Project Deliverable 1

Pandemic Insights: A Comprehensive COVID-19 Retrospective Analysis Tool

Group 7

UF UNIVERSITY *of* FLORIDA

Members

Vivek Milind Aher	vaher@ufl.edu
Dhruv R Makwana	dmakwana@ufl.edu
Sakshi Pandey	sakshi.pandey@ufl.edu
Vishal Prakash	vishalprakash@ufl.edu

Overview

The Coronavirus pandemic is the most significant public health crisis humans have faced since the 1918 influenza pandemic. It affected multiple aspects of human life, and this application aims to explore the factors and circumstances that led to the COVID-19 pandemic unfolding the way it did. The primary objective of this COVID-19 application is to empower users, including healthcare professionals, policymakers, environmental scientists, and hospitals, with the capability to conduct comprehensive retrospective analyses of the pandemic. This analysis aims to extract valuable insights to inform and enhance future practices and strategies.

Description

Considering the widespread impact of COVID-19, there are numerous stakeholders, and this application is being designed to address the needs of some of the most influential stakeholders, such as Doctors, Policy Makers, and Researchers who would benefit from further analysis of the different facets of the pandemic. These users are in positions of influence and can implement lessons learned from the pandemic to change future policies. Our application provides them with the tools to do so.

Our application will allow these users to perform computations and statistical tests on several interdisciplinary datasets and visualize these results so that they are more intuitive. These visualizations and analyses can serve as valuable resources for ongoing research and provide substantial support for driving future policy changes and improvements. The application will have an interactive interface where the user can select multiple parameters and time frames to get useful visualizations, which can be used to identify important and interesting data trends.

Motivation of the Database needs and the Potential User Interest in the Application

The primary users of this application will be Doctors, Policy makers, Environmental Scientists, Researchers, and Epidemiologists. This database will aim to provide different views of the COVID-19 pandemic, providing perspective on comorbidities and their effect on mortality, how the allocation of resources towards the COVID-19 pandemic may have contributed to further loss of life due to negligence in the treatment of other ailments, and how socioeconomics and race may have been a factor in the high mortality rates during the pandemic. This database application also aims to explore environmental changes caused during the COVID-19 pandemic. This application is designed to play a pivotal role in advancing policies and healthcare practices for future pandemics. Additionally, it seeks to leverage the lessons learned from the current pandemic to introduce innovative and effective strategies in the fields of environmental research and epidemiology.

Given the diverse user base, including doctors, policy makers, environmental scientists, researchers, and epidemiologists, the application will generate and require access to a vast amount of data. A well-structured database will efficiently manage this data, ensuring it's organized, secure, and easily retrievable. Data from multiple sources and disciplines will be necessary to provide comprehensive views of the COVID-19 pandemic and its various facets. A database can integrate and harmonize diverse datasets, allowing for cross-disciplinary analysis.

As the pandemic evolves and new data becomes available, a database can facilitate real-time updates, ensuring users can access the most current information for informed decision-making.

All users (doctors, epidemiologists, and policymakers) can benefit from complex trend queries to find correlations and patterns relating to the COVID-19 epidemic. Complex queries can delve into socioeconomic and racial factors influencing pandemic outcomes or explore shifts in AQI and its implications on the environment. By using such queries, we can analyze resource allocation trends, and the application can thus be used to highlight instances where the focus on COVID-19 may have inadvertently led to neglect in treating other health conditions.

In summary, database support and the capability to perform complex trend queries are vital for this application's success. They enable efficient data management, integration of diverse datasets, and in-depth analysis, empowering users to make informed decisions and advance policies and practices for future pandemics and related fields like environmental research and epidemiology.

Needed Web-based User Interface Functionality:

This project will aim to provide a comprehensive and user-friendly web-based platform for analyzing COVID-19 trends. The proposed project will aim to facilitate sophisticated functionalities that would help analyze and visualize the trends spanning the pandemic. Some of the features of the proposed project would be:

1. **User Authentication and Authorization:**

   Users will have a particular User Role assigned to them like doctor, researcher, admin, or guest. The data analysis results will be accessible only to the users who have the required permission. Implementing Authentication and Authorization helps in maintaining data integrity and aids in prevention against insider threats.

2. **Data Visualization:**

   Users will be able to visualize the trends of distinct factors that affected COVID-19 infections or got affected due to COVD-19 during the pandemic through interactive graphs and charts. Users will also have the ability to select and group the time periods/time frames for which the user wants to analyse the data. The visualization of these trends will give deeper insights into the impact of COVID-19. The project will include statistical methods and visualizations such as Box Plots, Histograms and Normal Distributions. These will provide a clear representation of data to aid the users in identification of patterns and aid them in using the insights they gain to make data-driven decisions.

   The project will also have a raw data table so the users can have a look at the data columns that would help them gain further contextual information. Displaying raw data ensures the transparency and quality of the data. It will also help users to perform custom analyses from the data table.

3. **Search, Filtering and Sorting:**

   The project will have a separate webpage for displaying the raw data table. The table will also have search, filtering, and sorting functionality so that the users can perform basic analysis from the user interface. Having these functionalities helps users to efficiently retrieve the data, perform data validation and view customization. These functionalities would aid users in retrieving specific data from a specific time frame to perform the analyses they require.

4. **Data Export:**

   The user will be able to export the data in csv format after performing further functionalities like searching and filtering. Exporting the data helps users integrate the endpoints of these projects into other data-visualization software like Tableau, Microsoft PowerBI etc. The export functionality enhances the usability, flexibility, and accessibility of the project, empowering users to leverage the data and insights generated within the system for various purposes, from research and decision-making to reporting and collaboration.

Five colloquial complex trend queries and their explanations:

Query 1: How many people died due to other causes during the pandemic time period (**cardiovascular death rate, other infectious disease deathrate**)? How many COVID-19 infections were successfully treated and how many COVID-19 patients died?

Description: The means of the successfully treated infections, COVID-19 mortalities, and the mean of deaths caused due to cardiovascular reasons or car accidents can be compared using ANOVA test for multiple samples.

Goal: Could hospital resources being allocated more towards the pandemic have led to a higher overall mortality rate due to other causes?

Interpretation: Let's say that a hospital such as Cedars-Sinai wants to know whether its resources were allocated efficiently during the pandemic so they can better prepare for future pandemics. They choose to analyse the cardiovascular death rate against the COVID-19 mortality rate and the COVID-19 cases successfully treated. The graph plotted may show a time period from 2018 to 2021(The height of the pandemic being in 2020).

If the data shows that the mean number of deaths in this time period due to cardiovascular causes were:

2018	**2019**	**2020**	**2021**
10,458	9,875	27,893	24,566

And shows the mean survival rates for COVID-19 cases to be:

2018	**2019**	**2020**	**2021**
0	7,567	59,847	87,533

And shows the mean mortality rates for COVID-19 cases to be:

2018	**2019**	**2020**	**2021**
0	2,349	14,892	10,566

Then using a statistical test such as One-way ANOVA we can clearly compare the means and deduce that the Cardiovascular death rate increased dramatically during the pandemic, and more patients were lost due to cardiovascular causes than were lost due to COVID-19 or saved due to the extensive resources given to the COVID-19 treatment. Or we can come to the conclusion that there is no difference between the three groups and therefore resource allocation was optimal during the pandemic and the same process should be followed in the event of another pandemic.

Query 2: What was the mortality rate of different races over time during the COVID-19 pandemic? For each region, which race had the highest mortality rate? How many people from that race had health insurance in each region vs. How many people from that race had no health insurance in each region?

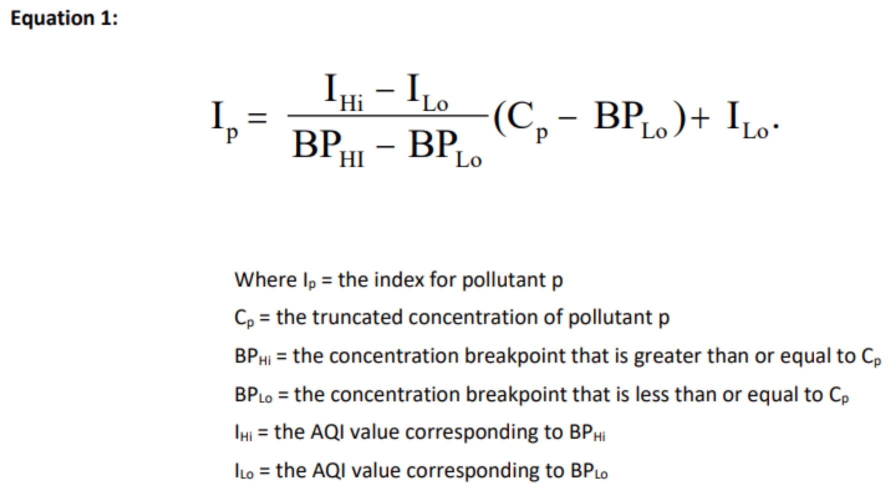
*region (northeast, midwest, south, and west)

Goal: Compare the race with the highest mortality vs. The race with no health insurance in each region.

Description: Find the mean mortality rate for each race (Hispanic, Asian, etc.) over a certain time period and then find which race had the highest mortality rate in each region. Then calculate the people from that race which had health insurance vs. No health insurance for each region over the same time period.

Interpretation: The health insurance rates were low for a particular region. Meaning that might be an area with low economic privilege and thus people without insurance may have sought treatment for COVID-19 too late. This data would help policy makers identify people in need earlier and design policies to aid and educate such people in the event of another pandemic.

Query 3: Find the mean AQI for each pollutant over a period of time for each state. Following this compare the AQI data points (for a year prior to COVID-19 with a year during the pandemic, ex: 2018 vs. 2020) with the Mann-Whitney U-test to determine whether the points are significantly different or not?



Equation 1:

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{HI} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo}.$$

Where I_p = the index for pollutant p

C_p = the truncated concentration of pollutant p

BP_{Hi} = the concentration breakpoint that is greater than or equal to C_p

BP_{Lo} = the concentration breakpoint that is less than or equal to C_p

I_{Hi} = the AQI value corresponding to BP_{Hi}

I_{Lo} = the AQI value corresponding to BP_{Lo}

Formula to calculate AQI.

Goal: Compare pollution rates before, during, and after the pandemic.

Description: This would involve finding the mean monthly AQI of all pollutants in the dataset for a period of time(2 years time) and then using the Mann-Whitney U-test to compare these data points to identify whether the pollution rates were higher or lower compared to the previous year.

Interpretation: This data could be helpful for environmental scientists to better identify pollutants to further study and understand why they were drastically reduced during the pandemic. This could also aid in the development of new policies to target and reduce these pollutants.

Query 4: Over time which condition group (ex. Circulatory diseases, Respiratory diseases), in each state, led to how many mean mortalities per month? Compare these against on another using the One-Way ANOVA test and determine what is the difference in mortality with reference to the other condition afflicting the patients?

Goal: To determine which conditions or comorbidities lead to the highest fatality rates following infection with COVID-19.

Description: This would involve finding the number of deaths for each disease/condition group and then carrying out the ANOVA test to find the difference in mortality with reference to the other condition afflicting the patients.

Interpretation: This would help healthcare professionals understand which conditions or diseases were exacerbated following infection by the COVID-19 virus and thus lead to mortality. This could also be used by researchers in healthcare to identify which diseases may contribute to fatality during a pandemic involving respiratory illnesses.

Query 5: What was the vaccination rate vs. infection rate over time? In the mortalities amongst people who were vaccinated did they have comorbidities? How many people who were infected and tested positive had comorbidities?

Goal: Find the efficacy of the vaccine and identify mortalities and infections following vaccination.

Description: Use a two-sample t-test to compare vaccination vs infection rate. Get the people who were vaccinated and following this got COVID-19 and passed away. Similarly, get the people who were vaccinated and then tested positive for the virus.

Interpretation: Help healthcare professionals develop a better understanding of the efficacy of vaccinations.

Description of the application goals regarding trend analysis

We have the following goals regarding trend analysis with respect to various queries:

1. Hospital and healthcare facilities had to allocate a significant portion of their resources to manage the surge in COVID-19 cases. Did this diversion of hospital resources towards the pandemic lead to increase in the overall mortality rate for other diseases?
2. Certain ethnic and racial groups experienced disproportionate impacts during the pandemic. To review this, we aim to find the race with the highest mortality and compare it with the race with no health insurance in each region.
3. Pollution is something the world has been striving to arrest and one positive aspect about COVID-19 was the reduction in pollution levels. To better understand this, we compare the pollution rates before, during, and after the pandemic by calculating the Air Quality Index of five different pollutants (SO2, NO2, Ozone, Pb, CO).
4. Comorbidities significantly influenced the mortality rates associated with COVID-19 and we would identify which comorbidities lead to the highest fatality rates following infection with COVID-19.
5. One important milestone in arresting the spread of COVID-19 was the introduction of the first vaccine. To check how effective this was, we would find the efficacy of the vaccine and identify mortalities and infections following vaccination.

Description of the Real-Word Data Forming the Basis of the Application and the Complex Trend Queries

Our project will use a comprehensive COVID-19 dataset with daily statistics and other interdisciplinary datasets encompassing a variety of critical metrics such as positive cases, deaths, comorbidities contributing to fatalities, vaccination rate, and air-quality index. The datasets are readily accessible, considerably maintained and are from recognized and trusted sources, hence highly dependable and accurate. This comprehensive dataset comprises over a hundred million tuples over various tables.

The datasets and their use cases are described below:

1. COVID-19 Dataset: This is a temporal dataset of the total number of COVID-19 cases in the United States. It contains information regarding the county, state, age, sex, and race of the people who contracted COVID-19. Every query revolves around this dataset.
   Source: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

2. Health Insurance Datasets: This source provides yearly datasets on the health insurance coverage of US citizens, categorized by age, sex, region, household income, education, work experience, health status, and other important datapoints. This dataset can be used to find the region-wise and race-wise comparison of COVID-19 related deaths between insured and uninsured people in the US.
   Source: https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hi/hi.html

3. Comorbidity Dataset: This is a dataset of comorbidities that contributed to COVID-19 deaths. It is categorized by age and state. This dataset can be used to analyze the effects of various other conditions on mortality rates following COVID-19 infection.

   Source: https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm

4. Vaccination Dataset: This is a temporal dataset of the county-wise data of vaccination rates in the United States. It contains information such as the percentage of people vaccinated, the type of vaccination administered, booster doses, and also the number of shots administered. This dataset can be used to find trends such as the decrease in the tendency of COVID-19 infections, and also the variation in the fatality of COVID-19 infections in vaccinated people compared to unvaccinated people.
   Source: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data

5. AQI Dataset: This is a temporal and demographic dataset of the concentrations of pollutants Ozone, CO, Pb, NO_2, SO_2, PM2.5, and PM10 from which we can calculate the Air Quality trends and analyze the environmental effects of the virus in different regions during the COVID-19 pandemic.
   Sources: https://www.epa.gov/outdoor-air-quality-data/download-daily-data
   https://www.epa.gov/outdoor-air-quality-data/air-data-daily-air-quality-tracker

6. Mortality Datasets: This source provides multiple temporal datasets on mortalities caused by various factors, such as cardiovascular diseases, other infectious diseases, and other such conditions. We can use these data sets to determine the number of deaths during the COVID-19 pandemic caused in conjunction with other illnesses or conditions.

Source: https://wonder.cdc.gov/

The aforementioned datasets are sufficient to perform the trend queries described in the trends section.

Intended use of public domain and/or proprietary software:

The major components of the project are the Frontend, Backend, and the Database. Frontend refers to all the web pages that we see on the web browser. Backend provides functionalities like fetching the data from database and server-side processing of user request. And DBMS is the most important component of the project because it deals with storing and managing the data and fetching the filtered data from database. The framework/libraries that we will be using for this project are:

Frontend:

HTML: It is a markup language which defines the basic structure of the Web page.

CSS: CSS is used to style and format the web-interface.

JavaScript: JavaScript provides interactivity to the web-interface.

React: React is a JavaScript library that uses components-based architecture to support fast rendering of the web page.

D3.js: D3.js is also a JavaScript library that is widely used to display interactive and dynamic graphs and charts on a web pages.

TailwindCSS: Tailwind CSS is used to rapidly design and style the project's web interface, ensuring an efficient and visually appealing layout.

Backend:

Flask: Flask serves as the back-end framework, facilitating server-side logic and data processing for the COVID-19 analysis application.

Database:

Oracle Database: It is an Industrial Relational Database System that is used by leading tech companies to store structured data. We will use it to store our data and dynamically fetching filtered data using complex trend queries.