CS 3
Lasata Tuladhar (lt9vx)
DS 4002
5/6/24

Analyzing National COVID-19 Data from 2020 to Early 2024

Topic Overview:

Starting from 2020, the COVID-19 pandemic has heavily impacted the world as a whole. It was the first global pandemic since the influenza pandemic of 1918-1919 and although it was also spread through animals at first, it was a new variant of disease that had not been studied before. This novelty the disease had, as well as some of its symptoms resembling the common cold, allowed it to grow into the widespread pandemic that it is known for today. Now, even in 2024, COVID-19 is still extremely prevalent, especially for the United States. Several countries such as Cameroon, Laos, and El Salavador have already brought down the number of new cases per day to nearly zero. America, however, has had several large spikes recently and the highest number of total cases in the world. Long COVID is also a growing issue in the US. Due to its lasting effects on the body and mind, and the symptoms varying between people, it poses a concern to the working class.

As a data scientist, you will analyze the trends in national COVID-19 data from the Centers for Disease Control and Prevention (CDC). This has the potential to determine which factors (such as weekly hospitalizations or weekly deaths) are highlighted in the data. This can also aid policymakers and hospital staff, and lead to a better understanding of how best to combat COVID-19 in the future. Using an Autoregressive Integrated Moving Average (ARIMA) model to train and test this data will ultimately create a time-series forecasting prediction on future COVID-19 counts in the United States.

The Deliverables:

As mentioned before, the CDC is seeking a data scientist who can effectively analyze their COVID-19 data from 2020 to early 2024, in order to determine the trends and create a time-series forecasting prediction for future COVID-19 counts. This task requires an initial cleaning and subsetting of the data. Then, you will analyze the autocorrelation of the data and based on that, create the ARIMA model. From there, you will determine the trends in the data by rolling the time-series forecasting. Finally, you will check the accuracy of the model by performing two additional tests — one for auto correlation and one for normality.