Project Report

Project Objective

This project aims to analyze a real-world COVID-19 dataset and identify the countries with the highest rates of positive COVID-19 tests relative to their total testing numbers.

Data Processing & Analysis Steps

1. Data Loading & Validation

- **Dataset dimensions and structure** were examined.
- **Unique country count** and **date range** were identified.
- **Missing values and negative values** were checked.

2. Data Cleaning

- Converted `Date` column to date format.
- Replaced negative values in numeric columns with absolute values.
- Removed infinite values.
- Filtered out records with missing values in `daily_tested` and `daily_positive`.

3. Aggregation & Sorting

- Data was grouped by `Country_Region` to calculate:
  - **Total Tests Conducted** (`total_tested`)
  - **Total Positive Cases** (`total_positive`)
  - **Positivity Rate** = `total_positive / total_tested`

- Countries were sorted in descending order by **positivity rate**.

4. Key Findings

- **Top 10 Countries** with the highest positivity rates were identified.

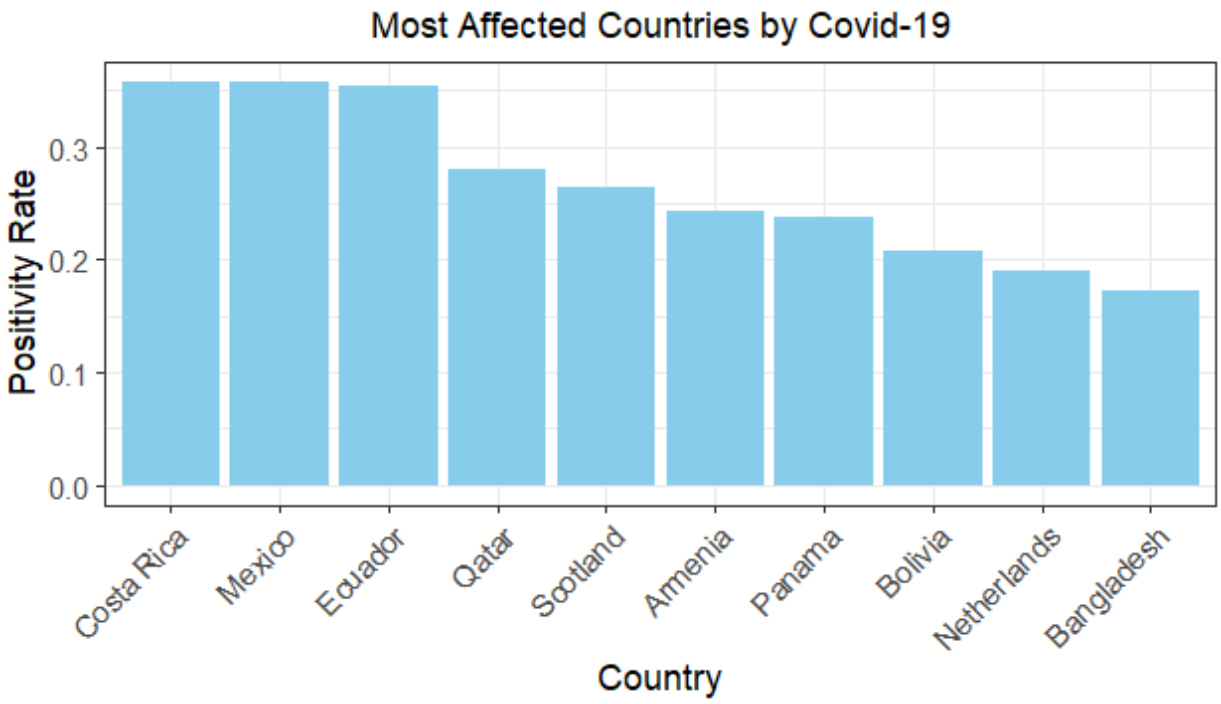
```
# A tibble: 10 × 4
   Country_Region total_tested total_positive positivity_rate
   <chr>                 <dbl>          <dbl>            <dbl>
 1 Costa Rica           320615         114791            0.358
 2 Mexico                15477           5528            0.357
 3 Ecuador               14895           5289            0.355
 4 Qatar                 19026           5314            0.279
 5 Scotland              13186           3491            0.265
 6 Armenia              438837         106424            0.243
 7 Panama                 2551            605            0.237
 8 Bolivia                 715            148            0.207
 9 Netherlands           43055           8153            0.189
10 Bangladesh          2442470         420235            0.172
```

- **Summary statistics** (average, max, min positivity rate) for these countries were computed.

```
# A tibble: 1 × 3
  avg_positivity max_positivity min_positivity
           <dbl>          <dbl>          <dbl>
1          0.266          0.358          0.172
```

5. Visualization

- A **bar chart** was created using `ggplot2` to visualize positivity rates for the top 10 countries.



6. Correlation Analysis

- A correlation matrix between **total tests** and **total positive cases** was computed.

```
> correlation_matrix
                total_tested total_positive
total_tested       1.0000000      0.9963484
total_positive     0.9963484      1.0000000
```

Project Outcomes & Insights

- The **top affected countries** (by positivity rate) were highlighted. The most affected country is **Costa Rica,** with a **0.358 positivity rate.**
- A **high and positive correlation (0.9963)** exists between total tests and positive cases.

- A **visual representation** shows that the top 3 affected countries are from the Americas continent, with a positivity rate greater than 0.3.

This analysis provides crucial insights into global COVID-19 testing trends and highlights the severity of the pandemic in different countries. 📊
