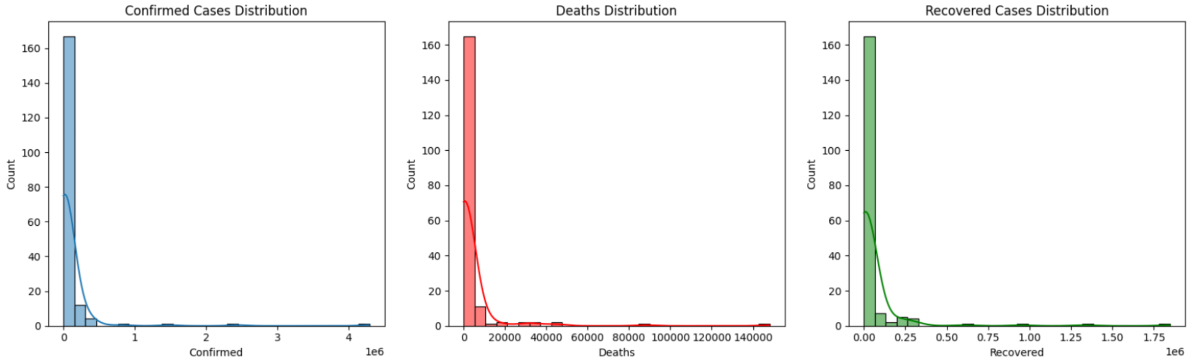
COVID-19 Feature Engineering & Data Visualization Report

1. Data Visualization with Seaborn and Matplotlib

To understand the spread and behavior of COVID-19 data, distribution plots were created for key numerical features:

- **Confirmed Cases:** Showed a right-skewed distribution with a few countries having extremely high case counts.
- **Deaths:** Distribution was heavily skewed, reflecting that a small number of countries suffered the majority of deaths.
- **Recovered:** Also skewed, indicating only a few countries experienced very high recovery rates.

These plots revealed that a small number of countries experienced the majority of the global COVID-19 impact.

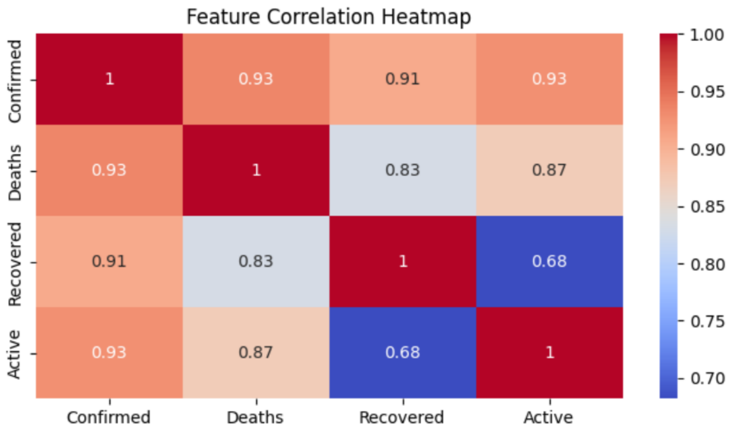


2. Feature Scaling

To normalize the range of values, two common scaling techniques were applied:

- **MinMax Scaling:** Rescaled features to the range [0, 1], preserving the original distribution.
- **Standardization (Z-score):** Centered values around 0 with a standard deviation of 1, helpful for algorithms sensitive to variance.

These methods are critical for machine learning algorithms that assume normally distributed or scaled input features.



3. Encoding Categorical Features

The WHO Region column (categorical) was transformed using **One-Hot Encoding**, which converts categories into binary features (0 or 1). This allows models to process and learn from categorical variables without introducing unintended ordinal relationships.

4. Correlation Analysis

A **correlation heatmap** was generated between major numerical features:

- Strong positive correlation was observed between Confirmed, Deaths, and Recovered cases.
- This indicates multicollinearity and linear relationships which can influence model behavior.

Visual correlation analysis helps in identifying which features might be redundant or provide unique information.

5. Conclusion

Feature engineering and visualization steps helped prepare the dataset for machine learning and advanced analytics. Through:

- Visualization,
- Normalization & Standardization,
- Encoding,
- Correlation Mapping,

the dataset has become cleaner, numerically consistent, and more suitable for predictive modeling.