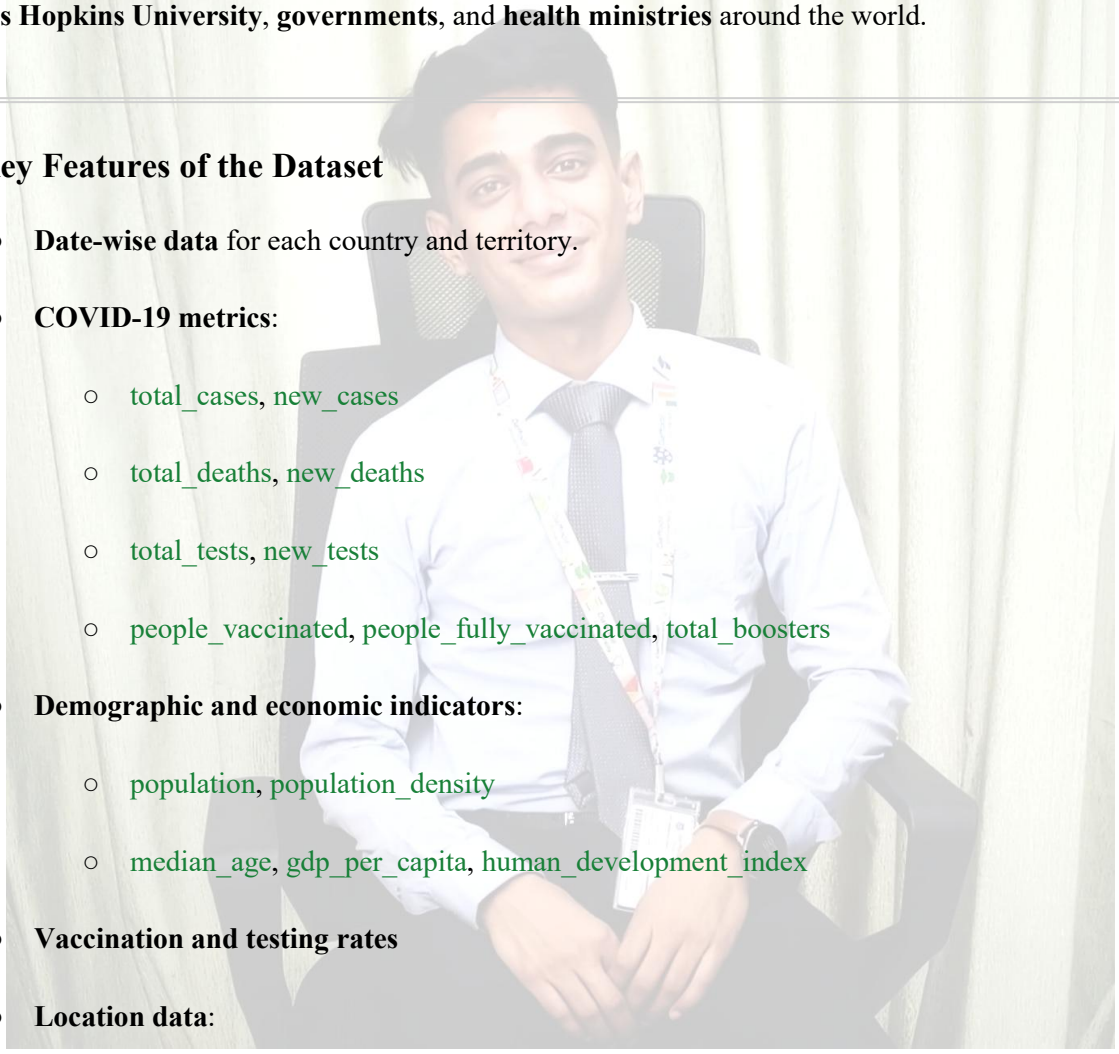
COVID-19_Data_Analysis_Sheet_01

The file owid-covid-data.csv is a comprehensive dataset provided by **Our World in Data (OWID)** that contains global information about the **COVID-19 pandemic**. This dataset is widely used for research, visualization, and policy analysis, offering a daily, country-level view of key pandemic-related metrics.

It includes data from the beginning of the outbreak (late 2019) and is updated regularly. The dataset combines information from various official sources such as the **World Health Organization (WHO)**, **Johns Hopkins University**, **governments**, and **health ministries** around the world.

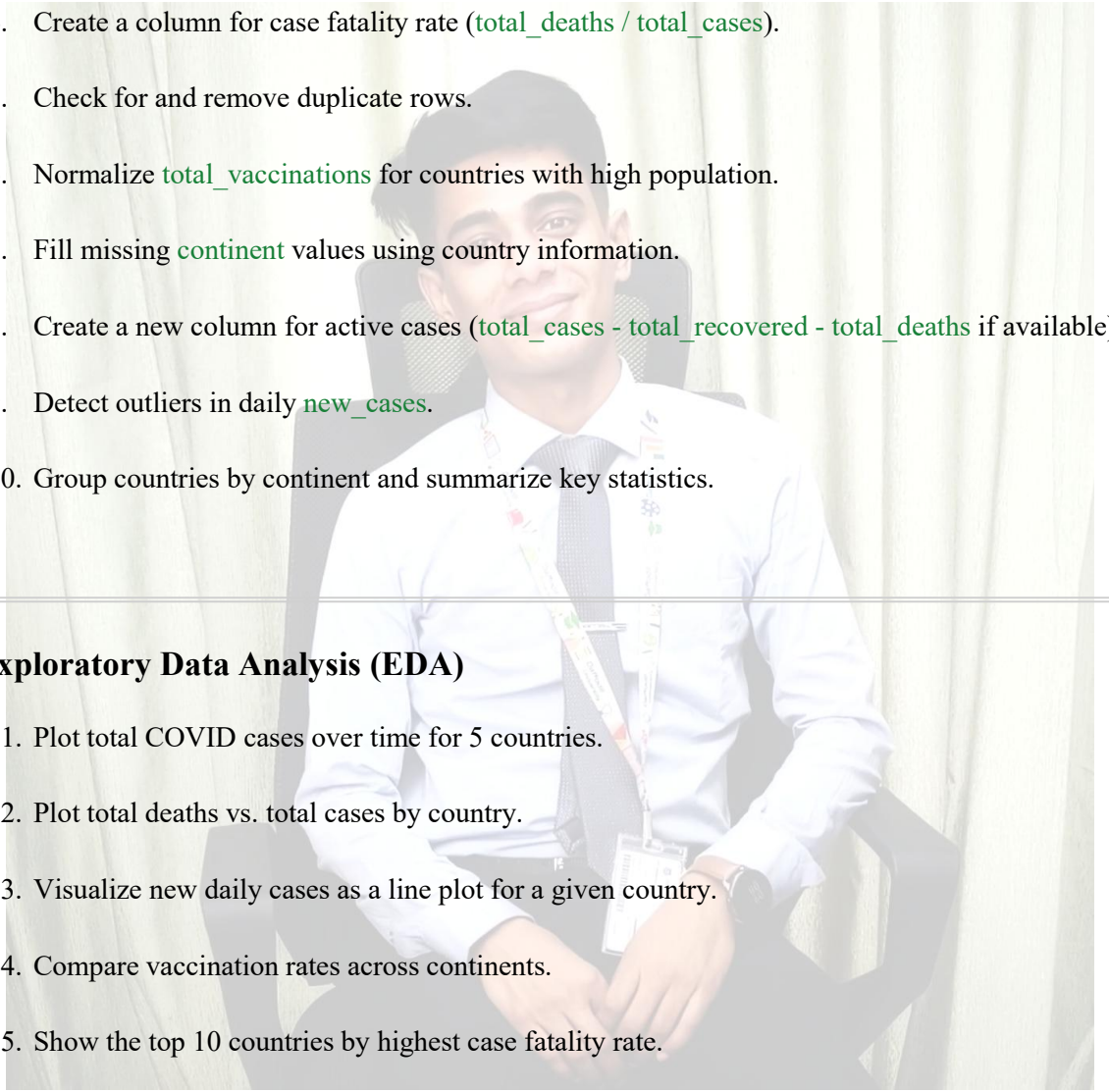


Key Features of the Dataset

- **Date-wise data** for each country and territory.
- **COVID-19 metrics**:
    - total_cases, new_cases
    - total_deaths, new_deaths
    - total_tests, new_tests
    - people_vaccinated, people_fully_vaccinated, total_boosters
- **Demographic and economic indicators**:
    - population, population_density
    - median_age, gdp_per_capita, human_development_index
- **Vaccination and testing rates**
- **Location data**:
    - location (country name), continent, iso_code

Data Cleaning & Preprocessing (Beginner)

1. Handle missing values in total_cases, new_cases, and total_deaths.
2. Convert date column to datetime format and extract year/month.
3. Filter data for a specific country (e.g., Bangladesh) and save it to a new file.
4. Create a column for case fatality rate (total_deaths / total_cases).
5. Check for and remove duplicate rows.
6. Normalize total_vaccinations for countries with high population.
7. Fill missing continent values using country information.
8. Create a new column for active cases (total_cases - total_recovered - total_deaths if available).
9. Detect outliers in daily new_cases.
10. Group countries by continent and summarize key statistics.



Exploratory Data Analysis (EDA)

11. Plot total COVID cases over time for 5 countries.
12. Plot total deaths vs. total cases by country.
13. Visualize new daily cases as a line plot for a given country.
14. Compare vaccination rates across continents.
15. Show the top 10 countries by highest case fatality rate.
16. Use histograms to show the distribution of new_cases.
17. Plot a correlation heatmap for numerical columns.
18. Compare median new cases before and after vaccination rollout.

19. Use boxplots to show daily new_cases spread by continent.
20. Create a bar plot showing total tests per country.

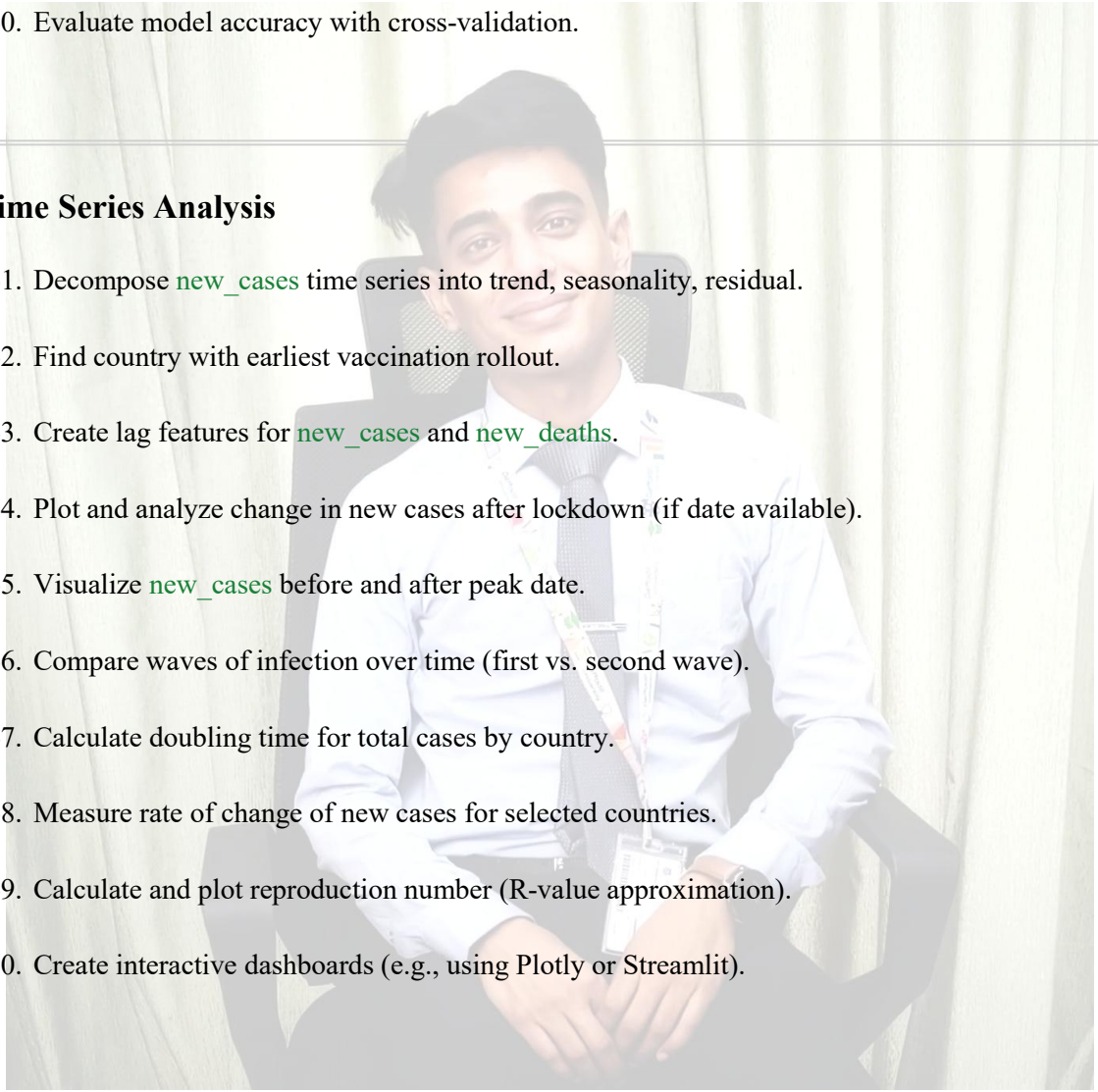
Descriptive & Inferential Statistics

21. Calculate mean and median of total_cases per continent.
22. Compute standard deviation of new_deaths for selected countries.
23. Test if there's a statistically significant difference in new_cases between two continents.
24. Compute rolling averages (7-day, 14-day) for new_cases.
25. Calculate z-scores to detect anomalies in daily new cases.
26. Test correlation between population and total cases.
27. Calculate quantiles and IQR for total_deaths.
28. Perform a chi-square test for association between continent and high/low fatality rates.
29. Calculate skewness and kurtosis for new_cases.
30. Use ANOVA to compare average daily new cases across continents.

Machine Learning / Prediction

31. Predict future new_cases using linear regression (per country).
32. Build a classification model to detect "high risk days" based on new_cases, new_deaths.
33. Cluster countries based on case trends and vaccination rates (K-means).
34. Use time series forecasting (ARIMA) for new_cases in one country.
35. Predict total deaths using features like total cases, vaccinations, and tests.

36. Build a model to predict continent based on country stats (classification).
37. Use decision trees to classify low-risk vs. high-risk countries.
38. Apply dimensionality reduction (PCA) on the dataset and visualize.
39. Build a random forest model to predict new_cases.
40. Evaluate model accuracy with cross-validation.



Time Series Analysis

41. Decompose new_cases time series into trend, seasonality, residual.
42. Find country with earliest vaccination rollout.
43. Create lag features for new_cases and new_deaths.
44. Plot and analyze change in new cases after lockdown (if date available).
45. Visualize new_cases before and after peak date.
46. Compare waves of infection over time (first vs. second wave).
47. Calculate doubling time for total cases by country.
48. Measure rate of change of new cases for selected countries.
49. Calculate and plot reproduction number (R-value approximation).
50. Create interactive dashboards (e.g., using Plotly or Streamlit).