10.020 Data Driven World

Design Thinking Project III

SC06 Group 4

Group members and their contributions

Yong Zheng Yew 1005155: Data cleaning, building model for tasks 1 and 2
Yeoh Siew Ning 1005471: Sourcing data, report writing for tasks 1 and 2
Devon Cheng 1005215: Excel for task 1, content creation for task 3
Guo Ziniu 1004890: Data cleaning, building model for tasks 1 and 2
Yeo Kai Wen 1005291: Content creation for tasks 2 and 3, report writing for task 2
(to include student ID and contributions to final report)

Task 1

## Introduction

Through our project, we aim to help the World Health Organisation (WHO) in projecting the death rate of COVID-19 in different countries. To do so, we have built a multiple linear regression model that predicts the number of deaths (over a period of time representing one wave of infection) in various countries due to COVID-19.

The primary data source used for Task 1 is taken from the Coronavirus Pandemic (COVID-19) research as conducted by Our World in Data (OWID). Specifically, data regarding COVID-19 cases and deaths is sourced from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

The data set is updated daily and collates data for countries worldwide from reputable sources, such as data from a country's ministry of health. The fact that COVID maps that track the global outbreak created by reputable sources such as CNN and BBC have referenced the COVID-19 Data Repository was evidence of this source's reputability and reliability.

## Selecting Countries and Selecting Time Period

Data regarding "Confirmed cases" and "Confirmed deaths" was collected from 196 countries. However, other variables which we considered as potential features were collected from varying numbers of countries; for example, variables regarding "Hospital & ICU" were collected from 38 countries, while those regarding "Other variables of interest" were sourced from 241 countries. Thus, it was necessary to consider a number of country-related factors:

- Absolute number of countries considered
- Number of variables that a certain country has missing entries for
    * For example, Afghanistan has 28 missing entries for `new_deaths`
- Number of countries that are missing a certain factor
    * For example, 222 countries are missing entries for `new_deaths`

Initially, our problem statement was: **Given a date and other statistics related to a single country, predict the number of COVID-related deaths in that country on that day**. To tackle this, we first chose `new_deaths` (the number of new deaths on each day) as our target, and so all other selected variables would be features.

For the sake of extensibility and modularity, we planned to choose countries and features programmatically, by setting a cutoff for a maximum amount of features that a country is allowed to lack statistics for, and a cutoff for a maximum amount of countries that could be missing a certain variable. We would automatically choose countries and variables that were relatively common.

However, we soon realized that many useful variables were amongst those which most countries were missing. We therefore abandoned this automated method in favor of manually choosing features that made logical sense in context. We ended up with plotting out many graphs, each depicting `new_deaths` against one feature, which gave us many unreadable graphs.

We then attempted to plot the same axes, but with only one country's data, in hopes of understanding the big picture better. We realized: For every day, each country would have a new `new_deaths` entry, and so we actually had to account somehow for the dimension of time.

Initially, we assumed time (`date`) would be just another feature, but further research showed this was incorrect. We then pivoted to a cumulative approach, changing our problem statement to: **Given statistics related to a single country, predict the total number of COVID-related deaths it will suffer over the course of a single COVID wave**.
(In scoping our problem statement as such, we are able to contextualise the problem with respect to a time frame of a single COVID wave, thus cumulative data would be able to provide information for each feature and target up after a single COVID wave.)
We repeated manual selection for cumulative features and chose `total_deaths` (cumulative deaths up to a certain point) as our target.

On further analysis of the graph of `total_deaths` against `date`, we realized that in the section of time which the given data covered, different countries were undergoing different phases in their respective waves of COVID-19 infections. Therefore, we should rather convert time to phase period and make the label no longer time-dependent.

The graph was sinusoidal, which represented each country’s waves of COVID-19 infections. The time period for each wave varied in start date and duration for every country. To aid in analysing the cumulative COVID-19 deaths, countries with similar start date and duration were selected. Our group decided to choose countries hailing from the European Union (EU) and surrounding countries. Geographically, these countries are located close to one another (and EU countries would even have less travel restrictions), hence they would go through a wave of COVID at similar times. Looking at the graph of `total_deaths_per_million` against `date`, we identified three different COVID waves.

GRAPH

We chose to study the middle wave, since it was better defined. Since it was surrounded on either side by the other periods, we could see where the middle wave started and ended, whereas the other two periods might be cut off by the edge of the graph where the data ends. The time period selected corresponded to a single wave of COVID-19 infections, from (both inclusive) 18 July 2020 to 26 August 2021.

The final list of countries chosen (total of 26) is as follows:
Austria, Belgium, Bulgaria, Croatia, Republic of Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden (Malta excluded due to extremely high population density).

Finally, we successfully plotted decent graphs and found that the trend of number of death cases versus different variables were noticeable.

## Preparation and Cleaning of Data Sets

### Target Variables
To predict the death rate of COVID-19 in different countries, the cumulative number of deaths due to COVID-19 during the time period was considered, the variable `new_total_deaths`. This was calculated by subtracting the value of `total_deaths` at the start of the cycle from its value after the cycle (since `total_deaths` is a cumulative statistic), giving us the amount of deaths that occured *during* that cycle.

### Selecting and Cleaning Predictor Variables
Initial selection of predictor variables was done based on logical intuition and background research. Other studies conducted to predict death rate due to COVID-19 mostly made use of medical data, most of which not made public or easily accessed. Instead, we tried to ensure that the predictor variables selected were representative of various aspects of society. These aspects include economic, social, medical, demographic and COVID-19 specific variables.

Taking into consideration all the possible variables (excluding variables on confirmed deaths), each variable was plotted against the target variable. Visually inspections were done to observe any obvious trends between the possible variables and target variables, upon which a number of variables were shortlisted.

To further ensure that the variables we chose were suitable, we used the r<sup>2</sup> coefficient as a metric to determine any correlation between the target variable (dependent variable) and the possible predictor variables (independent variable). As a baseline, variables with r<sup>2</sup> > 0 were chosen. (This is because r<sup>2</sup> < 0 indicates that the best-fit model fits worse than a straight line, which indicates that the feature and target likely do not follow a linear relationship.) (

[need explain why r<sup>2</sup> > 0?, I mean its possible to have r<sup>2</sup> thats negative, and that means that the model fit sucks that bad that a straight line is a better model)

Calculating r<sup>2</sup> coefficient aided in data transformation as well, as it revealed whether the predictor variable was linearly related to the target variable. Other kinds of relationships between the predictor variable and target variable were determined via trial and error.

Finally, we had chosen some features which logically made sense:

Variable Name	Reason
human_development_index	
extreme_poverty	
total_cases_per_million	
people_vaccinated_per_hundred	
total_tests_per_thousand	
stringency_index	

However, we needed to clean and manipulate this data into a usable form. Therefore, although they were not features in themselves, we required `location` and `date` in order to manipulate the data programmatically.

Since we were only considering

Thus, our finished set of features are:
- `new_total_cases_per_million`
- `new_total_tests_per_thousand`
- `new_people_vaccinated_per_hundred`
- `average_stringency_index`
- `extreme_poverty`
- `human_development_index`

Our target is `new_total_deaths_per_million`.

### Transformation of Data
Cumulative data type is required for variables that were an aggregate over the time period.

Transformation for the following four variables, which are "extreme_poverty", "life_expectancy", "new_total_tests", "average_stringency_index", is done by taking the natural log of the feature variables.

Variable Name	Description
new_total_cases	Cumulative new cases
new_total_deaths	Cumulative new deaths
average_icu_patients	Number of COVID-19 patients in intensive care units (ICUs)
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506
average_stringency_index	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
new_total_tests	Cumulative new tests for COVID-19
life_expectancy	Life expectancy at birth in 2019
hospital_beds	Hospital beds, most recent year available since 2010
new_total_people_vaccinated	Cumulative number of people were vaccinated

### Multiple Linear Regression Model
~~We use data transformation and Multiple Linear Regression (MLR) to find relations between different variables. The MLR Model we use includes functions of: data transformation, **something, something**~~
Multiple linear regression is used to determine the relationship between multiple independent variables and a dependent variable. The general mathematical equation for n independent variables would be:

ŷ\left(x\right)\ =\ \hat{\beta}_{0}+\hat{\beta}_{1}x_{1}+\hat{\beta}_{2}x_{2}\ +...+\ \hat{\beta}_{n}x_{n}

where ŷ = predicted dependent variable, x_{1}, x_{2}, …, x_{n} are independent variables and beta_{1}, beta_{2}, …, beta_{n} are the respective coefficients of the independent variables. It can be taken that $x_0 = 1$ with $\hat{\beta}_0$ as its coefficient.

In the context of Task 1, n = 6 and the equation is:

ŷ\left(x\right)\ =\ \hat{\beta}_{0}+\hat{\beta}_{1}x_{1}+\hat{\beta}_{2}x_{2}\ +\ \hat{\beta}_{3}x_{3}\ +\hat{\ \beta}_{4}x_{4}\ +\ \hat{\ \beta}_{5}x_{5}\ +\ \hat{\ \beta}_{6}x_{6}

where ŷ = predicted total number of COVID-related deaths, x_{1} = new_total_cases_per_million , x_{2} = new_total_tests_per_thousand, x_{3} = new_people_vaccinated_per_hundred, x_{4} = average_stringency_index , x_{5} = extreme_poverty, x_{6} = human_development_index, and beta_{1}, beta_{2}, beta_{3}, …, beta_{6}, are the coefficients of x_{1}, x_{2}, x_{3}, …, x_{6} respectively.

### Model Evaluation
The metric we chose is mean square error (MSE). Thus, it requires a huge number of iterations to finish the regression.

### Discussion and Analysis of Results
After performing MLR, we find that some relations of feature variables and number of death cases are intuitive, such as XXX

Some relations statistically make sense HOW?, such as death number by absolute date-time or locations, but are not meaningful in real life, so we should not try finding the regression of them.

On the other hand, some relations like the death number by GDP are quite counter-intuitive — the relation that we suspect they had (WHICH?) does not show up. This is explainable. Firstly, the p-value of such a dataset is so large that the data is meaningless. Secondly, some variables have no relation if we do not control other variables than them. Although this is a Multiple Linear Regression, which means that we can control most variables that may affect the regression, we are still not able to take everything into consideration.

### Future Improvements
- Taking data from different sources to get more updated and hence accurate data (eg. for hospital beds and HDI)
- More reasonable data cleaning: for Na and zero data, predict it instead of using it as training data.

Introduction
Through our project, we aim to help the World Health Organisation (WHO) in projecting the death rate of COVID-19 in different countries. To do so, we have built a multiple linear regression model that predicts the number of deaths (over a period of time) in various countries due to COVID-19.

The primary data source used for Task 1 is taken from the Coronavirus Pandemic (COVID-19) research as conducted by Our World in Data (OWID). Specifically, data regarding Covid-19 cases and deaths is sourced from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.
The data set is updated daily and colates data for countries worldwide from reputable sources, such as data from a country's ministry of health. Covid maps that track the global outbreak created by reputable sources such as CNN and BBC have referenced the COVID-19 Data Repository, thus [something validating the reliability of our chosen datasets].

Selecting countries and selecting time period
The data set for the metrics of "confirmed cases" and "confirmed deaths" contained data from 196 countries. However, other metrics that were being considered as predictor variables were from varying numbers of countries, ranging from 38 for "hospital & icu" to 241 for "other variables of interest". Thus, it was necessary to choose a suitable number of countries.

Initial attempts to select countries were done by considering the number of variables with empty data entries for each country, as well as the number of countries that had empty data entries for each variable. However, it was difficult to analyse the data and we decided to switch our approach entirely, realising that it would be easier to analyse the data visually through plots.

In trying to factor in time, we plotted out the [total_deaths] against time for each country. The graph was sinusoidal, which represented each country's waves of COVID-19 infections. The time period for each wave varied in start date and duration for every country. To aid in analysing the cumulative COVID-19 deaths, countries with similar start date and duration were selected. Our group decided to choose countries hailing from the European Union (EU) and surrounding countries. Geographically, these countries are located next to or close to one another, hence infection rates observed in each were similar. The time period selected corresponded to a single wave of COVID-19 infections, from 18 July 2020 to 26 August 2021.

[graph of the [total_deaths] against time for EU countries]

The final list of countries chosen (total of 26) is as follows:

- Austria, Belgium, Bulgaria, Croatia, Republic of Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden (Malta excluded due to high population density)

Preparation and Cleaning of data sets

Target variable
To predict the death rate of COVID-19 in different countries, the cumulative number of deaths due to COVID-19 during the time period was considered, the variable new_total_deaths. [why was this considered] This was calculated by summing up the new

deaths due to COVID-19 per day (new_death) over the time period. new_deaths as opposed to total_deaths as deaths that occured prior to the time period were not considered.

Selecting predictor variables

Initial selection of predictor variables was done based on logical intuition and background research. Other studies conducted to predict death rate due to COVID-19 mostly made use of medical data, most of which not made public or easily accessed. Instead, we tried to ensure that the predictor variables selected were representative of various aspects of society. These aspects include economic, social, medical, demographic and COVID-19 specific variables.

Taking into consideration all the possible variables (excluding variables on confirmed deaths), each variable was plotted against the target variable. Visually inspections were done to observe any obvious trends between the possible variables and target variables, upon which a number of variables were shortlisted.

To further ensure that the variables we chose were suitable, we used the r^2 coefficient as a metric to determine any correlation between the target variable (dependent variable) and the possible predictor variables (independent variable). As a baseline, variables with $r^2 > 0$ were chosen. [need explain why r^2 > 0?, I mean its possible to have r^2 thats negative, and that means that the model fit sucks that bad that a straight line is a better model)

[formula for r^2, not sure if needed]

Calculating r^2 coefficient aided in data transformation as well, as it revealed whether the predictor variable was linearly related to the target variable. Other kinds of relationships between the predictor variable and target variable were determined via trial and error.

Transformation of data

Cumulative data type was required for variables that were an aggregate over the time period.
Calculations for variables with naming convention “new_total_(variable)” and for “new_people_vaccinated” are as follows: [something]
(take sum,

Calculations for variables with naming convention “average_(variable)” are as follows: [something]
(on average, how strict a country was)

“Hospital_beds” was calculated by

Table outlining specific raw data used from the data set

Variable name	Description of variable
new_cases	New confirmed cases of COVID-19

new_deaths	New deaths attributed to COVID-19
icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day
stringency_index	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
new_tests	New tests for COVID-19 (only calculated for consecutive days) Note: implies that _
people_vaccinated	Total number of people who received at least one vaccine dose
location	Geographical location
date	Date of observation
population	Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv for full list of sources
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010
life_expectancy	Life expectancy at birth in 2019
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506

The final list of predictor variables used are as follows: human_development_index, extreme_poverty, average_icu_patients, life_expectancy, new_total_cases, new_total_tests, new_people_vaccinated, hospital_beds, average_stringency_index, new_total_deaths, gdp_per_capita

Table outlining cleaned data used for model

Variable name	Description of variable
new_total_deaths	Cumulative number of deaths due to COVID-19 during time period
new_total_cases	Cumulative COVID-19 cases during time period

average_icu_patients	Average number of COVID-19 ICU patients during time period
average_stringency_index	Average stringency index during time period
new_total_tests	Cumulative number of test for COVID-19 during time period
new_people_vaccinated	Cumulative number of people who received at least one vaccine dose during time period
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
hospital_beds	Total number of hospital beds available
life_expectancy	Life expectancy at birth in 2019
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506

Description of Multiple Linear Regression Model

- Include mathematical eqn and other math details

Model Evaluation

The metric we chose is mean square error (MSE). MSE gives the average of the squared distance between the predicted data points and the actual data points. We chose this metric because it could reveal how closely our predictions fit the actual data points.

[formula of mse]

where n = number of data points, y_i = actual number of deaths in data set , yhat_i = predicted number of deaths in data set

To achieve a good model, we would want to minimise MSE.

The other metric we chose is the adjusted r<sup>2</sup>. This differs from r<sup>2</sup> coefficient as it helps to identify features that are useful in predicting the target. Adjusted r<sup>2</sup> will decrease if there are features used that do not help in variation in the target variable.

[formula for adjusted r^2]

where n = number of data points, k = number of features, excluding the constant (x_0 = 1)

Similarly, we would want to maximise adjusted r<sup>2</sup> as it indicates that the features chosen are suitable in predicting the target.

Discussion/Analysis of results

- After performing MLR, we find that some relations of feature variables and number of death cases are intuitive. Some relations statistically make sense, such as death number by absolute date-time or locations, but are not meaningful in real life, so we should not try finding the regression of them. Some relations like the death number by GDP are quite counter-intuitive—the relation that we suspect they had does not show up. This is explainable. Firstly, the p-value of such a dataset is so large that the data is meaningless. Secondly, some variables have no relation if we do not control other variables than them. Although this is a Multiple Linear Regression, which means that we can control most variables that may affect the regression, we are still not able to take everything into consideration.

(There is a possibility that overfitting occurred, which means that the model is unable to generalise the dataset. This happens when the model is able to simulate a specific sample from a dataset, but is unable to simulate another sample from the same dataset. In the context of Task 1, this may be caused by the choice of countries used. Our sample of countries is mainly from the EU, which is not highly representative of all the countries in the world. Hence, trying to predict the number of COVID-related deaths in a country outside of the countries chosen previously may not give an accurate result.

To overcome this problem,)

Future improvements

- Taking data across different sources - to get more updated and hence accurate data - eg for hospital beds and HDI

Task 2

Introduction

With the Covid-19 pandemic hitting the global stage and affecting the lives of billions around the world, vaccinations are key to reducing the spread of the virus. However, the preparations for vaccines are straining the overworked healthcare sectors that have their hands full with handling Covid-19 patients and the healthcare needs of the general populations.

As such, our group has decided to predict vaccination rates for a chosen country, the United States of America (USA). Not only will it help hospitals in preparing resources for future vaccinations doses, the government can utilise our model to keep track of vaccination rates in respective states and use it to make adjustments in their policies if necessary.

We specifically chose USA as it has a lot of relevant data many countries lack that would greatly help in creating our model. Additionally, it only has about 59% of its population fully vaccinated, and is one of the countries that have been badly affected by the Covid-19 pandemic since the discovery of the virus.

(The problem statement that we have chosen is: **Given the statistics of a county in USA, predict the vaccination rate of that county.**)

Target

The data source used for vaccination rates, the target, in counties across the USA is taken from “COVID-19 vaccination data”. A study named “Moral Values Predict County-Level COVID-19 Vaccination Rates in the United States” referenced this data source, as it was determined that data provided by the Centers for Disease Control and Prevention (CDC) was “missing vaccination rates for some states and counties”, hence this data source used data from CDC and other sources to produce a complete data set. [which data source and why]

Drawing on lessons learned from Task 1, we determined it would be easier to analyse data from a specific snapshot in time, hence we decided to use the data set with the latest available for each county (data_county_current.csv), as opposed to data spanning across time period (data_county_timeseries.csv). [which specific data set and why]

Predictor variables chosen

Learning from Task 1, we determined that data from reputable sources would be most ideal, as it would likely be more reliable and accurate. County level data was taken from data compiled by the U.S. Department of Agriculture, which in turn sourced data from other governmental level organisations, such as U.S. Census Bureau. Data found was up to date (taken from 2019 or 2020) and unlikely to have as many empty data entries, as efforts to collect official data would be more uniform across a single country.

The indicators for which data was collected were “poverty”, “population”, “unemployment, and median household income”, and “education”.

The predictor variables chosen are as follows:

The predictor variables for our model are chosen based on

Preparation and Cleaning of data sets

At first the data sets for each category (complete coverage aka full vaccinated, partial coverage aka partially vaccinated, poverty, unemployment rate, education levels, population)were provided separately, hence there was a need to combine all the data sets into a single table for simple perusal.

Next, the education levels were separated into four categories: below high school level, high school level, college level, and bachelor's degree. It would be unnecessarily complicated to consider all the various education levels as variables for the model, so they were simplified into a percentage of the population indicating those who have at least college level education. *Initially, we wanted to add weight percentage on each education level, but we found it difficult to justify how much percentage we use. We also specifically chose college level education as the benchmark as it was observed how its numbers potentially correlate with the vaccination rates.*

Afterwards, we added data sets for confirmed cases and confirmed deaths onto the tables. Confirmed cases and deaths converted into percentage of the population *as population varied among different counties and it would be more meaningful to show the confirmed cases and deaths as a ratio in regards to the population as opposed to the absolute number.*

Confirmed cases were removed as a variable since it showed no relation to vaccination rates according to r^2 and observation of the scatter graph.

Task 2

Building Model

Describe your model. Is this Linear Regression or Logistic Regression? Put any other details about the model. Put the codes to build your model.

Similar to Task 1, we decided to use a Multiple Linear Regression model. This is a suitable model because in our problem statement, there is a clear target, which is the vaccination rates for counties in the USA, and the features, statistics such as indicators that provide information about different aspects of each county. Hence, the model that we have previously built in Task 1 would be suitable for Task 2 as well. The key difference in the two models would be the data that is used and how it is cleaned.

In the context of Task 2, the equation is:

ŷ\left(x\right)\ =\ \hat{\beta}_{0}+\hat{\beta}_{1}x_{1}+\hat{\beta}_{2}x_{2}\ +\ \hat{\beta}_{3}x_{3}\ +\hat{\ \beta}_{4}x_{4}

where ŷ = predicted vaccination rate, x_{1} = Unemployment_rate_2020 , x_{2} = PCTPOVALL_2019 , x_{3} = CollegeLevel , x_{4} = cases_and_deaths_percentage and beta_{1}, beta_{2}, beta_{3}, beta_{4} are the coefficients of x_{1}, x_{2}, x_{3}, x_{4} respectively. It can be taken that $x_0 = 1$ with $\hat{\beta}_0$ as its coefficient.

Evaluating the Model

As mentioned, the same MLR model from Task 1 was used. Thus, the same metrics, MSE and adjusted r<sup>2</sup>, were used to evaluate the performance of our model. This time, we fit the equations to the context of Task 2.

Discussion and analysis

The adjusted r<sup>2</sup> of the regression was not perfectly ideal.

Initially, we suspected that the reason why the target and feature showed weak relationship, in spite of the accuracy of our regression model, was that more factors needed to be taken into consideration. Thus, we tried including more variables as features since our model was a multiple linear regression.

However, the result was still not ideal, and then we realised that this was because they simply did not have a strong relationship in the real world context. Therefore, we decided to present the data as it is, as opposed to modifying the data in hopes of plotting a decent regression graph.

The MSE we computed is _, although admittedly it is difficult to determine whether it has been minimised without any reference. This is because each model is unique, hence there is no threshold as to what is a “good” MSE value to obtain.

Reference list

Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

Merritt, Alexes; Tiu, Andrew; Bansal, Shweta, 2021, "Integrated US COVID-19 Vaccination Data", https://doi.org/10.7910/DVN/BFRIKI, Harvard Dataverse, V1.

Andrew Tiu, Zachary Susswein, Alexes Merritt, Shweta Bansal. Characterizing the spatiotemporal heterogeneity of the COVID-19 vaccination landscape. MedRxiv.

Karimi-Malekabadi, F., Reimer, N. K., Atari, M., Trager, J., Kennedy, B., Graham, J., & Dehghani, M. (2021, September 15). Moral Values Predict County-Level COVID-19 Vaccination Rates in the United States. https://doi.org/10.31234/osf.io/z6kxm

The Visual and Data Journalism Team. (2021, November 22). *Covid map: Coronavirus cases, deaths, vaccinations by country*. BBC News. Retrieved November 22, 2021, from https://www.bbc.com/news/world-51235105

Pettersson, H., Manley, B., Hernandez, S. (2021, November 25). *Tracking Covid-19's global spread*. CNN. Retrieved November 26, 2021, from https://edition.cnn.com/interactive/2020/health/coronavirus-maps-and-cases/

Stephanie Glen (2021). "Adjusted R2 / Adjusted R-Squared: What is it used for?" Retrieved November 26, 2021, from StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/adjusted-r2/

Links to data sources

Task 1

Data	Source
Features	https://github.com/owid/covid-19-data/tree/master/public/data
Data regarding deaths	https://github.com/CSSEGISandData/COVID-19
Population	https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv
HDI	http://hdr.undp.org/en/indicators/137506

Task 2

Data	Source
Vaccination rates - target	https://github.com/bansallab/vaccinetracking/tree/main/vacc_data
Features	https://www.ers.usda.gov/data-products/county-level-data-sets/