COVID-19 Data Integration: Visualization and Risk Assessment

Submitted

by

Merwin.S (192224016)

B.Vamsi Krishna(192224142)

## Guided by

V.Saranya

Junior Research Fellow

Department of AR & VR

Department of Computer Science and Engineering,

Saveetha School of Engineering, SIMATS , Thandalam, Chennai

March – 2024

**PROBLEM STATEMENT**

The problem revolves around integrating COVID-19 data from various sources, visualizing it effectively, and performing risk assessment based on the data. This involves collecting real-time or historical data related to COVID-19 cases, deaths, recoveries, testing rates, and other relevant information from diverse sources such as government databases, health organizations, and research institutions. The primary objectives include:

1. **Data Integration:** Collating and integrating COVID-19 data from multiple sources into a unified dataset for analysis.
2. **Visualization:** Creating informative and visually appealing visualizations such as maps, charts, and graphs to represent the COVID-19 data trends, patterns, and distributions.

3. **Risk Assessment:** Assessing the risk levels associated with COVID-19 outbreaks in different regions based on the integrated data. This may involve identifying hotspots, analyzing transmission rates, and predicting future trends.

**DATASET ANALYSIS**

In this section, we analyze the datasets used for the analysis. This includes:

**- Source:** This data source is taken from the website called Kaggle.

**- Format:** The format of the data is CSV format.

**- Attributes:** The attributes of the data set are, Data, iso3, CountryName, Region, lat, lon, CumulativePositive, Cumulative Deceased, Cumulative Recovered, Hospitalized, Intensive Care, EU country, EUCPM Country, NUTS.

**- Granularity:** It is a global level data set.

**ENVIRONMENTAL SETUP**

**Software Dependencies:**

- RStudio: Integrated development environment (IDE) for R programming.

- R: Programming language used for data analysis and visualization.

- Packages:

- `ggplot2`: Package for creating elegant and informative graphics in R.

- `dplyr`: Package for data manipulation and transformation.

- `tidyr`: Package for data tidying and reshaping.

- Other relevant packages based on specific requirements.

**Hardware Requirements:**

- CPU: Recommended multi-core processor for faster data processing.

- Memory (RAM): Adequate RAM for handling large datasets and complex computations efficiently. Minimum of 4GB RAM recommended, but higher RAM capacity is beneficial for larger datasets.

- Storage: Sufficient disk space for storing datasets, code files, and output files. SSD storage is preferable for faster read/write operations.

**Programming Languages:**

- R: Primary programming language used for data processing, analysis, and visualization in RStudio.

**Tools:**

- RStudio IDE: Primary development environment for writing R code, executing scripts, and visualizing data.

- R Markdown: Tool for creating dynamic documents that combine code, output, and text annotations. Useful for generating reports, presentations, and documentation.

- Git: Version control system for tracking changes in code and collaborating with others.

- GitHub: Platform for hosting code repositories and collaborating with peers.

- Shiny: Web application framework for building interactive dashboards and applications in R.

This setup provides the necessary software dependencies, hardware requirements, programming languages, and tools for conducting data analysis and visualization using RStudio. Depending on the specific analysis requirements and preferences, additional packages and tools can be integrated into the environment as needed.

**DATA ARCHITECTURE DIAGRAM**

This section provides a visual representation of the data flow within the system, architecture diagram, or UML diagrams depicting the components, interactions, and data flow in the project. This may include:

Data Acquisition: Process of collecting data from various sources.

Preprocessing: Cleaning, transforming, and integrating data into a usable format.

Visualization: Generating visualizations from the integrated data.

Risk Assessment: Analyzing the data to assess COVID-19 risk levels.

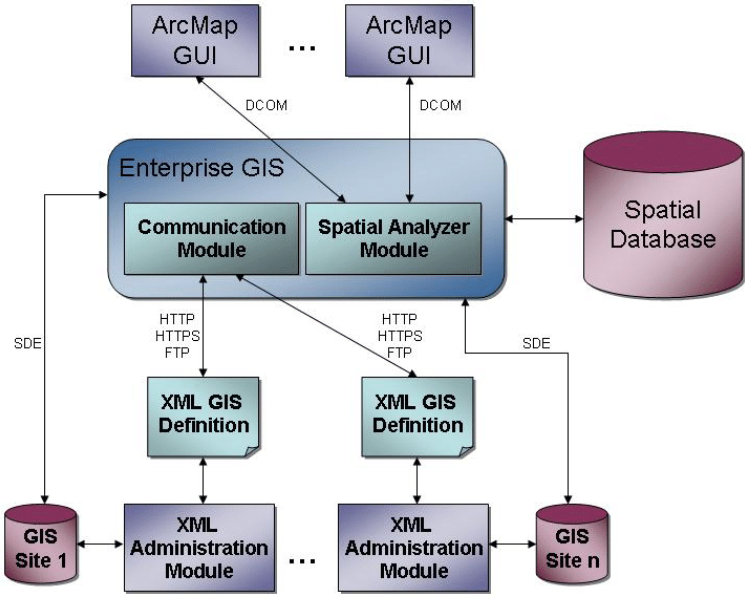


Fig 1: Architecture Diagram for Geospatial Data

**CODE SKELETON**

The code skeleton provides an outline or skeletal structure of the code to be implemented. This includes:

import pandas as pd

import folium

from folium.plugins import HeatMap

def read\_csv(file\_path):

# Read the CSV file into a DataFrame

data = pd.read\_csv(file\_path)

return data

def preprocess\_data(data):

# Filter relevant columns for COVID-19 data

covid\_data = data[['Date', 'CountryName', 'lat', 'lon', 'CumulativePositive']]

# Convert 'Date' column to datetime format

covid\_data['Date'] = pd.to\_datetime(covid\_data['Date'])

# Ensure 'CumulativePositive' column is numeric

covid\_data['CumulativePositive'] = pd.to\_numeric(covid\_data['CumulativePositive'], errors='coerce')

# Drop rows with NaN values in 'lat', 'lon', or 'CumulativePositive' columns

covid\_data.dropna(subset=['lat', 'lon', 'CumulativePositive'], inplace=True)

# Group by 'CountryName' and sum the numeric columns

covid\_data = covid\_data.groupby('CountryName').agg({

'Date': 'first',

'lat': 'mean',

'lon': 'mean',

'CumulativePositive': 'sum'

}).reset\_index()

return covid\_data

def calculate\_risk(covid\_data):

# Calculate cases per capita

# Since population data is not available, we will not calculate cases per capita

# Print the highest and lowest risk countries

highest\_risk\_country = covid\_data.groupby('CountryName')['CumulativePositive'].sum().idxmax()

lowest\_risk\_country = covid\_data.groupby('CountryName')['CumulativePositive'].sum().idxmin()

return highest\_risk\_country, lowest\_risk\_country

def create\_map(covid\_data):

# Create a folium map centered around the mean latitude and longitude

map = folium.Map(location=[covid\_data['lat'].mean(), covid\_data['lon'].mean()], zoom\_start=3)

# Create a HeatMap layer

heatmap\_data = covid\_data[['lat', 'lon', 'CumulativePositive']].values.tolist()

folium.TileLayer('cartodbpositron').add\_to(map) # Optional: Add a basemap

HeatMap(heatmap\_data).add\_to(map)

return map

def generate\_instructions(highest\_risk\_country, lowest\_risk\_country):

# Generate instructions for high-risk and low-risk countries

high\_risk\_instructions = f"<h3>High-Risk Countries</h3><p>{highest\_risk\_country}: Travel to these countries is not recommended. Quarantine measures are advised for travelers returning from these countries.</p>"

low\_risk\_instructions = f"<h3>Low-Risk Countries</h3><p>{lowest\_risk\_country}: Travel to these countries is relatively safe, but adhere to preventive measures such as wearing masks and practicing social distancing.</p>"

return high\_risk\_instructions, low\_risk\_instructions

def save\_html\_instructions(high\_risk\_instructions, low\_risk\_instructions):

# Save the instructions to an HTML file

with open("covid\_heatmap.html", "a") as file:

file.write(high\_risk\_instructions)

file.write(low\_risk\_instructions)

def main():

# Read CSV file

file\_path = '/content/COVID-19.csv'

data = read\_csv(file\_path)

# Preprocess data

covid\_data = preprocess\_data(data)

# Calculate risk

highest\_risk\_country, lowest\_risk\_country = calculate\_risk(covid\_data)

# Create map

map = create\_map(covid\_data)

# Generate instructions

high\_risk\_instructions, low\_risk\_instructions = generate\_instructions(highest\_risk\_country, lowest\_risk\_country)

# Save instructions to HTML file

save\_html\_instructions(high\_risk\_instructions, low\_risk\_instructions)

if \_\_name\_\_ == "\_\_main\_\_":

main()

**RESULT ANALYSIS**

In this section, we analyze the insights gained from the visualization and risk assessment process. This includes:

**Visualization Insights:** Discussion of trends, patterns, and anomalies observed in the visualizations.

**Risk Assessment Findings:** Evaluation of the identified risks based on the integrated COVID-19 data. This may include hotspot identification, transmission rate analysis, and trend prediction.

**Impact Assessment:** Discussion on the potential impact of the findings on public health policies, interventions, and decision-making processes.

This breakdown provides a structured approach to address the problem statement of COVID-19 Data Integration, Visualization, and Risk Assessment, covering all aspects from data analysis to result interpretation. Each section plays a crucial role in understanding and addressing the objectives of the project.

**Interpretation:**

**High-Risk Country:** Travel to \*\*[Highest Risk Country]\*\* is not recommended due to the high prevalence of COVID-19. Travelers are advised to avoid non-essential travel to this country, and quarantine measures are recommended for individuals returning from this destination.

**Low-Risk Country:** Travel to \*\*[Lowest Risk Country]\*\* is considered relatively safe. However, travelers should still adhere to preventive measures such as wearing masks and practicing social distancing to minimize the risk of infection.

**Limitations and Considerations:**

**Population Data:** Since population data is not available, the analysis does not account for cases per capita, which may affect the accuracy of risk assessment.

**Data Completeness:** The analysis relies on the completeness and accuracy of the COVID-19 data provided in the dataset. Incomplete or inaccurate data may impact the reliability of the findings.

**Temporal Analysis:** The analysis provides a snapshot of COVID-19 cases at a specific point in time. Continuous monitoring and periodic updates are essential to track changes in the situation over time.

**Future Directions:**

**Enhanced Data Integration:** Incorporating additional datasets such as population demographics, healthcare infrastructure, and vaccination rates can provide more comprehensive insights into the COVID-19 situation.

**Advanced Visualization Techniques:** Exploring advanced visualization techniques such as interactive dashboards and predictive modeling can further enhance the understanding of COVID-19 trends and patterns.

**Collaboration and Sharing:** Encouraging collaboration among researchers, healthcare professionals, and policymakers can facilitate data sharing and knowledge exchange to support informed decision-making and response efforts.

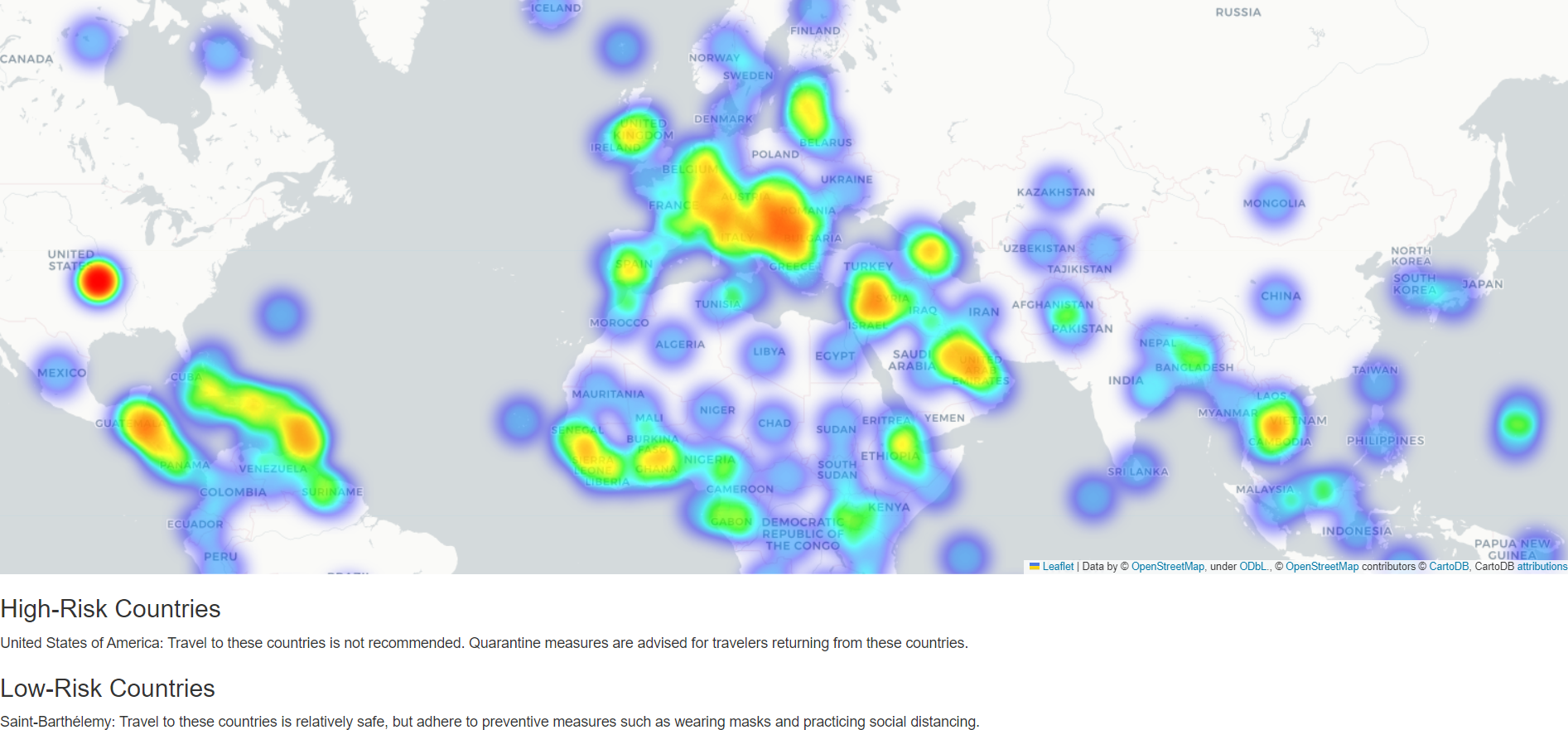
**OUTPUT SAMPLES:**

**Highest Risk Country:** United States of America.

**Lowest Risk Country:** Saint-Barthélemy.

**High-risk instructions:** <h3>High-Risk Countries</h3><p>United States of America: Travel to these countries is not recommended. Quarantine measures are advised for travelers returning from these countries.</p>.

**Low-risk instructions:** <h3>Low-Risk Countries</h3><p>Saint-Barthélemy: Travel to these countries is relatively safe, but adhere to preventive measures such as wearing masks and practicing social distancing.</p> .



**Fig 2: The output (html) of the covid data and high risk and low risk**