COVID-19 Descriptive Statistics and EDA Report

1. Introduction

The COVID-19 pandemic has had a profound impact on global health, and analyzing the spread and impact of the disease is crucial for effective policy-making. This project focuses on analyzing the **COVID-19 data** for Indian states and Union Territories, using **descriptive statistics** and **exploratory data analysis (EDA)** to uncover trends and patterns. The dataset provides insights into **active cases**, **discharged cases**, **death ratios**, and **other key metrics** that are vital for understanding the pandemic's impact.

Dataset Description

The dataset used in this project contains the most recent state-wise COVID-19 data as of **September 12, 2023**. The dataset includes the following key attributes:

- **State/UTs**: Names of Indian States and Union Territories.
- **Total Cases**: Total number of confirmed cases.
- **Active**: Total number of active cases.
- **Discharged**: Total number of discharged cases.
- **Deaths**: Total number of deaths.
- **Active Ratio (%)**: Ratio of active cases to total cases.
- **Discharge Ratio (%)**: Ratio of discharged cases to total cases.
- **Death Ratio (%)**: Ratio of deaths to total cases.
- **Population**: Population of the State/UT.

---

2. Objectives

The main objectives of this project are to:

- Perform **descriptive statistics** to summarize the dataset.
- Identify key **trends** and **patterns** in COVID-19 data across Indian states.
- Visualize the data using **bar charts**, **scatter plots**, **box plots**, and **heatmaps**.
- Conduct **advanced analysis**, including outlier detection, clustering, and hypothesis testing.
- Provide **data-driven insights** and **recommendations** for policymakers.

3. Methodology

3.1 Data Cleaning and Preprocessing

Before starting the analysis, the dataset was cleaned by:

- Checking for **missing values**, which were absent in the dataset.
- Verifying **data types** to ensure correct interpretation of numerical and categorical columns.
- **Handling duplicates**: Duplicate rows were removed to ensure accuracy.

3.2 Descriptive Statistics

The dataset was analyzed using descriptive statistics to provide:

- **Central tendency** (mean, median, mode) of key variables like Total Cases, Active Cases, and Death Ratios.
- **Measures of spread** (standard deviation, range, interquartile range) to understand variability.
- **Skewness** and **kurtosis** to identify data distribution and potential anomalies.

3.3 Data Visualization

Key visualizations were created to aid in understanding the data:

- **Histogram**: Used to visualize the distribution of Total COVID-19 Cases.
- **Boxplot**: Displayed the spread of Death Ratios and identified outliers.
- **Bar Chart**: Used to compare Death Ratios and Active Cases across states.
- **Scatter Plot**: Investigated the relationship between **Population** and **Total Cases**.

3.4 Advanced Analytics

- **Outlier Detection**: Used **Z-scores** to identify outliers in the Active Ratio (%) column. States with values above a Z-score threshold of 3 were flagged as potential outliers.
- **Clustering (K-Means)**: Applied K-Means clustering to segment states based on **Active Ratio**, **Discharge Ratio**, and **Death Ratio**. The clustering identified states with similar characteristics.
- **Hypothesis Testing**: Conducted a **t-test** to test the hypothesis: "States with higher populations have higher total cases." The test revealed no significant correlation, as the **p-value** was much greater than 0.05.

4. Key Findings

4.1 Top States with Highest Death Ratios

- **Punjab** and **Nagaland** exhibited the highest death ratios, indicating potential challenges in healthcare management and suggesting that more resources may be needed to reduce mortality.

4.2 Clustering Analysis

- **Cluster 0**: States like **Andhra Pradesh**, **Kerala**, and **Odisha** with high discharge rates and low active cases, indicating **good recovery rates** and effective COVID-19 management.
- **Cluster 1**: **Punjab** with a high death ratio and low recovery ratio, suggesting a **severe COVID-19 impact** and the need for urgent intervention.
- **Cluster 2**: States like **Maharashtra** and **Nagaland**, with moderate death ratios and recovery rates.

4.3 Hypothesis Testing

- The hypothesis that **higher population size** leads to **higher total cases** was **not supported** by the data. The **p-value** of 0.83 suggests that population size alone is not a statistically significant factor in determining the total number of COVID-19 cases.

---

5. Insights and Recommendations

5.1 High-Risk States

- **Punjab** stands out as a high-risk state, with a **high death ratio** and **low recovery rates**. The government should prioritize healthcare resources, **ICU capacity**, and **vaccination efforts**.
- **Nagaland** also requires focused attention due to its **moderate recovery rate** and **high death ratio**.

5.2 Clustering Insights

- States in **Cluster 0** show the best COVID-19 management, with **high discharge ratios** and **low active cases**. These states can serve as models for **best practices**.
- **Cluster 2** states, while facing moderate challenges, still require targeted intervention to **improve recovery** and **reduce death ratios**.

5.3 Policy Implications

- States with **high active cases** and **death ratios** should be prioritized for more resources, **healthcare support**, and **public health interventions**.
- The government should consider focusing on **health infrastructure** improvement in states with low recovery rates (like **Punjab**).
- Implement **region-specific strategies** based on clustering outcomes to optimize healthcare deployment.

---

6. Conclusion

This analysis provides valuable insights into the current state of COVID-19 across Indian states. By leveraging **descriptive statistics**, **clustering techniques**, and **hypothesis testing**, we have identified critical trends and provided actionable recommendations. This work can assist policymakers in **prioritizing resources** and **targeting interventions** for states facing the greatest challenges.

As the pandemic continues to evolve, this type of data-driven analysis will remain crucial for ongoing decision-making and public health management.

Next Steps:

- Further analysis can be conducted with **time-series data** to understand **trends over time**.
- **Predictive modeling** can help forecast future COVID-19 trends and improve resource allocation.