice how the Regional Groupings in the Score Plot also follow reality: NYC is a separate unit (in red, upper right quadrant), as it has had the largest number of Covid-19 cases and fatalities. Long Island and Mid Hudson are then grouped together (in blue, lower right quadrant) as they are next in cases, due to their proximity to NYC. Capital, Finger Lakes and Southern regions are grouped (in green, lower left quadrant) as medium level. Finally, Central NY, Mohawk, North Country and Western NY are also grouped together as they are similar (in black, upper right quadrant). Such regional groupings reflect current NYS metrics reality in Covid-19 infections.

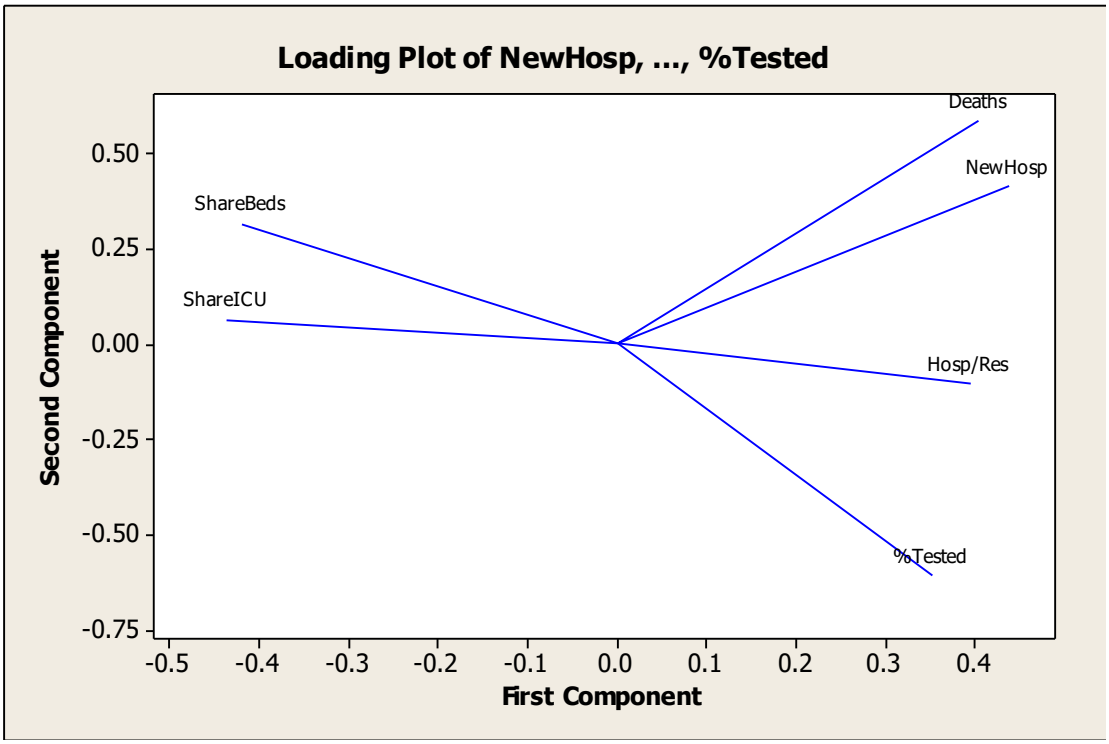
Using the First and Second Component coefficients (below), the six analysis variables are then plotted in the Loadings Plot, and then clustered:

Varables	CoefPC1	CoefPC2
NewHosp	0.436799	0.413887
Deaths	0.403741	0.585769
Hosp/Res	0.394209	-0.10521
ShareBeds	-0.41962	0.316652
ShareICU	-0.43615	0.064914
%Tested	0.352827	-0.60829

There are two clear Variable groups (Positive and Negative for First Component); and the first of the two groups, with two subgroups (Positive and Negative Second Component).

The First Component singles out as negative variables: Share of Hospital and of ICU Beds. They are consistent with strict Logic: as both, number of hospitalizations and deaths increase, because infection rate increases, then the number of Hospital and of ICU beds decreases.

Loading Plot of NewHosp, ..., %Tested



Discriminant Analysis:

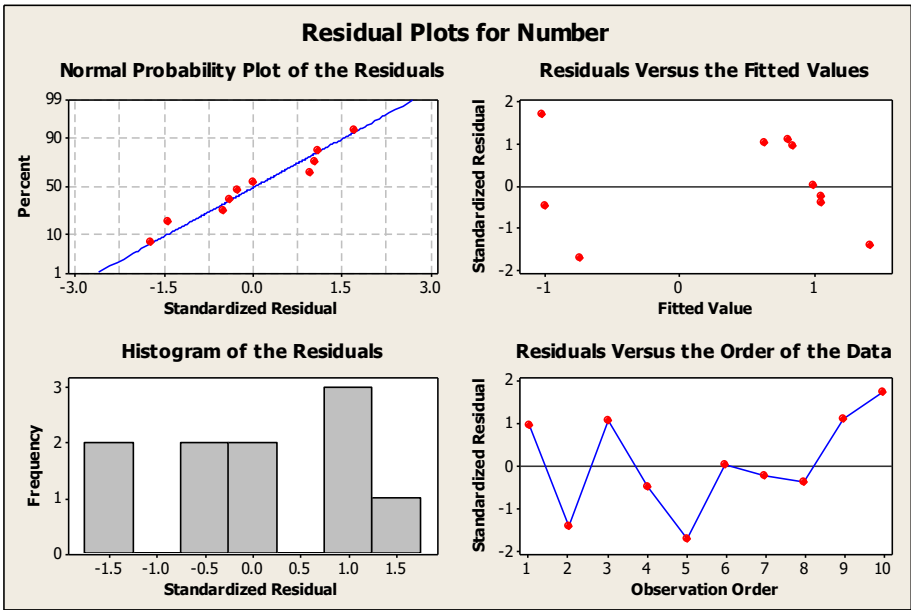
We have divided the NYS Regions into two groups, *infected* (i.e. not prone to open its economy) and with *manageable infection levels* (possibly, to open its economy), *based upon variable New Hospitalizations peer 1000 residents* (higher/lower than 2.0) of said Data. This division is totally arbitrary, done only to illustrate the value of the Discrimination method into Covid-19 analysis.

We used Regression approach to Discrimination Analysis, and regressed the Region variables vs. -1 and +1, according to which Regional group they had been assigned to. We considered all six variables, even when we expected such Discrimination Function, given so few regions, to fail.

```
The regression equation is:

Number = 4.53 - 0.00319 NewHosp + 0.00237 Deaths - 0.335 Hosp/Res
         + 1.97 ShareBeds - 3.63 ShareICU - 1.82 %Tested

Predictor       Coef   SE Coef      T      P
Constant       4.529     2.026   2.24  0.111
NewHosp    -0.003191  0.002319  -1.38  0.262
Deaths      0.002367  0.003107   0.76  0.502
Hosp/Res     -0.3349    0.2932  -1.14  0.336
ShareBeds      1.967     3.802   0.52  0.641
ShareICU      -3.628     2.578  -1.41  0.254
%Tested      -1.8173    0.6459  -2.81  0.067

S = 0.387895   R-Sq = 94.6%   R-Sq(adj) = 83.9%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	7.9486	1.3248	8.80	0.051
Residual Error	3	0.4514	0.1505		
Total	9	8.4000			



After analyzing several variable combinations, we found that only *Hospitalizations/residents* and *%testing* yielded a statistically significant Discrimination Function, able to separate well the two different Regional groups we had created:

The regression equation is: Number = 3.47 - 0.570 Hosp/Res - 1.75 %Tested

Predictor	Coef	SE Coef	T	P
Constant	3.4688	0.5813	5.97	0.001
Hosp/Res	**-0.5697**	**0.1523**	**-3.74**	**0.007**
%Tested	**-1.7484**	**0.4978**	**-3.51**	**0.010**

S = 0.368946 R-Sq = 88.7% R-Sq(adj) = 85.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	7.4471	3.7236	27.35	0.000
Residual Error	7	0.9529	0.1361		
Total	9	8.4000			

This Discrimination Function explains 85% of the problem and is able to correctly classify all the Regions in their respective group. The **Mahalanobis Distance** separates these two Region Groups already defined, and can be obtained in the following way:

For: n1=3; n2 = 7; and Lambda ^{2} = n1*n2/(n1+n2) = 21 / 10 = 2.1
Dp^2 = [(n1+n2-2)/Lambda ^{2}]*[R^2 / (1- R^2)] = ((10-2)/2.1)*(0.85 / (1-0.85)) = 21.58
Dp = Sqrt (21.58) = 4.64 => Prob (- ½ Dp) = Prob (- 4.64/2) ~ 0.0102

For completeness, we implemented the Minitab SW Discrimination method to this situation:

Predictors: %Tested, Hosp/Res

Group	No	Yes
Count	3	7

Linear Discriminant Function for Groups

	No	Yes
Constant	-92.428	-34.944
%Tested	81.404	52.046
Hosp/Res	20.240	10.674

Summary of classification

	True Group	
Put into Group	No	Yes
No	3	0
Yes	0	7
Total N	3	7
N correct	3	7
Proportion	1.000	1.000

N = 10 N Correct = 10 Proportion Correct = 1.000

Squared Distance Between Groups

	No	Yes
No	0.0000	29.7739
Yes	29.7739	0.0000

The Minitab SW Package Discrimination Function, using the same two variables defined, (Hospitalizations/residents and %testing) also separates well into two groups (according to Covid-19 impact). This corroborates how variables Hospitalizations/residents and %testing are the two best variables to use, with the present data.

Conclusions:

We have developed two powerful Multivariate Statistics methods: *Principal Components* and *Discriminant Analysis*, to examine and assess NYS Regional Covid-19 metrics, The data has been obtained from the NYS Web Site. *This is a preliminary analysis* and is subject to further validation and verification from subsequent and additional NYS data**.** *Its main objective is to demonstrate the use of Multivariate Analysis* in the struggle against Covid-19. *As more NYS Covid-19 Metrics regional data is collected*, these Principal Components and Discrimination Function *analyses can be validated and updated*.

Bibliography

Anderson, T.W. An Introduction to Multivariate Statistical Analysis. John Wiley & Sons. New York. Second Edition. 1971.

Romeu, J. L. *A Comparative Study of Goodness-of-Fit Tests for Multivariate Normality*. Journal of Multivariate Analysis. V. 46, No. 2. August 1993. 309--334.

Coronavarus Pandemic: *Leaders*, Going Global; *Briefing*, Flattening the Curve; *Graphic Detail*, Coronavarus Statistics. The Economist, February 29th, 2020.
https://web.cortland.edu/matresearch/CoronavarusEconomist.pdf

About the Author:

Jorge Luis Romeu was, for sixteen years, a Research Professor at Syracuse University. He is currently an Adjunct Professor of Statistics. Romeu retired Emeritus from the State University of New York and worked, as a Senior Research Engineer, with IIT Research Institute. He received seven Fulbright assignments in Mexico, Ecuador, Colombia and the Dominican Republic. Romeu has a doctorate in Statistics/O.R., is a C. Stat Fellow of the Royal Statistical Society, a Member of the American Statistical Society and of the American Society for Quality. He is Past ASQ Regional Director and holds Reliability and Quality ASQ Certifications. Romeu created and directs the Juarez Lincoln Marti Int'l. Ed. Project (https://web.cortland.edu/matresearch/) dedicated to support higher education in Ibero-America.