Covid-19 Regression Analysis Segment 2 - Dashboard

Akshaya Kamble, Tyler Engalla, Tommy Watson, Ray Hunt

Why Covid-19?

Covid-19 is a global pandemic which has rocked our society to its core. It has caused massive fatality, increased unemployment across the nation, and shutdown travel in many parts of the world. We have decided that as a team we will be analyzing Covid-19 data because of its relevance in everyday life around us. It has affected our health, our jobs, and our way of life. Over the next few weeks, our team will be focused on recognizing and analyzing trends due to covid-19.

To analyze Covid data across the United States and the world we will be pulling data from several sources. These covid-19 sources include: CDC data, Kaggle data, World Bank Data, and Our World In Data. These datasets include information about GDP, infection rates, vaccination rates, age distribution, death rates, Human Development Index, and much more.

Links to Sources:

CDC Link: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?name_desc=false
Kaggle Data: https://www.kaggle.com/gpreda/covid-world-vaccination-progress
World Bank Data: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
Our World In Data: https://github.com/owid/covid-19-data/tree/master/public/data
UNDP data for hdi : http://hdr.undp.org/en/data

Description of Data source

1. **country_gdp**
This file from world bank has GDP for all countries starting from year 1960 to 2019, Since we are using covid-19 data for the project we will use gdp from the year 2019.

2. **country_vaccinations**
This file from kaggle has vaccination data for all countries along with dates,As the vaccinations are still in progress some countries have advanced towards some small percent of population considered fully vaccinated while some countries no not have any data.

3. **world_population**
This file from world bank has populations for all countries from year 1960 to 2019. As per project requirements we have used data only for the year 2019.This data is added to the gdp data file

4. **human_dev_index**
This file from UNITED NATIONS DEVELOPMENT PROGRAMME has human development index(HDI) for all countries along with Human Development Type which are required for the project

5. **Infection_Data**
This file from github repository has updated data from Our World in Data.This file has data about the infections rates,deaths, new case,hospitalization rate, testing dates for different countries.We have filtered the dataset as per project requirements and included data like total cases and total deaths.

What we hope to accomplish with the Data

Given our strict timeline on this Covid-19 analysis, we would like to analyze how certain attributes such as age, Gross Domestic Product (GDP), Human Development Index (HDI), and population directly impact the spread of Covid-19.

If time allows, we would also be interesting in creating a model to help us predict the spread of Covid-19 in the future.

We would like to make the visuals extremely user friendly in the sense that the user will not have to navigate away from the page and be very easy to understand. All of our graphs will include titles, labels, and axises to make the message crystal clear.

Data Exploration Phase Description

1.Importing data from websites in csv and excel formats

2. Cleaning data in Jupyter notebook using the following methods
   a. Checking Data Types
   b. Replace NaN by 0
   c. Removing Nan
   d. Outlier detection with help of scatter plot
   e. Editing column names by removing spaces
   f. Changing column names to match the SQL schema
   g. Filter required columns as some tables contain irrelevant data for project

3. Making Database connection to Postgres and sending the cleaned Data Frames from Jupyter notebook to Postgres
4. Joining tables in postgres to create master data files
5. Making database connection to Postgres to import the joined tables in Jupyter notebook as new DataFrame.
6. Using the Data Frame for machine learning

Analysis Phase Description

As data cleaning is the primary requirement for getting a perfect machine learning model we have incorporated the following methods for cleaning data in the individual files.For the project our main focus was to obtain data for GDP,HDI,population and people fully vaccinated for each country so that we can find relations between the data we have.Before sending the data to machine learning the data was cleaned, sent to postgres for storing and joining data as required.The data will help us determine the relation between HDI vs Total cases per million, HDI vs Population,GDP per Capita vs Total Cases per Million,Population Density vs Total Cases per Million,Median Age vs Total Deaths,Aged 70 or older vs Total Deaths,GDP vs Vaccination Rates,HDI vs Vaccination Rates.

In some columns the null values are dropped because we require data for the respective countries and in some columns the null values were replaced by 0 as the model will not work with null values and 0 can be useful in some cases.The column names were changed in some files to match the exact column name in other files as these are required for joining data in postgres.The required columns are kept and some unwanted were deleted,Outliers were determined by plotting scatter plots and changing the axis range accordingly.

Tools to be Used

Our dashboard will be browser based.

We will build it using D3.js with the buildcharts() function.

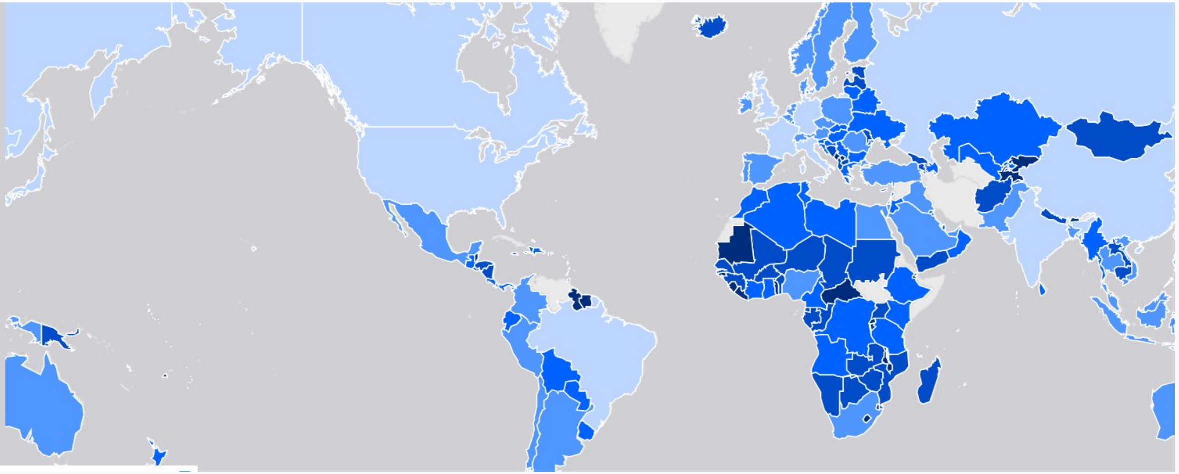
We will use JavaScript to populate and read a drop-down menu selection and update a bubble chart.

Finally, we will deploy our project to GitHub Pages.

Storyboard: World Map

Our first chart will be an interactive map of the world with data for select countries.

Based on filter selection, a tooltip will appear on hover-over to display that country’s name, population and GDP, Health Data Index or total vaccinations rate.

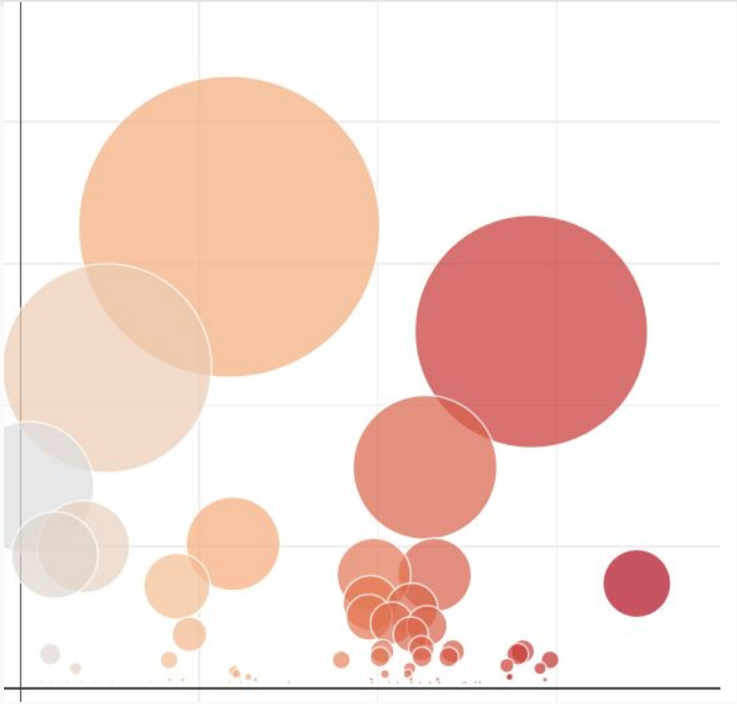


Storyboard: Bubble Chart

A drop-down menu will let the user select the top 10 countries by GDP, bottom 10 countries by GDP or all countries.

When the selection is made the bubble chart will update displaying countries as bubbles.

The Y axis is population, the X axis is GDP and the size of the bubble is determined by total vaccinations per hundred.



Machine Learning - Preliminary Data Preprocessing

- Acquired Datasets for Vaccinations, GDP, Infections, and HDI
- Imported Pandas , sklearn, matplotlib.pyplot, hvplot.pandas, statsmodels.api
- Imported the dataset that has gone through our ETL process
- Drop any null values still in the data. We only want countries that have statistics for everything we want to test for.
- Kept only columns that we were wanting to perform regression analysis on

Machine Learning - Preliminary feature engineering and preliminary feature selection, including their decision-making process

- For the Vaccination Data, we started by joining this Vaccination table with the GDP table on “country_name”
- There were multiple months of dates, but we only needed one to perform the regression analysis. So we filtered to get the date that gave us the most countries tied to it with the least amount of null values across the features
- Then we wanted to select the feature that gave us an idea at the rate countries were getting vaccinated, so we took “total_vaccination_per_hundred” and regressed it against other features such as GDP, HDI, Life Expectancy at Birth, and Population.
- We scaled the data, but found it wasn’t necessary for our regressions to take place

Machine Learning - Description of how data was split into training and testing sets

- For our Regression Analysis, there wasn't a need to split the data for training and testing sets.
- However, we did split the data into our independent and dependent variables to see how one features affects or how correlated it is with the other.
- Vaccination Rates was chosen as our dependent feature or data set, and GDP, HDI, and Population were chosen as our main independent features to see how they each affected Vaccination Rates.

Machine Learning - Explanation of model choice, including limitations and benefits

- We chose to perform Regression Analysis so that we can see how Vaccination Rates are either positively or negatively correlated with things like GDP, HDI, and Population. This would allow us to easily visualize this relationship and also give us an idea of how strongly they're having an impact on Vaccinations within different countries.
- Limitations with this model would be around outliers having a huge effect that that skews our findings. As well, we're only comparing two variables against each other at a time. So this gives us of an idea of the relationship between the two but it's not a complete picture of the relationship and what else might be impacting it.