COVID DATA PROJECT

Project Notes

1. About / Initial Thoughts

The COVID-19 dataset, with data collected and provided by the WHO and curated by ourworldindata.com (provide more) is vast. The roughly 126,000 entries cover (count countries) countries who report new cases, deaths, and vaccination rates to the WHO. Additionally, the data also covers, population, population density, rates of smoking, diabetes, cardiovascular disease, etc. With such an extensive dataset, there should be some interesting and enlightening data analysis to be extracted.

1. Cleaning the Data
   1. Removing Unwanted / Irrelevent Data

I decided to work on several critical factors that have been reported to play a decisive role in the spread of COVID-19. These columns include:

* Continent
* Location
  + Country
* Date
* Total Cases
  + Cumulative Count of COVID-19 infections
* New Cases
  + Total Infections Recorded on the Day
* Total Deaths
  + Cumulative Count of COVID-19 Deaths
* New Deaths
  + Total Infections Recorded on the Day
* ICU Patients
  + The number of ICU patients. This can vary by country. Most often is absent.
* Hospital Patients
  + The number of hospital patients. This can vary by country. Most often is absent.
* People Vaccinated
  + Total Number of people with at least one vaccination
* People Fully Vaccinated
  + Total Number of People Fully Vaccinated
* New Vaccinations
  + Total Number of new doses given to patients
* Stringency Index
  + Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
* Population
  + Total Number of People in each country
* Population Density
  + Population divided by area. A general impression of how closely people live together. This might need to be updated… For example, the population density of Saudi Arabia may be quite low. However, since most people in the country live in or near cities, this could be a misleading metric.
* Median Age
  + The middle of the ordered list of ages in a country
* Over 60, Over 70
  + The number of people that are over the ages of 60 and 70 respectively
* GDP per Capita
  + The GDP (Gross Domestic Product) of a country when controlled for population
* Cardiovascular Related Death Rate
  + The rates of death relating to cardiovascular disease in 2017 per 100,000
* Diabetes Rate
  + Diabetes prevalence in 2017 of people aged 20-79 per 100,000
* Smoking Rate / Male and Female
  + The current rate of smokers (will likely combine because there are no other gender specific columns in the set)
  1. Remove Duplicate Data

There was no duplicate data found. The members who compiled the data at open world data have already ensured this step was completed.

* 1. Fix Structural Errors
  2. Analyze Outliers
  3. Handle Missing Data
  4. Validation