Covid-19 Regression Analysis

Akshaya Kamble, Tyler Engalla, Tommy Watson and Ray Hunt

Why Covid-19?

- Very Relevant
- Data Availability
- Challenge
- Genuinely Curious

390M

Confirmed Cases Worldwide

838,260,663

Received at least 1 dose (0.2% of population)

Factors Affecting Vaccination Rates

- Gross Domestic Product (GDP)
- Human Development Index (HDI)
- Population
- Infection Rate

Technologies Used

⚙️	**LANGUAGES**	• Python Libraries - Pandas, Numpy, json • JavaScript
📊	**DATABASE**	• PgAdmin • Python Libraries - sqlalchemy, psycopg2
🧪	**STATISTICAL TOOLS**	• Python Libraries - sklearn.linear_model , sklearn, statsmodels
📈	**VISUALIZATION TOOLS**	• Python Libraries - plotly, hvplot, matplotlib, • HTML • JavaScript Libraries - Leaflet and D3.js • GitHub pages
🧫	**OTHER**	• Jupyter notebook • Visual studio code • GitHub • Git Bash

Database

csv, Jupyter notebook, pgAdmin

Data Sources

Name of our Files :

1. Country_gdp
2. Country_vaccinations
3. World_population
4. Human_dev_index
5. Infection_Data

Data Exploration

5

Different Data Sets

2019

This is the year !

150

Countries in Vaccination Dataset

82
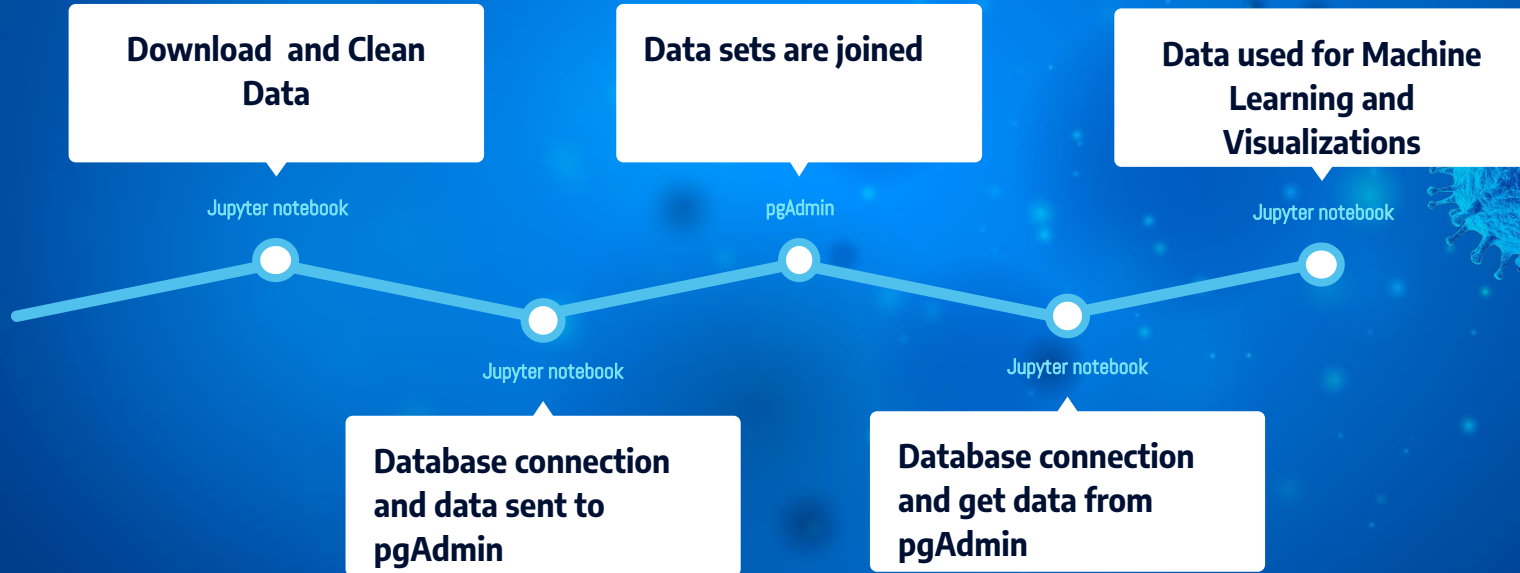
Countries with 0.2% people full vaccinated

138

Countries with people vaccinated

NaN Data

For some countries

Our Data Flow

Download and Clean Data

Data sets are joined

Data used for Machine Learning and Visualizations

Jupyter notebook

pgAdmin

Jupyter notebook

Jupyter notebook

Jupyter notebook

Database connection and data sent to pgAdmin

Database connection and get data from pgAdmin

Analysis Phase Description

DATA SETS CLEANING

TABLES JOINED

MACHINE LEARNING

1. Replace NaN by 0
2. Removing Nan
3. Detecting outlier
4. Editing column names
5. Changing column names
6. Filter required columns
7. Grouped data by country names

Data sets joined in pgadmin using inner join and Primary keys

1. Gdp_vaccination_join
2. Gdp_vaccination_hdi
3. country_group

Tables imported to Jupyter notebook as Dataframes from pgadmin for machine learning

Machine Learning

Linear Regression

Why this model?

Advantages

Disadvantages

⚙ Reliable method to identify variables that have an impact on our topic of interest

⚙ Outliers skewing the data

⚙ Overfitting

⚙ Easy way to visualize relationship

Dashboard
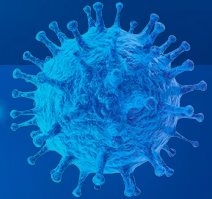
How'd our model do?

Based on GDP, HDI Rank, Population, how well can we predict:

2.9%
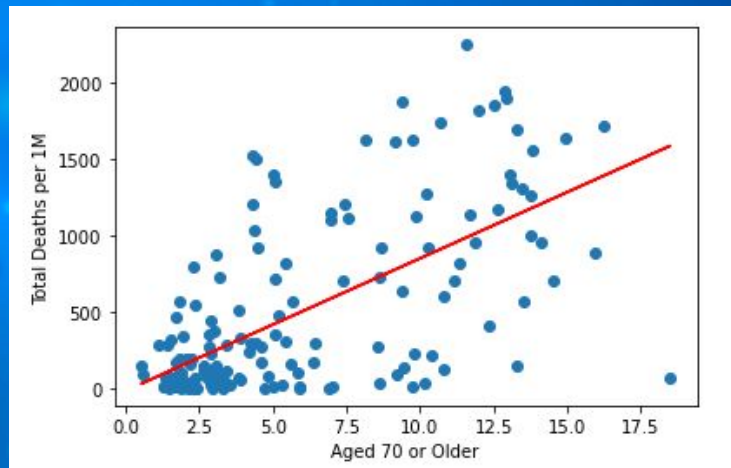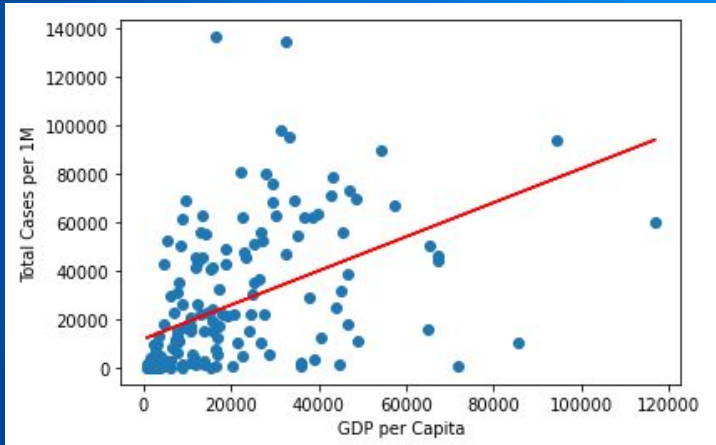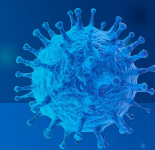
Vaccination Rate (Total Vaccinations per 100)

69%

Total Vaccinations

Exploration of Covid Case Data

Recommendations for Future Analysis

Data:

- Bigger Datasets

Machine Learning:

- Improve Accuracy by changing the test train ratio
- Check for more dependent variables
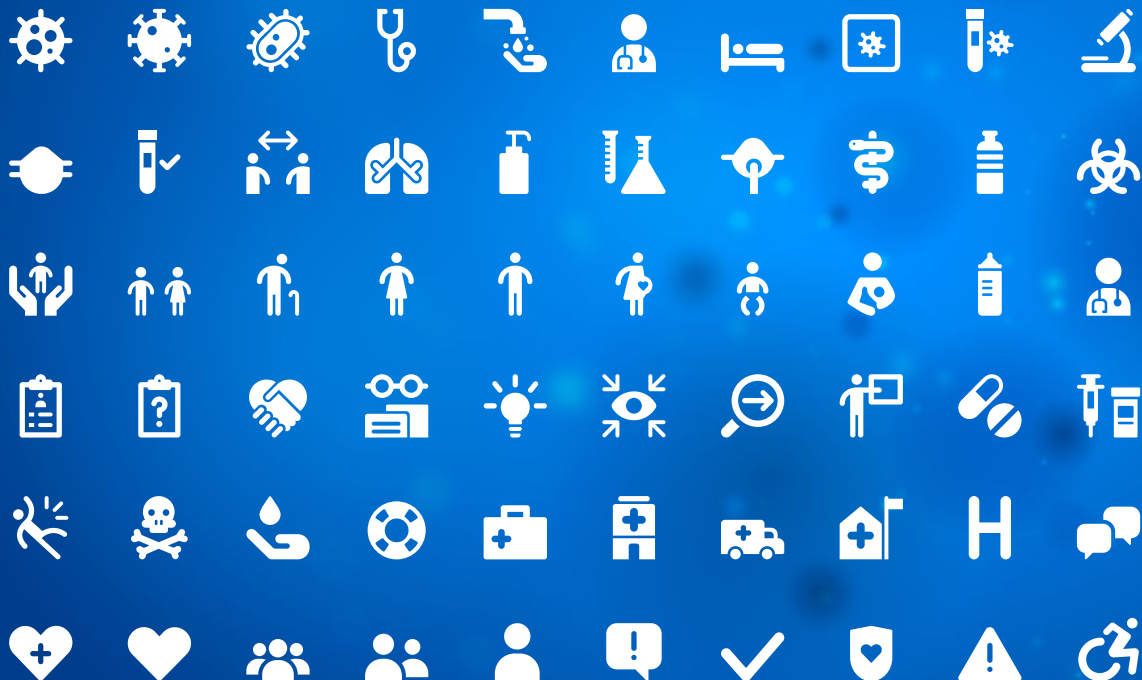
Visualization:

- Tableau

Anything done differently

1. Consider more dependent variables for better Accuracy
2. Use of Tableau
3. Highlight countries with colors using other libraries and Interactive Choropleth Map

Sources

- CDC Link: https://covid.cdc.gov/covid-data-tracker/#datatracker-home

- 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

Medical icons

19

SlidesCarnival icons are editable shapes.

This means that you can:

- ‣ Resize them without losing quality.
- ‣ Change fill color and opacity.
- ‣ Change line color, width and style.

Isn't that nice? :)

Examples:

Find more icons at
slidescarnival.com/extra-free-resources-icons-and-maps

20