Database Project Proposal

DTSC691: Applied Data Science
MUSUNGU KERRYL

Project Overview

Project Goals

Purpose : As part of our study, we shall be conducting a thorough analysis of the Covid-19 datasets. Our objective is to identify key details that are pertinent to our research topic. To achieve this, we shall scrutinize the data with utmost diligence and apply statistical methods where necessary to extract meaningful insights. The findings of this study will be instrumental in informing our research and contributing to the wider body of knowledge on this subject.

Project Focus : The aim of this project is to carefully monitor several critical consequences or variable outcomes related to the COVID-19 pandemic over a specific period, from January 22, 2020, to July 27, 2020. The variables in question include the number of cases, mortality rate, recovery rate, and others, in different countries and continents. The objective of this project is to analyze the effects of these variables, rather than their root causes.

Specific Goals : This research project will be dedicated to examining the impact of COVID-19 in various regions, as designated by the World Health Organization (WHO). The anticipated outcomes of this project will be a comprehensive list of questions that we intend to address to provide greater clarity and insight into the effects of COVID-19 in these regions. Our primary objective is to conduct a thorough and rigorous investigation, which will enable us to generate a detailed report on the key issues related to this pandemic. We are committed to delivering a well-researched and insightful report that will serve as a valuable resource for business and academic professionals alike. How the COVID-19 pandemic has changed over time in each geographical area (globe, continent, and WHO region). We shall endeavor to provide responses to the following primary inquiries later:

- The number of cases of COVID-19 per geographical area (country, continent, and WHO region).
- What is the population's overall death rate by geographic area (country, continent, and WHO region)?
- The number of individuals who returned home after recovering from the COvid-19 in each country, continent, and WHO area
- Assessing the efficacy of treatment modalities by geographical area (nation, continent, and WHO region)
- Analyzing the differences in the efficacy of COVID-19 treatment between the US and other nations
- Where was the COVID-19 pandemic least disruptive? (Nationality, continent, and WHO area)
- Where was the COVID-19 epidemic most prevalent? (nation, region, and WHO)

Project Description

Problem Domain : As our problem domain, we will examine the problem's demographic component. Stated differently, our primary focus will be on altering the composition of human populations as a result of the COVID-19 pandemic.

Database Design and Assumptions : We are going to give a brief description of different tables and the relationship

1. **Country_wise**

Columns: Country/Region, Confirmed, Deaths, Recovered, Active, New cases, New deaths, New recovered, Deaths/100 cases, Recovered/100 cases, Deaths/100 recovered, WHO Regions.

Shape: (187, 15)

We can obtain information from this database about the number of Covid-19 cases by nation and WHO region. We can ascertain the quantity of ongoing cases, fatalities, and recoveries in every nation.

2. **covid_19_clan_complete**

Columns: 'Country/Region', 'Lat', 'Long', 'Date', 'Confirmed', 'Deaths', 'Recovered', 'Active', 'WHO Region'

Shape: (49068, 10)

Here is data on the amount of persons who have perished, recovered, or are currently experiencing the COVID-19 pandemic on a daily basis across all countries. Additionally, the geolocation data is included in the Latitude and Longitude columns.

3. **day_wise**

Columns: 'Date', 'Confirmed', 'Deaths', 'Recovered', 'Active', 'New cases', 'New deaths', 'New recovered', 'Deaths / 100 Cases', 'Recovered / 100 Cases', 'Deaths / 100 Recovered', 'No. of countries'.

Shape: (188,12)

Once more, we know how many cases, fatalities, recoveries, or active cases there are. We also have the total number of countries where the COVID-19 pandemic was the cause of these cases.

4. **full_grouped**

columns: 'Date', 'Country/Region', 'Confirmed', 'Deaths', 'Recovered', 'Active', 'New cases', 'New deaths', 'New recovered', 'WHO Region'
Shape: (35156,10)

Case of deaths, recovered, active observed per day in both country and WHO wise, are information provided by this database

5. **usa_country_wise**

columns: 'UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Province_State', 'Country_Region', 'Lat', 'Long_', 'Combined_Key', 'Date', 'Confirmed', 'Deaths'

Shape: (627920,14)

This database gives us information on the total number of confirmed cases and fatal cases that occur daily in the Unite d States of America, split down by state or province. Furthermore, we own the geolocation information.

6. **worldometer_data**

columns: 'Country/Region', 'Continent', 'Population', 'TotalCases', 'NewCases', 'TotalDeaths', 'NewDeaths', 'TotalRecovered', 'NewRecovered', 'ActiveCases', 'Serious,Critical', 'Tot Cases/1M pop', 'Deaths/1M pop',

'TotalTests', 'Tests/1M pop', 'WHO Region'

Shape: (209,16)

The "worldometer" database gives us a more thorough picture of the overall number of cases, including fatalities, recoveries, and cases per continent, nation, and WHO. We also have the population data, which is quite helpful.

7. **owid-covid-data**

Columns: 'iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases', 'new_cases_smoothed', 'total_deaths', 'new_deaths', 'new_deaths_smoothed', 'total_cases_per_million', 'new_cases_per_million', 'new_cases_smoothed_per_million', 'total_deaths_per_million', 'new_deaths_per_million', 'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients', 'icu_patients_per_million', 'hosp_patients', 'hosp_patients_per_million', 'weekly_icu_admissions', 'weekly_icu_admissions_per_million', 'weekly_hosp_admissions', 'weekly_hosp_admissions_per_million', 'new_tests', 'total_tests', 'total_tests_per_thousand', 'new_tests_per_thousand', 'new_tests_smoothed', 'new_tests_smoothed_per_thousand', 'positive_rate', 'tests_per_case', 'tests_units', 'total_vaccinations', 'people_vaccinated', 'people_fully_vaccinated', 'new_vaccinations', 'new_vaccinations_smoothed', 'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred', 'new_vaccinations_smoothed_per_million', 'stringency_index', 'population', 'population_density', 'median_age', 'aged_65_older', 'aged_70_older', 'gdp_per_capita', 'extreme_poverty', 'cardiovasc_death_rate', 'diabetes_prevalence', 'female_smokers', 'male_smokers', 'handwashing_facilities', 'hospital_beds_per_thousand', 'life_expectancy', 'human_development_index'

Shape: (91026,59)

With 59 columns, this dataset is enormous, as can be seen. We have listed the titles of every column here for informational purposes, but we will not be using them all in our project because the information in some column does not help us achieve our project's objective. Alternatively, we might briefly discuss the dataset design. We will receive information on the number of cases in each category from this dataset. Age distribution, population density, number of incomplete and complete vaccinations, and test unit information. Details regarding the number of beds per thousand, certain behaviors, such as smoking, that might negatively affect a patient's health and previous medical conditions a patient may have, such as diabetes or heart failure.

8. ***covid_19_usa_city***

Columns: *'City', 'Total Cases', 'New Cases', 'Total Deaths', 'New Deaths', 'Active Cases', 'Total Cases /1M pop', 'Deaths /1M pop', 'Total Tests', 'Tests /1M pop', 'Date'*

Shape: (9660,11)

This dataset provides daily information on American cities' total cases, including deaths, active cases, and tests per milli on.

Relational Schema: Please find the schema on the last page of this document.

Database Implementation

Data Insertion:
As we are using real-life data, we already have a CSV file that we are going to import into different software.

Data Manipulation:
Given the use of real-life data, it is imperative that we maintain data quality by conducting a thorough data cleaning proc ess. This process involves scrutinizing data types and identifying any missing values or errors that may have been intro duced during data collection. By executing these steps, we can ensure the accuracy, dependability, and validity of our d ata. Such measures are critical for ensuring that our business or academic endeavors benefit from the most reliable and accurate data possible.

Query Examples:
We need to execute a minimum of fifteen SQL queries as required on our PostgreSQL/PgAdmin file.
However, for a better understanding, we can offer two examples of queries.

- If we want to see the top countries where Covid-19 caused the most deaths, we can sort the number of deaths by country or region in descending order.

    - *SELECT country_region, deaths FROM Country_wise*
      ORDER BY deaths DESC;

- In case we want to see the countries with the highest deaths_per100_recovered.

```
SELECT country_region, recovered_per100_cases FROM Country_wise
    ORDER BY recovered_per100_cases DESC;
```

Database Integration:
We are integrating Python for data analysis, visualization, and database integration using Python.

Capstone Complexity

Software:

To complete this project, we will use several software packages as per the task requirements. The initial phase involves creating a PostgreSQL database. We will then standardize input data and upload it into the database using Python. Finally, we will perform analytics using Python and Jupyter Notebooks, which offer multiple packages for efficient and quick analysis. We are going to use Excel and Tableau as well.

Project Completion Plan

Week 1: Choosing a suitable project idea and conducting thorough research to collect relevant data is crucial for a succ essful project. The information obtained will enable us to develop a comprehensive project plan that meets our needs.

Week 2: Our initial undertaking is to explore the dataset and conduct an exploratory analysis of the data. Once this phas e is complete, we intend to finalize the project's direction and subsequently develop a comprehensive list of formal busi ness requirements.

Week 3: Here, our primary focus will be on the development and unit testing of the PostgreSQL database. Our objective is to work on SQL Data Definition Language (DDL) and Data Manipulation Language (DML) statements, which will assis t us in creating the necessary database and tables. Upon completion of this task, our subsequent step will be to identify and select the requisite packages to connect both R and Python with our database. This will enable us to optimize our d atabase performance and enhance our data management capabilities.

Week 4-5: This week, we will configure data transformation tools and start transforming data. We can easily configure ta bles by ingesting CSV data, converting it to a data frame, modifying columns, renaming columns, removing duplicates, and reordering rows. Then, we will use R to upload transformed data to our PostgreSQL database.

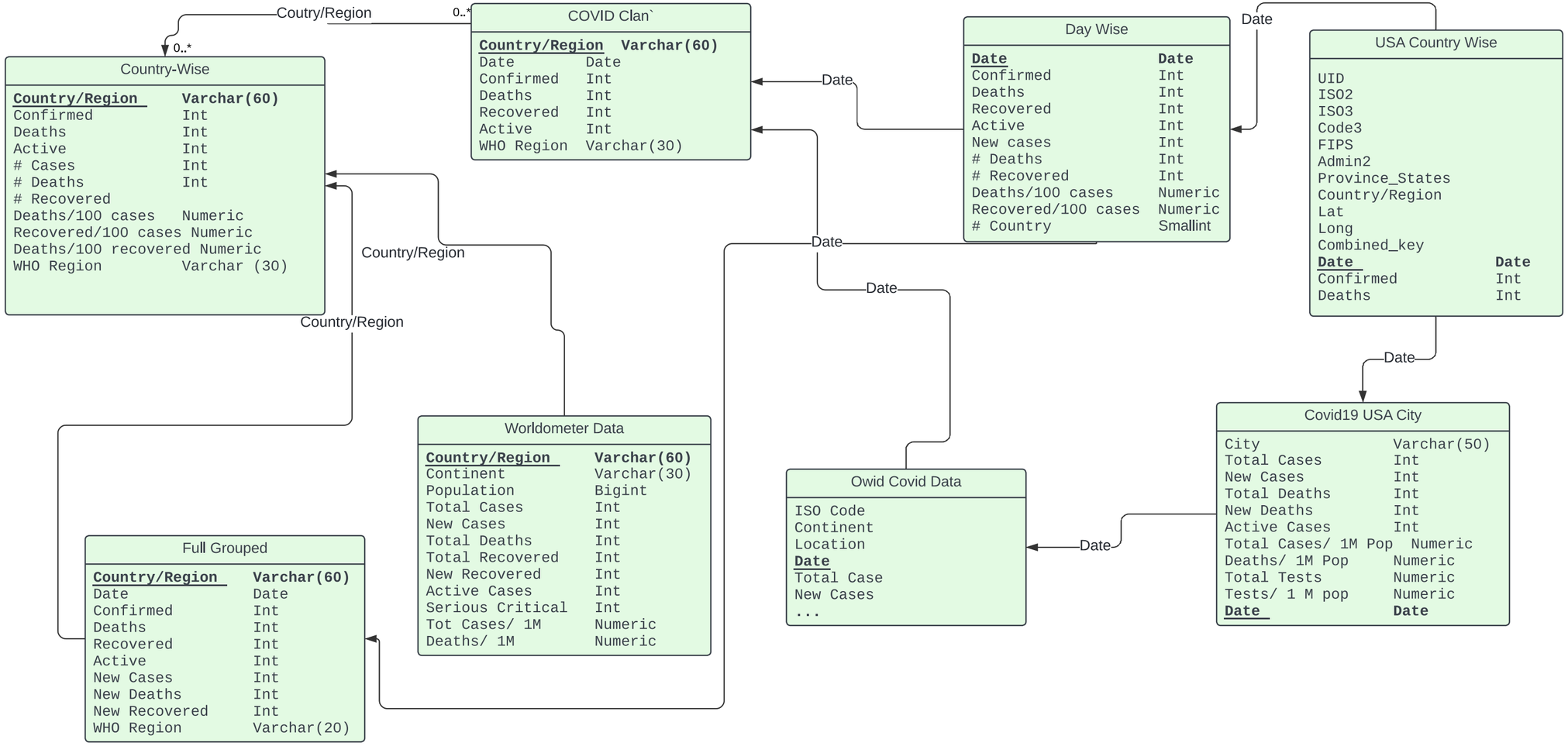
Week 6-7: During this phase of our project, we will be conducting both exploratory and explanatory analyses to address a variety of research inquiries related to the COVID-19 pandemic. We aim to answer all questions about our study. We w ill be using Tableau for data visualization and storytelling. Once the exploratory analyses are complete, we will shift our f ocus to finalizing the explanatory analyses. After that, we will apply machine learning algorithms to the dataset, and fine-tune the hyper parameters of the optimal model to achieve high performance.

Presentation Plan

The presentation is structured into three parts. Firstly, we will provide a comprehensive overview of the dataset, along w ith the business query that will be addressed subsequently. Secondly, a detailed, step-by-step explanation of the entire p rocess will be presented, including database design, data wrangling, data manipulation, and more. We will extensively d eliberate on the tools that have been employed for the project and the rationale behind their selection. Finally, upon con cluding the database segment of the project, We will showcase the analytics outcomes through a Jupyter Notebook and Tableau storytelling. The Jupyter Notebook will facilitate an artful representation of the manipulation of each data frame, as well as the resultant visualization.

Resources

- https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=country_wise_latest.csv
- https://www.kaggle.com/datasets/aditeloo/the-world-dataset-of-covid19


Coutry/Region
0..*
0..*
COVID Clan`
Country/Region Varchar(60)
Date Date
Confirmed Int
Deaths Int
Recovered Int
Active Int
WHO Region Varchar(30)
Country-Wise
Country/Region Varchar(60)
Confirmed Int
Deaths Int
Active Int
Cases Int
Deaths Int
Recovered
Deaths/100 cases Numeric
Recovered/100 cases Numeric
Deaths/100 recovered Numeric
WHO Region Varchar (30)
Date
Day Wise
Date Date
Confirmed Int
Deaths Int
Recovered Int
Active Int
New cases Int
Deaths Int
Recovered Int
Deaths/100 cases Numeric
Recovered/100 cases Numeric
Country Smallint
Date
USA Country Wise
UID
ISO2
ISO3
Code3
FIPS
Admin2
Province_States
Country/Region
Lat
Long
Combined_key
Date Date
Confirmed Int
Deaths Int
Country/Region
Country/Region
Date
Date
Date
Worldometer Data
Country/Region Varchar(60)
Continent Varchar(30)
Population Bigint
Total Cases Int
New Cases Int
Total Deaths Int
Total Recovered Int
New Recovered Int
Active Cases Int
Serious Critical Int
Tot Cases/ 1M Numeric
Deaths/ 1M Numeric
Owid Covid Data
ISO Code
Continent
Location
Date
Total Case
New Cases
...
Date
Covid19 USA City
City Varchar(50)
Total Cases Int
New Cases Int
Total Deaths Int
New Deaths Int
Active Cases Int
Total Cases/ 1M Pop Numeric
Deaths/ 1M Pop Numeric
Total Tests Numeric
Tests/ 1 M pop Numeric
Date Date
Full Grouped
Country/Region Varchar(60)
Date Date
Confirmed Int
Deaths Int
Recovered Int
Active Int
New Cases Int
New Deaths Int
New Recovered Int
WHO Region Varchar(20)