Project Metadata

Title: *Pandemic Trends: A Visual Exploration of COVID-19 Data*

Team Members:

Name	**Email**	**Spire ID**
Hemangani Nagarajan	hemanganinag@umass.edu	34036775
Kavisha Parikh	kavishaprana@umass.edu	34733769
Varshini Venkataraman	vvenkatarama@umass.edu	34829219

GitHub: https://github.com/HEMANGANI/571-Project

Background & Motivation

The COVID-19 pandemic has been one of the most significant global health crises, affecting millions worldwide. Understanding how the virus spreads across different regions, its impact on mortality rates, and recovery patterns is crucial for gaining insights into pandemic dynamics and informing future public health responses.

Our motivation for this project stems from the opportunity to work with an extraordinarily vast and multimodal dataset that captures numerous dimensions of the pandemic. The COVID-19 data includes temporal, geographic, demographic, and epidemiological information that, when viewed as raw numbers alone, cannot reveal the complex interrelationships and patterns that defined the pandemic's progression. By transforming these complex, multimodal datasets into accessible, interactive visualizations, we aim to uncover hidden patterns, unexpected correlations, and subtle, invisible trends in tabular data formats.

Working with this rich dataset provides an exceptional learning opportunity to develop our data visualization skills while potentially revealing insights that could contribute to pandemic understanding. The visual exploration of this data can illuminate spatial-temporal patterns of virus spread, identify unexpected hotspots, and demonstrate the effectiveness of different regional responses.

The Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) has maintained comprehensive datasets tracking COVID-19 cases, deaths, and recoveries worldwide. By leveraging these datasets, we can create meaningful visualizations beyond simple counts and ratios to reveal the complex story of how the pandemic unfolded across different geographical regions and timeframes.

Project Objectives

The primary goal of our project is to develop interactive visualizations that provide insights into the spread and impact of COVID-19 across different regions and timeframes. Through these visualizations, we aim to answer the following key questions:

1. **How did COVID-19 spread across different regions over time?**
    - By visualizing the progression of confirmed cases, we can identify when and where major outbreaks occurred and how quickly the virus spread in different areas.
2. **What were the mortality and recovery patterns across different regions?**
    - By comparing death and recovery rates across different countries and states, we can identify severely affected regions and those managed the pandemic more effectively.
3. **When did the spread of COVID-19 slow down in different regions?**
    - By analyzing trends in case numbers over time, we can identify when various regions reached their peak infection rates and when the spread began to slow.
4. **How did different waves of infection compare across regions?**
    - By examining time-series data, we can identify and compare different waves of infection across regions.
5. **What are the relationships between confirmed cases, deaths, and recoveries?**
    - By analyzing correlations between these metrics, we can gain insights into the overall impact of the pandemic.

Data

JHU CSSE COVID-19 Dataset:
GitHub Repository: https://github.com/CSSEGISandData/COVID-19.git

- **Important Note:** The dataset was not updated after March 10, 2023 (3/10/2023), as JHU CSSE concluded their data collection efforts. Our analysis will cover the pandemic from its beginning until this cutoff date.

- Key datasets include:
    - time_series_covid19_confirmed_global.csv:
        - Contains daily cumulative confirmed COVID-19 cases for all countries outside the US.
    - time_series_covid19_deaths_global.csv:
        - Records cumulative death counts for each country/region over time, following the same structure as the confirmed cases file.
    - time_series_covid19_recovered_global.csv:
        - Contains recovery data by country/region over time, though it's worth noting that recovery reporting has been less consistent.
    - time_series_covid19_confirmed_US.csv:
        - Provides county-level data for the United States, with significantly more geographic granularity.
    - time_series_covid19_deaths_US.csv:
        - Contains US county-level death data and includes population data for each county, which enables per capita calculations.

Data Processing

Our data processing approach will involve several key steps to ensure that the data is clean, consistent, and ready for visualization:

1. **Data Extraction and Integration:**
    - We will extract relevant data from the JHU CSSE time-series files.
    - For global data, we will combine information from the confirmed cases, deaths, and recoveries datasets.
    - For US data, we will integrate state-level information from the confirmed cases and deaths datasets.
2. **Data Cleaning and Standardization:**
    - Handling missing values by imputing them based on surrounding data points or excluding them from the analysis.
    - Standardizing country and region names to ensure consistency across datasets.
    - Converting date formats to a standardized format for easier time-series analysis.
3. **Compute Quantities:**
    - **Case Rates:** Calculating per capita cases by normalizing confirmed cases by population.
    - **Death Rates:** (Total Deaths / Total Confirmed Cases) × 100.
    - **Recovery Rates:** (Total Recoveries / Total Confirmed Cases) × 100.
    - **Growth Rates:** day-over-day and week-over-week growth rates in confirmed cases.
    - **Doubling Times:** Computing how quickly cases doubled in different regions.
    - **Peak Identification:** Identifying when different regions reached peak infection rates.
    - **First Onset Dates:** Determining when COVID-19 first appeared in different regions.
4. **Time-Series Processing:**
    - Converting cumulative data to daily new cases, deaths, and recoveries.
    - Creating rolling averages (7-day, 14-day) to smooth out reporting inconsistencies.
    - Identifying trends and patterns in the time-series data.
5. **Categorical Aggregation:**
    - Aggregating data by geographical regions (countries, states, continents).
    - Creating time-based aggregations (monthly quarterly summaries).
    - Classifying regions based on severity metrics (case rates, death rates).

We will implement this using JavaScript and D3.js for the web-based visualization, with preprocessing done in Python to generate clean, structured datasets ready for visualization.

Visualization Design

Our visualization design will focus on creating interactive, informative, and user-friendly representations of COVID-19 data. Based on our brainstorming and design exploration, we have developed the following visualization components:

Brainstroming Ideas

more useful for visualizations

COVID19 Data

- General dataset
- Daily reports
- Time Series Data (Sub Datasets)
  - Confirmed
  - deaths
  - recovered
  - global & US

Filter

Datapoints
- Date
- Location
- Cases
- Deaths
- Recovery
- Population
- Vaccin
- Source

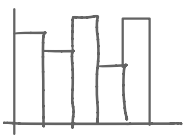
→ heat map
→ cases doubling
→ regions

Bit complex to understand ✗

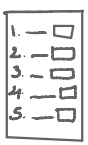
We can see in map view clearly

→ area map
→ growing bars

→ how covid progresses
→ add more features in next iterations ✓
→ confirmed cases old



→ top n regions
→ if cumbe cases, deaths or recovery
→ all a datasets are
→ changes according to different time periods ✓
→ location + other data used



→ alternative ideas to bar charts a list of countries
→ dynamically changes to time period
→ Region name [value]

→ But relative diff in have to notice with numbers ✗
→ graph is better

→ Region wise recovery / deaths percentage
→ can use these multiple regions
→ normalize to 100%
→ link it to bar graph ... ✓

→ definitely include map charts
→ easy to perceive the location
→ combine location and other datasets information
→ refined view ✓

Categorize

include trend lines area represents cases trend lins for deaths, recovery etc ...

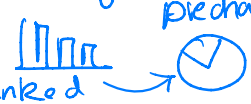
Trends in time series data

Analytical Visualization

Split into 2

Combine

clicking country → change pie chart

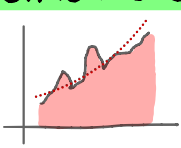


Geographic Visualization

Colour coded highlight hotspots

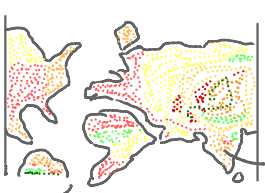
gradient to show spread and severity concentration

Combine & Refine



Trends Graph

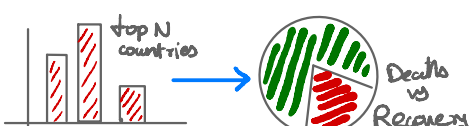
Different waves combe visualized how the trend change over time



Global Covid Spread

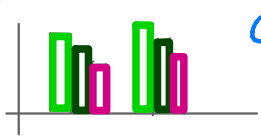
tool tip show detailed info

hot spots highlighted red



Linked charts

which region is more effective



Cases vs deaths vs recovery

Region wise

relationships in these variables

Focus.

Identification of highly affected areas.

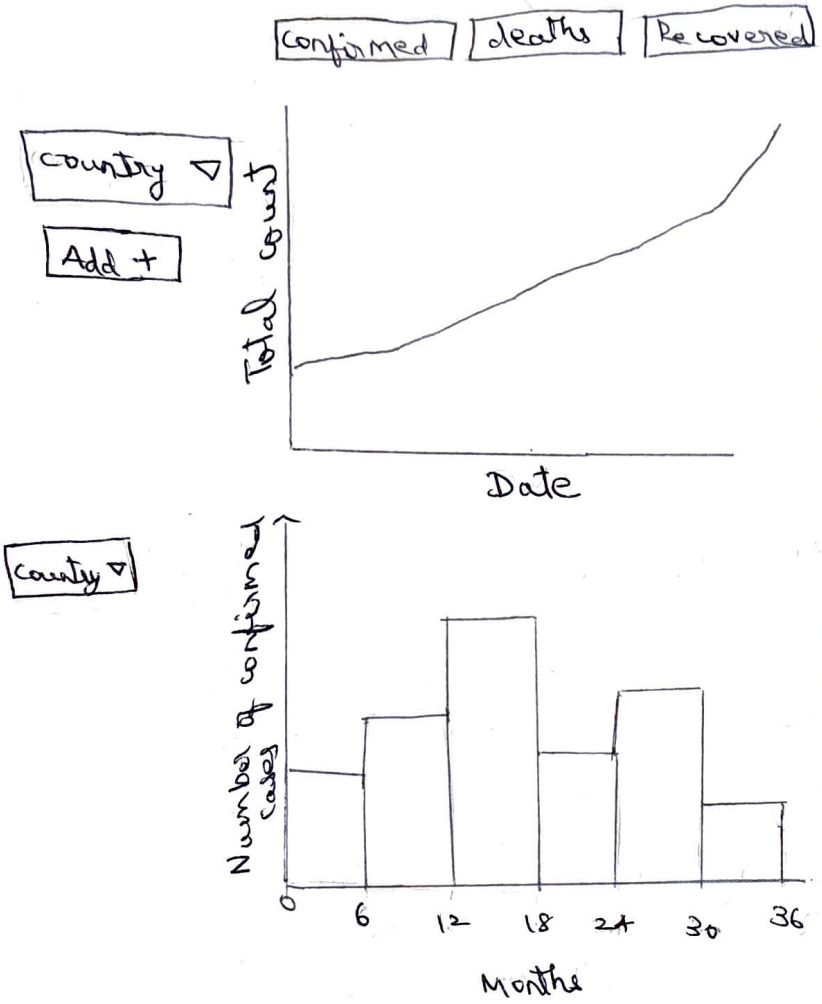
Time series analysis over different periods of time.

Metadata:

Title: Pandemic Trends: A visual Exploration of COVID-19 Data.

Sheet: 2

Task: Trend over different time periods



Operations

* The user can select country from dropdown for which he wants to see the graph.
* Add button: Adds a new country dropdown for comparisons (max 5).
* Confirmed, Deaths and recovered buttons add a new trend line for the country selected.
* Hover over a bar or line to get the exact value of that point.
* Hover over a country in the cluster column chart to get exact value of point.

Discussions

1) What are the advantages of having these graphs?
2) Do these accurately convey the trend?

Metadata:
Title:
Pandemic Trends: A Visual Exploration of COVID-19 Data
Sheet: 3
Task: Visualizing the Global Spread of COVID-19

Global COVID-19 Spread: Total Confirmed Cases, Deaths and Recoveries

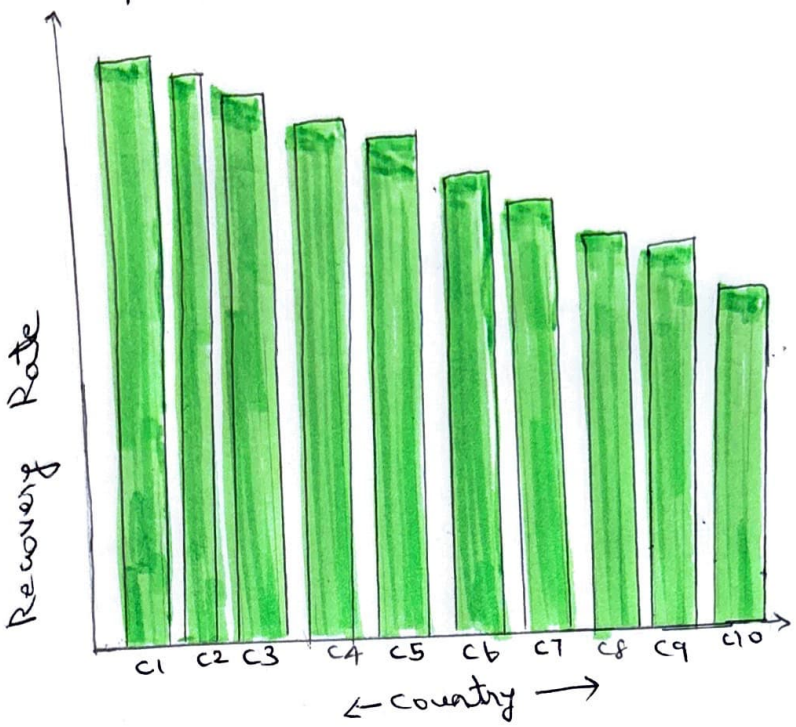
Operations:

* When a user hovers over a country, the country's name, the number of confirmed cases, deaths and recoveries are displayed.
* Additionally, the intensity of COVID-19 can be indicated using color coding: red for severe, yellow for moderate and green for safe.

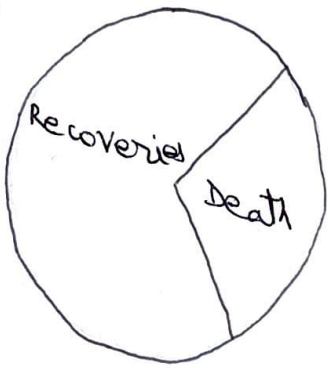
Discussions:

1) What type of graph, other than line and bar graphs, would better illustrate the spread of COVID-19 across the globe?
2) What kind of information should be included when the user hovers over the graph?
3) Would using color coding to represent confirmed cases be effective?

Top 10 countries with best recovery rate



Country ▼



Recoveries And Deaths

Metadata:

Title:

Pandemic Trends: A Visual Exploration of COVID-19 Data

Sheet: 4

Task: Displaying countries with best recovery rates

Operation:

* For the bar chart: When a user hovers over the bars, the value will be displayed.

* For the pie chart: The user can select the country they want to view the number of deaths or recoveries for.

Discussion:

1) How can we visualize the deaths and confirmed cases for all countries and will this be useful?

2) Should we display data about the top 10 countries with best recovery rates?

3) What are the advantages of these two charts?

Focus

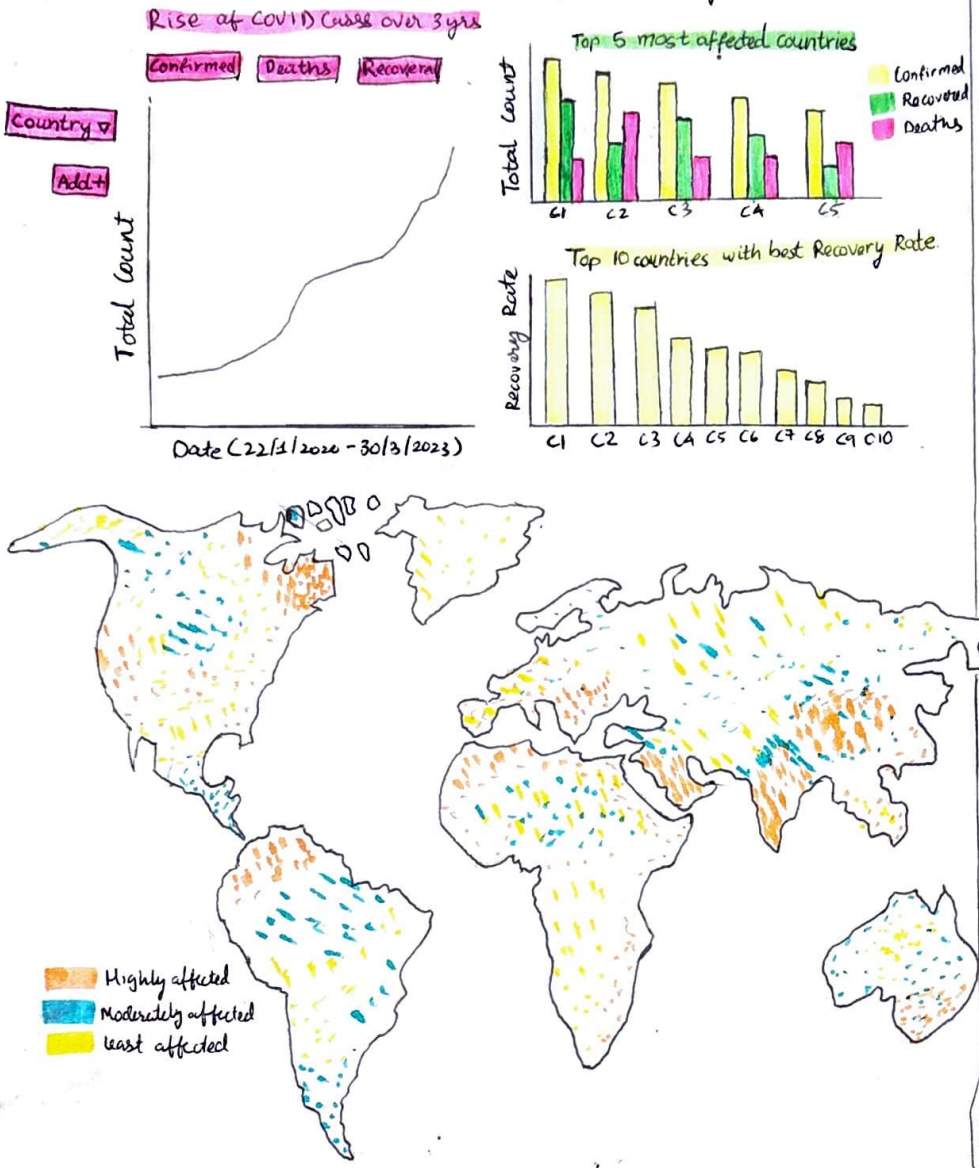
Dashboard Design, Chart selection, Extracting requirement and relevant Dataprocessing steps.

Meta Data

Title: Pandemic Trends: A visual Exploration of COVID-19 Data.

Names: Kavisha, Varshini, Hemangani.

Tasks: Dashboard Design, chart placement.



Operations

- Country Dropdown: User selects country from drop down for which he wants to see the graph.
- Add Button: Adds a new country dropdown for comparison (Max 5)
- Confirmed, Deaths and recoverd Buttons add a new trend line for the country selected.
- Hover over a bar or line to get the exact values of that point.
- Hover over a country in the chloropleth map to get exact number of confirmed cases, deaths and recoveries.

Detail.

Software Dependencies.

Frontend: D3.js, CSS, Bootstrap, HTML.

Backend: Excel, Python: Pandas, Numpy, Flask.

Calculations

Recovery Rate graph: Filtering countries with cases above a threshold and computing (total recoveries / total confirmed) * 100

6-month graph: Data is cumulative, so in order to get values for this graph, subtract every 6-month date column, eg, for bar chart, first bar value will be 2 July 2020 - Jan 22 2020

Pie chart calculation: Death and recovery ratio calculation relative to total outcomes.

Must-Have Features

The following features are essential for our COVID-19 data visualization project:

1. **Interactive Time Series Visualization**
    - Line chart showing progression of COVID-19 metrics over time
    - X-axis: dates
    - Y-axis: confirmed cases, deaths, or recoveries (user-selectable)
    - Users can select specific countries/regions to display
    - Visualization supports data from all five key datasets (global confirmed, global deaths, US confirmed, US deaths, global recovered)
2. **Global Choropleth Map**
    - Color-coded world map displaying either confirmed cases, deaths, or recoveries based on user selection
    - Toggle to view either total cumulative numbers or per capita rates
3. **Comparative Bar Chart by Time Period**
    - Bar chart showing progression over 6-month periods
    - X-axis: 6-month periods
    - Y-axis: selected metric (deaths, recoveries, or confirmed cases)
    - Users can select specific countries or view data for all countries
    - Clear visual distinction between time periods
4. **Top N Countries Clustered Column Chart**
    - Clustered bar chart showing the top N most affected countries
    - Three bars for each country representing total cases, deaths, and recoveries
    - Countries selected based on the highest total confirmed cases
    - Clear labeling and a consistent color scheme for each metric
5. **Recovery Rate Analysis Dashboard**
    - Visualization of top N countries based on recovery percentage
    - Adjustable threshold for minimum number of cases to be included
    - Clear presentation of recovery rate calculation methodology
    - Sorting functionality to arrange countries by the recovery rate
6. **Country-Specific Pie Chart**
    - Pie chart for the proportion of deaths and recoveries for a selected country
    - The total of deaths and recoveries represented as %
    - Clear labeling of each segment with percentages
    - Updates dynamically when the user selects different countries
7. **Basic Filtering and Selection Controls**
    - Country/region selector with search functionality
    - Metric selector (confirmed cases, deaths, recoveries)
    - Clear user interface elements for interacting with visualizations
    - Consistent layout and design across all visualizations

Optional Features

Our primary focus will be on implementing the must-haves, the following features would enhance the project if time permits:

1. **Enhanced Interactive Elements**
    - Hover functionality showing detailed statistics when hovering data points
    - Tooltips with comprehensive information about specific data points
    - Click interactions that update multiple visualizations simultaneously
    - Interactive legends that allow filtering by clicking on legend items
2. **Wave Analysis Visualization**
    - Identification and highlighting of pandemic waves across different regions
    - Comparison of wave intensity, duration, and timing between countries
    - Annotations indicating potential causal factors for different waves
3. **First Onset Visualization**
    - Visualization showing when COVID-19 first appeared in different regions
    - For US data: visualization of first onset dates by state
    - For global data: visualization of spread patterns from initial outbreaks
4. **Doubling Time Heatmap**
    - Color-coded grid showing how quickly cases doubled in different regions
    - Ability to track changes in doubling time as the pandemic progressed
    - Helps identify where the spread was accelerating or decelerating
5. **Mobile-Optimized Interface**
    - Responsive design tailored explicitly for smartphone and tablet viewing
    - Touch-friendly controls for all interactive elements
    - Simplified views for smaller screens
6. **Data Export Functionality**
    - Option to download visualized data in CSV or JSON format
    - Ability to save or share specific visualization states
    - Export high-resolution images of visualizations
7. **Informational Overlays**
    - Optional educational content explaining epidemiological concepts
    - Contextual information about major pandemic events
    - Methodology explanations for calculations like recovery rates

Project Schedule

Our project timeline spans 10 weeks with key milestones as follows:

Week 1 (March 3-9)
- Setup GitHub repository and project structure
- Explore and understand the COVID-19 datasets
- Initial data exploration and preprocessing

Week 2 (March 10-16)
- Implement core data processing functions
- Set up the D3.js framework
- Begin initial visualization components

Week 3 (March 17-23)
- Implement time series, choropleth map, and bar chart visualizations
- Create a basic user interface for visualization selection

Week 4 (March 24-30)
- Develop remaining visualizations (top countries, recovery analysis, pie chart)
- Integrate visualizations into a cohesive interface

Week 5 (March 31-April 6)
- Project milestone preparation and presentation
- Implement feedback from milestone review

Week 6 (April 7-13)
- Refine visualizations based on feedback
- Performance optimization and testing
- Begin implementing optional features

Week 7 (April 14-20)
- Continue optional feature implementation
- User testing and refinement

Week 8 (April 21-27)
- Final UI polish and optimization
- Complete documentation

Week 9 (April 28-May 4)
- Create project screencast
- Final testing and quality assurance

Week 10 (May 5-11)
- Final submission
- Group feedback and retrospective

Task Distribution:

- Hemangani Nagarajan
    - Data Processing and Time Series Visualization
- Kavisha Parikh
    - Geographic Visualizations and Integration
- Varshini Venkataraman
    - Analytics Visualizations and Testing