COVID-19 Data Handling & Preprocessing Report

1. Dataset Overview

The dataset country_wise_latest.csv from Kaggle contains COVID-19 statistics for countries worldwide. It includes columns such as:

- Country/Region
- Confirmed
- Deaths
- Recovered
- Active
- WHO Region

The goal is to prepare this data for analysis by handling missing values, removing duplicates, and understanding basic patterns.

Collapse Output

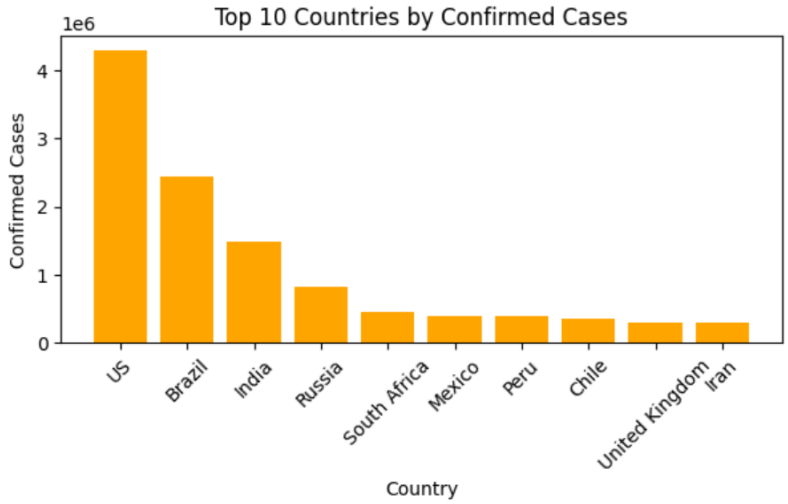
	...on	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region
0	Afghanistan	36263	1269	25198	9796	106	10	18	3.50	69.49	5.04	35526	737	2.07	Eastern Mediterranean
1	Albania	4880	144	2745	1991	117	6	63	2.95	56.25	5.25	4171	709	17.00	Europe
2	Algeria	27973	1163	18837	7973	616	8	749	4.16	67.34	6.17	23691	4282	18.07	Africa
3	Andorra	907	52	803	52	10	0	0	5.73	88.53	6.48	884	23	2.60	Europe
4	Angola	950	41	242	667	18	1	0	4.32	25.47	16.94	749	201	26.84	Africa

2. Data Cleaning

- **Missing values** were identified and filled using zero values for numerical consistency.
- **Duplicate rows** were checked and removed.
- The final dataset has a consistent shape with clean data, ready for further processing.

3. Basic Exploratory Data Analysis (EDA)

- The **top 10 countries** with the highest confirmed cases include the USA, Brazil, and India.
- **Deaths and recoveries** show similar trends to confirmed cases.
- A **correlation heatmap** showed strong correlation between Confirmed, Deaths, and Recovered.
- **Visualizations** like bar plots were used to show comparisons among countries.



4. Conclusion

This step completed essential data cleaning and basic exploratory analysis. The dataset is now ready for:

- Feature Engineering
- Scaling and Encoding
- In-depth Data Visualization