**EDA Report: COVID-19 Dataset Tracker**

**Prepared on March 25, 2025**

This report outlines the exploratory data analysis (EDA) and dashboard development process for a COVID-19 dataset sourced from the Ministry of Health. The analysis was conducted using Jupyter Notebook, with results visualized in an interactive Power BI dashboard. Each step—from data collection to insight documentation—is detailed below, providing a thorough understanding of the dataset’s characteristics, trends, and implications.

**1. Data Collection**

**Objective:** Obtain a reliable dataset capturing COVID-19 metrics.

Process: The dataset was acquired as a CSV file from the Ministry of Health, representing daily records across multiple countries.

**Details:**

* Fields: Includes date of reporting (`Date\_reported`), country identifiers (`Country\_Code`, `Country`), regional classification (`WHO\_region`), and key metrics (`New\_cases`, `Cumulative\_cases`, `New\_deaths`, `Cumulative\_deaths`).
* Size: Contains [Insert total rows from `df.shape[0]`] records.
* TimePeriod:Spans from [Insert `df['Date\_reported'].min()`] to [Insert `df['Date\_reported'].max()`], offering a broad view of the pandemic’s progression.
* Significance: This multi-dimensional dataset enables both temporal and geographic analysis of COVID-19’s spread and impact.

**2. Data Formatting**

**Objective:** Ensure the dataset is structured and typed correctly for analysis.

**Process:** Dates were transformed into a standardized format, and categorical fields were normalized to facilitate grouping and visualization.

**Actions:**

* Converted `Date\_reported` into a date-time format to support time-series trends.
* Standardized `Country\_Code` and `WHO\_region` as text fields, ensuring consistency across records (e.g., avoiding mixed formats like "US" vs. "USA").

**Outcome:** A well-formatted dataset, with dates ready for chronological analysis and categories aligned for regional comparisons.

**3. Correcting Inconsistencies**

**Objective:** Eliminate errors that could skew results, such as duplicates or illogical values.

**Process:** The dataset was scrutinized for duplicate entries and invalid data points, with corrections applied systematically.

**Findings:**

* Duplicates: Identified and removed [Insert number from `df.duplicated().sum()`] duplicate records, likely from redundant reporting.
* Negative Values: Found [Insert number] instances of negative `New\_cases` and [Insert number] negative `New\_deaths`, which were adjusted to zero. These anomalies may stem from data entry errors or corrections in reporting.

**Outcome:** A refined dataset free of redundancies and illogical entries, enhancing the reliability of subsequent analyses.

**4. Handling Missing Values**

**Objective:** Address gaps in the data to ensure completeness.

**Process:** Missing values were evaluated and filled based on reasonable assumptions tailored to the dataset’s context.

**Findings:**

* Extent: `New\_cases` had 18,729 missing entries, and `New\_deaths` had 30,306, while `Country\_Code` and `WHO\_region` had [Insert counts] missing values.
* Strategy:
* For `New\_cases` and `New\_deaths`, missing values were replaced with zero, assuming no report equates to no cases or deaths—a common practice in public health data when reporting is incomplete.
* For `Country\_Code` and `WHO\_region`, missing entries were labeled "Unknown" to retain these records without introducing bias.

**Outcome:** A fully populated dataset, eliminating null values while preserving data integrity for analysis.

**5. Descriptive Statistics**

**Objective:** Summarize the dataset’s key characteristics to understand its distribution and scale.

**Process:** Statistical measures were calculated for numerical fields, with additional breakdowns by region.

**Findings:**

Overall Metrics:

* `New\_cases`: Average daily cases were [Insert mean], with a median of [Insert median] and a peak of [Insert max], reflecting sporadic large outbreaks.
* `Cumulative\_deaths`: Total deaths averaged [Insert mean] per country, reaching a maximum of [Insert max].

Regional Insights:

* The [Insert top `WHO\_region`] region recorded the highest cumulative deaths at [Insert value], indicating a severe regional impact.
* Total new cases were highest in [Insert region], summing to [Insert value], suggesting concentrated infection spread.

**Significance:** These statistics reveal both the typical scale of the pandemic and its extreme outliers, guiding focus toward heavily affected areas.

**6. Correlation Analysis**

**Objective:** Explore relationships between variables to uncover dependencies.

**Process:** A correlation matrix was generated to assess how `New\_cases`, `Cumulative\_cases`, `New\_deaths`, and `Cumulative\_deaths` interrelate, visualized as a heatmap.

**Findings:**

* Strongest Link: [Insert pair, e.g., "Cumulative\_cases and Cumulative\_deaths"] showed a correlation of [Insert value, e.g., "0.92"], indicating that total cases closely predict total deaths over time.
* Daily Dynamics: `New\_cases` and `New\_deaths` had a correlation of [Insert value, e.g., "0.65"], suggesting a moderate but significant relationship—new infections often precede fatalities.

**Significance:** These relationships highlight the pandemic’s progression, where rising cases foreshadow increased mortality, critical for resource planning.

**7. Data Visualization**

**Objective:** Create visual representations to identify trends and patterns.

**Process:** Two primary visualizations were developed: a time-series plot and a bar chart.

**Findings:**

* Time Series (New Cases by WHO Region):
* Displayed daily `New\_cases` over time, segmented by `WHO\_region`.
* Key observation: A major peak occurred on [Insert date] with [Insert value] cases, predominantly in [Insert region], marking a significant wave.
* Trend: [Insert observation, e.g., "Multiple surges in AMRO, with quieter periods in AFRO"].
* Bar Chart (Top 10 Countries by Cumulative Deaths):
* Highlighted the 10 countries with the highest `Cumulative\_deaths`.
* Top performer: [Insert country, e.g., "United States"] with [Insert value, e.g., "1,000,000 deaths"], followed by [Insert next country and value].

**Outcome:** Visuals effectively communicate temporal peaks and geographic hotspots, saved as images for inclusion in this report.

**8. Dashboard Design**

**Objective:** Plan a user-friendly layout for Power BI visualization.

**Process:** A dashboard structure was conceptualized before implementation.

Design:

* Key Performance Indicators (KPIs): Total New Cases, Total Cumulative Cases, Total Deaths—positioned at the top for quick reference.

Charts:

* Line chart for tracking cases over time.
* Bar chart for comparing deaths by country.
* Map for visualizing geographic spread.

**Navigation:** Filters for date range and country selection planned for interactivity.

**Significance:** The layout balances overview stats with detailed trends, catering to both high-level and granular analysis.

**9. Visualization Creation**

**Objective:** Build the planned visuals in Power BI.

**Process:** The cleaned dataset (`cleaned\_covid\_data.csv`) was imported into Power BI, and visuals were constructed.

**Components:**

* KPIs: Displayed as cards showing the sum of `New\_cases`, `Cumulative\_cases`, and `Cumulative\_deaths`.
* Line Chart: Plotted `New\_cases` over `Date\_reported`, revealing infection trends.
* Bar Chart: Ranked the top 10 countries by maximum `Cumulative\_deaths`.
* Map: Used `Country\_Code` to map `Cumulative\_cases`, with bubble sizes reflecting case totals.

**Outcome:** A visually cohesive dashboard that mirrors EDA findings, ready for interactive exploration.

**10. Data Integration**

**Objective:** Merge multiple data sources if necessary.

**Process:** Assessed the need for additional datasets but found the single Ministry of Health file sufficient.

**Findings:** No external sources were integrated, as the dataset already included comprehensive temporal, geographic, and metric data.

**Outcome:** Analysis proceeded with a unified, self-contained dataset, simplifying the workflow.

**11. Interactive Features**

**Objective:** Enhance the dashboard with dynamic, user-driven elements.

**Process:** Added interactive components in Power BI to improve usability.

**Features:**

* Date Range Filter: A slider for `Date\_reported`, allowing users to focus on specific time periods.
* Country Filter: A dropdown menu for `Country`, enabling country-specific views.
* Drill-Down: Implemented on the line chart, letting users zoom from yearly to monthly case trends.

**Outcome:** An interactive dashboard that empowers users to customize their analysis, increasing its practical utility.

**12. Insight Documentation**

**Objective:** Summarize key takeaways for decision-makers.

**Process:** Distilled EDA and visualization findings into concise insights.

**Conclusion**

This project transformed raw COVID-19 data into a robust analytical product through systematic EDA and visualization. Key deliverables include:

Jupyter Notebook: `COVID19\_EDA.ipynb`, containing full analysis and visualizations.

Power BI Dashboard: `COVID 19\_Dashboard.pbix`, offering an interactive interface.

The findings reveal significant regional disparities, temporal patterns, and predictive relationships, laying a foundation for informed decision-making in pandemic response.