Contents

Objective: 3
What data sources were used? 3
Why you chose those data sources? 3
What target you chose? 4
Visualisation of Target variables against infection and vaccination rates. 5
Population Variables and Real GDP: 5
Climate Variables: 6
Stringency Index: 7
What techniques you did use? 8
Part one: The World Factbook 8
Part two: World Health Organisation, Oxford Covid-19 Government Response Tracker and OECD 9
Part 3: Global Historical Climatology Network daily (GHCNd) 10
Part 4: Converging the 3 data frames into the final Data frame 10
What difficulties you had to overcome to wrangle the data sources into the target data model? 12
Limitations 13

Objective:

This assignment aims to investigate whether population-based variables, GDP per capita, climate variables and strictness of government policy have influenced Covid-19 vaccination and infection rates.

What data sources were used?

The five web-based sources were:

- The World Factbook - Central Intelligence Agency
    - https://www.cia.gov/the-world-factbook/
- Global Historical Climatology Network daily (GHCNd)
    - https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily
- World Health Organisation (WHO)
    - https://covid19.who.int/WHO-COVID-19-global-table-data.csv
    - https://covid19.who.int/who-data/vaccination-data.csv
- Oxford Covid-19 Government Response Tracker
    - https://covidtrackerapi.bsg.ox.ac.uk/api/v2/stringency/date-range/{YYYY-MM-DD}/{YYYY-MM-DD}
- The Organisation for Economic Co-operation and Development (OECD)
    - https://www.oecd.org/about/members-and-partners/

Why you chose those data sources?

- The World Factbook was chosen as a source because it was created to provide insight to Government Officials. This source is also used regularly for scholarly research, as it is deemed reliable and is publicly available.
- The GHCNd was chosen as a source because GHCN data is extensive as it integrates numerous sources. These sources of data are subjected to quality assurance reviews before being included in the dataset. The database uses measurements from over 100,000 weather stations which span over 180 countries.
- The World Health Organisation was chosen as a source because they manage and maintain a comprehensive database on Covid-19. They also follow data principles that are designed to ensure the data is reliable.
- The Oxford Covid-19 Government Response Tracker was chosen as a source because Oxford is an elite educational institute. The data was collected by a team at Oxford University along with students from around the world.

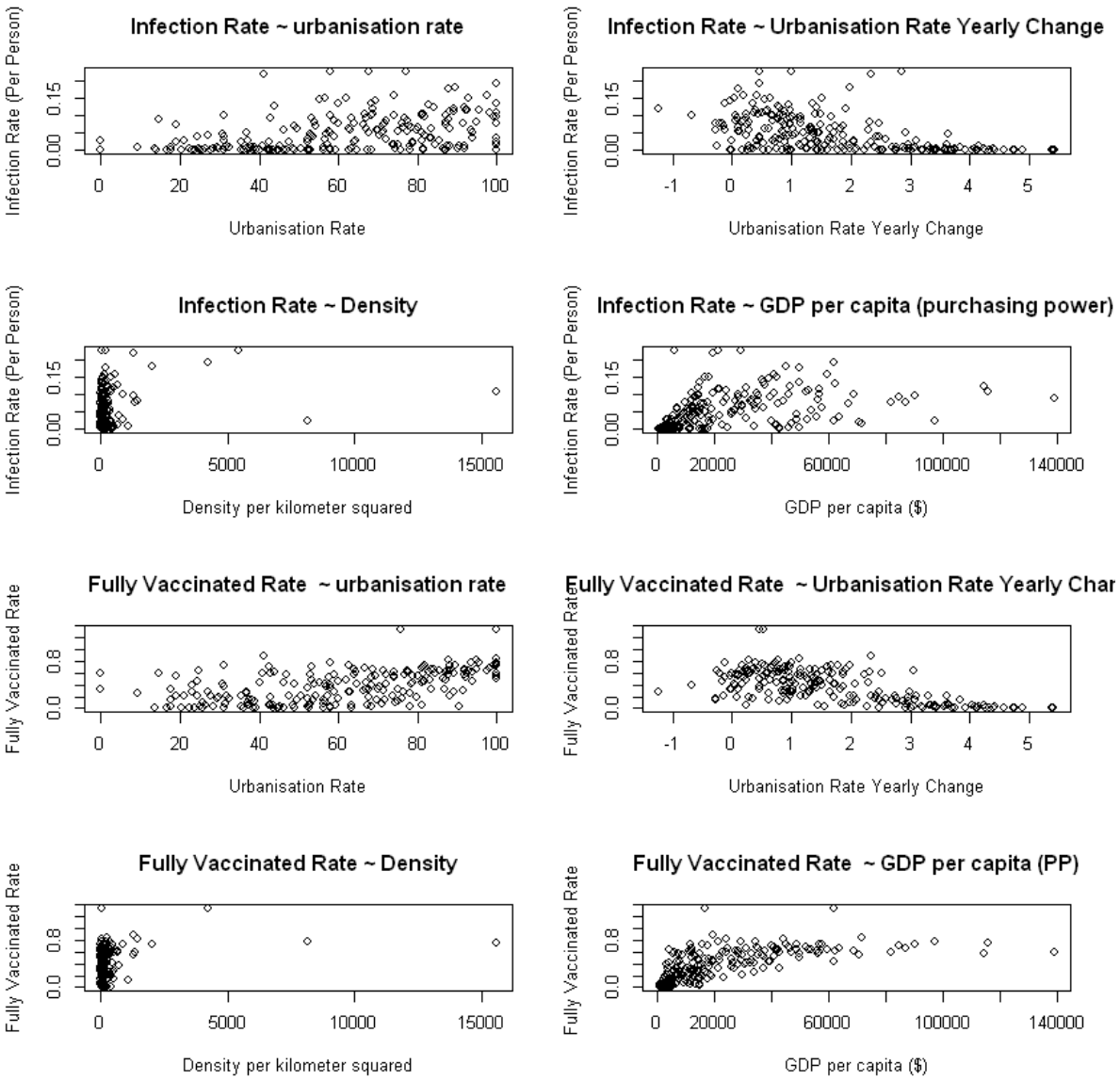
What target you chose?

The target of the database was to be able to determine the extent of correlations and multicollinearity between various country data and the Covid-19 data for each country. These variables used were:

- The variables scraped from The World Factbook:
    - o population
    - o urbanisation rate
    - o urbanisation yearly rate change
    - o Density (This variable was created by dividing population by size)
- The variables from the GHCNd's database:
    - o average temperatures
    - o precipitation
- The variable extracted from the Oxford Covid-19 Government Response Tracker:
    - o stringency index
- The variables extracted from the World Health Organisation Database:
    - o total number of cases
    - o total number of deaths
    - o the total number of vaccinations
    - o the total number of people who received the first dose
    - o the total number of people fully vaccinated
- The variable scraped from the OECD:
    - o List of member states
- The variable retrieved from the country code package or a dictionary:
    - o ISO3 country code

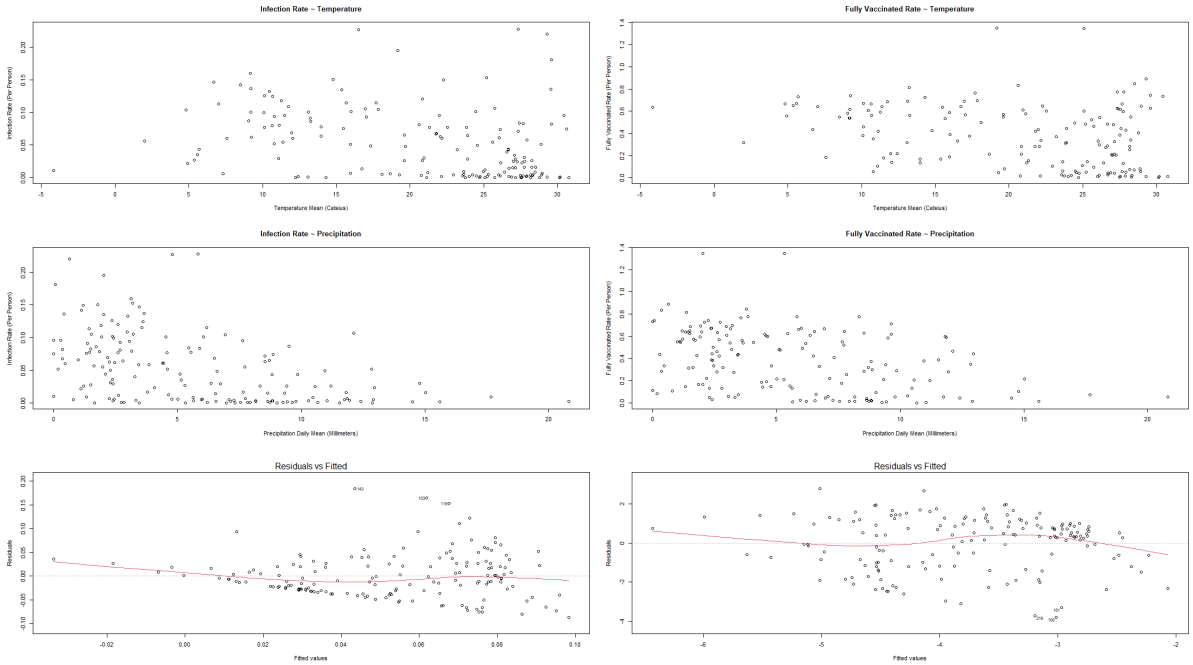
Visualisation of Target variables against infection and vaccination rates.

Population Variables and Real GDP:



The six plots show each of the variables gathered from the World Factbook website compared with the countries Covid-19 infected and fully vaccinated rates per person. The Urbanisation rate and real GDP per capita have a positive correlation. On the other hand, urbanisation rate yearly change has a negative correlation. The only variable collected that was not statistically significant (based on p-values) was the density rate. These trends might be linked to wealth as density is the only variable not associated with a country's wealth.

Climate Variables:

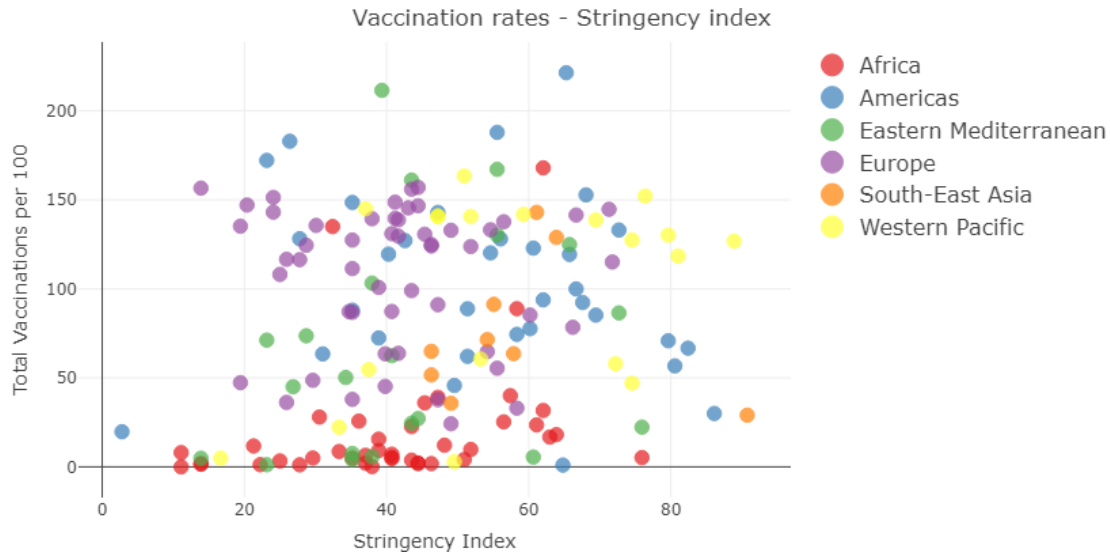



```
Call:
lm(formula = vacr ~ tavg + prcp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43697 -0.21573  0.01572  0.16590  0.99087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.630991   0.058095  10.861  < 2e-16 ***
tavg        -0.006141   0.002856  -2.150   0.0331 *
prcp        -0.023215   0.005305  -4.376 2.21e-05 ***
```

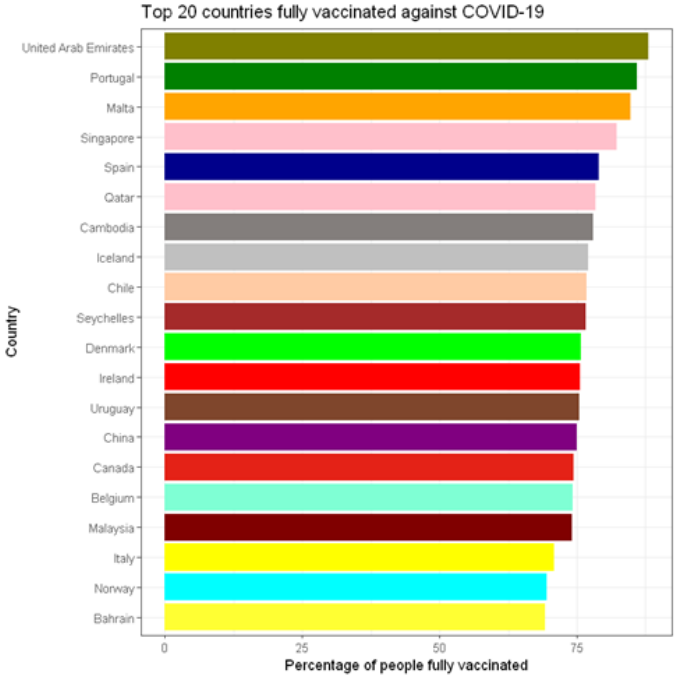
```
Call:
lm(formula = infr ~ tavg + prcp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.087024 -0.030041 -0.006784  0.024853  0.184154

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1066196  0.0105850  10.073  < 2e-16 ***
tavg        -0.0012188  0.0005085  -2.397   0.0177 *
prcp        -0.0051162  0.0009447  -5.416 2.21e-07 ***
```

The top four plots show each of countries averaged temperature and daily precipitation during 2020 and 2021 compared with the countries Covid-19 infected rate and fully vaccinated rate per person. Precipitation has a stronger (negative) correlation than temperature. From fitting a simple linear model with the residuals shown in the bottom two plots – it is apparent from the p-values that there is a statistically significant relationship for each of the climate variables with the target variables. The estimates show that there is a more noticeable trend between the climate data and infected rate rather than fully vaccinated rate. In both cases precipitation is a much larger factor than temperature.

Stringency Index:



The scatter plot was used to see whether there is any relationship between stringency index and Covid-19 vaccinations rate. These two variables have no relationship, has no direction and as the dots are scattered all over it has a non-linear form. There were no specific clusters even when countries were divided into regions. If we compare the African countries in the red dots with the European countries in the purple, they have similar stringency index ranging from 15 to 80 with a huge variation in vaccination rates. This clearly shows that the data need to be analysed further to uncover any relationship between the stringency index and vaccination rates. The below bar chart shows the top 20 countries that have fully vaccinated.



What techniques you did use?

Part one: The World Factbook

There were multiple techniques used when scraping the World Factbook website. First, the dynamic nature of the website meant the HTML script retrieved by rvest differed from the HTML script that was seen by the user in their browser. Therefore, a selenium server was set up using a google chrome web driver and the selenium program. The set-up of the server took place in the command prompt, not the notebook, because Java was required to set up the server. A port number was used to establish a connection between R and the server. This was done through the package of Rselenium. The server was then used to navigate to a URL and then pass the HTML code it retrieved into rvest.

A web browser selector was used to find tags that could select the desired values from the HTML code. These tags were used with rvest to extract the variables either in table format or as text. The stringer package was then used to extract the values from paragraphs or remove commas and dollar signs. This was done in order for the variables to be converted into a usable form, i.e., a numeric value or double value.

After Scraping all the variables, four data frames were produced.

- Data frame one contained the key variables: urbanisation rate and urbanisation yearly rate change.
- Data frame two contained the key variable: population.
- Data frame three contained the key variable: size in squared kilometres.
- Data frame four contained the key variable: real GDP per capita.

These four data frames were merged into a single data frame using the merge function and the joining variable country_name. After the data was merged, it was mutated to create a new density variable. This was created by dividing the population by size and then rounding the values to 2 decimal places. The country code column was created and bound to the data frame using the country code package. The international organisation created these codes for the purpose of standardisation. Country names are not standardised and can change from source to source, unlike country codes. Finally, the final data frame was saved in a CSV file. Below is a screenshot of the final data frame.

country	urbanisation_rate	urbanisation_yearly_change	gdp_per_cap	size	population	density	country_code
Afghanistan	26.3	3.34	2065	652230	37466414	57.44	AFG
Albania	63.0	1.29	13965	28748	3088385	107.43	ALB
Algeria	74.3	1.99	11511	2381740	43576691	18.30	DZA
American Samoa	87.2	0.26	11200	224	46366	206.99	ASM
Andorra	87.9	0.11	49900	468	85645	183.00	AND
Angola	67.5	4.04	6670	1246700	33642646	26.99	AGO
Anguilla	100.0	0.47	12200	91	18403	202.23	AIA
Antigua and Barbuda	24.4	0.87	21910	443	99175	223.87	ATG
Argentina	92.2	0.97	22064	2780400	45864941	16.50	ARG
Armenia	63.4	0.23	13654	29743	3011609	101.25	ARM
Aruba	43.9	0.77	37500	180	120917	671.76	ABW
Australia	86.4	1.27	49854	7741220	25809973	3.33	AUS
Austria	59.0	0.68	56188	83871	8884864	105.93	AUT
Azerbaijan	56.8	1.38	14404	86600	10282283	118.73	AZE
Bahamas, The	83.4	1.02	37101	13880	352655	25.41	BHS
Bahrain	89.6	1.99	45011	760	1526929	2009.12	BHR
Bangladesh	38.9	2.88	4754	148460	164098818	1105.34	BGD
Barbados	31.2	0.46	15639	430	301865	702.01	BRB
Belarus	79.9	0.28	19150	207600	9441842	45.48	BLR
Belgium	98.1	0.38	51934	30528	11778842	385.84	BEL
Belize	46.2	2.30	7005	22966	405633	17.66	BLZ

Part two: World Health Organisation, Oxford Covid-19 Government Response Tracker and OECD

Multiple techniques were used to extract information for the Covid-19 data set and OECD member states data set. First, the data was collected from Oxford Covid-19 Government Response Tracker database using an API. get_json_time() function creates an API URL query for the latest date, get_data_time() is used to retrieve the data. The country code and stringency columns were selected from this data set to create the desired data frame.

The World Health organisation's information was extracted from two CSV files. These two CSV files were extracted using the get_csv function. The first data frame created from a CSV file contained information about country name, region, covid cases, covid related deaths. This data frame was cleaned by selecting the relevant columns, renaming these columns, and filtering the values. The country codes were added to the data frame via the countrycode function. Some country names were not matched with the country codes like Saint Martin, which had to be replaced with Sint Marteen. Bonaire, Saba and Sint Eustatius shared the same country code – BES, so these rows had to be summarised. Rows with country names that did not exist were dropped.

The second data frame created from a CSV file contained information about vaccination rates. The columns related to vaccinations were selected. These columns were then renamed and joined with the first data frame using left_join() by country code and country name.

These three data frames were combined using country codes. This data frame was finally saved into a CSV file for further use. The final dataset variables were:

- country_code - ISO-3 Character
- country_name - Country name
- region - Country in a specific region, as the WHO Member States are grouped into six regions
- total_cases - Total number of COVID-19 cases
- total_deaths - Total number of deaths due to COVID-19
- total_vaccinations - Total number of vaccinations administered for COVID-19
- vaccinated_1dose - Total number of people vaccinated by a single dose
- fully_vacccinated - Total number of people fully vaccinated
- total_vaccinations_per100 - Total number of vaccinations per 100 people
- percent_full_vac - Percentage of people fully vaccinated
- first_vaccine_date - Date of first Vaccination
- last_updated_date - Last updated date
- stringency - Stringency index is the measure from 0 to 100 based on ordinal calculations of people's behaviour in terms of lockdowns during the pandemic. Like social distancing, facemasks, hand hygiene, banning or limited public place gatherings, closures of schools, parks and workplaces, economic aids, prioritising the vaccinations are some of the measures taken to minimise the spread of infection and mortalities in the community.

Julia is used to scrape a members' list from the OCED webpage. The HTML script was retrieved using the get() function. The HTML script was parsed into a useable form using gumbo. A web browser selector extension from Cascadia was used to find a tag that could select the desired values from the HTML code. The tag was then used in the function eachmatch() to return an array of all 36 member

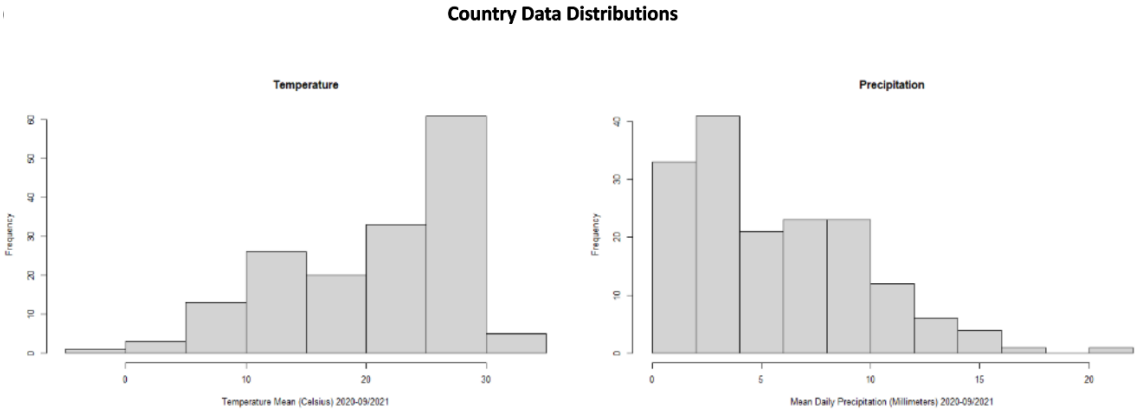
states. Country codes were added using the WorldBankData package and this data frame was saved as a CSV file.

Part 3: Global Historical Climatology Network daily (GHCNd)

Data from the GHCNd came as observations for each station, day, and measurement type. To get it into the desired format of daily average precipitation and temperature by country:

1. Obtained all data during Covid-19 period (2020 and 2021).
2. Append the two year-spanning files.
3. Filter set to only include temperature and precipitation measurement types.
4. Select station, date, measurement type and value columns.
5. Mutate a column of FIPS country codes which is a substring of each station ID.
6. Group by country and measurement type and aggregate mean value.
7. Transform to a wide data frame: measurement type to temperature and precipitation columns.
8. Join with ISO3 country code dictionary by FIPS codes so it can join the main data frame.

The distributions of the final average daily precipitation and temperature by country were:



These distributions are as we may expect – there are no obvious univariable novelties to consider.

Part 4: Converging the 3 data frames into the final Data frame

The data was processed using R. The packages used to clean, process, and finalise the data were ''dplyr'', ''tidyverse'', and country code.

First, the Covid, Population, Climate and OECD data sets were loaded into R using read.csv. the two data sets, Covid and Population, were joined together into one data frame using the ''inner_join'' function with the joining variable being country code. This is because country codes are consistent across all datasets, while the country's names vary. The Climate data set was then processed by changing the column names from name to ''country_name'' and iso3166 to ''country_code''. The values within the climate data set were rounded to 2 decimal places. This was done using the ''mutate_if'' function because there were NAs within this data frame. The Climate data set was then joined to the Covid and population data set using ''left_join'' function, with the joining variable being country code. These steps produced a final dataset with 220 rows and 21 columns.

Three columns were created and added to the final data frame using the Mutate function. These were:

- infection rate per 1000
- Deaths per 1000
- Total vaccinations per 1000

These were created instead of scraped from the WHO website because this helps to maintain consistency across the data frame. These variables were rounded to 2 decimal places to ensure New Zealand's values were not rounded to zero.

The final dataset also had a few duplicate rows which had to be removed. This was done using the ‘'Distinct'' function. Finally, the OECD data frame was joined to the final data frame using the ‘'left_join'' and ‘'case_when'' functions. This produced a column to show whether a country is in the OECD or not. This variable was added for the purpose of being able to further refine the dataset if required.

What difficulties you had to overcome to wrangle the data sources into the target data model?

- The Climate data was daily for each station, so countries needed to be identified and grouped and days averaged. Temperature and precipitation data were in specific rows with an identifier, so these rows needed to be filtered.

- The World Factbook did not use the required ISO3 country codes. Therefore, the country's name had to be converted into the relevant country codes. This was done through a function called countrycode. Second, the population density per square kilometre was not a statistic produced by the World Factbook. However, the population and country size variables were available, therefore, these variables could be mutated to create density. The third problem was the dynamic nature of the website. To resolve this issue, a selenium server was set up using a web driver and a java program. This server was connected to R through rselenium and a port number.

- The website WHO was a dynamic website. Even though a significant amount of time was spent trying to scrape this website, the information could not be extracted. Meanwhile, an issue arose with the Jupyter Notebook where the R Kernel stopped working, which had to be reinstalled. There was a problem with data quality as we initially used Our World in Data (OWID) for theCovid-19 dataset. However, this dataset was inconsistent. Therefore, we chose to extract the information from the WHO

- The four data sets were not similar and not all of them used the same reference column. The covid and population datasets were using three-letter country codes, while the climate data was using two-letter FIPS country codes. We had to find and use additional country code resources to be able to join different datasets together.

- Data obtained was not always as current or spanned exactly the same time as others which caused some inconsistencies between variables in the data. To overcome this: dates of data records were also recorded where possible and necessary.

Limitations

The Final dataset contains many missing values for variables: precipitation and temperature. This means there is less data to find relationships using the climate data. This could be greatly improved in future if reliable climate data does become available for these countries. Modelling work involving the climate data would require methods tolerant of missing values, or pre-processing steps such as data deletion or imputation. Also, the final dataset includes only the total number of cases, deaths, and vaccinations for each country until September 2021. Therefore, because of the fast-moving nature of Covid responses, the analysis may become quickly outdated.