COVID-19 STATE WISE DATA ANALYSIS

1.DATASET DESCRIPTION:

1.1 Source : Kaggle covid-19 state wise dataset(StatewiseTestingDetails.csv)-16,337

1.2 Columns :

Column Name	Data Type	Description
Date	String / Date	The date on which the testing data was recorded.
State	String	Name of the state or union territory.
TotalSamples	Integer / Float	Total number of samples collected or tested up to that date.
Negative	Integer / Float	Number of samples that tested negative.
Positive	Integer / Float	Number of samples that tested positive.
Total Tested	Integer / Float	Overall count of conducted tests (may be same or derived from TotalSamples).
Source1	String	Primary source/reference for the data.
Source2	String	Secondary source or additional reference.
Source3	String	Tertiary or backup data source reference.

1.3 Data Quality :

- Missing values exist in some columns, particularly State and TotalSamples.
- Duplicate records detected in a few Date and State combinations.
- Outliers observed in TotalSamples, Negative, and Positive (e.g., extremely high sample counts on certain dates).
- Overall, the dataset is large, diverse across states and dates, and representative of testing data trends.

---

2. OPERATIONS PERFORMED

2.1 Data Cleaning & Exploration

- Checked and handled missing/null values in columns like State, TotalSamples, Negative, and Positive.
- Removed or flagged duplicate entries based on Date and State.
- Summarized numerical columns (TotalSamples, Negative, Positive) with mean, median, standard deviation, and interquartile range (IQR).
- Identified outliers in TotalSamples, Negative, and Positive.

2.2 Descriptive Analytics

- Distribution of testing across **States** (bar chart).
- Daily testing trend (TotalSamples over Date) (line chart).
- Positive vs. Negative case distribution (stacked bar / pie chart).
- Positive case distribution over time (line chart / histogram).
- Negative case distribution over time (line chart / histogram).

2.3 Relationship Analysis

- **TotalSamples vs. Positive cases** (scatter plot) to observe correlation.
- **State vs. Positive/Negative cases** (stacked bar chart) to compare test outcomes across regions.
- **Date vs. Positive/Negative cases** (line chart) to analyze trends over time.
- **Positive Rate (%) vs. State** (calculated as Positive/TotalSamples) to identify high-risk areas.
- **Positive vs. Negative cases correlation** (scatter plot / heatmap).

---

3. KEY INSIGHTS

3.1 Temporal Trends

- Testing activity shows peaks on specific dates, indicating surges in testing demand.
- Daily TotalSamples fluctuate, with some days showing significantly higher testing volumes.

- Positive cases generally track with total samples, but positivity rates vary over time.

3.2 State-wise Distribution

- Certain states consistently report higher testing volumes (TotalSamples) than others.
- Positive cases are concentrated in a few states, indicating regional hotspots.
- Negative cases dominate overall, but state-level variations show differing positivity rates.

3.3 Positivity Insights

- Positivity rate (Positive / TotalSamples) varies by state, highlighting areas with higher infection prevalence.
- Some states show sudden spikes in positivity rate on specific dates, suggesting outbreak events or mass testing drives.

3.4 Outlier Observations

- Outliers in TotalSamples exist for dates with unusually high testing numbers.
- Some states report extremely high positive counts relative to total samples, which may indicate data reporting anomalies or testing surges.

3.5 Geographic Spread

- Urbanized states tend to report higher total testing numbers and positive cases.
- State-level distribution shows concentrated testing in major cities or capitals.
- Regional trends can guide resource allocation and testing strategies.

---

4. RECOMMENDATION

4.1 Targeted Testing & Resource Allocation

- Focus testing resources on states showing higher positivity rates to contain outbreaks.
- Allocate additional testing kits on dates with historically high testing demand.

- Prioritize testing in urban or high-population states where sample volumes are consistently high.

4.2 Data Accuracy & Reporting

- Investigate outliers in TotalSamples and Positive counts to ensure data quality and reporting accuracy.
- Standardize data entry procedures across states to reduce duplicate or inconsistent records.

4.3 Positivity Monitoring & Alerts

- Implement real-time dashboards to monitor positivity rates (Positive / TotalSamples) by state.
- Issue alerts when positivity rates exceed thresholds to guide public health interventions.

4.4 Geographic & Temporal Insights

- Expand testing in underrepresented states with low sample counts to ensure comprehensive coverage.
- Analyze trends over time to predict testing surges and prepare resources accordingly.

4.5 Future Analytics Opportunities

- Develop predictive models to forecast positive case counts based on historical TotalSamples and trends.
- Segment states by positivity trends to guide targeted testing campaigns.
- Track testing efficiency (e.g., Positive / TotalSamples ratio) to optimize resource allocation.