Exploratory Data Analysis (EDA) Report: COVID-19 Dataset

This report provides an exploratory analysis of a COVID-19 dataset, detailing its structure, statistical properties, and key insights. The analysis was performed as of **March 27, 2025**.

---

1. Dataset Overview

- **Total Rows:** 18,110
- **Total Columns:** 9
- **Columns:**
  - Sno: Serial number (unique identifier for each entry)
  - State/UnionTerritory: Region in India (e.g., states or union territories)
  - ConfirmedIndianNational: Number of cases among Indian nationals
  - ConfirmedForeignNational: Number of cases among foreign nationals
  - Cured: Number of recovered cases
  - Deaths: Number of deaths
  - Confirmed: Total confirmed cases (daily)
  - Datetime: Date and time of the entry
  - Cumulative_Confirmed: Cumulative total of confirmed cases up to that date

The dataset appears to track COVID-19 cases across various regions in India over time, with a focus on daily and cumulative metrics.

---

2. Missing Values Analysis

- **Missing Data:** None
  - All 18,110 rows have **non-null** values across all 9 columns.
  - This completeness simplifies analysis but should be verified against external sources to ensure no data was omitted during collection.

---

3. Data Types

- **Current Data Types:**
  - Sno: Integer
  - State/UnionTerritory: Object (string)
  - ConfirmedIndianNational: Integer
  - ConfirmedForeignNational: Integer
  - Cured: Integer

- Deaths: Integer
- Confirmed: Integer
- Datetime: Object (string)
- Cumulative_Confirmed: Integer

- **Recommendations:**
  - Convert Datetime from object to a datetime format (e.g., using pd.to_datetime in Python) to enable time-series analysis.

---

4. Summary Statistics

The table below summarizes the numerical columns in the dataset:

Metric	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed	Cumulative_Confirmed
Count	18,110	18,110	18,110	18,110	18,110	18,110
Mean	0.30	0.036	278,637	4,052	0.34	169.1
Standard Deviation (std)	3.87	0.60	614,891	10,919	4.12	279.03
Minimum (min)	0	0	0	0	0	0
25th Percentile (Q1)	0	0	3,360	32	0	4
Median (Q2)	0	0	33,364	588	0	40
75th Percentile (Q3)	0	0	278,869	3,643	0	231
Maximum (max)	177	14	6,159,676	134,201	180	1,160

Key Observations:

- **Low Means for Confirmed Cases:** The mean values for ConfirmedIndianNational, ConfirmedForeignNational, and Confirmed are near zero, indicating sparsity in daily case reporting.
- **High Variability:** Large standard deviations in Cured, Deaths, and Cumulative_Confirmed suggest significant regional or temporal variation.
- **Outliers:** Maximum values (e.g., 6,159,676 cured cases, 134,201 deaths) indicate extreme peaks in some regions or dates.

---

5. Trends and Patterns

Distribution Analysis:

- **Confirmed Cases (Confirmed, ConfirmedIndianNational, ConfirmedForeignNational):**
    - Predominantly zero values, suggesting many regions reported no new cases on most days.
    - Rare but significant spikes (e.g., max of 180 for Confirmed) indicate localized outbreaks.
- **Cured Cases:**
    - Highly right-skewed distribution, with a few regions or dates reporting millions of recoveries (max: 6,159,676).
- **Deaths:**
    - Also right-skewed, with most regions reporting low fatalities (median: 588) but some extreme cases (max: 134,201).
- **Cumulative Confirmed Cases:**
    - Shows a steady increase over time, consistent with the progressive nature of a pandemic.

Temporal Trends:

- After converting Datetime to a datetime format, a time-series plot could reveal:
    - Growth patterns in Cumulative_Confirmed.
    - Peaks in Deaths and Cured corresponding to waves of the pandemic.
- This EDA reveals a dataset with no missing values but significant skewness and variability in key metrics like Cured and Deaths. The predominance of zero values in daily confirmed cases suggests either under-reporting or a focus on cumulative tracking. Further analysis with visualizations and regional breakdowns could provide deeper insights into India's COVID-19 experience.