Project 2

COVID-19 Data Analysis and Visualization

Ram Gupta

Takeo AI

Submitted by

Sonu Tamang

Date: Jun 26, 2025

Milestone 1: Data Collection and Cleaning

In the first phase of my project, I started by importing the raw COVID-19 dataset and preparing it for analysis.

- I cleaned the date column by converting it into proper datetime format so I could later use it for trend analysis.
- I standardized country names by removing extra spaces and converting them to title case. E.g. united states to United States.
- I checked and confirmed that there were no duplicate rows in the dataset.
- I handled missing values mostly from the Province/State column, which I chose to keep since it wasn't critical.
- I also fixed invalid values by setting any negative values in the Active cases column to 0, because negative case counts don't make sense.
- Finally, I saved this cleaned version into a new CSV file called covid_19_cleaned.csv, which I used in the next milestone.

Milestone 2: Exploratory Data Analysis (EDA)

This milestone helped me explore the cleaned data visually and statistically to uncover useful insights.

Task 1: Descriptive Statistics

I used .describe() to generate summary stats like mean, min, max, and standard deviation.

This gave me a quick understanding of the spread and typical values in columns like Confirmed, Deaths, Recovered, and Active.

Task 2: Data Visualization

- Line Chart

I created a line chart to show how total global confirmed cases increased over time. At first, the chart looked cluttered and hard to read because it included every single date. I fixed this by limiting the x-axis ticks (using every 20th date), which made the chart much cleaner and easier to understand.

- Bar Chart

I plotted the top 10 countries by total confirmed cases.

The chart revealed that countries like the United States, Brazil, and India had the highest counts.

- Histogram

I created a histogram to show the distribution of active case counts. Most regions had lower active case counts, but a few had extremely high numbers, which I saw as outliers. I also tested limiting the x-axis, but kept the full view since it was still technically correct.

- Pie Chart

I used a pie chart to visualize confirmed cases by WHO Region. The Americas had the highest share of global cases (around 48%), followed by Europe. This matched global reports and helped me see regional impacts.

Task 3: Correlation Analysis

I created a correlation heatmap to check relationships between Confirmed, Deaths, Recovered, and Active cases. I found that confirmed and active showed a strong positive correlation, as expected. Confirmed and Deaths were also highly related. Recovered had some variation, likely depending on healthcare and reporting standards in different regions.