**A PROJECT REPORT ON**

**Impact of COVID-19 on Global Health and Economy**

**USING PYTHON AND ML**

****

****

**Course Name: MSc Agriculture Analytics**

**Module Name: Programming for geodata Processing**

**Name: Nidhi Chaudhary**

**IIRS Registration No.: DAIICT19004**

**Submission Date: 07/04/2025**

|  |  |  |
| --- | --- | --- |
| Sr.No. | Topic | Page No. |
| 1 | **Introduction & Datasets** | **2** |
| 2 | **Research Question** | **2** |
| 3 | **Analysis** | **3** |
| 4 | **Visualization** | **4** |
| 5 | **Feature engineering** | **8** |
| 6 | **Model fitting** | **9** |
| 7 | **Conclusion & Reference** | **10** |

**CHAPTER 1 : INTRODUCTION AND DATASETS:**

**ABSTRACTION:**  This project analyzes the socio-economic impact of covid-19 using a comprehensive dataset covering infection rates, mortality, vaccination progress and government policies. After preprocessing the data, statistical and time series methods were applied to uncover trends, forecast deaths and explore key influencing factors. The highlights anomalies and regional differences to better understand the pandemic’s diverse effects. Finding aims to inform policy, enhance preparation and demonstration how data science support effective crisis response.

**PROJECT IDEA:** This project aims to forecast COVID-19- related new deaths using time –series dataset using python code and some ML models like XGBoost, LSTM and ARIMA. By comparing model performance, it identifies the best approach for accurate problem statement. The goal is to support better healthcare planning and crisis preparedness through data- driven forecasting, addressing challenges like data inconsistency and complex feature relationships.

**DATA COLLECTION:**

**Our World in Data**: COVID-19 Dataset and COVID-19 dataset(Healthcare Expenditure dataset)

**URL:** [**https://ourworldindata.org/**](https://ourworldindata.org/) **and** [**https://data.worldbank.org/**](https://data.worldbank.org/)

**DATASET DESCRIPTION:** The dataset tracks COVID-19 trends in INDIA and including other countries, in this cases, deaths, testing, and vaccinations. It covers healthcare capacity (ICU, hospital data), government response(stringency index), demographics(GDP, poverty),and heath factors (smoking , diabetes). This allows comprehensive analysis of the pandemic’s impact and response over time.

**CHAPTER 2 : RESEARCH QUESTION:**

**Problem statement**: This project analyzes COVID-19 trends in India and other country to understand how the pandemic affected public health over time. Using real-world data and python-based EDA, it explores the relationship between case number, deaths and healthcare factors- aiming to support or expand claims about how healthcare readiness impact COVID-19 outcomes.

**Claim**: COVID-19 has led to significant disruptions in global health systems and economies, especially affecting countries with lower healthcare infrastructure and economic resilience.

**CHAPTER 3 ANALYSIS:**

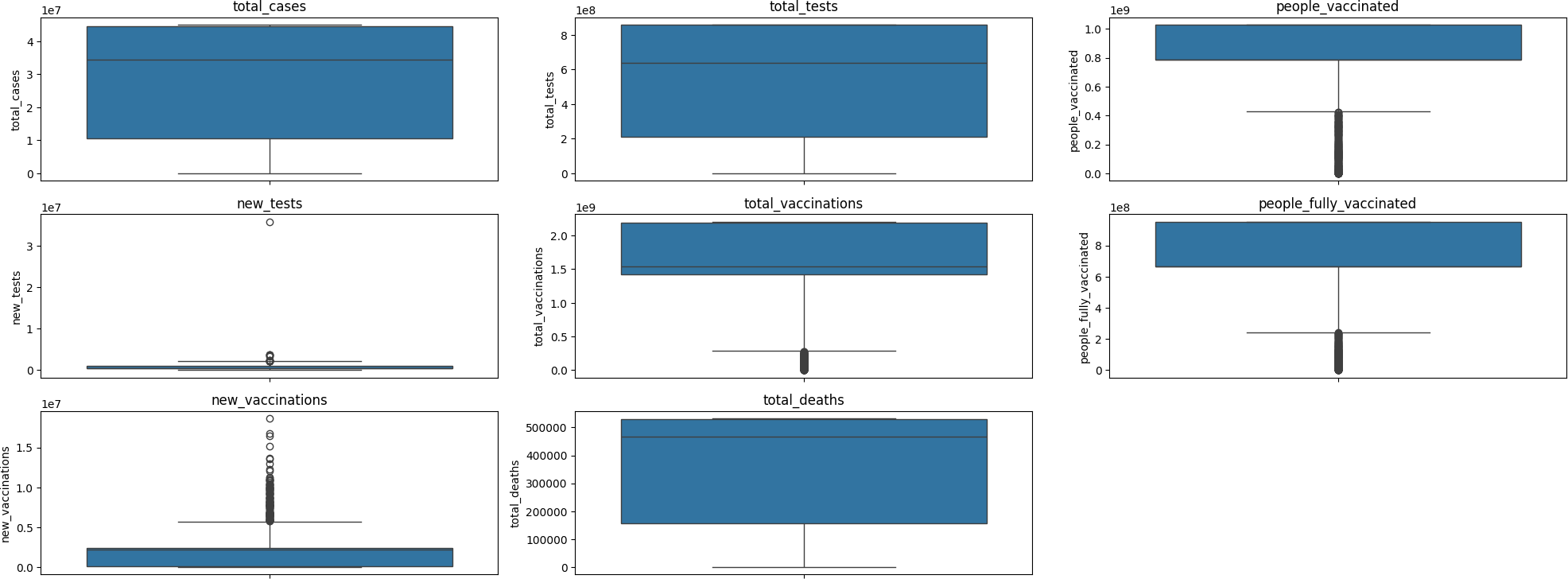
**3.1 DATA CLEANING AND PREPARATION:** To begin with, the dataset was loaded using the pandas library from a remote CSV file. Basic inspection revealed 1390 rows and 67 columns. We dropped irrelevant columns like iso\_code, continent, and location as they provide no analytics value in the context of a country study.

Then, addressed non- numeric columns. The tests units column, which had only one unique category and several missing values, was filled with the placeholder “Missing” and label encoded for analysis compatibility. The data column, initially in string format, was converted to data time to facilitate time-based operations.

**3.2 MISSING DATA HANDLING:** Missing data was analyzed visually using bar plots, matrix plots, and dendrograms from the missingno library. This visualization helped identify both the extent and patterns of missigness. Columns with over 50% missing data were removed, while remaining numeric columns were imputes using forward fill followed by mean substitution to ensure no missing values persisted.

**3.3 OUTLIER DECTION AND TREATMENT:** Outliers were detected using boxplots and interquartile range (IQR) method. Although extreme values were retained deliberately. As such spikes often reflect real- world pandemic events.Final boxplots confirmed that the data distribution remains balanced and reflective of actual trends.

The below figure of box plots depicts the distributions of COVID-19 cases, tests, vaccinations and deaths. Some metrics show a wide range of values, while others are more concentrated.

* **USING THE INTERQUARTILE RANGE(IQR) FOR OUTLIER DETECTION**

**IQR: Q3- Q1**

Where, Q1 is the first quartile (25th percentile) and Q3 is the third quartile (75th percentile)

**LOWER Bound: Q1 – 1.5\* IQR and UPPER BOUND: Q3+1.5 \* IQR**

**DECISION TO Retain Outliers:** After detecting the outlier, we made a deliberate decision not to remove them. In typical data analysis, outlies are often removed to avoid destroying statistical models. However , in the context of COVID-19 data, outliers are not always errors but may represent important events such as:- sudden surges in COVID-19 cases or deaths due to outbreaks or new variants. Mass testing efforts leading to large increases in reported tests. Government interventions that lead to abrupt increases in vaccinations or public health measures.

For example, if a particular day sees a dramatic increase in deaths or cases, it could reflect a significant event such as a new wave of infections or a backlog in reporting. These are crucial for under standing the course of the course of the pandemic and should be included in the true impact of COVID-19 and lead to inaccurate conclusion. Therefore, we decided to retain the outliers, allowing the dataset to maintain the integrity of the real-world phenomena it represents.

**CHAPTER 4: VISUALIZATION**

Univariate analysis was conducted to understand the distribution, central tendency and spread of individual variables such as total cases, new cases, deaths and vaccination rates in the COVID-19 dataset. We used histograms, boxplots, violin plots, and KDE plots to visualize each variable’s characteristics.

* **Histograms** revealed overall distribution shapes and helped identify peaks and gaps, such as multiple of infection. For instance, a histogram of daily new cases may reveal periods of high infection rates or the existence of multiple peaks, including pandemic waves.
* **Boxplots** highlighted variability and detected outliers, while often aligned with signification events like mass testing or vaccination drives.
* **Violin plots** provided details insights into data distribution and skewness, especially useful for understanding the spread of daily cases.
* **KDE plots** offered smooth density estimate, revealing underlying trends and peaks in case and deaths.
* **Univariates analysis provides** a foundation for understanding the distribution and variability of individual variables in the COVID-19 dataset. By leveraging multiple visualization techniques, we obtained a comprehensive view of the dataset, identified patterns and detected anomalies. These insights from the basis for further analysis , including bivariate and multivariate explorations.

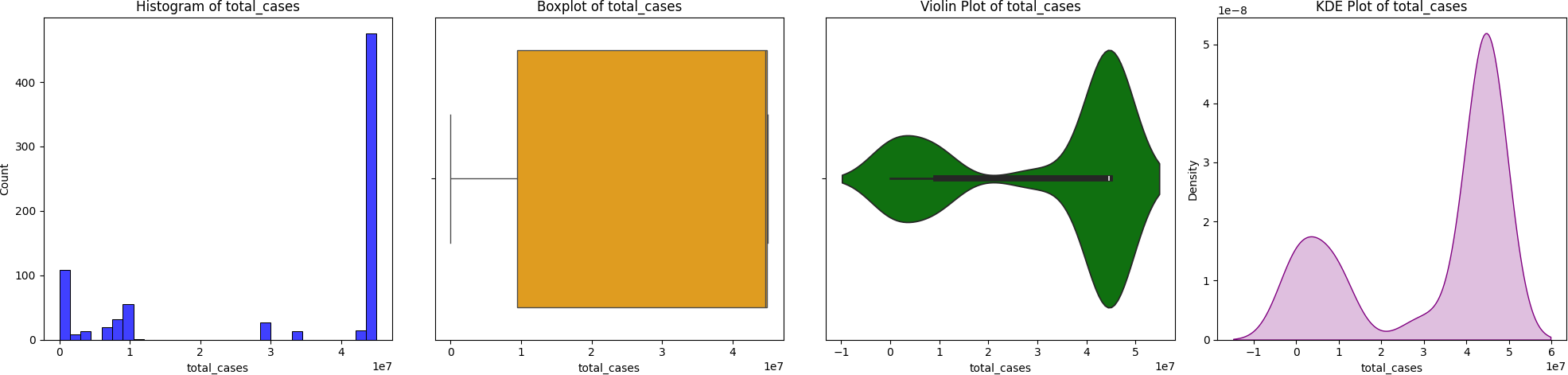
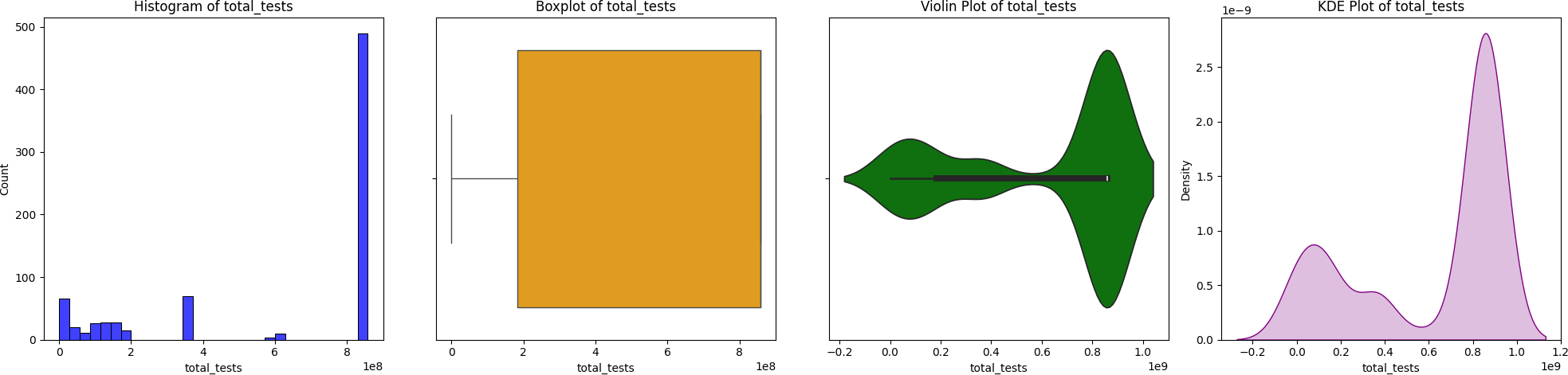
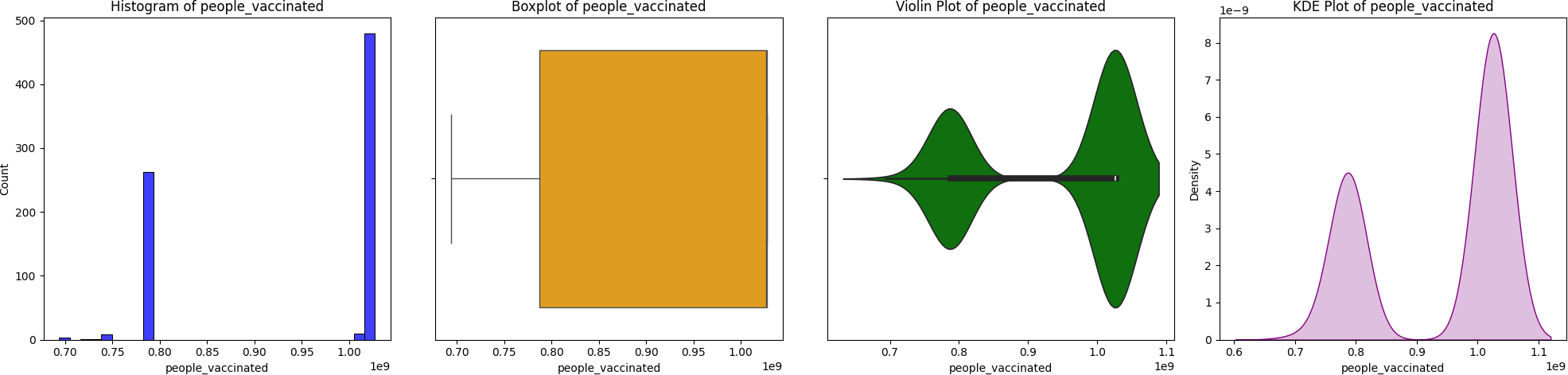
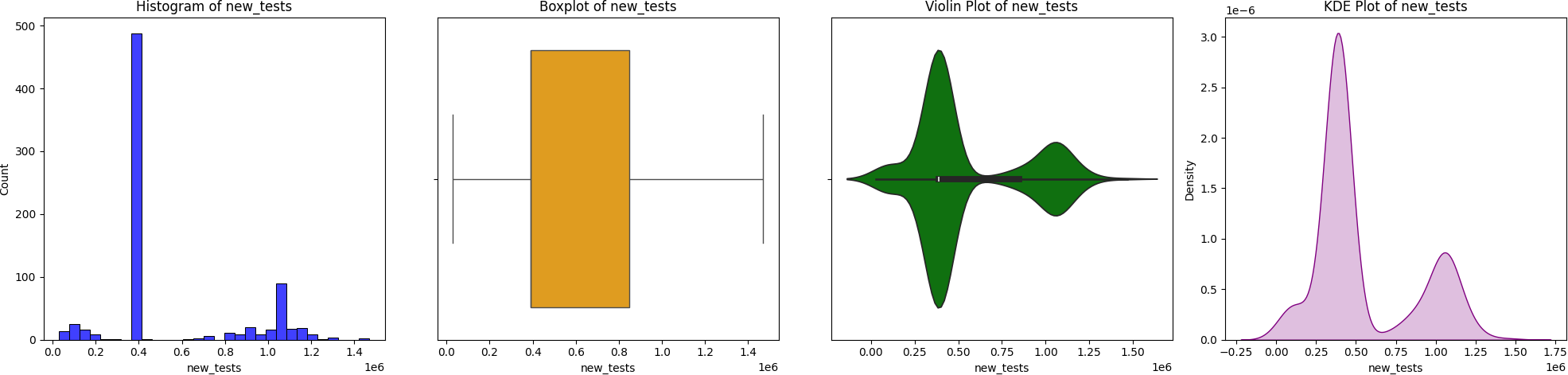
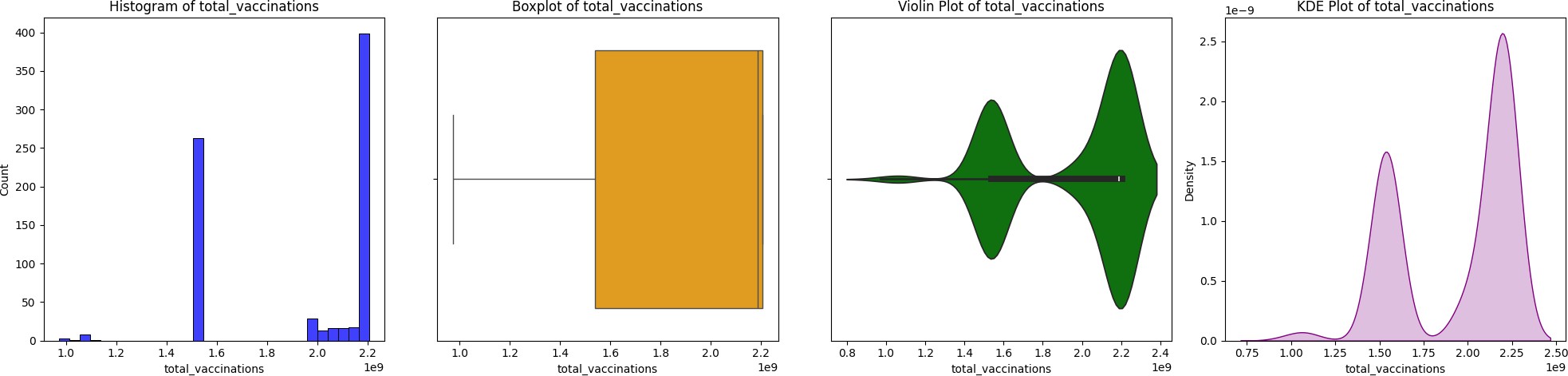


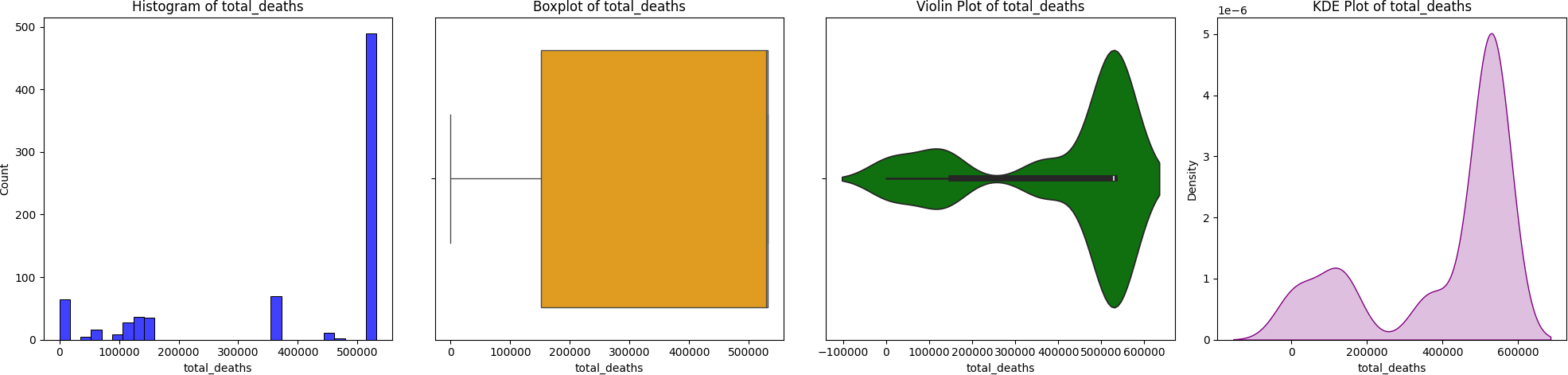
Figure 4.1: total cases

 Figure 4.2: total tests

 Figure 4.3: people vaccinate

**Figure: 4.4 new\_tests**

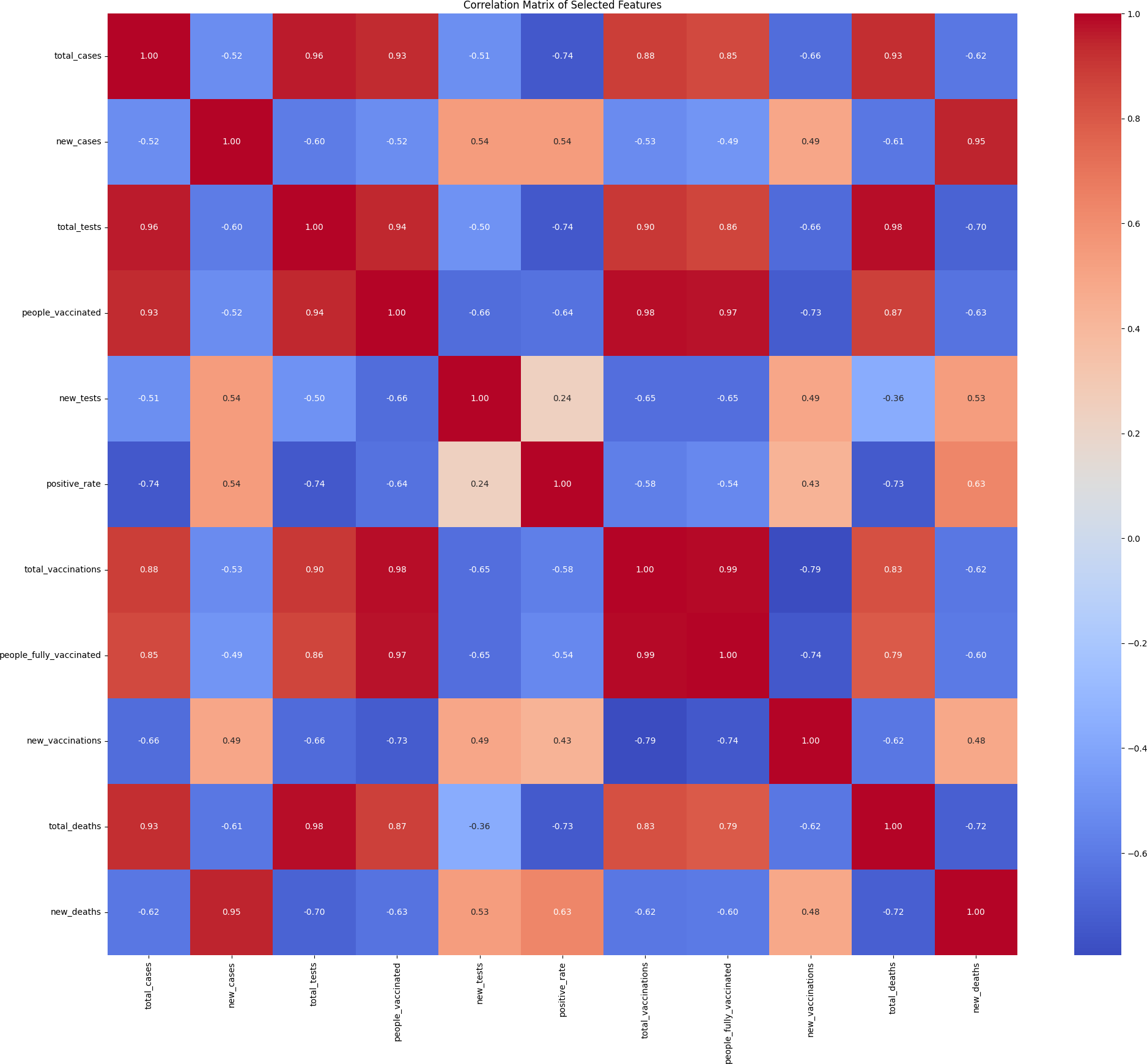


**Figure 4.5: total vaccination**

**Figure 4.6:** **total deaths**

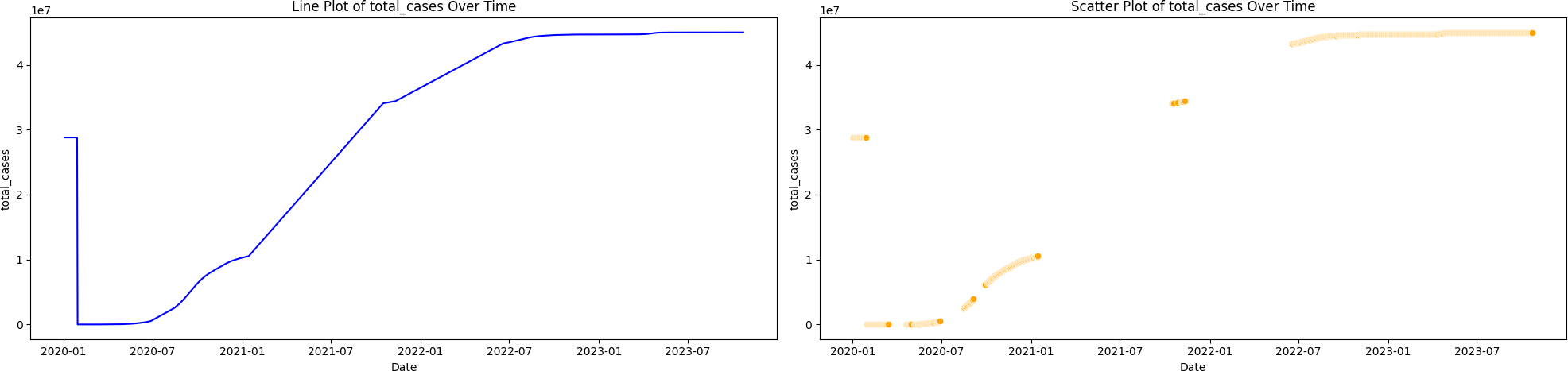
**MULTIVARIATE ANALYSIS:** Multivariate analysis was used to explore relationships between key COVID-19 variables like cases, deaths and vaccinations. Four visualization method- Heatmaps, line plots, scatter plots, and pair plots—were employed to uncover pattern and corrections.

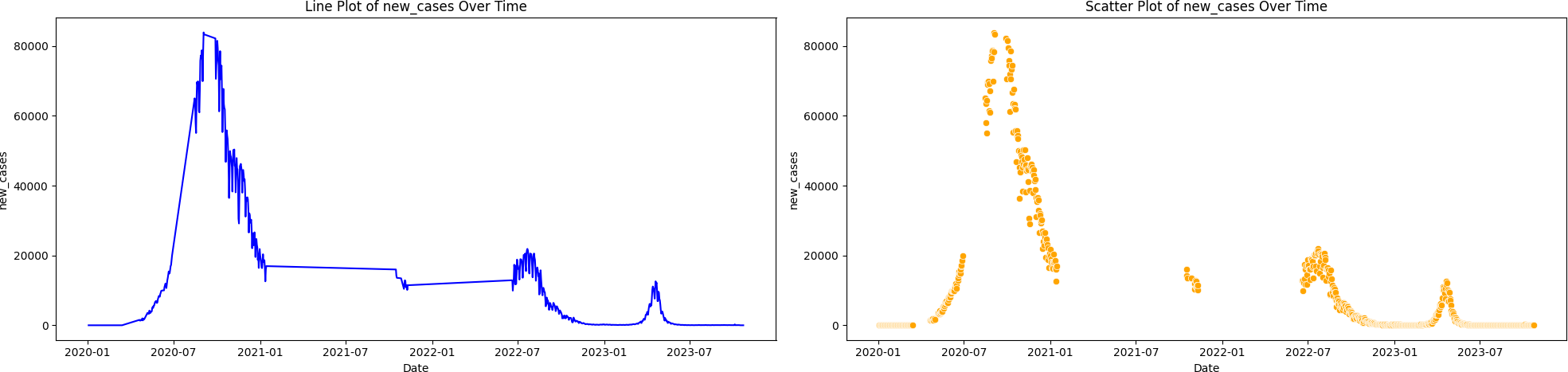
* **Heatmaps** illustrated the strength and direction of correlations between variables. For instance, strong positive correlation was observed besewn total cases and total deaths, including closely linked trends.

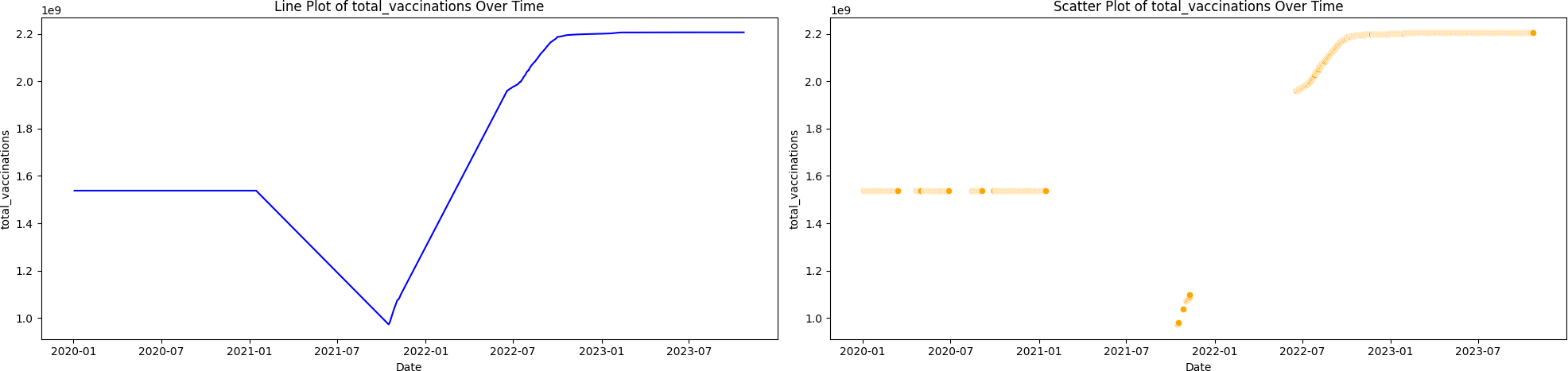
figure : 4.7 caption

**LINE PLOT:** Line plot was used to visualize time-based trends in daily cases, deaths, and vaccinations rates, they revealed key patterns such as pandemic waves, the impact of interventions, and the rise in vaccinations, while also highlighting anomalies like sudden surges or drops.

**Scatter plot:** Scatter plots are used to explore relationships between two countinuous variables. Each point in the plot represents an observation in the dataset, with its position determined by the values of the two variables being compared. Scatter plot illustrated relationship between pairs of variables, such ass total tests vs, total cases. They helped identify trends, clusters, and outliers , offering insights into correlations and unusual data points across regions or time periods.



Figure 4.8: total cases

Figure 4.9: new cases

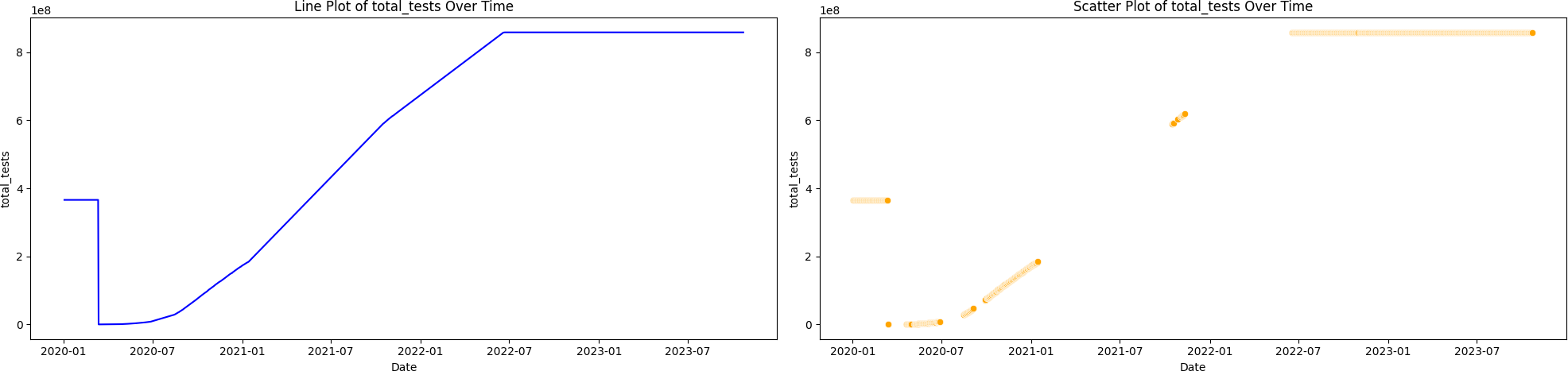
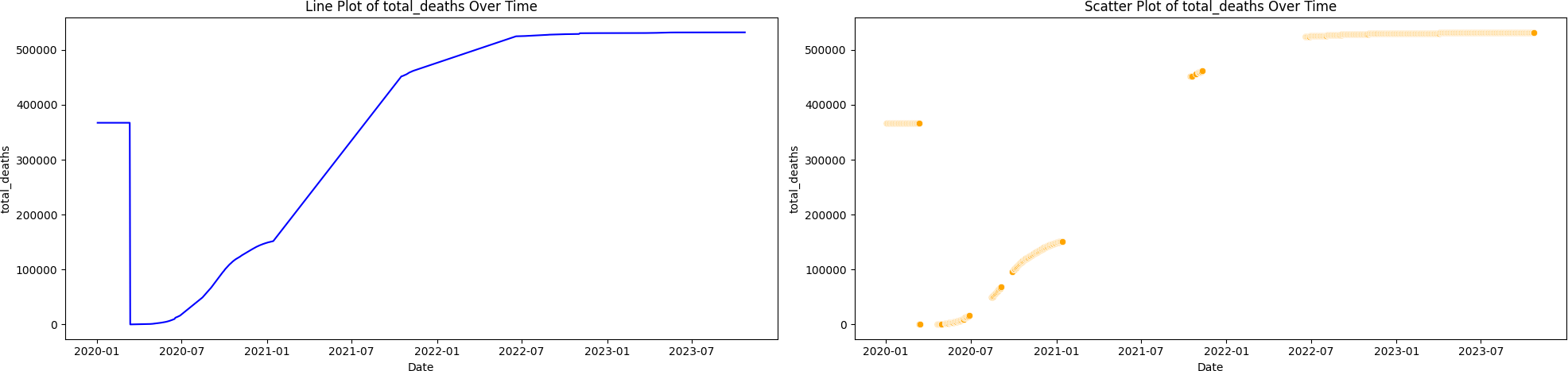
Figure 4.10: total tests

Figure 4.11: total vaccinations

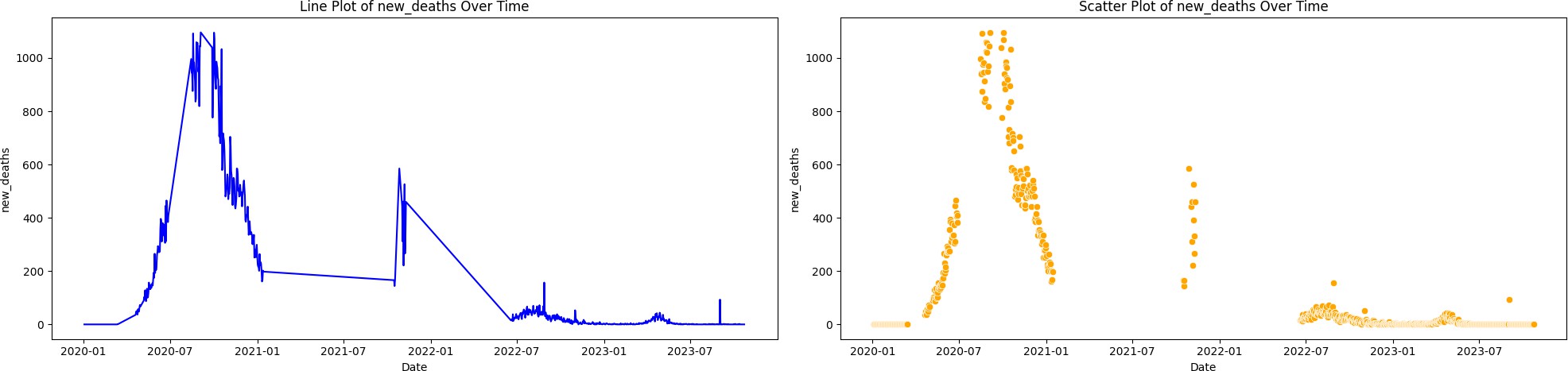
Figure 4.12: total deaths

Figure 4.13: new deaths

**PAIR PLOT:** Pair plot offered a comprehensive view of relationships among variables like total cases, deaths, tests, and vaccinations. They helped identify correlations, distribution pattern, clusters, and sign of multicollinearity.

**INSIGHTS FROM MULTIVARIATE ANALYSIS:** The heatmap revealed a strong positive correlation between total cases and total deaths, consistent with the severity of the pandemic.

* Line plots demonstrated the progression of daily cases and vaccinations, helping identify key turning points such as the impact of lockdowns or vaccine rollouts.
* Scatter plots showed a clear relationship between total tests and total cases, emphasizing the role of widespread testing in identifying cases.
* the pair plot provided a holistic view of how key variables relate, with clusters suggesting differences in pandemic outcomes across regions or time periods.

**CHAPTER 5: FEATURE ENGINEERING:**

**CORRELATION ANALYSIS:** Correlation matrix and heatmap were used to identify features with strong linear relationship to the target variable, helping prioritize relevant predictions.

**RANDOM FOREST:** A random forest regressor ranked feature based on their predictive power for new deaths. The most influential variables were identified using importance scores.

**K-BEST METHOD:** Using the select k-best method with mutual information regression, the top feature was selected based on their statistical relevance to the target, capturing both linear and non-linear dependencies.

**CHAPTER 6: MODEL FITTING:**

1. **ARIMA: TIME SERIES MOODEL:**

**ARIMA (**auto regressive integrated moving average) is a univariate model that predicts future values based on past observations and residuals.

* **Stationarity check:** ADF test confirms stationarity, differencing is used if required.
* **Parameter selection**: ACF and PACF plots determine AR(p) and MA(q) orders.
* **Diagnostics:** residuals are analyzed for normality and random via plots.
* **Forecasting:** 10 -days future prediction is generated after scaling.
* **Evolution:** ARIMA shows decent performance but struggled with nonlinear patterns.

**Process: Testing for stationarity (ADF Test)**

* The ADF test checks if the time series is stationary.
* If the P-value > 0.05, the data is non – stationary, and differencing is required to stabilize the mean. And if P-value <= 0.05, the data is stationary, and no differencing is needed.
* ARIMA assumes the time series is stationary for accurate modeling. The ADF statistic is less then the critical values (threshold-2.9 for 5% signification). The P-value is much less than 0.05, so that we reject the null hypothesis of non-stationarity.

1. **LSTM: LSTM** (long short-term memory) is an RNN architecture suited for sequence prediction.

* **Data Preparation:** new deaths are normalized and sequences are created for time – step learning.
* **Models:** two LSTM layers with dropout, trained using optimizer with MSE loss.
* **Results:**

**MSE:** 72,277.70

**MAE:** 82.69

**FORECASTS** Show good fit actual data, capturing temporal dependencies well.

**Data preparation:** Normalizing the target variable, new\_deaths between 0 and 1 ensures the data fits within a range suitable for LSTM, which performs better with normalized inputs. Minmaxscaler rescales the new\_deaths column so that all values are mapped between 0 and 1.

1. **XGBoost:**

XGBOOST is a robust ensemble learning method based on gradient boosting.

* **Strengths:** handles nonlinear relationships and works well on small datasets.
* **Result:**

**MAE:** 77.70

**MSE:** 28,504.92

**RMSE:** 168.83

**R²:** 0.924**—**explains 92.4% variance in new deaths.

* Performs well with limited preprocessing.

1. **MODEL COMPARISON:**

* LSTM excels due its ability to capture long-term temporal dependencies in time- series data.
* XGBOOST is strong on structure data but lacks sequential modeling capability.
* ARIMA is similar and interpretable but limited to linear to liner and stationary data.

**WHY LSTM over XGBOOST or ARIMA ? compared to XGBOOST:** XGBoost is great for tabular, structured data lacks the sequential understanding needed for time-series trends. While XGBoost can provide good results, it might require heavy feature engineering to capture temporal dependencies, which LSTM does inherently.

Compared to ARIMA: ARIMA assumes linear relationship and stationarity, which COVID-19 data often violates. It performs porly when the data has complex, nonlinear, or non-stationary pattern.

**CHAPTER 7: CONCLUSION:**

The analysis of the COVID-19 dataset reveled strong correlations between total cases , deaths and testing , supporting the claim that testing and timely interventions play a key role in managing the pandemic. Multivariate visualization highlighted patterns like rising deaths during infection sugers and improvement following vaccination rollouts. However, variations across time and regions showed that the data doesn’t always follow a uniform trend, indicating that external factors such as healthcare capacity, public policy and population behavior significantly affect outcomes. These insights support goal of understanding and predicting COVID-19 deaths, while also highlighting the complexity and variability in real world- data.

**OBSERVATIONS:**

* **potential for real- time forecasting:**  the model can bee integrated into live system to provide ongoing predictions and support proactive health interventions.
* **Scalability:** the model can be adapted for future pandemic or other diseases by incorporating additional features like vaccination rates and policy changes.
* **Temporal patterns:** the LSTM outperformed XGBoost and ARIMA, particularly in handling the sequential and non-liner data dynamic typical of pandemic data.

**REFERENCE:**

**Research paper link:** [**https://www.researchgate.net/publication/354946127\_COVID-19\_Exploratory\_Data\_Analysis\_of\_Coronavirus\_Outbreak**](https://www.researchgate.net/publication/354946127_COVID-19_Exploratory_Data_Analysis_of_Coronavirus_Outbreak)