**Capstone Project Presentation by Omolola Lawani**

I obtained my dataset from the Kaggle website, focusing on COVID-19 data. After unzipping the files, I selected three datasets: **Worldometer**, **USA Country**, and **COVID-19**. These datasets were chosen intentionally as they contained all the information I needed for my analysis. However, for the final visualization, I used only two datasets: **Worldometer** and **COVID-19**.

**Data Preparation**

I began by importing the necessary libraries and preparing the data. I read the datasets using .read\_csv() and examined their structure. Next, I dropped unnecessary columns using .drop() and checked for missing values with .isnull().sum(). To handle the missing values, I filled numeric columns with their median values and categorical columns with their mode. This approach preserved the dataset's integrity, as dropping rows would have significantly reduced the data.

Additionally, I removed rows that were not effective for my analysis. I then reviewed the **Worldometer** dataset's structure and renamed certain columns to make them more readable.

**Feature Engineering**

My primary objective was to predict the spread of COVID-19. To start, I analyzed the top 20 countries with the highest total cases using a horizontal bar chart. Bar charts are intuitive and easy to interpret. The visualization revealed that the **USA** had the highest number of COVID-19 cases.

Curious about the potential impact of population size, I examined the top 20 countries with the largest populations using another bar chart. Surprisingly, the results showed that although **India** had the highest population, the **USA** still had the highest number of cases. This finding indicated that factors beyond population size contributed to the high rate of COVID-19 cases.

To further validate the relationship between population size and total cases, I created a scatter plot. This visualization revealed outliers, with some countries reporting unusually high case numbers. I also explored whether a higher population correlated with an increased mortality ratio, recovery rate, and cases per population. For this analysis, I grouped data by **WHO regions** rather than individual countries.

Upon merging all countries in the Americas, I found that **Europe** had the highest mortality ratio, followed by the Americas. Unfortunately, due to the limitations of my dataset, I could not delve deeper into why the **USA** had the highest COVID-19 cases.

**Exploring the COVID-19 Dataset**

Next, I analyzed the **COVID-19** dataset. I read the data, cleaned it, handled missing values, dropped unnecessary columns, and renamed others where necessary. Afterward, I examined the relationships between variables in the **Worldometer** dataset.

A strong positive correlation was observed among variables such as **Total Cases**, **Total Deaths**, **Active Cases**, **Total Recovered**, and **Serious Critical** cases. An increase in **Total Cases** corresponded to an increase in these other variables. However, I noticed a weak correlation between **Population** and variables like **Total Cases** and **Total Deaths**, indicating that a higher population does not necessarily result in more cases or deaths.

**Time Series Analysis**

I intended to merge the **Worldometer** and **COVID-19** datasets for further analysis but faced challenges due to inconsistencies in their sizes. Instead, I proceeded to create a time series analysis using the **Prophet** library.

To prepare the data, I formatted the date column into a datetime format using pd.to\_datetime(). The time series forecast, based on observations from the past seven months, predicted trends for the next three months. The results showed a potential increase in COVID-19 cases during this period. The shaded area in the chart represented uncertainty in the predictions, influenced by factors such as improvements in healthcare systems, vaccine availability, and virus mutations.

**Model Performance and Conclusion**

Finally, I assessed the model's performance accuracy. Despite dataset inconsistencies, the forecast confirmed a likely increase in COVID-19 cases, as indicated by the upward-sloping trend line. The blue shaded area represented the range of possible future cases.

In summary, my analysis highlighted key insights into the spread of COVID-19, relationships between variables, and potential future trends. Further research with more comprehensive datasets could provide deeper understanding and more robust conclusions.