Introduction

In late 2019, the city of Wuhan in China, reported a number of sick individuals suffering from pneumonia like symptoms. Further, this was exacerbated by a late acknowledgement by the Chinese government itself to a rising viral spread throughout the region. At the beginning of 2020, the virus was isolated and named “Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)”, later on the WHO named Covid-19 a pandemic. The labelling of Covid-19 as a pandemic has severe implications due to the spread of the virus throughout the world. The disease is transmitted through multiple avenues: airborne, and through droplets. Through creating a globalized society, we have created an interlinked world where a virus such as Covid-19 can spread readily throughout the global population.

Problem Statement

Through the early months of the pandemic, the disease has transformed our lives. From lockdowns, to preventative health measures, and inevitable job losses. Countries are facing a pandemic while not at the lethality of the Spanish flu, it does affect certain populations more than others. Similarly, of those effected, many can be saved through medical intervention. The problem though is that in countries cannot meet the rising infection rates to perform medical interventions, and more developed countries are equally having a rough time. Through data analysis of cases, we can discern how countries around the world are dealing with the virus.

Intention

I intend to assess the hypothesis that covid testing and positive covid rates suffer from the class size paradox. The paradox is that when children are surveyed, the reported number is higher than what actually occurs in that the children being surveyed generally come from larger family homes. Similarly, as rates of covid testing go up what is expected is that there is an increase in positive rates. I will look at country data from multiple countries, parse through and observe by three measures of low, mid, and high covid testing, and see the rates covid positivity. I would like to also explore the effect of comorbidities between different sets of ages and how they are affected by Covid. Likewise, if there is a prominent comorbidity that is associated with covid deaths, that would be interesting to see. In performing the exploratory data analysis, I do not believe I will have much difficulty in doing the steps necessary to generate viable data. Steps such as: Replacing Headers, Formatting data into a more readable format, identify outliers and bad data, find duplicates, fix casing or inconsistent values, and Conduct Fuzzy Matching. I have in my current professional position as well in my academics performed these many times. If anything, it will be time consuming with such a large dataset. Therefore, I find it essential, that I load only relevant columns and indexes necessary for the pursuit of the hypothesis. I have been wanting to do something like this in one of my classes, and I look forward to discovering any unknown links! Below is a sample dataset that I will be using:

| **Sex** | **Age group** | **All Deaths involving COVID-19 (U07.1)1** | **Deaths from All Causes** | **Deaths involving Pneumonia, with or without COVID-19, excluding Influenza deaths (J12.0–J18.9)2** | **Deaths involving COVID-19 and Pneumonia, excluding Influenza (U07.1 and J12.0–J18.9)2** | **All Deaths involving Influenza, with or without COVID-19 or Pneumonia (J09–J11)3** | **Deaths involving Pneumonia, Influenza, or COVID-19 (U07.1 or J09–J18.9)4** | **Population5** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total |  |  |  |  |  |  |  |  |
|  | All ages | 347,131 | 3,329,907 | 331,354 | 161,474 | 8,726 | 524,629 | 328,239,523 |
|  | Under 1 year | 39 | 18,436 | 193 | 6 | 21 | 247 | 3,783,052 |
|  | 1–4 years | 21 | 3,388 | 120 | 3 | 59 | 197 | 15,793,631 |
|  | 5–14 years | 58 | 5,440 | 153 | 11 | 74 | 274 | 40,994,163 |
|  | 15–24 years | 525 | 34,922 | 710 | 188 | 80 | 1,123 | 42,687,510 |
|  | 25–34 years | 2,278 | 71,710 | 2,577 | 1,012 | 232 | 4,058 | 45,940,321 |
|  | 35–44 years | 5,991 | 102,128 | 5,872 | 2,718 | 353 | 9,465 | 41,659,144 |
|  | 45–54 years | 16,282 | 186,691 | 15,505 | 7,950 | 761 | 24,488 | 40,874,902 |
|  | 55–64 years | 40,758 | 431,615 | 41,990 | 20,616 | 1,608 | 63,508 | 42,448,537 |
|  | 65–74 years | 73,856 | 661,771 | 74,220 | 37,521 | 1,865 | 112,151 | 31,483,433 |
|  | 75–84 years | 95,848 | 809,670 | 92,804 | 46,562 | 1,887 | 143,734 | 15,969,872 |
|  | 85 years and over | 111,475 | 1,004,136 | 97,210 | 44,887 | 1,786 | 165,384 | 6,604,958 |

https://www.cdc.gov/nchs/nvss/vsrr/covid\_weekly/index.htm # Covid website

https://github.com/owid/covid-19-data/blob/master/public/data/ecdc/full\_data.csv # Covid CSV

with large data pool.